Encyclopedia of Cognitive Science

WILEY

Bartlett, Frederic Charles

Introductory article

Henry L Roediger, Washington University, St Louis, Missouri, USA

CONTENTS

Frederic C. Bartlett (1886–1969) was a distinguished British psychologist who spent most of his career at Cambridge University. He is chiefly known today for his book Remembering: A Study in Experimental and Social Psychology, *which laid the foundation for schema theory.*

INTRODUCTION

Frederic C. Bartlett (1886–1969) was a distinguished British psychologist who spent most of his career at the University of Cambridge. He was trained as an experimental psychologist and became the most prominent English psychologist of his generation through the influence of his writings, his work on applied problems, and the great students he trained who continued work in his tradition. He is chiefly remembered today for his 1932 book, *Remembering: A Study in Experimental and Social Psychology*, which laid the foundation for schema theory and pioneered the study of memory distortions. Bartlett was knighted in 1948 for his great accomplishments, which are described briefly below.

BIOGRAPHICAL DETAILS

Early Life and Education

Bartlett was born in Stow-on-the-Wold, a small country town in Gloucestershire. His father was a successful businessman who made shoes and boots, but the educational opportunities in town were slim. A severe illness when he was 14 years old made it impossible for him to attend a boarding school, so young Bartlett stayed in Stow and educated himself with the aid of his family (his father had a great library) and friends. He eventually took a distance course at the University of London and settled on psychology, logic, sociology, and ethics as topics of study. He received an MA from London in 1911 and continued to Cambridge, where he came under the influence of W. H. R. Rivers, Cyril Burt, and C. S. Myers. He obtained his doctorate with first-class honors in 1914, just as Burt decided to leave Cambridge. Myers then offered Bartlett Burt's vacated position, so Bartlett stayed in Cambridge.

The First World War and Bartlett's Development

The First World War broke out soon after Bartlett took up his position at Cambridge. Most of Bartlett's colleagues left to aid the war effort, but poor health prevented him from joining them. However, the absence of people senior to him thrust him into the role of leading the psychological laboratory. He threw himself into teaching and began writing a book based on his dissertation, although it would not appear for many years. Much of his research during this time focused on practical problems driven by the war, such as detecting weak auditory signals in noise (to help with the problem of detecting German submarines). His war work eventually culminated in a book, *The Psychology of the Soldier* (1927).

After the war, Rivers and Myers returned to Cambridge and became Bartlett's associates. However, in 1922 Rivers died suddenly and Myers retired, so Bartlett became director of the Cambridge Laboratories and built them into a research powerhouse over the years. In the 1920s Bartlett's research turned to social anthropology, an early interest, and he wrote *Psychology and Primitive Culture* (1923). His international reputation expanded and he came to know distinguished psychologists from around the world.

REMEMBERING

In 1932 Bartlett published his great book, which is still in print today. *Remembering* actually grew out of his dissertation experiments begun in 1913, so the gestation period was nearly 20 years. The book

introduced a very different tradition for studying memory from the scientific methods of Ebbinghaus with their emphasis on careful control and measurement of memory in rather unnatural conditions. Bartlett's methods were casual, almost anecdotal, compared with those of Ebbinghaus, yet he uncovered powerful truths about remembering that reverberate through the field even today. Bartlett tested people under fairly relaxed conditions and his 'data' consisted largely of verbal reports with which he sprinkled his writing. (*See* **Ebbinghaus, Hermann**)

The early chapters of *Remembering* actually consist of studies of perceiving. The great middle part of the book is directly concerned with memory. The last section of the book deals with social and anthropological factors in cultural transmission. The general thrust of the book is to emphasize the constructive nature of cognition. Perceiving, remembering, and all of thinking involve the individual as part and parcel of the cognitive process. For example, in perceiving an ambiguous stimulus that is briefly presented, one's past background and experience determine what is perceived as much as (or even more than) the stimulus that is presented.

Bartlett devised two methods to study remembering: repeated reproduction and serial reproduction. In his most famous work he read a native American folk tale, *The War of the Ghosts*, to his British participants and then later tested their memories. This bizarre and supernatural story was usually read twice, aloud. In the repeated reproduction technique Bartlett would have his listeners recall the story after an interval of about 15 min. Next he would test their memory for the story at various later times, but with no further presentations of the story. Thus, repeated reproduction involves the same individual repeatedly reproducing the story, as the name implies. Bartlett's interest centered on how people remembered the story and how their memories would change over time and repeated retellings.

Not surprisingly, people remembered less about the story as time passed – their reports became increasingly short. Of more interest was the content of what they did remember and what these recollections indicated about the workings of memory. Besides becoming shorter, the stories became simpler, supernatural elements dropped out and other bizarre items would be reinterpreted. Bartlett called this process 'rationalization' because people added material to explain unnatural elements, or dropped them out altogether if they did not seem to fit the person's past experience. Rationalization over repeated retellings caused the story 'to be robbed of all its surprising, jerky and inconsequential form, and reduced it to an orderly narration' (p. 153 of the 1932 edition of *Remembering*). Bartlett also referred to the 'effort after meaning' that occurred in his perception and memory experiments, whereby people try to convert or recode elements that are difficult to perceive or understand into forms that can be comprehended. People try to impose structure and order to understand the world around them, even when their experience does not conform neatly to their prior categories.

Bartlett wrote that 'the most general characteristic of the whole of this group of experiments was the persistence, for any single subject, of the form of his first reproduction', and the use of 'a general form, order and arrangement of material seems to be dominant, both in initial reception and in subsequent remembering' (p. 83). He named this general form that people use to encode and to remember experiences a 'schema', a term now used throughout the cognitive sciences. A schema is a general organization of a story of a typical event. So, for example, many old films about the American wild west follow a schema involving 'good guys', 'bad guys', crisis, and resolution. The schema can aid encoding and retention of details that are consistent with it, but details that do not fit may be forgotten or distorted to fit the schema. In remembering *The War of the Ghosts* some English participants seemed to use the schema of a fairy tale, a genre to which they were more accustomed. Some even tacked on a moral at the end of the story.

The method of serial reproduction, the other major technique Bartlett introduced, is like the children's game of rumor or telephone. One person hears *The War of the Ghosts* (or is exposed to some other material) and recalls it after a set period. This person's recollections are then read to a second person, who recalls it in turn. This second recall is then read to a third person for later recall, and so on, through as many instantiations as desired. The changes in recall across repeated tests using the serial reproduction method are much greater than those in repeated reproduction, although Bartlett thought the same types of memory processes were at work (but in greater force). The serial reproduction technique involves a human chain, and if there were to be one weak link in the chain – someone who was wildly inaccurate in recall – then there would be no hope of a person later in the chain correcting the false memory of the material because that person would never have been exposed to the correct version. Reading through the lengthy samples that Bartlett provided in

Remembering (chapters 7 and 8) leads to agreement with his basic claims. The serial reproduction technique was later championed by psychologists studying the transmission of rumors.

The serial reproduction technique also served, Bartlett believed, as a useful analogy for the way information might be handed down from one generation to another within a society or even for the spread of ideas from culture to culture. He dealt with these issues in some detail, although with anecdotal evidence, in the last section of his book.

Bartlett's *Remembering* provides many interesting ideas and quotable passages. The book was well known at the time, but his research tradition did not really catch on. Part of the reason for this is that, in his hands, the research was more anecdotal than experimental (despite the subtitle of his book). He has been criticized for this lack of careful empirical research to document his points, and it was not until recently that a successful replication of his basic findings using the repeated reproduction technique appeared in print. Bartlett's book came to the forefront of the field when Neisser adopted Bartlett's theme of the constructive nature of cognition for his 1967 text, *Cognitive Psychology*, which helped to usher in the cognitive revolution in psychology. In the early 1970s psychologists such as Elizabeth Loftus, John Bransford and Marcia Johnson became interested in errors of memory and Bartlett's ideas were invoked and his book was once again read by a new generation.

Throughout *Remembering*, Bartlett's message ran counter to the idea that memory should be conceived of as static memory traces that are called to mind and read off in a more or less accurate fashion. Memory does not work like a video recorder, tape recorder, or computer. In his words, 'Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces' (1932, p. 213). Rather, 'remembering appears to be far more decisively an affair of construction rather than reproduction' (p. 205). 'It is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a mass of organized past reactions or experiences' (p. 213). This credo still guides the field today in many ways.

LATER CONTRIBUTIONS

The Second World War confronted psychologists with many more practical problems to be solved. Bartlett and Kenneth Craik worked during the war on problems of skill acquisition, and Bartlett served on the Royal Air Force's Flying Personnel Research Committee, focusing his work on pilot training. They also studied related topics such as the effects of fatigue on performance. When Craik was tragically killed in an automobile accident two days before the war in Europe ended, Bartlett felt the loss keenly, because the men had become best friends as well as close collaborators.

After the war, Bartlett applied notions of skill learning to those of higher-order thinking, capitalizing on the insight that just as experts in a physical skill develop their exquisite expertise after many hours of practice, so do experts in thinking skills – problem-solving, reading X-ray graphs and so on. In 1958 he published *Thinking: An Experimental and Social Study*, which provided his insights on these topics. However, this book did not enjoy the earlier success of *Remembering*, although it too is an interesting treatise.

Bartlett retired from the chair of experimental psychology in Cambridge in 1952, but maintained his affiliation with the applied psychology unit which he had helped to found. His many students frequently called on him for advice and he continued to serve on national committees. Despite his early health difficulties, he remained generally robust in his later years, although he was bothered by hearing loss. He died after a brief illness on 30 September 1969.

CONCLUSION

Frederic Charles Bartlett wielded tremendous influence both nationally and internationally. Some commentators have remarked that this influence was out of proportion to his actual scholarly work. His contributions were good, but only one (*Remembering*) was of enduring importance. Rather, Bartlett's own charismatic character drew people to him and established his leadership, the power of his personality infecting those around him with his wit, his wisdom, his generosity, and his good nature.

Knighted, in 1948, Sir Frederic Bartlett received many other honors, including honorary doctorate degrees from seven universities in six countries. In Britain he was elected to the Royal Society in 1932 and received its Baly and Huxley medals in 1943. He was awarded the Royal Medal in 1952, the highest distinction a scientist in Britain can receive. In the USA Bartlett was elected to the American Philosophical Society, the National Academy of Sciences (as a foreign fellow) and the American Association of Arts and Sciences. Bartlett was a towering figure of twentieth-century psychology, and in recent years the study of human memory has come around to the approach he advocated so strongly in the 1930s.

Further Reading

Allport GW and Postman L (1947) *The Psychology of Rumor*. New York, NY: Holt.

Bartlett FC (1916) An experimental study of some problems of perceiving and imaging. *British Journal of Psychology* **8**: 222–266.

Bartlett FC (1923) *Psychology and Primitive Culture*. Cambridge, UK: Cambridge University Press.

Bartlett FC (1927) *Psychology and the Soldier*. Cambridge, UK: Cambridge University Press.

Bartlett FC (1932) *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.

Bartlett FC (1958) *Thinking: An Experimental and Social Study*. London, UK: Allen & Unwin.

Bergman E and Roediger HL (1999) Can Bartlett's repeated reproduction experiments be replicated? *Memory and Cognition* **27**: 937–947.

Kintsch W (1995) Foreword. In: Bartlett FC, *Remembering: A Study in Experimental and Social Psychology*. [reprint] Cambridge, UK: Cambridge University Press.

Neisser U (1967) *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts.

Roediger HL (2000) Sir Frederic Charles Bartlett: experimental and applied psychologist. In: Kimble GA and Wertheimer M (eds) *Portraits of Pioneers in Psychology*, vol. 4, pp. 149–161. Mahwah, NJ: Lawrence Erlbaum.

Bloomfield, Leonard

Introductory article

Stephen R Anderson, Yale University, New Haven, Connecticut, USA

CONTENTS

Leonard Bloomfield (1887–1949) was an American linguist whose contributions to general linguistics as well as to the study of a number of language families make him one of the central figures in the history of this field of study. His name is virtually synonymous with the American Structuralist approach to language through the 1950s.

INTRODUCTION

Few figures in the history of linguistics stand out as prominently as incarnations of their time and place as Leonard Bloomfield. Linguistics in America from the publication of his book *Language* in 1933 until the development of Generative Grammar in the 1960s is practically identifiable with his approach. This was in part because he represented the desire of linguists to be treated seriously as pursuing a scientific discipline with its own methods, goals, and results. Edward Sapir and Franz Boas, other major figures in the history of linguistics whose activity overlapped with Bloomfield's, studied languages within the theoretical framework of anthropology or psychology. Others studied particular languages and language families for their own sake. In contrast, Bloomfield thought of himself as a linguist, studying language for its own sake. In the process, he aligned himself with contemporary positions in philosophy (positivism) and psychology (behaviorism) that were seen as paving the road to a genuinely scientific view of language, in contrast to humanistic approaches.

As a result, linguistic theory as it developed during this time was largely formed either through Bloomfield's own work or by what his students and colleagues did in the name of his views. Although a good deal of 'post-Bloomfieldian' linguistics was not particularly close to Bloomfield's own positions, it was nonetheless felt that a scientific approach to language could be largely identified with the task of working out Bloomfield's views. American structural linguistics largely *was* Bloomfieldian linguistics.

BLOOMFIELD'S LIFE

Leonard Bloomfield was born in Chicago in 1887, and moved to Elkhart Lake, Wisconsin, in 1896 when his father bought a resort lodge there. During his childhood in Wisconsin he came into contact with the Menomini people and their language, a member of the Algonquian family, which would occupy much of his later attention. His father's brother, Maurice Bloomfield, was a noted Sanskritist and no doubt had an influence on Bloomfield's subsequent interest in this language and its grammatical tradition.

Bloomfield entered Harvard in 1903, received his AB degree in 1906, and went to the University of Wisconsin for graduate study. One of the first scholars Bloomfield met there was Edward Prokosch, one of the major names in Germanic studies, who interested him in historical work within the framework of Indo-European linguistics. In 1908 he moved to the University of Chicago, where he received his PhD in 1909 for a thoroughly traditional, philologically oriented thesis: *A Semasiological Differentiation in Germanic Secondary Ablaut*.

Most of Bloomfield's academic career was spent as a teacher of German: although he practiced general linguistics within the limits of such positions, it was not until he came to Yale in 1940 that he actually held a professorship of linguistics, as opposed to German. His first job was at the University of Cincinnati, from which he moved to the University of Illinois in 1910. He was told early on that while his department was enthusiastic about promoting him, a competing candidate had the edge by virtue of having studied in Germany, and that if Bloomfield wanted to get ahead, he would have to study in Germany too. Taking this advice to heart, he spent the year 1913–1914 in Leipzig and Göttingen, studying with such notable Indo-Europeanists as Leskien and Brugmann. In the process, he rubbed elbows (quite literally) with a number of other students who would later be important names in linguistics, such as Nikolai Trubetzkoy.

In 1914 he published his first book, *An Introduction to the Study of Language*. This general survey was based solidly in the introspectionist psychology of Wundt, influential at the time but virtually the antithesis of the approach he would later champion. This book is little read today, but interesting for understanding the later development of Bloomfield's views on, for instance, morphology (inflection and word formation).

The outbreak of the First World War led to an immediate and precipitous decline in German studies in the USA, and Bloomfield no doubt had a certain amount of time on his hands as a teacher of German. During the war, he worked with a student at Illinois who spoke Tagalog, the principal indigenous language of the Philippines. This work resulted in a book *Tagalog Texts with Grammatical Analysis*, which contains an extensive grammar of the language, though one that is difficult to use as a result of Bloomfield's explicit, conscious avoidance of traditional categories and terminology in describing a system far from the familiar structure of Indo-European languages.

In 1921, he moved to Ohio State University (not having been offered tenure at Illinois, despite having studied in Germany!) where he immediately became a full professor of German. Here one of his colleagues was Albert Weiss, a major figure in the early development of behaviorism, whose views on the mind largely determined Bloomfield's own for the rest of his career. In 1927, he was invited to the University of Chicago, again in the German department. Here one of his colleagues was Sapir, in the anthropology department: the two were professional collaborators (but uneasy friends) in the emerging discipline of general linguistics.

In addition to his work on German and general linguistics, Bloomfield was also occupied during this time with comparative Algonquian studies. This was not just an escape from the rigors of Germanic linguistics. Bloomfield brought the methods of Indo-European studies to work on American Indian (and by extension, other indigenous) languages. This was unusual: others had suggested that the methodology of comparative reconstruction, developed with respect to Indo-European, was substantially dependent on the fact that several languages (Vedic Sanskrit, Gothic, Homeric Greek, Hittite, Old Church Slavonic, etc.) of the family are attested at considerable time depth, and that the same techniques would not be effective in dealing with unwritten languages. Bloomfield showed that the methodology could be applied in establishing the comparative grammar of Algonquian (especially its central branch, based on data from Fox, Cree, Menomini and Ojibwa).

The clinching demonstration of this came when, in working out the system of consonant clusters in the system ancestral to these languages, he was left with one correspondence set that did not fit any known combination of segments. For this, he postulated an additional proto-Algonquian cluster which he wrote as *çk. Later, however, as data from other languages and dialects of the family became available, it became clear that exactly the words for which Bloomfield had posited this cluster showed consistent unique reflexes across the family; and indeed other words came to light that illustrated the same correspondence set. This was widely seen as providing a dramatic confirmation of the correctness and generality of the assumptions of comparative linguistics – as dramatic, in its way, as the confirmation provided by the analysis of Hittite for the prior assumption of 'laryngeal' segments in the phonology of proto-Indo-European.

Bloomfield's role in the professionalization of linguistics in the 1930s and 1940s was tremendously important. He worked hard for the establishment of the field's distinctive institutions, especially the Linguistic Society of America, its journal *Language*, and the annual summer institutes which it organized (at the time, virtually the only occasions when linguists gathered in significant numbers). In 1940 Bloomfield was invited to Yale, after the death of Sapir (who had preceded him there as Sterling Professor of Linguistics). He never really settled in New Haven: he and his wife were both attached to Chicago, and she suffered from severe depression when they left. To this was of course added the dislocation provoked by the Second World War, but he turned this to advantage, working actively in the army's Intensive Language Program during the war years and thereby providing useful work for a new generation of descriptive linguists. In 1946, Bloomfield suffered a severe stroke from which he never really recovered. He died in 1949.

BLOOMFIELD'S VIEW OF LANGUAGE AND THE MIND

Bloomfield's first writing dealing with general issues in the study of the mind was in his 1914 book *An Introduction to the Study of Language*, but he soon lost confidence in the explanatory power of the Wundtian psychology underlying that book. That point of view was soon supplanted by an ardent embrace of the behaviorist (or 'mechanist'

as he preferred to call it) psychology of his Illinois colleague Weiss. This was already evident in his 1926 paper, 'A set of postulates for the science of language' (*Language* **2**: 153–164), intended as a fairly direct calque on a paper by Weiss laying out an axiomatization of psychology, although it also shows considerable influence of the study of the Sanskrit grammatical tradition. More important perhaps than Bloomfield's intent to emulate Weiss's point of view, the paper's terminology in referring to psychological factors is enthusiastically behaviorist in tone, as when he defines the meanings of utterances as their 'corresponding stimulus–reaction features'.

A product of his times, Bloomfield's notion of a scientific explanation was one based solely on propositions relating observable events by principles of logic and mathematics alone. Throughout his career, he repeatedly ridiculed 'mentalistic' explanations as they appeared in the supposedly scientific literature on language and linguistics. Subsequent commentators (as well as many of his contemporaries) took this to imply a rejection of the existence of a mental life, but this is not at all what he intended. Rather, he meant to reject the notion that linguistic (or any other) phenomena are causally affected by a mysterious and unobservable entity (the 'mind') whose principal property is its nonobedience to normal laws of physical causation.

Early behaviorists insisted that if the mind were to be taken seriously as an object of scientific inquiry, it must be reduced to special cases of the activity of some observable physical system. There are, of course, alternatives to this, as the 'cognitive revolution' has made clear, but for Bloomfield, an attempt to ground the study of language in the properties of mental and cognitive organization seemed like an effort to evade the constraints of rational inquiry. Considering the excesses of romanticist approaches to the nature of language, and indeed the introspectionist psychology Bloomfield himself followed in his early years, these concerns were not entirely illusory. For Bloomfield, the only sensible alternative to antirational speculation about the mysteries of the soul was a denial of the scientific relevance of anything but the material embodiment of mind.

This restriction of scientific discourse, including all talk about 'meanings' apart from the framework of observable stimuli and responses, was not, as it is commonly seen, intended to deny that minds and meanings exist, or even that they might play a central role in human life. His point, rather, was that in the present state of science we have no way of cashing these notions out in strictly observable terms, and thus that talk about them necessarily falls outside real science. He did believe that a satisfactory account of meaning would need to be based on an encyclopedic knowledge of the world and its laws, down to the last detail – something obviously well beyond the scope of linguistics or perhaps any science. This belief that meaning ultimately has a comprehensive explanation in terms of sufficiently minute details concerning (potentially observable) electrochemical events within the nervous system is just as much a matter of faith on his part as the 'mentalist' picture is for others.

A language can be seen as a system that relates sounds and meanings, and it would thus seem that some account of meaning is necessary to linguistics. For this reason, Bloomfield introduces a mechanistic picture that seems naive even by comparison with other behaviorist work, but he also denies that the difference between such a view and the mentalist one has any significance. For him, the structural properties of language can be investigated perfectly well even if meaning is simply reduced to the status of a postulate, not treated in its substance (whatever that might be). In a 1944 article, he compared his 'antimentalism' to 'a community where nearly everyone believed that the moon is made of green cheese, [in which] students who constructed nautical almanacs without reference to cheese would have to be designated by some special term, such as *non-cheesists*.'

With complete impartiality, Bloomfield maintained that the concrete properties of sound are also, strictly speaking, irrelevant to an understanding of language; and thus neither phonetics nor semantics played a role in the sort of structuralist accounts he advocated. His actual practice involved appeals to our understanding both of sound and of meaning that were not significantly different from those he opposed: he simply maintained that these matters were not essential to an understanding of linguistic systems.

Bloomfield's views on the nature of mind and cognitive organization were surely much too simplistic, as generations of commentators have maintained. Nonetheless, his repeated insistence that the methodology and results of linguistics are independent in principle of any particular theory of psychology (his or another) should be taken at face value. The radical behaviorist views he advocated had much less influence on his own practice with regard to central areas such as phonology and morphology than his pronouncements would have on his own students and their immediate successors. Bloomfield was a solid scholar in a number of

areas, and one whose intuitions about linguistic structure took him far beyond the limitations of his stated basic principles.

Further Reading

Anderson SR (1985) *Phonology in the twentieth century*. Chicago: University of Chicago Press (pp. 250–276).
Bloomfield L (1914) *An Introduction to the Study of Language*. New York, NY: Henry Holt.
Bloomfield L (1933) *Language*. New York, NY: Henry Holt.
Hall RA (ed.) (1987) *Leonard Bloomfield: Essays on his life and work*. Amsterdam: Benjamins.
Hall RA Jr. (1990) *A life for Language*. Amsterdam: Benjamins.
Hockett CF (ed.) (1970) *A Leonard Bloomfield Anthology*. Bloomington, IN: University of Indiana Press.
Hymes DH and Fought J (1981) *American structuralism*. The Hague: Mouton.
Matthews PH (1993) *Grammatical theory in the United States from Bloomfield to Chomsky*. Cambridge, UK: Cambridge University Press.

Brazier, Mary A. B.

Introductory article

Mary Brown Parlee, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

Mary A. B. Brazier was an authority on electroencephalography who after World War II pioneered the use of correlational techniques and high-speed computers to study brain activity and behavior. She also authored influential books and articles on the history of neurophysiology which stimulated historical interest in the brain and behavioral sciences.

INTRODUCTION

Mary A. B. Brazier (1904–1995), an authority on electroencephalography, pioneered the use of computers to study brain physiology and brain–behavior relationships. She was also an influential editor, organizational leader and historian of neurophysiology (Figure 1). In collaboration with colleagues at Harvard University and the Massachusetts Institute of Technology, she was the first to use modern high-speed computers and correlational techniques for frequency analysis to study the electrical activity of the brain in relation to behavior. Author of a classic textbook, *The Electrical Activity of the Nervous System* (see Further Reading section), she edited several volumes of Macy Foundation-sponsored conferences on brain and behavior. These brought the latest research to the attention of a broad scientific readership at a time (the late 1950s and early 1960s) when the modern cognitive and neurosciences were emerging as new sciences from their interdisciplinary roots. Brazier's influence on these fields continued through her editorship (1975–84) of the journal *Electroencephalography and Clinical Neurophysiology* (of which she was a founder in 1949) and through her leadership in the International Brain Research Organization (IBRO) and other national and international scientific organizations concerned with brain and behavior. She also authored numerous articles and two books on the history of neurophysiology in the seventeenth, eighteenth and nineteenth centuries. Prior to her move in 1940 from England to the USA, where she made her reputation as a brain researcher and historian of science, Brazier conducted award-winning research in endocrinology. In the USA she was usually the first and only woman in the post-World War II scientific circles of neurophysiologists, neurologists, electrical engineers and mathematicians among whom she traveled and thrived.

LIFE AND WORK

Mary Agnes Burniston Brown, known as 'Mollie' to her friends, was born in Weston-super-Mare, near Bristol, England on 17 May 1904, the second of two children (she had an older brother) in a Quaker family. As a child she developed what would be lifelong loves of the sea and of science, and after attending school at Sidcot she entered the all-women Bedford College of the University of London, where she studied physiology, obtaining her BSc in 1926. In 1927 she took up an appointment as a Research Fellow in the Imperial College of Science and Technology, working with Frederick Golla at the Maudsley Hospital on separation of the products of protein hydrolysis. She continued this research for more than a decade, obtaining a PhD in biochemistry from the University of London in 1930. In 1928 she married electrical engineer Leslie J. Brazier, with whom she had a son, Oliver.

Mary Brazier's research at the Maudsley Hospital focused on measurement of the electrical changes in the skin which occurred in patients with diseases of the thyroid gland. The technique for measuring these changes, which Brazier named the *impedance angle test*, proved useful for diagnosing thyrotoxicosis. It led to her receiving research awards in 1934 from the (British) Institute of Electrical Engineers and the American Association for the Study of Goiter. A few years later, also at the Maudsley Hospital, research by W. Grey Walter demonstrated the diagnostic usefulness of another

Figure 1. Mary A. B. Brazier (1904–1995). Photo supplied courtesy of the History and Special Collections Division of the Louise M. Darling Biomedical Library, UCLA.

electrophysiological measure, namely the electroencephalogram (recordings of electrical activity of the brain through the unopened scalp and skull of humans). EEGs could be used to locate tumors of the brain. This finding, together with a report that petit mal epilepsy was associated with changes in the EEG, generated considerable interest in EEGs and their potential clinical applications. References to these and other articles – more than 100 are covered – can be found in Brazier's 1958 account of the development of concepts relating to electrical activity of the brain (see Further Reading).

In 1940, with London under heavy attack, Mary Brazier and her son moved to Boston, while Leslie Brazier remained in England. With a Rockefeller fellowship she secured an appointment as a neurophysiologist at Massachusetts General Hospital (MGH), in Harvard Medical School's Department of Psychiatry, headed by Stanley Cobb (she later held appointments in the Departments of Anesthesia and Neurology). One of the earliest EEG laboratories in the USA had been established at MGH in 1937 by Robert Schwab, and Brazier soon found her way there. She remained at MGH, later heading her own Neurophysiological Laboratory, for the next 20 years. During the war she collaborated with her new colleagues in national defense research on a variety of topics, including peripheral nerve injuries, aircraft pilot selection, war neuroses, electromyograms and muscle dysfunction in poliomyelitis. Her first publication on the subject of EEGs appeared in 1942, and during the next 7 years such research became the focus of her work, including the characteristics of normal EEGs: comparison of the EEGs of psychoneurotic patients and normal adults; and the effects of blood sugar levels, anoxia and various anesthetic agents on EEGs.

The first International EEG Congress was held in London in 1947, and there Brazier was a founding member of the International Federation of Societies for Electroencephalography and Clinical Neurophysiology. In 1950, her comprehensive 178-page *Bibliography of Electroencephalography, 1875–1948* (see Further Reading) was published as the first supplement to the novel field's new (two-year-old) journal, *EEG and Clinical Neurophysiology*. It became, as three of her long-time friends and colleagues later put it, 'a guiding beacon to the then newcomers to the field'. The same would prove true of Brazier's 1951 textbook, *The Electrical Activity of the Nervous System* (see Further Reading), which went through four editions and was translated into seven languages. During this period (the late 1940s to early 1950s), in addition to her research, writing and organizational work, Brazier began to develop her thinking about the relationships between nervous activity, consciousness and behavior.

Until the end of the 1940s, most research on EEGs involved 'eyeballing' the data, and it became apparent to Brazier and others that more precise and reliable techniques were needed. In 1946, Schwab and Brazier persuaded Grey Walter to visit Boston with the automatic low-frequency analyzer he had developed for use in his EEG research. He demonstrated the Walter analyzer at MGH at a meeting of the Eastern Association of Encephalographers, and the analyzer remained in the MGH Clinical EEG Laboratory. It was used there by Brazier and others to analyze human EEGs until the early 1950s, when it was superceded by new technologies that were developed in the post-World War II ferment of excitement about cybernetics, signal analysis and communication (information) theory.

This ferment was particularly intense in the Boston and Cambridge area, with MIT mathematician Norbert Wiener at its epicenter. After the war, Wiener published two influential books based largely on his wartime work on prediction theory, namely *Cybernetics or Control and Communication in*

the Animal and the Machine (published in 1948) and *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (published in 1949). Scientists and engineers in a wide range of fields regarded these texts as providing mathematically specifiable concepts (e.g. information, feedback, communication system, signal/noise) which could be used to analyze complex systems of all kinds (machine, biological, human–machine, social), including the central nervous system. Wiener's influence was particularly strong at MIT's interdepartmental Research Laboratory of Electronics (RLE), where in 1951 Walter Rosenblith established a Communications Biophysics Group to investigate sensory (primarily auditory) systems in animals and humans using the latest engineering technologies.

Beginning in 1948, when Brazier invited Wiener to speak to a group of MGH researchers in psychiatry and physiology, she became increasingly interested in using mathematical techniques developed by Wiener (autocorrelation and cross-correlation analyses, used to detect the presence of a weak signal embedded in noise) to analyze EEGs. These correlational techniques could be used both to identify naturally occurring rhythms in the brain's 'resting' state and to detect responses evoked by sensory stimulation, making them useful general tools for exploring questions about brain functioning and behavior. Stimulated by Wiener's work and his new-found interest in EEGs, Brazier and James Casby, an MIT undergraduate working in Brazier's MGH laboratory, published a paper on 'cross-correlation and autocorrelation studies of electroencephalographic potentials' in 1952. Brazier and Wiener discussed the application of these techniques to EEGs at the Third International Congress of Electroencephalography and Clinical Neurophysiology, which was held in Cambridge, Massachusetts in 1953.

'Application' of the mathematical techniques in EEG research required instrumentation, in particular devices for filtering the electrical activity recorded through EEG electrodes and 'correlators', namely machines for calculating the cross-correlations (between a stimulus presentation and brain activity, over repeated stimulus presentations) and autocorrelations (between a segment of the recorded activity and the same segment overlaid but displaced in time). (The latter technique enables the detection of cycles of unknown periodicities and small amplitudes relative to background activity.) Brazier's MGH group, which after 1951 included the then third-year Harvard Medical School student John Barlow, began a mutually fruitful collaboration with Walter Rosenblith's Communications Biophysics Group. An analog electronic correlator had been developed at the RLE in the late 1940s, and reliable, high-speed digital electronic computers were being developed by computer designers at MIT's Lincoln Laboratory, which had close ties to Rosenblith's laboratory. Both analog and digital machines could be used to perform the correlations required for the quantitative analyses of EEGs which Brazier sought, and she used both in research published in the early 1950s. When the first of a new generation of general-purpose digital computers was developed at Lincoln Laboratory in the mid-1950s (the TX-0), one of its earliest applications was in a device called the Average Response Computer (ARC-1). The latter was used in RLE's Communications Biophysics Laboratory by Rosenblith, Brazier, Barlow, Nelson Kiang and others to analyze electrophysiological recordings of nervous system activity (including EEGs). A technically detailed, first-hand account of the early history of EEG data processing by the MIT–MGH collaborators was published by Barlow in 1997 (see Further Reading). Brazier herself published a more extended overview with illustrations from her experiments on animals, normal human subjects, and neurological patients.

Interest in electroencephalography as a technique for investigating brain activity and consciousness continued to grow during the 1950s. Beginning in 1953, after Stalin's death, international exchanges between brain scientists in the West and in the former Soviet Union generated considerable excitement. Brazier participated in one of the earliest of these, namely the 1953 Laurentian Symposium on Brain Mechanisms and Consciousnesss (held in Ste Marguerite, Quebec), and in the subsequent Moscow Colloquium on Electroencephalography of Higher Nervous Activity, to which Brazier was invited as one of five US scientists. From the Moscow Colloquium, plans emerged for the formation of the International Brain Research Organization (IBRO), the first international organization to encompass all of the areas of what are now termed the neurosciences, including investigations of brain–behavior relationships. The first meeting was held in 1960, and Brazier served as Secretary General of IBRO from 1978 to 1982. Throughout her career Brazier actively promoted internationalism in brain research, not only through IBRO, but by serving successively as Treasurer, Secretary and President of the International Federation of Societies for Electroencephalography and Clinical Neurophysiology between 1953 and 1965. After her death, the International Federation established

a scientific award in clinical neurophysiology in her honor, the M.A.B. Brazier Young Investigator International Award.

The objective of Brazier's EEG research, in the words of long-time collaborators and friends, 'was to try to understand the nature of the EEG, as reflected in its statistical properties, as a signal in a communication system, i.e. the brain'. Throughout her career, Brazier continued to consider the implications of cybernetics and communication theory for brain research, sometimes contrasting models drawn from information theory with classical approaches ('deterministic models') in ways that presaged neural network models by many years. Brazier's interactions with Soviet brain scientists also influenced her research, and in the late 1950s and early 1960s she used correlational techniques to investigate EEGs in relation to the orienting, conditioning and habituation responses (phenomena of considerable interest to brain scientists working in the Pavlovian tradition). Brazier's bibliography of EEG research, published in 1950, had shown an interest in the historical background of her science, and in the 1950s she began to publish articles about the history of neurophysiology. In the 1980s she published her landmark books, *A History of Neurophysiology in the Seventeenth and Eighteenth Centuries* and *A History of Neurophysiology in the Nineteenth Century* (see Further Reading).

In 1958, the Macy Foundation sponsored the first of two invited Conferences on the Central Nervous system and Behavior, organized by Horace Magoun of the University of California at Los Angeles (UCLA). Magoun had been instrumental in establishing a brain and nervous system research unit when UCLA's School of Medicine was organized in 1950, and the unit was established as the Brain Research Institute (BRI) in 1959. Brazier's skillful editing of these conference proceedings led to a continuation of the series – now called Brain and Behavior Conferences – under the auspices of the American Institute of Biological Sciences, again with Magoun as organizer and Brazier as editor. These interdisciplinary conferences had international participation (including distinguished Soviet and East European brain scientists), and their rapid publication brought the latest research in the brain sciences to broad scientific audiences. Concurrently, between 1961 and 1965 the US Air Force Office of Scientific Research sponsored conferences on brain function. These, too, were organized by Magoun, edited by Brazier, and influential in the new field that was becoming known as 'neuroscience'. In 1961, Brazier left her MGH laboratory and her long-time MIT collaborators to join Magoun, Donald Lindsley, Louise Marshall and others at the BRI. Always impossible to pigeon-hole in a neat grid of established disciplines, Brazier was appointed professor in the Departments of Anatomy, Biophysics and Nuclear Medicine, and Physiology of UCLA's Medical School. Prior to her arrival, computers were not widely used by BRI researchers, and Brazier, who served on a National Institutes of Health (NIH) advisory committee on computers in research, led the development of a data-processing facility. Her EEG research continued at the BRI, expanding to include new clinical applications. She was active in university-wide affairs at UCLA, and also served as a consultant or advisor on several NIH and National Science Foundation committees, as well as those of national non-governmental organizations.

In 1988, with her eyesight failing, Brazier moved back to the East Coast, to the sea and to the house she had built on Cape Cod while she was living in Boston. There she gardened, sailed, traveled (at least once every year to Paris and London) and visited and corresponded with friends and family – her son Oliver and his family were nearby. She died on 9 May 1995, nine days before what would have been her ninety-first birthday.

HONORS AND AWARDS

Mary A. B. Brazier was elected to the American Academy of Arts and Sciences in 1956. In 1962 she received a Career Research Award from the National Institutes of Health, one of four scientists so honored in the first year of these awards. The University of London honored her with a doctorate (DSc, on the basis of her published works) in 1960, and she received an MD (honoris causa) from the University of Utrecht in 1976. In 1985, the British EEG Society awarded her the Grey Walter Medal. The most complete and detailed account of Brazier's life and work to date is the 1996 memorial tribute written by John Barlow, Robert Naquet and Hans van Duijn (see Further Reading).

Brazier's papers are archived in the History and Special Collections Division of the Louise M. Darling Medical Library, UCLA. Contemporary researchers are fortunate that her published work (she was the author or editor of almost 250 articles and books) speaks clearly for itself. She wrote well, and her sense of history – even when she was working within the conventional genres of scientific articles and chapters – enabled her to place her own work in an unusually broad and detailed scientific context. Reading Brazier's writings, both scientific and historical, will give contemporary

cognitive and brain scientists a richer and deeper understanding of their field and its conceptual and technical roots.

Acknowledgments

Dr John Barlow has very kindly allowed me to see some of the extensive correspondence he had with Mary Brazier's friends and colleagues while he was preparing the 1996 memorial tribute to her life and work. I thank him for doing so.

Further Reading

Barlow JS (1997) The early history of EEG data-processing at the Massachusetts Institute of Technology and the Massachusetts General Hospital. *International Journal of Psychophysiology* **26**: 443–454.

Barlow JS, Naquet R and van Duijn H (1996) In memoriam: Mary A.B. Brazier (1904–1995). *Electroencephalography and Clinical Neurophysiology* **98**: 1–4.

Brazier MAB (1950) Bibliography of electroencephalography, 1875–1948. *EEG and Clinical Neurophysiology* **Supplement 1**: 1–178.

Brazier MAB (1950) Neural nets and integration. In: Richter K (ed.) *Perspectives in Neuropsychiatry*, pp. 35–45. London, UK: HK Lewis.

Brazier MAB (1951) *The Electrical Activity of the Nervous System*. London, UK: Pitman.

Brazier MAB (1954) The Laurentian Symposium on the electrical activity of the cortex as affected by the brainstem reticular formation in relation to states of consciousness. *EEG and Clinical Neurophysiology* **6**: 355–359.

Brazier MAB (1958) The development of concepts relating to the electrical activity of the nervous system. *Journal of Nervous and Mental Diseases* **126**: 303–321.

Brazier MAB (1959) The historical development of neurophysiology. In: Field J, Magoun HW and Hall VE (eds) *Handbook of Physiology – Neurophysiology*, vol. 1, pp. 1–58. Washington DC: American Physiological Society.

Brazier MAB (1960) Some uses of computers in experimental neurology. *Experimental Neurology* **2**: 123–143.

Brazier MAB (1960) Long-persisting electrical traces in the brain of man and their possible relationship to higher nervous activity. *EEG and Clinical Neurophysiology* **Supplement 13**: 347–358.

Brazier MAB (1963) How can models from information theory be used in neurophysiology? In: Fields WS and Abbott W (eds) *Information Storage and Neural Control*, pp. 230–242. Springfield, IL: Thomas.

Brazier MAB (1984) *A History of Neurophysiology in the Seventeenth and Eighteenth Centuries: From Concept to Experiment*. New York, NY: Raven.

Brazier MAB (1988) *A History of Neurophysiology in the Nineteenth Century*. New York, NY: Raven.

French JD, Lindsley DB and Magoun HW (1984) *The Brain Research Institute, UCLA: An American Contribution to Neuroscience*. Los Angeles, CA: UCLA Brain Research Institute.

Marshall LH (1996) Early history of IBRO: the birth of organized neuroscience. *Neuroscience* **72**: 283–306.

Rosenblith WA (1966) From a biophysicist who came to supper. In: *R.L.E.: 1946 + 20*, pp. 42–50. Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Wiesner JB (1966) The communication sciences – those early days. In: *R.L.E.: 1946 + 20*, pp. 12–16. Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Broca, Paul

Intermediate article

Stanley Finger, Washington University, St Louis, Missouri, USA

CONTENTS

Paul Broca (1824–1880) was a French surgeon, pathologist, anatomist and anthropologist. He is remembered for localizing speech in the frontal lobes, recognizing cerebral dominance, and performing the first surgery based on localization, as well as his research on the brain and intellect, and for his insights and theories on ancient trepanned skulls.

INTRODUCTION

Paul Broca (Figure 1) was born in France on 29 June 1824 in Sainte-Foy-la-Grande, a small town near Bordeaux, to Protestant parents. Following in the footsteps of his father, he opted to study medicine. This decision led him to Paris, where he excelled in his studies and completed his medical degree in 1848. From the beginning he had a reputation for being extremely thoughtful, thorough, and for looking at medical and scientific problems from many perspectives. He was also considered a liberal – not by today's standards, but in the culture in which he lived. Broca first made a name for himself by showing that cancer cells can be spread through the blood, and with his studies on various diseases, including muscular dystrophy and rickets. His early research in pathology, coupled with his strong belief that laboratory and clinic must join forces to improve medicine, helped him to secure several desirable Paris hospital appointments, such as surgeon at the Bicêtre.

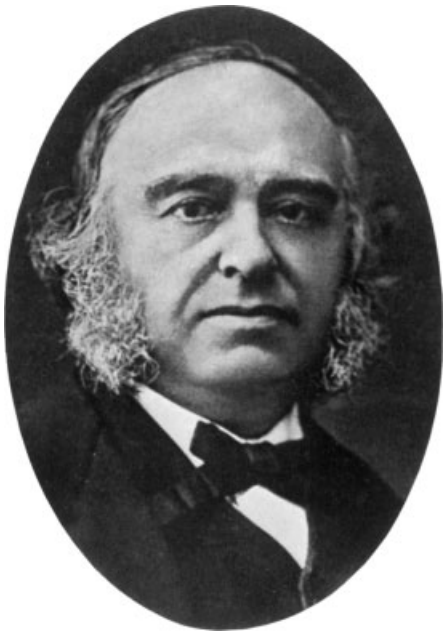


Figure 1. Paul Broca (1824–1880).

SPEECH AND THE FRONTAL LOBE

In 1859 Broca founded the world's first anthropological society, the Société d'Anthropologie. It was at the meetings of this fledgling society, where Broca served as secretary, that scientists discussed human groupings, intelligence, and the brain. It was also here that some French physicians interested in the effects of brain damage started to make the case for cortical localization of function. Two such individuals were Jean-Baptiste Bouillaud and his son-in-law Simon Alexandre Ernest Aubertin. They contended that damage towards the front of the cerebrum is more likely to disrupt speech than damage towards the back of the massive cerebral hemispheres. Others, however, including Pierre Gratiolet, disagreed, arguing that the cerebral hemispheres function as an indivisible unit.

On 12 April 1861 a 51-year-old man suffering from cellulitis and gangrene was transferred to Broca's surgical service at the Bicêtre. Described as mean and vindictive, the patient (named Leborgne) had suffered from epilepsy since youth and had been hospitalized at the age of 31, after losing his power to speak. He then developed a paralysis on the right side of his body with loss of sensitivity on the same side. Broca invited Aubertin

to examine Leborgne with him, to assess his speech and to see if he would exhibit damage to his frontal cortex when he died. Indeed, the patient was found to have great difficulty speaking. After he died an autopsy showed a chronic softening in the third frontal convolution, near the rolandic fissure of the left hemisphere. Broca first presented this patient's brain to the Société d'Anthropologie. A more detailed report was given to the Société d'Anatomie later in 1861. Broca used the French word *aphemie* (or aphemia) for Leborgne's inability to speak. It was Armand Trousseau who coined the more popular word 'aphasia' in 1864. This form of aphasia became known as Broca aphasia or motor aphasia. As for Monsieur Leborgne, since 'tan' was once of the few sounds he was able to make before he died, he was often to be called Tan in the later literature.

Broca used this case to argue for a special frontal cortical area that is responsible for fluent speech. He went out of his way to explain that this is located behind and below the one proposed by phrenologists, such as Gall and Spurzheim, who associated character and abilities with bumps on the skull and whose theories were in disrepute. Today we refer to this specialized cortical region as 'Broca's area'. Hence, two eponyms that helped make Broca famous stemmed from this one case. With the landmark case of Leborgne, Broca became fully committed to the cortical localization revolution and was looked upon as its champion. As a careful investigator, however, he worried that he might have gone too far on the basis of only a single case study. Over the next few years he was gladdened to find additional cases that were supportive of his frontal lobe localization for speech, beginning with the case of an old man named Lelong later in 1861.

CEREBRAL DOMINANCE

On 2 April 1863 Broca lectured about eight cases of loss of fluent speech. He remarked that, to his surprise, all exhibited lesions of the left hemisphere. Still, he felt that more cases were needed before he could make a definitive statement about the left hemisphere being special. The idea seemed likely to generate even more of a storm than cortical localization of function.

Broca's clearest statements and most important thoughts about cerebral dominance appeared in 1865, in an article in the *Bulletin de la Société d'Anthropologie*. In this he theorized that the left hemisphere is, in fact, dominant for language. Because it matures faster than the right hemisphere, it is better suited to take the lead. As was true with the concept of cortical localization of function, Broca was not the first to argue for this new way of looking at the brain, although his role was significant. That honor goes to Marc Dax, a physician from the south of France who wrote a memoir on the left hemisphere and speech in 1836, but never published it (he died a year later). Whether his material had indeed been presented orally at a congress in 1836, as was claimed by his son Gustave, is not certain. Broca himself could find no evidence that it was. What we do know is that the Marc Dax paper was sent to the Académie de Médecine in Paris in 1863. It was a part of a larger report by Gustave Dax which contained his own collection of cases supportive of cerebral dominance. The paper arrived and was announced by title (but not made public) just before Broca made his first tentative remark about the eight cases with left hemispheric lesions. The Dax report was then sent to a 'secret' committee, where it languished before it was severely criticized by the committee chairman (Lelut) late in 1864. Upset by the Lelut report, Gustave Dax then saw to it that both his father's report and his own data were published elsewhere. They appeared as two short but separate articles in a medical periodical in 1865, just weeks before Broca's own celebrated paper appeared in a different journal.

Today, both the Daxes and Broca are recognized for their seminal contributions to the development of the concept of cerebral dominance. For a while, however, Broca was given most – if not all – of the credit for this important discovery.

AGE, BRAIN DAMAGE AND THERAPY

In his 1865 paper on cerebral dominance, Broca was forced to deal with exceptions to the idea that the center for articulate language resides in the third left frontal convolution. One such case was that of a woman with epilepsy who was a patient at the largest Paris hospital, the Salpêtrière. This patient was probably born without a left Broca's area, but was able to learn to read, speak fairly well, and express her ideas without difficulty. Broca suggested that the healthy right hemisphere had taken over the role of the compromised left hemisphere, something that is accomplished more readily when the brain damage occurs early in life. He also postulated that a small percentage of healthy people might be born 'right-brained'. He then considered the question of why we do not see more sparing and recovery following damage to this part of the brain, postulating that one limiting factor might be that most aphasic patients also suffer

from intellectual deficiencies, limiting their ability to relearn (see below). This would be especially likely after strokes and injuries that affected more than just Broca's area in the frontal lobes.

Broca also pointed out that professionals did little to retrain their aphasic patients. He suggested teaching people with aphasia in the same way that a child learns to speak: therapy should begin with sounds of the alphabet, then words, then phrases, and eventually sentences. By working from the simple to the complex, Broca suggested, the right hemisphere might find it easier to take over from its injured counterpart on the left side. He tried speech therapy with one of his own adult patients, who was successful in relearning the alphabet and in working with syllables, but did not do well when it came to constructing longer words. Nevertheless, Broca was optimistic and expressed the hope that others would be able to devote more time to speech therapy than he had been able to do owing to his busy schedule.

SURGERY BASED ON LOCALIZATION

In 1865 Broca was elected president of the Paris Surgical Society and 3 years later he became professor of clinical surgery. In 1868 he introduced cranial cerebral topography, a technique that uses skull and scalp landmarks to localize underlying parts of the brain. Broca used his new method to open the skull in the right place and drain an abscess in a patient whose speech had become impaired after a closed head injury. Although the operation took place late in the 1860s, it was not reported until 1876. This was probably the first brain surgery to be based on the new theory of cortical localization of function.

INTELLECT, BRAIN AND RACE

Beginning in 1861, Broca also raised the possibility that the frontal lobes might serve executive functions other than speech, including judgment, reflection and abstraction. Indeed, when Leborgne's lesion was spreading throughout the frontal lobes, this patient showed signs of losing his intellect, not just his fluent speech. By arguing that the front of the brain is more 'intellectual' than the back, Broca was able to explain why there was not more relearning and recovery after large frontal lobe lesions that affect speech. He and the other localizationists who accepted this idea also had a good explanation for why some individuals with large skulls were not as intelligent as others with smaller skulls. For example, in 1873 he examined some recently unearthed Cro-Magnon specimens from central France. They had cranial capacities that far exceeded those of the modern French. To Broca and his colleagues in Paris, it was not that Cro-Magnon men and women were geniuses; they most certainly were not. Instead, the greater overall size of their crania only reflected the greater development of the more pedestrian back of the brain. Broca made precisely the same point when referring to some old exhumed Basque skulls that were sent north to Paris. They were also large relative to the skulls of modern Parisians, but this too was attributed to growth in the back of the brain, not the intellectual front.

Thus, although Broca initially believed that cranial capacity was a good physical correlate of intellect, he abandoned this view as he learned more about cortical localization of function. In addition, like many others at the time, Broca believed in multiple creations for the different human races. However, once caught up in the Darwinian revolution of 1859, he also embraced evolution, rejecting the older notion that the human groups are fixed entities, and with it the popular belief that only pure races could prosper. Moreover, he found slavery, even for people with small brains and low intelligence, inexcusable and repulsive.

TREPANATION

Broca's interest in trepanned skulls began in 1867 when he was asked to examine an Inca skull with cross-hatched cuts. This unusual cranium had recently been obtained in Peru by American diplomat-archeologist E. George Squier. Broca agreed with Squier that the cuts on the skull had been made on a living person prior to the European conquest, and that this individual had survived the operation by a few weeks. Thanks to Broca's help, this was the first case of trepanation from an ancient culture to be correctly and widely recognized as such.

Broca then became involved with the discovery of much older trepanned skulls in France. Many late Neolithic (New Stone Age) crania that had been trepanned were found, most of which are now estimated to be approximately 5000 years old (the Peruvian skull was judged to be only around 500 years old). Broca visited burial sites and unearthed some cranial specimens himself, but mostly studied the findings presented to him by others, especially one of his associates, Prunières. Broca postulated that that the openings in the Neolithic skulls had been made by scraping with a sharp stone, such as a piece of flint or obsidian,

and that the fibrous dura mater covering the brain was left intact. As for the smoothed surfaces where bone had been removed, they were the result of an extended period of healing. He further posited (from one skull in particular) that the operation was probably performed early in life.

During the mid-1870s Broca gave many talks and published a large number of papers on trepanation. One of his goals was to convince people that the holes in many of the unearthed French skulls were not due to accidents, combat, nature, or gnawing animals. Another was to associate the surgery with some sort of therapy and with the primitive mind. Broca held that the openings were not the result of surgical treatment of head wounds, since there would have been more openings over the facial areas, which were carefully avoided. Moreover, he did not see signs of fractures. Instead, he thought it more likely that the operations were done on the living to treat 'internal maladies'. After much thought, and after considering newer anthropological evidence, he suggested that the surgery might have originated as a way to treat benign infantile convulsions, such as seizures caused by fever spikes or teething. These were disorders that primitive people might have attributed to demons. Moreover, the children would have recovered anyway: an illusion of success would have been achieved, and the practice would have spread and perhaps generalized.

Broca's theory about trepanation, demonology and seizure disorders is still widely cited in books and papers on trepanation. Most researchers agree that he was probably on the mark when he suggested that the practice had something to do with medicine, the brain, and abnormal behaviors; but he was wrong to think that the operations were confined to children.

LATER YEARS

Paul Broca made his last statements about speech and the brain in 1877. At this time, he was much more interested in the family of man than in cortical localization of function. In addition, he was intrigued by the limbic lobe, a collection of brain parts then thought to be associated with olfaction, a subject on which he wrote in 1877 and 1878.

Broca died in 1880, only months after he was elected to the French Senate as a representative of science and medicine. He had been a perfectionist and a 'workaholic' who published over 500 books and articles during his intense scientific career. When he succumbed to heart disease, his wife, three children, and scientists and physicians around the world mourned the passing of a man who had contributed monumentally to many fields. In the neural and cognitive sciences he is best remembered for his theory of the cortical localization of speech, his recognition of cerebral dominance, his thoughts about intelligence and the races, and for his discoveries and insights bearing on the ancient practice of cranial trepanation.

Further Reading

Broca P (1861) Remarques sur le siège de la faculté du langage articulé; suivies d'une observation d'aphémie (perte de la parole). *Bulletins de la Société Anatomique (Paris)* **6**: 330–357, 398–407. Translated as 'Remarks on the seat of the faculty of articulate language, followed by an observation of aphemia' in von Bonin G (1960) *Some Papers on the Cerebral Cortex*, pp. 49–72. Springfield, IL: Charles C. Thomas.

Broca P (1865) Sur le siège de la faculté du langage articulé. *Bulletins de la Société d'Anthropologie* **6**: 337–393. Translated as 'Localization of speech in the third left frontal convolution' by Berker EA, Berker AH and Smith A (1986) in *Archives of Neurology* **43**: 1065–1072.

Broca P (1878) Anatomie comparée des circonvolutions cérébrales. Le grand lobe limbique et la scissure limbique dans la série des mammifères. *Revue d'Anthropologie Sériè 2* **1**: 385–498.

Clower WT and Finger S (2001) Discovering trepanation: the contributions of Paul Broca. *Neurosurgery* **49**: 1417–1425.

Dax M (1865) Lésions de la moitié gauche de l'encéphale coïncidant avec l'oubli des signes de la pensée (lu au Congrès méridional tenu à Montpellier en 1836). *Gazette Hebdomadaire de Médecine et de Chirurgie* **2** (series 2): 259–260.

Finger S (1994) *Origins of Neuroscience*. New York, NY: Oxford University Press.

Finger S (2000) *Minds Behind the Brain*, chap. 10, Paul Broca, pp. 137–154. New York, NY: Oxford University Press.

Finger S and Roe D (1996) Gustave Dax and the early history of cerebral dominance. *Archives of Neurology* **53**: 806–813.

Joynt RJ and Benton AL (1964) The memoir of Marc Dax on aphasia. *Neurology* **14**: 851–854.

Schiller F (1992) *Paul Broca: Founder of French Anthropology, Explorer of the Brain*. New York, NY: Oxford University Press.

Stone JL (1991) Paul Broca and the first craniotomy based on cerebral localization. *Journal of Neurosurgery* **75**: 154–159.

Descartes, René

Introductory article

Stephen Gaukroger, University of Sydney, Sydney, New South Wales, Australia

CONTENTS

Descartes pursued two different approaches to the question of the nature of the mind: one via psychophysiology and one via a theory of the different properties of mind and matter, construed as different substances.

DESCARTES' LIFE AND PHILOSOPHICAL DEVELOPMENT

René Descartes was born in 1596, and entered the Jesuit college of La Flèche as a boarder at the age of 10. He left it in 1614, and, after spending a year in Paris, completed his formal education by taking a degree in civil and canon law at the University of Poitiers in 1616. From 1619 onwards, he pursued a career as a gentleman soldier, first in the army of Maurice of Nassau and then in that of Maximillian I of Bavaria, before settling down to a life of science and scholarship in the early 1620s. He worked primarily in mathematics and natural philosophy in the 1620s and early 1630s, in Paris and elsewhere, moving to the Netherlands in 1628. In the mid-1630s he developed a skeptical form of epistemology, set out in the *Discourse on Method* (1637), in the *Meditations* (1641), and finally in the *Principles of Philosophy* (1644). It is this form of pure epistemological speculation for which Descartes is now principally remembered. In 1649 he moved to Sweden, where he died early in 1650.

We can trace three different strands of interest in Descartes' development. From 1619 to the late 1620s he pursued mathematics above all else. A precocious and original mathematician, his greatest contribution was to the discipline of analytic geometry, in which lines and curves are represented by equations through the use of coordinates. What he provided was a powerful unification of arithmetic and geometry, and it was from his treatise on the techniques that he had developed in this area, the *Geometry*, that Newton and others learned their advanced mathematics later in the seventeenth century.

Descartes also pursued an active research program in natural philosophy from 1619 onwards, moving from kinematics, hydrostatics, and optics to a general Copernican cosmology in the 1630s. Some time in the middle to late 1620s, he discovered a central law in geometrical optics, the law of refraction, which was crucial in the development of better telescope lenses. In the early 1630s, in *The World*, he developed a comprehensive physical cosmology – the most important seventeenth-century cosmological system before Newton – in which the problem of how the planets can revolve in stable orbits around a central sun was solved by proposing a model in which a revolving celestial fluid carries the planets along, their distance from the sun being a function of their size.

In the 1630s, Descartes began to develop a distinctive epistemology driven by skepticism. He focused on a number of problems, such as radical skepticism, the provision of foundations for knowledge, and the exact nature of the relation between mind and body, which were either new or treated in a new way. However, Descartes explicitly warned against an insulation of philosophy from empirical questions.

THE PSYCHOPHYSIOLOGY OF COGNITION

At various stages in his career, Descartes tried to describe the nature of various kinds of intellectual or psychological phenomena in psychophysiological terms. There are three that are of particular importance: mathematical cognition, perceptual cognition, and affective states.

Mathematical Cognition

In his *Rules for the Direction of the Mind*, the relevant parts of which were completed between about 1626 and 1628, Descartes was concerned with the question how a quantitative grasp of the world was

possible. The question is how to connect the contents of the world, which consists of material objects, with the contents of the intellect, which, in the case of mathematical cognition, consists of abstract mathematical structures, which may have arithmetical or geometrical interpretations, but which are neither arithmetical nor geometrical in themselves. If a quantitative grasp of nature is to be possible, mathematics must somehow be mapped onto the material world.

Descartes' solution is to suggest that such a mapping cannot be direct: a determinate representation of the abstract mathematical structures is mapped onto a *representation* of the world. As regards the representation of the world, Descartes sets out to show how qualitative differences, such as differences in color, can be represented purely in terms of different arrangements of lines: red as vertical lines, blue as horizontal lines, green as a combination of these, yellow as diagonal lines, etc. We might think of this as a form of encoding: qualitative differences can be encoded in a very economical form, namely in terms of lines. As regards abstract mathematical structures, Descartes argued that these can be represented in terms of line lengths, or combinations of line lengths – the basic arithmetical operations of addition, subtraction, multiplication, division, and root extraction can all be performed using line lengths, for example, and geometrical operations present no problem in this respect. The contents of the intellect are represented in the 'imagination' as line lengths, and the contents of the material world are represented there as configurations of lines, and the former are mapped directly onto the latter, thus allowing a quantitative grasp of nature.

There are a number of interesting features of this account. Firstly, there is the idea that sensory information must be encoded in some way if we are to be able to engage with it cognitively. Secondly, note that the 'imagination' is a material organ (Descartes will later identify it with the pineal gland, this being the unique central, unduplicated organ in the brain, and hence ideally suited as a site for central cognitive processing), and that this, rather than the intellect, is where the cognition actually takes place. In other words, we seem to have a material site for cognition.

Perceptual Cognition

A distinctive feature of Descartes' natural philosophy is his commitment to mechanical explanation. This is evident in his physics and astronomy, but it goes further. Without appealing to vital forces of any kind, Descartes reasoned that, except in the case of the exercise of judgment and free will, which require consciousness, physiological processes – including such psychophysiological cognitive functions as visual perception, memory, and habitual and instinctual responses – can be accounted for mechanically. In the *Treatise on Man*, Descartes set out one of his most daring projects: the complete mechanization of physiology, from nutrition, excretion and respiration up to memory and perceptual cognition. His treatment of the last two is particularly ingenious.

In the case of visual perception, he argues – against a long and deeply entrenched tradition stretching back as least as far as Aristotle – that a visual image need not resemble the object perceived. Not only is there nothing in the optics or physiology of vision that requires resemblance, but the fact that the retinal image is inverted, that the retina where the visual image must be represented is a two-dimensional concave surface, that it must be transmitted through the nerves, and so on, all indicate a form of encoding of information. Descartes also shows awareness of fundamental problems of information recognition: he shows how we must employ an innate or unconscious geometry in order to be able to gauge the distance of objects, since our visual stimulation results from a light ray which cannot carry information about how far it has traveled from the object to the eye.

Descartes' account of animal cognition is very sophisticated. His aim is to show that the structure and behavior of animal bodies are to be explained in the same way as we explain the structure and behavior of machines. In doing this, he wants to show how a form of genuine cognition occurs in animals, and that this can be captured in mechanistic terms. He does not want to show that cognition does not occur, that instead of a cognitive process we have a merely mechanical one. In more modern terms, his project is a reductionist, not an eliminativist, one.

In the case of memory, Descartes offers an account in which the memory images do not have to resemble what caused them, and they do not have to be stored faithfully and separately but only in a way that enables the idea to be represented in a recognizable form. Unlike his contemporaries, who were largely preoccupied with identifying the physical location of memory storage (a favored location was in the folds of the surface of the brain, because of the large surface area such folds created), Descartes' concern is with just what is needed for recall, and he provides a rudimentary

account of how memory works by means of association.

Affective States

In the *Passions of the Soul* (1649), Descartes provided an extensive account of affective states, or passions, in which he examined how the mind and the body interact to produce such states as fear, anger, and joy. One of his primary concerns here was to argue against the idea that there are higher and lower functions of the mind, that there is a hierarchy of appetites, passions, and virtues, with the will occupying a precarious position. What Descartes opposes is the idea of a fragmentation of the soul, whereby one loses a sense of how the agent can collect himself or herself together, and exercise true moral responsibility. This is particularly important for Descartes, because he sees the crucial part of ethics to be the difficult process of forming oneself into a fully responsible moral agent – this is what the control of one's passions is ultimately aimed at – rather than the question of how such an agent should behave.

THE METAPHYSICS OF MIND

The doctrine for which Descartes is most famous is 'Cartesian dualism'. He advocates a view of the mind whereby (1) the mind is a different substance from the body, by virtue of having different essential or defining properties, and (2) the mind can exist in its own right, independently of the body, and have an identity that distinguishes it from other minds. Modern versions of dualism usually restrict their claims to the first of the above claims, substance dualism, but Descartes' advocacy of substance dualism seems to be in large part motivated by the second claim, which, because mind is not subject to the physical processes that lead to death and corruption, is tantamount to the doctrine of personal immortality.

Substance dualism requires no commitment to the capacity for independent existence of the mind. The fact that mind and matter are separate substances does not in itself require us to imagine that mind might be able to exist independently of matter: we might conceive of the mind as the 'software' that runs the cognitive parts of the body, for example, thinking of it as something quite distinct from its material realization in a particular brain, or central nervous system, while at the same time arguing that it makes no sense to talk about such software independently of its being a program.

However, Descartes' concerns seem different. As he indicates in the dedicatory letter which prefaces the *Meditations*, he is concerned to defend the doctrine of personal immortality of the soul. This doctrine had been undermined by two different kinds of philosophical conception of the mind. The first was Alexandrism, which was in effect substance dualism without personal immortality. Alexander of Aphrodisias and his followers had argued that the mind or soul is the 'organizing principle' of the body, something essentially materially realized, so that with the death and corruption of the body, it goes out of existence. The second kind of conception was Averroism, whereby the mind can be separated from the body at death, but in undergoing such separation it loses any identifying features and becomes identical with 'mind' as such. The idea here is that what distinguishes my own mind from another is a set of features it has by virtue of being instantiated – my sensations, memories, and passions are easily sufficient to mark me out from everyone else, for example, but these are dependent on my having a body – and once it becomes separated from my body, it loses anything that might differentiate me from anything else, and so becomes one with a universal mind (God).

Descartes' challenge is to steer a middle path between Alexandrism, which denies immortality altogether to the soul, and Averroism, which denies it personal immortality. It is not clear that he is able to do this. His account of such processes as mathematical cognition, perceptual cognition, and affective states presuppose that the mind is instantiated in the body, and so are compatible with Alexandrism (conceived as a minimal substance dualism). He does not describe what a disembodied soul is like except to tell us that it contemplates universals, but that is what God does, and what Averroes' single mind does: there is nothing to distinguish disembodied souls from one another in this respect.

RELEVANCE OF DESCARTES' WORK TO COGNITIVE SCIENCE

Descartes was the first person to provide a comprehensive account of a mechanized psychophysiology. Although his ideas had some followers in the succeeding two centuries, the resources available – most importantly knowledge of brain physiology – were far from adequate, and Descartes' project looked like a dead end. As these resources were acquired, from the late nineteenth century

onwards, the situation changed. There remain a number of deep philosophical problems about the nature of cognition and the mind – perceptual cognition in unintelligent animals, what is involved in memory retrieval, and so on – which Descartes, because of his limited empirical resources, was forced to focus on in a way that draws attention to some of the conceptual problems that need to be addressed if one is to orientate one's empirical investigations in a fruitful direction.

Further Reading

Baker G and Morris K (1996) *Descartes' Dualism*. London, UK: Routledge.

Descartes R (1984–1991) *The Philosophical Writings of Descartes*, edited and translated by J Cottingham *et al*., 3 vols. Cambridge, UK: Cambridge University Press.

Descartes R (1998) *The World and Other Writings*, edited and translated by S Gaukroger. Cambridge, UK: Cambridge University Press.

Gaukroger S (1995) *Descartes: An Intellectual Biography*. Oxford, UK: Oxford University Press.

Gaukroger S (2002) *Descartes' System of Natural Philosophy*. Cambridge, UK: Cambridge University Press

Gaukroger S, Schuster J and Sutton J (eds) (2000) *Descartes' Natural Philosophy*. London, UK: Routledge.

Sepper D (1996) *Descartes's Imagination*. Berkeley, CA: University of California Press.

Sutton J (1998) *Philosophy and Memory Traces*. Cambridge, UK: Cambridge University Press.

Wolf-Devine C (1993) *Descartes on Seeing*. Carbondale, IL: Southern Illinois University Press.

Yolton JW (1984) *Perceptual Acquaintance from Descartes to Reid*. Oxford, UK: Blackwell.

Ebbinghaus, Hermann

Introductory article

Robert R Hoffman, Institute for Human and Machine Cognition, University of West Florida, Pensacola, Florida, USA
Michael Bamberg, Clark University, Worcester, Massachusetts, USA

CONTENTS

Hermann Ebbinghaus (1850–1909) was a German psychologist whose books, research, and ideas had a great effect on early psychological theory. He is often credited with founding the experimental psychology of the 'higher mental processes'.

INTRODUCTION

Hermann Ebbinghaus was born on 23 January 1850 in the industrial town of Barmen, in the Rhine Province of the kingdom of Prussia. He studied classics, languages, and philosophy, and completed his doctoral dissertation at the University of Bonn in 1873. After working for some years as a tutor, he happened to read about the new research on psychophysics, and became inspired to study the 'higher mental processes'. In 1878, Ebbinghaus began formal experiments on memory, conducted in his home. A monograph on the work was published in 1885. Within a year he was promoted to a salaried professorship at the Friedrich Wilhelm University in Berlin. Journals that published psychological research were beginning to spring up everywhere and Germany needed a general journal. Ebbinghaus helped establish the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* (Journal for the Psychology and Physiology of the Sense Organs) and served as its editor for 22 years. In 1893 Ebbinghaus took a professorship at Breslau University in the Prussian province of Silesia. Ebbinghaus's psychology textbook appeared in 1897, and was the most popular and widely used general psychology text for many years.

Ebbinghaus was known as an eloquent lecturer and excellent teacher. He was a man with vision, a champion of the view that psychology should be emancipated from philosophy, and the higher mental processes studied experimentally.

Ebbinghaus died of pneumonia in 1909. At that year's psychology conference at Clark University (to which Ebbinghaus had been invited), Cornell University psychologist Edward B. Titchener began with a eulogy: 'As I approach the topic of this lecture, what is uppermost in my mind is a sense of irreparable loss. When the cable brought the bad news, last February, that Ebbinghaus was dead … the feeling that took precedence even of personal sorrow was the wonder of what experimental psychology would do without him.'

EBBINGHAUS'S RESEARCH

The monograph *On Memory* has three aspects: the experiments themselves, a discussion of statistical analyses of data, and some theorizing. Much of the theorizing concerns the strength and vividness of associations and the search for 'mathematical rules for mental events'. Ebbinghaus's discussion of the basic statistical methods was so clear and exact that many psychologists had their students read Ebbinghaus's book just for its discussion of statistics. Each data point that entered into his analyses was an average of the learning times (or average number of repetitions needed to reach a learning criterion) over a large number of lists. The averages were used to compute a distribution of means, and results were then described in terms of standard errors: the percentage of cases falling under a given area of the distribution.

The monograph reported 19 studies conducted in the years 1879–1880 and 1883–1884. To conduct the research Ebbinghaus first prepared a pool of 'all possible syllables' – 2300 in all (quite a few of them were words in German, English or French). A few examples are heim, beis, ship, dush, noir, noch, dach, wash, born, for, zuch, dauch, shok, hal, dauf, fich, theif, hatim, shish, and rur. Pacing himself with a metronome, and reading the lists aloud with a poetic meter, he proceeded to memorize lists of syllables. Using a set of buttons on a string, he was able to keep track of the number of repetitions

he needed in order to learn a list to the point that he could give one perfect recitation. The experiments were an ambitious project and required great effort.

Ebbinghaus led a ritualistic, almost monastic life during these experiments, learning and recalling lists every day for months on end, dozens of experiments and replications of experiments, each involving multiple trials and hour after hour of data collection and careful record-keeping. Imagine learning 84 600 syllables in 6600 lists, taking more than 830 h! Although the number of list repetitions involved for every experiment cannot be determined for some of the studies on the basis of what Ebbinghaus said in his monograph, for experiment 2 alone Ebbinghaus engaged in 189 501 repetitions of lists.

The first two experiments had the goal of showing that the variability of the average learning times over a large number of lists was within limits that would be scientifically acceptable. He emphasized that the standard errors he obtained compared favorably with the precision of measurement in the physical and biological sciences (e.g. measurements of the speed of neural conduction, or measurements of the mechanical equivalent of heat). In fact, his 'probable errors' of about 7% were more precise than the physical measurements and very close to those for the biological measurements.

The list of Ebbinghaus's findings includes many of the basic phenomena that are discussed to this day in books on the psychology of memory. His research demonstrated the viability of the method of savings or 'ease of relearning' as a means of measuring the strength of association. He demonstrated the effects of fatigue and time of day on retention; the effect of list length on the number of repetitions it takes to learn material; the 'decay of memory' as a function of the delay between acquisition and memory test (with delays spanning hours, days, weeks and even years). Ebbinghaus demonstrated the effect of 'distributed versus massed' practice. He demonstrated what came to be called the 'serial position' effect (i.e. better memory for material that falls near the beginning and near the ending of a list). Ebbinghaus also measured what would come to be called the short-term memory span – 'the number of syllables which I can repeat without error after a single reading is about seven. One can, with a certain justification, look upon this number as a measure of the ideas of this sort which I can grasp in a single unitary conscious act.'

Textbooks on general and cognitive psychology preserve a myth about Ebbinghaus, which is that he conducted experiments in which he memorized 'nonsense' syllables. It was not the syllables that were nonsense – in the examples given above, the first 10 are all meaningful in one or another of the languages Ebbinghaus knew – it was the task of learning a list of semantically unconnected items that Ebbinghaus refers to as involving an 'impression of nonsense'. Indeed, the term he preferred for his lists, *Vorstellungsreihen* (literally, 'presentation series'), could just as well be translated as 'image series', and Ebbinghaus discussed at some length how the strength and vividness of memory images should be related to the effort taken in learning them.

Another contradiction to the myth is that Ebbinghaus did not begin his studies by attempting to memorize lists of syllables. Instead, he began with a task more familiar to teachers – and to the pupils whom Ebbinghaus taught – the memorization of poetry. His preliminary trials with poetry showed that the material was learned too quickly. He found no need for multiple repetitions (meaning that he could not obtain enough data about trials and time to criterion in order to generate statistically reliable laws) and he was also concerned that the material could not be systematically and quantitatively varied (lists of numbers did not afford enough variety either); hence his eventual choice of syllable lists. However, his research did not end with the memorization of syllable lists. Ebbinghaus memorized stanzas from Byron's *Don Juan* in order to address the question of the role of meaningfulness in the associative process. Over a period of 4 days he conducted seven separate tests, each test involving the learning of six stanzas. Each test took about 20 min and involved about eight repetitions of each stanza. Given that each stanza consisted of about 80 syllables, he could compute that meaningfulness resulted in a large advantage: about one-tenth the effort in terms of the number of repetitions needed to achieve one perfect recitation. Most important to Ebbinghaus was the fact that the findings with the poetry confirmed the findings for the syllables: general laws were in operation.

INTELLIGENCE TESTING

The school board of Breslau had commissioned Ebbinghaus to generate mental tests that could be used to determine the best distribution of study hours for schoolchildren. He invented the completion method to see how well children could perceive relationships, combine information, and arrive at correct conclusions. In the task, students would have to fill in the missing letters in sentences such

as 'WH__ WILLY ___ TWO _____ OLD, HE _____ __ _ RED FARM_____'. This type of task is still used in modern intelligence and aptitude tests. Along with digit memory and a rapid calculation task, the results showed a clear effect of age and individual differences. However, only the results from the method of combinations showed a relation to the children's grades. Ebbinghaus's work on intelligence testing was thus some of the very first research on this topic. According to Woodworth, the completion method was probably a better test of intelligence than any other method available at the time. Alfred Binet was working on mental testing at the time, and Binet was encouraged by Ebbinghaus' studies of school children. The original Binet–Simon scale included Ebbinghaus's method of relearning of lists (of words) as well as the sentence completion task.

IMPACT ON PSYCHOLOGY

Ebbinghaus's monograph received mixed reviews when it was published, but American psychologist William James praised the work, pointing out the author's 'heroic efforts'. To James, 'this particular series of experiments [was] the entering wedge of a new method of incalculable reach' (p. 199). Once Ebbinghaus's work became known in the USA, other psychologists began conducting studies of learning. Ebbinghaus became a model of the experimental psychologist, whose theoretical speculations were brief and cautious but whose research was rigorous in its method and its use of statistics. He provided a model for the use of experimental logic, including the testing of alternative hypotheses by setting up experimental situations where rival hypotheses would make differing predictions, and also a sensitivity to what are today called 'experimenter bias effects' (especially important to Ebbinghaus because he was his own subject). Finally, he provided experimental psychology with a model for preparing a research report: the now-traditional ordering of introduction, methods, results, and discussion sections.

Further Reading

Boring EG (1929) *A History of Experimental Psychology*. New York, NY: Appleton-Century-Crofts.

Ebbinghaus H (1885) *Über das Gedächtnis* ('On Memory') Leipzig: Duncker & Humblot. Translated by Ruger H and Busenius C (1913) New York, NY: Columbia University Teacher's College.

Fechner GT (1860) *Elemente der Psychophysik* ('Elements of Psychophysics'). Leipzig, Germany: Breithaus & Hartel.

Herrmann DJ and Chaffin R (1987) Memory before Ebbinghaus. In: Gorfein DS and Hoffman RR (eds) *Memory and Learning: The Ebbinghaus Centennial Conference*, pp. 35–56. Hillsdale, NJ: Lawrence Erlbaum.

Hoffman RR, Bringmann W, Bamberg M and Klein R (1987) Some historical observations on Ebbinghaus. In: Gorfein DS and Hoffman RR (eds), *Memory and Learning: The Ebbinghaus Centennial Conference*, pp. 57–76. Hillsdale, NJ: Lawrence Erlbaum.

Hothersall D (1984) *A History of Psychology*. New York, NY: Random House.

Jaensch ER (1909) Hermann Ebbinghaus. *Zeitschrift für Psychologie* **51**: 3–8.

James W (1885) Experiments in memory. *Science* **6**: 198–199.

Peterson J (1925) *Early Conceptions and Tests of Intelligence*. Chicago, IL: World Book Co.

Stigler SM (1978) Some forgotten work on memory. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 1–4.

Titchener EB (1910) The past decade in experimental psychology. *American Journal of Psychology* **21**: 404–421.

Woodworth RS (1909) Hermann Ebbinghaus. *Journal of Philosophy and Scientific Methods* **6**: 253–256.

Woodworth RS (1938) *Experimental Psychology*. New York, NY: Holt.

Fechner, Gustav Theodor

Introductory article

Stephen Link, University of California, San Diego, California, USA

CONTENTS

Gustav Theodor Fechner (1801–1887) was a German physicist and philosopher whose application of mathematical and scientific methods to psychological theory established the science of psychophysics and laid the foundations of experimental psychology.

INTRODUCTION

'In general, the clergy was strongly represented among our relatives, and I was also supposed to embark on this path, but somehow it turned out differently.' Gustav Fechner, son and grandson of pastors, was born on 19 April 1801 in the Saxony village of Großsärchen, now Zarki-Wielkie in Poland. At age 16 he matriculated at Leipzig University, where he spent 70 years first as a medical student, then as professor, chemist, physicist, psychophysicist, estheticist, nature philosopher, poet and satirist (Figure 1). His many theoretical and empirical discoveries enhanced the field of physics and created a basis for statistical hypothesis testing, descriptive statistics, experimental psychology and experimental esthetics.

Figure 1. Gustav Theodor Fechner (1801–1887).

LIFE AND WORKS

Following the Magisterexamen (rigorosum) at Leipzig University in 1823, Fechner spent ten years developing the theory of electricity, and his important experiments on Ohm's law that made him famous. His translations into German of books by the greatest French physicists of his day gave him the opportunity to meet them personally, to expand on their works and, in 1832, to publish the three-volume *Repertory of Experimental Physics*. These achievements led to his appointment as Professor in Ordinary at Leipzig University in 1834, the establishment of the Institute of Physics in Leipzig, and researches into the perception of color and afterimages.

In December 1839, overwork brought on by voluminous publications including the eight-volume, 7000-page *Hauslexicon* in 1837, and blindness caused by staring directly into sunlight to create vivid afterimages, led to a complete nervous collapse. Ellenberger characterized Fechner's illness as a 'sublime hypochondriasis, a creative illness from which a person emerges with a new philosophical insight and a transformation in their personality'. William James, a close follower of Fechner's work, described the illness as a 'habit neurosis'. For three years Fechner lived as a recluse, wearing a lead eye mask to prevent the pain caused by even the slightest illumination of his damaged eyes. His thoughts became uncontrollable, and his inability to consume food or liquids left him a skeleton near death.

Three years later, however, Fechner miraculously recovered. A transformation of his personality evidenced itself in his writings on the first law of the mind, 'the pleasure principle of action'. In 1846 Fechner argued that the search for pleasure and the avoidance of unpleasure were forces driving human behavior, in *Über das höchste Gut* (On the

highest good). The ideas appear fixed in *Über das Lustprincipl des Handelns* (1848). In 1848, his famous *Nanna, oder über das Seelenleben der Pflanzen* (On the soul life of plants) proposed that 'one can ask whether such a life (of animate creatures) pertains also to the plants, whether they too are animate individuals, combining in themselves impulses or sensations, or maybe more psychic experiences …. If this were so then plants along with men and animals would constitute a common contrast to stones and all things we call dead.'

Expanding on the idea of consciousness, in 1851 Fechner rendered a new version of *Zend-Avesta*, the sacred writings of the Persian prophet Zarathustra (Zoroaster). Fechner's *Zend-Avesta, über die Dinge des Himmels und des Jenseits: vom Standpunkt der Naturbetrachtung* (About heavenly things and the hereafter from the standpoint of contemplating nature) proposed that all life forms were self-aware and conscious. Even more, consciousness was in all and through all things. The Earth itself was conscious. Fechner also revealed that when he awoke on 22 October 1850 (now known as Fechner Day), he saw a relation between the body and mind, between the physical and mental, that became the basis for a new science and the first methods of mental measurement.

In particular, Fechner observed that a just noticeable difference (JND) in sensation is felt when a new stimulus increases in magnitude by a fixed proportion of the stimulus against which it is compared. For example, if a 312 g weight feels just noticeably different from a 300 g weight, then a 624 g weight will feel just noticeably different from a 600 g weight. Each JND is a unit of experience, as important to psychology as the mole is to chemistry or the quantum to physics.

Defining stimulus magnitude as a value S, and a just noticeable increment in stimulus magnitude as a value ΔS, Fechner developed the equation known as Weber's law:

$$\Delta S/S = \text{constant} \quad (1)$$

Fechner generalized this principle, and proposed that even smaller units of $\Delta S/S$ may be added together to create a measure of sensation magnitude. Letting dS represent a small increment in stimulus, Fechner created the differential equation

$$dS/S = \text{constant} \quad (2)$$

Integrating this equation adds together small relative increases in stimulation to form a total amount of sensation. The result is one of the most famous laws in psychology, relating sensation magnitude ψ to physical stimulus magnitude S, known as Fechner's law:

$$\psi = \log_e S \quad (3)$$

This formula set Fechner on a voyage of scientific discovery that marked the origin of experimental psychology.

THEORY OF MENTAL JUDGMENT

In *Elemente der Psychophysik*, published in 1860, the future of experimental psychology was set. In this two-volume, 907-page work, Fechner created the theory of judgment and methods of experimentation that still dominate psychological research. Fechner discovered that when comparing two stimuli, such as weights, a person could feel a just noticeable difference in weight but err in judging which of the two weighed more. For example, weights of 300 g and 312 g might feel just noticeably different but the 300 g weight might be judged heavier than the 312 g weight. This paradox led Fechner to propose a theory of judgment that predates the modern theory of statistical hypothesis testing.

In this regard, Fechner assumed that the sensory system is the mind's connective tunnel to the external world. A sensation or feeling such as heaviness must derive from the process of transforming into neural energy the energy of a physical stimulus. However, Fechner also assumed that the electrochemical sensory measurement system was not perfect. It suffered from the same form of inherent variability in measurement as Gauss had proposed in 1809 to affect physical measurement devices.

As shown in Figure 2, the transformation of weight to heaviness generates a Gaussian probability distribution of heaviness values. The true heaviness of a fixed 300 g weight equals $h(300)$, but because of sensory variability, the actual feeling of heaviness varies. Sometimes the weight feels heavier – and sometimes lighter – than the true heaviness of $h(300)$. Similarly, the true heaviness of a 312 g weight equals $h(312)$, but its heaviness varies according to the probability distribution shown in the upper part of Figure 2. The spread of heaviness values surrounding the mean values of $h(300)$ and $h(312)$ is characterized by the standard deviation of the Gaussian distribution, often denoted σ. Fechner suggested that this sensory variability was the basis for errors of judgment. As a theoretician, Fechner devised a theory of errors of judgment that allowed for a scientific measure of the unknown amount of sensory variability, σ.

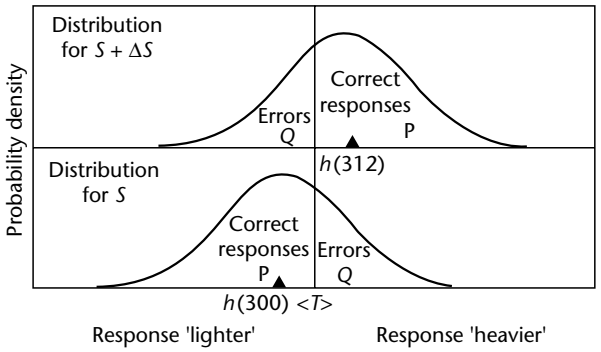


Figure 2. Fechner characterized the unseen variability within the nervous system as a Gaussian (normal) distribution. The abscissa represents heaviness, a psychological phenomenon. Two weights of 300 g and 312 g respectively generate average heaviness values of $h(300)$ and $h(312)$. The average of the two true heaviness values $h(300)$ and $h(312)$ equals T and is the threshold used (on the average) to determine which weight feels heavier.

The theory of judgment assumes that when two weights are compared, the average of the two heaviness values serves as a threshold for deciding which stimulus is heavier (or lighter). The stimulus that produces a heaviness greater than the average heaviness is judged to be the physically greater stimulus (*Elemente*, vol. 1). Converting these ideas into a mathematical form produces a measurement of the amount of variability in the nervous system in units of the physical stimulus. This extraordinary achievement, the first mental measurement, established psychology as a scientific discipline.

Figure 2 illustrates the essential features of Fechner's decision theory. Two weights of 300 g and 312 g are under comparison. On each comparison trial, the participant lifts each stimulus, experiences a sense of heaviness for each weight, and then reports which stimulus is greater in weight. According to the theory, Gaussian distributed variability perturbs the measure of heaviness. These distributions are shown in Figure 2 with heaviness means equal to $h(300)$ and $h(312)$. The threshold (criterion) for deciding which weight was heavier will change from trial to trial because the same heaviness values will not occur with each lifting of the weights. On the average, however, the threshold for deciding which weight is heavier is the average of the two mean heaviness values, equal to the value T shown in Figure 2.

The area to the right of T and under the Gaussian distribution for the heaviness of 312 g equals P. This is the probability of correctly judging the 312 g weight to be the larger weight. Also, the area Q, to the right of the criterion under the Gaussian distribution for the 300 g weight, equals the probability of an error in judging the smaller weight of 300 g to be the larger. Owing to the mirror symmetry of the Gaussian distributions with respect to T, and the requirement that areas under probability distributions must sum to 1, the values of P and Q must sum to 1.0. For a larger weight, say 324 g, the distribution of heaviness values shifts to the right. As a consequence, the threshold value T also shifts to the right, causing the value of P to increase and the value of Q to decrease.

To measure the amount of the unseen variability in the nervous system Fechner defined a measure of distance along the abscissa of Figure 2 in terms of the standard deviation, σ. Then he created mathematical tables showing how many errors of judgment occur for any value of T as measured in numbers of standard deviation units, σ. By running experiments to determine the probability of an error of judgment, and by comparing the error probability to entries of T in his tables, Fechner determined the number of standard deviation units separating T from $h(300)$. Using the Newtonian assumption that for very small differences in heaviness the function h is approximately linear, he determined that the threshold at T equaled $\frac{1}{2}(312 - 300)/\sigma$. Therefore,

$$\sigma = \tfrac{1}{2}(312 - 300)/T \qquad (4)$$

In this way Fechner measured the unseen force that resulted in errors of judgment. In this way he measured the unknown value of σ in units of the physical stimulus. In this way psychology became a science.

Elemente remained the basic work on experimental design until the 1935 appearance of R. A. Fisher's *Design of Experiments*. The theory still finds powerful applications. Thurstone developed a theory equivalent to Fechner's that yielded psychological measurement scales for such diverse stimuli as the seriousness of crimes, likeableness of vegetables and attitudes generally. The signal detection theory formulated by Tanner and Swets in 1954 allowed the value of T to vary as a function of experimenter inducements to bias judgments toward one response. Kinchla and Smyzer extended Fechner's theory in 1967 by providing a theory of visual position memory that predicted linear increases in σ^2 as a function of time.

The 'mirror effect' in recognition memory can be interpreted as a shift of Gaussian distributions of the memory strength along an abscissa of memory strength. As one distribution for memory strength shifts toward higher values, due to increased memory strength, the value of T must also shift

and the value of P in Figure 2 must increase while the value of Q must decrease. Stretch and Wixted showed that the value of T increases as word recognition memory strengthens, as Fechner's decision theory requires.

Elemente Volume 2 defines Fechner's law and describes 'inner psychophysics', the study of mind without regard to its sensory connections. The application of Weber's law, Fechner's law, and the threshold to inner psychophysics results in ideas about sleep and being awake, partial sleep, attention, and consciousness. Other chapters describe the wave scheme, relations between sensory and imagery phenomena, memory images, memory afterimages, the phenomena of sensory memory, psychophysical continuity and noncontinuity, hallucinations, illusions and dreams. Some fifty years later, Sigmund Freud commented, 'I … have followed that thinker on many important points.'

During his last 27 years Fechner created the field of experimental esthetics, continued his psychophysical investigations, and introduced ideas about descriptive statistics. In 1866 *Das Associationprincip in der Aesthetik* foreshadowed *Zur experimentellen Aesthetik* (1871) and the establishment of the field of experimental esthetics with the two-volume *Vorschule der Aesthetik* in 1876. Modern works on esthetic judgments by Beebe-Center, Hare, and Dorfman and colleagues extend and apply Fechner's and Thurstone's theories to art and emotion.

Returning to psychophysics, in 1877 Fechner published *In Sachen der Psychophysics*, in 1882 *Revision der Hauptpuncte der Psychophysik*, and in 1884 two more major experimental works. At Wilhelm Wundt's urging G. F. Lipps edited and published, posthumously, Fechner's last work *Kelletivemasslehre* (1897) a theory of data analysis that coined the term 'descriptive statistics'.

CONCLUSION

Fechner called upon the world to recognize the fundamental unity of the mind and physical reality. His many theoretical ideas changed over time and yet became foundations for a century and more of research and theory. What would psychology be without Fechner's psychophysics, without such important developments as signal detection theory, psychoanalysis, experimental memory, and attention and esthetics? Each of these mighty fields owes much to Fechner's originality of thought, integration of psychological theory with mathematics and experimental design, and firm belief that the physical and mental worlds form a single reality.

Further Reading

Anderson NH (1982) *Methods of Information Integration Theory*. New York: Academic Press.

Beebe-Center JG (1932) *The Psychology of Pleasantness and Unpleasantness*. New York: Van Nostrand.

Dorfman LYa, Leontiev DA, Petrov VM and Sozinov VA (1992) *Emotions and Art*. Perm: Institute for Arts and Culture.

Ellenberger HF (1970) *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. New York: Basic Books.

Falmagne JC (1985) *Elements of Psychophysical Theory*. New York: Oxford University Press.

Fechner GT (1966) *Elements of Psychophysics*, vol. 1, translated by HE Adler. New York: Holt, Rinehart & Winston. [Original work published 1860.]

Freud S (1935) *Autobiography*, translated by J. Strachey. New York: WW Norton.

Gauss CF (1809) *Theoria Motus Corporum Caelestium*. Hamburg: Certhes & Besser.

Gescheider GA (1997) *Psychophysics: The Fundamentals*. Mahwah, NJ: Erlbaum.

Green DM and Swets JA (1966) *Signal Detection Theory and Psychophysics*. New York: John Wiley.

Hare WF (1967) *Modern Aesthetics*. New York: Teachers College Press.

Kinchla RA and Smyzer F (1967) A diffusion model of perceptual memory. *Perception and Psychophysics* **2**: 219–229.

Laming DRJ (1986) *Sensory Analysis*. Orlando: Academic Press.

Link SW (1992) *The Wave Theory of Difference and Similarity*. Mahwah, NJ: Erlbaum.

Link SW (1992) Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science* **5**: 335–340.

Lowrie W (1946) *Religion of a Scientist: Selections From The Religious Writings of Gustav Theodor Fechner*. New York: Pantheon Books.

Luce RD (2000) *The Utility of Gains and Losses*. Mahwah, NJ: Erlbaum.

Macmillan NA and Creelman CD (1991) *Detection Theory: A User's Guide*. New York: Cambridge University Press.

Marks LE (1974) *Sensory Processes: The New Psychophysics*. New York: Academic Press.

Norwich KH (1993) *Information, Sensation, and Perception*. San Diego: Academic Press.

Sommerfeld E, Kompass R and Lachman T, eds (2001) *Fechner Day 2001*. Lengerich: Pabst.

Stevens SS (1975) *Psychophysics*. New York: John Wiley.

Stretch V and Wixted JT (1998) On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology, Learning, Memory and Cognition* **24**(6): 1–18.

Tanner WP and Swets JA (1954) A decision making theory of visual detection. *Psychological Review* **61**(6): 401–409.

Thurstone LL (1955) *The Measurement of Values*. Chicago: University of Chicago Press.

Vickers D (1979) *Decision Processes in Visual Perception*. New York: Academic Press.

Freud, Sigmund

Introductory article

Malcolm Macmillan, Deakin University, Burwood, Victoria, Australia

CONTENTS

Sigmund Freud founded psychoanalysis between 1895 and 1900 as a therapy for neuroses but developed it as a method for investigating mental processes and as a general psychology. He applied psychoanalysis to such fields as anthropology, sociology, and literature, and it became one of the most influential of all twentieth-century systems of thought. However, neither his original theory nor the variants descended from it have been widely accepted.

BACKGROUND, UPBRINGING, EDUCATION, AND INTERESTS

Sigmund Freud was born on 6 May 1856 in Příbor, Czechoslovakia. He was Jewish, his ancestors coming from Galacia. His family moved to Vienna when he was three, where he was educated and practiced medicine. When the Nazis invaded Austria in 1938 and persecuted the Jews there, they extorted a substantial ransom before allowing Freud and most of his family to go into exile in London. He died there on 23 September 1939. As a student Freud was well above average in academic ability. He read much general literature and philosophy and became deeply interested in biology and evolution but, as a Jew, medicine was one of the few careers open to him. Initially a neurohistologist, he began his medical practice as a neurologist.

NEUROLOGY AND PSYCHOLOGY

After beginning his medical studies in 1873, Freud enrolled in extra classes in physics, zoology, philosophy, and biology and Darwinism. He conducted histological work as a student and began his adult neurohistological research in 1876. Freud explicitly related his findings to the evolutionary-developmental framework in which he conducted this work.

Neurology and Affect

In 1885 Freud went to Paris to extend his neurohistological work at the clinic of Jean-Martin Charcot, where he became interested in Charcot's work on hysteria and hypnosis. Charcot produced anaesthesias and paralyses in hypnotic subjects and showed that the symptoms so produced were identical to those of hysteria. Using the minor trauma of indirect suggestion, by for example striking the subject on the arm, he seemed to show that hypnotic and hysterical symptoms formed when the affects of traumas caused a loss of ego control. Freud called the reflection in the symptoms of ideas and sensations from the trauma 'determining quality'. Together with the effects of the intense emotional state, it became central to his theories of mind.

Libido and Affect

After returning to Vienna in 1886 Freud practiced as a neurologist. Most of those of his patients with symptoms having no organic basis had neurasthenia, a supposed nervous weakness (asthenia), rather than hysteria. Sexual problems were frequently associated with it and Freud set out to establish its exclusively sexual aetiology. By the beginning of 1893 he claimed it was adolescent masturbation. Freud also differentiated anxiety neurosis – a sudden attack of anxiety with pronounced physiological manifestations such as increases in heart rate – from neurasthenia and proposed that its cause was incomplete sexual gratification (e.g. *coitus interruptus*).

Calling both neuroses 'actual neuroses', he theorized that they were caused by defective discharge of libido, that is, of the psychological sexual energy that formed when sexual ideas were charged or invested with physiological sexual energy. The loss of libido in neurasthenia caused general weakness; its deflection into organs like the heart caused

orgasm-like anxiety symptoms. Freud eventually generalized this thesis by proposing that the affects causing hysteria and obsessional neurosis were charges of libido.

FREUD AND THE PSYCHONEUROSES

Freud had first become interested in hysteria when Josef Breuer, his Viennese medical mentor, told him in 1882 how he had used talking to treat the hysterical symptoms of the pseudonymous Anna O. (Bertha Pappenheim). Under Freud's influence, Breuer explained Anna O.'s symptoms with 'French' dissociation theory: the 'traumatic' sensations that occurred in her elementary *secondary consciousness* returned as symptoms when it did. Her talking reconnected these secondary consciousness experiences with her *primary consciousness* and, according to Breuer, caused her symptoms to disappear.

Freud's Theory of Psychological Forces

Freud himself did not find that symptoms always formed in secondary states. Many patients recalled consciously trying to forget unacceptable ideas, which were then repressed by an unconscious mechanism. Freud's evidence for this ego force was the psychological effort he had to make to overcome the resistance patients had to recalling the unacceptable ideas. Initially Freud also drew on dissociation theory: repression separated the idea from its affect, pushed it into a secondary consciousness, and kept it there. Expressing the affect under light hypnosis re-established the previously subconscious association of the idea with normal consciousness and the symptom disappeared.

Freud soon abandoned hypnosis and dissociation theory. Pathogenic ideas could not always be recalled under hypnosis and repressed ideas could be retrieved in normal consciousness when patients fully reported what came to mind as they thought about their symptoms (*free association*). Both symptom formation and symptom removal took place in the waking state and these waking state concepts explained more: in hysteria the affect might be converted into physical symptoms and in obsessions displaced on to another idea, etc.

Affect, Sexuality, and Repression

Affect became central to Freud's theorizing when he concluded that effective therapy depended on how completely patients expressed or *abreacted* the emotions they had held back when their symptoms formed.

When first treating hysteria and obsessions (the psychoneuroses), Freud did not report that the unacceptable ideas had any particular content. But, from about 1895 they were always sexual, by 1896 they were always of perverse sexual experiences into which patients had been seduced as children, usually by an adult, and by 1897 the seducer was invariably the patient's father. Few patients recalled such experiences; most of the so-called 'memories' were constructed by Freud from their fragmentary recollections. This *childhood seduction theory* thus aligned the psychoneuroses with the actual neuroses. And, in accord with *Koch's postulates* from bacteriology, each neurosis had a specific sexual cause: masturbation for neurasthenia, incomplete gratification for anxiety neurosis, unpleasant childhood seductions for hysteria, and enjoyable for obsessions. Libido, not affect, thus powered repressed ideas.

NEURONAL AND NEURONAL-LIKE THEORIZING

In 1895 Freud outlined a physiological theory particularly directed to explaining why only sexual ideas were repressed. This unpublished work, called (in English) the *Project for a Scientific Psychology*, was much influenced by the *biophysics movement*, which tried to explain living phenomena by material processes. Freud eventually abandoned the *Project*: the deductions became too complex, some assumptions, especially about consciousness, became too *ad hoc*, and he never completed the section on repression.

A Neuronal-like Theory

In the next theory, the *topographic theory* of Chapter 7 of *The Interpretation of Dreams* (1900), Freud nevertheless retained the three kinds of neurons of the *Project*: permeable Φ neurons for perception, alterable Ψ neurons for recording memories, and ω neurons for consciousness, and a group of neurons with its own store of energy to delay responses (an ego). Apparently devoid of speculative physiology, the theory was still based on a teleological reflex and drew on notions of energy in flow and the cathexis or investment of ideas by it. Primary process was located in a system unconscious (*Ucs.*) and the secondary ego functions shared between the systems preconscious (*Pcs.*) and conscious (*Cs.*). *Pcs.* and *Cs.* repressed and controlled unacceptable ideas in *Ucs.*, allowing them only a disguised *Cs.* representation. Dreams, like symptoms, were fundamentally based on repressed sexual wishes.

Freud generalized these neuronal-like proposals and aligned them with his earlier theses that aspects of everyday mental life, such as faulty recall, slips of the tongue etc. (or parapraxes), resulted from conflicts between opposing conscious and unconscious forces. He could speak equally as well in the language of wishes as of neuronal activity.

SEXUAL AND OTHER INSTINCTUAL DRIVES

It was still a mystery why sexual ideas were so important and why only they were repressed. Freud tried to solve the problem by proposing that repression was a two-stage process that selectively affected the sexual drive.

Seduction Fantasies and Phylogenesis

Freud first proposed that the seduction 'memories' were fantasies caused by a childhood sexual drive. This childhood drive had separate *oral*, *anal*, and *phallic components*, each of which sought the same objectless, or *autoerotic*, perverse *modes* of satisfaction as in the perverse adult. Sucking and biting satisfied the oral component, fecal retention and expulsion the anal, and childhood masturbation the phallic. Freud claimed that some of these activities caused orgasm-like reactions. His analysis of symptoms also required such drives. A biological process of *primary repression* repressed each component drive according to a *phylogenetically* determined timetable.

Repression proper occurred when unacceptable ideas later revived these earlier, primarily repressed modes of satisfaction. Modes of satisfaction, just like other evolutionary developmental processes, could be fixated. Uncomplicated fixations explained perversions, and excessive repression and ego regression accounted for neuroses and psychoses respectively. Thus, schizophrenia and paranoia were based, respectively, on oral and anal fixations and regression, and hysteria on excessive repression of phallic sexuality. In this theory of psychosexual development, normal adult character traits were continuations of partial fixations. For example, eating, drinking, and dependence continued the oral mode, and orderliness, cleanliness, and parsimony the anal-retentive.

Finalizing the Instinctual and Structural Theories

Freud's theory logically required an *ego-instinctual drive* to repress the sexual instinctual drive in childhood when the ego was weak. He therefore recast basic mental conflict as one between the drives for self-preservation (ego) and preservation of the species (sexual). Self-love or narcissism undermined this polarity. Partly to resolve this difficulty and partly to explain the prominence of guilt and self-punishment in severe depression, Freud announced his final instinct theory. A death instinct, Thanatos, which aimed to return living matter to the inanimate state was now in conflict with a life instinct, Eros, comprising the old ego and sexual drives. Directed inward, Thanatos explained depression; outward explained aggression.

The two new drives were incorporated into Freud's third theory, the *structural theory* of 1923. Mental life was now an interaction between the *Id*, *Ego*, and *Superego*, and the Id was the repository of Thanatos and Eros as well as of repressed ideas. The Ego had many *Cs*. and *Pcs*. functions but was partly unconscious. It was powered by a *sublimated* form of Eros and only it sensed anxiety. The Superego was entirely new. Containing standards and conscience, it scrutinized behavior, punished infringements, and repressed sexual drives when the Ego generated signal anxiety. It formed in the *Oedipal phase* when the sexual and aggressive feelings of the child toward both parents were repressed. The child then *identified* with its lost sexual object and *incorporated* its standards into its Ego where they were cathected by Thanatos to form its own, harsh Superego. In this way the Oedipus complex was dissolved.

These drives, structures, and related developmental processes virtually completed Freud's general psychological theory.

SOME PROBLEMS WITH PSYCHOANALYSIS

Although Freud's influence is undeniable, the logical and empirical deficiencies of his central ideas means that their truth is still debated.

Structural Problems

First, psychoanalysts do not agree on the functions possessed by the structures of the third theory. Second, there is little agreement on when or how the Superego forms, and Oedipal identifications deliver feminine qualities to the boy's Superego and masculine to the girl's. Third, female sexuality develops tortuously, mainly because Freud insisted on a male starting point for both sexes. Fourth, how sublimation 'purged' libido of its overt sexuality is a mystery. Fifth, although

analysts do not agree on the sources of aggression most reject Thanatos.

Sexuality

Freud's clinical evidence for the role of sexuality is weak. There is also no direct evidence for a childhood sexual drive of the kind Freud's theory required. Observation shows that the supposed component drives are not sexual and that perverse adult behaviours do not lack objects or by themselves lead to orgasm. Empirical studies of the ways character traits cluster and whether they relate to early fixations also provide only weak support to the psychosexual theory. There is also a lack of evidence for the kind of repression that Freud proposed.

Free Association

Freud claimed that free association was as reliable and objective as the microscope. He insisted that he did not influence what the patient reported. However, analyses of psychotherapy sessions show that patients produce what their therapists search for, and the same thing appears to happen in the therapies conducted by psychoanalysts of different schools.

Interpretation

Freud believed that *interpretation* of the apparently meaningless ideas obtained by free association revealed their unconscious relations and uncovered their logically meaningful connections with the causes of the problem. For him these unconscious processes had the same reality as the unobservables of physics or chemistry. Psychoanalytic methods therefore allowed for trustworthy reconstructions of patients' histories.

Freud often compared interpreting free associations to deciphering or translating unknown languages. However, the translations of different psychoanalysts agree only minimally. Partly this is because there are no interpretive guidelines, but the more basic problem has been identified as the lack of knowledge of the grammar, syntax, and lexicon of the 'language' of the unconscious. When nothing is known about it, a language cannot be deciphered.

CONCLUSION

Were free association and interpretation as reliable and objective as Freud claimed, and were his theoretical concepts soundly based, analysts would collect and interpret essentially the same data in the same way and develop the theory from agreed-upon basic concepts. None of this happens. By 1930 irreconcilable differences had emerged among psychoanalysts, a fragmentation that has accelerated. Few analysts now accept all or even most of Freud's propositions. Nor can they define psychoanalytic theory or therapy. After more than a hundred years there should also be plenty of evidence to support at least one version of psychoanalysis, but there is none that has either a strong clinical or empirical base.

Further Reading

Ellenberger HF (1970) *The Discovery of the Unconscious*. New York: Basic Books.

Freud S (1953–1974) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vols. 1–24, edited and translated by J Strachey. London: Hogarth Press.

Jones E (1953–1957) *The Life and Work of Sigmund Freud*, Vols I–III. New York: Basic Books.

Macmillan M (1997) *Freud Evaluated: The Completed Arc*. Cambridge, MA: MIT Press.

Sulloway FJ (1992) *Freud, Biologist of the Mind: Beyond the Psychoanalytic Legend*, 2nd edn. Cambridge, MA: Harvard University Press.

Geschwind, Norman

Introductory article

Steven C Schachter, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
Orrin Devinsky, New York University School of Medicine, New York City, New York, USA

CONTENTS

Norman Geschwind (1926–1984) was a twentieth-century American neurologist considered to be the father of behavioral neurology because of his research and theories concerning higher cortical functions.

INTRODUCTION

Norman Geschwind (Figure 1) is a seminal figure in behavioral neurology because of his intensive clinical investigation and anatomical research on aphasia, isolation of the speech area, apraxia, agnosia, language-induced epilepsy, anatomical asymmetries of the brain, and cerebral dominance.

Geschwind was born in New York City in 1926. After graduating from Boys High School in Brooklyn, Geschwind attended Harvard College on a Pulitzer scholarship. He graduated *magna cum laude* with a concentration in mathematics in 1947, and then obtained a degree in medicine, *cum laude*, from Harvard Medical School in 1951. He interned in medicine at Boston's Beth Israel Hospital and then was awarded a Moseley scholarship to study at the National Hospital, Queen Square, London in 1952. Upon his return to the USA, he became chief resident on the Boston City Hospital neurological service under Derek Denny-Brown. After 2 years of axonal physiology research at the Massachusetts Institute of Technology, he joined the neurology service at Boston's Veterans Administration Hospital in 1958, becoming its chief in 1963. In 1966, with support from the National Institutes of Health, Geschwind started the Aphasia Research Center. That year he was also awarded the chair of the department of neurology at Boston University and began his research on higher cortical brain functions.

Returning to Harvard in 1969, Geschwind was named the James Jackson Putnam Professor of Neurology, succeeding Denny-Brown. The Harvard neurological unit moved from the Boston City Hospital to the Beth Israel Hospital in 1975, where Geschwind remained neurologist-in-chief until his death in 1984. At Beth Israel Hospital his interests flourished, particularly in behavioral neurology, language alterations, the relationship of one side of the brain to the other, the significance of large neuronal networks and their interconnections, and the relationships between brain development and other characteristics of both health and disease, particularly autoimmune illnesses.

Geschwind received three honorary degrees, two awards, and several honorary memberships and visiting lectureships. In 1978 he accepted a joint appointment at the Massachusetts Institute of Technology as a professor of psychology. He was elected president of the Boston Society of Neurology and Psychiatry and of the American Association of University Professors of Neurology. Geschwind was considered by many to be the father of behavioral neurology, and several of his many areas of interest, which extended over a remarkably long period, are outlined below. His work extended from the reestablishment of classic, predominantly French and German, theories to the creation of new frontiers in the understanding of brain–behavior relationships.

Geschwind was a physician, educator, colleague, advisor, role model, teacher, mentor, and creative genius. He inspired countless colleagues and students worldwide because of the brilliance of his ideas and his zeal for discussing them, his passion for history, an ability to see new relationships that others ignored, his appreciation for the value of detailed clinical and pathological case studies, and a knack for synthesizing observations from daily life into his hypotheses. His legacy in the field of

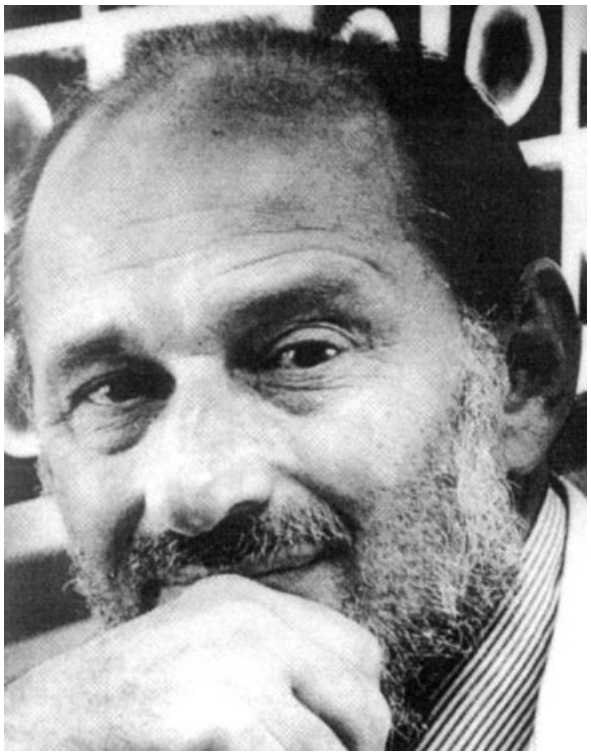


Figure 1. Norman Geschwind.

neurology is memorialized with the Norman Geschwind Prize in Behavioral Neurology, awarded annually by the American Academy of Neurology, and the annual Geschwind Visiting Professorship at the Beth Israel Deaconess Medical Center in Boston.

DISCONNECTION SYNDROMES

Geschwind's most powerful and persisting contribution to neurology was his work on disconnection syndromes in animals and humans. He resurrected the relevant German and French literature from the late nineteenth and early twentieth centuries and added new, carefully studied cases that revealed how neuroanatomy could shed light on the mechanisms of behavior.

Geschwind's interests in alexia, apraxia, and aphasia were largely responsible for the resurgence of attention of neuroscientists to brain–behavior relationships. Both alexia and apraxia provided insights into the cortical machinery of executing specific coordinated higher functions: reading, and skilled, learned movements, respectively. The role of the dominant parietal lobe in reading, writing, and praxis was further defined by these contributions, as were the effects of disconnecting left parietal input (alexia without agraphia) or output (one form of ideomotor apraxia). Although most neurologists considered apraxia to be rare, Geschwind emphasized that it was common, especially in patients with acute dominant hemisphere frontal lobe strokes. He stressed that the term should be used to describe a disorder characterized by impairment in executing a learned movement in response to a stimulus that would normally elicit the movement in the setting of intact sensation, strength, attention, and cooperation.

Geschwind revolutionized the conceptual framework of aphasic disorders. He subdivided aphasias into fluent and nonfluent groups, which largely corresponded with lesions that were anterior (nonfluent) or posterior (fluent) to the central sulcus. He then renewed interest in testing repetition, a simple maneuver that permitted identification of conduction or transcortical aphasias. Geschwind's approaches to the anatomy of language, the mechanisms underlying language dysfunction, the examination of aphasic patients, and the classification of aphasias continue to define the modern approach to examining and diagnosing patients with these disorders.

EPILEPSY

Geschwind studied the memory disorders that are commonly found in patients with epilepsy and emphasized the role of neuropsychological testing in the assessment of memory dysfunction. He was intrigued by personality changes associated with epilepsy and the schizophreniform psychosis that took many years to develop after its onset. He was concerned with elucidating the physiology of seizure triggers, language and epilepsy, behavioral changes following temporal lobectomy, and aggression in temporal lobe epilepsy (TLE). Most importantly, he described the interictal behavior syndrome of TLE. Geschwind felt strongly that psychiatrists should study TLE, because he contended that the associated personality changes might constitute the single most important condition in psychiatry. For Geschwind, TLE was a neurological model of psychiatric illness – a probe for the study of the physiology of emotion. He maintained that there was no other disorder characterized by alterations of behavior whose neurophysiological mechanisms were so well understood.

CEREBRAL DOMINANCE

Geschwind's interests in cerebral dominance, and the related topics of hemispheric asymmetries and 'handedness', grew out of his belief that the human

brain was endowed with anatomical asymmetries that could account for various aspects of cerebral dominance, such as language representation in the left hemisphere, which is found in 90% of right-handed people. Geschwind began his study of asymmetries and cerebral dominance with the temporal lobe. Together with a colleague, Geschwind cut 100 adult brains postmortem in the planes of the sylvian fissures and compared the appearance of the left and right plana temporale. In this landmark study, they found a larger left planum temporale in 65 of 100 brains, whereas in 35 brains, the plana were symmetric ($n = 24$) or the right planum temporale was larger ($n = 11$). This ratio of approximately 2:1 of left–right asymmetry was central to Geschwind's later concept of standard structural dominance.

During the early 1980s additional studies were performed and published, and Geschwind, in conjunction with his colleague Albert Galaburda, formulated a far-ranging hypothesis that integrated handedness, cerebral dominance, and autoimmune disease. The publication of this hypothesis, often called the Geschwind–Behan–Galaburda (GBG) hypothesis, stimulated innumerable additional investigations and papers, and generated considerable support and skepticism. As originally put forward by Geschwind and Galaburda, the hypothesis states, 'the most powerful factors [in determining the patterns of cerebral asymmetry] are variations in the chemical environment in fetal life and to a lesser extent in infancy and early childhood. The factors that modify cerebral dominance also influence the development of many other systems, e.g. the organs involved in immune response.'

The resistance to widespread acceptance of the GBG hypothesis derives from three sources. First, the study of the fetal origins of cerebral dominance is at odds with another influential handedness researcher and theorist, Marian Annett, who purports that the genetic theory of cerebral development is a more powerful determinant of cerebral development than the 'pathology model' of Geschwind. She argued that intrauterine factors are less important than genetic factors in determining brain structure. However, her theory and the GBG hypothesis actually complement one another. The latter approach may be able to shed light on the mechanism by which genes control cerebral development. For instance, identifying the fetal insults or factors (for example, testosterone) that modify brain structure and cerebral dominance from what would otherwise be predicted on strictly genetic grounds may provide clues to the mechanisms by which genes control normal brain development. Further, distinguishing the biologic associations of cerebral dominance and handedness may shed light on other genes that may be tightly linked with the 'right shift' gene, which according to Annett controls the development of handedness. For example, if autoimmune illness is associated with left-handedness, as was thought by Geschwind, then the search for the right shift gene could be directed toward those genes that are associated with autoimmune illness.

Second, the lack of consistency in handedness studies has resulted in a number of 'negative' studies of handedness and biological associations, casting further doubt on the GBG hypothesis. Yet many of these studies are inadequate because they do not include enough participants to demonstrate statistical significance, or they use measuring instruments that do not produce a handedness score that is proportional to the degree (for example, slightly right-handed versus strongly right-handed) and direction (left versus right) of handedness. Further, investigators often do not test the distribution of handedness scores for normality and yet use statistics for normally distributed data.

Third, and most importantly, the interpretation of the hypothesis has often been too narrow. Restatements of this hypothesis by other authors and critics often focus exclusively on the possible role of testosterone, though it is only one of many possible fetal factors. Geschwind did propose that testosterone was one of the major influences on brain development, but the essence of the hypothesis, the assertion that intrauterine factors influence fetal brain development, is a profound concept that further revises the nature–nurture argument. Conceptually, this was one of Geschwind's most insightful contributions.

Geschwind sought to increase our understanding of the interrelationships between brain structure and function. Like his other contributions, it was based on Geschwind's recognition of a connection – in this case between fetal factors, brain development, and human behavior. Though it was tragic, as pointed out by a colleague, that he died suddenly in 1984 before this opus was ever published in its complete form, neuroscientists around the world carry on his work, making new discoveries and thereby forging new frontiers made possible by Norman Geschwind.

Further Reading

Devinsky O and Schachter SC (eds) (1997) *Norman Geschwind: Selected Publications on Language, Behavior, and Epilepsy*. Boston, MA: Butterworth Heinemann.

Geschwind N (1965) Disconnexion syndromes in animals and man. I. *Brain* **88**: 237–294.

Geschwind N (1965) Disconnexion syndromes in animals and man. II. *Brain* **88**: 585–644.

Geschwind N (1972) Language and the brain. *Scientific American* **226**: 76–83.

Geschwind N (1973) Effects of temporal lobe surgery on behavior. *New England Journal of Medicine* **289**: 480–481.

Geschwind N (1974) *Selected Papers on Language and the Brain*. Boston, MA: Reidel.

Geschwind N and Behan P (1983) Left-handedness: association with immune disease, migraine, and developmental learning disorder. *Proceedings of the National Academy of Sciences of the USA* **79**: 5097–5100.

Geschwind N and Galaburda AM (1985) Cerebral lateralization. Biological mechanisms, associations, and pathology. I. A hypothesis and a program for research. *Archives of Neurology* **42**: 428–459.

Geschwind N and Galaburda AM (1985) Cerebral lateralization. Biological mechanisms, associations, and pathology. II. A hypothesis and a program for research. *Archives of Neurology* **42**: 521–552.

Geschwind N and Galaburda AM (1985) Cerebral lateralization. Biological mechanisms, associations, and pathology. III. A hypothesis and a program for research. *Archives of Neurology* **42**: 634–654.

Geschwind N and Levitsky W (1968) Human brain: left-right asymmetries in temporal speech region. *Science* **161**: 186–187.

Schachter SC (2000) Quantification and classification of handedness. In: Mandal MK, Bullman-Fleming B, and Tiwari G (eds) *Side-bias: A Neuropsychological Perspective*, pp. 155–174. Dordrecht, Netherlands: Kluwer.

Schachter SC and Devinsky O (eds) (1997) *Behavioral Neurology and the Legacy of Norman Geschwind*. Philadelphia, PA: Lippincott-Raven.

Waxman SG and Geschwind N (1975) The interictal behavior syndrome of temporal lobe epilepsy. *Archives of General Psychiatry* **32**: 1580–1586.

Waxman SG and Geschwind N (1980) Hypergraphia in temporal lobe epilepsy. *Neurology* **30**: 314–317.

Gibson, James J.

Introductory article

Eleanor Gibson, Cornell University, Ithaca, New York, USA

CONTENTS

James J. Gibson (1904–1979) was a leading American scholar and researcher in the field of perception. His dynamic and functional approach radically changed perceptual theories of the time and led to his own theory, the ecological approach to perception.

INTRODUCTION

James J. Gibson is known chiefly for his work on perception, which includes a novel theory that renounces the traditional view of perception as based on bare sensations that have to be supplemented by past experience or inferential processes in order to yield meaningful knowledge. He defined perception as an organism's means of acquiring information about the external world and his own relation to it, a functional view in contrast to the old, static sensation-based position, thus making him a rebel in the eyes of an older generation of perception theorists.

James Gibson was born in the American Midwest in 1904 and attended public schools in Wilmette, Illinois. He received his collegiate education at Princeton University and there obtained his PhD. He majored in both philosophy and psychology and was considerably influenced by one of his mentors, philosopher–psychologist E. B. Holt, a radical behaviorist. His first position was at Smith College as assistant professor of psychology, where he taught experimental psychology and with his students (one of whom later became his wife) performed many original experiments. Meanwhile, he began a program of research in perception. Experiments by his class on adaptation to curvature of lines in the layout, and other distortions caused by wearing prism spectacles, led him to examine the effect of inspecting curved and tilted lines without the spectacles. He found that after some seconds of inspection, there was adaptation to the curvature or tilt, and when a straight line was presented afterward, an after-effect occurred in the opposite direction. These experiments led him to the conclusion that spatial vision was not caused by stimulation of separate single receptors, but rather by larger structures such as gradients of change in the optic array – a new concept.

Wartime Research on Aviation

Early in the Second World War Gibson joined the US Army Air Corps as a research psychologist, and was assigned to a unit given the task of preparing tests for the selection of pilots and other aircrew such as bombardiers. Tests of the adequacy of spatial perception were especially important, but the research on perception of distance and orientation at that time was confined to the use of static cues such as retinal disparity, offering little for the task at hand. Gibson was sent to the Santa Ana Army Air Base in California, where he formed a film test unit. Films allowed him to study the role of motion in space perception, since it was clear that a flier needed to detect the location of other aircraft (and the ground) while in rapid movement. Gibson formulated a new approach to space perception based on information obtained from motion. He defined the role of what he named 'flow patterns' for judging the location of the flier in relation to the ground, other aircraft, and the external layout as a whole. He illustrated these flow patterns with diagrams (Figure 1); these diagrams now appear in all discussions of space perception. Not only fliers move; so do we all, moving our eyes, our heads and our positions. Other work of Gibson's during the war included research on aircraft identification, which eventually led him to a theory of perceptual learning as a process of differentiation rather than association, work that was followed up later with his wife, Eleanor Gibson.

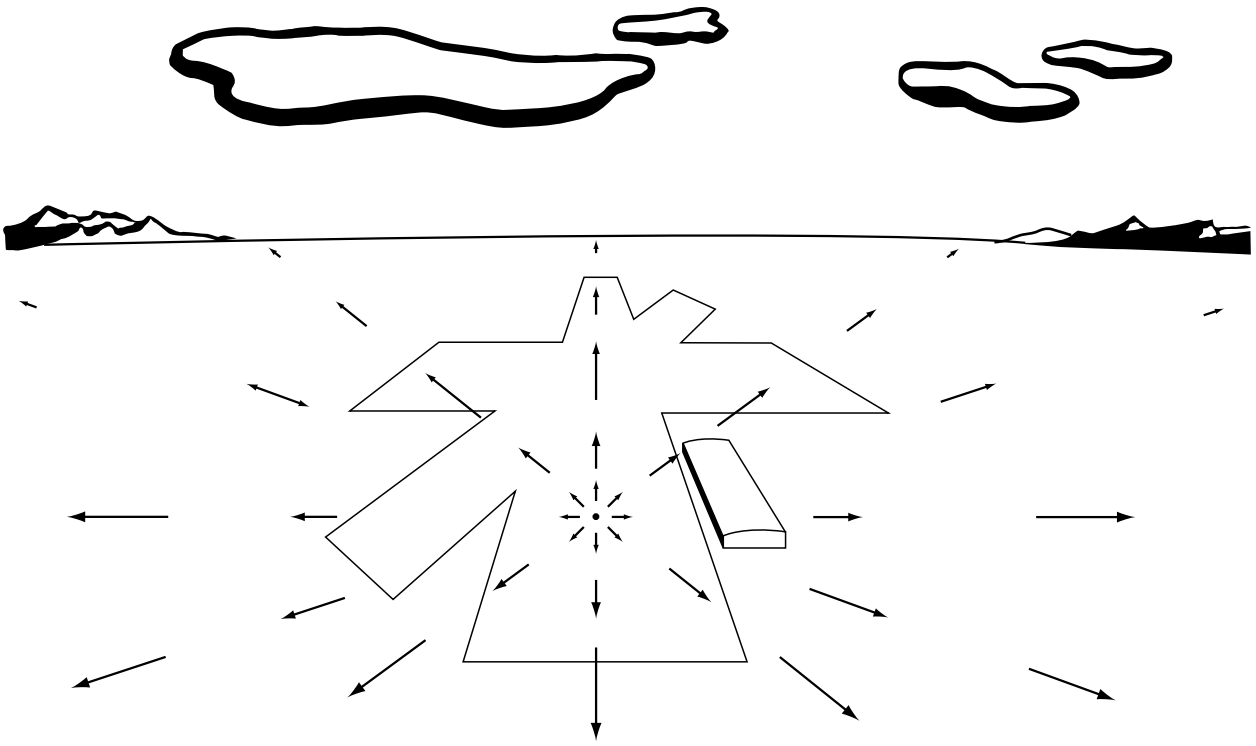


Figure 1. A flow pattern illustrating motion in an optical array when approaching a landing site.

A NEW THEORY OF PERCEPTION

Gibson's discoveries about the importance of motion as information for perceiving one's location in the spatial layout led him to reflect further on the nature of stimulation. He was by then sure that information for invariant properties of the environment must be extended over time, as well as space (which he had already discovered in the gradients that specify curvature and tilt of lines). The ground, in the natural human habitat, is also an essential ingredient in information for location and locomotion in the spatial layout. On his return from the war, Gibson began a book, *The Perception of the Visual World*, which would make these points and propose a new theory which would displace the traditional view that perception of the visual world begins with separate bits of stimulation giving rise to static, mosaic-like retinal images, and sensations, which are then supplemented by past experience and inference, to yield perception of a structured world in three dimensions. The old theory led to two views of the way the supplementation occurred, an 'empiricist' view and a 'nativist' view. Gibson's theory invoked neither one, since information for perception was rich and structured. The old research with static pictorial displays was useless and should be discarded. New research, pinning down the role of motion and dynamic displays that one directly perceived, was called for. The task for the psychologist was to find and describe the information. Three years later, when the book was finished, Gibson moved to Cornell University where he would find many graduate students to undertake the research called for.

THE PERCEPTUAL SYSTEMS

At Cornell, Gibson performed many experiments on the way motion gradients and transformations specified invariant structure in the world. Many of these experiments are described in his second book, *The Senses Considered as Perceptual Systems*. His view of perception as the work of systems emphasizes the active nature of perception, that information is obtained by perceptual systems, rather than stimulation falling passively on receptors. The book explains in detail other receptor systems besides the visual one, and shows how they cooperate in detecting the layout of the world and what is going on in it. Exploratory activity, such as looking around in visual search, and touching before seizing something or stepping on an unfamiliar surface, are examples.

According to this theory of active perception conducted by perceptual systems rather than

passive sensations, knowledge is certainly not innate, but the mechanisms of perceptual activity are to a great extent ready to go at birth. Research has shown that spontaneous activity, such as looking around and listening, is present at birth. Sensory equipment is not fully mature, as in the case of the visual system, but information about the external environment is obtained at once, and even newborns are motivated to obtain such information. This environment itself must be described in an account of what information is obtained. Gibson coined the term 'ecological optics' to refer to visual information in an ambient array, structured to specify the layout of things and the events occurring in it. Events, indeed, he considered the major source of information for perception.

THE ECOLOGICAL APPROACH

Radical as were these views, Gibson felt that he had not gone far enough in discarding the old view of perception based on a retinal image that had to be supplemented by past experience and reasoning processes.

What we actually perceive is a layout furnished with things that may or may not be useful to us and events that stir us to action. He felt that a sensory surface such as a retinal image interposed between the perceiver and this meaningful world was not of service for understanding a world full of action, including the perceiver's own. Action itself must be part of the repertoire of a perceiver, who shapes the environment by actions which are directly perceived as well as detected through the external events that they cause. The relation between exteroceptive and proprioceptive information is perceived and no inferential process is needed to detect it. Evolution has adapted us to detect proprioceptive information about our own actions and information ensuing from those actions. We are thus enabled to keep in touch with what is going on around us, allowing us to be attentive to what will happen next. The exploratory actions that are characteristic of perceiving result in perception that is prospective, heading toward the next action to be undertaken. Locomotion is a prime example. As we move towards some surface, information about the layout, where we are in it, and where we might be going are all specified at the same time. There is a reciprocity between action and perception that requires a description of both environmental structure and the perceiver's actions with respect to it.

The ecological approach to perception, as Gibson called his new view, retained the innovations expressed in his earlier books: that stimulation goes on over time, providing invariant information, with gradients over space and those carried by motion; that such information specifies structure in the world that is directly perceived; that perception is an active process of seeking this information; and that perceivers detect their own actions and their consequences.

The relation between a perceiver and the environment is thus one of reciprocity. This notion led to a new concept, that of 'affordances'. Gibson coined the word to refer to the resources of the environment with respect to their potential use to an organism, relating properties of the physical world to the functional capacities of a given organism. The 'graspability' of an object, the 'traversibility' of a ground surface, and the 'edibility' of substances are examples. When such affordances are perceived, the organism is perceiving the meaning of them. To quote a well-known sentence of Gibson's, 'The affordances of the environment are what it offers animals, what it provides or furnishes, for good or ill.' Surfaces may be 'stand-on-able' or 'fall-off-able'. They may be negative with respect to an animal's welfare as well as positive, harmful or beneficial. They may be social, such as the affordances for play, sex or communication. In any case, they are all specified in information obtainable by the perceptual systems of the animal – visual, auditory, proprioceptive and so on. They are real, as well as meaningful when they are picked up. Picking them up may involve a learning process, known as 'perceptual learning.'

It is evident that the relationship between an animal and its environment is one of mutuality, both specialized for each other, either by evolution or by alteration of the environment by acts such as the making of tools, or changing natural environments by artificial means such as heating or cooling. In either case, the relation is an invariant objective one, potentially detectable by an organism for which it has usefulness for existence or comfort. Affordances, Gibson affirmed, are the foundation of values.

Gibson's views on ecological psychology were received as a major contribution by many psychologists. One American reviewer (Restle) wrote in the journal *Contemporary Psychology*, 'This book comes forth as a major theory, as the culmination of the life's work of Jimmy Gibson, our one original irreplaceable creative genius.' Shortly after Gibson's death in 1979, friends and students, both American and European, formed the International Society for Ecological Psychology. At the time of writing, this

group continues to thrive and attract new members. It holds annual meetings and sponsors the *Journal of Ecological Psychology*.

CONCLUSION

The dean of American psychologists, E. G. Boring, wrote of Gibson's first book, 'The book's appearance becomes, in a sense, an event within a tradition: Goethe (brilliant, erudite, dogmatic, wrong); Purkinje (keen, usually right); Hering (argumatic, pedantic); and now Gibson creating the paradox of a phenomenology of vision which is both modern and American.' In the end, however, Gibson was hardly a phenomenologist, proposing an ecological psychology that included the world: a paradox, perhaps, for history.

Further Reading

Gibson EJ (2002) *Perceiving the Affordances: A Portrait of Two Psychologists*. Hillsdale, NJ: Lawrence Erlbaum.

Gibson EJ and Pick AD (2000) *The Ecological Approach to Perceptual Learning and Development*. New York, NY: Oxford University Press.

Gibson JJ (1967) James J. Gibson. In: Boring EG and Lindzey A (eds) *A History of Psychology in Autobiography*, pp. 127–143. New York, NY: Appleton-Century-Crofts.

Gibson JJ (1977) The theory of affordances. In: Shaw R and Bransford J (eds) *Perceiving, Acting and Knowing*, pp. 67–82. Hillsdale, NJ: Lawrence Erlbaum.

Gibson JJ and Crooks LE (1938) A theoretical field-analysis of automobile driving. *American Journal of Psychology* **51**: 453–471.

Lee DN (1980) The optic flow field: the foundation of vision. *Philosophical Transactions of the Royal Society Series B* **290**: 169–179.

Lombardo TJ (1987) *The Reciprocity of Perceiver and Environment*: *The Evolution of James J. Gibson's Ecological Psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Neisser U (1985) Toward an ecologically oriented cognitive science. In: Shlechter TM and Taglia MP (eds) *New Directions in Cognitive Science*, pp. 17–32. Norwood, NJ: Ablex.

Reed ES (1988) *James J. Gibson and the Psychology of Perception*. New Haven, CT: Yale University Press.

Reed ES and Jones R (eds) (1982) *Reasons for Realism: Selected Essays of James J. Gibson*. Hillsdale, NJ: Lawrence Erlbaum.

Turvey MT (1992) Affordances and prospective control: an outline of the ontology. *Ecological Psychology* **4**: 173–187.

Hebb, Donald Olding

Intermediate article

Raymond M Klein, Dalhousie University, Halifax, Nova Scotia, Canada

During his lifetime Donald Olding Hebb (1904–1985) was an extraordinarily influential figure in psychological science and behavioral neuroscience. In the middle of the twentieth century his principled opposition to radical behaviorism, his critical analyses of the shortcomings of the then dominant theories of learning and perception, and his emphasis on understanding what goes on between stimulus and response helped clear the way for a revolution in North American psychology. Hebb's view of psychology as a biological science helped rejuvenate interest in physiological psychology, and his simple and appealing neuropsychological cell-assembly proposal served as a magnet for creative scientists and a stimulus for theoretical and empirical advances. Since his death, Hebb's seminal ideas wield an ever-growing influence on scholars interested in mind (cognitive science), brain (neuroscience), and how brains implement mind (cognitive neuroscience).

Born to physician parents, in Chester, Nova Scotia, in the Maritime region of Canada, Hebb attended Dalhousie University, from which he graduated in 1925. Though he aspired to write novels, Hebb chose instead to earn a living as a public school teacher and soon became a school principal in the Province of Quebec. Hebb notes that his interest in psychology was excited by the writings of James, Freud, Watson, and Pavlov. Hebb was accepted as a part-time student to the McGill graduate program, and got a position as a research assistant with Dr L. A. Andreyev, a visiting scientist from Pavlov's laboratory. Though this led to a Master's degree, Hebb was unimpressed with Pavlov's program, and was, as he would say, 'softened up for my encounter with Kohler's Gestalt Psychology and Lashley's critique of reflexology'. Hebb went to work with Lashley at the University of Chicago; moved with Lashley to Harvard University; and in 1936 completed his PhD on the effects of early visual deprivation upon size and brightness perception in the rat. (*See* **Lashley, Karl S.**)

The first of two pivotal postdoctoral experiences was made possible when Wilder Penfield offered Hebb a fellowship at the Montreal Neurological Institute (MNI). At the MNI Hebb explored the impact of brain injury and surgery, particularly lesions of the frontal lobes, on human intelligence and behavior. From his observations that removal of large amounts of tissue seemed to have little impact on memory and intelligence, Hebb inferred a widely distributed neural substrate (Hebb, 1945).

Hebb then briefly joined the faculty at Queens University, Kingston, Ontario, Canada, where he developed human and animal intelligence tests, including the 'Hebb-Williams' mazes, which subsequently have been used in hundreds of studies to investigate the intelligence of a wide range of species, including even humans – thanks to the development of virtual versions of the mazes, making it the 'Stanford–Binet' of comparative intelligence (Brown and Stanford, 1997). Hebb's studies of human and animal intelligence led him to conclude that experience played a much greater role in determining intelligence than was typically assumed (Hebb, 1942). Though Hebb would later emphasize the contribution of heredity this was an explicit effort to reverse the pendulum of opinion on the nature/nurture issue, which Hebb thought had swung too far in the nurture direction (Hebb, 1953). Nevertheless, Hebb maintained that the question 'to what extent is a given piece of behavior dependent on one of these influences?' was nonproductive, and his cogent analysis is often summarized with the following analogy which he used to explain why:

> Is it fifty-percent environment, fifty-percent heredity, or ninety to ten, or what are the proportions? This is exactly like asking how much of the area of a field is due to its length, how much to its width. The only reasonable answer is that the two proportions are one-hundred-percent environment, one-hundred-percent heredity. They are not additive; any bit of behavior, whatever, is fully dependent on each. (Hebb, 1953 p. 44)

In 1942 Hebb left Queens University and rejoined Lashley, who had become director of the Yerkes Laboratories of Primate Biology in Florida. Although his plans to interact with Lashley and explore the effects of various brain lesions on the performance and temperament of adult chimpanzees did not materialize, Hebb's experiences at the Yerkes Laboratories were, like those at the MNI,

pivotal for the development and impending dissemination of his ideas. Instead of the effects of lesions, Hebb explored fear, anger, and other emotional processes in the intact chimpanzee. He would later say of this exposure that he 'learned more about human beings during that time than in any other 5-year period of my life except the first' (Hebb, 1980). During his time at the Yerkes, Hebb began writing a book, *The Organization of Behavior: A Neuropsychological Theory* (1949), in which he synthesized different lines of research into a 'general theory of behavior that attempts to bridge the gap between neurophysiology and psychology' (p. vii). Hebb credits 'the weekly colloquium and the persistent theoretical debate at the Yerkes Laboratories of Primate Biology' (p. viii) for providing an invaluable intellectual climate for the development and refinement of his ideas.

Hebb returned to McGill as Professor of Psychology and in 1948 was appointed chair of the department. The book he wrote while at the Yerkes Laboratories appeared in 1949. In the years after its appearance the impact of *The Organization of Behavior* was great and the influence of Hebb and his ideas grew steadily. McGill University became a North American mecca for scientists interested in the brain mechanisms of behavior, and the proposals put forward in Hebb's book led to many important discoveries and steered contemporary psychology onto a more fruitful path.

For Hebb 'the problem of understanding behavior is the problem of understanding the total action of the nervous system, and vice versa' (1949, p. xiv) and his advocacy of an interdisciplinary effort to solve this neuropsychological problem was his most general theme. Hebb's book provided a rallying point for those interested in the brain mechanisms of mind, and because there was a powerful movement in psychology to reject physiological concepts (Skinner, 1938), his book marked a turning point away from this trend. The history of attempts to explain behavior and thought is punctuated by metaphors put forward by scholars in the hope that understanding will be illuminated by reference to the properties of well-understood nonbiological entities. Psychological theory, for example, has its roots in the 'mental chemistry' of the British Associationists; hydraulic concepts figure prominently in ethological and psychodynamic conceptions of motivational influences upon behavior; and the Gestalt psychologists relied on magnetic fields to help understand perceptual organization. More recently, the cognitive revolution was inspired by the idea that the mind is like a general-purpose digital computer with inputs and outputs, storage devices, and a central processor that performs operations one at a time. Hebb shunned such metaphors and sought instead to develop a theory of human and animal behavior and thought in terms of the actual device which produces them – the neural machinery in the brain. In *The Organization of Behavior*, Hebb presented just such a neuropsychological theory.

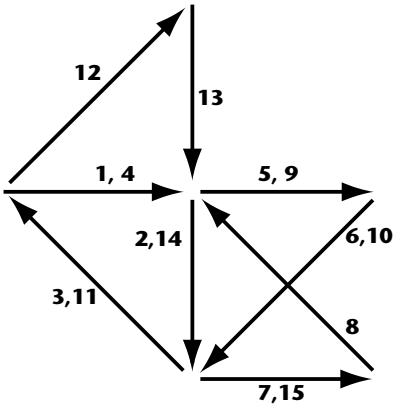


Figure 1. Schematic of Hebb's 'cell-assembly' hypothesis. 'Arrows represent a simple "assembly" of neural pathways or open multiple chains firing according to the numbers on each (the pathway "1,4" fires first and fourth, and so on), illustrating the possibility of an alternating reverberation which would not extinguish as readily as that in a simple closed circuit.' (Redrawn from Hebb, 1949, Figure 10, p. 73.)

There were three pivotal postulates. First, the connections between neurons increase in efficacy in proportion to the degree of correlation between pre- and post-synaptic activity. In Hebb's own words, from Chapter 4 of *The Organization of Behavior*: 'When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased' (p. 62). In neuroscience this proposal is referred to as the 'Hebb synapse'. Speculative and hypothetical in 1949, real examples were subsequently revealed in long-term potentiation (Bliss and Lomo, 1973) and kindling (Goddard *et al*., 1969). In cognitive science this postulate, which is known as the 'Hebb rule', provides the most basic learning algorithm for adjusting the strength of connections between nodes in artificial neural network models.

The second postulate is that when groups of neurons tend to fire together they form a *cell-assembly* whose activity can persist after, and serves to represent, the original triggering event. This proposal, that the joint activation of an assembly of

cells represented a sensation or idea (see Figure 1), foreshadowed the wide adoption of distributed representation in models of natural and artificial intelligence and is considered by some (Milner, 1986) to be Hebb's most important conceptual contribution.

The third postulate is that thinking is the sequential activation of sets of cell-assemblies. In the Introduction to *The Organization of Behavior* Hebb provides this summary of the theory that is developed in the subsequent chapters:

> Any frequently repeated, particular stimulation will lead to the slow development of a 'cell-assembly,' a diffuse structure comprising cells in the cortex and diencephalon (and also, perhaps, in the basal ganglia of the cerebrum), capable of acting briefly as a closed system, delivering facilitation to other such systems and usually having a specific motor facilitation. A series of such events constitutes a 'phase sequence' – the thought process. Each assembly action may be aroused by a preceding assembly, by a sensory event, or – normally – by both. The central facilitation from one of these activities on the next is the prototype of 'attention.' … The theory is evidently a form of connectionism … though it does not deal in direct connections between afferent and efferent pathways: not an S–R [stimulus–response] psychology, if R means *muscular* response. … It does not, further, make any single nerve cell or pathway essential to any habit or perception. (p. xix)

Hebb's book was greeted enthusiastically. Oliver Zangwill (1950) noted that Hebb has 'made an original and exciting beginning' towards a framework for linking psychological phenomena with principles of nervous system organization, and Fred Attneave (1950) called it 'the most important contribution to psychological theory in recent years'. The sheer magnitude of the contribution and the degree of excitement Hebb's book, ideas and leadership would generate was not anticipated, nor was the seminal role his proposals would attain.

Hebb acknowledged that his theory was speculative and incomplete. Missing from the model, for example, was neural inhibition (Milner, 1957), a concept Hebb later incorporated (Hebb, 1959) and to which he would attach great significance (Hebb, 1980a). At this early stage in the development of theories linking brain and mind (neuropsychological theories) speculation was a virtue and incompleteness a necessity. Hebb himself noted that a primary role of our merely momentarily correct theories is to stimulate scientific discovery and that 'one's only strategy is to interest intelligent people of diverse skills, interests and knowledge, in the problems as one sees them' (Hebb, 1959). Hebb was extremely fruitful in implementing this strategy. Many important psychologists and behavioral neuroscientists of the latter half of the twentieth century trained directly with Hebb, and other scholars who came to McGill during his era benefited from their contact with him. Whole literatures on the role of early experience in perceptual development (Hunt, 1979), on sensory deprivation (Zubek, 1969), self-stimulation (Olds and Milner, 1954), the stopped retinal image (Pritchard *et al.*, 1960), the neural basis of pain (Melzack, 1996), synaptic modifiability (Goddard, 1980), and learning without awareness (McKelvie, 1987), were provoked or fostered by Hebb's proposals.

Hebb's seminal ideas of 1949 are being applied not only in neurophysiology, neuroscience, and psychology, but also in engineering, robotics, and computer science. In these diverse scientific literatures references to Hebb, the Hebbian cell-assembly, the Hebb synapse, and the Hebb rule increase each year. This widespread influence is a testament to Hebb's scientific acumen, creativity, and courage to put forth a foundational neuropsychological theory of the organization of behavior. One of Hebb's professors at Harvard, E. G. Boring, described psychology as the mid-nineteenth-century offspring of the scientific method of physiologists and the preoccupation with mind of philosophers (Boring, 1950). In the mid-twentieth century, by providing a neural implementation of the Associationists' mental chemistry, Hebb's ideas nurtured the young and developing science and laid the foundation for neoconnectionism which seeks to explain cognitive processes in terms of connections between assemblies of neurons, real or artificial.

Hebb won many honors and awards and held many positions of leadership. Among these: he was named Fellow of the Royal Society of Canada and of the Royal Society (London); he won the American Psychological Association Award for Distinguished Scientific Contribution; and he served as President of the Canadian and American Psychological Associations. For the reader interested in learning more about Hebb's life and the evolution of his ideas, his own articles (Hebb, 1959, 1980b) and those by Glickman (1996), Klein (1980), and Milner (1986) are recommended.

References

Attneave F (1950) Review of *The Organization of Behavior* by D. O. Hebb. *American Journal of Psychology* **63**: 633–635.

Bliss TVP and Lomo T (1973) Long lasting potentiation of synaptic transmission in the dentate area of the

anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–356.

Boring EG (1950) *A History of Experimental Psychology*, 2nd edn. New York: Appleton-Century-Crofts.

Brown RE and Stanford L (1997) The Hebb–Williams Maze: 50 years of research (1946–1996). *Society for Neuroscience Abstracts* (#110.15) **23**: 278.

Glickman S (1996) Donald Olding Hebb: returning the nervous system to psychology. In: Kimble G, Boneau C and Wertheimer M (eds) *Portraits of Pioneers in Psychology*, vol. 2. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goddard GV (1980) Component properties of the memory machine: Hebb revisited. In: Jusczyk PW and Klein RM (eds) *The Nature of Thought: Essays in Honor of D. O. Hebb*, pp. 231–247. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goddard GV, McIntyre DC and Leech CK (1969) A permanent change in brain function resulting from daily electrical stimulation. *Experimental Neurology* **25**: 295–330.

Hebb DO (1942) The effects of early and late brain injury upon test scores, and the nature of normal adult intelligence. *Proceedings of the American Philosophical Society* **85**: 275–292.

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Hebb DO (1953) Heredity and environment in mammalian behavior. *British Journal of Animal Behavior* **1**: 43–47.

Hebb DO (1959) A neuropsychological theory. In: Koch S (ed.) *Psychology: A Study of a Science*, vol. 1. New York: McGraw-Hill.

Hebb DO (1980a) *Essay on Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hebb DO (1980b) D. O. Hebb. In: Lindzey G (ed.) *A History of Psychology in Autobiography*, vol. VII. San Francisco, CA: WH Freeman.

Hunt JM (1979) Psychological development: early experience. *Annual Review of Psychology* **30**: 103–143.

Klein RM (1980) D. O. Hebb: an appreciation. In: Jusczyk PW and Klein RM (eds) *The Nature of Thought: Essays in Honor of D. O. Hebb*, pp. 1–18. Hillsdale, NJ: Lawrence Erlbaum Associates.

McKelvie S (1987) Learning and awareness in the Hebb digits task. *Journal of General Psychology* **114**: 75–88.

Melzack R (1996) Gate control theory: on the evolution of pain concepts. *Pain Forum* **5**: 128–138.

Milner PM (1957) The cell assembly: Mark II. *Psychological Review* **64**: 242–252.

Milner PM (1986) The mind and Donald O. Hebb. *Scientific American* **268**: 124–129.

Olds J and Milner PM (1954) Positive reinforcement produced by electrical stimulation of the septal area and other regions of the rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–427.

Pritchard RM, Heron W and Hebb DO (1960) Visual perception approached by the method of stabilized images. *Canadian Journal of Psychology* **14**: 67–77.

Shore DI, Stanford L, Mac Innes WJ, Klein RM and Brown RE (2001) Of mice and men: using virtual Hebb–Williams mazes to compare learning across gender and species. *Cognitive, Affective and Behavioral Neuroscience* **1**: 83–89.

Skinner BF (1938) *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century.

Zangwill OL (1950) Review of *The Organization of Behavior* by D. O. Hebb. *Quarterly Journal of Experimental Psychology* **2**: 142–143.

Zubek P (1969) *Sensory Deprivation: 15 Years of Research*. New York: Meredith.

Helmholtz, Hermann von

Introductory article

Warren R Street, Central Washington University, Ellensburg, Washington, USA
Philip Tolin, Central Washington University, Ellensburg, Washington, USA

CONTENTS

Hermann von Helmholtz (1821–1894) was one of the most productive and versatile scientists of the nineteenth century. His experimental, theoretical, and technical accomplishments advanced the fields of optics, physiology of hearing and vision, epistemology, electrodynamics, hydrodynamics, meteorology, mathematics, music, and the emerging science of experimental psychology.

INTRODUCTION

Hermann Ludwig Ferdinand von Helmholtz was born on 31 August 1821 in Potsdam, Prussia, the most powerful of the states that merged to form the German empire in 1871. His mother, Caroline Penn Helmholtz, was a descendant of the American colonist William Penn and was a reserved, intelligent and encouraging parent. His father, August Ferdinand Julius Helmholtz, was a teacher of classical languages at the Potsdam gymnasium and an admirer of the philosophies of Immanuel Kant and Johann Gottlieb Fichte. (*See* **Kant, Immanuel**)

Helmholtz was frequently ill as a child and did not attend school until the age of seven. Bolstered by an intellectually stimulating home environment, he progressed rapidly through his studies. An inspiring mathematics teacher ignited a lifelong interest in physics and optics. In classes that did not interest him, Helmholtz sometimes ignored his assigned reading and worked on optics problems underneath his desk.

By the age of 14 years, Helmholtz had declared an interest in becoming a scientist, and at 17 years old he was admitted to the Friedrich Wilhelm Institute of Medicine and Surgery in Berlin. While physics was Helmholtz's preferred field, medicine provided a route to advanced studies that his family could afford. The institute provided full scholarships in exchange for 8 years' service as a military surgeon.

Helmholtz carried out his doctoral studies under the direction of the great physiologist Johannes Müller. Despite his exhaustive knowledge of physiology, Müller was a vitalist – a person who believes that life ultimately is directed by intangible forces that have no physical properties. Some of Helmholtz's earliest work provided evidence that these hypothetical vital forces were, at best, irrelevant to understanding living organisms. Helmholtz was one of a group of gifted physiology students who formalized their support of mechanism and opposition to vitalism by swearing allegiance with an oath that began, 'No other forces than the common physical-chemical ones are active within the organism'.

Helmholtz earned his medical degree in 1842 with research showing that nerve fibers are extended parts of nerve cells. In 1843 he was appointed assistant surgeon to the Royal Hussars at Potsdam and began discharging his military obligation. In his spare time he conducted studies in an improvised laboratory. This work culminated, in 1847, in a convincing argument for one of the fundamental principles of modern physics: the principle of conservation of energy.

Helmholtz clearly possessed greater promise as a scientist than as a military surgeon. In 1848, he was released from his remaining 3 years of military service to fill a teaching position in anatomy in Berlin. A year later he was invited to become extraordinary (associate) professor of physiology in Königsberg, now the Russian city of Kaliningrad, on the Baltic sea. His financial security permitted him to marry his fiancée of more than 2 years, Olga von Velten. Her fragile health declined in Königsberg, and the family was advised to move to a warmer climate. By the time he left in 1855 to take up the chair of physiology at Bonn, Helmholtz had invented the ophthalmoscope (1850), completed historic work on the speed of nervous

conduction (1850), and studied the focusing motion of the lens of the eye (1855). He had begun his work on color vision and his theory of unconscious inference in human perception.

In Helmholtz's 3 years at Bonn he completed the first volume of his comprehensive *Treatise on Physiological Optics* (1856) and began to study the physical and psychological nature of sound. An attractive offer to become a professor of physiology drew him to Heidelberg University. Olga Helmholtz's health continued to deteriorate, and she and Helmholtz's father both died in 1859. Helmholtz found solace in his research, and then married Anna von Mohl in 1861. His research led to the second (1860) and third (1867) volumes of the *Treatise on Physiological Optics* and, in 1863, his landmark volume on sound and hearing, *On the Sensations of Tone as a Physiological Basis of the Theory of Music*. He devoted increasing energy to important work in physics and mathematics and in 1871 was appointed to the chair of physics at Berlin, the nation's most prestigious physics position.

The last 23 years of Helmholtz's life were ones of fame and undiminished scientific productivity, primarily in physics. In 1882 Helmholtz was made a member of the nobility and the traditional 'von' was appended to his name. The psychophysics of Gustav Fechner attracted Helmholtz's interest, and he published four articles on the topic. Helmholtz suffered a severe fall returning from America in 1893, recovered only slowly, and in June 1894 suffered a stroke. His health rapidly declined, and he died on 8 September 1894, leaving a remarkable body of research and interpretation to modern science. (*See* **Fechner, Gustav Theodor**)

HELMHOLTZ ON VISION

Helmholtz's three-volume *Treatise on Physiological Optics* integrated previous work with his own studies of the physiological, sensory, and perceptual aspects of vision. For example, he measured many of the optical characteristics of the eye, demonstrated how the eye focuses on different distances, and developed a theory of visual space. He invented the ophthalmoscope, a device still used for visual examination of the retina. Most important, however, was his theory of color vision. (*See* **Color Vision, Neural Basis of**; **Color Perception, Psychology of**)

In 1802 Thomas Young had suggested that the retina contains three kinds of color-sensitive receptors and that the three associated 'primary' colors combine in some fashion to produce the complete range of colors that humans experience. Helmholtz's 1858 elaboration of Young's theory proposed that three types of color-sensitive light receptors in the retina give rise to experiences of red, green, or blue–violet light when stimulated. Each type of receptor contains a chemical that reacts with differing sensitivity to different wavelengths of light, exciting neurons and sending impulses to different parts of the brain. Helmholtz used his theory to explain a variety of visual experiences, including the results of mixing colors, color blindness, and negative afterimages. While the Young–Helmholtz theory has been modified, elaborated and corrected, it provided the foundation for the current understanding of vision.

HELMHOLTZ ON HEARING

Helmholtz made many contributions to understanding sound and hearing, including his explanation of the phenomenon of timbre, or tonal quality. A violin playing middle C has a timbre different from that of a trombone playing the same note. Helmholtz explained that the difference is created by different patterns of overtones, or harmonics, added to the base frequency. In typical fashion, Helmholtz did not simply advance a theory; he also provided inventive demonstrations. For example, he built a series of tuning forks to vary the intensity of different overtones and produce synthetically the timbres of various musical instruments. His work provided early theoretical and empirical bases for modern music and speech synthesizers.

In addition to advances in the physics of sound, Helmholtz made major discoveries about the anatomy and physiology of the ear. A portion of the inner ear, the basilar membrane, contains thousands of tiny hair cells that must be stimulated for sound to be heard. Helmholtz proposed that the basilar membrane vibrates, or 'resonates', when placed near a sound source. The membrane is narrow at one end and wider at the other, and Helmholtz theorized that high-pitched sounds cause the most resonance where the basilar membrane is narrowest, while low-pitched sounds cause the most resonance at the wide end of the membrane. A good model of Helmholtz's theory is a harp, with short strings at one end and long strings at the other. High-pitched sounds will cause short strings to resonate, while lower-pitched sounds will cause longer strings to resonate. As different parts of the basilar membrane resonate, hair cells are stimulated; impulses are sent to the brain, and the listener hears corresponding

frequencies of sound. The sound heard is related to the intensity of resonance at different places on the basilar membrane.

The greatest competition to Helmholtz's 'place theory' has come from 'frequency theory', which contends that the basilar membrane vibrates as a whole and that analysis of sounds into pitches occurs in the brain, rather than in the ear. Both theories only partially explain the phenomenon of hearing. Helmholtz's theory has been shown to be incorrect in some details, but it set an important foundation for later studies of hearing, such as those by Nobel laureate Georg von Békésy.

HELMHOLTZ'S EPISTEMOLOGY

Helmholtz's scientific achievements carried implications for epistemology, the branch of philosophy concerned with the nature and origins of human knowledge. Helmholtz's interest in philosophy was primed by his father's devotion to Kant and Fichte, although Helmholtz took issue with portions of their writings. (*See* **Epistemology**)

Helmholtz rejected nativism, the view that thought springs from innate sources, in favor of empiricism, the view that thought arises from sensory stimulation. He distinguished between sensation (the response of the sense organs to stimulation) and perception (the meaningful interpretation of sensations). Like Kant, Helmholtz thought that sensations are the raw material of meaningful perception, but Kant believed that fundamental perceptual interpretations, such as the location of events in space and time, and their quantity, existence and causality, were innate. Helmholtz contended that meanings were learned, not innate, with the exception of causality. He showed, for example, how space perception – one of Kant's innate categories – could be learned from the focusing motions of the eyes on near and distant objects. (*See* **Kant, Immanuel**; **Space Perception, Development of**; **Perception: Overview**)

Helmholtz was intrigued by differences between human perception and physical reality. He studied visual illusions, in which simple line drawings create predictable perceptual errors. Helmholtz was convinced that the senses do not transmit anything like a miniature replica of physical reality to the brain. Instead, the senses send 'tokens' or 'signs' of the physical world to the brain. Based on experience, the brain unconsciously constructs a meaningful perception from these signs. He called this process 'unconscious inference'. In his *Treatise on Physiological Optics* Helmholtz wrote, 'The sensations of the senses are tokens for our consciousness, it being left to our intelligence to learn how to comprehend their meaning.' Helmholtz's description of how sensory information unconsciously creates meaningful perceptions is sometimes regarded as the first information processing approach to cognition. (*See* **Unconscious Processes**; **Illusions**; **Information Processing**)

Helmholtz thought that the introspective psychology of his day was of no help in understanding perception because perception was an unconscious process and introspection produces reports of conscious experiences. As Helmholtz wrote, 'Here we have to do with mental activities, of which self observation cannot give us any information at all, whose existence can only be inferred from the physiological investigation of the sensory organs.' (*See* **Introspection**)

Helmholtz had an unwavering disregard for vitalism and the findings of many studies were brought to bear upon it. Helmholtz's support for the principle of conservation of energy (1847) is an early example. This principle, also proposed by Julius Robert Meyer and James Prescott Joule, is that energy may be changed from one form to another, but it is never created or destroyed. For example, electrical energy flowing to a light bulb is transformed into light and heat, but the amount of energy is unchanged. Helmholtz said that this principle also applied to living organisms. He argued that animals convert the chemical energy in nutrients into an equal amount of heat and work. Careful experiments with frog muscles showed that heat was produced when they contracted. If the principle of conservation of energy is true of living organisms, then there is no role for hypothetical energies that have no physical source, such as vital forces.

Shortly after his work on the conservation of energy, Helmholtz provided further evidence against vitalism by measuring the velocity of a nerve impulse (1850). Vitalists thought that the vital forces in nerves acted immeasurably rapidly, perhaps instantaneously. Helmholtz constructed an apparatus that held a frog's leg muscle and its motor nerve. When the nerve was electrically stimulated, the muscle contracted. The stimulus current also started a timing device and the muscle's contraction stopped it. Helmholtz stimulated the nerve at different distances from the muscle and estimated the speed of nervous conduction from the differences in reaction time. His average estimate was 27 meters per second. Modern studies have altered this estimate, but the most important outcome was to show that nerve impulses are not instantaneous: they are

measurable by ordinary instruments, and act on a scale consistent with physical and chemical processes. He carried out similar experiments with human reaction times, concluding again that nerve impulses are ordinary physical events, not the product of ethereal vital forces.

Helmholtz's commitment to the empirical study of a remarkable array of phenomena and his capacity to integrate and synthesize science and philosophy mark him as one of the most important scientists of the nineteenth century and as a seminal figure in the history of modern psychology.

Further Reading

Adler HE (2000) Hermann Ludwig Ferdinand von Helmholtz: physicist as psychologist. In: Kimble GA and Wertheimer M (eds) *Portraits of Pioneers in Psychology*, vol. 4, pp. 15–31. Washington, DC: American Psychological Association/Mahwah, NJ: Lawrence Erlbaum.

Cahan D (ed.) (1993) *Hermann von Helmholtz and the Foundations of Nineteenth-century Science*. Berkeley, CA: University of California Press.

Hatfield GC (1990) *The Natural and the Normative: Theories of Spatial Perception from Kant to Helmholtz*. Cambridge, MA: MIT Press.

Helmholtz H von (1995) *Science and Culture: Popular and Philosophical Essays* (Cahan D, ed.). Chicago, IL: University of Chicago Press.

Koenigsberger L (1965) *Hermann von Helmholtz*, translated by Welby FA. New York, NY: Dover. [Original edition published 1909, Oxford: Clarendon.]

Meyering TC (1989) *Historical Roots of Cognitive Science: The Rise of a Cognitive Theory of Perception from Antiquity to the Nineteenth Century*. Dordrecht, Netherlands: Kluwer.

Pastore N (1971) *Selective History of Theories of Visual Perception: 1650–1950*. New York, NY: Oxford University Press.

Turner RS (1974) Helmholtz, Hermann von. In: Gillespie CC (ed.) *Dictionary of Scientific Biography*, vol. 6, pp. 241–253. New York, NY: Scribner's.

Turner RS (1977) Hermann von Helmholtz and the empiricist vision. *Journal of the History of the Behavioral Sciences* **13**: 48–58.

Turner RS (1982) Helmholtz, sensory psychology and the disciplinary development of German psychology. In: Woodward WR and Ash MG (eds) *The Problematic Science: Psychology in Nineteenth-century Thought*. New York, NY: Praeger.

Turner RS (2000) Hermann von Helmholtz. In: Kazdin AE (ed.) *Encyclopedia of Psychology*, vol. 4, pp. 109–111. Washington, DC: American Psychological Association.

Warren RM and Warren RP (eds and transl.) (1968) *Helmholtz on Perception: Its Physiology and Development*. New York, NY: John Wiley.

Hull, Clark L.

Introductory article

Laurence D Smith, University of Maine, Orono, Maine, USA

CONTENTS

Clark L. Hull (1884–1952) was a prominent American behaviorist and learning theorist, known for his influential book Principles of Behavior*. He designed machine simulations of intelligence and proposed a mechanistic psychology that bridged behaviorism and cognitive psychology.*

INTRODUCTION

A leading experimental psychologist during the second quarter of the twentieth century, Clark C. Hull is best known for his 1943 book *Principles of Behavior*, an influential exposition of learning theory. As a behaviorist with longstanding interests in the higher mental processes, he also developed an early research program for machine simulations of intelligent behavior, producing a mechanistic psychology that bridged behaviorism and cognitive psychology. A portrait of Hull during his years at Yale is reproduced in Figure 1. (*See* **Animal Learning**)

BACKGROUND

Hull was born in rural New York in 1884 but grew up near Sickels, Michigan, where he attended a one-room school and worked on the family farm. In his youth he developed a passion for geometry, admiring its systematic arrangement in a hierarchy of postulates and theorems. While preparing for a career in engineering at nearby Alma College, he contracted poliomyelitis which left him able to walk only with the aid of a steel leg brace of his own design. Partly for reasons of health, and partly in response to a reading of William James's *Principles of Psychology*, Hull soon decided on a career in psychology. Significantly, he was attracted to the field by the opportunities it provided for systematic theorizing as well as for the design of mechanical apparatus.

While completing his undergraduate degree at the University of Michigan, Hull studied philosophy under Roy Wood Sellars and psychology under Walter Pillsbury and J. F. Shepard. To further his interest in the psychology of reasoning, he enrolled in Sellars's course in logic, for which he constructed a logic machine made of rotating metal plates that would generate the implications of various syllogisms. This device would prove to be the first in a series of Hull's mechanical simulations of cognitive and behavioral phenomena.

Pursuing graduate work at the University of Wisconsin, Hull studied under Joseph Jastrow and completed his doctorate in 1918 with a dissertation on concept formation. This research produced Hull's first quantitative learning curves and evinced the methodological rigor that would characterize his later work. Remaining at Wisconsin for the next decade, he taught courses and performed research in the areas of hypnosis and aptitude testing, eventually publishing well-received books on each of these topics. (*See* **Concept Learning**)

Hull came to regard his work on hypnosis and testing as a digression from his aim of discovering general laws of thought and behavior, but one aspect of his research on testing did contribute to the development of his behaviorism. To expedite the calculation of correlations between various tests, he designed and built a machine that computed correlation coefficients from data coded on punched paper tapes. Completed in 1925 with support from the National Research Council, the machine convinced Hull that a purely physical device with the proper arrangement of parts was capable of carrying out operations characteristic of higher-level mental processes. Armed with this insight, Hull underwent a conversion to behaviorism in the next few years, teaching seminars on it and devoting careful study to Ivan Pavlov's newly translated *Conditioned Reflexes*. As an admirer of the British empiricists, Hull revered Thomas Hobbes and David Hume as the philosophical forerunners of behaviorism, and regarded Pavlovian conditioned reflexes as the material analogues of the empiricists'

Figure 1. Clark L. Hull during his years at Yale University (Archives of the History of American Psychology, the University of Akron).

association of ideas. (*See* **Behaviorism, Philosophical**; **Hume, David**; **Pavlov, Ivan Petrovich**)

HULL'S CONTRIBUTION TO BEHAVIORISM

During the 1920s and 1930s, Hull blended behaviorist concepts with proposals for the machine simulation of intelligent behavior to form an ambitious program for a mechanistic psychology. The study of conditioning, he felt, would provide the basis for an experimental science of thought processes, conceived materialistically as mental habits. Accordingly, he enlisted the aid of several associates in designing and constructing machines that simulated various learning processes. The first of these was described in *Science* in 1929 in collaboration with the chemist H. D. Baernstein. Encouraged by the ensuing publicity (the journalist George Gray described the Hull–Baernstein machine as the forerunner of an entire generation of 'thinking machines'), Hull and his colleagues produced a series of such devices during the 1930s. Although the machines varied in internal structure (ranging from electrochemical to purely mechanical), they exhibited an impressive array of learning phenomena, including conditioned associations, extinction, spontaneous recovery, higher-order conditioning, trial-and-error learning, and maze learning. Significantly, Hull saw the need to incorporate a source of behavioral variability in the machines, as well as a system of hierarchical control to coordinate the parts and a means by which the machines could store representations of the environment. Such features, Hull stressed, would imbue the machines with a degree of 'ultra-automaticity' that would supersede the capacities of 'ordinary, rigid-type machines'. (*See* **Machine Learning**; **Simulation Theory**; **Artificial Intelligence, Philosophy of**; **Conditioning**; **Knowledge Representation, Psychology of**)

Although Hull was not the first to design and build machine simulations of intelligent behavior, he gave unusually cogent and explicit statements of the rationale for what he later came to call the 'robot' approach. As early as 1926, Hull's intellectual diaries – his 'idea books' – contained a clear explanation of the simulation method:

> It has struck me many times of late that the human is one of the most extraordinary machines – and yet a machine. And it has struck me more than once that so far as the thinking processes go, a machine could be built which would do every essential thing that the body does… [To] think through the essentials of such a mechanism would probably be the best way of analyzing out the essential requirements of thinking… [An] automaton might be constructed on the analogy of the nervous system which could learn and through experience acquire a considerable degree of intelligence.

Hull regarded the design of machines that could exhibit intelligent behavior as equivalent to the formulation of a theory of that behavior, an insight that would underlie later research on computer simulations in the field of artificial intelligence. Viewing intelligent machines as a vindication of materialist philosophy, he often cited them in his rhetorical attacks on 'subjectivists' such as the vitalist Hans Driesch and the Gestalt psychologist Kurt Koffka, both of whom taught at Wisconsin during Hull's years there. Significantly, Hull's interest in machine simulations led him into interactions with Nicolas Rashevsky and Warren McCulloch, two leading architects of subsequent developments in artificial intelligence, and his 'robot' approach influenced Kenneth Craik, one of the founders of British cybernetics in the 1940s. (*See* **McCulloch, Warren**)

Concurrent with his work on machine design, Hull began a series of articles on behavior theory that appeared in the journal *Psychological Review* and, along with his move to Yale University in 1929, served to raise his visibility as a leading behaviorist. Cast in the form of deductive systems containing postulates and empirically testable theorems, the 'miniature systems' presented in the papers covered various forms of adaptive behavior and introduced the concepts that became standard explanatory devices for Hullians in the decades to come. Among the concepts were fractional anticipatory goal responses, habit-family hierarchies, and pure stimulus acts. The pure stimulus acts were hypothetical implicit responses that ran in parallel to event sequences in the environment and could thus function as internal 'replicas' of the world, providing an organic basis of symbolic knowledge. Hull used such concepts in his attempts to account for the emergent phenomena of knowledge, purpose, and insight. They became key constructs for a later generation of mediational behaviorists, whose theorizing proved to be a bridge between the behaviorist era and the eventual cognitive revolution of the 1960s. (*See* **Epistemology**; **Animal Cognition**)

As Hull repeatedly reminded his readers, the miniature systems really amounted to exercises in machine design. His 1936 presidential address to the American Psychological Association presented a theory of trial-and-error learning and concluded with a demonstration of one of his conditioning machines. However, this would be his last public demonstration of such a machine, and Hull's mechanistic emphasis quickly receded from public view. The reasons for this shift are several. One important factor was that the logical positivist philosophy which swept through psychology in the 1930s eschewed analogical explanations in favor of logical explanations. As a result, the apparent rigor of Hull's deductive theoretical systems appealed to his psychological contemporaries, who, in contrast, failed to grasp the significance of simulational methods and remained largely unimpressed by his machine simulations.

Hull's theoretical papers of the 1930s laid the foundation for his *Principles of Behavior*, which appeared in 1943 and became the most-cited work in experimental psychology during the following years. Increasingly aware of the rhetorical power of logical methods, Hull cast the book in the form of an elaborate hypothetico-deductive system containing 16 postulates and more than a hundred corollaries and theorems. Reaction potential, the basic measure of learned behavior, was defined in terms of such now-familiar concepts as habit strength, drives, and incentive; reinforcement of stimulus–response connections was said to take place through drive reduction. The system served as a focal point of extended debates among learning theorists, pitting Hull against such adversaries as Edward C. Tolman and Kurt Lewin. Although the book contained a brief discussion of the robot approach, the debates over learning were by this time being framed in terms of such logical positivist concepts as postulate systems, correspondence principles and intervening variables. (*See* **Learning, Psychology of**; **Tolman, Edward C.**; **Lewin, Kurt**)

During the 1930s and 1940s, Hull's preeminent standing in American psychology was enhanced by his position of leadership in Yale's Institute of Human Relations, which was well funded by the Rockefeller Foundation during the Depression years. Through the Institute, Hull attracted a large number of young social scientists to pursue the application of behavioral principles to a range of topics in social learning, aggression, psychopathology, and personality theory. Various books produced by this group – notably the 1939 classic *Frustration and Aggression* – inspired much research among later generations of Hullians. The talented cast of collaborators who came under Hull's influence at Yale included Daniel Berlyne, John Dollard, Eleanor Gibson, Carl Hovland, Neal Miller, O. H. Mowrer, Charles Osgood, Robert Sears, and Kenneth Spence. Among these, Spence remained particularly close to Hull, refining the Hullian system and transmitting it to scores of doctoral students during his productive career at the University of Iowa.

Hull had long planned to supplement the *Principles* with sequel volumes on individual cognition and social behavior, but plans for the trilogy were never realized. When the learning theory debates of the 1940s revealed the complexity of even the simplest forms of adaptive behavior, he became immersed in the details of his system, repeatedly revising the 1943 postulates in light of anomalous findings. His final two books, completed in the years just before his death in 1952 of heart disease, never achieved the scope or influence of the *Principles*.

During its heyday, the prominence of Hull's system made it a target for widespread criticism. Some critics, notably Sigmund Koch, attacked the system on methodological grounds for its problematic handling of intervening variables and its failure to provide genuine deductions of testable implications. Others, including B. F. Skinner, criticized it for its empirical shortcomings and for its

premature attempt at logical rigor and quantitative precision. As the Hullian system fell from favor during the 1960s, its leading status among behaviorists was usurped by Skinner's relatively atheoretical tradition of operant psychology. Behaviorists of a Hullian stripe turned to circumscribed studies of topics in conditioning research or drifted towards a more cognitively oriented mediational behaviorism. At the same time, Hull's earlier program for machine simulations of intelligence was superseded by the postwar emergence of cybernetics, computer-based studies of artificial intelligence, and the information-processing perspective of cognitive psychology. (*See* **Skinner, Burrhus Frederic**)

CONCLUSION

As a transitional figure bridging behaviorism and cognitive psychology, Hull has received mixed treatment from historians. Considered as an example of behaviorist theorizing, his mechanistic psychology has been portrayed as the grand failure of behaviorism that allowed cognitive approaches to gain ascendancy during the cognitive revolution of the 1960s. Considered as a program for machine simulations of cognitive processes, his mechanistic psychology has also been viewed as a significant precursor to the work on artificial intelligence which had a central role in furthering that revolution. Taken as a whole, however, Hull's work stands as a notable, if sometimes flawed, effort to bring the study of mental life under the canopy of natural science and materialist philosophy. (*See* **Materialism**)

Further Reading

Amsel A and Rashotte ME (1984) *Mechanisms of Adaptive Behavior: Clark L. Hull's Theoretical Papers, with Commentary*. New York, NY: Columbia University Press.

Cordeschi R (2002) *The Discovery of the Artificial*. Dordrecht, Netherlands: Kluwer.

Gray GW (1936) Thinking machines. *Harper's Monthly Magazine* **172**: (March) 416–425.

Hilgard ER and Bower GH (1966) *Theories of Learning*, 3rd edn. New York, NY: Appleton-Century-Crofts.

Hull CL (1943) *Principles of Behavior: An Introduction to Behavior Theory*. New York, NY: D. Appleton-Century.

Hull CL (1952) Clark L. Hull. In: Boring EG, Werner H, Langfeld HS and Yerkes RM (eds) *A History of Psychology in Autobiography*, vol. 4, pp. 143–162. Worcester, MA: Clark University Press.

Hull CL (1962) Psychology of the scientist: IV. Passages from the 'Idea Books' of Clark L. Hull. *Perceptual and Motor Skills* **15**: 807–882.

Leahey TH (1992) *A History of Psychology: Main Currents in Psychological Thought*, 3rd edn. Englewood Cliffs, NJ: Prentice Hall.

Morawski JG (1986) Organizing knowledge and behavior at Yale's Institute of Human Relations. *Isis* **77**: 219–242.

Smith LD (1986) *Behaviorism and Logical Positivism: A Reassessment of the Alliance*. Stanford, CA: Stanford University Press.

Hume, David

Introductory article

John Biro, University of Florida, Gainesville, Florida, USA

CONTENTS

In his 'science of man', Hume set out to construct a complete account of the workings of the human mind, in many ways anticipating the aims, questions, and even some of the answers, of modern cognitive science.

INTRODUCTION

David Hume (1711–1776) was the last of the three great British empiricist philosophers of the late seventeenth and early eighteenth centuries, a Scot following an Englishman (John Locke, 1637–1704) and an Irishman (George Berkeley, 1685–1753). He developed and pushed to its limit the thesis advanced by his two predecessors that sense experience is the only source of the contents of the mind and that all our beliefs and whatever knowledge we may possess must thus be traceable to it. In doing so, he deployed arguments that seem to lead to skepticism concerning many things we normally take ourselves to know, though it can be also argued that he intended merely to show that our commonsense knowledge cannot, and need not, be based on the metaphysical claims of philosophers, especially not of rationalist ones.

In his best-known philosophical work, *A Treatise of Human Nature*, he announces himself as engaged in developing a 'science of man' on the model of Newton's science of matter, in which we would, through careful observation of our own and others' cognitive behavior, discover the fundamental properties and activities of the human mind: a prerequisite of understanding anything else. In its aims, in the questions it addresses, and even in some of its results, Hume's science looks forward to modern cognitive science.

HUME'S LIFE

David Hume was born in Edinburgh in 1711. His family had a small estate in the Borders, and Hume spent most of his early life there. He received little formal education, following his passion for reading and learning on his own, clear in his ambition to be a 'scholar and philosopher'. After some half-hearted attempts at a more practical career in the law and in commerce, Hume spent three years living modestly in France while writing the *Treatise*, a work he had been planning since his late teens. It was published when he was 28 years old, and met with little success. A more cautious and more accessible recasting of its doctrines in the two *Enquiries* (*An Enquiry Concerning Human Understanding* and *An Enquiry Concerning the Principles of Morals*) fared better; but it was only later, with the appearance of his *Essays, Moral and Political*, that Hume began to achieve that literary fame which he described in his brief and remarkably candid *My Own Life* as 'my ruling passion'. These essays earned him European fame and the correspondence of some of the leading intellects on the Continent. They also earned him financial independence. This was buttressed by his service from time to time in various military and diplomatic missions, both before and after the publication of his *History of Great Britain*, which, while controversial, cemented his reputation as a writer. He was able to retire to Edinburgh in comfort for the last half dozen years of his life, dying peacefully in 1776.

CENTRAL PHILOSOPHICAL VIEWS

In his own time and until recently, Hume's philosophy was seen as essentially, sometimes exclusively, skeptical. It was thought to consist in a string of arguments designed to show the impossibility of knowledge – indeed, even of reasonable belief – about many of the things that our commonsense view of the world takes for granted. According to commonsense, there is a world of bodies outside, and independent of, our minds – bodies that stand

in causal relations to one another, which we can discover through reason or experience. Thus we can predict their behavior, at least to some extent, giving us reason to act in one way rather than another. How we act matters, both practically and morally. We are cognizers, enduring subjects to whom knowledge or belief can be attributed, and agents who can be judged for their actions. Hume's famous discussions of external existence, of necessary connection, of personal identity, of the passions, and of morals, seemed to deny that our ordinary beliefs about any of these things could be supported by either reason or experience. In fact, no belief has any more warrant than any other, and we adopt the ones we do from 'habit' – from unexplainable and irresistible natural inclination.

Many readers of Hume today reject this picture of Hume's philosophy as at least misleading and oversimplified, if not distorting. They are more inclined to take Hume at his word when he declares in the introduction to the *Treatise* that he is pursuing a scientific goal, that of creating a new 'science of man' by 'introducing the experimental method into moral subjects'. As part of this project, he finds it necessary to attack the pretensions of philosophy, including its claim to be able to provide rational justifications of our commonsense beliefs. The skeptical arguments are Hume's way of reducing these pretended philosophical justifications to absurdity, thereby clearing the ground for a descriptive science of the workings of our cognitive and affective capacities. Hume the skeptic thus gives way to Hume the naturalist.

Hume's 'skeptical' arguments may also be seen in a different light in the context of his lifelong campaign against 'enthusiasm', the vehement and dogmatic adherence to religious doctrines, many of them underpinned by metaphysical speculation. In showing the absurdity of the latter, Hume hoped to draw the sting of religious and political extremism, whose dangers he saw both in history and in the politics of his own day. Far from being a threat to common sense, then, these skeptical arguments were deployed by Hume on its behalf.

VIEWS ON MIND, COGNITION, AND LANGUAGE

Mind

The philosophical and religious notion that there is a single and simple entity that is one's mind (soul), which endures unchanged amid the flux of one's thoughts and is the thinker of those thoughts, is, Hume maintains, a fiction, impossible to defend on either rational or empirical grounds. Instead, the mind be seen as the sum of its contents: just as a commonwealth is nothing over and above its constituent parts, so the mind is nothing but 'a bundle of perceptions'. Its identity, like that of a commonwealth, resides in its structure and its functional organization. It is governed by natural propensities that it cannot resist, except in infrequent and brief intervals of philosophical reflection, and then only in theory, not in practice.

Cognition

Hume maintains that all our concepts ('ideas') are derived, directly or indirectly, from sensation ('impressions'), outer or inner. We form our beliefs as a function of the strength of our impressions: as an idea approaches an impression in 'vivacity', 'liveliness', or 'force', it becomes a belief. Thinking consists in a transition from one idea to another. This transition, contrary to philosophical myth, is a matter of natural propensities, not of reflective, rational calculation. It is governed by three 'principles of association': resemblence, contiguity, and cause and effect.

In thinking of something, we naturally think of other things that it resembles (as a picture resembles its subject), adjoins (as my garden adjoins my house), or causes or is caused by (as drinking is caused by thirst or as heat causes ice to melt). Such transitions are the only three kinds of inference that can lead to belief, the last being by far the most important, as it is the only one by which 'the mind can go beyond what is immediately present to the senses … to discover the real existence or the relations of objects'.

However, causal inference is based only on the expectations generated by our experience of the 'constant conjunction' of two events, experience that contains no grounds for positing a necessary connection between the two of the sort that philosophers postulate. Thus no demonstration of a causal relation is possible, and our confidence in our causal beliefs and the predictions we must base on them to survive is not justifiable by rational argument: 'belief is more properly an act of the sensitive, rather than of the cogitative part of our natures'. So is our belief in the existence of bodies, which 'we must take for granted in all our reasonings', even if we cannot justify it by argument.

Language

Hume has little to say about language as a distinct cognitive faculty. He has a theory of meaning, according to which the meaningfulness of a word depends on its standing for an idea with appropriate empiricist credentials. (What counts as appropriate is a matter of considerable controversy.) His chief concern is to press home Berkeley's earlier criticism of Locke's theory of abstract ideas, that the meaningfulness of a general term, such as 'triangle', does not depend on there being a general idea of a triangle, one that is not of a particular triangle, of some specific sort. Given that for Hume, as for most philosophers of the early modern period, thinking and reasoning involve the linking together of ideas, not of words, this relative neglect of language is not surprising.

RESPONSES TO HUME

Hume's influence has been immense. Beginning with Thomas Reid's defense of common sense against the threat of Humean skepticism and Immanuel Kant's heroic attempt to confront Hume's challenge regarding the possibility of genuine causal knowledge, every philosopher since his time has realized that his skeptical arguments have to be faced one way or another: answered, circumvented, or reinterpreted. They have continued to be recognized as among the most penetrating investigations into the fundamental questions about the mind.

There is, however, little agreement about the outcome of those investigations. Hume is the most crystalline of writers, one of the greatest stylists in the English language. Yet the upshot of his discussions is often unclear. (The early twentieth-century philosopher and mathematician, Alfred North Whitehead, aptly remarked on Hume's 'local clarity and global obscurity'.) Is he a skeptic or a naturalist? Can one be both? Is he serious about a Newtonian science of the mind? Or does he really believe that reason 'is, and ought to be, a slave of the passions' and that all beliefs are equally unjustified? Which of his claims and arguments are to be taken at face value? Can they be somehow reinterpreted to make a coherent whole?

RELEVANCE OF HUME'S WORK TO COGNITIVE SCIENCE

Hume's announcement of a new science of the mind seems to be an anticipation of modern cognitive science, in its scope, ambitions, and methods. 'There is no question of importance, whose decision is not compriz'd in the science of man.' Rather than philosophizing about what the mind must be like *a priori*, the new science will be based on 'experience and observation'.

Even more interestingly, Hume's detailed descriptions of the mental operations and processes that generate our concepts and beliefs and explain the transitions among the latter sound remarkably modern. He takes the objects on which these operations and processes work to be representations ('ideas') and has a theory about their semantics. He insists that many of the operations and processes in question are non-reflective, non-conscious, and non-optional – in modern parlance, sub-doxastic and modular ('cognitively impenetrable', and 'informationally encapsulated'). He is acutely aware of the difficulties involved in saying anything sensible about the ownership of the representations and the agent of the cognitive operations and processes in which they figure – the modern 'problem of the homunculus'. In spite of his denial of the possibility of a philosophical justification of causal or of inductive inferences, he gives us both a descriptive and a normative theory of them, an insightful description of how we make them, and a set of rules for making them well. On all these matters, and on many others, he is engaging with the deepest and most difficult questions about the mind, and he gives answers that still command our interest and respect.

Further Reading

Biro J (1985) Hume and cognitive science. *History of Philosophy Quarterly* **2**(3).

Biro J (1993) Hume's new science of the mind. In: North DF (ed.) *The Cambridge Companion to Hume*. Cambridge, UK: Cambridge University Press.

Bricke J (1980) *Hume's Philosophy of Mind*. Princeton, NJ: Princeton University Press.

Easton P (ed.) (1997) *Logic and the Workings of the Mind: The Logic of Ideas and Faculty Psychology in Early Modern Philosophy*. Atascadero, CA: Ridgeview.

Flage D (1990) *David Hume's Theory of Mind*. London, UK: Routledge & Kegan Paul.

Garrett D (1997) *Cognition and Commitment in Hume's Philosophy*. New York, NY: Oxford University Press.

Hume D, Selby-Bigge LA and Niddich PH (eds) (1987) *A Treatise of Human Nature*. Oxford, UK: Clarendon Press.

Mossner EC (1954) *The Life of David Hume*. Austin, TX: University of Texas Press.

Noonan H (1998) *Hume on Knowledge*. London, UK: Routledge.

Owen D (1999) *Hume's Reason*. Oxford, UK: Oxford University Press.

Smith, J-C (1990) *Historical Foundations of Cognitive Science*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Stroud B (1977) *Hume*. London, UK: Routledge & Kegan Paul.

Winkler K (1991) The new Hume. *Philosophical Review* **100**: 541–579.

Jakobson, Roman

Introductory article

Ilse Lehiste, The Ohio State University, Columbus, Ohio, USA

CONTENTS

Roman Jakobson was one of the most creative and influential linguists of the twentieth century particularly in terms of his contributions to linguistic theory.

BIOGRAPHY

Roman Jakobson was born on 11 October 1896, in Moscow, and died on 18 July 1982, in Boston. He studied at the Lazarev Institute of Oriental Languages and the University of Moscow (1918). During his student years he was active in the Moscow Linguistic Circle (founded in 1915). After teaching briefly at the Higher School of Drama, he moved to Czechoslovakia (1920), where he continued his studies and obtained his doctoral degree (Charles University, 1930). While in Prague, he was a member of the Prague Linguistic Circle (founded in 1926), collaborating for about 10 years with its leading member N. Trubetzkoy (whom he had already met in 1914). In 1933 he was appointed as *docent* at the Masaryk University in Brno, and held the chair in Russian philology and Old Czech literature there from 1937 until he was forced to leave by the German invasion in the spring of 1939. He went first to Scandinavia (1939–1941), and then to the United States. From 1942 to 1946 he taught at the École Libre des Hautes Études in New York; in 1943 he started teaching at Columbia University, and in 1949 he received a professorial chair at Harvard University in Cambridge, Massachusetts (1949–1966). After 1957 he was also Institute Professor at the Massachusetts Institute of Technology (1957–1982).

In an essay published in the first volume of his *Selected Writings*, he describes the intellectual atmosphere at Moscow University during his years there as student. Already during his freshman year, his interests were broader than those of his teachers – representatives of what he characterizes as the orthodox Moscow linguistic school. The first stimulus for the study of modern phonology came from the 1912 monograph on Russian vowels by L.V. Ščerba, which introduced him to the concept of the phoneme. Somewhat later he learned about the work of Ferdinand de Saussure. During these years, students of psychology and linguistics at Moscow University were passionately discussing the philosophical concepts of the *signatum* (that what is signified) and the *denotatum* (that what is referred to), and learned to assign linguistic contents to the terms, first to the *signatum*, and then to its counterpart – the *signans* (the signifier – that which is performing the signifying).

In that autobiographical essay, Jakobson emphasizes the influence of the turbulent artistic movements of the early twentieth century on the changes in thinking about linguistics, singling out in particular the pictorial theory and practice of cubism, in which everything was based on relationships. In a later presentation he put it more categorically: 'There is an inventory of simple relations common to all tongues of the world.'

Jakobson was also inspired by the Russian poet Velimir Xlebnikov, who played with minimal pairs in his poetry in verses such as '/v,idil vid,il v'os,in vos,in,/' (*videl vydel vesen v osen*), where the initial consonants differ only by presence and absence of palatalization. Xlebnikov's search for the 'infinitesimals of the poetic word' prompted an intuitive grasp of what were to be called some decades later the ultimate phonemic units – *distinctive features*.

JAKOBSON'S LINGUISTIC CAREER: CONTRIBUTIONS TO PHONOLOGICAL THEORY

Jakobson started his linguistic career as a philologist, studying language history, including the historical study of literary texts. His doctoral dissertation dealt with the metrics of the South Slavic epic. While his literary interests continued into his later years, he gravitated more and more to the study of language, first of all Russian. But the study of specific languages was always illuminated

by insights derived from his developing linguistic theories, and he continued to search for linguistic universals in the structure of specific languages with which he was working.

The two concepts with which Jakobson will always be associated are distinctive features and binarism. Distinctive features – contrastive characteristics of sounds – are a concept that was under lively discussion in the Prague Linguistic Circle, of which Jakobson was an active member during his stay in Czechoslovakia. Phonemes – speech sounds that are associated with differences in meaning – were redefined as bundles of distinctive features. The first public presentation of the theory took place at the First International Congress of Linguists at The Hague in 1928. The paper was written by Jakobson in October 1927 in response to a question presented by the organizing committee of the congress; it was later approved and countersigned by S. Karcevski and N. Trubetzkoy, distributed to members of the congress in 1928, and published (in 1930) in the *Actes du 1er Congrès International de Linguistes du 10–15 avril, 1928*. The theory was elaborated and defended in a series of publications by Jakobson and Trubetzkoy, and it has been expanded, refined – and criticized – by many linguists ever since.

Jakobson's system differs from that of Trubetzkoy in several respects, of which the most significant for future linguistic theorization was his introduction of the principle of binarism. This means that every opposition consisted of two possible states or poles: presence or absence of a sound characteristic, or choice between two (and only two) possible characteristics. Jakobson presented the first arguments in this respect in 1938, in a communication to the Third International Congress of Phonetic Sciences in Ghent.

In the distinctive features theory, the universal set of speech sounds is organized according to a set of characteristics specifying place of articulation and/or manner of articulation. These characteristics – labeled features – were first described in articulatory terms, such as dental–labial, or voiced–voiceless. In a significant collaboration with the physicist Gunnar Fant, the features were assigned acoustic characteristics. In later studies, distinctive features have been related to human speech production and perception mechanisms. Distinctive features have become the basis of much of generative phonology (e.g. Chomsky and Halle 1968, and many later publications). Likewise, the notion that linguistic oppositions are binary has become almost an axiom in much of linguistic theory.

CONTRIBUTIONS TO OTHER SUBFIELDS OF LINGUISTICS

Jakobson's significance in the development of linguistic theory is by no means limited to his contributions to phonological theory. He was one of the first to look at language acquisition and loss from the point of view of what these two processes can reveal about the structure of language. In studying linguistic universals, he came to suggest that what he called the inventory of simple relations common to all tongues of the world pertained both to the early acquisition of children's language and to the most stable verbal properties in those types of aphasic regress which display a mirror picture of infants' development.

His continuing interest in the structure of Russian included the phonetics and phonology of Russian, historical morphology of Russian, dialectology of Russian, and even the Russian spelling system; of particular significance for grammatical theory are his studies of Russian morphology.

His interests in language typology led to observations of areal convergences – the development of *Sprachbünde* – areas where language contact has led to the emergence of typological similarities between unrelated languages. All through his life he maintained an interest in metrics and continued to make major contributions in this area.

As a Slavic philologist, he continued his investigations of Slavic epic poetry. He also published extensively in folklore and comparative mythology, especially the mythology of Slavic peoples. To quote Robert Austerlitz:

> Jakobson's contributions to linguistics should be viewed in the context of his interest and work in folkloristic, ritual, and literary texts: here he pursues both overt and subliminal messages with all the tools in his arsenal. This is where his most striking and innovative contributions to the study of the human instrument for communication lie. The inscription on his tombstone – *russkij filolog* – is therefore justified.

Jakobson was also a charismatic teacher. An entire generation of American Slavists was inspired and trained by him, as well as numerous general linguists. A Supplement to Volume XXVII of the *International Journal of Slavic Linguistics and Poetics* (1983), edited by Morris Halle, contains articles by nine scholars, offering appraisals on Jakobson's contributions to the various branches of linguistic science, and constitutes a fitting memorial to the scholar and teacher. But his influence reached across disciplines to such scholars as the anthropologist Claude Levi-Strauss, who was introduced to structural linguistics through Jakobson's work.

Jakobson discovered new and illuminating aspects in whatever linguistic problem he approached, and his creativity was matched by his productivity: the complete bibliography of his writings comprises 98 pages. Jakobson was one of the true intellectual giants of the twentieth century.

Further Reading

Austerlitz R (1996) Jakobson, Roman (Osipovič). In: Stammerjohann H (ed.) *Lexicon Grammaticorum*, pp. 471–474. Max Niemeyer Verlag: Tübingen.

Chomsky N and Halle M (1968) *The Sound Pattern of English*. New York: Harper & Row.

Halle M (ed.) (1983) Roman Jakobson: what he taught us. *International Journal of Slavic Linguistics and Poetics* **XXVII**: Supplement.

Jakobson R (1962) *Selected Writings* Vol. I: *Phonological Studies*, 2nd, expanded edn, 1971. The Hague: Mouton.

Jakobson R (1971) *Selected Writings* Vol. II: *Word and Language*, pp. XII, 1–752. The Hague: Mouton.

Jakobson R (1981) *Selected Writings* Vol. III: *Poetry of Grammar and Grammar of Poetry*. The Hague: Mouton.

Jakobson R (1966) *Selected Writings* Vol. IV: *Slavic Epic Studies*. The Hague: Mouton.

Jakobson R (1979) *Selected Writings* Vol. V: *On Verse, Its Masters and Explorers*. The Hague: Mouton.

Jakobson R (1985) *Selected Writings* Vol. VI: *Early Slavic Paths and Crossroads*. Part One: *Comparative Slavic Studies*. *The Cyrillo-Methodian Tradition*. Berlin: Mouton.

Jakobson R (1985) *Selected Writings* Vol. VI: *Early Slavic Paths and Crossroads*. Part Two, pp. VIII, 401–942. Berlin: Mouton.

Jakobson R (1985) *Selected Writings* Vol. VII: *Contributions to Comparative Mythology. Studies in Linguistics and Philology, 1972–1982*. Berlin: Mouton de Gruyter.

Jakobson R (1988) *Selected Writings* Vol. VIII: *Major Works 1976–1980*. Berlin: Mouton de Gruyter.

Jakobson R, Gunnar C, Fant M and Halle M (1952) *Preliminaries to Speech Analysis*. (Acoustics Laboratory, Technical Report 13.) Cambridge, MA: MIT Press. [Revised edn (1963) In: *Selected Works*, Vol. VIII, pp. 585–660.]

Rudy S (1990) *Roman Jakobson, 1896–1982: A Complete Bibliography of his Writings*. Berlin: Mouton.

Kant, Immanuel

Introductory article

Andrew Brook, Carleton University, Ottawa, Canada

CONTENTS

Immanuel Kant (1724–1804) has had an enormous influence on cognitive research: the dominant model of the mind in contemporary cognitive science is thoroughly Kantian. Nonetheless, some of his most distinctive ideas have played little role in it.

INTRODUCTION

Immanuel Kant may be the single most influential figure in the history of cognitive research before the twentieth century, indeed the intellectual grandfather of cognitive science. Kant held that cognition requires application of concepts as well as sensory input, and that synthesis and mental unity are central to cognition. He advanced a functionalist model of the mind almost 200 years before functionalism was articulated, and he had some highly original things to say about self-consciousness. However, while much of contemporary cognitive science is Kantian, there remain many ideas from Kant's model of the mind which cognitive science has not assimilated.

KANT'S LIFE

Kant was the last great thinker of the German Enlightenment. He focused on the human individual (rather than, say, society). Though said to be one-quarter Scottish (some think that 'Kant' is a Germanization of 'Candt'), Kant lived his whole life in Königsberg (now Kaliningrad) just south of Lithuania. His father was a saddle maker. He was devoutly religious but hostile to many conventional religious observances. By the time of his death, he had served some terms as Rector of the University of Königsberg and effectively the official philosopher of the German-speaking world.

Kant's most famous work is the *Critique of Pure Reason* of 1781 and 1787 (two editions). He was already 57 when he wrote the first edition, yet he went on to write the *Critique of Practical Reason* (1788) on moral reasoning, the *Critique of Judgement* (1790) – a work devoted to a number of topics including reasoning about ends, the nature of judgment, and aesthetics – and books on natural science, cosmology, history, geography, logic and anthropology. From the point of view of cognitive science, his two most important works are the *Critique of Pure Reason* and a small book composed from lecture notes late in his life, *Anthropology From a Pragmatic Point of View* (1798).

MAIN PHILOSOPHICAL VIEWS

Kant started out as a conventional rationalist. Then, memories of reading David Hume 'interrupted my dogmatic slumbers', as he put it. He called the new approach 'critical philosophy'. Critical philosophy asks the question: what must we be like to experience as we do? Kant's answer provided the framework for most subsequent cognitive research.

Philosophy of mind and knowledge were by no means the only areas in which Kant made seminal contributions. He founded physical geometry. His work on social ethics is the basis for modern liberal democratic theory. His deontological approach to the justification of ethical beliefs put ethics on a new footing, and remains influential. He taught metaphysics, ethics, physical geometry, logic, mechanics, theoretical physics, algebra, calculus, trigonometry, and history.

Kant aimed to do two principal things in the *Critique of Pure Reason*: to justify our conviction that physics, like mathematics, is a body of necessary and universal truth; and to insulate religion, including belief in immortality, and free will from the corrosive effects of this same science. Kant had not the slightest doubt that 'God, freedom and immortality' exist, but feared that if science is relevant to their existence it will show them not to exist.

Fortunately, as he saw it, science is quite irrelevant to the question of their existence.

MODEL OF THE MIND

It was the pursuit of the aim of putting physics on a secure footing that led Kant to his views about how the mind works. He approached the foundations of physics by asking: what are the necessary conditions of experience? Put simply, he held that for our experience, and therefore our minds, to be as they are, our experience must be tied together in the way physics says it is. But this also tells us a lot about what our minds must be like. As we will see, his attempt to insulate religion from science led him to some very original views about our awareness of ourselves.

Interestingly, Kant held that psychology (by which he meant the introspective study of the mind) could never be a science. Once, after saying that chemistry would never be a science, he went on, 'the empirical doctrine of the soul … must remain even further removed than chemistry from the rank of what may be called a natural science proper'. Kant thought we should study the mind by thinking through what the mind must be like and what capacities it must have to represent things as it does. This is his famous 'transcendental method'; as we will see, despite its nonempirical roots, it has become an essential method of cognitive science.

Kant made a number of substantive claims about the mind. The most famous is his claim that representation requires concepts as well as percepts; that is, rule-guided acts of cognition as well as signals of the senses. As he put it, 'concepts without intuitions are empty, intuitions without concepts are blind'. In more contemporary terms, the claim is that to discriminate anything from anything else, we need information; but for information to be of any use to us, we must also be able to organize information.

Kant held that to organize information requires two kinds of synthesis. The first ties the raw material of sensible experience together into objects. In terms of contemporary binding theory, colors, lines, shapes, textures, etc., are represented in widely dispersed areas of the brain. These dispersed representations have to be brought into relation to one another and integrated into a representation of a single object.

The second kind of synthesis ties these individual representations together into what might be called 'global representations', in such a way that to be aware of any of the representations thus tied together is to be aware of some of the others, too, and of the group of them as a single group. Kant thought that the capacity to form global representations is essential to both the kind of cognition that we have and the kind of consciousness that we have.

Though by no means all global representations are conscious (Kant is widely misunderstood on this point), the unity found in them is also a feature of consciousness, one that greatly interested Kant. In his view, unified global representations are the result of unifying acts of synthesis, and it takes a unified consciousness to perform such acts.

In addition to what he said about consciousness in general, Kant made some original claims about consciousness of self. These claims arose in the course of his attempt to insulate immortality (and God and free will) from the attacks of science. His rationalist predecessors thought that they could prove that the mind is substantial, and simple (without parts), and that it persists in a special way. This opened the door to a proof of immortality. Descartes, Leibniz, and Reid all took this approach. However, if arguments and evidence are relevant to determining the existence of immortality at all, then argument and evidence could also show it not to exist. For Kant, the best hope was to insulate such matters from argument and evidence entirely. That way, conclusions could be accepted on the basis of faith (and Kant did so accept them) without being at risk from science. Kant thought that introspection provides strong prima facie counterevidence to his anti-intellectual conclusions about what we can know about the nature of the mind. In introspection, one does appear to be substantial, simple and persisting, just as rational psychology says ('rational psychology' was Kant's name for these views). It was incumbent upon Kant to show that introspection gives us nothing of the sort.

In the course of his attack on introspection, Kant made a number of claims. He distinguished two quite different kinds of self-awareness, awareness of one's states and awareness of oneself as the subject of these states. He claimed that the cognitive and semantic machinery used to obtain awareness of self as subject is unusual. In it, we 'denote' but do not 'represent' ourselves. In other words, we designate ourselves without noting 'any quality whatsoever' in ourselves. He argued that the representational basis of awareness of self as subject is not a special experience of self but any experience of anything whatsoever. When one is aware of oneself as subject, he claimed, one is aware of oneself in a way that is not awareness of features of

oneself, a way in which 'nothing manifold is given'. Finally, he asserted that when we are aware of ourselves as subject, we are aware of ourselves as the 'single common subject' of a number of representations.

Kant's conception of the mind is an early example of functionalism. To model the mind, we must model what it does and can do, that is to say, its functions.

KANT'S INFLUENCE

The influence of Kant was enormous even in his own time. When he died, he was the dominant philosopher throughout the German-speaking world. His ideas had a major influence on empirical students of the mind in the nineteenth century. His influence waned during the heyday of behaviorism, but increased again with the revolution in cognitive science of the 1960s and 1970s, though this time the influence was indirect and not widely acknowledged.

KANT AND COGNITIVE SCIENCE

Kant influenced cognitive science via nineteenth-century cognitive researchers, Herbart and Helmholtz in particular. These figures are rightly regarded as the precursors of contemporary cognitive research, but it is seldom realized that they both regarded themselves as Kantians. Some of Kant's doctrines are built into the very foundations of cognitive science.

The first of these is the transcendental method. Transcendental arguments attempt to infer the conditions necessary for some phenomenon to occur. Translated into contemporary terms, this has become the method of postulating unobservable mental mechanisms in order to arrive at the best explanation of observed behavior. This approach completely supplanted introspection and rational inference from concepts, the two dominant methods of studying the mind before Kant, and is now the most important method of cognitive science.

Secondly, the doctrine that most representation requires concepts as well as percepts has become as orthodox in cognitive science as it was central to Kant.

Thirdly, the functionalist conception of the mind, and the claim that we can model cognitive function without knowing very much about the underlying structure, derive from Kant. Kant even shared functionalism's lack of enthusiasm for introspection. Indeed, he went further than contemporary functionalists in one respect: he thought that we could know nothing about the underlying structure, an application of his general doctrine that we can know nothing about anything as it is in itself.

Interestingly, other ideas equally central to Kant have been largely ignored in cognitive science. Recall Kant's claim that cognition requires two kinds of synthesis. The first kind of synthesis is widely studied in the form of binding. Indeed, one model, Anne Treisman's three-stage model, is very similar to Kant's. According to Treisman, object recognition proceeds in three stages: feature detection, location of features on a map of locations, and integration and identification of objects under concepts. This is similar to Kant's three-stage model of apprehension of features, association of features in something like clusters (Kant called this stage 'reproduction'), and recognition of these 'clusters' as objects falling under concepts. However, Kant's second kind of synthesis, the activity of tying multiple representations together into a global representation, has received little attention.

The same was true until recently of Kant's doctrine of the unity of consciousness. In Kant's view, unity in representation requires unity in the thing doing the representing. Indeed, the whole topic of mental unity has been neglected in cognitive science until recently.

Finally, Kant's views on consciousness of self have played little role in cognitive science. Kant did not consider consciousness of self to be essential to all forms of unified cognition, but he made a number of penetrating discoveries about it. Some closely related ideas have now reappeared in the philosophy of language.

Thus, while, the dominant model of the mind in contemporary cognitive science is Kantian, some of his most distinctive ideas have not yet been taken up. This article has discussed only a few of Kant's ideas about cognition. He also had a complex model of representation in space and time, and strong views on what we can and cannot know. Many philosophers now have serious doubts about these, and they have not played a significant role in cognitive science.

Further Reading

Ameriks K (1983) *Kant's Theory of Mind*. Oxford, UK: Oxford University Press.

Brook A (1994) *Kant and the Mind*. Cambridge, UK and New York, NY: Cambridge University Press.

Kant I (1970) *The Metaphysical Foundations of Natural Science*, translated by J Ellington. Indianapolis, IN: Library of Liberal Arts. [First published 1786.]

Kant I (1974) *Anthropology From a Pragmatic Point of View*, translated by M Gregor. The Hague: Martinus Nijhoff. [First published 1798.]

Kant I (1977) *Prolegomena to Any Future Metaphysics*, translated by P Carus. Indianapolis, IN: Hackett. [Revised with an introduction by James Ellington. First published 1783.]

Kant I (1997) *Critique of Pure Reason*, translated by P Guyer and A Woods. Cambridge, UK and New York, NY: Cambridge University Press. [First published 1781/1787.]

Kitcher P (1990) *Kant's Transcendental Psychology*. New York, NY: Oxford University Press.

Meerbote R (1989) Kant's functionalism. In: Smith JC (ed.) *Historical Foundations of Cognitive Science*. Dordrecht, Netherlands: Reidel.

Sellars W (1970) '… this I or he or it (the thing) which thinks …'. *Proceedings of the American Philosophical Association* **44**: 5–31.

Strawson PF (1966) *The Bounds of Sense*. London, UK: Methuen.

Köhler, Wolfgang

Introductory article

Alfred D Kornfeld, Eastern Connecticut State University, Connecticut, USA

CONTENTS

Wolfgang Köhler (1887–1967) was a German–American Gestalt psychologist. He conducted seminal research into insight, and believed that psychology should model itself after field physics.

INTRODUCTION

Wolfgang Köhler was born in Reval, Estonia, in 1887. He attended the universities of Tübigen, Bonn, and Berlin, and was awarded a PhD in 1909 by Berlin for a dissertation on psychoacoustics under the direction of Carl Stumpf. At Berlin Köhler also studied with the Nobel laureate physicist Max Planck, whose quantum field physics would profoundly influence Köhler's ideas about the relationship between conscious experience and brain processes.

While working as an assistant at the Psychological Institute in Frankfurt am Main, Köhler collaborated with Max Wertheimer and Kurt Koffka on investigations of apparent movement: the phi phenomenon. Wertheimer's publication of their findings in 1912 marks the beginning of the Gestalt psychology movement. Köhler, along with Wertheimer and Koffka, would later be recognized as a leader of this movement.

INSIGHT RESEARCH

In 1913, Köhler was appointed as director of the Prussian Academy of Sciences Anthropoid Station at Tenerife in the Canary Islands. After the onset of the First World War, he was stranded at Tenerife, and was unable to return to Germany until 1920. This enforced isolation proved to be fortuitous, because it enabled him to conduct ingenious investigations of animal problem-solving which would earn him international recognition.

Prior to Köhler, researchers in the field of animal learning had stressed the gradual and mechanical formation of connections between stimuli and responses following reinforcement of 'correct' responses. Their investigations were based on the view that animals reproduce learned responses according to their previous experiences and rewards. Köhler's research, which was reported in his 1917 book *The Mentality of Apes*, led him to reject this interpretation. He found that chimpanzees could utilize simple objects to solve problems, and this suggested to him that animals could learn by perceiving relationships.

Köhler observed that chimpanzees were able to learn to manipulate boxes and bamboo poles and sticks in order to reach fruit that had been placed beyond their grasp. For example, some of Köhler's chimpanzees were able to obtain fruit placed at the top of their cages by building three-box towers. Sultan, Köhler's favorite chimpanzee, would achieve fame in the psychological literature for his ability to insert a small hollow rod into a larger hollow rod in order to create a stick long enough to reach fruit that had been placed outside his cage. Such intelligent solutions, which Köhler termed 'insight', were characterized by the sudden emergence of the correct response following a period during which the animal would often quietly stare at the objects and apparently 'think' about the problem.

Köhler's analysis of the nature and significance of insight was criticized for ignoring the role of previous learning. In fact, Köhler did not deny the importance of experience, but contended that what is vital is how experience is integrated with the features of the current situation. Köhler's analysis was also criticized because it seemed to imply the existence of a 'mysterious' inner agent responsible for solving problems. Köhler, however, emphatically denied that his analysis required the assumption of an inner problem-solving agent. He would later claim that his introduction of the term 'insight' was meant to be a description, rather than an explanation, of intelligent problem-solving behavior.

While at Tenerife, Köhler also investigated discrimination learning. He found that chimpanzees and chickens could learn consistently to choose the darker of two stimuli. Köhler concluded that both insight and discrimination tasks involve relational learning. (*See* **Animal Learning**)

PSYCHOPHYSICAL ISOMORPHISM

The concept of psychophysical isomorphism reflects Köhler's belief that psychology should be based on models derived from physics. Wertheimer had already used this concept in discussions of the phi phenomenon, but it assumed a more prominent role for Köhler. He pointed out that field physics described natural systems as organized functional wholes dependent on the relations among local conditions but not derived from the separate actions of parts. Such systems tend towards a dynamic equilibrium characterized by both regularity and simplicity and governed by laws of self-distribution, such as those governing the forces maintaining the shape of a soap bubble.

Köhler hypothesized that self-distribution principles should also regulate consciousness and brain activity. The theory of psychophysical isomorphism posits a one-to-one correspondence between the spatial and temporal organizations of conscious experience and cerebral electrical patterns. However, it does not predict a literal structural or quantitative identity between the psychological and neurological realms. For example, the perception of a square enclosing a circle should be correlated with a pattern of brain currents that follows the general organization of this relationship, without, however, reproducing it geometrically. In this respect, the relationship that Köhler proposed between the electrical activity of the brain and conscious experience has been compared to the relationship between a map and the actual city that it represents.

Much later, Köhler was to explore psychophysical isomorphism by examining the relationships between cortical patterns of excitation and figural aftereffects in vision. The results of these investigations appeared to support the theory of psychophysical isomorphism. However, when the physiological psychologist Karl Lashley subjected psychophysical isomorphism to further investigation by laying strips of metal foil on the surface of the visual cortex of monkeys, he found no visual disturbances in the monkeys even though the visual cortical fields were disrupted. These results were widely interpreted as a refutation of psychophysical isomorphism. Gestalt psychologists would later attempt to demonstrate how psychophysical isomorphism could be applied to thinking and perception, but their efforts were largely ignored because of these negative findings. (*See* **Object Perception, Neural Basis of**; **Vision: Form Perception**; **Reasoning and Thinking, Neural Basis of**)

KÖHLER AND AMERICAN PSYCHOLOGY

In 1922, Köhler was appointed Professor of Psychology and Director of the Psychological Institute at the University of Berlin. He was a visiting professor at Clark University in the United States from 1925 to 1926. While at Clark, he delivered a series of lectures that presented his chimpanzee research as well as general Gestalt organizational principles. These lectures were subsequently published, along with those of his contemporary and rival John B. Watson, in a volume entitled *Psychologies of 1925*.

Köhler's increased familiarity with American psychology, especially Watsonian behaviorism, was reflected in his 1929 book *Gestalt Psychology*. This book is widely considered to be the standard exposition of such Gestalt concepts as insight, part–whole relationships, psychophysical isomorphism, and the nature of phenomenological experience. Employing a model strongly influenced by quantum field physics, Köhler proposed parallel structural organizations in consciousness, the nervous system, and the physical environment. In this framework, everyday conscious experience illuminates the nature of the myriad relationships between the psychological and physical realms. Köhler was very critical of behaviorism's reductionistic methods and rejection of consciousness. He insisted that the phenomenal, everyday world of consciousness must be the starting point for all scientific investigation.

Although *Gestalt Psychology* was very influential, its partisan tone, and especially its severe critique of behaviorism, offended some American psychologists and created the unfortunate impression that Köhler's program was largely negative in nature.

In spite of these reservations Köhler attained wide recognition in American psychological circles for the originality and significance of his achievements. Harvard University invited him to deliver the 1934–1935 William James Lectures. These lectures provided the basis for the volume *The Place of Values in a World of Facts*, which presented a view of knowledge that integrated psychology, physical

science, and the philosophy of values. Köhler proposed that the experience of 'rightness' and 'wrongness' was determined by a phenomenally objective gestalt quality. This quality, which he termed 'requiredness', did not just operate in the realm of perception and thinking, but also furnished a basis for understanding the nature and meaning of values. Köhler's lectures were generally well received, especially by the philosophers in the audience, and Harvard's joint philosophy–psychology department considered offering him a position. However, the distinguished psychologist E. G. Boring, who was the director of Harvard's psychological laboratory, was disappointed by the non-experimental and philosophical content of the lectures and seems to have played a major role in dissuading Harvard from doing so. (Boring would later recognize Köhler's importance for American psychology.)

With the Nazi rise to power, Köhler found himself in an untenable position. Frustrated by Nazi interference in his work and angry about the dismissal of Jewish and politically suspect non-Jewish colleagues, he resigned from the Berlin Psychological Institute in 1935. A man of considerable personal courage and integrity, Köhler is credited with making the last anti-Nazi statement published in Germany before the Second World War, in which he defended Jews for their contributions to the science and culture of Germany. Köhler and his family emigrated to the United States. He was appointed a professor of psychology at Swarthmore College in 1935, and would remain there until 1958. After his retirement from Swarthmore, Köhler lectured at Dartmouth College and Princeton University. He was elected president of the American Psychological Association in 1958, and was a recipient of the Distinguished Scientific Contribution Award of that association.

Köhler died at Enfield, New Hampshire, in 1967.

EVALUATION OF KÖHLER

For more than fifty years, Wolfgang Köhler was a forceful advocate for Gestalt psychology. As a Gestalt psychologist, Köhler believed in the primacy of organizational principles in thinking and perception. The distinguishing characteristics of his approach to Gestalt psychology were his seminal research into insight and his conviction that psychology should model itself after physics. Köhler's concept of insight has become standard textbook material and is still a point of departure for discussions of problem-solving. His concept of psychophysical isomorphism has not fared as well, and is regarded by many as a historical curiosity derived from a discredited model of central nervous system functioning. Kohler's speculations about the brain processes associated with psychophysical isomorphism are now generally thought to be irrelevant to understanding perceptual and cognitive processes. (It has, however, recently been suggested that a more sophisticated model of the brain, based on connectionist networks, might resurrect the isomorphism hypothesis.)

Köhler seems to have relished his confrontational role as an advocate of the Gestalt movement. He vehemently criticized the extreme positivistic views and Procrustean methodology that dominated the psychology of his time. As an alternative, he proposed more flexible and creative research methods that would enable psychologists to investigate complex and meaningful psychological processes. Underlying his research and theorizing was the belief that Gestalt psychology would provide the basis for a revolution in the methods and subject matter of psychology.

This call for sweeping reforms in the field seems to have had little effect on the practice of mainstream psychology. Perhaps the explanation is that Gestalt psychology itself was viewed by many as too philosophical and unscientific. Köhler's ideas about the role of relational learning in solving problems were much more influential. The second generation of behavioristic psychologists, the 'neobehaviorists', well understood the challenge that Köhler's findings posed for their stimulus–response models of learning. They developed imaginative experimental paradigms and a more advanced stimulus–response framework in an attempt to meet this challenge. Thus, Köhler can be credited with encouraging neobehavioristic psychologists both to study phenomena they might otherwise have ignored and to develop more sophisticated models of learning.

Köhler's influence on contemporary psychology is evident in the areas of thinking and problem-solving. His idea of the role played by configural factors in recall resembles contemporary views of memory organization. Köhler's investigations of problem-solving emphasized nonassociated factors and relational elements, and may thereby have influenced information processing models of thinking. Although Köhler's focus on the active organizational principles underlying perception and thinking anticipated a major theme of current cognitive psychology, it would be inaccurate to call him a cognitive psychologist. Köhler was rooted in an earlier period that did not possess the

information processing and structural models of the mind that define modern cognitive science. Nonetheless, Köhler occupies a position of honor in the history of psychology for his role in directing attention to the primacy of organized mental processes in intelligent human and animal behavior. (*See* **Reasoning**; **Theory of Mind**)

Further Reading

Henle M (ed.) (1971) *The Selected Papers of Wolfgang Köhler*. New York, NY: Liveright.

Köhler W (1925) *The Mentality of Apes*, translated by E. Winter. New York, NY: Harcourt Brace.

Köhler W (1938) *The Place of Values in a World of Facts*. New York, NY: Liveright.

Köhler W (1940) *Dynamics in Psychology*. New York, NY: Liveright.

Köhler W (1947) *Gestalt Psychology*. New York, NY: Liveright. [Revised edition.]

Köhler W (1959) *The Task of Gestalt Psychology*. Princeton, NJ: Princeton University Press.

Lashley, Karl S.

Introductory article

Jack Orbach, Queens College, Flushing, New York, USA

CONTENTS

Karl S. Lashley was a major figure of twentieth-century neuropsychology, whose work on the relation between brain function and behavior, especially learning and memory, has proved seminal.

INTRODUCTION

Karl Spencer Lashley, one of the titans of the twentieth century in neuropsychology, was born in 1890 in Davis, West Virginia and died in 1958 in Poitier, France. A small-town Appalachian boy with the heart of a naturalist, he grew up to become a world-class animal psychologist. His research and theorizing concerning the relation between brain and behavior has had a lasting influence on psychology and neurology. That influence was further enhanced by the contributions of some of his students, particularly Frank A. Beach, Donald O. Hebb, and Roger W. Sperry, who was awarded the Nobel Prize in medicine in 1981.

The major question Lashley set for himself was: how does the brain work in the process of remembering things? His research convinced him that contemporary theories of learning and memory were erroneous – that learning could never be explained in terms of the formation of simple connections at synapses of particular neurons. His contributions to neuropsychology are collected in a volume of his selected papers, and his research and theories are reviewed in two volumes by Jack Orbach (see the Further Reading section).

EARLY YEARS

From childhood, Lashley's mother encouraged him in intellectual pursuits. He learned to read at the age of four and made avid use of the household library. But reading was not the only source of his attraction to science. He had the instincts of an observer and displayed a deep interest in animal and plant life. A favorite boyhood pastime was to wander in the woods, observing and collecting various species, including butterflies, snakes, frogs, snails, mice, and raccoons. As this interest in animals persisted throughout his life, he was never without a pet. At one time he owned a cat and a parrot, but the combination created unexpected problems when the voluble bird displayed a tendency to adopt the cat's kittens. Other pets included cockateels, monkeys, and dogs. One dog, '*Till Eulenspiegel*', developed a fondness for daiquiris and 'pink ladies'.

Lashley had a mechanical aptitude that appeared early. He was fascinated by his mother's sewing machine and learned to use it efficiently, later, as an adult, making sails for his boats and drapes for his home. Meanwhile, as a substitute for sewing, his father bought him a jigsaw to use in woodworking. This started Lashley on a hobby that became a lifelong pleasure. He produced a steady flow of elegantly designed and finely executed articles, even living room furniture. After retirement, he continued his cabinet-making and repair work, including the remodeling of his house. As a laboratory researcher, Lashley found that this aptitude served him well, and he demanded similar skills from his students, requiring them to construct their own devices for experiments.

When Lashley was seven, his father fell prey to gold fever and decided to take his family northwest to prospect. Years later he recalled the excitement of the gold rush and meeting with such swashbucklers as Swiftwater Bill and Klondike Pete, who sported nugget jewelry and recounted tall tales about fabulous strikes and the dangers of encountering bird-sized Alaskan mosquitoes!

At college, Lashley's plan was to major in Latin. In order to fill a vacant hour in his schedule, he enrolled in a course in zoology. The teacher of this class had a profound influence upon the 16-year-old boy: 'Within a few weeks in his class I knew that I had found my life's work.' After

earning his degree at the University of West Virginia, Lashley was awarded a teaching fellowship in biology at the University of Pittsburgh and enrolled as a Master's candidate. His thesis was on the permeability of the eggshell. But his most important contact that year was with another teaching fellow, Karl Dallenbach, who later said of Lashley: 'Though he had never taken a course in psychology, he was permitted to take my laboratory course. In this small class, we worked intimately together on various experiments. Lashley was intensely interested and was an outstanding student. In this course, he showed the promise that he later fulfilled.' After receiving his Master's degree, Lashley accepted a fellowship at Johns Hopkins University to work with the eminent biologist H. S. Jennings on the invertebrate *paramecium*. His doctoral research was on asexual reproduction of *hydra*. He received his PhD in genetics in 1914. Though interested in learning as a topic, he remained a nativist all his life.

PASSION FOR MUSIC

At age 11, Lashley had a few piano lessons but he found practicing scales impossibly boring. Then, at 18, he picked up the violin and learned to play without formal instruction. He claimed that he first heard classical music at 20 and was immediately fascinated by it. His first wife, a pianist, introduced him to the literature of chamber music. It didn't take long before he taught himself to play the cello (Figure 1). He collected an extensive library of instrumental music, joined the Jacksonville Orchestra and organized a small group of Florida musicians to meet regularly at his home (called 'Fiddler's Cove') for the playing of chamber music. He was a trustee and benefactor of the Jacksonville College of Music. Always on the alert for behavior that might shed light on how the nervous system worked, he once calculated the speed of finger movements involved in playing a fast cadenza on the piano, and compared this with the known speed of neural transmission. The comparison revealed that the intervals between successive finger movements were too short to support the theory that each movement is aroused by motor impulses set off by sensory impulses from the preceding finger movement. There is just not enough time for a sensory message from the finger to go to the brain and pass to the motor area and then for a motor impulse to trigger the next finger movement. Lashley cited this example to support the notion of 'central patterning' of complex motor sequences in the brain (something that today we call 'motor programs').

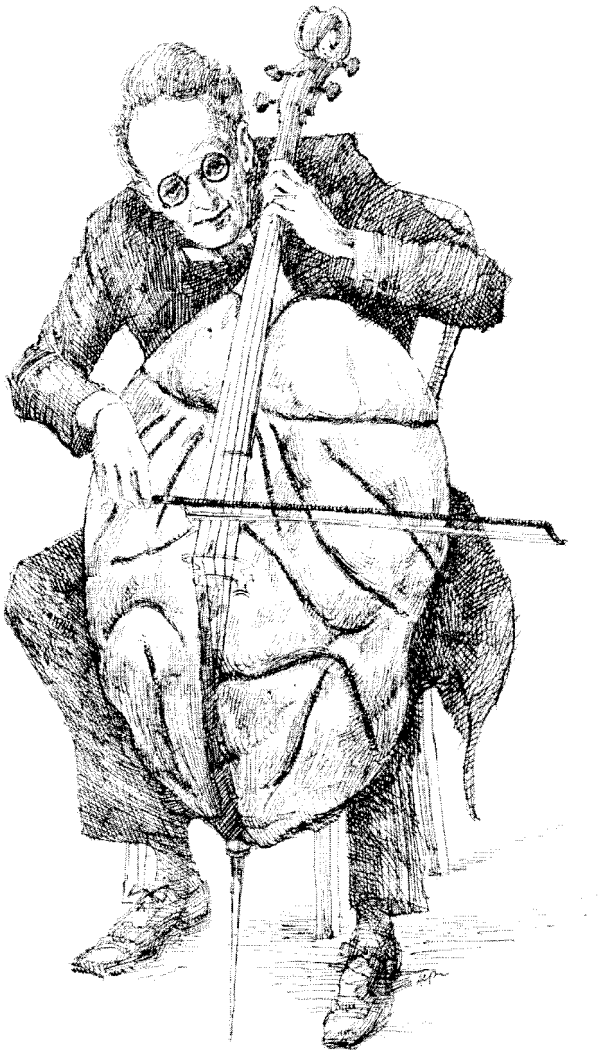


Figure 1. Karl S. Lashley as a cellist, fiddling with the monkey brain. Reprinted with permission from Lawrence Erlbaum Associates.

DEVELOPING INTEREST IN PSYCHOLOGY

Lashley's interest in psychology, first aroused in Pittsburgh, continued at Hopkins. While majoring in zoology, he took two minors: one with Adolph Meyer, professor of psychiatry; the other with John B. Watson, who became the father of behaviorism. Watson's impact on him was so great that, 44 years later, Lashley asserted: 'Anyone who knows American psychology today knows that its value derives from biology and from Watson.' In 1914, he joined Watson to carry out field experiments on homing, nesting, and reproductive behavior of sooty and noddy terns on the Dry Tortugas (west of Key West, Florida).

During the First World War, Lashley was assigned to educate the public and the military on the dangers of venereal diseases. Working together, he and Watson showed and discussed movies designed to further the campaign against these afflictions. The movies illustrated what damage could be wrought on the genitals. In later years, Lashley enjoyed telling of the time when they went to a small town and distributed advertisements announcing a free movie. The advertisements included no mention of the subject matter and, according to Lashley's account, he and Watson were fortunate to escape in one piece from the sheriff and enraged citizenry.

While holding a postdoctoral scholarship, Lashley continued to work with Watson, studying the effects of strychnine and other drugs on maze learning in rats. At the same time he journeyed frequently to Washington, DC to study the brain-lesioned monkeys of the psychologist Shepherd Ivory Franz. Eventually, he acquired the surgical and histological skills to embark on an ablation program of his own on the neural basis of learning and memory. This program brought him worldwide recognition, and his research career was solidly launched.

After stints in academic posts at Minnesota and Chicago, Lashley was chosen in 1935 for a chair at Harvard. The invitation came from a search committee charged by the President of the University to find 'the best psychologist in the world'. Not bad for someone who never took a didactic course in psychology! Finally, in 1942, he was appointed Director of the Yerkes Laboratories of Primate Biology in Orange Park, Florida, where he wrote some of his most memorable papers, including *In Search of the Engram* (1950), in which he concluded that he could not find the memory trace in any one place in the brain.

RESEARCH AND THEORY

In his 1929 monograph (see Further Reading), Lashley enunciated his controversial concepts of *mass action* and *equipotentiality.* Empirically, the term *mass action* summarizes the results of many brain ablation experiments – that the loss of the maze habit in rats is determined by the size of the lesion and not by its locus in the brain. Theoretically, *mass action* refers to a theory of how the cerebral cortex works, that it is the *pattern* of activity, independent of its *locus,* that is relevant, and that the memory trace is reduplicated and distributed in the brain. The concept *equipotentiality* refers to the fact that intact neurons can take over the function of destroyed cells. In his later years, Lashley discarded this concept and preferred to cite the facts of sensory and motor equivalence. Examples include the recognition of unfamiliar visual stimuli by the eye not used during monocular learning to recognize those stimuli (*interocular transfer*), and the performance of skilled movements by the hand not practiced during learning (*intermanual transfer*). Lashley loved to point out impishly that right-handers can write with their left hands, with their feet (on the beach), and even with their noses. In short, neurons that are not used during the course of learning can still show the effects of learning – that is, they can mediate memories (referred to as *Lashley's lesson* by Orbach, 1998). The recovery of function in patients after suffering from brain lesions also points to the same fact, that neurons inactive during learning can still show the effects of learning.

By the early 1920s, Lashley's research results on Pavlovian conditioning led him to break away from Watson's theorizing on learning. He found Watson's Stimulus–Response formula troubling because it failed to include the brain in the causal sequence. Does a stimulus cause a response? No, answered Lashley, a stimulus excites the brain, and it is the resultant activity in the brain that is responsible for the response. Thus, he revised Watson's formula to read Stimulus–Brain–Response. In this way, he provided for psychological functions that need to be sustained in the brain, such as selective attention and the memory trace. Unfortunately, the model of the day assumed that brain processes were linear – that is, the neural activity was thought to flow directly from input to output, as in a telephone line. It wasn't until 1938 that Lashley came up with a mechanism to explain how neural activity is sustained after the stimulus has ceased. This mechanism, called the reverberatory circuit, was borrowed from the neuroanatomical descriptions of Lorente de Nó. But it was Lashley's student Hebb who illustrated the wide application of reverberatory circuits for neuropsychological theory in a landmark book published in 1949. Hebb's theory of *cell assemblies* in the brain took the neuropsychological community by storm.

In 1952, Lashley wrote: 'I have never been able by any operation on the brain to destroy a specific memory. From such experiments, I have been forced to conclude that the memory trace is diffuse, that all memories are somehow represented in all or almost all parts of the cerebral cortex. Whatever the nature of the trace, it must be reduplicated throughout wide areas.'

Finally, in 1957, he wrote: 'The neuron is a living organism, subject to continual variation in its

functional capacity according to its metabolic state. It shows an all-or-none response in that it does or does not fire. But its 'all' may vary greatly from time to time. Comparison of the nervous system with a digital computer (and with soldered wire circuits) implies a uniformity in the action of neurons which is contrary to fact.' Thus, Lashley dismissed the rigid circuitry of early theories of artificial intelligence.

Further Reading

Beach FA *et al*. (1960) *The Neuropsychology of Lashley*. New York, NY: McGraw-Hill.

Hebb DO (1949) *The Organization of Behavior*. New York, NY: John Wiley & Sons.

Lashley KS (1929) *Brain Mechanisms and Intelligence*. Chicago, IL: University of Chicago Press.

Orbach J (1982) *Neuropsychology after Lashley*, chapters 1–5. Hillsdale, NJ: Lawrence Erlbaum Associates.

Orbach J (1998) *The Neuropsychological Theories of Lashley and Hebb*, chapters 1–8 and Epilogue. Lanham, MD: University Press of America.

Lewin, Kurt

Introductory article

Lenelis Kruse, University of Hagen, Hagen, Germany
Carl F Graumann, University of Heidelberg, Heidelberg, Germany

CONTENTS

Kurt Lewin (1890–1947) was one of the prominent figures in twentieth century psychology. His impact on psychology was fundamental to the development of experimental social psychology, cognitive psychology, and action research.

INTRODUCTION

Kurt Lewin is one of the most frequently quoted psychologists of the twentieth century. His work is usually associated with words such as field theory, life space, barrier, valence, success and failure, leadership style, group dynamics, and action research. Lewin is claimed to be the father of various developments in psychology, such as experimental social psychology, cognitive psychology, and action research.

Lewin became a prominent figure in both German and American psychology as a result of his engagement in many different areas of research as well as of application, and through his role as a highly influential teacher who gathered around him groups of talented students. The antithesis of an armchair psychologist, Lewin from his earliest student years was politically active in various areas of public life.

BIOGRAPHY

Born in 1890 in a small town in Prussia (now belonging to Poland), Kurt Lewin grew up as one of four children in a Jewish family. Later on the family moved to Berlin where Lewin graduated from a humanistic gymnasium. As a student he first concentrated on medicine and biology, but soon became interested in philosophy, theory of science, and psychology. His dissertation on mental activities in the inhibition processes of volition was published almost at the same time as another, quite different, paper on the 'war landscape' that reflected Lewin's experiences while he served in the army during the First World War. These papers anticipate the wide scope of Lewin's later research interests, which ranged from experimental research on volition to conceptualizations of ecological problems.

Lewin's interest in social problems became apparent when he joined a socialist student group and took an active part in social and educational activities for workers in agriculture and industry. From this interest in practical problems resulted two 'applied' publications, one on rationalization in agriculture, one on the socialization of the Taylor system.

On his return from the war he was appointed to the Psychological Institute of Berlin University, Lewin worked with Wolfgang Köhler, the leader of the Berlin Gestalt psychology group. Later on he became an assistant in the new section of applied psychology, where he remained until his emigration in 1933. Although he received the title of professor in 1927 it was impossible for a Jewish citizen to be appointed to a tenured position. During these years Lewin concentrated on experimental studies of will, affect, and action, which became the basis for the development of his field theory. He always worked with talented students, some of whom not only shared his later fate as refugees but also won international reputations by their association with Lewin (they included Dembo, Zeigarnik, and Rickers-Ovsiankina). The year 1929 became an important date in Lewin's life. Through an American student who had worked with Lewin and had made the work of the Lewin group known to American psychologists, Lewin received an invitation to attend an international congress at Yale University. His lecture on environmental forces (presented in German!) and a short film on the effect of these forces demonstrated by a small girl trying to sit on a stone apparently impressed his audience so much that he was invited as visiting professor to Stanford University in 1932. Lewin was forced to leave Germany when the Nazis seized power in 1933, and was fortunate to receive

a timely invitation to join the faculty at Cornell University. After 2 years he was offered a position at the Iowa University Child Welfare Research Station, where he worked for 8 years as a professor of child psychology. During these years Lewin won excellent students as disciples and later as colleagues, such as Barker, Cartwright, Festinger, French, Lippitt, White, and Zander. His research interests shifted to social psychology. In 1937 he became one of the founders of the Society for the Psychological Study of Social Issues.

Another institution which still exists at the University of Michigan in Ann Arbor is the Research Center for Group Dynamics. Originally founded at the Massachusetts Institute of Technology in 1945, it reflected the social engagement of Lewin and his increasing concern with the interaction between research and practice which culminated in his conception of 'action research'. Students and colleagues at the center developed methods of studying and changing human relations. This institution became the frame of Lewin's last period of life, ended abruptly by a heart attack at the age of 56 in 1947.

RESEARCH TOPICS

Since the Lewin heritage is mostly associated with field theory it is appropriate to take a closer look at its central themes and concepts. The key term of field theory is the 'life space' (or total situation); i.e. the sum of all facts that determine a person's behavior at a given time. From this conception the popular and much debated formula was derived that $B = f(LS)$: behaviour B is a function of the life space $f(LS)$ and the latter is the product of the interaction between a person P and the person's environment E. Theoretically and methodologically essential is the interdependence between P and E. As a nonmetric mathematical discipline, topology is of interest for field theory since it focuses on spatial relations of which the most important is the whole–part relation – a favorite topic of Gestalt theorists. Topological psychology describes 'regions' and 'boundaries' and the relations between them as well as 'locomotion' within and between such regions.

A few words are in order about the dynamics of the field, for which the central term is the 'tension system'. It exists whenever in P a psychological need or intention exists with respect to a quality of the psychological environment; i.e. the environment as it is experienced by a person. Such qualities are called 'valences' and are either positive (attractive) or negative (qualities to be avoided by P). Tension systems and their interdependence have been shown to be at work in a series of experiments dealing with a variety of dynamic or motivational phenomena, such as rigidity, substitution, satiation, frustration, forgetting and – of special interest to social psychologists – the studies of various forms of conflict resulting from the interaction between the 'driving' and 'restraining' forces of a psychological field, limited by a 'boundary' and 'barriers'. Since the conception of a field of forces, governed by the methodological principle of interdependence, is not restricted to individual persons, Lewin successfully transferred it to field experiments in which he and his group studied the phenomena and dynamics of group life, an innovation in the history of experimentation. Of these group experiments the ones dealing with leadership style and group 'atmosphere', such as 'autocratic', 'democratic' and '*laisser-faire*', have gained lasting notoriety and popularity.

In some of Lewin's social psychological studies his own fate is mirrored as a member of a minority group, as a Jewish emigrant, and as a person who had to adapt to another culture. Mainly these studies underline his interest in a synthesis of social research and social change.

ASSESSMENT

For several decades, mainly in the social psychological literature of the 1970s and 1980s, Kurt Lewin was almost unanimously hailed as one of the great men in this field. In this period the 'Lewin tradition' (Patnoe) or the 'Lewin legacy' (Stivers and Wheelan) seemed to be part and parcel of contemporary social psychology. The postwar history of this field, centered as it has been in the USA, seemed to be dominated by Lewin and his followers. Leon Festinger, himself a prominent associate of Lewin's, documented in 1980 how three generations of Lewin's students and their students had built and shaped modern social psychology over a period of more than thirty years. Festinger also tried to make it clear that Lewin's final creation, the Research Center for Group Dynamics, was an effective and lasting monument to him. It is remarkable, therefore, that when Festinger went on to assess Lewin's impact on social psychology he felt obliged to state that 'for at least two decades the social psychology literature has been virtually devoid of mention of … Lewinian concepts and terms'.

What is Lewin's stature at the beginning of the twenty-first century? Terms characteristic of Lewinian theorizing that were already being omitted

from the literature of thirty years ago are still rare; at any rate, they have lost the specificity that Lewin meant to give them. While 'field theory' is still a frequently used term it is no longer connected with the Lewinian components of 'topology', 'hodological space' and 'vectors', while many terms that were originally or temporarily defined by Lewin's theorems have now acquired (or regained) a much broader, almost loose meaning – e.g. 'force', 'valence', 'life space', and 'group dynamics'.

Irritating to an outside observer is also the presence of two traditions with little overlap. On the one hand there is the Lewin–Festinger–Schachter tradition as represented in Festinger's and Patnoe's books. It is the tradition of experimental social psychology from which the present mainstream interest in social cognition may be derived. On the other hand, under the (also Lewinian) title of 'group dynamics', there is the broad movement of sensitivity training techniques, interpersonal workshops, and encounter groups.

A common origin of both forms of group dynamics, the experimental and the therapeutic variety, is seen by some historians in Lewin's late interest in 'action research', in which 'combination of social research, action, and research evaluation action' (as described by Back) Lewin had tried to systematize experimentation with small groups for purposes of social change, an effort that he could not bring to a conclusion owing to his untimely death. Hence, a certain ambiguity has remained; in books and journals entitled 'group dynamics' we occasionally find both small group research and group therapy. However, most experimental social (and cognitive) psychologists have come to avoid the term for clarity's sake. Although, strictly speaking, field theory, topology, and action research no longer rank among the major methodologies of psychology, and Kurt Lewin is cited in a historical context rather than in the research chapters of modern textbooks, Lewin's overall impact on contemporary psychology must not be underrated. While his theory and methodology in the 1960s could still be offered as a 'contemporary systematic framework', it has meanwhile almost disappeared from the agenda of theories. Social psychology has increasingly become individualistic (again).

Historically, it is interesting to see that Lewin and his generation suffered the same fate that the 'Gestalt and field' generation inflicted upon their predecessors and their theories. Theories are not necessarily replaced by refuting them but very often by providing new perspectives that are comprehensive enough to allow for some kind of continuity. Lewin's (and the Gestalt psychologists') emphasis on the perceptual (cognitive) primacy of the self-organizing whole or field over its component 'elements' as well as on the interdependence between features of the person and the person's environment have become elements of the common stock of psychological knowledge.

Further Reading

Ash M (1995) *Gestalt Psychology in German Culture, 1890–1969: Holism and the Quest for Objectivity*. Cambridge, UK: Cambridge University Press.

Back KW (1973) *Beyond Words. The Story of Sensitivity Training and the Encounter Movement*. Baltimore, MD: Penguin Books.

Festinger L (ed.) (1980) *Retrospections on Social Psychology*. New York, NY: Oxford University Press.

Graumann CF (ed.) (1981) *Kurt Lewin Werkausgabe*, 7 vols. Bern: Huber.

Lewin K (1935) *A Dynamic Theory of Personality*. New York, NY: McGraw-Hill.

Lewin K (1936) *Principles of Topological Psychology*. New York, NY: McGraw-Hill.

Lewin K (1951) In: Cartwright D (ed.) *Field Theory in Social Science*. New York, NY: Harper.

Marrow AJ (1969) *The Practical Theorist. The Life and Work of Kurt Lewin*. New York, NY: Basic Books.

Patnoe S (1988) *A Narrative History of Experimental Social Psychology – The Lewin Tradition*. New York, NY: Springer-Verlag.

Stivers E and Wheelan S (eds) (1986) *The Lewin Legacy. Field Theory in Current Practice*. Berlin, Germany: Springer-Verlag.

Luria, Alexander R.

Introductory article

David E Tupper, Hennepin County Medical Center, Minneapolis, Minnesota, USA

CONTENTS

Alexander Romanovich Luria (1902–1977) was a Russian neurologist and psychologist recognized as one of the preeminent neuropsychologists of the twentieth century. His primary interest was the study of the cerebral localization of psychological functions, but he also was involved in developmental psychology, cross-cultural cognitive research, the study of frontal lobe impairments, aphasiology, educational and rehabilitative interventions, and personal biographies of individuals with unusual cognitive characteristics.

INTRODUCTION

Alexander Romanovich Luria (Figure 1) was born in Russia on 16 July 1902, in Kazan, a city east of Moscow. After a cultured upbringing in his early years, during which he learned German, French, and English and demonstrated quick acquisition of philosophy, history, and literature, Luria entered the University of Kazan in 1917. He remained a student in the newly formed department of social sciences until 1921, when he graduated with a degree in the humanities. Luria was an active student, and his study of sociology also prompted an interest in psychology. After graduation, Luria took a job as a laboratory assistant at the Institute of the Scientific Organization of Labor, and also continued to study in Kazan at the Pedagogical Institute and the medical school. Early in his career, Luria recognized that a valid approach to psychology needed to incorporate dynamic aspects of behavior and include both biological and cultural influences. He moved to Moscow at the end of 1921 and, soon thereafter, developed such a deep interest in psychoanalysis that he set up and chaired a psychoanalytic circle, and arranged the publication of important psychoanalytic works. Correspondence with Sigmund Freud confirmed permission for the circle to provide authorized translations into Russian of his important psychoanalytic writings, which was overseen by Luria. (*See* **Freud, Sigmund**)

Figure 1. Alexander R. Luria during the Eighteenth International Congress of Psychology, held in Moscow in 1966. Reproduced from E. D. Homskaya, *Alexander Romanovich Luria: A Scientific Biography*, with the permission of Kluwer Academic/Plenum Publishers, © 2001.

In 1922 Luria was offered a position at the Moscow Institute of Psychology, where he focused his research on more objective methods to study affective processes, rather than the more subjective methods used in psychoanalysis. This work led to his development of the 'combined motor method' by which Luria measured motor responses to affective stimuli in various clinical, criminal and student groups with a chronoscopic apparatus similar to a lie detector. The ideas developed in this research probably represent the early evolution of Luria's interest in self-regulatory mechanisms, which became prominent in his later research on

the frontal lobe. In 1923, Luria worked briefly in education at the Krupskaya Academy of Communist Education. He initiated new research in that setting, analyzing verbal associative reactions in children to study the development of speech and cognition.

Luria's life was changed in January 1924 when he met Lev Vygotsky, a technical school teacher from Gomel, at the Second All-Russian Congress on Psychoneurology in Leningrad. Luria was so impressed with Vygotsky's approach to psychology that he helped arrange an invitation for Vygotsky to work in Moscow. From the time that they first met to the time of Vygotsky's death in 1934, Luria and Vygotsky, along with Alexei Leontiev, a like-minded colleague, worked together to create a practical and Marxist-oriented Soviet psychology described as a 'cultural-historical' approach. The main principle that united Luria and Vygotsky in this new theory was the idea that psychology could view higher psychological functions (namely, conscious activity) only in the context of their development in historical and cultural processes and demonstrated via objective principles of brain function. This 'troika' of researchers, sharing a common theoretical base, collaborated on a vigorous research program in Moscow, and studied the mediated nature of psychological functions in children and adults, as well as the organization of psychological functions in patients with aphasia, Parkinsonism, and learning difficulties. The central aspect of the theory for the research was the role of cultural mediation in the constitution of human psychological processes, and the role of the social environment in structuring those processes. The researchers believed that higher mental processes are formed initially between people, in social interaction, and only later become internalized in individuals as inner speech and other cognitive structures. All higher mental functions are therefore initially culturally determined and mediated. Although Vygotsky's further elaboration of this theory was cut short by his premature death, Luria remained true to a dynamic, cultural-historical theory for the rest of his life, even in his neuropsychological practice. (*See* **Vygotsky, Lev**; **Learning and Instruction, Cognitive and Situative Theories of**)

CROSS-CULTURAL EXPEDITIONS

During the summers of 1931 and 1932, Luria organized two psychological expeditions to central Asia in an attempt to further support the cultural-historical theory. The research was planned with a number of prominent investigators, including Vygotsky, who was too ill to actually join the trips. The purpose of the investigations was to study the influence of the cultural and social environment on the development of psychological processes, as rapid sociological change was taking place in the villages of Uzbekistan and Kirghizia at that time. More specifically, the researchers were interested in changes in perception, problem-solving and memory associated with historical changes in economic activity and schooling, and a number of naturalistic experiments with both literate and illiterate people were designed and implemented during the expeditions. The investigators used observational and clinical methods to test their hypothesis that the structure of human cognitive processes differs according to the ways in which various social groups live out their lives. They found that people whose lives are dominated by concrete, practical activities have a different method of thinking from people whose lives require abstract, verbal, and theoretical approaches to reality. Despite a number of problems interpreting the data, the investigators concluded that cultural factors affect mental processes and that these differences vanish as soon as people are exposed to more industrialized situations. Unfortunately, results from the cross-cultural investigations were not fully analyzed or published immediately, as the Stalinist government considered such research racist; Luria's complete account of the research was not published in book form until 40 years later. (*See* **Culture and Cognitive Development**)

After Vygotsky's death in 1934, Luria returned to Moscow where he joined the Medico-Genetic Institute and studied the psychological development of twins, focusing on the relative contributions of biological and social determinants to cognition. Luria continued his twin research at the Medico-Genetic Institute until 1936 when such research was proclaimed illegal by the government. He then defended his reorganized manuscript on the combined motor method as a doctoral thesis in Tbilisi and reentered full-time medical school, subsequently graduating as a medical doctor in 1937. Luria thereby became one of the youngest psychologists and medical doctors in the country.

NEUROPSYCHOLOGICAL CONTRIBUTIONS

In the years before the Second World War, Luria worked mainly in the field of neuropsychology in the neurological clinic of the State Institute of Experimental Medicine. Working with many different

kinds of neurological and neurosurgical cases, Luria first developed his personal, creative, detective-like approach to neurological diagnosis, which involves searching for damaged psychological capabilities within functional psychological systems. Luria's approach to neuropsychological diagnosis was qualitative in nature, and used a variety of 'simple' tasks presented to the patient until a common defective cognitive link (factor) could be identified, which indicated the impaired cerebral region. Because of his earlier cultural-historical theorizing, Luria also became particularly interested in the cerebral organization of speech mechanisms, and the effects of aphasia on voluntary actions.

With the advent of war Luria was called away to organize a rehabilitation hospital in Kisegatch in the southern Urals. This renovated sanatorium was converted into a specialized rehabilitation facility for soldiers recovering from dysfunctions caused by brain wounds. Along with the medical and rehabilitation treatment provided there, Luria was particularly interested in the rehabilitation of mental activity in patients with localized brain damage, and he applied his theory of the dynamic systemic localization of higher psychological functions at this time. In contrast to narrow localizationist or nonspecific equipotentialist theories of cerebral organization, Luria's systemic theory maintained that one should conceptualize the higher mental functions as functional systems of various brain regions working in concert to complete socially determined activities. Luria was one of the first psychologists in the world to attempt an integration of major neuropsychological views, and he is considered by many to be the founder of neuropsychology in the Soviet Union. At the end of the war, Luria returned to the Burdenko Neurosurgical Institute and started teaching at Moscow State University, with which he kept in close academic contact for the rest of his life. (*See* **Brain Damage, Treatment and Recovery from**)

A joint session of the Academy of Medical Sciences and the Academy of Sciences took place in Moscow in 1950, and resulted in the interruption of Luria's neuropsychological work and the closing of his laboratory in the Neurosurgical Institute. The meeting was an ideological defeat of Soviet biological and medical sciences unless based on Pavlovian principles, and caused Luria to take refuge in a more sanctioned position in the Institute of Defectology of the Russian Federation's Academy of Pedagogical Sciences. In this challenging political climate, it was necessary for Luria to modify the nature of his research. He returned to his past use of the combined motor method and applied this method to the study of the development of verbal regulation in normal children and children with mental retardation. Based on earlier work initiated with Vygotsky, he continued to study the ontogenesis of the regulating function of oral and written speech. Luria also modified his clinical methods and created diagnostic tests and methods of educational instruction for these children. He more generally pursued his scientific interests within the context of a series of studies of the development of language and thought in children with mental retardation, and wrote several important monographs on these topics. (*See* **Pavlov, Ivan Petrovich**; **Neuropsychological Development**)

In 1959 Luria's laboratory at the Burdenko Neurosurgical Institute was reestablished, and he was permitted to return to the study of neuropsychology. In the remaining 18 years of his life, Luria's laboratory became a leading neuropsychological institution in the Soviet Union. In addition to his clinical activities and increased teaching responsibilities at Moscow State University, Luria continued to develop and expand his ideas concerning the cerebral organization of psychological processes. He became particularly interested in the analysis of speech and language disturbances, and executed further investigations into aphasia and neurolinguistics, the regulative role of speech and language in frontal lobe activities, and the diagnosis of local brain damage, all themes that had interested him in earlier years. (*See* **Language Disorders**; **Aphasia**; **Executive Function, Models of**)

Among the most important of Luria's contributions to neuropsychology is his conceptualization of the three functional units of the brain. Initially developed as a didactic tool, Luria's concept postulated three main functional cerebral blocks organized according to the dynamic localization of higher mental functions. Luria stated that the whole brain participates in each mental activity, with each of the three units or blocks of the brain contributing specific components. Thus, the first unit, housed in the brainstem, is responsible for general cerebral activation and arousal, and participates in wakefulness and attention; the second functional unit, located in posterior cerebral regions, serves to process sensory information, and provides for analysis, coding, and storage of the information; while the anterior and especially the frontal regions of the brain are responsible for the development and execution of intentions, plans and movements. Depending upon the area of the brain or unit affected by brain damage, any of a variety of functional psychological disorders could result. Luria's books *Higher Cortical Functions in*

Man and *The Working Brain* contain detailed descriptions of his neuropsychological diagnostic methods, his theoretical conceptualizations of functional units and dynamic localization, and historical and background information on the evolution of neuropsychological concepts. (*See* **Frontal Cortex**)

CASE STUDIES IN ROMANTIC SCIENCE

In the later years of his life Luria was a prolific writer on diverse psychological topics. Returning to his early ideal of creating a unified psychology, Luria contributed two detailed and personal-scientific summaries of unique individuals designed to provide a more synthetic, romantic scientific perspective for psychology, to contrast with classical, analytic approaches.

The first case study described S., an exceptional mnemonist who was endowed with a virtually limitless memory. Luria's account is documented in *The Mind of a Mnemonist*, where he considers not only the unique character of S.'s memory but also reveals important aspects of S.'s mind, behavior, and personality. Luria worked with S. for over 30 years and came to understand not only the exceptional gifts that S. possessed but also the unique handicaps that S. experienced in his personality and daily life. The second case study, *The Man With a Shattered World*, describes a young soldier, Zasetsky, who suffers a traumatic gunshot wound to the left hemisphere which impaired a number of higher cognitive abilities, and required Zasetsky to relearn many of these abilities in order to compensate for his lost faculties and try to live a normal life again. Over the course of 25 years, with the use of writing as his only means of expression, Luria assisted Zasetsky to rebuild his memories and his grasp on reality. These two cases illustrate Luria's interests in developing a person-centered psychology that would synthesize classical and romantic science.

During the latter part of his career, Luria received numerous awards and honors in the Soviet Union and internationally. He remained professor of psychology and chair of neuropsychology at Moscow State University for many years, and had a long association with the Burdenko Neurosurgical Institute, where he was director of the neuropsychology laboratory. Luria died from heart failure on 14 August 1977.

CONTEMPORARY IMPLICATIONS

Contemporary neuropsychology has benefited from Luria's work in a number of ways. Luria maintained a strong conceptual emphasis and internally consistent theory throughout his work, based on cultural-historical theorizing and culminating in his theory of the dynamic systemic localization of cerebral functions. This theory is widely accepted and used worldwide, and counters the more quantitative and atheoretical approaches to neuropsychology developed in North America. Luria's three-block model of cerebral organization has also provided the impetus for the creation of newer and more complex models of intellectual or cognitive functioning, as exemplified by J. P. Das's planning–attention–simultaneous–successive (PASS) model of cognition. A number of clinical neuropsychological batteries have evolved from Luria's diagnostic methods, including Luria's Neuropsychological Investigation developed by Christensen, and the Luria–Nebraska Neuropsychological Battery by Golden and colleagues. Finally, Luria's rich clinical and theoretical case descriptions of individuals with self-regulatory impairments from frontal lobe damage are unparalleled in contemporary neuropsychological literature, and have instigated significant research endeavors into these mysterious executive functions.

During his 75 years, Alexander Romanovich Luria became a respected, prolific and important neuropsychologist. His work has had a significant global influence on psychological theory and practice, particularly in the area of neuropsychology. As a contemporary of Lev Vygotsky, Luria was significantly influenced by Vygotsky's approach to understanding the mind as inseparable from the surrounding society and culture. Although his later work contributed greatly to the developing field of neuropsychology, the cultural-historical theme provided the basis for Luria's work all his life, and places him in a prominent position in the history of Soviet psychology.

Further Reading

Cole M (ed.) (1978) *The Selected Writings of A. R. Luria*. White Plains, NY: M. E. Sharpe.

Das JP (1994) Luria AR (1902–1977) In: Sternberg RJ (ed.) *Encyclopedia of Human Intelligence*, pp. 678–681. New York, NY: Macmillan.

Homskaya ED (2001) *Alexander Romanovich Luria: A Scientific Biography* (DE Tupper, ed., translated by D Krotova). New York, NY: Kluwer/Plenum.

Luria AR (1968) *The Mind of a Mnemonist*, translated by L Solotaroff. New York, NY: Basic Books.

Luria AR (1972) *The Man With a Shattered World*, translated by L Solotaroff. New York, NY: Basic Books.

Luria AR (1973) *The Working Brain: An Introduction to Neuropsychology*, translated by B Haigh. New York, NY: Basic Books.

Luria AR (1974) AR Luria. In: Lindzey G (ed.) *A History of Psychology in Autobiography*, vol. VI, pp. 251–292. Englewood Cliffs, NJ: Prentice-Hall.

Luria AR (1976) *Cognitive Development: Its Cultural and Social Foundations*, translated by M Lopez-Morillas and L Solotaroff. Cambridge, MA: Harvard University Press.

Luria AR (1979) *The Making of Mind: A Personal Account of Soviet Psychology*, translated by M Cole and S Cole. Cambridge, MA: Harvard University Press.

Luria AR (1980) *Higher Cortical Functions in Man*, 2nd edn, translated by B Haigh. New York, NY: Basic Books.

Tupper DE (ed.) (1999) International extensions of Luria's Neuropsychological Investigation, Parts I and II. *Neuropsychology Review* **9**(1): 1–56, **9**(2): 57–116 [special issues].

Marr, David

Introductory article

Lucia M Vaina, Boston University and Harvard Medical School, Boston, Massachusetts, USA

CONTENTS

David C. Marr (1945–1980) was a theoretical neurophysiologist and cognitive scientist whose work symbiotically integrated data from experimental neuroscience and psychophysics with novel computational models, thus providing an explicit foundation for the field of computational neuroscience.

INTRODUCTION

David Courtnay Marr (Figure 1) was born on 19 January 1945 in Essex, England. He went to the English public school Rugby, on a scholarship, and between 1963 and 1966 he studied at Trinity College, Cambridge, where he obtained his BA degree in mathematics with first-class honors. For his doctoral research he continued in theoretical neuroscience, under the supervision of Giles Brindley. His education involved training in neuroanatomy, neurophysiology, biochemistry and molecular biology. At Trinity College in 1971 he received an MA (with distinction) in mathematics and a PhD in theoretical neurophysiology. After obtaining his PhD, he accepted a research appointment at the Medical Research Council (MRC) Laboratory of Molecular Biology under Sydney Brenner and Francis Crick, and he retained an affiliation with the MRC until 1976.

CONTRIBUTIONS TO THEORETICAL NEUROSCIENCE: WHAT IS IT THAT THE BRAIN DOES?

In three successive papers that combine high-level theoretical speculation with meticulous synthesis of the available neuroanatomical data, Marr proposed a definite answer to this question for the cerebellum, archicortex and neocortex. Common to these three studies is the idea that the central function of the brain is statistical pattern recognition and association in a very high-dimensional space of 'elemental' features. The basic building block of all three theories is the codon, or a subset of features, with which a cell that is wired in such a way as to fire in the presence of that particular codon is associated.

A paper entitled 'A theory of cerebellar cortex', published in 1969 (see Further Reading), represents the essence of Marr's doctoral research. The paper is a theoretical model that made critical predictions elucidating how the cerebellum learns the motor skills involved in performing actions and maintaining posture and balance. The fundamental elements of the model are the known cell types in the cerebellum, their connectivities and the synaptic

Figure 1. David Marr (1945–1980).

actions of the cerebellar cortex. The process involves context recognition and learning. The former was described at the level of the mossy fiber–granule cell–Golgi cell circuitry, and the latter was described at the level of the parallel fiber–Purkinje cell synapse, heterosynaptically strengthened by the inferior olive climbing fiber. Linked through learning to the context of the previous movement in the sequence, the Purkinje cell, presumably implementing the codon representation, associates (through synaptic modification) a particular movement with the context in which it is performed. Subsequently, the context alone causes the Purkinje cell to fire, which in turn precipitates the next elemental movement. Basically the cerebellum model is a one-layer network (of granule cells) with fixed synapses, and an associative memory store and a set of conditioning inputs (the climbing fibers).

The second paper, entitled 'A theory for cerebral neocortex', published by Marr in 1970 (see Further Reading), extended the codon theory to encompass a more general type of statistical concept learning, which he assessed as being 'capable of serving many of the aspects of the brain's functions', in particular the formation and organization of networks capable of classifying and representing 'the world'. This hypothesis is an early attempt at a theory of unsupervised learning relating to methods of cluster analysis. The paper discusses the structure of the relationships which appear in the afferent information, and the usefulness to the organism of discovering them. These two ideas are combined by the 'fundamental hypothesis' which is based on the existence and prevalence in the world of a particular type of 'statistical redundancy'. The fundamental hypothesis, as set out in Marr's 1970 paper, states that:

> Where instances of a particular collection of intrinsic properties (i.e. properties already diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist which generalize over such instances. Further, properties often are grouped in this way.

The neocortex model keeps track of probabilities of events, and to do this it needs an extensive memory of a special kind, allowing retrieval that is based on the content rather than the location of the items. In his third theoretical paper, entitled 'Simple memory: a theory for archicortex', published in 1971 (see Further Reading), Marr considers the hippocampus as a candidate for fulfilling this function. In analyzing the memory capacity and recall characteristics of the hippocampus, Marr integrated combinatorial–mathematical constraints on the representational capabilities of codons with concrete data derived from neuroanatomical and neurophysiological studies. In modern terms, the hippocampal model consists of a recurrent network with two layers of trainable 'hidden' units that encode and classify input patterns connected to an associative memory store. The paper postulated the involvement in learning of synaptic connections modifiable by experience – a notion that originated from the research of Donald Hebb in the late 1940s. The paper is a mathematical proof of efficient partial content-based recall by the model, and it offered a functional interpretation of many anatomical structures in the hippocampus, together with concrete testable predictions.

'Truth, I believed, was basically neuronal, and the central aim of research was a thorough analysis of the structure of the nervous system'. This view expressed by Marr in 1982 in his book, *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information* (see Further Reading), combined with his initial training in mathematics, shaped the quantitative, analytical methodology that he applied in these three studies. In a letter to Francis Crick in 1977 he summarizes his fundamental views as follows:

> For a mathematician, understanding (or explanation) is all, yet in science proof is, of course, what counts. In the case of information-processing devices, understanding is very important; one can know a fact about a device for years without really understanding it, and part of the theoretician's job is to place into a comprehensible framework the facts that one already knows. I still think that the cerebellum is a good example. For sure, the idea that the parallel fiber–Purkinje cell synapse might be modifiable may not have been very difficult to arrive at, and other theories have since incorporated it, but that surely is only part of the story. I found the real impact of the story to lie in the combinatorial trick. That is, the granule cell arrangement, with associated inhibitory interneurons, had been right in front of people's eyes ever since Cajal (modulo inhibition and excitation), but its significance had not been appreciated. Of course my theory might be wrong, but if it is right, then I would regard a major part of its contribution as being explanatory. And also, that is almost inevitable.

Many of the ideas developed in these three papers were subsequently extended and adapted to be consistent with later neurobiological discoveries. This topic was addressed in *From the Retina to the Neocortex* by L. M. Vaina (see Further Reading).

A PIONEER OF COMPUTATIONAL NEUROSCIENCE

After the publication of these three fundamental papers, Marr moved to the Massachusetts Institute of Technology (MIT) Artificial Intelligence Laboratory where he was a visiting scientist in the group of Marvin Minsky and Seymour Papert. 'Since the facilities and the people were really impressive' (as he wrote to Sydney Brenner in 1973), Marr relocated from Cambridge, England to Cambridge, Massachusetts for a faculty appointment in the Department of Psychology at MIT, and in 1980 he was promoted to full professor with tenure.

While at MIT, his decision to break with the previous research was stated clearly in a letter to Giles Brindley (written in October 1973):

> I do not expect to write any more papers in theoretical neurophysiology – at least not for a long time. I do not regard the achievements of your 1969 or my papers negligible. At the very least, they contain techniques that anyone concerned with biological computer architecture should be aware of.

This decision was motivated by his realization that, without an understanding of specific tasks and mechanisms – the issues from which his earlier theories were 'once removed' – any general theory would always be incomplete.

He proposed a new methodology for understanding the brain by essentially inventing a field and a mode of study now referred to as *computational neuroscience*. In his opening remarks at a workshop organized in 1972 by Benjamin Kaminer at Boston University, Marr suggested an 'inverse square law' for theoretical research, according to which the value of a study varies inversely with the square of its generality – an assessment that favors top-down reasoning firmly supported by functional (computational) understanding, together with bottom-up work grounded in an understanding of the mechanism.

Proposing that the primary unresolved issue in brain science was what function must be implemented and why, Marr argued fiercely against the usefulness of the theoretical approaches to brain science adopted in the early 1970s, such as the catastrophe theory pioneered by Rene Thom, and neural nets (of that time). Instead, he proposed a fundamentally novel approach to biological information processing which required that any problem must be addressed at several different levels of abstraction. What exactly was the task executed by the system? On what properties of the world could a system performing this task be expected to rely? What methods could be shown to be effective in the performance of the task? Given a particular method, what are the appropriate algorithms for implementing it? Given a particular algorithm, what neural circuitry would be sufficient to perform it? These questions formed the core of Marr's research philosophy, and they were explicitly formulated as three levels of explanation of information processing. At the highest level is a computational theory – that is, a theory of how a task could be performed. The computational theory must specify what is being computed and why it is a useful thing to compute. At the next level is a representation and an algorithm (or a set of algorithms) to achieve that representation. At the third level lies the question of how the algorithm is actually implemented in the hardware of the system. A key point in Marr's approach is that the three levels should be considered relatively independently.

Marr's originality and depth of thinking stem from his emphasis on the computational theory level – not because it was the most important level, but because it had been generally neglected by most researchers. The computational theory of a task not only constrains the nature of the algorithm(s) for performing it, but also constrains the nature of the representation of the information at any given stage of processing. In addition, it specifies how the image is related to the outside world, by explicitly spelling out the limits on how the image can be interpreted. Knowledge of the constraints allows recovery from the image of the properties of the scene. For example, stereopsis depends on the constraint that only one point on the retina receives light from the same source as another (unique) point on the other retina, and that the changes in disparity will be small. Although there are some possible exceptions, in general this constraint holds because the world is largely composed of smooth surfaces.

A COMPUTATIONAL THEORY OF VISION: HOW DOES THE BRAIN SEE?

This theoretical framework was the 'signature' of the research conducted in the MIT Vision Group that was formed and inspired by David Marr. The group included many talented and creative students and colleagues, such as Tomaso Poggio, Shimon Ullman, Ellen Hildreth, Eric Grimson and Keith Nishihara. Together they were seeking computational insights into the working of the visual system, and they put them to the test of implementation as computer models. Within only a few years many ground-breaking papers on computational vision had been published, including a theory of

binocular stereopsis, a theory of low-level image representation, representation of direction selectivity in the cortex and a theory of the way in which shapes and actions are categorized.

Marr's book entitled *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (see Further Reading) is a lucid presentation of this work which proposes a general theory of the visual processing stages up to (but not including) object recognition. The framework of this theory is based on three main symbolic representations of the visual world which are created, maintained and interpreted by the process of vision. First, the *primal sketch* is mainly concerned with the description of changes in intensity of the image and their local geometry, on the grounds that intensity variations are likely to correspond to object boundaries or other physical realities. The primal sketch representation is constructed from symbolic primitives such as zero crossings, edges, contours and blobs. Secondly, the *two-and-a-half-dimensional sketch* ($2\frac{1}{2}$*-D sketch*) is a viewer-centered description of the relative distances, contours and orientations of surfaces. Thirdly, the *three-dimensional model* (*sketch*) is an object-centered representation of objects with the goal of later allowing manipulation and recognition. This representation must be initially related to and derived from the two-and-a-half-dimensional sketch, which means that there must be a relationship between the schema of an object and the way in which the organization of its surfaces appears to the perceiver.

Each of these representations is associated with algorithms used to produce them and computational theories describing specific modules in the visual system that are used to construct the sketches at each level. The idea of the vision process as a set of relatively independent modules is a powerful one from both computational and evolutionary perspectives, and some of the modules have been isolated experimentally.

DAVID MARR'S VISION

In the winter of 1978 David Marr was diagnosed with leukemia. He died on 17 November 1980 in Cambridge, Massachusetts. His entire work provided solid proof that in behavior and brain sciences a good theory does not have to sacrifice mathematical rigor for faithfulness to specific findings. More importantly, it emphasized the role of explanation over and above mere curve fitting, making it legitimate to ask why a particular brain process is taking place, and not merely what differential equation can describe it.

Through his published work, intellectual leadership, and the harmonious blend of insight, mathematical rigor and deep knowledge of neurobiology that characterizes his research, David Marr has given us a new intellectual landscape. More than two decades after his quest was cut short, research in neurobiology and cognitive sciences increasingly emphasizes the importance of elucidating the computations performed by the brain, and the most exciting developments are those prompted (or at least accompanied) by computational theories.

Further Reading

Marr D (1969) A theory of cerebellar cortex. *Journal of Physiology* **202**: 437–470.

Marr D (1970) A theory for cerebral neocortex. *Proceedings of the Royal Society of London* **176**: 161–234.

Marr D (1971) Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London* **262**: 23–81.

Marr D (1976) Early processing of visual information. *Philosophical Transactions of the Royal Society of London* **275**: 483–524.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman and Company.

Marr D and Poggio T (1976) Cooperative computation of stereo disparity. *Science* **194**: 283–287.

Marr D and Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London* **200**: 269–294.

Marr D and Poggio T (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London* **204**: 301–328.

Marr D and Hildreth E (1980) Theory of edge detection. *Proceedings of the Royal Society of London* **207**: 187–217.

Marr D and Vaina LM (1982) Representation and recognition of the movement of shapes. *Proceedings of the Royal Society of London* **214**: 501–524.

Vaina LM (1991) *From the Retina to the Neocortex. Selected Papers of David Marr*. Boston, MA: Birhauser, Springer Verlag.

McCulloch, Warren

Introductory article

Jerome Y Lettvin, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
Taffy Holland, Old Lyme, Connecticut, USA

Warren Sturgis McCulloch (1898–1969) was an important figure in the development of current views on the relation of brain, as mechanism, to spirit, as process. He was particularly interested in the physiology of the nervous system.

The concert of body and mind remains the major philosophical problem it has been since antiquity. But systems that act purposively, and modify their action to optimize performance, have proliferated since the mid-twentieth century. In some ways they have changed how we view the mind–body problem; and the paper 'A logical calculus of the ideas immanent in nervous activity', written by McCulloch and Pitts and published in 1943, helped make this change possible.

McCulloch was born and grew up in Orange, New Jersey. He showed early abilities in carpentry, building design, construction, blacksmithing, surveying, and sailing. He wrote poetry, and continued to do so throughout his life.

After high school he attended Haverford College for a year, where he concentrated on the philosophical problems of epistemology. He went on to Yale, where he took his BA in 1921 while in the Naval Reserve. After two years in the navy, he went on to Columbia University in New York, taking an MA in 1923 and an MD in 1927.

His internship and residency at Bellevue Hospital (1927–1928) launched his career in neurology. From 1928 to 1930 he worked on experimental epilepsy research in the Neurosurgical Laboratory and Department of Neurology at Columbia University. Then, from 1930 to 1931, he studied head injury under Foster Kennedy at Bellevue Hospital, while teaching physiological psychology at Seth Low Junior College.

From 1931 to 1932 McCulloch did graduate work in mathematical physics at New York University. What he learned during this break in his clinical work was to serve him well in his later studies in physiology.

From 1932 to 1934 he was Resident in Neurology at Rockland State Hospital (Orangeburg, New York State). In 1934 he committed himself to the study of neurophysiology at Dusser de Barenne's laboratory in Yale, where he was able to begin the main thread of his scientific life: the attempt to understand the brain.

In Holland, Dusser de Barenne had developed the technique of strychnine neuronography. By this technique it was possible to map out corticocortical connections in live mammalian brain, a prohibitively complex task by ordinary anatomical methods.

The cortex can be mapped, by cytoarchitectonic methods (such as used by Brodmann), into many distinct areas, surprisingly clearly bounded. These areas are functionally distinct, and their numbering has become the mode of reference by which clinicians and physiologists identify the locus of cortical representation of specific motor, sensory, and cognitive functions.

As his wife, Rook, said of McCulloch, his goal was set early in his life: namely, to understand man and man's understanding. His mission was not to unravel the sources of misery and madness in man but to explain man himself. To him, neurology was a necessary introduction to the philosophical questions, a tool rather than a profession.

For the next 15 years, McCulloch devoted himself to the localization of function in the brains of cats, macaques, and chimpanzees. Unfortunately, that work has largely been neglected with the development of more precise modern methods. Some of McCulloch's subtler observations about cortical function have almost been forgotten.

McCulloch and Dusser de Barenne wrote on the phenomenon of 'suppression', which they distinguished from 'inhibition' and 'extinction'. Between the motor and premotor areas of the cortex lies a strip which they called '4s'. When stimulation of this area preceded by several minutes a testing stimulus of the motor area, over a period of a few minutes, no response to the stimulation of the motor area was obtained. The area 4s coincided with that strip from which Marion Hines had obtained, by electrical stimulation, a cessation of movement and a relaxation of contracted muscles.

Suppressor strips (of which there are several in the cortex) are largely ignored, and left unexplained, in current accounts of cortical physiology. McCulloch, Snider, and Magoun associated them with activity of the reticular formation in the brain stem.

After the death of Dusser de Barenne, McCulloch stayed at Yale to finish and publish the papers describing their joint work. Then, in 1941, at the instigation of Percival Bailey and Gerhardt von Bonin (both of whom had worked with de Berenne at Yale), McCulloch became Director of the Laboratory of Basic Research at the Neuropsychiatric Institute in the University of Illinois College of Medicine. In 1945 he became Full Professor of Psychiatry and Clinical Professor of Physiology.

He and his family (Rook, his two daughters Taffy and Jean, his son David, and his adopted son George Duncan) took a house in a western suburb of Chicago. The family had a farm in Old Lyme, Connecticut, and every summer there would be scholars, artists, and scientists visiting from all over the world. There would also be the neighbors, to whom Warren and Rook were unstintingly generous. Rook was the farmer, Warren engaged in building, and guests joined in all the activities. The dam that Warren designed for the large pond was one of the few dams in the region untouched by the hurricane in 1938.

McCulloch's 12 years at the Neuropsychiatric Institute were productive. He, Garol, Bailey, Von Bonin, Roseman, Ward, Davis, and others worked on the strychnine neuronography of the cortex, finishing the effort begun at Yale.

Other research was also going on in McCulloch's laboratory. Fred and Erna Gibbs came from Boston City Hospital and set up their laboratory of encephalography. Craig Goodwin designed and maintained most of the electrical and electronic apparatus. Elwood Henneman began his studies on the spinal cord, which would occupy him for the rest of his life at Harvard Medical School.

At about the same time, Snider and Magoun at Northwestern University Medical School in Chicago were discovering the bulboreticular facilitatory and inhibitory nuclei. These large-celled groups in the brain stem act on all spinal reflexes, one greatly enhancing them, the other diminishing them to complete suppression. The whole reticular formation in the brain stem is very mysterious; it has global effects on the state of the organism, and feeds upward to the forebrain as well as downward to the spinal cord. This work opened a new view on the cortical suppressor strips, which has since fallen out of favor but is still debated.

In 1942, Walter Pitts came to the laboratory. When he visited McCulloch he was not yet 18; but he had already joined Rashevsky's group in mathematical biophysics at the University of Chicago. McCulloch instantly recognized Pitts's talents. Pitts was as well read in poetry, philosophy, and history as in mathematics and logic. The affinity between the two men was strong and obvious. Walter became part of the family, and lived with them in 1943.

McCulloch explained to Pitts the problems of studying the brain, the nature of reflexes, the physiology of synaptic connections, and the problem of making sense of it all. Turing's famous 1939 paper on the 'universal logical engine' was well known to both men. McCulloch pointed out that there was an analogy between the synaptic actions of excitation and inhibition and logical operations, and that the nervous system may be an engine performing logic. With great enthusiasm, they produced 'A logical calculus of the ideas immanent in nervous activity', which was highly influential.

Pitts met Norbert Wiener towards the end of 1945, and was invited to become Wiener's student at the Massachusetts Institute of Technology. He moved to Boston, but revisited Chicago to write a second paper with McCulloch – 'How we know universals: the perception of auditory and visual forms' – in 1947. This ambitious venture is as remarkable as the first, but less widely recognized.

By 1946, and until 1953, McCulloch was chairing the Josiah Macy Jr Foundation Meetings on Cybernetics, multidisciplinary meetings to deal with problems involved in the nature of information and its processing.

While he continued to oversee and guide the laboratory, his interests had changed. So when Jerome Wiesner, at Wiener's request, invited McCulloch to MIT in 1951, McCulloch gladly accepted. Here he could devote himself fully to a new venture with no administrative responsibilities. With him came Pat Wall and Jerry Lettvin.

At that time, the Research Laboratory of Electronics at MIT was a playground of new thought. Engineers, mathematicians, and scientists mingled with linguists and poets, artists and musicians, philosophers and critics, in a common creative effort, in a climate where ideas could be explored without the promise of results.

Two problems remained from McCulloch's years at the Neuropsychiatric Institute. First was the question of whether the brain could be regarded as (or imitated by) a logical engine – a project already started by John von Neuman and being actively pursued at MIT. The second question

concerned the nature of inhibition. Is it a specific synaptically mediated signal, or is it also mediated in other ways?

McCulloch devoted himself to the problem of developing a mode of logic for modeling neurons. He left the laboratory experimentation on nervous action to his group. However, he remained very involved in the design and prosecution of the experiments.

The laboratory studies on vision and on smell, on synaptic interactions, and on self-repair of the nervous system, all done using frogs and salamanders, are now well known. But one study, which occupied the first three years, has been almost ignored. The 1955 paper 'Reflex inhibition by dorsal root interaction' shows that electrical currents produced by the impulses in one set of nervous fibers affect the threshold for invasion of an impulse into the branchings of adjacent fibers. Every branching is a two-bit switch; the oncoming impulse in the fiber can invade one or other branch, or both, or neither. What happens at this switch depends on the electrical field produced by what is happening everywhere around it. This makes nervous processing very much more complex than had previously been thought. The implications of this work are potentially explosive.

Warren McCulloch made vital contributions to the study of the brain at a time when the field was undergoing revolutionary change. His work raises serious problems about the nature of process in living things as compared to logical engines. His pleasure in paradox spices all his writings, and his insistence on underlying metaphysics is the ghost in the living machines he studied. McCulloch was a 17th century figure in 20th century clothes.

Further Reading

McCulloch R (ed.) (1989) *The Collected Works of Warren S. McCulloch*, 4 vols. Salinas, CA: Intersystem Publications.

McCulloch WS (1965) *Embodiments of Mind*. Cambridge, MA: MIT Press.

Newell, Allen

Introductory article

Roheena Anand, Nature Publishing Group, London, UK

CONTENTS

Allen Newell (1927–1992) was a pioneer in the field of artificial intelligence. Much of the work he initiated continues to be of fundamental importance today.

INTRODUCTION

Allen Newell was one of the greatest thinkers in the field of artificial intelligence (AI). In recognition of the enormity of his achievements, he was made the first president of the American Association for Artificial Intelligence, and received one of the first awards for research excellence from the International Joint Conference on AI. In his lifetime Newell wrote and contributed to 250 publications, including ten books. His final book was entitled *Unified Theories of Cognition*, and the title represented his lifelong goal: to understand the workings of the human mind.

BEGINNINGS

Allen Newell was born in San Francisco in 1927. His father was Dr Robert R. Newell, a renowned professor of radiology at Stanford Medical School, and his mother was Jeanette Le Valley. Newell looked up to his father a great deal, seeing him as '… in many respects a complete man … He'd built a log cabin up in the mountains … He could fish, pan for gold, the whole bit. At the same time, he was the complete intellectual … Within the environment where I was raised, he was a great man. He was extremely idealistic. He used to write poetry.'

At 17, Newell, perhaps influenced by his father's outdoor pursuits, dreamt of being a forest ranger but a career in science became his main preoccupation after witnessing the atomic-bomb tests on the Bikini Atoll. Newell was 19 and had just been drafted into the US Navy. He had observed the tests on board a ship full of scientists; and his job afterwards involved making maps of the distribution of radiation over the atolls.

He completed his first degree in physics at Stanford in 1949, having already published his first paper (on the subject of X-ray optics). At Stanford, Newell took a course on mathematical methods in physical science given by the mathematician George Polya. This proved to be a significant influence on the course of his future studies – implicitly if not explicitly – by introducing him to the idea of solving problems using a method grounded in heuristic reasoning.

After a year studying pure mathematics at Princeton, Newell left to join the RAND Corporation, a think-tank based in Santa Monica. He worked for RAND for the next 11 years, though not always on site. It was here that he met Herbert A. Simon, then Professor of Industrial Administration at the Carnegie Institute of Technology (CIT), who was to play a large part in Newell's life and studies. (*See* **Simon, Herbert A.**)

At first he studied logistics systems and organizational science, specifically with reference to the military. The Systems Research Laboratory at RAND grew from Newell's first efforts (along with those of Bob Chapman, Bill Biel and John Kennedy) to study information handling and decision-making processes by military pilots for which it was necessary to construct a full-scale mock-up of an early warning station. Newell then worked with Cliff Shaw, a systems programmer, to produce a computer program to simulate a radar display of air traffic. This project brought about the realization for them that computers could be used for symbolic as well as for numerical processing.

There came a new epiphany for Newell in 1954 during a seminar by Oliver Selfridge on pattern recognition, when he realized that it would be possible to build more complex intelligent adaptive systems than had ever been built before, and importantly that they could be programmed on digital computers. From then on, his main preoccupation was the architecture of the human mind. (*See*

Pattern Recognition, Statistical; **Pattern Vision, Neural Basis of**)

BUILDING SYSTEMS

The way to learn about systems is to try to build one.

Newell's first steps towards this goal took the form of programming a computer to learn to play chess. Chess had always interested researchers in this area because of its perceived difficulty as a form of thought and its innate logic. Newell's chess program was reliant on heuristic search, and led to his first publication on this topic. However, while the design was promising, the restricted memory of computers of the time proved to be an obstacle. (*See* **Search**)

Newell's working relationship with Simon continued from there on. Instead of moving to Stanford's Behavioural Sciences unit, he joined CIT to complete a doctoral degree and work on simulations. Newell, Simon and Clifford Shaw joined forces in 1955 to build a computer program that would demonstrate complex information processing. They started to construct a system with the aim of working in geometry, but changed direction to work on propositional logic (this decision was due in part to Simon's owning a copy of *Principia Mathematica*). From this collaboration came 'logic theorist' (LT), one of the first working AI programs. It incorporated many of the ideas that have since become foundations of this topic, one of the most fundamental being that of heuristic search.

While building LT and their chess-playing programs, the team were faced with the problem that none of the current computer languages supported the symbolic processing that was so central to their programs; therefore they had to spend time building their own tools. In 1957, there emerged from these efforts the first implemented list-processing language, called simply 'information-processing language' (IPL). Over the next seven years, six different versions of this language were developed. The language introduced many of the ideas that became fundamental first to list processing and later to computer science in general: lists, associations, schemata, dynamic memory allocation, data types, recursion, associative retrieval, functions as arguments, and generators.

Newell received his Ph.D. from CIT in 1957. His thesis described his work on LT and chess. In it were ideas that would appear 25 years later in Soar (originally standing for 'state, operator and result'): dynamic subgoal creation, multiple evaluation, methods for controlling search, resource-limited reasoning, learning based on applying existing knowledge, and the close integration of learning and problem solving. (*See* **Soar**; **Problem Solving**).

Four years later, in 1961, Newell would join the faculty as a professor. He was instrumental in the creation of CIT's computer science department and its subsequent development into one of the best departments of its kind in the world.

After their work on LT, the team returned briefly to the realm of chess, and, using a version of IPL, wrote a chess-playing program called NSS (after the initials of the authors). This differed from others around at the time in that it aimed to simulate human players rather than just go for the win.

LT as a system exhibited intelligent behavior, but in a limited domain: all its methods and knowledge were specific to propositional logic. The next major step, rather than building more task-specific systems, was the generalization of these basic techniques to construct a single system that would model human behavior across many different domains. The 'general problem solver' (GPS) was thus conceived. This was a system that, in Newell's words, 'separated out the program structure for problem solving from the program structure to describe a particular task'.

GPS was constructed by studying human protocols for logic tasks. It was apparent from this that they required a more goal-directed strategy: means–ends analysis, where actions are selected based on their ability to achieve a goal. Thinking-aloud protocol analysis was used (and indeed resurrected, since it was then generally out of favor), and problem-behavior graphs tracked the performance of subjects, coding decisions that were being made and comparing them to those made by the computer.

However, there were problems with GPS. In Simon's words, it would 'burrow into a deep pit of successive subgoals with no way for the top program levels to regain control'. A potential solution lay in the use of production-system languages, where each instruction was a condition followed by an action. This and other aspects of GPS would eventually be generalized in Soar, the problem-solving architecture based on problem spaces and production systems that was developed by Newell, John Laird and Paul Rosenbloom.

In 1972 Newell and Simon published *Human Problem Solving*, the culmination of their investigations into complex problem solving. It covered protocol analysis, GPS, and production systems, presenting a computational theory of human problem solving based on heuristic search in problem spaces.

Simon and Newell's research paths diverged slightly after the writing of this book, with Simon in his own words, then concentrating on 'GPS, EPAM the sequence extrapolator and BACON, while Newell's emphasis was on computational architecture and attempts to model the control structure underlying intelligence'.

THEORIES OF MIND

> Diversions occur, make them count: salvage what is possible for the main goal.

In 1968, Newell, with Gordon Bell, wrote a book on computer systems, in the process of which they created languages for two different levels of computer design: the system level (PMS) and the instruction level (ISP). This work was a diversion from Newell's main preoccupation with intelligence, but it served to crystallize ideas of architectures and hierarchies of levels for analyzing computational systems.

Other work on computer and software systems design followed, along with a number of publications, up to 1982. The L* language was developed by Newell, George Robertson, Peter Freeman and Don McCracken with the system programmer in mind in that it was meant to facilitate the construction of a customized operating system and user interface.

Newell was also willing to observe rather than participate, and played an advisory role in the ARPA program of research on speech recognition in the 1970s. He wrote the final report for this, which proved to be very influential at the time.

In 1973 Newell became a consultant to Xerox PARC (dedicated to exploring digital electronic technologies in the context of office information systems). A year later he was joined by two of his former students (Stuart Card and Thomas Moran), thereby forming the Applied Information-Processing Psychology Project, part of his long-term project to apply psychological theory to human–computer interaction (or user-interface design). Existing psychological data were examined for regularities, and from this analysis an engineering-level model of routine cognitive skills, and a methodology for analyzing new tasks in terms of the basic processes required to perform them, were constructed. This work resulted in the publication of *The Psychology of Human–Computer Interaction*, and led Newell back to human mental architecture.

> An architecture is a fixed set of mechanisms that enable the acquisition and use of content in a memory to guide behavior in pursuit of goals. In effect, this is the hardware–software distinction: the architecture is the hardware that is supporting the software and the software is the collection of data structures that encode the content. This is the essence of the computational theory of mind.

SOAR

> Choose a final project to outlast you.

A unified theory of cognition is usually based around one central cognitive activity – with Soar (and previously GPS), this was problem solving. Soar allowed multiple problem spaces to be used in solving a single problem, thereby overcoming some of the constraints of the GPS system. Soar was a production system to which learning by chunking and a universal weak method were added. The concept of chunking had existed in previous AI programs where memory organization was studied in terms of chunks, and in learning by adaptive production systems.

Soar represents all long-term knowledge as productions and all short-term knowledge as attribute values. Problem solving is formulated as processing within problem spaces; all goals are generated from architectural 'impasses'; and all learning occurs via chunking.

Soar was to become the central focus of Newell's research for the rest of his life, and it quickly became clear that it had many of the properties required to model human cognition, for the above reasons. This realization prompted Newell to propose Soar as the basis for a unified theory of cognition. He did not mean it to be the fundamental 'theory of theories', but rather a vehicle for him to demonstrate and explore what a unified theory would look like.

Newell's lectures on Soar at Harvard in the late 1980s prompted worldwide research on the topic as a theory of cognition, and Newell took an active role in research on cognitive development, natural language, instruction taking, visual attention, human–computer interaction, and syllogistic reasoning. (*See* **Natural Language Processing**; **Cognitive Development, Computational Models of**; **Visual Attention**; **Human–Computer Interaction**)

NEWELL'S VISION

> What lives is what your scientific descendants must use to solve their problems.

Work still continues today on Soar, but Newell's contribution to the field of AI is broader – indeed, without him, it would not be what it is today. Herb

Simon, himself a giant in the field, had the following to say about Newell:

> He was a person who not only dreamt but gave body to his dream, brought it to life. He had a vision of what human thinking is. He spent his life enlarging that vision, shaping it, materializing it in a sequence of computer programs that exhibited the very intelligence they explained.

Further Reading

Bell CG and Newell A (1971) *Computer Structures: Readings and Examples*. New York, NY: McGraw-Hill.

Card S, Moran TP and Newell A (1983) *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

Chapman RL, Kennedy JL, Newell A and Biel WC (1959) The systems research laboratory's air defense experiments. *Management Science* **5**: 250–269.

Laird JE and Rosenbloom PS (1992) In pursuit of mind: the research of Allen Newell. *AI Magazine* **13**(4): 17–45.

Newell A (1955) The chess machine: an example of dealing with a complex task by adaptation. In: *Proceedings of the 1955 Western Joint Computer Conference*, pp. 101–108. New York, NY: Institute of Radio Engineers. [Also issued as Rand Technical Report P-620.]

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A, Shaw JC and Simon HA (1958) Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development* **2**: 320–325.

Newell A, Shaw JC and Simon HA (1960) Report on a general problem solving program. In: *Proceedings of the International Conference on Information Processing*, pp. 256–264. Paris: UNESCO.

Newell A and Simon HA (1956) The logic theory machine: a complex information processing system. *IRE Transactions on Information Theory* **IT-2**: 61–79.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Rosenbloom PS, Laird JE and Newell A (eds) (1993) *The Soar Papers: Research on Integrated Intelligence*. Cambridge, MA: MIT Press.

Simon HA (1997) Allen Newell: a biographical memoir. *Biographical Memoirs*, vol. 71. Washington, DC: National Academy of Sciences.

Simon HA (1998) Allen Newell 1927–1992. *IEEE Annals of the History of Computing* **20**(2): 63–76.

Olds, James

Introductory article

Phillip J Best, Miami University, Oxford, Ohio, USA

The discovery by James Olds (1922–1976) of reward systems in the brain was an important advance in our understanding of brain–behavior relationships. Olds later developed new techniques for recording the activity of individual brain cells in freely behaving animals, and used these techniques to study brain activity during learning.

After graduating from Amherst College in 1947, Olds went to Harvard University to study for a Ph. D. in Social Psychology under Talcott Parsons. At Harvard he became interested in animal learning and motivation, and brain mechanisms of behavior. The prevailing view at that time was that of the behavioral theorists like Clarke Hull and B. F. Skinner, who believed that learning occurs because the immediate consequences of a response reinforce a connection between that response and the preceding stimulus. According to Hull's 'drive reduction theory', physiological deficiencies, or needs, produce drives. The responses the animal makes that result in reduction of these drives are reinforced. Responses that do not reduce drives are not reinforced, and would therefore occur less frequently in the future. Responses that result in increased internal drives should result in negative reinforcement, or 'punishment'; and their frequency should decrease rapidly and significantly.

According to this point of view, behavioral theory had no need to hypothesize any internal, 'mental or emotional' processes like thinking or feeling. In fact, such terms were eschewed. Indeed, the behaviorist saw no need to investigate the brain to understand behavioral processes.

An alternative view to strict behaviorism was held by Edwin Tolman, who placed more emphasis on the information processing that should result from experiencing an important event. He also emphasized the significance of 'internal' processes in explaining behavior. Tolman was a visiting professor at Harvard during Olds's graduate career and profoundly influenced his thinking. Olds also liked to think about the brain in his attempts to explain the results of his behavioral experiments.

Olds was also strongly influenced by the writings of Donald O. Hebb, who was the most prominent biological psychologist in North America. One of his professors at Harvard, Richard Solomon, suggested that if Olds wanted so much to study the brain, he should go to Hebb's laboratory at McGill University.

When Olds arrived at Hebb's laboratory, he was assigned to work with a graduate student named Peter Milner on a project to investigate the behavioral effects of electrical stimulation of an animal's brain. Until that time, the most common method for studying brain–behavior relationships had been to examine the behavioral effects of brain damage. Most knowledge of brain function had been learned by inducing lesions in selected brain structures and carefully examining the consequences of the damage on specific behavioral processes: for example, sensory and motor deficits, or changes in arousal, motivation, learning, or memory.

A small number of studies had used electrical stimulation of the brain to study brain–behavior relationships. William Randolph Hess had just won a Nobel Prize for studies showing that electrical stimulation of a certain part of the brain, the hypothalamus, induces aggressive attack behavior. Another group of scientists had discovered that stimulation of another part of the brain, the midbrain reticular formation, induced arousal in sleeping animals. And a group at Yale had reported that stimulation of certain parts of the brain produced fear-like behavior.

Recognizing the value of these new electrical stimulation techniques, researchers at Hebb's laboratory began to study the effects of brain stimulation on behavior. When Olds arrived at McGill, Peter Milner and another student, Seth Sharpless, were studying the arousing effects of electrical stimulation of the midbrain reticular formation. Milner taught Olds how to implant the electrodes in rats.

While testing his first animal, with an electrode aimed at its midbrain reticular formation, Olds noticed that the animal did not avoid the places on the testing table where it had just received stimulation. In fact, the stimulus seemed to reinforce the responses that brought the animal back to those places where it had just been stimulated. Olds was able to use the stimulation to reinforce the animal for moving to any location on the testing

table. He realized that the brain stimulation possessed all of the positive reinforcing properties of natural rewards such as food or water to a hungry or thirsty rat: the animal behaved as if it 'enjoyed' the electrical brain stimulation, rather than finding it aversive.

Later, Olds discovered that the electrode was not after all in the midbrain reticular formation. It was a few millimeters away, in an area of the forebrain known as the septal area. In order to investigate this phenomenon objectively, Olds and Milner implanted other animals, and showed that electrical brain stimulation to the septal area acted as a positive reinforcer for lever-pressing behavior. Their study aroused such interest that it became the most cited paper in the field of psychology for the next two decades.

Soon thereafter, Olds and others demonstrated that electrical stimulation in various areas of the forebrain, notably areas related to the hypothalamus, induced a variety of consummatory behaviors, such as eating, drinking, sex, nest building, and parental behavior. Further, whenever a stimulus was found to elicit consummatory behavior, that stimulus was found to be rewarding. So, when a brain stimulus increased the drive for a consummatory behavior, it was also positively reinforcing. This result contradicted the fundamental premise of drive reduction theory, that reinforcement occurs only when drives are reduced. The results of brain stimulation experiments were so dramatic, and so counter to the prevailing behaviorist tradition, that they led to a flurry of research that produced a paradigm shift in the study of the biological mechanisms of behavior. Thereafter, brain stimulation became a standard method for investigating brain–behavior relationships.

After leaving Hebb's laboratory at McGill in 1955, Olds became a research associate in the laboratory of H. W. Magoun and D. B. Lindsley at UCLA. In 1958 he moved to the University of Michigan, where he and his wife Marianna (Nicky) and their students addressed a number of important questions about the nature of rewarding brain stimulation and its implications for motivation and reinforcement. They mapped the locations of the brain where electrical brain stimulation produced different degrees of reward, and showed that the reward produced by electrical stimulation could far exceed the reward produced by natural biological motivation. The rats would even cross an electrified grid floor to receive the stimulation, or starve themselves in the presence of abundant food when presented with a lever whose depression would result in rewarding brain stimulation.

Olds's group studied the relationship between rewarding brain stimulation, various hormonal conditions, and natural biological motivational behaviors, such as feeding, drinking, and sex. They found significant interactions between the motivational state of the animal and the rewarding effects of electrical stimulation. Stimulation produced the greatest reward in the lateral hypothalamus, a region traversed by the medial forebrain bundle. Lesions in this region had previously been shown to interfere with consummatory behaviors, such as feeding, drinking, sex, and nest building. Stimulation of the sites that produced consummatory behaviors was also generally rewarding.

This research program radically changed the way psychologists thought about motivation, reward, and reinforcement. It has also had a major influence on the study of drug addiction. Many of the areas of the brain where stimulation was found to be especially intense are now known to be the areas where addictive substances produce their highly motivating effects.

In the early 1960s Olds began to consider the limitations of the use of electrical and chemical brain stimulation in the study of brain–behavior relationships. He sought new techniques for recording the activity of individual brain cells in awake, freely behaving animals. Techniques had been developed to record the activity of individual neurons in invertebrate preparations, and some progress had been made in recording from individual brain cells in anesthetized or immobilized vertebrate animals. However, neither of these techniques was suited for recording for very long periods of time, or from individual brain cells in awake freely behaving animals. The only brain recording techniques being used at that time in freely behaving animals were the electroencephalography and multiple-unit techniques which recorded the activity of small clumps of cells but could not distinguish individual cells. Neither of these techniques provided enough resolution for Olds's purpose.

Olds eventually learned that large, blunt, soft, wire electrodes, could record from individual neurons in awake, freely behaving animals for reasonable periods of time. The task was formidable, because the cellular signals were of very low amplitude, the background noise level was generally very high, it was often difficult to distinguish the activity of one brain cell from another, and behavioral movements would often produce electrical signals in the recording wires that were difficult to distinguish from neuronal activity. Olds patiently solved each of these problems by

designing special electronic circuits to analyze the signals and to discriminate between true nerve signals and artifacts. He was one of the first to introduce general-purpose computers into the laboratory to perform real-time, online signal analysis.

His original intention was to use single-cell recording techniques to study brain mechanisms of reinforcement. Since graduate school, he had been interested in the idea that some change in the brain is responsible for changing the likelihood of a response or for increasing the strength of connections between a stimulus and a response or between two stimuli. While searching for the site, or sites, where reinforcement occurred, his early recording studies included the study of motivation and states of arousal. While his laboratory produced an impressive set of studies of cellular changes in different brain areas during changes in motivation and arousal, he was not able to identify the sites where reinforcement occurred. Since his behavioral paradigms typically included a learning component, he decided to investigate regional differences in cellular activity during learning per se. He was the first to show changes in single brain-cell activity during learning in freely behaving animals.

In 1969 Olds moved to the California Institute of Technology, where his research focused on isolating those parts of the brain that were critical for learning. There was particular interest in the hippocampus, because lesions there in humans produced severe memory deficits. Olds's laboratory was the first to report reliable changes in activity in hippocampal neurons during Pavlovian conditioning. He embarked on a large program to study the activity of many areas of the brain simultaneously during Pavlovian conditioning to determine which areas changed the most, which changed the earliest during learning, and which showed the shortest latency response to the stimuli. This program was very successful, making a number of important discoveries, including the discovery that the earliest conditioned changes in brain activity often occurred in sensory pathways.

Before his program had come to complete fruition, James Olds died, apparently of a heart attack, while swimming off Newport Beach, California, in 1976 at the age of 54.

Olds received many awards and honors during his career, including the Newcombe Cleveland Prize of the American Association for the Advancement of Science in 1956, the Hofheimer award from the American Psychiatric Association in 1958, the Warren Medal from the Society of Experimental Psychologists in 1962, the Distinguished Scientist Award from the American Psychological Association in 1967, and the Kittay Prize from the Kittay Scientific Foundation in 1976. He was elected to the National Academy of Science. If his career had not been cut so short he might well have received a Nobel Prize.

Further Reading

Milner P (1989) The discovery of self-stimulation and other stories. *Neuroscience and Biobehavioral Reviews* **13**: 61–67. [A personal account of the atmosphere in Hebb's laboratory at the time of Olds's arrival.]

Olds J (1975) Mapping the mind unto the brain. In: Worden FG, Swazey JP and Adelman G (eds) *The Neurosciences: Paths of Discovery*, pp. 375–400. Cambridge, MA: MIT Press. [An autobiographical account of Olds's career and the development of his ideas.]

Olds J (1977) *Drives and Reinforcements: Behavioral Studies of Hypothalamic Functions*. New York, NY: Raven Press. [A comprehensive statement of Olds's ideas concerning the brain mechanisms of motivation, reinforcement, and reward.]

Olds J (1980) Thoughts on cerebral functions: the cortex as an action system. In: Routtenberg A (ed.) *Biology of Reinforcement*, pp. 149–167. New York, NY: Academic Press. [A posthumously published article expressing Olds's later view of brain function.]

Olds J and Milner P (1954) Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–427. [The original article on the reinforcing effect of brain stimulation.]

Pavlov, Ivan Petrovich

Introductory article

Richard Bowen, Colorado State University, Fort Collins, Colorado, USA

CONTENTS

Ivan Petrovich Pavlov (1849–1936) was a Russian physiologist most famous for his research on the conditioned response. His work provided a firm experimental grounding for many of the concepts currently accepted in psychology.

INTRODUCTION

Ivan Petrovich Pavlov (1849–1936) was a Russian physiologist best known for his groundbreaking studies on 'conditioned responses'. Less well known are his contributions to other aspects of physiology, particularly those dealing with digestive functions. Indeed, it was while studying digestive secretions that Pavlov became interested in the integration of brain and body, which ultimately occupied the bulk of his career. In 1897 Pavlov published his experimental results and generalizations in a book called *Work of the Digestive Glands*, which included defining the role of the brain in control of digestive function. For this work, he became the first Russian and the first physiologist to receive the Nobel prize, which was awarded in 1904.

Pavlov was born in the Russian town of Ryazan. His father was a priest and, as was customary in such situations, when Pavlov came of age, he began studying for the priesthood. While at the seminary he was exposed to texts on physiology and to Darwin's work on evolution, which led him in 1870 to abandon his religious training and enroll in the University of St Petersburg. His first research project as a student involved investigation of pancreatic nerves. Pavlov continued his studies at the Military Medical Academy between the years of 1875 and 1879. He finished his dissertation and earned the degree of doctor of medicine in 1883. Pavlov gained the attention of such prominent researchers as Ludwig, Heidenhain, and Bofkin during the next several years, and was appointed a professor at the St Petersburg Institute of Experimental Medicine in 1895.

PAVLOV AS A PHYSIOLOGIST

Early Career

Pavlov's first independent work focused on the physiology of blood circulation, which began with studies on the influence of variations in blood volume on blood pressure. He also investigated the nervous control of the heart, and argued that four types of nerves controlled the rhythm and strength of cardiac contractions. He came to believe that in order to understand the true physiological mechanisms of an organ, that organ had to be observed as it functioned as a part of the whole body. Towards that end, Pavlov began to use unanesthetized, neurologically intact dogs for his studies, an experimental paradigm that would become the mainstay of his most important accomplishments. Pavlov's methodology involved training dogs to lie calmly on the operating table while he incised the skin and surface tissues, exposed the artery, and connected it to instruments for measuring blood pressure.

Pavlov's second independent work concerned digestive physiology. He started studying digestive processes as early as 1879, and it was his major focus from 1890 to 1897. The bulk of this work also was based on experimentation with dogs, and involved developing fistulas through which secretions from the salivary glands, stomach, pancreas and small intestine could be collected over time. His technique was truly unique in that he did not cut the nerve supply nor contaminate the secretions with food. Among other things, these experiments led to a description of the neural control of pancreatic secretion, the demonstration that chewing and swallowing alone stimulated gastric secretion, and the finding that the types and amounts of secretions from the stomach varied in response to different foods. Finally, and importantly, Pavlov observed that the mere sight of food stimulated secretion from both salivary glands and stomach.

The Salivary Reflex as a Window to the Brain

Pavlov's initial work on digestive secretions confirmed what others had noted: that introduction of food or mild irritants into the mouth led rapidly to the secretion of saliva. He and his students (the 'Pavlovians') subsequently observed that their dogs would also secrete saliva when food was near enough for them to see or smell it. Moreover, the same pattern of salivation was seen when the dogs were shown their empty food bowl, when the person who fed them entered the room, or even when they heard footsteps approaching their pens. Clearly, the cortical regions of the brain were directing such responses, and Pavlov recognized that such 'psychic secretion' provided a unique opportunity to study neural processing in the brain. He felt strongly that understanding higher brain function required a purely physiologic approach, and exhibited substantial disdain for studies by the animal psychologists of his day. In promoting his use of the salivary reflex, Pavlov stated:

> In this manner the investigation of the cerebral hemispheres is brought into line with the investigations conducted in other branches of natural science, and their activities are studied as purely physiological facts, without any need to resort to fantastic speculations as to the existence of any possible subjective state in the animal which may be conjectured on analogy with ourselves.

Investigations on Conditioned Reflexes

Between 1897 and 1936 Pavlov and his associates published more than five hundred papers dealing with conditioning of the salivary reflex. This massive body of work was an extension of his early experience in collecting saliva from conscious dogs in which a salivary duct was surgically diverted to the exterior surface of the skin and the saliva was collected into tubing. The chief advantage of this method over other approaches of the day was that it provided an objective and easily quantifiable end point – Pavlov could measure the time between stimulus and response, and (as volume) the magnitude of the response. Additionally, the stimuli applied to elicit the salivary reflex were of the type encountered every day and thus more likely to reveal normal function than stimuli such as electrical shocks, employed by some other investigators.

Recognizing the necessity of isolating his dogs from all but the stimulus under study, Pavlov arranged for the construction of a building dedicated to his investigations. The animal rooms within this building were soundproofed and allowed investigators to remain hidden behind a wall while collecting saliva, without the dogs being aware of their presence.

Introduction of bread, sand or an acidic solution into the mouth induces secretion of saliva. This is an example of an unconditioned reflex. The sight of food also will induce secretion of saliva; this is called a conditioned reflex or conditioned response because it requires experience. The Pavlovians demonstrated that conventionally raised dogs salivated in response to the sight of bread, but that dogs raised exclusively on milk did not salivate when they saw bread, until they learned to associate bread with something to eat. Development of a conditioned reflex is thus an example of associative learning.

Pavlov and his students performed hundreds of experiments to delineate the prerequisites and characteristics of the conditioned salivary reflex. Most famously, they conditioned dogs so that when a neutral stimulus such as the sound of a beating metronome was allowed to reach the animal, secretion of saliva would commence within a few seconds. In addition to salivating, the dog would also turn its head in the direction from which food was usually presented and begin to lick its lips vigorously, but these responses were difficult to quantify. Eliciting such a conditioned reflex was found to depend on several factors. First, the dog had to be hungry; even after conditioning, a satiated dog failed to secrete saliva in response to the stimulus. Second, the neutral stimulus had to be applied in the correct temporal sequence with respect to the true stimulus. That is, if the metronome (or bell, light or nonfood odor) was presented simultaneously with feeding or after feeding, it subsequently failed to elicit the conditioned reflex. This was demonstrated in numerous experiments. As one example, a loud buzzer was sounded 5–10 s after presentation of food on 374 successive occasions, but in none of these trials did the dog develop a conditioned salivary reflex to the buzzer. In contrast, applying the buzzer shortly before the meal on a single occasion was reported to successfully establish a conditioned reflex. Finally, a conditioned reflex can be lost if the neutral stimulus is applied a number of times without being followed by stimulating the unconditioned reflex (i.e. if the buzzer is sounded several times without presentation of food, the buzzer will ultimately fail to stimulate flow of saliva). Such loss of a conditioned response is called *extinction*.

One of Pavlov's most intriguing demonstrations of conditioned reflexes involved the use of a noxious instead of neutral stimulus. Unconditioned dogs respond to the prick of a needle by withdrawal, as one would expect. Pavlov demonstrated that dogs conditioned to a needle prick followed by food not only developed a normal conditioned salivary response, but did so without aversion to the needle prick.

CONTRIBUTIONS TO PSYCHOLOGY AND NEURAL SCIENCE

Pavlov began his work on conditioned reflexes at a time when much attention was focused on understanding brain function, with little progress to show for it. This period also coincided with the beginnings of psychology as a separate discipline, and psychologists frequently constructed elaborate theories of mental function without constraint by or support from experimental observations.

Signalization: Establishing a Conditioned Reflex

The concept of a reflex was recognized well before Pavlov began his work, but he was the first to apply rigorous experimental analysis to its understanding. He considered that unconditioned reflexes were the result of inborn connections in the nervous system, which enabled an innate coupling between stimulus and response. Complex unconditioned reflexes were equated to instincts. Conditioned reflexes, in contrast, required establishment of new connections within the brain; another name Pavlov used for these was 'acquired reflexes'.

Pavlov's greatest contribution was in developing the concept of *signalization*, and proposing its role in behavior and survival. Conditioned reflexes are established within cortical regions of the brain as alternative pathways to invoke preexisting unconditioned reflexes. An example of signalization in Pavlov's dogs was when the sound of a buzzer was sensed, propagated through the auditory nerve to the cortex, then relayed to nerves innervating the salivary glands, leading to secretion of saliva. A connection was thereby established between the environment and the unconditioned salivary reflex. Pavlov referred to signals from the environment as 'first signals', to distinguish them from language ('second signals'), which he considered a special case of great importance to humans. He explained signalization using a telephonic analogy:

> My residence may be connected directly with the laboratory by a private line; and I may call up the laboratory whenever it pleases me to do so; or on the other hand, a connection may have to be made through the central exchange. But the result in both cases is the same. The only point of distinction between the methods is that the private line provides a permanent and readily available cable, while the other line necessitates a preliminary central connection being established. In one case the communicating wire is always complete, in the other case a small addition must be made to the wire at the central exchange.

Unconditioned reflexes, particularly in chains, lead to behaviors. Signalization and conditioned reflexes allow sophisticated behaviors, including habits, to develop. By careful scientific study of the simple salivary reflex, Pavlov thus contributed a crucial insight to psychology.

A Darwinian View of Signalization and Conditioned Reflexes

Pavlov was influenced deeply by Darwinism and recognized that innate, unconditioned reflexes were critical for survival in any environment. The unconditioned reflex of withdrawing from fire rather than being attracted to it clearly contributes to fitness. However, because the environment is constantly changing, fitness can be greatly enhanced if organisms are endowed with the ability to form temporary connections in the brain that allow alternative pathways to invoke innate reflexes. Such temporary connections are, of course, the embodiment of Pavlov's signalization theory. Smelling smoke or seeing the glow of an approaching fire act essentially as neutral stimuli to condition animals to avoid fire. All in all, conditioned reflexes are of great adaptive significance, allowing animals to avoid predators, seek out mates, and find food.

CONCLUSION

Ivan Petrovich Pavlov was a self-proclaimed physiologist whose work ultimately had the most profound effects on psychology. Through intense study of salivary reflex conditioning, he and his students provided a firm experimental grounding for many of the concepts familiar to psychologists and neural scientists today and led the way toward scientific study of the brain.

Further Reading

Pavlov IP (1961) Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex [originally published 1926]. In: Thorne Shipley (ed.), *Classics in Psychology*, pp. 756–797. New York, NY: Philosophical Library.

Wells HK (1956) *Ivan P. Pavlov. Toward a Scientific Psychology and Psychiatry*. New York, NY: International Publishers.

Windholz G (1992) Pavlov's conceptualization of learning. *American Journal of Psychology* **105**: 51–67.

Windholz G (1997) Ivan P. Pavlov, an overview of his life and psychological work. *American Psychologist* **52**: 941–946.

Penfield, Wilder

Introductory article

Louise Fabiani, Montreal, Quebec, Canada

CONTENTS

Wilder G. Penfield (1891–1976), the founder of the Montreal Neurological Institute, possessed a rare talent for combining science and medicine in the study of the human brain. This approach made many advances possible, including the mapping of the human cerebral cortex, improvements in neurosurgical technique, and developments in the treatment of epilepsy and other neurological diseases.

BEGINNINGS

Wilder Graves Penfield was born in January 1891 in Spokane, Washington, USA. His father, Charles Penfield, was a physician at the time of Wilder's birth. However, a few years later he gave up his practice to answer 'the call of the wild'. The result was a separation from Wilder, his brother Herbert, his sister Ruth and their mother Jennie (née Jean Jefferson). In 1899, Charles Penfield moved to California to live in the woods, by hunting and fishing, while the rest of the family moved to Wisconsin, where Jennie's family lived.

Before long, both older siblings moved away to marry, leaving Wilder and his mother. The two were very close, and both attached a high value to religion, family, hard work and the greater good.

When Wilder was still a boy, Jennie encouraged him to aim for a Rhodes Scholarship. After graduating with a philosophy degree from Princeton University in 1913 with excellent grades and a fine record as a football player, Wilder did win the coveted prize on the second attempt. However, the First World War disrupted his plans and he did not commence his studies in Oxford, England, until 1915, after a brief stint at Harvard University and some volunteer work in the medical corps in France. Following many bouts of indecision, he eventually settled on medicine, the profession that had been practised by his grandfather and abandoned by his father.

The ship on which Wilder sailed back to England in 1915 was torpedoed by the Germans. His leg was shattered, and during the long period that it took to heal, he convalesced in the home of the famous Canadian physician, Sir William Osler, who was living in Oxford. Osler was considered to be the first Canadian neurologist, and he became one of Wilder's first mentors.

In 1917, Wilder Penfield married Helen K. Kermott, who was herself the daughter of a physician. They had four children: Wilder Jr, born in 1918, Ruthmary, born in 1919, Priscilla, born in 1926 and Amos Jefferson ('Jeff'), born in 1927.

After finishing his medical training in 1918 at Johns Hopkins University in Baltimore, Maryland, Wilder decided on a career in surgery. As an intern at the Peter Bent Brigham Hospital in Boston, Massachusetts, he focused on neurosurgery. His first job was in New York City.

Although he was to spend decades in medicine, Penfield devoted those early years to basic science – neurocytology and neuropathology. Curious about the most fundamental causes of epilepsy, he traveled to Spain and Germany, where he studied tissue-staining techniques with renowned neuroscientists. From those early experiences he retained an appreciation of the discipline of scientific research, as well as the spirit of collaboration. He was to become that rare kind of generalist – a physician who carefully observed and gathered data, a scientist who studied disease but who never forgot about human suffering.

A STRONG WILL AND A GENEROUS SPIRIT

In 1924, while working in New York, Wilder Penfield met William Cone, a fellow surgeon. The two were to become close friends and collaborators for 36 years, often performing the same operation. When Penfield and his family moved to Montreal in 1928 at the invitation of Edward Archibald, the director of the Royal Victoria Hospital (RVH), Cone went with him.

At the RVH, Penfield hoped to realize his dream of an institute where medical practice and research could go hand in hand. He discussed the concept with Archibald, and began to contact various philanthropists for building and maintenance funds. The combination of surgical practice, spending time with his growing family, and all that fundraising meant that he was a very busy man.

In 1934, the Montreal Neurological Institute and Hospital (MNI) officially opened. Penfield was its first director. At the MNI, those who worked with patient care communicated with those in research, who were in turn interdependent. Not only did this make sense from a strictly practical viewpoint (as the study of the brain is one of the most complex fields of research in the world, those who are working on a specific aspect of the subject depend heavily on each other), but also cooperation was good for morale. Penfield's strength of will, generous spirit and collaborative idealism energized the institute from the very beginning. Very soon, news of the MNI's successes spread around the world. Patients – and hopeful neurologists – flocked to Montreal.

Penfield would be neither slowed down nor dissuaded in the pursuit of his vision. In the building itself, which grew from a sketch he had made on scrap paper, his mark was everywhere. Careful floor specifications accounted for research facilities, wards and even the placement of the research-animal rooms (on the upper floors, with access to light and fresh air). An art lover, he also chose the Art Nouveau look of the lobby. The ceiling, which was decorated with pictures of ganglia, featured a ram's head (the sign of Aries is associated with the brain) encircled with Egyptian hieroglyphs. The cast-iron lampstands resembled spinal columns. The names of famous neuroscientists were inscribed along the top edge of the walls, and the centerpiece was an impressive white marble statue. (Years later, when a new wing opened in 1951, Penfield commissioned one of his operating nurses, Mary Filer, to paint a mural called 'The Advance of Neurology', which depicted Penfield, Osler, Hippocrates and other luminaries.)

One of Penfield's earliest and most difficult operations at the MNI was on his own sister. Ruth Inglis had suffered from headaches and 'spells' (epileptic events) since her teens. Decades later, the pain became unbearable. It was found that she had a brain tumor. Unless it was immediately removed, it would cause blindness and then death. Penfield was forced to break the medical taboo against treating close relatives, and in fact there was no better team for performing the dangerous operation. Penfield and Cone had to cut out part of Ruth's frontal lobe before they reached the tumor. To their amazement, Ruth lived – and with little impairment. (However, her brother did notice the signs of frontal lobe damage. For example, she no longer seemed able to plan and coordinate household duties.) Although the tumor eventually killed her, Penfield's risky procedure gave her an extra two years of life. (*See* **Frontal Cortex**; **Epilepsy**)

THE 'LITTLE MAN'

Researchers working on animals in the nineteenth and early twentieth centuries discovered that if they stimulated the open brain with a mild electrical current, a limb moved, for example, on the opposite side of the body. That is, a right-brain stimulation led to a left-body movement, and vice versa. Later, neurosurgeons applied the same technique in conscious patients (the brain itself cannot feel pain). In his search for epileptic foci (i.e. specific points in the brain that would cause seizures), Penfield began to detect patterns in patients' reactions to particular types of stimulation.

Penfield's 'Montreal procedure' came to associate function with location (functional neuroanatomy), depending on which side of the central sulcus was being tested. When the pre-central gyrus was stimulated with a light electrical current, Penfield detected a muscle twitch in a particular site on the opposite side of the body – for example, the big toe moved. On the post-central gyrus, a similar current provoked sensation, such as tingling, again in a specific site. Many patients and many painstaking trials later, Penfield and his associate, Theodore Rasmussen, were able to draw the body part most commonly associated with each stimulated point in the cortex. More sensitive areas, such as the thumb and lips, were disproportionately large (i.e. compared with the representations from less sensitive areas.) Linked together, the drawings formed two representations of the human body (motor and sensory). Each roughly drawn 'little man' was a homunculus – the brain's map of the body surface. (*See* **Cortical Map Formation**)

Penfield was forever indebted to the brave patients who responded to his tests. The resulting map was indispensable for neurosurgeons, as it enabled them to operate with minimal risk of damaging the primary motor cortex.

MORALS AND MIND

In the early 1950s, with his first career winding down, Penfield became yet another of a long line

of physicians to try his hand at fiction. He eventually published two novels. In 1960, he handed over the directorship of the MNI to Theodore Rasmussen and embarked on a second career of writing and travel. With family friends, Georges Vanier, the Governor-General of Canada, and his wife Pauline, Penfield co-founded the Vanier Institute of the Family, which was set up as a research and funding agency for the study and promotion of the solid values with which he was raised – values that seemed to be in jeopardy during the permissive 1960s. Penfield also accepted invitations to be a Canadian goodwill ambassador to many countries around the world.

Just before his death in April 1976, he completed two last books of non-fiction. One of them, *The Mystery of the Mind*, examines possible explanations for the relationship between mind and brain. Penfield had observed the numerous ways in which the mind – personality and behavior – could be radically altered by brain injury, surgery or disease (e.g. his sister's postoperative changes). Philosophically, he was a dualist – he regarded the mind and the brain as separate, although related. Decades of neurological study jostled with even older beliefs. At the contemplative end of his life, the spiritual won over the material. He concluded that the greatness of the human mind and the mystery of the soul could not be reduced to the activity of neuronal tissue – a thought-provoking opinion from one of the most famous neurosurgeons ever to have lived. (*See* **Cognitive Science: Philosophical Issues**; **Consciousness, Cognitive Theories of**; **Consciousness, Philosophical Issues about**; **Mind–Body Problem**; **Neural Correlates of Consciousness as States and Trait**; **Philosophy of Neuroscience**; **Philosophy of Mind**; **Dualism**)

SUMMARY

Wilder Penfield was one of the twentieth century's most outstanding neurologists, neurosurgeons and scientists. Under his leadership, the Montreal Neurological Institute, which he founded in 1934, became a place where science and medicine combined harmoniously. Penfield's search for an effective treatment for epilepsy motivated him to study the brain at many levels – from cellular to anatomical – and led to the development of the Montreal procedure. Through various specific electrical stimulations of the exposed brain, a sensory–motor map of the body surface could be drawn.

Further Reading

Lewis J (1981) *Something Hidden: A Biography of Wilder Penfield*. Toronto: Doubleday Canada Ltd.

Penfield WG (ed.) (1932) *Cytology and Cellular Pathology of the Nervous System*. New York: PB Hoeber Inc.

Penfield WG (1941) *Epilepsy and Cerebral Localization: A Study of the Mechanisms, Treatment and Prevention of Epileptic Seizures*. Springfield, IL: CC Thomas.

Penfield WG (1958) *The Excitable Cortex in Conscious Man*. Springfield, IL: CC Thomas.

Penfield WG (1975) *The Mystery of the Mind*. Princeton, NJ: Princeton University Press.

Penfield WG (1977) *No Man Alone: A Neurosurgeon's Life*. Boston, MA: Little, Brown & Co.

Penfield WG and Rasmussen T (1950) *The Cerebral Cortex of Man*. New York: Hafner Publishing Co.

Penfield WG and Jasper H (1954) *Epilepsy and the Functional Anatomy of the Human Brain*. Boston, MA: Little, Brown & Co.

Piaget, Jean

Introductory article

Kurt Fischer, Harvard Graduate School of Education, Cambridge, Massachusetts, USA
Ulas Kaplan, Harvard Graduate School of Education, Cambridge, Massachusetts, USA

CONTENTS

Jean Piaget, one of the founders of cognitive science, established that children build their intelligence through acting in the world and that they develop through a sequence of qualitatively different organizations of thought and action. He related the diverse fields of human knowledge to the developing minds of children.

INTRODUCTION

Jean Piaget helped lay the groundwork for cognitive science in the twentieth century and became one of the best-known and most influential cognitive scientists. The enormous body of research and theory that he created continues to drive many of the central questions about the mind today, especially how it develops and takes different forms. He studied how infants and children construct knowledge of objects and events in the physical world, investigating with his many collaborators a wide array of tasks and concepts. His own label for his approach was *genetic epistemology*, the study of the development (genesis) of knowledge in its diverse forms (epistemology). His most famous research involved two phenomena: (1) conservation – understanding what conserves, or remains constant, when an object is changed, such as when a clay ball is reshaped into a sausage, and (2) object permanence – understanding that an object continues to exist even when it is not perceived. However, he studied scores of other developmental phenomena, covering most topics in epistemology of the physical world.

KNOWLEDGE: ACTIVITY, DEVELOPMENT, AND LOGIC

Piaget showed in all these domains how infants and children build their own skills and knowledge: 'Comprendre, c'est inventer' ('To understand is to invent'). The most important organizing concepts in his work were: (1) knowledge is based in activity: people actively transform objects and events in order to know them; (2) these activities develop systematically through a series of qualitatively different patterns, which can be characterized as stages; (3) logic is the basic organizing force behind the mind, and the most clearcut stages involve distinct forms of logic in action. Piaget's theoretical work focused primarily on characterizing (a) the forms of logic that organize the mind as children develop, and (b) the processes of change that regulate children's developing activities, which he called *equilibration*. He also worked with scholars from many disciplines with the goal of capturing the whole array of human knowledge. (*See* **Piagetian Theory, Development of Conceptual Structure**)

PIAGET'S DEVELOPMENT

Jean Piaget was born in Neuchâtel, Switzerland, on 9 August 1896, as the oldest child of Rebecca Jackson and Arthur Piaget. He was a precocious child, showing an early interest in biology and publishing at the age of 11 his first article, on an albino sparrow. He studied mollusks during his adolescence, assisting the director of a natural history museum in Neuchâtel. As a result of his publications about mollusks, he became a well-known zoologist by the end of high school.

He went on to study natural sciences at the University of Neuchâtel and received his PhD at the age of 21 based on a study of mollusks in Swiss lakes. Meanwhile, his interdisciplinary interests developed from an early age, fostered by his father (a professor of medieval literature) and godfather Samuel Cornat, who stimulated his interest in philosophy, religion, and logic. He became intensely involved from an early age with not only biology but also psychology, logic, epistemology, philosophy of science, and religion, and published two

papers in philosophy. One of his early goals was to find a source of absolute truth, ranging from religion to logic. After his graduation in 1918, he studied psychoanalysis and experimental methodology at the University of Zurich, and then went to Paris to study at the Sorbonne. He came into contact with the new intelligence-testing movement founded by Alfred Binet and Théodore Simon, and under Simon's supervision he devised and administered tests to schoolchildren.

Errors and Qualitative Changes in Development of Intelligence

In this period, he became interested in the errors that children make in performing cognitive tests. Much of his future work built on using these errors to infer children's strategies of acting and reasoning. He found that at specific ages children share certain distinctive strategies as evidenced by their systematic errors, and that those strategies change regularly with age. What matters in intelligence is less the number of correct responses than the type of reasoning. As evidenced by children's errors as well as by their correct responses, older children do not merely know more, but they act and think differently from younger children, showing qualitative changes in intelligence. One of Piaget's major focuses for the rest of his career was characterizing the nature of these qualitative changes in intelligence, along with the mechanisms of change.

Piaget was a brilliant observer and had a great eye for insightful collaborators. He and his colleagues discovered hundreds of strategies and errors in children's activities that initially surprised people and eventually captivated the interest of cognitive scientists and educators. These observations convinced many people that children's knowledge is based in their own activity and undergoes powerful reorganizations with age. Piaget viewed children as having a logic of their own by which they actively construct their world, deriving all knowledge from their actions. Children's minds are not merely immature versions of the adult mind, but they operate with distinct principles based in earlier, less sophisticated logic. Children have fundamentally different ways of making sense of the world, and at different periods of life they create qualitatively distinct mental structures to make sense of their activities and experiences. Infants begin by building a logic of direct action on the world, what Piaget called *sensorimotor* intelligence. Eventually they build such complex action systems that they create concrete representations of objects at about age two (*preoperational* intelligence). By six to eight years children construct a logic of *concrete operations* on objects, events, and people. Operations are mental actions that can be done both in the mind and on the real world. Building on this concrete logic, children of 10 to 12 years invent a formal, hypothetical logic that Piaget called *formal operations*, which has much in common with the rational logic of scientists and mathematicians.

Building the Foundation for Genetic Epistemology

In 1921, Piaget became the director of the Jean-Jacques Rousseau Institute in Geneva, which had been founded by his mentor Édouard Claparède. In 1923 he published his first book, *Le langage et la pensée chez l'enfant* (*Language and Thought in the Child*), based on his talent at interviewing schoolchildren about their understanding of phenomena in the world.

In the same year he married Valentine Châtenay, who herself was a talented psychologist and observer of children. The couple had three children: Jacqueline, born in 1925, Lucienne in 1927, and Laurent in 1931. Jean and Valentine observed their children's development intensively in the natural context of their home. Building on earlier work by James Mark Baldwin, Paul Guillaume, and others, they kept detailed diaries of their observations and created many seminal tasks that formed the foundation for major paradigms in cognitive science, such as object permanence (search of hidden objects), the logic of movement in space, and verbal imitation. In this way, consistent with Piaget's general approach, he built his own knowledge of human development through his and his wife's activities with their children – an interaction between theory and nature.

Piaget considered his early publications, including his first five books, to be working papers, not finished scientific documents. His and Valentine's observations with their infants produced the beginnings of his theory as it is known today. He first presented the new conception of infant development in a paper to the British Psychological Society in 1927, which was elaborated in three seminal books, *La Naissance de l'intelligence chez l'enfant* (*The Origins of Intelligence in Children*) (published in 1936), *La Construction du réel chez l'enfant* (*The Construction of Reality in the Child*) (1937), and *La Formation du symbole chez l'enfant* (*Play, Dreams, and Imitation in Childhood*) (1945). These works not only described a series of stages of cognitive

development in infancy and early childhood, but more importantly, according to Piaget in his 1936 book *The Origins of Intelligence in Children*, they taught him 'in the most direct way how intellectual operations are prepared for by sensorimotor action, even before the appearance of language'. Among his many important collaborators in later work were Bärbel Inhelder and Alina Szeminska, who helped to formulate explanations of the later stages of intelligence – concrete and formal operations.

PIAGET'S FRAMEWORK AND THE BREADTH OF HUMAN KNOWLEDGE

Piaget saw himself not primarily as a psychologist but as a genetic epistemologist, a scientist who studies the origins and development of human knowledge by combining biology and logic. He sought to uncover the nature of the growth of human knowledge, not only in children but in history. At the center of his framework, he saw an individual actively and continuously making sense of the world: 'I am a constructivist. I think that knowledge is a matter of constant, new construction, by its interaction with reality, and that it is not pre-formed. There is a continuous creativity.' By acting on the world, whether on material objects or on other social beings, people are constantly adapting to the world they live in and simultaneously creating their own reality. Piaget emphasized that simultaneously (a) individuals actively construct their own worlds, and (b) they interact constantly with the world in this construction, adapting their knowledge to their experience.

In his effort to integrate philosophical questions about human knowledge with empirical questions about human development, Piaget emphasized the unified structure and functioning of the human mind in an era of emphasis on the fragmented nature of human knowledge, led especially by behaviorism. He created an empirically based, epistemological breakthrough connecting philosophy to cognitive science and the development of intelligence. In particular, he explicated the development of knowledge in terms of key philosophical and scientific categories, starting with object, space, causality, and time in the three books based on his own children's development and moving on to incorporate quantity (number), classes, relations, scientific experimentation, and of course logic. He also examined in depth the structure and growth of human knowledge in human history and across knowledge disciplines, although his work with children is more famous. According to Piaget, individual development connects closely with species development, with intriguing parallels in which individual development sometimes follows historical development and sometimes reverses it. The development of a human being is a reflection of the development of humankind. At the heart of this parallelism is people's active construction of knowledge and reality, which creates qualitative transformations of activity and knowledge in order to adapt to the richly varying world.

In 1955 Piaget founded the International Center for Genetic Epistemology in Geneva, Switzerland, to support his search for the origins and growth of knowledge, and he directed the Center until his death in 1980 in Geneva. He led a group of scholars from different disciplines, including psychologists, philosophers, mathematicians, logicians, scientists, and historians of science, to build his vision of the unity of knowledge. Experts would come to the Center to collaborate, often staying for a year and participating in seminars led by Piaget. To master the basic concepts and framework in the discipline chosen for that year, Piaget himself worked intensely to make connections with genetic epistemology. According to Gardner in his 1982 book, *Art, Mind, & Brain*, he was 'pursuing his own religion – the passion for truth, the search for the totality of knowledge'. In 1967, he published *Logique et connaissance scientifique* (*Logic and Scientific Knowledge*), an encyclopedia surveying all the sciences from the viewpoint of genetic epistemology, and for the rest of his life he continued to publish books and articles to fulfill this vision.

With his focus on the breadth of epistemology, Piaget came late in his life to abstain from debates about psychological questions, such as stages and mechanisms of cognitive development, even while his 95 books and hundreds of articles were powerfully shaping the core questions and concepts of psychology, cognitive science, philosophy, and education. At the same time, he defined the limitations of his own explanations of intelligence, suggesting in the posthumously published book *Le possible et le nécessaire* (*Possibility and Necessity*) and elsewhere that, contrary to his earlier hypotheses, intelligence is not founded in logic but instead constructs logic as one of its greatest accomplishments, building on practical activities to eventually create the soaring possibilities of human intelligence.

CONCLUSION

Piaget established one of the central frameworks in cognitive science, including the propositions that

people build knowledge through acting in the world and that children develop through a sequence of powerfully different forms of intelligence before they reach adulthood. He helped bring together the many disciplines relevant to mind and knowledge, moving the field away from narrowly defined disciplines. He created an agenda for understanding mind and knowledge that continues to drive a wide range of research today, especially in cognitive development, learning, psychology, philosophy, cognitive linguistics, and education. Arenas in which great progress remains to be made include analyzing the dynamic processes of cognitive growth, characterizing the fundamental structures of knowledge in general as they extend beyond logic and science, and connecting knowledge and intelligence more effectively to biology, including both brain and body.

Further Reading

Bringuier J-C (1980) *Conversations with Jean Piaget*. Chicago, IL: University of Chicago Press.

Gardner H (1982) *Art, Mind, and Brain: A Cognitive Approach to Creativity*. New York, NY: Basic Books.

Gruber HE and Vonèche J (eds) (1977) *The Essential Piaget: An Interpretive Reference and Guide*. New York, NY: Basic Books.

Piaget J (1952) Autobiography. In: Boring EG, Langfield H, Werner H and Yerkes R (eds) *The History of Science in Autobiography*, pp. 237–256. Worcester, MA: Clark University Press.

Piaget J (1952) *The Origins of Intelligence in Children*, translated by M Cook. New York, NY: International Universities Press. [Originally published 1936.]

Piaget J (ed.) (1967) *Logique et connaissance scientifique*. Paris, France: Gallimard.

Piaget J (1971) *Biology and Knowledge: An Essay on the Relations between Organic Regulations and Cognitive Processes*, translated by B Walsh. Chicago, IL: University of Chicago Press. [Originally published 1967.]

Piaget J (1987) *Possibility and Necessity*, translated by H. Feider. Minneapolis, MN: University of Minnesota Press. [Originally published 1981–1983.]

Piaget J and Garcia R (1989) *Psychogenesis and the History of Science*. Cambridge, UK: Cambridge University Press. [Originally published 1983.]

Piaget J and Inhelder B (1969) *The Psychology of the Child*. New York, NY: Basic Books. [Originally published 1966.]

Piatelli-Palmarini M (ed.) (1980) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Vidal F (1994) *Piaget before Piaget*. Cambridge, MA: Harvard University Press.

Sapir, Edward

Introductory article

Regna Darnell, University of Western Ontario, London, Ontario, Canada

CONTENTS

Edward Sapir (1884–1939) was the foremost linguist among the first generation of North American anthropologists trained by Franz Boas. His work on American Indian languages, linguistic relativity, and culture and personality defined the scope of linguistics within anthropology.

INTRODUCTION

Edward Sapir, a member of the first generation of students of Franz Boas who came to dominate North American anthropology during the first three decades of the twentieth century, was the primary linguistic specialist among these pioneering students of American Indian languages. His humanistic view of language also led him to consider the creative role of the individual in culture and society.

Sapir was born in Lauenberg, Pomerania (now Germany), in 1884, but grew up on the Lower East Side of New York City where his father was a cantor. He obtained a prestigious Pulitzer scholarship which paid his way to Columbia University. After obtaining his first two degrees in Germanic linguistics, Sapir switched to anthropology, purportedly because Boas convinced him of the urgency of recording endangered Native American languages. He obtained a BA degree in 1904, an MA in 1905, and a Ph.D. in 1909 with a dissertation on the Takelma language of Oregon.

First-hand field work was the basis of the Boasian paradigm, in both linguistics and ethnology. Accordingly, Sapir spent the years 1907 and 1908 in California doing fieldwork on Yana, and the years from 1908 to 1910 at the University of Pennsylvania where he worked on Ute and Southern Paiute, before accepting a position as the first Director of the Anthropological Division of the Geological Survey of Canada (part of the Department of Mines). He chose to specialize in the languages of the Northwest Coast, although the program of his division was designed to cover the entire Dominion of Canada.

Despite his growing dissatisfaction with museum work, the impossibility of securing a teaching position in Ottawa, and reductions in funding for research and publication during the First World War, Sapir remained in this position until 1925 when he moved to the University of Chicago. There he became increasingly involved in an interdisciplinary social science synthesis funded by the Rockefeller Foundation and primarily focused on sociology, psychology and psychiatry. Sapir, who insisted that anthropology should study everyday life in North America as well as the exotic and primitive, became the primary translator and synthesizer across these diverse disciplines. In 1931 he moved to Yale University to organize an interdisciplinary seminar for foreign fellows on 'the impact of culture on personality'. At Yale, he developed a school of linguistics which emphasized meaning and process. Although troubled by declining health and by anti-Semitism at Yale, Sapir remained there until his death in 1939 at the age of 55.

AMERICAN INDIAN LANGUAGES

Sapir is remembered within anthropological linguistics primarily for his production of grammars, texts and dictionaries based on his first-hand fieldwork with a series of American Indian languages. Although much of his work was incomplete at the time of his death, the diversity of the languages he studied – Wishram Chinook, Takelma, Chasta Costa, Yana, Kato, Catawba, Ute, Southern Paiute, Hopi, Nootka, Comox, Mohawk, Seneca, Tutelo, Delaware, Abenaki, Malecite, Micmac, Montagnais, Cree, Tlingit, Nass River, Kootenay, Thompson River, Lillooet, Shuswap, Okanagan, Haida, Tsimshian, Sarcee, Ingalik, Kutchin, Navajo, Hupa, Yurok, and Chimariko – provided him

with an extensive database for historical comparative work; this combined with an intuition for language which was called 'genius' by his contemporaries and successors alike.

In 1916, Sapir published a paper called 'Time perspective in Aboriginal American culture: a study in method', which exemplified the Boasian approach to reconstruction of cultural history using examples from some 60 (mostly Amerindian) languages. Sapir demonstrated by the diversity of these miniature analyses that language was more amenable to historical inference than the rest of culture, and that this enabled the student to distinguish between common past history and borrowing. (Boas himself rejected this distinction. His position diverged increasingly from Sapir's as the latter turned to historical inference as a framework for ethnology.)

Based on his own fieldwork, supplemented by that of other Boasians, especially Alfred Kroeber in California, by 1921 Sapir had reduced the 55 linguistic families of North America codified by the Bureau of American Ethnology in 1892 to six superstocks that he thought could be correlated with broad continental migration patterns. Within each of these larger units, Sapir distinguished more conservative subgroups. American anthropologists followed the Sapir classification as a framework for ethnology at least until 1963. Although current classifications tend to be more conservative, many of Sapir's broadest generalizations still preoccupy his successors.

Sapir's only book, *Language: An Introduction to the Study of Speech* (1921) was directed at a popular as well as an academic audience. His dependence on American Indian examples was virtually unique in the Indo-European-based linguistics of the day, and emphasized his anthropological commitment to the equal value and communicative capacity of all human languages.

LINGUISTIC RELATIVITY

Sapir followed his teacher Boas in insisting that the grammatical categories of familiar languages like Greek and Latin could not be applied to unrelated languages. Rather, each language and language family had to be analyzed in terms of its own unique categories. Moreover, Sapir was determined to incorporate the 'psychological reality' of a language for its speakers into his linguistic analysis. He defined the concept of the phoneme, the smallest meaningful unit of sound in a given language, in 1925 and in 1933 he elaborated on the basis of this linguistic distinction in the perceptions of native speakers. The objective character of sounds across languages was less important than their patterning within a single language at a given moment in time. This focus on meaning as construed by native speakers of a language foreshadowed the abstract post-Chomskyan linguistics that has dominated the discipline since the mid-1960s. In contrast, Boas and many of his students within anthropology continued to insist on recording sounds in as much detail as possible using a phonetic grid assumed to apply across languages.

Although the so-called Sapir–Whorf hypothesis of linguistic relativity is best known through the formulations of Sapir's student Benjamin Lee Whorf, many passages in the work of Sapir, and indeed of Boas, share the common-sense assumption that the categories of language influence the way native speakers think about the world around them. Stated in this rather weak form, linguistic relativity seems almost trivial. On the other hand, the strong claim that language determines thought cannot be demonstrated in any fully convincing way.

After Sapir's death in 1939, North American structural linguistics, under the influence of Leonard Bloomfield, moved away from the study of meaning. Even many of Sapir's former students moved towards a more behaviorist and positivist notion of both language and culture. Whorf's accounts of the vast differences between the grammatical cateogies of Hopi, based on his own fieldwork, and what he called 'Standard Average European' were less persuasive in the changed intellectual climate. Efforts to prove or disprove the Sapir–Whorf hypothesis by scientific means produced ambiguous results, leading the linguistic relativity hypothesis itself to fall into disrepute. Neither Sapir nor Whorf, however, understood linguistic relativity as a scientific hypothesis. Rather, it followed from Boasian cultural relativism that each culture, and language, should be understood first in its own terms. Only thereafter could translinguistic insights emerge from what Whorf called 'multilingual awareness.'

Recent returns to questions of linguistic relativity within cognitive linguistics, especially in the work of Stephen Levinson and John Lucy, cite Sapir and Whorf as significant precursors, but have shifted the focus of the argument away from Boasian preoccupations with differences across cultures and languages. Working in a post-Chomskyan linguistic context, they frame linguistic differences within universal grammatical constraints that seem to belie the very diversity that was the original reason

for conceptual relativity in the study of particular languages.

CULTURE AND PERSONALITY

From about 1910, Boasian anthropology began to turn from reconstruction of American Indian cultural histories towards the study of meaning for the individual; what Boas called 'psychology' was later 'called culture and personality'. Sapir preferred to speak about the influence of culture on personality, i.e., how the individual acquired a language already codified within his or her speech community while still retaining a creative ability to modify that language. Sapir was among the first to acknowledge the diversity of individual personalities and social positions within a single society, foreshadowing sociolinguistics and the ethnography of communication.

Few anthropologists or linguists have matched Sapir's extraordinary range of interests, which included American Indian languages, linguistic theory, historical linguistics, the individual in culture, psychology and psychiatry, life history, and folklore. (He also wrote poetry and composed music.) For Sapir, these interests were clearly integrated. For example, working closely with a small number of linguistic consultants on native language texts that expressed their point of view on cultural matters encouraged him to attend to differences between individuals and to consider how all native speakers had somewhat different perceptions of the common culture. Like Boas, Sapir trained consultants to write their own languages and record texts during his absence from the field. This Boasian text tradition encouraged use of the same texts as a database for linguistic, cultural and psychological analysis. Sapir defined culture in terms of symbols in people's heads; his ethnography, compared with that of his contemporaries, relied heavily on linguistic codifications of world views. The 'native point of view' was implicit in the texts.

ANTHROPOLOGICAL LINGUISTICS

Throughout his career, Sapir's professional identity oscillated between anthropology and linguistics. He approached the study of culture as a linguist, focusing on the words and connected texts produced by native speakers of American Indian languages. He approached the study of language by asking questions about the role of culture and socialization in forming the thoughts and actions of individuals: meaning, for him, was culturally constituted rather than universal, at least in its surface forms. Moreover, to the consternation of those linguists who assumed that civilization was a prerogative of literate cultures, Sapir insisted that the insights of Indo-European linguistics could be applied to the study of unwritten languages. Despite the absence of writing, sound changes and other historical processes operated in the same systematic way in American Indian languages. The cultural relativism of Sapir's anthropology, as expressed in Ruth Benedict's *Patterns of Culture* in 1934, applied also to the study of linguistic form.

After his return to academia in 1925, Sapir trained students in both linguistics and cultural anthropology, without apparent disjuncture. The anthropology students were exposed to Sapirian linguistics and the linguistics students worked on American Indian languages, with many of them pursuing careers in anthropology departments.

Sapir was an important figure in the genesis of linguistics as an independent professional discipline, arising from literature and language departments as well from anthropology. He was a founding member of the Linguistic Society of America in 1925 and the paper in which he defined the phoneme appeared in the first issue of its journal *Language* in the same year. Sapir's stature in theoretical as well as descriptive linguistics reinforced the links between the two disciplines. On the one hand, linguists increasingly acknowledged that an adequate science of language could not be based solely on the study of Indo-European languages. On the other hand, language provided anthropologists with their best means of access to unfamiliar cultural worlds. For anthropologists, language has remained the symbol system *par excellence*, although the technical skills of the linguist have become less accessible to most cultural anthropologists. Within both disciplines, Sapir remains the single most significant figure linking linguistics and anthropology in method, theory, and practice.

Further Reading

Benedict R (1934) *Patterns of Culture*. Boston, MA: Houghon Mifflin.

Carroll J (ed.) (1956) *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. New York, NY: Wiley.

Darnell R (1990) *Edward Sapir: Linguist, Anthropologist, Humanist*. Berkeley, CA and Los Angeles, CA: University of California Press.

Darnell R (2001) *Invisible Genealogies: A History of Americanist Anthropology*. Lincoln, NE: University of Nebraska Press.

Gumperz J and Levinson S (eds) (1996) *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.

Irvine JT (1994) *Edward Sapir's The Psychology of Culture*. Berlin: Mouton de Gruyter.

Lucy J (1992) *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge, UK: Cambridge University Press.

Mandelbaum D (ed.) (1949) *Selected Writings of Edward Sapir*. Berkeley, CA and Los Angeles, CA: University of California Press.

Sapir E (1921) *Language: An Introduction to the Study of Speech*. New York, NY: Harcourt Brace.

Saussure, Ferdinand de

Introductory article

W Terrence Gordon, Dalhousie University, Halifax, Nova Scotia, Canada

CONTENTS

Ferdinand de Saussure (1857–1913) is generally acknowledged as the father of modern linguistics. Observations and speculations about human language form a long tradition of philosophical inquiry dating back to antiquity, but it was Saussure who first focused on the systematic nature of language, thus launching twentieth-century linguistics and making it possible to put the discipline on a scientific footing.

INTRODUCTION

Nearly a century after Saussure's first contributions to linguistics, the discipline has passed through various phases of development and expansion which have changed its goals and terminology, although systematicity and the scientific orientation remain at its core. In North America, linguistics in the first half of the twentieth century was dominated by a structuralist approach refined by Leonard Bloomfield (who wrote a positive review of Saussure). Since the 1960s, the North American brand of linguistics has been dominated by transformational generative grammar and its manifold developments and reactions to it. Noam Chomsky's initial comment on Saussure's linguistics was that it focused on systems of elements rather than systems of rules. It is this distinction between the goal of descriptive adequacy (Saussure) and explanatory adequacy (Chomsky) that separates Saussure's linguistics from the dominant note in linguistics today.

Born into a Swiss family that had already produced generations of distinguished scholars, Saussure made his first mark at the age of 21 with a treatise demonstrating that proto-Indo-European (the source of a vast family of languages of Europe, India, and southwest Asia) did not have just three vowels, as had been widely supposed, but five. Saussure began his university teaching career in Paris in 1880, but returned to his homeland in 1891 to accept an invitation to teach at the University of Geneva. It was only late in his career that he agreed to give a course of lectures in general linguistics. This he did between 1906 and 1911, modifying the course from year to year without ever satisfying himself it was complete or worthy of publication. When he died in 1913, those who had followed his lectures thought otherwise, and collated and edited their own notes to produce the *Course in General Linguistics*, which has been in print ever since. The original French text has been translated into more than twenty languages.

The *Course in General Linguistics* has a profound coherence that is achieved by using a set of pairs of basic terms as an organizational principle. The terms of each pair relate to each other, each pair relates to the other pairs, and the entire set relates back to the concept of the linguistic sign. ('Sign', in this sense, refers to anything that tells us about something other than itself; the sign can be a word, a part of a word that carries meaning (the *-ly* in *quickly*, for example), or a group of words making up a complex sign (*do it quickly*).)

LANGUAGE AND SPEECH

Saussure defined the linguistic sign as belonging to language, the system of elements that speakers of a given language know and use. The sounds that actualize a sign belong to speech; i.e. not to system *as* system but in use. Saussure did not invent these terms of course, but he saw the first to insist that linguistics needed to be grounded in the distinction, in order to avoid the errors of scholarship that he thought characterized so much of the work of his predecessors.

SIGNIFIER AND SIGNIFIED

The terms 'signifier' and 'signified' designate the two parts of the linguistic sign for Saussure. The

signifier is a mental image of a recurring unit (usually a short sequence of sounds) that allows a language user to put that element of the language system to use. The signified is the concept connected to the signifier. In the original French, the terms are *signifiant* (literally translated as 'signifying') and *signifié* ('signified'). What may at first appear to be a potentially confusing terminology is, in fact, a reminder both of the necessary distinction between the two parts of the linguistic sign and of the distinction between those parts and the sign as a whole.

SYNCHRONIC AND DIACHRONIC ANALYSIS

The terms 'synchronic' and 'diachronic' describe two fundamentally different ways of analyzing language. Saussure emphasizes that a complete description of language in general or of any particular language must take account both of the community that uses it and of the effects of time on the language system. These two factors are radically different from each other and produce radically different consequences for language. A synchronic analysis takes account of relations that hold among coexisting elements in a language system at any given moment, and is thus independent of any factor attributable to the passage of time. It provides a snapshot of the state of the language under analysis. The idea of systematicity in language implies that, if the linguist's account is valid, it will present that state as a unified whole of interacting elements. By contrast, a diachronic analysis yields a description in which the evolution of only isolated elements of various states of a language at different times constitute the relevant data.

The synchronic and diachronic modes of analyzing linguistic data are autonomous but interdependent. The state of a language is not simply the sum of the changes it has undergone. Diachronic facts do not belong to a state of a language, because any such state is a synchronic phenomenon. But diachronic linguistics deals with forms that replace one another as a system evolves. There is thus a complementarity between synchronic and diachronic linguistics, which Saussure described as the diverging paths that linguistics can take.

FORM AND SUBSTANCE

As in other instances, Saussure imported the terms 'form' and 'substance' from everyday language as a starting point for finer distinctions that were subsequently expressed by more specialized analytical terms. Thus, the distinction between 'form' and 'substance' anticipates the distinctions between 'signification' and 'value' and between 'arbitrariness' and 'motivation'. The *Course in General Linguistics* has little to say about form and substance, but the distinction is crucial to Saussure's design for a sign-based approach to linguistics: the link between sound and thought in the linguistic sign produces form, not substance. Saussure calls the link between signifier and signified, as well as that among signs, 'pure form', indicating by 'pure' that it consists 'solely' of a relation. In this sense, no substance is essential to the formation of the linguistic sign, or to the endlessly complex relations among the signs of a language system, or to the communicative acts for which the system serves.

SIGNIFICATION AND VALUE

The terms 'signification' and 'value' distinguish between two types of meaning: signification is the meaning belonging to a sign taken individually, and value is the meaning that derives from the contrast between or among signs. Again, the terminology is intended as a reminder of both the contrast and the interplay between the terms involved: in this instance, reminding us that the meaning inherent in a sign by itself (signification) is subordinate or preliminary to the meaning emerging from signs in relation to each other (value).

Saussure draws attention to the paradox inherent in any system of values: in order to function, it must contain elements that are unlike whatever they may be exchanged for, and whose value they function to determine, but similar to those elements whose value they function to determine. The principle that distinguishes signification from value distinguishes forms from each other and creates meaning. The signification–value interplay thus harks back to form and substance.

DIFFERENCE AND OPPOSITION

According to Saussure, the meanings created out of distinctions of form are carried by those distinctions alone. That is what he means by 'difference'. No positive elements are required for the formation of a linguistic system; it functions to create perceptually discernible differences, and those differences function to distinguish among ideas for purposes of communication. Such differences in signifiers and signifieds taken separately are what Saussure calls 'pure' differences: purely negative. But, once the linguist's analysis moves to the level of the sign

as a functioning unit of communication, where signifier and signified must be taken together, it is no longer a matter of difference but of 'opposition'. The two components of the sign differ from the components of other signs, but the sign as a whole is merely distinct from others. The features of language structure as Saussure defined them, once he had added the concepts of difference and opposition to his analysis, are all based on the distinctiveness of the linguistic sign. In the language system there are only differences; among signs there are only oppositions.

SYNTAGMATIC AND ASSOCIATIVE RELATIONS

Syntagmatic relations are complex signs consisting of at least two components, such as *im-possible*, *possibl-y*, or *possible suspect*. Saussure does not specify any upper limit for what could be considered a complex sign, but syntagmatic relations of many different types operate in the formation of all signs from the level of compound words to phrases, full sentences, and, by implication, complete texts. The linear structure and the functional combination and integration that characterizes all such data distinguish syntagmatic relations from associative relations. The latter are the connections that language users spontaneously make among signs on the basis of signifiers or signifieds or both.

Saussure conceives of every sign in a linguistic system as the center of a constellation around which distinct groups of other signs may cluster in varying numbers, depending on the complexity of the sign at the center. These word groups are never spoken as such, unlike the groups that form syntagmatic relations are. They are related forms among which speakers choose in forming syntagmatic relations. Thus, at the level of functional units of communication, these two features of language structure again interact intimately with one another.

ARBITRARINESS AND MOTIVATION

In an early chapter of the *Course in General Linguistics*, the arbitrariness of the linguistic sign is mentioned for the first time. 'Arbitrariness' refers to the arbitrary connection between the signifier and the signified of a linguistic sign, in the sense that no language mandates the expression of any of its signifieds by a specific signifier. Conversely, no signifier is constrained by virtue of its inherent properties to signify a specific signified. It is not till all the other complementarities of the *Course in General Linguistics* have been established that the principle of the arbitrariness of the linguistic sign takes its place within the distinction arbitrariness–motivation, or arbitrary–motivated. The logic of this order of presentation emerges from the fact that the difference–opposition, signification–value, and syntagmatic–associative distinctions each operates in its own way to limit arbitrariness, thus creating signs that are at least partially motivated (not completely arbitrary).

CONCLUSION

The foundational complementarities developed in the portion of the *Course in General Linguistics* devoted to synchronic analysis take their place in the section denoted to diachronic analysis. There are additional sections devoted to what Saussure calls 'geographic' and 'retrospective' linguistics, but these are scarcely mentioned in the vast body of literature surrounding his work. Nor has much attention been paid to the obsessive and abstruse scholarship to which Saussure devoted substantial time and effort in an attempt to decode hidden words within visible words of Horace, Virgil, and other Latin authors. But both Saussurean scholarship and serious distortions of his seminal ideas flourish, a century after he first challenged an attentive audience to believe that the course of linguistics needed to be recharted and suggested how it could be done.

Further Reading

Bouquet S (1997) *Introduction à la Lecture de Saussure*. Paris, France: Payot.

Bouissac P (2002) *Cyber Semiotic Institute*. http://www.chass.utoronto.ca/epc/srb/cyber/cyber.html

Culler J (1986) *Ferdinand de Saussure*. Ithaca, NY: Cornell University Press.

Gordon WT (1996) *Saussure for Beginners*. New York, NY and London: Writers and Readers.

Harris R (1987) *Reading Saussure*. London, UK: Duckworth.

Harris R (2001) *Saussure and His Interpreters.* Edinburgh, UK: Edinburgh University Press.

de Saussure F (1959) *Course in General Linguistics*, translated by Wade Baskin. New York, NY: Philosophical Library.

de Saussure (1987) *Course in General Linguistics*, translated by Roy Harris. London, UK: Duckworth.

Starobinski J (1970) *Words Upon Words: The Anagrams of Ferdinand de Saussure*. New Haven, CT: Yale University Press.

Tallis R (1995) *Not Saussure*. London, UK: Macmillan.

Thibault P (1997) *Re-Reading Saussure*. London, UK: Routledge.

Shannon, Claude

Introductory article

Thomas J Carter, California State University, Stanislaus, California, USA

CONTENTS

Claude E. Shannon (1916–2001) is regarded as the founder of information theory. He developed a general model for communication systems, and theoretical tools for their analysis.

INTRODUCTION

Claude Elwood Shannon (1916–2001) is generally considered to be the founder of the contemporary theory of information. Shannon's work in mathematical engineering focused on problems associated with telecommunications and computer systems. His early work developed the application of Boolean algebra to telephone switching circuits. In the 1940s, while working at the Bell Telephone Laboratories, he developed and applied a theoretical model of the communication of information through a transmission channel. This theoretical model allowed him to clearly define a measure of the information content of a message, a measure of the capacity of a communication channel, and relationships between these measures. His theory allows rigorous analysis of the efficiency and reliability of communication through a noisy channel. A noisy channel in this context is one in which the receiver might have some doubt about exactly what message was sent. Shannon's work also forms the basis for the foundational paradigm of cognitive science in which human cognition is analyzed as an information-processing system.

Claude E. Shannon was born on 30 April 1916, in Gaylord, Michigan, to Claude Elwood and Mabel Wolf Shannon. His father was a judge, and his mother was principal of the town's high school. Shannon married Mary Elizabeth Moore on 27 March 1949. They had three children. Shannon died on 24 February 2001 in Medford, Massachusetts, of Alzheimer disease.

Shannon earned double BS degrees in electrical engineering and mathematics in 1936 at Michigan University. He earned his Masters in electrical engineering and his doctorate in mathematics at the Massachusetts Institute of Technology (MIT). Both degrees were awarded in 1940. Shannon's work for his Masters degree, done with Vannevar Bush on Bush's differential analyzer machine, established the relationship between Boolean logic and switching circuits. His Masters thesis, published in 1938 under the title 'A Symbolic Analysis of Relay and Switching Circuits', drew immediate attention, and has since been recognized as fundamental in laying the groundwork for the use of Boolean algebra in the theoretical analysis and design of digital computers. Shannon's work for his doctorate involved the application of mathematical methods to genetics. His doctoral thesis was entitled 'An Algebra for Theoretical Genetics'.

In 1940, Shannon was named a National Research Fellow, and spent a year at the Institute for Advanced Study at Princeton. He was awarded the National Medal of Science (1966), the Jacquard award (1978), the Kyoto Prize in Basic Science (1985), and numerous other awards and honors, including honorary degrees from such institutions as Carnegie-Mellon, Edinburgh, Oxford, Princeton, and Yale.

Bell Telephone Laboratories hired Shannon in 1941 as a research mathematician. He worked at the lab until 1956, and remained affiliated with the lab until 1972. In 1956 and 1957 he was a visiting professor at MIT. In 1958, he was appointed by MIT as Donner Professor of Science, where he remained until his retirement.

During the Second World War, Shannon worked in a team at Bell Labs developing anti-aircraft control and aiming devices. Shannon also continued his work on switching circuits, and developed his general theory of information and communication channels. In 1948, in the *Bell System Technical Journal*, Shannon published his most important work, 'A mathematical theory of communication'. During his career, Shannon applied his theoretical methods to telecommunication systems, computer systems, cryptography and cryptanalysis, the stock market,

and games such as chess-playing machines. A volume of Shannon's collected papers has been published, containing important works selected from 127 published and unpublished papers, including some that had formerly been classified secret, but since declassified.

Shannon had a life-long interest in games, toys, musical instruments, and devices of all sorts. He invented a rocket-powered Frisbee, devices which can juggle, a machine to solve Rubik's cube, and a computer that calculates in Roman numerals.

INFORMATION THEORY

In his classic 1948 paper, 'A mathematical theory of communication', Shannon laid the foundations for contemporary information, coding, and communication theory. He developed a general model for communication systems, and a set of theoretical tools for analyzing such systems. His basic model consists of three parts: a sender (or source), a channel, and a receiver (or sink) (see Figure 1). His general model also includes encoding and decoding elements, and noise within the channel. In his paper, Shannon analyzed both discrete (digital) and continuous systems. Here we will briefly outline the elements of Shannon's analysis of discrete communication channels.

Shannon's first important step was to develop a rigorous, formal, general definition of *information*. He focused on information carried by symbols (such as letters, words, or digits). After abstracting away incidental properties of the symbols (such as their size, color, typeface, etc.), Shannon characterized a symbol by its probability of occurrence. Thus, in general, Shannon's theory analyzes the information carried by a sequence of abstract events, each occurring with a specific probability. It is worth mentioning that Shannon's use of the term *information* is technical, and that successful use of his theory in a particular context depends on the applicability and relevance of his definitions. For example, with Shannon's definition of *information*, a completely random sequence of symbols (each equally likely) has the highest possible information content. This issue of technical applicability has sometimes led to unwarranted expectations of information theory, and an ebb and flow of its reputation over the years since its appearance in 1948.

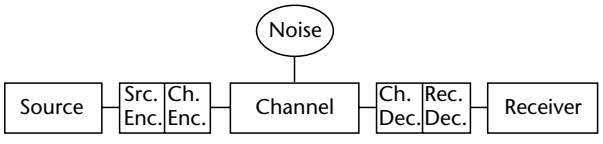


Figure 1. Shannon's general communication model. Fundamental elements are source, channel, and receiver/sink. The general model also includes channel noise, source and channel encoding elements, and channel and receiver decoding elements.

Shannon wanted a quantifiable measure of the information to be derived from the observation of an event occurring with probability p, and he wanted the measure to satisfy certain basic conditions. In particular, he wanted his information measure I(p) to be:

1. Non-negative – that is, $I(p) \geq 0$.
2. Continuous – that is, small changes in p should result in small changes in I(p).
3. Additive for independent events – that is, $I(P_1{}^* P_2) = I(p_1) + I(p_2)$.

The first property assures that the occurrence of an event can never reduce our information about the world (although it may cause us to revise some things we thought we knew). The second property is a standard mathematical requirement. The third property says, for example, that two books of the same size can contain twice as much information as one book, or that if we have two telephone lines, we can transmit twice as much information per unit time as with one line.

From these three properties, we can show that the information measure is given by (the negative of) the logarithm of p: $I(p) = -\log(p)$. Using different bases for the logarithm amounts to using different units of measurement. If we use base 2, then we are measuring information in units of *bits*, the standard unit of measurement in binary computer systems.

A useful way to think about Shannon's definition of *information* is that when we observe an event, our uncertainty about the world is reduced. The amount by which our uncertainty is reduced is the information content (for us) of the event. Thus, for example, before a coin is flipped, we are uncertain whether it will come up heads or tails. When we see that it has come up, say, heads, we have observed an event which has probability one-half. Thus, the information we have received (the reduction in our uncertainty about the world) is $I(1/2) = -\log_2(1/2) = \log_2(2) = 1$ bit. If a coin is flipped twenty times (or, alternatively, if twenty coins are each flipped once), we receive twenty bits of information. More generally, if an event is almost sure to happen (probability very close to 1), the information we receive from that event is very close to 0. If a very unlikely event occurs (probability very close to 0), we gain much more

information. If you learn tomorrow that someone won the lottery, you have learned little, since nearly always somebody wins the lottery. If you learn that you in particular won (a very unlikely event), then you have learned a lot!

ENTROPY

Using this basic definition of the amount of information carried by the occurrence of a single event of probability p, such as the observation of a particular symbol out of a set of possible symbols, Shannon went on to calculate the average amount of information carried per symbol in a stream of symbols. Thus, for example, we might analyze a sequence of words spoken by an individual, or a sequence of characters transmitted by a modem over a telephone line. If we have a finite set of symbols $\{a_i\}$, each of which occurs in the stream with probabilities $\{p_i\}$, then the average information carried per symbol in the stream is given by the weighted average:

$$H = -\Sigma_i \, p_i * \log(p_i)$$

Shannon called this quantity the *entropy* of the stream of symbols. This quantity is sometimes called the *information entropy*, to distinguish it from the *physical entropy* of thermodynamics, as encountered in physics or chemistry. In fact, however, these two notions of entropy are compatible with one another, and it is not a coincidence that they have the same general form. Shannon's methods have been successfully adapted to applications in the physical sciences.

Shannon's seminal paper, 'A mathematical theory of Communication', is a careful and insightful application of these two basic definitions of *information* and *entropy* to a variety of problems of communication.

COMMUNICATION THEORY

Having developed his definitions of information and entropy, and his general model of a communication system (sender–channel–receiver), Shannon addressed two of the most fundamental problems of communication theory.

The first general problem is overall efficiency of transmission, or, said slightly differently, data compression. If I want to send a particular message to someone, what is the shortest stream of symbols I can send to convey that message fully? For example, I once saw a sign which read 'f u cn rd ths, u cn gt a gd jb'. I was able to decode the message as 'If you can read this, you can get a good job'. Instead of sending 45 symbols (letters, spaces, and punctuation), the makers of the sign sent only 31, a reduction of almost one third.

Shannon proved that if we want to be able to encode the symbols of an arbitrary message for transmission, the average number of bits per encoded symbol will have to be at least as large as the entropy of our stream of symbols. That is, $H \leq$ average-code-length. Shannon also pointed out that if we want to come close to the minimum (to the most efficient code), we should encode frequently occurring symbols with short codes. A particular example of this is the Morse code. For example, the two most frequent letters in English text are 'e' and 't'. In Morse code, the code for 'e' is 'dot' and the code for 't' is 'dash', the two shortest possible Morse codes. Thus Morse code is a first step towards efficient coding of English text, although it is far from optimal since it does not take advantage of sequential dependencies among the letters.

The second general problem Shannon addressed is reliability of transmission when the channel is noisy. The simplest way to increase reliability is through repetition (if they might not have heard you right the first time, say it again, and, if necessary, again). Unfortunately, repetition decreases efficiency, since often the repetition will be wasted, if the message got through correctly the first time.

Shannon observed that noise in a channel has the effect of increasing the entropy of the message. He then developed a method for calculating the relative entropy of the channel, and defined the *capacity* of a channel as the difference between the actual entropy of the channel and the entropy of a channel of pure noise (where no message can get through). Shannon's main result was to prove that given a channel with non-zero capacity, there exists a way to encode messages with arbitrarily high reliability (nearly perfect transmission), but wasting an arbitrarily small amount of the channel capacity. This is a remarkable result. At first sight, it seems that if we want to improve the reliability of our transmissions through a noisy channel, we will in effect have to repeat our message, thus wasting some of the channel capacity. Shannon proved otherwise: there exist codes that give high reliability while wasting hardly any of the channel capacity.

Shannon's theoretical model of communication, and his information and entropy tools, have found wide applicability across the broad range of sciences and technology, and he stands as a central figure in the development of the Information Age. (*See* **Natural Language Processing, Disambiguation in**; **History of Cognitive Science and**

Computational Modeling; **Information Theory**; **Natural Language Processing, Statistical Approaches to**)

Further Reading

Campbell J (1982) *Grammatical Man – Information, Entropy, Language, and Life*. New York, NY: Simon and Schuster.

Gatlin LL (1972) *Information Theory and the Living System*. New York, NY: Columbia University Press.

Haken H (1988) *Information and Self-Organization: A Macroscopic Approach to Complex Systems*. Berlin/New York, NY: Springer-Verlag.

Hamming RW (1986) *Coding and Information Theory*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Pierce JR (1980) *An Introduction to Information Theory – Symbols, Signals and Noise*, 2nd revised edn. New York, NY: Dover Publications.

Roman S (1997) *Introduction to Coding and Information Theory*. Berlin/New York, NY: Springer-Verlag.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* **27**: 379–423; also 623–656.

Shannon CE (1993) *Collected Papers*, edited by NJA Sloane and AD Wyner. Piscataway, NJ: IEEE Press.

Slepian D (ed.) (1974) *Key Papers in the Development of Information Theory*. New York, NY: IEEE Press.

Simon, Herbert A.

Introductory article

Edward A Feigenbaum, Stanford University, Stanford, California, USA

CONTENTS

Herbert A. Simon (1916–2001) was an American scientist whose research ranged broadly over the cognitive and social sciences, computer science, economics, and the philosophy of science. For his fundamental, innovative, and penetrating contributions he received the highest research awards in the fields of economics, psychology, computer science and artificial intelligence, including the 1978 Nobel Prize for Economics for his model of bounded rationality in decision-making and problem-solving.

INTRODUCTION

Herbert A. Simon made seminal contributions to several fields of behavioral and social science, to computer science, and to the philosophy of science. His contemporaries often referred to him as the quintessential 'Renaissance man'. For the originality and profundity of his contributions, he received the highest awards of several sciences: the Distinguished Scientific Contribution Award of the American Psychological Association; the A. M. Turing Award of the Association for Computing Machinery; the Nobel Prize in Economics; and the Lifetime Research Achievement Award of the American Association for Artificial Intelligence.

Simon was born in Milwaukee, Wisconsin, in 1916. He did his doctoral dissertation research at the University of Chicago in the mid-1930s, studying organizational and administrative behavior. These studies led to two influential books: the early *Administrative Behavior* (continually revised into the 1990s) and the later (1958 collaboration with J. G. March) *Organizations*. These books laid the foundations of the social science discipline now called organization theory.

In the late 1940s, Simon taught at the Illinois Institute of Technology. He moved to the Carnegie Institute of Technology (now Carnegie Mellon University) in 1949 to help found the Graduate School of Industrial Administration. His most important contributions to economics, psychology, and computer science were made at Carnegie Mellon, and he remained there until his death in 2001. Although he began his career there as a Professor of Industrial Administration, for most of it he was a Professor of Psychology and Computer Science. In 1978, when he was awarded the Nobel Prize in Economics, he received the congratulatory messages from around the world in his office in the psychology department, his academic home.

He wrote 27 books, more than 600 scholarly papers and contributions, and an autobiography entitled *Models of My Life*. In the words of one of the memorial papers written about him, he was the 'model of a scholar'.

THEMES IN SIMON'S WORK

Before discussing the details of Simon's contributions, it is important to mention the themes that run through his contributions to so many fields. One might ask (as Simon himself might have done): what model of Simon could explain how he was able to make foundational contributions to such seemingly disparate fields?

Modeling Decision-making

The most important theme in Simon's work, which tied together his various contributions in different disciplines, was his focus on models of the decision-making behavior of individuals, and of organizations composed of individuals. The 'decision' was, in effect, the unit of behavior that he studied.

For example, in economics, the standard model had been 'economic man' who made choices using unlimited information and unlimited information-processing power. Simon replaced that with a model that had more empirical validity: 'behavioral man', who made decisions that were 'good

enough' using a person's limited perceptual, cognitive, and informational resources. This concept has been called 'satisficing', in contrast with the (unrealistic) 'optimizing' of economic man. In his work in cognitive psychology, Simon viewed problem-solving as a complex emergent of simple decision-making mechanisms; and he viewed the cognitive acts of decision-making as problem-solving. In studying organizational behavior, he believed that one must study the collective decision-making of individuals: the nature of those processes, and the influences on those processes (e.g., informational and motivational).

Empiricism

Secondly, Simon was an empiricist. His theories, and the models formulated to test them, were based on real data from empirical studies or experiments. Although his theories were often elegant, he was much less interested in elegance than in veracity – whether the theory or model represents a reasonable and testable induction from real data.

Information-processing Languages and Artificial Intelligence

The third major theme was Simon's choice of a fundamental framework (and hence, a fundamental language) with which to model all processes of decision-making and cognition in individuals and organizations: the information-processing framework. This view was implicit in his rejection of 'economic man' as a model for decision-making. But in 1955, he specifically chose as his fundamental framework the information processes of the digital computer – hence, computer languages – as the modeling medium for his theories of decision-making. Indeed, his modeling of human cognition using techniques of computer simulation (which he also helped to pioneer) resulted in the launch of a new field of science and engineering, that of artificial intelligence (AI). He is considered one of the four founders of AI along with Allen Newell, Marvin Minsky, and John McCarthy.

PROBLEM-SOLVING

In 1955, and for several decades thereafter, Simon collaborated with Allen Newell on building information-processing models of complex human problem-solving. Their first model, known as 'logic theorist' (LT), proved theorems of logic in much the same way that they were proven by Whitehead and Russell in *Principia Mathematica*. Methodologically, Newell and Simon insisted that a model, to be complete, must 'run' as a computer program. LT first 'ran' in 1956. Immediately thereafter, Newell and Simon began work on their model of human chess-playing behavior, seeking to achieve world-class levels of chess competence on a computer. This goal was achieved only with the advent of the very fast computers of the 1990s.

Together, LT and the models of chess problem-solving explicated a set of concepts and mechanisms of cognition that formed the basis of the fields of cognitive science and artificial intelligence. These concepts and mechanisms in effect defined the 'mainstream' of those fields. They included: the concept of a problem space of possible solutions that must be searched; the concept of heuristics, both general and task-specific; and the use of heuristics to control the search of the problem space, including heuristics for generating plausible moves, pruning search paths, and limiting the search (i.e., 'satisficing').

Simon had anticipated this pioneering work of the mid-1950s with two seminal papers in which he presented (as verbal concepts, not yet computer programs) his theory of rational decision-making and his theory of the influence of the task environment on complex problem-solving. He later synthesized these ideas with his famous metaphor of the 'ant on the beach'. The ant's goal is to reach some distant food. Its path to get there – its behavior – seems intricate and complex. But most of the (apparent) complexity is in the grains of sand to be traversed, not in the ant's simple mechanism for homing in on food. Simon believed that the observed complexity of human problem-solving behavior arises from simple and general underlying search mechanisms that are applied to a complex task environment.

In 1957 and 1958, Newell and Simon began to build (again, in information-processing language as a computer program) their 'general problem solver' (GPS). Their intention was that GPS should be applicable in any task environment. The method they postulated for achieving problem goals was called 'means–ends analysis'. The problem-solver analyzed 'differences' between the current state and the goal state, and chose 'operators' to reduce these differences, until there were no differences remaining. The task environment was represented as a table of associations between operators and the differences they affected.

GPS was tested in a variety of tasks by a series of computer simulation experiments that generated detailed sequences of behaviors. The scientific

goal of Newell and Simon was to show that these sequences matched, step by step, down to a level of great detail, the sequences of behaviors observed in people doing the same tasks in controlled laboratory experiments. Their book on this work, *Human Problem Solving*, is regarded as a landmark of cognitive science and of psychology in the twentieth century.

LEARNING, RECOGNITION, AND RECALL OF VERBAL MATERIAL AND SYMBOLIC OBJECTS

Few aspects of human learning have been studied as rigorously and extensively, in breadth and in depth, as the memorization, recognition, and recall of simple verbal material. To control for effects of word meaning on memorizing, three-letter sequences (usually pronounceable but meaningless) are often used as verbal materials. The major empirically observed phenomena are highly reproducible; yet they depend critically on the verbal materials and experimental conditions. These phenomena are seen by many cognitive psychologists as 'basic' or 'elementary'. The combination of stable reproducibility (implying a simple underlying mechanism) and strong dependence on the 'task environment' of materials and experimental regimes) recalls Simon's 'ant on the beach' metaphor.

In 1956, Simon began a collaboration with Edward Feigenbaum to develop an information-processing model (expressed, of course, as a computer program) that would: offer simple mechanistic explanations of the major phenomena of verbal learning; predict behavior in new experiments; and account for the observed dependence on the task environment using only simple mechanisms. The resulting model was called EPAM (for 'elementary perceiver and memorizer'). It has since been under continuous development, and extension to additional phenomena, primarily by Simon and his collaborator Howard Richman.

The central mechanism of the EPAM model was a network of discriminations (i.e., tests that represent branch points) among the symbols being learned. This network is not given *a priori*. It develops as the learner makes errors, indicating the need for finer discriminations among the symbols to be learned. Only partial information is memorized when that is adequate for performance (once again, 'satisficing'), for example when performing recognition tasks or memorizing symbolic cues to stimulus–response associations.

EPAM used a small 'immediate memory' of about seven symbolic 'chunks'. EPAM is itself a model of what is now called 'working memory', as opposed to long-term semantic memory. In postulating mechanisms for EPAM, Feigenbaum and Simon did not give EPAM any specific mechanisms for forgetting, even though forgetting was omnipresent in the behavior of human subjects. Nevertheless, EPAM did forget in the usual ways observed among human subjects. Forgetting emerged from the combination of the ever-growing network of memorized material and the use of 'satisficing' in storing only partial information as a cue to link stimulus and response. Sometimes (and temporarily, in a predictable way) the cue became inadequate to invoke the correct response.

EPAM was tested by being used as an 'artificial' subject in traditional verbal-learning experiments; it exhibited several of the known stable phenomena. The variety observed in its behavior was a consequence partly of its use of strategies for allocating its time and attention, and partly of its use of many different adaptive mechanisms.

EPAM research of the 1980s and 1990s added mechanisms of auditory short-term memory, phenomena involving articulation in verbal learning, a model of how certain experts at memorization learn extremely long lists of digits, and studies of the effects of context in letter perception. During this period, Simon and Richman extended EPAM to new domains such as chess, visual processing, and categorization.

The EPAM model is the most comprehensive, rigorously developed, and tested model of elementary human verbal learning in cognitive science.

OTHER CONTRIBUTIONS TO COGNITIVE SCIENCE

After his Nobel Prize in 1978, Simon turned part of his attention to economics. Yet he continued to make significant contributions to cognitive science.

Expert Knowledge

What is it that makes an expert different from a novice in performing a task such as playing chess? Several years of work by AI researchers building 'expert systems' pointed to the much larger body of knowledge used by experts, rather than their superior reasoning skills.

Using his own experiments and those of others, Simon studied the question in terms of chess-playing behavior. He reported that in very short visual presentations of complex chess positions, experts were able to encode most of the important

features, whereas novices were able to encode few. Simon inferred that chess experts use learned patterns of important chess features with which they could quickly 'parse' a complex chess position. He estimated the number of those patterns at about 50 000, learned over a period of 10 years.

Scientific Discovery

How are scientific theories formed? By what information processes were the great early discoveries of science made? For example, how did Kepler discover that planets travel around the Sun in elliptical orbits? Simon collaborated with Pat Langley, Gary Bradshaw, and Jan Zytkow to construct models of how several of the significant early discoveries in science may have been made. Simon and his colleagues created computer programs that 'rediscovered' the early scientific principles from data available to the early scientists.

Protocol Analysis

From 1956 until the publication of their book in 1972, Newell and Simon developed methods, called protocol analysis, for analyzing and measuring the details of the behavior that subjects were verbalizing about their problem-solving, moment by moment. The methodology for this intricate and difficult analysis was further developed in a book by Simon and Anders Ericsson in 1984 (revised in 1993).

CONCLUSION

Simon was a great unifier of models of seemingly disparate phenomena. His body of work represents a synthesis of essentially simple abstractions about decision-making. His scientific genius was to cut across the many and varied phenomena, and the (often obscuring) behavioral detail, to the simple but essential model. Simon created works not only of great scientific insight and power but of scientific beauty.

Further Reading

Feigenbaum EA and Simon HA (1984) EPAM-like models of recognition and learning. *Cognitive Science* **8**: 305–336.

Langley P, Simon HA, Bradshaw GL and Zytkow JM (1987) *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, MA: MIT Press.

March JG and Simon HA (1958) *Organizations*. New York, NY: John Wiley.

Newell A and Simon HA (1956) The logic theory machine. *IRE Transactions on Information Theory* **IT-2**(3): 61–79.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Simon HA (1969) *The Sciences of the Artificial*. Cambridge, MA: MIT Press. [The Karl Taylor Compton lectures.]

Simon HA (1977) *Models of Discovery*. Boston, MA: Reidel.

Simon HA (1982) *Models of Bounded Rationality*. Cambridge, MA: MIT Press.

Simon HA (1991) *Models of My Life*. New York, NY: Basic Books.

Skinner, Burrhus Frederic

Introductory article

A Charles Catania, University of Maryland, Baltimore County, Maryland, USA

CONTENTS

B. F. Skinner (1904–1990), an American psychologist, provided the experimental foundations of contemporary behavior analysis and its applications. He interpreted verbal behavior in terms of those foundations, and he was outspoken about the differences between the methods of behavior analysis and those of cognitive psychology.

BIOGRAPHY

Born on 20 March 1904, in Susquehanna, Pennsylvania, Burrhus Frederic Skinner (later known mostly as B. F.) grew up when inventions such as those of Thomas Edison were changing life in small-town America. In his school days, Skinner built gadgets to manage his own behavior; one blocked his bedroom door with a sign until he hung up his pajamas (thereby avoiding his mother's intervention). At Hamilton College, in Clinton, New York, he majored in English, also taking courses in science and philosophy. He sent some short stories to Robert Frost after graduation. Encouraged by Frost's reply, Skinner took a year off from academic pursuits to try a career in writing. Giving it up on the grounds that he had nothing to say, Skinner called the time his Dark Year.

Having read John B. Watson, Ivan P. Pavlov, and Bertrand Russell, he turned from English to psychology and entered the doctoral program at Harvard University, where he began a series of experiments on the behavior of rats that led to more than two dozen journal articles and culminated in his 1938 book, *The Behavior of Organisms*. In 1936, he moved to the University of Minnesota, where he continued basic research and also published work on verbal behavior. Then the Second World War occasioned a project on training pigeons to guide missiles. It got only to the demonstration stage, but a fringe benefit was the discovery of shaping, the technique for creating novel forms of behavior through differential reinforcement of successive approximations to a response.

Another product of those days was the Aircrib, which Skinner built for his wife and second daughter. The windowed space with temperature and humidity control improved on the safety and comfort of ordinary cribs while making childcare less burdensome. Despite rumours to the contrary, it was not used for conditioning the infant.

In 1945, Skinner became Chair of the Department of Psychology at Indiana University. After his 1947 William James Lectures at Harvard University on verbal behavior, he returned permanently to Harvard. His 1948 novel, *Walden Two*, described a utopia the experimental character of which was its most important feature: unsatisfactory practices were to be modified until more effective substitutes were found.

The Harvard Pigeon Laboratory provided many of Skinner's students with opportunities to start independent lines of research while he developed reinforcement schedules in collaboration with Charles B. Ferster, revised and expanded his William James Lectures in the 1957 book *Verbal Behavior*, and built his first teaching machines. A Division for the Experimental Analysis of Behavior (Division 25) was established within the American Psychological Association, and the *Journal of the Experimental Analysis of Behavior* began publication in 1958. Within a decade, increased activity in applications led to the companion *Journal of Applied Behavior Analysis*.

Skinner retired from the laboratory in 1962, returning to it only briefly nearly 20 years later, but he continued his writing throughout his life. Among many papers and books were 'Why I am not a cognitive psychologist' in *Behaviorism* (1977), 'The origins of cognitive thought' in *American Psychologist* (1989), and a three-volume autobiography.

Skinner learned of his leukemia in 1989. On 10 August 1990, at his final public appearance, he accepted an award from the American Psychological Association for lifetime achievement in psychology. His remarks criticized cognitive science as the creationism of the twentieth century, in that it sought causes of behavior inside the

organism instead of in the organism's environment. A week later, in hospital, Skinner put the finishing touches to his last paper, 'Can psychology be a science of mind?', for the *American Psychologist*. He died the next day, 18 August 1990. His last word, upon receiving a drink of water, was 'Marvelous'.

THE ANALYSIS OF BEHAVIOR

Skinner followed Pavlov in insisting on the primacy of data and the study of individuals rather than groups, but diverged from Pavlov in many theoretical and empirical ways. For Skinner, behavior was not to be taken as a symptom of something else. As interaction between organism and environment, it should be studied in its own right, not to resolve problems of physiology or to open the way to cognitive or other levels of analysis. Skinner did not disapprove of physiology, but argued that without a science of behavior neuroscientists would not know what to look for in the nervous system.

In Pavlovian conditioning, the conditional stimulus reliably precedes the unconditional stimulus and comes to produce behavior related to the responses elicited by the unconditional stimulus. The prototypical example is the elicitation of salivation by some stimulus that consistently precedes food. In Skinner's operant behavior, the contingencies are different: a discriminative stimulus sets the occasion on which responses have some consequence; in the absence of the stimulus, responses do not produce that consequence. The prototypical example is the rat whose lever presses produce food in the presence but not the absence of a light. The rat comes to press the lever only when the light is present. Discriminative stimuli, colloquially called signals or cues, do not elicit responses; instead, they set the occasions on which responses have consequences. Such behavior, called *operant* because it operates on the environment, does not entail associations or stimulus–response connections. The three-term contingency, in which discriminative stimuli set the occasion upon which responding has consequences, is not reducible to pairwise stimulus and response relationships.

Skinner complained that he was sometimes misunderstood to be a stimulus–response theorist, but though some behavior sequences can be analyzed as chains in which each response produces stimulus conditions that occasion the next, he recognized that others must be integrated so that responses appear in proper order without depending on stimuli produced by earlier responses.

Admitting the possibility of behavior without eliciting stimuli was crucial. Behavior has causes other than eliciting stimuli. It was a profoundly simple concept: behavior now depends on its past consequences. Saying it another way: the consequences of current responses reinforce or select the responses that will occur later. Selectionism had replaced associationism.

Operants are classes of responses defined by their environmental effects rather than by topography (what they look like). Consider a rat, a lever, and a device for delivering food. If lever pressing does nothing, the rat presses only occasionally. If pressing produces food, the rat presses more often. The food is a *reinforcer*. The effectiveness of reinforcers changes over time (as in deprivation or satiation) and a given reinforcer might reinforce some responses but not others. Furthermore, the effects of reinforcement are not permanent: responding decreases to its earlier levels, or extinguishes, when reinforcement is discontinued.

The discovery that behavior could be maintained easily even when only occasional responses were reinforced led to schedules of reinforcement, which arrange reinforcers based on number of responses, the times when they occur, or various combinations of these and other variables. Different schedules produce different temporal patterns of responding.

Reinforcement operates on populations of responses within individual lifetimes much as evolutionary selection operates on populations over successive generations in Darwinian natural selection. (Skinner also considered implications of a third variety of selection, sometimes called cultural or memetic selection, that occurs when behavior is passed on from one organism to another, as in imitation.)

Reinforcement as selection is illustrated by the procedure called shaping, which creates novel behavior through reinforcement of responses that successively approximate it. For example, if the strongest of a rat's initially weak lever presses are reinforced, the force of pressing will increase and the criterion for reinforcement can be moved up to the strongest of the new population. With continuing increases in force and changes in criterion, the rat soon presses with forces that would never have been observed without shaping.

Shaping opened up education, developmental disabilities, and behavioral medicine to applications of behavior analysis. When research produced variable results, solutions were sought not by averaging over more subjects but by refining procedural details to identify sources of variability. The applied analysis of behavior is recognized for both

effectiveness and accountability; treatment of early autism is among its several notable successes.

Consequences operate in the context of other sources of behavior. Behavior arising from an organism's evolutionary history or phylogeny often interacts with behavior arising from its experience or ontogeny. Imprinting, which occurs when a duckling sees some moving object shortly after hatching, provides an example. The stimulus acquires special significance for the duckling, which then follows wherever it goes. Imprinted stimuli had been said to elicit or release following. But ducklings that must stand still to keep an imprinted stimulus visible learn to do so. In natural environments, following usually keeps ducklings close to mother ducks, but in other environments ducklings can learn to do different things. Genetic histories made ducklings capable both of learning and of becoming imprinted, but environmental contingencies determined what they learned to do.

VERBAL BEHAVIOR

The analysis of verbal behavior examines the functions of words. It differs from linguistics, the study of language, in that linguists describe practices of verbal communities in terms of the grammars, vocabularies, and phonetic units characterizing different languages. These descriptions of language structure tell little about their functions.

This behavioral distinction is analogous to that between physiology and anatomy in biology. For example, walking could be studied structurally by examining coordinations among the legs and other body parts. Analyses of particular muscle interactions and their extensions to running and other gaits might provide a grammar of movements that distinguished possible from impossible gaits. However, that grammar would not predict when someone might switch from standing to walking to running. Structure and function are different topics. Behavior analysis, following from Skinner's work, deals mainly with function, whereas cognitive science deals more often with structure (e.g. organization in what is perceived or learned).

The relevance is that Skinner's 1957 book, *Verbal Behavior*, was mainly about language function. A critical 1959 review by the linguist Noam Chomsky was more concerned with language structure than with the functional content of Skinner's account. *Verbal Behavior* provided a taxonomy of function rather than structure (for example, identifying verbal classes by their effects rather than by their topographies), applying it to a broad range of verbal phenomena. Later expansions extended the taxonomy to the origins of novelty in verbal behavior. Skinner consistently treated words as behavior rather than as vehicles for something else (a pervasive metaphor in everyday language is of words as carriers of ideas or meanings, making it hard to talk about sentences without referring to them as containing other entities).

One function of verbal behavior is instruction; people often do things because they are instructed to do them. Following instructions has social consequences and is crucial to many social institutions – families, schools, industry, the military – so it is important to understand not only how instructions work but also how they can go wrong (e.g. as in following unethical orders without question). Skinner called behavior that depended on words 'rule-governed behavior', though it is now more often called 'verbally governed behavior'. In verbal governance, what we say about what we do often determines what we do. In particular, one way to get people to do something is to shape not their doing but what they say about doing it. Contemporary analyses of verbal behavior include experimental studies of how these relations come about and the conditions under which they occur.

BEHAVING AND KNOWING (BEHAVIOR AND COGNITION)

Skinner's analyses were also about how we come to know ourselves. We think we have privileged access to private events such as feelings and thoughts, but how do we learn to talk about them? Parents who see the colors that a child sees can respond appropriately to the child's color-naming and so can teach the names, but how can a verbal community without access to relevant stimuli create and maintain verbal responses? In referred pain, a bad tooth in the lower jaw may be reported as a toothache in the upper jaw; here the dentist is a better judge than the patient of where the bad tooth really is. If we can be mistaken even about the location of a toothache, how can other reports of private events be reliable?

Skinner did not deny the private, but started from the fact that common vocabularies can be based only on what is mutually accessible to and therefore shared by speakers and listeners. If private feelings do not have public correlates, how can one tell when anyone else has them? If one cannot tell, how can one ever teach appropriate words? This is a problem because much of the language of cognitive thought originates in the vocabulary of private events.

According to Skinner, many processes called cognitive (e.g. thinking, visualizing) are kinds of behavior, but difficulties arise when they are invoked as explanations instead of as kinds of behavior in themselves. Skinner explicitly eschewed dualism, the distinction between mental and physical worlds. He did not deny events taking place inside the skin, but maintained that they should be called private rather than mental.

Skinner regarded internal causes as surrogates for external ones. Circularity is obvious if one characterizes individuals based on behavior (e.g. that person is greedy) and then uses the characterization as an explanation (e.g. he did it because of greed). Skinner argued that many cognitive explanations that invoke images or other internal representations, though more complex, are essentially of this sort and discourage further inquiry.

When Skinner criticized representations in cognitive psychology, the issue was not whether lasting effects are produced by stimuli (an organism that has responded to a stimulus is a changed organism). Rather, it was about the form the change takes. Skinner opposed copy theories of behavior or perception. A representation is not necessarily a copy (a spoken letter may represent a seen one but has no visual properties in common with it), so it is of interest that the most successful cognitive accounts, such as those in terms of parallel distributed processing, do not involve representations that function as copies. In this regard, behavior analysis shares its views with cognitive scientists who are advocates of neural nets and connectionist systems.

Further Reading

Catania AC (1992) B. F. Skinner, organism. *American Psychologist* **47**: 1521–1530.

Catania AC and Harnad S (eds) (1988) *The Selection of Behavior: The Operant Behaviorism of B. F. Skinner*. New York, NY: Cambridge University Press.

Chomsky N (1959) Review of B. F. Skinner's *Verbal Behavior*. *Language* **35**: 26–58.

Day WF (1969) On certain similarities between the philosophical investigations of Ludwig Wittgenstein and the operationism of B. F. Skinner. *Journal of the Experimental Analysis of Behavior* **12**: 489–506.

Lattal KA (ed.) (1992) Reflections on B. F. Skinner and psychology. *American Psychologist* **47**: 1269–1533.

Skinner BF (1938) *The Behavior of Organisms*. New York, NY: Appleton-Century-Crofts.

Skinner BF (1957) *Verbal Behavior*. New York, NY: Appleton-Century-Crofts.

Skinner BF (1969) *Contingencies of Reinforcement*. New York, NY: Appleton-Century-Crofts.

Skinner BF (1977) Why I am not a cognitive psychologist. *Behaviorism* **5**: 1–10.

Skinner BF (1989) The origins of cognitive thought. *American Psychologist* **44**: 13–18.

Skinner BF (1990) Can psychology be a science of mind? *American Psychologist* **45**: 1206–1210.

Skinner BF (1999) *Cumulative Record*, 4th edn. Acton, MA: B. F. Skinner Foundation.

Todd JT and Morris EK (1983) Misconceptions and miseducation: presentations of radical behaviorism in psychology textbooks. *The Behavior Analyst* **6**: 153–160.

Sperry, Roger

Introductory article

Colwyn Trevarthen, University of Edinburgh, Edinburgh, UK

Roger Wolcott Sperry received the Nobel Prize for Physiology and Medicine in 1981. Famous for experiments on the embryology and regeneration of functional nerve connections, on perception and learning in split-brain cats and monkeys, and on hemispheric modes of consciousness in human beings following commissurotomy, his belief that awareness and memory depend on the generation of motor images in the brain is now central in brain theory, important in understanding both self-awareness and communication.

Roger Wolcott Sperry (1913–1994) (Figure 1), who received a Nobel Prize for Physiology and Medicine in 1981 for his research on brain science, devoted his scientific work to a search for the neural circuits of action and consciousness. From an undergraduate psychology project at Oberlin College with R. H. Stetson (an expert on timing in speech and music), to appointment as Hixon Professor of Psychobiology at the California Institute of Technology and fame as a neuropsychologist who discovered new ways to reveal the mental functions of the cerebral hemispheres, Sperry was a meticulous experimenter on the growth and functions of neural systems. A quiet-spoken, taciturn man, he seemed to have the skill of a magician in making basic laws of brain activity apparent in a new form. He was also a plain-speaking thinker about the power of consciousness in guiding what we do, and he wrote passionately about the nature of human values. As a young man he had accepted the doctrine of the philosopher John Dewey that thought is a form of conduct, with corresponding responsibilities, and this belief remained with him throughout his life.

Sperry's first experiment, the results of which were published when he was 26 years old, used electromyography (recording the electrical signals of contracting muscles) to demonstrate the infinitely varied yet coherent neural instructions for a simple human movement, namely making a circle in the air in different planes with an extended arm. The results led to a concept of the 'motor image', which anticipated the classical work on the 'Coordination and Regulation of Movement' of the Russian physiologist Nicholas Bernstein, and to the following statement of Sperry's concept of the role of movement in generation of awareness:

> An objective psychologist, hoping to get at the physiological side of behaviour, is apt to plunge immediately into neurology, trying to correlate brain activity with modes of experience. The result in many cases only accentuates the gap between the total experience as studied by the psychologist and neural activity as analysed by the neurologist. But the experience of the organism is integrated, organised, and has its meaning in terms of coordinated movement. (Sperry, 1939, p. 295)

A contemporary behavioral scientist may feel that this text ought to be displayed as a caution in all experimental psychology laboratories and functional magnetic resonance imaging (fMRI) units, where immobile subjects ponder prefabricated thoughts or are excited by stimuli intended to arouse named categories of perception or emotion.

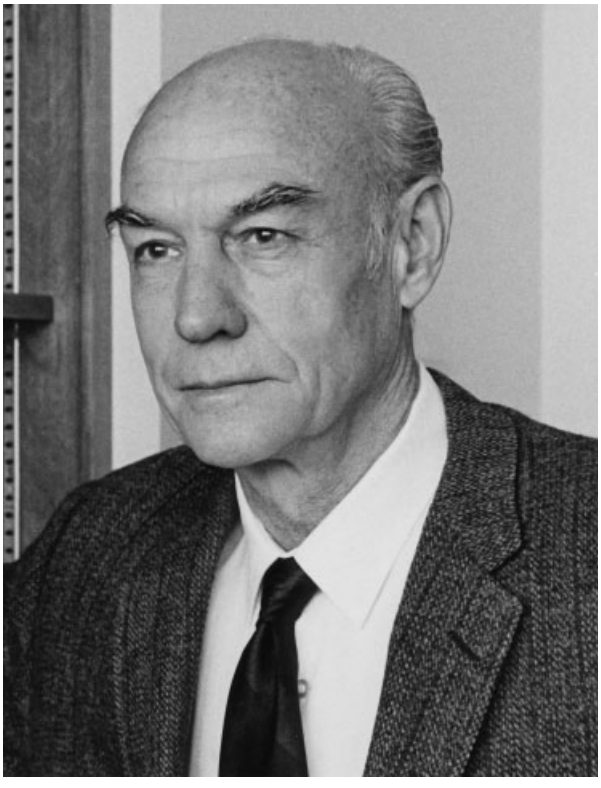


Figure 1. Portrait of Roger W. Sperry.

Sperry's graduate research, undertaken with Paul Weiss at the University of Chicago, initiated a series of experiments on plasticity of nerve–muscle systems and regrowth of motor nerves in rats and monkeys. Subsequently, his studies of the patterning of connections from eye to brain and from brain to muscles in fish and amphibia proved that chemical markers of some kind were guiding nerve axons to form the right connections. Experiments on the growth of eye-to-brain connections in fish and newts after surgical rotation of the eyes made Sperry world famous before he was 40 years of age (Figure 2(a)). They confirmed that the chemical markers would rebuild visual connections by inherent mapping principles, even when rearranged eye-to-body relationships produced totally nonfunctional reactions to stimuli. Visual areas map out the space for action in terms of body form automatically, without teaching from stimulation. This cast doubt on prevailing theories of the construction of functional brain circuits by learning. Sperry's 'chemoaffinity' principle is now accepted as being a primary factor in the early growth of nerve-cell connections. This is despite revolutionary advances in molecular biology and new models to explain order emerging in complex nonlinear dynamic systems of nerve cells, axonal and dendritic branches, and synapses, and selective retention of elements under the validation of environmental input. Emergent order and selection do create new functions and transform sensory and motor maps, but these maps retain a basic functional design of 'behavior space' – a product of a cell-to-cell communication that is set up in the embryo before the sensory nerves are connected to the brain.

Throughout his career, Sperry constantly tried to explain the creativity of 'experience', in the true sense of that word – that is, how conscious minds in active bodies 'try' to know. Cognitive psychology now favors analysis of human consciousness in terms of the data-transforming powers of intelligence – the processing and storage of information for mind work. Sperry had a deep conviction about and respect for the vitality and complexity of whole living organisms, and he emphasized the services of motor action to cognition. He attempted to determine how brains seek and generate information – not just logically processing what comes in, but how they create a useful sense of stimuli and how sensory information is taken in to monitor and guide intended movements. This interest led him to examine pathways within the visual cortex, and from visual areas to motor output. For every question Sperry found revolutionary answers. Step by step, he presented clear data on the living structure of cerebral pathways.

Mapping of skin or retina to the brain in a fish, a frog or a newt makes it possible for the animal to anticipate what will happen when it moves.

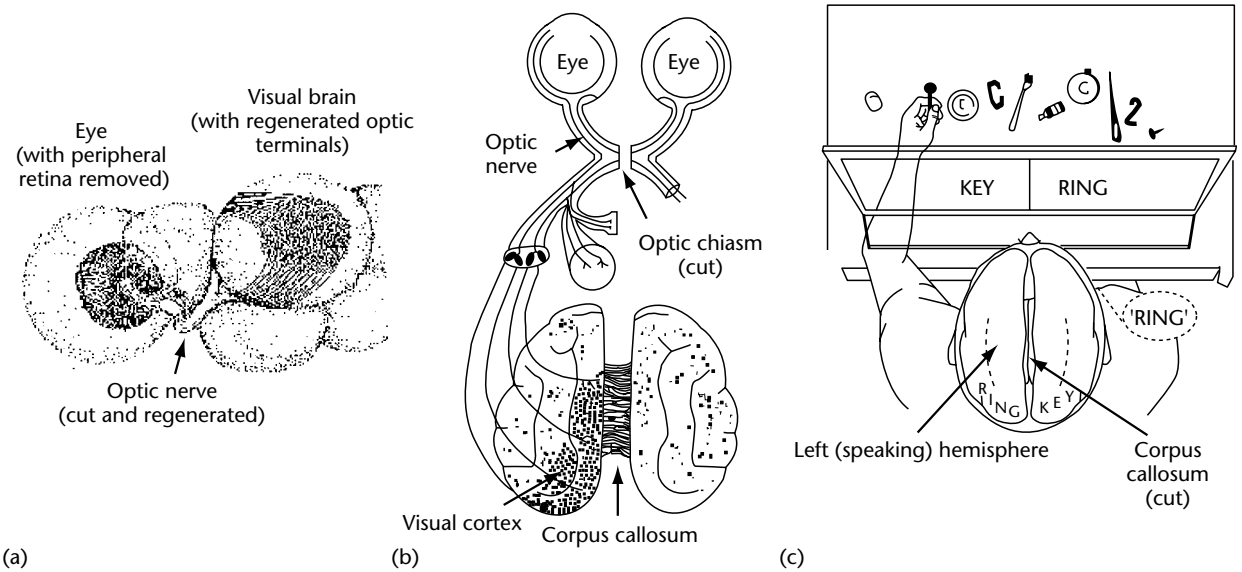


Figure 2. Drawings by Sperry showing three important experiments in the study of brain growth and function. (a) Selective regrowth of eye-to-brain connections in fish after removal of the peripheral parts of the retina and surgical section of the optic nerve. (b) Division of visual input to the cortices of the two cerebral hemispheres in cat brain. (c) Division of the corpus callosum in a human patient; the left hand (right hemisphere) feels a key, but the patient says (with left hemisphere) that he sees the word 'ring'.

Sperry's nerve regrowth experiments that were designed to determine how the maps are made are simple and elegant. The surgical skill necessary to operate under a binocular microscope on the nerves of animals only a few centimeters in length gave Sperry the means to conduct a brilliant study of the cortical mechanisms of vision in a mammal. He sliced through loop connections just beneath the gray matter of the visual cortex, making the incisions with tiny specially shaped knives through holes in the blood-vessel-rich membrane that covers the cerebral cortex. By showing that operated cats could perform difficult visual discriminations with little loss after this surgery, he disproved a theory of 'psychoneural isomorphism', that integration of perceptions was by local linking of features to make a neural analogue of a photographic image. He found that visual pattern recognition was mediated by deep cortico-cortical loops which the surface surgery left intact. This was one demonstration in a series that he designed to highlight the involvement of longer connections and separate brain parts in the unity of sight. A visual 'image' in the brain is not like a photograph or movie – it is more like a rumour spread by telephone calls between distant points.

Sperry's research on coordination of movements and awareness led him to investigate how intentions relate to consciousness in cats, monkeys, and humans. The 'split-brain' studies which eventually earned him his Nobel Prize were started in the University of Chicago, and continued at the California Institute of Technology after Sperry moved there in 1954. They germinated from an idea that was implanted in Sperry's mind in the 1940s while he was working with Karl Lashley at Yerkes Laboratory of Primate Biology in Florida. The aim was to find out how consciousness could be related to brain pathways. As Lashley interpreted his own extensive experiments, there was no clear location of learning and no function for the millions of fibers that bridge the gap between the two halves of the cortex in the corpus callosum, the largest single fiber tract in the brain. In Sperry's attack on this question, visual input to the cortices of the two cerebral hemispheres was divided by cutting the half of optic fibers that cross the midline under the brain in the optic chiasm (Figure 2(b)). Ronald Myers and Sperry showed that visual learning in a cat could be doubled by transecting both the optic chiasm and the corpus callosum, thereby producing the 'split brain'. A major advance in understanding how brain connections serve in awareness, and how the cerebral cortex works with brainstem structures that were not divided by the surgery, was made possible. Eventually it led Sperry to direct inventive testing of the effects of the operation, by dividing the corpus callosum in human patients, tests that demonstrated the different consciousness of the two human hemispheres (Figure 2(c)).

In 1960, the neurosurgeon Joseph Bogen suggested to Sperry that an improved understanding of the human mind could be obtained by methods based on those that had been perfected at the California Institute of Technology in the late 1950s for elucidating visual and touch learning in split-brain monkeys. The equivalent human operation, known as commissurotomy, had been used by neurosurgeons to prevent the interhemispheric spread of intractable epilepsy, but no effects on consciousness were reported. Most probably the tests had been inappropriate, or the surgery was incomplete. A new series of operations was medically justified for individuals with severe seizures, and better-controlled testing, like that developed for the studies on animals, could aid the post-surgical care of these patients. The first paper on surgically divided awareness in humans was published by Bogen and Sperry with Michael Gazzaniga in 1962, nine years after Myers and Sperry's first report on split-brain learning in cats. There followed a rich series of tests of the differences and interactions between images and memories, feelings and intentions in the left and right cerebral cortices of a small and devoted group of patients who were grateful for the improvement in their lives, and for the attention that they received in collaboration with Sperry's team.

Apart from firmly linking aspects of consciousness to identified parts of the brain, the results of experiments by researchers who joined Sperry at the California Institute of Technology clarified the functions of the brain areas involved in language and thinking. They proved that intentional 'mind sets' (anticipating the type of strategy that would be needed for the subject to carry a conscious act of any kind to a successful conclusion) were allocating particular territories of the cerebral cortex, directing attention and the information-processing of perception, and preparing the body to move in precisely directed ways, using intention, memory, and emotions to generate a receptive consciousness.

The California Institute of Technology commissurotomy studies stimulated a new era in neuropsychology of the hemispheres, and accelerated the testing of left-brain and right-brain mental functions in normal individuals. In Sperry's mind the findings generated a new line of philosophical inquiry and a new teaching venture that kept him hard at work until his death in April 1994. He

believed that the 'cognitive revolution' of the 1970s was inspired by the research on consciousness in commissurotomy patients, and he continued to spare no effort in explaining his thinking, despite increasing disability due to a serious central motor disorder that weakened him and progressively restricted his freedom of action and participation in academic meetings. This infirmity in no way diminished the acuteness of his thought or the fluency of its expression.

Roger Sperry's thoughts on the causal potency of consciousness were developed after 1965 in a series of philosophical papers on 'Mind, Brain and Humanist Values'. His scientific colleagues, many of whom were still persuaded that consciousness must be a notion beyond the reach of rigorous science, were puzzled. Was he becoming mystical? Surely he was being philosophically naive. However, anyone who had crossed logical swords with Roger Sperry knew the danger of assuming that he could be naive. He was too thorough and tenacious to miss a weakness in his own or another's argument. On the other hand, he was stubborn, and the notion of consciousness and its innate values as a causal force in nature, as a director of human experience and knowledge, was consistent with his original phenomenological beliefs, and it did become an abiding conviction. Now, of course, his message seems far less deluded, and more a prophetic realization of how universal 'autopoetic' or self-creative emergent principles must apply (and with great force) in the consciously monitored activities of the human brain and mind, directing not only the elementary physiology of the neurons in patterned arrays, but also the culturally contrived mechanisms of society in harmony with ancient requirements for human life on earth. And it is not so odd, in the twenty-first century, to say that if we do not find lasting values to guide our collective actions, things may go badly for our societies and for the world that coexists with our dangerously increasing numbers. In that context, Sperry's thinking on global causality and the place of the human mind and human conscience in the making of a liveable future has increased validity.

New notions of causality propounded by physicists with regard to the origins of the cosmos and the evolution of matter give scientific authority to the nonreductive psychology that Sperry had developed as a biologist who was aware of the special creativity of neural systems. However, physics can provide only the most general, abstract enlightenment about psychological processes. What Sperry set out to do was to challenge the powerful orthodoxy of psychometrics, experimental psychology, and reductive brain science that seeks linear explanations for the construction of behavior and consciousness, all deriving from the input of information from sources that have no intrinsic values and no moral force. As he challenged behavioristic notions of learning with his demonstrations of innate forces in the growth of nerve connections to form functional networks, and as he insisted that surgical operations could locate motives for learning in animals and show the real anatomy of human consciousness and its creative subdivisions, so he concluded that human values are not just convenient fictions with short-term validity. Rather, he believed that they are rooted in inherited psychobiological principles for acting in harmony with the physical and social world in which humans have evolved.

Sperry's writings are lucid, not least when his ideas challenge accepted theory. His later writings are motivated by strongly formed personal beliefs, but they remain a clearly articulated call for new scientific thinking. He always advocated keeping 'the big picture' in mind.

Further Reading

Evarts EV (1990) Foreword. Coordination of movement as a key to higher brain function: Roger W. Sperry's contributions from 1939 to 1952. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*, pp. xiii–xxvi. New York, NY: Cambridge University Press.

Hunt RK and Cowan WM (1990) The chemoaffinity hypothesis: an appreciation of Roger W. Sperry's contributions to developmental biology. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*, pp. 19–74. New York, NY: Cambridge University Press.

Levi-Montalcini R (1990) Ontogenesis of neural nets: the chemoaffinity theory, 1963–1983. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*, pp. 3–18. New York, NY: Cambridge University Press.

Levy J (1990) Regulation and generation of perception in the asymmetric brain. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*, pp. 231–248. New York, NY: Cambridge University Press.

Sperry RW (1939) Action current study in movement coordination. *Journal of General Psychology* **20**: 295–313.

Sperry RW (1950) Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology* **43**: 483–489.

Sperry RW (1952) Neurology and the mind–brain problem. *American Scientist* **40**: 291–312.

Sperry RW (1961) Cerebral organization and behavior. *Science* **133**: 1749–1757.

Sperry RW (1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Sciences of the USA* **50**: 703–710.

Sperry RW (1982) Some effects of disconnecting the cerebral hemispheres (Nobel lecture). *Science* **217**: 1223–1226.

Sperry RW (1983) *Science and Moral Priority*. New York, NY: Columbia University Press.

Sperry RW (1992) Paradigms of belief, theory and metatheory. *Zygon* **27**: 245–259.

Sperry RW (1993) The impact and promise of the cognitive revolution. *American Psychologist* **48**: 878–885.

Sperry RW, Gazzaniga MS and Bogen JE (1969) Interhemispheric relationships. The neocortical commissures: syndromes of hemisphere disconnection. In: Vinken PJ and Bruyn GW (eds) *Handbook of Clinical Neurology*, vol. 4, pp. 273–290. Amsterdam, Netherlands: North Holland.

Trevarthen C (1990) Editor's preface. Roger W. Sperry's lifework and our tribute. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*, pp. xxvii–xxxvii. New York, NY: Cambridge University Press.

Teuber, Hans-Lukas

Introductory article

Mary Brown Parlee, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston, Massachusetts USA
George Adelman, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston, Massachusetts USA

CONTENTS

Hans-Lukas Teuber (1916–1977) was a neuropsychologist known for his research on the effects of brain injuries on human sensation, perception, language, and memory. He founded the MIT Department of Psychology (now Brain and Cognitive Sciences) in 1964, an early instance of brain and behavioral scientists working together in a single graduate department in the US.

INTRODUCTION

Best known for his work on perception and his behavioral research on brain-injured World War II veterans, Hans-Lukas Teuber arrived at MIT in 1960 to establish its first department of Psychology. He organized the department around his vision of a science of brain and behavior that would emphasize the biological and physiological aspects of the field rather than the more traditional aspects of learning, psychophysics, personality, and social psychology. Through his organizational skills, and working with a faculty he selected that shared his view of a unified field of brain and behavior research, he was able to define and study problems and questions that still lie at the heart of the cognitive and brain sciences. The department he created was a prototype of many of the new 'neuroscience' departments that would be established at universities and medical schools throughout the United States.

LIFE AND WORK

Hans-Lukas Teuber was born in Berlin in 1916, and educated at the College Français, a Huguenot school where all subjects were taught in French. He studied Greek, Latin, ancient history and the natural sciences, receiving his baccalaureat in 1934. At home his parents Rose (Knopf) and Eugen Teuber stimulated his interests in music, literature, animal behavior, and mathematics. His father had studied under Wilhelm Wundt, and in 1913 had set up the primate research center at Tenerife (Canary Islands) where Wolfgang Köhler, one of the founders of Gestalt psychology and the next director of the primate center, carried out his famous experiments on intelligence in anthropoid apes. It was in the early 1930s in Berlin that Hans-Lukas Teuber first met another of his parents' friends, Kurt Goldstein, and became aware of Goldstein's studies of brain-injured German veterans of World War I.

Teuber entered the University of Basle in Switzerland in 1935. Continuing his search for integrative principles, which he later said was shaped by his father's example and disparate interests, Teuber studied philosophy of science, chemistry, biology and zoology, comparative anatomy, and embryology. Lectures by the embryologist Hans Spemann were particularly influential, awakening his interest in the possibility of using concepts and methods of experimental embryology to study central nervous system functions. At Basle Teuber participated in small interdisciplinary group discussions of the methodology of various sciences and how to bridge the gap between biological and social science. Here he met a fellow student (of art history), Marianne Liepe, and they soon made plans to marry.

At Kurt Goldstein's suggestion Teuber applied for the Holtzer Fellowship to study psychology at Harvard University. Liepe moved to the United States in 1939 to continue her study of art history at Vassar College. Teuber's entry to the United States was delayed until 1941 by the outbreak of World War II. Immediately on arrival, he married Liepe, and began his studies in psychology at Harvard. He and Liepe became naturalized US citizens in 1944.

While a graduate student Teuber supported himself and his family (Andreas Wolfgang was born in 1942, Christopher Lawrence in 1946) by working as a research assistant on a project on the effectiveness of psychological interventions – counseling of various kinds – in preventing juvenile delinquency. His Ph.D. dissertation grew out of this research (*Dyadic Groups: A Study in Counseling Relationships*) and was written under the direction of the social psychologist Gordon Allport. Teuber later said that his work on this project had been valuable in convincing him of the importance of control groups and of quantitative measures of behavior. His enthusiasm for social psychology as a science, however, was never great.

According to Teuber's later recollections, the most important influences on his thinking during his years at Harvard were Karl Lashley and Kurt Goldstein (Goldstein was a visiting professor at Harvard when Teuber entered). Through them he became acquainted with the work of W. B. Cannon, L. J. Henderson, and J. W. Gibbs. Teuber thought that Gibbs's work in physical chemistry might be applicable to both biological and social systems. Frequent personal contacts with Goldstein convinced Teuber, as he later wrote, of the strategic role of experimental neurology within the framework of general biological science, and suggested a reconsideration of the earlier German work (Bethe, von Uexküll, Weiss) in comparative physiology of nervous systems and problems of sensorimotor integration.

In 1944 Teuber interrupted his graduate studies to enter the US Navy, which classified him as a 'pharmacist's mate' and assigned him to the San Diego Naval Hospital. There Teuber met Morris Bender, who headed the neurology wards and was interested in sensory disturbances and causalgia following cerebral injuries. Teuber's background in experimental psychology and his knowledge of Goldstein's work led Bender, a gifted clinical neurologist, to propose a collaboration. The two were soon developing experiments on behavioral consequences of brain injuries sustained by servicemen in battles in the Pacific. Several papers on alterations in visual perception resulted. Teuber regarded their research as continuing in the tradition of Goldstein and Gelb and also of Poppelreuter, Head, and Holmes, who had studied brain-injured soldiers from World War I; he was interested not simply in the presence or absence of particular 'symptoms' associated with different sites of injury but in their scientific significance for a general understanding of normal behavior and brain function.

Teuber returned to Harvard in 1946 and completed his dissertation. He then moved to the Psychophysiological Laboratory, which he and Bender established at the Bellevue Psychiatric Hospital in New York City; Teuber became its director when Bender left to head the Neurology Department at the Mount Sinai College of Medicine. From 1948 to 1960 Teuber held faculty appointments at New York University's College of Medicine and at the Graduate School of Arts and Science, where he taught courses in neuroanatomy, physiological psychology, and social psychology of small groups. He was appointed Area Consultant to the US Veterans Administration in 1948 and received funding to study World War II veterans with injuries of the brain from penetrating missiles. In the same year he was invited to participate in the Macy Foundation's conferences on cybernetics, where a small interdisciplinary group (including Norbert Wiener, Warren McCulloch, John von Neumann, and Margaret Mead) met annually to see whether concepts from cybernetics, information theory, and computer theory could be developed into a common, transdisciplinary framework for integrating the social and natural sciences.

At the American Psychological Association meeting in Boston in 1948 Teuber gave a paper laying out his vision of a new field forming 'in the border region of psychology and neurology'. Using examples from his research with Bender, Teuber called this border region 'neuropsychology', a field in which the fundamental question is how does neural structure 'mediate' psychological functions, and in which the behavior of brain-injured patients plays a strategic role ('experiments of nature') as a source of clues to normal brain–behavior functions. This vision guided Teuber's empirical research in the Psychophysiology Laboratory at New York University, and would later be realized in the psychology department he founded at MIT.

As head of the Psychophysiological Laboratory, Teuber, and the research group he gathered around him (Sidney Weinstein, Lila Ghent Brain, Josephine Semmes, Mortimer Mishkin, William Battersby, Rita Rudel, and others), continued to recruit and test World War II veterans with brain injuries and a control group with peripheral nerve injuries. They devised quantitative testing procedures to assess somatosensory, visual, visuospatial, and cognitive impairments associated with injuries in different brain sites. In addition to setting high standards for methodological rigor in behavioral testing, research from the Psychophysiology Laboratory demonstrated the importance of testing large

populations of patients and of using research designs that permitted clear interpretations of relationships between behavioral deficits and site of injury. ('Double dissociation of symptoms' – demonstration of contrasting sets of behavioral symptoms with contrasting cerebral lesions – was one of Teuber's off-repeated methodological maxims.)

Research from the laboratory appeared in a steady stream of papers during the 1950s on: brain injuries and alterations in visual and visuospatial phenomena (extinction, completion, pattern perception, critical flicker frequency, eye movements, and visual searching); alterations in performance on complex visual tasks and in body schemata; judgments of visual and postural vertical; recognition of objects by touch; somatosensory thresholds; and performance on intelligence tests. Some of this work was summarized in two monographs published in 1960, *Visual Defects after Penetrating Missile Wounds of the Brain* (Teuber, Battersby, and Bender) and *Somatosensory Changes after Penetrating Brain Wounds in Man* (Semmes, Weinstein, Ghent, and Teuber). The latter has been credited with changing traditional concepts of hemispheric asymmetry in the mediation of somatosensory performance. Other work brought then-widely-accepted distinctions between specific and general behavioral effects of lesions into question and challenged widely held beliefs about effects of brain injuries on IQ test performance.

Teuber's theoretical views evolved with his empirical work, and he published two particularly influential synthetic reviews of the literature while at the NYU Psychophysiological Laboratory. One, 'Physiological psychology' (in the 1955 *Annual Review of Psychology*) was an overview which one leading neuroscientist and historian of neuroscience (Charles Gross) later said set the program of the field for the next decade. The other was the chapter on 'Perception' in the 1960 *Handbook of Physiology: Neurophysiology*, *vol 3* in which Teuber comprehensively reviewed research on classic problems of perception and located them in the broader context of sensory-isolation studies, ethological research, studies of behavior following brain injuries, and experimental embryology.

Drawing on work by von Holst and Sperry, Teuber developed throughout the 'Perception' chapter a critique of the assumption, on which both field theories and scanning theories of perception rely, that brain correlates of perception can be understood by viewing the nervous system as a passive receiver of sensory information. By contrast he wrote: 'Throughout this chapter we have stressed the potential role of a central corollary discharge which is postulated as coordinating efferent and afferent processes. This corollary discharge presumably travels from motor into sensory systems at the onset of every [self-initiated] bodily movement and thus permits anticipatory adjustment of the perceptual process … It can be seen that these concepts are closely related to the earlier neurologic postulates of "schemata" as the neural basis for awareness of posture and spatial orientation … .' Within such a theoretical framework, the distinction between voluntary and involuntary motor activity becomes important, while differences among perceptual, sensorimotor, and visuospatial processes become less so.

In 1960 Teuber was invited to MIT to head the Psychology Section of the Department of Economics and Social Science and to plan an independent doctorate-granting Department of Psychology at the Institute. Teuber seized the opportunity to organize his department according to his vision of a science of behavior and the brain and to bring together on the faculty people who shared his vision.

The first two senior appointments, together with Teuber himself, indicated the scope and direction of the department Teuber envisioned. Richard Held, an experimental psychologist from Brandeis University, had recently demonstrated through a series of ingenious experiments the importance of voluntary movement in developing and maintaining perceptual–motor coordination. Walle Nauta, a neuroanatomist then at the National Institutes of Health, had discovered a silver-staining technique that identified axonal degeneration following experimental brain lesions and permitted tracing of pathways in the brain. (The presence of a neuroanatomist, especially one as distinguished as Nauta, in a psychology department at a time when behaviorist learning theory dominated most psychology departments, was unprecedented.) Teuber's own work on perception in brain-injured adults continued at MIT, and his interest in ethology (again unusual for the time) encouraged his students (beginning with Robert Yin) to investigate the perception of human faces.

The junior faculty Teuber assembled used a variety of behavioral, physiological, and anatomical techniques to investigate a variety of psychological topics: memory, eye–hand coordination, space perception, visual backward masking, perceptual–motor development in animals reared in sensory isolation, learning, language, and others. What unified research in the MIT department and

distinguished it from other psychology departments at the time was its focus on behavior in relation to brain function. Investigations often involved both behavioral neurophysiological or neuroanatomical techniques; even when only behavioral data were gathered, they were usually interpreted in terms of hypotheses concerning underlying neural processes.

Early faculty in the department included Joseph Altman, Stephen Chorover, Peter Schiller, Alan Hein, Wayne Wickelgren, Merrill Garrett, Jerry Fodor, and Whitman Richards (the first student to receive a Ph.D. from the department). They were later joined by Gerald Schneider and Ann Graybiel (both of whom also received their degrees from the department), Molly Potter, Sue Carey-Block, and Emilio Bizzi. Before Teuber's untimely death in 1977 he had made plans to add David Marr to the faculty. Visiting scientists, research associates, postdoctoral fellows, and graduate students also contributed to the MIT Psychology Department's distinctive brain–behavior focus, including Charles Gross, Helen Mahut, David Ingle, Harvey Karten, Donald Stein, Thomas Bever, Suzanne Corkin, Donald Pfaff, and Larry Squire. Several generations of undergraduate students at MIT were exposed to Teuber's very popular introductory psychology course, in which he propounded his vision of a new kind of psychology in elegant, witty lectures.

Although the pace of his empirical work slackened at MIT, Teuber was much sought-after as a speaker and discussant at conferences because of his wide-ranging knowledge and ability to identify key unanswered questions. Throughout his career, Teuber was an active participant in international societies (e.g., International Neuropsychological Symposium, International Brain Research Organization, European Brain and Behavior Society, French Psychological Society) as well as North American organizations (e.g., American Psychological Society, Psychonomic Society, Society for Neuroscience). He played an important role in introducing the work of Alexander R. Luria and other Russian neuropsychologists to their counterparts in North America and Europe (Teuber and Luria visited each other's laboratories in the 1950s); his preface to the English translation of Luria's *Higher Cortical Functions in Man* (1966) thoughtfully analyzed Luria's contributions in the context of other national traditions of neuropsychological research. In 1970 Teuber was elected to associate membership in the Neurosciences Research Program, an interdisciplinary program founded by the MIT biophysicist Francis O. Schmitt to bring together world experts in the many disparate sciences and clinical disciplines involved in understanding how the nervous system mediates behavior. Teuber's broad knowledge of neurobehavioral research and his vigorous interaction with his fellow Associates helped strengthen the organization and its impact on the emerging new field of neuroscience.

While at MIT, Teuber wrote numerous syntheses of research on brain and behavior, many drawing on research in the department to exemplify and promote the new field he was working to institutionalize at MIT. In 1969 the Alfred P. Sloan Foundation launched its program in the neurosciences (the first programmatic funding for the new field) and made its first award (and largest overall) to the MIT Psychology Department, designating it a 'center of excellence' not only because of the strength of the department's faculty and staff but because of its unusually integrated approach. Teuber's creativity and effectiveness as an institution-builder were reflected in the many honors and awards he received before he died in a drowning accident in 1977 at the age of 61. He was a member of the Society of Experimental Psychologists (1960), the American Academy of Sciences (1962), the National Academy of Sciences (1972), the French Neurological Society (1968), and the National Institute of Neurology Faculty, Mexico (1967). He received the Karl Spencer Lashley Award for Research in Neurobiology (1966), the Apollo Achievement Award, NASA (1969), the Kenneth Craik Award in Experimental Psychology (1971), MIT's James R. Killian Faculty Achievement Award, Oxford University's Eastman Professorship, and honorary degrees from the Université Claude Bernard (Lyon, France) and the Université de Genève (Switzerland).

Teuber told an interviewer in 1966 that he attributed to his father's influence his own tendency not to feel at home in any single field, and his career demonstrates his powerful drive to synthesize research results and to bring disparate methods to bear on fundamental questions about behavior in relation to brain function. The name he gave to the new field he envisioned changed (largely for rhetorical reasons) several times during his lifetime: psychology, neuropsychology, psychophysiology, brain and behavior, behavioral biology, psychology and brain sciences. The unvarying thread was the inclusion of some term referring to an organism's behavior. It would probably have pleased him that the MIT Psychology Department is now called the Department of Brain and Cognitive Sciences.

Further Reading

Gross CG (1994) Hans-Lukas Teuber: a tribute. *Cerebral Cortex* **4**: 451–454.

Hecaen H (1979) H. L. Teuber et la fondation de la neuropsychologie experimentale. *Neuropsychologia* **17**: 119–124.

Held R (1979) Hans-Lukas Teuber. *Neuropsychologia* **17**: 117–118.

Hurvich LM, Jameson D and Rosenblith WA (1987) Hans-Lukas Teuber, 1916–1977. In: National Academy of Sciences, *Biographical Memoirs*, vol. LVII, pp. 461–490. Washington, DC: National Academy Press.

Milner B and Teuber H-L (1968) Alteration of perception and memory in man: reflections on methods. In: Weiskrantz L (ed.) *Analysis of Behavioral Change*, pp. 268–375. New York, NY: Harper & Row.

Parlee MB (2002) In Memoriam: Hans-Lukas Teuber. In: Stringer AY, Cooley EL and Christensen DL (eds) *Pathways to Prominence: Reflections on 20th Century Neuropsychologists*, pp. 77–98. Hove, UK: Psychology Press.

Pribram KH (1977) Hans-Lukas Teuber: 1916–1977. *American Journal of Psychology* **90**: 705–707.

Richards W (1978) Obituary: H-L Teuber, 1916–1977. *Vision Research* **18**: 357–359.

Semmes J, Weinstein S, Ghent L and Teuber H-L (1960) *Somatosensory Changes After Penetrating Brain Wounds in Man*. Cambridge, MA: Harvard University Press.

Teuber H-L (1955) Physiological psychology. *Annual Review of Psychology* **6**: 267–296.

Teuber H-L (1960) Perception. In: Field J, Magoun HW and Hall VE (eds) *Handbook of Physiology: Neurophysiology*, vol. III, pp. 1595–1688. Washington, DC: American Physiological Society.

Teuber H-L (1975) Effects of focal brain injury on human behavior. In: Tower DB (ed.) *The Nervous System*, vol. II: *The Clinical Neurosciences*, pp. 457–480. New York, NY: Raven Press.

Teuber H-L, Battersby WS and Bender MB (1960) *Visual Field Defects After Penetrating Missile Wounds of the Brain*. Cambridge, MA: Harvard University Press.

Weinstein S (1985) The influence of Hans-Lukas Teuber and the Psychophysiological Laboratory on the establishment and development of neuropsychology. *International Journal of Neuroscience* **25**: 277–288.

Tolman, Edward C.

Introductory article

Nancy K Innis, University of Western Ontario, London, Ontario, Canada

CONTENTS

American psychologist Edward C. Tolman (1886–1959) was author of the learning theory known as purposive behaviorism. Tolman is also noted for introducing important psychological concepts, such as the intervening variable and the cognitive map, which are widely used today.

INTRODUCTION

Edward Chace Tolman was an American psychologist who is perhaps best known today for introducing the idea of a cognitive map. He did so on 17 March 1947, when he delivered the 34th Annual Faculty Research Lecture at the University of California. The occasion celebrated his long and distinguished career in the Psychology Department at Berkeley. The title of his talk was 'Cognitive Maps in Rats and Men', and the concept caught on.

EARLY BACKGROUND

Edward Tolman, the son of Mary Chace and James Pike Tolman, was born in West Newton, Massachusetts, an affluent suburb of Boston, on 14 April 1886. He attended the excellent Newton public schools and then, following family tradition, enrolled at the Massachusetts Institute of Technology (MIT). James Tolman had been in the first graduating class at MIT and was a member of the Board of Trustees, and Edward's elder brother, Richard, was a graduate student there at the time Edward enrolled. Although their father wanted his sons to take over his prosperous cordage business, they both opted for academic careers. Richard would become an eminent physical chemist, spending most of his career at the California Institute of Technology.

After obtaining his undergraduate degree in electrochemistry from MIT in 1911, Edward entered the doctoral program at Harvard. His humanitarian interests, and a disinclination to compete directly with Richard, led him to the new science of psychology, a discipline that was beginning to apply experimental findings to human problems. He received his PhD in 1915 for studies of human memory, under the supervision of Hugo Munsterberg, Director of the Harvard Psychology Laboratory. Although Edward became an experimental, rather than an applied, psychologist, he always retained a strong interest in the human condition. Towards the end of his career he began applying his theoretical principles to personality and social psychology.

In 1915, Tolman accepted his first academic position, as an Instructor at Northwestern University in Evanston, Illinois. Here he continued his research on human memory. In 1918, his contract was not renewed, apparently because of wartime cutbacks; but Tolman always believed that his association with a pacifist student periodical led to his dismissal. However, that summer he was offered an appointment at the University of California, Berkeley, where he would remain for the rest of his career. It was a good move for Edward, both personally and academically. In California, he lost his New England reserve, revealing the warm, fun-loving nature that colleagues would always associate with him. His research interests changed too, and he began to study animal behavior. His first major project with rats was on the inheritance of the ability to learn. This marked the beginning of a long-term research program in behavior genetics at Berkeley. Initiated by Tolman, this work was continued under the direction of his graduate student, and later colleague, Robert Tryon. Tolman turned to more theoretical pursuits.

A NEW FORMULA FOR BEHAVIORISM

Just as Tolman was starting out in psychology, a new theoretical system – behaviorism – was

beginning to take root in America. Led by a young researcher, John Watson, behaviorists rejected the subjective approach of psychologists who used the method of introspection to study unobservables such as consciousness. Instead, they took the more objective position that psychology was the study of behavior. Their research examined the overt responses of subjects, often animals, to environmental stimuli, using classical conditioning methods. This led to their being labeled stimulus–response (S–R) psychologists. Tolman's interest in behaviorism was stimulated initially in a course at Harvard taught by comparative psychologist Robert Yerkes. Tolman liked the fact that this system was scientific and objective. However, he was never able to accept whole-heartedly the brand of behaviorism promoted by Watson.

Tolman believed that behavior was more than simple reflex reactions to stimuli. In 1922, he introduced his own 'new formula for behaviorism' with the aim of providing a scientific treatment of concepts, such as motive and purpose, that had been rejected as subjective and mentalistic by other behaviorists. His approach reveals the influence of two other Harvard professors, Edwin Holt and Ralph Barton Perry, both of whom had written about such concepts as purpose and cognition as objective terms. Over the next ten years, Tolman published articles providing objective definitions for emotions, ideas, and consciousness, as well as for purpose and cognition. He also supervised a number of students whose research with rats in mazes provided support for his theoretical position. He brought theory and data together in a book, *Purposive Behavior in Animals and Men*, published in 1932. Tolman's system of psychology, with its emphasis on the goal-directed nature of behavior, became known as *purposive behaviorism*.

PURPOSIVE BEHAVIORISM

In *Purposive Behavior* Tolman identified four factors that played a role in producing behavior – stimuli, heredity, training, and physiological state – now referred to as independent variables. Intervening between these causal factors and the observed behavior (the dependent variable) were motivational and cognitive behavior-determinants – purposes and cognitions. Tolman suggested that purposiveness was revealed in the behavior itself, by the fact that responding persisted until the goal was achieved. Cognitions were expectations about the relationship between environmental stimuli, referred to as 'signs', and the goals that they indicated. Two types of expectations were identified. Long-term expectations that depended on genetics or past experience (training), he called *means–end-readinesses*. Others, that were specific to the ongoing situation providing information about how to achieve current goals, were labeled *sign-gestalt-expectations*. Tolman used a lot of hyphenated terms like these to refer to his constructs, often making his theory difficult to understand. Soon the simpler term *expectation* began to be used by learning theorists to refer to both types of cognitions. However, Tolman believed that it was important to distinguish between them. A similar kind of distinction has been made more recently by comparative cognition researchers who talk about reference and working memory.

Tolman also attempted to provide an objective definition of conscious awareness or attention with the idea of *behavior-adjustments*. These were responses by means of which the animal sampled its environment before making a response. For example, if these were overt, a rat might wiggle its nose from side to side before choosing to run down an arm of the maze. Later these responses were labeled *vicarious-trial-and-error* (VTE) behavior and became an important feature of Tolman's only attempt to quantify his theory with his schematic sowbug model. The whimsical term merely indicated the tropistic nature of an organism's movement toward or away from a goal.

INTERVENING VARIABLES

Not long after his book was published, Tolman introduced the term *intervening variable* to refer to the motivational and cognitive determinants of behavior that he had identified. For Tolman, a theory was no more than a set of intervening variables, and the task of the psychologist was to operationally define them. This could be accomplished by means of standard experiments in which an independent variable correlated with the construct being studied is systematically varied while all others are held constant. Tolman referred to his revised position as 'operational behaviorism' because it used operational definitions and because the behavior being observed by the researcher involved an organism acting on – operating on – its environment.

Tolman's book and subsequent theoretical articles were well received by his colleagues, and in 1937 he was elected President of the American Psychological Association (APA). In his presidential address, after comparing his brand of behaviorism with that of the currently more popular

S–R theorists, such as Clark Hull, Tolman outlined his system for his colleagues. He concluded with the surprising statement that 'everything important in psychology … can be investigated in essence through the continued experimental and theoretical analysis of the determiners of rat behavior at the choice point in a maze'. In his talk he had shown in detail how operational behaviorism could account for such choice-point behavior. But, of course, he saw that his system had broader implications. Soon he turned his attention to its application to humans.

COGNITIVE MAPS IN RATS AND MEN

Tolman brought his ideas about the behavior of rats and humans together in his Faculty Research Lecture at Berkeley. Again he compared his theory of learning with that of the S–R psychologists, this time using compelling metaphors. In contrast to the 'telephone-switchboard' type of connections proposed by his S–R rivals, for Tolman learning involved 'something like a field map of the environment [being] established in the rat's brain'. The suggestion of a mental map in the brain was the first time that Tolman gave a direct physiological referent to one of his intervening variables. The map metaphor was very appropriate since much of the data supporting Tolman's position had come from studies of rats in mazes.

Two types of cognitive maps – 'broad and comprehensive' and 'narrow and striplike' – were identified. On mazes, rats with broad maps responded to the general location of the goal rather than learning a specific route to it; those with narrow maps became fixated on a particular route. Narrow maps, which were less adaptive, were established as a result of brain damage or when conditions were highly motivating or frustrating.

At the end of his Berkeley lecture, Tolman suggested that human social maladjustments could be interpreted as narrowings of cognitive maps as a result of strong motivation or frustration. He hoped that psychologists could help establish a society in which frustrations and motivations were modulated, one that would produce broad cognitive maps, and thereby appropriate psychological adjustment. Over the next few years, many of his articles were devoted to extending his theory to human social learning.

Tolman's own cognitive map was a broad map. He was eclectic in his approach, and open-minded and tolerant in his treatment of the ideas of others.

PERSONALITY AND SOCIAL PSYCHOLOGY

Tolman had a real affinity for personality and social psychology, and his later writings reflect this interest. He had first applied his ideas to human behavior during World War II. As a pacifist, he despised war and violence. In 1933–4, during a sabbatical year in Vienna, he witnessed with considerable dismay Hitler's rise to power. On returning to the USA he worked to help émigré psychologists fleeing from Europe to find positions in America. He was also involved with the Society for the Psychological Study of Social Issues (SPSSI), a group of psychologists committed to applying their expertise to problems arising from the economic depression and the looming possibility of war. In his address as Chairman of SPSSI in 1940, Tolman introduced ideas that were presented at greater length in a book, *Drives Toward War*, published in 1942. Although it started out as a motivation textbook, the major aim of the book was to identify mechanisms for harnessing human aggression to prevent future wars between nations. It illustrated Tolman's strong belief that basic psychological research could provide solutions to social problems. He would always remain optimistic about the role of psychology in meliorating the human condition.

THE FIGHT FOR ACADEMIC FREEDOM

Edward Tolman was a staunch defender of academic freedom. As the Communist scare swept America in the late 1940s, the Regents of the University of California demanded an anti-Communist oath of loyalty from the faculty. Tolman was one of the first to speak out against this violation of academic freedom and became leader of the Group for Academic Freedom, formed to oppose the oath. In 1950, he and about 20 others, who had been fired from the university for refusing to sign the oath, took the Regents to court. After a long struggle they eventually won their case and were reinstated. The oath controversy was widely reported in the media, and Tolman's principled stand gained him the respect of professors throughout the USA. The fight over the oath, however, took a personal toll on Tolman, and his health began to deteriorate. He died at his home in Berkeley on 19 November 1959.

Further Reading

Innis NK (1992) Tolman and Tryon: early research on the inheritance of the ability to learn. *American Psychologist* **47**: 190–197.

Innis NK (1992) Lessons from the controversy over the loyalty oath at the University of California. *Minerva* **30**: 337–365.

Innis NK (1998) Edward C. Tolman's purposive behaviorism. In: O'Donohue W and Kitchener R (eds) *Handbook of Behaviorism*, pp. 97–118. New York, NY: Academic Press.

Tolman EC (1922) A new formula for behaviorism. *Psychological Review* **29**: 44–53.

Tolman EC (1932) *Purposive Behavior in Animals and Men*. New York, NY: Century Co.

Tolman EC (1936) Operational behaviorism and current trends in psychology. *Proceedings of the 25th Anniversary Celebration of the Inauguration of Graduate Studies*, pp. 89–103. Los Angeles, CA: University of Southern California.

Tolman EC (1938)The determiners of behavior at a choice point. *Psychological Review* **45**: 1–41.

Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* **55**: 189–210.

Tolman EC (1951) *Behavior and Psychological Man*. Berkeley, CA: University of California Press.

Tolman EC (1952) Edward Chace Tolman. In: Boring EG, Langfeld HS, Werner H and Yerkes RM (eds) *A History of Psychology in Autobiography*, vol. 4, pp. 323–339. Worcester, MA: Clark University Press.

Turing, Alan

Introductory article

B Jack Copeland, University of Canterbury, Christchurch, New Zealand

CONTENTS

Alan Turing was a British mathematical logician who pioneered computer science, cognitive science, artificial intelligence, and artificial life.

OVERVIEW OF TURING'S WORK

The mathematical logician Alan Mathison Turing OBE (1912–1954) contributed to logic, mathematics, cryptanalysis, philosophy, mathematical biology, and formatively to computer science, cognitive science, artificial intelligence (AI), and artificial life. He was elected a Fellow of King's College, Cambridge in 1935 and in 1936 published his most important work, 'On computable numbers, with an application to the Entscheidungsproblem'. From 1936 to 1938 he studied for a PhD at Princeton University, returning to King's in 1938. At the outbreak of war with Germany (September 1939) he moved to Bletchley Park, the wartime headquarters of the Government Code and Cypher School (GC&CS). At Bletchley, Turing single-handedly broke Naval Enigma and was the principal designer of the Bombe, a large-scale electromechanical machine for revealing Enigma message keys by a process of high-speed search. From 1945 Turing worked at the National Physical Laboratory (NPL) in London, pioneering computer science and AI, and from 1948 at the University of Manchester as Deputy Director of the Computing Machine Laboratory (there was no Director), taking up a specially created Readership in the Theory of Computing in 1953.

RELEVANCE AND ROLE OF TURING'S WORK IN COGNITIVE SCIENCE

The Universal Turing Machine

As everyone who uses a personal computer knows, the way to make the machine perform some desired task is to open the appropriate program stored in the computer's memory. But the earliest large-scale electronic digital computers, the British Colossus (1943) and the American ENIAC (1945), did not store programs in memory. To set up these computers for a fresh task, it was necessary to modify some of the machine's wiring, rerouting cables by hand and setting switches. The basic principle of the modern computer – controlling the machine's operations by means of a program of coded instructions stored in the computer's memory – was thought of by Turing in 1935. His abstract 'universal computing machine', as he called it – it would soon become known as the universal Turing machine (UTM) – consists of a limitless memory in which both data and instructions are stored, in symbolically encoded form, and a scanner that moves back and forth through the memory, symbol by symbol, reading what it finds and writing further symbols. By inserting different programs into the memory, the machine is made to carry out different computations. The UTM (described in 'On computable numbers') was the first formal model of computation. Turing's idea of a universal stored-program computing machine was promulgated in the US by John von Neumann and in the UK by Maxwell Herman Newman. By 1945 groups in both countries were attempting to build an electronic stored-program universal digital computer: a Turing machine in hardware. (*See* **Computation, Formal Models of**)

The Church–Turing thesis

Before the advent of computing machines, a computer was a human being: a mathematical assistant whose task was to calculate by rote, in accordance with a systematic method supplied by an overseer. Like a filing clerk, the computer might have little detailed knowledge of the end to which the calculations were directed. The Church–Turing thesis (advanced independently by Turing and the American Alonzo Church in 1936) says in effect that the UTM can perform any calculation that can be done

by an idealized human computer (who lives forever and never runs out of paper and pencils).

The Turing machine and the brain

According to a prominent theory in cognitive science, the brain is a Turing machine, in the sense that any information processing of which the brain is capable can be done by the UTM. A stronger version of the theory holds that the brain is fundamentally similar to the UTM, in that both are 'symbol systems': machines that process formal symbols by means of a fixed repertoire of basic computational operations. (*See* **Symbol Systems**)

Uncomputability

In 'On computable numbers' Turing proved that some well-defined mathematical problems are uncomputable, in the sense that they cannot be solved by the UTM – and so, according to the Church–Turing thesis, cannot be solved by a human being working by rote. An example is the following. Some Turing machines print '1' at some stage in their computations. All the remaining Turing machines never print '1'. Consider the problem of deciding, given an arbitrary Turing machine, which of these two categories it falls into. Turing showed that this problem cannot be solved by the UTM.

Turing developed these ideas further while at Princeton, laying the foundations of the branch of mathematical logic that investigates and codifies problems 'too hard' to be solvable by the UTM, and introducing the concept of an oracle machine (or *o*-machine) – a 'new kind of machine' able to solve problems that the UTM cannot. In modern cognitive science, *o*-machines form the basis for a controversial new class of models of cognition termed 'hypercomputational'.

Computer Pioneer

Turing's technical report 'Proposed electronic calculator', dating from the end of 1945 and containing his design for the 'Automatic Computing Engine' (ACE), was the first relatively complete specification of an electronic stored-program digital computer. The earlier 'First draft of a report on the EDVAC' (May 1945), written in the US by von Neumann (familiar with the UTM since before the war), discussed at length the design of an electronic stored-program digital computer, but in fairly abstract terms, saying little about programming, hardware details, or even electronics. Turing's report, on the other hand, contained specimen programs in machine code, full specifications of hardware units, detailed circuit designs, and even an estimate of the cost of building the machine (£11,200). Had Turing's ACE been built as he planned, it would have been in a different league from the other early computers, but his colleagues at the NPL thought the engineering work too difficult to attempt and a considerably smaller machine was built. Known as the 'Pilot Model ACE', this machine ran its first program on 10 May 1950. With a clock speed of 1 MHz, it was for some time the fastest computer in the world. Computers derived from Turing's ACE design remained in use until about 1970, including the Bendix G-15, arguably the first personal computer. (Delays beyond Turing's control resulted in the NPL's losing the race to build the world's first stored-program electronic digital computer. That honour went to Newman's Computing Machine Laboratory at Manchester, where the 'Manchester Baby' ran its first program on 21 June 1948.)

Artificial Intelligence

Turing was the first to carry out substantial research in the field of AI. He was thinking about AI at least as early as 1941, and during the war circulated a typewritten paper on machine intelligence among his colleagues at GC&CS. Now lost, this was undoubtedly the earliest paper in the field of AI. It probably concerned machine learning, heuristic problem-solving, and machine chess, topics that Turing discussed extensively at GC&CS. His thinking on AI was probably influenced by his work on the Bombe, which involved him in the design of, for example, a 'majority vote gadget', mechanizing the process of evaluating a hypothesis on the basis of unreliable data, and a 'Ringstellung cut-out', an early example of the use of heuristics (rules of thumb) to constrain search.

In February 1947 Turing gave the earliest known public lecture to mention computer intelligence, providing an exciting glimpse of a new field. He discussed the prospect of machines acting intelligently, learning from experience, and beating humans at chess.

Intelligent machinery

In 1948 Turing wrote a report for the NPL entitled 'Intelligent machinery'. Described by an NPL bureaucrat as 'not suitable for publication', this farsighted paper was the first manifesto of AI. In it Turing gave a wide-ranging and imaginative survey of the prospects of AI and brilliantly introduced many of the concepts that were later to

become central in the field, in some cases after reinvention by others. These included the logic-based approach to problem-solving, now widely used in expert systems, and the idea, subsequently made popular by Allen Newell and Herbert Simon, that (as Turing put it) 'intellectual activity consists mainly of various kinds of search'. (The idea of solving a problem by means of a guided search through the space of possible solutions was central to the Bombe.) Turing anticipated the concept of a genetic algorithm in a brief passage concerning what he called 'genetical or evolutionary search'. Genetic algorithms employ methods analogous to the processes of natural evolution in order to produce successive generations of software entities that are increasingly fit for their intended purpose. 'Intelligent machinery' also contains the earliest description of (a restricted form of) what Turing was in 1950 to call the 'imitation game' and is now known simply as the Turing test. In 1952 he said of the imitation game that 'you might call it a test to see whether the machine thinks, but it would be better to avoid begging the question, and say that the machines that pass are (let's say) "Grade A" machines', and predicted that it would be at least 100 years before computers would stand a chance at the imitation game (with no questions barred). (*See* **Turing Test**)

Can machines think?

In 1950 Turing famously remarked that this question is 'too meaningless to deserve discussion'; however, his later accounts of the imitation game (in 1951 and 1952) reveal a milder attitude and contain liberal use of such phrases as 'programming a machine to think' and 'making a thinking machine'.

Gödelian arguments against AI

Turing was among the first to write about the Gödelian arguments against AI (which originated with Emil Post in 1921). Turing discerned a simple fallacy in such arguments. In 'Intelligent machinery' he wrote: (*See* **Artificial Intelligence, Gödelian Arguments against**)

> The argument from Gödel's and other theorems … rests essentially on the condition that the machine must not make mistakes. But this is not a requirement for intelligence.

In 'Proposed electronic calculator' (the earliest surviving written statement of Turing's views concerning machine intelligence), Turing wrote:

> There are indications however that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess.

The risk that the machine will produce incorrect results is the price of heuristic search.

Neural Simulation

Probably Turing's earliest surviving mention of his interest in neural simulation is in a letter to the cyberneticist W. Ross Ashby: 'In working on the ACE I am more interested in the possibility of producing models of the action of the brain than in the practical applications to computing.' The major part of 'Intelligent machinery' consists of an exquisite discussion of neural simulation and machine learning, in which Turing anticipated the modern approach known as connectionism. Donald Hebb and Frank Rosenblatt are widely regarded as the founders of connectionism and it is not widely realized that Turing had outlined much of the connectionist project as early as 1948. He introduced what he called 'unorganized machines', giving as examples networks of neuron-like elements connected together in a largely random manner; he described one type of network as 'the simplest model of a nervous system', and hypothesized that 'the cortex of the infant is an unorganized machine, which can be organized by suitable interfering training'. From a historical point of view, his idea that an initially unorganized neural network can be organized by means of 'interfering training' is of considerable significance: this idea did not appear in the earlier work of McCulloch and Pitts. In Turing's model, the training process renders certain neural pathways effective and others ineffective. He anticipated the modern procedure of using an ordinary digital computer to simulate neural networks and the process of training them. He claimed a proof (now lost) of the proposition that an initially unorganized network with a sufficient number of neurons can be organized to become a universal Turing machine with a given storage capacity. This proof raised the possibility, noted by Turing, that the human cognitive system may be a universal symbol processor implemented in a neural network. (*See* **Learning through Case Analysis**; **Neurons, Computation in**; **McCulloch–Pitts Neurons**; **Perceptron**)

Artificial Life

During his final years Turing pioneered artificial life. He used the Manchester Ferranti Mark I (the first commercially sold electronic stored-program

computer) to model biological growth, and he appears to have been the first person to engage in the computer-assisted exploration of nonlinear dynamical systems. He programmed the computer to simulate a chemical mechanism by which the genes of a zygote may determine the anatomical structure of the resulting animal or plant. (*See* **Artificial Life**)

Further Reading

Carpenter BE and Doran RW (eds) (1986) *A. M. Turing's ACE Report of 1946 and Other Papers*. Cambridge, MA: MIT Press.

Copeland BJ (2000) The Turing test. *Minds and Machines* **10**: 519–539.

Copeland BJ (ed.) (2003) *The Essential Turing: Core Papers in Philosophy, Logic, Artificial Intelligence and Artificial Life, Plus the Secrets of Enigma*. Oxford, UK and New York, NY: Oxford University Press.

Copeland BJ and Proudfoot D (1996) On Alan Turing's anticipation of connectionism. *Synthese* **108**: 361–377. [Reprinted in: Chrisley R (ed.) (2000) *Artificial Intelligence: Critical Concepts in Cognitive Science*, vol. II 'Symbolic AI'. London: Routledge.]

Copeland BJ and Proudfoot D (2000) What Turing did after he invented the universal Turing machine. *Journal of Logic, Language, and Information* **9**: 491–509.

Erskine R and Smith M (eds) (2001) *Action This Day*. London: Bantam Books.

Gottfried T (1996) *Alan Turing: The Architect of the Computer Age*. Danbury, CT: Franklin Watts.

Herken R (ed.) (1988) *The Universal Turing Machine: A Half-Century Survey*. Oxford, UK and New York, NY: Oxford University Press.

Hodges A (1983) *Alan Turing: The Enigma*. London: Burnett.

Turing AM and Copeland BJ (ed.) (1999) Posthumously published lectures. In: Furukawa K, Michie D and Muggleton S (eds) *Machine Intelligence* **15**: 381–475. Oxford, UK and New York, NY: Oxford University Press.

Tversky, Amos

Introductory article

Maya Bar-Hillel, The Hebrew University, Jerusalem, Israel

CONTENTS

Amos Tversky is best known for his pathbreaking work, along with Daniel Kahneman, on heuristics and biases in judgment under uncertainty, and for prospect theory, *a theory of decision under uncertainty. His research showed that people are 'irrational', namely, their judgments and decisions deviate in systematic ways from normative dictates.*

BIOGRAPHY

Amos Tversky was born on 16 March 1937, in Haifa, Israel, his parents' second child; his mother was a social worker (later to become a Member in Israel's first Parliament), and his father was one of the newly emerging State of Israel's first veterinarians. Tversky served as an officer in the paratroops regiment of the Israeli army, winning Israel's highest honor for bravery. As a student at The Hebrew University of Jerusalem he majored in psychology and philosophy, and after graduation in 1961 enrolled in Michigan University's mathematical psychology program. In 1965 he wrote an award-winning dissertation under the supervision of Clyde H. Coombs.

At Michigan he met future collaborators, such as Robyn Dawes, Ward Edwards, Dave Krantz, and Paul Slovic. There he also met his wife-to-be, Barbara (née Gans, later a Professor of Psychology at Stanford University). He spent a year as Fellow at Harvard University's Center for Cognitive Studies, after which he returned to The Hebrew University, serving on its faculty as Professor of Psychology until 1978. It was there that Tversky began his extraordinary collaboration with Daniel Kahneman, which was to last over a quarter of a century and yield some of the most influential and innovative ideas in the area of judgment and decision-making (JDM).

Soon after his return to Israel, the 1967 Six Day War broke out, and Tversky fought in it. Subsequently he also fought in the 1973 October War. During 1970–2 he spent a year as Fellow at Stanford's Center for Advanced Studies and a year at the Oregon Research Institute. In 1978 he moved to Stanford University, where he was the inaugural Davis-Brack Professor of Behavioral Science and Principal Investigator at the Stanford Center on Conflict and Negotiation until his untimely death on 2 June 1996. Throughout his years at Stanford, he maintained very close ties with Israel, visiting regularly a couple of times each year. In 1984/5 he spent a year at The Hebrew University, and from 1992 was Senior Visiting Professor of Economics and Psychology and Permanent Fellow of the Sackler Institute of Advanced Studies at Tel Aviv University. Many of his students themselves became Professors of Psychology in universities across the world.

Tversky won many awards during his short life: Distinguished Scientific Contribution Award of the American Psychological Association (1982); McArthur Prize (1984); Warren Medal of the Society of Experimental Psychologists (1995). He received Honorary Doctorates from the University of Goteborg (1985), State University of New York at Buffalo (1987), University of Chicago (1988), and Yale University (1994). He was a member of the Society of Experimental Psychologists (1979), the American Academy of Arts and Sciences (1980), the National Academy of Science (1985), and the Econometric Society (1993).

CONTRIBUTIONS

Probability Judgments

Until the early 1970s, the prevalent view of how people judge uncertainty was that they followed normative dictates to a good first approximation, although they were limited, fallible, and imperfect in doing so. Comparing notes from their experience in teaching college statistics, Tversky and Kahneman noted some common statistical errors among their students (e.g. insufficient appreciation for the role of sample size), as well as among their more sophisticated colleagues (e.g. insufficient attention

to sample error) which led them to discard this view of 'Man as an intuitive statistician'. They proposed an alternative picture whereby people intuitively and spontaneously replace judgments of probability or of frequency by quite different judgments, which are more cognitively natural and easy for the human mind to perform. Such a cognitive strategy is called a *heuristic*, and two were put forth.

First, when relying on representativeness, 'an event is judged probable to the extent that it represents the essential features of its parent population or generating process'. A famous example is Linda, who is '31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.' Most people think Linda is more likely to be a feminist bank-teller than a bank-teller – even though, of course, all feminist bank-tellers are bank-tellers. This normative violation is known as the *conjunction fallacy*. It follows from, and indicates that, people fail to consider the normative logical constraints their judgments should obey, instead following the heuristic judgment that Linda sounds more like a feminist – even a feminist bank-teller – than like a prototypical bank-teller.

Second, when relying on availability, people judge the likelihood of an event by the ease with which it can be brought to mind, whether by mental construction or by retrieval from memory. Consider people's inclination to think that there are fewer English words whose penultimate letter is N than words whose final letters are ING.

In 1974, the heuristics and biases work was published in *Science*, and became Tversky and Kahneman's single most cited paper. It also gave its name to a leading school within the area of JDM, and helped turn JDM from a marginal area in cognition into one of its fastest growing areas. Shortly thereafter, Tversky and Kahneman expanded into the study of decision-making, described in the following section.

In his later years, Tversky developed his *support theory*, positing that probabilities are attached not to actual events in the world, but to descriptions or mental representations thereof, called 'hypotheses'. A hypothesis' probability depends on the relative strength of its supporting evidence versus evidence supporting its alternative. For example, unpacking an event such as 'dying an unnatural death' into its constituents (e.g. road accident, homicide, suicide, drowning, other) greatly increases its perceived support, hence its perceived likelihood, by bringing to mind overlooked constituents and highlighting others. Hence, the notion of unpacking a hypothesis, which changes an event's description without changing the event, accounts for how the sum of the estimated probabilities of an event's components can exceed that of the event itself.

Decision and Choice

Tversky had begun to make his mark in both empirical and formal aspects of decision and choice already in his Michigan days. He had developed an elegant mathematical model to capture the notion that when faced with a multitude of options, we often reduce the possibilities by successively eliminating those options which are inferior on some aspect important to the choice. But it was within his remarkable collaboration with Kahneman that the full impact of his thinking came to fruition. In 1979, they published their seminal paper on prospect theory. The paper appeared in *Econometrica*, and at the time of Tversky's death this paper by two cognitive psychologists was the most cited paper of all time (1703 citations) to appear in this most prestigious of economics journals. One reason for this remarkable fact is the paper's popularity outside of economics – not only among cognitive and social psychologists, but also among jurists, political scientists, physicians, and more. Indeed, Kahneman and Tversky's work is arguably the most influential contribution of cognitive psychology to neighboring disciplines in the second half of the twentieth century.

Prospect theory (PT) offered a descriptively adequate alternative to subjective expected utility theory (SEU), the normative cornerstone in economic theory of individual decision-making under uncertainty. In SEU, utilities are derived from 'rational' (namely, consistent) choice, and the utility of a 'gamble' (namely, an uncertain choice) is the sum of the products of the utility of each of its possible final outcomes by the (subjective) probability of this outcome's occurrence. PT kept the form of the utility function, but substituted each of its components by a more psychologically realistic counterpart.

Final outcomes (also known as 'states of wealth') were replaced by gains or losses, namely by changes in final states. More accurately, the carriers of utility were posited to be descriptions, or 'frames', of changes in wealth as gains or as losses (recall the similar, though chronologically later, idea in support theory). Thus, they range from $-$ to $+$, not just from 0 to $+$ as in SEU. The 0 point is either the actual status quo or some frame-induced reference point. People's attitudes to risk are 'reflected' around this point, so that

where they are risk-averse in one domain, they are risk-seeking in the other, and vice versa. Probabilities, on the other hand, are subjected to a function which imposes on them decision weights.

This was a radical conceptual change, as there exists no mental operation that automatically converts changes of states into final states, whereas probabilities, even when objectively given, are distorted by the decision weights. That our decision intuitions are not based on the decision's impact on final wealth is immediately apparent: we usually have a poor idea, if any, of just what our state of wealth is at any given moment. Moreover, the intuitions of people whose wealth differs by orders of magnitude are often highly similar (e.g. almost everybody, rich or poor, would reject a gamble which offers even odds of gaining or losing $1000). That our reponse to probabilities is not linear can be demonstrated by the following classical example. Imagine you are playing a game of Russian roulette, in which you must fire a gun loaded with k live bullets and 6 − k empty chambers. You can deduct one live bullet by paying some large amount of money. When would you be more inclined to pay: if the gun has a single live bullet, or if it has four live bullets? The drop in the probability of death is the same – 1/6 – in both cases, but the greater willingness to pay in the former case than in the latter follows from the shape of the decision weights function. Altogether, at objective probabilities of 0 and 1 interesting discontinuities arise, which have no parallel elsewhere in the probability range. These are known as the *certainty effect*.

The dramatic effect which framing can have on decision is demonstrated in the following notorious problem. A virulent strain of flu is expected to hit next winter, and could claim 600 lives. Two alternative programs, A and B, have been proposed to combat it. If A is adopted, 200 of those lives can be saved (400 of those lives will be lost). If B is adopted, there is a 1/3 chance of saving all 600 lives, but a 2/3 chance of saving none (there is a 1/3 chance of nobody dying, but a 2/3 chance all 600 will die). Typically, most people prefer A under the description outside the brackets, but prefer B under the description within the brackets. This example clearly shows that people do not have a canonical representation of final outcomes, as the final outcomes described outside of and within the brackets are identical. They do not even have a canonical representation of gains and losses, as those, too, are identical within and outside the brackets. Rather, when given a pseudo 'gain-frame' (the outcomes are described in terms of lives saved), they passively accept it, exhibiting risk-aversion (preferring the sure-thing), and when given a pseudo 'loss-frame' (outcomes described in terms of lives lost), they become risk-seeking (reject the sure-thing). It is violations of invariance such as this that most clearly prove that a normative utility theory can never be descriptively adequate, because no normative theory can afford to give invariance up.

Another powerful notion to have come out of PT is that of 'loss aversion' (or: 'losses loom larger than gains'). To wit, for most people, offsetting a possible loss of $X requires a possible gain, at equal odds, considerably larger than $X, and often as large as $2X. Loss aversion has become a popular and useful tool in analyzing real-life situations outside the psychological laboratory. Consider, for example, the difficulty of resolving international conflicts by give and take. Suppose a seemingly symmetrical negotiation around arms reduction. An 'honest broker' believes that one of my ABCD missiles has roughly equal impact, in terms of arms reduction, to one of my enemy's MNOP missiles. But 'losing' one of mine can only be offset, for me, by 'gaining' about two of the enemy's – whereas for my enemy it is the other way around.

Additional Work

Tversky made substantial contributions in two additional areas, loosely related to his pathbreaking work in JDM – the axiomatic foundations of measurement, and the study of similarity judgments.

His feature-based theory of similarity challenged many common assumptions about similarity. Thus, Tversky posited that similarity need not be symmetrical: the son resembles the father more than the father resembles the son. Similarity also need not be the complement of dissimilarity: the USA and Britain may be simultaneously more similar to each other and more dissimilar from each other than North Korea and South Korea. His theory predicts that weights are given to features in a manner that is sensitive both to the task (similarity versus dissimilarity) and to which of A and B is the subject and which the object of the judgment.

His experiments and examples in this area are characterized by the same clarity, perceptiveness, and wit which characterize the examples shown before, and which helped popularize his work by their sheer appeal and catchiness. At the same time, his mathematical modeling is characterized by a unique style of using formal thinking to clarify psychological ideas. Even people with little mathematical sophistication can follow his reasoning

and benefit from the formal treatment, which invariably highlights essential conceptual distinctions and somehow cuts nature at the correct joints.

Tversky had a passion about 'getting things right'. Accordingly, his publications are notable not so much for their number as for their impact (at the time of his death, he had published more papers in *Psychological Review* than any previous author). His work will no doubt have a lasting, if not dominating, influence on JDM.

Further Reading

Coombs CH, Dawes RM and Tversky A (1970) *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman D and Tversky A (2000) *Choices, Values, and Frames*. Cambridge, UK: Cambridge University Press.

Kahneman D, Slovic P and Tversky A (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Krantz D, Luce RD, Suppes P and Tversky A (1971) *Foundations of Measurement*, vol. 1: *Additive and Polynomial Representations*. New York, NY: Academic Press.

Tversky A (1972) Elimination by aspects: a theory of choice. *Psychological Review* **79**: 281–299.

Tversky A (1977) Features of similarity. *Psychological Review* **84**: 31–48.

Tversky A and Koehler DK (1994) Support theory: a nonextensional representation of subjective probability. *Psychological Review* **101**: 547–567.

von Neumann, John

Introductory article

Arthur W Burks, University of Michigan, Ann Arbor, MI, USA

CONTENTS

John von Neumann (1903–1957) was an interdisciplinary thinker who not only brought fresh approaches to seemingly disparate fields, but discovered and developed their interrelationships. His greatest achievement in cognitive science was his contribution to the concept of the stored-program computer, where he applied the logical correspondence between nervous systems and computing systems to the logical design of electronic computers.

INTRODUCTION

John von Neumann was a Hungarian-born prodigy who became a world-famous mathematician, physicist, logician, economist, engineer, and computer scientist. From 1921 to 1926, he lived and studied variously in Budapest, Berlin, and Zurich. In Berlin, he studied chemistry under the Nobel laureate Fritz Haber. At the Swiss Federal Institute of Technology in Zurich, he earned a bachelor's degree in chemical engineering; while there, he often had talks with the distinguished mathematician Hermann Weyl. He received his doctoral degree in mathematics from the University of Budapest in 1926, and from 1927 to 1930 he taught at the University of Berlin as a *privatdocent*.

In 1927, von Neumann published his paper, 'Zur Hilbertschen Beweistheorie', having worked on set theory and mathematical logic for several years. This work contributed to David Hilbert's program to reduce mathematics to a rigidly formal system of logic and arithmetic. Von Neumann was invited to visit Princeton University in 1930; three years later he became a Full Professor at the nearby Institute for Advanced Study, as a colleague of Weyl, Albert Einstein, and other notable scholars.

Von Neumann was highly interdisciplinary. In mathematics, he contributed to the theory of rings of operators in Hilbert's multidimensional spaces. In quantum mechanics, he showed that Schrödinger's wave mechanics and Heisenberg's matrix mechanics are mathematically equivalent, even though they are intuitively very different.

Von Neumann published his seminal paper, 'Theory of games of strategy' in 1928. In 1944, he wrote *Theory of Games and Economic Behavior* jointly with the economist Oscar Morgenstern. This theory addresses a sophisticated component of human cognition and action, as it applies mathematics to a fundamental aspect of human activity: conscious intentional goal-seeking. The phenomenologist Franz Bretano attempted to distinguish the mental from the physical, asserting that only the mental could possess intentionality. However, several computer programs based on goal-directed strategies have been developed – for example, in checkers and chess – that take the first tiny step towards designing a computer with a general power of rational thought.

ELECTRONIC COMPUTERS

Von Neumann's interest in electronic computers was kindled by developments in the early years of the Second World War, during which a causally connected sequence of four electronic computers launched the modern computer revolution. These were: the special-purpose Atanasoff–Berry Computer (ABC) of Iowa State University; the manually programmed general-purpose Electronic Numerical Integrator and Computer (ENIAC) of the University of Pennsylvania's Moore School of Electrical Engineering; the stored-program Electronic Discrete Variable Computer (EDVAC), also of the Moore School; and the stored-program Institute for Advanced Study (IAS) Computer. Von Neumann entered the picture in August 1944, having been invited to consult for the Moore School team as the ENIAC was nearing completion and the EDVAC was being envisioned.

J. Presper Eckert and John W. Mauchly, who were planning for the EDVAC, had decided that the computer would have a memory separate from

and interacting with an arithmetic unit. The memory would be large enough to hold not only numerical data but also instructions in coded binary form, both to be entered automatically prior to a problem run. The memory, which would be regenerated periodically, would consist of mercury-delay-line tanks adapted by Eckert and T. Kite Sharpless from those used for radar detection of enemy aircraft. The arithmetic unit would perform addition, subtraction, multiplication, division, and perhaps square-rooting, all serially and in the binary number system.

When von Neumann arrived, he was occupied primarily (and secretly) with consultations on the two atomic bomb projects. Nevertheless, his contribution to the design of the EDVAC constitutes his most important contribution to cognitive science. His idea was to proceed in two stages: first, develop a logical design in terms of a new logical language of his own invention; then, have computer engineers work out the electronic circuitry according to that design. The idea, which was to become standard, was that the logical structure could be developed before the much more complex electronic circuitry was devised.

Figure 1(a) is a simple example of what von Neumann had in mind for the logical stage. It shows the structure of a binary serial adder, the core of the arithmetic unit, as it adds two binary streams $A(t)$ and $B(t)$ for $t = 0, 1, 2, \ldots$. At each moment of time t, the sum $S(t)$ is 1 just in case an odd number of the two input bits $A(t)$ and $B(t)$ and the carry-back bit $M(t)$ are 1. The carry bit $N(t)$ is 1 just in case at least two of its inputs $A(t)$, $B(t)$, and the delayed carry bit $M(t)$ are 1. These are the defining rules of binary addition. Figure 1(b) is the state transition table describing the behavior of 1(a).

Von Neumann's new computer logic is an 'iconic logic of switching and memory nets': 'iconic' because it gives a two-dimensional diagram of a computer circuit of any complexity, and 'switching and memory' because any clocked computer circuit can be constructed from complexes of switches and unit delays. The unit delay provides the memory, when put in a cycle with switches, that can remove a word and replace it with a new one. More generally, von Neumann realized from his background in logic that the structure of any clocked digital computer could be represented in his iconic logic of switching and memory nets. He realized also that this language could be used to settle many questions concerning the structure and the machine language of an electronic computer at this logical stage.

The iconic logic of switching and memory nets has other switching symbols not shown in figure 1, in particular, switches for 'not' and 'and'. When these are added, the logic becomes universal in the sense that the logical and computing structure of any clocked electronic digital computer can be represented in it.

COMPUTERS AND NEURAL NETS

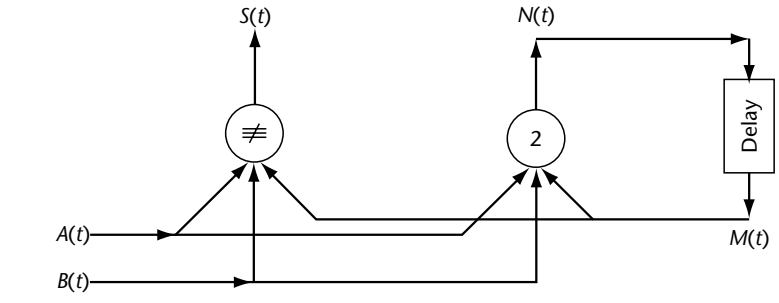
Von Neumann's first paper on the EDVAC, 'First draft of a report on the EDVAC', was distributed to selected people in the US and England in June 1945. In this paper, von Neumann states the source of his iconic logic of switching and memory nets. In a section entitled 'Neurons, synapses, excitatory and inhibiting types', where the threshold-2 switch of figure 1(a) is excitatory and a negation is inhibitory, he wrote:

> Following W. Pitts and W. S. McCulloch ('A logical calculus of the ideas immanent in nervous activity', *Bull. Math. Biophysics*, vol. 5 [1943], pp. 115–133) we ignore the more complicated aspects of neuron functioning: Thresholds, temporal summation, relative inhibition, changes of the threshold by after effects of stimulation beyond the synaptic delay, etc.

McCulloch and Pitts had seen a deep analogy between how the neural nets of nervous systems operate and a new mathematical logic symbolism in which each simple logical switch is followed by a unit delay. Von Neumann then saw that this logical structure, in which every switching action of 'not', 'or' and 'and' is followed by a unit delay, was an entirely new mathematical logic that could be applied to the electronic action of the planned EDVAC. This newly discovered correspondence between the logic of the EDVAC and the logic of neural nets shows the vital importance of physiology and neurology in shaping cognitive science.

Using his new logic of nets, von Neumann worked out the logical design of the memory and of the arithmetic unit of the EDVAC, together with the communication channel between them. This much of the EDVAC's architecture was a sufficient semantic basis for him to specify the machine language of the EDVAC, which he did in the last section of his paper, entitled 'The code'.

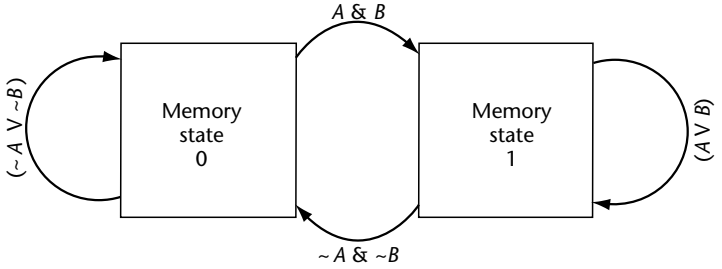
After the Second World War, von Neumann turned to the design of the IAS Computer, alternatively called the Von Neumann Computer. He had, in his paper on the EDVAC, suggested a new, better kind of computer memory. This entailed combining the functions of a standard cathode ray tube (as used in television sets) with those of an iconoscope (as used in television cameras) to make



(a)

Input state		Present memory state *M*	Next memory state *N*	Output state *S*
0	0	0	0	0
0	1	0	0	1
1	0	0	0	1
1	1	0	1	0
0	0	1	0	1
0	1	1	1	0
1	0	1	1	0
1	1	1	1	1

(b)



(c)

Figure 1. Local and global ways of representing a von Neumann logical net binary serial adder. Figure 1(a) shows a three-part logical net specifying an electronic circuit for serial addition. There is a logical inequivalence switch on the left. It signifies that the output sum, $S(t)$, is 1 whenever either just one or all three of the input bits $A(t)$, $B(t)$, and $M(t)$ are 1; but that otherwise $S(t)$ is 0. There is a threshold-2 switch in the middle, since a carry occurs when at least two of its three inputs are 1. Finally, there is a delay for carry-back on the right, which obeys the recursion: $M(0) = 0,\ M(t+1) = N(t)$. Figure 1(b) shows the state transition table (truth table) specifying the behavior of this logical net; and figure 1(c) shows the corresponding state transition diagram.

a computer storage tube. An iconoscope converts a pattern of varying light and dark into corresponding variations in electric charges; a cathode ray viewing tube converts those variations back into a (black-and-white) picture. Von Neumann reasoned by analogy that a single tube could be designed to carry out both functions and thus be a memory tube in which the bits 0 and 1 were distinguished at each of 1,024 spots on the inside of the front of the cathode ray tube by two different levels of electric charge. Frederick C. Williams of Manchester University invented such a tube in late 1946.

This form of computer memory allowed random access to the bits it held, so that a better way to treat binary words than the serial processing of the EDVAC was possible: one could now store, transfer, and process all the bits of a word in parallel. The June 1946 report by Burks, Goldstine and von Neumann, called 'Preliminary discussion of the logical design of an electronic computing instrument', laid out the complete general plan for the first computer with a random-access memory. There were to be 40 memory cathode ray tubes, each storing one bit of a 40–bit word, which could be either a 40–bit binary number or a pair of single-address instructions. Each tube would store 1,024 bits, but each word or instruction would be transferred in bit-parallel between the memory and the

arithmetic unit or the control, or from the arithmetic unit to the memory.

Moreover, this report contained the first simple formulation of a modern programming language, the descriptive design of the first modern program control, and the paradigm of the first modern computer architecture. In particular, the address-substitution instruction is stated much more clearly than in the EDVAC paper, and the control apparatus to execute it is described completely.

COMPUTER SELF-PROGRAMMING

The Moore School team had realized the great advantage the EDVAC would offer for setting up problems, compared with the ENIAC's manual plugboard system. It would now be possible to enter a program automatically, by transferring it from punched paper tape or magnetic wire into the memory of the computer. But von Neumann saw further that when instructions are stored in read–write memory, they also can be operated on by the program as the problem run progresses. His address-substitution instruction, invented for this very purpose, changes the address reference of an instruction without changing the operation it directs. This new instruction would be used to select a particular entry from a function table of entries. The fact that the EDVAC, the IAS Computer, and all their descendants can be instructed to write their own programs shows that modern electronic computers are potentially cognitive systems of considerable intellectual power.

Von Neumann's address-substitution instruction, which he had also proposed in his 1945 paper, was one of his most important inventions, for it led him to his 'Library of subroutines', the first automatic programming system. He and Goldstine developed this in their reports of 1947 and 1948 entitled 'Planning and coding of problems for an electronic computing instrument'. Von Neumann's idea was to have a library of the most commonly used subroutines, so that the programmer need only write a 'combining routine' to instruct this library as to what subroutines to use in what order, and the boundary conditions for compounding them into a single program.

SELF-REPRODUCING CELLULAR AUTOMATON

Soon after the war, von Neumann became engaged in advising the US government on computers and military policy. He was appointed to the Atomic Energy Commission, and for his contribution to this, and for his other achievements, President Eisenhower personally awarded him the Medal of Freedom in 1956. He continued to make scientific contributions, with the central goal now of creating a theory of automata. His first project was the design of a self-reproducing cellular automaton, which he partially worked out before his death from cancer in 1957. Burks completed and edited this design in a 1966 book, *Theory of Self-Reproducing Automata*.

Von Neumann's construction of a logical net diagram for a self-reproducing automaton was the first contribution to a subject now called 'artificial life'. This automaton operates in a potentially infinite two-dimensional space of squares – an infinite checkerboard array. Each square contains a 29-state automaton that is directly connected to its four contiguous neighbors, so that its state at time $t+1$ is determined by the states of these neighbors at time t.

These 29 states and their interaction are complicated, but the architecture of the self-reproducing automaton is straightforward. It has two main parts, a tape unit and a constructing unit, each of which is a finite automaton. The tape unit can construct an indefinitely extendable storage tape that it can write on, read from, or erase at the request of the constructing unit. The constructing unit uses the tape unit as its source of instructions for constructing an automaton in a blank area of the cellular space. It has an attached 'constructing arm', which it can extend out through any empty area of space, use to construct any automaton that is completely described on the tape, and then withdraw.

Suppose, finally, that the complete cell-by-cell description of the self-reproducing automaton is put on its own tape, and the automaton is started. It will then carry out the following operations in succession: first, send its construction arm out into an empty area and direct its construction tip to construct a copy of itself, including its tape; then, read its own tape and copy all the information on it onto the new tape it has just constructed.

The result will be a complete copy of the self-reproducing automaton in a new region of cellular space. This is automaton self-reproduction. Using a computer with a very large screen, one could program this process as a dynamic display with different colors representing the different states of each cell.

Von Neumann's final work was *The Computer and the Brain*, published in 1958. It was argued earlier that a modern electronic computer is a cognitive system, on the basis of von Neumann's derivation

of his iconic logic of switching and memory nets from the McCulloch and Pitts logic of neurons. *The Computer and the Brain* strengthens this thesis.

Further Reading

Aspray W and Burks AW (eds) (1987) *Papers of John von Neumann on Computing and Computer Theory*. Cambridge, MA: MIT Press.

Burks AR and Burks AW (1988) *The First Electronic Computer: The Atanasoff Story*. Ann Arbor, MI: University of Michigan Press.

Burks AW (1970a) *Essays on Cellular Automata*. Urbana, IL: University of Illinois Press. [Reprinted (1971) by the Library of Computer and Information Sciences, Riverside, NJ.]

Burks AW (1970b) *Von Neumann's Self-Reproducing Automata*. Urbana, IL: University of Illinois Press. [Reprinted in (Aspray and Burks, 1987).]

Burks AW (2001) Turing's theory of infinite computing machines (1936–1937) and its relation to the invention of finite electronic computers (1939–1949). In: Bandimi S and Worsch T (eds) *Theory and Practical Issues on Cellular Automata: Proceedings of the Fourth International Conference on Cellular Automata for Research and Industry, Karlsruhe, 4–6 October 2000*, pp. 179–197. London: Springer.

Burks AW and Burks AR (1981) The ENIAC: first general-purpose electronic computer. *Annals of the History of Computing* **3**: 310–399.

Feigenbaum EA and Feldman J (eds) (1983) *Computers and Thought*. New York, NY: McGraw-Hill.

Macrae N (1992) *John von Neumann*. New York, NY: Pantheon Books.

Metropolis N, Howlett J and Rota G (eds) (1980) *A History of Computing in the Twentieth Century*. New York, NY: Academic Press.

Vygotsky, Lev

Introductory article

Laura E Berk, Illinois State University, Normal, Illinois, USA
Sara Harris, Illinois State University, Normal, Illinois, USA

CONTENTS

Russian psychologist Lev Semenovich Vygotsky (1896–1934) originated sociocultural theory, which regards social interaction between children and more expert members of their culture as the wellspring of cognitive development. Vygotsky's theory and especially his most influential concept – the zone of proximal development – inspired research into many aspects of teaching and learning.

INTRODUCTION

Lev Semenovich Vygotsky laid the groundwork for the sociocultural perspective on cognitive development that rose to the forefront of the fields of child development and education during the last two decades of the twentieth century. Sociocultural theory emphasizes the vital connection between the individual's social and psychological worlds. Specifically, it regards communication between children and more expert members of their culture as the source of consciousness, of distinctly human, higher cognitive processes, and of the capacity to regulate thought and behavior.

Born in 1896 into a well-to-do Russian-Jewish family, Vygotsky experienced an intensely intellectual home life. His elementary education took place at home, through a private tutor who used a variation of the Socratic method – an experience that may have contributed to Vygotsky's later view of adult–child dialogue as central to cognitive development.

As an adolescent, Vygotsky attended public high school and then a selective private Jewish school, where he developed strong interests in literature and theater. After winning a lottery that determined which of a small number of Jewish students were to be admitted, Vygotsky enrolled at Moscow University and majored in law, one of the few fields that permitted Jews to live outside restricted areas. He also attended Shiniavsky People's University, an unofficial institution staffed by leading scholars who had been expelled from faculty posts for their political views. In 1917, the year of the Russian revolution, 21-year-old Vygotsky graduated with a firm grounding in history, philosophy, psychology and literature, the last of which remained his primary interest. He then returned to his hometown of Gomel to teach Russian language and literature at vocational schools for tradesmen and teachers. During those years he read widely in psychology, and set up a small laboratory at Gomel Teachers College for students to run simple psychological experiments – events that laid the groundwork for his transition to psychological research.

In 1919 Vygotsky contracted tuberculosis. Judging from the extraordinary productivity of his brief career, he may have understood that his days were numbered. Captivated by psychology, in 1924 he gave an influential address before the Russian psychological community and as a result was invited to join prominent Soviet psychologists at the Psychological Institute in Moscow. Thus, Vygotsky entered the field without formal training.

Collaborating with other talented researchers, Vygotsky set two lofty goals. The first was theoretical: creating a unifying perspective in psychology that would resolve the theoretical contradictions of his time. The second was practical: addressing serious problems of Soviet society, including reducing the high rate of illiteracy and improving the circumstances of children with physical disabilities and psychological problems. Vygotsky helped found the Institute of Defectology (the Russian term for the field of abnormal psychology) and in 1925 was appointed scientific leader of research and education for children with disabilities. That same year he finished his doctoral dissertation, on the psychology of art. He held various professorships and either directed or was affiliated with many psychological research and pedagogical organizations. Despite repeated bouts of illness, he

wrote prolifically, producing in 1929–1930 alone almost fifty works.

In 1934 Vygotsky suffered his last tuberculosis attack. Nevertheless, he continued to work at a frantic pace, finishing one of his most important works, *Thought and Language*, from his deathbed. He died in 1934 at the age of 37 years. For the next two decades his publications were banned in the Soviet Union because of Stalinist repression. In the 1950s they began to be reissued. They were first translated into English in the 1960s, and new facets of his 180 works continue to be disseminated around the world today.

VYGOTSKY'S THEORY

Massive social changes following the Russian revolution energized Vygotsky as he forged a new sociocultural perspective on child development and education consistent with Marxist principles. Just as human history evolves through revolutionary social movements and authentic social activities of labor and production, so the child develops through social interaction and participation in culturally meaningful endeavors.

According to Vygotsky, infants are endowed with basic perceptual, attentional, and memory capacities that they share with other animals. These follow a natural course of development as the infant makes direct contact with the environment. Once children become capable of mental representation, especially through language, their capacity to communicate with more knowledgeable members of their culture greatly expands. As experts guide, prompt, and explain during joint adult–child activities, they transfer to children the values, beliefs, customs, and skills of their cultural group. Through these experiences, basic mental capacities are transformed into uniquely human, higher cognitive processes. These include, among others, controlled attention, deliberate memorization and recall, categorization, planning, problem-solving, abstract reasoning, and self-regulation of thought and behavior.

In sum, Vygotsky claimed that any higher cognitive capacity first appears between people, in social interaction. Only later can children take over responsibility for that capacity, applying it on their own. Symbolic 'tools of the mind' – especially language, the most flexible and widely used representational system – enable this transfer of cognition from the social to the psychological plane. At first, the adult assists the child in regulating his or her activities, as illustrated by the following excerpt of communication between a 4-year-old child and his mother, who helps him master a difficult puzzle:

> Sammy: 'I can't get this one in.' [Tries to insert a piece in the wrong place]
>
> Mother: 'Which piece might go here?' [Points to the bottom of the puzzle]
>
> Sammy: 'His shoes.' [Looks for a piece resembling the clown's shoes, but tries the wrong one]
>
> Mother: 'Well, what piece looks like this shape?' [Pointing again to the bottom of the puzzle]
>
> Sammy: 'The brown one.' [Tries it, and it fits; then attempts another piece and looks at his mother]
>
> Mother: 'That's it. Try turning that one just a little.'
>
> Sammy: 'There!' [Puts in several more pieces while commenting to himself, 'Now a green piece to match,' 'Turn it' (the puzzle piece), as his mother watches]

Soon children take the language of these dialogues and direct it toward the self, at first engaging in audible self-talk called 'private speech.' Gradually, children internalize private speech, integrating socially derived ways of thinking into their inner speech – the silent verbal dialogues we carry on with ourselves while thinking and acting in everyday situations. As private speech becomes less audible, its structure modifies; it becomes more abbreviated and efficient. Simultaneously, the function of speech changes, from influencing others to clarifying thoughts and regulating (or gaining voluntary control over) behavior. During the preschool years, Vygotsky observed, children frequently speak aloud to themselves for the purpose of self-regulation. With cognitive maturity and mastery of many challenging activities, private speech is largely transformed into inner speech by middle to late childhood.

Vygotsky's most influential theoretical concept – the zone of proximal development (ZPD) – specifies the region in which this transfer of abilities from social interaction to individual functioning occurs. Vygotsky originally introduced the ZPD to argue against traditional intelligence tests, which assess static abilities rather than the ever-changing quality of human cognition. He suggested that tests should measure not what children already know and can do by themselves but what they can do with the assistance of a more knowledgeable partner. Hence, the ZPD refers to a range of tasks that the child cannot yet handle alone but can accomplish with the help of adults and more skilled peers.

The ZPD is closely related to a core assumption of Vygotsky's theory: that education leads, or elicits, cognitive development. Through collaboration with

teachers, parents, and more expert peers on tasks within the ZPD, the child actively constructs new competencies. Vygotsky argued that educative environments must utilize the ZPD, providing tasks carefully selected to awaken those capacities the child is ready to master with social support. As education leads to new knowledge and skills, it permits children to attain a new level of understanding, in which they become aware of their mental activities and regulate thought and behavior more effectively.

According to Vygotsky, the same general principles that govern normal child development also apply to children with disabilities. The most debilitating consequence of a physical or psychological problem, Vygotsky explained, is not the disability itself but its implications for the child's participation in culturally meaningful activities. When a disability interferes with opportunities to experience positive interaction with adults and peers, the child develops a more serious, secondary deficit in higher cognitive processes. Therefore, education must focus on improving social life, to ensure that children with special needs reach their cognitive potential.

CONTEMPORARY RESEARCH AND APPLICATIONS

Vygotsky's theory – especially the concept of the ZPD – has inspired a burgeoning contemporary research literature. Investigators have addressed the features of communication that create the ZPD, coining the term 'intersubjectivity' to describe the fine-tuned adjustments of social partners to one another's perspectives, which establish a common ground for dialogue. Another feature of communication that fosters cognitive progress is 'scaffolding': this refers to a changing quality of social support during a teaching session, in which adults adjust the assistance they provide to fit the child's current level of progress (see Sammy's puzzle-solving with his mother's help, above, for an illustration). Furthermore, numerous investigations have addressed parent–child conversation as the prime source of children's narrative competence. As children internalize their culture's narrative style, they organize personally relevant past experiences into a life story that is central to defining the self.

Energized by Vygotsky's suggestion that peers can create ZPDs for one another, researchers have examined the conditions in which peer collaboration fosters cognitive development. Findings indicate that a crucial factor is cooperative learning – structuring the peer group so that children benefit from one another's expertise and work toward common goals. Furthermore, Vygotsky-based school reform experiments reveal that peer collaboration in classrooms works best when it is supported by a culture of collaboration throughout the educational institution, in which administrators, specialized consultants and teachers cooperate to promote children's learning.

The ZPD has also stimulated a new approach to mental testing called 'dynamic assessment', in which the examiner introduces purposeful teaching into the testing situation to see what the child can attain with social support. Children's responsiveness to individualized teaching and capacity to transfer what they have learned to new tasks predicts future intellectual performance.

Vygotsky's claims about the role of self-directed communication in regulating thought and action have prompted a steady stream of research on private speech. Findings confirm that socially rich contexts and challenging tasks foster this speech-to-self. Consistent with Vygotsky's hypotheses about internalization, audible private speech declines over the late preschool and school years, while signs of inner speech (inaudible muttering and lip and tongue movements) increase. Moreover, children's use of task-relevant private speech is consistently related to focused attention and gains in task performance – outcomes that support the self-regulating function of private speech.

Vygotsky's theory has also sparked a spate of studies on make-believe play as a ZPD in which preschool children acquire a wide variety of culturally adaptive competencies. Findings reveal that pretense originates in joint caregiver–child play, which prepares children for playful cooperation with agemates. Among favorable outcomes of collaborative make-believe are enhanced memory, reasoning, language, early literacy, imagination, perspective taking, and self-regulation skills.

Children with disabilities have also been the focus of Vygotsky-inspired research. For example, because of greater access to supportive parent–child interaction, deaf children of deaf parents (who communicate through sign language) develop more favorably than do deaf children of hearing parents (who either do not use sign language or need time to learn to do so). Deaf children of deaf parents resemble hearing children in quality of parent–child interaction, development of private speech (manual signing to themselves) and self-regulation. In contrast, deaf children of hearing parents display diminished use of private speech and serious self-regulation difficulties. Similar evidence exists for children with attention deficit hyperactivity disorder (ADHD), whose biologically

based cognitive deficits interfere with supportive adult–child interaction and development of private speech.

Finally, Vygotsky's theory has served as the foundation for an expanding literature on cultural variation in cognitive development. A major finding is that each culture selects tasks for children's learning, and social interaction within those tasks leads to knowledge and skills essential for success in that culture. Consequently, children develop unique, culturally relevant cognitive strengths.

CONCLUSION

Vygotsky's theory has not gone unchallenged. Although he acknowledged the role of diverse symbol systems in the development of higher cognitive processes, he elevated language to highest importance. Yet cross-cultural research suggests that a strong emphasis on verbal dialogue and scaffolded teaching may be uniquely Western phenomena. Observations in several village societies reveal that children are expected to learn through keen observation and participation in adult activities, and adults rely more on demonstration and gesture than on verbal communication to transfer culturally adaptive ways of thinking to children. Furthermore, Vygotsky theorized that the natural line and the social line of development join, forming a single developmental pathway. Yet in focusing on the social line, he said little about the natural (biological) line. Investigators in the sociocultural tradition continue to pay less attention to the biological substrate of development than do researchers of other theoretical persuasions.

Sociocultural theory is unique in granting social experience a fundamental role in cognitive development. In its fine-grained examination of social collaboration within the ZPD, Vygotsky-inspired research has deepened our understanding of the everyday processes that spur cognitive development and permit children to become participating members of their communities. In this respect, sociocultural theory complements more traditional approaches to investigating children's cognition, which have focused largely on solitary experimentation and discovery.

Further Reading

Berk LE (2001) *Awakening Children's Minds: How Parents and Teachers Can Make a Difference*. New York, NY: Oxford University Press.

Diaz RM and Berk LE (eds) (1992) *Private Speech: From Social Interaction to Self-regulation*. Hillsdale, NJ: Lawrence Erlbaum.

Forman EA, Minick B and Stone CA (eds) (1993) *Contexts for Learning: Sociocultural Dynamics in Children's Development*. New York, NY: Oxford University Press.

Gauvain M (2001) *The Social Context of Cognitive Development*. New York, NY: Guilford.

Goncu R (ed.) (1999) *Children's Engagement in the World: Sociocultural Perspectives*. Cambridge, UK: Cambridge University Press.

Lidz CS and Elliott J (eds) (2000) *Dynamic Assessment: Prevailing Models and Applications*. New York, NY: JAI Press.

Rieber RW (ed.) (1993–1999) *The Collected Works of L. S. Vygotsky*, vols 1–6, translated by Hall MJ. New York, NY: Plenum.

Rogoff B (1998) Cognition as a collaborative process. In: Kuhn D and Siegler RS (eds) *Cognition, Perception, and Language*, 5th edn, *Handbook of Child Psychology*, vol. 2, pp. 679–744. New York, NY: John Wiley.

Tharp RG and Gallimore R (1988) *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context*. New York, NY: Cambridge University Press.

Wertsch JV (1985) *Vygotsky and the Social Formation of the Mind*. Cambridge, MA: Harvard University Press.

Wertsch JV and Tulviste P (1992) L. S. Vygotsky and contemporary developmental psychology. *Developmental Psychology* **28**(4): 548–557.

Walter, Grey

Introductory article

Walter J Freeman, University of California, Berkeley, California, USA

CONTENTS

An autonomous robot embodies the principles of goal-seeking and scanning that characterize animal behavior. Grey Walter (1910–1977), a physiologist adept in electronics engineering with major accomplishments in early electroencephalography, used his wartime exposure to radio detection and ranging (RADAR) to build a simple 'brain' that endowed his artificial 'turtle' with complex adaptive behaviors.

INTRODUCTION

Nobel prize-winning Sir Winston Churchill in his history of the Second World War wrote a chapter on the 'wizards' who had helped Britain win the war in the air by the development and use of radar. William Grey Walter (1910–1977) was one of those young wizards. He used his experience to design and construct a lifelike robot with a nonlinear dynamic brain, which offered an existence proof that brains are simpler than many of us have supposed.

BIOMEDICAL ENGINEERING

In the decade before the war Grey Walter had already done important work in a field we now call biomedical engineering by his discoveries in electroencephalography (EEG), a medical procedure in which the oscillating fields of electric potential on the scalp and in the brain are measured and interpreted. His first achievement was to identify correctly the source of the alpha rhythm (8–12 Hz). A German psychiatrist, Hans Berger, in 1929 had discovered brain waves by attaching one electrode to the forehead and another to the back of the head. He used a primitive Fleming electronic 'valve' ('vacuum tube' for Americans, the predecessor of the transistor) to amplify the potential difference. He erroneously inferred that alpha came from the frontal lobes. Walter used his knowledge of the theory of potential, volume conduction, and electronics to triangulate the waves and locate their source in the occipital lobe. He also invented the use of low-frequency delta waves (1–2 Hz) to locate brain tumors and abscesses, as well as foci of brain damage that triggered bouts of epileptic activity, then widely known as 'paroxysmal cerebral dysrhythmias'.

After the war he gathered a group of young engineers, along with surplus radar, radio, and other electronic equipment, to work in a laboratory at Bristol University, which rapidly became one of the world's leading centers for EEG research. One of their achievements was automated spectral analysis of EEG traces. A standard method for EEG analysis, since its discovery, had been to measure the power in various frequency bands, including alpha and delta, also beta (15–30 Hz) which Berger identified as the carrier of brain information, and theta (3–7 Hz) as well as alpha, which he concluded were gating frequencies of packets of brain information. Walter used his skills in analog electronics to conceive a device built by engineers that displayed the frequency content in an EEG trace, even as the trace was displayed with an ink-writing oscillograph, a pen whose fluctuations left a trace on moving paper that became the mainstay of electroencephalographers.

Another notable discovery was a very slow change in electrical potential at and around the vertex of the head, measured with respect to indifferent reference points such as the ear lobes. Walter named this event the contingent negative variation (CNV), because it was seen only after a warning signal had been given to a human subject, who would then plan a possible movement in anticipation of a second signal. German researchers discovered a comparable slow potential in a similar behavioral context, calling it the *Bereitsschaftpotential* ('readiness potential'). The intriguing aspect of these electrical potentials is that they permit the observer to predict that a subject will make a response within the next half to one second, before the subject is aware of an intention to act. Some psychologists regard this cerebral phenomenon as evidence that intentional actions are initiated

before awareness of such actions emerges, and that consciousness is involved in judging the values of actions rather than in the execution of them. Simply put, we learn what sort of person we are by observing not our good intentions but our own actions, which often surprise us.

Walter extended his temporal spectral analysis of time series to spatial analysis by conceiving a bank of amplifiers connected to an array of 22 oscilloscopes. This advance enabled him to show not only the amplitude but the phase difference of each trace of the alpha waves with respect to the others, by using cinemas of the oscilloscopes. With his 'toposcope' he visualized the spread of alpha waves across the surface of the brain in ways resembling the ebb and flow of tidal waves around the earth. Alpha activity has the peculiarity that it is most apparent when a human subject is at rest with eyes closed, and it disappears when the eyes are opened or if mental arithmetic is undertaken. Walter proposed that the alpha represented 'scanning' by the brain in search of local centers of activity when none was present, and that it stopped when a 'target' was found in the cortex. This 50-year-old hypothesis was and still is controversial, but it is still not disproven.

AUTONOMOUS, ADAPTIVE ROBOTS

Walter's greatest achievement stemmed from his wartime experience with electronics. Guided missiles with proximity fuzes were then one of two very active foci of interest, the other being devices for scanning the horizon for targets to be identified and intercepted. The scanning mechanism he helped develop was known as the 'plan position indicator', consisting of the point of light created by an electron beam that moved from the center to the edge of the oscilloscope screen and created a bar like the spoke of a wheel. The spoke rotated counterclockwise at the refresh rate of the screen. A likely radar target appeared as a bright spot, giving its direction and distance. This device is in widespread use today, for example, in ships, submarines, and air traffic control centers. It was the basis for Walter's toposcope when applied to alpha waves.

Walter had a very rich, speculative imagination. The concept of a machine that would define a goal and seek it by scanning resonated with his interest in brains as biological systems that evolved through learning from the consequences of their own goal-oriented actions. He undertook to incorporate these two cognitive operations, goal-seeking and scanning, into an electronic 'toy' that would simulate these most basic characteristics of animal (and human) behavior.

The outcome was fully spectacular, though at the time its significance was not recognized, and his device has been all but forgotten. It was a roving machine so lifelike, as he described it in his book *The Living Brain*, that an old lady who felt pursued by it ran upstairs and locked her door.

He named his device *Machina speculatrix* in order to distinguish it from passive devices, such as his earlier conception of *M. sopora* that incorporated Norbert Wiener's principle of stabilization of machine performance by negative feedback ('*Cybernetics*'), and from W. Ross Ashby's '*Homeostat*' that extended the principle of the stability of biological organisms by introducing adaptation through learning. These stable models used what the Harvard physiologist Walter Cannon called 'homeostasis', but unlike plants and sessile automata, *M. speculatrix* was continually on the prowl in search of its designer-endowed goal: moderate illumination. His three-wheeled vehicle, which came to be known as 'Grey Walter's turtle', had two motors, one for progression by the front wheel dragging the hind wheels like a child's tricycle and one for turning the front wheel. Its drive system, batteries and 'brain' were mounted on a chassis. Above it he hung a carapace from a center pole, so that it could swing inwardly from any direction and contact the chassis. This contact operated a switch, so that if the tortoise hit an obstacle or encountered an incline, it would stop, back up, turn, and eventually move around it or avoid it altogether. This 'receptor' gave the device the sense of touch, information about the direction of gravity, and the means to explore objects in its environment by touch.

He also gave *M. speculatrix* a photocell for sensing light, and designed its circuits to use homeostatic feedback to seek and maintain a moderate level of illumination, which varied with its location and orientation but also with the charge in its two batteries. Its hutch was brightly lit, and when its batteries were fully charged, that level was aversive, so it moved out into its contracted world, continually swinging in cycloid loops first away from the light, then, as its batteries ran down back toward the light in its hutch (see 'Illustrations' webpage). In a single charge cycle it could explore nearly 100 m^2, dealing with obstacles by pushing them aside or going around them, though sometimes straying too far and being found 'starved to death' behind a couch. When it regained its hutch, it turned itself off and took nourishment from electrical contacts on the floor.

As a means for detecting the internal state, Walter fixed a marker light on the carapace that stayed on when the turning motor was on but went out when turning stopped. When the turtle encountered its own light in a mirror, it stopped and oriented to its own light, but stopping turned out its light. Then it resumed circling, saw its light again, and stopped. This behavior continued until it had passed the mirror. If it encountered another of its own kind, attracted by the other light, a stately dance ensued of bumping and backing. Walter thought that these behaviors expressed self-recognition and recognition of conspecifics.

These complex and not fully predictable behaviors of exploration, negative and positive tropism, discrimination, adaptation to changing internal and external environments, optimization, and stabilization of the internal medium were done with a very simple brain: two miniature valves serving as 'neurons', two mechanical relays, two capacitors, two receptors, and two motors. Walter achieved this by ingenious circuit design. For example, when the turtle was in search mode, the two valves served as serial amplifiers. When it hit an obstacle, the circuit changed to an oscillator, then called a 'multivibrator', which generated the repetitive backing and butting. He went further with new circuitry, which he called the 'conditioned reflex analog' (CORA), to simulate the seven operations he identified in the formation of associative conditioned reflexes. He proposed to embody CORA in a more highly evolved *Machina docilis* ('easily taught') that could learn to go around an obstacle and would then continue to circumvent it after it had been removed. He used *M. docilis* in several prototypic variants to explore different types of memory and the importance of high-frequency oscillations, such as the beta activity he had observed in EEGs, for enriching memory stores. His career was cut short by his tragic motorcycle accident, in which he sustained massive brain damage, leading to his death seven years later.

EVALUATION AND SUMMARY

The significance of Walter's achievements can be understood by recognizing that these complex adaptive behaviors came not from a large number of parts, but from a small number ingeniously interconnected. His devices were autodidacts that could learn by trial and error from their own actions and mistakes. They remembered without internal images and representations. They judged without numbers, and recognized objects without templates. They were the first free-ranging, autonomous robots capable of exploring their limited worlds. They still provide a high standard of accomplishment. The reason is that, despite major advances in locomotion, particularly in simulations of bipedal, quadripedal and hexapedal gaits based on birds, mammals and insects, and despite advances in scanning and navigation that improve robotic comprehension of operating territories, less has been done towards implementation of goal-seeking. The essence of an intelligent machine is that it has within its brain a capacity to conceive desired future states, and it has the degrees of freedom needed to create and adapt its actions in pursuit of those goals in the unpredictable circumstances of the immediate and remote environments. These flexible brain functions that enable simple systems to function in infinitely complex environments are not achieved by rule-driven symbol manipulation, which is at the heart of cognitive science and conventional artificial intelligence. Moreover, Walter emphasized analog electronics to simulate neurodynamics at a time when most of his colleagues, such as John von Neumann, were developing digital computers to implement symbolic logic and deep arithmetic algorithms. His devices were the forerunners of currently emerging machines that are governed by nonlinear dynamics, and that rely on controlled instability, noise, and chaos to achieve continually updated adaptation to ever-changing and unpredictable worlds. He can well be said to have been the Godfather of truly intelligent machines.

Further Reading

Ashby WR (1952) *Design for a Brain*. London, UK: Chapman and Hall.

Cannon WB (1939) *The Wisdom of the Body*. New York, NY: W. W. Norton.

Churchill WS (1949) The wizard war. *The Second World War*, vol. 2, chap. 4. Boston, MA: Houghton Mifflin.

Clark A (1996) *Being There. Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Freeman WJ (1999) *How Brains Make Up Their Minds*. London, UK: Weidenfeld & Nicolson.

Hendriks-Jansen H (1996) *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. Cambridge, MA: MIT Press.

von Neumann J (1958) *The Computer and the Brain*. New Haven, CT: Yale University Press.

Walter WG (1953) *The Living Brain*. New York, NY: W. W. Norton.

Wiener N (1948) *Cybernetics; or, Control and Communication in the Animal and the Machine*. New York, NY: Wiley.

Whorf, Benjamin Lee

Introductory article

John B Carroll, University of North Carolina, Chapel Hill, North Carolina, USA

Benjamin Lee Whorf may justly be considered one of the major contributors to cognitive science in the twentieth century.

Benjamin Lee Whorf pioneered in evaluating the idea of what he called *linguistic relativity*, that is, that 'the structure of the language one habitually uses influences the manner in which one understands his environment', to quote the definition offered by Stuart Chase in his foreword to an edition of Whorf's writings. It was an idea proposed much earlier by such thinkers as Roger Bacon (1220–1292) and Wilhelm von Humboldt (1767–1835). It is not clear whether Whorf was even aware of the thoughts and writings of Bacon, von Humboldt, and others; apparently, it was Whorf's own natural genius that impelled him to think along the same lines as Bacon and von Humboldt. Whorf's contribution was to offer several kinds of evidence that the idea of linguistic relativity – the idea that people actually are influenced to think in particular ways by the structure of the language they speak – might indeed be true and valid.

Yet Whorf was not trained as a cognitive scientist. He developed and published his ideas largely independently of others. Only his major teacher Edward Sapir (1884–1939), a professor of linguistics at Yale University, where Whorf studied and taught for several brief periods starting in 1928, exerted a major influence on the development of his ideas.

Whorf was born in 1897 to a bright and talented family, and died of cancer in 1941. His father was a commercial artist who spent much of his time in such pursuits as stage designing, the production of plays, and the development of photolithography. One of Whorf's two younger brothers, John, became well known as an artist – a painter of watercolors – while the other, Richard, became a highly successful actor in motion pictures.

Whorf himself attended Massachusetts Institute of Technology, graduating in 1918 with a BS degree in chemical engineering. He obtained a position with a fire insurance company in Hartford, Connecticut, where he established himself as a specialist in helping chemical manufacturing companies avoid problems with fires. He enjoyed this work and believed that the generous income and free time available to him enabled him to study problems of interest to him to an extent that he could not have done if he had pursued a career as an academician.

For some years after his graduation from MIT he found many interesting things to read and study in the general field of language. For example, in an attempt to resolve what he regarded as conflicts between science and religion, he studied Hebrew, claiming to discover interesting, previously unnoticed relations between the phonetic structure of Hebrew roots and the meanings of those roots – relations that he called oligosynthetic. Interestingly, some current work on the origin of language makes use of an idea about phonetic meanings that is much like Whorf's oligosynthesis. Only around 1928 did Whorf give up his studies of oligosynthesis because Sapir advised him that such studies were at that time not considered proper scientific activities in linguistics. With Sapir's guidance Whorf began to study various Native American languages, mainly Aztec, Maya, and Hopi.

It was in his studies of these languages that Whorf became impressed with the fact that languages differ markedly in their grammatical structures. He envisaged the possibility that these structures influence the cognitive processes of their speakers. Whorf noted that Native American languages such as Aztec and Hopi are very different from what he called 'standard average European' languages like English, French, and Spanish. Often, it seemed, speakers of Native American languages are required to pay attention to aspects of experience that rarely enter the minds of speakers of 'standard average European' languages like English and French. Conversely, Whorf noticed that languages like English and French may require their speakers to attend to aspects of experience that are not featured in the structures of Native American languages, or other non-European languages.

It is useful for understanding this idea to look at an example of a structural difference between English and certain other languages, such as Chinese. Actually, some aspects of this example will be familiar to many speakers of English because like many other 'standard average European

languages', English has not one but two singular personal pronouns, *he* (generally meaning a male) and *she* (normally meaning a female). It has no singular pronoun, except possibly *they* (when it denotes a single person) that can apply either to a male or a female. Chinese does possess such a pronoun ('ta'); thus it would seem that speakers of Chinese are not continually required to indicate the sex or gender of people they speak about, as speakers of English often are.

When I was a young student in an American elementary school, in the 1920s, I was told that it was perfectly proper in many circumstances to use the male pronoun *he* (or its grammatical variants *him*, *his*) to apply to females, and I have been doing this for most of my life. (Note the first sentence of the second paragraph above, where I unconsciously accepted Stuart Chase's words 'the manner in which one understands his environment'. But perhaps I should have rewritten it as 'the manner in which one understands his or her environment'.) Anyway, in those days it was thought perfectly proper to write a sentence like 'Each child wrote his name in his book', even though many of the children might be female.

In recent years, however, many speakers of English would regard this sentence as improper, or even ungrammatical. They would prefer it to be rewritten, perhaps as 'Each child wrote his or her name in his or her book', or as 'The children wrote their names in their books'.

Research has shown that when English speakers hear or read the pronoun *he* (or its variants) they tend to assume that the person being talked about is a male, even though, according to prescriptive grammar, it could also refer to a female. Thus, the use of *he* tends to bias their thinking. For example, if somebody is trying to hire an assistant who could be either male or female, and places in a newspaper an advertisement that uses only variants of *he* in talking about the kind of person desired (e.g. 'He has good social skills'), the person who placed this advertisement can be expected to receive mainly nominations of males, with many names of female candidates being excluded because people are inclined to think that only male candidates are desired.

More generally, there are a fair number of instances in which it has been shown or believed that people can be influenced in their thinking by structures in the language they speak. Some of these instances, many in native American Indian languages, were offered by Whorf himself. For example, one can find in the Hopi language a special verb form that refers to actions that are highly repetitive, like shaking, pulsing, or meandering (like a river). Whorf suggested that speakers of Hopi were better prepared by their language to understand the many vibratory phenomena studied in modern physics.

Another instance considered by Whorf was the behavior of English-speaking people with reference to highly flammable substances like gasoline or oil. When drums of oil are described as 'full', people tend to be extremely careful in handling them, whereas when drums of oil are described as 'empty' people tend to be less careful, when in actuality they should be *more* careful because of the possible presence of highly explosive residual fumes.

A case found in the Navaho language is interesting. There are many structures in Navaho that require the speaker to use a special form of the verb that indicates the shape of the object being mentioned or talked about (e.g., whether it is round and flat). As a consequence, in a study done some years ago, it was found that Navaho children whose first language was Navaho tended to show awareness of the form or shape of objects earlier in their lives than Navaho children whose first language was English.

These and other such cases have been examined by scholars who conclude that promising evidence exists for the validity of Whorf's linguistic relativity theory. Psychologists and linguists caution, however, that much more evidence is needed, because there are often alternative explanations, other than linguistic relativity, for the findings. Consequently, the status of Whorf's hypothesis of linguistic relativity continues to be problematical.

Further Reading

Carroll JB (ed.) (1956) *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press. With foreword by Stuart Chase. [A second edition of this work is in preparation.]

Carroll JB and Casagrande JB (1958) The function of language classifications in behavior. In: Maccoby EE, Newcomb JM and Hartley EL (eds) *Readings in Social Psychology*, 3rd edn, pp. 18–31. New York: Holt, Rinehart & Winston.

Gumperz JJ and Levinson SC (eds) (1996) *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.

Harley TA (2001) *The Psychology of Language from Data to Theory*, 2nd edn. Hove, UK: Psychology Press.

Khosroshahi F (1980) Penguins don't care, but women do: a social identity analysis of a Whorfian problem. *Language in Society* **18**: 505–525.

Lee P (1996) *The Whorf Theory Complex: A Critical Reconstruction*. Amsterdam/Philadelphia: John

Benjamins. [Vol. 81, Amsterdam Studies in the Theory and History of Linguistic Science.]

Lucy JA (1992) *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. New York: Cambridge University Press.

Lucy JA (1996) *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge, UK: Cambridge University Press.

MacKay DG (1980) Psychology, prescriptive grammar, and the pronoun problem. *American Psychologist* **35**: 444–449.

Ruhlen M (1994) *The Origin of Language: Tracing the Evolution of the Mother Tongue*. New York: John Wiley & Sons.

Wittgenstein, Ludwig

Introductory article

Justin Leiber, University of Houston, Houston, Texas, USA

CONTENTS

For his views on language, logic, and mind, Ludwig Wittgenstein (1889–1951) is perhaps the most studied, most influential, and most puzzling philosopher of the twentieth century.

CENTRAL PHILOSOPHICAL VIEWS

Bertrand Russell, who turned the young Wittgenstein's attention from aeronautical engineering to logic, tells us that Wittgenstein had all the passionately mad drive, exultantly creative and savagely critical in his thinking, 'of genius as traditionally conceived'. Wittgenstein also displayed the naive arrogance, isolation, unconventionality, and stylistic peculiarities of genius 'as traditionally conceived'. His terse first book, *Tractatus Logico-Philosophicus*, presents seven cardinal numbered sentences, each of which (except the last) is followed by further sub-numbered and sub-sub- numbered commentaries. The seven cardinal sentences are:

1. The world is everything that is the case.
2. What is the case, the fact, is the existence of states of affairs.
3. A logical picture of facts is a thought.
4. A thought is a sentence with sense.
5. A sentence is a truth-function of elementary sentences.
6. The general form of a truth function is $[\bar{p}, \bar{\xi}, N(\bar{\xi})]$.
7. Whereof one cannot speak, thereof one must be silent.

Like Russell, Wittgenstein sharply distinguished logical and mathematical truths from the empirical, and therefore accidental, truths of sensory experience and their summations in natural science as truth functions of the elementary sentences. For Wittgenstein, all meaningful sentences – all that 'can be said' – picture some contingent 'state of affairs', and the totality of such states of affairs constitutes 'the world'. Since, for Wittgenstein, logical and mathematical 'truths' are true no matter what the actual facts are, they are tautologies and cannot be 'said meaningfully' but only shown. Indeed, just before his final sentence, Wittgenstein suggests that his reader understand that the sentences of *Tractatus* itself are, strictly speaking, meaningless, a rungless ladder that one climbs up to 'command a clear view'. Tautologies are the linguistic scaffolding, that must be common to any language in which the facts constituting the world can be stated.

Basic to the *Tractatus* is the view that what can be said – sentences – are 'pictures' of facts, of what, if true, is the case. Complex factual sentences break down completely into elementary sentences which picture facts by a pictorial correspondence between the arrangement of the simple names that constitute the sentences and the arrangement of simple objects of the world (whatever either of them are – for Wittgenstein, as a pure logician, could not give examples of actual names or objects – they just had to be there). Hence 'a sentence is a truth-function of elementary sentences'; and the formula '$[\bar{p}, \bar{\xi}, N(\bar{\xi})]$' is meant to enumerate all possible combinations of these elementary sentences. Although Russell, in his introduction to the *Tractatus*, suggested that Wittgenstein was describing a logically perfect language, Wittgenstein later insisted that he meant that our present languages, when meaningful, display just such logical properties and rest on just such foundations of elementary sentences. His insistence that all natural language sentences must have a consistent, determinate, and semantically complete 'deep structure' is echoed in the similar convictions of the 'generative semanticist' movement in linguistics of the 1970s. In his starkly minimalist program, Wittgenstein, unlike Russell, also wanted to do without the abstract mathematical objects of set theory and wanted to take universal statements, including 'laws' of empirical science, as indefinite conjunctions of elementary sentences.

Given his minimalism about what could be said, and hence what could be in the world, Wittgenstein held that all evaluative or mystical sentences were meaningless, or, as sentence 7 insists,

'unspeakable'; hence, necessarily, one cannot avoid being silent about such matters. But sentence 7, like the other sentences of the *Tractatus*, can exude a pregnant silence: one can show things about such matters, and they are of the highest importance, although such matters are not in the world. After all, the world is just the (entirely accidental) totality of everything that is the case; but that something is good or bad, right or wrong, cannot be a matter of factual accident. Similarly, my mind, as something that can will, also transcends and is powerless to act in the world, and what the solipsist wants to say – '*the world* is *my world*' – is correct but inexpressible.

Wittgenstein then left philosophy to become a village schoolteacher, and afterwards, rather more successfully, an architect. (Although he designed and built only one building, a private mansion, it is a classic of modern architecture. As a schoolteacher, he was a severe disciplinarian and did not get on with his colleagues or with the parents of his 9- and 10-year-old students. He once beat a girl, then denied he had done so, only to return years later to apologize to the villagers.) Wittgenstein returned to philosophy and Cambridge University in 1929 and, teaching small classes attended by luminaries such as G. E. Moore and Alan Turing, he continued gradually to reject the central claims of the *Tractatus*. Indeed, the oracular demands of the *Tractatus* now became the 'bewitchment of the mind by language'. Wittgenstein distilled these lectures and notes into the *Philosophical Investigations*, which he prepared for publication but which, because of delays and continual reworking, only appeared two years after his death.

While the *Tractatus* is a relentlessly systematic, masterful, impersonal monolog, *Investigations* traverses the 'landscape' of language and the mental in tightly woven vignettes, whose voices are both conversational and confessional: the *Tractatus* Wittgenstein, the later Wittgenstein, and the reader or listener who is introduced in lines that begin, 'You will think that…'. Philosophy now presents no theses, not even 'unspeakable' ones. Rather, it is an activity of dissolving characteristic pieces of nonsense from the *Tractatus* and elsewhere. The meaning of a word is rarely what it names. There is no 'general form of the proposition' but rather 'countless' uses of language which emphatically do not form a consistent system. There is no need to assume that whenever we say something our act has to be accompanied by some inner mental event. Philosophical avenues that lead to mind–body dualism and solipsism are energetically undermined: 'It is humiliating to have to appear like an empty tube which is simply inflated by a mind.'

VIEWS ON MIND, COGNITION, AND LANGUAGE

Because Wittgenstein insisted that philosophy was not empirical, not 'natural science', his philosophical views or vignettes constitute his thinking on mind, cognition, and language. But his 'grammatical investigations' reveal aspects of our cognitive life that are subject to empirical investigation.

For example, Wittgenstein points out that 'game' has no one meaning. Some games have rules, some do not; some have boards, pieces, cards, balls, or fields, some do not; some have winners and losers, some do not; some involve running about, some sitting still; and so on. Rather, 'game' picks out exemplars and a 'family of resemblances'. Similarly, Wittgenstein's earlier attempt to find out 'the' meaning of the proposition, and 'the' meaning of a word, was an illusory quest. However, what can mislead a philosopher can prove fertile ground for empirical researchers. In the diverse writings of researchers such as Eleanor Rosch and Amos Tversky, we find striking empirical accounts of the role of stereotypes and exemplars in our thinking. An account of the ways we may feel pulled down an illusory path also serves to illustrate the peculiar contours of our actual folk-psychological cognitive apparatus – or even illusions likely to bedevil cognitive scientists. Seeing how we can misunderstand everyday notions is also, importantly, seeing how they actually operate.

Similarly, Wittgenstein runs through a rich series of examples, drawn in part from Gestalt psychology, of 'seeing as'. Perhaps his best-known example is a line drawing which can be seen as a rabbit or as a duck. Wittgenstein insists that when someone simply sees one aspect, the proper report of him is 'he sees a rabbit' not 'he sees it as a rabbit'. Wittgenstein's point is that the empiricist philosopher's characteristic assumption that sense perception is the passively received foundation of our knowledge is mistaken. But in collecting and narrating our temptations to go wrong in such matters, Wittgenstein (1953) is again suggesting puzzles that can prove fruitful paths for modular cognitive science.

> When it looks as if there were no room for such a form between other ones you have to look for it in another dimension. If there is no room here, there is room in another dimension.

Only that would be a job for natural science, including today's modular sub-personal psychology

which investigates cognitive operations – such as visual and linguistic perception – that are not open to everyday conscious inspection.

In one of his most compelling metaphors, Wittgenstein compares our everyday language, our intentional talk about ourselves and others, to the tangled streets of the 'old city', while our scientific and technical talk he compares to neatly laid-out suburbs. His illusion in the *Tractatus* was to think that, deep down, the old city really had to be the suburbs, that our everyday talk had to translate into, or be revealed as, the 'totality of the true elementary propositions'. His view of language in the *Tractatus* is consonant with views still held by some linguistic philosophers and by some linguists, who assume that natural languages are consistent, interpreted formal systems. His critique of this view and the related claim that consciousness is peculiarly private and peculiarly foundational has inspired cognitive scientists as diverse as the philosopher Daniel Dennett and the linguist Noam Chomsky.

RESPONSES TO WITTGENSTEIN

During the 1920s, Wittgenstein occasionally talked with the physicists, mathematicians and philosophers of the Vienna Circle, many of whom admired the *Tractatus* and shared many of its views. However, the Vienna Circle confidently asserted that sensory experiences or observation sentences were elementary and foundational and constituted a criterion of meaningfulness or verifiability. This emphasis was absent from the austere *Tractatus*, although Wittgenstein's notebooks reveal that he had considered sensory experiences as candidates for elementary facts. The oracular tone of the *Tractatus* – as, later, the conversational narrative vignettes of *Investigations* – fascinated many people who had little professional interest in philosophy, logic, or the philosophy of science (Iris Murdoch's novels vividly present characters to whom reading Wittgenstein was a fundamental rite of passage). And a substantial number of philosophers, especially in the 1950s and 1960s, came to think that *Investigations* conveyed an authoritative, new and final way of doing philosophy. (Some even went as far as imitating Wittgenstein's classroom mannerisms.) Many quite various and contradictory theses have been attributed to *Investigations* by Wittgenstein zealots of one ilk or another. This was perhaps inevitable since Wittgenstein's naturally unsystematic cognitive narratives were meant, as he often announced, to stand in the way of any philosophical theses or antitheses; but that empowered the zealot to read whatever general view he liked into *Investigations*. The philosopher Stanley Cavell has compared the later Wittgenstein to the seventeenth-century epigrammatist La Rochefoucauld. As La Rochefoucauld tersely presents the paradoxical kaleidoscope of the human moral condition, so Wittgenstein sketches the scattered paradoxes, anomalies, and tempting illusions of our everyday cognitive condition. As Daniel Dennett remarks (Dennett, 1999):

> Wittgenstein continues to attract fanatics who devote their life to disagreeing with one another about the ultimate meaning of his words. These disciples cling myopically to their Wittgenstein, not realizing that there are many great Wittgensteins to choose from.

RELEVANCE OF WITTGENSTEIN'S WORK TO COGNITIVE SCIENCE

Alan Turing, who laid many of the foundations for modern computation and cognitive science, appears in the transcripts of Wittgenstein's 1939 lectures on the philosophy of mathematics. Among the dozens of transcripts of Wittgenstein's classes his students made, here alone we hear another forceful voice, Turing daring, and Wittgenstein permitting, even inviting, him, to interrupt and persistently question, and expound alternatives to, Wittgenstein's skeptical views about contradiction in formal systems and human rule following. Turing soon went on to propose the project of simulating the mentality of a human person on a computer, whether through programming, tweaking connectionist nets, equipping a 'child machine' with 'the best eyes and ears money can buy' and sending it to school, or even adding limbs and letting it 'roam the countryside'. Wittgenstein's *Investigations* can be read as a series of vivid narrative reminders of how daunting, complex, and fraught with illusions, such a project might be: how great is the gulf between the 'old city' of personal narrative experience and the formal system embodied in a machine.

Further Reading

Cavell S (1962) The availability of Wittgenstein's later philosophy. *Philosophical Review* **71**: 67–93.

Chomsky N (1995) Language and nature. *Mind* **104**(413): 1–61.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Dennett D (1999) Ludwig Wittgenstein. In: *People of the Century*, pp. 145–149. New York, NY: Simon & Schuster.

Diamond C (ed.) (1976) *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939: From the*

Notes of RG Bosanquet, Norman Malcolm, Rush Rhees, and Yorick Smythies. Ithaca, NY: Cornell University Press.

Leiber J (1991) *An Invitation to Cognitive Science*. Oxford: Blackwell.

Leiber J (1997) On what sort of speech act Wittgenstein's *Investigations* is and why it matters. *Philosophical Forum* **28**: 233–267.

Monk R (1990) *Ludwig Wittgenstein: The Duty of Genius*. New York, NY: Free Press and Maxwell Macmillan International.

Pears DF (1980) *Ludwig Wittgenstein*. Cambridge, MA: Harvard University Press.

Russell B (1968) *Autobiography*, vol. II. London: George Allen and Unwin.

Thorton T (1998) *Wittgenstein on Language and Thought: The Philosophy of Content*. Edinburgh: University of Edinburgh Press.

Wittgenstein L (1922) *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.

Wittgenstein L (1953) *Philosophical Investigations*. Oxford: Oxford University Press.

Wundt, Wilhelm

Introductory article

Robert W Rieber, John Jay College and the Graduate Center, City University of New York, New York, USA

CONTENTS

Wilhelm Wundt is generally recognized as the founder of experimental psychology. He taught the first formal academic psychology course and helped establish psychology as a separate science.

LIFE HISTORY

Wilhelm Wundt, the first modern psychologist, was born in 1832 near Heidelberg in Baden, the only child of a Lutheran pastor. He began to study under a Lutheran vicar, who was most likely his father's assistant, and in fact left home early to continue studying with him. At the age of 13, he started at a gymnasium – roughly equivalent to an American academic prep school – and at 19, he entered university. He studied medicine, but it is unlikely that he ever seriously considered practicing it. It is believed that he only did so to remain away from home and a couple of years later he switched to studying physiology. In 1855 Wundt earned his doctorate from Heidelberg in physiology. In 1857 he was appointed Dozent at Heidelberg. He served as an assistant to Hermann Helmholtz, who was a professor of physiology, between 1858 and 1864. During this time, a conception of psychology as a distinct science was beginning to emerge.

With Wundt's first publication in 1853 began the most extraordinary record of publications in the history of psychologists, with some books and almost 500 articles appearing in the next 60 or so years. Between 1858 and 1862 he published sections of his *Contributions to the Theory of Sensory Perception*, which formalized his ideas about psychology. Titchener, who was one of his greatest students, asserted that this volume outlined the program of Wundt's entire life. It dealt with a program similar to physics and chemistry for an experimental psychology. In this work, he emphasized the importance of method for scientific advancement. To him, the use of the experimental method, whenever possible, was mandatory. He also stressed in *Contributions* that psychology should begin with simple questions and then progress to the difficult metaphysical ones. He used physiological methodology in dealing with psychological problems. Wundt replaced the old method of meditation with a more exact and exacting method of introspection.

In 1864 Wundt was appointed assistant professor. In 1866 he was chosen to represent Heidelberg in the Baden Chamber – a distinguished scholarly group of the period – but he soon resigned because of time restraints. Wundt created and developed the first school of psychological thought, structuralism, whose basic building block was sensation. He advanced his status rapidly at Heidelberg and, starting in 1867, he taught a course at Heidelberg entitled 'Physiological Psychology', which was the first formal offering of an academic course of this nature. He believed that physiology follows one method to achieve knowledge and that psychology follows another; they are two bodies of knowledge. The manner of connection between elements in the subject's immediate experience is entirely different from those occurrences studied by physiology. One of his most important books, *Principles of Physiological Psychology*, was framed from his lecture notes and published in parts between 1873 and 1874. When the last edition was published in 1911, it had gone through six editions and comprised three large volumes. It is regarded as one of the most important books in the history of psychology.

Wundt was not seeking to study the relation between the body and the mind but the relation between sensation and the process of psychological judgement. The result was a purely psychological interpretation with no appeal to the relation of stimulus and sensation. Physiological psychology was concerned with the process of excitations from stimulation of the sense organs, through sensory neurons to the lower and higher brain centers, and from these centers to the muscles. He established introspection as psychology's distinguishing methodology. He proceeded

in a significantly different fashion from earlier versions of introspection, which actually should be called 'meditation'. Wundt refined the conscious elements of meditation and combined them with experiment. He claimed that in psychology pure self-observation is insufficient. He believed that in the laboratory, unless they can be related to an external or measurable response, observations have no scientific usefulness.

Within a year after the appearance of *Principles of Physiological Psychology*, he was offered a professorship at Zurich and later in 1875 he became professor of philosophy at Leipzig, where he worked for the next 45 years. He began one of the first experimental laboratories of psychology in the world. This laboratory became a focus for those with a serious interest in psychology. This institute of psychology, which he established at Leipzig, eventually became one of the most famous centers for the study of psychology in Europe. Philosophy and psychology students from all over the continent as well as from the United States sought out the opportunity to spend some time with or pursue a degree with Wundt at Leipzig. These students, though not always in agreement with Wundt in their systematic views, were careful to tow the line while at Leipzig, and once away from Wundt they often diverged from the details of his systematic psychology. What they shared with Wundt, however, was an enthusiasm for experimental, laboratory psychology. The laboratory atmosphere fostered specific research studies that appeared as articles in journals. All subsequent psychological laboratories in their early years were closely modeled on the Wundt design.

The availability of co-workers was important. Since a worker can hardly be the experimenter and observer at one and the same time, the experimenter of one study was available as a subject for another. Lest this point be dismissed as trivial, it is pertinent to indicate that introspection as practiced in Wundt's laboratory was not a skill acquired without a period of rigorous apprenticeship. Getting at the elements of experience required arduous training. Moreover, even if nonstudent assistants could be trained as subjects, the nature of the task to which they were assigned would have required payment.

Wundt began to publish a journal, *Philosophical Studies*, in 1881, which was the first journal in the German language committed equally to philosophy and psychology. It often included reports of experimental studies that had been done in his laboratory. About one hundred experimental studies appeared in *Philosophical Studies* during its 20-odd years, most of them having been made in his laboratory, or conducted by Wundt's students soon after they left Leipzig. Consequently, the research bore heavily the impress of Wundt's direction, since typically he assigned the problem on which a particular student was to work. About half of the research studies dealt with problems in sensation and perception, while others were about reaction, attention, feeling, and association.

In his 1893 edition of *Principles of Physiological Psychology*, he published the 'tridimensional theory of feeling': feelings were classified as pleasant or unpleasant, tense or relaxed, excited or depressed. A given feeling might equally be a combination of one of each of the categories.

A survey of the research from his laboratory shows that Wundt did not occupy himself with developing new kinds of experiments. The methods he used were not particularly new. Students of the psychology and physiology of the senses owe much to previous work, particularly to that of Helmholtz. Reaction time studies again owe something, not only to Helmholtz, but also to Donders, whereas the association study can be attributed to Galton. Even the study of feeling, where Wundt was at his most original, in a theoretical sense, depended on the extension of Fechner's method of impression to paired comparisons; studies of expression were linked to the utilization of already existing methods for studying pulse, breathing, and the like. There had been antecedent studies even for attention. Wundt's experimental contribution was to reduce to quantitative terms the research areas already extant.

In 1875, Wundt began sponsoring doctoral dissertations and by 1919 the total had reached 186. He accepted students from all over Europe and even America. These students managed to show extreme breadth of activity after the severely rigorous pure training they had received. Not many of them would follow the detail of Wundt's system. What they took away from Leipzig, however, was a belief in the importance of the laboratory and psychological research.

By 1902, with the fifth edition of the *Principles of Physiological Psychology*, there was no longer any doubt as to the legitimacy of his endeavors. Already, he was concerned with differing trends within the field itself; as well as work that deviated from his own. Wundt was not personally interested in the application of psychology, despite the fact that many of his students devoted much of their careers to it. He dismissed any area of psychology that dared to violate the rules of introspection. Like

many German professors, he had a very low tolerance for opinions that differed from his own. Work other than that of his students received severe criticism. He argued that one should not give the name of experimental psychology to each and every operation that brings about a change in consciousness.

He believed that, when problems more difficult than those of perception and memory are considered, experiment is not feasible. He believed that experimental methodology was not applicable to extremely elaborate processes such as reason, memory, thought, and the like. Cultural or ethnic psychology must be used for those problems. Wundt thought that language, myth, and custom should be used to study the higher mental processes. Language, he believed, was the way to understand thought. He began writing the *Volkerpsychologie*, which means cultural psychology, in 1900. Nine more volumes were published by 1920. Cultural psychology began as the investigation of the various, still-existing stages of mental development in humankind.

To Wundt, psychology is the science that investigates the facts of consciousness and cannot be based on metaphysical assumptions of any sort. He sought to create a systematic structure so that every possible experience was represented. He believed that all our experiences are complex and must be analyzed introspectively. Wundt attempted to measure experience so that others could repeat his procedures. He encouraged empiricism and rejected rationalism.

Wilhelm Wundt published his autobiography, *Erlebtes und Erkanntes*, in 1920, and then died on 31 August 1920, two weeks after his 88th birthday. Although it has been estimated that he published 53 000 pages, Wundt did not have many original ideas or perspectives; he simply refined older ideas so that a better, more complete picture emerged. His method of introspection did not remain as a fundamental tool of psychological experimentation beyond the early 1920s. Wundt's greatest contribution was to show that psychology could be a valid experimental science. Perhaps if he had been more original, psychology's appearance as a separate discipline would have been postponed.

Further Reading

Boring EG (1957) *History of Experimental Psychology*, 2nd edn. New York: Appleton/Century/Crofts.

Boring EG (1942) *Sensation and Perception in the History of Experimental Psychology*. New York: Appleton/Century/Crofts.

Krantz DL (ed.) (1969) *Schools of Psychology: A Symposium*. New York: Appleton/Century/Crofts.

Rieber RW and Salzinger K (eds) (1998) *Psychology of Theoretical Perspectives*. APA Press.

Wundt W (1894) *Lectures on Human and Animal Psychology*, 2nd German edn, translated by JE Creighton and EB Titchener. New York: Macmillan.

Wundt W (1904) *Principles of Physiological Psychology*, 5th German edn, vol. 1, translated by EB Titchener. New York: Macmillan.

Wundt W (1912) *An Introduction to Psychology*, 2nd edn, translated by R Pintner. New York: Macmillan.

Wundt W (in press) *History and the Making of a Scientific Psychology*. In: Rieber RW and Robinson D (eds). New York: Plenum.

ACT

Intermediate article

Christian Lebiere, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

CONTENTS

The ACT theory of cognition has been used to model a wide range of cognitive tasks. The latest version of the theory, ACT-R, is a hybrid architecture which combines a symbolic production system with a subsymbolic level of neural-like activation processes.

INTRODUCTION

In the twentieth century, cognitive psychology produced enormous amounts of data on all aspects of human behavior. The challenge of cognitive science in the twenty-first century is to provide a coherent, integrated and systematic scientific explanation of these phenomena. Computational modeling has become an increasingly popular tool in a wide range of sciences, from mathematics and physics to economics and the social sciences. By combining precision, lacking in many verbal theories, with the ability to deal with almost limitless complexity, a limitation of mathematical theories, computational modeling provides the most promising approach to a unified theory of cognition.

The 'adaptive control of thought' (ACT) theory of cognition attempts to provide such a unified framework by defining an integrated cognitive architecture that can be applied to a wide range of psychological tasks. The most recent version of the theory, ACT-R, is a hybrid architecture which combines elements of the symbolic and connectionist frameworks. By using a symbolic production system to provide the structure of behavior, it has a direct interpretive link to psychological data, and thus inherits a tractable level of analysis. By associating to each symbolic knowledge structure real-valued activation quantities that control its application, it enables the use of statistical learning mechanisms that provide neural-like adaptivity and generalization. The long-term goal of the ACT theory is to provide constrained computational models of a wide range of cognitive phenomena.

PRODUCTIONS AND CHUNKS AS ATOMIC COMPONENTS OF THOUGHT

How to represent knowledge is the first and most important question concerning the design of a cognitive system. As with data structures in a computer program, the knowledge structures assumed by a cognitive theory will fundamentally determine its characteristics: what knowledge can be represented, how it can be accessed and which learning processes can be used to acquire it. Since many cognitive systems, including ACT, are implemented as computer simulations, it is important to define which representational assumptions constitute theoretical claims and which are merely notational conventions. The ACT theory has from its inception made three theoretical assumptions regarding knowledge representation. The first assumption is known as the procedural–declarative distinction; that is, that there are two long-term repositories of knowledge, a procedural memory and a declarative memory. The second assumption is that 'chunks' are the basic units of knowledge in declarative memory. The third assumption is that production rules are the basic units of knowledge in procedural memory. These assumptions will be examined in detail in the rest of this section.

Unlike other cognitive theories and frameworks such as Soar or connectionism, ACT makes a fundamental distinction between declarative and procedural knowledge. This is not just a notational distinction, but a fundamental psychological claim about the existence of distinct memory systems

with different properties. The best way to state the procedural–declarative distinction is in terms of a production system framework. Declarative memory holds factual knowledge, such as the knowledge that George Washington was the first president of the United States or that $3 + 4 = 7$, while procedural memory holds rules that access and modify declarative memory. This distinction corresponds closely to the common operational definition that declarative knowledge is verbalizable while procedural knowledge is not, with the caveat that some declarative knowledge might exist without being directly reportable because one might lack the necessary procedural knowledge to access and express it. These two memory systems have some common features, such as the build-up and decay of strength, but they also have distinct properties: for example, more flexible access to declarative than to procedural knowledge, and different acquisition and retention characteristics. Some recent results in cognitive neuroscience can be interpreted as supporting the procedural–declarative distinction: for example, results showing that damage to the hippocampus inhibited the creation of new declarative memories but not of new procedural memories (Squire, 1992).

The basic units of declarative memory are called chunks. The purpose of a chunk is to organize a set of elements (either chunks themselves or more basic perceptual components) into a long-term memory unit. Chunks can only contain a limited number of elements: as few as two, often three (which it has been argued is a theoretical optimum), and seldom more than five or six. Elements in a chunk assume specific relational roles: for example, in the chunk encoding '$3 + 4 = 7$', '3' is the first addend, '+' is the operator, '4' is the second addend and '7' is the sum. Finally, chunks can be organized hierarchically, since the elements of a chunk can be chunks themselves. For example, to memorize a long sequence, chunks of finite length can be used to store short pieces of the sequence and can themselves be aggregated into other chunks to encode the full sequence. Chunks have two possible origins: either as direct encodings of objects in the environment or as long-term encodings of particular internal elaborations, called goals. Thus when reading a sentence, every word read, together perhaps with some environmental characteristics such as its position in the sentence, constitutes a chunk. But the understanding of the sentence, which is the goal of the processing, is also available as one or more chunks holding an elaboration of its meaning as a result of the cognitive processing.

Production rules are the basic units of procedural memory. Production rules encode cognitive skills as condition–action pairs. A production rule tests the contents of the current goal and perhaps of declarative memory, then executes one or more actions, which can include modifying the current goal or changing the external environment. Just as for chunks, the size and complexity of productions are limited in that they only perform a limited number of retrievals from declarative memory (usually a single one), and those retrievals are performed sequentially. ACT makes four claims related to production rules. The first is 'modularity', i.e, that procedural knowledge takes the form of production rules that can be acquired and deployed independently. Complex skills can be decomposed into production rules that capture significant regularities in human behavior (Anderson, Conrad and Corbett, 1989). For example, performance in the learning of a programming language, which doesn't show any regularities when organized with respect to the number of problems, exhibits a very regular learning pattern when organized with respect to the production rules used in solving each problem.

The second claim is 'abstraction', i.e., that productions are general, applying across a wide range of problems. The third claim is 'goal factoring', which moderates the claim of abstraction by making each production specific to a particular goal type. Thus, a production that performs addition can apply to any addition problem (abstraction) but only to addition problems (goal factoring).

The fourth claim is 'condition–action asymmetry'. For example, while a chunk holding a multiplication fact can be used to solve either a multiplication or a division problem, productions used to compute the answer to a multiplication problem (e.g. by repeated addition) cannot be used to find the answer to a division problem.

According to the ACT theory, chunks and procedural units thus constitute the atomic components of thought.

FROM THE HAM THEORY OF MEMORY TO A FRAMEWORK FOR MODELING COGNITION

ACT has its roots in the 'human associative memory' (HAM) theory of human memory (Anderson and Bower, 1973), which represented declarative knowledge as a propositional network. HAM was implemented as a computer simulation in an attempt to handle complexity and to specify precisely how the model applied to the task, thus

overcoming the major limitations of the mathematical theories of the 1950s and 1960s. Although it fell short of these goals, the theory was afterwards developed in significant ways.

The next step was the introduction of the first instance of ACT, ACTE (Anderson, 1976). It combined HAM's theory of declarative memory with a production system implementation of procedural memory, thus precisely specifying the process by which declarative knowledge was created and applied, and added basic activation processes to link procedural and declarative memory. While the distinction between procedural and declarative knowledge had little support at the time, it has found increasing popularity and support from recent neuroscientific evidence for a dissociation between declarative and procedural memories. The next major step in the evolution of the ACT theory was the ACT* system (Anderson, 1983), which added a more neural-like calculus of activation and a more plausible theory of production rule learning. ACT* was successfully applied to a wide range of psychological phenomena and had a profound influence on cognitive science, but, although some computer simulations were available, it existed primarily as a verbally specified mathematical theory.

The first computational implementation of the theory to be widely adopted was ACT-R (Anderson, 1993). ACT-R was introduced to capitalize on advances in skill acquisition (Anderson and Thompson, 1989) to improve production rule learning, and to tune the subsymbolic level to the structure of the environment to reflect the rational analysis of cognition (Anderson, 1990). Due to these theoretical advances as well as computational factors such as the standardization of the implementation language, Common LISP, and the exponentially increasing power of desktop computers, ACT-R has been adopted as a cognitive architecture by a growing group of researchers. The needs of that user community to apply ACT-R to an increasingly diverse set of tasks, and a growing need for neural plausibility, led to further advances which were embodied in a new version of the theory, ACT-R 4.0 (Anderson and Lebiere, 1998). As suggested by the use of version numbers, the changes in the theory were relatively minor and largely consisted of a reduction in the grain size of chunks and productions.

Through its widening range of applications and the increasing constraints on model development, ACT-R can be regarded as being well on the way to achieving HAM's long-term goal of providing a rigorous computational framework for modeling cognition. The technique of using computational simulation rather than mathematical analysis to provide tractability in complex domains has been widely adopted not only in cognitive modeling but in many other sciences as well. Because it tightly integrates a symbolic production system with a neural-like activation calculus, ACT can be termed a hybrid activation-based production system cognitive architecture. As such, it constitutes a general framework for modeling cognition (another example of such a framework being connectionism). While frameworks can often make general qualitative predictions, in order to obtain precise quantitative predictions for specific experiments they need to be instantiated into detailed theories, such as the various members of the ACT family, that exactly specify the system's mechanisms and equations. A serious potential problem resulting from the generality of frameworks is the possibility of instantiating them into competing theories and incorrectly citing those theories' accomplishments as support for the framework itself. To avoid that danger and to ensure cumulative progress in the development of the ACT theory, Anderson and Lebiere made available on the web every simulation described in Anderson and Lebiere (1998), and promised that the account provided by those models would remain valid in future instantiations of the theory (*The ACT Web*: see Further Reading list).

THE MECHANISMS OF ACT

While the specific mechanisms of the ACT theory have changed significantly over more than 20 years of evolution, the assumptions that underlie the theory have remained fairly constant. ACT operates in continuous time, predicting specific latencies for each step of cognition, such as production firing or declarative retrieval. At the symbolic level, the procedural–declarative distinction states that there are two memory systems, with a production rule component operating on a declarative component. Declarative knowledge is composed of chunks, each having a limited number of components, which can be described alternatively in terms of a propositional network. Procedural knowledge is composed of production rules, or condition–action pairs, which are the basic units of skills. Conditions apply to the state of declarative memory, while actions result in changes in declarative memory. Cognition is goal-directed and maintains at all times a current goal which a production must match in order to apply. At the subsymbolic level, real-valued quantities such as chunk activation and production utility control the application of that

symbolic knowledge. Those subsymbolic parameters are learned, and reflect the past use of their respective symbolic structures. Activation computation is a dynamic process that involves the spreading of activation from a set of sources, usually the contents of the current goal, to related nodes, as well as time-related decay.

ACQUIRING PRODUCTIONS: ACT* AND PUPS

The learning of skills in the form of production rules is one of the more complex mechanisms of the ACT architecture, and has seen the most fundamental changes over the course of the development of the theory. In ACT*, the first version of the theory to provide a comprehensive account of the learning of production rules, procedural learning is accomplished by a set of mechanisms that compile into production rules the process of interpreting declarative knowledge. In other words, procedural skills are learned by doing. For example, if one dials a telephone number or enters one's password by explicitly remembering the number and then iteratively identifying and keying each digit (or letter), the knowledge compilation process would create a production that directly encodes and keys each character without retrieving it from declarative memory. Knowledge compilation is accomplished by two separate processes: composition, which takes a sequence of production firings and compiles them into a single equivalent production; and proceduralization, which takes an existing production and encodes the result of declarative retrievals (in our example, the phone number or password) directly into a new production. Unlike the learning of a declarative fact, which can take place in a single episode, the learning of a procedural skill in the form of production rules requires many iterations for the new knowledge to be refined into its final form and ready to be applied. ACT* has three mechanisms that take new production rules and tune them for optimal performance. The generalization process broadens the applicability of new production rules by making their conditions more general. The discrimination process performs the opposite task, narrowing the applicability of new rules by making their conditions more specific (in our example, the production would be specific to the person or account associated with the number or password). Finally, the strengthening process increases the production strength, allowing it to apply faster and more often.

Anderson and Thompson (1989) introduced a different conception of production learning in their 'penultimate production system' (PUPS). Following empirical studies indicating that analogical problem solving played a fundamental role in skill acquisition, they proposed a new mechanism of production creation based on an analogy process. Instead of automatically creating new productions as a function of an interpretive process, PUPS creates productions to encode the solving of a current problem by analogy with a solution to a previous problem encoded in declarative memory. The analogy process discovers a mapping in declarative memory from problem to solution, which is then encoded in a new production that can solve the problem directly without referring to previous examples. Analogy can therefore be regarded as a mechanism for generalizing from examples. Since it operates on explicit memory structures instead of an automatic goal-based trace of operations (as in ACT*), it allows for the addition to the example of chunks of conditions and heuristics to guide the generalization and discrimination processes that produce the final form of the new productions. This results in a more reliable and controlled mechanism.

RATIONAL ANALYSIS AND ITS IMPLICATIONS FOR COGNITIVE ARCHITECTURE

Inspired by Marr's theory of information-processing levels, Anderson (1990) introduced his 'rational analysis' of human cognition, based on the assumption of a rational level used to analyze the computations performed by human cognition. The general 'principle of rationality' states that a cognitive system operates at all times to optimize the adaptation of the behavior of the organism. This does not imply that human cognition is perfectly optimal, but it helps explain why cognition operates the way it does at the algorithmic level given its physical limitations at the biological level and the optimum defined by the rational level that it attempts to implement. Anderson's analysis provides strong guidance on theory development, because given a particular framework (say, an activation-based production system) it strongly constrains the set of possible mechanisms to those that satisfy the rational level.

The rational analysis can be applied to several aspects of human cognition, including memory, categorization, causal inference and problem solving. The task of human memory is analyzed in terms of a Bayesian estimation of the probabilities of needing a particular memory at a particular point in time. The analysis accounts for effects of

recency, frequency and spacing; effects of context and word frequency; and priming and fan effects. It also provides an interpretation of the concept of activation in ACT-R as the logarithm of the odds that the corresponding chunk needs to be retrieved from memory. The rational analysis of categorization can be interpreted as supporting the creation of chunk types in ACT-R. Finally, the analysis of problem solving in terms of expected utilities of procedural operators led to the refinement of the conflict resolution process in ACT-R. In summary, the rational analysis of cognition provided strong guidance for the development of learning mechanisms in ACT-R (the R stands for 'rational') that automatically tune the subsymbolic level to the structure of the environment.

CONNECTIONIST CONSIDERATIONS: ACT-RN

Lebiere and Anderson (1993) describe ACT-RN, which is an attempt to implement ACT-R using standard connectionist constructs such as Hopfield networks and feedforward networks. Symbols are represented using distributed patterns of activation over pools of units. The current goal, or focus of attention, is located in a central memory that holds the components of the goal, which are the sources of activation in ACT-R. Separate declarative memories for each chunk type are implemented using associative memories in the form of simplified Hopfield networks with a separate pool of units for each component of the chunk. ACT-R's goal stack is implemented as a separate declarative memory that associates each goal to its parent goal. Procedural memory consists of pathways between central memory and declarative memories. Each production is represented as a single unit that tests the contents of the current goal in central memory, then, if successful, activates the proper connections to perform a retrieval from a single declarative memory (possibly the goal stack) and update the goal with the retrieval results. There is a rough correspondence between these constructs and neural locations. Goal memory can be associated with the prefrontal cortex, since damage in that area has been associated with loss of executive function. Declarative memory is distributed throughout the posterior cortex, with each type corresponding to hypercolumns or small cortical areas. Procedural memory might be located in the basal ganglia, which have extensive connections to all cortical areas and could therefore implement the functionality of production rules.

As a practical system, ACT-RN was found unsatisfactory in a number of ways. It only provided a partial implementation of ACT-R, with some features and mechanisms being too complex or difficult to map onto a connectionist substrate. While some models adapted well to the connectionist implementation, others, even simple ones, ran into computational hurdles. More fundamentally, it was an imperfect implementation of the ACT-R standard which only approximately reproduced the ACT-R mechanisms and equations. Nevertheless, ACT-RN provided the main impetus for further development of the theory. Some features of ACT-R that were too complex to be implemented in ACT-RN, such as complicated representational constructs in chunks and powerful pattern-matching primitives in productions, were abandoned in the later versions of the theory as being too computationally powerful for any neural implementation. This resulted in a welcome simplification of the theory. Conversely, a feature of ACT-RN which the connectionist implementation provided naturally – generalization based on distributed representations – was added to the theory in the form of similarity-based partial matching of production conditions (Lebiere *et al.*, 1994). Thus, although implementing ACT-R in a neural network did not directly result in a practical system, it did provide a functional theory of neural organization and imposed a strong direction on further developments of the ACT theory.

THE MECHANISMS OF ACT-R: INTEGRATING SYMBOLIC AND SUBSYMBOLIC PROCESSING AND LEARNING

The power of ACT-R as a hybrid architecture of cognition lies in its tight integration of the symbolic and subsymbolic levels. Because of this integration, it is able to combine the most desirable characteristics of symbolic systems, such as structured behavior and ease of analysis, with those of connectionist networks, such as generalization and fault-tolerance, while avoiding their most serious shortcomings, such as the overly deterministic behavior of symbolic systems and the intractability of learning characteristic of connectionist networks. At the symbolic level, ACT-R operates sequentially – only one production can fire in each cycle and only one chunk can be retrieved from memory at a time – corresponding to the basic sequential nature of human cognition. However, each of the basic steps of production selection and of declarative memory retrieval involves the parallel consideration of all

relevant productions and memory chunks, reflecting the massively parallel nature of the human brain.

A typical production cycle in ACT-R works as follows. The conflict resolution process attempts to select the best production that matches the current goal. To that effect, the expected production utilities of all matching productions are computed in parallel and the best production is selected. Typically, that production attempts a retrieval of information from declarative memory. The activations of the relevant memory chunks are computed concurrently, as: the sum of the base-level activations, reflecting the history of use of each chunk; the activation spreads from the components of the goals, reflecting the specificity of that chunk to the current context; a partial matching penalty, allowing generalization to similar patterns; and a noise component providing stochasticity. Once the retrieval is complete, the activation parameters of the chunks involved are automatically adjusted by the subsymbolic learning mechanisms to reflect this experience. The production then executes its action, which typically consists of modifying the goal to incorporate the results of the declarative retrieval and perhaps performing an external action. If a goal is accomplished, the subsymbolic parameters controlling the utility of the productions involved in solving that goal will be automatically learned, and a chunk encoding the results of the goal will enter declarative memory.

Thus, the production system part of ACT-R provides the basic synchronization of a massively parallel system into a meaningful sequence of cognitive steps, while the subsymbolic part is continuously tuned to the statistical nature of the environment to provide the adaptivity characteristic of human cognition.

MATCHING HUMAN PERFORMANCE IN DIVERSE DOMAINS

The idea of using computational precision to eliminate the looseness of the merely verbal mapping between model and task is embodied in what Anderson and Lebiere (1998) call the 'no-magic doctrine', which consists of the following six tenets:

1. Theories must be experimentally grounded. To avoid unprincipled degrees of freedom in the mapping between task stimuli and model representations, ACT-R includes a 'perceptual motor' component called ACT-R/PM which interacts with the task through the same interface as human subjects.
2. Theories must provide a detailed and precise accounting for the data. Because of its experimental grounding, ACT-R makes precise predictions about every aspect of empirical data, including choice percentages, response latencies, etc.
3. Models must be learnable through experience. ACT-R has mechanisms capable of learning symbolic chunks and productions as well as their subsymbolic parameters.
4. Theories must be capable of dealing with complex cognitive phenomena. ACT-R is applicable to tasks ranging from sub-second psychology experiments to complex environments involving substantial knowledge and learning that may take hours.
5. Theories must have principled parameters. The parameters attached to symbolic knowledge structures are learned from experience. The modeler sets the architectural parameters, but variations are increasingly constrained and understood across tasks.
6. Theories must be neurally plausible. ACT-R is situated at a level of abstraction higher than actual brain structures, but the need to provide a plausible mapping between theoretical constructs and actual brain structures imposes useful and powerful constraints on theory development.

ACT-R has been applied to an increasing variety of cognitive tasks in domains as diverse as memory, categorization, problem solving, analogy, scientific discovery, human–computer interaction, decision theory and game theory. In each domain, ACT-R predicts a wide range of measurable aspects of human behavior at a very fine scale, including latency, errors, learning, eye movements and individual differences.

SUMMARY

ACT is a hybrid cognitive architecture that combines a symbolic production system with a subsymbolic level of neural-like activation processes that control the application of the symbolic structures. Learning mechanisms provide for the acquisition of symbolic knowledge and the statistical tuning of the subsymbolic layer to the structure of the environment, as specified by the rational analysis of cognition. The ACT architecture has been successfully applied to a wide variety of cognitive tasks to accurately predict many aspects of human behavior. (*See* **Computer Modeling of Cognition: Levels of Analysis**)

References

Anderson JR (1976) *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Anderson JR (1983) *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson JR (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.
Anderson JR (1993) *Rules of the Mind*. Hillsdale, NJ: Erlbaum.
Anderson JR and Bower GH (1973) *Human Associative Memory*. Washington, DC: Winston and Sons.
Anderson JR, Conrad FG and Corbett AT (1989) Skill acquisition and the LISP Tutor. *Cognitive Science* **13:** 467–506.
Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.
Anderson JR and Thompson R (1989) Use of analogy in a production system architecture. In: Ortony A *et al*. (eds) *Similarity and Analogy*, pp. 367–397. New York, NY: Cambridge University Press.
Lebiere C and Anderson JR (1993) A connectionist implementation of the ACT-R production system. In: *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, pp. 635–640. Hillsdale, NJ: Erlbaum.
Lebiere C, Anderson JR and Reder LM (1994) Error modeling in the ACT-R production system. In: *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, pp. 555–559. Hillsdale, NJ: Erlbaum.
Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review* **99**: 195–232.

Further Reading

The ACT Web. [http://act.psy.cmu.edu]
Anderson JR (1987) Skill acquisition: compilation of weak-method problem solutions. *Psychological Review* **94**: 192–210.
Anderson JR (1991) The adaptive nature of human categorization. *Psychological Review* **98**: 409–429.
Anderson JR, Bothell D, Lebiere C and Matessa M (1998) An integrated theory of list memory. *Journal of Memory and Language* **38**: 341–380.
Anderson JR, John BE, Just MA *et al*. (1995). Production system models of complex cognition. In: *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pp. 9–12. Hillsdale, NJ: Erlbaum.
Anderson JR, Matessa M and Lebiere C (1997) ACT-R: A theory of higher level cognition and its relation to visual attention. *Human Computer Interaction* **12**: 439–462.
Anderson JR, Reder LM and Lebiere C (1996) Working memory: activation limitations on retrieval. *Cognitive Psychology* **30**: 221–256.
Anderson JR and Schooler LJ (1991) Reflections of the environment in memory. *Psychological Science* **2**: 396–408.
Corbett AT, Koedinger KR and Anderson JR (1997) Intelligent tutoring systems. In: Helander MG, Landauer TK and Prabhu P (eds) *Handbook of Human–Computer Interaction*, 2nd edn. Amsterdam: Elsevier.
Newell A (1973a) You can't play twenty questions with nature and win: projective comments on the nature of this symposium. In: Chase WD (ed.) *Visual Information Processing*, pp. 283–310. New York, NY: Academic Press.
Newell A (1973b) Production systems: models of control structures. In: Chase WG (ed.) *Visual Information Processing*, pp. 463–526. New York, NY: Academic Press.

Adaptive Resonance Theory

Advanced article

Stephen Grossberg, Boston University, Boston, Massachusetts, USA

CONTENTS

Adaptive resonance theory is a cognitive and neural theory about how the brain develops and learns to recognize and recall objects and events throughout life. It shows how processes of learning, categorization, expectation, attention, resonance, synchronization, and memory search interact to enable the brain to learn quickly and to retain its memories stably, while explaining many data about perception, cognition, learning, memory, and consciousness.

INTRODUCTION

The processes whereby our brains continue to learn about, recognize, and recall a changing world in a stable fashion throughout life are among the most important for understanding cognition. These processes include the learning of top-down expectations, the matching of these expectations against bottom-up data, the focusing of attention upon the expected clusters of information, and the development of resonant states between bottom-up and top-down processes as they reach an attentive consensus between what is expected and what is there in the outside world. It has been suggested that all conscious states in the brain are resonant states, and that these resonant states trigger learning of sensory and cognitive representations. The models which summarize these concepts are called 'adaptive resonance theory' (ART) models. ART was introduced by Grossberg in 1976 (see Carpenter and Grossberg, 1991), along with rules for competitive learning and self-organizing maps. Since then, psychophysical and neurobiological data in support of ART have been reported in experiments on vision, visual object recognition, auditory streaming, variable-rate speech perception, somatosensory perception, and cognitive–emotional interactions, among others (e.g. Carpenter and Grossberg, 1991, 1994; Grossberg, 1999a,b; and Grossberg and Merrill, 1996). In particular, ART mechanisms seem to be operative at all levels of the visual system, and these mechanisms may be realized by known laminar circuits of visual cortex. It is predicted that the same circuit realization of ART mechanisms will be found, suitably specialized, in the laminar circuits of all sensory and cognitive neocortex.

THE STABILITY–PLASTICITY DILEMMA

We experience the world as a whole. Although myriad signals relentlessly bombard our senses, we somehow integrate them into unified moments of conscious experience that cohere despite their diversity. Because of the apparent unity and coherence of our awareness, we can develop a sense of self that can gradually mature with our experiences of the world. This capacity lies at the heart of our ability to function as intelligent beings.

The apparent unity and coherence of our experiences is all the more remarkable when we consider several properties of how the brain copes with the environmental events that it processes. First and foremost, these events are highly context-sensitive. When we look at a complex picture or scene as a whole, we can often recognize its objects and its meaning at a glance, as in the picture of a familiar face. However, if we process the face piece by piece, as through a small aperture, then its significance may be greatly degraded. To cope with this context-sensitivity, the brain typically processes pictures and other sense data in parallel, as patterns of activation across a large number of feature-sensitive nerve cells, or neurons. The same is true for senses other than vision, such as audition. If the sound of the word 'go' is altered by clipping

off the vowel 'o', then the consonant 'g' may sound like a chirp, quite unlike its sound as part of the word 'go'.

During vision, all the signals from a scene typically reach the photosensitive retinas of the eyes at virtually the same time, so parallel processing of all the scene's parts begins at the retina itself. During audition, successive sounds reach the ear at different times. Before an entire pattern of sounds, such as the word 'go', can be processed as a whole, it needs to be recoded, at a later processing stage, into a simultaneously available spatial pattern of activation. Such a processing stage is often called a working memory, and the activations that it stores are often called short-term memory (STM) traces. For example, when you hear an unfamiliar telephone number, you can temporarily store it in working memory while you walk over to the telephone and dial the number.

In order to determine which of these patterns represent familiar events and which do not, the brain matches the patterns against stored representations of previous experiences that have been acquired through learning. The learned experiences are stored in long-term memory (LTM) traces. One difference between STM and LTM traces concerns how they react to distractions. For example, if you are distracted by a loud noise before you dial an unfamiliar telephone number, its STM representation can be rapidly reset so that you forget it. On the other hand, you will not normally forget the LTM representation of your own name.

How does new information get stably stored in LTM? For example, after seeing a movie just once, we can tell our friends many details about it later on, even though the scenes flashed by very quickly. More generally, we can quickly learn about new environments, even if no one tells us how the rules of the environments differ. We can rapidly learn new facts, without being forced to just as rapidly forget what we already know. We do not need to avoid going out into the world for fear that, in learning to recognize a new friend's face, we will suddenly forget our parents' faces. This is sometimes called the problem of catastrophic forgetting.

Many learning algorithms can forget catastrophically. But the brain is capable of rapid yet stable autonomous learning of huge amounts of data in an ever-changing world. Discovering the brain's solution to this problem is as important for understanding ourselves as it is for developing new pattern recognition and prediction applications in technology.

The problem of learning quickly and stably without catastrophically forgetting past knowledge may be called the stability–plasticity dilemma. It must be solved by every brain system that needs to rapidly and adaptively respond to the flood of signals that subserves even the most ordinary experiences. If the brain's design is parsimonious, then we should expect to find similar design principles operating in all the brain systems that can stably learn an accumulating knowledge base in response to changing conditions throughout life. The discovery of such principles should also clarify how the brain unifies diverse sources of information into coherent moments of conscious experience.

LEARNING, EXPECTATION, ATTENTION, AND RESONANCE

Humans are intentional beings, who learn expectations about the world and make predictions about what is about to happen. Humans are also attentional beings, who focus their processing resources upon a restricted amount of incoming information at any time. Why are we both intentional and attentional beings, and are these two types of process related? The stability–plasticity dilemma, and its solution using resonant states, provides a unifying framework for understanding these questions.

Suppose you were asked to 'find the yellow ball within half a second, and you will win a $10 000 prize'. Activating an expectation of 'yellow balls' enables more rapid detection of a yellow ball, and with a more energetic neural response, than if you were not looking for one. Sensory and cognitive top-down expectations lead to 'excitatory matching' with confirmatory bottom-up data. On the other hand, mismatch between top-down expectations and bottom-up data can suppress the mismatched part of the bottom-up data, and thereby focus attention upon the matched, or expected, part of the bottom-up data.

This sort of excitatory matching and attentional focusing on bottom-up data using top-down expectations may generate resonant brain states: When there is a good enough match between bottom-up and top-down signal patterns, between two or more levels of processing, their positive feedback signals amplify and prolong their mutual activation, leading to a resonant state. The amplification and prolongation of the system's fast activations are sufficient to trigger learning in the more slowly varying adaptive weights that control the signal flow along pathways from cell to cell. Resonance thus provides a global context-sensitive indicator that the system is processing data worthy of learning.

ART proposes that there is an intimate connection between the mechanisms that enable us to learn quickly and stably about a changing world and the mechanisms that enable us to learn expectations about such a world, test hypotheses about it, and focus attention upon information that we find interesting. ART also proposes that, in order to solve the stability–plasticity dilemma, resonance must be the mechanism that drives rapid new learning.

Learning within the sensory and cognitive domains is often 'match learning'. Match learning occurs only if a good enough match occurs between bottom-up information and a learned top-down expectation that is specified by an active recognition category, or code. When such a match occurs, previously learned knowledge can be refined. If novel information cannot form a good enough match with the expectations that are specified by previously learned recognition categories, then a memory search, or hypothesis testing, is triggered, which leads to selection and learning of a new recognition category, rather than catastrophic forgetting of an old one. (Figure 1 illustrates how this happens in an ART model.) In contrast, learning within spatial and motor processes could be 'mismatch learning' that continuously updates sensory–motor maps or the gains of sensory–motor commands. As a result, we can stably learn what is happening in a changing world, thereby solving the stability–plasticity dilemma, while adaptively updating our representations of where objects are and how to act upon them using bodies whose parameters change continuously through time.

It has been mathematically proven that match learning within an ART model leads to stable memories in response to arbitrary lists of events to be learned (Carpenter and Grossberg, 1991). However, match learning has a serious potential weakness: if you can only learn when there is a good enough match between bottom-up data and learned top-down expectations, then how do you ever learn anything that you do not already know? ART proposes that this problem is solved by the brain by using a complementary interaction between processes of resonance and reset, which are proposed to control properties of attention and memory search, respectively. These complementary processes help our brains to balance the complementary demands of processing the familiar and the unfamiliar, the expected and the unexpected. One of these complementary processes is hypothesized to take place in the What cortical stream, notably in the visual, inferotemporal, and prefrontal cortex. It is here that top-down expectations are matched against bottom-up inputs (Chelazzi *et al*., 1998; Miller *et al*., 1996). When a top-down expectation achieves a good enough match with bottom-up data, this matching process focuses attention upon those feature clusters in the bottom-up input that are expected. If the expectation is close enough to the input pattern, then a state of resonance develops as the attentional focus is established.

Figure 1 illustrates these ideas in a simple two-level example. Here, a bottom-up input pattern, or vector, I activates a pattern X of activity across the feature detectors of the first level F_1. For example, a visual scene may be represented by the features comprising its boundary and surface representations. This feature pattern represents the relative importance of different features in I. In Figure 1(a), the pattern peaks represent more activated feature detector cells, the troughs less activated feature detectors. This feature pattern sends signals S through an adaptive filter to the second level F_2 at which a compressed representation Y (a recognition category, or symbol) is activated in response to the distributed input T. T is computed by multiplying the signal vector S by a matrix of adaptive weights, which can be altered through learning. The representation Y is compressed by competitive interactions across F_2 that allow only a small subset of its most strongly activated cells to remain active in response to T. The pattern Y in the figure indicates that a small number of category cells may be activated to different degrees. These category cells, in turn, send top-down signals U to F_1 (see Figure 1(b)). The vector U is converted into the top-down expectation V by being multiplied by another matrix of adaptive weights. When V is received by F_1, a matching process takes place between I and V, which selects that subset X^* of F_1 features that were 'expected' by the active F_2 category Y. The set of these selected features is the emerging 'attentional focus'.

RECONCILING DISTRIBUTED AND SYMBOLIC REPRESENTATIONS USING RESONANCE

If the top-down expectation is close enough to the bottom-up input pattern, then the pattern X^* of attended features reactivates the category Y which, in turn, reactivates X^*. The network thus enters a resonant state through a positive feedback loop that dynamically links, or binds, the attended features across X^* with their category, or symbol, Y.

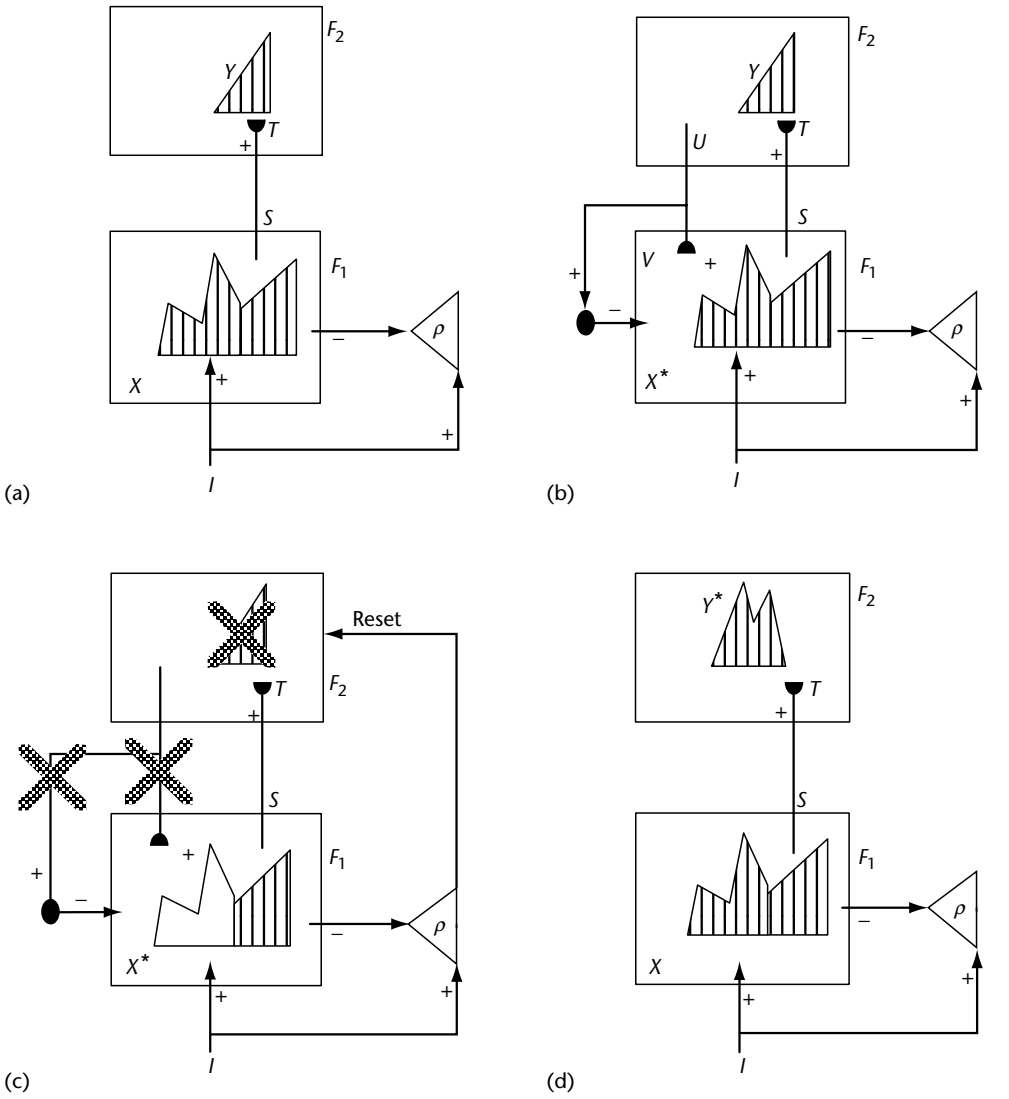


Figure 1. Search for a recognition code within an ART learning circuit. (a) The input pattern I is instated across the feature detectors at level F_1 as a short-term memory (STM) activity pattern X. I also nonspecifically activates the orienting system ρ; that is, all the input pathways converge on ρ and can activate it. X is represented by the hatched pattern across F_1. X both inhibits ρ and generates the output pattern S. S is multiplied by learned adaptive weights, which are long-term memory (LTM) traces. These LTM-gated signals are added at F_2 cells, or nodes, to form the input pattern T, which activates the STM pattern Y across the recognition categories coded at level F_2. (b) Pattern Y generates the top-down output pattern U, which is multiplied by top-down LTM traces and added at F_1 nodes to form a 'prototype' pattern V that encodes the learned expectation of the active F_2 nodes. Such a prototype represents the set of features shared by all the input patterns capable of activating Y. If V mismatches I at F_1, then a new STM activity pattern X^* is selected at F_1. X^* is represented by the hatched pattern. It consists of the features of I that are confirmed by V. Mismatched features are inhibited. The inactivated nodes, corresponding to unconfirmed features of X, are unhatched. The reduction in total STM activity which occurs when X is transformed into X^* causes a decrease in the total inhibition of ρ from F_1. (c) If inhibition decreases sufficiently, ρ releases a nonspecific arousal wave to F_2; that is, a wave of activation that activates all F_2 nodes equally. ('Novel events are arousing'.) This arousal wave resets the STM pattern Y at F_2 by inhibiting Y. (d) After Y is inhibited, its top-down prototype signal is eliminated, and X can be reinstated at F_1. The prior reset event maintains inhibition of Y during the search cycle. As a result, X can activate a different STM pattern Y^* at F_2. If the top-down prototype due to Y^* also mismatches I at F_1, then the search for an appropriate F_2 code continues, until an appropriate F_2 representation is selected. Such a search cycle represents a type of nonstationary hypothesis testing. When the search ends, an attentive resonance develops and learning of the attended data is initiated. (Adapted with permission from Grossberg, 1999b.)

The individual features at F_1 have no meaning on their own, just as the pixels in a picture are meaningless individually. The category, or symbol, in F_2 is sensitive to the global patterning of these features, but it cannot represent the 'contents' of the experience, including their conscious qualia, because a category is a compressed, or 'symbolic', representation. It has often been erroneously claimed that a system must process either distributed features or symbolic representations, but cannot process both. This is not true in ART. The resonance between these two types of information converts the pattern of attended features into a coherent context-sensitive state that is linked to its category through feedback. It is this coherent state, which joins together distributed features and symbolic categories, that can enter consciousness. ART proposes that all conscious states are resonant states. In particular, such a resonance binds spatially distributed features into either a synchronous equilibrium or an oscillation, until it is dynamically reset. Such synchronous states have recently attracted much interest after being reported in neurophysiological experiments. They were predicted in the 1970s in the articles that introduced ART (Grossberg, 1999b).

RESONANCE AS A MEDIATOR BETWEEN INFORMATION PROCESSING AND LEARNING

In ART, the resonant state, rather than bottom-up activation, is claimed to drive the learning process. The resonant state persists for long enough, and at a high enough activity level, to activate the slower learning processes in the adaptive weights that guide the flow of signals between bottom-up and top-down pathways between levels F_1 and F_2. This helps to explain how adaptive weights that were changed through previous learning can regulate the brain's present information processing, without learning about the signals that they are currently processing unless they can initiate a resonant state. Through resonance as a mediating event, one can see from a deeper viewpoint why humans are intentional beings who are continually predicting what may next occur, and why we tend to learn about the events to which we pay attention.

LEARNING AND HYPOTHESIS TESTING

A sufficiently strong mismatch between an active top-down expectation and a bottom-up input – for example, because the input represents an unfamiliar type of experience – can drive a memory search. Such a mismatch within the attentional system may activate a complementary 'orienting system', which is sensitive to unexpected and unfamiliar events. ART suggests that this orienting system includes the hippocampal system, which has long been known to be involved in mismatch processing, including the processing of novel events (e.g., Otto and Eichenbaum, 1992). Output signals from the orienting system rapidly reset the recognition category that has been specifying the poorly matching top-down expectation (Figure 1(b) and 1(c)). The cause of the mismatch is thus removed, thereby freeing the system to activate a different recognition category (Figure 1(d)). The reset event triggers memory search, or hypothesis testing, which automatically leads to the selection of a recognition category that can better match the input. If no such recognition category exists, say because the bottom-up input represents a truly novel experience, then the search process automatically activates an as-yet-uncommitted population of cells, with which to learn about the novel information.

This learning process works well under both unsupervised and supervised conditions (e.g., Carpenter and Grossberg, 1994). Under supervised conditions, a predictive error can force a cycle of hypothesis testing, or memory search, that might not have occurred under unsupervised conditions. For example, a misclassification of a letter F as an E could persist based just on visual similarity, unless culturally determined feedback forced the network to separate them into different categories. Such a search can discover a new or better-matching category with which to represent the novel data. Taken together, the interacting processes of attentive learning and orienting search achieve a type of error correction through hypothesis testing that can build an ever-growing, self-refining internal model of a changing world.

CONTROLLING THE GENERALITY OF KNOWLEDGE

What information is bound into object or event representations? Some scientists believe that exemplars, or individual experiences, can be learned and remembered, like familiar faces. But storing every exemplar requires huge amounts of memory, and leads to unwieldy memory retrieval. Others believe that we learn prototypes (Posner and Keele, 1970) that represent more general properties of the environment, for example, that everyone has a face. But then how do we learn specific episodic memories? ART provides an answer to this question.

ART systems learn prototypes whose generality is determined by a process of 'vigilance' control by environmental feedback or internal volition (Carpenter and Grossberg, 1991; Grossberg, 1999b). Low vigilance permits learning of general categories with abstract prototypes. High vigilance forces memory search to occur when even small mismatches exist between an exemplar and the category that it activates: for example, between letter exemplar F and letter category E. Given high enough vigilance, a category prototype may encode an individual exemplar. Vigilance is computed within the ART orienting system: see Figure 1. Here, bottom-up excitation from an input pattern I is balanced against inhibition from active features across level F_1. If a top-down expectation acts on F_1, then only the 'matched' features are active there. If the ratio of matched features in F_1 to all features in I is less than a vigilance parameter ρ (Figure 1(b)), then a reset, or 'novelty,' wave is activated (Figure 1(c)), which can trigger a search for another category.

The simplest rule for controlling vigilance during supervised learning is called match tracking. Here, a predictive error causes vigilance to increase until it is just higher than the ratio of active features in F_1 to total features in I. The error hereby forces vigilance to 'track' the degree of match between input exemplar and matched prototype. This is the minimal level of vigilance that can trigger a reset wave and thus a memory search for a new category. Match tracking realizes a minimax learning rule that maximizes category generality while minimizing predictive error. That is, it uses the least amount of memory resources that can prevent errors. ART models thus try to learn the most general category that is consistent with the data. This can lead to overgeneralization, like that seen in young children, until further learning causes category refinement. Benchmark studies of classifying complex databases have shown that the number of categories learned scales well with data complexity (e.g. Carpenter and Grossberg, 1994).

MEMORY CONSOLIDATION AND THE EMERGENCE OF RULES

As sequences of inputs are practiced over learning trials, the search process eventually converges upon stable categories. It has been mathematically proven (Carpenter and Grossberg, 1991) that familiar inputs directly access the category whose prototype provides the globally best match, while unfamiliar inputs engage the orienting subsystem to trigger memory searches for better categories, until they become familiar. This process continues until the available memory, which can be arbitrarily large, is fully utilized. The process whereby search is automatically disengaged is a form of memory consolidation that emerges from network interactions. Emergent consolidation does not preclude structural consolidation at individual cells, since the amplified and prolonged activities that subserve a resonance may be a trigger for learning-dependent cellular processes, such as protein synthesis and transmitter production. It has also been shown that the adaptive weights which are learned by some ART models can, at any stage of learning, be translated into if–then rules (e.g. Carpenter and Grossberg, 1994). Thus the ART model is a self-organizing rule-discovering production system as well as a neural network. These examples show that the claims of some cognitive scientists and AI practitioners that neural network models cannot learn rule-based behaviors are incorrect.

CORTICOHIPPOCAMPAL INTERACTIONS AND MEDIAL TEMPORAL AMNESIA

As noted above, the attentional subsystem of ART has been used to model aspects of inferotemporal cortex, while the orienting subsystem models part of the hippocampal system. The interpretation of ART dynamics in terms of inferotemporal cortex led Miller *et al*. (1991) to successfully test the prediction that cells in monkey inferotemporal cortex are reset after each trial in a working memory task. To illustrate the implications of an ART interpretation of inferotemporal-hippocampal interactions, we will review how a lesion of the ART model's orienting subsystem creates a formal memory disorder with symptoms much like the medial temporal amnesia that is caused in animals and human patients after hippocampal system lesions. In particular, such a lesion *in vivo* causes: unlimited anterograde amnesia; limited retrograde amnesia; failure of consolidation; tendency to learn the first event in a series; abnormal reactions to novelty, including perseverative reactions; normal priming; and normal information processing of familiar events. Unlimited anterograde amnesia occurs because the network cannot carry out the memory search to learn a new recognition code. Limited retrograde amnesia occurs because familiar events can directly access correct recognition codes. Before events become familiar, memory consolidation occurs, which utilizes the orienting subsystem (Figure 1(c)). This failure of consolidation would

not necessarily prevent learning. Instead, it would learn coarser categories, because of the failure of vigilance control and memory search. For the same reason, learning may differentially influence the first recognition category activated by bottom-up processing, much as amnesics are particularly strongly bound to the first response they learn. Perseverative reactions can occur because the orienting subsystem cannot reset sensory representations or top-down expectations that may be persistently mismatched by bottom-up cues. The inability to search memory prevents ART from discovering more appropriate stimulus combinations to attend. Normal priming occurs because it is mediated by the attentional subsystem. Data supporting these predictions are summarized by Grossberg and Merrill (1996), who also note that these are not the only problems that can be caused by such a lesion: hippocampal structures can also play a role in learned spatial navigation and adaptive timing functions.

Knowlton and Squire (1993) have reported that amnesics can classify items as members of a large category even if they are impaired on remembering the individual items themselves. To account for these results, the authors propose that item and category memories are formed by distinct brain systems. Grossberg and Merrill (1996) suggest that their data could be explained by a single ART system in which the absence of vigilance control caused only coarse categories to form. Nosofsky and Zaki (2000) have quantitatively simulated the Knowlton and Squire data using a single-system model in which category sensitivity is low.

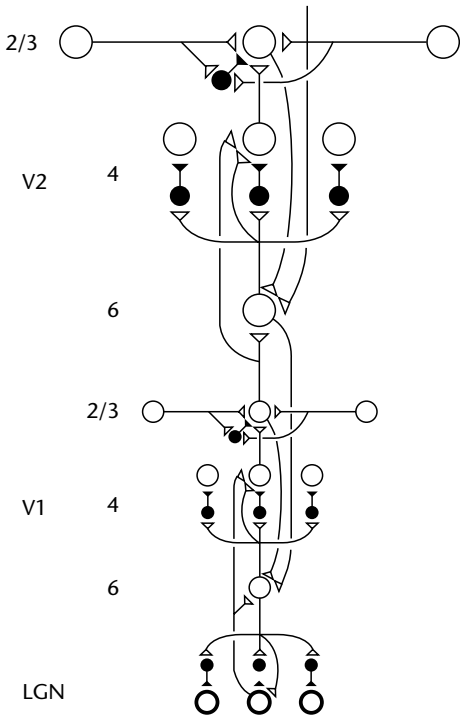


Figure 2. The LAMINART model. The model is a synthesis of feedforward (bottom-up), feedback (top-down), and horizontal interactions within and between the lateral geniculate nucleus (LGN) and visual cortical areas V1 and V2. Cells and connections with open symbols indicate excitatory interactions, and closed symbols indicate inhibitory interactions. The top-down connections from level 6 of V2 to level 6 of V1 indicate attentional feedback. (See Grossberg, 1999a and Grossberg and Raizada, 2000 for further discussion of how these circuits work.) (Adapted with permission from Grossberg and Raizada, 2000.)

CORTICAL SUBSTRATES OF ART MATCHING

How are ART top-down matching rules implemented in the cerebral cortex of the brain? An answer to this question has been proposed as part of a rapidly developing theory of why the cerebral cortex is typically organized into six distinct layers of cells (Grossberg, 1999a). Earlier mathematical work had predicted that such a matching rule would be realized by a 'modulatory top-down on-center off-surround' network (e.g. Carpenter and Grossberg, 1991; Grossberg, 1999b). Figure 2 shows how such a matching circuit may be realized in the cortex. The top-down circuit generates outputs from cortical layer 6 of V2 that activate layer 6 of V1 via the vertical pathway between these layers that ends in an open triangle (which indicates an excitatory connection). Cells in layer 6 of V1, in turn, activate an 'on-center off-surround' circuit to layer 4 of V1. In this circuit, an excitatory cell (open circle) in layer 6 excites the excitatory cell immediately above it in layer 4 via the vertical pathway from layer 6 to layer 4 that ends in an open triangle. This excitatory interaction constitutes the 'on-center'. The same excitatory cell in layer 6 also excites nearby inhibitory cells (closed black circles) which, in turn, inhibit cells in layer 4. This spatially distributed inhibition constitutes the 'off-surround' of the layer 6 cell. The on-center is predicted to have a modulatory, or sensitizing, effect on layer 4, due to the balancing of excitatory and inhibitory inputs to layer 4 within the on-center. The inhibitory signals in the off-surround can strongly suppress unattended visual features. This arrangement shows how top-down attention can sensitize the brain to prepare for expected information that may or may not actually occur, without actively firing the sensitized target cells and thereby inadvertently creating

hallucinations that the information is already there. When this balance breaks down, model 'hallucinations' may indeed occur, and these have many of the properties reported by schizophrenic patients.

CONCLUSION

Adaptive resonance theory is a neural and a cognitive theory of human and animal information processing. ART proposes how the processes whereby the brain can stably develop in the infant and learn throughout life constrain the form of perceptual and cognitive processes such as categorization, expectation, attention, synchronization, memory search, and consciousness in both normal and clinical patients. ART realizes a mechanistic unification of concepts about exemplar, prototype, distributed, symbolic, and rule-based processing. Recent models have shown how predicted ART matching properties may be realized in certain laminar circuits of visual cortex, and by extension in other sensory and cognitive neocortical areas.

Acknowledgments

Supported in part by the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-95-1-0409), the National Science Foundation (NSF IRI-97-20333), and the Office of Naval Research (ONR N00014-95-1-0657).

References

Carpenter GA and Grossberg S (1991) *Pattern Recognition by Self-Organizing Neural Networks*. Cambridge, MA: MIT Press.

Carpenter GA and Grossberg S (1994) Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction. In: Honavar V and Uhr L (eds) *Artificial Intelligence and Neural Networks: Steps Towards Principled Prediction*, pp. 387–421. San Diego, CA: Academic Press.

Chelazzi L, Duncan J, Miller EK and Desimone R (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology* **80**: 2918–2940.

Grossberg S (1999a) How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision* **12**: 163–186.

Grossberg S (1999b) The link between brain learning, attention, and consciousness. *Consciousness and Cognition* **8**: 1–44.

Grossberg S and Merrill JWL (1996) The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience* **8**: 257–277.

Grossberg S and Raizada RDS (2000) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Reseach* **40**: 1413–1432.

Knowlton BJ and Squire LR (1993) The learning of categories: parallel brain systems for item memory and category knowledge. *Science* **262**: 1747–1749.

Miller EK, Erickson CA and Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience* **16**: 5154–5167.

Miller EK, Li L and Desimone R (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* **254**: 1377–1379.

Nosofsky RM and Zaki SR (2000) Category learning and amnesia: an exemplar model perspective. In: *Proceedings of the 2000 Memory Disorders Research Society Annual Meeting*. Toronto.

Posner MI and Keele SW (1970) Retention of abstract ideas. *Journal of Experimental Psychology* **83**: 304–308.

Otto T and Eichenbaum H (1992) Neuronal activity in the hippocampus during delayed non-match to sample performance in rats: evidence for hippocampal processing in recognition memory. *Hippocampus* **2**: 323–334.

Further Reading

Clark EV (1973) What's in a word? On the child's acquisition of semantics in his first language. In: Morre TE (ed.) *Cognitive Development and the Acquisition of Language*, pp. 65–110. New York, NY: Academic Press.

Goodale MA and Milner D (1992) Separate visual pathways for perception and action. *Trends in Neurosciences* **15**: 20–25.

Grossberg S (2000) How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society* **6**: 583–592.

Lynch G, McGaugh JL and Weinberger NM (eds) (1984) *Neurobiology of Learning and Memory*. New York, NY: Guilford Press.

Mishkin M, Ungerleider LG and Macko KA (1983) Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences* **6**: 414–417.

Sokolov EN (1968) *Mechanisms of Memory*. Moscow: Moscow University Press.

Squire LR and Butters N (eds) (1984) *Neuropsychology of Memory*. New York, NY: Guilford Press.

Vinogradova OS (1975) Functional organization of the limbic system in the process of registration of information: facts and hypotheses. In: Isaacson RL and Pribram KH (eds) *The Hippocampus*, vol. II. pp. 3–69. New York, NY: Plenum Press.

Analogy-making, Computational Models of

Boicho Kokinov, New Bulgarian University, Sofia, Bulgaria
Robert M French, University of Liège, Liège, Belgium

CONTENTS

Analogy-making is the process of finding or constructing a common relational structure in the descriptions of two situations or domains and making inferences by transferring knowledge from the familiar domain (the 'base' or 'source') to the unfamiliar domain (the 'target'), thus enriching our knowledge about the latter.

INTRODUCTION

Analogy-making is crucial for human cognition. Many cognitive processes involve analogy-making in one way or another: perceiving a stone as a human face, solving a problem in a way similar to another problem previously solved, arguing in court for a case based on its common structure with another case, understanding metaphors, communicating emotions, learning, or translating poetry from one language to another (Gentner *et al*, 2001). All these applications require an abstract mapping to be established between two cases or domains based on their common structure (common systems of relations). This may require re-representation of one (or both) of the domains in terms of the other one (or in terms of a third domain). The first domain is called the base, or source, and the second is called the target.

Analogy-making is a basic cognitive ability. It appears to be present in humans from a very early age, and develops over time. It starts with the simple ability of babies to imitate adults and to recognize when adults are imitating them, progresses to children's being able to recognize an analogy between a picture and the corresponding real object, and culminates in the adult ability to make complex analogies between various situations. This seems to suggest that analogy-making serves as the basis for numerous other kinds of human thinking; hence the importance of developing computational models of analogy-making.

Analogy-making involves at least the following sub-processes: building a representation, retrieving a 'base' for the analogy, mapping this base to the 'target', transferring unmapped elements from the base to the target, thereby making inferences, evaluating the validity and applicability of these inferences, and learning from the experience – which includes generalizing from specific cases and, possibly, developing general mental schemata. There are, at present, no models that incorporate all of these sub-processes. Individual models focus on one or more of them.

Representation-building

The process of representation-building is absent from most models of analogy-making. Typically, representations are fed into the model. However, there are some models (e.g., ANALOGY, Copycat, Tabletop, Metacat) that do produce their own high-level representations based on essentially unprocessed input. These models (Mitchell, 1993; Hofstadter *et al*, 1995; French, 1995) attempt to build flexible, context-sensitive representations during the course of the mapping phase. Other models, such as AMBR (Kokinov and Petrov, 2001), perform re-representation of old episodes.

Retrieval

The retrieval process has been extensively studied experimentally. Superficial similarity is the most important factor in analogical retrieval: the retrieval of a base for analogy is easier if it shares similar objects, similar properties, and similar general themes with the target. Structural similarity, the familiarity of the domain from which the

analogy is drawn, the richness of its representations, and the presence of generalized schemata, also facilitate retrieval. Most models of retrieval are based on exhaustive search of long-term memory (LTM) and on the assumption that old episodes have context-independent, encapsulated representations. There are, however, exceptions (e.g., AMBR) that rely on context-sensitive reconstruction of old episodes performed in interaction with the mapping process.

Mapping

Mapping is the core of analogy-making. All computational models of analogy-making include mapping mechanisms, i.e., means of discovering which elements of the base correspond to which elements of the target. The difficulty is that one situation can be mapped to a second situation in many different ways. We might, for example, make a mapping based on the color of the objects in both the base and target (the red-shirted person in the base domain would be mapped to the red-shirted person in the target domain). This would, in general, be a very superficial mapping (but might, none the less, be appropriate on occasion). We could also map the objects in the two domains based on their relational structures. For example, we could decide that it was important to map the giver–receiver relationship in the first domain to the same relationship in the target domain, ignoring the fact that in the base domain the giver had a red shirt and in the target domain the giver was wearing a blue shirt.

Experimental work has demonstrated that finding this type of structural isomorphism between base and target domains is crucial for mapping (Gentner, 1983). Object similarity also plays a role in mapping, although generally a secondary one. A third factor is the pragmatic importance of various elements in the target: we want to find mappings that involve the most important elements in the target. Searching for the appropriate correspondences between the base and the target is a computationally complex task that can become infeasibly time-consuming if the search is unconstrained.

Transfer

New knowledge then has to be inserted into the target domain based on the mapping. For example, suppose a new brand of car appears on the market, and that this car maps well onto another brand of car that is small, fast, and handles well on tight curves. But you also know that this latter brand of car is frequently in need of repair. You then wonder whether the new brand of car will also be in frequent need of repair.

Transfer is present in some form in most models of analogy-making, and is typically integrated with mapping. Transfer is considered by some researchers as an extension of the mapping already established, adding new elements to the target.

Evaluation

Evaluation is the process of establishing the likelihood that the transferred knowledge will be applicable to the target domain. In the example above, the evaluation process would have to assign the degree of confidence we would have that the new car would also be in frequent need of repair. Evaluation is often implicit in the mechanisms of mapping and transfer.

Learning

Only a few models of analogy-making have incorporated learning mechanisms. This is somewhat surprising since analogy-making is clearly a driving force behind much learning. However, some models are capable of generalization from the base and target, or from multiple exemplars, to form an abstract schema, as in LISA (Hummel and Holyoak, 1997) and the SEQL model based on SME (Falkenhainer *et al*, 1989).

Below we will review a number of important computational models of analogy-making belonging to different classes and following different approaches. First the 'symbolic' models will be presented. These models employ separate local representations of objects, relations, propositions and episodes (e.g., 'John', 'chair', 'run', 'greater than', 'John ate fish', 'my birthday party last year'). Then, 'connectionist' models will be presented. Here the objects, relations, and episodes are represented as overlapping patterns of activation in a neural network. Finally, a third, hybrid class of models will be presented. These models combine symbolic representations with connectionist activations. They are based on the idea that cognition is an emergent property of the collective behavior of many simple agents.

SYMBOLIC MODELS

ANALOGY

The earliest computational model of analogy-making, ANALOGY, was developed by Thomas

Evans (1964). This program solves multiple-choice geometric analogy problems of the form '*A* is to *B* as *C* is to what?' taken from intelligence tests and college entrance examinations.

An important feature of this program is that the input is not a high-level description of the problem, but a low-level description of each component of the figure – dots, simple closed curves or polygons, and sets of closed curves or polygons. The program builds its own high-level representation describing the figures in *A*, *B*, *C*, and all given alternatives for the answer, with their properties and relationships – for example `((P1 P2 P3) . ((INSIDE P1 P2) (LEFT P1 P3) (LEFT P2 P3)))`. Then the program represents the relationship between *A* and *B* as a set of possible rules describing how figure *A* is transformed into figure *B* – for example, `((MATCH P2 P4) (MATCH P1 P5) (REMOVE P3))` which means that figure P_2 from *A* corresponds to figure P_4 from *B*, P_1 corresponds to P_5, and the figure P_3 does not have a corresponding figure and is therefore deleted. Then each such rule is applied to *C* in order to get one of the alternative answers. In fact, each such rule would be generalized in such a way as to allow *C* to be applied to *D*. Finally, the most specific successful rule would be selected as an outcome. Arguably, one of the most significant features of the program is its ability to represent the target problem on its own – a feature that has been dropped in most recent models.

Structure Mapping Theory

The most influential family of computational models of analogy-making have been those based on Dedre Gentner's (1983) 'structure mapping theory' (SMT). This theory was the first to explicitly emphasize the importance of structural similarity between base and target domains, defined by common systems of relations between objects in the respective domains. Numerous psychological experiments have confirmed the crucial role of relational mappings in producing convincing and sound analogies. There are several important assumptions underlying the computational implementation of SMT called SME (Falkenhainer *et al*, 1989): (1) mapping is largely isolated from other analogy-making sub-processes (such as representation, retrieval and evaluation) and is based on independent mechanisms; (2) relational matches are preferred over property matches; (3) only relations that are identical in the two domains can be put into correspondence; (4) relations that are arguments of higher-order relations that can also be mapped have priority, following the 'systematicity principle' that favors systems of relations over isolated relations; and (5) two or three interpretations are constructed by a 'greedy merge' algorithm that generally finds the 'best' structurally coherent mapping. Early versions of SME mapped only identical relations and relied solely on relational structure. This purely structural approach was intended to ensure the domain independence of the mapping process. Recent versions of SME have made some limited use of pragmatic aspects of the situation, as well as re-representation techniques that allow initially non-matching predicates to match.

The MAC/FAC model of analogical retrieval (Forbus *et al*, 1995) was intended to be coupled with SME. This model assumes that episodes are encapsulated representations of past events; they have a dual encoding in LTM: a detailed predicate-calculus representation of all the properties and relations of the objects in an episode and a shorter summary (a vector representation indicating the relative frequencies of the predicates that are used in the detailed representation). The retrieval process has two stages. The first stage uses the vector representations to perform a superficial search for episodes that share predicates with the target problem. The episode vectors in LTM that are close to the target vector are selected for processing by the second stage. The second stage uses the detailed predicate-calculus representations of the episodes to select the one that best matches the target. These two stages reflect the dominance of superficial similarity as well as the influence of structural similarity.

Gentner's ideas – in particular, their emphasis on the structural aspects of analogical mappings – have been very influential in the area of computational analogy-making and have been applied in contexts ranging from child development to folk physics. Various improvements and variants of SME have been developed, and it has been included as a module in various practical applications.

Other Symbolic Models

A number of other symbolic models have helped to advance our understanding of analogy-making. Jaime Carbonell proposed the concept of derivational analogy, where the analogy is drawn not with the final solution of the old problem, but with its derivation, i.e., with the way of reaching the solution, an approach developed further by Manuela Veloso. Smadar Kedar-Cabelli developed a model of purpose-directed analogy-making in

concept learning. Mark Burstein developed a model called CARL which learned from multiple analogies combining several bases. Mark Keane and his colleagues developed an incremental model of mapping, IAM, which helps explain the effects of order of presentation of the material observed in humans.

CONNECTIONIST MODELS

Research in the field of analogy-making has, until recently, been largely dominated by the symbolic approach, for an obvious reason: symbolic models are well equipped to process and compare the complex structures required for analogy-making. In the early years of the new connectionist paradigm, these structures were very difficult to represent in a connectionist network. However, advances in connectionist representation techniques have allowed distributed connectionist models of analogy to be developed. Most importantly, distributed representations provide a natural internal measure of similarity, thereby allowing the system to handle the problem of similar but not identical relations in a relatively straightforward manner. This ability is essential to analogy-making and has proved hard for symbolic models to implement.

Multiple Constraints Theory

The earliest attempt to design an architecture in which analogy-making was an emergent process of activation states of neuron-like objects was proposed by Keith Holyoak and Paul Thagard (1989) and implemented in a model called ACME. In this model, structural similarity, semantic similarity, and pragmatic importance determine a set of constraints to be simultaneously satisfied. The model is supplied with representations of the target and of the base, and proceeds to build a localist constraint-satisfaction connectionist network where each node corresponds to a possible pairing hypothesis for an element of the base and an element of the target. For example, if the base is 'train' and the target is 'car' then all elements of trains will be mapped to all elements of cars; there will therefore be hypothesis nodes created not only for 'locomotive → motor' but also for 'locomotive → license plate', 'locomotive → seat-belt buckle', etc. The excitatory and inhibitory links between these nodes implement the structural constraints. In this way, contradictory hypothesis nodes compete and do not become simultaneously active, while consistent ones mutually support each other. The network gradually moves towards an equilibrium state, and the best set of consistent mapping hypotheses (e.g., 'locomotive → motor', 'rails → road', etc.) wins. The relaxation of the network provides a parallel evaluation of all possible mappings and finds the best one, which is represented by the set of most active hypothesis nodes.

ARCS is another related model of retrieval. It is coupled with ACME and operates in a similar fashion. However, while mapping is dominated by structural similarity, retrieval is dominated by semantic similarity.

STAR

STAR-1 was the first distributed connectionist model of analogy-making (Halford *et al*., 1994). It is based on the tensor product connectionist models developed by Smolensky. A proposition like MOTHER-OF (CAT, KITTEN) is represented by the tensor product of the three vectors corresponding to MOTHER-OF, CAT, and KITTEN: MOTHER-OF $\otimes$ CAT $\otimes$ KITTEN. The tensor product in this case is a three-dimensional array of numbers where the number in each cell is the product of the three corresponding coordinates. This representation allows any of the arguments, or the relational symbol, to be extracted by a generalized dot product operation: (MOTHER-OF $\otimes$ CAT) $\bullet$ (MOTHER-OF $\otimes$ CAT $\otimes$ KITTEN) = KITTEN. The LTM of the system is represented by a tensor that is the sum of all tensor products representing the individual statements (the main restriction being that the propositions are simple and have the same number of arguments). Using this type of representation, STAR-1 solves proportional analogy problems like 'cat is to kitten as mare is to what?'

STAR-2 (Wilson *et al*, 2001) maps complex analogies by sequentially focusing on various parts of the domains – simple propositions with no more than four dimensions – and finding the best map for the arguments of these propositions by parallel processing in the constraint satisfaction network (similarly to ACME). The fact that the number of units required for a tensor product representation increases exponentially with the number of arguments of a predicate implies processing constraints in the model. Wilson *et al* claim that humans are subject to similar processing constraints: specifically, they can, in general, handle a maximum of four dimensions of a situation concurrently. The primary interest of the modelers is in exploring and explaining capacity limitations of human beings and achieving a better understanding of the development of analogy-making capabilities in children.

LISA

John Hummel and Keith Holyoak (1997) proposed an alternative computational model of analogy-making using distributed representations of structure relying on dynamic binding. The idea is to introduce an explicit time axis so that patterns of activation can oscillate over time (thus the timing of activation becomes an additional parameter independent of the level of activation). Patterns of activation oscillating in synchrony are considered to be bound together, while those oscillating out of synchrony are not. For example, 'John hired Mary' requires synchronous oscillation of the patterns for 'John' and 'Employer' alternating with synchronous oscillation of the patterns for 'Mary' and 'Employee'. Periodic alternation of the activation of the two pairs represents the whole statement. However, if the statement is too complex there will be too many pairs that need to fire in synchrony. Based on research on single-cell recordings, Hummel and Holyoak believe that the number of such pairs of synchronously firing concepts cannot exceed six. Representations in LISA's working memory are distributed over the network of semantic primitives, but representations in long-term memory are localist – there are separate units representing the episode, the propositions, their components, and the predicates, arguments, and bindings. Retrieval is performed by spreading activation, while mapping is performed by learning new connections between the most active nodes. LISA successfully integrates retrieval of a base with the mapping of the base and target, even though retrieval and mapping are still performed sequentially (mapping starts only after one episode is retrieved).

HYBRID MODELS

Two groups of researchers have independently produced similar models of analogy-making based on the idea that high-level cognition emerges as a result of the continual interaction of relatively simple, low-level processing units, capable of doing only local computations. These models are a combination of the symbolic and connectionist approaches. Semantic knowledge is incorporated in order to compute the similarity between elements of the two domains in a context-sensitive way.

Copycat and Related Architectures

The family of Copycat and Tabletop architectures (Mitchell, 1993; Hofstadter *et al*, 1995; French, 1995) was explicitly designed to integrate top-down semantic information with bottom-up emergent processing. Copycat solves letter-string analogies of the form '*ABC* is to *ABD* as *KLM* is to what?' and gives plausible answers such as *KLN* or *KLD*. The architecture of Copycat involves a working memory, a semantic network (simulating long-term memory) defining the concepts used in the system and their relationships, and the 'coderack' – the procedural memory of the system – a store for small, nondeterministic computational agents ('codelets') working on the structures in the working memory and continually interacting with the semantic network. Codelets can build new structures or destroy old structures in working memory. The system gradually settles towards a consistent set of structures that will determine the mapping between the base and the target.

The most important feature of these models of analogy-making is their ability to build up their own representations of the problem, in contrast with most other models which receive the representations of the base and target as input. Thus these models abandon traditional sequential processing and allow representation-building and mapping to run in parallel and continually influence each other. The partial mapping can influence further representation-building, thus allowing the gradual construction by the program of context-sensitive representations. In this way, the mapping may force us to see a situation from an unusual perspective in terms of another situation, and this is an essential aspect of creative analogy-making.

AMBR

AMBR (Kokinov, 1994) solves problems by analogy. For example, 'how can you heat some water in a wooden vessel, being in the forest?' The solution, heating a knife in a fire and immersing it in the water, is found by analogy with boiling water in a glass using an immersion heater.

The AMBR model is based on DUAL, a general cognitive architecture. The LTM of DUAL consist of many micro-agents, each of which represents a small piece of knowledge. Thus concepts, instances and episodes are represented by (possibly overlapping) coalitions of micro-agents. Each micro-agent is hybrid: its symbolic part encodes the declarative or procedural knowledge it is representing, while its connectionist part computes the agent's activation level, which represents the relevance of this knowledge to the current context. The symbolic processors run at a speed proportional to their computed relevance, so the behavior of the system

is highly context-sensitive. The AMBR model implements the interactive parallel work of recollection, mapping and transfer that emerge from the collective behavior of the agents and which produces the analogy. Recollection in AMBR-2 (Kokinov and Petrov, 2001) is reconstruction of the base episode in working memory by activating relevant aspects of event information, of general knowledge, and of other episodes, and forming a coherent representation which will correspond to the target problem. The model exhibits illusory memories, including insertions from general knowledge and blending with other episodes, and context and priming effects. Some of these phenomena have been experimentally confirmed in humans.

CONCLUSIONS

The field of computational modeling of analogy-making has moved from the early models, which were intended mainly to demonstrate that computers could, in fact, be programmed to do analogy-making, to complex models that make nontrivial predictions of human behavior. Researchers have come to appreciate the need for structural mapping of the base and target domains, for integration of and interaction between representation-building, retrieval, mapping and learning, and for systems that can potentially scale up to the real world. Computational models of analogy-making have now been applied to a large number of cognitive domains (Gentner *et al*, 2001). However, many issues remain to be explored in the endeavor to model the human capacity for analogy-making, one of our most important cognitive abilities.

References

Evans TG (1964) A heuristic program to solve geometric-analogy problems. In: *Proceedings of the Spring Joint Computer Conference* **25**: 327–338 [Reprinted in: Fischler M and Firschein O (eds) (1987) *Readings in Computer Vision*. Los Altos, CA: Morgan Kaufmann.]

Falkenhainer B, Forbus KD and Gentner D (1989) The structure-mapping engine: algorithm and examples. *Artificial Intelligence* **41**: 1–63.

Forbus K, Gentner D and Law K (1995) MAC/FAC: a model of similarity-based retrieval. *Cognitive Science* **19**(2): 141–205.

French R (1995) *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making*. Cambridge, MA: MIT Press.

Gentner D (1983) Structure-mapping: a theoretical framework for analogy. *Cognitive Science* **7**(2): 155–170.

Gentner D, Holyoak K and Kokinov B (eds) (2001) *The Analogical Mind: Perspectives From Cognitive Science*. Cambridge, MA: MIT Press.

Halford G, Wilson W, Guo J *et al* (1994) Connectionist implications for processing capacity limitations in analogies. In: Holyoak K and Barnden J (eds) *Advances in Connectionist and Neural Computation Theory*, vol. II 'Analogical Connections', pp. 363–415. Norwood, NJ: Ablex.

Hofstadter D and the Fluid Analogies Research Group (1995) *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York, NY: Basic Books.

Holyoak K and Thagard P (1989) Analogical mapping by constraint satisfaction. *Cognitive Science* **13**: 295–355.

Hummel J and Holyoak K (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* **104**: 427–466.

Kokinov B (1994) A hybrid model of analogical reasoning. In: Holyoak K and Barnden J (eds) *Advances in Connectionist and Neural Computation Theory*, vol. II 'Analogical Connections', pp. 247–318. Norwood, NJ: Ablex.

Kokinov B and Petrov A (2001) Integration of memory and reasoning in analogy-making: the AMBR model. In: Gentner *et al* (2001), pp. 59–124.

Mitchell M (1993) *Analogy-Making as Perception: A Computer Model*. Cambridge, MA: MIT Press.

Wilson W, Halford G, Gray B and Phillips S (2001) The STAR-2 model for mapping hierarchically structured analogs. In: Gentner *et al* (2001), pp. 125–159.

Further Reading

Barnden J and Holyoak K (eds) (1994) *Advances in Connectionist and Neural Computation Theory*, vol. III 'Analogy, Metaphor, and Reminding'. Norwood, NJ: Ablex.

Gentner D, Holyoak K and Kokinov B (eds) (2001) *The Analogical Mind: Perspectives From Cognitive Science*. Cambridge, MA: MIT Press.

Hall R (1989) Computational approaches to analogical reasoning: a comparative analysis. *Artificial Intelligence* **39**: 39–120.

Holyoak K and Barnden J (eds) (1994) *Advances in Connectionist and Neural Computation Theory*, vol. II 'Analogical Connections'. Norwood, NJ: Ablex.

Holyoak K, Gentner D and Kokinov B (eds) (1998) *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*. Sofia: New Bulgarian University Press.

Holyoak K and Thagard P (1995) *Mental Leaps*. Cambridge, MA: MIT Press.

Vosniadou S and Ortony A (eds) (1989) *Similarity and Analogical Reasoning*. New York, NY: Cambridge University Press.

Artificial Life

Intermediate article

Norman H Packard, Prediction Company, Santa Fe, New Mexico, USA
Mark A Bedau, Reed College, Portland, Oregon, USA

CONTENTS

Artificial life is the study of life and life-like processes through simulation and synthesis.

INTRODUCTION

Artificial life literally means 'life made by human artifice rather than by nature'. It has come to refer to a broad, interdisciplinary endeavor that uses the simulation and synthesis of life-like processes to achieve any of several possible ends: to model life, to develop applications using intuitions and methods taken from life, or even to create life. The aim of creating life in a purely technological context is sometimes called 'strong artificial life'.

Artificial life is of interest to biologists because artificial life models can shed light on biological phenomena. It is relevant to engineers because it offers methods to generate and control complex behaviors that are difficult to generate or control using traditional approaches. But artificial life also has many other facets involving *inter alia* various aspects of cognitive science, economics, art, and even ethics.

There is not a consensus, even among workers in the field, on exactly what artificial life is, and many of its central concepts and working hypotheses are controversial. As a consequence, the field itself is evolving from year to year. This article provides a snapshot and highlights some controversies.

HISTORY

The roots of artificial life are quite varied, and many of its central concepts arose in earlier intellectual movements.

John von Neumann implemented the first artificial life model (without referring to it as such) with his famous creation of a self-reproducing, computation-universal entity using cellular automata. At the time, the construction was surprising, since many had argued its impossibility, for example on the grounds that such an entity would need to contain a description of itself, and that description would also need to contain a description, *ad infinitum*. Von Neumann was pursuing many of the very issues that drive artificial life today, such as understanding the spontaneous generation and evolution of complex adaptive structures; and he approached these issues with the extremely abstract methodology that typifies contemporary artificial life. Even in the absence of modern computational tools, von Neumann made striking progress.

Cybernetics developed at about the same time as von Neumann's work on cellular automata, and he attended some of its formative meetings. Norbert Wiener is usually considered to be the originator of the field (Wiener, 1948). It brought two separate foci to the study of life processes: the use of information theory and a deep study of the self-regulatory processes (homeostases) considered essential to life. Information theory typifies the abstractness and material-independence of the approach often taken within both cybernetics and artificial life. Both fields are associated with an extremely wide range of studies, from mathematics to art. As a discipline, cybernetics has evolved in divergent directions; in Europe, academic departments of cybernetics study rather specific biological phenomena, whereas in America cybernetics has tended to merge into systems theory, which generally aims toward formal mathematical studies. Scientists from both cybernetics and systems theory contribute substantially to contemporary artificial life.

Biology (i.e. the study of actual life) has provided many of the roots of artificial life. The subfields of biology that have contributed most are microbiology and genetics, evolution theory, ecology, and development. To date there are two main ways that artificial life has drawn on biology: crystalizing

intuitions about life from the study of life, and using and developing models that were originally devised to study a specific biological phenomenon. A notable example of the latter is Kauffman's use of random Boolean networks (Kauffman, 1969, 1993). Biology has also influenced the problems studied in artificial life, since artificial life's models provide definite answers to problems that are intractable by the traditional methods of mathematical biology. Mainstream biologists are increasingly participating in artificial life, and the methods and approaches pioneered in artificial life are increasingly accepted within biology.

The most heavily represented discipline among contemporary researchers in artificial life is computer science. One set of artificial life's roots in computer science is embedded in artificial intelligence (AI), because living systems exhibit simple but striking forms of intelligence. Like AI, artificial life aims to understand a natural phenomenon through computational models. But in sharp contrast to AI, at least as it was originally formulated, artificial life tends to use bottom-up models in which desired behavior emerges in a number of computational stages, instead of top-down models that aim to yield the desired behavior directly (as with expert systems). In this respect, artificial life shares much with the connectionist movement that has recently swept through artificial intelligence and cognitive science. Artificial life has a related set of roots in machine learning, inspired by the robust and flexible processes by which living systems generate complex useful structures. In particular, some machine learning algorithms such as the genetic algorithm (Holland, 1975) are now seen as examples of artificial life applications, even though they existed before the field was named. New areas of computer science (e.g., evolutionary programming, autonomous agents) have increasingly strong links to artificial life. (*See* **Artificial Intelligence, Philosophy of**)

Physics and mathematics have also had a strong influence on artificial life. Statistical mechanics and thermodynamics have always claimed relevance to life, since life's formation of structure is a local reversal of the second law of thermodynamics, made possible by the energy flowing through a living system. Prigogine's thermodynamics of dissipative structures is the most modern description of this view. Statistical mechanics is also used to analyze some of the models used in artificial life that are sufficiently simple and abstract, such as random Boolean networks. Dynamical systems theory has also had various contributions, such as its formulation of the generic behavior in dynamical systems. And physics and dynamical systems have together spawned the development of synergetics and the study of complex systems (Wolfram, 1994), which are closely allied with artificial life. One of artificial life's main influences from physics and mathematics has been an emphasis on studying model systems that are simple enough to have broad generality and to facilitate quantitative analysis.

The first conference on artificial life (Langton, 1989), where the term 'artificial life' was coined, gave recognition to artificial life as a field in its own right, although it had been preceded by a similar conference entitled 'Evolution, Games, and Learning' (Farmer *et al*., 1986). Since then there have been many conferences on artificial life, with strong contributions worldwide (e.g., Bedau *et al*., 2000). In addition to the scientific influences described above, research in artificial life has also come to include elements of chemistry, psychology, linguistics, economics, sociology, anthropology, and philosophy.

CONCEPTS AND METHODOLOGY

Most entities that exhibit lifelike behavior are complex systems – systems made up of many elements simultaneously interacting with each other. One way to understand the global behavior of a complex system is to model that behavior with a simple system of equations that describe how global variables interact. By contrast, the characteristic approach followed in artificial life is to construct lower-level models that themselves are complex systems and then to iterate the models and observe the resulting global behavior. Such lower-level models are sometimes called agent- or individual-based models, because the whole system's behavior is represented only indirectly and arises merely out of the interactions of a collection of directly represented parts ('agents' or 'individuals').

As complex systems change over time, each element changes according to its state and the state of those 'neighbors' with which it interacts. Complex systems typically lack any central control, though they may have boundary conditions. The elements of a complex system are often simple compared to the whole system, and the rules by which the elements interact are also often simple. The behavior of a complex system is simply the aggregate of the changes over time of all of the system's elements. In rare cases the behavior of a complex system may actually be mathematically derived from the rules governing the elements' behavior, but typically a complex system's behavior

cannot be discerned short of empirically observing the emergent behavior of its constituent parts. The elements of a complex system may be connected in a regular way, such as on a Euclidean lattice, or in an irregular way, such as on a random network. Interactions between elements may also be without a fixed pattern, as in molecular dynamics of a chemical soup or interaction of autonomous agents. When adaptation is part of a complex system's dynamics, it is sometimes described as a complex adaptive system. Examples of complex systems include cellular automata, Boolean networks, and neural networks. Examples of complex adaptive systems include neural networks undergoing a learning process and populations of entities evolving by natural selection.

One of the simplest examples of a complex system is the so-called 'game of life' devised by the mathematician John Conway (Berlekamp *et al.*, 1982). The game of life is a two-state two-dimensional cellular automaton with a trivial nearest-neighbor rule. You can think of this 'game' as taking place on a two-dimensional rectangular grid of cells, analogous to a huge checkerboard. Time advances in discrete steps, and a cell's state at a given time is determined by the states of its eight neighboring cells according to the following simple 'birth–death' rule: A 'dead' cell becomes 'alive' if and only if exactly three neighbors were just 'alive', and a 'living' cell 'dies' if and only if fewer than two or more than three neighbors were just 'alive'. From inspection of the birth–death rule, nothing particular can be discerned regarding how the whole system will behave. But when the system is simulated, a rich variety of complicated dynamics can be observed and a complex zoo of structures can be identified and classified (blinkers, gliders, glider guns, logic switching circuits, etc.). It is even possible to construct a universal Turing machine in the game of life and other cellular automata, by cunningly configuring the initial configuration of living cells. In such constructions gliders perform a role of passing signals, and analyzing the computational potential of cellular automata on the basis of glider interactions has become a major research thrust.

Those who model complex adaptive systems encounter a tension resulting from two seemingly conflicting aims. To make a model 'realistic' one is driven to include complicated realistic details about the elements, but to see and understand the emergent global behavior clearly one is driven to simplify the elements as much as possible. Even though complex adaptive systems include systems whose elements and dynamical rules are highly complicated, the spirit of most artificial life work is to look for the complexity in the emergent global behavior of the system, rather than to program the complexity directly into the elements.

Computation is used extensively in the field of artificial life, usually to simulate models to generate data for studying those models. Simulation is essential for the study of complex adaptive systems for it plays the role that observation and experiment play in more conventional science. Having no access to significant computational machinery, Conway and his students first studied the game of life by physically mapping out dynamics with go stones at teatime. Now thousands of evolutionary generations for millions of sites can be computed in short order with a conventional home computer. Computational ability to simulate large-scale complex systems is the single most crucial development that enabled the field of artificial life to flourish and distinguish itself from precursors (such as cybernetics or systems theory).

The dependence of artificial life on simulation has led to debate within the field over the ontological status of the simulations themselves. One version of strong artificial life holds that life may be created completely within a simulation, with its own virtual reality, yet with the same ontological status as the phenomenon of life in the real world. Some hold, however, that simulated, virtual reality cannot possibly have the same ontological status as the reality we experience. These point out that a simulated hurricane can never cause us to become wet. They also believe that if artificial life is to achieve the status of reality, it must include an element of embodiment, an extension into the real, non-simulated world enabling an interaction with that world. Believers in the reality of simulation point out that a simulation has its own embodiment within a computer, that a simulation is not an abstract formula specifying a program but the actual running of a program in a real physical medium using real physical resources. The belief that artificial life has its own bona fide reality is particularly strong among those who generate experimental data with simulations.

Both living systems and artificial life models are commonly said to exhibit emergent behavior – indeed, many consider emergent behavior to be a hallmark of life – but the notion of emergence remains ill-defined. There is general agreement that the term has a precise meaning in some contexts, most notably to refer to the resultant aggregate global behavior of complex systems. The higher-level structures produced in Conway's

game of life provide a classic example of this kind of emergent behavior. In spite of clear examples like the game of life, there is no agreement regarding how one might most usefully define emergence. Some believe that emergence is merely a form of surprise. On this view, emergence exists only in the eye of the beholder and whether a phenomenon is emergent or not depends on the mindset of the observer. Others believe that there is an objective, observer-independent definition of emergence in terms of whether a phenomenon is derivable from the dynamical rules, even if it is often difficult to tell *a priori* what can be derived from the dynamical rules underlying complex systems. These difficulties lead some to argue that the term 'emergence' should simply be dropped from the vocabulary of artificial life. However, this advice is not widely heeded at present.

Complexity is another commonly recognized hallmark of life, and this notion has also so far eluded satisfactory definition. Apparently several different concepts are involved, such as structural complexity, interaction complexity, and temporal complexity. To some, it seems obvious that the biosphere is quite complex at present and that its complexity has increased on an evolutionary timescale. But the difficulties of defining complexity lead others to claim that life's present complexity and its increase over time are either illusory or a contingent artifact of our particular evolutionary history. Understanding complexity and its increase through the course of evolution are at the center of much research in artificial life. In fact, one of the field's main goals at present is to produce and then understand open-ended evolution, an ongoing evolutionary process with continually increasing complexity.

Darwin's view of evolution, with its emphasis on survival of the fittest, implied that the process of adaptation was the key to the creation of intelligent design through life's evolution. However, the role and significance of adaptation is controversial today. Some hold that adaptation is the main force driving the changes observed in evolution. Others maintain that most of evolution consists of non-adaptive changes that simply explore a complex space of morphological forms. Still others claim that much of the apparent intelligence of complex systems is a necessary result of certain complex system architectures. Artificial life may shed light on this debate by providing many diverse examples of evolutionary processes, with an attendant ability to analyze the details of those processes in a way that is impossible with the biosphere, because the analogous assaying of historical data is currently impractical and much of the historical data is simply unavailable.

Analysis of adaptation has led to the idea of a fitness landscape. Organisms (or agents in an artificial life model) are considered to be specified by a genome (or sometimes a set of model parameters). The interaction of the organism with other organisms as well as with its environment yields an overall fitness of the organism, which is often thought of as a real-valued function over the space of possible genomes (or model parameters). In various applications of evolutionary algorithms, such as the genetic algorithm, specifying a fitness function is an essential part of defining the problem. In such cases, adaptation is a form of optimization, 'hill climbing in the fitness landscape'. In artificial life models, however, fitness is often not specified explicitly, but is a property emerging from the interactions of an organism with its world.

The concept of a fitness landscape as an analytical device suffers various limitations. One is that a fitness landscape is generally an approximation; the fitness landscape itself can evolve when organisms in a population interact strongly with each other. Another reason is that on an evolutionary timescale, the space on which a fitness function is defined is changing with the advent of new elements to the genome or new model parameters for artificial organisms. Simulating agent-based artificial life models is a natural and feasible way to study these more general situations.

MODELS AND PHENOMENA

Generally, artificial life models choose a level of biological life to model. The lowest stratum may be thought of as analogous to the chemical level; higher stages include modeling of simple organisms such as bacteria, constituents of more complex organisms such as cells, complex organisms themselves, and varieties of complex organisms that can give rise to ecologies. One might consider a holy grail of artificial life to be the discovery of a single model that can span all these levels; so far the field has had difficulty producing a model that spans even one connected pair of levels.

The most primitive phenomenon explored by some artificial life models is self-organization. Such models study how structure may emerge from unstructured ensembles of initial conditions. Naturally, one aim is to discover the emergence of lifelike structure; some models explicitly aim to model the origin of life – such as chemical soups from which fundamental structures such as self-maintaining autocatalytic networks might be seen

to emerge. Models for the immune system are another example of a lifelike process emerging from chemical interactions. Self-organization has also been studied in models for higher-level living structures, such as metabolisms and cell networks, with Boolean networks whose dynamics converge to different structures depending on model meta-parameters (Kauffman, 1969, 1993).

A host of models target the organismic level, sometimes with significant interactions between organisms. These models typically allow changes in the organisms as part of the system's dynamics (e.g., through a genetic mechanism), and the most common goal of research using these models is to identify and elucidate structure that emerges in the ensuing evolutionary process. Some models fit in between the chemical level and the organismic level, aiming to understand development by modeling interacting cells. Other models are inter-organismic, in the sense that they aim explicitly to model interactions between different types of organisms or agents. These models often contain elements of game theory.

Many of the models studied in artificial life should be viewed as 'purely digital' models. Purely digital models drop any pretense of modeling any pre-existing biological structures; their elements are digital constructs having no direct biological reference. Such models seek to produce novel, purely digital instances of biological phenomena in their emergent behavior. Conway's game of life is a purely digital model at the physical or chemical level, embodying an extremely simple and unique form of 'chemical' interactions (the birth–death rule). The self-organization exhibited in the game of life is not a representation of chemical self-organization in the real world but a wholly novel instance of this phenomenon. Another chemical-level model is AlChemy (Fontana, 1992), which consists of a mixture of 'reacting chemical molecules' that are actually simple programs that produce new programs as output when one program is given as input to another program.

One example of a purely digital model on the 'organismic' level is Tierra (Ray, 1992), which consists of 'organisms' that are actually simple self-replicating computer programs populating an environment consisting of computer memory. Tierra was a mature version of earlier efforts of a model called Core Wars (Dewdney, 1984) and has been followed by more developed versions such as Avida (Adami and Brown, 1994). In Tierra, the world is a one-dimensional ring of computer memory, which may be populated with instructions that are much like idealized microprocessor assembly language instructions (e.g., copy, jump, conditional branch, etc.). The instructions are the microscopic components of the model, and the model's central processing unit (CPU) implements the instructions in memory, creating a chemistry from which structure in the model can emerge. The model is generally seeded with a primordial organism consisting of a group of instructions that can copy itself to another place in memory. The copying is accompanied by errors (mutations) that can enhance the functionality of the organisms.

The accomplishments and shortcomings of most artificial life models are exemplified by those of Tierra. On the side of accomplishments, Tierra shows clear evidence of evolution, and the resulting emergence of structure and organization that were not 'programmed' into the model explicitly. Careful analysis of the evolutionary results reveals computational features such as evolution of subroutines and versions of parasitism. On the negative side, the model shows only one level of emergence (e.g., the model must be seeded by a primordial organism; evolution of an unstructured soup has not yet produced an emergent viable organism). Secondly, the evolution of the digital organisms appears to 'level off', reaching a stage where increasingly insignificant innovations are absorbed into the population, instead of displaying the open-ended evolution of natural systems. Reasons for this limitation include (1) simplicity of the model's evolutionary driving force (the evolutionary value of replication with minimal CPU time), (2) structural limitations on the space of innovations possible, which create limitations on organism functionality, and (3) structural limitations on organisms' ability to interact with each other and their environment. Different artificial life models have different detailed reasons for the two limitations we have discussed in Tierra, but the limitations are generally prevalent.

Another important area of artificial life is not so much a modeling activity as much as an implementation activity. This work aims to produce hardware implementations of lifelike processes. Some of these implementations are practical physical devices. But some of this activity is primarily theoretical, motivated by the belief that the only way to confront the hard questions about how life occurs in the physical world is to study real physical systems. Again, there is an analogy with biological levels. The 'chemical' level is represented by work on evolvable hardware, often using programmable logic arrays (e.g., Breyer *et al*., 1998). The 'organismic' level is represented by recent work in evolutionary robotics (e.g., Cliff *et al*., 1993). An

'ecological' level might be represented by the Internet along with its interactions with all its users on computers distributed around the world.

Artificial life, like its antecedent, cybernetics, has a peculiarly broad cultural scope extending beyond cut and dried scientific progress. This breadth is best exemplified by the work of Karl Sims (Sims, 1991), who has coupled rich image-producing computational environments with interactions between those environments and people watching the images at an exhibit. The result is an evolutionary system that is not constrained to live within the confines of a particular model's framework, but rather that is a coupling of two evolutionary subsystems, one of which is natural (the audience). Sims' interactive evolutionary art has produced several visually striking results, and human interaction seems to give the evolutionary system an open-ended quality characteristic of natural evolution.

FUTURE DIRECTIONS

One broad direction artificial life will continue to take in the future is that of synthesis: the synthesis of significant biological phenomena either within the context of model simulation or hardware implementation. A grave difficulty facing progress in this area is the lack of any quantitative basis of comparison for many of the biological phenomena artificial life aims to model. An example of this difficulty is modeling open-ended evolution. How could we know when this is achieved? In general, measurable characterization of phenomena is a prerequisite to quantitative comparison, and much progress is needed in order to achieve this for many target phenomena.

Probably the largest goal of the field is to understand the nature of life itself. This will be furthered to some extent with the quantitative comparisons just mentioned, but there is also a broader goal of discerning what the boundaries of life are, and how the idea of life might be extended to phenomena beyond biological life. Is there a sense in which financial markets or sociotechnical networks are alive, independent of the lives of their biological constituents? Many in the field of artificial life believe that, if the concept of life is properly framed and understood, such questions may well have a precise affirmative answer.

References

Adami C and Brown CT (1994) Evolutionary learning in the 2D artificial life system 'Avida'. In: Brooks R and Maes P (eds) *Artificial Life*, vol. IV, pp. 377–381. Cambridge, MA: Bradford/MIT Press.

Bedau MA, McCaskill JS, Packard NH and Rasmussen S (eds) (2000) *Artificial Life VII: Proceedings of the Seventh International Conference on Artificial Life*. Cambridge, MA: MIT Press.

Berlekamp ER, Conway JH and Guy RK (1982) *Winning Ways*, vol. II. New York, NY: Academic Press.

Breyer J, Ackermann J and McCaskill JS (1998) Evolving reaction–diffusion ecosystems with self-assembling structures in thin films. *Artificial Life* **4**: 25–40.

Cliff D, Harvey I and Husbands P (1993) Explorations in evolutionary robotics. *Adaptive Behavior* **2**: 73–110.

Dewdney A (1984) In a game called core wars hostile programs engage in a battle of bits. *Scientific American* **250**: 14–23.

Farmer JD, Lapedes A, Packard NH and Wendroff B (eds) (1986) *Evolution, Games, and Learning: Models for Adaptation for Machines and Nature*. Amsterdam, Netherlands: North-Holland.

Fontana W (1992) Algorithmic Chemistry. In: Langton CG, Taylor CE, Farmer JD and Rasmussen S (eds) *Artificial Life II: Proceedings of the Workshop on Artificial Life*, pp. 159–209. Reading, CA: Addison Wesley.

Holland JH (1975) *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press. [Revised and expanded edition (1992) Cambridge, MA: MIT Press.]

Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology* **119**: 437–467.

Kauffman SA (1993) *The Origins of Order: Self-Organization and Selection in Evolution*. New York, NY: Oxford University Press.

Langton CG (ed.) (1989) *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*. Redwood City, CA: Addison Wesley.

Ray T (1992) An approach to the synthesis of life. In: Langton CG, Taylor CE, Farmer JD and Rasmussen S (eds) *Artificial Life II: Proceedings of the Workshop on Artificial Life*, pp. 371–408. Reading, CA: Addison Wesley.

Sims K (1991) Artificial evolution for computer graphics. *Computer Graphics* **25**: 319–328. (ACM SIGGRAPH '91 Conference Proceedings, Las Vegas, Nevada, July 1991.)

Wiener N (1948) *Cybernetics, or Control and Communication in the Animal and the Machine*. New York, NY: Wiley.

Wolfram S (1994) *Cellular Automata and Complexity: Collected Papers*. Reading, MA: Addison Wesley.

Further Reading

Adami C (1998) *Introduction to Artificial Life*. New York, NY: Springer.

Adami C, Belew RK, Kitano H and Taylor CE (1998) *Artificial Life VI: Proceedings of the Sixth International Conference on Artificial Life*. Cambridge, MA: MIT Press.

Bedau MA (1997) Weak emergence. *Philosophical Perspectives* **11**: 375–399.

Bedau M, McCaskill J, Packard N *et al*. (2000) Open problems in artificial life. *Artificial Life* **6**: 363–376.

Boden MA (1996) *The Philosophy of Artificial Life*. Oxford, UK: Oxford University Press.

Brooks RA and Maes P (1994) *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*. Cambridge, MA: MIT Press.

Floreano D, Nicoud J-D and Mondada F (1999) *Advances in Artificial Life: 5th European Conference, ECAL'99*. Berlin, Germany: Springer.

Husbands P and Harvey I (1997) *Fourth European Conference on Artificial Life*. Cambridge, MA: MIT Press.

Langton CG (ed.) (1995) *Artificial Life: An Overview*. Cambridge, MA: MIT Press.

Langton CG, Taylor CE, Farmer JD and Rasmussen S (eds) (1992) *Artificial Life II: Proceedings of the Workshop on Artificial Life*. Reading, CA: Addison Wesley.

Levy S (1992) *Artificial Life, The Quest for a New Creation*. New York, NY: Pantheon.

Koza JR (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press.

Mitchell M (1996) *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Varela FJ and Bourgine P (1992) *Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*. Cambridge, MA: MIT Press.

Attractor Networks

Intermediate article

Garrison W Cottrell, University of California, San Diego, California, USA

CONTENTS

Attractor networks are types of neural network that are often used to represent human subjects' content-addressable memory. That is, given part of a memory, such as a name, we can complete the pattern and recall other information about the individual associated with that name.

INTRODUCTION: COGNITION AS ACTIVATION DYNAMICS

How is it we recall information? Many psychologists have suggested that our memory is *reintegrative* – that is, we 're-integrate' the information each time we recall something. The correlate in computer science is called *content-addressable memory* (CAM). Standard retrieval of information from a computer's memory requires an *address*, the location of the information. One can, however, buy an expensive piece of hardware that will retrieve the bits at any address in the memory that match a partial pattern of bits; i.e., based on the *content* of that memory. What is expensive to a computer scientist seems effortless for a human. For example, I can tell you I am thinking of someone who is a former actor, a former President of the United States, a Republican, quite elderly, and a Rhodes scholar. You probably retrieved 'Ronald Reagan', even though part of the description is incorrect. You retrieved this memory through *pattern completion*: given part of the memory, you retrieved the rest of the memory.

In neural networks, systems that provide for the storage and retrieval of such patterns are called *attractor networks*, and have been used to model such diverse phenomena as memory, lexical access, reading, aphasia, dyslexia, associating names and faces, and face recognition. This article very briefly reviews some of the mathematics of such networks and how they are trained.

HOPFIELD NETWORKS

Hopfield networks, named after the physicist John Hopfield who studied them and proved many of their properties, are a particular kind of neural network in which the units are symmetrically connected. (Such networks had been studied earlier; see Cowan and Sharp (1988) for a review. Hopfield (1982) became associated with these networks by proving their stability properties using an energy function and by promoting the idea of them as memory models.) Hopfield networks are important because there is a great deal of elegant theory surrounding them. Unfortunately, the theory shows that as memories, they have a very small capacity. However, the important role they played in the development of the theoretical properties of neural networks makes them worthy of consideration. The following discussion owes much to Hertz *et al*. (1991).

A binary Hopfield network is a collection of N *units*, connected by weighted links. A unit is represented by its activation value, y_i:

$$y_i = g(u_i) \tag{1}$$

$$u_i = \sum_{j=0}^{N} w_{ij} y_j \tag{2}$$

$$g(x) = 1 \text{ if } x \geq 0, \text{ else } -1 \tag{3}$$

where $g(x)$ is called the *activation function* of the units (here, it is also known as the sign function), u_i is often called the *net input* to the unit, and w_{ij} is the weight from unit j to unit i. Also, $y_0 = 1$ by definition. w_{i0} is sometimes called the *bias* of the unit, and is equivalent to the negative of a threshold (which it is explicitly when y_0 is defined as -1 instead of 1). In Hopfield networks, there is a constraint that $w_{ij} = w_{ji}$; that is, the network is symmetrically connected. This is explained later.

This system has a *dynamics*. That is, if we start the system in some state (a pattern of activation on the y_i), it updates its state over time based upon the above equations. For a 'standard' Hopfield network, this is done asynchronously. That is, on each time step of the system, we randomly choose an i between 1 and N, and update the activation of unit i using the above equations.

Given this formulation, the question is: how can we store a set of patterns of activation such that, when presented with a new pattern, the system activations evolve to the closest stored pattern? This idea is illustrated abstractly in Figure 1. The coordinates of the graph are the activations of two continuously-valued units (note this is easier to draw than if we used $-1, 1$ units, as we are here). The 'X's represent stored activation patterns. The arrows represent the idea that, starting from nearby patterns, the system should move towards the nearest one. The boundaries around any pattern are called the *attractor basin* for that pattern.

Following Hertz *et al*. (1991), we start with one pattern, call it $\vec{x} \in \{-1, 1\}^N$, a vector of 1s and -1s of length N. We would at least like this pattern to be *stable*; that is, if we impose this activation pattern on the units by setting all of the y_i, the network should not move from that pattern. The requirement for stability is simply:

$$g\left(\sum_j w_{ij} x_j\right) = x_i \quad \forall_i \tag{4}$$

If the update rule is applied, then, nothing changes. The question, then, is how to set the weights w_{ij} to guarantee this. One common rule is:

$$w_{ij} = \frac{1}{N} x_i x_j \tag{5}$$

Then (assuming no bias):

$$\begin{aligned} u_i &= \sum_j w_{ij} x_j \\ &= \frac{1}{N} \sum_j x_i x_j x_j \\ &= \frac{1}{N} \sum_j x_i = x_i \end{aligned} \tag{6}$$

as required (note: $g(x_i) = x_i$). In fact, if fewer than half of the elements are 'wrong', they will be overruled by the majority that are correct. Thus this system has an attracting state that is the desired pattern. Given a partially correct pattern, it will 'complete' the pattern, just as the reader completed the 'Reagan' pattern. It should be noted that it also possesses another attractor, $-\vec{x}$, called the *reversed state* (Hertz *et al*., 1991). This is the other attracting state of the system.

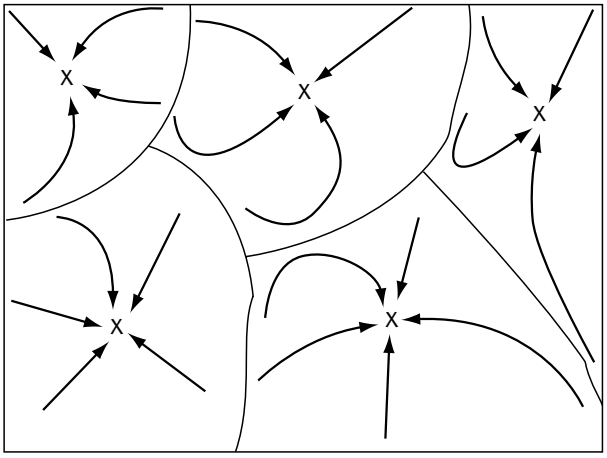


Figure 1. Attractor basins in an imaginary network. The x, y axes represent activations of two of the units, in this case, continuous valued.

For multiple patterns, we just overlay the weight prescription of each pattern:

$$w_{ij} = \frac{1}{N} \sum_p x_i^p x_j^p \tag{7}$$

where p ranges over all of the patterns. This is called the *outer product rule* or *Hebb rule* after Donald Hebb (1949), who proposed a similar idea in his classic book, *The Organization of Behavior*. If one imagines that these patterns are imposed (say, via perception) on the set of units, then this rule sets the weights according to the correlation between their firing. In neuroscience, the slogan is, 'neurons that fire together wire together'.

In a learning setting, we can imagine that upon presentation (or imposition) of a pattern (i.e., all y_i are set to the corresponding x_i^p), starting from some random initial weights, each weight is updated by:

$$w_{ij} := w_{ij} + \eta y_i y_j \tag{8}$$

where η is a (usually small) learning rate parameter. Over many presentations of the patterns, the weights will become proportional to the correlation between the elements of the patterns. They will also grow without bound, a problem that can be addressed by artificially limiting the size of the weights, or by adding a 'weight decay' (or 'forgetting') term that moves the weights slowly back towards 0.

Intuitively, we can think of the weights in the network as *constraints* between the units. Suppose,

for the sake of exposition, some unit represents the feature 'happy' and another unit represents the feature 'has a big smile'. We would like both of these units to be 'on' ($y_{happy} = 1, y_{smile} = 1$), or both of these units to be 'off' ($y_{happy} = -1, y_{smile} = -1$) in a stable pattern. Then they should have a positive weight between them, since the features are correlated. On the other hand, these units should have a negative connection to a unit representing 'sad'. Then if $y_{sad} = 1$, it will tend to try to turn off y_{happy} and vice versa. Hence the weights represent the way that features may 'vote' for their friends (positive weights) and vote against their enemies (negative weights).

Returning to the rule for the weights given in equation 7, the requirement for stability of a pattern becomes:

$$g\left(\sum_j w_{ij} x_j^p\right) = x_i^p \quad \forall i, p \tag{9}$$

Note that we can decompose the sum into:

$$\begin{aligned}\sum_j w_{ij} x_j^p &= \frac{1}{N} \sum_j \sum_{p'} x_i^{p'} x_j^{p'} x_j^p \\ &= x_i^p + \frac{1}{N} \sum_j \sum_{p' \neq p} x_i^{p'} x_j^{p'} x_j^p \end{aligned} \tag{10}$$

The second term is called the *crosstalk* term. The pattern will be stable if the crosstalk term is the same sign as x_j^p for all j, or if its magnitude is less than 1, so that it does not overwhelm x_j^p.

The question now becomes: what is the capacity of such a network? That is, how many such patterns can reliably be stored? First, by 'stored' we simply mean that the pattern is stable, not that a partial pattern will complete to a full pattern. Also, there are several ways in which we might define 'reliably'. For example, we might require that the patterns remain exactly as stored, or that they change only a small percentage of their bits, or that they are stable with some probability and some amount of distortion. It also matters how correlated the patterns are with one another. The details are beyond the scope of this article, but the general result agrees with Hopfield's empirical finding in his original article that the capacity is about $0.15N$. The most quoted figure for the capacity is $0.138N$, which corresponds to about 1.6% of the bits in a pattern changing to an incorrect setting before it stabilizes (Hertz *et al*., 1991). This means that to store, say, one pattern from every day of our lives until we reach 50 years, we would need of the order of 132,000 units. Looked at another way, given that we have 10^{11}–10^{12} brain cells, we could store of the order of 10^{10}–10^{11} memories – if the brain were a Hopfield network, and if it had to do nothing but store memories. However, even if we used only 1% of our brains for memory, we would still be able to store around 100 million memories by this calculation. So perhaps the view that they have small capacity is not such an worry.

ENERGY FUNCTIONS AND CONVERGENCE

One of the novel contributions of Hopfield's original paper was the proof of convergence of a Hopfield network. *Convergence* here means that when the network starts in some state of activation, using the update equations 1–3, the network will reach a point where no activations change. Note that this does not mean that it will converge to one of the stored memories! It simply means that the network will not oscillate, or become chaotic. The proof relies on the notion of an *energy function*, also known in physics as a *Lyapunov function*. The idea is that the energy function is a real-valued function of the state of the network, that it is bounded from below, and that the update equations of the network always make this number stay the same or go down. If one can come up with a Lyapunov function for a system, one can infer that the system reaches a point where nothing changes. (This is speaking rather loosely. There may be updates that move the state of the system to another state with the same energy value. However, in the proof below, we assume that only one unit changes its state, which avoids this problem.) The particular Lyapunov function Hopfield proposed is:

$$E = -\frac{1}{2} \sum_{ij} w_{ij} y_i y_j \tag{11}$$

This expression, without the negation sign, has also been called the 'goodness' function of the network by Rumelhart *et al*. (1986b), in which case it always goes up or stays the same.

Here we have excluded the contribution of any external input to the network, which can be easily incorporated into this expression. Given a set of weights, E clearly has a minimum value (it is bounded from below). Intuitively, we can think of this as a representation of how many constraints have been violated by the current activation state. That is, suppose two units are connected by a positive link. Then the constraint between them is that they should both be on or both be off, as in the 'happy/smile' example above. This 'constraintlet'

is represented by two terms (redundantly) in the above sum, $w_{happy\,smile} y_{happy} y_{smile}$ and its dual. If both units were on, then that would be a positive component of the sum, which would mean a more negative overall energy (more constraints satisfied). If one of them was 1 and the other -1, then the energy would be higher, all other things being equal. Similarly, if w_{ij} is negative, then the energy is lower when y_i and y_j are of opposite sign, versus when they have the same sign.

Another way to conceive of this is to imagine the values of E for different activation values of the network. If there are just two units in the network, we can lay out the possible states of the system in a plane. For some values of the units, E will be high, and for others, E will be low. Doing some injustice to the fact that we have only four discrete states, we can imagine that E forms an *energy landscape* over this two-dimensional state space of the system, as shown in Figure 2. The surface shows the height of the energy. The state of the network can be thought of as a ball on this surface. Updating the network state will cause the ball to roll downhill. This is just a more detailed picture of the same phenomenon shown in Figure 1.

Going back to the equation for E, it is intuitively clear why the operation of the network dynamics (equations 1–3) lowers this number. For example, if a unit's current value is -1, and the input from the rest of the network is positive, the unit will update its state to be $+1$, which will violate fewer constraints. To see this formally, suppose that a unit changes state upon updating (if it does not change state, the energy remains the same). Let y_i' represent the new state of unit i. Then the difference in energy with y_i in its new state, E', and the previous E only involves the terms with y_i and y_i' in them:

$$\begin{aligned} E' - E &= -\sum_j w_{ij} y_i' y_j + \sum_j w_{ij} y_i y_j \\ &= -y_i' \sum_j w_{ij} y_j + y_i \sum_j w_{ij} y_j \\ &= (y_i - y_i') \sum_j w_{ij} y_j \\ &= (y_i - y_i') u_i' \end{aligned} \tag{12}$$

By assumption, u_i' is of opposite sign to y_i, and must be the same sign as y_i'. Hence the difference is negative, and E' must therefore be of lower energy than E, as required.

BOLTZMANN MACHINES

One possible problem with Hopfield networks is that, since the update equation is deterministic, the energy always decreases. This means that the network will always move to the lowest nearby minimum, even if there is a better minimum

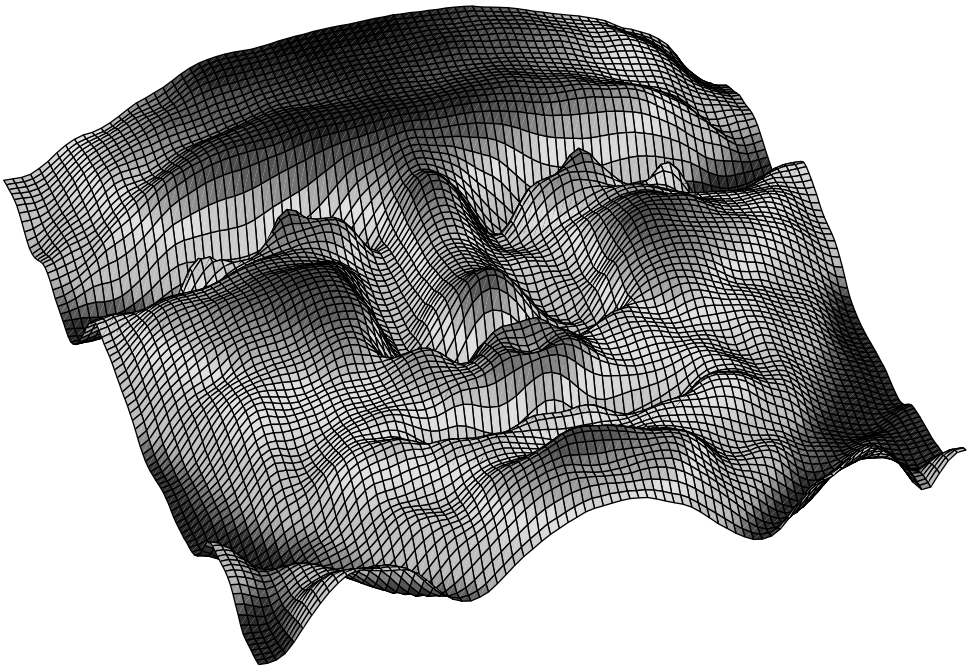


Figure 2. [*Figure is also reproduced in color section.*] An imaginary energy landscape.

somewhere else. The situation where this matters is where some of the units in the network are *clamped* – that is, their activation values are held fixed – and the problem is to find the 'best completion' of the pattern, given the constraints encoded in the network weights. One possible fix to this problem is to have the network sometimes go *uphill* in energy with some small probability. *Boltzmann machines* are one embodiment of this notion (Hinton and Sejnowski, 1986). The formulation of Boltzmann machines is somewhat different from the standard Hopfield network. First, the units are *stochastic*, and have states that are either 0 or 1:

$$s_i = \begin{cases} 1 \text{ with probability } g(u_i) \\ 0 \text{ with probability } 1 - g(u_i) \end{cases} \quad (13)$$

The probability of a unit being 1 is usually taken to be a *sigmoidal* function of the input:

$$g(u_i) = \frac{1}{1 + exp(\frac{-u_i}{T})} \quad (14)$$

The parameter T is called the *temperature*, because of an analogy between the operation of the system and spin glass models in statistical physics. It controls the steepness of the function g. T can be thought of as a noise parameter. The bigger T is, the less each unit will respond to the input from other units (the function g will be very flat). If T were infinite, each unit would flip between 0 and 1 with 50% probability. Thus, the basic idea in the operation of a Boltzmann machine is to start with a high T, and slowly lower T until it is near 0 and the system is practically deterministic (the function g will be very close to a threshold unit as in equations 1–3). Hinton has described the idea as follows (paraphrasing). Suppose you have a black box with a surface like that in Figure 2 inside, and a ball is dropped into the box. Your job is to get the ball to the lowest spot on the surface, obviously without being able to see inside the box. One way to do this is to shake the box vigorously, then slowly shake it less and less. The energetic shaking should get the ball into the largest well in the box, and as the shaking subsides, it should get into the lowest spot. One can prove that if you shake the box for an infinitely long time, and slow your shaking appropriately, then the ball will end up in the lowest spot in the box.

Another way to think about what T does is that when it is big, it smooths out the bumps in the energy landscape, as if you are squinting at it. As it is lowered, the smaller bumps will emerge, and the ball's behavior will depend more on the fine details of the landscape. It should be clear that no one is proposing that your brain 'heats up' as you are recalling memories! However, the idea that noise may help in reaching better states is not so far-fetched. It has been shown that laughing in the middle of a test improves performance.

Another novelty in Boltzmann machines is the idea of *hidden units* in the network. These are units that are not part of the stored patterns (which are placed on the *visible units*), but can be used to differentiate between patterns that might otherwise be confused because they are too similar. The introduction of a set of units that are not part of the stored patterns makes the use of the Hebb rule problematic, as the states of the hidden units are not specified. However, learning in Boltzmann machines nevertheless turns out to be relatively simple conceptually, but tends to be very slow. (Recently, Hinton has developed a relatively fast learning algorithm for a special case of Boltzmann machines (Hinton, 2002), but it is beyond the scope of this article.) The basic idea is to have two phases. In one phase, the patterns that are to be stored are clamped on the visible units. The network is run for a long time, while statistics are collected on how often units are on and off together. Then the network is run again without the patterns clamped. The weights are updated according to:

$$w_{ij} = \eta(< y_i^+ y_j^+ > - < y_i^- y_j^- >) \quad (15)$$

where the angle brackets mean averages over time, η is a learning rate parameter, y_i^+ refers to the clamped phase, and y_i^- refers to the unclamped phase. Notice that this rule will still result in symmetrically connected networks. Intuitively, one is subtracting off the statistics of the network running 'on its own' from the statistics of the network when the desired pattern is present. This has been compared to an 'awake' state (where the network is clamped by its perceptions of the environment) versus a 'sleep' state (the unclamped phase). Such evocative imagery has not yet been verified by neuroscientists, but one really wants it to be true.

One advantage of such networks is that, if they learn successfully, operating them without inputs will result in the network displaying on the visible units the entire *distribution* of the training environment. Unlike deterministic networks, then, these networks can be thought of as *generative models* of their environment. Also, if one thinks of the networks as simply learning by exposure to an environment, they can be thought of as *unsupervised* models that learn (on the hidden units) efficient encodings (features) of their environment. This has appeal as a model of how humans learn from 'mere exposure' to the environment. Of course, we

start with a highly structured neural network, based upon millions of years of evolution. Integrating such pre-structuring of the network with appropriate learning rules is the subject of a great deal of current research.

SEQUENTIAL ATTRACTORS: BACKPROPAGATION THROUGH TIME

While the networks described so far are attractive as models of pattern completion memory, they also seem flawed as a model of how brains might work. First of all, they require the connections in the network to be symmetric. There is little evidence that the neurons in brains, human or otherwise, are so connected. Second, if one takes seriously the notion of finding a stable state of the network, the idea that our neurons settle to a stable state is not particularly palatable. As Walter Freeman has remarked, 'the only stable neuron is a dead neuron'. One would like a model that perhaps reaches stable states transiently, and then progresses to a new state. While there has been some work in this area for Hopfield networks (see Hertz *et al*. (1991) for examples), there has been more work on training networks to go through sequences of states using *supervised*, or *error-correction*, learning techniques. The standard approach is called *backpropagation through time* (BPTT). While the full details are beyond the scope of this article, we can summarize some of the main points.

First, note that if we eschew stable states, the networks must *not* be symmetrically connected. Otherwise, there would be a Lyapunov function to describe their dynamics. Second, if one wants the network to go through different states, it must be told, in some way or another, what those states are. Hence, the training must be supervised – states are specified for the trajectory of the network, and the network is required to pass through those states in the order specified. We may retain the notion of attractors if we generalize it to cyclical behaviors. This means that, for example, if the network is somehow pushed out of the trajectory it has been trained to produce, it will move back towards that trajectory. Finally, such systems usually use *continuous* units, that take values in the range [0,1]. A standard equation for the activity of such units is the logistic equation:

$$g(u_i) = \frac{1}{1 + exp(-u_i)} \tag{16}$$

which bears a remarkable resemblance to the equation for the probability of a unit being 'on' in a Boltzmann machine. However, here this produces not a probability, but the actual activation of the unit. Otherwise, the u_i is calculated in exactly the same way as in equations 1–3.

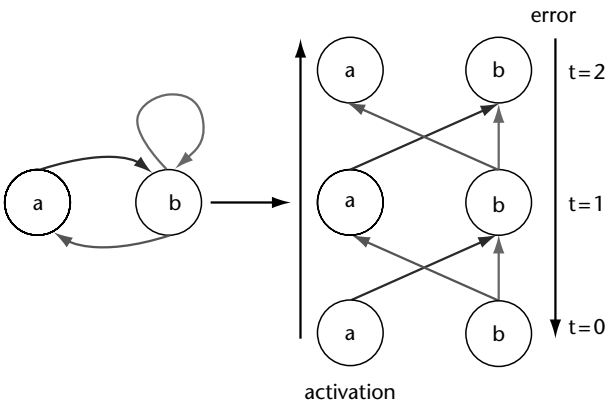


Figure 3. [*Figure is also reproduced in color section*.] Backpropagation in time.

BPTT uses the idea illustrated in Figure 3. On the left of the figure, there is a simple asymmetric recurrent network. This network is converted into a so-called feedforward network by 'unrolling' it in time, as shown on the right. Backpropagation (Rumelhart *et al*., 1986a) is a supervised learning technique that adjusts the weights in a network in order to reduce the error in the state of the network. Errors, in the form of target states, can be 'injected' into the network at any time step, and the weights are adjusted to make the state of the network closer to the target state. Errors are then propagated backwards through the network (hence the name 'backpropagation'), in this case through time (hence the name, BPTT).

The weighted links are color-coded to show how the links in the feedforward version relate to the links in the recurrent version. Also, since links of the same color in the feedforward version are, conceptually, the *same* link, this means that any weight adjustments to one link must be made to all of the links of the same color. Essentially, in BPTT, the weight changes to each link are added together for links of the same color. One of the most striking examples of BPTT in action was its use to train a system to back up a semi-tractor trailer, which required many steps and is a complex task that requires a very skilled human operator to perform (Nguyen and Widrow, 1989).

PHASE-SPACE LEARNING

BPTT in its basic form has problems in training systems with complicated attractors, especially if

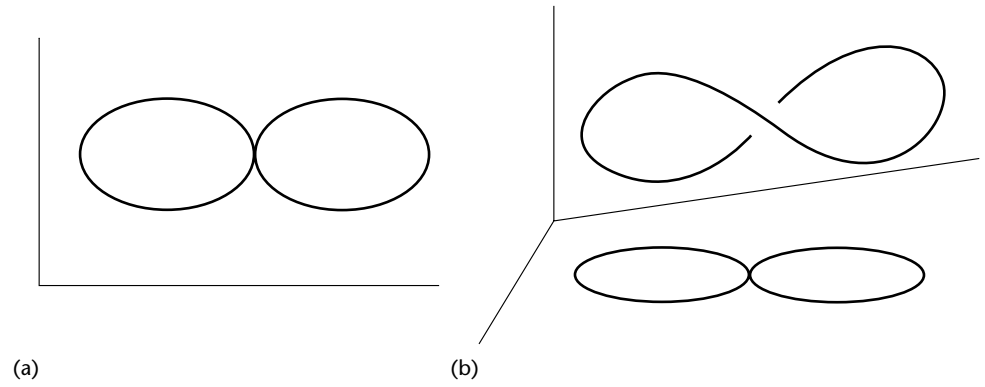


Figure 4. Embedding a '∞' from 2-D into 3-D.

one wants multiple sequential attractors in the same system (see Tsung and Cottrell (1995) for a discussion). A mild variation on BPTT produces surprising results. Tsung and Cottrell (1995) introduced a variation that applies in the case where the desired system dynamics are deterministic, which is often the case (an example where this is not the case is in machine translation; for example, the Spanish *casa* can be translated into English as *house* or *home*). It uses two ideas. The first, borrowed from time-series prediction, is to perform a *delay-space embedding* of the desired trajectory. This is useful in cases where the observed behavior does not appear deterministic, but could be, if more dimensions were introduced. Take, for example, a figure 8 shape. If the two visible units of a system (represented by the two axes in Figure 4(a)) are supposed to describe a figure 8 with their behavior, what happens at the middle point of the 8 is not determined – the system could go one of two ways. However, if there was a third dimension that basically raised one of the curves above the other (think of a raised highway), then the system would be deterministic. The idea is shown in Figure 4(b). Note that the 'shadow' of the trajectory in 2-D is still a figure 8 shape, but the system has a third variable that allows the crossing point to be separated in space.

Delay-space embedding is a technique for adding more state variables to avoid such crossing points, basically by taking the desired trajectory and adding more variables that are simply the ones we already have, but delayed in time. The skill is in picking the amount of delay, and the number of extra variables formed this way (Kennel *et al*., 1992). An important property of a proper embedding is that each point on the trajectory uniquely determines the next point on the trajectory. The space that the system is embedded into is called a *phase space*. In this space, time is represented by movement through the space. For example, the picture given in Figure 1 is a phase space picture: movement along the arrows is movement in time for the system. In this case, the memories given by the Xs in that picture are called *fixed point* attractors. We are interested here in attractors that may involve, for example, closed loops in phase space, which correspond to oscillations. An example of this has already been shown in Figure 4.

The second idea is, once the system has been made deterministic in this way, the mapping from one point to another is a *map*; i.e., there is some function, let's call it f, that produces the next point on the trajectory given the current point. Introducing t, a time variable, $f(\vec{y}(t)) \rightarrow \vec{y}(t + \delta t)$, for some increment δt of our choosing. We can use a feedforward neural network to learn f from examples. Essentially, we are doing BPTT only one δt step back in time. However, given the way we have arranged things via delay-space embedding, this is all that is needed. Now, given this map, we may *iterate* it. That is, once we start from some point, $\vec{y}(0)$, and obtain $f(\vec{y}(0)) = \vec{y}(\delta t)$, we can apply f again, to get the next point, and so on. The final idea is that we can make any trajectory an attractor by training the network to start from points near the desired attractor and making the target closer to the attractor. Tsung and Cottrell (1995) noted that this function f may be arbitrarily complex, so that hidden units between the input and the output may be needed. The idea is shown in Figure 5.

Thus, *phase-space learning* consists of: (1) embedding the trajectory to avoid crossing points, (2) sampling trajectory elements near the desired trajectory, and (3) training a feedforward network on these trajectory elements. Since feedforward

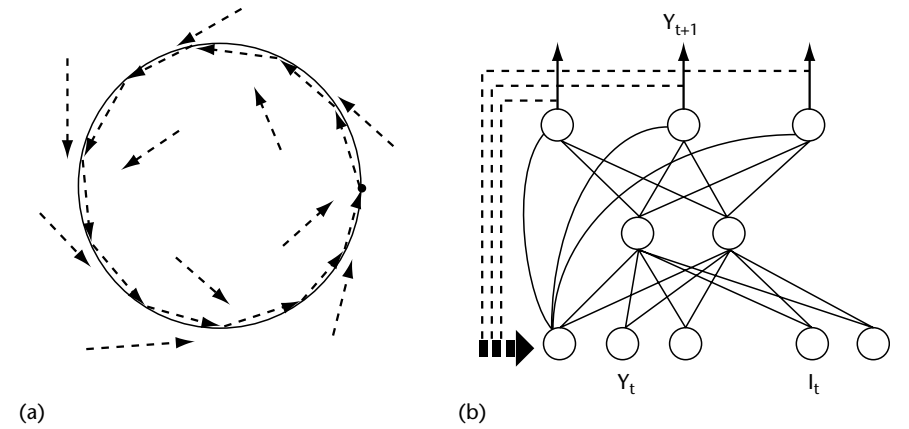


Figure 5. Phase-space learning. (a) The training set is a sample of the trajectory elements. (b) Phase-space learning network. Dashed connections are used after learning. From Tsung and Cottrell (1995), reprinted by permission of MIT Press.

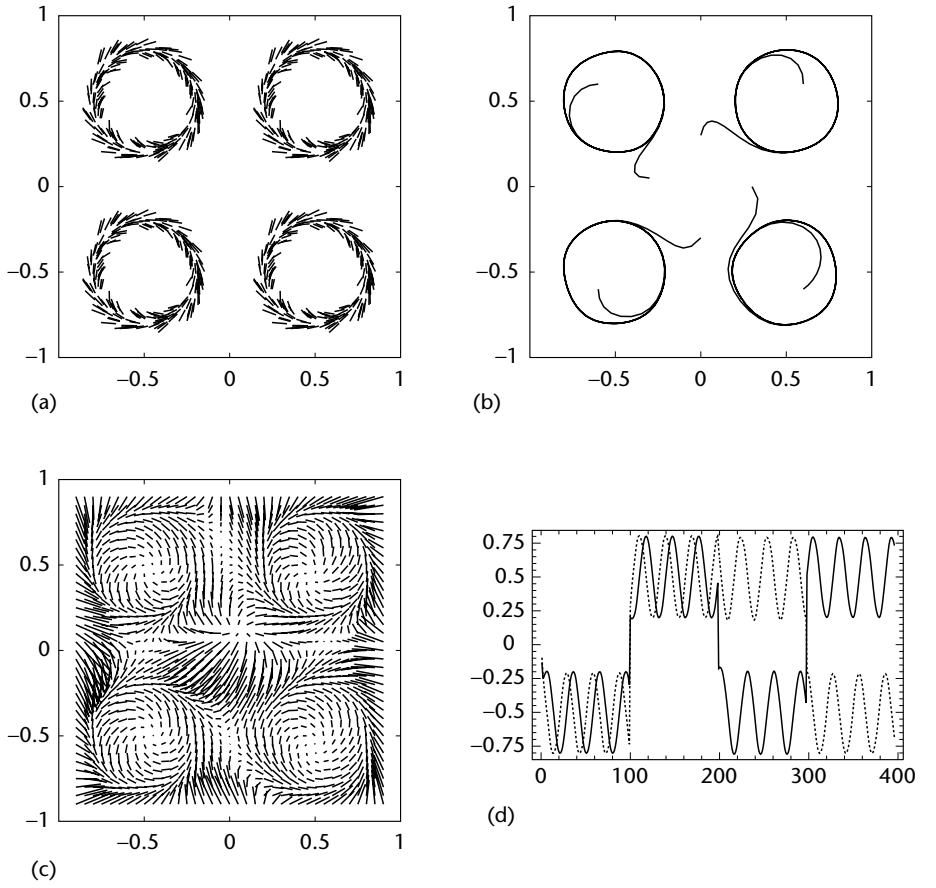


Figure 6. Learning four coexisting periodic attractors. The network had 2-20-20-2 units and was trained using back-propagation with conjugate gradient for 6000 passes through the training set: (a) the training set: 250 data pairs for each of the attractors; (b) eight trajectories of the trained network delineate the four attractors; (c) vector field of the network: this shows, for every little arrow, where the network would go next; (d) graph of the activations of the two visible units over time, as they are 'bumped' into different attractor basins. From Tsung and Cottrell (1995), reprinted by permission of MIT Press.

networks are universal approximators (Hornik *et al*., 1989), we are assured that, at least locally, the trajectory can be represented. The trajectory is recovered from the iterated output of the pre-embedded portion of the visible units (the ones we started with – e.g., the 'shadow' of the embedded system in the original space, as in Figure 4(b)). Additionally, we may extend the phase-space learning framework to also include time-varying inputs that affect the output of the system, as shown in Figure 5, bottom right.

The phase-space learning approach has no difficulties storing multiple attractors. Learning multiple attractors can be done in the same way a single attractor is trained; one simply includes a sufficient number of trajectory segments near all of the desired attractors. Figure 6 shows the result of training four coexisting oscillating attractors, one in each quadrant of the two-dimensional phase space. The underlying feedforward network has two inputs, two layers of 20 hidden units each, and two outputs. The network will remain in one of the oscillating regimes until an external force pushes it into another attractor basin. Such oscillating attractors are called *limit cycles* in dynamical systems theory.

Similarly, such a system can be used to avoid the problems inherent in standard Hopfield networks. The author has used phase-space learning to create a standard fixed-point attractor network to store 'meaning' patterns derived from co-occurrence counts. Specifically, 233 word vectors were used that were processed versions of vectors obtained from Curt Burgess at UC Riverside (Lund *et al*., 1995). The words were those used by Chiarello *et al*. (1990) in their priming experiments. They represented various words from 'ale' to 'wool'. The structure of the vectors was such that, for example, all of the food words were similar, all of the clothing words were similar, etc. The vectors were 36-dimensional (hence 36 visible units would be required to store them), and the elements were $+/-1$. Hence they were perfect for storing in a Hopfield network, except for one thing: there were over six times as many vectors as units, and a Hopfield network with 36 units should be able to store about five vectors. Instead, the author used phase-space learning, in the following way: a 36-70-36 feedforward network was used – that is, there were 36 inputs, 70 hidden units, and 36 outputs. On every training trial, one of the vectors was chosen to present on the input. 25% of the bits in the vector were probabilistically set to 0. The network was trained to produce the original vector from this nearby vector.

Once trained, it could be iterated by copying the outputs back to the inputs. One can think of this network as a recurrent network of 36 units, with 70 hidden units, by 'folding over' the output onto the input. The activation starts at the input, flows to the hidden units, and back to the input. There were also direct connections from the units to themselves (from the input to the output in the original network). This network was trained in about 10 minutes of Cray time (circa 1995), and produced the correct vector about 98% of the time. This demonstrates that the capacity of such networks is much higher than that of standard Hopfield networks.

CONCLUSION

There is not space in this article to cover several related topics. In particular, the reader should be aware that there are continuous-valued Hopfield networks (Hopfield, 1984); a deterministic Boltzmann learning algorithm (Hinton, 1989) that learns much faster than the standard algorithm; and recurrent networks that both recognize and produce sequences (Elman, 1990; Jordan, 1986) in cases where deterministic methods such as phase-space learning do not apply.

References

Chiarello C, Burgess C, Richards L and Pollock A (1990) Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't… sometimes, some places. *Brain and Language* **38**: 75–104.

Cowan J and Sharp DH (1988) Neural nets. *Quarterly Reviews of Biophysics* **21**: 365–427.

Elman JL (1990) Finding structure in time. *Cognitive Science* **14**(2): 179–212.

Hebb D (1949) *The Organization of Behavior*. New York, NY: John Wiley.

Hertz J, Krogh A and Palmer RG (1991) *Introduction to the Theory of Neural Computation*, volume I of *Lecture Notes in the Santa Fe Institute Studies in the Sciences of Complexity*. Redwood City, CA: Addison-Wesley.

Hinton G (1989) Deterministic boltzmann learning performs steepest descent in weight space. *Neural Computation* **1**: 143–150.

Hinton G (2002) Training products of experts by minimizing contrastive divergence. *Neural Computation* **14**(8): 1771–1800.

Hinton GE and Sejnowski TJ (1986) Learning and relearning in Boltzmann machines. In: Rumelhart D and McClelland J (eds) *Parallel Distributed Processing*, vol. I, chap. VII, pp. 282–317. Cambridge, MA: MIT Press.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA* **79**: 2554–2558.

Hopfield J (1984) Neurons with graded responses have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the USA* **81**: 3088–3092.

Hornik K, Stinchcombe M and White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359–366.

Jordan MI (1986) *Serial Order: A Parallel Distributed Processing Approach*. Technical report, Institute for Cognitive Science, UCSD.

Kennel M, Brown R and Abarbanel H (1992) Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A* **45**: 3403–3411.

Lund K, Burgess C and Atchley RA (1995) Semantic and associative priming in high-dimensional semantic space. In: Moore JD and Lehman JF (eds) *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 660–665. 22–25 July, University of Pittsburg. Hillsdale, NJ: Lawrence Erlbaum.

Nguyen D and Widrow B (1989) The truck backer-upper: An example of self-learning in neural networks. In: *Proceedings of the International Joint Conference on Neural Networks*, vol. II, pp. 357–363. Washington, DC.

Rumelhart D, Hinton G and Williams R (1986a) Learning representations by backpropagating errors. *Nature* **323**: 533–536.

Rumelhart D, Smolensky P, McClelland J and Hinton G (1986b) Schemata and sequential thought processes in PDP models. *Parallel Distributed Processing*, vol. II, pp. 7–57. Cambridge, MA: MIT Press.

Tsung F-S and Cottrell GW (1995) Phase-space learning. In: Tesauro G, Touretzky D and Leen T (eds) *Advances in Neural Information Processing Systems 7*, pp. 481–488. Cambridge, MA: MIT Press.

Further Reading

Amit DJ (1989) *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge, UK: Cambridge University Press.

Anderson JA (1993) The BSB Model: a simple nonlinear autoassociative neural network. In: Hassoun M (ed.) *Associative Neural Memories*. New York, NY: Oxford University Press.

Anderson JA and Rosenfeld E (eds) (1988) *Neurocomputing:Foundations of Research*. Cambridge, MA: MIT Press.

Cottrell GW and Plunkett K (1994) Acquiring the mapping from meanings to sounds. *Connection Science* **6**: 379–412.

Kawamoto AH, Farrar WT and Kello CT (1994) When two meanings are better than one: modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology, Human Perception and Performance* **20**: 1233–1247.

McClelland JL and Rumelhart DE (1988) *Explorations in Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Plaut DC (1995) Semantic and associative priming in a distributed attractor network. In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 37–42. Hillsdale, NJ: Lawrence Erlbaum.

Plaut DC and Shallice T (1993) Deep dyslexia: a case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**: 377–500.

Rumelhart DE, McClelland JL and the PDP Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vols 1 and 2. Cambridge, MA: MIT Press.

Backpropagation

Intermediate article

Paul Munro, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Backpropagation is a supervised training procedure for feedforward connectionist networks that is widely used by cognitive scientists to model learning phenomena.

LEARNING AND GENERALIZATION IN NEURAL NETWORKS

Initial approaches to developing general-purpose learning machines have generally been restricted to a form of learning known as supervised learning, i.e. abstracting the properties that underlie an input–output relation given a sample of input–output pairs. A standard test of successful learning is the ability of the system to generalize, which is measured by the average correctness of the system's responses to a set of novel stimuli.

Since the response properties of a neural network depend on the weights, or connections between pairs of units, models of learning are generally framed in terms of how the weights change as a function of experience. Donald Hebb (1949) suggested that changes in the biological correlates of these weights, the synapses between neurons, underlie human learning. In the late 1950s and early 1960s, learning rules were developed for networks. These rules were computationally limited, since both the units and the network architectures were simple (Rosenblatt, 1958; Widrow and Hoff, 1960). While the design of more powerful networks was well within the scope of scientific knowledge at that time, there was no known method for training them. (*See* **Hebb Synapses: Modeling of Neuronal Selectivity**; **Hebb, Donald Olding**)

A simple change to the processing function in the individual units led directly to the development of a learning rule for feedforward networks of arbitrarily complex connectivity. Notions from statistics can be applied to derive a learning procedure for networks with complex architectures. This was first done by Paul Werbos (1974), but his results went unnoticed by the scientific community until their almost simultaneous rediscovery by LeCun (1985), Parker (1982), and Rumelhart *et al*. (1986).

SINGLE-LAYER CONNECTIONIST NETWORKS WITH LINEAR THRESHOLD UNITS

The computational power of a connectionist network depends on the computational power of each unit and on the connectivity among the units. Early learning rules applied to the 'linear threshold unit' (LTU) model, which can respond to a stimulus pattern with one of just two possible values (usually these are 0 and 1). Each LTU performs a 'categorization' function on the space of possible stimuli: it divides the set of stimuli into those that generate a response of 1 and those that generate a response of 0. For an LTU, the boundary between these regions is linear. A 'layer' of LTUs performs independent categorization tasks. A category whose stimuli can be separated from stimuli outside the category by a linear boundary, and which is thus computable by an LTU, is called linearly separable (LS) in the stimulus space. Some LS categorization tasks are shown in Figure 1, along with some that are not LS. (*See* **Connectionism**)

LEARNING IN SINGLE-LAYER NETWORKS

Learning by an LTU is a process whereby the weights change in order to improve the placement of the category boundary. A 'supervised learning rule' operates on the parameters of a system (in this

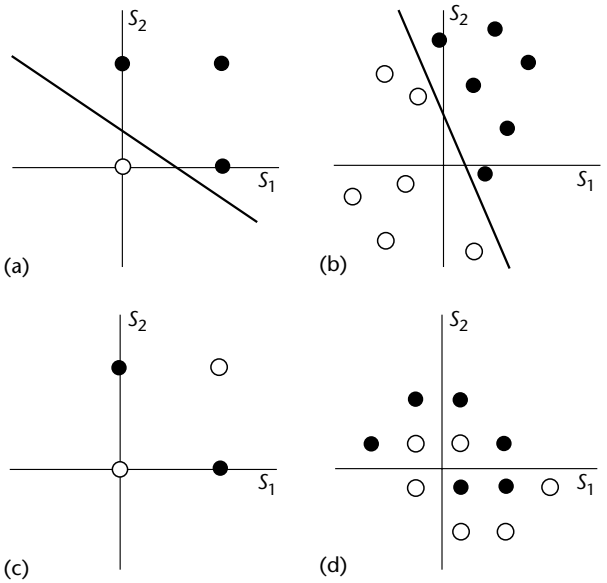


Figure 1. Linear separability. The graphs illustrate four categorization tasks in which patterns are plotted according to the stimulus coordinates (s_1, s_2). Filled dots represent stimuli that are in the category, and open dots represent stimuli not in the category. Categorization boundaries are drawn for the two linearly separable tasks: (a) the 'Boolean OR' task, and (b) a real-valued task. Two tasks that are not linearly separable are also shown: (c) the 'Boolean XOR' task, and (d) a 'double spiral' task.

case, the weights of the network) under the assumption that a set of labeled data (i.e. stimulus points for which the correct categorizations are known) is available. This 'training set' is used to tune, or train, the network weights in the hope that the system will generalize from the training so that it will classify novel stimuli appropriately. This approach resembles standard regression techniques from statistics.

Figure 2 illustrates weight changes of a linear threshold unit and the resulting categorizations on a two-dimensional stimulus space. Note that, in the initial state, some responses of the network are correct and some are incorrect. Eventually, the system finds a classification boundary that solves the given task as well as possible.

MULTI-LAYER NETWORKS

A network built of LTUs can compute a more general class of functions than a single LTU, which is restricted to computing functions that are LS. A typical network structure is the 'multi-layer perceptron' (MLP), originally proposed by Rosenblatt (1958). The MLP architecture first computes the responses of several units, each with different weights (the first layer) to a common stimulus. The pattern of responses is fed as a stimulus to a second layer, and so forth until the final (output) layer. The layers that precede the output layer are known as 'hidden layers'.

An MLP can compute functions that are not LS. Note that each unit in an MLP performs a linear separation on its direct input, but the category it computes on the network input might be more complex. By introducing one or more layers between the network stimulus and the ultimate response, the stimulus pattern is transformed to another 'representation'. A task that is not LS using the representations given at the stimulus level may become LS at a hidden layer. Thus, an MLP can compute a complex categorization task by changing the representation of the task. What is needed is to find a transformation under which the hidden-layer representations are LS. An example is shown in Figure 3.

BACKPROPAGATION OF ERROR

Until the publication of the 'backprop' procedure, there was no technique for training an MLP with more than a single layer. Like standard regression techniques, the derivation of backprop begins with the definition of an 'error measure' E which quantifies how closely the network approximates the given data as a function of its weight parameters.

The well-known 'gradient descent' technique is used to modify each weight, by an amount that is proportional to the derivative of E with respect to that weight. This approach eluded researchers in the 1960s, because of the abrupt shift in the value of the threshold function at the threshold, which renders the required derivatives undefined. The insight that enables the use of gradient descent is to replace the threshold function with a function that is differentiable but retains the important features of the threshold function (Figure 4).

The backpropagation learning rule is derived by applying the gradient descent technique to fit a feedforward network of sigmoid units to a set of data. The rule can be implemented as a process whereby an error value is first computed for each output unit. Subsequently, these errors are 'propagated backwards' through the weights of the network to determine an 'effective error' for all the hidden units in the network.

In its simplest form, the backprop procedure is implemented as follows (compare with the single-layer learning procedure described above):

	S_1	S_2	r	T
A	0	6	1	1
B	2	0	0	1
C	3	2	0	1
D	4	5	1	1
E	1	–4	0	0
F	–1	3	1	0
G	–3	4	1	0
H	–5	–1	0	0

	S_1	S_2	r	T
A	0	6	1	1
B	2	0	0	1
C	3	2	0	1
D	4	5	1	1
E	1	–4	0	0
F	–1	3	1	0
G	**–3**	**4**	**0**	**0**
H	–5	–1	0	0

	S_1	S_2	r	T
A	0	6	1	1
B	2	0	1	1
C	3	2	1	1
D	4	5	1	1
E	1	–4	0	0
F	–1	3	0	0
G	–3	4	0	0
H	–5	–1	0	0

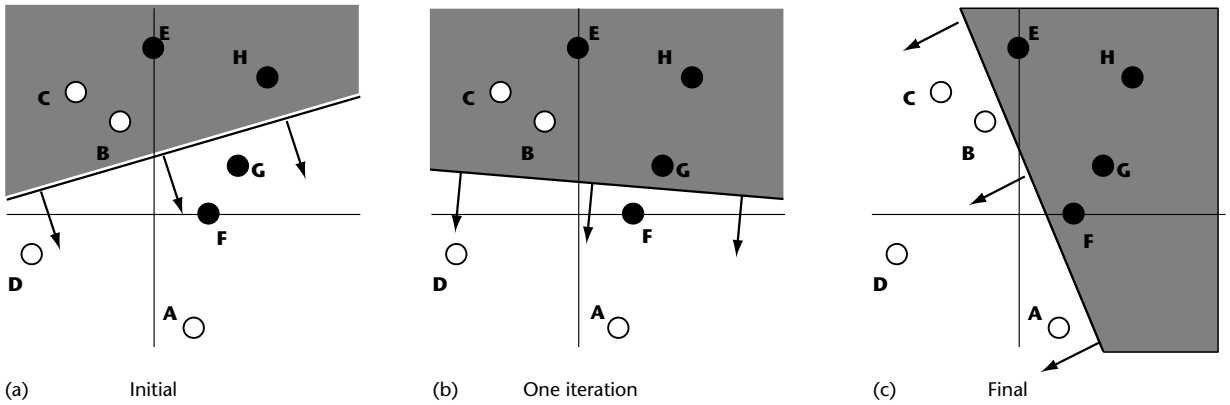
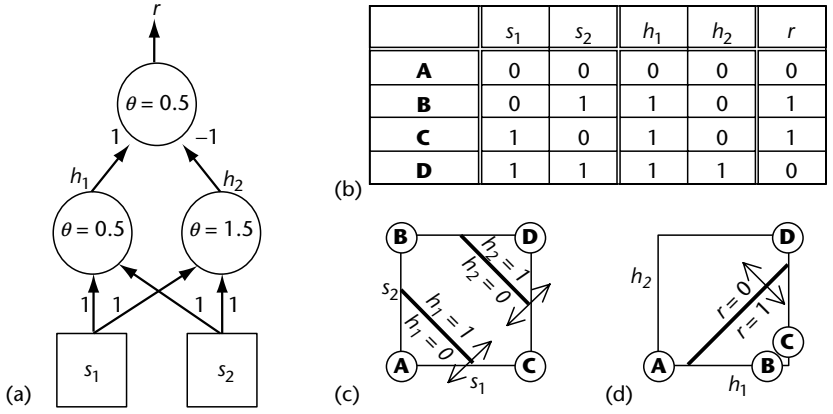


Figure 2. Learning by an LTU. Three stages during the learning process are shown. At each stage a table lists the stimulus values (s_1, s_2), the target (T), and the actual response (r), for the eight elements (A to H) of the training set. Each element is plotted according to its stimulus values, and is either open or filled depending on its target value. The line shows the discrimination boundary at each stage. (a) The initial (random) state. (b) The state after presentation of the first pattern, G. (c) The final state: all patterns are correctly classified.



	s_1	s_2	h_1	h_2	r
A	0	0	0	0	0
B	0	1	1	0	1
C	1	0	1	0	1
D	1	1	1	1	0

Figure 3. A simple MLP. (a) The network shown computes the 'XOR' function of the stimulus, which is not a linearly separable function. The intermediate (hidden) layer has two LTUs, which compute responses using the weights (arrow labels) and thresholds (θ) shown in the diagram. (b) The table shows the responses (h_1, h_2) of the hidden units and the response r of the output unit for each stimulus (s_1, s_2). (c) The linear classification boundaries of the hidden units are shown in 'S-space', in which the stimulus patterns are plotted. (d) The linear classification boundary of the output unit is shown in 'H-space', in which the representations of the patterns at the hidden layer are plotted.

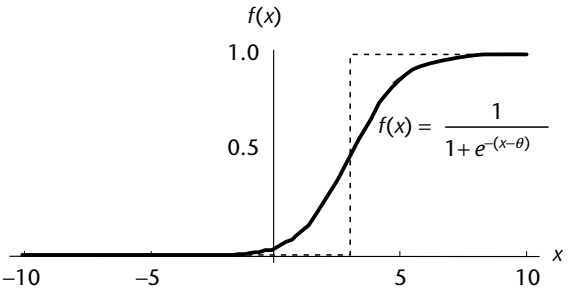


Figure 4. A sigmoid function. Backprop requires units that use differentiable functions. A sigmoid function is differentiable and has many of the important properties of the threshold function. Here, a sigmoid function is plotted (solid line) with a threshold function (dotted line). The value of θ is 3 for both functions. The sigmoid function only approaches its bounds asymptotically. The most common function used for this purpose is the 'logistic' function $f(x) = \frac{1}{1+e^{-x}}$.

1. Initialize the weights to random values.
2. Choose a random data item from the given set (stimulus and categorization value).
3. Compute the activities of the hidden units (first mapping), and from these the activities of the output units (second mapping). This is the 'forward propagation' of neural activity.
4. Compare the output responses with the target values and assign an error value to each output unit.
5. Compute the 'effective error' for each hidden unit as a function of the output unit errors and the hidden–output weights. This is the 'backward propagation' of error.
6. Modify the weights and biases.
7. Test the network on all items in the set of labeled data. If the number of incorrect classifications is acceptable, or if the number of iterations hits the maximum allowed, then stop – otherwise, go back to step 2.

Table 1 gives a more detailed description of the computations.

DEVELOPING INTERNAL REPRESENTATIONS FROM EXPERIENCE

In order for backprop to converge to a state that computes a 'difficult' task (such as one that is not LS), the mapping from the input to the hidden layer must give representations that simplify the task (e.g. rendering the task LS). Not only is convergence dependent on finding an appropriate set of internal representations, but so is the ability to generalize appropriately. Although backprop is not guaranteed to find an appropriate mapping of this kind, it does converge to a solution in many cases. In addition, it should be noted that an 'almost perfect' solution is acceptable for modeling many phenomena.

MODELING COGNITION

Networks have been trained with backprop to simulate cognitive functions ranging from perceptual tasks to high-order processes. These models address a broad range of scientific questions. In many cases, particularly in linguistics, the models have been used as counterexamples to assertions that certain cognitive capabilities must be innate or must require specific types of symbolic manipulation. Generally, backprop is used as an example of how a 'neural-like' system can extract statistical regularities from the environment, so that the performance of the system appears to follow 'rules' without any explicit encoding of those rules. The following examples from linguistics demonstrate the facility of backprop for developing models of cognitive tasks at several levels.

Mapping Text to Speech

In their simulation 'NetTalk', Sejnowski and Rosenberg (1987) trained a network to map English text to the corresponding phonological representation. Unlike some more regular languages, the correct pronunciation of a given letter in English is not always the same. However, it is not completely arbitrary, but depends in large part on the letters in the same neighborhood. Consider, for example, the pronunciations of the letter 'c' in the three nonword strings 'stince', 'stinch', and 'stinck'. The fact that most readers of English would agree on the pronunciation, even though they have never heard the words read aloud, indicates that there are 'rules' for pronunciation rather than arbitrary correspondences between letters and phonemes. NetTalk is trained on a corpus of text, and eventually is able to pronounce not only text from the

Table 1. The backprop algorithm

Algorithmic step	*Formulae*	*Pseudocode*

Hidden unit activities	$h_j = \frac{1}{1+e^{-x}}$
	where $x = b_j^{\text{hid}} + \sum_i w_{ij} s_i$
		`for j := 1 to nhid`
		`arg[j] := hidb[j] ;`
		`for i := 1 to ninput`
		`arg[j] := arg[j] + w[i,j] * s[i];`
		`h[j] := f(arg[j]) ;`
		`end`

Output unit responses	$r_k = \frac{1}{1+e^{-x}}$
	where $x = b_k^{\text{out}} + \sum_j v_{jk} h_j$
		`for k := 1 to noutput`
		`arg[k] := outb[k] ;`
		`for j := 1 to nhid`
		`arg[k] := arg[k] + v[j,k] * h[j] ;`
		`r[k] := f(arg[k]) ;`
		`end`

Output unit errors	$\delta_k^{\text{out}} = (T_k - r_k) r_k (1 - r_k)$	`for k := 1 to noutput`
		`dout[k] :=`
		`(T[k] – r[k]) * r[k] * (1 – r[k]) ;`

Hidden unit errors	$\delta_j^{\text{hid}} = \left[\sum_k v_{jk} \delta_k\right] h_j (1 - h_j)$	`for j := 1 to nhid`
		`dhid[j] := 0 ;`
		`for k := 1 to noutput`
		`dhid[j] := dhid[j] + v[j,k] * d[k] ;`
		`dhid[j] := dhid[j] * h[j] * (1 – h[j]) ;`
		`end`

Weight and bias adjustments	$\Delta w_{ij} = \eta \delta_j s_i$
	$\Delta v_{jk} = \eta \delta_k h_j$
	$\Delta b_j^{\text{hid}} = \eta \delta_j$
	$\Delta b_k^{\text{out}} = \eta \delta_k$
		`for j := 1 to nhid`
		`bhid[j] := bhid[j] + q * dhid[j] ;`
		`for i := 1 to ninput`
		`w[i,j] := w[i,j] + q * dhid[j] * s[i] ;`
		`for k := 1 to noutput`
		`bout[k] := bout[k] + q * dout[k] ;`
		`for j := 1 to nhid`
		`v[j,k] := v[j,k] + q * dout[k] * h[j] ;`

training corpus, but also text unseen during the training process. It extracts the mapping from text to speech without an explicit representation of the rules (see Figure 5).

Generating the Past Tense

The generation of past-tense verbs has been the subject of study by developmental psychologists because children almost universally go through similar stages on the path to adult competence in this task. A network simulation developed by Rumelhart and McClelland (1986), which did not use backprop, required a carefully designed representation in order to learn the task of mapping verbs from their present-tense forms to their past-tense forms. Their simulation successfully generalized from the training examples to novel verbs, both regular (e.g. *jump* and *jumped*) and irregular (e.g. *sing* and *sang*). In addition, it was able to mimic certain developmental stages of language learning in children over the course of its training. With the development of backprop, these results have been replicated in multi-layered networks that develop the requisite representations, rather than having them specified (Plunkett and Marchman, 1991).

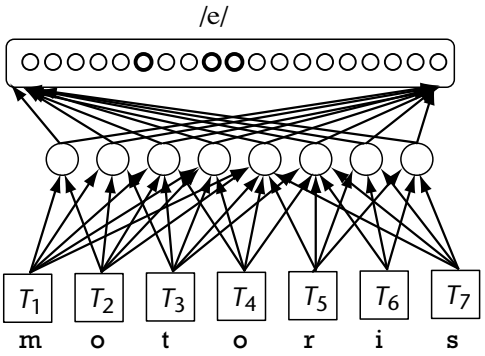


Figure 5. NetTalk. A sliding 7-character window of text (T_1, T_2, T_3, T_4, T_5, T_6, T_7) is presented as input to a network that is trained to generate a phonological representation of the central character (T_4). There are 29 input nodes for each of the 7 character positions (26 letters, space, period, and comma), a single hidden layer, and output units representing phonemic features.

Evolution of Language

Over the course of centuries, languages undergo subtle incremental changes that tend to reinforce regularity (Quirk and Wrenn, 1957) – that is, exceptions to rules are gradually lost, especially for verbs that occur with low frequency. Hare and Elman (1993) offer an explanation and support it by training a succession of networks on the past-tense task using backprop. The first network in their simulation is trained on present–past verb pairs from Old English. Before it achieves perfect performance, a second 'child' network is trained using the first network's computed past tenses. This process proceeds iteratively, each network learning imperfectly from its parent, and so the language evolves. While the process does not precisely follow the evolution of English, it exhibits similar properties with respect to increased regularization and the influence of word frequency.

ENHANCEMENTS TO BACKPROP

As a technique for training feedforward networks for classification tasks in many domains and for developing models of cognitive processes, backprop has been very successful. However, as a gradient-descent procedure, the pure form of the backprop procedure (commonly known as 'vanilla backprop') has some flaws, which can make it an inelegant, or even useless, approach. Some of these are discussed below.

Local Minima

Gradient-descent processes proceed along a path through the state space that reduces the objective function (in this case, the error), like water being driven down a hillside by gravity. The process converges to a state that is a local minimum, from which any direction leads uphill. Thus, the final value of the error is dependent upon the initial state, just as some mountain streams can lead to a high-altitude lake, while others flow to the sea. One imperfect, but simple, remedy is to add a 'momentum' term to the learning rule, which reinforces those components of weight changes that are common across learning trials.

Overfitting Training Data

The input–output items used for training are generally 'noisy'; that is, the output is not perfectly dependent on the input. A system with many adjustable parameters can be trained to fit noisy training data exactly, but only at the expense of its ability to generalize. By monitoring network performance on a set of 'test data' (consisting of items not used for training, but representative of the same sample set), training can be stopped before the system overfits the training data (Figure 6).

Slow Convergence

Gradient-descent processes tend to be slow. Learning of this kind is often criticized as a model of human learning for its inability to learn an item from a single exposure. The slow convergence is exacerbated by the complexity of the input–output relationship and the number of parameters. The speed of convergence is closely related to the value of the learning rate η.

Network Architecture

The determination of the best number of hidden units and their connectivity is a challenge, for which guesswork and trial-and-error are often relied upon. There are two primary approaches to optimizing network architecture. Networks can be 'grown' by starting with a minimal architecture and incrementally adding hidden units (Ash, 1989; Fahlman and Lebiere, 1990; Hanson, 1990) when the generalization error stops decreasing. At some point, the addition of more hidden units no longer benefits the network, or may even be a negative contribution.

Alternatively, one can begin with many more hidden units than are required and then 'prune'

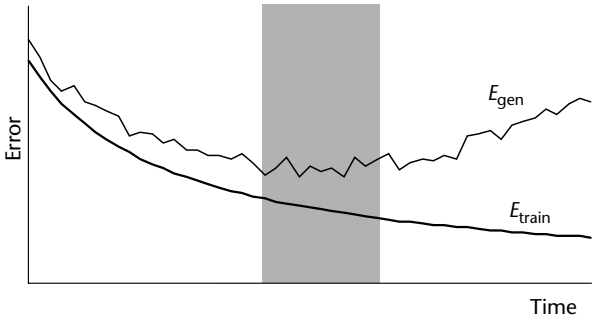


Figure 6. Two error measures. The error measured over the training set, E_{train}, steadily decreases over time. The error with respect to a set sampled from the same population, E_{gen}, also decreases during the first period of learning. Typically, for data sets with noise, E_{gen} eventually begins to steadily increase. Learning should be stopped when E_{gen} is at a minimum (at some point in the shaded region).

units that are deemed to have a low 'relevance' to the network task (Chauvin, 1989; Mozer and Smolensky, 1989; Le Cun *et al*., 1990; Hassibi and Stork, 1991; Demers and Cottrell, 1993). With each removal, the network is retrained and the generalization error is measured. The cycle of pruning and retraining is continued until the generalization performance begins to deteriorate. The objective function (the function minimized by the gradient descent) is typically augmented by an additional term related to the number of active hidden units.

LEARNING TO PERFORM TEMPORAL TASKS

The first implementations of backprop were confined to feedforward networks with static inputs and static outputs. Temporal processes are more naturally accommodated by networks that have recurrence (i.e. they are not feedforward). The fact that many cognitive processes are temporal in nature has led to the development of strategies to enable the application of backprop to temporal tasks. Three main approaches are described below.

Time Delay Neural Networks

In a 'time delay neural network' (TDNN), the input nodes encode consecutive items from a discrete temporal sequence. The sequence is shifted one item at a time, presenting the network with a sliding window on the entire temporal pattern. The NetTalk architecture (Figure 5) is a TDNN that processes a sliding window of text as input and produces a stream of phonemes as output. While a TDNN combines signals from different points in time to interact at the input level, the temporal interaction is limited by the size of the window.

Simple Recurrent Networks

Jordan (1986) introduced the idea of cycling the output back to the input in order to learn sequences. The next step was Elman's (1990) *simple recurrent network* (SRN), which learns to recognize patterns within a sequence by storing 'temporal context' in the hidden layer. With each iteration of the learning procedure, the hidden unit activities from the previous iteration are treated as if they were part of the input to the network (Figure 7(a)). Thus, the hidden layer acts as a memory that can retain information over several time steps. Servan-Schreiber *et al*. (1989) explored grammar learning by an SRN. They showed that the SRN could not only detect errors (with 100% accuracy in a simple grammar), but could generate novel sequences as well.

Backpropagation Through Time

The functionality of a given recurrent network **R** with N nodes can be approximated for T timesteps by a feedforward network **F** that has T layers with N nodes per layer (see Figure 7(b)). In the 'backpropagation through time' procedure (Williams and Zipser, 1989), each layer of **F** represents one timestep of **R**. Thus there are T units in **F** corresponding to each unit in **R**. Each connection from a unit i to a unit j in **R** is replicated by a number of copies in **F**. The implementation of backprop on **F** is subject to the constraint that the replicants of a given weight in **R** have the same value.

UNSUPERVISED LEARNING WITH BACKPROPAGATION

Simply defined, the *autoencoder* is a network trained to compute an identity map; that is, the target pattern is the same as the input. The standard form is a strictly-layered network (i.e. there are only connections between adjacent layers) with a single hidden layer of K units, and an equal number (N) of input and output units. Typically, $K < N$, and since the output layer only has access to the hidden layer, it must reconstruct the input pattern from the reduced (encoded) representation. In order to learn this task successfully, the hidden unit representations must evolve such that each pattern is unique.

One of the most obvious applications of this (assuming $K < N$) is to reduce the dimensionality of data representation. Any data item that can be successfully generated by the network can be stored in a compressed form, simply by presenting it as input to the network and storing the hidden-unit representation. The well-known data compression technique of principal components analysis is mathematically similar, though not exactly the same (Baldi and Hornik, 1989). The original item can be reconstituted at the output by activating the hidden layer with the compressed representation (see Figure 8). With this technique, the data compression is 'lossy': that is, the reconstructed data are not guaranteed to be accurate.

Some applications of auto-encoders trained with back propagation are described below.

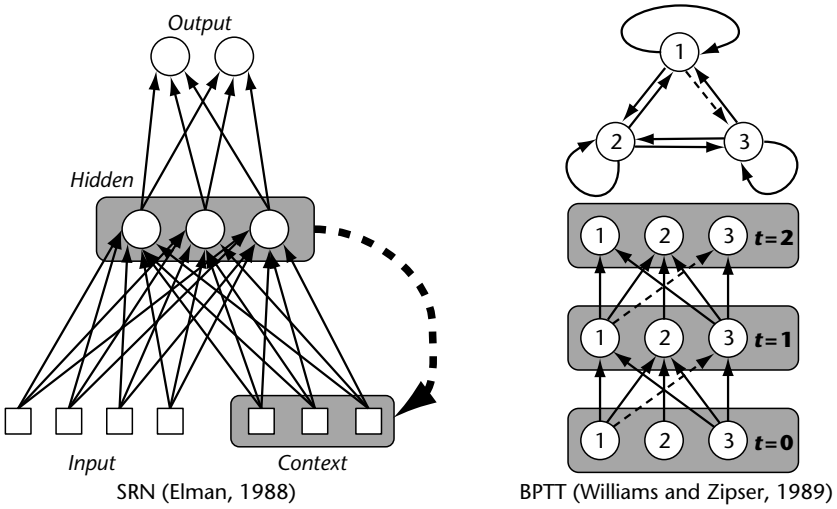


Figure 7. Architectures for temporal tasks. *Left* The simple recurrent network (SRN) maintains a history of the input sequence at the hidden layer by including the previous pattern of hidden unit activity as if it were part of a new input. *Right* A recurrent network (top) is approximated by a three-layer feed-forward architecture (bottom), where every node is replicated at every layer, and each layer corresponds to a different time step.

Image Compression

Cottrell *et al*. (1989) trained an auto-encoder on image data. After training, the network was able to represent the image using only 25% of the space required for the bitmap version. The reconstructed images had a very small deviation from the originals.

Novelty Detection

For some classification tasks, there are not enough data of one class to train a standard classifier (whether connectionist or not). After training an auto-encoder on items from the one class with sufficient data, any test item that shows low reconstruction error is classified as a member of the class used for training. If the reconstruction is poor, the item is classified as a member of the other class (we assume only two classes here). This approach has been used for predicting failure of electric motors using data from normally functioning motors only (Petsche *et al*., 1996).

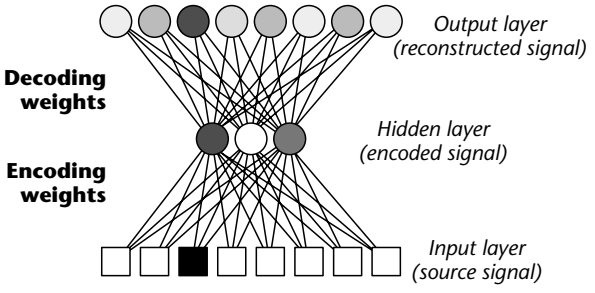


Figure 8. The autoencoder. This network is trained to reconstruct the input pattern at the output layer. Accurate reconstruction depends upon sufficient information in the hidden units, since they supply the only information available to the output units. Thus, the input is encoded by the input-hidden weights. The hidden unit representation is decoded by the hidden-output weights.

Random Access Auto-associative Memory

Pollack (1990) presented a technique by which an auto-encoder could develop a representation for a binary tree. In his scheme, each node in the tree is represented by an n-dimensional pattern of activity. An auto-encoding network with $2n$ input units, n hidden units, and $2n$ output units is used. Such 'random access auto-associative memories' have been applied as models of grammar learning and mental representations of music (Large *et al*., 1995).

BACKPROPAGATION AS A MODEL OF HUMAN LEARNING

Generally, there are two aspects of backprop training that are of potential interest to cognitive scientists: the dynamic process of learning, and the properties of the network after learning. While it is recognized as an important technique for cognitive modeling, the application of backprop to account for cognitive phenomena has been criticized on several grounds. These include: biological implausibility, the nature of the teaching signal,

and interference between learning old and new information.

Biological Plausibility

In part, the appeal of the neural-network approaches in artificial intelligence and cognitive science is their connection with biology. While there is ample neurobiological support for the notion that synaptic modification underlies learning, there is no specific mechanism known that corresponds to the error transmission implied by backprop. Furthermore, backprop has the inherent property of allowing weights to change sign, which would be analogous to an excitatory synapse becoming inhibitory or vice versa. A more biologically plausible procedure that is similar to backprop has been suggested by O'Reilly (1996). (*See* **Long-term Potentiation and Long-term Depression**)

Teaching Signals

In its pure form, backprop requires an explicit teaching signal to every output unit with every pattern presentation. However, a great deal of learning takes place with feedback that is much less specific, or entirely absent. A network with multiple output units can be trained with the minimal feedback of a scalar reward signal by introducing a second network that predicts the reward as a function of the first network's response (Munro, 1987).

Catastrophic Interference

If a network is trained with backprop on a set A of input–output pairs, and then trained on an independent set B, with no further training on the items from A, the performance on A deteriorates quickly. Eventually the items in A are forgotten (although they are on average more easily relearned than completely novel patterns). Partial remedies to the problem include: using two types of weights that change at different speeds (Hinton and Plaut, 1987); various forms of rehearsal (Ratcliff, 1990; Robins, 1995); and enforcing sparser hidden-layer representations (French, 1992; Krushke, 1993).

SUMMARY

Backprop has generated much attention, both inside and outside the cognitive science community. As a tool for cognitive modeling, it is still the best technique for abstracting the statistics of a task into a structure, and studying the internal representations that emerge and their influence on generalization performance. Thus, the acquisition of knowledge by the artificial system can be compared with human learning in several ways and over many modalities.

Aside from any relevance it has to cognition, backprop is also now a standard tool for machine learning, and performs well compared with other techniques. Variants of backprop are used in software in a broad range of application domains, from recognition of handwritten characters, to financial forecasting, to medical diagnosis. Of course, developers of commercial software are not concerned with cognitive and biological plausibility. They are generally more concerned with minimizing error rates than with emulating human patterns of error.

References

Ash T (1989) Dynamic node creation in backpropagation networks. *Connection Science* **1**: 365–375.

Baldi P and Hornik K (1988) Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks* **2**: 53–58.

Chauvin Y (1989) A back-propagation algorithm with optimal use of hidden units. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann.

Cottrell G, Munro P and Zipser D (1989) Image compression by back propagation: an example of extensional programming. In: Sharkey NE (ed.) *Models of Cognition: A Review of Cognitive Science*, vol. I, pp. 208–240. Norwood, NJ: Ablex.

Demers D and Cottrell G (1993) Non-linear dimensionality reduction. In: Hanson SJ, Cowan JD and Giles CL (eds) *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann.

Elman J (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Fahlman S and Lebiere C (1990) The cascade-correlation learning architecture. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 2*. San Mateo, CA: Morgan Kaufmann.

French R (1992) Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science* **4**: 365–377.

Hanson S (1990) Meiosis networks. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 2*. San Mateo, CA: Morgan Kaufmann.

Hare M and Elman JL (1993) From weared to wore: a connectionist account of language change. In: *Proceedings of the 15th Meeting of the Cognitive Science Society*, pp. 265–270. Princeton, NJ: Erlbaum.

Hassibi B and Stork DG (1993) Second order derivatives for network pruning: optimal brain surgeon. In: Hanson SJ, Cowan JD and Giles CL (eds) *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann.

Hebb D (1949) *The Organization of Behavior*. New York, NY: Wiley.

Hinton G and Plaut D (1987) Using fast weights to deblur old memories. In: *Proceedings of the 9th Meeting of the Cognitive Science Society*, pp. 177–186. Princeton, NJ: Erlbaum.

Jordan M (1986) Attractor dynamics and parallelism in a connectionist sequential machine. In: *Proceedings of the 8th Meeting of the Cognitive Science Society*, pp. 531–546. Princeton, NJ: Erlbaum.

Krushke J (1993) Human category learning: implications for backpropagation models. *Connection Science* **5**: 3–36.

Large E, Palmer C and Pollack J (1995) Reduced memory representation for music. *Cognitive Science* **19**: 53–96.

LeCun Y (1985) *Modeles Connexionnistes de l'Apprentissage*. PhD thesis, Université Pierre et Marie Curie, Paris, France.

LeCun Y, Denker J and Solla S (1990) Optimal brain damage. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 2*. San Mateo, CA: Morgan Kaufmann.

Mozer MC and Smolensky P (1989) Skeletonization: a technique for trimming the fat from a network via relevance assessment. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann.

Munro P (1987) Dual backpropagation: a scheme for self-supervised learning. In: *Proceedings of the 9th Meeting of the Cognitive Science Society*. Princeton, NJ: Erlbaum.

O'Reilly R (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation* **8**: 895–939.

Parker DB (1982) *Learning-Logic*. Invention Report S81-64, File 1. Stanford, CA: Office of Technology Licensing, Stanford University.

Petsche T, Marcantonio A, Darken C *et al*. (1996) A neural network autoassociator for induction motor failure prediction. In: Touretzky DS, Mozer MC and Hasselmo ME (eds) *Advances in Neural Information Processing Systems*, vol. VIII, pp. 924–930. Cambridge, MA: MIT Press.

Plunkett K and Marchman V (1991) U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition* **38**: 43–102.

Pollack J (1990) Recursive distributed representations. *Artificial Intelligence* **46**: 77–105.

Quirk R and Wrenn CL (1957) *An Old English Grammar*. London: Methuen.

Ratcliff R (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review* **97**: 285–308.

Robins A (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science* **7**: 123–146.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408.

Rumelhart DE, Hinton GE and Williams RW (1986) Learning internal representations by error propagation. In: (Rumelhart and McClelland, 1986), pp. 318–364.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I 'Foundations'. Cambridge, MA: MIT Press/Bradford.

Sejnowski TJ and Rosenberg CR (1987) Parallel networks that learn to pronounce English text. *Complex Systems* **1**: 145–168.

Servan-Schreiber D, Cleermans A and McClelland J (1989) Learning sequential structure in simple recurrent networks. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann.

Werbos P (1974) *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.

Widrow B and Hoff M (1960) Adaptive switching circuits. In: *1960 IRE WESCON Convention Record*, pp. 96–104. New York, NY: IRE.

Williams R and Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* **1**: 270–280.

Further Reading

Allman W (1990) *Apprentices of Wonder: Inside the Neural Network Revolution*. New York, NY: Bantam.

Anderson J (1995) *Introduction to Neural Networks*. Cambridge, MA: MIT Press.

Anderson J (2000) *Talking Nets: An Oral History of Neural Networks*. Cambridge, MA: MIT Press.

Ballard D (1999) *An Introduction to Natural Computation*. Cambridge, MA: MIT Press.

Bishop C (1996) *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Hebb D (1949) *The Organization of Behavior*. New York, NY: Wiley.

Reed R and Marks R (1999) *Neural Smithing*. Cambridge, MA: MIT Press.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I 'Foundations'. Cambridge, MA: MIT Press/Bradford.

Bayesian and Computational Learning Theory

Intermediate article

David H Wolpert, NASA Ames Research Center, Moffett Field, California, USA

CONTENTS

THE MATHEMATICS OF INDUCTIVE LEARNING

Inductive learning is the process of coming to statistical conclusions based on past experiences. Unlike deduction, with induction one is never perfectly sure of one's conclusion, instead arriving at a (hopefully highly probable) guess. Inductive learning is performed by the human brain continually: almost all of a brain's conclusions, from the 'simplest' ones involved in sensor-motor decisions, to the most 'sophisticated' ones concerning how one should live one's life, are based at least in part on inductive learning. Even science is ultimately inductive in nature, with the 'past experiences' its conclusions are based on being previous experimental data, and its 'conclusions' being theories that are always open to revision.

A lot of work has been directed at implementing inductive learning algorithmically, in computers. 'Adaptive computation', involving neural networks, fuzzy logic, and computational statistics, can be viewed as a set of attempts to do this. The topic of algorithmic induction also looms large in other fields, like artificial intelligence and genetic algorithms. Recently this work has fostered new research on the mathematical underpinnings of inductive learning. A thorough understanding of those would not only result in improvements in our applied computational learning systems; it would also provide us with insight into the scientific method, as well as human cognition.

This article surveys 'Bayesian learning theory' and 'computational learning theory'. These are the two principal mathematical approaches that have been applied to supervised learning, a particularly important branch of inductive learning. The form of supervised learning considered in this article is simplified, the aim being to highlight the distinctions between these two learning theories rather than to present either in its full form.

A mathematical framework that can encapsulate both learning theories is the 'extended Bayesian framework' (EBF) (Wolpert, 1997). A simplified version of it, sufficient for current purposes, can be roughly described as follows. Say we have a finite '*input space*' X and a finite '*output space*' Y, and a set of m input–output pairs $d = \{d_X(i), d_Y(i)\}$. Call d a '*training set*', and assume it was created by repeated noise-free sampling of an $X \to Y$ '*target function*' f. More formally, assume that the '*likelihood*' governing the generation of d from f is $P(d \mid f) = \prod_{i=1}^{m} \pi(d_X(i))\delta(d_Y(i), f(d_X(i)))$, where $\delta(.,.)$ is the Kronecker delta function, which takes the value 1 if its arguments are equal and equals 0 otherwise, and π is the '*sampling distribution*'. $P(f)$ is known as the '*prior distribution*' over targets, and $P(f \mid d)$ is known as the '*posterior distribution*'.

Let h be the $X \to Y$ function our learning algorithm produces in response to d. As far as learning accuracy is concerned, that learning algorithm is described by $P(h \mid d)$: the details of how the algorithm generates h from d are irrelevant. (One of the major reasons why formalisms other than EBF have limited scope is that they do not use $P(h \mid d)$ to describe the learning algorithm; there is no other quantity that can capture all possible learning algorithms.) When discussing multiple learning algorithms – i.e., multiple distributions $P(h \mid d)$ – we will sometimes distinguish the different algorithms with the notation $\gamma_1, \gamma_2, \ldots$. Note that learning algorithms only ever see d, never f (although they often make assumptions concerning f). Accordingly, $P(h \mid d, f) = P(h \mid d)$. Also note that $P(h) = \sum_{d,f} P(h \mid d)P(d \mid f)P(f)$, and in general need not equal the prior $P(f)$ evaluated for $f = h$.

Take s to be the fraction of $d_X(i)$ such that $d_Y(i) = h(d_X(i))$; i.e., s is the learning algorithm's average accuracy on the training set. We use c to

indicate an error value, with its dependence on the other variables indicated by $c(h,f,d)$. In particular, we write c_O for the average (according to π) across all $x \in X$ lying outside the training set of whether h and f agree on x. We call c_O the '*off training set*' (OTS) error; it is a measure of how well our learning algorithm generalizes from the training set. An alternative error function, indicated by c_I, is the 'independent, identically distributed' (IID) error function. It is the same average, but not restricted to $x \notin d_X$, so that a learning algorithm gets some credit simply for memorizing what it's already seen.

Extensions of these definitions to allow for other kinds of error functions – noise in the target, uncertain sampling distributions, different likelihoods, infinite input and output spaces, etc. – are all straightforward, though laborious; see Wolpert (1997). The next section presents some theorems which will help us to compare the Bayesian and computational theories of supervised learning.

A FORMALIZATION OF INDUCTIVE BIAS

We start with the following theorem (Wolpert, 1995), which specifies the expected generalization error after training on some particular training set:

Theorem 1. The value of the conditional expectation $E(c \mid d)$ can be written as a (non-Euclidean) inner product between the distributions $P(h \mid d)$ and $P(f \mid d)$: $E(c \mid d) = \sum_{h,f} c(h,f,d)P(h \mid d)P(f \mid d)$.

(Similar results hold for $E(c \mid m)$, etc.)

Theorem 1 says that how well a learning algorithm $P(h \mid d)$ performs is determined by how 'aligned' it is with the actual posterior, $P(f \mid d)$, where 'alignment' is quantified by the error function. This theorem allows one to ask questions like 'for what set of posteriors is algorithm γ_1 better than algorithm γ_2?' It also means that, unless one can somehow prove from first principles that $P(f \mid d)$ has a certain form, one cannot prove that a particular $P(h \mid d)$ will be aligned with $P(f \mid d)$, and therefore one cannot prove that the learning algorithm generalizes well.

There are a number of ways to formalize this impossibility of establishing the superiority of some particular learning algorithm with a proof from first principles, i.e. with a proof that is not implicitly predicated on a particular posterior. One of them is in the following set of '*no free lunch*' theorems (Wolpert, 1996a):

Theorem 2. Let $E_{\gamma i}(.)$ indicate an expectation value evaluated using learning algorithm i. Then for any two learning algorithms γ_1 and γ_2, independent of the sampling distribution:

1. Uniformly averaged over all f, $E_{\gamma_1}(c_O \mid f,m) - E_{\gamma_2}(c_O \mid f,m) = 0$.
2. Uniformly averaged over all f, $E_{\gamma_1}(c_O \mid f,d) - E_{\gamma_2}(c_O \mid f,d) = 0$ for any training set d.
3. Uniformly averaged over all $P(f)$, $E_{\gamma_1}(c_O \mid m) - E_{\gamma_2}(c_O \mid m) = 0$.
4. Uniformly averaged over all $P(f)$, $E_{\gamma_1}(c_O \mid d) - E_{\gamma_2}(c_O \mid d) = 0$, for any training set d.

According to these results, by any of the measures $E(c_O \mid d)$, $E(c_O \mid m)$, $E(c_O \mid f,d)$, or $E(c_O \mid f,m)$, all algorithms are equivalent, on average. The uniform averaging that goes into these results should be viewed as a calculational tool for comparing algorithms, rather than as an assumption concerning the real world. In particular, the proper way to interpret statement 1 is that, appropriately weighted, there are 'just as many' targets for which algorithm 1 has better $E(c_O \mid f,m)$ as there are for which the reverse is true. Accordingly, unless one can establish *a priori*, before seeing any of the data d, that the f that generated d is one of the ones for which one's favorite algorithm performs better than other algorithms, one has no assurances that that algorithm performs any better than the algorithm of purely random guessing.

This does not mean that one's algorithm must perform no better than random guessing in the real world. Rather it means that, formally, one cannot establish superiority to random guessing without making some assumptions. Note in particular that you cannot use your prior experience – or even the billion years or so of 'prior experiences' of your genome, reflected in the design of your brain – to circumvent this problem, since all that prior experience is, formally, just an extension to the training set d.

As an important example of the foregoing, consider assessing the validity of a hypothesis by using experimental data that were not available when the hypothesis was created. In the form of 'falsifiability', this is one of the primary tools commonly employed in the scientific method. It can be viewed as a crude version of a procedure that is common in applied supervised learning: choose between the two hypothesis functions $h_{\gamma 1}$ and $h_{\gamma 2}$, made by running two generalizers γ_1 and γ_2 on a training set d_1, by examining their accuracies on a distinct 'held out' training set d_2 that was generated from the same target that generated d_1.

Such a procedure for choosing between hypotheses seems almost unimpeachable. Certainly its crude implementation in the scientific method has

resulted in astonishing success. Yet it cannot be justified without making assumptions about the real world. To state this more formally, take any two learning algorithms γ_1 and γ_2, and consider two new algorithms based on them, S and T. S uses an extension of the choosing procedure outlined above, known as '*cross-validation*': given a training set d, S breaks d into two disjoint portions, d_1 and d_2; trains γ_1 and γ_2 on d_1 alone; sees which resultant hypothesis is more accurate on d_2; and then trains the associated learning algorithm on all of d and uses the associated hypothesis. In contrast, T uses anti-cross-validation: It is identical to S except that it chooses the learning algorithm the accuracy of whose associated hypothesis on d_2 was worst. By the 'no free lunch' theorems, we know that T must outperform S as readily as vice versa, regardless of γ_1 and γ_2. It is only when a certain (subtle) relationship holds between $P(f)$ and the γ_1 and γ_2 one is considering that S can be preferable to T (see Theorem 1). When that relationship does not hold, T will outperform S.

This result means in particular that the scientific method must fail as readily as it succeeds, unless there is some *a priori* relation between the learning algorithms it uses (i.e. scientists) and the actual truth. Unfortunately, next to nothing is known formally about that required relation. In this sense, the whole of science – not to mention human cognition – is based on a procedure whose assumptions not only are formally unjustified, but have not even been formally stated.

BAYESIAN LEARNING THEORY

Intuitively, the Bayesian approach to supervised learning can be viewed as an attempt to circumvent the 'no free lunch' theorems by explicitly making an assumption for the posterior. Usually, to do this it first restricts attention to situations in which the likelihood is known (which in the context of this article means there is no 'noise'). It then makes an assumption about the prior distribution, $P(f)$. Next Bayes' theorem is invoked to combine the prior with the likelihood to give us our desired posterior: $P(f \mid d) \propto P(d \mid f)P(f)$, where the proportionality constant is independent of f. (Besides these kind of assumptions concerning the prior, there are other kinds of assumptions which, when combined with the likelihood, fix the posterior (Wolpert, 1993). However, such assumptions have not yet been investigated in any detail.)

Given such a posterior, the value of $E(C \mid d)$ that accompanies any particular learning algorithm $P(h \mid d)$ is determined uniquely (see Theorem 1). In particular, one can solve for the $P(h \mid d)$ that minimizes $E(C \mid d)$. This is known as the '*Bayes-optimal*' learning algorithm. This algorithm is given by the following theorem (which is rather more general than we need):

Theorem 3. Let $c(h,f,d) = \sum_{x \in X} \pi'(x) G(h(x),f(x))$ for some real-valued function $G(.,.)$ and some real-valued $\pi'(.)$ that is nowhere negative (and may or may not equal the distribution $\pi(.)$ arising in $P(d \mid f)$). Then the Bayes-optimal $P(h \mid d)$ always guesses the same function h^* for the same d:

$$h^* = \{x \in X \rightarrow \arg\min_{y \in Y} \Omega(x,y)\}, \text{where}$$
$$\Omega(x,y) \equiv \sum_f G(y,f(x)) P(f \mid d).$$

(The function $\pi'(.)$ is allowed to vary with d, as it does in OTS error.)

$\Omega(x,y)$ is the contribution to the posterior expected error that arises if the learning algorithm outputs (an h having) the value y at point x. So intuitively, Theorem 3 says that for any x, one should choose the $y \in Y$ that minimizes the average 'distance' from y to $f(x)$, where the average is over all $f(.)$, according to the distribution $P(f \mid d)$, and 'distance' is measured by $G(.,.)$. Note that this result holds regardless of the form of $P(f)$, and regardless of what (if any) noise process is present: all such considerations are taken care of automatically, in the $P(f \mid d)$ term. Note also that h^* might be an f with zero-valued posterior: in the Bayesian framework, the output h of the learning algorithm does not really constitute a 'guess for the f which generated the data'.

This is all there is to the Bayesian framework, as far as foundational issues are concerned (Berger, 1985; Loredo, 1990; Buntine and Weigend, 1991; Wolpert, 1995). Everything else in the literature concerning the framework involves either philosophical or calculational issues. The philosophical issues usually concern what $P(f)$ 'means' (Wolpert, 1993). In particular, some Bayesians do not view the $P(f)$ they use to derive their learning algorithm as an assumption for the actual $P(f)$, which may or may not correspond to reality. Rather, in general they interpret the probability of an event as one's 'personal degree of belief' in that event, and therefore in particular they interpret $P(f)$ that way. According to this view, probability theory is simply a calculus for forcing consistency in one's use of probability to manipulate one's subjective beliefs. Accordingly, no matter how absurd a Bayesian's prior, under this interpretation practitioners of non-Bayesian approaches to supervised learning are by definition always going to perform worse

than that Bayesian (since the Bayesian determines $P(f)$ and therefore $P(f \mid d)$, and accordingly guesses in an optimal manner – see Theorem 1).

Unfortunately, there are algorithms that cannot be cast as the Bayes-optimal algorithm for some implicit prior and likelihood (Wolpert, 1996b). Accordingly, even if one accepts the 'fundamentalist' Bayesian's view of what $P(f)$ 'means', the rigidity of the framework makes it ill-suited to broad analysis of algorithms. More generally, often our prior knowledge does not concern targets directly, but rather concerns the relative performances of various (possibly non-Bayesian) algorithms, or the efficacy of a scheme (like cross-validation) for choosing among those algorithms. The conventional Bayesian framework provides no way to exploit that prior knowledge. In general, we need to introduce the random variable h for such an analysis – which is what is done in EBF. (See, however, Wolpert (1993) for a discussion of how one can sometimes employ Bayes' theorem to exploit such knowledge directly.)

Some of the calculational issues in the Bayesian framework involve evaluating the Bayes-optimal algorithm, given knowledge of the posterior $P(f \mid d)$. The problem is that using the Bayes-optimal algorithm requires evaluating (and then minimizing) the sum giving $\Omega(x, y)$. Since this can be difficult, people often settle for approximations to finding $\arg\min_{y \in Y} \Omega(x, y)$. For example, the 'maximum *a posteriori*' (MAP) estimator is $h(x) = [\arg\max_f P(f \mid d)](x)$. Evaluating it involves finding a peak of a surface (namely $P(f \mid d)$) rather than performing a sum, and is often simpler.

Even if one is willing to use a MAP estimator, there might still be difficulties in evaluating the surface $P(f \mid d)$. For example, if we have a 'hierarchical' likelihood, then $P(d \mid f) = \sum_\lambda P(d \mid \lambda, f)\, P(\lambda)$, where λ is a 'hyperparameter'. (An example is where we know that the data are generated via a particular kind of noise process, but don't know the noise level in advance – that noise level is a hyperparameter.) Sums being difficult, often we can't even evaluate such a hierarchical likelihood (and therefore can't evaluate the associated posterior, $P(f \mid d)$). Instead, often one makes an ad hoc estimate of the hyperparameter, λ', and replaces $P(d \mid f)$ with $P(d \mid \lambda', f)$. (See the discussion of empirical Bayes and ML-II in Wolpert (1995).)

COMPUTATIONAL LEARNING THEORY

The computational learning framework takes a number of forms, the principal ones being the statistical physics, PAC and VC (uniform convergence) approaches (Baum and Haussler, 1989; Vapnik, 1982; Wolpert, 1995). All three can be cast as bounds concerning a probability distribution that involves IID error, and that is conditioned on f (in contrast to the Bayesian framework, in which f is not fixed). In their most common forms they all have m rather than d fixed in their distribution of interest (again, in contrast to the Bayesian framework). This means that they do not address the question of what the likely outcome is for the training set at hand. Rather, they address the question of what the likely outcome would be if one had different training sets from the actual d. Such varying of quantities that are in fact fixed and known has been criticized by Bayesian practitioners on formal grounds, as violating any possible self-consistent principles for induction. (See Wolpert (1993), and the discussion of the 'honesty principles' in Wolpert (1995), for an overview of the conflict between the two learning theories.)

For purposes of illustration, we will focus on (a pared-down version of) the VC framework. Start with the following simple result, which concerns the '*confidence interval*' relating c and s, for the case where $H^\sim$, the h-space support of a learning algorithm's $P(h)$, consists of a single h (Wolpert, 1995):

Theorem 4. Assume that there is an h' such that $P(h \mid d) = \delta(h - h')$ for all d. Then:

$$P(c_{\mathrm{I}} > s + \varepsilon \mid f, m) < 2e^{-m\varepsilon^2}$$

(Recall that s is the empirical misclassification rate.) Note that this bound is independent of f, and therefore of the prior $P(f)$.

If $H^\sim$ instead consists of more than one h, the bound in Theorem 4 still applies if one multiplies the right-hand side by $\mid H^\sim \mid$, the number of functions in $H^\sim$. The major insight behind the '*uniform convergence*' framework was how to derive even tighter bounds by characterizing $P(h \mid d)$ in terms of its VC dimension (Baum and Haussler, 1989; Vapnik, 1982; Wolpert, 1995). (It is important to distinguish between this use of the VC dimension and its use in other contexts, as a characterization of $P(f)$.) For $Y = \{0, 1\}$ and our error function, the VC dimension is given by the smallest m such that, for any d_X of size m, all of whose elements are distinct, there is a d_Y for which no h in $H^\sim$ goes through d. (The VC dimension is this smallest number minus one.)

Common to all such extensions of Theorem 4 is a rough equivalence (as far as the likely values of c are concerned) between: (1) lowering s; (2) lowering the expressive power of $P(h \mid d)$ (i.e., shrinking its

VC dimension, or shrinking $|H^{\sim}|$); and (3) raising m. Important as these extensions of Theorem 4 are, to understand the foundational issues underpinning the uniform convergence framework it makes sense to restrict attention to the scenario in which there is a single h in $H^{\sim}$.

In general, since we can measure s and want to know c_{I} (rather than the other way around), a bound on something like $P(c_{\mathrm{I}} > k \,|\, s, m)$, perhaps with $k \equiv s + \varepsilon$, would provide some useful information concerning generalization error. With such a bound, we could say that if we observe the values of m and s to be M and S, then with high probability c_{I} is lower than $U(M, S)$ for some appropriate function $U(.,.)$. However, since both f and (for our learning algorithm) h are fixed in the probability distribution in Theorem 4, c_{I} is also fixed there, for IID error. (By contrast, in the Bayesian framework, c_{I} is only probabilistically determined.) In fact, in Theorem 4 what is varying is d_X (or more generally, when there is noise, d). So Theorem 4 does not directly give us the probability that c_{I} lies in a certain region, given the training set at hand. Rather, it gives the probability of a d_X (generated via experiments other than ours) such that the difference between the fixed c_{I} and (the function of d_X) s lies in a certain region.

It might seem that Theorem 4 could be modified to provide a bound of the type we seek. After all, since the value of c_{I} is fixed in Theorem 4, that theorem can be written as a bound on the 'inverse' of $P(c_{\mathrm{I}} > k \,|\, s, m)$, $P(s < \kappa \,|\, c_{\mathrm{I}}, m)$, where $\kappa \equiv c - \varepsilon$. How does $P(s \,|\, c_{\mathrm{I}}, m)$ relate to what we wish to know, $P(c_{\mathrm{I}} \,|\, s, m)$? The answer is given by Bayes' theorem: $P(c_{\mathrm{I}} \,|\, s, m) = P(s \,|\, c_{\mathrm{I}}, m) P(c_{\mathrm{I}} \,|\, m) / P(s \,|\, m)$.

Unfortunately, this result has the usual problem associated with Bayesian results: it is prior-dependent. Nor does it somehow turn out that that prior has little effect. Depending on $P(c_{\mathrm{I}})$, $P(c_{\mathrm{I}} > s + \varepsilon \,|\, s, m)$ can differ markedly from the bound on $P(s < c_{\mathrm{I}} - \varepsilon \,|\, m, c_{\mathrm{I}})$ given in Theorem 4. Even if, given a truth c_{I}, the probability of an s that differs substantially from the truth is small, it does not follow that given an s, the probability of a truth that differs substantially from that s is small.

To illustrate this point, suppose we have two random variables, A and B, which can both take on the values 'low' and 'high'. Suppose that the joint probability distribution is proportional to: $P(A = \text{high}, B = \text{high}) = 100$, $P(A = \text{high}, B = \text{low}) = 2$, $P(A = \text{low}, B = \text{high}) = 1$, $P(A = \text{low}, B = \text{low}) = 1$. Then the probability that A and B differ is quite small (3/104); we have a tight confidence interval relating them, just as in Theorem 4. Nonetheless, $P(A = \text{high} \mid B = \text{low})$ is 2/3: despite the tight confidence interval, if we observe $B = \text{low}$, we cannot infer that A is low as well. Replace 'A' with 'c_{I}', and 'B' with 's', and we see that results like Theorem 4 do not imply that having observed a low s, one can conclude that one has a low c_{I}.

A more concrete example of this effect in the context of supervised learning is the following result, established in Wolpert (1995):

Theorem 5. Let $\pi(x)$ be flat over all x and $P(f)$ flat over all f. For the noise-free IID likelihood, and the learning algorithm of Theorem 4:

$$P(c_{\mathrm{I}} \,|s, m) = \left[\binom{m}{sm} c_{\mathrm{I}}^{sm} (1 - c_{\mathrm{I}})^{m - sm}\right] \times \left[\binom{n}{nc_{\mathrm{I}}} (|Y| - 1)^{nc_{\mathrm{I}}}\right]$$

where $|Y|$ is the number of values in Y.

Theorem 5 can be viewed as a sort of compromise between the likelihood-driven 'something for nothing' results of the VC framework, and the 'no free lunch' theorems. The first term in the product has no c_{I}-dependence. The second and third terms together reach a peak when $c_{\mathrm{I}} = s$; they 'push' the true misclassification rate towards the empirical misclassification rate, and would disappear if we were using OTS error. These two terms are closely related to the likelihood-driven VC bounds. However, the last two terms, taken together, form a function of c_{I} whose mean is $1/|Y|$. They reflect the fact that any f is allowed with an equal prior probability, and are closely related to the 'no free lunch' theorems (despite the fact that IID error is being used). In this sense, our result for $P(c \,|\, s, m)$ is just a product of a 'no free lunch' term with a VC-type term.

In response to the formal admonitions of these theorems, one is tempted to make the following intuitive reply: 'Say we have a function f and a given hypothesis function h' that have no *a priori* relation with one another. A sample point is drawn from f, and it is found that h' correctly predicts that point. Then another sample point is drawn, with the same result. Based on such a sequence of points, you guess that h' will correctly predict the next sample point. And lo and behold, it does. You keep extending the original sequence this way, always getting the same result that h' makes the correct prediction (since s is small, and the full training set d consists of the extended sequence, not the original one). In other words, the generalizer given by the rule "always guess h'" has excellent cross-validation error. In this situation, wouldn't you believe that it is unlikely for h' and f to disagree on future sample points, regardless of the "no free lunch" theorems?'

To disentangle the implicit assumptions behind this argument, consider it again in the case where h' is some extremely complex function that was formed by a purely random process. The claim in the intuitive argument is that h' was fixed independently of any determination of f, d, or anything else, and is not biased in any way towards f. Then, so goes the claim, f was sampled to generate d (the sequence of points), and it just so happened that f and h' agree on d. According to the intuitive argument presented above, we should conclude in such a case that h' and f would agree on points not yet sampled. Yet in such a situation our first suspicion might instead be that the claims that were made are wrong, that cheating has taken place and that h' is actually based on prior knowledge concerning f. After all, how else could the 'purely random' h' agree with f? How else could there be agreement when h' was supposedly fixed without any information concerning d, and therefore without any coupling with f?

If, however, we are assured that no cheating is going on, then 'intuition' might very well lead one to say that the agreements between f and h' must be simple coincidence. They have to be, since, by hypothesis, there is nothing that could possibly connect h' and f. So intuition need not proclaim that the agreements on the data set mean that f and h' will agree on future samples. Moreover, if cheating did occur, then to formulate the problem correctly, we have to know about the *a priori* connection between f and h' in order to properly analyze the situation. This results in a (prior-dependent) distribution different from the one investigated in the uniform convergence framework. (In the real world, of the two alternatives of coincidence and cheating, the reason for low s is almost always 'cheating'. Almost always one uses prior knowledge of some sort to guide the learning, rather than generate hypotheses purely at random.)

CONCLUSION

In all forms of reasoning that do not proceed by strict logical deduction, some kind of statistical algorithm must be employed. One of the major types of such reasoning is 'supervised learning'. In this type of reasoning one is provided with a training set of input–output pairs, and must make a guess for the entire input–output function in such a way as to minimize the error between that guess and the actual function that generated the data.

Apart from conventional sampling theory statistics, there are two principal mathematical approaches to supervised learning: the Bayesian framework and the computational learning framework. We have examined the foundations of these two approaches, especially in light of the 'no free lunch' theorems which limit what *a priori* formal assurances one can have concerning a learning algorithm without making assumptions concerning the real world.

In the Bayesian framework the assumptions concerning the real world arise explicitly, as 'prior probabilities' to be used to calculate the optimal guess. Unfortunately, the fact that the formalism concentrates solely on this assumption-driven calculation prevents it from allowing easy and broad investigation of what happens when those underlying assumptions are incorrect. In particular, Bayesian analysis cannot be used to analyze most learning algorithms that do not make their assumptions explicit; its scope is limited by construction. In particular, this restricts the framework's ability to analyze perhaps the most common algorithm in science, cross-validation.

In contrast, the simplest version of the computational learning framework makes no explicit assumptions about the nature of one's learning algorithm, or about the priors. In this sense it is universally applicable. Whereas the Bayesian framework skirts the 'no free lunch' theorems by forcing the underlying assumptions concerning the problem domain to be explicit, the strategy of the computational learning framework is to instead focus on the counterfactual scenario in which one's data are not fixed (to whatever the data set currently in front of you happens to be), but are averaged over. Unfortunately, the resulting bounds on learning error cannot be modified to concern some particular scenario a learning practitioner is confronted with; they are by their nature concerned with an average over multiple scenarios, when only one actually exists.

References

Baum E and Haussler D (1989) What size net gives valid generalization? *Neural Computation* **1**: 151–160.

Berger J (1985) *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag.

Buntine W and Weigend A (1991) Bayesian back-propagation. *Complex Systems* **5**: 603–643.

Loredo T (1990) From Laplace to Supernova 1987a: Bayesian inference in astrophysics. In: Fougere P (ed.) *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer.

Vapnik V (1982) *Estimation of Dependences Based on Empirical Data*. New York, NY: Springer-Verlag.

Wolpert DH (1993) Reconciling Bayesian and non-Bayesian analysis. In: Heidbreder G (ed.) *Maximum Entropy and Bayesian Methods 1993*. Dordrecht: Kluwer.

Wolpert DH (1995) The Relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In: Wolpert DH (ed.) *The Mathematics of Generalization*, pp. 117–214. Reading, MA: Addison-Wesley.

Wolpert DH (1996a) The lack of a priori distinctions between learning algorithms. *Neural Computation* **8**: 1341–1390.

Wolpert DH (1996b) The bootstrap is inconsistent with probability theory. In: Hanson K and Silver R (eds) *Maximum Entropy and Bayesian Methods 1995*. Dordrecht: Kluwer.

Wolpert DH (1997) On bias plus variance. *Neural Computation* **9**: 1211–1243.

Bayesian Belief Networks

Intermediate article

Brendan J Frey, University of Toronto, Toronto, Canada

CONTENTS

A "Bayesian belief network" is a directed acyclic graph that specifies how stochastic variables in a complex system interact. The joint distribution is equal to the product of a set of conditional distributions. There is one conditional distribution for each node of the graph and each conditional distribution is conditioned on the parents of the corresponding node. Unlike Markov random fields, Bayesian networks do not require a partition function.

REASONING UNDER UNCERTAINTY

An intelligent agent that makes decisions in a realistic environment should take into account the uncertainties in the environment and the uncertainties introduced by incomplete knowledge of the environment. Also, a mathematical description of a physical system for inference should account for the uncertainties in physical systems.

Probability theory provides a way to account for uncertainty. A system is described by a set of random variables, and an instantiation of the variables is called a *configuration*. A numerical probability between 0 and 1 is associated with each configuration and this number corresponds to the relative frequency with which the configuration occurs, or possibly, in the Bayesian view, the chance that the configuration *will* occur. The sum over the probabilities of all configurations must be 1. If we are interested in only a subset of the variables, we can derive the probability of each sub-configuration by summing the probabilities of all configurations that have matching sub-configurations. In this way, a system can be viewed as being 'consistent' with a larger system. So, when building or inferring a system, we need not include all variables in the universe, but can instead include only a smaller, more tractable, subset.

For example, we may use $P(T=1)$ to represent the probability of the event that there is a tiger in the field of view of an intelligent agent, and $P(T=0)$ to represent the probability that there is not a tiger. In this case, $P(T=1)$ is equal to the sum over the probabilities corresponding to all configurations of the universe for which there is a tiger in the field of view of the intelligent agent.

It is often convenient to express probabilities in functional form. For example, the probabilities that $T=1$ and $T=0$ can be written $P(T)$. If we set $T=1$, then $P(T)$ is the probability that $T=1$. We refer to $P(T)$ as a *probability distribution*.

We use the conditional probability distribution

$$P(T|V=v) \tag{1}$$

to represent the probabilities that there is and is not a tiger, given that the random variable V representing the agent's visual input has the value v. If $P(T=1|V=v) \gg P(T=0|V=v)$, there is very probably a tiger in the scene and the agent ought to seriously consider running away.

It is natural to represent the random variable for the agent's visual input by a continuous vector (e.g., a vector of real-valued pixel intensities). For brevity, we will assume that all variables are discrete. In the case of continuous variables, probabilities can be replaced with probability density functions, and summations can be replaced by integrals. Many computations involving seemingly continuous quantities are in fact discrete (e.g., floating point computations on a computer), so not too much is lost by assuming the variables are discrete.

Suppose v is a visual scene containing black and orange stripes, and the probability of a scene containing black and orange stripes is greater if there is a tiger than if there is no tiger:

$$P(V=v|T=1) > P(V=v|T=0) \tag{2}$$

After observing $V=v$, it is tempting to conclude that there is probably a tiger in the scene, but this may not be so.

The probability that there is a tiger in the scene is given by the posterior probability distribution

$P(T|V=v)$. The posterior distribution can be computed from the values of $P(V=v|T=1)$ and $P(V=v|T=0)$ using Bayes' rule:

$$P(T|V=v) = \frac{P(V=v|T)P(T)}{\sum_{t=0,1} P(V=v|T=t)P(T=t)} \tag{3}$$

In this expression, $P(T|V=v)$ is called the *posterior distribution*, $P(T)$ is called the *prior distribution*, and $P(V=v|T)$ is called the *likelihood function*. Notice that all three expressions are functions of the unknown random variable T.

For example, on a jungle tour in India, we may have $P(T=1)=0.2$ and $P(T=0)=0.8$. However in Canada, we may have $P(T=1)=0.001$ and $P(T=0)=0.999$. So, for the scene with orange and black stripes, even though $P(V=v|T=1) > P(V=v|T=0)$, it may turn out that $P(T=1|V=v) < P(T=0|V=v)$, depending on the prior distribution over T.

MAKING DECISIONS UNDER UNCERTAINTY

Even though $P(T=1|V=v) < P(T=0|V=v)$, it may be a good idea to leap aside if orange and black stripes appear in the visual scene.

For every configuration of the unknown random variable ($T=1$ and $T=0$), we can specify a utility (benefit, negative cost) for every action, say 'Leap' and 'NoLeap'. Leaping aside when there is a tiger may mean survival, whereas not leaping aside when there is a tiger may mean death:

$$U^{\text{Leap}}(T=1) >>> U^{\text{NoLeap}}(T=1) \tag{4}$$

Leaping aside when there is no tiger will waste some energy, so

$$U^{\text{Leap}}(T=0) < U^{\text{NoLeap}}(T=0) \tag{5}$$

However, notice that there is much less difference in this inequality than in the previous inequality.

The agent can maximize its expected utility by choosing the decision that has highest expected utility. For Leap and NoLeap, we have

$$EU^{\text{Leap}} = \sum_{T=0,1} U^{\text{Leap}}(T)P(T|V=v) \tag{6}$$

$$EU^{\text{NoLeap}} = \sum_{T=0,1} U^{\text{NoLeap}}(T)P(T|V=v) \tag{7}$$

So, even in Canada, the second term in each formula may dominate, in which case the agent should leap aside if orange and black stripes appear in the visual scene.

BAYESIAN NETWORKS

Instead of a real tiger, orange and black stripes could be caused, say, by a child's stuffed toy tiger. Further, in a toy store, the cause is more likely to be a stuffed toy. In a zoo, the cause is more likely to be a real tiger. These random variables influence each other in a structured way, and Bayesian networks provide a graphical description of this structure.

A *Bayesian network* (Pearl, 1988) is a graphical description of the probability distribution on a system of random variables. It is a directed acyclic graph (DAG) on vertices corresponding to the random variables, along with a specification of the conditional distribution for each variable given its parents in the graph. In the context of directed graphs, 'acyclic' means there aren't any cycles *when edge directions are followed*. The probability distribution on the system of random variables is equal to the product of all of the conditional distributions. If the conditional distributions are not specified, the Bayesian network refers to the set of all probability distributions that can be derived by choosing numbers for the conditional distributions in the network. Further, all distributions described by the network satisfy certain conditional independence properties that can be determined directly from the graph.

Suppose $S=1$ indicates there is a stuffed toy tiger in the scene, whereas $S=0$ indicates there is not a stuffed toy tiger in the scene. The presence of orange stripes in the scene can be caused by either a tiger or a stuffed toy tiger. Figure 1(a) shows one possible Bayesian network that can be used to describe the relationship between the random variables T, S, and V. (Pearl uses an example where a burglar alarm is tripped by either a burglar or an earthquake.) This network describes the set of probability distributions that satisfy

$$P(T,S,V) = P(T)P(S)P(V|T,S) \tag{8}$$

the product of the conditional distributions for each child given its parents. Note that since T does not have parents, its 'conditional distribution' is written $P(T)$.

As described in Pearl (1988), the network can be examined to determine that all distributions described by the network have the property that T and S are independent random variables. This

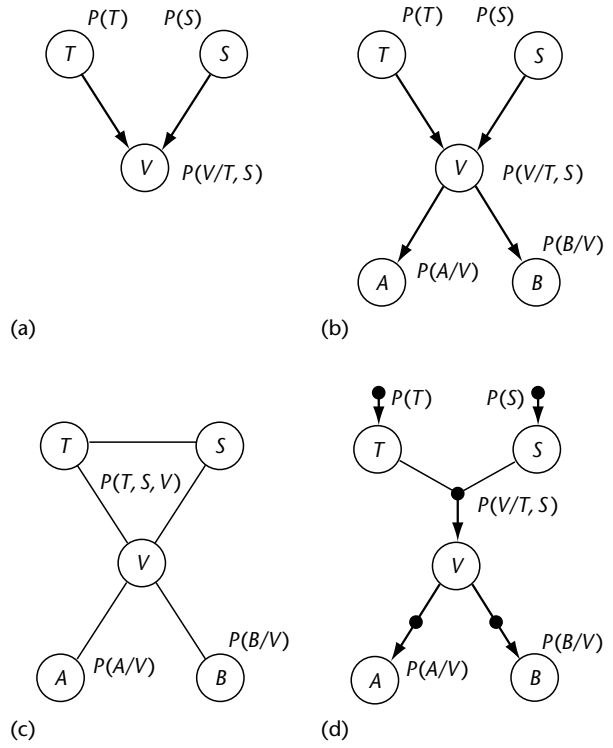


Figure 1. (a) A Bayesian network describing the probability distribution $P(T, S, V) = P(T)P(S)P(V|T, S)$ for the binary indicator variables tiger T, stuffed toy S, and a variable V that indicates the presence of orange stripes in the visual scene. (b) The network in (a) is extended to include variables that indicate whether Alice (A) and Bob (B) leap aside when orange stripes are present in the same scene. The joint distribution is $P(T, S, V, A, B) = P(T)P(S)P(V|T, S)\ P(A|V)P(B|V)$. (c) A Markov network, and (d) a factor graph that describe the same joint distribution.

example is simple enough that we can show this is true by summing over V to obtain the distribution $P(T, S)$:

$$P(T,S) = \sum_V P(T,S,V) = \sum_V P(T)P(S)P(V|T,S)$$
$$= P(T)P(S) \qquad (9)$$

It follows that $P(T|S) = P(T)$ and $P(S|T) = P(S)$, so T and S are independent. For example, the distribution over T is the same, whether or not S is known.

A particular distribution $P(T, S, V)$ is determined by specifying the numerical values of the conditional probability tables; e.g., $P(V = v|T = t, S = s)$ for all values of v, t, and s. Suppose $V = 1$ indicates the presence of orange stripes in the visual scene and $V = 0$ indicates that orange stripes are not present. In Canada, where toy tigers are more abundant than real tigers, we may have the following conditional probability tables:

$P(T)$	
$T = 0$	$T = 1$
0.999	0.001

$P(S)$	
$S = 0$	$S = 1$
0.999	0.001

		$P(V\|T,S)$	
T	S	$V = 0$	$V = 1$
0	0	0.999	0.01
0	1	0.5	0.5
1	0	0.1	0.9
1	1	0.05	0.95

Now, suppose the agent is accompanied by two other agents, Alice and Bob, and that in response to the presence of orange stripes in the scene, Alice and Bob independently choose to leap aside. $A = 1$ indicates that Alice has leapt aside, whereas $A = 0$ indicates that Alice has not leapt aside. Similarly, B indicates Bob's behavior. Figure 1(b) shows a Bayesian network that includes the behavior of Alice and Bob. The conditional probability table for $P(A|V)$ gives the probability that Alice leaps aside given the presence or absence of orange stripes in the scene.

This Bayesian network indicates that the joint distribution factorizes as follows: $P(T, S, V, A, B) = P(T)P(S)P(V|T, S)P(A|V)P(B|V)$. Also, various conditional independencies can be determined by studying the graph, as described in Pearl (1988). For example, Alice's and Bob's behaviors, A and B, are generally not independent. Their behavior is a consequence of a common cause, V. However, Alice's and Bob's behaviors are independent given the visual scene, V.

MARKOV RANDOM FIELDS

Like Bayesian networks, *Markov networks* (Markov random fields) provide a graphical description of the structure of the joint distribution. Unlike Bayesian networks, they are undirected.

A Markov network (Kinderman and Snell, 1980) is an undirected graph on vertices corresponding to the variables, along with a specification of potential functions defined on the variables in maximal *cliques*. A clique is a set of variables that are completely connected. A maximal clique is a clique that cannot be expanded to include an additional variable, without violating the condition that the variables be completely connected. Each potential function is a nonnegative function of the appropriate set of variables. Assuming all potentials are strictly positive, the probability distribution on the system of random variables is equal to the product of all of the clique potentials, multiplied by a normalizing constant.

The conditional independencies indicated by a Markov network are quite different from those indicated by a Bayesian network. In a Markov network, given the neighbors of a variable (the Markov blanket), the variable is independent of all other variables in the network. For a given set of variables, it is possible that there is a Bayesian network that can represent conditional independencies that cannot be represented by a Markov random field. The converse is also true.

For example, the Markov network for the above example is shown in Figure 1(c). The factor $P(V|T,S)$ requires that at least one maximal clique contain V, T, and S. Consequently, by examining the network without reference to the potentials, it is not possible to determine that T and S are independent.

FACTOR GRAPHS

Factor graphs subsume Bayesian networks and Markov networks, in the sense that there is a unique Bayesian network (or Markov network) for every factor graph. However, for a given Bayesian network (or Markov network), there may be multiple factor graphs. Since the Bayesian networks or Markov networks corresponding to the multiple factor graphs are the same, the different factor graphs do not indicate different conditional independencies. However, they can indicate more detailed factorizations of the joint distribution. Also, factor graphs have an extra set of nodes that identify factorization sites, and, in the algorithm described in the next section, computation sites for message-passing algorithms.

A *factor graph* (Kschischang *et al*., 2001) is a bipartite graph on *function vertices* and *variable vertices*. For each function vertex, a *local function* is specified, which is a function of the variables connected to the function vertex. 'Bipartite' means that each edge connects a variable vertex and a function vertex. The local functions may correspond to the conditional distributions in the Bayesian network, the clique potentials in the Markov network, or something else. The probability distribution on the system of random variables is equal to the product of all of the local functions, multiplied by a normalizing constant, if necessary.

If a local function is a conditional distribution over a variable, the edge connecting the local function to the variable may be directed toward the variable, to graphically indicate that the local function is a conditional distribution. This notation is useful for converting a factor graph to a Bayesian network.

Figure 1(d) shows a factor graph corresponding to the Bayesian network in Figure 1(b), where variable vertices are shown as white discs and function vertices are shown as black discs. This graph indicates that the joint distribution factors into the product $P(T)P(S)P(V|T,S)P(A|V)P(B|V)$.

A Bayesian network is converted to a factor graph by creating one variable vertex for each variable, creating one function vertex for each variable, connecting the function vertex for each variable to the variable and its parents, and setting the local function for each function vertex to P (variable|parents) from the Bayesian network. A factor graph is converted to a Bayesian network by removing the function nodes and using the directed edges to identify the child–parent relationships. Compare Figure 1(b) and Figure 1(d).

A Markov network is converted to a factor graph by creating one variable vertex for each variable, creating one function vertex for each clique potential, setting the local function for each function vertex to the corresponding clique potential, and connecting the function vertex to the variables on which it depends (i.e., to the variables in the corresponding clique of the Markov network). A factor graph is converted to a Markov network by considering each local function in turn, and creating a maximal clique from all variables connected to the local function. Compare Figure 1(c) and Figure 1(d).

From the above descriptions, it is clear that a factor graph has a unique Markov network, and it has a unique Bayesian network as well, if edge directions are provided.

PROBABILISTIC INFERENCE

From a joint distribution, we can compute the conditional distribution of one subset of variables given another subset of variables. To do this, we use the *chain rule*

$$P(A,B) = P(A|B)P(B) \tag{10}$$

and the rule for marginalization

$$P(A) = \sum_{B} P(A,B) \tag{11}$$

For the example where $P(T,S,V) = P(T)P(S)P(V|T,S)$, we can compute $P(T|V=v)$ as follows:

$$\begin{aligned} P(T|V=v) &= \frac{P(T,V=v)}{P(V=v)} \\ &= \frac{\sum_s P(T,S=s,V=v)}{\sum_{t,s} P(T=t,S=s,V=v)} \\ &= \frac{\sum_s P(V=v|T,S=s)P(T)P(S=s)}{\sum_{t,s} P(V=v|T=t,S=s)P(T=t)P(S=s)} \end{aligned} \tag{12}$$

While this 'direct' approach works for small problems, the number of additions needed grows exponentially with the number of variables. So, the direct approach becomes intractable when there are more than a few dozen binary variables, since the summations will then involve many billions of terms.

PROBABILITY (BELIEF) PROPAGATION, OR THE SUM–PRODUCT ALGORITHM

Probability propagation provides a way of computing the distribution of one variable given a set of observed variables. This computation is usually not as straightforward as the direct application of Bayes' rule, since there may be many other variables that must be properly accounted for. For example, to compute $P(A|B)$ from a distribution $P(A,B,C)$, we first sum over C to get $P(A,B) = \sum_c P(A,B,C)$.

The algorithm is easily understood from an example. Suppose the distribution over variables $A, B, \ldots, F$ factorizes as follows:

$$P(A,B,C,D,E,F) = P(A|B)P(B|C)P(C|D)P(D|E)P(E|F)P(F) \quad (13)$$

(Note that the Bayesian network, Markov network, and factor graph for this distribution all have the form of a chain.) Say we would like to compute the marginal distribution, $P(A)$:

$$P(A) = \sum_B \sum_C \sum_D \sum_E \sum_F \big(P(A|B)P(B|C)P(C|D) P(D|E)P(E|F)P(F)\big) \quad (14)$$

Computing this distribution directly takes roughly 2^5 multiplications and additions.

The distribution can be computed more efficiently by distributing the summations over the products:

$$P(A) = \sum_B P(A|B)\Bigg(\sum_C P(B|C)\Bigg(\sum_D P(C|D) \Bigg(\sum_E P(D|E)\Bigg(\sum_F P(E|F)P(F)\Bigg)\Bigg)\Bigg)\Bigg) \quad (15)$$

Computing the distribution $P(A)$ by successively computing 'partial distributions' takes roughly 2×5 summations, an exponential speed-up over the direct approach.

The computation of each 'partial distribution' can be thought of as a procedure that takes in messages, combines them with a conditional probability (or a potential or local function), performs a summation, and produces a new message. In the above example, the summation $\sum_F P(E|F)P(F)$ produces a message that is a real-valued function of E.

Probability propagation has a very simple form in factor graphs, so we describe the algorithm using factor graphs. Using the procedures described above, Bayesian networks and Markov networks can easily be converted to factor graphs. Also, the Bayesian network or Markov network corresponding to a given factor graph is quite obvious, so working with the factor graph does not obfuscate the original model.

Probability propagation consists of passing messages (implemented in a computer as short vectors of real numbers) on edges in the factor graph. Both function vertices and variables vertices combine incoming messages to produce outgoing messages on each of their edges. Each edge in the factor graph can pass a message in either direction, but *the number of values in a message is equal to the number of values its neighboring variable can take on*. For any edge, this number is unique, since in a factor graph, each edge is connected to only one variable vertex. So, we can think of each message as being a function of its neighboring variable.

For now, we'll assume that we are given a *message-passing schedule* that specifies which messages should be updated at each timestep. Think of each edge in the factor graph as having two message buffers (memory to store two messages) – one for each direction. Initially, we set all the messages (all the elements of all the vectors used to store the messages) to 1.

There are three types of computation that are performed in probability propagation:

1. *Propagating variable-to-function messages*. To produce an outgoing message on an edge, a variable computes the element-wise product of incoming messages on the other edges. For example, in Figure 2(a), $f(V)$ is the message sent from function 4 to variable V (note that it is a function of its neighboring variable, V). Let $g(V)$ be the message sent from function 5 to variable V. The message sent from variable V to function 3 is computed from:

$$h(V) = f(V)g(V) \quad (16)$$

That is, for each value of V, the corresponding elements $f(V)$ and $g(V)$ are multiplied together. Note that if a variable has just one edge (e.g., A in Figure 2(a)) its outgoing message is set to 1.

Propagating observations. If variable V is observed and has the value v, then an outgoing message is computed in the same fashion as described above, *except* that for all values $V \neq v$, we set the outgoing

message to *zero*. So, in the above example, we compute $h(V)$ as follows:

$$h(V) = \begin{cases} f(V)g(V) & \text{if } V = v \\ 0 & \text{if } V \neq v \end{cases} \quad (17)$$

2. *Propagating function-to-variable messages*. To produce an outgoing message on an edge, a function takes the product of its associated local function (conditional probability function, in the case of Bayesian networks, potential, in the case of Markov networks) with the incoming messages and sums over all variables in the conditional probability function, *except* the variable to which the message is being sent. For example, in Figure 2(b), $f(T)$ is the message sent from variable T to function 3 (note that it is a function of its neighboring variable), and $g(S)$ is the message sent from variable S to function 3. The message sent from function 3 to variable V is computed from

$$h(V) = \sum_T \sum_S P(V|T,S) f(T) g(S) \quad (18)$$

3. *Fusion*. Suppose the *incoming* messages to variable V are $f(V)$, $g(V)$ (from Figure 2(a)) and $h(V)$ (from Figure 2(b)). Variable V can fuse its incoming messages to compute an estimate of the *joint* probability of V and the observed variables:

$$\hat{P}(V, \text{Observations}) = f(V)g(V)h(V) \quad (19)$$

If V is observed to have the value v, we use

$$\hat{P}(V, \text{Observations}) = \begin{cases} f(V)g(V)h(V) & \text{if } V = v \\ 0 & \text{if } V \neq v \end{cases} \quad (20)$$

If $\hat{P}(V, \text{Observations})$ is then normalized with respect to V, we obtain an estimate of the *conditional* probability of V *given* the observed variables:

$$\hat{P}(V|\text{Observations}) = \hat{P}(V, \text{Observations}) / \sum_V \hat{P}(V, \text{Observations}) \quad (21)$$

Exact Inference Using Probability Propagation

If the factor graph is a tree, and if the messages arriving at a variable, say V, are based on the input from every other vertex in the graph (variable vertices and function vertices), then

$$\hat{P}(V, \text{Observations}) = P(V, \text{Observations}) \quad (22)$$

and

$$\hat{P}(V|\text{Observations}) = P(V|\text{Observations}) \quad (23)$$

That is, the fused messages give *exact* probabilistic inferences.

The Generalized Forward–Backward Algorithm

Suppose we wish to infer the probability for each and every variable in the network, given the observations. Clearly, for each and every variable in the network to receive messages from each and every other variable, at least roughly $2E$ messages must be computed, where E is the number of edges in the factor graph. (Slightly less than $2E$ messages may be needed, since, for example in Figure 2(a), we needn't pass a message from variable T to function 1.) The following procedure, called the *generalized forward–backward algorithm*, shows how we can achieve this bound – infer the probability for each and every variable in the network by passing $2E$ messages.

First, arbitrarily pick a vertex in the factor graph and call it the 'root'. Form a tree by arranging the vertices in layers, with the root at the top. Now, pass messages layer by layer *up* from the bottom to the top and then pass messages layer by layer *down* from the top to the bottom. Clearly, this procedure computes $2E$ messages and the messages arriving at each variable contain the input from each and

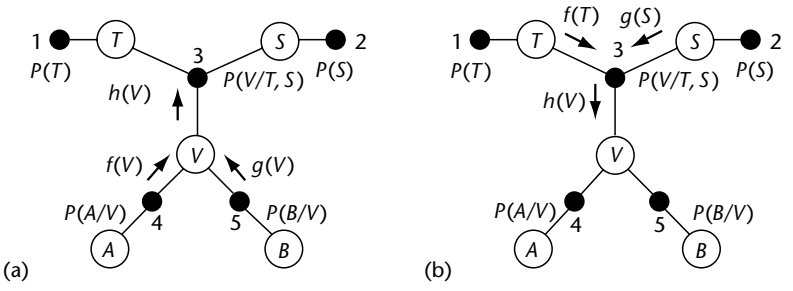


Figure 2. (a) Computing a variable-to-function message in the factor graph from Figure 1(d). The edge directions in the factor graph are dropped for visual clarity. (b) Computing a function-to-variable message.

every other vertex in the factor graph. Note that in this case, we need not initialize the messages.

Approximate Inference Using Probability Propagation

If the messages arriving at a variable do not contain the input from each and every other vertex in the factor graph, then $\hat{P}(A, \text{Observations})$ may not equal $P(A, \text{Observations})$. However, it *may* be a good estimate and in some cases can even be exactly correct. A more interesting case is when the factor graph is not a tree, but contains lots of cycles. In this case, even if the messages arriving at a variable contain the input from each and every other vertex in the factor graph, $\hat{P}(A, \text{Observations})$ will usually not be equal to $P(A, \text{Observations})$, because of the cycles. However, there are some very impressive applications where the approximation is astonishingly good (Frey and MacKay, 1998; Freeman and Pasztor, 2000; Frey *et al*., 2001). Also, new analysis is emerging that partly explains the approximation (Weiss and Freeman, 2001; Yedidia *et al*., 2001; Wainwright *et al*., 2002).

References

Freeman W and Pasztor E (1999) Learning low-level vision. *Proceedings of the International Conference on Computer Vision*, 1182–1189.

Frey BJ and MacKay DJC (1998) A revolution: {B}elief propagation in graphs with cycles. In: Jordan MI, Kearns MI and Solla SA (eds) *Advances in Neural Information Processing Systems*, *vol. 10*. Cambridge, MA: MIT Press.

Frey BJ, Koetter R and Petrovic N (2002) Very loopy belief propagation for unwrapping phase images. In: Dietterich TG, Becker S and Ghahraman Z (eds) *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Kinderman R and Snell JL (1980) *Markov Random Fields and Their Applications*. Providence state, USA: American Mathematical Society.

Kschischang FR, Frey BJ and Loeliger HA (2001) Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* **47**(2): 498–519.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Wainwright MJ, Jaakkola T and Willsky AS (2002) Tree-based reparameterization for approximate estimation on loopy graphs. In: Dietterich TG, Becker S and Ghahraman Z (eds) *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Weiss Y and Freeman W (2001) On the optimality of solutions of the max-product belief propogation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory* **47**(2): 736–744.

Yedidia J, Freeman WT and Weiss Y (2001) Generalized belief propogation. In: Dietterich TG, Becker S and Ghahraman Z (eds) *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Further Reading

Neapolitan RE (1990) *Probabilistic Reasoning in Expert Systems*. New York, NY: John Wiley and Sons.

Frey BJ (1998) *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: MIT Press.

Hinton GE, Dayan P, Frey BJ and Neal RM (1995) The wake-sleep algorithm for unsupervised neural networks. *Science* **268**: 1158–1161.

Hinton GE and Sejnowski TJ (1986) Learning and relearning in Boltzmann machines. In: Rumelhart DE and McClelland JL (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 282–327. Cambridge, MA: MIT press.

Jordan MI (2001) *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Binding Problem

Intermediate article

Jacques P Sougné, University of Liège, Liège, Belgium

CONTENTS

The binding problem is the problem of how a cognitive system (a brain or a computational model) groups a set of features together, associates a filler with a role, a value with a variable, an attribute with a concept, etc.

OVERVIEW OF THE BINDING PROBLEM

Humans effortlessly recognize objects, faces, sounds, tastes and so on, all of which are composed of many features. For example, a red ball has a particular shape and color. Since shape and color are not processed in the same cortical areas, there must be a mechanism able to bind the round shape with the red color and differentiate them from a blue cube nearby. A particular face is composed of a set of properties that have to be linked but must also be bound to a particular name and be distinguished from other faces and other names. Representing a predicate and its arguments in 'John loves Mary' requires correctly binding the filler 'John' to the role of 'lover', and the filler 'Mary' to the role of 'lovee', without confusing them. Representing a rule 'if *a* then *b*' requires correctly binding '*a*' to the antecedent and '*b*' to the consequent part of the rule. As these examples show, binding is an essential mechanism in a wide range of cognitive tasks.

How can an artificial neural network perform binding? Early critiques of connectionism raised this problem. How could a connectionist network represent the simple fact that 'the red rose is on a green table'? Since 'red', 'green', 'rose' and 'table' have distributed overlapping representations in the system, the problem is to correctly associate 'rose' with 'red' and 'table' with 'green' while avoiding 'crosstalk', i.e. avoiding the spurious associations between 'table' and 'red' and between 'rose' and 'green'.

The first idea that comes to mind for solving the problem in a connectionist setting is to increase the connection weights between 'rose' nodes and 'red' nodes and between 'table' nodes and 'green' nodes while decreasing other connection weights. But once these connection weights have been set, linked nodes cannot individually participate in other representations. Nodes must be reusable for representing another object like 'yellow rose'. Binding, therefore, must occur dynamically.

Fodor and Pylyshyn (1988) point out a difficulty that arises from the binding problem. They question the value of connectionism as a model of cognition since, according to them, connectionist models cannot display what they call 'systematicity'. The examples they give make reference to the binding problem. They point out that there are no people able to think that 'John loves Mary' but unable to think that 'Mary loves John'. Nobody is able to infer that 'John went to the store' from 'John, Mary, Susan and Sally went to the store' but unable to infer that 'John went to the store' from 'John, Mary and Susan went to the store'. One can define systematicity as the ability to apply a particular structure to any content.

Symbolic systems achieve systematicity very efficiently, but human systematicity does not prevail for complex logical forms. The ability to deduce $\sim A$ from the two statements $\sim B$ and $(A \supset B)$ is not linked to the ability to deduce A from the two statements B and $(\sim A \supset \sim B)$. These two inferences are both obtained by applying the same deduction rule (modus tollens), but they are not cognitively equivalent. Research in cognitive psychology has demonstrated that the first inference is easier (e.g. Evans, 1977; Wildman and Fletcher, 1977). Furthermore, children learning to talk do not display systematicity in their predicate use. (*See* **Symbolic versus Subsymbolic**)

BINDING IN SYMBOLIC SYSTEMS

Binding is not a problem in symbolic systems. A variable is defined and a value is associated with

('bound to') that variable. The lack of constraints on binding means that symbolic systems would not encounter the problems that humans sometimes have when doing binding.

For example, when the time of visual presentation is short, 'illusory conjunctions' are frequent. Illusory conjunction occurs when a particular feature of one object is incorrectly bound to another object. When people are rapidly shown a scene containing a blue ball and a yellow vase they may report having seen a yellow ball. Illusory conjunction is even more frequent if objects share common features.

While most connectionist systems of the 1980s were unable to do binding at all, symbolic systems were clearly too efficient in solving the binding problem compared with humans. Even if one could constrain the performance of a symbolic system with a parameter that would stochastically perturb binding, this would be far less psychologically persuasive than having binding be constrained as an emergent consequence of the architecture of the system.

Before examining the binding problem for connectionist systems, we will briefly review how the brain might achieve binding.

HOW THE BRAIN MIGHT ACHIEVE BINDING

How are neuron assemblies constituted? How are different assemblies bound and differentiated? How can we, as external observers, understand the messages involved in neural patterns of activity – and what code is used by the brain? There are two main hypotheses concerning this code: rate codes and pulse codes.

Rate Codes

There are three ways of considering rate coding, each of which uses a different averaging procedure.

Rate can be computed for a single cell firing over time. In this case, spikes are counted and their number divided by the time elapsed. The objection to this code is that behavioral reaction times are sometimes too small to allow the system to compute an average. If neurons fired at regular intervals, averages could theoretically be computed after two spikes. But noise is also a factor. To obtain a good estimate of the rate, it is necessary to compute the average over a longer period.

The second procedure for evaluating neural firing rate is to repeatedly average single cell spikes. The experimenter repeatedly records the spikes of a cell before and after a stimulation. The average is obtained by dividing the total number of spikes by the number of repetitions and the length of the recording intervals. However, this measure cannot be the code used by the organism to process information, since the reactions of most organisms are the consequence of single stimulus presentations.

The third procedure for computing rate involves recording several neurons before averaging. This rate represents the activity of a population of neurons. Some populations seem to react to particular classes of stimuli. If, after stimulation, an experimenter records the firing of each of these neurons and divides the sum by the number of neurons and by the length of the recording time window, a rate measure is obtained. The advantage of this procedure is that it allows the rate to be calculated for a short time window.

Rate coding has received empirical support. For example, Thomas *et al*. (2001) found that what was crucial for categorization was not category-specific neurons, but rather those neurons that respond more (at a higher rate) to one category than to another.

Pulse Codes

Pulse codes are based on precise timing of spikes. One possible pulse code is latency. The idea behind this is that the time separating a stimulus from the first spike of a neuron can carry information. Gawne *et al*. (1996) recorded activity of striate cortex cells and showed that spiking latency was a function of the visual stimulus contrast.

A second possible pulse code is phase coding. Oscillation of a population activity has been found in the hippocampus and cortical areas. This background oscillation can serve as a reference signal. A particular firing of a neuron can be compared to this background oscillation, and its location on the oscillation curve can serve as a code. This coding scheme has received empirical support. O'Keefe and Recce (1993), for example, showed that phase codes contained spatial information independently of spike rate in the rat hippocampus.

A third possible pulse code is synchrony. Neurons corresponding to microfeatures of a stimulus fire at the same time, and thereby allow the representation of the whole stimulus. This hypothesis has also received empirical support. Engel *et al*. (1991) showed that if several objects make up a scene, distinct clusters of synchrony are formed, each associated with a particular object in the scene. Synchrony is often associated with

oscillation since it has been shown that gamma oscillations enable synchronization.

It is important to note that rate codes and pulse codes are not mutually exclusive. Synchrony within a population of neurons over a short period of time also means that the population firing rate is high. It is also possible that different kinds of information could be coded by different coding schemes.

A final problem is how the brain reads the code expressed by neurons. People are capable of describing and observing their own thoughts, but exactly how this is done is not known.

CONNECTIONIST SOLUTIONS TO THE BINDING PROBLEM

Grandmother Cells

The first solution to the binding problem is to use one node for each possible binding. According to this purely localist solution, a binding is represented by a single cell which responds whenever this particular binding is used. This unique cell is called a 'grandmother cell'. This kind of representation poses a number of problems. First, imagine a soccer player banging his head against the goalpost and losing his 'soccer ball' cell. What would this player do after getting up? Would he no longer know anything at all about soccer balls? Second, an enormous number of feature combinations are necessary for representing the multitude of objects, concepts, etc., that humans deal with. If every specific combination had a particular corresponding cell, such a representation would require an impossibly large number of neurons. A final problem is generalization. If every new object would need a new representing neuron, similar objects cannot help in representation building.

Coarse–Fine Coding

At the opposite end of the spectrum from grandmother cells representations are fully distributed representations in which every node may participate in every object representation. Such representations also pose problems. Suppose that a particular object (a ball) has a representation and a particular color (blue) has another representation. Since every node participates in every representation, the conjunction 'blue ball' cannot be represented, since activations coming from 'blue' will be added to activations coming from 'ball' and the subsequent activations will not necessarily represent 'blue ball', 'blue' or 'ball', but could represent nothing or a ghost representation which can correspond to 'pear' or anything else. A solution to this problem is to use a finer coding in which a subset of the population is used to code a particular value. A node becomes active if the input falls in its receptive field. This partially distributed coding needs fewer nodes than purely localist codes to represent the same amount of information, therefore avoiding the problem faced by grandmother cells representation of needing an impossibly large number of cells. However, this solution, at least in its simplest form, only allows one variable to be bound at a time. This solution can represent the conjunction 'blue ball' as the set of 'blue' – responding nodes and 'ball'-responding nodes, but adding a 'yellow vase' to the scene to be represented would make everything bind together. Consequently, we would not be able to detect which color is associated with which object.

Tensor Products

Smolensky (1990) proposed the use of a tensor product representation of binding. Tensor products are similar to outer products of vectors (Figure 1a)). The outer product of two n-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ is the product $\mathbf{u}\mathbf{v}^{\mathrm{T}}$ of $\mathbf{u}$ and the transpose of $\mathbf{v}$ ($\mathbf{v}^{\mathrm{T}}$ is a matrix with one row and n columns). This is an $n \times n$ matrix. If one has to encode 'John loves Mary', the 'John' filler must be bound to the 'lover' role. John in the role of lover will be represented by the outer product of 'John' and 'lover' vectors. This solution can encode bindings, but it does not allow inferencing. It is also not clear what the neural correlate of outer products might be. One disadvantage of the tensor product is the increase of dimensionality for each additional tensor product (the tensor product of two vectors is a 2D array, the tensor product of a vector and a 2D array is a 3D array, etc.). To overcome this problem, other techniques have been proposed in which the binding of two vectors results in a vector (e.g. convolution – correlation: Plate, 1995). (*See* **Convolution-based Memory Models**)

Values Associated with Roles and Fillers

In the model ROBIN (Lange and Dyer, 1989) each filler has an associated node that outputs a particular constant value (called its signature). Each role has an associated object node or binding node. When a role object node has the same activation as that of a symbol's signature, this symbol is bound to the role (Figure 1(b)). A similar solution

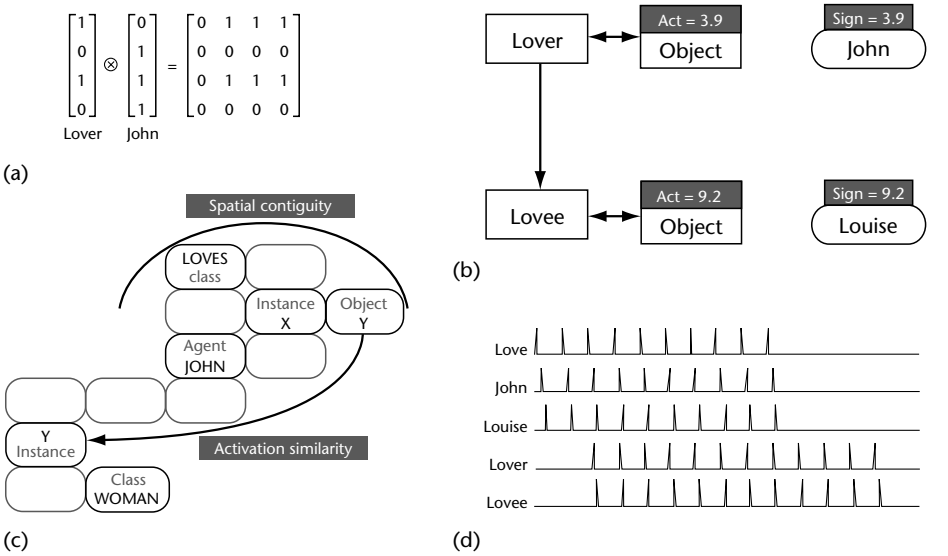


Figure 1. Connectionist solutions to the binding problem. (a) Tensor product representation (Smolensky, 1990) of 'John' bound to 'lover': 'lover' is represented by the (column) vector [1010] and 'John' by the (column) vector [0111]; their binding is represented by their outer product. (b) Values associated with role and filler, as in the model ROBIN (Lange and Dyer, 1989) in which binding is achieved by a match between activation of role object node and filler signature. For representing 'John loves Louise', the activation of the 'lover' object node has the same value as the signature of the 'John' node while the activation value of the 'lovee' object node has the same value as the signature of the 'Louise' node. (c) COMPOSIT binding by activation similarity and spatial contiguity, the representation of 'John loves some woman' in working memory. The register that contains the 'instance' flag with the X symbol denotes a particular instance of the adjacent 'loves' class. The agent of this loving situation is found in the adjacent register containing the 'agent' flag: 'JOHN'. The object of that loving instantiation is found in the adjacent register that contains the 'object' flag 'Y'. By activation similarity, 'Y' points to another register with the same 'Y' symbol and the 'instance' flag. This instance is adjacent to another register that contains the 'class' flag and the 'WOMAN' symbol, denoting that the object of the loving relation is some undefined woman. (d) Binding by synchrony. For representing 'John loves Louise', the firings of the nodes associated with the predicate 'Love' are followed by the 'John' nodes firing in synchrony with the 'lover' nodes while the 'Louise' nodes fire in synchrony with the 'lovee' nodes.

was proposed by Sun (1992) with his hybrid model CONSYDERR, which consists of a localist network and a distributed network. The localist network is composed of nodes representing symbols and links between these symbols representing rules. Each symbol node is linked to several nodes in the distributed network. The nodes of the distributed network represent the features of the symbol. Variable binding is achieved by the use of a particular value which is passed from a role node to a filler node along a link. These solutions lack psychological plausibility since the number of separated bindings is not constrained.

Activation Similarity and Spatial Contiguity

The model COMPOSIT (Barnden and Srinivas, 1991) uses two systems, a long-term memory and a working memory, both of which are connectionist networks. In this model, working memory is composed of several registers filled with activation patterns from long-term memory. Fillers and their roles are stored in registers as two vectors. One vector (the symbol vector) represents the filler and the other vector (the highlighting vector) represents the role. Predicates, related roles, and fillers are stored contiguously and thus constitute a distinguishable set that can be linked to a particular role pertaining to another predicate (Figure 1(c)). A role can be linked to another role by the similarity of their highlighting vectors. This is a very efficient solution that permits recursive predication. However, it is not clear what neural mechanism might correspond to the loading of registers.

The above solutions have various advantages and limitations, but they all lack neural plausibility. We will now explore more neurally plausible solutions.

Binding by Temporal Frequency

Lange and Dyer (1989) proposed the use of temporal frequency (instead of signatures) for binding, each signature being an unique frequency of node

spike. Nodes having the same firing frequency would be bound together. This model has not yet been explored by cognitive scientists, though there are neurobiological data that seem to be consistent with it. However, if the activation of a node is considered to be its firing rate, then binding by activation, as in CONSYDERR and ROBIN, can fall into this category.

Binding by Synchrony

For systems using temporal synchrony for variable binding, nodes can be in two different states: they can be firing ('on'), or they can be at rest ('off'). A node fires at a precise moment and transmits activation to other connected nodes. When a node's activation reaches threshold, it fires. Whenever two nodes (or two sets of nodes) representing two objects fire simultaneously, these objects are temporarily associated (Figure 1(d)). On the other hand, if two nodes (or two sets of nodes) fire in succession, they are distinguished. This is how the systems Shruti (Shastri and Ajjanagadde, 1993) and INFERNET (Sougné, 2001) solve the binding problem.

Synchrony has been used as a binding mechanism in various cognitive models, for perception, for attention, for spatial cognition, for memory, and for different types of inference. These models fit human data well, mainly because the number of distinguishable entities in these systems is constrained by the precision of synchrony. However, it is not clear how these models could represent recursive structures.

The Binding Problem and Distributed Representations

Most of the above models used localist representations. Binding in a fully distributed network is problematic, because as soon as two symbols are required at the same time, their representations may overlap, which could lead to crosstalk.

This problem is greatly reduced by using partially distributed representations, where each symbol is represented by an assembly. This solution avoids problems of 'grandmother' cells, and, since some assemblies share nodes, similarity effects related to binding can be achieved (Sougné, 2001).

THE PROBLEM OF MULTIPLE INSTANTIATION

Multiple instantiation involves the simultaneous use of the same parts of the knowledge base in different ways. Knowing that 'John is the father of Peter' and that 'Peter is the father of Paul', one can readily infer that John is the grandfather of Paul. To derive this conclusion, one must simultaneously instantiate the predicate 'father of' and the object 'Peter' twice. Precisely how this is done is the problem of multiple instantiation. This problem is also called the type–token problem. It is closely related to the binding problem. Solving the binding problem is not by itself sufficient to solve the problem of multiple instantiation; it is, however, necessary.

Symbolic models load copies of pieces of knowledge into a working area before processing them. For these models, there is no problem of multiple instantiation: they simply make several copies of the same content from the long-term knowledge base (LTKB) and store them in the working area. By contrast, for connectionist models that use the knowledge base itself as the place where symbols are associated and processed, multiple instantiation is a serious problem. How can the same object be associated with different roles at the same time without making several copies of this object? In general, multiple instantiation poses significant problems for distributed representations. Two closely related objects will, in principle, share nodes. Therefore, if both objects are needed simultaneously (for different roles), their common nodes must be associated with two different clusters of nodes.

People are able to cope with multiple instances of the same concept, unlike most connectionist models, but their performance when doing so is diminished. There is no naturally arising decrease in performance for symbolic models doing multiple instantiation.

Multiple Instantiation in Neural Networks

For localist networks (where each symbol is represented by a single node) that can represent predicates with more than one argument, the problem of multiple instantiation arises if two instantiations of a predicate have different bindings of their arguments. For example, 'John likes tennis and John likes football' does not require separate instances of the predicate 'likes' since this statement is equivalent to 'John likes tennis and football', in which 'tennis' and 'football' are bound to the same role. However, when two sets of two fillers must be associated with identical pairs of roles, the system must be able to handle two copies of the predicate and argument slots. For example, 'John likes football and Mary likes tennis' cannot be reduced to 'John and Mary like football and tennis',

because if it is one can no longer distinguish who likes what.

In distributed networks the problem of multiple instantiation arises as soon as one node must be used by clusters that have to be differentiated. If a predicate with more than one argument has to be represented, and if either the predicate's arguments or any of the arguments' fillers need to use a common node, then this node will have to be bound to different roles.

Relevance for Cognitive Science

In some sense, the connectionist limitations regarding the problem of multiple instantiation could be a blessing in disguise, because some tasks involving multiple instantiation are precisely those tasks that cause problems for humans. Empirical evidence of these difficulties comes from relational reasoning, from repetition blindness, and from the effects of similarity on working memory and on perception. These data show that multiple instantiation can indeed cause problems for humans and animals. In short, when confronted with multiple instantiation, people tend to be slower or to make more mistakes. A good model should not only be able to deal with multiply instantiated symbols, but should also reflect the difficulties that humans have with them.

Connectionist Solutions

There are three types of solution to the problem of multiple instantiation. The first type of solution is to use two systems, an LTKB and a working area into which copies of pieces of the LTKB are loaded. The second type of solution uses several copies of the same symbol in the LTKB. The third relies on superposing frequencies of oscillation.

Multiple copies loaded in a working area

The first solution is borrowed from symbolic models. Each additional instance of a symbol that is required will be represented by an additional copy of the symbol inside short-term memory (STM). An example is the COMPOSIT model (Barnden and Srinivas, 1991), illustrated in Figure 2(a). This solution is probably the most powerful one, and allows a broad range of high-level inferencing capabilities.

Unfortunately, however, this solution inadequately reflects the difficulties people have when they perform multiple instantiation. For these models, even if the STM has a limited capacity, it is as easy to load one copy as to fill the STM with copies of the same content (unless, of course, this is explicitly prevented). Other solutions have been developed, however, in which the STM is the activated part of the LTKB.

Multiple copies inside the LTKB

Lange (1992) describes a potential solution designed to handle multiple instantiation inside ROBIN. This solution involves each symbol having more than one signature and each role more than one activation (Figure 2(b). Multiple instantiation in Shruti (Shastri and Ajjanagadde, 1993) is achieved by the use of a bounded set of copies or banks of predicates and their argument slots (usually at most three).

One difficulty with the solution involving multiple copies inside LTKB is that the number of allowable copies is arbitrary. Another difficulty is the abrupt loss of performance. As long as a copy is available, performance will be perfect, but when no more copies are available, performance will decrease abruptly. This behavior does not accurately reflect human performance.

Period doubling

Sougné (2001) describes another solution to the problem of multiple instantiation. In INFERNET, nodes pertaining to a doubly instantiated object will support two oscillations while those singly instantiated will support one oscillation (Figure 2(c)). This makes doubly instantiated object nodes fire twice while singly instantiated objects fire once. This means that each new instance will occupy a new place in STM, thus avoiding crosstalk.

This solution can be compared to the neural phenomenon of bifurcation by period doubling, whereby a stable oscillatory state can lose its stability, giving rise to a new stable state with doubled period. Similar period doubling was obtained experimentally, *in vivo*, by Ishizuka and Hayashi (1996). They recorded cell activity in the somatosensory cortex of rats while increasing the stimulation frequency on a medial lemniscus fiber. As stimulation frequency increased, successive period doublings occurred, finally leading to a chaotic firing pattern.

It is therefore reasonable to assume that multiply instantiated symbol nodes could receive a higher frequency of input than singly instantiated ones; so that increasing the number of instantiations would increase the chance of period doublings.

Another advantage of this solution is its psychological plausibility. As the number of instantiations increases, cognitive performance decreases. The

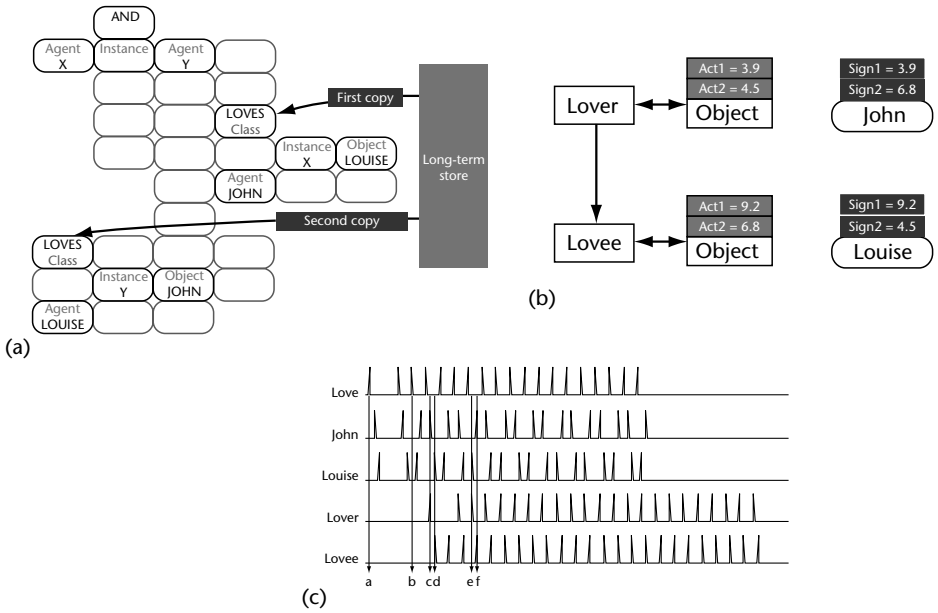


Figure 2. Connectionist solutions to the problem of multiple instantiation. Examples of how different solutions could encode 'John loves Louise and Louise loves John'. (a) The solution of the COMPOSIT model (Barnden and Srinivas, 1991), borrowed from symbolic artificial intelligence. Each additional instance occupies a new place in the 'working memory' area. The 'AND' symbol is instantiated and has two adjacent agents 'X' and 'Y' each pointing to their own instantiation of 'loves'. The register that contains the 'instance' flag with the 'X' symbol denotes a particular instance of the adjacent 'loves' class. The agent of this loving situation is found in the adjacent register containing the 'agent' flag: 'JOHN'. The object of that loving instantiation is found in the adjacent register that contains the 'object' flag: 'LOUISE'. The register that contains the 'instance' flag with the 'Y' symbol denotes a particular instance of the adjacent 'loves' class. The agent of this loving situation is found in the adjacent register containing the 'agent' flag: 'LOUISE'. The object of that loving instantiation is found in the adjacent register that contains the 'object' flag: 'JOHN'. (b) Multiple copies inside a long-term knowledge base. Handling multiple instantiation inside ROBIN requires that each symbol has more than one signature and each role more than one activation (Lange, 1992). For representing 'John loves Louise' and 'Louise loves John', one activation of the 'Lover' object node has the same value as one of the signatures of the 'John' node, and one activation of the 'Lovee' object node has the same value as one of the signatures of the 'Louise' node. Additionally, one activation of the 'Lover' object node has the same value as one of the signatures of the 'Louise' node, and one activation of the 'Lovee' object node has the same value as one of the signatures of the 'John' node. (c) Multiple instantiation by period doubling (Sougné, 2001). The fact 'John loves Louise' is introduced at time a. This pattern starts oscillating. At time c 'John' begins to fire in synchrony with 'Lover' and at time d 'Louise' starts firing in synchrony with 'Lovee'. The statement 'Louise loves John' is introduced at time b. This pattern starts oscillating. At time e 'Louise' begins to fire in synchrony with 'Lover' and at time f 'John' starts firing in synchrony with 'Lovee'. Therefore, the 'Love', 'John', 'Louise', 'Lover', and 'Lovee' nodes sustain two oscillations, enabling the 'John' and 'Louise' nodes to bind alternatively to 'Lover' and 'Lovee' nodes.

model can also simulate various similarity effects that humans display. However, it has not yet been proved to be able to simulate recursive predication.

SUMMARY

When looking at a field of poppies on a sunny day, how can we correctly associate the color red with the poppies, green with the grass, and blue with the sky, and avoid associating the color red with the grass and the color blue with the poppies? If we see Louise picking a red poppy, how can we correctly associate Louise with the picker and the red poppy with the picked object, without making the opposite and incorrect association? As long as the number of distinct ensembles remains small, these associations are easy for us, but how does the brain perform them correctly? The question of how a connectionist system binds a set of features together, associates a filler with a role, a value with a variable, an attribute with a concept, etc., is what is called 'the binding problem'. Related to this problem is the problem of multiple instantiation.

References

Barnden JA and Srinivas K (1991) Encoding techniques for complex information structures in connectionist systems. *Connection Science* **3**: 269–315.

Engel AK, Kreiter AK, König P and Singer W (1991) Synchronisation of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. *Proceedings of the National Academy of Sciences* **88**: 6048–6052.

Evans JStB (1977) Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology* **29**: 297–306.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Gawne TJ, Kjaer TW and Richmond BJ (1996) Latency: another potential code for feature binding in striate cortex. *Journal of Neurophysiology* **76**: 1356–1360.

Ishizuka S and Hayashi H (1996) Chaotic and phase-locked responses of the somatosensory cortex to a periodic medial lemniscus stimulation in the anesthetized rat. *Brain Research* **723**: 46–60.

Lange TE (1992) Lexical and pragmatic disambiguation and re-interpretation in connectionist networks. *International Journal of Man–Machine Studies* **36**: 191–220.

Lange TE and Dyer MG (1989) High-level inferencing in a connectionist network. *Connection Science* **1**: 181–217.

O'Keefe J and Recce M (1993) Phase relationship between hippocampal place units and the hippocampal theta rhythm. *Hippocampus* **3**: 317–330.

Plate T (1995) Holographic reduced representations. *IEEE Transactions on Neural Networks* **6**: 623–641.

Shastri L and Ajjanagadde V (1993) From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences* **16**: 417–494.

Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46**: 159–216.

Sougné J (2001) Binding and multiple instantiation in distributed networks of spiking nodes. *Connection Science* **13**: 99–126.

Sun R (1992) On variable binding in connectionist networks. *Connection Science* **4**: 93–124.

Thomas E, Van Hulle MM and Vogels R (2001) Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience* **13**: 1–11.

Wildman TM and Fletcher HJ (1977) Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology* **13**: 630–636.

Further Reading

Barnden JA and Pollack JB (1991) Introduction: problems for high-level connectionism. In: Barnden JA and Pollack JB (eds) *Advances in Connectionist and Neural Computation Theory*, vol. I 'High-Level Connectionist Models', pp. 1–16. Norwood, NJ: Ablex.

Dyer MG (1991) Symbolic neuroengineering for natural language processing: a multilevel research approach. In: Barnden JA and Pollack JB (eds) *Advances in Connectionist and Neural Computation Theory*, vol. I 'High-Level Connectionist Models', pp. 32–86. Norwood, NJ: Ablex.

Hadley RF (1996) Connectionism, systematicity and nomic necessity. In: Cottrell GW (ed.) *Proceedings of the Eighteenth Conference of the Cognitive Science Society*, pp. 80–85. Mahwah, NJ: Erbaum.

Maass W and Bishop CM (1999) *Pulsed Neural Networks*. Cambridge, MA: MIT Press.

Norman DA (1986) Reflections on cognition and parallel distributed processing. In: McClelland JL and Rumelhart DE (eds) *Parallel Distributed Processing*, vol. II, pp. 531–546. Cambridge, MA: MIT Press.

Singer W (1993) Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* **55**: 349–374.

Sougné J (1998) Connectionism and the problem of multiple instantiation. *Trends in Cognitive Sciences* **2**: 183–189.

Treisman A (1996) The binding problem. *Current Opinion in Neurobiology* **6**: 171–178.

Von der Malsburg C (1995) Binding in models of perception and brain function. *Current Opinion in Neurobiology* **5**: 520–526.

Catastrophic Forgetting in Connectionist Networks

Intermediate article

Robert M French, University of Liège, Liège, Belgium

CONTENTS

Unlike human brains, connectionist networks can forget previously learned information suddenly and completely ('catastrophically') when learning new information. Various solutions to this problem have been proposed.

INTRODUCTION

The connectionist paradigm in artificial intelligence came to prominence in 1986 with the publication of Rumelhart and McClelland's two-volume collection of articles entitled *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Some 20 years earlier, research on an elementary type of neural network, known as perceptrons, had come to a sudden halt in 1969 with the publication of *Perceptrons*, Minsky and Papert's careful mathematical analysis of the capacities of a particular class of single-layered perceptrons. Minsky and Papert's work demonstrated a number of fundamental theoretical limitations of elementary perceptrons. Multi-layered perceptrons and new learning algorithms, which overcome these limitations, were developed over the course of the next two decades. These new networks were able to do many tasks that presented serious problems for traditional symbolic artificial intelligence programs. For example, they were able to function appropriately with degraded inputs, they could generalize well, and they were fault-tolerant. In the late 1980s there were many attempts to apply these networks to tasks ranging from underwater mine detection to cognition, from stock market prediction to bank loan screening.

At the end of the 1980s, however, a problem with these multi-layered networks came to light. McCloskey and Cohen (1989) and Ratcliff (1990) showed that the very property – namely, using a single set of weights as the network's memory – that gave these networks such power caused an unsuspected problem: catastrophic interference. Grossberg (1982) had previously cast this problem in the more general context of 'stability–plasticity'. In short, the problem was to determine how to design network architectures that would be sensitive to new input without being overly disrupted by it.

Catastrophic interference occurs when a network has learned to recognize a particular set of patterns and is then called upon to learn a new set of patterns. The learning of the new patterns modifies the weights of the network in such a way that the previously learned set of patterns is forgotten. In other words, the newly learned patterns suddenly and completely – 'catastrophically' – erase the network's memory of the previously learned patterns.

CATASTROPHIC VERSUS NORMAL FORGETTING

Catastrophic forgetting is significantly different from normal human forgetting. In fact, catastrophic forgetting is almost unknown in humans. Barnes and Underwood (1959) conducted a well-known series of experiments in human forgetting. Subjects begin by learning a set of paired associates *A–B*, each consisting of a non-word and a word (e.g., *pruth–heavy*). Once this learning is complete, they learn to associate a new word with each of the original non-words (*A–C*). At various points during the learning of the *A–C* pairs, they are asked to recall the originally learned *A–B* associates. McCloskey and Cohen (1989) conducted a similar experiment using addition facts on a standard connectionist network. After five learning trials in the *A–C* condition, the network's knowledge of the *A–B* pairs had dropped to 1 per cent, and after 15 trials it had disappeared. The newly learned pairs

had catastrophically interfered with the previously learned pairs.

The problem for connectionist models of human memory – in particular, for those models with a single set of shared multiplicative weights (e.g. feedforward back propagation networks) – is that catastrophic interference is hardly observed in humans. This raises a number of issues of significant practical and theoretical interest. Arguably, the most important issue for cognitive science is understanding how the brain manages to overcome the problem of catastrophic forgetting. The brain is, after all, a distributed (or partially distributed) neural network, yet it does not exhibit anything like the catastrophic interference seen in connectionist networks. What neural architecture allows the brain to overcome catastrophic interference, and what characteristics of neural networks in general will allow them to overcome it?

At present, in order to avoid catastrophic interference, most connectionist architectures rely on learning algorithms that require the network to cycle repeatedly through all the patterns to be learned, adjusting the weights by a small amount at a time. After many cycles (called 'epochs') through the entire set of patterns, the network will (usually) converge on an appropriate set of weights for the set of patterns that it is supposed to learn. Humans, however, do not learn in this way. In order to memorize ten piano pieces, we do not play each piece once and then cycle repeatedly through all the pieces until we have learned them all. We learn piano pieces – and just about everything – sequentially. We start by learning one or two pieces thoroughly, then learn a new piece, then another, and so on. If a standard connectionist network were to do this, each new piece learned by the network would probably erase from its memory all previously learned pieces. By the tenth piece, the network would have no recollection whatsoever of the first piece. In order for connectionist networks to exhibit anything like human sequential learning, they must overcome the problem of catastrophic interference.

MEASURES OF CATASTROPHIC INTERFERENCE

The two most common measures of catastrophic interference are known as 'exact recognition' and 'relearning'. In both cases, the network is first trained to a given level on a set of patterns. It is then given a second set of patterns to learn. Once it has learned this second set of patterns, we use one of the two measures of forgetting to determine the effect on the original learning of having learned the second set of patterns. The exact recognition measure depends on the percentage of the original patterns that can still be recognized by the network. (We give the input part of the pattern to the network; and if it produces the correct output, it is considered to have 'recognized' the pattern.) The relearning measure, first proposed by Ebbinghaus for human memory in the late nineteenth century, depends on how long it takes the network to relearn the originally learned patterns. Thus, even if the rate of exact recognition is very low, the knowledge might lie 'just below the surface' and the network might be able to relearn it very quickly. The studies by McCloskey and Cohen (1989) and Ratcliff (1990) used an exact recognition measure. A study by Hetherington and Seidenberg (1989) using the relearning measure showed that, at least in some cases, catastrophic interference might be less of a problem than was thought because the network, even if it could not recognize the originally learned patterns exactly, could relearn them very quickly.

SOLUTIONS TO THE PROBLEM

As soon as the problem came to light, there were many attempts to solve it. One of the first was the suggestion by Kortge (1990) that the problem was due to the back propagation learning algorithm. He proposed a modified learning algorithm using what he called 'novelty vectors' that did, in fact, decrease catastrophic interference. The basic idea of novelty vectors was 'blame assignment'. Kortge's learning rule was developed for auto-associative networks, i.e., networks that, starting from a random weight configuration, learn to produce on output the vectors that they received on input. Each pattern to be learned was fed through the network and the output was compared with the intended output (i.e., the input). Kortge called the resulting difference vector a 'novelty vector' because the bigger the activation differences at each node, the more novel the input – for vectors that the network had already learned there would be little difference between output and input. The novelty vector indicated by how much to change the weights: the greater the novelty activation, the more the weights were changed. This technique significantly reduced catastrophic interference, but it applied only to auto-associative networks.

French (1992) argued that catastrophic forgetting was in large measure due to excessive overlap of internal representations. He claimed that the

problem lay with the fully distributed nature of the network's internal representations, and suggested that by developing algorithms that produced 'semi-distributed' internal representations (i.e., representations whose activation was spread only over a limited subset of hidden nodes) catastrophic interference could be reduced. To this end he suggested a learning algorithm, 'node sharpening', that developed much sparser internal representations than standard back propagation. The result was a significant reduction in catastrophic interference. However, the overly sparse representations developed by this technique resulted in a significant decrease in the network's ability to discriminate categories. What was needed was a means of making representations both highly distributed and well enough separated.

Brousse and Smolensky (1989) and McRae and Hetherington (1993) showed that the problem was closely related to the domain of learning. In domains with a high degree of internal structure, such as language, the problem is much less acute. McRae and Hetherington managed to eliminate the problem by pretraining the network on a random sample of patterns drawn from the domain. Because of the degree of structure in the domain, this sample was enough to capture overall regularities of the domain. Consequently, the new patterns to be learned were perceived by the network to be variants of patterns, already learned and did not interfere with previous learning.

The early attempts to solve the problem of catastrophic interference concentrated on reducing representational overlap on input or internally. Kortge (1990) and Lewandowsky (1991) modified the input vectors in an attempt to achieve greater mutual orthogonalization (this is equivalent to reducing the overlap among input vectors). French (1992), Murre (1992), Krushke (1992) and others developed algorithms that reduced internal representational overlap, thereby significantly reducing the amount of catastrophic interference.

Some researchers (e.g. Carpenter, 1994) have laid the blame for the problem of catastrophic interference on a particular architectural feature of the most widely used class of connectionist networks, namely their use of multiplicative connection weights. In the ART family of networks (Carpenter and Grossberg, 1987), new input does not interfere with previously learned patterns because the network is able to recognize new patterns as being new and assigns a new set of nodes for their internal representation.

Hopfield networks, and related architectures, have been shown to have critical saturation limits beyond which there is a steep fall in memory performance. For these networks, forgetting is gradual until the memory becomes saturated, at which point it becomes catastrophic.

REHEARSAL AND PSEUDOREHEARSAL

Most connectionist networks learn patterns concurrently. In terms of human cognition, this is a contrived type of learning. For a given set of n patterns $\{P_1,\ldots P_n\}$, the network will successively adjust its weights by a very small amount for all of the patterns and then will repeat this process until it has found a set of weights that allow it to recognize all n patterns. This, in itself, is a way of learning foreign to humans. In addition, if a new set of patterns $\{P_{n+1},\ldots P_m\}$ must then be learned by the network, the standard way of handling the situation is to mix the original set of patterns with the new set of patterns to be learned, creating a new set $\{P_1,\ldots P_n, P_{n+1},\ldots P_m\}$, and then train the network on this new expanded set. In this way, the new patterns will not interfere with the old patterns, but there is a major problem with this technique: namely, that in the real world, the originally learned patterns are often no longer available and cannot simply be added to the set of new patterns to be learned.

In 1995 Anthony Robins made a major contribution to research on catastrophic forgetting with a technique based on what he called 'pseudopatterns' (Robins, 1995). His idea was simple and elegant. Suppose that a connectionist network with n inputs and m outputs has learned a number of input–output patterns $\{P_1, P_2,\ldots P_N\}$ generated by some underlying function f. Assume that these original input–output vectors are no longer available. How could one determine, even approximately, what function the network had originally learned? One way would be to create a number M of random input vectors of length n, $\{\hat{\imath}_1,\ldots \hat{\imath}_M\}$. These pseudo-input vectors would be fed through the previously trained network, producing a corresponding set of outputs $\{\hat{o}_1,\ldots \hat{o}_M\}$. This would result in a set of 'pseudopatterns' $S = \{\psi_1, \psi_2,\ldots \psi_M\}$ where ψ_k: $\hat{\imath}_k \rightarrow \hat{o}_k$. This set of pseudopatterns would approximate the prior learning of the network. The accuracy of the pseudopatterns in describing the originally learned function would depend on the nature of the function. Thus, when the network had to learn a new set of patterns, it would mix in a number of pseudopatterns with the new patterns.

The pseudopattern technique was the basis of the dual memory models developed by French (1997) and Ans and Rousset (1997), which loosely

simulate the hippocampal–neocortical separation, considered by some to be the brain's way of overcoming catastrophic interference (McClelland *et al*., 1995). These models incorporate two separate, continually interacting, pattern processing areas, one for early processing and one for long-term storage, information being passed between the areas by means of pseudopatterns. This technique allows them to forget gradually and to perform sequential learning appropriately. Somewhat unexpectedly, these dual memory networks also exhibit over time a gradual representational 'compression' (i.e., fewer active nodes) of the long-term internal representations. If this can be shown to occur also in humans, it might help explain certain types of category-specific deficits commonly observed in amnesiacs (French and Mareschal, 1998).

OTHER TECHNIQUES FOR ALLEVIATING CATASTROPHIC FORGETTING IN NEURAL NETWORKS

A number of other techniques have been developed to address the problem of catastrophic interference. Notably, there have been attempts to combine auto-associative architectures with sparse representations. Some architectures use two different kinds of weights on the connections between nodes, one that decays rapidly to zero and another that decays much more slowly. Convolution-correlation models such as CHARM and TODAM, which are mathematically equivalent to certain types of connectionist networks (sigma–pi networks) seem to be relatively immune to catastrophic interference, at least up to a point. Cascade-correlation learning algorithms have also been tried as a means of alleviating catastrophic interference, with some success. For a more complete review of the various models that have been developed to handle the problem of catastrophic interference in connectionist networks, see (French, 1999).

SUMMARY

The problem of catastrophic interference in connectionist networks has been known and studied since the early 1990s. Sequential learning of the kind done by humans cannot be achieved unless a solution is found to this problem. In other words, network models of cognition must, as Grossberg has stressed, be sensitive to new input but not so sensitive that the new input destroys previously learned information. Certain types of patterns, such as those found in highly structured domains, are less susceptible to catastrophic interference than patterns from less well structured domains. Nature seems to have evolved a way of keeping new learning (hippocampal learning) at arm's length from previously learned information stored in the neocortex (neocortical consolidation), thus physically preventing new learning from interfering with previously learned information. Connectionist models have been developed that simulate this cerebral separation. This is certainly not the only way to tackle the problem of catastrophic interference, but its close relationship with the way in which the brain may have solved the problem makes further exploration of these dual memory models of particular interest.

References

Ans B and Rousset S (1997) Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Academie des Sciences: Sciences de la Vie* **320**: 989–997.

Barnes J and Underwood B (1959) 'Fate' of first-learned associations in transfer theory. *Journal of Experimental Psychology* **58**: 97–105.

Brousse O and Smolensky P (1989) Virtual memories and massive generalization in connectionist combinatorial learning. In: *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 380–387. Hillsdale, NJ: Erlbaum.

Carpenter G (1994) A distributed outstar network for spatial pattern learning. *Neural Networks* **7**: 159–168.

Carpenter G and Grossberg S (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing* **37**: 54–115.

French RM (1992) Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science* **4**: 365–377.

French RM (1997) Pseudo-recurrent connectionist networks: an approach to the 'sensitivity–stability' dilemma. *Connection Science* **9**: 353–379.

French RM (1999) Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences* **3**(4): 128–135.

French RM and Mareschal D (1998) Could category-specific anomia reflect differences in the distributions of features within a unified semantic memory? In: Gernsbacher A and Derry SJ (eds) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 374–379. Hillsdale, NJ: Erlbaum.

Grossberg S (1982) *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Boston, MA: Reidel.

Hetherington P and Seidenberg M (1989) Is there 'catastrophic interference' in connectionist networks? In: *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 26–33. Hillsdale, NJ: Erlbaum.

Kortge C (1990) Episodic memory in connectionist networks. In: *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 764–771. Hillsdale, NJ: Erlbaum.

Krushke J (1992) ALCOVE: an exemplar-based model of category learning. *Psychological Review* **99**: 22–44.

Lewandowsky S (1991) Gradual unlearning and catastrophic interference: a comparison of distributed architectures. In: Hockley W and Lewandowsky S (eds) *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*. pp. 445–476. Hillsdale, NJ: Erlbaum.

McClelland J, McNaughton B and O'Reilly R (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–457.

McCloskey M and Cohen N (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower GH (ed.) *The Psychology of Learning and Motivation*, vol. XXIV, pp. 109–164. New York, NY: Academic Press.

McRae K and Hetherington P (1993) Catastrophic interference is eliminated in pretrained networks. In: *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 723–728. Hillsdale, NJ: Erlbaum.

Murre J (1992) *Learning and Categorization in Modular Neural Networks*. Hillsdale, NJ: Erlbaum.

Ratcliff (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review* **97**: 285–308.

Robins A (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science* **7**: 123–146.

Further Reading

French RM (1999) Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences* **3**(4): 128–135.

Hetherington P (1991) *The Sequential Learning Problem*. Master's thesis, McGill University.

Classifier Systems

Intermediate article

Lashon B Booker, MITRE Corporation, McLean, Virginia, USA

CONTENTS

A classifier system is a parallel, message-passing, rule-based system designed to learn and use internal models of complex environments. Classifier systems are useful to cognitive modelers because they build representations that have both connectionist and symbolic qualities.

OVERVIEW

Real-world environments seldom provide salient, timely, complete, and unambiguous information. Therefore the correspondence between the unfolding complexity of the world and the representations used by a cognitive system cannot be taken for granted. Environmental properties, particularly complexity and uncertainty, are an important constraint on cognitive behavior. A complex environment may overwhelm a system with large amounts of information, not all of which is directly relevant to the system's appointed task. In order to function at all, a cognitive system must be discriminating and selective about what information it stores and uses. An uncertain environment is one in which it is unlikely that input configurations can be discerned or predictions made with accuracy. Under such circumstances, the only viable information-processing strategies are those capable of making pragmatic 'good guesses' about the true state of the world. In order to cope with both complexity and uncertainty, one must resist the temptation to try to know the environment in explicit detail. A system must focus on learning the basic concepts and regularities in the environment, their relationships, and their relevance to system goals. It is precisely this kind of economical, orderly arrangement of knowledge that constitutes an internal model of the environment. The notion that organisms can benefit from the use of internal models has long been recognized by psychologists as a powerful idea (Craik, 1943; Tolman, 1948).

Inductive processes generate and revise the constituent elements of internal models. Learning that leads to the development of an internal model can therefore be viewed as a pragmatic cognitive strategy for successful functioning in the real world. For this reason, induction is a central topic in cognitive science. In a comprehensive discussion of induction from this perspective, Holland *et al*. (1986), viewing induction broadly as any inferential process that expands knowledge in the face of uncertainty, describe a rule-based framework that seeks to answer the question: 'How can a cognitive system process environmental input and stored knowledge so as to benefit from experience?' Classifier systems are general-purpose rule-based systems designed to learn and use internal models in a manner consistent with this framework.

CLASSIFIER SYSTEMS AND COGNITIVE MODELING

Several properties of this framework for induction, and classifier systems in particular, are well suited to support cognitive modeling. In order to see why, it is helpful to begin with a broad characterization of internal representations and the mental operations that use them. The discussion by William James (1892) remains one of the most useful such characterizations available. James distinguishes two important aspects of cognitive representations and processes that are relevant here. Firstly, internal representations are dynamic, composite descriptions of the current situation and the expectations associated with it. The constituent elements of representations are associated with the many different aspects of the current stimuli and the overall context. The simultaneous, context-dependent activation of some combination of these elements constitutes the system's interpretation of the current situation. Secondly, the various elements that are candidates to participate in a representation interact with each other to determine which elements become active. The nature of these interactions is determined by the network

of associations connecting the elements together, and by mechanisms that direct the flow of activity from one element to another.

This notion of distributed representations emerging from dynamic patterns of interactions among large numbers of primitive elements is common to many connectionist approaches to cognitive science. The framework for induction proposed by Holland *et al*. takes a similar notion of representation, but uses simple condition–action rules as the basic elements. Rules interact by passing messages, and many rules can be active simultaneously. Inductive mechanisms organize rules into clusters that provide a multifaceted representation of the current situation and the expectations that flow from it. The structure of a concept is modeled by the organization, variability, and strength of the rules in a cluster. Simple rules thereby become building blocks for representing complex concepts, constraints, and problem-solving behaviors. Knowledge can be represented at many levels of organization, using rules and rule clusters as building blocks of different sizes and complexities. Because constituent elements compete to become active, aspects of a representation are selected only when they are relevant in a given context. Moreover, since rules are activated in parallel, new combinations of existing rules and rule clusters can be used to dynamically represent novel situations. (*See* **Connectionism**; **Distributed Representations**)

Because the framework for induction uses symbolic rules as representational primitives, there are important differences between this approach and the typical connectionist approach to cognitive modeling. The most obvious difference is that it is easier to construct representations having symbolic expressiveness. Another significant difference is in the way inductions are achieved. Modification of connection strengths is the only inductive mechanism available in most connectionist systems. Moreover, the procedures for updating strength are part of the initial system design, and cannot be changed, except perhaps by adjusting a few parameters. The framework for induction, on the other hand, permits a range of inductive mechanisms, from strength adjustments to analogies. In principle, many of these mechanisms can be controlled by or easily expressed in terms of condition–action rules. These rules can be evaluated, modified and used to build higher-level concepts. Overall, this paradigm offers a set of potential cognitive modeling capabilities that occupy an important middle ground between subsymbolic connectionist systems and the traditional symbolic paradigms of artificial intelligence.

Classifier systems were originally proposed as a rule-based implementation of this approach to cognitive modeling. The earliest proposals (Holland, 1976) and experiments (Holland and Reitman, 1978) described simple cognitive systems designed to adapt to their environment under sensory guidance. Most subsequent work has focused less on developing cognitive architectures and more on the algorithms and computational techniques needed to implement architectures of this kind. The remainder of this article discusses these algorithms and their properties from the standpoint of cognitive modeling.

ALGORITHMIC DESCRIPTION OF CLASSIFIER SYSTEMS

A classifier system is a parallel, message-passing, rule-based system designed to permit nontrivial modifications and reorganizations of its knowledge as it performs a task. The typical operating principles for a classifier system can be briefly summarized as follows. In the simplest version, all information processing is accomplished using messages that are encoded as binary strings of fixed length k. The set of messages to be processed at any given moment is stored on a 'message list'. Classifier systems process these messages using a population of rules, called 'classifiers'. Every classifier is a fixed-length string having the form

$$s_1, s_2, \ldots, s_n \Rightarrow m \qquad (1)$$

Each s_i is an input condition, represented as a string of length k in the ternary alphabet {0, 1, #}. The action part m is also a string of length k in this ternary alphabet. A condition is satisfied by a message whenever there is a 'match' – that is, whenever the 0 and 1 symbols in the condition are identical to the bit values at corresponding positions of the message. The # symbol is a 'don't care' place holder that matches any message bit value at the designated position. A classifier is eligible to become active only if each of its conditions matches a message on the message list. Once a classifier is activated, the action part of the rule specifies a message m which the rule will 'post' on the message list on the next timestep. When the # symbol appears in the action part it is interpreted as a 'pass through' place holder. This means that the designated position in the posted message is assigned the value of the corresponding bit from a message that matched one of the classifier's conditions.

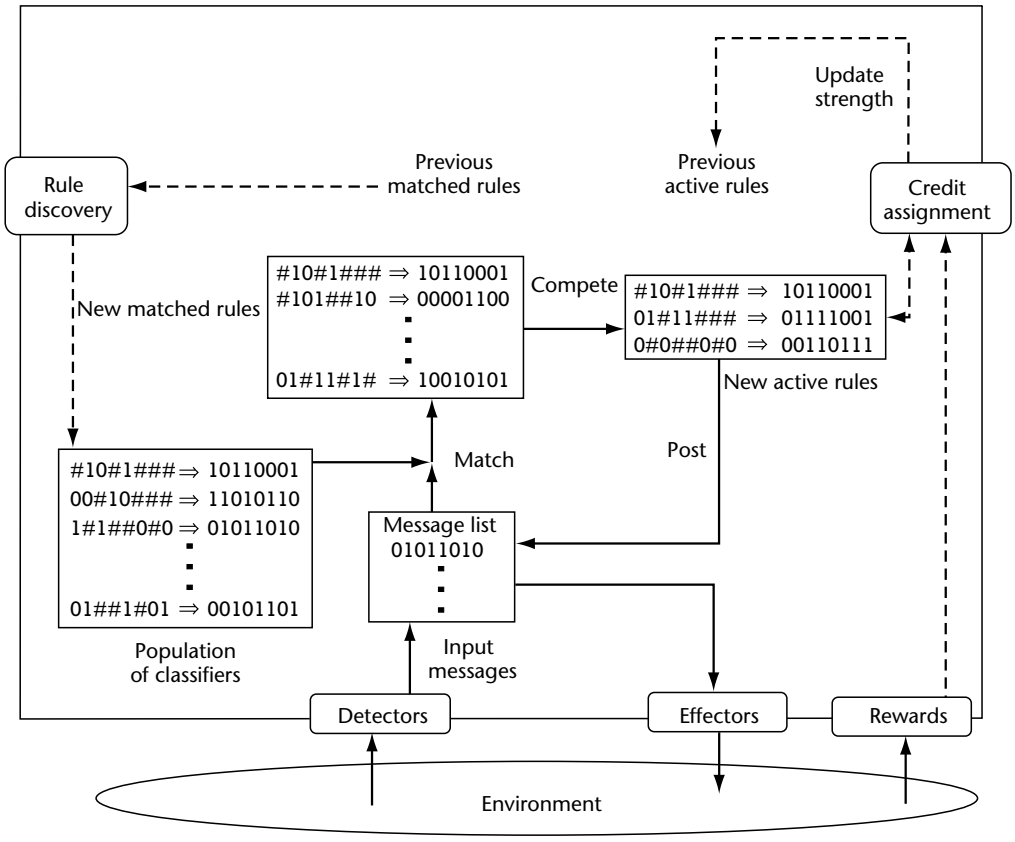


Figure 1. The organization of a typical classifier system. See the text for a detailed explanation.

In more detail, the typical classifier system can be described in terms of three interacting subsystems: a 'performance' system, a 'credit assignment' system, and a 'rule discovery' system (see Figure 1). The performance system is responsible for interacting with the external environment and generating behavior. It is assumed that an input interface is available to translate the state of the environment into messages (e.g. by using a set of detectors) and that an output interface interprets relevant messages as action specifications (e.g. by using a set of effectors). It is also assumed that the environment occasionally gives the system explicit performance feedback in the form of pay-off or reinforcement. Given these assumptions, the basic execution cycle for the performance system proceeds as follows:

1. Place messages from the input interface on the current message list.
2. Compare all messages to all conditions and conduct a competition among matching classifiers to determine which ones will become active.
3. For each active classifier, generate one message for the new message list.
4. Replace the current message list with the new message list.
5. Process the current message list through the output interface to produce system output.
6. Return to step 1.

Two important aspects of the performance system are worth noting here. Firstly, the system is highly parallel. Many classifiers can be active simultaneously. Since different classifiers match different subsets of messages, the classifiers can be thought of as building blocks which can be activated in many combinations to represent a variety of situations. Secondly, matching classifiers must compete to become active. This competition makes it possible to avoid imposing any consistency requirements on posted messages, and allows the system to insert new rules smoothly without disrupting existing capabilities. Classifiers are treated as tentative hypotheses about the effects of posting a message given the current conditions. Each of these 'hypotheses' is repeatedly tested according to the system's experience with the environment.

The competition mechanism assumes that a reliable assessment of the plausibility of each

hypothesis is available. The credit assignment system is responsible for computing those assessments. It is difficult to assign credit for successful problem-solving behavior in a system that uses many rules over several timesteps to generate behavior. The difficulty is even more pronounced when overt feedback from the environment is intermittent or rare. The only realistic strategy for classifier systems is to evaluate performance for behavioral sequences in local terms for each classifier involved. The most straightforward approach is to rely on a simple reinforcement principle: strengthen a hypothesis whenever the associated message leads to a favorable or rewarding situation. A considerable amount of theoretical work in psychology, adaptive control, and machine learning has used this principle as a starting point for understanding how to solve difficult credit-assignment problems. One computational approach designed for classifier systems is the 'bucket brigade' algorithm (Holland *et al*., 1986), one of many algorithms from the reinforcement learning literature that are suitable for this purpose. (*See* **Reinforcement Learning: A Computational Perspective**)

Credit assignment is only one of several learning mechanisms that are needed for a successful inductive process. It is also necessary to have a mechanism that can generate plausible new hypotheses. For rule-based systems this means generating new rules as candidates to replace existing rules that have not been particularly useful. In the classifier system framework, the rule discovery system is responsible for generating plausible new classifiers. 'Plausibility' in this context is tied to the notion of 'building blocks' mentioned in the discussion of parallelism above. The simple syntax used in classifier conditions makes it straightforward to view classifiers as strings composed of readily identifiable parts. One easily-specified and useful set of parts can be defined as follows. In each classifier condition and action, there are many 'components' which can be identified by specific combinations of symbols at designated positions. For example, the condition `0#011### ... 1#` begins with a `0` in position 1 and a `#` in position 2. This combination defines a template, or 'substring schema', which can be denoted by the string `0#****** ... **` (where the symbol `*` indicates positions not involved in the definition). Any condition beginning with this two-symbol prefix contains the corresponding substring schema as one of its parts. The utility of such a part can be estimated by the average strength of the rules containing that part. A new rule can be considered 'plausible' to the extent that it is composed of parts, or building blocks, having above-average utility. Thus, the rule-discovery process can generate plausible new rules by favoring above-average building blocks in the construction of new rules.

Note that every condition or action of length k uses 2^k distinct building blocks. Even given a moderately sized population of classifiers, it is computationally infeasible to evaluate explicitly and use information about all of the building blocks in all of the classifier conditions and actions. It is possible, however, to devise procedures that implicitly make use of this information. The mechanism used most often for this purpose in classifier systems is a 'genetic algorithm'. A genetic algorithm is a general-purpose search procedure that uses sample-based induction (Holland *et al*., 1986) to conduct the search. The algorithm repeatedly selects a sample of rules (the 'parents') from the population and recombines their building blocks to construct new rules (the 'offspring'). The new elements are constructed using genetic operators such as recombination and mutation. The selection criterion for parents is biased to favor high-strength classifiers, and new classifiers replace low-strength classifiers in the current population. (*See* **Evolutionary Algorithms**)

REPRESENTATIONS IN CLASSIFIER SYSTEMS

The starting point for representing knowledge in the classifier system framework is the use of simple message-passing rules as primitive elements. Each rule condition specifies a basic equivalence class or category of messages that match the condition. Because of parallelism and recombination, higher-level representations built using these primitives have an inherently composite nature. Rather than construct a syntactically complex representation of a symbolic concept that might be difficult to use or modify, a classifier system is designed to use clusters of rules as representations. The implementation of this approach to representing knowledge in classifier systems relies on principles and mechanisms operating at two levels of organization: rules and rule clusters. Each rule offers a range of representational power that is determined by the way messages are matched, encoded and processed. The representational power of rule clusters is determined by the way rules are organized in relation to each other and by their competitive and cooperative interactions. These issues are discussed briefly below.

Rule Syntax, Message Encoding and Matching

An input message is a string of feature values or primitive attributes describing some state configuration in the environment. Condition–action rules are a convenient way to represent states and transitions between states in the environment, generate predictions, and specify simple procedures. Rule conditions provide generalizations of messages that correspond to useful regularities and attribute-based concepts. These generalizations are the lowest-level building blocks available for constructing new rules. The generalizations that are possible in a classifier system depend on the classifier rule syntax and on the way messages are encoded and matched.

In the basic classifier 'language', conditions are fixed-length ternary strings that correspond to generalizations given by simple conjunctions of attribute values. Because the syntax is so simple, no extra mechanisms are needed to specialize or generalize these conditions. Specialization only requires changing a # to a 0 or 1, while generalization is accomplished by changing a 0 or 1 to a #. Changes of this kind are routinely generated by the genetic algorithm in the rule discovery system. Note, however, that simple classifier conditions of this kind cannot be used to express arbitrary, general relationships among structured attributes. There are two ways to increase the expressive power of these classifier conditions.

First, by using multiple conditions and multiple rules, it is possible to represent concepts involving the logical conjunction and disjunction of simple clauses. A classifier with multiple conditions can be activated only when each condition matches a message on the message list. The posting of that classifier's message therefore indicates an input configuration in which the conjunction of those conditions is true. When two or more classifiers have different conditions but identical actions, the posting of that particular message indicates an input configuration in which at least one of the conditions is true (disjunction) Additional expressive power is obtained by allowing some conditions to be satisfied when no matching message is on the list. This provides a way to represent logical negation, making it possible to represent arbitrary Boolean expressions.

The second way to increase the expressive power of individual classifiers is to use more powerful encoding schemes for messages. In the simplest encoding, each message bit corresponds to a single attribute–value pair or predicate. This encoding is adequate for categories defined by a set of critical features whose presence or absence is mandatory for category membership. More sophisticated encodings support generalizations about the range of ordinal values and disjunctions of nominal values. These encodings provide capabilities that compare favorably to the primitive language constructs used in many symbolic learning paradigms. Moreover, the basic building blocks supporting these generalizations can be easily recombined and otherwise manipulated by the genetic algorithm and other local, syntactic rule modification algorithms.

Even with more expressive encodings, however, there are limits to what can be represented with conditions using the standard syntax. For example, although continuous input values can be represented as bit strings, classifier conditions can only represent a subset of the generalizations that may be useful or required. Moreover, the standard matching algorithm does not allow for generalizations that require variable binding or parametrization. This means that it is not possible to express simple relations between subfields in a single condition or across multiple conditions. In order to overcome these limitations, some classifier system implementations have utilized more sophisticated symbolic expressions in rule conditions. This can be problematic, though, because it is not always clear what the buildings blocks are in these complex representations. One of the important open research questions is how to increase the expressive power of classifier conditions without sacrificing the efficiency and well-chosen building blocks characteristic of the basic classifier language syntax.

Multiple Rules and Default Hierarchies

As noted above, the expressive power of rule conditions can be enhanced by using a set of rules to represent disjunctive clauses in the description of an attribute-based concept. This is a form of implicit cooperation among rules, whereby different subsets of rules are responsible for different aspects of the overall representation. Many forms of cooperative interactions among rules are possible.

For example, another useful form of implicit cooperative interaction is tied to the competition mechanism. The competition for the right to post messages is usually implemented with a bias towards selecting the 'specific' rules whose conditions contain the most detail about the current situation. This makes it possible for rules to become organized into 'default hierarchies'. The simplest example of a rule-based default hierarchy consists

of two rules. The first ('default') rule has a relatively general condition and provides an action that is sometimes incorrect. The second (exception) rule is satisfied only by a subset of the messages satisfying the default rule, and its action generally corrects errors committed by the default. That is, the exception rule uses additional information (its more specific condition) to distinguish situations that lead the more general default rule astray. When the exception wins the competition it prevents the default from making a mistake and losing strength under the credit assignment mechanism. There is consequently a kind of symbiotic relationship between the two rules. Note that the default may have other exceptions, and each exception may, in turn, have exceptions, resulting in a hierarchy of interactions. Default hierarchies are an important feature of classifier systems from the standpoint of cognitive modeling. Classifier systems with default hierarchies have been used to model a wide variety of conditioning phenomena (Holyoak *et al*., 1990) and human performance on simple discrimination tasks (Riolo, 1991).

Another useful implicit interaction among rules is based on patterns of conditions and actions. If the action posted by rule R_1 matches the condition of another rule R_2, then the activation of R_1 will tend to result in the activation of R_2. In this case the two rules are said to be 'coupled'. Since rule coupling implies sequential activation, classifier systems can use representations in which the constituent rules are activated over more than one match–compete–post execution cycle. This is the first step needed to go beyond simple reactive behavior to a mode in which internal information processing can be used to generate responses. Only a few classifier systems have exploited implicit coupling for cognitive modeling purposes. Booker (1988) describes a system that learns a simple internal model in which the goal-relevance of environmental states can be retrieved from memory using internal messages.

Complex Knowledge Structures

The coupling mechanism becomes even more useful when the rule interactions are organized explicitly. Particular bits incorporated into a rule's condition – such as a suffix or prefix – can be used as a kind of identifier or address, called a 'tag'. Messages can be directed to a rule explicitly by incorporating the appropriate tag into those messages. If there are several rules sensitive to the same tag, then they will be activated together as a cluster. Consequently, any rule that contains that tag in its action part will be explicitly coupled to the entire cluster. Tags are the building blocks for representing sequences and associations. Because tags are simple parts of messages and rules, associations among rules can be constructed and modified with the standard repertoire of rule discovery mechanisms.

Moreover, a tag can be given semantics as a label indicating the origin (e.g. the input interface) or destination (e.g. the output interface) of a message. By identifying tags with semantic content, we can impose a hierarchical organization on classifier representations. Tags become identifiers for relationships within and between levels in a hierarchy, and for relationships between hierarchies. In principle, the use of semantically useful tags to couple rules makes classifier systems a powerful framework for representing complex knowledge structures. For instance, it has been shown (Forrest, 1991) that the knowledge contained in a standard semantic network description (e.g. KL-ONE or NETL) can be mapped into a set of classifiers that support the same information-retrieval operations. It remains to be seen whether these kinds of complex knowledge structures can be learned by a classifier system as a result of its experiences in an environment. However, experiments have demonstrated that classifier systems can learn simple associative knowledge structures (Riolo, 1990; Stolzmann and Butz, 2000). (*See* **Semantic Networks**)

PROBLEM SOLVING USING CLASSIFIER SYSTEMS

Problem solving in the classifier systems framework differs from more conventional approaches to problem solving in that a strong emphasis is placed on flexibly modeling the problem-solving context. The rationale for this emphasis on flexible problem representations is that cognitive systems are often confronted with problems that are poorly defined – that is, various aspects of the initial problem specification may be unknown or partially known. Instead of relying on stand-alone procedures for reformulating such problems, classifier systems can use building blocks, together with the simultaneous activation and combination of multiple representations, to recategorize the problem components. In this way, the system conducts a search for solutions to a problem both in the problem space and in the space of alternative problem representations.

The starting point for this approach is a good model that allows for prediction-based evaluation of the knowledge base, and the assignment of credit

to the model's building blocks. This makes it possible to modify, replace, or add to existing rules via inductive mechanisms such as the recombination of highly rated building blocks. After repeated experiences solving instances of a problem, the inductive mechanisms generate rules which are specialized or generalized as needed to adequately represent the typical elements of the problem space. Maintaining a varied repertoire of useful specific and general rules is essential to achieving problem-solving flexibility. When a novel situation is encountered, there is some chance that it will be matched by general rules that provide some useful, though perhaps imperfect, guidance about how to proceed. The simultaneous activation of specific and general rules defines an implicit default hierarchy in which default expectations can be overridden whenever a specific exception occurs. Novel situations can also be handled by activating associations that suggest recategorizations of structured concepts and relations. By coordinating multiple sources of knowledge, hypotheses and constraints in this way, problem solving can proceed opportunistically, guided by the integration of converging evidence and building on weak or partial results to arrive at confident conclusions.

In simple cases, problem solving can be achieved by learning a direct mapping between inputs and outputs. A large body of research within the reinforcement learning paradigm has been devoted to solving Markovian decision tasks using this strategy, and classifier systems can achieve comparable results. However, classifier systems used in this way are simple reactive systems that do not require any cognitive processing beyond basic categorization. One particularly sophisticated problem-solving mechanism that has been implemented in a classifier system is based on Baum's (Baum and Durdanovic, 2001) approach, which views rule clusters as 'post-production systems'. The system learns algorithms that solve instances of Rubik's cube and arbitrary block-stacking problems. (*See* **Markov Decision Processes, Learning of**)

SUMMARY

Classifier systems can be thought of as connectionist systems that use rules as the basic epistemic unit. Using simple rules in this way makes it possible to enjoy the advantages of distributed representations, and at the same time to represent nontrivial symbolic concepts and employ flexible problem-solving mechanisms. Consequently, classifier systems have the potential to occupy an important middle ground between the symbolic and connectionist paradigms. In order to realize this potential, more work must be done to understand how classifier systems might dynamically construct and modify the kinds of multifaceted representations described here. Research on small systems and simple problems has been a promising first step towards that goal.

References

Baum EB and Durdanovic I (2001) An artificial economy of post production systems. In: Lanzi PL, Stolzmann W and Wilson SW (eds) *Advances in Learning Classifier Systems*, pp. 3–20. Berlin, Germany: Springer.

Booker LB (1988) Classifier systems that learn internal world models. *Machine Learning* **3**: 161–192.

Craik KJW (1943) *The Nature of Explanation*. Cambridge, UK: Cambridge University Press.

Forrest S (1991) *Parallelism and Programming in Classifier Systems*. London, UK: Pitman.

Holland JH (1976) Adaptation. In: Rosen R and Snell FM (eds) *Progress in Theoretical Biology*, vol. IV, pp. 263–293. New York, NY: Academic Press.

Holland JH, Holyoak KJ, Nisbett RE and Thagard PR (1986) *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Holland JH and Reitman JS (1978) Cognitive systems based on adaptive algorithms. In: Waterman DA and Hayes-Roth F (eds) *Pattern-Directed Inference Systems*, pp. 313–329. New York, NY: Academic Press. [Reprinted in: Fogel B (ed.) (1998) *Evolutionary Computation: The Fossil Record*. IEEE Press.]

Holyoak KJ, Koh K and Nisbett RE (1990) A theory of conditioning: inductive learning within rule-based default hierarchies. *Psychological Review* **96**: 315–340.

James W (1892) *Psychology: The Briefer Course*. New York, NY: Henry Holt. [Reprinted edition edited by G. Allport, published by Harper & Row, New York, 1961.]

Riolo RL (1990) Lookahead planning and latent learning in a classifier system. In: Meyer JA and Wilson SW (eds) *From Animals to Animats 1. Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pp. 316–326. Cambridge, MA: MIT Press/ Bradford Books.

Riolo RL (1991) Modelling simple human category learning with a classifier system. In: Booker LB and Belew RK (eds) *Proceedings of the 4th International Conference on Genetic Algorithms (ICGA91)*, pp. 324–333. San Mateo, CA: Morgan Kaufmann.

Stolzmann W and Butz M (2000) Latent learning and action-planning in robots with anticipatory classifier systems. In: Lanzi PL, Stolzmann W and Wilson SW (eds) *Learning Classifier Systems: From Foundations to Applications*, pp. 301–317. Berlin, Germany: Springer.

Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* **55**: 189–203.

Further Reading

Booker LB, Goldberg DE and Holland JH (1989) Classifier systems and genetic algorithms. *Artificial Intelligence* **40**: 235–282.

Booker LB, Riolo RL and Holland JH (1994) Learning and representation in classifier systems. In: Honavar V and Uhr L (eds) *Artificial Intelligence and Neural Networks*, pp. 581–613. San Diego, CA: Academic Press.

Donnart J-Y and Meyer JA (1996) Learning reactive and planning rules in a motivationally autonomous animat. *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics* **26**(3): 381–395.

Holland JH (1986) A mathematical framework for studying learning in a classifier system. In: Farmer D, Lapedes A, Packard N and Wendroff B (eds) *Evolution, Games and Learning: Models for Adaptation in Machines and Nature*, pp. 307–317. Amsterdam: North-Holland.

Holland JH (1990) Concerning the emergence of tag-mediated lookahead in classifier systems. *Physica D* **42**: 188–201. [Republished in *Emergent Computation*, edited by S. Forrest, MIT Press/Bradford Books.]

Lanzi PL, Stolzmann W and Wilson SW (eds) (2000) *Learning Classifier Systems: From Foundations to Applications*. Berlin: Springer.

Lanzi PL, Stolzmann W and Wilson SW (eds) (2001) *Advances in Learning Classifier Systems*. Berlin, Germany: Springer.

Stolzmann W, Butz M, Hoffmann J and Goldberg DE (2000) First cognitive capabilities in the anticipatory classifier system. In: Meyer JA *et al*. (eds) *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, pp. 287–296. Cambridge, MA: MIT Press.

Wilson SW (1995) Classifier fitness based on accuracy. *Evolutionary Computation* **3**(2): 149–175.

Wilson SW and Goldberg DE (1989) A critical review of classifier systems. In: Schaffer JD (ed.) *Proceedings of the 3rd International Conference on Genetic Algorithms (ICGA89)*, pp. 244–255. San Mateo, CA: Morgan Kaufmann.

Cognitive Processing Through the Interaction of Many Agents

Intermediate article

Chris Jones, University of Southern California, Los Angeles, California, USA
Maja Matarić, University of Southern California, Los Angeles, California, USA
Barry Werger, Jet Propulsion Laboratory, Pasadena, California, USA

CONTENTS

A collection of interacting autonomous entities, called 'agents', may be capable of creating complex cognitive processes and behaviors, which could not be achieved by a single agent, without the need for outside centralized coordination or control.

INTRODUCTION

Several theories of cognition, most notably Minsky's 'society of mind', posit that intelligent behavior can be seen as the result of the interaction of simple processes. Minsky states: 'Very few of our actions and decisions come to depend on any single mechanism. Instead, they emerge from conflicts and negotiations among societies of processes that constantly challenge one another' (Minsky, 1986). The central tenet of such theories of cognition and behavior is that complex system-level behavior can emerge from the interaction of multiple, possibly numerous, components.

A canonical example of such emergence is the function of the human brain. The brain itself is made up of billions of simple neurons organized into a massively connected network (Nicholls *et al*., 2001). In general, an individual neuron acts as a comparatively simple processing unit that receives signals from a set of neighboring input neurons, and under appropriate conditions transmits signals to a set of its neighboring output neurons. From this network of interacting neurons emerges the complexity of human cognition and behavior. No single neuron or subset of neurons is responsible for this complexity; rather, it is the result of their interactions.

Several disparate research fields are actively involved in investigating the principles of interaction among a collection of components. These include cognitive science, computer networks, distributed systems, artificial life, collective robotics, multi-agent systems, as well as others. The rest of this article aims to explain conceptually, and show through examples, how complex cognition and behavior can emerge from the interaction of individual components and how those emergent behaviors can be used in a variety of ways.

AGENTS

The term 'agent' has become a popular choice for a nontrivial component of a system with many interacting components that result in emergent behavior. Precisely defining an agent remains difficult, as agents come in many guises. An agent could be a piece of software, a specific computer on the Internet, a mobile robot, or even a person. In general, an agent is an autonomous entity, situated in an environment, and equipped with some degree of intelligence.

Being 'situated' places a number of constraints on how the agent can operate (Brooks, 1991; Maes, 1990). It implies that the agent has some means of sensing its environment, but the sensing may be limited and inaccurate. For example, a mobile robot may be situated in an office environment and have a means of sensing the distance to nearby objects; thus its sensing capabilities are both limited and prone to error and noise. Situatedness also implies an interaction with the environment. Thus, the same robot may be able to drive around, affecting the environment with its placement, and perhaps move objects, intentionally or otherwise. Conversely, the actions of a situated agent are influenced by the environment. The objects the robot encounters affect what it senses and how it behaves. Another implication of being situated is that an agent is constrained by environmental

characteristics. For example, a mobile robot cannot drive through walls nor can it avoid falling if it drives down a set of stairs. Finally, the agent's characteristics – its computational, sensory, and actuation capabilities – influence how it interacts with its environment, which includes other agents.

SOCIETIES OF AGENTS

A collection of interacting agents is referred to as a 'society of agents'. Using this metaphor, a brain is a society of agents, as is a team of mobile robots cooperating on a task. Such agent societies are interesting for a number of reasons. First, for certain tasks and/or environments, a society of agents is the only viable or efficient solution. Second, even for tasks that can be handled by an individual agent, there may be more efficient, adaptive, and robust solutions performed by a society of agents.

A society of agents may consist of a homogeneous or heterogeneous collection of agents. In a homogeneous society, all agents are identical, while in a heterogeneous society, agents may have different characteristics. The variations in capabilities may result in hierarchies, specializations, or various other forms of social organization. Consequently, heterogeneous societies are generally more complex to control but are typically capable of a larger set of tasks.

The human immune system (Segel and Cohen, 2001) is an excellent example of a society of heterogeneous yet simple agents. It is capable of protecting the body against infection and invasion by foreign substances of all kinds, whether bacteria, virus, parasite, etc., which can be viewed as a very large set of different defensive 'tasks' to be accomplished. Each agent in this society is very specific, but the large number of agents of each kind and in total, combined with the ability to generate additional agents when needed, produces an unprecedented defensive functionality.

AGENT INTERACTION

Agents situated in a shared environment have ample opportunity to interact, by directly sensing each other, communicating, coordinating actions, and even competing. The shared environment in which the interactions take place may be an abstract data space, the physical world, or anything in between. A society of software agents may interact through a personal computer or even the entire Internet. Likewise, a society of mobile robots may interact in an office environment or on the surface of Mars. Furthermore, multiple environment types may be spanned in a single multi-agent system. For example, in multi-robot systems using behavior-based control (Matarić, 1994), individual robots are controlled by a collection of internal agents that interact through a computational environment, while the society of physical robots that contains them interacts in and through the physical world. The nature of the interactions in a society of agents depends upon such factors as agent capabilities, environmental constraints, and desired local and global behavior.

The spectrum of agent interaction is broad. At one end are methods that employ large numbers of simple, identical agents, connected together in patterns which lead to useful computation as a result of data flow through the system. At the other end are systems of complex, specialized agents which explicitly negotiate for task assignments and resources. Mechanisms for agent interaction can be broadly classified as fitting into the following, often overlapping categories: interaction through the environment, interaction through sensing, and interaction through communication. Each is described in turn.

Interaction Through the Environment

The first mechanism for interaction among agents is through their shared environment. This form of interaction is indirect in that it consists of no explicit communication or physical interaction between agents. Instead, the environment itself is used as a medium of indirect communication. This is a powerful method of interaction that can be utilized by very simple agents with no capability for complex reasoning or for direct communication.

Stigmergy is an example of interaction through the environment employed in a variety of natural insect societies. Originally introduced in the biological sciences to explain some aspects of social insect nest-building behavior, stigmergy is defined as 'the process by which the coordination of tasks and the regulation of constructions does not depend directly on the workers, but on the constructions themselves' (McFarland, 1985; Holland and Melhuish, 1999). The notion was originally used to describe the nest-building behavior of termites and ants (Franks and Deneubourg, 1997). It was shown that the coordination of building activity in a termite colony was not inherent in the termites themselves. Instead, the coordination mechanisms were found to be regulated by the task environment, in this case the growing nest structure. A location on the growing nest

stimulates a termite's building behavior, thereby transforming the local nest structure, which in turn stimulates additional building behavior of the same or another termite.

Examples of artificial systems in which agents interact through the environment include distributed construction (Bonabeau *et al*., 1994), sorting (Deneubourg *et al*., 1990), clustering (Beckers *et al*., 1994), optimization problems (Dorigo *et al*., 1999), object manipulation (Donald *et al*., 1993), analysis of network congestion (Huberman and Lukose, 1997), and phenomena such as the spread of computer viruses (Minar *et al*., 1998).

Interaction Through Sensing

The second mechanism for interaction among agents is through sensing. As described by Cao *et al*. (1997), interaction through sensing 'refers to local interactions that occur between agents as a result of agents sensing one another, but without explicit communication'. This form of interaction is also indirect, as there is no explicit communication between agents; however, it requires each agent to be able to distinguish other agents from miscellaneous objects in the environment.

Interaction through sensing can be used by an agent to model the behavior of another agent or to determine what another agent is doing in order to make decisions and respond appropriately. For example, flocking birds use sensing to monitor the actions of other birds in their vicinity in order to make local corrections to their own motion. It has been shown that effective flocking results from quite simple local rules followed by each of the birds in the society (flock), responding to the direction and velocity of the local neighbors (Reynolds, 1987). Such methods of interaction through sensing can be found in use in mobile robot flocking, following, and foraging (Matarić, 1995), robot soccer (Werger, 1999), robot formations (Fredslund and Matarić, 2002), and simulations of behaviorally realistic animations of fish schooling (Tu and Terzopoulos, 1994). Other applications of interaction through sensing include human-like physical or visual interaction between physical agents (Murciano and del R. Millan, 1996; Michaud and Vu, 1999; Nicolescu and Matarić, 2000), including the ability to understand and influence the motives of other physical agents (Breazeal and Scassellati, 1999; Ogata *et al*., 2000).

Interaction Through Communication

The third mechanism for interaction among agents is through direct communication. Unlike the first two forms of interaction described above, which were indirect, in interaction through communication agents may address other agents directly, either in a system-specific manner or through a standard agent communication protocol such as KQML (Finin *et al*., 1996) or CORBA (Vinoski, 1997). Such agent-directed communication can be used to request information or action from others or to respond to requests received from others. Communication may be task-related rather than agent-directed, in which case it is made available to all (or a subset) of the agents in the environment. Two common task-related communication schemes are blackboard architectures (Schwartz, 1995; Gelernter, 1991) and publish/subscribe messaging (Arvola, 1998). In blackboard architectures, agents examine and modify a central data repository; in publish/subscribe messaging, subscribing agents request to receive certain categories of messages, and publishing agents supply messages to all appropriate subscribers. In some domains, such as the Internet, communication is reliable and of unlimited range, while in others, such as physical robot interaction, communication range and reliability are important factors in system design (Arkin, 1998; Gerkey and Matarić, 2001).

FROM AGENT INTERACTION TO COGNITION AND BEHAVIOR

Given a society of interacting agents, how is complex system-level behavior achieved? Interaction among the society members is not sufficient in itself to produce an interesting or useful global result. In order for the interacting agents to produce coherent global behavior, there must be some overarching coordination mechanism that appropriately organizes the interactions in both space and time.

There are many coordination mechanisms by which to organize the various interactions among agents in order to produce coherent system-level behavior. Self-organization techniques are based on a 'set of dynamical mechanisms whereby structures appear at the global level of a system from interactions among its lower-level components. The rules specifying the interactions among the system's constituent units are executed on the basis of purely local information, without reference to the global pattern, which is an emergent property of the system rather than a property imposed upon the system by an external ordering influence' (Bonabeau *et al*., 1997). Methods such as genetic algorithms (Holland, 1975), machine learning techniques such as reinforcement learning (Sutton and Barto, 1998), and distributed constraint

satisfaction (Clearwater *et al*., 1991) can all be used to design agents and their interactions such that the resulting behavior meets desired system-level goals. Agents may also explicitly negotiate with each other for resources and task assignments in order to coordinate their behavior.

One such approach, employed in human as well as synthetic agent societies, is 'market-based' coordination, where individual agents competitively bid for tasks, which they must either complete or report as broken contracts. Auctions are a common coordination method in market-based techniques. In auctions, the most appropriate agents are continuously selected and (re)-assigned to various non-terminating roles (Tambe and Jung, 1999; Werger and Matarić, 2000). In contrast, more symbolic negotiation protocols based on distributed planning involve multiple stages, in which agents first share their plans, then criticize them, and finally update them accordingly (Bussmann and Muller, 1992; Kreifelt and von Martial, 1991; Lesser and Corkill, 1981). Game-theoretic approaches to negotiation have proven effective in situations where agents may be deceptive in their communication (Rosenschein and Zlotkin, 1994). In most complex models of negotiation-based coordination, agents reason about the beliefs, desires, and intentions of other agents, and influence those using specialized techniques (Brandt *et al*., 2000).

CONCLUSION

As was stated earlier, in systems where complex global behavior emerges from the interactions of a society of simple agents, as the complex function of the brain emerges from the interactions of a large society of neurons, the resulting complexity cannot be attributed to any single agent but instead to the interaction of all agents. Agents and their, often local, interactions with each other and with the environment generate the resulting global behavior of the system – no additional external coordination mechanism is needed. Interaction in the agent society can take place through several mechanisms, including interaction through the environment, through sensing, and through communication. In lieu of a central coordinator, a society of agents coordinates its interactions to produce desired system-level behavior through such mechanisms as self-organization, machine learning techniques, or more complex negotiation mechanisms.

The notion that complex behavior can arise from the interaction of simple agents has powerful and far-reaching implications. Many fields, ranging from biology, to artificial intelligence, computer networking, and business management, all find inspiration and motivation from these principles.

References

Arkin R (1998) *Behavior-Based Robotics*. Cambridge, MA: MIT Press.

Arvola C (1998) *Transactional Publish / Subscribe: The Proactive Multicast of Database Changes*. SIGMOD Conference.

Beckers R, Holland OE and Deneubourg JL (1994) *Proceedings, Artificial Life IV*. In: R Brooks and P Maes (eds), pp. 181–189. Cambridge, MA: MIT Press.

Bonabeau E, Theraulaz G, Arpin E and Sardet E (1994) The building behavior of lattice swarms. In: Brooks R and Maes P (eds) *Artificial Life* **4**: 307–312.

Bonabeau E, Theraulaz G, Deneubourg J-L and Camazine S (1997) Self-organisation in social insects. *Trends in Ecology and Evolution* **12**(5): 188–193.

Brandt F, Brauer W and Weiss G (2000) *Task Assignment in Multiagent Systems based on Vickrey-type Auctioning and Leveled Commitment Contracting*. Proceedings of the Fourth International Workshop on Cooperative Information Agents.

Breazeal C and Scassellati B (1999) *How to Build Robots that Make Friends and Influence People*. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-99), Kyongju, Korea.

Brooks RA (1991) *Intelligence without Reason*. Proceedings of the International Joint Conference on Artificial Intelligence, Sydney, Australia.

Bussmann S and Muller JA (1992) *Negotiation Framework for Cooperating Agents*. In: Deen SM (ed.) Proceedings CKBS-SIG, Dake Centre, University of Keele.

Cao Y, Fukunaga A and Kahng A (1997) Cooperative mobile robotics: Antecedents and directions. *Autonomous Robots* **4**: 7–27.

Deneubourg JL, Goss S, Franks, Sendova-Franks A, Detrain C and Chretien L (1990) *Proceedings, Simulation of Adaptive Behavior*, pp. 365–363, Cambridge, MA: MIT Press.

Donald BR, Jennins J and Rus D (1993) Proceedings, International Symposium on Robotics Research, Hidden Vallen, PA.

Dorigo M, Di Caro G and Gambardella LM (1999) Ant algorithms for discrete optimization. *Artificial Life* **5**(3): 137–172.

Finin T, Labrou Y and Mayfield J (1996) KQML as an agent communication language. In: Bradshaw JM (ed.) *Software Agents*. Cambridge, MA: AAAI/MIT Press.

Franks NR and Deneubourg J-L (1997) Self-organising nest construction in ants: individual worker behaviour and the nest's dynamics. *Animal Behaviour* **54**: 779–796.

Fredslund J and Matarić M (2002) *A General, Local Algorithm for Robot Formations*. IEEE Transactions on Robotics and Automation.

Gelernter D (1991) *Mirror Worlds*. New York, NY: Oxford University Press.

Gerkey BP and Matarić MJ (2001) Principled communication for dynamic multi-robot task allocation. In: Rus D and Singh S (eds) *Experimental Robotics VII*, pp. 253–362. Springer Verlag: Berlin.

Gerkey BP and Matarić MJ (2002) Sold! Auction methods for multi-robot coordination. *IEEE Transactions on Robotics and Automation*, special issue on multirobot systems.

Holland JH (1975) *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Holland OE and Melhuish C (1999) Stigmergy, self-organisation, and sorting in collective robotics. *Artificial Life* **5**(2): 173–202.

Huberman A and Lukose RM (1997) Social dilemmas and Internet congestion. *Science* **277**: 535–537.

Kandel ER, Schwartz JH and Jessell TM (1995) *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton and Lange.

Kreifelt T and von Martial FA (1991) A negotiation framework for autonomous agents. In: Demazeau Y and Muller JP (eds) *Decentralized A. I. 2*. Oxford, UK: Elsevier Science.

Lesser V and Corkill D (1981) Functionally accurate, cooperative distributed systems. *IEEE Transactions on Systems, Man and Cybernetics* **11**(1): 81–96.

Maes P (1990) Situated agents can have goals. *Robotics and Autonomous Systems* **6**: 49–70.

Matarić MJ (1994) *Interaction and Intelligent Behavior*. MIT EECS PhD Thesis, MIT AI Lab Tech Report AITR-1495.

Matarić MJ (1995) Designing and understanding adaptive group behavior. *Adaptive Behavior* **4**(1): 51–80.

McFarland D (1985) *Animal Behavior*. Menlo Park, CA: Benjamin Cummings.

Michaud F and Vu MT (1999) Managing robot autonomy and interactivity using motives and visual communication. In Proceedings Conference Autonomous Agents. Seattle, Washington.

Minar N, Kwindla HK and Pattie M (1999) *Cooperating Mobile Agents for Mapping Networks*, Proceedings of the First Hungarian National Conference on Agent Based Computation.

Minsky M (1986) *The Society of Mind*. New York, NY: Simon & Schuster.

Murciano A and del R Millán J (1996) Learning signaling behaviors and specialization in cooperative agents. *Adaptive Behavior* **5**(1).

Nicolescu M and Matarić M (2000) *Learning Cooperation From Human-Robot Interaction*. Proceedings, 5th International Symposium on Distributed Autonomous Robotic Systems (DARS), 4–6 Oct, Knoxville, TN.

Ogata T, Matsuyama Y, Komiya T *et al*. (2000) Development of emotional communication robot: WAMOEBA-2R -Experimental evaluation of the emotional communication between robots and humans. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS' 2000), pp. 175–180, November. Takamatsu, Japan.

Reynolds C (1987) Flocks, herds, and schools: a distributed behavioral model. *Computer Graphics* **21**(4): 25–34.

Rosenschein JS and Zlotkin G (1994) *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. Cambridge, MA: MIT Press.

Schwartz DG (1995) *Cooperating Heterogeneous Systems*. Dordrecht: Kluwer Academic.

Segel LA and Cohen IR (2001) *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Oxford, UK: Oxford University Press.

Sutton and Barto (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tambe M and Jung H (1999) The benefits of arguing in a team. *AI Magazine* **20**(4).

Tu X and Terzopoulos D (1994) *Artificial Fishes: Physics, Locomotion, Perception, Behavior*. Computer Graphics, SIGGRAPH 94 Conference Proceedings, pp. 43–50, July.

Vinoski S (1997) CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications Magazine*, February.

Werger B (1999) Cooperation without deliberation: a minimal behavior-based approach to multi-robot teams. *Artificial Intelligence* **110**: 293–320.

Werger B and Matarić MJ (2001) From insect to internet: Situated control for networked robot teams. *Annals of Mathematics and Artificial Intelligence* **31**(4): 173–198.

Further Reading

Bonabeau E, Dorigo M and Theraulaz G (1999) *Swarm Intelligence: From Natural to Artificial Systems*. Oxford, UK: Oxford University Press.

Brooks R (1999) *Cambrian Intelligence*. Cambridge, MA: MIT Press.

Clearwater HS, Huberman BA and Hogg T (xxxx) Cooperative solution of constraint satisfaction problems. *Science* **254**: 1181–1183.

Computability and Computational Complexity

Introductory article

Patrick Doyle, Stanford University, Stanford, California, USA

CONTENTS

Processes that can be specified precisely can be formalized as algorithms. The theory of algorithm analysis determines important properties of these algorithms, such as the resources they consume, and the theory of computational complexity categorizes the problems these processes solve according to those properties, revealing the fundamental limitations of computation.

MEASURING THE DIFFICULTY OF INFORMATION PROCESSING PROBLEMS

The major tasks of a computational system are the retrieval, processing, and presentation of information. The processing of information involves some set of operations designed to transform it from one form into another. A spreadsheet program may add up a column of numbers to determine a month's expenses. A mathematics package might compute the area under a curve. A buyer of a new car might analyze several alternative models to decide which has the best combination of features, safety, speed, and cost.

Some of these tasks are simpler than others. Most people would find it easier to add up a column of numbers than to compute the area under a curve. There are also usually many different ways to solve a particular problem, but some are preferable to others because they consume fewer valuable resources.

In order to solve a problem, there must be a procedure that will generate the answer to the problem, and that procedure must be executed. The branch of computer science known as the theory of algorithms deals with the analysis of such procedures, and attempts to find meaningful measures for comparing one procedure with another. The study of computational complexity extends this theory by organizing problems into categories according to the difficulty of the procedures needed to solve them.

WHAT IS AN ALGORITHM?

In order to solve a problem, some organized sequence of operations must be performed – some sort of procedure, process, routine, or recipe. These terms all capture an intuitive notion of algorithm, but in order to apply formal analysis, it is necessary to be more precise about just what such an activity entails.

An algorithm is a finite sequence of effective and exact instructions for transforming any set of appropriately expressed pieces of information (the input) into another set of pieces of information (the output) in a finite amount of time.

Consider the problem of sorting a hand of playing cards. The goal is to sort the cards from lowest rank to highest. One algorithm to solve this problem works as follows: given a hand of cards, go through the hand and find the card with the lowest rank (if there is more than one, take the first one). Place that card on top of a separate pile. Repeat this process until no cards are left in the hand. Now the pile contains the sorted cards.

This procedure, while inefficient, meets the criteria of an algorithm. The inputs and outputs are well defined as sequences of playing cards. The instructions are effective for a human being to

execute, since they involve moving cards in a hand. They are also exact, since they precisely and unambiguously explain what to do next in each case. The algorithm is guaranteed to terminate since there are only finitely many cards in a hand, and at each repetition of the loop there is one fewer card to examine.

The particular language in which an algorithm is written does not matter much. Generally, in computer science one finds algorithms described in formal programming languages that are designed for this purpose. However, so long as it meets the above requirements, it does not matter whether the algorithm is written as a flowchart, a Java program, English prose, or a Japanese haiku. It is only important that the environment in which the algorithm is to be executed can interpret and follow its instructions exactly and unambiguously. The above description of the card sorting algorithm, for example, is sufficient for an English-speaking human being, but not for a computer.

There are several natural questions about algorithms that have important and surprisingly complex answers. First: does every problem have an algorithm that will solve it? Second: are some algorithms 'better' than others for solving a certain problem? Last, and most practically: is there a way to measure the 'goodness' of an algorithm? The modern theory of computation is a result of the attempt to answer these questions.

COMPUTABILITY AND DECIDABLE PROBLEMS

In the early twentieth century, before the first electronic computers were built, several different formal models of computation were proposed. One of the simplest, and the one most used in computer science today, is the Turing machine, introduced by the mathematician Alan Turing in 1936. (*See* **Turing, Alan**)

A Turing machine is an imaginary mechanical device with three components: an infinitely long tape of paper, divided into cells that may each hold one of a finite set of symbols; a read–write head that can read and modify the contents of a cell; and a controller that may be in any of a finite set of states. The machine's behavior is directed by a fixed set of rules (the machine's program) that depend only on the current state and the symbol under the read–write head. The machine operates by repeatedly cycling through four steps:

1. Examine the symbol in the current cell.
2. Change the internal state (possibly to the same state it already occupies).
3. Write a symbol in the cell (possibly the same symbol already there).
4. According to the rules, move the head one cell to the left or to the right, leave it where it is, or halt the computation.

Although a Turing machine is a simple device, it has all the power and generality of a modern computer. That is, any problem that a modern computer can solve, a Turing machine could also solve. The Church–Turing thesis asserts that the Turing machine (or any equivalent model) exactly captures the notion of 'effective computability'.

A computable function is one for which there is an algorithm that will, for any input it is designed to understand, produce the correct output in a finite amount of time. Thus, according to the Church–Turing thesis, a Turing machine can solve any problem for which an algorithm can be provided. This thesis is only a hypothesis, however. It cannot be proven, since the notion of algorithm is not mathematically rigorous. The Church–Turing thesis says that our intuition about computability is captured by these formalisms.

Because their operations are so simple, Turing machines are frequently used in proving properties of computation. One important area of research into these properties deals with a special class of problems called decision problems. These are problems that have only 'yes' or 'no' answers. Any problem can be cast as a decision problem, and it is often simpler to prove properties of the decision version of a problem than to prove properties of the original problem directly. A decision problem may be 'decidable', 'semi-decidable', or 'undecidable'.

A problem is decidable if there exists some algorithm that is guaranteed either to accept (with a 'yes') or reject (with a 'no') any instance of the problem. Many obvious problems have decidable algorithms. Even the game of chess is a decidable problem: since there are only finitely many games (assuming no boundless repetition of useless moves), it is possible to find a strategy that will play perfectly just by exhaustively examining every possible game. The amount of computation is entirely impractical, but in principle, questions such as 'is white guaranteed to win in this position?' can be answered.

Semi-decidable problems have algorithms that are guaranteed to accept all 'yes' instances. If the instance is a 'no', the algorithm may reject the instance, or it may run forever without being able to determine that the instance should be rejected. An

example of a semi-decidable problem is the problem of determining whether a given number is the difference of two primes. If a problem has no algorithm that can be guaranteed to accept all 'yes' instances, it is said to be fully undecidable. Many important questions, even some about algorithms themselves, are undecidable.

GÖDEL'S INCOMPLETENESS RESULTS AND THE ALGORITHMIC NATURE OF COGNITION

At the beginning of the twentieth century, the mathematician David Hilbert posed the problem of finding an algorithm that could determine whether any mathematical proposition, expressed in the language of logic, was true or false. This question remained open until 1931, when the mathematician Kurt Gödel published his famous 'Incompleteness Theorem'. This theorem states that, in any sufficiently powerful logical system such as mathematics, there are propositions that cannot be proved true or false within the system itself. Gödel showed that mathematics is fundamentally incomplete because there are true propositions in mathematics that cannot be proved using the axioms of mathematics. Thus Hilbert's algorithm could not exist.

When Alan Turing introduced the Turing machine and its definition of computability in 1936, he used a version of Gödel's argument to show that there are certain problems that are not computable, in the sense that there exist no algorithms that can compute their solutions. Specifically, Turing proved that no Turing machine could solve the so-called halting problem, a decision problem that asks whether a given Turing machine will eventually halt on a given input.

The proof can be described informally. Suppose there does exist a Turing machine $D(M, x)$ that decides whether Turing machine M would halt when run on input x. Now construct a Turing machine $N(M)$ that takes a machine M as input. N operates as follows. First, $N(M)$ runs $D(M, M)$; that is, N uses D to decide whether machine M would halt when fed its own design as its input. If D replies that M would halt on input M, N enters an infinite loop. If D replies that M would never halt on input M, N halts.

What happens if N is run on itself? If D says that N would halt on input N, N actually enters an infinite loop. If D says that N is would not halt, N actually halts. This is a contradiction of the assumption that D can correctly decide the halting problem, so that assumption must be false. No such machine D can exist.

These remarkable results have led to philosophical debates about algorithms and their relationship to human information processing. Do human beings think in ways that can be expressed in algorithms, or is there some other kind of computation taking place? If humans do use algorithms, are they therefore bound by the same limitations as computers, and logically incapable of performing certain kinds of tasks? (*See* **Artificial Intelligence, Gödelian Arguments against**)

Several prominent philosophers have argued against this. Their arguments range from John Searle's 'Chinese room' argument that computation cannot produce understanding, to Roger Penrose's hypothesis that the brain has certain quantum-mechanical properties that allow it to function in ways no computer could mimic. These arguments have not convinced many computer scientists, and the branch of computer science known as artificial intelligence is devoted to building computers with abilities that equal or exceed those of human beings. (*See* **Artificial Intelligence, Philosophy of**; **Chinese Room Argument, The**; **Computation, Philosophical Issues about**)

ORDER OF MAGNITUDE AND COMPLEXITY MEASURES IN SPACE AND TIME

Given a problem together with an algorithm for solving it, the next task is to analyze the resources that the algorithm consumes. With each instance of the problem, we associate a 'size', which is the number of symbols required to describe the instance. The resources the algorithm consumes when operating on an instance of the problem can then be expressed in terms of the size of the instance. Generally the resource of interest is either the time it takes to run or the space the algorithm needs to perform its computations.

Recall the simple algorithm for sorting a hand of playing cards. It looks for the highest-ranked card and moves it to a separate pile. It repeats this process with each succeeding card until all the cards are sorted. If there are n cards in the hand, the first card must be compared against all $(n-1)$ remaining cards, the second card against $(n-2)$ cards, and so on down to the last two cards, which require only one comparison. This approach requires $(n-1)+(n-2)+\ldots+1=(n^2-n)/2$ comparisons altogether.

To determine how much time this algorithm would actually take would require detailed information about the computer on which it is run, including how much time it takes to read the input,

make a comparison, and so on. Ordinarily algorithm analysis is not interested in this level of detail; it is enough to determine the rate of growth, or order of growth, of the algorithm, a more abstract measure that will be identical on all machines that share the same fundamental principles of operation.

Several simplifying assumptions are made to find the algorithm's order of growth. First, constant-time overhead costs are ignored. Only operations that grow in number as the size of the input grows are considered. Second, only the dominant term ($\frac{1}{2}n^2$ in the above example) in the number of operations is used. As the size of the input grows, this term overwhelms the others: when $n = 10$, $\frac{1}{2}n = 5$ and $\frac{1}{2}n^2 = 50$; when $n = 1000$, $\frac{1}{2}n = 500$ but $\frac{1}{2}n^2 = 500\,000$. Finally, any coefficients on the dominant term are ignored, since such constants are less important than the overall rate of growth. This leaves n^2 as the term of interest. This order-of-growth bound is written as $O(n^2)$. This O notation is commonly used to describe the worst-case complexity of an algorithm, which is the amount of effort it will require on the most difficult possible input.

This formulation makes it straightforward to compare algorithms without having to consider the details of a particular computer or minor factors in the algorithm that do not have a significant effect on its overall efficiency. For the card-sorting problem, for example, there are many known algorithms, and the most efficient comparison algorithms have an order of growth of $O(n \ln n)$, a considerable improvement over $O(n^2)$ for large hands.

PROBLEM REDUCTION AND COMPLEXITY EQUIVALENCE CLASSES

The kind of analysis used in the previous section is helpful when examining a single algorithm, or comparing several algorithms, but the theory of computational complexity is concerned with understanding fundamental distinctions between classes of algorithms. There are two especially important classes, known as P and NP, which intuitively divide problems into 'easy' and 'hard'.

P is the class of all decision problems that take an amount of time that is polynomial or better in the size of the input, such as $O(n \ln n)$, $O(n^2)$ or $O(n^{357})$. These problems are regarded as 'easy' or 'tractable' because the amount of time grows relatively slowly with the size of the input.

The other important basic complexity class is called NP, for nondeterministic polynomial time. NP consists of decision problems whose answers can be verified in polynomial time. That is, given some instance of a problem and a guess for an answer, there is an algorithm that can check in polynomial time whether that guess is correct.

Clearly P is a subset of NP, since any problem that can be solved in polynomial time can have its solution checked in polynomial time. However, there are many NP decision problems for which no polynomial-time algorithm has been found: it is known how to check their solutions in polynomial time, but the best-known algorithms for solving them take exponential time, such as $O(2^n)$, or worse. These problems are in NP but seemingly not in P.

Many other complexity classes have been studied. Some deal with different time bounds: for example, EXP consists of problems solvable by exponential-time algorithms, and is a strict superset of NP. Others measure different resources, such as space. PSPACE is a class of problems whose algorithms are polynomially bounded (in the size of the problem instance) in the amount of space they use rather than in the number of operations they perform.

Proving that a decision problem is in a certain complexity class can be difficult. An important concept in complexity analysis is that of reducing one problem to another. In order to determine the complexity of some problem A, one can often show that it can be transformed into another problem B whose complexity is already known. Then the complexity of A is no worse than the complexity of B plus the complexity of turning instances of A into instances of B. Computer scientists have built up a large library of problems with known complexities, making this approach a common way to determine the complexity of a new problem.

NP-COMPLETE PROBLEMS AND THE 'P = NP' QUESTION

In 1971, Stephen Cook proved that there is a certain problem in NP, called the satisfiability problem, such that any other problem in NP can be reduced, in polynomial time, to an instance of it. Hence, if there is a fast algorithm for satisfiability, then there would be a fast algorithm for every problem in NP; conversely, if any problem in NP is intractable, then satisfiability is intractable.

The satisfiability problem can be stated informally as follows. Given a set of variables that can be either 'true' or 'false', and a formula combining

those variables with the logical connectors 'and', 'or' and 'not', is there an assignment to each variable that makes the entire formula true? For example, '((v_1 and v_2) or (v_2 and not v_3))' is true when either both v_1 and v_2 are true or when v_2 is true and v_3 is false. Any instance of any problem in NP can be rewritten, in polynomial time, as an instance of satisfiability.

Since Cook's proof, many other problems in NP have also been shown to have this property. Such problems are called NP-complete. They are, in a sense, the 'hardest' problems in NP, since finding a fast algorithm for any one of them would mean finding a fast algorithm for all problems in NP.

As mentioned above, the class P is contained within NP. Surprisingly, it is still not known whether P is actually equal to NP: that is, whether every problem in NP actually has a polynomial-time algorithm, and we have just not yet found one for any NP-complete problem.

If P = NP, then the huge range of important problems in NP for which no tractable algorithms have been found must all have tractable algorithms. After decades of research, however, it is widely believed that P $\neq$ NP. Many modern cryptographic systems base their security on the assumption that certain problems in NP are too difficult to solve in any reasonable span of time. Whether these classes are equivalent or not, a proof either way will have widespread implications for complexity analysis and algorithm design. This is one of the great unsolved questions in computer science.

COMBINATORIC EXPLOSIONS IN COMMON COGNITIVE TASKS AND THE USE OF APPROXIMATION ALGORITHMS

There are many important problems that are in NP. Many common cognitive tasks fall into this category, and yet human beings are able in their daily lives to perform tasks that are theoretically too difficult to solve in any reasonable time. One explanation is that humans perform a fundamentally different kind of computation than computers do. But there are some other possibile explanations.

A problem that is in NP may still be tractable in practice. There exist problems in NP for which algorithms with very low exponents are known; for example, $O(2^{0.00001n})$ might be an acceptable order of growth in most cases.

Another possibility is using a fast algorithm that provides an approximate solution to the problem. Often it is possible to find an algorithm in P that is guaranteed to provide solutions that are within some bounded range of the best solution. One well-known example is the travelling salesman problem, which provides a map of cities and the distances between them, and asks for the most efficient route for visiting all the cities. This problem is in NP, but there is a known polynomial-time algorithm that is guaranteed to find a route no worse than twice the length of the most efficient route. In many cases these approximation algorithms are good enough for practical needs, and they are useful when the exact algorithms are impractical.

STRONG COMPLEXITY CONSTRAINTS ON COGNITIVE MODELS

Complexity measures ordinarily assume that computations are being performed on a serial machine – that is, steps in the algorithm are executed one after another and one at a time. However, there are strong arguments that this is not a reasonable model of human cognition. One prominent argument, given by Jerome Feldman and Dana Ballard, is contained in what is called the '100-step rule'.

Simple cognitive tasks, such as identifying and naming an object, take human beings something on the order of half a second (500 ms) to perform. Since neurons, which are the basic computational components of the brain, take around 5 ms to act, the 100-step rule declares that these cognitive tasks cannot take more than about 100 sequential neuronal operations. (*See* **Computational Models of Cognition: Constraining**)

However, although a single neuron may take several ms to act, our brains contain many billions of neurons. Inspired by the structure of the brain, connectionist models of computation consist of many independent processors that are connected together, and that perform their computations in parallel, rather than serially. As long as the serial algorithm can be redesigned so that each one of the individual parallel processors has something to do in each time step, its speed can be improved by as many times as there are processors – in the case of the human brain, a parallel algorithm might run billions of times faster than a serial one. (See **Connectionist Architectures: Optimization**; **Connectionism**)

Although parallel algorithms can be more complex to design, the same tools that are used for ordinary serial algorithms can be applied. The complexity analyses on time and space are the same for

both parallel and serial systems, since a parallel machine only improves over a serial one by a constant factor, and such constant terms are ignored in order of magnitude analysis. However, in practice a linear speed-up of billions of times significantly increases the range of problems that can be solved.

SUMMARY

The design of algorithms is fundamental to computer science. The theory of algorithm analysis makes it possible to determine the amount of a resource that an algorithm requires. This allows the comparison of different algorithms to determine which are best in what situations. The theory of computational complexity has led to categorizing problems according to their difficulty, with the polynomial-time P problems being intuitively 'easy' and the nondeterministic polynomial-time NP-complete problems 'hard'. Many important problems have been shown to be NP-complete, but often it is possible to develop approximation algorithms that will give good, if not necessarily the best, answers to these problems. These concepts can be applied to algorithms that attempt to explain human brain processes. However, since the brain contains many billions of slow neurons operating in parallel, rather than a few fast processors, some argue that they may not be appropriate measures for human cognition.

Further Reading

Cook S (1971) The complexity of theorem-proving procedures. In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pp. 151–158. New York, NY: ACM Press.

Cormen T, Leiserson C and Rivest R (1996) *Introduction to Algorithms*. Cambridge, MA: MIT Press.

Hopcroft J and Ullman J (1979) *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Feldman J and Ballard D (1982) Connectionist models and their properties. *Cognitive Science* **6**: 205–54.

Garey M and Johnson D (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY: W. H. Freeman.

Johnson-Laird P (1988) *The Computer and the Mind: An Introduction to Cognitive Science*. Cambridge, MA: Harvard University Press.

Lewis H and Papadimitriou C (1981) *Elements of the Theory of Computation*. Englewood Cliffs, NJ: Prentice-Hall.

Papadimitriou C (1994) *Computational Complexity*. Reading, MA: Addison-Wesley.

Pinker S (1997) *How the Mind Works*. New York, NY: W. W. Norton.

Computation, Formal Models of

Introductory article

Arun Jagota, University of California, Santa Cruz, California, USA

CONTENTS

This is the study of abstract machines that compute, and of what they compute. This is also a study that characterizes what problems can be computed in principle, and which of the computable problems can be computed in practice.

THE MIND AS MACHINE?

For centuries, philosophers have wondered whether the mind is a 'mere' machine. We are no closer to answering this question now, at the beginning of the twenty-first century, than we were then. On the other hand, we have made tremendous progress in building and understanding *machines*. Now we are even able to put together ones that work in ways that might be considered somewhat intelligent. (*See* **Computational Models: Why Build Them?**)

THE BENEFITS OF A FORMAL MATHEMATICS OF COMPUTATION

The advances in building and understanding mind-like machines would not have occurred without the theoretical and algorithmic foundations provided by formal models of computation. Formal models of computation establish the limits of what can and cannot be computed. For a problem that is computable, complexity theory tells us whether it can be computed *efficiently* or not. Automata theory tells us what types of computations different types of machines can perform. Formal language theory presents to us a hierarchy of languages of increasing sophistication. A remarkable correspondence is found between the automata of automata theory and the languages of formal language theory. One of the most powerful abstract models of computation is the Turing machine. The study of its properties has led to major results, including the theory of NP-completeness.

Upto this point we have discussed only sequential models of computation. Parallel computation on the other hand can be much faster. Brain-inspired models of massively parallel computation have been developed and are now widely studied and used. A fundamentally different type of model of parallel computation – the quantum computer – is being studied in the abstract. Should it be realizable, it would radically transform the nature of computation as we know it today.

AUTOMATA

The word *automaton* really only means 'machine'. However, it is used to convey a sense that our interest lies in the computational properties of the machine, especially those that may be characterized formally. This section considers automata that recognize languages. Such an automaton reads a string of input symbols and returns 'yes' if the string is in the language associated with the automaton and 'no' if not.

There are three broad classes of automata: *finite-state automata (FSA)*, *pushdown automata (PDA)*, and *Turing machines (TMs)*. Finite-state automata are the simplest and, not surprisingly, the least powerful. Pushdown automata have more features; this makes them more powerful than finite-state automata. Turing machines have even more features; this makes them the most powerful. The power of a class of automata is measured by the richness of the class of languages associated with it (see the last paragraph of this section for more on this).

To appreciate the issues we need to understand what we mean by *features* in a machine, and what we mean by *power* of a machine. The features of a machine may be roughly classified as type of *memory*, and the *instruction set*. For instance, the only memory in a finite-state machine resides in its *states*. The only instruction that a finite-state automaton has is one that reads the next input symbol in the current state, and makes a transition to the next state. Both issues are illustrated in Figure 1.

In this diagram, let us imagine that we are presently in state A. This is the memory we are talking about. That is, the FSA *knows* that we are in state A. However, the FSA does *not* know how we got there; i.e. it does not remember states, if any, that were visited prior to reaching state A. Next, the machine reads the next letter in the input. If the letter is a 0, it moves to state B. If the letter is a 1, it moves to state C. These are the only types of actions this machine can take.

A pushdown automaton on the other hand has a more sophisticated memory mechanism – its memory resides not only in the present state it is in, but also in a *stack* that it is allowed to use. To make effective use of the stack, the PDA also has a richer instruction set, specifically, instructions to push items onto, and pop items off, the top of the stack as illustrated in Figure 2.

In this diagram, let us imagine that we are presently in state A. If the next letter in the input is a 0 and the top of the stack contains a Y, then we replace the Y on the top of the stack by Z and move to state B. If, on the other hand, the next letter in the input is a 1 and the top of the stack contains a W, then we replace the W on the top of the stack by X and move to state C. From this example, we see that the instructions of a PDA are more sophisticated than those of an FSA. It may appear that, like an FSA, the PDA also does not remember which states it visited before reaching A. This is not true. We can record these states in the stack. In fact, creative use of the stack is what endows a PDA with its added power over an FSA.

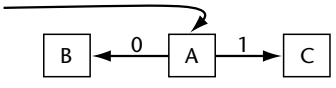


Figure 1. Finite-state automaton.

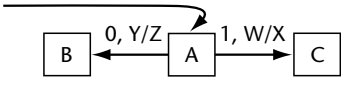


Figure 2. Pushdown automaton.

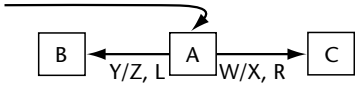


Figure 3. Turing machine.

One wonders whether a stack is enough to compute anything that is computable. Interestingly, it turns out that it is not. The Turing machine, which *is* enough, has an even more flexible memory, and instructions to go with it. A TM has a *tape*, which can be read in *both* directions. (By contrast, one can only push items onto, and pop items off, the top of a stack.) Figure 3 illustrates the use of this tape, especially the instructions that read, write, and use it. (*See* **Turing, Alan**)

In this diagram, let us imagine that we are presently in state A. Also, the tape head is on top of a certain cell on the tape (the tape is divided into cells). If the current cell – that the tape head is scanning – contains a Y, we replace this cell's content (Y) by Z, and move the tape head one cell to the left (this is what L means). If the current cell contains a W, we replace this cell's content by X, and move the tape head one cell to the right. From this example we see that, unlike an FSA or a PDA, a TM does not explicitly read the external input – it is assumed to be already on the tape before the TM execution is begun. We see that the TM has a richer feature set than a PDA, both in its memory and in its instructions.

Now one would hope that the automata with the richer feature sets are strictly more powerful than the ones with the poorer feature sets. Indeed this is the case. We explain this in the following setting. We measure computing power of an automaton in its use as a *language recognizer*. In this use, TMs recognize all languages that are decidable (the recursive languages), PDAs recognize only a *proper* subset of these languages (the context-free languages), and FSAs recognize only a *proper* subset of the languages that PDAs recognize (the regular languages). Various classes of languages mentioned above – recursive, context-free and regular – are discussed in the next section.

FORMAL LANGUAGES AND THE CHOMSKY HIERARCHY

As everyone knows, language is essential to communication. What one might not realize is that language is also essential to computation.

The Chomsky hierarchy of formal languages was a landmark event in linguistics and, as it turns out, also in computation. This hierarchy defines classes

of languages of strictly increasing power. In what follows we will let Σ denote the alphabet, Σ^* the set of all strings on the alphabet, and $L \subseteq \Sigma^*$ a language on Σ. (Notice that a language is a *particular subset* of all possible strings on an alphabet.)

At the lowest level in the hierarchy is the class of *regular languages*. This class, R, may be defined as follows. The language $L = \emptyset$ that contains no strings is in R. The languages $L = \{a\}, a \in \Sigma$ are in R. (Each of these languages contains a one-letter string, comprised of an alphabet symbol.) The language $L = \{\wedge\}$, which contains just one element – the empty string – is in R. If languages L_1, L_2 are in R then so are the languages $L_1 \cup L_2$, $L_1 \circ L_2$, and $L_1{}^*$. Here $L_1 \circ L_2 = \{x \circ y | x \in L_1, y \in L_2\}$ where the small circle symbol ($\circ$) denotes the concatenation of two strings, and $L_1{}^* = \bigcup_{i=0,\ldots,\infty} L_1^i$ where $L_1^i = \underbrace{L_1 \circ L_1 \circ \ldots \circ L_1}_{i \text{ times}}$. In other words, a regular language is built from simpler languages by the union of two regular languages, the concatenation of two regular languages, or the repeated concatenation – zero or more times – of the same language.

This simple structure suggests that regular languages should be easy to parse. Indeed this is the case. Regular languages are parsed by *regular expressions*. Regular expressions mirror the structure of the definition of the class of the regular languages; specifically they are composed of operators that correspond to the union and the concatenation of two languages and an operator for repeated concatenation of the same language.

Despite being at the lowest rung of the Chomsky hierarchy, a regular language can have a rich structure. On the other hand, precisely because the language is regular, this structure can be efficiently parsed by computer methods. These two facts collectively explain why regular languages are so popular, having many applications. These include free-text searches, especially in Unix and on the web, and computational biology. (*See* **Finite State Processing**)

At the next level in the hierarchy is the class of *context-free languages*. Every regular language is a context-free language. On the other hand, there are context-free languages that are not regular. The language formed by the set of all palindromes (say on the alphabet $\Sigma = \{0, 1\}$) is one such example. (A palindrome is a string that coincides with its reverse.)

Before we give a formal definition for this class of languages, it will help us to examine the closely related notion of a grammar. A *grammar* for a language is a set of rules which collectively allow us to determine which strings are in the language and which are not. A grammar is itself written in a *meta-language*. By imposing different restrictions on what forms the rules can take in this meta-language, we get different classes of grammars, hence different classes of languages.

In a *context-free grammar* we can have rules only of the type $A \Rightarrow x$. Here A is a single non-terminal symbol and x a string, possibly empty, composed of non-terminal and/or terminal symbols. The terminal symbols in the grammar are the alphabet symbols in the associated language.

Let us use an example to explain what a non-terminal symbol is, and how the set of rules comprising a grammar is applied to test whether a given string is in the language associated with the grammar or not. Here is one context-free grammar for the language of palindromes mentioned earlier.

$$S \Rightarrow 0S0 \quad S \Rightarrow 1S1 \quad S \Rightarrow 0 \quad S \Rightarrow 1 \quad S \Rightarrow \wedge$$

In this grammar, S is the only non-terminal symbol, 0 and 1 are the terminal symbols, and $\wedge$ denotes the empty string. This grammar has five rules. Let us now apply this grammar to verify that 0110 is a palindrome. We do this by starting with the non-terminal symbol S and applying a certain sequence of rules which transforms this S to the string 0110. This process is called a *derivation*. The derivation in this example is:

$$S \rightarrow 0S0 \rightarrow 01S10 \rightarrow 0110$$

Which rules were applied when is not made explicit in this derivation. In our example though, we can easily infer this from our derivation.

A context-free language may now be defined as one that is recognized by a context-free grammar.

At the next level in the Chomsky hierarchy is the class of *context-sensitive languages*. Every context-free language is a context-sensitive language. On the other hand, there are context-sensitive languages that are not context-free. The language $L = \{xx | x \in \Sigma^*\}$ is one such example.

A context-sensitive language is associated with a grammar in which we can have rules only of the type $xAy \Rightarrow xzy$. Here A is a non-terminal symbol and x, y, and z are strings composed of non-terminal and/or terminal symbols. (x and/or y can be empty, but z cannot be.) The meaning of this rule is that 'when A occurs with x on its left and y on its right, it may be replaced by the string z'. Notice that this rule is more general than one for a context-free grammar; the latter does not permit context-dependent replacement.

Finally, a *recursive language* is one whose characteristic function is decidable. That is, given a string x in Σ^*, there is an effective procedure for determining whether x is in the language or not.

THE CORRESPONDENCE BETWEEN AUTOMATA AND FORMAL LANGUAGES

What solidifies the connection between languages and computation is the precise correspondence between formal languages of the Chomsky hierarchy and the automata that we saw in a previous section. Specifically, it turns out that FSAs recognize regular languages, PDAs recognize context-free languages, and TMs recognize recursive languages. In fact, even context-sensitive languages have their automaton counterpart – the linear-bounded automaton, which is realized by imposing certain restrictions on a TM.

THE TURING MACHINE

To this point, we have described the Turing machine as a language recognizer. Unlike FSAs and PDAs, however, a Turing machine may also be used as a more conventional type of computer, to read input, do some task, and produce output. This is done in much the same way as a language would be recognized, the main difference being that after reading the input and doing its computation, the TM places the output on the tape at a specific location.

THE UNIVERSAL TURING MACHINE

A TM runs a particular program. This is cumbersome; to run different programs, one needs to construct different TMs, one for each program. The universal Turing machine (UTM) is a fix to this. A UTM accepts, as its input, both the program P and the input x to P. The UTM then runs the program P on input x, and writes out, to the tape, the output that P would have produced, on input x.

THE VON NEUMANN ARCHITECTURE

The von Neumann architecture may be thought of as a practically-inclined realization of the idea of a universal Turing machine. This architecture uses the key notion of a 'stored program concept'. A program, like data, is stored in memory. A program is indistinguishable from data, as it resides in memory. (*See* **von Neumann, John**)

In the von Neumann architecture, there is a 'central processing unit' that reads and executes the instructions of a program stored in memory, one by one. The data needed by an instruction is read from memory; the result produced by the instruction is written back to memory. For example, the instruction $z = x + y$ would be executed as follows. First, the instruction is read from memory. Next, the data items needed by the instruction, x and y, are read from memory. Then, the values of x and y are added together. Finally, the result is stored back into memory at location z.

INSTRUCTION SETS, COMPUTER LANGUAGES, AND THE IDEA OF THE 'VIRTUAL MACHINE'

The von Neumann architecture is the basis for every sequential computer built. The huge significance of the stored program concept – that it facilitates the execution of countless programs on the same computer – was recognized early on and profitably exploited since. Computers were developed with carefully designed instruction sets. The instruction sets had to be powerful enough so that any programs could be written in them. On the other hand, they had to be simple enough to keep the central processing unit from becoming overly complex.

These conflicting objectives were resolved by adopting 'minimalist' designs. Specifically, the instruction set was kept small, and composed of simple and 'orthogonal' instructions. The instructions were designed to be combinable. Thus, a user wanting a more elaborate instruction set could in effect create one by combining various instructions from the smaller set in various ways.

While this strategy solved the problem it was designed to solve, a different one remained. The instructions in a computer's instruction set were at a very low level; just right for the computer to understand them, but very tedious for humans to program in. A breakthrough came in the development of computer languages that humans could write in more easily. These languages had very high-level instructions. On the other hand, computers could not understand them. This dilemma was resolved by building translators, programs that take a program written in a high-level language and translate it to a program in the computer's (low- level) language.

Once we had freed ourselves from having to write programs at the level of the computer, we had also freed ourselves from its constraints to a considerable extent. Specifically, we no longer

needed to write programs that would fit in its memory, or use only its instructions. The translator in effect created a 'virtual machine' around the actual computer, one that was vastly richer. This notion of a 'virtual machine' revolutionized software development, freeing it to a great extent from the confines of a particular computer.

WHAT IS AN ALGORITHM?

An algorithm is a step-by-step procedure to solve a particular problem. An algorithm is required to terminate in any one run. How does this notion differ from that of a program? Well, for one thing, a program may not always terminate – loops in its control flow may cause it to run forever at times. Second, a program typically means a sequence of instructions (to do something) written in a particular language, while an algorithm is a higher-level, language-independent recipe for solving a problem in a particular way.

THE NOTIONS OF COMPUTABILITY AND COMPUTATIONAL COMPLEXITY

Is it possible to develop a program that can check whether or not any given program will terminate? This would be nice; we could use it to find 'infinite loops' in our programs. Unfortunately, it is impossible to do so. (On a practical computer with bounded resources such a program can be developed, though it will be impractical.) That is, the so-called halting problem is noncomputable.

More precisely, let us define a partial function $f{:}\mathrm{N} \rightarrow \mathrm{N}$ as a function that takes a natural number as input and returns a natural number $f(n)$ if f is defined on n, and returns 'undefined' if f is undefined on n. We say that a Turing machine T computes f if T reads the input n as 1^n and produces output $1^{f(n)}$ if $f(n)$ is defined and does not halt if $f(n)$ is undefined. We say that a partial function is computable if there exists a Turing machine that computes it. Since any given TM can compute at most one partial function, the number of TMs is countable, and the number of partial functions is uncountable, we see that there are many partial functions that are not computable.

The notion of computability establishes the limits of what can be computed in principle. It turns out that the limit of what can be computed in practice is far lower. The characterization of this limit is the subject of computational complexity.

Computational complexity is the study of the time needed to solve a particular problem as a function of the size of an instance of the problem. For example, consider the problem of factoring a composite number. An instance of this problem is specified by a particular positive integer. The size of an instance – the number n – is $\lceil \log_2 n \rceil$, the number of bits needed to describe n. The fastest known factoring algorithm runs in time exponential in n. The apparent difficulty of factoring has a very significant application: the development of public key cryptosystems such as those used at banks (and now for e-commerce over the internet) that seem unbreakable.

Now consider a different problem – of testing whether a given number occurs in a list of numbers. This problem can be easily solved in time proportional to n. If we assume that the instance size is also proportional to n – a valid assumption when the numbers are required to fit in a fixed word size – then this problem's computational complexity is linear.

Thus, we would conclude that factoring is computationally hard, while search is computationally easy.

How do we know that factoring is indeed computationally hard? Perhaps there is an efficient algorithm – we just have not found it. In the early 1970s this very sort of questioning led to the remarkable theory of NP-completeness. Informally speaking, the class NP-complete is the set of computable problems, none of which is known to be efficiently solvable, yet all are efficiently interconvertible (i.e. any NP-complete problem can be efficiently transformed into any other). This still does not mean that an NP-complete problem is definitely hard. However, the efficient interconvertibility yields an intriguing and significant property: If *any* NP-complete problem is easy, then *every* NP-complete problem is easy.

The practical import of this theory is as follows. Thousands of problems are known to be NP-complete. If one can show that a new problem of interest is NP-complete, then this suggests that it really is hard. If this problem were easy, then so would the thousands of other problems be. But since many able researchers have tried to devise efficient algorithms for many of these problems, over the past thirty years, and all have failed, this is highly unlikely.

PARALLEL COMPUTATION, ASSOCIATIVE NETWORKS, AND CELLULAR AUTOMATA

The von Neumann architecture, while pivotal to the evolution of computers, has one major drawback – it has a sequential bottleneck. Instructions must be

executed one at a time. On the other hand, everyone realizes that many algorithms are intrinsically parallel. For example, to test if a number x occurs in a list L of numbers, we could compare x against each number in L in parallel.

The observation that many algorithms are inherently parallel, coupled with the incessant need of humans to 'have things done faster', has spurred research on parallel computation. Many models of parallel computation have been proposed. Some have even been implemented. These days every computer – even a sequential one – is parallel. On your computer you can compose an e-mail message in one window at the same time that the computer is downloading a video off the internet. This is an example of parallel computation.

The brain has been one source of inspiration for many models of parallel computation. Many computations in the brain are massively parallel (many others are sequential). Groups of neurons fire in synchrony, or in asynchrony. Brain-inspired models such as artificial neural networks and cellular automata are composed of primitive computing elements interconnected together in various ways. (In the case of artificial neural networks, these are the neurons.) The power of these models arises from the precise way in which these primitive elements are interconnected. If we 'add up' the computations performed by the computing elements, they don't amount to much. If we now interconnect these same elements in particular ways, suddenly we have gained a lot of computing power. In these cases, the whole is indeed greater than the sum of its parts.

In artificial neural networks, the net input to neuron i is $n_i = \sum_j w_{ij} S_j$ where j indexes the neurons connected to neuron i, S_j is the output of neuron j, and w_{ij} is the weight of the connection from neuron j to neuron i. The output of neuron i is $S_i = g(n_i)$ where g is the neuron's transfer function, typically a sigmoid.

In a cellular automaton, the primitive computing element is a cell. By contrast to neural networks, the cells are usually interconnected in a lattice, typically with one or two dimensions (neural networks are typically more flexibly connected). A cell updates its state by examining the states of its neighbors on the lattice. Although a cellular automaton typically has more restricted connectivity than a neural network, the cells in a cellular automaton are more flexible in what they can compute. Specifically, their computations are not restricted to taking the weighted combinations of the outputs of their neighbors and following these by a transfer function.

QUANTUM COMPUTATION

Complexity theory tells us that there are many problems that seem to be inherently intractable. These results hold only on models of classical computation, however. Quantum computers, on the other hand, are capable of performing exponentially many computations in parallel, exponential in the size of the input. They are therefore able to break the intractability barrier of classical computation. However, not all NP-complete problems are known to be tractable on quantum computers. The class of tractable problems on quantum computers is still not well understood in relation to classical models.

Let us revisit the factoring problem discussed earlier. Recall that we mentioned that the best classical algorithm takes exponential time in the size $\lceil \log_2 n \rceil$ of the binary representation of the composite number n that is to be factored. By contrast, there exists an algorithm for factoring on a quantum computer that takes only roughly quadratic time in the size $\lceil \log_2 n \rceil$ of the input. Should quantum computation become a practical reality at some point in time, this would suggest that public key cryptosystems might be breakable. (The first famous paper demonstrating tractability of prime factorization on quantum computers is due to P. Shor.)

At this time, quantum computers are already a little more than an abstract model in that they have been realized in toy devices. It remains to be seen whether practical quantum computers can be built.

Further Reading

Davis MD and Weyuker EJ (1983) *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*. New York, NY: Academic Press.

Garey MR and Johnson DS (1979) *Computers and Intractability. A Guide to the Theory of NP-Completeness*. New York, NY: WH Freeman.

Hopcroft J, Ullman JD and Motwani R (2000) *Introduction to Theory of Neural Computation*. New York, NY: Addison-Wesley.

Hertz J, Krogh A and Palmer R (1991) *Introduction to the Theory of Neural Computation*. New York, NY: Addison-Wesley.

Martin JC (1991) *Introduction to Languages and the Theory of Computation*. New York, NY: McGraw-Hill.

Nielsen MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. New York, NY: Cambridge University Press.

Toffoli T and Margolus N (1987) *Cellular Automata Machines: A New Environment for Modeling*. Cambridge, MA: MIT Press.

Computational Models of Cognition: Constraining

Intermediate article

Terry Regier, University of Chicago, Chicago, Illinois, USA

CONTENTS

Computational models of cognition may be constrained in various ways. The three primary benefits of constraining a computational model are: (1) parsimony; (2) avoiding over-flexibility; and (3) potentially stronger motivation for the elements of the model.

OVERVIEW

Computational models of cognition are an essential tool in the study of the mind. One advantage of such models is their explicitness. The computational researcher must specify a mental mechanism in sufficient detail to allow the resulting model to be instantiated on a computer, and run as a cognitive simulation. This often requires that theoretically important elements of the model be described very clearly: the theory must be explicit enough to be described as a machine.

But what sort of machine? Computational machines range from the inflexible to the very flexible. An inflexible machine performs only a single predetermined task, such as addition, or multiplication, or accepting 65 cents in coins and returning a can of soda. A very flexible or general machine, in contrast, may be used for a wide variety of functions. Significantly, there exists an abstract machine, the Turing machine, that is so general that it has been taken as the formal equivalent of the very broad informal notion of an algorithm – thus, a Turing machine is taken to be capable of implementing any algorithm at all (Lewis and Papadimitriou, 1981). Modern programmable computers are approximations to Turing machines. Thus, the flexibility of modern computers gives a sense of the generality of Turing machines, and of the potential generality of computational machines as a class.

What is the significance of these issues of computational generality and flexibility for models of cognition? Often, generality is taken as a strength in a scientific theory. One would like to be able to account for a broad array of data, from a variety of sources, rather than provide a micro-theory of a very limited domain of cognition. There is, however, a danger of over-generality in model construction. If data are accounted for by a model that is very general and flexible, it is not always clear what to make of the model's success. The problem is that such a general model could also perhaps have fit other data, of a sort never empirically observed. That is, it may have succeeded in accounting for the empirical data because of its extreme flexibility, rather than because of a systematic match between the structures of the model and the mental processes at play. In this sense, a model's computational power may undermine its explanatory power. A more constrained, less computationally powerful model may in some instances provide more insight into the mental processes under study.

THE BENEFITS OF CONSTRAINTS

There are three primary benefits of a constrained model. These are: a more parsimonious characterization, an avoidance of over-flexibility, and potentially stronger motivation for the elements of the model. To illustrate these benefits, let us turn to a simple example.

Consider the task of silently counting from zero up to a given positive integer, and then pressing a button when done. This task involves the mental process of counting, with no overt behavioral clues as to the nature of the process other than the total time required. This time is shown as a function of the target number in Figure 1. These are informally collected data, meant for illustrative purposes only. The data represent averages over five testing sessions, from one subject. The figure also shows the fits of two models, one more constrained than the other.

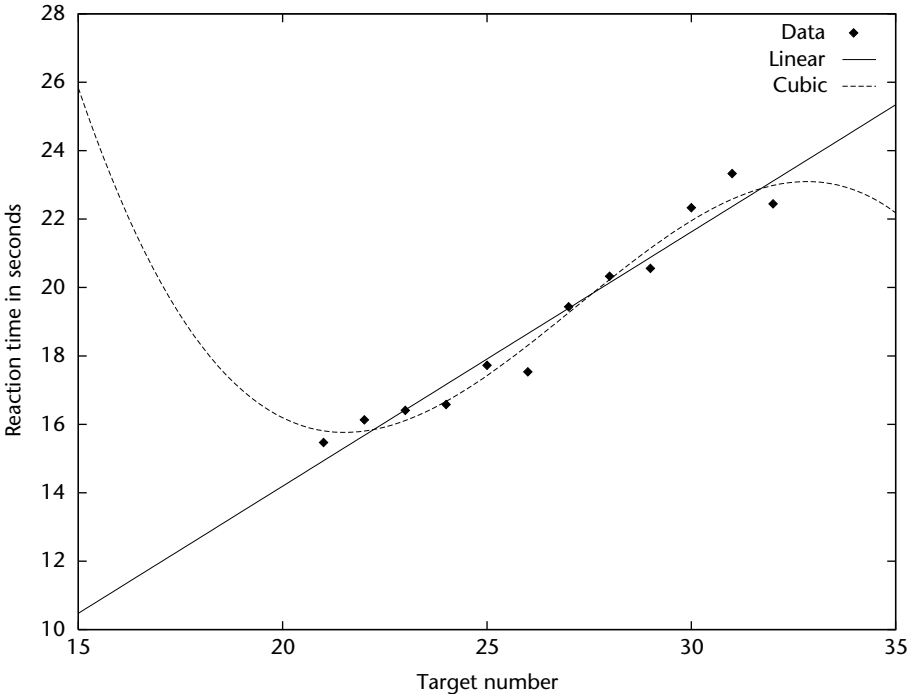


Figure 1. Mental counting. The data represent average time taken to count silently up to a given number. The best linear fit and best cubic fit are also shown.

The obvious computational characterization of mental counting is a simple iterative loop: begin with zero, then add 1 repeatedly until the target number is reached. We may reasonably assume that the counting will assume a rhythmic character, such that each iteration requires the same amount of time. Thus, overall reaction time T should be a linear function of the target number X:

$$T = aX + b \tag{1}$$

Here a and b are free parameters. Note that the parameter a has a straightforward psychological interpretation: the time taken for a single counting iteration. The parameter b denotes the amount of time needed to count from zero to zero – ideally this would be zero. Fitting this linear model to these data, we obtain a good fit, as shown in the figure ($R^2 = 0.9504$, $p < 0.0001$).

Let us compare this with the performance of a less constrained model, one based on a cubic function of the target number:

$$T = a_1X + a_2X^2 + a_3X^3 + b \tag{2}$$

This more general model provides a somewhat closer fit to the data ($R^2 = 0.9713$, $p < 0.0001$). The two new terms in the model approach significance. However, a comparison between the two models also highlights the three benefits of more constrained models.

Perhaps the most obvious benefit is that of parsimony. The linear model is simpler than the cubic model, which contains two terms in addition to those of the linear model. Model simplicity is aesthetically pleasing, and also makes the model easier to analyze.

A related issue is that the constrained model is less likely to be overly flexible. In this example, the less constrained cubic model overfits the data. With the obtained parameter settings, the cubic model predicts the observed data very well. But in doing so it also fits some of the noise, so that it generalizes poorly outside the observed range. As can be seen in the figure, this model predicts that it will take longer to count to 15 than to 30: an obviously false prediction. This problem is avoided by the more constrained linear model.

Finally, as we have seen, in this case the elements of the constrained linear model are motivated: they readily admit clear psychological interpretations, in the context of an iterative counting process. This is not true of the additional terms in the cubic model. The addition of such elements to a model, in the hope of increasing flexibility and thereby improving the fit to the data, comes at the price of arbitrariness.

These various benefits of constrained models need not all correlate perfectly. One could imagine, for instance, an ill-motivated model with few free parameters, or a well-motivated one with many. But the appeal of a constrained model lies in the prospect of accurately explaining a psychological phenomenon using a small number of well-motivated components.

These issues are relevant to a range of different styles of cognitive modeling, not just to simple models like that used in the example. Concerns about unconstrained models have been voiced in the context of symbolic models, and also in the context of connectionist or neural network models. For example, Miller *et al*. (1960) proposed a symbolic mental structure called a plan, which might be recursively created by other plans, yielding an overall system that was potentially quite complex and flexible. They anticipated that some might find their proposal too general and open-ended, and imagined their critics arguing as follows (quoted in Seidenberg, 1993):

> A good scientist can draw an elephant with three parameters, and with four he can tie a knot in its tail. There must be hundreds of parameters floating around in this kind of theory and nobody will ever be able to untangle them.

While Miller *et al*. defended their proposal against this critique, on the grounds that a complex system may in some instances be required, their sensitivity to the issue demonstrates an early concern with underconstrainedness in symbolic models. Another manifestation of the same concern may be found in Newell's (1990, p. 220) discussion of the computational universality of his production system framework SOAR, and the need for constraints in such a framework.

Some connectionist models have raised similar concerns. Massaro (1988) claimed that the multilayer perceptron – a commonly-used connectionist architecture – was too computationally powerful to be psychologically meaningful. His argument was based on the finding that a single connectionist model could simulate results generated by three mutually exclusive process models. Thus, the connectionist model appeared to Massaro to be overly flexible, and potentially unfalsifiable. And Cybenko (1989) showed formally that multi-layer perceptrons are in principle flexible enough to approximate arbitrarily well any continuous function over inputs that range from 0 to 1. Further discussion of constraints in connectionist models may be found in (McCloskey, 1991; Regier, 1996; Seidenberg, 1993; Siegelmann, 1995). We shall return to some of these issues below.

THE SOURCES OF CONSTRAINTS

There are many possible sources for constraints in a cognitive model. Four particularly important ones are treated here: constraints derived from known psychological structure, those derived from known biological structure, those derived from task structure, and those based on the nature of the input to the model.

Psychological Structure

It is natural for a computational model of one psychological phenomenon to be informed by – and constrained by – independent observations concerning another, related phenomenon. This would afford an explanation of the one mental process in terms of the other. For example, Gluck and Bower (1988) accounted for aspects of human categorization using a simple connectionist learning rule, the delta rule, which is ultimately motivated by studies of Pavlovian conditioning in animals (Rescorla and Wagner, 1972). This rule has known constraints: for example, it predicts blocking and overshadowing in learning. The strength of Gluck and Bower's presentation is that they find empirical evidence of these constraints in human category learning – thus mechanistically linking human categorization and animal conditioning. Another example of the use of psychological constraints may be found in Nosofsky's (1986) model of categorization. This model builds on an existing model of identification, or the discrimination of stimuli from one another (Shepard, 1957). Thus, the constraints of the original model – such as a monotonic decrease in generalization with increasing psychological distance – are incorporated into its successor. This implicity grounds an account of categorization in an existing account of identification.

Biological Structure

Biological constraints may also be brought to bear on cognitive models. This idea holds the potential of reduction, of explaining cognitive processes in terms of the neural structure that underlies them. Wilson *et al*. (2000) provide a concrete example. They present a model of how humans perceive the orientation of another person's head: whether the other person is facing one directly, or in partial profile. They are able to account for their empirical

findings in this domain using a model based on a population code of neurons found in area V4 of cortex, neurons sensitive to concentric and radial visual structure. Thus, the constraints of the underlying neural structure are used to explain a psychological phenomenon. Another example can be found in Regier and Carlson's (2001) study of spatial language. Participants were asked to rate the acceptability of linguistic spatial descriptions such as 'the dot is above the triangle', when shown pairs of objects in various spatial configurations. Their linguistic responses were well described by a model based in part on a neurobiological finding: the representation of overall direction as a vector sum. Thus, the linguistic data are partially grounded in a neurobiological constraint. More generally, one of the major appeals of connectionism as a modeling framework has been the prospect of bringing neural constraints to bear on psychological models (Feldman and Ballard, 1982; Seidenberg, 1993).

An important source of biological constraints is timing (Feldman and Ballard, 1982). Complex mentally-guided behavior can occur at a timescale of seconds. Assume, for example, that someone were to ask: 'Do you know where the registrar's office is?' It would take only a few seconds to hear the question, extract the intended meaning, recall where you believe the office is, prepare a linguistic response to the question, and deliver that response. Thus, a good amount of cognitive computational work is accomplished in a short timespan. But neurons, which are widely assumed to be the ultimate implementation of this computational work, operate on a timescale of a few milliseconds – relatively slowly by the standards of modern computers. Thus, the entire process of hearing and responding to the question must take place within only a few thousand neural computational timesteps. This constrains the number of iterations a model may realistically take when accomplishing such a behavioral or cognitive task. Since very little can be computed in a few thousand time steps using serial computation, this constraint strongly motivates parallel computation in cognitive models.

Task Structure

The nature of the psychological or behavioral task under study may also provide constraints on potential models. An example is the simple motor task of moving one's hand (or a pointer) to a target area of a specific size. Empirically, it has been found that the time required for this task varies as log (D/S), where D is the distance from starting point to end point of the motion and S is the size of the target region. This regularity is known as Fitts' law. An elegant model explaining this law has been given, based on the nature of the behavioral task. On this model, the most natural solution to the task is to launch the hand in the right general direction, and then iteratively correct during movement so as to bring the hand closer and closer to the target. This defines a recursion which, when solved analytically, yields the formula for Fitts' law (Keele, 1968; Meyer *et al.*, 1988; Newell, 1990). Another example is the simple mental counting task described earlier. Here, it is the intuitively obvious iterative nature of the counting task that motivates the linear model, the more constrained of the two models considered.

Input

An important source of constraint lying outside the model itself is the nature of the input supplied to it. Researchers investigating very flexible or general models often emphasize the constraints that reside in the model's input, at least as much as those residing in the structure of the model itself. Much connectionist modeling takes this approach. Specifically, as we have seen, multi-layer perceptrons are computationally quite general, and this generality has been a point of criticism. However, flexible mechanisms of this sort can be very useful scientifically, when considered together with the nature of the model input. A concrete example may be found in language acquisition. For many years, the study of language acquisition was dominated by the 'argument from poverty of the stimulus' (Chomsky, 1986). This view holds that the linguistic input heard by the language-learning child is too sparse, too impoverished to eventually give the child full knowledge of the syntactic structure of the language. Therefore, on this account, some elements of this knowledge must be innate. This account predicts that general-purpose learning mechanisms would not be able to learn the syntactic structure of natural language – as such mechanisms lack the requisite language-specific innate structure. This stance has been challenged recently. Very flexible, general-purpose connectionist networks have succeeded in learning artificial languages that resemble natural language in some important respects (Elman, 1993; Rohde and Plaut, 1999). This suggests that the input may not be as impoverished as had earlier been asserted, and that there may be no need to posit innate language-specific structure. Thus, the importance of these demonstrations lies

precisely in the unconstrainedness of the mechanism itself, and the substantial constraints present in the input.

Given this variety of possible sources of model constraints, is there a most appropriate source? Is there any reason to prefer a model constrained in one manner over a model constrained in another? The answer to this question ultimately lies in the nature of the scientific question being asked. A general mechanism with constrained input is useful in addressing the alleged necessity of innate structure. For other sorts of questions, however, it may be more informative to demonstrate that elements of already acknowledged mechanistic structure can explain a novel phenomenon, one to which they were not originally tied. In both cases, the data at hand are explained in terms of known and independently motivated constraints, whether these constraints reside in the mechanism itself or in environmental input.

SUMMARY

Over-generality is a potential danger in computational models of cognition. It is conceivable that a model may account well for a set of empirical data because of its extreme flexibility, rather than because of a clear match between its structures and those of the cognitive process under study. This possibility has been a concern in both symbolic and connectionist models of cognition. However, the problem can be avoided by constraining models in principled ways. This can yield a more parsimonious, less over-flexible, and more convincingly motivated model. There are several possible sources for constraints on cognitive models: existing knowledge concerning psychological or biological structure, the nature of the cognitive task itself, or the input to the model. The scientific question being posed will indicate the most appropriate source of constraint in a given modeling enterprise.

References

Chomsky N (1986) *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Cybenko G (1989) Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* **2**: 303–314. [Also available as report number 856, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, IL, USA.]

Elman JL (1993) Learning and development in neural networks: the importance of starting small. *Cognition* **48**: 71–99.

Feldman J and Ballard D (1982) Connectionist models and their properties. *Cognitive Science* **6**: 205–254.

Gluck M and Bower G (1988) From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General* **117**(3): 227–247.

Keele SW (1968) Movement control in skilled motor performance. *Psychological Bulletin* **70**: 387–403.

Lewis HR and Papadimitriou CH (1981) *Elements of the Theory of Computation*. Englewood Cliffs, NJ: Prentice-Hall.

Massaro D (1988) Some criticisms of connectionist models of human performance. *Journal of Memory and Language* **27**: 213–234.

McCloskey M (1991) Networks and theories: the place of connectionism in cognitive science. *Psychological Science* **2**(6): 387–395.

Meyer D, Abrams R, Kornblum S and Wright C (1988) Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological Review* **95**(3): 340–370.

Miller G, Galanter E and Pribram K (1960) *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart, and Winston.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Nosofsky R (1986) Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**(1): 39–57.

Regier T (1996) *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press.

Regier T and Carlson L (2001) Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General* **130**: 273–298.

Rescorla RA and Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: Black AH and Prokasy WF (eds) *Classical Conditioning II: Current Research and Theory*. New York, NY: Appleton-Century-Crofts.

Rohde D and Plaut D (1999) Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition* **72**(1): 67–109.

Seidenberg M (1993) Connectionist models and cognitive theory. *Psychological Science* **4**(4): 228–235.

Shepard R (1957) Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika* **22**: 325–345.

Siegelmann HT (1995) Computation beyond the Turing limit. *Science* **268**: 545–548.

Wilson HR, Wilkinson F, Lin L-M and Castillo M (2000) Perception of head orientation. *Vision Research* **40**: 459–472.

Computational Models: Why Build Them?

Introductory article

Herbert A Simon, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

CONTENTS

The term 'computational model' is used for a theory that is stated definitely enough that precise reasoning (not necessarily numerical) can be performed on it to infer the course of system behavior. Nowadays, such a theory is often expressed as a computer program, the computer tracing the system's path.

INTRODUCTION

A computational model traces a system's processes over time, determining its path from its present state and the influences impinging on it at each moment. By this definition, a clock, whose hour hand moves one degree of arc for every half-degree of rotation of the earth, is a computational model in astronomy. In physics, computational models commonly take the form of systems of differential equations: for example, Maxwell's equations for electromagnetism.

HOW COMPUTERS MODEL THE HUMAN MIND

Computational models are not new to the social sciences. Economics has been using them for more than a century and a half. Because much of economics is concerned with quantities and prices of commodities, many economic situations and can be translated directly into numerical form, and standard analytic methods, together with computer simulations, can be used to study them.

In cognitive science, simple numerical computational models have long been used in psychophysics and psychometrics: for example, in scaling stimulus–response relations and in factor analysis. A powerful new class of computational models, *physical symbol systems* (PSS), emerged when it was recognized (around 1950) that the patterns stored in computers could be used to represent not only numbers, but also symbols of any kind, including words in natural language or pictures and diagrams. A PSS, implemented by a computer program, may represent the successive states of the disks and pegs of a Tower of Hanoi puzzle while it is being solved, or successive states of the thought processes of the person solving it. This article focuses upon physical symbol systems, both serial and parallel (connectionist), which are the predominant forms of computational model in contemporary cognitive science; but it will comment on their relation to mathematical models as well.

To behave as a PSS, a computer carries out certain fundamental processes: receiving (*sensing*) patterns, encoding (*perceiving*) them, storing (*remembering*) them, evoking (*recognizing*) them on appropriate occasions, modifying symbol structures in memory (*reasoning*), and outputting symbol structures that initiate actions (*behaving*). Finally, the system's actions may depend upon what particular symbols it finds in memory (*branching* or *choosing*). In each case the symbolic process can be matched with the corresponding class of psychological processes.

Symbols, in this context, are any kinds of distinguishable patterns, which may be linguistic, diagrammatic, pictorial, or purely abstract-relational in character. Although 'symbol' is popularly used specifically for linguistic patterns, these have no privileged position in cognitive modeling. The essential characteristic of a pattern is that it be capable of denoting (pointing to) other patterns, as a line drawing may denote a cat, or a DNA sequence, or a protein.

The *physical symbol system hypothesis* postulates that any system possessing these capabilities, and only such a system, can be programmed, or

can program itself, to act intelligently. This is an empirical hypothesis to be confirmed or rejected by comparing the behaviors of such systems with human behavior, exactly as differential equations in physics are matched with physical systems.

A computer program, whether it processes numbers or non-numerical symbols, is formally a system of *difference equations*. Difference equations differ from differential equations only in that they move ahead by discrete increments of time instead of continuously (the increment being the time required to execute an instruction). Thus, the PSS hypothesis asserts that cognitive processes can be modeled by difference equations, numerical or symbolic or both in combination.

Computational models of cognition often not only simulate but actually carry out the solution process, and arrive at the solution (or fail to); they are both models of problem solving (or, more generally, thinking) and problem solvers. They describe and predict the course of thought, explaining its mechanisms while seeking to solve the problem. In this respect, they are different from the equations of physics, for equations representing the path of a planet do not move in such a path, but merely represent it, symbolically or graphically.

CONSTRUCTING AND TESTING MODELS

Some examples will show how models are constructed and what they do. Beginning with a very specific problem in algebra, the model is then generalized to a wide range of problems. This raises the question of how to accommodate the different ways in which different people may respond to the same problem: what invariants in cognitive processes constitute scientific laws of thinking?

An Algebraic Example

A program can model a student solving a linear equation in algebra, say, $7x - 15 = 3x + 9$. The solution takes the form $x = N$, where substitution of the number N for x satisfies the original equation. To discover the solution process, data are obtained from a student who writes down successive steps while talking aloud. In this case, she first adds 15 to both sides of the equation, obtaining $7x = 3x + 24$; then subtracts $3x$ from both sides, obtaining $4x = 24$. Finally, she divides both sides by 4, obtaining $x = 6$. At each step, she also collects terms (e.g. $9 + 15 = 24$). Each equation is transformed into another that has a closer resemblance to the final result.

A computer program can be constructed to follow this same procedure. The program applies a succession of *operators* that remove the differences between the current expression and the goal expression, leaving unchanged the number that satisfies the equation. The first step removes the unwanted number (-15) from the left side of the equation; the second step removes the unwanted term in x ($3x$) from the right side; the third step divides out the non-unitary coefficient of x (4). At each step the program performs the same operation on both sides of the equation and collects terms to simplify the result.

Deriving a More General Theory

Means–ends analysis and heuristic search

The method just described, *means–ends analysis*, is applicable to all sorts of problems, numerical or not. It is a powerful form of an even more general method called *heuristic search*. In heuristic search a problem is defined by a set of possible situations, or *problem space*. One of these is the *starting situation*, and any situation that satisfies a specified set of criteria is a *solution*. By applying a sequence of operators, the system can move through the space in search of a solution.

In the Tower of Hanoi we begin with three pegs *A*, *B* and *C* and a pyramid of disks, all stacked, say, on peg *A*. The disks must be moved, one by one, to other pegs, never placing a larger disk on a smaller one, until all have been moved to peg *C*. Each time a disk, starting with the largest, is moved successfully to peg *C*, a difference is removed between the initial and the final situation: fewer disks now remain to be relocated. The theory that people solve problems by using heuristic search, and means–ends analysis in particular, is captured in the computational model called the general problem solver (GPS). The theory has been shown to explain much human problem solving, but it clearly does not tell the whole story. More recently, GPS has been broadened into even more comprehensive models of cognition, including SOAR and the ACT-R family of programs.

Production systems

At a still more general level, the computational model proposes that the problem solving search is implemented by a mechanism called a *production system*. The choice of successive search steps results from the execution of 'if–then' rules (called *productions*), successors to the stimulus–response (S–R) pairs of behaviorist psychology. Whenever the

conditions of a rule ('ifs') are satisfied, its actions ('thens') are executed. The principal advance over S–R is that if–then rules, unlike S–R connections, can contain variables that are instantiated in each problem situation, and the stimuli and responses can be symbol structures within the brain as well as external sensory stimuli and motor responses. Thus, short-term memory at a given moment can supply one or more of the ifs of an if–then rule, and the action, the 'then', or part of it, may invoke knowledge in long-term memory. The algebraic problem described above might be solved mentally, holding the intermediate symbol structures internally and only writing down the answer.

Many problem spaces that humans search are very large. In chess, about 10^{20} branches emanate from each move. In the time available to make a move, a master can explore perhaps 1000 of these, a minute fraction. Similarly, in the algebraic problem, an infinite number of equations have the same value of x as the given one, and x has an infinite number of possible values. How does the student, after generating only three equations, find the right value?

The computer program of the example shows that knowledge of the operations of adding or subtracting the same quantity from both sides of an equation, and of multiplying or dividing both sides by the same quantity, combined with knowledge of means–ends analysis, is sufficient to solve the equation rapidly. The student's written work and verbal statements will show whether this is the process she used to solve it. Thus the computational model is a theory that can be tested in detail, and at several levels of generality, against the verbal and other behavior of human subjects, with a time resolution of about five seconds per action.

The model both tests the sufficiency of the theory's mechanisms to produce the observed behavior, and compares its information processes, point by point, with the observed processes of the human subject. This is precisely how we test any theory against data; a computational model is not special in this respect. The comparison tests its sufficiency to perform the task and the degree of its correspondence with human processes.

Model invariants

The model described above can only solve algebraic equations; but the architecture of the model, ignoring its knowledge of algebra, constitutes a broad theory of problem solving. The theory postulates that a problem solver defines a problem space for the given problem; then, operations appropriate to the problem domain, implemented as productions, search selectively through the space, guided by means–ends analysis. The validity of this architecture as a general theory of problem solving can be tested by constructing the appropriate problem spaces and if–then productions for various problem domains, then simulating the system's behavior for specific problems in these domains and comparing the behavior with that of human subjects.

The problem solving models, which make much use of information stored in semantic memory, can be joined to several programs, for example, the EPAM program, which serves as a perceptual 'front end', UNDERSTAND and ISAAC, which are capable of translating verbal problem statements into inputs to GPS-like problem solvers, and CaMeRa, which reasons from pictorial or diagrammatic information. In this way the theory can undergo successive stages of generalization.

Individual differences

Whereas laws of human behavior presumably describe invariants over our species, different people perform the same task differently, and a single person behaves differently on different occasions. A major source of variation is the difference between people's memory stores, which are augmented by new learning, and different subsets of which are invoked in each problem situation. There is also genetic and other variability in the biological structure of brains. Theories in cognitive science, including computational models, must somehow extract invariants from this variability.

Invariant relations may be obtained in several ways. Laws may describe nonidentical but similar behaviors: 'When people are hungry, they search for food.' This is a (weak) law of behavior. The food varies greatly between cultures, with availability, and from one time to another, as do the methods of search (hunting, fishing, visiting the supermarket, etc.). The invariant is that people take actions that are relevant to the goal of food gathering. A theory of problem solving (like those already discussed) must deal with goals, and with processes for attaining them, taking as boundary conditions the goal-arousing circumstances and priorities among competing goals, and as initial conditions the actor's knowledge of ways of attaining goals and skills in pursuing the paths that are seen.

Because much variability derives from differences in what subjects know and attend to, an important approach to variability, and to the independent estimation of the parameters needed for testing models, is to design experiments that

explicitly manipulate the knowledge available to subjects or modify the focuses of their attention or their goals, inducing them to adopt particular strategies.

These considerations are relevant whether behavior is studied at the level of symbolic processes or of neural activity. All human behavior is ultimately implemented by neurons, just as computer behavior is by chips. The neuronal activity is just as susceptible to individual variation deriving from inheritance, experience, and focus of attention as is the more aggregated symbolic activity it produces.

The variations in a system's structure, knowledge, and attention serve as initial and boundary conditions for a computational model, and (like parameter values in numerical models) must be supplied before the model can simulate the behavior of a particular subject in a particular situation. Of course, in many tasks, members of a group (say, college sophomores) will all behave in much the same way. To the extent that this is so, we can construct a model of a 'typical' subject by averaging over subjects in each experimental condition.

Unified theories of cognition

A model may contain so many parameters that it can fit almost any data, hence predict nothing. One solution to this problem is to identify a range of tasks throughout which a certain set of parameters play a significant role, then fit a single set of parameter values to data from all the tasks. This strategy can be pursued within unified theories of cognition, which aim to embody the whole range of cognitive processes, hence are testable in almost all task environments. For such theories, the ratio of number of data points to number of parameters will increase with the range of tasks.

A more conservative strategy is to take advantage of the cognitive system's subsystems, and to construct computational models of the subsystems, with the aim of ultimately joining them. This strategy reaps the usual dividends of 'divide and conquer', but needs to be carried out with due attention to the compatibility of the component models, so that they can later be joined, and to defining components that are capable of performing a variety of tasks, the more the better. As an example of design for compatibility, the UNDERSTAND system was constructed some years after GPS with the condition that the output of the former should be a suitable input to the latter. With respect to range of tasks, the EPAM system, using the same set of parameter values throughout, has been applied to perceptual, learning and expert memory tasks, and tasks of categorization and concept attainment.

The two most significant efforts to move towards a unified system are the Soar system of Newell, Rosenbloom and Laird, and J. R. Anderson's ACT systems. The 'divide and conquer' strategy has motivated the development of the set of programs that includes GPS, EPAM, UNDERSTAND, CaMeRa, and a model of scientific discovery.

All the problems that human variability, the complexity of the cognitive system and the multiplication of parameters pose for cognitive modeling are present to the same degree in other modes of theorizing about cognition. Vagueness may create the illusion of lawfulness; it cannot create the reality. Without the discipline of modeling, the sources of variability are not likely to be recognized as clearly or dealt with as carefully. The rigor and unambiguity of cognitive models brings these problems forcibly to mind, and allows them to be addressed with precision and clarity.

MODELS OF LEARNING

Because human (and animal) responses are highly modifiable, learning processes play a major role in cognitive psychology, and many computational models are theories of learning. In fact, learning processes are generally closer than task behaviors to being invariants of behavior, although learning processes are also modifiable as people learn to learn. Among the models of learning mechanisms are *adaptive production systems* (APS), *parallel distributed processors* (PDP), the 'elementary perceiver and memorizer' (EPAM), the *chunking* mechanism of Soar, and a number of similar learning processes in the ACT-R system. Several or all of these different mechanisms (and perhaps others) may be required to account for the full range of human learning capabilities.

Adaptive Production Systems

The algebraic example discussed earlier will illustrate how an adaptive production system can learn. An APS constructs new productions, and inserts them in its memory along with those already there. These new productions will henceforth be executed along with those previously stored. In each new task environment, the APS mechanism calls for a distinct set of differences and operators for building the new if–then rules.

Suppose an APS is presented with a set of worked-out examples of problems like the

algebraic problem examined above, having learned previously what operations can legitimately be performed on an equation. It examines the first two lines of a worked-out example to discover what change was made in the equation, what operator was used to accomplish it, and what difference between initial expression and solution was removed. It then constructs a new production whose 'if' clause is the difference that was removed and whose 'then' clause is the operator used to remove it. The clauses are then generalized to permit any numbers to replace the specific numbers in the example. Thus, the APS initially constructs the production: 'If -15 appears on the left side of the equation, then add 15 to both sides and simplify.' It then generalizes it to: 'If a numerical term, N, appears on the left side of the equation, then add N to both sides and simplify.' The first APS with these capabilities was constructed by D. Neves.

Parallel Distributed Processors

Parallel distributed processors, which derive their original inspiration from Hebb's 'cell assemblies', use elements that somewhat resemble simplified neurons, or collections of neurons. Each such element consists of a link connecting two nodes. Each node has a modifiable level of stimulation, and each link an activation level. The links are typically organized into an input layer, one or more 'hidden' layers, and an output layer, the layers being arranged sequentially. A stimulus (a set of features) activates one or more nodes (each corresponding to a feature) in the input layer to varying levels of stimulation; each node then contributes to the activation of the links it is connected to; these links contribute to the stimulation of the input nodes of the hidden layer; and so on. Each final node in the output layer corresponds to a possible output, and the PDP responds with the output of the node that has the strongest total stimulation. A PDP model can learn to recognize distinct stimuli, and to conceptualize, by grouping them together, stimuli that share certain properties.

To learn, for example, to associate each letter of the Roman alphabet with a distinct set of input features, the system is given feedback on the correctness of its response to each set of features presented. A correct response increases the activation of the links that contributed to it, and decreases the activation of the other links; after an incorrect response, the reverse occurs. This simple scheme has been used to construct systems that can discriminate and conceptualize.

Discrimination Nets

Other kinds of learning which are closely related but organized in serial instead of parallel fashion, are modeled by the 'elementary perceiver and memorizer'. EPAM, in learning, grows a network of nodes and links. Each node contains a test for the value of some feature of the stimulus; the system tests this feature, then moves to a new node along the link that represents the stimulus's value on the test. If, for example, a test discriminated among various sizes of birds (wren-size, robin-size, chicken-size, etc.), then EPAM would move to a new node, following the link corresponding to the size of the bird being recognized. At the new node, EPAM tests another feature (the bird's color, say), and repeats the process until a terminal node was reached.

If given feedback when it is right or wrong, EPAM can learn to recognize large numbers of different patterns (an EPAM net has learned to distinguish 300 000 different patterns of chess pieces on chessboards). It can also store relevant information at each node in the net; hence it can serve, for example, as a medical diagnosis system which, at each node, has information about the disease whose symptoms would cause it to be sorted there. EPAM is a computational model of substantial breadth and power which has modeled a wide range of the phenomena that have been observed in memorization, discrimination, and categorization experiments.

Learning by Compilation ('Chunking')

Yet another learning mechanism is employed by Soar, an adaptive production system that can improve the functionality of new productions it has acquired by 'chunking' them. When two or more processes are frequently performed in sequence, their performance can be greatly speeded up by 'automating' them: that is, by removing time-consuming perceptual tests that would otherwise be used to determine what to do next, but which are unnecessary if the sequential process becomes automatic.

Comparing Learning Models

The existing evidence suggests that human learning probably employs all of the methods that have been examined here, and perhaps others as well. Nevertheless, close comparison of the behavior of different learning systems in approaching the same task can advance our understanding of the

contribution of each mechanism to their performance, and can detect commonalities of mechanism among them.

For example, the EPAM and PDP models perform almost identically in a well-rehearsed experiment that found that letters embedded in four-letter words are recognized more rapidly than letters embedded in nonsense words. An investigation into how two such different architectures (serial versus parallel, symbolic versus numerical) could arrive at the same predictions showed that both models recognize words by recognizing the letters of which they are composed: EPAM by first recognizing letters, then words, by a recursive process; PDP systems using two serially arranged hidden layers with feedback from the 'word' layer to the 'letter' layer. Comparisons like this, of different models in performing the same task, shed valuable light on the functional equivalence of apparently quite different processes, and separate those features of theories that are important for accounting for the phenomena, from those that are merely artefacts of implementation. Such comparisons are important for choosing among proposed models and mechanisms.

SYMBOLIC AND NUMERICAL MODELS

This article has emphasized symbolic rather than numerical models. Removing the necessity of expressing the qualitative differences among architectures in numbers greatly extends the range of systems and realistic detail of mechanisms that can be modeled. For example, traditional mathematical learning models applied to conceptualization tasks generally compare feature vectors of two stimuli, using some kind of correlation between them as a single measure of similarity. As we have seen, EPAM and PDP systems employ much more detailed mechanisms of comparison, hence can answer much more specific questions about the process. No one would propose that computation of correlations is the actual psychological mechanism of comparison, whereas EPAM and PDP systems embody plausible candidate mechanisms. Observations of behavior may allow such claims to be proven.

LEVELS IN COMPLEX SYSTEMS

Not all computational models describe cognition at the same level of detail or in terms of processes on the same timescale. A similar diversity of theory is familiar in the other sciences. In physics, we have classical thermodynamics, which theorizes about heat and temperature averaged over bodies of substantial size. On the other hand, kinematic theories describe interactions of individual particles like atoms or molecules; while statistical mechanics treats of similar detail, but only on a probabilistic basis.

Levels in Scientific Theory

In any science there exist theories (and usually important theories) at each of several levels. This is possible because most complex systems observed in the world are constructed in a hierarchy of levels, and the elements are joined in such a way that the interaction between elements at any single level can be described without specifying any but very general properties of the elements at the next level below, and without considering dynamics at the next level above.

For example, a car can be described in terms of ignition, engine, transmission, and wheels, without describing anything about these components except their functions and their mutual connections. The ignition, at the proper time, fires the fuel from which the engine obtains energy to turn the drive shaft. The latter, connected to the wheels by the transmission, turns them, moving the car down the road. However, we can also fill a book describing each of these components in terms of their subcomponents, but still without mentioning details about their atomic structure.

Levels in Cognitive Theory

The human mind and brain are also organized as a hierarchy of components, each of these components made up of subcomponents, and so on. Thus we speak of the brain, then of the cerebrum, the cerebellum, the hippocampus, and so on; or of long-term memory, short-term memory, visual senses, auditory senses, and so on. Components can be recognized because there are stronger and more frequent interactions among the elements of a component than between distinct components on the same level.

The elementary actions and events among components at any given level usually take place on the same general timescale, shorter times as we go lower in the hierarchy, longer times as we go higher. Thus, a simple reaction to a sensory stimulus may take half a second, a step in solving an algebraic problem a couple of seconds, solution of a homework problem ten minutes, performance of a sequence of scientific experiments three weeks.

As in the biological and physical sciences, a theory of cognition can be built at each broad level to describe and explain the phenomena and processes at that level without attention to the detailed structure of the levels below, treating the latter as already in steady state and treating the levels above as constant constraints for the time interval of interest. This near-independence between levels follows mathematically from the comparative independence of components from each other as compared with their high level of internal interaction. Typically, descriptions refer to events at a particular level, whereas explanations account for events at one level in terms of processes at the next level below.

Three major levels of theory have been identified in cognitive science: the knowledge (or representational) level, of events that are minutes or more in duration, the symbolic level, whose events are seconds or tenths of seconds in duration, and the neural level, whose events may endure for only milliseconds or tens of milliseconds. Both the symbolic and the neural levels are possibly divisible into sublevels. Computational models of various aspects of thinking have been built at all of these levels.

The knowledge level

The behavior of a goal-oriented organism can often be predicted, in first approximation, by asking what actions would be effective, in its given environment, to attain its current goals. This is like predicting the shape of jelly from the shape of the mould in which it set. It requires information only about the goal and the environment, and none about the processes used by the organism to select the goal-attaining behavior. In ignoring the mechanisms, it is predictive and descriptive, rather than explanatory.

For example, the dominant theory in twentieth-century economics was that economic decisions maximize achievement of a goal called *utility*. A real system can behave in this way only if it has capabilities for determining in any situation which of all the available choices would yield this maximum. To predict its behavior, the goals embodied in the utility function and their relative weights have to be specified. The term 'rational' is often reserved for behavior that satisfies these criteria.

In some situations in real life, where we know what people value (what has utility), and they know what alternative actions are available and how to implement them, rationality defined in this way can be a powerful tool for prediction. This level of theory is called the 'knowledge level', calling attention to the centrality of an actor's knowledge of the environment as both generating and limiting the possible actions for reaching goals and deriving utility. If the environment is simple enough and the actors are rational, then their behavior (like the behavior of the jelly) can be predicted from the shape of the environment (the mold), specifying no other characteristics of the actors than their goals. The fulfillment of knowledge-level predictions provides no information about the mechanisms of human thought, except that there exist mechanisms of making simple goal-oriented decisions in these situations.

Thus, if we bring a college student into the laboratory with instructions to point to the taller of a five-foot and a six-foot person standing in the room, we can generally predict which person will be selected. We are assuming that the student has the goal of performing the task and sensory and mental capabilities for picking the taller person. However banal these conditions, they are at the core of descriptions of behavior at the knowledge level.

Two examples of knowledge-level theory in psychology are J. J. Gibson's theory of affordances, and the 'rational' (R) component of J. R. Anderson's ACT-R model. Gibson, for example, characterized a pilot landing a plane in terms of the characteristics of the landing field's surface (e.g. the size variation of ground patterns with distance), implicitly assuming that these characteristics were visible to the pilot, who understood their implications for the position of the plane. Similarly, Anderson's ACT-R model has built-in assumptions that certain learning processes will use information optimally. No strong claim is made that the mechanisms in the model for these processes correspond to the real mechanisms (those that would be discovered by exploration at the symbolic or neural levels).

An important argument for analysis at the knowledge level is that evolutionary processes tend to produce a fit between environmental properties and the behaviors of adaptive organisms in response to them. If sitting is sometimes an important goal, then the organism will become able to recognize objects that afford 'sittability', perhaps logs and chairs among them.

The symbolic level

At the next level below the knowledge level – the information processing, or symbolic, level – the theory takes account not only of the goals and external environment, but also of the actor's knowledge, and ability to use that knowledge to recognize, search and reason. For example, most

means–ends analysis is at this level. PDP and other 'neural net' models are sometimes said to belong to the neural level, or sometimes to combine both levels, denying any separation in this range.

Information processing theories have been especially successful in modeling complex, often ill-structured, professional-level problem solving and design tasks, including chess playing, medical diagnosis, and scientific discovery. Such tasks require intuition, insight, and even creativity. Similar programs compose music, draw, or paint. All of these theories share a common core: they employ recognition and heuristic search as the basic mechanisms in problem solving. Hence, they are instances of a more general theory.

Bounded rationality

Knowledge-level theories, particularly those that assume optimization, have been criticized as ignoring the boundedness of human rationality: limited knowledge, and limited ability to compute consequences and compare incommensurate goals. Defenders of optimization reply that the bounded rationality viewpoint merely calls attention to the occurrence of human irrationality without proposing an alternative; and bounded rationality has sometimes been interpreted incorrectly as declaring that human behavior is largely irrational.

The principal contact between the proponents of knowledge-level optimization and the proponents of bounded rationality has been through research, often published in the fields of economics or decision theory, that attacks specific assumptions of economic theory, especially the consistency of the utility function.

By introducing the variability of human behavior, symbolic-level theories become less general and require the specification of numerous parameters before they can be applied to specific situations. This issue has already been discussed, and ways described of maintaining a high ratio of data points to parameters. It should also be noted that, in actual applications, most optimization models that ignore the limits on human rationality are (and must be) applied to complex real-world situations only after much abstraction and simplification of details, and that these approximations are generally carried out in a very unsystematic way without direct measurement of their effects.

The neural level

It is an important goal of science to connect theories at adjoining levels: for example, symbolic-level psychology with neuroscience, or neuroscience with biochemistry. The major explanatory theories in science use the mechanisms at one level to account for the mechanisms at the next level above without challenging the value of the theory at either level. The second half of the twentieth century saw a rapid growth in knowledge about human cognition at both symbolic and neural levels, as well as at the linkage between neural and biochemical levels. Understanding of the connections between symbolic and neural levels has developed more slowly, but in recent years, more attention has been paid to these connections, and the instruments for discovering them have been greatly improved. The modeling of such connections between levels is likely to receive much attention in the years immediately ahead.

SUMMARY

Cognitive science aims to understand systems as complex as physical and other biological systems. To reach that understanding, scientific representations are needed that provide the same kind of precision and reasoning power that physics, chemistry and biology get from mathematics and from the formal notations that describe chemical structures and processes. Psychology was severely handicapped by the looseness of the language and associated inference processes traditionally employed in its theories. In the past, a major deterrent to the employment of clear representations was that they could only be used if the phenomena were translated into mathematical language.

The modern computer created, to the surprise even of its inventors, a new set of formalisms appropriate for describing and analyzing information processing systems, without needing to cast these descriptions in numerical form. Computer programs can simulate cognition at the level of its information processes, even to the point of actually carrying out the cognitive tasks. Today, we have theories, in the form of operative programs, that give clear and precise accounts of most of the human cognitive processes at the symbolic level that enable people to perform complex, often ill-structured, tasks at professional, and sometimes creative, levels. These accounts may be incorrect in places. But we know that expert behavior derives from knowledge-based recognition and search processes, and we know how this is accomplished in an important collection of task domains.

In cognitive science today, our capabilities for stating theories at the symbolic level far exceed the power of our tools for actually observing the cognitive phenomena in the detail required to test the theories rigorously. As observational means

improve, for example, by advances in the temporal resolution of functional magnetic resonance imaging (FMRI) studies, one can hope that the balance can be restored between our theory building and our observational capabilities.

Further Reading

Anderson JR (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Feigenbaum E (1961) The simulation of verbal learning behavior. In: *Proceedings of the Western Joint Computer Conference* **19**: 121–132.

Friedman M (1953) *The Methodology of Positive Economics*. Chicago, IL: University of Chicago Press.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Gigerenzer G and Todd PM (1999) *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.

Gobet F, Richman H, Staszewski J and Simon HA (1997) Goals, representations, and strategies in a concept attainment task: the EPAM model. In: Medin DL (ed.) *The Psychology of Learning and Motivation*, vol. xxxvii. San Diego, CA: Academic Press.

Hayes JR and Simon HA (1974) Understanding written problem instructions. In: Gregg LW (ed.) *Knowledge and Cognition*. Potomac, MD: Erlbaum.

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: John Wiley.

Hiller LA and Isaacson LM (1959) *Experimental Music: Composition With an Electronic Computer*. New York, NY: McGraw-Hill.

Iwasaki Y and Simon HA (1994) Causality and model abstraction. *Artificial Intelligence* **67**: 143–194.

Langley P, Simon HA, Bradshaw GL and Zytkow JM (1987) *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, MA: MIT Press.

McCorduck P (1991) *Aaron's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*. New York, NY: Freeman.

Medin DL and Smith EE (1981) Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory* **7**(4): 241–253.

Neves DM (1978) A computer program that learns algebraic procedures by examining examples and working problems in a textbook. In: *Proceedings of the Second Conference of Computational Studies of Intelligence*, pp. 191–195. Toronto: Canadian Society for Computational Studies of Intelligence.

Newell A (1973) You can't play 20 questions with nature and win. In: Chase WG (ed.) *Visual Information Processsing*, pp. 283–308. New York, NY: Academic Press.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Simon HA (1961) GPS: a program that simulates human thought. In: Billings H (ed.) *Lernende Automaten*, pp. 109–124. Munich: Oldenbourg.

Newell A and Simon HA (1976) Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery* **9**(3): 113–126.

Novak GS (1977) Representations of knowledge in a program for solving physics problems. In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Cambridge, MA.

Richman H and Simon HA (1989) Context effects in letter perception: comparison of two theories. *Psychological Review* **96**: 417–432.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 vols. Cambridge, MA: MIT Press.

Shortliffe EH (1974) *MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection*. Stanford, CA: Stanford University Press.

Simon HA (1983) *Reason in Human Affairs*. Stanford, CA: Stanford University Press.

Simon HA and Ando A (1961) Aggregation of variables in dynamic systems. *Econometrica* **29**: 111–138.

Tabachneck-Schijf HJM, Leonardo AM and Simon HA (1997) CaMeRa: a computational model of multiple representations. *Cognitive Science* **21**(3): 305–350.

Tversky A and Kahnemann D (1981) The framing of decisions and the psychology of choice. *Science* **211**: 353–358.

Computer Modeling of Cognition: Levels of Analysis

Introductory article

Michael RW Dawson, University of Alberta, Edmonton, Alberta, Canada

CONTENTS

According to Marr's three-level hypothesis, information processors must be described at the computational level, the algorithmic level, and the implementational level in order to be fully understood.

MODELING AT SEVERAL LEVELS: DESCRIBING COGNITION FROM MULTIPLE PERSPECTIVES

On 11 May 1997 a computer system called Deep Blue defeated world chess champion Garry Kasparov in a six-game match by a score of 3.5 to 2.5. For researchers interested in modeling high-level cognitive activities like thinking and problem solving, it is instructive to compare Deep Blue and Kasparov, and to consider in what ways these two chess players are different and in what ways they are similar.

At one level of analysis – that of their physical components – they are clearly quite different. Deep Blue is constructed from silicon circuitry, while Kasparov's chess playing ability is based in a biological brain. However, if we consider them from a more abstract perspective, we find that they are similar. At some level of description Deep Blue and Kasparov both fall into the same class of systems, because both can be described as 'chess players'. Furthermore, if one were to only observe a record of the moves made by these two systems, then it would appear that they were very similar because both played chess at a very high level of ability, and in many situations they would choose to make similar moves.

ANALYSIS OF TASK, ALGORITHM AND IMPLEMENTATION

From the example above, it is clear that when researchers wish to compare a computer model with a person or system being modeled, they must consider this comparison from a variety of perspectives. Similarly, when examining some phenomenon of interest in the hope of developing a computer simulation, the phenomenon must be investigated at a number of different levels of analysis.

The vision scientist David Marr argued convincingly that when information processors were being studied, three general levels of analysis were relevant. Each of these levels corresponds to asking a particular kind of question.

The first analysis proposed by Marr is at the *computational* level. At this level of analysis, a researcher is primarily concerned with the question: 'What information processing problem is being solved by the system?' In the chess example, this would involve specifying the abstract 'rules of chess', and classifying a system (e.g. Deep Blue or Kasparov) as being a chess player provided it obeyed those rules. Marr provided many examples in vision in which a computational analysis involved proving that the visual system could be described as solving a particular type of mathematical problem (such as a constrained optimization problem) when it was detecting some visual property (such as a pattern of motion). In general, because computational-level analyses often involve specifying abstract laws, they require researchers to use formal techniques such as mathematical or logical proof. In other words, the answer to the computational-level question is usually a mathematical statement that an information processing system is solving a particular problem; this statement can characterize the abstract laws that define the problem, or that impose constraints on how solutions to the problem are to be achieved. Such a statement can be applied either to a human subject or to a computer simulation.

The second analysis proposed by Marr is at the *algorithmic* level. At this level of analysis, a researcher is primarily concerned with the question: 'What procedures, program, or algorithm is being used by the system to solve the information processing problem?' There are many different ways of solving the same information processing problem. So, while Deep Blue and Kasparov may be equivalent when compared at the computational level, because they are both playing chess, they might at the same time be very different at the algorithmic level because they are using very different procedures to choose their next move.

Answering the algorithmic-level question is the primary goal of empirical cognitive scientists, such as experimental psychologists and psycholinguists. They use the research methodologies of the behavioral sciences to try and determine the procedures used by humans to solve information processing problems. This attempt usually separates into two different investigations. The first attempts to determine the general information processing steps (the program) being used to solve the problem. The second attempts to determine the primitive properties of symbols and the processes that act upon these symbols (the programming language). Many fundamental debates in cognitive science, such as the debate between proponents of symbolic connectionist models, are debates about the nature of the 'programming language' for human cognition. Zenon Pylyshyn has called this 'programming language' the functional architecture of cognition.

The third analysis proposed by Marr is at the *implementational* level. At this level of analysis, a researcher is primarily concerned with the question: 'What physical properties are responsible for carrying out the program or algorithm?' At the algorithmic level, it is sufficient to state what the components of the functional architecture are. At the implementational level, one has to explain how these components are 'brought to life' by an actual physical device. For human cognition, in which we assume that the brain is responsible for carrying out information processing, this requires researchers to determine how brain states and neural circuits are responsible for instantiating the functional architecture. Not surprisingly, the research methods of neuroscience provide the primary tools for answering the implementational-level question.

THE EXAMPLE OF COLOR PERCEPTION

To illustrate the three-level approach, let us consider color perception. In the seventeenth century, Sir Isaac Newton performed experiments with prisms and light and published the results in his book *Opticks*. As far as color vision was concerned, his great discovery was that some colors could be 'constructed' from combinations of others. Newton proposed that human color vision was based upon the perception of seven primary colors. Many alternative theories have been proposed. Goethe proposed a two-color theory in 1810; Young proposed a three-color theory in 1810; four-color theories were proposed by both Hering and Ladd-Franklin in the late nineteenth century.

The question of what is the minimum number of primary colors required for a complete theory of human color perception is a computational-level question. The physicist James Clerk Maxwell answered this question in 1856. Using a 'color solid' proposed by Newton, and an elegant geometric proof, Maxwell was able to show that any perceived color could be expressed in terms of three primary colors.

Maxwell's proof provides important information about color perception in principle, but leaves many unanswered questions about human color vision in practice. For instance, does the human visual system base itself upon the minimal number of primary colors? If so, what three primary colors does it use? These questions arise at the algorithmic level, and must be answered by experimental studies of human vision. In 1873, Helmholtz presented a series of lectures which popularized Young's three-color theory of color perception. This led to decades of psychophysical experimentation, such as having subjects differentiate pure colors from colors created by mixing combinations of light. The results of these experiments supported the Young–Helmholtz theory, indicated that Maxwell's proof applied to human color vision, and demonstrated that our perceptions of color arise from combining the sensations of the three primary colors red, green, and blue.

The answers to the algorithmic-level questions provide a great deal of information about the basic information processing steps used in color vision. However, they give no information about the physical mechanisms that are involved. Helmholtz himself was painfully aware of this. In 1873 he said: 'It must be confessed that both in man and in quadrupeds we have at present no anatomical basis for this theory of colors.' It wasn't until the second half of the twentieth century that a physical account of the Young–Helmholtz theory became possible. Researchers discovered three different kinds of cone cells in the retina, each containing a different light-sensitive pigment. A technique

called microspectrophotometry, in which a small beam of light is passed through individual receptors that have been removed from the retina, revealed that these pigments are sensitive to different wavelengths of light: one is most sensitive to red light, one to green light, and one to blue light.

This example illustrates two general points. First, it shows that a psychological phenomenon can be examined at a number of different levels. Indeed, complete understanding of a phenomenon requires computational, algorithmic, and implementational accounts. Second, this degree of understanding is not easy to achieve. Our detailed modern theories of color perception are based upon several centuries of research.

INVESTIGATING MIND AT A VARIETY OF SCALES

The three levels of analysis that have been described above are not the only approach to understanding and modeling cognitive phenomena. For example, neuroscientists have argued that cognitive science requires a complete understanding of the brain, and that this understanding requires that the brain be examined at a number of different levels. This means analyzing the brain at a number of different scales. The neuroscientist Gordon Shepherd calls these levels of organization. At the largest scale, one studies the behavior of the whole brain. Reducing the scale to the next level of organization, one studies large systems and pathways in the brain (e.g. sensory pathways). At the next level, one studies the properties of specific centers and local circuits. Reducing the scale once more, one studies the properties of neurons (for example, through single cell recording). At even more microscopic scales, one studies structures within neurons (for example, systems of connectivity involving dendrites or axons). Reducing the scale further, one studies individual synapses. Finally, one can study the molecular properties of membranes and ion channels.

Patricia Churchland and Terry Sejnowski have cited these levels of organization as an argument against the three-level approach proposed by Marr. They suggest that because there are so many levels of organization, one cannot talk about a single implementational level. Furthermore, because different implementational levels might have different task descriptions, there will be a multiplicity of algorithms, so that it is inappropriate to talk about a single algorithmic level.

However, Marr's three levels were proposed as properties of an investigation, and not of the system being investigated (that is, they are an epistemology, not an ontology). Marr did not claim that information processors 'have' an algorithmic level. Rather, he claimed that to understand an information processor, one must answer the algorithmic-level questions. These questions need not be applied monolithically to an entire system. Thus, one can ask Marr's three questions of any level of organization in the brain, provided the level in question can be usefully considered as being involved in information processing. The 'levels of analysis' and 'levels of organization' approaches are not mutually exclusive, but are complementary.

SMOLENSKY'S SYMBOLIC AND SUBSYMBOLIC LEVELS

An alternative view of levels has arisen from attempts to relate symbolic models of cognition to connectionist simulations. Paul Smolensky has argued that in the context of connectionist simulations one can distinguish between a symbolic account and a subsymbolic account. To say that a connectionist network is subsymbolic is to say that the activation values of its individual hidden units do not represent interpretable features that could be represented as individual symbols. Instead, each hidden unit is viewed as indicating the presence of a microfeature. Individually, a microfeature is unintelligible, because its 'interpretation' depends crucially upon its context (i.e., the set of other microfeatures that are simultaneously present). However, a collection of microfeatures represented by a number of different hidden units can represent a concept that could be represented by a symbol in a symbolic model. From this perspective, a symbolic account of a network is only an approximate account. A more accurate and complete account of how a network solves a problem can only be found by analyzing it at the subsymbolic level.

The distinction between symbolic and subsymbolic levels is both interesting and important. However, this distinction too is compatible with Marr's approach. One could argue that a subsymbolic account of a network (e.g. a description of its microfeatures) would represent a description of its functional architecture while a symbolic account of the network would represent a description of an algorithm (e.g. a description of symbolic representations), as well as how this algorithm is composed from the functional architecture (the microfeatures).

This observation shows that the three-level approach can be applied to connectionist models as

well as to more traditional symbolic computer simulations. Michael Dawson has argued that the three-level hypothesis applies equally to these two streams within cognitive science because connectionist and symbolic modelers both agree with the general assumption that cognition is information processing. This is an important point because many researchers are concerned that connectionist networks are not fully-fledged cognitive theories, but only provide an implementational account of cognitive phenomena.

SUMMARY

Information processing systems are very complex. In order to provide a complete account of an information processor, or to build a computer simulation of a cognitive phenomenon, three different levels of analysis are required. The computational level addresses the question: 'What information processing problem is being solved?' The algorithmic level addresses the question: 'What information processing steps are being applied to solve the problem?'. Thus the algorithm is distinguished from the functional architecture. The implementational level addresses the question: 'What physical properties are responsible for bringing these information processing steps into existence?' One reason why cognitive science is such a multidisciplinary undertaking is that the answers to these questions require vocabularies and methodologies from such diverse disciplines as mathematics, psychology and biology.

Further Reading

Bechtel W and Abrahamsen A (1991) *Connectionism and the Mind*. Cambridge, MA: Blackwell.

Dawson MRW (1998) *Understanding Cognitive Science*. Oxford: Blackwell.

Dennett DC (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Fodor JA (1968) *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York, NY: Random House.

Jackendoff R (1992) *Languages of the Mind*. Cambridge, MA: MIT Press.

Marr D (1982) *Vision*. San Francisco, CA: Freeman.

Medler DA (1998) A brief history of connectionism. *Neural Computing Surveys* **1**(2): 18–72.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Smolensky P (1988) On the proper treatment of connectionism. *Behavioural and Brain Sciences* **11**: 1–74.

Computer Vision

Introductory article

John K Tsotsos, York University, Toronto, Canada

CONTENTS

Computer vision is a discipline whose major concern is the development of computational methods, in terms of both hardware and software, to acquire, process, and interpret visual images.

INTRODUCTION

Computer vision systems attempt to recover shapes of objects, illumination direction, motion, object structure and identity, composition, and so on, from a digital image of a scene. The purpose of such systems is either to help us understand the contents of the image or to elicit an appropriate response from an autonomous system such as a robot. The interpretative process often occurs in the context of a particular domain. The goal of research in computer vision is to devise methods that approach or even exceed the performance of human vision. In addition, many researchers use the language of computation to formalize theories of biological vision.

Computer vision is difficult. At least half of human (or primate) cerebral cortex is involved in visual processing: this represents a huge amount of neural processing power. Some specific reasons for the difficulty of computer vision are described below.

Problems Inherent in the Analysis of Images

A camera is used to acquire a two-dimensional digital image of a three-dimensional scene. Imagine viewing an unknown scene with one eye and trying to estimate distances in that scene: you cannot do so reliably. Explicit information about depth is not present in a single image. The depth of objects in the scene must be inferred either indirectly or by using a binocular camera system. Indirect inference may use cues in the image such as occlusion (one object being in front of another).

Nor does an image contain explicit information about where the source of illumination is located. The computer system must infer this by considering positions and shapes of shadows, highlights, and shading on surfaces.

Often knowledge of where the scene is located and what the participants in the scene are doing assists greatly with solving ambiguities. For example, if you see a picture of a soccer game that includes a number of players and a ball in the air, from what direction did the ball come and who kicked it? Knowledge of the objects outside the scene and of the timing of the actions that led up to the scene can help with such a question.

The geometry of the camera system presents a different sort of difficulty, because it distorts the image. Extreme perspective effects are commonly seen if a wide-angle lens is used. The image is distorted on a finer scale by sources of noise in the imaging system. Such noise adds small variations to the true values of image intensity and color. For example, image sensors quantize light levels; therefore they do not represent those levels precisely.

Finally, it must be remembered that digital images represent large amounts of information. These bits of information must be compared with one another and combined with one another, and decisions must be made on all possible combinations in order to arrive at an interpretation of the image contents.

Although vision seems effortless for normally-sighted humans, it requires a large amount of processing power. Because of these difficulties, many have questioned whether a computer vision system could ever perform at the level of the human visual system.

Can Vision be Modeled Computationally?

An often-debated question is whether or not human or primate vision – and, more broadly, perception – can be modeled computationally. If not, then research on models of biological vision based on computational principles is sure to fail.

A proof of *decidability* is sufficient to guarantee that a problem can be modeled computationally. A decision problem has the form: given an element of a countably infinite set A, does that element belong to the set B (a subset of A)? Such a problem is decidable if there exists a Turing machine that computes 'yes' or 'no' for each element of A.

Perception, in general, has not yet been formulated as a decision problem, but many subproblems have been so formulated. Visual search, an important subproblem, has been formulated as a decision problem, and is decidable: it is an instance of the 'comparing' Turing machine.

Even if some other aspect of perception were determined to be undecidable, it would not follow that all of perception is undecidable, or that other aspects of perception cannot be modeled computationally. For example, one of the most famous undecidable problems is whether or not an arbitrary Diophantine equation has integral solutions (Hilbert's 10th problem). But this does not mean that mathematics cannot be modeled computationally. Similarly, another famous undecidable problem is the halting problem for Turing machines: it is undecidable whether a given Turing machine will halt for a given initial specification of its tape. This has important theoretical implications, but it certainly does not mean that computation cannot exist.

THE COMPONENTS OF A COMPUTER VISION SYSTEM

The processing stages depend strongly on the problem to be solved. For example, methods for processing color images would not be needed if only grayscale images are received. Similarly, if images from a microscope are analyzed, there is no depth inherent in the domain, so depth-specific processes can be omitted. The following discussion will assume no such domain-specific restrictions, and will describe the stages and components one might consider in the design of a general-purpose vision system. Note, however, that although the phrase 'general-purpose vision system' is often used, and may represent the ultimate goal of research in the area, general-purpose functionality has so far been elusive.

Levels of Analysis

It has been claimed that any information processing system must be understood at three different levels: the computational level, the representational and algorithmic level, and the hardware implementation level. The computational level is concerned with the goal of the computation and with the logic of the solution strategy. The representational and algorithmic level deals with the representations for the input, the output, and the intermediate computations, and with the algorithm that implements the solution strategy. The hardware implementation level concerns the physical realizations of the representations and algorithms – whether in a computer or in a brain. These levels of analysis have become part of the basic computational framework for the study of perception.

Acquiring Images

Cameras with arrays of sensing elements capture digital images. Each sensing element captures the color and brightness over a small region of the scene and translates it to a single value. (This value does not correspond directly to human perception of color and light.) For example, the quantity measured as brightness is irradiance, or power of light per unit area.

The result is an array of picture element values, or pixels, representing the scene. Note that the image is not an exact representation of the scene. The imaging geometry determines the size of the cone of light that is digitized by each pixel sensor; if that cone is large then detail from the scene is lost. There are also noise sources from within the electronic sensors.

The geometry of the imaging system relative to the scene is important if one wishes to recover the elements of the scene from a single image. A computer vision system must assume a projection model. Usually, this is the perspective model, but sometimes, for convenience, one assumes an orthographic projection model whereby the image is formed by rays from the object that are parallel to the optical axis of the camera. Perspective and orthographic projections do not differ much if the distance to the scene is much larger than the variation in distance among objects in the scene. Whatever model is chosen, if computations are made with the origin of a coordinate system on the camera, the frame of reference for the vision system is known as the 'camera frame'.

Often a single image is insufficient to recover the characteristics of a scene. Even an object as simple

as a cube can present difficulties: if, for example, you could see only one face of it, you could not know if it was a cube, a pyramid or some other kind of solid. Edges of the object align with one another, and this alignment obscures the object structure. Only if you rotated the cube, or moved your head, could you confirm the shapes and sizes of the other faces. Using computer-controlled camera systems, researchers have studied this seemingly natural and obvious tactic for deciding between competing recognition hypotheses. In 'active perception', a passive sensor is used in an active fashion, its state parameters being changed according to sensing strategies. In such systems, intelligent control is applied to the data acquisition process in a way that depends on the current state of data interpretation, including recognition.

Active vision is useful in order to see a portion of the visual field that is otherwise hidden, to increase spatial resolution (by moving in closer, or zooming), and to disambiguate aspects of the visual world (for example by analyzing induced motion or lighting changes). In all these uses, some hypothesize-and-test mechanism must be at work. Only if hypotheses are available can a particular action actually yield benefits; otherwise the number of possible interpretations and actions is too large. Furthermore, in practical implementations of active perception, the additional cost imposed on a perceptual system must be considered.

Preparing for Analysis

The image acquisition process leads to a scene representation that contains errors and noise. Usually one attempts to model noise in such a way that algorithms can account for it and compensate for its effects. If it can be assumed that the noise is additive and random, there are algorithms that suppress, smooth or filter it. Such algorithms are usually either applied before any further processing or integrated into other processing stages.

A second major issue is camera calibration. Without calibration, it is impossible to tell, for example, the length of a line in the scene from the length of its representation in the image. If the image acquisition system is calibrated, knowledge of the camera system parameters is related to known coordinates in the three-dimensional scene. As a result, measurements in images can be related to measurements in the scene. Usually, in order to calibrate a camera system, a known target in a known position and pose is captured and the projection equations are solved for the resulting image.

Extracting Features from an Image

Image features are meaningful structures of an image, that is, they are strongly related to elements of the scene being imaged. There are two types of features: global properties, such as the average gray level across the image, and local properties, such as lines, circles, or textures. Each feature corresponds to a particular geometric arrangement of image values. Features that are preserved under some transformation are known as invariants with respect to that transformation. The length of an object, for example, is invariant with respect to rigid movement. Invariants are especially valuable as indices for databases of object models. (*See* **Vision, Early**)

We will now look at some of the most important image features.

Edges

Edges are sets of pixels around which image values exhibit a sharp variation, i.e., discontinuity. Edge detectors are designed to detect such pixels. The first generation of edge detectors simply calculated a discrete approximation to the first derivative of image intensity with respect to position; peaks in this derivative are then candidates for edge points. However, noise can both mask true peaks in the derivative and give rise to false ones.

The second derivative of image intensity, for an ideal step edge, is zero at the point of the step. On one side of the step the value is negative while on the other it is positive. If the step is corrupted by noise, the positive and negative regions of the second derivative may be similarly corrupted, but the zero is much more stable. Edge detectors using this idea have somewhat better properties than the first generation of detectors. They usually smooth the image with a mathematical operation designed to remove noise and then compute an approximation to the second derivative. This is sometimes known as the 'difference of Gaussians' method because the overall process can be modeled by subtracting two Gaussian functions with appropriate parameters. The scale of intensity structure in the image must be carefully considered: for good detection, the scale of the edge detector must be matched to the scale of the image structure.

Regions

'Image segmentation' is the partitioning of an image into regions, or subsets of pixels, that satisfy some homogeneity criterion. The partitioning is constrained so that all pixels are accounted for, there is no overlap between regions, and no two

adjacent regions can be merged into a larger one that still satisfies the criterion. The criterion may relate, for example, to intensity, color, texture, or combinations of these.

How can such a segmentation be computed? Start with an arbitrary segmentation – say, into individual pixels. At each iteration, each region may or may not satisfy the criterion. If it does, then look at those of its neighbors that also satisfy the criterion and try to merge those regions into one. If the new region still satisfies the criterion, keep it; otherwise do not.

If a region does not satisfy the criterion, then split it into smaller regions. Continue the merging and splitting process until no further merges or splits are possible. (Note that regions consisting of single pixels must satisfy the homogeneity criterion, so the process terminates.)

This is known as the split-and-merge algorithm. Such regions are also known as 'connected components'. Segmentation remains one of the more stubborn problems in computer vision owing to the difficulty in defining comprehensive homogeneity criteria that correspond to objects.

Texture

The texture of a surface can be described as a statistical property. Texture elements ('texels') are visual primitives with certain invariant properties that occur repeatedly in different positions, deformations, and orientations inside a given area. The simplest approach to texture is to consider the first-order gray-level statistics of the image ('first-order' statistics are statistics of single pixels, as opposed to groups of pixels). These can be captured by the gray-level histogram (which specifies the frequency of occurrence in the image of each gray-level value). However, such histograms are insensitive to permutations of pixel positions, and thus have limited utility.

Second-order statistics describe pixel pairs in fixed geometric relationships. For example, how many pixels in an image have a given gray level while another pixel, a given distance away along a given direction, has another given value? Such statistics can be computed and used to classify textures such as those corresponding to images of wood, corn, grass, and water. Edge pairs – perhaps requiring certain angular relationships – can also be used to represent textures. (A cloth with a herringbone weave illustrates this.) Texture characteristics can be used as part of a segmentation algorithm to represent properties of regions.

Color

Color can be represented by associating red, green, and blue values with each pixel. Intensity is then the average of these three values. Other color spaces, such as the CIE system or the hue–saturation–intensity system, are also in common use. Each has advantages for different applications.

A major challenge in color processing is to model color constancy, the fact that humans see object colors consistently under widely varying illumination. 'Physics-based' methods attempt to calculate the spectral power distribution of the illumination and the objects. The physics of how light interacts with objects, and how colors are reflected from surfaces, is important in the design of color processing algorithms. The appearance of a surface depends strongly on the type of illumination, the direction of the illuminant, and other ambient light reflected off nearby objects.

Color is often also used as an aid in image segmentation. One technique uses knowledge of object color to classify pixels. For example, in a scene containing a basket of fruit, yellow may indicate banana, red may indicate apple, and so on. The yellow pixels can be grouped together, and if they are in the proper geometric configuration, the region may represent a banana. (*See* **Color Vision, Philosophical Issues about**; **Color Vision, Neural Basis of**; **Color Perception, Psychology of**)

Shape from shading

A smooth opaque object produces an image with spatially varying brightness even if the object is illuminated evenly and is made of material with uniform optical properties. Shading thus provides essential information about object shape. A smooth yellow billiard ball, illuminated with a single light source, has a surface that appears increasingly shaded as the angle of incidence of the light from the perpendicular increases. This effect can be mathematically modeled, and the resulting equation, describing the amount of light emitted from the surface, is called the image irradiance equation. Surface orientation has two degrees of freedom, while brightness – what is measured by image irradiance – has only one. Therefore, one needs to add a constraint in order to solve the image irradiance equation. This constraint may come from an overall smoothness metric on the desired solution, or from the use of multiple images (a technique known as 'photometric stereo').

Shape from texture

If an object's surface is textured, the variation in texture can be used to determine surface shape. Imagine a large white cloth, patterned with blue polka dots. Drape the cloth over a surface so that it is not flat. You can easily deduce the shape of the cloth from the shapes of the polka dots and their variation from true circles in the image seen by your eye. Shape-from-texture algorithms use this variation. Uniformly textured surfaces undergo two types of distortions due to imaging: as a surface recedes from the observer, increasingly large areas of the surface are isotropically compressed onto a given area of the image; and as a surface tilts away from the frontal plane, foreshortening causes anisotropic compression of texture elements. The rate of change of the projected size of texture elements – called the texture gradient – constrains the orientation of the plane.

A surface orientation can be represented by the slant – the angle between the normal to the surface and the projection direction – and the tilt – the angle between the normal's orthographic projection onto the image plane and the x-axis. If we know that a surface has blue dots as texture, then the variation in appearance of those dots in an image can be expressed as a slant and tilt and used to represent changes of local orientation of the surface.

Stereoscopic disparity

Normally sighted humans have two eyes. If you hold a finger in front of you and look at it first with the right and then with the left eye, it appears to shift position. The amount of displacement is the horizontal 'disparity'. Your visual system compares the two images and relates the disparity to the depth of the object (in the simplest case, by solving for the triangle formed by your eyes and the object). If you observe a scene with both eyes (a convergent imaging system), there is a single point on which both eyes fixate. At that point, there is zero horizontal and vertical disparity. For convergent systems, vertical disparity is zero only on the $x = 0$ and $y = 0$ planes. Horizontal disparity is zero only on the circle which contains the fixation point and the two eyes. The major problem is how to match a feature in one eye's image with the same feature in the other eye's image (the correspondence problem). For example, if you view a white picket fence from a close-up position, so that the fence covers the entire field of view, and if you look with each eye separately, assuming there is a uniform scene behind the fence (say, blue sky, or a green lawn), then you do not have enough information to know where a given picket in either image lies.

The geometry of the binocular imaging system can help. Given a point in the left image, the corresponding point in the right image is constrained to lie within a line in the right image, called the epipolar line. (This can be derived using the perspective projection transformation from one image to the other image.) In general, three important constraints help with this problem: similarity of corresponding features in the two views; the viewing geometry which constrains corresponding features to lie on epipolar lines; and piecewise continuity of surfaces in scenes. In the picket fence example, if you are close enough to the fence, matching the edges of the pickets will solve the problem if you can make assumptions about the width of the pickets and your viewing distance. (*See* **Depth Perception**)

Motion

Movies are natural representations of changes in object position, shape, and so on, and humans have no difficulty interpreting the apparent motion in an image sequence as equivalent to real motion, if the images are taken close enough in time. The simplest strategy for computing motion from a sequence of images is to consider the differences between successive images. The problem is to determine which pairs of points correspond to each other from one image to the next. This is another correspondence problem. For example, if the image sequence depicts a car moving along a road, it is easy to see that the points corresponding to the door handle in one image belong to those of the same door handle in the second. What about points on a smooth surface? What about the points of light in a fireworks display? Which ones are in correspondence, and what is the algorithm that determines the correspondence? In general, there are six translation and rotation parameters to recover for an object in motion. From two perspective images in time sequence, a computer system can recover only relative translation and not absolute translation of objects in motion at the scene. If the depth of objects in the scene is known, then absolute translation can be recovered.

Velocity can be measured only perpendicular to the contour of the moving object (that is, only one component of the velocity vector can be measured). This is known as the aperture problem. If one can assume that all edges in an image belong to the same rigid surface patch and that motion has constant velocity on the patch, then one can determine

the full velocity vector by solving for the correct velocity in the equations that minimize the error of the measured velocity relative to the correct velocity. The field of velocity vectors thus obtained is known as the optical flow associated with the apparent motion in an image sequence.

If an image sequence contains several moving objects as well as a background, and perhaps a moving camera system, the problem is more complicated. Suppose motion vectors could be defined along the contours and internal structures of each of the objects and of the background. This set of vectors will then contain 'groupings' for each motion: each object produces its own distribution of motion vectors. So in order to determine which groups of vectors correspond to single objects, a mixture of distributions made up of several separate motions can be proposed. The groups of vectors that best fit the distributions would describe single objects in motion. (*See* **Motion Perception, Neural Basis of**)

Representations

How a computer system represents images, extracted features, domain knowledge, and processing strategies is critical to its performance. A number of representational schemes have been proposed whose value has been proven in practice.

Image pyramids

In an image pyramid, an image is represented as a number of layers, each layer being an image. Successive images have smaller numbers of pixels representing the whole image. Each layer computes image properties (at successively coarser resolutions), and each computation communicates only with computations occurring in layers immediately above or below or with computations within the layer. This representation helps to reduce the amount of computation required: the coarser layers constrain the processing in the more detailed layers.

Intrinsic images

Intrinsic images form an intermediate level of representation, between images and object models. They consist of a number of separate feature maps that interact so that they can be computed unambiguously. The features they describe may include surface discontinuities, range, surface orientation, velocity, and color. These representations seem to be related to the processing stages in the primate cortex, where different kinds of computations seem to be grouped together in separate, but interacting, brain areas.

Image sketches and visual routines

Progressions of successively more abstract representations are often useful for coding complex visual information. One commonly-used sequence consists of the 'primal' sketch, the 'two-and-a-half-dimensional' sketch, and the 'three-dimensional' sketch. The primal sketch represents intensity changes and their organization in a two-dimensional image. The two-and-a-half-dimensional sketch represents the orientation and depth of surfaces, and discontinuity contours. Finally, the three-dimensional sketch represents shapes and their spatial organization.

Given the large number of feature types that might be computed on an image, and the representations that arise from them, how can a vision system integrate this information in a meaningful manner? The idea of visual routines addresses this question. These routines extract abstract shape properties and spatial relations from the early representations. Shape properties are characteristics of single items (e.g. length, orientation, area), and spatial relations are characteristics of two or more items (e.g. 'above', 'inside', 'longer than'). Shape properties and spatial relations are important for object recognition, visually guided manipulation, and abstract visualizations. Using a fixed set of basic operations, the visual system might assemble different visual routines to extract a variety of shape properties and spatial relations. Several specific operations have been proposed, including shifts of attentional focus, indexing to an 'odd man out' location, bounded activation, boundary tracing, and marking.

Object representations

Vision systems require the explicit representation of points, curves, surfaces, and volumes. There are a number of schemes that are employed, but there is no consensus yet on what constitutes an adequate set of primitives for spatial representations. Several popular methods exist for representing an object, including 'generalized cylinders', 'deformable models', 'geons', and 'aspect graphs'.

Strategies for Recognition

Several methods have proved useful for recognition of objects in images. The three most important are model-based, invariant-based, and appearance-based methods. The problem of recognition remains unsolved in general. At an abstract level,

recognition of objects in an image given a database of object types seems solvable in principle. One needs only to locate groups of features that might be objects, and then match each group to a model in the database. This is a hypothesize-and-test strategy: you first make a hypothesis, then test it; if it is false, you refine it and test it again, and so on until you are successful. If the image object is an instance of one of the models, then this procedure will find it.

This is a search task. However, a problem common to many search tasks arises: if it is not known in advance which parts of an image match with which parts of a model in a database, then all possible combinations of pixels must be checked, which is computationally infeasible.

Selective attention is an important mechanism for dealing with this problem. A visual attention mechanism can help with the selection of a region of interest in the visual field, the selection of feature dimensions and values of interest, the shifting from one selected region to the next, the integration of successive attentional fixations, interactions with memory, indexing into model bases, etc. (*See* **Vision, High-level**; **Object Perception, Neural Basis of**; **Vision: Object Recognition**; **Audition, Neural Basis of**; **Spatial Attention, Neural Basis of**; **Selective Attention**; **Vision: Top-down Effects**)

Model-based recognition

It is common for systems to employ databases of models to help with recognition. Important subtasks in recognition include deciding which models in the database have instances in the image, where those instances are, and their orientations.

Recognition can be viewed as a process that proceeds from the general to the specific and that overlaps with, guides, and constrains the derivation of a description from the image. For example, a database of models can be constructed using volumetric primitives and organized using a specialization hierarchy as well as a decomposition hierarchy. Models can be selected according to the characteristics of the extracted volumetric primitives.

Interpretation trees represent methods that use features of objects for recognition. Each model in the database is described by its component features, and the description must include the constraints between features. The interpretation tree enumerates all possible ways in which a given set of features found in an image can be matched with features of particular objects in the model base. Methods for searching this tree focus on limiting search so that not all of the tree need be examined (the trees can grow exponentially). An assignment can be verified by back-projecting the resulting model into the image (using the positions of features, and modeling the image formation process to see whether the object actually looks like the one in the image).

Much work has been done on understanding scenes with polyhedral objects. Edges are found and connected, and the resulting sequences labeled as concave or convex with respect to the polyhedron; shadows, occluding boundaries, and so on are also hypothesized. Labeled edges can then be grouped into hypothesized whole objects, and these hypotheses can be confirmed within an interpretation tree.

Invariant-based recognition

Invariants are properties of geometric configurations that do not change under certain transformations. For example, the length of an edge of a solid, rigid object is an invariant property of that object. Suppose several such invariants for a set of objects are defined. The database of objects can then have a feature vector associated with each model, where the features of the vector are all invariants. These features can be used as indices into the database. For example, if objects have planar sides and lines are straight or simple conics, then sets of contours may participate as invariant features under projective transformation.

Appearance-based recognition

Images themselves can also be considered as features. In order to make such features usable, several images of each object, taken from different viewpoints (and illumination conditions, if relevant), are needed. Object models in a database are defined by such a set of images. Face recognition is a particularly good application of this method.

Suppose you have a set of pictures, say, passport photographs. The size of the face in each is about the same; there is only a single face in each; and the lighting is approximately the same in each. Suppose now that you compute the average image: add up the pixel values at each location from all the pictures and divide by the number of pictures. Each picture can then be redefined as the sum of this average picture and a representation of how the image differs from the average in terms of a relatively small set of 'principal components' of the image set. This provides an economical way of representing the images. For any new picture, you compute this representation. Representations can be compared: if the new image is sufficiently close

to one of the stored images, then they are taken to represent the same person.

Choosing a recognition strategy

In each of the above methods, the computer system must include representations that formalize the knowledge of a domain. Performance depends critically on the representation, the control strategy that uses this knowledge, and the quality of the knowledge. Only if the knowledge of how to solve the task can be extracted, codified in a suitable formalism, and used in drawing conclusions during processing, will the method work. The choice of recognition strategy thus depends on the kinds and qualities of domain knowledge available. If only very general knowledge of the domain is possible then an invaraint strategy might be the best choice. If explicit models are known, then model-based methods are preferable. If control over imaging conditions and scene positioning is possible, then appearance-based methods may be appropriate. The methods may also be combined.

The role of knowledge and its application to guide processing is critical. It can be shown that basic problems in vision such as matching are potentially intractable if no knowledge is used to guide processing: that is, there is no guarantee that those problems can be solved using any existing or future computational resources. The intractability is due solely to the difficulty of selecting which parts of the input image are to be processed: without knowledge, the number of such image subsets increases exponentially.

Task guidance can be implicit (as in positioning a person so that the face is imaged in a particular way) or explicit (as in annotating interesting portions of a scene by hand before computer processing). Attentional selection, using knowledge to optimize processing, may determine which image parts to attempt to process first; if the first few selections are good ones, a great deal of searching can be avoided.

CONCLUSION

Although computer vision has been an active discipline since the early 1960s, progress has been slow. The 1990s saw some promising developments, in part due to the advent of sufficiently powerful computers and the resulting ability to search through more possibilities and to test more complex theories. Computer vision is now widely used in commercial applications. Still, progress on finding theories that address the capabilities of human vision remains slow. This represents the greatest intellectual challenge in the field.

Further Reading

Faugeras O (1993) *Three-Dimensional Computer Vision*. Cambridge, MA: MIT Press.

Fischler MA and Firschein O (1987) *Readings in Computer Vision: Issues, Problems, Principles and Paradigms*. Los Altos, CA: Morgan Kaufmann.

Marr D (1982) *Vision*. San Francisco, CA: WH Freeman.

Tsotsos JK (1990) Analyzing vision at the complexity level. *Behavioral and Brain Sciences* **13**: 423–445.

Tsotsos JK (1992) Image understanding. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 641–663. New York, NY: John Wiley.

Zucker SW (1992) Early vision. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 394–420. New York, NY: John Wiley.

Concept Learning and Categorization: Models

Intermediate article

John K Kruschke, Indiana University, Bloomington, Indiana, USA

CONTENTS

Category learning involves generalizing from one learned case to another in appropriate ways. Models of category learning have been based on various representations, including exemplars, prototypes, rules, and hybrids thereof.

CATEGORIZATION IN COGNITION

Categories pervade our cognition. We classify variously shaped printed squiggles into different letter categories. We classify a spectrum of acoustic signals into phonemic categories. We categorize people, animals, plants, and artefacts, and we base our actions on how we categorize.

A central function of learned categories is generalizing from a particular learned instance to novel situations. If learned knowledge consisted merely of isolated facts with no generalization, then the knowledge would be inapplicable except for the unlikely exact recurrence of the learned situation. For example, learning that a four-legged, striped, 1.0-meter-tall animal is a tiger would not generalize to inferring that a four-legged, striped, 1.1-meter-tall animal is also a tiger. The consequence of this failure to generalize a category could be an eaten learner. At the opposite extreme, if learned knowledge were to generalize too broadly, then complementary errors could be committed: learning that a four-legged, striped, 1.0-meter-tall animal is a tiger would lead to inferring that a zebra is also a tiger. The consequence of this over-generalization could be a starved learner.

An equally crucial goal of learning categories is retaining previously learned knowledge while quickly acquiring new knowledge. For example, after having learned about zebras, it could prove disastrous if learning about tigers required dozens of exposures. It could also be disastrous if the learning about tigers erased valid knowledge about zebras.

Category learning is critically important because it underlies essentially all cognitive activities; yet it is very difficult because: (1) learned categories must generalize appropriately; (2) learning must occur quickly; and (3) new learning must not overwrite previous knowledge. Understanding how learners accomplish these feats is the topic of this article.

Any theory of category learning must specify: (1) what information from the world is actually retained in the mind; (2) how that information is used and learned; and (3) why that particular learning algorithm is useful. These three issues are addressed in turn, for different theories. Each type of theory is initially described informally, to convey the basic motivating principles of the theory. It is then described in formal, mathematical terms. By being expressed mathematically, the theory gains: quantitative precision rather than vague verbal description; publicly derivable predictions rather than theorist-dependent intuitively derived predictions; stronger support when predictions are confirmed in quantitative detail; greater explanatory power when the formal mechanisms in the model have clear psychological interpretations; and greater applicability because of precise specification of relevant factors.

EXEMPLAR THEORIES

Perhaps the simplest way to learn is just to memorize the experienced instances. For example, a learner's knowledge of the category *dog* might consist of knowing that the particular cases named Lassie, Rin-Tin-Tin, Old Yeller and Pongo are exemplars of dogs. There is no derived representation of a prototypical dog, nor is there any abstracted set

of necessary and sufficient features that define what a dog is. As new cases of dogs are experienced, these cases are also stored in memory. Notice, however, that just because these exemplars of dogs are in memory, the learner need not be able to distinctly recall every dog ever encountered. Retrieving a memory might be quite different from using it for categorization.

According to these exemplar theories of categorization, a new stimulus is classified according to how similar it is to all the known instances of the various candidate categories. For example, a newly encountered animal is classified as a dog if it is more similar to known exemplars of dogs than it is to known exemplars of cats or horses, etc. The notion of similarity, therefore, plays a critical role in exemplar theories.

Selective attention also plays an important role in exemplar theories. Not all features are equally relevant for all category distinctions. For example, in deciding whether a novel animal is a dog or a cat, it might be more important to pay attention to size than to number of legs, because dogs and cats tend be of different sizes, but have the same number of legs.

In principle, exemplar encoding can accurately learn any possible category structure, no matter how complicated, because the exemplars in memory directly correspond with the instances in the world. This computational power of exemplar models is one rationale for their use. On the other hand, the uniform application of exemplar encoding can make learning slow, if highly similar instances belong to different categories. One way around this problem is to associate exemplars with categories only to the extent that doing so will improve accuracy of categorization. Analogously, features or stimulus dimensions may be attended to only to the extent that doing so will reduce error. Error reduction is one rationale for theories of learning.

Formal Models of Exemplar Theories

In a prominent exemplar-based model (Kruschke, 1992; Medin and Schaffer, 1978; Nosofsky, 1986), a stimulus is represented by its values on various psychological dimensions. For example, a tiger might be represented by a large numerical value on the dimension of size, and by another large numerical value on the dimension of ferocity, along with other values on other dimensions. The psychological value on the d^{th} dimension is denoted $\psi_d{}^{\text{stim}}$. For the m^{th} exemplar in memory, the psychological value on the d^{th} dimension is denoted $\psi_{md}{}^{\text{ex}}$. These psychological scale values can be determined by methods of multidimensional scaling (e.g., Kruskal and Wish, 1978).

The similarity of the stimulus to a memory exemplar gets larger as the distance between the stimulus and the exemplar in psychological space gets smaller. For psychological dimensions that can be selectively attended to, the usual measure of distance between the stimulus, s, and the m^{th} memory exemplar is given by $\text{dist}(s,m) = \sum_i \alpha_i |\psi_i{}^{\text{stim}} - \psi_{mi}{}^{\text{ex}}|$, where the sum is taken over the dimensions indexed by i, and $\alpha_i \geq 0$ is the attention allocated to the i^{th} dimension. When attention on a dimension is large, then differences on that dimension have a large effect on the distance, but when attention on a dimension is zero, then differences on that dimension have no effect on the distance. The distance is then converted to similarity by an exponentially decaying function: $\text{sim}(s,m) = \exp(-\text{dist}(s,m))$. Therefore, when the stimulus exactly matches the memory exemplar, the similarity is 1.0, and as the distance between the stimulus and the memory exemplar increases, the similarity decreases towards zero. (Shepard (1987) provides a review of the properties of the exponential similarity function.)

Each exemplar then 'votes' for the categories. The strength of an exemplar's vote is its similarity to the stimulus, and the exemplar's selection of categories is a continuous weighting given by its associative strengths to the categories. The associative strength from exemplar m to category k is denoted w_{km}, and the total 'voting' for category k is $v_k = \sum_m w_{km} \text{sim}(s,m)$. The overall probability of classifying the stimulus into category k is the total vote for category k relative to the total of votes cast. Formally, the probability of classifying stimulus s into category k is given by $p_k = v_k / \sum_c v_c$.

In laboratory experiments on category learning, after the learner makes his or her guess as to the correct categorization of a given stimulus, he or she is given corrective feedback, and then tries to learn this correct answer. The same procedure applies to learning in the model. The model adjusts its associative weights and attention strengths to reduce the error between its vote and the correct answer. Error is defined as $E = \sum_k (t_k - v_k)^2$, where t_k is the 'teacher' value: $t_k = 1$ if k is the correct category, and $t_k = 0$ otherwise. There are many possible methods by which the associative weights and attention strengths could be adjusted to reduce this error, but one sensible method is 'gradient descent' on error. According to this procedure, the changes that make the error decrease most rapidly are computed according to the derivative of the error with

respect to the associative weights and attention strengths. The resulting formula for weight changes is $\Delta w_{km} = \lambda(t_k - v_k)\ \mathrm{sim}(s, m)$, where λ is a constant of proportionality called the learning rate. This formula states that the associative weight between exemplar m and category k increases to the extent that the exemplar is similar to the current input and the category teacher is under-predicted. Notice that after the weight changes according to this formula, the predicted category will be closer to the correct category; i.e., the error will have been reduced. The formula for attentional changes is slightly more complicated, but essentially it combines information from all the exemplars to decide whether attention on a dimension should be increased or decreased (Kruschke, 1992; Kruschke and Johansen, 1999).

Variants of this exemplar model have been shown to fit a wide range of phenomena in category learning and generalization (e.g., Choi *et al*., 1993; Estes, 1994; Kruschke and Johansen, 1999; Lamberts, 1998; Nosofsky and Kruschke, 1992; Nosofsky, Gluck *et al*., 1994; Nosofsky and Palmeri, 1997; Palmeri, 1999).

PROTOTYPE THEORIES

Instead of remembering every exemplar of a category, the learner might construct a representation of what is typical of the category. For example, the mental representation of *dog* might be an average of all the experienced instances. The dog prototype need not necessarily correspond to any actually experienced individual dog. Alternatively, the representative summary could be an idealized caricature or extreme case that is maximally distinct from other categories, rather than the central tendency of the category.

According to prototype theories of categorization, a new stimulus is classified according to how similar it is to the prototypes of the various candidate categories. A newly encountered animal is classified as a dog if it is more similar to the dog prototype than it is to other category prototypes.

One rationale for this approach to categorization is that it is efficient: the entire set of members in a category is represented by just the small amount of information in the prototype.

Formal Models of Prototype Theories

Prototypes can be formally described in a similar way to exemplars. The prototype for category k has psychological value on dimension i denoted by $\psi_{ki}^{\mathrm{proto}}$, and this value represents the central tendency of the category instances on that dimension. The model classifies a stimulus as category k in a manner directly analogous to the exemplar model, so that the probability of classifying stimulus s as category k is given by $p_k = \mathrm{sim}(s, k) / \sum_m \mathrm{sim}(s, m)$. The sum in the denominator is over all category prototypes, instead of over all exemplars.

In one kind of prototype model, each prototype must be tuned to represent the central tendency of the instances in its category. For the first experienced instance of a category, the prototype is created and set to match that instance. For subsequently experienced instances of the category, the prototype changes from its current values slightly towards those of the new case. By moving towards the instances of the category as they are experienced, the prototype gradually progresses towards the central tendency of the instances.

One way of formalizing the learning of central tendencies is the following algorithm, closely related to so-called 'competitive learning' or 'clustering' methods. The idea is that a prototype should be adjusted so that it is as similar as possible to as many instances as possible; in this way the prototype is maximally representative of the stimuli in its category. Define the total similarity of the prototypes to the instances as $S = \sum_{k,s} \mathrm{sim}(s, k)$, where $\mathrm{sim(s,k)} = \exp(-\sum_{\mathrm{i}} \alpha_{\mathrm{i}} [\psi_{\mathrm{i}}^{\mathrm{stim}} - \psi_{\mathrm{ki}}^{\mathrm{proto}}]^2)$. (This summation across all instances does not require that all the instances be stored in memory, nor that the instances be simultaneously available for learning.) The question then is how best to adjust $\psi_{ki}^{\mathrm{proto}}$ so that the total similarity increases. One way to do this is gradient ascent: the prototype values are adjusted to increase the total similarity as quickly as possible. The resulting formula, determined as the derivative of the total similarity with respect to the coordinates, yields $\Delta\psi_{ki}^{\mathrm{proto}} = \lambda\ \mathrm{sim}(s, k)\alpha_i(\psi_i^{\mathrm{stim}} - \psi_{ki}^{\mathrm{proto}})$. This formula causes each prototype's values to move towards the currently experienced stimulus, but only to the extent that the prototype is already similar to the stimulus, and only to the extent that the dimension is being attended to. In this way, prototypes that do not represent the stimulus very well are not much influenced by the stimulus.

Some models allow multiple prototypes per category, to capture multimodal distributions (Anderson, 1991), and use other learning methods derived from Bayesian statistics. In the extreme case, there can be one prototype per exemplar, and such models become equivalent to exemplar models (Nosofsky, 1991). In exemplar models,

however, the coordinates of the exemplars typically do not get adjusted from one trial to the next.

In several studies that compare prototype and exemplar models, it has been found that prototype models do not fit data better than exemplar models (e.g. Ashby and Maddox, 1993; Busemeyer *et al*., 1984; Busemeyer and Myung, 1988; Nosofsky, 1992; but cf. Reed, 1972), but some have found evidence for prototypes either early or late in learning (Homa *et al*., 1993; Smith and Minda, 1998). Prototype theory is, however, intuitively appealing, and a challenge for cognitive scientists is to discover phenomena that are naturally addressed by prototype models but that cannot be adequately accounted for by exemplar-based models, or by rule-based models, which are described next.

RULE THEORIES

Yet another way of representing categories is with rules that specify strict necessary and sufficient conditions for category membership. For example, something is a member of the category 'bachelor' if it human, male, unmarried and eligible. Many natural categories are very difficult to specify in terms of rules, however (e.g. Rosch and Mervis, 1975). For example, the category 'game' has no necessary and sufficient features (Wittgenstein, 1953). Nevertheless, people are prone to look for features that define category distinctions, and people tend to believe that such defining features exist even if in fact they do not (Brooks, 1978; Brooks *et al*., 1998).

Rules for category definition are typically a single threshold on a single dimension: for example, a building is a skyscraper if and only if it is taller than 10 floors. Rules can also be logical combinations of such thresholds: for example, a building is a skyscraper if and only if it is taller than 10 floors and its facade is at least 60% glass. In some rule-based theories, rules can be more complicated boundaries: for example, a building is a skyscraper if and only if the number of floors multiplied by the percentage of glass in the facade exceeds the value 6.0. (By this multiplicative rule, a building only seven floors tall would be classified as a skyscraper if it had at least 86% glass in its facade, because $7 \times 86\% > 6.0$.)

In principle, categorization rules are absolute, and there is no 'gray area' around the boundary of the category. In practice, however, most rule-based models do incorporate some mechanism for blurring the category boundary, to accommodate real performance data.

Formal Models of Rule Theories

Traditionally, rule models have been referred to as 'hypothesis testing' or 'concept learning' models (for a review, see Levine, 1975). In these sorts of models, individual features are tested, one at a time, for their ability to account for the correct classifications of the stimuli. For example, the model might test the rule 'if it's red then it's in category K'. As long as the rule works, it is retained, but when an error is encountered, another rule is tested. As simple rules are excluded, more complicated rules are tried. A recent incarnation of this type of model is also able to learn exceptions to rules, by testing additional features of instances that violate an otherwise successful rule (Nosofsky, Palmeri and Mckinley, 1994). This model is also able to account for differences in behavior between people, because there can be different sets of rules and exceptions that equally well account for the classifications of the stimuli.

For stimuli that vary on continuous dimensions, there is a well-studied class of models for which the decision boundary is assumed to have a shape that can be described by a quadratic function, because a quadratic describes the optimal boundary between two multivariate normal distributions, and natural categories are sometimes assumed to be distributed normally (e.g. Ashby, 1992). In this approach, there are three basic postulates: (1) the stimulus is represented as a point in multidimensional space, but the exact location of this point is variable because of perceptual noise; (2) a stimulus is classified according to which side of a quadratic decision boundary it falls on; (3) the decision boundary is also subject to variability because of noise in the decision process. Thus, although the classification rule is strict and there is no explicit role in the model for similarity gradients, the model as a whole produces a gradation of classification performance across the boundary because of noise in perception and decision. There are many variations on this scheme of models, involving different shapes of boundaries, deterministic or probabilistic decision rules, and so on. (Ashby and Alfonso-Reese, 1995; Ashby and Maddox, 1993).

HYBRID REPRESENTATION THEORIES

It is unlikely that any one of these types of representation can completely explain the complexity of human category learning. A variety of work has shown that neither rule-based nor prototype models can fully account for human categorization (e.g. Ashby and Waldron, 1999; Kalish and

Kruschke, 1997). In particular, exemplar representation must be supplemented with rules to account for human learning and generalization (Erickson and Kruschke, 1998). Therefore, some recent theories combine different representations. A model constructed by Vandierendonck (1995) combines rectangular decision boundaries with exponentially decaying similarity gradients. A model constructed by Ashby *et al*. (1998) combines linear decision boundaries that involve single dimensions (corresponding to verbalizable rules) with linear decision boundaries that combine two or more dimensions (corresponding to implicitly learned rules). A model constructed by Erickson and Kruschke (1998) combines exemplars with single-dimension rules.

The challenge for hybrid representation theories is specifying the interaction of the various types of representation. If there are several representational types available, under which conditions is each type used? In the model of Erickson and Kruschke (1998), for example, which combines rules with examples, the representations compete for attention, so that the type of representation that reduces categorization error most quickly is the type that is used for that instance. Hybrid models will probably proliferate in the future.

ROLE OF SIMILARITY

Similarity is critical in exemplar and prototype theories, and also appears in hybrid rule-based theories as distance from the boundary (e.g. Vandierendonck, 1995). Some researchers have criticized the notion of similarity as being internally incoherent, or have argued that similarity does not always correlate with categorization.

Similarity can be empirically investigated in several different ways. One method is simply to ask people to rate the similarity of two items; another is to measure discriminability between items. Usually these different assessments agree, but sometimes they do not (e.g. Tversky, 1977). Similarity can be context-specific: in the context of hair, gray is more similar to white than to black, but in the context of clouds, gray is more similar to black than to white (Medin and Shoben, 1988). In general, models of similarity presume which features or dimensions are used for comparing the objects, without any explanation of why those features or dimensions are selected. Models of similarity do have parameters for specifying the attention allocated to different features, but the models do not describe how these attentional values arise (Goodman, 1972; Murphy and Medin, 1985).

Similarity is not always a clear predictor of categorization. Consider the category *things to remove from a burning house*. The items 'heirloom jewellery' and 'children' are both central members of this category, yet they have little surface similarity (Barsalou, 1983). On the other hand, if attention is directed only to the features *irreplaceable* and *portable*, then children and heirloom jewellery bear a strong similarity. Once again the question of what to attend to is crucial, but not addressed by current theories of similarity.

Despite these complexities, there are strong regularities in similarity and categorization data that should yield to formal treatment. Excellent reviews of these topics have been written by Goldstone (1994) and by Medin *et al*. (1993).

SUMMARY

Categorization is central to cognition. Different theories of category learning posit different representations for the information underlying categorization. Research has shown that no single type of representation can account for the full range of categorization observed in humans. Instead, recent models combine different types of representations in hybrid systems. Challenges for future research include determining how different representations interact, and how attention influences and is influenced by category learning.

References

Anderson JR (1991) The adaptive nature of human categorization. *Psychological Review* **98**: 409–429.

Ashby FG (1992) Multidimensional models of categorization. In: Ashby FG (ed.) *Multidimensional Models of Perception and Cognition*, pp. 449–483. Hillsdale, NJ: Erlbaum.

Ashby FG and Alfonso-Reese L (1995) Categorization as probability density estimation. *Journal of Mathematical Psychology* **39**: 216–233.

Ashby FG, Alfonso-Reese LA, Turken AU and Waldron EM (1998) A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**: 442–481.

Ashby FG and Maddox WT (1993) Relations between prototype, exemplar and decision bound models of categorization. *Journal of Mathematical Psychology* **37**: 372–400.

Ashby FG and Waldron EM (1999) On the nature of implicit categorization. *Psychonomic Bulletin and Review* **6**: 363–378.

Barsalou L (1983) Ad hoc categories. *Memory and Cognition* **11**: 211–227.

Brooks LR (1978) Nonanalytic concept formation and memory for instances. In: Rosch E and Lloyd BB (eds)

Cognition and Categorization, pp. 169–211. Hillsdale, NJ: Erlbaum.
Brooks LR, Squire-Graydon R and Wood TJ (1998) *The role of inattention in everyday concept learning: identification in the service of use*. [Available from L. R. Brooks, Department of Psychology, McMaster University, Hamilton, Ontario, Canada L8S 4K1.]
Busemeyer JR, Dewey GI and Medin DL (1984) Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory and Cognition* **10**: 638–648.
Busemeyer JR and Myung IJ (1988) A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* **14**: 3–11.
Choi S, McDaniel MA and Busemeyer JR (1993) Incorporating prior biases in network models of conceptual rule learning. *Memory and Cognition* **21**: 413–423.
Erickson MA and Kruschke JK (1998) Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127**: 107–140.
Estes WK (1994) *Classification and Cognition*. New York, NY: Oxford University Press.
Goldstone RL (1994) The role of similarity in categorization: providing a groundwork. *Cognition* **52**: 125–157.
Goodman N (1972) Seven strictures on similarity. In: Goodman N (ed.) *Problems and Projects*, pp. 437–447. New York, NY: Bobbs-Merrill.
Homa D, Goldhardt B, Burruel-Homa L and Smith JC (1993) Influence of manipulated category knowledge on prototype classification and recognition. *Memory and Cognition* **21**: 529–538.
Kalish ML and Kruschke JK (1997) Decision boundaries in one dimensional categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**: 1362–1377.
Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review* **99**: 22–44.
Kruschke JK and Johansen MK (1999) A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* **25**: 1083–1119.
Kruskal JB and Wish M (1978) *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.
Lamberts K (1998) The time course of categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 695–711.
Levine M (1975) *A Cognitive Theory of Learning: Research on Hypothesis Testing*. Hillsdale, NJ: Erlbaum.
Medin DL, Goldstone RL and Gentner D (1993) Respects for similarity. *Psychological Review* **100**: 254–278.
Medin DL and Schaffer MM (1978) Context theory of classification learning. *Psychological Review* **85**: 207–238.
Medin DL and Shoben EJ (1988) Context and structure in conceptual combination. *Cognitive Psychology* **20**: 158–190.
Murphy GL and Medin DL (1985) The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316.
Nosofsky RM (1986) Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**: 39–57.
Nosofsky RM (1991) Relation between the rational model and the context model of categorization. *Psychological Science* **2**: 416–421.
Nosofsky RM (1992) Exemplars, prototypes, and similarity rules. In: Healy AF, Kosslyn SM and Shiffrin RM (eds) *Essays in Honor of William K. Estes*, vol. II 'From Learning Processes to Cognitive Processes', pp. 149–167. Hillsdale, NJ: Erlbaum.
Nosofsky RM, Gluck MA, Palmeri TJ, McKinley SC and Glauthier P (1994) Comparing models of rule-based classification learning: a replication of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition* **22**: 352–369.
Nosofsky RM and Kruschke JK (1992) Investigations of an exemplar-based connectionist model of category learning. In: Medin DL (ed.) *The Psychology of Learning and Motivation*, vol. XXVIII, pp. 207–250. San Diego, CA: Academic Press.
Nosofsky RM and Palmeri TJ (1997) An exemplar-based random walk model of speeded classification. *Psychological Review* **104**: 266–300.
Nosofsky RM, Palmeri TJ and McKinley SC (1994) Rule-plus-exception model of classification learning. *Psychological Review* **101**: 53–79.
Palmeri TJ (1999) Learning categories at different hierarchical levels: a comparison of category learning models. *Psychonomic Bulletin and Review* **6**: 495–503.
Reed SK (1972) Pattern recognition and categorization. *Cognitive Psychology* **3**: 382–407.
Rosch EH and Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.
Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* **237**: 1317–1323.
Smith JD and Minda JP (1998) Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 1411–1436.
Tversky A (1977) Features of similarity. *Psychological Review* **84**: 327–352.
Vandierendonck A (1995) A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin and Review* **2**: 442–459.
Wittgenstein L (1953) *Philosophical Investigations*. New York, NY: Macmillan.

Further Reading

Estes WK (1994) *Classification and Cognition*. New York, NY: Oxford University Press. [A mathematically oriented survey.]
Lamberts K and Shanks D (eds) (1977) *Knowledge, Concepts and Categories*. Cambridge, MA: MIT Press. [An accessible collection of tutorials.]

Rosch E and Lloyd BB (eds) (1978) *Cognition and Categorization*. Hillsdale, NJ: Erlbaum. [A collection of statements of fundamental results and theoretical perspectives.]

Shanks DR (1995) *The Psychology of Associative Learning*. Cambridge, UK: Cambridge University Press. [A lucid review of issues in category learning.]

Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: Harvard University Press. [A very readable introduction to the field of categorization.]

Connectionism

Intermediate article

Jerome A Feldman, University of California, Berkeley, California, USA
Lokendra Shastri, University of California, Berkeley, California, USA

CONTENTS

Science must eventually explain how the properties of the brain determine cognitive behavior. Connectionist approaches to cognitive science are based on the belief that our present understanding of neural and brain physiology should guide cognitive research.

INTRODUCTION

Since the detailed anatomy of higher cognition is not known, most connectionist research involves mathematical and computational modeling of behavior. The same modeling techniques also have a wide range of biological, engineering and management applications, and the discipline called 'neural networks' or 'connectionism' now encompasses a wide range of efforts that have only a vague link to the brain or cognition. This article will describe the kinds of connectionist model used in cognitive science and their implications for how we think about thinking.

Connectionist computational models are almost always computer programs, but programs of a different kind from those used in, for example, word processing, or symbolic artificial intelligence (AI). Connectionist models are specified as networks of simple computing units, which are abstract models of neurons. Typically, a model unit calculates the weighted sum of its inputs from upstream units and sends to its downstream neighbors an output signal that is a nonlinear function of its inputs. Learning in such systems is modeled by experience-based changes in the weights of the connections between units.

EARLY DEVELOPMENTS

There is a distinguished prehistory to current connectionist modeling techniques. Many of the original papers are gathered in (Anderson and Rosenfeld, 1988).

Sigmund Freud and William James, arguably the greatest psychologists of the nineteenth century, both discussed explicit neural models of cognition. For much of the first half of the twentieth century, the Anglo-American study of cognition was dominated by the behaviorist paradigm, which rejected any investigation of internal mechanisms of mind. In the 1940s, there arose two new connectionist paradigms, which still shape the field. McCulloch and Pitts (McCulloch, 1988) emphasized the study of particular computational structures comprised of abstract neural elements, while Hebb (1949) focused on the properties of assemblies of cells. The study of bulk properties of abstract neural systems has continued through Hopfield networks (Hopfield, 1982) and the Boltzman machine (Ackley *et al*., 1985), but has had little impact in cognitive science. One lasting result of Hebb's work was the Hebbian model of learning, which will be discussed in the next section.

The notion of a computational model is now commonplace in all scientific fields and in many other areas of contemporary life. One builds a detailed software model of some phenomenon and studies the behavior of the model, hoping to gain understanding of the original system. While all fields use computational models, connectionist researchers also invent and study computational techniques for constructing models, presenting the results of simulations, and understanding the limitations of the simulation. From the time of the first electronic computers, people dreamed of making them 'intelligent' by two quite distinct means. The first is to build standard computer programs as models of intelligence. This remains the dominant paradigm in AI and has had considerable success. The second approach was to try to build hardware that was as brain-like as possible

and have it learn the required behavior, this was the origin of connectionist modeling.

LEARNING CONNECTION WEIGHTS

Connectionist models consist of simple computing units (or nodes) connected via weighted links. A node communicates with the rest of the network by transmitting its output to all the nodes immediately downstream of it. This output is a monotonic (and often nonlinear) function of the node's total input. The contribution of each input link to the receiving node's total input is the output of the node at its source multiplied by the weight of the link. The most commonly used input–output transfer functions compute the total input by summing the inputs contributed by all the weighted links and then passing this total input through a nonlinear function. An example of such a nonlinear function is the step function, whereby the node produces an output of 1 (i.e. fires) if and only if the total input exceeds a threshold. Another commonly used function is the sigmoid function, whereby the output of a node varies smoothly between -1 and $+1$ as a function of the total input (the shape of this response curve resembles a stretched S).

The behavior of a connectionist network is completely determined by the input–output transfer function, the pattern of interconnection among the nodes, and the link weights. Typically, the input–output transfer function is assumed to remain fixed for a given model. It is possible to 'sculpt' the interconnection pattern by starting with a completely connected network and then pruning it by setting the weights of certain links to zero. Thus, changes in the functionality of a connectionist network (learning) can be effected by changing only the link weights; and most mechanisms and theories of learning in connectionist networks focus on changes in link weights.

One of the most influential proposals about learning in neural networks was made by Hebb (1949) and is known as Hebb's rule. In its simplest form, this rule states that the weight of the link from node i to node j is strengthened if i repeatedly participates in the firing of j. This rule and its variants lie at the core of many learning rules investigated by neural network modelers and theorists. In its simplest form, the rule suffers from several technical difficulties. These include lack of specificity, and saturation (all link weights would increase until they eventually reached their maximum possible values). To tackle this problem, Hebb's rule has been supplemented with an anti-Hebbian learning rule which says that the weight from node i to node j decreases if node j fires but node i does not; or alternatively, if node i fires but node j does not. The 'BCM' rule (Bienenstock, Cooper and Munro, 1982) states that the weight vector of a node is tilted in the direction of the input vector if the output exceeds a threshold, or in the direction opposite to that of the input vector if the output is below that threshold. The threshold can be variable, and may depend on the time-averaged output of the node. Note that the weight vector of a node with n input links is simply the n-tuple consisting of the n link weights.

Another framework for modifying link weights is competitive learning (von der Malsburg, 1973; Grossberg, 1976; Rumelhart and Zipser, 1985). Informally, the competitive learning algorithm may be described as follows. Let each node compute its output to be the weighted sum of its inputs, and assume that weight and input vectors are normalized to one. Then the output of a node is given by $\cos \alpha$, where α is the angle between the weight vector and the input vector. Weight modification in response to external inputs proceeds as follows: (1) An input pattern is presented to the network; (2) the node with the highest response is identified (this can be achieved by a 'winner takes all' configuration (Feldman and Ballard, 1982)); (3) the weight vector of this winning node is rotated towards the input vector; (4) the weight vector is renormalized. These steps are repeated for each input pattern. As a result of this weight modification regime, the nodes in the network modify their link weights so as to respond maximally to frequently occurring input patterns. In effect, each cluster (or category) in the input space 'recruits' one or more nodes in the network that produce a strong response to input patterns from the associated cluster.

The competitive learning algorithm can be augmented by requiring that (1) nodes in the network are arranged in a low-dimensional lattice (for example, in a linear sequence, or in a two-dimensional grid) and (2) changes in the weight vector of a node are accompanied by similar changes in the weight vectors of neighboring nodes (for example, via lateral connections between neighboring nodes). Such an augmented neural network learning algorithm is called a self-organizing map (Kohonen, 1982). Self-organizing maps have been applied to a variety of problems in pattern recognition. They have the important capability of developing topology-preserving maps from a high-dimensional space to a low-dimensional space. Examples of such maps abound in the brain: for example, several visual features

(e.g., location in the visual field, orientation, direction of motion) are mapped to a layered two-dimensional cortex.

The 'adaptive resonance theory' architecture (Grossberg, 1980) also extends the competitive learning framework by incorporating several new features, such as separate layers for representing features and categories, top-down and bottom-up interactions between these layers, lateral inhibition, and mechanisms for regulating attention and detecting novelty. These mechanisms allow such models to control the granularity of classification, while achieving a balance between plasticity (the ability to learn new categories) and stability (the ability to retain categories that have already been learned).

All of the learning algorithms discussed above belong to the category of unsupervised, or self-organizing, algorithms, since their weight modification depends solely on the inputs. A different class of algorithms arises if the inputs are accompanied by some form of feedback indicative of the network's performance. This feedback can be used by the network to to guide the modification of link weights. Such feedback can be non-specific, providing a single measure of the network's performance (such as a simple positive or negative reinforcement signal); or highly specific, providing a detailed description of the desired network response (for example, a specification of the output pattern for each input pattern). Algorithms with non-specific feedback are called reinforcement learning algorithms (Sutton and Barto, 1998); while algorithms with specific feedback are called supervised learning algorithms (Rosenblatt, 1958; Block, 1962; Minsky and Papert, 1988; Rumelhart *et al*., 1986; Werbos, 1994).

In the most general terms, reinforcement learning means learning to act in a manner that maximizes the expected future reinforcement (reward). A simplified version of this problem can be formulated as follows. In a trial T, the system observes an input pattern (stimulus) $x(T)$ and responds by performing action $a(x(T))$. After a number of such moves, the system receives a positive or negative reinforcement. The objective of the learning algorithm is to discover a policy whereby at each trial, the system chooses an action that maximizes the expected total reward. This is difficult for several reasons. First, there is uncertainty associated with actions: given the complexity and variability of the environment, the result of action a_i in response to input $x(T)$ is not fixed. Second, there can be a significant delay between choosing an action and receiving a reinforcement. Third, examples of optimal actions are not provided and must be discovered by the system through trial and error. Several reinforcement learning algorithms have been developed. These include the 'temporal difference method' and 'Q-learning' (Sutton and Barto, 1998). The central idea in these learning algorithms is, for each reward signal, to assign appropriate credit or blame to the decisions that led to that signal. After enough training, the system acquires tables that list the best action (that which maximizes the expected reward) for each combination of state and input. These tables are taken to stand for weighted connections in a neural network.

Reinforcement learning is more plausible biologically than supervised learning, which requires a teacher to provide detailed answers rather than just a numerical score. It has been applied to models of motor control and related phenomena. But it is much slower than supervised learning in complex tasks, and has not been widely used in cognitive models.

THE PERCEPTRON AND ITS SUCCESSORS

The most striking early result about supervised learning was the perceptron learning theorem (Rosenblatt, 1958; Block, 1962; Minsky and Papert, 1988). The perceptron model uses a single layer of linear threshold units for categorizing simple visual patterns. Each unit calculates the weighted sum of its inputs and fires (outputs 1) if this sum is greater than the unit's threshold. The associated learning rule involves changing each weight if the unit's prediction about an input pattern was wrong. The theorem showed that any classification that could be computed by a perceptron could be learned by this simple rule.

This theorem gave rise to the belief that ever more complex behavior could be learned directly from feedback. This turned out to be overly optimistic as other scientists, notably Minsky and Papert (1988), showed that many simple distinctions could not be captured by such simple networks. This led to a relatively quiet period for connectionist modeling.

Around 1980, a variety of ideas from biology, physics, psychology, computer science and engineering coalesced to yield a 'new connectionist' approach to modeling intelligence, which has become a core field of cognitive science and also the basis for a wide range of practical applications (McClelland and Rumelhart, 1986). Among the important advances was a mathematical technique (back propagation) that extended the early work

on perceptron learning to a much richer set of network structures. Two ideas allowed error feedback to be used to train networks with multiple layers. The first idea was to replace the linear unit function of the perceptron with a smooth and bounded function, typically the sigmoid described above. The second idea exploits the chain rule for partial derivatives to assign appropriate amounts of credit and blame to model units that do not connect directly to the output nodes, where comparisons with training data are made. While it is theoretically possible to apply back propagation to arbitrary networks (McClelland and Rumelhart, 1986), this does not work well in practice, and essentially all of the back propagation work in cognitive science has been based on two architectural styles.

Most of the early connectionist learning models used strictly layered networks with no feedback at all (McClelland and Rumelhart, 1986). These have limited computational ability, but are still used as components of larger models, some of which can be quite elaborate (Plaut and Kello, 1999). A simple modification of the layer architecture (Elman *et al*., 1996) involves adding fixed-weight connections from a hidden layer at one timestep to the same layer at the next timestep. Because the feedback weights remain fixed, the back propagation learning algorithm remains the same, but these networks can learn some serial tasks beyond the capabilities of layered networks without feedback. Most of the current work on learning in unstructured tasks uses this architecture.

THE CONNECTIONIST APPROACH TO COGNITIVE SCIENCE

The basic connectionist style of modeling is now being used in different ways in neurobiology, in applications, and in cognitive science. Neurobiologists who study networks of neurons employ a wide range of computational models, from very detailed descriptions of the internal chemistry of the neuron to the abstract units described above.

In cognitive science, connectionist techniques have been used for modeling all aspects of language, perception, motor control, memory and reasoning. This universal coverage represents a potential breakthrough: previous computational models of, for example, early vision and of problem solving used entirely different mathematical and computational techniques. Since the brain is known to use the same neural computation throughout, it is not too surprising that models based on this paradigm can be applied to all behavior. The existing models are neither broad nor deep enough to ensure that the current set of mechanisms will suffice to bridge the gap between structure and behavior, but the work remains productive.

Connectionist models in cognitive science belong to two general categories, often called structured (or localist) and layered (or parallel distributed processor (PDP)) networks. Most connectionists are primarily interested in learning, which is modelled as experience-driven change in connection weights. There is a great deal of research studying different models of learning with and without supervision, different rules for changing weights, and so on. Since the focus of such an experiment is on what the network can learn, any imposed structure will weaken the results of the experiment. The standard approach is to use networks with unidirectional connections arranged in completely connected layers, sometimes with a very restricted additional set of feedback links. This kind of network contains a minimal amount of imposed structure and is also amenable to efficient learning techniques such as the back propagation method described above. Most researchers using totally connected layered models do not believe that the brain shares this architecture, but there is a controversy, which we will discuss later, about the implications of PDP learning models for theories of mind.

Structured connectionist models are usually focused less on learning than on the representation and processing of information. Essentially all the modeling done by neurobiologists involves specific architectures, which are known from experiment. For structured connectionist models of cognitive phenomena, the underlying brain architecture is rarely known in detail and sometimes not at all at the level of neurons and connections. The methodology employed is to experiment with computational models of the behavior under study that are consistent with the known biological and psychological data and are plausible in terms of the resources (neurons, computing time, etc.) they require. This methodology is very similar to what are called spreading activation models and are widely used in psycholinguistics. Some studies combine structured and layered networks (Regier, 1996) or investigate learning in networks with initial structure dependent on the problem area or the known neural architecture.

Another important difference between the structured and the layered approaches is that the

layered approach assumes that each 'item' (or mental object) is represented as a pattern of activity distributed over a common pool of nodes (van Gelder, 1992). This notion of 'holographic' representation suffers from some fundamental limitations. Consider the representation of 'John and Mary'. If 'John' and 'Mary' are represented as patterns of activity over the entire network such that each node in the network has two specific patterns for 'John' and 'Mary' respectively, then how can the network represent both 'John' and 'Mary' at the same time? The situation gets even more complex if the system has to represent relations such as 'John loves Mary', or 'John loves Mary but Tom loves Susan'. In contrast to the layered approach, the structured approach uses small clusters of nodes with distinct representational status.

An early example of a structured connectionist model was the interactive activation model for letter perception developed by McClelland and Rumelhart (1981). This model consisted of three layers of nodes, corresponding to visual features of letters, letters, and words. Nodes representing mutually exclusive hypotheses within the letter and word layers inhibited each other. For example, nodes representing letters in the same position inhibited each other, since only one letter can exist in a given position. A node in the feature layer was connected via *excitatory* connections (connections with positive weights) to nodes in the letter layer representing letters that contained that feature. Similarly, a node in the letter layer was connected via *excitatory* connections to nodes in the word layer representing words that contained that letter in the appropriate position. Additionally, there were reciprocal connections from the word layer to the letter layer. The interconnection pattern allowed bottom-up perceptual processing to be guided by top-down expectations. The model could explain a number of psychological findings about the preference of words and pronounceable non-words over other non-words and isolated letters. Other examples of early structured connectionist models included word sense disambiguation models (Cottrell and Small, 1983; Waltz and Pollack, 1985), and the semantic network model CSN (Shastri and Feldmanx, 1986). CSN encoded 'is a' relations using links, and property values using binder nodes that connected property, value and concept nodes. CSN could infer the most likely value of a specified property for a given concept, and find the concept that best matched a partial description.

The Binding Problem

A critical limitation of the early connectionist models was that they could not encode dynamic bindings. Consider the representation of the event 'John gave Mary a book'. This cannot be represented by simply activating the conceptual roles 'giver', 'recipient', and 'gift', and the entities 'John', 'Mary', and 'a book'. Such a representation would be identical to that of 'Mary gave John a book'. Unambiguous representation of an event requires the representation of bindings between the roles of an event (e.g., giver) and the entities that fill these roles in the event (e.g., John). In conventional computing, binding is carried out by variables and pointers, but these techniques have no direct neural counterparts.

Structured connectionist modelers have made significant progress towards neurally plausible solutions to the binding problem. Some of these models use the relative positions of active nodes and the similarity of firing patterns to encode bindings (Barnden and Srinivas, 1991). Some assign distinct activation patterns (signatures) and propagate these signatures to establish bindings (e.g., Lange and Dyer, 1989). Some models use synchronous firing of nodes to represent and propagate dynamic bindings (Shastri and Ajjanagadde, 1993; Hummel and Holyoak, 1997). The possible role of synchronous activity for feature binding had been suggested earlier by others (e.g., von der Malsburg, 1981), but a model called Shruti (Shastri and Ajjanagadde, 1993) offered the first detailed computational account of how such activity can be used to solve problems in the representation and processing of high-level conceptual knowledge and to carry out inference.

Shruti is a structured connectionist model that can encode semantic, causal, and episodic knowledge and perform inferences to establish referential and causal coherence. Shruti encodes relational knowledge using neural circuits composed of focal cell clusters. A systematic mapping between relations (and other rule-like knowledge) is encoded by highly efficacious links between focal clusters. Inference in Shruti results from the propagation of rhythmic activity across interconnected focal clusters. There is no interpreter or inference engine that manipulates symbols or applies rules of inference. In general, Shruti combines predictive inferences with explanatory (or abductive) inferences, instantiates new entities during inference, and unifies multiple entities by merging their phases of firing.

Recruitment Learning

In addition to incremental learning driven by repeated exposure to a large body of training data, structured models have also made use of one-trial learning using 'recruitment learning' (e.g., Feldman, 1982). Recruitment learning can be described as follows. Learning occurs within a structured network containing an unassigned group of randomly connected nodes. Recruited nodes in such a network acquire distinct 'meanings' (or functionality) by virtue of their strong connections to previously structured nodes. For example, a novel concept that is a conjunct of existing concepts x_1 and x_2 can be learned by (1) identifying free nodes that receive links from nodes representing both x_1 and x_2 and (2) 'recruiting' one or more such free nodes by strengthening the weights of links incident on such nodes from x_1 and x_2 nodes. In general, the recruitment process can transform a quasi-random network into a collection of nodes and circuits with specific functionalities.

It has been shown (Shastri, 1999) that recruitment learning can be firmly grounded in the biological phenomena of long-term potentiation and long-term depression, which involve rapid, long-lasting, and highly specific changes in synaptic strength.

Rules versus Connections

Perhaps the most visible contribution of connectionist computational models in cognitive science has been to provide a new conceptual framework for some long-standing debates on the nature of intelligence. Much of the current debate is being published in scientific journals. The question of nature versus nurture concerns how much of some trait, usually intelligence, can be accounted for by genetic factors and how much depends on postnatal environment and training. Some PDP connectionists have taken very strong positions suggesting that learning can account for everything interesting (Elman *et al*., 1996). In the particular case of grammar, an important group of linguists and other cognitive scientists take an equally extreme nativist position, suggesting that humans only need to choose a few parameters to learn grammar. A related question is the whether human grammatical knowledge is represented as general rules or just appears as the rule-like consequences of PDP learning in the neural network of the brain. There is ample evidence against both extreme positions, but the debate continues to motivate a great deal of thought and experimentation.

CURRENT AND FUTURE TRENDS

Quantitative neural models are playing a major role in cognitive science. The mathematical and computational ideas underlying learning in neural networks have also found application in a wide range of practical problems, from speech recognition to financial prediction. Given current computing power, back propagation and similar techniques allow large systems of nonlinear units to learn reasonably complex probabilistic relationships using labelled data. This general methodology overlaps not only with AI but also with mathematical statistics, and is part of a unifying theory called computational learning theory. There is also a large community of scientists and engineers who are working on neural networks and related statistical techniques for various tasks, and there are conferences and journals to support this effort.

The explosion of activity on the internet is affecting the whole field of computing. Two application areas that seem particularly important to cognitive science are intelligent web agents and spoken language interaction. As the range of users and activities on the internet continues to expand, there is increasing demand for systems that are both more powerful and easier to use. This is leading to increased efforts on the design of human–computer interfaces, including the modeling of users' plans and intentions – clearly overlapping with traditional concerns of cognitive science. One particularly active area is interaction with systems using ordinary language. While machine recognition of individual spoken words is relatively successful, dealing with the full richness of language is one of the most exciting challenges facing computing and cognitive science, and a problem of great commercial and social importance.

Looking ahead, technical work on connectionist modeling is likely to remain closely linked to statistics and learning theory. From the scientific perspective, it is likely that most interdisciplinary connectionist research in cognitive science will remain focused on specialized domains such as language, speech and vision. With the rapid advances in neurobiology, the field will increasingly connect with the life sciences, yielding great mutual benefits (e.g. Carter *et al*., 1998).

References

Ackley DH, Hinton GF and Sejnowski TJ (1985) A learning algorithm for Boltzman machines. *Cognitive Science* **9**: 147–169.

Anderson JA and Rosenfeld E (eds) (1988) *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.

Barnden J and Srinivas K (1991) Encoding techniques for complex information structures in connectionist systems. *Connection Science* **3**(3): 269–315.

Bienenstock EL, Cooper LN and Munro PW (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* **2**: 32–48. [Reprinted in Anderson and Rosenfeld, 1988.]

Block HD (1962) The perceptron: a model for brain functioning. *I Reviews of Modern Physics* **34**: 123–135. [Reprinted in Anderson and Rosenfeld, 1988.]

Carter CS, Braver TS, Barch DM *et al*. (1998) Anterior cingulate cortex, error detection and the on-line monitoring of performance. *Science* **280**: 747–749.

Cottrell GW and Small SL (1983) A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory* **6**: 89–120.

Elman J, Bates E and Johnson M (1996) *Rethinking Innateness: A Connectionist Perspective on Development (Neural Network Modeling and Connectionism)*. Cambridge, MA: MIT Press.

Feldman JA (1982) Dynamic connections in neural networks. *Biological Cybernetics* **46**: 27–39.

Feldman JA and Ballard DB (1982) Connectionist models and their properties. *Cognitive Science* **6**: 205–254. [Reprinted in Anderson and Rosenfeld, 1988.]

van Gelder T (1992) Defining 'distributed representation'. *Connection Science* **4**(3,4): 175–191.

Grossberg S (1976) Adaptive pattern classification and universal recoding. *Biological Cybernetics* **23**: 121–134. [Reprinted in (Anderson and Rosenfeld, 1988).]

Grossberg S (1980) How does a brain build a cognitive code? *Psychological Review* **87**: 1–51. [Reprinted in Anderson and Rosenfeld, 1988.]

Hebb DO (1949) *The Organization of Behavior*. New York, NY: Wiley.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79**: 2554–2558. [Reprinted in Anderson and Rosenfeld, 1988.]

Hummel JE and Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* **104**: 427–466.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**: 59–69. [Reprinted in Anderson and Rosenfeld, 1988.]

Lange TE and Dyer MG (1989) High-level inferencing in a connectionist network. *Connection Science* **1**(2): 181–217.

von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14**: 85–100. [Reprinted in Anderson and Rosenfeld, 1988.]

von der Malsburg C (1981) The correlation theory of brain function. Internal Report 81–2, Department of Neurobiology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception: part 1: an account of basic findings. *Psychological Reviews* **88**: 375–407. [Reprinted in Anderson and Rosenfeld, 1988.]

McClelland J and Rumelhart D (1986) *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

McCulloch WS (1988) *Embodiments of Mind*. Cambridge, MA: MIT Press. [Reprint edition.]

Minsky ML and Papert SA (1988) *Perceptrons: Introduction to Computational Geometry*. Cambridge, MA: MIT Press. [Expanded edition; first published 1969.]

Plaut DC and Kello CT (1999) The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In: MacWhinney B (ed) *The Emergence of Language*, pp. 381–415. Mahwah, NJ: Erlbaum.

Regier T (1996) *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408. [Reprinted in Anderson and Rosenfeld, 1988.]

Rumelhart D, Hinton GE and Williams RJ (1986) *Learning internal representations by error propagation*. In: McClelland and Rumelhart, 1986.

Rumelhart D and Zipser D (1985) *Feature discovery by competitive learning*. In: McClelland and Rumelhart, 1986.

Shastri L (1999) Biological grounding of recruitment learning and vicinal algorithms in long-term potentiation and depression. Technical Report TR-99-009, International Computer Science Institute, Berkeley, CA.

Shastri L and Ajjanagadde A (1993) From simple associations to systematic reasoning: a connectionist encoding of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences* **16**(3): 417–494.

Shastri L and Feldman JA (1986) Neural nets, routines and semantic networks. In: Sharkey N (ed.) *Advances in Cognitive Science*, pp. 158–203. Chichester, UK: Ellis Horwood.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Waltz D and Pollack J (1985) Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science* **9**: 51–74.

Werbos P (1994) *The Roots of Back-Propagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. New York, NY: Wiley.

Connectionist Architectures: Optimization

Advanced article

Marcus Frean, Victoria University of Wellington, Wellington, New Zealand

CONTENTS

A key issue in using connectionist methods is the choice of which network architecture to use. There are a number of ways this choice can be made automatically, driven by the problem at hand.

INTRODUCTION

If one takes a training set in the form of input–output pairs and trains a large connectionist network on it, the result is generally 'overfitting'. There are many functions which exactly fit the existing data and the act of learning arrives at just one of them, somewhat arbitrarily. The problem is compounded where the data is noisy, in which case the network uses its extra degrees of freedom to fit the noise rather than the underlying function generating the data. Conversely, if the network is too small it 'underfits', which is equally unsatisfactory. The real aim is usually not to get the training set correct, but to generalize successfully to new data. The model selection problem is to arrive at the network that gives the best possible predictions on new inputs, using only the available training data and prior knowledge about the task. (*See* **Machine Learning**)

There are several ways of controlling the complexity of mappings learned by neural networks. These include varying the number of weights or hidden units by building up or paring down an existing network, and direct penalties (otherwise known as regularization) on model complexity, such as weight decay. Other ideas include partitioning the input space into regions which are locally linear as in 'mixtures of experts', or using genetic algorithms to choose between different architectures.

CRITERIA FOR NETWORK OPTIMALITY

The optimality or otherwise of a network is, in many cases, determined by its ability to generalize. Almost by definition this ability is not directly observable, so in practice we have to make an educated guess at it and use that to choose between networks.

The simplest method takes part of the available data and sets it aside. Once the network has been trained on the remaining data it can be 'validated' by seeing how well it performs on the withheld data, thus giving an estimate of how well it will generalize. This estimate won't be very good unless the hold-out set is large, which wastes a lot of the data that could otherwise be used for training. To minimize this effect, 'cross-validation' applies the same idea repeatedly with different subsets of the data, retraining the network each time. In k-fold cross-validation, for example, the data is divided into k subsets. One at a time, these serve as hold-out sets, and the validation performance is then averaged across them to give an estimate of how well the network generalizes. 'Leave-one-out' cross-validation uses $k = N$, the number of samples, but $k = 10$ is typically used.

Another general approach, from conventional statistics, is to attempt to quantify the generalization performance of trained networks without a validation set at all. One prefers networks with a low 'prediction error'

$$C = C_{data} + C_{net} \quad (1)$$

Here C_{data} is the usual training error, such as the sum of squared errors, and C_{net} is taken to be a measure of the complexity of the network,

proportional to the effective number of free parameters it has. Assuming a nonlinear network is locally linear in the region of the minimum, an approximation to C_{net} can be calculated (Moody, 1992; Murata *et al*., 1994) given the Hessian matrix of second derivatives $H_{ij} = \partial^2 C_{data}/\partial w_i \partial w_i$, which can be found using a number of methods (Buntine and Weigend, 1994). For large training sets, leave-one-out cross-validation and the above are essentially equivalent, with the latter giving the same effect for much less computational effort. These approaches assume a single minimum however, so the estimate can be strongly affected by local minima. On the other hand, leave-one-out cross-validation also gets trapped in local minima, in which case 10-fold cross-validation is preferable.

A third approach is to use Bayesian model comparison to choose between networks, as well as to set other parameters such as the amount of weight decay. Bayesians represent uncertainty of any kind by an initial or 'prior' probability distribution, and use the laws of probability to update this to a 'posterior' distribution in the light of the training set. In this view we should choose between models based on their posterior probabilities given the available data – again this does not require that any data be set aside for validation (MacKay, 1995). In the fully Bayesian approach, ideally we should use not one set of weights and one structure but many sets and many architectures, weighting the prediction of each by their posterior probability. To the extent this averaging can be done, deciding on an 'optimal' model (and indeed all learning as it is usually thought of) becomes unnecessary. (*See* **Reasoning under Uncertainty**; **Pattern Recognition, Statistical**)

Generalization performance is not the only measure of usefulness of a given network architecture. Other potentially important measures are its fault tolerance, training time on the problem at hand, robustness to 'catastrophic forgetting', ease of silicon implementation, speed of processing once trained, and the extent to which hidden representations can be interpreted. (*See* **Catastrophic Forgetting in Connectionist Networks**)

PRUNING OF UNIMPORTANT CONNECTIONS OR UNITS

Pruning algorithms start by training an overly complex network before trimming it back to size. In other words, we knowingly overfit the data and then reduce the number of free parameters, attempting to stop at just the right point. Clearly the general model selection schemes described above (cross-validation, estimated prediction error, and Bayesian model comparison) can be applied to prune overly large networks; however a number of ideas have been formulated that are specific to pruning. Pruning algorithms can remove weights or whole units, and one can think of the choice of which element to remove as being driven by a measure of 'saliency' for that element. Each algorithm uses a different form for this saliency.

A simple measure of saliency to use for weights is their absolute value. However, while it may be true that removing the smallest weight affects the network the least, it doesn't follow that this is the best weight to remove to improve generalization. Indeed this seems completely opposite to weight decay (see below), which in effect 'removes' large weights by decaying them the most, and it performs poorly in practice.

A more principled idea, known as 'optimal brain damage' (Le Cun *et al*., 1990), approximates the change to the error function that would be caused by removal of a given connection or unit, and uses this measure to decide which to remove. To make this approximation, one trains the network until it is at a minimum of the usual error function, and then calculates the Hessian H. Ignoring the off-diagonal elements of this matrix, the saliency of the weight is given by $H_{ii}w_i^2$.

This idea has been further developed in 'optimal brain surgeon' (Hassibi and Stork, 1993), which avoids the assumption that the Hessian is diagonal. Interestingly this gives a rule for changing all the weights, with the constraint that one of these involves the setting of a weight to zero. One must first calculate the full inverse Hessian matrix however, which can make the algorithm slow and memory intensive for large problems.

Statistical tests can also be applied to detect non-contributing units that could be made redundant. For example if two units are in the same layer and are perfectly correlated (or anti-correlated) in their activity, we know the network can perform the same mapping with one of them removed. A particularly simple case is when a unit has the same output all the time, making it functionally no different from the bias unit.

WEIGHT DECAY

In networks whose output varies smoothly with their input, small weights give rise to outputs which change slowly with the input to the net, while large weights can give rise to more abrupt changes of the kind seen in overfitting. For this reason one response to overfitting is to penalize

the network for having large weights. The most obvious way to do this is by adding a new term

$$C_{net} = \frac{\beta}{2}\sum_i w_i^2 \quad 0 < \beta < 1 \tag{2}$$

to the objective function being minimized during learning. The total cost $C = C_{data} + C_{net}$ can then be minimized by gradient descent. For a particular weight we have

$$\Delta w \propto -\frac{\partial C}{\partial w} = -\frac{\partial C_{data}}{\partial w} - \beta w \tag{3}$$

The first term leads to a learning rule such as back propagation (depending on the form of C_{data}), while the second removes a fixed proportion of the weight's current value. Hence each weight has a tendency to decay toward zero during training, unless pulled away from zero by the training data. The 'decay rate' β determines how strong this tendency is. Clearly a major question is how to set β, for which the general methods described earlier are applicable. (*See* **Backpropagation**)

Weight decay helps learning in other ways as well as its effect on generalization – it reduces the number of local minima, and makes the objective function more nearly quadratic so quasi-Newton and conjugate gradient methods work better.

From a Bayesian perspective, weight decay amounts to finding a *maximum a posteriori* (MAP) estimate given a Gaussian prior over the weights, reflecting our belief that the weights should not be too large. Weight decay is not usually applied to bias weights, reflecting the intuition that we have no *a priori* reason to suppose the bias offset should be small. Depending on the nature of the problem, this may not be a particularly sensible prior – for instance we may actually believe that most weights should be zero but that some should be substantially nonzero. One expression of this to use a different weight cost such as

$$C_{net} = \beta \sum_i \frac{w_i^2}{c^2 + w_i^2} \tag{4}$$

This has been called weight elimination, because it is more likely to drive weights towards zero than simple weight decay. Very small weights can then be eliminated. c is a second parameter which needs to be set by hand. An interesting alternative is 'soft weight sharing' (Nowlan and Hinton, 1992) which implements MAP with a prior that is a mixture of Gaussians. The means (which need not be zero) and variances of these Gaussians can be adapted by the learning algorithm as training proceeds.

GENERATIVE ARCHITECTURES

Generative architectures, also called constructive algorithms, build networks from scratch to suit the problem at hand. Once each unit is trained, its weights are 'frozen' before building the next unit. An important advantage of this is that only single layers of weights are being trained at any one time. Accordingly the learning rules involved need only be local to the unit in question (unlike back propagation), which tends to make learning particularly fast and straightforward.

For simplicity each algorithm is described here as it applies to a single output unit – multiple outputs are trivial extensions to this, as described in the cited papers.

Upstarts

The upstart algorithm (Frean, 1990) is a method for constructing a network of threshold units. Imagine a single linear threshold unit (perceptron) that is trained to minimize the number of errors it makes on the training set, and then frozen. This unit, which we will call u, makes two kinds of error: it is either wrongly on, or wrongly off. In the upstart algorithm these errors are dealt with separately by recruiting two new units, which we could call u_- and u_+, one for each type of error. These new units receive the same inputs, but their outputs go directly to u (see Figure 1(a)). The role of u_- is to correct the 'wrongly on' errors by the parent unit u so it has a large negative output weight, while the output weight of u_+ is large and positive since its function is to correct the wrongly off errors. (*See* **Perceptron**)

It is easy to derive appropriate targets for u_-, given the original targets and u's responses: u_- is to be given the target 1 whenever u is wrongly on, while in all other cases its target should be 0. Similarly the target for u_+ is 1 whenever u is wrongly off, and 0 otherwise. Notice that the output of u_- (u_+) does not matter if u was already correctly off (on), so these can be omitted from the child node's training set. Should the child units be free of errors on their respective training sets, u will itself be error-free. If, however, either u_- or u_+ still make mistakes of their own, these errors are likewise of two types and we can apply the same idea, recursively. The result is a binary tree of units, grown 'backwards' from the original output unit. Child nodes spend their time loudly correcting their parent's mistakes, hence the algorithm's name.

Suppose u_- has the output 1 for just one of the wrongly on patterns of u, and is zero in all other

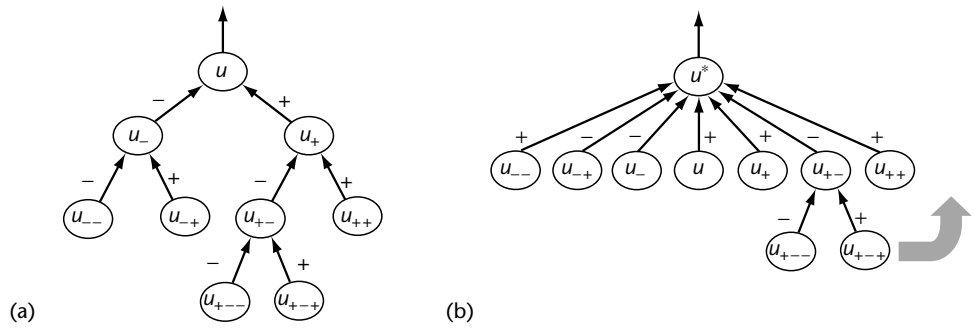


Figure 1. (a) A binary tree constructed by the upstart algorithm. All units have direct inputs, omitted here for clarity. The 'leaves' of the tree make no errors, so neither does the root node. (b) The same network being rearranged into a single hidden layer.

cases. For convex training sets (e.g. binary patterns) it is always possible to 'slice off' one pattern from the others with a hyperplane, so in this case it is easy for u_- to improve u by at least one pattern. Of course we hope that u_- and u_+ will confer much more advantage than this in the course of training. In practice a quick check is made that the number of errors by u_- is in fact lower than the number of wrongly on errors by u, to ensure convergence to zero errors.

Networks constructed using this method can be reorganized into a single hidden layer, if desired. That is, a new output unit can get zero errors by being connected to this layer with weights which are easily found, as shown in Figure 1(b).

For noise-free data this procedure usually produces networks that are close to the smallest that can fit the data, with attendant gains in generalization ability compared to larger networks. Notice however that the training set is learned without errors, so this is just as prone as any other algorithm to overfitting of noisy data (the idea has not been generalized to handle such noise, although there seems no reason why this couldn't be done). One can also use the same procedure to add hidden units to a binary attractor (Hopfield) network, thereby increasing its memory capacity from $\sim N$ to 2^N patterns.

The Pyramid Algorithm

Another algorithm for binary outputs is the pyramid algorithm (Gallant, 1993). One begins as before with a single binary unit, connected to the inputs and trained to minimize the number of errors. This unit is then 'frozen' and (assuming errors are still being made) a new unit is designated the output: this new unit sees both the regular inputs and any (frozen) predecessors as its input, as shown in Figure 2. (*See* **Perceptron**)

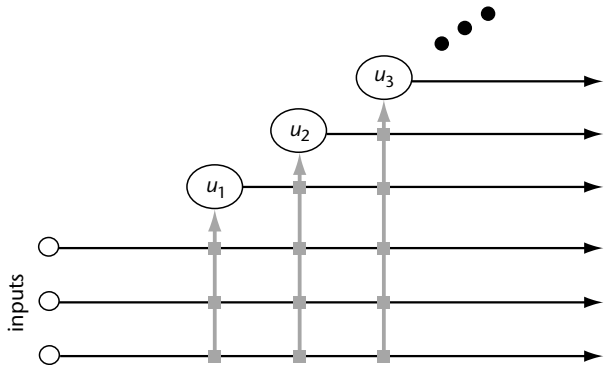


Figure 2. The architecture constructed by the pyramid algorithm. Vertical lines represent multiple connections (shown as squares) to the unit above. Each new unit assumes the role of output.

It is not hard to show that this new unit can achieve fewer errors than its predecessor, provided the input patterns are convex. If it sets its weights from the network inputs to zero and has a positive weight from the previous frozen unit, these two obviously make the same number of errors. As with upstarts, given convex inputs it could then easily reset its input weights so that this behavior was altered for just one input pattern where it was previously in error. This is the 'worst case' behavior, and appropriate weights can easily be predefined, to be improved by training (any method for arriving at good weights for a single unit is applicable). Despite its apparently 'greedy' approach to optimization and its extreme simplicity, the method can build concise networks. For example, given the N-bit parity problem (where the task is to output the parity of a binary input) the upstart algorithm generates a network with N hidden units, while the pyramid algorithm builds the apparently minimal network having only $(N+1)/2$

hidden units. Like the upstart algorithm, the method as it stands is prone to overfitting noisy data.

Cascade Correlation

Cascade correlation (Fahlman and Lebiere, 1990) can be applied to networks with real-valued outputs, and uses sigmoidal hidden units. We begin with a network having only direct connections between inputs and outputs, with no hidden units. These weights are trained using gradient ascent (the delta rule), or whatever learning procedure you like. We then introduce a hidden unit, with connections to the input layer. This unit sends its output via new weighted connections to the original output layer. In upstarts, the hidden unit is binary and is preassigned one of two roles, which determines how it is trained and the sign of its output weight – this is because being binary it can only correct errors of one type by the output unit, given its output weight. In this case, however, the hidden unit is real-valued and this means it can play a role in correcting errors of either sign by the output unit. Fahlman and Lebiere's idea is to train the hidden unit to maximize the covariance between its output and the existing errors by the output units. We can then use this gradient to learn input weights for the hidden unit in the usual way. When this phase of learning is deemed to have finished, all the output unit's connections are retrained, including the new ones. The process can now be repeated with a new hidden unit, with each such unit receiving inputs from the original inputs as well as all previous hidden units. Figure 3 shows the resulting cascade architecture.

A potential problem is that the output can make a lot of errors yet, after averaging, the correlation with a hidden unit can be very small. Despite this the method seems to work well in practice, and can be extended to recurrent networks.

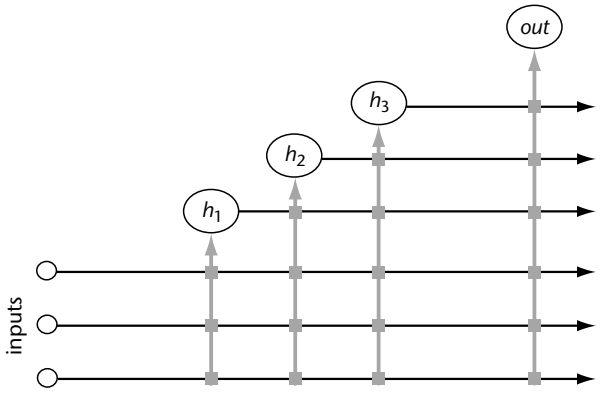


Figure 3. The architecture constructed by cascade correlation.

ADAPTIVE MIXTURES OF EXPERTS

In conventional back propagation networks, each sigmoid unit potentially plays a part in the network's output over its entire range of inputs. One way of restricting the power of the network is to partition the input space into distinct regions, and restrict the influence of a given unit to a particular region. Ideally we would like to learn this partition rather than assume it from the beginning.

A particularly appealing way to do this is known as the 'mixture of experts' architecture (Jacobs *et al.*, 1991). Each 'expert' consists of a standard feedforward neural network. A separate 'gating' network, with as many outputs as there are experts, is used to choose between them. The output of this network is chosen stochastically using the softmax activation function at its output layer,

$$Pr(i = 1) = \frac{e^{\phi_i}}{\sum_j e^{\phi_j}} \tag{5}$$

where ϕ_i is the weighted sum into the ith output unit. This reflects the fact that only one of the outputs is active at any given time. All of the nets (including the switch) are connected to the same inputs as shown in Figure 4, but only the expert that happens to be chosen by the switch is allowed to produce the output. A learning procedure can be found by maximizing the log likelihood of the network generating the training outputs given the inputs, in the same way as the cost function for back propagation is derived. Indeed the learning rule for the 'experts' turns out to be simply a weighted version of back propagation. During learning each expert network gets better at generating correct outputs for the input patterns that are

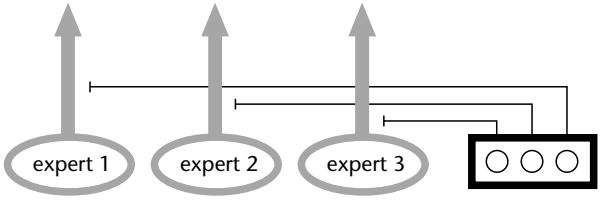


Figure 4. The mixtures of experts architecture. Each expert consists of a separate network, and may have multiple outputs. A separate gating network acts as a switch, allowing just one of the experts to generate the output. All of the experts, together with the gating network, have access to the input pattern.

assigned to it by the switch. At the same time, the switch itself learns to apportion inputs to the best experts. One can think of the switch as performing a 'soft' partition of the input space into sections which are learnable by individual experts. (*See* **Backpropagation**)

A further possibility is to treat each expert as a mixture of experts system itself (Jordan and Jacobs, 1994). A form of hierarchical decomposition of the task can thus be repeated for as many levels as desired. If simple linear units are used for the leaf nodes of the resulting tree-structured network, training can be achieved using a version of the expectation-maximization (EM) algorithm rather than gradient descent.

USING GENETIC ALGORITHMS TO EVOLVE CONNECTIONIST ARCHITECTURES

One criticism of both pruning and constructive algorithms is that they alter networks in only very limited ways, and as such they are prone to getting stuck in local optima in the space of possible architectures. Genetic algorithms offer a richer variety of change operators in the form of mutation and crossover between encodings (called 'chromosomes') of parent individuals in a population. The hope is that networks which are more nearly optimal may be found by evolving such a population of candidate structures, compared to making limited incremental changes to a single architecture. (*See* **Evolutionary Algorithms**)

In generating new candidate architectures, genetic algorithms choose parents based on their performance ('fitness'), which may be evaluated using the techniques described previously for determining network optimality. The main contribution of genetic algorithms then is their more general change operators, principally that of crossover, which operate on the chromosome rather than the network directly. Accordingly, the way in which architectures are mapped to chromosomes and vice versa is of central importance (Yao, 1999).

One approach is to assume an upper limit N to the total number of units and consider an $N \times N$ connectivity matrix, whose binary entries specify the presence or absence of a connection. Any units without outputs are effectively discarded, as are those lacking inputs. A population of such matrices can then be evolved, by training each such network using a learning algorithm initialized with random weights. To apply genetic operators, each matrix is simply converted to a vector by concatenating its rows. Restriction to feedforward networks is straightforward: matrix elements on and below the diagonal are set to zero, and are left out of the concatenation.

A drawback of this approach (though by no means unique to it) is that the evaluation of a given network is very noisy, essentially because the architecture is not evaluated on its own but in conjunction with its random initial weights. Averaging over many such initializations is computationally expensive, and one solution is to evolve both the connections and their values together. In this case an individual consists of a fully specified architecture together with the values of weights. On the other hand cross-over makes little sense for combining such specifications (unless the neural network uses localist units such as radial basis functions) because it destroys distributed representations.

Less direct encodings can be used, such as rules for generating networks, rather than the networks themselves. Evolutionary algorithms have also been used to change the transfer functions used by units (such as choosing between sigmoid and Gaussian for each unit), and even to adapt the learning rules used to set the weights.

References

Buntine WL and Weigend AS (1994) Computing second derivatives in feedforward networks: a review. *IEEE Transactions on Neural Networks* **5**(3): 480–488.

Fahlman SE and Lebiere C (1990) The cascade correlation learning architecture. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems*, vol. II, pp. 524–532. San Mateo, CA: Morgan Kaufmann.

Frean M (1990) The upstart algorithm: a method for constructing and training feedforward neural networks. *Neural Computation* **2**(2): 198–209.

Gallant SI (1993) Neural network learning and expert systems. Cambridge, MA: MIT Press.

Hassibi B and Stork DG (1993) Second-order derivatives for network pruning: optimal brain surgeon. In: Hanson SJ, Cowan JD and Giles CL (eds) *Advances in Neural Information Processing Systems*, vol. V, pp. 164–171. San Mateo, CA: Morgan Kaufmann.

Jacobs RA, Jordan MI, Nowlan SJ and Hinton GE (1991) Adaptive mixtures of local experts. *Neural Computation* **3**(1): 79–87.

Jordan MI and Jacobs RA (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**(2): 181–214.

Le Cun Y, Denker JS and Solla SA (1990) Optimal brain damage. In: Touretzky DS (ed.) *Advances in Neural Information Processing Systems*, vol. II, pp. 598–605. San Mateo, CA: Morgan Kaufmann.

Moody JE (1992) The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In: Moody JE, Hanson SJ and Lippmann RP (eds) *Advances in Neural Information Processing Systems*, vol. IV, pp. 847–854. San Mateo, CA: Morgan Kaufmann.

Murata N, Yoshizawa S and Amari S (1994) Network information criterion – determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks* **5**: 865–872.

Nowlan SJ and Hinton GE (1992) Simplifying neural networks by soft weight sharing. *Neural Computation* **4**(4): 473–493.

Yao X (1999) Evolving artificial neural networks. *Proceedings of the IEEE* **87**(9): 1423–1447.

Further Reading

Bishop C (1995) *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Neal R (1996) *Bayesian Learning for Neural Networks*. New York: Springer-Verlag.

Read RD and Marks RJ (1999) *Neural Smithing – Supervised Learning in Feedforward Artificial Neural Networks*. Cambridge, MA: MIT Press.

Connectionist Implementationalism and Hybrid Systems

Intermediate article

Ron Sun, University of Missouri, Columbia, Missouri, USA

CONTENTS

We may incorporate symbolic processing capabilities in connectionist models, including implementing such capabilities in conventional connectionist models and/or adding additional mechanisms to connectionist models.

INTRODUCTION

Many cognitive models have incorporated both symbolic and connectionist processing in one architecture, apparently going against the conventional wisdom of seeking uniformity and parsimony of mechanisms. It has been argued by many that hybrid connectionist–symbolic systems constitute a promising approach to developing more robust and powerful systems for modeling cognitive processes and for building practical intelligent systems. Interest in hybrid models has been slowly but steadily growing. Some important techniques have been proposed and developed. Several important events have brought to light ideas, issues, trends, controversies, and syntheses in this area. In this article, we will undertake a brief examination of this area, including rationales for such models and different ways of constructing them.

MODELING DIFFERENT COGNITIVE PROCESSES WITH DIFFERENT FORMALISMS

The basic rationale for research on hybrid systems can be succinctly summarized as 'using the right tool for the right job'. More specifically, we observe that cognitive processes are not homogeneous: a wide variety of representations and processes seem to be employed, playing different roles and serving different purposes. Some cognitive processes and representations are best captured by symbolic models, others by connectionist models (Dreyfus and Dreyfus, 1987; Smolensky, 1988; Sun, 1995). Therefore, in cognitive science, there is a need for 'pluralism' in modeling human cognitive processes. Such a need leads naturally to the development of hybrid systems, in order to provide the necessary computational tools and conceptual frameworks. For instance, to capture the full range of skill-learning capabilities, a cognitive architecture needs to incorporate both declarative and procedural knowledge. Such an architecture can be implemented computationally by a combination of symbolic models (which capture declarative knowledge) and connectionist models (which capture procedural knowledge). The development of intelligent systems for industrial applications can also benefit greatly from a proper combination of different techniques, because currently no one technique can do everything successfully. This is the case in many application domains.

The relative advantages of connectionist and symbolic models have been argued at length. (See, for example, Dreyfus and Dreyfus, 1987; Smolensky, 1988 and Sun, 1995 for various views.) The advantages of connectionist models include: massive parallelism; graded representation; learning capabilities; and fault tolerance. The advantages of symbolic models include: crisp representation and processing; ease of specifying detailed processing steps; and the resulting precision in processing. With these relative advantages in mind, the combination of connectionist and symbolic models is

easy to justify: hybrid systems seek to take advantage of the synergy of the two types of model when they are combined or integrated.

Psychologists have proposed many cognitive dichotomies on the basis of experimental evidence, such as: implicit versus explicit learning; implicit versus explicit memory; automatic versus controlled processing; incidental versus intentional learning. Above all, there is the well-known dichotomy between procedural and declarative knowledge. The evidence for these dichotomies lies in experimental data that elucidate various dissociations and differences in performance under different conditions. Although there is no consensus regarding the details of the dichotomies, there is a consensus on the qualitative difference between two types of cognition. Moreover, most researchers believe in the necessity of incorporating both sides of the dichotomies, because each side serves a unique function and is thus indispensable. Some cognitive architectures have been structured around some of these dichotomies.

Smolensky (1988) proposed a more abstract distinction of conceptual versus subconceptual processing; and he related the distinction to that between connectionist and symbolic models. Conceptual processing involves knowledge that possesses the following three characteristics: public access; reliability; and formality. This is what symbolic models capture. There are other kinds of cognitive capacities, such as skill and intuition, that are not expressible in linguistic forms and do not share the above characteristics. It has proved futile to try to model such capacities with symbolic models. These capacities should belong to a different level of cognition: the subconceptual level. The subconceptual level is better modeled by connectionist subsymbolic systems, which can overcome some of the problems faced by symbolic systems modeling subconceptual processing. Therefore, the combination of the two types of models can capture a significantly wider range of cognitive capacities. These ideas provide the justification for building complex hybrid cognitive architectures. For detailed accounts of a variety of examples of the synergistic combination of connectionist and symbolic processes, see Dreyfus and Dreyfus, 1987; Sun, 1995; Waltz and Feldman, 1986; and Wermter and Sun, 2000.

INTEGRATING CONNECTIONIST AND SYMBOLIC ARCHITECTURES

Hybrid models are likely to involve a variety of types of processes and representations, in both learning and performance. Therefore, they will involve multiple heterogeneous mechanisms interacting in complex ways. We need to consider how to structure these different components; in other words, we need to consider architectures. Questions concerning hybrid architectures include:

- Should hybrid architectures be modular or monolithic?
- For modular architectures, should we use different representations in different modules, or the same representations throughout?
- How do we decide whether the representation of a particular part of an architecture should be symbolic, localist, or distributed?
- What are the appropriate representational techniques for bridging the heterogeneity likely in hybrid systems?
- How are representations learned in hybrid systems?
- How do we structure different parts to achieve appropriate results?

Although many interesting models have been proposed, including some that correspond to the cognitive dichotomies outlined above, our understanding of hybrid architectures is still limited. We need to look at the proposed models and analyze their strengths and weaknesses, to provide a basis for a synthesis of the existing divergent approaches and to provide insight for further advances. Below we will provide a broad categorization of the existing architectures.

Architectures of hybrid models can be divided into 'single-module' and 'multi-module' architectures. Single-module systems can be further divided according to their representation types: symbolic (as in symbolic models); localist (i.e. using one distinct node for representing each concept – see, for example, Lange and Dyer, 1989; Sun, 1992 and Shastri and Ajjanagadde, 1993); and distributed (i.e. using a set of overlapping nodes for representing each concept – see, for example, Pollack, 1990 and Touretzky and Hinton, 1988). Usually, it is easier to incorporate prior knowledge into localist models, since their structures can be made to directly correspond to that of symbolic knowledge. On the other hand, connectionist learning usually leads to distributed representation (as in the case of back-propagation learning). Distributed representation has some useful properties.

Multi-module systems can be divided into 'homogeneous' and 'heterogeneous' systems. Homogeneous systems are similar to the single-module systems discussed above, except that they can contain several replicated copies of the same structure, each of which can be used for processing the same set of inputs, to provide redundancy for various reasons; alternatively, each module (of the

same structure) can be specialized for processing inputs of a particular type (of content).

For heterogeneous multi-module systems, several distinctions can be made. First, a distinction can be made in terms of the representations of the constituent modules. There can be different combinations of types of constituent modules: for example, a system may be a combination of localist and distributed modules (as in CONSYDERR, described in (Sun, 1995), or it may be a combination of symbolic and connectionist modules, either localist or distributed (as in CLARION, described in (Sun and Peterson, 1998)).

Second, a distinction can be made in terms of the coupling of modules: a set of modules may be loosely or tightly coupled. In loosely coupled architectures modules communicate with each other, primarily through message passing, shared memory locations, or shared files. This allows for some loose forms of cooperation among modules. One form of cooperation is in terms of the type of processing: while one or more modules take care of preprocessing (e.g. transforming input data) or postprocessing (e.g. rectifying output data), another module focuses on the main part of the task. Preprocessing and postprocessing are commonly done using a neural network, while the main task is accomplished by symbolic methods. Another form of cooperation is through a master–slave relationship: while one module maintains control of the task at hand, it can command other modules to handle some specific aspects of the task. For example, a symbolic expert system, as part of a rule, may invoke a neural network to make a specific classification decision. Yet another form of cooperation is an equal partnership of multiple modules. In this form, the modules (the equal partners) may represent complementary processes; functionally equivalent but structurally and representationally different processes; or differentially specialized and heterogeneously represented 'experts'.

In tightly coupled architectures on the other hand, the constituent modules interact through multiple channels (for example, various possible function calls); or they may even have node-to-node connections between modules (as in CONSYDERR (Sun, 1995) and ACT-R (Anderson and Lebiere, 1988)). As in the case of loosely coupled systems, there are several possible forms of cooperation among modules.

TIGHTLY COUPLED ARCHITECTURES

Let us examine briefly a tightly coupled, heterogeneous, multi-module architecture: CONSYDERR (Sun, 1995). It consists of a concept level and a microfeature level. The representation is localist at the concept level, with one node for each concept, and distributed at the microfeature level, with an (overlapping) set of nodes for representing each concept. Rules are implemented, at the concept level, using links between nodes representing conditions and nodes representing conclusions, and weighted sums are used for evaluating evidence. Rules are diffusely duplicated at the microfeature level in a way consistent with the meanings of the rules. Rules implemented at the concept level capture explicit and conceptual knowledge that is available to a cognitive agent, and diffused representations of rules at the microfeature level capture (to some extent) associative and embodied knowledge. Figure 1 shows a sketch of the model. There are two-way (gated) connections between corresponding representations at the two different levels;

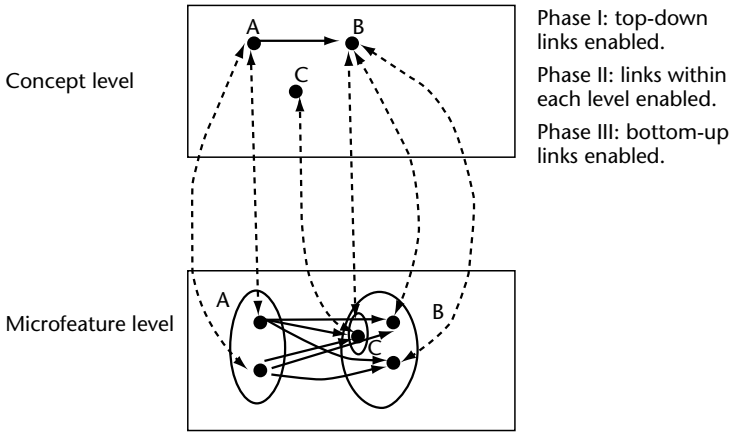


Figure 1. The CONSYDERR architecture.

that is, each concept is connected to all the related microfeature nodes, and vice versa. The operation of the model is divided into three phases: the top-down phase, the settling phase, and the bottom-up phase. In the top-down phase, microfeatures corresponding to activated concepts are themselves activated, enabling similarity-based reasoning at the microfeature level. In the settling phase, rule-based reasoning takes place at each level separately. Finally, in the bottom-up phase, the results of rule-based and similarity-based reasoning at the two levels are combined.

Because of the interaction between the two levels, the architecture is successful in producing, in a massively parallel manner, a number of important patterns of common-sense human reasoning: for example, evidential rule application, similarity matching, mixed rule application and similarity matching, and both top-down and bottom-up inheritance (Sun, 1995).

COMPLETELY INTEGRATED ARCHITECTURES

An even tighter coupling between symbolic and connectionist processes exists in ACT-R (Anderson and Lebiere, 1998). ACT-R consists of a number of symbolic components, including declarative memory (a set of structured chunks), procedural memory (a set of production rules), and goal stacks. Retrieval in declarative memory is controlled by activations of chunks, which spread in a connectionist fashion and are affected by the past history of activations, similarity-based generalization, and stochasticity. Learning of associations among chunks and selection of procedural knowledge also happen in a connectionist fashion. Thus, the learning and the use of symbolic knowledge are partially controlled by connectionist processes. Through this tight integration of the two types of process, ACT-R has been successful in modeling human learning in areas such as arithmetic, analogy, scientific discovery, and human–computer interaction.

LOOSELY COUPLED ARCHITECTURES

Loosely coupled multi-module architectures, unlike the tightly coupled models discussed above, involve only loose and occasional interaction among components. For example, CLARION (Sun and Peterson, 1998), a model for capturing human skill learning, consists of two levels: a symbolic rule level and a connectionist network level. The two levels work rather independently, but their outcomes are combined in decision-making. The network level consists of back-propagation networks, which work through spreading activation and learn by reinforcement. The rule level works according to symbolic rules, which are learned by extracting information from the network level. Through the loose, outcome-based interaction of the two types of processes, the system is able to model a variety of types of human skill learning.

LOCALIST IMPLEMENTATIONS OF RULE-BASED REASONING

Among single-module or homogeneous multi-module models, localist implementations of symbolic processes, especially rule-based reasoning, stand out as an interesting compromise between connectionist networks and purely symbolic models. The representational techniques described below are shared by a number of localist models of rule-based reasoning (see, e.g. Lange and Dyer, 1989; Sun, 1992 and Shastri and Ajjanagadde, 1993).

The simplest way of mapping the structure of a rule set into that of a connectionist network is by associating each concept in the rule set with an individual node in the network, and implementing a rule by connecting each node representing a concept in the condition of the rule to each node representing a concept in the conclusion of the rule. The weights and activation functions can be set to carry out binary logic or fuzzy evidential reasoning.

To express relations, especially relations between large numbers of variables, we need to introduce variables into rules in connectionist implementations. We can represent each variable in a rule as a separate node. We assign values to these variable nodes dynamically and pass values from one variable node to another, based on links that represent variable binding constraints. Such values can be simple numerical signs (Lange and Dyer, 1989; Sun, 1992) or activation phases (Shastri and Ajjanagadde, 1993).

For example, in first-order predicate logic, each argument of a predicate is allocated a node as its representation; a value is assigned to represent an object (i.e., a constant in first-order logic) and thus is a sign of the object. This sign can be propagated from one node to other nodes, when the object that the sign represents is being bound to other variables from an application of a rule.

For each predicate in the rule set, an assembly of nodes is constructed. The assembly contains $k + 1$ nodes if the corresponding predicate contains k arguments. We link up assemblies in accordance

with rules. With this network, we can perform forward-chaining inference. We first activate the assemblies that represent known facts; then activations from these assemblies will propagate to other assemblies to which they are connected. This propagation can continue to further assemblies. For backward chaining, we first try to match the hypothesis with conclusions of existing rules; if a match is found, then we use the conditions of the matching rule as our new hypotheses to be proved: if these new hypotheses can be proved, the original hypothesis is also proved. To implement backward chaining in assemblies, we need, in addition to a predicate node, another node for indicating whether the predicate node is being considered as a hypothesis. Backward flow of activation through hypothesis nodes leads to backward-chaining inference.

Why should we use connectionist models (especially localist ones) for symbolic processing, instead of symbolic models? There are two reasons in particular why researchers explore such models. First, connectionist models are believed to be a more apt framework for capturing many (or even all) cognitive processes (Waltz and Feldman, 1986). The inherent processing characteristics of connectionist models often make them more suitable for cognitive modelling. Second, learning may be more easily incorporated into connectionist models than symbolic models: using, for example, gradient descent and its various approximations, expectation maximization, or the Baun–Welch algorithm. This is especially true of distributed models, but is also true of localist ones to some extent.

DISTRIBUTED IMPLEMENTATIONS OF RULE-BASED REASONING

A stronger notion of integration emphasizes developing symbolic processing capabilities in truly connectionist models, rather than juxtaposing symbolic codes with neural networks, or adopting a compromise as in localist implementations. This approach is more parsimonious explanatorily and thus potentially a more interesting form of cognitive modelling if it can be properly developed. Hence there is considerable interest in symbolic processing capabilities of distributed (or 'true') connectionist models.

An early example is Touretzky and Hinton's (1988) DCPS, which implements a production system in connectionist models. There is a working memory, which stores initially known facts and derived facts; there are two clause components, each of which is used to match one of the two conditions of a rule (each rule is restricted to have two conditions); there is a rule component, which is used to execute the action of a matching rule in the working memory; and a bind component is used to enforce constraints that may exist in a rule regarding variables. Each component is a connectionist network. See Figure 2.

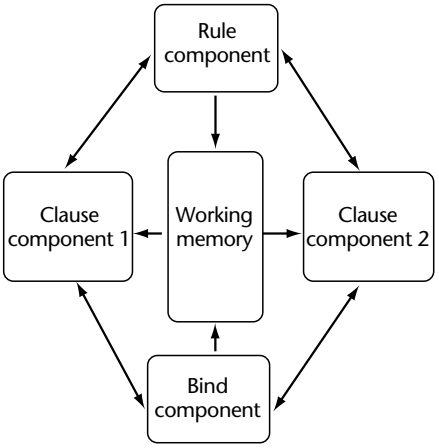


Figure 2. The overall structure of a connectionist production system.

The working memory consists of a large number of nodes, each of which has a randomly assigned 'receptive field'. A *triple* (a fact) is stored in the working memory by activating all the nodes that include the triple in their receptive fields. Many such triples can be stored in the working memory. The two clause components are used to 'pull out' two triples that can match two conditions of a rule. That is, they are used to match triples (in the working memory) with rules (in the rule component). Each node in working memory is connected to a corresponding node in each clause component. A clause component is a kind of 'winner takes all' network.

The rule component is made up of mutually inhibiting clusters. It is also a kind of 'winner takes all' network. Each rule is represented in the rule component by a cluster of identical nodes. The connections from the rule component to the clause components are used to help to pull out the triples that match a rule. In turn, these pulled-out triples also help a particular rule to win in the rule component. After successfully matching a rule with two triples in working memory, actions of the rule are carried out by the gated connections from rule nodes (in the rule component) to nodes in working memory. If the action of the rule includes adding a triple, then the gated connections will excite those nodes in the working memory that

represent the triple; if the action includes deleting a triple, then the gated connections will inhibit those nodes in the working memory that represent the triple to be deleted. Overall, it is a complex system designed specifically to implement a limited production system.

EXTRACTION OF SYMBOLIC KNOWLEDGE FROM CONNECTIONIST MODELS

Many hybrid models involve extracting symbolic knowledge, especially rules, from trained connectionist networks. For example, some researchers proposed a search-based algorithm to extract conjunctive rules from networks trained with backpropagation (see Fu, 1989 and Wermter and Sun, 2000). To find rules, the algorithm first searches for all the combinations of positive conditions that can lead to a conclusion; then, with a given combination of such positive conditions, the algorithm searches for negative conditions that should be added to guarantee the conclusion. In the case of three-layered networks, the algorithm can extract two separate sets of rules, one for each layer, and then integrate them by substitution. Other researchers (e.g. Towell and Shavlik, 1993) tried rules of an alternative form, the 'N of M' form: 'If N of the M conditions $a_1, a_2, \ldots a_M$ are true, then the conclusion b is true.' (It is believed that some rules can be better expressed in such a form, which more closely resembles the weighted-sum computation in connectionist networks, in order to avoid the combinatorial explosion and to discern structures.) A four-step procedure is used to extract such rules, by first grouping similarly weighted links and eliminating insignificant groups, and then forming rules with the remaining groups.

These early rule extraction algorithms are meant to be applied at the end of the training of a network. Once extracted, the rules are fixed; there is no modification 'on the fly', unless the rules are completely extracted again after further training of the network. In some more recent systems, rules can be extracted and modified dynamically. Connectionist learning and rule learning can work together, simultaneously. Thus the synergy of the two processes may be utilized to improve learning (Sun and Peterson, 1998). Dynamic modification is also suitable for dealing with changing environments, allowing the addition and removal of rules at any time.

SUMMARY

Overall, we can discern two approaches toward incorporating symbolic processing capabilities in connectionist models: combining symbolic and connectionist models; and using connectionist models for symbolic processing. In the first approach, the representation and learning techniques from both symbolic processing and neural network models are used to tackle complex problems, including modeling cognition, which involves modeling a variety of cognitive capacities. The second approach is based on the belief that one can perform complex symbolic processing using neural networks alone, with, for example, tensor products, RAAM, or holographic models (see Wermter and Sun, 2000). We may call the first approach 'hybrid connectionism' and the second 'connectionist implementationalism'.

Despite the differences between them, both approaches strive to develop architectures that bring together symbolic and connectionist processes, to achieve a synthesis and synergy of the two paradigms. Many researchers in this area share the belief that connectionist and symbolic methods can be usefully combined and integrated, and that such integration may lead to significant advances in our understanding of cognition.

References

Anderson J and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Dreyfus H and Dreyfus S (1987) *Mind Over Machine*. New York, NY: The Free Press.

Fu L (1989) Integration of neural heuristics into knowledge-based inferences. *Connection Science* **1**(3): 240–325.

Lange T and Dyer M (1989) High-level inferencing in a connectionist network. *Connection Science* **1**: 181–217.

Pollack J (1990) Recursive distributed representation. *Artificial Intelligence* **46**(1,2): 77–106.

Shastri L and Ajjanagadde V (1993) From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences* **16**(3): 417–494.

Smolensky P (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**(1): 1–74.

Sun R (1992) On variable binding in connectionist networks. *Connection Science* **4**(2): 93–124.

Sun R (1995) Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence* **75**(2): 241–295.

Sun R and Peterson T (1998) Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks* **9**(6): 1217–1234.

Touretzky D and Hinton G (1988) A distributed connectionist production system. *Cognitive Science* **12**: 423–466.
Towell G and Shavlik J (1993) Extracting rules from knowledge-based neural networks. *Machine Learning* **13**(1): 71–101.
Waltz D and Feldman J (eds) (1986) *Connectionist Models and Their Implications*. Norwood, NJ: Ablex.
Wermter S and Sun R (eds) (2000) *Hybrid Neural Systems*. Heidelberg: Springer.

Further Reading

Barnden JA and Pollack JB (eds) (1991) *Advances in Connectionist and Neural Computation Theory*. Hillsdale, NJ: Erlbaum.
Giles L and Gori M (1998) *Adaptive Processing of Sequences and Data Structures*. New York, NY: Springer.
Medsker L (1994) *Hybrid Neural Networks and Expert Systems*. Boston, MA: Kluwer.
Sun R (1994) *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. New York, NY: Wiley.
Sun R and Alexandre F (eds) (1997) *Connectionist Symbolic Integration*. Hillsdale, NJ: Erlbaum.
Sun R and Bookman L (eds) (1994) *Architectures Incorporating Neural and Symbolic Processes*. Boston, MA: Kluwer.
Wermter S, Riloff E and Scheler E (eds) (1996) *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*. Berlin: Springer.

Constraint Satisfaction

Intermediate article

Rina Dechter, University of California, Irvine, California, USA
Francesca Rossi, Università di Padova, Padova, Italy

CONTENTS

Constraints are a formalism for the representation of declarative knowledge that allows for a compact and expressive modeling of many real-life problems. Constraint satisfaction and propagation tools, as well as constraint programming languages, are successfully used to model, solve, and reason about many classes of problems, such as design, diagnosis, scheduling, spatio-temporal reasoning, resource allocation, configuration, network optimization, and graphical interfaces.

INTRODUCTION

A constraint satisfaction problem (CSP) consists of a finite set of variables, each associated with a domain of values, and a set of constraints. Each constraint is a relation, defined on some subset of the variables, called its scope, specifying their legal combinations of values. Constraints may be described by mathematical expressions, or by computable procedures.

A solution is an assignment of a value to each variable from its domain such that all the constraints are satisfied. Typical constraint satisfaction problems are to determine whether a solution exists, to find one or all solutions, and to find an optimal solution relative to a given cost function.

An example of a constraint satisfaction problem is the well-known k-colorability problem. The task is to color, if possible, a given graph with k colors only, in such a way that any two adjacent nodes have different colors. A constraint satisfaction formulation of this problem associates the nodes of the graph with variables; the sets of possible colors are their domains; and the 'not equal' constraints between adjacent nodes are the constraints of the problem.

Another well-known constraint satisfaction problem in logic concerns 'satisfiability', which is the task of finding a truth assignment to propositional variables such that a given set of clauses are satisfied. For example, given the two clauses $(A \vee B \vee \neg C)$, $(\neg A \vee D)$, the assignment of 'false' to A, 'true' to B, 'false' to C and 'false' to D is a satisfying truth value assignment.

The structure of a constraint problem is usually depicted by a 'constraint graph', whose nodes represent the variables. Two nodes are connected if the corresponding variables participate in the same constraint scope. In the k-colorability problem, the graph to be colored is the constraint graph. In the satisfiability problem above, the constraint graph has A connected with D, and A, B and C connected with each other.

Constraint problems have proved successful in modeling many practical tasks, such as scheduling, design, diagnosis, and temporal and spatial reasoning. The reason is that constraints allow for a natural, expressive and declarative formulation of what has to be satisfied, without the need to say how it should be satisfied.

In addition, many cognitive tasks, such as language comprehension, default reasoning and abduction, can be naturally represented as constraint satisfaction problems. Historically, constraints and constraint satisfaction have been used in many cognitive tasks related to vision (Waltz, 1975).

When an observer looks at a two-dimensional representation of a three-dimensional geometrical scene, each line and line intersection can be interpreted in many ways, but the physical world together with the laws of geometry put restrictions. It is natural to model lines as variables, with three possible values, indicating that a line can represent a convex, concave, or boundary line, and to have constraints among the lines which are incident to the same junction, forcing the usual laws of three-dimensional geometry.

This representation provides a natural modeling, and is useful for finding a geometrically plausible three-dimensional interpretation of the two-dimensional scene. In fact, by looking at just one constraint, one can automatically eliminate values from the domains of its variables, if such values do not agree with the constraint. This can trigger a chain reaction in which other constraints are considered (one at a time) and other values are deleted in the domains of other variables, until nothing more can be deleted. At this point, we are often left with just one value for each variable, that is, the only plausible interpretation of the scene.

This (generally incomplete) method is usually called 'arc consistency' and belongs to the class of constraint propagation techniques that we will describe in greater detail below.

In linguistics, there are constraint-based views of a language (Chomsky and Lasnik, 1992), constraints over the logic of typed feature structures and relations which are used to describe principles of phrase grammars (Pollard and Sag, 1994), and constraint-based linguistic theories, which perform constraint-assisted searches for deductive proofs, in the spirit of constraint logic programming languages (see below).

It should be noted, however, that constraints and constraint satisfaction are sometimes understood in different ways by computer scientists on one hand and linguists and cognitive scientists on the other. In fact, in most cognitive science applications, a constraint is usually interpreted as being universally quantified over all its entities (e.g., linguistic signs), while this is not the way constraints are defined and used in computer science.

CONSTRAINT SATISFACTION AS SEARCH

Complexity of Constraint-related Tasks

In general, constraint satisfaction tasks (like finding one or all solutions, or the best solution) are computationally intractable (NP-hard). Roughly, this means that there are likely to be problem instances requiring all the possible variable instantiations to be considered before a solution (or best solution) can be found, and this can take a time exponential in the size of the problem. However, there are some tractable classes of problems that allow for efficient solution algorithms of all the problems in the class. Moreover, even for intractable classes, many techniques exhibit a good performance in practice in the average case.

Techniques for Solving CSPs

The techniques for processing constraint problems can be roughly classified into two categories: search (also called conditioning), and consistency inference (or propagation). However, techniques can be combined, and in practice, constraint processing techniques usually contain aspects of both categories.

These two methods have a cognitive analogy in human problem solving. Conditioning search uses the basic operation of guessing a value of a variable and trying to solve a subproblem with the guess. Inference corresponds instead to thinking and deduction, with a view to simplifying the problem.

Search algorithms traverse the space of partial instantiations, building up a complete instantiation that satisfies all the constraints, or else they determine that the problem is inconsistent. By contrast, consistency inference algorithms reason through equivalent problems: at each step they modify the current problem to make it more explicit without losing any information (that is, maintaining the same set of solutions). Search is either systematic and complete, or stochastic and incomplete. Likewise, consistency inference algorithms may achieve complete solutions (e.g., by variable elimination), or incomplete solutions. The latter are usually called local consistency algorithms because they operate on local portions of the constraint problem.

Backtracking search

The most common algorithm for performing systematic search for a solution of a constraint problem is the so-called backtracking search algorithm. This algorithm traverses the space of partial solutions in a 'depth first' manner, and at each step it extends a partial solution (that is, a variable instantiation of a subset of variables which satisfies all the relevant constraints) by assigning a value to one more variable. When a variable is encountered none of whose possible values are consistent with the current partial solution (a situation referred to as a dead end), backtracking takes place, and the algorithm reconsiders one of the previous assignments. The best case occurs when the algorithm is able to successfully assign a value to each variable without encountering any dead ends. In this case, the time complexity is linear in the size of the problem (often identified with the number of its variables). In the worst case, the time complexity of this algorithm is exponential in the size of the problem. However, even in this case the algorithm requires only linear space.

Look-ahead schemes

Several improvements to backtracking have focused on one or both of the two phases of the algorithm: moving forward to a new variable ('look-ahead' schemes) and backtracking to a previous assignment ('look-back' schemes) (Dechter and Pearl, 1987). When moving forward to extend a partial solution, some computation (e.g., arc consistency) may be carried out to decide which variable, and which of the variable's values, to choose next in order to either fail quickly if there is no solution, or not fail at all if there is one. Variables that maximally constrain the rest of the search space are usually preferred, and therefore, the most constrained variable is selected. This method follows the so-called 'first fail' heuristics, which aim to force failures as early as possible, in order to cut down on the amount of backtracking. By contrast, for value selection, the least constraining value is preferred, in order to maximize future options for instantiations (Haralick and Elliot, 1980). A well-known look-ahead method is 'forward checking', which performs a limited form of consistency inference at each step, ruling out some values that would lead to a dead end. A currently popular form of look-ahead scheme, called MAC (for 'maintaining arc consistency'), performs arc consistency at each step and uses the revealed information for variable and value selection (Gaschnig, 1979).

Look-back schemes

Look-back schemes are invoked when the algorithm encounters a dead end. These schemes perform two functions. The first is to decide how far to backtrack, by analyzing the reasons for the current dead end, a process often referred to as 'backjumping' (Gaschnig, 1979). The second is to record the reasons for the dead end in the form of new constraints so that the same conflict will not arise again, a process known as 'constraint learning' and 'no-good recording' (Stallman and Sussman, 1977).

Local search

Stochastic local search strategies were introduced in the 1990s and are popular especially for solving propositional satisfiability problems. These methods move in a hill-climbing manner in the space of all variables' instantiations, and at each step they improve the current instantiation by changing the value of a variable so as to maximize the number of constraints satisfied. Such search algorithms are incomplete, since they may get stuck at a local maximum and they are not able to discover that a constraint problem is inconsistent. Nevertheless, when equipped with some heuristics for randomizing the search, or for revising the guiding criterion function (e.g., constraint reweighting), they have been shown to be reasonably successful in solving many large problems that are too hard to be handled by a backtracking search (Selman *et al*., 1992). A well-known local search algorithm for optimization tasks is called 'simulated annealing'.

Evaluation of Algorithms

The theoretical evaluation of constraint satisfaction algorithms is accomplished primarily by worst-case analysis, that is, determining a function of the problem's size that represents an upper bound of the algorithm's performance over all problems of that size. However, the trade-off between constraint inference and search is hardly captured by such analysis. This is because worst-case analysis, by its nature, is very pessimistic, and often does not reflect the actual performance. Thus in most cases an empirical evaluation is also necessary. Normally, an algorithm is evaluated empirically on a set of randomly generated problems, chosen in such a way that they are reasonably hard to solve (this is done by selecting them from the phase transition region (Selman *et al*., 1992)). Several benchmarks, based on real-life applications such as scheduling, are also used.

CONSTRAINT PROPAGATION

Constraint propagation (or local consistency) algorithms (Montanari, 1974; Mackworth, 1977; Freuder, 1982) transform a given constraint problem into an equivalent one which is more explicit, by inferring new constraints which are added to the problem. Therefore, they can make explicit inconsistencies that were implicitly contained in the problem specification. Intuitively, given a constraint problem, a constraint propagation algorithm will make any solution of a small subproblem extensible to some surrounding variables and constraints. These algorithms are interesting because their worst-case time complexity is polynomial in the size of the problem, and they are often very effective in discovering local inconsistencies.

Arc and Path Consistency

The most basic and most popular propagation algorithm, called arc consistency, ensures that any value in the domain of a variable has a legal

match in the domain of any other variable. This means that any solution of a one-variable subproblem is extensible in a consistent manner to any other variable. The time complexity of this algorithm is linear in the size of the problem. Another well-known constraint propagation algorithm is path consistency. This algorithm ensures that any solution of a two-variable subproblem is extensible to any third variable. As would be expected, it is more powerful than arc consistency in discovering and removing inconsistencies. It also requires more time: its time complexity is cubic in the number of variables.

i-Consistency

Arc and path consistency can be generalized to *i*-consistency. In general, *i*-consistency algorithms guarantee that any locally consistent instantiation of $i - 1$ variables is extensible to any i^{th} variable. Thus, arc consistency is just 2-consistency, and path consistency is just 3-consistency. Enforcing *i*-consistency can be accomplished in time and space exponential in *i*: if the constraint problem has n variables, the complexity of achieving *i*-consistency is $O(n^i)$.

Global Consistency

A constraint problem is said to be globally consistent if it is *i*-consistent for every *i*. In this case, a solution can be assembled by assigning values to variables (in any order) without encountering a dead end, that is, in a backtrack-free manner.

Adaptive Consistency as Complete Inference

In practice, global consistency is not necessary to have backtrack-free assignment of values: it is enough to have directional global consistency relative to a given variable ordering. For example, an 'adaptive consistency' algorithm, which is a variable elimination algorithm, enforces global consistency in a given order only, so that every solution can be extracted with no dead ends along this ordering. Another related algorithm, called tree clustering, compiles the given constraint problem into an equivalent tree of subproblems whose respective solutions can be efficiently combined into a complete solution. Adaptive consistency and tree clustering are complete inference algorithms that can take time and space exponential in a parameter of the constraint graph called the 'induced width' (or 'tree width') (Dechter and Pearl, 1987).

Bucket Elimination

Bucket elimination (Dechter, 1999) is a recently proposed framework for variable elimination algorithms which generalizes adaptive consistency to include dynamic programming for optimization tasks, directional resolution for propositional satisfiability, Fourier elimination for linear inequalities, and algorithms for probabilistic inference in Bayesian networks.

Constraint Propagation and Search

Some problems that are computationally too hard for adaptive consistency can be solved by bounding the amount of consistency enforcing (e.g., applying only arc or path consistency) and embedding these constraint propagation algorithms within a search component, as described above. This yields a tradeoff between the effort spent in constraint propagation and that spent on the search, which can be exploited and which is the focus of empirical studies.

TRACTABLE CLASSES

In between search and constraint propagation algorithms, there are the so-called structure-driven algorithms. These techniques emerged from an attempt to topologically characterize classes of constraint problems that are tractable (that is, polynomially solvable). Tractable classes are generally recognized by realizing that enforcing low-level consistency (in polynomial time) guarantees global consistency for some problems.

Graph-based Tractability

The basic constraint graph structure that supports tractability is a tree. This has been observed repeatedly in constraint networks, complexity theory and database theory. In particular, enforcing arc consistency on a tree-structured constraint problem ensures global consistency along some orderings. Most other graph-based techniques can be viewed as transforming a given network into a meta-tree. Among these, we find methods such as tree clustering and adaptive consistency, the cycle cutset scheme, and the biconnected component decomposition. These lead to a general characterization of tractability that uses the notion of induced width (Dechter and Pearl, 1987).

Constraint-based Tractability

Some tractable classes have also been characterized by special properties of the constraints, without

any regard to the topology of the constraint graph. For example, tractable classes of temporal contraints include subsets of the qualitative interval algebra, expressing relationships such as 'time interval A overlaps or precedes time interval B', as well as quantitative binary linear inequalities over the real numbers of the form $X - Y \leq a$ (Meiri *et al.*, 1990). In general, we exploit notions such as tight domains and tight constraints, row-convex constraints (van Beek and Dechter, 1995), implicational and max-ordered constraints, and causal networks. A connection between tractability and algebraic closure has been discovered and intensively investigated in recent years, yielding a nice theory of tractability (Cohen *et al.*, 2000).

SOFT CONSTRAINT SATISFACTION AND CONSTRAINT OPTIMIZATION

Constraint processing tasks include not only problems of constraint satisfaction, but also problems of constraint optimization. Such problems arise when the solutions are not equally preferred. The preferences among solutions can be expressed via a cost function (also called an objective function), and the task is to find the best-cost solution or a reasonable approximation to it.

For example, we may have the constraints $X \leq Y$ and $Y \leq 10$, with the objective function $f = X + Y$, to be maximized. The best solution (unique in this example) is: $X = 10,\ Y = 10$. All other solutions (e.g. $X = 5,\ Y = 6$), although satisfying all the constraints, are less preferred. Cost functions are often specified as a sum of cost components, each defined on a subset of the variables.

Branch and Bound

Adapting the backtracking search algorithm for the task of selecting the most preferred (best cost) solution yields the well-known branch and bound algorithm. Like backtracking, branch and bound traverses the search tree in a 'depth first' manner, pruning not only partial instantiations that are inconsistent, but also those that are estimated to be inferior to the current best solution. At each node, the value of the current partial solution is estimated (by an evaluation function) and compared with the current best solution; if it is inferior, search along the path is terminated. When the evaluation function is accurate, branch and bound prunes substantial portions of the search tree.

Soft Constraints

One way to specify preferences between solutions is to attach a level of importance to each constraint or to each of its tuples. This technique was introduced because constraints in real problems often cannot be described by a set of 'true or false' statements only. Often constraints are associated with features such as preferences, probabilities, costs, and uncertainties. Moreover, many real problems, even when modeled correctly, are overconstrained. Constraints that have varied levels of importance are called soft constraints. There are several frameworks for soft constraints, such as the semi-ring formalism (Bistarelli *et al.*, 1997), whereby each tuple in each constraint has an associated element taken from a partially ordered set (a semi-ring); and the valued constraint formalism, whereby each constraint is associated with an element from a totally ordered set. These formalisms are general enough to model classical constraints, weighted constraints, 'fuzzy' constraints, and overconstrained problems. Current research effort is focused on extending propagation and search techniques to these more general frameworks.

CONSTRAINT PROGRAMMING

The constraint satisfaction model is useful because of its mathematical simplicity on the one hand, and its ability to capture many real-life problems on the other. Yet, to make this framework useful for many real-life applications, advanced tools for modeling and for implementation are necessary. For this reason, constraint systems (providing some built-in propagation and solution algorithms) are usually embedded within a high-level programming environment which assists in the modeling phase and which allows for some control over the solution method.

Logic Programming

Although many programming paradigms have recently been augmented with constraints, the concept of constraint programming is mainly associated with the logic programming (LP) framework (Lloyd, 1993). Logic programming is a declarative programming paradigm whereby a program is seen as a logical theory and has the form of a set of rules (called clauses) which relate the truth value of an atom (the 'head' of the clause) to that of a set of other atoms (the 'body' of the clause). The clause

```
p(X, Y) :- q(X), r(X, Y, Z)
```

says that if atoms `q(X)` and `r(X, Y, Z)` are true, then also atom `p(X, Y)` is true. For example, the clauses

```
reach(X, Y) :- flight(X, Y)

reach(X, Y) :- flight(X, Z), reach(Z, Y)
```

describe the 'reachability' between two cities (`X` and `Y`) via a sequence of direct flights.

Logic programming and search

Executing a logic program means asking for the truth value of a certain predicate, called the goal. For example, the goal `:- p(X, Y)` asks whether there are values for the variables `X` and `Y` such that `p(X, Y)` is true in the given logic program. The answer is found by recursively 'unifying' the current goal with the head of a clause (by finding values for the variables that make the two atoms equal). As with constraint solving, the algorithm that searches for such an answer in LP involves a backtracking search.

Constraint Logic Programming

To use constraints within LP, one just has to treat some of the predicates in a clause as constraints and to replace unification with constraint solving. The resulting programming paradigm is called 'constraint logic programming' (CLP) (Jaffar and Maher, 1994; Marriott and Stuckey, 1998). A typical example of a clause in CLP is

```
p(X, Y) :- X < Y + 1, q(X), r(X, Y, Z)
```

which states that `p(X, Y)` is true if `q(X)` and `r(X, Y, Z)` are true and the value of `X` is smaller than that of `Y + 1`. While the regular predicates are treated as in LP, constraints are manipulated using specialized constraint processing tools. The shift from LP to CLP permits the choice among several constraint domains, yielding an effective scheme that can solve many more classes of real-life problems. Some examples of CLP languages are ECLiPSe (IC-PARC, 1999), CHIP (Dincbas *et al.*, 1988) SICStus Prolog (Carlsson and Widen, 1999), CHRs (Fruhwirth, 1995), and GNU Prolog (Codognet and Diaz, 1996; Diaz, 2000).

Specialized algorithms for CLP

CLP languages are reasonably efficient, due to their use of a collection of specialized solving and propagation algorithms for frequently used constraints and for special variable domain shapes. Global constraints and bounds consistency are two aspects of such techniques that are incorporated into most current CLP languages.

Global constraints

Global constraints are constraints, usually nonbinary, for which there exist specialized and efficient solution methods. An example is the constraint `alldifferent`, which requires all the involved variables to assume different values, and which can be efficiently solved using a bipartite matching algorithm. Global constraints are used to replace a set of other constraints, usually involving fewer variables and belonging to some special class. For example, a set of binary inequality constraints seldom gives rise to useful constraint propagation and thus may require a complete search. This expensive search is avoided by replacing these constraints with a single `alldifferent` constraint that is accompanied by an efficient propogation algorithm. Most current CLP languages are equipped to handle several kinds of global constraints.

Bounds consistency

Bounds consistency is one of the major contributions of CLP to the field of constraint propagation. When the variable domains are sets of integers, they are usually represented, to save space, by intervals. In this way, one can store only their minimum and the maximum elements. However, constraint propagation techniques like arc consistency could destroy this representation by removing elements internal to the intervals. Therefore, CLP languages usually use an approximation of arc consistency, called bounds consistency, which removes an element from a domain only if the resulting domain is still an interval. This technique is used in most CLP languages, since it is both efficient (in time and space) and powerful in terms of propagation.

CONCLUSION

The study of CSPs is interdisciplinary, since it involves ideas and results from several fields, including artificial intelligence (where it began), databases, programming languages, and operations research. While it is not possible to cover all the lines of work related to CSPs, this article has covered the main ideas.

Current investigations include: identification of new tractable classes; studying the relationship

between search and propagation; extending propagation techniques to soft constraints; and developing more flexible and efficient constraint languages.

References

van Beek P and Dechter R (1995) On the minimality and decomposability of row-convex constraint networks. *Journal of the ACM* **42**: 543–561.

Bistarelli S, Montanari U and Rossi F (1997) Semiring-based constraint solving and optimization. *Journal of the ACM* **44**(2): 201–236.

Carlsson M and Widen J (1999) *SICStus Prolog Homepage*. http://www.sics.se/sicstus/.

Chomsky N and Lasnik H (1992) Principles and parameters theory. In: Jacobs J, von Stechow A and Sternefeld W (eds) *Syntax: An International Handbook of Contemporary Research*. Berlin: Walter de Gruyter.

Codognet P and Diaz D (1996) Compiling constraints in clp(FD). In: *Journal of Logic Programming* **27**(3): 185–226.

Cohen D, Jeavons P, Jonsson P and Koubarakis M (2000) Building tractable disjunctive constraints. *Journal of the ACM* **47**(5): 826–853.

Dechter R (1999) Bucket elimination: a unifying framework for reasoning. *Artificial Intelligence* **13**(1–2): 41–85.

Dechter R and Pearl J (1987) Network-based heuristics for constraint satisfaction problems. *Artificial Intelligence* **34**: 1–38.

Diaz D (2000) *The GNU Prolog web site*. http://gnu-prolog.inria.fr/

Dincbas M, van Hentenryck P and Simonis M *et al*. (1998) The constraint logic programming language CHIP. In: *Proc. International Conference on Fifth Generation Computer Systems*. Tokyo: Ohmsha Ltd.

Freuder EC (1982) A sufficient condition for backtrack-free search. *Journal of the ACM* **29**(1): 24–32.

Fruhwirth T (1995) Constraint simplification rules. In: *Constraint Programming: Basics and Trends*. New York, NY: Springer-Verlag.

Gaschnig J (1979) *Performance Measurement and Analysis of Search Algorithms*. PhD thesis, Pittsburgh, PA: Carnegie Mellon University.

Haralick M and Elliot GL (1980) Increasing tree-search efficiency for constraint satisfaction problems. *Artificial Intelligence* **14**: 263–313.

IC-PARC (1999) *The ECLiPSe Constraint Logic Programming System*. http://www.icparc.ic.ac.uk/eclipse/

Jaffar J and Maher MJ (1994) Constraint logic programming: a survey. *Journal of Logic Programming* **19/20**: 503–581.

Mackworth AK (1977) Consistency in networks of relations. *Artificial Intelligence* **8**(1): 99–118.

Marriott K and Stuckey PJ (1998) *Programming with Constraints: An Introduction*. Cambridge, MA: MIT Press.

Meiri I, Dechter R and Pearl J (1990) Temporal constraint networks. *Artificial Intelligence* **49**: 61–95.

Montanari U (1974) Networks of constraints: fundamental properties and applications to picture processing. *Information Science* **7**(66): 95–132.

Pollard C and Sag IA (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Selman B, Levesque H and Mitchell D (1992) A new method for solving hard satisfiability problems. In: *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 440–446. Menlo Park, CA: AAAI Press.

Stallman M and Sussman GJ (1977) Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. *Artificial Intelligence* **9**(2): 135–196.

Waltz DL (1975) Understanding line drawings of scenes with shadows. In: Winston P (ed.) *The Psychology of Computer Vision*. New York, NY: McGraw-Hill.

Further Reading

Arnborg S and Proskourowski A (1989) Linear time algorithms for np-hard problems restricted to partial *k*-trees. *Discrete and Applied Mathematics* **23**: 11–24.

Beldicenau N and Contejean E (1994) Introducing global constraints in CHIP. *Journal of Mathematical and Computer Modeling* **12**: 97–123.

Dechter R (1990) Enhancement schemes for constraint processing: backjumping, learning and cutset decomposition. *Artificial Intelligence* **41**: 273–312.

Dechter R (1992) *Constraint networks*. In: *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 276–285. New York, NY: Wiley.

Golumbic MC and Shamir R (1993) Complexity and algorithms for reasoning about time: a graph-theoretic approach. *Journal of the ACM* **40**: 1108–1133.

Lloyd JW (1993) *Foundations of Logic Programming*. New York, NY: Springer-Verlag.

Mackworth AK (1992) Constraint satisfaction. In: *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 285–293. New York, NY: Wiley.

Maier D (1983) *The theory of relational databases*. Rockville, MD: Computer Science Press.

Tsang E (1993) *Foundation of Constraint Satisfaction*. London: Academic Press.

Wallace M (1996) Practical applications of constraint programming. *Constraints: An International Journal* **1**: 139–164.

Convolution-based Memory Models

Intermediate article

Tony A Plate, Black Mesa Capital, Santa Fe, New Mexico, USA

CONTENTS

Convolution-based memory models are mathematical models of neural storage of complex data structures using distributed representations. Data structures stored range from lists of pairs, through to sequences, trees, and networks.

HOLOGRAPHIC MEMORY: THE BASIC IDEA

Convolution-based memory models (CBMMs) are mathematical models of storage for lists of paired items, and more complex data structures such as those needed to support language and reasoning capabilities. CBMMs use distributed representations in which items are represented as vectors of binary or real numbers (a *pattern*). (*See* **Distributed Representations**)

CBMMs can store information about items arranged in a great variety of relationships, such as lists of paired items, and sequences, networks, and tree structures. All CBMM storage schemes use a convolution operation to associate or *bind* two (or more) patterns together in a memory *trace*, which is also a pattern.

CBMMs are sometimes called *holographic* memory models because the properties and underlying mathematical principles of CBMMs and light holography are very similar. One of the most striking similarities is that both can reconstruct an entire pattern (in a noisy form) in response to a noisy or partial cue. This ability is a consequence of the *distributed* and *equipotential* nature of storage in both holograms and CBMMs: information about each element of an item or region of an image is distributed across the entire storage medium.

RAPIDLY BINDING TOGETHER COMPONENTS OF A MEMORY

CBMMs use two operations for composing patterns: superposition and binding. For patterns of real numbers, superposition is ordinary element-wise addition; for patterns of binary numbers, superposition is element-wise binary-OR. Superposition is useful for forming unstructured collections of items. However, associations or *bindings* between items cannot be represented using superposition alone because of the *binding problem*. (*See* **Binding Problem**; **Distributed Representations**)

CBMMs use *convolution* as a binding operation: convolution binds two patterns together into one. If $\mathbf{x}$ and $\mathbf{y}$ are n-dimensional pattern vectors (subscripted 0 to $n-1$), then the circular convolution of $\mathbf{x}$ and $\mathbf{y}$, written $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$, is also an n-dimensional pattern vector and has elements

$$z_i = \sum_{k=0}^{n-1} x_k y_{(i-k) \bmod n} \qquad (1)$$

Circular convolution can be viewed as a compression of the outer (or tensor) product of the two vectors, where compression is achieved by summing particular elements, as shown in Figure 1. (Other variants of convolution can be viewed as slightly different ways of compressing the outer product.)

A list of paired items can be represented as the superposition of pairs of items bound together by a convolution. For example, a simple way of representing the list of two pairs 'red-square and blue-circle' is as the pattern $(\mathbf{red} \otimes \mathbf{circle}) + (\mathbf{blue} \otimes \mathbf{square})$. This pattern is quite different from the one

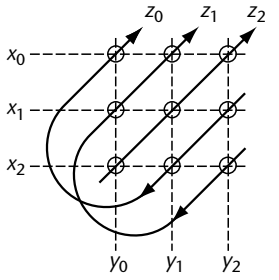


Figure 1. The *circular convolution* $\mathbf{z}$ of vectors $\mathbf{x}$ and $\mathbf{y}$ can be expressed as the sum of elements of their outer product.

that results from a different pairing of the same items such as (**blue** $\otimes$ **circle**) + (**red** $\otimes$ **square**).

In CBMMs, as in many other memory models that use vector or distributed representations, *similarity* is computed by either the *dot product* $\mathbf{x}\cdot\mathbf{y}$, or *cosine* (a scaled version of the dot product) of two pattern vectors:

$$\mathbf{x}\cdot\mathbf{y} = \sum_{i=0}^{n-1} x_i y_i \tag{2}$$

$$\text{cosine}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{n-1} x_i y_i}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=0}^{n-1} x_i y_i}{\sqrt{\sum_{i=0}^{n-1} x_i^2}\sqrt{\sum_{i=0}^{n-1} y_i^2}} \tag{3}$$

One the most important properties of convolution is *similarity preservation*: if patterns **red** and **pink** are similar, then the bindings **red** $\otimes$ **square** and **pink** $\otimes$ **square** will also be similar, to approximately the same degree.

RAPID RETRIEVAL AND INTERFERENCE EFFECTS

Convolution bindings can be easily decoded using inverse convolution operations. For example, using exact inverses, $\mathbf{red}^{-1} \otimes \mathbf{red} \otimes \mathbf{circle} = \mathbf{circle}$. However, the exact inverse can be numerically unstable and is not always the best choice for decoding. For many vectors, such as those whose elements have independent Gaussian statistics with mean zero and variance $1/n$, an approximate inverse can be used. The approximate inverse of $\mathbf{x}$ is denoted by $\mathbf{x}^{\mathrm{T}}$ (this notation is chosen because the approximate inverse is closely related to the matrix transpose). It is a simple rearrangement of the elements of $\mathbf{x}$: $x_i^{\mathrm{T}} = x_{(-i) \bmod n}$. Reconstruction using the approximate inverse is noisy ($\mathbf{red}^{\mathrm{T}} \otimes \mathbf{red} \otimes \mathbf{circle}$ is only approximately equal to **circle**), but is usually more stable in the presence of noise than reconstruction using the exact inverse. If necessary, exact reconstructions can be provided by passing the noisy result through a clean-up memory, which returns the closest matching pattern among the patterns it contains.

Decoding still works when multiple associations are superimposed. For example:

$$\begin{aligned}\mathbf{blue}^{\mathrm{T}} \otimes ((\mathbf{red} \otimes \mathbf{circle}) + (\mathbf{blue} \otimes \mathbf{square})) \\ = (\mathbf{blue}^{\mathrm{T}} \otimes \mathbf{red} \otimes \mathbf{circle}) \\ + (\mathbf{blue}^{\mathrm{T}} \otimes \mathbf{blue} \otimes \mathbf{square}) \\ \approx \mathbf{square}\end{aligned} \tag{4}$$

Because of the randomizing properties of convolution, the first term on the right in the expansion ($\mathbf{blue}^{\mathrm{T}} \otimes \mathbf{red} \otimes \mathbf{circle}$) is not similar to any of **blue**, **red**, **circle**, or **square** and can be regarded as noise. The second term on the right ($\mathbf{blue}^{\mathrm{T}} \otimes \mathbf{blue} \otimes \mathbf{square}$) is a noisy version of **square**. The sum of these two terms is an even noisier, but still recognizable, version of **square**. When larger numbers of bindings are superimposed together the interference effects can become significant, though increasing the vector dimension can reduce interference effects. For further discussion and quantitative analysis, see Murdock (1982), Metcalf-Eich (1982), or Plate (1995).

TODAM

Murdock's (1982) 'theory of distributed associative memory' model (TODAM) is intended to model patterns of human performance on memorization tasks, focusing on tasks involving lists of paired associates. For example, a subject might be asked to memorize the list 'cow-horse, car-truck, dog-cat, and pen-pencil' and then answer such questions as 'Did *car* appear in the list?' (recognition), or 'What was *cat* associated with?' (cued recall). Subjects' relative abilities to perform these and other tasks under different conditions, and the types of errors they produce, give insight into the properties of human memory. Some of the conditions commonly varied are the number of pairs, the familiarity of items, the similarity of items, and the position of recall or recognition targets within the list.

The TODAM formula for sequentially constructing a memory trace for a list of pairs ($\mathbf{x}_i$, $\mathbf{y}_i$) of item patterns is as follows:

$$\mathbf{T}_j = \alpha\mathbf{T}_{j-1} + \gamma_1\mathbf{x}_j + \gamma_2\mathbf{y}_j + \gamma_3\mathbf{x}_j \otimes \mathbf{y}_j \tag{5}$$

where $\mathbf{T}_j$ is the memory trace pattern (a vector) representing pairs 1 through j (with $\mathbf{T}_0 = \mathbf{0}$). The scalars α, γ_1, γ_2, and γ_3 are adjustable parameters of the model, taking values between 0 and 1.

TODAM uses an 'unwrapped' version of convolution which expands the size of vectors each time it is applied, but TODAM could use any convolution operation.

For example, the memory trace for the list of three pairs (**a, b**), (**c, d**), and (**e, f**) is built as follows:

$$\mathbf{T}_1 = \gamma_1\mathbf{a} + \gamma_2\mathbf{b} + \gamma_3\mathbf{a} \otimes \mathbf{b} \tag{6}$$

$$\mathbf{T}_2 = \gamma_1\mathbf{c} + \gamma_2\mathbf{d} + \gamma_3\mathbf{c} \otimes \mathbf{d} + \alpha(\gamma_1\mathbf{a} + \gamma_2\mathbf{b} + \gamma_3\mathbf{a} \otimes \mathbf{b}) \tag{7}$$

$$\mathbf{T}_3 = \gamma_1\mathbf{e} + \gamma_2\mathbf{f} + \gamma_3\mathbf{e} \otimes \mathbf{f} + \alpha(\gamma_1\mathbf{c} + \gamma_2\mathbf{d} + \gamma_3\mathbf{c} \otimes \mathbf{d}) + \alpha^2(\gamma_1\mathbf{a} + \gamma_2\mathbf{b} + \gamma_3\mathbf{a} \otimes \mathbf{b}) \tag{8}$$

Item recognition is done by comparing an item with the trace: item $\mathbf{x}$ was stored in trace $\mathbf{T}$ if $\mathbf{x} \cdot \mathbf{T} > t$ (if the dot product of $\mathbf{x}$ and $\mathbf{T}$ is greater than some threshold t).

Cued recall is accomplished by decoding the trace with the cue: if item $\mathbf{x}$ was stored in trace $\mathbf{T}$, then $\mathbf{x}\#\mathbf{T}$ is a noisy reconstruction of the partner of $\mathbf{x}$ (where $\mathbf{x}\#\mathbf{T}$ is another way of writing $\mathbf{x}^{\mathrm{T}} \otimes \mathbf{T}$).

Some of the predictions of TODAM that are supported by evidence in the psychological literature are as follows:

- Performance decreases with increasing list length.
- Cued recall is symmetric: the recall of $\mathbf{x}$ given $\mathbf{y}$ from a trace containing the pair $\mathbf{x} \otimes \mathbf{y}$ is as accurate as the recall of $\mathbf{y}$ given $\mathbf{x}$ from the same trace.
- There is no primacy effect, only a recency effect, because forgetting is geometric in α.
- Cued recall for a particular item can be superior to recognition for that same item – it can be possible to recall an item that cannot be recognized. This is because weights can be defined so that associative information is stronger than item information.

CHARM

The 'composite holographic associative recall model' (CHARM) (Metcalfe-Eich, 1982) was specifically intended to address the effects of similarity among items in cued recall from lists of paired associates. CHARM uses an even simpler storage method than TODAM – it stores only associative information and no item information. The memory trace for a list of pairs ($\mathbf{x}_i$, $\mathbf{y}_i$) of item patterns is constructed as follows:

$$\mathbf{T} = \sum_{i=1}^{k} \mathbf{x}_i \otimes \mathbf{y}_i \tag{9}$$

CHARM uses a truncated version of the non-wrapped convolution used in TODAM so that the patterns for memory traces are the same size as for items.

As with TODAM, the process for performing cued recall in CHARM begins by correlating a composite memory trace with the cue; e.g., to find the item corresponding to $\mathbf{x}_1$ in $\mathbf{T}$, $\mathbf{x}_1\#\mathbf{T}$ is computed. The resulting pattern will be a noisy version of the pattern associated with $\mathbf{x}_1$ in $\mathbf{T}$, which is passed through a clean-up memory. For the purposes of Metcalfe's experiments, the clean-up memory contained patterns for items stored in the memory trace, and patterns for some other items not stored in the memory trace.

One type of retrieval phenomenon modeled with CHARM is the reduced ability to accurately recall items from a list whose members are similar, versus from a list whose members are dissimilar. For example, performance on a pair such as Napoleon–Aristotle is worse when the pair is embedded in a list of pairs of names of other famous people (a homogenous list) than when it is embedded in a list containing items conceptually unrelated to it, such as red–blue. Furthermore, with homogenous lists, incorrect recall of an item that is similar to the correct response and that was also in the list with an associated item similar to the cue is a frequent type of error in both CHARM and with human subjects.

BINDING VIA FULL TENSOR PRODUCTS

A list of paired items is a very simple set of relationships. Many cognitive tasks demand the ability to store more complicated relationships. For example, understanding language requires the ability to work with recursive structures: a phrase can have a verb, a subject and an object, but the object could be a phrase itself, which could even contain further subphrases. For example, the sentence 'I believe that politicians will say whatever will help them to get elected' contains at least three levels of recursion.

One of the first concrete descriptions of such a scheme was given by Smolensky (1990). Smolensky used tensor products to bind roles and fillers together in a recursive manner. For example, the sentence 'Politicians tell stories' could be represented as the rank-2 tensor $\mathbf{T} = \mathbf{politicians} \otimes \mathbf{tell}_{\mathrm{agent}} + \mathbf{stories} \otimes \mathbf{tell}_{\mathrm{object}}$, where **politicians** is a pattern representing politicians, $\mathbf{tell}_{\mathrm{agent}}$ is a pattern for the agent role of 'tell', etc., and $\otimes$ is the tensor product (a generalization of the outer product). Tensors can be superimposed and decoded in a manner similar to convolution traces; the role pattern $\mathbf{tell}_{\mathrm{agent}}$ can be used to decode the tensor $\mathbf{T}$ to retrieve the pattern **politicians**. What makes the use of tensors interesting is that the rank-2 tensor $\mathbf{T}$ can

be used as the filler in some higher-level role-filler binding, such as representing the meaning of the sentence 'I know politicians tell stories'. This higher-level binding is a rank-3 tensor.

HOLOGRAPHIC REDUCED REPRESENTATIONS

Holographic reduced representations (HRRs) (Plate, 1995, 2000b) use convolution-based role-filler bindings to construct patterns representing a recursive structure.

The HRR for the proposition 'Politicians tell stories' is constructed as follows:

$$\mathbf{P}_{\text{tell}} = \mathbf{tell} + \mathbf{politicians} + \mathbf{stories} + \mathbf{tell}_{\text{agt}} \otimes \mathbf{politicians} + \mathbf{tell}_{\text{obj}} \otimes \mathbf{stories} \quad (10)$$

If we have the pattern $\mathbf{P}_{\text{tell}}$ and know the role patterns, then we can reconstruct a filler pattern by convolving $\mathbf{P}_{\text{tell}}$ with the approximate inverse of a role pattern. For example, $\mathbf{tell}_{\text{agt}}^{\text{T}} \otimes \mathbf{P}_{\text{tell}}$ gives a noisy version of **politicians** which can be put through a clean-up memory to provide an accurate reconstruction.

The HRR pattern $\mathbf{P}_{\text{tell}}$ is a *reduced representation* for the proposition 'Politicians tell stories' and can be used as a filler in a higher-order proposition. For example, the HRR $\mathbf{P}_{\text{know}}$, representing 'Bill knows politicians tell stories', is constructed as follows:

$$\mathbf{P}_{\text{know}} = \mathbf{know} + \mathbf{bill} + \mathbf{P}_{\text{tell}} + \mathbf{know}_{\text{agt}} \otimes \mathbf{bill} + \mathbf{know}_{\text{obj}} \otimes \mathbf{P}_{\text{tell}} \quad (11)$$

Such higher-level HRRs can be decoded in the same way as first-order HRRs. For example, the filler of the know-object role is decoded as follows:

$$\mathbf{P}_{\text{know}} \otimes \mathbf{know}_{\text{obj}}^{\text{T}} \approx \mathbf{P}_{\text{tell}} \quad (12)$$

This reconstructed filler is a proposition. To discover its fillers it could be cleaned up and then decoded again. (*See* **Distributed Representations**)

HRRs are similar if they merely involve similar entities or predicates. Because of the similarity-preserving properties of convolution, they will be even more similar if the entities are involved in similar roles. Thus it turns out that the similarity of HRRs can reflect both superficial and structural similarity in a way that neatly corresponds to the data on human analog retrieval (Plate, 2000).

IMPLEMENTING CONVOLUTION-BASED MEMORIES IN CONNECTIONIST NETWORKS

The various operations used in convolution-based memory models – convolution, correlation, approximate inverse, dot-product, and clean-up memory – are easily implemented in connectionist networks. Convolution encoding and decoding can be implemented by suitably connected networks of 'sigma–pi' neurons. Figure 2 shows a network that computes the circular convolution of two three-element vectors.

$$z_0 = x_0\, y_0 + x_1\, y_2 + x_2\, y_1$$
$$z_1 = x_0\, y_1 + x_1\, y_0 + x_2\, y_2$$
$$z_2 = x_0\, y_2 + x_1\, y_1 + x_2\, y_0$$

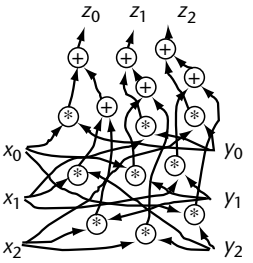


Figure 2. The *circular convolution* **z** of vectors **x** and **y** drawn as a network of three sigma–pi neurons. Each sigma–pi neuron computes the sum of three products as shown on the left.

The pattern of connections in the sigma–pi network that computes circular convolution may seem unrealistically intricate and precise for a biological circuit. However, Plate (2000a) shows that sigma–pi networks that sum random products of pairs of elements from **x** and **y** can also function as encoding and decoding networks with similar properties to convolution.

For computation of similarity, a dot product can be computed by a single sigma–pi neuron. Clean-up memory can be implemented in several ways, such as with Kanerva's (1988) sparse distributed memory, or Baum *et al*.'s (1988) various associative content-addressable memory schemes.

References

Baum EB, Moody J and Wilczek F (1988) Internal representations for associative memory. *Biological Cybernetics* **59**: 217–228.

Kanerva P (1988) *Sparse Distributed Representations*. Cambridge, MA: MIT Press

Metcalfe-Eich J (1982) A composite holographic associative recall model. *Psychological Review* **89**: 627–661.

Murdock BB (1982) A theory for the storage and retrieval of item and associative information. *Psychological Review* **89**(6): 316–338.

Plate TA (2000a) Randomly connected sigma–pi neurons can form associator networks. *Network: Computation in Neural Systems* **11**(4): 321–332.

Plate TA (2000b) Structured operations with vector representations. *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks* **17**(1): 29–40.

Plate TA (1995) Holographic reduced representations. *IEEE Transactions on Neural Networks* **6**(3): 623–641.
Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46**(1–2): 159–216.

Further Reading

Anderson JA (1973) A theory for the recognition of items from short memorized lists. *Psychological Review* **80**(6): 417–438.
Borsellino A and Poggio T (1973) Convolution and correlation algebras. *Kybernetik* **13**: 113–122.
Halford G, Wilson WH and Phillips S (1998) Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences* **21**(6): 803–831.
Kanerva P (1996) Binary spatter-coding of ordered k-tuples. In: von der Malsburg C, von Seelen W, Vorbruggen J and Sendhoff B (eds) *Artificial Neural Networks–ICANN Proceedings*, vol. 1112, pp. 869–873. Berlin, Germany: Springer.
Murdock B (1993) TODAM2: a model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review* **100**(2): 183–203.
Rachkovskij DA and Kussul EM (2001) Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation* **13**(2): 411–452.
Van Gelder TJ (1999) Distributed versus local representation. In: Wilson R and Keil F (eds) *The MIT Encyclopedia of Cognitive Sciences*, pp. 236–238. Cambridge, MA: MIT Press.
Willshaw D (1981) Holography, associative memory, and inductive generalization. In: Hinton GE and Anderson JA (eds) *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Cooperative and Collaborative Learning

Intermediate article

Angela M O'Donnell, Rutgers, The State University of New Jersey, New Jersey, USA

CONTENTS

Forms of cooperative or collaborative learning have been used for centuries and a variety of peer learning techniques have emerged. The underlying premise of these techniques is that learning is enhanced by peer interaction.

THEORETICAL APPROACHES TO COOPERATION AND COLLABORATION

Cooperative/collaborative learning refers to a variety of instructional arrangements that have the common characteristic of students working together to help one another learn. The term 'cooperative learning' is often used to describe particular techniques such as Slavin's *Student Teams Achievement Divisions* or the Johnsons' *Learning Together*. However, cooperative learning is also a process involving collaboration.

'Collaboration' is a term generally used to describe the process of shared learning and understanding. Because of the close relationship of the two terms, they will be used interchangeably in this article. Cooperative and collaborative learning techniques have been used for instructional purposes for centuries and constitute some of the oldest forms of instruction. The most recent meta-analysis of cooperative learning studies included five hundred studies (Johnson and Johnson, 1989) and provides strong support for the positive benefits of working with peers, including positive effects on achievement, self-confidence, peer relationships, and the inclusion of children with special needs. Although a great deal of work has been conducted since 1989, this meta-analysis remains the most comprehensive integration of the literature.

A full discussion of the various forms of instructional uses of cooperation and collaboration is beyond the scope of this chapter. These approaches to understanding the benefits or other effects of cooperative and collaborative learning differ in whether they emphasize the individual (as in peer tutoring), the group (as in distributed cognition), or the reciprocal influence of the individual and the group (as in problem-based learning). The predominant approach to understanding the effects of collaborative or cooperative learning has been in terms of a focus on individual achievement. Other approaches that emphasize learning communities and apprenticeship models emphasize the group culture and the emergence of group cognition. A number of different approaches to understanding collaborative/cooperative learning are illustrated here. Specific cooperative and collaborative learning techniques may involve elements from more than one perspective.

SOCIAL PSYCHOLOGICAL APPROACHES

A variety of theories can be used to account for the benefits associated with cooperative/collaborative learning. Most of the published research on cooperative learning is influenced by social-motivational theory originating with the work of Morton Deutsch.

According to Deutsch, cooperation is one form of interdependence in which individuals' outcomes are linked. One person cannot succeed in a cooperative group unless all participants succeed. From this perspective, group learning is expected to be more productive than individual learning as a result of interdependence among group members that increases motivation within the group. The key

mechanism by which cooperation can lead to successful outcomes is through motivation.

Interdependence can be created by providing group rewards (e.g., Slavin's *Student Teams' Achievement Divisions* (STAD)) or by developing group norms of caring and mutual helping (e.g., the Johnsons' *Learning Together*). In STAD, students work together in heterogeneous groups to help one another learn material presented by the teacher. Students' improvement scores on individual tests result in points earned for the team. Team averages are computed and the teams with the highest points are rewarded with some form of recomputed group average. The teams with the highest group points receive tangible recognition for their work in the form of certificates or other rewards valued by the group members. In contrast, cooperative learning techniques that depend on mutual care and concern as the basis for creating interdependence rarely use overt rewards. In these techniques, instructors spent a lot of time teaching students social skills and effective strategies for communicating and supporting one another.

DEVELOPMENTAL PSYCHOLOGICAL APPROACHES

Theories related to the mechanisms underlying effective collaboration can also be found in the developmental psychological literature.

Piaget described cognitive development as occurring through a process of adaptation in which a child's existing knowledge and concepts interacted with the world. Experience with the world provides the opportunity to create conflict with existing conceptual structures and the child will attempt to reduce the experience of conflict. From a Piagetian perspective, collaborative groups provide the possibility for cognitive development because group members may prompt cognitive conflict that can result in disequilibration and subsequent conceptual growth. Disequilibration occurs when the learner recognizes a conflict between new information or experience and existing conceptual structures. The effort to restore harmony to one's cognitive structures results in adaptation to the new information or experience. Conceptual growth may occur through this process, although other processes such as denial can achieve the restoration of equilibrium.

Many educators rely on the possibility that learners will bring different perspectives to a task. Difficulties can arise depending on the composition of a group. In groups in which there is a status hierarchy, students may defer to those with higher status, agreeing readily but experiencing little in the way of cognitive conflict. Without the experience of disequilibrium, conceptual growth is unlikely to occur. Piaget proposed that learning was most likely to occur when peers were mutually influential: that is a student was as likely to influence others or be influenced by them. In creating collaborative groups from a Piagetian standpoint, groups consisting of members who are relatively homogeneous are more desirable than very heterogeneous groups. Piaget did not write much about collaboration among peers and his focus was on the individual child's growth and development. Collaboration among peers provided a context for development.

Another developmental psychologist provides a contrasting perspective. Vygotsky emphasized the crucial role of society and culture in shaping the cognitive skills of children. The community is the venue in which cognitive skills are first observed and subsequently internalized by the developing learner. Learning occurs from a Vygotskian perspective when a more skilled individual supports the performance of less skilled or younger learners. The weaker partner can perform tasks with the support of the more skilled partner that he or she could not do alone. The difference between what the child can accomplish alone and with a skilled partner is called the 'zone of proximal development'. This process is one way in which learners become participants in a community of practice. In other words, this is a process in which learners come to acquire the competencies already available in the community. Unlike Piaget, who viewed mutuality in power and influence as crucial to effective collaboration, Vygotsky seems to require an imbalance in power and expertise. The more expert student or adult scaffolds the learning of the more novice individual.

Peer Tutoring as an Example of Collaborative Learning

Peer tutoring is one of the most enduring forms of instruction and typically involves a more skilled individual teaching a less skilled individual. The relationship between learners is one of inequality with respect to knowledge of the target subject. Vygotskian theory might be drawn upon to explain the benefits of this strategy. Interactions between tutor and tutee have the goal of improving the performance of the tutee and are characterized by efforts to prompt the tutee to higher levels of achievement. Tutoring works – it is a unique form

of individualized instruction that improves student achievement (Cohen *et al*., 1982). However, understanding how tutoring works is not a simple task. Empirical studies of tutoring have not been conducted under the rubric of a common theory that details the processes underlying particular this particular form of instruction.

Effects on tutees and tutors

Cohen *et al*. conducted the only available comprehensive review of the effects of tutoring in 1982. Their meta-analysis provided convincing evidence that peer tutoring is effective. Strong effects on student learning were found in studies of short duration that involved structured tutoring related to lower-level skills, often in mathematics. Very few studies examined affective outcomes from tutoring, although in those studies that did, tutored students generally had positive attitudes. Tutees are not the only beneficiaries of tutoring. In 38 studies included in the 1982 meta-analysis related to tutor achievement, the average effect size for tutor achievement was a moderate one of 0.33.

Few studies examine the joint effects on tutors and tutees. One exception to this is the work of Connie Juel (1996). In this study, underachieving student athletes spent about four hours a week preparing materials for tutorial sessions with at-risk first grade students. Both tutors and tutees made substantial progress during the course of a year. The literacy levels of both groups of students increased.

Research on tutoring in the period since 1982 has not focused on achievement *per se* but has concentrated instead on the processes by which tutors engage in their task of scaffolding the learning of their tutees. Even when tutors were naive or provided wrong answers, tutees still benefited. There was general acceptance of the utility of tutoring as an instructional strategy and a shift towards studies designed to understand the processes involved in tutoring rather than documenting outcomes from such tutors. An important strand of this research concerned comparisons of human and computer tutors.

Human and computer tutors

Human tutors allow students to do much of the work and maintain a sufficient feeling of control, provide sufficient guidance to keep the tutee from being frustrated, monitor students' reasoning and intervene to keep problem solving on track, are flexible in how they interact, and are motivating (Merrill *et al*., 1992). Naive tutors often provide indiscriminate feedback, use unsophisticated strategies, and employ politeness strategies that may interfere with tutoring effectiveness. Despite these inadequacies, even naive human tutors produce achievement gains for tutees. Many possible reasons might be offered to explain such findings. It may be that the one-to-one interaction is motivating and the tutee makes more of an effort to engage the material. Perhaps normal classroom instruction may provide such little benefit to those who end up as tutees that even a poor tutor is better than existing instruction.

Computer tutors are often designed to compare the student's problem-solving steps to those of a domain expert. In other words, the student's performance is compared to an expert model and the steps he or she uses are traced. Computer tutors attend to more of the components of recovering from error than humans and provide more diagnostic information, more accurate feedback, and provide interventions as the learner deviates from an expert model. They have limited flexibility in response to learner errors, although recent innovations in the design of computer tutors have improved this functionality. Recent innovations in the development of computer tutors use what is known about the human tutoring process and include strategies and prompts to student learning that are more typical of human tutors.

Future research

The published research on tutoring provides little analysis of how the subject matter influences the nature of the tutoring. Much of the work on tutoring is done in curricular areas that are highly constrained, such as mathematics or physics. Little is known about the effects of tutoring on more ambiguous domains such as social studies or writing. The kinds of outcome that were included in the 1982 meta-analysis were mostly low-level outcomes. The effect of tutoring on more high-level skills has not received the same kind of attention.

An additional area of potentially fruitful work is the comparison of the effectiveness of expert tutors and tutors who are subject matter experts. Tutors develop models of tutees and direct their efforts based on those models. Expert tutors are likely to have very refined models of tutee differences and have a range of tutorial strategies from which to select. Subject matter experts who are less experienced as tutors will likely bring a very different set of skills to the tutorial process.

INFORMATION-PROCESSING APPROACHES

Learning, from an information-processing perspective, occurs as a result of active processing of meaningful materials and experiences. Cooperative/collaborative learning approaches influenced by information-processing theory focus on using the group context for learning as a way for amplifying the active processing of information. Such approaches are often considered cognitive/elaborative approaches and suggest that the group interaction provides an opportunity for deeper processing of material and restructuring of ideas and understanding. Cognitive/elaborative approaches attempt to maximize the participation of each student in cognitive strategies and tasks.

Scripted Cooperation

Scripted cooperation (see O'Donnell and King, 1999) is an example of a cooperative technique that relies on elaborative processing of information by a cooperating dyad. In this strategy, pairs of students read a portion of a text. One partner then summarizes the information to his or her partner who, in turn, provides feedback on the accuracy and completion of the information summarized. Both partners work together to elaborate on the material read, generating connections to other knowledge they have and developing techniques for making the information more memorable. The students then proceed to the next section of the text in which they switch roles with the summarizer from the first section now acting as the person who detects errors and omissions. The use of this technique has been shown to be very successful. The size of the cooperating unit (pairs of students) promotes active participation by the partners, and the requirement to switch roles provides each student with practice of requisite cognitive skills.

Reciprocal Teaching

Another example of a cooperative technique that is influenced by information-processing theory, and also by Vygotskian theory, is reciprocal teaching.

Reciprocal teaching was designed for use with students who showed a significant disparity between their ability to decode and comprehend text (Palinscar and Herrenkohl, 1999). It involves guided instruction in the use of four strategies to assist text comprehension. The teacher and students take turns in leading discussions about a shared text. Before reading the text, students make predictions about what the text will be about. Second, everyone reads a segment of the text and participants then discuss the text content, asking and answering questions. The third strategy used is summarization. The fourth strategy involves clarifying difficult concepts and, finally, new predictions are made for the next segment of the text. The focus in this technique is on teaching students explicit strategies to aid comprehension that they can use with any new text. Instruction in strategy use proceeds through a dialogue between the teacher and the students. Students participate by elaborating on other students' responses or by commenting on one another's questions. Initially, the teacher is very much in charge, modeling the strategies and monitoring students' use of them. With practice, students assume a more central role in leading the discussions, with the teacher participating as a member of the group. In this role, the teacher can remind students about how to use the strategies. He or she provides a scaffold or support to the students' efforts. Finally, student leaders can conduct discussions about a segment of text without assistance from the teacher. The teacher's support has thus gradually been withdrawn from the group interaction.

The model of collaboration exemplified in reciprocal teaching is strongly influenced by Vygotskian theory. Skills are available in the social world before they can be internalized. The entire transition from the teacher-led discussions and modeling of strategies to the student-led discussions and execution of these strategies typifies the internalization of skill. Learners first observing these skills as performed and modeled by an expert practice the skills with guidance from the expert, and finally internalize them as part of their own cognitive repertoires.

Effects on student achievement

Rosenshine and Meister (1994) included 16 studies (only four were published) in their meta-analysis of studies using reciprocal teaching. Studies were selected if they included experimental and control groups and were considered high-, medium-, or low-quality studies, depending on the quality of implementation of the treatments, student assessment, and the quality of the reciprocal teaching dialogue. Across the 16 studies, the median effect size associated with reciprocal teaching was 0.32 when using standardized tests as the outcome measure, and 0.88 when using experimenter developed assessments. Reciprocal teaching was significantly better than a control comparison in only two of nine studies when standardized tests were

used. When experimenter-developed assessments were used, reciprocal teaching was significantly better in four out of five studies. Based on an analysis of the reading materials provided in a standardized test and those provided in experimenter tests, Rosenshine and Meister concluded that the experimenter-developed texts were easier because they were longer, were almost always organized in a topic-sentence-and-supporting-detail form, and answering the questions required less background and searching of text.

No conclusions could be drawn from this meta-analysis about which strategies involved in reciprocal teaching were most effective in assisting students. Many of the studies included provided explicit instruction in each of the component strategies prior to engaging in reciprocal teaching, while others embedded the instruction in cognitive strategies in the use of reciprocal teaching. Both strategies seem to work.

Although the dialogue that occurs in a reciprocal teaching group is expected to play a key role in the benefits that may be derived from reciprocal teaching, few studies analyze this dialogue or use it to detect what kinds of cognitive strategy students are learning and using. Palinscar and Herrenkohl (1999) suggest that the explicit goal of jointly making sense of text encourages an intersubjective attitude in which learners feel they are part of a community. This notion of community is further emphasized by the need for each participant to take a turn in leading the discussion and contributing to the discussion.

DISTRIBUTED-COGNITION APPROACHES

Another view of collaboration/cooperation can be found in the perspectives drawn from distributed cognition that describes a process by which cooperation or collaboration can be accomplished. The types of collaborative/cooperative learning described previously are primarily concerned with individual cognition and the effect of collaborative interactions on individual achievement. The techniques described can be viewed as examples of distributed cognition in that cognitive activity is distributed within an interactive unit. A simple version of distributed cognition and collaboration is the use of scripted cooperation described previously. In this technique, the cognitive processing of information is divided among members of a dyad. One partner engages in rehearsal (summarizing information), while the other partner provides an elaboration of the information. In this way, cognitive processing is distributed.

Some views of distributed cognition consider the distribution to be an extension of the individual's competency, a person-plus approach. Others view the distribution of cognition within a group as inseparable from the activity and functioning of the group, a social-only perspective. The focus of the 'in-the-head' or person-plus approach is on the cognitive residue that remains from group interaction. The group can serve to distribute cognitive processing (e.g., metacognition) and reduce cognitive load. The role of scribe or recorder that is specifically built into some cooperative techniques may reduce working memory demands on other participants. Other cooperative techniques in which groups of students become expert in subtopics of a major topic and are then responsible to teach this content to other members of their group involve a distribution of expertise. This person-plus approach to distributed cognition is modeled on what is known about individual cognition, and functioning within the group is described in terms of individual cognitive functions.

The social-only perspective on distributed cognition views the group members and their work as indivisible. This perspective has strong influences from cultural-historical psychology that emphasizes the role of culture and history on cognition. Thus, distributed cognition approaches that involve the social-only perspective have much in common with a Vygotskian approach to cooperation/collaboration. Interpreted from this perspective, the activity of the individual is mediated by the tools available (to include both symbols and actual artefacts), the rules of the community, and the accepted divisions of labor. Tools represent and embody the expertise of others and interaction with these tools involves interaction with the wisdom and experience of others. For example, when a doctor conducts a history and physical examination of a patient and uses a checklist of steps to guide this process, the doctor is not simply performing an act of individual cognition. The steps included in the checklist represent the embodied expertise of other doctors whose expertise and wisdom in practice has been distilled to a checklist that can be used by others. Thus, the doctor in using this tool participates in the practices of a community that extend beyond the individual interaction with a particular patient.

The notion of distributed cognition has enormous appeal. We do not wish to fly in an aircraft where the work is considered divisible. There are

products where the individual contributions to the final product should be seamless and invisible. Educational systems, however, focus on individual accomplishment, and viewing collaborative groups without concern for individual achievement is not realistic.

An Educational Example of Distributed Cognition

The examples of peer tutoring and scripted cooperation represent versions of distributed cognition and are used in many schools. There are fewer examples of the social-only perspective. One such example is the computer support for intentional learning environments (CSILE: Scardamalia *et al*., 1994). The goal of this project was to support students' intentional and autonomous learning in a community of learners. An environment was created thought the use of a networked set of computers in which students contributed to a shared database. Students could add graphical or text notes to a shared database. They could add comments to notes in the database. Only the author of a note could edit the note or delete it. Authors were notified when a comment was added to a note they contributed. Students could use CSILE for a range of tasks from very traditional schoolwork in which they rehearsed information or for more open-ended student-initiated tasks. Knowledge was public and could be added to, critiqued, revised, and reformulated by anyone in the class. The task for students was the social construction of knowledge in which individual contributions were less important than the adequacy of the emerging knowledge base. Unlike most cooperative or collaborative learning techniques, the focus was not on students improving their own skills.

In one example of students' use of CSILE, students generated notes and comments on the inheritance of characteristics or why a child might look more like one parent than another. Students developed a complicated database of information about inheritance that connected to both current uses of genetics and previous history of research on genetics. Despite the fact that the focus in CSILE classrooms is on community knowledge, students in such classes do in fact improve their individual skills and perform better on standardized tests than do students in control classrooms (Scardamalia *et al.*, 1994).

Problem-based Learning

Both person-plus and social-only aspects of distributed cognition are present in problem-based learning (PBL). PBL must be distinguished from individual problem-solving even when the problems are authentic. Problem-based learning has three key features: a rich and authentic problem that supports student inquiry, student-centered learning, and learning that is done collaboratively with group members, scaffolded by the availability of a tutor or facilitator.

In medical schools the goals of PBL are to develop clinical reasoning, integrate clinical reasoning and basic biomedical knowledge, and promote self-directed learning. The use of PBL in medical schools involves a group of five to seven students who meet with a facilitator to discuss a problem. The facilitator provides students with a small amount of information about a patient's case. One student assumes the role of scribe, recording the decisions and hypotheses of the group. This record is typically visible to all. The group evaluates and defines various aspects of the case and develops an understanding of the causes of the problem. The students do so by identifying key information, generating and evaluating hypotheses, and formulating learning issues that they deem relevant and in need of further explanation. They conduct research to find information relevant to the identified learning issues and eventually draw conclusions about the patient's problem. The process concludes with reflections upon what was learned. The role of the facilitator is crucial to this process.

Problem-based learning exemplifies a number of aspects of distributed cognition. The knowledge needed for problem solution is distributed among the members of the group, library resources, other experts in the school, and prior records of patient cases. Learning requires connecting to a community of practice and prior experience in pursuit of a solution to a particular problem. In this sense, the work of the group is an example of social-only distributed cognition in that understanding of the patient case requires joint activity. The technique also provides an example of person-plus approaches to distributed cognition in that cognitive load is reduced by the use of a single scribe. Metacognition is partially offloaded to the facilitator who adopts some of the monitoring function typical of individual cognition.

Effects on student achievement

Early evaluations of problem-based learning examined traditional learning outcomes such as performance on the medical board examinations. Students from PBL programs score below those from more traditional curricula, but there is some evidence to suggest they do better on the clinical

portions of the examinations. However, comparisons are confounded by selection problems associated with different kinds of students selecting to participate in different kinds of curricula. Hmelo (1998) compared students from a PBL program with those from a traditional curriculum on measures of problem solving. She concluded that the PBL students were more accurate, constructed better explanations, and were more likely to use hypothesis-driven reasoning strategies than students from the traditional curriculum.

Applications outside medicine

Problem-based learning is being used in contexts other than in medical training. Science departments in many universities in the United States have adopted PBL methods in basic courses at the undergraduate level as part of ongoing efforts to make science meaningful and relevant to large groups of students. Many inquiry-based programs in elementary and secondary schools include elements of PBL. In particular, anchoring knowledge acquisition in authentic problems is common and reflects a shift in theories of learning towards a view of knowledge acquisition that has social interaction as a key feature.

COMMON GROUND AMONG APPROACHES

All of the collaborative/cooperative learning techniques described here (reciprocal teaching, problem-based learning, scripted cooperation, and peer tutoring) involve some type of scaffolded instruction. Despite differences in approaches, most collaborative or cooperative techniques promote student engagement with one another or with tasks such that effective cognitive processing and productive discourse occur. Efforts to scaffold instruction are attempts to promote such processes.

Instruction can be scaffolded by setting up the initial context for learning, specifying the cognitive or other roles that students adopt during collaboration, delineating interaction rules such as taking turns, and monitoring or regulating the interactions. Task structures drive the nature and quality of interaction. Authentic tasks such as those found in problem-based learning scenarios engage student interest and provide meaningful challenges.

Instruction can also be scaffolded by the assignment of cognitive roles within a group. In problem-based learning groups, the roles of facilitator and scribe are important in assisting the group members to focus on the learning issues at hand. The scribe records the decisions of the group and this visible record of group hypotheses and decisions about needed information provides important metacognitive processing opportunities for the group. The external record of discussions allows group members to recheck their choices and monitor progress towards their goals. In many cooperative contexts (e.g., the Johnsons' *Learning Together*), specific cognitive roles are assigned (checker, recorder, question generator, etc.), and the performance of these roles provides a scaffold that moves the group towards goal completion. In some instances such as tutoring, the tutor quite deliberatively scaffolds the tutee's processing as he or she directs attention to features of the task and strategies for accomplishing the task. Depending on the content and pedagogical expertise of the tutor, the scaffolding that is provided will be more or less effective. Reciprocal teaching provides scaffolded instruction through dialogues between the teacher who models the desired cognitive strategies and the students who practice the strategies. Support is gradually removed as students become more expert in using the targeted cognitive strategies. Scaffolded instruction may include strategies for maintaining all group members' involvement in group dialogue. Examples of such strategies include requiring turn taking with specific roles or tasks such that the same individual is not always responsible for the same task. Thus, opportunity to participate is deliberately distributed.

Key features of scaffolded instruction include making the use of cognitive strategies visible and providing opportunity for modeling, practice, and feedback related to the performance of those strategies. Collaborative techniques vary depending on whose strategies are made visible (other students, experts) and how the modeling, practice, and feedback cycles occur. They share common goals of student engagement and enhancement of learning.

References

Cohen PA, Kulik JA and Kulik CC (1982) Educational outcomes of tutoring: a meta-analysis of findings. *American Educational Research Journal* **19**: 237–248.

Hmelo CE (1998) Problem-based learning: effects on the early acquisition of cognitive skill in medicine. *Journal of the Learning Sciences* **7**: 173–208.

Johnson DW and Johnson RT (1989) *Cooperation and competition: theory and research*. Edina, MN: Interaction Book Co.

Juel C (1996) What makes literacy tutoring effective? *Reading Research Quarterly* **31**: 268–289.

Merrill DC, Reiser BJ, Ranney M and Trafton JG (1992) Effective tutoring techniques: a comparison of human tutors and intelligent tutoring systems. *Journal of the Learning Sciences* **2**: 277–305.

O'Donnell AM and King A (1999) (eds) *Cognitive Perspectives on Peer Learning*. Mahwah, NJ: Lawrence Erlbaum.

Palinscar AS and Herrenkohl LR (1999) Designing collaborative contexts: lessons from three research programs. In: O'Donnell AM and King A (eds) *Cognitive Perspectives on Peer Learning*, pp. 151–177. Mahwah, NJ: Lawrence Erlbaum.

Rosenshine B and Meister C (1994) Reciprocal teaching: a review of the research. *Review of Educational Research* **64**: 479–530.

Scardamalia M, Bereiter C and Lamon M (1994) The CSILE project: trying to bring the classroom into the world. In: McGilly K (ed.) *Classroom Lessons: Integrating Cognitive Theory*, pp. 201–228. Cambridge, MA: MIT Press.

Further Reading

Dillenbourg P (1999) (ed.) *Collaborative Learning: Cognitive and Computational Approaches*. Oxford, UK: Elsevier.

Evensen DH and Hmelo CE (2000) (eds) *Problem-Based Learning: A Research Perspective on Learning Interactions*. Mahwah, NJ: Lawrence Erlbaum.

Graesser AC, Person NK and Magliano JP (1995) Collaborative dialogue patterns in naturalistic one-to-one tutoring sessions. *Applied Cognitive Psychology* **9**: 1–28.

Salomon G (1993) (ed.) *Distributed Cognitions: Psychological and Educational Considerations*, New York, NY: Cambridge University Press.

Webb NM and Palinscar AS (1996) Group processes in the classroom. In: Berliner DC and Calfee RC (eds) *Handbook of Educational Psychology*, pp. 841–873. New York, NY: Macmillan.

Distributed Representations

Intermediate article

Tony Plate, Black Mesa Capital, Santa Fe, New Mexico, USA

CONTENTS

Distributed representations are a way of representing information in a pattern of activation over a set of neurons, in which each concept is represented by activation over multiple neurons, and each neuron participates in the representation of multiple concepts.

THE IMPORTANCE OF REPRESENTATION

The way information is represented has a large impact on the type of operations that are easy or practical to perform with it. Researchers working on neural models of cognitive tasks have taken representational issues especially seriously. It is a great challenge to work out how to effectively utilize the potentially vast computational power of the human brain, in the face of the unreliability and slowness of individual neurons. It does seem clear that any neural computational scheme that performs at near-human levels must make use of neural representations that make it possible to perform relatively high-level tasks in just a few 'steps' of neural computation. There is simply not enough time for many steps of computation in the time that people take to act when having a conversation, playing sports, etc.

PROPERTIES OF DISTRIBUTED REPRESENTATIONS

In distributed representations, concepts are represented by patterns of activity over a collection of neurons. This contrasts with local representations, in which each neuron represents a single concept, and each concept is represented by a single neuron. Researchers generally accept that a neural representation with the following two properties is a distributed representation (e.g., Hinton *et al.*, 1986):

- Each concept (e.g., an entity, token, or value) is represented by more than one neuron (by a pattern of neural activity in which more than one neuron is active).
- Each neuron participates in the representation of more than one concept.

Another equivalent property is that in a distributed representation one cannot interpret the meaning of activity on a single neuron in isolation: the meaning of activity on any particular neuron is dependent on the activity in other neurons (Thorpe, 1995).

The distinction between local and distributed representations is not always as clear as it might initially seem (see van Gelder, 1991, for discussion). Is a standard eight-bit binary encoding for numbers between zero and 255 a local or a distributed code? At the level of numbers each 'concept' (number) is represented by multiple 'neurons' (bits), and each neuron participates in representing many concepts. However, this encoding can also be viewed as a local representation of powers of two: 1, 2, 4, 8, etc. This is an example of where a representation is distributed at one level of interpretation but in which individual neurons represent finer-grained features or 'micro-features' in a localist fashion.

Similarity and Generalization

The similarity of two patterns is very important in distributed representations. Similarity of two patterns is usually computed as their dot-product or cosine. The cosine is preferable when total neural activity differs significantly across patterns.

Neural networks typically respond in a similar manner to similar inputs. Distributed representations are generally designed to take advantage of this; inputs that should result in similar responses are represented by similar activation patterns, and inputs that should result in different responses are represented by quite different activation patterns. One sees the same principle at work when a

network develops distributed representations for itself, as in Hinton's (1986) family-tree learning network. Note that domain similarity can depend upon the task to be performed: two inputs that can be treated as the same for one task may need to be treated as different for another task. At the most basic level, learning is about determining which similarities and differences between inputs are and are not important for a particular task. A network that is either provided with, or that can learn, a good set of features to represent its input will often be able to generalize well from limited data.

Superposition, Multiple Concepts, Interference, and Ghosting

Using distributed representations, multiple concepts can be represented at the same time on the same set of neurons by superimposing their patterns together. Mathematically, superposition means some sort of addition, possibly followed by a thresholding or a normalizing operation.

How can we tell whether a particular pattern $\mathbf{x}$ is part of a superposition of patterns, denoted by $\mathbf{y}$? The simplest, and most commonly used way is to check whether the similarity between $\mathbf{x}$ and $\mathbf{y}$, exceeds some predetermined threshold. Another technique is discussed later in this article.

As more patterns are superimposed together, it can become difficult to tell whether or not a particular pattern is part of the superposition. When patterns appear in the result of a superposition, but were in fact not part of the superposition, this is known as interference or ghosting. The number of patterns that can be superimposed before ghosting becomes a problem depends on several aspects of the representation: the number of neurons (more neurons means less ghosting); the number of distinct patterns (more patterns means more ghosting); the degree of noise tolerance we wish the representation to have (higher noise tolerance means more ghosting); and the density of patterns.

Density and Sparse Representations

The total level of activity in a pattern is referred to as the sparseness, or alternatively, the density, of the representation. For a binary representation, this is the fraction of neurons that are active. For a continuous representation the Euclidean length of vectors is often used as the density. Patterns chosen to represent concepts in a model are often restricted to have the same density. Density of patterns in a representation can be tuned to optimize the properties of the representation, for example minimizing ghosting and maximizing capacity.

Sparse binary distributed representations have the attractive property that they can be superimposed using the binary-OR rule with little ghosting until the number of superimposed patterns becomes large. Sparse representations are also of interest because neurophysiological evidence suggests that representations used in the brain are quite sparse.

Advantages of Distributed Representations

Distributed representations are often held to have many advantages compared to symbolic (e.g., Lisp data structures) and local representations. The most common and important are as follows:

1. *Representational efficiency*. Distributed representations form a more efficient code than localist representations, provided that only a few concepts are to be represented at once. A localist representation using n neurons can represent just n different entities. A distributed representation using n binary neurons can represent up to 2^n different entities (using all possible patterns of zeros and ones).
2. *Mapping efficiency*. A microfeature-based distributed representation often allows a simple mapping (that uses few connections or weights) to solve a task. For example, suppose we wish to classify 100 different colored shapes as to whether or not they are yellow. Using a localist representation, this would require a connection from each neuron representing a yellow shape to the output neuron. Using a feature-based distributed representation, all that is required is a single connection from the neuron encoding 'yellow' to the output neuron. In general, a mapping that can be encoded in relatively few weights operating on a feature-based distributed representation can be learned from a relatively small training sample and will generalize correctly to examples not in the training set.
3. *Continuity* (in the mathematical sense). Representing concepts in continuous vector spaces allows powerful gradient-based learning techniques such as back propagation to be applied to many problems, including ones that might otherwise be seen as discrete symbolic problems.
4. *Soft capacity limits and graceful degradation*. Distributed representations typically have soft limits on how many concepts can be represented simultaneously before ghosting or interference becomes a serious problem. Also, the performance of neural networks using distributed representations tends to degrade gracefully in response to damage to the network or noise added to activations. Many researchers find these properties compellingly similar to performance observed in people.

On the other hand, local representations are far simpler to understand, implement, interpret, and work with. If distributed representations do not provide significant advantages for a particular application, it may be more appropriate to use a local representation. Page (2000) argues forcefully that this is the case in many cognitive modeling applications.

Major Areas of Research in Distributed Representations

The major areas of research in distributed representations are: (1) techniques for representing data more complex than simple tokens, such as data with compositional structure, continuous data, probability distributions; (2) properties of representational schemes, such as capacity, scaling, reconstruction accuracy; and (3) techniques for learning distributed representations.

CODING PROBLEMS AND TECHNIQUES

The Binding Problem

The problem of keeping track of which features or components belong to which objects is known as the 'binding problem'. Consider trying to represent the presence of several colored shapes. A simple local representation could have one set of features for color and another set of units for shape. A red circle would be represented by activity on the red unit and on the circle unit (Figure 1(a)). However, when we try to represent two colored objects simultaneously, we encounter a binding problem: does red + blue + circle + triangle mean red circle and blue triangle or blue circle and red triangle (Figure 1(b))? The binding problem can arise with local or distributed representations when features or components are represented independently and can belong to different objects.

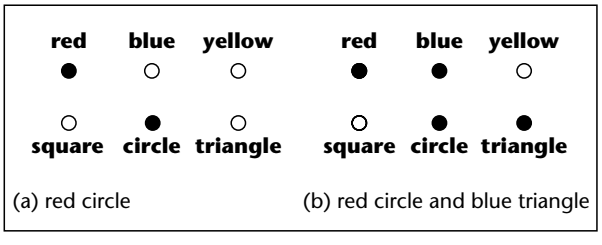


Figure 1. The binding problem: representations of multiple objects on independent feature sets can lose information about which features belong to which objects.

Conjunctive Coding

Conjunctive codes are a general approach to solving the binding problem in neural representations. A simple local conjunctive code has one neuron for every possible combination (conjunction) of a single value from each feature set. For example, to represent colored shapes with 10 possible colors and five possible shapes we would use 50 units. Figure 2 shows an example of this with just three colors and three shapes. This technique can also be used when the value on each feature dimension is encoded with a distributed representation: form the outer product of the patterns for each feature dimension. This outer-product operation underlies such well known associative memory schemes as the Willshaw net (Willshaw, 1989) and the Hopfield net (Hopfield, 1982).

This simple kind of conjunctive code is an example of how a neural code for a composite object can be computed in a systematic manner from the codes for its constituent parts or features. This is an important aspect of Hinton's influential idea of reduced representations, discussed later in this article.

There are two serious problems that occur with simple outer-product conjunctive codes. The first is inefficiency: resource requirements grow exponentially with the number of feature classes. With k feature classes (e.g., color, shape, size, etc.) each having n possible values, the total number of neurons required for a complete conjunctive code is n^k. The informational efficiency of such a code is very poor if only a handful of objects are to be represented simultaneously. Another way of looking at this is that the same set of neurons could represent far more information about each object, or about more objects, if a more efficient code were used. The second problem is that the conjunctive code can hide what makes objects

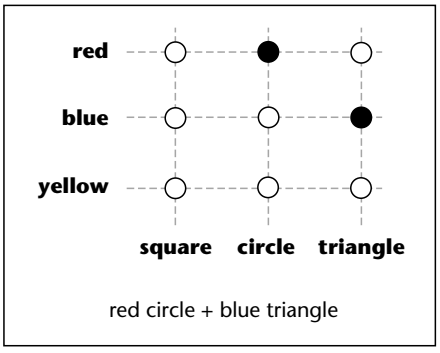


Figure 2. A conjunctive code representing two objects simultaneously.

similar, which places high resource demands on mapping and can make learning difficult. If the independent features of the input are useful for solving the problem or constructing the mapping, learning and mapping are simpler when input to a network represents each feature dimension independently rather than in a conjunctive manner.

Coarse Coding

One of the apparent paradoxes of neural codes is that a stimulus or entity can be represented more accurately by a collection of neurons with broad or coarse response functions than by a collection of neurons with more finely-tuned respone functions. This applies to the representation of both discrete and continuous stimuli or entities.

For representing continuous data, such as a position in space (with two or more dimensions), this means that one can increase the overall accuracy of coding schemes by decreasing the accuracy with which individual neurons represent a data point (i.e., by making the neurons code the data more coarsely). Consider a simple representation for a point in space, where each of n neurons has a randomly chosen centre (in input space), and responds to a data point within a radius r of its centre (its *receptive field*). A two-dimensional version of this is shown in Figure 3. Hinton *et al*. (1986) show that for neurons representing regions in a k-dimensional space, the inaccuracy of a distributed representation like this is proportional to $1/r^{k-1}$. For example, in a six-dimensional space, doubling the radius of each receptive field reduces the inaccuracy of the representation by a factor of 32 on each dimension.

When we look at the system from an information-theoretic viewpoint, it is not surprising that coarsening the receptive fields increases rather than decreases overall accuracy. With very small receptive fields, each neuron is active for only a small fraction of possible stimuli and thus each neuron carries very little information. With larger receptive fields, each neuron is active for a larger fraction of possible stimuli and thus carry more information.

Coarse coding can also be applied to representing discrete entities. The principle is the same: a single neuron is active for a moderately large proportion of possible stimuli, and different neurons cover dissimilar subsets of stimuli. For example, a distributed representation for animals based on semantic features might have one unit responding to animals that are small-to-medium in size AND moderate-to-high in ferocity; another unit might respond to animals that are medium-to-large in size AND moderate-to-high in ferocity, etc.

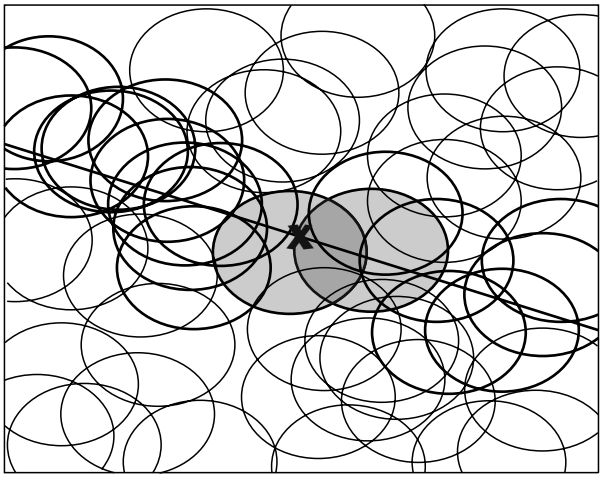


Figure 3. Coarse coding of two-dimensional positions. Each neuron is represented by an area showing its receptive field. Neurons that respond to the point X have gray receptive fields. Neurons that respond to some point along the solid line are shown with bolder receptive fields.

REPRESENTATIONS FOR INFORMATION WITH COMPLEX STRUCTURE

Many cognitive tasks involve understanding and processing information that has a very complex structure. The concepts communicated by human language almost always have a compositional nature, and quite often, the composition is hierarchical. In the sentence 'Joan watched Sam cook the eggs', there are a number of different concepts involved: 'Joan', 'watching', 'Sam', 'cooking', 'eggs'. Representing the overall concept communicated by the above sentence involves representing its components and their relationships. If the relationships are not represented we are not able to distinguish between 'Joan watched Sam cook the eggs' and 'Sam watched Joan cook the eggs'. This is another example of a binding problem; we need to bind the concepts 'Sam' and 'eggs' to the relational concept 'cooking'. One possibility is to allocate a set of neurons for each role or aspect of the relation: a relation name set containing the code for 'cooking', a subject set containing the code for 'Sam', and an object set containing the code for 'eggs'. However, there is also a hierarchical, or recursive, aspect to the

sentence 'Joan watched Sam cook the eggs'. Once we have constructed an appropriate representation for 'Sam cooking eggs' we then need to bind it and 'Joan' to the relational concept 'watching'. An approach utilizing a fixed-size set of neurons for each role prevents any neat solution for representing the recursive aspects of the whole problem because the number of neurons in the representation of a relation is larger than the number of neurons allocated to represent the filler of a role.

The compositional and recursive nature of concepts is not limited to obviously linguistic tasks, but is widespread in human cognition; visual scene understanding, and analogy recall and matching are two examples. Although representing hierarchical compositional structures in the activations and/or weights of a neural network is not easy, it is important as it could eventually allow the application of powerful neural-network learning techniques to difficult problems such as language understanding and acquisition.

Reduced Descriptions

Hinton (1990) introduced the idea of 'reduced description' as a way of representing hierarchical compositional structure in fixed-size distributed representations. The basic idea is that a relation can be represented in two different ways: (1) as an expanded representation in which the fillers of each role are represented on separate groups of neurons (e.g., a relation with two arguments and a relation name could use three groups of n neurons); and (2) as a compressed or reduced description over just n neurons. Since the reduced description for a relation occupies just n neurons, it can be used as a filler in some other relation, which in turn could have a reduced description, and so on, allowing arbitrary levels of nesting. A reduced representation behaves like a pointer in symbolic data structures in that it gives a way of referring to another structure. An important difference is that unlike a pointer, a reduced description should carry some information about its contents that can be accessed without expanding the reduced description into a full description. This allows processing to be sensitive to components nested within reduced descriptions without having to unpack multiple levels of nested relations.

In the 1990s researchers developed a number of concrete neural schemes for implementing reduced descriptions. Many use some form of conjunctive role-filler bindings in which both the roles and fillers of a relation are represented with distributed patterns. Many of these schemes differ mainly in the binding operation they use; it turns out that there are many alternatives to a straight outer-product for forming a conjunctive code of two patterns.

RAAMs

Pollack (1990) used a bottleneck auto-encoder network to learn reduced descriptions for relations, which he called 'recursive auto-associative memory' (RAAM). A full, expanded relation was represented across the input units of the network: each filler on one group of input units, and possibly the relation name on a final group of input units. The hidden layer was the same size as one of the groups of input units and was intended to contain a reduced description of the full relation. The network learned, using back propagation, to compress the full relation down to a reduced description, which could then be expanded back out to a full relation. During learning the network had to discover simultaneously how to create and how to decode reduced descriptions for relations. Pollack showed that a network like this could reliably learn to represent hierarchically structured relations that were several levels deep.

Tensor Product Representations

Smolensky (1990) proposed a tensor-product formalism for representing recursive structure. This extends the outer-product role-filler binding operation to higher dimensions. For example, suppose a role and a filler are each represented by a pattern over a line of n neurons. Then their outer-product binding is a pattern over a square of n^2 neurons. This can be bound with another role vector (again a pattern over n neurons) by forming the tensor product, which in this situation is a pattern over a cube of n^3 neurons. This can be taken to arbitrarily deep levels of nesting. However, the number of neurons required increases exponentially with the depth of conceptual nesting.

Holographic Reduced Representations

Holographic reduced representations (HRRs) (Plate, 1995), are a role-filler binding scheme for recursive compositional structure based on conjunctive coding implemented using circular convolution. In HRRs, roles, fillers, labels, and entire relations are all represented as patterns over n neurons. Circular convolution is defined as $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$, where $z_i = \sum_{k=0}^{n-1} x_k y_{(i-k) \bmod n}$ (where n is the length of the vectors). Circular convolution

is a conjunctive code that keeps dimensionality constant: given role and filler patterns over n neurons each, their circular convolution is also a pattern over n neurons. This makes it simple to build recursive structures. The encoding of a relation is the superposition of role-filler bindings (and possibly a relation label) and thus is also a pattern over n neurons and is easily used as a filler in another relation. For example, a reduced representation for the sentence 'Joan watched Sam cooking eggs' can be constructed as follows:

$$\mathbf{S_1} = \mathbf{cook} + \mathbf{cook}_{\text{agt}} \otimes \mathbf{Sam} + \mathbf{cook}_{\text{obj}} \otimes \mathbf{eggs} \tag{1}$$

$$\mathbf{S_2} = \mathbf{watch} + \mathbf{watch}_{\text{agt}} \otimes \mathbf{Joan} + \mathbf{watch}_{\text{obj}} \otimes \mathbf{S_1} \tag{2}$$

where all the variables are patterns over n neurons (**cook** and **watch** are relation labels; **Sam, eggs**, etc. are patterns representing entities; and $\mathbf{cook}_{\text{agt}}$, etc. are patterns representing the roles of the relations). (*See* **Convolution-based Memory Models**)

A filler of a role in a relation in a reduced representation may be decoded by convolving with the approximate inverse of the role pattern. The approximate inverse of $\mathbf{x}$, denoted by $\mathbf{x}^{\text{T}}$, is a simple permutation of elements: $x_i^{\text{T}} = x_{-i \bmod n}$, and has the property that $\mathbf{x}^{\text{T}} \otimes (\mathbf{x} \otimes \mathbf{y}) \approx \mathbf{y}$.

For example, to recover the filler of the cook-agent role in $\mathbf{S_1}$, we compute $\mathbf{S_1} \otimes \mathbf{cook}_{\text{agt}}^{\text{T}}$, which results in the vector $\mathbf{Sam} + \mathbf{noise}$, because

$$\begin{aligned}
\mathbf{cook}_{\text{agt}}^{\text{T}} \otimes \mathbf{S_1} &= \mathbf{cook}_{\text{agt}}^{\text{T}} \otimes (\mathbf{cook} + \mathbf{cook}_{\text{agt}} \\
&\quad \otimes \mathbf{Sam} + \mathbf{cook}_{\text{obj}} \otimes \mathbf{eggs}) \\
&= \mathbf{cook}_{\text{agt}}^{\text{T}} \otimes \mathbf{cook} + \mathbf{cook}_{\text{agt}}^{\text{T}} \\
&\quad \otimes \mathbf{cook}_{\text{agt}} \otimes \mathbf{Sam} + \mathbf{cook}_{\text{agt}}^{\text{T}} \\
&\quad \otimes \mathbf{cook}_{\text{obj}} \otimes \mathbf{eggs} \\
&= \mathbf{noise} + \mathbf{Sam} + \mathbf{noise}
\end{aligned}$$

In order to recognize $\mathbf{Sam} + \mathbf{noise}$ as the pattern **Sam**, we must pass it through an autoassociative clean-up memory that can take $\mathbf{Sam} + \mathbf{noise}$ as input and return the pattern **Sam**. All potential decoding targets, such as lower-level patterns like **cook,** $\mathbf{cook}_{\text{agt}}$, **Sam**, and higher-level patterns representing chunks, such as $\mathbf{S_1}$ and $\mathbf{S_2}$, must be stored in this long-term clean-up memory. A clean-up memory is also necessary to identify when a decoding target is not present. Without a clean-up memory it would be impossible to tell whether or not there was anything bound to $\mathbf{watch}_{\text{agt}}$ in $\mathbf{S_1}$ because $\mathbf{S_1} \otimes \mathbf{watch}_{\text{agt}}^{\text{T}}$ is a pattern with similar statistical properties to any other pattern. With a clean-up memory containing all potential decoding targets, $\mathbf{S_1} \otimes \mathbf{watch}_{\text{agt}}^{\text{T}}$ can be identified as noise because it almost certainly will not be similar to anything in the clean-up memory.

Binary Spatter Codes

Kanerva's (1996) 'binary spatter code' is a scheme for encoding complex compositional structures in binary distributed representations. Binary spatter codes use binary vectors as patterns, element-wise exclusive-OR for encoding and decoding bindings, and a thresholded sum for superposition. They have similar properties to HRRs.

Holistic Processing

A major reason for interest in distributed representations of a complex structure is the potential for performing structure-sensitive processing without having to unpack hierarchical structures.

Determining similarity is one of the simplest types of processing. Plate (2000) shows that the dot-product of HRRs reflects both superficial similarity (similarity of components) and structural similarity (similarity of structural arrangement of components). The dot-product of HRRs composed of similar entities is higher if the entities are arranged in an analogical (isomorphic) structure. This means that HRRs can be used for fast (but approximate) detection of structural similarity. HRR computations can also be used to rapidly but imperfectly identify corresponding entities in isomorphisms (Plate, 2000; Eliasmith and Thagard, 2001).

Various authors have demonstrated that a variety of structure-sensitive manipulations can be performed on distributed representations without unpacking them. Pollack (1990) trained a feedforward network to transform reduced descriptions for propositions like (LOVED X Y) to ones for (LOVED Y X) where the reduced descriptions were found by a RAAM. Chalmers (1990) trained a feedforward network to transform reduced descriptions of simple passive sentences to reduced descriptions of active sentences, where the reduced descriptions were found by a RAAM. Niklasson and van Gelder (1994) trained a feedforward network to do material conditional inference, and its reverse, on reduced descriptions found by a RAAM. Legendre *et al*. (1991) showed how tensor product representations for active sentences could be transformed to ones for passive sentences (and

vice-versa) by a pre-calculated linear transform. Neumann (2001) trained networks to perform holistic transformations on a variety of representations, including RAAMs, HRRs, and binary spatter codes.

INTERPRETATION OF DISTRIBUTED REPRESENTATIONS

It is straightforward to tell what is represented in a model that uses symbolic or local neural representations. This is not the case with models that use distributed representations. It is important to note that for the purposes of processing the information present in a distributed representation, it is usually NOT necessary to identify what is represented in terms easily understandable to people. Indeed, the whole point of having a distributed representation is that it makes further processing simpler than performing the same computations on a more easily interpretable representation. Interpretation of a distributed representation is typically necessary either when outputs need to be computed in a readily interpretable form or when a person wants to gain insight into the internal workings of a model.

There are two quite different senses in which a distributed representation can be interpreted: (1) determine which items are represented in a pattern of activation, where patterns for individual items are known; and (2) determine what features a network has learned to use to represent a set of items.

Algorithms for Determining Which Items are Represented

A simple and common algorithm for determining the items present in a distributed pattern of activation is to compute the dot-product of each item with the pattern of activation. A feedforward neural network with a localist output representation performs this computation in its final layer of weights. Often it is useful to apply a thresholding or a winner-take-all function to the outputs to cut down the noise.

It is also possible to take an inferential approach to identifying the items present in a distributed code. The idea is to find the best explanation for the observed pattern of activities, in terms of the items that could be present. Zemel *et al*. (1998) do this in analyzing the potential of sparse distributed representations for representing probability distributions. This requires three pieces of knowledge, which can be combined using Bayes' rule: (1) the set of all possible probability distributions that could be represented, and the respective prior probability of each probability distribution (prior to knowledge of current activities); (2) the currently observed pattern of activity; and (3) for each possible probability distribution, the probability that it would generate the currently observed pattern of activity. In Zemel *et al*.'s approach, the probability distribution represented by a pattern of activity is the one with the maximum a-posteriori (MAP) probability (the posterior probability of a distribution is the product of its prior probability and the probability that it would have generated the currently observed pattern of activity).

Understanding Learned Representations

Principal components analysis is often used to gain understanding into a learned distributed representation. Elman (1991) trained a network to predict the next word in sentences generated from a simple recursive English-like language. The predictions made by the network indicated that it had managed to learn something about the recursive nature of the language. Elman used principal components analysis to gain some understanding of how the network represented state (i.e., its memory of what it had seen so far); he showed that a particular phrase caused a similar shape of trajectory through the principal component space independent of its level of embedding, and that level of embedding determined the position of the trajectory.

CONCLUSION

Distributed representations offer powerful representational principles that can be used in neural network approaches to learning and to modeling human cognitive performance. Research continues on three main aspects of distributed representation: schemes for representing data; properties of representational schemes; and ways of learning distributed representations. The ability to learn and use distributed representations of concepts, and to compose complex data structures out of these representations, offers the potential of building general-purpose computers that combine the power of symbolic manipulation with the robustness, learning, and generalization ability of neural networks.

References

Chalmers DJ (1990) Syntactic transformations on distributed representations. *Connection Science* **2**(1–2): 53–62.

Eliasmith E and Thagard P (2001) Integrating structure and meaning: a distributed model of analogical mapping. *Cognitive Science* **25**: 245–286.

Elman JL (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* **7**: 194–220.

van Gelder T (1991) What is the 'D' in 'PDP'? An overview of the concept of distribution. In: Stich S, Rumelhart D and Ramsey W (eds) *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum.

Hinton GE (1986) Learning distributed representations of concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Sciences*, pp. 1–12. Hillsdale, NJ: Lawrence Erlbaum.

Hinton GE (1990) Mapping part–whole hierarchies into connectionist networks. *Artificial Intelligence* **46**: 47–75.

Hinton GE, McClelland JL and Rumelhart DE (1986) Distributed representations. In: Rumelhart DE and McClelland JL and the PDP research group (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, pp. 77–109. Cambridge, MA: MIT Press.

Hopfield J (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA* **79**: 2554–2558.

Kanerva P (1996) Binary spatter-coding of ordered k-tuples. In: von der Malsburg C, von Seelen W, Vorbruggen J and Sendhoff B (eds) *Artificial Neural Networks – ICANN Proceedings*, vol. 1112 of Lecture Notes in Computer Science, pp. 869–873. Berlin, Germany: Springer.

Legendre G, Miyata Y and Smolensky P (1991) Distributed recursive structure processing. In Touretzky DS and Lippman R (eds) *Advances in Neural Information Processing Systems 3*, pp. 591–597. San Mateo, CA: Morgan Kaufmann.

Neumann J (2001) *Holistic Processing of Hierarchical Structures in Connectionist Networks*. PhD thesis, University of Edinburgh. [http://www.cogsci.ed.ac.uk/~jne/holistic_trafo/thesis.pdf.]

Niklasson LF and van Gelder T (1994) Can Connectionist Models Exhibit Non-Classical Structure Sensitivity? *Proceedings of the Sixteenth Annual Conference of The Cognitive Science Society*, pp. 664–669. Hillsdale, NJ: Lawrence Erlbaum.

Page (2000) Connectionist modeling in psychology: a localist manifesto. *Behavioral and Brain Sciences* **23**: 443–512.

Plate TA (1995) Holographic reduced representations. *IEEE Transactions on Neural Networks* **6**(3): 623–641.

Plate TA (2000) Structured operations with vector representations. *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks: Special Issue on Connectionist Symbol Processing* **17**(1): 29–40.

Pollack JB (1990) Recursive distributed representations. *Artificial Intelligence* **46**(1–2): 77–105.

Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46**(1–2): 159–216.

Thorpe S (1995) Localized versus distributed representations. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

Willshaw D (1989) Holography, associative memory, and inductive generalization. In: Hinton GE and Anderson JA (eds) *Parallel Models of Associative Memory* (updated edition), pp. 99–127. Hillsdale, NJ: Lawrence Erlbaum.

Zemel R, Dayan P and Pouget A (1998) Probabilistic Interpretation of Population Codes. *Neural Computation* **10**(2): 403–430.

Further Reading

Baldi P and Hornik K (1989) Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks* **2**: 53–58.

Baum EB, Moody J and Wilczek F (1988) Internal representations for associative memory. *Biological Cybernetics* **59**: 217–228.

Bourlard H and Kamp Y (1988) Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics* **59**: 291–294.

Deerwester S, Dumais ST, Furnas GW, Landauer TK and Harshman R (1990) Indexing by latent semantic analysis. *Journal of the American Society For Information Science* **41**: 391–407.

Halford G, Wilson WH and Phillips S (1998) Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences* **21**(6): 803–831.

Hinton GE, Dayan P, Frey BJ and Neal R (1995) The wake–sleep algorithm for unsupervised neural networks. *Science* **268**: 1158–1161.

Hinton GE and Ghahramani Z (1997) Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London* **352**: 1177–1190.

Hummel JE and Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* **104**(3): 427–466.

LeCun Y, Boser B, Denker JS *et al*. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1**(4): 541–551.

Olshausen BA and Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**: 607–609.

Rachkovskij DA (2001) Representation and processing of structures with binary sparse distributed codes. *IEEE Transactions on Knowledge and Data Engineering* **13**(2): 261–276.

Rumelhart DE and McClelland JL (1986) On learning the past tenses of English verbs. In: McClelland JL,

Rumelhart DE and the PDP research group (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, pp. 216–271. Cambridge, MA: MIT Press.

Zemel RS and Hinton GE (1995) Learning population codes by minimizing description length. *Neural Computation* **7**(3): 549–564.

Dynamical Systems Hypothesis in Cognitive Science

Intermediate article

Robert F Port, Indiana University, Bloomington, Indiana, USA

CONTENTS

The dynamical systems hypothesis in cognitive science identifies various research paradigms applying the mathematics of dynamical systems to understanding cognitive function.

OVERVIEW

The dynamical approach to cognition is allied with and partly inspired by research in neural science since the 1950s, in which dynamical equations have been found to provide excellent models for the behavior of single neurons (Hodgkin and Huxley, 1952). It also takes inspiration from work on gross motor activity by the limbs (e.g. Bernstein, 1967; Fel'dman, 1966). In the early 1950s, Ashby made the startling proposal that all of cognition might be accounted for with dynamical system models (Ashby, 1952), but little work followed directly from his speculation because of a lack of appropriate mathematical methods and computational tools to implement practical models. More recently, the connectionist movement (Rumelhart and McClelland, 1986) has provided insights and mathematical implementations of perception and learning, for example, that have helped revive interest in dynamical modeling.

The dynamical approach to cognition is also closely related to ideas about the embodiment of mind and the environmental situatedness of human cognition, since it emphasizes connections between behavior in neural and cognitive processes, on the one hand, and physiological and environmental events, on the other. The most important such connection is the dimension of time shared by all of these domains. This permits real-time coupling between domains, whereby the dynamics of one system influence the timing of events in another. Humans often couple many systems together: for example, when dancing to music, one's auditory perception system is coupled with environmental sound, and the gross motor system with both audition and musical sounds. Because of this connection between the world, the body and cognition, the method of differential equations is applicable to events at all levels of analysis over a wide range of timescales. This approach emphasizes change over time of relevant system variables. (*See* **Embodiment**; **Dynamical Systems, Philosophical Issues about**)

MATHEMATICAL CONTEXT

The mathematical models employed in dynamical systems research derive from many sources in biology and physics. Two schemas will be discussed out here. The first is the neural network idea, partially inspired by the remarkable equations of Hodgkin and Huxley (1952) which account for many neuronal phenomena in terms of the dynamics of cell membranes. Hodgkin and Huxley proposed a set of differential equations describing the flow of sodium and potassium ions through the axonal membrane during the passage of an action potential down the axon. These equations, which apply with slight modification to all neurons, led to attempts to account for whole cells (rather than patches of membrane) in terms of their likelihood of firing given various excitatory and inhibitory inputs. Interesting circuits of neuron-like units were constructed and simulated on computers. The Hodgkin–Huxley equations inspired many psychological models, such as those of Grossberg (1982, 1986), the connectionist network models (Rumelhart and McClelland, 1986), and models of neural oscillations (Kopell, 1995). (*See* **Hodgkin–Huxley**)

In this schema, it is hypothesized that each cell or cell group in a network follows an equation like

$$\frac{dA}{dT} = -\gamma A(t) + \delta(aE(t) - bI(t) + cS(t)) + k \qquad (1)$$

indicating that the rate of change of activation (i.e., likelihood of firing) A at time t depends on the decay γ of A and a term representing inputs from other cells that are either excitatory, $E(t)$ (tending to increase the likelihood of firing), or inhibitory, $-I(t)$ (tending to decrease the likelihood of firing). For some units there may be an external physical stimulus, $S(t)$. A nonlinear function, $\delta(x)$, encourages all-or-none firing behavior, and the bias term k adjusts the value of the firing threshold. An equation of this general form can describe any neuron. Networks of units like these can exhibit a wide variety of behaviors, including many specific patterns of activity associated with animal nervous systems. (*See* **Neurons, Computation in**)

A second schema inspiring the dynamical approach to cognition is the classical equation for a simple oscillator like a pendulum. Indeed, it is obvious that arms and legs have many of the properties of pendulums. Pendular motion is a reasonable prototype for many limb motions. A similar system (lacking the complication of arc-shaped motion) is that of a mass and spring. It is described by the equation

$$m\frac{d^2x}{dt^2} + d\frac{dx}{dt} + k(x - x_0) = 0 \qquad (2)$$

which specifies simple harmonic motion in terms of the mass m, the damping d, the spring's stiffness k, and the neutral position x_0 of the mass. Fel'dman (1966) used heavily damped harmonic motion to model a simple reach with the arm. If the neutral position x_0 (the attractor position when damped) can be externally set to the intended target position (for example, by adjusting the stiffness in springs representing flexor and extensor muscles), then movements from arbitrary distances and directions towards the target can occur – simply by allowing the neuromuscular system for the arm to settle to its fixed point, x_0. A number of experimental observations – for example, reaching maximum velocity in the middle of the gesture, higher maximum velocity for longer movements, automatic correction for an external perturbation, and the naturalness and ease of oscillatory motions at various rates – can be accounted for with a model using a mass and a spring with controllable stiffness, rest length and damping. (*See* **Motor Control and Learning**)

In the most general terms, a dynamical system may be defined as a set of quantitative variables (e.g., distances, activations, rates of change) that change simultaneously in real time due to influences on each other. These mutual influences can be described by differential or difference equations (van Gelder and Port, 1995). Newton's equations of motion for physical bodies were among the earliest dynamical models. Until the 1950s, the analysis of dynamical models was restricted to linear systems (such as eqns 1 and 2) containing no more than two or three variables. Since the 1970s, mathematical developments, simulations by digital computer programs and computer graphics have revolutionized modeling possibilities, and practical methods for studying nonlinear systems with many variables are now possible (Strogatz, 1994).

PERCEPTUAL MODELS

Dynamical models seem particularly appropriate to account for motor control and for perceptual recognition. In particular, there is a large body of research on temporal aspects of perception. (*See* **Perception: Overview**; **Perceptual Learning**; **Reaction Time**)

One well-known example of a dynamical model for general perception is the adaptive resonance theory (ART) model of Grossberg (1995). This neural network is defined by a series of differential equations, similar to eqn 1 above, describing how the activation of any given node is increased or decreased by stimulus inputs, excitation and inhibition from other nodes, and intrinsic decay. This depends on weights (represented as matrices for a, b and c in eqn 1) which are modified by previous successful perceptual events (simulating learning from experience). The model can discover the low-level features that are most useful for distinguishing frequent patterns in its stimulus environment (using unsupervised learning) and identify specific high-level patterns even from noisy or incomplete inputs. It can also reassign resources whenever a new significant pattern appears in its environment, without forgetting earlier patterns. Notions like 'successful perception' and 'significant pattern' are provided with mathematical specifications that drive the system toward greater 'understanding' of its environment.

To recognize an object such as a letter of the alphabet from visual input, the signal from a spatial retina-like system excites low-level 'feature' nodes. The pattern of activated features here feeds excitation through weighted connections to a higher set of 'identification' nodes. These nodes compete through mutual inhibition to identify the pattern. The best matching unit quickly wins by suppressing all its competitors. When the match is good enough, a 'resonance loop' is established between

some sensory feature units and a particular classification unit. Only at this point is successful (and, according to Grossberg, conscious) identification achieved. This perceptual model is dynamic because it depends on differential equations that increase or decrease the activation of nodes in the network at various rates. Grossberg's research group has shown that variants of this model can account in a fairly natural way for many phenomena of visual perception, including those involving backward masking and reaction time. (*See* **Perceptual Systems: The Visual Model**)

HIGH-LEVEL MODELS

Dynamical models have also been applied to higher-level cognitive phenomena. Grossberg and colleagues have extended the ART model with mechanisms such as 'masking fields', so that, for example, the model can recognize words from temporally arriving auditory input. Several time-sensitive phenomena of speech perception can be successfully modeled in this way (Grossberg, 1986).

Models of human decision-making have traditionally applied the theory of expected utility, whereby evaluation of the relative advantages and disadvantages of each choice is made at a single point in time. But Townsend and Busemeyer (1995) have developed a 'decision field theory' that not only accounts for the likelihood of each eventual choice, but also accounts for many time-varying aspects of decision making, such as 'approach–avoidance' or vacillatory effects, and the fact that some decisions need more time than others. (*See* **Decision-making**; **Choice Selection**)

Some phenomena that at first glance seem to depend on high-level reasoning skills may in fact reflect more low-level properties of cognition. One startling result of this kind is in the 'A-not-B problem'. Infants (9 to 12 months old) will sometimes reach to grab a hidden object; yet when the object is moved to a new location, they often reach to the first location again. This puzzle was interpreted by Piaget (1954) as demonstrating a lack of the concept of 'object permanence', that is, that children have an inadequate understanding of objects, thinking that they somehow intrinsically belong to the place where they are first observed. However, Thelen *et al*. (2001) demonstrated a dynamical model for control of reaching that predicted sensitivity to a variety of temporal variables in a way that is supported by experimental tests. Thus what seems at first to be a property of abstract, high-level, static representations may turn out to result from less abstract time-sensitive processes, which are naturally modeled using dynamical equations. (*See* **Object Concept, Development of**)

RELATION TO SITUATED COGNITION AND CONNECTIONISM

From the perspective of 'situated cognition', the world, the body and the cognitive functions of the brain can all be analyzed using the same conceptual tools. This is important because it greatly simplifies our understanding of the mapping between these domains, and is readily interpreted as an illustration of the biological adaptation of the body and brain to the environment on short-term and long-term time scales. (*See* **Perception, Direct**)

Connectionist models are discrete dynamical systems, as are the learning algorithms used with them. But not all connectionist models study phenomena occurring in continuous time. Neural network models are frequently used to study time-varying phenomena, but other dynamical methods that do not employ connectionist networks are also available. The development of connectionist modeling since the 1980s has helped to move the field in the direction of the dynamical approach, but connectionist models are not always good illustrations of the dynamical hypothesis of cognition. (*See* **Language, Connectionist and Symbolic Representations of**)

CONTRAST WITH TRADITIONAL APPROACHES

The most widespread conceptualization of the mechanism of human cognition proposes that cognition resembles computational processes, like deductive reasoning or long division, by using symbolic representations (of objects and events in the world) that are manipulated by cognitive operations, modeling time only as serial order. These operations reorder or replace symbols, and draw deductions from them (Haugeland, 1985). The computational approach has been articulated as the 'physical symbol system hypothesis' (Newell and Simon, 1976). The theoretical framework of modern linguistics (Chomsky, 1963, 1965 and Chomsky and Halle, 1967) also falls squarely within this tradition since it views sentence generation and interpretation as a serially ordered process of manipulating word-like symbols (such as 'table' or 'go'), abstract syntactic symbols (such as 'noun phrase' or 'sentence') and letter-like symbols representing minimal speech sounds (such as /t/, /a/ or features like 'voiceless' or 'labial') in discrete time. The computational approach,

applied to skills like the perceptual recognition of letters of the alphabet and sounds, or recognizing a person's distinctive gait, or the motor control that produces actions like reaching, walking or pronouncing a word, hypothesizes that essentially all processes of cognition are computational operations that manipulate digital representations in discrete time. The mathematics of such systems is based on the algebra of strings and graphs of symbol tokens. Chomsky's work on the foundations of such abstract algebras (Chomsky, 1963) served as the theoretical foundation for computer science as well as modern linguistic theory. (*See* **Representation, Philosophical Issues about**; **Computation, Philosophical Issues about**; **Symbol Systems**; **Syntax**)

It should be noted that the dynamical systems hypothesis for cognition is in no way incompatible with serially ordered operations on discrete symbols. However, proponents of the dynamical systems approach typically deny that most cognition can be satisfactorily understood in computational terms. They propose that any explanation of human symbolic processing must sooner or later include an account of its implementation in continuous time. The dynamical approach points out the inadequacy of assuming that a 'symbol processing mechanism' is available to human cognition, as a computer happens to be available to a programmer. In the dynamical framework, the discrete time of computational models is replaced with continuous time; first and second time derivatives are meaningful at each instant; and critical time points are specified by the environment or the body rather than by a discrete-time device jumping from one time point to the next. (*See* **Symbol Systems**)

STRENGTHS AND WEAKNESSES OF DYNAMICAL MODELS

Dynamical modeling offers many important advantages over traditional computational cognition. First, the biological implausibility of digital, discrete-time models remains a problem. How and where in the brain might there be a device that would behave like a computer chip, clicking along infallibly performing operations on digital units? One answer that has often been put forward is 'we don't really know how the brain works, anyway, so this hypothesis is as plausible as any other' (Chomsky, 1965 and 2000). Such an answer does not seem as reasonable today as it did in the 1960s. Certainly neurophysiological function exhibits many forms of discreteness; but this fact does not justify the postulation of whatever kind of units and operations would be useful for a digital model of cognition. (*See* **Computation, Philosophical Issues about**; **Categorical Perception**)

Secondly, temporal data can, by means of dynamical models, be incorporated directly into cognitive models. Phenomena such as processing time (e.g., reaction time, recognition time, response time), and temporal structure in motor behavior (e.g., reaching, speech production, locomotion, dance), and in stimulation (e.g., speech and music perception, interpersonal coordination while watching a tennis match), can be linked together if critical, events spanning several domains can be predicted in time. (*See* **Perception: The Ecological Approach**; **Perception, Direct**)

The language of dynamical systems provides a conceptual vocabulary that permits unification of cognitive processes in the brain with physiological processes in our bodily periphery and with environmental events external to the organism. Unification of processes across these fuzzy and partly artificial boundaries makes possible a truly embodied and situated understanding of all types of human behavior. Discrete-time models are always forced to draw a boundary somewhere to separate the discrete-time, digital aspects of cognition from continuous-time physiology (as articulated in Chomsky's, 1965 distinction between 'competence' and 'performance'). (*See* **Neuropsychological Development**; **Performance and Competence**)

Thirdly, cognitive development and 'run-time' processing can now be integrated, since learning and perceptual and motor behavior are governed by similar processes even if on different timescales. Symbolic or computational models were forced to treat learning and development as separate processes unrelated to motor and perceptual activity.

Fourthly, trumping the advantages given above, dynamical models include discrete-time, digital models as a special case. The converse is not true: the sampling of continuous events permits discrete simulation of continuous functions, but the simulation itself remains discrete, and only models a continuous function to an accuracy dependent on its sampling rate (Port *et al*., 1995). Thus, any actual digital computer is also a dynamical system with real voltage values in continuous time that are discretized by an independent clock. Of course, computer scientists prefer not to regard them as continuous-valued dynamical systems (because it is much simpler to exploit their digital properties), but computer engineers have no choice. Hardware engineers have learned to constrain computer dynamics to be governed reliably by powerful

attractors for each binary cell, ensuring that each bit settles into either one of two states before the next clock tick.

These strengths of dynamical modeling are of great importance to our understanding of human and animal cognition. But there are several weaknesses of dynamical modelling. First, the mathematics of dynamical models are more inscrutable and less developed than the mathematics of digital models. It is much more difficult to construct actual models, except for carefully constrained simple cases.

Second, during some cognitive phenomena (for example, performing long division, or designing an algorithm, and possibly some processes in the use of language) humans appear to rely on ordered operations on discrete symbols. Although dynamical models are capable of exhibiting digital behavior, how a neurally plausible model could perform these tasks remains a puzzle. It seems that computational models are simpler and more direct, even if they remain inherently insufficient.

DISCRETE VERSUS CONTINUOUS REPRESENTATIONS

Intuitively one of the major strengths of the traditional computational approach to cognition has been the seeming clarity of the traditional notion of a cognitive representation. Since cognition is conceived as functioning somewhat like a program in LISP, the representations are constructed from parts that resemble LISP 'atoms' and 's-expressions'. (*See* **Symbol Systems**; **Computation, Philosophical Issues about**)

A representation is a distinct data structure that has semantic content (with respect to the world outside or inside the cognitive system). Representations can be moved around or transformed as needed. Such tokens have an undeniable resemblance to words and phrases in natural language (Fodor, 1975). Thus, if one considers making a sandwich from bread and the ham in the refrigerator, one can imagine employing cognitive tokens standing for bread, the refrigerator, and so on. Thinking about sandwich assembly might be cognitively modeled using representations of sandwich components. Similarly, constructing the past tense of 'walk' can be modeled by concatenating the representation of 'walk' with the representation of '-ed'. However, this view runs into difficulties when we try to imagine thinking about actually slicing the bread or spreading the mayonnaise. How could discrete, word-like representations be deployed to yield successful slicing of bread? If this is instead to be handled by a nonrepresentational system (such as a dynamical one), then how could we combine these two seemingly incompatible types of systems? (*See* **Representation, Philosophical Issues about**; **Language of Thought**)

In the 1980s, the development of connectionist models, employing networks of interconnected nodes, provided the first alternative to the view of representations as context-invariant, manipulable tokens. In connectionist models, the result of a process of identification (of, say, an alphabetic character or a human face) is only a temporary pattern of activations across a particular set of nodes (modeling cells or cell groups), not something resembling a context-free object. The possibility of representation in this more flexible form led to the notion of distributed representations, where no apparent 'object' can be found to do the work of representing, but only a particular pattern distributed over a set of nodes that are also used for many other patterns. Connectionists emphasized that such a representation would not seem to be a good candidate for a *symbol token*, as conceived in the formalist or computational tradition, yet can still function as a representation for many of the same purposes.

The development of dynamical models of perception and motor tasks has led to further extension of the notion of the representational function to include time-varying trajectories, limit cycles, coupled limit cycles, and attractors towards which the system state may tend. From the dynamical viewpoint, static, computational representations will play a far more limited role in cognition. Indeed, some researchers have denied that static representations are ever needed for modeling any cognitive behavior (Brooks, 1997).

References

Ashby R (1952) *Design for a Brain*. London, UK: Chapman-Hall.

Bernstein N (1967) *The Control and Regulation of Movements*. London, UK: Pergamon Press.

Brooks R (1997) Intelligence without representation. In: Haugeland J (ed.) *Mind Design II*, pp. 395–420. Cambridge, MA: MIT Press.

Chomsky N (1963) Formal properties of grammars. In: Luce RD, Bush RR and Galanter E (eds) *Handbook of Mathematical Psychology*, vol. II, pp. 323–418. New York, NY: Wiley.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky N (2000) Linguistics and brain science. In: Marantz A, Miyashita Y and O'Neil W (eds) *Image, Language and Brain*, pp. 13–28. Cambridge, MA: MIT Press.

Chomsky N and Halle M (1967) *The Sound Pattern of English*: Harper and Row.

Fel'dman AG (1966) Functional tuning of the nervous system with control of movement or maintenance of a steady posture – III. Mechanographic analysis of the execution by man of the simplest motor tasks. *Biophysics* **11**: 766–775.

Fodor J (1975) *The Language of Thought*. Cambridge, MA: Harvard University Press.

van Gelder T and Port RF (1995) It's about time: overview of the dynamical approach to cognition. In: Port RF and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 1–43. Cambridge, MA: MIT Press.

Grossberg S (1982) Studies of mind and brain: neural principles of learning, perception, development, cognition, and motor control. Norwell, MA: Kluwer.

Grossberg S (1986) The adaptive self-organization of serial order in behavior: speech, language and motor control. In: Schwab NE and Nusbaum H (eds) *Pattern Recognition by Humans and Machines: Speech Perception*, pp. 187–294. Orlando, FL: Academic Press.

Grossberg S (1995) Neural dynamics of motion perception, recognition, learning and spatial cognition. In: Port RF and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 449–490. Cambridge, MA: MIT Press.

Haugeland J (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* **117**: 500–544.

Kopell N (1995) Chains of coupled oscillators. In: Arbib M (ed.) *Handbook of Brain Theory and Neural Networks*, pp. 178–183. Cambridge, MA: MIT Press.

Newell A and Simon H (1976) Computer science and empirical inquiry. *Communications of the ACM* **19**: 113–126.

Piaget J (1954) *The Construction of Reality in the Child*. New York, NY: Basic Books.

Port RF, Cummins F and McAuley JD (1995) Naive time, temporal patterns and human audition. In: Port RF and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 339–371. Cambridge, MA: MIT Press.

Rumelhart D and McChelland J (1986) *Parallel Distributed Processing*, vols 1 and 2. Bradford Books, MIT Press.

Strogatz SH (1994) *Nonlinear Dynamics and Chaos With Applications to Physics, Biology, Chemistry, and Engineering*. Reading, MA: Addison-Wesley.

Thelen E, Schöner G, Scheier C and Smith LB (2001) The dynamics of embodiment: a field theory of infant perseverative reaching. *Behavioral and Brain Sciences* **24**: 1–34.

Townsend J and Busemeyer J (1995) Dynamic representation of decision making. In: Port RF and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 101–120. Cambridge, MA: MIT Press.

Further Reading

Abraham RH and Shaw CD (1982) *Dynamics: The Geometry of Behavior*, vol. I. Santa Cruz, CA: Ariel Press.

Clark A (1997) *Being There: Putting the Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Haugeland J (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Kelso JAS (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Port RF and van Gelder T (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Thelen E and Smith LB (1994) *A Dynamical Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Dynamical Systems: Mathematics

Intermediate article

Jerome R Busemeyer, University of Indiana, Bloomington, Indiana, USA

CONTENTS

Dynamical systems are mathematical models for describing how the state of a biological or an artificial system changes over time.

INTRODUCTION

A critical factor facilitating the 'cognitive revolution' of the early 1960s was the capability of formulating rigorous (computational or mathematical) models of how the inputs and outputs of mental processing systems change over time. The earliest approach was to view the mind as a dynamical cybernetic feedback system (Miller, Gallanter and Pribram, 1960). But this was abandoned in favour of another approach, which was to view the mind as a rule-based symbol processor (Newell and Simon, 1972). Most recently, developments in neural and connectionist networks have revived interest in a dynamical systems approach. More broadly, dynamical systems theory has been adopted by a wide range of fields in cognitive science, including the study of perceptual motor behavior, child development, speech and language, and artificial intelligence. There are excellent presentations of mathematical dynamical systems theory in Beltrami (1987), Luenberger (1979), Padulo and Arbib (1974), and Strogatz (1994).

ELEMENTS OF DYNAMICAL SYSTEMS

Generally speaking, a *dynamical system* is composed of three parts. The first part is the *state*, which is a representation of all the information about the system at a particular moment in time. As an example, the state of a computer can be summarized by a long list of the binary bits of information stored in the registers and memory banks at a given moment. The state of a brain model can be represented as a large-dimensional vector of positive real numbers, representing all the neural activations at a given moment. In general, the notation $X(t) = [x_1(t), \ldots x_n(t)]$ will be used to denote the state of a system at time t.

The second part is the *state space*. This is a set that contains all of the possible states to which a system can be assigned. The state space of a computer is the set of all of the possible configurations for the n-element binary-valued list representing its state (a set of size 2^n). The state space of a brain model is the set of points contained in the positive region of the n-dimensional Cartesian vector space IR^n. The symbol Ω is used to denote the state space of a dynamical system, and $X(t) \in \Omega$.

The third part is the *state transition function*, which is used to update and change the state from one moment to another. For example, the state transition function of a computer is defined by the production rules that change the bits of information from the state at one step $X(t)$ to the state at the next step $X(t+1)$. The state transition function for a brain model is a continuous function of time that maps the brain state $X(t)$ at time t to another state $X(t+h)$ at a later moment. The symbol $\boldsymbol{T}$ is used to denote the state transition function that maps an initial state $X(t)$ into a new state $X(t+h)$:

$$X(t+h) = \boldsymbol{T}(X(t),\, t,\, t+h) \qquad (1)$$

Whenever the state transition function is assumed to be a differentiable function of time, then we can define the *local generator* as the time derivative

$$\frac{\mathrm{d}\boldsymbol{T}}{\mathrm{d}t} = \lim_{h \to 0} \frac{X(t+h) - X(t)}{h} = \boldsymbol{f}(X(t), t) \qquad (2)$$

Given an initial starting state $X(0)$, the local generator is used to generate a *trajectory* $X(t)$ for all $t > 0$. Provided that the local generator satisfies certain

smoothness properties, a local generator is guaranteed to produce a unique trajectory from any given starting state. The objective of dynamical systems analysis is to understand all the possible trajectories produced by a local generator.

EXAMPLES OF DYNAMICAL SYSTEMS

Before describing the analysis of dynamical systems in more detail, it will be helpful to describe some well-studied examples.

Logistic Growth Model

Consider the following simple dynamical model of growth. To be concrete, suppose we wish to model a student's probability of performing a task correctly as a function of training (e.g., successfully playing a piece of classical music on the piano). Let $p(t)$ be the probability of correctly performing the task, and assume that this preference changes from one time point t to a later time point $t+h$ according to the following simple difference equation:

$$\frac{p(t+h)-p(t)}{h} = \alpha p(t) \times (1-p(t)) \tag{3}$$

This model can be understood as a product of two parts. The second part, $1-p(t)$, represents the amount that remains to be learned. The first part, $\alpha p(t)$, can be interpreted as the learning rate, which is an increasing function of the amount already learned. Note that as the probability approaches zero or one, then the change approaches zero; in effect, the probability remains bounded between zero and one.

This difference equation can be reformulated as a differential equation by allowing the time interval h to approach zero in the limit:

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \lim_{h\to 0}\frac{p(t+h)-p(t)}{h} = \alpha p(t)(1-p(t)) \tag{4}$$

By separating the variables and integrating, we can solve this differential equation to yield the solution (Braun, 1975)

$$p(t) = (1+c\mathrm{e}^{-\alpha t})^{-1} \tag{5}$$

where the constant c depends on the initial state, $p(0)$. Figure 1 is a time series plot that illustrates performance over time (with $c=100$ so that $p(0)$ is very close to zero). The model produces a family of smooth S-shaped curves that gradually rise from the initial level and approach the line of stable equilibrium $p=1$. The growth rate, α, determines the steepness of the curve.

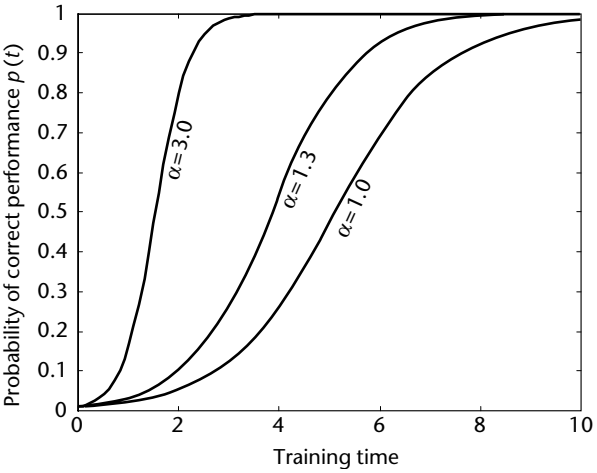


Figure 1. A time series plot showing the trajectories produced by the logistic growth model. The horizontal axis represents training time, and the vertical axis represents the probability of correct task performance.

Linear System Model

Next consider the problem of developing a simple dynamical model of motivation to perform a task (Atkinson and Birch, 1970; Townsend and Busemeyer, 1989). Suppose a student is trying to decide how much effort to invest in a task over time, for example, the amount of time to allocate towards achieving an athletic, academic, or social goal. Let $q(t)$ be the amount of effort actually expended at time t, and let $p(t)$ be the student's strength of motivation for doing the task at time t.

First we assume that the rate of change in effort is directly influenced by the motivation state at a given time. But due to fatigue, the rate of change in effort also decreases in proportion to the current effort. These assumptions are incorporated into the following model:

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = p(t) - \alpha q(t) \tag{6}$$

Secondly, we assume that the rate of change in motivation is determined by two factors. One is the valence difference between the anticipated rewards for success and punishments for failure, denoted by $v(t)$. Another is a satiation effect, which causes the motivation to decrease in proportion to the current effort. The following model is used to represent these assumptions:

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = v(t) - \beta q(t) \tag{7}$$

Given the initial states, $p(0)$ and $q(0)$, and given the valence differences, $v(t)$, $t \geq 0$, eqns 6 and 7 define

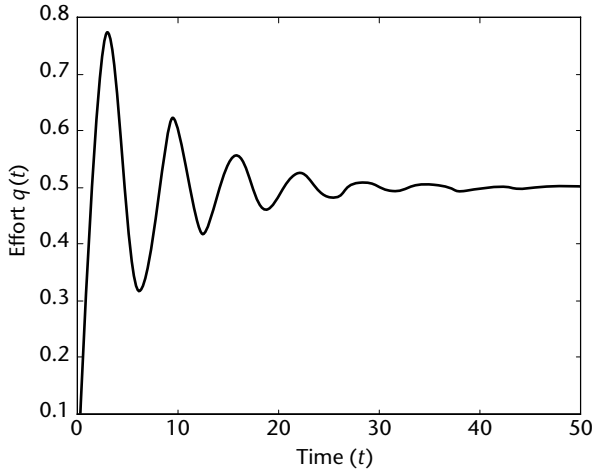


Figure 2. A time series plot showing the trajectory produced by the simple dynamic model of effort. The horizontal axis represents time elapsed on the task, and the vertical axis represents the amount of effort as a function of time.

a simple dynamical system that can be used to model the behavior of a student's effort on a task over time. The valence $v(t)$, is called the *input* or *forcing term* of the system, and the coefficients α and β are called the *parameters* of the system.

Methods for solving *coupled systems* of linear differential equations, such as eqns 6 and 7, are covered in standard texts on ordinary differential equations (see e.g. Braun, 1975). In the special case when $p(0) = q(0) = 0$ and the input valence is constant $(v(t) = v)$, the solution for effort is

$$q(t) = \frac{v}{\beta}\left(1 - e^{-\frac{1}{2}\alpha t}(\sin\theta t + \frac{\alpha}{2}\cos\theta t)\right) \quad (8)$$

where $\theta^2 = \beta - (\frac{\alpha}{2})^2$. (This can be checked by differentiating eqn 8 with respect to t and showing that it satisfies both eqns 6 and 7 and the initial conditions.)

The time series plot in figure 2 illustrates the behavior predicted by the model (with the parameters $\alpha = 0.25$, $\beta = 1.00$ and $v = 0.50$). The horizontal axis shows the time elapsed on the task, and the vertical axis shows the effort spent. The curve oscillates up and down (between zero and one), approaching the asymptotic level of effort $q = 0.50$.

Predator–Prey Model

Suppose a researcher is trying to decide how much effort to spend on generating and testing ideas. Let x_1 be the number of candidate ideas generated for testing, and let x_2 be the number of ideas tested. Note that an idea cannot be tested until it has been generated; and once it has been tested it is eliminated from the candidate pool. Consider the following simple nonlinear model of this process:

$$\frac{dx_1}{dt} = \alpha x_1 - \beta x_1 x_2 - \gamma x_1^2 \quad (9)$$

$$\frac{dx_2}{dt} = \phi x_1 x_2 - \lambda x_2 \quad (10)$$

In eqn 9, the coefficient α allows candidate ideas to grow over time, the coefficient β reflects depletion of candidate ideas that are generated and tested, and the coefficient γ represents interference between ideas. In eqn 10, the coefficient ϕ represents the increase in testing produced by a successful test and the coefficient λ provides for a fatigue effect from testing ideas. All of these coefficients are assumed to be positive.

Eqns 9 and 10 form what is called a predator–prey model. In this example, the idea-testing process is playing the role of the predator, which is preying on the idea-generating process. When dealing with nonlinear differential equations such as this, the most practical method for finding solutions is to use numerical integration routines, which are available in mathematical programming languages such as Matlab® or Mathematica®.

Figure 3 shows a time series plot for the ideas generated and tested (with $\alpha = 0.2$, $\beta = 0.3$, $\gamma = 0.1$, $\phi = 0.6$, $\lambda = 0.5$, and $x_1(0) = x_2(0) = 0.1$). As seen in the figure, generated ideas must build up first,

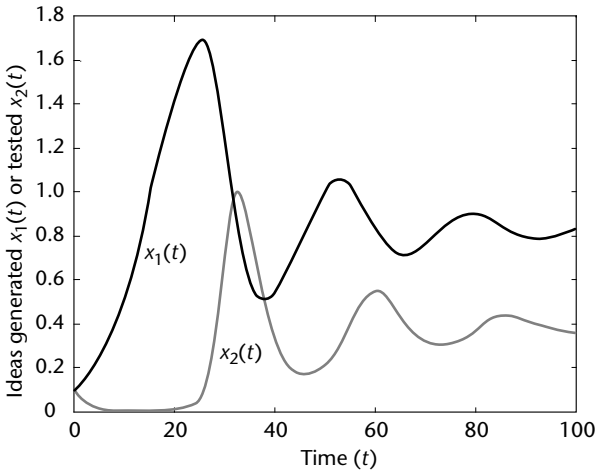


Figure 3. A time series plot showing the trajectories produced by the predator–prey model of idea generation and testing. The horizontal axis represents time elapsed on the task, and the vertical axis represents the number of ideas generated and tested, where generation leads testing.

and testing ideas dominates later. Eventually, both processes approach a steady state.

PROPERTIES OF DYNAMICAL SYSTEMS

Dynamical systems describe the general laws that systems obey. Various special cases of the general law can be derived simply by changing either the initial state $X(0)$, or the system parameters. For a given initial state and a fixed set of parameter values, a dynamical system generates a unique trajectory or path through the state space as a function of time (like that shown in figure 1). The primary goal of dynamical systems theory is to develop analytic methods for studying all the possible trajectories produced by a dynamical system, and to understand how these trajectories change as a function of the initial state and system parameters.

Phase Portraits and Attractors

Figure 4 illustrates these ideas using the simple linear system model described above. The four panels shown in the figure are called *phase portraits*. The two-dimensional plane within each portrait represents the state space of the system. The four different portraits were computed from eqns 6 and 7 by varying the parameter α in eqn 6. The remaining parameters of the model were fixed at $\beta = 1$ and $v = 1$.

Each arrow inside each phase portrait is a velocity vector representing the directional rate of change in motivation and effort that occurs at a particular state of the system. In other words, the head of an arrow indicates where the state will move in the next instant, given the current state indicated by tail of the arrow. Thus the arrows indicate the flow of the dynamical system. The smooth curve following the flow within each portrait indicates the trajectory produced by the system with initial state $p(0) = q(0) = 0$. Different trajectories would result from different choices of initial state. In fact, each initial state defines a unique trajectory, and therefore the trajectories never cross (provided that the local generator satisfies certain smoothness conditions).

The four panels illustrate how the dynamical properties of the model depend on the parameter α. The top left panel shows all of the arrows flowing

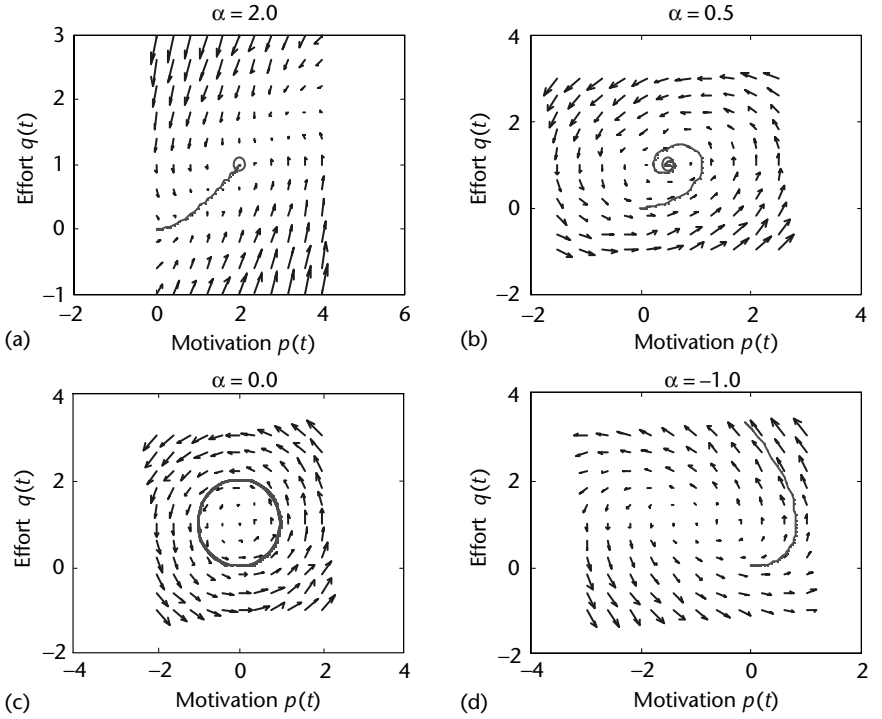


Figure 4. A display of four different phase portraits, one for each setting of the parameter α in Eqns 6 and 7. The horizontal axis within each portrait represents the preference state, and the vertical axis represents the level of effort. Each arrow is a vector indicating the direction and rate of change in the system at a particular point in the state space. The smooth curve within each portrait shows the trajectory produced by setting the initial state equal to zero for preference and effort.

towards a stable equilibrium point located at state $[p, q] = [2, 1]$. In this case, each trajectory moves steadily towards the equilibrium without any oscillation. The top right panel shows the arrows spiralling in towards a stable equilibrium point located at $[p, q] = [0.5, 1]$. The bottom left panel shows the arrows flowing in a circular manner around a central point located at $[p, q] = [0, 1]$. In this case, each trajectory oscillates indefinitely, like a clock. Finally, the bottom right panel shows the arrows spiralling away from an unstable equilibrium point located at $[p, q] = [-1, 1]$. In this case, the system shoots off towards infinity.

All four of these phase portraits contain a special state called an *equilibrium point*. But the nature of the equilibrium varies from one to another. An equilibrium point $\boldsymbol{X}^*$ has the special property that the local generator is zero at this point:

$$\boldsymbol{f}(\boldsymbol{X}^*) = 0 \tag{11}$$

Thus no change occurs when the system is in this state. The equilibrium points for the top left and top right portraits are called *stable* equilibrium points or *attractors* because the system eventually tends towards these equilibrium points whenever the state of the system starts within a close proximity of these points. The equilibrium point for the bottom right panel is an *unstable* equilibrium point or *repellor* because if the system is placed an arbitrarily small distance away from that point, it eventually drifts further away (For rigorous definitions, see any of the references cited at the end of the Introduction.)

Stability Analysis

The model defined by eqns 6 and 7 is an example of a linear system. A linear system enjoys the special property of allowing only a single equilibrium point (provided that the system is nonsingular). The stability of equilibrium points for linear systems can be easily determined by checking the eigenvalues of the linear equations (see (Braun, 1975)). Nonlinear systems, however, allow multiple equilibrium points, of which some may be stable and others unstable, and a more general method of *stability analysis* is required to determine their properties.

There are several general mathematical techniques for studying the qualitative properties of equilibrium points. One of the most powerful is based on the construction of what is called a *Liapunov function* for the dynamical system. A Liapunov function maps each state of the system to a real number

$$\boldsymbol{V}: \Omega \to \mathbb{R} \tag{12}$$

and it has the special property that its time derivative never increases:

$$\frac{\mathrm{d}\boldsymbol{V}}{\mathrm{d}t} \leq 0 \tag{13}$$

for all t. In physics, the Liapunov function can be interpreted as the potential function of a conservative system; and in engineering it can be interpreted as the objective function that a control system is designed to minimize. If there is a Liapunov function for the system, and an equilibrium point $\boldsymbol{X}^*$ is a local minimum of this function, then $\boldsymbol{X}^*$ is a stable attractor. The *basin of attraction* for $\boldsymbol{X}^*$ is the largest possible region for which $\boldsymbol{X}^*$ serves as the attractor. Thus, once the system enters the basin of attraction for $\boldsymbol{X}^*$, then it never leaves, and it converges towards the attractor. When a Liapunov function is defined over the entire state space, then the state space can be partitioned into a collection of attraction basins with a single stable attractor located within each basin (see figure 5).

To illustrate the idea of a Liapunov function, consider the predator–prey model described earlier. Figure 6(a) shows the phase portrait for this model.

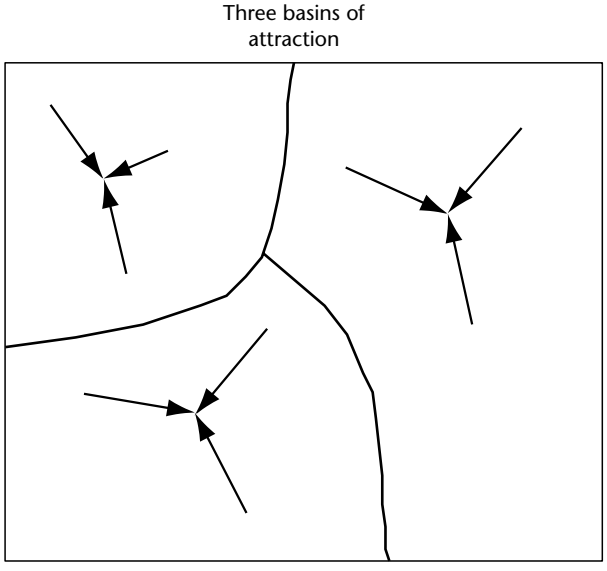


Figure 5. An illustration of a two dimensional state space that is divided into three basins of attraction. The arrows indicate the direction pointing downhill, where the Liapunov function is decreasing. The meeting point of the three arrows within each basin represents the stable attractor at the local medium. The curves indicate the boundaries that separate each basin.

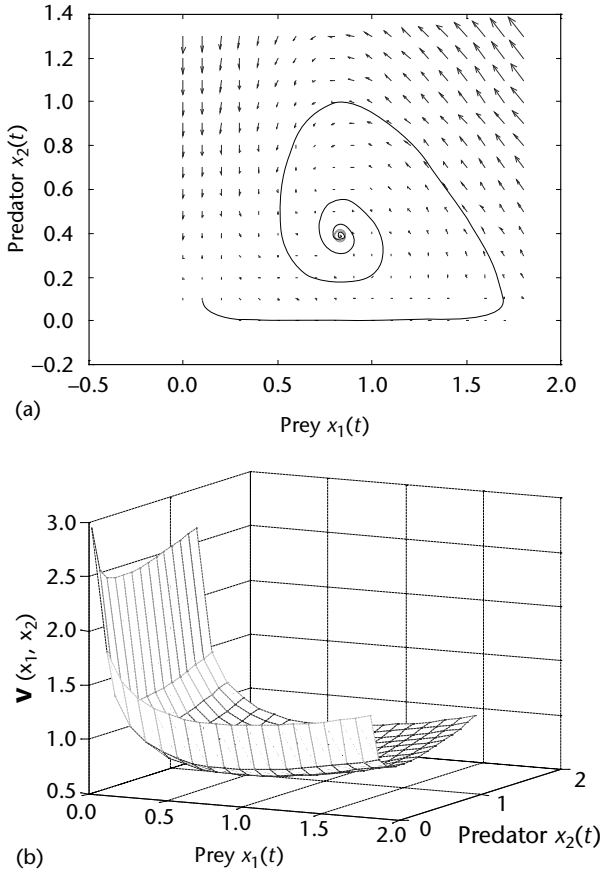


Figure 6. (a) An illustration of the phase portrait produced by the predator–prey model of effort for generating and testing ideas. (b) An illustration of the Liapunov function corresponding to this model. In (b), the plane represents the state space of the model, and the surface on the vertical axis above the plane represents the value of the Liapunov function for each point in the state space.

This nonlinear model has three equilibrium points – [0, 0], $[\frac{\alpha}{\gamma}, 0]$ and $\left[\frac{\lambda}{\phi}, \frac{\phi\alpha-\lambda\gamma}{\beta\phi}\right]$ – but only the last one is stable. For convenience, define $[x_1^*, x_2^*] = \left[\frac{\lambda}{\phi}, \frac{\phi\alpha-\lambda\gamma}{\beta\phi}\right]$. (In the case of figure 6 $[x_1^*, x_2^*] = [0.83, 0.39]$.)

It can be shown that the Liapunov function for this example is

$$\boldsymbol{V}(x_1, x_2) = \lambda\left(\frac{x_1}{x_1^*} - \ln\frac{x_1}{x_1^*}\right) + \alpha\left(\frac{x_2}{x_2^*} - \ln\frac{x_2}{x_2^*}\right) \tag{14}$$

The time derivative of this function is

$$\frac{\mathrm{d}\boldsymbol{V}(x_1, x_2)}{\mathrm{d}t} = \frac{-\lambda^2\gamma}{\phi}\left(1 - \frac{x_1}{x_1^*}\right)^2 \leq 0 \tag{15}$$

which is nonincreasing for all positive values of the state variables. The partial derivative of $\boldsymbol{V}$ with respect to the state variable is zero at $[x_1^*, x_2^*]$, so this point is a local minimum of $\boldsymbol{V}$; hence it is a stable attractor for all positive values of the state variables.

Figure 6(b) shows the surface of the Liapunov function over the state space (using the same parameters). The surface has a minimum at the equilibrium point $[x_1^*, x_2^*] = [0.83, 0.39]$. This point is the stable attractor associated with the basin of attraction inside the positive region of the state space.

Bifurcation Analysis and Catastrophe Theory

A *bifurcation* is said to occur if the equilibrium points undergo qualitative changes as a result of small continuous changes in the model parameters. The parameter values at which these bifurcations occur are called bifurcation points. To illustrate the idea of bifurcation, consider the following example of a slightly more complex version of the predator–prey model:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (0.4)x_1 - (0.04){x_1}^2 - (0.2)\frac{x_1x_2}{1+x_1} \tag{16}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \alpha x_2\left(1 - (0.5)\frac{x_1}{x_2}\right) \tag{17}$$

This example has only one free parameter, α, which is the focus of this bifurcation analysis. For this model, the location of the positive-valued equilibrium point $[x_1^*, x_2^*] = [2.7, 5.4]$ is independent of the parameter α (see (Beltrami, 1987)). Also, for large values of α, this equilibrium point is a stable attractor. The trajectory of the model looks very much like that shown in figure 6(a).

As the parameter α decreases, a qualitative change in the dynamics appears. The equilibrium point changes from an attractor to a repellor. A new behavior appears in which the asymptotic trajectory is attracted towards a *limit cycle* or asymptotic periodic orbit. In this case, the attractor is the set of points of the limit cycle. Figure 7 shows the phase portrait for this case, and the trajectory produced when the system is started at the initial state $[x_1(0), x_2(0)] = [0.5, 0.5]$ and $\alpha = 0.1$. This is an example of what is known as a *Hopf bifurcation*.

Catastrophe theory (Zeeman, 1977) is concerned with bifucations that result in discontinuous jumps in stable equilibrium points. Figure 8 illustrates the idea of a catastrophe. In this figure, each

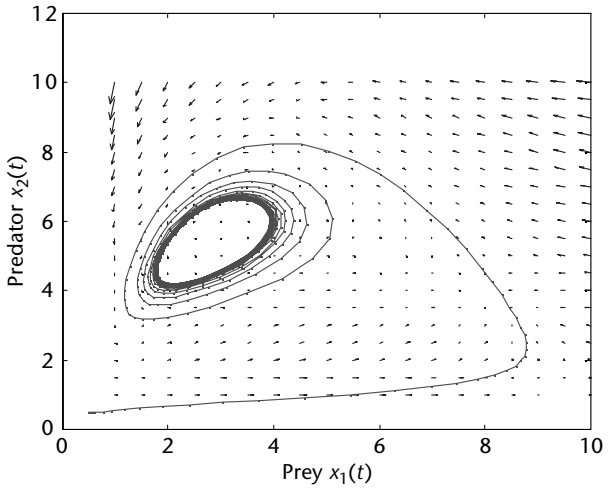


Figure 7. Phase portrait and trajectory produced by a modified predator–prey model that exhibits limit cycle behavior.

of the five curves represent a Liapunov function defined over a one-dimensional state space. The variations between the curves are produced by making small changes in a single parameter of the dynamical system. The upward-pointing arrow indicates the starting position of the system. The first curve, on the far left, exhibits an attractor on the right to which the system converges in the limit. Following a birfucation, the second, third and fourth curves exhibit two attractors, on the left and right, and a repellor in the middle. However, since the initial state lies inside the right basin, the system continues to converge to the attractor on the right. Finally, for the last curve, the change in parameter has eliminated the attractor on the right, so that the basins of attraction combine and the system converges to the attractor on the left. Exactly the opposite jump would occur if the change in parameter were reversed and the system started on the left, producing a *hysteresis* effect.

CHAOTIC SYSTEMS

So far we have encountered three types of asymptotic behavior exhibited by dynamical systems: (1) the system is attracted towards a single attracting state; (2) the system is attracted towards a limit cycle and oscillates indefinitely along some periodic orbit; or (3) the system shoots off towards infinity. However, some dynamical systems exhibit aperiodic behavior that does not fall into any of

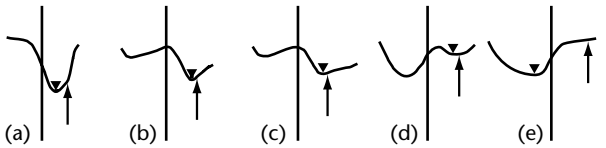


Figure 8. Five hypothetical Liapunov functions produced by small changes in a parameter of a dynamic system. The upward pointing arrow indicates the starting position. The solid triangle indicates the final equilibrium state. Note that the equilibrium makes a sudden jump into a new basin of attraction in the last case.

these three categories. This new category of *strange attractors* is exhibited by *chaotic dynamical systems*.

Chaotic behavior can arise from what appear to be very simple dynamical models. Let us reconsider the discrete-time version of the logistic growth model, given by eqn 3. (Recall that the continuous-time version produced well-behaved and easily understood trajectories for all values of the parameter α).

Setting $h = 1$ for the discrete-time model yields:

$$p(t+1) - p(t) = \alpha p(t)(1 - p(t)) \tag{18}$$

If we define $\beta = 1 + \alpha$, and $x(t) = \frac{\alpha}{1+\alpha} p(t)$, we can write the equation as.

$$x(t+1) = \beta x(t)(1 - x(t)) \tag{19}$$

When $\beta < 1$, the system described by eqn 19 decays to zero; when $1 < \beta < 3$, the system grows towards a stable equilibrium point, as in the continuous-time model. However, as β increases above 3, the system becomes periodic, and for $\beta > 3.57$ the system breaks down and becomes aperiodic or chaotic. Figure 9 shows a time series plot of the behavior of the model with $x(0) = 0.1$, and various values of β.

One of the defining features of a chaotic dynamical system is sensitive dependence on initial conditions. This means that an arbitrarily small change in the initial state is eventually magnified into a large change in future states. This is sometimes referred to as the *butterfly effect*, after the idea that the fluttering of a butterfly's wings in Brazil can eventually set off a tornado in Texas.

For example, if we set $\beta = 4$, then $x(0) = 0.100\,0$ yields $x(10\,000) = 0.209\,8$, but $x(0) = 0.100\,1$ yields $x(10\,000) = 0.981\,9$. Thus, although both of these trajectories were computed from the same equation and started from almost the same initial state, the trajectories they eventually produce are quite different.

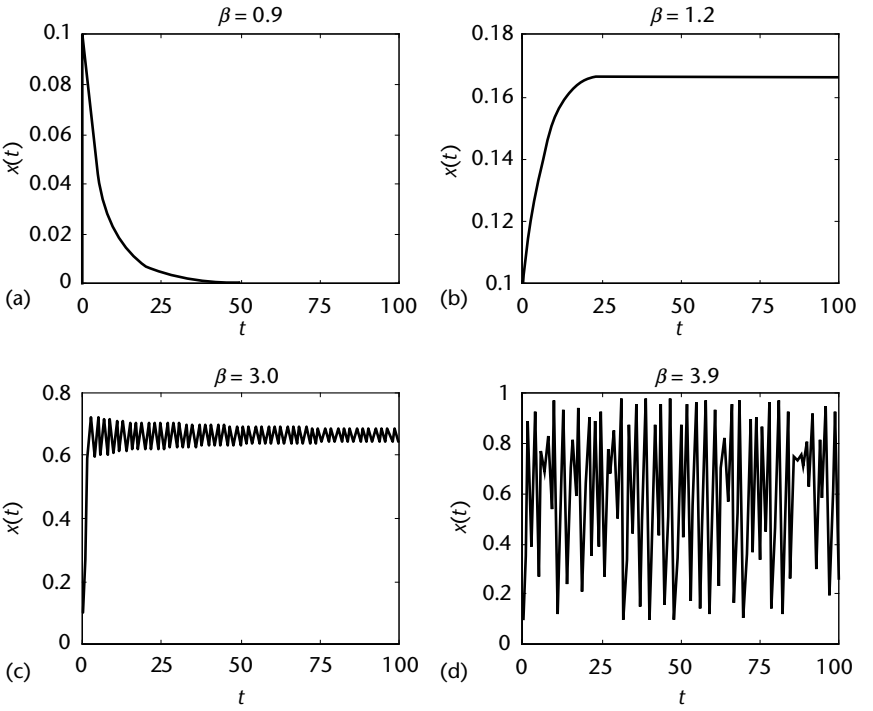


Figure 9. A display of four different time series plots produced by changing the parameter β in the discrete-time logistic growth model.

A rigorous method for identifying chaotic dynamical systems is based on an index, called the Liapunov index, λ:

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln|f'(x_i)| \qquad (20)$$

The function f in eqn 20 refers to the local generator for a univariate discrete dynamical system. For example, for the logistic model f is defined by the right-hand side of eqn 19, and $f'(x) = \beta(1 - 2x)$ in this case. This index provides a measurement of sensitivity to initial conditions. A dynamical system is chaotic when the Liapunov index is positive ($\lambda > 0$).

Figure 10 shows the Liapunov index plotted as a function of β for the logistic model (using $x(0) = 0.10$; however, the pattern does not depend on this starting position). Note that the index is never positive until $\beta > 3.57$, at which point the model becomes chaotic. It is also interesting to note that the system occasionally returns to periodic behavior at a few higher values of β.

The discrete-time logistic model is the simplest example of a chaotic dynamical system. However, chaotic behavior is not limited to discrete-time systems. Many (and more complex) examples of continuous-time chaotic systems have also been studied.

Chaotic behavior may seem to be an 'undesirable' property for a dynamical system. But these models do provide an alternative to stochastic dynamical models for describing the unpredictability and variability manifested by many complex biological systems.

TYPES OF DYNAMICAL SYSTEMS

Cognitive scientists have employed many different types of dynamical systems, ranging from production rule models of computers (Newell and Simon, 1972) to artificial neural network models of the brain (Grossberg, 1982; Rumelhart and McClelland, 1986). These models differ according to some basic characteristics as described below.

Discrete Versus Continuous State Spaces

A computer system has a *discrete* state space, which contains a countable number of possible states. The state space for the brain model is an example of a *continuous* vector space, which contains an uncountable number of states. The latter is also endowed with additional properties, including scalar multiplication of states ($aX \in \Omega$ for any real number a and state X), addition of states

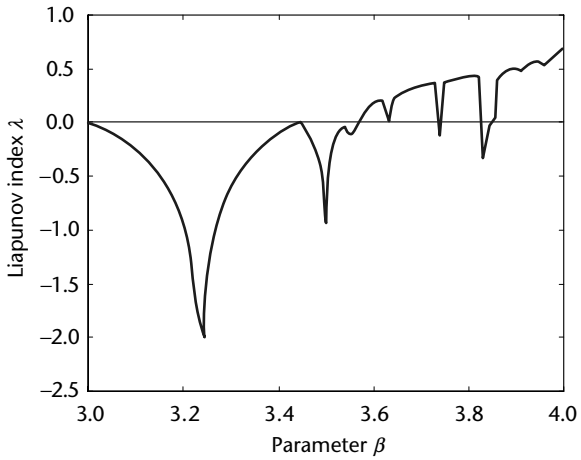


Figure 10. A plot of the Liapunov index as a function of the parameter β for the discrete-time logistic growth model.

$(X_1 + X_2 \in \Omega$ for any X_1 and $X_2 \in \Omega)$, and distances between states $(\| X_1 - X_2 \|)$. Dynamical systems theory usually assumes that state spaces are vector spaces.

Discrete Versus Continuous Time Indices

The computer model is an example of a discrete-time system, in which a countable set of time points is indexed by the set of natural numbers $\{0, 1, 2, 3, \ldots\}$. The brain model is an example of a continuous-time system, in which an uncountable set of time points is indexed by the set of nonnegative real numbers $[0, \infty]$. Originally dynamical systems theory was concerned only with continuous-time systems, but now both types are studied in parallel.

Linear Versus Nonlinear Systems

The computer model is an example of a discrete nonlinear system in which the production rules produce jumps from state to state. Some early neural models used continuous linear state transition functions, but more recent neural models use continuous nonlinear transition functions. In general, a dynamical system is linear if the local generator $\boldsymbol{f}$ satisfies the following superposition property for two arbitrary states X_1 and X_2 and scales a and b:

$$\boldsymbol{f}(aX_1 + bX_2, t) = a\boldsymbol{f}(X_1, t) + b\boldsymbol{f}(X_2, t) \tag{21}$$

In this case, the local generator can be written as a linear combination of the state variables:

$$\boldsymbol{f}_j(X(t), t) = \sum_{i=1}^{n} \alpha_i(t) x_i(t) \tag{22}$$

The coefficients used to define the linear combination are called the system parameters. The model of effort defined in eqns 6 and 7 is an example of a linear model. The logistic and predator–prey models are both examples of nonlinear models.

Time-invariant Versus Time-varying Systems

The computer model is an example of a time-invariant system, in which the state transition function does not change over time. A brain model, on the other hand, may require a time-varying system to allow for growth, development and aging. In general, the system is time-invariant if the local generator is independent of the time index:

$$\boldsymbol{f}(X, t) = \boldsymbol{f}(X) \tag{23}$$

One way to redefine a time-varying system as a time-invariant system is to add an extra state variable and set it equal to the time index. For example, the model of effort defined in eqns 6 and 7 contains a possibly time-varying input, $v(t)$. However, this two-dimensional time-varying system $[p, q]$ can be transformed into a three-dimensional time-invariant system $[p, q, x_3]$ by defining a third state variable, $x_3 = t$. The new three-dimensional system can then be described by three equations:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = p - \alpha q \tag{24}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = v(x_3) - \beta q \tag{25}$$

$$\frac{\mathrm{d}x_3}{\mathrm{d}t} = 1 \tag{26}$$

Deterministic Versus Stochastic Systems

The computer model is an example of a deterministic system: if we know the exact state of the system at time t then we can predict the state of the system at time $t + 1$. Our simple model of effort (eqns 6 and 7) was also formulated as a deterministic system. However, models of the brain must account for the inherent unpredictability of human behavior. In the past, this was usually accomplished by allowing a subset of the variables in

the state vector to be stochastic, or by allowing the initial state vector to be a random vector. Recently, deterministic but chaotic dynamical systems have been explored as alternatives to stochastic dynamical systems.

Bhattacharya and Waymire (1990) provide an introduction to stochastic dynamical systems theory.

SUMMARY

Dynamical systems theory was originally developed to solve problems arising in physics and engineering. Now cognitive scientists are making use of this approach, especially in applications of connectionist and neural models of cognition. Most applications are much more complex than the examples given in this article. Nevertheless, cognitive scientists have successfully applied these ideas to various substantive areas including pattern recognition, motor behavior, cognitive development, learning, thinking, and decision-making. There is little doubt that dynamical systems theory has much to contribute to cognitive science.

References

Atkinson JW and Birch D (1970) *The Dynamics of Action*. New York, NY: Wiley.

Bhattacharya RN and Waymire EC (1990) *Stochastic Processes With Applications*. New York, NY: Wiley.

Beltrami E (1987) *Mathematics for Dynamic Modeling*. New York, NY: Academic Press.

Braun M (1975) *Differential Equations and Their Applications*. New York, NY: Springer.

Grossberg S (1982) *Studies of Mind and Brain*. Reidel.

Luenberger DG (1979) *Introduction to Dynamic Systems*. New York, NY: Wiley.

Miller G, Galanter E and Pribram KH (1960) *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart, Winston.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Padulo L and Arbib MA (1974) *System Theory*. Philadelphia, PA: Saunders.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I. Cambridge, MA: MIT Press.

Strogatz SH (1994) *Nonlinear Dynamics and Chaos*. Reading, MA: Addison-Wesley.

Townsend JT and Busemeyer JR (1989) Approach-avoidance: return to dynamic decision behavior. In: Izawa C (ed.) *Current Issues in Cognitive Processes: The Flowerree Symposium on Cognition*, pp. 107–133. Hillsdale, NJ: Erlbaum.

Zeeman EC (1977) *Catastrophe Theory: Selected Papers 1972–1977*. Reading, MA: Addison-Wesley.

Further Reading

Anderson JA (1997) *Introduction to Neural Networks*. Cambridge, MA: MIT Press.

Gleick J (1987) *Chaos: Making a New Science*. New York, NY: Viking Press.

Golden RM (1996) *Mathematical Methods for Neural Network Design and Analysis*. Cambridge, MA: MIT Press.

Grossberg S (1988) *Neural Networks and Natural Intelligence*. Cambridge, MA: MIT Press.

Haykin S (1994) *Neural Networks*. New York, NY: Macmillan.

Port RF and Van Gelder T (1995) *Mind As Motion*. Cambridge, MA: MIT Press.

Evolutionary Algorithms

Intermediate article

Jennifer S Hallinan, University of Queensland, St Lucia, Queensland, Australia
Janet Wiles, University of Queensland, St Lucia, Queensland, Australia

CONTENTS

Evolutionary algorithms make direct use of adaptive change in a population of computational agents in order to solve problems or model complex systems.

INTRODUCTION

So far, the only process that has produced intelligence is evolution. Consequently, the prospect of incorporating the principles underlying biological evolution into models of the development and performance of cognitive systems is attractive to many cognitive scientists. The basic principles of evolution are straightforward. Even before the mechanical details of DNA and the machinery of molecular biology were understood, it was apparent to any observer that individuals of the same species differ in minor ways. Darwin (1859) realized that these minor differences could affect the number of offspring left by an individual, particularly when there are more individuals of a species than its habitat can comfortably support. Natural selection therefore acts to reduce the genetic contribution of less fit individuals to future generations, gradually causing a population to become better adapted to its environment.

Evolution, working in a blind and directionless fashion, has produced the astonishingly orderly and robust complexity of the biosphere. It is evidently a powerful optimizer, and the idea of incorporating evolutionary principles into computer algorithms has been used since the 1940s by engineers in search of optimization tools for use in complex, nonlinear systems. Evolutionary algorithms (EAs) are also a natural choice for modeling systems to whose development biological evolution has been fundamental, such as the human cognitive architecture.

COMPUTATIONAL PARADIGMS FOR EVOLUTION

'Evolutionary algorithm' is an umbrella term, covering a family of algorithms which take inspiration from the principles of biological evolution. An EA is any algorithm that makes use of adaptive change in a population of computational agents in order to solve problems or model complex systems. There are several varieties of EA (Mitchell, 1996).

Genetic Algorithms

A genetic algorithm (GA) involves a population of individual 'chromosomes'. A chromosome may consist of a string of bits, a string of real numbers, or a more complex composition, depending on its purpose. Each chromosome can be assigned a numerical 'fitness', as defined by a fitness function, which measures how well the solution encoded in that chromosome solves the problem at hand. Chromosomes are chosen to contribute to the next generation in a fitness-dependent manner, so that fitter chromosomes have more offspring than less fit chromosomes. New chromosomes are produced by copying the parents and applying genetic operators based on biological mutation and crossover. The fitness of the new generation is then assessed, and the process iterated until a good solution is developed (Figure 1).

The term 'simple genetic algorithm' is often used to refer to an algorithm as outlined in Figure 1 using a bit-string chromosome.

There are endless variations on the basic GA. The outcome may depend on the genetic operators used, how the problem is represented, selection

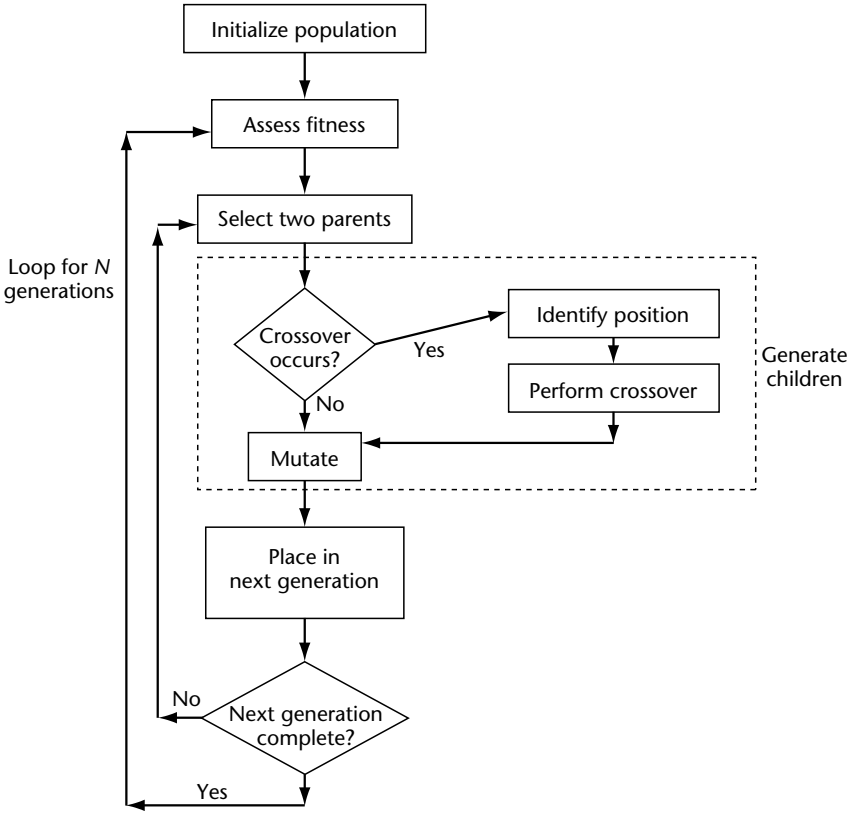


Figure 1. A simple genetic algorithm.

strategies, the order of application of the operators, and how new generations are constructed.

Other Paradigms

In addition to the simple GA, there are several other evolutionary computation paradigms. The major ones are described briefly below. Despite the differences in implementation, the various algorithms are alike in utilizing stochastic, fitness-dependent selection over random variations between individuals as a tool to develop individuals optimally adapted to their environment.

Evolutionary programming

'Evolutionary programming' (Fogel, 1999) was developed by L. J. Fogel in the early 1960s. It does not use a 'genomic' representation. Each individual in the population is an algorithm, chosen at random from an appropriate sample space. Mutation is the only genetic operator used; there is no crossover.

Evolutionary strategies

'Evolutionary strategies' (Schwefel, 1995) were developed by H. -P. Schwefel, also in the 1960s, as an optimization tool. They use a real-valued chromosome, with a population of one, and mutation as the only genetic operator. In each generation, the parent is mutated to produce a descendant; if the descendant is fitter it becomes the parent for the next generation, otherwise the original parent is retained.

Classifier systems

In a 'classifier system' (Holland, 1992), a classifier takes inputs from the environment and produces outputs indicating a classification of the input events. New classifiers are produced through the action of a genetic algorithm on the system's population of classifiers. (*See* **Classifier Systems**)

Genetic programming

The aim of 'genetic programming' (Koza, 1999), developed by J. Koza in the late 1980s, is the automatic programming of computers: allowing programs to evolve to solve a given problem. The population consists of programs expressed as parse trees; the operators used include crossover, mutation, and architecture-altering operations patterned after gene duplication and gene deletion in nature.

Evolution as a Search Process

The fundamental process underlying all evolutionary algorithms is a heuristic search through a search space, or fitness landscape, defined by the problem representation in conjunction with the fitness function.

The concept of a 'fitness landscape' was introduced by S. Wright (1932) in the context of biological evolution. He suggested that for a given set of genes, each possible combination of gene values ('alleles') could be assigned a fitness value for a particular set of conditions. The entire genotype space can then be visualized as a landscape, with genotypes of high fitness occupying peaks and those of low fitness forming troughs. Such a fitness landscape is generally very high-dimensional, but fitness landscape can be visualized in two dimensions (Figure 2).

The fitness landscape metaphor has proven to be powerful, and is applicable to computational as well as biological evolution. A population whose members have slightly differing genotypes is represented as a set of points in a fitness landscape. Mutation and natural selection will act to drive a population up to its nearest local maximum, which may or may not be the global maximum for the landscape. Under strong selection pressure, the population may become 'trapped' on a suboptimal local maximum. Because of the stochastic nature of the evolutionary process, however, the population will be spread out over the landscape to some extent, and different individuals may find themselves on the slopes of different maxima, depending on the ruggedness of the environment. The population is, in effect, performing a parallel search of the fitness landscape. (*See* **Machine Learning**)

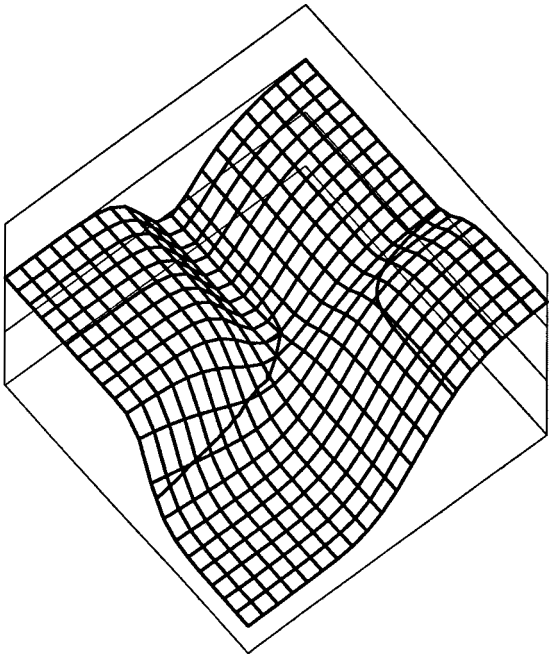


Figure 2. A two-dimensional fitness landscape. It shows the global maximum and one local maximum.

The Schema Theorem

Much of the research into evolutionary algorithms has been purely empirical, with EAs being used for optimization of problems involving multiple parameters, for which the problem domain, and hence the search space, is poorly understood. There are continuing efforts, however, to formulate a theoretical basis for evolutionary computation. One of the earliest attempts to produce such a formulation was the 'schema theorem' proposed in 1975 by Holland (Holland, 1992) and restated by Goldberg (1989) as 'short, low-order, above-average schemata receive exponentially increasing trials in subsequent generations'. Briefly, the idea is that a schema is a set of building blocks which can be described by a template comprising ones, zeros, and asterisks, with the asterisks representing 'wild cards' which can take on any value. The evolutionary algorithm proceeds by identifying schemas of high fitness and recombining them using crossover in order to produce entire individuals having high fitness. The theory is attractive because, for a given EA, schemata can be identified, and the effects of mutation and crossover on schemata in a population of a given size can be quantified, potentially providing useful insight into the way in which the EA functions.

There has been a large amount of investigation into the schema theorem, with often inconclusive and controversial results. Mitchell *et al*. (1991) designed a class of fitness landscapes which they called the 'royal road' functions. These functions produce a hierarchical fitness landscape, in which crossover between instances of fit lower-order schemata tends to produce fit higher-order schemata. However, these researchers found that the presence of relatively fit intermediate stages could in fact interfere with the production of fit higher-order solutions, since once an instance of a fit higher-order schema is discovered its high fitness allows it to spread quickly throughout the population, carrying with it 'hitchhiking' genes in positions not included in the schema. Low-order schemata are therefore discovered more or less sequentially, rather than in parallel.

The extent to which the schema theorem applies in practice remains controversial. For example, Vose (1999) places little credence in the general applicability of the schema theorem, and offers an alternative mathematical approach to analyzing the behavior of the simple genetic algorithm.

Exploration Versus Exploitation

An EA attempts to find optimal areas of its search space by discovering new solutions (exploration) as well as by maintaining fit solutions that have already been discovered (exploitation). Achieving a good balance between exploration and exploitation is important to the efficacy of the algorithm: too little exploration means that important areas of the search space may be ignored, while too little exploitation increases the risk of prematurely discarding good solutions. The trade-off between exploration and exploitation is often studied in the context of the 'n-armed bandit problem'.

The n-armed bandit is an extension of the one-armed bandit, or casino slot machine. Instead of having a single lever which can be pulled, the bandit has n levers, each of which has an expected reward. Since the expected reward associated with each lever is unknown, a strategy must be developed to balance exploitation of knowledge already gained (by playing various arms) with exploration of the behavior of untested arms. Exploitation of the best action discovered so far will maximize expected reward on a single play, but exploration may lead to a better total reward in the long run. The n-armed bandit provides a basis for a mathematical formulation of the exploration–exploitation trade-off for an EA. The optimal strategy involves exponentially increasing the probability of selecting the best solutions over time. (See Holland, 1992, chap. 5 and chap. 10 for a detailed discussion of the n-armed bandit and its application to schema theory.)

Evolution Versus Hill Climbing

The power of an EA is often assumed to lie in its 'implicit parallelization': by maintaining a population of candidate solutions that are modified by mutation or crossover, the algorithm is, in effect, exploring different regions of its search space in parallel. The simplest alternative to an EA is a 'hill climber', an algorithm which maintains a population of one individual and performs a strictly local search using mutation. There are many variations of the hill climber algorithm; most involve mutating the current best candidate and accepting the mutated individual if it is fitter than the original. A hill climber will thus typically move only to the top of the nearest peak in the fitness landscape, which may not be a global optimum.

An EA is no more efficient than multiple random restarts of a hill climber, in terms of the number of evaluations performed. An EA with a population size of 100 running for 1000 generations performs 100 000 evaluations, as does an algorithm with a single population member restarted 100 times for 1000 generations each time. The EA would, however, be expected to outperform the hill climber if the action of the genetic operators used in the EA provided advantages over local search. This would be the case if the schema theorem applied as described above, with useful partial solutions discovered by different individuals being recombined to produce fitter individuals more rapidly than could be done by mutation alone. The EA would also be expected to outperform the hill climber if the structure of the fitness landscape was such that the 'implicit memory' of a population-based algorithm (i.e., the memory encoded into the structure of the population itself as a result of evolution) allowed it to concentrate its search in areas of high fitness in a manner that would not be possible for a hill climber.

In practice, hill climbers with multiple restarts often perform at least as efficiently as population-based algorithms (Mitchell *et al.*, 1994).

COEVOLUTION

The evolutionary algorithms described so far are primarily heuristic optimizers based upon a simple model of evolution. In biological systems, however, evolution in a particular population occurs in the context of a complex mixture of endogeneous and exogenous factors. The simple EA has been extended in several directions inspired by further observations from the realm of biology.

Biological evolution does not occur in isolation. Individuals of a particular species live, breed, and die in collaboration and competition with other organisms of their own and other species, and evolutionary change in one population affects the fitness landscapes of other populations. Kauffman (1996) uses the analogy of the fitness landscape as a rubber sheet. Evolutionary change in one species deforms the sheet, not only for itself, but for all species existing in that fitness landscape.

'Coevolution' is often used in an EA as a mechanism for the prevention of premature convergence to a suboptimal fitness peak. A population of problems and a population of solutions evolving

together should, in theory, produce better solutions, since as good solutions are found the problems against which they are tested become harder, and an evolutionary 'arms race' develops (Dawkins and Krebs, 1979). For an interesting overview of the background to coevolution in evolutionary computation, see Rosin and Belew (1997).

LEARNING AND EVOLUTION

The interaction between learning and evolution has been the subject of extensive research, both by those interested in using learning as a local search operator to improve the optimization performance of the algorithm, and by those interested in understanding the evolution and utility of learning in biological systems.

The so-called Baldwin effect (Baldwin, 1896) is based upon the idea that learning on the part of individuals could guide the course of evolution in the population as a whole. A particular trait may be learned, or it may be innate (a term which can be taken, in this context, to indicate that it is genetically determined). A learned trait has the advantage of providing flexibility, but the disadvantage of being slow to acquire; an innate trait is present from birth, but inflexible. Baldwin suggested that traits that are initially learned may, over time, become encoded in the genotype of the population.

Although this suggestion appears, at first glance, to be tantamount to Lamarckism, in fact no Lamarckian phenotype-to-genotype information flow is required. For the Baldwin effect to operate, two conditions must be met. Firstly, the trait in question (which may be a behavioral or a physical trait) must be influenced by several interacting genes, so that a mutation in one of these genes will make the phenotypic expression of the trait more likely. Secondly, an individual bearing such a mutation must be able to learn to express the trait.

Under these conditions, learning acts to provide 'partial credit' for a mutation. An individual carrying a mutation that predisposes it towards an advantageous phenotype will learn the trait more easily than its less fortunately genetically endowed conspecifics, and thus will tend to survive and pass on more copies of the relevant allele to the next generation. Over time, multiple mutations for the desirable trait will accumulate in the genes, and the trait will thus become innate in the population.

Hinton and Nowlan (1987) were the first to demonstrate the feasibility of the Baldwin effect, at least in a simplified, computational model, and much research has been done in this area in the intervening years. See Turney (2001) for a comprehensive online bibliography of publications on the Baldwin effect.

Learning may also be incorporated into an EA by making a learning system, such as a neural network, the object that is evolved. When evolving a neural network, the network architecture, weights, learning rules, and input features may all be subject to evolution. Modified genetic operators may be required in order to avoid disruption of the network architecture. See Yao (1999) for a comprehensive review of recent research in the evolution of artificial neural networks. (*See* **Connectionist Architectures: Optimization**)

EVOLUTIONARY COMPUTATION AND COGNITIVE SCIENCE

Evolutionary computation plays several roles in cognitive science. It has been used both as a modeling framework for exploring ideas inspired by biological evolution, and for the optimization of computational models. (*See* **Artificial Life**)

Evolutionary algorithms have been used both to understand the development of well-known cognitive mechanisms, and to create new mechanisms with desired emergent behaviors. They are used in diverse domains, including language, memory, reasoning, motor control, and the analysis of social interactions.

Modeling contributes to cognitive science in several ways: in providing converging evidence for empirical studies, in testing the internal consistency and completeness of theories, and in investigating the functioning and emergent properties of complex systems, which are often intractable to mathematical analysis. Evolutionary models have often been shown to exhibit complex dynamics emerging from the interaction between computational agents: dynamics that are not inherent in the behavior of a single such agent.

As a modeling framework, evolutionary algorithms are most widely used within evolutionary psychology, which takes as its starting point the hypothesis that the mind contains a set of evolved systems, each one designed to solve a task that was important for survival in the human ancestral environment. (*See* **Evolutionary Psychology: Theoretical Foundations**; **Human Brain, Evolution of the**)

The Evolution of Altruism: The Prisoner's Dilemma

Many organisms exist in social groups, interacting frequently with others of their own and other

species. While an individual may benefit from cooperative interactions with others of its species, conspecifics are also any animal's fiercest competitors for food, mates, and territory. Despite this inescapable competition, altruistic behavior (i.e., behavior that benefits another at some cost to the altruist) is often observed in natural populations.

In an attempt to understand how such apparently paradoxical behavior could arise, a simple game known as the 'prisoner's dilemma' has been much studied. Suppose that two criminals have been arrested on suspicion of committing a crime. They are held incommunicado, and each offered a bargain: if the prisoner admits to the crime (defects) while the other prisoner keeps silent (cooperates), the defector will go free while the cooperator gets a long sentence. If both prisoners keep silent, they will both receive short sentences; and if they both admit to the crime they will each get an intermediate sentence. A possible pay-off matrix for the prisoner's dilemma is shown in Figure 3.

What makes the prisoner's dilemma interesting is that, while the best outcome for the prisoners as a pair is for both to cooperate, the best decision for each prisoner, in the absence of knowledge about the other prisoner's decision, is to defect. Prisoner's dilemmas arise frequently in real life – in any situation in which the action that most benefits an individual harms the group as a whole.

Analysis of the prisoner's dilemma may appear to support the conclusion that altruism cannot arise as a consequence of evolution, which requires that individuals act 'selfishly' in order to pass on their own genes to the next generation. However, Axelrod (1984) studied an iterated version of the prisoner's dilemma, in which individuals play against each other repeatedly, and have a memory of the past behavior of other individuals, with the opportunity to adjust their strategy based on this past history. In a computer tournament of strategies for the iterated prisoner's dilemma, the clear winner was 'tit for tat', which starts by cooperating, and then copies whatever its opponent did in the previous round. Once cooperation becomes established, this strategy will continue to cooperate indefinitely. This strategy has been proven to be highly robust.

	***B* cooperates**	***B* defects**
***A* cooperates**	1 year, 1 year	3 years, 0 years
***A* defects**	0 years, 3 years	2 years, 2 years

Figure 3. Pay-off matrix for the prisoner's dilemma. In each cell, the first amount is *A*'s sentence and the second amount is *B*'s sentence.

Considerable research has been conducted into the evolution of altruism, using the iterated prisoner's dilemma and other models. For an overview of the literature, see Ridley (1996). (*See* **Social Processes, Computational Models of**; **Game Theory**)

The Evolution of Language

A second area where evolutionary algorithms are being increasingly applied is the study of the evolution of language. Human language leaves no fossils, and no other animals have communication systems utilizing such extensive symbolic structures, so studying the evolution of language poses particular problems for cognitive scientists. Human languages share features (such as phonology and syntax) to such an extent that a universal grammar has been conjectured to explain the similarities, but their origins have long been controversial. (*See* **Language Learning, Computational Models of**)

Recently, questions about the evolutionary origins of language and the extent to which it is determined by the cognitive architecture of the young child have been addressed using evolutionary algorithms. Groups of simple language users, modeled as computational agents, have been programmed to evolve communication systems. The emergent behaviors of such systems provide converging evidence on the possibilities for the evolution of language.

Two levels of approach to the evolution of language phenomena have been proposed. In 'micro' evolutionary modeling, learners are modeled by computational agents; a language is a set of utterances (sequences of symbols); global properties of languages are emergent properties; and either the learners or the set of utterances evolves. By contrast, in 'macro' evolutionary modeling, learners are modeled as bundles of parameters; a language is an abstract entity; global properties are explicit parameters; and the distribution of parameters evolves. The micro and macro approaches differ in terms of model fidelity versus analytic tractability, and the role that emergence plays in explaining phenomena. (*See* **Emergence**)

EAs have been used to study language at three levels: phonology (e.g. the self-organization of sound systems); lexicon (e.g. learning concepts, relating words to meanings, and convergence of a population on common conventions for word meanings); and syntax (e.g. the emergence of compositional structure from initially unstructured

utterances). (*See* **Phonology and Phonetics, Acquisition of**; **Semantics, Acquisition of**; **Syntax, Acquisition of**)

The main conclusion that can be drawn from simulations of language evolution is that over time weak functional constraints on language use and acquisition can give the appearance of strong constraints on language structure. Interaction itself accounts for many aspects of coordinated communication of groups. See Hurford *et al*. (1998) and Wray (2002) for an introduction to the breadth of work in this area.

EVOLUTIONARY ROBOTICS

Evolutionary robotics is a rapidly expanding field based on EA techniques. Both fixed and autonomous robots have many parameters that require optimizing, for example in kinematics (topology and dimensions), dynamics, components, and control. EAs are currently being used in a wide variety of projects, as a means for optimization of many parameters in parallel, with relatively unconstrained problems. Autonomous robots are being evolved to develop their own skills via direct interaction with the environment and without human intervention. See Pfeiffer and Scheier (1999) for an introduction to this area.

CONCLUSION

Evolutionary computation draws inspiration from the study of biological evolution in order either to model the evolutionary process, or to use simplified evolutionary principles in order to solve complex, nonlinear optimization tasks. Evolutionary ideas have been used in computing since the 1940s, but it is only with the recent availability of cheap, powerful desktop computers that EAs have become a tool readily available to researchers in all fields. In the field of cognitive science, they are used both to model the development of those cognitive structures and processes that are supposed to be the product of biological evolution – such as human mental structures, language, and altruistic behavior – and to optimize the parameter settings of cognitive models in general.

References

Axelrod R (1984) *The Evolution of Cooperation*. New York, NY: Basic Books.

Baldwin JM (1896) A new factor in evolution. *The American Naturalist* **30**: 441–451. [Reprinted in: Belew R and Mitchell M (1996) *Adaptive Individuals in Evolving Populations*, pp. 59–80. Reading, MA: Addison-Wesley.]

Darwin C (1859) *On the Origin of Species in Means of Natural Selection*. London, UK: John Murray.

Dawkins R and Krebs JR (1979) Arms races between and within species. *Proceedings of the Royal Society, Series B* **205**: 1161.

Fogel LJ (1999) *Intelligence Through Simulated Evolution: Four Decades of Evolutionary Programming*. New York, NY: John Wiley.

Goldberg D (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley.

Hinton GE and Nowlan SJ (1987) How learning guides evolution. *Complex Systems* **1**: 495–502.

Holland JH (1992) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA: MIT Press.

Hurford JR, Studdert-Kennedy M and Knight C (eds) (1998) *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge, UK: Cambridge University Press.

Kauffman SA (1996) *At Home in the Universe*. London, UK: Penguin.

Koza J (1999) *Genetic Programming*, vol. III: *Darwinian Invention and Problem Solving*, San Francisco, CA: Morgan Kaufmann.

Mitchell M (1996) *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Mitchell M, Forrest S and Holland JH (1991) The royal road for genetic algorithms: fitness landscapes and GA performance. In: Vorela FJ and Bourgine P (eds) *Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pp. 245–254. Cambridge, MA: MIT Press.

Mitchell M, Holland JH and Forrest S (1994) When will a genetic algorithm outperform hill climbing? In: Cowan DJ, Tesauro G and Alspector J (eds) *Advances in Neural Information Processing Systems*, vol. VI, pp. 51–58. San Mateo, CA: Morgan Kaufman.

Pfeiffer R and Scheier C (1999) *Understanding Intelligence*. Cambridge, MA: MIT Press.

Ridley M (1996) *The Origins of Virtue*. London, UK: Penguin.

Rosin CD and Belew RK (1997) New methods for competitive coevolution. *Evolutionary Computation* **5**(1): 1–26.

Schwefel H-P (1995) *Evolution and Optimum Seeking*. New York, NY: John Wiley.

Turney P (2001) *The Baldwin Effect: A Bibliography*. http://www.ai.mit.edu/~joanna/baldwin.html.

Vose MD (1999) *The Simple Genetic Algorithm*. Cambridge, MA: MIT Press.

Wray A (2002) *The Transition to Language*. Oxford, UK: Oxford University Press.

Wright S (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: Jones DF (ed.) *Proceedings of the Sixth International Congress of Genetics*, vol. I, pp. 356–366. [Reprinted In: Ridley M (1997) *Evolution*. Oxford, UK: Oxford University Press.]

Yao X (1999) Evolving artificial neural networks. *Proceedings of the IEEE* **87**: 1423–1447.

Further Reading

Davis L (ed.) (1991) *Handbook of Genetic Algorithms*. New York, NY: Van Nostrand.

Dawkins R (1976, 2nd edn. 1989) *The Selfish Gene*. Oxford, UK: Oxford University Press.

Fogel D (1995) *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence*. Piscataway, NJ: IEEE Press.

Fogel D (1998) *Evolutionary Computation: the Fossil Record*. Piscataway, NJ: IEEE Press.

Langdon WB (1998) *Genetic Programming and Data Structures*. Hingham, MA: Kluwer.

Michalewicz Z (1992) *Genetic Algorithms + Data Structures = Evolution Programs*. New York, NY: Springer-Verlag.

Expert Systems

Introductory article

Larry R Medsker, American University, Washington, DC, USA
Theodor W Schulte, American University, Washington, DC, USA

CONTENTS

Expert systems are specialized computer programs that capture the knowledge of human experts. Problems are solved by retrieving knowledge and performing automated logical reasoning.

INTRODUCTION

An expert system is a computer program that stores the knowledge of one or more human experts concerning a limited field of expertise and is able to retrieve that knowledge as well as do automated reasoning on it. During the 1950s, computer scientists started to move computation from mere data manipulation to simulation of activities associated with human intelligence. The field of artificial intelligence (AI) encompasses a number of heterogeneous disciplines, including machine learning, robotics, machine vision, evolutionary computation, natural language processing, and knowledge-based techniques such as expert systems. After basic techniques of knowledge representation and automated reasoning were developed in the 1960s and 1970s, functional expert systems were developed in the late 1970s and commercial applications appeared in the early 1980s. Starting in the late 1980s and early 1990s, expert systems have been combined with other intelligence technologies to form hybrid systems, and from the late 1990s computer networks and the internet have allowed existing software to be further integrated to increase its functionality.

The interest in expert systems derives from the importance and value of human expertise. During a long process of training and experience, a human expert gains special knowledge in a limited field (usually called a 'domain' in computer science). Human expertise typically includes rules of thumb, or heuristics. Heuristics allow human experts to solve complex tasks that cannot be solved by strict mathematical formulas or algorithms. Experts are generally rare and expensive, and their expertise is important in many areas of science, business, engineering, medicine, and so on. Expert systems have been successfully implemented in all of these areas even though they have not been able to abolish or significantly reduce the need for human experts. Often, the role of an expert system is to cover a large proportion of routine situations so that the expert can concentrate on more unusual and interesting cases.

THE COMPONENTS OF EXPERT SYSTEMS

A central component of an expert system is its knowledge base. In rule-based systems, the knowledge base consists of facts and rules (Figure 1). Facts are simple statements that are believed to be true. Rules are complex statements of the form 'if premise then conclusion'. The logic of the domain knowledge is separated from the logic of the general program. The user of the expert system interacts via a user interface, which formats the questions of the user and passes them on to an inference engine. The inference engine uses its programmed algorithms for automated reasoning to find an answer or adequate response to the question or input. It uses a scheduling mechanism to handle multiple inputs, facts, and rules that are active. The interpreter executes the item chosen by the scheduler by applying the rules and facts in the knowledge base. The final result of the reasoning is passed on to the user interface, which formats the output for the user. An expert system may also have an explanation component that can tell the user how the conclusion was derived.

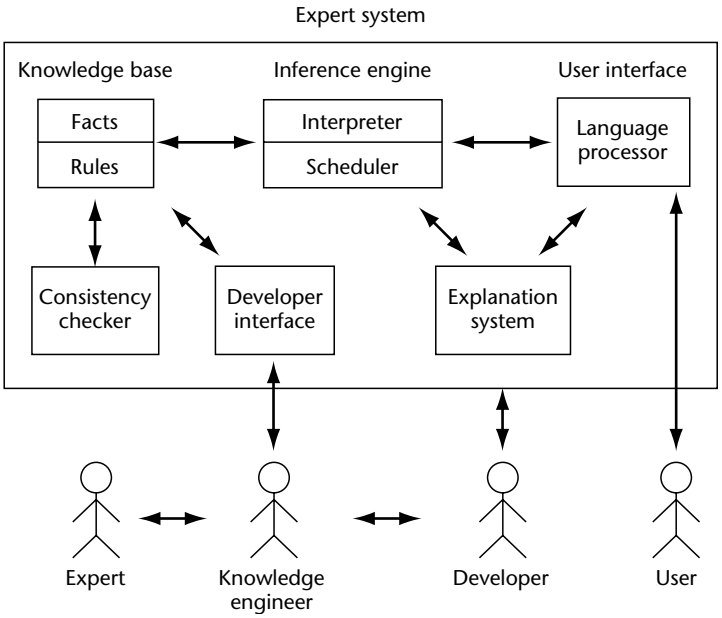


Figure 1. Components of a rule-based expert system.

The development of an expert system consists of two separate tasks: the development of the expert system tool (often called an expert system shell), and the definition of the expert knowledge (rules and facts), a process called 'knowledge engineering'. Expert system shells can be implemented in general-purpose computer languages, such as C or C++, which allow very fast computation, or in computer languages developed for AI programming, such as LISP or PROLOG, which often allow faster development or prototyping.

KNOWLEDGE REPRESENTATION AND AUTOMATED REASONING

A data structure commonly used to represent knowledge is the 'frame', which represents an entity that is being described by a number of attributes ('slots'). Frames can have procedures attached to them that determine how to process certain data. Also, frames can be linked by relationships, such as 'is a' or 'is a part of'. The special case of a supertype–subtype relationship between two frames is called 'inheritance' because the more specific subtype inherits components from the more general supertype. Frames are somewhat similar to 'objects' in object-oriented programming, which is a popular paradigm in the design and development of computer programs. Objects are parts of computer programs that facilitate modularity of large programs, reuse of code, and representation of real-life objects in software. Objects are formed from classes, patterns which determine the attributes (data stored in variables), the behavior (procedures, usually called 'methods'), and inheritance of the objects.

Frames are also somewhat similar to semantic nets. The nodes of the net represent the entities that are being described and the links between the nodes represent the relationships between these entities. Semantic nets have been used to model rich semantic relationships between entities such as 'agent', 'object', 'instrument' and 'recipient'.

In addition to facts, knowledge is often represented as if–then rules that allow reasoning based on the facts. The premises of a rule are called its left-hand side and the conclusions its right-hand side. If the pattern of the left-hand side of a rule can be matched with the facts in the knowledge base, the rule is activated ('fired'). The conclusions of the rule can establish new facts and activate new rules, thus cascading through a line of reasoning to a final conclusion.

Rule systems can be used in two different ways, which are called 'forward chaining' and 'backward chaining'. In forward chaining, the input consists of a set of facts describing a specific problem or situation that demands a solution or response. These facts activate rules, which again activate new rules, until the solution is found (a 'data-driven' approach). In backward chaining, the user first establishes a hypothesis or goal. The system tries to find a fact that matches the goal, and if no fact is found, it looks for a right-hand side of a rule. If a

right-hand side of a rule is found that matches the question, the system tries to match the left-hand side of this rule, and so on until finally all hypotheses are matched by facts (a 'goal-driven' approach). If a required fact cannot be found in the knowledge base, the interface may be used to query the user for the fact.

Often, expert systems have to use rules or facts that are not certain but can be stated only with a certain likelihood. One mathematical technique to handle uncertain information is probability theory. The probability of an event can be any number between 0, for an impossible event, and 1, for a certain event. Probabilities of facts and rules can be combined. Although probabilities can be based on calculations of the likelihood of random events, statistics, or even subjective judgments of experts, it often proves difficult to find good estimates of probabilities.

The most important alternative approach to the calculation of probabilities is the use of certainty factors. Certainty factors describe how confident human experts feel about facts and rules. This approach reflects the point that human experts often use rules of thumb and experience-based intuition rather than exact probabilities. Certainty factors can range between -1, for impossible events, and 1, for certain events. The mathematical formulas used to combine conclusions of rules that have certainty factors are simpler than methods using probabilities.

Finally, fuzzy logic, a mathematical theory developed by Lotfi Zadeh, provides a means of describing inexact knowledge and semantic terms such as 'very certain', 'tall', 'medium', or 'close to 18 years'. Exact values and fuzzy terms can be related to each other using membership functions. Systems based on fuzzy rules use procedures to combine fuzzy terms, define inexact rules, and perform reasoning based on these concepts.

KNOWLEDGE ENGINEERING

The development of a new expert system starts with a number of questions. The need for a particular kind of expert system needs to be established. Is human expertise scarce in a particular field, justifying the efforts of development? Is there a market for a new expert system; or will it help to save money and justify the investment in its production? Next, the feasibility of the expert system needs to be established. Can the task to be solved be limited in such a way that it will be not too large, not too difficult, and well enough understood? Do enough human experts exist, and are these experts willing to collaborate on the project? One problem is a possible need for common sense, which has never been successfully implemented in expert systems so far.

Once the need for and feasibility of the new expert system have been established, the knowledge engineer has to define the project's goal and the main problems and sub-problems. At least one human expert has to be found to be part of the team. Usually, an expert system is not built from scratch, but a tool, often called an expert system shell, is used. The right tool for the task at hand has to be chosen. The knowledge engineer interviews the human expert and turns the information gleaned into facts and rules that are appropriate for the expert system tool. Once a significant part of the expert system has been implemented, it needs to be tested to see if it produces the expected output. Usually, the human expert has to validate the answers produced by the new expert system.

It is important to have an experienced knowledge engineer on the team who knows the strengths and weaknesses of the various expert system tools and is able to avoid pitfalls. A clear separation of the roles of the knowledge engineer and the human expert is also important. Competent domain experts often find it difficult to describe their knowledge, especially the exact ways by which they solve problems. The knowledge engineer helps the domain expert to formalize the problem-solving methods and to separate the domain-specific knowledge from the system-specific reasoning strategies.

Expert systems are expensive and their development is time-consuming. Some expert systems have taken more than 30 person-years to build. As with other kinds of software, the total development time of a new expert system is reduced by rapid prototyping and early testing of the system, followed by further cycles of planning, knowledge acquisition, coding, and system evaluation.

EARLY EXPERT SYSTEMS AND DEVELOPMENT ENVIRONMENTS

The development of expert systems was pioneered in the 1960s at Stanford University (by Edward Feigenbaum) and at Carnegie-Mellon University (by Allen Newell and Herbert Simon). Among the first successful systems was MYCIN, which was developed at Stanford in the 1970s to diagnose serious bacterial infections in hospitalized patients and help to find the right antibiotic treatment. MYCIN used information about patient history, symptoms, and laboratory results, suggested a diagnosis, and recommended a drug treatment.

MYCIN was implemented in LISP, and used if–then rules and backward chaining. Uncertainty was handled using certainty factors. MYCIN reached the stage of a research prototype and proved the viability and usefulness of the concept of expert systems in narrow knowledge domains.

Subsequently, the domain knowledge of MYCIN was removed in order to create an expert system shell called EMYCIN. Since the inference engine used backward chaining, it was especially appropriate for diagnosis and classification tasks. EMYCIN has been used to develop several expert systems, including PUFF, which diagnosed lung diseases on the basis of pulmonary function data. PUFF was developed at Stanford and reached the stage of a production prototype. The development of PUFF required only 5 person-years, compared with 20 person-years for MYCIN, because of its use of an existing expert system shell.

TEIRESIAS was developed at Stanford as a research system. It was implemented in a LISP dialect. In order to integrate domain knowledge into an EMYCIN expert system, the knowledge had to be coded in an EMYCIN-specific, LISP-like code. TEIRESIAS helped with this process by analyzing rules, checking their completeness and consistency, and making suggestions for how to improve and troubleshoot them. It facilitated research in knowledge acquisition and expert system development and maintenance.

XCON was a commercial expert system developed as a collaboration between Carnegie-Mellon University and Digital Equipment Corporation (DEC). Configuring the DEC VAX family of computer systems was a difficult task that often led to costly mistakes. The expert system tool OPS5, developed at Carnegie-Mellon, was a rule-based system that used forward chaining and was based on the efficient Rete pattern-matching algorithm developed by Charles Forgy. XCON was developed using OPS5, and reached a production stage in the late 1970s. It grew over the following years to a size of over 3000 rules. It reached a performance level of 90 to 95 percent, while demonstrating that commercial viability was consistent with a limited rate of mistakes. Development and maintenance required four person-years per year, but the system saved DEC a significant amount of money by enabling faster and more accurate processing of new orders.

Following the success of the early expert systems, a number of expert system tools were developed by various commercial companies. These include ART, a rule-based knowledge engineering language developed by Inference Corporation. ART was written in LISP and supported forward and backward chaining and certainty factors. It featured a debugging aid and a graphical monitor. S.1, which was developed in a LISP dialect by Teknowledge, supported not only rule-based representations, but also frame-based and procedure-oriented methods. It supported certainty factors and graphical debugging tools. KEE, developed by Intellicorp in a LISP dialect, supported frames and both rule-based and procedure-based knowledge representations. It featured a modular knowledge base, forward and backward chaining, and a graphical debugging tool. These and other tools were used to develop a large number of expert systems in a variety of areas including agriculture, chemistry, computer systems, engineering, geology, law, financial analysis, mathematics, medicine, and physics.

Besides these commercial tools, free or inexpensive tools have been developed, many of which can be downloaded from the internet. The most important one is CLIPS, an OPS-like tool developed by NASA. CLIPS is implemented in C, it runs on many platforms, and its rules are coded in a LISP-like language. CLIPS supports only forward chaining and does not handle uncertainty. Extensions such as FuzzyCLIPS and AgentCLIPS have been developed, as well as an implementation in Java called JESS.

HYBRID SYSTEMS

Progress in other fields of AI suggested the combination of expert systems with other intelligent systems to create hybrid systems. The most common hybrid systems combine expert systems with artificial neural networks. Other useful hybrid systems are systems based on fuzzy rules and combinations of expert systems with genetic algorithms.

While much progress has been made in the areas of knowledge representation and automated reasoning, knowledge acquisition remains a bottleneck in the development of expert systems. Machine learning techniques can be used to generate knowledge from data with less involvement of a human expert. Induced decision trees are a machine learning technique pioneered by J. Ross Quinlan, who developed the ID3 and the C4.5 algorithms. Decision-tree systems form a hierarchical structure by dividing a data set according to the values of the most informative attribute. This process is repeated, and leads to a predictive rule-based model.

Artificial neural networks are another machine learning technology that was inspired by biological

brains. Neural networks are used for pattern extraction, classification, and prediction of quantitative data. Although they do not produce rules, neural networks equivalent to rule-based systems can be developed. Also, neural networks can be used in combination with knowledge-based systems. Because these two types of system have complementary characteristics, the use of human knowledge and symbolic reasoning of expert systems can fruitfully be combined with the pattern-oriented, data-based nature of neural networks and statistical techniques.

Case-based reasoning is an alternative technique, related to expert systems. Instead of creating rules based on expert knowledge, cases are stored using feature descriptors. Case-based reasoning resembles learning by analogy. The computer system searches through its store of cases and develops a solution based upon the most similar case in the case base. The main difficulties are in defining correct criteria for retrieving, matching, and applying similar cases.

Intelligent natural language processing systems are being used to improve the user interface of expert systems. The goal is to develop an interface that allows questions to be asked in English without the knowledge of a special syntax. The system should also be able to formulate answers in English. Teaching computers to understand natural languages is a difficult problem, which has only partially been solved. The various efforts in natural language processing and computational linguistics are beyond the scope of this article.

USE OF KNOWLEDGE-BASED SYSTEMS FOR DECISION SUPPORT AND ENTERPRISE SOFTWARE

Changes in computer system architecture have influenced the way corporate data are handled and how expert systems are integrated into enterprise software. While in the past a mainframe or minicomputer was connected to a number of dumb terminals, the dominant architecture now is a network system in which client computers have significant computing capabilities of their own, and can use programs and data stored on larger server machines for more complex tasks. In a distributed system, various database and application servers are connected on a network and can cooperate on tasks and share resources. The size and power of networks have been greatly increased by the advent of the internet and the subsequent development of intranets (networks restricted to one organization) and extranets (networks open to partner organizations but not the general public), and intelligence technologies have been a growing part of network operations and applications.

Decision support systems manage corporate data and knowledge to facilitate management decisions. The functionality of these systems has substantially increased over time. Transaction processing and other organizational functions create raw data that are stored in databases. These data can be queried, aggregated, analyzed, and visualized using spreadsheets, reports, graphs, and more sophisticated tools such as online analytical processing. Besides using databases, decision support tools usually contain modules for data modeling, visualization, simulation, and forecasting. Corporate knowledge derived from these large databases is stored in text-based systems and expert systems.

In recent years, decision support systems have tended to include data warehouses, where raw data from various sources are stored permanently, after cleaning, integration, and transformation of the data in order to support 'knowledge discovery'. Statistical and machine learning tools have been adapted to glean knowledge from these large amounts of data ('data mining'). Results take the form of classifications and rules with predictive power or clusters and patterns that are useful for analyzing and understanding the data. Rules derived from knowledge discovery in databases can be integrated into expert systems once they have been validated.

Knowledge has become one of the most important assets of business enterprises as well as of other large organizations. This has led to the concept of knowledge management, which includes the identification and analysis of knowledge and knowledge-related processes as well as the development of new knowledge strategies. In large organizations a common knowledge vocabulary needs to be developed and different knowledge sources and knowledge models need to be harmonized and integrated. A culture of knowledge sharing supported by software tools is essential in a context of rapid fluctuations of personnel, massive amounts of raw data, and organizations that are spread out over locations that are far apart.

Other examples of how expert systems can be integrated into mission-critical enterprise software are systems that monitor and control production processes. Data are typically collected through sensors connected to production machinery, then stored in databases and automatically analyzed, often using rule-based expert systems. Output can include diagnostic messages and alarms as well as feedback signals for tuning the production

machinery through motors. Often changes have to be made in a small time frame, which requires specialized real-time computer systems. An example is Gensym's G2 software, which offers a graphical, object-oriented environment that supports rule-based expert systems as well as real-time systems. It has visual programming and debugging tools, and other intelligent applications such as neural networks and genetic algorithms. G2 can be connected to other software and hardware and is typically used to monitor, diagnose, and control dynamic production environments or to perform simulations.

Distributed computer systems and the internet have created interest in an intelligent software technology called the 'intelligent agent'. An agent is a program that can take action on behalf of its user. It is mobile, and can travel through the network, perceiving its environment using sensors. It communicates with other agents by sending and receiving messages; it stores information; and it performs reasoning and takes action using 'effectors'. Expert systems are often a component of intelligent agents: the technologies developed for expert systems can support knowledge representation, reasoning, and communication in intelligent agents. The rapid development of the internet and its associated computer languages and standards (especially the 'extensible markup language' XML) will facilitate the development of agent systems. In business applications agents can collect information, compare offers, and even negotiate with other software agents.

STRENGTHS, WEAKNESSES, AND APPLICATIONS OF MODERN EXPERT SYSTEMS

Modern expert systems are different in certain ways from the systems in the late 1970s and 1980s. While many of the early expert systems were written in LISP and required special hardware, current expert systems are adapted to UNIX and Windows environments and are often implemented in general-purpose computer languages such as C or C++. Modern expert systems have several features that simplify development, such as graphical support, debugging and explanatory devices, and rule integrity checking. Systems that use the Rete algorithm for efficient pattern matching or PROLOG for backward chaining are still in use. Handling of uncertainty, and the use of fuzzy logic, are often supported. While most expert systems are rule-based, other paradigms exist. An example is the commercial Hugin software, which develops expert systems based on probabilistic networks, an extension of Bayesian mathematics due to Steffen Lauritzen and David Spiegelhalter. Integration with and embedding in other software are important features in today's internet-oriented client–server architectures.

A large number of vendors offer expert systems and products, ranging from small inexpensive systems to large systems for mission-critical enterprise software. The internet and recent publications are a good resource for further information in the rapidly changing field of computer software. Because of the importance of internet commerce, specialized expert systems in this area are popular. For example, a financial institution can use expert system technology for applications such as credit evaluation and fraud detection, as well as for web-based helpdesk applications and support for employees in call centers. Software support for customer relations management is one of the most important areas in electronic commerce today.

The strengths of expert systems lie in their ability to store and use human knowledge. Knowledge plays an important role in many fields, and is of a practical, heuristic nature, often based on many years of human experience. Expert systems have been shown to be efficient and reliable when used in well-described problem areas. The abundance of data in large organizational systems can be used only if the data are transformed into information and the information organized as knowledge and summarized as meta-knowledge. In this knowledge management enterprise, machine learning and automated knowledge discovery in databases can be combined effectively with expert systems. Concepts and techniques for knowledge representation and automated reasoning have been developed over many years. Modern rule-based systems are modular, allow efficient explanations of their results, and provide debugging tools.

Still, a number of problems with expert systems have not been solved, and weaknesses remain. One problem is the knowledge acquisition bottleneck. Building expert systems is still time- and personnel-intensive, although improvements have been achieved with more powerful expert system development tools. A second major problem (sometimes called 'brittleness') is that towards the boundaries of the expert knowledge, expert systems, rather than showing a gradual decline in performance, often show a sudden drop from good function to no function. Missing information and unexpected values cannot be handled well. Expert systems can work only in well-defined and limited areas. Unlike human experts, expert systems cannot fall

back on general principles and common sense, or create or evaluate new solutions. Expert systems do not learn by themselves but have to be updated by knowledge engineers. The handling of numeric data is often more efficiently managed by statistical or neural network techniques, although these can possibly be combined with expert systems.

AI techniques including expert systems were a major focus of research and development in the 1970s and 1980s, but the focus in the 1990s shifted to database and data-mining strategies, machine learning, internet-based technologies including internet commerce, and the integration of computer systems. The internet has made expert systems available to many users worldwide, and has created a need for automated helpdesk and information systems in the growing number of knowledge-based organizations. In the future, well-focused, quickly-developed expert systems will be needed, and expert systems will play a growing role in knowledge management and knowledge discovery in databases.

Further Reading

Cowell RG, Dawid AP, Lauritzen SL and Spiegelhalter DJ (1999) *Probabilistic Networks and Expert Systems*. Berlin, Germany: Springer-Verlag.

Durkin J (1994) *Expert Systems: Design and Development*. New York, NY: Macmillan.

Firebaugh M (1988) *Artificial Intelligence: A Knowledge-Based Approach*. Boston, MA: Boyd & Frazer.

Giarratano J and Riley G (1989) *Expert Systems: Principles and Programming*. Boston, MA: PWS Publishing.

Jackson P (1999) *Introduction to Expert Systems*. Harlow, UK: Addison-Wesley.

Kandel A (ed.) (1992) *Fuzzy Expert Systems*. Boca Raton, FL: CRC Press.

Kolodner JL (1993) *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Liebowitz J (ed.) (1997) *The Handbook of Applied Expert Systems*. Boca Raton, FL: CRC Press.

Medsker L (1995) *Hybrid Intelligent Systems*. Norwell, MA: Kluwer.

Medsker L and Liebowitz J (1994) *Design and Development of Expert Systems and Neural Networks*. New York, NY: Macmillan.

Waterman DA (1986) *A Guide to Expert Systems*. Reading, MA: Addison-Wesley.

Frame Problem, The

Intermediate article

Murray Shanahan, Imperial College, London, UK

CONTENTS

The frame problem is a difficulty that arises when formal logic is used to represent the effects of actions. The challenge is to avoid having to represent explicitly large numbers of common-sense facts about what does not change when an action occurs.

INTRODUCTION

The frame problem, in the strictly technical sense of the term coined by McCarthy and Hayes (1969), is a difficulty that arises when we attempt to formalize the effects of actions in mathematical logic. Put briefly, the problem is this. How is it possible to represent the changes brought about by an action without having to explicitly enumerate all the non-changes? For example, suppose I tell you that taking certain medication cures a given ailment. Then common sense licenses the conclusion that taking that medication doesn't cause the color of the walls to change, that it doesn't cause it to rain, doesn't cause the government to collapse, doesn't cause your neighbor to turn into a frog, and so on. Unfortunately, formal logic does not inherently license any of these perfectly reasonable implications. Somehow, our logical formalization of the effects of taking the medication has to have the apparatus for drawing such common-sense conclusions deliberately built into it.

This article will discuss the variety of ways by which this difficulty may be overcome. Some of these will employ non-monotonic forms of inference, while others will adhere to classical, monotonic logic. In either case, a viable solution will have to be robust in the presence of certain phenomena, such as concurrent actions, actions with nondeterministic effects, actions with indirect effects, and continuous change. This article will not address the question of the desirability or otherwise of using logic as a representational formalism in the first place; nor will it discuss the frame problem in the wider sense of the term that has become current among philosophers (Pylyshin, 1987).

HOW THE FRAME PROBLEM ARISES

Let us take a closer look at how the frame problem arises. We first need to settle on a formalism for representing the effects of actions. Since the most prevalent formalism in the literature is the 'situation calculus' (McCarthy and Hayes, 1969), this will be our initial choice. The choice of formalism has only a minor bearing on the way the problem manifests itself and the possible methods for addressing it. It is important to see through the details of the presentation to the underlying issue. On the other hand, without a precise description of the problem as our starting point, it would be easy to misunderstand the underlying issue altogether.

The language of the situation calculus includes actions, fluents, and situations. A fluent is any property whose value is subject to change, such as the color of a block, the location of a robot, or the height of a ball. The most basic kind of formula in the situation calculus expresses the fact that, in a situation that results from performing a given type of action, a given fluent holds. Consider the following example:

$$\forall s\ [\mathrm{Holds}(\texttt{Cured},\mathrm{Result}(\texttt{TakeDrug},s)) \leftarrow \neg\mathrm{Holds}(\texttt{Anemic},s)] \quad (1)$$

This formula says that, in a situation that results from performing the action `TakeDrug`, the fluent `Cured` will be true, provided the fluent `Anemic` does not hold beforehand. (Universal quantifiers that range over an entire formula will be dropped from now on.) In general, Result(e,s) denotes the situation that results from carrying out action e in situation s, while Holds(f,s) says that fluent f is true in situation s. Now suppose we include the following three formulae representing the fluents that hold in the initial situation S_0:

$$\neg\mathrm{Holds}(\texttt{Cured},S_0) \quad (2)$$

$$\neg\mathrm{Holds}(\texttt{Anemic},S_0) \quad (3)$$

$$\mathrm{Holds}(\texttt{Alive},S_0) \quad (4)$$

Naturally, it follows from the conjunction of formulae 1 to 4 that

$$\text{Holds}(\texttt{Cured},\text{Result}(\texttt{TakeDrug},S_0)) \quad (5)$$

However, it does not follow from these formulae that

$$\text{Holds}(\texttt{Alive},\text{Result}(\texttt{TakeDrug},S_0)) \quad (6)$$

To obtain this conclusion, we could augment our attempted formalization with the following formula:

$$\text{Holds}(\texttt{Alive},\text{Result}(\texttt{TakeDrug},s)) \leftarrow \text{Holds}(\texttt{Alive},s) \quad (7)$$

In other words, the fluent `Alive` is unchanged by the action `TakeDrug`.

This kind of formula, describing a non-effect of an action, is known as a *frame axiom*. The strategy of adding frame axioms will allow us to make the desired inferences. But clearly, to formalize any domain of reasonable size, a very large number of frame axioms will be required. Even to formalize this simple domain properly requires four further frame axioms:

$$\text{Holds}(\texttt{Cured},\text{Result}(\texttt{TakeDrug},s)) \leftarrow \text{Holds}(\texttt{Cured},s) \quad (8)$$

$$\text{Holds}(\texttt{Anemic},\text{Result}(\texttt{TakeDrug},s)) \leftarrow \text{Holds}(\texttt{Anemic},s) \quad (9)$$

$$\neg\text{Holds}(\texttt{Alive},\text{Result}(\texttt{TakeDrug},s)) \leftarrow \neg\text{Holds}(\texttt{Alive},s) \quad (10)$$

$$\neg\text{Holds}(\texttt{Anemic},\text{Result}(\texttt{TakeDrug},s)) \leftarrow \neg\text{Holds}(\texttt{Anemic},s) \quad (11)$$

In general, since most fluents are unaffected by most actions, the number of frame axioms required will be around $2 \times N \times M$, if there are N actions and M fluents in the domain. This is the frame problem, as it arises in the situation calculus. The challenge is to find a way to capture the non-effects of actions without having to use a large number of explicit frame axioms.

A natural reaction to the frame problem is to attribute it to the monotonicity of classical logic. A form of inference is 'monotonic' if the set of conclusions that follow from a collection of premises can only ever increase with the addition of new premises: once a conclusion has been drawn, it can never be retracted on the basis of new information. Classical logic is monotonic in this sense. So perhaps the frame problem arises because the monotonicity of classical logic makes it impossible to say: 'assume an action doesn't affect a fluent unless the contrary is stated explicitly'. With a non-monotonic inference method, the addition of a premise describing some new effect of an action may cause a previous default conclusion about its non-effects to be revised. So if we want to express the assumption above, a non-monotonic form of inference is required.

NON-MONOTONIC SOLUTIONS

Circumscription (McCarthy, 1986) is one form of non-monotonic inference. Formally, if Σ is a set of formulae, the formula CIRC[Σ ; P] denotes the 'circumscription' of Σ, minimizing the set P of predicates. The effect of 'minimizing a predicate' is to keep its extension – the set of things of which it is true – as small as possible. So the formulae that follow from CIRC[Σ ; P] are those that follow from Σ alone, together with a number of extra default conclusions that follow only if the extensions of the predicates in P are constrained in this way.

For example, suppose Σ comprises the formulae `Happy(Fred)` and `Happy(Jill)`. Then, although ¬`Happy(Albert)` does not follow from Σ alone, it does follow from CIRC[Σ ; `Happy`], since the circumscription constrains the extension of `Happy` to exclude `Albert`. Note that ¬`Happy(Albert)` no longer follows if we add `Happy(Albert)` to Σ. In other words, ¬`Happy(Albert)` does not follow from CIRC[$\Sigma \cup$ {`Happy(Albert)`} ; `Happy`]. So the set of consequences of the circumscription has diminished, although the set of premises has increased. This is the sense in which circumscription is a non-monotonic form of inference.

Let us now consider the application of circumscription to the frame problem. Can we use it to formalize the default assumption that the only effects of actions are the explicitly enumerated effects? One way to do this is to formalize the so-called 'common-sense law of inertia' (McCarthy, 1986), according to which change is abnormal, and persistence is to be expected in the absence of information to the contrary. A natural way to formalize this law in the context of the situation calculus is as follows:

$$[\text{Holds}(f,\text{Result}(e,s)) \leftrightarrow \text{Holds}(f,s)] \leftarrow \neg\text{Ab}(e,f,s) \quad (12)$$

Intuitively, Ab(e,f,s) means that action e is abnormal with respect to fluent f in situation s, in the sense that the value of f is subject to change when e is performed in s. Now, for this 'universal frame axiom' to have its desired effect, we must conjoin it to a description of the effects of actions in the

problem domain, and then circumscribe minimizing the predicate Ab. So, for the drugs example above, if we let Σ consist of formulae 1 to 4 and formula 12, we are interested in what follows from the circumscription CIRC[Σ ; Ab]. In this case, we have

$$\begin{aligned}&\mathrm{Ab}(e,f,s) \leftrightarrow \\ &\quad [e = \mathtt{TakeDrug} \wedge f = \mathtt{Cured} \\ &\quad \wedge \neg \mathrm{Holds}(\mathtt{Cured},s) \\ &\quad \wedge \neg \mathrm{Holds}(\mathtt{Anemic},s)]\end{aligned} \tag{13}$$

from which, by formula 12, it follows that

$$\mathrm{Holds}(\mathtt{Alive},\mathrm{Result}(\mathtt{TakeDrug},S_0)). \tag{14}$$

Unfortunately, the naive application of circumscription can lead to a difficulty known as the Hanks–McDermott problem, in which the circumscription yields two alternative ways of minimizing change, only one of which is intuitively correct (Hanks and McDermott, 1987). The problem is traditionally presented by means of a benchmark known as the Yale shooting scenario. In this scenario, there are three actions: `Load`, `Shoot`, and `Sneeze`. The `Load` action causes a gun to be loaded; the `Shoot` action causes the gun to be fired; and the `Sneeze` action makes no change. Formally, these actions affect two fluents: `Alive` and `Loaded`. In the initial situation, `Alive` holds. The initial situation and the effects of subsequent actions can be represented as follows:

$$\mathrm{Holds}(\mathtt{Alive},S_0) \tag{15}$$

$$\mathrm{Holds}(\mathtt{Loaded},\mathrm{Result}(\mathtt{Load},s)) \tag{16}$$

$$\begin{aligned}&\neg \mathrm{Holds}(\mathtt{Alive},\mathrm{Result}(\mathtt{Shoot},s)) \\ &\quad \leftarrow \mathrm{Holds}(\mathtt{Loaded},s)\end{aligned} \tag{17}$$

Now, let Σ consist of the formulae 12, 15, 16, and 17, and consider what follows from CIRC[Σ ; Ab]. In particular, consider the sequence of actions: `Load`, `Sneeze`, `Shoot`. Clearly `Loaded` follows immediately after the `Load` action, from axiom 16. And since we have incorporated the common-sense law of inertia, we might expect `Loaded` to continue to hold through the `Sneeze` action, thus ensuring that the `Shoot` action causes `Alive` to stop holding, by axiom 17. But this intuitive conclusion does not follow, because the minimization of the Ab predicate is ambiguous. In addition to the intuitive minimization, in which the only effects of actions are that `Load` modifies `Loaded` and `Shoot` modifies `Alive`, there is another equally valid minimization in which the `Sneeze` action miraculously unloads the gun, thus preventing the `Shoot` action from affecting the `Alive` fluent.

This scenerio is interesting because it is exemplary. The lesson is that the common-sense law of inertia, if formalized incorrectly, fails to rule out unintended effects of actions, whether in the context of circumscription or of other forms of default reasoning. Moreover, the Hanks–McDermott problem arises not only with the situation calculus, but also with other formalisms for representing actions, such as the event calculus (Shanahan, 1999b). In the late 1980s, many researchers devised techniques for overcoming the Hanks–McDermott problem, including chronological minimization (Shoham, 1988), which works by imposing temporal directionality on the minimization process, and causal minimization (Lifschitz, 1987), in which special causal predicates are introduced. Other researchers have produced effective solutions by modifying the original circumscriptive formalization (Baker, 1991). (In our description above, we have, for the sake of simplicity, ignored certain formal details, such as 'uniqueness of names' axioms and the question of which predicates are allowed to vary in circumscription.) All this work is surveyed in detail in Shanahan (1997).

FILTERING

With hindsight, two related approaches stand out for being both simple and robust in the presence of diverse phenomena such as nondeterminism and actions with indirect effects. These two approaches are called 'filtering' (Sandewall, 1994) and 'explanation closure' (Haas, 1987).

Although filtering can be successfully applied to the situation calculus (Kartha and Lifschitz, 1995), it is easiest to explain in the context of a formalism that employs explicit predicates for encapsulating the effects of actions, such as the event calculus (Kowalski and Sergot, 1986; Shanahan, 1999b). The idea of filtering is to isolate formulae describing the effects of actions, and apply default reasoning to them, separately from their application to a given scenario.

We will see how this works below, after briefly introducing the event calculus. An event calculus theory comprises two parts, a description of the effects of actions using the predicates 'Initiates' and 'Terminates', and a description of a narrative of events using the predicates 'Initially' and 'Happens'. These two sets of formulae are conjoined with some generic axioms to yield, as logical consequences, formulae describing the values of fluents at different time points. The formula Initiates(e,f,t) means that immediately following the occurrence of an action of type e at time t, the fluent

f starts to hold. Similarly, Terminates(e,f,t) means that immediately following an action of type e at t, the fluent f ceases to hold. The formula HoldsAt(f,t) means that fluent f holds at time t. So the effects of the actions in the Yale shooting scenario are represented as follows:

$$\text{Initiates}(\texttt{Load},\texttt{Loaded},t) \tag{18}$$

$$\text{Terminates}(\texttt{Shoot},\texttt{Alive},t) \leftarrow \text{HoldsAt}(\texttt{Loaded},t) \tag{19}$$

The narrative of events in the Yale shooting scenario can be given the following representation. The formula Initially(f) means that the fluent f is true in the initial situation, and the formula Happens(e,t) means that an action of type e occurs at time t.

$$\text{Initially}(\texttt{Alive}) \tag{20}$$

$$\text{Happens}(\texttt{Load},1) \tag{21}$$

$$\text{Happens}(\texttt{Sneeze},2) \tag{22}$$

$$\text{Happens}(\texttt{Shoot},3) \tag{23}$$

Now, to get results, we need some domain-independent axioms:

$$\text{HoldsAt}(f,t_2) \leftarrow \text{Initially}(f) \wedge \neg\exists\, e,t_1\ [\text{Happens}(e,t) \wedge t_1 < t_2 \wedge \text{Terminates}(e,f,t_1)] \tag{24}$$

$$\text{HoldsAt}(f,t_3) \leftarrow \text{Happens}(e_1,t_1) \wedge \text{Initiates}(e_1,f,t_1) \wedge t_1 < t_3 \wedge \neg\exists e_2,t_2[\text{Happens}(e_2,t_2) \wedge t_1 < t_2 < t_3 \wedge \text{Terminates}(e_2,f,t_2)] \tag{25}$$

$$\neg\text{HoldsAt}(f,t_3) \leftarrow \text{Happens}(e_1,t_1) \wedge \text{Terminates}(e_1,f,t_1) \wedge t_1 < t_3 \wedge \neg\exists\, e_2,t_2[\text{Happens}(e_2,t_2) \wedge t_1 < t_2 < t_3 \wedge \text{Initiates}(e_2,f,t_2)] \tag{26}$$

In the context of the situation calculus, we tried to tackle the frame problem by minimizing the Ab predicate. The analogous thing to do with the event calculus is to minimize the Initiates and Terminates predicates – this corresponds to the default assumption that there are no unknown effects of actions. In addition, the Happens predicate must be minimized – this corresponds to the default assumption that there are no unknown action occurrences. But if we naively conjoin all the relevant formulae (18 to 26), and circumscribe them minimizing these predicates, the Hanks–McDermott problem arises. We get just the same unintended possibilities as we did with the Yale shooting scenario in the situation calculus.

Fortunately, a straightforward technique will overcome this problem. We just need to isolate the various components of the formalization, and circumscribe them separately. This technique is known as filtering (Sandewall, 1994), since the idea is to filter out those parts of the formalization that should not be subject to default reasoning. This ensures that the assignment of fluents to actual time points cannot interfere with the process of minimizing the effects of actions. In the case of the Yale shooting scenario, this means that the `Shoot` action has the effect of terminating `Alive` if `Loaded` holds – and this conditional effect is retained in the minimization of the Terminates predicate, irrespective of which fluents actually hold at what times. This effect cannot be replaced by a miraculous unloading by the `Sneeze` action. Formally, if we let EC be the conjunction of formulae 24, 25, and 26, Σ consist of formulae 18 and 19, and Δ consist of formulae 20 to 23, then we are interested in

$$\text{CIRC}[\Delta\ ;\ \text{Happens}] \wedge \text{CIRC}[\Sigma\ ;\ \text{Initiates},\ \text{Terminates}] \wedge \text{EC} \tag{27}$$

This, at last, yields the desired results, and among the logical consequences of this theory we have HoldsAt(`Loaded`,3) and ¬HoldsAt(`Alive`,4). This technique works for any simple theory describing the effects of actions with preconditions. Moreover, it can be used for theories involving concurrent actions, actions with nondeterministic effects, and actions with indirect effects (Shanahan, 1997). Solving the frame problem in the context of actions with indirect effects is the most challenging of these tasks, and is known as the ramification problem. We will return to the ramification problem below.

EXPLANATION CLOSURE

A crucial property of the circumscription CIRC[Σ ; Initiates, Terminates] is that it reduces, in the case of the Yale shooting scenerio, to the predicate completions of Initiates and Terminates (Clark, 1978), because of the simple form of Σ (Lifschitz, 1994). That is:

$$\text{Initiates}(e,f,t) \leftrightarrow [e = \texttt{Load} \wedge f = \texttt{Loaded}] \tag{28}$$

$$\text{Terminates}(e,f,t) \leftrightarrow [e = \texttt{Shoot} \wedge f = \texttt{Alive} \wedge \text{HoldsAt}(\texttt{Loaded},t)] \tag{29}$$

The above formulae, which are in biconditional ('if and only if') form, directly encode the assumption that the only effects of actions are the known effects. So it is natural to ask whether we can use this style of formula directly. Instead of writing a

set of formulae describing the effects of actions and then circumscribing them, why not omit the circumscription process altogether and proceed directly to a set of biconditional formulae? One of the main attractions of this approach to the frame problem is that it altogether avoids the use of non-monotonic formalisms, with their attendant mathematical difficulties.

A similar maneuver is possible for the situation calculus, as was first noted by Haas (1987). Instead of using a form of non-monotonic reasoning, we can simply rewrite the frame axioms in a more parsimonious form, known as *explanation closure axioms*. Consider a slightly extended version of the Yale shooting scenario, in which the `Shoot` action unloads the gun, and in which there is also an `Unload` action with the same effect. We then have the following effect axioms in addition to formulae 16 and 17:

$$\neg\text{Holds}(\texttt{Loaded},\text{Result}(\texttt{Shoot},s)) \tag{30}$$

$$\neg\text{Holds}(\texttt{Loaded},\text{Result}(\texttt{Unload},s)) \tag{31}$$

The following two explanation closure axioms say, respectively, that the only way to terminate the `Alive` fluent is by a `Shoot` action while `Loaded` holds, and that the only way to terminate the `Loaded` fluent is by either a `Shoot` action or an `Unload` action:

$$[\text{Holds}(\texttt{Alive},s) \wedge \neg\text{Holds}(\texttt{Alive}, \text{Result}(e,s))] \rightarrow [e = \texttt{Shoot} \wedge \text{Holds}(\text{Loaded},s)] \tag{32}$$

$$[\text{Holds}(\texttt{Loaded},s) \wedge \neg\text{Holds}(\texttt{Loaded}, \text{Result}(e,s))] \rightarrow [e = \texttt{Shoot} \vee e = \texttt{Unload}] \tag{33}$$

Two further explanation closure axioms concern the conditions under which the fluents are initiated. Clearly, explanation closure axioms subsume frame axioms: taken together with formulae 15, 16, and 17, the above axioms yield the expected conclusions for the Yale shooting scenario. However, since only two explanation closure axioms are required per fluent, the total number required for a domain of N actions and M fluents is around $2 \times M$. The length of an individual explanation closure axiom will typically increase with the size and complexity of the domain. But this increase will reflect the amount of change being captured, not the amount of non-change. This is why explanation closure axioms, though strictly monotonic, offer a parsimonious alternative to frame axioms.

An effect axiom and an explanation closure axiom can be combined into a single formula, known as a *successor state axiom* (Reiter, 1991). One such formula is required for each fluent, defining exactly the circumstances under which that fluent does and does not hold. Here, for example, are the successor state axioms for the Yale shooting scenario with an `Unload` action:

$$\text{Holds}(\texttt{Alive},\text{Result}(e,s)) \leftrightarrow [\text{Holds}(\texttt{Alive},s) \wedge \neg[e = \texttt{Shoot} \wedge \text{Holds}(\texttt{Loaded},s)]] \tag{34}$$

$$\text{Holds}(\texttt{Loaded},\text{Result}(e,s)) \leftrightarrow [e = \texttt{Load} \vee [\text{Holds}(\texttt{Loaded},s) \wedge e \neq \texttt{Unload} \wedge e \neq \texttt{Shoot}]] \tag{35}$$

One possible criticism of monotonic solutions to the frame problem is that they lack elaboration-tolerance. A formalism is elaboration-tolerant to the extent that it can easily accommodate the addition of new information (McCarthy, 1988). In the case of a formalism for reasoning about action, we would expect it to smoothly absorb new facts about the effects of actions – both previously unknown effects of known actions and effects of previously unknown actions. Using a non-monotonic formalism, such as circumscription, the assimilation of new information is simply a matter of conjoining it to the old theory and recircumscribing. By contrast, adding new information to a set of successor state axioms requires the dismantling and reconstruction of the whole theory. It is possible, however, to introduce elaboration tolerance by formally defining operators that take a set of standard situation calculus effect axioms and return the corresponding successor state axioms (Reiter, 1991). Such an operator is itself non-monotonic in the same sense as the circumscription operator.

THE SCOPE AND LIMITS OF A SOLUTION

One of the lessons of the Hanks–McDermott problem is that no formalization can be assumed to yield the expected results without a proper investigation of its properties. The Hanks–McDermott problem arises when the process of projecting fluents along a narrative line is permitted to interfere with the process of non-monotonically minimizing the effects of actions. We can be confident that the Hanks–McDermott problem has been solved when these two processes are separated, as they are in filtering and (implicitly) in explanation closure. But how can we establish rigorous criteria for the class of problem scenarios to which a proposed solution to the frame problem can be applied?

Two approaches to this question are 'action description languages' (Gelfond and Lifschitz, 1993) and the 'features and fluents' framework (Sandewall, 1994).

An action description language is a formal language for describing the effects of actions, together with a corresponding entailment relation. The idea is that the action description language should be used as a standard against which other formalisms couched in general-purpose logic are measured. Many such languages have been designed, to handle various phenomena.

Similarly, in the features and fluents framework, a hierarchical class of domains is defined, together with formal criteria for demonstrating that an action formalism correctly captures the expected logical consequences in that domain. Sandewall's classes range from domains that are restricted to deterministic actions with no concurrency or indirect effects, to domains allowing concurrent actions, continuous change, and actions with nondeterministic effects.

Many of these phenomena can be handled by existing formalisms, including the situation calculus, the event calculus, and others (Sandewall, 1994; Shanahan, 1997). But the issue that has proved most difficult is the ramification problem, which is the challenge of supplying a solution to the frame problem that works in the presence of actions with indirect effects. One way to represent the indirect effects of actions is through state constraints. A state constraint, in effect, rules out certain simultaneous combinations of fluents. For example, in the situation calculus, we might write the following state constraint:

$$\mathrm{Holds}(\texttt{Walking},s) \rightarrow \mathrm{Holds}(\texttt{Alive},s) \quad (36)$$

This formula rules out the possibility of the `Walking` fluent holding while the `Alive` fluent does not hold. So if an action has the direct effect of terminating `Alive`, it has the indirect effect of terminating `Walking`. The explanation closure approach to the frame problem can be extended to handle state constraints. The above formula, for example, in combination with the effect axiom 17, can be incorporated into a successor state axiom for the `Walking` fluent as follows (McIlraith, 2000):

$$\begin{aligned}&\mathrm{Holds}(\texttt{Walking},\mathrm{Result}(e,s)) \leftrightarrow \\ &\quad [\mathrm{Holds}(\texttt{Walking},s) \wedge \neg[e = \texttt{Shoot} \\ &\quad \wedge \mathrm{Holds}(\texttt{Loaded},s]] \end{aligned} \quad (37)$$

However, state constraints can lead to unexpected conclusions. For example, if we include a `Walk` action that initiates the `Walking` fluent, then we cannot avoid the conclusion that after the sequence of actions '`Load`, `Shoot`, `Walk`', the `Alive` fluent holds. In other words, the `Walk` action will bring about the unexpected resurrection of the victim of the `Shoot` action. The reason for this anomaly is that the relationship between `Alive` and `Walking` has a causal dimension – any action that terminates `Alive` indirectly causes `Walking` to be terminated – and state constraints cannot capture the directionality of the cause–effect relation. The issue is made more complicated by the possibility of concurrent actions with multiple, interacting indirect effects (Lin, 1995; Thielscher, 1997). However, many researchers have supplied formalisms that successfully address both the frame problem and the ramification problem, even in such extreme scenarios, by capturing the causal character of the effects of actions (e.g., McCain and Turner, 1997; Shanahan, 1999a).

References

Baker AB (1991) Nonmonotonic reasoning in the framework of the situation calculus. *Artificial Intelligence* **49**: 5–23.

Clark KL (1978) Negation as failure. In: Gallaire H and Minker J (eds) *Logic and Databases*, pp. 293–322. New York, NY: Plenum Press.

Gelfond M and Lifschitz V (1993) Representing action and change by logic programs. *Journal of Logic Programming* **17**: 301–322.

Haas AR (1987) The case for domain-specific frame axioms. In: Brown F (ed.) *Proceedings of the 1987 Workshop on the Frame Problem*, pp. 343–348.

Hanks S and McDermott D (1987) Nonmonotonic logic and temporal projection. *Artificial Intelligence* **33**: 379–412.

Kartha GN and Lifschitz V (1995) A simple formalization of actions using circumscription. In: Mellish CS (ed.) *Proceedings IJCAI 95*, pp. 1970–1975.

Kowalski RA and Sergot MJ (1986) A logic-based calculus of events. *New Generation Computing* **4**: 67–95.

Lifschitz V (1987) Formal theories of action. In: Brown F (ed.) *Proceedings of the 1987 Workshop on the Frame Problem*, pp. 35–57.

Lifschitz V (1994) Circumscription. In: Gabbay DM, Hogger CJ and Robinson JA (eds) *The Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. III, *Nonmonotonic Reasoning and Uncertain Reasoning*, pp. 297–352. Oxford, UK: Oxford University Press.

Lin F (1995) Embracing causality in specifying the indirect effects of actions. In: Mellish CS (ed.) *Proceedings IJCAI 95*, pp. 1985–1991.

McCain N and Turner H (1997) Causal theories of action and change. In: Pollack ME (ed.) *Proceedings AAAI 97*, pp. 460–465.

McCarthy J (1986) Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence* **26**: 89–116.

McCarthy J (1988) Mathematical logic in artificial intelligence. *Daedalus* **117**(1): 297–311.

McCarthy J and Hayes PJ (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Michie D and Meltzer B (eds) *Machine Intelligence*, vol. IV, pp. 463–502. Edinburgh, UK: Edinburgh University Press.

McIlraith S (2000) Integrating actions and state constraints: a closed-form solution to the ramification problem (sometimes). *Artificial Intelligence* **116**: 87–121.

Pylyshyn ZW (ed.) (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex.

Reiter R (1991) The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In: Lifschitz V (ed.) *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy* pp. 359–380. San Diego, CA: Academic Press.

Sandewall E (1994) *Features and Fluents: The Representation of Knowledge about Dynamical Systems*, vol. I. New York, NY: Oxford University Press.

Shanahan M (1997) *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA: MIT Press.

Shanahan M (1999a) The ramification problem in the event calculus. In: Dean T (ed.) *Proceedings IJCAI 99*, pp. 140–146.

Shanahan M (1999b) The event calculus explained. In: Woolridge MJ and Veloso M (eds) *Artificial Intelligence Today*, pp. 409–430. Berlin, Germany: Springer-Verlag.

Shoham Y (1988) *Reasoning About Change: Time and Causation From the Standpoint of Artificial Intelligence*. Cambridge, MA: MIT Press.

Thielscher M (1997) Ramification and causality. *Artificial Intelligence* **89**: 317–364.

Further Reading

Hanks S and McDermott D (1987) Nonmonotonic logic and temporal projection. *Artificial Intelligence* **33**: 379–412.

Reiter R (2001) *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. Cambridge, MA: MIT Press.

Fuzzy Logic

Intermediate article

Ronald R Yager, Machine Intelligence Institute, Iona College, New Rochelle, New York, USA

CONTENTS

Fuzzy logic provides a methodology that enables the modeling of human reasoning.

INTRODUCTION

Humans have a remarkable capability to perform a wide variety of physical and mental tasks without precise measurements. Examples of such tasks are driving a car, deciding what to wear, and playing many types of sport. In performing such tasks, humans mainly use imprecise information, perceptions, rather than precise readings. Some statements involving this kind of perception-based information are: the weather will turn cold in the early afternoon; he appears intent on attacking; and the vehicle is moving rapidly. The use of perceptions, in addition to being necessitated by the finite ability of humans to resolve detail, facilitates the acquisition, expression, and manipulation of knowledge. A fundamental aspect of many types of perceptions is their granularization and fuzziness (gradularization). In considering an attribute such as temperature, a perception would be it is *cold* rather than a precise measurement such as 20°F. One clearly sees the granular nature of cold in that it involves a collection or clump of temperatures. The fuzzy nature of perceptions is reflected in the fact that the common use of the word *cold* does not entail a crisp boundary distinguishing cold from not cold but more naturally reflects a gradual decay at its boundaries. The modeling of human cognition requires the manipulation of perceptions of many different kinds of variables. Physical variables such as time, direction, speed, and shape as well as intellectual constructs such as probability, causality, and stability are but a few examples. Fuzzy logic and the related idea of approximate reasoning is a technology expressly developed to help in this task. The idea of a fuzzy subset was originally introduced by L. A. Zadeh in 1965 (Zadeh, 1965). In Yager *et al*. (1987) the editors provide a selection of seminal papers by Zadeh. A collection of the pioneering papers on fuzzy logic by various authors can be found in Dubois *et al*. (1993). Recent works have focused on the development of a computational theory for reasoning with perceptions, and a framework for computing with words (Zadeh, 1999, 2001). A number of books provide good introductions to this subject (see Further Reading).

BASIC IDEAS AND CONCEPTS

The concept of a fuzzy subset was originally introduced by Zadeh as a generalization of the idea of an ordinary set. A fuzzy subset A on the universe X is defined by a function called the membership function. For each x the membership function A (x) can assume any value in $[0, 1]$, the unit interval, indicating the degree of membership of x in A. This extension from the classic binary/crisp case, in which the membership grade is restricted to be either zero or one, allows for the description of concepts in which the boundary between satisfying and not satisfying the concept is not sharp. Here we shall define some of the fundamental operations and concepts of fuzzy set theory and fuzzy logic.

Basic Operations on Fuzzy Sets

Here we begin with some of the basic operations. We shall assume A and B are two fuzzy subsets of X and use them to describe these basic fuzzy set operations:

- A is said to be a subset of B, denoted $A \subset B$, if $B(x) \geq A(x)$ for all $x \in X$
- The union A and B is a fuzzy subset $C = A \cup B$ such that $C(x) = \text{Max}[A(x), B(x)]$
- The intersection A and B is a fuzzy subset $D = A \cap B$ such that $D(x) = \text{Min}[A(x), B(x)]$
- The negation of A is a fuzzy subset $\bar{A}$ such that $\bar{A}(x) = 1 - A(x)$

It is easily seen that these operations reduce to the ordinary set operations when we restrict the membership grades to be one or zero. The operations we have introduced are the standard definitions of these operations. A lively literature exists in this field describing alternative ways of extending the set operations to the fuzzy domain.

Measuring Fuzziness

Since the concept of a fuzzy set was first introduced, various authors have attempted to define a measure of fuzziness of a fuzzy subset A, **FUZ**(A). DeLuca and Termini (1972) stated conditions which a measure of fuzziness should satisfy. Among these conditions are that $\mathbf{FUZ}(\mathrm{A}) = 0$ if A is a crisp set and **FUZ**(A) is maximum for a fuzzy subset where $\mathrm{A}(x) = 0.5$ for all x. Yager (1979) suggested that fuzziness is related to the lack of distinction between a set and its negation. Based on this he suggested

$$\mathbf{FUZ}(\mathrm{A}) = \frac{1}{n^{1/2}} \left(\sum_{i=1}^{n} |\mathrm{A}(x_i) - \overline{\mathrm{A}(x_i)}|^2 \right)^{1/2}.$$

Other measures for fuzziness have been suggested in the literature; among these is one suggested by Kosko (1986).

Linguistic Values and Computing with Words

A concept which plays a crucial role in many applications of fuzzy logic is the idea of a linguistic value. Consider the statement: Mary is *young*. This statement involves the assignment of a value to a variable. The variable is *the Age of Mary* which we shall denote as V. Associated with V is a set X, called its universe of discourse, consisting of the set of allowable values for the variable, $\mathrm{X} = \{1, 2...., 100\}$. The value assigned to V is *young*. In our case, rather than knowing that Mary's age is precisely one particular element in the set X, we are presented with the imprecise information, perception, that her age is *young*. Young in this framework is what is called a linguistic value. The concept of *young* can in turn be defined as a fuzzy subset A. In this fuzzy subset, the membership grade $\mathrm{A}(x)$ is the compatibility of the age x with the concept of *young*. The fuzzy subset A can be viewed as a constraint on the allowable value of the variable. Here we are viewing information as constraints on the allowable values of variables. A common notational convention is to use the form V *is* A to express this association of a value with a variable.

Approximate reasoning, sometimes called fuzzy logic or computing with words, provides a unified structure in which human reasoning can be modeled. This reasoning system uses the idea of a linguistic value and its representation as a fuzzy subset as a central knowledge representation tool. The theory of approximate reasoning has two fundamental components, 'translation rules' and 'inference rules'. The translation rules provide a mechanism for the translation of natural language statements into fuzzy subsets. This translation step can be seen as a bridge between human expression of information and the formal representation required for computerized manipulation of this information. The inference component of approximate reasoning provides rules for the manipulation of knowledge, which allow us to make inferences and in other ways process information.

Possibility and Certainty

Our representation of this type of linguistic information using fuzzy sets allows us to process this information formally. One type of processing involves answering questions. When one has information such as V *is* x^* and is asked if some other statement V *is* B is true, one can calculate 'the degree of truth' as the membership grade of x^* in B, $\mathrm{B}(x^*)$. In situations in which our knowledge involves a linguistic value, a fuzzy subset, the question of determining the truth is not as straightforward. Consider the knowledge *John is in his twenties*. If we are asked is John 33, we can easily answer no. If we are asked is John over fifteen, we can easily answer yes. But if we are asked is John 25, the best we can say is that it is possible but we are not certain. In order to handle this situation, where our information has a granular nature, two measures are introduced. We shall assume that A and B are fuzzy subsets of the set X corresponding to some linguistic concepts. Assume we have the knowledge that V *is* A and we are interested in confirmation of the statement V *is* B. One measure of this confirmation is called the *possibility* of B given A. It is denoted as Poss[B/A] and defined as $\mathrm{Poss}[\mathrm{B/A}] = \mathrm{Max}_x[\mathrm{D}(x)]$ where $\mathrm{D} = \mathrm{A} \cap \mathrm{B}$. A second measure of this confirmation is called the *certainty* of B given A. It is denoted as Cert[B/A] and defined as $\mathrm{Cert}[\mathrm{B/A}] = 1 - \mathrm{Poss}[\bar{\mathrm{B}}/\mathrm{A}]$. We can view the possibility and certainty measures as providing optimistic and pessimistic measures on the confirmation of B given A. The possibility and certainty measures can also be viewed as providing upper and lower bounds on the truth of B given A.

Information in a Fuzzy Set

When using linguistic values one becomes concerned with the issue of measuring the amount of information contained in the associated possibility distribution. For example, knowing that Rachel is 27 is more informative than knowing that she is in her twenties. In the framework of probability theory, the Shannon entropy is commonly used to measure the amount of information contained in a probability distribution. In fuzzy logic the measure of specificity can be used in an analogous manner. Assume A is a fuzzy subset corresponding to a linguistic value, the specificity of A, denoted Sp(A), can be defined as the difference between the maximal membership in A minus the average membership grade in A, $\mathrm{Sp(A)} = \mathrm{Max}_x(\mathrm{A}) - \mathrm{Ave}_x(\mathrm{A})$. A fundamental feature of this measure is that it attains its biggest value when A just contains one element with membership grade one and all other elements have membership zero. Another feature is that it decreases in value as membership grades increase.

PROTOTYPICAL APPLICATIONS OF FUZZY LOGIC

While fuzzy logic has found numerous applications, many of these involve two prototypical methodologies. The first of these is fuzzy systems modeling and the second is criteria satisfaction. In the following we describe these prototypical applications.

Fuzzy Systems Modeling

Fuzzy systems modeling is a technique for modeling complex nonlinear relationships using a rule-based methodology. Central to this approach is a partitioning of the model into fuzzy subregions in which we have knowledge of the model's performance. Consider a relationship $\mathrm{U} = f(\mathrm{V}, \mathrm{W})$. Here V and W are the input variables and U is called the output variable. Typically we are interested in obtaining the value of U for a given value of V and W. In fuzzy systems modeling we represent this relationship by a collection of fuzzy if–then rules of the form:

If V is *Large* and W is *About 25* then U is *Medium*

At a formal level we have a collection of rules expressed as:

If V is A_i and W is B_i then U is D_i

where the A_is, B_is and D_is are fuzzy subsets over the universes X, Y and Z. In using fuzzy systems modeling we are essentially partitioning the input space $\mathrm{X} \times \mathrm{Y}$ into fuzzy regions $\mathrm{A}_i \times \mathrm{B}_i$ in which we know the output value, $\mathrm{D_i}$.

Given values for the input variables, $\mathrm{V} = a$ and $\mathrm{W} = b$, we calculate the value of U as a fuzzy subset E by using a process called fuzzy inference:

1. For each rule we find the firing level $\lambda_i = \mathrm{Min}[\mathrm{A}_i(a), \mathrm{B}_i(b)]$.
2. We calculate the effective output of each rule $\mathrm{E_i}$.
3. Combine individual effective rule outputs to get overall system output E.

Two different paradigms have been typically used for implementing steps two and three in the above procedure. The first paradigm, called the Min-Max inference procedure, uses $\mathrm{E}_i(z) = \mathrm{Min}[\lambda_i, \mathrm{D}_i(z)]$ for the effective rule outputs. It then uses a union of these outputs to get the overall output $\mathrm{E} = \cup_{i=1}^{n} \mathrm{E}_i$, hence $\mathrm{E}(z) = \mathrm{Max}_i[\mathrm{E}_i(z)]$. The second paradigm uses arithmetic operations instead of the Min-Max operation. In this approach $\mathrm{E}_i(z) = \lambda_i * \mathrm{D}_i(z)$ and $\mathrm{E}(z) = \frac{1}{\mathrm{T}} \sum_{i=1}^{n} \mathrm{E}_i(z)$ where $\mathrm{T} = \sum_{i=1}^{n} \lambda_i$. As a simplified expression of this we have $\mathrm{E}(z) = \sum_{i=1}^{n} \mathrm{E}_i(z)$ where $\mathrm{E}_i(z) = w_i * \mathrm{D}_i(z)$ with $w_i = \frac{\lambda_i}{\mathrm{T}}$. We shall call this the arithmetic inference procedure.

When we desire a crisp output value z* rather than a fuzzy one we use a defuzzification step such as the center of area (COA) method where we calculate

$$z^* = \frac{\sum_z z\mathrm{E}(z)}{\sum_z \mathrm{E}(z)}$$

The most commonly used approach to fuzzy systems modeling, due to Takagi and Sugeno (1985) and called the TSK method, involves the use of the arithmetic method with a further simplification in the model. In this approach the outputs of the rules rather than being fuzzy sets are assumed to be specific values. In this case a rule becomes *If V is A_i and W is B_i then U is d_i*. Here d_i is a specific value from the space Z. This assumption leads to a simplification of the process of obtaining the output so that the output is obtained directly as

$$z^* = \sum_{i=1}^{n} w_i * d_i.$$

Criteria Satisfaction

The second major fuzzy technology focuses on criteria satisfaction problems. This technology has found applications in areas such as decision-

making, database-querying, and internet search and information retrieval.

Assume we have a collection of criteria which we desire to satisfy by selecting an action from a set of alternative actions X. The starting point of this approach is the representation of each of the criteria by a fuzzy subset. Specifically, a criterion can be represented by a fuzzy subset C_i in which the membership grade $C_i(x)$ indicates the degree to which alternative x satisfies the criterion. In this environment we let $D(x) = \text{Agg}(C_1(x), C_2(x), \ldots, C_n(x))$ be the fuzzy subset representing the overall satisfaction of alternative x to the collection of criteria. In some applications, such as decision-making, we select the alternative with the largest membership in D. In other applications, such as information retrieval, we provide an ordered listing of the objects based on their membership in D.

A central problem here is the formulation of the function D. The structure of the function Agg is a reflection of the desired relationship between the criteria. In many cases the relationship between the criteria can be most conveniently expressed by the decision maker in terms of natural language. In these environments we are able to take advantage of the strong connection between fuzzy logic and natural language to obtain formal representations of the decision function. If we desire to satisfy *all* the criteria, the decision function can be linguistically expressed as requiring C_1 *and* C_2 *and* C_3 *and*… *and* C_n. Using the definition for the *anding* (intersection) of fuzzy subsets we obtain $D(x) = \text{Min}_i[C_i(x)]$. In some situations, rather than requiring that all the criteria be satisfied an agent may only require that some portion of the criteria be satisfied. For example, one may find a solution acceptable if it satisfies *most* of the criteria. We can express this as $D = \textit{Most}[C_1, C_2, \ldots, C_n]$. The introduction of a quantifier such as *most* provides a softening of the decision function by allowing for a relaxation to the satisfaction of some of the criteria while still obtaining a good solution. More complex quantifiers-based decision rules can be envisioned, for example All of $[C_1, \ldots, C_k]$ and some of $[C_{k+1}, \ldots, C_n]$. A considerable body of literature has focused on the modeling of complex relationships between criteria. The Ordered Weighted Averaging (OWA) operator (Yager, 1988) has played a major role in this task.

SUMMARY

Human cognition uses information that is often imprecise. Fuzzy logic provides a formal framework for the modeling and manipulation of this kind of information. A fundamental idea in the use of fuzzy logic is the idea of the linguistic value and its representation as a fuzzy set.

References

DeLuca A and Termini S (1972) A definition of a non-probabilistic entropy in the setting of fuzzy sets. *Information and Control* **20**: 301–312.

Dubois D, Prade H and Yager RR (1993) *Readings in Fuzzy Sets for Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Kosko B (1986) Fuzzy entropy and conditioning. *Information Sciences* **40**: 165–174.

Takagi T and Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* **15**: 116–132.

Yager RR (1979) On the measure of fuzziness and negation part I: membership in the unit interval. *International Journal of General Systems* **5**: 221–229.

Yager RR (1988) On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics* **18**: 183–190.

Yager RR, Ovchinnikov S, Tong R and Nguyen H (eds) (1987) *Fuzzy Sets and Applications: Selected Papers by L. A. Zadeh*. New York, NY: John Wiley.

Zadeh LA (1965) Fuzzy sets. *Information and Control* **8**: 338–353.

Zadeh LA (1999) From computing with numbers to computing with words–From manipulation of measurements to manipulations of perceptions. *IEEE Transactions on Circuits and Systems* **45**: 105–119.

Zadeh LA (2001) Toward a logic of perceptions based on fuzzy logic. In: Novak W and Perfilieva I (eds) *Discovering the World with Fuzzy Logic*, pp. 4–28. Heidelberg: Physica-Verlag.

Further Reading

Kecman V (2001) *Learning and Soft Computing*. Cambridge, MA: MIT Press.

Klir GJ and Yuan B (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ: Prentice Hall.

Kosko B (1993) *Fuzzy Thinking*. New York, NY: Hyperion.

Kosko B and Isaka S (1993) Fuzzy logic. *Scientific American* **269**: 76–81.

Pedrycz W and Gomide F (1998) *An Introduction to Fuzzy Systems*. Cambridge, MA: MIT Press.

Slowinski R (1998) *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*. Norwell, MA: Kluwer Academic Publishers.

Yager RR and Filev DP (1994) *Essentials of Fuzzy Modeling and Control*. New York, NY: John Wiley.

Zadeh LA and Kacprzyk J (1999) *Computing with Words in Information/Intelligent Systems 1*. Heidelberg: Physica-Verlag.

Zimmermann HJ (2001) *Fuzzy Set Theory and its Applications*. Dordrecht: Kluwer Academic.

Game-playing Programs

Intermediate article

Susan L Epstein, Hunter College and The Graduate Center of The City University of New York, New York, USA

CONTENTS

Game-playing programs rely on fast, deep search and knowledge to defeat human champions. For more difficult games, simulation and machine learning have been employed, and human cognition is under consideration.

GAME TREES

Board games are not only entertaining – they also provide us with challenging, well-defined problems, and force us to confront fundamental issues in artificial intelligence: knowledge representation, search, learning, and planning. Computer game playing has thus far relied on fast, deep search and vast stores of knowledge. To date, some programs have defeated human champions, but other challenging games remain to be won.

A *game* is a noise-free, discrete space in which two or more agents (*contestants*) manipulate a finite set of objects (*playing pieces*) among a finite set of locations (the *board*). A *position* is a world state in a game; it specifies the whereabouts of each playing piece and identifies the contestant whose turn it is to act (the *mover*). Examples appear in Figure 1. Each game has its own finite, static set of rules that specify legal locations on the board, and when and how contestants may *move* (transform one state into another). The rules also specify an *initial state* (the starting position for play), a set of *terminal states* where play must halt, and assign to each terminal state a *game-theoretic value*, which can be thought of as a numerical score for each contestant.

As in Figure 1, the search space for a game is typically represented by a *game tree*, where each node represents a position and each link represents one move by one contestant (called a *ply*). A *contest* is a finite path in a game tree from an initial state to a terminal state. A contest ends at the first terminal state it reaches; it may also be terminated by the rules because a time limit has been exceeded or because a position has repeatedly occurred.

The goal of each contestant is to reach a terminal state that optimizes the game-theoretic value from its perspective. An *optimal move* from position p is a move that creates a position with maximal value for the mover in p. In a terminal state, that value is determined by the rules; in a non-terminal state, it is the best result the mover can achieve if subsequent play to the end of the contest is always optimal. The game-theoretic value of a non-terminal position is the best the mover can achieve from it during error-free play. If a subtree stops at states all of which are labeled with values, a *minimax algorithm* backs those values up, one ply at a time, selecting the optimal move for the mover at each

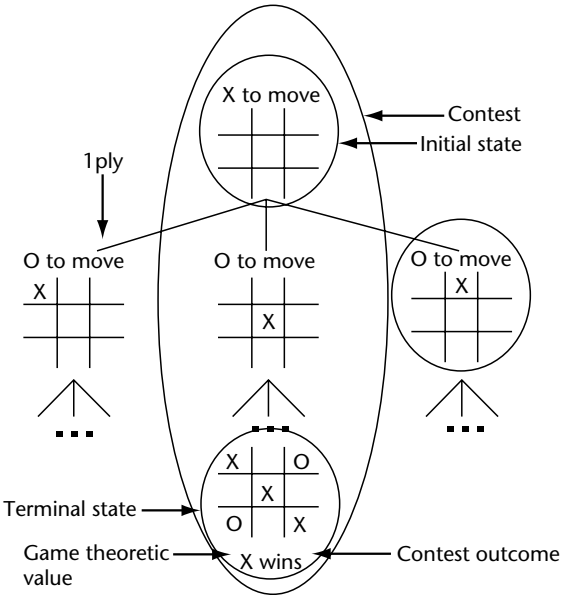


Figure 1. A game tree and basic game-playing terminology.

node. In Figure 2, for example, each possible next state in tic-tac-toe is shown with its game-theoretic value; minimax selects the move on the left.

Retrograde analysis backs up the rule-determined values of all terminal nodes in a subtree to compute the game-theoretic value of the initial state. It minimaxes from all the terminal nodes to compute the game-theoretic value of every node in the game tree, as shown in Figure 3. The number of nodes visited during retrograde analysis is dependent both on a game's *branching factor* (average number of legal moves from each position) and the depth of the subtree under consideration. For any challenging game, such as checkers (draughts) or chess, retrograde analysis to the initial state is computationally intractable. Therefore, move selection requires a way to compare alternatives.

An *evaluation function* maps positions to values, from the perspective of a single contestant. A *perfect* evaluation function preserves order among all positions' game-theoretic values. For games with relatively small game trees, one can generate a perfect evaluation function by caching the values from retrograde analysis along with the optimal moves. Alternatively, one might devise a way to compute a perfect evaluation function from a description of the position alone, given enough knowledge about the nature of the game. In this approach, a position is described as a set of *features*, descriptive properties such as piece advantage or control of the center. It is possible, for example, to construct, and then program, a perfect, feature-based evaluation function for tic-tac-toe. Given a perfect evaluation function, a game-playing program searches only one ply – it evaluates all possible next states and makes the move to the next state with the highest value. For a challenging game, however, the identity of the features and their relative importance may be unknown, even to human experts.

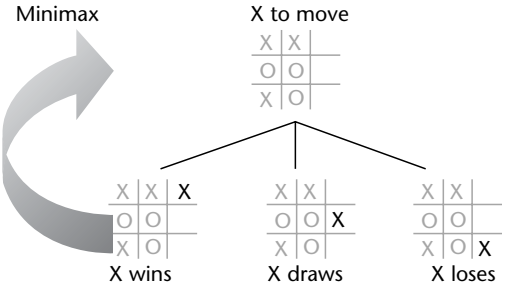


Figure 2. A minimax algorithm selects the best choice for the mover.

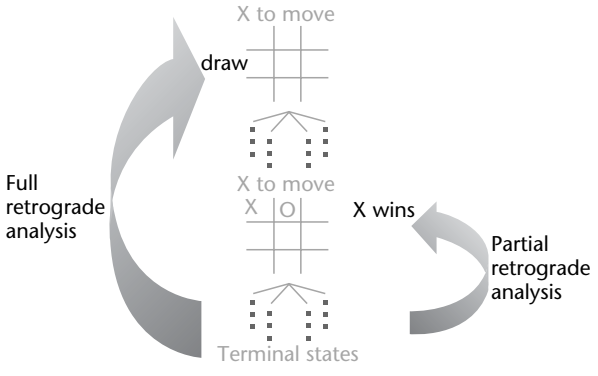


Figure 3. Retrograde analysis backs up rule-determined values.

SEARCH AND KNOWLEDGE

Confronted with a large game tree and without a perfect evaluation function, the typical game-playing program relies instead on heuristic search in the game tree. The program searches several ply down from the current state, labels each game state it reaches with an estimate of its game-theoretic value as computed by a heuristic evaluation function, and then backs those values up to select the best move. Most classic game-playing programs devote extensive time and space to such heuristic search. The most successful variations preserve exhaustive search's correctness: a *transposition table* to save previously evaluated positions, the α-β algorithm to *prune* (not search) irrelevant segments of the game tree, extensions along promising lines of play, and extensions that include forced moves. Other search algorithms take conservative risks; they prune unpromising lines early or seek *quiescence*, a relatively stable heuristic evaluation in a small search tree. Whatever its search mechanisms, however, a powerful game-playing program typically plays only a single game, because it also relies on knowledge.

Knowledge is traditionally incorporated into a game-playing program in three ways. First, formulaic behavior early in play (*openings*) is prerecorded in an *opening book*. Early in a contest, the program identifies the current opening and continues it. Second, knowledge about features and their relative importance is embedded in a heuristic evaluation function. Finally, prior to competition, the program calculates the true game-theoretic values of certain nodes with exhaustive search and stores them with their optimal moves (*endgame database*). Because a heuristic evaluation function always returns any available endgame values, the larger that database, the more accurate the evaluation and the better the search is likely to perform.

EARLY ATTEMPTS AT MECHANIZED GAME PLAYING

Chess has long been the focus of automated game playing. The first known mechanical game player was for a chess endgame (king and rook against king), constructed about 1890 by Torrès y Quevedo. In the 1940s many researchers began to consider how a computer might play chess well, and constructed specialized hardware and algorithms for chess. Work by Shannon, Turing, and de Groot was particularly influential. By 1958 a program capable of playing the entire game was reported, and by the mid-1960s computers had begun to compete against each other in tournaments (Marsland, 1990).

At about the same time, Samuel was laying the foundation for today's ambitious game-playing programs, and much of machine learning, with his checkers player (Samuel, 1959, 1967). A game state was summarized by his program in a vector of 38 feature values. The program searched at least three ply, with deeper searches for positions associated with piece capture and substantial differences in material. The checker player stored as many evaluated positions as possible, reusing them to make subsequent decisions. Samuel tested a variety of evaluation functions, beginning with a prespecified linear combination of the features. He created a compact representation for game states, as well as a routine to learn weighted combinations of 16 of the features at a time. Samuel's work pioneered rote learning, generalization, and coevolution. His program employed α-β search, tuned its evaluation function to book games played by checker masters, constructed a library of moves learned by rote, and experimented with nonlinear terms through a signature table. After playing 28 contests against itself, the checker program had learned to play tournament-level checkers, but it remained weaker than the best human players.

For many years it was the chess programs that held the spotlight. The first match between two computer programs was played by telegraph in 1967, when a Russian program defeated an American one 3–1. Although they initially explored a variety of techniques, the most successful chess programs went on to demonstrate the power of fast, deep game-tree search. These included versions of Kaissa, MacHack 6, Chess 4.6, Belle, Cray Blitz, Bebe, Hitech, and a program named Deep Thought, the precursor of Deep Blue. As computers grew more powerful, so did chess-playing programs, moving from Chess 3.0's 1400 rating in 1970 to Deep Blue's championship play in 1997.

BRUTE FORCE WINS THE DAY

Brute force is fast, deep search plus enormous memory directed to the solution of a problem. In checkers and in chess, brute force has triumphed over acknowledged human champions. Both programs had search engines that rapidly explored enormous subtrees, and supported that search with extensive, efficient opening books and endgame databases. Each also had a carefully tuned, human-constructed, heuristic evaluation function, with features whose relative importance were well understood in the human expert community.

In 1994, Chinook became the world's champion checker player, defeating Marion Tinsley (Schaeffer, 1997). Its opening book included 80,000 positions. Its 10-gigabyte endgame database, constructed by exhaustive forward search, included about 443 billion positions, every position in which no more than 8 pieces (checkers or kings) remain on the board. The frequency with which Chinook's search reached these game-theoretic values was in large measure responsible for the program's success.

In 1997 Deep Blue defeated Garry Kasparov, the human chess champion. Deep Blue's custom chess-searching hardware enabled it to evaluate 200 million moves per second, sometimes to depths over 30 ply. In the year immediately before its victory, the program benefited from a substantial infusion of grandmaster-level knowledge, particularly in its evaluation function and its opening book. Deep Blue's endgame database included all chess positions with five or fewer pieces, but it was rarely reached.

SIMULATION AND MACHINE LEARNING, THE ALTERNATIVES

There are, however, games more difficult than chess, games where programs require more than brute force to win. Consider, for example, Shogi and Go, played on the boards in Figures 4(a) and 4(b), respectively. Although the branching factor for chess is 35, for Shogi it is 80–150, and for Go it is 250. Such a large branching factor makes deep search intractable. Games with very long contests also reduce the opportunity for search to reach an endgame database, where the evaluation function would be perfect. For example, the typical checkers contest averages about 100 moves, but the typical Go contest averages more than 300. In games that include imperfect information (e.g. a concealed hand of cards) or nondeterminism (e.g. dice), the

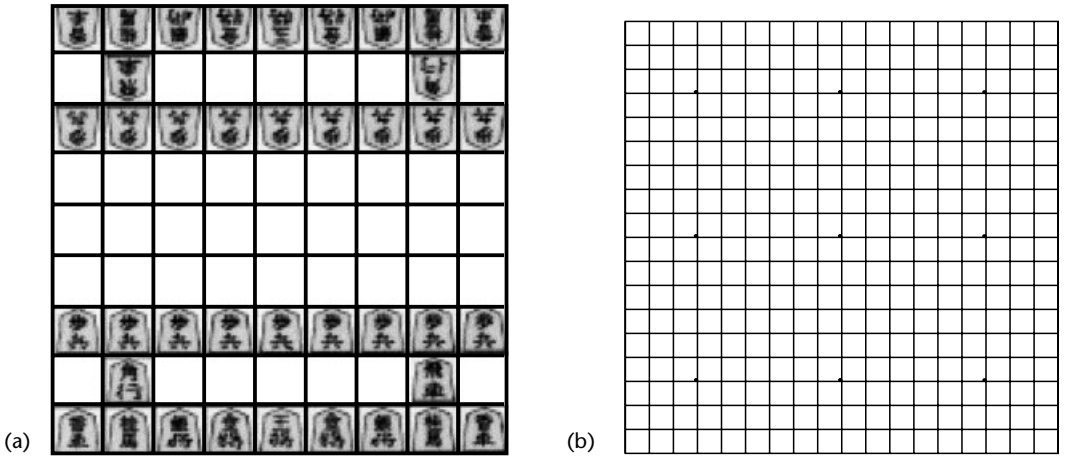


Figure 4. (a) The starting position in Shogi. (b) The Go board.

brute force approach represents each possibility as a separate state. Once again, the branching factor makes deep search intractable. In bridge, for example, after bidding the declarer can see 26 cards, but there are more than 20 million ways the other 26 cards may be distributed between the opponents' hands.

In a game with a very large branching factor, rather than search all possibilities exhaustively, a program can sample the game tree by simulation. Simulation generates the unknown information (e.g. the opponents' hands) at random and evaluates game states based on that assumption. Since a single random guess is unlikely to be correct, simulation is repeated, typically thousands of times. The evaluation function is applied to each state resulting from the same move in the simulated trees, and averaged across them to approximate the goodness of a particular move. Simulation can be extended as many ply as desired. Maven, for example, plays Scrabble®, a game in which contestants place one-letter tiles into a crossword format. Scrabble® is nondeterministic because tiles are selected at random, and it involves imperfect information because unplayed tiles are concealed. Nonetheless, Maven is considered the best player of the game, human or machine (Sheppard, 1999). Instead of deep search, Maven uses a standard, game-specific move generator (Appel and Jacobson, 1988), a probabilistic simulation of tile selection with three-ply search, and the B* search algorithm in the endgame.

When people lack the requisite expert knowledge, a game-playing program can learn. A program that learns, executes code that enables it to process information and reuse it appropriately. Rather than rely upon the programmer's knowledge, such a program instead acquires knowledge that it needs to play expertly, either during competition (*on-line*) or in advance (*off-line*). A variety of learning methods have been directed toward game playing: rote memorization of expert moves, deduction from the rules of a game, and a variety of inductive methods. An approach that succeeds for one game does not necessarily do well on another. Thus a game-learning program must be carefully engineered.

A game-playing program can learn openings, endgame play, or portions of its evaluation function. Openings are relatively accessible from human experts' play. For example, Samuel's checkers player acquired a database of common moves on-line, and Deep Blue learned about grandmasters' openings off-line. Endgame database computations are costly but relatively straightforward; Chinook's endgame database, learned off-line, was essential to its success. In a game whose large branching factor or lengthy contests preclude deep search, however, an endgame database is rarely reached during lookahead.

Machine learning for game playing often focuses on the evaluation function. TD-gammon is one of the world's strongest backgammon players. The mover in backgammon rolls a pair of dice on every turn; as a result, the branching factor is 400, precluding extensive search. TD-gammon models decision-making with a neural network whose weights are acquired with temporal difference learning in millions of contests between two copies of the program. Given a description of the position with human-supplied features, the neural net serves as an evaluation function; during competition, TD-gammon uses it to select a move after a two- to three-ply search (Tesauro, 1995).

Ideally, a program could learn not only weights for its evaluation function, but also the features it references. Logistello plays Othello (Reversi); in 1997 it defeated Takeshi Murakami, the human world champion, winning all 6 contests in the match (Buro, 1998). Logistello's heuristic evaluation function is primarily a weighted combination of simple patterns that appear on the board, such as horizontal or diagonal lines. (Which player has the last move and how far a contest has progressed are also included.) To produce this evaluation function, 1.5 million weights for elaborate conjunctions of these features were calculated with gradient descent during off-line training, from analysis of 11 million positions. Although it uses a sophisticated search algorithm and a large opening book, Logistello's evaluation function is the key to its prowess. Its creator supplied the raw material for Logistello's evaluation function, but the program learned features produced from them, and learned weights for those features as well.

COGNITION AND GAME-PLAYING PROGRAMS

Although no person could search as quickly or recall as accurately as a champion program, there are some aspects of these programs that simulate human experts. A good human player remembers previous significant experiences, as if the person had a knowledge base. A good human player expands the same portion of a game tree only once in a contest, as if the person had a transposition table. A good human player has a smaller, but equally significant, opening book and recognizes and employs endgame knowledge.

There are also features of human expertise that programs generally lack. People plan, but planning in game playing has not performed as well as heuristic search. People narrow their choices, but simulation or exhaustive search, at least for the first few ply, have proved more reliable for programs. People construct a model of the opposition and use it to guide decision-making, but most programs are oblivious of their opposition. People have a variety of rationales for decisions, and are able to offer explanations for them, but most programs have opaque representations. Skilled people remember *chunks* (unordered static spatial patterns) that could arise during play (Chase and Simon, 1973), but, at least for chess programs, heuristic search ultimately proved more powerful. Finally, many people play more than one game very well, but the programs described here can each only play a single game. (One program, Hoyle, learns to play multiple games, but their game trees are substantially smaller than that for chess.)

The cognitive differences between people and programs become of interest in the face of games, such as shogi and Go, games that programs do not yet play well at all. These games do not yield readily to search. Moreover, the construction of a powerful evaluation function for these games is problematic, since even the appropriate features are unknown. In shogi, unlike chess, there is no human consensus on the relative strength of the individual pieces (Beal and Smith, 1998). In Go there are thousands of plausible features (often couched in proverbs) whose interactions are not well understood. Finally, because the endgame is likely to have at least as large a branching factor as earlier positions, the construction of a useful endgame database for either game is intractable. Although both games have attracted many talented researchers and have their own annual computer tournaments, no entry has yet played either game as well as a strong amateur human.

Timed photographs of a chess player's brain demonstrate that perception is interleaved with cognition (Nichelli *et al.*, 1994). Although Go masters do not appear to have chunks as originally predicated (Reitman, 1976), there is recent evidence that these people do see dynamic patterns and readily annotate them with plans. Moreover, Go players' memories now appear to be cued to sequences of visual perceptions. As a result, despite their inferiority for chess programs, work in Go continues to focus on patterns and plans. Another promising technique, foreshadowed by the way human experts look at the Go board, is decomposition search, which replaces a single full search with a set of locally restricted ones (Muller, 1999).

The challenges presented by popular card games, such as bridge and poker, have also received attention recently. Both involve more than two contestants and include imperfect information. Bridge offers the challenge of pairs of collaborating opponents, while poker permits tacit alliances among the contestants. At least one bridge program has won masters' points in play against people, relying on simulation of the concealed card hands. Poker pits a single contestant simultaneously against many others, each with an individual style of play. Poki plays strong Texas Hold'em poker, although not as well as the best humans (Billings *et al.*, 1999). The program bases its bets on probabilities, uses simulation as a search device, and has begun to model its opponents.

Finally, a synergy can develop between game-playing programs and the human experts they

simulate. Scrabble® and backgammon both provide examples. Maven has hundreds of human-supplied features in its evaluation function. The program learned weights for those features from several thousand contests played against itself. Since their 1992 announcement, Maven's weights have become the accepted standard for both human and machine players. Meanwhile, TD-gammon's simulations, known as *rollouts*, have become the authority on the appropriateness of certain play. In particular, human professionals have changed their opening style based on data from TD-gammon's rollouts.

SUMMARY

Game-playing programs are powerfully engineered expert systems. They often have special-purpose hardware, and they employ concise representations designed for efficiency. Where the branching factor permits, a game-playing program relies on fast, deep, algorithmic search, guided by heuristics that estimate the value of alternative moves. If that is not possible, simulation is used to determine a decision. Champion programs play a single game, and benefit from vast stores of knowledge, knowledge either provided by people or learned by the programs from their experience. Nonetheless, challenging games remain, games that humans still play best.

References

Appel AW and Jacobson GJ (1988) The world's fastest Scrabble program. *Communications of the ACM* **31**(5): 572–578.

Beal D and Smith M (1998) *First Results from Using Temporal Difference Learning in Shogi*. Proceedings of the First International Conference on Computers and Games, Tsukuba, Japan.

Billings D, Pena L, Schaeffer J and Szafron D (1999) *Using Probabilistic Knowledge and Simulation to Play Poker*. Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp. 697–703.

Buro M (1998) *From Simple Features to Sophisticated Evaluation Functions*. Proceedings of the First International Conference on Computers and Games, Tsukuba, Japan.

Chase WG and Simon HA (1973) The mind's eye in chess. In: Chase WG (ed.) *Visual Information Processing*, pp. 215–281. New York, NY: Academic Press.

Marsland TA (1990) A short history of computer chess. In: Marsland TA and Schaeffer J (eds) *Computers, Chess, and Cognition*, pp. 3–7. New York, NY: Springer-Verlag.

Muller M (1999) *Decomposition Search: A Combinatorial Games Approach to Game Tree Search, with Applications to Solving Go Endgames*. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 578–583, Stockholm.

Nichelli P, Grafman J, Pietrini P *et al*. (1994) Brain activity in chess playing. *Nature* **369**: 191.

Reitman JS (1976) Skilled perception in Go: deducing memory structures from inter-response times. *Cognitive Psychology* **8**: 336–356.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 210–229.

Samuel AL (1967) Some studies in machine learning using the game of checkers. II – recent progress. *IBM Journal of Research and Development* **11**: 601–617.

Schaeffer J (1997) *One Jump Ahead: Challenging Human Supremacy in Checkers*. New York, NY: Springer-Verlag.

Sheppard B (1999) Mastering Scrabble. *IEEE Intelligent Systems* **14**(6): 15–16.

Tesauro G (1995) Temporal difference learning and TD-Gammon. *CACM* **38**(3): 58–68.

Further Reading

Berlekamp ER, Conway JH and Guy RK (1982) *Winning Ways for Your Mathematical Plays*. London, UK: Academic Press.

Conway JH (1976) *On Numbers and Games*. New York, NY: Academic Press.

Fürnkranz J and Kubat M (eds) (2001) *Machines that Learn to Play Games*. Huntington, NY: Nova Science.

Holding D (1985) *The Psychology of Chess Skill*. Hillsdale, NJ: Lawrence Erlbaum.

Hill-climbing Search

Intermediate article

Bart Selman, Cornell University, Ithaca, New York, USA
Carla P Gomes, Cornell University, Ithaca, New York, USA

CONTENTS

Many computational and cognitive tasks involve a search through a large space of possible solutions. Hill-climbing, or local search, is one strategy for searching such a solution space.

LOCAL VERSUS GLOBAL SEARCH

There are many problems that require a search of a large number of possible solutions to find one that satisfies all the constraints. As a basic example consider the problem of composing a classroom schedule given a set of constraints on the availability of classrooms, and various time constraints involving students and lecturers. Even with a relatively simple set of constraints, one may be forced to search through many possible schedules in order to find one that satisfies all constraints. In certain cases, computer scientists have found clever algorithms to solve such computational problems without having to search through the space of all potential solutions. However, in many other cases, such a search cannot be avoided. We refer to the search space as a *combinatorial* search space. A key question is what is the best way of searching a combinatorial space. The answer often depends on the underlying problem to be solved. In order to illustrate the two main techniques for searching a combinatorial space, we consider the N-queens problem as an example. (*See* **Search**; **Constraint Satisfaction**; **Problem Solving**)

The N-queens problem is that of placing N queens on an N by N chess board, so that no two queens can attack each other. Carl Friedrich Gauss considered this problem in its original form on an 8 by 8 chess board (Campbell, 1977). Given that a queen can move horizontally in its row, it follows that we can have at most one queen in each row. In fact, because we need to place N queens, a solution will require us to place exactly one queen in each row. Similarly, because of the movement of a queen in its column, the placement of queens is such that there is exactly one queen in each column. What makes the problem difficult is the fact that queens can also move diagonally. Therefore, we have to find a placement with exactly one queen per row and column where no two queens share a diagonal.

To search for a solution to the N-queens problem, there are two fundamentally different techniques. Both techniques search through the space of placements of N queens on a chess board, but in dramatically different ways.

One strategy is referred to as a *global* (or *systematic*) search strategy. In this strategy, a solution is constructed incrementally. That is, we place one queen at a time, starting with one in the first row, then one in the second row etc. For each placement, we look for a position in the row that is not under attack from any of the previously placed queens. One difficulty is that after placing several queens, we may be unable find such an 'open' position in a row (i.e. one that is not being attacked). When we encounter such a situation, say in row i, we need to go back to row $i-1$, and shift the queen in that row to another open position in the row. We may find that even with the queen in the new position in row $i-1$, we still cannot place a queen in row i. We then again shift the queen in the $(i-1)^{\text{th}}$ row to another open position. We may, of course, run out of open positions to move to in row $i-1$, in which case we have to revisit the placement of the queen in row $i-2$, etc. The process of shifting previously placed queens is called *backtracking*. We literally revise or 'backtrack' on our earlier placement choices. Such a search technique will eventually search the full space of all possible placements of queens. So, if a solution exists, it will eventually be found. Unfortunately, the search space is exponentially large and a backtrack technique can therefore be quite inefficient. Using sophisticated heuristics to try the 'most promising' available positions first, one can solve the N-queens problem for up to around 100 queens using backtrack search (Stone and Stone, 1987).

Hill-climbing or *local search* provides a very different search strategy for this problem. In a local

search approach, we first start with a 'random' guess at a solution, for example by placing a queen in each row, where the position within a row is chosen randomly. Because of the random placements, it is quite likely that we have one or more pairs of queens that can attack each other either because they share a column or a diagonal. This means that our initial placement does not yet give us a valid solution to the N-queens problem. We now select one of the queens on the board that is under attack from one or more other queens. We move this queen to another location in its row, giving preference to the positions that are not attacked by any other queen. If all positions in a row are under attack, we select a position that is under attack from the least number of queens. If after shifting a queen, which is referred to as a 'local move' or 'local modification', we still have queens under attack, we again select a queen randomly from the ones that are under attack and move that queen, each time trying to further minimize the number of conflicts on the board. This basic strategy is surprisingly effective. It can solve the N-queens problem for over a million queens in less than a minute on a standard PC (Sosic and Gu, 1991; Minton *et al*., 1992).

In general, the key benefits of hill-climbing search are that it requires only a limited amount of memory (only the current state is stored), it is easy to implement efficiently, and, if one or more solutions exists in the search space, hill-climbing search can be surprisingly effective at finding it. Perhaps the first successful use of a hill-climbing strategy was that of Lin and Kernighan (1973; Lin, 1965) for finding good solutions for the traveling salesperson problem. In this problem, the goal is to find the shortest route for a salesperson to take to visit a given set of cities. Starting with an arbitrary tour that visits the cities, Lin and Kernighan showed how one can quickly reduce the length of the tour by making a series of local changes to the route. Since the work by Lin and Kernighan, local search has become one of the main practical techniques for tackling combinatorial optimization problems (Papadimitriou and Steiglitz, 1982).

An important drawback of local search is its inability to detect the unsolvability of a problem instance. That is, if no solution exists, a local search method will simply continue to make local modifications indefinitely. (When dealing with an optimization problem, such as the traveling salesperson problem, the difficulty is that local search cannot be used to determine whether the solution found is globally optimal.) In principle, one could memorize all previously visited problem states and force the local search method to never explore states it has explored before. Provided the local modifications are general enough, such a search would eventually explore the full search space underlying the problem. However, given the combinatorial nature of these problems, this is not feasible on instances of practical interest. Interestingly, for global search techniques, there are memory-efficient ways of keeping track of the space explored so far. Therefore, global search techniques, in contrast to local search methods, can tell us that no solution exists after the method has explored the full search space and no solution has been found.

LOCAL SEARCH STRATEGIES

In hill-climbing search, we are optimizing a certain objective function. For example, in our N-queens problem, our objective function is $O(board) = (N^2/2) - L$, where L is the number of pairs of queens that attack each other on the current board. ($N^2/2$ is the number of pairs of queens.) A solution corresponds to finding a board with the largest possible value for the objective function, i.e. $N^2/2$.

In hill-climbing search, we select any local change that improves the current value of the objective function. *Greedy local search* is a form of hill-climbing search where we select the local move that leads to the largest improvement of the objective function. Traditionally, one would terminate hill-climbing and greedy search methods when no local move could further improve the objective function. Upon termination, the search would have reached a local, but not necessarily global, optimum of the objective function. For many optimization problems, such as the traveling salesperson problem, such a local optimum is quite acceptable, since it often is a reasonable approximation of the global optimum value. However, when a globally optimal solution is required – such as in our N-queens example – local optima present a serious problem for local search methods.

In recent years, it has been found, perhaps somewhat surprisingly, that simply allowing the local search to continue, by accepting 'sideway' or even 'downhill' moves, i.e. local moves to states with, respectively, the same or worse objective values, one can often eventually still reach a global optimum. For example, such 'non-improving' moves are a key component of local search methods for the Boolean satisfiability (SAT) problems, such as GSAT and Walksat (Selman *et al*., 1992, 1994; Gu, 1992). These local search methods for SAT

outperform the more traditional backtrack search strategies for a number of SAT problem classes. Many combinatorial problems can be encoded effectively as Boolean SAT problems, so these local search algorithms provide a general mechanism for solving combinatorial problems.

The local search framework allows for a large number of different realizations, and literally hundreds of variants have been explored over the years. The main variants are tabu search, simulated annealing, and genetic algorithms.

Simulated annealing (Kirkpatrick *et al*., 1983) is another example of a local search technique that incorporates sideway and downhill moves. In simulated annealing, downhill moves are accepted with a probability based on the size of the change in the objective function, with the 'worst' moves becoming the least likely. The fraction of accepted local changes is also controlled by a formal parameter, called the 'temperature'. At a high temperature, almost any possible local move is accepted. When lowering the temperature parameter, fewer moves are accepted that worsen the objective value, and the search starts to resemble a purely greedy strategy. In simulated annealing, inspired by the physical process of annealing, the temperature starts high and is slowly lowered during the search process. Simulated annealing has been successfully applied in a wide range of applications.

In tabu search (Glover, 1989), a 'tabu' list is added to the local search method. The tabu list contains the most recent local moves; the local search method is prevented from undoing those modifications. The tabu list is of limited length, and is updated after each local move. Therefore, any particular change is only prevented for a limited period of time. A tabu list provides a powerful way of forcing a local search method to explore a larger part of the search space.

Another form of local search can be found in so-called genetic algorithms (Holland, 1992). Genetic algorithms are inspired by the natural selection process encountered in evolution. A genetic algorithm can be viewed as a strategy for running a large number of local searches in parallel. Aside from local modifications, the states are also modified through a process called 'crossover', in which states from different local searches are combined to provide a (hopefully) better starting point for a new local search.

Hill-climbing search also has a long tradition in the area of continuous optimization. In continuous search spaces, one generally uses the gradient of the objective function to take local steps in the direction of the greatest possible improvement. There are many refinements on the basic gradient search techniques, such as dynamically varying the local step size during the search (Miller, 2000).

In conclusion, hill-climbing search provides a powerful strategy for exploring combinatorial search spaces. In many applications, hill-climbing outperforms global search. In recent work on randomizing backtrack search methods (Gomes *et al*., 2000), we see how some of the ideas from local search are being used to boost the performance of global search methods.

References

Campbell PJ (1977) Gauss and the eight queens problem: a study in miniature of propagation. *Historia Mathematica* **4**: 397–404.

Glover F (1989) Tabu search – part I. *ORSA Journal on Computing* **1**(3): 190–206.

Gomes C, Selman B, Crato N and Kautz H (2000) Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *Journal of Automated Reasoning* **24**(1/2): 67–100.

Gu J (1992) Efficient local search for very large-scale satisfiability problems. *Sigart Bulletin* **3**(1): 8–12.

Holland JH (1992) *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Kirkpatrick S, Gelatt CD and Vecchi MP (1983) Optimization by simulated annealing. *Science* **220**: 671–680.

Lin S (1965) Computer solutions of the traveling salesman problem. *BSTJ* **44**(10): 2245–2269.

Lin S and Kernighan BW (1973) An effective heuristic for the traveling-salesman problem. *Operational Research* **21**: 498–516.

Miller RE (2000) *Optimization*. New York, NY: John Wiley.

Minton S, Johnston M, Philips AB and Laird P (1992) Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems. *Artificial Intelligence* **58**: 161–205.

Papadimitriou CH and Steiglitz K (1982) *Combinatorial Optimization*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Selman B, Kautz HA and Cohen B (1994) *Noise Strategies for Improving Local Search*. AAAI-94, Seattle, Washington, pp. 337–343.

Selman B, Levesque HJ and Mitchell D (1992) *A New Method for Solving Hard Satisfiability Problems*. Proceedings AAAI-92, San Jose, California, pp. 440–446.

Sosic R and Gu J (1991) 3,000,000 queens in less than one minute. *Bulletin* **2**(2): 22–24.

Stone HS and Stone JM (1987) Efficient search techniques – an empirical study of the *N*-queens problem. *IBM Journal of Research and Development* **31**: 464–474.

Further Reading

Aarts EHK and Lenstra JK (1997) *Local Search in Combinatorial Optimization*. London, UK: John Wiley.

Garey MR and Johnson DS (1979) *Computers and Intractability, A Guide to the Theory of NP-Completeness*. New York, NY: WH Freeman.

Gent I, van Maaran H and Walsh T (eds) (2000) *SAT 2000*. IOS Press.

Papadimitriou C and Steiglitz K (1998) *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications.

Rayward-Smith VJ, Osman IH and Reeves CR (eds) (1996) *Modern Heuristic Search Methods*. London, UK: John Wiley.

Sait SM and Youssef H (1999) *Iterative Computer Algorithms with Applications in Engineering*. 1EEE Computer Science Press.

History of Cognitive Science and Computational Modeling

Introductory article

Whit Schonbein, Washington University, St Louis, Missouri, USA
William Bechtel, University of California, San Diego, California, USA

CONTENTS

The various strands of computational modeling in cognitive science are simultaneously grounded in foundational work in the theory of computation and in recent developments in computational neuroscience and dynamical systems theory. All these approaches face important challenges, and to some degree they stand opposed in their treatment of role of computation in modeling.

COMPUTATION

In 1936 – the birth year of the theory of computation – various strands of mathematical logic had all converged on a common set of questions. Which problems can be solved through a finite number of applications of a finite number of rules? For example, is there a general procedure – an algorithm – for multiplying two integers? Is there an algorithm for deciding whether a given well-formed formula is a theorem of a particular set of axioms? Are there problems that cannot be solved by applying a finite set of rules a finite number of times?

In that year, Alan Turing, Alonzo Church and Emil Post independently proposed equivalent answers. They argued that in fact the power of algorithms is limited. There are problems that cannot be solved in this way. Turing's proof has come to play a central role in current accounts of computation, and therefore deserves special emphasis.

Turing focused on what it was that humans do when they solve calculation problems (such as multiplication or division) using pencil and paper. Turing constructed a formal 'machine' with three parts. The first, corresponding to the paper used by the human, is a 'tape'. The tape is unbounded, or as long as it needs to be (as if the human had as much paper as he or she needed), and divided into discrete squares arranged linearly. Each square can contain a single 'symbol' (e.g. an 'x', a '1', or a squiggle).

The second part of Turing's machine is a 'controller', corresponding to the brain of the human. It is responsible for keeping track of where in a given calculation the machine is at any moment. For example, when multiplying large numbers, one must sometimes carry a digit. In this case, a person would write a numeral on the paper, and make a 'mental note' to add it to the result of the next multiplication. The paper is the tape, and the 'mental note' is a state of the controller. So a Turing machine (TM) would write a symbol (or symbols) to the tape and change the state of the controller to one representing the need to carry.

Finally, corresponding to the eyes and hands of the human, a TM has a read–write head that can read and write symbols from and to the tape. At each moment, the head is over one square of the tape, can read the symbol written on that square, can write a symbol to the square, and can move to another square.

For the TM to actually do anything, we need to specify how it behaves in various circumstances. This is done by means of a 'transition table'. This is a finite collection of rules, each of which is of the form: 'If the current symbol is W_1, and the current controller state is Q_1, then write symbol W_2 to the tape, change to state Q_2, and move the tape head one square to the left or right.'

So a TM is a tape, a controller, a read–write head, and a set of rules describing its behavior. To run a TM, one simply provides it with an input and starts it. The machine continues to follow the rules

contained in its table, until it halts, leaving the result printed on the tape. Thus, after a finite number of applications of a finite number of rules, the TM provides an answer.

Using this formalism, Turing proved a number of interesting results. First, in some cases the TM will not halt – it may continue forever. Second, there are limits to what a TM can do. There are more functions than there are TMs, so there are functions that lack algorithms to compute them; some functions are uncomputable. Finally, Turing demonstrated that it is possible to define a TM that takes as an input the specification of any other TM. This 'universal' TM then simulates the behavior of that other TM. In other words, Turing discovered the possibility of a machine that could be programmed to compute arbitrary Turing-computable functions.

Turing's machines constitute one attempt at formalizing the intuitive notion of an algorithm. Assuming they are adequate, important consequences follow for what can be accomplished using algorithms, for example, not all functions can be calculated using algorithms. Thus the TM constitutes Turing's simple, elegant, and fruitful contribution to mathematical logic. (Other attempts include those of Emile Post and Alonzo Church, both from 1936. Post's formalization has been demonstrated to be equivalent to Turing's; Church's has not.) In addition, however, TMs embody a particular conception of computation: they compute by manipulating symbols stored on a tape according to a set of rules as indicated in the transition table. This conception of computation, as we shall see, is adopted in traditional models in cognitive science.

TURING AND HUMAN INTELLIGENCE

During the Second World War Turing was part of a team at Bletchley Park attempting to break the increasingly complex codes employed in German communications. Turing and his colleagues built a series of machines to aid in the tedious (and otherwise practically impossible) job of considering possible solutions to these cryptographic puzzles. Some of these machines were purely mechanical, but others were composed of electronic components. This experience indicated to Turing the possibility of actually building a general computing device, designed along the lines of the universal TM.

Immediately after the war, Turing was involved in the design and implementation of such a device (the ACE), but he soon lost control of the project to other interested parties. (Similar projects were underway in the United States, leading to the construction of ENIAC and EDVAC. Turing's role in the conception and design of these devices is a matter of some dispute.) However, unlike many of his contemporaries (who saw a general-purpose computer as a way of solving mathematical problems), Turing had a vision of the role the computer could play in the science of human behavior. Turing viewed the human problem-solving capacity – presumed by many to be uniquely human – as nothing more than what TMs do. That is, Turing believed that humans solve problems – they act intelligently – by manipulating symbols according to rules. In a personal letter from 1947, he wrote: 'In working on the ACE I am more interested in the possibility of producing models of the action of the brain than on the practical applications to computing.'

Indeed, the idea that the logical formalization of the notion of an algorithm could be fundamental to the study of human intelligence was in place as early as 1936, when Emil Post proposed a mechanism for computation equivalent to Turing's machines, and possessing the same basic features. Post explicitly considers the relevance of his account for 'psychological fidelity', or the accuracy of its relation to human psychological mechanisms. He proposes that his system captures 'wider and wider formulations', in that extensions will be logically reducible to it. This suggests that his system puts an 'upper limit' on how to conceive of problem-solving, and if this is the case, then the hypothesis is not just a definition or an axiom but a 'natural law'. In other words, Post hypothesizes that no physical device – including human beings – can fail to be describable in terms of one of his systems. By equivalence, this holds for TMs as well. (The thesis that no physical device can compute a function not computable by a TM or Post system is only a hypothesis, and has not been proven.) So the capacity of humans to act intelligently can be accounted for in terms of a system such as those proposed by Turing or Post, i.e., in terms of the manipulation of symbols according to rules.

THE TURING TEST FOR INTELLIGENCE

In 1950, Turing published a provocative paper on the capacity for computers to exhibit human intelligence in which he proposed a 'test' for machine intelligence. Passing this test, he argued, is sufficient for the possession of genuine intelligence.

The test is an imitation game. A human 'examiner' sits in front of a keyboard and screen. By typing questions and reading responses, the examiner can interact with two other entities. One of these entities is a computer programmed to mimic human

responses. The other is an actual human being. The goal of the examiner is to distinguish the human from the computer by their responses to questions the examiner poses. If a computer can fool the examiner into believing it is the human, the computer has passed the 'test', and, according to Turing, is intelligent. Turing predicted that computers would pass the test by the year 2000.

Much can be (and has been) said about the Turing test. Some have objected that the test measures only verbal behavior and ignores other possible indicators of intelligence, such as a capacity to construct tools, navigate in messy environments, engage in long-term planning, or interact in social groups. Furthermore, it neglects the actual processes by which humans manifest the intelligent behaviors they do.

The distinction here is between systems that behave intelligently because they execute the same processes by which humans achieve that behavior, and systems that 'merely' behave intelligently. John Searle has called this the distinction between 'strong' and 'weak' artificial intelligence (AI). Suppose, for example, that we have a computer that can pass the Turing test in the restricted domain of reading a story and answering questions about it. According to Searle, if the program mimics the relevant human behavior (i.e., answers the questions as a human would), then it is a successful instance of weak AI. In contrast, strong AI requires that the computer actually be intelligent in the way we are. Thus, it would have to understand the story and employ that understanding in answering the questions.

In strong AI, the computer program is a model of the psychological processes underlying the production of intelligent behavior in human beings. The program explains how humans behave by making explicit the symbols (and collections of symbols) that humans think with, and the rules they use to manipulate them. Indeed, this was the goal of a summer workshop held at Dartmouth University in the summer of 1956, which gave birth to the discipline of artificial intelligence. According to the original proposal, the workshop's aim was to pursue the hypothesis that 'every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it'. This is essentially Turing's prophecy of 1950.

THE 'GENERAL PROBLEM SOLVER'

At the 1956 Dartmouth workshop, Alan Newell and Herbert Simon demonstrated their 'logic theorist', which constructed its own proof of a theorem from Whitehead and Russell's *Principia Mathematica*. It was the first AI system actually implemented in a physical machine. Newell and Simon's goal was not merely to have the program construct proofs, but to generate them in the same way that humans do, thereby realizing strong AI.

In 1963, Newell and Simon extended their endeavor beyond logic to reasoning in general with their 'general problem solver' (GPS). This employs a set of symbols that designates a solution to a problem (e.g. a theorem in sentential calculus) and a set of symbols that constitutes the current knowledge possessed by the system (e.g. a set of axioms). Its goal is to generate the solution by applying a finite number of rules to the initial knowledge base. For example, to make a move in chess, it would begin with a symbolic representation of the initial board configuration and apply various operations to it that would correspond to legal moves, and other rules that evaluated the positions they generated (with the ultimate goal of checkmate). In this way, intelligent behavior is conceived of as a search through a problem space. Newell and Simon recognized that 'blind' search is too costly, and proposed that intelligence does not result from considering all possible operations, but rather in choosing the best rules to apply in a given situation. Thus GPS uses a heuristic, or 'rule of thumb', for selecting a rule to apply. GPS compares representations of its goal state and current state, determines the differences, and identifies which of its stored rules would be most likely to minimize these differences. This process is repeated until the solution is reached. Such 'means–ends analysis' is only one of a number of possible ways to make a search through a problem space more efficient and more 'clever'.

Protocol Analysis

To support their claim that GPS is a model of human problem-solving, Newell and Simon compared its behavior with how humans solve the same problems, using a process called protocol analysis. A subject is given a problem, such as transforming one logical expression into another using a finite set of valid rules, and asked to report aloud their thoughts as they solve the problem. A 'protocol' is then a verbatim record of all that the subject or the experimenter said during the experiment. The protocol is compared to a 'trace' of GPS's performance (i.e. a report of the rules it applied in order to generate the solution). Newell and Simon held that if the trace were close enough to the

protocol then the program would constitute a good theory of the subject's problem-solving. Although 'logic theorist' did not fit the behavioral data, GPS exhibited some interesting correspondences.

Physical Symbol Systems

In their 1976 address to the Association for Computing Machinery, Newell and Simon characterized GPS and computational systems like it as 'physical symbol systems', which they define as follows:

> A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of processes that operate on expressions to produce other expressions: processes of creation, modification, reproduction, and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures.

Several points in this passage need clarification.

First, the distinction Newell and Simon make between symbols and symbol strings has been left implicit in the above accounts, but the basic idea can be illustrated through a simple example. Suppose we have the symbols 'A', '&', and 'B'. These symbols can be arranged in different ways – one way is to concatenate them: 'A & B'. Newell and Simon note that operations transform complexes of symbols. For example, a rule might transform 'A & B' to 'A'. This is in contrast to the basic definition of a TM, which defines rules over single symbols. However, as Turing pointed out in his 1936 article, the definition of a TM allows one to use finite sequences of rules to simulate the operation of a single rule defined over symbol strings. For example, if a system has a rule that transforms 'A & B' to 'A', a TM might use four rules: three to erase the individual symbols 'A', 'B', and '&', and one to write 'A'. It is this capacity that allows one to abstract symbols away from whatever primitives a particular machine might use (just as we treat the ones and zeros in a computer as characters, numbers, addresses, and so on, rather than just as strings of ones and zeros.)

Secondly, the definition of a physical symbol system says nothing about what a set of symbols means. For example, the collection of words 'There's gold in them than hills' is a collection of symbols (words) related to one another (through concatenation) in a structure (a sentence). And it clearly means something – in English. But there is nothing about its being part of a physical symbol system that makes it have the meaning it does: the meaning of that collection of symbols is open. Therefore, transformations over expressions do not depend on the meanings of the symbols but simply on their form. That is, transforming 'A & B' into 'A' depends on the fact 'A' is next to '& B', not on anything these symbols might mean.

Finally, the definition is silent on the question of how to choose the appropriate rule. Newell and Simon address this question in a separate hypothesis, the 'heuristic search hypothesis':

> The solutions to problems are represented as symbol structures. A physical symbol system exercises intelligence in problem-solving by search – that is, by generating and progressively modifying symbol structures until it produces a solution structure.

Having characterized physical symbol systems, Newell and Simon make the following strong claim: 'A physical symbol system has the necessary and sufficient means for general intelligent action'. In other words, physical symbol systems provide resources sufficient for manifesting intelligent behavior, and furthermore, all intelligent behaviors are necessarily the product of a physical symbol system. The first claim is consistent with weak AI. The second, however, is a clear declaration of strong AI. Returning to the GPS as a theory of human problem-solving, the claim becomes that the reason human subjects exhibit behavior that corresponds with the trace of GPS solving the same problem is that those individuals realize a physical symbol system like the one implemented in GPS. GPS is not merely a way to mimic human behavior, it is a way of explaining the mechanisms underlying that behavior.

COMPUTATIONAL MODELS OF LANGUAGE PROCESSING

Since language use is a quintessentially intelligent human activity, it provides a natural target of computational modeling. One goal has been to devise computational systems that can understand sentences presented in a natural language such as English, and one measure of understanding is the ability of the system to generate reasonable responses to such sentences.

Language processing programs developed in the 1960s used relatively simple template matching. For example, Bert Green's BASEBALL program answers questions about one season of American League games by transforming simple queries into

a canonical form, and using this form to search a prespecified database. Bertram Raphael's SIR program matches English inputs to templates (e.g., * IS PART OF *, HOW MANY * DOES * HAVE?). If the input is declarative, it extracts the symbol string for later use in answering queries; if the input is a question, it searches stored strings for an appropriate response. Daniel Bobrow applied the same approach in his program STUDENT, designed to solve high school algebra story problems. The most famous of the template matching language processing systems is Joseph Weizenbaum's ELIZA, which simulates a (Rogerian) therapist by extracting pre-identified terms from the user's inputs and using these to formulate questions to continue the 'dialogue'. The system includes a number of 'tricks' that allow the substitution of one term for another (e.g. 'mother' for 'family'), yielding a modicum of unpredictability in its responses.

These relatively simple systems were followed in the 1970s by more substantial programs, three of which are described here.

SHRDLU

Terry Winograd objected to the exclusive emphasis on syntax in the programs described above. He proposed that to really understand language, a program also must have access to the meanings of words. He designed SHRDLU with this in mind. SHRDLU both answered questions and could 'manipulate' a simulated world made of blocks. For example, SHRDLU could answer the question whether there was a blue box under a green pyramid, or could stack a blue pyramid on a yellow box.

Winograd used procedural representations to capture aspects of the meaning of terms or phrases in English. Roughly speaking, he attached actions to words. For example, the concept CLEARTOP (associated with commands to move a block, or to determine whether a block has anything on top of it) can be used to describe a situation (i.e., as a predicate), but, if false, can also be 'run' so as to generate a sequence of actions that result in the desired block being clear on top.

Although SHRDLU is limited to 'toy problems', by linking words to actions and thereby incorporating information beyond that contained in the syntax of the natural language it represented a clear advance over template-matching approaches. Subsequent language understanding programs have followed the approach of supplementing syntactic information with rich internal representations so as to expand their behavioral repertoires.

MARGIE

In his book *Conceptual Information Processing*, Roger Schank expressed the view that the basis of human thought is 'a representation of meaning of natural language that does not involve any of the words of the language'. This language-free representation is an 'interlingua' – so named because of its possible role as an intermediary in translating between natural languages. Thus it 'should be extractable from any language and be capable of being generated into any other language'. In a later work, Schank and his colleague Robert Abelson express this view in their 'basic axiom' of conceptual dependency: 'For any two sentences that are identical in meaning, regardless of language, there should be only one representation'. MARGIE was an attempt to put this principle into practice.

When presented with a sentence in English, MARGIE extracts a conceptual dependency representation: a symbol structure that captures the meaning of the sentence, rather than its syntactic profile. This involves, for example, representing the roles of agent and patient. Furthermore, as Schank and Abelson note, MARGIE operates on the principle that 'any information in a sentence that is implicit must be made explicit in the representation of the meaning of that sentence'. Thus, if MARGIE is presented with the sentence 'Bob smokes a cigarette', the system constructs a representation that includes implicit information such as the fact that MOUTH is the recipient of the cigarette smoke.

SAM

Semantics is attached not only to words or sentences, but also to larger units such as paragraphs or pages. To deal with these, Schank and Abelson developed SAM, a program designed to read and understand stories. They proposed that stories are organized around 'scripts', or prototypical scenarios. When one hears about a person going to a restaurant, one expects the person to be seated, to be given a menu, to have his or her order taken by a waiter or waitress, to have food delivered, to eat the food, to be presented with a bill, and to pay the bill before leaving. All of these are stereotypical components of a visit to a restaurant.

Schank and Abelson proposed that scripts could be realized as abstract data structures that included representations of the components, with slots and default fillers for the various ways these components might be filled. When SAM reads a story, it encodes the events in such a script. Using the default slot fillers, it is able to make inferences from

partial accounts of actions in a given setting. For example, if a story states that Dabney entered a restaurant, ordered a hamburger, and caught the 6 o'clock bus, SAM can infer that Dabney ate the hamburger, since this is part of the script. In addition to scripts, Schank and Abelson found that other conceptual dependency data structures were also useful. For example, 'plans' were used to help SAM to identify the goals of the main actors and to propose actions required to fulfill them. By using plan structures, SAM could guess what an agent would do next.

Both MARGIE and SAM are exercises in strong AI. As accounts of human cognition, they propose that conceptual dependency representations are what humans think with and that understanding the meaning of a sentence involves forming such a representation.

Like GPS, these programs all work by manipulating symbols and collections of symbols according to sets of rules. What was innovative was the kind and extent of the information that the programmers encoded in these symbols, structures, and rules. This observation brings us to two important challenges to strong AI: the frame problem, and Searle's Chinese room argument.

THE FRAME PROBLEM

In essence, the frame problem is the problem of determining what information is relevant to a given task. This difficulty is apparent in MARGIE: by following conceptual dependencies between internal representations, the system is designed to make information implicit in a sentence explicit. But this process could go on indefinitely: with the sentence 'Bob smokes a cigarette', MOUTH is the recipient of the smoke, so TEETH come in contact with the smoke, so they turn more BROWN, and so on. One solution is simply to place an arbitrary limit on the length of inferential chains. But given a large knowledge base, even series of limited length may result in an infeasibly large number of inferences being explored – so large, indeed, that effective action may be blocked. This is especially important in real-world contexts, which usually have built-in time constraints.

The frame problem runs even deeper than this. Not only must an intelligent system determine which of the facts it possesses are relevant to a given task; it must also decide which changes in the world are relevant. For example, when SHRDLU clears the top of one block, its actions may cover the top of another. If this other block is relevant to the task, SHRDLU must detect this; otherwise the change should be ignored. But how do we determine whether a fact should be ignored? We might add a rule (a 'frame axiom') to the effect that, if the task is of a certain sort, and the fact has a certain feature, then it is not relevant. But adding more rules to cover judgments of relevance is tantamount to adding more inferences that must be explored: so, instead of simply inferring that Bob's teeth turn brown, MARGIE might expend resources inferring that the fact that their brownness is irrelevant to the current task (or not). In general, the problem is successfully keeping track of relevant changes while discarding irrelevant ones, without a rampant proliferation of frame axioms.

In some ways the frame problem is a practical problem for which the solution is to design more effective abstract representations of knowledge and heuristics for the manipulation of those representations, and a wide variety of solutions have been proposed. Some theorists, however, have suggested that the frame problem is intrinsic to – and unsolvable by – traditional computational approaches to cognitive modeling: that it is the reliance on the rule-based manipulation of representations that generates the frame problem, and therefore an alternative approach must be adopted. A second challenge to traditional strong AI – the Chinese room argument – can be interpreted as leading to a similar conclusion.

THE CHINESE ROOM ARGUMENT

In 1980, John Searle proposed a criticism of strong AI in an argument as simple and as elegant as Turing's machine. Searle offered the following simple thought experiment. You are a native English speaker, and have no knowledge of Chinese. You sit in an empty room, where there is a slot in one wall. Three batches of cards are passed through the slot. The first is a set of Chinese characters. By hypothesis, they are meaningless to you. The second is another collection of Chinese characters, together with some rules in English (which you understand, of course), instructing you how to correlate the new batch of Chinese characters with those of the first. Finally, more Chinese characters are fed through the slot, along with some further instructions in English. These instructions tell you how to correlate the newly arriving symbols with those of the first two batches and sometimes to feed cards containing Chinese characters back through the slot.

Unbeknown to you, the first batch of Chinese characters was a *script*, the second batch was a story, and the third batch was a set of questions.

By virtue of the rules you followed, the characters you fed back out through the slot were cogent answers to these questions. However, you did not know what any of the Chinese characters meant, so you did not understand the story.

Searle concludes that strong AI is effectively refuted. First, neither you in the Chinese room nor SAM understands the stories you have read. Since you can realize SAM without understanding the story, realizing SAM is not sufficient for understanding. Secondly, since there is no relation between realizing SAM and understanding the story, there is no reason to think that running a program like SAM is necessary for understanding.

Searle thus rejects the fundamental idea, evinced by Turing, Post, Newell, Simon, Schank and others, that human cognitive processes can be captured by appeals to the manipulation of symbols according to rules. When you manipulate Chinese characters according to the rules given in English, you cannot identify those characters in terms of what they mean since, by hypothesis, you do not understand Chinese. The same holds for TMs as well. A rule in a TM transition table operates over controller states and symbols. Symbols are not distinguished on the basis of what they might mean, but rather on their formal properties. Likewise, physical symbol systems are defined in terms of purely formal symbols, and so are the language processing systems considered above. Searle's claim is therefore that human psychological states such as understanding cannot be accounted for by rules that transform symbols.

If not symbol manipulation, then what does explain the understanding achieved by cognitive agents? Searle suggests that we need to stop thinking in terms of computation and consider the physical properties of the brain. But there is another possibility: to examine the computational capacities of the brain. One approach to this task is to design and study the capacities of artificial brains.

EARLY NEURAL NETWORKS

In 1943, inspired by the work of Turing, the neurophysiologist Warren McCulloch and the logician Walter Pitts formulated a model for computing using networks of artificial neurons. An artificial neuron could receive input from others, and when the sum of the inputs to it exceeded a specified threshold, it would 'fire' and pass input to other neurons to which it was connected. McCulloch and Pitts proved that networks of these elements were equivalent to finite state automata (essentially, TMs without a tape), and suggested that, with the addition of a tape, they would be the same as Turing machines.

Although McCulloch and Pitts were initially interested in instantiating logic operations in a network, they and others soon began to explore the capacities of networks more broadly. In particular, researchers began to think of networks as devices that could 'learn' to perform desired mappings between inputs and outputs by modifying the strengths of connections between units. In 1949, Donald Hebb proposed what has turned out to be a powerful rule for adjusting weights. In essence, it allows for the strengthening of connections between units whenever they are both active together ('units that fire together wire together').

In 1962, Frank Rosenblatt introduced the formal notion of a 'perceptron', a network of linear-threshold artificial neurons (similar to those of McCulloch and Pitts) in which some constituted an input layer, others an output layer, with connections between input and output units. Rosenblatt introduced a learning algorithm that he proved would allow these networks to learn any mapping for which there was a possible set of weights. In particular, Rosenblatt trained perceptrons to classify various characters presented on a two-dimensional array.

Initially perceptrons seemed to offer great potential as tools for modeling cognitive operations; but in 1969 Marvin Minsky and Seymour Papert demonstrated that there was a large class of problems that could not be solved by (two-layer) perceptrons. An example is the logical 'exclusive or' operation, whose value is true when one proposition is true and the other false. The values of weights required to determine the proper output for some instances (both true) is incompatible with those required for other instances (both false). Technically, these are known as problems that are not linearly separable.

Minsky and Papert's objections focused not only on nonlinear problems. In some cases, perceptrons were able to learn a problem at a small scale, but unable to learn the problem at a larger scale. The reason, they argued, is that the learning procedure used by perceptrons suffers from exponential growth. That is, as the size n of the problem (e.g. the size of the image to be classified) increases, the amount of processing required of the perceptrons increases exponentially (as x^n where x is some constant): thus, the amount of processing required quickly becomes infeasible. Partly as a result of Minsky and Papert's analysis, interest in artificial neural networks waned for over a decade.

CONNECTIONISM

Rosenblatt (as well as Minsky and Papert) recognized that by inserting an additional layer of units (known as hidden units) between input and output units, one could overcome the problem of linear separability. But he failed to discover a learning rule that could determine the appropriate weights feeding into the hidden units. In 1986 three researchers independently discovered such a learning rule (back propagation). In the same year, a two-volume compilation of articles edited by David Rumelhart and James McClelland entitled *Parallel Distribution Processing: Explorations in the Microstructure of Cognition* heralded the revival of interest in artificial neural networks. Such 'connectionist' networks have since been employed to model a great variety of cognitive activities, including simple pattern matching or pattern detection, but also including cognitive tasks like problem-solving and reasoning that had been the preserve of symbolic models. However, one of the problems noted by Minsky and Papert is still unresolved: how to scale from 'toy' problems to the sorts of problems that real cognitive agents confront.

The rapid expansion of connectionist modeling in the cognitive sciences raises a basic theoretical question: do these models count as computational? Since, prima facie, they do not operate by manipulating symbols according to rules, they do not compute in the same manner as Turing machines or symbolic models. But they are frequently employed for similar problems (including problems of cognitive modeling). Moreover, under appropriate constraints, they can perform the same input–output mappings as symbolic systems, and indeed they can be configured to literally implement a Turing machine. Some theorists argue that it is only at a high level that they are characterizable in terms of symbol processing, proposing that instead they should be viewed as performing subsymbolic computation.

In part, the question of whether connectionist networks are computational is a semantic issue that depends on how 'computation' is defined. But one idea that has been proposed by some theorists is that by construing neural networks as computational devices, we increase the range of computable functions. If we follow Turing and Post in thinking of human cognition in terms of computing function, this raises the possibility that their hypothesis that Turing- or Post-computable functions set an upper bound on what humans could do may be false: neural networks may be able to do things – realize and explain behaviors – that traditional computational automata cannot.

Expanding the notion of computation, however, is not without cost. If whatever any physical device does is construed as computation, then characterizing an activity as computation becomes vacuous. This concern is illustrated by a recent challenge to both symbolic and connectionist models of cognition.

THE DYNAMICAL APPROACH

A number of theorists have recently argued that computational approaches to modeling cognition are misguided. They propose instead that cognitive science should follow the model of physics, where explanation consists of putting forward a set of equations describing a system's behavior. In 1995, Timothy van Gelder illustrated the contrast between the computational and dynamical approaches by describing two ways in which James Watt could have designed a governor to control the steam flow in a steam engine. Had Watt been a computer programmer, he could have written a symbolic program that would have regularly calculated the difference between the target speed of the engine and the actual speed and determined how much to open or close the value in order to reduce the difference to zero. In fact, Watt designed a physical device in which a spindle with arms was attached to the flywheel of the engine. Centrifugal force caused the arms to open or close in proportion to the engine's speed. All Watt then needed to do was add a mechanical linkage so that when the arms opened too far, the opening of the steam valve was reduced, and when they closed too far, it was increased. Van Gelder argued that Watt's solution was not merely a necessity given inadequate computational devices, but in fact was a superior solution.

Van Gelder's argument raises two questions. The first is whether a mechanical solution such as Watt's can be extended to human cognition, where reasoning and planning are the objects of interest: recall that Turing and Post developed their conception of computation by reflecting on these sorts of human activities. The second is whether the Watt governor and the various dynamical systems that have been proposed to account for cognition really are noncomputational. One common strategy to account for why these devices achieve their purposes is to try to understand their intermediate states in terms of what information

they carry (the angle of the arms carries information about the speed of the flywheel) and how the system then utilizes the information to generate behavior. Giving such accounts seems to require us to construe the devices as computational. This response, however, seems almost to imply that every physical system is computational. Thus, the problem remains of specifying just what constitutes computation.

Further Reading

Anderson JA and Rosenfeld E (eds) (1988) *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.

Cummins R and Cummins DD (eds) (2000) *Minds, Brains, and Computers: The Foundations of Cognitive Science*. Oxford: Blackwell.

Davis M (ed.) (1965) *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions*. Hewlett, NY: Raven Press.

Haugeland J (ed.) (1997) *Mind Design II*. Cambridge, MA: MIT Press.

Hodges A (1983) *Alan Turing: The Enigma*. New York, NY: Simon and Schuster.

Hopcroft JE and Ullman JD (1979) *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

McCartney S (1999) *ENIAC: The Triumphs and Tragedies of the World's First Computer*. New York, NY: Walker.

Implicit Learning Models

Intermediate article

Axel Cleeremans, Université Libre de Bruxelles, Brussels, Belgium

CONTENTS

Implicit learning is the process through which we become sensitive to certain regularities in the environment (1) in the absence of intention to learn about those regularities, (2) in the absence of awareness that one is learning, and (3) in such a way that the resulting knowledge is difficult to express.

INTRODUCTION

Implicit learning – broadly construed, learning without awareness – is a complex, multifaceted phenomenon that defies easy definition. Frensch (1998) lists as many as eleven definitions in a recent overview – a diversity that is undoubtedly symptomatic of the conceptual and methodological challenges that continue to pervade the field 35 years after the term first appeared in the literature (Reber, 1967). According to Berry and Dienes (1993), learning is implicit when we acquire new information without intending to do so, and in such a way that the resulting knowledge is difficult to express. In this, implicit learning thus contrasts strongly with explicit learning (e.g. as when learning how to solve a problem or learning a concept), which is typically hypothesis-driven and fully conscious.

Since the early 1990s, the field of implicit learning has come to embody ongoing questioning about three fundamental issues in the cognitive sciences, namely (1) consciousness (how we should conceptualize and measure the relationships between conscious and unconscious cognition); (2) mental representation (in particular the complex issue of abstraction); and (3) modularity and the architecture of the cognitive system (whether one should think of implicit and explicit learning as being subtended by separable systems of the brain or not). Computational modeling plays a central role in addressing these issues, most importantly by offering principled ways of deconstructing early characterizations of implicit learning as involving the unconscious acquisition of abstract knowledge.

IMPLICIT COGNITION: THE PHENOMENA

Everyday experience suggests that implicit learning is a ubiquitous phenomenon. For instance, we often seem to know more than we can tell. Riding a bicycle, playing tennis or driving a car all involve mastering complex sets of motor skills that we find very difficult to describe verbally. These dissociations between our ability to report on cognitive processes and the behaviors that involve these processes are not limited to action but also extend to high-level cognition. Most native speakers of a language are unable to articulate the grammatical rules they nevertheless follow when uttering expressions of the language. Likewise, expertise in domains such as medical diagnosis or chess, as well as social or esthetic judgments, all involve intuitive knowledge that one seems to have little introspective access to.

We also often seem to tell more than we can know. In a classic article, social psychologists Nisbett and Wilson (1977) reported on many experimental demonstrations that verbal reports on our own behavior often reflect reconstructive and interpretative processes rather than genuine introspection. While it is generally agreed that cognitive *processes* are not in and of themselves open to any sort of introspection, Nisbett and Wilson (1977) further claimed that we can sometimes be '(a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response' (p. 231).

Demonstrations of dissociations between subjective experience and various cognitive processes have now been reported in many domains of cognitive science. For instance, dissociations have been reported between conscious awareness and memory. Memory for previous events can be expressed explicitly, as a conscious recollection, or implicitly, as automatic, unconscious influences on behavior. Numerous studies have demonstrated dissociations between implicit and explicit memory, both in normal participants (see Schacter, 1987) as well as in special populations. Amnesic patients, for instance, who exhibit severe or total loss in their ability to explicitly recall previous experiences (conscious recollection) nevertheless retain the ability to learn novel procedural skills or to exhibit sensitivity to past experiences of which they are not conscious.

Findings of 'learning without awareness' have also been reported with normal subjects (Cleeremans *et al*., 1998). It is Arthur Reber, in a classic series of studies conducted in 1965 (Reber, 1967), who first suggested that learning might be 'implicit' to the extent that people appear to be able to learn new information without intending to do so and in such a way that the resulting knowledge is difficult to express. Implicit learning contrasts with implicit memory in that it typically involves sensitivity to relationships between events rather than sensitivity to single events, and with subliminal perception in that it typically involves supra-liminal stimuli.

Implicit learning research has essentially been focused on three experimental paradigms: artificial grammar learning, dynamic system control, and sequence learning. Additional paradigms that will not be discussed further include probability learning, hidden covariation detection, acquisition of invariant characteristics, or visual search in complex stimulus environments.

In Reber's seminal study of artificial grammar learning (Reber, 1967), subjects were asked to memorize meaningless letter strings generated by a simple set of rules embodied in a finite-state grammar (Figure 1). After this memorization phase, subjects were told that the strings followed the rules of a grammar, and were asked to classify novel strings as grammatical or not. In this experiment and in many subsequent replications, subjects were able to perform this classification task better than chance despite remaining unable to describe the rules of the grammar in verbal reports. This dissociation between classification performance and verbal report is the finding that prompted Reber to describe learning as implicit, for subjects appeared sensitive to and could apply knowledge that they remained unable to describe and had had no intention to learn.

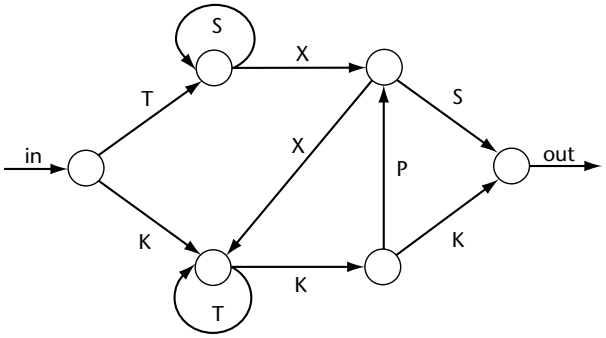


Figure 1. A finite-state grammar is a simple directed graph consisting of nodes connected by labeled arcs. Sequences of symbols can be generated by entering the grammar through an 'in' node, and by moving from node to node until an 'out' node is reached. Each transition between a node and the next produces the label associated with the arc linking the two nodes. Concatenating the symbols together produces strings of symbols, in this case letters of the alphabet.

In a series of studies that attracted renewed interest in implicit learning, Berry and Broadbent (1984) showed that success in learning how to control a simulated system (e.g. a 'sugar factory') so as to make it reach certain goal states was independent from ability to answer questions about the principles governing subjects' inputs and the system's output. Practice selectively influenced ability to control the system, whereas verbal explanations about how the system works selectively influenced ability to answer questions.

Today, in 2002, another paradigm – sequence learning – has become dominant in the study of implicit learning. In sequence learning situations (Clegg *et al*., 1998), participants are asked to react to each element of a sequentially structured visual sequence of events in the context of a serial reaction time task. On each trial, subjects see a stimulus that appears at one of several locations on a computer screen and are asked to press as fast and as accurately as possible on the key corresponding to its current location. Nissen and Bullemer (1987) first demonstrated that subjects progressively learned about the sequential structure of the stimulus sequence in spite of showing little evidence of being aware that the material contained structure. Numerous subsequent studies have indicated that subjects can learn about complex sequential relationships despite remaining unable to fully deploy this knowledge in corresponding direct tasks.

DEMONSTRATING IMPLICIT COGNITION

These findings all suggest that unconscious influences on behavior are pervasive. This raises the question of how to best characterize the relationships between conscious and unconscious processes, and in particular whether one should consider that mental representations can be unconscious. Because there is no accepted operational definition of what it means for an agent to be conscious of something, three challenges need to be overcome when attempting to contrast conscious and unconscious cognition: a definitional challenge (what is it that we want to measure?), a methodological challenge (which measure offers an appropriate index of awareness?), and a conceptual challenge (what are the implications of dissociation findings?).

The Definitional Challenge

Addressing this first challenge involves delineating which aspects of consciousness 'count' when assessing whether a subject is aware or not of a particular piece of information: awareness of the presence or absence of a stimulus, conscious memory for a specific previous processing episode, awareness of one's intention to use some information, or awareness that one's behavior is influenced by some previous processing episode. Consciousness is not a single process or phenomenon, but rather encompasses many dimensions of experience. Different aspects of conscious processing are thus engaged by different paradigms. In subliminal perception studies, for instance, one is concerned with determining whether stimuli that have not been consciously encoded can influence subsequent responses. In contrast, implicit memory research has been more focused on retrieval processes, that is, on the unintentional, automatic effects that previously consciously perceived stimuli may exert on subsequent decisions. In studies of implicit learning, it is the relationships between ensembles of consciously processed stimuli that remain purportedly unconscious.

These subtle differences in which specific aspects of the situation are available to awareness emphasize the need to distinguish carefully between awareness during encoding and awareness during retrieval of information. Further, both encoding and retrieval can concern either individual stimuli or relationships between sets of stimuli, and both can either be intentional or not. Even so, there is continuing disagreement about exactly which criteria for conscious awareness should be used. In a recent review for instance, Butler and Berry (2001) found little evidence for implicit memory if one takes 'implicit' to imply that performance (1) results from an unintentional retrieval strategy, and (2) is not accompanied by conscious recollection. Likewise, in a landmark review article dedicated to implicit learning, Shanks and St. John (1994) failed to identify convincing evidence for learning without awareness and therefore concluded that 'Human learning is almost invariably accompanied by conscious awareness' (p. 394).

The Methodological Challenge

This second challenge consists of devising an appropriate measure of awareness. Most experimental paradigms dedicated to exploring the relationships between conscious and unconscious processing have relied on a simple quantitative dissociation logic aimed at comparing the sensitivity of two different measures to some relevant information: a measure *C* of subjects' awareness of the information, and a measure *P* of behavioral sensitivity to the same information in the context of some task. Unconscious processing, according to the quantitative dissociation logic, is demonstrated whenever *P* exhibits sensitivity to some information in the absence of correlated sensitivity in *C*. There are several important pitfalls with this reasoning, however.

First, the measures *C* and *P* cannot typically be obtained concurrently. This 'retrospective assessment' problem (Shanks and St. John, 1994) means that finding that *C* fails to be sensitive to the relevant information need not necessarily imply that information was processed unconsciously during encoding, but that, for instance, it might have been forgotten or otherwise distorted before retrieval. It is therefore important, but often impossible – short of resorting to on-line measures of awareness such as are made possible through brain imaging techniques – to obtain concurrent measures of learning and awareness, partly because measuring awareness concurrently to processing entails a form of 'observer paradox' in which what is measured comes to be influenced by the measurement itself.

A second issue is to ensure that the information revealed through *C* is in fact relevant to perform the task. As Shanks and St. John (1994) suggested, many studies of implicit learning have failed to respect this 'information' criterion. For instance, successful classification in an artificial grammar learning task need not necessarily be based on

knowledge of the rules of the grammar, but can instead involve knowledge of the similarity relationships between training and test items. Subjects asked about the rules of the grammar would then understandably fail to offer relevant explicit knowledge, not because they have no awareness of the relevant rules, but simply because these rules are not necessary to perform the classification task successfully.

A third issue is to ensure that *C* and *P* are both equally sensitive to the relevant information (the 'sensitivity' criterion; Shanks and St. John, 1994). At first sight, verbal reports and other *subjective measures* such as confidence ratings would appear to offer the most direct way through which to assess the contents of subjective experience. The use of subjective measures to assess awareness was first advocated by Cheesman and Merikle (1984), who also introduced the notions of subjective and objective thresholds. Performance on a given task (i.e. identification) is said to be below the subjective threshold if one can show that performance is better than chance while subjects indicate they are guessing through subjective measures such as confidence judgments. Performance is said to be below the objective threshold if it fails to differ from chance on *objective measures* of awareness such as forced-choice tests (e.g. recognition, presence-absence decisions, or identification). Unconscious perception, for instance, would thus be demonstrated whenever performance is below the subjective threshold and above the objective threshold. This logic can also be applied to the domain of implicit learning, and several studies have now applied these ideas in the domains of artificial grammar learning and sequence learning. Overall, these studies indicate that the knowledge acquired by participants in these empirical situations can indeed be implicit to the extent that it is 'below the subjective threshold'.

Even if the different criteria briefly overviewed above are fulfilled, however, it might be optimistic to hope to be able to obtain measures of awareness that are simultaneously *exclusive* and *exhaustive* with respect to knowledge held consciously. In other words, finding null sensitivity in *C*, as required by the dissociation paradigms for unconscious processing to be demonstrated, might simply be impossible because no such absolute measure exists. A significant implication of this conclusion is that, at least with normal participants, it makes little sense to assume that conditions exist where awareness can simply be 'turned off'.

It might therefore instead be more plausible to assume that any task is always sensitive to both conscious and unconscious influences. In other words, no task is *process-pure*. Hence, according to this logic, even performance on forced-choice, objective tests such as recognition might be influenced by unconscious influences. These tests could therefore overestimate explicit knowledge. This is the *contamination* problem. Two methodological approaches that specifically attempt to overcome the conceptual limitations of the dissociation logic have been developed.

The first approach was introduced by Reingold and Merikle (1988), who suggested that the search for absolute measures of awareness should simply be abandoned in favor of approaches that seek to compare the sensitivity of *direct* measures and *indirect* measures of some discrimination. Direct measures involve tasks in which the instructions make explicit reference to the relevant discrimination, and include objective measures such a recognition or recall. In contrast, indirect measures, such as stem completion in memory tasks, make no reference to the relevant discrimination. By assumption, direct measures should exhibit greater or equal sensitivity than indirect measures to consciously held task-relevant information, for subjects should be expected to be more successful in using conscious information when instructed to do so than when not. Hence, demonstrating that an indirect task is more sensitive to some information than a comparable direct task can only be interpreted as indicating unconscious influences on performance. This 'relative sensitivity approach' has been successfully applied in the study of subliminal perception and implicit memory. Jiménez *et al*. (1996) first applied this logic to the study of implicit learning by comparing direct and indirect measures of the knowledge acquired by participants trained on an SL (sequence learning) task. Using detailed correlational analyses, Jiménez *et al*. were able to show that some sequence knowledge tended to be expressed exclusively through choice reaction decisions.

The second approach – Larry Jacoby's 'Process Dissociation Procedure' (Jacoby, 1991) – is based on the argument that, just as direct measures can be contaminated by unconscious influences, indirect measures can likewise be contaminated by conscious influences: particular tasks cannot simply be identified with particular underlying processes. The process dissociation procedure thus aims to tease apart the relative contributions of conscious and unconscious influences on performance. To do so, two conditions are compared in which conscious and unconscious influences either both contribute to performance improvement, or act

against each other. For instance, subjects might be asked to memorize a list of words and then, after some delay, to perform a stem completion task in which word stems are to be completed either so as to form one of the words memorized earlier (the *inclusion* condition) or so as to form a different word (the *exclusion* condition). If the stems nevertheless tend to be completed by memorized words under exclusion instructions, then one can only conclude that memory for these words was implicit, since if subjects had been able to consciously recollect them, they would have avoided using them to complete the stems. Numerous experiments have now been designed using the process dissociation procedure. They collectively offer convincing evidence that performance can be influenced by unconscious information in the absence of conscious awareness. In the context of implicit learning research, Destrebecqz and Cleeremans (2001) adapted the process dissociation procedure to sequence learning, asking trained participants to either generate a sequence that resembled the training sequence (inclusion) or a sequence that was as different as possible from the training sequence (exclusion). Results indicated that under certain conditions, participants were unable to exclude familiar sequence fragments, thus suggesting that they had no control over the knowledge acquired during training. Destrebecqz and Cleeremans (2001) concluded that this knowledge was best described as implicit, for its expression is not under conscious control.

Despite the considerable methodological advances achieved over the past decade or so, assessing awareness in implicit learning and related fields remains particularly challenging. The central issue of the extent to which information processing can occur in the absence of conscious awareness remains as controversial today as it was in 1967.

The Conceptual Challenge

The third conceptual challenge is to determine how best to interpret existing dissociation results. Dunn and Kirsner (1988) pointed out that even crossed double dissociations between two tasks do not necessarily indicate the involvement of separable, independent processes. Other authors have appealed to theoretical and simulation work to call the dissociation logic into question. Many authors have described non-modular architectures that can nevertheless produce double dissociations. Plaut (1995) explored these issues in the context of cognitive neuropsychology. In a compelling series of simulation studies, Plaut not only showed that lesioning a single connectionist network in various ways could account for the double dissociations between concrete and abstract word reading exhibited by deep dyslexic patients, but also that lesions in a single site produced *both* patterns of dissociations observed with patients. In other words, the observed dissociations can clearly not be attributed to architectural specialization, but can instead be a consequence of *functional specialization* (functional modularity) in the representational system of the network. These issues are also debated in the context of implicit learning research.

THE ROLE OF COMPUTATIONAL MODELING

Computational modeling has played a central role in deconstructing early verbal theories of implicit learning (1) by offering 'proof of existence' demonstrations that elementary, associative learning processes (as opposed to rule-based learning) are in fact often sufficient to account for the data, (2) by making it possible to cast specific predictions that can then be contrasted with those of competing models, and (3) by making it possible to explore how specific computational principles can offer novel, unitary accounts of the data.

Detailed computational models have now been proposed for all three main paradigms of implicit learning. Two families of models are currently most influential: neural network models and fragment-based models. Neural network models typically consist of simple auto-associator models (Dienes, 1992) or of networks capable of processing sequences of events, such as the simple recurrent network (SRN, see Figure 2) introduced by Elman

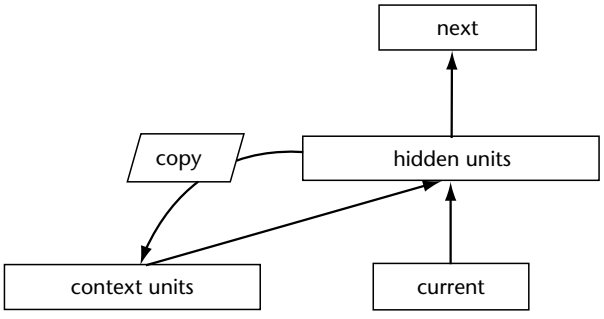


Figure 2. The simple recurrent network (SRN) introduced by Elman (1990). The network takes the current element of a sequence as input, and is trained to predict the next element using back propagation. Context units, which on timestep contain a copy of the activation pattern that existed over the network's hidden units on the previous timestep, enable previous information to influence current predictions.

(1990) and applied to sequence learning by Cleeremans and McClelland (1991). Such models assume that over the course of training, information about the statistical structure of the stimulus material is stored in the connection weights between the model's processing units. This information is thus subsymbolic to the extent that it is embedded in the same structures that are used to support processing itself. Fragment-based models (e.g. Perruchet and Vinter, 1998), in contrast, are variants of exemplar-based models which assume that learning results in the acquisition of memory structures such as whole exemplars or fragments thereof.

While no type of model can currently claim generality, both approaches share a number of central assumptions: (1) learning involves elementary association or recoding processes that are highly sensitive to the statistical features of the training set, (2) learning is viewed essentially as a mandatory by-product of ongoing processing, (3) learning is based on the processing of *exemplars* and produces *distributed* knowledge, and (4) learning is *unsupervised* and *self-organizing*.

More recently, hybrid models that specifically attempt to capture the relationships between symbolic and subsymbolic processes in learning have also been proposed. Sun (2001), for instance, has introduced models that specifically attempt to link the subsymbolic, associative, statistics-based processes characteristic of implicit learning with the symbolic, declarative, rule-based processes characteristic of explicit learning.

All these models have been essentially directed at addressing two questions: (1) What knowledge do people acquire in implicit learning situations? and (2) To what extent should demonstrated dissociations be interpreted as reflecting the involvement of separable learning systems?

Rules vs. Statistics

Early characterizations of implicit knowledge have tended to describe it as abstract based essentially on findings that subjects exhibit better-than-chance transfer performance, as when asked to make grammaticality judgments on novel strings in the context of artificial grammar learning situations (Reber, 1989). Likewise, it has often been assumed that the reaction time savings observed in sequence learning tasks reflect the acquisition of 'deep' knowledge about the rules used to generate the stimulus material (Lewicki *et al.*, 1987). These abstractionist accounts have generally left unspecified what the exact form of the acquired knowledge may be, short of noting that it must somehow represent the structure of the stimuli and their relationships, and be independent of the surface features of the material. The latter claim was further substantiated by findings that artificial grammar learning knowledge transfers to strings based on the same grammar but instantiated with a different letter set, or even across modalities, as when training involves letter strings but transfer involves tone sequences.

However, there now is considerable evidence that non-abstractionist mechanisms are largely sufficient to account for the data. Brooks (1978) first suggested that subjects in artificial grammar learning experiments were classifying novel strings based not on abstract knowledge of the rules, but simply on the extent to which novel grammatical or ungrammatical strings are similar to *whole exemplars* memorized during training. Perruchet and colleagues (Perruchet and Pacteau, 1990) showed that the knowledge acquired in both artificial grammar learning and sequence learning tasks might consist of little more than explicitly memorized short fragments or chunks of the training material such as bigrams or trigrams, or simple frequency counts. Both learning and transfer performance can then be accounted for by the extent to which novel material contains memorized chunks. Figure 3 illustrates some of the computational possibilities that have been suggested in the context of artificial grammar learning tasks, ranging from purely exemplar-based approaches to neural network models.

More recently, accounts that assume separate memory systems for representing general or specific knowledge in artificial grammar learning tasks have been proposed based on evidence that significant sensitivity to grammaticality remains even when similarity and fragment overlap is carefully controlled.

Overall, while it is clear that the knowledge acquired in typical implicit learning situations need not be based on the unconscious acquisition of symbolic rules, significant areas of debate remain about the extent to which unitary, fragment-based mechanisms are sufficient to account for sensitivity to both the general and specific features of the training material. Simulation models have generally been suggestive that such mechanisms are in fact sufficient to account simultaneously for both grammaticality and similarity effects.

Based on these properties of successful models of implicit learning, it is appealing to consider it as a complex form of priming whereby experience continuously shapes memory, and through which stored traces in turn continuously influence further

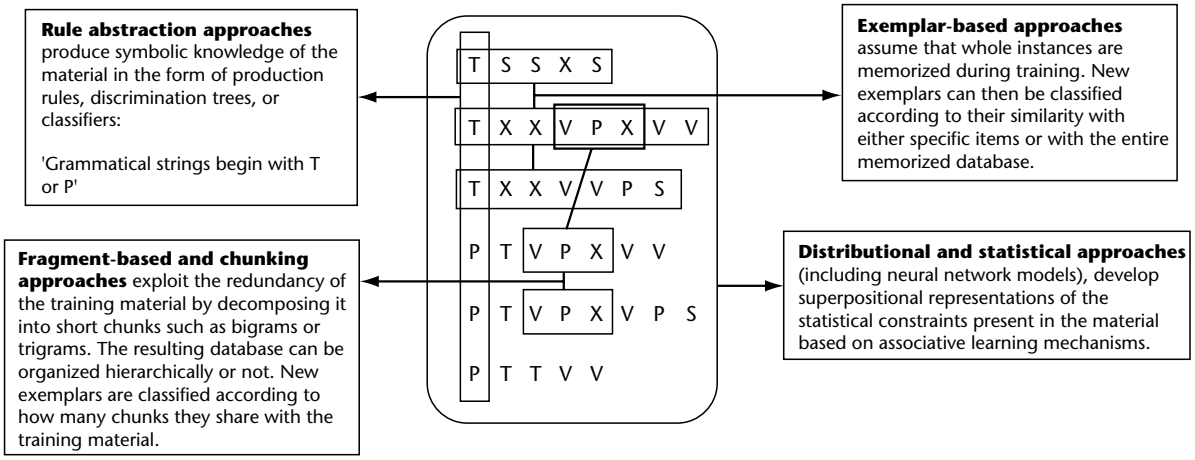


Figure 3. A representation of different computational approaches to artificial grammar learning.

processing. Such priming, far from involving the sort of passive and automatic acquisition of abstract structure that were previously assumed to lie at the heart of implicit learning, is in fact highly dependent on task demands and attentional processing during acquisition, as well as on the congruence between learning and transfer conditions (Whittlesea and Dorken, 1993).

Finally, while both fragment-based and neural network models make it clear how sensitivity to the distributional properties of an ensemble of stimuli can emerge out of the processing of exemplars, they differ in whether they assume that the shared features of the training materials are represented as such or merely computed when needed. This *locus of abstraction* issue is a difficult one that is unlikely to be resolved by modeling alone. Thus overall, it appears that the knowledge acquired through implicit learning is best described as lying somewhere on a continuum between purely exemplar-based representations and more general, abstract representations – a characteristic that neural network models are particularly apt at capturing.

Separable Systems?

Dissociations between implicit and explicit learning or processing have often been interpreted as suggesting the existence of separable memory systems. For instance, Squire and collaborators have shown that artificial grammar learning is largely preserved in amnesia (e.g. Knowlton *et al*., 1992), to the extent that amnesic patients perform at the same level as normal controls when asked to classify strings as grammatical or not, but are severely impaired when asked to discriminate between familiar and novel instances (or fragments) of the strings. These results suggest that the processes that subtend declarative and non-declarative memory depend on separable brain systems respectively dedicated to representing either information about the specific features of each encountered exemplar on the one hand (the hippocampus and related structures), and information about the features shared by many exemplars on the other hand (the neocortex).

In this case also, however, computational modeling often casts the empirical findings in a different light. For instance, Kinder and Shanks (2001) were able to simulate the observed dissociations by tuning a single parameter (the learning rate) in a simple recurrent network trained on the same material as used in the behavioral studies, and therefore concluded that a single-system account is in fact sufficient to account for the data.

CONCLUSIONS

The study of differences between implicit and explicit processing is a major endeavor for the cognitive neurosciences. Indeed, as our knowledge of how the brain works accumulates, our knowledge about how the mind works is changing rapidly. In particular, many existing distinctions previously described in purely functional, binary terms, such as the implicit–explicit distinction, the controlled–automatic distinction, or the declarative–procedural distinction, are now being explained anew in terms of graded characterizations of the computational problems that corresponding

brain systems have been evolved to solve. In this respect, the study of implicit learning, from the perspective of computational cognitive neuroscience, has a bright future, for it is through the development of sensitive paradigms through which to explore the differences between conscious and unconscious cognition that one can best contribute to the search for the neural, behavioral, and computational correlates of consciousness.

References

Berry DC and Broadbent DE (1984) On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology* **36**: 209–231.

Berry DC and Dienes Z (1993) *Implicit Learning: Theoretical and Empirical Issues*. Hove, UK: Lawrence Erlbaum Associates.

Brooks LR (1978) Non-analytic concept formation and memory for instances. In: Rosch E and Lloyd B (eds) *Cognition and Concepts*, pp. 16–211. Mahwah, NJ: Lawrence Erlbaum Associates.

Butler LT and Berry DC (2001) Implicit memory: intention and awareness revisited. *Trends in Cognitive Sciences* **5**: 192–197.

Cheesman J and Merikle PM (1984) Priming with and without awareness. *Perception and Psychophysics* **36**: 87–395.

Cleeremans A and McClelland JL (1991) Learning the structure of event sequences. *Journal of Experimental Psychology: General* **120**: 235–253.

Cleeremans A, Destrebecqz A and Boyer M (1998) Implicit learning: news from the front. *Trends in Cognitive Sciences* **2**: 406–416.

Clegg BA, DiGirolamo GJ and Keele SW (1998) Sequence learning. *Trends in Cognitive Sciences* **2**: 75–281.

Destrebecqz A and Cleeremans A (2001) Can sequence learning be implicit? New evidence with the Process Dissociation Procedure. *Psychonomic Bulletin & Review* **8**: 343–350.

Dienes Z (1992) Connectionist and memory-array models of artificial grammar learning. *Cognitive Science* **16**: 41–79.

Dunn JC and Kirsner K (1988) Discovering functionally independent mental process: the principle of reversed association. *Psychological Review* **95**: 91–101.

Elman JL (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Frensch PA (1998) One concept, multiple meanings: on how to define the concept of implicit learning. In: Stadler MA and Frensch PA (eds) *Handbook of Implicit Learning*, pp. 47–104. Thousand Oaks, CA: Sage Publications.

Jacoby LL (1991) A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language* **30**: 513–541.

Jiménez L, Mendéz C and Cleeremans A (1996) Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* **22**: 948–969.

Kinder A and Shanks DR (2001) Amnesia and the Declarative/Nondeclarative distinction: a recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience* **13**: 648–669.

Knowlton BJ, Ramus SJ and Squire LR (1992) Intact artificial grammar learning in amnesia: dissociation of classification learning and explicit memory for specific instances. *Psychological Science* **3**: 172–179.

Lewicki P, Czyzewska M and Hoffman H (1987) Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition* **13**: 523–530.

Nisbett RE and Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review* **84**: 231–259.

Nissen MJ and Bullemer P (1987) Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology* **19**: 1–32.

Perruchet P and Pacteau C (1990) Synthetic grammar learning: implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General* **119**: 264–275.

Perruchet P and Vinter A (1998) PARSER: a model for word segmentation. *Journal of Memory and Language* **39**: 246–263.

Plaut DC (1995) Double dissociation without modularity: evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology* **17**: 291–326.

Reber AS (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* **6**: 855–863.

Reber AS (1989) Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* **118**: 219–235.

Reingold EM and Merikle PM (1988) Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* **44**: 563–575.

Schacter DL (1987) Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**: 501–518.

Shanks DR and St. John MF (1994) Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* **17**: 367–447.

Sun R (2001) *Duality of the Mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

Whittlesea BWA and Dorken MD (1993) Incidentally, things in general are particularly determined: an episodic-processing account of implicit learning. *Journal of Experimental Psychology: General* **122**: 227–248.

Further Reading

Berry DC (1997) *How Implicit is Implicit Learning?* Oxford, UK: Oxford University Press.

Berry DC and Dienes Z (1993) *Implicit Learning: Theoretical and Empirical Issues*. Hove, UK: Lawrence Erlbaum Associates.

Cleeremans A (1993) *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing*. Cambridge, MA: MIT Press.

French RM and Cleeremans A (2002) *Implicit Learning and Consciousness: An Empirical, Computational and Philosophical Consensus in the Making*. Hove, UK: Psychology Press.

Reber AS (1993) *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford, UK: Oxford University Press.

Stadler MA and French PA (1998) *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage Publications.

Inference using Formal Logics

Intermediate article

David McAllester, AT&T Labs-Research, Florham Park, New Jersey, USA

CONTENTS

Formal logics are used fundamentally in mathematics, natural language syntax and semantics, computer programming languages, and artificial intelligence.

INTRODUCTION

Logic is fundamental to a variety of disciplines. Logic provides insight into the nature of mathematics and human mathematical reasoning. Logic provides insight into the syntax and semantics of human language. Logic is an important tool in the design and implementation of computer programming languages. Logic also holds the promise of playing an important role in the development of artificial intelligence (AI). We will briefly consider each of these motivations in more detail.

Metamathematics

The idea that human thought is governed by linguistic rules started with the development of syllogisms by Aristotle. A syllogism is a pattern of inference. For example, if all men are mortal, and Socrates is a man, then Socrates is mortal. The general pattern is, of course, if all X are Y and Z is an X, then Z is a Y. In modern terminology, such a pattern is called a 'rule of inference'. The study of inference rules, begun by Aristotle, led to the development in the 1920s of the system of rules and axioms known as ZFC – Zermelo–Fränkel set theory with the axiom of choice. It is now generally accepted by mathematicians that all the provable truths of mathematics can be formally derived from the nine axioms of ZFC using the inference rules of first-order logic. Gödel's incompleteness theorem, proved in 1930, implies that if ZFC is sound then there are true statements not provable in ZFC. Gödel's theorem is consistent with the idea that the theorems provable by human mathematicians are precisely those provable in ZFC: true statements unprovable in ZFC are presumably also unprovable by human mathematicians. Gödel's theorem is itself provable in ZFC. A general treatment of logic as metamathematics can be found in Bell and Machover (1977).

Metaprogramming

A compiler is a computer program that takes as input another computer program and compiles that input into low-level instructions executable by computer hardware. A compiler is a metaprogram – a program that manipulates and analyzes other programs. Logic plays an important role in metaprogramming. Computer programs are composed of symbolic expressions, which have well-defined meanings (which may be determined by the meaning of their subsidiary expressions). (*See* **Symbol Systems**; **Computation, Philosophical Issues about**)

Compilers often analyze (reason about) the programs being compiled. For example, many compilers can determine that a pointer variable must be non-null (not equal to zero) at a certain place in the program. Logic is fundamental to such software analysis. A survey of the theory and practice of compilers can be found in Appel (1997).

Human Language

All human languages have sentences which form 'complete thoughts'. Furthermore, in all languages sentences have phrase structure, e.g. a sentence may be composed of a noun phrase and a verb phrase each of which in turn can be composed of subsidiary phrases. It seems that the meaning of a sentence is constructed from the meanings of its phrases and the meaning of a phrase is constructed from the meaning of its subsidiary phrases. The phrase 'book on the table' has a meaning that is determined by the meaning of the word 'book' and

the meaning of the phrase 'on the table'. The ability to combine meanings by putting words together into phrases, and phrases together into larger phrases, allows for a very large number of different ideas to be expressed – the number of ideas expressible in a fixed number of words grows exponentially in the number of words allowed. Logic provides a clear (though approximate) mathematical model of the way meanings are assigned to phrases and the way the meaning of a phrase is determined by the meanings of its subphrases. A detailed study of the logical treatment of human language can be found in Carpenter (1997). (*See* **Categorial Grammar and Formal Semantics**; **Generalized Quantifiers**)

Knowledge Representation

Since the early days of AI, logic has been proposed as a foundation for the representation of knowledge about the world (Lifschitz, 1998). Logic-based mechanical deduction has not produced human-level intelligence in machines. But alternatives such as neural networks, genetic algorithms, and reinforcement learning have also failed to produce human-level intelligence (e.g. machines that that can have sustained intelligent conversations in human language). Because of the fundamental importance of logic to metamathematics, metaprogramming, human language, and knowledge representation, logic remains relevant to the study of intelligence. (*See* **Computation, Formal Models of**; **Knowledge Representation**; **Knowledge Representation, Psychology of**)

WHAT IS LOGIC?

Logic is the study of the manipulation of symbolic expressions. A symbolic expression is a concise representation of a much larger (often infinite) set of behaviors or responses. For example, the symbolic expression 'x^2+2x+1' is a concise representation of an infinite input–output relation: the relation mapping input x to output x^2+2x+1. Algebraic rules of inference allow one to deduce that if $2y+4=4x+6$ then $y=2x+1$. The symbolic expression '$2y+4=4x+6$' is a proposition – an expression that, given values for its variables, is either true or false. To say that a proposition Φ 'entails' a proposition Ψ means that in any situation where Φ is true, Ψ must also be true. For example, $2y+4=4x+6$ entails $y=2x+1$.

Formally, a symbolic expression can be defined as a tree where each node is labeled by a character string. For example, the expression '$2y+4=4x+6$' is a tree whose root is labeled with '$=$' and has two children, one representing '$2y+4$' and one representing '$4x+6$'. An expression whose root is labeled with the string f and with children trees $t_1, \ldots, t_n$ is often written as $f(t_1, \ldots, t_n)$. For example, the expression $2y+4=4x+6$ can be written as '$=(+(*(2,y),4),+(*(4,x),6))$'. The leaves of this tree are labeled with the numerals '6', '4' and '2' and the variables 'x' and 'y'. The internal nodes are labeled with '$=$', '$+$' and '$*$'.

An expression is usually used to represent a function whose inputs are the variables occurring in the expression and whose output is the value of the expression. There are many different ways of assigning values to expressions. Different ways of evaluating expressions correspond to different logics. There are logics of time, modal logics, logics of knowledge, logics of communicating systems, logics of imperative computer programs, and many more. Historically, however, the best-known logic is first-order logic.

FIRST-ORDER LOGIC

In first-order logic, one assumes a set of variables, a set of constants, a set of function symbols, and a set of predicate symbols where each function and predicate symbol is associated with a specified 'arity' (number of arguments). There are two types of expressions: terms and formulas. The terms t and formulas Φ can be defined by the following grammar:

$$\begin{aligned} t &:= x|c|f(t_1, \ldots, t_n) \\ \Phi &:= P(t_1, \ldots, t_n)|t_1 = t_2|\Phi_1 \vee \Phi_2|\Phi_1 \wedge \Phi_2| \\ &\quad \Phi_1 \rightarrow \Phi_2|\neg\Phi|\forall x\Phi|\exists x\Phi \end{aligned} \tag{1}$$

In the grammar for terms, x ranges over variables, c ranges over constants, and f ranges over function symbols. (We assume some unspecified way of classifying character strings into variables, constants and function symbols with specified arities.) For example, if f is a function symbol of two arguments, g is a function symbol of one argument, x and y are variables, and b and c are constants, then $f(b,f(g(x),f(c,y)))$ is a term. In the grammar for formulas, P ranges over predicate symbols.

The defining feature of most logics is the 'semantics' assigned to the expressions, i.e. the way in which values are assigned to expressions. In the case of first-order logic, the semantics is defined relative to a 'structure'. A structure $\mathcal{M}$ consists of: a set $D_{\mathcal{M}}$ called the domain of $\mathcal{M}$; a meaning $\mathcal{M}(c) \in D_{\mathcal{M}}$ for each constant c; a meaning $\mathcal{M}(f)$ for each function symbol f, such that if f has arity n then $\mathcal{M}(f)$ is a function taking n arguments in

$D_{\mathcal{M}}$ and returning an element of $D_{\mathcal{M}}$; and finally a meaning $\mathcal{M}(P)$ for each predicate symbol P, such that if P has arity n then $\mathcal{M}(P)$ is a function taking n arguments in $D_{\mathcal{M}}$ and returning a truth value (either T or F). Meaning can be assigned to terms and formulas if, in addition to the meanings provided by the model, one is also given a meaning for each variable. A variable interpretation ρ relative to a structure $\mathcal{M}$ is a mapping from variables to elements of $D_{\mathcal{M}}$. For any structure $\mathcal{M}$ and variable interpretation ρ relative to $\mathcal{M}$, and any term t, we can define the value of the term t under $\mathcal{M}$ and ρ, $V(t, M, \rho)$, by the following equations:

$$\begin{aligned} &V(c, \mathcal{M}, \rho) = \mathcal{M}(c) \\ &V(x, \mathcal{M}, \rho) = \rho(x) \\ &V(f(t_1, \ldots, t_n), \mathcal{M}, \rho) = \mathcal{M}(f)(V(t_1, \mathcal{M}, \rho), \ldots, \\ &\quad V(t_n, \mathcal{M}, \rho)) \end{aligned} \tag{2}$$

As an example, let $\mathcal{N}$ be a structure whose domain $D_{\mathcal{N}}$ is the integers, that interprets the binary function symbols '$+$' and '$*$' as addition and multiplication respectively, and that interprets numerical constants such as '2' and '7' as the appropriate integers in $D_{\mathcal{N}}$. For any variable interpretation ρ and domain element d we write $\rho\ [x{:=}d]$ for the variable interpretation identical to ρ except that it maps the variable x to the value d. As an example of the semantic evaluation of terms, we have:

$$V(+(*(2, x), 3), \mathcal{N}, \rho[x := d]) = 2d + 3 \tag{3}$$

While the value of a term is a domain element, the value of a formula is a truth value. For a given structure $\mathcal{M}$, and variable interpretation ρ relative to $\mathcal{M}$, the values of formulas are defined by the following equations:

$$\begin{aligned} &V(P(t_1, \ldots, t_n), \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ \mathcal{M}(P) \\ &\quad (V(t_1, \mathcal{M}, \rho), \ldots, V(t_n, \mathcal{M}, \rho)) = \mathrm{T} \\ &V(t_1 = t_2, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ V(t_1, \mathcal{M}, \rho) = \\ &\quad V(t_2, \mathcal{M}, \rho) \\ &V(\Phi_1 \vee \Phi_2, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ V(\Phi_1, \mathcal{M}, \rho) = \mathrm{T} \\ &\quad \text{or } V(\Phi_2, \mathcal{M}, \rho) = \mathrm{T} \\ &V(\Phi_1 \wedge \Phi_2, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ V(\Phi_1, \mathcal{M}, \rho) = \mathrm{T} \\ &\quad \text{and } V(\Phi_2, \mathcal{M}, \rho) = \mathrm{T} \\ &V(\Phi_1 \rightarrow \Phi_2, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ V(\Phi_1, \mathcal{M}, \rho) = \mathrm{F} \\ &\quad \text{or } V(\Phi_2, \mathcal{M}, \rho) = \mathrm{T} \\ &V(\neg\Phi, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ V(\Phi, \mathcal{M}, \rho) = \mathrm{F} \\ &V(\forall x\Phi, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ \text{for all } d \in D_{\mathcal{M}}, \\ &\quad V(\Phi, \mathcal{M}, \rho[x := \mathrm{d}]) = \mathrm{T} \\ &V(\exists x\Phi, \mathcal{M}, \rho) = \mathrm{T} \ \text{ iff } \ \text{there exists } d \in D_{\mathcal{M}} \\ &\quad \text{such that } V(\Phi, \mathcal{M}, \rho[x := d]) = \mathrm{T} \end{aligned} \tag{4}$$

Inference is a computational process for determining symbolic consequences of symbolic assertions. For example, consider the following formula:

$$\forall x, y, z(x \leq y \wedge y \leq z) \rightarrow x \leq z \tag{5}$$

Here $\forall x, y, z\ \Phi$ is being used as an abbreviation for nested universal quantification and $x \leq y$ is being used as a more legible notation for the formula $\leq(x, y)$ where $\leq$ is a binary predicate symbol. Since $\leq$ is a binary predicate symbol, in any structure $\mathcal{M}$, $\mathcal{M}(\leq)$ must be a binary relation on the set $D_{\mathcal{M}}$. Formula 5 is true in structure $\mathcal{M}$ provided that $\mathcal{M}(\leq)$ is transitive. Note that all variables in formula 5 are quantified – we say that all variable occurrences are 'bound'. A variable occurrence not bound by a quantifier is called a 'free occurrence'. A formula with no free occurrences of variables is called 'closed'.

The truth value of a closed formula (such as formula 5) is determined by the structure and is independent of the variable interpretation. Formula 5 is true in a structure whose domain is the natural numbers and that interprets the predicate symbol $\leq$ as the standard ordering predicate on natural numbers. However, formula 5 can be false in other structures.

We say that formula Φ 'entails' formula Ψ (written $\Phi \models \Psi$) if, for any pair $\langle\, \mathcal{M}, \rho \,\rangle$ such that $V(\Phi, \mathcal{M}, \rho) = \mathrm{T}$ we have $V(\Psi, \mathcal{M}, \rho) = \mathrm{T}$. If Σ is a set of formulas then we say that a pair $\langle\, \mathcal{M}, \rho \,\rangle$ satisfies Σ if $V(\Phi, \mathcal{M}, \rho)) = \mathrm{T}$ for every Φ in Σ. We say that Σ entails Ψ (written $\Sigma \models \Phi$) if for every pair $\langle\, \mathcal{M}, \rho \,\rangle$ satisfying Σ we have $V(\Psi, \mathcal{M}, \rho) = \mathrm{T}$. As we shall see below, for a given set Σ of formulas representing 'knowledge' of some sort, and given a question formula Φ, symbolic processing (theorem proving) can, in many cases, determine whether $\Sigma \models \Phi$. Now we turn to the significance of this symbolic processing in early logicist models of AI.

THE SITUATION CALCULUS

In logicist models of artificial intelligence, human world knowledge is modeled by a large set of first-order formulas (Lifschitz, 1998). These formulas are taken to express 'commonsense knowledge'. A fundamental concept in early AI models is the notion of a situation. Intuitively, a situation is a certain place at a certain time: for example, the coffee room in the mathematics department at about 4 p.m. on a certain day. In addition to variables that range over situations, early AI models often used variables and terms intended to denote actions. An action is something that an agent can choose to do. The formula CAN (x, a, s) can be used to represent the statement that in situation s agent x can choose to take action a. The term RESULT (x, a, s) can be used to represent the situation that results from agent x performing action a in

situation s. The formula $\mathrm{ON}(x,y,s)$ can be used to represent the assertion that object x is on object y in situation s. An example of a formula expressing world knowledge might be the following:

$$\forall x, y, z, s\ \mathrm{CAN}(x, \mathrm{PUTON}(y, z), s) \rightarrow \mathrm{ON}(y, z, \mathrm{RESULT}(x, \mathrm{PUTON}(y, z), s)) \quad (6)$$

In the early logicist models, symbolic inference from world knowledge is presumed to be the computational method by which people 'understand' the world, i.e. the way they make predictions, answer questions, and plan actions. (*See* **Learning Rules and Productions**; **Deductive Reasoning**)

RESOLUTION THEOREM PROVING

We have defined semantic entailment for first-order formulas, but we have not defined any corresponding computational process. Early AI research focused on the resolution principle as a means of mechanizing general-purpose symbolic inference. Resolution is based on the notion of a clause. Formula 5, expressing the transitivity of the predicate $\leq$, is equivalent to the following:

$$\forall x, y, z\ \neg(x \leq y) \vee \neg(y \leq z) \vee x \leq z \quad (7)$$

To see the equivalence, one should try to verify that in any model where formula 5 is true, formula 7 is true, and vice versa. Formula 7 is in a special form, called a clause. Formally, an atom is defined to be a formula of the form $P(t_1, \ldots, t_n)$ and a literal is defined to be either an atom or the negation of an atom. The formula $x \leq z$ is an atom (which is a special case of a literal) and the formula $\neg(x \leq y)$ is a literal. A clause is defined to be a closed formula of the following form, where each Φ_k is a literal:

$$\forall x_1, \ldots, x_n\ \Phi_1 \vee \Phi_2 \vee \ldots \vee \Phi_k \quad (8)$$

It turns out that, in a technical sense, clauses suffice for arbitrary first-order knowledge representation and reasoning. To state the technical sense in which this is true, we will introduce some notation. For closed formulas Φ and Ψ we have $\Phi \models \Psi$ if and only if there is no model $\mathcal{M}$ such that $V(\Phi, \mathcal{M}) = \mathrm{T}$ but $V(\Psi, \mathcal{M}) = \mathrm{F}$, or equivalently, the formula $\Phi \wedge \neg\Psi$ is unsatisfiable (is not true in any model). This means that any question of entailment can be reduced to a question of satisfiability. The following lemma implies that any question of entailment can be reduced to a question about clauses.

Lemma 1. Any closed first-order formula Φ can be converted, in time linear in the size of Φ, into a finite set Σ of clauses such that Φ is satisfiable if and only if Σ is satisfiable (there is a model satisfying all of the clauses in Σ).

We will not prove this lemma here, but the basic idea is to introduce predicates representing subformulas of Φ. For example, if $\Psi[x, y]$ is a subformula of Φ involving the variables x and y then one introduces a predicate P_Ψ and adds clauses to Σ ensuring that in any model of the clauses we have that $P_\Psi(x, y)$ is true if and only if $\Psi[x, y]$ is true. The clauses defining the predicate for $\Psi_1 \wedge \Psi_2$ can be stated in terms of the predicates for Ψ_1 and Ψ_2. A similar construction applies to other formula types.

When a clause representation is desired, world knowledge is often written directly as clauses and no mechanical translation is used. The existence of a mechanical translation shows, however, that clauses suffice as a representation language.

'Resolution' is a mechanical inference procedure for deriving new clauses from a given set of clauses. It can be viewed as composed of two inference rules: factoring and binary resolution (these two rules can be combined into a single rule, with some loss of clarity). In discussing these rules, we will omit the universal quantifier in front of clauses, but it should be understood that the variables in a clause are universally quantified as in formula 8. (*See* **Resolution Theorem Proving**)

The factoring rule can be applied to clauses where two of the literals in the clause can be interpreted as the same literal. For example, consider the clause $P(x, y) \vee P(y, x)$. By considering the case where $x = y$ we can derive the clause $P(x, x)$. To formally state the general factoring rule we first need to define some terminology. A substitution is a mapping from variables to terms. For any substitution σ and literal Φ we define $\sigma[\Phi]$ to be the result of replacing each variable x in Φ by $\sigma(x)$. Two literals Φ_1 and Φ_2 are unifiable if they can be interpreted as the same literal, i.e. if there exists a substitution σ such that $\sigma[\Phi_1]$ is the same literal as $\sigma[\Phi_2]$. In this case σ is called a 'unifier' of Φ_1 and Φ_2. If Φ_1 and Φ_2 are unifiable, then a 'most general unifier' of Φ_1 and Φ_2 is a unifier γ such that any other unifier σ is a specialization of γ, i.e. there exists a substitution δ such that for any variable x we have that $\delta[\gamma(x)] = \sigma(x)$. The most general unifier of two literals is unique up to an arbitrary renaming of variables in the range of the substitution, and can be computed in time linear in the size of the two given literals. The general form of the factoring rule states that, given a clause $\Phi_1 \vee \ldots \vee \Phi_n$ containing two unifiable literals Φ_i and Φ_j, one can derive the clause $\gamma[\Phi_1] \vee \ldots \vee \gamma[\Phi_n]$ where γ is the most general unifier of Φ_i and Φ_j. The resulting clause always has fewer literals.

The other fundamental inference rule of resolution theorem proving is binary resolution. Here

we will assume that the literals in a clause can be reordered arbitrarily (a clause is viewed as a set of literals), and that the variables in a clause can be renamed arbitrarily at any time (they are universally quantified). The binary resolution rule states that given two clauses

$$\begin{array}{l}\Phi_1 \vee \ldots \Phi_{n-1} \vee \Phi_n \\ \neg\Psi_1 \vee \Psi_2 \ldots \vee \Psi_k\end{array} \tag{9}$$

where no variable occurs in both clauses (the second clause can be renamed to make this true) and where Φ_n and $\neg\Psi_1$ are unifiable, one can derive the following, where γ is the most general unifier of Φ_n and $\neg\Psi_1$:

$$\gamma[\Phi_1] \vee \ldots \vee \gamma[\Phi_{n-1}] \vee \gamma[\Psi_2] \vee \ldots \vee \gamma[\Psi_k] \tag{10}$$

As an example, consider the simple statement of transitivity in formula 5. Converting this formula to the clause 7 and applying binary resolution between this clause and a renamed version of itself results in the following:

$$\neg(x \leq y) \vee \neg(y \leq z) \vee \neg(z \leq w) \vee x \leq w \tag{11}$$

This clause is equivalent to the following implication:

$$x \leq y \wedge y \leq z \wedge z \leq w \rightarrow x \leq w \tag{12}$$

So the resolution rule can be used to compose the transitivity clause with itself to yield a three-step transitivity statement.

As another very simple example, suppose we want to show that $P \wedge (P \rightarrow Q) \models Q$. Here P and Q are predicate symbols of no arguments, i.e. they are formula constants. This can be formulated as the problem of showing that the set of formulas $\{P, P \rightarrow Q, \neg Q\}$ is unsatisfiable. The formula $P \rightarrow Q$ can be replaced by the clause $\neg P \vee Q$. The binary resolution rule can be applied to the clause P and the clause $\neg P \vee Q$ to yield the clause Q. The binary resolution rule can then be applied to the clause Q and the clause $\neg Q$ to yield the empty clause. A derivation of the empty clause shows that the given clause set is unsatisfiable; hence the original entailment holds.

The resolution rules are complete in the following sense.

Theorem 2. If Σ is an unsatisfiable set of clauses then the rules of factoring and binary resolution can be used to derive the empty clause.

The proof of this theorem is somewhat involved, and will not be given here – a complete proof can be found in Leitsch, (1997). Lemma 1 and theorem 2 show that resolution theorem proving is, in a technical sense, sufficient for first-order inference.

Researchers soon discovered that mechanical resolution theorem proving was in many cases impractical: while the empty clause was derivable in principle, the system might take the age of the universe to derive it. Over the years there has been much research into methods for implementing more efficient variants of resolution theorem proving. Today, resolution theorem provers are very good at solving certain kinds of problems particularly amenable to first-order reasoning Leitsch (1997). However, first-order logic itself seems to be of limited value in most important applications of logic. Other logics seem more relevant.

THE BOYER–MOORE LOGIC

As mentioned above, logics differ in the way they associate values with expressions. One way of assigning values to expressions is inspired by computer programming. Most programming languages allow expressions that include procedure calls. For example, if f is a defined procedure then the expression '$x + f(y)$' computes to the sum of the value of x and the value returned by the procedure f when applied to the argument y. As in programming languages, the Boyer–Moore logic requires function and predicate symbols (procedure names) to be defined: the precise input–output relation of each function and predicate symbol must be completely specified. This is not true of first-order logic, where, for example, formula 5 can be assumed and used without stating any definition of the meaning of the predicate $\leq$. In the Boyer–Moore logic, the specification of the meaning of a function symbol is directly analogous to the definition of a procedure in a programming language. In fact, the Boyer–Moore logic is a kind of programming language together with inference rules for proving general assertions, such as $\forall x \quad f(f(x)) = f(x)$ where f is a defined procedure. Note that no amount of testing formally establishes a desired property on all of the (infinitely many) possible inputs to a procedure. Establishing the truth of a statement for all inputs requires logical inference.

The Boyer–Moore logic, and its descendent ACL2 (Kaufmann *et al*., 2000) are complex languages with many features that have been used for a large number of mathematical proofs and hardware and software verifications. Here we will explain the essential ideas of this logic by presenting a simpler but similar logic, the 'simplified Boyer–Moore logic' (SBM). The terms of SBM are defined by the following grammar, where "s"

ranges over character string constants, `if`, `pair`, Π_1, Π_2, `stringp`, and `equal` are special function symbols (primitives of the language), and f ranges over procedure names.

$$t ::= \text{“}s\text{”} \mid x \mid \texttt{if}(t_1,t_2,t_3) \mid \texttt{pair}(t_1,t_1) \mid \Pi_1(t) \mid \Pi_2(t) \mid \texttt{stringp}(t) \mid \texttt{equal}(t_1,t_2) \mid f(t_1,\ldots,t_n) \quad (13)$$

The method of assigning values to terms differs from that of first-order logic in that here the evaluation process must be given a definition for each procedure name. A definition is an expression of the form '$f(x_1,\ldots,x_n) \equiv t$' where t is a term not containing variables other than $x_1,\ldots,x_n$. A set D of definitions will be called self-contained if every procedure symbol appearing on the right-hand side (the body) of a definition is itself defined in D. Note that a self-contained set of definitions allows recursive definitions. For example, under the standard way of representing sequences with pairs, we can define a procedure `append` for concatenating two sequences as follows:

$$\texttt{append}(x,y) \equiv \texttt{if}(\texttt{stringp}(x), y, \texttt{pair}(\Pi_1(x), \texttt{append}(\Pi_2(x), y))) \quad (14)$$

For the remainder of this discussion we assume a given self-contained set of definitions. A term all of whose procedure names have definitions (in the given definition set) will be called 'fully-defined'. Values are themselves taken to be terms defined by the following grammar:

$$v ::= \text{“}s\text{”} \mid \texttt{pair}(v_1, v_2) \quad (15)$$

In other words, a value is either a string constant or a pair of values. If t is a fully-defined closed term (one not containing variables), we say that t has value v (written $t \Rightarrow v$) if a certain evaluation process generates value v for t. The value relation $\Rightarrow$ should not be confused with the logical implication relation $\rightarrow$ – evaluation maps an expression to a value (e.g. a number), while implication is a logical relationship between two statements.

The value relation can be made more precise by giving specific inference rules for deriving values. For any string constant “s” we have “s” $\Rightarrow$ “s”. If $t_1 \Rightarrow v_1$ and $t_2 \Rightarrow v_2$ then `pair` $(t_1, t_2) \Rightarrow$ `pair` (v_1, v_2). Note that these two rules together imply $v \Rightarrow v$ for any value v. If $t \Rightarrow$ `pair` (v_1, v_2) then $\prod_1(t) \Rightarrow v_1$ and $\prod_2(t) \Rightarrow v_2$. If $t \Rightarrow$ “s” then `stringp` $(t) \Rightarrow$ “`true`” and if $t \Rightarrow$ `pair` (v_1, v_2) then `stringp` $(t) \Rightarrow$ “`false`”. If $t_1 \Rightarrow$ “`true`” and $t_2 \Rightarrow v$ then `if` $(t_1, t_2, t_3) \Rightarrow v$. If $t_1 \Rightarrow$ “`false`” and $t_3 \Rightarrow v$ then `if` $(t_1, t_2, t_3) \Rightarrow v$. If $t_1 \Rightarrow v$ and $t_2 \Rightarrow v$ then `equal` $(t_1, t_2) \Rightarrow$ “`true`”. If $t_1 \Rightarrow v_1$ and $t_2 \Rightarrow v_2$ with v_1 different from v_2 then `equal` $(t_1, t_2) \Rightarrow$ “`false`”. Finally, if '$f(x_1, \ldots, x_n) \equiv s$' is one of the given definitions, and v_1, v_n are values, we write $s[v_1/x_1, \ldots, v_n/x_n]$ for the result of replacing each variable x_i with the value v_i in s. If $t_1 \Rightarrow v_1, \ldots, t_n \Rightarrow v_n$, the procedure f is defined by $f(x_1, \ldots, x_n) \equiv s$ and $s\,[v_1/x_1, \ldots, v_n/x_n] \Rightarrow \omega$ then $f(t_1, \ldots, t_n) \Rightarrow \omega$. These inference principles do not always allow a value to be derived. For example, if f is defined by $f(x) \equiv$ `pair` $(x, f(x))$ then no value can be derived for f (“a”). However, if a value can be derived for a term then that value is unique.

The given system of definitions is called 'terminating' provided that a value can be derived for every fully-defined closed term. The Boyer–Moore logic enforces the constraint that the given system of definitions must be terminating – a definition is only accepted after termination has been proved. The methods for proving termination in the Boyer–Moore logic involve well-founded orders on values (Kaufmann *et al.*, 2000).

SBM can express mathematical facts as universally quantified terms. For example, given the above definition of `append`, the following universally quantified statement holds:

$$\forall x,y,z\ \texttt{equal}(\texttt{append}(x, \texttt{append}(y,z)), \texttt{append}(\texttt{append}(x,y), z)) \quad (16)$$

In general, a formula of the form $\forall x_1, \ldots, x_n\, t$ is true if, for all values $v_1, \ldots, v_n$, we have $t[v_1/x_1, \ldots, v_n/x_n] \Rightarrow$ “`true`”. It is easy to give SBM definitions of the Boolean operators `and`, `or`, `not` and `implies`. It is also easy to define natural arithmetic operations on positive integers, so that, for example, Fermat's last theorem can be formulated as follows, where `and5` is a five-argument conjunction operator:

$$\forall\, n,x,y,z\ \texttt{implies}(\texttt{and5}(\texttt{nat}(n), \texttt{nat}(x), \texttt{nat}(y), \texttt{nat}(z), \texttt{greater}(n, \texttt{two}())), \texttt{not}(\texttt{equal}(\texttt{sum}(\texttt{exp}(x,n), \texttt{exp}(y,n)), \texttt{exp}(z,n)))) \quad (17)$$

It turns out that a wide variety of statements can be formulated as universally quantified terms of SBM. The Boyer–Moore logic and its descendent ACL2 have been used to state and prove many theorems of number theory and to formulate and verify specifications for microprocessors, machine-language assemblers and compilers. All of these theorems and specifications can be easily formulated (though not easily verified) in SBM.

DEFINITIONS, COMPLETENESS, AND GÖDEL'S THEOREM

Gödel's incompleteness theorem states that for any sufficiently expressive logic there cannot exist any sound and complete system of inference. Intuitively, 'sufficiently expressive' means that the language is expressive enough to make simple statements about the integers, such as the statement that a certain polynomial does not equal zero for any integer values of its variables. The logic SBM is clearly expressive enough to make these kinds of statements. It follows that no sound and complete inference system exists for SBM. Here, 'sound' means that every provable statement is true, and 'complete' means that every true statement is provable. In practice one works with inference rules that are sound but incomplete. Although no precise characterization has been given of the statements that are provable using the inference rules provided by the Boyer–Moore system, it is clear that these inference rules are powerful enough to prove much of number theory and most specifications for hardware and software systems.

Resolution theorem proving provides a sound and complete system of inference for first-order logic. It follows immediately from Gödel's incompleteness theorem that first-order logic is not sufficiently expressive. In particular, first-order logic is not expressive enough to define addition and multiplication on the natural numbers – the natural numbers themselves cannot be defined in first-order logic. More specifically, for any set of first-order statements about the natural numbers there exist 'nonstandard' models of those statements: models different from (not isomorphic to) the natural numbers in which those statements are also true.

It is possible to introduce a notion of 'nonstandard model' for the Boyer–Moore logic and replace the notion of truth with the notion of being true in all possible (nonstandard) models. This reduces the expressive power of the logic: one can no longer define the set of natural numbers and the standard arithmetic operations on them. However, it allows a complete system of inference to be given: all statements that are true in all the (nonstandard) interpretations can be proven. On the other hand, the natural semantics seems easier to understand, and allows, in principle, the construction of ever more powerful proof systems. In practice the construction of ever more powerful proof systems is limited by the fact that all human proofs can be expressed in ZFC, so that no proof system more powerful than ZFC can be proven sound by human mathematicians.

HIGHER-ORDER LOGIC

Second-order logic allows quantification over predicates on individuals. Third-order logic allows quantification over predicates that take predicates as arguments. Higher-order logic allows n^{th}-order quantification for any n. Higher-order logic is of central importance in the theory of computer programming languages and in the study of the syntax and semantics of human language.

To formally define higher-order logic, one must first define types. The type of an expression determines its order. For example, a predicate P of one argument has type $D \to B$. This means that it takes an individual (a first-order object) as its input and returns a truth-value as its output. One writes $P{:}D \to B$ to indicate that P has type $D \to B$. One can also write $x{:}D$ to indicate that x is a first-order value, or $f{:}(D \to B) \to B$ to indicate that f takes a predicate as an argument and returns a truth-value. More generally, one writes $f{:}\tau$ to indicate that f has type τ where τ is a type expression as defined by the following grammar:

$$\tau ::= D|B|\tau_1 \to \tau_2 \tag{18}$$

Here D is the type of individuals (first-order values), B is the type of truth values, and $\tau_1 \to \tau_2$ is the type of functions that take as input values of type τ_1 and produce as output values of type τ_2. For each type expression we assume an infinite number of variables and constants of that type. We will write x^τ and c^τ to signify variables and constants of type τ. The terms of higher-order logic are defined by the following grammar:

$$t ::= x^\tau|c^\tau|\texttt{apply}(t_1, t_2)|\lambda(x^\tau, t) \tag{19}$$

The term $\texttt{apply}\ (t_1, t_2)$ represents the application of the operator t_1 to the argument t_2. The term $\lambda\ (x^\tau, t)$ represents the function that takes an input of type τ and produces as output the value of t when x^τ is assigned the input value. We write $t{:}\tau$ to indicate that the term t is a well-typed term of type τ. This well-typedness relation is defined by certain simple rules. First, we have $x^\tau{:}\tau$ and $c^\tau{:}\tau$. Next, if $t_1{:}\sigma \to \tau$ and $t_2{:}\sigma$ then $\texttt{apply}\ (t_1, t_2){:}\tau$. Finally, if $t{:}\tau$ then $\lambda\ (x^\sigma, t){:}\sigma \to \tau$. A term t is called well-typed if there exists a type τ such that these rules derive $t{:}\tau$. Note that if t is well-typed then there is only one type τ such that $t{:}\tau$.

Just as there are many different logics, so there are many different higher-order type systems. The system presented above is called the simply-typed lambda calculus with explicit types. In other systems, the same variable might be assigned different types and the single term $\lambda(x,\ x)$ can be assigned any type of the form $\tau \rightarrow \tau$. Still other systems allow type schemes and quantified types.

Several abbreviations are standard in treatments of higher-order logic. The term `apply` $(t_1,\ t_2)$ is usually abbreviated to $(t_1\ t_2)$. Types are usually left implicit on variables and constants, so that x denotes a typed variable. The term $\lambda(x^\tau,\ t)$ is often written as $\lambda x.t$. It turns out that functions of more than one argument can be represented by functions of one argument. The term $\lambda x,y.t$ is used as an abbreviation for $\lambda x.\lambda y.t$. The notation $\sigma_1 \times \sigma_2 \rightarrow \tau$ is an alternative notation for $\sigma_1 \rightarrow (\sigma_2 \rightarrow \tau)$; and the two-argument application $(f\ x\ y)$ is an abbreviation for $((f\ x)\ y)$.

In higher-order logic we assume the standard logical Boolean constants: $\wedge{:}B \times B \rightarrow B$, $\wedge{:}B \times B \rightarrow B$, $\neg{:}B \rightarrow B$, and $\rightarrow{:}B \times B \rightarrow B$ (we rely on context to distinguish Boolean implication from the type-theoretic function constructor). These constants are assigned their standard logical meaning. In higher-order logic we also assume constants representing existential and universal quantification. Thus, $\forall x^\tau\ \Phi$ (where $\Phi{:}B$) is an abbreviation for $\forall^\tau(\lambda x^\tau \Phi)$ where $\forall_\tau$ is a constant with the appropriate standard meaning of type $(\tau \rightarrow B) \rightarrow B$. Similarly, $\exists\ x^\tau\ \Phi$ is an abbreviation for $\exists_\tau\ (\lambda x^\tau \Phi)$ where $\exists_\tau$ is an appropriate logical constant.

As in first-order logic, the value of a well-typed higher-order term is defined relative to a model, where the model specifies a domain – a set of first-order values – and a meaning for each non-logical constant. Given a domain of first-order values, each type expression is associated with a well-defined set: D denotes the set of first-order values; B denotes the set {T, F}; and $\tau_1 \rightarrow \tau_2$ denotes the set of (all semantic) functions that take as input an element of the set denoted by τ_1 and produce as output an element of the set denoted by τ_2. The value for each non-logical constant c^τ must be an element of the set denoted by τ. A variable environment ρ assigns a value to each variable, in such a way that the value assigned to x^τ is in the set denoted by τ. Given a model $\mathcal{M}$, a variable environment ρ, and a well-typed term t, the value $V\ (t,\ \mathcal{M},\ \rho)$ can be defined in a manner similar to that used for first-order logic. If we have $t{:}\tau$ then $V\ (t,\ \mathcal{M},\ \rho)$ is an element of the set denoted by τ.

The formalism of higher-order logic, and its uses in the study of the syntax and semantics of human language, are described in more detail in Carpenter (1997).

HIGHER-ORDER LOGIC AND GÖDEL'S THEOREM

Although higher-order logic does not explicitly support recursive definitions, it is sufficiently expressive for Gödel's theorem to apply. In higher-order logic one can write the following induction principle for the natural numbers:

$$\forall P \quad [P(0) \wedge (\forall x\ P(x) \rightarrow P(x+1))] \rightarrow \forall x\ P(x) \tag{20}$$

Combined with a few more simple axioms, this formula yields a set of formulas (the second-order Peano axioms) whose only model (up to isomorphism) is the natural numbers. So no sound system of inference for higher-order logic can be complete. As with the Boyer–Moore logic, it is possible to give an 'unnatural' semantics, making sound and complete inference possible. However, the natural semantics seems easier to understand and in principle allows for the construction of ever more powerful inference systems.

It is possible to approximate higher-order logic using first-order schemata. For example, the induction principle for natural numbers can be expressed by the infinite set of all first-order formulas of the following form, where $\Phi\ [x]$ represents an arbitrary first-order formula with a single free variable x:

$$[\Phi[0] \wedge (\forall x\ \Phi[x] \rightarrow \Phi[x+1])] \rightarrow \forall x\ \Phi[x] \tag{21}$$

This, however, is only an approximation to the second-order statement given in formula 20: there are subsets of the integers that are nameable in richer (higher-order) languages but not nameable by a first-order formula $\Phi[x]$. Hence there are uses of the true second-order induction principle that are not covered by the first-order schema. There are natural statements of number theory that are known to be true (they are provable in ZFC) but are not provable from the first-order version of Peano's axioms. The completeness theorem for first-order logic implies that if Φ is not provable in first-order Peano arithmetic then there must be nonstandard models of these axioms where Φ is false, even if Φ is true of the actual natural numbers. In the case of Peano arithmetic it seems reasonable to work with the higher-order formulation even though it is sufficiently expressive for Gödel's theorem to apply.

SUMMARY

Logic is fundamental to our understanding of mathematics, computer programming, human language, and knowledge representation. It has been proposed as a foundation for artificial intelligence. Although no theoretical framework has led to the construction of machines matching human thought, it seems clear that logic remains relevant to the study of intelligence. The relationship between logic and statistics is currently an active area of research (Getoor *et al*., 2001).

References

Appel AW (1997) *Modern Compiler Implementation in Java*. New York, NY and Cambridge, UK: Cambridge University Press.

Bell J and Machover M (1977) *A Course in Mathematical Logic*. Amsterdam: North-Holland.

Carpenter B (1997) *Type-Logical Semantics*. Cambridge, MA: MIT Press.

Getoor L, Friedman N, Koller D and Pfeffer A (2001) Learning probabilistic relational models. In: Dzeroski S and Lavrac N (eds) *Relational Data Mining*, pp. 307–335. Berlin: Springer.

Kaufmann M, Manolios P and Moore JS (2000) *Computer-Aided Reasoning: An Approach*. Boston, MA: Kluwer.

Leitsch A (1997) *The Resolution Calculus*. Berlin: Springer.

Lifschitz V (ed.) (1998) *Formalizing Common Sense: Papers by John McCarthy*. Oxford: Intellect.

Further Reading

Barwise J (ed.) (1977) *Handbook of Mathematical Logic*. Amsterdam: North-Holland.

Boyer R and Moore J (1979) *A Computational Logic*. New York, NY: Academic Press.

Kunen K (1980) *Set Theory: An Introduction to Independence Proofs*. Amsterdam: North-Holland.

Intelligent Tutoring Systems

Introductory article

Alan Lesgold, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Intelligent tutoring systems are computer programs that use artificial intelligence techniques to interact with students and to provide experiences tailored to students' immediate needs. Originally, the term included only systems that modeled the student's developing knowledge, but today the term often includes systems that adapt to the pattern of a student's activity and to estimates of what the student knows.

INTRODUCTION

Learning involves intelligent interactions – with teachers, with fellow students, and with environments like laboratories. When researchers began to consider the possibilities of machine intelligence, they soon realized that instruction was an area in which an intelligent computer could play a useful role. Today, there are a number of intelligent computer systems, which help people learn a variety of things, from basic skills like reading and mathematics to specific technical jobs like maintaining complex equipment in a factory.

Intelligent tutoring systems vary in the details of how they work, but they all do some of the things that human teachers do. For example, one system tracks students as they solve algebraic word problems. It tries to infer from a student's performance whether he or she is on the right track towards solving a problem, and uses accumulated knowledge about the student's past performance to decide what advice to offer if not. Another system listens to young children as they try to read aloud and offers corrective feedback if they have trouble pronouncing particular words. A third system provides a simulation of a complex machine and assigns trainees tough problems, like determining what is wrong when the machine fails to work properly. If the trainees get stuck, the system offers advice based upon what actions have already been taken in trying to solve the problem. (*See* **Learning from Advice**)

These systems take advantage of several important advances in the world of computer science. They use realistic multimedia presentation, both to motivate the student and to provide realistic scenarios for learning. They promote 'learning by doing', providing feedback as students try to apply what they already know and to learn from experience. And a major part of their intelligence is based upon representations or models of the environment or task on which the student is working, what the student knows that is relevant to the problem, and what an expert would be thinking about the problem.

MULTIMEDIA PRESENTATION FOR EDUCATION

Intelligent tutoring systems try to achieve cognitive realism. That is, they present information in ways that support thought processes relevant to the desired learning and that highlight features of an environment that merit special notice by the learner. For example, an intelligent system helping teachers to improve their skills might present video clips of classroom activity. To maximize the efficiency of learning, the video might skip uninteresting moments, and it might not be high-fidelity, but it would make relevant parts of a situation clear. To facilitate attending to the important aspects of discussion, the system might simultaneously include a running transcript of classroom talk – similar to closed captioning on television. (*See* **Instructional Design**)

Sometimes sketches and diagrams can be more useful than complete realism. For example, one system built for industrial training lets trainees click on various parts of schematic diagrams to

access menus of actions they can ask a simulated machine to perform. Another system presents intelligent diagrams that can have different segments emphasized or shown in more detail, depending on where the system wants to focus student attention.

A few intelligent systems need very high fidelity to be effective. For example, a system that trains captains to steer submarines into port has full 'surround sound' capability, since expert steering sometimes responds to the direction and loudness of harbor bell buoys. Virtual-reality components, whereby scenes change in response to student movement, are also used by some intelligent training systems. For example, in one system being developed, teachers can use a mouse to move around a classroom and look for interesting activity, and then zoom in on areas of interest. Here, the system intelligence lies in how the rough sketch of the classroom from which interactions start calls attention to important information. Such systems can also adapt to the pattern of where students look.

Some intelligent instructional systems have substantial ability to understand student activity that involves movement or speech. For example, although computers are not yet perfect at understanding arbitrary speech, they are quite able to decide whether or not specific words have been spoken. One system listens to children reading aloud, decides what they are saying, and offers corrective feedback when the student misses a word. Another system, used to train special military forces, tracks the movements of trainees and modifies the behavior of simulated environments according to where they move and in which direction they look.

FORMS OF FEEDBACK

As mentioned above, the source of learning is interaction. Teachers, in particular, work by interacting with students, offering various forms of feedback. The simplest feedback is an evaluation of a performance. Just as teachers mark problems as correct or incorrect or give grades on short essays, some intelligent instructional systems offer this kind of feedback. In the simplest case, instructional computer systems may have the correct answer to every problem stored in memory, but intelligent instructional systems can do much more. By comparing the actions taken by a student in solving a problem to those modeled by an expert system, they can determine whether the student was expert-like in his or her performance. (*See* **Expert Systems**)

Of course, many tasks have multiple correct answers, so an expert system may not be able to 'score' a student's performance if the student displayed originality that the expert system did not have. There are a number of intelligent actions a computer system might take even when the range of possible answers is less constrained. First, it is possible in some cases to infer logically that an answer is correct. In this case, an intelligent system uses a combination of inference techniques and knowledge it has about the problem domain. For example, in courses where students must prove theorems, alternative proof strategies can be evaluated this way. Another possibility is to evaluate the solution, if some dimension of value exists. For example, a training system that asks people to work in a simulated environment to repair or reconfigure machinery might have a knowledge base that allows it to determine the cost of a student's proposed solution. If there is also an expert model that can solve the problem in a different way, then the cost of the student's solution can be compared with that of the expert model, and the system can give feedback accordingly (e.g. 'your solution is correct, but doing it that way would cost about $500, and there is a way to solve the problem for $75').

For written essays, a technique called latent semantic analysis can be used. Actual essays that are considered to be of high quality are used to develop a statistical profile of which words are mentioned in close proximity to which others. This profile can then be compared to that derived from a student's essay, and a score computed. Since essays can be compared to each of a number of partial responses, a system can get a reasonably good sense of where the student's weaknesses are. If each of the standard partial essays is associated with a knowledge base, that knowledge base can then be used to generate further advice to the student.

A good teacher can do more than just evaluate a student's performance. The best teachers explain why one approach is better than another and how complex systems work. They tailor explanations and advice to the particular needs of students. They adapt the complexity and extent of their advice to the student's current capabilities. We will now explore how this can be done.

MODELING AND ADAPTING TO THE MIND OF THE STUDENT

Overlay Models

In order for a computer system to be able to adapt to a student's current knowledge level, it must

somehow construct a model or representation of the student's knowledge. Such representations can take a number of different forms. The simplest approach is to list everything an expert ought to know about a domain (the expert model) and then check which items of knowledge on the list the student appears to know (the student model). Finally, the system must decide what to say to the student to help him or her learn more. (*See* **Knowledge Representation**; **Knowledge Representation, Psychology of**)

This method will work if the form in which pieces of knowledge are listed is both standardized enough to fit available computational tools and sufficient to support the needed coaching. We call this approach an 'overlay' model, since the model of the student's knowledge is some subset of the expert model.

The expert model must be stated in terms that capture everything the tutor wishes to teach, and the individual elements of the expert model must be identifiable with aspects of student performance. That is, there must be a way of determining which of the knowledge elements the student knows. One approach to this is called model tracing. The basic idea is that each performance by the student is simulated by the intelligent tutoring system, using the expert model. For this to work, the expert model needs to be expressed as a production system. A production system is a set of contingent action rules, or productions, stated in the form of 'if–then' statements (if a condition is satisfied, then carry out an action). (*See* **Learning Rules and Productions**)

Production systems can easily produce a trace of how an expert would tackle a given task. The system alternately determines which productions have a match between the current situation and their 'if' conditions and carries out the actions associated with those conditions. (*See* **Production Systems and Rule-based Inference**)

Production systems used as expert models in intelligent tutors tend to involve explicitly stated goals. Most productions in such a system have among their conditions that a particular goal be present. Many productions have among their actions the changing of a goal. For example, a production to dial a phone number might set as goals dialing the area code and then dialing the local number. The productions that dial the area code would have among their conditions that the goal be to dial the area code and among their actions to set the goal to now dial the local number.

Modeling a student's performance with such an expert system amounts to going through the student's actions step by step and asking, for each action, whether a sequence of one or more expert productions would have led to that action. If so, this constitutes prima facie evidence that the student knows those productions. If not, then this suggests that the student does not yet know the productions an expert would have used.

Once it is known which productions the student still needs to learn, it is possible to provide appropriate feedback aimed at teaching the student to apply the missing production (note that being able to state the 'if–then' relationship is not the same as knowing to apply it whenever it is relevant). Generally, such feedback is generated through a combination of text prepared in advance, sentence frames, and inferred explanations. This method is often effective, since it can take account of the specific rule the student did not apply. It can be made more effective by tailoring feedback to the specifics of the situation in which it is offered. This is straightforward when productions contain variables that are matched to specific characteristics of the current situation. The values matched by those variables can then be used to make an explanation more specific.

Here is a simple example. Suppose that a geometry expert system has a rule that has as a condition that two angles be complementary; for example: 'if angle α is complementary to angle β, then compute $\alpha = 90^\circ - \beta$'. (That is, if two angles are complementary, then the measure of one is equal to 90° minus the measure of the other.) Suppose that a student fails to apply this rule when it is appropriate. It would be more effective if the intelligent tutor could refer to the specific angles that are complementary in the specific problem the student is working on, rather than just to the rule as an abstraction. Since the specific angles must be identified anyway in order for the expert production to apply, it is quite straightforward to substitute references to the angles of the problem situation into a coaching frame that helps the student understand the rule. For example:

> You need to determine the measure of $\angle\mathbf{ABC}$. You know that $\angle\mathbf{ABC}$ and $\angle\mathbf{CBD}$ are complementary, and you know that $\angle\mathbf{CBD} = \mathbf{37}^\circ$. So, since two complementary angles sum to 90°, $\angle\mathbf{ABC}$ must equal $90^\circ - \mathbf{37}^\circ = \mathbf{53}^\circ$. (All of the substitutions for variables in the expert rule are shown in bold face.)

Bayesian Models

An alternative to symbolic inference of student knowledge using ordinary logical reasoning is to infer the knowledge of various rules or concepts

from student actions via a web of conditional probabilities. This approach allows for student modeling that captures both understanding (sometimes called declarative knowledge) and performance capability (sometimes called procedural knowledge). The basic idea is to develop a set of estimates of the conditional probability of knowing one piece of knowledge given evidence that the student knows a second piece. For example, if a student knows the definition of complementary angles, then it is likely that he or she knows the rule mentioned above.

Ideally, an intelligent tutor would possess complete knowledge of all the conditional probabilities that interrelate all the knowledge segments. This represents a lot of information (if there are n knowledge pieces, then there could be $(n-1) \times n$ conditional probabilities interconnecting them). Actually, the scheme will work as long as there are at least a few interconnections between any one piece and the rest of the knowledge base.

But there is another, potentially more severe, problem. This is the computational complexity of updating a student model of this kind as new information comes in. Once there are a number of conditional probabilities specified among knowledge elements, the exact probabilities that the student knows all the different pieces of knowledge must match the conditional probability constraints. Getting all the numbers perfect is beyond the reach of any computer if the number of knowledge pieces is large.

However, efficient schemes have been developed for approximating the student model given a network of conditional probabilities and information about which pieces of knowledge have been demonstrated. Today, there are intelligent tutoring systems (e.g. for the mechanics portion of introductory physics courses) that use Bayesian technology to improve their student modeling. These systems can make good guesses not only about procedural capabilities of students (what they can do) but also about conceptual capabilities (what they understand).

Other Schemes

There are other student modeling schemes that focus only on the context of a particular task. They do not maintain a detailed model of what the student knows, but rather maintain a model of the task given to the student and of the implications of actions the student has already taken. For example, systems have been built that track the actions of a student trying to solve a problem (medical diagnosis, or repairing a piece of equipment) in a simulated environment, and attempt to determine what an expert would do next given what the student has done already. When advice is needed, such systems tend to offer the student a choice of different kinds of advice, depending on the situation.

A related approach is to observe the actions of the student and, when they fit a known pattern, offer general advice relating to the kind of situation that usually produces that pattern. Sometimes this kind of advice is in the form of anecdotes, from which the student might infer some new knowledge or new procedural options. This approach can be very effective, especially in situations where the task is only partly specified or where many very different approaches might all be reasonable.

ENCOURAGING EXPLORATION, EXPERIMENTATION AND ACTIVE LEARNING

Intelligent tutoring systems can encourage active learning. Many researchers now believe that all personal knowledge is constructed by the student, and that drill can sometimes interfere with the student constructing rich knowledge and understanding that will be useful outside the context of tightly specified examinations and schoolwork. Even those who do not accept the more radical forms of constructivism (e.g. that nothing can be learned just by being told) still believe that active engagement of the student with a body of knowledge and its applications is essential to the achievement of significant levels of learning. (*See* **Learning, Psychology of**; **Learning Aids and Strategies**)

The earliest computer-based instruction schemes provided limited active learning possibilities, but since they worked by simple scoring of students' 'answers' rather than by assessing the patterns of their ongoing work, those possibilities were limited. By modeling the knowledge needed to solve a problem, and possibly even the knowledge apparently possessed by the student, intelligent tutors are able to go further, providing more specific advice and hence allowing students to tackle more substantial problems. Thus schooling can proceed rather like sports learning, where much of the activity consists in engaging the whole body of knowledge at once (one learns to play tennis mostly by playing tennis, not by doing tiny exercises that can be scored as right or wrong).

Active learning means more than just doing large tasks. Sometimes one has to test one's knowledge

by experimenting with the environment. Just as scientists test their theories through experiments, so, according to constructivist theory, each of us tests our new knowledge through combinations of experiments and interactions with others. We may hope that intelligent systems will, over time, become better able to support such broader knowledge gathering. Some preliminary efforts already seem productive. For example, one system, developed for limited laboratory use, helps students to access web pages with various hypotheses and data on them, store these in an 'intelligent notebook', and then annotate the connections between these different sources of information. An intelligent coach observes the patterns of annotation and makes a few broad suggestions about further research that might be useful. For example, a student might access both a collection of data relevant to a question (e.g. 'why are there mountains?') and discussions of alternative theories. If the student then draws a diagram showing the ways in which various data support different hypotheses, then the student's reasoning may be sufficiently examinable that inferences can be made about what coaching is needed.

ASSESSMENT OF TUTORING SYSTEMS

However sensible the ideas behind intelligent instructional systems may seem, it is important to test the systems to see if they really work – particularly given the high cost of producing them. Generally, those systems developed by researchers with sound understanding of how learning happens have been shown to be substantially more effective than extant forms of group instruction, but substantially less effective than the best human one-to-one tutoring (assessment based upon quizzes given before and after tutoring).

In the business world, evaluations of tutors tend to be most useful if they translate into measures of the cost of a worker's performance and the savings produced by the changes in work performance after tutoring. For example, one intelligent tutor was built for a technical job in which the biggest concern was the time it takes to repair a piece of equipment that was vital to a production line. It was possible to assign a time cost to every action that might be taken in trying to diagnose and repair this equipment. Then, the tutor could be evaluated by estimating the total time that each trainee's solution to each of several realistic problems might have taken in real life, where some actions take hours to complete. The tutor was shown to produce an average decrease in the time required to repair the equipment of about ten hours for difficult problems. The business managers knew how often such difficult problems occur, and they were able to determine that the approach was cost-effective.

For educational applications, assessment poses a special problem. Since intelligent tutors focus on more complex performances, it is likely that their main effect will be on the ability to apply complex bodies of knowledge successfully in novel situations. However, for a variety of reasons, routine measures of student educational achievement tend not to tap this high level of knowledge. Standardized tests need to take account of the many different curricula used in different schools. They also need to be fair to students coming from different backgrounds. In subtle ways, these requirements force much of standardized testing to focus on ability to use knowledge in the simplest situations (for example, complex situations are hard to standardize, so they tend to favor students with certain life experiences over those with others). Thus, the general pattern of evaluation results on intelligent tutors for use in schools has been a modest positive effect on standardized tests and a very positive effect on more demanding tasks, including some of those appearing in recently-published standards for mathematics and science.

INTELLIGENT TUTORING SYSTEMS AS PSYCHOLOGICAL RESEARCH TOOLS

Since intelligent tutoring systems carefully track the performance of students engaged in well-specified but rich problem solving, the technology behind these tutors has become useful in cognitive research. Given a hypothesis concerning how people solve certain kinds of problems or acquire certain knowledge, it is possible to specify the hypothesis as the expert system for an intelligent tutor. The tutor can then be operated in a mode where it does not provide feedback, providing a means of automatically scoring participants' performances and matching them against the hypothesized performance model. Essentially, instead of giving advice, the system tells the experimenter about when and how the student deviates from the hypothesized performance model. Such studies can be very helpful in establishing how people address complex tasks and the specific circumstances in which the tasks are addressed in a given way. (*See* **Computation, Formal Models of**; **Computer Modeling of Cognition: Levels of Analysis**)

INTELLIGENT TUTORING SYSTEMS AND OTHER FORMS OF INSTRUCTION

Intelligent tutoring systems are an important tool for education and training. They are especially effective in promoting 'learning by doing', since they permit students to tackle larger tasks while still having the support of an intelligent coach. In many respects, they follow a 'sink or swim' approach in which sinking (i.e. being overwhelmed by a large real-life task) is prevented by intelligent coaching. On the other hand, the intelligent tutors developed so far cannot effectively replace the learning that comes from critical interactions among people. We still learn much of what we know by discussing our ideas with teachers and peers in a respectful but critical way. We challenge each other's thinking and thereby make our own thinking stronger. Intelligent tools to help with this kind of learning are still at an early stage of development.

Further Reading

Anderson JR, Boyle CF and Reiser BJ (1985) Intelligent tutoring systems. *Science* **228**: 456–462.

Anderson JR, Corbett AT, Koedinger KR and Pelletier R (1995) Cognitive tutors: lessons learned. *Journal of Learning Sciences* **4**(2): 167–207.

Beck J, Stern M and Haugsjaa E (1996) Applications of AI in education. *ACM Crossroads* **3**(1): 11–15. [Available from http://www.acm.org/crossroads/xrds3-1/aied.html.]

Brown JS, Collins A and Duguid P (1989) Situated cognition and the culture of learning. *Educational Researcher* **18**(1): 32–41.

Corbett AT and Anderson JR (1992) LISP Intelligent Tutoring System: research in skill acquisition. In: Larkin JH and Chabay RW (eds) *Computer-Assisted Instruction and Intelligent Tutoring Systems: Shared Goals and Complementary Approaches*, pp. 73–109. Hillsdale, NJ: Erlbaum.

Forbus K and Feltovich P (eds) (2001) *Smart Machines in Education: The Coming Revolution in Educational Technology*. Menlo Park, CA: AAAI/MIT Press.

Gott SP and Lesgold AM (2000) Competence in the workplace: how cognitive performance models and situated instruction can accelerate skill acquisition. In: Glaser R (ed.) *Advances in Instructional Psychology*, vol. V 'Educational Design and Cognitive Science', pp. 239–327. Hillsdale, NJ: Erlbaum.

Govindaraj T, Su YLD, Vasandani V and Recker MM (1995) Training for diagnostic problem solving in complex engineered systems: modeling, simulation, intelligent tutors. In: Rouse WB (ed.) *Human/Technology Interaction in Complex Systems*, vol. VIII, pp. 1–66. Greenwich, CT: JAI Press.

Lesgold A and Nahemow M (2001) Tools to assist learning by doing: achieving and assessing efficient technology for learning. In: Klahr D and Carver S (eds) *Cognition and Instruction: Twenty-Five Years of Progress*, pp. 307–346. Mahwah, NJ: Erlbaum.

Polson MC and Richardson JJ (1988) *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: Erlbaum.

VanLehn K, Niu Z, Siler S and Gertner A (1998) Student modeling from conventional test data: a Bayesian approach without priors. In: Goettle BP, Halff HM, Redfield CL and Shute VJ (eds) *Intelligent Tutoring Systems: 4th International Conference, ITS98*, pp. 434–443. Berlin, Germany: Springer.

Wenger E (1987) *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. Los Altos, CA: Morgan Kaufmann.

Knowledge Representation

Introductory article

Stuart C Shapiro, State University of New York, University at Buffalo, USA

CONTENTS

Knowledge representation is a subarea of artificial intelligence concerned with understanding, designing, and implementing ways of representing information in computers so that programs can use this information to: derive information that is implied by it; to converse with people in natural languages; to plan future activities; and solve problems in areas that normally require human expertise.

INTRODUCTION

Knowledge representation is a subarea of Artificial Intelligence concerned with understanding, designing, and implementing ways of representing information in computers so that programs can use it:

- to derive information that is implied by it,
- to converse with people in natural languages,
- to plan future activities,
- to solve problems in areas that normally require human expertise.

Deriving information that is implied by the information already present is a form of reasoning. Because knowledge representation schemes are useless without the ability to reason with them, the field is usually known as 'knowledge representation and reasoning'. (*See* **Language of Thought**; **Artificial Intelligence, Philosophy of**; **Representation, Philosophical Issues about**; **Implicit and Explicit Representation**; **Deductive Reasoning**; **Knowledge Representation, Psychology of**; **Reasoning**)

Many philosophers consider knowledge to be justified true belief. Thus, if John believes that the world is flat, we would not say that John knows that the world is flat, because he is wrong—'the world is flat' is not true. Also, it may be that Sally believes that the first player in chess can always win, Betty believes that the second player can always win, and Mary believes that, with optimal play on both sides, chess will always end in a tie. One of them is correct, but we would still not say that any of them knows the answer, because their belief cannot have been justified by a complete analysis of the game. A computer system could not limit its information to knowledge in this strict sense, so it would be more accurate to say that the topic being discussed is belief representation rather than knowledge representation. Nevertheless, we will continue to use 'knowledge representation', because that has become accepted as the name of this subject. (*See* **Epistemology**)

REPRESENTING COMMON-SENSE KNOWLEDGE

The field of knowledge representation began, around 1958, with an investigation of how a computer might be able to represent and use the kind of common-sense knowledge we have when we decide that to get from our house to the airport, we should walk to our car and drive to the airport rather than, for example, drive to our car and then walk to the airport.

In the 1960s and 1970s, much knowledge representation research was concerned with representing and using the kind of information we get from reading and talking to other people; that is, the information that is often expressed in natural languages, and that underlies our ability to understand and use natural languages. For example, we probably understand each of the sentences in the first column of Table 1 as shown in the second column, by adding our 'background knowledge'

Table 1. Some sentences and how we understand them

Sentence	*How we understand it*
John likes ice cream.	John likes to eat ice cream.
Mary likes Asimov.	Mary likes to read books by Isaac Asimov.
Bill flicked the switch. The room was flooded with light.	Bill moved the switch to the 'on' position, which caused a light to come on, which lit up the room Bill was in.
Betty opened the blinds. The courtyard was flooded with light.	Betty adjusted the blinds so that she could see through the window they were in front of, after which she could see that the courtyard on the other side of the window was bright.

to what the sentences explicitly say. Moreover, our understanding of English includes our being able to make the following inferences. (*See* **Natural Language Processing**; **Meaning**; **Semantic Memory: Computational Models**)

Every student studies hard. Therefore every smart student studies.

On Tuesday evening, Jack either went to the movies, played bridge, or studied. On Tuesday evening, Jack played bridge. Therefore, Jack neither went to the movies nor studied on Tuesday evening. (1)

In the 1970s and 1980s, researchers became increasingly concerned with knowledge about specific domains in which human experts operate, such as medical diagnosis and the identification of chemical compounds from mass spectrometry data, and also with the other extreme – knowledge about the everyday world that everyone knows, such as the fact that when you tip over a glass of water, the water will spill on the floor. (*See* **Expert Systems**; **Expertise**)

In the 1980s and 1990s, these concerns focused on the details of specific subdomains of everyday knowledge, such as theories of time and space, and also on the general structure of our knowledge of everyday terms, leading to the construction of large and general purpose 'ontologies'. For example, the Cyc Project has devoted many staff-years to the organization of a computer-usable representation of all the knowledge that is *not* contained in encyclopedias (thus the name 'Cyc,' from 'encyclopedia') but is assumed to be already known by people who read them, and Lycos is using such an ontology to organize searches of the World Wide Web. (*See* **Spatial Representation and Reasoning**)

All these threads continue into the 2000s.

PREDICATE CALCULUS AND OTHER LOGICAL REPRESENTATION SCHEMES

In the late 1800s and early 1900s, various formal systems were developed by people who hoped to turn human reasoning into a kind of calculation. From our perspective, we can now see that what these people were engaged in was research in knowledge representation. The formal systems they developed were systems of logic, a topic which has been studied since the days of Plato and Aristotle. We may consider logic to be the study of correct reasoning. The systems of logic developed in the late 1800s and early 1900s consist of three basic components:

- syntax: the specification of a set of atomic symbols, and the grammatical rules for combining them into well-formed expressions;
- semantics: the specification of the meaning of the atomic symbols, and the rules for determining the meanings of well-formed expressions from the meanings of their parts;
- proof theory: the specification of a set of rules, called 'rules of inference', which, given an initial collection, called a 'proof', of well-formed expressions, called 'axioms', specify what other well-formed expressions can be added to the proof. (*See* **Inference using Formal Logics**)

There are two kinds of 'meaning' determined by the semantics of a system of logic. In one, we might say that the meaning of G is the claim made by the sentence, 'The moon is made of green cheese.' For this notion of meaning, the meaning of $\neg G$, as shown in Table 2, would be the same as 'It is not the case that the moon is made of green cheese' or 'The moon is not made of green cheese.' The other sense of 'meaning' is a truth value. Different systems of logic have different truth values, and

Table 2. A set of propositional connectives and their meaning

Propositional connective	*Sample use*	*Meaning*
$\neg$	$\neg P$	It is not the case that P
$\wedge$	$P \wedge Q$	P and Q
$\vee$	$P \vee Q$	P or Q, or both
$\Rightarrow$	$P \Rightarrow Q$	If P then Q

Table 3. The quantifiers and their meanings

Quantifier	*Sample use*	*Meaning*
$\forall$	$\forall x\ P\ (x)$	Every x is a P
$\exists$	$\exists x\ P\ (x)$	Some x is a P

even different numbers of truth values. There are two-valued logics, three-valued logics, four-valued logics, and even logics with an infinite number of truth values. Two-valued logics usually call their truth values 'True' and 'False'. Some logicians would say that in such a two-valued logic any sentence either means True or False. Less strictly, one might say that in such a logic, the semantics assigns a truth value of True or False to every sentence. In this notion of meaning, if some sentence P happened to be (or be assigned the truth value of) True, then $\neg P$ would be (or be assigned the truth value of) False, and if P were False, then $\neg P$ would be True. So, if G meant (by the first sense of meaning) 'The moon is made of green cheese,' then G would be (or mean, in the second sense of meaning) False, so $\neg G$ would be (or mean, or have the truth value of) True.

Although the proof theory considers only the syntax of the expressions, not their semantics, it is usually the case that if the semantics assigns a truth value of True to the axioms, all expressions that the rules of inference add to the proof will also be True. Logics that have this property are called *sound*. Soundness seems to capture the notion of correct reasoning, which is what the study of logic is all about.

Many different logics have been described and investigated. Propositional (or 'sentential') logics do not analyze information below the level of the proposition (or sentence), but use 'propositional connectives,' such as are shown in Table 2 to build more complex sentences from simpler sentences. For example, the sentence '*Students who study hard get good grades*' could not be represented in more detail than $P \Rightarrow Q$, where P represents '*Students study hard*' and Q represents '*Students get good grades*'. First-order logics (predicate logics) continue the analysis down to objects, classes, properties, and relations, with the aid of the quantifiers shown in Table 3. So in some first-order logic, '*Students who study hard get good grades*' might be represented as $\forall x\ (Student(x) \wedge study\ (x, hard) \Rightarrow get\ (x, good\ (grades))$. Some first-order logics allow functions. In one of them, this sentence might be represented as $\forall x\ (Student(x) \wedge study\ (x, hard) \Rightarrow get\ (x, good\ (grades)))$. Second-order logics allow functions, classes, properties, and relations, themselves, to be the arguments of other functions, classes, properties, and relations. In one of them, this sentence might be represented as $\forall x\ (Student\ (x) \wedge hard\ (study)\ (x) \Rightarrow get(x, good\ (grades)))$. (*See* **Representations Using Formal Logics**)

A logical sentence that always evaluates to True regardless of the meanings of its atomic parts is called *valid*. In most standard logics, the sentence $P \wedge \neg P \Rightarrow Q$ is valid, meaning that a contradiction implies anything whatsoever, but in some logics, called 'paraconsistent' logics, that sentence is not valid. In most standard logics the sentence $P \vee \neg P$ is valid, meaning that any sentence is either True or False, but in some logics, called 'intuitionistic' logics, that sentence is not valid.

Someone who uses a propositional logic to formalize some domain of interest chooses the proposition symbols to be used to represent the sentences of the domain, and their semantics – what sentence each symbol will represent. Someone who uses a predicate logic to formalize some domain of interest chooses the syntax and semantics of the individual constants that represent objects in the domain, the function symbols that represent functions in the domain, and the predicate symbols that represent classes, properties, and relations in the domain. The logic itself determines the propositional connectives and the quantifiers, and how they are to be used, along with function and predicate application, to determine the nonatomic expressions, and their meaning. The rules of inference also operate only on nonatomic expressions, and pay attention only to the logical constants. This is the sense in which people consider these logics to be 'formal' logics that pay attention only to the form, and not to the content, of the logical expressions. When selecting a logic to use, one is choosing the formal apparatus supplied by that logic.

Any knowledge representation and reasoning system consists of two parts – a knowledge representation language and a reasoning component.

If the knowledge representation language is well-defined, it will have a well-defined syntax to describe the atomic symbols and how well-defined symbol structures may be constructed from them, and a well-defined semantics to describe what the atomic symbols and the symbol structures are supposed to mean. The reasoning component is a program, often called an 'inference engine', that, given a 'knowledge base' of symbol structures, adds additional symbol structures to that knowledge base according to the rules implemented in the program. Clearly these components – syntax, semantics, inference engine, knowledge base – correspond to the components of logics – syntax, semantics, proof theory, proof. So we may view any knowledge representation and reasoning system as a logic, and ask what kind of logic it is, what formal apparatus it supplies, and whether or not it is sound. The user of a knowledge representation and reasoning system, like the user of a logic, must choose a system, and then design the representations that are not at the level of knowledge representation constructs that the system deals with. In the knowledge representation world, this person is called a 'knowledge engineer'.

PROCEDURAL REPRESENTATIONS

In the mid-1970s, knowledge representation researchers were embroiled in what was called the 'declarative/procedural controversy'. Although this controversy has largely been resolved (in a sort of compromise), it is worthwhile understanding these two approaches to knowledge representation.

Firstly, we must recognize that there are several kinds of knowing, among which are *knowing that*, *knowing who*, and *knowing how*. *Knowing that* is the kind of knowledge we have of propositions. For example, we may know that Seattle is north of San Francisco. *Knowing who* is acquaintance with a person, animal, object, etc. We may say that we know a person even though we might not know some important facts about that person, for example their birthdate. On the other hand, we may know many facts about a person without being able to say we know that person. For example, many of us know many facts about Bill Clinton, but how many of us can truly say, 'I know Bill Clinton'? *Knowing how* is knowledge of how to do things, for example how to swim or ride a bicycle.

There has not been much work in knowledge representation on *knowing who*, and everyone would agree that a procedural representation is appropriate for *knowing how*, though more on this later. The declarative/procedural controversy was about how to represent *knowing that*. The declarativists were in favor of representing propositions that are known (believed) by some agent as a symbol structure with declarative semantics, for example a well-formed expression of propositional or predicate logic, stored in the agent's knowledge base. The proceduralists were in favor of representing such propositions as small programs. For example, when the early (simulated) robot SHRDLU was told, 'I own blocks which are not red, but I don't own anything which supports a pyramid', it represented that information as two small procedures in the PLANNER programming language. When, later, SHRDLU was asked 'Do I own anything in the box?', it ran those two procedures to determine the answer. The problem with maintaining this distinction between declarative and procedural representations of *knowing that* is the well-known equivalence of data and program. A procedure can be written in a declarative programming language such as Lisp or Prolog, and can thereby be viewed as a symbol structure with declarative semantics. On the other hand, when a declarative representation is used by the inference engine to draw some inference, the declarative representation may be viewed as a program in the programming language interpreted by the inference engine. In this sense, whether a representation is declarative or procedural depends on how one views it. (*See* **SHRDLU**)

We might resurrect the declarative/procedural distinction by considering our own knowledge of how to do things. Many of us know how to ride a bicycle. However, few of us can describe how to ride a bicycle, for example in order to instruct someone else. We might consider this knowledge procedural *only*. In this view, all knowledge may be viewed as procedural knowledge, but only knowledge that can be expressed in a declarative language by the knower may be viewed as declarative knowledge. As another example, the restaurant script (see below) is a representation of what typically happens in a restaurant. There have been programs that, supplied with the restaurant script, could fill in details about what happened in restaurant stories. For example, given the story, 'John went to a restaurant and ordered a steak', such a program could infer that John was seated and given a menu between entering and ordering. However, most of these programs could not answer questions about the restaurant script itself, such as 'What typically happens in a restaurant after the patron is seated?' It is, therefore,

reasonable to say that for these programs the restaurant script is not represented declaratively, but only procedurally. (*See* **Story Understanding**)

PRODUCTION SYSTEMS

Production systems are a subclass of knowledge representation and reasoning systems. The knowledge base of a production system is divided into two parts, a working memory and a rule memory. The working memory consists of a set of symbol structures not containing variables. The rule memory consists of a set of pattern-action rules. Each pattern-action rule has a 'left-hand side', which is a set of patterns, and a 'right-hand side', which is a set, or sequence, of actions. The patterns and actions may contain variables as long as every variable in the right-hand side of a rule is also in the left-hand side of the same rule. If every pattern in a rule matches some symbol structure in working memory, with a consistent substitution of constants for the variables in the patterns, then the rule is said to be 'triggered'. A triggered rule may 'fire', in which case every action in the right-hand side is performed after replacing the variables with the constants they were matched to in the left-hand side. The actions allowed in the right-hand sides of rules vary among different production systems, but they generally include adding structures to working memory and removing structures from working memory. They may also include adding and removing rules, and interacting with a user. (*See* **Production Systems and Rule-based Inference**; **Working Memory, Computational Models of**; **Rule-based Thought**)

Since the firing of a rule may change the structures in working memory, it may also change which other rules are triggered. So if several rules are triggered at the same time, the ultimate behavior of the production system may depend on which triggered rule fires first. This is determined by a conflict resolution strategy. Typical strategies are: don't fire any rule more than once on the same variable substitutions; fire the rule that was triggered most (or least) recently; fire a more specific rule before a less specific rule.

Production systems were first designed to be a model of the human mind. For this purpose, the size of working memory and the allowed actions were restricted. However, they have been a popular architecture for expert systems, for which purpose those restrictions were lifted.

It should be noted that the symbol structures in working memory and the patterns and actions in rule memory must be formulated in some knowledge representation language. What the production system architecture provides is a particular style of reasoning and acting using that language.

SEMANTIC NETWORKS

Semantic networks are a variety of labeled, directed acyclic graph in which the nodes denote entities and labeled directed arcs denote relations between the nodes they connect. Two kinds of semantic networks have been developed, inheritance networks and propositional semantic networks. Figure 1 illustrates an inheritance network which is intended to represent the following information:

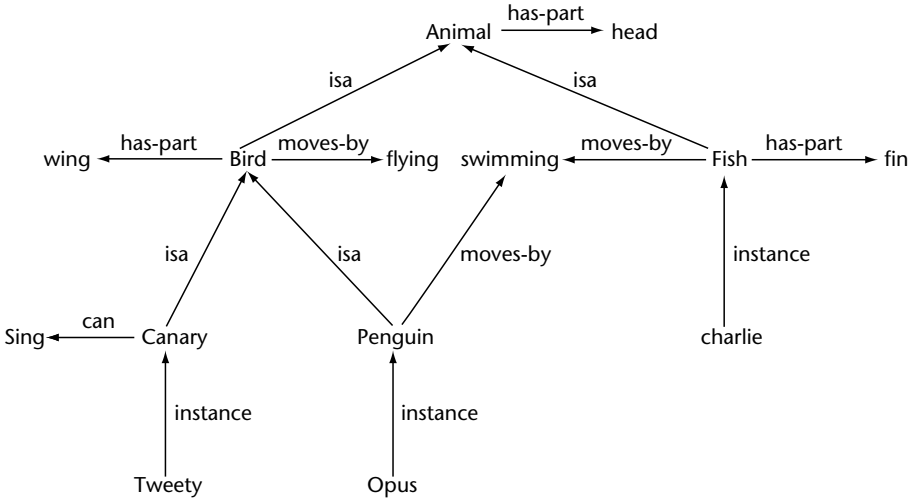


Figure 1. An inheritance-style semantic network.

Birds and fish are animals. Canaries and penguins are birds. Animals have heads. Birds have wings. Fish have fins. Birds move by flying. Fish and penguins move swimming. Canaries can sing. Tweety is a canary. Opus is a penguin. Charlie is a fish. (2)

Notice that some of the nodes represent categories (Animal, Bird, Fish, Canary, and Penguin), and others represent individuals (Tweety, Opus, and Charlie). The relation between an individual and the categories it is a member of (instance) is different from the relation between a category and its supercategories (isa). (*See* **Semantic Networks**)

Early presenters of inheritance networks did not make the semantics of the relations very clear. For example, it is not clear in Figure 1 what it means when arcs with the same label emanate from a category and one of its supercategories. Surely, birds have both wings and heads, but penguins swim and do not fly. Moreover, although the 'has-part' relation seems simple, the 'has-part' relation from 'Animal' to 'head' must mean *Every instance of Animal has a part which is an instance of head*. These semantic confusions were clarified in the successors to simple inheritance networks, the most prominent of which are called *description logics*.

Description logics are a variety of inheritance networks in which categories (called 'concepts') and relations (called 'roles') between them can be defined without the semantic confusions of earlier inheritance networks. For example, using a formalism called $\mathcal{KL}$, which was designed to reflect many of the features common to different description logics, a parent can be defined as a person with at least one child who is also a person, as follows.

```
(cdef PARENT (and PERSON (c-some
  Child PERSON)))                    (3)
```

Here, PARENT is the concept being defined, PERSON is a concept which, presumably, has already been defined, and Child is a role. This is also an example of a concept defined with necessary and sufficient conditions. That is, if Ken is said to be a PARENT, it is *necessary* that Ken be a PERSON with at least one Child who is a PERSON. So the description logic system can infer that Ken is a person, has at least one child, and that child is a person. On the other hand this same definition says that if Judi is a PERSON with at least one Child who is a PERSON, that is *sufficient* information to conclude that Judi is a Parent. Natural kinds, such as birds, fish, and animals, cannot be given necessary and sufficient conditions, so *primitive concepts* can be defined with only necessary conditions. The $\mathcal{KL}$ definitions of ANIMAL and FISH from Figure 1 are:

```
(cprim ANIMAL (and top
  (c-some Part HEAD)
  (c-atmost 1 Part HEAD)))
(cprim FISH (and ANIMAL
  (c-some Part FIN)
  (c-some Moves-by SWIMMING)))       (4)
```

This says that every FISH has one head, by inheritance from ANIMAL, and, in addition, has one or more FINs. Since description logic roles accumulate in this way, the only way to say that birds fly, but penguins are birds that swim instead, is to separate flying birds from swimming birds:

```
(cprim BIRD (and ANIMAL
  (c-atleast 2 Part WING)
  (c-atmost 2 Part WING)))
(cdef FLYING-BIRD (and BIRD
  (c-some Moves-by FLYING)))
(cprim PENGUIN (and BIRD
  (c-some Moves-by SWIMMING)))
(cprim CANARY (and FLYING-BIRD
  (c-some Can SING)))                (5)
```

(*See* **Concepts, Philosophical Issues about**; **Conceptual Representations in Psychology**; **Natural Kinds and Artifacts**)

All these $\mathcal{KL}$ constructs define concepts, and are considered part of the description logic *terminological component*. To actually make assertions about individuals, most description logics also have an *assertional component*. Assertions in the assertional component are usually written in a syntax that looks like normal first-order predicate logic in which defined concepts can be used as unary predicates and defined relations can be used as binary relations. For example, we might have:

CANARY(Tweety) WING(Tweety-left-wing)
PENGUIN(Opus)
SWIMMING(Opus-swimming-style) (6)

Part(Tweety, Tweety-left-wing)
Moves-by(Opus, Opus-swimming-style) (7)

Besides the confused semantics of their relations, another deficiency of inheritance networks is that since information can only be represented about nodes, one cannot represent information about relations, such as that the 'isa' relation is transitive. Nor can one represent information about beliefs, such as that the encyclopedia is the source of the belief that canaries are birds. Description logics

do represent information about relations, but they do not represent information about beliefs. This deficiency is solved by propositional semantic networks, in which nodes are used to represent beliefs (propositions) as well as the individuals, categories, and properties represented by nodes in inheritence networks. Figure 2 illustrates a propositional semantic net in which 'M1!' represents the proposition that canaries are birds, 'M2!' represents the proposition that 'isa' is a transitive relation, and 'M3!' represents the proposition that the source of 'M1!' is the encyclopedia.

SCHEMAS, FRAMES, AND SCRIPTS

Some researchers felt that semantic networks used a representation that was too fine-grained and too passive. Instead, they argued for representational structures that contain more information about the entities being represented, and also incorporate active processes. They adapted the notion of schemas (or 'schemata') from psychological literature. The two most widely used schema representation systems are frames and scripts. (*See* **Schemas in Psychology**)

Frames were originally proposed as a representation of structured visual information about complex objects. For example, if you open a door to an office, you expect to see certain things, such as a desk, chairs, etc. You would be surprised to see a tennis court, a beach, and a swimming pool in the office. On the other hand, if you opened a door to a bedroom, you would expect to see a bed, a chest of drawers, etc. The proposal was that the 'office frame' would contain pointers to the representations of objects you would expect to be in an office, the 'bedroom frame' would contain pointers to the representation of objects you would expect to be in a bedroom, etc. As frame representation systems were implemented, they became more similar to semantic networks, but with the labelled arcs, now called 'slots', pushed into the nodes, now called 'frames', and the nodes pointed to by the arcs, now called 'slot fillers'. For example, Figure 3 shows the information of Figure 1 as a frame system.

One feature frame systems tend to have that semantic networks do not is *procedural attachment*. Instead of a slot being filled by a pointer to a frame or a set of such pointers, the slot could be filled by an *if-needed* or an *if-added* procedure. If a slot containing an if-needed procedure is accessed, the procedure is executed and is expected to compute and return the slot filler. If a slot containing an if-added procedure is filled, the procedure is executed and is expected to fill other slots that depend on the new information being added to this slot. If-needed and if-added procedures are procedural versions of inference by backward chaining and forward chaining, respectively.

Scripts, like frame systems, were designed to be structured representations, but of activities rather than objects. For example, the restaurant script contains a representation of all the activities that typically occur when one visits a restaurant. If you read a story about someone going to a restaurant and ordering a steak, you fill in the information about being seated and being given a menu from your restaurant script. (*See* **Natural Language Processing: Models of Roger Schank and his Students**)

PICTORIAL REPRESENTATIONS

Some people think mostly linguistically; others think mostly in pictures. Everyone can probably do both, even though they usually do one or the other. Try this: Think of an elephant. Which way is it facing? If, when you thought of an elephant, you pictured one in your mind, you should have a definite answer to that question.

Just as people can represent entities in their minds either linguistically or pictorially, we can

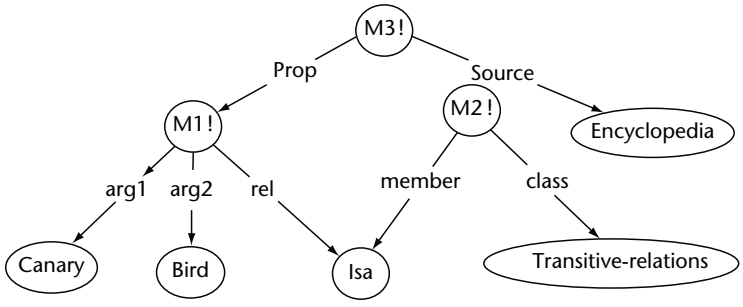


Figure 2. A propositional semantic network.

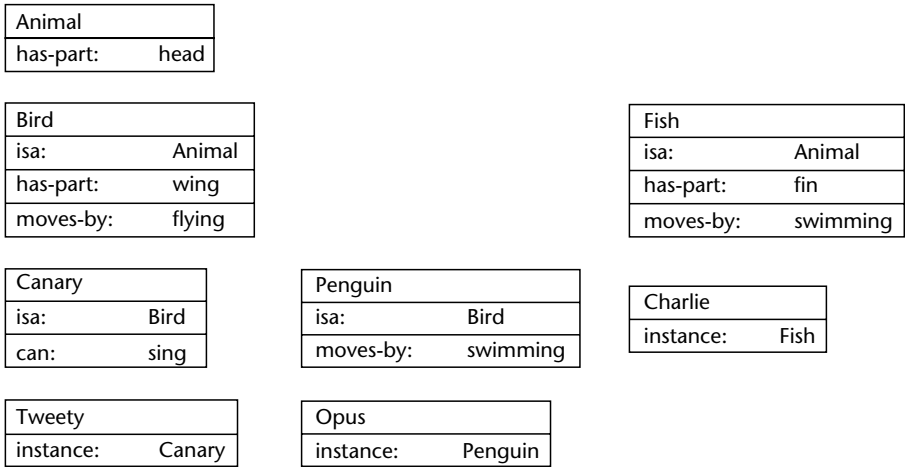


Figure 3. Figure 1 as a frame system.

use linguistic and pictorial representations in other media, including computers. The distinction is also often termed digital vs. analog, as in digital clocks vs. analog clocks. (*See* **Mental Imagery, Philosophical Issues about**)

The best way to distinguish analog from digital representations is to compare the domain of the representation (syntax) to the domain of what is represented (semantics). An analog representation has a syntactic operation that is a direct analogue of a semantic representation. Consider clocks. What is represented is time. On an analog clock, the representation is the rotation of the clock hands around the circle of the clock face. The difference between the representation of 10.15 a.m. and that of 10.30 a.m. is a 90 degree rotation of the minute hand, which is one quarter of the complete 360 degree rotation. The complete 360 degree rotation represents one hour, and one quarter of a rotation represents one quarter of an hour. On a digital clock, however, the times 10.15 a.m. and 10.30 a.m. are represented with different numerals. Nothing about the difference between the two sets of numerals indicates what the difference in the represented times is, unless one moves to the separate semantic domain of numbers, and subtracts 15 from 30.

Analogue representations can be constructed in computers by using a data structure whose operations are analogues of the relations being represented. For example, consider a predicate logic representation of items arranged in a row: *Left (desk, chair), Left(sofa, desk), Left (chair, stool)*, where *Left (x, y)* means that *x* is to the left of *y*. To decide the left-to-right arrangement of the stool and the sofa requires a certain amount of search and inference. However if, instead, the Left relation were represented by order in a list, the four relations would be captured by the list *(sofa, desk, chair, stool)*, and the left-to-right arrangement of the stool and the sofa could be decided by a linear search. Some researchers have created systems that can reason about diagrams or visual scenes by representing them in two-dimensional data structures where it is easy to rotate or otherwise manipulate the figures. (*See* **Analogical Reasoning, Psychology of**)

CONNECTIONIST REPRESENTATIONS: LOCAL AND DISTRIBUTED

Connectionist representations are designed to model the brain by using a large collection of intercommunicating units, each of which is a model of a neuron. These units are organized in layers: an input layer, an output layer, and one or more 'hidden' layers. Each unit maintains an activation level and connections to other units, which may be on the same layer (in some versions) or on layers closer to the output layer. Each connection is also given some weight, which might be negative or positive. The network as a whole makes some decision or characterizes its input. Input is achieved by adjusting the activation level of the units in the input layer. When the activation of a unit exceeds some threshold (which may be different for different units), an activation is passed to all outgoing connections, where it is adjusted by the connection weights, and passed to the connected units, etc. The final decision or characterization is read off the units in the output layer. Networks are trained by adjusting the weights on the connections by one of several possible feedback mechanisms. (*See* **Connectionism**; A00068; A00163)

Local connectionist representations are distinguished by the requirement that each decision or characterization is represented by a single unit, and each input unit also represents some concept of the input. For example, in a lexical decision task, each input unit might represent a particular letter in a particular position, and each output unit might represent a particular word.

In a distributed connectionist representation, each represented decision or characterization is represented, not by a single unit, but by a pattern of unit activations. Distributed representations have been found to generalize what they have learned better than local representations do.

Connectionist representations are considered subsymbolic rather than symbolic representations. As such, they are not as capable of representing and reasoning about beliefs as the other representation techniques discussed in this article. (*See* **Symbolic versus Subsymbolic**; **Bayesian Belief Networks**; **Language, Connectionist and Symbolic Representations of**)

MANAGING CHANGE

Consider again some of the information in Figures 1 and 3, namely that birds fly, but penguins do not. If you learn that Opus is a bird, you are justified in concluding that Opus can fly. However, if you then learn that Opus is a penguin, you must reject your conclusion that Opus can fly. This is an example of an interesting phenomenon where a new piece of information causes the rejection of a previous conclusion. It is sometimes said that, in this case, the new piece of information *defeats* the old conclusion. This phenomenon often occurs in the presence of general information to which there are exceptions. The general information is sometimes referred to as *default* knowledge, and conclusions drawn from the general information are sometimes said to be *defeasible*. From the point of view of classical propositional and predicate logic, this situation is most unusual, since these logics are *monotonic*, meaning that if a conclusion can be drawn from some set of beliefs, it can also be drawn from any superset of those beliefs. (Just ignore the extra beliefs.) Attempts to formalize defeasible reasoning have been made by knowledge representation researchers, and this remains an active area of research. (*See* **Non-monotonic Logic**)

Removing default conclusions that have been defeated by more specific information is just one possible reason that information might have to be removed from a knowledge base. If the knowledge base is a model of the world, or a model of some agent's beliefs about the world, it may be that the world changes because of the action of the agent or some other agent. If the knowledge base is a model of some agent, or a collection of beliefs input by some agent or agents, it may be that the agent or agents have changed their beliefs. If the knowledge base is a collection of facts and 'laws' of some developing theory, it might be found that some of the facts and laws are contradictory, and the theory must be revised. Removing information from a knowledge base seldom involves merely removing a single proposition (fact, rule, law). If additional propositions have been derived from the one to be removed, they might need to be found and removed also. If the proposition to be removed was derived from other propositions in the knowledge base, or could be rederived from them, they must be found, and at least one of them must be removed or else the removed proposition could be reintroduced. The first systems that knowledge representation researchers implemented to handle these complications of removing information from knowledge bases were called 'truth maintenance systems'. More formal studies, carried out by computer scientists, logicians, and philosophers go under the name 'belief revision'.

Using belief revision or truth maintenance to deal with a changing world is appropriate if the knowledge base always represents the *current* time, and should be changed as time moves. However, this eliminates the possibility of representing what was the case at past times. To do that, time must be represented explicitly, and propositions that hold only for some specific time must indicate so explicitly. To do this, specialized logics including temporal logics and modal logics have been used. Another logic for this purpose, popular among knowledge representation researchers, is situation calculus, in which predicates that can change are given an extra argument that ranges over situations. For example, if a particular book is on a particular table in a particular situation, this might be represented as *On(book 1, table3, S5)*. In situation calculus, an action is considered to be a function from the situation before the action is performed to the situation afterward. (Some versions of situation calculus use slight variations of this.) For example, the action *pickup*(*book1, S5*) might represent the situation that exists after picking up *book1* in situation *S*5. We would then have ¬*On* (*book1, table3, pickup*(*book1, S5*)). Stating the effects of actions is fairly straightforward. However, stating what is *not* changed by an action is more involved. For example, if it is the case that *In*(*table3, room2, S5*), is it the case that *In*(*table3, room2, pickup*

(*book1*, *S5*))? The problem of specifying what is not changed by an action has been called the 'frame problem', not to be confused with the frames used as schema representations. (*See* **Frame Problem, The**)

Further Reading

Addanki S (1992) Connectionism. In: Shapiro SC (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 268–274. New York, NY: John Wiley.

Bobrow DG and Collins A (eds) (1975) *Representation and Understanding: Studies in Cognitive Science*. New York, NY: Academic Press.

Brachman RJ and Levesque HJ (eds) (1985) *Readings in Knowledge Representation*. San Mateo, CA: Morgan Kaufmann.

Cercone N and McCalla G (eds) (1987) *The Knowledge Frontier: Essays in the Representation of Knowledge*. New York, NY: Springer-Verlag.

Davis E (1990) *Representations of Commonsense Knowledge*. San Mateo, CA: Morgan Kaufmann.

Hobbs JR and Moore RC (eds) (1985) *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex.

Gärdenfors P (ed.) (1992) *Belief Revision*. Cambridge, UK: Cambridge University Press.

Iwańska ŁM and Shapiro SC (eds) (2000) *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. Menlo Park, CA: AAAI Press/MIT Press.

Kramer B and Mylopoulos J (1992) Knowledge representation. In: Shapiro SC (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 743–759. New York, NY: John Wiley.

Lehmann F (ed.) (1992) *Semantic Networks in Artificial Intelligence*. Oxford, UK: Pergamon Press.

Levesque HJ and Lakemeyer G (2000) *The Logic of Knowledge Bases*. Cambridge, MA: MIT Press.

Lifschitz V (ed.) (1990) *Formalizing Common Sense: Papers by John McCarthy*. Norwood, NJ: Ablex.

Reichgelt H (1991) *Knowledge Representation: An AI Perspective*. Norwood, NJ: Ablex.

Rumelhart DE and McClelland JL (eds) (1986) *Parallel Distributed Processing* 2 vols. Cambridge, MA: MIT Press.

Sowa JF (ed.) (1991) *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Los Altos, CA: Morgan Kaufmann.

Sowa JF (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Language Learning, Computational Models of

Intermediate article

Michael Gasser, Indiana University, Bloomington, Indiana, USA

CONTENTS

Computational models of language acquisition range from symbolic approaches based on linguistic theory and relying on built-in (innate) constraints to statistical and connectionist approaches relying only on general-purpose learning mechanisms.

INTRODUCTION

Probably no single issue had more influence on the development of cognitive science in the second half of the twentieth century than the question of how natural languages are learned. As in other areas of cognitive science, much progress has been made through attempts to build formal models of the processes and to implement these as computer programs. Language acquisition is also of practical interest within computational linguistics: even fragments of natural languages are complex enough to make hand-coding of grammars and lexicons cumbersome or impossible. Thus systems that learn, at least in part, may be the only usable language processing systems. This article examines formal models of language acquisition, focusing on models that have actually been implemented. (*See* **Language Acquisition**)

Computational models of language acquisition vary in at least the following ways: the nature of the task (what theory of language and language behavior is assumed, what the learner is learning to do, and what characterizes the input to learning); and the nature of the model (what is built in ('innate'), and what sort of learning architecture and algorithm are used).

THE TASK

Theories of Language and Language Behavior

In most linguistic theories, language, and each level of linguistic inquiry, is treated as a module, something that can be studied without consideration of other domains. This view corresponds to learning models in which the modeled behavior is purely linguistic. In cognitive linguistics, on the other hand, as well as in some work on natural language within artificial intelligence, it is felt that language can be understood only in terms of its place in cognition. The corresponding goal in models of processing and learning is the grounding of linguistic behavior in psychological processes or robotics. (*See* **Cognitive Linguistics**)

More fundamentally, theories of language differ in what they take language to be, what it means to know a language. Everyone accepts the general view that a language is a set of patterns, associating features of linguistic form (phonetic, phonological, syntactic) with each other and with features of linguistic function (semantic and pragmatic). The simplest view, then, would be that language is just associations, of varying complexity and generality, and that it is these associations that are learned. This idea lends itself to particular kinds of learning algorithms, especially connectionist and memory-based algorithms, which are discussed below.

The idea of knowledge as associations and learning as rote memorization or statistical correlation makes sense to most researchers when applied to certain aspects of language: the relationship between acoustic or articulatory parameters and phonetic categories, the arbitrary forms that morphemes take in a given language, morphological irregularities, and the details of lexical semantics. More controversial are aspects of language that are thought to be systematic, those aspects that make up grammar: the mapping between underlying and surface phonological forms, the formation of regular morphological inflections, the assignment of words to syntactic categories, the grouping of words into syntactic constituents, the mapping between syntax and sentence-level semantics, grammatical relationships between constituents within

sentences, and relationships between sentences. Nearly all accounts of these aspects of language treat them as rule-governed. In general, these accounts assume structured representations, built up from atomic symbols, including both constants and variables. (*See* **Symbol Systems**)

Knowledge in these symbolic theories of language often takes the form of formal grammars; thus, learning is often grammatical induction, in which the learner attempts to arrive at a grammar characterizing the target language on the basis of examples provided by the environment. An important concern in computational linguistics has been which type of grammar, or equivalently which type of automaton, in the Chomsky hierarchy is adequate for natural language. Interest has centered on context-free grammars and, to a lesser extent, regular (finite-state) grammars, which may be adequate for morphology and phonology. Some symbolic models also make use of transformations that can move elements within the tree structures generated by context-free grammars. There has been considerable interest in what would be necessary in order for a learning model to learn these different classes of grammars. (*See* **Syntax**)

To capture the commonalities among languages and deal with some of the learnability issues discussed below, some modern theories within generative linguistics express the possible range of variation within natural language grammars in terms of a finite set of parameters. In learning models that rely on these theories, it is the parameter settings, rather than explicit grammatical rules, that are learned (Roeper and Williams, 1987). One symbolic theory, optimality theory, posits a set of universal constraints which apply to abstract, underlying forms to yield surface forms. Languages differ in the priority that these constraints have, and it is this priority that an optimality-theoretic learner learns (Tesar and Smolensky, 1998). (*See* **Optimality Theory**; **Learnability Theory**)

Thus, to some researchers, learning a language means learning patterns of associations. To others, learning a language means learning both rote associations and the regularities that make up grammar (explicit rules, parameter settings, or constraint rankings). As described below, there are correspondingly two sorts of views on what the learner needs to know beforehand and what sort of learning algorithms are needed: the 'symbolic' and the 'subsymbolic' (or 'connectionist') views. Language grounding models are usually of the subsymbolic type, because the extralinguistic domains they take into consideration are often best handled by such models. To the extent that symbolic and subsymbolic models have been applied to similar phenomena, they have often been seen as competitors. (*See* **Symbolic versus Subsymbolic**; **Language, Connectionist and Symbolic Representations of**)

What the Learner Learns to Do

Language behavior includes both analysis (parsing, recognition, comprehension) and production. A 'performance-oriented' model is designed to learn one or other or both of these behaviors directly. For example, given an input form produced by a speaker, an analysis model attempts to assign a category, a structure, or a meaning representation to it. A 'competence-oriented' model is designed to learn the knowledge (usually some form of grammar) which, together with analysis and production algorithms, permits language behavior. For these models, the overt task of the model may be to determine whether an input is grammatical or not. (*See* **Language Comprehension**; **Performance and Competence**)

For grammatical induction, there is also the question of what would constitute success. The process would have to converge on the adult grammar. Gold (1967) formalized this notion: if, in the learning of a target language, the learner's hypothesized grammar never changes beyond some point, and that grammar correctly generates the target language, then the learner is said to have 'identified the language in the limit'.

Obviously no model, whether of the competence or the performance type, can be expected to learn phenomena across the full range of linguistic enquiry. Specific areas that have been addressed include word recognition, segmentation, phonemes, stress, regular and irregular inflectional morphology, word meaning, syntactic categories, formal grammars, grammatical parameter settings, verb subcategorization frames, and long-distance dependencies. Several ambitious projects have addressed both syntactic and semantic phenomena. An early symbolic example is Anderson's (1977) 'language acquisition system', which learns a procedural grammar in the form of an augmented transition network. A subsymbolic example is Miikkulainen's (1993) model, which learns distributed syntactic and semantic representations for lexical items.

The Input to Learning

In theories of human language acquisition, the character of the input that the learner receives is a

crucial issue, one that computational models must address. One way in which tasks vary with respect to the available input is in the degree of supervision that is available. Does the environment provide the learner with feedback concerning the correctness of the learner's response? (*See* **Bayesian and Computational Learning Theory**)

Purely phonetic or phonological acquisition is usually treated as unsupervised, as it is for human learners. Inputs are not labeled with their phonological categories, and word boundaries are not provided. Word recognition, on the other hand, is usually treated as a supervised task. Given a form, the learner responds with a word, and the environment provides feedback on whether the response is correct. The acquisition of word meaning may be treated as supervised or unsupervised, depending on what 'meaning' is taken to be. If the meaning of a noun is taken to be its referent noun, for example, a model may provide the meaning along with the word being learned. If word meaning is seen as a property to be extracted on the basis of the interrelationships among words, on the other hand, then the learning may be unsupervised. (*See* **Word Recognition**; **Lexical Semantics**)

For grammatical induction, both approaches have been considered. The unsupervised approach simply presents grammatical forms. The supervised approach presents both grammatical and ungrammatical forms with their grammaticality labeled; or, in the case of optimality theory, pairs of underlying forms and corresponding surface forms. When ungrammatical forms are labeled as such, the learner has access to a form of negative evidence. As noted below, negative evidence can greatly constrain the space of possible grammars hypothesized by the learner. On the other hand, it has been argued that children do not normally have access to negative evidence, and that when they do, they ignore it. For this reason, most grammatical induction models make use of positive evidence only.

In connectionist models, grammatical learning is sometimes a side effect of exposure of a network to an unsupervised task. One such task is prediction. A network is presented with one element in a sequence (for example, a phonetic segment or a word) and trained to predict the next element in the sequence. Networks trained on prediction can learn about the structure of the input as a side effect of the overt prediction task (Elman, 1991).

Models also vary in the extent to which they make use of real data. While most models still use artificial training sets, statistical models, especially those acquiring phonetic categories, segmentation strategies, syntactic categories, probabilistic grammars, and lexical semantics, often learn on the basis of real data.

TYPES OF MODELS

Innateness

The negative evidence problem

Probably the most contentious question within the study of language acquisition concerns what knowledge or mechanisms are assumed to be in place before learning takes place – that is, to be innate. Early work on language identification in the limit by Gold (1967) has been very influential in this regard. Gold proved that, even with a very powerful learning algorithm which tries each of the possible grammars until it hits on the right one, negative evidence is required to learn the sorts of grammars that characterize natural language.

Consider the acquisition of the passive in English. The learner hears several examples: *the dog got caught*, *he'll get hurt*, *it got eaten*. The hypothesis might be that a verb phrase can consist of a form of the verb *get* followed by the past participle form of another verb. But in the absence of negative evidence, how would the learner know what restrictions there are on the possible verbs that can take this pattern? For example, would *they got died* and *I get seemed* be grammatical?

Similar arguments, going back at least to the influential work of Quine (1960), have been made regarding the learning of word meaning. The problem is that too many meanings are apparently compatible with words when they are presented to children. Without being told explicitly what a word does not mean, it again appears impossible to settle on the correct meaning.

Within the field of computational language acquisition, there have been two sorts of responses to the apparent insufficiency of the input. One is to treat these idealizations of the learning task as missing essential information, specifically semantics and pragmatics in the case of the learning of grammar. This is the view taken by Anderson (1977) in his model mentioned above. Another is to constrain the learner to particular sorts of hypotheses on the basis of innate mechanisms or knowledge. This innate help may be specific to language, or it may be applicable to cognitive processes more generally. (*See* **Innateness and Universal Grammar**)

Innate constraints specific to language

One well-known proposal for constraining the set of possible hypothesized grammars is the 'subset principle' (Berwick, 1985). By this principle, when there are two grammars that accommodate the input examples, and one is a proper subset of the other, the smaller grammar is to be preferred. For example, if the learner, for other reasons, had already divided verbs into transitive and intransitive verbs and had then been presented with the passive sentences given above, the hypothesis could be that only transitive verbs could take part in the passive rule. This would obviate the need for negative examples such as **they got died*.

Most other proposals for innate linguistic knowledge constrain the set of possible languages more directly, limiting them to what is compatible with 'universal grammar', the purported system of properties that all natural languages share. Universal grammar is supposed to be built into the language learner, overcoming the limitations found by Gold because the size of the search space is much smaller. The best-known current proposals of this type are the parameter-setting models mentioned above. The learner starts with an innate finite set of parameters, each with a finite set of possible values, normally including a default initial value for each.

Innate general cognitive constraints

Other models are based on the hypothesis that language acquisition is possible because of more general constraints: cognitive mechanisms, regularities in the world, and properties of the input. Most such models are subsymbolic.

Regier (1996) provides an example from the domain of word meaning. In his connectionist model of the acquisition of spatial terms such as *in* and *through*, most of the constraints come from a built-in visual processing module. This module extracts information such as the relative positions of the centers of mass of the objects in the input. This processing is compatible with what is known about the human visual system, and this constrains what can be learned by the linguistic part of the system.

Some connectionist models are best described in terms of what is *not* innate in them. In the acquisition of word meaning, it has been proposed that the learner is guided by possibly innate biases: for example, the 'shape' bias whereby nouns referring to solid objects are generalized on the basis of shape. There has been much connectionist research designed to show that such biases emerge from general learning mechanisms in combination with the properties of the input itself. (*See* **Word Learning**)

Learning Architectures and Algorithms

Symbolic models

Symbolic models are designed to learn grammar, whether in the form of phonological rules, syntactic rules, parameter settings, or constraint rankings. These models work by maintaining an explicit candidate grammar and modifying it when it fails to handle an input form.

The model must have a way of determining whether the grammar fails to handle an input. Many models have a parser as a separate module for this purpose. The parser may also provide structural information that is relevant to the change that would take place when the grammar fails. For example, if the model is learning how to set the 'pro-drop' parameter, which reflects whether the language permits zero subjects, it must be able to identify the subject noun phrase and the verb in sentences. (*See* **Learnability Theory**)

Models vary in how they respond to input forms. There are two major issues. Firstly, which aspect of the grammar is changed when the grammar fails to handle the input? Some parameter-setting models select a parameter randomly, change its current value to some other value, attempt to parse the form again, and if they succeed, change the parameter setting in the evolving grammar. Others, such as Dresher and Kaye's (1990) model, select a parameter according to the type of input that is presented: thus, there is further innate knowledge specifying which sorts of forms convey information about which parameters.

Secondly, how much information is taken into account when a parameter is changed? In the simplest treatment, the learner takes only a single input into account each time (e.g. Gibson and Wexler, 1994). The danger in this approach is that the learner may jump to a conclusion and make an inappropriate change to the grammar. This is possible when there is more than one parameter change that can make the form parsable, when the form is an exception to the rules, or when the form contains an error. One solution is to give the learning model some form of memory: then, rather than responding to a single item, it responds to an accumulated set of items. In the most complex treatment, the learner has access to all of the relevant data (e.g. Dresher and Kaye, 1990). A model may also maintain a count of different categories of relevant data, and use a statistical algorithm to make adjustments in its grammar.

Connectionist models

Connectionist (neural) networks consist of simple processing units joined by weighted connections. Short-term memory is represented by the patterns of activation across the units, long-term memory is represented by the pattern of weights on the connections, and learning consists in adjusting the weights in response to inputs, and output targets if these are available. Connectionist models are most appropriate for cognitive tasks that involve parallel constraint satisfaction, the solving of best-match problems, and similarity-based generalization, and for which robustness is desirable. They have been applied to a wide range of language acquisition tasks, from word recognition to syntax. (*See* **Connectionism**)

Connectionist networks have two apparent limitations. Firstly, internal representations in neural networks consist simply of patterns of activation across units; they have no overt internal structure. Secondly, a connectionist network responds to a novel pattern entirely on the basis of its similarity to familiar patterns. This means that variables, which should match any item within a domain regardless of its similarity to familiar items, are excluded from connectionist models. As noted above, however, structured representations, variables, and the formal rules they make possible, have long been thought to be characteristic of human language. (*See* **Language, Connectionist and Symbolic Representations of**)

Connectionist modelers have argued that the connectionist alternative to structured representations and variables may be better able to model human linguistic behavior and learning. In numerous studies, they have shown that networks, especially simple recurrent networks, that have been trained on inputs that embody internal structure, can generalize to novel patterns that are similarly structured (e.g. Morris *et al*., 2000). The internal representations that are learned by such a network, rather than being general-purpose structures, as in a symbolic model, are optimized for the training set. Thus, these models make predictions about human performance that are different from those of the more powerful symbolic models.

Similarly, connectionist modellers have attempted to show that networks can learn to approach variable-like behavior, though the nature of the training set will always influence the network's performance. They have argued that the connectionist account may more closely resemble the performance of humans. Beginning with the well-known model of Rumelhart and McClelland (1986), connectionists have attempted to demonstrate that a connectionist network could learn both regular and irregular morphology in a manner that resembles the behavior of human language learners. The idea is that the rule or rules embodied in the regular patterns are implicit in the weights of the trained network, rather than being explicit as in a symbolic model. (*See* **Morphology**)

This research has led to a debate between proponents of purely associationist models and proponents of dual-mechanism models embodying both associationist and rule-based components for the learning of morphology (and of grammar generally). The debate has inspired new behavioral experiments, as well as refinements to connectionist architectures and modes of representation. It is still not clear whether connectionist networks are adequate for handling apparently rule-governed linguistic behavior. For recent positions on either side of the debate, see Marcus (2000) and Nakisa *et al*. (2000).

Hidden Markov models

Hidden Markov models (HMMs) are applicable to problems in which there are sequences of input elements ('observations') selected from a finite set. An HMM is an abstract characterization of such a sequence in terms of a set of 'hidden states'. Each hidden state has an 'output probability' associated with each of the possible input elements, a 'transition probability' associating it with each other hidden state, and an 'initial probability' representing its likelihood of beginning a sequence. An HMM learning algorithm assigns these three types of probabilities to the states in a set of HMMs in such a way that the regularities in the available observation sequences are captured. (*See* **Markov Decision Processes, Learning of**)

A set of HMMs can be used to classify input sequences. For example, each HMM may represent a particular word that is input in the form of a sequence of phonemes. Alternately, an HMM can be used to estimate the probability that each element in a sequence belongs to a particular category. For example, each hidden state in an HMM could represent a particular syntactic category; a path through the HMM would then represent a sequence of syntactic categories associated with an observation sequence in the form of an input sentence.

HMMs have been applied to the acquisition of phonetic categories, word forms, syntactic categories, and probabilistic grammars. Since they are a finite-state formalism, they are limited in terms of the sorts of regularities in sequences that they can handle. They have not been proposed as a

general-purpose model of natural language processing and acquisition.

Other models

Other statistical learning techniques, especially those involving 'dimensionality reduction' or some information-theoretic construct such as relative entropy (Resnik, 1996), have been applied to the problem of learning lexical semantics. One view of lexical semantics is that words get their meanings in part from their pattern of co-occurrences with other words. 'Latent semantic analysis' (Landauer and Dumais, 1997) is one well-known example of a model that learns about word meaning by accumulating co-occurrences in corpora. It maintains a table of co-occurrences of words with contexts (for example, encyclopedia articles). For each word in the corpus this yields a vector representing its co-occurrences with the various contexts. This vector is subjected to singular value decomposition, a technique for reducing dimensionality. The vector for each word is a point in a high-dimensional space which is taken to represent its semantics. (*See* **Information Theory**; **Latent Semantic Analysis**)

There has been some recent interest in memory-based learning models for language acquisition (Daelemans, 1999). These models, borrowed from machine learning, base generalization on stored exemplars, rather than on abstractions over exemplars, as in a symbolic model (which induces rules) or a connectionist model (which combines the exemplars in its weights).

FUTURE PROSPECTS

There is deep disagreement over some of the most basic issues in computational language acquisition. Yet the ideas that have been generated by the modeling of human language acquisition from such diverse perspectives represent real progress. We can expect the debate over whether language requires an explicit rule-learning mechanism and innate linguistic constraints to continue, with increasingly specific predictions coming from the different models. In addition, future work should address the following outstanding issues.

Firstly, how could a single model learn to handle a wide range of linguistic behaviors? Because of the different sorts of mechanisms that are required, and the potential for interference when a single device is presented with more than one task, some modularity will presumably be needed in any large-scale model, whether it be symbolic or connectionist. (*See* **Modularity**)

Secondly, how well do the models handle real data? Symbolic models will need to adopt statistical learning components to overcome their brittleness, and connectionist models will need to address the problem of how later learning can interfere with earlier learning. (*See* **Language Acquisition and Language Change**)

Finally, how do language acquisition, historical language change, and the evolution of human language interact? Modelers have begun to build computational models of aspects of language evolution (e.g. Cangelosi and Parisi, 2001) and of historical language change (e.g. Niyogi and Berwick, 1997). An integrated computational picture of change at all three timescales may help to solve some remaining problems concerning language acquisition.

References

Anderson J (1977) Induction of augmented transition networks. *Cognitive Science* **1**: 125–157.

Berwick RC (1985) *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Cangelosi A and Parisi D (eds) (2001) *Simulating the Evolution of Language*. London, UK: Springer-Verlag.

Daelemans W (ed.) (1999) *Journal of Experimental and Theoretical Artificial Intelligence* **11**(3). [Special issue on memory-based language learning.]

Dresher E and Kaye JD (1990) A computational learning model for metrical phonology. *Cognition* **34**: 137–195.

Elman JL (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* **7**: 195–225.

Gibson E and Wexler K (1994) Triggers. *Linguistic Inquiry* **25**: 407–454.

Gold E (1967) Language identification in the limit. *Information and Control* **10**: 447–474.

Landauer T and Dumais S (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 211–240.

Marcus GF (2000) Children's overregularization and its implications for cognition. In: Broeder P and Murre J (eds) *Models of Language Acquisition: Inductive and Deductive Approaches*, pp. 154–176. Oxford, UK: Oxford University Press.

Miikkulainen R (1993) *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

Morris WC, Cottrell GW and Elman JL (2000) A connectionist simulation of the empirical acquisition of grammatical relations. In: Wermter S and Sun R (eds) *Hybrid Neural Models*, pp. 175–193. Heidelberg, Germany: Springer-Verlag.

Nakisa R, Plunkett K and Hahn U (2000) Single- and dual-route models of inflectional morphology. In: Broeder P and Murre J (eds) *Models of Language*

Acquisition: Inductive and Deductive Approaches, pp. 201–222. Oxford, UK: Oxford University Press.
Niyogi P and Berwick RC (1997) A dynamical systems model for language change. *Complex Systems* **11**: 161–204.
Quine WVO (1960) *Word and Object*. Cambridge, MA: MIT Press.
Regier T (1996) *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press.
Resnik P (1996) Selectional constraints: an information-theoretic model and its computational realization. *Cognition* **61**: 127–159.
Roeper T and Williams E (eds) (1987) *Parameter Setting*. Dordrecht, Netherlands: Reidel.
Rumelhart DE and McClelland JL (1986) On learning the past tense of English verbs. In: McClelland JL and Rumelhart DE (eds) *Parallel Distributed Processing*, vol. II. Cambridge, MA: MIT Press.
Tesar B and Smolensky P (1998) Learnability in optimality theory. *Linguistic Inquiry* **29**: 229–268.

Further Reading

Broeder P and Murre J (eds) (2000) *Models of Language Acquisition: Inductive and Deductive Approaches*. Oxford, UK: Oxford University Press.
Charniak E (1993) *Statistical Language Learning*. Cambridge, MA: MIT Press.
Elman JL, Bates E, Johnson M *et al*. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
MacWhinney B (ed.) (1999) *Emergence of Language*. Hillsdale, NJ: Lawrence Erlbaum.

Learning from Advice

Intermediate article

David C Noelle, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

Computer models of how humans learn from direct instruction by a knowledgeable teacher have demonstrated how language processing mechanisms may be leveraged to rapidly acquire and apply new knowledge, as well as how such explicit knowledge may be refined through experience.

INTRODUCTION

There is a wide variety of learning strategies that appear to be shared by humans and non-human animals. The advent of language, however, has given rise to a distinctly human form of learning. Through language, new knowledge may be directly transfered from one individual to another in the form of instructions or advice. Learning from advice (also called learning from direct instruction, instructed learning, or learning by being told) has a number of advantages over other learning strategies. Compared with learning from induction over many example experiences, new knowledge can be acquired very rapidly when presented via direct instruction. Furthermore, fairly abstract knowledge can be communicated using language. Instructors can explicitly specify the situations in which their advice might be fruitfully deployed, helping the learner determine the generality of the instructions. This ability to make advice conditional on hypothetical situations also allows learners to gain knowledge about useful courses of action in rare or dangerous situations, in which personal experience is difficult or risky to obtain (for example, what to do if your parachute fails to open). Also, the speed with which abstract knowledge may be acquired is conducive to very flexible learning. Knowledge may be modified with the same rapidity as its initial acquisition, given appropriate instructions. Lastly, making knowledge explicit through advice facilitates uniformity and coordination of behavior between individuals. Learning from advice is a critical component of the transmission of cultural knowledge to each member of a society.

Researchers interested in the computational structure of human cognition recognized the importance of advice taking very early. In his most widely cited paper on artificial intelligence, Alan Turing discussed the benefits of instructed learning over learning from rewards and punishments (Turing, 1950); and within a decade of Turing's paper John McCarthy began laying the foundations for an 'advice taker' computer program (McCarthy, 1958). Despite this early interest, however, relatively little work has been done on computational models of learning from advice. This article describes some approaches that have been taken to the modeling of instructed learning.

OPERATIONALIZATION

The central problem of learning from advice is the translation of statements in a human language into internal representations that can influence behavior in beneficial ways. This process is analogous to compiling a computer program into executable machine code. Learning from direct instruction can be more difficult than program compilation, however, because commonly offered advice is rarely as precise and complete as the instructions of a computer program. The cognitive mechanisms underlying instructed learning must be able to use established knowledge to infer appropriate behaviour from vague advice (e.g. being told that 'to win at this card game, you should generally avoid taking points' does not specify exactly which card to play at each step of the game – this must be inferred), and these mechanisms must be able to determine the range of situations in which the given advice is applicable (e.g. there could be specific situations in which it is good to take points). This process of transforming instructions into behaviorally efficacious internal representations – into conditionalized plans of action – is called operationalization. In humans, operationalization is accomplished quickly and produces knowledge

representations that can be flexibly refined by experience.

TRANSFORMING ADVICE INTO RULES

Some theories of cognition posit internal mental representations encoded in a 'language of thought'. From this perspective, the process of operationalization initially involves translating advice provided in a natural language into this internal language of logical propositions and rules. If the format of these sentential representations is similar to that of natural language, this initial translation process may be simple. Once this process is complete, standard rule-based reasoning mechanisms may be used to transform the given advice into action. Also, rule refinement techniques may be used to generalize from provided instructions and to adapt the agent's knowledge in response to experience.

One of the earliest investigations of this approach involved a computer program called FOO (First Operational Operationalizer) (Mostow, 1983). FOO was designed to learn to play a card game called 'hearts'. The focus of this work was on the operationalization of abstract and incomplete forms of advice, such as 'avoid taking points'. Statements of this kind were encoded as propositional goals, and a heuristic backward chaining search was conducted over a knowledge base of rules to find chains of actions that, when applied, resulted in the attainment of the given goals. The knowledge base was then augmented with rules that performed these action plans under the appropriate conditions. One of the main lessons drawn from the work on FOO was how computationally difficult it is to translate general preferences and constraints into plans of action. The system required long chains of inference in order to generate such plans, suggesting that human learners, who appear to cogitate less on such instructions, may postpone the burden of operationalizing such advice until times when action is required.

The suggestion that learning mostly occurs in the context of actually performing a task became the focus of a subsequent computer model of instructed learning called Instructo-SOAR (Huffman, 1994). While executing a 'blocks world' construction task, this program could solicit and receive natural language instructions related to the current task context. A natural language understanding system translated instructions into production rules, which could directly influence behavior. As in FOO, advice invoked an effort to fabricate general rules of action, but this effort was limited in Instructo-SOAR. Given an instruction to take a particular action in a particular situation, Instructo-SOAR would use a form of explanation-based learning to find a chain of rules that related the advised action to the current goals. If found, this chain would be abstracted into new general knowledge. If such an explanation was not easily found, however, the program would abandon its attempt to generalize the given advice and would record the instructions as pertinent only within the current context. One result of this approach was speedy response to instructions, with the effort to generalize knowledge distributed over multiple practice sessions. Instructions also became integrated with other sources of knowledge through SOAR's chunking mechanism, a technique for abstracting rules based on experience. In brief, work on Instructo-SOAR showed that the computational problem of translating instructions into an operational form is greatly simplified when such instructions appear in the context of a specific situation.

TRANSFORMING ADVICE INTO WEIGHTS

Artificial neural networks have been successfully used to model various forms of inductive learning in humans and non-human animals, but learning from advice poses a particular challenge to this computational modeling framework. Connectionist theories of cognition present behaviour as arising from the parallel and distributed flow of activity between very simple processing units, with knowledge encoded in the strengths of connections between units. The internal representations used by connectionist networks are profoundly different in structure from natural language utterances, and it is difficult to imagine how direct instruction might be translated into knowledge encoded as connection weights.

Techniques have been developed, however, for encoding a collection of propositional rules in a connectionist network in a manner which allows inference to be performed through the propagation of activity between units. For example, if the system is told that a particular action should be taken in a particular situation, separate processing units may be used to represent the features of the situation and the features of the prescribed action, and strongly weighted connections may be placed between the situation units and the action units, causing the perception of the given situation to activate a representation of the appropriate action. Logical operators, such as 'and' and 'or', may be

implemented by selecting appropriate weights for hidden units, and such hidden units may be used to group situations and actions according to provided rules.

These techniques suggest a hybrid approach to modeling learning from advice. Instructions may be translated into propositional rules, and these rules may, in turn, be used to construct a connectionist network which performs the given task. The connection weights may then be further refined based on experience, using standard connectionist methods for inductive learning. This approach naturally supports generalization of knowledge provided via instruction, since similar situations will tend to activate similar candidate actions through the normal process of activation propagation.

Such a hybrid model of instructed learning, incorporating a neural network construction module along with the resulting connectionist network, was developed in RATLE (Reinforcement and Advice-taking Learning Environment) (Maclin, 1995). This system was given the task of successfully playing a computer game called 'Pengo'. Advice was provided using an artificial rule-based language, and instructions were compiled into networks which could then be refined using a connectionist reinforcement learning algorithm. RATLE was able to operationalize instructions involving plans and iterative procedures as well as condition–action rules, and the system handled incomplete instructions robustly, learning from experience what was not explicitly provided by the teacher. This system demonstrated how the inductive learning strengths of connectionist networks could be augmented with the ability to learn from explicit advice.

TRANSFORMING ADVICE INTO ACTIVATION PATTERNS

Some connectionist modelers view processing elements as an abstraction of neural tissue in the brain. From this perspective, responding to instructions by rapidly constructing networks whose weights encode the given advice seems like an implausible account of human learning. Although neurogenesis (i.e., the appearance of new neural cells) is thought to occur in the adult cortex, there is little evidence for mechanisms that would be capable of promptly recruiting these cells and wiring them together appropriately. There are also problems with a perspective in which instruction quickly sets weight values in an established network architecture. Such rapid weight changes can introduce performance difficulties, such as catastrophic forgetting of previously acquired knowledge.

An alternative connectionist approach to learning from advice avoids these problems by encoding explicitly provided knowledge in the activation states of units rather than in weight values. The insight behind this approach is that the behaviour of a recurrently connected connectionist network is determined not only by connection weights but also by the activation levels of its units. These activation levels can be substantially influenced by previous inputs to the network, such as verbal instructions. An input activation pattern that encodes advice may be actively maintained over time, through the use of attractor networks, and the actively maintained internal representation of the given instructions can immediately influence how the network performs some task.

Early work on parallel distributed processing models suggested this segregation of implicit knowledge encoded in weight values and more explicit knowledge encoded in activation levels (Rumelhart *et al*., 1986). More recently, an 'activation space' model of instructed learning (Noelle, 1997) was proposed and simulated in the context of a category learning task. This model inductively learned to form internal representations of advice which was provided as input to the network, and these internal representations could be actively maintained over time. On receiving input instructions, these internal representations could immediately guide behaviour by propagating activity to units involved in the selection of actions. Also, the knowledge provided via advice could be refined through the continued modification of weights based on experience and feedback. While these simulations involved only very restricted types of instruction, they do suggest that connectionism might provide a biologically plausible account of learning from advice.

SUMMARY

Learning from advice is unlike other learning strategies in that it allows for the very rapid acquisition of abstract knowledge. The central problem faced by a computational account of instructed learning is that of operationalization: of transforming explicit advice into internal knowledge representations that can directly influence behavior. Computational models of learning by being told have been proposed in which instructions are translated into symbolic rules, connectionist weight values, and patterns of activity in a parallel distributed

processing network. Research continues on this topic, focusing on the biological foundations of instruction following and on integrating learning from advice with other learning strategies.

References

Huffman SB (1994) *Instructable Autonomous Agents*. PhD thesis, University of Michigan, MI.

Maclin RF (1995) *Learning From Instruction and Experience: Methods for Incorporating Procedural Domain Theories Into Knowledge-Based Neural Networks*. PhD thesis, University of Wisconsin, WI.

McCarthy J (1958) Programs with common sense. In: *Proceedings of the Symposium on Mechanisation of Thought Processes*, vol. I, pp. 77–84. London, UK: HMSO.

Mostow DJ (1983) Machine transformation of advice into a heuristic search procedure. In: Michalski RS, Carbonell JG and Mitchell TM (eds) *Machine Learning: An Artificial Intelligence Approach*, vol. I, pp. 367–403. Palo Alto, CA: Tioga.

Noelle DC (1997) *A Connectionist Model of Instructed Learning*. PhD thesis, University of California, CA.

Rumelhart DE, Smolensky P, McClelland JL and Hinton GE (1986) Schemata and sequential thought processes in PDP models. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. II, pp. 7–57. Cambridge, MA: MIT Press.

Turing AM (1950) Computing machinery and intelligence. *Mind* **59**: 433–460.

Further Reading

Hayes-Roth F, Klahr P and Mostow DJ (1981) Advice-taking and knowledge refinement: An iterative view of skill acquisition. In: Anderson JR (ed.) *Cognitive Skills and Their Acquisition*, pp. 231–253. Hillsdale, NJ: Erlbaum.

Huffman SB and Laird JE (1993) Learning procedures from interactive natural language instructions. In: Utgoff P (ed.) *Machine Learning: Proceedings of the Tenth International Conference*, pp. 143–150. Amherst, MA: Morgan Kaufmann.

Huffman SB and Laird JE (1995) Flexibly instructable agents. *Journal of Artificial Intelligence Research* **3**: 271–324.

Huffman SB, Miller CS and Laird JE (1993) Learning from instruction: a knowledge-level capability within a unified theory of cognition. In: *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 114–119. Boulder, CO: Erlbaum.

Maclin R and Shavlik JW (1994) Incorporating advice into agents that learn from reinforcements. In: *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 694–699. Seattle, WA: AAAI Press.

Maclin RF and Shavlik JW (1996) Creating advice-taking reinforcement learners. *Machine Learning* **22**: 251–281.

Noelle DC and Cottrell GW (2000) *Individual differences in exemplar-based interference during instructed category learning*. In: Gleitman LR and Joshi AK (eds) *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pp. 358–363. Philadelphia, PA: Erlbaum.

Learning Rules and Productions

Intermediate article

Niels A Taatgen, University of Groningen, Groningen, Netherlands

CONTENTS

Learning rules and productions involves creating new rules through inference, induction and compilation and is used in both machine learning and cognitive modeling.

INTRODUCTION

Rules are a popular means of knowledge representation used in several different domains of cognitive science. Not only are they a powerful form of representation, they differ from many other types of representation in the sense that they incorporate both the knowledge itself, and the way to use this knowledge. (*See* **Knowledge Representation, Psychology of**)

Different Types of Rules

Rules are small units of knowledge that, although they work in concert with other rules, are relatively independent, as opposed to a line of code in an arbitrary programming language. The attractive property of this independence is that knowledge can be built up incrementally. The focus of this article will be on two types of rules, logical rules and production rules.

Logical rules are a subclass of first-order logic expressions, often Horn clauses. A Horn clause is an expression of the form:

$$(L_1 \text{ and } L_2 \text{ and} \ldots \text{and } L_n) \rightarrow H \qquad (1)$$

In this expression, L_1 to L_n and H are all literals. Horn clauses closely resemble production rules, which have the following form:

IF condition_1 and condition_2 and ... and condition_n THEN action_1, action_2 ... action_m

For the present discussion, we will largely ignore the difference between the two, but the reader must be aware that they have different properties. As an example, production rules are generally used left-to-right: once the conditions of the rule are satisfied, the action can be carried out. Horn clauses on the other hand, when used in a Prolog context, are used the other way around: in order to satisfy some predicate, a rule is selected that has the predicate as a conclusion.

Goals of Learning Rules

Before examining mechanisms for rule learning, it is useful to characterize the context in which rules are learned, and the goals of rule learning. Two broad fields can be distinguished: machine learning and cognitive modeling. (*See* **Machine Learning**; **Computational Models: Why Build Them?**)

Machine learning

The main focus in this field is to learn rules that characterize concepts. The goal is to find a set of rules that can decide whether or not a particular example is an instance of a certain concept. A concept can correspond to a natural category like a bird, a dog or a chair, or concepts like 'paper accepted at the cognitive science conference'. Suppose we want to characterize this latter category, papers accepted at a certain conference. The final part of the decision process is whether or not to accept a paper given the judgments of the reviewers. Table 1 gives an example of judgments. (*See* **Concept Learning**; **Concept Learning and Categorization: Models**)

The goal here is to find a rule or set of rules that characterizes the concept of 'accepted paper'. Given the five examples we may come up with the following rules:

$$(\text{Overall A} = \text{Good}) \wedge (\text{Overall B} \neq \text{Poor}) \rightarrow \text{Accept} = \text{yes} \qquad (2)$$

$$(\text{Overall B} = \text{Good}) \wedge (\text{Overall A} \neq \text{Poor}) \rightarrow \text{Accept} = \text{yes} \qquad (3)$$

Table 1. Five judgments in the conference example

	Relevance A	*Technical A*	*Overall A*	*Relevance B*	*Technical B*	*Overall B*	*Accept?*
1	Good	Fair	Good	Fair	Fair	Good	Yes
2	Good	Good	Good	Fair	Poor	Poor	No
3	Good	Good	Good	Fair	Poor	Fair	Yes
4	Fair	Poor	Fair	Good	Good	Good	Yes
5	Good	Fair	Poor	Good	Good	Fair	No

The assumption in this and later examples will be that if Accept is not set to 'yes' by some rule, Accept will be 'no'. Although this may look like a very plausible characterization of the five examples, it is by no means the only one. It may be too general, as it does not take into account attributes other than Overall A and Overall B.

The following three rules also characterize the five examples:

$$\begin{aligned}&(\text{Relevance A} = \text{Good}) \wedge (\text{Technical A} = \text{Fair})\\&\quad\wedge(\text{Overall A} = \text{Good})\wedge\\&(\text{Relevance B} = \text{Fair}) \wedge (\text{Technical B} = \text{Fair})\\&\quad\wedge(\text{Overall B} = \text{Good}) \rightarrow \text{Accept} = \text{yes}\end{aligned} \quad (4)$$

$$\begin{aligned}&(\text{Relevance A} = \text{Good}) \wedge (\text{Technical A} = \text{Good})\\&\quad\wedge(\text{Overall A} = \text{Good})\wedge\\&(\text{Relevance B} = \text{Fair}) \wedge (\text{Technical B} = \text{Poor})\\&\quad\wedge(\text{Overall B} = \text{Fair}) \rightarrow \text{Accept} = \text{yes}\end{aligned} \quad (5)$$

$$\begin{aligned}&(\text{Relevance A} = \text{Fair}) \wedge (\text{Technical A} = \text{Poor})\\&\quad\wedge(\text{Overall A} = \text{Fair})\wedge\\&(\text{Relevance B} = \text{Good}) \wedge (\text{Technical B} = \text{Good})\\&\quad\wedge(\text{Overall B} = \text{Good}) \rightarrow \text{Accept} = \text{yes}\end{aligned} \quad (6)$$

This set of rules, though technically correct, is much less satisfactory, because it just lists examples 1, 3 and 4, and is therefore too specific. The goal of a good rule learning algorithm is to find the right set of rules for a certain set of positive and negative examples.

The general procedure in machine learning is that the learning algorithm is trained on a set of examples for which the answers are provided. After learning, the rule set that has been developed is tested on a new set of examples, the test set. The quality of the algorithm is judged by its efficiency, and the score on the test set.

Cognitive modeling

If rules are considered not merely as convenient representations, but as atomic components of human knowledge (Anderson and Lebiere, 1998) then such a theory of human knowledge has to include mechanisms to learn these rules. Whereas the goal of machine learning is to find a rule set that characterizes some concept, the focus in cognitive modeling is more on the learning process than the learning outcome. A cognitive modeler tries to produce a computer simulation that mimics human learning as closely as possible. Cognitive modeling approaches that use rules have to answer the question of how these rules are learned, and what the effects of rule learning are on performance. (*See* **Unified Theories of Cognition**; **Production Systems and Rule-based Inference**; **Learning, Psychology of**)

An issue in cognitive modeling is how task-specific rules can be learned on the basis of general rule knowledge on the one hand and instruction and experience on the other. Anderson (1987) uses the example of learning to program in Lisp. When novices have to learn a new skill like programming, they rely not only on general instruction but also on examples that can be used as templates. (*See* **Skill Learning**; **Learning, Psychology of**)

In an experiment, participants were given a template on how to define Lisp functions, and an example:

```
(DEFUN <function name>
  (<parameter 1><parameter 2>...
  <parameter n>)
  <process description>)
(DEFUN F-TO-C (TEMP)
  (QUOTIENT (DIFFERENCE TEMP 32)
  1.8))
```

They were then given the assignment to write a Lisp definition of a new function FIRST that returns the first element of a list. Many participants came up with the following definition:

```
(DEFUN FIRST (LIST1)
  (CAR (LIST1)))
```

They had produced this definition by using both the general template and the example, but had wrongly generalized the (DIFFERENCE TEMP 32) part to produce (LIST1) in the answer. The paren-

theses are present because DIFFERENCE is itself a function call. No parentheses are needed in the case of LIST1 as this is one of the parameters. The correct solution is:

```
(DEFUN FIRST (LIST1)
  (CAR LIST1))
```

Anderson's rule learning system, which we will examine in detail later on, produced the following two rules in a simulation of the knowledge acquisition in this task:

```
IF   the goal is to write a function
     of one variable
THEN write (DEFUN function (vari-
     able) and set as a subgoal to
     code the relations calculated
     by this function and then
     write)
IF   the goal is to code an argument
     and that argument corresponds
     to a variable of the function
THEN write the variable name
```

Note that these functions are generalizations of both examples, but that the first rule is a specialization of the general template, as it only applies to functions of one variable.

In the Lisp example, general knowledge and a single example are used to find the solution to a new example. Except for the template and the example, general strategies like analogy are assumed in the model. Almost as a by-product, rules are learned. Although this setting is different from the machine learning perspective, there is a strong resemblance: rules are learned on the basis of examples, in this case with some domain knowledge. In cognitive modeling the focus is also on errors and speed, so the cognitive model also has to make the same errors people do, and show the same increase in performance due to practice.

ALGORITHMS FOR CONCEPT LEARNING

Single Hypothesis Learning

An early rule learning algorithm was developed by Winston (1970). Many variants have been produced, but the basic idea is very simple. A single rule or rule set is maintained, and this rule is adjusted as new examples arrive.

When a new example is presented, it is first checked to see if the rule is already consistent with the example. When the example is inconsistent, we have to update the rule. If the example is a positive example, we have to generalize the rule, so that it will include the new example. In the case of a negative example, the rule has to be specialized to exclude the new example.

Table 2. Simplification of the conference example

	Reviewer A	*Reviewer B*	*Accept?*
1	Good	Good	Yes
2	Good	Poor	No
3	Good	Fair	Yes
4	Fair	Good	Yes
5	Poor	Fair	No

Consider a simplification of the conference acceptance problem in Table 2.

A possible rule based on the first example is:

$$\text{A} = \text{Good} \rightarrow \text{Accept} = \text{yes} \ (1) \tag{7}$$

The second, negative example is inconsistent with this rule, so specialization is needed, for example by adding a condition:

$$\text{A} = \text{Good} \wedge \text{B} \neq \text{Poor} \rightarrow \text{Accept} = \text{yes} \ (2) \tag{8}$$

The third example is consistent with this rule, so no further modification is necessary. The fourth example, however, again requires a generalization of the rule. This may be achieved by dropping a condition, resulting in:

$$\text{B} \neq \text{Poor} \rightarrow \text{Accept} = \text{yes} \ (3) \tag{9}$$

Example five is again inconsistent with the rule, requiring a final specialization:

$$\text{B} \neq \text{Poor} \ \wedge \text{A} \neq \text{Poor} \rightarrow \text{Accept} = \text{yes} \ (4) \tag{10}$$

A problem with this approach is that with each generalization or specialization we have to make sure all the previous examples are still consistent with the current rule. Also, finding a new hypothesis can become very hard, as there are many possible generalizations and specializations, and they are not all as easy to derive from the current rule.

Version Space

In the example two methods were used to update the rule: generalization and specialization. Rules can be ordered with respect to generality: rule p is said to be more general than rule q when all instances that are included in the concept by rule q are also included by rule p. This ordering is a partial order. The most general rule is the rule that includes all instances of the concept, and the most specific rule is the rule that includes no instances at all. All other rules are in between these extremes.

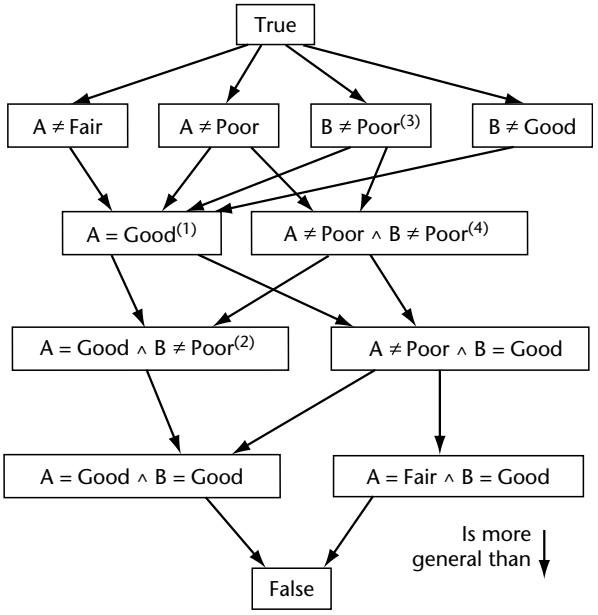


Figure 1. Part of the hypothesis space of the conference problem.

This view on hypotheses offers a handle for a more systematic approach than the one offered by single hypothesis learning. Figure 1 shows part of the hypothesis space for the conference problem.

The most general hypothesis is that all papers are accepted (True), and the most specialized hypothesis is that none are accepted (False). All other hypotheses are ordered in between. The numbers in Figure 1 refer to the four hypotheses we had in the previous example: each time a positive example is not included in the current hypothesis, we have to generalize and move up in hypothesis space, and each time a negative example is not excluded we have to move down. The version space, the space of all plausible versions of the concept, is the subset of all possible hypotheses that are still consistent with the examples we have seen up to now.

Version Space Learning

Instead of maintaining a single hypothesis about the target concept, an alternative is to represent all hypotheses that are still consistent with the present set of examples. The naive version would be to list all the hypotheses, but fortunately this is not necessary. Due to the fact that hypotheses are partially ordered by the more-general-than relation, it is only necessary to represent the set of the most general hypotheses that are still consistent with all examples, and the set of the most specific hypotheses that are still consistent with all examples. These two sets, usually designated *G* and *S* respectively, are often called boundary sets, as everything more general than *G* or more specialized than *S* is not consistent with the examples, but everything in between is. The algorithm that maintains these boundary sets is known as version space learning (Mitchell, 1977). *G* is initialized to (True), and *S* to (False). For each example, *G* and *S* are updated along the following lines.

For positive examples:

- Remove all members of *G* that do not include the example.
- For each member of *S* that does not include the example, replace it by all immediate generalizations of that member that do include the example, and are specializations of some member of *G*.

For negative examples:

- Remove all members of *S* that do not exclude the example.
- For each member of *G* that does include the example, replace it by all immediate specializations of that member that do not include the example, and are generalizations of some member of *S*.

Figure 2 shows how version space learning would handle the conference acceptance example. The algorithm starts out with *G* set to (True) and *S* set to (False). The first example is a positive example: `A=Good ∧ B=Good`. This means that False is no longer the most specific hypothesis, so it is generalized to `A=Good ∧ B=Good`. The second example is a negative example, which means that the True hypothesis needs specialization. Although there are two specializations of True consistent with the second example `(A≠Good, B≠Poor)`, only

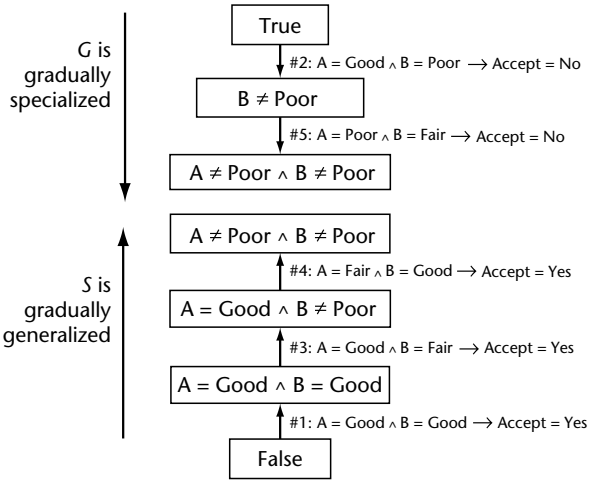


Figure 2. Version space learning operating on the conference example.

B ≠ Poor is also a generalization of a member of *S*. The process continues until, after example 5, both *G* and *S* have converged to the hypothesis A ≠ Poor ∧ B ≠ Poor. Note that in this example both *G* and *S* only contain one hypothesis at a time; this is not true in general.

A disadvantage of version space learning is that it cannot handle noise: it relies on the fact that all examples are correct. Another problem is that it cannot handle disjunction very well. In our example we have only used a limited form of disjunction (in the sense that B ≠ Poor means B = Good ∨ B = Fair). The case A = Fair ∧ B = Fair is classified as a positive example, although it has never been presented as an example, and it would be perfectly plausible for it to have been a negative example (in the case the criterion is 'At least one Good and no Poors').

Other Concept Learning Algorithms

Besides single hypothesis learning and version space learning, many other algorithms have been developed for concept learning. Decision tree learning, for example, learns a decision tree for a concept. Learning is, however, not incremental, because the algorithm learns the whole tree on the basis of a set of examples. As a decision tree is not really a rule-based representation, it is beyond the scope of this article. Another family of algorithms, often shared under the heading of inductive logic programming, infers sets of rules to characterize concepts instead of single rules. The advantage of using a set of rules is that single rules only cover part of the concept, so there is no problem of overly specific rules. Inductive logic programming algorithms also operate on the set of examples as a whole.

LEARNING WITH DOMAIN KNOWLEDGE

A property of the algorithms discussed in the previous paragraph is that they operate on the basis of examples only. In addition to examples there may be other knowledge that may guide the learning process, for example common sense knowledge, or a complete set of domain knowledge. An algorithm that includes the use of domain knowledge is explanation-based learning (EBL) (DeJong, 1981).

Explanation-Based Learning

The assumption of explanation-based learning (EBL) is that we have a complete set of knowledge, the domain knowledge, that is in principle enough to make decisions, but may be computationally intractable. The goal of EBL is, given a proof for a single example, to derive new rules of intermediate computational complexity that are generalizations of the example, but specializations of the domain theory. Suppose we have the following rule set to make conference decisions:

$$\text{reviewer}(X) \wedge \text{review}(X, \text{poor}) \rightarrow \text{negative} \tag{11}$$

$$\text{reviewer}(Y) \wedge \text{review}(Y, \text{good}) \rightarrow \text{positive} \tag{12}$$

$$\text{negative} \rightarrow \text{decide}(\text{no}) \tag{13}$$

$$\text{positive} \wedge \neg\text{negative} \rightarrow \text{decide}(\text{yes}) \tag{14}$$

Given a specific example, for instance reviewer(a) ∧ reviewer(b) ∧ review(a, good) ∧ review(b, good) a general problem solver like Prolog can derive decide(yes). EBL now uses the proof (or explanation, hence the name) of decide(yes) to generate a new rule that is a generalization of the example but a specialization of the domain theory:

$$\neg\text{review}(a, \text{poor}) \wedge \text{review}(b, \text{good}) \rightarrow \text{decide}(\text{yes}) \tag{15}$$

This newly learned rule does not contribute anything new, as all the knowledge is already contained in the domain knowledge. However, if the domain knowledge itself is very inefficient to use, new efficient rules may effectively extend the capabilities of the system by allowing new proofs that were previously computationally unachievable. Take the example of mathematics: the natural numbers can be defined by a small set of axioms, but mathematicians use many derived rules to solve actual problems.

In EBL new rules are added to the rule set, so in that sense it differs from algorithms discussed earlier where learned rules replaced old rules. This introduces a new problem: adding rules to the system may improve its performance because rules are tailored to certain often-occurring situations, but may also decrease performance if they are never used. This *utility problem* is an issue different from correctness: knowledge that is true may still be undesirable because it is useless. Possible solutions to the utility problem are to develop procedures that estimate the cost/benefit properties of a rule in advance, or to just introduce them into the system and keep track of how they fare (Minton, 1988).

LEARNING RULES IN COGNITIVE MODELS

Rule learning in cognitive modeling has aspects in common with both the purely inductive algorithms like version space learning and deductive algorithms like explanation-based learning. Human learners have a large store of background knowledge, strategic knowledge and domain knowledge, but still have to make generalizations from examples that are not fully deducible from the domain knowledge. A general view in cognitive modeling is that humans have a set of weak methods that, when supplied with some background knowledge and a particular case to work on, will produce the desired performance, and are also the basis for learning. Weak methods are problem-solving strategies that are independent of the particular problem, and are generally applicable. Examples of these strategies are: means-ends analysis, forward-checking search, analogy, etc.

In the Lisp learning example in the introduction, the weak method of analogy was used to generate a new Lisp program on the basis of an example and some background knowledge of Lisp. During this process some new rules were learned, and the mechanisms cognitive modelers use to achieve this rule learning show a resemblance to explanation-based learning. The difference is that learning and problem-solving happen at the same time, and domain knowledge is often incomplete.

Chunking in Soar

Newell and Rosenbloom (1981) proposed a rule mechanism called *chunking* that became an important component of the Soar cognitive architecture (Newell, 1990). Within Soar, learning rules are tied to impasses and subgoaling. Whenever Soar reaches a state in which it runs into some impasse (no applicable rules, an irresolvable choice between operators, etc.), it automatically creates the subgoal to resolve this impasse. When the subgoal is successfully completed and the impasse is resolved, a specialized rule is learned that summarizes all the processing required to achieve that subgoal. If Soar later encounters a similar impasse, it no longer needs a separate subgoal to process it. Instead it can use the learned rule to solve it in a single step. (*See* **Soar**)

Although Soar's basic set of methods encompasses several weak methods, the one reported most often is forward checking. Suppose we have the blocks-world problem in Figure 3.

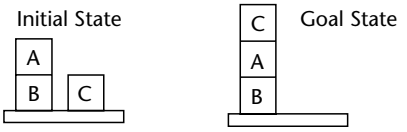


Figure 3. Example blocks-world problem.

If Soar were to solve this problem, it would discover that there are two possible operators in the initial state: 'Move block A to C' and 'Move block C to A'. If Soar had no additional knowledge on how to choose between these operators (called search-control knowledge), it would run into a so-called tie impasse. To resolve this impasse, Soar would create a subgoal by evaluating these two operators. Evaluation can be done by 'mental simulation': what would happen if each of the tied operators was applied? In this mental simulation Soar would quickly discover that 'Move C to A' is the best operator as it immediately accomplishes the goal. So the subgoal created by the impasse comes back with the resolution 'Move C to A is best', and in the main goal this operator is applied.

Learning occurs directly after the subgoal is resolved. The rule that is created has as a condition the circumstances in which the impasse occurred, and as action the result of the subgoal, in this case the 'best preference' for the operator that accomplishes the goal:

```
IF the problem-space is a simple-
blocks-world
  and the state has block X and block Y
  clear,
  and block X is on the table,
  and the goal state has block X on
  block Y,
THEN make a best preference for the
operator that move block X onto
block Y
```

The consequence of this rule is that whenever Soar encounters the situation described in the rule, an impasse no longer occurs and an operator is chosen immediately.

A problem with Soar's rule learning mechanism is that there is no solution to the utility problem: Soar can produce over-specific rules that are expensive to match and has no way to get rid of them once learned.

Production Learning in ACT

Anderson (1983) proposed a set of four rule learning mechanisms that were specified to work in the

ACT* cognitive architecture. An aspect of ACT* is that it has not only a memory for rules (procedural memory), but also a memory for facts (declarative memory). The following four mechanisms were used to learn new rules: (*See* **ACT**)

1. Proceduralization. If a rule accesses declarative memory, and uses the knowledge from declarative memory in its action, then learn a new rule that is identical to the old rule but has the retrieved knowledge instantiated into the rule, eliminating the declarative retrieval.
2. Composition. Collapse a sequence of rules into a single rule that performs all the actions of the individual rules of the sequence.
3. Generalization. If there are two rules that are similar, create a generalized version of these rules. This can be done by removing conditions or by substituting variables for constants.
4. Discrimination. If a rule is successful in one situation but not in another, add conditions to the rule to make a more restrictive version that only applies in successful situations.

In the Lisp learning example, proceduralization and composition are used to learn the two rules to write Lisp functions on the basis of the weak method of analogy.

Despite the relative success of the mechanisms in ACT*, they were too unconstrained, and were able to produce too many invalid rules and too many rules with poor utility. In the latest version of ACT (ACT-R 5.0), a constrained version of the four mechanisms is introduced. This single mechanism combines the proceduralization and composition mechanisms from ACT*. In ACT-R, the expressive power of production rules is more constrained when compared to earlier versions of ACT and to Soar. Each production rule is only allowed a single access to declarative memory. Also, this access is composed of two steps. Firstly, a rule has to issue a request to declarative memory for a certain fact as part of its action side, and then a second rule can match the retrieved fact on its condition side. Suppose we want to add three numbers. In the older ACT and Soar systems, this would require only a single rule. In ACT-R, we need three rules to accomplish this as shown in Table 3.

Using one of the older composition mechanisms, one general rule could be learned out of these three rules to do three-number additions in one step. From a cognitive perspective this is not desirable, as people generally cannot do these types of additions in one step, although they have ample experience with them. Also, in ACT-R it would no longer be possible, as only one retrieval from declarative memory is allowed in each rule.

Rule learning in ACT-R is aimed at composing two rules that fire in sequence into one new rule, while maintaining the constraint that only one retrieval from declarative memory is allowed. This is done by eliminating a retrieval request in the first rule and a retrieval condition in the second rule. The fact that has been retrieved is entered in the combined action of the new rule. Suppose that in the example above, the three numbers that are added are 1, 2 and 3, then this would produce two new rules, a combination of rules 1 and 2, and a combination of rules 2 and 3. Each of these two new

Table 3. ACT-R rules for adding three numbers

Rule 1	*Rule 2*	*Rule 3*
IF the goal is to add three numbers THEN send a retrieval request to declarative memory for the sum of the first two numbers	IF the goal is to add three numbers AND the sum of the first two numbers is retrieved THEN send a retrieval request to declarative memory for the sum of the currently retrieved sum and the third number	IF the goal is to add three numbers AND the sum of the first two numbers and the third number is retrieved THEN the answer is the retrieved sum

Table 4. Combined rules in ACT-R

Rule 1 & 2	*Rule 2 & 3*	*Rule 1 & 2 & 3*
IF the goal is to add 1, 2 and a third number THEN send a retrieval request to declarative memory for the sum of 3 and the third number	IF the goal is to add three numbers and the third number is 3 AND the sum of the first two numbers is retrieved and is equal to 3 THEN the answer is 6	IF the goal is to add 1, 2 and 3 THEN the answer is 6

rules can be combined with one of the original rules to learn a rule that combines all three rules as shown in Table 4.

Compared to the original rules, these rules are very specialized: they work only for certain numbers. The implication is that people will only learn specific rules for very common additions: it is likely that one immediately knows the sum of 1, 2 and 3, but not of 9, 3 and 4.

Although the example is about addition, production learning can also be used to transform general production rules into task-specific rules, and to incorporate previous experiences into rules. The utility problem is handled in two ways: by constraining the size of the rules, rules with many conditions are impossible; rules that are learned are introduced only gradually in the system, ensuring a relatively slow but safe proceduralization.

SUMMARY

Rule learning is a fundamental issue in both machine learning and cognitive modeling. This article has focused on incremental learning mechanisms, in which a current hypothesis is continuously updated on the basis of examples, involving both processes of generalization and specialization. Learning can be with or without domain or background knowledge. If no background knowledge is provided, in algorithms like single-hypothesis learning and version space learning, learning is purely inductive. Explanation-based learning on the other hand deduces its knowledge from the domain theory, guided by examples.

Learning in cognitive modeling often employs a mixture of inductive and deductive methods, as background knowledge is often available but almost never complete. The most common method employed is to instantiate general problem-solving methods with specific knowledge and examples, producing task-specific rules.

References

Anderson JR (1983) *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson JR (1987) Skill acquisition: compilation of weak-method problem solutions. *Psychological Review* **94**: 192–210.

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

DeJong G (1981) Generalizations based on explanations. Proceedings of the Seventh International Joint Conference on Artificial Intelligence, pp. 67–70.

Minton S (1988) *Learning Search Control Knowledge: An Explanation-Based Approach*. Boston, MA: Kluwer.

Mitchell TM (1977) Version spaces: a candidate elimination approach to rule learning. Proceedings of the Fifth International Joint Conference on AI, pp. 305–310. Cambridge, MA: MIT Press.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Rosenbloom PS (1981) Mechanisms of skill acquisition and the law of practice. In: Anderson JR (ed.) *Cognitive Skills and Their Acquisition*, pp. 1–55. Hillsdale, NJ: Erlbaum.

Winston PH (1970) *Learning Structural Descriptions from Examples*. PhD dissertation, MIT Technical Report AI-TR-231.

Further Reading

Anderson JR (1982) Acquisition of cognitive skill. *Psychological Review* **89**: 369–406.

Mitchell TM (1997) *Machine Learning*. New York: McGraw-Hill.

Learning through Case Analysis

Intermediate article

David B Leake, Indiana University, Bloomington, Indiana, USA
Janet L Kolodner, Georgia Institute of Technology, Atlanta, Georgia, USA

CONTENTS

Case-based reasoning is a process for reasoning by remembering prior experiences and adapting their lessons to new circumstances. Cognitive science research on case-based reasoning has developed computational models of how humans learn from experience, and suggests principles for providing more effective support for human learning.

REASONING FROM REMEMBERED EVENTS

When people justify their decisions, or give advice to others, they often support their conclusions with anecdotes about prior experiences. A study of the use of anecdotes in medicine, for example, found that doctors confronted with difficult problems often call upon other doctors with requests 'not for the latest news of research from the journals but for … anecdote[s]: "Anybody had any experience with this?"' (Hunter, 1986). Likewise, when people encounter new situations, they often remember prior experiences and apply their lessons. Remembered prior cases can guide the interpretation of new situations, help to predict new outcomes, suggest solutions to new problems, and warn of potential failures to avoid. In medicine, for example, experiences with prior patients help doctors to interpret their patients' descriptions of their problems, predict the outcomes of diseases, select appropriate treatments, and predict difficulties that may arise.

The process of reasoning and learning from experiences is called 'case-based reasoning' (CBR). Case-based reasoners interpret new situations and solve new problems by adapting the lessons of prior cases. They learn by integrating their own new experiences, and experiences presented to them, into their memories for future use. Their proficiency at reasoning comes from having the right cases in memory, being able to access them at the right times, and being able to apply them in the right ways.

Case-based reasoning has been studied both as a model of human cognition and as an artificial intelligence approach. Cognitive science has developed computational models of the CBR process to specify theories of human cognitive processing, provide testable hypotheses about these processes, and illuminate the constraints on human reasoning relating to information processing. The resulting principles have been the basis of computer systems that reason and learn, as well of systems that aid human reasoning and learning by providing people with useful cases with which to augment their own memories. More recently, the model of cognition developed in CBR research has been the basis of interactive learning environments and theories of how to help students to extract concepts, acquire skills, and improve their ability to transfer knowledge to new situations. This article introduces the CBR process, discusses issues and methods arising from CBR research, and describes the ramifications of the CBR model for understanding and supporting human reasoning and learning.

CASE-BASED REASONING AND THE UTILITY OF ANALOGY

Analogical reasoning transfers lessons from one domain to another, by comparing similar objects, processes, or ideas in order to map information from one to the other. (For a collection of recent research on analogy, see Gentner *et al.*, 2001.)

Case-based reasoning is an analogical reasoning process that focuses on making analogies to one's previous experiences. CBR research studies the memory processes required to retrieve analogous experiences (e.g. Schank, 1999), the processes for comparing them to new situations and adapting their lessons to new needs, and the learning processes that index new experiences in memory to be retrieved when they are relevant in the future.

Analogical reasoning is essential both to everyday reasoning and to more specialized reasoning tasks, and is used by novices and experts alike. When similar problems have been encountered before, it enables reasoners to propose solutions to problems quickly, avoiding the need to derive those answers from scratch. It also enables reasoners to propose solutions in domains that are not completely understood; it provides a method for evaluating candidate solutions when no algorithmic method is available; and it enables interpretation of open-ended and ill-defined concepts by analogy to their prior use. For example: an architect with a difficult assignment may save time by adapting a previous design for similar requirements, rather than starting from scratch; a doctor may choose between treatments by considering their effectiveness with similar patients; an employer may evaluate the promise of a prospective employee based on experience with a similar candidate in the past; a lawyer may try to show that there are mitigating circumstances for a crime by analogy with similar cases in which a sentence was reduced; a parent trying to interpret what a child means by a 'fun' present might do so by considering the other presents the child has enjoyed.

Storing and reasoning from specific experiences, rather than more abstract rules or patterns, provides two other advantages. First, it enables a simple learning process: experiences can be stored without distilling them into abstract rules. Second, cases are often easier to apply than general rules, because they store information at an operational level: they include the specific details which may be useful to reapplying the solution. For example, when a doctor remembers the treatment for a patient with a rare disease, reapplying specifics (e.g. the name of a local physical therapist with appropriate specialized training) may save effort or improve results. Numerous studies provide evidence for human reasoning and learning from specific examples, for tasks such as learning to use a word processor, learning computer programming, mathematical problem solving, medical diagnosis, explanation of events, and decision-making (Kolodner, 1993).

Models of case-based reasoning have been developed for a wide range of tasks including problem solving, planning, design, understanding and explanation, diagnosis, parsing, and legal reasoning (Kolodner, 1993; Leake, 1996). These tasks can be divided into two broad classes: problem solving and interpretation. Problem-solving CBR uses cases to suggest solutions; interpretive CBR uses cases as the basis for interpretation or classification. Both types of CBR involve four main steps, as summarized in Figure 1: retrieval of relevant prior cases; adaptation of their lessons to new needs; evaluation and application of candidate solutions; and learning by storing new cases arising from experience.

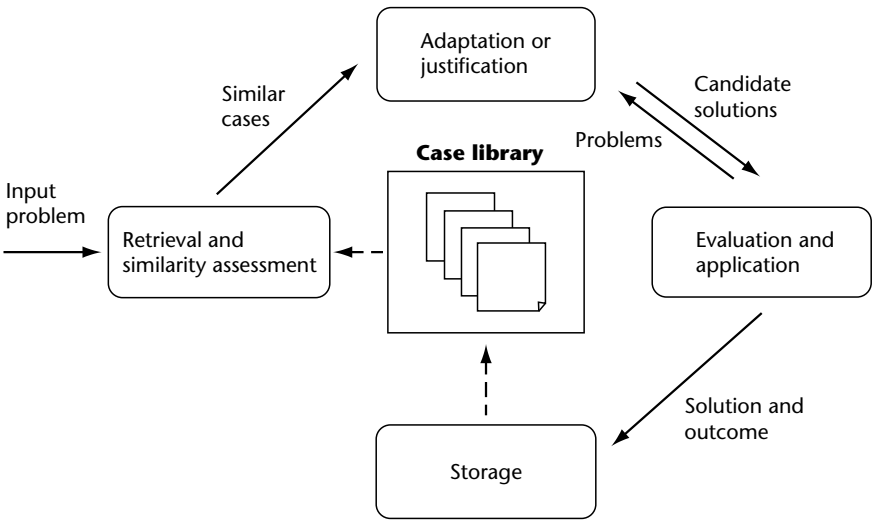


Figure 1. The case-based reasoning cycle.

These four steps are described in outline below. Note that cycles of steps may be repeated or restarted, and each step may draw on case-based reasoning or other methods to support its own processing.

Case Retrieval and Similarity Assessment

Case-based reasoning begins by retrieving cases relevant to the current problem, based on features of the current situation. In problem-solving CBR, cases are retrieved based on the problem and the constraints on the solution (for example, cases for medical treatments might be retrieved based on both the disease to treat and factors such as the patient's allergies). In interpretive CBR, cases are retrieved according to either the situation to classify, or a candidate interpretation to support (for example, when a lawyer is arguing to support a client's viewpoint). In this process, the reasoner generates cues to guide the search for relevant cases in memory, establishes correspondences between the current situation and the retrieved cases, and assesses the similarity of the retrieved cases and the new situation, to identify differences to address.

Case Adaptation or Justification

In problem-solving CBR, the solution to the previous problem is adapted to the new situation (for example, a doctor may adjust a remembered treatment based on the patient's age or weight), or the partial solutions suggested by several cases are merged. In interpretive CBR, a 'justification structure' is constructed to compare and contrast the case with the new situation and show how its lessons apply. Justification structures record the relationships between the prior case and the new situation, and provide reasoning to support a new conclusion in terms of the original conclusion and the relevant similarities and differences. Thus, a justification structure in a legal domain might show how differences between the current circumstances and a prior crime (for example, that one was premeditated and the other was not) suggest that the new crime should be classified as more serious.

Case Evaluation and Application

When a candidate solution or interpretation has been generated, it is evaluated to identify problems needing revision, and a cycle of evaluation and adaptation or justification is repeated until an acceptable solution has been generated (or the process is abandoned). The candidate solution is then applied to the real problem (or a simulation), providing feedback on the solution's performance. For example, after a doctor selects a candidate treatment, it may be necessary to revise it to avoid bad interactions with drugs that the patient is taking for other complaints. The results of applying the treatment provide the doctor with feedback on the effectiveness of different treatments on particular patients.

Case Storage

Every reasoning episode provides a new case for storage and future use. Success cases provide solutions to reuse; failure cases provide warnings of potential problems and suggest possible repairs when those problems occur. In case storage, execution feedback is analyzed and explained to identify useful lessons, which – in conjunction with predictions of when the case may apply – determine the indices under which the case is stored. For example, if an initial treatment caused a bad reaction to another drug that the patient was taking, requiring a change in medication, the new case would be indexed as useful for avoiding the interaction.

MEMORY ORGANIZATION AND RETRIEVAL

In order to build a useful memory of cases, case-based reasoners must extract the lessons of new cases, predict how those lessons might be used in the future, and encode the cases in ways that will make them recognizably useful in those situations. The process of encoding cases to be accessible at the right times is known as 'indexing': the features that predict the cases' applicability are regarded as indices organizing the memory of cases. The appropriateness of the indices used to label a case will depend on the adequacy of the reasoner's background knowledge for analyzing the case's potential applicability, and on the effort put into the case analysis.

At retrieval time, the reasoner must generate cues to retrieve cases that are relevant to the current circumstances, based on the reasoner's goals and understanding of the new situation. The retrieval process begins with the reasoner performing 'situation assessment' to identify indices that might have been used to index relevant stored cases. The reasoner then uses those indices as a probe into memory, and assesses the results of retrieval to decide which retrieved cases are most appropriate.

The extent to which a reasoner is able (and willing) to analyze the new circumstances determines the quality of the probe into memory. By using the right probes, a reasoner may be able to retrieve useful cases that were encoded in quite different contexts. Seeing a corporate lawsuit as a contest of strength, for example, might lead to retrieving sports cases that suggest useful competitive strategies. The more creative a reasoner is at interpreting a situation, the more likely the reasoner will be to find new and surprising connections.

The indices used to describe cases make up the reasoner's 'indexing vocabulary'. In general, indexing vocabularies have two parts: a set of features, and a set of associated values. For straightforward retrievals, indexing features should be those that are naturally articulated in the process of reasoning or doing the tasks for which the case will be relevant. Indices may be specific to a domain, as in 'has pneumonia', which (in conjunction with other indices, such as 'is a toddler') might be used to retrieve cases suggesting treatments. Indices may also be abstract, applying to many domains. For example, the index 'bad balance interaction' can apply to domains ranging from medicine (when the dosage of two medicines was not properly coordinated), to cooking (when an addition to a soufflé results in too much liquid for the quantity of eggs), to manufacturing (when the force of a press for stamping metal was not commensurate with the size of the piece to be stamped).

CBR research uses two methods to determine the features and values used to index cases. The 'functional' approach examines a corpus of cases and the tasks that must be supported, considering what each case can be used for and the ways in which it needs to be described to make it available. The 'reminding' approach examines the kinds of remindings that are natural among humans who do the designated task, looking for similarities between new situations and retrieved cases, but not with cases that are not retrieved. Both approaches provide testable hypotheses for the kinds of indices that are important in human retrieval.

Once indices have been selected, cases are retrieved according to the similarities of their relevant features. One model of this process is 'nearest-neighbor' retrieval, in which the similarity between cases is based on a weighted sum of the distances between their corresponding features. (For example, a case-based reasoner seeking cases to suggest medical treatments might weight symptoms heavily but give no weight to the patients' names.) If the features considered for retrieval of two cases C and C' are each represented as a set of N attributes $\{a_i\}$ and $\{a'_i\}$, and each attribute's importance is assigned a nonnegative weight w_i, and d_i is a distance metric over the set of possible attributes of the i^{th} feature, then the degree of dissimilarity between C and C' may be defined as $d(C, C') = \sqrt{\sum_{i=1}^{N} w_i \times d_i(a_i, a'_i)^2}$.

For large sets of cases, another model of retrieval, 'discrimination trees', provides efficient retrieval at the cost of decreased flexibility for changing the features used for retrieval. Discrimination trees model case access as a traversal of a tree representing an ordered sequence of questions about indexing features.

ADAPTING CASES TO NEW NEEDS

After a relevant case has been retrieved, it is often necessary to perform additional reasoning to apply its lessons to the new circumstances. In legal case-based reasoning, for example, it may be necessary to show why the desired case is more relevant than the cases that opposing lawyers have proposed to support different conclusions (Rissland, 1998). In problem-solving CBR, the retrieved solution must be adapted to provide a suitable solution for the new problems. For example, a doctor may need to replace the drug used in a treatment with an alternative drug, in order to avoid aggravating a patient's allergy, or may need to perform parameter adjustment to adapt the dose for a previous adult patient to suit a child.

Research into case adaptation focuses on understanding the types of adaptations used by case-based reasoners and the knowledge those adaptations require. The operations may range from limited substitutions to complex structural changes; the knowledge required may include specialized knowledge, general heuristics, or domain models. The case adaptation process is often modeled by rule-based production systems using fixed adaptation knowledge, but researchers are also investigating how this knowledge may itself be learned by case-based reasoning or other methods. Flexibility, and sometimes even creativity, results from retrieving cases in unexpected contexts and adapting them to new needs.

WHAT TO REMEMBER, WHAT TO FORGET, AND WHAT TO GENERALIZE

Case storage presents a range of challenges. First, as discussed above, enabling access to cases in apparently disimilar situations depends on indexing cases to make them accessible whenever they are relevant. Thus careful case analysis is needed to

predict the relevant features for indexing. For example, if a newly-prescribed drug has the side effect of interfering with a drug previously prescribed to a patient, it may be appropriate to index the case for the doctor's response both by the specific drugs involved and by very general indices such as 'side effect disables needed condition'. This would enable the case to be retrieved in situations with very different specifics, but where the case's solution (for example, separating the interacting elements by administering them at different times) might still be applicable.

Secondly, it may not be beneficial to retain all new cases. As a case-based reasoner's memory grows, the reasoner has access to more cases, increasing the likelihood of being able to retrieve very similar cases, but also increasing retrieval costs, possibly negating the efficiency benefits from additional cases, or exceeding available storage. These problems are being studied in research on 'case-base maintenance', which focuses on how to determine which cases to remember and which to forget, as well as how to update the case library and other system knowledge sources over time.

Thirdly, when several cases are indexed in the same way and all predict the same solution, it may be useful to generalize from them. One model of the interaction between abstract and specific knowledge is proposed by the 'memory organization packages' (MOP) model proposed in dynamic memory theory (Schank, 1999). In this model, episodes in memory are organized by hierarchical MOPs at different levels of generality that collect constituent 'scenes'. Scenes are lower-level components which normally describe events that take place in a single location, with a single purpose, and in a single time interval, and can be shared by a number of MOPs (for example, the act of payment after a professional office visit corresponds to a 'pay' scene shared by the MOPs for 'doctor visits' and 'dentist visits'). Specific case details are stored under the scenes to which they refer, so that the only case information that must be stored is deviations from the routine expectations for the scene. When a case must be retrieved, its details are reconstructed; when several cases share details, the shared details provide the basis for a generalized MOP capturing the common features.

RELATIONSHIP TO OTHER FORMS OF LEARNING

The learning in case-based reasoning contrasts with both inductive and theory-driven models of learning. In contrast to traditional symbolic and neural network approaches to inductive learning, which define concepts by generalizations and discard the examples on which the generalizations are based, case-based reasoners retain and reason from specific prior cases. This enables new information take effect immediately – there is no need to 'retrain' the reasoner – and enables effective incremental learning. No matter how few cases are contained in the case library, performance on the problems those cases address will be correct, and as soon as a case has been stored by a CBR system, that case is available for use.

Theory-driven learning (often called explanation-based learning) uses general rules about a domain to form an explanation of the relevant features, in order to use that explanation to guide generalization. Unlike inductive generalization, explanation-based generalization can reliably learn from single examples. However, it cannot learn without a domain theory, and it performs all generalization at storage time, when it may not be clear which generalizations will be useful, or whether any generalizations will be useful. Instead of generalizing cases at storage time, CBR adapts cases in response to new problems, and makes only the changes needed to address those problems. The analysis of a current experience still plays an important role in the quality of learning, however, by determining the quality of the indexing for a case (to ensure that it will be retrieved in appropriate future situations) and for understanding the current situation and execution feedback (to ensure that the right lessons are stored).

IMPLICATIONS FOR HUMAN–MACHINE SYSTEMS AND THE LEARNING SCIENCES

Psychological studies show that case-based reasoning is a natural reasoning process for people, and computational models of the CBR process illuminate specific requirements for any case-based reasoning process to be effective. An active area of current research and applications is to bring these two strands together: to design interactive computer systems that exploit understanding of the CBR process to provide more effective support for human case-based reasoning. Two important applications are interactive case-based aiding systems, which support human problem solving by making relevant cases available to users to supplement their own experiences, and educational systems, in which computational models of CBR provide concrete suggestions about what makes a good problem, how courses of instruction should

be designed, and the kinds of resources that should be made available to student learners.

Case-based aiding systems support problem solving by providing external memories of solutions and warnings from stored prior experiences. Many case-based aiding systems have been developed to aid the design process by providing access to suggestions and lessons from prior designs, and many systems are in use to aid diagnosis and other tasks. For example, in aeronautical engineering, case-based aiding systems have been developed to guide the positioning of composite parts in curing ovens and to help with diagnosis of jet engine problems. ASK systems (Ferguson *et al*., 1992) enable users to browse a case library of experts' stories, in a rather conversational way. For example, one ASK system has been developed to capture the expertise of military transportation planners, providing a resource of first-person stories that current planners can apply to their problems. When the user browses a story, the system presents a set of specific questions that the story might raise, such as questions about the context of the story, results, or warnings arising from it. Each question provides a link for interested users to follow to access cases relevant to that question.

Goal-based scenarios (GBS) (Schank *et al*., 1993/1994) apply the lessons of CBR research to the design of environments for skill-centered learning by doing. Students learn by case acquisition as they perform activities in pursuit of compelling goals. In addition, CBR methods are used to retrieve anecdotes of others' prior experiences to present to learners at useful times. For example, in the GBS 'Broadcast News', students play the role of a newsreader, editing a report on some current or past event to make sure that it is understandable to the viewing public. They practice writing skills, particularly describing events and people in ways that others can understand and summarizing to make sure the most important points are made. Meanwhile they are learning about current events or history. As they are working, they may ask for help with a specific editing problem, and a short video will play describing how some expert newsreader or editor dealt with a similar editorial decision. When they have finished, they present their stories to the rest of the class, and the class discusses both the content being presented and the skills involved in getting to that presentation.

'Learning by design' (LBD) (Hmelo *et al*., 2000; Kolodner *et al*., 1998) focuses on learning scientific theory and practices (skills) in the context of design challenges. Middle-school students generate questions and conduct investigations in the context of a design challenge (for example, to design a propulsion system that will propel a miniature vehicle over several hills (for learning about forces and motion), or to design a way to save a basketball court from erosion by dirt from an adjacent hill (for learning about managing erosion)). LBD provides a set of classroom rituals, practices, and activities, informed by case-based reasoning's cognitive model (Schank, 1999), that work together to enable deep and transferable learning. For example, the 'gallery walk' is a time when students share experimental results or experiences, providing each other with a variety of examples of how some scientific law applies and is applied in the world. The discussions that accompany gallery walks help students to identify good and bad examples and ways of improving their practices and designs, and to explain the limitations of their designs. These discussions also provide an opportunity to draw out scientific principles from their experiences. Pilot and field tests have shown that LBD students learn scientific theory as well as or better than other students and that, in addition, they learn the value of collaboration and gain skill at many of the practices associated with being a scientist.

Learning by design and goal-based scenarios use the principles of CBR to support and refine students' own use of case-based reasoning. Both are designed to ensure that learners have the kinds of experience that can lead to the given learning objectives, and provide activities orchestrated so that attempting them, and interpreting the resulting experiences, will lead to learning the right lessons and selecting good indices for future recall and transfer. Likewise, because failure motivates explanations and focuses a reasoner on what he or she needs to learn, both LBD and GBS place students in circumstances in which they can fail safely and recognize failure, by giving them goals to achieve, helping them to generate ideas about how to achieve those goals, and giving them the opportunity to try out their ideas and experience the results. Both models promote iterative refinement of solutions – moving gradually towards better solutions and understanding of the underlying theory – through explanation of imperfect solutions, and refinement of both beliefs and solutions as a result of those explanations. LBD goes further by orchestrating sequences of experiences, sequencing challenges so that students refer back to and attempt to reuse what they learned in earlier challenges.

The goal of helping learners become better case-based reasoners provides guidance for teachers in

an LBD or GBS environment, or indeed in any environment where students are learning from experiences. CBR suggests that teachers should help students interpret their experiences and store them as well-articulated and well-indexed cases in their memories: helping learners to make connections between their intentions, plans, outcomes, and explanations; helping them to extract lessons from their experiences and associated applicability conditions; making sure they get timely feedback on the decisions they make; helping them explain their mistakes or poor predictions; and helping them rethink and revise the encodings of beliefs that were responsible for those mistakes. CBR suggests, also, that teachers should help learners learn to use cases effectively as they reason: providing students with opportunities to practice retrieving applicable cases from their memories, to judge which of several potential cases might be applicable in a new situation, and to merge, adapt, and apply lessons in new situations. Teachers should also help them to learn how to learn: providing a scaffolding to help students notice similarities and general rules by comparing cases, helping them discern applicability conditions for lessons they extract from cases, helping them to make abstractions for use in more sophisticated reasoning, and helping them to be aware of the reasoning they are doing.

SUMMARY

Cases provide a rich source of information at an operational level, and can be learned and reapplied in a wide range of circumstances, both to improve reasoning efficiency and to enable effective reasoning in domains that are imperfectly understood. Case-based reasoning research in cognitive science develops computational models of the memory, analogical reasoning, and learning processes required to reason and learn from cases. These models provide testable hypotheses about human reasoning and learning, and illuminate the constraints on the case-based reasoning process relating to information processing. This research has led to useful artificial intelligence technologies, often inspired by human case-based reasoning and sometimes exploiting knowledge about human CBR to assist humans' own processes of learning from cases. In educational applications, the CBR approach redefines 'transfer' as spontaneous reminding and use of experiences in new situations. This process can be improved by structuring learning environments to support and assist students' analysis and assimilation of cases and to encourage students to compare their cases with the cases of others, by sharing and examining anecdotes about prior experiences. (*See* **Analogy-making, Computational Models of**; **Machine Learning**; **Expert Systems**; **Natural Language Processing: Models of Roger Schank and his Students**)

References

Ferguson W, Bareiss R, Osgood R and Birnbaum L (1992) ASK systems: an approach to the realization of story-based teachers. *Journal of the Learning Sciences* **2**(1): 95–134.

Gentner D, Holyoak K and Kokinov B (2001) *The Analogical Mind: Perspectives From Cognitive Science*. Cambridge, MA: MIT Press.

Hmelo C, Holton D and Kolodner J (2000) Designing to learn about complex systems. *Journal of the Learning Sciences* **9**(3): 247–298.

Hunter KM (1986) 'There was this one guy': anecdotes in medicine. *Biology in Medicine* **29**: 619–630. [Reprinted in: Hunter KM (ed.) (1991) *Doctors' Stories: The Narrative Structure of Medical Knowledge*. Princeton, NJ: Princeton University Press.]

Kolodner J (1993) *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Kolodner J, Crismond D, Gray J, Holbrook J and Puntembakar S (1998) Learning by design from theory to practice. In: *Proceedings of the Third International Conference on the Learning Sciences*, pp. 16–22. Charlottesville, VA: AACE Press.

Leake D (1996) CBR in context: the present and future. In: Leake D (ed.) *Case-Based Reasoning: Experiences, Lessons, and Future Directions*, pp. 3–30. Menlo Park, CA: AAAI Press/MIT Press.

Rissland E (1998) Legal reasoning. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 722–733. Oxford: Blackwell.

Schank R (1999) *Dynamic Memory Revisited*. Cambridge, UK: Cambridge University Press.

Schank R, Fano A, Bell B and Jona M (1993/1994) The design of goal-based scenarios. *Journal of the Learning Sciences* **34**: 305–345.

Further Reading

Kolodner J (1997) Educational implications of analogy: a view from case-based reasoning. *American Psychologist* **52**(1): 57–66.

Leake D (ed.) (1996) *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, CA: AAAI Press/MIT Press.

Leake D (1998) Cognition as case-based reasoning. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 465–476. Oxford: Blackwell.

Pu P and Maher ML (1997) *Issues and Applications of Case-Based Reasoning to Design*. Mahwah, NJ: Erlbaum.

Riesbeck C and Schank R (1999) *Inside Case-Based Reasoning*. Hillsdale, NJ: Erlbaum.

Schank R and Cleary C (1995) *Engines for Education*. Hillsdale, NJ: Erlbaum.

Watson I (1997) *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. San Mateo, CA: Morgan Kaufmann.

Machine Learning

Intermediate article

Thomas G Dietterich, Oregon State University, Corvallis, Oregon, USA

CONTENTS

Machine learning is the study of methods for programming computers to learn.

INTRODUCTION

Machine learning is the study of methods for programming computers to learn. Computers are applied to a wide range of tasks, and for most of these it is relatively easy for programmers to design and implement the necessary software. However, there are many tasks for which this is difficult or impossible. For these tasks, machine learning provides a way to construct successful computer systems. These can be divided into four general categories.

First, there are problems for which there exist no human experts. For example, in modern automated manufacturing facilities, there is a need to predict machine failures before they occur by analyzing sensor readings. Because the machines are new, there are no human experts who can be interviewed by a programmer to provide the knowledge necessary to build a computer system. A machine learning system can study recorded data and subsequent machine failures and learn prediction rules.

Second, there are problems where human experts exist, but where they are unable to explain their expertise. This is the case in many perceptual tasks, such as speech recognition, hand-writing recognition, and natural language understanding. Virtually all humans exhibit expert-level abilities in these tasks, but none of them can describe the detailed steps that they follow as they perform them. Fortunately, humans can provide machines with examples of the inputs and correct outputs for these tasks, so machine learning algorithms can learn to map the inputs to the outputs.

Third, there are problems where phenomena are changing rapidly. In finance, for example, people would like to predict the future behavior of the stock market, of consumer purchases, or of exchange rates. These behaviors change frequently, so that even if a programmer could construct a good predictive computer program, it would need to be rewritten frequently. A learning program can relieve the programmer of this burden by constantly modifying and tuning a set of learned prediction rules.

Fourth, there are applications that need to be customized for each computer user separately. Consider, for example, a program to filter unwanted electronic mail messages. Different users will need different filters. It is unreasonable to expect each user to program his or her own rules, and it is infeasible to provide every user with a software engineer to keep the rules up-to-date. A machine learning system can learn which mail messages the user rejects and maintain the filtering rules automatically.

Machine learning addresses many of the same research questions as the fields of statistics, data mining, and psychology, but with differences of emphasis. Statistics focuses on understanding the phenomena that have generated the data, often with the goal of testing different hypotheses about those phenomena. Data mining seeks to find patterns in the data that are understandable by people. Psychological studies of human learning aspire to understand the mechanisms underlying the various learning behaviors exhibited by people (concept learning, skill acquisition, strategy change, etc.). In contrast, machine learning is primarily concerned with the accuracy and effectiveness of the resulting computer system. To illustrate this, consider the different questions that might be asked about speech data. A machine learning approach focuses on building an accurate and efficient speech recognition system. A statistician might collaborate with a psychologist to test

hypotheses about the mechanisms underlying speech recognition. A data mining approach might look for patterns in speech data that could be applied to group speakers according to age, sex, or level of education.

ANALYTICAL AND EMPIRICAL LEARNING TASKS

Learning tasks can be classified along many different dimensions. One important dimension is the distinction between empirical and analytical learning. Empirical learning is learning that relies on some form of external experience, while analytical learning requires no external inputs. Consider, for example, the problem of learning to play tic-tac-toe (noughts and crosses). Suppose a programmer has provided an encoding of the rules for the game in the form of a function that indicates whether proposed moves are legal or illegal and another function that indicates whether the game is won, lost, or tied. Given these two functions, it is easy to write a computer program that repeatedly plays games of tic-tac-toe against itself. Suppose that this program remembers every board position that it encounters. For every final board position (i.e. where the game is won, lost, or tied), it remembers the outcome. As it plays many games, it can mark a board position as a losing position if every move made from that position leads to a winning position for the opponent. Similarly, it can mark a board position as a winning position if there exists a move from that position that leads to a losing position for the opponent. If it plays enough games, it can eventually determine all of the winning and losing positions and play perfect tic-tac-toe. This is a form of analytical learning because no external input is needed. The program is able to improve its performance just by analyzing the problem.

In contrast, consider a program that must learn the *rules* for tic-tac-toe. It generates possible moves and a teacher indicates which of them are legal and which are illegal as well as which positions are won, lost, or tied. The program can remember this experience. After it has visited every possible position and tried every possible move, it will have complete knowledge of the rules of the game (although it may guess them long before that point). This is empirical learning, because the program could not infer the rules of the game analytically – it must interact with a teacher to learn them.

The dividing line between empirical and analytical learning can be blurred. Consider a program like the first one that knows the rules of the game. However, instead of playing against itself, it plays against a human opponent. It still remembers all of the positions it has ever visited, and it still marks them as won, lost, or tied based on its knowledge of the rules. This program is likely to play better tic-tac-toe sooner, because the board positions that it visits will be ones that arise when playing against a knowledgeable player (rather than random positions encountered while playing against itself). So *during the learning process*, this program will perform better. Nonetheless, the program did not *require* the external input, because it could have inferred everything analytically.

The solution to this puzzle is to consider that the overall *learning task* is an analytical task, but that the program solves the task empirically. Furthermore, the task of playing well against a human opponent *during the learning process* is an empirical learning task, because the program needs to know which game positions are likely to arise in human games.

This may not seem like a significant issue with tic-tac-toe. But in chess, for example, it makes a huge difference. Given the rules of the game, learning to play optimal chess is an analytical learning task, but the analysis is computationally infeasible, so methods that include some empirical component must be employed instead. From a cognitive science perspective, the difference is also important. People frequently confront learning tasks which could be solved analytically, but they cannot (or choose not to) solve them this way. Instead, they rely on empirical methods.

The remainder of this article is divided into five parts. The first four parts are devoted to empirical learning. First we discuss the fundamental questions and methods in supervised learning. Then we consider more complex supervised learning problems involving sequential and spatial data. The third section is devoted to unsupervised learning problems, and the fourth section discusses reinforcement learning for sequential decision-making. The article concludes with a review of methods for analytical learning.

FUNDAMENTALS OF SUPERVISED LEARNING

Let us begin by considering the simplest machine learning task: *supervised learning* for classification. Suppose we wish to develop a computer program that, when given a picture of a person, can determine whether the person is male or female. Such a program is called a *classifier*, because it assigns a class (i.e. male or female) to an object (i.e. a photograph). The task of supervised learning is to

construct a classifier given a set of classified training examples – in this case, example photographs along with the correct classes.

The key challenge for supervised learning is the problem of *generalization*: after analyzing only a (usually small) sample of photographs, the learning system should output a classifier that works well on all possible photographs.

A pair consisting of an object and its associated class is called a *labeled example*. The set of labeled examples provided to the learning algorithm is called the *training set*. Suppose we provide a training set to a learning algorithm and it outputs a classifier. How can we evaluate the quality of this classifier? The usual approach is to employ a second set of labeled examples called the *test set*. We measure the percentage of test examples correctly classified (called the *classification rate*) or the percentage of test examples misclassified (the *misclassification rate*).

The reason we employ a separate test set is that most learned classifiers will be very accurate on the training examples. Indeed, a classifier that simply memorized the training examples would be able to classify them perfectly. We want to test the ability of the learned classifier to generalize to new data points.

Note that this approach of measuring the classification rate assumes that each classification decision is independent and that each classification decision is equally important. These assumptions are often violated.

The independence assumption could be violated if there is some temporal dependence in the data. Suppose, for example, that the photographs were taken of students in classrooms. Some classes (e.g. early childhood development) primarily contain girls, other classes (e.g. car repair) primarily contain boys. If a classifier knew that the data consisted of batches, it could achieve higher accuracy by trying to identify the point at which one batch ends and another begins. Then within each batch of photographs, it could classify all of the objects into a single class (e.g. based on a majority vote of its guesses on the individual photographs). These kinds of temporal dependencies arise frequently. For example, a doctor seeing patients in a clinic knows that contagious illnesses tend to come in waves. Hence, after seeing several consecutive patients with the flu, the doctor is more likely to classify the next patient as having the flu too, even if that patient's symptoms are not as clearcut as the symptoms of the previous patients.

The assumption of equal importance could be violated if there are different costs or risks associated with different misclassification errors. Suppose the classifier must decide whether a patient has cancer based on some laboratory measurements. There are two kinds of errors. A *false positive* error occurs when the classifier classifies a healthy patient as having cancer. A *false negative* error occurs when the classifier classifies a person with cancer as being healthy. Typically false negatives are more costly than false positives, so we might want the learning algorithm to prefer classifiers that make fewer false negative errors, even if they make more false positives as a result.

The term supervised learning includes not only learning classifiers but also learning functions that predict numerical values. For example, given a photograph of a person, we might want to predict the person's age, height, and weight. This task is usually called *regression*. In this case, each labeled training example is a pair of an object and the associated numerical value. The quality of a learned prediction function is usually measured as the square of the difference between the predicted value and the true value, although sometimes the absolute value of this difference is measured instead.

An Example Learning Algorithm: Learning Decision Trees

There are many different learning algorithms that have been developed for supervised classification and regression. These can be grouped according to the formalism they employ for representing the learned classifier or predictor: decision trees, decision rules, neural networks, linear discriminant functions, Bayesian networks, support vector machines, and nearest-neighbor methods. Many of these algorithms are described in other articles in this encyclopedia. Here, we will present a topdown algorithm for learning decision trees, since this is one of the most versatile, most efficient, and most popular machine learning algorithms. (*See* **Connectionism**; **Backpropagation**; **Bayesian Belief Networks**; **Perceptron**; **Pattern Recognition, Statistical**; **Learning through Case Analysis**)

A decision tree is a branching structure as shown in Figure 1. The tree consists of nodes and leaves. The *root node* is at the top of the diagram, and the leaves at the bottom. Each node tests the value of some *feature* of an example, and each leaf assigns a class label to the example. This tree was constructed by analyzing 670 labeled examples of breast cancer biopsies. Each biopsy is represented by 9 features such as Clump Thickness (CT), Uniformity of Cell Size (US), and Uniformity of Cell Shape (USh). To

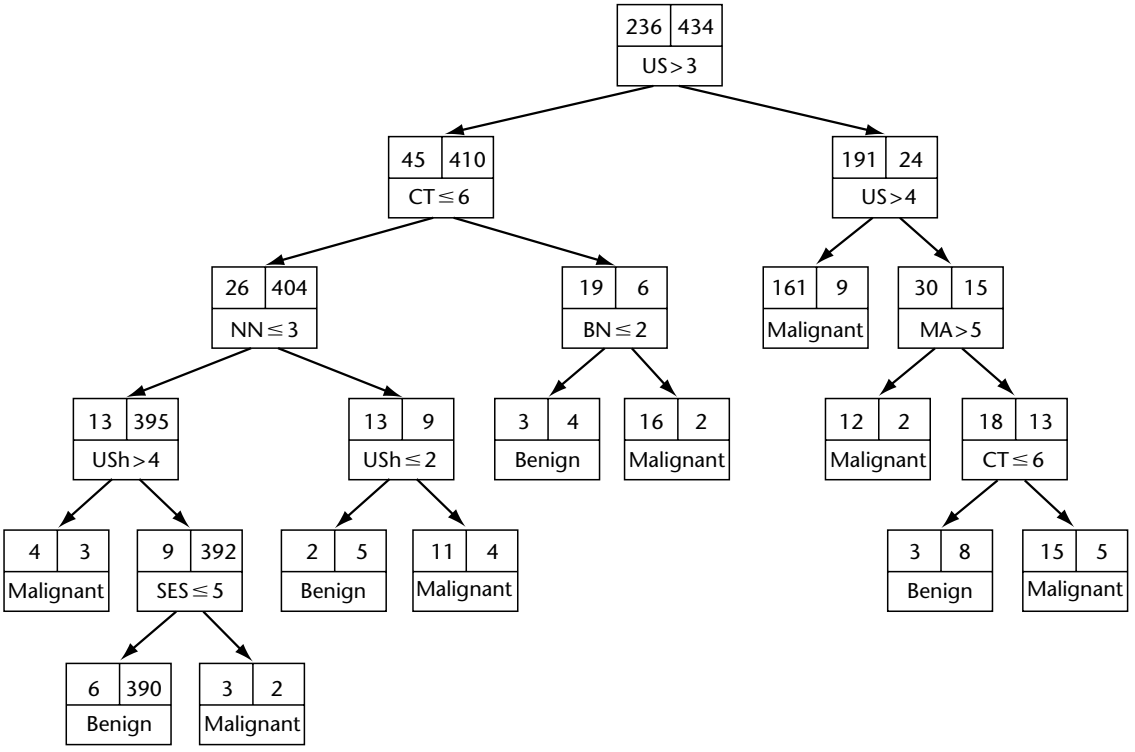


Figure 1. Decision tree for diagnosing breast cancer. US = Uniformity of Cell Size; CT = Clump Thickness; NN = Normal Nucleoli; USh = Uniformity of Cell Shape; SES = Single Epithelial Cell Size; BN = Bare Nuclei; MA = Marginal Adhesion.

understand how the decision tree works, suppose we have a biopsy example with US = 5, CT = 7, and BN = 2. To classify this example, the decision tree first tests if US > 3, which is true. Whenever the test in a node is true, control follows the left outgoing arrow; otherwise, it follows the right outgoing arrow. In this case, the next test is CT ≤ 6 which is false, so control follows the right arrow to the test BN ≤ 2. This is true, so control follows the left arrow to a leaf node which assigns the class 'Benign' to the biopsy.

The numbers in each node indicate the number of Malignant and Benign training examples that 'reached' that node during the learning process. At the root, the 670 training examples comprised 236 Malignant cases and 434 Benign cases. The decision tree is constructed top-down by repeatedly choosing a feature (e.g. US) and a threshold (e.g. 3) to test. Different algorithms employ different heuristics, but all of these heuristics try to find the feature and threshold that are most predictive of the class label. A perfect test would send all of the Benign examples to one branch and all of the Malignant examples to the other branch. The test US > 3 is not perfect, but it is still very good: the left branch receives 410 of the 434 Benign cases and only 45 of the 236 Malignant ones, while the right branch receives 191 of the 236 Malignant cases and only 24 of the Benign ones. After selecting this test, the algorithm splits the training examples according to the test. This gives it $45 + 410 = 455$ examples on the left branch and $191 + 24 = 215$ on the right. It now repeats the same process of choosing a predictive feature and threshold, and splitting the data until a termination rule halts the splitting process. At that point, a leaf is created whose class label is the label of the majority of the training examples that reached the leaf.

One advantage of decision trees is that, if they are not too large, they can be interpreted by humans. This can be useful both for gaining insight into the data and also for validating the reasonableness of the learned tree.

The Triple Trade-off in Empirical Learning

All empirical learning algorithms must contend with a trade-off among three factors: (1) the size or complexity of the learned classifier, (2) the amount of training data, and (3) the generalization accuracy on new examples. Specifically, the

generalization accuracy on new examples will usually increase as the amount of training data increases. As the complexity of the learned classifier increases, the generalization accuracy first rises and then falls. These trade-offs are illustrated in Figure 2. The different curves correspond to different amounts of training data. As more data are available, the generalization accuracy reaches a higher level before eventually dropping. In addition, this higher level corresponds to increasingly more complex classifiers.

The relationship between generalization accuracy and the amount of training data is fairly intuitive: the more training data given to the learning algorithm, the more evidence the algorithm has about the classification problem. In the limit, the data would contain every possible example, so the algorithm would know the correct label of every possible example, and it would generalize perfectly.

In contrast, the relationship between generalization accuracy and the complexity of the learned structure is less obvious. To understand it, consider what would happen if we allowed the decision tree to grow extremely large. An advantage of such a large tree is that it can be proved that if the tree becomes large enough, it can represent *any* classifier. We say that such trees have *low bias*.

Unfortunately, however, such a very large tree would typically end up having only a few training examples in each leaf. As a result, the choice of the class label in each leaf would be based on just those few examples, which is a precarious situation. If there is any noise in the process of measuring and labeling those training examples, then the class label could be in error. The resulting classifier is said to have *high variance*, because a slight change in the training examples can lead to changes in the classification decisions. Such a decision tree has merely memorized the training data, and although it will be very accurate on the training data, it will usually generalize very poorly. We say that it has 'overfit' the training data.

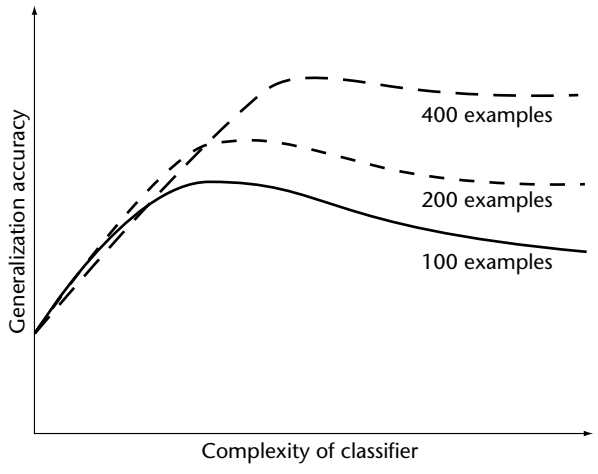


Figure 2. Generalization accuracy as a function of the complexity of the classifier, for various amounts of training data.

At the other extreme, suppose we consider a degenerate decision tree that contains only one decision node and two leaves. (These are known as 'decision stumps'.) The class label in each leaf is now based on hundreds of training examples, so the tree has *low variance*, because it would take a large change in the training data to cause a change in the classifier. However, such a simple decision tree might not be able to capture the full complexity of the data. For diagnosing breast cancer, for example, it is probably not sufficient to consider only one feature. Formally, such a classifier is said to have *high bias*, because its representational structure prevents it from representing the optimal classifier. Consequently, the classifier may also generalize poorly, and we say that it has 'underfit' the training data.

An intuitive way of thinking about this trade-off between bias and variance is the following. A learning algorithm faces a choice between a vast number of possible classifiers. When very few data are available, it does not have enough information to distinguish between all of these classifiers – many classifiers will appear to have identical accuracy on the training data and if it chooses randomly among such apparently good classifiers, this will result in high variance. It must reduce the number of possible classifiers (i.e. by reducing their complexity) until it does have enough data to discriminate among them. Unfortunately, this reduction will probably introduce bias, but it will reduce the variance.

In virtually every empirical learning algorithm, there are mechanisms that seek to match the complexity of the classifier to the complexity of the training data. In decision trees, for example, there are *pruning procedures* that remove branches from an overly large tree to reduce the risk of overfitting. In neural networks, support vector machines, and linear discriminant functions, there are *regularization methods* that place a numerical penalty on having large numerical weights. This turns out to be mathematically equivalent to limiting the complexity of the resulting classifiers.

Some learning algorithms, such as the naive Bayes method and the perceptron algorithm, are not able to adapt the complexity of the classifier. These algorithms only consider relatively simple

classifiers. As a result, on small training sets, they tend to perform fairly well, but as the amount of training data increases, their performance suffers, because they underfit the data (i.e. they are biased).

Prior Knowledge and Bias

Most machine learning algorithms make only very general and very weak assumptions about the nature of the training data. As a result, they typically require large amounts of training data to learn accurate classifiers. This problem can be solved by exploiting prior knowledge to eliminate from consideration classifiers that are not consistent with the prior knowledge. The resulting learning algorithms may be able to learn from very few training examples.

However, there is a risk to introducing prior knowledge. If that knowledge is incorrect, then it will eliminate all of the accurate classifiers from consideration by the learning algorithm. In short, prior knowledge introduces bias into the learning process, and it is important that this bias be correct.

SUPERVISED LEARNING FOR SEQUENCES, TIME SERIES, AND SPATIAL DATA

Now that we have discussed the basic supervised learning problem and the bias–variance trade-off, we turn our attention to more complex supervised learning tasks.

Consider the problem of speech recognition. A speech recognition system typically accepts as input a spoken sentence (e.g. 5 seconds of a sound signal) and produces as output the corresponding string of words. This involves many levels of processing, but at the lowest level, we can think of a sentence as a sequence of labeled examples. Each example consists of a 40 ms segment of speech (the object) along with a corresponding phoneme (the label). However, it would be a mistake to assume that these labeled examples are independent of each other, because there are strong sequential patterns relating adjacent phonemes. For example, the pair of phonemes /s/ /p/ (as in the English words 'spill' and 'spin') is much more common than the pair /s/ /b/ (which almost never appears). Hence, a speech recognition system has the opportunity to learn not only how to relate the speech signal to the phonemes, but also how to relate the phonemes to each other. The hidden Markov model (HMM) is an example of a classifier that can learn both of these kinds of information. (*See* **Speech Recognition, Automatic**; **Speech Perception and Recognition, Theories and Models of**)

A similar problem arises in time series analysis. Suppose we wish to predict the El Niño phenomenon, which can be measured by the temperature of the sea surface in the equatorial Pacific Ocean. Imagine that we have measurements of the temperature every month for the past 20 years. We can view this as a set of labeled training examples. Each example is a pair of temperatures from two consecutive months, and the goal is to learn a function for predicting the temperature next month from the temperature in the current month. Again it is a mistake to treat these examples as independent. The relationship between adjacent months is similar to the relationship between adjacent phonemes in speech recognition. However, unlike in speech recognition, we must make a prediction every month about what the next month's temperature will be. This would be like trying to predict the next word in the sentence based on the previous word.

Spatial data present learning tasks similar to sequential data, but in two dimensions. For example, a typical spatial task is to predict the type of land cover (trees, grasslands, lakes, etc.) on the ground based on satellite photographs. Training data consist of photographs in which each pixel has been labeled by its land cover type. Methods such as Markov random fields can be applied to capture the relationships between nearby pixels.

Supervised Learning for Complex Objects

So far we have discussed the task of classifying single objects and the task of classifying a one- or two-dimensional array of objects. There is a third task that is intermediate between these: the task of classifying complex objects. For example, consider the problem of deciding whether a credit card has been stolen. The 'object' in this case is a *sequence* of credit card transactions, but the class label (stolen or not stolen) is attached to the entire sequence, not to each individual transaction. In this case, we wish to analyze the entire sequence to decide whether it provides evidence that the card is stolen.

There are three ways to approach this problem. The first method converts it into a simple supervised learning problem by extracting a set of features from the sequence. For example, we might compute the average, minimum, and maximum dollar amounts of the transactions, the variance of the transactions, the number of transactions per day, the geographical distribution of the transactions, and so on. These features summarize the

variable length sequence as a fixed length feature vector, which we can then give as input to a standard supervised learning algorithm.

The second method is to convert the problem into the problem of classifying labeled sequences of objects. On the training data, we assign a label to each transaction indicating whether it was legitimate or not. Then we train a classifier for classifying individual transactions. Finally, to decide whether a new sequence of transactions indicates fraud, we apply our learned classifier to the entire sequence and then make a decision based on the number of fraudulent transactions it identifies.

The third method is to learn explicit models of fraudulent and non-fraudulent sequences. For example, we might learn a hidden Markov model that describes the fraudulent training sequences and another HMM to describe the non-fraudulent sequences. To classify a new sequence, we compute the likelihood that each of these two models could have generated the new sequence and choose the class label of the more likely model.

UNSUPERVISED LEARNING

The term unsupervised learning is employed to describe a wide range of different learning tasks. As the name implies, these tasks analyze a given set of objects that do not have attached class labels. In this section, we will describe five unsupervised learning tasks.

Understanding and Visualization

Given a large collection of objects, we often want to be able to understand these objects and visualize their relationships. Consider, for example, the vast diversity of living things on earth. Linnaeus devoted much of his life to arranging living organisms into a hierarchy of classes with the goal of arranging similar organisms together at all levels of the hierarchy.

Many unsupervised learning algorithms create similar hierarchical arrangements. The task of *hierarchical clustering* is to arrange a set of objects into a hierarchy so that similar objects are grouped together. A standard approach is to define a measure of the similarity between any two objects and then seek clusters of objects which are more similar to each other than they are to the objects in other clusters. *Non-hierarchical clustering* seeks to partition the data into some number of disjoint clusters.

A second approach to understanding and visualizing data is to arrange the objects in a low-dimensional space (e.g. in a 2-dimensional plane) so that similar objects are located nearby each other. Suppose, for example, that the objects are represented by 6 real-valued attributes: height, width, length, weight, color, and density. We can measure the similarity of any two objects by their Euclidean distance in this 6-dimensional space. We wish to assign each object two new dimensions (call them x and y) such that the Euclidean distance between the objects in this 2-dimensional space is proportional to their Euclidean distance in the original 6-dimensional space. We can then plot each object as a point in the 2-dimensional plane and visually see which objects are similar.

Density Estimation and Anomaly Detection

A second unsupervised learning task is density estimation (and the closely related task of anomaly detection). Given a set of objects, $\{e_1, e_2, \ldots, e_n\}$, we can imagine that these objects constitute a random sample from some underlying probability distribution $P(e)$. The task of density estimation is to learn the definition of this probability density function P.

A common application of density estimation is to identify anomalies or outliers. These are objects that do not belong to the underlying probability density. For example, one approach to detecting fraudulent credit card transactions is to collect a sample of legal credit card transactions and learn a probability density $P(t)$ for the probability of transaction t. Then, given a new transaction t', if $P(t')$ is very small, this indicates that t' is unusual and should be brought to the attention of the fraud department. In manufacturing, one quality control procedure is to raise an alarm whenever an anomalous object is produced by the manufacturing process.

Object Completion

People have an amazing ability to complete a fragmentary description of an object or situation. For example, in natural language understanding, if we read the sentences, 'Fred went to the market. He found some milk on the shelf, paid for it, and left,' we can fill in many events that were not mentioned. For example, we are quite confident that Fred picked up the milk from the shelf and took it to the cash register. We also believe that Fred took the milk with him when he left the market. We can complete this description because we know about 'typical' shopping episodes.

Similarly, suppose we see the front bumper and wheels of a car visible around the corner of a building. We can predict very accurately what the rest of the car looks like, even though it is hidden from view.

Object completion involves predicting the missing parts of an object given a partial description of the object. Both clustering and density estimation methods can be applied to perform object completion. The partial description of the object can be used to select the most similar cluster, and then the object can be completed by analyzing the other objects in that cluster. Similarly, a learned probability density $P(x_1, x_2)$ can be used to compute the most likely values of x_2 given the observed values of x_1. A third approach to object completion is to apply a supervised learning algorithm to predict each attribute of an object given different subsets of the remaining attributes.

Information Retrieval

A fourth unsupervised learning task is to retrieve relevant objects (documents, images, finger prints) from a large collection of objects. Information retrieval systems are typically given a partial description of an object, and they use this partial description to identify the K most similar objects in the collection. In other cases, a few examples of complete objects may be given, and again the goal is to retrieve the K most similar objects.

Clustering methods can be applied to this problem. Given partial or complete descriptions of objects, the most similar cluster can be identified. Then the K most similar objects can be extracted from that cluster.

Data Compression

There are many situations in which we do not want to store or transmit fully detailed descriptions of objects. Each image taken by a digital camera, for example, can require 3 megabytes to store. By applying image compression techniques, such images can often be reduced to 50 kilobytes (a 60-fold reduction) without noticeably degrading the picture. Data compression involves identifying and removing the irrelevant aspects of data (or equivalently, identifying and retaining the essential aspects of data). Most data compression methods work by identifying commonly occurring subimages or substrings and storing them in a 'dictionary'. Each occurrence of such a substring or subimage can then be replaced by a (much shorter) reference to the corresponding dictionary entry.

LEARNING FOR SEQUENTIAL DECISION-MAKING

In all learning systems, learning results in an improved ability to make decisions. In the supervised and unsupervised learning tasks we have discussed so far, the decisions made by the computer system after learning are non-sequential. That is, if the computer system makes a mistake on one decision, this has no bearing on subsequent decisions. Hence, if an optical character recognition system misreads the postal code on a package, this only causes that package to be sent to the wrong address. It does not have any effect on where the next package will be sent. Similarly, if a fraud detection system correctly identifies a stolen credit card for one customer, this has no effect on the cost (or benefit) of identifying the stolen credit cards of other customers.

In contrast, consider the problem of steering a car down a street. The driver must make a decision approximately once per second about how to turn the wheel to keep the car in its lane. Suppose the car is in the center of the lane but pointed slightly to the right. If the driver fails to correct by turning slightly to the left, then at the next time step, the car will move into the right part of the lane. If the driver again fails to correct, the car will start to leave the lane. In short, if the driver makes a mistake at one decision point, this affects the situation that he or she will confront at the next decision point. The hallmark of sequential decision-making is that the decision-maker must live with the consequences of his or her mistakes.

Sequential decision-making tasks arise in many domains where it is necessary to control a system (e.g. steering of robots, cars, and spacecraft; control of oil refineries, chemical plants, power plants, and factories; management of patients in intensive care). The system under control is also referred to as 'the environment'.

Many of these control problems can be modeled as *Markov decision problems (MDPs)*. In a Markov decision problem, the environment is said to have a current 'state' that completely summarizes the variables in the system (e.g. the position, velocity, and acceleration of a car). At each time step, the controller observes the state of the environment and must choose an action (e.g. steer left or steer right). The action is then executed, which may cause the environment to move to a new state. The controller then receives an immediate 'reward' which is a measure of the cost of the action and the desirability of the current state or the new state. For example, in steering a car, there might be a reward of zero for all actions as long as the car remains in

its lane. But the controller would receive a penalty whenever the car departed from the lane. The controller makes its decisions according to a control *policy*. The policy tells, for each state of the environment, what action should be executed. The optimal policy is one that maximizes the sum of the rewards received by the controller.

Reinforcement learning is the task of learning a control policy by interacting with an *unknown* environment. There are two main approaches to reinforcement learning: model-based and model-free methods.

In model-based reinforcement learning, the learner executes a control policy for the purpose of learning about the environment. Each time it executes an action a in state s and observes the resulting reward r and next state s', it collects a four-tuple $\langle s, a, r, s' \rangle$ that records the experience. After collecting a sufficient number of these four-tuples, the learner can learn a *probability transition function* $P(s'|s, a)$ and a *reward function* $R(s, a, s')$. The probability transition function says that if action a is executed in state s, then the environment will move to state s' with probability $P(s'|s, a)$. The reward function gives the average value of the reward that will be received when this happens: $R(s, a, s')$. Given these two functions, it is possible to apply *dynamic programming* algorithms to compute the optimal control policy.

Model-free reinforcement learning algorithms learn the policy directly by interacting with the environment without storing experience four-tuples or learning a model (i.e. without learning P and R). The best-known model-free algorithm is Q-learning, but researchers have also explored actor/critic and policy gradient algorithms that directly modify the control policy to improve its performance. (*See* **Reinforcement Learning: A Computational Perspective**)

In many control problems, the entire state of the environment cannot be observed at each time step. For example, when driving a car, the driver cannot simultaneously observe all of the other cars driving on the same street. Similarly, a robot controller can rarely observe the entire state of the world. The task of controlling such partially observable environments is known as a Partially Observable Markov Decision Problem (POMDP). As with MDPs, model-based and model-free methods have been developed.

Model-based methods for POMDPs must posit the existence of underlying (but unobservable states) in order to explain (and predict) the way that the observable parts of the state will change. (*See* **Markov Decision Processes, Learning of**)

SPEEDUP LEARNING

We now turn our attention to analytical learning. Because analytical learning does not involve interaction with an external source of data, analytical learning systems cannot learn knowledge with new empirical content. Instead, analytical learning focuses on improving the speed and reliability of the inferences and decisions that are performed by the computer. This is analogous in many ways to the process of skill acquisition in people.

Consider a computation that involves search. Examples include searching for good sequences of moves in chess, searching for good routes in a city, and searching for the right steps in a cooking recipe. The task of speedup learning is to remember and analyze past searches so that future problems can be solved more quickly and with little or no search.

The simplest form of speedup learning is called *caching* – replacing computation with memory. When the system performs a search, it stores the results of the search in memory. Later, it can retrieve information from memory rather than repeating the computation. For example, consider a person trying to bake a cake. There are many possible combinations of ingredients and many possible processing steps (e.g. stirring, sifting, cooking at various temperatures and for various amounts of time). A cook must search this space, trying various combinations, until a good cake is made. The cook can learn from this search by storing good combinations of ingredients and processing steps (e.g. in the form of a recipe written on a card). Then, when he or she needs to bake another cake, the recipe can be retrieved and followed.

Analogous methods have been applied to speed up computer game playing. Good sequences of moves can be found by searching through possible game situations. These sequences can then be stored and later retrieved to avoid repeating the search during future games. This search for good move sequences can be performed without playing any 'real' games against opponents.

A more interesting form of speedup learning is generalized caching – also known as *explanation-based learning*. Consider a cook who now wants to bake bread. Are there processing steps that were found during the search for a good cake recipe that can be re-used for a good bread recipe? The cook may have discovered that it is important to add the flour, sugar, and cocoa powder slowly when mixing it with the water, eggs, and vanilla extract. If the cook can identify an *explanation* for this part of the recipe, then it can be generalized. In this case,

the explanation is that when adding powdered ingredients (flour, sugar, cocoa) to a liquid batter (water, eggs, and vanilla extract), adding them slowly while stirring avoids creating lumps. This explanation supports the creation of a general rule: add powdered ingredients slowly to a liquid batter while stirring.

When baking bread, the cook can retrieve this rule and apply it, but this time the powdered ingredients are flour, salt, and dry yeast, and the liquid batter is water. Note that the explanation provides the useful abstractions (powdered ingredients, liquid batter) and also the justification for the rule. Explanation-based learning is a form of analytical learning, because it relies on the availability of background knowledge that is able to explain why particular steps succeed or fail.

Retrieving a rule is usually more difficult than retrieving an entire recipe. To retrieve an entire recipe, we just need to look up the name ('chocolate cake'). But to retrieve a rule, we must identify the relevant situation ('adding powdered ingredients to liquid batter'). Sometimes, the cost of evaluating the rule conditions is greater than the time saved by not searching. This is known as the *utility problem*. One solution to the utility problem is to restrict the expressive power of rule conditions so that they are guaranteed to be cheap to evaluate. Another solution is to approximate the rule conditions with different conditions that are easier to evaluate, even if this introduces some errors. This is known as *knowledge compilation*. Explanation-based learning mechanisms have been incorporated into cognitive architectures such as the Soar architecture of Newell, Rosenbloom, and Laird and the various ACT architectures of Anderson. (*See* **Soar**; **ACT**)

A third form of speedup learning is to learn search control heuristics. These heuristics typically take the form of evaluation functions, pruning rules, preference rules, or macros. An evaluation function assigns a numerical score to each situation (e.g. to each proposed cooking step). A pruning rule evaluates a proposed step and indicates whether it is likely to succeed. If not, the step can be pruned from the search. A preference rule compares two different proposed steps and indicates which is better. It can be applied to rank-order the proposed steps. A macro prescribes a *sequence* of steps to take when certain conditions are satisfied. All of these search control heuristics are typically learned by applying empirical learning methods.

Evaluation functions are frequently employed in game-playing programs. For example, the best computer backgammon program (Tesauro's TD-gammon) applies the empirical $TD(\lambda)$ reinforcement learning algorithm to learn an evaluation function for backgammon. (*See* **Samuel's Checkers Player**; **Reinforcement Learning: A Computational Perspective**)

Pruning rules can be learned as follows. Suppose our cook has kept records of all of the cooking experiments along with information about which experiments succeeded and which ones failed. We can consider each step in each recipe as a training example, and assign it a label of 'succeeds' or 'fails' depending on whether it was a good step to execute. Then supervised learning algorithms can be applied to learn a classifier for recipe steps. When developing a new recipe, this classifier can be applied to evaluate proposed steps and indicate which ones are likely to succeed – thus avoiding search.

Similarly, preference rules can also be learned by supervised learning. One training example is constructed for each *pair* of proposed actions (e.g. proposed cooking steps). The example is labeled as 'Better' if the first step is better than the second, and 'Worse' if the second step is better than the first. Then a supervised learning algorithm can learn a classifier for ranking the proposed actions.

Finally, macros can be learned in many different ways. One interesting method is known as the 'peak-to-peak' method. It can be applied to convert an imperfect evaluation function into a set of macros. A perfect evaluation function gives a high score to every good move from the starting state to the final goal (i.e. from the raw ingredients to the finished cake). An imperfect evaluation function will give high scores to some steps but then incorrectly give poor scores to a sequence of steps before again giving high scores. We can view this as a 'valley' between two peaks. The valley can be removed by introducing a macro that says when you reach the first peak, execute the *sequence* of steps that will lead to the next peak. By executing the macro as a single 'macro step', we are able to move right through the valley without becoming confused by the evaluation function errors.

Further Reading

Anderson JR (1989) A theory of the origins of human knowledge. *Artificial Intelligence* **40**: 313–352.

Bishop CM (1996) *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press.

Breiman L, Friedman JH, Olshen RA and Stone CJ (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Chatfield C (1996) *The Analysis of Time Series: An Introduction*, 5th edn. Baton Rouge, LA: Chapman and Hall/CRC.

Cristianini N and Shawe-Taylor J (2000) *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.

Dietterich TG (1997) Machine learning research: four current directions. *AI Magazine* **18**(4): 97–136.

Hand D, Mannila H and Smyth P (2001) *Principles of Data Mining*. Cambridge, MA: MIT Press.

Hastie T, Tibshirani R and Friedman J (2001) *The Elements of Statistical Learning*. New York, NY: Springer-Verlag.

Jelinek F (1999) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Jordan M (ed.) (1999) *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Mitchell TM (1997) *Machine Learning*. New York, NY: McGraw-Hill.

Newell A (1994) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Quinlan JR (1992) *C4.5: Programs for Empirical Learning*. San Mateo, CA: Morgan Kaufmann.

Shavlik J and Dietterich TG (1990) *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Sutton R and Barto AG (1998) *Introduction to Reinforcement Learning*. Cambridge, MA: MIT Press.

Machine Translation

Intermediate article

Christian Boitet, Université Joseph Fourier, Grenoble, France

CONTENTS

Machine translation is the subset of automated translation which aims to build programs capable of taking over translation per se, whether fully automatically or with the assistance of monolingual users.

INTRODUCTION

Machine translation (MT) was the first non-numerical application of computers. After initially promising demonstrations in the US around 1954, it was realized that high quality, fully automatic machine translation (HQFAMT) would in general be impossible. Less ambitious tasks were then attacked by lowering one or more requirements. The result was several kinds of automated translation (AT) systems. There exist many LQFAMT (low quality…) systems, producing rough translations used for accessing information in foreign languages. HQFAMT systems for very restricted kinds of texts are less common, but do exist. There are also MT systems for restricted text types which are high quality but not fully automatic (HQMT); these produce raw translations good enough to be revised cost-efficiently by professional revisers. HQMT can also be obtained by asking end users to assist the system. Finally, translation aids (TA), combining on-line dictionaries, term banks, and translation memories, are used extensively by professionals.

It is important to realize that human translation is difficult and diverse, and that automation is needed not only by end users but also by translators and interpreters. Also, automation itself comes in many forms. After expanding upon these points, we will move on to survey various linguistic and computational aspects of automated translation. This survey will yield an array of criteria for categorizing existing AT systems and for evaluating the state of the art. Finally, we present perspectives of future research, development, and dissemination.

Misconceptions about Translation

Translation is more difficult than usually believed, because of conceptual, cultural, and structural differences between languages; unavoidable ambiguities on many levels; and the need for considerable contextual knowledge. Underestimates concerning the inherent difficulty can distort evaluations of automated translation.

Further, the term 'translation' is often used imprecisely, as translation directions and purposes can vary widely. (According to the situation, the required translation may be 1-to-1, that is, from one language into another one, e.g. Russian–English; or 1-to-N, M-to-1, or M-to-N. The purpose of a translation may be to disseminate technical knowledge, to provide access to general information, etc. The purposes of speech translation, or interpretation, are again quite different.) Misleading generalizations about human translation may again distort perceptions about the practicality or worth of MT.

Misconceptions about human translation quality and costs are also common. While it is often believed that perfect understanding is needed to produce high quality translations, perfect understanding is rare, especially in rapidly developing specialties. In reality, experienced translators can often produce imperfect but high quality translations with less than full understanding. In any case, 'raw' translations produced by junior translators must usually be revised by senior translators. Thus the quality of even human translation is often lower than commonly supposed. Similarly, the work of interpreters is often judged much more by its speed and regularity than by its exactitude and completeness. As for translation cost, it is often underestimated. If it is done in-house, essential costs for such requirements as training, meetings, and research may not be taken into account. If it is subcontracted, different departments often use different subcontractors, and these expenses are not consolidated.

Why Automation is Needed

While misconceptions about translation remain widespread, the automation of translation and interpretation has become increasingly necessary since the 1950s as new requirements have arisen.

First, there can never be enough translators to cover the perceived needs. For example, not even the US army could train enough translators to skim through all Soviet literature in the Cold War era. With increasing globalization and the growth of the Internet, the need for all kinds of translation, from rough to raw to refined, is growing. The European Community now has eleven official languages, but still about 1200 in-house translators, the same number it had in the 1970s, when there were only eight official languages.

Second, many translation tasks are so boring or stressful that translators want to escape them. Automation can thus be seen as a way to free translators from menial tasks and promote them to revisers and co-developers of the automated systems.

Third, there is also increasing need for interpretation, especially to assist travelers abroad. There are many situations where sufficient knowledge of a common language is lacking: visiting a doctor, booking tickets for travel or leisure, asking for help on the roadside, calling motels, etc.

AUTOMATED TRANSLATION TASKS

When a translation job is automated, several tasks or subtasks may be affected. Computers can help to prepare translation resources, such as aligned bilingual texts; they can aid in preparation of the source text, e.g. by segmenting it; they can carry out all or part of the actual translation; and they can assist revision of a target text draft.

Preparation of Resources

Computers may first help in preparing resources for translation: they can (1) help to build bilingual terminological lexicons; (2) help to discover similar texts or dialogs, if possible already translated; (3) supply *aligners*, in order to build *bitexts* (aligned bilingual texts) from translations; or (4) provide terminology extractors working on monolingual texts or bitexts.

Preparation of Source Texts

Computers may also be used to prepare the texts to be translated. This preparation may first include correcting the texts using spellcheckers and grammar checkers – a crucial step, since bad source text quality is a principal source of translation errors. Second, it is often useful to *normalize* the source text: style and terminology checkers can be applied; pronouns with distant referents can be replaced by their referents; and implicit elements, such as subjects or objects in Japanese sentences, can be inserted. Third, texts can be segmented into maximally homogenous translation units. For example, long and complex sentences, even if they are admissible in the genre at hand, may be broken into simpler ones to help translation. Finally, it is also possible to annotate the text by inserting tags to help the analyzer.

Translation Proper

Translation *per se*, i.e. the passage from one language to another, can be totally, partially, or 'only apparently' automated.

'Pure' machine translation operates fully automatically, and its result is used by the end user. As mentioned, high quality of raw output can be obtained only by tuning systems to a restricted domain or genre. More generally, in order to achieve professional quality output and to give feedback to the machine translation developers, human revision must be used. It may take anywhere from one minute per page for very restricted domains (METEO ((Chandioux, 1988), 1976–), ALT/Flash (NTT, 1999–)…) to 10–15 minutes for broader domains such as technical manuals or administrative documents (ENGSPAN & SPANAM ((Vasconcellos and León, 1988), 1980–), CATALYST ((Nyberg and Mitamura 1992), 1992–)…).

In semi-automatic translation, one or more humans help the system translate. There are several variants. The oldest methods (ITS (BYU, Provo, 1974–81), N-Trans (Whitelock *et al*., 1986)) are highly interactive: as soon as the system encounters a problem in a local context (e.g. ambiguity in analysis, synonymy in generation, or both in transfer), a question is issued, and the user must give an answer. This procedure can be annoyingly demanding, and costly if specialized human knowledge is required. More recent semi-automatic systems (such as JETS at IBM-Japan (Maruyama *et al*., 1990; see also Boitet and Blanchon, 1994; Wehrli, 1992)) avoid these problems by delaying questions until one or two specific points in the translation process.

We say translation is 'only apparently' automated when the automation consists only of retrieving past translations from a *translation*

memory. When an exact match is found, the system may propose not only one, but several translations, produced in different contexts. In principle, all of them are very good, being of professional quality. When a 'fuzzy' or partial match is found, however, translations of similar but different source language segments are proposed, all of them in principle inadequate and requiring editing by the translator to become correct. In typical situations with high redundancy, such as successive versions of the same set of manuals, one may find 20–30% exact matches and 40–60% partial matches.

These approaches may be combined, e.g. by providing translators with suggestions from a translation memory and from an MT system.

Revision

To assist the revision of a first translation draft, both standard and specialized tools can be used. Standard tools would include various kinds of checkers of the target language. Specialized aids include: (1) flagging by the MT system of dubious passages (for which no complete analysis has been found, or in which unreliable choices had to be made); (2) production of two or more alternate translations of ambiguous terms, with appropriate formatting; and (3) development of special macro commands in the text editor to automate the most frequent correction operations, such as the permutation of three not necessarily contiguous blocks of texts (…A…B…C… to …B…C…A…).

COMPUTATIONAL ASPECTS OF AUTOMATED TRANSLATION

We now examine several computational aspects of automated translation. We will discuss linguistic architectures; various approaches to analysis of the source text; the handling of ambiguities; the use of specialized software languages; and the acquisition of linguistic resources.

At the outset, we should emphasize that imitating humans is no more a viable approach for translation than for voice recognition or other complex mental processes such as chess playing, since today's computers have very little in common with human brains. However, observation of human practice and introspection are nevertheless very useful: *what* is done can be reproduced independently of *how* it is done, and translator rules of thumb can be more or less exactly formalized as heuristic rules.

Linguistic Architectures

The possible linguistic architectures of MT systems are best understood with the help of Vauquois' triangle (Figure 1).

The triangle represents translation as proceeding from left to right. Thus we can imagine source language structures on the left, and target language structures on the right. The arrows which traverse the triangle from left to right represent translation or *transfer* processes which receive source language structures as input, manipulate them, and produce target language structures as output. Notice, however, that such linguistic structures can be defined at various *levels of abstraction or depth*: they may be shallow structures near the broad base of the triangle; or they may be deeper (more abstract) structures nearer to the narrow peak.

Each level can have its own translation or transfer process, shown in the figure as a horizontal arrow. The figure also contains ascending and descending arrows, representing the possibility that the transfer process could yield a structure more, or less, abstract than the input structure.

Direct and statistical MT: lexical replacement plus 'massaging'

Historically, the first MT paradigm was based on the supposed similarity of translation and deciphering. In this approach, a Russian text is viewed as the result of some encoding of an original English text, using word replacement and permutations of groups of words. To translate it into English, then, one tries to find the best English equivalents for the Russian words or groups of words by looking at the context in the utterance, and 'massage' the result into a grammatical English utterance.

This technique does not work very well, but human readers can often compensate for mistranslation and agrammaticality in the output if they are reading for comprehension only. The deciphering, or *direct*, approach is still used in many systems today (the majority of Systran language pairs until very recently, along with GlobaLink, Webtranslator, and others).

'Pure' statistical MT is a modern version of the direct approach. The IBM group, which pioneered the successful use of statistical methods for speech recognition has pioneered in the statistical translation area as well. The idea is to learn correspondences between (groups of) words of two languages from a large set of aligned sentences, such as the Hansard bilingual transcriptions of the debates of the Canadian Parliament, and then use these learned correspondences to translate new

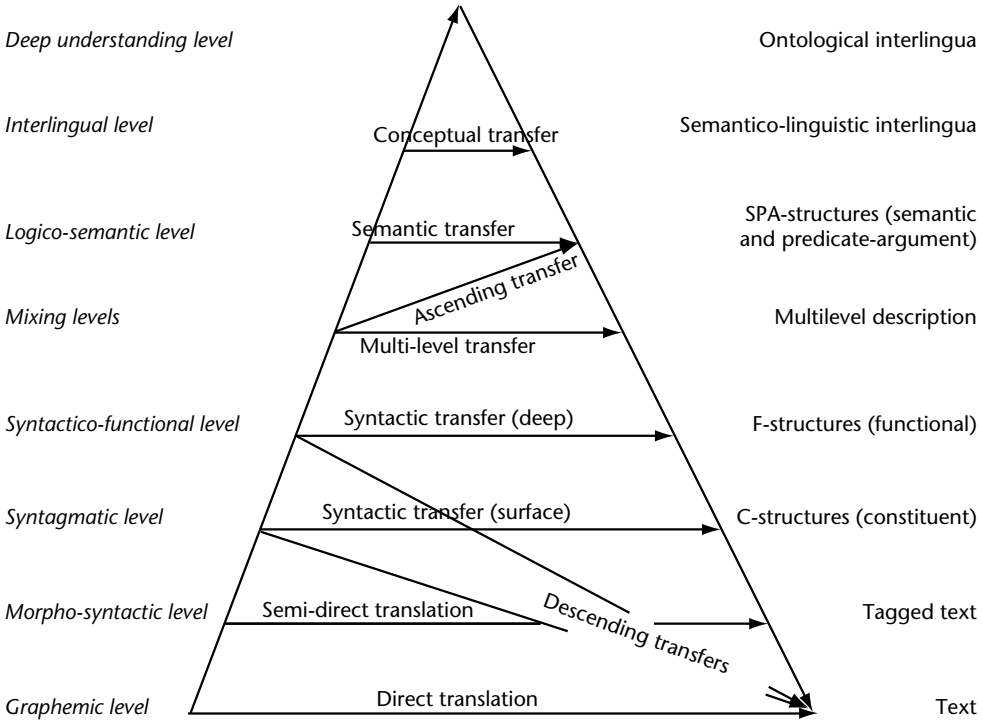


Figure 1. Vauquois' triangle.

sentences. Because of the huge number of possibilities to be examined (the combinatorial explosion), sentences used for training cannot be very long, and the word correspondences are usually limited to 1-1, 1-0, 0-1, and 1-N, although M-N (many-to-many) correspondences would be desirable.

Experiments have demonstrated that this statistical technique is inferior to the older handcrafted direct approach. But its proponents have been able to use the same ideas at higher levels of linguistic interpretation, this time with success (e.g. using 'stochastic grammars' at the syntagmatic, or syntactic, level, and more recently employing 'semantic grammars'), in effect moving from a semi-direct approach to a *transfer* or *pivot* approach.

Transfer approaches to MT

In the 'strict' transfer approach to MT, there are three steps: (1) a strictly monolingual analysis into a descriptive structure of the source language at some level L of linguistic interpretation; (2) a bilingual transfer (normally decomposed into a lexical and a structural phase) into a descriptive structure of the target language at the same level L; and (3) a monolingual generation process. One can distinguish between surface (e.g. TDMT (ATR)) and deep syntactic transfer (e.g. LMT (IBM)), semantic transfer (e.g. MU (University of Kyoto), and its successor MAJESTIC (JST)), and multi-level transfer (if the descriptive structure contains both syntactic and semantic information (e.g. GETA's Russian–French system (Boitet and Nédobejkine, 1981))). Whichever variant is used, the transfer-based translation technique makes it possible to reuse the analysis and generation steps in several language pairs (English–German and Russian–German).

Due to a lack of appropriate tools, some developers have started with the strict transfer approach in mind, but have not been able to develop the 'structural' part of a strict transfer process – that part which represents correspondences between arbitrarily complex source language and target language structures. The resulting systems (e.g. METAL (Slocum, 1985), Shalt-1 (IBM)) mix transfer and syntactic generation in one step, usually implemented by a recursive descent of the analysis tree structure, thus achieving a sort of 'descending' transfer, in which the transfer output is less abstract than the input.

By contrast, B. Vauquois introduced the technique of 'ascending' transfer, in which the transfer output is more abstract than the input. In this approach, the analysis produces a *multi-level structure* containing both semantic and syntactic structures. That is, the structures contain both a language-neutral level of logico-semantic interpretation and some information about surface levels,

such as syntagmatic categories (noun phrase, verb phrase, etc.) and syntactic functions (subject, direct object, etc.). During transfer, some surface-level information may be transferred from the source language structure to that of the target language; but, during generation, any remaining surface information is recomputed from the language-neutral semantic level. At this generation stage, the goal is to minimize the transfer of structural surface information – to rely as far as possible on the transfer of semantic information, so that the target language can express this semantic information idiomatically.

Pivot approaches to MT

Strictly speaking, the *pivot* approach simply consists in using some standard intermediate format to represent the information passed during translation from any language into any other language. But the nature of this pivot format may vary: it may be a natural language text, or it may be a structure describing the utterance being translated. In the latter case, the structure may or may not include language-specific information. Further, the intended coverage may be broad, ideally handling the relevant languages at large; or only restricted domains may be addressed.

Using natural language text as a pivot structure implies using the output of one MT process as the input for one or more later processes. Clearly, this approach is practical only if the output is of high quality and grammatically correct. But it always encounters the problem of the inherent ambiguity of the pivot language.

As mentioned, the next possibility is to use structural descriptors of a natural language L0 as pivotal elements, rather than normal text of L0. To translate between L0 and another language Lk, a standard transfer approach is employed, as explained above. However, to translate between two other languages Lj and Lk, one first analyzes Lj to obtain a structural description at some level; then performs a transfer to a corresponding structure of L0 (at the same level or a more abstract one); then performs a *second* transfer operation from the L0 structure to a structure of Lk; and finally generates target text from the Lk structure. If this procedure is followed, text of the intermediate language L0 need never be generated or analyzed, so any additional errors or ambiguity which might have been introduced during this generation or analysis can be avoided.

This structural pivot technique was initially proposed and tried by CETA in Grenoble (Vauquois, 1975) in the 1960s, under the name of *hybrid pivot*. It was then used successfully in the DLT project at BSO research between 1982 and 1988. The pivot language was Esperanto, and the pivot descriptors were Esperanto utterances augmented with structural and grammatical tags. The same technique is currently used by the IPPI team in Moscow (Boguslavskij *et al*.) to translate from UNL (see below) into Russian.

When the structural pivot approach is adopted, some means is needed to represent the lexical elements, or words, of the source and target structures. These *lexical units* can be dictionary forms (or *lemmas*) extracted from dictionaries of the relevant languages. Since natural language words generally have several meanings, it will in this case be necessary to specify the intended meaning, e.g. by using numerical indexes (so that *observe.1* indicates the watching meaning of this verb, and *observe.2* indicates the obeying meaning). Lexical units can also be derivational families, e.g. the unit 'OBSERVE-2' might contain the lemmas *observe, observable, observant, observance*.

However, it is also possible to create a language-neutral set of word meaning symbols (or *acceptions*). If such neutral symbols are used within structures which themselves are sufficiently abstract to contain no language-specific information, the result is a *linguistic interlingua*.

The best example today is UNL (the Universal Networking Language). The UNL project is concerned with multilingual communication over the Internet. It was initiated in 1996 by the University of the United Nations (based in Tokyo), and addressed fourteen languages as of 2000. The representation of an utterance is a hypergraph with a unique *entry node*, where normal nodes bear *Universal Words* (UWs, or interlingual word senses) with semantic attributes, and arcs bear semantic relations (deep cases, such as agt, obj, etc.).

Figure 2 gives an example of a UNL graph. The semantic attributes, marked by '@', represent phenomena like speech acts, modality, time, etc.

Finally, it is possible to develop purely semantic, task-oriented pivot representations within restricted domains such as transport and hotel reservation, schedule arrangement, etc. This approach has been taken in several speech translation projects such as CSTAR, which employs the Interface Format (IF) for intermediate representation. An utterance by a client such as *Hello Tom, how are you today?* may be represented by the IF expression 'C: greetings(time = morning, level = familiar)', which could then be re-expressed as *Good morning!*

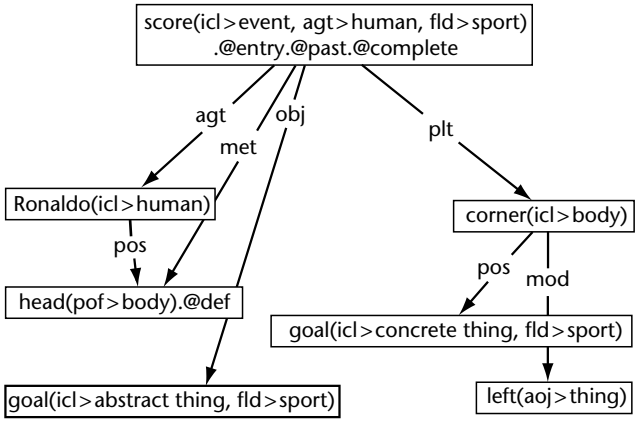


Figure 2. A simple UNL graph.

Analysis of Input: Symbolic, Numerical, and Hybrid Techniques

All operational MT systems use symbolic (rule- or procedure-based) techniques, sometimes enhanced with numerical (statistical) techniques. Their linguistic programs are thus based either on formal grammars or on automata.

The grammatical approach uses regular or context-free grammars, extended with attributes and abstract tree-building mechanisms, much as is done to build compilers.

Various sorts of *transducing* automata – abstract machines which incrementally transform input structures into output structures – are also used directly: (1) Finite State Transducers (FSTs) are extensively used for morphological and bounded syntactic processes. (2) In 1970, W. Woods introduced string-to-tree transducers called Augmented Transition Networks (ATNs) to build analyzers, and used them in the LUNAR project. ATN-based analyzers are used today in the Prompt (Softissimo), ENGSPAN/SPANAM (PAHO) and AS-TRANSAC (Toshiba) MT systems. (3) G. Stewart extended the model to Q-graphs transducers for the TAUM-Aviation project, but his remarkable REZO language has yet to find a successor. (4) Finally, various types of transducers for annotated trees (transformational systems) have been used for structural analysis, transfer, and generation. The first was built at CETA (Grenoble) between 1965 and 1966. Quite a few MT systems use some variants today, including ROBRA in GETA's Ariane-G5-based MT systems, L. C. Tong's GWR in Tapestry at KRDL in Singapore, Nakamura's GRADE in MU/Majestic in Japan, and HICATS of Hitachi.

Numerical, or statistical, techniques are increasingly used for source language analysis in speech translation. The end-to-end evaluation of the German VerbMobil project (1992–2000) has shown that, after comparable development efforts, they can lower the error rate to 30%, while the symbolic techniques produced errors in at least 60% of all cases. This advantage is due not only to the noisy nature of the input (a word lattice produced by the speech recognizer), but also to the fact that numerical techniques can 'learn' the weights of the transitions of a stochastic automaton far more complex than any automaton which could be built by hand.

Another promising avenue for source language analysis is the 'hybrid' approach, where symbolic and numerical techniques are mixed at different points. An exemplary case is the MT system developed by Microsoft (2000–01).

Handling Ambiguities

Traditionally, the goal of *source language analysis* is to produce one or several abstract representations (annotated or 'decorated' trees, feature structures, conceptual graphs, logical formulae, etc.), each of them *disambiguated*.

In the standard *combinatorial approach* to analysis, *all* possible analyses are computed. Subsequently, since a single best solution or small set of solutions is normally desired, filters, weights, or preferences are used. Filtering consists in applying increasingly stricter conditions of well-formedness until the set of solutions is satisfactorily reduced. Weights may be computed for solutions in a number of ways.

In the *heuristic approach*, by contrast, only *some* possible analyses are computed. The worst heuristic technique is to use a standard search algorithm (e.g. backtracking) and stop once the first solution is reached, because in this case potential ambiguities are simply ignored. A better technique is to use weights to guide the search and to seek more than one solution.

These two search techniques do not actually *handle* all candidate solutions; they are just ways to *eliminate* some solutions from further consideration. The use of preferences is more to the point: one tries to recognize the cause of ambiguity and choose among all candidates on some principled basis. But then all ambiguity cases and their solutions must be foreseen, a very difficult task. Also, principled selection of the best candidate based upon preferences is local, which makes it difficult to guarantee an overall best solution, and perhaps explains why other techniques often perform as well in practice. In practical working systems, the heuristic and the preferential approaches are combined.

The main problem with usual search techniques is that they produce a set, or a (weighted) list, of complete, fully disambiguated solutions, and *discard* all trace of the source of the ambiguities. B. Vauquois' solution to this problem of information loss was to produce abstract structures that *preserve certain internal ambiguities*, either directly, or through 'tactical' annotations. This technique enables a system to produce 'warnings' in the output which show the presence of ambiguities in a factorized way. However, ambiguities are coded in the same structures (annotated trees) used to describe disambiguated solutions, which leads to more complex transfers and generations.

Specialized Programming Languages

Specialized languages for linguistic programming (SLLPs, earlier called 'metalanguages') have been around since the 1960s, aiding in the production of MT dictionaries. Linguistic programming has often been done using classical programming languages (macroassembler and the C-macroprocessor for Systran, LISP for parts of METAL, etc.).

The idea is simply to define a symbolic language familiar to the linguist, and then to compile linguistic data or programs written in this SLLP into interpretable or directly executable binary code. The first complete SLLP was COMIT, developed at MIT in the 1960s. Later languages have been based upon theoretical models, or have emphasized powerful data structures and control structures.

Linguistic Resource Acquisition

Large-scale lexical acquisition has been a major problem since the beginnings of MT, in the 1960s. There are three main approaches, each with its context, methodology, and pros and cons. The first approach consists in working directly on dictionaries specialized for MT; the second in creating specialized lexical databases, generally asymmetrical and proprietary, and sometimes usable for applications different from MT; and the third in building lexical databases which can not only be used in multiple ways by both humans and machines, but are also intrinsically symmetrical, linguistically very detailed, potentially containing a great many entries and languages, and open. For the future, if MT systems are to cover all pairs of languages and all domains, it seems that the only approach which might circumvent prohibitive costs is that of Linux: a collaborative, Web-based approach to the creation and usage of lexical resources.

For any of these methods, it is necessary to use corpora to obtain primary information and to test refined information. In recent years, corpus linguistics has led to the development of very powerful automatic tools, based on low-level linguistic processing and sophisticated statistical techniques, which enormously help lexical acquisition. For example, there are now effective terminology extractors which work on monolingual documents and on bilingual aligned texts. However, their results must be refined by human labor: as many as 40% of the suggestions may be wrong, and information automatically obtained from the remaining 60% must still be revised.

The acquisition of syntactic and semantic knowledge remains mostly a human responsibility, reserved for specialists. However, in hybrid techniques, programs can to some degree be used to learn from examples, as at T. Matsumoto's lab at NAIST (Nara).

STATE OF THE ART

To assess current MT systems, it is useful to distinguish three main goals: MT aiming at rough translation of texts, MT aiming at quality translation of texts, and MT of speech.

Rough MT for Comprehension

Many MT systems are currently available, at low prices, for basic comprehension. The obtained translations are rough, but often adequate. They cannot practically be directly revised to obtain

quality translation, but readers understand the gist of the information, or at least its topics, and translators can use the result as a suggestion, just as they use suggestions from translation memories. The main uses for comprehension MT now are for surfing the Web and for seeking information, although military, economic, and scientific intelligence were early users, and demand continues in these areas. The number of available language pairs is in fact very low compared to the needs. For example, an official of the EU reported at LREC-2000 that EC-Systran had only nineteen pairs after 24 years of development, eight of them 'almost satisfactory'. However, there are 110 language pairs in the EC. In Japan (and similarly in China), very few language pairs are offered besides English ↔ Japanese and English ↔ Chinese. Russian is offered for two or three pairs, and Thai only for English ↔ Thai. Some websites claim to offer many language pairs by translating through English. Unfortunately, the results are usually terrible, for reasons explained above.

The identified obstacles here are:

- the cost of developing the first commercial version of a new language pair (at least 40 person-years according to the CEO of Softissimo).
- the direct approach, which makes it impossible to combine two systems without dramatically lowering the quality.
- the law of diminishing returns: each new language pair to be developed usually corresponds to a lesser need than the previous one, hence there are fewer users/buyers, all expecting to pay no more than the cost of already available language pairs.

Quality Raw MT for Dissemination of Information

We find here specialized systems for rare niches, such as METEO (Chandioux, 1988), ENGSPAN, SPANAM (Vasconcellos and León, 1988), METAL (Slocum, 1985), LMT (McCord, IBM), CATALYST (Caterpillar-CMU), perhaps some LOGOS systems, etc. In Japan, we might mention ALT/Flash (the NTT system for Nikkei stock market flash reports) and perhaps some specialized systems, mostly ENG-JAP, used internally for translating technical manuals (AS-Transac at Toshiba, ATLAS-II at Fujitsu, CrossRoad at NEC, SHALT at IBM, Pensée at OKI, etc.). In Europe, few such systems are now available, due to the relatively small market, and to the negative attitude of the EC and all governments towards funding quality MT.

Quality MT systems for information dissemination are very rare. However, these systems are indeed very good and very useful (30 Mwords/year for METEO, 75% English–French and 25% French–English, with 1mn revision per page, 0.15 cents/word for final output), because they are quite specialized.

It is extremely difficult to prepare comparative benchmarks for such systems because, like expert systems, they are very good in their domain, and fail miserably on other tasks. The best way to measure them is through some combined assessment of the sale, maintenance, and evolution prices, and through consideration of the human time needed to obtain a professional result.

Technically, these systems almost always have a separate analysis component, producing a syntactic or syntactico-semantic descriptor of the source unit of translation (usually an annotated tree). Almost all use some flavor of the transfer approach (even systems like ATLAS-II by Fujitsu or PIVOT-CROSSROAD by NEC). In most cases, there is no syntactico-semantic descriptor of the target unit of translation, transfer and generation being merged into a single phase using recursive descent of the analysis tree. Hence, changing the source language implies redoing all the work. (Compare the difficulties experienced by Siemens with METAL in the 1990s, which contributed to this company's exit from the scene.)

The identified obstacles here are:

- the cost of developing the first commercial version of a new language pair: at least 100 person-years according to H. Sakaki, the main author of KATE at KDD, and to Dr. Nagao, director of the MU project; and perhaps 300 person-years with large dictionaries, as H. Uchida estimated for ATLAS-II at Fujitsu.
- the impossibility of factorizing generation processes, when the situation changes from $1 \rightarrow N$ to $M \rightarrow N$.
- the apparent lack of demand for high quality translation for many language pairs, even though structural transfer systems could be combined, as explained above, to produce multiple pairs.

Speech Translation

Current commercially available technology makes speech translation already possible and usable for social interchanges such as chat. Such systems are usually built by combining speech recognition (SR), text MT, and speech synthesis. NEC demonstrated a system for JAP ↔ ENG at Telecom'99. Linguatec, a subsidiary of IBM, markets Talk & Translate (using IBM's Via Voice and Logic-Programming based Machine Translation (LMT) for ENG ↔ GER). The quality is of course not very high in all components, but this drawback is compensated for by the broad

coverage, by some feedback (e.g. editable output of SR and written reverse translation), and by human intelligence and desire to communicate.

At the research level, the aim is to obtain higher quality while allowing more spontaneous speech in task-oriented situations. The large German Verb-Mobil project (1992–2000) showed the feasibility of reaching these goals, and compared many alternative methods in the same task setting (Wahlster, 2000). The goals can also be reached in a multilingual setting, as demonstrated by the C-STAR II consortium in intercontinental public demonstrations with large media coverage in July 1999. C-STAR II used a kind of semantico-pragmatic pivot designed to represent the dialogue utterances of participants in a limited set of situations (e.g. getting tourism information, booking hotels or tickets for events and transports, etc.).

The identified obstacles here are:

- the great difficulty of developing an adequate pivot (such as the IF or Interface Format of CSTAR).
- the cost of building the necessary lexical resources, as for MT of texts.
- the difficulty of handling the context (pragmatic, discursive, linguistic, and lexical), and in particular of computing correctly the speech acts and the referents of anaphoras and elisions.

PERSPECTIVES: KEYS TO THE GENERALIZATION OF MT IN THE FUTURE

Despite considerable investment since the 1960s, only a few language pairs are covered by MT systems designed for information access, and even fewer are capable of quality translation or speech translation. To open the door towards MT of adequate quality for all languages (at least in principle), four keys are needed. On the technical side, one should (1) dramatically increase the use of learning techniques which have demonstrated their potential at the research level, and (2) use pivot architectures, the most universally usable pivot being UNL. On the organisational side, the keys are (3) the cooperative development of open source linguistic resources on the Web, and (4) the construction of systems where quality can be improved on demand by users, either a priori through interactive disambiguation, or a posteriori by correcting the pivot representation through any language, thereby unifying MT, computer-aided authoring, and multilingual generation.

References

Boitet C and Nédobejkine N (1981) Recent developments in Russian-French machine translation at Grenoble. *Linguistics* **19**: 199–271.

Boitet C and Blanchon H (1994) Multilingual dialogue-based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation* **9**(2): 99–132.

Chandioux J (1988) 10 ans de METEO (MD). In: Abbou A (ed.) *Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires*, pp. 169–173. Paris: Observatoire des Industries de la Langue (OFIL).

Maruyama H, Watanabe H and Ogino S (1990) *An Interactive Japanese Parser for Machine Translation*. Proceedings of COLING-90, 20–25 August, ACL.

Nyberg EH and Mitamura T (1992) *The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains*. Proceedings of COLING-92, 23–28 July, ACL.

Slocum J (1985) A survey of machine translation: its history, curent status, and future prospects. *Computational Linguistics* **11**(1): 1–17.

Vasconcellos M and León M (1988) SPANAM and ENGSPAM: machine translation at the Pan American Health Organization. In: Slocum J (ed.) *Machine Translation Systems*, pp. 187–236. Cambridge, UK: Cambridge University Press.

Vauquois B (1975) *Some Problems of Optimization in Multilingual Automatic Translation*. Proceedings of First National Conference on the Application of Mathematical Models and Computers in Linguistics, May, Varna.

Wahlster W (2000) Verbmobil: Foundations of Speech-to-Speech Translation. *Artificial Intelligence*. Berlin: Springer.

Wehrli E (1992) *The IPS System*. Proceedings of COLING-92, 23–28 July, Nantes.

Whitelock PJ, Wood MM, Chandler BJ, Holden N and Horsfall HJ (1986) *Strategies for Interactive Machine Translation: The Experience and Implications of the UMIST Japanese Project*. Proceedings of COLING-86, 25–29 August, Bonn.

Further Reading

Boitet C (1993) La TAO comme technologie scientifique: le cas de la TA fondée sur le dialogue. In: Clas A and Bouillon P (eds) *La traductique*, pp. 109–148. Montreal: Presses de l'université de Montreal.

Boitet C (1996) (Human-aided) machine translation: a better future? In: Cole R, Mariani J, Uszkoreit H, Zaenen A and Zue V (eds) *State of the Art of Human Language Technology*, pp. 251–256. Pisa: Giardini.

Hutchins WJ (1986) *Machine Translation: Past, Present, Future*. Chichester, UK: John Wiley & Sons.

Lehrberger J and Bourbeau L (1988) *Machine Translation. Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. John Benjamins.

Planas E (1999) *Formalizing Translation Memories*. Proceedings of MT Summit VII, 13–17 September, Singapore.

Markov Decision Processes, Learning of

Advanced article

Kevin P Murphy, University of California, Berkeley, California, USA

CONTENTS

Markov models are probabilistic models of dynamical systems, which can be used to predict the future and the consequences of one's actions.

INTRODUCTION

The ability to predict the future is fundamental to probably all intelligent systems. Since in general we cannot hope to make perfect predictions, it makes sense to adopt a probabilistic approach. For example, although we may not be able to predict the exact price of a stock tomorrow, we may be able to predict its *expected* price; deviations around this mean will be modeled as 'noise', assumed to result from all the unknown factors that were omitted from the model. Similarly, although we may not be able to exactly predict the next word that someone will speak or type at a keyboard, we can predict a set of likely or probable words based on what we have heard or seen so far. (*See* **Reasoning under Uncertainty**)

It is standard to model the system to be predicted as a dynamical system. That is, the system is assumed to be in some state, $X_t \in \Omega$, at each time step t; the next state of the system is determined by the state transition function, $X_{t+1} = f(X_t)$. (In this article, we restrict our attention to discrete time dynamical systems.) Typically we do not know the exact dynamics of the system, so instead we consider a probabilistic state transition function: $P(X_{t+1}|X_t)$. Such a probabilistic formulation will be particularly useful when we try to *learn* the model from data, since although there may be no model that fits the data perfectly, some models might be more probable than others. The state space, Ω, might be discrete (finite) or continuous (infinite). For example, we might try to predict the probability that a stock goes up or down, in which case $\Omega\{\uparrow, \downarrow\}$, or we might try to predict its expected value, in which case $\Omega = \mathbb{R}$. Of course, the state X_t may also be vector-valued. (*See* **Dynamical Systems: Mathematics**)

A (first-order) Markov process is one such that $P(X_{t+h}|X_{1:t}) = P(X_{t+h}|X_t)$, where $X_{1:t} = X_1, \ldots, X_t$ is the past observation sequence and $h > 0$ is the amount of lookahead (the prediction horizon). This is often paraphrased as: the future X_{t+h} is independent of the past $X_{1:t-1}$ given the present X_t. A k'th order Markov process is such that $P(X_{t+h}|X_{1:t}) = P(X_{t+h}|X_{t-k+1:t})$. $X_{t-k+1:t}$ is called a sufficient statistic, since it can be used to predict the future with the same accuracy as would be obtained if we conditioned on the whole sequence of past observations. A k'th order Markov model can always be converted to a first-order Markov model by creating a 'mega variable' containing the last k observations (a sliding window). For example, if the system is second-order Markov, we can convert it to first-order by defining a new state space, $\tilde{X}_t = (X_t, X_{t-1})$, and set $P(\tilde{X}_t = (x_t, x_{t-1})|\tilde{X}_{t-1} = (x'_{t-1}, x_{t-2})) = \delta(x_{t-1}, x'_{t-1})\, P(x_t|x_{t-1}, x_{t-2})$. The δ function ensures that $\tilde{X}_t$ and $\tilde{X}_{t-1}$ assign the same value to the variables that they share (in this case, X_{t-1}).

Often a system has no finite (or reasonably small) sufficient statistic, which means we need to base our predictions on all the past data; this is clearly impractical, especially for high-dimensional data such as audio or video signals. However, we may posit the existence of a hidden or latent variable, which generates the observations at each step, and whose transition dynamics do satisfy the (first-order) Markov property. This is called a hidden Markov model (HMM) if the latent variable is discrete, or a state-space model if the latent variable is continuous. We will denote the hidden variable by X_t and the observed variable by Y_t. (Since the value of X_t is hidden, the model is called 'partially observed'.) In addition to the transition function $P(X_t|X_{t-1})$, we must now specify the observation function, $P(Y_t|X_t)$. We will explain this in more detail below.

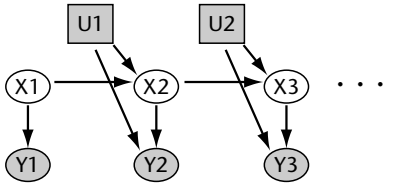


Figure 1. A generic discrete-time dynamical system represented as a dynamic Bayesian network (DBN). U_t is the input X_t is the hidden state, and Y_t is the output. Shaded nodes are observed, clear nodes are hidden. Square nodes are fixed inputs (controls), round nodes are random variables.

In addition to hidden and observed variables, we may have control or input variables, U_t. In this case, the transition function becomes $P(X_t|X_{t-1}, U_t)$ and the observation model becomes $P(Y_t|X_t, U_t)$. (Notice how what we see, Y_t, may depend on the actions that we take, U_t: this can be used to model active perception.) The resulting model can be represented as a Bayesian network, as shown in Figure 1. (*See* **Bayesian Belief Networks**)

In the following sections, we will explain how to learn Markov models from data. We start with the case where the system is fully observed, and then consider the harder case in which X_t is hidden, and we only get to observe Y_t, which is a probabilistic function of X_t.

FULLY OBSERVED MODELS

Discrete State Spaces: Markov Chains

In a finite state Markov chain, the system can be in one of S states, $X_t \in \Omega = \{1, 2, \ldots, S\}$; the probability of a transition from state i to state j is specified by a transition matrix, $A_t(i,j) \stackrel{\text{def}}{=} P(X_t = j|X_{t-1} = i)$. If A_t is independent of time, then this is called a homogeneous or stationary Markov chain; we shall assume this throughout. The initial state distribution is specified by $\pi(i) \stackrel{\text{def}}{=} P(X_1 = i)$. Since π is a probability distribution, it must satisfy the constraint that $\sum_{i=1}^{S} \pi(i) = 1$. Similarly, each row of A must satisfy $\sum_j A(i,j) = 1$; such a matrix is called stochastic.

A simple example of a finite state Markov chain is a bigram model, widely used in statistical approaches to natural language processing. In this case, Ω is the set of words, and the goal is to predict the next word given the previous. An n-gram model uses the last $n-1$ words to predict the next word, e.g. in a trigram model, the goal is to learn $P(X_t|X_{t-1}, X_{t-2})$ (see Model selection). (*See* **Natural Language Processing, Statistical Approaches to**)

Another example of a finite state Markov chain is a Markov decision process (MDP), widely used in reinforcement learning. In an MDP, the state transitions are 'triggered' by an input U_t. If we suppose the inputs are also discrete valued, we may write $P(X_t = j|X_{t-1} = i, U_{t-1} = k) = A_k(i,j)$; that is, the input just specifies which transition matrix to use. This can be represented as in Figure 1 by omitting the observation nodes Y_t (since we are assuming that we can directly observe X_t; if this were not the case, the model would be a *partially observed* MDP, or POMDP). (*See* **Reinforcement Learning: A Computational Perspective**)

Parameter estimation

Given a training set D of sequences, the goal of parameter learning (also known as system identification) is usually defined as finding the maximum likelihood estimate of the parameters: $\hat{\theta}_{ML} \stackrel{\text{def}}{=} \arg\max_\theta P(D|\theta)$, where $\theta = (\pi, A)$. If the training data contain N independent sequences, we can express the likelihood as $P(D|\theta) = \prod_{n=1}^{N} P(D^{(n)}|\theta)$. To compute the likelihood of an individual sequence, consider an example: $X_{1:4} = (1, 2, 1, 2)$. This has likelihood

$$\begin{aligned} P(X|\theta) &= P(X_1 = 1)P(X_2 = 2|X_1 = 1) \\ &\quad P(X_3 = 1|X_2 = 2)P(X_4 = 2|X_3 = 1) \\ &= \pi(1)A(1,2)A(2,1)A(1,2) \\ &= \pi(1)A(1,2)^2A(2,1)^1 \end{aligned}$$

In general, the likelihood is

$$P(D|\theta) = \prod_{i=1}^{S} \pi(i)^{\#1(i)} \prod_{i=1}^{S} \prod_{j=1}^{S} A(i,j)^{\#(i \to j)}$$

where $\#(i \to j)$ is the number of times an i to j state transition is observed in the whole training set:

$$\#(i \to j) \stackrel{\text{def}}{=} \sum_{n=1}^{N} \sum_{t=2}^{T_n} I(X_{t-1}^{(n)} = i, X_t^{(n)} = j)$$

Here, $I(e)$ is the indicator function that is 1 if event e occurs, and is 0 otherwise. $\#_1(i)$ is defined analogously to be the number of times the first state is observed to be i.

Maximizing the log-likelihood is equivalent to maximizing the likelihood (since log is a monotonic transformation), but is mathematically more convenient. The log-likelihood of a sequence is given by

$$\log P(D|\theta) = \sum_{i=1}^{S} \#_1(i) \log \pi(i) + \sum_{i=1}^{S} \sum_{j=1}^{S} \#_{i \to j} \log A(i,j)$$

To maximize the log-likelihood we must introduce Lagrange multipliers to enforce the sum-to-one constraints. Some simple calculus yields the intuitive results that the parameter estimates are just the normalized counts:

$$\hat{A}_{ML}(i,j) = \frac{\#(i \to j)}{\sum_k \#(i \to k)}$$

$$\hat{\pi}_{ML}(i) = \frac{\#_1(i)}{\sum_k \#_1(k)} = \frac{\#_1(i)}{N}$$

One problem with maximum likelihood (ML) estimation is that it assigns a probability of 0 to any event that was not seen in the training data; this is called the 'sparse data' problem. To see why this is a problem, consider the case of bigram models of language, where Ω is the set of all words. If the training corpus contains the phrase 'good dog' but not 'bad dog', the ML parameter estimates will assign a probability of 0 to any test sentence that contains 'bad dog', even though intuitively this ought to be as probable as a sentence that contains 'good dog'.

A simple and well principled solution to the sparse data problem is to compute the maximum a posteriori (MAP) estimate instead of the ML estimate: $\hat{\theta}_{MAP} \stackrel{\text{def}}{=} \arg\max_\theta P(\theta|D)$. By Bayes' rule

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

or, in words,

$$\text{posterior} = \frac{\text{conditional likelihood} \times \text{prior}}{\text{likelihood}}$$

If we use a Dirichlet prior $P(\theta)$ (Heckerman, 1998), it can be shown that the MAP estimate of the transition matrix of a Markov chain is given by

$$\hat{A}_{MAP}(i,j) = \frac{\#(i \to j) + \alpha(i,j)}{\sum_k \#(i \to k) + \alpha(i,k)}$$

where the $\alpha(i,j)$ have an intuitive interpretation in terms of 'pseudo counts', i.e. $\alpha(i,j)$ is the number of times an $i \to j$ transition was observed in some prior 'virtual' training set. Using the uninformative prior $\alpha(i,j) = 1$ (also known as Laplace's prior) has the effect of preventing any 0s in the estimated parameters, but otherwise letting the data 'speak for itself'. (*See* **Reasoning under Uncertainty**)

We can estimate the parameters of a k'th order Markov model in a similar way. Unfortunately, the number of parameters that we need to specify $P(X_t|X_{t-k:t-1})$ is now $O(S^k)$, making the sparse data problem particularly severe. One solution is to approximate the conditional distribution $P(X_t|X_{t-k:t-1})$ as a mixture of k bigram models (Saul and Jordan, 1999). We can learn the parameters of such a mixture model using the EM algorithm (see The EM algorithm). Another solution is to use more complex parameter priors (MacKay and Peto, 1995), or to cluster words into similar categories; in this case, the category is a hidden variable, making the model partially observable.

Model selection

The problem of choosing the appropriate number of steps of history to maintain, k, is called *model order determination*, and is a simple example of *model selection*. Maximum likelihood is not an appropriate metric, since the model with the highest likelihood on the training set will always be the one with the largest possible value of k (since this has the largest number of parameters, and hence can fit the data the best: this is called overfitting). One approach to model selection is to use cross-validation, i.e. to choose k by measuring performance on a hold-out set (as opposed to the training set). Alternatively, we can can choose the model M that maximizes a penalized log-likelihood function: $M^* = \arg\max_M \log P(D|\hat{\theta}_M) - g(d_M)$, where $\hat{\theta}_M$ is the ML estimate of the parameters for model M, d_M is the number of parameters in M, and $g(d)$ is some complexity penalty that increases with d. Some popular choices for this penalty function are the Akaike information criterion (AIC), $g(d) = d$, the Bayesian information criterion (BIC), $g(d) = \frac{d}{2}\log N$, where N is the number of samples, and the minimum description length (MDL) criterion, which is the same as BIC (Heckerman, 1998).

Continuous State Spaces

The most common form for a Markov process with a continuous state space is $X_t = f(X_{t-1}, U_{t-1}) + W_t$ where f is some function, and W_t is a zero-mean random variable, representing noise.

Linear systems

A widely studied special case is when f is a *linear* function, and the noise is Gaussian, in which case we can write $X_t = AX_{t-1} + BU_{t-1} + W_t$ where $W_t \sim N(0, \sum)$. Equivalently, we may write

$$P(X_t = x_t|X_{t-1} = x, U_{t-1} = u) = N\left(x_t; Ax + Bu, \sum\right)$$

where the notation $N(x; \mu, \sum)$ means evaluating a normal (Gaussian) probability density function with means μ and covariance $\sum$ at the point (vector) x. The distribution of the initial state is $P(X_1 = x) = N(x; \mu_1, \sum_1)$. It is possible to estimate

the parameters of such a linear system using techniques based on linear regression (Ljung, 1987).

Nonlinear systems

The usual approach to learning nonlinear dynamical systems is to represent the f function using a feedforward neural network (also called a multilayer perceptron). We create a training set from the matrices

$$X = \begin{pmatrix} X_1' & U_1' \\ \vdots & \\ X_{T-1}' & U_{T-1}' \end{pmatrix}, \quad Y = \begin{pmatrix} X_2' \\ \vdots \\ X_T' \end{pmatrix}$$

where the t'th row of X is the t'th input vector, and the t'th row of Y is the corresponding output. (X_t' denotes the transpose of the vector X_t.) Now we use some nonlinear learning procedure, such as back propagation or a conjugate gradient method (Bishop, 1995), to estimate the parameters of the f function. The ML estimate of $\sum$ is the empirical covariance matrix of the residuals (i.e. the difference between the predicted output and the actual output). (*See* **Perceptron**)

Higher order models

It is easy to convert a k'th-order linear model to a first-order model. For example, consider the following autoregressive process of order 2:

$$X_t = A_1 X_{t-1} + A_2 X_{t-2}$$

(This is called autoregressive since X_t is computed by using linear regression applied to 'older' values of itself.) This can be represented as a first-order model as follows:

$$\tilde{X}_t \stackrel{\text{def}}{=} \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} = \begin{pmatrix} A_1 & A_2 \\ I & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \end{pmatrix} \stackrel{\text{def}}{=} \tilde{A}\tilde{X}_{t-1}$$

Now we should estimate each block in $\tilde{A}$ separately. (Note that, in contrast to the case of discrete state spaces discussed in Model selection, the number of parameters increases *linearly* with k, not exponentially.) The most common way to choose k is to use the AIC scoring metric; this is built in to many software packages.

For a nonlinear model, we can use a tapped delay line, i.e. we feed a window of the last k time slices into f. (Note that this does not exploit the prior knowledge that successive input vectors (windows) are overlapping, and hence highly correlated.) In this case, k is usually chosen by cross-validation.

PARTIALLY OBSERVED MODELS

Consider trying to predict a speech signal, represented, for example, as an m-dimensional vector encoding the instantaneous power of the signal in m different frequency bands (typically $m \approx 20$). One approach would be to build a nonlinear model to directly predict Y_{t+1} as a function of Y_t, Y_{t-1}, etc. Such a model would probably need a very large number of parameters.

An alternative approach would be to try to infer the word that the person is saying, based on the acoustic evidence, then predict the next word that they are likely to say, and finally predict the corresponding sound. (In practice, it is more common to work with phonemes instead of words, since phonemes have less acoustic variability, and suffer less from the sparse data problems mentioned previously.) This is the basis of the hidden Markov model (HMM) approach to speech recognition (Rabiner, 1989; Jelinek, 1997), which has proved to be quite successful in practice. We will discuss how to learn such models later.

Another reason to consider models with hidden variables is that often we are more interested in inferring the hidden causes of a signal than in predicting the signal itself. As in Bayesian networks, a good approach is to build a generative model of the observed signal (a mapping from X_t to Y_t) and then to apply Bayes' rule to 'invert the causal arrow', i.e. to infer $P(X_t|y_{1:t})$. We will discuss how to do this below. (*See* **Bayesian Belief Networks**)

The EM Algorithm

Parameter learning in the case of partially observed models is much harder than in the case of fully observed models because the likelihood surface can have multiple maxima. (A simple way to see this is that we can permute the labels of the hidden states without affecting the likelihood function.) One approach is to use gradient ascent (e.g. Binder *et al.*, 1997), but there is a simpler and often more effective algorithm called expectation maximization (EM) (Dempster *et al.*, 1977; Neal and Hinton, 1998). EM is guaranteed to converge to a local maximum of the likelihood (or MAP) surface, and works roughly as follows. We first perform probabilistic inference to try to estimate the hidden state given the observation sequence (this is the E step). We then use the estimated values of the hidden variables as if they were observed, in order to update the parameters using the techniques discussed for fully observed models (this is the M step). Since the inferred values depend on the parameter settings, we then need to re-estimate the hidden values, and then re-estimate the parameters, repeating this until convergence. We will explain the EM algorithm in more detail below.

(Note: when applied to HMMs, EM is also known as the Baum-Welch algorithm.)

Discrete State Spaces: Hidden Markov Models

A hidden Markov model (HMM) is a partially observed Markov process. The dynamics of the hidden variable X_t are modeled using a standard finite state Markov chain; the observed variable Y_t is some function of X_t. For example, in speech recognition, in which $Y_t \in \mathbb{R}^m$, we assume that the distribution of Y_t is Gaussian (or perhaps a mixture of Gaussians), whose parameters are determined by the hidden state X_t: $P(Y_t = y|X_t = i) = N(y; \mu_i, \sum_i)$. If Y_t is discrete, we specify the observation model using another matrix: $P(Y_t = j|X_t = i) = B(i,j)$.

The complete log-likelihood of a sequence is given by

$$\log P(X_{1:T}, Y_{1:T}) = \log P(X_1) + \sum_{t=1}^{T} \log P(Y_t|X_t) + \sum_{t=2}^{T} \log P(X_t|X_{t-1})$$

Using the trick discussed in Parameter estimation, we may rewrite the last term as

$$\sum_{t=2}^{T} \sum_{i=1}^{S} \sum_{j=1}^{S} I(X_{t-1} = i, X_t = j) \log A(i,j)$$

and similarly for the other terms.

We cannot maximize this directly, since X_t is hidden. Instead, we will try to maximize the expected complete-data log-likelihood. Taking expectations with respect to the hidden variables X_t, using the current parameter settings θ, we find

$$E\ \log P(X_{1:T}, Y_{1:T}) = \ldots + \sum_{i=1}^{S} \sum_{j=1}^{S} E\left[\sum_{t=2}^{T} I(X_{t-1} = i, X_t = j)\right] \log A(i,j)$$

where we have omitted terms not involving A for simplicity. The terms inside the square brackets are the expected number of times we see an $i \to j$ transition:

$$E(i,j) \stackrel{\text{def}}{=} E\left[\sum_{t=2}^{T} I(X_{t-1} = i, X_t = j)\right] = \sum_{t=2}^{T} P(X_{t-1} = i, X_t = j|y_{1:T})$$

If we could compute these expected sufficient statistics (ESS), we could update the parameters to

$$\hat{A}_{ML} = \frac{E(i,j)}{\sum_k E(i,k)}$$

by analogy with the fully observed case. We could then use the new parameters to recompute the ESS, and so on, until convergence.

The forwards-backwards algorithm

We will now explain how to compute $\gamma_t(i) \stackrel{\text{def}}{=} P(X_t = i|y_{1:T})$ and $\xi_t(i,j) \stackrel{\text{def}}{=} P(X_{t-1} = i, X_t = j|y_{1:T})$ in $O(TS^2)$ time and $O(TS)$ space using dynamic programming, where T is the length of the sequence and S is the number of states.

In the forwards pass, we compute

$$\begin{aligned}\alpha_t(j) &\stackrel{\text{def}}{=} P(X_t = j, y_{1:t}) \\ &= P(y_t|X_t = j, y_{1:t-1})\ P(X_t = j|y_{1:t-1}) \\ &= P(y_t|X_t = j) \sum_i P(X_t = j|X_{t-1} = i, y_{1:t-1}) \\ &\quad P(X_{t-1} = i|y_{1:t-1}) \\ &= P(y_t|X_t = j) \sum_i A(i,j)\, \alpha_{t-1}(i)\end{aligned}$$

Notice how this algorithm contains the two key steps of sequential Bayesian updating: predict (i.e. computing the one step-ahead prediction, $P(X_t|y_{1:t-1})$), and update (combining the prediction with the likelihood $P(y_t|X_t)$ to get the updated estimate $P(X_t|y_{1:t})$).

If we define a diagonal matrix B_t in which $B_t(j,j) \stackrel{\text{def}}{=} B(j, y_t)$, we may rewrite the above equation as a simple vector-matrix multiplication: $\alpha_t = B_t A' \alpha_{t-1}$.

In the backwards pass, we compute

$$\begin{aligned}\beta_t(i) &\stackrel{\text{def}}{=} P(y_{t+1:T}|X_t = i) \\ &= \sum_j P(X_{t+1} = j|X_t = i) \\ &\quad P(y_{t+1:T}|X_t = i, X_{t+1} = j) \\ &= \sum_j P(X_{t+1} = j|X_t = i)\, P(y_{t+1}|X_{t+1} = j) \\ &\quad P(y_{t+2:T}|X_{t+1} = j) \\ &= \sum_j A(i,j)\ P(y_{t+1}|X_{t+1} = j)\ \beta_{t+1}(j)\end{aligned}$$

Again, we may rewrite this as a simple vector-matrix multiplication: $\beta_t = AB_{t+1}\beta_{t+1}$.

Finally, we combine the results of the forwards and backwards steps to obtain

$$\gamma_t(i) \stackrel{\text{def}}{=} P(X_t = i|y_{1:T}) \propto P(y_{t+1:T}|X_t = i, y_{1:t})$$
$$P(X_t = i, y_{1:t}) = \beta_t(i)\ \alpha_t(i)$$

and

$$\xi_t\ (i,j) \stackrel{\text{def}}{=} P(X_{t-1} = i, X_t = j|y_{1:T})$$
$$\propto P(X_{t-1} = i|y_{1:t-1})\ P(X_t = j|X_{t-1} = i)$$
$$P(y_t|X_t = j)\ P(y_{t+1:T}|X_t = j)$$
$$= \alpha_{t-1}(i)\ A(i,j)\ B_t(j,j)\ \beta_t(j)$$

The ESS for A can be computed using $E(i,j) = \sum_{t=2}^{T} \xi_t(i,j)$, and similarly for the other parameters.

The Viterbi algorithm

A useful task in a variety of applications is to compute the most probable state sequence given the observations: $\arg\max_{x_{1:T}} P(x_{1:T}|y_{1:T})$. For example, in speech recognition, this might be the most likely sequence of words given the acoustic signal, and in biology, it might be the most likely alignment of a protein sequence to some model. This quantity can be computed using the Viterbi algorithm, which is like forwards-backwards except that we replace the $\sum$ operator with max in the forwards pass, and do pointer-following in the backwards pass: see Rabiner (1989) for details.

Classifying sequences using HMMs

The normalization constant used to compute $\gamma_t(i) = P(X_t = i|y_{1:T})$ is equal to the likelihood of the data:

$$P(y_{1:T}) = \sum_i P(y_{t+1:T}|X_t = i, y_{1:t})\ P(X_t = i, y_{1:t})$$

This is a useful quantity, because it can be used to classify a sequence. For example, suppose we train up 10 HMMs, one for each spoken digit 0 to 9; given a test sequence $y_{1:T}$, we compute the likelihood that HMM i generated $y_{1:T}$; we then classify $y_{1:T}$ as digit i if HMM i has higher likelihood than any of the rival models.

Factored state spaces

An HMM represents the state of the system using a single discrete random variable X_t. To represent k bits of information, we need to use 2^k values; this results in high computational complexity (i.e. inference using forwards-backwards is slow) and high sample complexity (i.e. the amount of data needed by EM is large). An alternative is to use a distributed representation of state, i.e. to use k binary variables. Such a model is called a factorial HMM (Ghahramani and Jordan, 1997), and is shown in Figure 2. This model has exponentially fewer parameters, and hence is easier to learn. Unfortunately, inference is still slow, because the k hidden random variables become correlated: intuitively, they 'compete' to 'explain' the observation, a phenomenon called 'explaining away' (Pearl, 1988). However, it is possible to exploit the structure of the model to enable efficient approximations.

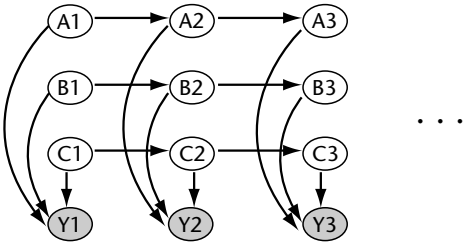


Figure 2. A factorial HMM. The hidden state is represented in a distributed fashion, in terms of three discrete random variables, $X_t = (A_t, B_t, C_t)$.

A factorial HMM is an example of a more general class of models called dynamic Bayesian networks (Dean and Wellman, 1991; Ghahramani, 1998; Murphy, 2002).

Continuous State Spaces

Linear systems and the Kalman filter

A linear dynamical system (LDS) has the generic form

$$X_t = AX_{t-1} + BU_{t-1} + W_t$$

$$Y_t = CX_t + DU_t + V_t$$

where $W_t \sim N(0, Q)$ and $V_t \sim N(0, R)$ are independent random variables representing Gaussian noise.

The Kalman filter is an algorithm to compute $P(X_t|y_{1:t}, u_{1:t})$ in an LDS. It is analogous to the forwards algorithm for HMMs (Minka, 1999). (In particular, it contains the same kind of predict-update cycle, which, it has been claimed (Rao, 1997), has analogs in the brain in terms of top-down and bottom-up visual processing.) For off-line learning, one can get better performance by estimating the hidden states based on 'future' data, as well as past, i.e. by using the 'smoothed' quantities $P(X_t|y_{1:T}, u_{1:T})$. The Rauch–Tung–Striebel (RTS) algorithm is a way to compute $P(X_t|y_{1:T}, u_{1:T})$ in an LDS, and is analogous to the forwards-backwards algorithm for HMMs. In the general case, the RTS algorithm takes $O(Tm^3)$ time and $O(Tn^2)$ space, where $Y_t \in \mathbb{R}^m$ and $X_t \in \mathbb{R}^n$, because at each step, the algorithm must invert an $m \times m$ matrix and must store an $n \times n$ covariance matrix. To do

parameter estimation in an LDS, we can use the EM algorithm, using the RTS algorithm as a subroutine to compute the required expected sufficient statistics (Roweis and Ghahramani, 1999).

Nonlinear systems

Inference, and therefore learning, is much harder in systems which are nonlinear and/or have non-Gaussian noise, and one typically has to resort to approximate methods. A popular method for computing $P(X_t|y_{1:t})$ is particle filtering (Doucet *et al*., 2001) or the extended Kalman filter (Bar-Shalom and Fortmann, 1988); for off-line computation of $P(X_t|y_{1:T})$, one can use Gibbs sampling (Gilks *et al*., 1996) or the extended Kalman smoother (Roweis and Ghahramani, 2001). These inference routines can be used as subroutines for the E step of EM.

References

Bar-Shalom Y and Fortmann T (1988) *Tracking and Data Association*. Academic Press.

Binder J, Koller D, Russell SJ and Kanazawa K (1997) Adaptive probabilistic networks with hidden variables. *Machine Learning* **29**: 213–244.

Bishop CM (1995) *Neural Networks for Pattern Recognition*. Clarendon Press.

Dean T and Wellman M (1991) *Planning and Control*. Morgan Kaufmann.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **34**: 1–38.

Doucet A, de Freitas N and Gordon NJ (2001) *Sequential Monte Carlo Methods in Practice*. Springer Verlag.

Ghahramani Z (1998) Learning dynamic Bayesian networks. In: Giles C and Gori M (eds) *Adaptive Processing of Sequences and Data Structures. Lecture Notes in Artificial Intelligence*, pp. 168–197. Springer-Verlag.

Ghahramani Z and Jordan M (1997) Factorial hidden Markov models. *Machine Learning* **29**: 245–273.

Gilks W, Richardson S and Spiegelhalter D (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Heckerman D (1998) A tutorial on learning with Bayesian networks. In: Jordan M (ed.) *Learning in Graphical Models*. MIT Press.

Jelinek F (1997) *Statistical Methods for Speech Recognition*. MIT Press.

Ljung L (1987) *System Identification: Theory for the User*. Prentice Hall.

MacKay D and Peto L (1995) A hierarchical Dirichlet language model. *Natural Language Engineering* **1**(3): 1–19.

Minka T (1999) *From Hidden Markov Models to Linear Dynamical Systems*. Technical report, MIT.

Murphy K (2002) *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley.

Neal RM and Hinton GE (1998) A new view of the EM algorithm that justifies incremental and other variants. In: Jordan M (ed.) *Learning in Graphical Models*. MIT Press.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2): 257–286.

Rao P (1997) Kalman filter model of the visual cortex. *Neural Computation* **9**(4): 721–763.

Roweis S and Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Computation* **11**(2): 305–345.

Roweis S and Ghahramani Z (2001) Learning nonlinear dynamical systems using the EM algorithm. In: Haykin S (ed.) *Kalman Filtering and Neural Networks*. Wiley.

Saul L and Jordan M (1999) Mixed memory Markov models: decomposing complex stochastic processes as mixture of simpler ones. *Machine Learning* **37**(1): 75–87.

Means–Ends Analysis

Introductory article

Wayne Iba, Kanisa Inc., Cupertino, California, USA

CONTENTS

Means–ends analysis is a simple, yet powerful, approach to problem-solving whereby progress towards a goal or solution is achieved by incrementally searching a state space.

DIFFERENCE REDUCTION IN A STATE SPACE

The basic idea of means–ends analysis (MEA) relies on two notions. First, the problem solving domain may be represented and explored as a state space. Specifically, the state of the world at the start of the problem-solving exercise, the goal state in which the problem has been 'solved', and any intermediate states, can all be represented within a state space. Second, MEA takes advantage of a correspondence between the state space representation and the definitions of the operators that transform the world from one state to another.

Representing the World

In many domains, one can solve problems by representing the world as a state consisting of propositions and relations. That is, the important aspects of the world – those that are relevant to the problem at hand – can be captured as a set of objects, predicates describing these objects' characteristics, and relations that exist between these objects.

For example, in the context of a world where structures such as towers and arches must be constructed from toy building blocks, one particular state of the world could be captured as: `((block A) (block B)(block C)(on-table A) (on-table C)(on B C))`. This state description ignores characteristics such as color and size, but captures the notion that there are three blocks, *A*, *B*, and *C*, that blocks *A* and *C* are on the table, and that block *B* is stacked on top of block *C*.

However, representing the state of the world is only the first step in establishing a state space. We also need to represent the operators that transform one state into another. In traditional state space representations for MEA, operators are represented as a set of preconditions and consequences. The preconditions specify properties and relationships that must be true in a state in order to apply the given operator. The consequences specify the properties and relationships that become true and are added to the state, as well as those that are no longer true and are deleted from the state, when the operator is applied.

For a given representation language and a finite set of objects, we can imagine the set of all possible state descriptions. Then, for a finite set of operators, we can construct directed links between pairs of states, according to whether the preconditions of the operator are true in the initial state and the differences between the initial state and the resulting state correspond to the consequences of the operator.

Goals and Sub-goals

Given a state space description of a problem-solving domain, a specific problem to solve can be thought of as a pair of states: the current state of the world, and the desired, or goal state. Stages of solving the problem can be thought of as sub-goals. For example, if your goal is to buy a car, one solution might be to get some money, go to the car showroom, and drive home in a car. Your initial state is without car and your goal state is with car. A sub-goal is to get some money, which might involve activities like getting a job and saving your earnings. Each of these may also involve sub-goals such as applying for and securing a well-paid job and opening a bank account. Sub-goal formation is guided by the preconditions of higher-level goals. For example, a precondition of getting a well-paid job may be having a good education, so getting a good education may become a new sub-goal.

At the end of a problem-solving activity, the result is a path from state to state through the state space. This path corresponds to a sequence of actions each of which accomplishes the transformation of one state into the next. Any state in this path may be thought of as both a sub-goal and as the initial state for some other sub-goal.

FORWARD AND BACKWARD REASONING

Searching for Solutions

Having described world states, operators, state spaces, and solutions, we now turn to the task of finding a solution (i.e. a path through the state space from the initial problem state to the goal state). This task is best thought of as a search problem. Techniques such as 'breadth-first' and 'depth-first' search exhaustively explore a search space until the goal is found. In the context of problem-solving this amounts to a search through the space of paths, where the desired item is a path that starts at our initial state and ends at the goal state. Unfortunately, itemizing and organizing the space of all paths through the state space is prohibitively expensive; thus, short cuts are needed.

Typically, a search scheme will construct a path incrementally by starting at the initial state, considering states reachable from the current one, and repeating this process until the goal state is found. This approach is known as forward search, and more generally 'forward reasoning'. Alternatively, 'backward reasoning' employs a scheme that starts at the goal state, considers all possible states that can reach the goal state in question, and repeats this process until the problem's initial state is found. Regardless of the direction of search, the biggest limitation for solving problems of any significant complexity is the size of such a search, which increases exponentially. For example, if there are just 10 reachable states from each state, there are 10^{10} (10 billion) paths of length 10. Both the branching factor and the length of solution are often much larger than 10: in fact, it is not uncommon to search a space that contains more states than the number of elementary particles in the universe. Thus, problem-solving schemes must employ heuristics, or rules of thumb, which selectively guide the consideration of states.

Search Strategies

Typical search strategies prioritize the consideration of possibilities by their cost or by their effectiveness. Usually, the cost of an action is known directly, but the effectiveness of an action, with respect to progress towards the goal, must be estimated by heuristics.

MEA integrates forward and backward search, and estimates of operator effectiveness, by selecting operators based on the differences that the operators reduce. Differences exist between two states when a predicate or relation between objects is true in one state but not the other. When the current state has been transformed into a state that has no differences from the goal state, the search has found a solution. For example, suppose there are 20 possible actions that can be executed in a current state, but the consequences of only one of them changes the state in a way that reduces the differences between the current state and the goal state. It makes sense to consider this action first before any of the others. Of course, it is not guaranteed that this action will actually participate in an eventual solution: one of the other operators may establish a precondition for some other action that makes more effective progress towards the goal. In MEA, when the strategy fails to yield progress, the search reverts to 'weak methods', which require no domain-specific knowledge.

MEA's search scheme is neither forward nor backward, but rather a blend of both. Based on differences, a selected action might transform the current state or reach the goal state. Alternatively, however, it might 'float' somewhere in the middle, creating a new sub-goal determined by the action's preconditions and a new initial state determined by the action's consequences. In this case, MEA then tackles two smaller problems: find a path from the initial state to the preconditions of the selected action, and find a path from the consequences of the action to the goal state. Finally, these two solutions can be joined by the originally selected action, thus yielding a complete solution to the problem.

OPERATOR–DIFFERENCE CORRESPONDENCE

MEA works by establishing a correspondence between an operator's consequences and state-pair differences. When the preconditions and consequences of operators are explicitly defined by their state features, it is straightforward to establish this correspondence: the state-pair's differences (predicates and relations that must hold in the goal state but that do not hold in the initial state) can be compared to the operator's consequences (also described as predicates and relations); and the operator that resolves the most differences is selected.

If instead the actions are not directly inspectable, background domain knowledge must be provided in the form of an operator–difference table. For each operator, this table specifies the differences it reduces. Now instead of comparing an action's consequences to a set of differences, the differences are used directly to look up actions in the table.

Sometimes, it is also desirable to specify preferences among differences or to attach relative importance to particular differences. For example, in a problem-solving domain such as our earlier car-buying scenario, first reducing differences in resource availability (bank account balance) may more effectively guide the search for a solution than focusing initially on differences such as being at a dealer, or deciding on a model or color. This kind of domain-specific knowledge can be conveniently included in the operator–difference table.

RELATION TO OTHER SEARCH TECHNIQUES

Heuristic Search

The idea of heuristic search is to focus exploration of the state space in a particular direction. Often, the heuristic avoids exponential search sizes by estimating the total cost of a solution based on the actual cost of the partial solution found thus far combined with the expected cost for completing the partial solution. The primary heuristic in MEA is using differences to guide search: instead of estimating the cost of a solution, the operator–difference heuristic estimates which solutions would be found more quickly, by reducing the biggest differences first. Thus, MEA is an instance of a particular type of heuristic search known as 'greedy' search.

Exhaustive Search

Heuristic search is not guaranteed to generate an efficient solution, nor even to find an inefficient solution quickly. However, because of MEA's ability to fall back on weak methods, it is guaranteed to find a solution eventually if one exists. Thus, MEA may be thought of as an exhaustive search where the order in which operators are considered is dynamically determined.

MEANS–ENDS ANALYSIS IN HUMAN REASONING

The MEA approach was developed to model and study human problem-solving. The working hypothesis was that humans solve problems using the operator–difference heuristic. To validate this hypothesis, the technique of 'protocol analysis' has been employed to observe and record the reasoning behavior of a human in the process of solving a problem. In this manner, mental problem-solving can be inspected and a virtual search trace can be extracted; this search trace can be compared with that generated by MEA. The stronger the correspondence between the two search traces, the more confident we are that the working hypothesis is correct. In experiments, the fit between protocol analyses and means–ends traces has been compelling, particularly in puzzles such as 'missionaries and cannibals', the Tower of Hanoi, and various water jug problems. These puzzles, for which this correspondence has been demonstrated, are difficult for people because their solutions can be a lengthy sequence of operators but are also tractable for protocol analysis because they inhabit a narrowly constrained state space.

SYSTEMS

GPS

The 'general problem-solver' (GPS) contained the first implementation of means–ends analysis and was designed as a theory of human problem-solving. It was intended to work on any problem that could be expressed using a state space and a set of operators. However, in order to move beyond very simple problems it relied on an operator–difference correspondence table for each problem. GPS exerted a strong influence on both cognitive science and artificial intelligence planning and problem-solving systems. It was also, in a sense, a precursor to Soar, both because of its conceptual framework and because of the influence and involvement of Allen Newell.

Prodigy

The Prodigy framework is a GPS-like system that integrates several machine learning methods. Prodigy's architecture encompasses problem-solving, planning, and learning. Several learning methods are included, such as explanation-based generalization, abstraction learning, and derivational analogy. When Prodigy has solved a particular problem, it might decide to store the solution trace, depending on how 'interesting' the system considers the solution to be. If the built-in heuristics predict that the solution will be useful in solving similar problems, and the solution is sufficiently

distinct from others that have already been stored, the trace will be saved for possible application during future problem-solving. Given a new problem to solve, if the initial and goal conditions are sufficiently similar to those of a saved trace, then the saved solution can be used as if it were a single operator with its own preconditions and consequences.

ICARUS

The ICARUS architecture also incorporates a general problem solver and learning, with many of the strengths of GPS and of Prodigy. The goal of ICARUS was to reduce the amount of domain-specific knowledge required, such as the operator–difference correspondence table. Instead of combining several learning methods to obtain flexible improvements in performance, ICARUS includes a single architectural learning method: probabilistic hierarchical concept formation. As with GPS and Prodigy, means–ends analysis serves as the basic problem-solving method. However, instead of relying on domain-specific knowledge, such as an operator–difference table, ICARUS learns the relevant knowledge through experience. Instead of providing derivational analogy or other special-purpose mechanisms, this architecture relies on the combination of its hierarchical memory organization and its memory retrieval mechanism to get analogical reasoning and other behaviors to emerge from the system's innate behavior.

Further Reading

Carbonell JG, Knoblock CA and Minton S (1989) PRODIGY: an integrated architecture for planning and learning. Techical Report CMU–CS–89–189. Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Langley P, McKusick KB, Allen JA, Iba WF and Thompson K (1991) A design for the ICARUS architecture. *SIGART Bulletin* **2**: 104–109.

Langley P, Thompson K, Iba W, Gennari J and Allen JA (1989) An integrated cognitive architecture for autonomous agents. Technical Report 89–28. University of California, Department of Information and Computer Sciences, Irvine, CA.

McDermott D (1996) A heuristic estimator for means–ends analysis in planning. In: Drabble B (ed.) *Proceedings of the Third International Conference on Artificial Intelligence Planning Systems* (AIPS96), pp. 142–149. Menlo Park, CA: AAAI Press.

Minton S (ed.) (1992) *Machine Learning Methods for Planning and Scheduling*. San Francisco, CA: Morgan Kaufmann.

Newell A and Simon H (1963) GPS, a program that simulates human thought. In: Feigenbaum EA and Feldman J (eds) *Computers and Thought*, pp. 279–293. New York, NY: McGraw-Hill.

Simon HA and Newell A (1971) *Human Problem Solving*. Princeton, NJ: Prentice-Hall.

Modeling Coordinate Transformations

Introductory article

Alexandre Pouget, University of Rochester, Rochester, New York, USA
Lawrence H Snyder, Washington University, St Louis, Missouri, USA

CONTENTS

Sensorimotor transformations can often be decomposed into a series of intermediate coordinate transforms. These may be modeled using units computing a product of sigmoidal tuning curves to sensory and posture signals.

INTRODUCTION

Coordinate transforms occur mostly in the context of sensorimotor transformations; i.e. the transformation of sensory signals into motor commands. For instance, to reach for an object in view, the brain needs to compute the changes in angles of the joints of the arm that bring the hand to the location of the object. This change of joint angles is obtained by computing the difference between two vectors: the current and the desired position of the hand, both expressed in terms of joint angles; i.e. in joint-centered coordinates. The current position of the hand is provided by visual and proprioceptive signals, but the desired position can be computed only by combining the information provided by the visual system – the retinal, or eye-centered, coordinates of the object – with posture signals, such as the position of the eyes in the head (eye position for short) and the position of the head with respect to the trunk (head position for short). Therefore, reaching requires a transform from eye-centered to joint-centered coordinates. Similar coordinate transforms take place in all other sensorimotor transformations.

LINEAR COORDINATE TRANSFORMS

In many cases, sensorimotor transformations can be decomposed into a series of intermediate coordinate transforms. For instance, instead of directly computing the change in joint-centered coordinates required to reach an object from its eye-centered coordinates, one can first compute the head-centered coordinates of the target, then the body-centered coordinates, and finally the joint-centered coordinates (Figure 1(a)). In this example, many of these transformations are linear (Figure 2): they simply involve summation of the signals involved. For example, the transformation from eye-centered to head-centered is (approximately) linear because the head-centered location of an object ($\mathbf{H}$) is approximately equal to the sum of the eye-centered location ($\mathbf{V}$) of the object and the current eye position ($\mathbf{E}$): $\mathbf{H} = \mathbf{V} + \mathbf{E}$ (we use bold type to denote vectors).

For these transformations we may consider a linear representational scheme which involves the coding of each vector, such as $\mathbf{V}$ and $\mathbf{E}$, through the activity of N units, where each unit computes a dot product between the unit's preferred direction ($\mathbf{V}_i$) and the vector to be encoded ($\mathbf{V}$) – see eqn [1] for the definition of a dot product. For instance, in the case of the two-dimensional retinal location of an object (we ignore the position of the object in depth and consider only the horizontal and vertical position, noted respectively, v^x and v^y, for vector $\mathbf{V}$), the activity (a) of the ith unit takes the form

$$\begin{aligned} a_i &= \mathbf{V}_i \cdot \mathbf{V} \\ &= v_i^x v^x + v_i^y v^y \\ &= \|\mathbf{V}_i\| \|\mathbf{V}\| \cos(\theta_i - \theta) \end{aligned} \quad (1)$$

where $\|\mathbf{V}\|$ is the Euclidean norm, or length, of $\mathbf{V}$, θ is the angle of the retinal vector, and θ_i is the preferred retinal direction for neuron i. The second line uses cartesian coordinates, while the third line uses polar coordinates. Redish and Toureztky call this representation a *sinusoidal array* to refer to the cosine tuning of the units (as can be seen

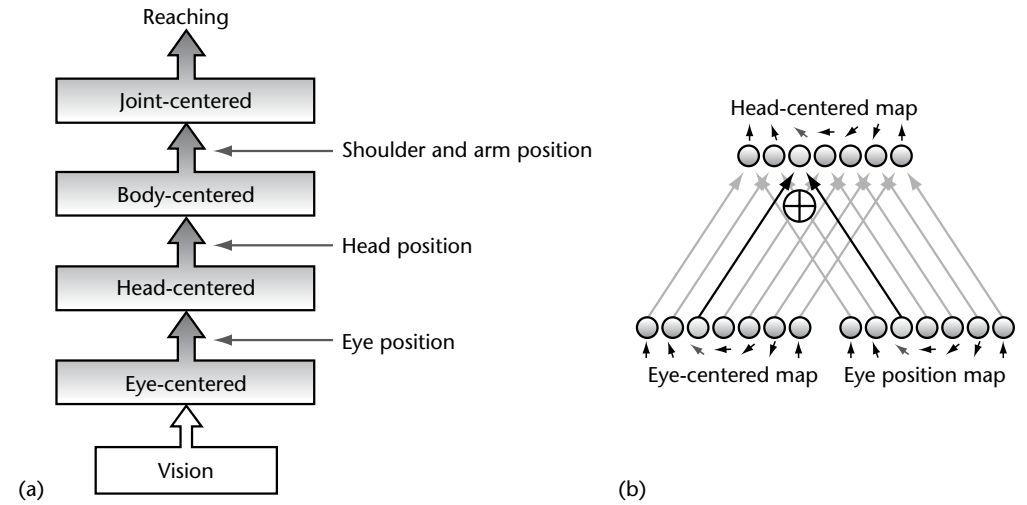


Figure 1. (a) During the process of reaching for an object, the cortex must transform the eye-centered coordinates of the object into the change of joint angle of the arm (joint-centered coordinates). This transformation can be decomposed into a series of coordinate transforms in which the position of the object is successively recoded into a set of intermediate frames of reference. Some of these transformations are linear in the sense that they require only a sum of the signals involved. For example, the head-centered location of an object is simply equal to the sum of the eye-centered and eye position signals. (b) A network implementation of the linear transformation from eye-centered to head-centered. Each vector is encoded with a sinusoidal array. For instance, the current eye-centered location of the object is encoded through the activity of N units. Each unit has a cosine tuning to the eye-centered location of the object and is characterized by a preferred direction (indicated by the small arrows). The preferred directions are uniformly distributed along a circle. When the units are activated with the current eye-centered position of the object, the pattern of activity across the neuronal array follows a cosine profile due to cosine tuning curves. The amplitude of this cosine pattern of activity encodes the eye-centered eccentricity of the stimulus, $\|\mathbf{V}\|$, while the phase encodes the angle of the eye-centered vector, θ. This representation is quite compact since it manages to encode a two-dimensional vector, $\mathbf{V}$, with a single-dimensional array. A similar code is used in the other maps. Each unit in the head-centered map takes the sum of two units with the identical preferred direction in the eye-centered and eye position maps (an example of which is shown in darker shading). This connectivity ensures that the network performs a vector addition.

on the last line of eqn [1]). Note that if we sum the activities of two neurons, one tuned to $\mathbf{V}$ and the other tuned to $\mathbf{E}$ with identical preferred directions ($\mathbf{V}_i = \mathbf{E}_i = \mathbf{H}_i$), we obtain a neuron tuned to $\mathbf{H}$ with the same preferred direction $\mathbf{H}_i$

$$\begin{aligned} a_i &= a_i^V + a_i^E \\ &= \mathbf{H}_i \cdot \mathbf{V} + \mathbf{H}_i \cdot \mathbf{E} \\ &= \mathbf{H}_i \cdot (\mathbf{V} + \mathbf{E}) \\ &= \mathbf{H}_i \cdot \mathbf{H} \end{aligned} \qquad (2)$$

Therefore, linear coordinate transforms can be performed with simple addition of activities (Figure 1(b)). Whether the brain uses a similar scheme remains controversial. A key piece of information, currently lacking, is whether individual neurons in areas such as the posterior parietal cortex have retinal and postural signals tuned to the same preferred directions. Colinear tuning would be an essential ingredient of a coding scheme in which transformations were performed by summing the activities of the neurons (eqn [2]), but to date no evidence has been obtained for such colinearity in cortical neurons.

It has been argued that the primary motor cortex uses sinusoidal arrays to encode the direction of hand movements, but this has never been thoroughly tested. Thus, Sanger has pointed out that the statistical test used by Georgopoulos *et al.* cannot distinguish between cosine tuning curves and almost any other alternative tuning curves. The evidence for cosine tuning functions is more convincing in the nervous systems of leeches and crickets, suggesting that sinusoidal arrays might provide a good conceptual framework for sensorimotor transformations in invertebrates.

NONLINEAR COORDINATE TRANSFORMS

Although some of the intermediate transforms involved in computing the change in joint angles of the arm are linear, this is not true for all sensorimotor transformations. In particular the

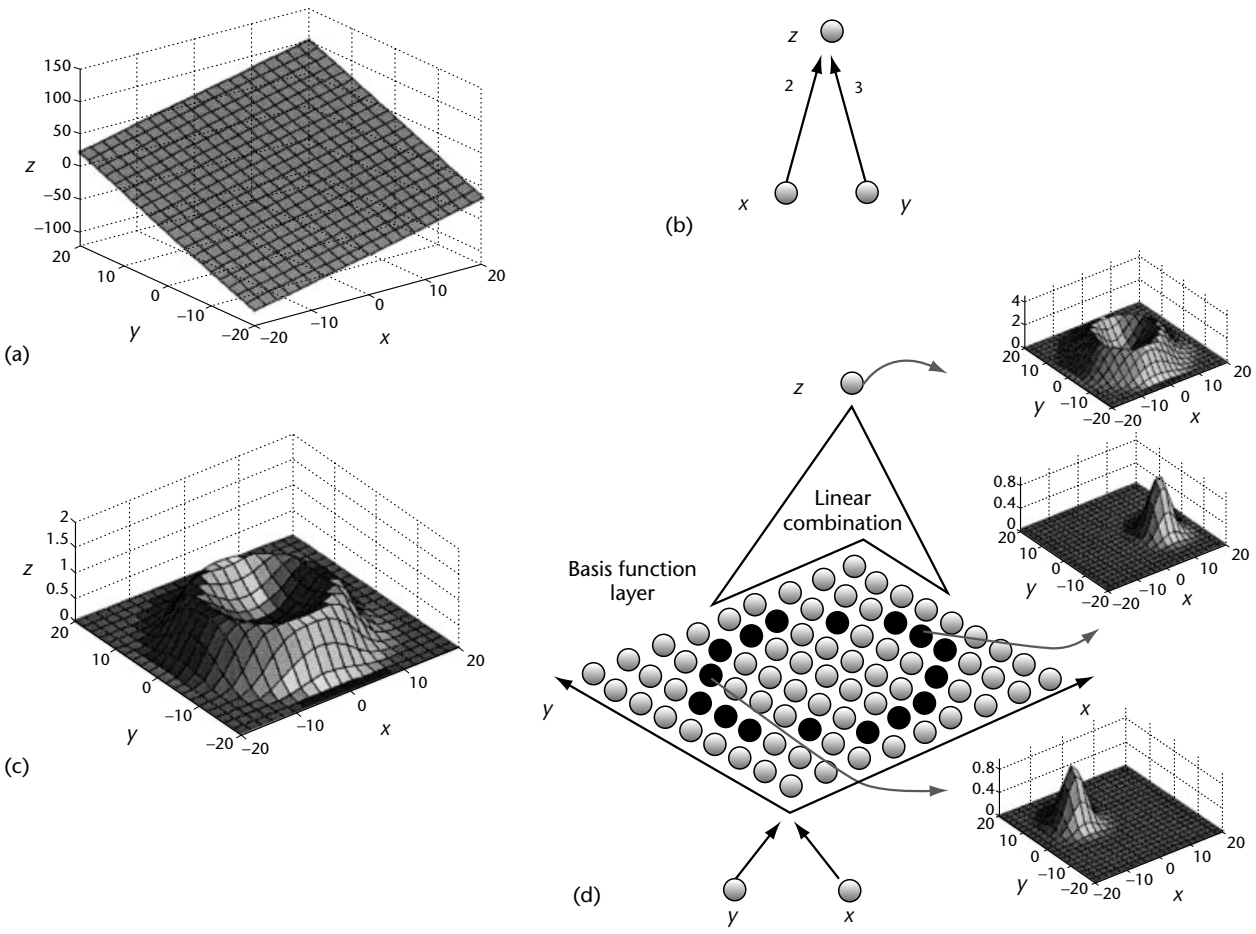


Figure 2. Linear versus nonlinear functions. (a) A function is said to be linear if it can be written as a weighted sum of its input variables plus a constant. All other functions are called nonlinear. For instance, $z = 2x + 3y$ is a linear function of x and y. This linear function forms a planar surface (as do all other linear functions). (b) Linear functions can be implemented in two-layer networks. The network shown corresponds to the linear function in (a). The output unit simply takes the sum of the input unit activity weighted by the coefficient indicated next to the connections. (c) A plot of the nonlinear function $z = \exp(-(x^2 + y^2 - 100)^2/1000)$. Nonlinear functions can take arbitrary shapes, in this case a circular ridge. (d) A neural network implementation of the nonlinear function shown in (c) using Gaussian basis functions in the intermediate representation. The basis function units are organized to form a map in the x–y plane. Two representative response functions of these basis function units are shown on the right (middle and bottom). The activity of the output unit is obtained by taking a linear sum of the basis function units. In this example, the weights of the black units onto the output unit are set to 1, while all the other units have a weight of 0. As a result, the output unit adds a set of Gaussian functions arranged along a circle in the x–y plane. This leads to the response function illustrated on the right (top), which is similar to the circular ridge shown in (c). The same idea can be applied to sensorimotor transformations. Unlike the examples above, sensorimotor transformations involve more than two input variables (e.g., x and y retinal position from each retina, plus many postural signals). Similarly, the output is also multidimensional. Yet, these high-dimensional transformations can be computed in a way exactly analogous to the low-dimensional computation illustrated here. One advantage of the basis function layer is that it can be used to control, and coordinate, several behaviors aiming at the same object. Indeed, motor commands are nonlinear functions of the sensory inputs and, as such, any motor command can be obtained by a linear combination of the basis functions.

transformation from body-centered coordinates to the required change in joint angles is nonlinear because it cannot be written as a weighted sum of the body-centered position and the current arm position. The exact equation involved in this case is complicated but the details do not matter for the theory we are about to develop. The critical point is that the transformation involves more than just simple summation. As a result, the overall transformation from eye-centered to joint-centered coordinates is also a nonlinear transformation. One way to approach nonlinear transformations is to

consider them in the general framework of nonlinear function approximation. Indeed, the coordinate transforms required for reaching are equivalent to the computation of a mathematical function which takes as input the visual position of the object and the posture signals, and produces as output the joint-centered coordinates of the object. Using the notation **V**, **P**, and **J** to refer to the visual position of the object, the posture signals and the change in joint angles, the coordinate transform can be expressed as a function f mapping **V** and **P** into **J**

$$\mathbf{J} = f(\mathbf{V}, \mathbf{P}) \tag{3}$$

Unlike linear functions, a nonlinear function cannot be implemented in two-layer networks. One needs at least one intermediate layer (or 'hidden' layer) to recode the sensory inputs before they can be transformed into motor commands (Figure 2(c) and (d)). One solution involves using intermediate units which compute basis functions. By definition, most functions of interest can be approximated using a linear combination of the functions from a basis set. The classic example of a basis set is all possible cosine and sine functions. This is the principle behind the Fourier transform: any function can be expressed as the sum of a series of cosine and sine functions with all possible amplitudes and frequencies. This result is not restricted to sine and cosine functions. Other families of functions such as Gaussian and sigmoid functions also form basis sets (Figure 2(d)).

When applied to sensorimotor transformations, and in particular to the example of reaching toward a visual target, the idea is as follows: the reaching motor command **J** can be obtained by a weighted sum of basis functions $\{B_i(\mathbf{V}, \mathbf{P})\}_{i=1}^{N}$ of the visual and posture signals, **V** and **P**

$$\mathbf{J} = \sum_{i=1}^{N} w_i B_i(\mathbf{V}, \mathbf{P}) \tag{4}$$

The set of weights, $\{w_i\}_{i=1}^{N}$, is specific to the reaching motor command being computed and, as we will see later, can be determined using simple learning rules.

Many choices are available for the basis functions. For instance, one can use the set of all Gaussian functions of **V** and **P**, which are a subset of a larger family known as radial basis functions (Figure 2(d)). An alternative choice for a basis set is the product of a Gaussian function of **V** and a sigmoid function of **P**. Figure 3(a) illustrates the response of an idealized neuron whose response function follows a Gaussian function of the eye-centered position of an object multiplied by a sigmoid function of eye position. Figure 3(b) shows a mapping of the receptive field of such a neuron for three different positions of the eyes. The receptive field is eye-centered; i.e. it remains at the same eye-centered location regardless of the position of the eye; however, the gain (or amplitude) of the response changes with the position of the eyes.

One of the problems from a biological point of view with eqn [4] is the format of the input and output vectors involved. For instance, we used a vector **V** which stores the three-dimensional polar coordinates of the object, yet no such explicit vector exists in the cortex. Instead, the visual position of objects is encoded by the activity of a large number of binocular neurons forming the retinotopic (or eye-centered) maps in the early visual areas. This does not mean that we cannot use the basis function framework; one simply needs to replace the vector **V** with a new vector, $\mathbf{V}^{\mathrm{A}}$, whose components are the activities (e.g. firing rates) of the neurons in response to the image of the object. This new vector has as many dimensions as there are neurons in the retinotopic maps – instead of three dimensions for **V**. Likewise, the vectors **P** and **J** can be replaced by the corresponding neuronal patterns of activities, $\mathbf{P}^{\mathrm{A}}$ and $\mathbf{J}^{\mathrm{A}}$. The function to be computed is now from $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{P}^{\mathrm{A}}$ onto $\mathbf{J}^{\mathrm{A}}$. When this function is nonlinear, which it is in this particular case, we can once again use the basis function approach. Figure 3(c) illustrates a basis function network which uses map-like representations, $\mathbf{V}^{\mathrm{A}}$, $\mathbf{P}^{\mathrm{A}}$, and $\mathbf{J}^{\mathrm{A}}$. Many neural network models of sensory motor transformations rely on such basis function representations in their intermediate layer.

Basis functions of $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{P}^{\mathrm{A}}$ can be used to compute $\mathbf{J}^{\mathrm{A}}$ as well as any other motor command (i.e. function) depending on $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{P}^{\mathrm{A}}$. One simply needs to adjust (or learn, see below) the weights $\{w_i\}$ in eqn [4] for the desired motor command. This means that basis function neurons are ideally placed to coordinate different behaviors depending on the same set of variables, such as moving the eyes and hand to the same object, despite the fact that these movements must be programmed using distinct coordinates.

Biological Plausibility of the Basis Function Approach

The basis function approach requires that the tuning curves of neurons in intermediate stages of computation provide a basis function set. Neurons whose response can be described by a Gaussian function of retinal location multiplied by

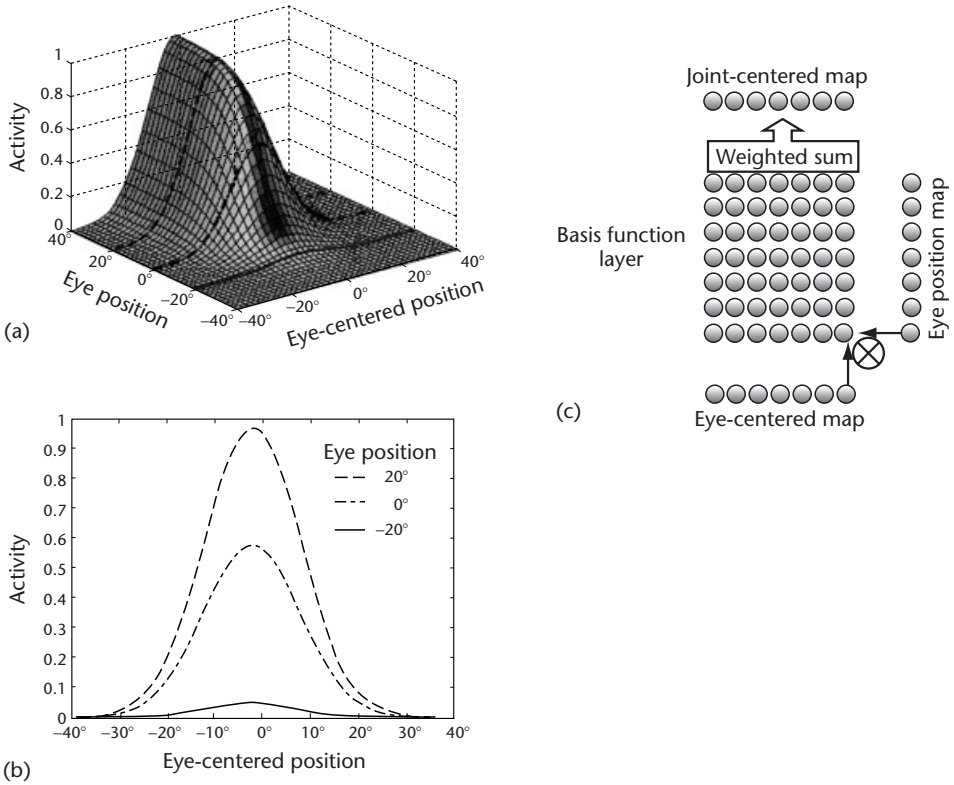


Figure 3. Basis function units. (a) The response function of a basis function unit computing a product of a Gaussian function of retinal location multiplied by a sigmoidal function of the position of the eyes. (b) A mapping of the retinotopic receptive field derived from a unit with the properties shown in (a) for three different eye positions. The bold lines in (a) correspond to the three curves shown here. The receptive field always peaks at the same retinal location but the gain (or amplitude) of the response varies with eye position. Gain modulations similar to this are found in numerous cortical areas, from V1 to the premotor cortex. (c) A neural network model for nonlinear sensorimotor transformations using basis functions. The input layer encodes the retinal location of an object and the current eye position while the output layer encodes the change in joint angles of the arm. The tuning curve of the eye-centered and eye position units are assumed to follow Gaussian and sigmoid functions, respectively. This nonlinear sensorimotor transformation requires an intermediate layer. In the case illustrated here, the intermediate layer uses basis function units. Each unit computes the product of the activity of one input unit from the eye-centered map and one input unit from the eye position map. As a result, the response of the basis function units is as shown in (a) and (b). The transformation from the basis function to the output units involves a simple linear transformation, namely a weighted sum of the activity of the basis functions units. This is the main advantage of this approach: once the basis functions are computed, nonlinear transformations become linear.

a sigmoidal function of eye position would qualify (see Figure 3(a) and (b)). Such gain-modulated neurons have been reported in large numbers in several cortical areas within the parietal lobe. Gain modulations between sensory and posture signals have been also reported in occipital and premotor cortices, suggesting that basis function representations may be used throughout the cortex.

In early studies gain modulation by posture signals was reported to be linear, not sigmoidal. Linear gain modulation is incompatible with the basis function hypothesis because it does not result in a basis set. These experiments, however, were designed to detect an effect, not to distinguish the precise form of the gain field. A linear model of gain fields was simple and lent itself most easily to statistical testing. However, more recent experiments and new analyses have revealed significant nonlinearities consistent with sigmoidal modulation.

LEARNING SENSORIMOTOR TRANSFORMATIONS

It is conceivable that a few sensorimotor transformations are already wired at birth and require little training (such as the eye blink reflex). In most cases, however, the mapping from sensory to motor coordinates must be learned and updated through life

as the eyes, arms, and body parts change in size and weight.

The goal of learning is to find out how to generate the proper motor commands in response to a desired sensory state. How do we learn to map the sensory coordinates of objects into motor coordinates? Piaget proposed that babies learn the appropriate mapping by associating spontaneous motor commands with the sensory consequences of those spontaneous actions. Consider how this would work in a two-layer network trained to control arm movements. We assume that the input layer encodes the visual location of the hand, while the output layer represents reaching motor commands in joint-centered coordinates. Note that we now define the motor command as the position of the arm expressed in joint-centered coordinates, as opposed to the changes in joint angles of the arm required to reach the target. We do so because learning is easier to explain with this new definition, but the points we make apply just as well to our previous definition. On each trial the network generates a spontaneous pattern of activity in the motor layer. This pattern is fed to the arm which moves accordingly. The network receives visual feedback regarding the position of the hand as a result of this movement. At this point, the system can learn the association between patterns in the sensory and output layers. In particular, the Hebb rule can be used to increase the weights between sensory and motor units that are coactive.

One can even use the current weights between the sensory and motor layers to compute the reaching motor command that the system would generate in response to the observed visual position of the hand. We will refer to this command as the predicted reaching command. If the predicted and spontaneous motor commands are the same, the network can already accurately bring the hand to the desired location and no learning is required. If the two commands differ, the difference between the two can be used as an error signal to adjust the weights (Figure 4(a)). For instance, one could use a learning rule known as the *delta rule*, which takes the form

$$\delta w_{ij} = \alpha\, a_i(a_j^* - a_j) \tag{5}$$

where δw_{ij} is the change in the weight between the

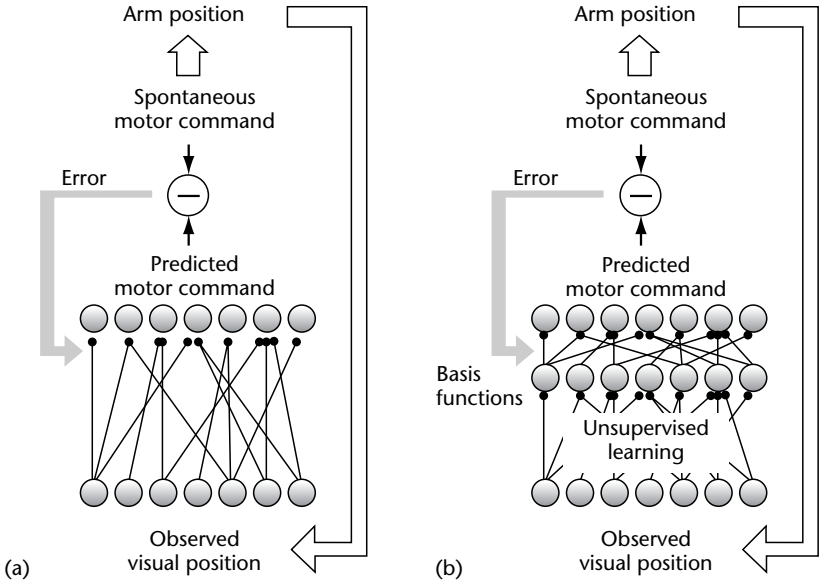


Figure 4. Learning sensorimotor transformations in neural networks. (a) Learning motor commands with spontaneous movements. The motor layer generates a spontaneous pattern of activity; this activity is fed into the arm, resulting in a new arm position, which is observed in the input sensory layer. The sensory activity is passed through the weights to compute a predicted motor command. The weights can then be adjusted according to an error signal obtained by computing the difference between the spontaneous and predicted motor command. By repeating this procedure many times, one can learn the appropriate sensorimotor mapping. (b) When the transformation is nonlinear, as is the case for arm movements, an intermediate layer of units is required. If basis functions are used in the intermediate layer, learning can be done in two stages. The weights to the basis function can be learned with an unsupervised or self-organizing rule because the basis functions depend only on the input, not on the motor command computed in the output layer. The weights to the output layer are equivalent to the ones in the network shown in (a) and can be learned using the same error signal.

presynaptic sensory unit i and postsynaptic motor unit j, α is a learning rate, a_i is the activity of the presynaptic unit, a_j^* is the spontaneous postsynaptic motor activity, and a_j is the predicted postsynaptic motor activity.

This strategy works well if the sensorimotor transformation is linear (that is, it can be implemented in a two-layer network), such as learning to make an eye movement to a visual target. Indeed, the retinal location of a target and the saccade vector required to obtain that target are identical. The transformation from a sensory map (e.g. cortical area V1) to a motor map (e.g. the superior colliculus) is therefore an identity mapping, which is a linear transformation.

Unfortunately, however, most sensorimotor transformations are nonlinear, and the networks that compute them require at least one intermediate layer. We have seen in the first section that a good choice for the intermediate representation is to use basis functions. This turns out to be a good choice for learning as well. Indeed, with basis functions, we can decompose learning into two independent stages:

- learning the basis functions
- learning the transformation from basis functions to motor commands (Figure 4(b)).

To learn the basis functions, one can use a purely unsupervised learning rule; that is to say, the basis functions can be learned without regard to the motor commands being computed – before the baby even starts to move his arm. Indeed, since by definition any basis function set can be used to construct any and all motor commands (eqn [4]), the choice of the basis set is independent of the exact set of motor commands that will eventually be learned. The choice, however, is constrained by general properties of motor commands: for example, motor commands tend to be smooth functions in the sense that similar sensory inputs usually lead to similar motor commands. It is therefore helpful to use smooth basis functions, such as Gaussian functions, which provide smooth interpolation between training samples. Learning constraints can also play a part. Hence, learning the transformation from the basis functions to the motor command is easier when the basis functions are local; i.e. when the functions are nonzero over a limited range of inputs, like Gaussian functions (a counterexample would be polynomial functions like x^2). Indeed, changing the weight of a local basis function affects the motor command only over the range of inputs that activates the basis function, while leaving intact what has already been learned for other sensory inputs, leading to a modular form of learning. Finally, some choices are more consistent with what we know of neuronal responses; this is true in particular of Gaussian and sigmoid functions which fit quite well the tuning responses of many neurons in the cortex. Several biologically plausible self-organizing rules are available for learning such basis functions.

Learning the transformation from basis functions to motor commands is easy, since motor commands are linear combinations of basis functions (see Figure 3(c)). We need to learn only the linear weights in eqn [4] which can be done as outlined above with a simple delta rule (Figure 4(b)). Other problems arise during learning when the nonlinear sensorimotor transformation is a one-to-many mapping, but these issues lie beyond the scope of this article.

CONCLUSION

Coordinate transforms for sensory motor transformations fall into two broad classes, linear and nonlinear. The linear transformations can be performed by networks of units computing simple dot products. The computation of the dot product leads to cosine tuning curves consistent with what has been reported in sensory and motor neurons of invertebrates. Most transformations, however, are nonlinear and require multilayer networks with intermediate representations. Several choices are available for these intermediate representations, but an efficient way to proceed is to use units computing a product of bell-shaped, or sigmoidal, tuning curves to sensory and posture signals. The key property is that units with bell-shaped tuning curves provide basis functions which, when combined linearly, make it easy to compute and learn nonlinear mappings. This solution also has the advantage of being biologically plausible since basis function units show a gain modulation of sensory evoked activity by posture signals, similar to what has been reported for many neurons in the prestriate, parietal, and premotor cortices. This tight link between the observed neurophysiological responses and the theory makes the issue of coordinate transforms one of the best-understood problems in computational neuroscience.

Clearly, basis function networks are only the beginning of the story. Unresolved issues abound, starting with the fact that basis function representations are subject to combinatorial explosion; i.e. the number of neurons required increases exponentially with the number of signals being integrated. One solution to this problem is to use multiple modules of basis functions. It is clear that the brain is indeed using multiple cortical modules

for sensorimotor transformations and it will be interesting to identify the computational principles underlying this modular architecture.

Further Reading

Andersen R, Essick G and Siegel R (1985) Encoding of spatial location by posterior parietal neurons. *Science* **230**: 456–458.

Boussaoud D, Barth T and Wise S (1993) Effects of gaze on apparent visual responses of frontal cortex neurons. *Experimental Brain Research* **93**: 423–434.

Bremmer F, Ilg U, Thiele A, Distler C and Hoffman K (1997) Eye position effects in monkey cortex. I: Visual and pursuit-related activity in extrastriate areas MT and MST. *Journal of Neurophysiology* **77**: 944–961.

Burnod Y, Grandguillaume P, Otto I *et al.* (1992) Visuomotor transformations underlying arm movements toward visual targets: a neural network model of cerebral cortical operations. *Journal of Neuroscience* **12**: 1435–1453.

Galletti C and Battaglini P (1989) Gaze-dependent visual neurons in area {V3a} of monkey prestriate cortex. *Journal of Neuroscience* **9**: 1112–1125.

Georgopoulos A, Kalaska J and Caminiti R (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience* **2**: 1527–1537.

Groh J and Sparks D (1992) Two models for transforming auditory signals from head-centered to eye-centered coordinates. *Biological Cybernetics* **67**: 291–302.

Hinton G and Brown A (2000) Spiking boltzmann machines. *Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press.

Jordan M and Rumelhart D (1992) Forward models: supervised learning with a distal teacher. *Cognitive Science* **16**: 307–354.

Kuperstein M (1988) Neural model of adaptative hand-eye coordination for single postures. *Science* **239**: 1308–1311.

Lewis J and Kristan W (1998) A neuronal network for computing population vectors in the leech. *Nature* **391**: 76–79.

Mazzoni P and Andersen R (1995) Gaze coding in the posterior parietal cortex. In: Arbib MA (ed.) *The Handbook of Brain Theory*, pp. 423–426. Cambridge, MA: MIT Press.

Moody J and Darken C (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation* **1**: 281–294.

Piaget J (1952) *The Origins of Intelligence in Children*. New York, NY: Norton Library.

Poggio T (1990) A theory of how the brain might work. *Cold Spring Harbor Symposium on Quantitative Biology* **55**: 899–910.

Pouget A and Sejnowski T (1997) Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience* **9**(2): 222–237.

Salinas E and Abbot L (1994) Vector reconstruction from firing rate. *Journal of Computational Neuroscience* **1**: 89–108.

Salinas E and Abbot L (1995) Transfer of coded information from sensory to motor networks. *Journal of Neuroscience* **15**: 6461–6474.

Sanger T (1994) Theoretical considerations for the analysis of population coding in motor cortex. *Neural Computation* **6**: 29–37.

Snyder L, Batista A and Anderson R (1997) Coding of intention in the posterior parietal cortex. *Nature* **386**: 167–170.

Snyder L, Grieve K, Brotchie P and Anderson R (1998) Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**: 887–891.

Theunissen F and Miller J (1991) Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *Journal of Neurophysiology* **66**: 1690–1703.

Touretsky D, Redish A and Wan H (1993) Neural representation of space using sinusoidal arrays. *Neural Computation* **5**: 869–884.

Trotter Y and Celebrini S (1999) Gaze direction controls response gain in primary visual-cortex neurons. *Nature* **398**: 239–242.

Westheimer G and Tanzman I (1956) Qualitative depth localization with diplopic images. *Journal of the Optical Society of America* **46**: 116–117.

Zipser D and Andersen R (1988) A back-propagation programmed network that stimulates reponse properties of a subset of posterior parietal neurons. *Nature* **331**: 679–684.

Modeling Individual Neurons and Small Neural Networks

Intermediate article

Elizabeth Thomas, Research Centre for Cellular and Molecular Neurobiology, Liège, Belgium

CONTENTS

Realistic models of neurons range in complexity from the perfect integrate-and-fire model to conductance models in which the ionic currents are computed from the kinetics of the underlying gating particles.

INTRODUCTION

This section outlines the construction of neuronal models which describe in more detail their intrinsic properties as well as their interactions with neighboring neurons. It starts with the integrate-and-fire neuron as a model at the less detailed end of the spectrum, and concludes with models in which the ionic currents emerge from the kinetics of the underlying gating particles. The complexity of the model that is chosen for a particular study would largely depend on the type of question being investigated. For example, one cannot fully investigate the effects of many drugs on a neuronal network without using the Hodgkin–Huxley model, the reason being that in many cases the effect of the drug is known in some detail. For example, benzodiazepines mainly affect the decay of γ-aminobutyric acid A ($GABA_A$)-mediated postsynaptic potentials and not the maximum conductance (Otis and Mody, 1992). On the other hand, if one is investigating auditory processing, much is still unknown about the manner in which information is encoded in a simple network, and constructing a complicated conductance model only serves to confound the problem. The extra details may also be irrelevant to the questions that are being addressed.

In describing various 'realistic' neuronal models by starting out with the simplest, this article begins with the integrate-and-fire model, before moving on to the leaky integrate-and-fire model, the rate model of a neuron, and finally conductance models. The conductance models can be constructed either with macroscopic descriptions of ionic currents or with more detailed models that outline the kinetics of each of the particles gating the channel. A study of network interactions requires models of synaptic activity. The article therefore describes two commonly used synaptic models, namely those with prefixed functions of postsynaptic activity and kinetic models. Finally, since phenomena such as learning and memory require the investigation of neuronal plasticity, the article describes biological models of both short- and long-term plasticity.

INTEGRATE-AND-FIRE MODEL

The integrate-and-fire model of neurons does not include any detailed information on the intrinsic properties of the neuron. Input to the neuron is integrated, and a spike or action potential is generated once a fixed threshold has been reached. The spiking property of the neurons allows the study of the role of factors such as spike timing and spike synchrony in network activity. Some recent connectionist models incorporating the spiking activity of neurons have been found to yield certain advantages over a network of non-spiking neurons (Sougné, 1998). The integrate-and-fire neuronal model is described by the following type of equation:

$$C_m \frac{dV_m}{dt} = I(t) \qquad (1)$$

where C_m is the capacitance of the neuron, V_m is the membrane potential and $I(t)$ is the summated input

current due to synaptic input from neighboring neurons as well as any external current source such as an electrode. The neuron fires a fixed action potential when a threshold is reached. The membrane potential is then set to a prefixed value after the spike. Implicit in such a model is the idea that the timing of a spike, and not its form, is important in carrying information about the system.

The simplest way of representing a spike from the neuron is a simple delta function $\delta(t - \tau_k)$ for a neuron which fires at time τ_k. Although this is the simplest representation that can be used, various complications can be added to the model in order to characterize more realistically the neuronal spike and, more importantly, its activity immediately afterwards. Perhaps the most important of these is the refractory period immediately after an action potential. During this period no stimulus is able to elicit a second action potential from the neuron. The refractory period is due to the fact that the sodium channels which are largely responsible for generating a spike are in an inactivated state. The presence of this constraint places an upper limit on the frequencies at which a neuron may fire an action potential. If the refractory period is t_{ref}, the maximum firing frequency of the neuron, $f_{\max}$, is $1/t_{\text{ref}}$. The neuronal model may also include a relative refractory period immediately after the refractory period, during which the threshold for evoking a spike is increased.

LEAKY INTEGRATE-AND-FIRE MODEL

These models add a new level of complexity to the behavior of a neuronal model by including a leak current. The neuronal membrane in this instance can be represented by the equivalent circuit shown in Figure 1. The current flowing across the resistance of the circuit represents the leak current. The total current I_m flowing across the neuronal membrane is the sum of the capacitance current I_C and the leak current I_l:

$$I_m = C_m \frac{dV_m}{dt} + \frac{V_m}{R_m} \tag{2}$$

The simple integrate-and-fire model would linearly integrate two temporally separated inputs provided that their sum does not exceed threshold. However, the presence of a leak introduces a more realistic behavior in which the occurrence of an input is gradually forgotten over time. The earlier occurring input therefore makes a less important contribution to the summated neuronal response at a later time.

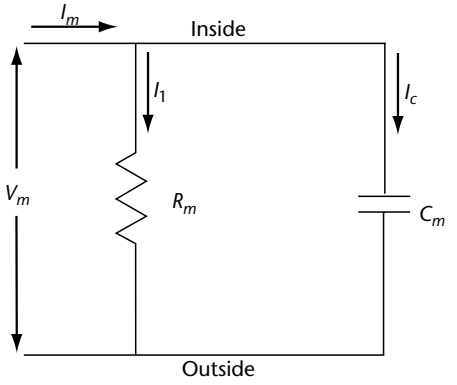


Figure 1. Equivalent circuit for leaky integrate-and-fire model of neuron.

In the presence of a leak term, the membrane potential in the absence of any input current will gradually decay to its initial value. In response to a constant current pulse I_m, the leaky integrate-and-fire model would attain the steady-state value $V_m(\infty) = I_m R_m$. However, the value of the membrane potential before this steady state is reached is given by:

$$V_m(t) = I_m R_m (1 - e^{-t/\tau_m}) \tag{3}$$

where $\tau_m = R_m C_m$ and is the membrane time constant representing the time taken for V_m to reach 63% of its final value.

RATE MODEL

In many models of large networks, the neuronal output is not represented in the form of individual spikes. Instead, the information that is taken into account is the firing rate of a neuron. In this case, the neuronal output is computed from a sigmoidal function relating membrane potential to the neuronal firing rate. The instantaneous firing rate $z_i(t)$ of a neuron is usually obtained from the neuronal membrane potential $V_m(t)$ with the use of a sigmoid function:

$$z_i(t) = \frac{z_{\max}}{1 + e^{-b(V_m(t) - \theta)}} \tag{4}$$

The use of this function allows for two factors contributing to nonlinearities in the neuronal output. Due to the refractory period, there is a maximal rate at which the neurons can fire, $z_{\max}$. The second factor is a threshold to reflect the fact that a minimum level of excitation is required for the neuron to begin firing. The factors b and θ contribute to setting the threshold of excitation and also the rate at which the firing increases with membrane potential.

Of course, the use of a rate model does not allow for the use of information that might be represented in the timing of neuronal firings or the correlation of firing times between neurons. The question of whether a rate code or temporal code is used for information coding in the nervous system has been much debated during the last few years. The investigators who have presented evidence for the representation of stimulus information in the temporal modulation of the neuronal spike train have included Bialek *et al*. (1991), who worked on the fly visual system, and Richmond *et al*. (1987), who investigated the inferior temporal cortex. The view that spike timing in neuronal firing is not very important has been taken by investigators such as Shadlen and Newsome (1998) and Thomas *et al*. (2000). The latter investigators have presented evidence to demonstrate that a good level of category information can be represented in the inferior temporal cortex without information on spike timing.

SYNAPTIC INPUT TO THE INTEGRATE-AND-FIRE NEURON

When they are present in a network, the neurons receive input from the surrounding neurons. At a given timestep, all of the input from the surrounding neurons is summed linearly to give the resulting activity V_i of a neuron i:

$$C_m \frac{dV_i}{dt} = \frac{V_i}{R_m} + \sum_j \sum_k \omega_{ij} g_{ij}(t - \tau_k)\delta_j(t - \tau_k)[V_i - V_{eq}] \tag{5}$$

The double summation is to take into account not only the input from all of the presynaptic neurons j, but also to sum over each incidence k of a presynaptic neuron firing. The term ω_{ij} is the sign and the strength of connections between neurons i and j. The function $g_{ij}(t - t_0)$ usually takes the form of an alpha function:

$$g_{ij}(t - t_0) = \frac{(t - t_0)}{\tau} e^{-(t-t_0)/\tau} \tag{6}$$

where t_0 is the time of transmitter release. The function $g_{ij}(t - t_0)$ decays with a time constant τ after reaching a peak. The values of τ are chosen to reflect the various fast and slow excitatory and inhibitory synapses that are seen in the nervous system. In many realistic network models these include the fast inhibitory synapse $GABA_A$, the slow inhibitory synapse $GABA_B$, fast excitation mediated by alpha-amino-3-hydroxy-5-methylisoxazole 4-proprionic acid (AMPA) receptors, and a slow excitation mediated by *N*-methyl-D-aspartate (NMDA) receptors. Each synapse also has its own characteristic value of V_{eq}, where $(V - V_{eq})$ represents the driving force on each synaptic current.

CONDUCTANCE MODELS

Many studies in neuroscience involve investigation of the neuronal ionic channels. The effects of various drugs on neurons are understood in terms of their actions on these ionic channels. The framework that is most widely used to describe these channels is the Hodgkin–Huxley model for action potentials (Hodgkin and Huxley, 1952). One of the main advantages of using this model is the availability of a large number of the required parameters from the experimental literature, as many of the experimental investigations themselves are conducted using the Hodgkin–Huxley framework. Some of the disadvantages of using the model include the large number of parameters that are required to characterize the system, and the long computer run times that are required to treat such unwieldy systems.

Ionic channels in the Hodgkin–Huxley model are represented as resistances (or conductances) arranged in parallel (Figure 2). The equivalent circuit for such a system is an elaboration of what has already been presented in Figure 1. In this case, additional resistances are added to simulate the other currents in the neuronal membrane in addition to the leak current. The circuit in Figure 2 denotes the currents that are typically represented in models of thalamocortical (TC) cells (Lytton and Thomas, 1997; Thomas and Lytton, 1997; Thomas and Grisar, 2000). The following equation describes such a circuit:

$$C_m \frac{dV}{dt} = -I_T - I_h - I_{Na} - I_K - I_l - I_{GABA_A} - I_{GABA_B} \tag{7}$$

The currents I_T, I_h, I_{Na}, I_K, and I_l are intrinsic currents, while I_{GABA_A} and I_{GABA_B} are synaptic currents. The values of these ionic currents are obtained from Ohm's law:

$$I = g\bar{g}(V - E_{eq}) \tag{8}$$

where E_{eq} is the reversal potential of the channel and $\bar{g}$ is the maximum conductance of a channel. Both of these values are fixed. However, the value of g, the channel conductance, is frequently dependent on factors such as the membrane potential,

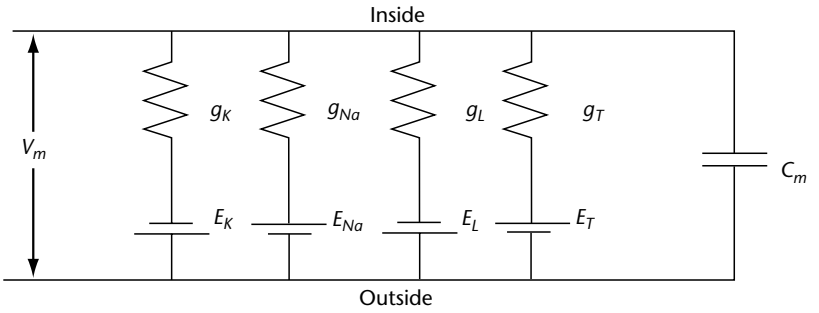


Figure 2. Equivalent circuit for conductance model of thalamocortical neuron.

time, or the internal calcium concentration, and has to be generally computed for each channel at each timestep. The leak current is usually an exception to this.

In the case of the intrinsic channels in which g changes, the value of g depends on the state of the gating particles m and h. While m is an activation gate, h is an inactivation gate. If N is the number of activation gates:

$$g = m^N h \tag{9}$$

The state of the gating particles (m or h) is a function of membrane potential as well as of time. At any time t, its value can be computed from the following equation:

$$gate_t = gate_\infty - (gate_\infty - gate_{t-1})\exp\left[\frac{-\Delta t}{\tau_{gate}}\right] \tag{10}$$

The values $gate_\infty$ and τ_{gate} depend on the membrane potential. In the case of many channels, their values as a function of membrane potential are directly available from the experimental literature.

In other cases they have to be computed from α and β, the rate constants for the reaction between open and closed states for gating particles:

$$gate_{open} \underset{\beta}{\overset{\alpha}{\rightleftarrows}} gate_{closed} \tag{11}$$

In this case:

$$gate_\infty(V) = \frac{\alpha(V)}{\alpha(V) + \beta(V)} \tag{12}$$

$$\tau_{gate}(V) = \frac{1}{\alpha(V) + \beta(V)} \tag{13}$$

As indicated by the name, the value $gate_\infty$ determines the magnitude of gate opening when the membrane potential is held at a particular value for time infinity. The value of τ_{gate} determines the amount of time required for the gate opening.

KINETIC MODELS OF IONIC CHANNELS

The Hodgkin–Huxley representation of ionic currents is at a macroscopic level. Many models also portray the currents at a more detailed level by describing the ionic channel based on the state transitions of each of the gating particles that constitute the channel. The value of an ionic current therefore emerges from the underlying dynamics of the system. There are two key assumptions that underlie classical kinetic theory which is used to describe the state transitions of the gating particles. First, gating is a Markov process – the rate constants of the transition of a gating particle from one state to the other (e.g. open to closed) are independent of the previous history of the system. Secondly, the transitions can be described by first-order differential equations with a single time constant.

A construction of such models can be illustrated by the model of a potassium (K^+) channel previously described by Hodgkin and Huxley. The channel is composed of four n gates. Each of these gates can be in either an open (1) or closed (0) state. The channel is conducting only if all four n gates are in an open state (1111). Starting from the condition where all of the gates are closed (0000), the system can therefore go through 14 intermediate closed states before it finally becomes open (1111) (Figure 3). By making the assumption that the four gates are independent and kinetically identical, the web of equations that lead from the nonconducting to the conducting state of the K^+ ionic channel can be very much reduced. For example, all states with the same number of closed gates are kinetically identical and can therefore be treated together. Using such assumptions, the entire tree of reactions can be collapsed to the transitions between five major states, with state 4 being the only conducting state. The values 4α, 3α, 2α, and α are the rate constants for all of the forward reactions

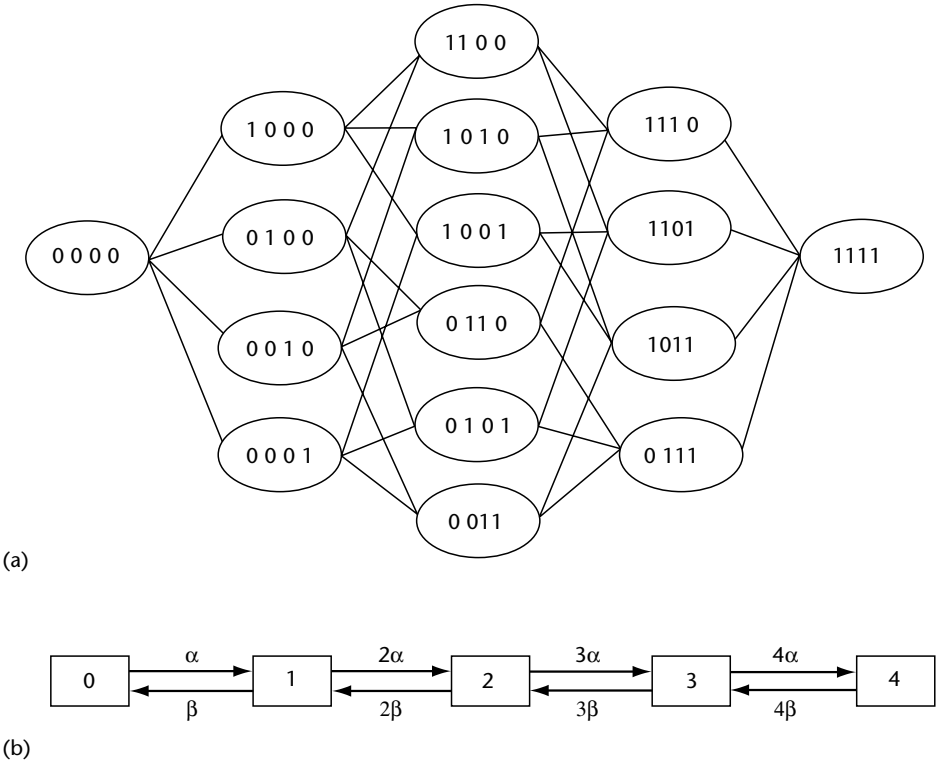


Figure 3. (a) The K^+ channel consisting of four n particles can go from the closed state (0000) to the only open state (1111) via 14 intermediate states which are all closed. (b) By pooling together states which are kinetically identical, the complex state diagram in (a) can be reduced to five states (Hille, 1991).

leading to state 4, while 4β, 3β, 2β, and β are the rate constants for the backward reactions leading from open state 4 to the closed state 0.

SYNAPTIC INPUT TO CONDUCTANCE MODELS

Synaptic input in many conductance models is represented in the form of the alpha function introduced in the earlier section on synaptic input to the integrate-and-fire neuron. As mentioned earlier, one of the problems with using this model is the fact that the time of occurrence of each spike has to be stored in a queue, and the corresponding exponentials have to be calculated for each time-step. Another criticism of these models is that although they imitate the observed time course of postsynaptic potentials in electrophysiological recordings, they are not based on any of the underlying mechanisms in the process. Furthermore, these models do not naturally provide for a saturation that could occur at the synapse.

A kinetic model of receptor binding described by Destexhe *et al*. (1994a) is able to resolve many of these problems. In a two-state kinetic model, neurotransmitter molecules T bind to the postsynaptic receptor R according to the following first-order scheme:

$$R + T \underset{\beta}{\overset{\alpha}{\rightleftarrows}} TR^* \tag{14}$$

If r is the fraction of bound receptors, then:

$$\frac{dr}{dt} = \alpha[T](1 - r) - \beta r \tag{15}$$

If $[T]$ occurs as a pulse, the above equation can be solved to obtain the following expression for the value of r during the pulse ($t_0 < t < t_1$) and for the period following the pulse ($t > t_1$).

During the pulse:

$$r(t - t_0) = r_\infty + (r(t_0) - r_\infty)\exp\left[\frac{-(t - t_0)}{\tau_r}\right] \quad (t_0 < t < t_1) \tag{16}$$

After the pulse:

$$r(t - t_1) = r(t_1)\exp[-\beta(t - t_1)] \quad (t > t_1) \tag{17}$$

where both r_∞ and τ_r are functions of the rate constants α and β in eqn [14].

The value of r is then used to calculate the synaptic current in the same manner as an intrinsic current:

$$I_{syn}(t) = \overline{g}_{syn} r(t)\left[V_{syn}(t) - E_{syn}\right] \tag{18}$$

Although these two state models capture many of the important characteristics of postsynaptic potentials, a larger number of states can be used to model them more accurately (Destexhe *et al*., 1994b).

MODELS OF SYNAPTIC MODIFICATION

The efficacy of connections between the neurons of the nervous system has been found to be a parameter that is not constant but which fluctuates according to the dynamics of the network. These changes can be both long- and short-term, and can cause synaptic efficacy to increase or decrease. While the short-term changes last for the order of seconds to minutes, the long-term changes may persist for an hour or more.

A model for short-term synaptic facilitation and depression has been provided by Abbott *et al*. (1997). In this model, the amplitude of neuronal response K_1 is adjusted by a factor $A(t_i)$. If R is the response of a neuron to a train of spikes:

$$R(t) = \sum A(t_1) K_1(t - t_i) \tag{19}$$

If there was an isolated spike, and previous spikes did not have an influence on the postsynaptic response, A would have a value of 1. However, as a result of a presynaptic spike, A would either increase in the case of facilitation or decrease in the case of depression. This could take place in either an additive or a multiplicative fashion. The value of A would change according to $A \rightarrow fA$ in the multiplicative case, or according to $A \rightarrow A + (f-1)$ in the additive case. A value of $f > 1$ would correspond to facilitation, while a value of $f < 1$ would lead to depression. In order to produce biologically realistic behavior, Abbott recommends the usage of a multiplicative expression for depression ($f < 1$) and an additive description for facilitation ($f > 1$).

Another model of short-term depression is that of Markram *et al*. (1998). This model aims to explain the frequency dependence of synaptic transmission. Both Markram and Abbott have suggested that these changes in synaptic strength have implications for the type of encoding that is used by the nervous system. A rate code is favored when depression is slow and a temporal code is favored when depression is fast. A neuron which was initially sensitive to the firing rate of a rapidly firing presynaptic terminal, as a result of synaptic depression, now reaches a steady state which is no longer a function of the presynaptic firing rate. However, several depressed excitatory postsynaptic potentials (EPSPs) are generated before this steady state is reached. During this period, the neuron shows an increased sensitivity to temporal coherence in the presynaptic input (Abbott *et al*., 1997; Tsodyks and Markram, 1997).

The long-term changes that are observed in the strengths of synaptic connectivity are long-term potentiation (LTP) and long-term depression (LTD). The experimental paradigm that is usually used to produce LTP is a train of high-frequency stimulation, while LTD is produced by a train of low-frequency stimuli. LTP has been found to be mediated by both NMDA and non-NMDA synapses. Long-term potentiation mediated by NMDA synapses generally obeys Hebbian rules for induction. The Hebbian rule was proposed by Donald Hebb as a mechanism for learning. He suggested that the concurrent excitation of pre- and postsynaptic elements would result in an increase in synaptic efficacy. In the NMDA synapse the requirement for concurrence arises because the postsynaptic terminal has to be depolarized in order to remove an Mg^{2+} block. The presynaptic terminal has to be depolarized for the release of neurotransmitter. Realistic models of the NMDA synapse must therefore take into account the concentration of Mg^{2+} ions. The following is an example of an equation for an NMDA-mediated current in a model of working memory in the cortical network (Wang, 1999):

$$I_{NMDA} = \frac{g_{NMDA}(V_m - V_E)}{(1 + [Mg^{2+}]e^{-kV_m})} \tag{20}$$

More recent experimental research has revealed more details of the way in which the relative timing of pre- and postsynaptic firing can change the sign as well as the strength of synaptic modification. If the presynaptic action potential precedes postsynaptic firing by no more than about 50 ms, then an increase in synaptic strength occurs. The smaller the temporal separation between the presynaptic firing and the postsynaptic action potential, the larger is the increase in synaptic strength. On the other hand, presynaptic action potentials that follow postsynaptic spikes lead to a depression in synaptic strength. Models of this type of spike-timing-dependent plasticity have been developed by Song *et al*. (2000).

MULTICOMPARTMENTAL MODELS

The highly branched dendritic trees of a neuron confer highly complex spatial structures on the neuron. Neuronal models which attempt to understand the role of these structures make use of multicompartments (Figure 4). The compartmental approach to study of the flow of ionic currents in dendritic trees was pioneered by Rall (1964), who modeled the dendritic tree by using cable equations. These equations could be solved analytically under extremely limited circumstances. This approach became difficult and sometimes impossible when most realistic dendritic structures were taken into account. Membrane properties that are voltage dependent and which generate action potentials are also problematic. Rall had pointed out that in such cases compartmental models should be used. A good description of such models can be found in Segev and Burke (1999). In this instance each neuron is composed of a number of compartments. Each compartment is considered to be isopotential and its properties are considered to be uniform. For such models, the property of each compartment is not only the consequence of currents intrinsic to the compartment, but also depends on currents which enter and leave from adjacent compartments.

The membrane potential for a neuronal compartment k with adjacent compartments $k+1$ and $k-1$ is then computed as follows:

$$C_m \frac{dV_k}{dt} + I_{ion} = \frac{V_{k-1} - V_k}{r_{k-1,k}} - \frac{V_k - V_{k+1}}{r_{k,k+1}} \qquad (21)$$

Although in many instances unicompartment models are sufficient to generate the activity of interest, in some situations they fail. An example of this is provided by a study described by Abbott and Marder (1998). These researchers demonstrated how in some bursting cells in the somatogastric ganglion of certain invertebrates it is important to isolate the primary neurite from the spike initiation zone on the axon. If this is not done, the slow oscillating potentials that are generated in the primary neurite can prevent the de-inactivation of the fast Na^+ conductance which terminates the action potential in the spike initiation zone. However, once these two areas are electrotonically separated the problem is resolved.

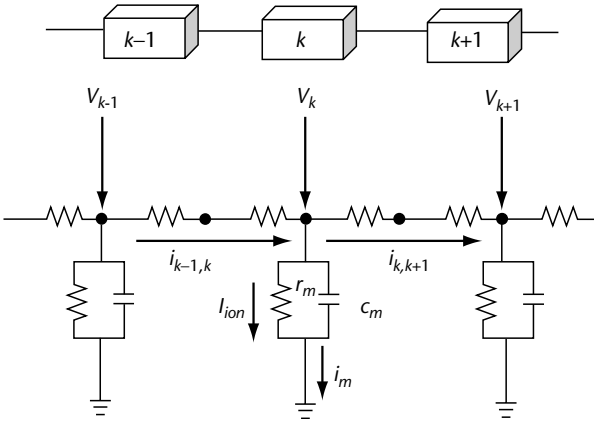


Figure 4. Equivalent circuit for multicompartment model (Segev and Burke, 1999).

References

Abbott L and Marder E (1998) Modeling small networks. In: Koch C and Segev I (eds) *Methods in Neuronal Modeling*, pp. 372–373. Cambridge, MA: MIT Press.

Abbott LF, Varela JA, Sen K and Nelson SB (1997) Synaptic depression and cortical gain control. *Science* **275**: 220–224.

Bialek W, Rieke F, de Ruyter van Stevenick RR and Warland D (1991) Reading a neural code. *Science* **252**: 1854–1857.

Destexhe A, Mainen Z and Sejnowski T (1994a) An efficient method for computing synaptic conductances based on a kinetic model of receptor binding. *Neural Computation* **6**: 14–18.

Destexhe A, Mainen Z and Sejnowski T (1994b) Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism. *Journal of Comparative Neuroscience* **1**: 195–230.

Hille B (1991) *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer Associates Inc.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* **117**: 500–544.

Lytton W and Thomas E (1997) Modeling thalamocortical oscillations. In: Ulinski PS and Jones EG (eds) *Cerebral Cortex*, vol. 13. *Models of Cortical Circuitry*, pp. 479–509. New York, NY: Plenum Press.

Markram H, Pikus D, Gupta A and Tsodyks M (1998) Potential for multiple mechanisms and algorithms for synaptic plasticity at single synapses. *Neuropharmacology* **37**: 489–500.

Otis T and Mody I (1992) Modulation of decay kinetics and frequency of $GABA_A$-receptor-mediated spontaneous inhibitory postsynaptic currents in hippocampal neurons. *Neuroscience* **49**: 13–32.

Rall W (1964) Theoretical significance of dendritic trees for neuronal input–output relations. In: Reiss RF (ed.) *Neural Theory and Modeling*, pp. 73–94. Stanford, CA: Stanford University Press.

Richmond BJ, Optican L, Podell M and Spitzer H (1987) Temporal encoding of two dimensional patterns by single units in primate inferior temporal cortex. I. Response characteristics. *Journal of Neurophysiology* **57**: 132–146.

Segev I and Burke R (1999) Compartmental models of complex neurons. In: Koch C and Segev I (eds) *Methods in Neuronal Modeling*, pp. 93–137. Cambridge, MA: MIT Press.

Shadlen MN and Newsome W (1998) The variable discharge of cortical neurons: implications for connectivity, computation and information coding. *Journal of Neuroscience* **18**: 3870–3896.

Song S, Miller K and Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience* **3**: 919–926.

Sougné J (1998) Connectionism and the problem of multiple instantiation. *Trends in Cognitive Sciences* **2**: 183–189.

Thomas E and Lytton W (1997) Computer model of anti-epileptic effects mediated by alterations in $GABA_A$-mediated inhibition. *Neuroreport* **9**: 691–696.

Thomas E and Grisar T (2000) Increased synchrony with increase of a low-threshold calcium conductance in a model thalamic network: a phase-shift mechanism. *Neural Computation* **12**: 1553–1573.

Thomas E, Van Hulle M and Vogels R (2000) Encoding of categories by non-category-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience* **13**: 190–200.

Tsodyks MV and Markram H (1997) The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences of the USA* **94**: 719–723.

Wang XJ (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *Journal of Neuroscience* **19**: 9587–9603.

Further Reading

Bower J (1999) *Computational Neuroscience: Trends in Research 1999*. The Netherlands: Elsevier.

De Schutter E (2000) *Computational Neuroscience: Realistic Modeling for Experimentalists*. Boca Raton, FL: CRC Press.

Destexhe A and Sejnowski T (2001) *Thalamocortical Assemblies*. Oxford, UK: Oxford University Press.

Hille B (1991) *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer Associates Inc.

Johnston D and Wu S (1995) *Foundations of Cellular Neurophysiology*. Cambridge, MA: MIT Press.

Koch C (1999) *Biophysics of Computation*. Oxford, UK: Oxford University Press.

Koch C and Segev I (1999) *Methods in Neuronal Modeling*. Cambridge, MA: MIT Press.

Traub RD and Miles R (1991) *Neuronal Networks of the Hippocampus*. Cambridge, UK: Cambridge University Press.

Motor Control: Models

Advanced article

Liana E Brown, Pennsylvania State University, University Park, Pennsylvania, USA
David A Rosenbaum, Pennsylvania State University, University Park, Pennsylvania, USA

CONTENTS

The term 'motor control' refers to the ability of biological and artificial systems to plan, initiate, maintain, monitor, and correct movements to attain physically realizable goals. A 'model' is a system or process that permits predictions.

INTRODUCTION

As in most areas of cognitive science, models play an important role in motor control research. This is fitting because motor control itself relies on models which predict the outcomes of actions. This article reviews the main challenges faced by motor control systems, and various conceptions of how these challenges might be met.

DEFINITIONS AND CHALLENGES

The term 'motor control' refers to the ability of biological and artificial systems to plan, initiate, maintain, monitor, and correct movements to attain physically realizable goals. The term 'model' refers to a description of a system or process that permits predictions. Models relevant to motor control permit predictions, for example, of the mechanical effects on the forearm of contracting the biceps muscles, of the optical effects on the retina of rotating the eye, and of the haptic effects of gripping a cup of tea. As these examples show, models apply over a wide variety of events.

In motor control research, it is typically assumed that the actor has some goal that can be satisfied physically. Selection of the goal is usually considered a motivational rather than a motor-control problem. The choice of goal may depend, however, on knowledge of what the actor can achieve motorically.

When a goal has been specified, the nature of the action to be carried out is usually not fully determined. If an apple is in reach, for example, there may be an infinite number of movements that allow it to be obtained. One challenge for motor control research is to explain how one movement is chosen from the plethora that are possible. This is an example of the redundancy problem. Redundancy is present at all levels of the motor control system. The redundancy problem (the problem of finding one motor solution when more than one is possible) can be approached by considering body positions while ignoring the forces behind them (kinematics), or by considering the forces involved (dynamics). Considering kinematics alone may be acceptable in computer animations or in abstract computational models. Dynamics may be considered without regard to the underlying muscle recruitment patterns, or, for even more physiological realism, with regard to how muscles are recruited and even how neurons fire.

A second challenge in motor control is learning. The learning problem is how to model relations between movements and their effects. Some learned models are 'forward' (e.g. 'What happens to my elbow if my biceps muscle contracts?'); others are 'inverse' (e.g. 'What muscle events caused my elbow to flex?').

A third challenge in motor control is perceptual–motor integration. It is important to understand how feedback (information about past performance) is used to correct errors, and also how feedforward (information about forthcoming performance) is used to prevent errors. Note that feedback control and feedforward control may each rely on either forward or inverse models.

A fourth challenge in motor control is movement sequencing. This is often called the serial order problem. This problem has been studied mainly in connection with such tasks as walking, speaking, typewriting, and drawing. Because the serial order of motor acts is implied by their timing, the problem of timing is closely related to the problem of

serial order. These two problems will be treated together in this article.

MODELS FOR REDUNDANCY

There are many ways to carry out elementary actions. Consider the task of reaching for an apple. The apple can be grasped in many ways, because more degrees of freedom characterize the state of the limb than the state of the apple. The components needed to characterize the position of the apple are six: the apple's position on the horizontal, vertical, and depth axes, and its pitch, roll, and yaw. But the apple can be grasped in infinitely many ways because the body has more than six mechanical degrees of freedom. The arm has more than six degrees of freedom because the shoulder can rotate in three spatial dimensions, the elbow can rotate in two, and the wrist can also rotate in two. Taking into account the fact that the fingers and trunk add their own degrees of freedom, it is apparent that there is much redundancy in the apple-grasping task. Thus, the question arises of how one particular grasping motion is chosen out of the infinite number that are possible in principle.

Models that have addressed this problem have been of four main types: models that emphasize properties of the peripheral neuromotor system; models that emphasize effector interactions; models that emphasize geometric restrictions; and models that emphasize cost reduction. Each will now be considered in turn.

Models Based on the Peripheral Neuromotor System

Models of the first type recognize that muscles and reflexes have properties that allow effectors to be treated as mass–spring systems with adjustable equilibria (Bizzi *et al*., 1992; Feldman, 1986). In this approach, desired positions are represented as modifiable equilibrium positions to which the body is drawn. There is controversy about how equilibria are represented. One view is that viscoelastic properties of muscle are set through feedforward control (Bizzi *et al*., 1992). Another view is that feedback properties of neuromuscular reflexes are changed to allow effectors to alter their resting states (Feldman, 1986). In any case, evidence exists for the general proposition that terminal positions are planned as goals rather than as mere consequences of movements. Firstly, limbs can reach terminal positions despite perturbations, even when feedback is removed (Polit and Bizzi, 1978). Secondly, position variability increases during movement but then decreases near terminal positions (Harris and Wolpert, 1998; Rosenbaum *et al*., 2001). Thirdly, memory for final positions is more long-lasting and more accurate than memory for movements to those final positions (Marteniuk and Roy, 1972).

Models Based on Effector Interactions

The second class of models that addresses the redundancy problem emphasizes effector interactions (Bernstein, 1967). These models assert that dependencies between limb segments limit action possibilities. For example, within one arm it is more difficult to steadily coordinate extension of the wrist with flexion of the elbow than to steadily coordinate extension of the wrist with extension of the elbow. Similarly, people find it difficult to draw a circle with one hand while drawing a rectangle with the other. Interestingly, callosotomy (split-brain) patients show no such spatial coupling (Franz *et al*., 1996). Some effector interactions are dependent on the perceptual feedback provided. Thus, neurologically normal individuals can produce bimanual movement patterns that are otherwise difficult if special visuo-motor displays are used (Mechsner *et al*., 2001).

Other interactions are task-dependent. If a downward force is lightly applied to the lower lip as it ascends during simple utterances such as 'aba, aba, aba', the upper lip descends more than usual, and with remarkably short latencies, to complete the lip closure (Abbs *et al*., 1984). By contrast, if the same light downward force is applied to the lower lip as it ascends during simple utterances such as 'ala, ala, ala', the upper lip does not descend more than usual. In saying 'ala, ala, ala', bilabial closure (bringing the lips together) is unnecessary. This demonstrates that coupling of effectors depends on the task being performed. Sophisticated neural control mechanisms must exist to permit such functionally adaptive coupling.

Models Based on Geometric Restrictions

The third class of models that addresses the redundancy problem emphasizes geometric restrictions. Models in this class focus on the observation that the range of adopted postures is less than the range of postures that are possible. When the eyes rotate to gaze at objects in different locations, for example, they do not adopt all possible ocular positions (Tweed and Vilis, 1990, 1992). Although in principle

the eyes can rotate freely about their vertical, horizontal, and torsional (depth) axes, eye movements conform to 'Listing's law': the axes of rotation actually used by the eyes tend to lie in a plane. When the upper arm rotates about the shoulder, the positions it adopts are similarly restricted (Gielen *et al*., 1997). Geometric restrictions may reflect a control strategy for dealing with motor redundancy.

Models Based on Cost Reduction

The fourth class of models that address the motor redundancy problem emphasizes cost containment. Much of the work illustrating this approach has focused on hand movements. Because such movements tend to be smooth, investigators have suggested that they implicitly satisfy a smoothness constraint such as minimization of the mean rate of change of acceleration (Flash and Hogan, 1985) or minimization of the mean rate of change of torque (Uno *et al*., 1989).

Much as effector interactions may be task-dependent, smoothness constraints may also be elective or 'soft' rather than physically necessary or 'hard'. This view is supported by the fact that skilled actors may intentionally switch between smooth and jerky movements. For example, a violinist may switch between bowing in a staccato or legato fashion. Recent models have emphasized flexible adjustment of movement properties (Rosenbaum *et al*., 2001). Such models rely on multiple constraints whose priorities can be changed according to task demands.

MODELS FOR LEARNING

As stated earlier, motor learning may use forward and inverse models. Acquisition of forward models is generally more straightforward than acquisition of inverse models. Acquiring a forward model entails predicting sensory outcomes of motor commands, comparing those outcomes to obtained feedback, and using error signals to alter the forward model (see Figure 1). By contrast, acquiring an inverse model entails predicting what change in motor commands will correct a mismatch between intended and obtained results. Only when the correct inverse model is available will the prediction be consistently correct. Because of the significant nonlinearities of the system, finding the correct inverse model may be difficult (Wolpert and Kawato, 1998).

Regardless of how models are learned and used, one fact that has clearly emerged from recent research on motor learning is that actors do indeed learn 'smart' functions relating performance to perception. In a demonstration of this capability, participants made point-to-point hand movements in artificial force fields (Conditt *et al*., 1997). Initially the participants' hand paths were curved, but soon they became essentially as straight as observed in normal circumstances. Even when the participants had to generate hand translations in new parts of the workspace, they could immediately produce straight movements, provided they had learned to compensate for the force field in the originally trained part of the workspace. This shows that the motor system does not learn the dynamics of the arm by rote memorization of past experience. Instead, it learns models that allow for transfer to new situations. Such models are needed not only for adaptation to artificial force fields but also for adaptive performance in the everyday environment. To maximize the effectiveness of models (forward and inverse), one approach has been to allow for the operation of many internal models, each of which may contribute to the planned movement in proportion to how well the current environment fits the environment for which it was developed. In this approach, the multiple internal models allow the system to respond appropriately in many different contexts (Wolpert and Ghahramani, 2000).

MODELS FOR PERCEPTUAL–MOTOR INTEGRATION

Moving skillfully requires anticipation of the perceptual consequences of one's own actions (feedforward), and rapid correction for perceived errors (feedback). These two forms of discrimination are considered below.

Feedforward

Anticipating the perceptual consequences of one's own actions (feedforward) is a prerequisite for effective motor control. That this capability is fundamental is shown by the fact that even insects benefit from it. This was discovered through an ingenious experiment that involved turning a fly's head so that the top of the fly's head was on the bottom and the bottom of its head was on the top, gluing the rotated head against the fly's thorax, and then observing the fly's behavior (von Holst and Mittelstaedt, 1950). When the fly's head was flipped, it walked haltingly, if at all, when it was in lit surroundings, though it could walk perfectly well in darkness. Why did the fly not walk in its normal way when the lights were on? The explanation was

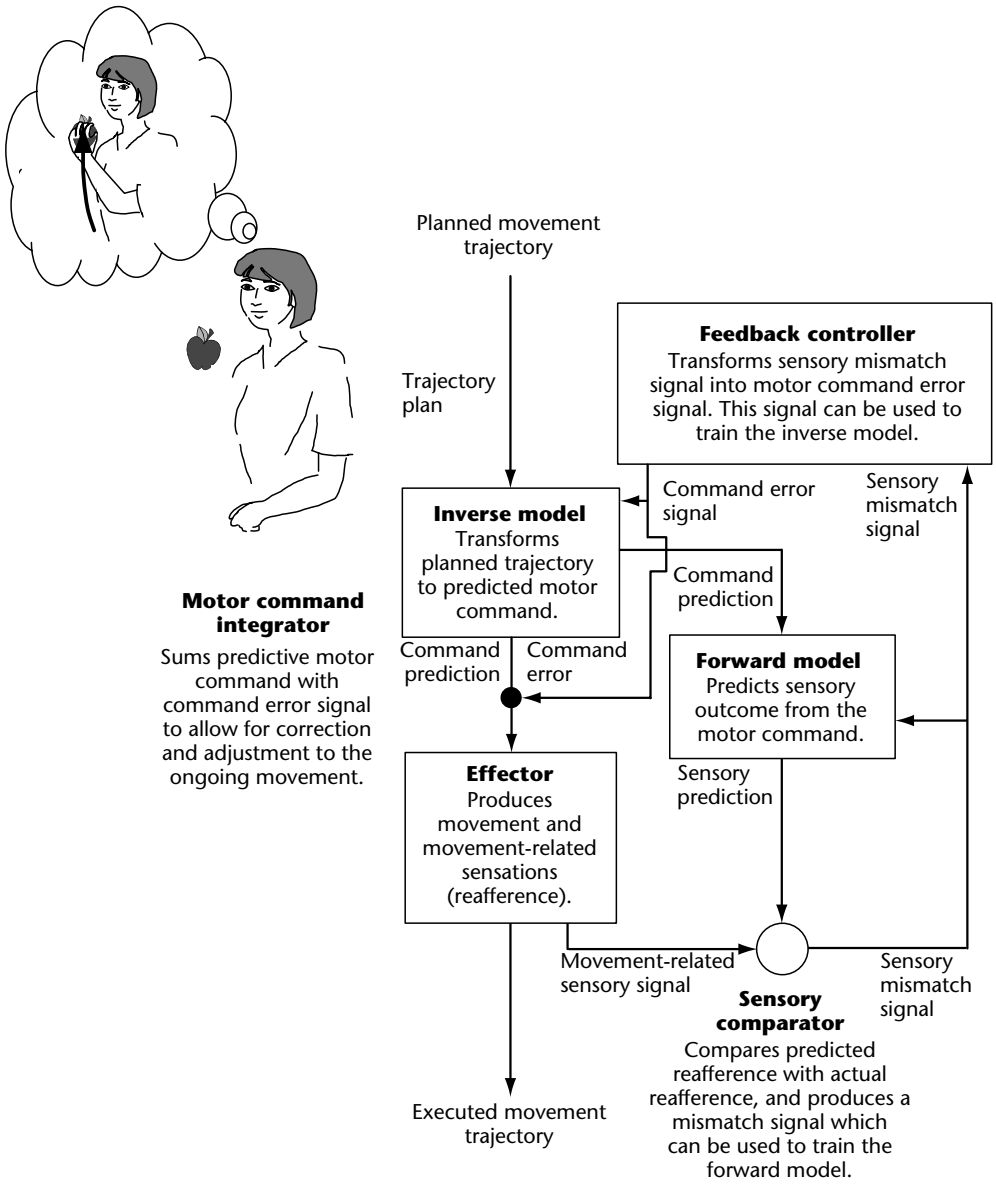


Figure 1. An example movement control system in which an inverse internal model and a forward internal model cooperate for movement control and learning. The inverse and forward models provide feedforward control by producing the predicted motor command and sensory consequences, respectively. The command signal is transmitted both to the forward model ('efference copy') and to the effector. The forward model uses the command signal to predict movement-related sensory feedback ('reafference') for the planned trajectory. The effector produces the trajectory and movement-related reafference. Predicted and actual reafference are compared and movement error is coded as a sensory mismatch signal. The sensory mismatch signal provides feedback control to the ongoing movement and is used to train the forward and inverse models. The forward model is trained directly on the basis of the sensory mismatch signal. The inverse model is trained after the sensory error signal is transformed to a motor command error signal, which is equivalent to the command that would correct for error in the executed movement trajectory. Adapted from Wolpert *et al*. (1998) and Wolpert and Kawato (1998).

that the fly expected visual shifts opposite to its own displacement. When the obtained visual shifts were in the same direction as the fly's movement, they signaled to the fly that the world was moving, which in turn caused the fly to behave conservatively – standing still rather than walking. In other words, the fly anticipated what it would see when it prepared to move.

Such anticipation also occurs in higher animals. Pushing gently on one's eyeball causes the seen

world to move even when it is stationary. In this situation, the normal commands for moving the eyes are absent and the induced retinal shift is interpreted as resulting from external motion. By implication, when the eyes are controlled by normal oculomotor commands, the seen stable world appears stationary because the oculomotor commands give rise to anticipated visual shifts. Brain changes reflecting such anticipations have been recorded neurophysiologically (Duhamel *et al*., 1992). Such data, coupled with behavioral results like the ones just reviewed, indicate that motor control is accompanied by anticipated perceptual changes.

Recent research shows that it is possible to anticipate not only what motor-related perceptual changes will occur, but also when they will occur. The capacity for time-related anticipation is important because long feedback delays can make it difficult to correct for errors. Feedback delays for normal eye–hand coordination usually range from 100 to 250 milliseconds, though in some circumstances the delays can be much longer (e.g., while interacting with a slow computer). One method for dealing with such temporal alignment problems is to use a model that estimates the delay of the feedback – a so-called 'Smith predictor' (Miall *et al*., 1993). Here a controller not only sends signals to the muscles to bring about immediate perceptual changes; it also preserves a copy of the expected perceptual changes and compares them with actual feedback signals when they arrive. Simulation results and behavioral data support the biological reality of this method. Miall *et al*. hypothesized that the cerebellum may play a central role in such prediction.

Feedback

Even when prediction works well, it rarely works perfectly. For a task as simple as moving a stylus to a target, the movement time M to get the stylus to the target increases as the distance D to be covered increases and as the width W of the target narrows. A quantitative relation among these variables was formulated by Paul Fitts (1954) and proved to be so reliable that it came to be called Fitts' law:

$$M = a + b\log(D/W) \quad (1)$$

The functional mechanisms underlying Fitts' law awaited explication by later investigators (Meyer *et al*., 1990), who noted that the higher the velocity of movement (and, in turn, the greater the forces generated), the greater the variability of the endpoints that are reached. Performers confronted with this velocity–variability trade-off must find a minimum time that allows D to be covered with an acceptable range of endpoints. Meyer *et al*. showed that Fitts' law represents an optimal trade-off between distance and time when increases in velocity, D/M, cause proportional increases in endpoint distributions. The argument presented by Meyer *et al*. emphasizes the inherent variability of motor control and the importance of developing internal models that optimally compensate for this variability. A similar view has been advanced by Harris and Wolpert (1998).

MODELS FOR SERIAL ORDER AND TIMING

One domain where understanding of variability has yielded good results is in models of the control of serial order and timing. Early in the study of movement sequencing it was suggested that movements are simply triggered by feedback from earlier movements. This view was rejected when it was found that interruption of feedback does not prevent movements from unfolding normally (Lashley, 1951). Another finding that dispelled the hypothesis was that early aspects of movement sequences often reflect later aspects, indicating that plans are established for forthcoming behavior.

Three sources of behavioral evidence for advance planning have been adduced. Firstly, the time to initiate a series of motor responses (e.g. keystrokes in typewriting) increases with the number of responses to be produced (up to a limit of about seven), provided the response rate is high (Sternberg *et al*., 1978). Secondly, performance errors (e.g. slips of the tongue) reveal implicit knowledge of what will be said or done (Lashley, 1951; Norman, 1981). Thirdly, the properties of correct actions reflect advance information of what will be done later. For example, the way an object is picked up changes as a function of where it will be placed (Rosenbaum *et al*., 1993). Detailed analyses of such effects have led to hierarchical models of motor plans. According to such models, serial ordering of behavior is achieved by the unfolding of high-level goals into lower-level constituents (e.g., MacKay, 1987). This general conception has been supported by studies of brain activity prior to movement (Gazzaniga *et al*., 1998).

Studies of timing likewise suggest hierarchical levels of organization. One of the most influential models of motor timing assumes two levels for timing control: a time-interval or 'clock' level, and a motor-execution level (Wing and Kristofferson,

1973). Patterns of timing variability from tasks where subjects try to generate responses at fixed rates (e.g. steady finger tapping) generally confirm predictions based on this two-tier model. More complex models that permit hierarchically nested timing relations have also been adduced to explain production of complex rhythmic sequences (Vorberg and Wing, 1996).

The fact that hierarchical models apply to timing as well as to serial ordering of behavior fits with the view that timing is essential for perceptual–motor integration, as mentioned above in connection with the Smith predictor. Further studies using brain-imaging and other psychophysical techniques have shown that predictive capabilities make it possible to engage in remarkably vivid motor imagery (Jeannerod, 1994). Such imagery can be useful for modeling one's own performance in tasks calling for critical action decisions (e.g., deciding where to grab an outcropping on a cliff while scaling it) and for improving one's perceptual–motor skills when it is impossible to practice those skills physically (see Schmidt and Lee (1999) for a review).

References

Abbs JH, Gracco VL and Cole KJ (1984) Control of multimovement coordination: sensorimotor mechanisms in speech motor programming. *Journal of Motor Behavior* **16**: 195–231.

Bernstein N (1967) *The Coordination and Regulation of Movements*. London, UK: Pergamon.

Bizzi E, Hogan N, Mussa-Ivaldi FA and Giszter S (1992) Does the nervous system use equilibrium-point control to guide single and multiple joint movements? *Behavioral and Brain Sciences* **15**: 603–613.

Conditt MA, Gandolfo F and Mussa-Ivaldi FA (1997) The motor system does not learn the dynamics of the arm by rote memorization of past experience. *Journal of Neurophysiology* **78**: 554–560.

Duhamel J-R, Colby CL and Goldberg ME (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**: 90–92.

Feldman AG (1986) Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of Motor Behavior* **18**: 17–54.

Fitts PM (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* **47**: 381–391.

Flash T and Hogan N (1985) The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience* **5**: 1688–1703.

Franz E, Eliassen J, Ivry R and Gazzaniga M (1996) Dissociation of spatial and temporal coupling in the bimanual movements of callosotomy patients. *Psychological Science* **7**: 306–310.

Gazzaniga MS, Ivry RB and Mangun GR (1998) *Cognitive Neuroscience: The Biology of the Mind*. New York, NY: WW Norton.

Gielen CC, Vrijenhoek EJ and Flash T (1997) Principles for the control of kinematically redundant limbs. In: Fetter M, Misslisch H and Tweed D (eds) *Three-Dimensional Kinematics of Eye-, Head-, and Limb-Movements*, pp. 285–297. Chur, Switzerland: Harwood.

Harris CM and Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* **394**: 780–784.

von Holst E and Mittelstaedt H (1950) Das Reafferenzprinzip: Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften* **37**: 464–476. [English translation in: von Holst E (1973) *The Behavioural Physiology of Animal and Man: The Collected Papers of Erich von Holst*, translated by R. Martin. London, UK: Methuen.]

Jeannerod M (1994) The representing brain: neural correlates of motor intention and imagery. *Brain and Behavioral Science* **17**: 187–245.

Lashley KS (1951) The problem of serial order in behavior. In: Jeffress LA (ed.) *Cerebral Mechanisms in Behavior*, pp. 112–131. New York, NY: John Wiley.

MacKay DG (1987) *The Organization of Perception and Action: A Theory for Language and Other Cognitive Skills*. New York, NY: Springer-Verlag.

Marteniuk RG and Roy EA (1972) The codability of kinesthetic location and distance information. *Acta Psychologica* **36**: 471–479.

Mechsner F, Kerzel D, Knoblich G and Prinz W (2001) Perceptual basis of bimanual coordination. *Nature* **414**: 69–73.

Meyer DE, Smith JEK, Kornblum S, Abrams RA and Wright CE (1990) Speed–accuracy tradeoffs in aimed movements: toward a theory of rapid voluntary action. In: Jeannerod M (ed.) *Attention and Performance*, vol. XIII, pp. 173–226. Hillsdale, NJ: Lawrence Erlbaum.

Miall RC, Weir DJ, Wolpert DM and Stein JF (1993) Is the cerebellum a Smith predictor? *Journal of Motor Behavior* **25**: 203–216.

Norman DA (1981) Categorization of action slips. *Psychological Review* **88**: 1–15.

Polit A and Bizzi E (1978) Processes controlling arm movements in monkeys. *Science* **201**: 1235–1237.

Rosenbaum DA, Meulenbroek RG, Vaughan J and Jansen C (2001) Posture-based motion planning: applications to grasping. *Psychological Review* **108**: 709–734.

Rosenbaum DA, Vaughan J, Jorgensen MJ, Barnes HJ and Stewart E (1993) Plans for object manipulation. In: Meyer DE and Kornblum S (eds) *Attention and Performance*, vol. XIV: *A Silver Jubilee: Synergies in Experimental Psychology, Artificial Intelligence and Cognitive Neuroscience*, pp. 803–820. Cambridge, MA: MIT Press/Bradford Books.

Schmidt RA and Lee TD (1999) *Motor Control and Learning: A Behavioral Emphasis*, 3rd edn. Champaign, IL: Human Kinetics.

Sternberg S, Monsell S, Knoll RL and Wright CE (1978) The latency and duration of rapid movement sequences: comparisons of speech and typewriting. In: Stelmach GE (ed.) *Information Processing in Motor Control and Learning*, pp. 117–152. New York, NY: Academic Press.

Tweed D and Vilis T (1990) Geometric relations of eye position and velocity vectors during saccades. *Vision Resarch* **30**: 111–127.

Tweed D and Vilis T (1992) Listing's law for gaze-directing head movements. In: Berthoz A, Graf W and Vidal PP (eds) *The Head–Neck Sensory–Motor System*, pp. 387–391. Chichester, UK: John Wiley.

Uno Y, Kawato M and Suzuki R (1989) Formation and control of optimal trajectory in human multijoint arm movement: minimum torque-change model. *Biological Cybernetics* **61**: 89–101.

Vorberg D and Wing A (1996) Modeling variability and dependence in timing. In: Heuer H and Keele SW (eds) *Handbook of Perception and Action*, vol. III: *Motor Skills*, pp. 181–261. London, UK: Academic Press.

Wing AM and Kristofferson AB (1973) Response delays and the timing of discrete motor responses. *Perception and Psychophysics* **14**: 5–12.

Wolpert DM and Ghahramani Z (2000) Computational principles of movement neuroscience. *Nature Neuroscience* **3**: 1212–1217.

Wolpert DM and Kawato M (1998) Multiple paired forward and inverse models for motor control. *Neural Networks* **11**: 1317–1329.

Wolpert DM, Miall RC and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Science* **2**: 338–347.

Further Reading

Gallistel CR (1999) Coordinate transformations in the genesis of directed action. In: Bly BM and Rumelhart DE (eds) *Cognitive Science*, pp. 1–42. San Diego, CA: Academic Press.

Gazzaniga MS, Ivry RB and Mangun GR (1998) *Cognitive Neuroscience: The Biology of the Mind*. New York, NY: WW Norton.

Gielen CC, Vrijenhoek EJ and Flash T (1997) Principles for the control of kinematically redundant limbs. In: Fetter M, Misslisch H and Tweed T (eds) *Three-Dimensional Kinematics of Eye-, Head-, and Limb-Movements*, pp. 285–297. Chur, Switzerland: Harwood.

Kawato M, Furawaka K and Suzuki R (1987) A hierarchical neural network model for the control and learning of voluntary movements. *Biological Cybernetics* **56**: 1–17.

Kawato M and Gomi H (1992) The cerebellum and VOR/OKR learning in models. *Trends in Neuroscience* **15**: 445–453.

Meyer DE, Smith JEK, Kornblum S, Abrams RA and Wright CE (1990) Speed–accuracy tradeoffs in aimed movements: toward a theory of rapid voluntary action. In: Jeannerod M (ed.) *Attention and Performance*, vol. XIII, pp. 173–226. Hillsdale, NJ: Lawrence Erlbaum.

Rosenbaum DA (1991) *Human Motor Control*. San Diego, CA: Academic Press.

Rosenbaum DA and Collyer CE (eds) (1998) *Timing of Behavior: Neural, Psychological, and Computational Perspectives*. Cambridge, MA: MIT Press.

Wolpert DM, Miall RC and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Science* **2**: 338–347.

Natural Language Processing: Models of Roger Schank and his Students

Intermediate article

Roger C Schank, Northwestern University, Evanston, Illinois, USA
David B Leake, Indiana University, Bloomington, Indiana, USA

CONTENTS

Over many years, Roger Schank and his students developed a series of theories and computer models of natural language processing, guided by two main tenets: that natural language processing cannot be studied in isolation from the purposes for which language is used, and that meaning and world knowledge may be crucial at even the earliest points in the understanding process. Through an interplay between study of human reasoning, development of theories, and experiments with computer models, this work has identified important research questions and shed light not only on natural language processing but on a broad range of cognitive processes such as planning, learning, explanation, and even creativity.

INTRODUCTION

A central question in the study of natural language processing (NLP) is what the relationship is between NLP and of other aspects of cognition. One possibility, pursued in a substantial body of research, is to consider NLP to be a fairly independent area, concerned primarily with linguistic issues that can be studied independently from questions of how people use the content of that language – how they understand, remember, learn, and achieve their goals in the world. Over several decades, Roger Schank's research group has developed a succession of theories guided by the opposite viewpoint: that language and thought are inextricably connected. In this view, communication is at the heart of language; the goal of NLP is to understand how people extract and generate the content of communication. This, in turn, requires addressing a wide range of general issues beyond language *per se*. For example, the group has studied questions such as how conceptual expectations guide understanding, the role of inference in the understanding process, the knowledge structures into which new information is placed, how knowledge is organized, how it is learned, and how it is brought to bear on future understanding. The group's methodology combines artificial intelligence, cognitive science, and psychology. Theory building and testing are guided by a continual interplay between psychological data and experiments with computer models.

This article describes some of the major milestones of the group's models of parsing, inferencing, understanding, learning, memory, and explanation, and explains the tenets of their work. This succession of projects has both answered questions, and, equally importantly, raised important new questions. This process has not only sharpened our understanding of NLP, but also provided theories of basic processes of thought.

THE INTEGRATED PROCESSING HYPOTHESIS

In the 1950s and 1960s, the availability of computers stimulated attempts to develop machine translation systems. However, the early attempts could not handle problems such as ambiguity and implicit content. As it became clear that a theoretical foundation was needed, in the mid-1960s many researchers turned to linguistics and theories of syntax, developing approaches with varying degrees of separation between syntax and semantics. At one extreme were theories of autonomous

syntax, which treated the extraction of syntactic structure as an autonomous process that preceded, and provided input for, later semantic processing. Other approaches treated syntax and semantics as largely 'decomposable', with processing controlled by syntax but syntactic processing occasionally querying a semantic component; or as following separate, but more cooperative processes.

It seems clear that semantic and pragmatic knowledge (which we will call conceptual knowledge) play an important role during human language processing. For example, someone with a limited grasp of a foreign language can often make sense of magazine articles in that language, and finds it much easier to do so than to paraphrase the article in the foreign language: knowledge of the domain can guide understanding, even when syntactic knowledge is weak. Likewise, there is strong evidence that humans generate conceptual expectations as they understand. For example, most people who read 'the old man's glasses were filled with sherry' are surprised by the last word: their knowledge of old age makes eyeglasses the expected meaning for 'glasses', before the conclusion of the sentence is read. Such observations shaped the 'integrated processing hypothesis' that has guided the work of Schank's group: 'meaning and world knowledge are often crucial at even the earliest points in the process of understanding language' (Schank and Birnbaum, 1984, p. 212). Contrary to some misunderstandings of their work, this hypothesis does not reject the use of syntactic knowledge, nor does it claim that early processing is purely semantically based. It means that understanders use whichever available knowledge is most useful and apply it in a unified way, with a single mechanism controlling the application of both types of knowledge. We describe below a series of projects exploring this hypothesis in a series of computer models with increasing levels of integration, guided by increasingly sophisticated theories of the knowledge that guides understanding and how it is applied.

CONCEPTUAL DEPENDENCY THEORY

Modeling the role of language as communication requires theories of how to represent the content of language, as well as theories of the processes used to map language to and from that representation. Conceptual dependency (CD) theory, one of the first theories developed by Schank's group, is often thought of as the theory of a specific set of primitives for knowledge representation, but in fact it is more: it is a theory of how knowledge drives the understanding process.

CD theory has four main tenets (Schank and Riesbeck, 1981, p. 13):

1. The representation of events is done apart from the words used to encode those events.
2. The task of the understander is to represent the underlying events, rather than to represent the sentences themselves.
3. Such representations are governed by a set of rules regarding the fitting of inputs into a predefined representation scheme.
4. The rules for 'filling in the slots' of those representations are the rules that are the basis of language understanding.

CD represented the meaning elements necessary to represent actions, causal relations, states and state changes, in a language-independent form. The initial set of primitive acts included 11 acts, such as:

- PTRANS: to change the location of an object.
- MTRANS: to transfer information, either within memory or between people.
- PROPEL: to apply a force to an object, in a given direction.
- INGEST: to take something inside an animate object.

In CD, similar events are represented in similar ways, regardless of differences in the sentences used to express them. For example, because both reading a newspaper and watching the news on television involve transmission of information, their CD representations both use MTRANS. Consistently with the aim of CD to represent the underlying events, rather than the sentences themselves, a single CD primitive may be used to represent a wide range of verbs (for example, PTRANS can represent the events underlying verbs such as 'go', 'drive', 'fly', or 'carry' a cargo to a destination). Each primitive licenses certain inferences, and the meaning of each primitive may be seen as coming from what a reasoner may infer from the act (e.g., that PTRANS involves a change of state from one location to another).

Each primitive act has associated conceptual cases which modify the action. PTRANS has five cases: the 'objective' case (the object whose location is being changed), the 'directive' cases TO and FROM (describing the change in location), the 'actor' (who effects the change), and the 'instrumental' case (how the change is effected, e.g., an object might be moved by the actor PROPELing it).

The MARGIE system (Schank, 1975) was developed to test the usefulness of CD and to

demonstrate the ability of an understanding system to successfully map natural language to – and back from – an independent conceptual representation. In performing this process, MARGIE could perform machine translation, but the fundamental research focus was on modeling understanding. The system was composed of three modules: a parser, Christopher Riesbeck's ELI, which performed conceptual analysis, taking natural language input and using the low-level semantic expectations from CD structures to generate a representation in CD; a memory module, Charles Rieger's MEMORY, to make plausible inferences from the CD representation; and Neil Goldman's natural language generator BABEL, which took CD input from either of these modules to produce either paraphrases of the input or sentences expressing the inferences in a variety of natural languages.

CONCEPTUAL ANALYSIS

In conceptual analysis, the task of parsing is seen primarily as connecting concepts referenced in a sentence and selecting appropriate meanings to resolve potential ambiguities. In this process, as explored in ELI and later in Larry Birnbaum and Mallory Selfridges's CA (Schank and Riesbeck, 1981, pp. 318–353), the conceptual cases from incomplete CD representations provide expectations that are used to extract meaning from sentences and resolve potential ambiguities. For example, consider the conceptual analysis process for the sentence 'Mary ate a macintosh'. When the conceptual analyzer encounters the word 'ate' it maps that word to the conceptualization INGEST, which provides the expectation that its object will be a food. Because the food sense of 'macintosh' (a type of apple) satisfies the expectation, that meaning is selected instead of alternatives that are unlikely in context, such as a type of raincoat or computer.

Conceptual analysis provided a first, partial step towards realizing the integrated processing hypothesis. Although it did not yet bring higher-level knowledge to bear, it constructed a conceptual representation using a unified control structure to apply conceptual knowledge throughout its parsing process, without requiring an independent syntactic analysis phase (Schank and Birnbaum, 1984).

CAUSAL COHERENCE AND INFERENTIAL MEMORY

Extracting the meaning of individual sentences is only part of understanding: understanding their connections and ramifications is just as important. For example, a person would not be said to understand 'John won a million dollars and took a trip around the world', without also understanding that John's new wealth enabled the trip; or to understand 'John bought a book', without knowing that it is likely that he wants to read it; or to understand 'Mary sold a car' without knowing that she no longer has the car, and that someone else does. MARGIE's memory module generated plausible inferences from CD representations generated by ELI, in a spontaneous inference process. For example, given the input 'John gave Mary an aspirin', it generated inferences such as 'John believes that Mary wants an aspirin', 'Mary is sick', 'Mary wants to feel better', and 'Mary will ingest the aspirin'. This process identified the connections between new information and information in memory, in order to establish coherence.

MARGIE succeeded in bringing inferential memory to bear on understanding, but also demonstrated important limitations of general-purpose inference chaining models. First, despite the usefulness of general-purpose knowledge, such knowledge is not sufficient in itself: some inferences depend on highly context-specific knowledge. For example, a program (or person) with no knowledge of Japanese culture will be unable to infer why customers remove their shoes when entering a Japanese restaurant. Secondly, MARGIE's memory module identified a serious processing issue: that unguided inferencing is computationally infeasible. As inferences are drawn from an input, and then from those inferences, and so on through many inference levels, the rapid increase in the number of possible inferences will soon overwhelm any reasoning system. This limitation suggested that a new theory was needed to account for the ease of human inferencing.

SCRIPTS, PLANS, AND GOALS

The theory of scripts (Schank and Abelson, 1977) was developed as a solution to the problems of controlling inference cost and providing context-specific guidance. A script is a stereotyped sequence of events that take place in a particular context, such as the events involved in going to a restaurant (wait for a table, sit down, order, etc.). The theory of scripts proposed that in familiar situations, people rely on scripts to provide overall structure and context-relevant expectations for what will happen next.

Scripts distinguish normative features (e.g. that restaurant visits normally include ordering), from

unique features of a given instantiation (e.g. what was ordered on a given restaurant visit). This knowledge provides a basis for summarization of stories: in good summaries, the noteworthy points of a story are described, and events that can be routinely filled in are left unstated.

The application of script theory to understanding was demonstrated in Richard Cullingford's program SAM (Schank and Riesbeck, 1981, pp. 75–119). The system used ELI to generate a conceptual representation of sentences, to which scripts were then applied. The system could read newspaper articles and summarize them in a number of languages. In addition, the question-answering system QUALM, developed by Wendy Lehnert, could use high-level, script-based knowledge to fill in missing information in a focused way, and to answer questions going far beyond information contained explicitly in the texts it processed, even to the extent of understanding and answering questions about what did not happen in the stories. For example, if John ordered a hamburger, but left because it was burnt, SAM could use its knowledge of normative events to correctly answer the question: 'Why didn't John eat the hamburger?'

Script theory provides a model of understanding for routine, stereotyped events, and experiments by Bower *et al* (1979) provided psychological support for the existence of scripts as culturally shared human knowledge structures. However, script theory does not attempt to account for understanding the more interesting stories that concern novel situations. Understanding such stories requires being able to understand their characters' motivations: their plans and goals. This led Schank's group to develop a theory of the requirements for understanding goal-based behaviors: of plans and goals, and of themes, which provide background knowledge about why people have particular goals (for example, role themes provide expectations for the goals of actors in certain societal roles, such as waiters, doctors, or police). In conjunction with the study of plans, goals, and themes, the group developed theories of the processes involved in understanding plan-based stories, as demonstrated in Robert Wilensky's PAM (Schank and Riesbeck, 1981, pp. 136–179).

Study of scripts, plans, and goals increased our understanding of the high-level knowledge needed in understanding; the next step was to develop models of how this knowledge could be brought to bear at the first phases of the understanding process. The group's first attempt to integrate high-level knowledge into early understanding was Gerald DeJong's program FRUMP (Schank and Abelson, 1977, pp. 204–210). FRUMP used its scripts directly to guide processing of stories, skimming their texts for script-relevant content. This strongly expectation-based system was sufficiently fast and robust to process and summarize news stories directly from the newswire, in real time.

DYNAMIC MEMORY THEORY

People learn from what they read, and what they learn colors their interpretations of later experiences. SAM, PAM, and FRUMP, however, did not learn: no matter how many times they processed a particular story, they processed it in exactly the same way. The next major effort of Schank's group was to model the relationship between understanding, learning, and memory. One of the findings they sought to address was a surprising result from the experiments of Bower *et al* that people who read script-based stories with similar components may confuse their components when asked to recall events. For example, if a story involves visits to both a dentist and a doctor, picking up a magazine in one waiting room might be remembered as taking place in the other. This result suggested that when events are stored in human memory, they are not associated with independent scripts, but instead with knowledge structures that share components.

This insight led to the development of the theory of memory organization packets (MOPs) (Schank, 1982). Like scripts, MOPs characterize event sequences; unlike scripts, they allow a sequence to be composed of a number of lower-level components which can be shared by a number of MOPs. The basic components, which are called scenes, normally describe events that take place in a single location, with a single purpose, and in a single time interval. For example, the MOP for professional office visits contains steps that take place in the office, such as waiting in the waiting room, receiving the service, and paying the bill. Scenes such as waiting in the waiting room are specific to professional office visits, and are directly associated with that MOP; other steps in the sequence, such as receiving the service, are filled in by other MOPs or by scenes from other MOPs that are simultaneously active. For example, the 'get service' step for a doctor's visit is provided by the MOP `M-Doctor-Service`, and the 'pay' scene is supplied by `M-Contract`.

MOPs are organized hierarchically, inheriting properties of more abstract MOPs: `M-Doctor-Service` is itself a specification of the general

MOP `M-Use-Service`. MOPs also organize information about initiating and follow-up steps – for example, how to get to the appointment – which may also be filled in by other MOPs. Figure 1 illustrates the MOP and scene organization for some of the components required to process a visit to a doctor's office.

The act of payment after a professional office visit is usually standard regardless of the particular profession involved. Although `M-Doctor-Visit` and `M-Dentist-Visit` are distinct MOPs, both share the same `Pay` scene. When these MOPs are applied during understanding, their details are reassembled by combining their constituent scenes. This principle of sharing structures has a significant functional benefit: it enables cross-contextual learning. Once learning in one context has changed a scene, that learning is automatically reflected in all MOPs using that scene.

MOPs are not only a processing structure, but also a memory structure. Events are organized in memory under the scenes used to process them, and an instantiation of a MOP is retrieved by reassembling it from its constituent scenes. When those scenes are shared across different MOPs, events that occurred in one MOP may be remembered as occurring in another, explaining the observed recall confusions. Thus MOP theory explains both the cross-contextual learning capabilities and script confusions exhibited by people: both are consequences of a dynamic memory.

MOP theory is intimately related to how remindings occur. When people encounter new situations, they often find that other similar situations spontaneously come to mind. Because events are stored in memory under the scenes used to process them, when a MOP is used to understand a new situation, records of those events may become available. This is especially useful in helping us to understand expectation failures, when a new event deviates from a MOP's expectations and the understander is reminded of a previous event that deviated in a similar way. Remembering a past failure and its repair, for example, may suggest a way to repart a new failure.

MOP theory and memory organization were studied in Janet Kolodner's CYRUS (Schank, 1982, pp. 207–218), a system that studied the use of memory to answer questions after understanding. In addition to modeling how new MOPs are built and organized in memory, the system examined the process of memory search in which questions are generated, refined, and elaborated to answer questions in flexible ways. Michael Lebowitz's system IPP modeled MOP-based learning in an integrated understander (Schank, 1982, pp. 197–207). Rather than relying on static knowledge structures, IPP formed inductive generalizations from events it read about to build new MOPs, which it then used to interpret later stories. This enabled it to recognize routine events and to point out potentially interesting deviations. As an example of its processing, consider the following story from the *New York Times*:

> An Arabic speaking gunman shot his way into the Iraqi Embassy here [Paris] yesterday morning, held hostages through most of the day before surrendering to French policemen, and then was shot by Iraqi security officials as he was led away by French officers.

IPP's language processing was guided directly by its MOPs. Consequently, it could ignore words without significant conceptual content (e.g., the word 'way' in 'shot his way'), and could immediately process significant words to generate high-level expectations to guide future processing (e.g., processing the word 'hostages' activated a MOP for terrorist acts).

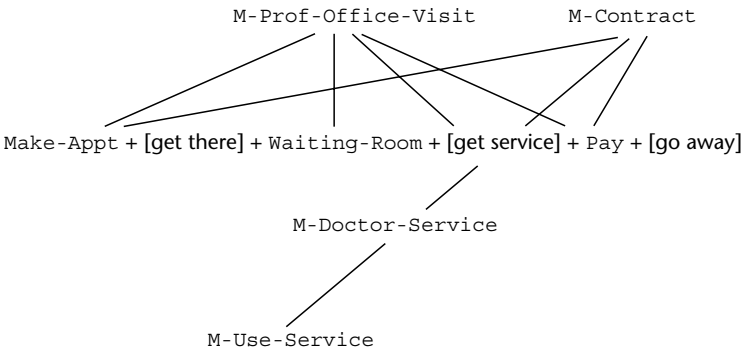


Figure 1. Some of the MOPs and scenes involved in understanding a visit to a doctor's office (Adapted from Schank and Birnbaum, 1984).

The beginning of this story is a fairly routine story of terrorism, but it is startling that the Iraqi officials shot the gunman after he had surrendered. IPP could recognize from experience that this shooting was exceptional, but could not reconcile the novel event with its other knowledge by explaining why it took place. How to model this aspect of understanding became the next research focus of Schank's group.

EXPLANATION PATTERNS AND CREATIVITY

When surprising things happen, people try to understand them by generating explanations. These explanations can in turn help guide the generation of new knowledge structures, the revision of existing knowledge, or the re-indexing of knowledge in memory. Work on explanation patterns, begun in the mid-1980s, combines the study of how to detect anomalies, how to generate new explanations, and how to adapt prior knowledge in novel ways – a form of creativity.

Building explanations from scratch is computationally intractable: it involves searching an enormous space of alternatives, often with very limited information to help narrow the set of possibilities. Thus it is a difficult question how to account for human facility at generating plausible explanations. In the SWALE project, the group investigated explanation using case-based reasoning: building new explanations by retrieving and adapting prior explanations (Schank, 1986; Schank and Leake, 1989). The SWALE system, developed by Alex Kass, David Leake and Chris Owens, was a story understanding system that used MOP-based understanding to account for routine situations and case-based explanation to explain novel events when its MOP-based understanding failed.

SWALE's namesake example was the story of a 3-year-old racehorse named Swale. Swale was in peak form, and had recently won two major victories, when a shocking headline appeared on the front page of the *New York Times*: Swale was dead. The death attracted much interest, both within and outside the racing community. A vast range of possible causes could be hypothesized for Swale's death, and little information was available to constrain the alternatives. Nevertheless, human explainers had little trouble generating plausible hypotheses, and prior experiences were often cited as the reasons for the hypotheses proposed. For example, one veterinarian's immediate reaction to the news was: 'This sounds like an aneurysm. I've seen this sort of thing before.' Swale's death reminded a Yale student of the death of the runner Jim Fixx, who died when the exertion of recreational jogging overtaxed a hereditary heart defect; the student therefore hypothesized that Swale's death was also caused by a heart attack. The explanation for Fixx's death does not apply directly – Swale was unlikely to do recreational jogging – but a minor adaptation of the explanation, substituting horse racing for jogging, produces the plausible explanation that the stress of running in a race overtaxed a hereditary heart defect. This process of retrieving and adapting prior explanations is at the heart of SWALE's understanding process. Later experiments by Read and Cesa (1991) provided psychological support for a human tendency to favor explanations supported by prior cases.

With sufficiently flexible retrieval and adaptation, the case-based explanation process also models a form of creativity. For example, Swale's death reminded one student of the death of Janis Joplin. Joplin was driven to recreational drug use by the stress of fame, and died from accidentally taking an overdose of recreational drugs. Little of that explanation applies to Swale: racehorses do not take recreational drugs. However, an essential part of the explanation is applicable: that Swale might have died of a drug overdose. By searching through its world knowledge to find potential supports, SWALE generates the hypothesis that Swale died of an overdose of performance-enhancing drugs administered by a trainer.

LATER MODELS

We have described a few landmark steps in the development of these theories of understanding, but many other models have been developed, and continue to be developed, by Schank, his students, and their own students. Charles Martin and Christopher Riesbeck's 'direct memory access parsing' (DMAP) (Riesbeck and Schank, 1999, pp. 319–352), for example, fully integrates parsing and understanding. In the DMAP model, expectations from a MOP-based memory are applied by a single, uniform parallel process. In this model, the goal of parsing is to interpret and impose structure – summarization is replaced with the formation of opinions – and the result of parsing is a new state of memory.

As the theories described above go beyond language *per se*, their ramifications also go beyond language. For example, the work on reminding and memory provided a foundation for research in case-based reasoning, which has had extensive influence beyond language processing for tasks

such as planning, design, and interpretation (Riesbeck and Schank, 1999; Kolodner, 1993; Leake, 1996). Looking at how people communicate led the group to study stories and storytelling: why people tell stories, how stories are indexed in memory, how they are understood, and the relationship between stories and intelligence (Schank, 1990). (*See* **Learning through Case Analysis**)

Recently the group has applied theories of human learning to education. This has led to the development of educational systems that support students' case acquisition by learning by doing in goal-based scenarios. Goal-based scenarios provide rich learning environments in which students learn skills and conceptual knowledge by performing activities in pursuit of compelling goals (Schank, 1998). Will Fitzgerald's research on indexed concept parsing has developed methods for robust, practical conceptual parsers to enable students to communicate with these educational systems.

CONCLUSION

Over many years, the work of Roger Schank and his students on natural language processing has been driven by the view that a theory of NLP must be a theory of how language is used to communicate information, which in turn requires addressing the cognitive processes underlying understanding, memory, and learning. Hence, in their view, the essential issues of natural language processing include issues such as the types of knowledge structures memory contains, how they are brought to bear, and how they change with experience.

The work on CD, scripts, plans, goals, and MOPs has all been aimed at understanding the knowledge structures that enable people to understand the world that language describes, and the processes by which these knowledge structures are applied and refined. This research has contributed both to artificial intelligence, by advancing the performance of computer systems, and to the understanding of human reasoning and learning. These models have generated a succession of research questions as they have identified issues that are central not only to NLP, but also to human cognition in general. These questions continue to drive research.

References

Bower G, Black J and Turner T (1979) Scripts in memory for text. *Cognitive Psychology* **11**: 177–220.

Kolodner J (1993) *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Leake D (ed.) (1996) *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, CA: AAAI Press/MIT Press.

Read S and Cesa I (1991) This reminds me of the time when …: Expectation failures in reminding and explanation. *Journal of Experimental Social Psychology* **27**: 1–25.

Riesbeck C and Schank R (1999) *Inside Case-Based Reasoning*. Hillsdale, NJ: Erlbaum.

Schank R (1975) *Conceptual Information Processing*. Amsterdam: North-Holland.

Schank R (1982) *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge, UK: Cambridge University Press.

Schank R (1986) *Explanation Patterns: Understanding Mechanically and Creatively*. Hillsdale, NJ: Erlbaum.

Schank R (1990) *Tell Me a Story: A New Look at Real and Artificial Memory*. New York, NY: Scribner's.

Schank R (1998) *Inside Multi-Media Case-Based Instruction*. Hillsdale, NJ: Erlbaum.

Schank R and Abelson R (1977) *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

Schank R and Birnbaum L (1984) Memory, meaning, and syntax. In: Berer TG, Carroll JM and Mil LA (eds) *Talking Minds: The Study of Language in Cognitive Science*, pp. 209–251. Cambridge, MA: MIT Press.

Schank R and Leake D (1989) Creativity and learning in a case-based explainer. *Artificial Intelligence* **40**(1–3): 353–385. [Reprinted in: Carbonell J (ed.) (1990) *Machine Learning: Paradigms and Methods*. Cambridge, MA: MIT Press.]

Schank R and Riesbeck C (1981) *Inside Computer Understanding: Five Programs with Miniatures*. Hillsdale, NJ: Erlbaum.

Further Reading

Kolodner J (1993) *Case-based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Leake D (ed.) (1996) *Case-based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, CA: AAAI Press/MIT Press.

Riesbeck C (1986) From conceptual analyzer to direct memory access parsing: *an overview*. In: Sharkey N (ed.) *Advances in Cognitive Science*, chap. 8, pp. 236–258. New York: Wiley.

Schank R (1986) *Explanation Patterns: Understanding Mechanically and Creatively*. Hillsdale, NJ: Lawrence Erlbaum.

Schank R (1998) *Inside Multi-Media Case-Based Instruction*. Hillsdale, NJ: Lawrence Erlbaum.

Schank R (1999) *Dynamic Memory Revisited*. Cambridge, UK: Cambridge University Press.

Schank R and Langer E (1994) *Beliefs, Reasoning, and Decision Making: Psycho-Logic in Honor of Bob Abelson*. Hillsdale, NJ: Lawrence Erlbaum.

Schank R, Riesbeck C and Kass A (eds) (1994) *Inside Case-Based Explanation*. Hillsdale, NJ: Lawrence Erlbaum.

Natural Language Processing, Disambiguation in

Intermediate article

Graeme Hirst, University of Toronto, Toronto, Ontario, Canada

CONTENTS

Ambiguity pervades all levels of language. Any computer system that uses the meaning of natural language must disambiguate its input.

INTRODUCTION

Language is rife with ambiguity: a single utterance can, in principle, have many different interpretations or meanings. Usually, however, the speaker or writer intends just one of these meanings – usually, only one of them will make sense – and humans are adept at rapidly determining which one was intended, not even consciously noticing the others. Computer language-processing systems that use the meaning of an utterance in their task must therefore, like people, disambiguate their input. (*See* **Natural Language Processing**)

For example, an English-to-French machine translation system must decide whether to translate the word *duty* as *douane* or *devoir*, depending on whether its meaning in the utterance relates to an import tax or to a moral or legal obligation. A speech recognition system that hears [djuti] must decide whether it should be rendered as *duty* or *due tea*. Suppose a text interpretation system comes across this excerpt (from Jane Austen's *Persuasion*):

> The Admiral, after taking two or three refreshing turns about the room with his hands behind him, …

It must then decide whether the prepositional phrase *with his hands behind him* describes the room or the Admiral's manner of taking refreshing turns. And in either case, it must also decide exactly whose hands are behind whom. (*See* **Machine Translation**)

As these examples show, ambiguity can occur in many ways and at all levels of language. This article will cover ambiguities of syntax and of word senses, including the special case of pronouns. Most of the article applies to both spoken and written text, and the terms *speaker* and *writer* will be used interchangeably. Ambiguity can be treated more specifically at the phonetic level and at the morphological level. (*See* **Speech Perception and Recognition, Theories and Models of**; **Morphological Processing**)

LEXICAL DISAMBIGUATION

There are two kinds of lexical ambiguity – that is, of ambiguity of words. The first is ambiguity as to the syntactic category, or part of speech, of a word in an utterance: for example, the word *flies* can be used as either a verb or a noun. While only about 12 percent of the word types of English are ambiguous as to category, they tend to be the more common words, representing about 40 percent of the word tokens that are uttered (DeRose, 1988). The second kind of lexical ambiguity is ambiguity of meaning even after the part of speech is determined. (*See* **Lexical Ambiguity Resolution**)

Part-of-speech Tagging

Ambiguities of syntactic category are resolved as part of the process of 'part-of-speech tagging' – labeling each word in an input sentence with its category – which is the first stage of processing in many applications of natural language processing. The rules of grammar constrain the allowable sequences of syntactic categories – in English, for example, the base form of a verb may not immediately follow the definite article *the* – and about 60 percent of word tokens are not ambiguous as to category (including, in English, the articles *the* and *a*). Consequently, the category of a word can be resolved with a high degree of accuracy, 97 percent or better, just by looking at the categories of a few preceding words. It is easy to construct examples in which a much greater number is required, but such cases are rare in practice. Part-of-speech tagging is usually regarded as a probabilistic process in

which the most likely category is chosen in light of the two preceding words and the potential categories of the word under consideration. Two sources of information are thus required: a lexicon of words, listing the allowable categories of each, and knowledge of allowable sequences of categories along with their probabilities of occurrence. (*See* **Natural Language Processing, Statistical Approaches to**; **Lexicon**; **Lexicon, Computational Models of**)

The Nature of Word Senses

A glance at a page of a dictionary reminds us that it is only a small minority of words – mostly technical terms – that have only a single sense. When the senses of a word are closely related, the word is said to be polysemous. For example, *window* can mean either an opening in a wall (*crawled through the window*) or the glass that fits in the opening (*broke the window*). When the senses are completely different, the word is said to be homonymous. For example, *ash* can mean either a tree or the residue from combustion. A single word may be both homonymous and polysemous: *bank* is homonymous in its senses relating to financial institutions and to the edge of a watercourse; but, when pertaining to a financial institution, it is polysemous in that it can denote both the institution and the building in which the institution does business. In speech recognition, word sense ambiguity arises from similarity of sound rather than spelling; thus disambiguation is required between *see* and *sea*.

The conventional view of word senses assumes that for each word there is a fixed inventory of senses to decide among, and that for any particular utterance, to disambiguate is to choose exactly one of these senses as 'correct'. The assumption of a fixed inventory of senses has been challenged by many researchers (e.g. Kilgarriff, 1997), who point to the wide disparities in treatments of the same word by different lexicographers in different dictionaries, especially with regard to fine-grained distinctions between polysemous senses: what is one sense for one lexicographer might be two or three for another. And people often find it hard to decide which fine-grained dictionary sense best represents the meaning of a word in its context (Kilgarriff, 1992); often, they will say that a word is being used in two senses at once. For example, *Nadia visited the bank to get some money* seems to invoke *bank* as both building and institution simultaneously. Can we reasonably expect or require a computer to be more precise or decisive about word senses than people are? Often, it doesn't matter: in many applications of natural language processing, very fine-grained word sense disambiguation is unnecessary, and it suffices to resolve homonymy and perhaps coarse-grained polysemy at a level where there is reasonable agreement as to the inventory of senses. For example, a program that translates English to French needs to know whether an occurrence of *bank* is used in a financial or river-related sense in order to choose the correct translation; but, if it is used in a financial sense, the program need not decide between the institution and the building as the translation is the same in either case. (*See* **Word Meaning, Psychology of**; **Word Recognition**)

Methods of Word Sense Disambiguation

When an ambiguous word is a member of more than one syntactic category, part-of-speech tagging allows senses not associated with the category in which the word is being used to be eliminated from consideration; occasionally this yields a unique sense. A unique sense can also be assigned if the word is recognized as part of an unambiguous lexicalized compound; for example, if *private school* is listed as a phrase with its own meaning, there is no need to choose among the different senses of *private* and of *school*. But usually more sophisticated methods are required.

Many such methods are based on selectional restriction, relationship to the topic of the text, or both. In addition, the relative frequency of two different senses may be used to break a tie between them when other methods are unable to choose. In particular, when one or more of a word's senses are relatively rare, they may be eliminated from consideration unless there is positive evidence for them: for example, the noun *email* in its now-rare sense of enamel should be ruled out in favor of the electronic mail sense unless there is some particular reason to suppose that enamel is intended. Data on how frequent each sense of a word is can be derived from large corpora of text that have been disambiguated by humans or by semi-automatic methods with human verification. The accumulation of sufficient sense frequency data to adequately cover a language is an enormous task, however, and so far only relatively small sense-tagged corpora exist for English (Resnik and Yarowsky, 1999). Like relative frequency, the 'one sense per discourse' heuristic (Yarowsky, 1995) can serve as an adjunct to any other method. This heuristic relies on the fact that it is rare in practice for a homonym to be used in more than one sense within

the same text or discourse; so if, for example, the word *crane* occurs five times in a text and some disambiguation method deems it to be a bird in four instances and construction equipment in the other, then the latter is almost certainly wrong and should be corrected to the majority vote.

Selectional restrictions are the semantic constraints that a word sense may place on the senses of other words that combine with it. For example, the verb *eat* requires in literal language that its subject be an animate being and its object be something edible; so in *the mouse ate the corn*, we favor *mouse* as rodent rather than computer equipment and *corn* as cereal rather than callus. Metaphor and other kinds of nonliteral language can violate selectional restrictions (*the photocopier ate my report*), so such restrictions are helpful but not absolute constraints. Selectional restrictions vary in their degree of specificity: *elapse* accepts only time or a unit of time as its subject, whereas many different kinds of things can *grow*.

For a natural language processing system to use selectional restrictions, it first needs a knowledge base of the restrictions pertaining to each word sense, but no such knowledge base yet exists, and the creation of such a resource would be a large and poorly defined lexicographical task (the FrameNet project (Johnson and Fillmore, 2000) is a step in this direction). Resnik (1998) has proposed a process that can construct such a knowledge base automatically from a parsed corpus and an online hierarchical thesaurus such as WordNet (Fellbaum, 1998). For example, if the corpus contains examples of forms of the verb *drink* with objects such as *coffee, wine*, and *water*, the process can learn, by looking up these words in the thesaurus, that *drink* tends to select objects that are beverages or liquids. Resnik's experiments with the process showed that the information it derived was helpful, but of course not by itself sufficient for reliable disambiguation of word senses.

Many methods of word sense disambiguation have tried to capture the intuition that a good disambiguation cue, especially for homonyms, is the existence of a general semantic relationship between one of the candidate senses and those of nearby words in the text. For example, in proximity to the words *garden* and *pest*, the word *mole* is much more likely to refer to a mammal than to a skin blemish or a chocolate sauce. More generally, the topic of the text as a whole can be a helpful cue. The problem is how to make this idea precise and determine the semantic relationships.

Lesk (1986) proposed that dictionary definitions could be used for this. From an online dictionary, the definitions of all content words within, say, 100 words of the target word are found. Regarding this set of definitions as nothing more than a 'bag of words', with no consideration of the structure of the sentences or even the order of the words, the candidate sense of the target word is chosen that contains in its own definition more of the words in the bag than any of the other candidates. For a simplified example with just one word of context, consider the word *keyboard* in the phrase *the keyboard of the terminal*: its dictionary definition includes, in one of its senses, the word *computer*, as does one of the senses of *terminal*; accordingly, this sense of *keyboard* is chosen. Observe that *terminal* is similarly disambiguated. Lesk's method is surprisingly effective given its simplicity, and serves as the baseline against which more complex methods are compared (Kilgarriff and Palmer, 2000).

An example of a more complex method is the use of naive Bayesian classification to classify words according to which sense of each ambiguous word they tend to be associated with. For example, *money* tends to be associated with the financial sense of *bank*, and so do the words *loan* and *mortgage*, but *time* does not and *grass* is probably a contraindication. By looking at a very large corpus of text in which each word is tagged with its correct sense, and counting the number of times that each sense occurs with various other words in its proximity, we can compute the probability of any given word occurring in the proximity of each sense. Then, when disambiguation is necessary, the probability of each sense can be computed in the context of the nearby words, even if those words do not all indicate the same sense, and the sense with the greatest probability can be chosen. This method assumes that all the words in the context are conditionally independent of one another: the probability of seeing one word in context is independent of seeing any other word in the same context. Obviously, this is not true in practice, almost for the very reason that we want to use this method: words of related meaning tend to cluster. None the less, the method gives reasonable results. (*See* **Machine Learning**; **Natural Language Processing, Statistical Approaches to**)

However, this method is limited by the need for sense-tagged corpora as training data. Sufficient data do not exist to cover English, let alone less studied languages. Researchers have sought methods of circumventing this limitation. Yarowsky (1992) proposed that naive Bayesian classification could be used if the goal is not to determine the fine-grained sense of an ambiguous

word but merely an indication of the topic with which it is associated: in effect, resolution of homonyms, which, while coarse-grained, is none the less useful in many applications such as information retrieval. For example, instead of having to determine separately the probability that the word *money* indicates a certain sense of *bank* and so do *deposit* and *account* and *river* and *canal* and *creek*, we instead determine that any word related to finance indicates one sense of *bank* (or one group of senses) and any word related to watercourses indicates another. Yarowsky used the categories of *Roget's Thesaurus* as his set of topics. In an experiment on 12 ambiguous words that appeared in a total of 39 thesaurus categories, he determined, from a corpus of 10 million words, what other words were both frequent and salient as indicators of each of the thesaurus categories in which those words appeared; he then used these words in a naive Bayesian process to classify occurrences of the same 12 words in a test corpus. The results were very good for words such as *mole*, whose senses are generally topic-specific, but not for words such as *interest*, whose senses tend to cut across topics.

But while this method avoids the need for a sense-tagged corpus, it still requires supervised training – that is, its learning phase is still based on some predefined knowledge source, in this case the thesaurus. Yarowsky (1995) has also proposed a method by which decision lists for disambiguation can be learned by unsupervised training. A decision list is an ordered sequence of very specific conditions for classifying a word by meaning: for example, a decision list for the word *bass* might include the conditions 'if the next word is *player*, the topic is music' and 'if the next word is *are*, the topic is fish'. The list is derived from an extremely large corpus, along with a 'seed' – an extremely strong cue – for each sense of the ambiguous word (*bird* and *construction* could be seeds for *crane*). Because the corpus is so large – 460 million words in Yarowsky's experiments – the seeds are sufficient to indicate a number of definite occurrences of each sense, whose context words, in turn, suggest additional cues to each sense. When some of the data have been thus tagged, a classification algorithm is used to find additional rules. The process then iterates, alternating with the 'one sense per discourse' heuristic, until most or all occurrences of the ambiguous word in the corpus have been tagged. The resulting decision lists give a disambiguation accuracy similar to that of the thesaurus-based method.

Both of Yarowsky's methods require separate training for each ambiguous word, so in practice they have been tried only on a few test words. The task of using these methods to cover all ambiguous words of a language remains a daunting one.

STRUCTURAL DISAMBIGUATION

Structural ambiguity is ambiguity of the structure of the utterance itself, as seen in the 'Admiral' example above, in which the prepositional phrase *with his hands behind his back* could be a modifier of *taking*, describing the manner in which the turns around the room were taken, or of *room*, describing the room. The ambiguity in this example is often referred to as one of 'prepositional phrase attachment', as the problem is determining which node in the parse tree of the sentence the prepositional phrase should be attached to. There are many kinds of structural ambiguity (the attachment point of relative clauses is another important one) but prepositional phrase attachment in English has received the most study and will be the example used here. (*See* **Sentence Processing**; **Sentence Processing: Mechanisms**)

Because the ambiguity is reflected in the parse tree of the sentence, its resolution is part of the process of syntactic analysis of the sentence, or parsing. A common way to conceive of the problem is that the parser determines, from the rules of syntax of the language, what the possible attachment points are, and then asks some other process to determine which is most likely to be correct in context (Hirst, 1987). (However, in the case of lexically conditioned statistical parsers, such as that of Collins (1996), no distinction is made between attachment decisions that are mandated by the grammar of the language and those, the kind that are of interest here, that are 'discretionary'; in both cases, the decisions are based on the probabilities of syntactic dependencies between the particular words.) Structural ambiguity and lexical ambiguity are independent phenomena, but clearly resolution of either one interacts with resolution of the other: the best attachment point might depend on the meaning of a word, and the most likely meaning of a word might depend on a structural decision. In practice, however, the two ambiguities are usually considered separately. (*See* **Parsing**)

Prepositional phrase attachment ambiguity, in its simplest form, has three or four elements: a verb (e.g. *await*), the head noun of its object (e.g. *approval*), the preposition (e.g. *from*), and, in some methods, the head noun of the prepositional phrase (e.g. *government*). The disambiguation process must choose between the verb and the object head noun as the attachment point. Recent approaches have

tried, in various ways, to use the relative frequency of each attachment, as determined by statistics gathered from a large corpus of sentences. Because manual annotation of the corpus is not required, more data are available for this than for the analogous problem in lexical disambiguation. However, the problem is harder because even in very large corpora, most combinations of three elements occur rarely if at all; using four elements instead of three increases the potential accuracy of the method at the expense of exacerbating the sparseness of the data.

Taking the three-element problem, Hindle and Rooth (1993) achieved about 80 percent accuracy with an unsupervised training method based on the attachment probabilities that were observed in a 13-million-word corpus of newswire text. The corpus had been almost fully parsed but lacked resolution of its ambiguous prepositional phrase attachment points. The method computed 'lexical association (LA) scores', defined as the logarithm (base 2) of the relative likelihood of verb and noun attachment for triples: for example, LA(*send*, *soldier*, *into*) was found to be approximately 5.81, meaning that verb attachment is 56 (i.e. $2^{5.81}$) times more likely than noun attachment in the sentence *Moscow sent more soldiers into Afghanistan*. These scores were determined by first gathering data from cases of unambiguous prepositional phrase attachment in the corpus (such as attachments to subjects of sentences and attachments to verbs without objects) and then, where strong lexical associations were found, using these data to resolve ambiguous cases; the procedure iterated until as many scores as possible were computed. Ratnaparkhi (1998) subsequently obtained similar results from an unsupervised method that required only part-of-speech tagging of the corpus, not parsing, by improving the heuristics by which the unambiguous training cases could be identified in the corpus.

Taking the four-element form of the problem, Brill and Resnik (1994) also obtained about 80 percent accuracy with a set of disambiguation rules that were derived from a corpus by means of the same supervised transformation-based learning method that Brill had earlier used for part-of-speech tagging. The rules state conditions under which a particular attachment is more likely: for example, 'the attachment point is the verb if the preposition is *in* and the noun of the prepositional phrase is a measure, quantity, or amount' or 'the attachment point is the object noun if the verb is a form of *to be*'. The semantic categorization of words for the rules (for example, characterizing a word as a measure, quantity, or amount for the rule mentioned above) is based on the WordNet electronic thesaurus (Fellbaum, 1998). Disambiguation initially assumes that the attachment is to the object noun. The rules are then applied, in a sequence of increasing specificity, to possibly change that; the provisional choice of attachment point might alternate several times as the rules are applied.

PRONOUN RESOLUTION

The ambiguity of pronouns (and anaphora in general) is different from the word sense ambiguity treated above in that there is no fixed inventory of candidate senses. Rather, the pronoun has an antecedent in the text with which it corefers; to disambiguate the pronoun is to find its antecedent, and the candidates are those elements of the preceding text that are 'available' for pronominal reference (in a sense that we will make more precise below). (*See* **Anaphora**; **Anaphora, Processing of**)

The antecedent of a pronoun is distinguished from its referent in that the antecedent is an element of text and the referent is an object in the world. Consider the following text from Charles Dickens's *Our Mutual Friend*:

> 'Let me', says the large man, trying to attract the attention of his wife in the distance, 'have the pleasure of presenting Mrs Podsnap to her host.'

The antecedent of *his* is the noun phrase *the large man*, and the referent of both is Mr Podsnap. The antecedent of a pronoun may be another pronoun; thus a text of the form *Mr Podsnap … He … He … He* (all about Mr Podsnap) creates a 'chain' of coreference, with the second and third pronouns each having the previous pronoun as its antecedent. Although one might have instead said that *Mr Podsnap* is independently the antecedent of all three pronouns, viewing antecedence as a chain appeals to our intuition that the antecedent of a pronoun must be recent within the text. (*See* **Story Understanding**)

The resolution of a pronoun is thus a two-stage process: determining the candidate antecedents, and then, if there is more than one, choosing among them. We will consider each stage in turn.

Candidates for Antecedence

While textual recency is a criterion for antecedence, it is neither necessary nor sufficient; indeed, there need be no single explicit textual antecedent. What matters most is that the referent be in the 'focus of attention' at the point at which the pronoun is uttered. That the antecedent need not be recent

was shown by Grosz (1977) in her studies of people engaged in task-oriented dialogues, such as an instructor guiding an apprentice. Grosz found that when a partially completed task was resumed after a long intervening subtask, the speakers would often refer by pronouns to antecedents in the earlier discourse about the task; what mattered was that the particular task was again in the speakers' attention. That recency is not sufficient for an element to be available as an antecedent can be seen in this text:

> John put the wine on the table. It was brown and round.

Readers generally find this text to be somewhat odd, with the antecedent of *it* seeming to be *the wine*, even though *the table* is more recent and a table is more likely than wine to be brown and round. (In the terminology of centering theory, to be introduced below, this text is an example of a 'rough shift'.) That there need not be a single explicit antecedent can be seen in this text from Charles Dickens's *Our Mutual Friend*:

> Mrs Lammle bestowed a sweet and loving smile upon her friend, which Miss Podsnap returned as she best could. They sat at lunch in Mrs Lammle's own boudoir.

The antecedent of *they* is *Mrs Lammle* and *Miss Podsnap* together – a set that the reader must construct from separate elements of the text.

Choosing from Multiple Candidates

When there is more than one candidate antecedent, the choice among them is based on linguistic constraints and preferences and on common-sense knowledge of the world.

In most languages, pronouns are marked for gender, number, or both, and candidates that do not match these features are therefore immediately ruled out by these constraints; in English, a reference to a person can be eliminated as a possible antecedent for the pronoun *it*. Syntax also puts various restrictions on antecedence. For example, in English syntactic structures, if a nonreflexive pronoun functions as a complete noun phrase, its antecedent cannot be any node of the parse tree that is immediately dominated by another node that also dominates the pronoun. It is this rule that precludes *Nadia* being the antecedent of *her* in *Nadia baked her a cake*.

Syntactic structure can also determine a preference for one candidate over another. For example, a candidate antecedent that plays the same syntactic role as the pronoun is preferred over one that doesn't, especially if the sentences in which they occur exhibit 'syntactic parallelism'. For example, in *Nadia waved at Emily and then she shouted at her*, the preferred interpretation is that *she* is *Nadia* (both are subjects of their verb) and *her* is *Emily* (both are objects of their verb). Notice that if the pronouns are stressed heavily, the pattern is reversed; stress on a pronoun generally indicates that its antecedent is not the one that would normally be preferred (Kameyama, 1999).

One particularly influential theory of antecedence preference is centering theory (Walker *et al*., 1998). Centering theory relates the form chosen for a referring expression – such as the speaker's choice of whether or not to use a pronoun – to the focus of attention within the discourse, the syntactic structure of the text, and the difficulty of interpretation of the utterance. In the theory, each sentence within a discourse is said to have a 'center', which is, roughly, its topic or its most salient element; and each sentence of the discourse makes elements available, including its center, that could become the center of the subsequent sentence. These potential centers are ranked by their syntactic position – subject is ranked highest, then object, then other positions – and if any of these potential centers are indeed mentioned in the subsequent sentence, then the one that ranks highest in the first sentence is the actual center of the second sentence, regardless of its position in that sentence. Now, the center must always be pronominalized if any other element of the sentence is; thus, if a sentence contains just one pronoun and its antecedent cannot be found in the same sentence, that pronoun must be the center, and so its antecedent is therefore the highest-ranking potential center from the previous sentence. The following example is simplified from Thomas Bulfinch's *The Age of Fable*:

> Orpheus was presented by Apollo with a lyre and taught to play upon it. He did so to such perfection that nothing could withstand the charm of the music.

He is the only pronoun in the second sentence, so it must be the center, and its antecedent is *Orpheus*, which as subject of the first sentence outranks *Apollo*. This rule also explains the problem of the 'brown and round' example earlier; *it* is the only pronoun in the second sentence, and in the first sentence, *the wine* outranks *the table* as a potential center. Transitions like this, to a new center that is neither the center of the previous sentence nor its highest-ranking potential center, are very rare in naturally occurring text (Di Eugenio, 1998).

In observations such as these, centering theory thus provides a set of preferences that can be

employed in pronoun resolution and an explanation of the difficulty that people experience when the expectations that these preferences entail are not fulfilled (Hudson-D'Zmura and Tanenhaus, 1998).

By applying constraints and preferences such those just described, a natural language system can often determine a unique antecedent for a pronoun. Systems differ in the exact rules that they apply, the order in which they apply them, and how they trade off conflicting constraints and preferences. The system developed by Lappin and Leass (1994), for example, achieved an overall success rate of 89 percent on within-sentence antecedence and 74 percent on cross-sentence antecedence; since within-sentence antecedence is more common, the overall success rate was 86 percent. But an error rate of 14 percent is still too high for most practical uses.

It is not surprising that a system using only syntactic constraints and preferences will make mistakes relatively often, as knowledge of what 'makes sense' is often required to choose the correct antecedent of a pronoun. Lappin and Leass give this example (from a computer manual):

> This green indicator is lit when the controller is on. It shows that the DC power supply voltages are at the correct level.

Their system incorrectly chooses *controller* over *green indicator* for *it*; the two alternatives are rated equally in all respects (each is the subject of its verb) except for recency, which favors *controller*. Clearly, *indicator* is, in general, a 'better' subject for the verb *show* than *controller* is; this suggests the use of selectional restrictions, as used for lexical ambiguity, as an additional constraint. An approximation to this, frequency of co-occurrence, is proposed by Dagan and Itai (1990): statistics gathered from a large corpus would be used to give preference to the candidate antecedent that occurs more frequently as the subject of the verb *show*. Incorporating this and other heuristics into a single process, Mitkov (1998) has achieved anaphor resolution accuracy approaching 90 percent.

References

Brill E and Resnik P (1994) A rule-based approach to prepositional phrase attachment disambiguation. In: *Proceedings, 15th International Conference on Computational Linguistics, Kyoto*, pp. 1198–1204.

Collins MJ (1996) A new statistical parser based on bigram lexical probabilities. In: *Proceedings, 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, California*, pp. 184–191.

Dagan I and Itai A (1990) Automatic processing of corpora for the resolution of anaphora references. In: *Proceedings, 13th International Conference on Computational Linguistics, Helsinki*, vol. III, pp. 330–332.

DeRose SJ (1988) Grammatical category disambiguation by statistical optimization. *Computational Linguistics* **14**: 31–39.

Di Eugenio B (1998) Centering in Italian. In: Walker *et al*. (1998), pp. 115–137.

Fellbaum C (ed) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Grosz BJ (1977) *The Representation and Use of Focus in Dialogue Understanding*. PhD thesis, University of California, Berkeley, CA.

Hindle D and Rooth M (1993) Structural ambiguity and lexical relations. *Computational Linguistics* **19**: 103–120.

Hirst G (1987) *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge, UK: Cambridge University Press.

Hudson-D'Zmura S and Tanenhaus MK (1998) Assigning antecedents to ambiguous pronouns: the role of the center of attention as the default assignment. In: Walker *et al*. (1998), pp. 199–226.

Johnson C and Fillmore CJ (2000) The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In: *Proceedings, 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle*, pp. 56–62.

Kameyama M (1999) Stressed and unstressed pronouns: complementary preferences. In: Bosch P and van der Sandt R (eds) *Focus: Linguistic, Cognitive, and Computational Perspectives*, pp. 306–321. Cambridge, UK: Cambridge University Press.

Kilgarriff A (1992) Dictionary word sense distinctions: an enquiry into their nature. *Computers and the Humanities* **26**: 365–387.

Kilgarriff A (1997) I don't believe in word senses. *Computers and the Humanities* **31**: 91–113.

Kilgarriff A and Palmer M (eds) (2000) *Computers and the Humanities*, **34**: 1–243. [Special issue on SENSEVAL.]

Lappin S and Leass HJ (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics* **20**: 535–561.

Lesk ME (1986) Automatic sense disambiguation using machine-readable dictionaries: how to tell a pine cone from an ice cream cone. In: *Proceedings, 5th International Conference on Systems Documentation, Toronto*, pp. 24–26. New York, NY: Association for Computing Machinery.

Mitkov R (1998) Robust pronoun resolution with limited knowledge. In: *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal*, pp. 869–875.

Ratnaparkhi A (1998) Statistical models for unsupervised prepositional phrase attachment. In: *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal*, pp. 1079–1085.

Resnik P (1998) WordNet and class-based probabilities. In: Fellbaum (1998), pp. 239–263.
Resnik P and Yarowsky D (1999) Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering* **5**: 113–133.
Walker M, Joshi AK and Prince EF (eds) (1998) *Centering Theory in Discourse*. Oxford: Clarendon Press.
Yarowsky D (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: *Proceedings, International Conference on Computational Linguistics, Nantes, France*, pp. 454–460.
Yarowsky D (1995) Unsupervised word sense disambiguation rivaling supervised methods. In: *Proceedings, 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA*, pp. 189–196.

Further Reading

Grosz BJ and Sidner CL (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics* **12**: 175–204.
Hirst G (1981) *Anaphora in Natural Language Understanding: A Survey*. Berlin: Springer.
Ide N and Véronis J (eds) (1998) *Computational Linguistics* **24**: 1–165. [Special issue on word sense disambiguation.]
Jurafsky D and Martin JM (2000) *Speech and Language Processing*. Upper Saddle River, NJ: Prentice-Hall.
Mitkov R (2002) *Anaphora Resolution*. London: Longman.
Palmer M and Light M (eds) (1999) *Natural Language Engineering* **5**(2): i–iv and 113–218. [Special issue on semantic tagging.]
Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* **11**: 95–130.
Schütze H (1997) *Ambiguity Resolution in Language Learning*. Stanford, CA: CSLI.
Webber BL (1978) *A Formal Approach to Discourse Anaphora*. New York, NY: Garland.

Natural Language Processing, Statistical Approaches to

Intermediate article

Christopher D Manning, Stanford University, Stanford, California, USA

CONTENTS

Statistical approaches to natural language processing use large text corpora to allow rapid, robust, and accurate handling of the ambiguities in human languages.

INTRODUCTION

In a statistical approach to natural language processing (NLP), the operations of the system are based in some way on counts that have been made over a text corpus. Such a collection of texts is usually large, so that accurate estimates can be made from it. It may just contain normal sentences, or it may be text that has already been annotated with linguistic information, such as information about parts of speech, syntactic structure, or coreference.

What is counted depends on the particular model being used, and can range from simple word adjacency to sophisticated grammatical relationships – for example, how often a prepositional phrase modifies a noun phrase when the noun phrase contains a superlative adjective. Almost any theoretical model of natural language processing can be enriched with such counts, which provide the additional information of how often the various structures licensed by the theory occur.

These counts are normally used to set the parameters of a specific probabilistic model of language. Such a model can be of two types. Most research has used generative models, which place a probability mass function over the structural descriptions of a language that are allowed by the model. Other research, focusing on particular decisions (such as deciding the part of speech of a word) has used discriminative models, which simply learn to distinguish between the different parts of speech of a word. (*See* **Natural Language Processing**)

REASONS FOR ADOPTING A STATISTICAL APPROACH

Statistical approaches have transformed the field of NLP. Rather than hand-building categorical grammar-based models, much work now involves the induction of linguistic knowledge based on evidence from corpora. Even strongly grammar-based work is increasingly making use of corpus frequencies to encode preferences.

The widespread adoption of statistical methods in the 1990s coincided with broader trends in artificial intelligence and cognitive science towards the use of probability distributions to represent knowledge, and conditional distributions to infer knowledge. Computing power and online textual resources had grown sufficiently for statistical approaches to be feasible and successful. (*See* **Reasoning under Uncertainty**)

Such methods are very suitable for NLP because sentences of human languages are highly ambiguous: most words have various possible meanings and parts of speech, and often sentence structures are ambiguous as to how one part of a sentence relates to another. Thus the process of language understanding (and, indeed, also language generation and acquisition) frequently involves integrating and reasoning from incomplete and uncertain information, using not just the actual words uttered but also the prior discourse context and world knowledge. Probabilistic methods are a powerful approach to incorporating such diverse knowledge sources, and have been much more successful in practice than the heuristics, ad hoc algorithms, and grammatical approaches that were widely used previously. Moreover, rather than simply giving a structure to sentences, they provide a probability mass over sentence structures, representing how likely various structures are. This information has

proven vital in applications ranging from speech recognition to sense and sentence structure disambiguation. For instance, a speech recognition system will choose between several word sequence hypotheses (that sound similar to the acoustic signal being processed) based on which word sequence is more likely to occur in the language being recognized. (*See* **Natural Language Processing, Disambiguation in**)

HIDDEN MARKOV MODELS FOR ASSIGNING SYNTACTIC CATEGORIES

For many purposes, such as terminology extraction, partial parsing, lexical acquisition, and linguistically informed text searching, a useful level of linguistic analysis is to assign parts of speech (that is, syntactic categories like 'noun', 'verb', etc.) to each word in a text. This operation is generally referred to as 'tagging'. Tagging is a straightforward classification problem for which there is abundant training data available, and almost every machine learning method has been applied to the tagging task (Van Halteren, 1999). Hidden Markov models provide a natural generative probabilistic model for this problem (Bahl and Mercer, 1976), and have become the standard approach, though in general their performance is not significantly better than that of various other machine learning methods.

Consider the task of assigning parts of speech to the (*New York Times*) headline *Fed raises interest rates 0.5% in effort to control inflation*. Many of these words can have multiple parts of speech: *raises*, *interest*, and *rates* can be nouns or verbs, *to* can be a preposition or the infinitive marker, *in* is usually a preposition but has rare adjectival (*the in crowd*) and nominal (*he had some sort of in with the boss*) uses.

The two obvious sources of information to exploit in disambiguating such tokens are facts about the *a priori* distribution of the words (*raises* is used as a verb 9 times out of 10), and information about tag sequences ('determiner–noun' is a common sequence, while 'determiner–verb' is not). While the second of these sources of information may seem most obvious to a linguist, it turns out to be of limited use (for example, in the sentence above, *raises*, *interest*, and *rates* can all be either nouns or verbs, and either part of speech can follow another noun). The first source of information – which has no value as a categorical constraint, but only as a statistical constraint – turns out to have greater value (Charniak *et al*., 1993).

A hidden Markov model combines these two sources of information, as illustrated in Figure 1. When tagging, we wish to find the most likely tag sequence $t_1, \ldots, t_n$ for a sequence of words $w_1, \ldots, w_n$, which we assume to have been generated from an underlying (hidden) Markov process, where the states of the Markov model are taken to represent part-of-speech tags.

Using Bayes' rule, and the Markov assumption, we obtain:

$$\begin{aligned}
&\arg\max_{t_1,\ldots,t_n} P(t_1, \ldots, t_n | w_1, \ldots, w_n) \\
&\quad = \arg\max_{t_1,\ldots,t_n} P(w_1, \ldots, w_n | t_1, \ldots, t_n) P(t_1, \ldots, t_n) \\
&\quad = \arg\max_{t_1,\ldots,t_n} \prod_{i=1}^{n} [P(w_i|t_i) P(t_i|t_{i-1})]
\end{aligned}$$

The parameters of such a model can be easily estimated from a corpus of text that has been tagged by hand, although careful probability smoothing and estimation is necessary to deal with rare uses and previously unseen words. The most likely tag sequence for a new sentence can then be determined, on the basis of the above equations, by an efficient dynamic programming algorithm known as the *Viterbi algorithm*. When trained and tested on homogeneous corpora, modern taggers commonly get over 95% of part-of-speech tags right. However, performance can be considerably worse when the training and testing text differ in style or subject matter. (*See* **Statistical Methods**)

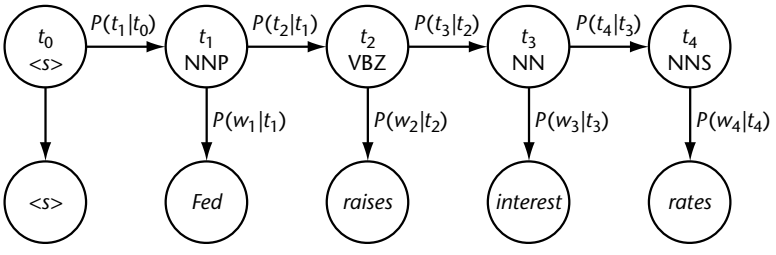


Figure 1. A hidden Markov model applied to tagging the sentence *Fed raises interest rates*. Arrows show direct probabilistic dependencies between part-of-speech states and words.

STATISTICAL PARSING

Parsing is the process of determining the correct structure for a sentence within the terms of a grammatical theory. Commonly, this involves a phrase structure tree representation. Because sentences' structures are highly ambiguous, a grammar that covers a fair proportion of general English (such as that found in newspapers) will typically produce dozens or even hundreds of possible parses for moderately long sentences. For example, in a sentence like *The board approved the acquisition [of the stores] [in Toronto] [at its monthly meeting]*, each of the three prepositional phrases at the end of the sentence is syntactically allowed to modify the verb or any preceding noun, so there are many possible structures. Whereas traditional parsers find all such structures, and then (sometimes) attempt to disambiguate on the basis of preferences or categorical selectional restrictions, statistical parsers attempt to chose the most likely parse for a sentence, by making use of statistics on how likely different words and phrases are to combine together in a sentence.

In the 1990s, new statistical parsers capable of successful, robust, rapid, broad-coverage parsing were developed (Charniak, 1997; Collins, 1997). The improvements were of three kinds. Firstly, since probabilistic language models use measures of likelihood instead of hard grammar rules, such parsers will find *some* analysis for any sentence. They will not fail to parse a sentence because of limitations of the grammar, or because it falls outside some narrow conception of grammaticality. Secondly, these parsers can work quickly. Traditional parsers exhaustively find all possible analyses for a string, but statistical parsers keep track of the most likely partial parses and expand just those. This generally involves orders of magnitude less work. Indeed, by using various methods to prune the search space, several modern probabilistic parsers parse in time linear in the sentence length (categorical parsers are typically cubic or worse).

The third and most important improvement was in accuracy. Statistical parsers use facts about word relationships to resolve ambiguities as to which parse structure to assign. The central statistics used in most current parsers are 'bilexical head dependencies', which are commonly collected from 'treebanks' of hand-parsed sentences. For example, if we are considering how to parse *Fed raises interest rates*, then on the hypothesis that *raises* is a verb, we would consider how frequently the subject relationship [*Fed, raises*] and the object head relationship [*raises, rates*] occur, whereas on the hypothesis that *interest* is the main verb, we would ask how likely it is that *raises* is the subject of *interest* and *rates* is the object.

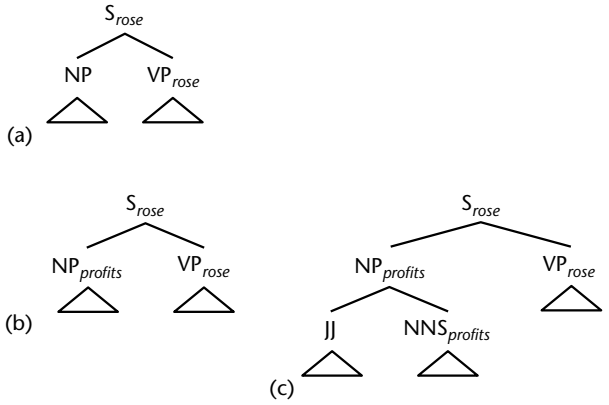


Figure 2. In the lexicalized probabilistic context-free grammar model of Charniak (1997), the generative model proceeds from a state such as that shown in diagram (a). It then determines the head $h =$ 'profits' of a non-head child given its category $c =$ NP and the head $h_P =$ 'rose' and category $c_P =$ S of the parent, using the probability function $P(h|h_P, c, c_P)$, to give diagram (b). Then it determines a rule r to expand this child using the probability function $P(r|h, c, c_P)$, to give diagram (c). This strategy is then iterated recursively.

Evaluation of these dependencies is generally achieved by lexicalizing a context-free grammar, so that the head word of each phrase is annotated on the parent node. Bilexical head relationships are then parent–child relationships. An example is [*profits, rose*] in Figure 2 (which follows the approach of Charniak (1997)). A generative probabilistic model predicts how likely nodes are to be headed and expanded in different ways. Although the probabilistic model uses top-down probabilistic conditioning, the actual computation is done using a bottom-up chart parser, which makes extensive use of search space pruning. Because of the extreme sparsity of bilexical head pairs, a lot of work goes into robust statistical estimation of the probability distributions shown in the figure, commonly via various forms of clustering and back-off. On newswire sentences (with an average length around 25 words), the best versions of such statistical parsers get about 90% of the postulated nodes in the parse tree right; and they get the overall constituency structure of the whole sentence right about 70% of the time (Charniak, 2000).

STATISTICAL EXTRACTION OF LEXICAL SEMANTICS FROM CORPORA

While knowing parts of speech and syntactic structures is helpful to human language interpretation,

most people apart from linguists are mainly concerned to understand not the structure but the meaning of human language sentences. In order to approach this goal, there has recently been a lot of interesting work on learning knowledge, in particular lexical semantic knowledge about word meanings, from large corpora. The idea is that a computer program, rather than having to have facts about word meanings encoded into it, can learn them by 'reading' a large corpus. For example, it can learn that a 'durian' is a kind of fruit, that 'image' and 'picture' mean roughly the same thing, or that people frequently 'repeat' a 'comment' but not typically an 'article'. All statistical NLP models need a large amount of information about individual words, many of which are quite rare, and whose meaning and manner of use varies between different genres or time periods.

Many approaches to this task have been explored. Some methods are similar to information extraction, where particular textual patterns that mark certain semantic relationships are detected. For example, Hearst (1992) used such methods to learn the hyponym ('is a kind of') relationships of a semantic hierarchy by matching frames such as '[*X*], a kind of [*Y*] …' in text corpora.

Other researchers have followed statistical approaches. Context vectors, which are counts of the words around instances of particular words, can be used with some success as a surrogate representation for the meaning of a word (Schütze, 1997). Richer syntactic relationships can be used to represent finer-grained semantic relationships. For example, clustering nouns based on just their modifiers can be an effective way to group nouns that share the same essential properties (Grefenstette, 1996); and clustering between pairs of verbs and the head noun of their object can pick out semantic classes of verbs, such as verbs of scalar motion or verbs of disposition (Rooth *et al*., 1999).

While all of these methods have had some success at learning partial semantic representations from large quantities of unannotated text, their success has so far been limited, both in terms of accuracy and in terms of the richness of the semantic information acquired. Finding effective methods for acquiring semantic information from large unannotated text corpora remains one of the central challenges for future NLP work. (*See* **Statistical Methods**)

References

Bahl LR and Mercer RL (1976) Part-of-speech assignment by a statistical decision algorithm. In: *IEEE International Symposium on Information Theory, Ronneby, Sweden*, pp. 88–89.

Charniak E (1997) Statistical parsing with a context-free grammar and word statistics. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 598–603. Menlo Park, CA: AAAI.

Charniak E (2000) A maximum-entropy-inspired parser. In: *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 132–139. San Francisco, CA: Morgan Kaufmann.

Charniak E, Hendrickson C, Jacobson N and Perkowitz M (1993) Equations for part-of-speech tagging. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 784–789. Menlo Park, CA: AAAI.

Collins MJ (1997) Three generative, lexicalized models for statistical parsing. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 16–23. San Francisco, CA: Morgan Kaufmann.

Grefenstette G (1996) Evaluation techniques for automatic semantic extraction: comparing syntactic and window-based approaches. In: Boguraev B and Pustejovsky J (eds) *Corpus Processing for Lexical Acquisition*, pp. 205–216. Cambridge, MA: MIT Press.

Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the Fourteenth International Conference on Computational Linguistics (Coling'92)*, pp. 539–545. San Francisco, CA: Morgan Kaufmann.

Rooth M, Riezler S, Prescher D, Carroll G and Beil F (1999) Inducing a semantically annotated lexicon via EM-based clustering. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111. San Francisco, CA: Morgan Kaufmann.

Schütze H (1997) *Ambiguity Resolution in Language Learning*. Stanford, CA: CSLI Publications.

Van Halteren H (ed.) (1999) *Syntactic Wordclass Tagging*. Dordrecht, Netherlands: Kluwer.

Further Reading

Abney S (1996a) Statistical methods and linguistics. In: Klavans JL and Resnik P (eds) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pp. 1–26. Cambridge, MA: MIT Press.

Abney S (1996b) Part-of-speech tagging and partial parsing. In: Young S and Bloothooft G (eds) *Corpus-Based Methods in Language and Speech Processing*, pp. 118–136. Dordrecht, Netherlands: Kluwer.

Armstrong S, Church KW, Isabelle P *et al*. (1999) *Natural Language Processing Using Very Large Corpora*. Dordrecht, Netherlands: Kluwer.

Charniak E (1993) *Statistical Language Learning*. Cambridge, MA: MIT Press.

Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Ng HT and Zelle J (1997) Corpus-based approaches to semantic interpretation in NLP. *AI Magazine* **18**: 45–64.

Natural Language Processing

Introductory article

James F Allen, University of Rochester, Rochester, New York, USA

CONTENTS

Natural language processing is a field that explores computational methods for interpreting and processing natural language, in either textual or spoken form.

INTRODUCTION

Computers that can speak and understand natural language have long been a key component in science fiction. Actually building machines that can understand and produce speech, however, has proven to be exceptionally difficult. While research started in the first days of functional computers in the 1950s, the complexity of the problem thwarted researchers for decades. Recently, however, significant progress has been made and, while machines are far from understanding language like humans do, useful applications involving language are now possible. We see dictation systems that allow someone to dictate a letter to a computer, internet search engines that look for pages with certain content, automated telephone systems that allow you to dial numbers and make collect calls just by speaking, and systems that can answer simple questions about the weather, traffic or stock quotes over the telephone. In addition, more sophisticated language processing systems are currently under development and will become usable in the next few years.

It is important to realize that there are several different motivations for work in natural language processing. On one side, there are engineering goals, where one is interested in finding techniques that allow computers to perform practical tasks. On the other are the cognitive science goals, where one is interested in the insight that computational models can provide for understanding human language processing. While there is overlap in research towards these goals, this article will focus mainly on issues of relevance to the cognitive science goals.

INTRODUCTION TO RELEVANT LINGUISTIC NOTIONS

The principal difficulty in processing natural language is the pervasive ambiguity that is present. Ambiguity is found at all levels of the problem. For example, all natural languages involve:

- acoustic ambiguity (given a speech signal, what words were actually spoken?);
- lexical ambiguity (e.g. 'duck' can be a noun (the animal or the cloth) or a verb (to avoid something thrown));
- structural or syntactic ambiguity (e.g. in 'I saw the man with a telescope', the telescope might be used for the viewing or might be held by the man being observed);
- semantic ambiguity (e.g. 'go' as a verb has well over ten distinct meanings in any dictionary);
- pragmatic ambiguity (e.g. 'Can you lift that rock?' may be a yes/no question or a request to lift the rock); and
- referential ambiguity (e.g. in 'Jack met Sam at the station. He was feeling ill…' it is not clear who is ill, although the remainder of the sentence might suggest a preferred interpretation).

All these forms of ambiguity may interact, producing an extremely complex interpretation process. It is the prevalence of ambiguity that distinguishes natural languages from precisely defined artificial languages such as logic and programming languages. It also makes most of the techniques developed in programming language grammars, parsing and semantics, ineffective unless significantly modified.

It will help to break apart language processing into different phases as shown in Figure 1. One side involves the understanding processes, in which we move from speech to intentions. The other side involves generation, moving from intentions to speech. The stages of understanding language consist of transforming sound into words (speech recognition), words into meaning (understanding), and meaning into knowledge, intention, and action

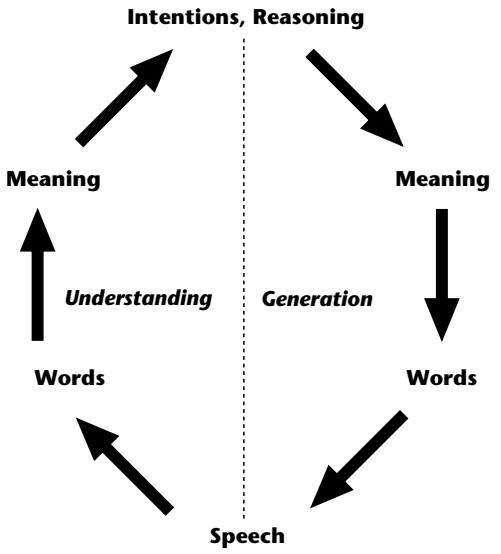


Figure 1. The stages of natural language processing.

(pragmatics and reasoning). For instance, say someone comes up and speaks to you. Using speech recognition you identify the words as *Do you know the time?* Using understanding techniques, you identify the meaning as a question about whether you know the time. Finally, using pragmatics and reasoning you decide that the stranger wants you to tell him the time and decide to answer the question. The stages of language generation account for the opposite process of generating a response. You decide to respond and form an intention to tell him the time, this is mapped into a meaning (content planning), then meaning to words, say, 'It's three p.m.' (generation), and finally words to sound (speech synthesis).

Each of these steps will be considered below. For many applications, we need only one or two stages. For instance, current dictation systems do only the first phase, mapping speech to words, and do not interpret the words any further.

SPEECH RECOGNITION: SOUNDS TO WORDS

While speech recognition systems are in use in practical applications now, they work well only in some specific situations. The important factors affecting the performance of recognition systems include:

1. the size of the vocabulary to be recognized,
2. the range of speakers that need to be handled,
3. the quality of the sound capture devices (e.g. microphones and recording), and
4. the style of the speech itself.

In general, the smaller the vocabulary, the smaller the number of speakers, the better the sound capture, and the more controlled the style of the speech, the better the recognition is. In practice, however, we usually have to balance one aspect against another to get acceptable performance in a specific application. For example, the systems that allow you to make long distance calls entirely by voice must handle hundreds of thousands of customers. They compensate for the problems caused by the large number of speakers by using a very small vocabulary (e.g., just digits and words such as 'yes' and 'no'), and by constraining the style of the speech by giving specific instructions to the person about what to say when. By doing this, the applications are usable and are saving telephone companies millions of dollars a year in operation costs. A very different case involves dictation systems. These must handle large vocabularies, say 40 000 to 100 000 words. To compensate for this, they are trained for a single speaker using a microphone held close to the mouth, and require people to speak carefully. Under these conditions, such systems can attain 95% word accuracy or better for many users.

As we start to relax the constraints more, accuracy starts to plummet. For example, currently the best speech recognition systems, when trying to recognize spontaneous speech in real time between two people having a conversation, can recognize only half of the words correctly.

Most speech recognition systems use the same basic set of techniques. The acoustic signal is digitized and converted into a sequence of sound frequency–intensity plots covering small segments of the speech. These plots are used to extract out a vector of measurements, such as intensity in different frequency bands, changes of intensity in different frequency bands from one window to the next, and the acceleration of changes of intensity in the different frequency bands. The vectors are then classified into categories based on their similarities to 'prototype' vectors. Typically, systems use 256 or 512 different categories, which comprise the *codebook*. At this stage, the signal has been converted into a sequence of symbols from the codebook.

The next phase of analysis uses models of the relevant linguistic events you want to recognize, such as words or phonemes. For large vocabulary recognition, word-based models are too hard to construct, and phoneme-based models are too inaccurate. As a result, most large vocabulary recognition systems use subword models, an intermediate level about the size of syllables, larger than the phoneme, but smaller than a word. Each

subword is represented by a Hidden Markov Model (HMM), from which one can compute the probability that a certain sequence of codebook symbols might have been generated in producing that subword. To perform speech recognition on an input signal represented as a sequence of codebook symbols, one simply finds the sequence of subword that maximizes the probability of producing that input signal.

The power of HMMs is that there are known effective algorithms for estimating the transition probabilities from a corpus of speech data, and for finding the path through a given HMM model that maximizes the probability of observing the input (the Viterbi algorithm).

So far, the description of the system has considered acoustic information only. Using only acoustic information, speech recognizers do not perform well. They must also use higher level linguistic information to predict what word sequences are likely to occur. This is called the language model. There are two types of language models in common use. The first is called an *n-gram model*, which is a probability model that specifies the probability of a given word w following a sequence of the previous $n-1$ words. A 2-gram (or bigram) model, for instance, simply provides the probability that the word w will follow the previous word. Such n-gram models are built by estimating the probability distribution from a large corpus of transcripts of speech. In an ideal world with unlimited data, one would use large amounts of previous context. In practice, however, speech recognition systems typically only use bigram or trigram models. Given an n-gram model, one can construct an HMM and use the training and search algorithms developed for HMMs to drive the recognition process. The other type of language model is the *probabilistic finite-state machine*. This is simply a probabilistic graph that specifies all possible sequences of words that can occur in interactions. This technique is common in practical applications where the vocabulary and range of language is not large.

APPROACHES TO LANGUAGE PROCESSING

As we consider the other levels of language processing, especially language understanding, we have a problem. With speech recognition, we know what the desired result is, namely what words were spoken. But we do not have such a clearly defined notion of what meaning and understanding are. This means it is hard to evaluate how well we are doing. The classic answer to this problem has been the Turing test. A system would be said to understand if a set of human judges could not distinguish a machine from a human by asking any questions they wished for as long as they wished. The problem with this test is that it is an all-or-nothing evaluation. It is hard to use this method to help drive incremental progress since we are so far from the goal. Today, most researchers in the field take a pragmatic approach and consider how well language systems can perform a specific task. For example, say the task is to find relevant webpages given natural language input such as the sentence 'I want to find sources that deal with economic growth in the late nineteenth century'. For such an application, we evaluate how well it does the task, say by checking how many relevant pages it finds (the recall) and what proportion of the pages found are relevant (the precision).

In the early decades of research (1960–1990), it was believed that machines would have to acquire a capability for 'deep' understanding of language before the technology could be useful. Deep understanding involves constructing a precise specification of the meaning of the sentence and being able to perform reasoning on the information. In the 1990s, however, researchers found that 'shallow' statistical analysis techniques that capture only some structure and content of language could be useful in a wide range of applications. For example, most webpage search engines just look for documents that contain the words or short phrases found in the query. While such techniques produce robust systems with reasonable recall, the precision is often a problem. You may find that the system returns so many irrelevant webpages that finding the appropriate ones can be very difficult. In addition to information retrieval, there are many practical applications, which are becoming feasible using statistical approaches, that fall far short of 'deep' understanding including information extraction (e.g., generating a database of merger and acquisitions activity from newspaper reports), information access (e.g., using speech over the telephone to find out information on airline schedules), rough machine translation (e.g., automatically translating equipment maintenance manuals), and writing tools (e.g., spelling correction and grammar checking tools in word processors).

For some period of time, many researchers viewed the deep and statistical processing techniques to be in opposition to each other. More recently, most researchers realize that each area of work addresses different aspects of the problem

and that both are needed for long-term success in the field.

COMPUTATIONAL PHONOLOGY AND MORPHOLOGY

Language is made of words, either realized as sound or in written form. Sometimes words are considered atomic elements of language, but they actually have a rich structure. Many languages use a wide variety of word forms to indicate how words relate to each other in the sentence. In English, we see word forms to encode number and person agreement between pronouns and verbs (*I am, you are, she is, they are, …*), grammatical role in pronouns (*he, his, him*), tense and aspect (*go, went, gone, going*), parts of speech (*destroy, destruction*, *destroyable, destructive, destructively*) and semantically related words (*untighten, retighten, pretighten, overtighten*). The study of the structure of words and how different forms relate to each other is called morphology, and the development of algorithms for processing word forms is computational morphology.

A very useful computational technique is called *lemmatizing*, which identifies the root form (or lemma) and *affixes* of a word for further processing. For example, such a system might break apart the words weave, wove, woven, and weaving into a root form *weave* plus features such as *present, past, pastparticiple*, and *presentparticiple*. Stemming is important as a first pass for producing deeper structural and semantic analyses of sentences as the lemmas typically identify key components of the semantic content. Another application of stemming is in information retrieval, where it would be useful for all the forms of *weave* to map to the same lemma. Thus if you look for documents that include the word *weave*, the program would also find documents that contain the word *wove*.

One of the most influential techniques for morphological analysis uses finite-state transducers (FSTs), which are finite-state machines that produce an output for every input. Such a device would take a word like *happier* as input and produce the output *happy* + *er* using an FST that maps the first four letters to themselves, then maps the *i* to a *y*, inserts a space, and maps the last two letters, *e* and *r* to themselves. One of the most common frameworks for morphological analysis uses a set of FSTs that are run in parallel; all of them must simultaneously accept the input and agree on the output.

Phonology studies the relationship between words and how they are realized in speech. The idealized sounds of language are captured by an alphabet of *phonemes*, each phoneme representing an abstraction of a primitive meaningful sound in the language. In actual fact, the same phoneme can be pronounced very differently in different contexts. These variants within a single phoneme are called *allophones*. Computational models of phonology are important in both speech synthesis and speech recognition work. It is not feasible, for instance, for a speech synthesizer that can read books to have a mapping of all words to their pronunciation. Rather, it often has to produce a phonetic spelling from the orthographic spelling of the word. In general, the same computational techniques used for morphological processing are also used for phonological processing.

PARSING

The process of parsing involves mapping from a sequence of words to an internal representation. A *syntactic* parser takes a grammar and a word sequence, and finds one or more combinations of rules in the grammar that could produce the word sequence. The output then is a parse tree that shows the different phrasal constituents that make up the sentence. A *semantic* parser maps a word sequence to a representation of its meaning in some formal meaning representation language. Most modern grammars and parsers combine these two tasks and construct a syntactic structure and a meaning representation simultaneously using *rule-by-rule compositional semantics*. In such an approach, the rules in the grammar specify not only structural relations (e.g., a noun phrase can consist of an article ('A') followed by a noun ('dog')), but also a fragment of the meaning representation, say, in some form of logic. As rules combine smaller phrases into larger ones, the meaning fragments of the smaller phrases are combined into larger fragments for the entire phrase. It is possible to separate these processing phases and use a grammar that produces a solely syntactic structure, and then define another set of rules for mapping the syntactic structures into meanings. But so far researchers have found little advantage in doing so. It is also possible to forgo syntactic processing altogether, and attempt to build a meaning representation directly from the word sequence. Such approaches are typically only used in quite limited applications in which the sentences remain fairly simple, and the meaning representation is specialized to the particular application and lacks generality.

The most common grammatical formalisms used in computational systems are augmented

context-free grammars. Such grammars have a core context-free grammar, but each rule is associated with a set of features that capture grammatical phenomena such as number agreement and encode semantic information. Most of these systems use *feature unification* and can use quite elaborate typed feature structure systems. For example, the rule

```
NP[AGR ?x] -> ART[AGR ?x] N[AGR ?x]
```

states that a noun phrase (NP) can consist of an article (ART) followed by a noun (N), but only the ART and N have compatible agreement feature (AGR) values. Furthermore, the NP acquires the same agreement feature. For example, the word *the* allows both third person singular and plural forms, so its AGR feature would be a set consisting of 3s (third person singular) and 3p (third person plural). The word *dog*, on the other hand, allows only third person singular, so the AGR feature would be the singleton set {3s}. Feature unification is like set intersection, so these two features can be unified and the result is {3s}. This becomes the AGR value of the newly introduced NP constituent. Likewise, agreement would be enforced between the subject and the verb by the S rule (a rule for sentences) that combines an NP followed by a verb phrase (VP). Note that with the AGR feature, the grammar will not accept sentences such as *A dogs barked* or *The dogs barks* even though the word sequences are fine based simply on the basic word categories. Figure 2 shows a parse tree that is the analysis of the sentence *The dog barked*, showing the basic categories and the AGR feature.

Context-free grammars augmented with feature unification provide the power to express much of natural language in fairly concise ways, and have become the most common grammatical formalism for computational grammars. They have also provided the formal underpinnings for linguistic theories such as head-driven phrase structure grammar (HPSG) and lexical-functional grammar (LFG).

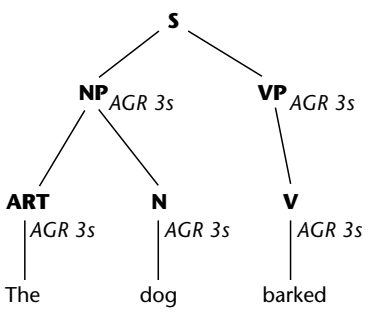


Figure 2. A parse tree.

A parser performs a search process through the possible combinations of grammatical rules to find trees that can account for the input. Most modern parsing systems use some form of chart parsing. A *chart* is a data structure that keeps track of the work done so far in parsing: both the possible constituents that have been found so far, and the rules that have partially matched so far. The parser then runs by constantly updating the chart with new information. There are two properties of chart parsers that make them efficient. The first is that the parser never does the same work twice. Before doing any work, it checks the chart to see whether the work has been done before. The second property is that information is shared across different possible interpretations. For example, a particular noun phrase might be able to be used in a number of different overlap interpretations. That noun phrase, however, appears only once in the chart.

SEMANTICS: LEXICAL AND COMPOSITIONAL

Semantics concerns the meaning of sentences. We need to define a language in which to express that meaning, and we need some method of computing the meaning from the parse tree. To start this process, we need to associate meanings with words. Words typically have many senses (e.g., 'seal' can denote an animal, a gasket for making containers airtight, a symbol representing some authority, or an action of closing an envelope). Such meanings can often be arranged hierarchically according to their properties. For example, a 'seal' is a particular type of animal, which is a living thing, which is a physical object. The properties of being a seal arise from particular properties of seals (e.g., seals bark), from more general properties (e.g., animals eat, physical objects have weight). Such representations are often called *semantic networks*. These properties are important for helping determine the appropriate sense while parsing. For instance, if we have a sentence 'The seal ate a fish', then we can conclude that we are talking about a seal as an animal, and not the seal of a canning jar, or the symbol of the president of the United States.

Other words tend to have more complex semantic structures. A verb, for instance, typically requires a set of arguments to complete its meaning. For instance, the verb 'put' describes some sort of physical action, which is determined once we know who did the action, what was acted upon, and where it went. In many computational representations a complex object is defined as a *type* and a set of *roles*. The roles for the PUT action might

be called AGENT, THEME, and TO-LOC. Such representations have been proposed in philosophy by Davidson, in linguistics by Fillmore, and are often called *frames* in computational work. Expressed in first-order predicate calculus, the meaning of the sentence *Jack put the book in the box* might be

```
∃e . PutEvent(e) & AGENT(e, J1)
  & THEME(e, B1) & TO-LOC (e, B2)
```

where the constants J1, B1, and B2 are defined elsewhere and represent Jack, the book, and the box. In a frame-like knowledge representation, this might be represented as

```
[PUTEVENT e
  [AGENT J1]
  [THEME B1]
  [TO-LOC B2]]
```

but the semantics remains the same.

Compositional Semantics

One of the key ideas used in many modern parsers is compositional semantics, where the semantic form of the sentence is constructed simultaneously with the syntactic form. Each rule in the grammar specifies both syntactic and semantic operations. To make this work, systems use either feature unification or lambda abstraction. For instance, the meaning of a VP 'ate the pizza' might be

```
λ x . [EAT [AGENT x] [THEME P1]]
```

which is a function that takes an argument and places it the AGENT slot. With this interpretation, the semantic component of a grammatical rule for declarative sentences, namely

```
S -> NP VP
```

would apply the meaning of the VP (like *ate the pizza* above) to the meaning of the NP to produce the meaning of the sentence. While the lambda calculus provides a nice formal model for compositional semantics, many parsers do the equivalent thing using feature unification. For example, we could write the same rule as

```
S[LF ?lf] ->
  NP[LF ?subj] VP[SUBJ ?subj LF ?lf]
```

Here, the VP rule would produce a constituent of the form

```
VP[SUBJ ?s LF [EAT [AGENT ?s] [THEME
  P1]]
```

and when the S rule is applied, the *?subj* variable will be bound to the meaning of the subject NP and the LF feature of the new S constituent would be bound to the completed LF in the VP.

Robust Interpretation

In many applications, it is not feasible to require a full syntactic analysis of each sentence. For example, in a spoken dialogue system, speech recognition errors may prevent the parser ever receiving all the correct words spoken. In a system that retrieves relevant articles from a newspaper database, we can't expect to have a grammar that has full grammatical coverage. In these cases, it is important to be able to extract out fragments of meaning from the input. Using a bottom-up chart parser provides one obvious approach. Even if a full syntactic analysis is not found, the chart will contain analyses of parts of the input – the major noun phrases, for instance. And given these analyses it may be possible to infer the intended content of the entire utterance. For example, consider a system that interacts with a user in spoken language to identify good routes for trains. Say the conversation has concerned a particular train TR1. The next utterance is *now send it on to Avon via Bath*, but say the speech recognition output is 'NOISE ENDED TO AVON VIA BATH'. From the fragments, one can recognize a path (to Avon via Bath), and from the discourse context, one knows the topic was train TR1. Thus, it is likely that the speaker intended to send TR1 from Avon to Bath.

For applications such as information retrieval and information extraction from textual databases, most systems forgo full parsing altogether and depend entirely on pattern-based techniques that look for patterns in the input that relate to information of concern.

PRAGMATICS: SPEECH ACTS AND DISCOURSE

Applications that involve interaction in natural language require significant processing beyond deriving the semantic interpretation of an utterance. In everyday language, a conversational system needs to recognize what the intentions of the user are in order to respond appropriately. For instance, a sentence such as *Do you know the time* might be a question about your state of knowledge, a request that you tell the speaker the time, or a reminder that it is late. Each of these different acts is called a *speech act*. Typically, the linguistic structure of the utterance does not identify which speech act has been performed. We need to consider the larger context of the conversation, namely the current situation and

what we know of the user's plans and goals, in order to identify the correct speech act.

The context of a conversation, or discourse in general, is captured by several different components. Two of the most important are the attentional state and the intentional state. The attentional state includes the recent history of the conversation, especially information about what was said in the last utterance. The objects mentioned in the last utterance are the most likely candidates for pronominal reference in the current sentence. In addition, the structure of the last sentence is needed to interpret forms of ellipsis (e.g., *I went to the store. John did too*.) One of the more influential models for pronoun interpretation is *centering theory*. This model places a high reliance on the objects mentioned in the immediately previous sentence, and distinguishes one object as most relevant.

The intentional state captures the motivations of the speakers engaged in the conversation. In general, computational work has focused on practical, or task-oriented, dialogue in which the participants are speaking to each other in order to perform a certain task, be it obtaining information, learning mathematics, getting help on home repair, or designing kitchens. In all these domains, we can represent the reasoning processes of the participants as a form of *planning*. The study of planning has a long history in artificial intelligence and is typically formalized as a search process through possible future actions. Actions are represented as operators that change the world, and are typically described in terms of their preconditions (what must be true when the action is attempted) and effects (what will be true after it is successfully executed). The process of planning starts with a library of actions, an initial world state, a statement of the goal conditions, and searches for a sequence of actions that can be executed starting in the initial state and resulting in a state that satisfies the goal conditions. By modeling the participants' planning processes, we can interpret their utterances by finding out how their utterances further their goals. These techniques can be used to address many problems, including word sense disambiguation (which reading makes sense in the plan), pronoun reference (which objects are most likely to be used in the plan), and most importantly, what speech act is intended. To do this last task we must model speech acts as operators in a planning system. The effects of a speech act are changes in the speaker's and hearer's mental state, namely their beliefs and goals. For example, an effect of a REQUEST act is that the hearer knows the speaker wants to get the hearer to do the requested act. By considering each possible speech act and how their effects might fit into the current task, we can determine which interpretation makes the most sense in context (e.g., a process called *plan inference*). A simple plan-based conversational agent uses plan inference to identify plausible goals of the user, and then uses planning to plan a response that is most helpful given the current task.

GENERATION: MEANING TO WORDS

The generation side of natural language processing has the same distinctions and levels of processing as the interpretation side. It is often viewed as a planning problem, where the input is a set of communicative goals and the output is a series of utterances that realize those goals. The process is typically broken into at least two levels: content planning, in which the system decides *what* needs to be communicated, and surface generation, in which the system determines the details of *how* it is to be communicated.

Content planning can roughly be thought of as mapping from abstract communicative goals (e.g., describe how to find the car in the parking lot) down to a sequence of specific speech acts (e.g., REQUEST that the person first go to the library entrance, then …). Rather than planning from first principles, generation is usually driven by larger scale *schemas* that relate conversational goals and specific circumstances to particular strategies for attaining those goals. By repeatedly refining conversational goals to more specific levels, one eventually ends up at the concrete speech act level. In addition to determining the speech, one must also plan the content of the act. A particular issue is determining what content to mention in order to produce a successful referring expression. In general, one is looking for a small set of properties that will successfully distinguish the intended object from other competitors in the context. For instance, in a setting in which there is a large red ball (B1), a large blue ball (B2), and a small red ball (B3), one might refer to B2 as 'the blue ball', not mentioning size, and B3 as 'the small ball', not mentioning color.

Surface generation can be viewed as the reverse process of parsing. It is given a logical form, the same or similar to the meanings produced by a parser. The challenge is to capture a given meaning in a natural sounding sentence of extended discourse. Some grammatical frameworks are designed in a way that they can be used both for parsing and for generation. Often, however, even when this is the case, different grammars are used.

One reason is that that there is a different set of concerns facing generation. For example, consider the passive construct (e.g., *John baked a cake* vs. *A cake was baked by John)*. A grammar for parsing might simply have rules that map both these sentences to the same logical form. A generator, on the other hand, must make a decision about which to use, and picking the wrong choice can lead to clumsy interactions or utterances that produce incorrect implications (e.g., consider *A cake was baked by John* as the answer to the question *What did John cook?*). In addition, the surface generator may have to make choices between different lexical realizations (e.g., *John donated $5 million to the museum* vs. *John gave the museum a $5 million contribution*).

CONCLUSION

Natural language processing (NLP) technology is becoming an area of great practical importance and commercial application. Within the foreseeable future, NLP will revolutionize the way we use and think about computers in a profound way. On the theoretical side, as the computational models become more sophisticated, we can expect new insights into possible models of human processing that can then become the subject of empirical experimentation.

Further Reading

Allen JF (1995) *Natural Language Understanding*. Redwood City, CA: Benjamin Cummings.

Allen JF, Miller B, Ringger E and Sikorski T (1996) *A Robust System for Natural Spoken Dialog*. Paper presented at the 31st Meeting of the Association for Computational Linguistics, Santa Cruz, CA.

Carpenter B (1992) *The Logic of Typed Feature Structures*. Cambridge, UK: Cambridge University Press.

Davidson D (1967) The Logical Form of Action Sentences. In: Rescher N (ed.), *The Logic of Decision and Action*. Pittsburgh, PA: University of Pittsburgh Press.

Grosz BJ and Sidner CL (1986) Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* **12**(3): 175–204.

Hobbs JR, Appelt D, Bear J *et al*. (1996) FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In: Roche E and Schabes Y (eds) *Finite-state Devices for Natural Language Processing*, pp. 383–406. Cambridge, MA: MIT Press.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Manning C and Schutze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Pollard C (1994) *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press.

Rabiner L and Juang BH (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Reiter E and Dale R (2000) *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.

Sproat R (1993) *Morphology and Computation*. Cambridge, MA: MIT Press.

Navigation

Intermediate article

Ulrich Nehmzow, University of Essex, Colchester, UK

CONTENTS

Navigation is the ability of a mobile agent (living being or machine) to move in a goal-directed way.

INTRODUCTION

Definition of Navigation

The ability of any mobile agent to move from its current position towards a specific location is arguably one of the most important competences it may possess. Only through the ability to navigate can the advantages of mobility be fully exploited. Without goal-directed motion the mobile agent has to resort to random motion, usually an inefficient strategy.

The word 'navigation' derives from the Latin *navis agere* (to steer a ship), and encompasses all skills required to stay operational (e.g. obstacle avoidance), to localize, to build internal representations ('maps'), to interpret these representations, and to plan paths. For the purpose of this article, 'staying operational' is considered an enabling sensor–motor competence, without which no navigation would be possible, rather than a navigational skill, and therefore it is not discussed in detail. The four fundamental competences of navigation are therefore:

- map building
- map interpretation
- self-localization
- path planning.

These four fundamental components of navigation are achieved by techniques which are discussed below. This article gives a general introduction to techniques commonly used by navigating agents, such as dead reckoning, piloting, and use of senses, and gives examples of their application in living beings (especially insects) and technological systems (especially mobile robots).

Terminology

We use the term 'agent' to mean any acting system, be it biological or technological. The term 'map' is used to mean a representation of the agent's physical space. Such a representation could be a conventional map, a topological representation of neighborhood relationships between discernible landmarks, or indeed any item for which there is a one-to-one mapping from that item to the agent's environment.

NAVIGATION TECHNIQUES

Dead Reckoning

The term 'dead reckoning' comes from 'deduced reckoning', a navigation method commonly used by sailors to determine the ship's position, and often entered in the logbook as 'ded. reckoning'. Because of its simplicity, it is the most widely used navigation technique in animals, humans and robots.

Dead reckoning allows the navigator to determine the current position within the map, provided the following information is available:

- map
- known starting location
- current heading
- either current speed or distance traveled.

Figure 1 illustrates the mechanism of dead reckoning.

Starting from a known position (x_0, y_0), the navigator measures the distance traveled (L) and the current heading, the angle α. If L cannot be measured, it can be determined if the speed v and travel time t are measured, using the equation:

$$L = vt \tag{1}$$

In sailing, it was common to determine the ship's speed through the water by throwing a floater attached to a knotted line overboard. The number of knots paid out per unit time determined the ship's speed, hence the term 'knots' for nautical speeds. This method can only measure speed across the water, and the effects of currents on the true

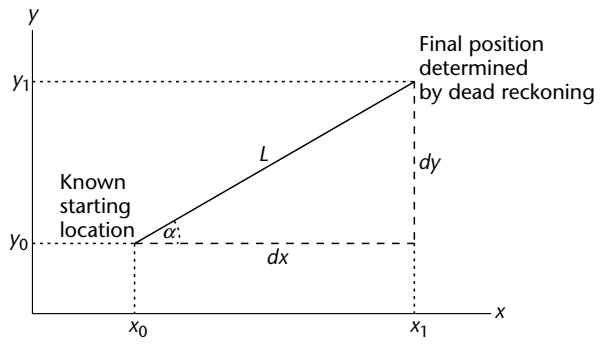


Figure 1. Dead reckoning. The map coordinates are represented by the x and y axes.

speed over ground have to be estimated by the navigator, introducing a dead reckoning error.

Taking the measurements L and α, the navigator can then calculate his current position by eqns 2 and 3 below, with dx and dy given by eqns 4 and 5 respectively. Dead reckoning is therefore also often referred to as 'path integration'.

$$x_1 = x_0 + dx \tag{2}$$

$$y_1 = y_0 + dy \tag{3}$$

$$dx = L\cos\alpha \tag{4}$$

$$dy = L\sin\alpha \tag{5}$$

Problems of dead reckoning

While it is not difficult to obtain a known starting location – every navigator starts at some prominent location, whose coordinates are known – it is difficult to obtain L and α with sufficient accuracy. It is especially important to obtain α as accurately as possible, because every heading error, however small, soon leads to large errors in the position estimate, because of the multiplication with the distance traveled (eqns 4 and 5). Figure 2 shows a robot's position estimate obtained by dead reckoning, completing four identical circuits in its environment.

A typical odometry error for both technological and biological systems is 5 to 10 per cent of the distance traveled (Gallistel, 1990).

Methods of estimating the distance traveled include measuring the speed of a float in water (mentioned above), measuring time while assuming a constant speed, or using optic flow (the apparent movement of images across the retina, which is proportional to speed).

The heading α is most commonly measured using a compass, but can also be determined using celestial bodies, or distant landmarks such as mountains.

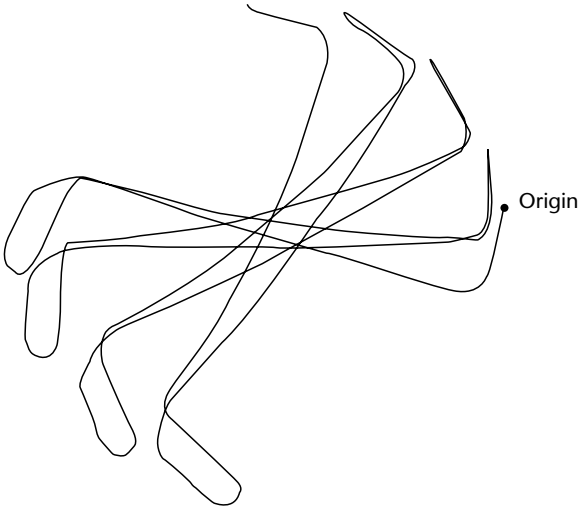


Figure 2. Accumulated dead reckoning error in a mobile robot.

Piloting

'Piloting' refers to navigation by observable features of the landscape – 'landmarks' – rather than by (blind) path integration. It allows self-localization (and therefore map building) and path planning, as shown in Figure 3.

Examples of technological systems that use radio beacons as landmarks are the Loran, Decca and Omega navigation systems, which determine the distance from shore transmitters by measuring the time differences of synchronized atomic clock signals being transmitted from each beacon. Two time signals determine two hyperbolae, whose intersection indicates the receiver's position. Satellite-based global positioning systems like the American GPS, the Russian GLONASS or the European Galileo project exploit the same principle, but give three-dimensional position information, using fixes with (at least) four satellites.

If the distances to the landmarks can be measured, then one distance and the angle is sufficient to determine the current position. In either case, a map is needed.

Problems of piloting

To use landmarks for navigation requires the unambiguous identification of the landmarks. This can be a problem. In natural environments, for instance, landscape features such as trees or mountains can be mistaken. The problem of 'perceptual aliasing' (i.e. the confusion of one landmark for another) in unmodified environments is particularly pronounced in technological systems such as mobile robots, because of the low sensory acuity

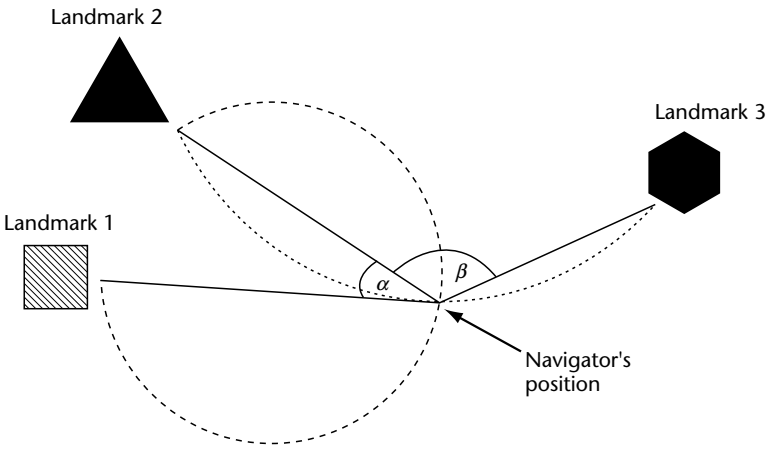


Figure 3. Piloting. The navigator's position is identified by the angles α and β between two clearly identified landmarks. Each angle determines a circle on which the navigator's position lies. The intersection of the two circles determines the navigator's position uniquely.

of robot sensors (Nehmzow, 2000). Navigation beacons (e.g. radio transmitters or lighthouses) therefore transmit identifying signals.

Spatial Representation in Animals and the Cognitive Map

Single-cell recordings in rats (and recently also in monkeys) show that in the CA3 and CA1 fields of the rat's hippocampus there are cells that fire when the rat is in a specific, small area of its environment. These cells are therefore referred to as 'place cells'. In the dorsal presubiculum, another area of the hippocampus, cells have been found whose behavioral correlate is head direction. These cells are therefore called 'head direction cells'.

Such cells are building blocks of navigation behavior, and a number of models have been suggested regarding the hippocampus's role in navigation (Burgess *et al*., 1995).

In 1948, Edward Tolman observed in rats the ability to determine novel shortcuts in a maze – an ability that was later also demonstrated in humans – and introduced the term 'cognitive map' for the rat's spatial representation. According to Tolman's definition, a 'cognitive map' is a representation of the space an agent inhabits, internal to the agent, that allows the agent to perform spatial reasoning, and in particular to determine novel shortcuts. O'Keefe and Nadel (1978) later argued that this cognitive map is located in the hippocampus.

Gould (1986) presented experiments in which honey-bees showed the ability to short-cut between two foraging sites, and proposed that honey-bees possessed cognitive maps. Considerable controversy arose over this claim, and several groups either were unable to replicate Gould's results, or could not rule out the explanation that landmarks seen across the short cut were used by the bees. Many groups argued that there are simpler explanations for the bees' ability to determine novel shortcuts. A summary of these arguments is presented in (Bennett, 1996).

Bennett argues that there are three simpler hypotheses that need to be eliminated first, before the presence of cognitive maps in bees can be considered proven: (1) that the short cut is not truly novel; (2) that path integration has been used, and (3) that familiar landmarks have been recognized from a new angle.

Gallistel (1990) defines 'cognitive map' more loosely as a spatial representation that facilitates spatial reasoning. That animals possess a cognitive map in this sense, i.e. an internal representation of space that is used for any type of navigation, is widely accepted. The debate continues, however, over whether they possess a cognitive map as defined by Tolman, i.e. a memory of landmarks that permits the determination of novel short cuts.

Other Techniques

Dead reckoning and piloting without maps

Both dead reckoning and piloting can be used without a map. Desert ants *Cataglyphis bicolor* (who do not lay pheromone trails), for instance, use path integration on both outward and homeward journeys from and to the nest (Wehner and Srinivasan,

1981), in featureless desert terrain. This is clearly shown when the ant is displaced at the beginning of the return journey, and returns to a position correspondingly displaced from the location of the nest. It then starts a search behavior centered around that position.

Likewise, piloting is possible without using maps. Again, experiments with *Cataglyphis bicolor* show that ants learn the position of a food source located between two visible pillars that serve as landmarks. If just the distance between the pillars is doubled, the ants search near either pillar for food (they cannot use the landmarks any longer for navigation), but if the size of the pillars is doubled as well, they search again between the pillars, at the right location (Wehner and Räber, 1979).

Specially controlled motion

In technological systems (e.g. robot arms), 'guarded motion' can be used to direct the arm reliably to its destination. Guarded motion entails taking the arm to an easily locatable edge, and then trailing it along the edge until the destination is reached.

The equivalent of guarded motion in the animate world is route following, where the navigator uses a naturally occurring route such as a river, a path, a shoreline or a mountain range for direction. Route following is commonly used in animal navigation. Canvasback ducks, for instance, follow the Mississippi river during migration (Waterman, 1989).

Sensors and Sensory Perceptions Commonly Used for Navigation

Of the two fundamental sources of information used in dead reckoning – direction and distance – direction is arguably the more important, given that even very small errors in estimating the current heading will quickly lead to large errors in the position estimate (see eqns 4 and 5, and figure 1). This 'compass sense' could be provided by an ordinary magnetic compass, but there are many other methods of obtaining a reference direction.

Carrier pigeons and many other birds, for instance, possess two independent compass senses. They are born with a magnetic inclination compass that measures the angle between the magnetic field lines and the Earth's surface, so that they can navigate 'pole-wards' and 'equator-wards'. Within the first 12 weeks of their life, they learn the relationship between the sun's movement at their home location and their magnetic compass. The sun compass acquired in this manner is later their preferred source of directional information (Wiltschko and Wiltschko, 1998).

Many insects, for example ants and bees (Waterman, 1989), use the polarization pattern of the sky, the so-called 'electric vector' ('e-vector') for orientation. Using the e-vector is another way of navigating by the sun, as the polarization pattern of the sky is dependent on the sun's position.

Celestial bodies can also provide a compass sense. The apparent motion of the celestial bodies is centered around the visually motionless celestial pole (which is near Polaris in the northern hemisphere). By observing the night sky over an extended period of time, the celestial pole can thus be determined, without any knowledge of constellations. Planetarium experiments with migrating birds, in which arbitrary constellations and movement patterns of celestial bodies can be created, demonstrate that migrating birds can use the night sky to determine the correct heading towards their goal (Waterman, 1989).

Locally occurring gradients can be used by many animals for navigation. Salmon, for instance, find their way back to the area where they were born by following a chemical gradient, the scent of their home waters (Waterman, 1989). They are able to detect minute differences in concentration of certain chemicals, and have a very high success rate in homing.

Carrier pigeons use locally occurring gradients (perhaps introduced through the Earth's magnetic field, or through infrasound from mountain ranges or the sea) to determine the correct heading for their loft (Wiltschko and Wiltschko, 1998).

One special form of path integration uses inertial navigation, doubly integrating the acceleration of the moving agent over time, to determine the navigator's position in space. In technological systems, accelerometers are used for this purpose. These are freely suspended masses, which exert forces onto force meters (e.g., strain gauges) when subjected to acceleration. As an example of biological navigators, rats possess a sense of direction (head direction cells in the anterior thalamus) and angular velocity, facilitating path integration (O'Keefe and Nadel, 1978).

Human Navigation

The seafarers in Polynesia and the Caroline islands routinely cover large distances (several hundred kilometers) between islands, across open sea, without using compass, sextant, or any other navigational instruments traditionally used in the Western world (Kyselka, 1987; Gladwin, 1970).

They achieve this by combining a range of information sources, rather than relying solely on any one navigational tool.

Celestial bodies

Obviously, the sun can be used to give directional information, but the navigators also make extensive use of the celestial bodies of the night sky.

Near the equator, stars rise and set almost vertically. This means that a rising or setting star can be used to indicate a particular direction while it is near the horizon. Navigators in Polynesia and the Caroline islands associate the rising and setting positions of certain stars with certain islands, and can thus use the stars to indicate the correct headings for their destinations.

The declination of a star is that latitude at which the star will pass through the zenith. Although Polynesian navigators do not use longitude and latitude for navigation, they do use the fact that a star passing through the zenith indicates latitude. By associating zenith stars with islands, they can determine whether they have arrived at the right latitude yet.

Landmarks and seamarks

When near the coast, either at departure or arrival, visible landmarks are aligned to give directions. The choice of landmarks obviously depends on the destination, but also on the wind and current situation: sailors select their landmarks to compensate for drift.

Once at sea, the navigators are able to derive direction from the low-frequency swell (not the choppy waves superimposed on the swell), which usually comes from a small number of general directions, because it is caused by (repeatable) weather patterns in a region. Whether the boat is aligned at right angles with the swell can easily be determined by detecting a gyrating movement of the boat, rather than a sideways or pitching motion.

Sea life is an important source of information. Many birds (for example, the booby) fly straight back home at dusk, having foraged over the sea. Certain islands have populations of turtles, or conspicuous fish, and can thus be detected from afar.

Fundamental to all these cues is the navigator's strategy, which is conservative. Sailors in Polynesia and the Caroline islands aim to avoid risk, for instance by selecting courses that will take them to the 'broad' side of an island or an island chain, or by voluntarily making detours via an easily accessible island. They do not rely on a single information source, but combine all the information available (sometimes even the taste of the water!).

Robot Navigation

Navigation strategies

The navigation strategies used in mobile robotics are variations of the methods used by biological navigators: route following, dead reckoning and piloting.

The simplest navigation method, employed by automated guided vehicles, is to detect and follow routes marked in the environment, for instance through buried induction loops, markers on floors, walls or ceilings, or radio beacons. Such navigation systems are robust, but expensive to set up, and inflexible.

Navigation systems for *autonomous* mobile robots, i.e. robots that have no power, control or computing connection to the outside world and that operate in unmodified environments, use navigation techniques similar to those of animals.

Dead reckoning by itself can only be used over very short distances (typically less than 10 meters), because of the accumulation of incorrigible drift error (see Figure 2). Navigation through dead reckoning therefore requires frequent recalibration at a known location, which can limit the active range of a mobile robot.

Landmark-based navigation systems aim to overcome the drawbacks of dead reckoning by representing space and planning motion in relation to identifiable landmarks (piloting). Provided landmarks are identified uniquely, such navigation is reliable, does not suffer from accumulation of incorrigible error, and works over long distances. The main problem landmark-based robot navigation systems face, however, is that, due to low sensor resolution, perceptual aliasing occurs, which leads to confusion and possibly to navigation errors.

Sensors for robot navigation

Dead reckoning, i.e., path integration, is most commonly performed using wheel encoders, which indicate the rotating angle and rotating direction of a wheel. Wheel encoders are purely proprioceptive, meaning that they do not take into account any external information. This is the reason why error accumulated by integration cannot be corrected using dead reckoning alone.

To determine a robot's heading, either wheel encoders or compasses can be used. Of the latter, a number of designs exist, notably flux gate compasses and Hall effect compasses.

Gyroscopes, in particular laser gyroscopes, can be used to measure rotation, and micro-accelerometers can be used to measure translation. They are therefore also suitable for path integration.

'Landmarks' in robot navigation are typically not conspicuous places, but sensory patterns that are recognizable and unique for a particular physical location.

Some commonly used types of sensors are described briefly below.

Sonar sensors

Sonar sensors measure the distance to the nearest object within a cone of typically 25° in front of the sensor, by determining the time it takes for a sound pulse to emanate from a transmitter, be reflected by the object and return to a receiver.

Laser range finders

Laser range finders work on the same principle as sonar sensors, but use light as the transmitted medium. Instead of using time of flight to determine range, phase shift between transmitted and received signal can also be used.

Cameras

Cameras using charge-coupled devices obtain a gray level or color intensity distribution of the image plane, and can be used effectively for landmark identification. However, it is computationally expensive to convert an image consisting of hundreds of thousands of picture elements ('pixels') into a meaningful, concise and identifiable representation.

Infrared proximity sensors

Infrared proximity sensors detect the reflection of emitted infrared light off a nearby object, and can thus detect the presence of (light-colored) objects in the vicinity of the sensor. They do not function well over distances much over a meter, or for detecting dark objects which do not reflect much light.

Tactile sensors

Tactile sensors detect immediate contact with an object, and are therefore only useful for navigation tasks such as route following (e.g. wall following along a corridor).

Representation of space and path planning

Unless the only navigational task is that of returning home, navigation using dead reckoning needs to use a metric map. Such a map contains metric coordinates (for example, Cartesian or polar coordinates) of locations. Simple trigonometry can then be used to compute the path between any two locations.

Landmark-based systems do not necessarily need a metric map: a topological map will usually suffice. Such a map represents the topological (neighborhood) relationships between landmarks, and is sometimes augmented by distance and direction information regarding the landmarks. A pure topological map still allows the determination of paths between arbitrary locations, but does not by itself allow determination of the shortest path.

For many navigation tasks, no map is needed. For instance, to perform a search (e.g. for food), and to return home, path integration and de-integration are sufficient. Similarly, if the home location is marked by a landmark or, more generally, an attractor (e.g. smell), simple taxis (moving towards an attractor) is sufficient.

To plan paths, either metric maps are used directly, and the required heading and distance determined by trigonometry, or the space is first tessellated (partitioned) into cells. Common methods of partitioning space include the Voronoi tessellation and its dual, the Delaunay triangulation. Search techniques (for example, those used in artificial intelligence to search through game trees) can then be used to find a free path.

Such search techniques can also be applied to path planning using topological maps. Reaction–diffusion dynamics are an example of a simple search technique suitable for path planning: a 'token' is placed on the goal location, and propagated in all possible directions at constant speed. The cell neighboring the navigating agent that receives the token first is the one indicating the shortest route to the goal. The process is repeated until the goal is reached.

SUMMARY

The advantages of mobility cannot be fully exploited without the capability to move in a goal-directed way, i.e., the ability to navigate. Virtually all mobile agents, biological or technological, therefore have some degree of navigational capability.

Common navigation strategies include path integration (dead reckoning), taxis (moving towards attractors), using landmarks, and using identifiable routes. These methods are used both by living beings and in robot navigation.

One interesting difference between biological and technological navigation is that biological navigators combine different information sources to a far greater degree than is currently done in robotics. This increases their reliability and robustness.

The combination of sensor signals from different sources (sensor fusion), and the combination of

different navigational strategies to achieve higher degrees of robustness and reliability, are still subjects of research, and a challenge for future robot navigation systems.

References

Bennett A (1996) Do animals have cognitive maps? *Journal of Experimental Biology* **199**: 219–224.

Burgess N, Recce M and O'Keefe J (1995) Spatial models of the hippocampus. In: Arbib M (ed.) *The Handbook of Brain Theory and Neural Networks*, pp. 468–472. Cambridge, MA: MIT Press.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA and London: MIT Press.

Gladwin T (1970) *East Is a Big Bird*. Cambridge, MA: Harvard University Press.

Gould J (1986) The locale map of honeybees: do insects have cognitive maps? *Science* **232**: 861–863.

Kyselka W (1987) *An Ocean in Mind*. Honolulu, HI: University of Hawaii Press.

Nehmzow U (2000) *Mobile Robotics: A Practical Introduction*. Berlin, Heidelberg and New York, NY: Springer.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

Waterman TH (1989) *Animal Navigation*. New York, NY: Scientific American Library.

Wehner R and Räber F (1979) Visual spatial memory in desert ants *cataglyphis bicolor*. *Experientia* **35**: 1569–1571.

Wehner R and Srinivasan M (1981) Searching behaviour of desert ants, genus cataglyphis (formicidae, hymenoptera). *Journal of Comparative Physiology* **142**: 315–338.

Wiltschko W and Wiltschko R (1998) The navigation system of birds and its development. In: Balda RP, Pepperberg IM and Kamil A (eds) *Animal Cognition in Nature*, pp. 155–199 London and New York, NY: Academic Press.

Further Reading

Baker R (1978) *The Evolutionary Ecology of Animal Migration*. London: Hodder and Stoughton. [An introduction to animal navigation.]

Borenstein J, Everett H and Feng L (1996) *Navigating Mobile Robots*. Wellesley, MA: AK Peters. [Technological foundations (mainly sensing techniques) of robot navigation.]

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA and London: MIT Press. [A broad and informative introduction to navigation which discusses mechanisms and presents experimental evidence for navigation mechanisms in living beings.]

Gladwin T (1970) *East Is a Big Bird*. Cambridge, MA: Harvard University Press. [An account of the navigation of the seafaring nations of Polynesia and neighboring areas.]

Kortenkamp D, Bonasso R and Murphy R (eds) (1998) *Artificial Intelligence and Mobile Robots*. Cambridge MA: MIT Press. [Examples of navigating robot systems.]

Kyselka W (1987) *An Ocean in Mind*. Honolulu, HI: University of Hawaii Press. [An account of the navigation of the seafaring nations of Polynesia and neighboring areas.]

Nehmzow U (2000) *Mobile Robotics: A Practical Introduction*. Berlin, Heidelberg and New York, NY: Springer. [Examples of navigating robot systems.]

Waterman TH (1989) *Animal Navigation*. New York, NY: Scientific American Library. [Detailed information on animal navigation.]

Journal of Experimental Biology **199**(1) (1996). [A special issue on biological navigation, especially migration and housing. A good source of information on the mechanisms used by biological navigators.]

Neural Behavior: Mathematical Models

Advanced article

Bard Ermentrout, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Neuronal networks are systems of counted equations meant to approximate the connectivity and function of biological neurons.

INTRODUCTION

The range of models of cognitive and neural processes spans many levels of detail, from single-channel dynamics, through models involving different brain regions, to models of social interactions between individuals. The appropriate level of detail depends in part on the types of questions asked. However, there comes a point, when addressing higher-level processes such as cognitive function, where the connection between the models and the actual machinery that underlies neural activity (e.g., channels, dendrites, synapses) appears to be very tenuous. This dissociation is not necessary, nor, in light of the recent rapid progress in experimental neuroscience (especially at the single-cell and molecular levels) is it desirable. Mathematical methods allow us to create bridges that span these different levels.

In order to study behavior and cognition, modelers generally study abstract networks of very simple units. For example, the traditional architecture for 'connectionist' models (McClelland and Rumelhart, 1988) consists of several layers of units that are connected in a feedforward manner by weights and simple nonlinearities. That is, the activity of unit j in layer n is determined by the sum of the activities of the units in layer $n-1$ that are connected to it. By modifying the weights of these interactions through a 'learning rule', it is possible to train these networks to perform a wide range of 'cognitive-like' tasks. Furthermore, by partially disrupting connections between units, a variety of psychopathologies can also be simulated and understood at least at this somewhat crude level.

Individual neurons exist in many shapes and sizes, have thousands of individual ionic channels and synapses, and are connected in networks of millions of different cells. Neurophysiologists are able to determine the channel properties (such as density and kinetics) on only a small piece of a small fraction of the neurons that comprise any given circuit. The synaptic interactions between neurons, and how their past activity influences them, remains a vigorous field of research. The gap between the complex and detailed models of single neurons and the simplified units alluded to in the previous paragraph would appear to be insurmountable. However, under certain assumptions (and whether these are in fact reasonable remains a subject of lively debate), it is sometimes possible to reduce complex units to much simpler versions which resemble the simple units favored by modelers of cognitive and behavioral psychology.

FROM NEURONS TO NEURAL NETWORKS

The usual models for neural networks (e.g. Amari, 1977; Hopfield, 1984; McClelland and Rumelhart, 1988; Wilson and Cowan, 1973) have the form

$$\tau_i \frac{dV_i}{dt} = -V_i + \sum_j W_{ij} F_j(V_j) \tag{1}$$

or

$$\tau_i \frac{du_i}{dt} = -u_i + F_i\left(\sum_j W_{ij} u_j\right) \tag{2}$$

The parameters τ_i represent a time constant, W_{ij} is the strength of the connection to j from i, and F_i is a nonlinearity often related to the firing rate of a neuron. In eqn 1, V_i often represents the voltage

or potential of the unit, while in eqn 2, u_i represents the activity or firing rate of the unit. The philosophy underlying these models is that it is the connections that matter and not the details of the neurons. This is true to an extent for models that do not involve dynamics such as action potentials or oscillations. However, as soon as temporal behavior is introduced, the details of the connections (such as latency) become far more important in determining the behavior.

There are many different connectivities possible for these neural networks. Connectionist models often work with feedforward mechanisms, as illustrated in Figure 1(a). That is, within a 'layer' there are no connections between units, nor are there feedback connections from layers above. In these models, the weights are adjusted by some type of learning rule in order to produce desired output when presented with some set of inputs. To study pathologies, modelers disrupt a fraction of the connections and then analyze the resulting network. In contrast to the simple feedforward network is the fully recurrent network shown in Figure 1(b). In this network, any unit can connect to any other unit. Such models have been used to study associative learning, and many other neural phenomena. Later in this article, we will discuss an example of this type of network as a model for working memory. If the connections between units in a fully recurrent network are symmetric (that is, $W_{ij} = W_{ji}$) then all solutions to eqns 1 and 2 go to equilibrium values (Hopfield, 1984). Thus, given some initial configuration, the network will always converge to a pattern of voltage or activity. This behavior allows the network to associate an input with a final output. Pattern completion is an often-exploited property of these networks.

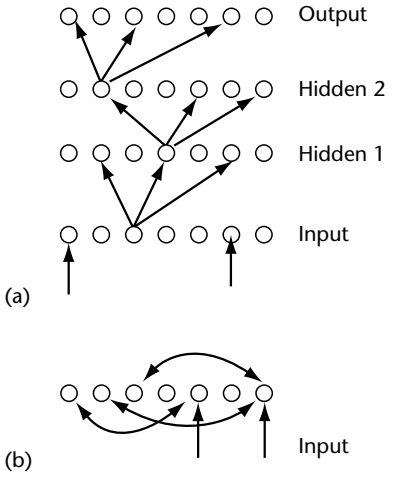


Figure 1. Common types of neural networks. (a) Feedforward network. Inputs come into the lower layer which makes forward connections to a sequence of hidden layers which terminates at the output layer. (b) Recurrent network. Any cell can connect to any other cell.

There are several ways to arrive at eqns 1 and 2 starting from physiological assumptions. The simplest way is to consider a pair of neurons coupled as in Figure 2. The somatic potential of the first neuron is $V_1(t)$. The potential difference between the soma and the axon hillock results in a current, and this current is sufficient to produce a train of action potentials along the axon. These occur at a rate $f(t)$ which is a function of the difference between the somatic potential and the resting potential of the axon hillock. Thus:

$$f(t) = F((V_1(t) - V_A)/R) \tag{3}$$

where $F(I)$ is the firing rate as a function of injected current (see below), V_A is the potential of the axon hillock, and R is the resistance between the two compartments. After a possible delay, the action potentials along the axon terminate at the synapse and allow neurotransmitter to be released. This results in a postsynaptic potential on the dendrite of the postsynaptic neuron. The postsynaptic potential from successive action potentials is summed and possibly filtered by dendritic properties. The result is that the somatic potential of the receiving neuron, $V_2(t)$, is changed to a new value. Assuming linear summation and a linear dendrite, the total potential contributed by the connection is

$$v_{1\to 2}(t) = \int Q(t-s)f(s)\mathrm{d}s \tag{4}$$

where $Q(t)$ is the response measured at the soma to an impulse delivered at the synapse. $Q(t)$ takes into account both the dynamics of the synaptic

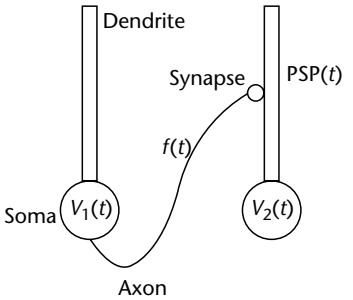


Figure 2. How neural network models are derived from biological neurons. Potentials $V_1(t)$ at the soma produce action potentials which travel down the axon at a rate $f(t)$, producing postsynaptic potentials PSP(t) in the dendrites. These potentials contribute to the somatic potentials $V_2(t)$.

transmission and the cable properties of the dendrite. It is often as simple as a decaying exponential, but more generally is a sum of decaying exponentials vanishing for $t \geq 0$ and vanishes for $t < 0$:

$$Q(t) = Q_0(\exp(-at) - \exp(-bt)) \tag{5}$$

Thus, if there were nothing else in the network but this connection, we would have

$$V_2(t) = V_{2,\mathrm{r}} + \int_{-\infty}^{t} Q(t-s) \times F((V_1(s) - V_\mathrm{A})/R)\mathrm{d}s \tag{6}$$

where $V_{2,\mathrm{r}}$ is the resting potential.

The network analogue of this model is

$$V_i(t) = \sum_j \int_{-\infty}^{t} Q_{ij}(t-s)F_j(V_j(s))\mathrm{d}s \tag{7}$$

where the resting potentials are all set to zero and the resistance is absorbed into the function F. Eqns 1 and 2 can both be derived from this general network under different assumptions. Suppose first that the form of the function Q_{ij} depends only on the postsynaptic neuron and is a single exponential. That is:

$$Q_{ij}(t) = W_{ij}\mathrm{e}^{-t/\tau_i} \tag{8}$$

Then eqn 7 can be rewritten as eqn 1. The interpretation of this is that since the response only depends on the postsynaptic neuron, either the synapses must be fast or the response of the neuron must be slow. Thus, the parameter τ_i is the membrane time constant of the neuron.

On the other hand, suppose that the form of Q_{ij} depends on the presynaptic neuron, that is:

$$Q_{ij}(t) = W_{ij}\mathrm{e}^{-t/\tau_i} \tag{9}$$

Then eqn 7 can be transformed into eqn 2. In this case, the interpretation is that the membrane time constant is short and the synaptic decay rate is slow. The parameter τ_i is the synaptic time constant.

An alternative derivation of eqn 2 starts with a biophysically detailed model with channels, such as the Hodgkin–Huxley model (Ermentrout, 1994). If the synaptic interactions are slow, then a mathematical method called averaging is used to produce equations of the form

$$\tau_\mathrm{i}\frac{\mathrm{d}s_i}{\mathrm{d}t} = -s_i + F_i\left(\sum_j W_{ij}s_j + I_i\right)(1 - s_i) \tag{10}$$

This differs in form from eqn 2 only by the factor $(1 - s_i)$. The interpretation of the variables in this model is quite different, s_i represents the fraction of open synapses occurring as a result of the firing of neuron i, F_i is proportional to the firing rate of the postsynaptic cell as a function of the current, I_i is the external input, and $W_{ij} = g_{ij}(V_{j,\mathrm{syn}} - V_{i,\mathrm{r}})$ is the maximum possible current due to the firing of the synapse. g_{ij} is the conductance of the synapse from j to i, $V_{j,\mathrm{syn}}$ is the reversal potential of the synapse, and $V_{i,\mathrm{r}}$ is the resting potential of the postsynaptic neuron. Thus, it is possible to connect detailed biophysical models with the abstract neural networks that are used in connectionism and other cognitive models.

Models such as eqn 10 can be extended to incorporate other processes such as synaptic depression or facilitation and spike frequency adaptation. For example, adaptation is incorporated by adding a term of the form $-K_i z_i$ to the inputs in eqn 10 and adding one more differential equation:

$$\tau_\mathrm{A}\frac{\mathrm{d}z_i}{\mathrm{d}t} = -z_i + (1 - z_i)G_i\left(\sum_j W_{ij}s_j - K_i z_i + I_i\right) \tag{11}$$

Firing Rates

When a neuron is injected with current, it often fires repetitively. The steady state firing rate is usually a monotonic function of the current. A number of firing rate functions are commonly used. The most common (although it does not directly follow from biophysical models) is the logistic function:

$$F_{\log}(I) = \frac{F_{\max}}{1 + \exp(-r(I - I_\mathrm{T}))} \tag{12}$$

where r is the gain and I_T is the threshold. In the limit as r gets large, this function approaches the step function. Another commonly-used model is the rectified linear model:

$$F_{\mathrm{lin}}(I) = r\max(0, I - I_\mathrm{T}) \tag{13}$$

For $I < I_\mathrm{T}$ the neuron does not fire but for $I > I_\mathrm{T}$ it fires at a rate that is linear. This model has some biological justification: there are a number of neuron models that have a nearly linear firing rate as a function of applied current. A firing rate function related to F_{lin} which fits a variety of cortical inhibitory neurons is the square-root model:

$$F_{\mathrm{sqr}}(I) = r\sqrt{\max(0, I - I_\mathrm{T})} \tag{14}$$

The Naka–Rushton model

$$F_{\mathrm{nr}}(I) = F_{\max}\frac{\max(0, I)^2}{{I_\mathrm{T}}^2 + I^2} \tag{15}$$

is saturating like the logistic model and is used in psychophysical models.

Finally, there is a very simple model for spiking neurons called the integrate-and-fire model. The integrate-and-fire model has enjoyed popularity for many years mainly due to its simplicity. Many variants have been developed but the basic structure of the model remains the same. The equations are

$$\tau \frac{dV}{dt} = -V + RI(t) \tag{16}$$

with the following condition. If $V(t) = V_T$ then the neuron fires and contributes some current to the cells to which it is connected. The voltage $V(t)$ is then immediately reset to $-V_R$. If the injected current I is held constant and is larger than V_T/R then the neuron will fire repetitively at a rate

$$F_{iaf}(I) = \frac{1}{\tau \ln\left[\frac{RI+V_R}{RI-V_T}\right]} \tag{17}$$

For low currents this function is similar to F_{sqr} but at high currents the rate is essentially linear. Several of the above firing rates are shown in Figure 3.

Modeling With Small Neural Networks

Using either the firing rate formulation or the voltage formulation, it is possible to model many neural systems. Suppose that there is no spatial structure to the connections between neurons. Then it is often desirable to reduce the system to as few as two populations of cells. A network of N groups of neurons all mutually inhibiting each other often exhibits what is called 'winner-take-all' behavior. That is, the network will select the group with the largest input and suppress all the others. Wilson (1999) uses one such network to explain the delay in picking a particular object out of a group of distractor objects. Since all distractor neurons are treated equally, their behavior can be combined into one unit. Let T be the activity of the group corresponding to the target and D be the activity corresponding to the distractors. Then the equations are:

$$\tau \frac{dT}{dt} = -T + f(E_T - ND) \tag{18}$$

$$\tau \frac{dD}{dt} = -D + f(E_D - (N-1)D - T) \tag{19}$$

The function f is one of the firing rate functions and N is the number of distractor neurons. The target network has a larger input, E_T. This network behaves as a winner-take-all network in that if both T and D start at zero, both will begin firing but eventually T will fire at a maximal rate and all the distractor neurons will be suppressed. Wilson shows that as the number N of distractor neurons increases, the amount of time for the target to 'win' increases in accordance with the psychophysical data.

By adding adaptation, fatigue, or some other slow negative feedback, it is possible to model perceptual oscillations such as those observed in binocular rivalry or the Necker cube. In these illusions, two figures compete equally for attention in a winner-take-all network. However, the presence of adaptation causes the 'winner' to weaken, allowing the 'loser' to gain strength. This cycle continues with regular periodicity. These simple models enable one to compute switching times and compare them to experimental results. Similar models have been suggested to explain competition at the neural level in recordings of visual tasks in awake monkeys.

Another simple network which has proved useful in the understanding of cortical functioning consists of a population of excitatory (E) and inhibitory (I) neurons:

$$\tau_E \frac{dE}{dt} = -E + f_E(w_{EE}E - w_{EI}I + T_E(t)) \tag{20}$$

$$\tau_I \frac{dI}{dt} = -I + f_E(w_{IE}E - w_{II}I + T_I(t)) \tag{21}$$

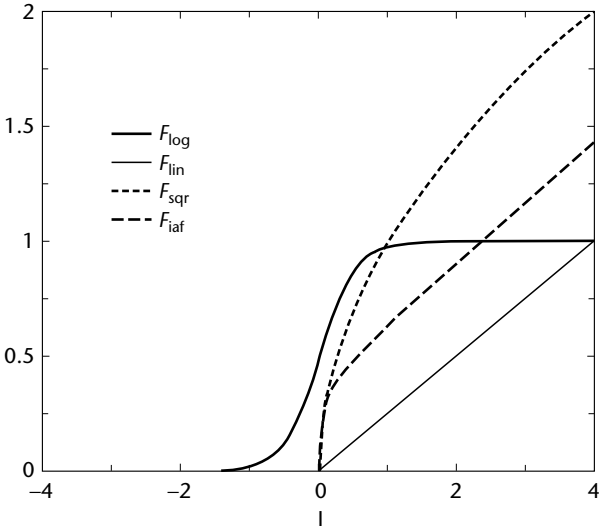


Figure 3. Different firing rate functions for the same applied current.

The functions $T_E(t)$ and $T_I(t)$ represent inputs from the thalamus. Since all excitatory and inibitory activity is combined into just two quantities, this simple system is a good illustration of the local dynamics of a single cortical column. Below, we will connect models of this form into spatial networks. One recent application of eqn 20 has been to the understanding of the response transformations of sensory inputs to the whisker barrel of the rat (Pinto *et al.*, 1996). Barrels are structures within the somatosensory area of the rat which receive inputs from a unique whisker. Eqn 20 was used in a network to predict responses of the barrel to temporally variable stimuli to the whisker corresponding to that barrel. The model predicted that the barrel acts like a differentiator, a prediction that was subsequently verified experimentally.

NETWORKS OF OSCILLATING NEURONS

The neural network models work under the assumption that it is the population rate that matters in cognition and sensory processing. An alternative point of view is that the timing of individual spikes is important. This idea is based on the fact that humans and other animals are able to make perceptual decisions over timescales that preclude the occurrence of more than one or two spikes per neuron in the relevant pathways. It is difficult to imagine how a rate code could work at that time scale. Furthermore, a code based on spike timing provides additional degrees of freedom that are not possible in a code based solely on the rate. Recent experimental work suggests that attention to a salient feature in a visual scene (or in other sensory scenarios) is accompanied by regular oscillations in a relatively narrow frequency band. Von der Malsburg and Schneider (1986) suggested that oscillations could be used to solve the so-called binding problem: how do different features (e.g. color, shape, position) of an object to be recognized get bound together as a coherent whole? They suggested that features could be grouped together if they shared a common phase; that is, if their neural correlates oscillated synchronously. Synchronous activity would result in the activation of specific target regions which shared these features. These theories, along with experimental evidence that supports them to some extent (Gray, 1994), have led to an interest in the types of computations that are possible with spiking and, in particular, oscillating neural networks.

Modeling large networks of spiking neurons presents a computational challenge, since details of the spike production and the synapses must be incorporated. The question of how detailed a model to use must also be addressed. There are several methods for simplifying the analysis and calculations associated with networks of spiking neurons. One idea is to use a very simple model for spiking and synapses. The model favored in most recent theoretical and computational work is the integrate-and-fire model. A network of these neurons has the form

$$\tau_i \frac{dV_i}{dt} = -V_i + R_i I_i(t) + R_i \sum_{j,\ell} \alpha_{ij}(t - t_j^\ell) \qquad (22)$$

with the condition that the voltage of each neuron is reset once it crosses a threshold. The numbers t_j^ℓ are the firing times of the jth neuron, that is, the times at which the neuron crosses its threshold. The functions $\alpha_{ij}(t)$ represent the postsynaptic potentials (similar in form to $Q(t)$ in eqn 5) resulting from the firing of the presynaptic neuron. Thus, suppose that at $t = t_j^\ell$ neuron j crosses threshold for the ℓth time. Then neuron j is immediately reset to a new voltage and each of the neurons to which neuron j projects receives a new contribution, $\alpha(t - t_j^\ell)$. The general behavior of these networks is difficult to analyze. However, it is often possible to find conditions for which waves or synchronous solutions exist, since the pattern of firing times is very regular. Bressloff and Coombes (2000) provide a comprehensive analysis of this class of models.

Another approach to spike timing is to assume a network of neurons each of which is coupled weakly to the other neurons (Hoppensteadt and Izhikevich, 1997). This means that the firing of other neurons does not destroy the oscillation of the target neuron; rather it serves to shift the timing of its next spike. Neurons in the network that are not firing at similar frequencies or not firing at all are effectively decoupled and do not participate in the timing pattern. The advantage of the weak coupling approach is twofold: the equations governing the timing can be derived from the fully detailed cellular models or from experiments; and the resulting equations are amenable to simulation and analysis. Each neuron (or group of neurons) in the circuit is represented by a single number lying on a circle, θ_i, the phase. Typically, zero phase represents the time at which a spike is emitted, but this is arbitrary as long as it is the same for each neuron. The equations have the form

$$\frac{d\theta_i}{dt} = \omega_i + \sum_j H_{ij}(\theta_j - \theta_i) \qquad (23)$$

where $H_{ij}(\phi)$ is a periodic function such as $\sin\phi$. The main quantities of interest are the timing or phase differences between neurons, $\Psi_{ij} = \theta_i - \theta_j$. If these are all zero, then the network is synchronous.

Kelso (1995) applied this class of model in experiments where he asks a subject to tap his fingers in an alternating rhythm and gradually increases the frequency at which the tapping must take place. He finds that at sufficiently high frequencies, the subject spontaneously switches from an alternating rhythm to a synchronous rhythm. In associated electroencephalographic recordings, he sees a qualitative switch in the pattern of the scalp potentials when the switch is made. He suggests a simple phase model. Let θ_L and θ_R be the phases of the left and right fingers. Then

$$\frac{d\theta_L}{dt} = \omega + H(\theta_R - \theta_L, f) \tag{24}$$

$$\frac{d\theta_R}{dt} = \omega + H(\theta_L - \theta_R, f) \tag{25}$$

where f is the frequency required for tapping. It is well documented in the theoretical literature that the shape of the functions H depends on the frequency of the underlying oscillators. (It can be expected that the time constants of the synapses connecting neural circuits are fixed. Changing the underlying frequency of the oscillations thus affects how these synaptic potentials alter the phase of the potentials. For example, it may be that stimuli that directly follow a neuron spike have a much smaller effect on shifting the time of the next spike than do stimuli that come later in the cycle. Since the persistence of a synapse is invariant, changing the frequency of the underlying oscillation drastically alters when the maximum stimulus comes in the cycle.) Since the quantity of interest is the phase difference between the left and right fingers, let $\phi = \theta_R - \theta_L$. Then

$$\frac{d\phi}{dt} = H(-\phi, f) - H(\phi, f) \equiv G(\phi, f) \tag{26}$$

Kelso suggests a particularly simple form for G:

$$G(\phi, f) = -\sin(\phi) - C(f)\sin(2\phi) \tag{27}$$

where $C(f) > 0$ and decreases with f. Clearly, $\phi = 0$ and $\phi = \pi$ are both equilibrium points of eqn 27. As long as $C(f) > 0$, the synchronous solution $\phi = 0$ is stable. For $C(f) > 1/2$ the alternating solution $\phi = \pi$ is also stable, but becomes unstable as f increases and thus $C(f)$ decreases. For $0 < C(f) < 1/2$ the only stable state is synchrony, so that if the subject starts in the alternating mode, then at sufficiently high frequencies, he or she must switch to the synchronous mode.

While the mathematics of the general systems represented by eqns 22 and 23 are reasonably well developed, the application of spiking and oscillatory networks to cognitive processes is only in its infancy. The main reason for this is that the level of modeling that is favored by cognitive theorists does not require detailed knowledge of spike times. As better neurophysiological data become available, spiking models may play a greater role in the modeling of cognitive data.

SPATIALLY DISTRIBUTED NETWORKS

The rapid development of imaging technology has made it possible to ask questions about the spatial distribution of activity at large and small scales. Thus, it is expected that neural network models with explicit spatial structure will play a role in our understanding of the biological mechanisms of cognition. Spatial models are no different from the general models represented by eqns 1 and 2 or eqns 22 and 23 except that the coupling weights of interactions depend on the spatial locations x_i of neurons in the network. In many networks, the coupling depends only on the distances $|x_i - x_j|$ between neurons. Much has been written on the behavior of spatially distributed networks (see Ermentrout (1998) for a review).

A kind of spatially localized self-sustaining activity known as 'bumps' has attracted much recent attention. Bumps are elicited in a spatial neural network by introducing a transient stimulus. Once the stimulus is removed, a local region of sustained firing appears and persists until some other process makes it disappear. The interest in this behavior derives from attempts to model delayed-response visual tasks. In these tasks, a monkey fixates on a target location. A brief stimulus is presented in the periphery and then removed. After a delay of several seconds, the monkey must make a saccade to the point where the peripheral stimulus appears. Thus, the task is a test of short-term or working memory. Cortical recordings during this task indicate that there are spatially localized groups of neurons which fire repetitively during the delay and stop firing once the task is completed. These are believed to be the neuronal correlates of working memory. Bumps have been used to model this neurophysiological behavior. There are many ways to produce these bumps of activity, ranging from simple neural networks such as eqn 2 to networks of spiking neurons with multiple channels and synapses. Since all the mechanisms

for sustained localized activity are based on the same general principles, we will consider the simplest implementation. Three conditions must be satisfied to obtain bump solutions in a neural network: (1) there must be a threshold, that is, small stimuli die out; (2) the spatial connectivity must be laterally inhibitory; (3) the network activity must be self-limiting. The first and third conditions preclude simple linear networks. The second condition says that local interactions are excitatory while the surround is inhibited. A network that is able to produce bumps is the spatial analog of the winner-take-all network with the additional condition that activity be sustained after the stimulus is removed. In a spatially distributed network, the localized bump of activity can be centred at any spatial location. Thus it uniquely specifies the position of the stimulus. Note that the position code in the cortical network could easily correspond to some other sensory modality. For example, auditory frequency is coded as a spatial map in the cortex.

The simplest model for bumps was proposed by Amari (1977). He goes to a continuous space limit and obtains the following equation for the potential:

$$\frac{\partial V(x,t)}{\partial t} = -V(x,t) + \int_{-\infty}^{\infty} J(x-y)F(V(y,t))\mathrm{d}y \tag{28}$$

For the firing rate function he chooses the step function: $F(V) = 0$ if $V < V_T$ and $F(V) = 1$ if $V \geq V_T$. The neurons in this network are either not firing or firing at a maximal rate. The connection function has the form shown in Figure 4(a). Nearby neurons are turned on and distant neurons are turned off. For simplicity, the network is one-dimensional. A bump consists of a localized region of activity and can be assumed to be centred at the origin. The bump is above threshold for $|x| < a$ and below threshold for $|x| > a$, as shown in Figure 4(c). Since the bump is independent of time and $F(V(x))$ is zero for $|x| > a$, the potential distribution satisfies

$$V(x) = \int_{-a}^{a} J(x-y)\mathrm{d}y = W(x+a) - W(x-a) \tag{29}$$

where

$$W(x) = \int_{0}^{x} J(y)\mathrm{d}y \tag{30}$$

The width of the bump, $2a$, is determined from the requirement that at the edges of the bump, the potential is at threshold, V_T:

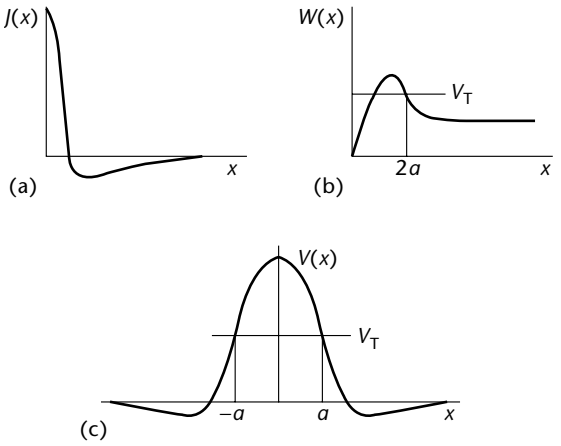


Figure 4. Construction of the 'bump' for a simple neural network. (a) The connection function $J(x)$ between two neurons a distance x apart. (b) The integral of $J(x)$. Intersections with the threshold V_T give the width of the bump, $2a$. (c) A bump of width $2a$ centered at the origin.

$$V(a) = V_T = W(2a) \tag{31}$$

Figure 4(b) shows that with a lateral inhibitory connection function, there are two possible bump widths for a range of thresholds. Amari shows that the bump is stable only if $J(2a) < 0$. Thus, although it is mathematically possible to get a bump without lateral inhibition, the bump will not be stable.

This calculation illustrates the general principles that underly the formation of localized bumps in neural networks. The same ideas have recently been applied to networks of spiking neurons. The main additional requirement for spiking neurons is that the synapses persist for a sufficiently long time. Similar ideas have been exploited to model head direction cells (Zhang, 1996), motor planning (Kopecz and Schoener, 1995), and the representation of the eye position in motor cortex (Salinas and Abbott, 1996).

SUMMARY AND CONCLUSIONS

By thinking about neural networks as simplifications of more complex biophysical systems, we have been able to connect neural networks with the underlying biophysics, simplify the necessary computations, and put ourselves in a position to analyze wide classes of networks which may be relevant to cognitive processes. The approach described in this article is meant to bridge the gap between the 'realistic' many-channel models and the simple 'connectionist' models that are used for modeling many cognitive processes. These

methods are useful because in many cases it is possible to produce models that provide a quantitative connection to the biology while maintaining the computational simplicity of abstract neural networks.

Acknowledgment

Author supported by NIMM and NSF.

References

Arbib M (1995) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

Bressloff PC and Coombes S (2000) A dynamical theory of spike train transitions in networks of integrate-and-fire oscillators. *SIAM Journal of Applied Mathematics* **60**: 820–841.

Ermentrout GB (1994) Reduction of conductance based models with slow synapses to neural nets. *Neural Computation* **6**: 679–695.

Ermentrout GB (1998) Neural networks as spatio-temporal pattern forming systems. *Reports of Progress in Physics* **61**: 353–430.

Gray CM (1994) Synchronous oscillations in neuronal systems: mechanisms and functions. *Journal of Computational Neuroscience* **1**: 11–38.

Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences* **81**: 3088–3092.

Hoppensteadt F and Izhikevich E (1997) *Weakly Connected Neural Nets*. Berlin: Springer.

Kelso JS (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kopecz K and Schoener G (1995) Saccadic motor planning by integrating visual information and pre-information on neural, dynamic fields. *Biological Cybernetics* **73**: 49–64.

McClelland JL and Rumelhart DE (1988) *Explorations in Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Pinto DJ, Brumberg JC, Simons DJ and Ermentrout GB (1996) A quantitative population model of whisker barrels: re-examining the Wilson–Cowan equations. *Journal of Computational Neuroscience* **3**: 247–264.

Salinas E and Abbott LF (1996) A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Academy of Sciences* **93**: 11956–11961.

von der Malsburg C and Schneider W (1986) A neural cocktail-party processor. *Biological Cybernetics* **54**: 29–40.

Wilson HR (1999) *Spikes, Decisions, and Actions*. Oxford: Oxford University Press.

Wilson HR and Cowan JD (1973) A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetic* **13**: 55–80.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of a head-direction cell ensemble: a theory. *Journal of Neurosciences* **16**: 2112–2126.

Further Reading

Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics* **27**: 77–87.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

Koch C and Segev I (1998) *Methods in Neuronal Modeling*, 2nd edn, chaps. VII and XIII. Cambridge, MA: MIT Press.

Tuckwell HC (1988) *Introduction to Theoretical Neurobiology: Nonlinear and Stochastic Theories*. Cambridge, UK: Cambridge University Press.

Wilson HR (1999) *Spikes, Decisions, and Actions: The Dynamical Foundations of Neuroscience*. Oxford, UK: Oxford University Press.

Non-monotonic Logic

Advanced article

Michael Gelfond, Texas Tech University, Lubbock, Texas, USA
Richard Watson, Texas Tech University, Lubbock, Texas, USA

CONTENTS

A logic is called non-monotonic if assertions made in a theory in the logic may be retracted when new information is added to the theory.

THE NON-MONOTONIC CHARACTER OF COMMONSENSE REASONING

The use of formal logic as a means for computer programming was first seen in the work of Newell, Simon, and Shaw with their 'logic theory machine' (Newell and Simon, 1956). The work described a system for 'discovering proofs for theorems in symbolic logic'. A few years later, John McCarthy (1959) advocated a logical approach to artificial intelligence (AI), which would lead to the development of non-monotonic logics.

According to this approach, an intelligent agent should have knowledge of its world and its goals and the ability to use this knowledge to infer its course of action. To successfully perform tasks in a changing environment the agent should be able to make tentative conclusions based on available information but be prepared to withdraw some of these conclusions at a later time when further information becomes available.

For instance, I may know that my car is normally in working condition and plan to use it tomorrow for driving to work. Suppose in the morning I discover that the lights were left on and the battery of the car is dead. I then have to abandon my previous plan and look for alternative ways to get to work. The logic-based approach to AI suggests that a mathematical model of an agent capable of such behavior should contain a formal language capable of expressing commonsense knowledge about the world, a precise characterization of the valid conclusions that can be derived from theories stated in this language, and a means by which the agent can arrive at these conclusions. (*See* **Knowledge Representation**)

In the late 1970s, AI researchers tried to build such a model in the framework of classical logic. They quickly discovered that the task is far from trivial. The difficulty is rather deep, and is related to the so-called *monotonicity* of the entailment relation of classical logic. An entailment relation $\models$ between sentences of some logical language $\mathcal{L}$ is called monotonic if for any formulae A, B, and C of $\mathcal{L}$, if $A \models C$ then $A \,\&\, B \models C$ (i.e., if A entails C then so does any extension of A). Monotonicity is natural for classical logic, which was primarily aimed at formalization of mathematical reasoning, where axioms are rarely extended and inferences are long and complex. In mathematics, truth is not invalidated by the addition of new axioms; once proven, a theorem stays proven. In commonsense reasoning, additions to an agent's knowledge are frequent and inferences are usually short but are often based on tentative conclusions. These properties show the inherent non-monotonicity of this type of reasoning and suggest that it may be better modeled by logics with non-monotonic entailment relations. Early definitions of such relations were connected with attempts to clarify reasoning in taxonomic hierarchies (Touretzky, 1986), the meaning of the closed world assumption in databases (Reiter, 1978), and the semantics of the 'negation as failure' operator of logic programming (Clark, 1978). More powerful and general non-monotonic reasoning formalisms – notably circumscription (McCarthy, 1980), default logic (Reiter, 1980), and non-monotonic modal logics (McDermott and Doyle, 1980; Moore, 1985) – appeared almost simultaneously in the early 1980s.

TAXONOMIC HIERARCHIES

Often the knowledge of a reasoning agent contains a collection of classes of objects organized in a taxonomic hierarchy. In the hierarchies illustrated

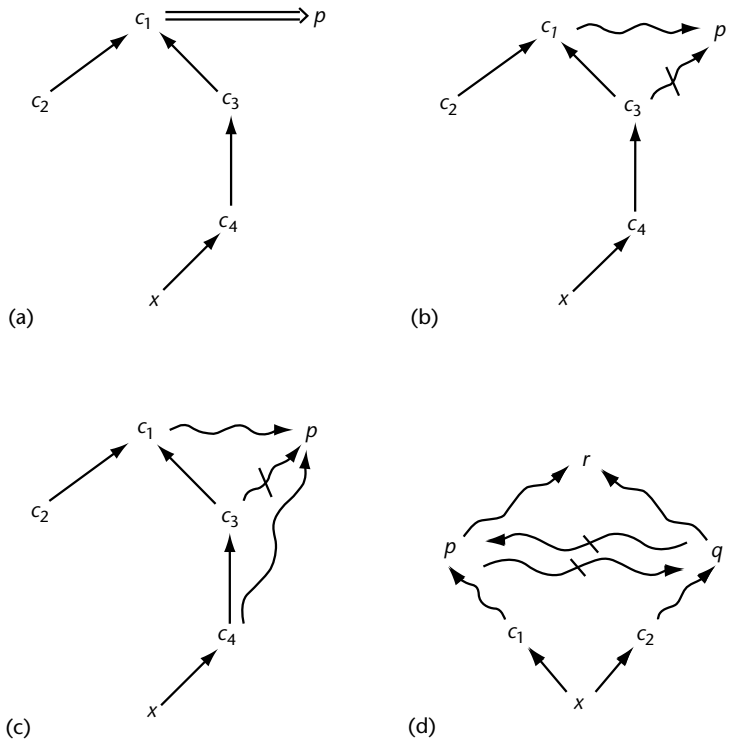


Figure 1. Taxonomic hierarchies. In diagram (a) we can conclude that x has property p. Diagrams (b) and (c) illustrate how the inheritance principle can cause us to modify our conclusions based on new information: in (b) we conclude that x does not have p, but in (c) we conclude again that x has p. Diagram (d) shows an ambiguous situation in which the inheritance principle provides no guidance.

in Figure 1, c_1, c_2, c_3, and c_4 denote classes of objects, $c_i \rightarrow c_j$ indicates that c_i is a proper subclass of $c_j (c_i \subset c_j)$ and $x \rightarrow c$ indicates that x is a member of $c (x \in c)$. Double arrows, such as in Figure 1(a), are used to link classes to properties: $c \Rightarrow p$ means that every element of class c has property p. Hierarchies containing only the elements mentioned above are called *strict*. They are often used to encode the hierarchical structure of a domain in a compact way. To establish that x has property p (denoted by the formula $p(x)$) it is sufficient to find a path from x to p. If no such path exists then x does not have property p. This reasoning can easily be justified by translating a hierarchy H into a first-order theory $T(H)$, such that for every object x of the hierarchy, $T(H) \models p(x)$ iff H contains a path from x to p. Here $\models$ denotes the entailment relation of classical logic.

In Figure 1(b), a link $c \rightsquigarrow p$ indicates only that elements of class c 'normally' have property p. Similarly, $c \not\rightsquigarrow p$ indicates that normally elements of c do not have property p. Statements of this form are called 'defaults'. They do not occur in mathematics, but seem to constitute a large portion of our commonsense knowledge about the world. A substantial part of our education consists of learning defaults and their exceptions, as well as various ways of reasoning with this knowledge. Given the hierarchy of Figure 1(b) a rational agent will have no difficulty in following the path $x \rightarrow c_4 \rightarrow c_3 \not\rightsquigarrow p$ which leads to the conclusion that x does not satisfy p. This is despite the fact that the path $x \rightarrow c_4 \rightarrow c_3 \rightarrow c_1 \rightsquigarrow p$ corresponds to an argument which contradicts this conclusion. This seems to happen because of the use of a commonsense idea known as the *inheritance principle* (Touretzky, 1986) which states that, in our reasoning, more specific information prevails over less specific information. Since c_3 is a proper subset of c_1, we know that the default $c_3 \not\rightsquigarrow p$ is more specific than $c_1 \rightsquigarrow p$, and so we prefer the first argument. Inheritance reasoning is non-monotonic. Expanding the hierarchy by a link $c_4 \rightsquigarrow p$ (see Figure 1(c)) will lead to the creation of a new preferred path, $x \rightarrow c_4 \rightsquigarrow p$. Given this new information we will be forced to change our view and conclude that x satisfies p after all.

The inheritance principle was one of the first useful principles of commonsense reasoning discovered by AI researchers. Attempts to formalize this principle (i.e. to precisely characterize the valid

conclusions that can be drawn by a rational agent whose knowledge is represented by an inheritance hierarchy) led to two distinct approaches to the problem. Direct theories of inheritance focus on paths of the network, viewed as possible arguments, and on defining the relative strengths of these arguments. For the hierarchy in Figure 1 (b), the argument $x \to c_4 \to c_3 \to c_1 \rightsquigarrow p$ is defeated by the argument $x \to c_4 \to c_3 \not\rightsquigarrow p$, which is not defeated by any other argument. Hence the conclusion $p(x)$ is justified. Figure 1(d) shows a more complex hierarchy in which properties p and q are normally mutually exclusive but elements having either one of them normally have property r. Here neither the argument $x \to c_1 \rightsquigarrow p$ nor $x \to c_2 \rightsquigarrow q \not\rightsquigarrow p$ is more specific than the other and hence neither defeats the other. The same is true for the property q; therefore we can conclude neither $p(x)$ nor $q(x)$. The answer given by this hierarchy to the query $r(x)$ depends on the precise definition of plausible counterargument. A 'sceptical' reasoner will argue that since the truths of $p(x)$ and $q(x)$ are unknown we should refrain from making any conclusions about $r(x)$. A 'credulous' reasoner can take a view that the net from Figure 1(d) sanctions the conclusion $p(x) \vee q(x)$, and hence $r(x)$ must be true. A detailed discussion of direct theories of inheritance can be found in Horty (1994).

Indirect approaches to inheritance are based on mapping the network together with the form of inheritance principle used in its semantics to more general theories of non-monotonic reasoning. Some insight into the relationship between direct and indirect approaches can be found in You *et al*. (1999).

THE CLOSED WORLD ASSUMPTION

The following example illustrates another typical situation requiring non-monotonic reasoning. Suppose the list of faculty staff of a small computer science department is posted on the wall of the main office. One may naturally assume that the list is complete: if Michael is not mentioned on it then he is not a faculty member in this department. This is, of course, only an assumption. (It is possible, for example, that the name of a recently appointed faculty member has not yet been added to the list.) Assumptions like this are called *closed world assumptions* (Reiter, 1978). They are frequently used by people acting in everyday situations. Such assumptions are also built into the semantics of databases and logic programming languages such as Datalog and PROLOG. In databases, a table describing a relation r contains the objects satisfying r. Objects that do not occur in the table are assumed not to satisfy r. In logic programming, an answer 'no' to a query q is often interpreted as 'q is false' (instead of a more precise but less useful 'PROLOG interpreter cannot prove q from information given in a program'). The closed world assumption is expressible in practically all general-purpose non-monotonic formalisms, including those discussed below.

NON-MONOTONIC REASONING IN LOGIC PROGRAMMING

Some of the earliest non-monotonic reasoning systems were developed in the framework of logic programming. Originally, logic programs were viewed as collections of clauses (or rules) of the form

$$p \leftarrow \Gamma \tag{1}$$

which is read as 'p if Γ'. Here the head, p, is an atom and the body, Γ, is a sequence of atoms. A clause with an empty body is called a 'fact'. With the advent of practical logic programming languages the rules became more complex. For instance, in the logic programming language PROLOG, the body of a rule is allowed to contain expressions of the form `not q`, interpreted as 'PROLOG interpreter failed to prove q'. Note that 'not' here is different from the negation $\neg$ in classical logic where $\neg q$ is interpreted as 'q is false'. Thus, a program π_0 consisting of the two rules

$$p(X) \leftarrow \texttt{not}\ q(X);\ p(X) \leftarrow r(X) \tag{2}$$

will answer 'yes' to a query $p(a)$. The PROLOG interpreter will attempt to prove $q(a)$, fail to do so, and conclude $p(a)$. If we extend π_0 by the new fact $q(a)$, then the previous conclusion will be withdrawn and the answer to $p(a)$ will become 'no'. This connective `not` is called 'negation as finite failure', and was introduced primarily as a procedural device. The first declarative semantics of this connective was suggested by Clark (1978). Under this semantics, the collection of rules with the head $p(X)$ in π_0 is regarded as a definition of p: a shorthand for the formula

$$\forall X\ p(X) \equiv (\neg q(X) \vee r(X)) \tag{3}$$

This idea is used to translate a program π into a set $\mathrm{comp}(\pi)$ of first-order formulae called the *predicate completion* of π. The theory $\mathrm{comp}(\pi_0)$ consists of statement 3 above, as well as the statements

$$\forall X \neg r(X) \tag{4}$$

and

$$\forall X \neg q(X) \tag{5}$$

and the collection of equality axioms guaranteeing that distinct names in the language represent distinct objects of the domain. The completion comp(π_1) of the program $\pi_1 = \pi_0 \cup \{q(a)\}$ is obtained by replacing statement 5 in comp(π_0) by $\forall X\ q(X) \equiv (X = a)$. By definition, a logic program π entails ($\models$) a literal l (not containing variables) iff l is classically entailed by comp(π). Note that comp(π_0) $\models p(a)$ but comp(π_1) $\models \neg p(a)$.

Work on the semantics of logic programs continued in several directions. The meaning of `not` was refined to make it less dependent on the particular inference mechanisms associated with the PROLOG interpreter (Gelfond and Lifschitz, 1988), and to allow abductive reasoning (Kakas *et al.*, 1992; Denecker and De Schreye, 1992). These approaches were generalized to accommodate programs with rules formed using more complex formulae. The resulting logical languages and entailment relations provide powerful means for representing and reasoning with commonsense knowledge. Suppose we wish to represent the facts that Tweety is a bird and birds normally fly. This can be done by the program consisting of two rules:

$$\texttt{fly}(X) \leftarrow \texttt{bird}(X), \texttt{not} \neg \texttt{fly}(X);\ \texttt{bird}\ (\texttt{tweety}) \tag{6}$$

Since there is no reason to believe that Tweety does not fly, the program concludes `fly (tweety)`. If we learn that Tweety is not a flying bird and expand our program by a new fact, $\neg$ `fly (tweety)`, the resulting program will be consistent and entail that Tweety does not fly.

CIRCUMSCRIPTION

The basic ideas and intuition behind another powerful non-monotonic system called *circumscription* were described and formalized by John McCarthy (1977, 1980). In circumscription, theories are written in classical first-order logic, but the entailment relation is not classical. A formula F is entailed by a circumscriptive theory T if it is true in all minimal models of T. A model M of a theory T is called minimal with respect to some partial ordering $<$ of models if there is no model N of T such that $N < M$. A circumscription policy is used to determine the particular partial ordering $<$ used to circumscribe a theory. In the basic case, a single predicate P is chosen to be circumscribed. Given two models M_1 and M_2, which differ only in their interpretations of P, we say that $M_1 \leq M_2$ if the extent of P in M_1 is a subset of its extent in M_2. We write `circ` (T; P) $\models F$ to indicate that F is entailed from T by circumscription with the policy of minimizing P. Here `circ` (T; P) can be viewed as a second-order formula expressing the above definition. (Recall that second-order formulae allow quantifiers over relations.)

Let us apply this idea to a version of the flying birds example above. We can attempt to express that Tweety is a bird and birds normally fly by first-order sentences

$$T = \{\texttt{bird tweety and}\ \forall X\ ((\texttt{bird}(X) \wedge \neg \texttt{ab}(X)) \supset \texttt{fly}(X))\} \tag{7}$$

where `ab` (X) means that X is abnormal with respect to flying. Obviously, classical first-order logic does not allow us to reach the desired commonsense conclusion that Tweety can fly. If, however, we use circumscription and circumscribe `ab`, then all minimal models under this policy contain `fly (tweety)`; hence `circ` (T; `ab`) $\models$ `fly (tweety)`. This basic form of circumscription is often too restrictive, so many other circumscriptive policies have been formalized. One common policy is to specify certain predicates which are allowed to vary. In this case, models are comparable if they differ in the extent of the varied predicates as well as the circumscribed predicate. As before, the ordering of comparable models is based solely on the extent of the circumscribed predicate in the models. Suppose we add to our example the fact that penguins are birds which are abnormal with respect to flying: (`penguin`(X) $\supset$ `bird`(X) `and` `penguin`($X \supset$ `ab`(X)). Since a model M_0 in which Tweety is a penguin is minimal with respect to ordering defined by the first policy, we no longer have `circ` (T; `ab`) $\models$ `fly (tweety)`. If we modify the policy so that `penguin` can vary, the model M_0 will not be minimal with respect to the new ordering. Under this policy, T concludes `fly (tweety)`. Selection of the right circumscriptive policy lies at the heart of representing knowledge in circumscriptive theories. Even though computing consequences of circumscribed theories are generally intractable (even in propositional cases), for some theories there are reasonably efficient algorithms based on reducing the circumscribed theory to a logic program or a set of first-order formulae.

DEFAULT LOGIC

Another non-monotonic formalism which uses classical first-order logic as its base is default logic (Reiter, 1980). A default theory is a pair (D, W)

where W is a collection of statements of first-order logic and D is a set of default rules, which are statements of the form

$$\frac{A : \mathtt{M}B_1, \ldots, \mathtt{M}B_n}{C} \tag{8}$$

where A, each B_i, and C are classical formulae. A default rule can be read as 'if A is provable and each B_i is possible then conclude C', and can be used as a non-monotonic rule of inference. For instance, a proof of r in a default theory $T_1 = (\{\frac{q:\mathtt{M}r}{r}\}, \{p, p \supset q\})$ consists of a 'classical' step, deriving q, and an application of the default rule of T_1, deriving r. The second step is justified by T_1's inability to prove $\neg r$. A collection of formulae that can be derived from a default theory T in this fashion is called an *extension* of T. Extensions are often interpreted as sets of beliefs which can be formed by a rational agent on the basis of T. The default theory T_1 above has one such extension, Cn (p, $p \supset q$, r), where Cn(S) stands for the set of all 'classical' consequences of S. Default theory $T_2 = (\{\frac{r:\mathtt{M}\neg p}{q}, \frac{r:\mathtt{M}\neg q}{p}\}, \{r\})$ has two extensions, Cn (r, q) and Cn(r, p), and default theory $T_3 = (\{\frac{:\mathtt{M}\neg p}{p}\}, \phi)$ has none. (The latter fact can be attributed to the irrationality of the default rule of T_3.) The precise notion of an extension of a default theory was first defined by a fixed-point construction in Reiter (1980). Although some variants of Reiter's notion of an extension have been advocated, the basic idea of the construction remains the same. Fixed-point constructions somewhat similar to that of Reiter were used in several other early non-monotonic systems. In particular they are used to define the semantics of the 'modal' non-monotonic logics of McDermott and Doyle (1980) and their variants (Moore, 1985). For more details see Marek and Truszczyński (1993).

Default logic is a powerful tool which can be used to give semantics of other non-monotonic formalisms. For instance, the hierarchy in Figure 1(d) can be represented by a default theory $T_4 = (D, W)$ where $D = \{\frac{c_1(X):\mathtt{M}p(X)}{p(X)}, \frac{c_2(X):\mathtt{M}q(X)}{q(X)}, \frac{p(X):\mathtt{M}r(X)}{r(X)}, \frac{q(X):\mathtt{M}r(X)}{r(X)}\}$ and $W = \{c_1(x), c_2(x), \forall X(p(X) \supset \neg q(X))\}$. (Here the rules of T_4 containing variables should be read as a shorthand for the set of their instantiations.) The theory has two extensions: one containing $\{p(x), \neg q(x), r(x)\}$ and another containing $\{q(x), \neg p(x), r(x)\}$. The set of valid conclusions that can be made by a 'credulous' inheritance reasoner on the basis of such hierarchies can now be defined as the set of formulae that belong to all extensions of the corresponding default theory. Default theories can also be used to give a semantics to logic programs. A rule $r = l_0 \leftarrow l_1, \ldots, l_m, \mathtt{not}\ l_{m+1}, \ldots, \mathtt{not}\ l_n$, where the l_i are literals, can be translated to a default rule

$$d(r) = \frac{l_1, \ldots, l_m : \mathtt{M}\bar{l}_{m+1}, \ldots, \mathtt{M}\bar{l}_n}{l_0} \tag{9}$$

Note that $\bar{l}$ stands for the complement of l; i.e. if l is an atom then $\bar{l} = \neg l$, and if $l = \neg a$ where a is an atom then $\bar{l} = a$. A program π consisting of such rules entails a formula F if F belongs to all the extensions of the default theory $d(\pi) = (\{d(r) : r \in \pi\}, \phi)$. It can be shown that the semantics defined in this way coincides with the answer set semantics (Gelfond and Lifschitz, 1991) of logic programs. There is also a close connection with truth-maintenance systems (TMSs) (Doyle, 1979): programs which manage sets of nodes and sets of justifications. A node can be viewed as a set of formulae, while justifications correspond to logic programming (or default) rules. Each justification states that one node should be believed if certain others are believed and certain others disbelieved. The precise mapping establishing this correspondence helped increase our understanding of what is computed by TMSs and at the same time linked theoretical work on non-monotonic reasoning with implemented reasoning systems.

It remains to be seen whether the impressive power of default logic, circumscription, and other 'superclassical' non-monotonic systems will make them tools of choice for representing commonsense knowledge, or if weaker but simpler languages (similar to those based on logic programming) will be preferred by knowledge engineers. In the 1990s substantial progress was made in building practical systems capable of non-monotonic reasoning. In addition to PROLOG and TMSs, we now have efficient systems, such as Smodels (Niemelä and Simons, 1997), CCALC (McCain, 1997), and dlv (Citrigno *et al.*, 1997), capable of computing answer sets of various classes of logic programs and default theories. These systems form the basis for a style of programming in which application problems are encoded so that their solutions are fully described by answer sets of the corresponding program. There are applications of this approach to AI problems such as planning and diagnosis. Non-monotonic systems also played an important role in the development of theories of action and change. In such theories one needs to describe causal laws, which define the effects of performing actions, as well as causal relations between the fluents (propositions whose values depend on time). One also needs to specify which fluents do

not change. The problem of finding a compact representation of what changes and what does not is called the 'frame problem'. In 1986, McCarthy suggested that the second part of this problem might be solved by the use of a default rule that states that actions normally do not change values of fluents. Attempts to solve this and other related problems have motivated much work on non-monotonic logics. More details on this subject can be found in Shanahan (1997).

NON-MONOTONICITY AND PROBABILISTIC REASONING

The formalisms discussed above allow reasoning about the truth or falsity of logical statements. In some cases the truth or falsity can be determined only with a certain degree of plausibility. AI researchers have used a variety of different approaches, such as Bayesian theory, Dempster–Shafer theory, fuzzy logic, and certainty factors, to develop formalisms for reasoning with such information. These formalisms are closely related to work on non-monotonic reasoning. A collection of important papers on probabilistic reasoning methods can be found in Pearl and Shafer (1990). To illustrate the use of such approaches, we consider Bayesian belief networks. Belief networks are applicable in situations where the plausibility of a statement can be given by the mathematical probability of the corresponding event. (*See* **Reasoning under Uncertainty**; **Bayesian Belief Networks**; **Fuzzy Logic**)

Consider the following situation. A person has a burglar alarm in his house and lives in an area where burglaries are fairly infrequent. If there were a burglary, the alarm would probably sound. False alarms sometimes occur as well. Suppose that, while at work, the person gets a call from a neighbor who says that the alarm is sounding. Suppose the person also knows that, while this particular neighbor could be trusted to call if the alarm really was on, the neighbor is also a practical joker and may say the alarm is ringing is a prank. The person has a second neighbor whom he can ask about the alarm. This neighbor lives further away from the person's house and, because of the distance, may not hear the alarm even if it is active. A belief network formalizing information from this example is shown in Figure 2.

In the figure, events are depicted by ovals. Beside each oval is a table giving the conditional probabilities of the event. The statement that 'burglaries are fairly infrequent' is expressed by assigning a very low probability to the event ($P(B) = 0.001$). An

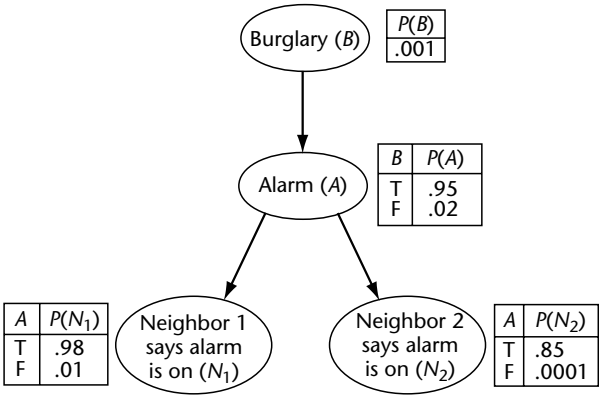


Figure 2. A Bayesian belief network. In this example, an agent knowing N_1 would have a 68% belief in A. This belief drops to 24% of the agent then learns that N_2 is false.

arrow from event X to event Y indicates that Y can be caused by X. The chances of this happening are given by the conditional probabilities in the corresponding table. For example, the table next to the alarm event states that the alarm goes off 95% of the time when there is a burglary and 2% of the time when there is not one. Using the information given in the belief network, algorithms based on Bayesian theory can be used to compute the probability of an event given knowledge of the occurrences of other events. The agent's belief is then based on the computed probability. In the example, Bayesian theory implies that the probability that the alarm sounded given the fact that the first neighbor called is 68%, and hence the agent would believe that the alarm was on. This approach is non-monotonic in that if the agent gets further information he may well change the computed probability, and may therefore alter his beliefs. For example, if the second neighbor calls and says that she does not hear the alarm then the computation returns only a 24% probability that the alarm is on, and therefore the agent changes his belief. A more complete discussion of the relationship between non-monotonic logics and probabilistic reasoning methods can be found in Goldszmidt and Pearl (1996).

References

Citrigno S, Eiter T, Faber W *et al*. (1997) The dlv system: model generator and application frontends. In: Bry(F), Freitag B and Seipel D (eds) *Proceedings of the 12th Workshop on Logic Programming*, pp. 128–137. Munich, Germany: LMU.

Clark K (1978) Negation as failure. In: Gallaire H and Minker J (eds) *Logic and Data Bases*, pp. 293–322. New York, NY: Plenum.

Denecker M and De Schreye R (1992) SLDNFA: an abductive procedure for normal abductive logic programs. In: Apt K (ed.) *Proceedings of JICSLP*, pp. 686–700. Cambridge, MA: MIT Press.

Doyle J (1979) A truth maintenance system. *Artificial Intelligence* **12**(3): 231–272.

Gelfond M and Lifschitz V (1988) The stable model semantics for logic programming. In: Kowalski R and Bowen K (eds) *Logic Programming: Proceedings of the 5th International Conference and Symposium*, pp. 1070–1080. Seattle, WA: MIT Press.

Gelfond M and Lifschitz V (1991) Classical negation in logic programs and disjunctive databases. *New Generation Computing* **9**: 365–385.

Goldszmidt M and Pearl J (1996) Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence* **84**(1–2): 57–112.

Horty J (1994) Some direct theories of nonmonotonic inheritance. In: Gabbay D, Hogger C and Robinson J (eds) *Handbook of Logic in Artificial Intelligence and Logic Programming*. vol. III: *Nonmonotonic Reasoning and Uncertain Reasoning*. New York, NY: Oxford University Press.

Kakas A, Kowalski R and Toni F (1992) Abductive logic programming. *Journal of Logic and Computation* **2**: 719–771.

Marek V and Truszczyński M (1993) *Nonmonotonic Logic*. Berlin, Germany: Springer-Verlag.

McCain N (1997) *Causality in Commonsense Reasoning About Actions*. PhD thesis, University of Texas.

McCarthy J (1959) Programs with common sense. In: *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pp. 75–91. London, UK: HMSO.

McCarthy J (1977) Epistemological problems of artificial intelligence. In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence (IJCAI–77)*, pp. 1038–1044. Cambridge, MA: MIT Press.

McCarthy J (1980) Circumscription: a form of non-monotonic reasoning. *Artificial Intelligence* **13**(1–2): 27–39, 171–172.

McDermott D and Doyle J (1980) Non-monotonic logic I. *Artificial Intelligence* **13**(1–2): 41–72.

Moore R (1985) Semantical considerations on nonmonotonic logic. *Artificial Intelligence* **25**(1): 75–94.

Newell A and Simon H (1956) The logic theory machine: a complex information processing system. *IRE Transactions on Information Theory* **2**(3): 61–79.

Niemelä I and Simons P (1997) Smodels: an implementation of the stable model and well-founded semantics for normal logic programs. In: Dix J, Furbach U and Nerode A (eds) *Proceedings of the 4th International Conference on Logic Programming and Non-Monotonic Reasoning*, pp. 420–429. Dagstuhl, Germany: Springer-Verlag.

Pearl J and Shafer G (eds) (1990) *Readings in Uncertain Reasoning*. San Mateo, CA: Morgan Kaufmann.

Reiter R (1978) On closed world databases. In: Gallaire H and Minker J (eds) *Logic and Data Bases*, pp. 119–140. New York, NY: Plenum.

Reiter R (1980) A logic for default reasoning. *Artificial Intelligence* **13**(1–2): 81–132.

Shanahan M (1997) *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA: MIT Press.

Touretzky D (1986) *The Mathematics of Inheritance Systems*. Los Altos, CA: Morgan Kaufmann.

You J, Wang X and Yuan L (1999) Compiling defeasible inheritance networks to general logic programs. *Artificial Intelligence* **113**(1–2): 247–268.

Further Reading

Brewka G, Dix J and Konolige K (1997) *Nonmonotonic Reasoning: An Overview*. Stanford, CA: CSLI Publications.

Gabbay D, Hogger C and Robinson J (eds) (1994) *Handbook of Logic in Artificial Intelligence and Logic Programming*. vol. III: *Nonmonotonic Reasoning and Uncertain Reasoning*. New York, NY: Oxford University Press.

Marek V and Truszczyński M (1993) *Nonmonotonic Logic*. Berlin, Germany: Springer-Verlag.

McCarthy J and Lifschitz V (eds) (1990) *Formalizing Common Sense: Papers by John McCarthy*. Norwood, NJ: Ablex.

Shanahan M (1997) *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA: MIT Press.

Pattern Recognition, Statistical

Intermediate article

Thomas Hofmann, Brown University, Providence, Rhode Island, USA

CONTENTS

Statistical pattern recognition deals with the problem of automatically classifying objects as belonging to one or more classes from a set of possible classes. Objects are typically represented by raw data that may include measured or known object properties and which are summarized in a feature vector. The general goal is to infer suitable classification rules that map feature vectors to class labels based on a training set of objects with known class memberships.

INTRODUCTION

Definition and Challenges

Humans solve pattern recognition problems all the time. Our visual system constantly identifies and recognizes objects and people under highly variable conditions such as pose, viewing angle and illumination. Our auditory system distinguishes sounds and phonetic units from a noisy acoustic data stream. We constantly group, structure, and interpret sensory inputs and integrate them with past experience to guide us in making decisions.

Statistical pattern recognition is a discipline at the intersection of computer science, statistics and engineering that deals with the question of how learning from examples can improve pattern recognition systems. The central problem is that of inductive inference: how can one infer a general recognition rule from a finite set of examples? How can a correct generalization from the available training examples be guaranteed? What algorithms should be used to implement automatic pattern recognition systems?

Pattern recognition research also involves certain problems that are presumed in inductive inference, namely the availability of a suitable representation for patterns (typically as a vector in some 'feature space') and the specification of suitable target classes or concepts.

Supervised Classification and Concept Learning

Formally, one commonly distinguishes between supervised and unsupervised classification. In supervised classification, it is assumed that patterns belong to one or more of a given set of classes and that a training set with correctly labeled examples is available. In unsupervised classification, the main challenge is to identify what the relevant classes are in the first place. The intention is usually to find classes that are in some way natural for the given domain, a task that is most commonly carried out by grouping patterns into clusters of similar patterns ('data clustering').

This article will focus on supervised learning, and for the most part deal with the simple but generic case of binary classification or 'concept learning'. In concept learning, an example can either be positive (being an example of the concept) or negative (not being an example of the concept); formally, the goal is to find a mapping $f: X \to \Omega$ that associates patterns $\mathbf{x}$ from some input space X with their corresponding class or concept labels $f(\mathbf{x})$. The classification function f is also called a 'discriminant function' and in the two-class case ($\Omega = \{-1, 1\}$) it is referred to as a 'dichotomy'. (A problem closely related to classification is regression, where the response variable is continuous and real-valued. In fact, many classification methods have regressional analogs.) Geometrically, one can often visualize a classification function by the decision boundaries it implies. For example, a dichotomy will define a boundary that separates the positive and the negative patterns (see Figure 1).

In statistical pattern recognition, the classification function has to be inferred from a set of examples, the so-called training set $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, where each example consists of a pattern $\mathbf{x}_i$ and an associated class label y_i.

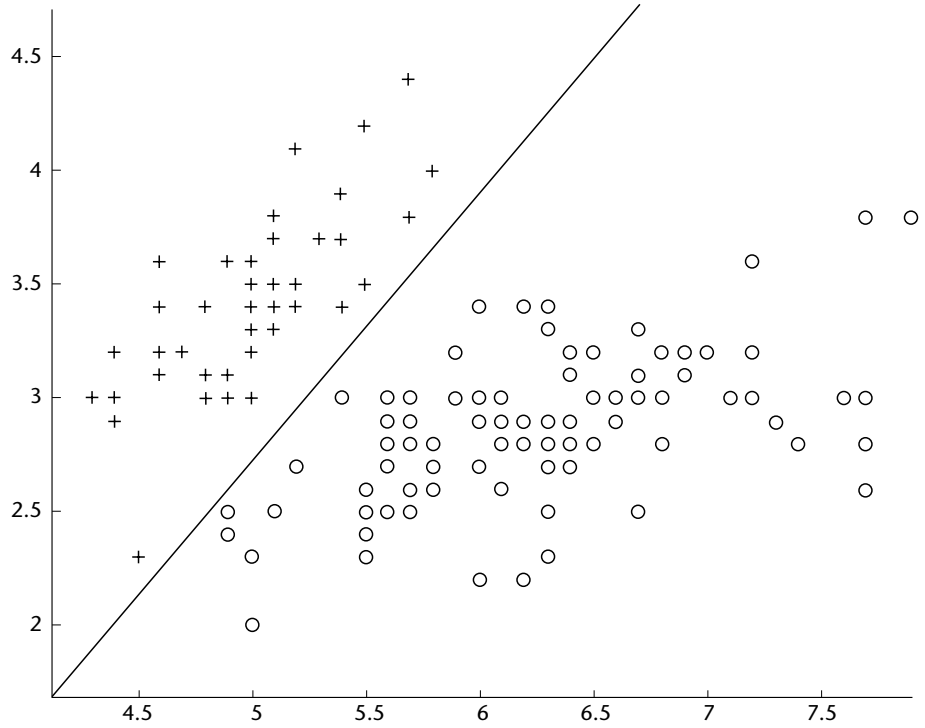


Figure 1. An example of a (linear) dichotomy on a two-dimensional data set. The linear discriminant function is shown as a solid line. 'Positive' examples lie on one side of the line and 'negative' examples on the other.

Classification problems arise in a vast range of applications. For example, in object recognition (e.g. optical character recognition), given a finite set of object classes, images must be classified according to the classes of objects they depict. In speech recognition, waveforms must be mapped to discrete phonetic or syntactic units. In text categorization, documents must be categorized or annotated as dealing with specific topics, being of interest to specific people, or being of a specific genre. In molecular biology, proteins or DNA sequences must be classified as belonging to known families of proteins or known types of signals or genes.

THEORY OF PATTERN RECOGNITION

Bayes-Optimal Classification

Before designing or even engineering a system, one needs to investigate the theoretical limits of the problem and identify the (perhaps unachievable) optimal solution. In order to quantify the goodness of a classifier, one commonly uses the zero–one loss function which assigns a loss of 0 to a correct classification ($f(\mathbf{x}_i) = y_i$) and a loss of 1 to a misclassification ($f(\mathbf{x}_i) \neq y_i$).

One typically assumes that the observed pattern–label pairs are generated independently from some unknown but fixed probability distribution P. The goal in statistical pattern recognition then is to find a classification function that minimizes the expected zero–one loss. The corresponding risk functional in the binary case of $\Omega = \{-1, 1\}$ can be written as

$$\begin{aligned} R(f) &= \int |f(\mathbf{x}) - y| \mathrm{d}P(\mathbf{x}, y) \\ &= \frac{1}{2} \int (1 - f(\mathbf{x})y) \mathrm{d}P(\mathbf{x}, y) \end{aligned} \tag{1}$$

A classifier that minimizes the risk, and hence achieves the best classification accuracy in expectation, is called a 'Bayes-optimal' classifier. It can be proven that a Bayes-optimal classifier assigns a pattern $\mathbf{x}$ to the class $y \in \Omega$ with the highest posterior probability $P(y|\mathbf{x})$, which by Bayes' theorem is given as

$$P(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_{y' \in \Omega} p(\mathbf{x}|y')P(y')} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{2}$$

Here the 'prior probability' $P(y)$ refers to the probability of sampling a pattern belonging to

class y, and the 'likelihood' $p(\mathbf{x}|y)$ is the class-conditional probability density of patterns $\mathbf{x}$, i.e. it represents the distribution of patterns $\mathbf{x}$ belonging to a particular class y. The normalization in the denominator ('evidence') is the same for all classes and can be ignored in most cases. To illustrate the meaning of Bayes' theorem in this context, assume for simplicity uniform class-prior probabilities. In this case, one would simply favor the class y that has the highest probability (density) to generate the observed pattern.

Since neither the probability distribution $P(\mathbf{x},y)$ nor the derived class-conditional probability densities $p(\mathbf{x}|y)$ are known, the Bayes-optimal classifier represents an unattainable 'gold standard'.

Generalization

The central problem in classification is the problem of generalization: how can the classification risk be minimized without direct access to the underlying distribution that generates the patterns and class labels? Is there an appropriate substitute for the classification risk that could be used to train classifiers and to make formal guarantees about the achieved classification accuracy? In particular, how is the empirical risk, i.e. the classification accuracy achieved on the training sample, related to the true risk? The potentially large discrepancy between the accuracy on training and new test data is often referred to as 'overfitting'.

These fundamental questions are dealt with in computational learning theory. Here we will briefly discuss only the basic philosophy. The Vapnik–Chervonienkis theory (Vapnik, 1999) attempts to derive upper bounds on the difference between the true risk and the empirical risk. This would guarantee (with high probability) that the classifier showing the best performance on the training data will also perform well on new patterns. Ideally, such bounds will be distribution-independent, so that no additional assumptions about P have to be made. Instead, generalization performance will depend on the complexity of the hypothesis class, i.e. the set of classification functions that are considered as candidates. If the hypothesis class is 'too large', it is impossible to guarantee that a finite training set will be sufficient to identify a well-performing classification function (overfitting). If the hypothesis class is 'too small', generalization might be guaranteed, but the training accuracy might be poor (underfitting). In practice, one has to find a balance between these two situations. Vapnik–Chervonienkis theory makes precise the mathematical form of these bounds and how to measure the complexity of a hypothesis class.

More recently, advances have been made in deriving data-dependent generalization bounds. Here one aims at defining appropriate quantities that can be constructively evaluated on the training set and that lead to bounds on the generalization performance. Examples of such data-dependent quantities are geometrical margins (related to 'support vector machines') and ensemble margins (related to boosting algorithms).

In practice, one often prefers an empirical approach to assess the classification accuracy, rather than theoretical bounds. In 'n-fold cross-validation' one partitions the data set into n subsets (a typical value being $n = 10$) and trains a classifier on $(n - 1)$ subsets, while testing classification accuracy on the remaining subset not used for training. This is repeated n times (once for each choice of $(n - 1)$ subsets) and the average of the test errors is used as an estimate of the generalization error. While practical and simple, this procedure does not provide any conceptual insight into how to design classifiers and learning architectures.

Feature Selection and Dimension Reduction

Another important factor is the representation of patterns used for classification. The achievable classification accuracy depends on how much information about the correct class can be derived from the feature representation. This will be reflected in the optimal Bayes error, which depends on the chosen pattern representations. Choosing an appropriate representation requires domain knowledge. For example, how should images, documents, speech waveforms, or molecules be represented? What features should be extracted? Pattern recognition methods need at least some representation as a starting point.

One approach to selecting a feature representation is to include all features that could possibly be relevant to solving the classification problem. Indeed, Bayesian decision theory predicts that extracting more features will generally reduce the Bayes error, since a feature can only help in discriminating classes (in the worst case it can simply be ignored). Yet in practice the situation is more difficult. It is harder to learn in higher-dimensional spaces, and the performance of a classifier may degrade significantly with increasing dimensionality.

A common way to address this problem is to perform feature selection or dimension reduction

as a preprocessing step prior to training the classifier. In feature selection, one often selects or removes one feature at a time, guided either by cross-validation techniques (a computationally expensive procedure, since several classifiers have to be trained at each step) or by surrogate quantities like mutual information between feature and class labels. An alternative is feature reduction by techniques like principal components analysis that map patterns to some lower-dimensional representation.

SUPERVISED CLASSIFICATION METHODS

Bayes Plug-In Classifier

One approach to classification is to take the Bayes-optimal classifier as an ideal starting point and to replace the unknown probability distributions by estimates derived from the training data. The estimated probability distributions can then be used as a 'plug-in' estimator for the true distribution, and the classification problem reduces to the problem of probability density estimation. Additional simplifying assumptions – for example, about the parametric form of the class-conditional probability distributions – are commonly made, in which case the statistical inference problem becomes one of estimating the parameters for each class.

The main problem with this approach is that it is nondiscriminative, i.e. our model might focus on approximating aspects of the distribution of patterns in each class that are irrelevant for discriminating between classes. In particular, it is unclear how the accuracy in estimating class-conditional distributions relates to the performance of the classifier.

Nearest-Neighbor Methods

Nearest-neighbor classification is a 'nonparametric' classification technique, i.e. it does not make any assumptions about the parametric forms of the discriminant functions or the class-conditional distributions. Instead, it relies on the assumption that as the pattern density increases, neighboring patterns will be likely to belong to the same class. The 'k-nearest-neighbor' (kNN) rule determines the class label of a test point by identifying the most popular class among the k closest patterns in the training set. In the simplest case of 1NN, each pattern simply inherits the label of the closest pattern in the training set. In geometric terms, the input space is partitioned into regions of patterns that share the same neighborhood. In the case of 1NN this coincides with the Voronoi tessellation of the input space induced by the training examples. Within each region the class label is constant (see Figure 2).

Despite its apparent simplicity, the asymptotic classification error of the 1NN classifier is known to be at most twice as high as the optimal Bayes error rate (Devroye *et al*., 1996). But this only holds as the number of training data points tends

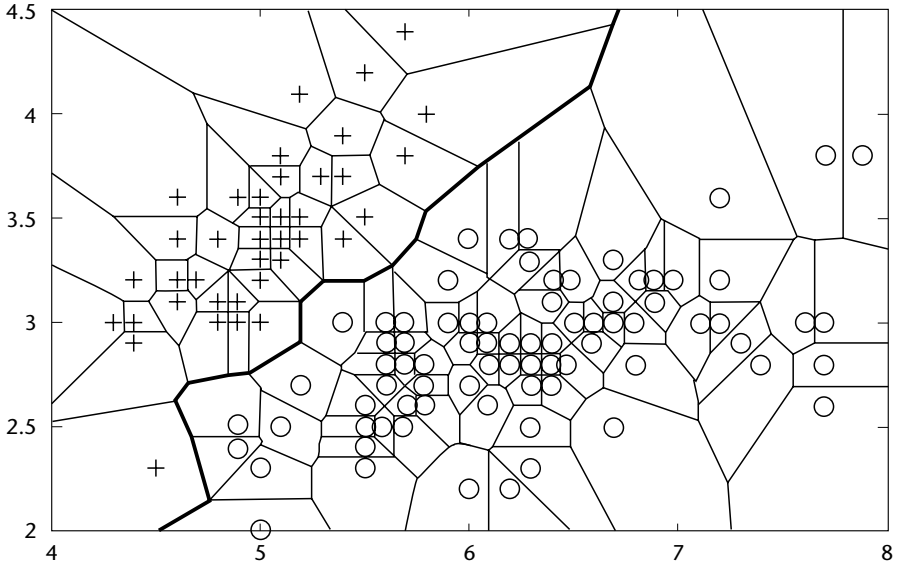


Figure 2. Nearest-neighbor classification for the data set depicted in Figure 1. The thin lines indicate the induced Voronoi tessellation, and the bold line indicates the resulting classifier.

to infinity and patterns densely fill the input space. In practice, this is usually unrealistic, especially in high-dimensional feature spaces. Whether or not kNN classification works well in practice depends strongly on the number of available training examples compared with the dimensionality of the feature space. The choice of a meaningful metric, or similarity function, is also essential: different similarity functions may result in very different classification accuracies.

Linear Classifiers

A common approach in pattern recognition is to restrict the set of classification functions to some family or hypothesis class at the outset. Then, the classifier within that family that performs best in the light of the training data is selected. For example, one may select a classification function from the hypothesis class that has the smallest training error (this is known as 'empirical risk minimization').

In linear classification, the hypothesis class consists of linear discriminants, of the form $f(\mathbf{x}; \mathbf{w}, b) = \text{sign}(\langle \mathbf{x}, \mathbf{w} \rangle + b)$, where $\mathbf{w}$ is a weight vector and b is a scalar offset called the bias. Since the equation $\langle \mathbf{x}, \mathbf{w} \rangle + b = 0$ defines a unique hyperplane with $\mathbf{w}$ being a normal vector, linear classifiers effectively partition the feature space into two half-spaces. In effect, the weight vector specifies a direction in feature space and patterns are projected onto the one-dimensional subspace spanned by $\mathbf{w}$. Within this subspace, features are separated by thresholding at b (see Figure 1).

Given a set of training patterns, how can we determine the optimal linear classifiers? Several methods are available for this purpose.

Perceptron algorithm

The perceptron algorithm is a simple iterative procedure to find a linear classifier, which produces zero training error for linearly separable data sets (i.e. if the positive and negative examples can be perfectly separated by a hyperplane in feature space). The basic idea is as follows. Cycle through the training patterns in some fixed or randomized order. Every time a pattern $\mathbf{x}_i$ is misclassified by the current choice of $\mathbf{w}$ and b, update the weights and bias according to

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} + y_i \mathbf{x}_i \\ b &\leftarrow b + y_i (\max_i ||\mathbf{x}_i||)^2 \end{aligned} \tag{3}$$

Thus, depending on the label of the misclassified pattern $y_i \in \{-1,1\}$, the pattern is added to or subtracted from the current weight vector and the bias is incremented or decremented by some constant. It can be proven that this will converge towards a solution that separates the training data (Novikoff's theorem). Moreover, it is easy to see from the update procedure that starting from zero weights and bias the ultimate solution can be written as an expansion of the form $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$. Here, the nonnegative coefficient α_i corresponds to the number of times an update has been performed for pattern $\mathbf{x}_i$.

Maximum-margin and soft-margin classifiers

For linearly separable data sets, many different linear classifiers will separate the training data perfectly. Is there a good reason to prefer one linear classifier to another? The idea in maximum-margin classification is that each training pattern should not only be correctly classified, but should also be as far away as possible from the decision boundary. For a separating hyperplane, this distance is called the 'margin' of a pattern. The maximum-margin hyperplane is the one that maximizes the minimal margin with respect to a set of training patterns. Geometrically speaking, it will pass through the 'middle' of the space between the convex hulls spanned by the positive and negative examples (see Figure 3).

This very simple idea can be a given a precise mathematical formulation. The maximum-margin hyperplane ($\mathbf{w}^*$, b^*) will fulfill

$$(\mathbf{w}^*, b^*) = \max_{\substack{\mathbf{w}, b \\ ||\mathbf{w}||=1}} \min_i (y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b)) \tag{4}$$

The above ideas can be generalized to deal with cases where the data may not be linearly separable or where a perfect separation of training patterns is not necessarily desirable. The resulting classifier is called a 'soft-margin' classifier.

Without explicitly performing the numerical optimization one can show that the optimal weight vector can be written as an expansion over training patterns, $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, with nonnegative weights α_i. Notice that the functional form is similar to the solution found by the perceptron algorithm. The coefficients α_i will be zero for patterns that are on the correct side of the hyperplane and have a margin that is strictly larger than the minimal. Patterns for which $\alpha_i > 0$ are called the 'support vectors'. Intuitively, they are the 'hard' training instances and they determine the decision boundary.

There are strong arguments from learning theory that support the use of maximum-margin or soft-margin classifiers. There exist upper bounds on

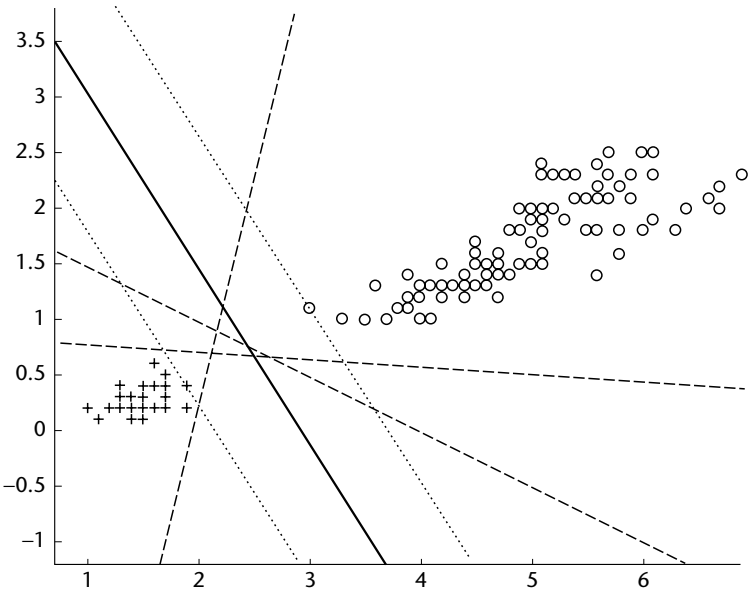


Figure 3. Different linear discriminant functions for a two-dimensional data set. All the lines separate the data without training error, but the maximum-margin hyperplane (bold line) achieves the largest separation margin. The margin is shown by dotted lines; the data points on the dotted lines are the support vectors.

the generalization performance that depend on the maximum margin achieved or the distribution of margins over training examples. Other results bound the generalization error by the sparseness of the solution, i.e. the proportion of training patterns that are support vectors. Detailed discussions can be found in Cristianini and Shawe-Taylor (2000) and Schölkopf and Smola (2002).

Logistic regression

Logistic regression is a statistical technique based on conditional likelihood or cross-entropy. The posterior probability of a class is approximated by a 'generalized linear model':

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\langle \mathbf{x}, \mathbf{w} \rangle - b)} \qquad (5)$$
$$P(y = -1|\mathbf{x}) = 1 - P(y = 1|\mathbf{x})$$

The logistic function maps an affine function of the input nonlinearly to a probability. Geometrically, the equation $\langle \mathbf{x}, \mathbf{w} \rangle + b = 0$ determines a hyperplane on which the posterior probabilities are exactly equal to ½. With increasing distance from this hyperplane, the posterior probabilities approach the extreme values (zero or one).

In logistic regression, $\mathbf{w}$ and b are determined so that the following conditional log-likelihood function is maximized:

$$(\mathbf{w}^*, b^*) = \arg\max_{\mathbf{w}, b} \sum_i \log P(y_i|\mathbf{x};\mathbf{w}, b) \qquad (6)$$

Informally, we are looking for parameter values that maximize the average posterior probability for the correct class.

The remaining numerical optimization problem can be solved, for example, by an iterative algorithm called iteratively reweighted least squares, which performs Newton–Raphson update steps.

Nonlinear Classifiers

Linear classifiers are relatively well understood, but they are severely limited in the type of classification they can achieve. Many interesting problems may not allow a good linear class separation even in an approximate sense. The standard (toy) example is the 'XOR' problem, where we have two positive patterns located in the plane at $(-1,-1)$ and (1,1) and two negative patterns at $(-1,1)$ and $(1,-1)$. No linear classifier can correctly classify these four data points. A whole generation of researchers therefore abandoned linear classifiers to develop neural networks based on the perceptron.

Neural networks

The use of neural networks for pattern recognition is inspired by biological systems. Neural networks implement nonlinear classifiers by explicitly constructing a nonlinear hypothesis class, corresponding to the architecture and topology of the network. There are several types of neural network

architectures, but the most popular is the 'multilayer perceptron' (MLP) with sigmoid transfer functions. An MLP classifier consists of a layer (or layers) of so-called hidden units, which compute new features $\phi(\mathbf{x})$ from the input representation. Each feature is parametrized by a logistic function

$$\phi_i(\mathbf{x}) = \text{logit}(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i) = \frac{1}{1 + \exp(-\langle \mathbf{w}_i, \mathbf{x} \rangle - b_i)} \quad (7)$$

Thus, each hidden unit is a logistic regression of the input pattern. Finally, an MLP has an output layer with units that compute a function of the output of the hidden units. In the case of binary classification, this is just another single logistic unit $f(\mathbf{x}) = \text{logit}(\langle \phi(\mathbf{x}), \mathbf{v} \rangle + b_0)$. Each hidden unit can be thought of as a feature extractor, and the output unit computes a logistic regression over these new features.

There is no (known) closed-form solution for how to choose the weights of an MLP optimally – for example, to maximize the conditional likelihood function used in logistic regression. In fact, one has to deal with a very high-dimensional nonlinear optimization problem. Standard training procedures are based on gradient information which can be computed efficiently using the 'back propagation' algorithm. This is basically a way to compute the gradient of the error with respect to all weights in the neural network by making use of the chain rule. It propagates error terms from the output layer back to the hidden layers. Of course, gradient methods are only suitable for finding local optima, and this is usually all one can hope for. (*See* **Backpropagation**)

There are a number of issues that have to be addressed in order to utilize neural networks in practical applications. The most important ones are model selection and regularization to avoid overfitting. In model selection, one has to decide on the optimal topology, i.e. the number of hidden units and the connectivity between neurons (such as the number of interconnected layers), a process that can involve significant search effort. In order to regularize the network training, one often uses techniques like weight decay or training with noise. Another popular heuristic is early stopping, whereby the training is stopped before a local optimum is reached.

Decision trees

Decision-tree classification is a technique that is very popular in artificial intelligence, and to some extent in statistics, mainly because of the interpretability of decision trees. A decision tree is a rooted tree – in the simplest and most popular case a binary tree – with a simple test function on patterns associated with each node. Starting at the root node, patterns are transmitted through the tree to a leaf node. The branch a pattern takes at an inner node depends on the outcome of the test at that node. Decision trees are analogous to the 20-question game: each node corresponds to a question and after a finite number of questions a leaf node is reached and a decision is made. Note that the questions asked subsequently depend on the outcomes of previous tests. Usually the types of tests that are performed at inner nodes are restricted to a single feature or attribute. With numerical features, an inner node will typically select one of the features along with a threshold, and transmit patterns to the left branch if the feature is below the threshold, and to the right branch if the feature is above the threshold. Decision trees effectively partition the feature space, each leaf node corresponding to one of the regions in the partition.

Most procedures to train decision trees have two ingredients: a node-purity function which measures how pure the class labels of training patterns transmitted to immediate descendants are (splits that create higher-purity subsets are preferred); and stopping and pruning heuristics which counteract overfitting by limiting the number of nodes in the decision tree.

Kernel-based classifiers and support vector machines

Kernel-based classifiers are an extension of linear classifiers. They exploit the fact that many linear classifiers can be defined purely in terms of inner products between data points. A kernel is a function $(\mathbf{x},\mathbf{y}) \mapsto K(\mathbf{x},\mathbf{y}) \in \mathbb{R}$ that corresponds to an inner product in some new feature space. A kernel can thus be thought of as the composition of two functions, a (possibly nonlinear) mapping to a new feature representation, $\phi(\mathbf{x})$, and an inner product in this new representation, $K(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. In many cases, it is possible to compute effectively a kernel function without explicit access to the embedding ϕ. Then K is an 'implicit mapping' into a new feature space.

Kernels transform the original feature vectors by implicitly mapping them into a new, typically higher-dimensional, feature representation. Linear classifiers in the high-dimensional representation will then correspond to nonlinear classification functions in the original feature space. This simple 'trick' of replacing the inner product in the original representation with an arbitrary kernel yields a

much more flexible and powerful family of classifiers. Some of the most popular kernels are polynomial kernels of degree d, $K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^d$, and 'radial basis function' kernels, $K(\mathbf{x}, \mathbf{y}) = \exp(-\| \mathbf{x} - \mathbf{y} \|^2/\sigma^2)$.

Combining kernels with the maximum-margin principle leads to a class of nonlinear classifiers known as 'support-vector machines'. Using the dual representation of the optimal weight vector, these yield solutions of the form $f(\mathbf{x}; \alpha, b) = \Sigma_i \alpha_i y_i K\langle \mathbf{x}_i, \mathbf{x} \rangle + b$. Thus, support-vector machines are linear (soft) maximum-margin classifiers in some implicitly defined feature space. Since the implicit data transformation can be nonlinear, they give rise to nonlinear discriminant functions in the original input representation.

Boosting

Boosting is a technique for combining many 'weak' classifiers (each of which might perform a little better than random guessing) to form a more powerful ensemble. We will focus on a version of boosting called AdaBoost.

In boosting, one incrementally adds weak classifiers to the classifier ensemble in what is called a 'boosting round'. In each round, a relative weight for each pattern in the training set is computed. Then a weak classifier f_t is trained from the weighted examples. This classifier then gets a relative weight α_t and is added to the ensemble. The final classifier will perform a weighted majority vote to determine the class of new test patterns: formally, $F(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$ where

$$g(\mathbf{x}) = \frac{\sum_t \alpha_t f_t(\mathbf{x})}{\sum_t \alpha_t} \tag{8}$$

and t refers to the boosting round.

In AdaBoost, both the weights for the training examples and the weights for the weak learners can be explicitly computed from simple equations. It can be shown that AdaBoost is a greedy optimization scheme to fit an 'additive' (i.e. ensemble) model based on the exponential loss $\sum_i \exp(-y_i\, g(\mathbf{x}_i))$. The negative exponent $y_i\, g(\mathbf{x}_i)$ is called the (ensemble) margin of the pattern $\mathbf{x}_i$; it measures how 'confident' the ensemble is in its vote.

SUMMARY

Statistical pattern recognition deals with the question of how learning from examples can help in the design and implementation of automatic pattern recognition systems. In the 1960s and 1970s, research focused mainly on methods derived from Bayesian decision theory, density estimation, perceptron classifiers, and nearest-neighbor classification. In the late 1980s and early 1990s, the advent of neural networks led to an increased interest in the field and many advances in building nonlinear pattern classifiers. More recently, substantial progress has been made by combining results from computational learning theory with practical algorithms such as support-vector machines and boosting algorithms.

References

Cristianini N and Shawe-Taylor J (2000) *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.

Devroye L, Györfi L and Lugosi G (1996) *A Probabilistic Theory of Pattern Recognition*. New York, NY: Springer.

Schölkopf B and Smola AJ (2002) *Learning with Kernels*. Cambridge, MA: MIT Press.

Vapnik VN (1999) *Statistical Learning Theory*. New York, NY: Wiley.

Further Reading

Bishop C (1995) *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Duda RO, Hart PE and Stork DG (2001) *Pattern Classification*, 2nd edn. New York, NY: Wiley.

Hastie T, Tibshirani R and Friedman J (2001) *The Elements of Statistical Learning*. New York, NY: Springer.

Jain AK, Duin RPW and Mao J (2000) Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1): 4–37.

Mitchell TM (1997) *Machine Learning*. New York, NY: McGraw Hill.

Ripley BD (1996) *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.

Perceptron

Intermediate article

Simon Haykin, McMaster University, Hamilton, Ontario, Canada

CONTENTS

The perceptron is a neural network designed to simulate the ability of the brain to recognize and discriminate.

INTRODUCTION

The *single-layer* perceptron is the simplest form of a neural network used for the classification of patterns said to be *linearly separable* (i.e., patterns that lie on opposite sides of a hyperplane). Basically, it consists of a single neuron with adjustable synaptic weights and bias. The algorithm used to adjust the free parameters of this neural network first appeared in a learning procedure developed by Rosenblatt in 1958 for his perceptron brain model. The generic 'perceptron' was originally defined by Rosenblatt as a set of input units called 'S-units' that are connected to a second set of units called 'A-units'. The definition of a 'single-layer perceptron' introduced in this article is different from that of Rosenblatt.

If the patterns used to train the perceptron are drawn from two linearly separable classes, then the perceptron algorithm converges and positions the decision surface in the form of a hyperplane between the two classes. The perceptron built around a single neuron is limited to performing pattern classification with only two classes (hypotheses). By expanding the output (computation) layer of the perceptron to include more than one neuron, we may correspondingly perform classification with more than two classes. However, the classes have to be linearly separable for the perceptron to work properly. The important point is that insofar as the basic theory of the perceptron as a pattern classifier is concerned, we need consider only the case of a single neuron. The extension of the theory to the case of more than one neuron is trivial.

The single neuron also forms the basis of an *adaptive filter*, a functional block that is basic to the ever-expanding subject of signal processing. The development of adaptive filtering owes much to the classic paper of Widrow and Hoff in 1960 for pioneering the so-called least-mean-square (LMS) algorithm, also known as the *delta rule*. The LMS algorithm is simple to implement yet highly effective in application. Indeed it is the workhorse of linear adaptive filtering, linear in the sense that the neuron operates in its linear mode when the weights and bias are fixed. Adaptive filters have been successfully applied in such diverse fields as radar, sonar, communications, and biomedical engineering.

Structural extensions of a single-layer perceptron lead to the development of *multi-layer perceptrons*, which distinguish themselves by the use of one or more *hidden layers* of computation nodes. Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the *error backpropagation algorithm*, which may be viewed as a generalization of the LMS algorithm.

Basically, error backpropagation learning consists of two passes through the different layers of the network: a forward pass and a backward pass. In the forward pass, an activity pattern (input vector) is applied to the sensory nodes of the network and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass, the synaptic weights of the network are all *fixed*. During the backward pass, on the other hand, the synaptic weights are all *adjusted* in accordance with an error correction rule. Specifically, the actual response of the network is subtracted from a desired (target) response to produce an error signal. This error signal is then propagated backward through the network, hence the name 'error backpropagation'. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense.

The multi-layer perceptron has three distinctive characteristics:

1. The model of each neuron in the network includes a nonlinear activation function. The important point to emphasize here is that the nonlinearity is smooth (i.e., differentiable everywhere) as opposed to the hard-limiting used in Rosenblatt's perceptron. A commonly used form of nonlinearity that satisfies this requirement is a sigmoidal nonlinearity defined by the logistic function:

$$y_j = \varphi_j(v_j) = \frac{1}{1 + \exp(-v_j)}$$

   where v_j is the induced local field (the weighted sum of all synaptic inputs plus the bias) of neuron j, $\varphi_j(v_j)$ is the activation function of neuron j, and y_j is the output of the neuron. The presence of nonlinearities is important because otherwise the input–output relation of the network could be reduced to that of a single-layer perceptron. Moreover, the use of the logistic function is biologically motivated, since it attempts to account for the refractory phase of real neurons.
2. The network contains one or more layers of hidden neurons that are not part of the input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input patterns.
3. The network exhibits a high degree of connectivity, determined by the synapses of the network. A change in the connectivity of the network requires a change in the population of synaptic connections or their weights.

It is through the combination of these characteristics, together with the ability to learn from experience through training, that the multi-layer perceptron derives its computing power. These same characteristics, however, are also responsible for the deficiencies in our present state of knowledge on the behavior of the network. First, the presence of a distributed form of nonlinearity and the high connectivity of the network make the theoretical analysis of a multi-layer perceptron difficult to undertake. Second, the use of hidden neurons makes the learning process harder to visualize. In an implicit sense, the learning process must decide which features of the input pattern should be represented by the hidden neurons. The learning process is therefore made more difficult because the search has to be conducted in a much larger space of possible functions, and a choice has to be made between alternative representations of the input pattern.

The development of the backpropagation algorithm represents a landmark in neural networks in that it provides a computationally efficient method for the training of multi-layer perceptrons. Although it cannot be claimed that the backpropagation algorithm provides an optimal solution for all solvable problems, it has put to rest the pessimism about learning in multi-layer machines that may have been inferred from the book by Minsky and Papert in 1969.

PERCEPTRON CONVERGENCE THEOREM

Figure 1 presents a signal-flow graph of the simplified version of Rosenblatt's perceptron built around a 'winner takes all' neuron, known as the *McCulloch–Pitts model* of a neuron (1943). Such a neuronal model consists of a linear combiner followed by a hard limiter (performing the signum function). The summing node of the model computes a linear combination of the inputs applied to its synapses, and also incorporates an externally applied bias. The resulting sum, that is, the induced local field, is applied to a hard limiter. Accordingly, the neuron produces an output equal to $+1$ if the hard limiter input is positive, and -1 if it is negative.

In the signal-flow graph model of Figure 1, the synaptic weights of the perceptron are denoted by $w_1, w_2, \ldots, w_m$. Correspondingly, the inputs applied to the perceptron are denoted by $x_1, x_2, \ldots, x_m$. The externally applied bias is denoted by b. The induced local field of the neuron is

$$v = \sum_{i=1}^{m} w_i x_i + b$$

The goal of the perceptron is to correctly classify the set of externally applied inputs into one of two classes C_1 or C_2. The decision rule for the classification is to assign the point represented by the inputs to class C_1 if the perceptron output y is $+1$ and to class C_2 if it is -1.

To develop insight into the behavior of a pattern classifier, it is customary to plot a map of the decision regions in the m-dimensional signal space spanned by the m input variables $x_1, x_2, \ldots, x_m$. In the simplest form of the perceptron there are

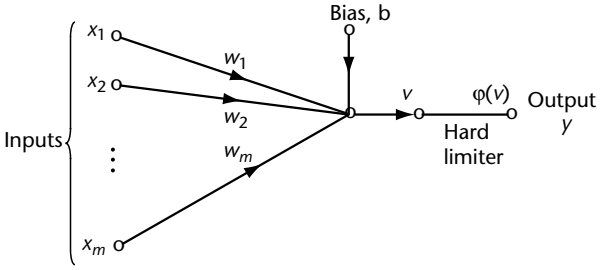


Figure 1. Signal-flow graph of the perceptron.

two decision regions separated by a hyperplane defined by

$$\sum_{i=1}^{m} w_i x_i + b = 0$$

This is illustrated in Figure 2 for the case of two input variables x_1 and x_2, for which the decision boundary takes the form of a straight line. A point (x_1, x_2) that lies above the boundary line is assigned to class C_1 and a point (x_1, x_2) that lies below the boundary line is assigned to class C_2. Note also that the effect of the bias b is merely to shift the decision boundary away from the origin.

The synaptic weights of the perceptron can be adapted on an iteration-by-iteration basis, using an error-correction rule known as the *perceptron convergence algorithm*:

1. If, at iteration n, the nth member of the training set, $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_m(n)]^{\mathrm{T}}$ where superscript T denotes transposition, is correctly classified by the weight vector $\mathbf{w}(n) = [w_1(n), w_2(n), \ldots, w_m(n)]^{\mathrm{T}}$ then no correction is made to the weight vector of the perceptron in accordance with the rule:

   $\mathbf{w}(n+1) = \mathbf{w}(n)$ *if* $\mathbf{w}^{\mathrm{T}}\mathbf{x}(n) > 0$ and $\mathbf{x}(n)$ belongs to class C_1

   $\mathbf{w}(n+1) = \mathbf{w}(n)$ *if* $\mathbf{w}^{\mathrm{T}}\mathbf{x}(n) \leq 0$ and $\mathbf{x}(n)$ belongs to class C_2

2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule:

   $\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\mathbf{x}(n)$ if $\mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n) > 0$ and $\mathbf{x}(n)$ belongs to class C_2

   $\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\mathbf{x}(n)$ if $\mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n) \leq 0$ and $\mathbf{x}(n)$ belongs to class C_1

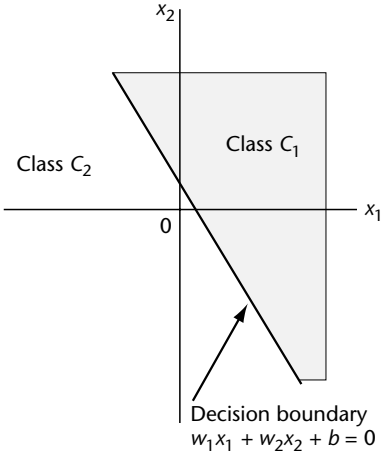


Figure 2. Illustration of the hyperplane (in this example, a straight line) as the decision boundary for a two-dimensional, two-class pattern-classification problem.

where the learning-rate parameter $\eta(n)$ controls the adjustment applied to the weight vector at iteration n.

If $\eta(n) = \eta > 0$, where η is a constant independent of the iteration number n, we have a fixed increment adaptation rule for the perceptron.

We may now state the fixed-increment convergence theorem for Rosenblatt's perceptron.

> Let the subsets of training vectors X_1 and X_2 be linearly separable, from which the inputs presented to the perceptron originate. Starting from the null initial condition $\mathbf{w}(0) = \mathbf{0}$, the perceptron converges after some n_0 iterations in the sense that
>
> $\mathbf{w}(n_0) = \mathbf{w}(n_0 + 1) = \mathbf{w}(n_0 + 2) = \ldots$
>
> is a solution vector for $n_0 \leq n_{\max}$.

The use of an initial value $\mathbf{w}(0)$ different from the null condition merely results in a decrease or increase in the number of iterations required to converge. Regardless of the value assigned to $\mathbf{w}(0)$, the perceptron is assured of convergence.

LEAST-MEAN-SQUARE ALGORITHM

Figure 3 depicts the signal-flow graph of an adaptive filter designed to deal with an unknown dynamical system. The basic difference between this model and that of Figure 1 is the substitution of a linear neuron for the McCulloch–Pitts neuron. Let $d(n)$ denote a desired response at iteration n. According to Figure 3, the corresponding error signal is

$$e(n) = d(n) - \sum_{i=1}^{m} w_i(n)x_i(n) = d(n) - \mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n)$$

The LMS algorithm is based on the use of instantaneous values for the cost function

$$E(\mathbf{w}) = e^2(n)/2$$

Differentiating $E(\mathbf{w})$ with respect to $\mathbf{w}$ yields

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}(n)} = -\mathbf{x}(n)e(n)$$

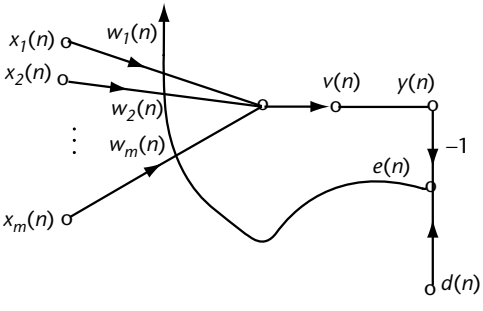


Figure 3. Signal-flow graph of adaptive model for the LMS algorithm.

Using this latter result as an estimate for the gradient vector, we may formulate the LMS algorithm as

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) - \eta\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}(n)} \\ &= \mathbf{w}(n) + \eta\,\mathbf{x}(n)e(n)\end{aligned}$$

where η is the learning-rate parameter. The feedback loop around the weight vector $\mathbf{w}(n)$ in the LMS algorithm behaves like a low-pass filter, passing the low-frequency components of the error signal and attenuating its high-frequency components. The average time constant of this filtering action is inversely proportional to the learning-rate parameter η. Hence, by assigning a small value to η, the adaptive process will progress slowly. More of the past data are then remembered by the LMS algorithm, resulting in a more accurate filtering action. In other words, the inverse of the learning-rate parameter η is a measure of the memory of the LMS algorithm.

In the LMS algorithm the weight vector $\mathbf{w}(n)$ traces a random trajectory. For this reason, the LMS algorithm is sometimes referred to as a 'stochastic gradient algorithm'. As the number of iterations in the LMS algorithm approaches infinity, $\mathbf{w}(n)$ performs a random walk (Brownian motion) about an optimum solution known as the Wiener solution. The important point is the fact that, unlike the deterministic method of steepest descent that yields the Wiener solution, the LMS algorithm does *not* require knowledge of the statistics of the environment.

BACKPROPAGATION ALGORITHM

Figure 4 shows the architectural graph of a multi-layer perceptron with two hidden layers and an output layer. To set the stage for a description of the multi-layer perceptron in its general form, the network shown here is fully connected. This means that a neuron in any layer of the network is connected to all the nodes/neurons in the previous layer. Signal flow through the network progresses in a forward direction, from left to right and on a layer-by-layer basis.

Figure 5 depicts a portion of the multi-layer perceptron, where two kinds of signals are identified:

1. *Function (input) signals*. A function signal is an input signal, a stimulus that comes in at the input end of the network, propagates forward (neuron by neuron) through the network, and emerges at the output end of the network as an output signal. We refer to such a signal as a 'function signal' for two reasons. First, it is presumed to perform a useful function at the output of the network. Second, at each neuron of the network through which a function signal passes, the signal is calculated as a function of the inputs and associated weights applied to that neuron.
2. *Error signals*. An error signal originates at an output neuron of the network, and propagates backward (layer by layer) through the network. We refer to it as an 'error signal' because its computation by every neuron of the network involves an error-dependent function in one form or another.

The output neurons constitute the output layer of the network. The remaining neurons constitute hidden layers of the network. Thus the hidden units are not part of the output or input of the

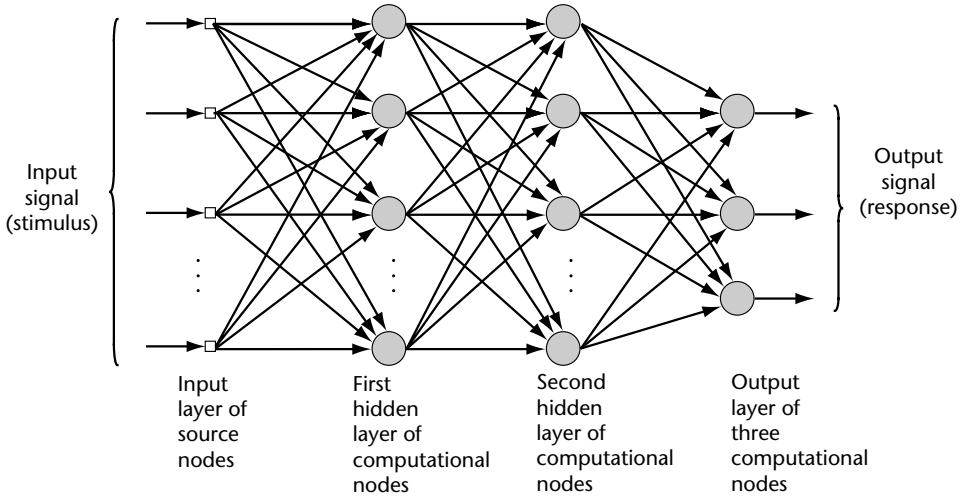


Figure 4. Architectural graph of a multi-layer perceptron with two hidden layers.

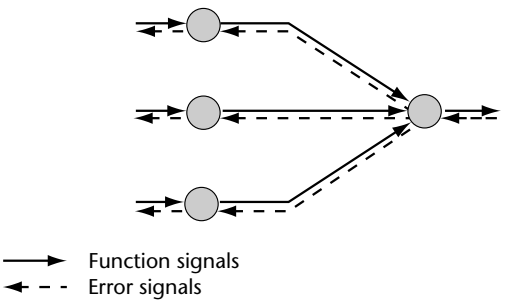


Figure 5. Illustration of the directions of function- and error-signal flows in a multi-layer perceptron.

network – hence their designation as 'hidden'. The first hidden layer is fed from the input layer made up of sensory units (source nodes); the resulting outputs of the first hidden layer are in turn applied to the next hidden layer; and so on for the rest of the network.

Each hidden or output neuron of a multi-layer perceptron is designed to perform two computations:

1. The computation of the function signal appearing at the output of a neuron, which is expressed as a continuous nonlinear function of the input signal and synaptic weights associated with that neuron.
2. The computation of an estimate of the gradient vector (i.e., the gradients of the error surface with respect to the weights connected to the inputs of a neuron), which is needed for the backward pass through the network.

The error signal at the output of neuron j at iteration n (i.e., presentation of the nth training example) is defined by:

$$e_j(n) = d_j(n) - y_j(n), \quad \text{neuron } j \text{ is an output node}$$

We define the instantaneous value of the error energy for neuron j as $[e_j^2(n)]/2$. Correspondingly, the instantaneous value $E(n)$ of the total error energy is obtained by summing $[e_j^2(n)]/2$ over all neurons in the output layer; these are the only 'visible' neurons for which error signals can be calculated directly. We may thus write

$$E(n) = \frac{1}{2}\sum_{j=C} e_j^2(n)$$

where the set C includes all the neurons in the output layer of the network. Let N denote the total number of patterns (examples) contained in the training set. The *average squared error energy* is obtained by summing $E(n)$ over all n and then normalizing with respect to the set size N, as shown by

$$E_{\text{av}} = \frac{1}{N}\sum_{n=1}^{N} E(n)$$

The instantaneous error energy $E(n)$, and therefore the average error energy, E_{av}, is a function of all the free parameters (i.e., synaptic weights and bias levels) of the network. For a given training set, E_{av} represents the *cost function* as a measure of learning performance. The objective of the learning process is to adjust the free parameters of the network to minimize E_{av}. To do this minimization, we use an approximation similar in rationale to that used for the derivation of the LMS algorithm. Specifically, we consider a simple method of training in which the weights are updated on a pattern-by-pattern basis until one epoch, that is, one complete presentation of the entire training set, has been dealt with. The adjustments to the weights are made in accordance with the respective errors computed for *each* pattern presented to the network. The arithmetic average of these individual weight changes over the training set is therefore an estimate of the true change that would result from modifying the weights based on minimizing the cost function E_{av} over the entire training set.

Summarizing the relations that govern the operation of the back propagation algorithm, we may say the following. The correction $\Delta w_{ji}(n)$ applied to the synaptic weight connecting neuron i to neuron j is defined by the delta rule:

$$\begin{pmatrix} \text{weight} \\ \text{correction,} \\ \Delta w_{ji}(n) \end{pmatrix} = \begin{pmatrix} \text{learning-} \\ \text{rate parameter,} \\ \eta \end{pmatrix} \cdot \begin{pmatrix} \text{local} \\ \text{gradient,} \\ \delta_j(n) \end{pmatrix} \cdot \begin{pmatrix} \text{input signal} \\ \text{of neuron } j, \\ y_i(n) \end{pmatrix}$$

The local gradient $\delta_j(n)$ depends on whether neuron j is an output node or a hidden node:

1. If neuron j is an output node, $\delta_j(n)$ equals the product of the derivative $\varphi_j'(v(n))$ and the error signal $e_j(n)$, both of which are associated with neuron j. Here, the prime denotes differentiation of the activation function $\varphi_j(v_j\,(n))$ with respect to the induced local field $v_j(n)$ at iteration n.
2. If neuron j is a hidden node, $\delta_j(n)$ equals the product of the associated derivative $\varphi'_j(v_j\,(n))$ and the weighted sum of the δs computed for the neurons in the adjacent hidden or output layer that are connected to neuron j.

As remarked previously, in the application of the backpropagation algorithm, two distinct passes of computation are distinguished: the forward pass, and the backward pass. In the *forward pass* the

synaptic weights remain unaltered throughout the network, and the function signals of the network are computed on a neuron-by-neuron basis. The function signal appearing at the output of neuron j is computed as

$$y_j(n) = \varphi(v_j(n))$$

where

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n)$$

where m is the total number of inputs (excluding the bias) applied to neuron j, $w_{ji}(n)$ is the synaptic weight connecting neuron i to neuron j, and $y_i(n)$ is the input signal of neuron j or equivalently, the function signal appearing at the output of neuron i. If neuron j is in the first hidden layer of the network, $m = m_0$ and the index i refers to the ith input terminal of the network, for which we write

$$y_j(n) = x_i(n)$$

where $x_i(n)$ is the ith element of the input vector (pattern). If, on the other hand, neuron j is in the output layer of the network, $m = m_L$ and the index j refers to the jth output terminal of the network, for which we write

$$y_j(n) = o_j(n)$$

where $o_j(n)$ is the jth element of the output vector $\mathbf{x}(n)$. This output is compared with the desired response $d_j(n)$, obtaining the error signal $e_j(n)$ for the jth output neuron. Thus the forward phase of computation begins at the first hidden layer by presenting it with the input vector, and terminates at the output layer by computing the error signal for each neuron of this layer.

The backward pass, on the other hand, starts at the output layer by passing the error signals leftward through the network, layer by layer, and recursively computing the local gradient δ for each neuron. This recursive process permits the synaptic weights of the network to undergo changes in accordance with the delta rule. For a neuron located in the output layer, the δ is simply equal to the error signal of that neuron multiplied by the first derivative of its nonlinearity. The recursive computation is continued, layer by layer, by propagating the changes to all synaptic weights in the network.

Note that for the presentation of each training example, the input pattern is fixed ('clamped') throughout the round-trip process, encompassing the forward pass followed by the backward pass.

DESIGN PRINCIPLES FOR MULTI-LAYER PERCEPTRONS

A particular form of the multi-layer perceptron known as *convolutional network* is designed specifically to recognize two-dimensional shapes with a high degree of invariance to translation, scaling, skewing, and other forms of distortion. This difficult task is learned in a supervised manner by means of a network whose structure includes the following forms of constraints:

1. *Feature extraction*. Each neuron takes its synaptic inputs from a local receptive field in the previous layer, thereby forcing it to extract local features. Once a feature has been extracted, its exact location becomes less important so long as its position relative to the features is approximately preserved.
2. *Feature mapping*. Each computational layer of the network is composed of multiple feature maps, with each feature map being in the form of a plane within which the individual neurons are constrained to share the same set of synaptic weights. This second form of structural constraint has the following beneficial effects:
   - *Shift variance*, forced into the operation of a feature map through the use of convolution with a kernel of small size, followed by a sigmoid (squashing) function.
   - *Reduction in the number of free parameters*, accomplished through the use of weight sharing, which makes it possible to implement the convolutional network in parallel form; weight sharing refers to a common set of synaptic weights that is shared by all the neurons in a given layer of the network.
3. *Subsampling*. Each convolutional layer is followed by a computational layer that performs local averaging and subsampling, whereby the resolution of the feature map is reduced. This operation has the effect of reducing the sensitivity of the feature map's output to shifts and other forms of distortion.

The development of convolutional networks, as just described, is neurobiologically motivated, which goes back to the pioneering work of Hubel and Wiesel in 1962 on locally sensitive and orientation-selective neurons in the visual cortex of a cat.

The lesson to be learned from convolutional networks is two-fold. First, a multi-layer perceptron of manageable size is able to learn a complex, high-dimensional, nonlinear mapping by constraining its design through the incorporation of prior knowledge about the task at hand. Second, the synaptic weights and bias levels can be learned by cycling the simple backpropagation algorithm through the training set.

CONCLUSION

This article has presented a description of the family of neural networks known collectively as the perceptrons, which encompasses Rosenblatt's single-layer perceptron, Widrow and Hoff's single-neuron adaptive filter, and their generalization, namely, multi-layer perceptrons. The single-layer perceptron is trained by means of the perceptron convergence algorithm, the single-neuron adaptive filter is trained by the simple least-mean-square (LMS) algorithm, and the multi-layer perceptron is trained by the backpropagation algorithm, that is a generalization of the LMS algorithm. The article also described a special form of multi-layer perceptrons known as the convolutional network whose design is inspired by neurobiological considerations; a convolutional network is capable of accomplishing complex learning tasks through successive use of convolution and subsampling. Moreover, through the clever use of weight sharing, the number of adjustable weights is significantly reduced despite the use of a large number of neurons.

References

Hubel DH and Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–154.

McCulloch WS and Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **55**: 115–133.

Minsky ML and Papert SA (1969) *Perceptrons*. Cambridge, MA: MIT Press.

Rosenblatt F (1958) The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408.

Widrow B and Hoff ME Jr (1960) Adaptive switching circuits. *IRE WESCON Convention Record*, pp. 96–104.

Further Reading

Haykin S (1995) *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice Hall.

Haykin S (2002) *Adaptive Filter Theory*, 4th edn. Englewood Cliffs, NJ: Prentice Hall.

LeCun Y (1989) *Generalization and network design strategies*. Technical report CRG-TR-89-4, Department of Computer Sciences, University of Toronto, Canada.

LeCun Y and Bengio Y (1995) Convolutional networks for images, speech, and time series. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press.

Problem Solving

Intermediate article

Marsha C Lovett, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

CONTENTS

Problem solving is the analysis and transformation of information towards a specific goal. This goal may be more or less well-defined; it may be directly attainable or require some insight; and the solution may rely on a large or small body of domain-specific knowledge.

INTRODUCTION

The processes of problem solving cover a broad range of situations. Some problems are 'routine' in that the solver easily recognizes a path to a solution. Other problems require insight to work around a perplexing obstacle. There are other respects in which problems can vary. One of these is the amount of domain-specific knowledge required for solution: knowledge-rich problems come from domains, such as physics, medical diagnosis, and aeroplane flying, that tend to require much specialized knowledge; whereas knowledge-lean problems often come from everyday tasks or general-interest puzzles. Problems can also be classified according to how well-defined the problem is (Reitman, 1965). A well-defined problem clearly specifies what can be taken as 'given' and what constitutes a solution. In contrast, an ill-defined problem leaves one or both of these components at most vaguely specified.

To understand problem solving completely, one must consider all the varieties of problems that can be encountered. What do all of these problem-solving situations have in common? All problems require some process of analysis in which the problem situation is understood and internally represented. From there, the solver manipulates and transforms the situation, seeking a specific solution. In accordance with this process-oriented view, this article defines problem solving as the analysis and transformation of information towards a specific goal. A central goal of problem-solving research, then, is to specify when and how the processes of analysis and transformation are applied as people solve problems. For example, an open question is whether it is exactly the same set of processes, or variants of certain processes, that are involved in solving problems of different types.

MODELING INSIGHT AND THE DISCOVERY OF CREATIVE SOLUTIONS

In 1913, the Gestalt psychologist Wolfgang Kohler travelled to Tenerife to study the problem-solving abilities of chimpanzees. Kohler was interested in chimpanzees' potential to demonstrate insightful problem solving, because he considered insight to be a distinctively human capacity. The most famous of Kohler's observations is of the chimpanzee Sultan 'discovering' a new tool by putting two bamboo sticks together, and then using this tool to retrieve bananas otherwise out of reach (Kohler, 1925).

Kohler's research, along with that of other Gestalt psychologists including Selz, Duncker, Luchins, Maier, and Katona, marked the beginning of modern research on problem solving. Because of the Gestaltists' interest in perception, it is not surprising that they focused their problem-solving research on insight problems, where perception of the problem is critical. When solving insight problems, solvers typically experience an 'aha!' moment when they switch from their initial, 'natural' way of viewing the problem to a new way that lets them see the path to a quick solution.

The central finding of the Gestalt psychologists concerning problem solving was that people have great difficulty in changing the way they represent a problem. This difficulty manifests itself in two

commonly observed phenomena: 'functional fixedness' and 'set effects'. Functional fixedness is the tendency among solvers to view an object only in its typical role. For example, if one's goal involves building a small shelf for a candle and one is given a box of nails, a box of matches, and the candle, it is not immediately apparent that the box of nails (once emptied) could relinquish its role as a container for that of a shelf (Duncker, 1945). Similarly, set effects refer to solvers' tendency to get stuck in a 'mental rut' where a familiar problem-solving approach continues to be applied even when it is no longer appropriate (Luchins, 1942).

The body of work produced by the Gestaltists revealed many interesting problem-solving results (see Duncan, 1959, for a review). However, the approach taken was descriptive, so it did not constitute a systematic exploration of the processes of problem solving.

PROBLEM SOLVING AS SEARCH

With the cognitive revolution of the 1950s and 1960s, psychologists once again embraced problem-solving research, this time with the aim of discerning the mental constructs and processes that could explain intelligent behavior. What developed from their work was a metaphor of the mind as an information-processing system; and a natural analytical framework for theorizing about such a system was the computer. In this context, the seminal work of Newell and Simon (1972) provided a new framework for analyzing the process of problem solving. The basic idea in this framework is that problem solving is equivalent to searching through a space of connected problem states, where each (directed) connection between two states represents the action that would move the solver from one state to the other. At every step in a problem, then, the solver faces the same question: which action should be chosen next?

The basic abstraction in this framework is the problem space, the set of all possible states in a problem along with the state-to-state transitions that connect them. Newell and Simon realized, however, that when a solver encounters a new problem, the entire problem space is not laid out in full. They claimed that the solver requires only three pieces of information to proceed: the initial state, the goal state, and the set of possible operators (actions along with conditions for their applicability). At each step, the solver may face a choice among many possible operators; so there is a need for a mechanism to resolve such conflicts. Choice of the next operator must often be made without domain-specific knowledge, so Newell and Simon developed heuristics for making good choices without such knowledge. They implemented their ideas in a computer program called the 'general problem solver' (GPS), which highlighted the domain-independent structure of problem solving and was able to solve problems requiring different kinds of specialized knowledge (e.g., logical proofs, trigonometric identities, formal integration, and sentence parsing).

PROBLEM SPACES, IMPASSES AND GOAL-DIRECTED SEARCH

Following the success of Newell and Simon's approach, more sophisticated computational models of problem solving were developed in the 1970s and 1980s. Most of these models were designed as 'production systems', in which knowledge for performing a task is represented in the form of 'production rules'. A production rule specifies the conditions of applicability and actions for each operator described in a task's instructions, and takes the form 'If conditions apply, then execute actions.' Production systems typically distinguish between long-term memory, represented in terms of production rules, and working memory, represented by the goals and facts that specify the solver's current mental state. For example, when starting a problem, working memory would include the goal to solve the problem and any facts specified in the problem statement. Each cycle of processing then involves the following steps. First, all of the production rules are matched against the contents of working memory (in particular, the solver's current goal) in order to determine which rules' conditions are met; second, a choice is made among the matching rules to determine which rule or rules will fire; third, the actions of the firing rules are executed. (The simplest case involves selecting a single production rule from the matching set and executing its actions.) Because a production rule's actions tend to change the contents of working memory (through a change to the solver's physical environment or mental state), a different set of production rules will be in the next matching set, and so progress is made in solving the problem.

A variety of production system models were developed to solve problems across a wide range of domains. Nevertheless, it was argued that more progress in understanding human problem solving could be made by developing these domain-specific models under the framework of overarching cognitive architectures. A cognitive architecture defines a specific way of representing knowledge and a

fixed set of mechanisms for processing and acquiring knowledge, just as the brain presumably employs a common set of mechanisms across a variety of tasks. Examples of cognitive architectures, each of which encompasses a family of related production system models, are ACT-R (Anderson and Lebiere, 1998), EPAM (Richman *et al*., 1996), and Soar (Newell, 1990).

One of the mechanisms that a complete cognitive architecture must specify is that of knowledge acquisition. In Soar, there is a single knowledge acquisition mechanism, called chunking. Whenever Soar cannot decide on the next step to take because of a gap in existing knowledge, an 'impasse' is said to have occurred. At this point, a sub-goal is initiated to determine the next step based on general problem-solving heuristics. When this sub-goal is achieved, the chunking process creates a new production rule with its action side being the next step and its conditions side being the set of pre-impasse problem features that were used in determining this next step. Thus, chunking is a way of 'caching' the many steps that were involved in deciding the next step into a single unit of knowledge that is available for future use. The next time Soar encounters the same (or a similar) situation, it will not face an impasse, but can employ the new production rule and take the next step directly.

In ACT-R, there are multiple knowledge acquisition mechanisms, mainly because ACT-R posits two separate long-term memory systems: procedural memory, which represents knowledge of skills in terms of production rules, and declarative memory, which represents knowledge of facts in terms of chunks (not to be confused with the 'chunking' described above). The most important of these mechanisms, called 'compilation', is engaged when no existing production rule is available to be applied next. The compilation mechanism begins by finding a problem in declarative memory whose features are similar to those of the current problem. It proceeds to build an analogy between the two problems, maps the solution step from the past problem to a viable step in the current problem, and constructs a production rule corresponding to that step. The condition side of the new production rule takes into account the specific facts that arose in the two problems, but generalizes them to a certain degree, making the new production rule more broadly applicable.

Both chunking and compilation have been proposed as explanations of the power-law increase in speed observed as people gain practice in solving problems in a domain.

MODEL-BASED REASONING AND PROBLEM SOLVING

While rule-based processing offers a good, general mechanism for problem solving, there may be particular situations in which people choose to invoke other processes. For example, in solving categorical syllogisms, the problem is to determine what implications may be validly derived from a set of logical statements, such as 'all artists are beekeepers; some beekeepers are clever'. Johnson-Laird and Steedman (1978) proposed a theory of reasoning through these problems that did not involve explicit rule-based processing, but rather the generation and evaluation of informal mental models. For example, one participant in their study said: 'I thought of all the little … artists in the room and imagined they all had beekeeper's hats on'. This kind of mental image can be inspected in order to generate new statements about the situation (e.g. 'some of the artists are clever' would be implied by a mental model in which only artist-beekeepers were imagined and some of these were mentally labeled as clever). Such a process does not make any use of formal rules of logical implication. Nevertheless, a production-rule-based model of this process has been developed with the cognitive architecture Soar.

Johnson-Laird and his colleagues found that different solvers choose different (more or less appropriate) mental models to represent a given problem. This choice, in turn, affects the correctness of their conclusions. Moreover, some syllogism problems require multiple mental models in order to adequately represent a possible ambiguity in the original premises. (For example, which of the beekeepers – all artists, all non-artists, or some of both – are the clever ones?) Thus, mental model theory proposes a metric for problem difficulty in terms of the number of mental models required. Empirical research has indeed found that solvers are quick and accurate in solving single-model syllogisms but that they produce more errors as the complexity of the model is increased. Additionally, syllogisms with high imagery value (e.g. those involving artists and painters) are more accurately and easily solved than syllogisms with low imagery value (e.g. those involving artists and 'clever people').

SOLVING PROBLEMS BY ANALOGY

In analogical problem solving, the solution to the current problem (the target) is determined by a past problem-solving episode stored in memory (the

source). There are three main processes that occur during analogical problem solving: (1) noticing the potentially analogous relationship between the target and a particular source; (2) mapping the correspondences between the basic elements of the target and the source; and (3) applying a solution to the target based on that mapping, possibly with some adaptation.

The main empirical result regarding the process of noticing a potential source is that solvers often miss opportunities to do it. That is, when solvers are given a helpful source problem in the same experimental context as the target, the vast majority do not even consider making an analogy (Gick and Holyoak, 1980). Two factors that facilitate this process are: giving solvers a hint to think back to the previous problem (Gick and Holyoak, 1980); and making the source and target problems superficially similar (Holyoak and Koh, 1987). The main factor determining success in the mapping process is structural similarity between the source and target (Gentner and Gentner, 1983), although certain kinds of superficial similarity can play a role here as well (Ross, 1989).

These empirical results have motivated the development of many models of analogical problem solving. Earlier models focused on computational methods for mapping the source to the target (e.g. Gentner, 1983). More recent models have successfully covered the entire process, from selecting a source to solving the target (e.g. Hummel and Holyoak, 1997).

CONNECTIONIST APPROACHES TO PROBLEM SOLVING

The connectionist framework represents knowledge in the form of continuously valued weights connecting nodes in a prespecified network. In connectionist models the main quantity of interest is the activation level of each node; this activation spreads through the network from an input layer, through one or more 'hidden' layers, and eventually to the output layer. The amount of activation propagated along a connection is proportional to the strength of that connection, and the amount of activation at a given node is the sum of the propagated activation along all connections going to that node. Because of this spreading activation mechanism, connectionist models very naturally represent information flow. Nevertheless, until recently these models were mainly applied to lower-level cognitive processes such as attention and perception. (*See* **Connectionism**)

Connectionist networks usually employ distributed rather than symbolic representations of information, so individual nodes in the network tend to represent individual features of a problem, not entire objects. For example, the input layer of a network often represents the features of a problem's initial state, and this information is processed through the subsequent layers until an 'answer' is obtained at the output layer. Connectionist models very naturally encode the similarity between problems in terms of the degree of overlap between sets of features. Not surprisingly, therefore, they have been most extensively applied to analogical problem solving, in which the retrieval of a similar past example is important (see, Holyoak and Barnden, 1994, for a collection of papers).

Connectionist models also specify a learning mechanism for updating the connection weights based on experience. This experience often comes in the form of 'teacher input', which essentially tells the model the correct answer for a given set of input features. The weights are gradually updated, based on repeated feedback of this kind, until the model can automatically generate the correct output pattern in response to a set of input features. Then, because of the model's distributed representations, it is able to produce similar answers to similarly represented inputs, i.e., to solve problems that are not a part of its direct experience. In addition, this mechanism of updating weights allows connectionist models to develop higher-level features in the intermediate layers of the network. These new features can be especially helpful for problems where the input layer's representation is insufficient for generating a solution and hence requires the solver to represent the task in a new way. The 'hidden' nature of these intermediate features is suggestive of the nonconscious aspects of representational change exhibited by solvers (see next section).

COMBINATORIC SEARCH VERSUS PATTERN RECOGNITION

Spurred on by the increasing sophistication of computational models and complexity of the tasks being modeled, empirical problem-solving research in the 1970s and 1980s moved towards a greater emphasis on learning processes. Comparisons between experts and novices were a natural starting point in the study of learning because experts and novices lie at the extreme ends of the learning process (see (Chi *et al*., 1988) for a collection of articles and (Riemann and Chi, 1989) for a

review of the expert–novice literature). Many studies observed experts and novices solving the same problems in a particular task domain, such as chess, physics, or bridge. The kinds of differences that emerged were fairly independent of the domain: experts tend to exhibit qualitatively different problem solving methods from novices. For example, when physics experts approach a problem, they look for a familiar pattern in the problem that determines the relevant physical principle and, from there, take a direct solution approach. Physics novices, on the other hand, tend to conduct a deliberate search over the various physics equations they know, trying one equation for a while, backing up and trying another, and so on.

The expert behavior described by these studies has been explained in terms of a pattern recognition process, whereby a particular pattern of features in the current problem triggers a relevant solution template that is instantiated by the current situation, thereby producing a solution. To gather more data regarding the complex patterns that experts use in solving problems, researchers used additional paradigms that do not ask participants to solve problems at all. For example, experts and novices would be given tasks relating to memory, problem classification, and similarity rating, where the stimuli were generated from problem-solving situations in the experts' domain. The chief result of the memory studies was that experts significantly outperform novices in their recall of realistic problem stimuli, but are indistinguishable from novices in their recall of random problem stimuli. Such findings have been conducted in the domains of chess (Chase and Simon, 1973), bridge (Charness, 1979; Engle and Bukstel, 1978), and medical diagnosis (Patel *et al*., 1986).

More evidence of schema-based representations comes from categorization paradigms. Expert and novice solvers are asked to sort a set of problem descriptions into meaningful groups. Whereas novices tend to sort based on superficial features, experts tend to do so in terms of the problems' solution structures. For example, Chi *et al*. (1981) found that experts sorted problems based on the physical principle that would be invoked in a solution (e.g. Newton's second law), and novices sorted more on the basis of superficial features (e.g. whether the problem involved inclined planes or pulleys). This difference suggests that experts were categorizing problems according to their problem-solving schemas, but novices were not. These findings have been replicated in mathematics (Silver, 1979; Schoenfeld and Herrmann, 1982) and computer programming (Weiser and Sherta, 1983).

SUMMARY

Early research by Gestalt psychologists stimulated interest in problem solving. A substantial development in the theory of problem solving came with Newell and Simon's (1972) framework, which considers problem solving as search through a problem space. Under this framework, choice of the next problem-solving step is fundamental.

Problem-solving research has made much theoretical progress by developing and testing computational models of the processes of problem solving. Much of this work is in the form of production system models, in which skills are represented as production rules and choice of the next problem-solving step is implemented as a choice among matching production rules. Specific kinds of problem solving that have received attention are model-based reasoning and analogical problem solving. Although the processes specified for these two modes of problem solving differ from each other and from the earlier work, they too need to address issues of choice; for example, which mental model to use to represent the syllogism, and, which source problem to draw an analogy from.

Recently, the focus of research has shifted from processes of choice to processes of learning. New empirical research on expert problem-solving behavior and the novice–expert transition shows qualitative differences. This raises the challenge for future models to achieve a qualitative change within learning theories that emphasize the role of gradual, quantitative processes.

References

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Charness N (1979) Components of skill in bridge. *Canadian Journal of Psychology* **33**: 1–16.

Chase WG and Simon HA (1973) Perception in chess. *Cognitive Psychology* **4**: 55–81.

Chi MTH, Feltovich PJ and Glaser R (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science* **5**: 121–152.

Chi MTH, Glaser R and Farr MJ (1988) *The Nature of Expertise*. Hillsdale, NJ: Erlbaum.

Duncan CP (1959) Recent research on human problem solving. *Psychological Bulletin* **56**: 397–429.

Duncker K (1945) On problem-solving. *Psychological Monographs* **58** (5): 113.

Engle RW and Bukstel L (1978) Memory processes among bridge players of differing expertise. *American Journal of Psychology* **91**: 673–689.

Gentner D (1983) Structure-mapping: a theoretical framework for analogy. *Cognitive Science* **7**: 155–170.

Gentner D and Gentner DR (1983) Flowing waters or teeming crowds: mental models of electricity. In:

Gentner D and Stevens AL (eds) *Mental Models*, pp. 99–130. Hillsdale, NJ: Erlbaum.
Gick ML and Holyoak KJ (1980) Analogical problem solving. *Cognitive Psychology* **12**: 306–355.
Holyoak KJ and Barnden JA (1994) *Advances in Connectionist and Neural Computation Theory: Analogical Connections*. Norwood, NJ: Ablex.
Holyoak KJ and Koh K (1987) Surface and structural similarity in analogical transfer. *Memory and Cognition* **15**: 332–340.
Hummel JE and Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* **104**: 427–466.
Johnson-Laird PN and Steedman M (1978) The psychology of syllogisms. *Cognitive Psychology* **10**: 64–99.
Kohler W (1925) *The Mentality of Apes*. London: Routledge and Kegan Paul.
Luchins AS (1942) Mechanization in problem solving: the effect of Einstellung. *Psychological Monographs* **54** (6): 195.
Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.
Patel VL, Groen GJ and Frederiksen CH (1986) Differences between medical students and doctors in memory for clinical cases. *Medical Education* **20**: 3–9.
Reitman WR (1965) *Cognition and Thought: An Information-Processing Approach*. New York, NY: Wiley.
Richman HB, Gobet F, Staszewski JJ and Simon HA (1996) Perceptual and memory processes in the acquisition of expert performance: The EPAM model. In: Ericsson KA (ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games*, pp. 167–187. Mahwah, NJ: Erlbaum.
Riemann P and Chi MTH (1989) Human expertise in complex problem solving. In: Gilhooly KJ (ed.) *Human and Machine Problem Solving*. Plenum Publishers.
Ross BH (1989) Distinguishing types of superficial similarities: different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**: 456–468.

Further Reading

Greeno JG and Simon HA (1988) Problem solving and reasoning. In: Atkinson RC, Herrnstein RJ, Lindzey G and Luce RD (eds) *Stevens' Handbook of Experimental Psychology*, vol. II 'Learning and Cognition', 2nd edn, pp. 589–672. New York, NY: Wiley.
Holyoak KJ (1995) Problem solving. In: Smith EE and Osherson DN (eds) *Thinking: An Invitation to Cognitive Science*, vol. III, 2nd edn, pp. 267–296. Cambridge, MA: MIT Press.
Hunt E (1994) Problem solving. In: Sternberg RJ (ed.) *Thinking and Problem Solving*, pp. 215–232. San Diego, CA: Academic Press.
Lovett MC (2002). Problem solving. In: Medin DL and Pashler H (eds) *Stevens' Handbook of Experimental Psychology*, vol. II, 3rd edn. New York, NY: Wiley.
Mayer RE (1992) *Thinking, Problem Solving, Cognition*. New York, NY: Freeman.
Medin DL and Ross BH (1989) The specific character of abstract thought: categorization, problem solving, and induction. In: Sternberg J (ed.) *Advances in the Psychology of Human Intelligence*, vol. V, pp. 189–223. Hillsdale, NJ: Erlbaum.
VanLehn K (1989) Problem solving and cognitive skill acquision. In: Posner MI (ed.) *Foundations of Cognitive Science* pp. 527–579. Cambridge, MA: MIT Press.

Production Systems and Rule-based Inference

Intermediate article

Gary Jones, University of Derby, Derby, UK
Frank E Ritter, Pennsylvania State University, University Park, Pennsylvania, USA

CONTENTS

Production systems are computer programs that reason using production rules. They have been used to create expert systems and models of human behavior.

REASONING AS RULE APPLICATION

A production system is a computer simulation of one or more tasks. Normally, the tasks have some form of goal to accomplish (for example, making a medical diagnosis based on symptom data). From a given starting position in the task the production system successively applies rules, each of which transforms the current task position, until a goal is reached. With certain sets of further constraints, production systems can be used to simulate how people perform tasks (particularly those of a problem-solving nature, which are well suited to the framework of production systems). Some examples of production system frameworks that have been used to simulate human behavior are Soar (Laird *et al*., 1987), ACT-R (Anderson and Lebiere, 1998), and OPS5 (Forgy, 1981).

A production system consists of three components: a long-term memory (in the form of a rule base), a working memory, and an inference engine. The rule base contains rules, the conditions of which must be matched to elements in working memory. The inference engine determines which of the rules in the rule base have all their conditions matched to objects in working memory, and then decides which rule to apply. The application of a rule will usually cause elements in working memory to be added, removed, or altered. Further rules can then be matched by the inference engine.

What distinguishes a production system from any other system that simulates problem-solving behavior (such as a connectionist system) is its use of rules. A rule is an 'if–then' construct, where the 'if' part contains conditions that must be met in order for the rule to be considered for use. If the rule gets used ('fired' or 'applied'), the actions in the 'then' part are performed. The conditions are sometimes called the left-hand side (LHS) of the rule, and the actions the right-hand side (RHS).

Production system behavior can be explained using an example problem of moving a sphere to the exit of a room. Figure 1 shows the layout of a room as a seven-by-eight grid where black squares represent obstacles. Given this scenario there are a variety of rules that will help in moving the sphere to the exit, such as:

`IF` the exit is above the sphere `AND` there isn't an obstacle directly above the sphere `THEN` move the sphere upwards one space. (1)

There are two conditions in rule 1 that need to be met before the action part of the rule is applied: the exit position must be above the sphere, and the position directly above the sphere must be free to

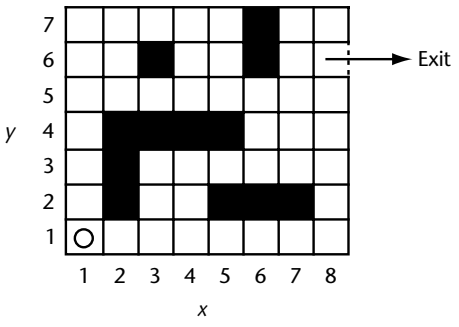


Figure 1. Problem scenario where a sphere has to be guided to the exit of a grid.

move into (i.e. it is not obstructed). If these two conditions apply to the current situation, then the sphere will be moved upwards one space.

Rules are often called productions, or production rules, because they necessarily produce something when they get used (normally changing something within working memory).

At any one time, there may be more than one rule whose conditions are satisfied. The process of collecting all applicable rules into a 'conflict set', and then selecting one of the rules in the conflict set to be fired, is called the recognize–act cycle. The production system stops either when no more rules are able to fire (e.g. ACT-R) or when the task is known to be complete (e.g. Soar).

In addition to being at the core of many expert systems, production systems are often used to simulate human behavior (e.g. Kieras and Polson, 1985; Klahr *et al*., 1987; Newell, 1990; Newell and Simon, 1972; Young, 1976). By successfully simulating human behavior on a task, the production system suggests the processes that may be used when humans perform the task. Although some researchers believe that human thinking need not be rule-based (e.g. Rumelhart *et al*., 1986; Brooks, 1991), there are many (such as Langley, Anderson, and Newell) who believe that it is, or at least can be, fruitfully viewed that way. The question revolves around whether or not human thinking deals with symbols. Rule-based systems use symbols, whereas others, such as connectionist systems, do not. The 'symbolic versus subsymbolic' question is not covered here, but is still widely debated within the cognitive science community.

COMPONENTS AND MECHANISMS OF PRODUCTION SYSTEMS

Two of the three components of a production system are distinct types of memory: working memory (facts) and long-term memory (the rule base). The third component is the inference engine.

Working memory usually contains facts about the world that are relevant to the task the production system is performing. Working memory is usually represented by attribute–value pairs. One element in working memory can have several attribute–value pairs. Using the sphere example, one element (or fact) in working memory will be the position of the exit of the grid. This element might have two attribute–value pairs: the first attribute will be 'x-coordinate', having the value '8', and the second attribute will be 'y-coordinate' having the value '6'.

Long-term memory usually consists of rules that govern the behavior of the production system. The inference engine combines working memory and long-term memory by finding rules whose conditions are matched by elements in working memory. When this happens, the rule can fire.

Production systems work in a cycle of production firings (the recognize–act cycle). Normally, only one production is fired on each cycle. The action of the production often changes what is in working memory and thus enables another production rule to fire. The resulting change may in turn enable a further production rule to fire. This is how the production system produces behavior (for example, maneuvering the sphere to the exit). When no further production rules can be fired, the system halts.

Let us work through an example illustrating this cycle and showing how production application works. Given the problem scenario in Figure 1, we noted that rule 1 matched working memory and could be applied. In that rule, there is a condition that specifies that the exit should be above the sphere's current position. If we did not already know that this was the case, then we would need a rule to work out whether or not the exit was above the sphere. The following rule could accomplish this:

`IF` the y-coordinate of the exit is greater than the y-coordinate of the sphere `THEN` put in working memory that the exit is above the sphere. (2)

Similarly we could have a rule to determine whether there was an obstacle directly above the sphere:

`IF` there is no obstacle having a y-coordinate of 1 less than the y-coordinate of the sphere `THEN` put in working memory that there isn't an obstacle directly above the sphere. (3)

Now suppose the starting position of the room is as shown in Figure 1. Certain elements would be in working memory, for example, the position of the sphere, the position of the exit, and the positions of obstacles.

Rules 2 and 3 can now be matched. The y-coordinate of the exit can be compared with the y-coordinate of the sphere; and, as it is in fact greater, rule 2 can fire. This places the element 'exit is above the sphere' in working memory. Rule 3 can also fire: all the obstacles can be checked as to whether they are directly above the sphere; as none are, the element 'no obstacle directly above

sphere' is placed in working memory. Note that these two rules could have fired at the same time, but most production systems would fire them in sequence (i.e. one recognize–act cycle for each). We will see later, when considering conflict resolution, how the production system determines which rule to fire first.

After these two new elements have been placed in working memory, rule 1 can fire, which moves the sphere upwards one space. Figure 2 shows how working memory changes during the three recognize–act cycles.

INFERENCE THROUGH FORWARD AND BACKWARD CHAINING

The normal way in which a production system processes rules is by matching elements in working memory in the 'if' part of the rule, and then applying the actions in the 'then' part of the rule. This is how the production system in the example above has cycled.

The production system can be seen as working forwards from the starting position (e.g. the problem position in Figure 1) towards a finishing position (the goal), in a series (a chain) of rule applications. This is called forward chaining or data-driven reasoning. Forward chaining is used in the majority of production system domains, including cognitive models of problem solving (an instructive example is the Tower of Hanoi (Anzai and Simon, 1979)).

The opposite of forward chaining is to match the actions of rules to working memory, and if all can be matched, to apply the conditional part of the rule. This method works backwards from the goal to the starting position (i.e. the initial data) and is therefore known as backward chaining, or goal-driven

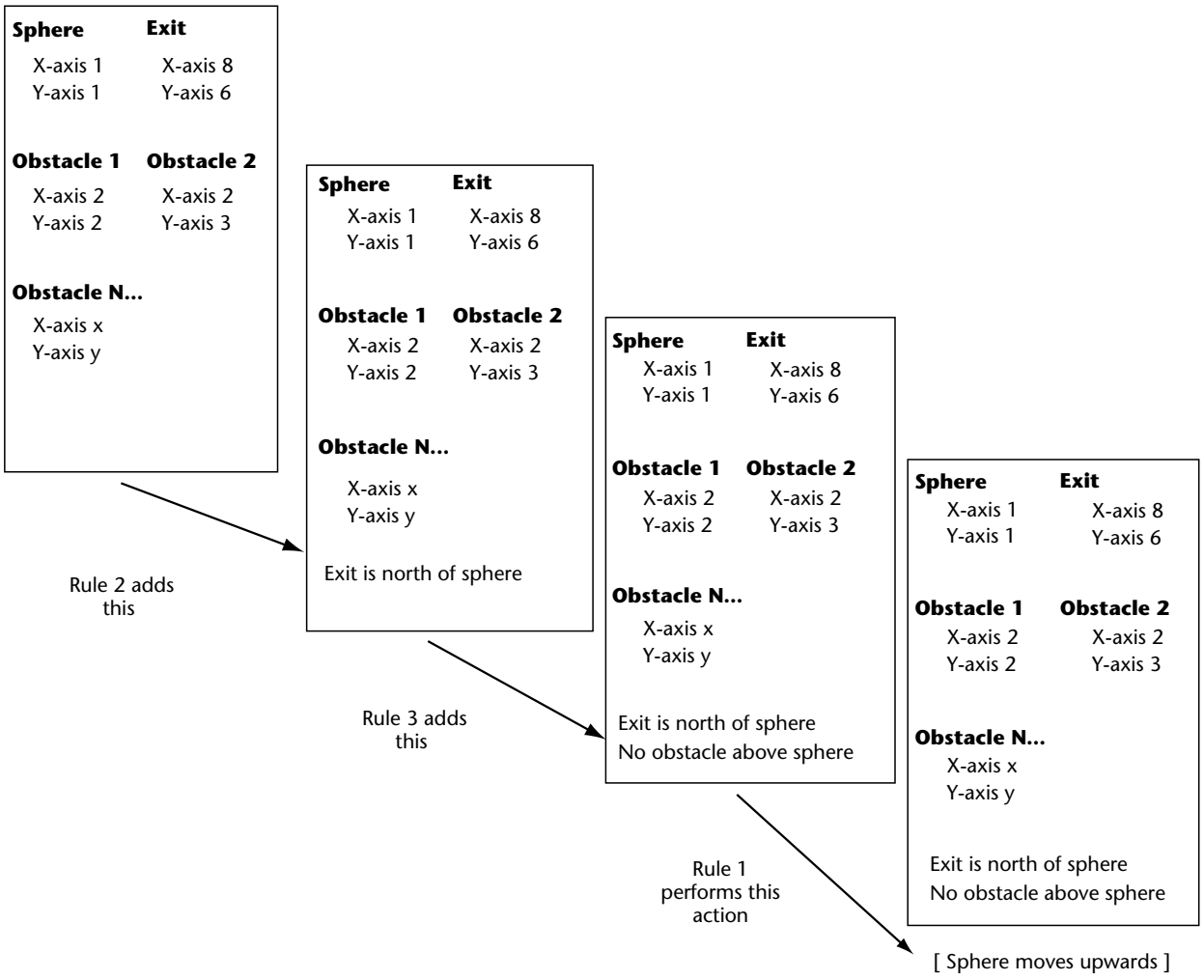


Figure 2. How working memory changes as the sphere production system rules fire.

reasoning. Backward chaining is normally used in domains where you wish to reason backwards from the final state of the world, such as discovering what preconditions led to a patient showing their current set of symptoms. The 'actions' of rules in these cases are often either changes to working memory or questions (to the user of the system) which need to be answered in order to add more facts into working memory. Expert systems often incorporate backward chaining so that they are able to explain how they arrived at their decisions.

Rules and knowledge can be represented at multiple levels. For example, Able (Larkin *et al*., 1980) represents how learning in formal domains such as physics shifts novices from a backward chaining approach of trying to find the values of variables in problem solving, to a forward chaining approach of experts where unknown variables are simply and directly derived from known variables. This model has been implemented in Soar (Ritter *et al*., 1998), which is normally seen as a forward chaining system because it applies its production rules towards a goal. In this case, the domain knowldge rules are implemented as sets of Soar rules. The domain knowledge is initially applied in a backward chaining way, searching from the target variables of the physics problem back towards the givens. With learning, the direction of processing on the knowledge level reverses.

CONFLICT RESOLUTION AND PARALLEL PRODUCTION FIRING

There are occasions when more than one rule can be fired for a given set of working memory elements. In our example above, rules 2 and 3 could both be matched. Figure 3 shows how these two rules could be matched for the elements in working memory that we started with in the example. The general approach in production systems is for the matching process to occur in parallel but the fining process to occur in serial. In the matching process, the inference engine determines which rules have all their conditions matched by elements in working memory. The resulting set of all rules that can be applied is called the conflict set.

When it is possible to fire more than one rule for a given situation, the production system is said to be in conflict. Firing all of the rules in the conflict set in parallel can give rise to inconsistent knowledge and results. Production systems generally resolve this problem by selecting only one of the rules to fire. The selection is made by the inference engine, using 'conflict resolution'. Production systems have used a variety of approaches including:

Textual order. This is the simplest resolution of conflict: simply choose the rule that comes first in the rule base.

Refractoriness. The same rule cannot be applied to the same working memory elements more than once. The inference engine needs to keep track of when elements in working memory were added or changed, in order to calculate whether a rule is being applied on exactly the same elements of working memory or whether there has been a change to those elements.

Recency. Apply the rule whose conditions match the most recently added working memory elements. This technique encourages adaptivity.

Specificity. Choose the rule that is either the most specific (i.e. has the largest number of conditions) or is the least specific (i.e. has the smallest number of conditions). Which of these is chosen is dependent upon the type of domain that is being

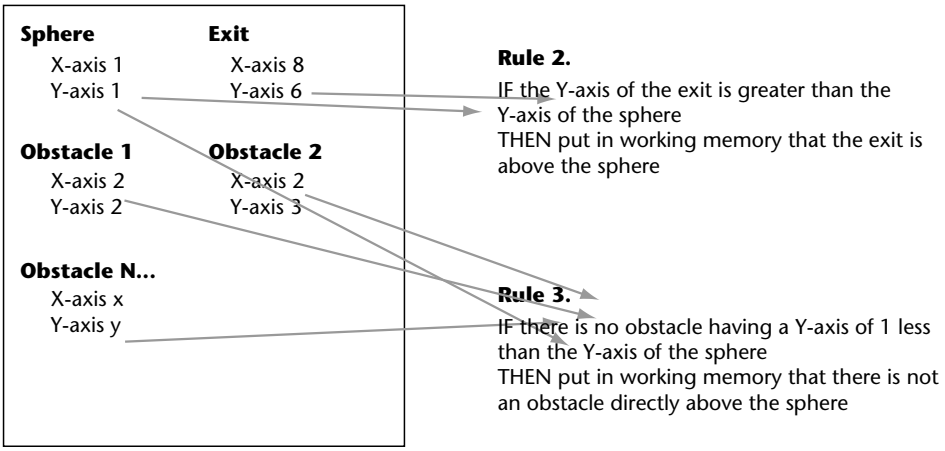


Figure 3. How working memory elements can be matched to conditions in rules.

modeled. For example, in a medical domain we may wish to be as certain as we can when applying a rule, and so we may set the most specific rule to be applied; in a domain involving search, we may wish to be less specific so that the system is less likely to arrive at a dead end.

Saliency. This allows the person writing the production system to set (numerically) how important each rule is. The rule with the highest saliency is selected. Conflict resolution is therefore primarily the responsibility of the production system designer.

Meta-rules. This allows the set of rules that are in conflict to be pruned or reordered based on a higher-order rule. For example, if the conflict set contains two rules, one which mentions high blood pressure leading to a possible heart condition and the other mentioning high blood pressure leading to a possible high temperature, then the production system designer could add a meta-rule which stated that if these two rules are in conflict, then choose the former over the latter.

Although these techniques are usually used independently of each other, there are occasions when more than one needs to be used. For example, it would be possible, when using refractoriness, to be unable to select a unique rule (for example, at the beginning of a production system run). Often, more than one type of conflict resolution is used, but one type is given priority.

Soar has taken a different approach. All matched rules are allowed to fire in parallel, but they are not allowed to modify working memory directly. They provide suggestions for changes to working memory, and a preference calculus is used to resolve contradictions and implement the changes.

STRUCTURE OF PRODUCTIONS

Productions can be represented in numerous ways. They can be represented as plain sentences (like rules 1–3 above); they can be represented in a structured editor as objects; or they can be represented as a list of clauses, which is how Soar and ACT-R represent them. The code sample below shows how rule 2 could be represented in ACT-R:

```
(p check-if-exit-is-above-sphere
  =goal>
    ISA           move-sphere-to-exit
    to-move       =sphere
    exit          =exit
  =exit>
    ISA           Exit
    Y-coordinate  =exit-y-coordinate
  =sphere>
    ISA           Sphere
    Y-coordinate  =sphere-y-coordinate
  =greater-than-fact>
    ISA           Greater-than-fact
    big-num       =exit-y-coordinate
    small-num     =sphere-y-coordinate
 = =>
  =new-working-memory-element>
    ISA           Sphere-location-fact
    location      exit-is-above-sphere
)
```

Parentheses are used to delineate the rule, which is introduced by `p` and then a name. This rule has four conditions, each consisting of a single working memory element; these are marked by = and then a name (in ACT-R – other systems use different conventions). Each condition must be matched to elements that are present in the working memory of the system.

Every element in working memory in ACT-R has an `ISA` ('is a') attribute–value pair, which defines the working memory element type. All variables in ACT-R are preceded by =. The first condition of a rule in ACT-R is always the `goal` condition. (The goal is often part of the conditions of rules in goal-directed production systems.) The `goal` condition states that there must be a goal in working memory of the type `move-sphere-to-exit`. If this does exist, it is matched; the value of the `to-move` attribute is placed in the `=sphere` variable, and the value of the `exit` attribute is placed in the `=exit` variable. Note how these same variables are then used in the next two conditions, signifying that what is `to-move` should be a memory element of the type `Sphere`, and where it is moved to should be a memory element of the type `Exit`.

The second condition specifies a match to an `Exit` element in working memory, and if this element exists it must have a `Y-coordinate` value (this will be stored in the `=exit-y-coordinate` variable). The third condition specifies a match to a `Sphere` element in working memory, and if this element exists it must also have a `Y-coordinate` value (this will be stored in the `=sphere-y-coordinate` variable). The fourth condition specifies a match to a `Greater-than-fact` element in working memory. This denotes knowledge of which numbers are larger than others. The variables that were set in the second and third conditions are used in the fourth condition to see if `=exit-y-coordinate` is greater than `=sphere-y-coordinate`. If the rule can be fired, then a

new working memory element is created, which specifies that the exit is above the sphere.

The number of input or output clauses is not limited by this syntax. In ACT-R the variables are bound (i.e. matched) in the order in which they are written, whereas in Soar they need not be. Rules in ACT-R and some other systems also have weights associated with them. These weights are updated by learning algorithms to represent the rule's probability of success, its cost, and other attributes. The weights may be used in a variety of ways, for example, to choose the most useful rule. Soar does not include these attributes.

THE RETE ALGORITHM

Most production systems include hundreds of rules; some include thousands. One includes nearly a million rules (Doorenbos, 1994). If each rule is checked individually to see if it matches, the time taken to create the conflict set will depend linearly on the number of clauses in the whole rule set. With small rule sets, this is not a problem; but as the size of models in production systems has grown, this causes a significant bottle-neck, particularly where the clauses have to be matched against sets of objects.

The RETE algorithm (Forgy, 1982) was created as a way to speed up the matching process by taking advantage of the reuse of clauses and the fact that working memory elements change slowly. For example, consider the three rules below, based on the sphere example:

IF the exit is above the sphere's current position AND there isn't an obstacle directly above the sphere THEN move the sphere upwards one space. (4)

IF the exit is above the sphere's current position THEN move the sphere upwards one space. (5)

IF there isn't an obstacle directly above the sphere THEN move the sphere upwards one space. (6)

These rules overlap, both in their conditions and in their actions. (The example is for illustration only: it would be unwise to move the sphere upward by only checking if the exit was above it, as rule 5 does.)

In essence, the RETE algorithm creates a network representing the whole rule set. The matching is then based on changes to working memory. As elements leave or enter, the state of the network is updated. Thus the time taken by the matching process depends linearly on the number of changes to working memory instead of on the number of rules.

For example, the RETE network for rules 4–6 would combine the first clauses of rules 4 and 5 as a top node. If the exit's position changed, then and only then would the clause be updated. When the conflict set is needed, all the clauses that are matched are already noted in the network. So if the exit is above the sphere's current location, then rule 5 would be waiting in the conflict set. Whether rule 4 was in the conflict set would depend on whether the working memory element matching its second clause had been added. The addition of the working memory element would also have satisfied rule 6 at the same time.

The presence of the RETE algorithm is not always noticed by people who use production systems, but it has drastically improved their speed and therefore their usefulness. RETE is used by both OPS5 and Soar. Further refinements of match optimization have also been developed.

PRODUCTION SYSTEMS AS COGNITIVE ARCHITECTURES

A cognitive architecture proposes a theory of the human information processing apparatus. Some production systems, such as ACT-R and Soar, have been used to implement such theories. Their intention is to model the types of process and structures that generate human behavior (e.g. what the constraints on memory are), and use these models to simulate human behavior. If the same architecture can be used to accurately simulate behavior across domains, then this provides evidence that the human brain may resemble the cognitive architecture. Cognitive architectures have been used to simulate the behavior of both adults (e.g. Jones *et al.*, 1999) and children (e.g. Jones *et al.*, 2000).

SUMMARY

Productions systems have been used to organize and apply knowledge in a variety of domains. While there remain questions as to whether knowledge can be modeled directly as structures in a production system, they provide a fruitful way to think about human behavior.

References

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anzai Y and Simon HA (1979) The theory of learning by doing. *Psychological Review* **86**: 124–140.

Brooks RA (1991) Intelligence without representation. *Artificial Intelligence* **47**: 139–159.

Doorenbos RB (1994) Combining left and right unlinking for matching a large number of learned rules. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*. Menlo Park, CA: AAAI Press.

Forgy CL (1981) *OPS5 User's Manual*. [Technical Report CMU-CS-81–135, Department of Computer Science, Carnegie-Mellon University, Pittsbugh, PA.]

Forgy CL (1982) Rete: a fast algorithm for the many pattern / many object pattern match-problem. *Artificial Intelligence* **19**: 17–37.

Jones G, Ritter FE and Wood DJ (2000) Using a cognitive architecture to examine what develops. *Psychological Science* **11**: 93–100.

Jones RM, Laird JE, Nielsen PE *et al*. (1999) Automated intelligent pilots for combat flight simulation. *AI Magazine* **20**: 27–41.

Kieras D and Polson PG (1985) An approach to the formal analysis of user complexity. *International Journal of Man–Machine Studies* **22**: 365–394.

Klahr D, Langley P and Neches R (eds) (1987) *Production System Models of Learning and Development*. Cambridge, MA: MIT Press.

Laird JE, Newell A and Rosenbloom PS (1987) Soar: an architecture for general intelligence. *Artificial Intelligence* **33**: 1–64.

Larkin JH, McDermott J, Simon DP and Simon HA (1980) Models of competence in solving physics problems. *Cognitive Science* **4**: 317–345.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ritter FE, Jones RM and Baxter GD (1998) Reusable models and graphical interfaces: realising the potential of a unified theory of cognition. In: Schmid U, Krems J and Wysotzki F (eds) *Mind Modeling: A Cognitive Science Approach to Reasoning, Learning and Discovery*, pp. 83–109. Lengerich, Germany: Pabst Scientific Publishing.

Rumelhart DE, McClelland JL and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I 'Foundations'. Cambridge, MA: MIT Press.

Young RM (1976) *Seriation by Children: An Artificial Intelligence Analysis of a Piagetian Task*. Basel: Birkhauser.

Rate versus Temporal Coding Models

Introductory article

Michael N Shadlen, University of Washington, Seattle, Washington, USA

CONTENTS

Information is represented in the brain by the electrical activity of neurons. A number of current theories attempt to explain this neural code of information and how it is used by the brain to achieve perception, action, thought, and consciousness.

INTRODUCTION

All theories about how the brain functions are based on the idea that information is represented by the electrical activity of neurons. The question of how neurons represent information is therefore fundamental to all branches of neuroscience. What is the neural code of information, and how is it used by the brain to achieve perception, action, thought, and consciousness? In other words, which aspects of a neuron's electrical activity convey information about the environment and our mental states?

It is common practice in many laboratories to display the electrical activity from one or more neurons in an animal while it looks at, hears, and reacts to its environment. This neural activity appears as a sequence of very brief events, known as action potentials or spikes, separated by gaps of variable duration (Figure 1). The intervals between spikes can be as long as a few tenths of a second or shorter than a hundredth of a second. The spikes and the intervals between them convey the neuron's message. If we wish to decipher the neural code, we need to know how to read these messages.

Certain facts are well established. The only message that one neuron can give another neuron in another part of the brain about what it has computed must be represented in the sequence of spikes that are transmitted along its axon. The timescale for neural computations involved in perception, thought, and action is too short to allow gene expression, protein synthesis, and chemical cascades to play a part in carrying information. Spikes are the only items in the alphabet, but unlike letters, there is only one kind. The spikes are all-or-nothing events – size does not matter. The question is how to read this sequence of spikes emitted by neurons as a function of time.

In its broadest sense, temporal coding refers to three types of problem. First and most obvious, neurons must code information that changes as a function of time. Stimuli come and go and change their intensity; behavior – and the thought behind it – is dynamic. Insofar as spikes code information, they must code information that changes as a function of time. Second, there is a possibility that the time structure of neural activity could be used to represent information. That is, spike times and intervals could expand the alphabet that the brain uses to encode stimuli, sequences of actions, and ideas. Third, neurons must ultimately code time itself: how much has elapsed, when an expected event is likely to occur, and so on. This article

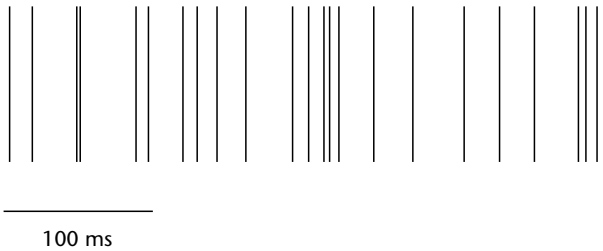


Figure 1. Neural spike trains. Neurons encode information using electrical impulses known as action potentials or spikes. A sequence of impulses is called a spike train. Information processing in the brain occurs through the interactions of neurons that communicate with one another by transmitting spikes. A sequence of spikes like this one was emitted by a neuron in the visual cortex in response to a moving light. Notice the irregular intervals between the spikes. This is a common feature of spike trains in the cerebral cortex.

examines the first two forms of temporal coding; the coding of time itself is only just beginning to be studied at the neural level.

A MENU OF TEMPORAL CODES

It makes sense to confine speculation about the neural code to neurons whose role in perception and behavior are known to at least some degree. Although the properties of neurons in much of the brain remain mysterious, the function of many neurons has been elucidated in some detail, especially in the primary sensory and motor areas of the brain. Thus far, this knowledge has rested on the fact that neurons emit more spikes in less time when a restricted set of conditions holds – such as when light of a particular orientation is present in a tiny region of the visual field and moving in a particular direction. Put simply, the neurons that we know the most about signal information by changing the rate at which they emit spikes.

The spike 'rate code' is thus well established as the vanguard neural code, and we will consider its role in temporal coding in a moment. However, there may be other ways to encode information in spike trains in addition to spike rate. In principle, these alternatives could expand the alphabet that the brain uses to encode information. Table 1 looks at the evidence for four putative neural codes: the time-varying rate code, which uses spike rate to code time-varying signals in the environment or in the state of a calculation; the spike bar code and the rate waveform code, which use time itself to evoke information about objects; and the synchrony code, which uses a pattern of spikes across many neurons to encode information. Each of these codes would promote a different scheme for encoding information in trains of spikes. We will consider each of them in turn, keeping in mind that they are not mutually exclusive ideas.

It should be stated at the outset that many ideas about temporal coding are just that: ideas. This is an active area of research that is being updated continually on the basis of new information obtained at all levels of inquiry, from the channels responsible for electrical activity in parts of the neuron, up to recordings from intact brains in behaving animals. Therefore, in addition to describing the ideas behind the putative temporal codes, this article will try to evaluate the evidence in support of each code.

The evidence for a temporal code can be divided broadly into four categories (see Table 1). First, the code can be detected reliably from neurons upon repeated exposure to the coded stimulus or upon repeated actions or circumstances. Second, the neural signal – its presence, absence, intensity or quality – is associated with variation in an animal's perception or behavior. If a neural signal putatively codes the color red, the animal should be less likely to 'report' (via its behavioral response) that it has seen red when the neural signal is absent or degraded. Third, when the signal is introduced to the brain artificially through electrical microstimulation, it causes an animal to act or perceive in accordance with the information thus encoded. Fourth, the properties of neurons – synapses, passive and active electrical properties – must be capable of preserving the coding scheme. If a proposed code were to require that action potentials occur within 1 ms of each other, we could reject the code on the basis of the fact that neurons have refractory periods of at least 1 ms before they can fire a second action potential.

Note that the first three items require experiments on the intact brain, although the first does not require the animal actually to do anything. The last item is informed by experiments that address how neurons work, namely, the reliability of their synapses and the way in which electric current is gathered by the dendrites and converted to action potentials at the initial segment of the axon.

With these guidelines in mind, let us turn to the four putative temporal codes and evaluate them based on the four categories of evidence.

Table 1. Evidence for four candidate temporal codes

	Code			
Type of evidence	*Rate*	*Spike bar*	*Waveform*	*Synchrony*
Reproducibility of response	✓	✓	✓	✓
Correlation of response with behavior	✓	✓		
Brain stimulation mimics inferred message	✓	✓	✓[a]	
Biophysical plausibility	✓	✓	✓	✓

[a]Evidence comes from one experiment in peripheral nerve responsible for taste.

Time-varying Rate Code

Neurons throughout the brain, spinal cord, and peripheral nervous system alter their rate of spike discharge to represent a change in intensity. Examples include pressure at a spot of skin, contrast in a spot on the visual field, the salience of a moving visual display, the proximity of an intended eye movement to a place in the visual field, and the force exerted by a muscle. The time-varying rate code is thus the best-known example of a temporal code. It is clearly the dominant principle of activation in the peripheral nervous system: the more spikes, the greater the intensity of the stimulus or the greater the force of the muscle contraction.

The rate code applies in the cortex as well, but there is a catch. Because the spikes emitted by a neuron occur at irregular intervals, the spike rate cannot be discerned from the interval between two spikes. From the point of view of a neuron that is receiving the message about rate, it would need to wait for several spikes to get a sense of their average rate. However, simple perceptual tests show that the brain is capable of processing information about changes in sensory stimuli that occur very quickly, as fast as one or two spike intervals. Evidence from neuroanatomy and neural recording experiments suggests that the cortex solves this problem by representing rate using several neurons. Because the receiving neuron obtains many samples of the same message, all with different erratically spaced spikes, it can estimate the spike rate by averaging across neurons in a short period (Figure 2).

The time-varying ensemble spike rate is a simple temporal code that allows the brain to keep track of dynamic changes in the sensory environment. In principle, it can be used to represent the intermediate stages of neural computations that underlie changes in mental states, idea formation, decision-making, and emotions.

Of the four categories of evidence described above, all support a role for spike rate in the coding of information.

1. Changes in spike rate are reliably reproduced in laboratory conditions when the same stimuli (or behavior) can be presented repetitively.
2. Variability in the spike rate predicts sensitivity to weak stimuli; for example, when monkeys make difficult judgments about sensory stimuli, their rate of errors can be predicted by the variability in the spike rate of appropriate neurons in the sensory cortex. In parts of the association cortex, it is possible to know which way an animal will decide about an ambiguous stimulus by measuring the spike rate.

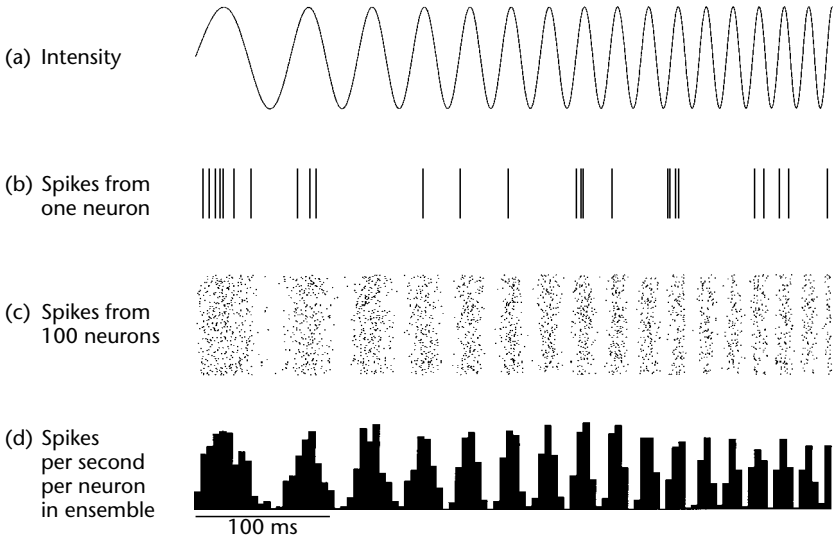


Figure 2. Time-varying rate code. How do neurons encode information that changes as a function of time? (a) A time-varying signal is shown at the top of the figure. This could represent a changing light intensity, or a quantity used in a calculation for directing atttention, say, to a bouncing ball. In this example, the signal intensity changes slowly at first (left) but eventually the changes occur more rapidly. (b) A single neuron emits spikes more often when the intensity is high, but the intervals between spikes preclude encoding the rapid changes in signal intensity. (c) If there are many neurons that emit spikes in this fashion, then their average (d) provides a reasonable approximation to the signal. This ensemble coding takes advantage of the presence of many neurons that encode the same information by changing their rate of spike discharge.

3. Electrically stimulating small groups of neurons (300–1000) to higher rates causes movements and, in a few cases, apparent changes in perception. In one remarkable example from the visual cortex, investigators measured the spike rate from individual neurons in association with visual stimuli. On the basis of changes in spike rate, they deduced the message represented by that neuron and its neighbors. This deduction is really a hypothesis about the neural code. In one of the few direct tests of this hypothesis, the investigators stimulated the neurons by passing small amounts of alternating current through the electrode. This stimulation caused animals to report perception of the stimulus, consistent with the message the investigators had inferred from the spike rate.
4. There is ample evidence from biophysics that neurons increase their rate of spike discharge when they receive more excitatory input. Exactly how neurons achieve a stable rate of firing in response to synaptic input is an active area of research. The source of variability in the spike discharge, hence the reliability of the rate code, is also an active area of investigation. As mentioned earlier, the variable spike discharge seems to be a property of cortical neurons, which for unknown reasons are much more variable in their response than are neurons in subcortical structures.

In short, there is ample evidence that the brain uses a time-varying rate code to mark the intensity of a stimulus, action or variable in a computation. The actual identity of a stimulus (e.g. its color, location, direction of movement), the particular action, or the fact that a variable is subtracted or added is coded by the neuron's location in the brain, namely its connections to other neurons and presumably its position in space in the cortex. According to this view of temporal coding, the rich set of symbols that the brain uses to encode information derives mainly from wiring in the patterns of neurons that can be activated under a variety of conditions (through development and learning). The rate code is just a way to represent the degree of this activation – the amount of evidence for a proposition that is represented by the identity of the neuron or neural ensemble. This idea is sometimes referred to as a labeled-line or place code.

In contrast, we now turn to two putative neural codes that use the pattern of spikes as a function of time to encode different features of the environment. By using time to code something else, they have the potential to achieve a much higher degree of complexity than the rate code.

Spike Bar Code

The intervals between spikes in cortex tend to be quite variable, but depending on one's point of view, this variability can be seen as a nuisance (as above) or as a potential code. Figure 3 illustrates

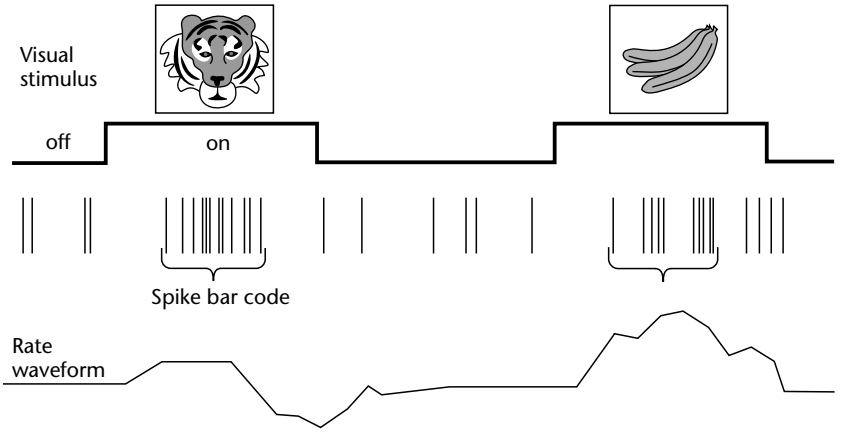


Figure 3. Spike bar code and rate waveform code. Spike activity from an idealized neuron occurs in association with the appearance of two visual stimuli. Brackets identify patterns of spikes that could provide distinct labels for 'tiger' and 'banana' akin to a bar code used to identify merchandise in retail stores. This idea and other related spike interval codes are improbable because specific spike patterns do not occur with any systematic regularity in the cortex. The jagged trace below the spikes represents the ensemble spike rate that might be obtained by averaging the activity from many neurons increasing and decreasing their rate of discharge in a manner similar to the spike train shown. According to this scheme, the exact pattern of spikes from each neuron would not carry information; rather, each neuron contributes to the average spike rate (as in Figure 2). The spike rate waveform has a different shape in association with the tiger and banana. It has been suggested that this rate waveform could encode stimuli. Note that both the spike bar code and the rate waveform codes would allow the same neuron to encode a variety of messages and that both use time to encode stimuli that are not changing.

a rather extreme idea for a temporal code, which this author terms a 'spike bar code'. The example shows a neuron that emits a different pattern of spikes when different pictures are presented to vision. The particular pattern of intervals separating the spikes could in principle symbolize a complex object. There are many possible patterns of spikes that a neuron could emit in a 300 ms window of time – roughly the amount of time between successive scanning eye movements. Therefore, the bar code could provide a rich alphabet for coding information.

The spike bar code illustrates an intriguing idea, but at present there is little evidence to support its use in the brain (see Table 1).

1. There is no example of a neuron in the cortex that emits the same pattern of spikes in association with a particular stimulus or behavior. There are occasional reports of patterns of spikes that seem to occur more often than expected by chance in some parts of the cortex, but not in association with a particular stimulus or action, and the claim that the patterns do not occur by chance has been contested.
2. Without clear evidence of specific patterns, it is impossible to test whether they predict errors in perception or variability in behavior.
3. To the minimal extent that anyone has tried to mimic patterns of activity in the brain, the effort has not yielded anything of interest.
4. The main evidence in favor of such a code is that it is not ruled out by what we know about how neurons generate action potentials.

In particular, neurons seem to be capable of emitting action potentials with a precision of below 1 ms in response to the same amount of current. The argument then goes like this: if the intervals between spikes can be controlled so precisely, then the brain must be using these intervals for coding. Of course, we do not really know that the intervals between spikes can be controlled precisely; we know only that the source of imprecision is not the part of the neuron that converts the current it collects to an action potential.

Besides the lack of experimental evidence for the spike bar code, there are two additional problems. First, the signal would need to be deciphered by neurons at the receiving end. It is hard to imagine a mechanism that would allow a neuron to respond selectively to a spike pattern that extends by more than a few tens of milliseconds or a few spikes. Second, it takes time to decode such a message. At a minimum, it would take the length of the message itself (about 300 ms for the messages depicted in Figure 2). It is hard to reconcile such a scheme with the rapidity of sensory processing.

Rate Waveform Code

The rate waveform code is related to the bar code in that it also uses the temporal changes in spike production to encode information. In this case, it is not the precise pattern of spikes and intervals but rather the rate of spike production that codes information. In the example in Figure 3, a benign stimulus (the banana) causes the spike rate to rise and then return to baseline, whereas in response to a threatening stimulus (the tiger) the rate rises and then falls below baseline before returning to baseline (i.e. it has a positive and negative phase). The spikes themselves occur more or less randomly, but with greater or lesser frequency in accordance with these rate waveforms. As noted earlier, this implies that many neurons undergo similar rate fluctuations.

There is better evidence for the rate waveform code than for the bar code.

1. In the visual cortex, some stimuli give rise to transient increases in spike rate followed by a rapid return to baseline, whereas others lead to more sustained responses. Usually, different neurons respond in these modes, but there are examples of the same neuron responding to one kind of stimulus with sustained activity and to another more transiently. The best example of a temporal code of this sort is found in the taste system, where different tastes (e.g. sweet and bitter) cause the same neuron to undergo different patterns of rate change.
2. The second type of evidence is lacking, however. If different waveforms connote different messages, then one would like to witness a correlation between the variation in response waveform and an animal's perception. Experiments of this type have not been tried or have been unsuccessful.
3. The third type of evidence, manipulation of the code, has been tried in the taste system of rats. When the brainstem nucleus that receives taste information is stimulated, rats respond as if they had tasted bitter or sweet, depending on the pattern of firing rate change induced by the stimulation. Experiments of this type have not been tried in the cortex. The closest example is a negative finding: changing the temporal pattern of activity in the somatosensory cortex does not interfere with a monkey's ability to discriminate flutter vibration frequency.
4. Because neurons can modulate their firing rate, there are no obvious theoretical obstacles to the coding of information with a rate waveform.

Although we do not yet know how a rate waveform would be decoded, it is in principle no more difficult a problem than decoding any time-varying function, as occurs in visual neurons that respond to stimuli moving in a particular direction and speed.

Synchrony Code

The last item on the menu is the synchrony code. This is a popular example of a code that would exploit temporal and spatial relationships between spikes to encode information (spatial in this context refers to places in the brain). The idea is that simultaneous spikes from two or more neurons might encode information. A natural candidate for the kind of information to be encoded is the relation between parts of objects or elements of movement. That is, synchronous spikes could represent combinations of features that are themselves encoded by single neurons or ensembles of neurons with similar properties.

Like the spike bar code, the synchrony code exploits the irregular intervals between spikes. As illustrated in Figure 4, another implication of irregularly spaced spikes is that the odds of any two neurons emitting a spike at the same moment (say, within 3 ms of each other) is a relatively rare chance event. It has been suggested that under particular circumstances, neurons can produce synchronous spikes in excess of the rate expected by chance. These synchronous spikes could constitute a special code. In vision, for example, it has been proposed that synchronous spikes are used to bind together separate features of objects into coherent wholes (Figure 4) and even to promote the representation of vision to conscious awareness.

Despite the enthusiasm for this code, there is little evidence to support it.

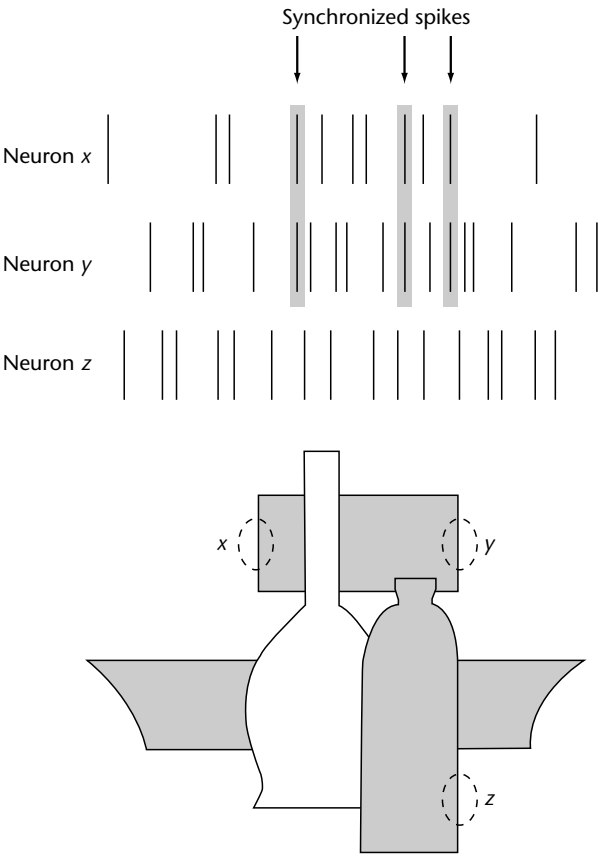


Figure 4. Synchrony code. Idealized responses from three neurons to a visual scene consisting of bottles and shapes. The three neurons respond to vertical contours marked by the dashed ellipses. The ellipses are not part of the scene. The spikes from each of the neurons exhibit variable intervals. On three occasions, spikes from neurons x and y occur within 3 ms of each other (arrows and gray highlight). Using this criterion, there are no synchronous spikes between neuron z and the other neurons. It has been proposed that synchronous spikes could encode information about the scene. For example, the synchronous spikes could indicate that the contours at the locations represented by x and y are part of the same object, whereas the contours represented by neurons y and z are not part of the same object. Despite much enthusiasm for this idea, the experimental evidence is weak.

1. There are many examples of pairs of neurons that emit synchronous spikes reliably. Typically, these are neurons that lie near each other in the brain and receive common synaptic input. There are also reports of neurons that encode different visual features but tend to respond synchronously when the two features are bound to a common object.
2. The few attempts to correlate synchronous spikes with behavior have failed to provide supporting evidence for a synchrony code. For example, synchronous spikes occur just as often whether a pair of features represented by the neurons is bound to a single object in the foreground or is split, with one feature in the background.
3. Experiments in which synchrony has been mimicked by means of stimulation have been performed in the motor system, but there is no report of any experiments in which sensory neurons have been stimulated in and out of synchrony. This might be technically difficult because stimulation always tends to synchronize the neurons near the stimulating electrode. On the other hand, some investigators have tried to disrupt synchrony by imposing asynchronous flicker to different component features of an object. This seems to have no effect on perception and thus casts doubt on the idea that synchronous discharge encodes anything special.
4. Much of what we know about synaptic physiology would indicate that synchronous spikes are more effective than asynchronous spikes at inducing a response from a postsynaptic neuron. However, cortical neurons receive many hundreds of excitatory inputs for every spike they emit. It is far from clear how synchronous spikes that convey information are to be distinguished from the rest of these spikes.

CONCLUSION

The topic of temporal coding is fundamental to basic and clinical neuroscience. The ability to read the neural code will help neuroscientists understand how the brain represents information and uses it in novel computations that underlie thought and behavior. Understanding the neural code may one day provide the ability to use brain recordings to control prosthetic devices in people who suffer from spinal cord and nerve injury.

The evidence at hand favors the use of spike rate to encode the intensity of values whose meaning is given by the identity of the neuron or group of neurons emitting spikes. This ensemble rate code is a temporal code because it provides a means to represent intensity (or magnitude) as a function of time. The alternative codes use time to encode meaning. These putative codes could expand the brain's ability to represent information in the same way that the Morse code allows two symbols (dot and dash) to encode all the letters of the alphabet (or the computer binary character code, which allows two numbers, 1 and 0, to encode all typographic symbols). There is some evidence for spike timing codes in the peripheral nervous system and in brainstem structures that are specialized for processing sound, but the idea is largely unsupported in the cerebral cortex.

Further Reading

Bair W (1999) Spike timing in the mammalian visual system. *Current Opinion in Neurobiology* **9**: 447–453.

Britten KH, Shadlen MN, Newsome WT and Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience* **12**: 4745–4765.

DeCharms RC and Zador A (2000) Neural representation and the cortical code. *Annual Review of Neuroscience* **23**: 613–647.

Di Lorenzo PM and Hecht GS (1993) Perceptual consequences of electrical stimulation in the gustatory system. *Behavioral Neuroscience* **107**: 130–138.

Lamme VAF and Spekreijse H (1998) Neuronal synchrony does not represent texture segregation. *Nature* **396**: 362–366.

Leon MI and Shadlen MN (1998) Exploring the neurophysiology of decisions. *Neuron* **21**: 669–672.

McClurkin JW, Optican LM, Richmond BJ and Gawne TJ (1991) Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science* **25**: 675–677.

Ogawa H, Yamashita S and Sato M (1974) Variation in gustatory nerve fiber discharge pattern with change in stimulus concentration and quality. *Journal of Neurophysiology* **37**: 443–457.

Parker AJ and Newsome WT (1998) Sense and the single neuron: probing the physiology of perception. *Annual Review of Neuroscience* **21**: 227–277.

Rieke F, Warland D, de Ruyter van Steveninck RR and Bialek W (1997) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Romo R, Hernandez A, Zainos A and Salinas E (1998) Somatosensory discrimination based on cortical microstimulation. *Nature* **392**: 387–390.

Romo R, Hernandez A, Zainos A, Brody CD and Lemus L (2000) Sensing without touching: psychophysical performance based on cortical microstimulation. *Neuron* **26**: 273–278.

Salzman CD, Murasugi CM, Britten KH and Newsome WT (1992) Microstimulation in visual area MT: effects on direction discrimination performance. *Journal of Neuroscience* **12**: 2331–2355.

Shadlen MN and Movshon JA (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* **24**: 67–77.

Shadlen MN and Newsome WT (1994) Noise, neural codes and cortical organization. *Current Opinion in Neurobiology* **4**: 569–579.

Shadlen MN and Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation and information coding. *Journal of Neuroscience* **18**: 3870–3896.

Singer W (1999) Neuronal synchrony: a versatile code for the definition of relations? *Neuron* **24**: 49–65.

Softky WR and Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience* **13**: 334–350.

Real-valued Computation and Complexity

Advanced article

Felipe Cucker, City University of Hong Kong, Hong Kong
Steve Smale, City University of Hong Kong, Hong Kong

CONTENTS

The theory of real-valued computation and complexity deals with foundational aspects of scientific computation. Here the basic unit of information is the floating-point number and a computational model whose basic unity of information is a real number reasonably captures this feature.

INTRODUCTION

In problems of learning and intelligence the traditional mathematical foundations are derived from Gödel and Turing. This is reflected in the debate over whether machines will ever be able to think as humans do. Most recently, the discussion about Roger Penrose's book (Penrose, 1989) reflected this foundational issue.

However, developments in the theory of learning in both machines and humans, revolve mainly around continuous mathematics. For example, real number analysis plays a role via the Hodgkin–Huxley equations and statistical learning theory.

For this reason one is led to consider alternative foundations for intelligence modeling, learning and computation.

AN ALTERNATIVE FORMALISM FOR COMPUTATION

'Scientific' computation is a domain of computation based mainly on the equations of physics. For example, from the equations of fluid mechanics, scientific computation helps us to design better aeroplanes, or predict the weather. The theory underlying this kind of computation is called numerical analysis.

There is a substantial conflict between theoretical computer science and numerical analysis. These two subjects have common goals but have grown apart. For example, computer scientists are uneasy with the differential calculus, while numerical analysts thrive on it. On the other hand numerical analysts see no use for the Turing machine.

This conflict has at its roots another age-old conflict, that between the continuous and the discrete. Computer science is oriented by the digital nature of machines and by its discrete foundations (Turing machines). On the other hand, differential equations are central to numerical analysis and this discipline depends heavily on the continuous nature of the real numbers.

The developments that began with the work on the theory of computable functions by Gödel, Turing and others in the 1930s have given a firm foundation to computer science as a rigorous discipline. Turing machines provide a unifying, formalized concept of algorithm. Thus computer scientists have been able to develop a complexity theory which permits discussion of lower bounds of all algorithms without ambiguity.

The situation in numerical analysis is quite the opposite. Algorithms are primarily a means to solve practical problems. There is not even a formal definition of algorithm in the subject.

A major obstacle to reconciling scientific computation and computer science is the present view of the machine, i.e. the digital computer. As long as the computer is seen simply as a finite or discrete object, it will be difficult to systematize numerical analysis. The belief that the Turing machine is inadequate as a foundation for real number algorithms led to the machine model proposed in Blum *et al*. (1989). This model takes its inputs from $\mathbb{R}^\infty$, the disjoint union of $\mathbb{R}^n$ for $n \geq 1$, and returns as outputs elements of this space. During the computation it performs arithmetic operations and comparisons and has the ability to 'move' information (i.e. real numbers) around within its

state space (roughly, its 'tape'). This last feature, equivalent to the addressing instructions on a RAM, allows for some form of management. The *size* of a point $x \in \mathbb{R}^\infty$ is the unique $n = \text{size}(x) \in \mathbb{N}$ such that $x \in \mathbb{R}^n$. Running time, which is defined as the number of operations, comparisons and movements, is then considered as a function of the input size.

REAL-VALUED COMPUTATIONAL COMPLEXITY

Probably the most important result in Blum *et al*. (1989) is the existence of NP-complete problems over the reals (and other rings as well). The 'P = NP' question had been at the heart of research in discrete complexity theory and a specific goal of Blum *et al*. (1989) was to extend this question to computations involving real numbers.

A set S is *decidable in polynomial time*, or in the class P, if there is a machine M deciding S whose running time is bounded, for all inputs of size n, by a polynomial function of n. Classically, S is a set of strings over a finite alphabet. Over the reals, the same definition can be used by taking S to be a set of points in $\mathbb{R}^\infty$. In this case we write $\text{P}_\mathbb{R}$ instead of P. Similarly, a function $\varphi : \mathbb{R}^\infty \to \mathbb{R}^\infty$ is *computable in polynomial time* if there is a machine M computing $\varphi(x)$, for all $x \in \mathbb{R}^\infty$, in time bounded by a polynomial in size (x).

The definition of the class NP over $\mathbb{R}$ is slightly more elaborate, and can be done in several ways. Here we modify the machine model and endow it with the ability to guess points $y \in \mathbb{R}^\infty$ (with cost size (y)). We call such an extended machine *nondeterministic*.

A set S (of points in $\mathbb{R}^\infty$) is in $\text{NP}_\mathbb{R}$ if there exists a nondeterministic machine M satisfying the following two conditions:

1. For all $x \in \mathbb{R}^\infty$, the running time of M with input x is bounded by a polynomial in size (x).
2. For all $x \in \mathbb{R}^\infty$, $x \in S$ if and only if there exists a *guess* $y \in \mathbb{R}^\infty$ such that the computation of M with input x and guess y ends in an accepting state (or, equivalently, returns 1).

COMPLEXITY OF A PROBLEM OVER THE REALS

As an example of a computational problem over the reals, we now focus on feasibility of real polynomials.

Consider polynomials $f_1, \ldots, f_k \in \mathbb{R}[X_1, \ldots, X_n]$. The problem at hand is to decide if there exists a common root $\xi \in \mathbb{R}^n$. Thus, one seeks an algorithm, in fact an algebraic algorithm over $\mathbb{R}$, which on input $f = \{f_1, \ldots, f_k\}$ returns 1 if and only if there is a $\xi \in \mathbb{R}^n$ such that $f_i(\xi) = 0$ for all i, and returns 0 otherwise.

By an algebraic algorithm we have in mind a machine whose computations involve the basic arithmetic operations and whose branching depends on order comparisons.

The input f to such a machine can be thought of as the vector of coefficients of the f_i in $\mathbb{R}^N$ where N is given by the formula

$$N = \sum_{i=1}^{k} \binom{n + d_i}{n} \tag{1}$$

where d_i is the total degree of the polynomial f_i, $i = 1, \ldots, k$. Thus, this N represents the size $S(f)$ of the input f.

A particular feature of the real numbers enables us to consider the same problem with only one polynomial at the cost of slightly increasing the degree. We associate to the polynomials $f_1, \ldots, f_k$ the single polynomial $g = \sum_{i=1}^{k} {f_i}^2$. Now g has the property that for every $\xi \in \mathbb{R}^n$, ξ is a common root of all the f_i if and only if $g(\xi) = 0$. Therefore, solving our problem for the f_i is equivalent to solving it for g. Moreover, if not all the f_i are linear then the degree of g is at least 4. Let us restrict our attention to degree-4 polynomials and consider the following problem, which we call 4-FEAS:

> Given a degree-4 polynomial in n variables with real coefficients, decide whether it has a real zero.

Again, the input g for this problem can be regarded as a vector in $\mathbb{R}^N$ where

$$N = \binom{n+4}{4} \tag{2}$$

is the size $S(g)$ of this input.

An algorithm for solving this problem was first given by Tarski in the context of exhibiting a decision procedure for the theory of real numbers. In the context of complexity theory, Tarski's algorithm is highly intractable. The number of arithmetic operations performed by this algorithm grows in the worst case by an exponential tower of n 2s. Collins devised another algorithm that solved 4-FEAS within a number of arithmetic operations bounded by

$$2^{2^{s(g)}} \tag{3}$$

More recent algorithms have achieved single exponential bounds.

Since evaluating g at a point $x \in \mathbb{R}^n$ has cost bounded by a linear function of the size of g, 4-FEAS is in NP over $\mathbb{R}$. The machine considers a guess x and evaluates $g(x)$, accepting if $g(x) = 0$. A guess leading to acceptance proves the existence of a zero of g.

A result in Blum *et al*. (1989) shows a certain universality of 4-FEAS with respect to this property of being NP-complete over $\mathbb{R}$. As a consequence, 4-FEAS belongs to $\mathrm{P}_{\mathbb{R}}$ if and only if $\mathrm{P}_{\mathbb{R}} = \mathrm{NP}_{\mathbb{R}}$.

It is an open question whether the 4-FEAS problem is intractable; that is, whether there is no algorithm solving 4-FEAS in polynomial time. This would imply $\mathrm{P}_{\mathbb{R}} \neq \mathrm{NP}_{\mathbb{R}}$.

FUTURE DIRECTIONS

The real machines in Blum *et al*. (1989) assume exact arithmetic. But rounding is a part of how machines and biological intelligence work. Thus, one needs to extend the above to accommodate rounding and approximate reasoning (see, for example, Cucker and Smale (1999)).

Acknowledgment

Part of the material contained in this article has been taken from Blum *et al*. (1996). We are grateful to World Scientific for kindly allowing us to do so.

References

Blum L, Cucker F, Shub M and Smale S (1996) Complexity and real computation: a manifesto. *International Journal of Bifurcation and Chaos* **6**: 3–26.

Blum L, Cucker F, Shub M and Smale S (1998) *Complexity and Real Computation*. New York, NY: Springer.

Blum L, Shub M and Smale S (1989) On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society* **21**: 1–46.

Cucker F and Smale S (1999) Complexity estimates depending on condition and round-off error. *Journal of the ACM* **46**: 113–184.

Penrose R (1989) *The Emperor's New Mind*. New York, NY: Oxford University Press.

Further Reading

Cucker F (1999) Real computations with fake numbers. In: Wiedermann J, van Emde Boas P and Nielsen M (eds) *26th ICALP*. **1644**: 55–73. Springer-Verlag.

Cucker F and Smale S (2002) On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* **39**: 1–49.

Meer K and Michaux C (1997) A survey on real structural complexity theory. *Bulletin of the Belgian Mathematical Society* **4**: 113–148.

Shub M (1994) Mysteries of mathematics and computation. *The Mathematical Intelligencer* **16**: 10–15.

Smale S (1990) Some remarks on the foundations of numerical analysis. *SIAM Review* **32**: 211–220.

Smale S (1997) Complexity theory and numerical analysis. In: Iserles A (ed.) *Acta Numerica*. Cambridge, UK: Cambridge University Press.

Real-valued Mathematical Tools for Cognitive Modeling

Intermediate article

Erkki Oja, Helsinki University of Technology, Espoo, Finland

CONTENTS

Many cognitive models use continuous mathematical formalisms like optimization in vector spaces by gradient descent, multivariate probability distributions, and data compression by coding.

THE DIVERSITY OF MATHEMATICAL FORMALISMS USED IN COGNITIVE MODELS

A great variety of mathematical models are used in cognitive science. The central concept is a symbol processing system. Thus, linguistic models, logic, and algorithms are essential tools. The linguistic models mostly employ concepts from discrete mathematics, although probabilistic models are becoming more popular, for example in natural language processing.

However, many connectionist models like artificial neural networks (Rumelhart and McClelland, 1986) are based on distributed and real-valued (contrary to discrete-valued) representations for which the most natural mathematical framework is a *vector space*. A vector is simply an ordered set of scalars, which are called the elements of the vector. In this article, a vector is denoted in boldface, and the elements of an n-dimensional vector $\mathbf{x}$ are denoted by x_i, $i = 1, \ldots, n$. Transformations between such vectorial representations are appropriate *continuous functions*, linear or nonlinear, applied on the vectors. Optimizing these functions is a central problem. Optimization is usually based on some cost function that must be minimized; hence techniques based on gradient descent are widely used. (*See* **Distributed Representations**)

Often, especially in models of sensory processes, real-world stimuli are modeled, and their vast variability is best described by *probabilistic models*. These are given as multivariate probability distributions. An essential problem is the propagation of the stimuli through the sensory pathways. A high degree of data compression is necessary to remove the redundancies. For modeling this compression, the theory of communication channels and *information theory* can be used. It is based on the coding length of a message or signal, with the key question being the optimality of the code used. (*See* **Pattern Recognition, Statistical**)

SEARCH IN CONTINUOUS SPACES: NUMERICAL OPTIMIZATION

The typical approach to many of the problems in cognitive science is based on parameterized functions. In neural networks, we search for the best nonlinear mapping between given inputs and outputs. In parametric statistics, we search for the density model that best describes our data within a family of parametric densities. In information theory, we may be seeking a transformation of signals transmitted in parallel such that the mutual information between the parallel signals is minimized, giving optimal coding. In all these cases we are searching for the best parameter vector in a continuous multivariate space.

The solution method is based on *cost functions*, also called objective functions. These are scalar functions $J(\mathbf{w})$ of a multivariate vector $\mathbf{w}$, whose elements w_i are the unknown parameters of the

model. The function $J(\mathbf{w})$ must be carefully designed so that the optimal vector of parameters that we are seeking occurs at a minimum. For example, in neural network training, the cost function is the sum of the squared distances between the outputs given by the network for the given input vectors, and the desired outputs that are associated with those input vectors. The smaller the value of this cost function, the more closely the true outputs follow the desired outputs. In this case, the unknown parameters, or elements of the vector $\mathbf{w}$, are all the weights and offsets of all the neurons in the network. The dimension of $\mathbf{w}$ can be very high, in the thousands or even more, so solving this problem is a demanding task. An often occurring problem is that the function $J(\mathbf{w})$ does not have just one global minimum but several local minima.

Minimization of multivariate functions, possibly under some constraints on the solutions, is the central problem of *optimization theory* (Fletcher, 1987; Luenberger, 1969). Mostly, the algorithms are based on the *gradients* of the cost functions. If the solution is constrained so that the parameter vector is not allowed to have all possible values, then often the method based on Lagrange multipliers is the most feasible one (Luenberger, 1969). Sometimes, the cost function $J(\mathbf{w})$ is so complex that it has no gradient, or computing the gradient is computationally too expensive. Then one must resort to random search methods, like simulated annealing or genetic algorithms (Goldberg, 1989).

GRADIENT FOLLOWING METHODS

In *gradient descent*, we minimize a function $J(\mathbf{w})$ iteratively by starting from some initial point $\mathbf{w}(0)$ of the high-dimensional parameter space, computing the gradient of $J(\mathbf{w})$ at this point, and then moving in the direction of the negative gradient or the steepest descent by a suitable distance, to get a new point $\mathbf{w}(1)$. Once there, we repeat the same procedure at the new point, and so on. For iteration step $t = 1, 2, \ldots$, we have the update rule

$$\mathbf{w}(t) = \mathbf{w}(t-1) - \alpha(t)[\partial J/\partial \mathbf{w}]|_{\mathbf{w}=\mathbf{w}(t-1)}$$

with the gradient $\partial J/\partial \mathbf{w}$ taken at the point $\mathbf{w}(t-1)$. The gradient of $J(\mathbf{w})$ is a vector, whose elements are the partial derivatives of $J(\mathbf{w})$ with respect to the elements w_i of $\mathbf{w}$. The parameter $\alpha(t)$ gives the length of the step taken in the direction of the negative gradient. It is called the *step size* or sometimes the learning rate. The iteration is continued until it converges, which in practice happens when the distance between two consequent solutions $\mathbf{w}(t)$ and $\mathbf{w}(t-1)$ goes below some small tolerance level.

Geometrically, a gradient descent step means going downhill. The graph of $J(\mathbf{w})$ is the multidimensional equivalent of a mountain terrain, and we are always moving downward in the steepest descent. This shows a possible disadvantage of gradient descent: it will lead to the closest valley or local minimum instead of the global minimum. The method offers no way to escape from a local minimum. So, good initial values $\mathbf{w}(0)$ are very important in initializing the algorithm.

An example of gradient descent is the error back propagation algorithm widely used in neural network training (Rumelhart and McClelland, 1986). The nonlinear cost function used is notorious for its local minima and there is a large body of literature devoted to the minimization problem. (*See* **Backpropagation**)

SECOND-ORDER METHODS AND DYNAMIC ADJUSTMENT OF STEP SIZE

A drawback of the simple gradient descent optimization method is the speed of convergence: a large number of iteration steps may be needed to reach a minimum of the cost function. The speed can be controlled with the step size $\alpha(t)$, but it can be shown that unless the step size is carefully adjusted to the shape of $J(\mathbf{w})$, the method converges *linearly* (Luenberger, 1969). Defining the error at step t as the Euclidean distance between $\mathbf{w}(t)$ and the true solution at the local minimum $\mathbf{w}^*$, linear convergence means that at the next step $t+1$, the error is a linear fraction of the error at step t.

Faster convergence can be achieved by a *second-order method*, in which the error at step $t+1$ is proportional to the *square* of the error at the previous step. When the error becomes small, close to convergence, its square at the next iteration step can be orders of magnitude smaller still, leading to very fast convergence. The most widely used second-order iteration method is Newton's method. Applied to minimizing the cost function $J(\mathbf{w})$, Newton's method also uses the second-order partial derivatives $\partial^2 J(\mathbf{w})/\partial w_i \partial w_j$ of $J(\mathbf{w})$, that make up the so-called Hessian matrix. It is often denoted by $[\partial^2 J(\mathbf{w})/\partial \mathbf{w}^2]$. The iteration for minimizing the function $J(\mathbf{w})$ is now

$$\mathbf{w}(t) = \mathbf{w}(t-1) - \left[\partial^2 J(\mathbf{w})/\partial \mathbf{w}^2\right]^{-1} [\partial J/\partial \mathbf{w}]\,|_{\mathbf{w}=\mathbf{w}(t-1)}$$

Note how the scalar-valued step size $\alpha(t)$ in the usual gradient descent update rule is now replaced with a matrix that is the inverse of the Hessian,

computed at the point $\mathbf{w}(t-1)$. The algorithm can make use of the local shape of $J(\mathbf{w})$, which results in quadratic or second-order convergence (Luenberger, 1969).

On the other hand, Newton's method is computationally much more demanding per iteration than the steepest-descent method. The inverse of the Hessian has to be computed at each step, which is prohibitively heavy for many practical cost functions in high dimensions. It may also happen that the Hessian matrix becomes close to a singular matrix, which has no inverse. One remedy is to regularize the algorithm by adding small positive numbers to the diagonal; this is the basis of the Marquardt–Levenberg algorithm. Sometimes the Hessian is approximated by its diagonal only, and then the inversion becomes very simple, yet the algorithm is a clear improvement over the usual steepest descent. Newton's method can be seen as the optimal way for *dynamic adjustment* of the step size as the iteration goes on.

For error functions that can be expressed as sums of squares, one can apply the Gauss–Newton method instead of Newton's method. Also the conjugate gradient method provides a compromise between the steepest-descent and Newton methods (Luenberger, 1969). All these algorithms are fairly complex to use in practice, and it is advisable to resort to well tested numerical software to perform the computations.

LINEAR PROGRAMMING

Perhaps the simplest special case for the cost function $J(\mathbf{w})$ is a *linear function*: a weighted sum of the elements w_i. Many practical problems result in linear cost functions, at least as good approximations, which are recommended because of their mathematical tractability. Now the problem of minimizing $J(\mathbf{w})$ is meaningless, however, unless some extra conditions are added. A linear function has no minima or maxima, but approaches plus or minus infinity just like a (non-horizontal) straight line in the two-dimensional coordinate system. What are needed to make the problem well defined are constraints on the unknown parameters w_i. This leads to the standard *linear programming problem* (Dantzig, 1963):

$$\text{minimize } J(\mathbf{w}) = \sum\nolimits_i c_i w_i \text{ under the } N \text{ constraints } \sum\nolimits_k a_{jk} w_k \leq b_j, \quad j = 1, 2, \ldots, N$$

The notation $\sum_i$ means taking the sum over all the appropriate indices i. Here, the given numbers c_i determine the cost function, and the given numbers a_{jk} and b_j determine the constraints on the solution vector $\mathbf{w}$.

The equations have an easy geometrical interpretation. Imagine the usual two-dimensional coordinate plane, but now the two axes are the two components w_1 and w_2. Then the constraints will in practice enclose a convex polygonal region, the so-called feasible set for the solution. All the solutions $\mathbf{w} = (w_1, w_2)$ must be within this closed region. The cost function is linearly decreasing within this region in some direction that is determined by the two coefficients c_1 and c_2. It is easy to see that the minimum is usually attained in a corner of the feasible region, in which just two of the borders of the region meet, determined by two of the N constraints. The solution of the problem is based on finding out which of the corners gives the minimum. The same geometrical intuition generalizes to any dimensions for $\mathbf{w}$.

There is a highly popular and successful algorithm for solving the linear programming problem, the so-called Simplex method (Dantzig, 1963). It is available in a large number of software packages.

DEALING WITH UNCERTAINTY: THE BASICS OF PROBABILITY THEORY

Probability theory deals with random events. If A is a random event, it is meaningful to give it a *probability* $P(A)$. It usually has the interpretation that, in a very long sequence of trials, with A one of the possible outcomes – tossing a coin or dice is the most classical example – the probability $P(A)$ gives the relative frequency for the occurrence of A. It is also possible to interpret the probability subjectively, as the degree of belief in the given outcomes. If $A_1, \ldots, A_n$ are all the possible outcomes of one trial, then their probabilities sum up to one: $\sum_i P(A_i) = 1$. If there is another trial with the possible outcomes $B_1, \ldots, B_m$, then basic probability theory will tell us how to formulate probabilities for various combinations of possible outcomes of the two trials.

Basic probability theory suffices for the case of discrete events. However, in continuous mathematical models, the theory has to be extended to deal with random real-valued variables and vectors. The central mathematical concept is then the probability density.

SAMPLE SPACES, PROBABILITY MASS, AND PROBABILITY DENSITY

A scalar *random variable* $x = x(\omega)$ is defined over a *sample space* Ω. The sample space totally determines

the probability structure of x. It is hidden in the sense that only the values of the variable x are observable, not the argument ω. The concept of the sample space is necessary to rigorously develop the probability theory for continuous variables, starting from measure theory (Doob, 1953). For example, the *distribution function* of x, defined as $F(\xi) = P(x(\omega) \leq \xi)$, actually means the probability measure of that region in Ω over which this inequality is true. However, in practice the sample space is not explicitly mentioned at all and the random variable is just denoted by x.

The derivative of the distribution function, if it exists, is called the *probability density function* of x (or just the density) and denoted by $f(\xi) = dF(\xi)/d\xi$. It then clearly follows that

$$P(\xi_1 \leq x \leq \xi_2) = \int_{(\xi_1 \leq u \leq \xi_2)} f(u)\mathrm{d}u$$

The integral of the density is taken over the interval (ξ_1, ξ_2). The variable u is just a dummy variable, necessary in integration, but having no significance. The integral is only a function of the end points ξ_1 and ξ_2. Through this kind of integral, the probability density function $f(\xi)$ can be used in practice to compute the *probability mass* of any interval of x. It follows from this definition that the integral of the density over the whole x-axis is equal to one.

All this generalizes in a straightforward way to multivariate random entities called random vectors. In cognitive modeling, we quite often wish to model probabilities of several interrelated variables. The distribution of a vector $\mathbf{x}$ with n elements x_i is defined as $F(\xi) = P(\mathbf{x} \leq \xi)$, where the vectorial inequality must be understood element by element. The variable ξ also has n elements ξ_i. The density is the derivative of this with respect to all the components ξ_i, defined as $f(\xi) = \partial^n F(\xi)/\partial\xi_1\partial\xi_2 \ldots \partial\xi_n$ Again, given a region Γ in the n-dimensional ξ-space, its probability mass is given by the integral of the density $f(\xi)$ over the region Γ. The integral over the whole n-dimensional space is equal to one (see, e.g. Morrison, 1967).

COMMON PROBABILITY DISTRIBUTIONS

When computing with probability densities, they must be given explicit mathematical forms. The most often used probability density is the normal or *gaussian* density (see Papoulis, 1991). It is defined for one random variable x as

$$f(x) = 1/(\sigma\sqrt{2}\pi)\exp[-(x-\mu)^2/(2\sigma^2)]$$

We have here adopted the simplified notation, widely in use, where the symbol for the random variable x is also used as the argument in the density, instead of some auxiliary variable ξ. The two parameters σ and μ that completely define this function have a very clear meaning: μ is the *mean value* of x, and σ is the standard deviation, whose square is called the *variance* of x. These are the two first moments of this density. It can be shown from the above equation that these parameters satisfy

$$\mu = E\{x\} = \int x\, f(x)\mathrm{d}x$$

$$\sigma^2 = E\{(x-\mu)^2\} = \int (x-\mu)^2 f(x)\mathrm{d}x$$

The integrals are taken from minus infinity to plus infinity and define the *expectations* (denoted by E{}) of x and $(x-\mu)^2$, respectively. These are the theoretical definitions; if we have available a sample from a random variable x that we assume to be gaussian, then the mean and variance are easily estimated by the sample average and the sample standard deviation, respectively. The importance of the gaussian distribution is largely due to the central limit theorem, stating that a sum of independent random variables, under very general conditions, tends to be gaussian. For instance, random noise can often be considered to arise from a large number of small independent additive effects, and is therefore naturally modeled as gaussian.

The multivariate generalization of the gaussian density is given by

$$f(\mathbf{x}) = \text{constant} \times \exp[-1/2(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\ \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

where $\boldsymbol{\mu}$ is the mean vector of $\mathbf{x}$ and $\mathbf{C}$ is the *covariance matrix*. The covariance matrix has the variances of the elements x_i on the diagonal, and the off-diagonal elements are the covariances.

Two other important densities (given here for a scalar case only, with zero means) are the *Laplace density* $f(x) = \lambda/2\exp(-\lambda|x|)$, where λ is a parameter controlling the variance, and the *uniform density* $f(x) = 1/(2a)$ on the interval $x \in [-a, a]$. These are representatives of two different non-gaussian densities: the Laplace density has thicker 'tails' than the gaussian, and hence its higher-order moments are larger. The uniform density is totally concentrated to a limited interval and has no tails at all.

Density estimation, given only a sample of variables or vectors, is often necessary when developing probabilistic models. This always means approximations. If none of the above densities suffices, then the most usual approximation is a

mixture of gaussians: for scalar x, this is $f(x) = \sum_i P(i)\, f(x|\mu_i, \sigma_i)$ where $f(x|\mu_i, \sigma_i)$is one of the component gaussians, also called a kernel function, with its own mean and standard deviation, and $P(i)$ is the probability that the random vector $\mathbf{x}$ is distributed according to the ith component density. The $P(i)$ sum up to one, which implies that the integral of $f(x)$ over the whole x-axis is equal to one as it should be. Any density can, in principle, be approximated by a mixture of gaussians, and again for this problem there exists good software, usually based on the expectation-maximization algorithm.

CONDITIONAL PROBABILITIES AND BAYES' RULE

In many estimation and inference approaches, we may ask the following question: we have two random vectors, $\mathbf{x}$ and $\mathbf{y}$, and we know that they are dependent. If we are able to measure one of the vectors, say $\mathbf{x}$, and we have a sample of observations available, what can we say about the other, dependent vector $\mathbf{y}$? To answer this kind of questions in a disciplined way, the notion of *conditional probability* is necessary.

First of all, let us denote by $f_{x,y}(\mathbf{x}, \mathbf{y})$ the joint density of $\mathbf{x}$ and $\mathbf{y}$. If $\mathbf{x}$ has n elements and $\mathbf{y}$ has m elements, then this is a function of $m + n$ elements. The density of $\mathbf{x}$, denoted by $f_x(\mathbf{x})$, is obtained from this by integrating over all values of $\mathbf{y}$, and vice versa for $f_y(\mathbf{y})$, the density of $\mathbf{y}$. We define the conditional density of $\mathbf{x}$, given $\mathbf{y}$, as

$$f_{x|y}(\mathbf{x}|\mathbf{y}) = f_{x,y}(\mathbf{x}, \mathbf{y})/f_y(\mathbf{y})$$

This is in complete analogy with discrete probabilities where the conditional probability of an event A, given another event B, is the joint probability of A and B divided by the probability of B. The function $f_{x|y}(\mathbf{x}|\mathbf{y})$ is a proper density function for $\mathbf{x}$, whose integral over all $\mathbf{x}$ is equal to one, as it is very easy to see. In this function, the conditioning quantity $\mathbf{y}$ is usually considered as a fixed parameter vector, although it is actually a random vector, too.

Because a symmetrical definition holds for the conditional density of $\mathbf{y}$, given $\mathbf{x}$, we have two ways to express the joint density: either as $f_{x|y}(\mathbf{x}|\mathbf{y}) f_y(\mathbf{y})$ or as $f_{y|x}(\mathbf{y}|\mathbf{x})\, f_x(\mathbf{x})$. These must be equal, which written in a slightly different way gives

$$f_{y|x}(\mathbf{y}|\mathbf{x}) = [f_{x|y}(\mathbf{x}|\mathbf{y})/f_x(\mathbf{x})]\, f_y(\mathbf{y})$$

This important equation is called the *Bayes' rule* (Morrison, 1967; Papoulis, 1991). It has a central role in Bayesian estimation and inference. The most common interpretation is that $\mathbf{y}$ is a vector-valued quantity that we wish to estimate based on observations contained in vector $\mathbf{x}$. The rule says how the *prior density* of $\mathbf{y}$, $f_y(\mathbf{y})$, is transformed by the observation data $\mathbf{x}$ into a *posterior density* $f_{y|x}(\mathbf{y}|\mathbf{x})$. The function $f_{x|y}(\mathbf{x}|\mathbf{y})$ that is central in this transformation is often called the *likelihood function* of the data $\mathbf{x}$. For instance, based on the observations, a very wide and non-informative gaussian prior density $f_y(\mathbf{y})$ can change into a sharp non-gaussian posterior density $f_{y|x}(\mathbf{y}|\mathbf{x})$ that gives us valuable information about the most likely values of $\mathbf{y}$. (*See* **Bayesian and Computational Learning Theory**)

MEASURING STRUCTURE: THE BASICS OF INFORMATION THEORY

Probability theory gives one approach to modeling the structure of observations and quantities through their multivariate probability distributions. An alternative, but related, approach is provided by *information theory*. Here the emphasis is on coding. By an efficient coding of observations, they can be stored in memory or transmitted over a communication channel. Finding a suitable code depends on the statistical properties of the data. A completely random quantity cannot be coded with a shorter code than the measurements themselves. However, if there is structure in the quantity, like a speech waveform or a picture, then a much shorter code is possible.

THE COMMUNICATION CHANNEL, NOISE, AND THE BIT

The basic framework for information theory is the *communication channel*: a string of symbols are transmitted from a transmitter to a receiver. Let us denote by X the symbol transmitted at a given moment. It is a random variable that has a known number of discrete symbols, say $s_1, s_2, \ldots, s_n$, as possible values. The probability (or relative frequency) of s_i occurring in the symbol string is denoted by $P(X = s_i)$. Then the *entropy* of X is defined as (see Cover and Thomas, 1991)

$$H(X) = -\sum_i P(X = s_i)\ \log P(X = s_i)$$

The minus sign in front makes entropy nonnegative, because $P(X = s_i)$ is between 0 and 1 and its logarithm is therefore negative or zero. The base of the logarithm is traditionally 2. Then the unit of entropy is called a *bit*. If we have only two possible

symbols, with equal probabilities, it is easy to see that the entropy has exactly the value of 1, or one bit. The entropy gives the degree of information that we obtain by observing the value of X. For two symbols, the maximal information is one bit in the case when the two possibilities are equally likely and no intelligent guess can be made. On the other hand, if one of the probabilities is 1, the other 0, then the entropy is equal to 0. In this case, the outcome can be completely predicted and no additional information comes from the observation.

Usually, in any realistic communication channel, there is *noise* present that will occasionally distort the symbols and cause errors. Some noise detection and correction mechanism has to be introduced.

OPTIMAL CODES

In computers and digital systems, binary codes are used. Everything is represented by just two symbols, traditionally noted as 0 and 1. An entity having these two possible values is also called a bit, although originally a bit is the measuring unit for entropy. In digital circuit hardware, a bit is formed using two different voltage levels. A binary string of length eight, a byte, can represent 256 different discrete symbols, for instance 256 different magnitudes of a signal amplitude or pixel gray level. A very relevant question is, whether we actually need eight bits for this or is there some way to use a shorter code.

One way to solve this is to use codes of variable length: for symbols occurring often, a shorter code is used. Then the codes for infrequent symbols become longer, in order not to mix up the codes. For instance, if 0 alone is the code for the most often occurring symbol, then the other codes must start with 1. There is a deep central result saying that the minimum average length of the binary code is given by the entropy of the symbols. Such a code is called an *optimal code*. There exist many practical coding methods that try to get as close to this lower bound as possible (Cover and Thomas, 1991). If noise is present that may introduce errors in the codes, then error detection and correction are necessary. This means adding extra redundant bits by which the errors can be handled, but then the code will also be longer in practice.

The entropy bound is not limited to digital communication systems, but must hold for any communication channel, even for the sensory systems in man and animals. No matter what coding the systems use, they cannot escape this bound. Essentially, this means that the more random or unpredictable the signals to be transmitted are, the higher is their entropy and the less compression is possible.

ENTROPY, MUTUAL INFORMATION, AND MINIMUM DESCRIPTION LENGTH

Entropy was defined above for discrete probability distributions. It can be extended to multivariate continuous densities by the concept of *differential entropy*, defined as (Cover and Thomas, 1991)

$$H(\mathbf{x}) = -\int f(\mathbf{x}) \log\, f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

where $\mathbf{x}$ is a random vector. The *mutual information* of the elements x_i of $\mathbf{x}$ is defined as

$$I(x_1, \ldots, x_n) = \sum\nolimits_i H(x_i) - H(\mathbf{x})$$

One interpretation is based on coding lengths: mutual information approximates the reduction in coding length that results from coding the whole vector $\mathbf{x}$ using some vector code, as compared to coding each individual element x_i separately. Another interpretation is obtained by writing the mutual information in the form

$$I(x_1, \ldots, x_n) = \int f(\mathbf{x}) \log\, [f(\mathbf{x}) / \prod\nolimits_i f(x_i)]\, \mathrm{d}\mathbf{x}$$

where $\prod_i f(x_i)$ is the product of the marginal densities of $\mathbf{x}$. From this it is seen that the mutual information becomes zero when the joint density $f(\mathbf{x})$ equals the product of its marginal densities. But then the elements x_i are independent and give no information about each other. The more dependent the elements are, the higher is the mutual information and the more efficient it is to code the vector $\mathbf{x}$ as a whole.

Often, the observations $\mathbf{x}$ are assumed to follow some model from a family of possible models. Then the question arises which is the best or optimal model. This has been answered in the *minimum description length principle* (Rissanen, 1978) and related approaches. We define the best model as the one giving the shortest code when the model itself is coded first, and then the observation data are coded based on the model. If the model is given by the parameters $\mathbf{M}$ of a parametric density for $\mathbf{x}$, then we get from Bayes' rule

$$-\log\, P(\mathbf{M}|\mathbf{x}) = -\log\, P(\mathbf{x}|\mathbf{M}) - \log\, P(\mathbf{M}) + \text{constant}$$

There the expectations of $-\log\, P(\mathbf{M})$ and $-\log\, P(\mathbf{x}|\mathbf{M})$ give the optimal coding lengths of the models and the data, given the model. Minimizing

these means actually finding the model whose posterior probability is the highest.

References

Cover TM and Thomas JA (1991) *Elements of Information Theory*. New York, NY: John Wiley.

Dantzig GB (1963) *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.

Doob JL (1953) *Stochastic Processes*. New York, NY: John Wiley.

Fletcher R (1987) *Practical Methods of Optimization*. New York, NY: John Wiley.

Goldberg DE (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison Wesley.

Luenberger D (1969) *Optimization by Vector Space Methods*. New York, NY: John Wiley.

Morrison D (1967) *Multivariate Statistical Methods*. New York, NY: McGraw-Hill.

Papoulis A (1991) *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill.

Rissanen J (1978) Modeling by shortest data description. *Automatica* **14**: 465–471.

Rumelhart DE and McClelland JL (eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Further Reading

Amari SI and Kasabov N (eds) (1998) *Brain-like Computing and Intelligent Information Systems*. Singapore: Springer.

Hinton GE and Anderson JA (eds) (1989) *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Hyvarinen A, Karhunen J and Oja E (2001) *Independent Component Analysis*. New York, NY: John Wiley.

Kohonen T (2001) *Self-organizing Maps*. Berlin, Germany: Springer.

Smolensky P, Mozer MC and Rumelhart DE (eds) (1996) *Mathematical Perspectives of Neural Networks*. Mahwah, NJ: Lawrence Erlbaum.

Reasoning about Time

Intermediate article

Paolo Terenziani, Univ. Piemonte Orientale 'Amedeo Avogadro', Italy

CONTENTS

Reasoning about time is a fundamental task in many 'intelligent' activities. From the standpoint of artificial intelligence, it includes representing and making inferences about the dynamic phenomena of the world, change, persistence and temporal constraints between activities.

INTRODUCTION

The human way of perceiving and understanding the world incorporates at a deep level the notion of time. Time seems to be a primitive entity to which objects in the world are related. In particular, the notions of action, causality and change are intrinsically related to the notion of time. Events occur in a given state of the world and cause changes in that state. Only in a static world with no changes would there be no notion of time. Thus, time plays a fundamental role in any model of reality, and reasoning about time is essential in most intelligent activities. For instance, dealing with time is important in many different areas of computer science, including process simulation, databases, protocol, guideline and workflow management, and many areas of artificial intelligence (AI), such as planning, scheduling, diagnosis, expert systems, natural language understanding and knowledge representation. Therefore, many researchers in AI have attempted to define suitable formalisms to represent time-related phenomena and to reason with them.

TEMPORAL PRIMITIVES AND THE STRUCTURE OF TIME

The definition of the basic primitives (i.e., the ontological analysis) of time is a matter of debate within the AI community. First of all, one has to choose between time points and time intervals as the basic temporal entities. There are intuitive and technical arguments in favour of both choices. Time points are taken as the basic primitives, for example, in the situation calculus (McCarthy and Hayes, 1969). Other authors have suggested that time intervals are more intuitive than the abstract mathematical notion of points, and have pointed out some technical problems arising from the use of time points (Allen, 1983).

Next, one has to specify a suitable structure for the temporal domain. For instance, time may be regarded as a discrete collection of temporal elements or, on the other hand, as a dense collection (so that between any two temporal elements there is always a third). Moreover, time may be unbounded (infinite in one or both directions) or bounded. Some ordering relation (temporal precedence) must be defined on the set of temporal elements. This may be linear (i.e., the temporal structure is totally ordered) or branching (for example, each branch may represent a possible future evolution of the given world). Circular time is also used in some applications. (See Van Benthem (1983) for a detailed analysis of these alternatives.)

Naturally, different approaches have been proposed and different choices made, depending on the specific phenomena and tasks at hand. However, one can distinguish two broad streams in research about time within the AI community, henceforth called 'general-purpose' and 'constraint-based' approaches.

GENERAL-PURPOSE (LOGICAL) APPROACHES

General-purpose approaches mainly focus on the definition of a formalism general enough to represent a wide range of temporal phenomena. Such approaches aim at dealing with the dynamic aspects of the world, at describing the internal structure of actions and events occurring in the world, and at modeling how the world changes in response to such actions and events. Although very different formalisms have been devised most of them are first-order or modal logical formalisms.

The Situation Calculus

An important example is the situation calculus (McCarthy and Hayes, 1969). The situation calculus adopts a very natural (and widely used) way of introducing time in first-order logic: functions and predicates are extended with an additional argument representing the time at which they are to be interpreted. Specifically, in the situation calculus, the temporal argument denotes a situation, i.e., 'the complete state of the universe at an instant of time' (McCarthy and Hayes, 1969). 'Propositional fluents' (i.e. functions from situations to truth-values) are used to describe facts. For instance, a situation s in which a person p is in place x and it is raining in place x can be described as follows:

$$\text{at}(p, x, s) \wedge \text{raining}(x, s) \tag{1}$$

Causation is modeled using the propositional fluent $F(a, s)$ stating that situation s will be followed by a situation that satisfies the propositional fluent a. For instance, one could use F to assert that if a person is out in the rain he will get wet:

$$\forall x \forall p \ \forall s \ \text{raining}(x, s) \wedge \text{at}(p, x, s) \wedge \\ \text{outside}(p, s) \Rightarrow F(\lambda s' \ \text{wet}(p, s'), s) \tag{2}$$

Actions are modeled using the 'situational fluent' *result*. result(p, a, s) denotes the situation that results when the person p carries out the action a in the situation s. For instance, the following formula represents the axiom schema that, if in a situation s a person p has a key k that fits the safe sf, then in the situation resulting from the action of opening sf with the key k, the safe is open.

$$\text{has}(p, k, s) \wedge \text{fits}(k, sf) \wedge \text{at}(p, sf, s) \\ \Rightarrow \text{open}(sf, \text{result}(p, \mathit{open}(sf, k), s)) \tag{3}$$

Action and Change

Difficult problems arise when dealing with the interplay between actions and changes of the world. Many researchers have realised that one is not able to infer everything about a situation resulting from an action without having a very large number of axioms describing how the action relates to the world and changes it. First, one has to specify all the conditions that a situation s must satisfy in order that an action a may be performed in that situation. The number of such conditions tends to be very large even for simple actions (for example, I can drive a car if I have a car, a driving license, petrol in it … but also I must have hands, I must not be blind and the car must have wheels …). This is the so-called 'qualification problem'. Secondly, one has to specify everything that changes as a result of the action (if I am driving, I am moving, the the atmosphere is changing because of the exhaust produced by the car …). This is called the 'ramification problem'. The specification of everything that does not change is called the 'frame problem'. In classical first-order logic, it is not easy to express notions such as 'everything that is not explicitly asserted to change remains unchanged'. Dealing with the qualification, ramification and frame problems involves a certain capacity for 'nonmonotonic' reasoning. That is, one must be able to make inferences not only from known facts about the world (situation), but also from assumed facts. For instance, if I parked my car in front of my house yesterday, I can assume that the car is still there today, and draw conclusions on the basis of such an assumption (called a 'persistence assumption'). However, such conclusions are defeasible, in the sense that they may have to be retracted if further information about the world becomes available (for example, the police telephone to advise me that my car has been stolen). In the AI literature, many different logical frameworks have been devised in order to deal with nonmonotonicity (see the survey in Lee, Tannok and Williams, (1993). In recent years, many approaches have been devised that deal with the above problems by taking advantage of the 'negation by failure' of logic programming. For instance, a formulation of the situation calculus based on logic programming that deals with the frame problem has been recently proposed (Kowalski and Sadri, 1997).

Temporal (Modal) Logic

Another popular way of introducing the notion of time into a logical framework is by means of modal logics dealing with time (usually called 'temporal logics'). Modal logics were originally introduced to study the different 'modes' of truth. For example, a property p may be false in the present world, and yet the assertion 'possibly p' may be true if there is another possible world where p is true. A temporal logic is a type of modal logic in which each possible world is associated with a temporal element. A temporal precedence relation (akin to the 'accessibility' relation in modal logics) relates temporal elements, and the standard modal operators are reinterpreted over the temporal context. Any temporal logic must provide some temporal operators in order to describe and reason about how the truth-values of assertions vary with time. Different temporal logics have been devised to deal with

linear time or branching time; and both propositional and first-order temporal logics have been devised. A simple example of a temporal logic is propositional linear temporal logic (PLTL) (Emerson, 1990). PLTL adopts linear time, and introduces the temporal operators Fp ('sometime p'), Gp ('always p'), Xp ('next time p') and pUq ('p until q'). Given the current situation s_{now}, the above assertions evaluate to true if and only if the following conditions hold:

Fp: p is true in at least one of the situations that follow s_{now}.

Gp: p is true at s_{now} and in all situations following s_{now}.

Xp: p is true in the situation that immediately follows s_{now}.

pUq: p is true at s_{now} and in all situations between s_{now} and the first situation (following s_{now}) in which q is true.

A survey of different temporal logics can be found in Emerson, 1990.

CONSTRAINT-BASED APPROACHES

Constraint-based approaches mainly focus on the definition of a representation formalism and of reasoning techniques to deal specifically with temporal constraints between temporal entities (time points or intervals), independently of the events and states associated with them. For instance, given three time intervals I_1, I_2 and I_3, if I_1 is before I_2 and I_2 is before I_3, then one can infer that I_1 is before I_3, independently of the events that occurred in I_1, I_2 and I_3. By restricting the problem in this way, one can obtain better results than general-purpose approaches. For example, with a careful definition of the temporal constraint language, one can define specialized reasoning techniques that make inferences such as the one above in a more efficient way than, say, a standard theorem-prover for first-order logic. There is a trade-off between the expressiveness of the constraint language and the computational complexity of the inference techniques that are possible with it, and this is a central topic of research (see the survey in (Vila, 1994)).

Because dealing with temporal constraints is a fundamental task in so many AI applications, there has been much research aimed at building application-independent and domain-independent temporal constraints managers. These are intended to be *specialized knowledge servers* that represent and reason with temporal constraints, and that cooperate with other software modules in order to solve problems in different applications (Console and Terenziani, 1999). For instance, a temporal constraints manager could cooperate with a planner, in order to check incrementally the temporal consistency of a plan. Specialized temporal constraints managers are very useful from the computational point of view, as they allow programmers to focus on problems specific to their domain and application and to design modular software architectures. Temporal constraints managers can be distinguished on the basis of the types of temporal constraints they deal with.

Representing and reasoning with qualitative temporal constraints

Qualitative temporal constraints concern the relative positions of time points or intervals. J.F. Allen's interval algebra (IA) was an attempt to formalize these. Allen defined the 13 primitive qualitative relations between time intervals shown in Figure 1 (Allen, 1983).

These relations are exhaustive and mutually exclusive, and can be combined to represent ambiguities. For example, the constraint I_1 (BEFORE, CONTAINS,OVERLAPS)I_2 represents the fact that the time interval I_1 is before I_2, overlaps I_2 or contains I_2.

Temporal inference is based on two algebraic operations on relations between time intervals: intersection and composition. Given two possibly ambiguous relations R_1 and R_2 between two time intervals I_1 and I_2, the temporal intersection $R_1 \cap R_2$ is the most constraining relation between I_1 and I_2. For example, the temporal intersection between

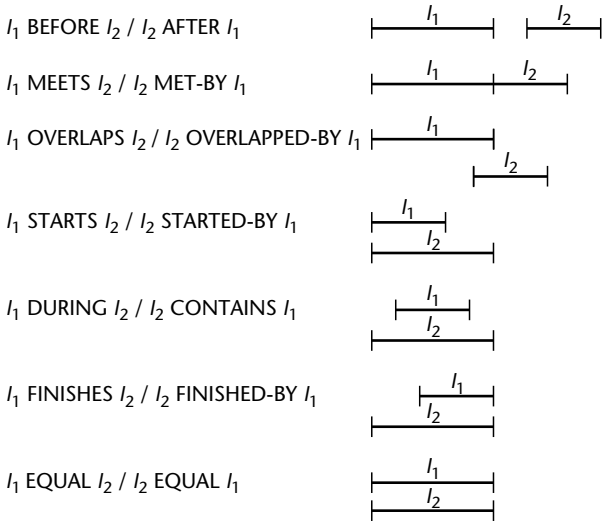


Figure 1. Qualitative relations between intervals in Allen's Interval Algebra.

$$R_1 : I_1(\text{BEFORE}, \text{CONTAINS}, \text{OVERLAPS})I_2 \quad (4)$$

and

$$R_2 : I_1(\text{BEFORE}, \text{MEETS}, \text{OVERLAPS})I_2 \quad (5)$$

is

$$R_1 \cap R_2 : I_1(\text{BEFORE}, \text{OVERLAPS})I_2 \quad (6)$$

Given a relation R_1 between I_1 and I_2 and a relation R_2 between I_2 and I_3, the composition $R_1@R_2$ gives the resulting relation between I_1 and I_3. For example, the composition of

$$R_1 : I_1(\text{BEFORE})I_2 \quad (7)$$

and

$$R_2 : I_2(\text{BEFORE}, \text{CONTAINS}, \text{OVERLAPS})I_3 \quad (8)$$

is

$$R_1@R_2 : I_1(\text{BEFORE})I_3 \quad (9)$$

In Allen's approach, temporal inference is performed by a path consistency algorithm that computes the transitive closure of the constraints by repeatedly applying intersection and composition. Abstracting from many optimizations, such an algorithm can be schematized as follows. Let R_{ij} denote the (possibly ambiguous) relation between I_i and I_j. For all triples of time intervals $< I_i, I_k, I_j >$, do $R_{ij} \leftarrow R_{ij} \cap (R_{ik}@R_{kj})$.

The time taken by Allen's algorithm increases as the cube of the number of time intervals. However, such an algorithm is not complete for IA (checking the consistency of a set of temporal constraints in IA is NP-hard). Other researchers have tried to design less expressive but more tractable formalisms. For example, the point algebra (PA) is defined similarly to IA, but the temporal elements are time points. Thus, there are only three primitive relations ($<$, $=$, and $>$), and four ambiguous relations (($<$, $=$), ($>$, $=$), ($<$, $>$), and ($<$, $=$, $>$)) between them. In PA, the constraint closure can be computed in polynomial time by algorithms that are both sound and complete. The price to be paid for tractability is a loss of expressive power: not all (ambiguous) relations between time intervals can be mapped onto relations between their endpoints. Another interesting algebra is the continuous point algebra (CPA), which consists of all relations in PA except for the inequality relation ($<$, $>$). Allen's path consistency algorithm is both correct and complete for such an algebra (see the survey in Vila, 1994).

Representing and reasoning with quantitative temporal constraints

Quantitative temporal constraints involve metric time and include *dates* ('John arrived on 10 October 1999 at 10.00'), *durations* ('John worked for 3 hours') and *delays* ('John arrived 10 minutes after Mary'). Different approaches have been developed to deal with quantitative temporal constraints, and their complexity depends on the types of constraints taken into account. The simplest case is when temporal information is available in the form of dates that exactly locate points and intervals on the timeline. However, in many cases, metric temporal information is not so precise: one can have approximate dates, durations, and delays:

The time interval I_1 ended on 10 October 1999, between 10.00 and 10.15. (10)

I_1 lasted between 20 and 30 minutes. (11)

I_2 started between 20 and 40 minutes after the end of I_1. (12)

In such cases, temporal reasoning is important in order to infer new temporal constraints and to detect inconsistencies. For instance, from the three constraints above one can infer that I_2 started on 10 October 1999, between 10.20 and 10.55, so that the set of constraints above is inconsistent with the constraint

I_2 started at 10.10. (13)

Problems become more complex if disjunctions of temporal constraints are taken into account:

I_2 started $20-30$ or $50-60$ minutes after the end of I_1. (14)

Many approaches have been developed to deal with these problems. For instance, the temporal constraint satisfaction problem (TCSP) model (Dechter, Meiri and Pearl, 1991) is based on the primitive notions of time points and distances between time points. A TCSP is a set of constraints of the form

$$P_i([l_1, u_1], \ldots, [l_n, u_n])P_j \quad (15)$$

where P_i and P_j denote time points ranging over a dense domain, and l_I and u_i ($l_i \leq u_i$) are real numbers, stating that the temporal distance between P_i and P_j is between l_i and u_i for some i. For instance, if we denote by I^- and I^+ the starting and ending points of a time interval I, the temporal constraint sentence 14 can be represented as:

$$I_1{}^{+}([20,30],\ [50,60])I_2{}^{-} \quad (16)$$

TCSP can easily be applied to graphs, where nodes represent time points and arcs are labelled with distances. Dechter, Meiri and Pearl developed an optimized path-finding algorithm to perform temporal reasoning with a graph representation. Their algorithm takes exponential time. However, if disjunctions of distances are not allowed (i.e., all constraints are of the form $P_i([l_1, u_1])\ P_j$), the resulting constraint problem (called the simple temporal problem (STP)) can be solved in cubic time using a standard algorithm for finding the shortest path between each pair of nodes in the graph (such as Floyd Warshall's algorithm, see Papadimitriou and Steiglitz, 1982). These algorithms are both correct and complete for STP. Finally, notice that constraints (10–12) (but not constraint (14)) can be easily mapped onto an STP. In particular, dates can be represented as distances from a given reference time which is unique for the overall set of constraints.

Hybrid approaches

Recently, various integrated approaches have been devised in order to deal with both qualitative and quantitative temporal constraints. Some researchers have proposed combining a standard specialized manager dealing only with qualitative temporal constraints with another manager dealing only with quantitative temporal constraints, and to let the two managers cooperate by exchanging pieces of information whenever new constraints are inferred. However, it has not been proven that the exchange of information between the two managers ever terminates.

Others have proposed high-level constraint languages in which both qualitative and quantitative temporal constraints can be represented, using reasoning techniques based on a homogeneous constraint framework. For example, in the LaTeR temporal constraint manager (Console and Terenziani, 1999), the high-level language allows one to deal with both time points and time intervals, and to express both quantitative and qualitative temporal constraints. The language has been carefully restricted so that all the constraints can be mapped onto an STP framework. In particular, this means that LaTeR can deal with quantitative temporal constraints such as those in sentences 10–12 (but not sentence 14), with all qualitative constraints between points except the inequality relation (i.e., with CPA), and with all qualitative constraints between time intervals that can be mapped onto CPA. While the user interacts with LaTeR through the high-level language, LaTeR internally translates all constraints into an STP framework, in which it performs temporal reasoning.

Recently (Jonsson and Backstrom, 1998), a homogeneous framework has been proposed, based on linear programming, that deals with all the types of constraints discussed above, and that also allows one to express constraints on the relative durations of events, such as: 'John takes at least 30 minutes more than Fred to drive to work.'

FURTHER ISSUES

L. Vila's survey (Vila, 1994) gives an overview of other issues concerning temporal reasoning techniques.

Terenziani has presented a high-level language to model user-defined periodicity (e.g., 'the first Monday of each month'), and a constraint language which extends IA to consider qualitative relations between periodic events (Terenziani, 1997). Terenziani's approach deals with constraints such as: 'Between 1 January 1999 and 31 December 1999, on the first Monday of each month, Mary went to the post office before going to work.' Temporal reasoning over such constraints can be performed by a path consistency algorithm which extends Allen's one. Such an algorithm is sound and operates in polynomial time, but is not complete for the high-level constraint language.

SUMMARY

Many logical approaches have been devised by AI researchers to deal with time, action and change, including the situation calculus, modal temporal logics and nonmonotonic logics. Other approaches have focused on the treatment of qualitative or quantitative temporal constraints, or both, in order to provide efficient domain-independent representation and reasoning techniques.

References

Allen JF (1983) Maintaining knowledge about temporal intervals. *Communications of the ACM* **26**(11): 832–843.

Console L and Terenziani P (1999) Efficient processing of queries and assertions about qualitative and quantitative temporal constraints. *Computational Intelligence* **15**(4): 442–465.

Dechter R, Meiri I and Pearl J (1991) Temporal constraint networks. *Artificial Intelligence* **49**: 61–95.

Emerson AE (1990) Temporal and modal logic. In: Van Leeuwen (ed.) *Handbook of Theoretical Computer Science*, **vol. B**, pp. 997–1072. Amsterdam: Elsevier.

Jonsson P and Backstrom C (1998) A unifying approach to temporal constraint reasoning. *Artificial Intelligence* **102**: 143–155.

Kowalski R and Sadri F (1997) Reconciling the event calculus with the situation calculus. *The Journal of Logic Programming* **31**(1–3): 34–58.

Lee H, Tannok J and Williams JS (1993) Logic-based reasoning about actions and plans in artificial intelligence. *Knowledge Engineering Review* **8**(2): 91–120.

McCarthy J and Hayes PJ (1969) Some philosophical problems from the standpoint of AI. *Machine Intelligence* **4**: 463–502.

Papadimitriou C and Steiglitz K (1982) *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs, NJ: Prentice Hall.

Terenziani P (1997) Integrating calendar-dates and qualitative temporal constraints in the treatment of periodic events. *IEEE Transactions on Knowledge and Data Engineering* **9**(5): 763–783.

Van Benthem J (1983) *The Logic of Time*. Dordrecht: Kluwer.

Vila L (1994) A survey on temporal reasoning in artificial intelligence. *AI Communications* **7**(1): 4–28.

Further Reading

Allen JF (1991) Time and time again: the many ways to represent time. *Intelligent Systems* **6**(4): 341–355.

Hajnicz E (1995) *Time structures. Lecture Notes in Artificial Intelligence 1047*. Berlin: Springer.

Kowalski R and Sergot M (1986) A logic-based calculus of events. *New Generation Computing* **4**: 67–95.

Lifschitz V (ed.) (1997) *Journal of Logic Programming* **31**(1–3). Special issue on reasoning about action and change.

Long D (1989) A review of temporal logics. *Knowledge Engineering Review* **4**(2): 141–162.

Sandewall E (1994) *Features and Fluents: The Representation of Knowledge About Dynamical Systems*. Oxford: Oxford University Press.

Snodgrass RT (ed.), Ahn I, Ariav G, Batory D *et al*. (1995) *The Temporal Query Language TSQL2*. Boston, MA: Kluwer.

Vilain M and Kautz H (1986) Constraint propagation algorithms for temporal reasoning. In: Kehler T, Rosenschein S, Filman R and Patel-Schneider P (eds), *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI'86)*, pp. 377–382. Philadelphia, PA.

Vilain M, Kautz H and van Beek P (1990) Constraint propagation algorithms for temporal reasoning: a revised report. In: Weld DS and de Kleer J (eds) *Readings in Qualitative Reasoning About Physical Systems*, pp. 373–381. San Mateo, CA: Morgan Kaufmann.

Reasoning under Uncertainty

Intermediate article

Francisco J Díez, UNED, Madrid, Spain
Marek J Druzdzel, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Most artificial intelligence applications, especially expert systems, have to reason and make decisions based on uncertain data and uncertain models. For this reason, several methods have been proposed for reasoning with different kinds of uncertainty.

INTRODUCTION

We often have to make decisions based on uncertain knowledge, not only in our private lives (which job to take, which house to buy, where to invest our money) but also in professional activities, such as medicine, economics, politics, engineering, and education. Therefore, any reasoning method that tries to replicate human reasoning must be able to draw conclusions from uncertain models and uncertain data. Models may be uncertain because of indeterminism in the real world or because of our lack of knowledge. Furthermore, data may be incomplete (pieces of information may be not available in a diagnostic case), ambiguous (a pronoun in a sentence may refer to different subjects), erroneous (patients may lie to their doctors, or sensors may be faulty), or imprecise (because of the limited precision of measuring devices, subjective estimations, or natural language).

This article reviews some of the uncertain reasoning methods that have been proposed in the field of artificial intelligence.

NAIVE BAYES

The oldest method applied in uncertain reasoning is probability theory. Probabilistic reasoning concentrates basically on computing the *posterior probability* of the variables of interest given the available *evidence*. In medicine, for example, the evidence consists of symptoms, signs, clinical history, and laboratory tests. A diagnostician may be interested in the probability that a patient suffers from a certain disease. In mineral prospecting, we may wish to know the posterior probability of the presence of a certain deposit given a set of geological findings. In computer vision, we might be interested in the probability that a certain object is present in an image given observation of certain shapes or shadows.

The probability of the diagnoses given the available evidence can be computed by the generalization of Bayes' theorem to several variables. However, the direct application of this method would need a prohibitive number of parameters (probabilities), which grows exponentially with the number of variables involved. Two assumptions were introduced to simplify the model. The first assumption is that the diagnoses are mutually exclusive; i.e. each patient can suffer from at most one disease and each device can have at most one failure at a time. It is then possible to consider a variable D taking n values, as many as the number of possible diagnoses. The second assumption is that the findings are conditionally independent given each diagnosis d_i, so:

$$P(f_1,\ldots,f_m \mid d_i) = P(f_1 \mid d_i)\cdot\ldots\cdot P(f_m|d_i), \quad 1 \le i \le n \tag{1}$$

where f_k represents one of the possible values of a finding F_k. Under these assumptions, the posterior probability of d_i can be computed as follows:

$$P(d_i \mid f_1,\ldots,f_m) = \frac{P(d_i)\cdot P(f_1|d_i)\cdot\ldots\cdot P(f_m|d_i)}{\sum_j P(d_j)\cdot P(f_1|d_j)\cdot\ldots\cdot P(f_m|d_j)} \tag{2}$$

In this simplified model, the number of parameters is proportional to the number of variables: it requires n prior probabilities $P(d_i)$ plus, in the case of m dichotomous findings, $n \times m$ conditional probabilities $P(f_k \mid d_i)$.

Although this method was used in the construction of diagnostic medical systems in the 1960s, it was severely criticized because its assumptions are usually unrealistic (Szolovits and Pauker, 1978). In fact, the assumption of exclusive diagnoses is a reasonable approximation only when the probability of the simultaneous presence of two diseases or two failures is very low. In medicine, however, it is common for a patient to suffer from multiple disorders. Also, the assumption of conditional independence is unrealistic when there are causal associations among findings other than those due to the diagnoses included in the model. Because of the crudeness of such assumptions, this method is often called 'naive Bayes' or 'idiot Bayes'. Even under these assumptions, the model still requires a large number of parameters, which may be difficult to obtain.

MYCIN'S CERTAINTY FACTORS

MYCIN was a rule-based expert system developed in the 1970s as a decision support tool for antibacterial therapy (Buchanan and Shortliffe, 1984). To accommodate uncertainty, MYCIN associated a certainty factor with each rule and, consequently, with each proposition.

The certainty factor of each rule 'if E then H', $CF(H, E)$, is a measure of the degree to which evidence E confirms hypothesis H. When E increases the probability of H, so that $P(H|E) > P(H)$, then $0 < CF(H, E) \leq 1$. The higher the increase in probability, the higher the certainty factor. When E contributes evidence against H, so that $P(H|E) < P(H)$, then $-1 \leq CF(H, E) < 0$. The value of each certainty factor in MYCIN's rules was obtained from human experts when formulating the rules.

Analogously, MYCIN assigned a certainty factor to each assertion. The result was a set of quadruplets of the form illustrated in Table 1. Thus we know with absolute certainty that the patient's name is John Smith; there is strong evidence indicating that the form of organism-1 is rod, weak evidence that it is staphylococcus, weak evidence that it is not a streptococcus, and absolute certainty that the form of organism-2 is not rod.

When the user introduces a piece of evidence, such as $A =$ 'the form of organism-1 is rod', MYCIN states that $CF(A) = 1$. Given the rule

Table 1. Certainty factors of assertions as represented in MYCIN

Object	*Attribute*	*Value*	*CF*
patient	name	John Smith	1.0
organism-1	morphology	rod	0.8
organism-1	identity	staphylococcus	0.2
organism-1	identity	streptococcus	−0.3
organism-2	morphology	rod	−1.0

'if A then B' with certainty factor $CF(B, A)$, the certainty factor for B can be computed as $CF(B) = CF(A) \cdot CF(B, A)$. If there is a second rule 'if B then C', then $CF(C) = CF(B) \cdot CF(C, B)$. This value $CF(C)$ might, in turn, be used in the application of a third rule 'if C then D', and so on.

There was also an equation for combining convergent rules, such as 'if E_1 then H' and 'if E_2 then H', which support the same hypothesis H. The certainty factor of a composed antecedent, such as 'if A and not B', was computed by applying these equations:

$$CF(\text{not } E) = 1 - CF(E) \tag{3}$$

$$CF(E_1 \text{ and } E_2) = \min(CF(E_1), CF(E_2)) \tag{4}$$

$$CF(E_1 \text{ or } E_2) = \max(CF(E_1), CF(E_2)) \tag{5}$$

Although the performance of MYCIN was comparable to that of human experts in the field of infectious diseases, the certainty factor model was soon criticized for its mathematical inconsistencies. One of them is that it does not consider correlation between propositions. For instance, if there are two hypotheses, $H_1 =$ 'organism-1 is a streptococcus', with $CF(H_1) = 0.6$, and $H_2 =$ 'organism-1 is a staphylococcus', with $CF(H_2) = 0.3$, then $CF(H_1 \text{ and } H_2) = \min(0.6, 0.3) = 0.3$, whereas it should be $CF(H_1 \text{ and } H_2) = 0$, because the hypotheses are mutually exclusive.

Another problem is the lack of sensitivity in eqns 4 and 5; for example $\min(0.2, 0.9) = \min(0.2, 0.2)$ and $\max(0.9, 0.9) = \max(0.9, 0.2)$.

Furthermore, MYCIN might assign a higher certainty factor to a hypothesis H_2 even if it is less probable than another hypothesis H_1. This anomaly could occur because certainty factors of rules were defined as measures of confirmation (the relative increase in belief), while certainty factors of propositions were interpreted as measures of absolute belief.

It was also pointed out that the combination of convergent rules is valid only under some conditions. These conditions resemble those of

conditional independence in the naive Bayes model. However, while the assumption of conditional independence might in particular cases be justified by means of causal arguments, no argument can be made for the assumptions that are implicit in the combination of rules. Therefore, MYCIN's model, rather than solving the main problem of the naive Bayes, ran into more serious troubles (see Buchanan and Shortliffe, 1984, ch. 10–12; Pearl, 1988, sec. 1.2). (*See* **Learning Rules and Productions**; **Expert Systems**; **Resolution Theorem Proving**; **Deductive Reasoning**; **Rule-based Thought**)

PROSPECTOR'S BAYESIAN MODEL

PROSPECTOR was an expert system for geological prospecting developed in the 1970s (Duda *et al.*, 1976). Like MYCIN, PROSPECTOR used if–then rules, but instead of using ad hoc certainty factors, it was based on the theory of probability. Each rule had two parameters, namely the likelihood ratios, which were obtained from human experts' estimations.

The propagation of evidence in PROSPECTOR consisted in computing the probability of the consequence of a rule given the probability of the antecedent. In order to simplify the computation, PROSPECTOR made several assumptions of conditional independence, in addition to approximations and interpolations aimed at smoothing the inconsistencies in the probabilities elicited from human experts.

PROSPECTOR became the first commercial success of artificial intelligence when it assisted in the discovery of a deposit of molybdenum worth about one million dollars. However, this achievement did not lessen criticism of the use of probabilistic methods in expert systems. The assumptions and approximations required by PROSPECTOR were still largely unjustified.

DEMPSTER–SHAFER THEORY

In 1968, Dempster proposed a probabilistic framework based on lower and upper bounds on probabilities. In 1976, Shafer developed a formalism for reasoning under uncertainty that used some of Dempster's mathematical expressions, but gave them a different interpretation: each piece of evidence (finding) may support a subset containing several hypotheses. This is a generalization of the 'pure' probabilistic framework in which every finding corresponds to a value of a variable (a single hypothesis).

The main criticism of this theory from a semantic point of view is the lack of robustness of the combination of evidence. Given three single hypotheses and two findings, it may happen that a hypothesis receiving almost no support from any individual finding is confirmed by the combination of them, while the other two hypotheses are discarded (Zadeh, 1986). Also, a small modification of the evidence assignments may lead to a completely different conclusion. This paradox poses no problem for Dempster's (1968) interpretation (lower and upper probabilities) or to Pearl's (1988, sec. 9.1) interpretation (probability of provability), but seems counterintuitive in Shafer's (1976) interpretation, and for this reason some researchers have proposed alternative formalisms based on different combination rules.

The main problem of the Dempster–Shafer theory in its original formulation is that its computational complexity grows exponentially with the number of hypotheses. One of the solutions proposed consists in building a network of frames of discernment (in fact, a network of random variables), whose axiomatic definition is reminiscent of the properties of conditional independence of Bayesian networks.

BAYESIAN NETWORKS

A Bayesian network (Pearl, 1988) is a probabilistic model that consists of a finite set of random variables $\{V_i\}$ and an acyclic directed graph whose nodes represent those variables and whose arcs represent, roughly speaking, probabilistic dependences between variables.

Dependences are quantified by means of a set of conditional probability distributions (CPDs). Each CPD involves a node and its parents in the graph: $P(v_i \mid pa(v_i))$. (Node V_i is a parent of V_j if there is a link $V_i \rightarrow V_j$ in the graph.) In the case of discrete variables, each CPD is given by atable of probabilities. The product of all the CPDs is a joint probability over all variables, from which it is possible to obtain any other marginal or conditional probability, such as the *a posteriori* probability of any variable given a set of evidence.

There are basically two ways of building a Bayesian network. The *automatic* process involves taking a database and applying one of the many algorithms that yield both the structure and the conditional probabilities. The *manual* process involves two stages: building the structure of the network by selecting the variables and drawing causal links among nodes, as in Figure 1; and then

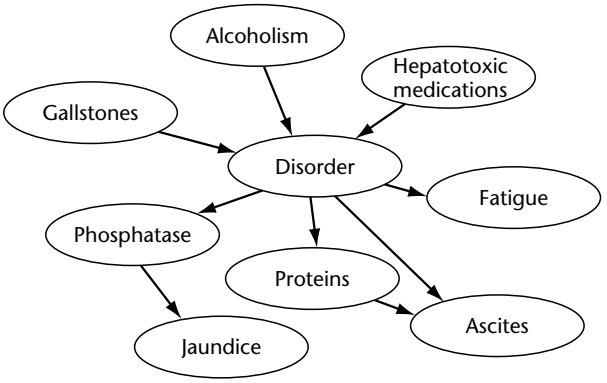


Figure 1. A simplified fragment of HEPAR-II, a medical Bayesian network.

estimating the corresponding conditional probability distributions. (*See* **Machine Learning**)

Ideally, those probabilities should be obtained from objective data, such as databases or epidemiological studies, but in practice the lack of objective data often forces the knowledge engineer to obtain the probabilities from human experts' estimations.

Bayesian networks overcome the limitations of the naive Bayes method in two ways. Firstly, they can diagnose the simultaneous presence of several diseases or failures, because each disease can be represented by a different node. Secondly, the properties of conditional independence are justified either by the statistical independencies in the database, in the case of automatic construction, or by the use of a causal graph elicited from a human expert.

Bayesian networks are also superior to PROSPECTOR in the justification of conditional independencies, but at the price of an increase in computational complexity (the time spent in computing the posterior probabilities).

Causal Bayesian networks have additional advantages, such as the ease with which the model can be extended or refined and its reasoning explained to users.

The main criticisms of Bayesian networks are the difficulty of building the network and the computational complexity of evidence propagation, which is NP-hard: the time required by exact probability updating algorithms depends mainly on the structure of the network, while the complexity of stochastic algorithms depends mainly on the numerical parameters (probabilities). (*See* **Computability and Computational Complexity**; **Bayesian Belief Networks**)

INFLUENCE DIAGRAMS

Influence diagrams are extensions of Bayesian networks which, in addition to random variables, capture available decision options and preferences (utilities). Random variables are represented by circles or ovals, decision nodes as squares or rectangles, and utility nodes as diamonds or parallelograms. Influence diagrams are decision support tools. They permit one to select the optimal decision, the decision that maximizes the expected utility. They overcome the limitations of Bayesian networks in their explicit representation of utilities and in the possibility of selecting the questions to ask or the tests to perform (goal-oriented reasoning). Influence diagrams are closely related to decision trees and Markov decision processes (see Pearl (1988) chapter 6). (*See* **Markov Decision Processes, Learning of**; **Decision-making**)

FUZZY LOGIC AND FUZZY SETS

Some of the sentences that we use in our daily life, such as 'it is cold today', are neither completely true nor completely false. These propositions are called *fuzzy*. In fact, most of the adjectives that we use daily could be interpreted as fuzzy predicates (e.g., 'young', 'rich', 'tall', 'happy', 'healthy', 'big', 'good', 'cheap', 'dark', 'crowded', 'heavy', 'fast', 'modern').

Since in classical logic it is usual to assign 0 to false propositions and 1 to true propositions, some logicians have built multivalued logics in which $v(p)$, the truth-value of proposition p, might also take values between 0 and 1. The truth-value of a composed proposition (negation, conjunction, disjunction, implication, etc.) is a function of the truth-values of the propositions that compose it: for instance, $v(p \wedge q) = f_{\wedge}(v(p), v(q))$.

Different choices of these logical functions lead to different logics. For instance, Lukasiewicz logic, Kleen logic, and standard fuzzy logic take the 'minimum' function for $f_{\wedge}$. Other fuzzy logics may use different triangular norms for $f_{\wedge}$. (A triangular norm is any function $f_{\wedge}$ that is commutative, associative, and monotone, and satisfies $f_{\wedge}(1, a) = a$.)

Similarly, there are several implication functions $f_{\rightarrow}$, all of which satisfy certain conditions. In principle, it would be possible to do inference with fuzzy propositions and fuzzy predicates, but in practice fuzzy inference is usually based on fuzzy sets and fuzzy relations, as shown below.

Given a set A and an element x, the truth-value of the proposition '$x \in A$' is usually called *membership*

degree and is represented by $\mu_A(x)$. Whereas in the case of crisp (classical) sets $\mu_A(x)$ is either 0 or 1, in the case of fuzzy sets $\mu_A(x)$ may be any number in the interval [0, 1].

Each operation on fuzzy sets corresponds to a fuzzy-logical operation. For instance, intersection corresponds to conjunction, since $\mu_{A\cap B}(x) = v(x\in A \wedge x\in B) = f_\wedge(\mu_A(x), \mu_B(x))$. Therefore, fuzzy logic can be viewed as the basis of fuzzy set theory. However, it is more usual to view fuzzy set theory as the basis for fuzzy logic.

The rule 'if $P_A(x)$ then $P_B(y)$', where A and B are fuzzy sets and P_A and P_B are their associated predicates, is translated into a fuzzy relation given by $\mu_{A\to B}(x,y) = f_\to(\mu_A(x), \mu_B(y))$. Modus ponens consists in combining this rule with an assertion $P_{A'}(x)$ in order to obtain a new set B'. This process is performed by composing the set A' with the relation $\mu_{A\to B}$. The resulting set B' depends on the logical functions involved in the composition. In many applications of fuzzy logic this choice is made more or less arbitrarily – a typical choice is min as a norm, max as a conorm and Lukasiewicz's implication – and leads to inconsistencies and counterintuitive results (Fukami *et al*., 1980). The correct way to approach this problem consists in determining the desirable properties of fuzzy inference and then selecting $f_\vee$, $f_\wedge$, and $f_\to$ coherently in order to ensure such properties (see, for instance, Trillas and Valverde, 1985).

The main criticism of fuzzy logic is the lack of a clear semantics, which leads to an arbitrariness in the application of fuzzy techniques. In particular, there are several definitions of degrees of membership, but apparently none of them is used when building real-world applications. All fuzzy systems use numbers between 0 and 1, but the semantics of those numbers and the way of assigning them differ significantly from application to application. Accordingly, there is no clear criterion for determining which norm or conorm to use in each case.

Additionally, there are several techniques of fuzzy reasoning. We have already mentioned that the properties of the fuzzy inference depend on the choice of logical functions, and knowledge engineers are often unaware of the inconsistencies that may result from an arbitrary choice. There are also other patterns of inference with fuzzy rules that are not based on the composition of relations, and other reasoning techniques, such as those involving fuzzy numbers and fuzzy clustering, not presented in this article. As a result, fuzzy logic consists in practice of a toolbox of heterogeneous techniques without clear indications for deciding which tool to use in any particular case. Users of fuzzy logic often devise ad hoc solutions for the representation and combination of knowledge and data. (*See* **Fuzzy Logic**; **Vagueness**)

ROUGH SETS

Rough sets (Pawlak, 1991) have been developed since the 1980s as a tool for data analysis. In simple terms, the starting point is a data table which represents, for each object, the values of some attributes and a label c: $(a_1, \ldots, a_n, c)$. The final objective of the analysis is to infer some classification rules of the form: 'If the attributes of a new object take values $(a_1, \ldots, a_n)$ then this object is c.'

Both the theoretical foundations and the interpretation of rough sets are completely different from those of fuzzy sets. The lack of a precise boundary of a fuzzy set is a consequence of the vagueness of some concepts, such as 'big' or 'tall', and does not necessarily entail uncertainty. In contrast, the lack of a precise boundary of a rough set derives from the coarseness of the background knowledge implicit in the data table. However, the two theories are complementary and are usually categorized together as 'soft computing'.

NON-MONOTONIC LOGICS

Non-monotonic logics are an attempt to model a pattern of human reasoning that consists in making plausible, although fallible, assumptions about the world, and revising them in the light of new evidence. Upon hearing that 'Tweety is a bird', one might assume that 'Tweety can fly' just because 'most birds fly'; however, further evidence that 'Tweety is a penguin' will lead one to retract this assumption. The name 'non-monotonic' refers to the fact that the addition of new clauses may lead to either adding or retracting a previous conclusion. In classical logic new information never invalidates previous conclusions.

Several non-monotonic logics have been proposed, such as Reiter's logic, McCarthy's circumscription, Doyle's truth maintenance systems, and Cohen's theory of endorsements, each based on some implicit assumptions about uncertainty and preferences. Virtually all have been shown to lead to undesirable behavior under some circumstances.

All these logics are based on qualitative reasoning patterns, and thus differ from the other methods described above, which use and propagate numerical information. Qualitative models are easier to build, because they do not need to estimate quantitative parameters. On the

other hand, they are unable to weight conflicting evidence. (*See* **Quantitative Reasoning**)

CONCLUSION

Probability is the oldest formalism for reasoning under uncertainty, and served as the first foundation for computer-based reasoning systems in the 1960s. The unrealistic assumptions of the naive Bayes model used in the first diagnostic systems, the desire to differentiate the evidence in favor of a hypothesis from the evidence against it, and the need for a rule-based reasoning method, led to the development of MYCIN's model of certainty factors. It was later shown that this model had serious inconsistencies and required even more unrealistic assumptions than those of naive Bayes. PROSPECTOR used a Bayesian framework for reasoning with rules, but it also relied on unjustified assumptions and approximations. Dempster–Shafer theory was an attempt to overcome some of the limitations of probabilistic reasoning, but its computational complexity prevented it from being used in practice. Fuzzy sets and fuzzy logic emerged as tools for representing the vagueness of natural language, and led to numerous applications in engineering, medicine, and many other fields.

These four methods – MYCIN, PROSPECTOR, Dempster–Shafer, and fuzzy sets – were developed in the 1970s. In those years, many computer scientists were convinced that probability was not an adequate framework for the problems addressed by artificial intelligence and expert systems. However, the emergence of Bayesian networks and influence diagrams in the 1980s proved that it was possible to build probabilistic models for real-world problems. Many expert systems and software packages based on these techniques became commercial products in the 1990s.

Some researchers nowadays take the position that probability is the only correct framework for uncertain reasoning. Others, while agreeing that probability theory is the best technique when there is enough statistical information, argue that it is hard to use in many practical cases, and for this reason artificial intelligence still needs alternative formalisms.

References

Buchanan BG and Shortliffe EH (eds) (1984) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Dempster AP (1968) A generalization of Bayesian interference. *Journal of the Royal Statistical Society*, series B **30**: 205–247.

Duda RO, Hart PE and Nilsson NJ (1976) Subjective Bayesian methods for rule-based interference systems. *Proceedings of the AFIPS National Computer Conference*, New York **45**: 1075–1082.

Fukami S, Mizumoto M and Tanaka K (1980) Some considerations on fuzzy conditional interference. *Fuzzy Sets and Systems* **4**: 243–273.

Pawlak Z (1991) *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht, Netherlands: Kluwer.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann. [Revised second printing.]

Shafer G (1976) *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.

Szolovits P and Pauker SG (1978) Categorical and probabilistic reasoning in medicine. *Artificial Intelligence* **11**: 115–144.

Trillas E and Valverde L (1985) On mode and implication in approximate reasoning. In: Gupta MM, Kandel A, Bandler W and Kiszka JB (eds), *Approximate Reasoning and Experts Systems*, pp. 157–166. Amsterdam, Netherlands: North-Holland.

Zadeh LA (1986) A simple view of the Dempster–Shafer theory of evidence and its implication for the rule of combination. *AI Magazine* **7**: 85–90.

Further Reading

Castillo E, Gutiérrez JM and Hadi AS (1997) *Expert Systems and Probabilistic Network Models*. New York, NY: Springer-Verlag.

Cowell RG, Dawid AP, Lauritzen LS and Spiegelhalter DJ (1999) *Probabilistic Networks and Expert Systems*. New York, NY: Springer-Verlag.

Dubois D, Yager RR and Prade H (1993) *Readings in Fuzzy Sets for Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Jensen FV (2001) *Bayesian Networks and Decision Graphs*. New York, NY: Springer-Verlag.

Klir GJ and Yuan B (1995) *Fuzzy Sets and Fuzzy Logic. Theory and Applications*. Upper Saddle River, NJ: Prentice Hall.

Krause P and Clark D (1993) *Representing Uncertain Knowledge. An Artificial Intelligence Approach*. Oxford, UK: Intellect Books.

Shafer G and Pearl J (1990) *Readings in Uncertain Reasoning*. San Mateo, CA: Morgan Kaufmann.

Shapiro SC (ed.) (1992) *Encyclopedia of Artificial Intelligence*, 2nd edn. New York, NY: John Wiley.

Reinforcement Learning: A Computational Perspective

Intermediate article

Peter Dayan, University College London, UK
Christopher JCH Watkins, Royal Holloway College, London, UK

CONTENTS

Reinforcement learning is one of the means by which animals and artificial systems can learn to optimize their behaviour in the face of rewards and punishments. Reinforcement learning algorithms have been developed that are closely related to methods of dynamic programming, which is a general approach to optimal control. Reinforcement learning phenomena have been observed in psychological studies of animal behaviour, and in neurobiological investigations of neuromodulation and addiction.

INTRODUCTION

One way in which animals acquire complex behaviours is by learning to obtain rewards and to avoid punishments. Reinforcement learning (RL) theory is a formal computational model of this type of learning. To see that some theory is needed, consider the fact that in many environments animals need to perform unrewarded or unpleasant preparatory actions in order to obtain some later reward. For example, a mouse may need to leave the warmth and safety of its burrow to go on a cold and initially unrewarded search for food.

Is it possible to explain learning to take such an unpleasant decision in terms of learning to obtain a large but distant reward? What computational abilities does an agent need in principle for this type of learning? What capacity for episodic memory is necessary? Does an agent need to be able to plan ahead by envisioning future situations and actions? If an agent is doing something without immediate reward, how does it learn to recognize whether it is making progress towards a desirable goal? It is these questions that the theory of reinforcement learning seeks to answer.

RL theory comprises a formal framework for describing an agent interacting with its environment, together with a family of learning algorithms, and analyses of their performance.

THE REINFORCEMENT LEARNING FRAMEWORK

In the standard framework for RL, a learning agent – an animal or perhaps a robot – repeatedly observes the state of its environment, and then chooses and performs an action. Performing the action changes the state of the world, and the agent also obtains an immediate numeric payoff as a result. Positive payoffs are rewards and negative payoffs are punishments. The agent must learn to choose actions so as to maximize a long-term sum or average of the future payoffs it will receive.

The agent's payoffs are subjective, in the sense that they are the agent's own evaluation of its experience. The ability to evaluate its experience in this way is an essential part of the agent's prior knowledge. The payoffs define what the agent is trying to achieve; what the agent needs to learn is how to choose actions to obtain high payoffs. For this framework to be sensible, the agent must be able to visit the same or at least similar states on many occasions, to take advantage of its learning.

A standard assumption is that the agent is able to observe those aspects of the state of the world that are relevant to deciding what to do. This assumption is convenient rather than plausible: it leads to great simplification of the RL problem. We will not here consider the important problem of how the agent could learn what information it needs to observe.

In the conventional framework, the agent does not initially know what effect its actions have on the state of the world, nor what immediate payoffs its actions will produce. In particular it does not know what action it is best to perform. Rather, it

tries out various actions at various states, and gradually learns which one is best at each state so as to maximize its long-term payoff. The agent thus acquires what is known as a *closed-loop control policy*, or a rule for choosing an action according to the observed current state of the world.

From an engineering perspective, the most natural way to learn the best actions would be for the agent to try out various actions in various states, and so learn a predictive model of the effects its actions have on the state of the world and of what payoffs its actions produce. Given such a model, the agent could plan ahead by contemplating alternative courses of action and the states and payoffs that would result. However, this approach to learning does not seem at all plausible for animals. Planning ahead involves accurately envisioning alternative sequences of actions and their consequences: the animal would have to imagine what states of the world would result and what payoffs it would receive for different possible sequences of actions. Planning ahead is particularly difficult if actions may have stochastic effects, so that performing an action may lead to one of several different possible states.

One of the most surprising discoveries in RL theory is that there are simple learning algorithms by means of which an agent can learn an optimal policy without ever being able to predict the effects of its actions or the immediate payoffs they will produce, and without ever planning ahead. It is also surprising that, in principle, learning requires only a minimal amount of episodic memory: an agent can learn if it can consider together the last action it took, the state when it chose that action, the payoff received, and the current state.

The simplest RL algorithms require minimal episodic memory of past states and actions, no prediction of effects of actions, and very simple computations. The speed of learning, and efficiency of use of experience, can be improved if the agent has greater abilities. A continuum of improvement is possible. If an agent has or constructs partial models of the effects of its actions, or if an agent can remember and process longer sequences of past states and actions, learning can be faster and the agent's experience can be used more efficiently.

Markov Decision Problems

Formally, an RL agent faces a Markov decision problem (MDP). An MDP has four components: states, actions, and transition and reward distributions.

The state at time t, for which we use the notation $x(t)$, characterizes the current situation of the agent in the world. For an agent in a maze, for instance, the relevant state would generally be the location of the agent. The action the agent takes at time t is called $a(t)$. Sometimes there may be little or no choice of actions. One consequence of taking the action is the immediate payoff or reward which we call $r(x(t), a(t))$, or sometimes just $r(t)$. If the rewards are stochastic, then $\hat{r}(x(t), a(t))$ is its mean. The other consequence of taking the action is that the state changes. Such changes are characterized by a *transition* distribution, which allows them to be stochastic. In the simplest RL problems, state-transition distributions and the rewards depend only on the current state and the current action, and not on the history of previous actions and states. This restriction is called the Markov property, and ensures that the description of the current state contains all information relevant to choosing an action in the current state. The Markov property is critically necessary for the learning algorithms that will be described below.

We typically use a set of matrices $P_{xy}(a)$ to describe the transition structure of the world. Here, $P_{xy}(a)$ is the probability that the state changes from x to y given that action a is performed. Conventionally, the agent starts off not knowing the transition matrices or reward distributions. Over the course of repeated exploration, the agent has to work out a course of action that will optimize its return.

As an example of a Markov decision process, consider a hypothetical experiment in which a rat presses levers to obtain food in a cage with a light. Suppose that if the light is off, pressing lever A turns the light back on with a certain probability, and pressing lever B has no effect. When the light is on, pressing lever A has no effect, but pressing lever B delivers food with a certain probability, and turns the light off again. In this simple environment there are two relevant states: *light on* and *light off*. Lever A may cause a transition from *light off* to *light on*; in *light on*, lever B may yield a reward. The only information that the rat needs to decide what to do is whether the light is on or off. The optimal policy is simple: in *light off*, press lever A; in *light on*, press lever B.

The goal for an agent is solving a Markov decision problem is to maximize its expected, long-term payoff, known as its *return*. A mathematically convenient way to formalize return is as a discounted sum of payoffs. That is, starting from state $x(0) = x$ at time $t = 0$, it should choose actions $a(0), a(1), \ldots$ to maximize

$$\left\langle \sum_{t=0}^{\infty} \gamma^t r(x(t), a(t)) \right\rangle_{x,r} \quad (1)$$

where the $\langle\rangle_{x,r}$ indicates that an average is taken over the stochasticity in the states and the payoffs. The factor $0 \leq \gamma < 1$ is called the *discount factor* and controls the weighting of payoffs that happen sooner relative to those that happen later. The larger γ, the more important are distant payoffs, and, typically, the more difficult the optimization problem.

Other definitions of return are possible. A Markov decision problem is said to be *absorbing* if there is a state or set of states which define the end of a trial (like the goal of the maze), allowing a new trial to start. The infinite sum of expression 1 is effectively truncated at these states, which is equivalent to specifying degenerate transition matrices and reward distributions there. In this case, it is possible to learn to optimize the sum total payoff that the agent receives before reaching the terminal state.

If, on the other hand, the MDP is such that it is always possible to return to every state, it is possible to learn to optimize the average payoff. A policy that optimizes average payoff will also optimize discounted payoff for a value of γ sufficiently close to 1. RL algorithms can be adapted to use any of these definitions of return; we will consider only total discounted reward here. For convenience, we also assume that the problem is finite, i.e. there are only finitely many states and possible actions.

DYNAMIC PROGRAMMING

The obvious problem with optimizing a goal such as that in expression 1 is that the action taken at time 0 can affect the rewards at time $t \gg 0$. In effect, one has to take into consideration all possible sequences of actions. However, the number of such sequences of length t typically grows exponentially as a function of t, making it impractical to consider each one in turn. Dynamic programming (Bellman, 1957) offers a set of methods for solving the optimization problem while (generally) avoiding this explosion, by taking advantage of the Markov property. The two main methods are called policy iteration and value iteration, and we discuss them in turn.

Policy Iteration

A *policy* is an assignment of actions to states – e.g. a recipe that says: push lever A if the light is off and push lever B if the light is on. In general, policies can be stochastic, specifying the probability of performing each action at each state. It can be shown that the solution to the optimization problem of expression 1 can be cast in terms of a deterministic policy which is constant over time, so the agent performs the same action every time it gets to a given state. Policies without time-dependence are called *stationary*. We will, for the moment, consider stationary deterministic policies. We use the notation $\pi(x)$ for the action that policy π recommends at state x.

In policy iteration, a policy is first *evaluated*, and then *improved*. Policy evaluation consists of working out the value of every state x under policy π, i.e. the expected long-term reward obtained by starting from x and following policy π. That is

$$V^{\pi}(x) = \left\langle \sum_{t=0}^{\infty} \gamma^t r(t) \right\rangle_{x,r} \quad (2)$$

where the states follow from the transitions with $x(0) = x$ and $r(t) = r(x(t),\ \pi(x(t)))$. We can separate the first and subsequent terms in the sum,

$$V^{\pi}(x) = \hat{r}(x, \pi(x)) + \gamma \left\langle \sum_{t=0}^{\infty} \gamma^t r(t+1) \right\rangle_{x,r} \quad (3)$$

which, using the Markov property, is

$$= \hat{r}(x, \pi(x)) + \gamma \sum_{y} P_{xy}(\pi(x)) V^{\pi}(y) \quad (4)$$

The second term in both these expressions is just γ times the expected infinite-horizon return for the state at time $t = 1$, averaged over all the possibilities for this state. Eqn 4 is a linear equation for the values $V^{\pi}(y)$, and so can be solved by a variety of numerical methods. Later, we will see how to find these values without knowing the mean rewards or the transition distributions.

Policy improvement uses the values $V^{\pi}(x)$ to specify a policy π' that is guaranteed to be at least as good as π. The idea is to consider at state x the non-stationary policy which consists of taking action a (which may or may not be the action $\pi(x)$ that policy π recommends at x) and then following policy π thereafter. By the same reasoning as above, the expected value of this is

$$Q^{\pi}(x, a) = \hat{r}(x, a) + \gamma \sum_{y} P_{xy}(a) V^{\pi}(y) \quad (5)$$

The new policy is

$$\pi'(x) = \max_{a} \{Q^{\pi}(x, a)\} \quad (6)$$

from which it is obvious that $Q^{\pi}(x,\ \pi'(x)) \geq Q^{\pi}(x,\ \pi(x))$. Since the actions at all states are thus

only improving, the overall policy can also only improve, simultaneously at every state. If policy π' is the same as policy π, then it is an optimal policy (of which there may be more than one). The values $V^{\pi}(x)$ and $Q^{\pi}(x, a)$ associated with an optimal policy are called the optimal values or the optimal action values and are often written $V^{*}(x)$ and $Q^{*}(x, a)$. We will see later how to improve a policy without having explicitly to maximize $Q^{\pi}(x, a)$. Since the problem is finite, and the policy is improved at each iteration, policy iteration is bound to converge to the optimal policy. Although, in the worst case, policy iteration can take an exponential number of steps, it is generally very fast.

Value Iteration

Value iteration is the main alternative to policy iteration. For one version of this, a set of values $Q(x, a)$ (which are called Q-values) is updated simultaneously according to the formula

$$Q'(x,a) = \hat{r}(x,a) + \gamma \sum_{y} P_{xy}(a) \max_{b} Q(y,b) \qquad (7)$$

Although the Q-values are not necessarily the Q^{π}-values associated with any policy π, as in eqn 5, one can show that iterating eqn 7 infinitely often will lead to the optimal Q-values. Then, eqn 6 can be used to find the associated optimal policy. The proof that value iteration works depends on a contraction property. That is, a particular measure (called the L_{∞} norm) of the distance between $Q(x, a)$ and $Q^{*}(x, a)$ is greater than that between $Q'(x, a)$ and $Q^{*}(x, a)$ by a factor of at least $\gamma < 1$. Thus, the values converge exponentially fast to the optimal values. The optimal policy can actually be derived from them even before convergence.

TEMPORAL DIFFERENCES

Implementing either policy or value iteration requires the agent to know the expected rewards and the transition matrices. If the agent does not know these, but rather has to learn just by interacting with the environment (e.g. by pulling the levers), then what can it do? Methods of temporal differences were invented to perform prediction and optimization in these circumstances. There are two principal kinds of temporal difference method, an actor-critic scheme (Barto, Sutton and Anderson, 1983), which parallels policy iteration, and has been suggested as being implemented in biological RL, and a method called Q-learning (Watkins, 1989), which parallels value iteration.

Methods of temporal differences were invented (Sutton, 1988; Sutton and Barto, 1987) to account for the behaviour of animals in psychological experiments involving prediction; the links with dynamic programming were only made much later (Watkins, 1989; Barto, Sutton and Watkins, 1990). A number of related suggestions were made by Werbos (1990).

Actor–Critic Learning

In actor–critic learning, the *actor* specifies a policy, which the *critic* evaluates. Consider first a fixed policy or actor π. The critic has to find a set of values $V^{\pi}(x)$ that satisfy eqn 4 based on samples $\{x(t), r(t)\}$ from the transition and reward structure of the world (writing $r(t) = r(x(t), a(t))$ for convenience). It does this by maintaining and improving an approximation $V(x)$ to these quantities.

Two ideas underlie the temporal difference algorithm. One is to use averages from random samples to determine means; the second is to use a form of bootstrapping, substituting the incorrect estimates $V(y)$ as approximations for $V^{\pi}(y)$ in eqn 4. First, consider the quantity

$$r(0) + \gamma V^{\pi}(x(1)) \qquad (8)$$

which comes from the random reward and transition consequent on choosing action $\pi(x)$ at state $x(0) = x$. The expected value of this is

$$\langle r(0) + \gamma V^{\pi}(x(1)) \rangle_{r(0), x(1)} = \hat{r}(x, \pi(x)) + \gamma \sum_{y} P_{xy}(\pi(x)) V^{\pi}(y) \qquad (9)$$

which is just $Q^{\pi}(x, \pi(x))$, the quantity on the left-hand side of eqn 3. Therefore,

$$r(0) + \gamma V^{\pi}(x(1)) - V(x) \qquad (10)$$

could be used as a sampled error in the current estimate of the value of state x, namely $V(x)$, and an algorithm such as

$$V(x) \rightarrow V(x) + \varepsilon[r(0) + \gamma V^{\pi}(x(1)) - V(x)] \qquad (11)$$

when applied over the course of many trials might be expected to make $V(x) \approx V^{\pi}(x)$. Here, ε is a *learning rate*. In fact, this is a form of a stochastic, error-correcting, learning rule like the delta rule. The same is true for all subsequent timesteps.

Of course, expression 8 contains $V^{\pi}(x(1))$, which is exactly the quantity that we are trying to estimate, only at state $x(1)$ rather than $x(0)$. The second key idea underlying temporal difference methods is to substitute the current approximation $V(x(1))$

for this quantity. In this case, the sampled error parallel expression 10 is

$$\delta(0) = r(0) + \gamma V(x(1)) - V(x) \tag{12}$$

which is called the *temporal difference*. The temporal difference term for subsequent timesteps is defined similarly:

$$\delta(t) = r(t) + \gamma V(x(t+1)) - V(x(t)) \tag{13}$$

The learning rule

$$V(x(t)) \rightarrow V(x(t)) + \varepsilon\delta(t) \tag{14}$$

is called the temporal difference learning rule. It can be considered as a prediction error or a measure of the inconsistency between the estimates $V(x(t))$ and $V(x(t+1))$.

Despite the apparent approximations, circumstances are known under which $V(x) \rightarrow V^{\pi}(x)$ as the number of trials increases, so implementing policy evaluation correctly. In fact, this is also true in the case that the policy is stochastic, that is $\pi_a(x)$ is a probability distribution over actions a associated with states x.

The other half of policy iteration is policy improvement, patterned after eqn 6. Here, the idea is to use a stochastic policy (rather than the deterministic ones that we have hitherto considered), which is defined by a set of parameters, called the action matrix M_{ax}. Then, rather than implement the full maximization of eqn 6, we consider changing the parameters, using the result of policy evaluation in order to increase the probabilities of actions that are associated with higher values of $Q^{\pi}(x,\ a)$. A natural way to parametrize a stochastic policy is to use a softmax function:

$$\pi_a(x) = \frac{\exp(M_{ax})}{\sum_b \exp(M_{bx})} \tag{15}$$

The normalization forces $\sum_a \pi_a(x) = 1$ for all states x. Here, the larger M_{ax} is compared with M_{bx}, the more likely action a is at state x.

Consider the temporal difference δ after convergence, i.e. when $V(x) = V^{\pi}(x)$. In this case,

$$V(x) = V^{\pi}(x) = \sum_a \pi_a(x) Q^{\pi}(x,a) \tag{16}$$

is the average value of the actions specified by policy π. If action a is better than average, then $Q^{\pi}(x,\ a) > V(x)$; if action a is worse than average, then $Q^{\pi}(x,\ a) < V(x)$. Thus, π could be improved by changing the action matrix according to

$$M_{ax} \rightarrow M_{ax} + \varepsilon[Q^{\pi}(x,a) - V(x)] \tag{17}$$

However, if the agent takes action $a(0) = a$ at state $x(0) = x$, then

$$\delta(0) = Q^{\pi}(x,a) - V(x) \tag{18}$$

and so the actor learning rule is

$$M_{a(0)x(0)} \rightarrow M_{a(0)x(0)} + \varepsilon\delta(0) \tag{19}$$

using just the same temporal difference term as in eqn 14.

Normally, the action values are changed according to eqn 19 before the critic has converged, and there is no proof that the combined estimation and optimization procedure is guaranteed to find an optimal policy. Nevertheless, it is found to work well in practice. Note that learning is based on following, and simultaneously trying to improve, a policy.

Q-Learning

Q-learning is a temporal difference version of value iteration. It applies to eqn 7 the same two ideas underlying actor–critic learning. That is, it maintains and improves values $Q(x,\ a)$, employing

$$r(0) + \gamma \max_b Q(x(1), b) \tag{20}$$

as a sampled version of the right-hand side of eqn 7. The overall update at time t is

$$Q(x(t), a(t)) \rightarrow Q(x(t),\ a(t)) + \varepsilon\left[r(t) + \gamma \max_b Q(x(t+1), b) - Q(x(t), a(t))\right] \tag{21}$$

As in value iteration, a policy can be defined at any time according to

$$\pi(x) = \max_a Q(x,a) \tag{22}$$

Unlike the actor–critic scheme, circumstances are known under which Q-learning is guaranteed to converge to the optimal policy.

Exploration and Exploitation

One of the most important issues for the temporal difference learning algorithms is maintaining a balance between exploration and exploitation. That is, the agent must sometimes choose actions that it believes to be suboptimal in order to find out whether they might actually be good. This is particularly true in problems which change over time (which is true of most behavioural experiments), since actions that used to be good might become bad, and vice versa.

Temporal difference algorithms that can be guaranteed to work usually require much experimentation: all actions must be repeatedly tried in all states. In practice, it is common to choose policies that employ some exploration (such as choosing a random action some small fraction of the time, but otherwise the action currently believed to be best).

EXTENSIONS TO TEMPORAL DIFFERENCE LEARNING

We have presented a very simple form of temporal difference learning algorithm, but it has been extended in various ways. First, it is possible to learn the transition matrices and rewards and then use standard dynamic programming methods on the resulting estimated model of the environment (this is called a model-based method). It is also possible to integrate the learning of a model with the temporal difference learning methods, as a way of using the samples of transitions and rewards more effectively. There is substantial evidence that animals build models of their environments, although it is not clear how they use this information to learn appropriate behaviours.

Second, the idea behind eqn 8 is to use $V^{\pi}(x(1))$ as a stochastic sample of $\sum_y P_{xy}(\pi(x))V^{\pi}(y)$ to replace all but the first reward term in the infinite sum of eqn 3. One could equally well imagine collecting two actual steps of reward $r(0) + \gamma r(1)$, and using $V^{\pi}(x(2))$ as a sample of the all but the first two terms in eqn 3. Similarly, one could consider all but the first three terms, etc. It turns out to be simple to weigh these different contributions in an exponentially decreasing manner, and this leads to an algorithm called TD(λ), where the value of λ is the weighting, and the value $\lambda = 0$ corresponds to eqn 8. The remaining bootstrap approximation is the same as before. The significance of λ is that it controls a trade-off between bias and variance. If λ is near 1, then the estimate is highly variable (since it depends on long sample paths), but not strongly biased (since real rewards from the reward distribution of the environment are considered). If λ is near 0, then the estimate is not very variable (since only short sample paths are involved), but it can be biased, because of the bootstrapping.

Third, we described the temporal difference algorithm using a unary or *table-lookup* representation in which we can separately manipulate the value $V(x)$ associated with each state $x(t)$. One can also consider parametrized representations of the values (and also the policies), for instance the linear form

$$V(x) = \mathbf{w} \cdot \boldsymbol{\phi}(x) \tag{23}$$

where $\mathbf{w}$ is a set of parameters, and $\boldsymbol{\phi}(x)$ is a population representation for state x. In this case, the temporal difference update rule of eqn 14 is changed to one for the parameters:

$$\mathbf{w} \rightarrow \mathbf{w} + \varepsilon\delta(t)\boldsymbol{\phi}(x(t)) \tag{24}$$

Again, this is similar to the delta rule. Representational schemes more sophisticated than eqn 23 can also be used, including neural networks, in which case $\nabla_{\mathbf{w}} V(x(t))$ should be used in place of $\boldsymbol{\phi}(x(t))$.

FORMALIZING LEARNING IN THE BRAIN

We have stressed that animals and artificial systems face similar problems in learning how to optimize their behaviour in the light of rewards and punishments. Although we have described temporal difference methods from the perspective of the engineering methods of dynamic programming, they are also interesting as a way of formalizing behavioural and neurobiological experiments on animals. Briefly, it has been postulated that a prediction error signal for appetitive events with some properties like that of the temporal difference signal $\delta(t)$ (eqn 13) seems to be broadcast by dopaminergic cells in the ventral tegmental area, to control the learning of predictions, and by dopaminergic cells in the substantia nigra to control the learning of actions. The model is, as yet, incomplete, failing, for example, to specify the interaction between attention and learning, which is critical in accounting for the results of behavioural experiments. (*See* **Reinforcement Learning: A Biological Perspective**)

References

Barto AG, Sutton RS and Anderson CW (1983) Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics* **13**: 834–846.

Bellman RE (1957) *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Sutton RS (1988) Learning to predict by the methods of temporal difference. *Machine Learning* **3**: 9–44.

Sutton RS and Barto AG (1987) A temporal-difference model of classical conditioning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*. Seattle, WA.

Watkins CJCH (1989) *Learning from Delayed Rewards*. PhD Thesis, University of Cambridge.

Werbos PJ (1990) Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks* **3**: 179–189.

Further Reading

Barto AG, Sutton RS and Watkins CJCH (1990) Learning and sequential decision making. In: Gabriel M and Moore J (eds) *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge, MA: MIT Press/Bradford Books.

Bertsekas DP and Tsitsiklis JN (1996) *Neuro dynamic Programming*. Belmont, MA: Athena Scientific.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Representations Using Formal Logics

Intermediate article

Selmer Bringsjord, Rensselaer Polytechnic Institute, Troy, NY, USA
Yingrui Yang, Rensselaer Polytechnic Institute, Troy, NY, USA

CONTENTS

LOGIC AND THE LANGUAGE OF THOUGHT

Johnson-Laird and Savary (1995) present the following 'illusion' (illusion 1):

1. If there is a king in the hand then there is an ace, or if there isn't a king in the hand then there is an ace (but not both).
2. There is a king in the hand.

Given these premises, what can one infer?

Almost certainly your verdict is this: one can infer that there is an ace in the hand. And you reached this verdict despite the fact that we introduced the problem as an 'illusion', which no doubt, at least to some degree, warned you that something unusual was in the air. Why do we refer to it as an *illusion*? Because your verdict seems correct, even perhaps obviously correct, and yet a little logic suffices to show not only that are you wrong, but that in fact what you can infer is that there *isn't* an ace in the hand.

Of course, not everyone is tricked. How do we explain the fact that Jones is, while Smith is not? The explanation offered by traditional cognitive science is that both Jones' and Smith's cognition involves knowledge: knowledge that is represented in logic (or logic-like systems), and knowledge that is processed by reasoning. So, an explanation of why some subjects 'crack' the illusion and others do not must be couched in terms of such representation and reasoning. According to one view, thinking, at least of the 'high-level' sort, takes place in a logic-like language; in other words, logic is the 'language of thought' for *Homo sapiens* (Braine, 1998; Fodor, 1975). We are not concerned here with whether or not this view is true. (For an argument in its favor, see Bringsjord and Ferrucci (1998).) Our concern is rather with presenting the simplest of those formal languages that are popular candidates for the language of thought. We want to explain logic as a means for both representing knowledge and reasoning with that knowledge.

To begin, we note that modern symbolic logic has three main components: one is purely syntactic, one is semantic, and one is metatheoretical in nature. The syntactic component includes specification of the alphabet of a given logical system, the grammar for building well-formed formulae (WFFs) from this alphabet, and a proof theory that precisely describes how and when one formula can be proved from a set of formulae. The semantic component includes a precise account of the conditions under which a formula in a given system is true or false. The metatheoretical component includes theorems, conjectures, and hypotheses concerning the syntactic component, the semantic component, and connections between them. The two simplest and most used logics for representation in cognitive science are the propositional calculus (also known as 'propositional logic' or 'sentential logic') and the predicate calculus. The second of these subsumes the first, and is often called 'first-order logic' (FOL). We now proceed to characterize the three components for both the propositional calculus and FOL, starting with the former.

PROPOSITIONAL LOGIC

Grammar

The alphabet for propositional logic is simply an infinite list $(p_1, p_2, \ldots)$ of propositional variables (traditionally p_1 is p, p_2 is q, and p_3 is r), and the five familiar truth-functional connectives $\neg$, $\rightarrow$, $\leftrightarrow$, $\wedge$, and $\vee$. These connectives can, at least provisionally, be read, respectively, as 'not', 'implies' (or 'if… then…'), 'if and only if', 'and', and 'or'. In

cognitive science it is often convenient to use propositional variables as mnemonics that help one remember what they are intended to represent. For an example, recall illusion 1. Instead of representing 'there is an ace in the hand' as 'p_i', for some $i \in \{1, 2, \ldots\}$, it would be convenient to represent this proposition as 'A'. Now, the grammar for propositional logic is composed of the following three rules:

1. Every propositional variable p_i is a WFF.
2. If ϕ is a WFF, then so is $\neg\phi$.
3. If ϕ and ψ are WFFs, then so is $(\phi * \psi)$, where $*$ is one of $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$. (We allow outermost parentheses to be dropped.)

This implies, for example, that $p \rightarrow (q \wedge r)$ is a WFF, but $\rightarrow q$ isn't. To represent the declarative sentence 'if there is an ace in the hand, then there is a king in the hand' we could use $A \rightarrow K$.

Syntactic Proofs (Proof Theory)

A number of proof theories are possible. One such system is an elegant Fitch-style system of natural deduction, $\mathcal{F}$ (Barwise and Etchemendy, 1999). (Such systems are commonly referred to as 'natural' systems.) In $\mathcal{F}$, each of the truth-functional connectives has a pair of corresponding inference rules, one for introducing the connective, and one for eliminating the connective. Proofs in $\mathcal{F}$ proceed in sequence line by line, with successive line numbers incremented by 1. Each line includes a line number, a formula (the one deduced at this line), and, in the rightmost column, a rule cited in justification for the deduction. We use a vertical ellipsis ⋮ to indicate the presence of 0 or more lines in the proof not explicitly shown.

Here is the rule for eliminating a conjunction:

⋮	⋮	⋮
k	$\phi \wedge \psi$	
⋮	⋮	⋮
m	ϕ	k $\wedge$-Elim
⋮	⋮	⋮

Intuitively, this rule says that if at line k in some derivation you have somehow obtained a conjunction $\phi \wedge \psi$, then at a subsequent line m, one can infer to either of the conjuncts alone.

Now here is the rule that allows a conjunction to be introduced; intuitively, it formalizes the fact that if two propositions are true then the conjunction of these two propositions is also true:

⋮	⋮	⋮
k	ϕ	
⋮	⋮	⋮
l	ψ	
⋮	⋮	⋮
m	$\phi \wedge \psi$	k,l $\wedge$-Intro
⋮	⋮	⋮

An important rule in $\mathcal{F}$ is 'supposition', according to which you are allowed to assume any WFF at any point in a derivation. The catch is that you must signal your use of supposition by setting it off typographically, as follows:

⋮	⋮	
k	ϕ	supposition
	⋮	
⋮	⋮	

Often a derivation will be used to establish that from some set Φ of propositional formulae a particular formula ϕ can be derived. In such a case, Φ will be given as suppositions (or, as we sometimes say, 'givens'). To say that ϕ can be derived in $\mathcal{F}$ from a set of formulae Φ we write

$$\Phi \vdash_{\mathcal{F}} \phi$$

When it is clear from context which system the deduction is to take place in, the subscript on $\vdash$ can be omitted. Here is a proof that puts to use the rules presented above and establishes that $((p \wedge q) \wedge r) \vdash_{\mathcal{F}} q$:

1	$(p \wedge q) \wedge r$	given
2	$(p \wedge q)$	1 $\wedge$-Elim
3	q	2 $\wedge$-Elim

Now here is a slightly more complicated rule, one for introducing a conditional. It basically says that if you can carry out a subderivation in which you suppose ϕ and derive ψ, you are entitled to close this subderivation and infer to the conditional $\phi \rightarrow \psi$:

⋮		⋮
k		ϕ supposition
⋮		⋮
m		ψ
⋮		⋮
n	$\phi \rightarrow \psi$	k–m $\rightarrow$-Intro

Figure 1. A proof of modus tollens in $\mathcal{F}$, constructed in Hyperproof.

As we said, in a Fitch-style system of natural deduction, the rules come in pairs. Here is the rule in $\mathcal{F}$ for eliminating conditionals:

k	$\phi \rightarrow \psi$	
⋮	⋮	⋮
l	ϕ	
⋮	⋮	⋮
m	ψ	k, l $\rightarrow$-Elim

Here is the rule for introducing $\vee$:

⋮	⋮	⋮
k	ϕ	
⋮	⋮	⋮
m	$\phi \vee \psi$	k $\wedge$-Intro
⋮	⋮	⋮

And here is the rather more elaborate rule for eliminating a disjunction:

⋮	⋮	
k	$\phi \vee \psi$	
⋮	⋮	
l	ϕ supposition	
⋮	⋮	
m	χ	
⋮	⋮	
n	ψ supposition	
⋮	⋮	
o	χ	
⋮	⋮	
p	χ	k, l–m, n–o $\vee$-Elim

The rule $\vee$-Elim is also known as 'constructive dilemma'. The intuition behind this rule is that if one knows that either ϕ or ψ is true, and if one can show that χ can be proved from ϕ alone, and from ψ alone, then χ must be true.

Next, here is a very powerful rule corresponding to proof by contradiction (sometimes called 'indirect proof' or 'reductio ad absurdum'). Notice that in $\mathcal{F}$ this rule is ¬-Intro:

⋮	⋮	
k	ϕ supposition	
⋮	⋮	
m	$\psi \wedge \neg \psi$	
⋮	⋮	
n	$\neg \phi$	k–m ¬-Intro

Sometimes a natural deduction system can be a little awkward, because by insisting that inference rules come exclusively in the form of pairs for each truth-functional connective, it leaves out certain rules that are exceedingly useful. Two examples are *modus tollens* and DeMorgan's laws. The former rule allows one to infer $\neg\phi$ from $\phi \rightarrow \psi$ and $\neg\psi$. This rule can be established through a proof in $\mathcal{F}$, as shown in Figure 1. This proof was constructed in the Hyperproof construction environment (Barwise and Etchemendy, 1994). The core of this proof is *reductio ad absurdum*, or ¬-Intro.

DeMorgan's laws for propositional logic sanction moving from a formula of the form $\neg(\phi \wedge \psi)$ to one of the form $\neg\phi \vee \neg\psi$, and vice versa. The laws also allow an inference from a formula of the form $\neg(\phi \vee \psi)$ to one of the form $\neg\phi \wedge \neg\psi$, and vice versa. When, in constructing a proof in $\mathcal{F}$, we want to use modus tollens or DeMorgan's laws, or some other time-saving rule, we can make the inference, using the rule of 'tautological consequence' as a justification. This rule (abbreviated as 'Taut Con' in Hyperproof) is designed to allow the human proof constructor to declare that a given inference is obvious, and could with more work be fully specified using only the rules of $\mathcal{F}$. Hyperproof responds with a check to indicate that an attempted inference is in fact correct. In Figure 2, Hyperproof approves of our use of 'Taut Con', which corresponds in this case not just to DeMorgan's law, but also to the useful inference of $\phi \wedge \neg\psi$ from $\neg(\phi \rightarrow \psi)$.

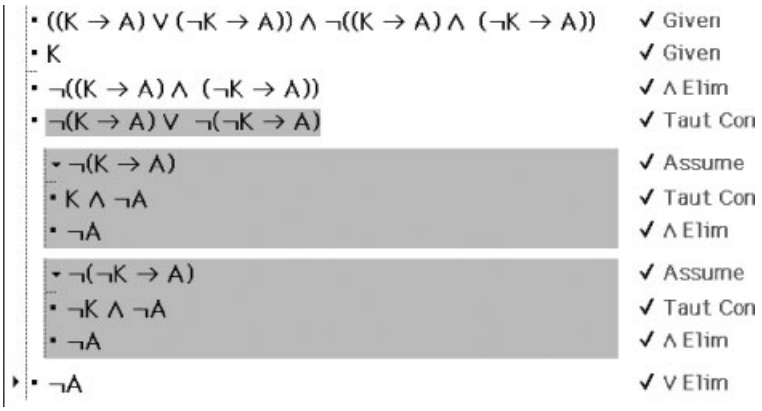


Figure 2. A proof in $\mathcal{F}$ that there is no ace in the hand. The premises are shown in the first two lines.

A formula provable from the null set is called a 'theorem', and where ϕ is such a formula we write $\vdash \phi$ to express this fact. Here are two examples: $\vdash (p \wedge q) \rightarrow q$; $\vdash (p \wedge \neg p) \rightarrow r$. We say that a set Φ of formulae is 'syntactically consistent' if and only if no contradiction can be derived from Φ.

Semantics (Truth Tables)

The precise meaning of the five truth-functional connectives of the propositional calculus is given via truth tables, which tell us what the truth-value of a statement is, given the truth-values of its components. The simplest truth table is that for negation, which informs us, unsurprisingly, that if ϕ is T then $\neg\phi$ is F and if ϕ is F then $\neg\phi$ is T:

ϕ	$\neg\phi$
T	F
F	T

Here are the remaining truth tables:

ϕ	ψ	$\phi \wedge \psi$
T	T	T
T	F	F
F	T	F
F	F	F

ϕ	ψ	$\phi \vee \psi$
T	T	T
T	F	T
F	T	T
F	F	F

ϕ	ψ	$\phi \rightarrow \psi$
T	T	T
T	F	F
F	T	T
F	F	T

ϕ	ψ	$\phi \leftrightarrow \psi$
T	T	T
T	F	F
F	T	F
F	F	T

Notice that the truth table for disjunction says that when both disjuncts are true, the disjunction is true. This is called 'inclusive' disjunction. In 'exclusive' disjunction, it's one disjunct or another, but not both. This distinction becomes particularly important if one is attempting to symbolize parts of English (or any other natural language). It would not be natural to represent the sentence 'George will either win or lose' as '$W \vee L$', because under the English meaning there is no way both possibilities can be true, whereas by the meaning of $\vee$ it would be possible that W and L are both true. We can use $\vee_x$ to denote exclusive disjunction, which we define through the following truth table:

ϕ	ψ	$\phi \vee_x \psi$
T	T	F
T	F	T
F	T	T
F	F	F

It is worth mentioning another issue involving the meaning of English sentences and their corresponding symbolizations in propositional logic: the issue of the oddity of 'material conditionals' (formulae of the form $\phi \rightarrow \psi$). Consider the following English sentence:

> If the Moon is made of green cheese, then Dan Quayle will be the next President of the United States.

Is this sentence true? If we were to ask 'the man on the street', the answer might well be: 'Of course not!' Or perhaps we would hear: 'This isn't even a meaningful sentence; you're speaking nonsense.' However, when represented in the propositional

calculus, the sentence turns out true. Why? The sentence is naturally represented as $G \rightarrow Q$. Since G is false, the truth table for $\rightarrow$ classifies the conditional as true. Results such as these have encouraged some to devise better (but much more complicated) accounts of the conditionals seen in natural languages (e.g. Goble, 2001). These accounts are beyond the scope of this article; we will be content with the conditional as defined by the truth table for $\rightarrow$ presented above.

Given a truth-value assignment υ (i.e., an assignment of T or F to each propositional variable p_i), we can say that υ 'makes true' or 'models' or 'satisfies' a given formula ϕ; this is written $\upsilon \models \phi$.

Some formulae are true on all models. For example, the formula $((p \vee q) \wedge \neg q) \rightarrow p$ is in this category. Such formulae are said to be 'valid' and are sometimes referred to as 'validities'. To indicate that a formula ϕ is valid we write $\models \phi$.

Another important semantic notion is 'consequence'. An individual formula ϕ is said to be a consequence of a set Φ of formulae provided that every truth-value assignment on which all of Φ are true is also one on which ϕ is true; this is written $\Phi \models \phi$.

The final concept in the semantic component of the propositional calculus is the concept of consistency: we say that a set Φ of formulae is 'semantically consistent' if and only if there is a truth-value assignment on which all of Φ are true.

Cracking Two Illusions

We now have at our disposal enough logic to 'crack' illusion 1. In this illusion, 'or' is to be understood as exclusive disjunction, so (using obvious symbolization) the two premises become $((K \rightarrow A) \vee (\neg K \rightarrow A)) \wedge \neg((K \rightarrow A) \wedge (\neg K \rightarrow A))$ and K.

Figure 2 shows a proof in $\mathcal{F}$, constructed in Hyperproof, that demonstrates that from these two givens one can conclude $\neg A$.

Now, consider another illusion (illusion 2):

1. The following three assertions are either all true or all false:
   - If Billy is happy, Doreen is happy.
   - If Doreen is happy, Frank is as well.
   - If Frank is happy, so is Emma.
2. Billy is happy.

Can it be inferred that Emma is happy?

Most people answer 'yes', but for the wrong reasons. They notice that since Billy is happy, if the three conditionals are true, one can 'chain' through them to arrive at the conclusion that Emma is happy. But this is only part of the story, and the other part has been ignored: it could be that all three conditionals are false. Other people realize that there are two cases to consider (conditionals all being true, and conditionals all being false), and because they believe that when the conditionals are all false one cannot prove that Emma is happy, they respond with 'No'. But this response is also wrong. The correct response is 'yes', because in both cases it can be proved that Emma is happy. This can be shown using propositional logic; the proof, again constructed in Hyperproof, is shown in Figure 3. This proof establishes

$$\{\neg(B \rightarrow D), \neg(D \rightarrow F)\} \vdash E$$

Note that the trick is exploiting the inconsistency of the set $\{\neg(B \rightarrow D), \neg(D \rightarrow F)\}$ in order to get a contradiction. Since everything follows from a contradiction, E can then be derived.

Metatheoretical Results

At this point we can give some metatheory for the propositional calculus. In general, metatheory would deploy logical and mathematical techniques in order to answer such questions as whether or not provability implies consequence, and whether or not the converse holds. When provability implies consequence, a logical system is said to be 'sound'. This fact can be expressed as: 'if $\Phi \vdash \phi$ then $\Phi \models \phi$'. Roughly, a logical system is sound if true formulae can only yield (through proofs) true formulae; one cannot pass from the true to the false.

When consequence implies provability, a system is said to be 'complete'. This is expressed by: 'if $\Phi \models \phi$ then $\Phi \vdash \phi$'.

The propositional calculus is both sound and complete. It follows that all theorems in the propositional calculus are valid, and all validities are theorems. This last fact is expressed more formally as: '$\models \phi$ if and only if $\vdash \phi$'.

FIRST-ORDER (PREDICATE) LOGIC

A Logical Illusion in Quantified Reasoning

Consider another illusion (illusion 3), a more complicated one that cannot be adequately represented in propositional logic. Here it is, adapted slightly from Yang and Johnson-Laird (2000):

Only one of the following statements is true:
- At least one of the beads is red.
- None of the beads are red.

Is it possible that none of the red things are beads?

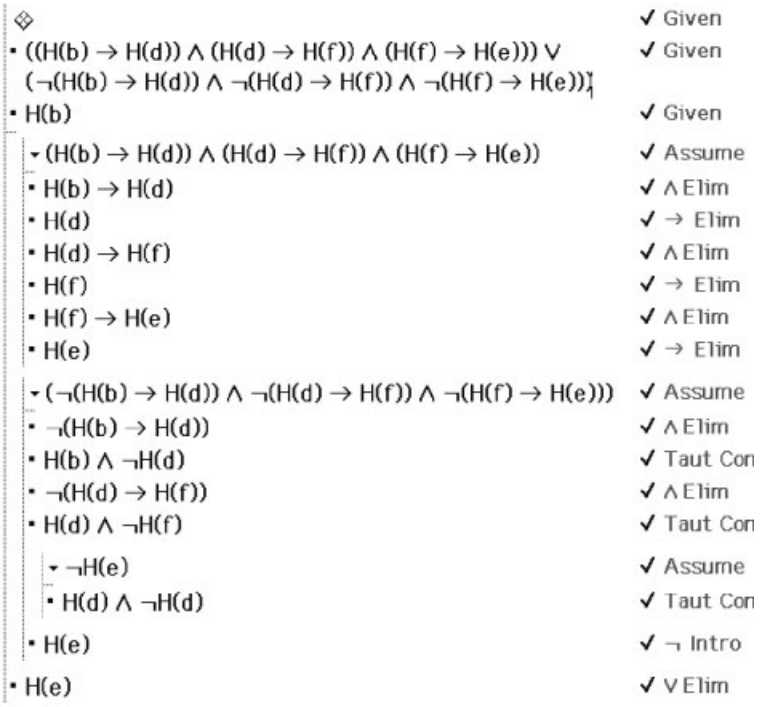


Figure 3. A proof in $\mathcal{F}$ that 'Emma is happy'.

We can remove all the mystery by turning to FOL, which we will introduce after explaining why the propositional calculus isn't expressive enough to represent this puzzle. The propositional calculus can represent propositions, but it cannot represent the internal structure of propositions: for example, propositions to the effect that certain objects have certain properties. In the propositional calculus, 'at least one of the beads is red' would be represented by some propositional variable, say N. We know that from 'at least one of the beads is red' it follows that 'at least one of the beads is red or blue'. The second statement here would be represented by some other propositional variable, say B. Clearly, $\{N\} \nvdash B$ in the propositional calculus. What is needed is some way to represent the fact that N says that at least one of a particular kind of object has a specific property, viz., being red. Only with such machinery can we get to the bottom of illusion 3.

The Syntactic Machinery of FOL

Our alphabet will now be augmented to include the following: the identity or equality symbol $=$; variables x, y,…; constants ('proper names' for objects) c_1, c_2,…; relational symbols R, G,… (e.g., R for 'being red'); functors (functions) f_1, f_2,…; the existential quantifier $\exists$ ('there exists at least one …'); the universal quantifier $\forall$ ('for all …'); and the familiar truth-functional connectives $\neg$, $\vee$, $\wedge$, $\rightarrow$, and $\leftrightarrow$.

Predictable 'formation rules' are introduced to allow us to represent propositions like those in illusion 3. With these rules, we can now write such things as $\exists x(Bx \wedge Rx)$, which says that there exists at least one thing x that has property B and property R. As in propositional logic, sets of formulae (say Φ), given certain 'rules of inference', can be used to prove individual formulae (say ϕ); such a situation is expressed by expressions having exactly the same form as those introduced above (e.g., $\Phi \vdash \phi$). The rules of inference for FOL in such systems as $\mathcal{F}$ include those we saw for the propositional calculus, and also new ones: two corresponding to the existential quantifier $\exists$, and two corresponding to the universal quantifier $\forall$. For example, one of the rules associated with $\forall$ says, intuitively, that if you know that everything has a certain property, then any particular thing a has that property. This rule, known as 'universal elimination' (or 'universal introduction') allows us to move from some formula $\forall x\phi$ to a formula with $\forall x$ dropped, and the variable x in ϕ replaced with the constant of choice. For example, from 'all beads are red', that is, $\forall x(Bx \rightarrow Rx)$, we can infer by $\forall$-Elim that $Ba \rightarrow Ra$, and if we happen to know that in fact Ba we can now infer by familiar propositional reasoning that Ra. The rule $\forall$-Elim in $\mathcal{F}$ is

$$\begin{array}{l|ll} k & \forall x \phi & \\ \vdots & \vdots & \vdots \\ l & \phi(\frac{a}{x}) & k\,\forall\text{-Elim} \end{array}$$

where $\phi(\frac{a}{x})$ denotes the result of replacing occurrences of x in ϕ with a.

Semantics (Interpretations)

FOL includes a semantic side, which systematically provides meaning (i.e., truth or falsity) for formulae. Unfortunately, the formal semantics of FOL are more tricky than the truth tables that are sufficient for the propositional level. In FOL, formulae are said to be true (or false) on 'interpretations' or 'models'; that some formula ϕ is true on an interpretation $\mathcal{I}$ is often written as $\mathcal{I} \models \phi$. (We say that $\mathcal{I}$ satisfies, or models, ϕ.) For example, the formula $\forall x \exists y Gyx$ might mean, on the standard interpretation for arithmetic, that for every natural number n, there is a natural number m such that $m > n$. In this case, the 'domain' is the set $\mathbb{N}$ of natural numbers; and G symbolizes 'is greater than'. Much more could of course be said about the formal semantics (or 'model theory') for FOL; but this is beyond the scope of the present article. For a full discussion using the traditional notation of model theory, see Ebbinghaus *et al*. (1984). There it is shown that FOL, like the propositional calculus, is both sound and complete.

Cracking the Illusion

We now have the tools to crack illusion 3. Yang and Johnson-Laird (2000) found that very few untrained reasoners answered the question in illusion 3 correctly – but these authors tacitly assumed that 'none of the beads are red' should be represented in a manner that entails that there do exist some beads. If we side with this interpretation, then the question becomes whether in both cases it can be proved that it's not possible that none of the red things are beads. That is, can it be proved that in both cases a contradiction arises if one assumes that none of the red things are beads? The first case is shown as an explicit Hyperproof-constructed proof in $\mathcal{F}$ in Figure 4.

This proof can be expressed informally as follows. We begin by assuming in the first line that the first statement of illusion 3 is true and the second statement is false. In the next line we assume that there are some red things, and that none of them are beads. Next, we isolate the proposition that there are some red beads, by deriving it from the first line. In the fourth line, we assume, in keeping with the third line, that some arbitrary object a is a red bead. Now we derive that all red things are not beads directly from line 2, by the $\wedge$-Elim rule. From this it follows by universal elimination that if the particular object a is red, it can't be a bead. Since we know that (under our assumptions) a is red, we can infer by modus ponens ($\rightarrow$-Elim, here) that a isn't a bead. But we are operating under the assumption that a is a bead (see line 4), so we have a contradiction. We use the propositional variable Z to represent an explicit contradiction. (Often the symbol $\perp$ is used for this purpose.) Since everything follows from a contradiction, we simply deduce Z from the contradiction of $B(a)$ and $\neg B(a)$. The reason is that we need to obey the rule of existential elimination, which insists that if something ϕ follows from assuming that some arbitrary thing (a in the proof) has some property, then we can infer that ϕ follows from the general existential claim that something x has that property – provided a doesn't occur in ϕ. At this point we have shown that Z, that is, a contradiction, arises if

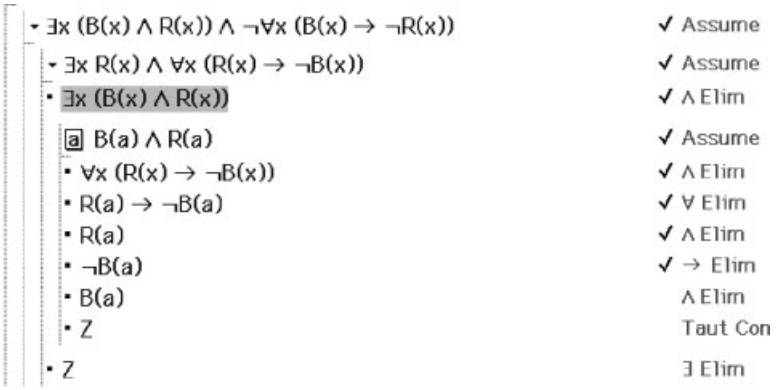


Figure 4. A derivation of a contradiction in $\mathcal{F}$.

we assume that none of the red things are beads. We leave it to our readers to ascertain whether the second case of illusion 3 can be dealt with in a similar way.

Representations in Logics Beyond FOL

Many declarative sentences cannot be represented in FOL. Consider the sentence: 'If two things x and y are identical, then for every property F that x has, y has it as well, and vice versa.' This is known as Leibniz' law (LL), and it seems self-evident. But LL cannot be represented in FOL. However, LL can be represented in second-order logic (SOL), in which one can quantify not only over individual objects, but over properties as well. In SOL, LL becomes:

$$\forall x \forall y (x = y \leftrightarrow \forall X (Xx \leftrightarrow Xy))$$

Note that this formula contains a part meaning 'for every property X', which cannot be expressed in FOL. FOL permits quantification only over objects, not over the properties they can have. For an introduction to SOL, see Ebbinghaus *et al*. (1984); for a discussion of logics that can represent even more difficult constructions, see Goble (2001).

THE SITUATION CALCULUS

FOL is not only suitable for representing static information. It has long been used with great success to represent dynamic environments. One of the schemes for doing this is the 'situation calculus', which we will briefly describe, in connection with the 'wumpus world' test-bed. An example of a situation in this world is shown in Figure 5. The objective of the agent in this world is to find the gold and bring it back without getting killed. Pits are always surrounded by breezes or by other pits, the wumpus is always surrounded on at least three sides by a stench, and the gold glitters in the square in which it's positioned. The agent dies if it enters a square with a pit or a wumpus in it. (In the figure, the agent has managed to reach the gold.)

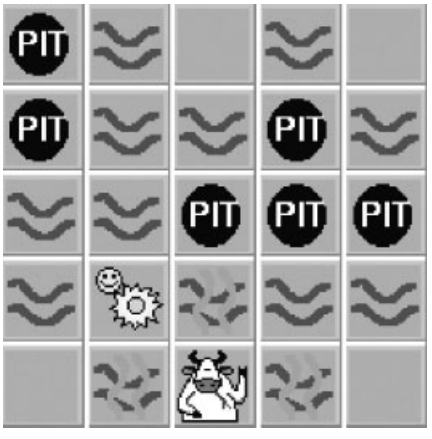


Figure 5. A typical wumpus world.

We can use the situation calculus to represent change in this world. First, we conceive of the world as consisting of a sequence of 'situations', essentially 'snapshots' of the world. Situations are generated from previous situations by actions. Properties in the world that can change over time are represented by inserting an extra constant to denote situations into the relevant formulae. For example, to describe the location of an agent a in the wumpus world at two situations s_0 and s_1, we could say that a is 'At' the square at row 1 and column 1, i.e., (1, 1), in situation s_0, and 'At' location (1, 2) in s_1. This would be written by:

$$\text{At}(a, (1,1), s_0) \wedge \text{At}(a, (1,2), s_1)$$

In order to use the situation calculus it is necessary to represent how the world changes from one situation to the next, by using the function

$$\text{Result}(\text{action}, \text{situation})$$

to refer to the situation that results from performing an action in some initial situation. With this function we can have sequences like that shown in Figure 6, in which, as a start, we have:

$$\text{Result}(\text{Forward}, s_0) = s_1$$
$$\text{Result}(\text{Backward}, s_1) = s_2$$
$$\text{Result}(\text{Turn}(\text{right}), s_2) = s_3$$
$$\text{Result}(\text{Forward}, s_3) = s_4$$

The 'Result' function also allows us to use FOL to represent such general rules about the wumpus world as that the agent isn't holding anything after a 'Release' action:

$$\forall x \forall s \neg \text{Holding}(x, \text{Result}(\text{Release}, s))$$

For more on the situation calculus for the wumpus world and beyond, see Russell and Norvig (1994), which includes discussion of the infamous 'frame problem', which plagues such general rules.

THE CHALLENGE TO LOGIC-BASED REPRESENTATION IN COGNITIVE SCIENCE

Illusion 2 shows that cognizers sometimes conceive of 'disproofs'. Specifically, a cognizer fooled by illusion 2 imagines an argument for the view that

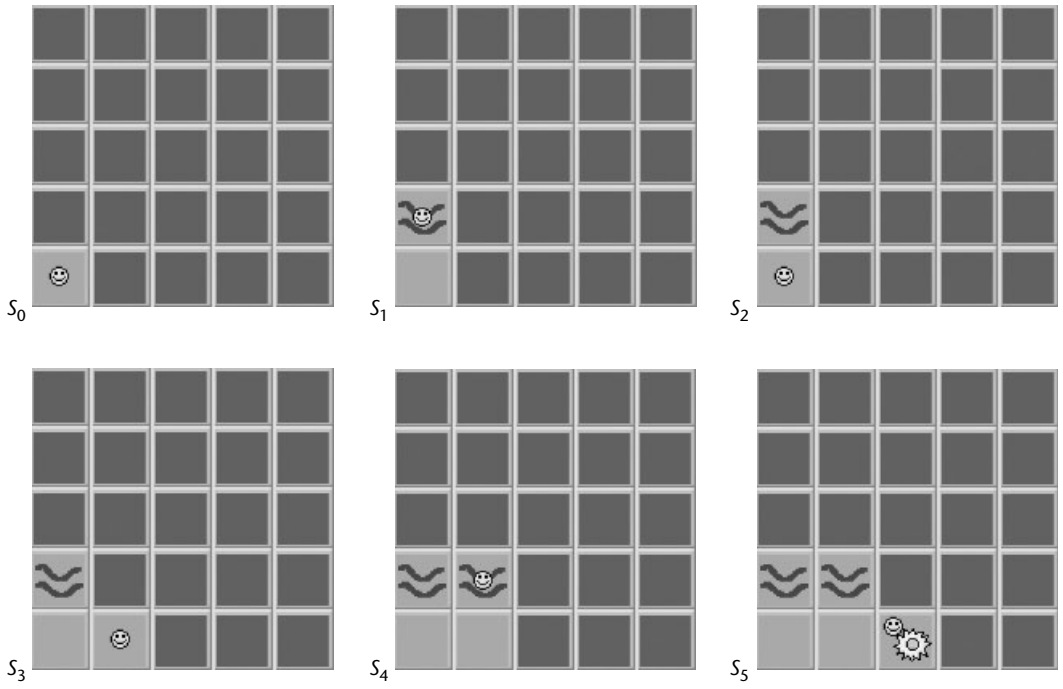


Figure 6. Situations s_0 to s_5 of a sequence in the wumpus world, linked by actions. (Program available at Escobar, 2002.)

there is no way to derive 'Emma is happy' from the negated trio of conditionals and 'Bill is happy'. Such arguments cannot be expressed in $\mathcal{F}$. However, a new theory, 'mental metalogic' (Yang and Bringsjord, 2001a,b; Rinella *et al*., 2001), provides a mechanism in which step-by-step proofs (including disproofs) can at once be syntactic and semantic, because situations can enter directly into line-by-line proofs (see Figure 7). Hyperproof can be viewed as a simple instantiation of part of this theory. In Hyperproof, one can prove such things as that $\Phi \nvdash \psi$, not only such things as $\Phi \vdash \psi$.

Suppose that in illusion 2, a 'tricked' cognizer moves from a correct representation of the premises when the conditionals are all true to an incorrect representation when the conditionals are false. Suppose, specifically, that the negated conditionals give rise to a situation, envisaged by the cognizer, in which four people *b*, *d*, *f*, and *e* are present, the sentence 'Billy is happy' is explicitly represented by a corresponding formula, and the question is whether it follows from this given information that Emma is happy. This situation is shown in Figure 8. Notice that the full logical import of the negated conditionals is nowhere to be found in this figure.

Next, given this starting situation, a disproof in Hyperproof is shown in Figure 9. Notice that a new, more detailed situation has been constructed, one which is consistent with the information given

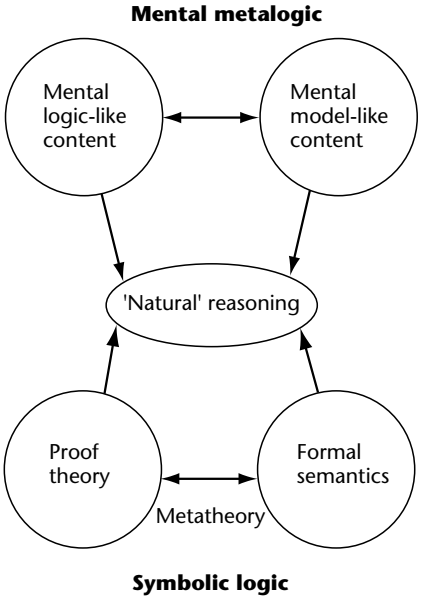


Figure 7. The symmetry of mental metalogic with symbolic logic.

(hence the 'CTA' rule which, if correctly used, **c**hecks the **t**ruth of the **a**ssumptions) and in which it is directly observed that Emma isn't happy. This demonstrates that Emma's being happy can't be deduced from the given information. So, though untrained human reasoning may not conform to normative patterns, erroneous thinking can

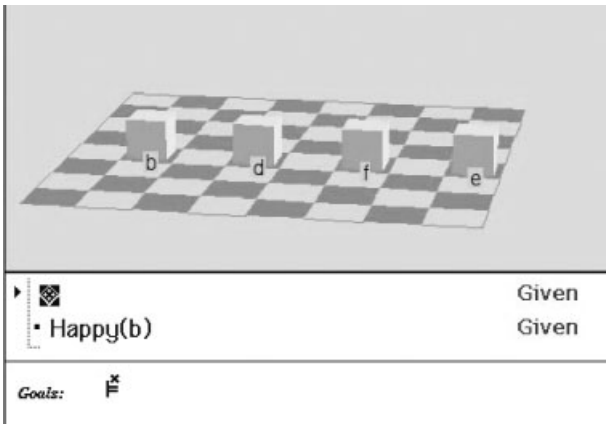


Figure 8. The start of a disproof that may be in the mind of cognizers.

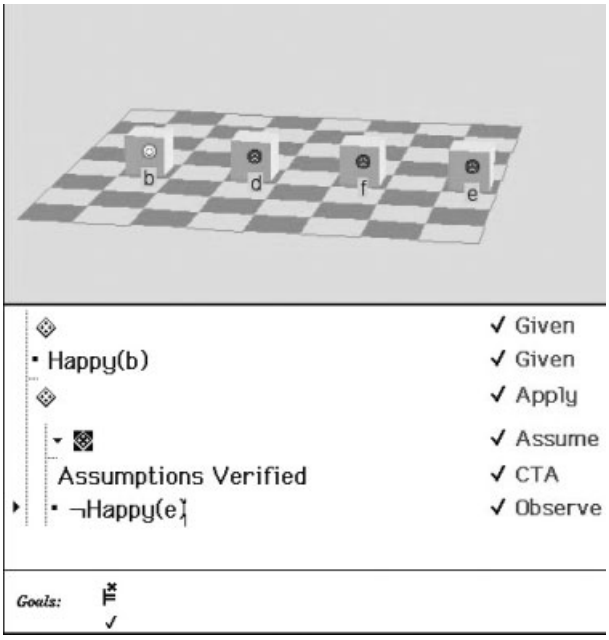


Figure 9. The completed disproof that may be in the mind of cognizers.

nonetheless be represented by logic, even when it involves pictorial reasoning. (For the reasons why Hyperproof derivations like that in Figure 9 cannot be reduced to purely syntactic reasoning, see Barwise and Etchemendy (1995).)

References

Barwise J and Etchemendy J (1994) *Hyperproof*. Stanford, CA: CSLI.

Barwise J and Etchemendy J (1995) Heterogeneous logic. In: Glasgow J, Narayanan N and Chandrasekaran B (eds) *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, pp. 211–234. Cambridge, MA: MIT Press.

Barwise J and Etchemendy J (1999) *Language, Proof, and Logic*. New York, NY: Seven Bridges.

Braine M (1998) How to investigate mental logic and the syntax of thought. In: Braine M and O'Brien P (eds) *Mental Logic*, pp. 45–61. Mahwah, NJ: Lawrence Erlbaum.

Bringsjord S and Ferrucci D (1998) Logic and artificial intelligence: divorced, still married, separated…? *Minds and Machines* **8**: 273–308.

Ebbinghaus HD, Flum J and Thomas W (1984) *Mathematical Logic*. New York, NY: Springer.

Escobar J and Bringsjord S (2002) *Wumpus World*. http://kryten.mm.rpi.edu/otter/wumpus/Wumpus.html.

Fodor J (1975) *The Language of Thought*. Cambridge, MA: Harvard University Press.

Goble L (ed.) (2001) *The Blackwell Guide to Philosophical Logic*. Oxford, UK: Blackwell.

Johnson-Laird P and Savary F (1995) How to make the impossible seem probable. In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 381–384. Hillsdale, NJ: Lawrence Erlbaum.

Rinella K, Bringsjord S and Yang Y (2001) Efficacious logic instruction: people are not irremediably poor deductive reasoners. In: Moore JD and Stenning K (eds) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp. 851–856. Mahwah, NJ: Lawrence Erlbaum.

Russell S and Norvig P (1994) *Artificial Intelligence: A Modern Approach*. Saddle River, NJ: Prentice Hall.

Yang Y and Bringsjord S (2001a) Mental metalogic: a new paradigm for psychology of reasoning. In: *Proceedings of the Third International Conference on Cognitive Science (ICCS 2001)*, pp. 199–204. Hefei, China: Press of the University of Science and Technology of China.

Yang Y and Bringsjord S (2001b) The mental possible worlds mechanism: a new method for analyzing logical reasoning problems on the GRE. In: *Proceedings of the Third International Conference on Cognitive Science (ICCS 2001)*, pp. 205–210. Hefei, China: Press of the University of Science and Technology of China.

Further Reading

Bringsjord S, Bringsjord E and Noel R (1998) In defense of logical minds. In: *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp. 173–178. Mahwah, NJ: Lawrence Erlbaum.

Johnson-Laird P (1997a) Rules and illusions: a critical study of Rips's *The Psychology of Proof*. *Minds and Machines* **7**: 387–407.

Johnson-Laird P (1997b) And end to the controversy? A reply to Rips. *Minds and Machines* **7**: 425–432.

Yang Y and Johnson-Laird P (2000) How to eliminate illusions in quantified reasoning. *Memory and Cognition* **28**: 1050–1059.

Resolution Theorem Proving

Intermediate article

Larry Wos, Argonne National Laboratory, Argonne, Illinois, USA

CONTENTS

Resolution theorem proving is the finding of proof of theorems, usually by contradiction, by applying inference rules each of which relies on unification.

INTRODUCTION

Puzzle solving, chess playing, computer program writing, and theorem proving in mathematics are satisfying endeavors. What is required to succeed in such activities? The essential factor is careful and deep reasoning: the ability to make useful, and often difficult, inferences.

Such fine reasoning must avoid various pitfalls, such as reliance on a hidden and erroneous assumption. For example, what is the problem with the following two 'facts'?

Plato likes absolutely everybody.
Absolutely nobody likes Plato.

Imagine what would be needed for a computer program to consider this question, and deeper questions. To submit questions and problems of various types to a computer program, a general language is needed. To draw conclusions in a rigorous manner, sound inference rules are needed. To 'know' when the question has been answered or the problem solved, a test for assignment completion is needed. And, at least for deep questions and hard problems, much more is needed for the program to be successful. (*See* **Problem Solving**)

This article describes such a language, diverse inference rules, a test for success, and other elements including the use of strategy. For modern computers are startlingly fast, a strategy to control the program's reasoning is necessary. Without some strategy, the program will almost certainly fail, especially if the task is a deep question or hard problem.

The objective of the field called 'automated reasoning' is to design and implement a program that applies flawless logical reasoning. At the heart of automated reasoning is 'resolution theorem proving': finding proofs of theorems, usually by contradiction, by applying inference rules each of which relies on 'unification' (see below). A number of automated reasoning programs have been produced: some of them are very powerful; some are freely available; and some have been used to make significant contributions to mathematics, logic, microchip design, and program verification.

Typically, an automated reasoning program given our two 'facts' about Plato would find a flaw (in the form of a contradiction) in a few milliseconds. In the program's 'clause language' (in which logical 'not' is denoted by '−'), the program would consider the two statements `LIKES(Plato,x)` (for all `x`) and `-LIKES(y,Plato)` (for all `y`) and substitute `Plato` for `x` in the first and `Plato` for `y` in the second and recognize that a contradiction had been found. The typical test for successful assignment completion is the discovery of a contradiction.

PROOF AND REFUTATION

In trying to prove a theorem of the form '*P* implies *Q*', there are two obvious approaches. In the first approach (often taken by a person, and almost never by a reasoning program), one begins with the elements of *P* and makes deductions until one recognizes that *Q* has been deduced. In the second approach (often taken by a mathematician, and almost always by a reasoning program), one assumes *Q* to be false and begins reasoning from that assumption together with *P* until a contradiction is found.

Thus, the reasoning program, when it seeks proofs by contradiction, seeks refutations of

inconsistent sets. Therefore, whether the context is puzzle solving, circuit design, program verification, or theorem proving in mathematics, the user is asked to phrase the question or problem in a manner that permits the program to seek a refutation in order to complete the assignment.

THE CLAUSE LANGUAGE

Except for the concept of equality (which is a built-in concept in some of the more powerful automated reasoning programs), the program knows nothing. Whatever information is thought to be needed must be supplied. It can be supplied in first-order predicate calculus or in the clause language. (Some programs instead rely on higher-order logic, on LISP, or on some other language.) If the information is supplied in first-order predicate calculus, many programs then convert the supplied information into the clause language.

The clause language relies on just two explicit logical connectives, 'not' and 'or', here denoted respectively by '−' and '|', and one implicit connective, 'and'. (There are various different notational conventions.) Other connectives are expressed in terms of 'not' and 'or'. For example, 'if *P* then *Q*' is replaced by (the logically equivalent) 'not *P* or *Q*'.

The language also relies on constants (such as `a`, `b`, `c`, `0`, and `Plato`), functions (such as `f`, `g`, `h`, `mother`, `youngest_daughter`, and `sum`), and relations called 'predicates' (such as `female`, `subset`, and `equals`). Variables (conventionally denoted by expressions beginning with lower-case `u` through `z`) are implicitly universally quantified, each meaning 'for all', and their scope is just the clause in which they occur. For example, the fact that everybody is female or male, and the fact that nobody is both female and male are, respectively, conveyed to the program with the following two clauses:

```
FEMALE(x)|MALE(x).
-FEMALE(x)|-MALE(x).
```

In these clauses, each of the items separated by 'or' is called a 'literal', and a clause is the 'or' of its literals with the requirement that no literal may appear more than once in a clause. The 'empty clause' contains no literals. The occurrence of the variable `x` in the two given clauses is, in effect, coincidental: the program will behave in the same way if the second clause has its occurrences of `x` replaced by `y`.

As for existence, the language relies on appropriate (Skolem) functions; existentially quantified variables are not acceptable. For example, the fact that for every `x` there exists a `y` greater than `x` is expressed with the following clause:

```
GREATERTHAN(f(x),x).
```

The dependence of `y` on `x` – the fact that as `x` varies, `y` may vary – is reflected in this clause. In contrast, when such a dependence is not present, as in the statement that there exists a nonnegative number `y` such that for all nonnegative `x`, `y` is less than or equal to `x`, a constant (function of no arguments) suffices:

```
LESSOREQUAL(a,x).
```

An important, and sometimes overlooked, subtlety of the language is the contrast between the use of functions and predicates. Functions require uniqueness (they must be unambiguous and well defined), whereas predicates do not. For example, a function can be used to convey information to the program about maternal grandfather because there is no ambiguity; but a predicate is required if the information is merely about grandfather because then there is ambiguity, in the sense that more than one choice exists. For a second example, the notion of successor (to a number) is unique and well defined, and a function suffices; but the concept of divisor requires the use of a predicate.

The clause language is a frugal language; and from the user's or researcher's perspective, a richer language would appear to be preferable. However, its lack of richness, rather than being a hindrance to the program and to automated reasoning in general, is an advantage. More richness would interfere with the formulation of computer-oriented inference rules, and, more important, with the introduction of powerful and effective strategies for controlling the reasoning. (*See* **Knowledge Representation**)

UNIFICATION

Computer-oriented reasoning differs from person-oriented reasoning in the former's emphasis on generality. The use of 'unification', a procedure essential to many aspects of automated reasoning including inference rules, illustrates this difference. Unification takes two expressions and seeks a most general substitution of terms for variables that, when applied to each expression, yields two (possibly) new expressions that are identical (Robinson, 1965b). In the earlier example about Plato, unification succeeded, yielding the substitution of `Plato` for both `x` and `y` to discover a contradiction. Of

course, unification can fail, as is the case with the following two clauses:

```
P(x,x).
-P(f(y),y).
```

No contradiction can be derived from these two clauses: ignoring the 'not' symbol, there is no substitution that can be applied to the two to yield identical expressions.

The following three clauses serve to illustrate the generality of a program's reasoning (in contrast to that often employed by a person):

```
-P(x,y,z)|Q(y,z,x).
P(a,b,u)
Q(b,z,a).
```

Given the first two clauses and the appropriate inference rule (that which unifies the first literals of the first two clauses), a reasoning program would correctly deduce the third clause. Various other conclusions could have been drawn by sound reasoning, for example the following:

```
Q(b,b,a).
Q(b,a,a).
Q(b,c,a).
Q(b,h(b,a,b),a).
```

However, these less general conclusions would not have been drawn by a reasoning program because of its preference for generality and because unification would not have yielded the appropriate substitution.

For a second example (of a type familiar to logicians): where the program might deduce a clause that (in effect) says that '*x* implies *x*', a person might prefer to deduce that '(*y* implies *z*) implies (*y* implies *z*)'. The latter deduction is an 'instance' of the former: it can be obtained by replacing the variable '*x*' by the term '(*y* implies *z*)'. A person's preference is often based on experience, intuition, and knowledge of what is likely to suffice for the task at hand; the program lacks all three of these advantages. Conversely, while the generality of the program's approach contributes to its effectiveness, such generality can actually interfere with a person's attempt to solve a problem.

INFERENCE RULES

Unification plays an essential role in inference rules. Any inference rule is required to be 'sound': to yield conclusions that follow inevitably from the hypotheses to which it is applied. Some inference rules apply to pairs of hypotheses in the form of clauses; some apply to more than two; and one important inference rule applies to a single hypothesis. One commonly used inference rule treats the relation of equality as understood (built in).

The inference rule that dominated the field in the early 1960s was 'binary resolution' (Robinson, 1965b), sometimes loosely called 'resolution'. This rule requires two clauses as hypotheses, focuses on one literal in each clause with the same predicate and opposite sign, and seeks to unify the chosen literals. If unification succeeds, the two literals are canceled (ignored), and the unifier (substitution) is applied to the remaining literals in both clauses, which are then used to form a new clause by taking their 'or', ignoring duplicates. The example above involving *P* and *Q* illustrates the use of binary resolution. For another example, consider the following clauses.

```
-P(a,x)|-Q(x).
P(y,b)|-R(y).
-Q(b)|-R(a).
```

The third clause can be obtained from applying the inference rule to the first two clauses.

As defined, binary resolution requires the presence of another inference rule, *factoring*, in order to guarantee that assignments that should be completable can be completed. In factoring, the program focuses on a single clause, chooses two of its literals that are alike in both predicate and sign, and seeks to unify the two. If the program is successful, it applies the unifier to the clause to yield the conclusion. For example, factoring applied to the first of the following four clauses yields (depending on the two literals being unified) three different conclusions, the second through the fourth.

```
Q(f(x),y)|Q(f(u),g(v))|Q(f(x),
  g(v)).
Q(f(x),g(v)).
Q(f(x),g(v))|Q(f(u),g(v)).
Q(f(x),y)|Q(f(x),g(v)).
```

As the following simple example shows, binary resolution without factoring may not find the desired refutation for two inconsistent clauses:

```
P(x)|P(y).
-P(u)|-P(v).
```

Among other commonly used inference rules, 'hyperresolution' (Robinson, 1965a) requires that the deduced clause be free of 'not'; 'UR-resolution' (McCharen, 1976) requires that the conclusion be nonempty and free of 'or'; and 'paramodulation' (Robinson, 1969) provides equality-oriented reasoning. Hyperresolution and UR-resolution can consider as hypotheses two or more clauses at a time; paramodulation focuses on pairs of clauses. These inference rules are generally much more

powerful than binary resolution. Many have been generalized to a class of 'linked inference rules' (Veroff and Wos, 1992), replacing syntactic criteria with semantic criteria for conclusion acceptance.

Paramodulation is an example of a type of reasoning well suited to the computer but ill suited to a person. Consider the two clauses below:

```
EQUAL(sum(x,minus(x)),0).
EQUAL(sum(y,sum(minus(y),z)),z).
```

Unification is applied to the term `sum(x,minus(x))` in the first clause and to the term `sum(minus(y),z)` in the second clause, yielding a substitution that asks for `x` to be replaced by `minus(y)` and `z` by `minus(minus(y))`. The application of this substitution produces two new clauses that contain two identical expressions. By the nature of equality, one can then substitute from the first new clause into the second new clause to obtain

```
EQUAL(sum(y,0),minus(minus(y))).
```

STRATEGY

The introduction of strategy has contributed more than any other component to the advance and successes of automated reasoning. Some strategies restrict the program's reasoning; some direct it; and some affect it in some other way. Without strategy, the program can continue rather aimlessly and without success, often becoming overwhelmed by information that proves useless for the task at hand. (Of course, the same is true for a person studying a deep question or hard problem.)

The various strategies that have proven powerful in automated reasoning are dependent neither on the clause language nor on the program that implements them. However, most reasoning programs implement only a small subset of strategies.

Of the restriction strategies, the 'set of support' strategy has proven the most powerful (Wos *et al.*, 1965). With this strategy, the user chooses a subset T of the input clauses, and the program is not allowed to draw a conclusion from hypotheses all of which are input clauses not in T. In other words, the program is restricted to drawing conclusions that are recursively traceable to one or more clauses in T. In effect, the input clauses not in T are used only to complete the application of an inference rule.

The set of support strategy meshes beautifully with proof by contradiction, in that a reasonable choice for T is 'not Q' where the theorem to be proved is 'P implies Q'. A more effective choice of T is 'not Q' together with those elements of P that are not basic information (axioms). Another possible choice for T, often more effective still, is just those elements of P that are not axioms.

A very different type of restriction strategy requires the user to place an upper bound on the complexity of deduced information that is retained. In the same spirit, the user can forbid the program to retain any conclusion in which more than k distinct variables occur, where k is chosen by the user. The program can also be forbidden to retain new conclusions that contain any term designated as undesirable. For example, if the function n denotes negation, the user may decide that new conclusions in which $n(n(t))$ is present for some term t are to be discarded. This strategy has enabled a reasoning program to discover proofs that had been missing for many decades, perhaps by exploring paths that were counterintuitive to human minds.

Other strategies direct a program's reasoning and focus. For example, R. Overbeek's 'weighting' strategy (McCharen, 1976) enables the user to impose knowledge, experience, and intuition on the program's actions. Templates can be included that (in effect) assign priorities to various types of term. For example, the user can instruct the program to give high priority to sums of products, lower priority to terms in which 'minus' occurs, and little or no priority to terms in which '0' occurs. (*See* **Learning from Advice**)

The 'resonance' strategy (Wos, 1995) also enables the user to impose knowledge, experience, and intuition on the program. The user supplies equations or formulae whose functional pattern (treating all variables as indistinguishable) is conjectured to merit preference.

A naive and in some sense natural direction strategy is breadth first ('first come, first served') search. Its opposite, depth-first search, is found in programs based on logic programming (see below). Although such programs can prove simple theorems quickly, for even moderately deep theorems they usually lack sufficient power. Nevertheless, a breadth-first search can occasionally prove profitable. The related 'ratio' strategy combines complexity preference (through weighting) with breadth-first searching. (*See* **Search**)

Unlike restriction strategies and direction strategies, the 'hot list' strategy (Wos and Pieper, 1999) enables the program to repeatedly revisit some user-chosen items among those input. For example, consider the theorem that asserts commutativity for rings in which $xxx = x$ for all x. If the clause equivalent of '$xxx = x$' is placed in the hot list, then each time a new clause is retained, and before any other reasoning action is taken, the hot list strategy

will apply the chosen inference rules to the new clause together with the clause equivalent of '$xxx = x$'. This strategy can radically reorder the space of retained conclusions and, in turn, the clauses chosen to drive the program's reasoning. Sometimes proofs that would otherwise have been out of reach are easily proved by a program using a hot list.

SUBSUMPTION AND REDUNDANCY

The fact that 'Kim's daughter is female' is a trivial consequence and instance of the fact that 'everybody's daughter is female'. If the latter information is present and the former deduced, 'subsumption' (Robinson, 1965b) will purge the less general bit of information. The program prefers generality. If both bits of information were retained, then there would be redundancy, which would interfere with effectiveness. Subsumption significantly reduces the amount of information useless to the program among the set of conclusions that can easily be deduced and retained.

To discover whether one clause subsumes a second, the program seeks a substitution of terms for variables in the first that yields a subset of the second, even if the first clause consists of more literals than does the second. In the following example, the first clause subsumes the second:

```
P(a,x)|P(y,b).
P(a,b).
```

DEMODULATION AND SIMPLIFICATION

Although a person often automatically simplifies expressions (for example, replacing '$0 + t$' by 't', where t is a term) a reasoning program does not, unless instructed to do so. Therefore, if appropriate actions are not taken, there can be much semantic redundancy within the database of deduced information (for example, '$a + b = c$', '$a + b = c + 0$', '$0 + a + b = c$'). Subsumption cannot deal with this type of redundancy. Rather, it is avoided by means of 'demodulation' (Wos *et al.*, 1967).

The inclusion in the input of appropriate demodulators (rewrite rules) enables a reasoning program to simplify newly deduced conclusions and to put them in canonical forms. For example, the program can automatically right-associate expressions, rewriting '$(a + b) + c$' as '$a + (b + c)$'. Sometimes, the set of input demodulators can be supplemented by new ones deduced during the program's study of a question or problem.

SOUNDNESS AND COMPLETENESS

Inference rules must be sound; that is, any conclusion that is drawn must follow inevitably from the hypotheses to which the rules are applied. (Of course, a false conclusion can be drawn if the program is given false information, such as 'finches are mammals'.) This soundness property guarantees that the proofs the program finds are without flaw, barring a bug in the program or the inclusion of erroneous items. However, in case one suspects that a supplied proof is not sound, many programs can provide detailed accounting of their reasoning.

Typically, automated reasoning programs may be refutation complete; that is, they may be configured in such a way that when given an inconsistent set of statements, they will, in theory, eventually complete a proof by contradiction, although an infeasible amount of time may be required in practice. Such completeness is often sacrificed for the sake of effectiveness. Indeed, many combinations of strategies, inference rules, and procedures are not refutation complete but are nevertheless widely used. In such cases, the counterexamples to completeness are usually bizarre and not relevant to serious studies. The set of possible proofs of a theorem is often so large that the absence of refutation completeness is not a practical problem.

'Deduction-completeness' is the guarantee that a consequence of the given hypotheses will eventually be deduced. Few reasoning programs are deduction complete. For example, even though Kim is mentioned in a puzzle and the program is told that everybody is female or male, the program is unlikely to explicitly deduce that Kim is female or male. The inference rule of instantiation would suffice to draw this conclusion from the more general fact, but that rule is rarely offered by a reasoning program because of the difficulty of adequately controlling its use. There is usually an infinite set of instances, and often it is not the simplest instance that is needed.

LOGIC PROGRAMMING

Logic programming languages have instruction sets based on the direct use of clauses. PROLOG is the best known. There are many similarities between logic programming and automated reasoning (Wos and McCune, 1991). In the mid-1980s, some researchers thought that the most effective way to automate logical reasoning was to base the design of a program on logic programming.

Such a program is very fast at proving simple theorems (Stickel, 1988). However, it lacks the power needed to tackle deep questions and hard problems. One of the main drawbacks is the absence of a diverse set of strategies. Implicitly, such a program does rely on a depth-first search and on the set of support strategy; but many additional strategies are needed in order to achieve the power required to prove interesting theorems.

Another serious drawback is the lack of information retention. Experiments strongly suggest that a program must retain new conclusions if it is to succeed in tackling deep questions and hard problems.

SUCCESSFUL APPLICATIONS

Despite the obstacles that seemed to many to be insurmountable even in the late 1970s, automated reasoning has prospered. Two distinct areas of success should be mentioned: design and verification, and mathematics and logic. (*See* **Problem Solving**)

The Boyer–Moore theorem prover has been used to prove the invertibility of the Rivest–Shamir–Adleman public key encryption algorithm (Boyer and Moore, 1984). Several major chip manufacturers now use automated reasoning programs to facilitate chip design by proving theorems.

In mathematics and logic, open questions have been answered, missing proofs found, and existing proofs improved upon in various ways. Of the open questions that have been answered, the most famous success was the proof that every Robbins algebra is a Boolean algebra (McCune, 1997). More recently, new two-axiom systems for Boolean algebra using the Sheffer stroke have been found (Veroff, 2001). Automated reasoning programs have also made substantial contributions to the theory of groups (Hart and Kunen, 1995; Kunen, 1992), combinatory logic (Glickfeld and Overbeek, 1986; Wos, 1993), abstract algebra (McCune and Padmanabhan, 1996), and logical calculi (Harris and Fitelson, 2001; Fitelson and Wos, 2001).

Perhaps the future will witness a combination of applications from each of these two areas: for example, the application of proof-shortening techniques used for mathematics and logic to the design of simpler chips and circuits.

Acknowledgment

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, US Department of Energy, under Contract W-31-109-Eng-38.

References

Boyer R and Moore JS (1984) Proof checking the RSA public key encryption algorithm. *American Mathematical Monthly* **91**: 181–189.

Fitelson B and Wos L (2001) Missing proofs found. *Journal of Automated Reasoning* **27**(2): 201–225.

Glickfeld B and Overbeek R (1986) A foray into combinatory logic. *Journal of Automated Reasoning* **2**: 419–431.

Harris K and Fitelson B (2001) Distributivity in L_{N_0} and other sentential logics. *Journal of Automated Reasoning* **27**: 141–156.

Hart J and Kunen K (1995) Single axioms for odd exponent groups. *Journal of Automated Reasoning* **14**: 383–412.

Kunen K (1992) Single axioms for groups. *Journal of Automated Reasoning* **9**: 291–308.

McCharen J, Overbeek R and Wos L (1976) Complexity and related enhancements for automated theorem-proving programs. *Computers and Mathematics with Applications* **2**: 1–16.

McCune M and Padmanabhan R (1996) *Automated Deduction in Equational Logic and Cubic Curves*. New York, NY: Springer-Verlag.

McCune W (1997) Solution of the Robbins problem. *Journal of Automated Reasoning* **19**: 263–276.

Robinson G and Wos L (1969) Paramodulation and theorem proving in first-order theories with equality. In: Meltzer B and Michie D (eds) *Machine Intelligence*, vol. IV, pp. 135–150. Edinburgh, UK: Edinburgh University Press.

Robinson JA (1965a) Automatic deduction with hyper-resolution. *International Journal of Computer Mathematics* **1**: 227–234.

Robinson JA (1965b) A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery* **12**: 23–41.

Stickel M (1988) A Prolog technology theorem prover. *Journal of Automated Reasoning* **4**: 353–380.

Veroff R and Wos L (1992) The linked inference principle, I: the formal treatment. *Journal of Automated Reasoning* **8**: 213–274.

Veroff R (2001) Solving open questions and other challenge problems using proof sketches. *Journal of Automated Reasoning* **27**: 157–174.

Wos L (1993) The kernel strategy and its use for the study of combinatory logic. *Journal of Automated Reasoning* **10**: 287–343.

Wos L (1995) The resonance strategy. *Computers and Mathematics with Applications* **29**: 133–178.

Wos L and McCune W (1991) Automated theorem proving and logic programming: a natural symbiosis. *Journal of Logic Programming* **11**(1): 1–53.

Wos L and Pieper GW (1999) The hot list strategy. *Journal of Automated Reasoning* **22**: 1–44.

Wos L, Robinson G and Carson D (1965) Efficiency and completeness of the set of support strategy in theorem proving. *Journal of the Association for Computing Machinery* **12**: 536–541.

Wos L, Robinson G, Carson D and Shalla L (1967) The concept of demodulation in theorem proving. *Journal of the Association for Computing Machinery* **14**: 698–709.

Further Reading

Chang C and Lee R (1973) *Symbolic Logic and Mechanical Theorem Proving*. New York, NY: Academic Press. [A thorough treatment of the clause language paradigm, and proofs of the theorems that establish the necessary logical properties of various inference rules and strategies.]

Fitting M (1996) *First Order Logic and Automated Theorem Proving*. Berlin, Germany: Springer-Verlag. [A rigorous graduate-level text that presents fundamental concepts and results of classical logic.]

Kalman J (2001) *Automated Reasoning with Otter*. Princeton, NJ: Rinton. [Covers such topics as how to convey a problem to an automated reasoning program, how to find a proof by contradiction, and how to reason about equality. Includes a CD-ROM with the automated reasoning program OTTER (designed and implemented by W. McCune).]

Loveland D (1978) *Automated Theorem Proving: A Logical Basis*. Amsterdam, Netherlands: Elsevier. [A classic text.]

Siekmann J and Wrightson G (1983) (eds) *The Automation of Reasoning: Collected Papers from 1957 to 1970*. New York, NY: Springer-Verlag. [A history of the field and relevant papers.]

Wos L and Pieper GW (1999) *A Fascinating Country in the World of Computing: Your Guide to Automated Reasoning*. Singapore: World Scientific. [An introductory book that guides the reader through the basics of automated reasoning and its applications and presents numerous open questions and research topics. Includes a CD-ROM with the automated reasoning program OTTER (designed and implemented by W. McCune).]

Wos L and Pieper GW (2000) *The Collected Works of Larry Wos*. Singapore: World Scientific. [A history of the field and a convenient source for all of the papers by Wos cited in this article.]

Wos L and Veroff R (1992) Resolution, binary: its nature, history, and impact on the use of computers. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 1341–1353. New York, NY: John Wiley. [A review of the field.]

Wos L and Veroff R (1994) Logical basis for the automation of reasoning: case studies. In: Gabbay DM, Hogger CJ and Robinson JA (eds) *Handbook of Logic in Artificial Intelligence and Logic Programming*, pp. 1–40. Oxford, UK: Oxford University Press. [A review of the field.]

Robotics

Introductory article

Luc Steels, Vrije Universiteit, Brussels, Belgium
Andreas Birk, International University, Bremen, Germany

CONTENTS

Robotics is the science of the construction and control of programmable machines suitable for flexible handling in various physical tasks like manipulation and locomotion.

INTRODUCTION

The science of robotics as well as the concept of a robot itself is not well defined from an industrial nor a scientific viewpoint. The Robotics Industries Association (RIA) defines a robot as a 're-programmable multi-functional manipulator designed to move material, parts, tools, or other specialized devices through variable programmed motions for the performance of a variety of tasks'. Nowadays, almost any device used in automation fits this definition to some degree. But the prototypical idea behind this definition is the most common form of a robot, the robot arm. The main applications of robot arms in industry are production tasks like welding and assembly, followed by packaging and storage tasks. Industry automation is still by far the largest market for robots, though slowly, new applications are on the rise like medical robotics or service and entertainment robots.

From a cognitive science point of view, robots are a platform for developing and testing precise theories on how sensorimotor behaviors might work. They do not prove whether certain mechanisms are used by animals or humans but they help to examine the plausibility and consequences of these mechanisms. At the same time ideas from the empirical study of living systems, both their capacities and the underlying brain structures and processes, are a continuing source of inspiration for building new types of robots and new software systems to control them. So there is a profound interaction between the robotics research carried out by artificial intelligence workers and the investigations of sensorimotor intelligence in other areas of cognitive science.

The History of Robotics 1: Science Fiction

In contrast to the industrial reality of the robot arm as the most common robotic device, the popular notion of a robot is strongly influenced by science fiction, which even coined the name of the device as well as the name of the field. The term 'robot' was introduced in 1921 by the Czech writer Karel Capek in his satirical drama '*R.U.R., Rossum's Universal Robots*'. The robots in this play are artificial humanlike creatures forced to work; the Czech word 'robota' means work. Capek's robots are, from a science fiction viewpoint, not really robots but Cyborgs as their control or 'brain' was supposedly based on re-engineered biological matter.

The field of robotics was invented by another science fiction writer called Isaac Asimov. His stories dealt mostly with the philosophical and moral implications of artificial humanlike creatures. The main ingredients of Capek's and Asimov's vision include a human-shaped body, superhuman strength, high cognitive capabilities, and a devotion to serve or – depending on the genre – to destroy mankind. Popular perception is strongly influenced by these visions but it is far removed from the reality of robotics research.

The History of Robotics 2: Beginnings

The industrial roots of robotics started in 1958 with the foundation of a pioneering company called Unimation (Universal Automation). Unimation produced a five-axes hydraulic robot arm based on a patent by George Devol and financed by Joseph Engelberger. From then on, robot arms started to become more and more important in automation.

In the scientific field of robotics, there was an early connection with artificial intelligence (AI). In

the Stanford Artificial Intelligence Laboratory (SAIL), founded in 1965, a robot arm was built in 1968 and very soon mobile robots followed, for example the famous robot Shakey, built in the Stanford Research Institute (SRI) from 1966 onwards, which was designed to receive instructions in natural language, and then plan and carry out a trajectory of movements in a real-world environment perceived through a camera.

Shakey was equipped with quite elaborate sensors for that time. Cameras, wireless communication with a host computer, optical range sensors, and bumper sensors were all pioneered on this robot. Most of the processing was done remotely on a DEC PDP-11, a supercomputer at that time, but less powerful than today's hand-held computers. Given the available computer power it is not surprising that the robot's behavior was very slow and hence observers viewed it as a failure. Nevertheless, Shakey was a major advance in the field, particularly from the viewpoint of overall architecture, and interaction between sensorimotor intelligence and cognition. Most subsequent projects explored the same general principles, constantly improving all aspects of the system and aided by steady advances in all enabling technologies of robotics: batteries, sensors, computers and mechanical components.

The History of Robotics 3: Successful Applications and Humanoid Robots

From the early 1990s, the cumulative effect of all these advances gave rise to a steady stream of functioning robots. Most of them were designed for specific purposes: cutting the grass, inspecting pipes, washing windows in high-rise buildings, checking and repairing telephone wiring, and of course various forms of factory automation. The advances have started to give rise to new applications for handicapped humans, such as artificial arms, hands, legs, etc. that are sometimes driven by signals captured from neuronal circuits. Research activities have also benefited from the massive advances in enabling technology. It has become feasible to build robots that test a theory on how a certain biological organism performs specific sensorimotor functions. This has given rise to the field of animal robotics, which operates in close collaboration with ethology and neuroscience. Only low-level organisms, typically insects, have been studied extensively so far but one can expect a steady rise in complexity until mammals or aspects of human sensorimotor intelligence come within reach.

The most recent exciting development concerns humanoid robots or humanoids. These are robots with a human shape (two legs, two arms, a torso, and head with two cameras), designed to exhibit human-like behaviors such as biped walking, sharing of attention by eye gaze following and gesture recognition, etc. The first humanoids have now been demonstrated in the laboratory and the first applications can be seen in the area of entertainment robotics.

The remainder of this article divides robotics research into five areas: the execution of action; the interpretation of the world through various sensors; behaviors that integrate sensing and actuating into useful building blocks; the architecture that integrates sensing and actuating into coherent goal-driven behavior adapted to the environment; and interactions among robots.

EFFECTORS, ACTUATORS, AND DEGREES OF FREEDOM

Any robot interacts with the physical world in one way or another, typically by manipulating objects or by changing their location. The parts of a robot designed for this purpose are known as effectors. An end-effector is the effector at the end of a robot arm, for example in the form of a gripper.

Actuators

Actuators provide the mechanical power that drives the active joints of a robot. There are three main classes of actuators, based on electromechanics, hydraulics, and pneumatics.

Hydraulic and pneumatic actuators are mainly used for linear motion based on pistons, though some special-purpose rotational motors of both kinds exist. Hydraulics are mainly used in applications where very high forces are needed, especially in large industrial robot arms. Pneumatic actuators suffer from the fact that air is – in contrast to hydraulic liquids – compressible, which makes control more difficult. Nevertheless, pneumatic actuators are commonly used as a lightweight alternative to hydraulics in applications where electrical motors cannot be used. Examples are environments where the potential sparking of electrical motors is a hazard, like chemical plants, or environments where sensitivity to radiation plays a role, like nuclear power plants.

A special form of pneumatic actuator is the pneumatic muscle, or McKibben actuator, which consists of an extendable bladder with a net woven around it. When the bladder is inflated,

the muscle evenly contracts and exhibits a pulling force at its endpoints.

Electric motors are the most common source of mechanical power for robots. Almost all electric motors are based on electromagnetism. The few exceptions are, for example, motors based on the piezo-effect – that certain ceramics change their shape when a voltage is applied to them – and electrochemical effects of some experimental classes of polymers.

Electric motors are seldom used on their own as actuators; exceptions are so-called direct drive set-ups. Electric motors tend to provide high rotational speed and low torque. Therefore they have to be combined with gears to be useful as an actuator. Gears serve several purposes. First, they can be used to convert speed to torque. Second, they can convert the type of motion, for example changing a rotation into a translation, or inverting the direction of a rotation. The main problems with gears are power and precision losses due to friction and the backlash, i.e. the small imprecision between the teeth of intermeshing gears.

Robot Arms

Robot arms are the prototypical industrial robots used for manipulation. Their usage is dominated by three tasks, namely welding, assembly, and packaging.

The structure of robot arms is similar to that of human arms with shoulder, elbow, and wrist. Each of these parts can be implemented in different ways, leading to different possible forms of motion. Each part of a robot arm can have several degrees of freedom (DOF). The most common approach is two DOF in the shoulder, one DOF in the elbow, and two to three DOF in the wrist. To the wrist, a hand or other end-effector is attached, which itself has additional degrees of freedom, ranging from a single one in a gripper to dozens in some so-called dextrous hands.

Components for Wheeled Locomotion

Most mobile robots use wheeled locomotion with a differential drive, although legged robots are becoming more and more common. A differential drive consists of two actively driven wheels on the same axis. Typically, this axis crosses the center of the drive platform. One or sometimes two passive, so-called castor-wheels are used to support the drive platform. A differential drive has many advantages. It can turn on the spot, it is easy to control, and the mechanical design is rather simple. Therefore, it is quite popular for autonomous devices, especially for mobile robots.

The Ackerman drive consists of four wheels on two axles, as in a car. The rear axle supplies the propulsion for the vehicle, powered by a single motor. The front axle is used for steering. For the steering operation, the two passive front wheels can be set in two steering angles, reducing slippage of the wheels.

When turning, not just one of the passive front wheels but also one of the rear drive-wheels has to turn faster than the other. This problem is solved by a passive mechanical solution in the form of a so-called differential gear. The Ackermann drive is useful for cars with a human operator at the steering wheel and a single combustion-based engine for propulsion. But it is much less suited for robots as the mechanical set-up is rather complex and it is difficult to control. Its main advantage is that it needs a single motor for propulsion, which can be quite powerful.

The syncho-drive is especially designed for mobile robots. Each of its three or four wheels is actively driven and steered. So the driving force is evenly applied to the ground, preventing problems from passive castors or steering wheels. The disadvantage of the synchro-drive is its mechanical complexity. Every wheel has to be actively oriented in the same direction, and it has to turn with exactly the same speed as the others.

The synchronization of the rotation and angle of the wheels is achieved by mechanical coupling through two belts. A single propulsion and single steering motor drive each belt respectively. The problem is to transmit the propulsion to the wheels while being able to steer them. This is done by special drive units. They consist of metal tubes through which an axle transmits the propulsion power vertically. At the lower end of this axle, a conical gear transmits power to the wheels. On the outside of the tube, a gear is mounted that allows the tube to rotate with the wheel axle, thus enabling the robot to be steered.

SENSORS: VISION, PROPRIOCEPTION, FORCE SENSING, TACTILE SENSING, SONAR

Initially, robots were primarily open loop devices. Detailed models of a robot were used to compute the necessary activations of actuators to reach a given target objective like, for example, the desired position of an end-effector in a welding application. Closed loop control, in contrast, uses error information derived via sensors to achieve the target

objective. Closed loop control is inherently more stable than open loop control and it frees the designer to some extent from the onerous task of modeling the robot in minute detail. Moreover sensor technology has greatly benefited from the developments in microelectronics making sensors smaller, mass production making them cheaper, and information processing making the output of sensors more reliable and ready to use. Consequently closed loop control is now more and more common in industrial robotics.

Vision

Most sensor systems are tailored to register a particular physical property, like infrared reflection, and hence are useful only for a particular task, such as obstacle avoidance. A vision system, in contrast, is more versatile. Visual stimuli can be used in many ways, for example to estimate distance to objects as well as for object recognition. Vision sensors or cameras provide a flow of large amounts of data with detailed information about the outside world and strong spatial and temporal resolution. The processing of this data is typically done by software, which can be changed and adapted much more easily than the more hardware-dependent processing of other sensor systems.

This advantage of being based on large amounts of data from the environment is also at the same time the major downside of vision. It is far from obvious how meaningful information can be extracted from the mass of data that a camera generates. In addition, it is questionable whether vision is just a kind of sensor that passes information to higher cognitive functions or whether vision is strongly embedded in cognition and action, a position known as active vision.

Vision is usually structured into the following stages. The first stage is image acquisition, where an image as an array of pixels of color or grayscale values is captured as a snapshot of the environment. Image processing is the second stage, where a series of transformations on the original image is performed. At every step, a new image is generated out of images from previous processing stages. Noise is eliminated and interesting properties are amplified, for example through edge detection, segmentation, shape from shading, etc. The third stage is high-level vision where an image is transformed into an abstract representation that can be used for planning or natural language dialogs.

Research on vision is typically divided into machine vision and artificial vision. The field of machine vision deals with vision from an application-oriented perspective, whereas artificial vision is more oriented towards cognitive aspects. The differences in the emphasis of both fields manifests itself mainly in the level of image acquisition. In machine vision, structural changes to the environmental set-up are commonly exploited, especially in commercial robotics applications. For example, special illumination eases segmentation, structured light improves the estimation of distances, background illumination enhances the contrast between objects and their surroundings, artificial landmarks are put in the environment, etc.

For both fields, machine vision and artificial vision, image acquisition and image processing are regarded as the least difficult parts. Especially for image processing, there are a large number of well established techniques like thresholding of pixel values or edge detection procedures. On the other hand, the extraction of qualitative descriptions from a visual image is extremely difficult, mostly because there is usually not enough information in the image itself to perform the interpretation. So context, prior expectation, and convention must be used in a top-down manner to guide the vision system, which is why many researchers now argue for active vision.

Proprioception

Proprioception is concerned with locating the parts of a mechanism in space. It is the basis for kinematic and dynamic control of robot arms as well as localization and navigation of mobile robots. Proprioception relies on sensors for measuring the positions of motors. These are typically encoders mounted on the drive shaft.

There are two main types of encoder, absolute and incremental. An absolute encoder detects the absolute angular position of the axle. The output of an incremental encoder is a pulse whenever the axle has turned by some fixed amount, so the output is a rectangular wave with a frequency proportional to the rotational speed of the axle. Incremental encoders typically have two output signals with a 90 degrees phase shift, enabling the detection of the direction of rotation of the axle. An incremental encoder often has an additional index channel that outputs a single pulse at one fixed absolute position for calibration purposes.

Encoders are technically implemented in many different ways. Mechanical solutions, such as those based on switches, are very rare for professional equipment due to wear problems. Serious encoders are contactless and based on optical sensors, such as via disks with slits or black and white patterns.

For mobile robots, direction is important proprioceptive information. When using shaft encoders on the drive motors and dead reckoning (vector addition of driven distances), there is a cumulative error: the longer the robot drives, the less precisely its position and orientation are known. Therefore, special purpose sensors like magnetic compasses and gyroscopes are often used in addition to shaft-encoders to measure direction.

Force Sensing

There are two main types of force sensor: strain gauge and pressure sensors. Strain gauges are based on Lord Kelvin's discovery in 1856 that the resistance of a conductor is proportional to its length and inversely proportional to its area, depending on a conductivity constant of the material. So, when a conductor is strained, its length increases and its area decreases thus leading to an overall increase of its resistance. To amplify this effect, a strain gauge contains a wire which is looped several times over the carrier material so that several parts of the wire are strained at the same time. The carrier material is usually some sort of plastic which protects the wire from damage. The quality of the sensor strongly depends on this material, because cheap versions wear out very quickly. Strain gauges are typically based on semiconductors to minimize the material constant, such that the overall resistance of the component is strongly dominated by the actual strained part.

Elastomers (compressible plastics) are a popular basis for pressure sensors. With the addition of metallic particles, elastomers can be made conductive. When this material is sandwiched between two electrodes, the resistance of this sensor decreases when it is pressed. Two effects are taking place. First, the distance between the two electrodes decreases and so, linearly, does the resistance. In addition the conductivity of the elastomer increases as the conductive particles come closer to each other. Depending on the design of the sensor, this second effect can lead to nonlinear sensitivity of the sensor.

The force sensors presented so far have in common that the range of forces to which they react is defined through the properties of the underlying materials of the sensors. In addition, it is possible to measure forces based on springs. A wide range of springs is available, allowing sensors for very small as well as very large forces. Within the admissible range of a spring, a force leads to an extension of the spring proportional to the strength of the force. This extension can be measured for example with a linear potentiometer or with a shaft encoder.

Tactile Sensing

Getting 'in touch' with the outside world is one of the most basic needs for a robot. The most simple touch sensors are switches. There are very small microswitches as well as very large ones that can handle high currents. Also the mechanical activation part of switches can differ substantially. The two most common forms are buttons and levers that react when they are pressed. The main disadvantage of switches is that they are mechanical, so they suffer from wear.

Proximity sensors react to the presence of objects that are very close without being in direct contact. Proximity sensors are typically used like switches, but they have the major advantage that they do not suffer from mechanical wear. The major disadvantage of proximity sensors is that they are limited with respect to the materials that they can sense. Therefore, they are mainly used in industrial applications where the parameters of the objects involved are known.

Tactile mats or 'artificial skins' are sensors that provide high-resolution information about contacts with a surface. They are still a basic research topic as they suffer from fundamental unsolved technical problems, especially with respect to manufacturing and meaningful data processing.

Sonar

Sonar sensors are active sensing systems. They emit a signal to probe the environment. The 'speaker' emits a short sound, called a ping, and a microphone listens for a reflection of the sound from the nearest surface. The ping is in the ultrasound range for convenience as these frequencies are too high for humans to perceive. Some animals, especially dogs, are nevertheless very sensitive to these frequencies.

Based on the time-of-flight between emission and reception of the sound, and the speed of sound in air, the distance to the object that caused the reflection can be calculated. In addition, it is possible to use the Doppler-effect to measure the relative speed between the sensor and the object. If both move towards each other the sound is compressed and its frequency increases. When they move away from each other the opposite effect takes place and the frequency decreases. Bats are heavy users of sonar sensing, but exploit the Doppler-effect and not the time-of-flight.

Sonar can be used, for example, on mobile systems for obstacle avoidance and map building by accumulating distances to objects from different positions and angles. In addition, sonar can be used for absolute positioning based on beacons. There are two ways to calculate the time the sound takes to travel from the beacon. Either a calibration signal like a radio pulse is emitted to mark the time when the ping is sent. Or the pings are sent in a fixed temporal pattern and the time differences between received pings are used to calculate relative position changes. Next, the actual position has to be determined. With one beacon and two receivers, distance from and orientation towards the beacon can be determined. With two or three beacons using different sounds, simple or proper triangulation can be used to determine the absolute position of the robot.

Issues in Sensing

The problems involved in interpreting sonar sensor data are representative for the key issue in sensing, namely how can a sufficiently accurate interpretation of the world be obtained to engage in appropriate action. In general it is never straightforward and usually impossible to derive absolutely reliable accurate information about the environment in sufficient time.

Thus for sonar sensing, there is the problem of reflection strength. The echo of the sound heavily depends on the type of surface from which it is reflected. Best suited are plain, straight walls. Corners or walls covered with an irregular structure, like a bookshelf, swallow the sound, so there is insufficient reflection. Next there is the problem of multiple echoes. There can be several reflections of the same ping from different objects located at different distances. Therefore, the system has to wait between repeated samples until all echoes of the first ping have (most probably) died out. As a result, the sampling rate is rather low. This is especially a problem when sonar is used for obstacle avoidance on a fast-moving robot. Third, sonar sensing gives low angular resolution. The opening angle of an ultrasound sensor is rather large. If the robot features several sonar sensors then either different frequencies for each sensor have to be used or the sensors have to be activated one after another, causing an additional slow-down of the sampling rate. Next, there is the problem for any active sensing method, that several similar systems may cause sensor pollution; that is, the active signals from different systems may cause interference. Therefore, the sound frequencies have to be adjusted if several systems use ultrasound in the same environment.

Finally, ultrasound sensors are affected by environmental conditions. The speed of sound in air is not constant, but depends on the temperature, the humidity, and atmospheric pressure. Within office environments these factors are nearly constant. But for precise operations, particularly outdoors, the time-of-flight calculations have to be calibrated. The most critical effect on the speed of sound is usually caused by temperature changes.

These various problems are encountered in all types of sensing. One of the main lessons of robotics is that the relationship between the physical world and its interpretation by a cognitive agent is very complex.

BEHAVIORS

Behaviors integrate sensing and actuating into a coherent functionality that can be reliably called upon by higher-level modules. Typical examples of behaviors are manipulation, locomotion, and navigation. Each behavior requires some representation of the world and sensing to fine-tune the behavior while it is being performed.

Manipulation

Kinematics is the general science of mechanical motion. On an abstract mechanical level, a robot consists of stiff parts called links and movable parts called joints. Active joints are powered by actuators, otherwise they are called passive. The degrees of freedom (DOF) of a mechanical system is the number of variables needed to locate every part of the system in a fixed coordinate system. A rigid body has, for example, six degrees of freedom in a general three-dimensional coordinate system, namely three DOF for its position along the x-, y- and z-axes and three DOF for its angular orientation along each of these axes. The rotational components are commonly denoted by the roll, pitch, and yaw of the body. The combined vector of the location and orientation components is known as the *pose* of the body.

Given a precise mechanical model of a robot, kinematics attempts to answer two questions. First, there is the problem of forward kinematics: given a value for each of its DOF, compute the resulting pose for at least one part of the robot. Typically, this part is the end-effector in the case of robot arms. Second, there is the problem of inverse kinematics: given the desired pose of at least

one part of the robot, compute a feasible set of values for each DOF that leads to this pose.

The state of an n-DOF robot is characterized from a kinematics viewpoint by n real-valued numbers, each of them describing the momentary state of the robot according to its DOF. The state or 'configuration' of the robot can thus be seen as a point p in the n-dimensional space of real numbers $\mathbb{R}^n$. The subspace of $\mathbb{R}^n$ that consists of all feasible states of the robot is known as the *configuration space*. A movement of the robot from one state A to another state B is hence possible only if there is a continuous trajectory from A to B in the configuration space.

Obstacles in the real world can be mapped to parts of the configuration space. The remaining part of the configuration space that is not occupied by obstacles is the configuration free space, or free space for short. Finding a feasible motion of the robot corresponds to finding a trajectory in the free space. This approach is most feasible when the obstacles are static, as the free space has to be computed only once.

The term 'generalized configuration space' is applied to methods where the DOF of objects other than just the robot are included in the configuration. These objects are obstacles that move, or objects that change their shape due to manipulation by the robot. The configuration space of a general rigid object in 3D is the six-dimensional space of its pose; i.e., its location and orientation. The generalized configuration space of parts A and B is the Cartesian product of each individual generalized configuration space. It follows that the number of dimensions and the size of the generalized configuration space is usually large.

There are several ways to deal with this problem. One possibility is to partition the space finitely into many states to which standard planning techniques can be applied. The problem with this approach is that no general method for this partitioning is known. A second possibility is to plan the object motions first and then plan the robot motions. The main problem is that this approach is not complete, so it might lead to an omission of feasible solutions. Finally, there is the third possibility: to place constraints on the motions of the objects to simplify planning.

Walking

Walking is a fascinating option for the locomotion of robots. Unfortunately, the technology of artificial walking is still in its infancy although some impressive results have been achieved. Walking allows robots to negotiate rough outdoor terrain, and is required for access to standard indoor environments. A wheeled robot cannot for example cope with a simple staircase.

The main problems with walking are its mechanical complexity and its energy consumption. The mechanical complexity is high because many degrees of freedom are needed by default. A minimal six-legged walker has 12 DOF, namely two in the hip of each leg; and a two-legged walker typically needs at least 10 DOF, namely two in the hip, one in the knee and two in the ankle. Energy consumption is high for several reasons. First, walking includes, in addition to propulsion, some lift. Second, it is very difficult to control the dynamics of walking so that not too much energy is wasted in the form of heat, for example through stress. Nature has optimized natural walkers tremendously in this respect, using not only active components like muscles, but also passive components like tendons to absorb and reuse energy.

There are two basic forms of walking: static walking and dynamic walking. In the case of static walking, the center of gravity of the walker is permanently within the support area, that spanned by the parts of the robot that touch the ground. In the case of dynamic walking, the center of gravity is not supported constantly. Therefore, dynamic walking is also called 'controlled falling'. Static walking is obviously much easier to control. In a static environment, there are no time constraints on static walking. The system can be switched off at any time in the walking process without falling over.

Dynamic walking is more difficult as severe timing constraints apply and the dynamics as well as the kinematics of the system have to be taken into account. But dynamic walking can be much more energy-efficient than static walking. Therefore, it will probably be adopted for autonomous robots in the long run.

Navigation

Navigation is getting from one location to another. It implies two other tasks, namely mapping and localization. *Mapping* deals with the construction of a spatial representation of the environment. *Localization* deals with the determination of the current spatial position of the robot in the map. Navigation can be structured into furthera behaviors, namely path planning and obstacle avoidance.

Maps

Maps are world models used by mobile robots to navigate through their environment. They restrict the representation of the environment to a two-dimensional plane. The 'free space' is the part of the environment where the system can move without restrictions, that is where no obstacles hinder the free movement of the system. Landmarks are locations in the free space which can be distinguished by the sensors. They serve as orientation points for navigation.

The representation of obstacles in maps is usually chosen such that the robot can be treated as a point. First a virtual boundary for the mobile system is chosen that represents it as a sphere of a fixed radius. Checking collisions between the virtual boundary and obstacles is efficiently done by expanding the obstacles by the radius of the robot and treating the robot as a point.

The two major types of maps currently used are Euclidean and topological maps. With a Euclidean map, each point in space is represented according to its metric distance to all other points in the space. Given two locations in a Euclidean map, their exact metric distance can be determined. A topological map represents locations and their adjacency. This representation does not necessarily include an exact metric.

Topological maps are typically represented as a graph. The nodes are distinct places and the edges represent adjacency relations between locations. A labeled edge from A to B can for example represent by which type of motion the robot gets from A to B, but without exact distance information. In this case, the topological map represents with which action the system can transit from one node to an other.

One option for Euclidean maps is to use a list of the exact locations and boundaries of the obstacles. Polygons are a common form of representation of the obstacles. Another alternative is to use a decomposition of the plane into cells. In doing so, each cell is labeled as free space or obstacle. A grid is a uniform partition of the space. It is used as a basis for efficient processing in many algorithms. The quadtree representation is an example of an uneven partition of the space into cells. Computations on this form of representation are usually lengthy for path planning, but the memory requirements are much less than needed for a grid.

Path Planning

A map is only useful if it is combined with appropriate algorithms for navigation. In the case of topological maps, where the environment is represented by graphs, general graph search techniques, like Dijkstra's algorithm for finding the shortest path, can be used.

For a Euclidean map, it is possible to compute a graph-like representation that is also known as a skeleton. This can, for example, be a Voronoi diagram or a visibility graph. A Voronoi diagram consists of curves, where each point on the curve is equidistant from two or more obstacles. In a visibility graph, the obstacles are typically represented as polygons. The corners of the polygons are the nodes of the graph. Two nodes are connected via an edge if and only if they are visible to each other; that is, a straight line connection between them goes only through the free space.

Potential field methods constitute another family of approaches to the path planning problem. Attraction and repulsion obtained from electrical fields are used on a grid representation of the world. The goal is modeled as an attractor and obstacles have repellent fields. The different fields overlay each other and give a gradient towards the goal while avoiding obstacles. The main difficulty with potential fields is that the robot might get stuck in a local optimum.

ARCHITECTURE

The role of the robot's architecture is to invoke appropriate behaviors to satisfy certain goals. This typically involves planning. Two kinds of architectures have been proposed. The classical or logic-based planners start from a description of the goal and use logical inference to decompose the goal into subgoals based on a description of the pre- and postconditions of each action. The preconditions describe what the state of the world has to be before an action can be performed and the postconditions describe the effect of an action. The decomposition generates a search space when there is more than one way to decompose a goal into subgoals, and heuristic criteria must be used to find the fastest or optimal solution.

This very general and high-level approach is very effective when a logical description of the world is available but has been proven problematic on real-world robots for several reasons.

First of all, the real-world environment of a robot is much too complex for any detailed *a priori* planning. Hence, planning for real-world robots must include options to react to environmental conditions during the execution of the plan. Second, computation resources – processing power, disk space, etc. – are typically very limited for

robots, especially for mobile ones. It is therefore necessary to adopt a minimal approach to planning, depending on the workload of the robot and the urgency of a task. Third, a planning module for a robot is not an independent component, but it must be grounded in the available functions of the robot. A classic planner usually does not care about the actual execution of primitives in the plan, the physical state of the robot (e.g., the level of the battery), and often relies on a high level of abstraction. Last but not least, a planner for a robot has to deal with concurrency. There are always many more than just one or two goals or jobs to do to keep the robot operational.

One alternative to classical planning is known as reactive planning. It starts with a crude default plan which typically comes from a library of template plans or is generated by some fast off-line method. While the default plan is executed, there is continuous revision going on, trying to improve the current plan depending on sensor input and available computational resources. The plan can for example be modified on the basis of a set of plan transformations from a library leading to a higher utility, or simply due to changed environmental conditions.

Another alternative to classical planning is a behavior-based architecture which relies as much as possible on emergent behaviors, motivational systems and strong reactivity to changes in the environment. Behavior-based architectures emphasize dynamics. Sensory states are directly coupled to actuator states and modulated by motivational states. This architecture is strongly influenced by animal decision-making models from ethology.

Another way to classify robot architectures focuses on the flow of data from sensors to motors. Planning, world-modeling, and sensor/motor-control can be seen as three levels, ordered by a descending amount of abstraction and decreasing time-horizons at each level. A vertical flow of data over these levels then corresponds to the 'sense–model–plan–act' cycle underlying early robots like Shakey. A horizontal flow of data from the sensors to the motors allows short, fast paths or reactiveness as promoted in behavior-based architectures. Typically, a successful system will incorporate both types of data streams. This is known as a hybrid architecture.

INTERACTIONS AMONG ROBOTS

Research in robotics is not confined to single robots. Cooperation among robots and collective robotics is a recent and very active research field. Work in this area has focused on one hand on handling the additional complexity due to the higher number of robots. An example for such a coordination problem is joined path planning where the motion of several robots together has to be scheduled such that they do not conflict with each other. Another group of researchers is attempting to exploit the benefits of multiple robots, to use the added value of cooperation. Examples include cooperative sensing, where different robots collectively construct an interpretation of the environment, or cooperative problem-solving like box-pushing tasks.

Cooperation has been explored using two approaches. The first one is self-organization. Self-organization achieves global coherence based on only local behaviors and local interactions between the elements of a system. An example is path formation in ant societies or swarming based on local attraction and repulsion. Coordination of multiple robots can also be achieved through pre-programmed explicit group structures based on assigned roles and hierarchical commands. For many tasks that have to be performed as a team, like playing soccer, coordination can be based on a separation of space, like midfield, defence area, and so on, and on roles that are assigned to team members depending on their capabilities and the requirements of the role in each spatial area.

CONCLUSION

Robotics is a very difficult field because it requires the integration of many different technologies and implementation techniques. The value of robotics for cognitive science is that it allows the precise testing of specific theories about sensorimotor intelligence and embodied action in the world.

Further Reading

Steels L (ed.) (1995) *The Biology and Technology of Intelligent Autonomous Agents*. Berlin, Germany: Springer-Verlag.

Samuel's Checkers Player

Introductory article

Jonathan Schaeffer, University of Alberta, Edmonton, Alberta, Canada

CONTENTS

Arthur Samuel's checkers-playing program of the 1960s was a milestone in the history of computer science and artificial intelligence.

INTRODUCTION

Arthur Samuel was one of the pioneers in artificial intelligence research. His seminal articles on machine learning are classics, and they laid the foundation for the field of reinforcement learning. His ideas were demonstrated in his checkers-playing program, arguably the first successful artificial intelligence program ever built.

Samuel's interest in checkers was piqued in 1948 after hearing about Claude Shannon's work with chess. Samuel decided to build a program to play checkers, a game that required less infrastructure to program than did chess, as a means of creating a visible project that could attract research dollars. It is difficult today to imagine the conditions under which Samuel began his 30-year quest to build a strong checkers program. In his (unpublished) autobiography he wrote:

> I started writing a program for a machine that did not yet exist using a set of computer instructions that I dreamed up as they were needed. … My first checkers program for the University of Illinois' Illiac was never actually run because the Illiac was still only a paper design when I left that University for IBM in 1949. It was not until 1952 that I had my program running on the experimental model of the IBM 701. Incidentally, this first program was written directly in machine code – before we even had a symbolic assembler.

However, the credit for building the first working checkers program went to Christopher Strachey who demonstrated it in 1952.

Samuel, although not a strong checkers player, found the prospect of harnessing the new technology to create an 'intelligent' checkers-playing machine irresistible. After moving to IBM and finally getting access to the computing resources he needed, he became determined to succeed with his program:

> IBM in those days did not take kindly to one of their engineers wasting company time playing checkers, even if it was against a machine, and so most of my checkers work had to be done in my own time. I dressed my efforts up with a certain amount of respectability by adding a learning feature to the program but even then it was the use of the program as a computer test vehicle which could be adjusted to run continuously that got me the machine time that I needed to try out my experimental learning routines.

Samuel gave a successful demonstration of his program on television in 1956. Thomas Watson, president of IBM, arranged for the program to be exhibited to IBM shareholders. He predicted that it would result in a 15-point rise in the price of IBM stock. It did.

Samuel's program is best remembered for a famous match against Robert Nealey, the US blind champion, in 1962. The program won one of the games in the match, thereby creating a milestone in the history of computer science. Samuel's early success, and in particular the Nealey game, had the unexpected side-effect of checkers being labeled a 'solved' game. From that point on, research into building high-performance game-playing programs switched almost exclusively to chess.

Although it was Samuel's early success with his checkers program that ensured his place in computing history, his contributions to machine learning have stood the test of time. His ideas of rote learning and reinforcement learning are the basis for many techniques that are in common use today.

ROTE LEARNING

Samuel followed Shannon's recipe for building a chess program. First, the program used an alpha-beta search algorithm to analyze possible move sequences. For example, the program might consider all possible legal sequences for both sides five moves into the future as part of its decision-making process. Then, at nodes where the search stops (so-called *leaf* nodes), knowledge would be used

to assess how good or bad the checkers position was, by means of an *evaluation function*. The evaluation scores would be minimized and maximized back to the root by the alpha-beta algorithm. The move in the current position receiving the highest alpha-beta score would be selected to be played in the game.

Samuel wanted his program to learn from experience. He accomplished this using rote learning.

For each move in a game, the program would do an alpha-beta search to decide on its best move. Suppose that this search considered all possibilities five moves into the future. After completing the search, the program would then record the position and the result of the search (the move chosen and an assessment of how good it was) in a database.

During subsequent games, the database of past search results would be queried. Suppose the program was searching five moves ahead, and three moves into the search it encountered a position that was in the database. Instead of continuing the search, the previous result for that position was retrieved from the database and used instead. This approach has two advantages. Firstly, the search is faster since some of the search effort has been eliminated and replaced by the database information. Secondly, the result of the search is more accurate because the database result represents more accurate information than would have been obtained by continuing the search. In this example, since three moves had been played, an additional two moves ahead would have been searched to arrive at a result. Instead, the database result provides a five-moves-ahead assessment.

The database of game moves would grow with every game played, creating a valuable repository of knowledge. However, the computers Samuel used had limited memory (measured in kilobytes, not megabytes) and he had to restrict the amount of data kept. He introduced the idea of 'forgetting', allowing the program to delete (or forget) game moves that had not been useful in a long time.

Samuel's ideas about rote learning are still applied today in most major game-playing programs. However, the availability of inexpensive memory has changed the scope. Instead of only the final result from a search being saved, whole subtrees of a search are now cached in a table (often called the transposition table) which is accessed dynamically throughout the search. The transposition table allows rote learning within a search (creating many more opportunities for increased search depth), the elimination of move cycles, the detection of transposed move sequences, and improved move ordering. Rote learning between games, as used by Samuel, is used in many commercial game products, allowing the program to avoid repeating mistakes.

EVALUATION FUNCTIONS

The evaluation function used in Samuel's program was based on consultations with strong checkers players and the extensive literature on the game. Samuel identified over 40 pieces of knowledge that might be useful for assessing the strength of a position. The problem then becomes how to combine these pieces of knowledge (or *features*) into a numeric assessment.

Initially, Samuel combined the features linearly:

$$\text{evaluation} = \sum_{i=1}^{n} w_i \times f_i \tag{1}$$

where n is the number of features, f_i is a feature, and w_i is the *weight* (or importance) of the feature. For example, in checkers two features that are correlated with success are the piece count (the relative number of pieces on the board for each side) and centre control (which side controls the central squares on the board). Piece count (f_1) is usually much more important than centre control (f_2), and hence has a much higher weighting ($w_1 >> w_2$).

Samuel wanted his program to learn the weights automatically. This would allow him to add knowledge to the program without having to evaluate how good or bad the knowledge was. During the automatic tuning, if a piece of knowledge was bad or inconsequential, the weight would drift towards a value of 0, effectively becoming irrelevant (and possibly being removed completely from the program).

A search is intended to produce a result that is a (good) predictor of the final result of the game. Presumably, a search conducted on move $i + 1$ of a game will give a better approximation of the final game result than a search conducted on move i. The weights of the evaluation function can be modified (some increased and others decreased) so that the move-i search result can be closer to the move-$(i + 1)$ search result. This technique is commonly known as *hill climbing*, where the weights are incrementally modified to try and give the best results over the (large) set of test data. Thus what happens in a game can be used to adjust the program to be a better predictor.

The weight adjustments were automated. This allowed Samuel's program to play against itself and improve its weights without human intervention. In effect, the program could learn to play better checkers by itself.

Samuel later modified the evaluation function to include what he called *signature tables*. Subsets of features were grouped together and separately weighted. The results of these subgroups could be hierarchically combined to produce a final evaluation score. The hierarchical evaluation function allowed for nonlinear relationships between terms, giving greater flexibility in what could be achieved in the evaluation function. Samuel reported a significant improvement in his program's performance with signature tables compared with the simple linear evaluation function.

Samuel's work with a learning system that automatically adjusted the weights of features in an evaluation function was a precursor of *temporal difference learning* methods. However, this was only recognized as a valuable learning algorithm after Sutton extended and formalized this work. One remarkable success of this type of learning was Gerry Tesauro's backgammon program, which played over a million games to learn its evaluation function weights. The result is a program that is as good as (and possibly better than) the strongest human players.

AFTERMATH

In 1966 Samuel's program played exhibition matches with the world champion Walter Hellman and his challenger Derek Oldbury. The program lost every game. Samuel sought advice from strong checkers players on how to strengthen his program. Herschel Smith met Samuel in 1967 and advised him to add databases of strong opening moves and common endgame sequences. Smith writes of that encounter:

> [Samuel's] response was: 'I cannot do that. That reduces the game of checkers to simple table look-up.' He believed the machine-learning concept was essential to his program and would not compromise. He put it to me this way: 'I did not teach the computer how to play checkers; I taught it how to learn to play checkers.'

Eventually, Samuel acquiesced and added databases of opening moves to his program, but it never approached master level. The Nealey result had been a fluke, but the damage was done: the notion that checkers is 'solved' persists to this day.

Samuel continued to tinker with his program into the 1970s, but it was eclipsed by the Duke program. Samuel's dream of a checkers program that could take on the best human players was not realized until 1990, when the program Chinook earned the right to play a match for the world championship by coming second to world champion Marion Tinsley in the US Championship. Ironically, Arthur Samuel died a few days before the Chinook success. In 1994, Chinook became the first game-playing program to win a human world championship by winning the World Man-Machine Checkers Championship.

Samuel's pioneering work in machine learning remains a milestone in the artificial intelligence literature. His two seminal papers are still frequently cited in the literature, four decades after the research was originally done. In the fast-moving field of computing science, there are few papers that remain relevant a few years after they appear let alone a few decades.

Eric Weiss, in his 1992 article, eloquently put Samuel's accomplishments into perspective:

> Clearly the accomplishment for which he is most famous is his checkers program. He lavished the most effort on it over the longest period and it is…recognized as the world's first self-learning computer program. I will go further and claim that it is the first functioning artificial intelligence program. Thus in spite of Samuel's own opinion…that he considered his [engineering] patents to be more important, the world will remember him for his Great Game.

Further Reading

Samuel A (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 210–229.

Samuel A (1967) Some studies in machine learning using the game of checkers: recent progress. *IBM Journal of Research and Development* **11**: 601–617.

Samuel A (1980) *Personal Computing*, March [Letter appearing in the *Computer Games* section.]

Samuel A (1983) AI, where it has been and where it is going. *International Joint Conference on Artificial Intelligence* **3**: 1152–1157.

Schaeffer J (1997) *One Jump Ahead: Challenging Human Supremacy in Checkers*. Springer Verlag.

Scherzer T, Scherzer L and Tjaden D (1990) Learning in Bebe. In: Marsland T and Schaeffer J (eds) *Computers, Chess, and Cognition*, pp. 197–216. Springer-Verlag.

Shannon CE (1950) Programming a computer for playing chess. *Philosophical Magazine* **41**: 256–275.

Strachey C (1952) Logical or non-mathematical programmes. *Proceedings of the Association for Computing Machinery Meeting*, pp. 46–49.

Sutton R (1988) Learning to predict by the methods of temporal differences. *Machine Learning* **3**: 9–44.

Sutton R and Barto A (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tesauro G (1995) Temporal difference learning and TD-Gammon. *Communications of the ACM* **38**(3): 58–68.

Truscott T (1979–1980) The Duke checkers program. *Journal of Recreational Mathematics* **12**(4): 241–247.

Weiss E (1992) Eloge: Arthur Lee Samuel (1901–1990). *IEEE Annals of the History of Computing* **14**(3): 55–69.

Search, Adversarial

Intermediate article

Dana S Nau, University of Maryland, Maryland, USA

CONTENTS

Adversarial search, or game-tree search, is a technique for analyzing an adversarial game in order to try to determine who can win the game and what moves the players should make in order to win.

INTRODUCTION

Adversarial search is one of the oldest topics in artificial intelligence (AI). The original ideas for adversarial search were developed by Shannon (1950) and independently by Turing in 1951, in the context of the game of chess – and their ideas still form the basis for the techniques used today. Computer programs based on adversarial search techniques are now as good as, or better than, humans in several popular board games and card games (see Table 1).

GAME TREES

Adversarial search is based on the notion of a *game tree*, a mathematical structure that represents the positions that might result from every possible series of moves in a game. Game trees can be used to model any game that satisfies the following restrictions:

- The game is played by two players, who make moves sequentially rather than simultaneously. This excludes most popular video games, and also the kinds of games studied in the branch of economics known as classical game theory.
- The game is *finite*. At each turn a player can choose from only a finite number of possible moves, and the game is guaranteed to end within some finite number of moves.
- The game is *zero-sum*; i.e. the amount that one player wins is precisely equal to the amount that the other player loses.
- The game contains no elements of chance (although we will later look at how to incorporate certain kinds of chance elements into game-tree searching).

Figure 1 shows a simple example of a game tree for a game between two players called Max and Min. Square nodes represent positions where it is Max's move, and round nodes represent positions where it is Min's move. The leaf nodes represent positions where the game has ended, and the numbers below them represent the payoffs for Max (recall that Min's payoff is always the negative of Max's payoff).

THE MINIMAX ALGORITHM

At the node n_4 of Figure 1, Max's best move is to go to n_8 to get the payoff of 5. At the node n_5, Max's best move is to go to n_{10} to get the payoff of 8. If we know that Max will always make the best move, then this means that at the node n_2, Min's best move is to go to n_4 so that Max's payoff will be 5 rather than 8. Generalizing this analysis, it follows that if both players play perfectly from some node n onward, then the payoff for Max is the *minimax value* of n, which is the value computed by the following algorithm:

```
procedure minimax(n)
  if n is a leaf node then return the
    payoff for Max at n;
  compute the children n_1,...,n_k
    of n;
  for i=1,...,k,
    let m_i=minimax(n_i);
  if it is Max's turn to move at n then
    return max(m_1, m_2,...,m_k);
  else return min(m_1, m_2,...,m_k)
```

The algorithm's name comes from the famous *minimax theorem* of von Neumann and Morgenstern. However, the principle embodied in the algorithm – that a player should maximize the minimum payoff for every alternative – is what decision theorists call the *maximin decision criterion*.

As an example, Figure 2 gives the minimax values for the tree in Figure 1. These values can be

Table 1. Popular board games that computer programs can play as well as, or better than, humans

Game	*Best programs*	*Ability*	*Primary techniques*
Checkers	Chinook (Schaeffer, 1997)	Better than best humans	Game-tree search
Othello	Hannibal (Geoffroy, 2000) and Logistello (Buro, 2000)	Better than best humans	Neural networks, game-tree search
Scrabble	Maven and CrossWise (Alexander, 2001)	Comparable to best humans	Simulation
Chess	Deep Blue (IBM Corp. 1997)	Comparable to best humans	Game-tree search
Backgammon	TD-gammon (Tesauro, 1995)	Comparable to best humans	Neural networks

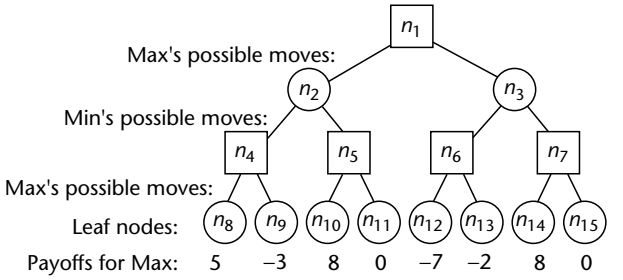


Figure 1. A game tree for a very simple game between two players, Max and Min. This game tree has two properties that most game trees do not have: at each position before the end of the game, each player always has exactly two moves; and the game always ends after exactly three moves.

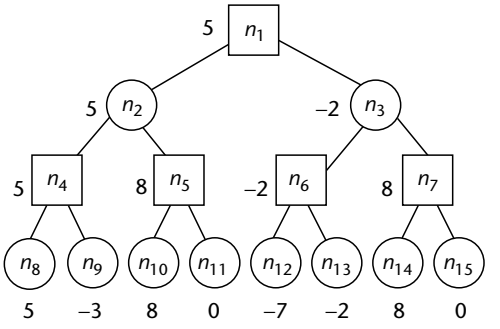


Figure 2. Minimax values for the nodes of the game tree of Figure 1.

used to decide what move to make at any node of the tree: Max should always move to the child that has the largest minimax value, and Min should always move to the child that has the smallest minimax value. Theoretically, this rule is only correct against infallible opponents (Pearl, 1984), but in practice it works well enough against any good opponent.

ALPHA–BETA PRUNING

Suppose that we are computing the minimax value for the node n_1 of Figure 1, and that we have

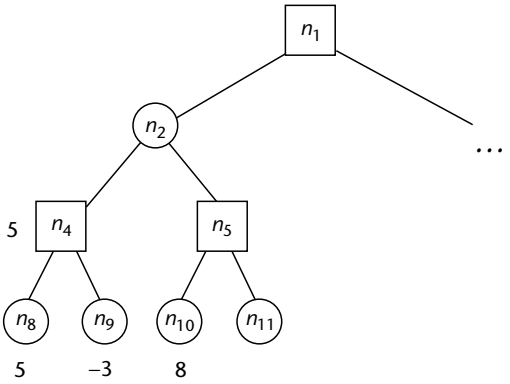


Figure 3. An example of alpha–beta pruning. In this example, it is not necessary to compute minimax(n_{11}), because it cannot possibly affect minimax(n_2) and minimax(n_1).

already computed $\texttt{minimax}(n_4) = 10$ and $\texttt{minimax}(n_{10}) = 11$ as shown in Figure 3. Then

$$\texttt{minimax}(n_2) = \texttt{min}(5, \texttt{minimax}(n_5)) \leq 5;$$
$$\texttt{minimax}(n_5) = \texttt{max}(8, \texttt{minimax}(n_{11})) \geq 8.$$

It follows that we do not need to know what minimax(n_{11}) is, because there is no way for minimax (n_{11}) to affect minimax(n_2) and minimax(n_1). Generalizing this analysis gives us the following algorithm, which is called the *alpha–beta pruning algorithm*:

```
procedure alphabeta (n, α, β)
  if n is a leaf node then
    return the payoff for Max at n
  else if it is Max's move at n then
    let n1, ..., nk be the children of n
    for i = 1, ..., k do
      α = max (α, alphabeta(ni, α, β))
      if α: ≥ β then return α
    end
    return α
  else /* it must be Min's move */
    let n1, ..., nk be the children of n
    for i = 1, ..., k do
      β = min(β, alphabeta(ni, α, β))
```

```
      if β≤α then return β
    end
    return β
  end if
end alphabeta
```

Alpha–beta pruning is guaranteed to compute a node's minimax value, and in general it will visit far fewer nodes than the minimax algorithm. There also are several other algorithms having this same property (Pearl, 1984), but because of its simplicity and low overhead, alpha–beta pruning is the one that is most widely used. In the worst case, alpha–beta pruning will visit every node that the minimax algorithm visits, but in the best case, it can search a tree of roughly twice the depth as the minimax algorithm in roughly the same amount of time (Knuth and Moore, 1975).

LIMITING THE DEPTH OF THE SEARCH

Most game trees contain far too many nodes to allow the entire tree to be searched quickly, even with a tree-pruning algorithm such as alpha–beta. Another way to visit fewer nodes is to stop searching at some arbitrary search depth d, and instead to estimate the minimax values of the nodes at this depth using a *static evaluation function*. This gives the following modification to the alpha–beta algorithm:

```
procedure alphabeta (n, d, α,β)
  if n is a leaf node or d=0 then
    return e(n)
  else if it is Max's move at n then
    let n1,...,nk be the children of n
    for i=1,...,k do
      α=max(α, alphabeta(ni,d–1,
        α,β))
      if α≥β then return α
    end
    return α
  else /* it must be Min's move */
    let n1,...,nk be the children of n
    for i=1,...,k do
      β=min(β, alphabeta(ni,d–1,
        α,β))
      if β≤α then return β
    end
    return β
  end if
end alphabeta
```

A search to depth d is likely to yield inaccurate results if something happens (such as an exchange of material in the game of chess) that makes a large change in the minimax values just beyond the cutoff depth d. In order to avoid this *horizon effect*, it is best to modify the above algorithm so that instead of stopping when $d = 0$, it keeps searching until it reaches a node that is *quiescent* (in chess, a node where there are no pending exchanges of material). However, this means that at different branches of the search tree, the evaluation function may be applied to nodes of different depths. Since a move usually strengthens the position of the player who makes the move, nodes at which Max has just moved are likely to get higher evaluations than nodes where Min has just moved. This can cause the computer program erroneously to prefer branches that end at an odd depth. To overcome this, practical implementations of game-tree search will usually include a *biasing factor* that is added to node evaluations at even depths and subtracted from node evaluations at odd depths.

SPEEDING UP THE SEARCH

Although there exist games and game trees in which searching deeper will produce worse decision-making (Nau, 1983), a deeper search will produce significantly better decisions in nearly all practical situations (Kaindl, 1988). However, in order to do a deeper search, a search algorithm must visit exponentially many more nodes, which makes it impossible to search very deeply unless the algorithm is extremely fast. Here are several techniques for speeding up the search.

Databases of standard openings. Sometimes it is possible to develop databases of various standard ways to start a game. As long as the computer's opponent uses one of these standard openings, the computer can make moves very quickly by retrieving them from the database.

Thinking on the opponent's time. In order to decide what move to make, a search algorithm will need to predict what node the opponent is likely to move to. After making its move, a computer program can immediately begin a game-tree search starting at this node, thereby saving some time if the opponent actually makes the predicted response.

Iterative deepening. One problem with the procedure described in the previous section's problem is deciding what value to use for the cutoff depth d. If we make d too large, then the search will take too long – but if we make d too small, then it will reduce the accuracy of the decision-making. One way to solve this problem is to choose a value for d that we know is not too large. If we finish our search more quickly than expected, then we can increment d and search again. Since the number of nodes visited usually grows exponentially with the search depth, this usually will not take significantly

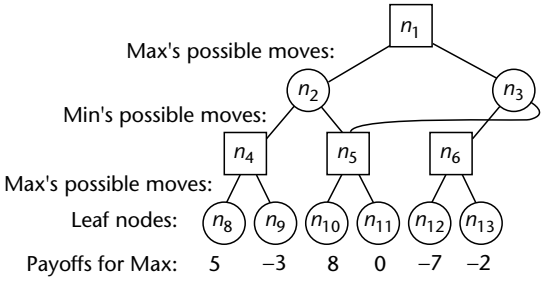


Figure 4. A game graph having the same minimax values as the game tree of Figure 1.

more time than if we had used the correct value of d the first time around.

Transposition tables. The term 'game tree' is a misnomer: most games actually correspond to graphs such as the one in Figure 4. We can take advantage of this by using a *transposition table*, a cache of recently visited positions and their minimax values. If we visit a node that we visited a short time ago, we can retrieve its value from the transposition table rather than computing it again.

Move ordering. For alpha–beta pruning, the difference between the best case and the worst case depends on the order in which the moves are considered. At a node where a cutoff is to occur, the best case is if the cutoff occurs at the first child that is visited – which occurs if the children appear in order of decreasing minimax values when it is Max's move, or in order of increasing minimax values when it is Min's move. A way to make this occurrence more likely is to sort a node's children in decreasing order of their evaluation-function values when it is Max's move, and in increasing order of their evaluation-function values when it is Min's move.

Specialized hardware. An expensive but effective technique, used in the Deep Blue program, is to design special-purpose computer hardware to speed up its game-tree search.

OTHER KINDS OF GAMES

Not all games are amenable to the 'brute-force' game-tree search techniques described above. Here are several examples:

- *Backgammon* incorporates a chance element, namely the rolls of the dice. The game-tree model can be generalized to model the dice rolls, but it produces too large a game tree to be searched in a reasonable amount of time. However, through the use of machine-learning techniques, static evaluation functions can be created that are good enough to work well with only a shallow game-tree search. Thus, the best computer backgammon programs play as well as the best human beings (Tesauro, 1995).
- *Bridge* incorporates both chance and imperfect information: the cards are dealt randomly, and no player knows all of the other players' cards. As in backgammon, the game-tree model can be extended to model this, but it produces too large a game tree to search in a reasonable time. Research in bridge has produced a variety of techniques for reducing the size of this game tree, ranging from Monte Carlo simulation to AI planning. Bridge programs are not yet as good as the best humans, but may get there in the next few years.
- *Poker* includes not only chance and imperfect information, but also presents the difficulty of recognizing what kind of betting strategy the opponent is using. Current programs play much worse than good human players.
- *Go*, unlike the above games, fits the game-tree model perfectly. However, its game tree is huge because there are hundreds of possible moves at each node. This makes it impossible to search the game tree to any significant depth. Currently, the best available computer programs are ranked below the level of an average amateur.
- Most current *video games* are action-oriented games that emphasize hand-to-eye coordination. However, a few of them (warcraft, command and conquer, civilization, etc.) involve longer-term strategic reasoning. These games incorporate several features not discussed above. For example, one may need to plan actions that have various time durations, which may overlap those of the opponents' moves. Computer-science game-playing research is only just beginning to address such problems, and thus the strategic reasoning capabilities of such computer programs still remain very poor.

CONCLUSION

The biggest successes for adversarial search techniques have been in board games such as chess, checkers, othello, and backgammon. The best computer programs for these games are now as good or better than the best human players. However, the techniques used in most of these programs work rather differently from human thought processes: the programs work by examining as many board positions as possible within the time constraints, often examining many thousands of positions in order to decide which move to make – and in contrast, the best humans examine at most a few dozen positions in order to decide upon their moves. How humans are able to do this is still ill-understood.

References

Alexander S (2001) Scrabble FAQ. [http://www.teleport.com/~stevena/scrabble/faqtext.html.]

Buro M (2000) LOGISTELLO's homepage. [http://www.neci.nj.nec.com/homepages/mic/log.html.]

Geoffroy L (2000) Hannibal homepage. [http://www.cam.org/~bigjeff/Hannibal.html.]

IBM Corporation (1997) *Kasparov Versus Deep Blue: The Rematch*. [http://www.research.ibm.com/deepblue/home/html/b.html.]

Kaindl H (1988) Minimaxing: theory and practice. *AI Magazine*, Fall, pp. 69–76.

Knuth D and Moore R (1975) An Analysis of alpha–beta pruning. *Artificial Intelligence* **6**: 293–326.

Nau D (1983) Pathology on game trees revisited, and an alternative to minimaxing. *Artificial Intelligence* **21**(1,2): 221–244.

Pearl J (1984) *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Reading, MA: Addison Wesley.

Schaeffer J (1997) *One Jump Ahead: Challenging Human Supremacy in Checkers*. New York, NY: Springer-Verlag. [Further information available at http://www.cs.ualberta.ca/~chinook/.]

Shannon C (1950) Programming a computer for playing chess. *Philosophical Magazine* (Series 7) **41**: 256–275.

Tesauro G (1995) Temporal difference learning and TD-gammon. *Communications of the ACM* **38**(3): 58–68. [Also available at http://www.research.ibm.com/massive/tdl.html.]

Further Reading

AAAI (2000–2002) *Games and Puzzles*. http://www.aaai.org/AITopics/html/games.html

Hirsh H (ed.) (1999) Playing with AI. *IEEE Intelligent Systems* November/December pp. 8–18. http://www.computer.org/intelligent/ex1999/pdf/x6008.pdf

Nau D (1999) AI game-playing techniques: are they useful for anything other than games? *AI Magazine* **20**(1): 117–118.

Newborn M and Newborn M (1996) *Kasparov versus Deep Blue: Computer Chess Comes of Age*. New York, NY: Springer Verlag.

Russell S and Norvig P (1995) *Artificial Intelligence, a Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall.

Schaeffer J (2001) A gamut of games. *AI Magazine* **22**(3): 29–46.

Search

Introductory article

Richard E Korf, University of California, Los Angeles, California, USA

CONTENTS

A search is a trial-and-error exploration of alternative solutions to a problem.

INTRODUCTION

Search is a universal problem-solving mechanism in artificial intelligence (AI). In AI problems, the sequence of steps required for solution of a problem are not known *a priori*, but often are determined by a systematic trial-and-error exploration of alternatives. The kinds of problems addressed in this article include single-agent problems and two-player games.

Sliding-tile puzzles, including the 3×3 'eight puzzle' (see Figure 1) and its larger relatives the 4×4 'fifteen puzzle' and 5×5 'twenty-four puzzle', are often used in the AI literature as examples of single-agent problems. The eight puzzle consists of a 3×3 square frame containing eight numbered square tiles, and an empty position called the blank. The legal operators slide any tile that is horizontally or vertically adjacent to the blank into the blank position. The problem is to rearrange the tiles from some random initial configuration into a particular goal configuration.

The sliding-tile puzzles are common test-beds for research in search algorithms, because they are very simple to represent and manipulate, yet finding optimal solutions to larger-sized problems is computationally very expensive. Other well-known examples of single-agent problems include Rubik's cube, theorem proving, the traveling salesman problem, vehicle navigation, and the wiring of VLSI ('very large-scale integrated') circuits. In each case, the task is to find a sequence of operations that map an initial state to a goal state.

The second major class of search problems includes two-player games such as chess, checkers, Othello, go, and backgammon.

First to be described here is the problem-space model on which search algorithms are based. The article then considers brute-force searches, including breadth-first, uniform-cost, depth-first, depth-first iterative-deepening, and bidirectional search. Next, it examines various heuristic searches, including pure heuristic search, the A* algorithm, iterative-deepening A*, and depth-first branch-and-bound. Finally, it considers two-player game searches, including minimax and alpha–beta pruning. The efficiency of these algorithms, in terms of the quality of the solutions they generate, the amount of time they take to execute, and the amount of computer memory they require, are of central concern throughout. Since search is a universal problem-solving method, what limits its applicability is the efficiency with which it can be performed.

THE PROBLEM-SPACE MODEL

A 'problem space' is the environment in which a search takes place. A problem space consists of a set of possible states of the problem, and a set of operators that change the state. For example, in the eight puzzle, the states are the possible permutations of the tiles, and the operators slide a tile into the blank position. A 'problem instance' is a problem space together with an initial state and a goal condition. In the case of the eight puzzle, the initial state would be whatever permutation the puzzle starts in, and the goal condition would be the desired permutation. The problem-solving task is to find a sequence of operators that map the initial state to the goal state.

A 'problem-space graph' is often used to represent a problem space. The states of the space are represented by nodes of the graph, and the operators by edges between nodes. Edges may be undirected or directed, depending on whether their corresponding operators are reversible or not. The task in a single-agent problem is to find a path in the graph from the initial node to a goal node. Figure 1 shows a small part of the problem-space graph for the eight puzzle.

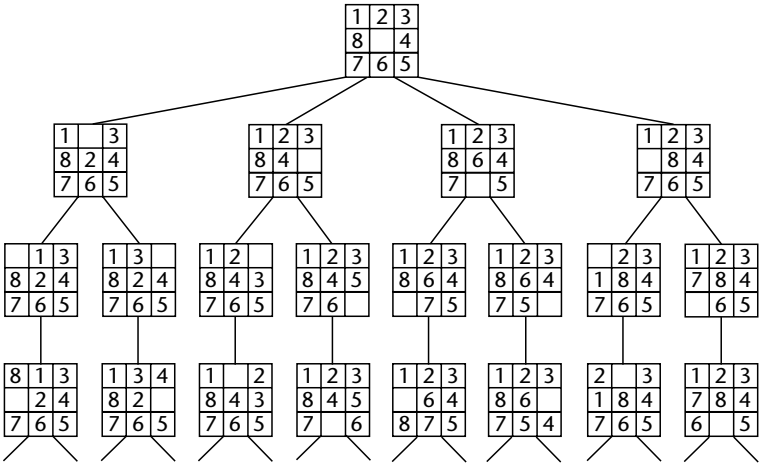


Figure 1. A small fragment of the problem-space graph for the eight puzzle, presented as a search tree.

Although most problem spaces correspond to graphs with more than one path between any pair of nodes, for simplicity they are often represented as trees, where the initial state is the root of the tree. The cost of this simplification is that any state that can be reached by two different paths will be represented by duplicate nodes, increasing the size of the tree. The benefit of a tree is that the absence of cycles simplifies many search algorithms. Here we will restrict our attention to trees, but there exist more general graph versions of most of the algorithms described.

One feature that distinguishes AI search problems from many other graph-searching problems is the size of the graphs involved. For example, the entire chess graph is estimated to contain over 10^{40} nodes. Even a simple problem like the twenty-four puzzle contains almost 10^{25} nodes. Therefore, the problem-space graphs of AI problems are never represented by listing each state, but rather by specifying an initial state and a set of operators to generate new states from existing states. Furthermore, the size of an AI problem is rarely expressed as the number of nodes in its problem-space graph. Rather, the two parameters of a search tree that are used to determine the efficiency of search algorithms are its 'branching factor' and its 'solution depth'. The branching factor is the average number of children of a node. For example, the average branching factor of the eight puzzle search tree is $\sqrt{3}$, or about 1.732, assuming that the parent of a node is not included as one of its children. The solution depth of a problem instance is the length of a shortest path from the initial state to a goal state, or the length of a shortest sequence of operators that solves the problem. Thus, if the goal were in the bottom row of Figure 1, the depth of the problem instance represented by the initial state at the root would be three.

BRUTE-FORCE SEARCH

The most general search algorithms are 'brute-force' searches, which do not require any domain-specific knowledge. All that is required for a brute-force search is a specification of the state space, the initial state, and the goal condition. The most important brute-force techniques are breadth-first, uniform-cost, depth-first, depth-first iterative-deepening, and bidirectional search.

In the descriptions below, to 'generate' a node means to create the data structure corresponding to that node, and to 'expand' a node means to generate all the children of that node.

Breadth-First Search

Breadth-first search expands nodes in order of their distance from the root, generating one level of the tree at a time until a solution is found (see Figure 2). If we take a family tree as an analogy, a breadth-first search would expand all the individuals in one generation before expanding the individuals of the next generation. It is most easily implemented by maintaining a queue of nodes, initially containing just the root, and repeatedly removing the node at the head of the queue, expanding it, and adding its children to the tail of the queue.

Since it never generates a node in the tree until all the nodes at shallower levels have been generated, breadth-first search always finds a shortest path to a goal. Assuming each node can be generated in

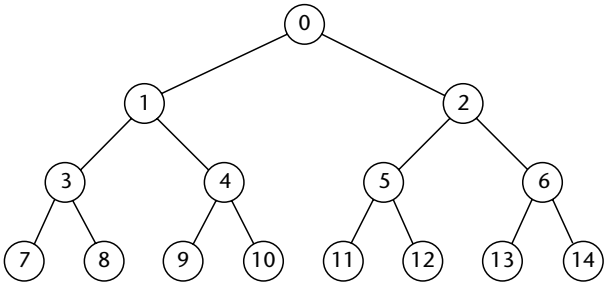


Figure 2. Order of node generation for breadth-first search.

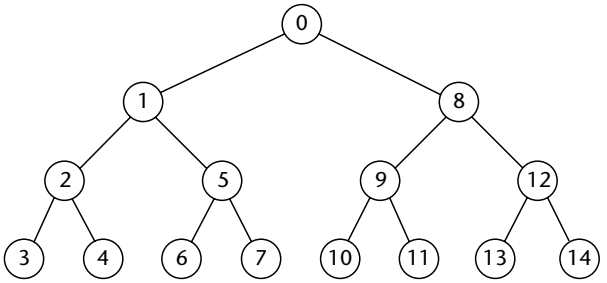


Figure 3. Order of node generation for depth-first search.

constant time, the amount of time used by breadth-first search is proportional to the number of nodes generated, which is a function of the branching factor b and the solution depth d. Since the number of nodes at level d is b^d, the total number of nodes generated in the worst case is $b + b^2 + b^3 + \ldots + b^d$, and the asymptotic time complexity of breadth-first search is $O(b^d)$.

The main disadvantage of breadth-first search is its memory requirement. Since each level of the tree must be saved in order to generate the next level, and assuming the amount of memory is proportional to the number of nodes stored, the space complexity of breadth-first search is also $O(b^d)$. As a result, breadth-first search is severely space-limited in practice, and will exhaust the memory available on typical computers in a matter of minutes.

Uniform-Cost Search

In some cases, a cost can be associated with each edge of a problem-space graph. For example, if the graph represents a network of roads, then the cost of an edge may be the distance of the corresponding road segment. In such cases, uniform-cost search, also known as Dijkstra's single-source shortest-path algorithm, is often used. Instead of expanding nodes in order of their depth in the tree, uniform-cost search expands nodes in order of their total cost, where the total cost of a node is the sum of the edge costs from the root to the node. At each step, uniform-cost search expands next a node whose total cost is lowest, among all nodes generated but not yet expanded.

Whenever a node is chosen for expansion by uniform-cost search, a lowest-cost path to that node has been found. The worst-case time complexity of uniform-cost search is $O(b^{c/m})$, where c is the cost of an optimal solution and m is the minimum edge cost. Unfortunately, uniform-cost search suffers the same memory limitation as breadth-first search. Indeed, if all edges have the same cost, then uniform-cost search reduces to breadth-first search.

Depth-First Search

Depth-first search overcomes the space limitation of breadth-first search by always generating next a child of the deepest unexpanded node (see Figure 3). Returning to our family tree analogy, depth-first search expands a complete line of descendents, from parent to child to grandchild and so on, before backing up and generating the next line. After all the descendents of one child are generated, then the next child is generated.

Both breadth-first and depth-first search can be implemented using a list of unexpanded nodes; breadth-first search manages the list as a 'first-in, first-out' queue, whereas depth-first search manages the list as a 'last-in, first-out' stack. More commonly, depth-first search is implemented recursively, with the recursion stack taking the place of an explicit node stack.

The advantage of depth-first search is that its space requirement is only linear in the maximum search depth, rather than exponential as in breadth-first search. The reason is that the algorithm only needs to store a stack of nodes on the path from the root to the current node. The time complexity of a depth-first search to depth d is $O(b^d)$, as in breadth-first search, since it generates the same set of nodes albeit in a different order. Thus, in practice, depth-first search is time-limited rather than space-limited.

The disadvantage of depth-first search is that it may not terminate on an infinite tree, but simply go down one path forever. Even a finite graph can generate an infinite tree. The usual solution to this problem is to impose a cut-off depth on the search. Although the ideal cut-off is the solution depth d, this value is rarely known in advance of actually solving the problem. If the chosen cut-off depth is less than d, the algorithm will fail

to find a solution, whereas if the cut-off depth is greater than d, a large price is paid in execution time, and the first solution found may not be an optimal one.

Depth-First Iterative Deepening

Depth-first iterative deepening (DFID) combines the best features of breadth-first and depth-first search. DFID first performs a depth-first search to depth one, then performs a depth-first search to depth two, and continues to perform depth-first searches to successively greater depths, until a solution is found (see Figure 4).

Since it never generates a node until all shallower nodes have been generated, the first solution found by DFID is guaranteed to be a shortest path. Furthermore, since at any given point it is executing a depth-first search, saving only a stack of nodes, and the algorithm terminates when it finds a solution at depth d, the space complexity of DFID is only $O(d)$.

Although it may appear that DFID wastes a great deal of time in the iterations prior to the one that finds a solution, this extra work is usually insignificant. To see this, note that the number of nodes at depth d is b^d, and each of these nodes is generated just once, during the final iteration. The number of nodes at depth $d-1$ is b^{d-1}, and each of these is generated twice, once during the final iteration and once during the penultimate iteration. In general, the number of nodes generated by DFID is $b^d + 2b^{d-1} + 3b^{d-2} + \ldots + db$. This is asymptotically $O(b^d)$ if b is greater than one, since for large values of d the lower-order terms become insignificant. In other words, most of the work goes into the final iteration, and the cost of the previous iterations is relatively small. The ratio of the number of nodes generated by DFID to the number generated by breadth-first search on a tree is approximately $b/(b-1)$. In fact, DFID is asymptotically optimal in terms of time and space among all brute-force shortest-path algorithms on a tree.

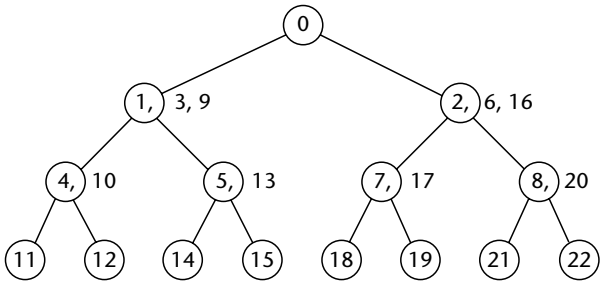


Figure 4. Order of node generation for depth-first iterative deepening.

If the edge costs vary, then one can run an iterative-deepening version of uniform-cost search, where the depth cut-off is replaced by a cut-off on the total cost of a node. At the end of each iteration, the threshold for the next iteration is set to the minimum total cost of all nodes generated but not expanded on the iteration just completed.

In a tree, there is only one path from the root to any given node. In more general graphs, however, such as a network of roads, there is often more than one path between a given pair of nodes. In that case, breadth-first search may be much more efficient than any depth-first search. The reason is that a breadth-first search can detect multiple paths to the same state, eliminating the duplicate paths, whereas a depth-first search cannot. Thus, the complexity of breadth-first search grows only as the number of nodes at a given depth, while the complexity of depth-first search grows as the number of paths of a given length. For example, in a square grid, the number of nodes within a radius r of the center is $O(r^2)$, whereas the number of paths of length r is $O(3^r)$, since there are three children of every node, not counting its parent. Thus, in a graph with a large number of such multiple paths, breadth-first search is preferable to depth-first search, if sufficient memory is available.

Bidirectional Search

Bidirectional search is a brute-force algorithm that requires an explicit goal state, not just a test for a goal condition. The main idea is to simultaneously search forward from the initial state and backward from the goal state until the two search frontiers meet. The path from the initial state is then concatenated with the inverse of the path from the goal state to form the complete solution path.

Bidirectional search guarantees optimal solutions. Assuming that the comparisons for identifying a common state between the two frontiers can be done in constant time per node, by hashing for example, the time complexity of bidirectional search is $O(b^{d/2})$ since each search need only proceed to half the solution depth. Since at least one of the searches must be breadth-first in order to find a common state, the space complexity of bidirectional search is also $O(b^{d/2})$. Thus, bidirectional search is space-limited in practice.

Combinatorial Explosion

The problem with all brute-force search algorithms is that their time complexities grow exponentially with problem size. This phenomenon is called

combinatorial explosion, and it severely limits the size of problems that can be solved with these techniques. For example, while the eight puzzle, with about 10^5 states, is easily solved by brute-force search, the fifteen puzzle contains over 10^{13} states, and cannot be solved with brute-force techniques. Development of faster machines will not have a significant impact on this problem: the 5×5 twenty-four puzzle contains almost 10^{25} states, for example.

HEURISTIC SEARCH

In order to solve larger problems, domain-specific knowledge must be added to improve search efficiency. The term 'heuristic search' has both a general meaning and a more specialized technical meaning. In a general sense, the term 'heuristic' is used for any advice that is often effective, but isn't guaranteed to work in every case. Within the heuristic search literature, however, the term usually refers to a 'heuristic evaluation function'.

Heuristic Evaluation Functions

In a single-agent problem, a heuristic evaluation function estimates the cost of an optimal path between a pair of states. For example, Euclidean or 'airline' distance is an estimate of the highway distance between a pair of locations. A common heuristic function for the sliding-tile puzzles is called Manhattan distance. It is computed by counting the number of moves along the grid that each tile is displaced from its goal position, and summing these values over all tiles. For a fixed goal state, a heuristic evaluation is a function that estimates the 'distance' from any given node to the given goal state.

The key properties of a heuristic evaluation function are that it estimate actual cost, and that it be inexpensive to compute. For example, the Euclidean distance between a pair of points can be computed in constant time. The Manhattan distance between a pair of states in a sliding-tile puzzle can be computed in time proportional to the number of tiles.

In addition, most common heuristic functions are lower bounds on actual cost. A heuristic function that is guaranteed to be less than or equal to the quantity it is estimating is said to be 'admissible'. For example, the airline distance is a lower bound on the road distance between two points, since the shortest path between a pair of points is a straight line. Similarly, Manhattan distance is a lower bound on the actual number of moves necessary to solve a sliding-tile puzzle instance, since every tile must move at least its Manhattan distance to its goal position, and each move only moves one tile.

A number of algorithms make use of heuristic functions. These include pure heuristic search, the A* algorithm, iterative-deepening A*, and depth-first branch-and-bound. Heuristic information can also be employed in bidirectional search.

Pure Heuristic Search

The simplest of these algorithms, pure heuristic search, expands nodes in order of their heuristic values. It maintains a 'closed list' of those nodes that have already been expanded, and an 'open list' of those nodes that have been generated but not yet expanded. The algorithm begins with just the initial state on the open list. At each cycle, a node on the open list with the minimum heuristic value is expanded, generating all of its children, and then placed on the closed list. The heuristic function is applied to the children, and they are placed on the open list in order of their heuristic values. The algorithm continues until a goal node is reached.

In a graph with cycles, multiple paths will be found to the same node, and the first path found may not be the shortest. When a shorter path to an open node is found, it is saved and the longer path discarded. When a shorter path to a closed node is found, the node is moved to the open list and the shorter path is saved. The main disadvantage of pure heuristic search is that it is not guaranteed to find optimal solutions.

Breadth-first search, uniform-cost search, and pure heuristic search are all special cases of a more general class of algorithm called 'best-first search'. In each cycle of a best-first search, a node that is best according to some cost function is expanded. These best-first algorithms differ only in their cost functions: the depth for breadth-first search, the total cost for uniform-cost search, and the heuristic function for pure heuristic search.

The A* Algorithm

The A* algorithm combines features of uniform-cost search and pure heuristic search to efficiently compute optimal solutions. A* is a best-first search in which the cost associated with a node n is $f(n) = g(n) + h(n)$, where $g(n)$ is the cost of the path from the initial state to node n, and $h(n)$ is the heuristic estimate of the cost of a path from node n to a goal. Thus, $f(n)$ estimates the lowest total cost of any solution path passing through node n. At each cycle, a node with lowest f value is expanded. Ties

among nodes of equal f value are broken in favor of nodes with lower h values. The algorithm terminates when a goal node is chosen for expansion.

A* finds an optimal path to a goal if the heuristic function h is admissible, meaning that it never overestimates actual cost. For example, since Manhattan distance never overestimates the actual number of moves required in the sliding-tile puzzles, A* using this evaluation function will find optimal solutions to these problems. In addition, A* makes the most efficient use of a given heuristic function in the following sense: among all shortest-path algorithms using a given heuristic function, A* expands the fewest number of nodes.

The main disadvantage of A*, and indeed of any best-first search, is its memory requirement. Since the open and closed lists must be saved, A* is severely space-limited in practice, and is typically no more practical than breadth-first search. For example, while it can be run successfully on the eight puzzle, it quickly exhausts available memory on the fifteen puzzle.

Iterative-Deepening A*

Just as depth-first iterative deepening solved the space problem of breadth-first search, iterative-deepening A* (IDA*) solves the space problem of A*, without sacrificing solution optimality. Each iteration of the algorithm is a depth-first search that keeps track of the cost, $f(n) = g(n) + h(n)$, of each node generated. When the cost of the last node on a path exceeds a threshold for that iteration, the path is pruned, and the search backtracks before continuing. The cost threshold is initialized to the heuristic estimate of the initial state, and in each successive iteration is increased to the total cost of the lowest-cost node that was pruned during the previous iteration. The algorithm terminates when a goal state is reached whose cost does not exceed the current threshold.

Since IDA* performs a series of depth-first searches, its memory requirement is linear in the maximum search depth. Furthermore if the heuristic function is admissible, IDA* finds an optimal solution. Finally, by an argument similar to that presented for DFID, IDA* expands the same number of nodes, asymptotically, as A* on a tree, provided that the number of nodes grows exponentially with the solution cost. These facts, together with the optimality of A*, imply that IDA* is asymptotically optimal in time and space over all heuristic search algorithms that find optimal solutions on a tree. In addition, IDA* is much easier to implement, and often runs faster, than A*, since it does not incur the overhead of managing the open and closed lists.

Depth-First Branch-and-Bound

For many problems, a maximum search depth is known in advance, or the search tree is finite. For example, the traveling salesman problem (TSP) is to visit each of a set of cities once, and return to the starting city, in a tour of shortest total distance. The natural problem space for this problem consists of a tree where the root node represents the starting city, the children of the root represent all the cities that could be visited first, the nodes at the next level represent all the cities that could be visited second, and so on. In this tree, the maximum depth is the number of cities, and all solutions occur at this depth. In such a space, a simple depth-first search guarantees finding an optimal solution using space that is only linear in the number of cities.

Depth-first branch-and-bound (DFBB) makes this search more efficient by keeping track of the best solution found so far. Since the cost of a partial tour is the sum of the costs of the edges traveled on it, whenever a partial tour is found whose cost equals or exceeds the cost of the best complete tour found so far, the branch representing the partial tour can be pruned, since all its descendents must have equal or greater cost. Whenever a lower-cost complete tour is found, the cost of the 'best' tour is amended to this lower cost. To increase the pruning, an admissible heuristic function, such as the cost of the minimum spanning tree of the remaining unvisited cities, can be added to the cost so far of a partial tour. Finally, by carefully ordering the children of a given node from smallest to largest estimated total cost, a lower-cost solution can be found more quickly, further improving pruning efficiency.

IDA* and DFBB exhibit complementary behavior. Both are guaranteed to return an optimal solution using only linear space, assuming that their cost functions are admissible. In IDA*, the cost threshold is a lower bound on the optimal solution cost, and increases in each iteration until it reaches the optimal cost. In DFBB, the cost of the best solution found so far is an upper bound on the optimal solution cost, and decreases until it reaches the optimal cost. While IDA* never expands a node whose cost exceeds the optimal cost, its overhead consists of expanding some nodes more than once. While DFBB never expands a node more than once, its overhead consists of expanding some nodes whose costs exceed the

optimal cost. For problems whose search trees are of bounded depth, or for which it is easy to construct a good solution, such as the traveling salesman problem, DFBB is the algorithm of choice for finding an optimal solution. For problems with infinite search trees, or for which it is difficult to construct a low-cost solution, such as the sliding-tile puzzles or Rubik's cube, IDA* is the best choice. For example, using appropriate admissible heuristic functions, IDA* can optimally solve random instances of the twenty-four puzzle and Rubik's cube.

TWO-PLAYER GAMES

A second major application of heuristic search algorithms in AI is to two-player games. One of the original challenges of AI was to build a program that could play chess at the level of the best human players. In May 1997, a computer called Deep Blue, evaluating about 200 million chess positions per second, defeated Gary Kasparov, the world champion, in a six-game tournament.

Minimax Search

The standard algorithm for two-player perfect-information games, such as chess, checkers or Othello, is minimax search with heuristic static evaluation. The algorithm searches forward to a fixed depth in the game tree, limited by the amount of time available per move. At this 'search horizon', a heuristic function is applied to the frontier nodes. In a two-player game, a heuristic evaluation is a function that takes a board position and returns a number that indicates how favorable the position is for one player relative to the other. For example, a simple heuristic evaluator for chess might count the total number of pieces on the board for one player, appropriately weighted by their relative strengths, and subtract the weighted sum of the opponent's pieces. Thus, large positive values would correspond to strong positions for one player, called Max, whereas large negative values would represent advantageous situations for the opponent, called Min.

Given the heuristic evaluations of the frontier nodes, values for the interior nodes in the tree are recursively computed according to the 'minimax rule'. The value of a node where it is Max's turn to move is the maximum of the values of its children, while the value of a node where Min is to move is the minimum of its children's values. Thus, at alternate levels of the tree, the minimum or the maximum values of the children are backed up. This continues until the values of the immediate children of the current position are computed, at which point a move is made to the child with the maximum or minimum value, depending on whose turn it is to move. Figure 5 shows an example of a completely evaluated minimax tree.

Alpha–Beta Pruning

One of the most elegant of all AI search algorithms is alpha–beta pruning. The idea, similar to branch-and-bound, is that the minimax value of the root of a game tree can be determined without examining all the nodes at the search frontier.

Figure 6 shows an example of alpha–beta pruning. At the square nodes Max is to move, while at the circular nodes it is Min's turn. The search proceeds depth first to minimize the memory required, and evaluates a node only

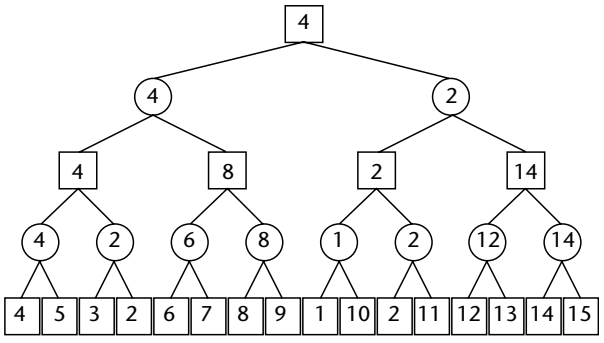


Figure 5. A minimax tree. At the square nodes it is Max to move; at the circular nodes, it is Min to move.

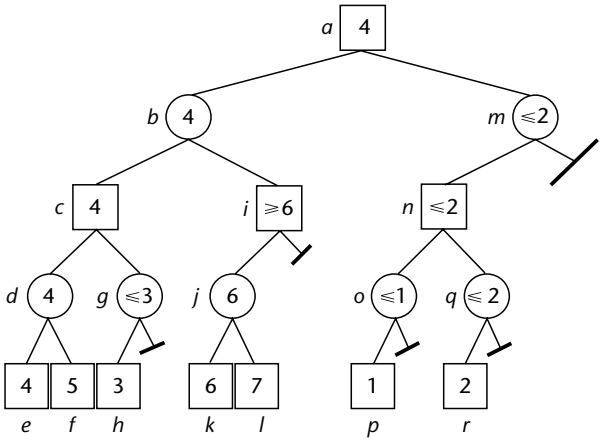


Figure 6. Alpha–beta pruning example.

when necessary. First, nodes e and f are statically evaluated as 4 and 5 respectively, and their minimum value is backed up to their parent node d. Node h is then evaluated as 3. The value of its parent node g must therefore be less than or equal to 3, since it is the minimum of 3 and the unknown value of its right child. The value of node c must then be 4, because it is the maximum of 4 and a value that is less than or equal to 3. Since we have determined the minimax value of node c, we do not need to evaluate or even generate the sibling of node h.

Similarly, after statically evaluating nodes k and l as 6 and 7 respectively, the backed-up value of their parent node j is the minimum of these values. This tells us that the minimax value of node i must be greater than or equal to 6, since it is the maximum of 6 and the unknown value of its right child. Since the value of node b is the minimum of 4 and a value that is greater than or equal to 6, it must be 4, and we have achieved another cut-off.

The right half of the tree shows an example of 'deep pruning'. After evaluating the left half of the tree, we know that the value of the root node a is greater than or equal to 4, the minimax value of node b. Once node p is evaluated as 1, the value of its parent node o must be less than or equal to 1. Since the value of the root is greater than or equal to 4, the value of node o cannot propagate to the root, and hence we need not generate the sibling of node p. A similar situation exists after the evaluation of node r as 2. At that point, the value of node o is less than or equal to 1 and the value of node q is less than or equal to 2, and hence the value of node n, which is the maximum of the values of nodes o and p, must be less than or equal to 2. Furthermore, since the value of node m is the minimum of the values of node n and its sibling, and node n has a value less than or equal to 2, the value of node m must also be less than or equal to 2. This inference allows the sibling of node n to be pruned, since the value of the root node a is greater than or equal to 4. Thus, we have computed the minimax value of the root of the tree to be 4, by generating only seven of sixteen leaf nodes.

Since alpha–beta pruning performs a minimax search while pruning much of the tree, its effect is to allow a deeper search with the same amount of computation. How much does alpha–beta improve performance? The best way to characterize the efficiency of a pruning algorithm is in terms of its 'effective branching factor'. The effective branching factor is the dth root of the number of frontier nodes that must be evaluated in a search to depth d, in the limit of large d.

The efficiency of alpha–beta pruning depends on the order in which nodes are encountered at the search frontier. In the worst case, alpha–beta will not perform any cut-offs at all. In that case, all frontier nodes must be evaluated and the effective branching factor is b, the brute-force branching factor.

On the other hand, in the best case of perfect ordering, every possible cut-off is realized. In that case, the effective branching factor is reduced from b to $b^{1/2}$, the square root of the brute-force branching factor. In other words, in the best case, one can search twice as deep with alpha–beta pruning as without, using the same amount of computation.

In between worst-possible ordering and perfect ordering is random ordering, which is the average case. Under random ordering of the frontier nodes, alpha–beta pruning reduces the effective branching factor to approximately $b^{3/4}$. This means that one can search 4/3 as deep with alpha–beta; that is, one can achieve roughly a 33% improvement in search depth.

In practice, however, the effective branching factor of alpha–beta is closer to the best case of $b^{1/2}$ because of 'node ordering'. Instead of generating the tree from left to right, we can reorder the tree based on static evaluations of the interior nodes. In other words, the children of Max nodes are expanded in decreasing order of their static values, while the children of Min nodes are expanded in increasing order of their static values.

CONCLUSION

We have examined search algorithms for two different classes of problems. In the case of single-agent problems, the task is to find a sequence of operators that map an initial state to a desired goal state. Much of the work in this area has focused on finding optimal solutions to such problems, often making use of admissible heuristic functions to speed up the search without sacrificing solution optimality. In the case of two-player games, finding optimal solutions is infeasible, and research has focused on algorithms for making the best possible move decisions given a limited amount of computing time. While these two types of problem are different, similar ideas, such as brute-force searches and heuristic evaluation functions, can be applied to both.

Acknowledgment

This work was supported in part by NSF Grant IRI-9619447.

Further Reading

Bolc L and Cytowski J (1992) *Search Methods for Artificial Intelligence*. London, UK: Academic Press.

Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische Mathematik* **1**: 269–271.

Hart PE, Nilsson NJ and Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* **4**(2): 100–107.

Kanal L and Kumar V (eds) (1988) *Search in Artificial Intelligence*. New York, NY: Springer-Verlag.

Knuth DE and Moore RE (1975) An analysis of alpha–beta pruning. *Artificial Intelligence* **6**(4): 293–326.

Korf RE (1998) Artificial intelligence search algorithms. In: Atallah MJ (ed.) *Algorithms and Theory of Computation Handbook*, pp. 36-1 to 36-20. Boca Raton, FL: CRC Press.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Pearl J (1984) *Heuristics*. Reading, MA: Addison-Wesley.

Samuel AL (1963) Some studies in machine learning using the game of checkers. In: Feigenbaum E and Feldman J (eds) *Computers and Thought*, pp. 71–105. New York, NY: McGraw-Hill.

Shannon CE (1950) Programming a computer for playing chess. *Philosophical Magazine* **41**: 256–275.

Turing AM (1950) Computing machinery and intelligence. *Mind* **59**: 433–460. [Reprinted in: Feigenbaum E and Feldman J (eds) (1963) *Computers and Thought*. New York, NY: McGraw-Hill.]

Semantic Networks

Intermediate article

John F Sowa, Vivo Mind LLC, New York, USA

CONTENTS

A semantic network is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs.

INTRODUCTION

A *semantic network* or *net* is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology, and linguistics.

What is common to all semantic networks is a declarative graphic representation that can be used either to represent knowledge or to support automated systems for reasoning about knowledge. Some versions are highly informal, but other versions are formally defined systems of logic. The following are six of the most common kinds of semantic network, each of which is discussed in detail in this article:

1. *Definitional networks* emphasize the *subtype* or *is-a* relation between a concept type and a newly defined subtype. The resulting network, also called a *generalization* or *subsumption* hierarchy, supports the rule of *inheritance* for copying properties defined for a supertype to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true.
2. *Assertional networks* are designed to assert propositions. Unlike definitional networks, the information in an assertional network is assumed to be contingently true, unless it is explicitly marked with a modal operator. Some assertional networks have been proposed as models of the *conceptual structures* underlying natural language semantics.
3. *Implicational networks* use implication as the primary relation for connecting nodes. They may be used to represent patterns of beliefs, causality, or inferences.
4. *Executable networks* include some mechanism, such as marker passing or attached procedures, which can perform inferences, pass messages, or search for patterns and associations.
5. *Learning networks* build or extend their representations by acquiring knowledge from examples. The new knowledge may change the old network by adding and deleting nodes and arcs or by modifying numerical values, called *weights*, associated with the nodes and arcs.
6. *Hybrid networks* combine two or more of the previous techniques, either in a single network or in separate, but closely interacting, networks.

Some of the networks have been explicitly designed to implement hypotheses about human cognitive mechanisms, while others have been designed primarily for computer efficiency. Sometimes, computational reasons may lead to the same conclusions as psychological evidence.

Network notations and linear notations are both capable of expressing equivalent information, but certain representational mechanisms are better suited to one form or the other. Since the boundary lines are vague, it is impossible to give necessary and sufficient conditions that include all semantic networks while excluding other systems that are not usually called semantic networks. The last section of this article discusses the syntactic mechanisms used to express information in network notations and compares them to the corresponding mechanisms used in linear notations.

DEFINITIONAL NETWORKS

The oldest known semantic network was drawn in the third century AD by the Greek philosopher Porphyry in his commentary on Aristotle's categories. Porphyry used it to illustrate Aristotle's method of defining categories by specifying the *genus* or general type and the *differentiae* that distinguish different subtypes of the same supertype. Figure 1 shows a version of the *Tree of Porphyry*, as it was drawn by the logician Peter of Spain circa 1329. It illustrates

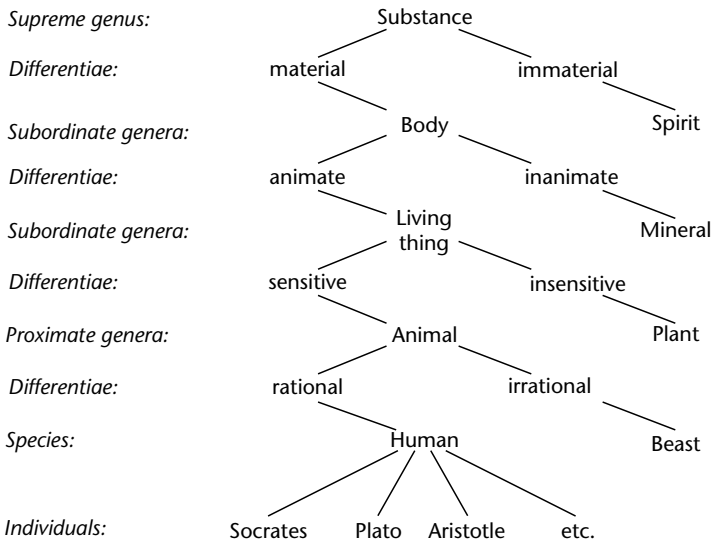


Figure 1. Tree of Porphyry, as drawn by Peter of Spain (1329).

the categories under Substance, which is called the *supreme genus* or the most general category.

Despite its age, the Tree of Porphyry represents the common core of all modern hierarchies that are used for defining concept types. In Figure 1, the genus Substance with the differentia material is Body and with the differentia immaterial is Spirit. The modern rule of *inheritance* is a special case of the Aristotelian syllogisms, which specify the conditions for inheriting properties from supertypes to subtypes: Living Thing inherits material Substance from Body and adds the differentia animate; Human inherits sensitive animate material Substance and adds the differentia rational. Aristotle, Porphyry, and the medieval logicians also distinguished the categories or *universals* from the individual instances or *particulars*, which are listed at the bottom of Figure 1. Aristotle's methods of definition and reasoning are still used in artificial intelligence, object-oriented programming languages, and every dictionary from the earliest days to the present.

The first implementations of semantic networks were used to define concept types and patterns of relations for machine translation systems. Silvio Ceccato (1961) developed *correlational nets*, which were based on 56 different relations including subtype, instance, part-whole, case relations, kinship relations, and various kinds of attributes. He used the correlations as patterns for guiding a parser and resolving syntactic ambiguities. Margaret Masterman's system at Cambridge University (1961) was the first to be called a semantic network. She developed a list of 100 primitive concept types such as Folk, Stuff, Thing, Do, and Be. In terms of those primitives, her group defined a conceptual dictionary of 15 000 entries. She organized the concept types into a lattice, which permits inheritance from multiple supertypes.

Among current systems, the *description logics* include the features of the Tree of Porphyry as a minimum, but they may also add various extensions. They are derived from an approach proposed by Woods (1975) and implemented by Brachman (1979) in a system called Knowledge Language One (KL-ONE). As an example, Figure 2 shows a KL-ONE network that defines the concepts Truck and TrailerTruck as subtypes of Vehicle.

Figure 2 has nine ovals for concept nodes and nine arrows which represent different kinds of links. The white ovals represent *generic concepts* for the types, as distinguished from the shaded oval, which is an *individual concept* for the instance 18. The oval marked with an asterisk* indicates that Integer is a built-in or primitive type. The concepts Truck and TrailerTruck are defined in Figure 2, but Vehicle, Trailer, WtMeasure, and VolMeasure would have to be defined by other KL-ONE diagrams.

The double-line arrows represent subtype–supertype links from TrailerTruck to Truck and from Truck to Vehicle. The arrows with a squared circle in the middle represent *roles*. The Truck node has four roles labeled UnloadedWt, MaxGrossWt, CargoCapacity, and NumberOfWheels. The TrailerTruck node has two roles, one labeled HasPart and one that restricts the NumberOfWheels role of Truck to the value 18. The notation *v/r* at the target

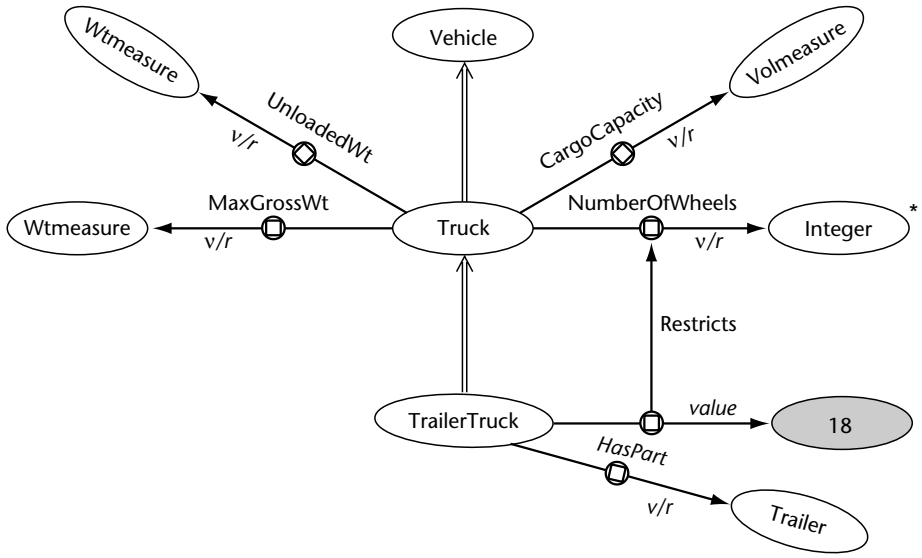


Figure 2. Truck and TrailerTruck concepts defined in KL-ONE.

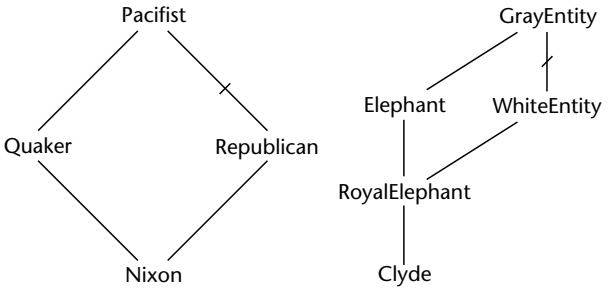


Figure 3. Conflicting defaults in a definitional network.

end of the role arrows indicates *value restrictions* or type constraints on the permissible values for those roles.

The Tree of Porphyry, KL-ONE, and many versions of description logics are subsets of classical first-order logic (FOL). They belong to the class of *monotonic logics*, in which new information monotonically increases the number of provable theorems, and none of the old information can ever be deleted or modified. Some versions of description logics support *non-monotonic reasoning*, which allows *default rules* to add optional information and *canceling rules* to block inherited information. Such systems can be useful for many applications, but they can also create problems of *conflicting defaults*, as illustrated in Figure 3.

The *Nixon diamond* on the left shows a conflict caused by inheritance from two different supertypes: by default, Quakers are pacifists, and Republicans are not. Does Nixon inherit pacifism along the Quaker path, or is it blocked by the negation on the Republican path? On the right is another diamond in which the subtype RoyalElephant cancels the property of being gray, which is the default color for ordinary elephants. If Clyde is first mentioned as an elephant, his default color would be gray, but later information that he is a RoyalElephant should cause the previous information to be retracted. To resolve such conflicts, many developers have rejected local defaults in favor of more systematic methods of *belief revision* that can guarantee global consistency.

Although the basic methods of description logics are as old as Aristotle, they remain a vital part of many versions of semantic networks and other kinds of systems. Much of the ongoing research on description logics has been devoted to increasing their expressive power while remaining within an efficiently computable subset of logic. Two recent description logics are DAML and OIL, which are intended for representing knowledge in the *semantic web* – a giant semantic network that spans the entire Internet.

ASSERTIONAL NETWORKS

Gottlob Frege (1879) developed a tree notation for the first complete version of first-order logic – his *Begriffsschrift* or *concept writing*. Charles Sanders Peirce (1880, 1885) independently developed an algebraic notation which, with a change of symbols by Peano, has become the modern notation for predicate calculus. Although Peirce invented the algebraic notation, he was never fully satisfied with it. As early as 1882, he was searching for a

graphic notation, similar to the notations used in organic chemistry, that would more clearly show 'the atoms and molecules of logic'. Figure 4 shows one of his *relational graphs*, which represents the sentence '*A Stagirite teacher of a Macedonian conqueror of the world is a disciple and an opponent of a philosopher admired by Church Fathers*.'

Figure 4 contains three branching *lines of identity*, each of which corresponds to an existentially quantified variable in the algebraic notation. The words and phrases attached to those lines correspond to the relations or predicates in the algebraic notation. With that correspondence, Figure 4 can be translated to the following formula in predicate calculus:

$$(\exists x)(\exists y)(\exists z)(\text{isStagirite}(x) \wedge \text{teaches}(x,y) \wedge \text{isMacedonian}(y) \wedge \text{conquersTheWorld}(y) \wedge \text{isDiscipleOf}(y,z) \wedge \text{isOpponentOf}(y,z) \wedge \text{isAdmiredByChurchFathers}(z))$$

As this formula illustrates, a relational graph represents only two logical operators: the conjunction $\wedge$ and the existential quantifier $\exists$. Other operators, such as negation $\sim$, disjunction $\vee$, implication $\supset$, and the universal quantifier $\forall$, are more difficult to express because they require some method for demarcating the *scope* – that part of the formula that is governed by the operator. The problem of representing scope, which Peirce faced in his graphs of 1882, also plagued the early semantic networks used in artificial intelligence 80 years later.

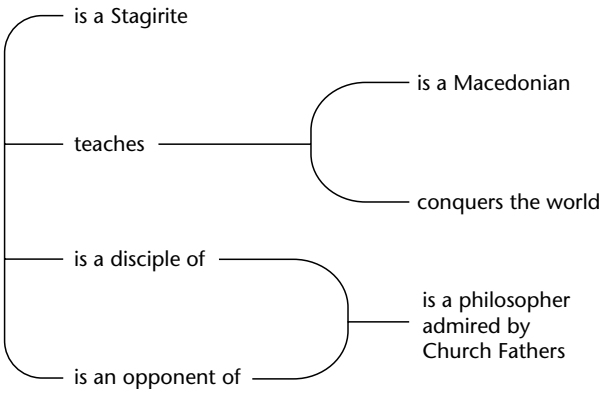


Figure 4. A relational graph.

In 1897, Peirce made a simple but brilliant discovery that solved all the problems at once: he introduced an oval that could enclose and negate an arbitrarily large graph or subgraph. Then combinations of ovals with conjunction and the existential quantifier could express all the logical operators used in the algebraic notation. That innovation transformed the relational graphs into the system of *existential graphs* (EG), which Peirce called 'the logic of the future'. The implication $\supset$, for example, could be represented with a nest of two ovals, since $(p \supset q)$ is equivalent to $\sim(p \wedge \sim q)$. At the left of Figure 5 is an existential graph for the sentence '*If a farmer owns a donkey, then he beats it*.'

The outer oval of Figure 5 is the antecedent or *if* part, which contains *farmer*, linked by a line representing $(\exists x)$ to *owns*, which is linked by a line representing $(\exists y)$ to *donkey*. The subgraph in the outer oval may be read *If a farmer x owns a donkey y*. The lines x and y are extended into the inner oval, which represents the consequent, *then x beats y*. Figure 5 may be translated to the following algebraic formula:

$$\sim (\exists x)(\exists y)(\text{farmer}(x) \wedge \text{donkey}(y) \wedge \text{owns}(x,y) \wedge \sim \text{beats}(x,y))$$

This formula is equivalent to

$$(\forall x)(\forall y)((\text{farmer}(x) \wedge \text{donkey}(y) \wedge \text{owns}(x,y)) \supset \text{beats}(x,y))$$

For comparison, the diagram on the right of Figure 5 is a *discourse representation structure* (DRS), which Hans Kamp (Kamp and Reyle, 1993) invented to represent natural language semantics. Instead of nested ovals, Kamp used boxes linked by arrows; and instead of lines of identity, Kamp used variables. But the logical structures are formally equivalent, and the same techniques for

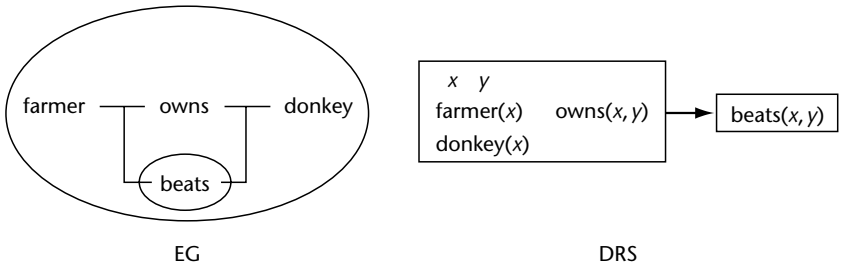


Figure 5. EG and DRS for 'If a farmer owns a donkey, then he beats it.'

representing logic in natural language can be adapted to either notation.

In linguistics, Lucien Tesnière (1959) developed graph notations for his system of *dependency grammar*. Figure 6 shows one of his graphs for the sentence '*L'autre jour, au fond d'un vallon, un serpent piqua Jean Fréron*' (The other day, at the bottom of a valley, a snake stung Jean Fréron). At the top is the verb *piqua* (stung), from which the words that depend directly on the verb are hanging: the subject (*serpent*), the object (*Jean*), and two prepositional phrases. The bull's-eye symbol indicates an implicit preposition (*à*). Every word other than *piqua* is hanging below some word on which it depends.

Tesnière has had a major influence on linguistic theories that place more emphasis on semantics than on syntax. The dependency theories have also been strongly influenced by *case grammar* (Fillmore, 1968), which provides a convenient set of labels for the arcs of the graphs.

Roger Schank adopted the dependency approach, but shifted the emphasis to concepts rather than words (Schank, 1975). Figure 7 shows a *conceptual dependency graph* for the sentence '*A dog is greedily eating a bone*.' Instead of Tesnière's tree notation, Schank used different kinds of arrows for different relations, such as ⇔ for the agent–verb relation or an arrow marked with *o* for object. He replaced the verb *eat* with one of his primitive acts *ingest*; he replaced adverbs like *greedily* with adjective forms like *greedy*; and he added the linked arrows marked with *d* for direction to show that the bone goes from some unspecified place into the dog (the subscript 1 indicates that the bone went into the same thing that ingested it). Conceptual dependencies were primarily suited to representing information at the sentence level, but Schank and his colleagues later developed notations for representing larger structures, in which the sentence-level dependencies occurred as nested substructures. To learn or discover the larger structures automatically, *case-based reasoning* has been used to search for commonly occurring patterns among the lower-level conceptual dependencies (Schank *et al*., 1994).

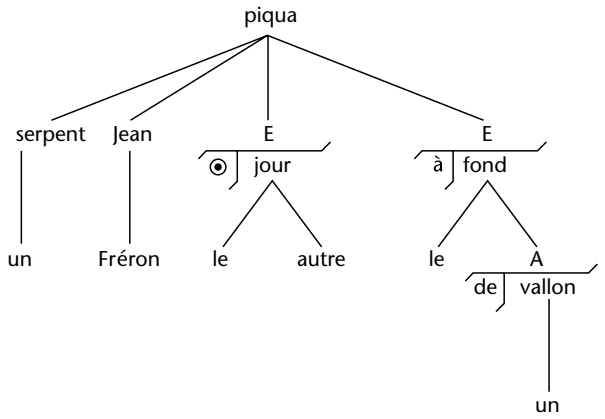


Figure 6. A dependency graph in Tesnière's notation.

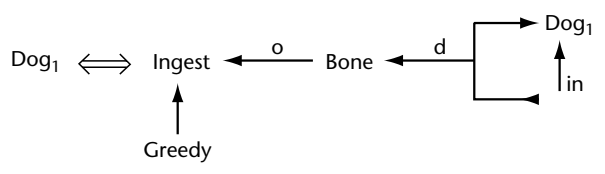


Figure 7. Schank's notation for conceptual dependencies.

Logically, Tesnière's dependency graphs have the same expressive power as Peirce's relational graphs of 1882: the only logical operators they can represent are conjunction and the existential quantifier. Even when those graphs have nodes marked with other logical operators, such as disjunction, negation, or the universal quantifier, they fail to express their scope correctly. During the 1970s, various network notations were developed to represent the scope of logical operators. The most successful approach was the method of adding explicit nodes to show propositions. Logical operators would connect the propositional nodes, and relations would either be attached to the propositional nodes or be nested inside them. By those criteria, Frege's *Begriffsschrift*, Peirce's existential graphs, and Kamp's discourse representation structures could be called *propositional semantic networks*. In Figure 5, for example, the two EG ovals and the two DRS boxes represent propositions, each of which contains nested propositions.

The first propositional semantic network to be implemented in artificial intelligence (AI) was the MIND system, developed by Stuart Shapiro (1971). It later evolved into the *Semantic Network Processing System* (SNePS), which has been used to represent a wide range of features in natural language semantics (Shapiro, 1979). Figure 8 shows the SNePS representation for the sentence '*Sue thinks that Bob believes that a dog is eating a bone*.' Each of the nodes labeled M1 through M5 represents a distinct proposition, whose relational content is attached to the propositional node.

The proposition M1 states that Sue is the experiencer (Expr) of the verb *think*, whose theme (Thme) is another proposition M2. For M2, the experiencer is Bob, the verb is *believe*, and the theme is a

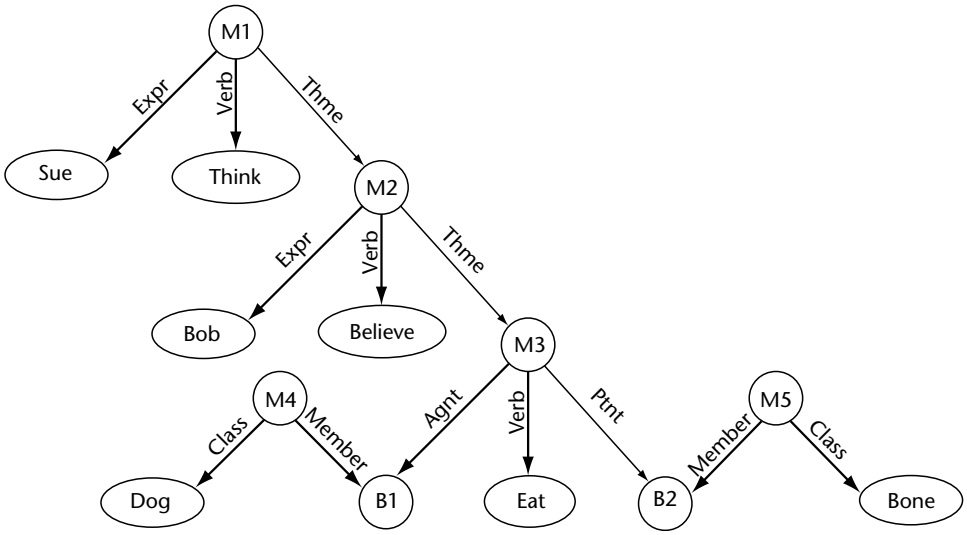


Figure 8. Propositions represented in SNePS.

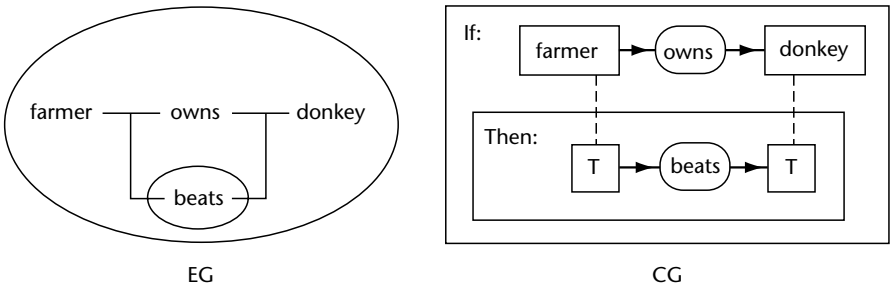


Figure 9. Comparison of the EG from Figure 5 with a CG for the same sentence.

proposition M3. For M3, the agent (Agnt) is some entity B1, which is a member of the class Dog, the verb is *eat*, and the patient (Ptnt) is an entity B2, which is a member of the class Bone. As Figure 8 illustrates, propositions may be used at the metalevel to make statements about other propositions: M1 states that M2 is thought by Sue, and M2 states that M3 is believed by Bob.

Conceptual graphs (Sowa, 1984, 2000) are a variety of propositional semantic networks in which the relations are nested inside the propositional nodes. They evolved as a combination of the linguistic features of Tesnière's dependency graphs and the logical features of Peirce's existential graphs with strong influences from the work in artificial intelligence and computational linguistics. Figure 9 shows a comparison of Peirce's EG from Figure 5 with a conceptual graph (CG) that represents the sentence '*If a farmer owns a donkey, then he beats it*.'

The most obvious differences between the EG and the CG are cosmetic: the ovals are squared off to form boxes, and the implicit negations in the EG are explicitly marked *If* and *Then* for better readability. The more subtle differences are in the range of quantification and the point where the quantification occurs. In an EG, a line of identity represents an existential quantifier ($\exists x$) or ($\exists y$), which ranges over anything in the domain of discourse; but in a CG, each box, called a *concept*, represents a quantifier ($\exists x$:Farmer) or ($\exists y$:Donkey), which is restricted to the *type* or *sort* Farmer or Donkey. In the CG, the arcs with arrows indicate the arguments of the relations (numbers are used to distinguish the arcs for relations with more than two arguments). Nodes such as [T] represent the pronouns *he* or *it*, which are linked to their antecedents by dotted lines called *coreference links*. As another example, Figure 10 shows the CG that corresponds to the SNePS diagram in Figure 8.

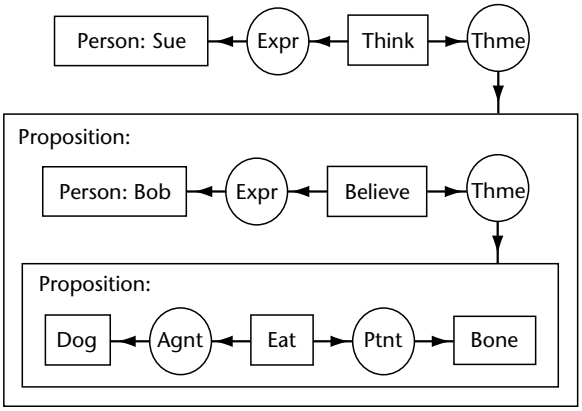


Figure 10. A conceptual graph that corresponds to Figure 8.

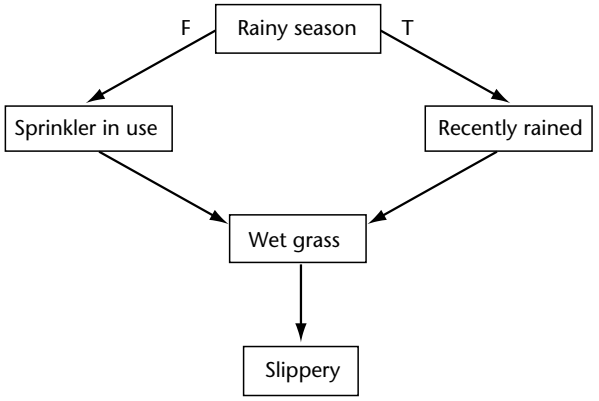


Figure 11. An implicational network for reasoning about wet grass.

Figures 8 and 10 both represent the sentence '*Sue thinks that Bob believes that a dog is eating a bone*'. The SNePS proposition M1 corresponds to the entire CG in Figure 10; M2 corresponds to the concept box that contains the CG for the nested proposition *Bob believes that a dog is eating a bone*; and M3 corresponds to the concept box that contains the CG for the more deeply nested proposition *A dog is eating a bone*. Each concept box in a CG could be considered a separate proposition node that could be translated to a complete sentence by itself. The concept [Dog] could be expressed by the sentence '*There exists a dog*', which corresponds to the SNePS proposition M4. The concept [Person: Sue] expresses the sentence '*There exists a person named Sue*'. By such methods, it is possible to translate propositions expressed in SNePS or CGs to equivalent propositions in the other notation. For most sentences, the translations are nearly one-to-one, but sentences that take advantage of special features in one notation may require a more roundabout paraphrase when translated to the other.

Different versions of propositional semantic networks have different syntactic mechanisms for associating the relational content with the propositional nodes, but formal translation rules can be defined for mapping one version to another. Peirce, Sowa, and Kamp used strictly nested propositional enclosures with variables or lines to show coreferences between different enclosures. Frege and Shapiro attached the relations to the propositional nodes (or lines in Frege's notation). Gary Hendrix (1979) developed a third option: *partitions* that enclose the relational content, but with the option of overlapping enclosures if they have common components. Formally, Hendrix's solution is equivalent to Shapiro's; but as a practical matter, it is not possible to draw the partitions on a plane sheet if multiple enclosures overlap in complex ways.

IMPLICATIONAL NETWORKS

An implicational network is a special case of a propositional semantic network in which the primary relation is implication. Other relations may be nested inside the propositional nodes, but they are ignored by the inference procedures. Depending on the interpretation, such networks may be called *belief networks, causal networks, Bayesian networks*, or *truth-maintenance systems*. Sometimes the same graph can be used with any or all of these interpretations. Figure 11 shows possible causes for slippery grass: each box represents a proposition, and the arrows show the implications from one proposition to another. If it is the rainy season, the arrow marked T implies that it recently rained; if not, the arrow marked F implies that the sprinkler is in use. For boxes with only one outgoing arrow, the truth of the first proposition implies the truth of the second, but falsity of the first makes no prediction about the second.

Suppose someone walking across a lawn slips on the grass. Figure 11 represents the kind of background knowledge that the victim might use to reason about the cause. A likely cause of slippery grass is that the grass is wet. It could be wet because either the sprinkler had been in use or it had recently rained. If it is the rainy season, the sprinkler would not be in use. Therefore, it must have rained.

The kind of reasoning described in the previous paragraph can be performed by various AI systems. Chuck Rieger developed a version of *causal networks*, which he used for analyzing problem descriptions in English and translating them to a network that could support metalevel reasoning.

Benjamin Kuipers, who was strongly influenced by Rieger's approach, developed methods of *qualitative reasoning*, which serve as a bridge between the symbolic methods of AI and the differential equations used in physics and engineering. Judea Pearl, who has developed techniques for applying statistics and probability to AI, introduced *belief networks*, which are causal networks whose links are labeled with probabilities.

Different methods of reasoning can be applied to the same basic graph, such as Figure 11, sometimes with further annotations to indicate truth values or probabilities. Following are two of the major approaches:

- *Logic*. Methods of logical inference are used in *truth-maintenance systems* (TMS). A TMS would start at nodes whose truth values are known and propagate them throughout the network. For the case of the person who slipped on the grass, it would start with the value T for the fact that the grass is slippery and work backwards. Alternatively, a TMS could start with the fact that it is now the rainy season and work forwards. By combinations of forward and backward reasoning, a TMS propagates truth values to nodes whose truth value is unknown. Besides deducing new information, a TMS can be used to verify consistency, search for contradictions, or find locations where the expected implications do not hold. When contradictions are found, the structure of the network may be modified by adding or deleting nodes; the result is a kind of non-monotonic reasoning called *belief revision*.
- *Probability*. Much of the forward and backward reasoning used with a TMS can also be adapted to a probabilistic interpretation, since truth can be considered a probability of 1.0 and falsity as 0.0. The continuous range of probabilities from 1.0 to 0.0, however, raises the need for more subtle interpretations and more complexity in the computations. The most detailed study of probabilistic reasoning in causal or belief networks has been done by Pearl. For Figure 11, a two-valued {T, F} interpretation is only a rough approximation, since it doesn't rain every day in a rainy season and a sprinkler might not be used even in a dry season. Pearl analyzed various techniques for applying Bayesian statistics to derive a causal network from observed data and to reason about it.

In both the logic-based and the probabilistic systems, the relational information that was used to derive the implications is ignored by the inference procedures. Doyle developed the first TMS by extracting a subgraph of implications from the rules of an expert system. Martins and Shapiro extracted a TMS from SNePS by analyzing only the Boolean connectives that link propositional nodes. Similar techniques could be applied to other propositional networks to derive an implicational subgraph that could be analyzed by logical or probabilistic methods.

Although implicational networks emphasize implication, they are capable of expressing all the Boolean connectives by allowing a conjunction of inputs to a propositional node and a disjunction of outputs. Gerhard Gentzen (1935) showed that a collection of implications in that form could express all of propositional logic. Following is the general form of an implication written in Gentzen's *clause form*:

$$p_1, \ldots, p_n \Rightarrow q_1, \ldots, q_m$$

The ps are called the *antecedents* of the implication, and the qs are called the *consequents*. The generalized rule of modus ponens states that when every one of the antecedents is true, at least one of the consequents must be true. In effect, the commas in the antecedent have the effect of *and* operators, and the commas in the consequent have the effect of *or* operators. Doyle's original TMS allowed only one term in the consequent; the resulting form, called *Horn-clause logic*, is widely used for expert systems. To support full propositional logic, later versions of TMS have been generalized to allow multiple *or* operators in the consequent.

EXECUTABLE NETWORKS

Executable semantic networks contain mechanisms that can cause some change to the network itself. The executable mechanisms distinguish them from networks that are static data structures, which can change only through the action of programs external to the net itself. Three kinds of mechanisms are commonly used with executable semantic networks:

1. *Message passing* networks can pass data from one node to another. For some networks, the data may consist of a single bit, called a *marker, token*, or *trigger*; for others, it may be a numeric *weight* or an arbitrarily large *message*.
2. *Attached procedures* are programs contained in or associated with a node that performs some kind of action or computation on data at that node or some nearby node.
3. *Graph transformations* combine graphs, modify them, or break them into smaller graphs. In typical theorem provers, such transformations are carried out by a program external to the graphs. When they are triggered by the graphs themselves, they behave like chemical reactions that combine molecules or break them apart.

These three mechanisms can be combined in various ways. Messages passed from node to node may

be processed by procedures attached to those nodes, and graph transformations may also be triggered by messages that appear at some of the nodes.

An important class of executable networks was inspired by the work of the psychologist Otto Selz (1913, 1922), who was dissatisfied with the undirected associationist theories that were then current. As an alternative, Selz proposed *schematic anticipation* as a goal-directed method of focusing the thought processes on the task of filling empty slots in a pattern or *schema*. Figure 12 is an example of a schema that Selz asked his test subjects to complete while he recorded their verbal protocols.

The expected answers for the empty slots in Figure 12 are the supertypes of the words at the bottom: the supertype of Newspaper and Magazine is Periodical, and the supertype of Periodical and Book is Publication. This task is actually more difficult in German than in English: Selz's subjects tried to find a one-word supertype for *Zeitung* (Newspaper) and *Zeitschrift* (Magazine), but the correct answer in German is the two-word phrase *periodische Druckschrift*.

The similarity between Selz's method of schematic anticipation and the goal-directed methods of AI is not an accident. Two of the pioneers in AI, Herbert Simon and Allen Newell, (1972) adopted Selz's method of protocol analysis for their study of human problem solving. Their student, Ross Quillian, combined Selz's networks with the semantic networks used in machine translation. Quillian's most significant innovation was the *marker passing* algorithm for *spreading activations*, which was adopted for later systems, such as NETL by Scott Fahlman (1979) and the *massively parallel* algorithms by Hendler (1992).

The simplest networks with attached procedures are *dataflow graphs*, which contain passive nodes that hold data and active nodes that take data from *input nodes* and send results to *output nodes*. Figure 13 shows a dataflow graph with boxes for the passive nodes and diamonds for the active nodes. The labels on the boxes indicate the data type (Number or String), and the labels on the diamonds indicate the name of the function (+, ×, or convert string to number).

For numeric computations, dataflow graphs have little advantage over the algebraic notation used in common programming languages. Figure 13, for example, would correspond to an assignment statement of the following form:

$$X = (A + B)^* S2N(C)$$

Graphic notations are more often used in an *Integrated Development Environment* (IDE) for linking multiple programs to form a complete system. When dataflow graphs are supplemented with a graphic method for specifying conditions, such as *if-then-else*, and a way of defining recursive functions, they can form a complete programming language, similar to *functional programming languages* such as Scheme and ML.

Petri nets, first introduced by Carl Adam Petri, are the most widely-used formalism that combines marker passing with procedures. Like dataflow diagrams, Petri nets have passive nodes, called *places*, and active nodes, called *transitions*. In addition, they have a set of rules for marking places with dots, called *tokens*, and for executing or *firing* the transitions. To illustrate the flow of tokens, Figure 14 shows a Petri net for a bus stop where three tokens represent people waiting and one token represents an arriving bus.

At the upper left of Figure 14, each of the three tokens represents one person waiting at the bus stop. The token at the upper right represents an arriving bus. The transition labeled *Bus stops*

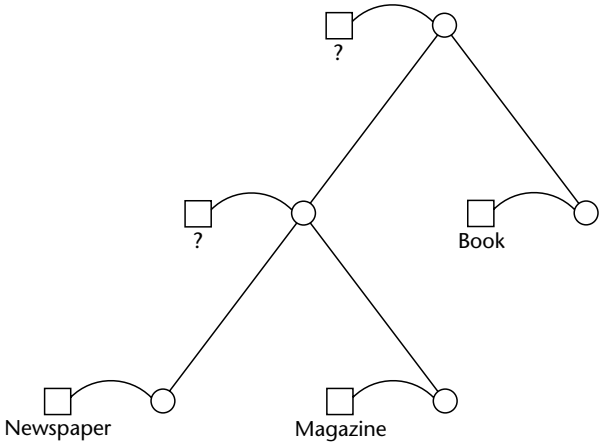


Figure 12. A schema used in Otto Selz's experiments.

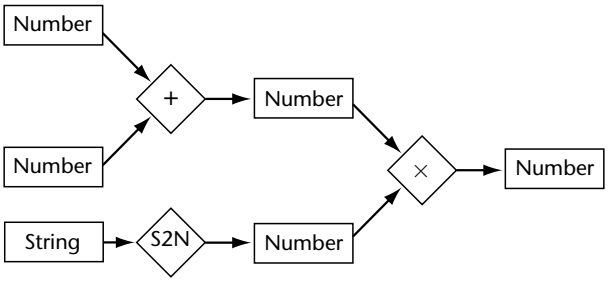


Figure 13. A dataflow graph.

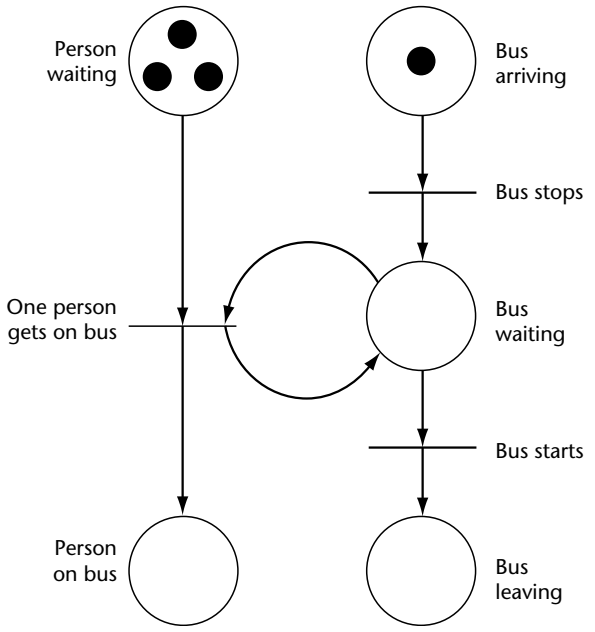


Figure 14. Petri net for a bus stop.

represents an event that fires by removing the token from the arriving place and putting a token in the waiting place. When the bus is waiting, the transition labeled *One person gets on bus* is *enabled* because it has at least one token in both of its input places. It fires by first removing one token from both of its input places and putting one token in both of its output places (including the *Bus waiting* place from which one token has just been removed). As long as the bus is waiting and there are more people waiting, that transition can keep firing. It stops firing when either there are no more people waiting or the *Bus starts* transition fires by removing the token for the waiting bus and putting a token in the place for *Bus leaving*.

Each place in a Petri net represents a precondition for the transitions that use it as an input and a postcondition for the transitions that use it as an output. A token in a place asserts that the corresponding condition is true. By removing a token from each input place, the firing of a transition retracts the assertions of its preconditions. By adding a token to each output place, the firing asserts that each of the postconditions has become true. Petri nets can be used to model or simulate physical events, as in the example of Figure 14. They can also be used to model processes that take place in computer hardware and software; they are especially useful for designing and modeling distributed parallel processes. In Figure 14, each token represents a single bit of information, but an extension, called *colored Petri nets*, can associate an arbitrary amount of data with each token. With such extensions, Petri nets can represent arbitrarily many dataflow graphs running in parallel or simulate the various marker-passing algorithms used in semantic networks in the Quillian tradition.

Although dataflow graphs and Petri nets are not usually called semantic networks, similar techniques have been implemented in *procedural semantic networks* (Levesque and Mylopoulos, 1979). Their systems incorporate definitional networks for defining *classes*, assertional networks for stating facts, and procedures similar to the *methods* of object-oriented programming languages. For conceptual graphs, Sowa allowed some relation nodes to be replaced by *actors*, which are functions that form the equivalent of a dataflow graph.

Besides markers and procedures, the third method for making networks executable is to let them grow and change dynamically. Peirce and Selz could also be considered pioneers of that approach. Peirce said that the inference operations on existential graphs could be considered 'a moving picture of thought'. For schematic anticipation, Selz considered a schema to be the cause of the neural activity that generates a solution to a problem. Formally, transformations on networks can be defined without reference to the mechanisms that perform the transformations. In Petri nets, for example, the definition states that a transition may 'fire' when each of its input nodes contains a token; the mechanism that performs the firing could be internal or external to the transition. For a computer implementation, it may be convenient to treat the networks as passive data structures and to write a program that manipulates them. For a cognitive theory, however, the transformation could be interpreted as network operations initiated and carried out by the network itself. Either interpretation could be found consistent with the same formal definitions.

LEARNING NETWORKS

A learning system, natural or artificial, responds to new information by modifying its internal representations in a way that enables the system to respond more effectively to its environment. Systems that use network representations can modify the networks in three ways:

1. *Rote memory*. The simplest form of learning is to convert the new information to a network and add it without any further changes to the current network.
2. *Changing weights*. Some networks have numbers, called *weights*, associated with the nodes and arcs. In an implicational network, for example, those weights

might represent probabilities, and each occurrence of the same type of network would increase the estimated probability of its recurrence.

3. *Restructuring*. The most complex form of learning makes fundamental changes to the structure of the network itself. Since the number and kinds of structural changes are unlimited, the study and classification of restructuring methods is the most difficult, but potentially the most rewarding if good methods can be found.

Systems that learn by rote or by changing weights can be used by themselves, but systems that learn by restructuring the network typically use one or both of the other methods as aids to restructuring.

Commercially, rote memory is of enormous importance, since the world economy depends on exact record keeping. For such applications, information is sometimes stored in tables, as in relational databases, but networks are also used. Either representation could be converted to the other. For better efficiency and usability, most database systems add indexes to speed up the search, and they support query languages, such as SQL, which perform transformations to extract and combine the information necessary to answer a request. Since a learning system must be able to distinguish common features and exceptions among similar examples, another feature is essential: the ability to measure *similarity* and to search the database for networks that are similar, but not identical to any given example.

Neural nets are a widely used technique for learning by changing the weights assigned to the nodes or arcs of a network. Their name, however, is a misnomer, since they bear little resemblance to actual neural mechanisms. Figure 15 shows a typical neural net, whose input is a sequence of numbers that indicate the relative proportion of some selected features and whose output is another sequence of numbers that indicate the most likely concept characterized by that combination of features. In an application such as optical character recognition, the features might represent lines, curves, and angles, and the concepts might represent the letters that have those features.

In a typical neural network, the structure of nodes and arcs is fixed, and the only changes that may occur are the assignments of weights to the arcs. When a new input is presented, the weights on the arcs are combined with the weights on the input features to determine the weights in the *hidden layers* of the net and ultimately the weights on the outputs. In the learning stage, the system is told whether the predicted weights are correct, and various methods of *back propagation* are used to adjust the weights on the arcs that lead to the result.

Rote memory is best suited to applications that require exact retrieval of the original data, and methods of changing weights are best suited to pattern recognition. For more versatile and creative kinds of learning, some way of restructuring the network is necessary. But the number of options for reorganizing a network is so vast that the full range of possibilities is largely unexplored. Here are some examples:

1. Patrick Winston (1975) used a version of relational graphs to describe structures, such as arches and towers. When given positive and negative examples of each type of structure, his program would generalize the graphs to derive a definitional network for classifying all the types that were considered.
2. Haas and Hendrix (1983) developed the NanoKlaus system that would learn definitional networks by being told. Unlike Winston's system, which required

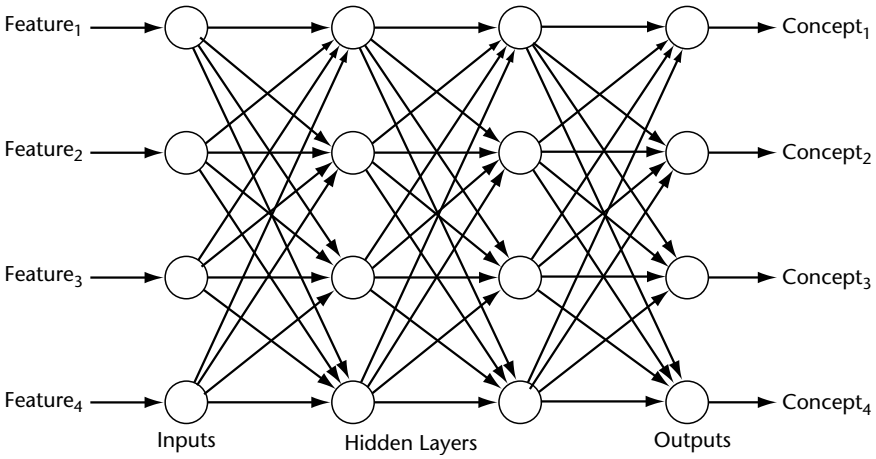


Figure 15. A neural net.

a set of examples that included all significant features, NanoKlaus would carry on a dialogue until the features it had been told were sufficient to distinguish all the specified types.

3. George Lendaris (1988) developed a two- stage learning system that combined conceptual graphs with neural networks. Both stages used a neural network with back propagation, but in the first stage, the inputs were features, and the outputs were concepts, as in Figure 15. In the second stage, each input represented a conceptual graph constructed from the concepts recognized by the first stage, and the outputs represented complex scenes described by those graphs. The two-stage system had a significantly reduced error rate and a faster learning rate than networks that matched features to scenes directly.
4. In *case-based reasoning* (Schank *et al*., 1994), the learning system uses rote memory to store various cases, such as medical diagnoses or legal disputes. For each case, it stores associated information, such as the prescribed treatment and its outcome or the legal argument and the court judgment. When a new case is encountered, the system finds those cases that are most similar to the new one and retrieves the outcome. Crucial requirements for the success of case-based reasoning are good similarity measures and efficient ways of searching for similar cases. To organize the search and evaluate similarity, the learning system must use restructuring to find common patterns in the individual cases and use those patterns as the keys for indexing the database.
5. Basili *et al*. (1996) developed methods of learning semantic patterns from a corpus of natural-language text. They started with a syntactic parser supplemented with a lexicon that had a limited amount of semantic information about the lexical patterns expected for each word. Then they used the parser to analyze the corpus and derive more detailed networks that represented the semantic patterns that occurred in the text. The system generalized those patterns to hypothesize better definitions of the lexical semantics for the words, which a linguist would verify before adding them to the lexicon. The system could then use the revised lexicon to reparse the corpus and further refine the definitions.
6. Robert Levinson (1996) developed a general system for learning to play board games such as chess or checkers. For each kind of game, the system was given the rules for making legal moves, but no further information about which moves are good or bad and no information about how to determine whether a game is won or lost. During the learning phase, the system would play games against a tutor (usually another program that plays very well, such as Gnu Chess). At the end of each game, the tutor would inform the system of a win, loss, or draw. For the learning phase, Levinson used a combination of rote learning as in case-based reasoning; restructuring to derive significant generalizations; a similarity measure based on the generalizations; and a method of back propagation to estimate the value of any case that occurred in a game. For playing chess, the cases were board positions represented as graphs. Every position that occurred in a game was stored in a generalization hierarchy, such as those used in definitional networks. At the end of each game, the system used back propagation to adjust the estimated values of each position that led to the win, loss, or draw. When playing a game, the system would examine all legal moves from a given position, search for similar positions in the hierarchy, and choose the move that led to a position whose closest match had the best predicted value.

These examples don't exhaust all the ways of using restructuring, but they illustrate its potential for learning sophisticated kinds of knowledge.

HYBRID NETWORKS

Many computer systems are hybrids, such as a combination of a database system for storing data, a graphics package for controlling the user interface, and a programming language for detailed computation. For knowledge representation, the Krypton system (Brachman *et al*., 1983) was a hybrid of a definitional network based on KL-ONE with an expert system that used a linear notation for asserting rules and facts. By the criteria used for calling Krypton a hybrid, most object-oriented (O-O) programming languages could be considered hybrids: the C++ language, for example, is a hybrid of the procedural language C with a definitional language for defining types or *classes*. Systems are usually called hybrids if their component languages have different syntaxes. Conceptual graphs, for example, include a definitional component for defining types and an assertional component that uses the types in graphs that assert propositions. But CGs are not usually considered hybrids because the syntax of the definitional component is the same as the syntax of the assertional component.

The most widely used hybrid of multiple network notations is the *Unified Modeling Language* (UML). Although UML is not usually called a semantic network, its notations can be classified according to the categories of semantic networks discussed in this article:

- Central to UML is a definitional network for defining object types. It includes the basic features of the Tree of Porphyry shown in Figure 1: type–subtype links, type–instance links, attributes that serve as differentiae for distinguishing a type from its supertype, and the inheritance of attributes from supertype to subtype.

- UML includes two kinds of executable networks that can be considered special cases of Petri nets: *statecharts*, which are special cases of Petri nets that do not support parallelism; and *activity diagrams*, which are almost identical to Petri nets, except that they do not use tokens to fire the transitions.
- The other networks in the UML family can be considered versions of relational graphs that are specialized for representing metalevel information. They include, for example, a version of entity-relation diagrams which are relational graphs designed for expressing the cardinality constraints and parameter types of relations.
- The most general of all the UML notations is a linear notation called the *Object Constraint Language* (OCL). It is a version of first-order logic with a notation that has syntactic features similar to some of the O-O programming languages. As an example, the following OCL statement says that all parameters of an entity have unique names:

```
self.parameter -> forAll (p1, p2 |
  p1.name = p2.name implies p1 = p2).
```

In OCL, *self* refers to the current entity being defined, and the names of functions are written after the entity to which they apply. In predicate calculus, the order would be interchanged: p1.name would be written name(p_1). Following is a translation of the OCL statement to predicate calculus with the symbol #self representing the current entity:

$$
\begin{aligned}
&(\forall p_1)\ (\forall p_2) \\
&\quad (\ (p_1 \in \texttt{parameter (\#self)} \\
&\quad \wedge\ p_2 \in \texttt{parameter (\#self)} \\
&\quad \wedge\ \texttt{name}\ (p_1) = \texttt{name}\ (p_2)\ ) \\
&\qquad \supset p_1 = p_2)
\end{aligned}
$$

This formula says that for every p_1 and p_2, if p_1 is a parameter of self and p_2 is a parameter of self and the name of p_1 is equal to the name of p_2, then p_1 is equal to p_2.

UML has been criticized for its lack of a formal definition, which has resulted in inconsistencies between its various notations. Work is currently underway to redesign UML with formal definitions for all the notations. One approach would be to extend OCL with sufficient features to define all the other notations. Another possibility would be to design a propositional semantic network that was equivalent to full first-order logic plus metalevel extensions, and use it to define everything in UML.

GRAPHIC AND LINEAR NOTATIONS

Graph notations and linear notations can express logically equivalent information, but with different syntactic conventions. The relational graph in Figure 4 and its translation to a formula in predicate calculus illustrate the differences between the two kinds of notations:

1. Both notations have seven occurrences of relation names, such as *isStagirite* or *teaches*.
2. They both have three occurrences of existential quantifiers, represented by three branching lines in the graph and by ($\exists x$), ($\exists y$), and ($\exists z$) in the formula.
3. The major difference lies in the way the connections from the quantifiers to the relations are shown: each line is directly connected to the relations, but 13 occurrences of the variables x, y, and z are scattered throughout the formula.

The chief advantage of graph notations is the ability to show direct connections. Linear notations must rely on repeated occurrences of variables or names to show the same connections.

As another example, Petri nets, which are usually expressed in a graphic notation, are formally equivalent to a notation called *linear logic*. Although Petri nets and linear logic were independently developed by different researchers for different purposes, a commonly used version of Petri nets happens to be isomorphic to a commonly used version of linear logic. Following is a translation of the Petri net of Figure 14 to that version of linear logic:

```
BusStops:
  BusArriving ⇒ BusWaiting

OnePersonGetsOnBus:
  PersonWaiting & BusWaiting ⇒
  PersonOnBus & BusWaiting

BusStarts:
  BusWaiting ⇒ BusLeaving

InitialAssertions:
  PersonWaiting. PersonWaiting.
  PersonWaiting. BusArriving.
```

Each arrow ($\Rightarrow$) in this example represents one of the transitions in the Petri net. The feature of linear logic that distinguishes it from classical first-order logic is the treatment of implication. For comparison, following is an application of modus ponens in classical FOL, its replacement in linear logic, and the rule for firing a transition in Petri nets:

- *Classical FOL*. Given propositions p and q and an implication $p \wedge q \supset r \wedge s$, conclude r and s. Everything that was previously true remains true.
- *Linear logic*. Given propositions p and q and an implication $p \wedge q \Rightarrow r \wedge s$, conclude r and s and retract the truth of p and q.
- *Petri nets*. Given tokens in the places p and q and a transition $p, q \rightarrow r, s$, add one token to each of the places

r and *s* and erase one token from each of the places *p* and *q*.

When the presence of a token in a place of a Petri net is interpreted as meaning the truth of the corresponding proposition, the rule for firing a transition is equivalent to using an implication in linear logic. Therefore, any collection of implications and assertions in linear logic can be represented by a Petri net, and any proof in linear logic corresponds to an execution of the Petri net.

One of the major arguments for graphic notations is human readability, but proponents of linear notations often argue that their notations are also highly readable. Each rule in linear logic, for example, is quite readable, but the relationships between rules are not as immediately readable as the direct connections in the Petri net.

As an example, consider the proposition Bus Waiting, which is represented by a single place in the Petri net, which has two inputs and two outputs. That fact can be seen immediately from Figure 14, but a reader would have to search through all of the rules in the linear logic example to verify that the name *BusWaiting* occurs four times. As examples become larger, any notation becomes more difficult to read, but the graphic notations still have an advantage over linear notations. Petri nets have been implemented with many thousands of nodes, but it is always possible to look at any node and see immediately how many inputs and outputs it has and where they are linked.

Besides readability, graphic notations often have heuristic value in helping human readers (either students or researchers) to discover patterns that would be difficult or impossible to see in the linear form. The reason why Peirce called his existential graphs 'the logic of the future' was not so much their readability as the direct insight they provided into the structure of proofs. With EGs, Peirce invented the simplest and most elegant rules of inference ever developed for any version of logic. Peirce's rules, which he discovered in 1897, are a simplification and generalization of the rules of *natural deduction* that Gentzen (1935) re-invented many years later. Even today, Peirce's rules lead to insights that have eluded logicians for many years.

References

Basili R, Pazienza MT and Velardi P (1996) An empirical symbolic approach to natural language processing. *Artificial Intelligence* **85**: 59–99.

Brachman RJ (1979) On the epistemological status of semantic networks. In: Findler NV (ed.) *Associativze Networks: Representation and Use of Knowledge by Computers*, pp. 3–50. New York, NY: Academic Press.

Brachman RJ, Fikes RE and Levesque HJ (1983) Krypton: a functional approach to knowledge representation. *IEEE Computer* **16**(10): 67–73.

Ceccato S (1961) *Linguistic Analysis and Programming for Mechanical Translation*. New York, NY: Gordon & Breach.

Fahlman SE (1979) *NETL: A System for Representing and Using Real-World Knowledge*. Cambridge, MA: MIT Press.

Fillmore CJ (1968) The case for case. In: Bach E and Harms RT (eds) *Universals in Linguistic Theory*, pp. 1–88. New York, NY: Holt, Rinehart & Winston.

Frege G (1879) *Begriffsschrift*, In: J van Heijenoort (ed.) *From Frege to Gödel*, pp. 1–82. Cambridge, MA: Harvard University Press.

Gentzen G (1935) *Untersuchungen über das logische Schließen*, translated by Szabo ME. In: Szabo ME (ed.) *The Collected Papers of Gerhard Gentzen*, pp. 68–131. Amsterdam, Netherlands: North-Holland.

Haas N and Hendrix GG (1983) Learning by being told. In: Michalski RS, Carbonell JG and Mitchell TM (eds) *Machine Learning*, pp. 405–427. Palo Alto, CA: Tioga Publishing Co.

Hendler JA (1992) Massively-parallel marker-passing in semantic networks. In: Lehmann F (ed.) *Semantic Networks in Artificial Intelligence*, pp. 277–291. Oxford, UK: Pergamon Press.

Hendrix GG (1979) Encoding knowledge in partitioned networks. In: Findler NV (ed.) *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 51–92. New York, NY: Academic Press.

Kamp H and Reyle U (1993) *From Discourse to Logic*. Dordrecht, Netherlands: Kluwer.

Lendaris GG (1988) Neural networks, potential assistants to knowledge engineers. *Heuristics* **1**.

Levesque H and Mylopoulos J (1979) A procedural semantics for semantic networks. In: Findler NV (ed.) *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 93–120. New York, NY: Academic Press.

Levinson RA (1996) General game-playing and reinforcement learning. *Computational Intelligence* **12**(1): 155–176.

Masterman M (1961) Semantic message detection for machine translation, using an interlingua. *International Conference on Machine Translation of Languages and Applied Language Analysis*, pp. 438–475. London, UK: HMSO.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Peirce CS (1880) On the algebra of logic. *American Journal of Mathematics* **3**: 15–57.

Peirce CS (1885) On the algebra of logic. *American Journal of Mathematics* **7**: 180–202.

Rational Software (1997) *UML Semantics*. [http://www.rational.com/media/uml/resources/media/ad970804_UML11_Semantics2.pdf.]

Schank RC (ed.) (1975) *Conceptual Information Processing*. Amsterdam, Netherlands: North-Holland.
Schank RC, Kass A and Riesbeck CK (1994) *Inside Case-Based Explanation*. Hillsdale, NJ: Lawrence Erlbaum.
Selz O (1913) *Über die Gesetze des geordneten Denkverlaufs*. Stuttgart, Germany: Spemann.
Selz O (1922) *Zur Psychologie des produktiven Denkens und des Irrtums*. Bonn, Germany: Friedrich Cohen.
Shapiro SC (1971) A net structure for semantic information storage, deduction and retrieval. Proceedings IJCAI-71.
Shapiro SC (1979) The SNePS semantic network processing system. In: Findler NV (ed.) *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 263–315. New York, NY: Academic Press.
Sowa JF (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.
Sowa JF (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole Publishing Co.
Tesnière L (1959) *Éléments de Syntaxe Structurale*, 2nd edn. Paris, France: Librairie C. Klincksieck.
Winston PH (1975) Learning structural descriptions from examples. In: Winston PH (ed.) *The Psychology of Computer Vision*, pp. 157–209. New York, NY: McGraw-Hill.
Woods WA (1975) What's in a link: foundations for semantic networks. In: Bobrow DG and Collins A (eds) *Representation and Understanding*, pp. 35–82. New York, NY: Academic Press.

Further Reading

Findler NV (ed.) (1979) *Associative Networks: Representation and Use of Knowledge by Computers*. New York, NY: Academic Press.
Lehmann F (ed.) (1992) *Semantic Networks in Artificial Intelligence*. Oxford, UK: Pergamon Press.
Minsky M (ed.) (1968) *Semantic Information Processing*. Cambridge, MA: MIT Press.
Sowa JF (ed.) (1991) *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann.

SHRDLU

Introductory article

Nigel Ward, University of Tokyo, Tokyo, Japan

CONTENTS

SHRDLU was an integrated artificial intelligence system which could make plans and carry on simple conversations about a set of blocks on a table.

INTRODUCTION

SHRDLU, created by Terry Winograd, was a computer program that could understand instructions and carry on conversations about a world consisting of some toy blocks on a table. For example:

Person: '*Pick up a big red block.*'
Computer: 'OK.' (In the display, SHRDLU's animated robot hand does this.)
Person: '*Grasp the pyramid.*'
Computer: 'I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.'
Person: '*Find a block which is taller than the one you are holding and put it into the box.*'
Computer: 'BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.'
Computer: 'OK.' (Does it.)
Person: '*What does the box contain?*'
Computer: 'THE BLUE PYRAMID AND THE BLUE BOX.'
...

SHRDLU was perhaps the first integrated artificial intelligence (AI) system. It incorporated modules for parsing, planning, question-answering, dialog management, etc., which worked together to perform a complete task.

PLANNING

SHRDLU controlled a simulated robot hand with which to arrange the blocks as the user requested. This was not trivial, since the robot could directly perform only three actions: `MOVETO` (a location), `GRASP` (the block currently under the hand) and `UNGRASP` (the block currently in the hand). Thus SHRDLU, which accepted high-level commands, such as 'put a red block on the green block', had to discover sequences of basic actions that would achieve its goals. It did this with the 'Planner' program.

The key idea, as in many planning systems, was a technique now called backward chaining or means-ends analysis. For example, if the goal was to have 'a red block on the green block', the Planner would first check to see if this was true already, and if not, it would try to find a way to achieve it. To find a way, it would search through the various operations in its knowledge base. In this example, this would yield `UNGRASP` as a possible way to satisfy the goal, provided that three preconditions held: (1) the green block had nothing on top; (2) the hand was holding a red block; and (3) the hand was over the green block. The Planner would then check to see if these preconditions were true, and if not, develop sub-plans to achieve each precondition. For example, if the green block had something on top, the Planner would proceed to find an operation or plan to achieve the 'nothing-on-green' goal. The concatenation of all the operations found in this process gave the sequences of actions needed to achieve the final goal.

SHRDLU also kept track of these chains of reasoning, so that if the user later asked, for example, 'Why did you pick up the brown pyramid?', SHRDLU could give an answer like 'SO THAT THE GREEN BLOCK WOULD BE CLEAR SO I COULD PUT THE RED BLOCK ON IT'.

SHRDLU was also able to backtrack when a partial plan turned out to be infeasible. That is, it was able partially to undo some line of reasoning, and start looking for an alternative sub-plan without having to start again from scratch.

PROCEDURAL SEMANTICS

Planner used a special internal representation. Thus, 'a red block in the box' would be represented as something like

```
(THGOAL (#IS ?X #BLOCK)) (THGOAL
(#COLOR $?X #RED)) (THGOAL (#IN $?X
:BOX))
```

With some familiarity with the conventions of logic, and a little imagination, it should be possible for an intelligent reader to see what this means. However, for this to be meaningful for a program requires something more, Winograd noted. SHRDLU needed to be able to reason about the basic concepts such as #RED and #IN. At the time, the representation of reasoning as theorem-proving, in which formal procedures manipulated logical symbols without using any knowledge of what the symbols 'meant', was prevalent. Winograd, however, argued for a 'procedural' view of semantics. In a vivid illustration of this approach to meaning, Winograd showed how the concept CLEARTOP, corresponding to the English phrase 'clear off', can be expressed as a procedure:

1. Start.
2. Does X support an object Y? If not, go to 5.
3. Move Y off X.
4. Go to 2.
5. Assert that X is CLEARTOP.

Similarly, the meaning of 'a red cube which supports a pyramid' can be described as a procedure for looping through all pairs of blocks on the table while checking their dimensions, colors, and relations to each other. This viewpoint, called 'procedural semantics', sparked excitement at the time: it seemed that this fresh perspective, inspired by practical programing considerations, would allow AI to resolve or bypass the messy problems that had troubled philosophers of meaning for centuries. In practice, however, most of the procedural knowledge in SHRDLU was represented as simple declarative facts, goals, or operators, and a sharp distinction between procedural and declarative approaches turned out to be both elusive and unimportant.

Building SHRDLU involved a lot of clever programming, not just the implementation of a mathematical or psychological theory. For this reason, SHRDLU is a landmark 'hacker' AI system. Even today, a lecture about SHRDLU is a common way to excite and inspire computer science undergraduates. However, the 'hacker' approach tends to lead to messy programs and confused models, and modern AI researchers are generally more careful to distinguish between the systems-building and the theory-building aspects of AI.

ANALYZING THE INPUT

Since SHRDLU's inputs were English sentences, it first had to parse them: to work out where the noun phrases began and ended, which words went with with which, and so on.

In designing the language analysis model, Winograd looked beyond purely practical considerations. Reasoning that in a psychologically realistic grammar, syntax and semantics were interdependent, he chose to base SHRDLU's grammar on systemic grammar. Winograd thus denied the claim, common in linguistics that it is possible or desirable to have an autonomous theory of syntax, formulated without regard to meaning.

SHRDLU also embodies a highly 'interactionist' view of the relation between syntactic and semantic processing. For example, consider the command 'Put the blue pyramid on the block in the box', which is ambiguous (as to whether 'on the block' describes the current location of the pyramid or its desired location). For this sentence, SHRDLU would first recognize that 'the blue pyramid' was a syntactically possible noun phrase, then immediately check to see if this interpretation made sense semantically. Because SHRDLU knew about the locations of all the blocks, it could determine whether there was a unique blue pyramid. If so, the parser would know that the rest of the sentence, 'on the block in the box', had to be a single prepositional phrase, specifying where to put the pyramid. If, on the other hand, there was more than one blue pyramid, the parser would deduce that 'on the block' was part of the noun phrase; then SHRDLU would perform another semantic check, to make sure that 'the blue pyramid on the block' was a valid description of a unique object. (*See* **Natural Language Processing, Disambiguation in**)

In this way, SHRDLU used syntactic and semantic information in concert to efficiently arrive at the correct interpretation of the input: by using semantic information early, it avoided wasting time considering interpretations which were syntactically possible but actually impossible given the configuration of the blocks. Thus SHRDLU demonstrated that an AI system with multiple modules could be tightly integrated, for the sake of both cognitive plausibility and efficiency. At the same time, SHRDLU was a fairly modular system, with the different kinds of data structures and processing algorithms neatly separated into components, in

accordance with general software engineering principles, making the system easier to develop, debug, and extend. Of course, there is something of a trade-off between integration and modularity, and dealing with this was a central concern of the new sub-field of 'AI architectures' for many years.

RESEARCH STRATEGY

In 1972 SHRDLU was an amazing achievement. It succeeded partly because it worked in a 'micro-world', that is a limited domain, consisting of some colored blocks on a table. It could only understand sentences about blocks, only make plans about blocks, and only ask questions about blocks. This may seem like a retreat from real-world complexities, but Winograd argued that it was a reasonable first step. After all, small children play with blocks before going on to more complex things. Similarly, AI should perhaps start by developing systems for various microworlds. Each system would be a simple but complete application. As time went on, more complex microworlds would be tackled, and the inventory of AI knowledge and techniques would grow, and ultimately all would be combined into a fully intelligent system. This approach is not universally accepted, and many have argued that AI for the real world cannot be developed by this strategy. Nevertheless, today all useful AI systems operate in microworlds. Indeed, the successful deployment of AI techniques is known to require the identification of some domain of expertise (such as engine diagnosis, chess playing, or answering payroll questions) that can be dealt with in isolation from the unbounded complexity of the 'real world'.

SHRDLU's microworld allowed direct mappings from lexical categories (in the input sentences) to meaning elements, and from syntactic structures to rules for combining meaning elements, as seen in table 1. Of course language is more complex than this in general: for example, '*destruction*' is a noun, but refers to an activity, not an object; and '*fake gun*' does not refer to the subset of guns that are fake.

DIALOG MANAGEMENT

Perhaps the most appealing aspect of SHRDLU was that it could hold a reasonably involved conversation and not lose the thread – something that is not trivial even for people. For example, in the dialog above SHRDLU deals with words like 'it' and 'that' which refer back to things it heard or did or said earlier. SHRDLU was also able to formulate follow-up questions if the input was ambiguous. As such, it was probably the first system to incorporate what is now called a 'dialog manager' and face up to the problems that arise when dealing with discourse, not just individual sentences.

IMPACT AND LIMITATIONS

SHRDLU was one of the first AI systems to perform a realistic task. Although moving colored blocks around on tables is not a commercially important activity, SHRDLU-like dialogs are very plausible when querying databases. Thus SHRDLU inspired, in the early 1970s, one of the periodic rushes to commercialize AI. There were several efforts to build 'natural language interfaces' to databases, so that people could get the information they wanted (for example, 'a list of all single women in zip code 94705 with cats' or 'a list of all people who earn more than their managers') without having to write programs or arcane database queries.

This turned out to be easy to achieve for certain hand-chosen inputs, but not for the sorts of things that real users input. To build a natural language interface for a new domain, such as a shoe shop inventory database, would require months of effort tuning the lexicon, grammar and code, and even then users tended to lapse into a narrow subset of English after discovering that fragmentary, idiomatic or unusual phrasings were seldom understood correctly. The nail in the coffin for natural language interfaces was the invention of graphical user interfaces (GUIs). Most users prefer GUIs, with on-screen views of the data and with the available commands laid out in buttons and menus, over natural language. Today, however, there is a revival of interest in natural language interfaces for spoken language interaction: for example, to get airline flight information over the telephone. Systems capable of carrying on rudimentary con-

Table 1. Some semantic interpretation rules exploited in SHRDLU

Language	*Meaning*
noun	object
adjective	a property of an object
'*the*'	expect to find a unique object
verb	action
'*it*'	the most recent topic of conversation
question mark	compute the answer
adjective–noun	a subset: all members of the set that have the specified property
verb–noun phrase	perform the action on the set of objects specified by the noun phrase

versations about such information are now becoming common.

SHRDLU was also one of the first AI systems to capture the popular imagination. Winograd's style of research, taking inspiration from linguistics and psychology and using low-level programming tricks, was in accordance with the intellectual climate of the time. The view he propounded – that programming considerations can tell us how a human reasoning system might work, and conversely, that knowledge of human cognitive processing could inform AI system development – implied that AI researchers, linguists and psychologists could achieve great things if they worked together. This excitement helped to stimulate the new field of cognitive science.

Further Reading

Boden MA (1987) *Artificial Intelligence and Natural Man*, 2nd edn. Cambridge, MA: MIT Press.

Jurafsky D and Martin JH (2000) *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Marr D (1977), Artificial intelligence – a personal view. *Artificial Intelligence* **9**: 37–48.

McDermott D (1981) Artificial intelligence meets natural stupidity. In: Haugeland J (ed.) *Mind Design: Philosophy, Psychology, Artificial Intelligence* pp. 143–160. Cambridge, MA: MIT Press. *SHRDLU*. http://hci.stanford.edu/cs147/examples/shrdlu/

Winograd T (1972) *Understanding Natural Language*. New York, NY: Academic Press. Also published in *Cognitive Psychology* **3**(1): 1–191.

Winograd T (1973) *A procedural model of language understanding* In: Schank R and Colby K (eds) *Computer Models of Thought and Language* pp. 152–186. New York, NY: WH Freeman [Reprinted in: Grosz BJ, Jones KS and Webber BL (eds) (1986) *Readings in Natural Language Processing*. Los Altos CA: Morgan Kaufmann.

Situated Cognition

Intermediate article

Kevin O'Connor, University of Wisconsin, Madison, Wisconsin, USA
Arthur M Glenberg, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Theories of situated cognition maintain that intelligent human action has evolved within and is shaped by and adapted to the specific forms of activity within which it occurs, and that cognition must therefore be understood and studied as an aspect of embodied practical activity.

INTRODUCTION

Situated cognition is an evolving interdisciplinary project to describe the relationship between embodied agents and the social, semiotic, and material settings of their activity. Contemporary theories of situated cognition have arisen and taken shape largely against the background of traditional cognitivist perspectives in psychology and artificial intelligence, which have tended to conceive of cognition in terms of abstract and disembodied reasoning processes taking place in the heads of individuals. While many of the topics addressed within situated cognition – such as the nature of knowledge and cognitive processes, and the relationship between knowledge and intelligent action – are shared with cognitivist approaches, the ways in which these topics are understood and studied diverge fundamentally from cognitivism. Theories of situated cognition maintain that intelligent human action has evolved within and is shaped by and adapted to the specific forms of activity within which it occurs, and that cognition must therefore be understood and studied as an aspect of embodied practical activity. Situated cognition is not just a type of cognition that can be distinguished from 'non-situated' cognition; rather, theories of situated cognition claim that human cognition is inherently situated, and that attempting to study cognition apart from the specific settings of activity within which it takes place destroys the phenomenon being studied.

TRADITIONAL COGNITIVISM AND A CARTESIAN SEPARATION FROM THE WORLD

Traditional cognitivist approaches in psychology and artificial intelligence are ultimately based on Cartesian ideas about, on one hand, an objective external world whose character is determinate and independent of human knowledge and concerns, and on the other hand, the human mind as a rational instrument that is able to discover and use the principles according to which the external world is structured. Cognitivism has developed these tenets into a number of assumptions about the nature of the mind and its relationship to action. (*See* **Descartes, René**; **Mind–Body Problem**)

First, the mind and the world are viewed as fundamentally separate from and impenetrable to one another. The world is taken to have an objective and stable character, and the mind is understood as capable of knowing the world by representing it in explicit and formal symbolic terms.

Second, according to cognitivism, this explicit and formal knowledge about the world underlies and is necessary for action; thus, action can only be properly understood in terms of symbolic mental representations of the world and rules for manipulating them. (*See* **Symbol Systems**)

Third, mental representations are taken to involve abstraction from the particularities of specific situations: the more abstract the representation is, the more generally applicable it is across situations; thus, greater abstraction is a mark of greater potential for intelligent action.

Fourth, the proper level of explanation for cognition involves descriptions of the processing of abstract symbols. Thus, the details of the realization of symbols in the brain are taken to be irrelevant to cognitive explanation: it is on this basis that

computers, despite their physical differences from the human brain, are understood to provide a potential model for human intelligence. Moreover, the body is understood not to be involved in symbol processing in any fundamental way; its role is limited to one of providing input to the cognitive system, through perceptual processes, and output, through motor actions.

The cognitivist analysis of the mind and its relation to action is evident in its view of action as the outcome of a process of problem-solving (e.g. Newell and Simon, 1972). According to this view, action in the world results from a process involving several discrete steps. First, a situation in the external world is symbolically represented in the mind as a problem to be solved. The mind then performs rule-based operations on the problem to produce a different symbolic representation – a solution to the problem. This solution then produces some action in the world. Note that the relevant context in the problem-solving process is not the world, but a mental representation of it – the problem – which is, in principle, fully specifiable in terms of formal symbolic structures. Thus, action in the world is caused by underlying cognitive processes. The proper focus of study, then, is not action in the world but rather the underlying knowledge, the mental representations and rules for manipulating them, that cause action. (*See* **Problem Solving**; **Newell, Allen**; **Simon, Herbert A.**)

Influenced by this analysis of action, cognitivists have tended to focus on, and indeed have been quite successful at modeling, problem-solving in situations that lend themselves to description in terms of the manipulation of abstract symbols. Two common types of problem used in cognitivist research can serve to illustrate this. One type of problem, of which the well-known 'tower of Hanoi' is an example, is intended to provide a window to purportedly general cognitive processes by examining problem-solving in simple abstract situations, which often bear little resemblance to the kinds of activities people pursue in much of their everyday life. Another type of problem is drawn from activities, such as chess, that are chosen because they are suitable for producing formalizable cognitive tasks with clear criteria for determining appropriate solutions.

SITUATED ACTIVITY THEORY

Beginning around the 1980s, a number of theoretical and practical difficulties with the cognitivist paradigm, along with empirical observations that were anomalous from the cognitivist point of view, led researchers to explore alternatives to the traditional emphasis on cognition as the purely mental manipulation of abstract symbolic representations. Work in this new perspective, a loose convergence of a variety of anti-Cartesian traditions, placed primary emphasis upon activity situated in its everyday contexts.

Some of this work pointed to a theoretical problem associated with the cognitivist assumption that symbols are manipulated solely on the basis of their abstract form, and suggested that cognitivism provides no account of how symbols ultimately come to take on meaning in human action and experience. The abstract symbols (e.g. nodes in a semantic network) are equivalent to the zeros and ones in a computer. What differentiates one concept from another is solely the set of relations among the symbols corresponding to the two concepts. Thus, one symbol is defined solely in terms of other symbols. In contrast, Harnad (1990), Searle (1980) and others have argued that meaning (e.g. what concepts are about) cannot, in principle, arise from such a system. For example, Harnad asks us to consider how we could understand a sentence written in a foreign language using only a dictionary for that language. We would look up in the dictionary the first word in the sentence (the first abstract symbol) and see that it is defined in terms of relations to other abstract symbols. On looking up the first word in the definition, we find only more abstract symbols. Clearly, we will never be able to discover the meaning of the first word in the sentence, let alone the meaning of the sentence. Yet abstract symbol theory claims just this: that meaning arises from the relations among the symbols. If abstract symbols are to be meaningful, they must be grounded in the world through perception and action. But Hilary Putnam has formally proved that starting just from a set of abstract symbols, there is an infinite number of exactly corresponding objects and relations in the world, so it is impossible to ground the abstract symbols under these conditions (Lakoff, 1987, pp. 229–259). (*See* **Symbol-grounding Problem**; **Meaning**; **Chinese Room Argument, The**)

Others have suggested that the cognitivist requirement that the world be represented in explicit symbolic terms presents a further theoretical difficulty. Among the major practical limitations of the cognitivist tradition is its difficulty in modeling common-sense knowledge. While cognitivism has been very successful at modeling certain kinds of highly formalized cognitive tasks, modeling some very basic forms of everyday human intelligence has proved to be much more problematic.

Consider, for example, a pedestrian deciding when to cross a street. There is a simple rule: wait for a green crossing light. But this rule needs to be modified to fit almost every conceivable situation: if you are in a great hurry, then cross against the light, except if traffic is extremely heavy, in which case wait, except if it is a real emergency, in which case cross anyway, except if traffic is extremely heavy, in which case wait, except if the traffic is moving slowly, in which case cross, except if you have a twisted ankle, and so on. Some researchers have argued that this practical difficulty is a reflection of the impossibility in principle of explicitly modeling common-sense knowledge (e.g. Dreyfus, 1979). Dreyfus, drawing on phenomenology and the later philosophy of Wittgenstein, takes issue with the cognitivist assumption that explicit symbolic representation of the world is necessary to produce action, and argues that action in the everyday world depends on an enormous body of background knowledge, combined with perception of affordances (Gibson, 1979), that is, how a particular type of body constrains actions occurring in a particular situation and structured in part by particular goals. Given that the affordances for action will change with momentary changes in bodily abilities (such as adrenaline surges) and with changes in the goals, no set of rules can be constructed to guide actions in situations treated as members of abstract categories such as 'street-crossing situation'. Since all situations are unique, abstract rules cannot be constructed to ensure the adequacy of action in the situation in general. In this view, then, even if the explicit models offered by cognitivism are able to provide insight into certain aspects of cognition, these insights are strictly limited to formalizable aspects of cognition and do not provide us with a plausible general model of human cognition. (*See* **Phenomenology**; **Perception: The Ecological Approach**; **Wittgenstein, Ludwig**; **Gibson, James J.**)

These weaknesses and limitations of cognitivism, along with others to be discussed below, have led some researchers to conclude that the complexity of the relationship between the mind and its environment is such that considering them separately involves unjustifiable oversimplification, and that primary analytical emphasis should instead be placed on cognition as it occurs in situated activity. While not all researchers would commit themselves equally to each of the following points, theories of situated cognition have proposed several possible alternative assumptions about the nature of cognition and its relationship to action:

- The mind and the world are fundamentally interconnected with one another. The character of the world is not given, but is instead dependent upon construal by agents in the course of activity. These construals, furthermore, are both enabled and constrained by the structure of the brain, by the body, and by their occurrence in specific physical and social contexts.
- Representations of the world are not abstract, disembodied, and detached from the world, but arise in the course of situated activity.
- Knowledge is located in the evolving relationships between people and artifacts in culturally evolved systems of activity.
- Cognition is mediated by artifacts and distributed across people in systems of activity.
- Cognition is opportunistic and improvized. Successful participation in everyday activity, rather than engagement in tasks based on formal logic, is the primary phenomenon to be explained.

BODY, ENVIRONMENT AND SOCIAL CONTEXT

This shift in focus from abstract and logical thinking to successful participation in everyday activity has led to a corresponding shift of attention towards phenomena that have been considered by cognitivists to be of marginal importance. For example, whereas cognitivism treats sensorimotor processes as peripheral to and separate from the cognitive system, theorists of situated cognition have begun, in various ways, seriously to study the consequences of physical embodiment for the nature of cognitive processing. (*See* **Embodiment**)

One area of research has focused on individual cognition, and has attempted to identify universal aspects of cognition that are grounded in the nature of the human body and its capacity to act in a physical environment. This work argues that cognitive processes, even symbolic processes, are not abstract and are intimately linked to the embodied nature of the organisms carrying out those processes. Other researchers emphasize the fact that physical bodies are a part of the physical world in which they act, and claim that this locatedness requires alternative ways of thinking about cognition.

Recent research in cognitive psychology and cognitive linguistics, while retaining a primary focus on the cognitive processes of individuals, has attempted to understand how cognition has evolved and how it functions in the service of embodied organisms who act in a physical environment. In this work, symbolic mental representations have remained at the center of studies of cognition. However, instead of being viewed as abstract or

amodal (that is, as bearing only an arbitrary connection to experience in the world) symbols are viewed as inherently perceptual in that they modally or analogically replicate some aspect of experience. Thus, this work takes representational units to develop within activity and to remain linked to activity. Lakoff (1987) suggests that even our most abstract concepts – logical rules, such as '*p* or *q* but not both' – arise from experience in the world. For example, our experience with containers and containment provides just the structure for the rule: something can be in the container (*p*) or not in the container (*q*), but not both. Barsalou (1999) has demonstrated that the representations of many concepts have components related to the way the corresponding objects are perceived and acted upon, rather than consisting of amodal abstract symbols. (*See* **Cognitive Linguistics**)

Glenberg and Robertson (2000) demonstrate that understanding simple sentences requires a consideration of how human actions can combine. Thus, people find sentences such as 'Bill stood on the tractor to paint the top of the barn wall' much more sensible than sentences such as 'Bill stood on the hammer to paint the top of the barn wall'. From the point of view of abstract symbols, the two sentences are virtually identical: tractors and hammers are both inanimate concrete nouns, both are tools, and both are unrelated to painting in past experience. The difference is that knowledge of how people can interact with objects – even when used in atypical and never-before-encountered ways – allows people to determine which sentence is sensible and which is nonsense.

Recent work in artificial intelligence and robotics has contributed in a somewhat different way to an understanding of the importance of embodiment for cognition (e.g. Agre and Chapman, 1987; Brooks, 1991). This work has argued that situated activity by embodied agents in a complex physical environment is a nontrivial accomplishment, and one that must be viewed as central to intelligence. However, it has not been modeled successfully by approaches that are grounded in traditional artificial intelligence. Thus, some researchers have attempted to design agents that are able to respond effectively, in real time, to the demands of complex, unpredictable, and changing environmental conditions. An important strategy has been to replace the traditional cognitivist focus on perception, modeling, planning and action as discrete steps in a problem-solving process, by proposing a tight coupling of sensing and action. These embodied agents exploit features of the physical environment and their own relationship to it so as to avoid the need to form a complete and explicit representation of the environment before acting. Instead, they opportunistically exploit aspects of the structure of the environment that offer clues about how to proceed and when a task has been completed. (*See* **Situated Robotics**)

Similar opportunism and context-embeddedness has been demonstrated in research, largely inspired by the cultural–historical tradition begun by L. S. Vygotsky, that examines humans engaged in a wide variety of practical tasks. This work has led Sylvia Scribner to argue that the environment plays a 'constitutive role' in cognition. Scribner (1997, p. 329) argues that 'skilled practical thinking incorporates features of the task environment (people, things, information) into the problem-solving system. It is as valid to describe the environment as part of the problem-solving system as it is to observe that problem-solving occurs "in" the environment.' (*See* **Cultural Psychology**; **Vygotsky, Lev**)

This work challenges a number of ideas fostered by cognitivist models of problem-solving, including: that problems are given to the problem-solver in a complete form; that problems of the same logical class will be solved by the same sequence of operations on all occasions; and that learning involves increasing independence from the concrete particularities of a context. Instead, studies of problem-solving in a wide range of everyday activities have shown that what appear, in theory, to be formally identical problems are, in practice, flexibly formulated and solved by the problem solver according to the contingencies of the environment. For example, Scribner has shown that dairy workers reformulate abstract computational problems into problems that depend on the concrete physical array of the product they are working with. She has shown, in addition, that these reformulations differ depending on the nature of the activity (e.g. 'filling orders' versus 'inventory'), on the specific point in the flow of activity, and on the values and goals of the problem-solver. Similar claims have been made regarding the skilled practical thinking of shoppers, cooks, bartenders, street vendors, and participants in a range of other activities.

This work indicates that, compared with novices, experts in a domain of activity rely more, not less, on environmental sources of information. This movement towards increasing contextualization of practical thinking not only contradicts cognitivist claims suggesting that cognition should become more powerful as it becomes less contextualized; it also suggests that cognitivist models of thinking as a purely mental activity will find analysis of many

practical thinking problems intractable (Scribner, 1997). It is also important to note, lest one conclude that the cognition of what Lave (1988) ironically called 'just plain folks' represents an inferior and deficient mode of thought, that the very same kinds of context-embeddedness and opportunism have been demonstrated in careful studies of the work of those who, to cognitivists, represent the height of rationality, such as scientists, mathematicians, engineers, and medical diagnosticians (e.g. Goodwin, 1995; Latour, 1987; Norman and Brooks, 1997).

In addition to emphasizing the role of the environment in cognition, work in situated cognition has emphasized that it is necessary to understand several ways in which the social context influences cognitive processing. One of these is closely related to the opportunism and flexibility of practical problem-solving. When Scribner's dairy workers transformed abstract computational problems into problems that were tied to the physical environment in which they were situated, they did so by using materials (e.g. milk crates) designed for noncognitive purposes (e.g. holding milk bottles) to mediate their formulation and solution of a cognitive task. This is an instance of a more general capacity of human beings to modify their environment by creating mediating artifacts. These artifacts, moreover, can transform subsequent possibilities for activity by embodying 'partial solutions to frequently encountered problems' (Hutchins, 1995, p. 374) of an individual or group, and can preserve these solutions to be passed on to subsequent generations. In this sense, use of a mediating artifact situates its user within the cultural tradition that developed, transmitted, and continues to maintain that artifact.

A second, and closely related, way in which the social context is important in cognitive processing is connected with the idea that individual mental functioning has its origins in social interaction. A great deal of research, much of it again building on the ideas of Vygotsky, has demonstrated that social interaction provides the conditions within which skilled practical thinking is mastered by individuals. According to this work, other actors provide the individual with support in performing cognitive tasks that he or she would be unable to perform alone. Through this kind of 'apprenticeship' (Rogoff, 1990), individuals develop mental structures and processes that can be traced to their origins in social interaction. This is not to claim that forms of individual cognition directly copy processes found in social interaction: individuals transform cognitive structures and processes in the course of internalization, and cognition is opportunistically adapted by agents to their specific circumstances of activity (Wertsch, 1998). (*See* **Culture and Cognitive Development**)

Finally, some cognitive tasks are more properly viewed as carried out by social groups than by individuals, so the cognitive processes of individuals engaged in such tasks must be viewed in terms of their relationship to the larger cognitive system.

These points about the importance of the social context for cognitive processing have been of central importance in recent work on cognitive artifacts and distributed cognition.

COGNITIVE ARTIFACTS AND DISTRIBUTED COGNITION

Vygotsky's primary concern was with semiotic artifacts, or signs, and the ways in which they mediate higher mental functioning in humans. D. A. Norman, drawing on this emphasis within the cultural–historical tradition, introduced the idea of 'cognitive artifacts'. For Norman (1991, p. 17), cognitive artifacts are 'those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance'. Examples of cognitive artifacts include language, inscriptional systems for representing language, maps, lists, and calculators.

A well-known example can illustrate the importance of cognitive artifacts in mediating cognitive processes (Rumelhart *et al.*, 1986; Wertsch, 1998). Multiplication of multi-digit numbers – for example, multiplying 343 by 822 – represents a highly abstract conceptual problem. However, such a problem can be transformed by first arranging the digits in a spatial array and then carrying out a sequence of concrete operations on the digits, as shown in Figure 1.

Here, the spatial array serves as a cognitive artifact that transforms an abstract and difficult

```
  343
  822
 ----
  686
 6860
274400
 ----
281946
```

Figure 1. An example of a cognitive artifact. An abstract and difficult problem (multiplying 343 by 822) is transformed into a relatively simple one by a process of arrangement of the digits and simple operations on them.

problem into a relatively simple one that most of us become very good at. It is important to note, however, that while an individual problem-solver must have knowledge of how to use this cognitive artifact, the knowledge required to solve this problem is not wholly a property of the individual. That is, much of the knowledge required for solving multi-digit multiplication problems is contained in the cognitive artifact, in the form of historically evolved ways of arranging and operating on digits. Moreover, while the cognitive processing involved in solving this problem does involve the manipulation of symbols, this symbol manipulation is not likely to be performed solely in the head, but will also depend on the use of materials in the world, such as paper and pencil. Indeed, as Wertsch (1998) points out, many individuals who are able easily to solve problems like this are unaware of the abstract principles involved in the cognitive artifact, and would in fact be unable to solve the problem if the use of the artifact were not possible – for example, without the use of paper and pencil in a noisy room.

Observations of this sort have led some theorists of situated cognition to propose that cognition is 'distributed – stretched over, not divided among – mind, body, activity, and culturally organized settings (which include other actors)' (Lave, 1988, p. 1). Theorists of distributed cognition argue that the study of cognition cannot involve only the study of what is in individual heads, but must extend 'beyond the skin' (Hutchins, 1995) to examine how individuals and groups of individuals use historically evolved artifacts in carrying out their activities. For example, Hutchins' study of the navigation team of a large naval vessel shows that, while there is a great deal of computation involved in the navigation of a large ship, this computation does not take place in the head of any individual, but in a coordination of the minds of different individuals with navigational artifacts, such as landmarks, maps, phone circuits, and organizational roles. The cognition of individuals often involves little more than what is involved in reading numbers or drawing lines. The symbol processing necessary for navigation is performed by the 'functional system' of a collection of individuals acting with cognitive artifacts.

Work in distributed cognition suggests that understanding how cognition is mediated by artifacts and realized in activity and social interaction leads ultimately to the study of how cognition is tied to and partially constitutive of cultural systems. Proponents of distributed cognition have called for a redefinition of the disciplinary boundaries of cognitive science, with a much more central role given to such fields as anthropology, sociology, and cultural psychology, whose methods and interests are well suited to the kinds of detailed 'cognitive ethnography' called for in the study of distributed activity systems.

References

Agre PE and Chapman D (1987) Pengi: an implementation of a theory of activity. *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pp. 268–272. Los Altos, CA: Morgan Kaufmann.

Barsalou LW (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* **22**: 577–660.

Brooks RA (1991) Intelligence without representation. *Artificial Intelligence* **47**: 139–159.

Dreyfus H (1979) *What Computers Can't Do: A Critique of Artificial Reason*. New York, NY: Harper and Row.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Glenberg AM and Robertson DA (2000) Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* **43**: 379–401.

Goodwin C (1995) Seeing in depth. *Social Studies of Science* **25**: 237–274.

Harnad S (1990) The symbol grounding problem. *Physica D* **42**: 335–346.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Lakoff G (1987) *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

Latour B (1987) *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.

Lave J (1988) *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. New York, NY: Cambridge University Press.

Newell A and Simon H (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norman DA (1991) Cognitive artifacts. In: Carroll JM (ed) *Designing Interaction: Psychology at the Human–Computer Interface*, pp. 17–38. New York, NY: Cambridge University Press.

Norman GR and Brooks LR (1997) The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education* **2**: 173–184.

Rogoff B (1990) *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford, UK: Oxford University Press.

Rumelhart DE, Smolensky P, McClelland JL and Hinton GE (1986) Schemata and sequential thought processes in PDP models. In: McClelland JL, Rumelhart DE and the PDP Research Group (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. II, 'Psychological and Biological Models', pp. 7–57. Cambridge, MA: MIT Press.

Scribner S (1997) Thinking in action: some characteristics of practical thought. In: Tobach E, Falmagne RJ, Parlee MB, Martin LMW and Kapelman AS (eds) *Mind and Social Practice: Selected Writings by Sylvia Scribner*, pp. 319–337. New York, NY: Cambridge University Press.

Searle JR (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**: 417–424.

Wertsch JV (1998) *Mind as Action*. New York, NY: Oxford University Press.

Further Reading

Agre PE (1996) *Computation and Human Experience*. New York, NY: Cambridge University Press.

Bowker GC, Star SL, Turner WL and Gasser L (eds) (1997) *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Mahwah, NJ: Erlbaum.

Clancey WJ (1997) *Situated Cognition: On Human Knowledge and Computer Representations*. New York, NY: Cambridge University Press.

Clark A (1997) *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Cole M (1996) *Cultural Psychology: A Once and Future Discipline*. Cambridge, MA: Harvard University Press.

Cole M, Engestrom Y and Vasquez O (eds) (1997) *Mind, Culture and Activity: Seminal Papers from the Laboratory of Comparative Human Cognition*. New York, NY: Cambridge University Press.

Gee JP (1992) *The Social Mind: Language, Ideology, and Social Practice*. New York, NY: Bergin and Garvey.

Greeno JG, Chi MTH, Clancey WJ and Elman J (eds) (1993) *Cognitive Science* **21**: 1–147. [Special issue on situated action.]

Kirshner D and Whitson JA (eds) (1997) *Situated Cognition: Social, Semiotic, and Psychological Perspectives*. Mahwah, NJ: Erlbaum.

Lave J and Wenger E (1991) *Situated Learning: Legitimate Peripheral Participation*. New York, NY: Cambridge University Press.

Nardi BA (ed.) (1996) *Context and Consciousness: Activity Theory and Human–Computer Interaction*. Cambridge, MA: MIT Press.

Rosenschein SJ and Kaelbling LP (1995) A situated view of representation and control. *Artificial Intelligence* **73**: 149–173.

Stanfield RA and Zwaan RA (2001) The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science* **12**: 153–156.

Suchman LA (1987) *Plans and Situated Actions: The Problem of Human Machine Communication*. New York, NY: Cambridge University Press.

Tomasello M (1999) *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Vygotsky LS (1978) *Mind in Society*. Cambridge, MA: Harvard University Press.

Winograd T and Flores F (1986) *Understanding Computers and Cognition: A New Foundation for Design*. Reading, MA: Addison-Wesley.

Situated Robotics

Intermediate article

Maja J Matarić, University of Southern California, Los Angeles, California, USA

CONTENTS

Situated robotics is the study of robots embedded in complex, often dynamically changing environments. The complexity of the robot control problem is directly related to how unpredictable and unstable the environment is, to how quickly the robot must react to it, and to how complex its task is.

INTRODUCTION

Robotics, like any concept that has grown and evolved over time, has eluded a single, unifying definition. What was once thought of as a replacement for repetitive, manual labor has grown into a large field that includes applications as diverse as automated car assembly, space exploration and robotic soccer. Although robotics includes teleoperation, in which the robot itself may be merely a remotely operated body, in most interesting cases the system exists in the physical world, senses its environment, makes autonomous decisions, and acts on them, again in the physical world, typically in ways involving movement. Situated robotics focuses specifically on robots that are embedded in complex, challenging, often dynamically changing environments. 'Situatedness' refers to existing in, and having one's behavior strongly affected by, such an environment. Examples of situated robots include autonomous robotic cars on the highway or on city streets (Pomerleau, 1989), teams of interacting mobile robots (Matarić, 1995), and a mobile robot in a museum full of people (Burgard *et al*., 2000). Examples of unsituated robots, which exist in fixed, unchanging environments, include assembly robots operating in highly structured, strongly predictable environments. The predictability and stability of the environment largely determines the complexity of the robot that must exist in it; situated robots present a significant challenge for the designer.

Embodiment is a concept related to situatedness. It refers to having a physical body and interacting with the environment through that body. Thus, embodiment is a form of situatedness: an agent operating within a body is situated within it, since the agent's actions are directly and strongly affected by it. Robots are embodied: they must possess a physical body in order to sense their environment, and act and move in it. Thus, in principle every robot is situated. But if the robot's body must exist in a complex, changing environment, the situatedness, and thus the control problem, are correspondingly complex.

TYPES OF ROBOT CONTROL

Robot control is the process of taking information about the environment, through the robot's sensors, processing it as necessary in order to make decisions about how to act, and then executing those actions in the environment. The complexity of the environment, i.e., the level of situatedness, clearly has a direct relation to the complexity of the control (which is directly related to the task of the robot): if the task requires the robot to react quickly yet intelligently in a dynamic, challenging environment, the control problem is very hard. If the robot need not respond quickly, the required complexity of control is reduced. The amount of time the robot has to respond, which is directly related to its level of situatedness and its task, influences what kind of controller the robot will need.

While there are infinitely many possible robot control programs, there is a finite and small set of fundamentally different classes of robot control methodologies, usually embodied in specific robot control architectures. The four fundamental classes of control program are: reactive control ('don't think, react'), deliberative control ('think, then act'), hybrid control ('think and act independently in parallel'), and behavior-based control ('think the way you act').

Each of these approaches has its strengths and weaknesses, and all play important and successful roles in certain problems and applications. Different approaches are suitable for different levels of

situatedness, natures of the task, and capabilities of the robot, in terms of both hardware and computation. Robot control involves the following unavoidable trade-offs:

Thinking is slow, but reaction must often be fast. (1)

Thinking allows looking ahead (planning) to avoid bad actions. But thinking too long can be dangerous (e.g., falling off a cliff, being run over). (2)

To think, the robot needs potentially a great deal of accurate information. Information must therefore actively be kept up to date. But the world keeps changing as the robot is thinking, so the longer it thinks, the more inaccurate its solutions. (3)

Some robots do not 'think' at all, but just execute preprogrammed reactions, while others think a lot and act very little. Most lie between these two extremes, and many use both thinking and reaction. Let us review each of the four major approaches to robot control, in turn.

Reactive Control

'Don't think, react'. Reactive control is a technique for tightly coupling sensory inputs and effector outputs, to allow the robot to respond very quickly to changing and unstructured environments (Brooks, 1986). Reactive control is often described as its biological equivalent: 'stimulus–response'. This is a powerful control method: many animals are largely reactive. Thus, this is a popular approach to situated robot control. Its limitations include the robot's inability to keep much information, form internal representations of the world (Brooks, 1991a), or learn over time. The trade-off is made in favor of fast reaction time and against complexity of reasoning. Formal analysis has shown that for environments and tasks that can be characterized *a priori*, reactive controllers can be highly powerful, and, if properly structured, capable of optimal performance in particular classes of problems (Schoppers, 1987; Agre and Chapman 1990). But in other types of environments and tasks, where internal models, memory, and learning are required, reactive control is not sufficient.

Deliberative Control

'Think, then act.' In deliberative control, the robot uses all of the available sensory information, and all of its internally stored knowledge, to reason about what actions to take next. The reasoning is typically in the form of planning, requiring a search of possible state–action sequences and their outcomes. Planning, a major component of artificial intelligence, is a computationally complex problem. The robot must construct and then evaluate potentially all possible plans until it finds one that will tell it how to reach the goal, solve the problem, or decide on a trajectory to execute. Planning requires the existence of an internal representation of the world, which allows the robot to look ahead into the future, to predict the outcomes of possible actions in various states, so as to generate plans. The internal model, therefore, must be kept accurate and up to date. When there is sufficient time to generate a plan, and the world model is accurate, this approach allows the robot to act strategically, selecting the best course of action for a given situation. However, being situated in a noisy, dynamic world usually makes this impossible. Thus, few situated robots are purely deliberative.

Hybrid Control

'Think and act independently in parallel.' Hybrid control combines the best aspects of reactive and deliberative control: it attempts to combine the real-time response of reactivity with the rationality and efficiency of deliberation. The control system contains both a reactive and a deliberative component, and these must interact in order to produce a coherent output. This is difficult: the reactive component deals with the robot's immediate needs, such as avoiding obstacles, and thus operates on a very short timescale and uses direct external sensory data and signals; while the deliberative component uses highly abstracted, symbolic, internal representations of the world, and operates on a longer timescale. As long as the outputs of the two components are not in conflict, the system requires no further coordination. However, the two parts of the system must interact if they are to benefit from each other. Thus, the reactive system must override the deliberative one if the world presents some unexpected and immediate challenge; and the deliberative component must inform the reactive one in order to guide the robot towards more efficient trajectories and goals. The interaction of the two parts of the system requires an intermediate component, whose construction is typically the greatest challenge of hybrid system design. Thus, hybrid systems are often called 'three-layer systems', consisting of the reactive, intermediate, and deliberative layers. A great deal of research has been

conducted on how to design these components and their interactions (Giralt *et al*., 1983; Firby, 1987; Arkin, 1989; Malcolm and Smithers, 1990; Connell, 1991; Gat, 1998).

Behavior-based Control

'Think the way you act.' Behavior-based control draws inspiration from biology, and tries to model how animals deal with their complex environments. The components of behavior-based systems are called behaviors: these are observable patterns of activity emerging from interactions between the robot and its environment. Such systems are constructed in a bottom-up fashion, starting with a set of survival behaviors, such as collision avoidance, which couple sensory inputs to robot actions. Behaviors are added that provide more complex capabilities, such as wall following, target chasing, exploration, and homing. New behaviors are introduced into the system incrementally, from the simple to the more complex, until their interaction results in the desired overall capabilities of the robot. Like hybrid systems, behavior-based systems may be organized in layers, but unlike hybrid systems, the layers do not differ from each other greatly in terms of the timescale and representation used. All the layers are encoded as behaviors, processes that quickly take inputs and send outputs to each other.

Behavior-based systems and reactive systems share some similar properties: both are built incrementally, from the bottom up, and consist of distributed modules. However, behavior-based systems are fundamentally more powerful, because they can store representations (Matarić, 1992), while reactive systems cannot. Representations in behavior-based systems are stored in a distributed fashion, so as to best match the underlying behavior structure that causes the robot to act. Thus if a robot needs to plan ahead, it does so in a network of communicating behaviors, rather than using a single centralized planner. If a robot needs to store a large map, the map is likely to be distributed over several behavior modules representing its components, like a network of landmarks (Matarić, 1990), so that reasoning about the map can be done in an active fashion, for example using message passing within the landmark network. Thus, the planning and reasoning components of the behavior-based system use the same mechanisms as the sensing and action-oriented behaviors, and so operate on a similar timescale and representation. In this sense, 'thinking' is organized in much the same way as 'acting'.

Because of their capability to embed representation and plan, behavior-based control systems are not an instance of 'behaviorism' as the term is used in psychology: behaviorist models of animal cognition involve no internal representations. Some argue that behavior-based systems are more difficult to design than hybrid systems, because the designer must directly take advantage of the dynamics of interactions rather than minimize interactions through traditional system modularity. However, as the field is maturing, expertise in complex system design is growing, and principled methods of distributed modularity are becoming available, along with behavior libraries. Much research has also been conducted in behavior-based robot control.

COMPARISON AND DISCUSSION

Behavior-based systems and hybrid systems have the same expressive and computational capabilities: both can store representations and look ahead. But they work in very different ways, and the two approaches have found different niches in mobile robotics problem and application domains. For example, hybrid systems dominate the domain of single robot control, unless the domain is so time-demanding that a reactive system must be used. Behavior-based systems dominate the domain of multiple-robot control, because the notion of collections of behaviors within the system scales well to collections of such robots, resulting in robust, adaptive group behavior.

In many ways, the amount of time the robot has (or does not have) determines what type of controller will be most appropriate. Reactive systems are the best choice for environments demanding very fast responses; this capability comes at the price of not being able to look into the past or the future. Reactive systems are also a popular choice in highly stochastic environments, and environments that can be properly characterized so as to be encoded in a reactive input–output mapping. Deliberative systems, on the other hand, are the best choice for domains that require a great deal of strategy and optimization, and in turn search and planning. Such domains, however, are not typical of situated robotics, but more so of scheduling, game playing, and system configuration, for instance. Hybrid systems are well suited to environments and tasks where internal models and planning can be employed, and the real-time demands are few, or sufficiently independent of the higher-level reasoning. Thus, these systems 'think while they act.' Behavior-based systems, in

contrast, are best suited to environments with significant dynamic changes, where fast response and adaptivity are necessary, but the ability to do some looking ahead and avoid past mistakes is required. Those capabilities are spread over the active behaviors, using active representations if necessary (Matarić, 1997). Thus, these systems 'think the way they act.'

We have largely treated the notion of 'situated robotics' here as a problem: the need for a robot to deal with a dynamic and challenging environment it is situated in. However, it has also come to mean a particular class of approaches to robot control, driven by the requirements of situatedness. These approaches are typically behavior-based, involving biologically-inspired, distributed, and scalable controllers that take advantage of a dynamic interaction with the environment rather than of explicit reasoning and planning. This body of work has included research and contributions in single-robot control for navigation (Connell, 1990; Matarić, 1990), models of biological systems ranging from sensors to drives to complete behavior patterns (Beer *et al*., 1990; Cliff, 1990; Maes, 1990; Webb, 1994; Blumberg, 1996), robot soccer (Asada *et al*., 1994; Werger, 1999; Asada *et al*., 1998), cooperative robotics (Matarić, 1995; Kube and Zhang, 1992; Krieger *et al*., 2000; Gerkey and Matarić, 2000), and humanoid robotics (Brooks and Stein, 1994; Scassellati, 2001; Matarić 2000). In all of these examples, the demands of being situated within a challenging environment while attempting to safely perform a task (ranging from survival, to achieving the goal, to winning a soccer match) present a set of challenges that require the robot controller to be real-time, adaptive, and robust.

The ability to improve performance over time, in the context of a changing and dynamic environment, is also an important area of research in situated robotics. Unlike in classical learning, where the goal is to optimize performance over a typically long period of time, in situated learning the aim is to adapt relatively quickly, achieving greater efficiency in the light of uncertainty. Models from biology are often considered, and reinforcement learning models are particularly popular, given their ability to learn directly from environmental feedback.

This area continues to expand and address increasingly complex robot control problems. There are several good surveys on situated robotics which provide more detail and references (e.g. Brooks, 1991b; Matarić, 1998).

References

Agre P and Chapman D (1990) What are plans for? In: Maes P (ed.) *Designing Autonomous Agents*, pp. 17–34. Cambridge, MA: MIT Press.

Arkin R (1989) Towards the unification of navigational planning and reactive control. In: *Proceedings of the American Association for Artificial Intelligence Spring Symposium on Robot Navigation*, pp. 1–5. Palo Alto, CA : AAAI Press/MIT Press.

Asada M, Stone P, Kitano H *et al*. (1998) The RoboCup physical agent challenge: Phase I. *Applied Artificial Intelligence* **12**: 251–263.

Asada M, Uchibe E, Noda S, Tawaratsumida S and Hosoda K (1994) Coordination of multiple behaviors acquired by a vision-based reinforcement learning. In: *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems*, pp. 917–924. Munich.

Beer R, Chiel H and Sterling L (1990) A biological perspective on autonomous agent design. *Robotics and Autonomous Systems* **6**: 169–186.

Blumberg B (1996) *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT.

Brooks A (1991a) Intelligence without representation. *Artificial Intelligence* **47**: 139–160.

Brooks A (1991b) Intelligence without reason. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, Sydney, Australia, pp. 569–595. Menlo Park, CA: Morgan Kaufmann.

Brooks R (1986) A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation* **2**: 14–23.

Brooks R and Stein L (1994) Building brains for bodies. *Autonomous Robots* **1**: 7–25.

Burgard W, Cremers A, Fox D *et al*. (2000) Experiences with an interactive museum tour-guide robot. *Artificial Intelligence* **114**: 32–149.

Cliff D (1990) The computational hoverfly: a study in computational neuroethology. In: Meyer J-A and Wilson S (eds) *Proceedings, Simulation of Adaptive Behavior*, pp. 87–96. Cambridge, MA: MIT Press.

Connell J (1990) *Minimalist Mobile Robotics: A Colony Architecture for an Artificial Creature*. Boston, MA: Academic Press.

Connell J (1991) SSS: a hybrid architecture applied to robot Navigation. In: *Proceedings of the International Conference on Robotics and Automation*, Nice, France, pp. 2719–2724. Los Alamitos, CA: IEEE Computer Society Press.

Firby J (1987) An investigation into reactive planning in complex domains. In: *Proceedings of the Sixth National Conference of the American Association for Artificial Intelligence Conference*, pp. 202–206. Seattle, WA: AAAI/MIT Press.

Gat E (1998) On three-layer architectures. In: Kortenkamp D, Bonnasso RP and Murphy R (eds) *Artificial Intelligence and Mobile Robotics*. Menlo Park, CA: AAAI Press.

Gerkey B and Matarić M (2000) Principled communication for dynamic multi-robot task allocation. In: Rus D and Singh S (eds) *Proceedings of the International Symposium on Experimental Robotics 2000*, Waikiki, Hawaii, pp. 341–352. Berlin, Germany: Springer-Verlag.

Giralt G, Chatila R and Vaisset M (1983) An integrated navigation and motion control system for autonomous multisensory mobile robots. In: *Proceedings of the First International Symposium on Robotics Research*, pp. 191–214. Cambridge, MA: MIT Press.

Krieger M, Billeter J-B and Keller L (2000) Ant-like task allocation and recruitment in cooperative robots. *Nature* **406**: 992.

Kube R and Zhang H (1992) Collective robotic intelligence. In: *Proceedings, Simulation of Adaptive Behavior*, pp. 460–468. Cambridge, MA: MIT Press.

Maes P (1990) Situated agents can have goals. *Robotics and Autonomous Systems* **6**: 49–70.

Malcolm C and Smithers T (1990) Symbol grounding via a hybrid architecture in an autonomous assembly system. *Robotics and Autonomous Systems* **6**: 145–168.

Matarić M (1990) Navigating with a rat brain: a neurobiologically-inspired model for robot spatial representation. In: Meyer J-A and Wilson S (eds) *Proceedings, From Animals to Animats 1, First International Conference on Simulation of Adaptive Behavior*, pp. 169–175. Cambridge, MA: MIT Press.

Matarić M (1992) Integration of representation into goal-driven behavior-based robots. *IEEE Transactions on Robotics and Automation* **8**(3): 304–312.

Matarić M (1995) Designing and understanding adaptive group behavior. *Adaptive Behavior* **4**(1): 51–80.

Matarić M (1997) Behavior-based control: examples from navigation, learning, and group behavior. *Journal of Experimental and Theoretical Artificial Intelligence* **9**: 323–336.

Matarić M (1998) Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior. *Trends in Cognitive Science* **2**(3): 82–87.

Matarić M (2000) Getting humanoids to move and imitate. *IEEE Intelligent Systems* **15**(4): 18–24.

Pomerleau D (1989) ALVINN: an autonomous land vehicle in a neural network. In: Touretzky D (ed.) *Advances in Neural Information Processing Systems 1*, pp. 305–313. San Mateo, CA: Morgan Kaufmann.

Scassellati B (2001) Investingating models of social development using a humanoid robot. In: Webb B and Consi T (eds) *Biorobotics*, pp. 145–168. Cambridge, MA: MIT Press.

Schoppers M (1987) Universal plans for reactive robots in unpredictable domains. In: *Proceedings, IJCAI-87*, pp. 1039–1046. Menlo Park, CA: Morgan Kaufmann.

Webb B (1994) Robotic experiments in cricket phonotaxis. In: *Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, pp. 45–54. Cambridge, MA: MIT Press.

Werger B (1999) Cooperation without deliberation: a minimal behavior-based approach to multi-robot teams. *Artificial Intelligence* **110**: 293–320.

Further Reading

Arkin R (1998) *Behavior-Based Robotics*. Cambridge, MA: MIT Press.

Brooks R (1999) *Cambrian Intelligence*. Cambridge, MA: MIT Press.

Maes P (1994) Modeling adaptive autonomous agents. *Artificial Life* **2**(2): 135–162.

Russell S and Norvig P (1995) *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

Skill Acquisition: Models

Intermediate article

Todd R Johnson, University of Texas Health Science Center, Houston, Texas, USA
Hongbin Wang, University of Texas Health Science Center, Houston, Texas, USA
Jiajie Zhang, University of Texas Health Science Center, Houston, Texas, USA

CONTENTS

Skill acquisition is the process whereby a learner progresses from slow, deliberate, memory-intensive, error-prone performance on a task, to rapid, automatic, near-error-free performance. The process applies to activities ranging from simple perceptual–motor tasks, such as pushing a button whenever a light comes on or shifting gears in a car, to complex cognitive tasks, such as playing chess or diagnosing diseases.

AUTOMATICITY AND THE POWER LAW OF PRACTICE

Automaticity is an essential aspect of skill acquisition. Skills are thought to consist largely of collections of automatic processes. Automatic processing often occurs without attention. It is often fast, effortless, stereotypic, autonomous, and unavailable to conscious awareness (Shiffrin and Schneider, 1977). (*See* **Automaticity**; **Attention**; **Attention, Models of**; **Consciousness and Attention**)

Skill acquisition is thought to proceed through three stages (Fitts, 1964; Fitts and Posner, 1967; Anderson, 1982). The first stage is the cognitive stage, in which the learner remembers a set of declarative facts (e.g., instructions and cue–feedback relations) relevant to the skill and uses domain-general problem-solving methods to perform the task. For instance, a novice skier might need to recall and even rehearse the ski instructor's instructions about how to turn and stop. While turning, the skier must carefully monitor and adjust his or her stance so as to achieve the desired result. Errors at this stage can result from misinterpreted instructions, failure to correctly recall instructions, and the relatively long time required to recall and interpret the instructions and carefully monitor task execution. (*See* **Implicit Learning**; **Implicit Learning Models**; **Implicit and Explicit Representation**)

In the second stage, called the associative stage, declarative facts are gradually transformed into procedural knowledge, which often represents a strengthened, direct link from perceptual stimuli to appropriate actions and therefore is much more efficient. In addition, general-purpose problem-solving methods are replaced by domain-specific procedures. For instance, when a person first learns a new card game, the person must try to remember all the rules and carefully consider how each possible card might affect the desired goal; but in the associative stage, the player would much more quickly select an appropriate move in any given situation.

The final stage in skill acquisition is the autonomous stage, in which performance becomes increasingly automated and rapid. As skill improves in this stage, cognitive involvement decreases, and the learner often loses the ability to verbally describe how he or she does the task. For instance, if an experienced touch typist is asked to fill in a blank typewriter keyboard with all the letters, he may need to imagine typing words in order to recall where the letters are located. Likewise, once someone becomes proficient at skiing, she may need to carefully observe herself turning in order to tell a novice how to turn. (*See* **Motor Control and Learning**; **Motor Control: Models**; **Memory: Implicit versus Explicit**)

It is generally accepted that automaticity and skill acquisition result from practice. A robust empirical finding in skill acquisition is that the time to perform a task is a power function of the amount of practice on the task. This is known as the *power law of practice*. Formally:

$$T = aP^{-b} \quad (1)$$

where T is the time to do a task, P is the amount of practice, and a and b are constants. This power law of practice is also called the log–linear learning law because the above equation can be transformed into

$$\ln T = a - b \ln P \quad (2)$$

indicating that there is a linear relationship between the logarithm of the time and the logarithm of the amount of practice. It has been shown that the power law of practice is universally valid in a wide range of human tasks, including both perceptual–motor and cognitive tasks (Newell and Rosenbloom, 1981; Anderson and Schooler, 1991).

Various different models of automaticity and skill acquisition have been proposed. This article will describe several of these models and their relationship to the three levels of skill acquisition and the power law of learning.

MATHEMATICAL MODELS OF AUTOMATICITY

Schneider and Shiffrin's (1977; Shiffrin and Schneider, 1977) model of automaticity assumes that automaticity is automation of attentional processes in terms of automatic encoding and automatic response. Their model was mainly based on the findings of different effects of consistent and varied mapping conditions in memory search tasks. In a memory search task, subjects are presented with a list of one or more letters (e.g. A, E, K, T), called the memory set, followed by a probe (e.g. E), and are instructed to make a timed response to indicate whether the probe is in the memory set (e.g. Kristofferson, 1972). Probes not in the memory set are called *distracters* (e.g. Q in the above example), whereas those in the memory set are called *targets*. The typical result for a memory search task is that the time it takes to locate a target in a memory set increases with the number of items in the set. In the consistent mapping condition, the items used as targets are never used as distracters in any trials. For example, as E is a target in the example above, E would never appear as a distracter in any other trial. In contrast, in the varied mapping condition, target items on one trial are also used as distracters on other trials. For example, probe T is a target in a trial for memory set (T, M, X, U) but a distracter in another trial with memory set (A, H, W, D). The basic finding in the consistent mapping experiments is that after enough practice the search time for a target does not increase with the number of items in the memory set, that is, the search time is a flat function of the number of items. In varied mapping conditions, however, the search time for a target still increases with the number of items in the memory set even after considerable practice. Based on such results, Schneider and Shiffrin argued that attentional processes were automated in the consistent mapping condition through automation of encoding and automation of responses. In the varied mapping condition, encoding and responses could not be automated because of the inconsistency of targets and distracters from trial to trial.

Schneider and Shiffrin's model of automaticity has been challenged by several researchers. For example, Hirst *et al*. (1980) showed that people could learn to perform two complex tasks simultaneously without automation of encoding and response in either task.

A more serious challenge for Schneider and Shiffrin's model is Logan's instance theory (Logan, 1988). Instead of relating automaticity to attention in terms of encoding and response, Logan's theory relates automaticity to memory retrieval. Specifically, it argues that performance is automatic when it is based on direct memory retrieval of past solutions from memory, whereas performance is not automatic when it is based on general-purpose algorithms or procedures. A novice starts with general algorithms. As he or she gains experience with practice, specific solutions are learned and remembered. When these specific solutions are later directly retrieved from memory to solve problems, the performance becomes automatic. Therefore, 'automatization reflects a transition from algorithm-based performance to memory-based performance' (Logan, 1988, p. 493). Compared to Fitts's stage model, solving problems using general algorithms corresponds to the cognitive stage, and solving problems using direct memory retrieval roughly corresponds to the associative and autonomous stages. Logan's theory assumes that whenever a stimulus is encountered, it is encoded, stored, and retrieved separately as an instance. It is this accumulation of specific episodic instances with practice that is behind the algorithm-to-memory transition and makes automaticity possible. In particular, the chances of solving a problem by direct memory retrieval are higher later in practice because of the greater number of acquired candidate instances.

The instance theory is consistent with the power law of practice. There is a race between memory and the algorithm; whichever finishes first controls the response. With practice, memory comes to

dominate the algorithm because more and more instances enter the race, and the more instances there are, the more likely it is that at least one of them will win the race. This model naturally yields the power law of practice.

Crossman (1959) proposed a theory of skill acquisition in terms of a trial-and-error method. Specifically, a novice performs a task by randomly selecting a method from a pool of possible methods and noting its speed of performance. Different methods have different speeds. If a method proves to be faster, its probability of being selected increases. In the long run, the fastest method would have the highest probability of being selected. This theory, though different from Logan's, is also consistent with the power law of practice: it is easier to improve the performance by finding a faster method early in practice than later.

CHUNKING IN SOAR

The mechanism of chunking was first proposed by Newell and Rosenbloom (1981), and is fully developed in Soar, a unified theory of cognition (Newell, 1991). Chunking is one of a class of learning mechanisms for converting declarative knowledge into procedural knowledge, a process called 'knowledge compilation'. Chunking models the transition from the cognitive stage to the associative stage by automatically converting multiple cognitive processing steps for achieving a goal into a single production rule, called a chunk, that represents procedural knowledge for achieving the goal. For instance, according to this theory, if a learner initially adds 2 and 3 by starting with 2 and then counting up three numbers to arrive at 5, that person will then acquire a rule that directly produces the answer 5 when presented with the goal of adding 2 and 3. The chunking theory also includes simple mechanisms that can produce a general rule from a single task instance. For example, Soar can learn a chunk that encodes '$m + 0 = m$' by solving a single instance such as '$5 + 0 = 5$'. (*See* **Soar**; **Learning Rules and Productions**)

Chunking produces power-law speed-up for tasks that can be decomposed into smaller, frequently recurring subtasks. For example, if Soar first solves '$112 + 112$' by using counting to add the two digits in each decimal place, then it would acquire three chunks: '$2 + 2 = 4$', '$1 + 1 = 2$', and '$112 + 112 = 224$'. The third chunk is specific to a single problem, so it has no effect on the speed of solving other problems; however, the first two chunks will speed up any problems in which either two ones or two twos must be summed. As Soar solves additional problems, it will quickly acquire chunks for other pairs of digits. Eventually, speed-up may be due to quickly solving problems by applying more specific rules, such as '$112 + 112 = 224$'.

Chunking is somewhat similar in spirit to Logan's instance theory. Both theories bypass processing by directly using stored answers to previously solved problems. However, Soar stores only one instance, in the form of a production rule, whereas, Logan's theory stores an instance, as a declarative memory trace, each time the problem is solved. Furthermore, in Soar there is never a race between chunks and algorithms for solving a problem: if a chunk is available Soar always uses it. Finally, chunking can produce general chunks (such as '$m + 0 = m$'), which apply to more than one instance of a task.

LEARNING IN ACT-R

ACT-R is a cognitive architecture that accounts for a wide range of cognitive phenomena, including perceptual–motor tasks, memory, and problem-solving (Anderson and Lebiere, 1998). ACT-R contains five learning mechanisms to account for the progression of skill from the cognitive to the autonomous stage. ACT-R's knowledge compilation mechanism is production compilation. It produces procedural knowledge (in the form of production rules) by summarizing and generalizing one or more problem-solving steps based on a declarative representation of the learner's understanding of how one state of problem-solving led to a subsequent state. For example, suppose that ACT-R discovers that it can add 5 and 2 on a calculator by typing the keys `5 ENTER 2 +`. To learn procedural knowledge from this example, ACT-R must set the goal of building a declarative representation that links the goal of adding 5 and 2 to typing `5 ENTER 2 +`. This representation includes the goal, the solution, and the fact that + is a symbol for addition. The result is a production rule of the form: 'If the goal is to perform a mathematical operation on two numbers n_1 and n_2, and o is a symbol for that operation, then type n_1 `ENTER` n_2 o.' This mechanism is similar to Soar's chunking; but Soar's mechanism works automatically and continuously, whenever a goal is achieved, whereas ACT-R's mechanism requires an explicit goal to understand how one problem-solving state led to another. (*See* **ACT**)

ACT-R models continuous improvement from the associative to the autonomous stage through the speed-up of declarative fact retrieval and two

rule-tuning mechanisms (discussed below). Research on human memory shows that the time to retrieve a declarative fact follows a power law. ACT-R contains two mechanisms for speeding up the retrieval of declarative facts from memory: base-level learning and associative strength learning. The time taken to retrieve a declarative fact in ACT-R is proportional to its activation, which depends on two quantities: its base-level activation, which is determined by its retrieval history, and activation it receives from other facts in working memory. A fact's base-level activation reflects the prior probability that it will be retrieved. In general, facts that are more frequently and more recently retrieved have higher activation. A fact's activation is adjusted (either upwards or downwards) by means of associative links with facts already in working memory. Associative links are unidirectional and have a strength that reflects the conditional probability that a fact will be needed, given that an associated fact is in working memory. The equations that govern activation growth and decay directly produce power-law response curves.

The activation of a declarative fact also affects the probability of its retrieval in ACT-R. Some facts may have such low activation that they will never be retrieved. In addition, ACT-R can model retrieval errors. Suppose that a child has two contradictory facts, '$5+4=9$' and '$5+4=8$', where '$5+4=8$' has been retrieved half as often as the correct fact. Since activation controls the probability of retrieval, the child will occasionally recall the incorrect fact.

The time taken to execute a rule in ACT-R is determined by the rule's strength, which reflects the probability that the rule will be needed based on the rule's past history of use. The equation governing this value is identical to that for base-level activation. As a result, the time taken to execute a rule is a power function of its frequency of execution.

The mechanisms described above lead to faster response times, but skill acquisition also involves learning to select an appropriate action. ACT-R models this using two mechanisms that estimate a production rule's cost and probability of success based on previous experience. These estimates are then used to select from among competing production rules by estimating an expected gain for each rule. Rules with higher expected gains have higher probabilities of selection. The expected gain is $PG-C$, where P is the probability of eventual success, G is the value of achieving the goal (typically specified as the number of seconds one is willing to spend to achieve the goal), and C is the cost of achieving the goal using the selected action. For example, when children are first learning to add, they often begin by counting on their fingers, then gradually shift to retrieving the sum from memory. ACT-R models this using two competing production rules: one that says to count to determine the sum and another that says to retrieve the sum from memory. Suppose that at an early stage a child's past experience indicates that counting is successful 95% of the time and takes 5 seconds, whereas memory retrieval is successful 75% of the time and takes 1 second. Suppose further that the child is willing to spend up to 30 seconds to correctly add the two numbers. In this example, the expected gain from counting is $0.95 \times 30 - 5 = 23.5$, whereas the expected gain from recalling the sum is $0.75 \times 30 - 1 = 21.5$. In this case, ACT-R predicts that the child would favor counting over memory recall. As the child experiences more success with recall, the expected gains will shift in its favor, causing the child to shift to direct recall.

NEURAL NETWORKS

Artificial neural networks, or connectionist models, have also been used to model a range of cognitive processes including learning, memory, and problem-solving (Rumelhart and McClelland, 1986). Instead of serial and symbolic computation, a neural network model employs parallel distributed processing, whereby processing occurs in parallel across a large number of simple units. Neural networks are capable of learning. The essence of learning in a neural network is to tune the connection strengths (weights) between units so that multiple constraints from the environment and the network structure can be simultaneously satisfied. Thus the network gradually achieves more efficient computations. In supervised learning (or learning from examples), where for every input the desired output is provided by an external teacher, a network learns by seeking to minimize the difference between its actual output and the desired output. In unsupervised learning, where no information about the desired output is available, the network has to discover for itself the best way to respond. This discovery may involve extracting the correlation among the inputs, or satisfying constraints, such as only one unit being allowed to fire for any given input (competitive learning). Both types of learning can be used to model human skill acquisition. (*See* **Backpropagation**)

A neural network is trained with a limited number of examples, in the hope that it will

generalize from them and achieve good performance on novel inputs. The performance of a neural network is typically indicated by the error, a measure of the difference between desired outputs and actual outputs. When the expected error is plotted against the number of training examples, a learning curve results. Both theoretical calculations (e.g. Seung *et al*., 1992) and experimental results (e.g. Cortes *et al*., 1994) have shown that the learning curves of layered feedforward neural networks typically follow a power function.

REINFORCEMENT LEARNING

Reinforcement learning (a term borrowed from the field of animal learning) is a learning model in which a system learns a mapping from situations to actions, by interacting with the environment and attempting to maximize the reward (reinforcement) from the environment (Sutton and Barto, 1998). Reinforcement learning is different from supervised learning in that the learner is not told the desired action. Reinforcement learning is also different from unsupervised learning in that the learner does receive a feedback for the action it performs. The feedback is evaluative, and may be delayed. The learner has to discover the actions that yield the greatest long-term reward through trial and error. It has been shown that reinforcement learning captures several fundamental concepts underlying human skill acquisition, including learning from interaction, trial and error, and temporal credit assignment. Reinforcement learning has been applied in many arenas, including robot learning and game playing. (*See* **Reinforcement Learning: A Computational Perspective**; **Reinforcement Learning: A Biological Perspective**)

OTHER COMPUTATIONAL FRAMEWORKS FOR SKILL ACQUISITION

Posner *et al*. (1997) have studied the brain mechanisms of skill acquisition. They claim that skill acquisition represents circuitry change in the brain. Specifically, there are four mechanisms that work together to generate skill acquisition. Firstly, automaticity is a result of an increase in the links between previously isolated subsystems. Secondly, when skills are acquired, the size or the number of brain areas involved increases. As a result, the computation within modules is improved. Thirdly, skill acquisition could be due to the replacement of the initial components by other more efficient components. Finally, certain brain areas can start performing tasks formerly implemented by different areas, indicating circuit change. All four mechanisms have been found using sophisticated brain imaging techniques. (*See* **Reward, Brain Mechanisms of**)

Anderson and Schooler (1991) provide evidence that the power law of learning reflects an optimal adaptation to the environment. They analyzed information access and retrieval in a variety of natural settings and found that the probability of a specific fact recurring is a power function of its past frequency of use. For example, they found that the probability that a word would appear in the 101st utterance of a speaker was a power function of the word's frequency of use in the speaker's last 100 utterances. They found similar patterns when studying electronic mail messages, and newspaper headlines. (*See* **Rational Models of Cognition**)

Wixted and Ebbesen (1991, 1997) suggest that the power function best describes the process of forgetting. (But see Anderson and Tweney (1997) for a different view.) (*See* **Memory Distortions and Forgetting**)

APPLICATIONS IN EDUCATION

Skill acquisition models have several important uses in education. The models often suggest ways to structure educational material so as to maximize learning. To achieve the greatest speed-up, knowledge compilation theories suggest that tasks should be hierarchically decomposed into subtasks and then taught by starting with the more general tasks and working towards more specific tasks. ACT-R's theory of knowledge compilation through understanding suggests ways to structure examples that might help the learner achieve better performance. For instance, for ACT-R to properly acquire a rule (namely, '$m+0=m$') from the example '$5+0=5$', it must note that 0 is a special case, otherwise it would induce the rule '$m+n=m$'. ACT-R's theory of production rule speed-up and memory retrieval suggests that each rule and each declarative memory fact follows the power law of learning, so that continued practice on all rules and facts used in a task is necessary to improve performance or maintain skill. (*See* **Learning Aids and Strategies**; **Education, Learning in**)

Skill acquisition models can also be used to construct models of specific learning tasks. Once constructed, these models can guide instructional design or be used as part of an intelligent tutoring system that monitors student progress and

presents appropriate examples to enhance learning. By including specific erroneous knowledge and rules in such models, an intelligent tutoring system can detect incorrect learning and propose appropriate intervention.

TYPES OF LEARNING

The three stages of skill acquisition proposed in Fitts's model can help us to understand the relevance of a model to the overall acquisition process, regardless of the model's particular mechanism. An alternative, and equally useful, way to view skill acquisition models is in terms of the type of learning they support. Based on salient features of the task, learning can be classified into different categories, including supervised and unsupervised, deductive and inductive, and symbolic and subsymbolic (or neural). The latter distinction is particularly important, because symbolic and neural models seem to be suited to very different kinds of learning tasks. Symbolic learning models are good at learning abstract rules from one example, but these rules are often not resistant to noisy data. Neural models require many learning trials, but result in networks that can handle noisy and incomplete data. Accordingly, many researchers have proposed hybrid symbolic–neural learning models of skill acquisition that incorporate neural networks for classifying perceptual information and symbolic learning systems for the overall control of problem-solving (Sun and Alexandre, 1997).

References

Anderson JR (1982) Acquisition of cognitive skill. *Psychological Review* **89**: 369–406.

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson JR and Schooler LJ (1991) Reflections of the environment in memory. *Psychological Science* **2**: 396–408.

Anderson RB and Tweney RD (1997) Artifactual power curves in forgetting. *Memory and Cognition* **25**: 724–730.

Cortes C, Jackel LD, Solla SA, Vapnik V and Denker JS (1994) Learning curves: asymptotic value and rate of convergence. In: Cowan JD, Tesauro G and Alspector J (eds) *Advances in Neural Information Processing Systems*, vol. VI. San Francisco, CA: Morgan Kaufmann.

Crossman ERFW (1959) A theory of the acquisition of speed-skill. *Ergonomics* **2**: 153–166.

Fitts PM (1964) Perceptual–motor skill learning. In: Melton AW (ed.) *Categories of Human Learning*. New York, NY: Academic Press.

Fitts PM and Posner MI (1967) *Human Performance*. Belmont, CA: Brooks Cole.

Hirst W, Spelke ES, Reaves CC, Caharack G and Neisser U (1980) Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General* **13**: 208–217.

Kristofferson M (1972) When item recognition and visual search functions are similar. *Perception & Psychophysics* **12**: 379–384.

Logan GD (1988) Toward an instance theory of automatization. *Psychological Review* **95**: 492–527.

Newell A (1991) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Rosenbloom PS (1981) Mechanisms of skill acquisition and the law of practice. In: Anderson JR (ed.) *Cognitive Skills and Their Acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Posner MI, DiGirolamo GJ and Fernandez-Duque D (1997) Brain mechanisms of cognitive skills. *Consciousness and Cognition* **6**: 267–290.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I, *Foundations*. Cambridge, MA: MIT Press.

Schneider W and Shiffrin RM (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* **84**: 1–66.

Seung HS, Sompolinsky H and Tishby N (1992) Statistical mechanics of learning from examples. *Physical Review A* **45**: 6056–6091.

Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* **84**: 127–190.

Sun R and Alexandre F (1997) (eds) *Connectionist Symbolic Integration*. Mahwah, NJ: Lawrence Erlbaum.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Wixted JT and Ebbesen EB (1991) On the form of forgetting. *Psychological Science* **2**: 409–415.

Wixted JT and Ebbesen EB (1997) Genuine power curves in forgetting: a quantitative analysis of individual subject forgetting functions. *Memory and Cognition* **25**: 731–739.

Further Reading

Anderson JR (2000) *Learning and Memory*. New York, NY: Wiley.

Anderson JR and Lebiere C (1998) *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Levine DS (2000) *Introduction to Neural and Cognitive Modeling*. Mahwah, NJ: Lawrence Erlbaum.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Soar

Intermediate article

Frank E Ritter, Pennsylvania State University, Pennsylvania, USA

CONTENTS

Soar is a unified theory of cognition, and a cognitive architecture, realized as a production system, a type of expert system. It is designed to model human behaviour on multiple levels.

UNIFYING COMPUTATIONAL MECHANISMS TO FORM A THEORY OF COGNITION

Soar is a unified theory of cognition (UTC) realized as a computer program. It can be considered in three mutually complementary ways. First, it can be seen as a theory of cognition realized as a set of principles and constraints on cognitive processing: a cognitive architecture (Newell, 1990). In this respect, Soar provides a conceptual framework for creating models of how people perform tasks, typically assisted by the corresponding computer program. In this view Soar can be considered as an integrated architecture for knowledge-based problem solving, learning, and interacting with external environments. It is thus similar to other unified theories in psychology, such as ACT-R, EPIC, PSI, and CAPS. (*See* **Skill Acquisition: Models**; **Learning and Memory, Models of**)

Second, Soar can be seen as the computer program that realizes a particular theory of cognition. There are debates as to whether and how the theory is different from the computer program, but it is fair to say that they are at least highly related. It is generally acknowledged that the program implements the theory, and also that there are commitments that are not in the theory but that must be made in the program to create a running system. In this way it is similar to other cognitive theories realized as computer programs, such as ACT-R, and connectionist models of specific tasks realized as programs.

Third, Soar can be seen simply as a specialized AI programming language. In this view, what matters is only that it performs the task in an intelligent way. In this respect it is similar to expert system tools such as OPS5 and CLIPS. (*See* **Expert Systems**)

The deliberate combination of these approaches to understand intelligence has been fruitful. Researchers interested in creating cognitive models have used Soar primarily to model human behaviour in detail, to suggest new uses of existing mechanisms to create behaviour, and to propose improvements to the Soar programming interface. Researchers interested in creating AI programs have contributed to the efficiency, functionality, and generality of Soar as a programming language and provided information on the functional requirements of working systems.

SOAR AS A UNIFIED THEORY OF COGNITION

Soar was proposed by Newell (1990) as a candidate UTC. Newell presents a full description of the virtues of unification. Three of the most important are: coherence in theorizing ('it is one mind that minds it all'); bringing to bear multiple constraints from empirical data; and reducing the number of theoretical degrees of freedom. (*See* **Unified Theories of Cognition**)

Being a UTC does not mean that there is only a single mechanism for each task or behaviour, although in most places in Soar there is only one. It does mean that the set of unifying principles and mechanisms must work together to support all of cognition: there is not a big switch or a set of disjoint modules (Newell, 1992). ACT-R is another unified theory, although the set of mechanisms it proposes to account for all of human behaviour is longer. (*See* **ACT**)

Unified theories represent a grand vision. None of them can yet provide even a verbal explanation

for all of human behavior in terms of architectural mechanisms, let alone provide implemented models, and few have yet covered more than a small set of regularities. This name may even be a misnomer, for they are attempting to unify all of behavior, not just cognition. Their intention is to cover larger amounts of data than have been covered before, and to bind the different areas of cognition together through a common set of mechanisms. A common and unproductive criticism is that an architecture is wrong because all areas are not yet covered. All theories suffer from this limitation. A much more valid and valuable criticism would be that an important aspect of a given area cannot be accounted for by the current architecture.

GOAL-DIRECTED SEARCH IN HIERARCHICAL PROBLEM SPACES BASED ON PRODUCTION RULES

Soar – as a theory, as a cognitive modeling language, and as an AI programming language – incorporates problem spaces as a single framework for all tasks and subtasks to be solved, production rules as the single representation of permanent knowledge, objects with attributes and values as the single representation of temporary knowledge, automatic subgoaling as the single mechanism for generating goals, and chunking as the single learning mechanism. Specifically, Soar provides a general scheme for control – deciding what to do next – that is hypothesized to apply to all cognition. These mechanisms can be used in different ways, however. For example, chunking can be used to learn both declarative and procedural knowledge.

Soar can be viewed at three levels. At the highest level, it approximates a knowledge-level system (Newell, 1982). This is an abstract level where a system is described in terms of its knowledge, and which is only approximated by any realized system, including Soar. The two lower levels are the problem space level and the symbol level. These work together to support learning.

The Problem Space Level

Figure 1 illustrates the two lower levels. These are similar to Marr's lower two level of analysis. The higher of these levels is the problem space level, where behaviour is seen as occurring in a problem space made up of goals, problem spaces, states, and operators. Note that these terms refer to specialized constructs in Soar, which are related to, but not strictly equivalent to, their usual meanings in cognitive science. (*See* **Computer Modeling of Cognition: Levels of Analysis**)

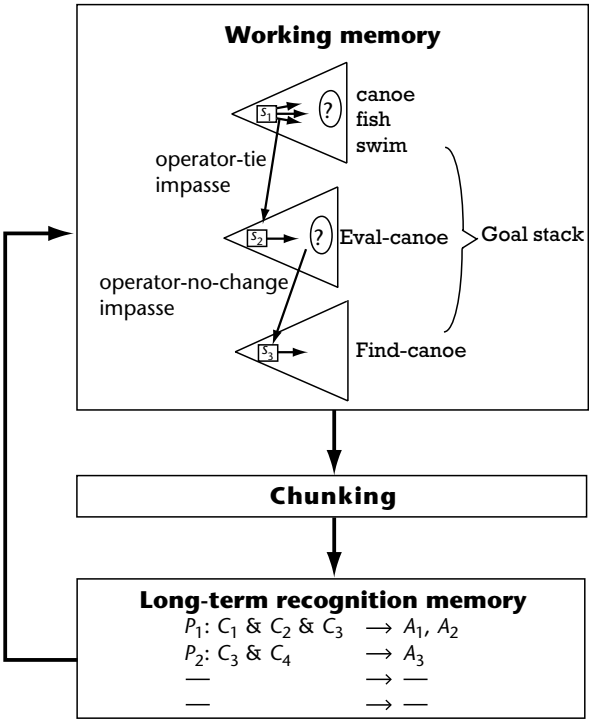


Figure 1. Structures in Soar.

A problem space is a set of representations for a problem, the structures for states, and all the operators relevant to those representations. The operators may be implicit and shared with other spaces. There can be several problem spaces active at any one time. A state may lack some required knowledge and have a state created to help it find the knowledge it needs, and similarly be providing knowledge itself. In figure 1 this state relationship is shown in the states S_1, S_2, and S_3. The main reason for organizing knowledge into problem spaces is that it reduces the search for information. This approach has also been used successfully as a software design technique.

While solving a problem there is a current state structure that specifies the situation of the problem solver in each problem space. For example, in a blocks world, the state might be 'block A is on top of block B, and block B is on the table'.

Fluent, expert behavior consists of a repeated cycle on the problem space level in which an operator is selected and is applied to the current state to produce a new (modified) current state. The process of choosing and applying a new operator (or creating a new state) is called a decision cycle. So in the example above, we could have applied an operator to move block A onto the table, after which

the current state would include the fact that block A and block B are both on the table.

The Symbol Level

The problem space level is realized by a lower level, a symbol level. At this level of analysis long-term recognition memory, realized as production rules, is compared to the current set of contexts. Rules (shown as P_1 and P_2 in figure 1) will have their conditions (C_1, C_2, etc.) matched to the current context. Their actions (A_1, A_2, etc.) will act on the problem space level to generate operators, propose how to choose between operators, implement operators, or augment the state with known inferences. Each cycle of rule application is called an elaboration cycle. There may be several elaboration cycles in each decision cycle. All rules whose conditions are satisfied are allowed to apply. If they make conflicting suggestions, the architecture sorts them out using an impasse (see below).

The rules are structured to match objects in the architecture. The rules can test the contents of states, and test for operators by name and by their contents. The rules' outputs are constrained to be in terms of the problem space structures. These constraints on the representation of the rules are part of what makes the system a cognitive architecture and not simply a free-form programming language.

Soar uses a modified RETE algorithm to apply the production rules. The time this algorithm takes to match a rule set is proportional to the number of memory elements that change, not the number of rules. This leads to very little slowdown as larger rule sets are used (but requires more computer memory). The largest systems created have had over a million rules with little or no slow down with additional rules (Doorenbos, 1995; Doorenbos *et al*., 1992).

Learning and Chunking

But what happens if something prevents the process of operator application from continuing smoothly? For example, perhaps the current knowledge in the Soar model cannot propose any operators to apply to the current state. Or the model may know of several applicable operators, but has no knowledge of how to choose between them. In such cases, the Soar model encounters an *impasse*. There is a limited number of types of impasse defined by the architecture, which primarily arise through a lack of knowledge (inability to apply or select an operator) or through inconsistent knowledge (conflict among operator proposals).

When Soar encounters an impasse in context level 1, it sets up a subcontext, a subgoal, at level 2, which has associated with it a new state, which may end up with its own problem space and operators. Note that the operators at level 2 could well depend upon the context at level 1. The goal of the level 2 context is to find knowledge sufficient to resolve the higher impasse, allowing processing to resume there. For example, we may not have been able to choose between two operators, so the level 2 subgoal may simply try one operator to see if it solves the problem, and if not, try the other operator.

The processing at level 2 might itself encounter an impasse, set up a subgoal at level 3, and so on. The problem solver usually has a stack of such levels, each generated by an impasse in the level above. Each level can have its own state, problem space, and operators.

In Figure 1, there were several operators proposed for the pond, including canoeing and fishing, and no knowledge was available to choose between them, so a new context was created to allow the architecture to consider this problem explicitly in a selection problem space, through what is called an 'operator-tie impasse'. Knowledge was available in that space, which suggested testing the canoeing operator and seeing how it would play out. The operator `Eval-canoe` was attempted, but nothing happened, so another impasse (an 'operator-no-change' impasse) was declared and an operator could be proposed in an evaluation problem space.

Whenever processing in the subgoal generates results that allow a higher level to continue, for example, if the operator `Find-canoe` allows `Eval-canoe` to continue, the architecture notices this, and automatically generates a new rule (also called a chunk) to summarize this problem solving. This rule's conditions are based on backtracking through the problem solving to find out what aspects of the initial situation were used, and the rule's actions are the output of `Find-canoe` that removed the higher-level impasse. In this case it would probably be a change to the `Eval-canoe` operator or to its state.

The next time such a condition occurs, the rule will match and update the operator or state, and the impasse will be avoided. This is the basic learning mechanism in Soar. This approach provides a strong theory of when and how learning and transfer will occur.

Chunking has been used to create a wide range of higher-level learning – including explanation-based learning, declarative learning, instruction

taking, and proceduralization – by varying the type of impasse and the knowledge used to resolve it.

THE HISTORY OF SOAR

The intellectual origins of Soar can be found in the seminal work of Newell and Simon on human problem solving. This builds upon work on production system architectures in the 1970s onwards, particularly Newell's work on the problem space as a fundamental category of cognition (Newell, 1980). Soar as a unified theory of cognition has some of its theoretical roots in the Model Human Processor (Card *et al*., 1983).

The first implementation of Soar was built by Laird, modifying Rosenbloom's XAPS architecture. Impasses were introduced in Soar 2, a reimplementation of Soar in OPS5, which allowed rules to fire in parallel and included the problem space decision mechanism. The original motivation was both functional (to create an architecture that could support problem solving using many different weak methods arising from the knowledge that was available) and structural (to create an architecture that integrated problem spaces and production systems). 'SOAR' was originally an acronym for 'state operator and result', but it is no longer recognised as being an acronym because the theory is more complex.

A major watershed in the development of Soar was Newell's William James lectures at Harvard (Newell, 1990). These lectures defined what a unified theory in psychology should include, proposed Soar as a candidate unified theory, and extended the Soar theory, providing some detailed examples. (*See* **Newell, Allen**)

The recent development of Soar has been driven by applications. Soar models have been applied to real-time domains such as flying simulated aircraft (Jones *et al*., 1999). Analyses of running models showed that the state and problem space in the original Soar theory were not being used as had been initially imagined: in most cases they did not vary and were simply reiterations of the goal. Later versions of Soar have dropped problem spaces and states as explicit reserved structures but allowed the modeller to represent them in the goal. This has led to faster systems that allow several models on a single computer to interact in real time, performing complex tasks. Because these context slots were not being used by models, their removal did not lead to changes in behavior.

Architectural work on Soar is currently focused on improving its interface, introducing new learning algorithms built upon the chunking mechanisms, tying Soar to external worlds (including behaviour moderators like stress), and the implications of interaction for problem solving and learning. Future work could include reviving the Neuro-Soar project (Cho *et al*., 1991). This project showed that it was possible to realize the symbol level of Soar with a connectionist network, although modelling so many theoretical levels made it slow.

MATCHING HUMAN PERFORMANCE IN DIVERSE DOMAINS

One of the strengths of Soar is that it predicts the action sequences and times to perform tasks (Newell, 1990). The parameters chosen by Newell have been gradually refined. The Soar philosophy has been to retain the same constraints from problem to problem, rather than having numerous parameters that can be adjusted for a specific task or data set.

For cognitive modelling, Soar is most effective at modelling deliberate cognitive human behavior at timescales greater than 50 ms. Published models include human–computer interaction tasks, typing, arithmetic, categorization, video game playing (i.e., rapid interaction), natural language understanding, concept acquisition, learning by instruction, verbal reasoning, driving, job shop scheduling, and teamwork.

Soar has also been used for modelling learning in many of these tasks, many of which involve interaction with external environments. Soar does not yet have a standard model for low-level perception or motor control, but two systems that could be integrated, EPIC-Soar (Chong and Laird, 1997) and Sim-eyes and Sim-hands (Ritter *et al*., 2000), have been created. Learning adds significant complexity to the structure of the model, however. (*See* **Learning Rules and Productions**)

One of the signature data regularities modeled in Soar is the learning curve. The learning curve predicts that the time to do a task decreases according to a power law (or perhaps an exponential decay). Soar's prediction of the power law of practice for a task arises from how models in Soar do the task and what they learn.

The first, and probably the simplest, way in which the power law of learning has been modelled in Soar is for the Seibel task. This simple task is to push the buttons on a panel corresponding to lights that are on. There are ten lights, therefore 1023 possible patterns of lights where at least one light is on. The model proposes two operators to do a left and a right subregion. If these are not individual lights, then an impasse occurs, and each subregion

gets two operators. This continues until a single light is a subregion. The model can then return a chunk that does both subregions, initially, two lights. Early trials generate two-light patterns that occur often and are useful. Later trials can build larger patterns, with more lights, that occur less often but save more time.

The Seibel model was one of the first learning models in Soar, and represents probably the simplest approach to learning in Soar. It does not represent more complex and accurate learning methods. Current models include learning by instruction, learning by following others, modeling transfer between tasks, and learning category knowledge. We are now at the point where, if we can model performance on a task in Soar, we expect to be able to model learning. Nearly all of the cognitive models in Soar are models that learn, and a majority of these have been compared with data.

SOAR AS AN EXPERT SYSTEM DEVELOPMENT ENVIRONMENT

Soar has also been used to create a variety of classification expert systems, that is, systems that classify situations. These including lift planning, production scheduling, diagnosis, robotic control, and computer configuration. It has been used in the Sisyphus knowledge elicitation comparisons.

Perhaps the greatest success for Soar expert systems has been in a procedural domain, flying simulated aircraft in a simulated hostile military environment. In one experiment Soar flew all of the US aircraft in an international 48-hour simulation exercise (Jones *et al*., 1999). The simulated pilots talked with each other and with ground control, and carried out over 700 sorties with up to 100 planes in the air at once.

For building artificial intelligence (AI) and expert systems Soar's strengths are in: integrating knowledge; planning; the ability to react quickly by modifying its internal state or changing its goal stack; search; and learning within a very efficient architecture. It also has the ability, used in a model that plays Quake®, to create a state mirroring its opponent's state, and consider what the opponent will do by considering what it would do itself in the same situation.

CHALLENGES FOR SOAR AND OTHER UTCs

Like any unified theory of cognition realized as a program, Soar faces major challenges. Work continues on applying Soar to a wider range of tasks and including learning and interaction in these models. Meanwhile, usability is becoming increasingly important as Soar moves out of the academic world into the world at large.

Soar has been developed and used by a community of researchers. Keeping a group of up to 100 researchers together intellectually has been difficult. Explicit mechanisms are necessary, such as repositories of papers and programs, regular meetings, mailing lists, frequently asked question lists (FAQs), and websites.

SUMMARY

There are a number of relatively unique capabilities that arise out of the combination of the structures and mechanisms in Soar. First, problem solving and learning are tightly intertwined: chunking depends on the problem solving, and most problem solving would not work without chunking. Secondly, interruptibility is available as a core aspect of behaviour. Rules are matched against the whole context stack. Processing can thus proceed in parallel on several levels. If the situation changes, rules can fire quickly, suggesting new operators at the level most appropriate for dealing with the change. Thirdly, it is possible to create large rule systems because they can be organized in problem spaces; and the architecture makes them fast to build and to run. Fourthly, planning can be integrated with reacting as well as with dynamic decomposition of tasks.

It takes effort to learn Soar. More practice is needed than for other, simpler, systems. Those projects that have used Soar successfully have often been able to solve problems that were previously unsolvable or unmodellable, but not without hard work on the part of the modellers.

Soar is, perhaps uniquely, appropriate for creating large cognitive models or expert systems, or for projects where learning or interaction (or both) are important.

References

Card S, Moran T and Newell A (1983) *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Erlbaum.

Cho B, Rosenbloom PS and Dolan CP (1991) Neuro-Soar: a neural-network architecture for goal-oriented behavior. In: *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pp. 673–677. Hillsdale, NJ: Erlbaum.

Chong RS and Laird JE (1997) Identifying dual-task executive process knowledge using EPIC-Soar. In: *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 107–112. Mahwah, NJ: Erlbaum.

Doorenbos R, Tambe M and Newell A (1992) Learning 10,000 chunks: what's it like out there? In: *Tenth National Conference on Artificial Intelligence (AAAI'92)*, pp. 830–836. Menlo Park, CA: AAAI.

Doorenbos RB (1995) *Production Matching for Large Learning Systems*. PhD thesis, Carnegie-Mellon University. Tech. Report CMU-C5-95-113.

Jones RM, Laird JE, Nielsen PE *et al*. (1999) Automated intelligent pilots for combat flight simulation. *AI Magazine* **20**(1): 27–41.

Newell A (1980) Reasoning, problem solving and decision processes: The problem space as a fundamental category. In: Nickerson RS (ed.) *Attention and Performance VIII*. Hillsdale, NJ: Erlbaum.

Newell A (1982) The knowledge level. *Artificial Intelligence* **18**: 87–127.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press. [Précis: Unified theories of cognition. *Behavioral and Brain Sciences* **15**: 425–492.]

Ritter FE, Baxter GD, Jones G and Young RM (2000) Supporting cognitive models as users. *ACM Transactions on Computer-Human Interaction* **7**(2): 141–173.

Further Reading

Laird JE and Rosenbloom PS (1992) In pursuit of mind: the research of Allen Newell. *AI Magazine* **13**(4): 17–45.

Laird JE and Rosenbloom PS (1995) The evolution of the Soar cognitive architecture. In: Steier DM and Mitchell TM (eds) *Mind Matters*, pp 1–50. Hillsdale, NJ: Erlbaum.

Rosenbloom PS, Laird JE and Newell A (1992) *The Soar Papers: Research on Integrated Intelligence*, 2 vols. Cambridge, MA: MIT Press.

Ritter FE, Baxter GD, Avraamides M and Wood AB (2001) *Soar FAQ*. [http://ritter.ist.psu.edu/soar-faq.html.]

The Soar Group: http://ai.eecs.umich.edu/soar/soar-group.html.

Sparse Distributed Memory

Intermediate article

Tim A Hely, University of Edinburgh, Edinburgh, UK

CONTENTS

The sparse distributed memory is a mathematical technique for storing and retrieving large random binary patterns. The memory has been used in simple visual and linguistic applications, and has been proposed as a model of cerebellum function.

BACKGROUND

In 1988, Pentti Kanerva outlined a simple and mathematically elegant formulation for an associative memory called the 'sparse distributed memory' (SDM) (Kanerva, 1988). The SDM has been used in a wide range of studies and applications, and numerous extensions to the original model have been developed. It displays many of the characteristics of human memory, such as content-addressability and the association of closely related concepts when presented with a single input pattern. Furthermore, the architecture of the memory can be mapped onto that of the cerebellar cortex.

OVERVIEW

The SDM is a generalized random-access memory. In simple terms, it is a 'pattern computer'. All memories, including the SDM, have a large number of 'locations' where information can be stored. In the SDM, memory locations are randomly distributed in the input space (the set of all possible input patterns). The address of a memory location is a long binary pattern (e.g. 1000 bits) which has the same form as the input patterns presented to the memory. Input patterns may be presented to the memory to be stored (writing to memory), or to access information from parts of the memory (reading from memory). Like other networks, such as the Hopfield network (see below), the SDM works best when the input pattern consists of random zeros and ones.

A binary input consisting of only 2 bits can represent 4 possible patterns: `00`, `01`, `10`, and `11`. This input can also represent the coordinates of one of the $4 = 2^2$ corners of a 2-dimensional square: {(0,0), (0,1), (1,0), (1,1)}. Pairs of points that differ by only 1 bit lie closer together than pairs that differ by 2 bits. The number of bits that are different in two binary words (of the same length) is known as the Hamming distance. A 3-bit input can represent one of the $8 = 2^3$ corners of a 3-dimensional cube. Similarly, the number of different patterns that can be generated by a binary word 1000 bits long is 2^{1000}, and these form the corners of a 1000-dimensional hypercube. (The number of atoms in the universe is about 2^{250}.)

It is not feasible to assign a memory location to each possible input, as the storage requirements are too great. Instead, in the SDM a large number of binary patterns (e.g. a million) are chosen randomly as the addresses of the memory locations (also called 'hard locations'). This gives a representative sample of the input space; however, the probability that a random input will correspond exactly to a memory location is effectively zero (a million divided by 2^{1000} is about 10^{-295}). Rather, the addresses are distributed amongst multiple locations, some of which will probably be close to the input.

The SDM model depends fundamentally on subtle, nonintuitive properties of high-dimensional metric spaces and random distributed representations. It is not so important that the dimensions are binary.

CONTENT-ADDRESSABLE MEMORY

In a conventional computer, a memory location is accessed by its address, not by its contents. However, in the SDM, the input patterns can serve both as addresses to the memory locations and as the

data to be stored. This feature gives rise to a form of content-addressability, which is a characteristic of human memory. For example, to find out where the quote 'to be or not to be' comes from, a conventional computer program would have to search all the files stored in the computer until it found a matching string. But in the SDM, an input pattern is stored in the region of memory where similar information has previously been stored, and the quote directly triggers the play 'Hamlet'.

HOW THE SDM WORKS

In the SDM, each hard location consists of an N-bit address, which corresponds to a point in the memory space, and a sequence (or row) of signed integer 'counters' to store data. An input to the memory consists of two binary vectors: an 'address vector', which determines which locations are accessed, and a 'data vector', which is stored in the counters of the hard locations. The lengths of the address and data vectors may be different. When storing data in the SDM, first of all the Hamming distance between the input address vector and the address of each hard memory location is calculated. A copy of the input data vector is added to the data already contained in the memory counters of all locations where this distance is less than or equal to a preset value (the 'activation radius'). In this way, similar data are stored at locations that have similar addresses. If the input data vector to be stored is the same as the input address vector, the SDM functions as an autoassociative memory.

In the original SDM, the locations have integer counters initially set at 0 (i.e. storing no information). At each memory counter of a location that has been selected in a write operation, a 1 in the corresponding position of the input data vector increments that counter value by 1, and a 0 decrements it by 1. Figure 1 (adapted from Rogers (1988)) is a representation of how the pattern is stored in the SDM. It shows the state of the SDM after only one write to the memory, when the counters are storing only a single item of data. (The diagram also shows the operation of reading from the memory.)

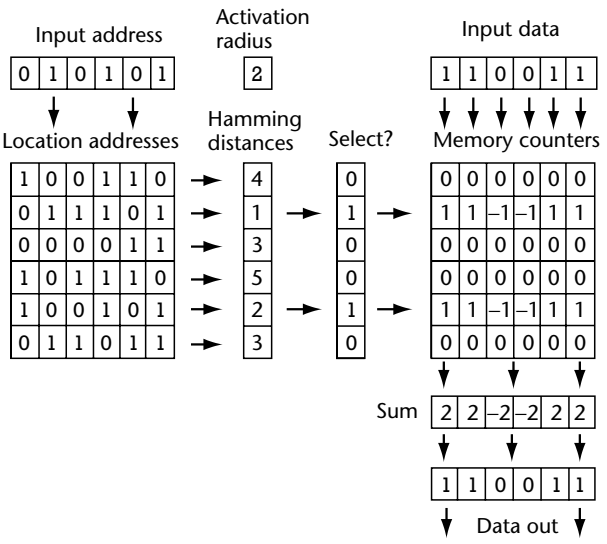


Figure 1. Writing to and reading from the SDM. (Adapted from Rogers (1998).)

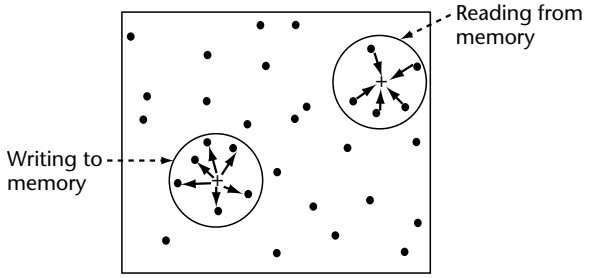


Figure 2. Writing to and reading from the SDM. The box represents the entire memory space (all possible addresses or cues). Each dot is the address of a memory location. The cross represents an address cue. When writing to memory, the system adds the contents of the input vector to all locations within the activation radius of the cue. When reading from memory, the system pools the contents of all locations within the activation radius of the cue.

The Write Operation

The word written to the memory is stored at those locations that lie sufficiently close to the input address vector. In Figure 1 the activation radius has been set to 2. For a location to be selected, the Hamming distance between the input address and the location address must be less than or equal to this value. (In the figure, these locations have a value of 1 in the 'Select?' column.) The counters of these memory locations are then updated with the input data pattern. In Figure 1, just one item of data has been stored, so the counters store only that item. Note that the input address pattern can be different from the input data pattern.

Figure 2 gives an alternative representation of the write and read processes. In the write process, the data are distributed to all locations within the activation circle.

The Read Operation

The word that is read from the memory at the reference address is the average of the contents of

all locations lying within a certain Hamming distance from it. The selection process is the same as for the write operation; however, the counters are not altered. The activated vectors of counters are added to produce an overall sum. At each position in the sum, a negative value sets the output bit to 0, while a nonnegative value sets the output bit to 1.

When the write process is repeated, the contents of a particular memory location will consist of the sum of all previous inputs. New input data are simply added to the existing values of the counters. As old patterns are not removed when a new word is written to the memory, the SDM stores copies of every data pattern presented to it.

If random binary data are stored in the memory, the probabilities of a counter value increasing by 1 or decreasing by 1 are equal. However many patterns a location may store, the statistical expectation of each counter will be zero. Figure 3 shows an SDM where three input data vectors **x**, **y** and **z** have been written to locations within the activation radius of input address points **A**, **B** and **C** respectively. (Note that points **A**, **B** and **C** do not correspond to any hard memory locations.)

In Figure 3, 10 copies of each of the patterns **x**, **y** and **z** have been stored around points **A**, **B** and **C** respectively. Reading from **A** will then retrieve the 10 copies of the original signal **x** plus 3 copies each of **y** and **z**, which make up the noise term. Pooling the data results in a pure signal term from the original input and a noise term due to other neighboring inputs. If random patterns are being stored, the noise term tends to be relatively small. It is only

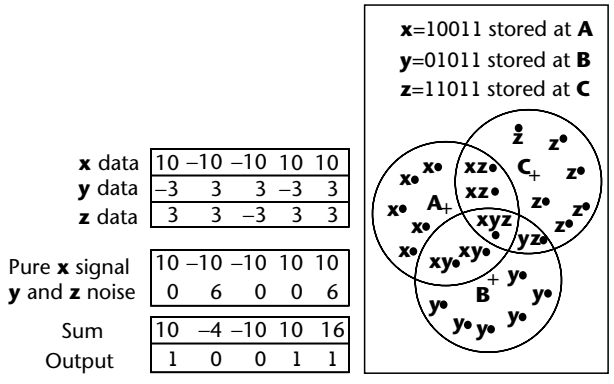


Figure 3. Retrieving a signal subject to noise. Reading at **A**, within the activation radius there are 10 locations storing **x**, 3 locations storing **y** and 3 locations storing **z**. Despite the noise due to the **y** and **z** signals, the original signal **x** is correctly retrieved.

when this noise term has a value opposite to and greater than the signal that an error will occur. Kanerva (1988) showed that this method results in the retrieval of the correct word, for appropriate values of the activation radius, with statistical reliability.

Before using the SDM, we must select a value for the activation radius. The memory performs best when the read and write radii are the same. In choosing the radius, there is a trade-off between obtaining a good signal over a reasonable range of uses and minimizing the noise. It turns out that a radius that selects slightly less than the square root of the total number of locations is a good solution. (The optimum probability of a 'hit' is $(2MT)^{-1/3}$ when T patterns are stored in a memory of M locations (Kanerva, 1993, p. 60).) The radius can also be set dynamically at each write or read operation to ensure that approximately the same number of locations are accessed each time.

ATTRACTORS IN THE SDM

When random input patterns are originally stored in an empty SDM, the point of each pattern in input space may be considered as an attractor (which is embedded in the particular set of memory locations to which the data are written). Reading data from nearby points will also retrieve the target pattern. This will not work if the memory is saturated (i.e. storing too many patterns), or if many almost identical copies of a pattern have previously been stored. (In practice, the latter situation is unlikely to occur unless the input patterns are low-dimensional or nonrandom.) If noisy versions of the same input pattern are stored in the memory, the attractor may actually represent a 'clean' version of the original pattern, as the noise can cancel out across the set (Kanerva, 1993).

An interesting situation can occur when the SDM is used as an autoassociative memory and multiple patterns have been stored in the network. If a noisy input pattern is fed into the SDM, the resulting output can be iteratively fed back into the memory until the network converges on a stable solution. Kanerva showed that the original target pattern is obtained if the Hamming distance between the input and target patterns initially decreases as the output is fed back into the network. However, if the Hamming distance initially increases, the memory will converge to a different attractor. Kanerva described this as a 'self-propagating search' and noted a similarity between this convergence process and the 'tip of the tongue' memory phenomenon.

RELATED MODELS

The SDM architecture can be mapped onto a standard 'two-layer' feedforward neural network design. Traditionally the input layer is not counted. The input units convey patterns to the memory. The weights of the hidden units store the hard locations and the weights of the output units store the memory contents. Other related associative memory models include the Willshaw and Hopfield networks (summarized below) and the Kohonen self-organizing network, which is usually used to map high-dimensional input data onto a one- or two-dimensional output. (*See* **Connectionism**)

The Willshaw network (Willshaw *et al*., 1969) has one of the simplest learning rules. In the fully connected network with binary inputs of length M and data of length N, there are $M \times N$ binary weights W_{ij} ($1 \leq i \leq M$, $1 \leq j \leq N$), initially all set to 0. On presentation of a pattern pair (**A**, **B**), W_{ij} is set to 1 if $A_i = B_j = 1$, where A_i is the i^{th} bit of input **A**. This is related to Hebbian learning, where coincidental firing of both presynaptic (input) and postsynaptic (output) cells is required to modify a synapse. Once activated, however, a weight remains on (set to 1) forever. This network performs best under sparse activity conditions (much fewer ones than zeros), otherwise it quickly saturates. A 'winner takes all' approach can be used to retrieve the output pattern. Calculate the dendritic sum, $d_i = \sum_j W_{ij}A_j$. Then select the required number of output units with the highest dendritic sums. The associative network has 69% of the capacity of a random access store with no associative capability. The network works well under sparsely connected conditions, and has been proposed as an abstract model of the hippocampus, which also has sparse connectivity (Graham and Willshaw, 1995). (*See* **Synapse**; **Hebb Synapses: Modeling of Neuronal Selectivity**; **Hippocampus**)

The Hopfield network (Hopfield, 1982) has N fully connected units with values of $\pm$ 1 and a (symmetric) weight update rule given by $W_{ij} = (\sum_n A_i^n A_j^n)/N = W_{ji}$, where the A_i^n represent p input patterns ($1 \leq n \leq p$), and $1 \leq i, j \leq N$. Usually there is a training phase, in which the weights are allowed to change, followed by a test phase, in which the weights remain fixed. In the test phase, the network can be used as an autoassociative, content-addressable memory. Stored patterns can be dynamically retrieved by asynchronously updating units (randomly choosing one at a time) according to the rule $A_i = \text{sgn}(\sum_j W_{ij}A_j)$ (where sgn $(x) = 1$ if $x \geq 0$, -1 otherwise) (cf. the dendritic sum in the Willshaw network above). If the test pattern is close enough to (a small Hamming distance away from) one of the stored patterns, the network will converge on the correct pattern, which acts as an attractor in a manner analogous to Kanerva's 'self-propagating search'. The Hopfield network has been the focus of extensive research. The maximum information density of the network that allows (random) patterns to be stored and dynamically retrieved in this way is $0.138N$.

To summarize, the SDM and the Willshaw and Hopfield associative networks use different methods to store (and retrieve) multiple copies of the input pattern. In the SDM, data are stored at those units closest to the input address; in the Willshaw network, the data are stored at all simultaneously active units; and in the Hopfield network, each data pattern is stored across the entire memory.

EXTENSIONS TO THE SDM

There have been many variations of the original SDM model (see Kanerva (1993) for a comprehensive summary). Many of these models aim to improve the theoretical capacity of the original memory which was calculated by Chou (1989). Rogers (1988, 1989) used a method called 'data tagging' to improve the efficiency of the SDM, but this method requires a greater storage capacity and takes longer to run. Jaeckel (1989a, 1989b) developed several sparsely connected versions of the SDM, including the 'selected coordinate' and 'hyperplane' designs. In a selected coordinate design with, say, 1000-bit binary addresses, only 10 of the 1000 bits might be 'set' to 0 or 1. The other 990 bits are not involved in the activation process. A location is activated only when all 10 of its selected bits match the corresponding bits in the input address vector. The hyperplane design uses a similar approach but with fewer selected bits, perhaps 3. This works well with nonrandom or skewed data, e.g. when there are on average 100 ones and 900 zeros in a 1000-bit address. Both designs are less computationally intensive than the original SDM and can perform better than it in certain circumstances.

Ryan and Andreae (1995) have extended the SDM to overcome some of the difficulties associated with storing and retrieving nonrandom input data. When reading data from the memory, they use a variable threshold for the values in the sum, which reflects the actual ratio of ones and zeros in the data set. If there are an equal number of ones and zeros, the threshold is 0, as before. However, if

there are more (or fewer) ones than zeros, the threshold is increased (or decreased) commensurably, resulting in an improved performance. To overcome some of the problems associated with a fixed activation radius, an alternative approach was developed by Hely *et al*. (1997). In this approach, the input signal 'propagates' throughout the memory, and its strength is attenuated as it encounters memory locations further from the input address.

Much recent research on the SDM has been carried out at the Swedish Institute of Computer Science, using sparse address patterns. Karlsson (1995) has developed a fast variant of the Jaeckel hyperplane design which makes it possible to simulate large SDMs in standard computer architecture. A more efficient way of reading the SDM memory using implicit information to remove noise has been developed by Sjoedin (1996). Kristoferson (1998) has shown that SDM performance does not scale up in the original SDM (the ability to tolerate noise in the retrieval cue decreases with memory size), but does scale up if sparse input address vectors are used.

The original SDM architecture has been used in many simple memory applications. These have included two-dimensional character recognition (Kanerva and Olshausen, 1989) and speech recognition (Prager and Fallside, 1989). In the character-recognition model, 16 $\times$ 16-pixel black-and-white images were converted into 256-bit binary patterns. These patterns were then stored and accessed using the normal operation of the SDM. Other applications have included converting text into phonemes: the address of a location corresponded to the text pattern, while the counters stored the phonemes associated with that word (Joglekar, 1989).

COMPARISON BETWEEN THE SDM AND THE CEREBELLUM

In the early 1970s, Kanerva noted the similarity between the SDM circuits and those of the cerebellar cortex. The cerebellum is involved in coordinating movements, and is located behind and below the cerebral hemispheres. Its highly regular design and repetitive architecture make it an ideal subject for theoretical analysis. It consists of the cerebellar cortex and deep cerebellar nuclei (the latter will not be discussed further here). The cerebellar cortex is organized into two layers: the inner or granular layer, and the outer or molecular layer. The principal cell and fiber types are shown in Figure 4 (adapted from Kanerva (1988)). (*See* **Cerebellum**)

Caution should always be exercised in attempting a one-to-one comparison between an abstract high-level model, such as the SDM, and the human brain, whose complexity is far from understood. However, it is instructive to summarize the functions of the cells and fibres in the cerebellum alongside the interpretation developed by Kanerva within the SDM framework.

Mossy Fibers

Mossy fibers originate from cells outside the cerebellar cortex. They form synapses with the granule cells.

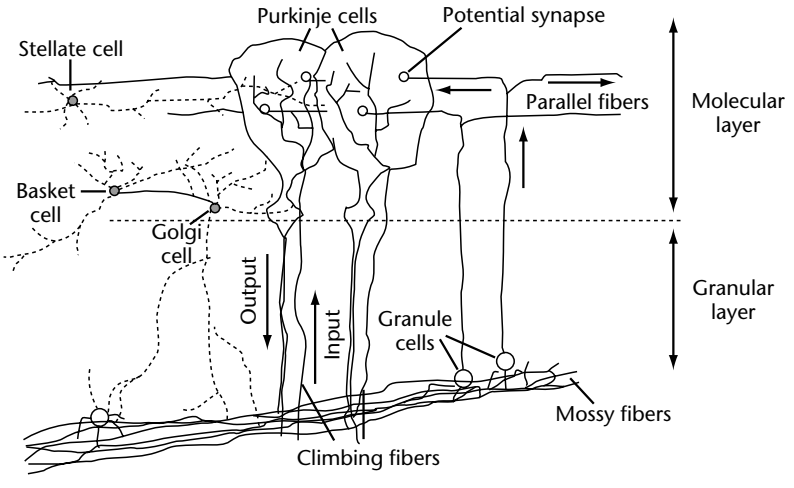


Figure 4. Simplified structure of the cerebellar cortex of the brain. Horizontal lines are continuations of parallel fiber axons. Single-headed arrows indicate direction of information flow (direction of action potential propagation along the axon).

In the SDM interpretation, they act as address lines to transmit the input address vectors or cues to the memory.

Granule Cells

Granule cells are the most abundant kind of neuron in the nervous system (there are as many as 10^{11} in humans). They occupy the inner layer, and receive input from the mossy or granular fibers and the Golgi cells.

In the SDM interpretation, their dendrites and bodies correspond to the hard memory address locations. The firing of a granule cell can be considered equivalent to the activation of a location in the SDM model.

Parallel Fibers

Parallel fibers are the axons of granule cells. They form synapses with the Purkinje cells, and also with the other cell types of the cerebellum. A single parallel fiber passes through between 200 and 450 of the Purkinje cells' dendrite planes.

In the SDM interpretation, the row of counters for a memory location is found along a parallel fiber and is activated by it. The synapse between a parallel fiber and a Purkinje cell acts as the counter for a given bit position. The other interneurons use the parallel fiber input to provide negative feedback when adjusting the thresholds of the Purkinje cells.

Purkinje Cells

The Purkinje cell axon provides the only output from the cerebellar cortex. The dendrite system of one Purkinje cell intersects as many as 400 000 parallel fibers and forms synapses with some or all of them. It has inputs from the stellate and basket cells and the climbing fibers, and weak connections with other Purkinje cells.

In the SDM interpretation, the Purkinje cell pools data for a single bit of output. The cell sums the inputs from parallel fibers and fires if the sum is over the threshold.

Stellate and Basket Cells

The stellate and basket cells are known as interneurons, and they reside in the outer layer. Both types of cell receive input from the parallel fibers, and their output has an inhibitory effect on the Purkinje cells. Albus (1971) hypothesized that synaptic weakening occurs at the parallel fiber synapses on basket and stellate dendrites. This effectively provides both positive and negative training. Positive adjustments occur by weakening excitatory synapses on inhibitory interneurons, and negative adjustments by weakening excitatory synapses on the Purkinje output cells.

In the SDM interpretation, these cells adjust the thresholds of the Purkinje cells to decide whether zeros or ones are more frequent in the pooled data.

Climbing Fibers

A Purkinje cell receives input from a single climbing fiber which branches extensively throughout the Purkinje dendritic tree. The firing of a climbing fiber guarantees the firing of the Purkinje cell output.

In the SDM interpretation, these are the data-input lines. They pair with the outputs (Purkinje cells) and they go to the vicinity of the bit locations (the Purkinje cell joins synaptically with the parallel fibers). In the microcircuit model (see below), the climbing fibers indicate when learning is to take place. Coincident parallel-fiber and climbing-fiber activity indicates where learning is to occur.

The granule cells, mossy fibers and climbing fibers are excitatory. The Purkinje, Golgi, basket and stellate cells are inhibitory. Kanerva does not discuss the functioning of the Golgi cells in relation to the SDM. A more generalized version of the SDM (the 'microcircuit architecture' model) has been developed, which also incorporates cerebellar interneurons (Miles and Rogers, 1993). In the microcircuit model, the basket cells also connect correlated cells responding to conditioned and unconditioned stimuli, facilitating learning.

THE SDM AS A MODEL OF HUMAN MEMORY

Like other neural network models, the SDM shares many features with human memory at the conceptual level: it is massively parallel; it is content-addressable; it degrades smoothly as storage locations are gradually removed; it can handle noisy or corrupt input patterns; and it processes high-dimensional data. In addition, information is widely distributed, and each memory location in the SDM encodes for multiple stored data patterns.

Although recent neurophysiological evidence supports the hypothesis that the cerebellum learns from experience, the cerebellum is not generally considered as a memory area of the brain. Indeed, the SDM was never intended as a biologically plausible model of short-item ('working') or long-term

memory, and the analogy should not be pushed too far. The SDM has, however, been considered alongside the theoretical work of Marr (1969) and Albus (1971) on the cerebellum. Interestingly, Kanerva (1993) has shown that the models of Marr and Albus can both be represented in terms of a slightly modified SDM architecture, and together they form the basis for a general theory of the cerebellum. (*See* **Working Memory, Computational Models of**; **Hippocampus**; **Neural Basis of Memory: Systems Level**; **Encoding and Retrieval, Neural Basis of**; **Learning and Memory, Models of**; **Semantic Memory: Computational Models**)

SUMMARY

The SDM was originally developed by Pentti Kanerva as a robust method for storing and retrieving large binary patterns. Later, the similarity between the architecture of the SDM and the cerebellar cortex was noted. There have been many applications and extensions to the original SDM, to handle nonrandom data, continuous variables, sparse connectivity, and so on; but many possible extensions and revisions remain to be explored. These include improved methods for avoiding memory saturation ('forgetting'), embedding multiple SDM architectures within a single framework, and storing spatio-temporal patterns. It is likely that the SDM will continue to inspire new memory architectures which will combine features of conventional computer and biological memories.

References

Albus J (1971) A theory of cerebellar functions. *Mathematical Biosciences* **10**: 25–61.

Chou PA (1989) The capacity of the Kanerva Associative Memory. *IEEE Transactions on Information Theory* **35**(2): 281–298.

Graham B and Willshaw D (1995) Improving recall from an associative memory. *Biological Cybernetics* **72**: 337–346.

Hely TA, Willshaw DJ and Hayes GM (1997) A new approach to Kanerva's Sparse Distributed Memory. *IEEE Transactions on Neural Networks* **8**(3): 791–794.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79**: 2554–2558.

Jaeckel LA (1989a) An alternative design for a Sparse Distributed Memory. [Technical Report RIACS TR 89.28, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Jaeckel LA (1989b) A class of designs for a Sparse Distributed Memory. [Technical Report RIACS TR 89.30, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Joglekar U (1989) *Learning to Read Aloud: A Neural Network Approach Using Sparse Distributed Memory*. Master's thesis, UC Santa Barbara. [Reprinted as RIACS Technical Report TR 89.27, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Kanerva P (1988) *Sparse Distributed Memory*. Cambridge, MA: MIT Press/Bradford Books.

Kanerva P (1993) Sparse Distributed Memory and related models. In: Hassoun MM (ed.) *Associative Neural Memories: Theory and Implementation*, chap. 3, pp. 50–76. New York, NY: Oxford University Press.

Kanerva P and Olshausen B (1989) Two-dimensional shape recognition using Sparse Distributed Memory. [Technical Report RIACS Memo 89.3, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Karlsson R (1995) A fast activation mechanism for the Kanerva SDM memory. In: *Proceedings of the 95 RWC Symposium, Tokyo, June 1995*, RWC Technical Report TR-95001, pp. 69–70.

Kristoferson J (1998) Some results on activation and scaling of Sparse Distributed Memory. In: Braga AP and Ludermir TB (eds) *SBRN '98: Proceedings Vth Brazilian Symposium on Neural Networks, Belo Horizonte, Brazil, December 1998*, vol. I, pp. 157–160. IEEE Computer Society.

Marr D (1969) A theory of cerebellar cortex. *Journal of Neurophysiology (London)* **202**: 437–470.

Miles C and Rogers D (1993) A biologically motivated associative memory architecture. [Technical Report, Department of Neurology, Baylor College of Medicine, Houston, TX, USA.]

Prager R and Fallside F (1989) The modified Kanerva model for automatic speech recognition. *Computer Speech and Language* **3**: 61–81.

Rogers D (1988) Using data tagging to improve the performance of Kanerva's Sparse Distributed Memory. [Technical report RIACS TR 88.01, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Rogers D (1989) Statistical prediction with Kanerva's Sparse Distributed Memory. [Technical Report RIACS TR 89.02, Research Institute for Advanced Computer Science, NASA Ames Research Center.]

Ryan S and Andreae J (1995) Improving the performance of Kanerva's associate memory. *IEEE Transactions on Neural Networks* **6**(1): 125–130.

Sjoedin G (1996) Getting more information out of SDM. In: von der Malsburg C, von Seelen W, Vorbruggen JC and Sendhoff B (eds), *Artificial Neural Networks: ICANN Proceedings*, pp. 477–482. Berlin: Springer.

Willshaw D, Buneman O and Longuet-Higgins H (1969) Non-holographic associative memory. *Nature* **222** 960–962.

Further Reading

Bechtel W and Abrahamsen A (1991) *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, MA and Oxford: Blackwell.

Churchland P (1989) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press/ Bradford Books.

Churchland P and Sejnowski T (1994) *The Computational Brain*. Cambridge, MA: MIT Press.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

Fausett LV (1994) *Fundamentals of Neural Networks*. Englewood Cliffs, NJ: Prentice-Hall.

Haykin S (1998) *Neural Networks: A Comprehensive Foundation*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Hertz J, Krogh A and Palmer RG (1991) *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley.

Rose S (1992) *The Making of Memory*. London: Bantam.

Schmahmann JD (ed.) (1997) *The Cerebellum and Cognition*. San Diego, CA: Academic Press.

Shepherd G (1998) *The Synaptic Organization of the Brain*, 4th edn. New York, NY: Oxford University Press.

Spatial Representation and Reasoning

Advanced article

Jerry R Hobbs, Artificial Intelligence Center, SRI International, Menlo Park, California, USA
Srini Narayanan, Artificial Intelligence Center, SRI International, Menlo Park, California, USA

CONTENTS

Spatial representation and reasoning is a way of representing the facts about spatial properties and relationships, contiguity, dimensionality, shape, and motion, in a way that is convenient for a computer to reason about them.

INTRODUCTION

A robot moving through an environment, an interface to a geographical database, and a natural language program for giving directions all need means for representing and reasoning about spatial properties and relations. These include shape, size, distance, orientation, and relative location. The most precise and highly developed system of spatial representation is the mathematics of Euclidean space. But very often this is too precise for the purposes at hand. We may not have exact information and we may not need precise answers. For example, if we are telling someone how to get to the post office, we do not need to be more precise than the streets to follow. For this reason, much work in artificial intelligence has focused on *qualitative* spatial representation and reasoning.

Indeed, studies of human problem-solving and language understanding (Talmy, 1983; Herskovits, 1986; Langacker, 1987; Lakoff, 1987; Tversky, 1993; Landau and Jackendoff, 1993) indicate that we draw fairly subtle and important qualitative spatial distinctions. It is accepted that a topological description of the environment is central to building a cognitive map which is developmentally prior to a metrical description. Infants have been shown (Landau *et al.*, 1992) to be sensitive to shape distinctions and invariants very early in childhood, suggesting that some qualitative representation of space precedes, or at least coexists, with a more metrical one.

Research in primate vision (Ungerleider and Mishkin, 1982; Milner and Goodale, 1995) further suggests the existence of two separate pathways for visual information processing: a dorsal pathway which projects from the visual cortex to the posterior parietal cortex and a ventral pathway which projects into the inferotemporal cortex. The dorsal pathway is referred to as the 'where' system and is suggested as primarily computing task-related spatial information (such as location in egocentric coordinates or hand pre-shaping for a grasping task) while the ventral or 'what' pathway is suggested as computing primarily object-related characteristics (such as shape or other visual features). While there seems to be robust evidence (including double dissociation evidence) of these two different pathways, it is also clear that there are complex interactions through multiple cortico-cortical connections between them. Based on these findings, Landau and Jackendoff (1993) suggest the existence of related subsystems in language; a 'where' system corresponding to spatial predications (such as prepositions) and a 'what' system corresponding to object naming and nouns.

From a computational perspective, the most general and most challenging problem in spatial representation and reasoning is route planning. What paths can an object of a given shape, size and location follow through an environment with obstacles of particular shapes, sizes, locations, and trajectories? Most researchers have studied restricted versions of this problem. For example, the problem may be merely to place an object within

an environment and not to move it. Or the size and shape of an object may be viewed as negligible as it moves through the environment.

Thus, the first broad problem that must be addressed is to devise ways for representing and reasoning about relations and measures between objects at a range of granularities, from the purely topological to fully metric. The second problem is how to represent and reason about shape in two and three dimensions, again at a range of granularities. Finally, there is the problem of how to represent and reason about objects of particular shapes when they are moving.

LEVELS OF STRUCTURE

Spatial and other quantities can be represented at a range of granularities. The coarsest level in common use is what is known as the 'sign algebra' (e.g. Forbus, 1984); the real line is divided into the negative numbers, zero, and the positive numbers, and what is known about a quantity is only which of the three regions it lies in. This has proven very useful in qualitative physics for reasoning about increases and decreases in quantities and direction of flow. A more refined structure can be imposed by dividing the real line into further intervals, with ordering relations between them. For example, for a pot of water sitting on a stove, we might want to distinguish the intervals between the freezing point, the boiling point, and the temperature of the stove, as well as the transition points themselves. Moreover, 'landmarks' can be set during the reasoning process. For example, to determine whether an oscillation is damping, we should compare successive maxima; these would be the landmarks (Kuipers, 1986).

Some work has been done on providing logarithmic structure to scales. This is known as order-of-magnitude reasoning (Raiman, 1991). We may not know whether Los Angeles is closer to San Diego or Santa Barbara, but we certainly know that it is closer to both than it is to Chicago, because these distances are of different orders of magnitude. Most work in order-of-magnitude reasoning has sought to exploit the fact that quantities at lower orders of magnitude can be ignored in operations involving higher orders of magnitude. Thus, we know that adding a stamp to a letter will not increase its weight enough that more postage will be required.

A very fine-grained representation of space is one that places ϵ-neighborhoods around points and considers points indistinguishable if they are within the same ϵ-neighborhood (Roberts, 1973).

More generally, the level of structure we want to view space at will depend on functionality. If we are planning a land trip and are only concerned with getting the right visas, then we can view countries as nodes in a graph, regardless of their size, where each country is connected by an arc to each of its neighbors. When we are traveling in the country, we need a finer-grained view.

DIMENSIONALITY

The most important feature that distinguishes space from other quantitative domains is the fact that it has more than one dimension. The minimal condition that is required before a notion of dimensionality makes sense is that there must be two or more scales where the order of elements on one scale cannot be determined by their order on the other scales. We cannot infer from the fact that Chicago is east of Denver which of the two is farther north. The scales need not be fully numeric. A spatially represented bar graph may have a numeric vertical dimension while its horizontal dimension is a discrete space of alphabetically ordered named entities. The dimensionality of an entity is dependent on perspective. A road, for example, may be viewed as one-dimensional when we are planning a trip, two-dimensional when we are driving on it, and three-dimensional when we hit a pothole.

When one admits several dimensions, problems arise involving objects of different dimensions. For example, can we have an object consisting of a volume and a line segment? Does the boundary of an object have a lower dimension, and is it part of the object? These problems are related to more general problems concerning open and closed sets. Galton (1996) has developed an interesting solution to these problems. In his system, an object cannot be *part* of an object of a different dimension, but objects are *bounded* by objects of lower dimension.

Where there are dimensions, there must be frames of reference for describing locations in each of the dimensions. In human cognition, as evidenced by language, there are a number of frames of reference characterized by how they are anchored (Carlson and Irwin, 1993). There is the self-anchored frame of reference, right-left-front-back, and the world-anchored frame of reference, north-south-east-west. There can be frames of reference anchored on the vehicle one is in, port-starboard-bow-stern, or one's geographical region – the Hawaiian language has prepositions for 'toward the center of the island' and 'toward the

ocean'. There can be frames of reference determined by the forces that are acting on one, such as windward-leeward and upriver-downriver. Neurobiologists have also shown the existence of multiple reference frames including ones anchored on specific joints (Rizzolatti *et al*., 1997).

REGIONS

In recent years, there has been a good deal of research in purely qualitative representations of space (Randall *et al*., 1992; Gotts *et al*., 1996; Gotts 1996; Cohn, 1997; Lemon and Pratt, 1997; Cohn and Hazarika, 2001). Much of this work attempts to build axiomatic theories of space that are predominantly topological in nature and based on taking *regions* rather than *points* as primitives. Topological relationships between regions in two-dimensional space have been formalized, with transitivity inferences, similar to those used in temporal reasoning, identified for vocabularies of relations (Randall *et al*., 1992; Cohn, 1997). The basic primitive is a notion of two regions x and y being connected if the closures of x and y share at least one point.

This primitive `connect(x,y)` has been shown to be extremely powerful and has led to the development of a rich calculus of spatial predicates and relations referred to as RCC (Region Connection Calculus). It has been shown (Gotts *et al*., 1996; Gotts, 1996) how RCC can describe and distinguish between complicated objects such as loops, figure-of-8s and doughnuts with degenerate holes. However, this expressiveness is costly and reasoning with the general RCC calculus is undecidable (Cohn, 1997; Lemon and Pratt, 1997). There has been some work on tractable subclasses. The best known is based on identifying a pairwise disjoint, jointly exhaustive set of eight spatial relations called the RCC8 calculus. The RCC8 calculus is able to describe the relations *inside (touching the boundary or not), outside, touching, overlapping* for objects describable as *regular sets*; regular sets cannot have holes. The RCC8 subset also has a well-founded semantics (unlike the RCC calculus) based on Euclidean geometry (Lemon and Pratt, 1997) and has been shown to be complete under this interpretation.

ORIENTATION

In Euclidean geometry, the representation of orientation is straightforward. There is a fixed set of axes, and a direction corresponds to a vector, normalized to length 1, from the origin to some point in the space. That vector can be specified by projecting its endpoint onto the axes.

Qualitative representations of orientation can be inherited from qualitative representations on the axes. For example, if the structure of each of the two axes in a two-dimensional system is $\{+, 0, -\}$, then the corresponding representation of orientation maps the direction into one of eight values – along one of the four axes or in one of the four quadrants.

A finer structure on the axes yields a finer structure on orientations. For example, if each axis has regions corresponding to 'slightly positive' and 'slightly negative', then it will be possible to define the notion of angles being 'slightly acute' (Liu, 1998).

SHAPE

Shape is one of the most complex phenomena that must be dealt with in the qualitative representation of space. There is a range of possibilities. In topology, a circle, a square, and an amoeba-shaped blob are all equivalent, as long as they do not have holes in them. In Euclidean geometry, a large square and the same square with a slight nick in one side are different. Shape is very important in common-sense reasoning because very often the shape of an object is functional. The shapes of objects and obstacles determine possible paths. It is important for a door and a doorframe to be the same shape, and the fact that they are the same shape is the reason doors must be open for things to pass through them.

One form of representation for shape is the use of *occupancy arrays* that encode the location of an object in a 2-D or 3-D grid (Funt, 1980; Glasgow *et al*., 1995). These representations have a number of attractive properties including ease of visualization and a natural parallel implementation of operations such as intersection, translation, and computing spatial relations between objects. However, occupancy arrays are inflexible and cannot express abstractions and partial knowledge.

Another way to characterize shapes is by the shapes of their boundaries. In one approach (Hoffman and Richards, 1982), the sides of a complex figure are classified as straight or curving in or out, and the vertices as angling in or out. Boundary information can be combined with qualitative length information so that we can distinguish between 'thin' and 'fat' rectangles, for example, or between rectangles that are wider than they are tall and ones that are taller than they are wide. Boundary information can also be combined with

theories of orientation, so that we can get qualitative measures of angles between sides.

Another representation of shape is in terms of its 'bounding box', or smallest containing quadrangle, or in terms of its convex hull. These representations are useful in moving objects through an environment. To determine whether a car will fit through a garage doorway, you need to take the side mirrors into account, but not the cavities due to open windows.

A commonly used representation for complex three-dimensional objects is the cylinder (Davis, 1990). This representation is subject to variations in granularity. For example, a human being can be represented at a very coarse grain as a cylinder. At a finer grain, the person is resolved into a cylinder for the torso, cylinders for each arm and leg, and a cylinder for the neck and head. At even finer grains, there are cylinders for the upper and lower arms or for individual fingers.

MOTION

For any quantity or quality that can be represented at an instant in time, we can also imagine it changing across time. Topological relations between entities can change as the entities move. The distance between two objects, the orientations of two lines, the shape of two objects, can all change with time.

In general, given any qualitative theory of a spatial feature and any qualitative theory of time, we can develop a qualitative theory of how the spatial feature changes with time. In most instances, this will involve transforming what is a continuous motion in Euclidean space into a motion of discrete jumps in the qualitative space, as when an entity suddenly crosses the boundary between region A and region B of a plane. Various qualitative theories impose various constraints on the possible motions, or possible changes of state. Thus, we cannot go from positive to negative without passing through zero. Representing the behavior as a transition graph (fully qualitative or with metric information attached to the nodes), allows for algorithms that generate and recognize behaviors using a variety of analytic and simulation techniques. Modeling spatially distributed motion is more complex and some efforts have attempted to use spatial aggregation based on motion field invariants and other methods from computer vision to represent qualitative structures of fluid motion (Yip, 1995) and geometric features in phase space (Zhao, 1994).

IMAGISTIC AND PROPOSITIONAL REPRESENTATIONS

There is a long-standing debate among researchers in spatial representation and reasoning about whether representations of spatial relationships should be imagistic or propositional, that is, more like pictures or more like linguistic descriptions (Pylyshyn, 1981; Shepard and Cooper, 1982; Kosslyn, 1994). Proponents of propositional representations argue that imagistic representations reduce to propositional ones, but impose constraints on what situations can be represented, which make them of very limited use (Lemon and Pratt, 1997).

On the other hand, imagistic representations are common in models of human spatial reasoning. In human communication, imagistic representations such as sketches, maps and figures are commonly used to communicate information about shape, size, routes, and spatial arrangements. Images and diagrams are also commonly used to communicate and comprehend abstract structures ranging from data structure visualizations and process flow diagrams to corporate hierarchies. Reasoning with imagistic representations has often been argued to be fundamental in human cognition (Kosslyn, 1994; Farah, 1995). It has been shown that infants, very early, perhaps pre-attentively, do grouping based on similarity of shapes, orientations and sizes (Julesz, 1984) as well as shape closure (Treisman, 1982, 1985). There is evidence that even symbolic tasks such as language understanding may use imagistic representations (Langacker, 1987; Lakoff, 1987). A computational model of the role of such representations in the acquisition of spatial prepositions can be found in Regier (1992).

There is growing evidence that acquiring, storing and reasoning with spatial concepts requires the coordinated use of heterogeneous representation and inferential processes that involve both propositional and imagistic components (qualitative and metric) (Glasgow *et al*., 1995), and much of current research is exploring how this can be accomplished computationally.

References

Carlson-Radvansky LA and Irwin DE (1993) Frames of reference in vision and language: where is above? *Cognition* **46**: 223–244.

Cohn AG (1997) *Qualitative Spatial Representation and Reasoning Techniques*. Proceedings of KI-97. Brewka G, Habel C and Nebel B (eds) LNAI, **1303**: 1–30. Springer Verlag.

Cohn AG and Hazarika SM (2001) Qualitative spatial representation and reasoning: an overview. *Fundamental Informaticae* **46**(1–2): 1–29.
Davis E (1990) *Representations of Commonsense Knowledge*. San Mateo, CA: Morgan-Kaufmann.
Farah M (1995) The neural bases of mental imagery. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 963–975. Cambridge, MA: MIT Press.
Forbus K (1984) Qualitative process theory. *Artificial Intelligence* **24**: 85–168.
Funt B (1980) Problem solving with diagrammatic representations. In: Glasgow J, Karan B and Narayanan N (eds) *Diagrammatic Reasoning*, pp. 33–38. Cambridge, MA: AAAI Press/MIT Press.
Galton A (1996) *Taking Dimension Seriously in Qualitative Spatial Reasoning*. Proceedings, 12th European Conference on Artificial Intelligence, August, Budapest, Hungary, pp. 501–505.
Glasgow J, Karan B and Narayanan N (eds) (1995) *Diagrammatic Reasoning*. Cambridge, MA: AAAI Press/MIT Press.
Gotts NM (1996) *Topology from a Single Primitive Relation: Defining Topological Properties and Relations in Terms of Connection*. Report 96.23, School of Computer Studies, University of Leeds.
Gotts NM, Gooday JM and Cohn AG (1996) A connection based approach to common-sense topological description and reasoning. *The Monist* **79**(1): 51–75.
Herskovits A (1986) *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge, UK: Cambridge University Press.
Hoffman DD and Richards WA (1982) *Representing Smooth Plane Curves for Recognition: Implications for Figure-Ground Reversal*. Proceedings, National Conference on Artificial Intelligence, August, Pittsburgh, PA, pp. 5–8.
Julesz B (1984) A brief outline of the texton theory of human vision. *Trends in Neurosciences* **7**: 41–45.
Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.
Kuipers B (1986) Qualitative simulation. *Artificial Intelligence* **29**: 289–338.
Lakoff G (1987) *Women, Fire and Dangerous Things*. Chicago, IL: University of Chicago Press.
Landau B and Jackendoff R (1993) 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences* **16**: 217–265.
Landau B, Smith L and Jones S (1992) Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language* **31**: 807–825.
Langacker RW (1987) An introduction to cognitive grammar. *Cognitive Science* **10**(1): 1–40.
Lemon O and Pratt I (1997) Spatial logic and the complexity of diagrammatic reasoning. *Machine Graphics and Vision* **6**(1): 89–109. [Special Issue on Diagrammatic Reasoning.]
Liu J (1998) A method of spatial reasoning based on qualitative trigonometry. *Artificial Intelligence Journal* **98**: 137–168.
Milner DA and Goodale MA (1995) The visual brain in action. *Oxford Psychology Series* 27. Oxford, UK: Oxford University Press.
Pylyshyn ZW (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **88**: 16–45.
Raiman O (1991) Order of magnitude reasoning. *Artificial Intelligence* **51**: 11–38.
Randall DA, Cui Z and Cohn AG (1992) *A Spatial Logic Based on Regions and Connection*. Proceedings 3rd International Conference on Knowledge Representation and Reasoning, pp. 165–176. San Mateo, CA: Morgan Kaufmann.
Regier T (1992) *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. PhD thesis, Computer Science Department, University of California, Berkeley.
Rizzolatti G, Fadiga L, Fogassi L and Gallese V (1997) The space around us. *Science* **277**: 190–191.
Roberts FS (1973) Tolerance geometry. *Notre Dame Journal of Formal Logic* **14**(1): 68–76.
Shepard RN and Cooper LA (1982) *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.
Talmy L (1983) How language structures space. In: Acredolo H and Acredolo L (eds) *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press.
Treisman A (1982) Perceptual grouping and attention in visual search for features and objects. *Journal of Experimental Psychology: Human Perception and Performance* **8**(2): 194–214.
Treisman A (1985) Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing* **31**: 156–177.
Tversky B (1993) Spatial mental models. In: Power GH (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 27. New York: Academic Press.
Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA and Mansfield RJW (eds) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.
Yip K (1995) *Reasoning about Fluid Motion I: Finding Structures*. Proceedings of IJCAI-95, pp. 1782–1788. Cambridge, MA: AAAI Press.
Zhao F (1994) Extracting and representing qualitative behaviors of complex systems in phase space. *Artificial Intelligence* **69**: 51–92.

Further Reading

Davis E (1990) *Representations of Commonsense Knowledge*. San Mateo, CA: Morgan-Kaufmann.
Stock O (ed.) (1997) *Spatial and Temporal Reasoning*. Dordrecht: Kluwer Academic.

Speech Perception and Recognition, Theories and Models of

Intermediate article

Dominic W Massaro, University of California, Santa Cruz, California, USA

CONTENTS

Multiple sources of information are exploited to perceive and understand spoken language.

INTRODUCTION

It is not unusual to have the impression that foreign languages are spoken much more rapidly than our own, and without silent periods between the words and sentences. Our own language, on the other hand, is perceived at a normal pace (or even too slowly at times) with clear periods of silence between the words and sentences. In fact, languages are spoken at approximately the same rate, and these experienced differences are solely due to the perceptual and memory structures and psychological processes involved in speech perception. Thus we define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input. The empirical and theoretical investigation of speech perception has become an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics.

PSYCHOPHYSICS OF SPEECH PERCEPTION

In any domain of perception, one goal is to determine the stimulus properties responsible for perception and recognition of the objects in that domain. The study of speech perception seems to be even more challenging than other domains of perception because there appears to be a discrepancy between the stimulus and the perceiver's experience of it. For speech, we perceive mostly a discrete auditory message composed of words, phrases, and sentences. The stimulus input for this experience, however, is a continuous stream of sound (and facial and gestural movements in face-to-face communication) resulting from the speech production. Somehow, this continuous input is transformed into a more or less meaningful sequence of discrete events.

One long-standing issue in research has been whether there is an invariance between the speech signal and its category membership. One of the most obvious perceptual categories for speech is the phoneme. Phonemes are the minimal units in speech that can change the meaning of a word. The word *ten* has three phonemes: we can change the /t/ to /d/ to make *den*, the /e/ to /æ/ to make *tan*, and the /n/ to /l/ to make *tell*. If phonemes were invariant perceptual categories, we would expect to find an orderly relationship between properties of the speech signal and phoneme categories. We would expect to find some constant characteristic in the speech signal for a given phoneme. However, this appears not to be the case. Figure 1 gives a visual representation of the sounds /di/ and /du/. Given that /d/ is the first phoneme of both sounds, we might expect to see the same signal at the beginning. However, we do not. The second visible band of energy from the bottom (the second formant) rises in /di/ and falls in /du/. One of the original arguments for the special nature of speech perception implicated this uncertain relationship between properties of the speech signal and a given phonemic category. It was emphasized that, in contrast to other domains of pattern recognition, one could not delineate a set of acoustic properties that uniquely defined a phoneme.

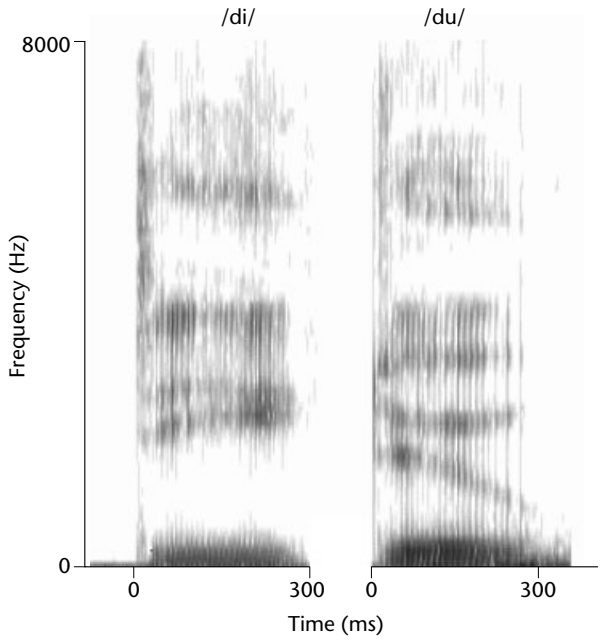


Figure 1. Spectrograms of the syllables /di/ and /du/, illustrating the lack of invariance between the acoustic signal and the phoneme. The second visible band of energy from the bottom (the second formant) rises in /di/ and falls in /du/, illustrating that the same phoneme /d/ has different acoustic characteristics in the different vowel contexts.

Not only is there a lack of invariance between the phoneme and the speech signal, but it does not at first seem to be possible to isolate the phoneme in the speech signal. Consider the syllable /da/: it has two phonemes, /d/ and /a/. If we listen to this syllable in isolation, we hear /da/. Now if we repeatedly shorten the syllable by removing short segments from the end, we should eventually hear just /d/. But this is not what happens. Our percept changes from /da/ to nonsense, not from /da/ to /d/. Therefore, some 'magic' must be involved in hearing both /d/ and /a/ given the syllable /da/.

One way to instate some systematicity between the speech signal and perception is to postulate a somewhat larger unit of speech perception. Open syllables are defined as V, VC or CV items, where V is a vowel and C can be either a single consonant or a consonant cluster. There is evidence that these items might function as units of perception. Subjects can easily identify shortened versions of V, VC and CV syllables when most of the vowel portion is eliminated. Although there are clearly contextual effects on the signal properties of these syllables, the influences are much more minor than those on the phoneme. Some support for the relative invariance of the open syllable comes from concatenative speech synthesis systems that use diphones (pairs of adjacent phonemes) rather than phonemes as units that are concatenated. Based on psychological research, synthesis would be even better if units for synthesis were open syllables.

Regardless of the units that are used for perception, there is still controversy over the ecological properties of the speech input that are actually functional in speech perception. One issue, revived by recent findings, is whether the functional properties in the signal are static or dynamic (changing with time). Static cues (such as the location of formants (bands of energy in the acoustic signal related to vocal tract configuration), the distribution of spectral noise as in the onset of *saw* and *shawl*, and the mouth shape at the onset of a segment) have been shown to be effective in influencing speech perception. Dynamic cues (such as the transition of energy between a consonant and the following vowel) have also been shown to be important. For example, recent research has shown that the second formant (F2) transition, defined as the difference between the F2 value at the onset of a consonant–vowel transition and the F2 value in the middle of the following vowel, is a reliable predictor of the category describing place of articulation (Sussman *et al.*, 1998).

Controversy arises over the type of cue on which speech perception primarily depends. For example, investigators recently isolated short segments of the speech signal and reversed the order of the speech within each segment (Saberi and Perrott, 1999). In this procedure, a sentence is divided into a sequence of successive segments of a fixed duration, such as 50 ms. Each segment is time-reversed and these new segments are recombined in their original order, without smoothing the transitions between the segments. Thus, the sentence could be described as locally time-reversed. Saberi and Perrott claimed that the speech was still intelligible when the reversed segments were relatively short (about 50 to 65 ms). Their conclusion was that our perception of speech was primarily dependent on higher-order dynamic properties rather than the short static cues assumed by most current theories. However, most successful research in psychology is better framed within the framework of *ceteris paribus* (other things being equal). There is good evidence that perceivers exploit many different cues in speech perception, and attempting to isolate a single functionally sufficient cue is futile.

There is now a large body of evidence indicating that multiple sources of information are available to support the perception, identification, and interpretation of spoken language. Auditory and visual

cues from the speaker, as well as the situational, social, and linguistic context contribute to understanding. There is an ideal experimental paradigm that allows us to determine which subset of the many potentially functional cues are actually used by human observers, and how these cues are combined to achieve speech perception (Massaro, 1998). The experiment involves the independent manipulation of two or more sources of information in a factorial or expanded factorial design. This systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro, 1998). It addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

AMBIGUITY IN SPEECH PERCEPTION

When listening to speech, we have the impression of perceiving discrete categories. This fact has contributed to the development of the popular 'categorical perception' hypothesis, that listeners can discriminate syllables only to the extent that they can recognize them as different phoneme categories. This hypothesis was quantified in order to predict discrimination performance from the identification judgments. Many researchers concluded that discrimination performance was fairly well predicted by identification. However, discrimination performance is consistently better than predicted by identification, contrary to the predictions of the categorical perception hypothesis (Massaro, 1987).

In many areas of inquiry, a new experimental paradigm enlightens our understanding by helping to resolve theoretical controversies. Rating experiments were used to determine if perceivers indeed have information about the degree of category membership. Rather than ask for categorical decisions, perceivers are asked to rate the stimulus along a continuum between two categories. A detailed quantitative analysis of the results indicated that perceivers have reliable information about the degree of category membership (Massaro and Cohen, 1983). Although communication forces us to partition the inputs into discrete categories for understanding, this does not imply that speech perception is categorical. To retrieve a toy upon request, a child might have to decide between *ball* and *doll*; however, the child can have information about the degree to which each toy was requested.

Although the categorical perception hypothesis has been refuted, it is often reinvented under new guises. Recently, the 'perceptual magnet effect' hypothesis has generated a great deal of research (Kuhl, 1991; Iverson and Kuhl, 2000). The idea is that the discriminability of a speech segment is inversely related to its category goodness. Ideal instances of a category are supposedly very difficult to distinguish from one another relative to poor instances of the category. If we understand that poor instances of one category will often tend to be at the boundary between two categories, then the perceptual magnet effect hypothesis is more or less a reformulation of the categorical perception hypothesis. That is, discrimination is predicted to be more accurate between categories than within categories. In the perceptual magnet effect framework, it is also necessary to show how discrimination is directly predicted by a measure of category goodness. We can expect category goodness to be related to identification performance. Good category instances will tend to be identified equivalently, whereas poor instances will tend to be identified as instances of different categories. Lotto *et al*. (1998) found that discriminability was not poorer for vowels with high category goodness, in contrast to the predictions of the perceptual magnet effect hypothesis. They also observed that category goodness ratings were highly context-sensitive, because they changed systematically with changes in the task context. This reliable context-sensitivity is a problem for the perceptual magnet effect hypothesis. If category goodness is functional in discrimination, it should be reasonably stable across different contexts.

It was once commonly believed that speech is perceived categorically. Current research suggests, however, that speech is perceived continuously and not categorically (Massaro, 1987, 1998). These results challenge views of language acquisition that attribute to the infant and child discrete speech categories (Eimas, 1985; Gleitman and Wanner, 1982). Most importantly, the case for the special nature of speech perception is weakened considerably, because of its dependence on the assumption of categorical perception (Liberman and Mattingly, 1985). Several neural network theories, such as single-layer perceptrons, recurrent network models, and interactive activation, have been developed to predict categorical perception (Damper, 1994; Damper and Harnad, 2000), and its probable nonexistence poses great problems for these models.

Given the existence of multiple sources of information in speech perception, each perceived

continuously, a new type of theory is needed. The theory must describe how each of the many sources of information is evaluated, how the many sources are combined or integrated, and how decisions are made. A promising theory has evolved from sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been accurately described within the fuzzy-logical model of perception (FLMP).

THE FUZZY-LOGICAL MODEL OF PERCEPTION

The three processes involved in perceptual recognition are illustrated in Figure 2. They are evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration to some response alternative. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: (1) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives; (2) the sources of information are evaluated independently of one another; (3) the sources are integrated to provide an overall continuous degree of support for each alternative; and (4) perceptual identification and interpretation follows the relative degree of support among the alternatives. The FLMP appears to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and no one source specifies completely the appropriate interpretation.

MULTIMODAL SPEECH PERCEPTION

Speech perception has traditionally been viewed as a unimodal process, but in fact appears to be multimodal. This is best seen in face-to-face communication. Experiments have shown conclusively that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as by the actual sound of the speech (Massaro, 1998). Consider a simple syllable

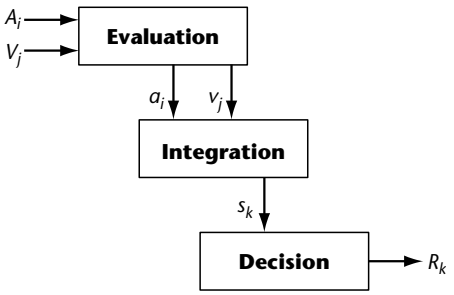


Figure 2. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown proceeding from left to right (in time) to illustrate their successive but overlapping operation. These processes make use of prototypes stored in long-term memory. Sources of information are represented by upper-case letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lower-case letters a_i and v_j). These sources are then integrated to give an overall degree of support s_k for each speech alternative k. The decision operation maps the outputs of integration into some response alternative R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

identification task. Synthetic visible speech and natural audible speech were used to generate the consonant-vowel syllables /ba/, /va/, /ða/, and /da/. Using an expanded factorial design, the four syllables were presented audibly, visibly, and bimodally. Each syllable was presented alone in each modality for $4 \times 2 = 8$ unimodal trials. For the bimodal presentation, each audible syllable was presented with each visible syllable for a total of $4 \times 4 = 16$ unique trials. Thus, there were 24 types of trial. Of the bimodal syllables, 12 had inconsistent auditory and visual information. The 20 participants in the experiment were instructed to watch and listen to the talking head and to indicate the syllable that was spoken.

Figure 3 shows the accuracy for unimodal and bimodal trials when the two syllables were consistent with one another. Performance was more accurate given two consistent sources of information than given either one presented alone. Consistent auditory information improved visual performance about as much as consistent visual information improved auditory performance. Given inconsistent information from the two sources, performance was poorer than observed in the unimodal conditions. These results show a strong influence of both modalities on performance, with a stronger influence from the auditory than from the visual source of information.

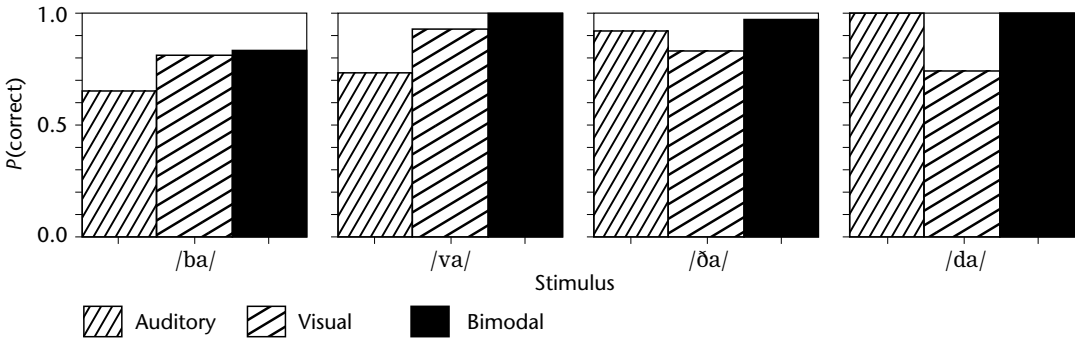


Figure 3. Probabilities of correct identification of syllables in unimodal (auditory and visual) and bimodal consistent trials for the four test syllables /ba/, /va/, /ða/ and /da/.

Although the results demonstrate that perceivers use both auditory and visible cues in speech perception, they do not indicate how the two sources are used together. There are many possible ways the two sources might be used. We first consider the predictions of the FLMP.

In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to the feature value. For bimodal trials, the predicted probability of a response $P(/\mathrm{da}/)$ is

$$P(/\mathrm{da}/) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)} \quad (1)$$

In previous work, the FLMP has been contrasted against several alternative models, including a weighted averaging model (WTAV), which is an inefficient algorithm for combining the auditory and visual sources. For bimodal trials, the predicted probability of a response $P(/\mathrm{da}/)$ is

$$P(/\mathrm{da}/) = \frac{w_1 a_i + w_2 v_j}{w_1 + w_2} = w a_i + (1 - w) v_j \quad (2)$$

where W_1 and W_2 are the weights and $W = \frac{W}{W_1 + W_2}$.

The WTAV predicts that two sources can never be more informative than one. In direct comparisons, the FLMP has consistently and significantly outperformed the WTAV (Massaro, 1998).

Furthermore, the results are well described by the FLMP, an optimal integration of the two sources of information (Massaro and Stork, 1998). A perceiver's recognition of an auditory–visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *my bab pop me poo brive* is paired with the visible sentence *my gag kok me koo grive* the perceiver is likely to hear *my dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro and Stork, 1998).

Recent findings show that speechreading, or the ability to obtain speech information from the face, is not compromised by oblique views, partial obstruction or visual distance. Humans are fairly good at speechreading even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998).

CONTEXTUAL, HIGHER-ORDER, OR TOP-DOWN INFLUENCES

There is now a substantial body of research showing that speech perception is influenced by a variety of contextual sources of information. Bottom-up sources have a direct mapping between the sensory input and the representational unit in question. Top-down sources, or contextual information, come from constraints that are not directly mapped to the unit in question. An example of a bottom-up source would be the stimulus presentation of a test word after the presentation of a top-down source, a sentence context. A critical question for both integration and autonomous (modularity) models is how bottom-up and top-down sources of information work together to achieve word recognition. For example, an important question is how early contextual information can be integrated with acoustic and phonetic information. A large body of research shows that several bottom-up sources are evaluated in parallel and integrated to achieve recognition (Massaro, 1987, 1994). Another important question is whether

top-down and bottom-up sources are processed in the same manner. A critical characteristic of autonomous models might be described as the language user's inability to integrate bottom-up and top-down information. An autonomous model must necessarily predict no perceptual integration of top-down with bottom-up information.

Pitt (1995) studied the joint influence of phonological information and lexical context in an experimental paradigm developed by Ganong (1980). A speech continuum is made between two alternatives, and the contextual information supports one alternative or the other. The initial consonant of a CVC syllable was varied, taking six values between /g/ and /k/ inclusive. The following context was either /Ift/ or /Is/. The context /Ift/ supports initial /g/ because *gift* is a word whereas *kift* is not. Similarly, the context /Is/ supports initial /k/ because *kiss* is a word whereas *giss* is not.

The lines in Figure 4 give the predictions of the FLMP. The model generally provides a good description of the results of this study. For each of these individuals, the model captures the observed interaction between phonological information and lexical context: the effect of context was greater to the extent that the phonological information was ambiguous.

The model tests establish that perceivers integrate top-down and bottom-up information in language processing, as described by the FLMP. This means that sensory information and context are integrated in the same manner as are several sources of bottom-up information. These results pose problems for autonomous models of language processing.

CONNECTIONIST MODELS

In connectionist models, information processing occurs through excitatory and inhibitory interactions

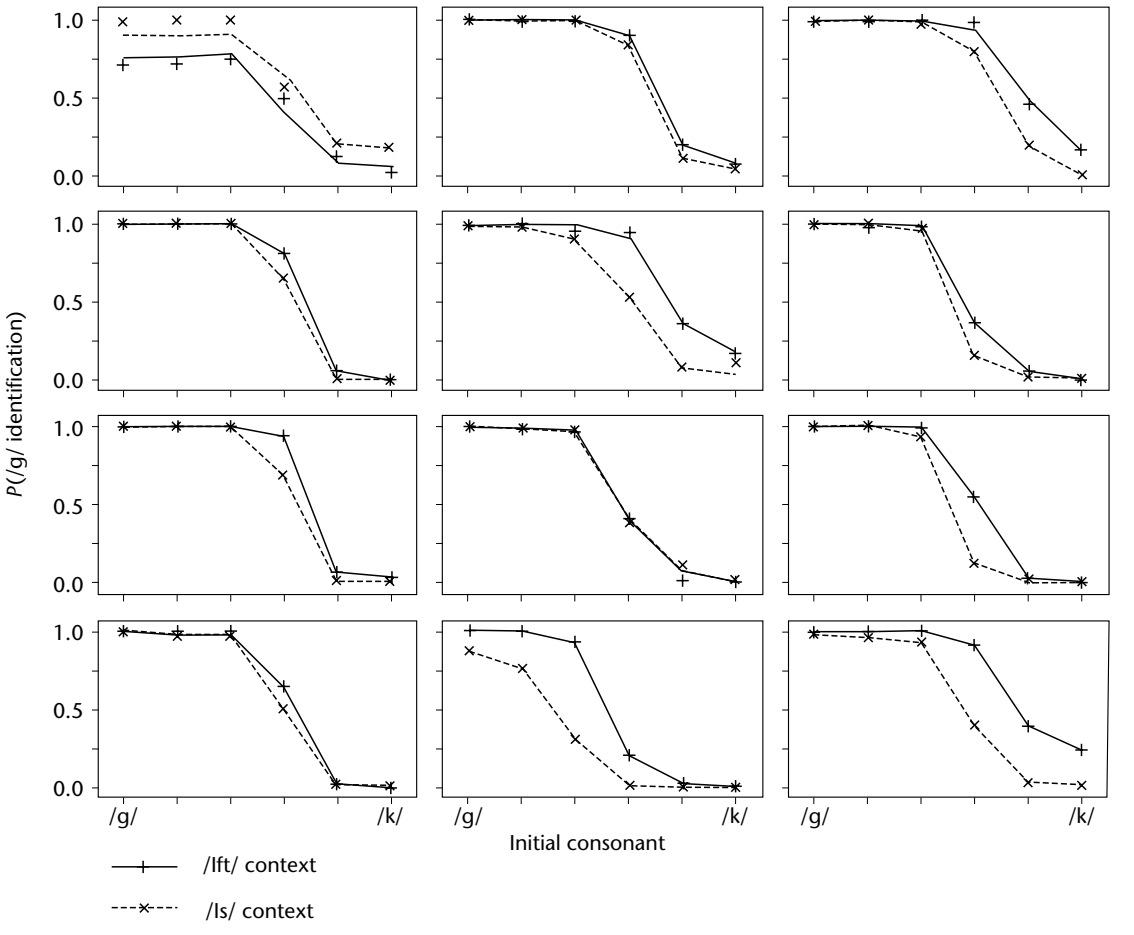


Figure 4. Observed probabilities (points), and FLMP's predictions (lines), of /g/ identifications for /Ift/ and /Is/ contexts as a function of the speech information of the initial consonant which ranges from /g/ to /k/. Results for 12 subjects from Pitt (1995).

among a large number of simple processing units. These units are meant to represent the functional properties of neurons or neural networks. Three levels or sizes of units are used in the TRACE model of speech perception (McClelland and Elman, 1986; McClelland, 1991): feature, phoneme, and word. Features activate phonemes which activate words, and activation of some units at a given level inhibits other units at the same level. In addition, an important assumption of interactive activation models is that activation of higher-order units activates their lower-order units; for example, activation of the /b/ phoneme would activate the features that are consistent with that phoneme.

In the TRACE model, word recognition is mediated by feature and phoneme recognition. The input is processed online in TRACE, all words are activated by the input in parallel, and their activation is context-dependent. In principle, TRACE is continuous, but its assumption about interactive activation leads to categorical-like behavior at the sensory (featural) level. In the TRACE model, a stimulus pattern is presented and activation of the corresponding features sends more excitation to some phoneme units than others. Given the assumption of feedback from the phoneme to the feature level, the activation of a particular phoneme feeds back and activates the features corresponding to that phoneme (McClelland and Elman, 1986, p. 47). This effect enhances sensitivity around category boundaries, exactly as predicted by the categorical perception hypothesis. Evidence against phonemes as perceptual units and against the categorical perception hypothesis is, therefore, evidence against the TRACE model.

PHONEMIC RESTORATION

In the original type of phonemic restoration study (Warren, 1970), a phoneme in a word is removed and replaced with some other stimulus, such as a tone or white noise. Subjects perceive the word as relatively intact and have difficulty indicating what phoneme is missing. This illusion has been taken to support interactive activation: the lexical information supposedly modifies the sublexical auditory representation. This interpretation contrasts with that given by the FLMP, in which the lexical context simply provides an additional source of information. There is no reason to assume that the auditory representation was modified. In terms of signal detection analysis, the lexical influence is one of bias and not sensitivity. Several experimenters have addressed this issue. Although Samuel (1981) concluded that his results favored interactive activation, the changes in top-down context were confused with different stimulus conditions (Massaro, 1989).

Repp (1992) carried out a systematic set of experiments to clarify the locus of the phonemic restoration phenomenon. Essentially, he asked the question whether the restoration was localized at the auditory level or at a higher linguistic level. Within the framework of the FLMP, the effect is attributed to the independent contribution of lexical constraints, with no top-down influence on the auditory level. Subjects were asked to judge the perceived pitch (timbre, brightness) of the extraneous noise that replaced the speech sound. If restoration involved auditory processing, we would expect these pitch judgments to be influenced by lexical context. The overall results of five experiments were negative: auditory effects were not observed, even though phonemic restoration did occur. Repp concludes that perceivers' reports of what they are hearing cannot be taken as an index of auditory process or representation (the McGurk effect). A positive influence of a top-down constraint on perceptual report does not necessarily mean that the representation of the bottom-up influences has been modified. Thus there is no evidence for top-down activation of a lexical representation on a lower-level auditory representation.

HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs) are used in speech recognition when we conceptualize the signal in speech as a sequence of states (Rabiner, 1989; Jurafsky and Martin, 2000). The goal is to determine which model best accounts for the observed sequence of states. For example, the states might correspond to a sequence of phonemes or to a sequence of smaller segments, and the goal is to determine which word best accounts for the observed sequence. A pedagogical example involves an observation of a series of coin tosses, such as HHTHHTTTH. We know there are many possible models that could generate these observed results. The most obvious (and parsimonious) model would be to assume that only a single (possibly biased) coin was being tossed. This is an observable model, and the only parameter needed to completely specificy the model is the bias value. In another model, one could assume that two different coins are being tossed. In this case, it is necessary to specify both the biases of the two coins and how the system moves from state to state (coin to coin). A

third model assumes three biased coins, and choosing among the three on the basis of some probabilistic event. The three models vary in the number of free parameters: the first model requires only a single free parameter; the second model requires four free parameters; and the third model requires nine free parameters. Given that the goal is simply speech recognition accuracy rather than psychological validity, it is not a problem that larger HMMs will necessarily provide a better description of the observed sequence of events. The only limitation on the size of the HMMs is computational complexity.

The three basic problems to solve given HMMs are: (1) computing the probability of an observed sequence for a given model; (2) choosing a state sequence that is optimal given the observation sequence and the model; and (3) determining how the model's parameters are adjusted to maximize the probability of an observed sequence for a given model. What are the fundamental assumptions in using HMMs? It is assumed that the observations are independent, so that the sequence of observations can be written as a product of individual observations. Furthermore, the probability of being in state t depends only on the state at time $t-1$. Finally, the distributions of individual observation parameters can be represented by a mixture of Gaussian or autoregressive densities.

HMMs have found limited application in psychology perhaps because they are not easily applied to empirical studies of speech perception, they are not grounded in psychological processes, and they may not be falsifiable.

CONCLUSION

The study of speech perception is an interdisciplinary endeavor, which involves a varied set of experimental and theoretical approaches. It includes the fundamental psychophysical question of what properties of spoken language are perceptually meaningful and how these properties signal the message. Independent variation of several properties, along with a quantitative theoretical analysis, is a productive paradigm to address not only the psychophysical question but also the issue of how multiple cues are used together for perception and understanding. Spoken language consists of multiple sources of information from several modalities, and people have continuous information from these many sources. In addition to these bottom-up sources, higher-order context is exploited in speech processing. There are several productive theoretical approaches to the complex question of how we so easily communicate with one another.

Acknowledgement

This work was supported in part by NSF CHALLENGE grant CDA-9726363, Public Health Service grant PHS R01 DC00236, National Science Foundation grant 23818, Intel Corporation, and the University of California Digital Media Innovation Program.

References

Damper RI (1994) Connectionist models of categorical perception of speech. In: *Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks*, vol. I, pp. 101–104. Hong Kong.

Damper RI and Harnad SR (2000) Neural network models of categorical perception. *Perception and Psychophysics* **62**(4): 843–867.

Eimas PD (1985) The perception of speech in early infancy. *Scientific American*: **252**: 46–52.

Ganong WF (1980) Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance* **6**: 110–125.

Gleitman LR and Wanner E (1982) Language acquisition: the state of the state of the art. In: Wanner E and Gleitman LR (eds) *Language Acquisition: The State of the Art*, pp. 3–48. Cambridge, UK: Cambridge University Press.

Iverson P and Kuhl PK (2000) Perceptual magnet and phoneme boundary effects in speech perception: do they arise from a common mechanism? *Perception and Psychophysics* **62**(4): 874–886.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kuhl PK (1991) Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perceptual Psychophysics* **50**: 93–107.

Liberman AM and Mattingly IG (1985) The motor theory of speech perception revised. *Cognition* **21**: 1–36.

Lotto AJ, Kluender KR and Holt LL (1998) Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* **103**: 3648–3655.

Massaro DW (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum.

Massaro DW (1989) Testing between the TRACE model and the Fuzzy Logical Model of speech perception. *Cognitive Psychology* **21**: 398–421.

Massaro DW (1994) Psychological aspects of speech perception: implications for research and theory. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*, pp. 219–263. New York, NY: Academic Press.

Massaro DW (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
Massaro DW and Cohen MM (1983) Categorical or continuous speech perception: a new test. *Speech Communication* **2**: 15–35.
Massaro DW and Stork DG (1998) Sensory integration and speechreading by humans and machines. *American Scientist* **86**: 236–244.
McClelland JL (1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* **23**: 1–44.
McClelland JL and Elman JL (1986) The TRACE model of speech perception. *Cognitive Psychology* **18**: 1–86.
Pitt MA (1995) The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory and Cognition* **21**: 1037–1052.
Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**: 257–286.
Repp BH (1992) Perceptual restoration of a 'missing' speech sound: auditory induction or illusion? *Perception and Psychophysics* **51**: 14–32
Saberi K and Perrott DR (1999) Cognitive restoration of reversed speech. *Nature* **398**: 760.
Samuel AG (1981) Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General* **110**: 474–494.
Sussman HM, Fruchter D, Hilbert J and Sirosh J (1998) Linear correlates in the speech signal: the orderly output constraint. *Behavioral and Brain Sciences* **21**: 241–299.
Warren RM (1970) Perceptual restoration of missing speech sounds. *Science* **167**(3917): 392–393.

Further Reading

Allport DA, MacKay DG, Prinz W and Scheerer E (eds) (1987) *Language Perception and Production: Shared Mechanisms in Listening, Speaking, Reading and Writing*. London: Academic Press.
Altmann GTM (ed.) (1990) *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.
Campbell R, Dodd B and Burnham D (eds) (1998) *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*. Hove, UK: Psychology Press.
Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.
Massaro DW (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
Schouten MEH (ed) (1992) *The Auditory Processing of Speech*. Berlin: Mouton de Gruyter.
Speech Processing: Perception, Analysis, and Synthesis. [http://www.ccp.uchicago.edu/overview/language/speech/]
Speech Research. [http://mambo.ucsc.edu/psl/speech.html]
Tohkura Y, Vatikiotis-Bateson E and Sagisaka Y (eds) (1992) *Speech Perception, Production and Linguistic Structure*. Tokyo: Ohmsha.
UCLA Speech Processing and Auditory Perception Laboratory. [http://www.icsl.ucla.edu/~spapl/]

Spreading-activation Networks

Introductory article

Lokendra Shastri, International Computer Science Institute, Berkeley, California, USA

CONTENTS

Spreading-activation networks are networks of 'active' nodes that compute by propagating activation levels along links. Such network models are inspired by the idea that meaning can be expressed via associations between concepts. They have been used to model cognitive phenomena ranging from low-level vision to common-sense reasoning and language understanding.

INTRODUCTION

Spreading-activation networks derive their inspiration from associationism – the idea that meaning can be expressed via associations between concepts. This idea can be traced through the writings of John Locke, David Hume, William James, and Donald Hebb. But the notion of associative representations came into its own when modern computers allowed it to be articulated and realized in the form of computational models. Today, spreading-activation network models exist for the full range of cognitive phenomena: from low-level vision to reasoning and language understanding.

Much of the early work on spreading-activation network models was motivated by a desire to model common-sense knowledge and common-sense reasoning. A simple example from this domain will illustrate certain aspects of spreading-activation networks. Figure 1 represents a semantic memory encoding the following knowledge: Sylvester is a cat, Tweety is a canary, canaries are birds, cats and birds are animals, canaries are yellow, cats move around by walking, and birds move around by flying. Note that concepts (instances, categories, and attribute values) are represented as nodes, and relations between concepts are encoded by labeled links. For example, `Sylvester` is linked to `Cat` by a link labeled `instance-of`, indicating that Sylvester is a cat; `Cat` is linked to `Animal` by a link labeled `is-a`, indicating that cats are animals, and `Cat` is linked to `Walking` by a link labeled `has-mode-of-transportation`, indicating that cats move around by walking.

The network depicted in Figure 1 may be viewed as a graphical notation (or language) for expressing knowledge. In this view, the network is a passive data structure, and the retrieval of knowledge encoded in the network involves the traversal of this data structure by a program (or interpreter). For example, to answer the question 'What is the mode of transportation of *X*?' (where *X* is a concept), the interpreter examines the links emanating from *X* and looks for a link labeled `has-mode-of-transportation`. If it finds such a link it follows it and retrieves the answer from the node at the end of the link. If it does not find such a link, it traverses the `instance-of` or `is-a` links emanating from *X*, and looks for a `has-mode-of-transportation` link emanating from concepts reached from *X* by traversing these links. For example, to answer the question 'What is the mode of transportation of Tweety', the program visits `Tweety`, `Canary` and `Bird`, and finds the answer `Flying` at the end of the `has-mode-of-transportation` link emanating from `Bird`.

Similarly, to answer the question 'What is yellow and flies?', the interpreter marks the node `Yellow` with a tag `a`, and the node `Flying` with a different tag `b`. Next, it traverses all links that terminate at `Yellow` and have the label `has-color`, and marks nodes at the tails of these links with the tag `a`. Similarly, it traverses all links that terminate at `Flying` and have the label `has-mode-of-transportation`, and marks nodes at the tails of these links with the tag `b`. Thereafter, it propagates tags `a` and `b` by traversing `is-a` and `instance-of` links in the reverse direction, and marks nodes at the tails of these links with the tags marking the heads of these links. Any node marked by both `a`

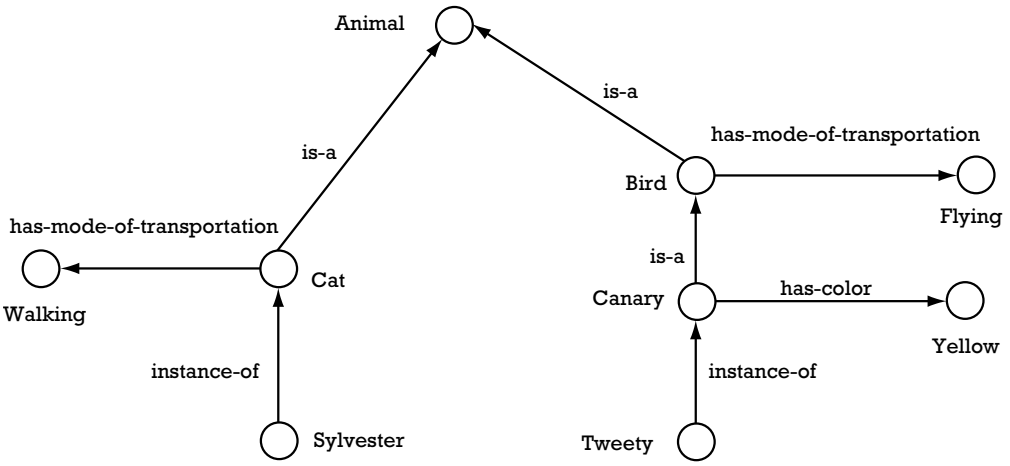


Figure 1. A graphical representation of a small domain of semantic knowledge.

and `b` tags constitutes an answer to the question. Given the domain depicted in Figure 1, the interpreter will identify `Canary` and `Tweety` as the answers to the question 'What is yellow and flies?'.

Note that in finding answers to the two questions discussed above, the system essentially computes set intersections and transitive closures of `is-a` and `instance-of` relations (and their inverses) by traversing a network.

SPREADING ACTIVATION AS A BASIS OF RETRIEVAL AND INFERENCE

A little reflection suggests a more efficient way of realizing the retrieval process described above. In this realization, each node in the graph corresponds to an active processing unit, and each link in the graph corresponds to a physical connection between nodes. A node receives messages along incoming connections, performs some simple operations on the received messages, and propagates the outcome of these operations as messages along outgoing connections. Such a device could perform the sorts of retrieval operations discussed above by a parallel propagation of messages among active processors.

For such a parallel realization to be effective, however, the messages exchanged by nodes and the operations performed by nodes must be simple. This preference for simplicity is motivated by basic computational considerations as well as by the desire to fashion nodes and links after neurons and synapses. (*See* **Connectionism**)

One can trace the idea of such parallel spreading-activation networks to a model of semantic memory proposed by M. Ross Quillian in 1968. Quillian's model represents 'word concepts' in terms of their associations with other word concepts, and finds common properties of two word concepts A and B by propagating discrete 'activation tags' starting at A and B and identifying nodes where tags originating from A and B intersect. Since Quillian's work on semantic memory, spreading-activation networks have been proposed for a wider range of cognitive tasks including semantic memory, word-sense disambiguation, language understanding, speech recognition, speech production, visual object recognition, reading, inference, and problem-solving. (*See* **ACT**; **Semantic Networks**)

Spreading-activation models can be broadly classified into 'marker-passing' networks and 'activation-passing' networks. In marker-passing networks, nodes propagate and store discrete markers, and perform Boolean operations on stored markers. In activation-passing networks, nodes propagate graded levels of activity, and links may have weights associated with them. These weights can be positive (excitatory) or negative (inhibitory). A node typically computes the weighted sum of its inputs (i.e., it multiplies the weight of each incoming link by the output of the node at the source of the link, and computes the sum over all incoming links), and produces an output activity determined by a simple, though often nonlinear, function of that weighted sum. Two such functions that are commonly used are the step function (output activity equals 1 if the weighted sum of inputs exceeds a threshold and zero otherwise) and the sigmoid function (output activity is a monotonically increasing function of the weighted sum of inputs and varies smoothly between -1 and $+1$ in the shape of an S). A node's output function in an

activity-passing system can be viewed as an 'evidence combination function' which performs a graded integration of evidence incident on the node.

MARKER-PASSING NETWORKS

The NETL system proposed by Scott Fahlman in 1979 is an example of a marker-passing system. It consists of a parallel network of simple processors and a serial central computer that controls the operation of the parallel network. Each node can store a small number of markers and perform Boolean operations on stored markers in response to commands issued by the central controller. Nodes can also propagate markers along links in parallel under the supervision of the central controller. Thus NETL performs parallel marker passing and effectively computes transitive closures and set intersections. Consequently, it can rapidly answer inheritance questions (such as 'Do sparrows fly?') and categorization questions (such as 'Which animal lives in Africa and has large ears?').

One limitation of a marker-passing network such as NETL is that markers and operations on markers are discrete ('yes or no'). Consequently, such a network cannot support evidential (or probabilistic) inference. In particular, such a network has no notion of 'best match' or of 'partial match'. Finding a concept that matches a given description amounts to finding a concept that has all the properties specified in the description.

A different type of marker-passing network has been proposed by Engene Charniak and James Hendler. These networks use complex messages that also encode information useful for controlling marker propagation. Typically, such a message contains path information, which can be stored at a node and used to detect loops, extract paths traced by a marker, and evaluate the relevance of these paths. Consequently, nodes in such networks are fairly complex.

ACTIVATION-PASSING NETWORKS

An early activation-passing network incorporating parallelism, graded activation, and evidence combination is the 'interactive activation model' proposed by James McClelland and David Rumelhart in 1981 for the perception of words and letters in visually presented stimuli.

The model consists of three layers of nodes, corresponding to visual letter features, letters, and words (see Figure 2). The letter features consist of horizontal, vertical, and diagonal bars similar to the ones used in some electronic alphanumeric displays. Feature and letter nodes are replicated four times, once for each position in a word, to enable the network to represent words of length up to four. The activity level of a node indicates the belief in the hypothesis that the node stands for. For example, the activation level of the node standing for the letter T in position 2 indicates the network's belief that the presented stimulus has a T in its second position.

Nodes representing mutually consistent hypotheses support one another, and hence are connected via excitatory links. Thus a feature node is connected via excitatory links to letter nodes representing letters that contain that feature. Similarly, a letter node (e.g. T in position 1) is connected via excitatory links to word nodes representing words containing that letter in the appropriate position (e.g. 'TAKE').

Nodes representing mutually inconsistent hypotheses are connected via inhibitory links. Thus all word nodes inhibit one another, all letter nodes for a given position inhibit one another, a feature node inhibits all letter nodes representing letters that do not contain that feature, and a letter node inhibits all word nodes representing words that do not contain that letter in the appropriate position.

Additionally, there are reciprocal (feedback) connections from a word node to all letter nodes that comprise the word (e.g. from 'TAKE' to the node for T in position 1, the node for A in position 2, and so on). These connections allow bottom-up perceptual processing to be reinforced by top-down expectations.

A visual input is presented to the model by activating feature nodes corresponding to it. The activation emanating from feature nodes excites and inhibits appropriate letter nodes. Activated letter nodes inhibit other letter nodes, and excite and inhibit appropriate word nodes. Activated word nodes, in turn, inhibit other word nodes, and excite appropriate letter nodes. Thus activation propagates through the network and results in a consistent set of letter nodes and a word node becoming highly active and forming a 'stable coalition'.

The model supports perceptual completion and disambiguation. For example, given an input consisting of the letters W, O, and R in positions 1, 2, and 3 respectively, and a partial input in position 4 consistent with both R and K, the network determines that the ambiguous letter in position 4 is K, and the word being presented is 'WORK'. An informal explanation of how this happens is as

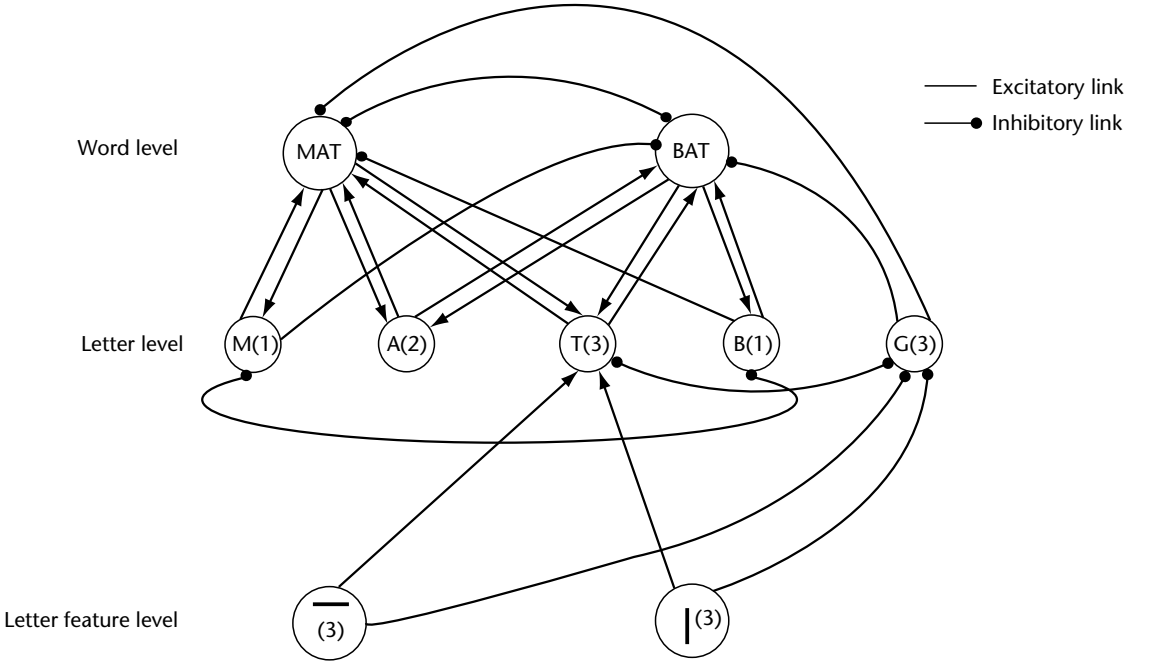


Figure 2. A fragment of a network showing the connectivity of the interactive activation model for word perception. The numbers in parentheses indicate the position of a letter or feature within an input word. The feature nodes depict the visual features of the letter T.

follows. The input provided to the feature nodes leads to the strong activation of letter nodes W, O, and R in positions 1, 2, and 3 respectively, and the weak activation of letter nodes K and R in position 4. All other letter nodes are inhibited. The activity of letter nodes leads to a strong activation of the 'WORK' node since this is the only word consistent with the ongoing activity of letter nodes. The 'WORK' node therefore dominates the activity in the word level, and sends excitatory feedback to letter nodes W, O, R, and K in the appropriate positions. This additional activity received by K in position 4 raises its activation level above that of R in position 4. Given the inhibitory link from R to K, the stronger activity of K suppresses the activity of R, resulting in the ambiguous and partial input being interpreted by the system as 'WORK'.

The model also explains psychological findings such as the ease of processing words and pronounceable non-words compared with other non-words, and the 'word superiority effect', whereby subjects are better at detecting a letter when the letter is presented as part of a word than when the letter is presented in isolation.

Applications to Language and Speech

The spreading-activation networks described above deal only with static inputs: the complete input is available at one time, and stays constant during processing. The treatment of time-varying signals is problematic in such a network. The TRACE model of McClelland and Jeffrey Elman shows how spreading-activation networks can be extended to process a time-varying signal such as speech. The model consists of three layers, corresponding to acoustic features, phonemes, and words. Feature and phoneme nodes are replicated several times to account for their occurrences in different temporal positions within an utterance. Moreover, since phonemes and words extend over time, inputs to phoneme and word nodes span multiple time slices.

Most words have multiple senses, but we are able to exploit contextual and syntactic (e.g. word-order) information to disambiguate them. Spreading-activation networks have modeled such disambiguation. For example, David Waltz and Jordan Pollack have developed a network that disambiguates between the 'celestial object' and 'famous personality' senses of the word 'star' in 'the astronomer married the star'. Similarly, Garrison Cottrell and Steven Small's network assigns correct senses to 'ball' and 'threw' in the sentences 'John threw a ball' and 'John threw a ball for charity'. These networks incorporate several representational levels, such as word, word-sense, and

thematic-role levels, and excitatory and inhibitory interconnections to express support and conflict.

Gary Dell has proposed a spreading-activation network for converting a sequence of words into a sequence of phonemes. The network consists of a word layer and a phoneme layer. Each word node has excitatory connections to nodes representing its phonemes, and each phoneme node has excitatory connections to all word nodes containing that phoneme. The word node corresponding to the current word in a sequence is activated to a high level, and word nodes corresponding to upcoming words in the same phrase as the current word are activated to a low level. This leads to the formation of a stable coalition of mutually consistent word and phoneme nodes. This network explains empirical findings about speech errors such as phoneme anticipation errors (e.g. pronouncing 'deal back' as 'beal back'), phoneme exchange errors (e.g. pronouncing 'deal back' as 'beal dack'), and the lexical bias effect (e.g. pronouncing 'dean bad' as 'bean bad' is more likely than pronouncing 'deal back' as 'beal back' because both 'bean' and 'bad' are words, whereas 'beal' is not a word).

NETWORKS WITH DYNAMIC BINDINGS

Consider the following simple narrative: 'John fell in the hallway. Tom had cleaned it. He got hurt.' Upon hearing this narrative, most of us would infer that Tom had cleaned the hallway, John fell because he slipped on the wet hallway floor, and John got hurt because of the fall. These inferences allow us to establish causal and referential coherence among the events and entities involved in the narrative. They help us to explain John's fall by means of plausible inferences, that the hallway floor was wet as a result of Tom cleaning the hallway, and that John fell because he slipped on the wet floor. They help us to causally link John's hurt to his fall, and help us to determine that 'it' in the second sentence refers to the hallway, and 'he' in the third sentence refers to John and not to Tom. Our ability to draw such inferences during language understanding suggests that we are capable of performing a wide range of inferences rapidly, spontaneously, and without conscious effort. This remarkable human ability poses a challenge for cognitive science: How can the brain represent a large body of common-sense knowledge and perform a wide range of inferences with the requisite speed?

The sort of reasoning described above is particularly difficult to model using spreading-activation networks because it requires the representation of relational structures and a solution to the binding problem. Consider the representation of the event 'John gave Mary a book'. Call this event *E*. *E* is a specific instance of a relation in which John, Mary, and book play specific roles. It cannot be represented by simply activating the conceptual roles 'giver', 'recipient', and 'given object' and the entities 'John', 'Mary', and 'book'. Such a representation would be identical to that of 'Mary gave John a book'. An unambiguous representation of *E* requires the representation of bindings between the roles of *E* (e.g. 'giver') and the entities that fill these roles in the event (e.g. 'John'). Furthermore, inferring that Mary has the book as a result of John giving the book to her involves rapidly establishing additional bindings between the role 'owner' and 'Mary' and between the role 'owned object' and 'book'. Thus reasoning about events and situations also requires a solution to the problem of propagating bindings.

None of the spreading-activation models discussed above solves the dynamic binding problem. For example, the TRACE model requires multiple banks of letter (or phoneme) nodes because it cannot dynamically bind a letter (or phoneme) to a position.

It is straightforward to represent role–entity bindings using additional nodes and links (e.g., using 'binder' nodes encoding conjunctions of roles and entities). But while it is feasible to use additional nodes and links to encode long-term knowledge, it is implausible that binder nodes could be recruited for representing large numbers of dynamic bindings arising rapidly during language understanding and visual processing. In standard computing, bindings are expressed using variables and pointers, but these techniques have no direct analogue in spreading-activation networks.

Activation-passing network modelers have devised several solutions to the binding problem. One solution, proposed by John Barnden and Kankanahalli Srinivas, makes use of the relative positions of active nodes and the similarity of their firing patterns to encode bindings. Another solution, proposed by Trent Lange and Michael Dyer and by Ron Sun, assigns a distinct activation pattern (a signature) to each entity, and propagates these signatures to establish role–entity bindings. In 1989, Venkat Ajjanagadde and Lokendra Shastri proposed a biologically plausible solution for expressing and propagating role–entity bindings. They suggested that, for example, the role–entity binding `John=giver` be expressed by the synchronous firing of `John` and `giver` nodes. The

use of synchrony for binding perceptual features during visual processing had been suggested earlier by Christoph von der Malsburg, but the Shruti model of Shastri and Ajjanagadde offered a detailed account of how synchronous activity can be harnessed to represent complex relational knowledge, and to carry out rapid inference with respect to such knowledge.

The Shruti model encodes relational knowledge using 'focal node clusters' (see Figure 3). The focal cluster of a relation such as `fall` consists of a `+:fall` node, whose activity indicates that the network is making an assertion about a fall, a `?:fall` node, whose activity indicates that the network is seeking an explanation about a fall, and role nodes – one for each role associated with `fall`. Links between the '+' and '?' nodes of a focal cluster enable the network to automatically seek explanations for active assertions. The focal cluster of an entity also contains a '+' node and a '?' node.

The event 'John fell in the hallway' is represented by the synchronous firing of the nodes `+:John` and `fall-patient`, the synchronous firing of the nodes `+:Hallway` and `fall-location`, and the sustained firing of the node `+:fall`.

From Simple Associations to Systematic Reasoning

A systematic mapping between relations (and other rule-like knowledge) is encoded by high-efficacy links between focal clusters (see Figure 3). For example, the rule-like knowledge 'If one slips, one is likely to fall' is encoded by bidirectional links between the patient roles of `slip` and `fall` and between the object roles of `slip` and `fall`, and by directed links from `+:slip` to `+:fall` and from `?:fall` to `?:slip`. Since each link has a weight associated with it, and since each node uses a graded activation combination function to generate its output based on its inputs, Shruti can encode evidential rules and perform evidential (probabilistic) reasoning.

Given the connectivity shown in Figure 3, the activity encoding 'John slipped in the hallway' automatically evolves so that `+:fall` starts firing, the 'patient' role of `fall` starts firing in synchrony with the 'patient' role of `slip` (and hence, with `+:John`), and the 'object' role of `fall` starts firing in synchrony with the 'object' role of `slip` (and hence, with `+:Hallway`). The resulting pattern of activity encodes not only 'John slipped in the hallway' but also the inferred event 'John fell in the hallway' (see Figure 4). Similarly, the pattern of activity encoding 'John fell in the hallway' automatically evolves to a pattern of activity that also encodes 'Did John slip in the hallway?'.

Inference in Shruti occurs automatically as a result of the propagation of rhythmic activity across interconnected focal clusters. There is no interpreter or central controller that manipulates symbols or applies rules of inference. The network encoding is best viewed as an internal model of the environment. When the nodes in this model are

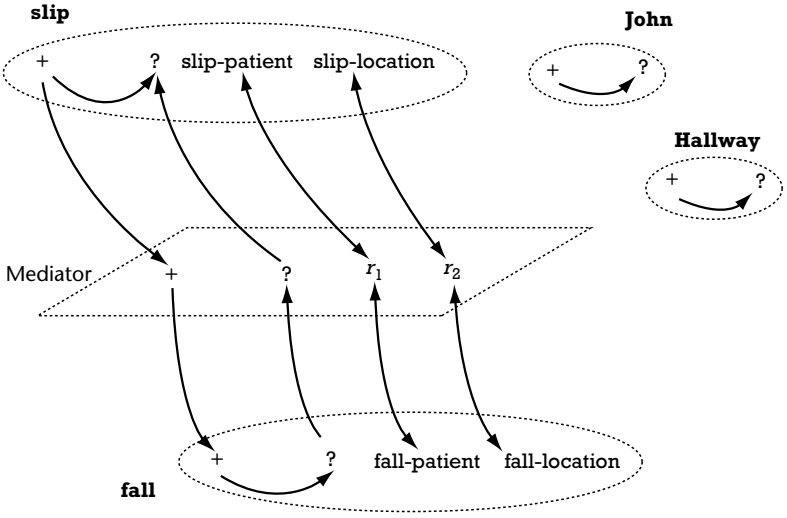


Figure 3. A network fragment illustrating Shruti's encoding of the causal rule-like knowledge 'if one slips, one is likely to fall'. Each ellipse encloses a focal node cluster associated with a relation. Labels within such a cluster denote nodes. The activity of a '+' node indicates an assertion and that of a '?' node indicates a search for an explanation. Other nodes within a focal cluster are role nodes (e.g. slip-patient and slip-location).

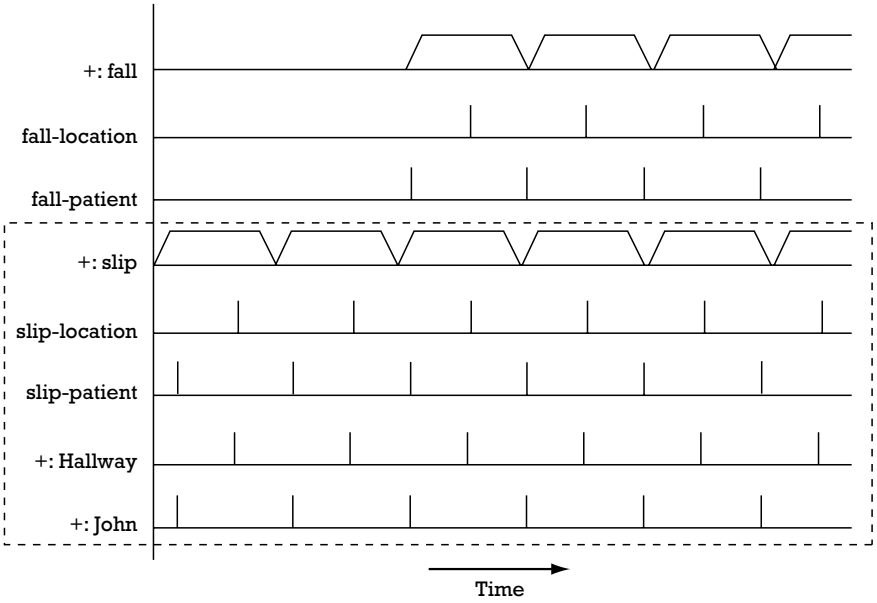


Figure 4. Rhythmic node firings in Shruti. The boxed activities correspond to the encoding of the input sentence 'John slipped in the hallway'. Bindings are encoded by the synchronous firing of role and entity nodes. This activity evolves to include the activation-based encoding of 'John fell in the hallway'.

activated to reflect a given state of affairs, the model spontaneously simulates the behavior of the external world, and in doing so, finds explanations, makes predictions, and draws inferences.

In addition to the representational machinery described above, Shruti can also encode episodic and semantic facts. These facts are 'coincidence' and 'coincidence error' detector circuits situated between the '?' and '+' nodes of focal clusters. A fact becomes active, and enables the flow of activity from the '?' to the '+' node of the associated relational focal cluster, whenever the bindings encoded in the fact match the dynamic bindings expressed in the ongoing flux of activity. Finally, during inference, Shruti can also instantiate new entities, unify multiple entities by merging their phases of firing, and exhibit priming effects by rapid changes in link weights.

With all the above machinery in place, Shruti can rapidly draw a wide range of inferences in parallel, including bridging inferences such as 'Tom had cleaned the hallway', 'John fell because he slipped on the wet hallway floor', and 'John got hurt because of the fall', in response to the input: 'John fell in the hallway. Tom had cleaned it.'

An Enriched Notion of Spreading-activation Networks

In the representational context of events and situations, the operative representational unit in a spreading-activation network is often a circuit of nodes rather than a node. Only some of the nodes in such a circuit correspond to cognitively meaningful entities; other nodes perform a processing function or serve an ancillary representational role. For example, the persistent encoding of the event *E* ('John gave a book to Mary') involves not only nodes corresponding to 'John', 'Mary', 'book', 'giver', 'recipient', and 'given object', but also functional nodes, such as a node for asserting belief in *E*, a node for querying *E*, binder nodes for encoding role–entity bindings in *E*, and nodes for linking *E* to perceptual–motor schemas for the 'give' action.

CURRENT RESEARCH

Spreading-activation networks provide a natural and computationally effective framework for encoding complex evidential interactions. These networks have grown in their sophistication and representational power to encompass not only perceptual and associative phenomena, but also high-level cognitive activities such as reasoning, language understanding, and problem-solving. A promising development has been the blending of spreading-activation models with connectionist models that are explicitly guided by behavioral, biological, and computational constraints. Having resolved some difficult representational problems, the focus of the field is shifting towards the study of structured adaptive networks grounded in perception and action.

Further Reading

Barnden J and Pollack L (1991) (eds) *Advances in Connectionist and Neural Computation Theory*, vol. I. Norwood, NJ: Ablex.

Cottrell GW and Small SL (1983) A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory* **6**: 89–120.

Dell GS (1985) Positive feedback in hierarchical connectionist models: applications to language production. *Cognitive Science* **9**: 3–23.

Engel AK and Singer W (2001) Temporal binding and neural correlates of sensory awareness. *Trends in Cognitive Sciences* **5**: 16–25.

Fahlman SE (1979) *NETL: A System for Representing and Using Real-World Knowledge*. Cambridge, MA: MIT Press.

Feldman JA and Ballard DH (1982) Connectionist models and their properties. *Cognitive Science* **6**: 205–254. [Reprinted in: Anderson JA and Rosenfeld E (eds) (1988) *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.]

McClelland J and Rumelhart D (1986) *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Page M (2001) Connectionist modelling in psychology: a localist manifesto. *Behavioral and Brain Sciences* **23**: 443–467.

Quillian RM (1968) *Semantic memory*. In: Minsky MA (ed.) *Semantic Information Processing*. Cambridge, MA: MIT Press.

Shastri L (1988) *Semantic Networks: An Evidential Formulation and Its Connectionist Realization*. London, UK: Pitman. Los Altos, CA: Morgan Kaufmann.

Shastri L and Ajjanagadde A (1993) From simple associations to systematic reasoning. *Behavioral and Brain Sciences* **16**: 417–494.

Waltz DL and Pollack JB (1985) Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science* **9**: 51–74.

Story Understanding

Intermediate article

Erik T Mueller, IBM Thomas J Watson Research Center, Yorktown Heights, New York, USA

CONTENTS

Story understanding researchers have built computer programs that read and answer questions about simple stories. They have proposed a number of knowledge structures and processing mechanisms useful for this task.

INTRODUCTION

The field of story understanding is concerned with building computer programs that can understand stories. It also investigates how people understand stories. Thus far, the field has dealt mainly with short stories a few paragraphs long, rather than full-length novels, which represent a much harder problem. By 'understanding' it is usually meant that the program should be able to answer questions about a story it reads. It should also be able to generate summaries.

STORY GRAMMARS

An early formalism for stories was the story grammar (Rumelhart, 1975). A story grammar consists of a context-free grammar, with semantic rules associated with each rule of the context-free grammar. An example of a simple story grammar is:

Story → Setting + Episode
Setting → State (Setting ALLOWS Episode)
Episode → Event + Emotion (Event CAUSES Emotion)
Event → Event + Event (Event CAUSES Event)
Event → Episode
Event → Action (1)

Consider the story:

1. Althea was in the playroom.
2. She sat on the balloon.
3. It popped.
4. She was pleased. (2)

The story grammar can be used to parse the story as follows:

Story ⇒
Setting ALLOWS Episode ⇒
State ALLOWS Episode ⇒
1 ALLOWS Episode ⇒
1 ALLOWS (Event CAUSES Emotion) ⇒
1 ALLOWS ((Event CAUSES Event) CAUSES Emotion) ⇒
1 ALLOWS ((Action CAUSES Action) CAUSES Emotion) ⇒
1 ALLOWS ((2 CAUSES 3) CAUSES 4) (3)

The complete parse of the story is:

Althea was in the playroom ALLOWS ((She sat on the balloon CAUSES It popped) CAUSES She was pleased) (4)

This parse can then be used to answer questions:

Why did the balloon pop? Because she sat on it. (5)

Why was she pleased? Because she sat on the balloon and it popped. (6)

The problem with story grammars is that they tie together form and content. According to the above story grammar, the form `Event1 + Event2` corresponds to the content `Event1 CAUSES Event2`. Yet it is quite possible that `Event1 + Event2` instead corresponds to `Event2 CAUSES Event1`, as in the text

The balloon popped. She sat on it. (7)

Although story grammars are useful for capturing the structure of certain story forms such as folk tales, they do not account for the content of a story. They fail to address how an understander is able to make sense of a story despite the variety of

ways of expressing the story. The ingredient missing from story grammars is the understander's knowledge about the way the world works.

KNOWLEDGE-DRIVEN STORY UNDERSTANDING

Stories do not specify everything down to the last detail. Rather, to understand a story one must 'fill in the blanks' and make inferences. Given the text

Althea shook the piggy bank. Three dimes fell out. (8)

a reader easily infers that the dimes fell out because Althea shook the piggy bank, even though this was not explicitly stated. This inference can be made because the reader knows that coins are often stored in piggy banks, that piggy banks have a slot through which coins and other small objects can pass, that shaking helps those objects pass through that slot, and that unsupported objects fall. Such information, known to the reader but not contained in the text, is variously referred to as 'world knowledge', 'general knowledge', 'common-sense knowledge', or simply 'knowledge'.

Much research has focused on identifying the types of knowledge required for story understanding, representing that knowledge within a computer program, and building programs that make use of the knowledge.

Demons

Early work on story understanding (Charniak, 1972) used a single mechanism, called 'demons', for representing and applying knowledge. A demon consists of a test and an action. The test specifies a condition to await. The action specifies an action to perform when the condition becomes true. The following demon allows a program to make the correct inference regarding the piggy bank:

Test : Person *P* shakes piggy bank *B* and money *M* comes out of *B*.
Action : Assert that *M* comes out of *B* because *P* shakes *B*. (9)

That is, this demon generates the inference:

Three dimes fell out BECAUSE Althea shook the piggy bank. (10)

Using a large number of demons, a story understanding program will be able to make a large number of inferences. However, demons can go off on a tangent, generating inferences of doubtful relevance to a story:

The dimes were in the piggy bank BEFORE they fell out. (11)

The dimes were somewhere else BEFORE they were in the piggy bank. (12)

The dimes were minted BEFORE they were somewhere else. (13)

Althea picked up the piggy bank BEFORE she shook the piggy bank. (14)

Althea was somewhere else BEFORE she picked up the piggy bank. (15)

and so on. A type of knowledge structure, called 'scripts', was therefore proposed (Schank and Abelson, 1977) for capturing the relevant inferences in a typical situation.

Scripts, Plans, and Goals

Scripts are bundles of information about situations or activities that are common in a given culture, such as eating at a restaurant, attending a birthday party, or taking the subway. Scripts consist of roles, props, settings, entry conditions, results, and scenes. Here is a short version of the restaurant script:

Roles: customer *C*, waiter *W*
Props: table *T*, menu *M*, food *F*, bill *B*, money *D*
Settings: restaurant *R*
Entry conditions: *C* is hungry, *C* has *D*
Results: *C* is satiated
Scenes:
1. Entering: *C* goes to *R*, *C* sits at *T*
2. Ordering: *C* reads *M*, *C* chooses *F*, *W* comes to *T*, *C* requests *F* from *W*
3. Eating: *W* brings *F* to *C*, *C* eats *F*
4. Exiting: *W* brings *B* to *C*, *C* gives *D* to *W*, *C* leaves *R*

This script may be used to fill in missing information. Told that someone went to a restaurant, ordered lobster, paid the bill, and left, the listener infers (unless told otherwise) that the person ate the lobster.

Of course, the above is not the only possible sequence for eating in a restaurant. Certain types of variations can be accommodated by scripts. A script may have several 'tracks'. The restaurant script has a fast food track, a cafeteria track, and a fancy restaurant track. In the fast food track, the customer pays for the food before eating and may eat the food inside or outside the restaurant. A script may contain alternative paths. If the service is poor, the customer leaves a smaller tip.

However, suppose the story begins:

Suzy was hungry. She went to the Zagat website. (16)

Though the restaurant script mentions hunger, it does not mention visiting a particular website. Knowledge structures for plans and goals were therefore proposed (Schank and Abelson, 1977) to deal with story events that do not follow an existing script. A person has a goal to reduce hunger, and one plan for achieving this goal is toeat at a restaurant. Another plan is to eat at home. A subgoal of eating at a restaurant is to go to the restaurant. A subgoal of going to the restaurant is to know the address of the restaurant. One plan for knowing the address of a restaurant is to read a restaurant guide. Another plan is to ask a friend. One plan for reading a restaurant guide is to read it online. Zagat is a restaurant guide. There are many other plans and subgoals for many other goals. (*See* **Natural Language Processing: Models of Roger Schank and his Students**)

Knowledge of these plans and goals may be used to explain why Suzy went to the Zagat website, namely, in order to read the Zagat guide, so that she could know the address of a restaurant, so that she could go to the restaurant, so that she could eat, so that she could satisfy her hunger.

Themes

Scripts, plans, and goals allow a reader to connect up elements of a story locally. But stories usually have some overall point, moral, or theme. That is, stories are coherent globally. Researchers have proposed a number of related knowledge structures for capturing the point of a story: thematic organization packets, story points, plot units, thematic abstraction units, and planning advice themes.

Consider a story about a man who loses his job, runs out of money, and then happens to save a wealthy person who gives him a large reward. The essence of this story is captured by the story point called 'fortuitous solution' (Wilensky, 1982):

Person *P* is in a negative state.
An incidental event *E* occurs.
E results in a positive state for *P*. (17)

Plot units (Lehnert, 1982) also capture the essence of stories. Plot units are graphs consisting of linked positive events, negative events, and mental states. For example, the 'retaliation' plot unit describes any story in which person *A* causes a negative event for person *B*, which motivates *B* to cause a negative event for *A* (see Figure 1).

Space and Time

Recent approaches to story understanding (Duchan *et al*., 1995) stress the importance of the reader's immersion in the story world. Readers imagine they are inside stories and vicariously experience them. As a story unfolds, the reader keeps track of the shifting 'where', 'when', 'who', and 'what' of the story.

Two-dimensional grids have been proposed (Mueller, 1998) for representing typical locations such as a grocery store, theatre, or hotel room, and for keeping track of the 'where' in a story understanding program. Given the text

Jim was in his hotel room. (18)

an imagined view from above is represented as a square grid in which certain cells are assigned to the various elements of the room: bed, door, minibar, wall, phone, table, and Jim.

Grids allow a story understanding program to make inferences regarding the distance between objects, the relative position of objects (left, right, front, back), whether story characters can see or hear each other, and whether there is a path from one location to another. The program answers questions by consulting the grid:

Was Jim near a bed? Yes. (19)

The 'when' of a story can be represented using absolute timestamps (e.g. '11 a.m. GMT on 12 January 1997') or relations on time intervals (e.g. '*A* happened before *B*', '*A* happened during *B*').

One possible organizational structure for the 'where' and 'when' is the scenario participant map (Dyer, 1983). This is a graph consisting of settings (such as the elevator or hotel room where the story action takes place) connected by transitions (such as walking through the hallway to get to the hotel room).

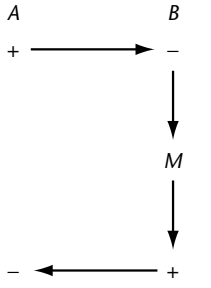


Figure 1. The 'retaliation' plot unit. Person *A* causes a negative event for person *B*, which motivates *B* (via mental state *M*) to cause a negative event for *A*.

USING SCRIPTS, PLANS, GOALS, STORY POINTS, AND PLOT UNITS

A story understanding program builds representations of knowledge structures such as scripts, plans and goals while reading a story. Those representations can then be used for question answering, paraphrasing, and summarization. Table 1 lists some of the story understanding programs that have been built over the years. Some of these programs can be downloaded from the internet (see 'Further Reading').

Script Application

The 'script applier mechanism' or SAM program (Cullingford, 1978) uses scripts to understand stories as follows. Suppose the first sentence of a story is

Fred went to a restaurant. (20)

SAM must activate the restaurant script. Script headers are attached to scripts to assist in script activation. In the above case, a locale header activates the restaurant script because the text mentions that a story character (Fred) went to the setting of the script (restaurant). A precondition header activates a script when the text mentions the main entry condition of a script, as in *Fred was hungry*. A direct header activates a script when the text simply states that the script occurred, as in *Fred ate at a restaurant*. A script is also activated when the text mentions an event of the script.

Alternatively, script activation may be viewed as a text categorization problem and handled using statistical natural language processing techniques. The task is to assign a segment of text to one of many scripts.

When the restaurant script is activated, 'Fred' is assigned to the customer role and 'restaurant' is assigned to the restaurant setting. SAM's representation of the story after reading the first sentence is:

restaurant script, $C =$ 'Fred', $R =$ 'restaurant'
last matched event = 'C goes to R' (21)

As further sentences of the story are read, they are matched to events of the script and additional assignments are made as necessary. For example, given *he ordered lobster*, the representation is updated to:

restaurant script, $C =$ 'Fred', $R =$ 'restaurant', $F =$ 'lobster'
last matched event = 'C requests F from W' (22)

In order to answer a question about the story, the program tries to locate an event of an active script that both matches the question and occurs in the script at or before the last matched event. (Later events have not yet happened in the story.) The script event 'C chooses F' matches the question *What did Fred choose?* and occurs before the last matched event. The variables C and F in the event are replaced by their values, resulting in the answer *Fred chose lobster*.

Several scripts can be active in SAM at a time, and the program can also handle certain deviations from a script, such as being served a burnt hamburger.

Tracking Plans and Goals

Since human behavior is to a large extent goal-directed, tracking the plans and goals of story characters is essential to understanding. If one is told that someone is making marinara sauce, one assumes the person has the goal of eating and is in the middle of preparing a meal.

Given observations of the behavior of story characters, the process of plan recognition (Kautz, 1991) produces explanation graphs such as that shown in Figure 2.

Table 1. Representative story understanding programs

Program	*Year*	*Knowledge structures and mechanisms*
Ms Malaprop	1977	frames
SAM	1978	scripts
PAM	1978	plans, goals
BORIS	1982	scripts, plans, goals, emotions, themes, integration
AQUA	1989	asking and answering questions while reading, explanation
DISCERN	1991	scripts, architecture of subsymbolic networks
TACITUS	1993	axioms, weighted abduction
AbMaL	1994	emotions
SNePS/Cassie	1995	propositional semantic networks, beliefs
ThoughtTreasure	1998	plans, goals, emotions, grids, simulation

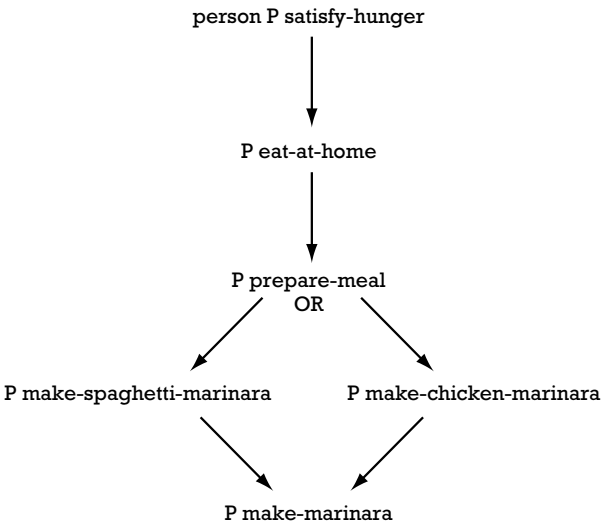


Figure 2. An explanation graph for the information that someone is making marinara sauce.

Since two recipes known to the understander use marinara sauce, the explanation graph contains these alternatives. The top-level goal is to satisfy hunger. In order to satisfy this goal, a subgoal of eating at home is activated. This in turn activates a subgoal to prepare a meal, which activates a subgoal to prepare spaghetti marinara or chicken marinara (the understander does not know which), which activates a subgoal to make marinara sauce. Producing such graphs requires a library of common plans.

Stories often mention states leading to goals. The program thus creates links from states to goals activated by those states:

$$\texttt{P hungry} \rightarrow \texttt{P satisfy-hunger} \quad (23)$$

Emotions are intertwined with goals, and stories often mention the emotional reactions of story characters (Dyer, 1983). Goal successes result in positive emotions, while goal failures result in negative emotions. The program therefore tracks goal outcomes, and creates links from goal outcomes to their resulting emotions:

$$\texttt{P obtain-employment} \rightarrow \texttt{P happy (succeeded)} \quad (24)$$

The program answers questions about a story by consulting the explanation graphs:

Why was Joan making marinara sauce?
She was hungry and wanted to eat. (25)

Why was Jim happy? Because he was hired for a job. (26)

Using Story Points and Plot Units

Story points and plot units are derived from explanation graphs, extended with further goal situations. Goal conflict is the situation in which several goals of a single character interfere with each other. Goal competition is the situation in which the goals of several characters interfere. Goal concord is the situation in which the goals of several characters are compatible. Goal subsumption is the situation in which one goal continually enables satisfaction of another goal.

Plot unit graphs map to explanation graphs as follows. Positive events correspond to goal successes or positive mental states, negative events correspond to goal failures or negative mental states, and mental states correspond to active goals. Named plot units are recognized by building plot unit graphs and matching those graphs to a library of named plot units such as retaliation or fortuitous problem resolution.

Themes such as story points and plot units are useful for predicting what might come next in a story. If a sufficiently large portion of a theme is recognized, the program anticipates the events predicted by the remainder of that theme.

Themes are also useful for producing remindings. If a story is recognized as being an instance of a given theme, the understander may be reminded of another story with that theme.

Finally, themes are useful for summarization. The retaliation plot unit leads to the summarization template:

Because *A* caused a negative event for *B*,
B later caused a negative event for *A*. (27)

For example, a summary that might be produced from the above template is:

Because Harrison turned Debra down for a date, she later turned him down when he changed his mind. (28)

INTEGRATING SENTENCE ANALYSIS WITH STORY UNDERSTANDING

As a story understanding program reads a text, it must incorporate new information into its existing representation of the story. There are two ways this might be performed. In batch interpretation, the program updates its representation after reading each sentence. In incremental interpretation, the program updates its representation after reading each word. The incremental approach is valid from a cognitive standpoint since people appear

to be able to interpret words in real time as they are read (Just and Carpenter, 1980).

There are two ways the story understanding program might be structured: as a series of modules with distinct responsibilities, or as one large process. The trend is towards modular processing, because it is easier to build and understand modular programs, and some self-contained modules for natural language tasks such as part-of-speech tagging now exist.

Modular processing has often been associated with batch interpretation, though this need not be the case. For example, syntactic and semantic parsing modules may cooperate to produce an interpretation incrementally (Mahesh *et al*., 1999).

Let us adopt a modular, batch approach in order to sketch out a complete story understanding program. The first module of a story understanding program is the sentence boundary detector, which segments an input text into sentences. The next module is the entity recognizer. This module segments a sentence into words, phrases, and other entities such as places, numbers, dates, and times. The next module is the tagger, which assigns a part of speech to each entity. The next module is the syntactic parser, which takes a stream of tagged entities and produces syntactic parses such as:

```
[S
  [NP [Name Jim]]
  [VP
    [V set]
    [NP [Det the] [N milk]]
    [PP [Prep on] [NP [Det the]
    [N table]]]]]
```

The next module is the semantic parser, which takes syntactic parse trees and converts them into logical formulae such as:

```
set(Jim,milk,on(table))        (29)
```

Ambiguities are recognized by each module and passed along to the next module. For example, a word may have several possible parts of speech, and a sentence may have several possible syntactic and semantic parses. (*See* **Natural Language Processing**; **Natural Language Processing, Disambiguation in**; **Parsing**; **Parsing: Overview**)

The understanding modules include a script applier, a plan recognizer, and a theme recognizer. Logical formulas from the semantic parser are fed to the understanding modules, which then update their representation or understanding of the story. The understanding modules must agree among themselves how to resolve the ambiguities that were introduced by previous modules, as well as any newly encountered ambiguities.

The question answerer and summarizer take questions from the semantic parser, and examine representations produced by the understanding modules to produce answers and summaries. They use a generator to convert representations into natural language.

The capabilities that have been proposed by various researchers as necessary for story understanding include: to extract themes and morals; to go back and reread; to look for hidden messages; to model naive physics; to model physical objects, devices, and settings; to model the minds of story characters (theory of mind); to pose questions during reading and answer them; to read according to some goal for reading; to read creatively and invent new explanatory frames; to recognize a typical situation (scripts); to reconcile conflicting interpretations; to revise an interpretation; to track emotions of story characters; to track plans and goals of story characters; to track temporal relationships; to track the shifting 'where', 'when', 'who', and 'what'; to use discourse markers; to use imagery or visual representations; to use past experiences to guide understanding; to zoom in on detail; and to zoom out from detail. This list does not include capabilities normally assumed in natural language processing, such as syntactic parsing and anaphora resolution.

The processing mechanisms used in various story understanding programs include: abduction (inference to the best explanation); backward chaining; constraint satisfaction; demons; discrimination nets; finite-state automata; forward chaining; indexing; logic; neural networks; pattern matching; plan recognition; production systems; simulation; society of agents; spreading activation or marker passing; and working memory. These vary in scope. For example, it has been claimed that all levels of natural language processing, including syntax, semantics, and pragmatics, can be handled using abduction.

SPREADING ACTIVATION AND CONNECTIONIST APPROACHES

Dissatisfaction with strictly symbolic mechanisms has led some researchers to experiment with mechanisms inspired by the physiology of the brain and experimental results in psychology such as those from priming experiments. (*See* **Syntax and Semantics, Neural Basis of**; **Connectionism**; **Priming**)

Spreading Activation

Consider the text

> Suzy wanted to eat dinner. She got in her car. (30)

One way to relate these two sentences is to conduct two simultaneous searches through the plan library. One search starts from `eat-dinner` while the other starts from `enter-car`. When one search meets the other, a path has been found from `eat-dinner` to `enter-car`, as shown in Figure 3.

This method is known as 'spreading activation' or 'marker passing' (Charniak, 1986). Once a path is found, it must be verified for correctness. It must be checked, for instance, that the person who wants to eat dinner is the same person who got in the car. Marker passing with verification is one method for doing plan recognition to produce an explanation graph.

Connectionist Approaches

So far we have discussed symbolic representations of knowledge structures such as scripts, plans, and goals. An alternative kind of representation is a subsymbolic, or connectionist, one, in which concepts are represented not as discrete entities but by the pattern of activity in a neural network.

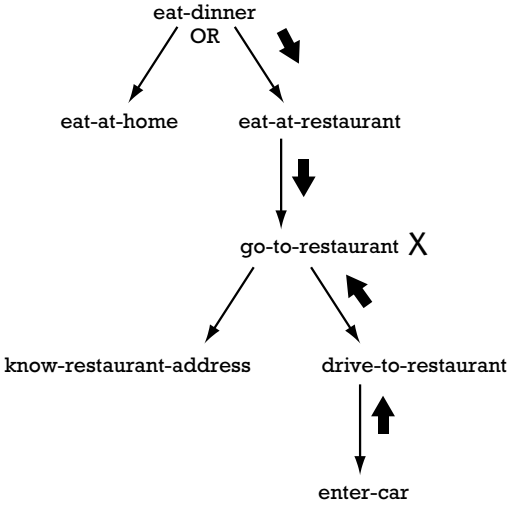


Figure 3. The statements 'Suzy wanted to eat dinner' and 'she got in her car' can be related by conducting two simultaneous searches through the plan library, one from `eat-dinner` and the other from `enter-car`. These are shown by the downward- and upward-facing stubby arrows, respectively. The point of intersection is marked by the 'X'.

The connectionist approach has several advantages. Firstly, neural networks are neurally inspired, so a story understanding program implemented in neural networks is more likely to model how understanding is actually implemented in the brain. Secondly, neural networks can be trained on a set of examples in order to learn representations and make generalizations automatically, reducing the need for a programmer to specify knowledge structures. Thirdly, the performance of neural networks generally degrades in a steady fashion.

On the other hand, the connectionist approach has several disadvantages. Firstly, neural networks have difficulty with novel inputs, since they need to be trained on a large number of examples of their inputs. Secondly, it is difficult to implement in neural networks certain operations such as role assignment and composition that are easy to implement in symbolic programs.

Miikkulainen (1993) used the connectionist approach to build a complete program called DISCERN that reads and answers questions about script-based stories. The program is built from independent connectionist modules that communicate using distributed representations. The modules are: lexicon, episodic memory, sentence parser, sentence generator, story parser, story generator, cue former, and answer producer.

The episodic memory stores and generalizes script-based stories. It is organized by the programmer into a fixed three-level architecture. The top level represents the script class, the middle level represents the script track, and the bottom level represents the script roles. The neural networks at each level are self-organized by training on an artificially generated set of stories involving the restaurant, shopping, and travel scripts. In testing, the distributed representation of an input story, which may contain missing role bindings, is fed to the episodic memory. The episodic memory is able to fill in any missing role bindings using the generalizations it made during training. The cue former retrieves answers to questions about input stories from the episodic memory.

DISCERN can only handle one script per story and is unable to handle deviations from a script. Thus it is not as sophisticated as SAM, one of the early symbolic story understanding programs.

ASSESSMENT

Since the 1970s, we have learned much about story understanding. Yet it is still not known how to scale up a story understanding program so that it can understand more than 'toy' stories. The problem

does not appear to be what symbolic or subsymbolic mechanism is used for processing – probably a number of mechanisms will do the job – but how to get a story understanding program to work at the human level at all.

It is hardly surprising that story understanding is a difficult problem. It is a task that calls the entire mind into play. All of the explanatory frames, skills, and mechanisms used to deal with everyday life can be invoked in story understanding. A story can be about almost anything, from picking up and holding an apple (motor skills) to observing the subjective redness of the apple (consciousness). (*See* **Language Comprehension, Methodologies for Studying**; **Discourse Processing**)

Programmers have difficulty managing the complexity of building and debugging story understanding programs. Even if a library of common-sense knowledge is available (Cyc and ThoughtTreasure are attempts at building such libraries), it is still difficult to build processing mechanisms that apply the library in understanding. It appears that the knowledge library needs to be developed with the understanding program in mind, yet it is not clear how this can be done. Story understanding might take a cue from statistical language processing and information extraction. Success has been achieved in these fields by building modules that address well-defined subproblems, such as part-of-speech tagging or filling templates about terrorism news stories. By putting together many modules that address parts of the story understanding problem, it may be possible to reach a complete solution. (*See* **Natural Language Processing, Statistical Approaches to**)

It is easy to forget how ambiguous natural language is. A sentence has many possible interpretations. Many of those possibilities are implausible, but the program does not always know that. Furthermore, the possible interpretations of each sentence must be considered in light of the possible interpretations of previous sentences. So if there are two interpretations of the first sentence, there might be four after reading the second, eight after reading the third, and so on. Knowledge structures such as scripts were designed to prevent such problems, but in practice they do not always work.

It is a mystery how people are able to avoid this proliferation of possible interpretations and understand stories. Cognitive psychologists have conducted experiments investigating how inferences are made during narrative comprehension (Graesser *et al*., 1994). Cognitive neuroscientists are beginning to address discourse comprehension (Beeman and Chiarello, 1998). Someday it may be possible to use a brain scan with high spatial and temporal resolution to trace a behavior such as a verbal response to a question back to its causes during reading.

References

Beeman M and Chiarello C (eds) (1998) *Right Hemisphere Language Comprehension*. Mahwah, NJ: Erlbaum.

Charniak E (1972) *Toward a Model of Children's Story Comprehension*. AI Laboratory Technical Report 266, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Charniak E (1986) A neat theory of marker passing. In: *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 584–588. Menlo Park, CA: AAAI Press.

Cullingford RE (1978) *Script Application: Computer Understanding of Newspaper Stories*. Research Report 116, Computer Science Department, Yale University.

Duchan JF, Bruder GA and Hewitt LE (eds) (1995) *Deixis in Narrative*. Hillsdale, NJ: Erlbaum.

Dyer MG (1983) *In-Depth Understanding*. Cambridge, MA: MIT Press.

Graesser AC, Singer M and Trabasso T (1994) Constructing inferences during narrative text comprehension. *Psychological Review* **101**(3): 371–395.

Just MA and Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. *Psychological Review* **87**(4): 329–354.

Kautz HA (1991) A formal theory of plan recognition and its implementation. In: Allen JF, Kautz HA, Pelavin RN and Tenenberg JD (eds) *Reasoning About Plans*, pp. 69–125. San Mateo, CA: Morgan Kaufmann.

Lehnert WG (1982) Plot units: a narrative summarization strategy. In: Lehnert WG and Ringle MH (eds) *Strategies for Natural Language Processing*, pp. 375–412. Hillsdale, NJ: Erlbaum.

Mahesh K, Eiselt KP and Holbrook JK (1999) Sentence processing in understanding: interaction and integration of knowledge sources. In: Ram A and Moorman K (eds) *Understanding Language Understanding*, pp. 27–72. Cambridge, MA: MIT Press.

Miikkulainen R (1993) *Subsymbolic Natural Language Processing*. Cambridge, MA: MIT Press.

Mueller ET (1998) *Natural Language Processing With ThoughtTreasure*. New York, NY: Signiform. [Available from: http://www.signiform.com/tt/book/]

Rumelhart DE (1975) Notes on a schema for stories. In: Bobrow DG and Collins AM (eds) *Representation and Understanding: Studies in Cognitive Science*, pp. 211–236. New York, NY: Academic Press.

Schank RC and Abelson RP (1977) *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum.

Wilensky R (1982) Points: a theory of the structure of stories in memory. In: Lehnert WG and Ringle MH (eds) *Strategies for Natural Language Processing*, pp. 345–374. Hillsdale, NJ: Erlbaum.

Further Reading

Bartlett FC (1932) *Remembering*. Cambridge, UK: Cambridge University Press.

DISCERN. [ftp://ftp.cs.utexas.edu/pub/neural-nets/software/discern.2.1.1.tar.Z.] [The *DISCERN* program developed by Miikkulainen].

Hobbs JR, Stickel ME, Appelt DE and Martin P (1993) Interpretation as abduction. In: Pereira FCN and Grosz BJ (eds) *Natural Language Processing*, pp. 69–142. Cambridge, MA: MIT Press.

ICU. [http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/nlu/icu/0.html.] [Miniature version of the SAM model developed by Cullingford.]

Lenat DB (1995) Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM* **38**(11): 33–48.

McDypar. [http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/nlu/mcdypar/0.html.] [Miniature version of the parser used in the BORIS program developed by Dyer.]

McKoon G and Ratcliff R (1992) Inference during reading. *Psychological Review* **99**(3): 440–466.

ter Meulen AGB (1995) *Representing Time in Natural Language*. Cambridge, MA: MIT Press.

O'Rorke P and Ortony A (1994) Explaining emotions. *Cognitive Science* **18**: 283–323.

Ram A and Moorman K (eds) (1999) *Understanding Language Understanding*. Cambridge, MA: MIT Press.

Schank RC and Riesbeck CK (1981) *Inside Computer Understanding*. Hillsdale, NJ: Erlbaum.

SNePS. [ftp://ftp.cse.buffalo.edu/pub/sneps.] [The SNePS program.]

Thought Treasure. [http://www.signiform.com/tt/htm/tt.htm.] [The Thought Treasure program and database developed by Mueller.]

STRIPS

Introductory article

David Furcy, Georgia Institute of Technology, Atlanta, Georgia, USA
Sven Koenig, Georgia Institute of Technology, Atlanta, Georgia, USA

CONTENTS

STRIPS was the planning component of the agent architecture for the robot Shakey. Its representation language is still widely used for planning.

THE PROBLEM OF GOAL-DIRECTED ACTION

Intelligent agents, including humans, need to determine how to achieve their goals. Reasoning about how the execution of actions can change the current state of the world to a desired state is called 'planning'. How to automate this task has been a major focus of research in artificial intelligence (AI).

This article describes seminal work on planning in AI, and the influence of that work on current research. The 'Stanford Research Institute problem solver' (STRIPS) was developed by Fikes, Nilsson, Hart, Sacerdoti and their colleagues as the planning component of the agent architecture for the robot Shakey. Begun in the late 1960s, the Shakey project studied how to control a robot to push boxes from room to room. The project had a major influence on subsequent research on planning and plan execution in AI. For example, although current planners are able to solve planning problems that are much larger than those solved by Shakey, many of them use variants of the knowledge representation language originally used by STRIPS. (*See* **Robotics**)

We will first describe the representation language of STRIPS, and then its search strategy, called means–ends analysis, which had originally been developed by Newell, Simon and their colleagues in the context of the 'general problem solver'. Next, we will discuss other important components of the agent architecture of Shakey, including macro-operator learning, plan monitoring, and hierarchical planning. Finally we will briefly discuss how current planners use and extend the representation language of STRIPS. (*See* **Newell, Allen**; **Simon, Herbert A.**)

PROPOSITIONAL WORLD MODELS AND THE ROBOT NAVIGATION TASK

A planner needs to represent world states in order to reason about them. STRIPS represents world states as sets of propositions, called world models. Each proposition in a world model represents a fact that is true in the corresponding world state. STRIPS assumes that propositions that are not part of a world model correspond to facts that are false in the corresponding world state (the 'closed-world assumption'). STRIPS also represents the goal as a set of propositions. The objective of planning, then, is to change the current world model to one that includes all propositions of the goal (but may contain additional propositions). The goal thus represents a set of world models, one of which needs to be reached. (*See* **Knowledge Representation**)

We will use Shakey's environment to illustrate the representation language of STRIPS. Symbols, such as `Box1` or `Door1`, represent objects. Propositions represent either properties of objects, such as `Pushable(Box1)` and `Open(Door1)`, or relationships among two or more objects, such as `Connects(Door1,Room1,Room2)` or `In(Shakey,Room1)`. (We will refer to `Pushable`, `Connects`, and so on as the 'proposition names'.) Some propositions are part of all world models, for example because they encode background knowledge: examples are `Pushable(Box1)` and `Connects(Door1,Room1,`

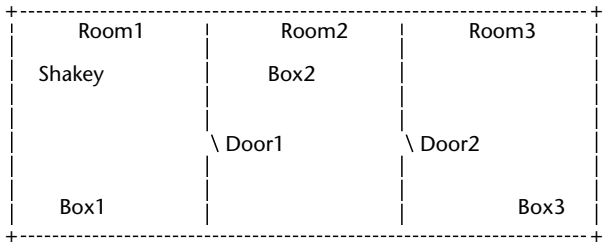


Figure 1. Diagram of a world state consisting of 3 rooms with 2 connecting open doors.

Room2). Other propositions encode facts that can change and are thus not part of all world models: examples are Open(Door1) and In (Shakey,Room1).

Imagine a world state with three rooms in a row, connected with two open doors (Figure 1). Each room contains a pushable box, and Shakey is in the first room. The corresponding world model is:

```
Open(Door1) and Open(Door2) and
Connects(Door1,Room1,Room2) and
  Connects(Door2,Room2,Room3) and
Connects(Door1,Room2,Room1) and
  Connects(Door2,Room3,Room2) and
Pushable(Box1) and Pushable(Box2)
  and Pushable(Box3) and
In(Shakey,Room1) and In(Box1,
  Room1) and In(Box2,Room2) and
  In(Box3,Room3)
```

A possible goal would be:

```
In(Box2,Room1)
```

This goal is satisfied by any world state where Box2 is in Room1.

The domain designer chooses which objects and propositions are used to represent world states. For example, we have not included the color of the boxes in the example, since this feature is irrelevant for pushing them. The representation language, together with the chosen objects and propositions, determines which planning problems a planner can solve and how quickly it can solve them. Many current planners use the propositional representation language presented here. STRIPS actually uses a slightly more expressive representation language. For example, it can represent in one short statement the fact that all boxes are pushable. However, its more expressive representation language requires it to use a theorem prover during planning, which increases the planning time.

PLANNING OPERATORS

A planner also needs to represent actions in order to reason about them. It needs to know, for example, in which world states they apply and which world states result from their execution. STRIPS assumes that each action execution instantaneously and deterministically transforms the current world state into another world state. STRIPS represents actions as operators. Each operator has a 'precondition list', an 'add list', and a 'delete list'. An operator can be applied in a world model if the world model contains all propositions of its precondition list. The world model that results from its application is obtained by deleting all propositions of its delete list from the current world model and then adding all propositions of its add list. Thus, the precondition list specifies which facts must be true in the current world state in order for an action to be executable. The add list specifies which facts its execution makes true; and the delete list specifies which facts its execution makes false. STRIPS assumes that the truth values of all other facts remain unchanged (the 'STRIPS assumption').

STRIPS actually uses operator schemata ('STRIPS rules') in order to represent actions concisely and reason about them in a general way. Operator schemata generalize operators, since they can use variables where operators use object symbols. The operators are then obtained by instantiating the operator schemata, that is, by replacing each variable within an operator schema with the same object symbol. Thus, an operator schema corresponds to several operators.

For example, the following operator pushes Box1 from Room1 through Door1 to Room2:

- Name: PUSHTHRU-Box1-Door1-Room1-Room2
- Precondition list: In(Shakey,Room1) and In(Box1,Room1) and Pushable(Box1) and Connects(Door1,Room1,Room2) and Open(Door1)
- Add list: In(Shakey,Room2) and In(Box1, Room2)
- Delete list: In(Shakey,Room1) and In(Box1,Room1)

The corresponding operator schema, which pushes any object from one room through some door to another room, is:

- Name: PUSHTHRU
- Parameter list: box, door, room1, room2
- Precondition list: In(Shakey,room1) and In(box,room1) and Pushable(box) and

Connects(door,room1,room2) and Open(door)
- Add list: In(Shakey,room2) and In(box, room2)
- Delete list: In(Shakey,room1) and In(box, room1)

(Note that object names start with a capital letter while variable names do not.) The following operator schema moves Shakey from one room through a door to another room:

- Name: GOTHRU
- Parameter list: door, room1, room2
- Precondition list: In(Shakey,room1) and Connects(door,room1,room2) and Open(door)
- Add list: In(Shakey,room2)
- Delete list: In(Shakey,room1)

PLANNING AS SEARCH

STRIPS takes as input the operator schemata, the initial world model, and the goal. Its objective is to output an operator sequence that transforms the initial world model into one that satisfies the goal. This is a difficult problem: in general, it is P-space complete to determine whether such an operator sequence exists. In principle, planning can be formulated as searching the state-space graph, whose vertices correspond to world models (states). An edge corresponds to an operator, and connects a world model in which the operator can be applied to the world model that results from its application. The objective of the search is to find a path from the vertex that corresponds to the initial world model to any vertex that corresponds to a world model that satisfies the goal. The high complexity of planning can be explained by the size of the state-space graph: the number of vertices is exponential in the number of propositions used to describe the planning problem. (*See* **Computability and Computational Complexity**)

Planners that search the state-space graph are called 'state-space' or 'situation-space' planners. The size of the graph makes uninformed searches of the graph intractable even for relatively small planning problems. STRIPS therefore searches the graph without representing all of it explicitly, using the structure of the world models and some heuristics to guide its search. (*See* **Search**)

PLAN GENERATION THROUGH MEANS–ENDS ANALYSIS

The following piece of pseudo-code shows the principle of means–ends analysis:

```
achieve-propositions(world-model
W1, propositions S)
{
  if S is a subset of W1, return the
    empty operator sequence;
  choose a proposition P such that P
    is part of S but not W1 (choice
    point);
  operator-sequence O1 := achieve-
    proposition(W1, P);
  world-model W2 := the world model
    that results from applying O1 in
    W1;
  operator-sequence O2 := achieve-
    propositions(W2, S);
  return the concatenation of O1 and
    O2
}
achieve-proposition(world-model
  W, proposition P)
{
  choose an operator O1 whose add
    list contains P (choice point);
  if no such operator exists,
    backtrack to the previous choice
    point;
  operator-sequence O2 := achieve-
    propositions(W, precondition
    list of O1);
  return O2 with O1 appended
}
```

STRIPS uses means–ends analysis as its search strategy. Means–ends analysis focuses on the propositions that need to be achieved (the ends) and the operators that can achieve them (the means). (*See* **Means–Ends Analysis**)

Initially, STRIPS calls achieve-propositions with the initial world model and all propositions that are part of the goal. It tries to achieve the desired propositions one after the other, combining forward and backward searches. If a desired proposition has not already been achieved, it chooses an operator that adds it and then tries to achieve the preconditions of the operator.

Later versions of means–ends analysis only choose operators whose main (primary) effects are to add the proposition. For example, moving the box is the main effect of the PUSHTHRU-Box1-Door1-Room1-Room2 operator. Moving Shakey is only a side effect.

Thus, means–ends analysis constructs plans by concatenating sub-plans, each of which achieves a desired proposition. However, a sub-plan can delete desired propositions that were achieved by earlier sub-plans in the sequence. Since there are planning problems where this happens no matter

the order in which means–ends analysis tries to achieve the propositions (Sussman anomaly), it might have to reachieve them. How to do this is a difficult problem, and means–ends analysis is not guaranteed to find an operator sequence that achieves the goal from the initial world model even if one exists.

PLAN GENERALIZATION

Since planning is time-consuming, STRIPS caches plans for later reuse. However, each cached plan can only be reused for one specific planning problem. Thus, a large number of plans need to be cached, which can make plan retrieval as time-consuming as planning itself (the 'utility problem'). To address this problem, STRIPS generalizes plans before caching them.

For example, the problem of the boxes in the three rooms, discussed above, is solved by the operator sequence `[GOTHRU(Door1,Room1, Room2), PUSHTHRU(Box2,Door1, Room2,Room1)]`. STRIPS first stores it in a triangle table, which is a compact way of recording the dependencies of the preconditions of each operator on the effects of the operators that precede it in the operator sequence. STRIPS then uses a theorem prover to generalize the operator sequence by replacing as many object symbols as possible with variables while maintaining the dependencies recorded in the triangle table. Maintaining the recorded dependencies guarantees that the preconditions of all operators are satisfied during the application of the operator sequence, no matter which object symbols the variables are replaced with.

This results in the generalized operator sequence `[GOTHRU(door1,room1,room2), PUSHTHRU(box,door2,room2,room3)]`. Finally, STRIPS could create a new operator (a 'macro-operator') (with a computer-generated name) that has the same effect as the operator sequence:

- Name: `FOO1653`
- Parameter list: `door1, door2, room1, room2, room3, box`
- Precondition list: `In(Shakey,room1)` and `In(box,room2)` and `Pushable(box)` and `Connects(door1,room1,room2)` and `Connects(door2,room2,room3)` and `Open(door1)` and `Open(door2)`
- Add list: `In(Shakey,room3)` and `In(box, room3)`
- Delete list: `In(Shakey,room1)` and `In(box, room2)`

This macro-operator could then be used like any other operator, allowing the program to achieve any proposition that is part of the add list of the macro-operator with only one operator application (instead of two, in our example). This reduces the number of recursive calls to `achieve-propositions`, and thus the number of choice points for backtracking.

In fact, STRIPS stores the triangle table instead of the macro-operator, so that it can reuse not only the generalized operator sequence, but also parts of it.

Unfortunately, the number of sub-sequences grows exponentially with the length of the operator sequence in a triangle table. Therefore, there is a trade-off between the number and size of stored triangle tables.

DYNAMIC MONITORING OF ACTIONS

The execution of plans can fail if STRIPS makes wrong assumptions. For example, if STRIPS were erroneously told that a door was open, or if the door were closed during plan execution, then the resulting plan might direct Shakey to pass through the door without opening it first. In this case, its execution would fail. Therefore, it is important that plan execution be monitored. Shakey's plan monitor, PLANEX, can change the operator sequence that Shakey executes. If this does not resolve the problem, PLANEX invokes STRIPS to find a different plan that achieves the goal from the current world state.

For example, if the preconditions of an operator are not satisfied, PLANEX can insert operators before it to ensure that the operator sequence can be applied and that it continues to achieve the goal. In our example, plan execution fails when Shakey tries to pass through a closed door. Therefore, PLANEX inserts an operator that opens the door and then proceeds to execute the operator sequence.

PLANEX also optimizes an operator sequence during execution. It uses a triangle table to find the shortest suffix of the operator sequence that achieves the goal, and executes it instead of the whole operator sequence. This allows PLANEX to skip operators if some of the operators returned by means–ends analysis are redundant or if their execution is unnecessary for some other reason. For example, Shakey might try to open a door that is already open because STRIPS was erroneously told that the door was closed. In this case, PLANEX removes the operator that opens the door from the operator sequence, and then proceeds to execute the sequence.

HIERARCHICAL PLANNING IN ABSTRIPS

Hierarchical planning can reduce the planning time by first finding a plan for a simplified version of the planning problem and then using it to speed up finding a plan for the more complex problem. ABSTRIPS is a hierarchical planner based on STRIPS. It was the first planner to demonstrate the computational advantages of hierarchical planning. It simplifies the planning problem by deleting details, then finds a plan for the more abstract planning problem, and finally refines it by filling in operators that achieve the deleted details.

For each proposition, ABSTRIPS stores a criticality value, which measures how easy it is to achieve the proposition. These values are determined by a combination of human input and automatic reasoning. For example, `Open(Door1)` has a low criticality value since closed doors can easily be opened once Shakey is next to them. `In(Room1)` has a higher criticality value since Shakey might have to traverse several rooms and open several doors in order to reach the room. ABSTRIPS first deletes all propositions with criticality values below a given threshold from all world models and operators. This makes the state-space graph significantly smaller, and therefore the planning problem easier to solve. ABSTRIPS then uses STRIPS to find a plan that solves this simplified planning problem. (In our example, this plan is a sequence of `GOTHRU` and `PUSHTHRU` operators.) It does not solve the original planning problem if the preconditions of some operators or some goal propositions are not satisfied, for example if it requires Shakey to go through closed doors. In such cases, ABSTRIPS might have to refine the plan. It uses STRIPS to find plans that achieve the unsatisfied propositions, and inserts each plan into the operator sequence where it is needed. (In our example, this adds the necessary `OPEN` operators to the operator sequence.) If the plan cannot be refined by inserting additional operators, ABSTRIPS attempts to find a different plan for the simplified planning problem (via backtracking) and refine it.

This planning scheme with two levels of abstraction can easily be generalized to multiple levels of abstraction with decreasing thresholds of criticality values. The resulting search strategy is called 'length-first' search, because it constructs a complete plan at each level of abstraction before it refines the plan at the next lower level of abstraction.

THE INFLUENCE OF STRIPS ON COMPUTER MODELS OF GOAL-DIRECTED ACTION

STRIPS had a major influence on subsequent research in AI planning. Research into topics such as plan representation, plan generalization and learning macro-operators, execution monitoring and replanning, and hierarchical planning, is continuing. For instance, the plan generalization method of STRIPS is an example of a class of machine-learning methods later called explanation-based learning. (*See* **Machine Learning**)

STRIPS trades off the expressive power of the representation language and the efficiency of reasoning. In principle, planning problems can be expressed in first-order logic (situation calculus) and solved with theorem provers. The designers of STRIPS decided to sacrifice some expressive power of the representation language to be able to plan more efficiently with specialized reasoning methods, such as means–ends analysis. STRIPS combined means–ends analysis with theorem proving; whereas many current planners use planning methods (without theorem proving) that work on the simplified propositional representation language presented here. This version of the STRIPS representation language allows one to represent simplified versions of many interesting planning problems, and retains the important property that the domain designers do not need to represent those facts that are not affected by action executions. (*See* **Frame Problem, The**)

However, few current planners use means–ends analysis as a search strategy. The success of some of them even casts doubt on a basic assumption behind means–ends analysis, namely, that planning has to use the structure of world models to be efficient. Some recent planners, for example, are based on techniques from heuristic search, constraint satisfaction, or model checking. Others do not even search the situation space. Planners that use hill-climbing methods from boolean satisfiability, for example, search the space of partial assignments of truth values to propositions; and partial-order planners search the space of incomplete plans. (*See* **Hill-climbing Search**)

However, most planners solve the same kind of problems as STRIPS, and make the same assumptions. They assume, for example, that planning problems have discrete world states and actions, that the goal can be expressed as a set of world states, that the goal is known in advance and does not change over time, that actions have known

effects and change the world state deterministically and instantaneously, that the agent always observes the current world state completely and executes actions sequentially, and that the states change only due to actions executed by the agent.

The representation language of STRIPS has been extended in various ways to relax some of these assumptions. For example, the 'planning domain definition language' is more general since it can model continuous resources, such as time or energy, and durative actions with continuous effects. There are also variants of STRIPS that can express planning problems where actions have nondeterministic effects or the current world state is only partially observable, in which case planners must decide when and what to sense and which actions to execute depending on the sensed information ('conditional planning').

SUMMARY

STRIPS was an early planner, consisting of a language that represents world states and actions, and a planning method that uses means–ends analysis to search the space of world models. It had a major influence on subsequent research in AI planning. Although current planners are more efficient and solve more realistic planning problems, the representation language of STRIPS is still widely used.

Further Reading

Cohen PR (1982) Planning and problem solving: STRIPS and ABSTRIPS. In: Cohen PR and Feigenbaum EA (eds) *Handbook of Artificial Intelligence*, vol. III, chap. XV, section B, pp. 523–530. Reading, MA: Addison-Wesley.

Gardner A (1981) Search: STRIPS and ABSTRIPS. In: Barr A and Feigenbaum EA (eds) *Handbook of Artificial Intelligence*, vol. I, chap. II, sections D5–D6, pp. 128–139. Reading, MA: Addison-Wesley.

Knoblock CA (1993) *Generating Abstraction Hierarchies: An Automated Approach to Reducing Search in Planning*. Norwell, MA: Kluwer.

Nilsson NJ (1980) *Principles of Artificial Intelligence*, chap. VII, pp. 275–319. Palo Alto, CA: Tioga.

Nilsson NJ (1998) *Artificial Intelligence: A New Synthesis*, chap. XXII, pp. 373–399. San Francisco, CA: Morgan Kaufmann.

Russell SJ and Norvig P (1995) *Artificial Intelligence: A Modern Approach*, chaps. XI–XIII, pp. 335–412. Upper Saddle River, NJ: Prentice Hall.

Yang Q (1997) *Intelligent Planning: A Decomposition and Abstraction Based Approach*. New York, NY: Springer.

Supervenience

Intermediate article

Brian P McLaughlin, Rutgers University, New Brunswick, New Jersey, USA

CONTENTS

There is supervenience when a difference in one respect requires a difference in another respect. For example, the mental characteristics of individuals supervene on their physical characteristics if and only if two individuals cannot differ in respect of their mental characteristics without differing in respect of their physical characteristics.

INTRODUCTION

The notion of supervenience has been invoked in virtually every area of philosophy. It has been invoked to try to characterize the relationship between mental and physical properties, the relationship between moral and natural properties, and the relationship between aesthetic and natural properties, to name just a few. There have been attempts to formulate the doctrine of physicalism as a psychophysical supervenience thesis. And the positions of internalism and externalism in psychosemantics – the theory of meaning for mental representations – are usually characterized in terms of supervenience theses.

WHAT IS SUPERVENIENCE?

Sometimes one difference requires another. Two sets cannot differ in respect of their cardinalities without differing in respect of what members they contain; two elements cannot differ in respect of their chemical properties without differing in respect of their electronic shells; and two machines cannot differ in respect of what algorithms they can execute without differing in respect of what programs they can run. If one sort of difference, difference in *A*-respects, requires another sort of difference, difference in *B*-respects, then, and only then, *A*-respects supervene on *B*-respects (McLaughlin, 1995). For example, mental respects supervene on physical respects if and only if there cannot be a difference in mental respects without a difference in physical respects; aesthetic respects supervene on natural respects if and only if there cannot be a difference in aesthetic respects without a difference in natural respects.

Supervenience is a relation. As such it is reflexive and transitive, but it is neither symmetrical nor asymmetrical. Specifically, supervenience is the relation of dependent variation (McLaughlin, 1995). For *A*-respects supervene on *B*-respects just in case variation in *A*-respects *depends* on variation in *B*-respects. The dependency is purely modal: from the fact that *A*-respects supervene on *B*-respects it does not follow that something is the way it is in *A*-respects by virtue of its being the way it is in *B*-respects. Equivalently, in the relevant sense of independence, *A*-respects can vary independently of variation in *B*-respects if and only if it is possible to have variation in *A*-respects without variation in *B*-respects. So, for example, mental respects fail to supervene on physical respects if and only if mental respects can vary independently of variation in physical respects (Jackson, 1993). That is just to say that mental respects fail to supervene on physical respects when and only when it is possible to have variation in mental respects without variation in physical respects.

When A-respects supervene on B-respects, *A*-respects are *supervenient* relative to *B*-respects and *B*-respects are *subvenient* relative to *A*-respects (Kim, 1984). Given that it is impossible to have variation in supervenient *A*-respects without variation in subvenient *B*-respects, it follows that exact similarity in *B*-respects suffices for exact similarity in *A*-respects. For example, two sets exactly alike in respect of what members they contain must be exactly alike in respect of what cardinalities they have; two machines exactly alike in respect of what programs they can run will be exactly alike in respect of what algorithms they can execute; and two elements exactly alike in their electronic shells must be exactly alike in their chemical properties. In the first two cases, the 'must' is that of logical necessity; in the last, it is that of nomological necessity.

Call two things *A-twins* if and only if they are exactly alike in every *A*-respect. (Note that everything is trivially an *A*-twin of itself.) Then, *A*-respects supervene on *B*-respects if and only if *B*-twins cannot fail to be *A*-twins (McLaughlin, 1995). So, for example, mental respects supervene on physical respects just in case physical twins cannot fail to be mental twins. If physical twins can fail to be mental twins, then, and only then, mental respects fail to supervene on physical respects; for, then and only then, variation in mental respects is possible without variation in physical respects.

VARIETIES OF SUPERVENIENCE

Various definitions have been proposed that define determinates of the determinable relation of dependent variation, and so define various subcategories of supervenience.

Strong and Weak Supervenience

Quantifying over possible worlds, two much-discussed categories of supervenience can be defined as follows (Kim, 1987; McLaughlin, 1995):

Weak supervenience: *A*-respects weakly supervene on *B*-respects if and only if for any possible world w, *B*-twins in w are *A*-twins in w. (1)

Strong supervenience: *A*-respects strongly supervene on *B*-respects if and only if for any possible worlds w_1 and w_2 and any individuals x_1 and x_2, if x_1 in w_1 is a *B*-twin of x_2 in w_2 then x_1 in w_1 is an *A*-twin of x_2 in w_2. (2)

Strong implies weak, but not conversely. Weak supervenience is concerned with intraworld relations, with whether it is the case that in every world, *B*-twins are *A*-twins. Strong supervenience is concerned with interworld as well as intraworld relations, with whether any possible *B*-twins, from the same or different worlds, are *A*-twins within their respective worlds.

The range of possible worlds quantified over can be restricted; for example, it can be restricted to the causally possible worlds, or the nomologically possible worlds. These restrictions yield subcategories of weak and strong supervenience.

Modal Operator Supervenience

While the focus of this article is supervenience understood as the relation of dependent variation, other relations are sometimes called supervenience relations. Jaegwon Kim (1984, 1987) has attempted to define notions of supervenience using modal operators, rather than quantification over possible worlds. He offers the following definitions, where *A* and *B* are non-empty sets of properties:

Modal operator weak supervenience: *A* weakly supervenes on *B* if and only if necessarily, if anything has some property *F* in *A*, then there is at least one property *G* in *B* such that that thing has *G*, and everything that has *G* has *F*. (3)

Modal operator strong supervenience: *A* strongly supervenes on *B* if and only if necessarily, if anything has some property *F* in *A*, then there is at least one property *G* in *B* such that that thing has *G*, and necessarily everything that has *G* has *F*. (4)

Since definitions 3 and 4 differ only in that the latter contains one more occurrence of 'necessarily' than does the former, operator strong implies operator weak, but not conversely. Kim (1993) has argued that modal operator weak and strong supervenience are equivalent, respectively, to weak and strong supervenience for properties, when: (a) 'necessarily' is understood as equivalent to universal quantification over worlds; (b) the same worlds are quantified over in each occurrence of 'necessarily' within the respective definition, and (c) *A* and *B* are sets of properties closed under complementation, infinitary conjunction, and infinitary disjunction.

Consider (c). It is controversial whether the complement of a *B*-property is itself a *B*-property (Post, 1987). Descartes took having spatial extension to be a physical property, and maintained that minds lack spatial extension. It is odd to say that he was, thereby, committed to the view that minds have a physical property, namely, the property of lacking spatial extension (McLaughlin, 1995). It is, moreover, controversial whether complementation is even a property-forming operation; that is to say, whether if *P* is a property, then not-*P* is too. It is, furthermore, controversial whether even finitary disjunction (let alone infinitary disjunction) is a property-forming operation. Some metaphysicians would deny that if *P* and *Q* are both properties, then $P \vee Q$ is too (Armstrong 1978).

In any case, even if Kim's conclusion is true, it is not the case that weak and strong supervenience are equivalent, respectively, to modal operator weak and strong supervenience, when *A* and *B* are understood to be simply non-empty sets of properties. When *A* and *B* are so understood,

operator weak and strong supervenience imply weak and strong supervenience, respectively, but the converse implications fail. For both operator weak and operator strong imply that something cannot have a supervenient property without having a subvenient property, and neither weak nor strong supervenience has that implication. For example, one way in which two things can be exactly alike in respect of every *B*-property is by lacking any *B*-property whatsoever. For a dramatic example of the failure of equivalence, suppose that complementation is indeed a property-forming operation. Then property *P* strongly supervenes on property not-*P*, and conversely. For any individuals x_1 and x_2 and any worlds w_1 and w_2, x_1 in w_1 and x_2 in w_2 will be exactly alike in respect of property *P* if and only if they are exactly alike in respect of property not-*P*. But property *P* fails to either operator-weakly or operator-strongly supervene on not-*P* since if something has *P* then it is not the case that it has not-*P*. Thus weak and strong supervenience fail to imply, respectively, operator weak and strong supervenience. Indeed, strong supervenience fails even to imply modal operator weak supervenience. Strong supervenience and operator weak supervenience are thus logically independent: neither implies the other (McLaughlin, 1997a).

Unlike the definitions of weak and strong supervenience, the definitions of operator weak and operator strong supervenience do not define determinates of the determinable relation of dependent variation. Rather, they define what might be called dependence-determination relations: something's having any *A*-property depends on its having some *B*-property that determines that it has that *A*-property (McLaughlin, 1995). Dependence-determination relations are of interest in their own right. For example, realization requires a dependence-determination relation. Thus, the claim that *A*-properties are realized by *B*-properties implies that *A*-properties modal operator strongly supervene on *B*-properties.

Multiple Domain Supervenience

Sometimes two individuals *x* and *y* cannot differ in some *A*-respect without their differing in some *B*-respect. Then, there is single domain supervenience. The definitions of weak and strong supervenience are definitions of kinds of single domain supervenience. There is, however, also multiple domain supervenience (Kim, 1988). There is multiple domain supervenience when some *x* and *y* cannot differ in some *A*-respect without distinct things, *x** and *y**, that bear appropriate relationships to *x* and *y* respectively, differing in some *B*-respect. Consider clay statues. A reason for claiming that clay statues are not identical with the lumps of clay that compose them is that the statues and the lumps of clay can differ in their temporal properties; for example, the lump of clay might come into existence before the statue and continue to exist after the statue ceases to exist. And even if a lump of clay and a statue come into existence and cease to exist at the same time, lumps of clay and clay statues have different persistence conditions; and so, a clay statue and the lump of clay it is made of will still differ in their modal properties. For example, the lump of clay has the property of being capable of surviving flattening, while the statue may lack this property. For these reasons, it is often claimed that clay statues are not identical with lumps of clay, but only materially constituted by them. Suppose that this claim is correct. It would still be the case that two clay statues cannot differ in respect of their shapes without the lumps of clay that compose them differing in respect of their shapes, that clay statues cannot differ in their masses without the lumps of clay that compose them differing in their masses, and so on. If clay statues are indeed not identical with lumps of clay, then these are cases of multiple domain supervenience (McLaughlin, 1995).

Kim (1988) has formulated definitions of multiple domain weak and strong supervenience. Let D_1 and D_2 be two nonempty domains, and let *R* be a relation whose domain is D_1 and whose range is a subset of D_2. For any member *x* of D_1, let R/x denote the 'image' of *x* under *R* (that is, the set of all objects in D_2 to which *x* is related by *R*). For a world *w*, let the expression 'R/z in w' designate the image of *x* that *R* picks out in *w*. Multiple domain weak and strong supervenience can then be formulated as follows:

> *Multiple domain weak supervenience*: $\{A, D_1\}$ weakly supervenes on $\{B, D_2\}$ relative to relation *R* if and only if for any world *w* and for any *x* and *y* in D_1, if R/x and R/y in *w* are *B*-twins in *w* then *x* and *y* are *A*-twins in *w*. (5)

> *Multiple domain strong supervenience*: $\{A, D_1\}$ strongly supervenes on $\{B, D_2\}$ relative to relation *R* if and only if for any worlds w_1 and w_2 and for any x_1 and x_2 in D_1, if R/x_1 in w_1 is a *B*-twin of R/x_2 in w_2, then x_1 in w_1 is an *A*-twin of x_2 in w_2. (6)

When *R*, the coordinating relation, is identity, there

is single domain supervenience. But R need not be identity; it can, for instance, be the relation of material constitution, in which case there is multiple domain supervenience. Multiple domain supervenience is of interest when discussing dependent variation relations among properties of objects and the lumps of material that materially constitute them, or macrostructures (e.g. tables, trees, mountains) and the microstructures (e.g. structures of atoms) that materially constitute them.

Global Supervenience

So far, this article has focused on 'local' supervenience. Another much-discussed category of supervenience is global supervenience. Global property supervenience can be formulated as follows:

> *Global property supervenience*: A supervenes on B if and only if for any worlds w and w^*, if w and w^* have exactly the same worldwide pattern of distribution of B-properties, then they have exactly the same worldwide pattern of distribution of A-properties. (7)

This formulation raises the question of what is it for two worlds to have the same worldwide pattern of distribution of properties of a certain sort. In exploring this question, it is fruitful to appeal to the notion of there being an isomorphism between worlds, without actually specifying one. Let us say that an isomorphism I between two worlds w and w^* preserves B-properties if and only if for any x in w, x has a B-property F in w if and only if the image of x under I has F in w^* (McLaughlin, 1996, 1997a). Appealing to the notion of a property-preserving isomorphism between worlds, two kinds of global property supervenience can be distinguished (Stalnaker, 1996; McLaughlin, 1997a; Sider, 1999):

> *Weak global supervenience*: A-respects weakly globally supervene on B-respects if and only if for any worlds w and w^*, for every B-preserving isomorphism between w and w^*, there is an A-preserving isomorphism between them. (8)

> *Strong global supervenience*: A-respects strongly globally supervene on B-respects if and only if for any worlds w and w^*, every B-preserving isomorphism between w and w^* is an A-preserving isomorphism between them. (9)

Strong implies weak, but not conversely.

HISTORY

As McLaughlin (1995) notes:

> The term 'supervenience' has a vernacular use: Dr Samuel Johnson's *A Dictionary of the English Language* (1775), Vol 2 reports that 'supervene' derives from the Latin 'super', meaning 'on', 'above', or 'additional', and from the Latin verb 'venire' meaning 'to come'. And Dr Johnson's Dictionary defines 'supervene' as 'to come as an extraneous addition' and 'supervenient' as 'added, additional'. More recently, *Webster's New International Dictionary*, 3rd edition (1986), defines 'supervene' as 'coming or occurring as something additional, extraneous, or unexpected'.

The philosophical sense of 'supervenience' is different from its vernacular sense. It was introduced into the philosophy of mind in the following passage by Donald Davidson:

> Mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events exactly alike in all physical respects but differing in some mental respect (Davidson, 1970).

Davidson has been reported as saying that he got the term 'supervenience', used in a related way, from R M. Hare (Hare, 1952). Hare (1984) reports that he did not himself introduce the term into philosophy, but that the term was being used at Oxford in the 1940s. There has been some speculation that the British philosopher–biologist Lloyd Morgan (1923) introduced the term 'supervenience' in something like its current philosophical sense in the context of emergent evolution (Kim, 1990). Morgan spoke of emergent phenomena as 'supervening' on the phenomena from which they emerge. This appears, however, to be the vernacular sense: emergent phenomena occur unexpectedly or unpredictably from the phemomena from which they emerge. This vernacular sense of 'supervenience' persisted in the philosophy of mind until the late 1950s (McLaughlin, 1997b).

As McLaughlin (1997b, pp. 41–42) notes:

> Harry Lewis (1985, p.159n) reports that [Peter] Geach suggested to him 'that the term "supervenient" entered our philosophical vocabulary by way of Latin translations of Aristotle's *Nichomachean Ethics* 1174B31-3'. The Greek at 1174B31-3 reads: 'hos epiginomenon ti telos, hoion toise akmaiois he hora'. Robert Grosseteste's Latin translation of this passage translated 'epiginomenon' as 'supervenire' [Gauthier, 1973]. Sir David Ross translated 'epiginomenon' as 'supervenient'. In Ross's English 1174B31-3 becomes: 'as an end which supervenes as the bloom of youth does on those in the flower of their age'. The passage

occurs in the context of Aristotle's talking of certain properties 'naturally following' from other properties.

This sense of the term 'supervenience' is related to Hare's sense in (Hare, 1952). Perhaps Hare encountered the term 'supervenience' at Oxford in the 1940s in discussions of the *Nichomachean Ethics*, a book that he frequently cites. In any case, as noted above, Hare's use of 'supervenience' inspired Davidson's related use.

While Davidson introduced the term 'supervenience' in its current philosophical sense into the philosophy of mind, it should be noted that psychophysical supervenience theses appear fairly early on in discussions of the mind–body problem. For example, in 1778 Joseph Priestley claimed: 'different systems of matter, organized exactly alike, must make different beings, who would feel and think exactly alike in the same circumstances'. Their minds, therefore, would be exactly 'similar', but 'numerically different' (Priestley 1819/1977, p. 47). Indeed, supervenience theses – claims of dependent variation – can be found in many areas of western philosophy throughout its history.

SUPERVENIENCE AND REDUCTION

Supervenience is usually taken not to imply reduction, so that *A*-respects can supervene on *B*-respects without reducing to *B*-respects. If reduction has an explanatory dimension, that is surely right. For the fact that *A*-respects supervene on *B*-respects does not imply that *A*-respects are in any sense explainable in terms of *B*-respects. Also, if reduction is irreflexive, then supervenience fails to suffice for reduction. Some hold, however, that both strong supervenience of one sort of property on another sort of property across all metaphysically possible worlds and strong global supervenience of one sort of property on another across all such worlds, imply reduction in a non-epistemic, ontological sense of reduction. Be this as it may, reducibility implies supervenience. If *A*-respects reduce to *B*-respects, then *A*-respects supervene on *B*-respects in even the strongest sense of supervenience defined above.

Given that supervenience is required for reducibility, one way to argue that *A*-respects fail to reduce to *B*-respects is by arguing that *A*-respects fail to supervene on *B*-respects. A single possible example of *B*-twins that are not *A*-twins will show that. Thus, supervenience theses provide a useful way of testing claims of reducibility. Consider, for example, the claim that intentional mental properties (such as believing that *P*, desiring that *Q*, etc.) reduce to neurophysiological properties. That claim implies that two individuals exactly alike in their neurophysiological properties cannot differ in their intentional properties. Some philosophers would appeal to a 'twin Earth' thought experiment of the sort designed by Hilary Putnam (1976) and Tyler Burge (1979) to argue that this supervenience claim is false since there can be neurophysiological twins that are not intentional twins. Arguments that attempt to refute claims of reducibility by offering counterexamples to the supervenience theses implied by the claims are called arguments by appeal to false implied supervenience theses.

SUPERVENIENCE AND COGNITIVE SCIENCE

The philosophy of mind is an area of cognitive science in which the notion of supervenience is frequently used. There have been various attempts to define physicalism as a psychophysical supervenience thesis. For example, appealing to the notion of a non-alien world (i.e., a possible world that has only the same perfectly natural properties found in the actual world), David Lewis (1983) has defined physicalism as follows: any two non-alien worlds that are indiscernible physically are indiscernible in every respect. And there have been other attempts to define physicalism by appeal to supervenience (see, for example, (Jackson, 1993) and (Chalmers, 1996)).

Whether any psychophysical supervenience thesis that counts as physicalism is true is, of course, a controversial question. Another controversial question is whether, in order to be justified in accepting a psychophysical supervenience thesis, there must be some explanation of why the thesis is true. It has been said that when a supervenience thesis is explainable, there is 'superdupervenience' (Horgan, 1993). The question is whether a justifiable physicalism would require a superdupervenience psychophysical claim.

Supervenience has also been invoked in one of the major debates in psychosemantics. Internalism in psychosemantics is the thesis that the meaning (or content) of any kind of mental representation strongly supervenes with metaphysical necessity on intrinsic properties of psychological subjects that possess mental representations of that kind. Externalism in psychosemantics, which is often defended by 'twin Earth' thought experiments, is the thesis that the meanings of certain kinds of mental representation fail to so supervene on intrinsic properties of psychological subjects that possess representations of that kind. Externalists

typically claim that the meanings of mental representations fail to even weakly supervene on intrinsic properties of psychological subjects.

References

Armstrong DM (1978) *A Theory of Universals*. Cambridge, UK: Cambridge University Press.

Burge T (1979) Individualism and the mental. In: French P, Euhling T and Wettstein (eds) *Midwest Studies in Philosophy*, vol. IV. Minneapolis, MN: University of Minnesota Press.

Chalmers D (1996) *The Conscious Mind*. New York, NY: Oxford University Press.

Davidson D (1970) Mental events. In: Foster L and Swanson JW (eds) *Experience and Theory*, pp. 79–101. Amherst, MA: University of Massachusetts Press.

Gauthier RA (1973) *Aristotleles Latinus*, vol. XXVI. Leiden, the Netherlands: Brill Academic Publishers.

Hare RM (1952) *The Language of Morals*. Oxford, UK: Oxford University Press.

Hare RM (1984) Supervenience. *Proceedings of the Aristotelian Society* Supplementary Volume **58**: 1–16.

Horgan T (1993) From supervenience to superdupervenience: meeting the demands of a material world. *Mind* **102**: 555–586.

Jackson F (1993) Armchair metaphysics. In: O'Leary-Hawthorne J and Michael M (eds) *Philosophy in Mind*. Dordrecht, the Netherlands: Kluwer.

Kim J (1984) Concepts of supervenience. *Philosophy and Phenomenological Research* **45**: 153–176.

Kim J (1987) 'Strong' and 'global' supervenience revisted. *Philosophy and Phenomenological Research* **48**: 315–326.

Kim J (1988) Supervenience for multiple domains. *Philosophical Topics* **16**: 129–150.

Kim J (1990) Supervenience as a philosophical concept. *Metaphilosophy* **21**: 1–27.

Kim J (1993) Postscripts on supervenience. In: Kim J (ed.) *Supervenience and Mind: Selected Philosophical Essays*, pp. 161–171. Cambridge, UK: Cambridge University Press.

Lewis DK (1983) New work for a theory of universals. *Australasian Journal of Philosophy* **61**: 343–377.

Lewis HA (1985) Is the mental supervenient on the physical? In: Vermazen B and Hintikka M (eds) *Essays on Davidson: Actions and Events*, pp. 159–172. Oxford, UK: Clarendon Press.

McLaughlin BP (1995) Varieties of supervenience. In: Savellos E and Yalcin U (eds) *Supervenience: New Essays*, pp. 16–59. Cambridge, UK: Cambridge University Press.

McLaughlin BP (1996) Supervenience. In: Borchert (ed.) *Encyclopedia of Philosophy Supplement*. London, UK: Macmillan.

McLaughlin BP (1997a) Supervenience, vagueness, and determination. *Philosophical Perspectives* **11**: 209–230.

McLaughlin BP (1997b) Emergence and supervenience. *Intellectica* **25**: 25–43.

Morgan L (1923) *Emergent Evolution*. London, UK: Williams & Norgate.

Post J (1987) *The Faces of Existence*. Ithaca, NY: Cornell University Press.

Priestley J and Price R (ed.) (1819/1977) *A Free Discussion of the Doctrines of Materialism, and Philosophical Necessity, in the Theological and Miscellaneous Works of Joseph Priestley*, vol. IV. Millwood, NJ: Kraus Reprint Co.

Putnam H (1976) The meaning of 'meaning'. In: Gunderson K (ed.) *Language, Mind, and Knowledge*. *Minnesota Studies in the Philosophy of Science*, vol. VII. Minneapolis, MN: University of Minnesota Press.

Sider TR (1999) Global supervenience and identity across worlds. *Philosophy and Phenomenological Research* **59**: 913–937.

Stalnaker R (1996) Varieties of supervenience. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. X, pp. 221–241. Cambridge, MA: Blackwell.

Further Reading

Haugeland J (1982) Weak supervenience. *American Philosophical Quarterly* **19**: 93–101.

Hellman G and Thompson F (1975) Physicalism, ontology, determination, and reduction. *Journal of Philosophy* **72**: 551–564.

Horgan T (1982) Supervenience and microphysics. *Pacific Philosophical Quarterly* **63**: 29–43.

Kim J (1993) *Supervenience and Mind: Selected Philosophical Essays*. Cambridge, UK: Cambridge University Press.

Klagge J (1988) Supervenience: ontological and ascriptive. *Australiasian Journal of Philosophy* **64**: 461–470.

Paull CP and Sider TR (1992) In defense of global supervenience. *Philosophy and Phenomenological Research* **32**: 830–845.

Savellos E and Yalcin U (eds) *Supervenience: New Essays*. Cambridge, UK: Cambridge University Press.

Symbol-grounding Problem

Intermediate article

Stevan Harnad, University of Quebec, Montreal, Quebec, Canada

CONTENTS

The symbol-grounding problem is related to the problem of how words get their meanings, and of what meanings are. The problem of meaning is in turn related to the problem of consciousness, or how it is that mental states are meaningful.

WORDS AND MEANINGS

We know since Frege (1952/1892) that the thing that a word refers to (its referent) is not the same as its meaning. This is most clearly illustrated using the proper names of concrete individuals (but it is also true of names of kinds of things and of abstract properties): (1) 'Tony Blair', (2) 'the UK's current prime minister', and (3) 'Cherie Blair's husband' all have the same referent, but not the same meaning.

Some have suggested that the meaning of a (referring) word is the rule or features one must use in order to pick out its referent. In that respect, (2) and (3) come closer to wearing their meanings on their sleeves, because they seem to be explicitly stating a rule for picking out their referents (find whoever is the UK's current PM, or whoever is Cherie's current husband). But that does not settle the matter, because there is still the problem of the meaning of the components of the rule ('UK', 'current', 'PM', 'Cherie', 'husband'), and how to pick *them* out.

Perhaps 'Tony Blair' (or better still, just 'Tony') does not have this component problem, because it points straight to its referent, but how? If the meaning is the rule for picking out the referent, what is that rule, when we come down to non-decomposable components?

It is probably unreasonable to expect us to know the rule, explicitly at least. Our brains need to have the 'know-how' to follow the rule, and actually pick out the intended referent, but they need not know how they do it consciously. We can leave it to cognitive science and neuroscience to find out and then explain how.

THE MEANS OF PICKING OUT REFERENTS

So if we take a word's meaning to be the means of picking out its referent, then meanings are in our brains. If we use 'meaning' in a wider sense, we may want to say that meanings include both the referents themselves and the means of picking them out. So if a word (say, 'Tony-Blair') is located inside an entity, then its meaning consists of both the means that that entity uses to pick out its referent, and the referent itself: a big causal nexus between a head, a word inside it, an object outside it, and whatever 'processing' is required to connect the inner word to the outer object.

But what if the 'entity' in which a word is located is not a head but a piece of paper? What is its meaning then? Surely all the (referring) words on this page, for example, have meanings, just as they have referents.

CONSCIOUSNESS

Here is where the problem of consciousness rears its head. For there would be no connection at all between scratches on paper and any intended referents if there were no minds mediating those intentions, via their internal means of picking out those referents.

So the meaning of a word on a page is 'ungrounded', whereas the meaning of a word in a head is 'grounded' (by the means that cognitive neuroscience will eventually reveal to us), and thereby mediates between the word on the page and its referent.

COMPUTATION

What about the meaning of a word inside a computer? Is it like the word on a page or like the word in a head? This is where the symbol-grounding problem comes in. Is a dynamic process transpiring in a computer more like the static paper page, or more like another dynamical system, the brain?

There is a school of thought according to which the computer is more like the brain – or rather, the brain is more like the computer. According to this view, called 'computationalism', that future theory about how the brain picks out its referents, the theory that cognitive neuroscience will eventually arrive at, will be a purely computational one. A computational theory is a theory at the software level; it is essentially a computer program. And software is 'implementation-independent'. That means that whatever it is that a program is doing, it will do the same thing no matter what hardware it is executed on. The physical details of the implementation are irrelevant to the computation; any hardware that can run the computation will do (Pylyshyn, 1984).

THE TURING TEST

A computer can execute any computation. Hence once computationalism finds the right computer program, the same one that our brain is running when there is meaning transpiring in our heads, then meaning will be transpiring in that computer too. How will we know that we have the right computer program? It will have to be able to pass the Turing Test (TT) (Turing, 1950). That means it will have to be capable of corresponding with any human being for a lifetime as a pen-pal, without ever being in any way distinguishable from a real pen-pal.

SEARLE'S CHINESE ROOM ARGUMENT

It was in order to show that computationalism is incorrect that Searle (1980) formulated his celebrated 'Chinese Room Argument', in which he pointed out that if the Turing test were conducted in Chinese, then he himself, Searle (who does not understand Chinese), could execute the same program that the computer was executing without knowing what any of the words he was processing meant. So if there's no meaning going on inside him when he is implementing the program, there's no meaning going on inside the computer when it is the one implementing the program either, computation being implementation-independent.

How does Searle know that there is no meaning going on when he is executing the TT-passing program? Exactly the same way he knows whether there is or is not meaning going on inside his head under any other conditions: he understands the words of English, whereas the Chinese symbols that he is manipulating according to the program's rules mean nothing to him. And there is no one else in there for them to mean anything to. They are like the ungrounded words on a page, not the grounded words in a head.

Note that in pointing out that the Chinese words would be meaningless to him under those conditions, Searle has appealed to consciousness. Otherwise one could argue that there *would* be meaning going on in his head under those conditions, but he would simply not be aware of it. This is called the 'System Reply', and Searle rightly rejects it as simply a reiteration, in the face of negative evidence, of the very thesis that is on trial in his thought experiment: are words in a running computation like the ungrounded words on a page, meaningless without the mediation of brains, or are they like the grounded words in brains?

In this either/or question, the (still undefined) word 'ungrounded' has implicitly relied on the difference between inert words on a page and consciously meaningful words in our heads. And Searle is reminding us that under these conditions (the Chinese TT), the words in his head would not be consciously meaningful, hence they would still be as ungrounded as the inert words on a page.

So if Searle is right, that (1) the words on a page, and in any running computer program, including a TT-passing computer program, are meaningless in and of themselves, and hence that (2) whatever it is that the brain is doing to generate meaning, it cannot be just implementation-independent computation, then what *is* the brain doing to generate meaning (Harnad, 2001a)?

FORMAL SYMBOLS

To answer this question we have to formulate the symbol-grounding problem (Harnad, 1990). First we have to define 'symbol': a symbol is any object that is part of a symbol system. (The notion of symbol in isolation is not a useful one.) A symbol system is a set of symbols and rules for manipulating them on the basis of their shapes (not their meanings). The symbols are systematically interpretable as having meanings, but their shape is arbitrary in relation to their meaning.

A numeral is as good an example as any: numerals (e.g. '1', '2', '3') are part of a symbol system (arithmetic) consisting of formal rules for combining them into well formed strings. '2' means what we mean by 'two', but its shape in no way resembles 'two-ness'. The symbol system is systematically interpretable as making true statements about numbers (e.g. '$1 + 1 = 2$').

It is critical to understand that the symbol-manipulation rules are based on shape rather than meaning (the symbols are treated as primitive and undefined, insofar as the rules are concerned), yet the symbols and their rule-based combinations are all meaningfully interpretable. It should be evident in the case of formal arithmetic, that although the symbols make sense, that sense is in our heads and not in the symbol system. The numerals in a running desk calculator are as meaningless as the numerals on a page of hand calculations. Only in our minds do they take on meaning (Harnad, 1994).

This is not to deprecate the property of systematic interpretability: we select and design formal symbol systems (algorithms) precisely because we want to know and use their systematic properties; the systematic correspondence between scratches on paper and quantities in the universe is a remarkable and extremely powerful property. But it is not the same as meaning, which is a property of certain things going on in our heads.

NATURAL LANGUAGE AND THE LANGUAGE OF THOUGHT

Another symbol system is natural language. On paper, or in a computer, it too is just a formal symbol system, manipulable by rules based on the arbitrary shapes of words. In the brain, meaningless strings of squiggles become meaningful thoughts. I am not going to be able to say what had to be added in the brain to make them meaningful, but I will suggest one property, and point to a second.

One property that the symbols on static paper or even in a dynamic computer lack that symbols in a brain possess is the capacity to pick out their referents. This is what we were discussing earlier, and it is what the hitherto undefined term 'grounding' refers to. A symbol system alone, whether static or dynamic, cannot have this capacity, because picking out referents is not just a computational property; it is a dynamical (implementation-*de*pendent) property.

To be grounded, the symbol system would have to be augmented with non-symbolic, sensorimotor capacities – the capacity to interact autonomously with that world of objects, events, properties, and states that its symbols are systematically interpretable (by us) as referring to. It would have to be able to pick out the referents of its symbols, and its sensorimotor interactions with the world would have to fit with the symbols' interpretations (Cangelosi *et al*., 2000).

The symbols, in other words, need to be connected directly to (i.e. grounded in) their referents; the connection must not be dependent only on the connections made by the brains of external interpreters like us. The symbol system alone, without this capacity for direct grounding, is not a viable candidate for being whatever it is that is really going on in our brains (Cangelosi and Harnad, 2001).

ROBOTICS

The necessity of groundedness, in other words, takes us from the level of the pen-pal Turing test, which is purely symbolic (computational), to the robotic Turing test, which is hybrid symbolic/sensorimotor (Harnad, 2000). Meaning is grounded in the robotic capacity to detect, identify, and act upon the things that words and sentences refer to. (*See* **Categorical Perception**)

But if groundedness is a necessary condition for meaning, is it a sufficient one? Not necessarily, for it is possible that even a robot that could pass the Turing test, 'living' amongst the rest of us indistinguishably for a lifetime, would fail to have in its head what Searle has in his: it could be a zombie, with no one home, feeling feelings, experiencing meanings.

And that is the second property, consciousness, towards which I wish merely to point, rather than to suggest what functional capacities it must correspond to (I have no idea what those might be – I rather think it is impossible for consciousness to have any independent functional role except on pain of telekinetic dualism). Maybe robotic TT capacity is enough to guarantee it, maybe not. In any case, there is no way we can hope to be any the wiser (Harnad, 2001b).

References

Cangelosi A and Harnad S (2001) The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. *Evolution of Communication* (Special Issue on Grounding). [http://cogprints.soton.ac.uk/documents/disk0/00/00/20/36/index.html]

Cangelosi A, Greco A and Harnad S (2000) From robotic toil to symbolic theft: grounding transfer from

entry-level to higher-level categories. *Connection Science* **12**(2): 143–162. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/47/index.html]

Frege G (1952/1892) On sense and reference. In: Geach P and Black M (eds) *Translations of the Philosophical Writings of Gottlob Frege*. Oxford, UK: Blackwell.

Harnad S (1990) The symbol grounding problem. *Physica D* **42**: 335–346. [http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.sgproblem.html]

Harnad S (1994) Computation is just interpretable symbol manipulation: cognition isn't. *Minds and Machines* **4**: 379–390. [http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad94.computation.cognition.html]

Harnad S (2000) Minds, machines and Turing: the indistinguishability of indistinguishables. *Journal of Logic, Language, and Information* **9**(4): 425–445. [http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.turing.html]

Harnad S (2001a) What's wrong and right about Searle's Chinese room argument? In: Bishop M and Preston J (eds) *Essays on Searle's Chinese Room Argument*. Oxford, UK: Oxford University Press. [http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.searle.html]

Harnad S (2001b) No easy way out. *The Sciences* **41**(2): 36–42. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/24/index.html]

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT/Bradford.

Searle JR (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* **3**(3): 417–457. [http://www.cogsci.soton.ac.uk/bbs/Archive/bbs.searle2.html]

Turing AM (1950) Computing machinery and intelligence. *Mind* **49**: 433–460. [Reprinted in *Minds and Machines*. A Anderson (ed.), Engelwood Cliffs, NJ: Prentice Hall, 1964.] [http://cogprints.soton.ac.uk/abs/comp/199807017]

Further Reading

Barsalou LW (1999) Perceptual symbol systems. *Behavioural and Brain Sciences* **22**(4): 577.

Bartell B and Cottrell GW (1991) A model of symbol grounding in a temporal environment. *AAAI Sprint Symposium Workshop on Connectionism and Natural Language Processing*, pp. 142–147.

Cummins Robert (1996) Why there is no symbol grounding problem. In: *Representations, Targets, and Attitudes*, chap. IX. Cambridge, MA: MIT Press.

Freeman WJ (2000) A neurobiological interpretation of semiotics: meaning, representation, and information. *Information Sciences* **124**(1–4): 93–102.

Glenberg AM and Robertson DA (2000) Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* **43**(3): 379–401.

Grumbach A (1996) Grounding symbols into perceptions. *Artificial Intelligence Review* **10**(1–2): 131–146.

Jackson SA and Sharkey NE (1996) Grounding computational engines. *Artificial Intelligence Review* **10**(1–2): 65–82.

Jung D and Zelinsky A (2000) Grounded symbolic communication between heterogeneous cooperating robots. *Autonomous Robots* **8**(3): 269–292.

MacDorman KF (1998) Feature learning, multiresolution analysis, and symbol grounding. *Behavioral and Brain Sciences* **21**(1): 32.

McKevitt P (1996) From Chinese rooms to Irish rooms: new words on visions for language. *Artificial Intelligence Review* **10**(1–2): 49–63.

Malcolm C and Smithers T (1990) Symbol grounding via a hybrid architecture in an autonomous assembly system. In: Maes P (ed.) *Designing Autonomous Agents*, pp. 145–168. Cambridge, MA: MIT Press.

Plunkett K, Sinha C, Moller MF and Strandsby O (1992) Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science: Journal of Neural Computing, Artificial Intelligence & Cognitive Research* **4**(3–4): 293–312.

Prem E (1995) Symbol grounding and transcendental logic. In: Niklasson L and Boden M (eds) *Current Trends in Connectionism*, pp. 271–282. Hillsdale, NJ: Lawrence Erlbaum.

Sun R (2000) Symbol grounding: a new look at an old idea. *Philosophical Psychology* **13**(2): 149–172.

Takeuchi I and Furuhashi T (1998) Acquisition of manipulative grounded symbols for integration of symbolic processing and stimulus-reaction type parallel processing. *Advanced Robotics* **12**(3): 271–287.

Tani J (1996) *Does Dynamics Solve the Symbol Grounding Problem of Robots? An Experiment in Navigation Learning. Learning in Robots and Animals – Working Notes*. AISB'96 workshop, Brighton, UK.

Thompson E (1997) Symbol grounding: a bridge from artificial life to artificial intelligence. *Brain and Cognition* **34**(1): 48–71.

Symbolic versus Subsymbolic

Intermediate article

Chris Eliasmith, University of Waterloo, Waterloo, Ontario, Canada
William Bechtel, Washington University, St Louis, Missouri, USA

CONTENTS

There are two competing approaches to computational modeling of cognition: the symbolic approach, based on language-like representations, and the subsymbolic (connectionist) approach, inspired by neuroscience.

INTRODUCTION

Explanations in cognitive science often take the form of computational models. Since the inception of the discipline in the 1950s, there has been a competition between two approaches to computational modeling: the symbolic approach, and the subsymbolic, connectionist, or 'artificial neural network' approach. These two approaches differ in their conception both of computation and of the representations that figure in computations. Not surprisingly, then, symbolic and connectionist explanations of cognitive behavior tend to be quite different.

HISTORY

Both approaches have a long pedigree. Indeed, both were advocated in talks on 11 September 1956 at the Symposium on Information Theory, which George Miller (1979) has identified as the birth event of cognitive science. Exemplifying the symbolic approach, Newell and Simon presented a paper describing a 'logic theory machine', while Rochester, Holland, Haibt, and Duda presented a connectionist implementation of Hebb's neuropsychological theory of cell assemblies.

Even in these prototypical instantiations, the symbolic and subsymbolic computational models are clearly distinct. Symbolic models typically employ syntactically structured representations and logic-like rules that are sensitive to those structures. Thus, Newell and Simon's model adopted the language of symbolic logic, and relied on algorithms that were sensitive to the logical operators in that language. Subsymbolic models, on the other hand, typically employ neuron-like units that excite or inhibit each other according to their current activation values.

However, specifying precisely what makes a computational model symbolic or connectionist is not easy. Before attempting this, it is helpful to examine how the two approaches have developed historically.

Historical Development of the Symbolic Tradition

The early history of the symbolic approach is one of successes. Not only did Newell and Simon's 'logic theorist' serve as an early existence proof that computers could be employed to model processes thought to require intelligence in humans, but shortly afterwards they developed an even more sophisticated artificial intelligence program called 'general problem solver' (GPS) that did more than just prove theorems. GPS was applied to a number of reasoning problems, including the Tower of Hanoi problem and problems of cryptoarithmetic. A central goal of this research was to model human reasoning, and GPS was able to provide a detailed 'trace' of state transitions that closely matched the verbal reports of a human working on the same problem (Newell and Simon, 1976b).

GPS and other early symbolic programs (of which several examples are presented in (Minsky, 1968) were applied to linguistically-posed problems, but there were also symbolist successes in other domains, such as interacting with physical objects. Around 1970, Terry Winograd developed a program called SHRDLU which functioned in a simulated world of blocks. The program responds to natural language queries about what blocks are present and how they are arranged, and it can carry

out simulated manipulations of the virtual blocks. Advocates claimed that the program seemed to understand its limited domain: it carries out appropriate actions based on commands, and responds correctly to questions about its environment. (For an argument that it did not really exhibit understanding, see (Dreyfus, 1972.)

These successes (albeit in limited domains such as logic and block worlds) led a number of researchers to begin developing general frameworks for symbolic cognitive modeling. Two early examples, which have been developed over the course of several decades, are John Anderson's ACT and ACT* (Anderson, 1983 and 1990), and Allen Newell's Soar (Newell, 1990). Both have, at their heart, a production system, that is, a system of 'if–then' (condition–production) rules that operate on syntactically structured symbolic representations stored in working memory.

The physical symbol system hypothesis

The early successes and seemingly unlimited promise of the symbolic approach led two of its early proponents, Newell and Simon (1976a), to propose what they called the physical symbol system hypothesis, according to which a physical symbol system has the necessary and sufficient means for general intelligent action.

According to Newell and Simon, a physical symbol system is subject to the laws of physics and operates on physical patterns. These patterns, or symbols, can be physically related so as to comprise expressions (symbol structures). The system itself is composed of such structures and a collection of processes that operate on these structures to produce other structures. With unlimited memory and unlimited time, physical symbol systems are capable of universal computation (Newell, 1990), just like the more famous Turing machine (Turing, 1950).

The physical symbol system hypothesis advances a very strong claim that (1) any intelligent system is a physical symbol system, and (2) any physical symbol system of 'sufficient size' can exhibit intelligence (Newell and Simon, 1976a). Newell (1990) also introduces the notion of a knowledge system. The symbols in such a system encode knowledge about that system's goals, actions and environment, and the relations between these. Knowledge systems are governed by a single law: take actions to attain goals using all of the knowledge available to the system. Newell's statement of the relation between knowledge systems and symbol systems is clear: symbol systems are supposed to 'realize knowledge systems by implementing representation laws so that the symbol structures encode the knowledge about the external world'. Newell's conclusion is that 'humans are symbol systems that are at least modest approximations of knowledge systems'. Thus, for the purposes of explaining cognition, a description in terms of a physical symbol system is a complete description and, also for these purposes, the fact that such a system might be implemented in a brain (i.e. a bunch of interconnected neurons) is irrelevant.

This view was widely accepted by many early cognitive scientists, for a number of reasons. Firstly, in computer science, universal computers had been proven to be able to compute all computable functions. If physical symbol systems are universal computers, and people don't compute noncomputable functions, then symbol systems have the computational resources necessary for exhibiting human intelligence. Secondly, in linguistics, Chomsky argued that grammars of natural languages required precisely that kind of computational power. Thirdly, in psychology and philosophy, the 'language of thought' hypothesis – the hypothesis that all thought is a language-like 'mentalese' and subject to a mental grammar – suggested that symbol processing was the basis of human cognition (Fodor, 1975). Lastly, such a picture of human function is consistent with our everyday folk psychology, which characterizes people's mental states as attitudes (e.g., beliefs, desires) towards propositions. If these very propositions are stored and operated on in symbolic models, then those models offer an explanation of why accounts of people's behavior in terms of beliefs and desires succeed much of the time.

Historical Development of the Subsymbolic Connectionist Approach

Even at its heyday, the symbolic approach was not the only one. One of the major early developers of the alternative, subsymbolic approach to understanding cognition was Frank Rosenblatt, the developer of a computational device called a 'perceptron' (Rosenblatt, 1958). This device consists of a grid of photocells, randomly connected to associators that collect and process the electrical impulses produced when objects are placed in front of the photocells. The associator units are, in turn, connected to response units by weighted connections. The weights determine how much each associator unit influences each response unit. These weights are modifiable by a learning procedure designed to correct any errors in the response

units. Rosenblatt (1962) proved the 'perceptron convergence theorem', which states that a certain learning procedure will succeed in finding weights that permit the network to produce correct responses, provided such weights exist. In fact, the device was successful at recognizing various letters of the alphabet, and provided an interesting alternative to symbolic models. However, excitement over this less logic-like approach soon diminished, mostly as a result of Minsky and Papert's (1968) demonstration that there are classes of computations that models of this kind can't perform. Rosenblatt knew that this problem could be overcome by inserting additional layers of trainable weights. However, he failed to develop a generalized learning rule for finding these weights.

It wasn't until the mid-1980s that a learning rule – back propagation – was developed that applied to multiple layers of connections. Around the same time, deep difficulties with the symbolic approach began to become apparent. Consequently, the subsymbolic approach began to attract more attention. Now known by a variety of names, including 'connectionism', '*artificial neural networks*', and 'parallel distributed processing', the subsymbolic approach has enjoyed its own history of successes (McClelland and Rumelhart, 1986; Bechtel and Abrahamsen, 1991). New and better models of language processing, object recognition, speech recognition, and motor control were proposed by connectionists. The stage was set for a conflict between computational paradigms.

THE CONFLICT BETWEEN THE SYMBOLIC AND CONNECTIONIST APPROACHES

Differences Between the Symbolic and Connectionist Approaches

The differences between these two approaches can be grouped under two headings: representational differences and computational differences.

Representational differences

In symbolic models, the basic units of representation (i.e., symbols) are semantically interpretable, and are often actually words in natural languages. By contrast, in the subsymbolic approach, the basic representational units are below the level of semantic interpretation (Smolensky, 1988). Only patterns of activation distributed over large populations of network units are semantically interpretable. (Such patterns are called vectors.) Moreover, any individual unit can be a constituent of many vectors, each of which has a different interpretation.

On occasion, the difference between symbolic and subsymbolic approaches is presented as a difference between continuous and discrete, or digital and analog representation. Paul Churchland has construed the difference in this way:

> So-called 'digital' or *discrete-state* computing machines are limited by their nature to representing and computing mathematical functions that range over the *rational* numbers. … This is a potentially severe limitation on the abilities of a digital machine. … Therefore, functions over real numbers cannot strictly be computed or even represented within a digital machine. They can only be approximated. (Churchland, 1995)

In fact, subsymbolic models almost always employ digital representations. There is nothing inherent in the 'digitalness' of these models that makes them only approximations of 'actual' subsymbolic models. Furthermore, there is evidence that biological networks do not employ continuous representations (Eliasmith, 2000). Thus, it is misleading and inappropriate to present the distinction between symbolic and subsymbolic systems as one between digital and analog representations.

Computational differences

Computational differences between symbolic and connectionist modeling can be understood as turning on the preferred metaphor for cognition. Symbolic theorists typically see cognizers as analogous to serial digital computers: there is a central processing unit (i.e., higher cognition), input systems (i.e., peripheral sensory systems), and output systems (i.e., motor systems) (see, for example (Newell, 1990)). Connectionists, by contrast, see cognizers as inherently brain-like: massively parallel processors that rely on uniform functional units. Thus, they claim that they work with a 'neurally inspired model of cognitive processes' (McClelland and Rumelhart, 1986), or with 'brain-style computation' (Rumelhart, 1989).

This apparent distinction is a result of a deeper distinction between the symbolic and subsymbolic approaches. Indeed, we could translate symbolic models to run on parallel computers. Such a translation is far from trivial since it requires decomposing a problem into independently solvable pieces and maintaining communication between parallel processors working on these pieces. However, there is a more basic difference: the kind of computation in a parallel instantiation of a symbolic program is not at all like that in a subsymbolic model. Computation in a symbolic system is governed by

many different, explicitly encoded rules (as in a production system). In subsymbolic models, computation is governed by one or a few mathematical rules that rely on numeric parameters to determine how the activation of one unit affects the activation of other units. Although in simple models these parameters might be predetermined, as models become more complex, researchers typically rely on learning rules to determine them. In addition, the set of possible computations a connectionist model can realize is governed by the overall design of the system (for example, what sets of nodes are connected to what other sets, and how densely). Generally, this system 'architecture' has been determined by the designer, but some connectionist models now rely on tools such as genetic algorithms to develop the pattern of connectivity in the model.

Connectionist Critiques of Symbolic Models

As noted above, by the 1980s some cognitive scientists had begun to feel frustrated by the perceived limitations of the symbolic approach, limitations which the subsymbolic connectionist alternative seemed able to overcome. The first limitation stemmed from the very fact that symbolic representations are generally sentential. Initially, this seemed to be a strength, since the problems that people solve are often easily represented linguistically.

However, many basic cognitive tasks, such as the classification of sensory stimuli (like taste, touch, smell, sound, and sight) are not presented linguistically. These tasks involve responding to statistical regularities in the environment, to which symbolic representations are generally not sensitive (Smolensky, 1995). Subsymbolic representations, on the other hand, are equally well suited to modeling representations in any modality, and have been successfully applied to visual (Qian and Sejnowski, 1988; Raeburn, 1993), olfactory (Skarda and Freeman, 1987), auditory (Lazzaro and Mead, 1989) and tactile problems. Moreover, Kosslyn (1980 and 1994) and others have convincingly argued that a large class of reasoning problems entail operations on images, which are a paradigmatically nonlinguistic representation (although symbolists such as Pylyshyn (1973, 1981) have argued for a propositional construal of visual images). Subsymbolic models are well suited to modeling this kind of perceptual processing.

Furthermore, symbolic systems tend to be brittle. Any degradation of a symbolic representation (such as a piece of computer code) can radically alter or destroy the whole system's functionality. Such brittleness is psychologically unrealistic. People often exhibit behavior that indicates partial retrieval of representations: 'tip of the brain' recall, prosopagnosia (loss of face recognition), and 'blindsight' are all instances in which mental representations are incomplete. Although such incomplete representations reduce performance, they generally don't completely destroy it or radically alter it. Human performance tends to degrade gradually as information is lost.

Subsymbolic connectionist models are not very brittle. Whereas minor damage to a symbolic conceptual network causes loss of entire concepts, damage to a connectionist network causes only a loss in accuracy, as observed in human subjects (Churchland and Sejnowski, 1992).

There are also difficulties in modeling learning in symbolic models. In general, learning in symbolic systems depends entirely on the current set of symbols in the system. The most common example of this kind of learning is 'chunking', in which many related symbols are grouped together and represented by another single symbol. Chunking can effectively capture some kinds of human learning, but is deficient for explaining how new primitive symbols themselves are learned. By contrast, learning in subsymbolic models involves strengthening or weakening connections, an approach which is well suited to modeling low-level perceptual learning: the kind of learning that may explain the development of new symbols.

Finally, symbolic systems suffer from their natural suitability for serial processing. Serial processing is quite compatible with the discrete, nonstatistical character of symbolic representations. If the relations between elements are irrelevant to their effect on cognition, then processing them one at a time makes sense. However, if the relations between elements are important for determining how they can or should be processed (i.e. if 'holistic' considerations are relevant) then parallel processing is more appropriate. Many cognitive problems involve soft constraints (constraints that ideally would all be satisfied, but some of which may have to be violated to arrive at the best solution). For example, generating analogies depends on the ability to match a current situation to another situation we have encountered before, where no previous situation is a perfect match.

In general, the representational and computational commitments of the subsymbolic connectionist approach have avoided many of the difficulties with the symbolic approach. Holistic processing,

handling of soft constraints, statistical sensitivity and gradual degradation are all more successfully incorporated into subsymbolic systems.

Symbolist Critiques of Connectionist Models

Many symbolists claim that, whatever the shortcomings of the symbolic approach, the cost of the subsymbolic connectionist approach is too high. Fodor and Pylyshyn (1988) argue that the failure to employ structured representations and structure-sensitive operations leaves subsymbolic models unable to explain truly cognitive phenomena.

Fodor and Pylyshyn focus on two features of cognition that they claim connectionist models cannot explain: productivity and systematicity. Productivity is the capacity to produce arbitrarily many expressions. This is achieved in symbolic systems by means of recursion (e.g. 'John told Sue who told Mary who told Bob …'). A physical symbol system can construct these sorts of structures since the rules used to compose representations employ (possibly recursive) processes defined over atomic elements (symbols). Fodor and Pylyshyn contend that by eschewing compositional rules and not recognizing the importance of syntactic structure, subsymbolic connectionist approaches fail to be productive.

Systematicity concerns the relations that exist between representations. Fodor and Pylyshyn note that anyone who has mastered English must admit all of the following as sentences: 'John loves the girl', 'the girl loves the girl', 'John loves John', and 'the girl loves John'. They claim that in order to explain the systematicity found both in language and thought, symbolists can appeal to syntactic structure. Structured schemas such as 'noun-phrase transitive-verb noun-phrase' can be applied to any noun-phrase and transitive-verb. Symbolic models routinely employ such structured representations and so can readily explain systematicity. However, since subsymbolic connectionist models do not employ this kind of representation, Fodor and Pylyshyn argue that they lack the resources needed to explain systematicity.

In characterizing the representational structures that can explain the systematicity and productivity of cognition, Fodor and Pylyshyn rely on the notion of a compositional syntax and semantics. They thereby emphasize that the particular syntactic rules that are employed are appropriate for ensuring semantic coherence. A compositional representation is one in which meaning is an additive function of the meaning of its components. Thus, in a sentence like 'the ball is red', the meaning of the structure can be derived from the meaning of each of the component lexical items (words). Furthermore, each lexical item makes the same semantic contribution to each expression in which it occurs. This, claims the symbolist, is typical of both language and thought. This 'compositionality' is a problem for subsymbolic connectionist models because population activation vectors, the semantic units for such models, have no particular structure. It is not clear how an activation vector that represents a sentence would be related to an activation vector that represents each of the words in the sentence. The sentence vector is not built up as a concatenation of word vectors. Connectionist models, then, will have difficulty accounting for some basic aspects of human cognition. All of the failings of the subsymbolic connectionist approach derive from the lack of structured representations.

New Kinds of Connectionist Model

Symbolic theorists have identified potential serious shortcomings of the subsymbolic connectionist approach. Connectionists cannot solve these problems merely by implementing a physical symbol system in a connectionist architecture. Such an implementation is in principle possible, since in both approaches it is possible to design universal computational systems (i.e. universal Turing machines). However, for a connectionist model merely to implement a symbolic one is for the connectionist to admit that subsymbols are themselves irrelevant for modeling cognition, and that nothing is gained by turning to subsymbols for a new and different understanding of cognition.

However, some theorists do think that an adequate response to the symbolist's challenge does require some appeal to symbolic systems. They advocate hybrid systems in which symbolic structures are implemented in subsymbolic structures in such a manner that the subsymbolic implementation plays a cognitive role. The goal in such hybrid systems is for the subsymbolic implementation to provide statistically sensitive processing, while allowing the implemented symbol system to provide structurally sensitive processing. Some researchers have developed interesting hybrid models (Barnden, 1995; Sun and Alexandre, 1997). But it is difficult to develop a principled means of determining the right combination of structurally sensitive and statistically sensitive processes.

Another, more purely subsymbolic, response to the symbolists' objections is to attempt to develop a

significant degree of structural sensitivity directly in subsymbolic models. Two such approaches have achieved some successes.

First, Pollack (1990) has developed 'recurrent auto-associative memory' (RAAM) networks for recursively creating compressed vectors of input vectors. If linguistic strings (structured representations) are employed as inputs, these networks can create representations that are not explicitly compositional, but from which the compositional structure of the input can be recovered. Moreover, other structurally sensitive operations (for example, transforming compressed representations of active sentences into compressed representations of passive sentences) can be performed directly on the compressed strings without first extracting the compositional structures. Since the compressed representations are not themselves syntactically structured but seem to perform tasks that require information about such structures, they are construed as functionally compositional (van Gelder, 1990).

A second way to account for productivity and systematicity in subsymbolic connectionist models employs distributed representation schemes based on vector products (Smolensky, 1990). Essentially, these schemes include operations for combining and extracting vectors. Two vectors, *A* and *B*, can be combined to produce a third vector *C*. From *C* either *A* or *B* can subsequently be extracted, given the other. This provides a way of building up structures in vector representations. Because of the generality of the defined operations, problems of systematicity no longer arise for this representational scheme. Productivity is achieved because these operations can be recursively applied.

In fact, these kinds of schemes have been successfully used to model high-level cognitive function. For example, Eliasmith and Thagard (2001) have proposed a model of analogy based on Plate's representational scheme (Plate, 1994). Analogical mapping is notoriously structurally sensitive and is a typically cognitive task (Gentner and Toupin, 1986; Holyoak and Thagard, 1995). This kind of model is thus an existence proof of the ability of the subsymbolic connectionist approach to handle structural sensitivity.

However, these subsymbolic attempts to account for productivity and systematicity do not provide for unlimited realizations of either of these features of cognition. For example, as longer strings are recursively supplied to a RAAM network, errors result. And the vector product approach restricts the size of the vectors. The product vector *C*, is not a concatenation of the vectors *A* and *B*. Rather, *A* and *B* are partially encoded into *C*. So the meaning of *C* is not a 'linear' function of the meanings of its component parts.

Whether those limitations mean that connectionists have failed to respond adequately to the symbolists' challenges depends on the nature of compositional human mental representations, and how productive and systematic human thought is. The suggestion that mental representations might be less compositional than symbolists claim is supported by examples taken from natural language, a typically symbolist source of data. The meanings of most colloquial expressions ('it's raining cats and dogs', 'what a couch potato', 'break a leg') are clearly not additive functions of the meanings of their components. Meanings are so flexible that, for example, the words 'unravel' and 'ravel' can mean the same thing despite opposite 'compositional' meanings. Given such examples, it is not so clear that lack of complete compositionality is a valid objection to the subsymbolic approach. Since compositionality is intended to explain productivity and systematicity, and since subsymbolic models only provide a measure of these properties to a degree, subsymbolic theorists prefer to ask just how much productivity and systematicity is needed to model human cognitive abilities.

SUMMARY

The symbolic and subsymbolic connectionist approaches provide distinct ways of modeling cognitive phenomena. Both have had significant successes at explaining certain aspects of human cognition. But both have been the subject of serious criticism, and neither approach obviously explains more aspects than the other. Although recent advances in connectionist models have made some headway into symbolist domains, it is not yet clear whether these advances achieve sufficient systematicity and productivity to model human cognition.

References

Anderson JR (1983) The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson JR (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Barnden JA (1995) High-level reasoning, computational challenges for connectionism, and the Conposit solution. *Applied Intelligence* **5**: 103–135.

Bechtel W and Abrahamsen A (1991) *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Oxford: Blackwell.

Churchland P (1995) *The Engine of Reason, the Seat of the Soul: A Philosophical Journey Into the Brain*. Cambridge, MA: MIT Press.
Churchland PS and Sejnowski TJ (1992) *The Computational Brain*. Cambridge, MA: MIT Press.
Dreyfus HL (1972) *What Computers Can't Do: A Critique of Artificial Reason*. New York, NY: Harper and Row.
Eliasmith C (2000) Is the brain analog or digital? The solution and its consequences for cognitive science. *Cognitive Science Quarterly* **1**(2): 147–170.
Eliasmith C and Thagard P (2001) Integrating structure and meaning: a distributed model of analogical mapping. *Cognitive Science* **25**(2): 245–286.
Fodor J (1975) *The Language of Thought*. New York, NY: Crowell.
Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.
Gentner D and Toupin C (1986) Systematicity and surface similarity in the development of analogy. *Cognitive Science* **10**: 277–300.
Holyoak K and Thagard P (1995) *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.
Kosslyn S (1980) *Image and Mind*. Cambridge, MA: Harvard University Press.
Kosslyn S (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.
Lazzaro J and Mead C (1989) A silicon model of auditory localization. *Neural Computation* **1**: 47–57.
McClelland JL and Rumelhart DE (eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. vol. 2. Cambridge, MA: MIT Press/ Bradford.
Miller GA (1979) *A Very Personal History*. Cambridge, MA, MIT Center for Cognitive Science, Occasional Paper No. 1.
Minsky M and Papert S (1968) *Perceptrons*. Cambridge, MA: MIT Press.
Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
Newell A and Simon H (1976a) Computer science as empirical enquiry: symbols and search. *Communications of the Association for Computing Machinery* **19**: 113–126.
Newell A and Simon HA (1976b) GPS, a program that simulates human thought. In: Feigenbaum E and Feldman J (eds) *Computers and Thought*. New York, NY: McGraw-Hill.
Plate TA (1994) Distributed representations and nested compositional structure. [Technical Report, Department of Computer Science, University of Toronto.]
Pollack J (1990) Recursive distributed representations. *Artificial Intelligence* **46**: 77–105.
Pylyshyn Z (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin* **80**: 1–24.
Pylyshyn Z (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **87**: 16–45.
Qian N and Sejnowski TJ (1988) Learning to solve random-dot seterograms of dense and transparent surfaces with recurrent backpropagation. *Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann.
Raeburn P (1993) Reverse engineering the human brain. *Technology Review*.
Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408.
Rosenblatt F (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.
Rumelhart DE (1989) The architecture of mind: a connectionist approach. In: Posner MI (ed) *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.
Rumelhart DE and McClelland JL (eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. vol. 1. Cambridge, MA: MIT Press/ Bradford.
Skarda CA and Freeman WJ (1987) How brains make chaos to make sense of the world. *Behavioral and Brain Sciences* **10**: 161–195.
Smolensky P (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–23.
Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46**: 159–217.
Smolensky P (1995) Computational models of mind. In: Guttenplan S (ed) *A Companion to the Philosophy of Mind*. Cambridge, MA: Blackwell.
Sun R and Alexandre F (1997) *Connectionist–Symbolic Integration: From Unified to Hybrid Approaches*. Mahwah, NJ: Erlbaum.
Turing AM (1950) Computing machinery and intelligence. *Mind* **59**: 433–460.
van Gelder T (1990) Compositionality: a connectionist variation on a classical theme. *Cognitive Science* **14**: 355–84.

Further Reading

Anderson JA and Rosenfeld E (eds) (1988) *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.
Anderson JA, Pellionisz A and Rosenfeld E (1990) *Neurocomputing 2: Directions for Research*. Cambridge, MA: MIT Press.
Ballard DH (1997) *An Introduction to Natural Computation*. Cambridge, MA: MIT Press.
Bechtel W and Abrahamsen A (2002) *Connectionism and the Mind II: Parallel Processing, Dynamics, and Evolution in Networks*. Oxford: Blackwell.
Clark A (1989) *Micocognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press.
Cummins R and Cummins DD (2000) *Minds, Brains, and Computers: The Foundations of Cognitive Science*. Oxford: Blackwell.

McLeod P, Plunkett K and Rolls E (1998) *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Minsky M (1968) *Semantic Information Processing*. Cambridge, MA: MIT Press.

Rolls ET and Treves A (1998) *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rosenbloom PS, Laird JE and Newell A (eds) (1993) *The Soar Papers: Research on Integrated Intelligence*. Cambridge, MA: MIT Press.

Wagman M (1997) *The General Unified Theory of Intelligence: Its Central Conceptions and Specific Application to Domains of Cognitive Science*. Westport, CT: Praeger.

Syntax and Semantics: Formal Approaches

Advanced article

Glyn Morrill, Universitat Politècnica de Catalunya, Barcelona, Spain

CONTENTS

Syntax concerns the relations between words in sentences; semantics concerns the meanings of words and sentences.

INTRODUCTION

Language is a vehicle for the expression of meanings via forms: an association between forms and meanings. On the side of forms we could speak of phonetic form, phonological form, prosodic form, signifier, and so on, depending on the construal of the term and the detail targeted. Let us suppose that we have a set of vocabulary items or words; then forms will comprise at least a string of words together with a prosodic contour. For the purposes of this article we assume that they comprise only strings of words, and refer to them as 'expressions'.

Consider the set of all expressions over some vocabulary. Some will be sentences, some not; some will be verb phrases, some not; and so on. We can picture language as a family of subsets of this domain, one member of the family for each part of speech, with a distinguished member of sentences. There will be interrelations between the members of this family: for example, sentences that contain noun phrases, and noun phrases that contain sentences. Mathematically, we can look at the family of all sets of expressions as an algebra in which there are operations such as language concatenation or product:

$$L_1 \cdot L_2 = \{s_1 + s_2 | s_1 \in L_1 \text{ and } s_2 \in L_2\} \quad (1)$$

On the side of meanings we could speak of content, sense, logical form, semantic form, signifier, and so on, depending on the construal of the term and the detail targeted. Let us suppose that, whatever meanings are, they comprise at least truth conditions, and that those of sentences are related by logical consequence. For the purposes of this article we will consider meanings as just terms of higher-order logic, and refer to them as 'logical forms'.

Consider the set of all logical forms, which will be partitioned into different types: those that represent propositions for sentences, those that represent properties for verb phrases, and so on. Now we can picture language as a family of relations between expressions and logical forms. Such a view of language as (a family of) associations between forms (signifiers) and meanings (signifieds) derives from de Saussure (1916), wherein each association signifier–signified is called a 'sign'. Note that the association between forms and meanings can be one-to-many (ambiguity) and many-to-one (paraphrase).

Thinking just of forms, the following program suggests itself: to try to specify the class of expressions that are sentences of a language. Such a program of 'formal syntax' was launched by Chomsky (1957), instantiated by his 'transformational grammar'. Thinking also of meanings, the program would include: to try to specify in addition the logical forms of the expressions. Such a program of 'formal semantics' was launched by Montague (1974).

One methodological option would be to observe that since meaning is the less tangible dimension, and forms represent a hard enough challenge, we had better focus our sights on just syntax and in this way divide the task at hand. Another would be to observe that since the expression of meaning is the point of language, we had better attempt to pursue both tasks in parallel, for surely it is the expressive and communicative *raison d'être* of language that must guide us through the many alternatives that will present themselves and in which explanation for why things are so will reside.

This article illustrates some basic tools and methods of formal syntax and semantics and shows how they can work together. In the process it indicates various courses that can be and have been taken in formal syntax and formal semantics.

PHRASE STRUCTURE GRAMMAR

Let us assume some category symbols for different parts of speech as in Table 1.

We are picturing a language as a family of sets of expressions (strings of words) indexed by such category symbols: thus, S = {**John** + **runs, Mary** + **runs, John** + **loves** + **Mary,** …}, N = {**John, Mary**, …}, etc. Let there be lexical assignments w: A of words to categories, meaning that $w \in A$:

- **John**: N
- **Mary**: N
- **man**: CN
- **that**: C
- **runs**: VP
- **loves**: TV
- **gives**: TTV
- **thinks**: SV
- **that**: RELPRO

Then relations between categories can be expressed by rules of the form $A \rightarrow A_1 \ldots A_n$, meaning that $A_1 \cdot \ldots \cdot A_n \subseteq A$. For example:

- S → N VP
- CP → C S
- VP → TV N
- VP → TTV N N
- VP → SV CP
- CN → CN RELCLS

The lexicon together with the rules entail certain facts about the language model. For example, **John** ∈ N, **runs** ∈ VP, and N · VP ⊆ S entail that **John** + **runs** ∈ S. Such reasoning can be structured into derivation trees ('phrase markers') like those in Figure 1.

Table 1. Category symbols for parts of speech

Symbol	*Meaning*
S	declarative sentence
N	proper name
CN	count noun
C	complementizer
CP	complementizer phrase (complementized sentence)
VP	verb phrase
TV	transitive verb
TTV	ditransitive verb
SV	sentential verb
RELCLS	relative clause
RELPRO	relative pronoun

Such phrase markers are the syntactic structures of phrase structure grammar: concrete geometrical representations of the essential structure by which expressions are deemed to be grammatical.

Consider object relativization:

(a) man that$_i$ Mary loves e_i
(b) man that$_i$ John thinks that Mary loves e_i
(c) man that$_i$ Mary thinks that John thinks that Mary loves e_i (2)

The noun *man* modified by the relative clause is construed as the object of the transitive verb *loves*, which may be embedded arbitrarily far from the relative pronoun. Such unbounded dependencies were originally taken as motivating, in transformational grammar, transformations mapping from deep phrase markers, in which the left-extracted element appears at the extraction site, to surface phrase markers, in which it appears in its surface position, by means of 'movement'. However, Gazdar (1981) showed how a simple generalization of phrase structure grammar can characterize unbounded dependency.

Let the category indices include structured category symbols B/A where B and A are categories (read: B 'slash' A). Roughly, B/A means 'a B missing an A'. Add a 'metarule' for deriving rules from rules:

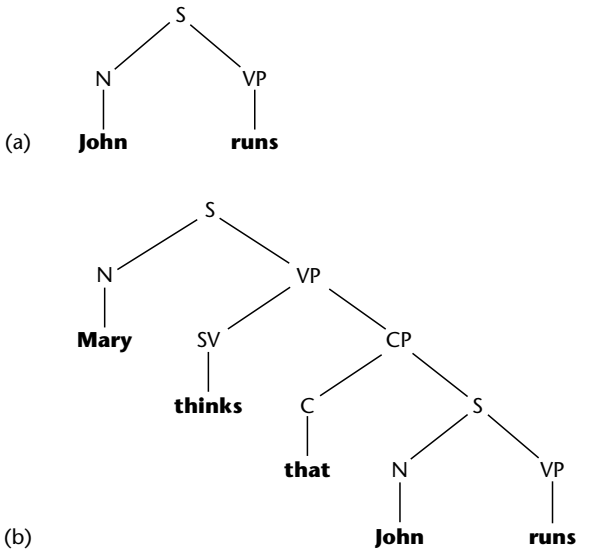


Figure 1. Examples of phrase markers. (a) Phrase marker for the sentence *John runs*. (b) Phrase marker for the sentence *Mary thinks that John runs*.

$$\frac{C \rightarrow \Gamma\ B}{C/A \rightarrow \Gamma\ B/A} \tag{3}$$

Add also rules for rewriting A/A as the empty string (after all, the empty expression is 'an A missing an A'), and a rule combining a filler with an S/N:

$$\mathrm{N/N} \rightarrow \varepsilon \tag{4}$$

$$\mathrm{RELCLS} \rightarrow \mathrm{RELPRO\ S/N} \tag{5}$$

Then example 2(b), and unbounded dependencies in general, are generated as illustrated in Figure 2.

The line of thought according to which perhaps, after all, transformations are not indispensable to unbounded dependencies, and phrase structure grammar might suffice, developed into 'generalized phrase structure grammar' (Gazdar *et al.*, 1985) and 'head-driven phrase structure grammar' (HPSG) (Pollard and Sag, 1987). (*See* **Construction Grammar**; **Phrase Structure Grammar, Head-driven**)

Another line of thought, according to which, for example, passivization is treated nontransformationally (by a lexical rule), led to 'lexical-functional grammar' (LFG) (Bresnan, 1982). Both HPSG and LFG make extensive use of the notion of 'unification', a device for treating universal quantification in automatic theorem proving, and central to logic programming languages such as PROLOG. Transformational grammar developed in the 1980s into 'government–binding theory' and in the 1990s into 'minimalism'. (*See* **Government–Binding Theory**; **Lexical-Functional Grammar**)

CATEGORIAL GRAMMAR

Consider the following operations in the algebra of sets of expressions (Lambek, 1958, 1988) (the first is just language concatenation):

$$A \bullet B = \{s_1 + s_2 | s_1 \in A \text{ and } s_2 \in B\} \tag{6}$$

$$A\backslash B = \{s | \text{ for all } s' \in A,\ s' + s \in B\} \tag{7}$$

$$B/A = \{s | \text{ for all } s' \in A,\ s + s' \in B\} \tag{8}$$

We read the operators $\bullet$, $\backslash$, and $/$ as 'product', 'under', and 'over', respectively. We assume some 'primitive types', for example S (declarative sentence), N (proper name), CN (count noun) and CP (complementizer phrase). We refer to the terms built over the primitive types by the operators as 'types'.

Let there be lexical assignments w: A of words to types, meaning that $w \in A$:

- **John**: N
- **Mary**: N
- **man**: CN
- **that**: CP/S
- **runs**: N\S
- **loves**: (N\S)/N
- **gives**: (N\S)/(N • N)
- **thinks**: (N\S)/CP
- **that**: (CN\CN)/(S/N)

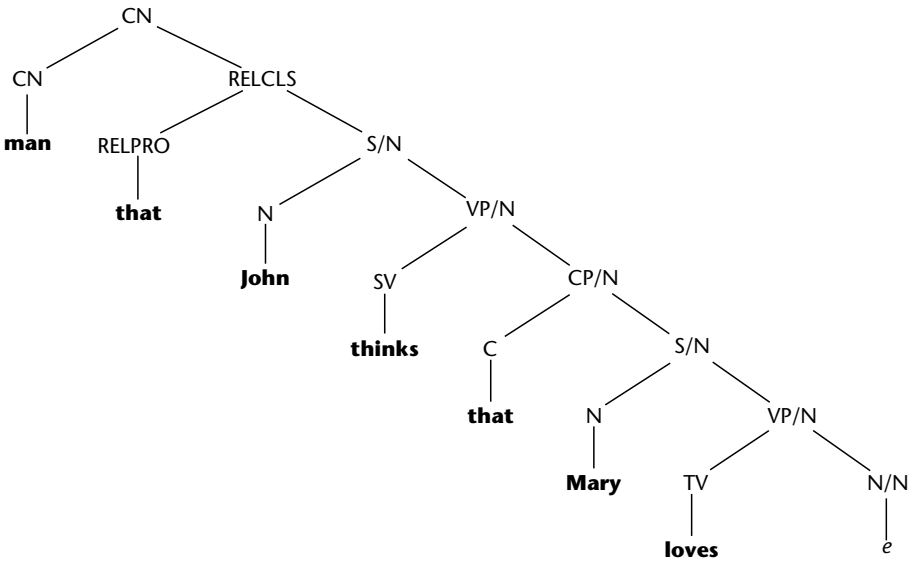


Figure 2. Phrase structure analysis of *man that$_i$ John thinks that Mary loves e$_i$*.

Then other facts are entailed. For example, from **John** $\in$ N and **runs** $\in$ N\S it follows that **John** + **runs** $\in$ S. Similarly, we have from the assignments above that **Mary** + **thinks** + **that** + **John** + **runs** $\in$ S. Note that in contrast to phrase structure grammar, the language model is determined solely by the lexicon (together with the meanings of the operators as operations): there are no syntactic rules in the definition. That is, categorial grammar is lexicalist whereas phrase structure grammar is nonlexicalist.

The grammar already generates the unbounded dependencies of example 2. Thus, sentence 2(b) is obtained because **John** + **thinks** + **that** + **Mary** + **loves** $\in$ S/N. Indeed the metarule of equation 3 is valid when we read 'slash' as 'over': if $\Gamma \cdot B \subseteq C$ then $\Gamma \cdot B/A \subseteq C/A$.

The syntactic structures of categorial grammar (i.e., the concrete geometrical representations of the essential structure by which expressions are deemed to be grammatical) are 'proof nets' (Roorda, 1991), which originated in linear logic (Girard, 1987). A 'polar type' is a type together with a 'polarity' – input ($^\bullet$) or output ($^\circ$). A 'polar type tree' is the result of unfolding a polar type into 'links' as shown in Figure 3.

For example, the polar type tree for (CN\CN)/(S/N)$^\bullet$ is shown in Figure 4.

To derive the theorem that words of types $A_1, \ldots, A_n$ constitute (in that order) an expression of type A, construct first the 'proof frame' comprising the sequence of polar type trees $A^\circ, A_1^\bullet, \ldots, A_n^\bullet$. A 'proof structure' is the result of connecting (by an 'axiom link') each leaf in a proof frame to exactly one other with the same primitive type of opposite polarity. A proof net is a proof structure that satisfies certain conditions. Examples of proof nets are shown in Figure 5.

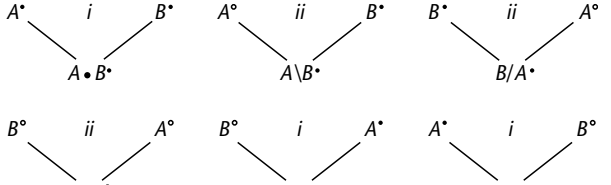


Figure 3. Unfolding a polar type into links.

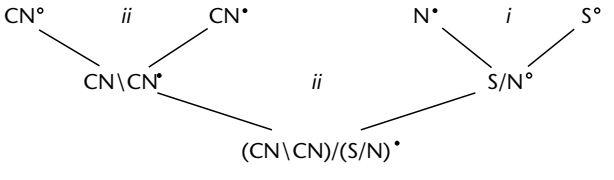


Figure 4. The polar type tree for (CN\CN)/(S/N)$^\bullet$, obtained by unfolding up to primitive leaves according to Figure 3.

Note that these proof nets are planar: that is, they can be drawn in the half-plane without crossing lines. Here, for a proof structure to be a proof net it must be planar; in addition, every cycle must cross both edges of some *i*-link, which is why we label links *i* or *ii*.

LAMBDA CALCULUS AND HIGHER-ORDER LOGIC

Given sets X and Y, the 'functional exponentiation' X^Y is the set of all functions from Y to X, and the 'Cartesian product' $X \times Y$ is the set of all ordered pairs comprising an element of X (first) and an element of Y (second).

Assume some set of primitive (semantic) types, for example o and e (for propositions and entities). The 'semantic types' are the terms that can be built over the primitive semantic types by the operators $\rightarrow$ and &. Given sets D_o and D_e associated to the primitive semantic types, a domain D_τ is associated to each semantic type τ as follows:

$$D_{\tau \rightarrow \tau'} = D_{\tau'}^{D_\tau} \tag{9}$$

$$D_{\tau \& \tau'} = D_\tau \times D_{\tau'} \tag{10}$$

Let there be a set of 'constants' and a set of 'variables' of each type, a 'valuation' mapping each constant of each type τ to an element of D_τ, and an 'assignment' mapping each variable of each type τ to an element of D_τ. Then the 'terms' ϕ of the 'typed lambda calculus' of each type τ and their denotations $[\![\phi]\!]_f^g \in D_\tau$ are defined as follows:

- If a is a constant of type τ, then a is a term of type τ and $[\![a]\!]_f^g$ is $f(a)$.
- If x is a variable of type τ, then x is a term of type τ and $[\![x]\!]_f^g$ is $g(x)$.
- If ϕ is a term of type $\tau \rightarrow \tau'$ and ψ is a term of type τ, then $(\phi\ \psi)$ is a term of type τ' and $[\![(\phi\ \psi)]\!]_f^g$ is $[\![\phi]\!]_f^g([\![\psi]\!]_f^g)$ (functional application).
- If x is a variable of type τ and ϕ is a term of type τ', then $\lambda x\phi$ is a term of type $\tau \rightarrow \tau'$, and $[\![\lambda x\phi]\!]_f^g$ is the function $h \in D_{\tau \rightarrow \tau'}$ such that $h(m) = [\![\phi]\!]_f^{g'}$, where g' is just like g but maps x to m (functional abstraction).
- If χ is a term of type $\tau \,\&\, \tau'$, then $\pi_1\chi$ is a term of type τ, $\pi_2\chi$ is a term of type τ', $[\![\pi_1\chi]\!]_f^g$ is $\mathrm{fst}([\![\chi]\!]_f^g)$, and $[\![\pi_2\chi]\!]_f^g$ is $\mathrm{snd}([\![\chi]\!]_f^g)$ (first and second projection).
- If ϕ is a term of type τ and ψ is a term of type τ', then (ϕ, ψ) is a term of type $\tau \,\&\, \tau'$ and $[\![(\phi\ \psi)]\!]_f^g$ is $\langle [\![\phi]\!]_g^f, [\![\psi]\!]_f^g \rangle$ (pairing).

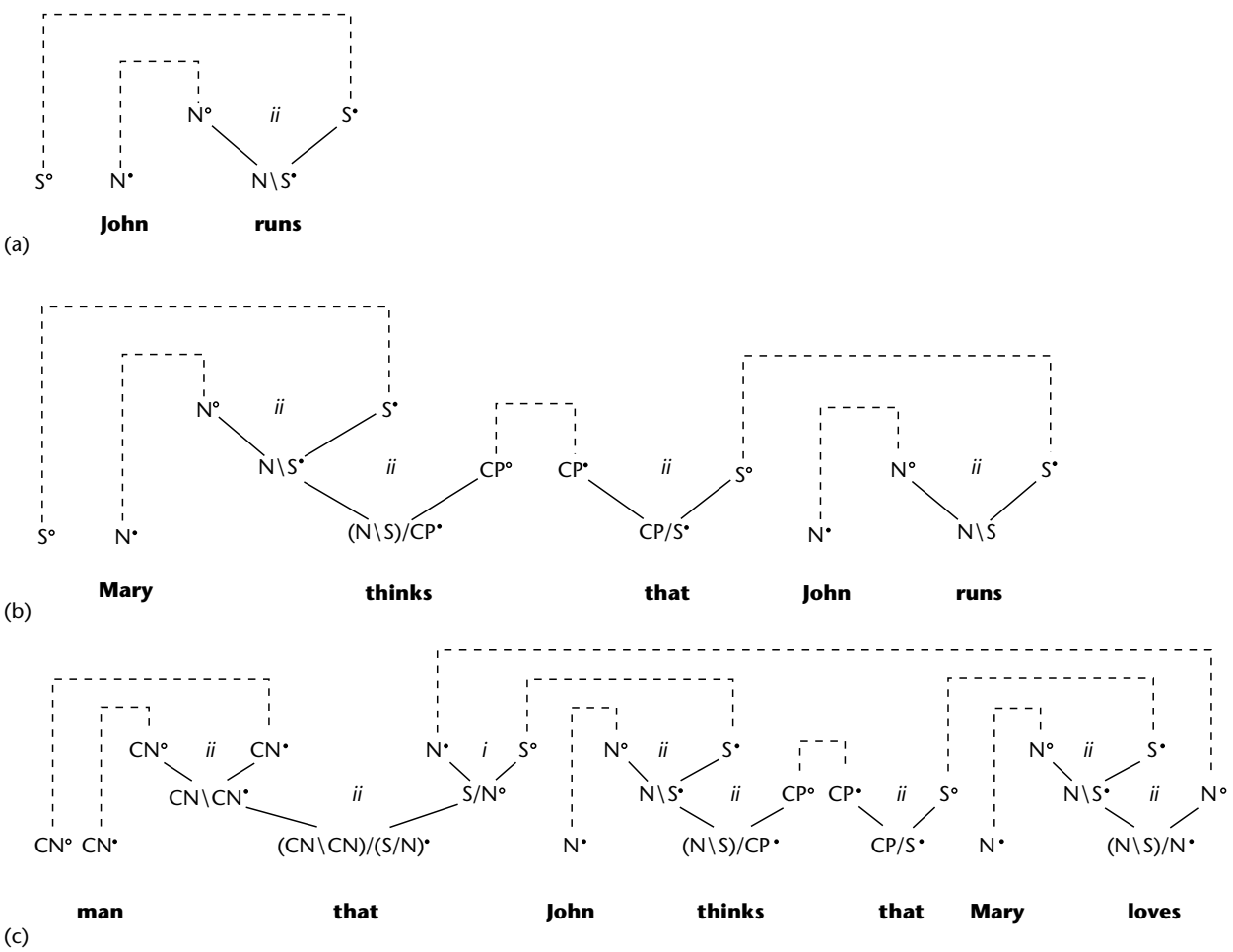


Figure 5. Examples of proof nets. (a) Proof net for *John runs*. (b) Proof net for *Mary thinks that John runs*. (c) Proof net for *man that*$_i$ *John thinks that Mary loves* e_i.

Let $\phi\{\psi/x\}$ be the result of substituting by ψ the free occurrences of x in ϕ; we say it is 'free' if and only if no variable becomes bound in the process of substitution. The following terms are equivalent (with respect to every valuation and assignment they share denotations):

- $\lambda x\phi \equiv \lambda y(\phi\{y/x\})$ if y is not free in ϕ and $\phi\{y/x\}$ is free (α-conversion)
- $(\lambda x\phi\ \psi) \equiv \phi\{\psi/x\}$ if $\phi\ \{\psi/x\}$ is free (β-conversion)
- $\lambda x(\phi\ x) \equiv \phi$ if x is not free in ϕ (η-conversion)
- $\pi_1(\phi, \psi) \equiv \phi$ and $\pi_2(\phi, \psi) = \psi$ (β-conversion)
- $(\pi_1\phi, \pi_2\phi) \equiv \phi$ (η-conversion)

We go from typed lambda calculus to higher-order logic by taking some of the constants to have denotational constraints (making them 'logical constants'). For example, if the elements of D_o are themselves sets, we might assume a logical constant AND of type $(o \& o) \rightarrow o$ which, rather than being assigned any function in this type, is assigned always the function of intersection on its arguments. And a consequence relation $\models$ is induced on the propositional terms: $\phi \models \psi$ if and only if $[\![\phi]\!]^g_f \subseteq [\![\psi]\!]^g_f$ for every f and g. In this way, terms of higher-order logic can be associated with expressions in an attempt to model their logical properties. For example, we might associate logical forms with expressions as follows. (*See* **Semantics and Pragmatics: Formal Approaches**)

(a) Mary thinks that John runs
(b) $((think\ (run\ j))\ m)$ (11)

(a) man that John thinks that Mary loves
(b) $\lambda z(\text{AND}\ ((man\ z), ((think\ ((love\ z)\ m))\ j)))$ (12)

The attempt to associate expressions of natural language systematically with such logical forms was initiated by Montague (1974), working with the notion of 'possible world' (Dowty *et al*., 1981); an alternative approach, centered on the notion of

'situation', was considered by Barwise and Perry (1983). Montague's focus was on semantics, and his syntax, though precise, was informal. Gamut (1991) dubbed the kind of grammar that associates logical forms to expressions 'logical grammar'.

LOGICAL GRAMMAR

We can develop a logical grammar associating logical forms with expressions by assigning semantic operations to each rule of a phrase structure grammar. Such a design, keying semantic interpretation on syntactic structure, is referred to as 'compositional'. First, we will need to fix a semantic type for each category. For example, we could use the associations in Table 2:

Lexical items will be associated with closed terms of the corresponding types. For example:

- **John** – j : N
- **Mary** – m : N
- **man** – *man* : CN
- **that** – λxx : C
- **runs** – *run* : VP
- **loves** – *love* : TV
- **gives** – *give* : TTV
- **thinks** – *think* : SV
- **that** – $\lambda x \lambda y \lambda z$(AND (($y$ z), (x z))) : RELPRO

Rules can be associated with semantic operations by assigning distinct variables to each daughter and assigning to the mother a term the free variables of which are those assigned to the daughters. For example:

- (y x): S $\rightarrow$ x: N y: VP
- (x y): CP $\rightarrow$ x: C y: S
- (x y): VP $\rightarrow$ x: TV y: N
- (x (y, z)): VP $\rightarrow$ x TTV y: N z: N
- (y x): CN $\rightarrow$ x: CN y: RELCLS

Then the examples in Figure 1 have the following derivations (we represent trees growing upwards, and lambda terms are simplified when possible):

$$\dfrac{\dfrac{\textbf{John}}{j\text{: N}} \quad \dfrac{\textbf{runs}}{run\text{: VP}}}{(run\ j)\text{: S}} \tag{13}$$

$$\dfrac{\textbf{Mary} \quad \dfrac{\textbf{thinks} \quad \dfrac{\dfrac{\textbf{that}}{\lambda xx\text{: C}} \quad \dfrac{\dfrac{\textbf{John}}{j\text{: N}} \quad \dfrac{\textbf{runs}}{run\text{: VP}}}{(run\ j)\text{: S}}}{(run\ j)\text{: CP}}}{(think\ (run\ j))\text{: VP}}}{((think\ (run\ j))\ m)\text{: S}} \tag{14}$$

The metarule annotated with semantics is:

$$\frac{\chi : C \rightarrow \Gamma\ y : B}{\lambda x \chi\{(z\ x)/y\} : C/A \rightarrow \Gamma\ z : B/A} \tag{15}$$

And:

$$\lambda xx : \text{N/N} \rightarrow \varepsilon \tag{16}$$

$$(x\ y)\text{: RELCLS} \rightarrow x\text{: RELPRO}\ y\text{: S/N} \tag{17}$$

Then example 2(b) is generated as in Figure 6.

The combination of semantics with syntax in categorial grammar is dubbed by Morrill (1994) 'type logical grammar' (Moortgat, 1997; Carpenter, 1997). (*See* **Categorial Grammar and Formal Semantics**)

Table 2. Semantic types associated with categories

Category	*Semantic type*
S	o
N	e
CN	e $\rightarrow$ o
C	o $\rightarrow$ o
CP	o
VP	e $\rightarrow$ o
TV	e $\rightarrow$ (e $\rightarrow$ o)
TTV	(e & e) $\rightarrow$ (e $\rightarrow$ o)
SV	o $\rightarrow$ (e $\rightarrow$ o)
RELCLS	(e $\rightarrow$ o) $\rightarrow$ (e $\rightarrow$ o)
RELPRO	(e $\rightarrow$ o) $\rightarrow$ ((e $\rightarrow$ o) $\rightarrow$ (e $\rightarrow$ o))

TYPE LOGICAL GRAMMAR

For categorial grammar the type map T from syntactic types to semantic types is such that:

$$T(A \bullet B) = T(A)\ \&\ T(B) \tag{18}$$

$$T(A\backslash B) = T(B/A) = T(A) \rightarrow T(B) \tag{19}$$

A mapping for primitive semantic types must be chosen – for example, $T(\text{S}) = \text{o}$, $T(\text{N}) = \text{e}$, $T(\text{CN}) = \text{e} \rightarrow \text{o}$, $T(\text{CP}) = \text{o}$.

When we include semantics, the categorial operations are over signs, which are pairs of forms and meanings:

$$A \bullet B = \{(s_1 + s_2, (m_1, m_2)) | (s_1, m_1) \in A \text{ and } (s_2, m_2) \in B\} \tag{20}$$

$$A\backslash B = \{(s, m) | \text{ for all } (s', m') \in A, (s' + s, m(m')) \in B\} \tag{21}$$

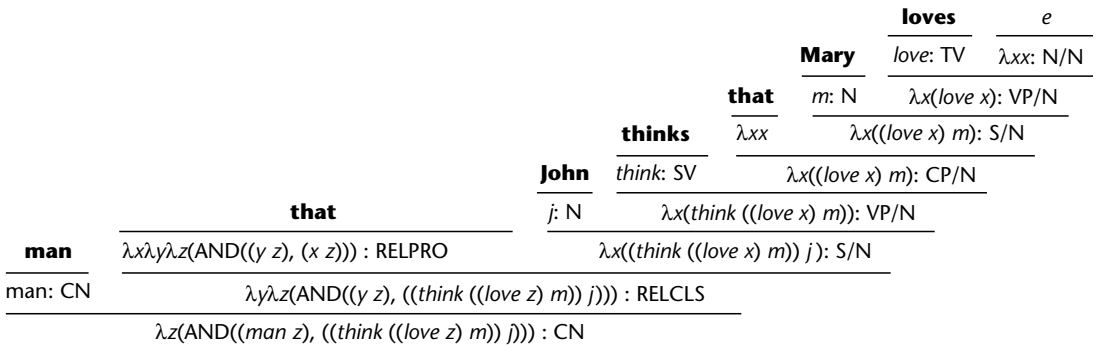


Figure 6. Logical analysis of *man that*$_i$ *John thinks that Mary loves* e_i.

$$B/A = \{(s,m) | \text{ for all } (s',m') \in A, \\ (s+s', m(m')) \in B\} \quad (22)$$

Lexical items are associated with closed semantic terms of the corresponding type, as for phrase structure logical grammar:

- **John** – j : N
- **Mary** – m : N
- **man** – *man* : CN
- **that** – λxx : CP/S
- **runs** – *run* : N\S
- **loves** – *love* : (N\S)/N
- **gives** – *give* : (N\S)/ (N • N)
- **thinks** – *think* : (N\S)/CP
- **that** – $\lambda x \lambda y \lambda z$ (AND ((y z), (x z))) : (CN\CN)/(S/N)

The semantics of derivation is contained in the proof net analysis. It is recovered by a deterministic semantic 'trip', which starts upwards at the output root of the proof net, visits each node twice, once traveling upwards and once traveling downwards, and finishes back down at its origin (de Groote and Retoré, 1996). The link of each $A\backslash B^\circ$ and B/A° mother-node is labeled with a distinct variable, then the semantics as a lambda term is generated by successive left-to-right production of its symbols during the trip, which is made according to the instructions shown in Figure 7.

Thus, the semantics extracted from Figure 5(a) is

$$(run\ j) \quad (23)$$

The semantics extracted from Figure 5(b) is

$$((think\ (\lambda xx\ (run\ j)))\ m) \quad (24)$$

which simplifies to

$$((think\ (run\ j))\ m) \quad (25)$$

The semantics extracted from Figure 5(c) is

$$((\lambda x \lambda y \lambda z(\text{AND}\ ((y\ z), (x\ z)))\ \lambda w((think \\ (\lambda xx\ ((love\ w)\ m)))\ j))\ man) \quad (26)$$

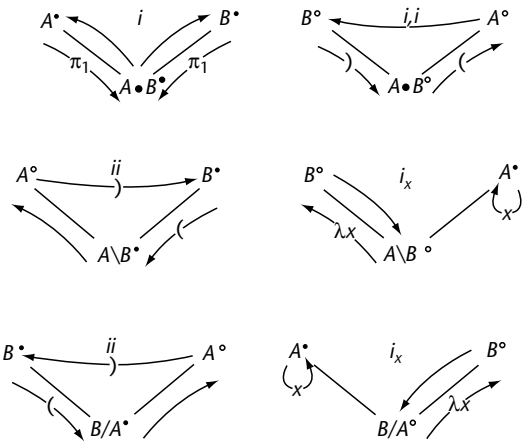


Figure 7. Instructions for a semantic 'trip' around a proof net that recovers the semantics of derivation. Travel starts upwards at the output root, follows the arrows at each link, picks up lexical semantics and bounces at input roots, and ends down backwards back at the output root.

which simplifies to

$$\lambda z(\text{AND}\ ((man\ z), ((think\ ((love\ z)\ m))\ j))) \quad (27)$$

CONCLUSION

We have looked at some approaches to formal grammar with reference to the 'action at a distance' of unbounded dependencies. To mention just one other phenomenon representing the challenge of mismatch between syntax and semantics, consider the quantification in a sentence like:

Everyone loves someone. (28)

The usual judgment is that the sentence is ambiguous, with one reading giving wide scope to the universal quantification (cf. 'everyone loves his mother') and another giving wide scope to the

existential quantification (cf. 'the queen is loved by everyone'). The puzzle is that if the architecture of grammar is compositional, that is, if semantics is determined by syntactic analysis, the sentence cannot be semantically ambiguous without having distinct syntactic analyses, and it is not obvious what syntactic alternation one should entertain.

Montague (1973) treated quantifier phrases as 'quantifying in' to sentences with variables in noun-phrase positions: different orders of quantifying in then yield different quantifier scopes. Cooper (1983) distinguishes from semantics proper a semantic 'quantifier store' from which quantifiers can be transferred at sentence nodes; again the order of transference effects alternative scopings. Treatments in contemporary phrase structure grammar and categorial grammar can usually be construed in one or other of these terms.

The range of phenomena that present themselves for syntactic and semantic study is truly vast, and this is hardly surprising: a huge part of our cognitive life, with all its subtlety and expanse, is linguistic, and we cannot expect a lot of easy answers. But formal syntax and formal semantics are good scientific programs, in that the nature of their failures, as well as the form of their successes, can reveal much about language and cognition.

References

Barwise J and Perry J (1983) *Situations and Attitudes*. Cambridge, MA: MIT Press.

Bresnan J (ed.) (1982) *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Carpenter B (1997) *Type-Logical Semantics*. Cambridge, MA: MIT Press.

Chomsky N (1957) *Syntactic Structures*. The Hague, Netherlands: Mouton.

Cooper R (1983) *Quantification and Syntactic Theory*. Dordrecht, Netherlands: Reidel.

Dowty DR, Wall RE and Peters S (1981) *Introduction to Montague Semantics*. Dordrecht, Netherlands: Reidel.

Gamut LTF (1991) *Logic, Language and Meaning*, vol. I and II. Chicago, IL: University of Chicago Press.

Gazdar G (1981) Unbounded dependencies and coordinate structure. *Linguistic Inquiry* **12**: 267–283.

Gazdar G, Klein E, Pullum G and Sag I (1985) *Generalized Phrase Structure Grammar*. Oxford, UK: Blackwell.

Girard J-Y (1987) Linear logic. *Theoretical Computer Science* **50**: 1–102.

de Groote P and Retoré C (1996) On the semantic readings of proof-nets. In: Kruijff G-J, Morrill G and Oehrle D (eds) *Proceedings of Formal Grammar*, pp. 57–70. Prague, Czech Rep.: FoLLI.

Lambek J (1958) The mathematics of sentence structure. *American Mathematical Monthly* **65**: 154–170. [Reprinted in: Buszkowski W, Marciszewski W and van Benthem J (eds) (1988) *Categorical Grammar*, pp. 153–172. Amsterdam: John Benjamins.]

Lambek J (1988) Categorial and categorical grammars. In: Oehrle RT, Bach E and Wheeler D (eds) *Categorial Grammars and Natural Language Structures*, pp. 297–317. Dordrecht, Netherlands: Reidel.

Montague R (1973) The proper treatment of quantification in ordinary English. In: Hintikka J, Moravcsik JME and Suppes P (eds) *Approaches to Natural Language*, pp. 221–242. Dordrecht, Netherlands: Reidel. [Reprinted in: Thomason RH (ed.) (1974) *Formal Philosophy: Selected Papers of Richard Montague*, pp. 247–270. New Haven, CT: Yale University Press.]

Montague R (1974) *Formal Philosophy*. New Haven, CT: Yale University Press.

Moortgat M (1997) Categorial type logics. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 93–177. Amsterdam, Netherlands: Elsevier.

Morrill G (1994) *Type Logical Grammar: Categorial Logic of Signs*. Dordrecht, Netherlands: Kluwer.

Pollard C and Sag I (1987) *Information-Based Syntax and Semantics*. Stanford, CA: CSLI.

Roorda D (1991) *Resource Logics: Proof-Theoretical Investigations*. PhD thesis, Universiteit van Amsterdam.

de Saussure F (1916) *Cours de Linguistique Générale*. Lausanne-Paris: Payot. [Translated in: Harris R (1983) *Course in General Linguistics*. London: Duckworth.]

Unified Theories of Cognition

Introductory article

Ronald S Chong, George Mason University, Fairfax, Virginia, USA
Robert E Wray, Soar Technology, Inc., Ann Arbor, Michigan, USA

CONTENTS

A collection of task-independent mechanisms that can be used to explain, through computational modeling, human cognitive behaviors such as problem solving, decision making, learning, memory, and interaction via perceptual and motor systems.

INTRODUCTION

In 1973, Allen Newell characterized the modus operandi of psychology as moving from phenomenon to phenomenon, producing vast numbers of studies and data either to formulate new theories or to support or refute existing theories. It appeared that the future of psychology would be a continuation of the same. Unlike physics, psychology had no practitioners pursuing the long-term goal of reconciling and synthesizing the various theories into a single unifying theory able to address all cognitive behavior.

Almost twenty years later, Newell, in his influential book *Unified Theories of Cognition*, called for psychology to begin the search for a unified theory of cognition (UTC). According to Newell, a UTC is 'a single set of mechanisms for all of cognitive behavior'. Examples of cognitive behaviors include problem solving, decision making, all forms of learning, skilled behavior, language, memory, perception, motor behavior, and perhaps even errors, motivation, emotions, imagining, and dreaming. Examples of behavior-independent mechanisms include memory systems, knowledge representation schemes, reasoning systems, perceptual and motor systems, and learning procedures.

Newell proposed several principles to guide the use and development of UTCs. The first is making a commitment to a UTC's mechanisms: new mechanisms must not be introduced just to address new phenomena. The reason for making a commitment is that it enforces parsimony and provides constraints on the modeling process, and hence leads to principled and informative models. Secondly, to accommodate UTC changes, Newell advocated 'listening to the architecture'. He believed that an architecture will indicate when it is wrong; i.e., when a phenomenon cannot plausibly be made to conform to the UTC's mechanisms. Only then should existing mechanisms be modified or new ones added. The final principle is that a UTC must be instantiated as a cognitive architecture. Newell proposed Soar as a candidate UTC. The remainder of this article will briefly describe Soar and two other computational architectures, ACT-R and EPIC.

SOAR

The Soar cognitive architecture, developed by Allen Newell, John Laird, and Paul Rosenbloom, has its roots in work that sought to understand human problem solving and decision making, or more generally, the nature of intelligent behavior. Soar is a production system which has been significantly extended to include mechanisms and structures believed to be functionally necessary for producing intelligent behavior. (*See* **Soar**; **Production Systems and Rule-based Inference**)

Soar incorporates a mechanism to automatically change the problem-solving context when a model reaches an 'impasse' – a point where progress towards accomplishing a goal cannot continue because of a lack of knowledge. In the new context, called a *subgoal*, processing is focused on resolving the impasse so that progress towards the initial goal can be resumed.

For example, if a Soar model of web browser use needs to display the bookmarks but does not know under which menu this command is located, the architecture would create a subgoal in which a 'search menus' method might be used to check each menu. When the menu item is found, the

impasse would be resolved and behavior can resume.

A by-product of resolving an impasse is the creation of new rules by 'chunking', Soar's sole learning mechanism. When processing in a subgoal succeeds in resolving an impasse, the chunking mechanism automatically compiles the processing in the subgoal into new productions. These learned rules could be applied the next time Soar is in the same or a similar situation: for example, it would remember how to get to the 'bookmarks' menu item. This single learning mechanism has been found to be sufficient for producing a variety of learning including concept learning, learning from instruction, and correcting faulty knowledge.

ACT-R

ACT-R, developed by John Anderson, grew out of an attempt to produce a computational theory of human memory. At the symbolic level, ACT-R is a production system. ACT-R also includes several subsymbolic mechanisms each of which addresses a specific form of cognitive adaptation. These mechanisms are justified by 'rational analysis' – the principle that cognition adapts to the structure of the environment. They modulate the availability of symbolic elements (declarative and procedural) during processing, as well as influencing the processing cycle time. (*See* **ACT**; **Symbolic versus Subsymbolic**)

One such subsymbolic mechanism assigns to each declarative memory element (DME) a number called an 'activation'. An activation learning mechanism varies the activation of each DME as a function of its frequency and recency of use. DMEs that are frequently or recently used will tend to have higher activations, and can be retrieved more quickly, than those less frequently or recently used. Any DME whose activation falls below a threshold ceases to be retrievable and is effectively forgotten.

For example, consider the situation of changing one's residence and having to learn a new phone number. Initially the new number may be difficult to recall, but as its use increases (by repeating the number to memorize it or giving it to friends), it becomes easier to recall. As the new number is being learned, the old number typically becomes harder to recall and may eventually be forgotten. ACT-R's activation learning mechanism can account for the gross aspects of this observation. As the use of the new-number DME increases, its initially low activation increases, causing it to be more readily recalled. At the same time, the high activation of the old-number DME decreases because of infrequent use, and in time may become inaccessible.

Another subsymbolic mechanism assigns an expected utility value to each production rule, and enables the system to learn which rules are best suited to accomplishing the goal that a model is currently trying to achieve.

EPIC

In *Unified Theories of Cognition*, Newell writes: '… one thing wrong with much theorizing about cognition is that it does not pay much attention to perception on the one side or motor behavior on the other. … The result is that the theory gives up the constraint on … cognition that these systems could provide. The loss is serious – it assures that theories will never … be able to tell the full story about any particular behavior.'

Unlike Soar and ACT-R, which are purely cognitive systems, EPIC, developed by David Kieras and David Meyer, addresses the full arc of behavior – from perception, to cognition, to motor behavior – with particular emphasis on the constraints provided by the perceptual and motor systems.

EPIC includes perceptual and motor mechanisms, each representing a synthesis of much empirical evidence. EPIC also contains a cognitive system, implemented using a production system. EPIC's cognitive system does not have any of the additional mechanisms found in Soar and ACT-R. This reflects the developer's commitment to parsimony: to avoid cognitive mechanisms or assumptions that may not be needed when perceptual and motor constraints are considered.

EPIC contains three perceptual systems: visual, auditory, and tactile. The visual system, the most highly developed, incorporates a representation of the eye's retinal regions. Consequently, the system accounts for the varying availability of an object's characteristics as a function of the object's eccentricity: for example, the color of an object is readily perceivable when the object is in the center of one's gaze, but is less so as it moves out to the periphery.

There are three motor systems: ocular, manual, and vocal. The two most developed processors, the ocular and manual motor processors, control the movement of a simulated eyeball and a pair of simulated hands, respectively.

The perceptual system conveys information about events and objects in a simulated task environment to the cognitive system. Using these inputs, the cognitive system then performs task-related

processing. When appropriate, it issues commands to the motor system to interact with a device in the simulated task environment.

Consider a very simple fictitious task where single text characters are sequentially presented at random locations on a screen and the user is to press a button when the '&' character appears. An EPIC model of this task would consist of productions that: first, send an ocular motor command to cause the eye to move to an object that just appeared; then, after the eye has moved to the object, think about the object to determine if it is an '&'; and finally, if it is an '&', send a manual motor command to cause the hand to press a button in the simulated task environment.

One of the advantages of incorporating perceptual and motor systems in a cognitive model is that they impose constraints such as the availability of perceptual information or the time it takes to perform a motor action. In the above example, the motor activity (eye movements and button presses) requires the same amount of time as observed in humans. Because of these and other constraints, more principled, informative, and predictive models can be developed.

EVALUATING UNIFIED THEORIES OF COGNITION

Newell outlines a method for evaluating UTCs in terms of cumulation: a UTC becomes more powerful as it explains more phenomena. A UTC would be 'better' than another if it could explain more phenomena while also enforcing more constraints.

This approach has suffered from two main problems. Firstly, most architectures provide little constraint on the creation of individual models, and thus offer little actual explanation. Furthermore, constraint is often imposed differently in different models, even within the same architecture. Some have argued that because all architectures can be made to exhibit any given phenomenon, the choice of one architecture over another must perhaps be made according to how easily the architecture explains some behavior, rather than empirical veracity alone.

A second problem concerning evaluation arises because UTCs are continually evolving. It is often unclear if the explanations and predictions of a model produced in an earlier version of an architecture remain valid in later versions. Anticipating this problem, Newell suggested that researchers build databases of models, so that previous models could always be validated in later architecture versions. However, this methodology has not been generally adopted. Thus, the cumulation of psychological evidence for UTC evaluation is not being undertaken in any rigorous way.

BENEFITS AND DANGERS OF UNIFIED THEORIES OF COGNITION

In *Unified Theories of Cognition*, Newell lists many benefits of UTCs: they provide unification in a wide-ranging field of research; they can increase the rate of cumulation of behavioral explanations; they can help constrain assumptions in models; they can help identify new areas for empirical investigations. UTCs have not yet provided all these benefits. However, another benefit of UTCs is their use in practical applications. Applications based on comprehensive cognitive models are used, for example in the evaluation of user interfaces, intelligent tutoring, and simulation of human actors for training.

UTCs also present dangers. Perhaps the most significant is the tendency to focus on cognition at the exclusion of other systems that are likely to impose strong constraints on cognition. Newell warns against the exclusion of perceptual and motor systems. The same warning might also apply to language, emotion, and even neurobiology. (*See* **Connectionism**)

CONCLUSION

Soar, ACT-R, and EPIC all satisfy the definition of a UTC in that they each make commitments to a small fixed set of mechanisms for producing a wide range of cognitive behavior. However, each of these architectures has traditionally had its own niche: Soar has mainly been used for modeling knowledge-level behavior and production learning via chunking; ACT-R for modeling the influence of memory on cognitive behavior via its subsymbolic mechanisms; and EPIC for modeling and explaining interactive behaviors via its perceptual and motor systems.

Note that these mechanisms are essentially orthogonal. One short-term approach to increasing both behavioral coverage and computational constraints is to combine established mechanisms from different architectures. An instance of this approach is the EPIC–Soar integrated architecture. This system merges Soar with the perceptual and motor mechanisms of EPIC and some of the subsymbolic mechanisms of ACT-R. Another instance of this approach is ACT-R/PM, an extension of ACT-R that incorporates some of the perceptual and motor systems of EPIC.

Further Reading

Anderson JR and Lebiere C (1998) *Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Chong RS and Laird JE (1997) Identifying dual-task executive process knowledge using EPIC-Soar. In: Shafto M and Langley P (eds) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 107–112. Hillsdale, NJ: Lawrence Erlbaum.

Hornof AJ and Kieras DE (1997) Cognitive modeling reveals menu search is both random and systematic. *Proceedings of ACM CHI 97: Conference on Human Factors in Computing Systems*, pp. 107–114. New York, NY: ACM.

Kieras D and Meyer DE (1997) An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. *Human–Computer Interaction* **12**: 391–438.

Meyer DE and Kieras DE (1997a) A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review* **104**: 3–65.

Meyer DE and Kieras DE (1997b) A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review* **104**: 749–791.

Newell A (1973) You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In: Chase WG (ed.) *Visual Information Processing*. New York, NY: Academic Press.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A (1992) Precis of unified theories of cognition. *Behavioral and Brain Sciences* **15**: 425–492.

Pew RW and Mavor AS (eds) (1998) *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

User Interface Design

Introductory article

Andrew Dillon, University of Texas at Austin, Texas, USA

CONTENTS

The design of computer interfaces that are usable and easily learned by humans is a non-trivial problem for software developers. As information technologies mediate many of the activities we now perform routinely, the process of human–computer interaction is of fundamental importance.

INTRODUCTION

Since much human–computer interaction (HCI) is cognitive in nature, involving perception, representation, problem-solving, navigation, query formulation and language processing, the theories and methods of cognitive science are directly relevant to it. Thus an applied cognitive science for software design has emerged.

Traditional cognitive science approaches to HCI and user interface design model the user as made up of three basic components: the psychomotor, perceptual, and cognitive subsystems. Recent treatments of HCI have extended this model to include the social system as an essential component of the user and have placed greater emphasis on group dynamics and social context in examining what users do with technology. Any full treatment of user psychology must embrace all these components, though cognitive issues dominate most research in HCI.

HUMAN–COMPUTER INTERACTION

The success of any computer application is dependent on it providing appropriate facilities for the task at hand in a manner that enables users to exploit them effectively. Whereas the provision of facilities is an issue of functionality, the user interface is the means by which the functionality can be used, and here we are directly concerned with usability.

For interaction to proceed, the human user must input a signal to the computer and perceive changes in the interface. Since most current interaction involves physical input and visual perception of output, any basic interactive device must incorporate an input device and a screen. However, to determine the appropriate input the user must have some representation of a goal or intention of an outcome to attain, necessitating the employment of memory, both short-term (to handle current status information) and long-term (to enable the planning and interpretation of the interactive sequence).

Information undergoes transformation at each stage of progression from a perceived stimulus (e.g. a visual change in the interface) to a comprehended cue (e.g. a recognized sign), leading to an active response (e.g. pushing a mouse button). All of this happens repeatedly and rapidly as the human interacts with the system. This process can be simply envisioned using a variant of Neisser's perceptual cycle model, whereby users engage in an ongoing cycle of information exchange involving exploration of a changing information environment (see Figure 1).

For example, an interface normally provides a response to user input which signals to the user that his or her input has caused an action. Through direct manipulation, users can select and move objects on screen, open and close windows with a mouse click, or jump from place to place within documents via hyperlinks. In each case, the user must initiate an action and the interface must communicate its change of state through appropriate feedback to the user. Where feedback is vague, too rapid or nonexistent, the user is likely to be confused. Ideally feedback communicates to the user that the system status has altered in the intended manner and that the user is closer to his or her goal as a result.

We can usefully understand HCI as a cognitive process by considering the human as possessing general knowledge structures (e.g. mental models or schemata) which organize the user's task-oriented information exploration and use. This

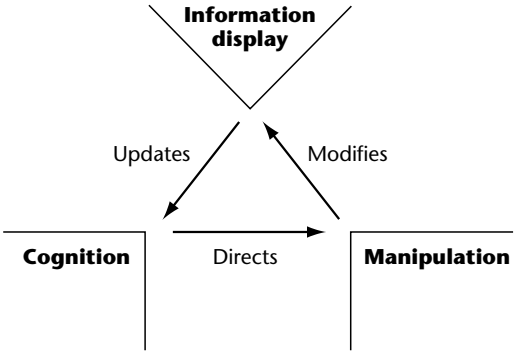


Figure 1. Human–computer interaction as an iterative process.

exploration, which may involve a variety of psychophysical actions such as key presses, scrolling, link selection, or command input, exposes the user to samples of the information space contained within the software, the perception of which must be interpreted and categorized, before it in turn can influence the subsequent actions of the user, and so on in an iterative fashion.

This is a very general portrayal of user psychology. Its purpose is to convey the active nature of cognition and show how humans select information on the basis of expectation and prior experience and how the selected information can itself modify the knowledge structures within the human. Thus, while the general model holds true, in the interface design context, each user is likely to have unique experiences and knowledge, which will influence the user's perception of, and interactions with, a computer. Software designers are now very aware of the effect that interface design has on users' experience of their products, and have sought guidance from cognitive scientists as to how best to design interfaces for usability and human acceptance.

The response from the cognitive science community has been threefold: the derivation of design guidelines to aid designers; the formulation of theoretical models to predict user response in specific instances; and the development of design methods and evaluation techniques to improve the process of user-centered design.

COGNITIVE DESIGN GUIDELINES

Since cognitive science has made important progress in understanding the mechanisms and processes underlying perception, memory, attention, categorization, decision-making, comprehension, and related processes and structures, it is reasonable to assume that such findings are relevant to practical design issues. For interfaces, we need to consider whether users will perceive actions and behaviors, how they will interpret them, what demands our designs place on their attention, and what knowledge they will need in order to interpret and respond in a manner acceptable to the software. There have been many attempts to bridge the gap between scientific findings and software design, and a full set of guidelines derived from studies of cognition would be very long. Much reinterpretation of cognitive science research has taken the form of general guidelines for designers to consider. Below are described several that have gained broad acceptance by most interface designers.

Screen Readability and Image Quality

To perform most information tasks with a computer, the user must be able to extract and process visual stimuli reliably and quickly. Early computer screens suffered from a variety of technological limitations that resulted in much slower reading of electronic text than of text on paper, or constrained the range of representations possible on screen. Current design guidance recommends the use of high-resolution screens with strong image polarity (preferably dark on light) to enhance human perception. Standard design advice is to produce all interfaces so that they can work in monotone and to add color sparingly to guide visual processing, attract attention, and aid chunking. As we interact more and more with screens, both large and small, the importance of readability becomes paramount. Poor visual ergonomics can prevent otherwise sophisticated software from being fully exploited.

Manipulation and Input Devices

While the typical computer of today is a desktop model with a keyboard and a mouse, there are many variants and alternatives, such as personal digital assistants and laptop computers, which are designed for mobility and use a stylus or trackpad as an input mechanism. Where once punched cards were the primary medium of interaction, the emergence of so-called 'direct manipulation' interfaces has advanced the exploitation of natural mappings between pointing and positioning found in the real world and the control of objects in a digital environment. 'Immersive' environments take this to the stage of creating a virtual world where users perform physical actions, much as in the real world, to produce responses from the software.

Studies of input devices have revealed that a mouse is fairly optimal for most standard point-and-click tasks, and that user reaching and target selection follow the basic principles of Fitt's law (a simple model of human psychomotor performance in rapid, aimed pointing tasks). However, where few on-screen selections are available, or where the use of a mouse would not work well (e.g. in a mobile application), touch screens with predefined tabbing zones have proven suitable, and there are claims of rapid input speeds with stylized input such as Graffiti (a simplified, single-stroke script designed for writing in a natural pen-based manner with the Palm series of devices). The qwerty keyboard remains dominant despite evidence that more efficient keyboard layouts could be designed, hinting at one important aspect of design that is beyond the control of science: the influence of precedent, habit, and market forces.

Supporting Accurate Mental Model Formation

Current work on HCI uses the concept of mental model extensively. The basic assumption is that users must try to understand what is happening with a system when they issue commands, and since much of the activity is hidden, they have to rely on inference. Depending on their knowledge of computing or the task being performed, users may infer correctly or incorrectly. Each user develops a personal image of the technology and how it works, though these images may be broadly similar across many users.

Research suggests that designers should regard models as 'mental scaffolding' upon which users hang their ideas about how the system works. The user's model is a personal, often idiosyncratic view of what the system does and how it does it. The designer should seek to make important aspects of the design transparent, coherent, and supportive. Another source of the user's mental model is prior experience, particularly with related products. If the user has worked with another system or an older version of the existing system then that experience is bound to influence the user's perception of the new technology since an existing schemata will be brought to bear on initial interactions. The user's experience of performing the tasks is also a contributing factor in the development of a personal model, and designers are advised to exploit the language, mappings, relationships among concepts, and procedures used by the target audience in creating an interface.

Use of Metaphors to Enhance Comprehension and Learning

Another generic aspect of human cognition that seems readily exploitable by designers is the reliance of human thinking on metaphors and analogies. Linked to the general tendency to model and learn by analogy, metaphors enable users to draw on their existing knowledge to act on a new domain.

There has been much discussion of the merits of the metaphor approach in dialogue design. It is argued that there are two dimensions relevant for understanding the information metaphors convey: scope and level of description. A metaphor's scope refers to the number of concepts to which it relates. A metaphor of broad scope in the domain of HCI is the 'desktop' metaphor common to many computing interfaces. Here, many of the concepts a user deals with when working on the system can be easily dealt with cognitively in terms of physical 'desktop' manipulations. The metaphor of the internet as an 'information superhighway' is also broad. The 'typewriter' metaphor that was often invoked for explaining word processors is far more limited in scope. It offers a basic orientation to using word processors (you can use them to create good-quality printouts) but is very limited beyond that as in many ways word processors do not behave like typewriters (e.g. typewriters do not save and store files, allow easy reformatting of text, or make instant copies of documents).

The metaphor's level of description refers to the type of knowledge it is intended to convey. This may be very high-level information, such as how to think about the task and its completion, or very low-level, such as how to think about particular command syntax in order to best remember it. Theorists in HCI distinguish four levels – task, semantic, lexical, and physical – which refer to different kinds of general question – 'can I do it?', 'what does this command do?', 'what does that term mean?', and 'what activities are needed to achieve that?', respectively.

Few, if any, metaphors convey information at all levels, but this does not prevent them being useful to users. Few users ever expect metaphors to offer full scope and all levels of description, so any metaphor employed should have its limitations and exceptions clearly pointed out. If the user cannot easily distinguish between the metaphorical aspects and functional relations that are and are not essential to its use, then the power of the metaphor will be greatly reduced. Of all the cognitive science

concepts used in HCI, metaphor has proved one of the most durable and accepted.

Learning by Doing

The most successful systems are those that enable a user to get something done as soon as possible. Users tend to be very resistant to reading any accompanying documentation and often want to get on with real tasks immediately rather than follow any training guides. Hence, error-free performance is not considered a real goal. Instead, cognitive scientists emphasize the importance of clear and informative feedback, and the ability to undo actions, to support the user through the learning process.

Having gained some knowledge by using one part of a system, users will expect to be able to apply this throughout the system. Particular attention should be paid to the consistent use of terms, colors, and highlighting techniques, and the positioning of task-related zones on the screen, so as to support generalization by the user. Consistency between systems can also be important to maintain – for example, between the old and the new versions. The benefits of a new system can easily be obscured if users feel that their existing knowledge is redundant and they must learn the new system from scratch.

Minimizing Attentional and Cognitive Load

Some theoretical insights into cognitive architecture emphasize the memory and attentional constraints of humans. These lessons have been learned by the HCI community who argue that interaction sequences should be designed to minimize the load on short-term memory (e.g. not asking a user to choose from an excessive number of menu items, or requiring the user to remember numbers or characters from one screen to another). Since recognition memory is superior to absolute recall, the use of menus is now the norm in design, as opposed to the command-line interfaces of the 1980s, which required users to memorize control arguments.

Another related contribution of cognitive science to user interface design has been in the area of task sequencing. User interface designers are encouraged to minimize the number of steps for which information must be retained by the user. Instead, designers are encouraged to provide all necessary information in the interface for the user to exploit as needed. The use of animation is recommended only where a process is being explained, although many designers deliberately exploit the natural human perceptual tendency to attend to movement by using animation to capture attention, particularly for advertisements in commercial internet sites.

Using Images and Icons

Screen 'real estate' is a limited commodity, so designers seek means of conveying concepts and actions through the medium of signs, images, and symbols. Another reason for iconic interfaces is their independence of language and their presumed ability to cross cultural boundaries.

Semiotic approaches to design have been invoked to help designers create appropriately comprehensible icons, but the results have been mixed. Current interfaces make extensive use of graphic capabilities and iconic representations but couple these with pop-up text labels that explain their meaning to users who find the representations difficult to decipher.

COGNITIVE THEORIES AND MODELS IN HCI

It is not yet possible to talk of a complete theory of human–computer interaction, given the many activities, processes, and tasks that computers support. However, to overcome the piecemeal approach that results from repeated empirical tests of evolving interface features, attempts have been made to produce stronger theoretical models to guide interface designers. This approach has worked best where it has been constrained to specific or localized interactive phenomena rather than the full range of user responses to information technology.

Interaction as Serial Information Processing

Cognitive scientists have derived many findings about human information processing, and this knowledge has been distilled in the area of HCI into a form of engineering model of the user that can be exploited by designers. Generally referred to as the 'model human processor', this cognitive model enables interface designers to predict the time a user will take to complete a task sequence given an analysis of the cognitive, perceptual, and psychomotor components that are applied at each step. For example, to determine how long it would take a user to complete the task sequence involving

saving a file to hard disk, consider the data in Table 1, derived from laboratory studies of humans.

To apply such a model, the designer would first list the basic steps a user must take with an interface. We can imagine a proposed design that requires the user to locate the mouse (*Th*), move the mouse to a menu (*Tp*), select the 'save' command ($Tp + Tk$), allow the system to respond with a prompt (*Tr*), input the filename ($Tk \times$ (number of letters)), and then hit a save button (*Tk*). The designer could quickly use the estimates from Table 1 to calculate how long a user would take to perform this sequence, and use these data to evaluate the proposed design.

The exact values of these estimates can be debated, but the principle of the model human processor is constant; i.e. decompose the task into its constituent actions and calculate the time involved in the serial processing of these actions. Multiple applications of this method have confirmed its value in estimating expert or error-free task completion times for repetitive, nondiscretionary tasks.

However, this model has its limitations. We cannot use it to estimate how long users will spend on tasks that are not highly practiced, or that require decision-making, planning, or learning. Similarly, where tasks involve parallel processing, it is easy to overestimate times by assuming simple serial processing of the task actions. However, as an applied model of cognition for a limited range of routine and well-practiced tasks, such a technique is clearly useful.

There have been several extensions of this approach, most notably to cover learning. Based on a production system analysis (describing the behavior a user must learn to complete a task in terms of a series of '*if–then*' rules of interactive sequences; e.g. 'if file is "new" then select menu option "new file" '), cognitive complexity theory enables calculation of the estimated time it would take a user to learn a new procedure. According to experimental findings, each new if–then rule production will take a typical user about 25 seconds to learn. Armed with such knowledge, designers could estimate, for example, the costs involved in changing procedures or violating consistency of interaction with new designs. This is obviously a gross estimate, but for many proceduralized tasks the data indicate the underlying regularity of human performance.

Table 1. Time estimates for completion of basic interactive tasks by a human operator

Label	*Action*	*Time estimate (seconds)*
Tk	Enter a keystroke	0.23
Th	Move hand to mouse	0.36
Tp	Point mouse	1.5
Tm	Retrieve from memory	1.2
Tr	Computer to respond	1.2

Sociocognitive Analyses of HCI: Activity Theory and Acceptance Models

An alternative application of cognitive theory has emerged as HCI researchers have become interested in user acceptance of computers and the exploitation of technology by groups of users. Such research draws less on the traditional base of laboratory findings within cognitive science and more on its social and anthropological traditions. These theories may be called 'sociocognitive' theories.

Activity theory aims to bring a closer reading of cultural forces to bear on our analyses of interaction. Users are seen as situated within a context that exerts strong forces on their actions. Furthermore, such users are dynamic, changing as their experience and application of technology changes. Taking an activity-theoretic approach to HCI, it is important to extend analyses of interface usability to cover the contexts in which the technology is used (or rejected).

Typical activity-theoretic approaches examine HCI in terms of the praxis, or situated context – e.g. a banking organization, a teaching scenario, or a medical process – in which the various levels of interaction can take place, from automatic individual operations to collective ventures or activities that define the group's purpose. The analysis and design of any technology needs to be grounded in such a broader perspective to ensure it is appropriate and usable by the intended user community. One can see activity theory as extending traditional cognitive approaches rather than replacing them.

Other socially-oriented approaches to HCI that consider cognition include the general class of acceptance theories that seek to predict whether a user, given a choice, will utilize a technology. Such models emphasize the perceived value that users place on the new technology, and measure the relationship between such ratings and subsequent behavior in context. For example, it is now known that if users perceive a new tool as having direct usefulness for them in their work, they will be more likely to choose it, and may tolerate a

certain difficulty of use for the sake of the power it affords them. Such perceptions by users seem to be formed very quickly, often within minutes of interacting for the first time. This means that early impressions due to aesthetics, implementation style, and related factors are particularly important.

Theoretical developments in HCI have not kept pace with developments in technology, partly because of the speed of technological change, but also because of the difficulty of translating cognitive science into rich theoretical models that predict human behavior in multiple contexts. Nevertheless, although some dismiss the theoretical approach as too limited for practical application, most HCI professionals are of the view that long-term progress is possible only with increased effort at deriving and applying cognitive science theories to the problems of user interface design.

DEVELOPING USER-CENTERED DESIGN METHODS

Where design guidelines and theoretical models of interaction fail to provide sufficient answers, designers resort to usability tests of their user interfaces. Cognitive scientists have contributed to this effort by providing the methodological and analytical perspective that informs evaluation practice.

Within HCI there are three basic evaluation methods: expert-based, model-based, and user-based. Expert-based methods assess an interface for compliance with known design principles and guidelines. Most of these guidelines are the products of research related to cognitive science. Model-based methods, including the methods mentioned above, are the application of theoretical models to specific design questions, and are almost always applied by those who have received some training in cognitive science. Both of these approaches are relatively fast and cost-effective, but they are limited in what they can predict.

User-based methods involve testing an interface with a sample of representative users in an appropriate context. There are as many variations on these methods as there are of methods of enquiry in cognitive science, ranging from controlled laboratory trials akin to psychology experiments, to field-based studies drawing on anthropological methods.

The pressures of design place demands on HCI professionals to produce fast answers, and cognitive scientists have worked on ways of improving the reliability and validity of test methods. There is a need for better expert-based evaluation methods to overcome the rather poor validity of most such methods. (Testers employing these methods tend to overestimate the number of problems users actually experience; that is, they label as 'problems' many aspects of interfaces that users subsequently perceive as acceptable.) Similarly, researchers have tried to package formal methods into tools that can be used effectively by non-specialists to predict usability. The aim of this approach is to develop software tools that designers could use to estimate learning effort or time to perform a task, without having to know the details of how such an estimate is derived. This would be analogous to the use engineers can make of the principles of physics. However, few such tools have yet made the transition from research laboratory to design practice.

CONCLUSION

User interface design is a complicated process that requires detailed analysis of human performance and preference. Furthermore, developments in technology require an understanding of emotional and 'trust' aspects of interaction that have yet to be studied in detail by cognitive scientists. As a form of applied cognitive science, interface design is a fruitful testing ground for a range of cognitive theories and methods.

Further developments in digital technologies will create new computing devices that will surround us at work, in leisure, and in our public and private lives. Cognitive scientists will be called on to help with the design and to study the impact of such technologies, and theory will meet practice in a manner that is likely to be important for our future wellbeing.

Further Reading

Card S, Moran T and Newell A (1983) *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum.

Dillon A (2002) *Designing Usable Electronic Text*, 2nd edn. London, UK: Taylor & Francis.

Helander M, Landauer T and Prabhu P (eds) (1997) *Handbook of Human–Computer Interaction*. Amsterdam, the Netherlands: Elsevier.

Landauer T (1995) *The Trouble With Computers*. Cambridge, MA: MIT Press.

Norman D (1986) *The Design of Everyday Things*. New York, NY: Basic Books.

Rosson M and Carroll J (2001) *Usability Engineering*. San Francisco, CA: Morgan Kaufmann.

Vision, Early

Intermediate article

Carlo Tomasi, Duke University, Durham, North Carolina, USA

CONTENTS

Early vision algorithms in humans and computers process images from the eye and from electronic video cameras respectively. They infer the shape, appearance, and motion of objects in the world. Conventionally, the lack of semantic interpretation distinguishes between 'early' and higher levels of vision.

INTRODUCTION

The crystalline lens in the human eye focuses the entering light onto the array of receptors in the retina, forming an image of the world. This retinal image encodes the color and brightness of the light that surfaces in the world reflect from light sources into the eye. The retinal image changes over time as objects or light sources move relative to the observer. The human early vision system analyzes these changing patterns of color and brightness to determine the position, shape, motion, and appearance of objects in the world. Conventionally, vision is said to be 'early' when it implies little or no semantic interpretation of the scene. Early vision therefore excludes higher cognitive aspects like object recognition or event interpretation.

Computer vision systems make similar inferences from the images produced by electronic video cameras. The basic computational elements and the overall architecture of human and computer early vision systems differ greatly. However, the abstract nature of the computations they both perform does not depend on the mechanisms of their implementation in man or machine.

The first step in vision is the formation of images, either in a camera or in the eye. Thereafter, images are analyzed and summarized in terms of edges, colors and textures, in order to provide a description of images that is more compact and depends to a lesser extent on changes of lighting or viewpoint. When changes of an image over time are considered, the motion of points in the field of view provides valuable information about the world. Image motion results from both observer motion and the movements and deformations of objects in the field of view. Its analysis allows distinguishing foreground from background, reconstructing the geometry of the three-dimensional world, and computing the motion of the observer within the environment. Additional sources of information about the world's geometry are: stereoscopic vision, which employs two cameras or eyes; the variations in the shading of visible surfaces; and the analysis of how edges meet one another in simple scenes. Although effortless to humans, early vision is the very difficult task of forming a stable representation of the world from the variable images seen by a moving observer.

IMAGE FORMATION

In the 'pinhole camera' model of perspective projection, the rays of light passing through a point O in space intersect an image plane, which records the intensity and color of each ray. The point O represents the optical center of the eye or camera, and the image plane stands for the retina or the camera sensor. Let a point P on a visible surface in the world have coordinates (X, Y, Z) in a Cartesian reference system with origin at O and Z-axis orthogonal to the image plane. If the focal distance (the distance from O to the image plane) is f, the image p of P has coordinates $x = f\,X/Z$ and $y = f\,Y/Z$. These coordinates are measured in a Cartesian image reference system, whose origin is the image point nearest to O.

The pinhole camera model captures the essential property of image formation: each point on the image plane corresponds to a line in the world, called the projection ray of that point. This model

does not account for secondary properties of real lenses, such as their imperfections, their limited ability to form sharp images, or the dependency of image brightness on lens size.

To model the finite number of sensing elements in real vision systems, the image coordinates x and y are discretized into integer pixel values within a finite range: $i = q(x/s)$ and $j = q(y/s)$ $(-n \leq i, j \leq n)$, where $q(a)$ is the integer nearest to a, s is the size of a sensing element, and n is a positive integer. The function q captures the discrete nature of images, and the bound n on image coordinates accounts for the finite field of view of a real vision system. This model does not account for possible overlap between adjacent sensing elements, or for sensors not on a square grid.

In the human eye, two types of photoreceptors at pixel (i, j) encode the color and brightness of incoming light. The 'rods' are highly sensitive to all wavelengths in the visible spectrum, but cannot distinguish colors. The 'cones' are less sensitive, but exist in three types, responsive to different, (but overlapping) bands of the visible spectrum. In electronic color cameras, red, green, and blue filters are superimposed on three brightness sensors at each pixel. Black-and-white cameras have one sensor per pixel.

CONVOLUTION WITH LINEAR FILTERS

Convolution is a ubiquitous operation in early vision. Intuitively, it amounts to additively blending the pixel values of small image neighborhoods to form a new pixel value. For instance, the blurring of a lens can be described as a weighted average of neighboring pixels in an ideal, sharp image. In this case, the blurred image is the convolution of the sharp image with an operator that averages pixel values together. At a somewhat higher level, an edge can be detected by comparing neighboring image pixels with an edge template. This is, again, a convolution, between the input image and the template.

Another example of convolution is image smoothing. To reduce the effects of noise in images, it is often useful to replace each pixel in an image with a weighted average of the intensity values that surround it. This averaging operation can be described by saying that the image is convolved with an operator, or 'kernel', that contains the weights to use for the average. As a simple example, if we compute the average of a pixel and its eight immediate neighbors, the kernel is a 3-by-3 matrix all of whose elements are equal to 1/9. These nine numbers are multiplied by the nine pixels in question, and the products are added together to yield the output average value. This procedure is repeated everywhere in the image.

Formally, let $L(i, j)$ be a function of pixel coordinates (i, j), and let $I(i, j)$ be the output, say, of a rod or cone. The convolution J of I with L is defined as

$$J(i,j) = \sum_a \sum_b I(i-a, j-b)\ L(a,b) \tag{1}$$

where the sums are performed over the domain of definition of the 'filter kernel' L. Thus, the output J at (i, j) is a linear combination of the values of the input I in some neighborhood of (i, j) and the values of the filter kernel L provide the combination coefficients. Similarly, single or triple summations appear in the definitions of convolutions of functions of one or three variables with filter kernels of one or three variables respectively.

Convolution with a bell-shaped kernel smoothes the input. A well-known example of a smoothing kernel is the isotropic Gaussian function

$$G(i,j) = k\mathrm{e}^{-(i^2+j^2)/2} \tag{2}$$

and its one- and three-dimensional equivalents. 'Isotropic' here means that the function G is rotationally symmetric, that is, it has the same shape in all directions. For instance, an out-of-focus lens blurs in approximately the same way in all directions, and its output is often approximated by the convolution of an isotropic Gaussian with the input image.

Convolution with the derivative of G with respect to one of its arguments approximates the partial derivative of the input with respect to the corresponding variable. This operation is very useful for edge detection, described next.

EDGE DETECTION

Edges are curves in the image across which image brightness or color changes abruptly. They are caused by shadow boundaries, contours of objects, or changes in surface orientation or reflectance. Standard algorithms for edge detection convolve the brightness $I(i, j)$ of the input image separately with the two partial derivatives G_i and G_j of the Gaussian kernel G to approximate the spatial gradient (I_x, I_y) of I. Some algorithms (Canny, 1986) then define edges to be ridges in the magnitude of the gradient. Others (Marr and Hildreth, 1980) convolve again I_x with G_i and I_y with G_j and add the resulting images together to obtain the Laplacian of image brightness. Edges are then zero crossings of the Laplacian.

Noise in the sensing elements produces random fluctuations of perceived brightness, which can cause spurious edges to be detected. A single threshold on the gradient magnitude cannot be used to suppress these edges: if the threshold is too high, good edges are removed as well, and if it is too low, spurious edges persist. Some algorithms (Canny, 1986) require edges to contain some elements above a high threshold, but are then extended to all edge elements that are connected to the former and are above a lower threshold. Other algorithms link edge elements into curves, and preserve only those curves that are longer than a given length.

OPTICAL FLOW AND MOTION PERCEPTION

As a point in the world moves relative to the observer, so does its image. When specified for every visible point, this image motion is called the motion field. The true motion field cannot be determined unambiguously from measurements in a very small image neighborhood. For example, if the neighborhood straddles a vertical edge and the edge moves in any direction, one only sees the horizontal component of motion. The edge may also be moving up or down, but this motion cannot be seen in the small aperture of this neighborhood. This inability to observe the motion field directly is called, somewhat awkwardly, the aperture problem.

The *optical flow* is defined to be the smallest image motion that is consistent with local image measurements. In the example above, the optical flow along the edge is $(u, 0)$. This example shows that the optical flow is not always the same as the motion field: the latter is the true, projected motion; the former is the apparent motion.

In computer vision, approximate motion fields are computed by combining values of the optical flow over several neighboring pixels, assumed to share the same motion field (Lucas and Kanade, 1981), or by imposing smoothness constraints on the field itself (Horn and Schunck, 1981). The human visual cortex computes the motion field by comparing the outputs of filters tuned to different orientations in space-time, that is, to different sets of spatial and temporal frequencies and directions of motion. Accuracies of 5 percent for velocities between 2 and 15 degrees per second have been reported for humans (McKee and Welch, 1985).

If only the observer is moving, the motion field at as few as five points is sufficient in principle to compute both the translation and the rotation of the observer, except for an overall scale factor, as well as the distance of the five points in the world from the observer (Thompson, 1959). However, this computation is very sensitive to inaccuracies in the field measurements, and reliable results require more motion field values. This computation is called 'structure from motion'. Accuracies as good as one degree of visual angle for observer translation have been reported both in psychophysics and in computer vision, where algorithms have been proposed also for reconstruction from more than two images.

COLOR AND TEXTURE PERCEPTION

Color

The three types of cone in the human eye are sensitive to three different bands in the visible spectrum of light, between 370 and 730 nm. Sensitivities peak at about 440, 540 and 560 nm for the three types, with much overlap in particular between the latter two bands. The distribution of the approximately 5 million cones in each eye varies from about 160 000 mm^{-2} in the fovea to about 20 000 mm^{-2} at the periphery, with the short-wavelength receptors being about 10 times sparser than those of the other types, consistently with the greater amount of blurring that the crystalline lens introduces at shorter wavelengths. The density in the fovea corresponds to a separation between cones of about half a minute of visual angle.

The spectrum of light impinging on a set of cones is the product of the spectrum of the light source and the reflectance of the visible surfaces. Yet if the color of the light source is changed, humans perceive surface colors as if the change in the light source were only about half as much as the actual change. This ability of the human visual system to compensate for changes in the color of the light source is called color constancy. To achieve color constancy, the visual system must estimate the color of the light source at least approximately. Some theories (Land, 1986) propose that the visual system selects a color for the light source that corrects the average color perceived by all the cones to gray. Other theories (D'Zmura and Iverson, 1993) propose that the visual system detects specular reflections in the image, and takes them to reflect the color of the light source unchanged.

Both psychophysical and physiological evidence suggest that colors are perceptually organized into pairs of 'opponent' colors in the visual cortex. This scheme for color encoding posits three mechanisms, each responding to a pair of sensations considered to be opposite to each other: light and dark, red and green, and blue and yellow.

Texture

An image region has visual texture when the distribution of its brightness values exhibits periodicity, either deterministic or stochastic. For instance, tiles on a floor are deterministically periodic, and leaves on a tree are stochastically periodic. When a texture on a complex surface is projected to an image, the deformations caused by foreshortening and by the changing normal to the surface modulate the periodic components of the texture.

Psychophysics and psychology indicate that humans can recognize materials and objects based on visual texture (texture classification), discriminate image regions with different textural properties (texture discrimination), and infer the shape and slant of a surface in the world from the modulation of its texture (shape from texture). Statistical representations (Caelli and Julesz, 1978) describe visual textures by the first- and second-order distributions of image brightness values in small regions of the image. Brightness histograms have been used in computer vision to capture the complete first-order empirical distribution. Summaries of the second-order distribution have included co-occurrence matrices, the fractal dimension of the spatial distribution of brightness values, and the conditional densities of an underlying Markov random field model.

Structural representations describe visual textures as the repetition of a basic pictorial element, or *texton*, according to some placement rule. This rule can take the form of the description of a grid on which textons are arranged, or it can be a formal grammar, either deterministic or stochastic, that generates the grid points. For instance, the texton of a tiled floor could be a single tile, and the grammar is the description of a regular grid of squares.

Current models of human texture analysis favor filter-based representations of texture (Bergen and Adelson, 1988). These are derived from the responses of a bank of linear filters tuned to different sizes and possibly orientations of the image intensity patterns. The integral of the magnitude of these responses over a small image neighborhood measures the energy contents of that neighborhood at the sizes and orientations that characterize the filters. A final stage of computation groups neighborhoods that have similar energy responses.

If the texture on a surface in the world is uniform, the different distances and slants of different surface patches relative to the observer produce gradual variations of the corresponding image texture, as described by the equations of perspective projection. Shape from texture solves these equations for either distance or slant, and infers the three-dimensional shape of the surface.

REGION ANALYSIS AND SEGMENTATION

The technique of segmentation partitions an image into regions such that different parts of the same region are similar to each other in some sense. For example, gray-level segmentation may require that the greatest difference between two pixels in the same region be less than a fixed threshold. A good segmentation would then have regions that are as large as possible given this constraint. Similarly, segmentation can be based on color or texture features. The results of segmentation differ from those of edge detection, mainly because edges are generally open curves while regions are bounded by closed contours.

In computer vision, images are often segmented in the hope that the resulting regions belong to different objects, or to different parts of an object. This is, however, rarely the case, because the varied colors of objects, shadows, shading, and variations in lighting produce large variations within objects , while at the same time similar colors in adjacent objects are not uncommon. Even so, describing an image as a collection of disjoint regions can offer advantages for later stages of processing, at least in terms of computational complexity.

Several segmentation methods are based on repeated splitting or merging, and sometimes on a combination of both. In a splitting method, for instance, the whole image may initially be considered as a single region, to be split, say, in half if the region as a whole is not homogeneous enough. The same procedure is then applied recursively to the resulting regions, until all regions are sufficiently homogeneous. Merging methods proceed in the other direction, starting with each pixel being considered as its own region, to be merged with neighbors as long as the resulting region is homogeneous.

'Stochastic relaxation' has also been used for segmentation. The class of images of interest is modeled as a Markov random field, which specifies the probability that a pixel has a certain value given the values of its neighbors. Given a particular image, relaxation iteratively adjusts pixel values to maximize the likelihood of the image having being drawn from the class in question (Geman and Geman, 1984). For segmentation, the underlying Markov random field assigns highest probability to noisy piecewise-constant images. As a result, relaxation turns the input image into a

piecewise-constant one, whose discontinuities are the segmentation boundaries.

BINOCULAR STEREOPSIS

'Binocular stereopsis' computes distances (called 'depths') to the visible surfaces by comparing the images of the world seen by two eyes, in humans, or by two cameras, in computer vision. This computation assumes knowledge of the relative position and orientation of the eyes or cameras. First, a stereoptic correspondence module finds pairs of points in the two images that are projections of the same point in the world. Then, the depth for each pair is computed by triangulation.

Stereoptic Correspondence

Two matching points in the two images of a stereoptic pair are likely to look similar to each other. However, this is not always the case. For instance, a point on a glossy surface may have the color of the surface when viewed from one eye, but the color of the light source in the other eye because of a specular reflection. Furthermore similar points do not necessarily match. On a blank wall, for instance, many points look the same. Thus, similarity of appearance, the main criterion for matching two points, is neither necessary nor sufficient for a match. Correspondence is also complicated by the fact that points visible in one image may not be visible in the other because of an intervening occluding object, so that not every image point need have a match.

Yet the human visual system can establish correspondences effortlessly and on the basis of minimal information. Random-dot stereograms like the one in Figure 1 serve to demonstrate this ability, and show that no prior recognition is necessary for stereoptic correspondences to be established (Julesz, 1960).

Knowledge of the relative position and orientation of the two eyes, or cameras, restricts matches for any given point in one image to be on a known line in the other. This is because the two optical centres and any one point P in the world define a plane, the so-called epipolar plane of P. This plane intersects the two image planes at two lines called the epipolar lines, which pass through the two images of P. Hence the 'epipolar constraint': the match for any point on one epipolar line must be on the corresponding epipolar line. The angle formed by the projection rays of two matching points is called the disparity.

One way to establish correspondences is to compare small image patches along corresponding epipolar lines in the two images. Sums of squared differences can quantify the comparison between a small image patch P_L centered at (i_L, j_L) in the left image I_L and a patch P_R centered at (i_R, j_R) in the right image I_R:

$$s = \sum_a \sum_b [I_L(i_L + a,\ j_L + b) - I_R(i_R + a, j_R + b)]^2 \quad (3)$$

where the summation indices range over the patches.

Even for image pairs as ambiguous as those in Figure 1, and even with imperfect image measurements, a small value of s indicates a likely match, as long as the patches being compared are not too small. When these local comparisons fail to determine matches unambiguously, more global criteria must be invoked. For instance, matches may be required to correspond to smooth, or at least continuous, surfaces. In computer vision, these more

Figure 1. The images in this random-dot stereogram are identical, except that a central square in the right image has been shifted slightly to the right, and the resulting gap filled with random dots. When viewed with eyes crossed so as to fuse the two images into one, a square floating in front of a plane should appear after a few seconds.

global requirements have been enforced by the use of various techniques, including stochastic relaxation, dynamic programming, and network flow optimization methods.

Triangulation

The depth of a point in the world is easily computed from a pair of corresponding points. Let α be the angle that the left projection ray of point P forms with the optical center of the left camera. Let δ be the disparity, and let the baseline, that is, the distance between the two optical centers, be b. Then, a simple geometric construction shows that the distance between the left optical centre and point P is

$$\rho = b(\cos\alpha + \sin\alpha\cot\delta) \tag{4}$$

SHAPE FROM SHADING

The perceived brightness of a surface varies with the orientation of the surface relative to illumination and viewing directions, among other factors. As a consequence, the shading on a uniformly colored surface conveys some information about the shape of the surface itself. The locations of concavities and convexities are examples of qualitative information that can be gathered in this way, but the shape of the surface can be reconstructed even quantitatively if its reflectance map is known.

The reflectance map of a surface expresses surface brightness as a function of surface orientation for a particular distribution of the light sources. If the depth of a surface is represented as a function $z(x,y)$ of the image coordinates x and y, the two partial derivatives p and q of z encode the orientation of the surface, since the surface is orthogonal to the vector $(-p, -q, 1)$ everywhere. Therefore, the reflectance map can be expressed as a function $R(p,q)$ that gives the brightness of a surface patch with normal proportional to $(-p, -q, 1)$ in viewer coordinates. If the image at (x, y) has brightness $I(x,y)$, one thus obtains the equation

$$R(p,\ q) = I(x,\ y) \tag{5}$$

The two functions $R(p,\ q)$ and $I(x,\ y)$ are known (the former by assumption, the latter by measurement), and the unknown depth $z(x,\ y)$ appears through its partial derivatives p and q. The equation above is therefore a partial differential equation, which can be solved for z by numerical means.

Because the reflectance map combines information about light and surface, it is often hard to determine *a priori*. However, it is conceivable that one could learn approximate reflectance maps for surfaces of various materials and, say, known position of the sun in the sky.

LINE LABELING IN POLYHEDRAL SCENE ANALYSIS

The remarkable human ability to understand line drawings inspired early computer vision researchers to separate the image interpretation task into two stages. In the first stage, edge detection transforms an image into a line drawing, which the second stage then interprets.

In a world of polyhedral objects with no surface markings, a line segment in the drawing can only represent the convex or concave edge between two visible faces, or an edge that occludes some surface at a greater depth. Two or more line segments meet at junctions corresponding to vertices in the world. If the number of lines meeting at a junction is bounded (typically by a maximum of three), junctions can be classified into a finite taxonomy depending on the number and angles of the meeting lines. For instance, two lines form a V junction. In Y junctions, three lines meet at angles that are all smaller than 180° in the image; and in W junctions one of the three angles exceeds 180°.

A line drawing can then be interpreted by assigning labels – convex, concave, or occluding – to all line segments, making sure that no impossible junctions result. For instance, a concave edge cannot meet a convex edge at a V junction; and three occluding edges cannot meet at a W or Y junction. Remarkably, these rules usually restrict the interpretation of a line drawing to one or very few possibilities, assuming there are no accidental alignments of features. Huffman (1971) developed an elegant algorithm for finding a possible labeling of a line drawing, and started a very active area of investigation.

However, attempts to extend these results to complex images failed for several fundamental reasons. First, objects in the world are not always polyhedral. Second, an edge detector also detects surface markings, shadow contours, and other curves, for which simple consistency rules cannot be given. Third, edges computed from images are usually broken, and junctions are hard to pinpoint. Because of these difficulties, this line of research has been abandoned.

SIZE AND POSITION INVARIANCE

As an object moves relative to the viewer, the size and position of its projection on the image

change. However, objects are recognized in spite of these changes, and also in spite of changes of illumination or viewing angle. The methods for three-dimensional reconstruction described in the previous sections form one basis for explaining this invariance of recognition performance under changing stimuli. According to this explanation, the visual system would compute invariant representations of the image, such as three-dimensional object models in world coordinates.

An alternative explanation seems to be more consistent with physiological evidence, and posits instead the existence of mechanisms that compensate for variations in size, position, and other factors (Ito *et al*., 1995). Almost half of the neurons in the anterior part of monkey inferotemporal cortex seem to respond to the same object in the field of view under large changes in object size and position, although different cells respond within different ranges of variation. On the other hand, other neurons in the same area respond only when the object is presented in a narrow range of sizes and positions. Invariance in the former type of neuron may be achieved by convergence of multiple cells of the latter type.

These two explanations of perceptual invariance lead to entirely different theories of how the world is represented in the visual system. The proposal based on compensation is more consistent with pictorial representations of objects, in which images are transformed and aligned to achieve constancy, than the miniature world of the reconstruction-based approach. But the evidence is still insufficient to allow us to conclude definitely either way.

CONCLUSION

Some of the problems of early vision are problems of image processing or statistical estimation. For instance, edge detection amounts to template matching in the presence of noise; and geometric reconstruction estimates three-dimensional quantities from noisy measurements of image motion. However, vision is inherently a process of inference, and early vision is no exception: several assumptions must be made in order to reconstruct aspects of the three-dimensional world from its two-dimensional projections on eye or camera, and finding assumptions that are adequate in most situations is still a challenge to computer vision researchers. Vision is a form of cognition, and the study of early vision may be one of the best approaches towards understanding intelligence.

References

Bergen JR and Adelson EH (1988) Early vision and texture perception. *Nature* **333**: 363–364.

Caelli T and Julesz B (1978) On perceptual analyzers underlying visual texture discrimination: Part I. *Biological Cybernetics* **28**(3): 167–175.

Canny J (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6): 679–698.

D'Zmura M and Iverson G (1993) Color constancy. I. Basic theory of two-stage linear recovery of spectral descriptions for lights and surfaces. *Journal of the Optical Society of America (A)* **10**: 2148–2165.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.

Horn BKP and Schunck BG (1981) Determining optical flow. *Artificial Intelligence* **17**: 185–203.

Huffman DA (1971) Impossible objects as nonsense sentences. *Machine Intelligence* **6**: 295–323.

Ito M, Tamura H, Fujita I and Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology* **73**: 218–226.

Julesz B (1960) Binocular depth perception of computer-generated patterns. *Bell System Technical Journal* **39**: 1125–1162.

Land EH (1986) Recent advances in retinex theory. *Vision Research* **26**: 7–22.

Lucas BD and Kanade T (1981) An iterative image registration technique with an application to stereo vision. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 674–679. San Mateo, CA: Morgan Kaufmann.

Marr D and Hildreth E (1980) Theory of edge detection. *Proceedings of the Royal Society of London (B)* **207**: 187–217.

McKee SP and Welch L (1985) Sequential recruitment in the discrimination of velocity. *Journal of the Optical Society of America (A)* **2**: 243–251.

Thompson EH (1959) A rational algebraic formulation of the problem of relative orientation. *Photogrammetric Record* **3**(14): 152–159.

Further Reading

Gibson JJ (1950) *The Perception of the Visual World*. Boston, MA: Houghton Mifflin.

Hering E (1878/1920) *Handbuch der gesammter Augenheilkunde*, part 1, chap. XII. Berlin: Springer. [Originally published as: *Grundzüge der Lehre vom Lichtsinn*.]

Horn BKP (1986) *Robot Vision*. Cambridge, MA: MIT Press.

Kanatani K (1993) *Geometric Computation for Machine Vision*. Oxford: Clarendon.

Longuet-Higgins HC (1981) A computer algorithm for reconstructing a scene from two projections. *Nature* **293**: 133–135.

Marr D and Poggio T (1976) Cooperative computation of stereo disparity. *Science* **194**: 283–287.

Pollard SB, Mayhew JEW and Frisby JP (1985) PMF: a stereo correspondence algorithm using a disparity gradient constraint. *Perception* **14**: 449–470.

Svaetichin G (1956) Spectral response curves from single cones. *Acta Physiologica Scandinavica* **134**: 17–46.

Zucker SW (1976) Toward a model of texture. *Computer Graphics and Image Processing* **5**: 190–202.

Vision, High-level

Intermediate article

Florin Cutzu, Indiana University, Bloomington, Indiana, USA

CONTENTS

The main functions of high-level vision are the recognition and categorization of visual objects and navigation guidance.

INTRODUCTION

The main functions of high-level vision are the recognition and categorization of visual objects and the guidance of navigation in the environment. It must be noted that these tasks are, however, not exclusively visual in nature. Object recognition can occasionally be based on olfactory (the recognition of a skunk) or auditory (the recognition of a seagull) or haptic (the recognition of tree bark) information. More significantly, in navigation and object manipulation navigation, the information conveyed by proprioceptors is essential.

This review discusses in detail object recognition and categorization, and only briefly vision-based navigation and object manipulation.

OBJECT RECOGNITION

Object recognition refers to the process whereby a previously seen object is identified in an input image, possibly under novel viewing conditions. Categorization is the assignment of a (possibly previously unseen) input object to a known object class. Of the two, recognition is better understood computationally. Recognition is in principle achieved by comparing input image data to a number of candidate internal object models, and selecting the model best matching the image data. There exist multiple sources of information in an image, and multiple pathways to recognition: an object can be recognized on the basis of its shape, texture, color, size, motion, or scene context. Only shape-based recognition is considered here.

Recognition proper must be preceded by two preparatory stages: visual selection and image segmentation.

Visual selection is the process whereby an object of interest is detected and localized in the input image. This represents a computationally difficult problem. One difficulty is that the location, scale and orientation of the sought object are unknown, resulting in a large search space. Another difficulty is that the target object may be only partially visible, being occluded by other objects in the scene.

The problem of selection has been studied in computer vision especially in connection with the detection of faces and people in images. In human vision research, visual selection refers to locating the *salient* regions of the input image, not to the detection of a target object. A salient image region is likely to contain a biologically relevant visual stimulus and consequently attracts visual attention. The salience of an image location is a measure of how different it is from the rest of the image in terms of features such as color, orientation, spatial frequency, etc. (*See* **Visual Attention**)

Segmentation is the grouping of image pixels into regions corresponding to the different objects in the scene. Although performed effortlessly by humans, computationally it is a hard problem. Segmentation techniques can be either contour-based, which work by linking image edges into smooth closed contours, or region-based, which seek an optimal partitioning of the image into regions of homogeneous texture or color. Segmentation is too vast a subject to be even summarized here; the interested reader is referred to Shi and Malik (2000), and Malik *et al*. (2001).

Following detection and segmentation, the problem is reduced to recognizing an isolated, non-occluded object presented under arbitrary viewpoint and illumination.

The computational and psychophysical studies of object recognition have mainly focused on the problem of recognizing rigid objects under variable viewpoint. Although humans effortlessly recognize objects from different orientations, the underlying computational problem is quite difficult. The image

change resulting from rotation in 3-D followed by projection onto the image plane can be very substantial, rendering direct approaches based on storing 'enough' views impractical.

The following sections review the most important computational methods for recognizing rigid objects under variable viewpoint.

Object Recognition by Part Decomposition

Many objects appear to be arrangements of simple shapes such as cylinders, cones, or blocks. These volumetric primitives have been modeled traditionally in computer vision and graphics by *generalized cylinders*. A generalized cylinder (Binford, 1987) is a 3-D volume generated by a planar shape (the cross-section) swept along a curve (the axis) according to a sweeping rule specifying how the cross-section changes during the sweep. By changing the shape of the cross-section, the curve, and the sweeping rule, a wide variety of shapes can be generated. Influential early 3-D shape representation schemes based on generalized cylinders are Binford (1971), Marr and Nishihara (1978), and Brooks (1981).

An important class of parametrized generalized cylinders are the deformable superquadric ellipsoids (Pentland, 1986). By adjusting some 20 shape control parameters of a superquadric, a large repertoire of volumetric primitives can be generated. The use of deformable superquadrics in object part modeling has been limited to fitting depth images (Dickinson *et al*., 1997b); attempts to fit superquadrics to intensity images have not been very successful.

Biederman (Biederman, 1985, 1987) developed the generalized cylinder shape representation idea into a theory of human visual object recognition called Recognition by Components (RBC). A basic assumption of RBC is that objects are arrangements of simplified generalized cylinders called geons. The shape differences between geons are not metrical, but qualitative. Geon axes can be straight or curved (it does not matter how curved), cross-sections can be either polygonal or smooth, and the cross-section sweep function can be either constant or increasing/decreasing. The resulting geon repertoire is small (a few tens), and includes shapes such as cylinders, cones, wedges, and blocks. The level of shape detail in the RBC scheme is such that most everyday objects are decomposed into less than ten geons.

The RBC scheme provides qualitative descriptors of the spatial relations among object geons (such as on-top-of, perpendicular-to, etc.) thus providing a complete structural object description. The resulting encoding of shape is qualitative: under the RBC scheme a dog is indistinguishable from, say, a horse.

To be used in recognition RBC must provide a method for segmenting an object into component geons and for recognizing the segmented geons in an input image. The RBC scheme assumes that an idealized edge detection stage provides a line drawing of the object. The object is segmented starting at the deep concavities of its outline, and using the constraint that there exists one pair of such concavities per geon, the silhouettes of the component geons are extracted. The qualitative shape of a geon can be inferred from its silhouette almost independently of viewpoint by using silhouette properties that directly reflect 3-D shape and are quasi-viewpoint invariant. For example, a straight image contour indicates a straight line in space (and thus a straight geon), two parallel image curves indicate parallel space lines (and thus a constant-section geon), (skew) symmetry in the image indicates a symmetric cross-section, etc. Such image properties are termed *nonaccidental* (Witkin and Tenenbaum, 1983), and are reliable indicators of 'true' 3-D shape. While a straight image segment can in principle be the projection of a curved line – say, an arc of circle – it is very unlikely, requiring a viewing direction perfectly aligned with the plane of the curve. Any perturbation of this special viewpoint will replace the straight line segment with an arc of ellipse.

The RBC scheme has two basic advantages over image-based recognition methods. First, it can be used for analyzing unfamiliar objects by decomposing them into familiar parts. Second, it can handle certain deformable objects – those composed of articulated rigid geons.

RBC also has serious limitations (see Dickinson *et al*. (1997a) for an in-depth discussion). First, it is very difficult to recover geons from real-world intensity images – mainly because available edge detectors cannot produce good quality line drawings. Implementations of RBC have sidestepped this difficulty by assuming the line drawings to be given; an example is the neural network model described in Hummel and Biederman (1992). A second problem is that many objects (a shoe for example) are not composed of geons. Third, RBC provides only a qualitative description of shape and thus is not suitable for object identification; consequently, RBC can be considered a theory of basic level shape classification.

However, RBC was intended to be not so much a computer vision recognition system as a model of human object recognition. In this connection, one of the most important – and disputed – predictions of RBC is that object recognition performance should be nearly viewpoint invariant when diagnostic component geons are not occluded or severely foreshortened. Biederman and his collaborators have published several psychophysical studies supporting this prediction; a representative study is Biederman and Gerhardstein (1993). In contrast, view-based models of object representation hold that recognition should be strongly viewpoint dependent. Many psychophysical studies indicate that object recognition is viewpoint dependent; two relevant papers are Tarr and Bülthoff (1995), and Tarr *et al*. (1997).

Object Recognition Using the Alignment Method

The alignment method, like RBC, uses 3-D models to represent objects, but, unlike RBC, it represents shapes quantitatively and holistically rather than qualitatively and by parts.

Under the alignment scheme an object is recognized if its image(s) can be reproduced by rotating some internal, pre-stored, 3-D object model. Let $\mathcal{M} = \{M_1, \ldots, M_n\}$ be the set of internal 3-D object models. The models are subjected to certain transformations, typically 3-D rotation, scaling, translation, and imaging transformations (perspective or orthographic projection). The recognition of an input image amounts to finding the element of $\mathcal{M}$ that under an allowable transformation best fits the input image. Given an input image I, the recognized 3-D model $M_{\hat{k}}$ best reproduces the input image when subjected to the optimal transformation $T(M_{\hat{k}})$:

$$\hat{k} = \arg\min_{k} \mathcal{D}[I - T(M_k)M_k] \qquad (1)$$

where $\mathcal{D}$ is some image distortion (distance) measure. The minimization in equation 1 is performed over the discrete set of stored models and for each model over the set of allowable transforms T. In general, T consists of a rotation R, a translation t, scaling s, and the 3D-2D imaging projection P. When homogeneous coordinates are employed, the compound transform is expressed as a product of matrices $T = PtRs$. The optimal aligning transformation depends both on the input image I and on the optimal model $M(\hat{k})$, leading to a nontrivial minimization problem.

The alignment scheme provides a method for performing this optimization. A key assumption of alignment is that several localized features (called anchor points) can be identified both in the input image and the models under consideration, and that it is known which point in the image corresponds to which point in the model. The image and the models are said to be *in correspondence*. Establishing correspondence is in itself a difficult problem, typically requiring an extensive search even when geometrical constraints are used to eliminate candidate feature pairings.

Good feature points must be reliably identifiable over a large range of viewpoints, that is, to be quasi-viewpoint invariant. Examples of such features are maxima of object surface curvature (corners), inflection points along contours, surface markings, etc.

In the alignment scheme the object models are, essentially, object views enhanced with depth values. Due to self-occlusion, a single 3-D object requires several such view-specific models. Model information includes the 3-D coordinates of the anchor points and other object features, such as internal contours and silhouette points, visible at the orientation the object is modeled from. The anchor points are used for aligning the model to the image, and the other features are used in evaluating the degree of agreement between the model and the input image.

Once correspondence between the image and model anchor points is established, the optimal alignment transformation is easily determined for *all* internal models, both correct and incorrect.

The transformation $T(M_k)$ necessary for aligning 3-D model M_k to the input image can be determined from the coordinates of the corresponding features in the image and model. Under orthographic projection, three pairs of corresponding feature points are sufficient, the only constraint being that the points not be co-linear. $T(M_k)$ can be determined given the image-plane coordinates (u_i, v_i) of the three feature-points and the 3-D coordinates (x_i, y_i, z_i), $i = 1, 2, 3$ of the same feature-points in a coordinate system attached to the internal model. Determining the transform T amounts to determining the rotation matrix R (three degrees of freedom), the translation t (two degrees of freedom) and the scaling factor s (one degree of freedom).

Following alignment, the minimization in equation 1 is reduced to searching over the discrete set of object models only, and selecting the model that best matches the input image under the measure $\mathcal{D}$.

Typically, matching is performed by comparing the locations of contour corners and inflection points in the image and the model (Huttenlocher and Ullman, 1990).

Determining feature correspondence

The preceding discussion assumes that correspondence has been established between the image and object features; in practice, computing correspondence is the most difficult stage of alignment. A simple approach to correspondence computation is to try all possible pairings of image and model features, selecting the one resulting in the best post-alignment match. This strategy can be improved by restricting the pairings to like features, for example corners to corners, blob centers to blob centers, etc. For polyhedral objects, Huttenlocher and Ullman also used connectivity information to narrow down matches: a pair of point features (polyhedron vertices) joined by an image contour (polyhedron edge) is matched to a pair of point features in the model only if they are joined by a straight model contour.

Dealing with self-occlusion

Not all surface points of a solid 3-D object are visible from all viewpoints, because the object is not transparent – it self-occludes. This poses a problem for alignment, as there exist object views in which the anchor points are not visible. Therefore, for such a view the alignment transform cannot be computed. The solution is to store, for a given 3-D object, *several* models and feature triplets. A single object model handles the range of viewing directions from which its anchor points are visible. Several such models are necessary to cover the entire viewing sphere. Each model represents the object from a certain range of viewpoints, and from any viewpoint a model and a feature triplet are available for alignment. It must be noted that establishing correspondence and tackling self-occlusion are difficult, basic problems that any 3-D object recognition technique must address.

View-based Recognition Schemes

Recognition by linear combination of views

In contrast to alignment, which uses 3-D object models, the linear combination of views approach uses only 2-D (image-plane) information.

Consider a rigid cloud of N 3-D points, rotating in space, and projected (imaged) orthographically onto the XY plane. A view v of the point cloud is specified by the XY coordinates of the image points: (x_{vi}, y_{vi}), $i = 1,\ldots,N$. It can easily be shown that any view v of the cloud can written as a linear combination of three, arbitrarily selected, basis object views α, β, γ:

$$x_{vi} = a_{v1}x_{\alpha i} + a_{v2}x_{\beta i} + a_{v3}x_{\gamma i}$$

$$y_{vi} = b_{v1}y_{\alpha i} + b_{v2}y_{\beta i} + b_{v3}y_{\gamma i}$$

where the coefficients $a_{v1,2,3}$, $b_{v1,2,3}$, can be interpreted as the 'coordinates' of the view v in the space spanned by the three basis views. The only constraint is that three 'basis' views must not have been generated by a rotation around the same axis. The fact that the views are generated by rotation of a *rigid* point cloud followed by *orthographic* projection imposes certain constraints on the values of the coefficients a, b, c. Note that correspondence is given and that there is no self-occlusion, the object being a cloud of discrete points.

This scheme can be directly applied to rotating polyhedral objects, for which the features are image contours corresponding to the edges and corners of the polyhedron. The linear combination of views technique assumes that the same feature points are visible in all the views. However, this is generally not possible due to self-occlusion. The solution is to store a number of view triplets, linear combinations thereof representing the object from all possible viewing directions.

Smooth objects are harder to handle by the alignment or the linear combination of views schemes than polyhedral objects: for smooth objects the linear combination method requires five rather than three basis views.

In conclusion, the set of all possible views of any object (whether smooth or not) can be expressed as the combination of a small number of its views.

The linear combination principle can be used for deciding whether a new view is a possible view of a stored model. Consider an input view for which a small number of features can be put in correspondence to those of the basis views. First, these corresponding features are used for recovering the coefficients of the input view. Next, these coefficients are used to generate a correct object view by combining the basis views. The recognition decision is based on comparing the generated view to the input image.

Recognition using nonlinear image combinations

An object view is a feature vector, that is, a point in a multidimensional feature space. According to the linear combination of views principle, the views of

an object (assuming no feature occlusion) occupy a low-dimensional linear subspace of the feature space.

The radial basis function (RBF) interpolation/approximation methods (Poggio, 1990; Poggio and Girosi, 1990; Poggio and Edelman, 1990) use nonlinear (rather than linear) interpolation between feature vectors of the basis views. The RBF method also requires that feature correspondence between views be established.

The idea of view recognition by RBF-based interpolation is simple. Views are N-dimensional vectors $\mathbf{x}$. The views of the target object are to be recognized by performing a certain computation on their feature vectors resulting in a characteristic constant value, c; when a view of a different object is subjected to the same computation the result c' should substantially differ from c. An (approximately) constant function $f(\mathbf{x}) = c$ must thus be defined on the set of views of the target object. M training object views $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are available, and therefore for these views it is known that $f(\mathbf{x}_k) = c$. The value of the function f for an arbitrary view $\mathbf{x}$ is calculated by taking a weighted sum of M radial basis functions (typically N-dimensional Gaussians) centered on the training views $\mathbf{x}_k$. The resulting function f, being a superposition of Gaussians, is smooth. If $\mathbf{x}$ is a view of the target object, the resulting value for $f(\mathbf{x})$ will not be very different from c, essentially because $\mathbf{x}$ cannot be far from the training views $\mathbf{x}_1,\ldots,\mathbf{x}_M$. On the contrary, if $\mathbf{x}$ is not a target view, $f(\mathbf{x})$ will differ from c substantially.

The disadvantage of this method is that it ignores the linearity of the viewspace of rigid objects under orthographic projection; the advantage is that it can handle nonrigidity, image distortions and perspective projection.

Object Recognition Using Geometric Invariants

A geometric invariant is a function defined on images that (1) does not change with changes in the imaged object's orientation, or, more generally, with changes in viewing conditions (2) changes only if object shape changes. Finding such an invariant would solve the problem of recognizing objects under variable viewing conditions. However, it is easy to see that geometric invariants do not exist in the general case. Consider two objects that look identical from a certain viewpoint and are otherwise different: condition (1) is satisfied but (2) is not.

In addition, a genuine geometrical invariant ought to be calculable from a single object image. However, in Burns *et al*. (1992) it is shown that it is not possible to define an invariant based on a single view of a cloud of 3-D points; two such views are necessary.

The definition above can be relaxed to allow for functions that are approximately invariant to viewpoint change and differentiate only among objects from a certain, restricted, category. In practice, a number of different invariants are used together to form a vector that can be used to index a database of object models, or can be further analyzed by statistical pattern classification techniques (Duda *et al*., 2001).

The geometric invariant approach has been most successful when applied to the recognition of planar or nearly planar objects such as airplanes. For planar objects it is straightforward to derive geometrical invariants from a single object view. Consider a planar object of polygonal shape imaged by *orthographic projection* onto the image plane. Orthographic projection is an adequate image formation model only when object size is small relative to its distance from the camera. When the object polygon rotates in 3-D, the image polygon undergoes a so-called *affine transformation*. Under an affine transformation parallel lines remain parallel, length ratios of parallel segments and ratios of areas of triangles are preserved. The values of these ratios are shared by all the views of the planar object and by the object itself. These simple properties provide the basis for more sophisticated invariants used for viewpoint invariant recognition of planar shapes under orthographic projection.

If the polygon object is imaged by a pin-hole camera, that is, by *perspective projection*, the transformation in the image plane is no longer affine, but projective. Projective transformations are more general than affinities and correctly model camera image formation. Projective transformations do not preserve parallelism or the other affine invariants described above. Projective transformations preserve only the *cross-ratio*, which is defined as a ratio of ratios of lengths or areas. Affine transformations are a special case of projective transformations, so cross-ratio is also an affine invariant.

The invariants discussed above are appropriate for polygonal lines and require the establishment of point correspondence across the different images. Affine invariants based on image moments are global quantities based on the entire image and thus correspondence is not required; in addition, image moments are defined for gray-level images, not just for contours. The disadvantage of moment invariants is that being global image properties, they are affected by image occlusion.

In conclusion, planar objects can be reliably recognized by geometric invariants. An in-depth treatment of recognition of flat shapes by invariants is given in Reiss (1993).

The problem of finding invariants of gray-level images of 3-D objects is much harder. In the last few years a number of single view invariants for 3-D smooth surfaces have been developed; these advances are beyond the scope of this review and are presented in Rothwell (1995); Mundy *et al*. (1994).

VISUAL CATEGORIZATION

Categorization is arguably a more difficult task than identification, because in addition to changing viewing conditions it has to handle certain shape variations. Note that recognition does not imply categorization: an unusual object can be recognized without being classifiable.

Much of the current understanding of human (visual) categorization is due to the theoretical and experimental work of Eleanor Rosch (Rosch *et al*., 1976; Rosch, 1978).

Natural object categories have a hierarchical structure (for example: Siamese $\subset$ cats $\subset$ felines $\subset$ quadrupeds $\subset$ mammals $\subset$ animals).

Class members have different degrees of typicality. Each category has a prototype, which is its most typical (average) member. Visual categorization is faster for more typical objects.

An object is recognized (named) spontaneously by observes at an intermediate level of its class hierarchy, called by Rosch the *basic level*. For example, a Siamese cat will be spontaneously recognized as *cat*, not as *feline* or *Siamese*. Categories above the basic level are called superordinate and categories below the basic level are called subordinate.

Further research (Jolicoeur *et al*., 1984) has shown that atypical objects are first classified at a subordinate level, and the term *entry level* has been proposed for designating the class to which an object is spontaneously assigned.

Interestingly, just like the different members of a class, the different views of a 3-D object are consistently judged by observers as being more or less typical; the most typical view is called the canonical view (Palmer *et al*., 1981). Categorization is fastest when the object is presented in the canonical orientation.

Categorization is based on visual similarity at the basic level and below; above the basic level categorization is not based on appearance but on abstract criteria. Understanding visual categories thus requires understanding *visual similarity*. Biederman's RBC scheme can be viewed as a theory of basic-level object categorization, as objects from the same basic level class share the same geons and geon spatial relations. RBC does not, however, offer a computational account of shape similarity. Not all objects in a basic level class are equally similar to each other; there are metric differences that RBC fails to capture. On the other hand, RBC does not define the similarity of objects that have different geon structural descriptions.

A model of visual similarity must take metrical properties of shape into account. One possibility (Shepard, 1980) is to represent a shape as a point in a multidimensional metric feature space and shape dissimilarity as distance in this space. Perceived similarity as well as similarity of long-term memory traces for a class of complex 3-D objects has been found to be highly correlated to feature-space distance (Cutzu and Edelman, 1998).

An in-depth discussion of visual similarity and its role in object representation is given in Edelman (1999).

USING VISION FOR OBJECT MANIPULATION AND NAVIGATION

Besides recognition, vision is used for guiding the navigation of an organism in the environment and for guiding object-grasping and manipulation.

Visual Control of Object Manipulation

An active research area, with practical applications in robotics, is the visual control of grasping. One problem is the determination of the best, most stable, grasp points for an object that must be picked up and handled. The simple variant of this problem uses an opposing, two-fingered grasp model, while the more complex variants use multi-fingered grasps. A related problem is the visual guidance of the robot gripper to the grasp points. This represents a computationally difficult visual task, requiring the tracking of the fingers and the object, and detection of object contacts and collisions in three dimensions.

Generally, vision alone is insufficient for hand–eye coordination and object manipulation control; feedback from proprioceptors, haptic and force sensors is essential. Purely vision-based solutions must be considered given that most robotic manipulators generally lack non-visual senses.

Visual Control of Navigation

Vision-based navigation (Borenstein *et al*., 1996) involves *goal specification, localization*, and *path*

planning and *following*. Localization is the process whereby the navigator's position is identified on a map, or more generally, in a stored model of the navigation environment. Path planning is the determination of the sequence of movements leading to the final destination. Path following is the actual execution of the planned movements, and involves issues such as dealing with moving obstacles, and the correction of errors.

Location determination requires extracting (point or contour) features from both model and input image, establishing correspondence between them, and the computation of the position of the observer with respect to environment on this basis. Necessary 3-D information is usually supplied by stereoscopic vision. The problem is difficult in unstructured outdoor environments, mainly due to the unreliability of edge-based feature extraction and ambiguities in feature correspondence; the problem is simplified in structured indoor environments by the presence of linear structures and unambiguous landmarks.

Path following requires the determination of both ego-motion and motion of mobile objects in the scene. Consider a pair of video cameras (a stereo head) mounted on a mobile robot moving in a rigid environment. From an image pair taken at an initial location, a stereo algorithm yields the depth (relative to the camera coordinate system) of the relevant image features (for example, rectilinear image contours can be corners for office scenes). After a robot move, another image pair is captured and feature depth values are computed at the new vantage point. The ego-motion of the robot is calculated as the rigid transformation (rotation and translation) aligning the corresponding features in the two consecutive depth images. In the case where some objects in the scene are also moving, their features will not be matched across views by the algorithm calculating ego-motion, which calculates the dominant rigid motion field in the scene. To determine the movement of the mobile objects in the scene, the features matched by the ego-motion algorithm are eliminated, and the ego-motion algorithm is re-applied to the remaining, previously unmatched features. The resulting rigid transformation is subtracted from the rigid transformation corresponding to ego-motion, obtaining the displacement of the mobile object (Faugeras, 1993).

Research in this area has the goal of developing robots capable of autonomous navigation. While most research has focused on navigation in indoor (structured) environments, the problem of navigation in outdoor, unstructured environments is of greater interest, given its multiple applications, ranging from cruise missile control to robotic rover planetary exploration.

References

Biederman I (1985) Human image understanding: recent research and a theory. *CVGIP: Image Understanding* **32**: 29–73.

Biederman I (1987) Recognition by components: a theory of human image understanding. *Psychological Review* **94**: 115–147.

Biederman I and Gerhardstein PC (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance* **19**(6): 1162–1182.

Binford TO (1987) Generalized cylinder representation. In: Sahpiro SC (ed.) *Encyclopedia of Artificial Intelligence*, pp. 321–323. New York, NY: John Wiley & Sons.

Binford TO (1971) *Visual Perception by Computer*. IEEE Conference on Systems and Control, Miami Beach, Florida.

Borenstein J, Everett HR and Feng L (1996) *Navigating Mobile Robots: Systems and Techniques*. Wellesley, MA: A K Peters Ltd.

Brooks RA (1981) Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence* **17**: 285–348.

Burns J, Weiss RS and Riseman EM (1992) The non-existence of general-case view-invariants. In: Mundy JL and Zisserman A (eds) *Geometric Invariance in Computer Vision*, pp. 120–131. Cambridge, MA: MIT Press.

Cutzu F and Edelman S (1998) Representation of object similarity in human vision: psychophysics and a computational model. *Vision Research* **38**(15/16): 2229–2258.

Dickinson S, Bergevin R, Biederman I *et al*. (1997a) Panel report: the potential of geons for generic 3-D object recognition. *Image and Vision Computing* **15**(4): 277–292.

Dickinson S, Metaxas D and Pentland A (1997b) The role of model-based segmentation in the recovery of volumetric parts from range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(3): 259–267.

Duda RO, Hart PE and Stork DG (2001) *Pattern Classification*. New York, NY: Wiley Interscience.

Edelman SE (1999) *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.

Faugeras O (1993) *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press.

Hummel JE and Biederman I (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* **99**: 480–517.

Huttenlocher D and Ullman S (1990) Recognizing solid objects by alignment with an image. *International Journal of Computer Vision* **5**(2): 195–212.

Jolicoeur P, Gluck M and Kosslyn S (1984) Pictures and names: making the connection. *Cognitive Psychology* **16**: 243–275.

Malik J, Belonge S, Leung T and Shi J (2001) Contour and texture analysis for image segmentation. *International Journal of Computer Vision* **43**(1): 7–27.

Marr D and Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional structure. *Proceedings of the Royal Society of London B* **200**: 269–294.

Mundy J, Zisserman A and Forsyth D (1994) Applications of invariance in computer vision. *Lecture Notes in Computer Science*, vol. 825. Berlin: Springer-Verlag.

Palmer SE, Rosch E and Chase P (1981) Canonical perspective and the perception of objects. In: Long J and Baddeley A (eds) *Attention and Performance IX*, pp. 135–151. Hillsdale, NJ: Erlbaum.

Pentland A (1986) Perceptual organization and the representation of natural form. *Artificial Intelligence* **28**: 293–331.

Poggio T (1990) *A Theory of How the Brain Might Work*. Cold Spring Harbor Symposia on Quantitative Biology, LV, pp. 899–910.

Poggio T and Edelman S (1990) A network that learns to recognize three-dimensional objects. *Nature* **343**: 263–266.

Poggio T and Girosi F (1990) Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247**: 978–982.

Reiss TH (1993) Recognizing planar objects using invariant image features. *Lecture Notes in Computer Science*, vol. 676. Berlin: Springer-Verlag.

Rosch E (1978) Principles of categorization. In: Rosch E and Lloyd B (eds) *Cognition and Categorization*, pp. 27–48. Hillsdale, NJ: Erlbaum.

Rosch E, Mervis CB, Gray WD, Johnson DM and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology* **8**: 382–439.

Rothwell C (1995) *Object Recognition through Invariant Indexing*. Oxford, UK: Oxford University Press.

Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* **210**: 390–397.

Shi J and Malik J (2000) Normalized cuts and image segmentation. *PAMI* **22**(8): 888–905.

Tarr MJ and Bülthoff HH (1995) Is human object recognition better described by geon structural descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance* **21**(6): 1494–1505.

Tarr MJ, Bülthoff HH, Zabinski M and Blanz V (1997) To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science* **8**(4): 282–289.

Witkin AP and Tenenbaum JM (1983) On the role of structure in vision. In: Beck, Hope and Rosenfeld (eds) *Human and Machine Vision*, pp. 481–543. Academic Press.

Further Reading

Basri R and Ullman S (1993) The alignment of objects with smooth surfaces. *CVGIP* **57**(3): 331–345.

Fleuret F and Geman D (2001) Coarse-to-fine face detection. *International Journal of Computer Vision* **41**: 85–107.

Freeman WT (1993) *Exploiting the Generic View Assumption to Estimate Scene Parameters*. Proceedings of the 3rd International Conference on Computer Vision, pp. 347–356, Washington, DC.

Hoffman DD and Richards WA (1984) Parts of recognition. *Cognition* **18**: 65–96.

Koenderink JJ and van Doorn AJ (1979) The internal representation of solid shape with respect to vision. *Biological Cybernetics* **32**: 211–217.

Nitzberg M, Mumford D and Shiota T (1993) Filtering, segmentation and depth. *Lecture Notes in Computer Science*, vol. 662. Berlin: Springer-Verlag.

Weinshall D, Werman M and Tishby N (1993) Stability and likelihood of views of three-dimensional objects. In: Basri R, Schild U and Stein Y (eds) Proc. 10th Israeli Symposium on Computer Vision and AI, pp. 445–454.

Yang M, Ahuja N and Kriegman D (2002) Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1): 34–58.

Working Memory, Computational Models of

Intermediate article

Stephan Lewandowsky, University of Western Australia, Nedlands, Western Australia, Australia
Simon Farrell, Northwestern University, Evanston, Illinois, USA

CONTENTS

Comparison of localist, distributed, and rule-based models of working memory gives insight into the essential processes underlying short-term retention of information.

VARIETIES OF MEMORY

Memory researchers often distinguish between two manifestations of memory, one dedicated to the retention of information for very short periods, known as short-term memory (STM) or working memory, and one dedicated to long-term storage, known as long-term memory (LTM). Working memory is involved when information is needed briefly but immediately; for example, when dialling a telephone number after looking it up in the directory. Long-term memory, by contrast, is involved when information needs to be available for repeated access across larger timescales, for example one's own telephone number.

Many theorists view working memory as a distinct functional and structural entity that requires examination and explanation in its own right. In support of this view, many empirical dissociations between working memory and LTM have been reported. For example, whereas working memory is adversely affected by phonological similarity (e.g., memory for the list 'B D T V G' is poorer than for the list 'T K X S M'), this has little effect on LTM. Conversely, semantic similarity has little effect on working memory whereas its effects on LTM can be pronounced.

This article focuses on models of working memory, in particular those that are formulated at a computational rather than a verbal level. Computational models provide a quantitative account of the data (e.g., through computer simulation), and thus offer more theoretical precision than verbal models. Although contemporary computational models encompass a diverse range of theoretical approaches and assumptions, they all focus on performance in the immediate serial recall task, in which individuals must repeat back short sequences of verbal items in the correct order.

Much of the research involving this task was stimulated by Baddeley's (1986) pioneering verbal theory of working memory, in particular his conception of the so-called phonological loop. Many contemporary models retain at least an empirical connection with Baddeley's proposal.

Baddeley's Phonological Loop

In Baddeley's (1986) model of working memory, verbal information such as lists of digits, letters, or words is maintained in a system known as the 'phonological loop'. The phonological loop involves two components: a phonological store, which is a receptacle, of limited capacity, of information that is subject to constant decay, and an articulatory control process, which is a rehearsal mechanism that counteracts decay by periodically refreshing the items in the phonological store. Because the representational code is phonological, similar-sounding items interfere with each other, thus producing the observed disadvantage for lists of phonologically similar items.

At a verbal level of explanation, these simple assumptions account for numerous results, for example the fact that memory span (the number of items for which immediate serial recall is possible) is a linear function of pronunciation rate of the material. Memory span is smaller for items that take longer to pronounce, presumably because this increases their exposure to decay in between rehearsals. The phonological loop also

accommodates the observed elimination of this effect when subjects recite irrelevant material during list presentation. Under these 'articulatory suppression' conditions, memory span for visually presented lists is poor, but equal for items with short and long pronunciation durations, presumably because the rehearsal that could prevent trace decay is suppressed.

WORKING MEMORY AS ACTIVE LOCALIST UNITS

Several computational models have been proposed that formalize some of the concepts arising from verbal theorizing within the phonological loop framework. Although these models handle a largely overlapping set of phenomena, they differ in the processes thought to underlie maintenance and retrieval of information. We first consider 'localist' models, in which each item is represented by a distinct unit whose activation represents memorial strength.

A Localist Connectionist Network Model of the Phonological Loop

Burgess and Hitch (1999) presented a network model of the phonological loop that combines localist representations with a time-driven 'competitive queuing' mechanism for ordered retrieval. Thus, when an item is presented for study, phoneme units that correspond to its pronunciation pattern are activated, and that activation is fed forward to a layer of representational units. The representational unit whose activation is greatest, because it is excited by all presented phonemes, is associated with a temporal 'sliding window' representation of context. Context is represented by numerous units that are organized in temporal sequence, so that a contiguous cluster is activated at any given time. Across time, the set of activated units slides along the temporal dimension, thus producing some overlap between the sets of units that are active at adjacent timesteps. At each timestep, the weights between the activated representational unit and the context units are strengthened through Hebbian learning. Connection weights are also assumed to decay with time, thus capturing the transient nature of the phonological loop.

At recall, the context signal is reset, and its subsequent advance along the original temporal sequence is used to cue retrieval. The context signal activates a given representational unit to the extent that it overlaps with the time signal present at study. Because temporally adjacent context representations overlap, several units will be partially activated. The resultant competition among representational units is resolved through a 'winner takes all' process based on mutual inhibition. Once an item has been selected for report, its associated output phoneme pattern is activated, which simulates spoken recall. Immediately upon recall, the representational unit is suppressed and thus removed from subsequent competitions. This item selection process is known as 'competitive queuing'.

Several components of the model can be linked to Baddeley's verbal theory. Competitive queuing implements (part of) the articulatory control process within the phonological loop; and the phoneme units that activate representational units instantiate the phonological store.

The model handles the effects of pronunciation time and articulatory suppression mentioned above in connection with Baddeley's model. Moreover, in important contrast to other computational models, Burgess and Hitch can accommodate a large body of neuropsychological literature by simulating various different types of brain damage through selective lesioning of connections. Finally, the model handles many basic aspects of serial recall, such as the bowed shape of the serial position curve, error patterns, and so on. An earlier version of the model, using the same basic architecture and representation of context, also accommodated the effect of temporal grouping of items on the serial position curve.

Two drawbacks of the model can be identified. Firstly, the model mixes representational approaches: representational units are localist (each item is represented by a single unit) whereas context is represented in a distributed manner. It is unclear on what grounds this mixed representation could be justified. Secondly, the model is complex: it includes four sets of weights, connecting context to items, input phonemes to items, items to output phonemes, and input phonemes to output phonemes. All but the last set of connections are adjustable through Hebbian learning, and each adjustable set in turn involves a variable short-term weight that decays over time as well as a static long-term weight. A large number of parameters is required to govern and coordinate these multiple sets of weights, which renders it difficult to assess the overall parsimony of the approach and the power of its core assumptions. The latter issue can be examined by considering a much simpler model that also uses localist representations and competitive queuing.

The Primacy Model: Parsimonious Competitive Queuing

The primacy model (Page and Norris, 1998) implements competitive queuing without a context or timing signal. At study, localist item nodes are activated to an extent determined by the positions of the corresponding items in the list. The first list item is maximally activated, and the activation of all other nodes decreases for successive serial positions, thus setting up a 'primacy gradient'. Mimicking the presumed properties of the phonological loop, the activation of a node begins to decay immediately after the item has been presented. If rehearsal occurs, decay is reversed and activations are restored to their original levels as dictated by the primacy gradient.

At recall, the strongest item is reported and then immediately suppressed. Thus, because the first list item is maximally active, it will also be recalled first. Because recall is followed by suppression, that item is unlikely to be recalled again, and the second most active item is recalled next. This process of recall followed by suppression continues until all list items have been recalled. To allow the model to generate errors, noise is introduced into activations before selecting an item for report.

The model, though relying on very few parameters, can account for many results. It predicts the standard serial position curve, with an advantage for items at the start of the list (the primacy effect) and a similar, though smaller, advantage for the last few items in a list (the recency effect). The model also predicts the associated patterns of transposition errors, in particular the 'locality constraint', which refers to the fact that items tend to be reported in proximity to their correct positions. The model also handles modality effects (recall is enhanced for words presented auditorily), the effects of pronunciation time discussed earlier, and list length effects (recall is worse for longer lists).

One shortcoming of the primacy model concerns 'fill-in'. This is the observed phenomenon whereby if an item is recalled too early (e.g. item 2 is recalled first), the displaced item is more likely to be recalled next than the item following the transposed item (i.e. item 1 is more likely to be recalled next than item 3). The primacy model handles the phenomenon of fill-in, but greatly over-predicts its extent, as an earlier item not recalled will be a formidable competitor at the next output position.

Another limitation of the primacy model is that it cannot account for grouping effects. Temporal separation of a list into distinct small groups is known to change the serial position curve and patterns of transpositions (see below); this is often taken as evidence of a hierarchical or multi-level representation. As the primacy model can order items only along the single dimension of node activations, grouping effects are beyond its purview. A model that overcomes these limitations is the 'start–end model'.

The Start–End Model

The start–end model (SEM) (Henson, 1998) assumes that, at presentation, a new token is created in short-term memory that contains a representation of the list item and information about its list position. That positional context information is provided by a start marker and an end marker whose strengths decrease and increase, respectively, across list positions. When considered together, the strengths of the start and end markers define a unique list position, and those strengths are stored together with the list item. At retrieval, the positional context for each position is reinstated, and the token whose encoded context is most similar to the cue is recalled. Thus the success of recall depends on the overlap between the positional context provided by the start and end markers for that position and all other possible positions. As in the primacy model, the assessment of overlap is noisy, and report of an item is followed by its suppression.

SEM matches the explanatory power of the primacy model. It also reproduces retention interval effects and the finding that protrusion errors (i.e. incorrectly recalling an item from list $n-1$ on list n) tend to maintain their position from list $n-1$. Notably, SEM gives a thorough account of grouping effects. Grouping increases overall list recall and yields a scalloped serial position curve, with small primacy and recency effects within each group. SEM accounts for these grouping effects by incorporating two sets of markers, item markers and group markers. SEM also handles the intricate pattern of transposition errors in grouped lists, for example the fact that terminal group items are likely to transpose with each other even if group sizes are unequal. For example, in a list presented as a group of three followed by a group of four, the third item in the first group, when erroneously recalled in the second group, is more likely to be reported in group position 4 than position 3. This supports SEM's assumption of relative coding involving end markers; and it casts doubt on absolute coding, such as that used in the Burgess and Hitch model, which would predict that transpositions between groups should maintain their positions within the groups.

Most problematic in SEM is the end marker, whose strength increases exponentially across serial positions, with a predetermined maximum value for the last input position. This requires that the list length be known ahead of presentation, which is an unreasonable assumption: the experimental data remain largely unchanged even if people cannot anticipate list lengths. Henson and Burgess (1997) suggest a solution to this problem that involves competition among multiple autonomous oscillators of different frequencies. When list presentation is complete, only those oscillators retain a role in item coding whose distinctiveness has reached a peak at the end of the list.

WORKING MEMORY AS A DISTRIBUTED NETWORK

Distributed models differ from the localist models described above by assuming that items are represented not by identifiable single units but as patterns of activation across a collection of units. Typically, each unit is involved in the representation of several items, and information storage is accomplished by adjusting the weights between layers of such units.

Unlike some of the localist models, distributed models typically reject any explicit theoretical link with the phonological loop and its notions of decay and rehearsal. Instead, they typically ascribe forgetting to interference. These models also postulate that memory is inherently associative and thus requires some form of cue to elicit recall.

The Oscillator-based Model of Associative Recall

In the 'oscillator-based model of associative recall' (OSCAR) of Brown *et al*. (2000), study items are represented by vectors that consist of many elements. Each element takes on an activation value that is randomly sampled from a Gaussian distribution. Upon presentation, each item is associated with a contextual cue, and, through Hebbian learning, is superimposed upon all earlier memories in a matrix of weights. The contextual cue is provided by a collection of temporal oscillators with differing frequencies whose overall activation pattern uniquely specifies a given point in time.

At recall, the temporal signal is reproduced by first resetting the oscillators and then letting them oscillate. Items are retrieved by cueing the memory matrix with the temporal context at each timestep. Owing to the distributed representations and the use of Hebbian learning, the retrieved item is 'noisy' and needs to be 'redintegrated' into an overt response. (Redintegration is the process by which partial memorial information is disambiguated and hence rendered available for output.) In OSCAR, redintegration is performed by comparing the retrieved item with a pool of possible responses and choosing the one that is most similar. Once a response has been chosen, it is typically suppressed and thus unavailable for further report.

The explicit representation of temporal context is related to the approach chosen by Burgess and Hitch (1999), and it allows OSCAR to handle time effects like recency judgments and various retention intervals. However, the strict adherence to temporal oscillators brings the model into conflict with the observation that transpositions between groups tend to maintain their relative positions within groups. Another limitation, typical of distributed models, is the fact that redintegration and response suppression affect a different form of representation (i.e. discrete responses among the pool of competitors) from that used to encode serial order (i.e. the weight matrix). A recent solution to this problem is presented below.

The Theory of Distributed Associative Memory

Brief consideration must be given to the 'theory of distributed associative memory (TODAM) (Lewandowsky and Murdock, 1989). Although this theory has been superseded by contemporary approaches, the reasons for its obsolescence are informative.

In contrast to all models discussed so far, TODAM rejects the idea of positional encoding; instead, successive items are associated with each other in a chain. Recall involves probing memory with each successive list item. In its simplest version, this 'chaining' model is immediately ruled out because it postulates that recall ceases with the first error, and thus cannot accommodate recency. TODAM circumvents this problem because its distributed representations, as in OSCAR, lead to 'noisy' retrievals that require subsequent redintegration. This can generate recency because the retrieved approximation can correctly cue the next item even though redintegration yields an incorrect response. However, like OSCAR, TODAM implements redintegration by comparing the retrieved response candidate with discrete representations of all possible responses. As with OSCAR, this solution is unsatisfactory because the representational assumptions of an important retrieval component are at odds with the essential properties of a distributed memory system.

TODAM handles many of the serial order effects reviewed in the context of the earlier models. Notwithstanding that apparent success, several recent studies cast doubt on the plausibility of chaining as a mechanism for ordering items (e.g., Henson *et al*., 1996). Chaining of items predicts catastrophic errors in lists of mixed confusable and non-confusable items, as the large overlap between similar items should cause much interference among cues, and recall of the following items should be impaired. The data very clearly do not exhibit this pattern, and the failure of TODAM thus strongly suggests that associations between items do not play a crucial role in the representation of serial order information.

Redintegration With Distributed Representations

The redintegration difficulties mentioned in connection with OSCAR and TODAM have been, at least in principle, resolved by a recent dynamic attractor model (Lewandowsky, 1999). This model was not intended as a complete description of serial recall, but instead provides a mechanism for the redintegration of the noisy outputs of distributed models, such as TODAM and OSCAR, into overt responses. The model redintegrates a partial response by iteratively feeding it back into a weight matrix composed of the self-associations of all study items. After enough iterations, the partial response is mathematically guaranteed to be disambiguated into some overt response, though the probability of the correct response being chosen depends (roughly) on the similarity between the partial response and the correct item.

Lewandowsky's (1999) redintegration model, with minimal assumptions about how partial information is retrieved from memory, can account for many aspects of serial recall. It produces the recency portion of the serial position curve, the correct shape of the transposition gradient, and the relative incidence of other errors across lists and output positions.

RULE-BASED MODELS OF WORKING MEMORY

In contrast to all the models described above, rule-based models represent a list as a collection of propositions, and as such are only tenuously related to Baddeley's (1986) notion of the phonological loop. The principal model of this type is the ACT-R theory of Anderson (Anderson and Matessa, 1997), according to which a list is encoded in a hierarchical structure composed of propositions that encode the identities of items and their positions in groups and lists.

Retrieval is coordinated by 'production rules', which are condition–action pairs: that is, they execute particular functions when their conditions are true. Errors occur through partial matching of the condition of a production; hence items sharing similar codes may both be activated, leading to potential confusion when noise is incorporated.

ACT-R handles the standard serial position curve, list length effects, the effects of pronunciation duration, the pattern of transpositions, phonological similarity effects, and the role of articulatory suppression. Some of these effects (e.g., list length) occur because there is a limited 'pool' of activation, so that the activation available per item decreases as list length increases. Importantly, the model also accommodates the scarce available latency data from serial recall tasks.

The strength of ACT-R is that it integrates explanations for serial recall performance into a wider theory of cognition. There are, however, also noticeable shortcomings. Although the model qualitatively explains many phenomena, an associated quantitative fit is lacking (Anderson and Matessa, 1997, Figure 8). In particular, ACT-R fails to capture the extent of primacy and recency in the standard serial position curve.

CONCLUSION

Several general comments can be made about the state of computational modeling in working memory. Firstly, there is little connection between the introspectively appealing notion of rehearsal that was central to Baddeley's (1986) phonological loop and the mechanisms embodied in current models. Even the theory that is most closely allied to the phonological loop, that of Burgess and Hitch (1999), does not contain a process that retains the intuitive character of subvocal rehearsal. A reasonable interpretation of this state of affairs is that introspection and intuition can be poor guides to theorizing.

Secondly, there are some interesting similarities between theories. For example, virtually all models assume some form of response suppression to prevent the same unit of information being repeatedly accessed. Another unifying assumption in many of the models is the superiority of encoding of early list items, which results in the primacy effect. In OSCAR, TODAM, SEM, the primacy model, and Lewandowsky's (1999) redintegration model, this is explained using some parameter which causes the

quality of storage to decay across serial positions. Similarly, most models explain the phonological similarity effect by relying on several stages to allow independent item and order confusions. This strategy is followed in the primacy model, SEM, the Burgess and Hitch (1999) model, and ACT-R.

The differences between the models are also enlightening. For example, it appears highly unlikely that a model without any kind of associations (e.g., the primacy model) will be able to account for grouping effects. Similarly, a comparison of TODAM with the other models suggests that pure chaining cannot underlie serial recall and that positional coding is a more likely alternative. A comparison of the Burgess and Hitch model and SEM suggests further that that positional encoding is relative, not absolute.

Some of the present models characterize working memory, STM, as a separate system, differing from LTM in processes and representation of items (e.g., the primacy model and SEM). These models probably cannot accommodate the ample evidence that LTM can affect performance on working memory tasks (e.g. Hulme *et al*., 1997). The issue of an LTM contribution is important because it relates to the purpose of working memory. There is now much empirical evidence that a separate verbal short-term system plays an integral role in vocabulary acquisition, or that vocabulary acquisition and immediate serial recall have a common substrate (Baddeley *et al*., 1998). Although the models discussed here have not been applied to vocabulary acquisition, improved knowledge of how temporal order is represented in the serial recall paradigm may be critical to our understanding of language development.

References

Anderson JR and Matessa M (1997) A production system theory of serial memory. *Psychological Review* **104**: 728–748.

Baddeley AD (1986) *Working Memory*. Oxford, UK: Oxford University Press.

Baddeley AD, Gathercole SE and Papagno C (1998) The phonological loop as a language learning device. *Psychological Review* **105**: 158–173.

Brown GDA, Preece T and Hulme C (2000) Oscillator-based memory for serial order. *Psychological Review* **107**: 127–181.

Burgess N and Hitch GJ (1999) Memory for serial order: a network model of the phonological loop and its timing. *Psychological Review* **106**: 551–581.

Henson RNA (1998) Short-term memory for serial order: the Start–End Model. *Cognitive Psychology* **36**: 73–137.

Henson RNA and Burgess N (1997) Representations of serial order. In: Bullinaria JA, Glasspool DW and Houghton G (eds) *4th Neural Computation and Psychology Workshop*, pp. 283–300. London, UK: Springer-Verlag.

Henson RNA, Norris D, Page MPA and Baddeley D (1996) Unchained memory: error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology* **49A**: 80–115.

Hulme C, Roodenrys S, Schweickert R *et al*. (1997) Word-frequency effects on short-term memory tasks: evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**: 1217–1232.

Lewandowsky S (1999) Redintegration and response suppression in serial recall: a dynamic network model. *International Journal of Psychology* **34**: 434–446.

Lewandowsky S and Murdock BB (1989) Memory for serial order. *Psychological Review* **96**: 25–57.

Page MPA and Norris D (1998) The primacy model: a new model of immediate serial recall. *Psychological Review* **105**: 761–781.

Further Reading

Conway MA (ed.) (1997) *Cognitive Models of Memory*. Cambridge, MA: MIT Press.

Gathercole SE (ed.) (1996) *Models of Short-Term Memory*. Hove, UK: Psychology Press.

Gupta P and MacWhinney B (1997) Vocabulary acquisition and verbal short-term memory: computational and neural bases. *Brain and Language* **59**: 267–333.

Levy JP, Bairaktaris D, Bullinaria JA and Cairns P (eds) (1995) *Connectionist Models of Memory and Language*. London, UK: UCL Press.

Miyake A and Shah P (1999) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge, UK: Cambridge University Press.

Asset Market Experiments

Intermediate article

Daniel Friedman, University of California, Santa Cruz, California, USA
Hugh M Kelley, Indiana University, Bloomington, Indiana, USA

CONTENTS

Laboratory evidence indicates that prices of financial assets such as stocks and bonds respond to changes in the assets' fundamental value but are also sometimes distorted by investors' cognitive and other biases.

INTRODUCTION

Any society allocates some resources to current consumption and some to investment, to building a better future. Asset markets determine the extent and form of investment in modern economies. Non-market allocation procedures such as those once used in Communist countries clearly worked less well and became less prevalent in the late twentieth century. Asset markets now have global scope and significance.

By definition, asset markets are efficient when asset prices reflect all relevant information about investment opportunities. Theory shows that efficient asset markets lead society to choose only the most productive investment prospects, and to choose the best overall level of investment.

The efficient asset price is called *fundamental value*. Actual asset prices are set by fallible human investors in imperfect markets, and thus may contain other components, called *bubbles*, that can lead to inefficient resource allocation and impair future wellbeing.

Laboratory and field evidence sheds light on asset market efficiency. Asset markets sometimes compensate for investors' cognitive biases, but at other times they amplify them and produce large bubbles. Laboratory experiments help to test policies intended to improve asset market efficiency.

FUNDAMENTAL ASSET VALUE

An asset is anything that provides its owner with a stream of benefits over time. Its economic value is the monetary equivalent of the net benefits it provides. Valuation of a real asset (such as a house, a pizza delivery car or a microprocessor production facility) involves estimating prices for the services the asset generates and for the resources required to maintain its productivity. This article will focus on financial assets, such as stocks and bonds, which are easier to value since they directly promise a monetary income stream. A stock, for example, promises annual per share dividends chosen by the company's board of directors.

Valuing a known income stream is straightforward. One computes the discounted present value: the amount of cash that, if put on deposit now and withdrawn over time, could replicate the income stream. For example, if the asset is an annuity (a simple bond) that promises \$1000 every year for 30 years, and the annual interest rate is 7%, then the asset value is $(\$1000)((1.07)^{-1} + (1.07)^{-2} + \ldots + (1.07)^{-30})$ or, summing the geometric series and simplifying, $(\$1000)(1 - (1.07)^{-30})/(0.07) = \$12{,}409$.

Valuation is more interesting when the income stream is uncertain, either because the promise is vague (as for a stock) or because the promise might not be kept (as in the case of a 'junk' bond). The economic value is defined to be the mathematical expectation of the income stream, discounted at the interest rate appropriate for the associated risk. For example, if the \$1000 annual payment in the previous example had a 50% independent probability of being canceled every year, and this requires a 2% risk premium or 9% interest rate, then the expected annual income is \$500 and the asset value is $(\$500)(1 - (1.09)^{-30})/(0.09) = \$5137$.

Two questions arise. Firstly, what is the appropriate interest rate (or risk premium)? This question stimulated the development of modern finance theory in the 1950s and 1960s, but we will not pursue it here. Secondly, how can people with diverse information arrive at a common expectation

of the income stream (or its discounted present value)? Some people may know a lot about the technical performance of a company's new product, others about customer demand, or competing products, or production costs, or the company's management and financing capacity. Usually nobody has all the relevant knowledge.

'Fundamental asset value' is the present value expectation incorporating all existing information regarding future income. The definition seems to assume that people immediately share all knowledge and combine it without bias or distortion. But people may not share information; they often are cognitively biased. Cognitive biases might cause the market price to deviate from fundamental value.

COGNITIVE BIASES

Which cognitive biases might distort people's estimates of asset value? Some of the main biases are summarized below (Camerer, 1993; Rabin, 1998; Thaler, 1992).

Firstly, judgment biases may distort the aggregation of diverse information (Massaro and Friedman, 1990) and income stream estimates. In particular, investors may:

- neglect some pieces of information and over-weight others, as in cue competition;
- overestimate the resemblance of the future to the immediate past, as in the well-known availability and representativeness heuristics or the anchoring and inertia biases;
- over-weight new information and neglect old information, as in overreaction to news and base rate neglect;
- regard ambiguous news as reinforcing current beliefs, as in the confirmatory bias;
- overrate the precision of their own information, relative to other traders' information, as in overconfidence;
- indulge in the gambler's fallacy or magical thinking, perceiving patterns in random data;
- react incorrectly to increasing information precision, or switch biases depending on state, for example, by overreacting to news when asset prices are volatile but underreacting otherwise.

Secondly, when moving from income estimates to asset value estimates, investors may apply hyperbolic discounting, using too high interest rates when comparing current income with near future income, and too low interest rates when comparing income received at two distant dates (Ainslie, 1992).

Thirdly, investors may make decision errors when they buy and sell assets, such as overvaluing assets they currently hold, as in the endowment effect, or making inconsistent risky choices, as in prospect theory, regret theory, or (more generally) decision field theory (Busemeyer and Townsend, 1993).

Fourthly, investors may learn by trial and error, some faster than others, creating additional asset price deviations from fundamental value (Kitzis *et al*., 1998; Busemeyer *et al*., 1993).

Note, however, that individual investors' biases do not translate directly into asset price biases. Market prices are set by subtle interactions between investors that may either attenuate (e.g., Friedman, 2001) or amplify (e.g., Akerlof and Yellin, 1985; DeLong *et al*., 1990) the individual biases and learning heterogeneities. Thus, whether (or when) asset price follows fundamental value is an empirical question in its own right.

FIELD EVIDENCE

Evidence from existing asset markets is interesting but inconclusive. Historians point to dramatic episodes of asset price increase and collapse, from the South Sea bubble and 'tulipmania' in the sixteenth century to Japan's 'bubble' economy of the late 1980s, various 1990s financial crises (in Western Europe, Latin America, East Asia, and Russia), and the 'dot com' bubble of 2000.

Some economists argue that these episodes are merely unusual movements in fundamental value (Garber, 2000). Economists cannot observe the private information held by traders in the field, and therefore have no direct measure of fundamental value or bubbles. Some indirect field evidence, however, favors the bubbles interpretation.

Shiller (1981) and later writers surveyed in LeRoy (1989) show that stock market indices are much more volatile than would be justified by subsequent changes in dividends. Roll (1984) argues that changes in the fundamental value of US orange juice futures come predominantly from two observables (Florida weather hazard and Brazil supply) but account for only a small portion of the actual price variability. Additionally, nonfundamental effects consistently observed, such as the day of the week or month and January effects, also suggest that stock prices are more volatile than fundamental values. Finally, in a natural field experiment, the stock prices (but arguably not the fundamental values) are far less volatile on weekends and days when the exchange is closed for upgrades. On the other hand there is also field evidence suggesting efficient information aggregation. The Iowa Electronic Market has consistently outperformed major polls in predicting election outcomes (Forsythe *et al*., 1999) and

similarly with the Hollywood Stock Exchange for box office revenues and Oscar winners (Pennock *et al*., 2001).

Field evidence is indirect and can seem either to support or to undermine claims of market efficiency.

LABORATORY EVIDENCE

Laboratory asset markets provide direct evidence, because the experimenter can control the information available to traders and can always compute the fundamental value. The first generation of asset market experiments in the early 1980s used oral double auction trading procedures similar to the traditional Chicago trading pits. To sharpen inferences, the laboratory markets are much simpler than field markets: for example, they often allow only 8 to 16 subjects to trade a single asset. Most of the results described below involve experienced traders who are fully adapted to the laboratory market. See Sunder (1993) for an excellent early survey, and Holt (1999) for an online bibliography.

Market Attenuation of Traders' Biases

There are several distinct forces that tend to attenuate biases. Firstly, people can learn to overcome their biases when the market outcomes make them aware of their mistakes. Secondly, to the extent that biased traders earn lower profits (or make losses), they will lose market share and will have less influence on asset price. Thirdly, institutions evolve to help people overcome cognitive limitations: for example, telephone books mitigate the brain's limited digital storage capacity. Trading procedures such as the oral double auction evolved over many centuries and seem to enhance market efficiency.

Oral double auctions allow all traders to observe other traders' attempts to buy and sell, and might enable them to infer other traders' information. Moreover, the closing price is not set by the most biased trader or even a random trader. The most optimistic traders buy (or already hold), and the most pessimistic traders sell (or never held), the asset, so the closing price reflects the moderate expectations of 'marginal' traders, the most reluctant sellers and buyers.

These attenuating forces may explain the surprisingly rapid convergence of asset price to fundamental value in the first generation of laboratory experiments. Forsythe *et al*. (1982) obtained such convergence for assets that paid different dividends to different traders over two periods. Likewise, Plott and Sunder (1982) obtained convergence to an efficient asset price for a single-period asset even when some traders had inside information. Friedman *et al*. (1984) found that simultaneous operation of spot and futures markets improved convergence to an efficient asset price and allocation when assets paid different dividends to different traders over three periods and traders knew only their own dividend schedule. Generally, convergence was first observed in the last dividend period, then in the middle period as traders correctly anticipated last-period prices, and finally in the first trading period as traders learned the asset's present value.

Market Amplification of Traders' Biases

Later experiments detected systematic discrepancies between price and fundamental value in more complex environments. Copeland and Friedman (1991) found that in a computerized double auction with three information events and eight states, a model of partial aggregation predicted price better than full aggregation or fundamental value.

Several experimental teams found that insider information was incorporated into asset price less reliably and less quickly when the asset paid the same dividend stream to each trader and the number (or presence) of insiders was not publicly known. Some data suggest the following scenario. An uninformed trader A observes trader B attempting to buy (due to some slight cognitive bias, say) and mistakenly concludes that B has favorable inside information. Then A tries to buy. Now trader C concludes that A (or B) is an insider and tries to mimic their trades. Other traders follow, creating a price bubble.

Such 'information mirages', or 'herding' bubbles, amplify the biases of individual traders, but they cannot be produced consistently, since incurred losses teach traders to be cautious when they suspect the presence of better-informed traders. (This lesson does not necessarily improve market efficiency, however, since excessive caution impedes the aggregation of information.)

Smith *et al*. (1988) found large positive bubbles and crashes for long-lived assets and inexperienced traders. Their interpretation invokes the 'greater fool' theory, another bias amplification process. Traders who themselves have no cognitive bias might be willing to buy at a price above fundamental value because they expect to sell later at even higher prices to other traders dazzled by rising prices. Subsequent studies confirmed that such dazzled traders do exist, and that bubbles are

more prevalent when traders are less experienced (individually and as a group), have larger cash endowments, and have less conclusive information.

Other mechanisms that amplify biases in laboratory asset markets include firm managers' discretionary release of information and fund managers' rank-based incentives (James and Isaac, 2000).

Policy Studies

We have only a tentative and fragmentary understanding of when asset markets amplify or attenuate investors' cognitive biases. The impact of proposed policy changes often cannot be predicted reliably, and must be assessed empirically by regulators, asset market makers contemplating reform, and entrepreneurs creating new asset markets. Laboratory markets offer helpful guidance at low cost. Recent relevant research includes performance assessment of:

- alternative market formats, including oral (or face-to-face) double auctions, anonymous electronic double auctions, call or uniform price periodic auctions, and fragmented opaque (or bilateral search) markets;
- trader privileges, such as price posting and access to order flow information;
- transaction taxes, price change limits, and trading suspensions intended (usually ineffectively) to mitigate price bubbles and panics;
- new derivative securities such as call and put options and state-contingent claims.

Current Research

Research continues at a rapid rate along all the lines mentioned. One promising new line of research investigates learning, and judgment and decision biases, in environments similar to asset markets. These environments clearly do attenuate some sorts of biases (Kelley and Friedman, forthcoming) and amplify others (Ganguly *et al*., 2000); eventually such work should clarify the patterns.

Other promising new lines of research integrate agent-based simulation models (e.g. Epstein and Axtell, 1996) into asset markets that may include human traders. The simulated agents, or 'bots', incorporate specified cognitive limitations and the simulations examine the market-level influence of these (e.g. Arthur *et al*., 1997). Gode and Sunder (1993) showed that simple (non-asset) double auction markets are efficient even when populated by 'zero intelligence' agents, bots that are constrained not to take losses but are otherwise quite random. Research is beginning to show how these and more intelligent bots affect efficiency in various sorts of asset markets and how they interact with humans.

The laboratory asset market evidence is more direct than the field evidence, but is still mixed. The laboratory evidence clearly demonstrates that asset price bubbles exist and persist under some circumstances, but that under other circumstances asset prices closely track fundamental values. Future work promises to identify more clearly the circumstances that promote or impair market efficiency. This should lead to improved policy and a better-functioning economy.

References

Ainslie G (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. New York, NY: Cambridge University Press.

Akerlof G and Yellen J (1985) Can small deviations from rationality make significant differences to economic equilibria? *American Economic Review* **75**: 708–720.

Arthur WB, Holland J, LeBaron B, Palmer R and Tayler P (1997) Asset pricing under endogenous expectations in an artificial stock market. In: *The Economy as an Evolving Complex System*, vol. II, pp. 15–44. Reading, MA: Addison-Wesley.

Busemeyer JR, Myung IJ and McDaniel MA (1993) Cue competition effects: theoretical implications for adaptive network learning models. *Psychological Science* **4**: 196–202.

Busemeyer J and Townsend J (1993) Decision field theory: a dynamic cognition approach to decision making. *Psychological Review* **100**: 432–459.

Camerer C (1993) Individual decision making. In: Kagel JH and Roth AE (eds) *The Handbook of Experimental Economics*, pp. 587–704. Princeton, NJ: Princeton University Press.

Copeland T and Friedman D (1991) Partial revelation of information in experimental asset markets. *Journal of Finance* **46**: 265–295.

DeLong JB, Shleifer A, Summers L and Waldmann RJ (1990) Noise traders and risk in financial markets. *Journal of Political Economy* **98**: 703–738.

Epstein JM and Axtell R (1996) *Growing Artificial Societies: Social Science From the Bottom Up*. Cambridge, MA: MIT Press.

Forsythe R, Palfrey T and Plott C (1982) Asset valuations in an experimental market. *Econometrica* **50**: 537–568.

Forsythe R, Rietz TA and Ross TW (1999) Wishes, expectations and actions: price formation in election stock markets. *Journal of Economic Behavior and Organization* **39**: 83–110.

Friedman D (2001) Towards evolutionary game models of financial markets. *Quantitative Finance* **1**: 177–185.

Friedman D, Harrison G and Salmon J (1984) The informational efficiency of experimental asset markets. *Journal of Political Economy* **92**: 349–408.

Ganguly AR, Kagel JH and Moser DV (2000) Do asset market prices reflect traders' judgement biases? *Journal of Risk and Uncertainty* **20**: 219–245.

Garber PJ (2000) *Famous First Bubbles: The Fundamentals of Early Manias*. Cambridge, MA: MIT Press.

Gode D and Sunder S (1993) Allocative efficiency in markets with zero intelligence (ZI) traders: market as a partial substitute for individual rationality. *Journal of Political Economy* **101**: 119–137.

Holt C (1999) *Y2K Bibliography of Experimental Economics and Social Science: Asset Market Experiments.* http://www.people.virginia.edu/∼cah2k/assety2k.htm.

James D and Isaac M (2000) Asset markets: how are they affected by tournament incentives for individuals? *American Economic Review* **90**: 995–1004.

Kelley H and Friedman D (forthcoming), Learning to forecast price. *Economic Inquiry*.

Kitzis S, Kelley H, Berg E, Massaro D and Friedman D (1998) Broadening the tests of learning models. *Journal of Mathematical Psychology* **42**(2): 327–355.

LeRoy S (1989) Efficient capital markets and martingales. *Journal of Economic Literature* **27**: 1583–1621.

Massaro D and Friedman D (1990) Models of decision making given multiple sources of information. *Psychological Review* **97**: 225–252.

Pennock DM, Lawrence S, Giles CL and Nielsen FA (2001) The real power of artificial markets. *Science* **291**: 987–988.

Plott C and Sunder S (1982) Efficiency of experimental security markets with insider information: an application of rational expectations models. *Journal of Political Economy* **90**: 663–698.

Rabin M (1998) Psychology and economics. *Journal of Economic Literature* **36**: 11–46.

Roll R (1984) Orange juice and weather. *American Economic Review* **74**: 861–880.

Shiller RJ (1981) Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* **71**: 421–436.

Smith V, Suchanek G and Williams A (1988) Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica* **56**: 1119–1151.

Sunder S (1993) Experimental asset markets: a survey. In: Kagel JH and Roth AE (eds) *The Handbook of Experimental Economics*, pp. 445–500. Princeton, NJ: Princeton University Press.

Thaler RH (1992) *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. Princeton, NJ: Princeton University Press.

Further Reading

Friedman D (1998) Monty Hall's three doors: construction and deconstruction of a choice anomaly. *American Economic Review* **88**: 933–946.

Hogarth R and Reder M (eds) (1987) *Rational Choice: The Contrast Between Economics and Psychology*. Chicago, IL: University of Chicago Press.

Mackay C (1841) *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds*. London, UK: Bentley.

Shefrin H (2000) *Beyond Greed and Fear*. Boston, MA: Harvard Business School Press.

Shiller R (2000) *Irrational Exuberance*. Princeton, NJ: Princeton University Press.

Bayesian Learning in Games

Intermediate article

JS Jordan, Pennsylvania State University, University Park, Pennsylvania, USA

CONTENTS

Bayesian learning is a method by which players in a game attempt to infer each other's future strategies from the observation of past actions. Under certain assumptions, learning is successful in the sense that players' expectations converge to Nash equilibria of the game.

INTRODUCTION

A Nash equilibrium of a game consists of a strategy for each player that maximizes the player's expected pay-off against the strategies of the other players. It is natural to assume that players seek to maximize their expected pay-offs, but the assumption that each player correctly anticipates the strategies of the others is more problematic. If the players' pay-off functions are public knowledge, then presumably each player could compute the equilibrium strategies, provided that some selection criterion is adopted in the case of multiple equilibria. In most applications of game theory to economics, however, individual characteristics are private information not directly observable by others. This raises the question of whether players might learn from experience.

Suppose that the game is repeated over time while pay-off functions remain fixed. If the actions taken by each player at each repetition are publicly observable, then players might, through some form of inductive inference, learn to form the correct expectations. In game theory, the canonical model of expectation formation is Bayes' theorem, so Bayesian learning provides a natural model of inductive inference. In a Bayesian learning model, each player has probabilistic beliefs about the sequence of actions that the repeated game will generate, and updates those beliefs at each iteration in response to the actions observed up to that point. The question is whether, over time, each player's beliefs about the future actions of the others converge to the correct Nash equilibrium expectations.

If all of the players except one were machines that take the same, possibly randomized, action each time, the inference problem for the single real player would be a straightforward problem of consistent statistical inference. Instead, the fact that all players are learning from each other's actions means that Bayesian learning produces complex interactive dynamics. Despite this complication, the Bayesian learning model provides some encouraging results on the possibility of learning Nash equilibrium expectations.

It is useful to distinguish between two different versions of the Bayesian learning model, which in this article will be termed 'sophisticated' and 'naive'. In the sophisticated Bayesian learning model, players are assumed to be knowledgeable game theorists who derive their expectations of the other players' actions from prior beliefs about the other players' pay-off functions. After each repetition, the observed actions cause them to revise their beliefs and update their expectations. In the naive Bayesian learning model, the players' prior beliefs are arbitrary probability distributions over sequences of actions. Expectations of future actions are updated directly from observed actions via Bayes' theorem, with no knowledge of game theory required.

SOPHISTICATED BAYESIAN LEARNING

Myopic Behavior

Since the game is repeated, players receive a stream of pay-offs over time. If the players ignore future pay-offs in each period, and seek to maximize their one-period pay-off against the expected actions of the other players, then it is natural to ask whether expectations approach correct Nash equilibrium expectations for the one-shot game that is being played each period. In contrast, if players anticipate future repetitions and seek to maximize the

expected discounted sum of future pay-offs, then it is more natural to ask whether expectations approach correct Nash equilibrium expectations for the full repeated game. The case of myopic behavior will be discussed first.

The actions available to player i at each iteration lie in a finite set S_i. There are n players, and, following interation t, each player observes the entire n-tuple $s_t = (s_1, \ldots, s_{nt})$ of chosen actions. A learning process for player i is a sequence of expectation functions e_{it} which associate with any observed finite history of play $h_t = (s_{1t}, \ldots, s_t)$ an expectation $e_{it}(h_t) \in \Delta(S_{-i})$, where $\Delta(S_{-i})$ is the set of probability distributions over $(n-1)$-tuples $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ of actions to be chosen by the other players in period $t+1$. After each iteration t, each player i receives the pay-off $u_i(s_t)$. Player i knows the pay-off function u_i, which remains fixed through time, but does not know the pay-off function of any other player. Player i chooses s_{it} in period t to maximize the one-period expected value of u_i against the expectation $e_{i(t-1)}(h_{t-1})$ of the other players' actions. The question is whether the expectations sequence $(e_{it}(h_t))_{t=1}^{\infty}$ approaches the set of Nash equilibrium expectations.

A Nash equilibrium consists of a strategy $\sigma_i^* \in \Delta(S_i)$ for each player i that is expected-pay-off-maximizing against the strategies of the other players, that is,

$$\sum_{s_i \in S_i} \sigma_i^*(s_i) \sum_{s_{-i} \in S_{-i}} u_i(s_i, s_{-i}) \sigma_{-i}^*(s_{-i}) \geq \sum_{s_i \in S_i} \sigma_i(s_i) \sum_{s-i \in S-i} u_i(s_i, s_{-i}) \sigma_{-i}^*(s_{-i})$$

for every other strategy $\sigma_i \in \Delta(S_i)$. (We write $u_i(s_i, s_{-i})$ for the pay-off resulting from player i using strategy s_i and the other players using strategies s_{-i}.) Given a Nash equilibrium $(\sigma_1^*, \ldots, \sigma_n^*)$, player i's Nash equilibrium expectation is the probability distribution $\sigma_{-i}^* = \sigma_1^* \times \cdots \times \sigma_{i-1}^* \times \sigma_{i+1}^* \times \cdots \times \sigma_n^* \in \Delta(S_{-i})$. Learning has the desired convergence property if, as $t \to \infty$, the distance between $e_{it}(h_t)$ and the nearest Nash equilibrium expectation σ_{-i}^* goes to zero for every player i. Since many games have multiple Nash equilibria, the sequence of expectations may have multiple limit points because it has subsequences converging to different Nash equilibrium expectations. The desired convergence property is that each limit point be a Nash equilibrium.

The expectation functions that constitute Bayesian learning can be derived as follows. The fact that player i knows u_i but not u_j for all $j \neq i$ means that the players face a game of incomplete information, in which each player's private type is the player's pay-off function. The incomplete information is modeled by supposing that player i believes that, prior to the first iteration, Nature chooses each pay-off function u_j according to a probability distribution μ_j over pay-off functions. Since the domain of pay-off functions S is finite, pay-off functions lie in the finite-dimensional vector space $\mathbb{R}^S$. Since best-response strategies are invariant with respect to positive linear transformations, we can reduce the space of possible pay-off functions to the unit sphere $B \subset \mathbb{R}^S$, which is an innocuous but convenient normalization. The pay-off functions determine the game that is being learned, so let $G = B^n$ denote the space of all possible games. All players believe that a game $(u_1, \ldots, u_n) \in G$ is initially chosen by nature according to the prior distribution $\mu = \mu_1 \times \cdots \times \mu_n$ over G. In particular, it is assumed that players share the common belief μ, under which the pay-off functions of different players are independently distributed. These assumptions can be relaxed, as will be discussed below.

The initial expectations e_{i0} are derived as a Bayesian Nash equilibrium of the incomplete information game in which each player i knows u_i and believes that each u_j is distributed according to μ_j. Given an initial expectation $e_{i1} \in \Delta(S_{-i})$, there is a best response $b_i(u_i) \in \Delta(S_i)$ associated with each possible pay-off function u_i. Together with the initial distribution of pay-off functions μ_i, a best-response function $b_i(\cdot)$ determines a distribution of player-i actions $\sigma_i \in \Delta(S_i)$, as $\sigma_i(s_i) = \int_B b_i(u_i)(s_i)\, \mu_i(\mathrm{d}u_i)$. The Bayesian Nash equilibrium condition is that, for each i, $e_{i0} = \sigma_{-i}$, that is, each player's expectation is simply the derived distribution of the best responses of the other players to their expectations.

The actions $(s_{11}, \ldots, s_{n1})$ are chosen at the first iteration as best responses to the expectations e_{i0} for the true pay-off functions $(u_1, \ldots, u_n)$. Each player i then observes each s_{j1} and revises the prior distribution μ_j according to Bayes' theorem. The revised distribution $(\mu_1 = \mu_{11} \times \cdots \times \mu_{n1})$ over G results in new Bayesian Nash equilibrium expectations $e_{i1}(h_1)$, where h_1 is the one-period history $h_1 = (s_{11}, \ldots, s_{n1})$, and so on. At each iteration t, the expectations $e_{it}(h_t)$ are derived as a Bayesian Nash equilibrium for the revised beliefs μ_t. Bayesian Nash equilibria need not be unique, so the expectation functions are not generally determined uniquely. The Bayesian learning model encompasses all expectation functions derived in this way from all initial beliefs $\mu = \mu_1 \times \cdots \times \mu_n$, although many of the convergence results mentioned below place restrictions on the initial beliefs.

Although the expectation functions are not unique, the players' best-response strategies have a useful deterministic property. A mixed strategy can only be a best response if there are two or more actions that each maximize the player's expected pay-off. In particular, there must be actions s_i and s_i' that yield the same expected pay-off, $\sum_{s_{-i}} u_i(s_i, s_{-i}) e_{it}(s_{-i}|h_t) = \sum_{s_{-i}} u_i(s_i', s_{-i}) e_{it}(s_{-i}|h_t)$. Given the expectations $e_{it}(h_t)$, this equation is satisfied only for a proper linear subspace of the space of pay-off functions $\mathbb{R}^S$, and therefore only for a set of pay-off functions having Lebesgue measure zero, i.e., probabilistically negligible. For each t, the set of possible t-period histories is finite; so the set of all histories is countable. Therefore, given any sequence of expectation functions, all pay-off functions except for a set of Lebesgue measure zero have unique best-response actions in every period for every history. If the prior distributions μ_i are absolutely continuous with respect to Lebesgue measure on the sphere $B \subset \mathbb{R}^S$, then the set of pay-off functions having unique best responses throughout the learning process has probability one. Given the expectation functions $e_{it}(\cdot)$, Bayesian-learning players never play mixed strategies, except for a set of games having Lebesgue measure zero.

2×2 games

In general, the derivation of expectations as Bayesian Nash equilibria makes their explicit computation problematic. However, in the case of 2×2 games with uniform prior beliefs, the derivation of expectations is straightforward and provides a useful illustration.

Let player 1 be the row player, player 2 the column player, $S_1 = \{U, D\}$, and $S_2 = \{L, R\}$. Then player i's pay-off function is a 2×2 matrix with the entries $u_i(U, L)$, $u_i(U, R)$, $u_i(D, L)$, and $u_i(D, R)$. However, it will be convenient to employ a normalization, which reduces each player's space of possible pay-off functions to the unit circle $C \subset \mathbb{R}^2$. To motivate this normalization, suppose that player 1 anticipates that player 2 will play L with probability $\sigma_2(L)$. Then player 1's optimal strategy is U, D, or both, according as the quantity

$$\sigma_2(L)[u_1(U, L) - u_1(D, L)] + (1 - \sigma_2(L))[u_1(U, R) - u_1(D, R)]$$

is positive, negative, or zero. Therefore we can subtract the second row of player 1's pay-off matrix from each row, so that the top row is $(u_1(U, L) - u_1(D, L), u_1(U, R) - u_1(D, R))$ and the bottom row is $(0, 0)$. Applying the same normalization to the columns of player 2's pay-off matrix produces the pay-off bimatrix in Table 1. If we ignore the measure-zero possibility that $a = b = 0$, that is, $u_1(U, L) = u_1(D, L)$ and $u_1(U, R) = u_1(D, R)$, then we can further normalize (a, b) to the unit circle without affecting player 1's best response to any mixed strategy played by player 2. Thus we can assume that $a^2 + b^2 = 1$, and (for player 2) that $\alpha^2 + \beta^2 = 1$. Under this normalization, each player's pay-off function is a point on the unit circle, and each 2×2 game is a point on the torus $C \times C$ (we will continue to exclude the degenerate cases $a = b = 0$ and $\alpha = \beta = 0$).

An obvious choice for the priors is the uniform distribution on the unit circle C. For this prior distribution, we will derive the expectations along the two-period history $((U, R), (D, L))$. More precisely, we will compute the first-period expectations $e_0(\cdot)$ and second-period expectations $e_1(\cdot|U, L)$, which are uniquely determined, and show that there are three possible choices for the third-period expectations $e_2(\cdot|(U, L), (D, R))$. First, let σ_2 denote the initial probability distribution on $S_2 = \{L, R\}$ facing player 1. That is, $\sigma_2(L)$ is the probability that player 2 will play L in the first period. Then player 1 will play U in period 1 if player 1's pay-off function, represented by (a, b), satisfies $a\sigma_2(L) + b(1 - \sigma_2(L)) > 0$. That is, the set of player-1 'types' that play U in period 1 is a semicircle $\{(a, b) : a\sigma_2(L) + b(1 - \sigma_2(L)) > 0\}$. The set $\{(a, b) : a\sigma_2(L) + b(1 - \sigma_2(L)) = 0\}$ has prior probability zero and thus can be ignored. Hence, for any expectation $\sigma_2(L)$, the measure of the set of player-1 types that play U, which is simply the prior probability of a semicircle, equals $\frac{1}{2}$. Since this reasoning applies to both players symmetrically, the unique first-period Bayesian Nash equilibrium expectations are $e_{20}(U) = e_{20}(D) = \frac{1}{2}$ and $e_{10}(L) = e_{10}(R) = \frac{1}{2}$.

Now suppose that the actions (U, R) are played in period 1. This reveals that player 1's pay-off function lies in the semicircle $\mathcal{T}_1 = \{(a, b) : \frac{1}{2}a + \frac{1}{2}b > 0\}$, and that player 2's pay-off function lies in the semicircle $\mathcal{T}_2 = \{(\alpha, \beta) : \frac{1}{2}\alpha + \frac{1}{2}\beta < 0\}$. To solve for the Bayesian Nash equilibrium conditional expectations $e_1(\cdot|(U, R))$, let $x = e_{21}(U|(U, R))$, the conditional measure of the set of

Table 1. Normalized pay-off bimatrix for a 2×2 game

	L	*R*
U	*a*, α	*b*, 0
D	0, β	0, 0

player-1 types that play U in period 2, and let $y = e_{11}(L|(U,R))$. Given y, x is simply $\mu_1(\{(a,b) \in \tau_1 : ay + b(1-y) > 0\})/\mu_1(\mathcal{T}_1)$, where μ_1 denotes the uniform distribution on the unit circle. Thus x is simply the relative arc length given by the formula

$$x = (\pi - |\theta(y)|)/\pi$$

where $\theta(y)$ is the angle between the expectations vectors $(\frac{1}{2}, \frac{1}{2})$ and $(y, 1-y)$ (in radians). The analogous formula for y as a function of x is

$$1 - y = (\pi - |\theta(x)|)/\pi$$

These two equations have a unique solution, $x^* \approx 0.82$ and $y^* \approx 0.18$, so the unique Bayesian Nash equilibrium expectations are $e_{21}(U|(U,R)) = x^*$ and $e_{11}(L|(U,R)) = y^*$.

Now suppose that the strategies (D, L) are played in period 2. This reveals that (a, b) and (α, β) satisfy the following conditions:

$$\frac{1}{2}a + \frac{1}{2}b > 0$$

$$y^*a + (1-y^*)b < 0$$

$$\frac{1}{2}\alpha + \frac{1}{2}\beta < 0$$

$$x^*\alpha + (1-x^*)\beta > 0$$

Since $y^* > \frac{1}{2}$ and $x^* > \frac{1}{2}$, equations 5 to 8 imply that $a > 0$, $b < 0$, and $\alpha > 0$, $\beta < 0$. It follows that (U, L) and (D, R) are pure-strategy Nash equilibria for every game $((a,b), (\alpha,\beta))$ that generates the two-period history $((U, R), (D, L))$. It also follows that there is a mixed-strategy Nash equilibrium, but the equilibrium mixed strategies are not yet revealed.

Thus there are three possible Bayesian Nash equilibrium expectations:

- $e_{22}(U|(U,R),(D,L)) = e_{12}(L|(U,R),(D,L)) = 1$
- $e_{22}(U|(U,R),(D,L)) = e_{12}(L|(U,R),(U,L)) = 0$
- $e_{22}(U|(U,R),(D,L)) \approx 0.61$, $e_{12}(L|(U,R),(U,L)) \approx 0.39$

The first two are 'pure strategy' Bayesian Nash equilibria. They correspond to the pure-strategy Nash equilibria and reveal no further information about each player's pay-off function. The third is a 'quasi-mixed' Bayesian Nash equilibrium. It further partitions the pay-off types revealed by the history $((U, R), (D, L))$.

This multiplicity of Bayesian Nash equilibria continues for all future periods. If the 'quasi-mixed' equilibrium is selected infinitely often, the expectations in those periods will converge to the mixed-strategy Nash equilibrium determined by the true pay-off types (a, b) and (α, β).

Far-sighted Behavior

The learning model described above addresses the question of whether players can learn from repeated experience to form correct Nash equilibrium expectations for the true game determined by the pay-off functions $(u_1, \ldots, u_n)$. However, the repetitions introduce the possibility that players may anticipate future pay-offs and take account of the effect that their current actions may have on the future actions of the other players. This can be modeled by supposing that player i has an additional pay-off characteristic consisting of a discount factor $1 > \delta_i \geq 0$, and chooses an action s_{it} in period t to maximize the expected discounted sum of current and future pay-offs $(1-\delta_i)\sum_{\tau=t}^{\infty} \delta_i^{\tau-t} u_i(s_\tau)$. Myopic behavior corresponds to the special case $\delta_i = 0$ (under the convention $0^0 = 1$). The prior beliefs μ_i can be extended to cover the discount factors δ_i as well as the stage-game pay-off functions u_i. The game that the players are learning about is now the full repeated game determined by the discount factors together with the stage-game pay-off functions. Like the stage game, the true repeated game is stationary over time, so the true repeated game is itself being repeated over time in the sense that each period is the first period of the game to be played from that period on.

A strategy for player i in a repeated game is a sequence of functions $f_{it} : S_1 \times \cdots \times S_t \to \Delta(S_i)$ that determine player i's (possibly randomized) action in period $t+1$ as a function of the previous history of play. The sequence of functions $(f_{it})_{t=0}^{\infty}$ is called a 'behavior strategy'. A Nash equilibrium for a repeated game consists of a behavior strategy for each player that maximizes the expected discounted sum of pay-offs against the behavior strategies of the other players. Any Nash equilibrium of the stage game, if repeated every period, is also a Nash equilibrium of the repeated game. More precisely, if $(\sigma_1^*, \ldots, \sigma_n^*) \in \Delta(S_1) \times \cdots \times \Delta(S_n)$ is a Nash equilibrium of the stage game, then the constant-behavior strategies in which player i plays σ_i^* in every period after every history is also a Nash equilibrium of the repeated game. However, repeated games typically have many more Nash equilibria, in which players can influence one another's actions over time, especially if the discount factors are near unity.

The derivation of expectations as Bayesian Nash equilibria extends directly to repeated games

(Jordan, 1995). In this setting, player i's initial expectation e_{i0} is a probability distribution over the behavior strategies of the other players. A theorem due to Kuhn simplifies the analysis by establishing that any probability distribution over behavior strategies generates the same distribution over action sequences as a single appropriately chosen behavior strategy. Hence the expectation e_{i0} can be represented as an $(n-1)$-tuple of behavior strategies $((f^i_{jt})^\infty_{t=1})_{j\neq i}$, and the subsequent expectations $e_{it}(h_t)$ reduce to the behavior strategies $((f^i_{j\tau}(h_t,\cdot))^\infty_{\tau=t})_{j\neq i}$. The question is whether these expected behavior strategies approach Nash equilibrium behavior strategies for the true pay-off characteristics $((u_1,\delta_1),\ldots,(u_n,\delta_n))$.

NAIVE BAYESIAN LEARNING

The naive Bayesian learning model drops the assumption that expectations are derived as Bayesian Nash equilibria from underlying prior beliefs about private pay-off characteristics. Instead, players are assumed to have arbitrarily given prior beliefs about the strategies of the other players, and to form expectations by conditioning their prior beliefs on the histories of observed actions. This model of expectation formation accommodates both far-sighted and myopic behavior. As in the sophisticated Bayesian learning model, the players choose actions in each period as best responses to their expectations, which generate the histories that the players observe over time.

In the case of myopic behavior, naive Bayesian learning accommodates all possible expectation functions. Given an arbitrary sequence of expectation functions $e_{it}(\cdot)$, one can construct a prior distribution μ_i over S^∞ using the expectations $e_{it}(h_t)$ as successive conditional distributions over $s_{-i(t+1)}$. The only restriction on the expectation functions is that player i expects players j and k to choose their next-period actions independently, that is, $s_{j(t+1)}$ and $s_{k(t+1)}$ are independently distributed under $e_{it}(h_t)$. Subject to this restriction, naive Bayesian learning formally includes all learning models in which players choose actions in each period as one-period best responses to expectations. The naive Bayesian learning model that is perhaps most in the spirit of Bayesian inference is fictitious play (e.g. Krishna and Sjöström, 1998). In fictitious play, player i expects player j's next-period action to be drawn randomly, and uses the observed frequency distribution of player j's past actions as the expected distribution.

A general model of naive Bayesian learning with far-sighted players is formulated by Kalai and Lehrer (1993a). Player i has an arbitrarily given prior belief μ_i over the behavior strategies of the other players. Given player i's pay-off characteristics (u_i,δ_i) and the expected behavior strategies $((f^i_{jt})^\infty_{t=1})_{j\neq i}$, player i chooses a best-response behavior strategy $(f^i_{it})^\infty_{t=1}$. Kalai and Lehrer do not assume that prior beliefs are common across players, so players i and j may have different expectations about the behavior strategies of player k. A behavior strategy specifies what a player would do in response to every possible history, and Nash equilibrium requires that each player have the correct expectations about the behavior strategies of all other players. In particular, this requires players i and j to have the same expectations about what player k would do in response to every possible history. Kalai and Lehrer (1993b) broaden the concept of Nash equilibrium to 'subjective equilibrium' by allowing players to have different expectations about the behavior strategies of third players for histories that occur with probability zero in equilibrium. Asymptotic expectations are more appropriately compared to subjective equilibria than Nash equilibria, since the observed history may not be sufficient to resolve disparities in expectations about behavior for all histories.

CONVERGENCE TO NASH EQUILIBRIUM

In each of the Bayesian learning models described above, players form expectations about the actions of the other players and choose their own actions in each period as best responses to their expectations. The best responses, which may be mixed, constitute the probability distribution of action sequences that the players will observe over time. In the case of sophisticated Bayesian learning, the expectations are derived as Bayesian Nash equilibria from a prior belief over the pay-off characteristics that constitute the true game. Each game generates a best-response sequence, so the prior distribution over games, together with the best-response process for each game, generates a joint distribution over games and action sequences. In this context, one can ask whether a random draw of a game and action sequence from this joint distribution will produce expectations that approach Nash equilibria of the drawn game.

In the case of naive Bayesian learning, the true game is taken as given and players' expectations are given initially rather than derived as Bayesian Nash equilibria. In this case, one can ask whether a random draw from the best-response distribution over action sequences will produce expectations

that approach Nash equilibria of the true game. One can also ask this question in the case of sophisticated Bayesian learning. That is, one can take the view, common in Bayesian statistics, that the derivation of expectations from a prior distribution is merely a heuristic device for obtaining expectations, and regard the expectations and the true game as given in the same sense as in myopic Bayesian learning.

The answers that have been obtained thus far to these and other questions are described below. For games with multiple Nash equilibria, convergence to Nash equilibrium will be understood to mean that the sequence of expectations has one or more limit points, all of which are Nash equilibria.

Sophisticated Bayesian Learning

In the model of sophisticated Bayesian learning with far-sighted behavior, with probability one, a randomly drawn game and best-response path has expectations that converge to Nash equilibrium. This was proved by Jordan (1995) under the assumptions that all players have the same prior beliefs about other players' pay-off characteristics, and that the pay-off characteristics of different players are independently distributed under the common prior. The common-prior assumption implies that any two players have the same expectations about the behavior strategies of third players, ensuring that limit points are Nash equilibria as opposed to subjective equilibria. This result is inherited under myopic behavior as a special case of far-sighted behavior with zero discount factors, but in the case of myopic behavior, Nyarko has substantially generalized the assumptions on prior beliefs. Nyarko (1998) retains the independence assumption but generalizes the common prior assumption to a condition ensuring that each player's prior belief is absolutely continuous with respect to a common distribution. The concept of prior belief is also extended to a hierarchy of beliefs, beliefs about beliefs and so on. Nyarko (1994) drops the independence assumption. In this case, a player's knowledge of his or her own pay-off characteristics may provide information about the pay-off characteristics of other players. Nyarko shows that with probability one, expectations converge to correlated equilibria.

These results imply that the set of games for which expectations do not converge to equilibria has probability zero according to the prior beliefs from which the expectations are derived. For the case of myopic behavior, a stronger assumption on prior beliefs makes it possible to derive expectations that converge for every game. Suppose that the prior beliefs, in addition to being the same for all players and satisfying the independence of pay-off characteristics across players, are 'smooth', in the sense that the prior distribution has a density function with respect to Lebesgue measure that is bounded from above and bounded from below away from zero. The uniform prior, as in the 2×2 example discussed above, is the natural example. Jordan (1991) establishes that in this case, for every game, the distribution of best-response paths generates expectations that converge to Nash equilibria of the game with probability one. Except for a set of games having Lebesgue measure zero, best-response actions are unique in every period (even when there are multiple Nash equilibria), and convergence is guaranteed along the unique best-response path. Moreover, under the same assumptions, except for a set of games having Lebesgue measure zero, expectations converge to Nash equilibria at an exponential rate (Jordan, 1992).

These results can be illustrated in the 2×2 game represented by the pay-off bimatrix in Table 2. The unique Nash equilibrium of this game is the mixed equilibrium in which the row player chooses T with probability 0.4 and the column player chooses L with probability 0.6.

Under the expectations derived from the uniform prior distribution (with pay-offs normalized as above), Table 3 shows the expectations, rounded to four decimal places, and the best-response actions for the first 12 iterations.

The table illustrates the rapid convergence of expectations to the unique Nash equilibrium. However, it also illustrates a typical disparity between expectations and actions in the case of convergence to mixed equilibrium. Player 1 is rapidly learning to predict that player 2 will choose L with probability 0.4, but in every period, player 2 actually chooses L with either probability one or probability zero. Since $e_{2(t-1)}(U \mid h_{t-1})$ is never exactly 0.6, player 2 always has a unique best-response action. This raises the question whether player 1 could eventually recognize that s_{2t} is not a random draw from the expected distribution $e_{1(t-1)}(h_{t-1})$. There is a precise sense in which the answer to this question is 'no'.

Table 2. A pay-off bimatrix representing a 2×2 game

	L	*R*
U	−2, 3	3, 0
D	0, −2	0, 0

Table 3. Expectations and best-response actions for the first 12 iterations of the game in Table 2

t	s_t	$e_{1(t-1)}(L \mid h_{t-1})$	$e_{2(t-1)}(U \mid h_{t-1})$
1	*U*, *L*	0.5000	0.5000
2	*D*, *L*	0.8192	0.8192
3	*D*, *R*	0.7624	0.1494
4	*D*, *R*	0.6514	0.3917
5	*U*, *L*	0.5808	0.4552
6	*D*, *L*	0.6203	0.4314
7	*D*, *L*	0.6037	0.4158
8	*U*, *L*	0.5944	0.4061
9	*U*, *L*	0.6000	0.4004
10	*D*, *R*	0.6022	0.3970
11	*D*, *R*	0.6013	0.3990
12	*D*, *R*	0.6007	0.3998

The actual performance of any probabilistic forecasting procedure can be evaluated by means of 'calibration tests'. For example, one can ask whether it actually rained on more than half the days for which the forecast probability of rain was greater than 50%. In the case of myopic behavior, any given calibration test comparing sophisticated Bayesian expectations with the actual sequence of best-response actions will be passed for every game with the exception of a set of games having Lebesgue measure zero. This result, due to Turdaliev (2002), assumes that the prior distribution is common, independent, and smooth.

Naive Bayesian Learning

In the naive Bayesian learning model, the true game and players' expectations of each other's actions are both arbitrary, so no convergence results are possible without further assumptions. In the case of myopic behavior, fictitious play is known to converge in zero-sum games and 2×2 games, but otherwise can fail to converge (e.g. Krishna and Sjöström, 1998).

For the more general case of far-sighted behavior, Kalai and Lehrer (1993a) prove convergence under the assumption that the distribution of best-response paths determined by the true game is absolutely continuous with respect to each player's expectations. Under this assumption, they prove that each player's expectations approach the true best-response distribution, and therefore that the players' best responses are approximately Nash. In the case of myopic behavior, this is enough to ensure that each player's expectations converge to Nash equilibria. In the more general case of far-sighted behavior, two players may continue to differ in their expectations of a third player's future response to actions off the best-response path, so that expectations are only guaranteed to converge to subjective equilibria.

The absolute-continuity assumption ensures that players' best-response strategies converge to Nash equilibrium, unlike the situation demonstrated in Table 3. If, as in Table 3, expectations converge to a mixed equilibrium while the best-response actions in each period are unique, the absolute-continuity assumption is clearly violated. If the true game is the game in Table 2, then the absolute-continuity assumption requires that, after some finite number of periods, each player's expectations are equal to the unique Nash equilibrium. Otherwise the players' best-response strategies would fail to be approximately Nash infinitely often.

The convergence of best-response strategies as well as expectations means that any calibration test comparing expectations with a randomly drawn sequence of best-response actions will be passed with probability one (Kalai *et al.*, 1999). This result applies to the general case of far-sighted behavior, for every game, provided the expectations and best-response distributions satisfy the absolute-continuity assumption.

WHAT IS BEING LEARNED?

Bayesian learning enables players to learn Nash equilibrium expectations over time without knowing the pay-off functions of the other players. It is natural to ask what players must already know in order to be capable of Bayesian learning. In all Bayesian learning models, players choose their actions as best responses to their expectations, so players must know their own pay-off functions. Naive Bayesian learning takes the players' expectations about future play paths as given, although each player is aware of the separate identities of the other players, in the sense that different players are expected to choose their actions independently conditional on past play. No additional knowledge is explicitly assumed, but the absolute-continuity assumption imposes a strong implicit relation between the players' expectations and the paths of best-response actions determined by the true pay-off functions.

Sophisticated Bayesian learning can be viewed as assuming that the players are knowledgeable game theorists who derive their expectations as Bayesian Nash equilibria from a common prior distribution over the unknown pay-off functions. In addition to having a common (or at least mutually consistent)

prior distribution, players share a common selection process for choosing among multiple Bayesian Nash equilibria. The resulting structure of expectations avoids the need for the absolute-continuity assumption, albeit at the expense of losing the convergence of best-response strategies in the case of mixed equilibria.

Alternatively, the sophisticated Bayesian expectations could be viewed as having been derived by a game theorist who does not know the players' payoff functions. If the players rely on the theorist's predictions, their faith will be vindicated by the convergence of the expectations to Nash equilibria of the true game, at least in the case of myopic behavior and the theorists' use of a prior distribution satisfying the smoothness and independence conditions. Thus the sophisticated Bayesian learning model can be viewed simply as providing a class of expectation functions that can be used by myopic players in any game to learn Nash equilibrium expectations.

References

Jordan J (1991) Bayesian learning in normal form games. *Games and Economic Behavior* **3**: 60–81.

Jordan J (1992) The exponential convergence of Bayesian learning in normal form games. *Games and Economic Behavior* **4**: 202–217.

Jordan J (1995) Bayesian learning in repeated games. *Games and Economic Behavior* **9**: 8–20.

Kalai E and Lehrer E (1993a) Rational learning leads to Nash equilibrium. *Econometrica* **61**: 1019–1045.

Kalai E and Lehrer E (1993b) Subjective equilibrium in repeated games. *Econometrica* **61**: 1231–1240.

Kalai E, Lehrer E and Smorodinsky R (1999) Calibrated forecasting and merging. *Games and Economic Behavior* **29**: 151–169.

Krishna V and Sjöström T (1998) On the convergence of fictitious play. *Mathematics of Operations Research* **23**: 479–511.

Nyarko Y (1994) Bayesian learning leads to correlated equilibria in normal form games. *Economic Theory* **4**: 821–841.

Nyarko Y (1998) Bayesian learning without common priors and convergence to Nash equilibria in normal form games. *Economic Theory* **4**: 821–841.

Turdaliev N (2002) Calibration and Bayesian learning. *Games and Economic Behavior*.

Further Reading

Fudenberg D and Levine D (1998) *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Jordan J (1997) Bayesian learning in games: a non-Bayesian perspective. In: Bicchieri C, Jeffrey R and Skyrms B (eds) *The Dynamics of Norms*, pp. 149–174. Cambridge, UK: Cambridge University Press.

Myerson R (1991) *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Bidding Strategies in Single-unit Auctions

Intermediate article

Bart Wilson, George Mason University, Fairfax, Virginia, USA

CONTENTS

Economists have developed models of auctions to enhance our understanding of one of the most commonly observed exchange institutions. Auction models are used to examine the allocation of products from an auctioneer to a number of bidders. The optimal bidding strategies and efficiency of four different auction institutions (first-price sealed bid, second-price sealed bid, English, and Dutch) can be analysed.

INTRODUCTION

Economists have developed models of auctions to enhance our understanding of one of the most commonly observed exchange institutions. Auction models are used to examine the allocation of products from an auctioneer to a number of bidders. For example, a manufacturer may procure inputs by soliciting sealed bids that name each supplier's contract price. In addition to their use in industrial procurement, auctions have been used to allocate products as varied as art, cattle, produce, government securities, and offshore mineral rights. As the internet reduces the costs of conducting auctions, this type of exchange will become still more pervasive.

AUCTION INSTITUTIONS

In this introduction to auctions and bidding, the discussion is limited to a few simplifying, yet often realistic, assumptions that will facilitate our analysis. The first is that an auctioneer is selling a single unit of a good to only one of several potential buyers, who are the bidders in the auction. The second assumption is that each bidder i knows the value v_i from consuming the item. This assumption allows each bidder to ignore any information from rivals in establishing a value estimate. The buyer's value indicates the maximum amount that the bidder is willing to pay for the item. The third assumption is that the bids b_i are statistically independent and derived from a known distribution, presumably estimated from history. Lastly, we will assume that the bidders are risk-neutral; i.e., the bidders are indifferent between playing a lottery and receiving the expected pay-off from the lottery with certainty.

The winning bidder w wins the auction, pays the price p, and receives a pay-off of $v_w - p$. All other buyers receive no pay-off. The seller's profit is $p - c$, where c is the (opportunity) cost to the seller of supplying the item. Hence, the total surplus from the exchange is $v_w - c$. The 'efficiency' of the auction – the metric by which we evaluate auction mechanisms – is defined as $(v_w - c)/(v_h - c)$, where v_h is the highest realized v_i. If the bidder with the highest value wins the auction, then the auction is perfectly efficient, or equivalently, the auction maximizes the gains from exchange.

The auctioneer chooses the institutional rules by which the bids determine who is the winning bidder and what price the winning bidder pays. As an indication of the extent to which the details of the institution matter in an auction, consider the following four different auction institutions: first-price sealed bid, second-price sealed bid, English, and Dutch auctions.

In a first-price sealed bid auction, the buyers independently submit a private sealed bid to the seller. The seller then awards the item to the highest bidder at a price equal to the highest bid. Ties are broken randomly. Note how this institution forces the buyers to simultaneously condition their bids on the other buyers' expected bids. The bidders do not use any information from the actual auction process in forming their bids.

The 'outcry' version of an English auction differs significantly from the first-price sealed bid auction in that the bidders call out increasing bids until only the highest bidder remains. Again, the highest bidder wins the auction and pays a price equal to his or her bid, but in this case the bids are conditioned on buyers' actual bids.

The second-price sealed bid auction, as its name suggests, is similar to the first-price auction in that the seller awards the item to the bidder with the highest (simultaneously submitted) bid. However, the price that the highest bidder pays is equal to the second-highest submitted bid. (In the case of a tie, one of the bidders is randomly chosen and pays the price of the other tied bidder's identical bid.)

In a Dutch auction, the auctioneer begins with a very high price which none of the bidders is willing to pay, and then in real time lowers the bid until one bidder claims the good, paying a price exactly equal to the bid called by the auctioneer.

OPTIMAL BIDDING STRATEGIES

Let us first analyze the optimal bidding strategy for the first-price sealed bid auction. For ease of exposition suppose that there are two bidders ($i = 1, 2$) whose values are independently and uniformly distributed on the interval [0, 1]. We constrain the bids to be nonnegative. Bidder i's pay-off is

$$u_i = \begin{cases} v_i - b_i & \text{if } b_i > b_j \\ \frac{(v_i - b_i)}{2} & \text{if } b_i = b_j \\ 0 & \text{if } b_i < b_j \end{cases} \quad (1)$$

A player's strategy is a function that maps values (v_i) to actions (b_i). To find the optimal bidding strategy we will assume that player 1's strategy $b_1(v_1)$ is a best response to player 2's strategy $b_2(v_2)$, and vice versa. Player 1 thus chooses the strategy b_1 to maximize the expected pay-off from the auction, or $(v_1 - b_1)\ P(b_1 > b_2(v_2))$. (Technically, we should include the pay-off $\frac{1}{2}(v_1 - b_1)\ P(b_1 = b_2(v_2))$, but since the values are continuously distributed, the probability of identical values is zero.) That is, in choosing a bid b_1, player 1 weighs the probability that b_1 will be greater than player 2's bid, which is a function of player 2's value. (*See* **Games: Coordination**)

Rasmusen (2001) provides an accessible derivation of the optimal bidding function in equilibrium, namely that a player's best-response strategy is $b_i(v_i) = \frac{v_i}{2}$. That is, each bidder submits a bid equal to one-half of the value. Notice how this optimal bidding function weighs the bidder's fundamental trade-off in the auction. The lower the bid, the more likely that the bidder will win the gain from an increased pay-off from winning the auction; but the higher the bid, the more likely that the bidder will actually win the auction. Note that the bidder with the highest value will always win the auction, so that the auction is perfectly efficient.

In contrast to a risk-neutral bidder, a risk-averse bidder will prefer to increase the probability of winning the auction at the cost of a lower pay-off by submitting a bid greater than $\frac{v_i}{2}$. In particular, it is possible that a more risk-averse bidder with a lower value will submit a bid greater than that of the bidder with the highest value who is less risk-averse. This results in an inefficient allocation if the highest bidder does not win the auction.

Now consider the English auction. If the auctioneer starts the bidding at the price of zero, every bidder is willing to purchase the item and is willing to submit a bid of some minimum increment, say 0.01. Each bidder continues to raise the bid until the standing bid exceeds his or her value. This process continues until the bidder with the second-highest value drops out of the auction at price equal to v_i (± 0.01). Hence, the bidder with the highest value wins the auction and pays a price equal to the second-highest value (± 0.01). This auction is also perfectly efficient. Furthermore, risk aversion does not change this result. Because the bidding occurs in real time, a bidder always knows whether or not he or she holds the highest bid and what is necessary to become the current highest bidder. Trying to determine the optimal bid as a function of this value is unnecessary: a bidder can always respond with a highest bid until the current highest bid exceeds his or her value. Hence, the cognitive costs of participating in this auction are much lower. (*See* **Markets, Institutions and Experiments**)

Even though it is a one-shot game, bidding incentives in a second-price auction lead to outcomes identical to those in an English auction. The optimal bidding strategy in a second-price sealed bid auction is a simple one: $b_i = v_i$. To understand why, first note that it is never profitable for a bidder i to submit a bid b_i greater than v_i, because if the second-highest bid happens to be less than b_i but greater than v_i then bidder i wins the auction but incurs a negative pay-off since the price (the second-highest bid) is greater than the value v_i. To eliminate such a possibility, the bidder should reduce his or her bid to v_i. A bidder should also never submit a bid less than the value, because the bidder gains nothing in terms of pay-off should he or she win: the price is determined by the second-highest bid, submitted by another bidder. Moreover, lowering a bid only reduces the likelihood that the bidder actually submits the highest bid and is declared winner. Hence, the optimal bid in a second-price sealed bid auction is $b_i = v_i$.

An interesting feature of this result is that bidding one's value is optimal, independently of the

other bidders' actions. A strategy that is always a best response to any strategy employed by the other players is known as a dominant strategy (see Rasmusen (2001) for a discussion). Notice that when every bidder plays the dominant strategy, the bidder with the highest value wins the auction and pays the price equal to the second-highest bid, which is equivalently the second-highest value. This outcome is identical to that of the English auction. Notice also that because setting $b_i = v_i$ is optimal regardless of what the other players do, the outcomes in this auction institution are robust to risk aversion: a bidder can never lose and only gain by playing the dominant strategy. Hence, the second-price sealed bid auction is perfectly efficient with risk-averse or risk-neutral bidders.

While seemingly quite different from the first-price sealed bid auction, the Dutch auction is strategically equivalent to it. We will call the successively lower prices offered by the auctioneer the 'bid' in a Dutch auction. When the auction begins at a very high bid, a bidder will first consider claiming the item when the bid falls to v_i. If another bidder has not already claimed the item by the time the bid reaches v_i, then the bidder must consider allowing the auctioneer to continue lowering the bid to increase the pay-off to the bidder from winning the auction, but the further the bid drops, the more likely it becomes that another bidder will actually win the auction at a price less than the bidder's own value. A bidder evaluates precisely the same trade-off in the first-price auction. Because the winner in a Dutch auction pays a price equal to his or her bid, the optimal bidding strategy, and concomitant results for efficiency and prices, in a Dutch auction are isomorphic to those in a first-price sealed bid auction.

EXPERIMENTAL RESULTS

There have been many experimental studies of bidding behavior in auctions (for a more comprehensive discussion, see Kagel and Roth (1995)). Two early studies test the strategic equivalence of first-price and Dutch auctions. Coppinger *et al*. (1980) and Cox *et al*. (1982) both find that prices are higher in first-price auctions than in Dutch auctions, and that bidding is consistent with risk-averse behavior. Cox *et al*. (1982) suggest two alternative models for the lower bidding in Dutch auctions. The first conjecture is that the bidders derive utility from the 'suspense' associated with the anticipation of purchasing at a lower and lower price, which is added to the pay-off from buying at a price less than the value for it. The second conjecture is that the real-time nature of the Dutch auction leads bidders to mistakenly update the estimates of their rivals' values to be lower when no one has taken the item at successively lower prices. Cox *et al*. (1983) test these alternative explanations and find that the probability miscalculation hypothesis cannot be rejected in favor of the suspense hypothesis. (*See* **Asset Market Experiments**; **Choice under Uncertainty**)

The predicted isomorphism between English and second-price auctions also fails to be observed. Coppinger *et al*. (1980) and Kagel *et al*. (1987) find that bidding in the English outcry auction conforms to the theoretical predictions, while in the one-shot second-price auction bidders consistently bid higher than the dominant-strategy prediction, even with experience in the auction mechanism (Kagel and Levin, 1993). One explanation for the deviation in behavior is that the real-time nature of the English auctions produces strong, immediate, and overt feedback as to what a bidder should and should not bid. The English auction makes immediately transparent the potential of a negative pay-off from bidding above one's value. The one-shot nature of the second-price auction obscures this realization. If the highest-value buyer submits a bid slightly greater than his or her value and still wins the auction, but the price (equal to the second-highest bid) is still less than that value, then the winning bidder is no worse off for having bid above the value. Since the price equals the second-highest bid, Kagel *et al*. speculate that bidders are deceived by the illusion that bidding higher is a low-cost means of increasing the probability of winning. This raises an interesting question for further research. Is this illusion based on familiarity with the standard first-price sealed bid and English auctions, where the buyer's own bid simultaneously determines the price and improves his or her chances of winning? Or does bidding one's own value present a cognitive impediment?

References

Coppinger V, Smith V and Titus J (1980) Incentives and behavior in English, Dutch and sealed-bid auctions. *Economic Inquiry* **43**: 1–22.

Cox J, Roberson B and Smith V (1982) Theory and behavior of single object auctions. In: Smith V (ed.) *Research in Experimental Economics*, vol. II, pp. 1–43. Greenwich, CT: JAI Press.

Cox J, Smith V and Walker J (1983) A test that discriminates between two models of the Dutch–first auction non-isomorphism. *Journal of Economic Behavior and Organization* **14**: 205–219.

Kagel J, Harstad R and Levin D (1987) Information impact and allocation rules in auctions with affiliated private values: a laboratory study. *Econometrica* **55**: 1275–1304.

Kagel J and Levin D (1993) Independent private value auctions: bidder behavior in first-, second-, and third-price auctions with varying numbers of bidders. *Economic Journal* **103**: 868–879.

Kagel J and Roth A (1995) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Rasmusen E (2001) *Games and Information: An Introduction to Game Theory*. Malden, MA: Blackwell.

Further Reading

Davis D and Holt C (1993) *Experimental Economics*. Princeton, NJ: Princeton University Press.

Gibbons R (1992) *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press.

Milgrom P (1989) Auctions and bidding: a primer. *Journal of Economics Perspectives* **3**(3): 3–22.

Smith V (1991) *Papers in Experimental Economics*. Cambridge, UK: Cambridge University Press.

Choice under Uncertainty

Advanced article

Mark J Machina, University of California, San Diego, California, USA

CONTENTS

The standard theory of individual choice under uncertainty consists of the joint hypothesis of expected utility risk preferences and probabilistic beliefs. Experimental work by both psychologists and economists has uncovered systematic departures from both hypotheses, and has led to the development of alternative, usually more general, models.

INTRODUCTION

Decisions under uncertainty take place in two types of settings. In settings of 'objective uncertainty', the probabilities attached to the various outcomes are specified in advance, and the objects of choice consist of 'lotteries' of the form $P = (x_1, p_1; \ldots; x_n, p_n)$, which yield outcomes or monetary pay-offs x_i with probability p_i, where $p_1 + \ldots + p_n = 1$. Examples include games of chance involving dice and roulette wheels, as well as ordinary lotteries.

In settings of 'subjective uncertainty', probabilities are not given, and the objects of choice consist of 'bets' or 'acts' $f(\cdot) = [x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n]$, which yield outcomes or pay-offs x_i in event E_i, for some mutually exclusive and exhaustive collection of events $\{E_1, \ldots, E_n\}$ which can be thought of as a partition of the set S of all possible 'states of nature'. Examples include bets on horse races or the weather, as well as standard insurance contracts.

Under objective uncertainty, choices are determined by an individual's attitudes towards risk. Under subjective uncertainty, they are additionally determined by the individual's subjective beliefs about the likelihoods of the various states and events.

EXPECTED UTILITY THEORY AND EXPERIMENTAL EVIDENCE

Axiomatic and Normative Foundations of Expected Utility Theory

The earliest formal hypothesis of individual attitudes towards risk, proposed by Pascal, Fermat and others in the seventeenth century, was that individuals evaluate monetary lotteries $P = (x_1, p_1; \ldots; x_n, p_n)$ simply on the basis of their mathematical expectation $E[P] = \sum_{i=1}^{n} x_i \cdot p_i$. This hypothesis was dramatically refuted by Daniel Bernoulli's 'St Petersburg paradox'. In this game, a fair coin is repeatedly flipped until it lands heads. If it lands heads on the first flip, the player wins \$1; if it does not land heads until the second flip, the player wins \$2; and in general, if it does not land heads until the i^{th} flip, the player wins $\$2^{i-1}$. Most people would prefer to receive a sure payment of, say, \$50 than a single play of the St Petersburg game, even though the expected pay-off of the game is $\frac{1}{2}\cdot\$1 + \frac{1}{4}\cdot\$2 + \frac{1}{8}\cdot\$4 + \ldots = \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \ldots = \$\infty$. In the first of what has turned out to be a long series of such developments, Bernoulli weakened the prevailing expected-value hypothesis by positing that individuals instead evaluate lotteries on the basis of their 'expected utility' $\sum_{i=1}^{n} U(x_i) \cdot p_i$, where the utility $U(x)$ of receiving a monetary amount x is probably subproportional to x. Bernoulli himself proposed the form $U(x) = \ln(x)$, which leads to an evaluation of the St Petersburg game consistent with typical actual play.

The expected utility hypothesis came to dominate decision theory on the twin bases of its elegant and highly normative axiomatic development (von Neumann and Morgenstern, 1944; Marschak, 1950)

and its analytical power (Arrow, 1965; Pratt, 1964). In the modern approach, risk preferences are denoted by the individual's 'weak preference' relation $\succeq$ over lotteries, where $P^* \succeq P$ reads 'P^* is weakly preferred to P', and its implied 'strict preference' relation $\succ$ (where $P^* \succ P$ iff $P^* \succeq P$ but not $P \succeq P^*$) and 'indifference' relation $\sim$ (where $P^* \sim P$ iff $P^* \succeq P$ and $P \succeq P^*$). The preference relation $\succeq$ is said to be 'represented' by an expected utility preference function $V_{\mathrm{EU}}(P) = \sum_{i=1}^{n} U(x_i) \cdot p_i$ if $P^* \succeq P \Leftrightarrow V_{\mathrm{EU}}(P^*) \geq V_{\mathrm{EU}}(P)$. $U(\cdot)$ is called the 'von Neumann–Morgenstern utility function'.

The axiomatic and normative underpinnings of expected utility theory are based on the notion of a 'probability mixture' $\alpha \cdot P + (1-\alpha) \cdot P^*$ of two lotteries $P = (x_1, p_1; \ldots; x_n, p_n)$ and $P^* = (x_1^*, p_1^*; \ldots; x_{n^*}^*, p_{n^*}^*)$, which is the lottery that would be generated by a coin flip yielding the lotteries P and P^* as prizes with respective probabilities α and $1-\alpha$, and where both stages of uncertainty (the coin flip and the resulting lottery) are realized simultaneously, so that we can write $\alpha \cdot P + (1-\alpha) \cdot P^* = (x_1, \alpha \cdot p_1; \ldots; x_n, \alpha \cdot p_n; x_1^*, (1-\alpha) \cdot p_1^*; \ldots; x_{n^*}^*, (1-\alpha) \cdot p_{n^*}^*)$. A preference relation $\succeq$ will then be represented by an expected utility preference function $V_{\mathrm{EU}}(\cdot)$ for some utility function $U(\cdot)$ if and only it satisfies the following axioms:

- *Completeness*. For all lotteries P and P^*, either $P \succeq P^*$ or $P^* \succeq P$, or both.
- *Transitivity*. For all lotteries P, P^* and P^{**}, if $P \succeq P^*$ and $P^* \succeq P^{**}$ then $P \succeq P^{**}$.
- *Mixture Continuity*. For all lotteries P, P^* and P^{**}, if $P \succ P^*$ and $P^* \succ P^{**}$ then $P^* \sim \alpha \cdot P + (1-\alpha) \cdot P^{**}$ for some $\alpha \in (0,1)$.
- *Independence Axiom*. For all lotteries P, P^* and P^{**} and all $\alpha \in (0,1)$, if $P \succeq P^*$ then $\alpha \cdot P + (1-\alpha) \cdot P^{**} \succeq \alpha \cdot P^* + (1-\alpha) \cdot P^{**}$.

Completeness and Transitivity are standard axioms in preference theory, and Mixture Continuity serves as the standard Archimedean property in the context of choice over lotteries. The key normative and behavioral axiom of the theory is the Independence axiom. Behaviorally, it corresponds to the property of separability across mutually exclusive events. Normatively, it corresponds to the following argument: 'Say you weakly prefer P to P^*, and have to choose between an α:$(1-\alpha)$ coin flip yielding P if heads and P^{**} if tails, or an α:$(1-\alpha)$ coin flip yielding P^* if heads and P^{**} if tails. Now, either the coin will land tails, in which case your choice won't have mattered, or it will land heads, in which case you are back to a choice between P and P^*, so you should weakly prefer the first coin flip to the second.'

The tension between the compelling nature of the Independence axiom and its systematic violations by experimental subjects has led to a sustained debate over the validity of the expected utility model, with some researchers continuing to posit expected utility maximization, and others developing and testing alternative models of risk preferences.

Analytics of Expected Utility Theory

Analytically, the expected utility hypothesis is characterized by the simplicity of its representation (involving the standard concepts of utility and mathematical expectation) as well as by the elegance of the correspondence between standard features of risk preferences and mathematical properties of $U(\cdot)$. The most basic of these properties is 'first-order stochastic dominance preference', which states that raising the level of some pay-off x_i in a lottery $P = (x_1, p_1; \ldots; x_n, p_n)$ – or alternatively, increasing its probability p_i at the expense of a reduction in the probability p_j of some smaller pay-off x_j – will lead to a preferred lottery. An expected utility maximizer's preferences will exhibit first-order stochastic dominance preference if and only if $U(\cdot)$ is an increasing function of x.

A second property is 'risk aversion'. Originally, this was defined as the property whereby the individual would always prefer receiving the expected value of a given lottery with certainty, rather than bearing the risk of the lottery itself. This is equivalent to the condition that the individual's 'certainty equivalent' $\mathrm{CE}(P)$ of a nondegenerate lottery $P = (x_1, p_1; \ldots; x_n, p_n)$ – that is, the value that satisfies $U(\mathrm{CE}(P)) = \sum_{i=1}^{n} U(x_i) \cdot p_i$ – is always less than the expected value of P. In modern treatments, risk aversion is defined as an aversion to all 'mean-preserving spreads' from any (degenerate or nondegenerate) lottery, where a mean-preserving spread consists of a decrease in the probability of a pay-off x_i by some amount Δp, and increases in the probabilities of some higher and lower pay-offs $x_i + \alpha$ and $x_i - \beta$ by the respective amounts $\Delta p \cdot \beta/(\alpha+\beta)$ and $\Delta p \cdot \alpha/(\alpha+\beta)$. This 'spreads' the probability mass of the lottery in a manner that does not change its expected value, so it can be thought of as a 'pure increase in risk'. An expected utility maximizer will be risk-averse in both the original and the modern senses if and only if $U(\cdot)$ is a strictly concave function of x. If $U(\cdot)$ is twice continuously differentiable, strict concavity is equivalent to a negative second derivative $U''(\cdot)$. Although the widespread purchase of actuarially unfair state lottery tickets is evidence of the opposite property of 'risk preference', the even

more widespread purchase of insurance and the prevalence of other risk-reducing instruments has led researchers to hypothesize that individuals are for the most part risk-averse.

After these basic characterizations, the most important analytical result in expected utility theory is the Arrow–Pratt characterization of 'comparative risk aversion', which states that the following four conditions on a pair of risk-averse von Neumann–Morgenstern utility functions $U_A(\cdot)$ and $U_B(\cdot)$ are equivalent:

- *Comparative Concavity*. $U_A(\cdot)$ is an increasing concave transformation of $U_B(\cdot)$, that is, $U_A(x) \equiv \rho(U_B(x))$ for some increasing concave function $\rho(\cdot)$.
- *Comparative Arrow–Pratt Measures*. $-U_A''(x)/U_A'(x) \geq -U_B''(x)/U_B'(x)$ for all x.
- *Comparative Certainty Equivalents*. For any lottery $P = (x_1, p_1; \ldots; x_n, p_n)$, if $\mathrm{CE}_A(P)$ and $\mathrm{CE}_B(P)$ satisfy $U_A(\mathrm{CE}_A(P)) = \sum_{i=1}^n U_A(x_i)\cdot p_i$ and $U_B(\mathrm{CE}_B(P)) = \sum_{i=1}^n U_B(x_i)\cdot p_i$, then $\mathrm{CE}_A(P) \leq \mathrm{CE}_B(P)$.
- *Comparative Demand for Risky Assets*. For any initial wealth W, constant $r > 0$, and random variable $\tilde{x}$ such that $E[\tilde{x}] > r$ but $P(\tilde{x} < r) > 0$, if γ_A^* and γ_B^* respectively maximize $E[U_A(\gamma\cdot\tilde{x} + (W-\gamma)\cdot r)]$ and $E[U_B(\gamma\cdot\tilde{x} + (W-\gamma)\cdot r)]$, then $\gamma_A^* \leq \gamma_B^*$.

(Note: here and elsewhere we write $P(\cdot)$ for the probability of an event. This should not be confused with the use of P to stand for a lottery.)

Each of these conditions can be interpreted as saying that $U_A(\cdot)$ is at least as risk-averse as $U_B(\cdot)$. The first condition extends the above characterization of risk aversion by the concavity of $U(\cdot)$ to its comparative version across individuals, and the second shows that this can be expressed in terms of a numerical index $-U''(x)/U'(x)$, known as the 'Arrow–Pratt index of absolute risk aversion'. The third condition extends the original notion of risk aversion as low certainty equivalents (lower than the mean) to its comparative form.

The fourth condition involves comparative optimization behavior. Consider an individual with initial wealth W to be divided between a riskless asset yielding gross return r, and a risky asset whose gross return $\tilde{x}$ has a higher expected value, but offers some risk of doing worse than the riskless asset. This condition states that the less risk-averse utility function $U_B(\cdot)$ will always choose to invest at least as much in the risky asset as will the more risk-averse $U_A(\cdot)$.

The equivalence of the above four conditions, the first two mathematical and the second two behavioral, and their numerous additional behavioral equivalencies and implications, has made the Arrow–Pratt characterization one of the central theorems in the analytics of expected utility theory, with applications in insurance, financial markets, auctions, the demand for information, bargaining, and game theory.

Experimental Evidence on the Independence Axiom

Experimental testing of the expected utility hypothesis has centered on the Independence axiom, either directly or via its implication that the expected utility preference function $V_{\mathrm{EU}}(P) = \sum_{i=1}^n U(x_i)\cdot p_i$ is linear in the probabilities p_i. One of the best-known tests is the 'Allais paradox' (Allais, 1953). An individual is asked to rank each of the following two pairs of lotteries (where \$1M = \$1,000,000):

- $a_1 = \{1.00$ chance of \$1M versus
  $a_2 = \begin{cases} .10 \text{ chance of } \$5\text{M} \\ .89 \text{ chance of } \$1\text{M} \\ .01 \text{ chance of } \$0 \end{cases}$
- $a_3 = \begin{cases} .10 \text{ chance of } \$5\text{M} \\ .90 \text{ chance of } \$0 \end{cases}$ versus
  $a_4 = \begin{cases} .11 \text{ chance of } \$1\text{M} \\ .89 \text{ chance of } \$0 \end{cases}$

The expected utility hypothesis implies that the individual's choices from these two pairs must either be a_1 and a_4 (whenever $.11\cdot U(\$1\text{M}) > .10\cdot U(\$5\text{M}) + .01\cdot U(\$0)$), or else a_2 and a_3 (whenever $.11\cdot U(\$1\text{M}) < .10\cdot U(\$5\text{M}) + .01\cdot U(\$0)$). However, when presented with these choices, most subjects choose a_1 from the first pair and a_3 from the second, which violates the hypothesis. Only a small number violate the hypothesis in the opposite direction, by choosing a_2 and a_4.

Although the Allais paradox was originally dismissed as an 'isolated example', subsequent experimental work by psychologists, economists and others has uncovered a similar pattern of violation over a range of probability and pay-off values, and the Allais paradox is now seen to be just one example of a type of systematic violation of the Independence axiom known as the 'common-consequence effect'. It is observed that for lotteries P, P^* and P^{**}, pay-off c, and mixture probability $\alpha \in (0,1)$, such that P^{**} first-order-stochastically dominates P^* and c lies between the highest and lowest pay-offs in P, preferences depart from the Independence axiom in the direction of exhibiting $\alpha\cdot P + (1-\alpha)\cdot P^* \succ \alpha\cdot c + (1-\alpha)\cdot P^*$ yet $\alpha\cdot P + (1-\alpha)\cdot P^{**} \prec \alpha\cdot c + (1-\alpha)\cdot P^{**}$. (In the Allais paradox, these constructs are $P = (\$5\text{M}, 10/11; \$0, 1/11)$, $P^* = (\$0, 1), P^{**} = (\$1\text{M}, 1), c = \$1\text{M}$ and $\alpha = .11$.)

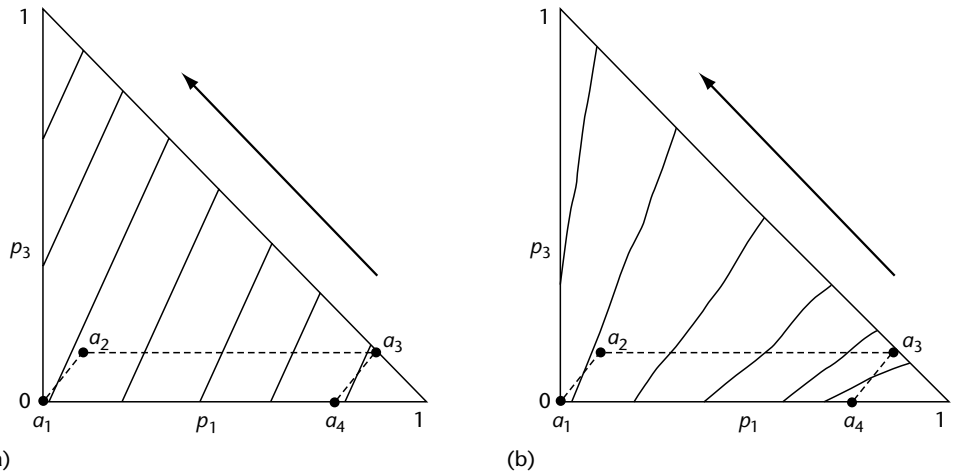


Figure 1. Indifference curves in the probability triangle. (a) Expected utility indifference curves, which are parallel straight lines. (b) Non-expected utility indifference curves, which 'fan out', illustrating the common-consequence effect.

Both the implications of the Independence axiom and the nature of this violation can be illustrated in the special case of all lotteries $P = (\bar{x}_1, p_1; \bar{x}_2, p_2; \bar{x}_3, p_3)$ over a triple of fixed pay-off values $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$. Since we can write $P = (\bar{x}_1, p_1; \bar{x}_2, 1 - p_1 - p_3; \bar{x}_3, p_3)$, each such lottery is uniquely associated with a point in the (p_1, p_3) triangles of Figures 1(a) and 1(b). Since we can write $V_{\mathrm{EU}}(P) = U(\bar{x}_1)\cdot p_1 + U(\bar{x}_2)\cdot(1 - p_1 - p_3) + U(\bar{x}_3)\cdot p_3$, the loci of constant expected utility ('expected utility indifference curves') consist of parallel straight lines as in Figure 1(a). Since upward shifts in the triangle represent increases in p_3 at the expense of p_2, and leftward shifts represent reductions in p_1 to the benefit of p_2, first-order stochastic dominance preference implies that these indifference curves will be upward-sloping, with increasing levels of preference in the direction indicated by the arrows.

Fixing the pay-offs at $\bar{x}_1 = \$0$, $\bar{x}_2 = \$1\mathrm{M}$ and $\bar{x}_3 = \$5\mathrm{M}$, the Allais paradox lotteries a_1, a_2, a_3 and a_4 are seen to form a parallelogram when plotted in the probability triangle, which explains why parallel straight-line expected utility indifference curves must either prefer a_1 and a_4 (as illustrated for the relatively steep indifference curves of Figure 1(a)) or else prefer a_2 and a_3 (for relatively flat expected utility indifference curves). Figure 1(b) illustrates 'non-expected utility indifference curves' that 'fan out', and are seen to exhibit the typical Allais paradox rankings of $a_1 \succ a_2$ and $a_3 \succ a_4$.

Another type of systematic experimental violation of the Independence axiom that has been uncovered is known as the 'common-ratio effect'. For pay-offs $x^* > x > 0$, probabilities $p^* < p$ and $r \in (0, 1)$, preferences depart from the Independence axiom in the direction of exhibiting $(x^*, p^*; 0, 1 - p^*) \prec (x, p; 0, 1 - p)$ yet $(x^*, r\cdot p^*; 0, 1 - r\cdot p^*) \succ (x, r\cdot p; 0, 1 - r\cdot p)$. For losses $0 > -x > -x^*$, with $p^* < p$ and $r \in (0, 1)$, preferences depart in the reflected direction of $(-x^*, p^*; 0, 1 - p^*) \succ (-x, p; 0, 1 - p)$ yet $(-x^*, r\cdot p^*; 0, 1 - r\cdot p^*) \prec (-x, r\cdot p; 0, 1 - r\cdot p)$.

With the pay-offs $\bar{x}_1 = 0$, $\bar{x}_2 = x$ and $\bar{x}_3 = x^*$, the line segment between the lotteries $b_1 = (x^*, p^*; 0, 1 - p^*)$ and $b_2 = (x, p; 0, 1 - p)$ in the probability triangle of Figure 2(a) is seen to be parallel to that between $b_3 = (x^*, r\cdot p^*; 0, 1 - r\cdot p^*)$ and $b_4 = (x, r\cdot p; 0, 1 - r\cdot p)$, and the common-ratio-effect rankings of $b_1 \prec b_2$ and $b_3 \succ b_4$ again suggests that indifference curves depart from expected utility by fanning out. For losses, with $\bar{x}_1 = -x^*$, $\bar{x}_2 = -x$ and $\bar{x}_3 = 0$ (to maintain the ordering $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$), the reflected rankings of $-b_1 \succ -b_2$ and $-b_3 \prec -b_4$ again suggest fanning out, as in Figure 2(b). Fanning out is consistent with other observed forms of departure from the Independence axiom, although it is not universal across subjects, and seems to be more pronounced near the edges of the triangle than in its central region.

GENERALIZATIONS OF EXPECTED UTILITY THEORY

Non-Expected Utility Functional Forms

The above phenomena, as well as other systematic departures from linearity in the probabilities, have prompted researchers to develop more general models of preferences over lotteries, primarily by generalizing the functional form of the lottery preference function $V(P) = V(x_1, p_1; \ldots; x_n, p_n)$. The earliest of these attempts, which used the form $V(P) = \sum_{i=1}^{n} U(x_i)\cdot\pi(p_i)$, was largely abandoned

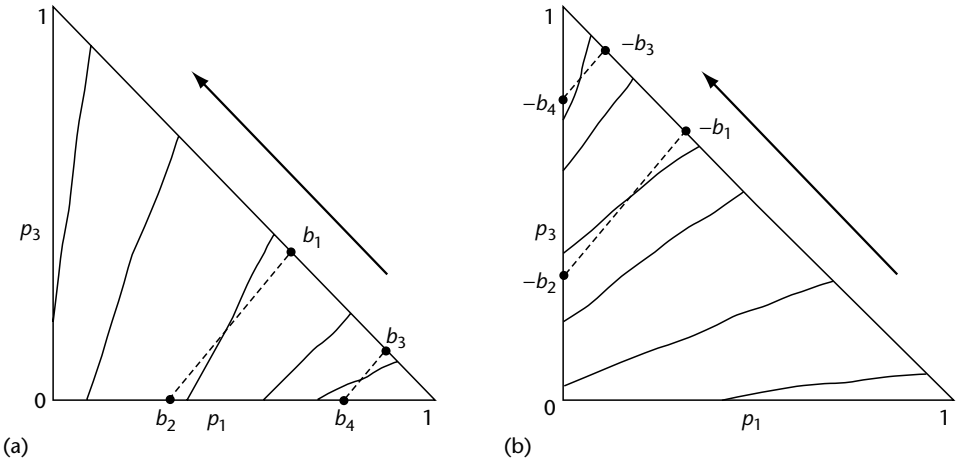


Figure 2. Probability triangles illustrating the common-ratio effect. (a) Positive pay-offs. (b) Negative pay-offs (losses).

when it was realized that, except for the case $\pi(p) \equiv p$ when it reduced to expected utility, it was inconsistent with the property of first-order stochastic dominance preference. Current models include the following:

- *Weighted Utility*. $V(P) = \sum_{i=1}^{n} U(x_i)\cdot\pi(p_i)/\sum_{i=1}^{n} S(x_i)\cdot\pi(p_i)$
- *Moments of Utility*. $V(P) = F(\sum_{i=1}^{n} U(x_i)\cdot p_i, \sum_{i=1}^{n} U(x_i)^2\cdot p_i, \sum_{i=1}^{n} U(x_i)^3\cdot p_i)$
- *Rank-Dependent Expected Utility*. $V(P) = \sum_{i=1}^{n} U(x_i)\cdot(G(\sum_{j=1}^{i} p_j) - G(\sum_{j=1}^{i-1} p_j))$for $x_1 \preceq \ldots \preceq x_n$
- *Quadratic in the Probabilities*. $V(P) = \sum_{i=1}^{n}\sum_{j=1}^{n} T(x_i, x_j)\cdot p_i\cdot p_j$

Under the appropriate monotonicity or curvature assumptions on their constituent functions $U(\cdot)$, $\pi(\cdot)$, $G(\cdot)$, etc., each of these forms is capable of exhibiting first-order stochastic dominance preference, risk aversion and comparative risk aversion, as well as many of the types of observed systematic violations of the Independence axiom. Researchers have also used these forms to revisit many of the applications previously modeled by expected utility theory (e.g. insurance, financial markets, auctions), to determine which of the earlier expected utility-based results are crucially dependent on preferences exhibiting the expected utility functional form, and which are robust to departures from expected utility.

Generalized Expected Utility Analysis

An alternative branch of research on non-expected utility preferences does not rely on any specific functional form, but links properties of attitudes towards risk directly with the probability derivatives of a general (i.e. not necessarily expected utility) preference function $V(P) = V(x_1, p_1; \ldots; x_n, p_n)$ over lotteries. Such analysis reveals that the basic analytics of the expected utility model as outlined above are in fact quite robust to general smooth departures from linearity in the probabilities. It proceeds from the correspondence between the properties of a linear function as determined by its coefficients and the properties of a nonlinear function as determined by its partial derivatives – in this case, between the 'probability coefficients' $U(x_1), \ldots, U(x_n)$ of the expected utility form $\sum_{i=1}^{n} U(x_i)\cdot p_i$ and the 'probability derivatives' $\partial V(x_1, p_1; \ldots; x_n, p_n)/\partial p_1, \ldots, \partial V(x_1, p_1; \ldots; x_n, p_n)/\partial p_n$ of a general smooth preference function $V(x_1, p_1; \ldots; x_n, p_n)$. Under such a correspondence, most of the fundamental analytical results of expected utility theory pass through directly (Machina, 1982). For example:

- *First-Order Stochastic Dominance Preference*. Under expected utility, this is equivalent to $U(x)$ (the coefficient of $P(x)$) being an increasing function of x. For a general smooth $V(\cdot)$, if $\partial V(P)/\partial P(x)$ is an increasing function of x at every lottery P, then for any pay-offs $x_i > x_j$ we will have $\partial V(P)/\partial p_i > \partial V(P)/\partial p_j$ at each P, so any (small or large) rise in p_i and matching fall in p_j will lead to an increase in $V(P)$ and hence will be preferred.
- *Risk Aversion*. Under expected utility, this is equivalent to $U(x)$ being a strictly concave function of x. For a general smooth $V(\cdot)$, if $\partial V(P)/\partial P(x)$ is a strictly concave function of x at each P, then for any pay-offs $x_i - \beta < x_i < x_i + \alpha$ we will have $[\partial V(P)/\partial P(x_i) - \partial V(P)/\partial P(x_i - \beta)]/\beta > [\partial V(P)/\partial P(x_i + \alpha) - \partial V(P)/\partial P(x_i)]/\alpha$ at each P, which implies that each mean-preserving spread over the pay-offs $x_i - \beta < x_i < x_i + \alpha$ will lead to a reduction in $V(P)$ and hence will be dispreferred.
- *Comparative Risk Aversion*. Under expected utility, this is equivalent to $U_A(\cdot)$ being an increasing concave transformation of $U_B(\cdot)$. For general smooth $V_A(\cdot)$ and

$V_B(\cdot)$, if at each P the function $\partial V_A(P)/\partial P(x)$ is some increasing concave transformation of $\partial V_B(P)/\partial P(x)$, then $V_A(\cdot)$ and $V_B(\cdot)$ will exhibit the above conditions for comparative certainty equivalence and comparative demand for risky assets.

In addition to the above theoretical results, this approach also allows for a direct characterization of the fanning-out property in terms of how the probability derivative $\partial V(P)/\partial P(x)$, treated as a function of x, varies with the lottery P. Namely, the indifference curves of a preference function $V(\cdot)$ will fan out for all pay-offs $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$ if and only if $\partial V(P^*)/\partial P(x)$ is a concave transformation of $\partial V(P)/\partial P(x)$ whenever P^* first-order-stochastically dominates P.

Regret Theory

Another type of non-expected utility model dispenses with the assumption of an underlying preference order $\succeq$ over lotteries, and instead derives choice behavior from the underlying psychological notion of 'regret' – that is, the reaction to receiving an outcome x when an alternative decision would have led to a preferred outcome x^* (Loomes and Sugden, 1982). The opposite experience, namely of receiving an outcome that is preferred to what the alternative decision would have yielded, is termed 'rejoice'. The primitive for this model is a 'rejoice function' $R(x, x^*)$ which is positive if x is preferred to x^*, negative if x^* is preferred to x, and zero if they are indifferent, and satisfies the skew-symmetry condition $R(x, x^*) \equiv -R(x^*, x)$.

In the simplest case of pairwise choice over two lotteries $P = (x_1, p_1; \ldots; x_n, p_n)$ and $P^* = (x_1^*, p_1^*; \ldots; x_{n^*}^*, p_{n^*}^*)$ that are realized independently, the individual's expected rejoice from choosing the lottery P over the alternative lottery P^* is given by the formula $\sum_{i=1}^{n}\sum_{j=1}^{n^*} R(x_i, x_j^*)\cdot p_i \cdot p_j^*$, and the individual is predicted to choose P if this value is positive, to choose P^* if it is negative, and to be indifferent if it is zero. Various proposals for extending this approach beyond pairwise choice have been made, including a formal result that shows that for any finite collection of lotteries, one of these lotteries or some randomization over them will exhibit nonnegative expected rejoice with respect to every other lottery or randomization.

As with the non-expected utility functional forms listed above, various monotonicity and curvature assumptions on the rejoice function $R(\cdot,\cdot)$ can be shown to correspond to various properties of risk preferences, such as risk aversion and comparative risk aversion, as well as to the general fanning-out property. Since this model derives from the pairwise comparison of lotteries rather than from their individual evaluation by some preference function, it allows pairwise choice to be intransitive, so that an individual could choose P over P^*, P^* over P^{**}, and P^{**} over P. Although some have argued that such cyclic choice allows for the phenomenon of 'money pumps', it also allows the model to solve the problem of 'preference reversal' described below.

SUBJECTIVE EXPECTED UTILITY AND AMBIGUITY

Axiomatic and Normative Foundations of Subjective Expected Utility

The expected utility model of choice under subjective uncertainty – sometimes called the 'subjective expected utility' model – hypothesizes that the individual's preference relation $\succeq$ over subjective acts $f(\cdot) = [x_1 \text{ on } E_1; \ldots; x_n \text{ on } \mathrm{E_n}]$ can be represented by a preference function of the form $W_{\mathrm{SEU}}(f(\cdot)) = \sum_{i=1}^{n} U(x_i)\cdot\mu(E_i)$ for some von Neumann–Morgenstern utility function $U(\cdot)$ and 'subjective probability measure' $\mu(\cdot)$ over events. Thus, both attitudes towards risk and subjective beliefs are specific to the individual, and the values $\mu(E_1), \ldots, \mu(E_n)$ are sometimes called 'personal probabilities'. By virtue of its independent representation of risk attitudes by the utility function $U(\cdot)$, and beliefs by the subjective probability measure $\mu(\cdot)$, the subjective expected utility model is sometimes described as achieving a 'separation of preferences and beliefs'.

By analogy with the probability mixture of two objective lotteries, the axiomatic and normative underpinnings of the subjective expected utility model are based on the notion of a 'subjective mixture' $[f(\cdot) \text{ on } E; f^*(\cdot) \text{ on } \sim E]$ of two acts $f(\cdot) = [x_i \text{ on } E_1; \ldots; x_n \text{ on } E_n]$ and $f^*(\cdot) = [x_1^* \text{ on } E_1^*; \ldots; x_{n^*}^* \text{ on } E_{n^*}^*]$, which is the act that would yield the same outcome as $f(\cdot)$ should the event E occur, and the same outcome as $f^*(\cdot)$ should the event $\sim E$ occur, so that we can write $[f(\cdot) \text{ on } E; f^*(\cdot) \text{ on } \sim E] = [x_1 \text{ on } E \cap E_1; \ldots; x_n \text{ on } E \cap E_n; x_1^* \text{ on } \sim E \cap E_1^*; \ldots; x_{n^*}^* \text{ on } \sim E \cap E_{n^*}^*]$. An event E is said to be 'null' for the individual if $[x^* \text{ on } E; f(\cdot) \text{ on } \sim E] \sim [x \text{ on } E; f(\cdot) \text{ on } \sim E]$ for all outcomes x and x^* and all acts $f(\cdot)$, so that the individual effectively treats E as if it had zero likelihood. Since we can identify each outcome x with the 'constant act' $[x \text{ on S}]$, we can write $x^* \succeq x$ if and only if $[x^* \text{ on S}] \succeq [x \text{ on S}]$. The individual's preferences over subjective acts can then be represented by a subjective expected utility preference function

$W_{\text{SEU}}(f(\cdot)) = \sum_{i=1}^{n} U(x_i) \cdot \mu(E_i)$ for some $U(\cdot)$ and $\mu(\cdot)$ if and only they satisfy the following axioms (Savage, 1954):

- *Completeness*. For all acts $f(\cdot)$ and $f^*(\cdot)$, either $f(\cdot) \succeq f^*(\cdot)$ or $f^*(\cdot) \succeq f(\cdot)$, or both.
- *Transitivity*. For all acts $f(\cdot)$, $f^*(\cdot)$ and $f^{**}(\cdot)$, if $f(\cdot) \succeq f^*(\cdot)$ and $f^*(\cdot) \succeq f^{**}(\cdot)$ then $f(\cdot) \succeq f^{**}(\cdot)$.
- *Eventwise Monotonicity*. For all outcomes x^* and x, non-null events E and acts $f(\cdot)$, $[x^* \text{ on } E; f(\cdot) \text{ on } \sim E] \succeq [x \text{ on } E; f(\cdot) \text{ on } \sim E]$ if and only if $x^* \succeq x$.
- *Weak Comparative Probability*. For all events A and B and outcomes $x^* \succ x$ and $y^* \succ y$, $[x^* \text{ on } A; x \text{ on } \sim A] \succeq [x^* \text{ on } B; x \text{ on } \sim B]$ implies $[y^* \text{ on } A; y \text{ on } \sim A] \succeq [y^* \text{ on } B; y \text{ on } \sim B]$.
- *Small-Event Continuity*. For all acts $f(\cdot) \succ g(\cdot)$ and outcomes x, there exists a partition $\{E_1,\ldots,E_n\}$ such that $f(\cdot) \succ [x \text{ on } E_i; g(\cdot) \text{ on } \sim E_i]$ and $[x \text{ on } E_i; f(\cdot) \text{ on } \sim E_i] \succ g(\cdot)$ for each $i = 1,\ldots,n$.
- *Sure-Thing Principle*. For all events E and acts $f(\cdot)$, $f^*(\cdot)$, $g(\cdot)$ and $h(\cdot)$, $[f^*(\cdot) \text{ on } E; g(\cdot) \text{ on } \sim E] \succeq [f(\cdot) \text{ on } E; g(\cdot) \text{ on } \sim E]$ implies $[f^*(\cdot) \text{ on } E; h(\cdot) \text{ on } \sim E] \succeq [f(\cdot) \text{ on } E; h(\cdot) \text{ on } \sim E]$.

Completeness and Transitivity are as before, and Eventwise Monotonicity is the subjective analogue of first-order stochastic dominance preference. Weak Comparative Probability essentially ensures that the individual's 'revealed likelihood ranking' of a pair of events A and B, as given by their preference for staking the more preferred of a pair of prizes on A versus staking it on B, is stable in the sense that it does not depend on the particular prizes involved. Small-Event Continuity serves as the standard Archimedean property in the context of choice over subjective acts. The key normative and behavioral axiom of subjective expected utility theory is the Sure-Thing Principle. Behaviorally, it once again corresponds to the property of separability across mutually exclusive events. Normatively, it corresponds to the same argument as for the Independence axiom, with the objective randomization by the α:$(1-\alpha)$ coin replaced by the 'subjective randomization' via the events E and $\sim E$.

State-dependent Utility

In some subjective settings, the individual's valuation of outcomes may depend on the source of uncertainty itself. Thus, for the mutually exclusive and exhaustive events ('rain','shine') and prizes ('umbrella', 'sun lotion'), each of which is preferred to \$0, the individual may well exhibit the preferences [umbrella on rain; \$0 on shine] $\succ$ [umbrella on shine; \$0 on rain] and [sun lotion on rain; \$0 on shine] $\prec$ [sun lotion on shine; \$0 on rain], which violates the Weak Comparative Probability axiom for $x^* = \text{umbrella}$, $y^* = \text{lotion}$, $x = y = \$0$, $A = \text{rain}$ and $B = \text{shine}$, and hence is inconsistent with the subjective expected utility preference function $W_{\text{SEU}}(\cdot)$. This phenomenon, known as 'state dependence', can occur even when the outcomes are monetary pay-offs: if the state of nature is the individual's health, the utility of a \$50,000 prize may be very high in states where the individual requires a \$50,000 operation to survive, much lower in states where the individual requires much more than that for the operation, and somewhere in between in states of good health.

The subjective expected utility model can be easily adapted to accommodate the phenomenon of state dependence, by allowing the utility function $U(\cdot|E)$ to depend upon the event or state of nature, so that the preference function over acts takes the 'state-dependent expected utility' form $W_{\text{SDEU}}(x_1 \text{ on } E_1; \ldots x_n \text{ on } E_n) = \sum_{i=1}^{n} U(x_i|E_i) \cdot \mu(E_i)$. Most of the analytics of the standard (i.e. 'state-independent') form $W_{\text{SEU}}(\cdot)$ extend to the state-dependent case (Karni, 1985). However, under state dependence, subjective probabilities cannot be uniquely inferred from preferences over acts: for any state-dependent preference function $W_{\text{SDEU}}(f(\cdot)) = \sum_{i=1}^{n} U(x_i|E_i) \cdot \mu(E_i)$, and any distinct subjective probability measure $\mu^*(\cdot)$ that satisfies $\mu^*(E) > 0 \Leftrightarrow \mu(E) > 0$, $W_{\text{SDEU}}(\cdot)$ is indistinguishable from the preference function $W^*_{\text{SDEU}}(f(\cdot)) = \sum_{i=1}^{n} U^*(x_i|E_i) \cdot \mu^*(E_i)$ with $U^*(\cdot|\cdot)$ defined by $U^*(x|E) = U(x|E) \cdot [\mu(E)/\mu^*(E)]$.

Ambiguity and Nonprobabilistic Beliefs

A more serious departure from the notion of well-defined probabilistic beliefs arises from the phenomenon of 'ambiguity', which is distinct from the phenomenon of state dependence and much more difficult to model. The best-known example is the 'Ellsberg paradox' (Ellsberg, 1961). An individual must draw a ball from an opaque urn that contains 30 red balls and 60 black or yellow balls in an unknown proportion, and is offered four possible bets on the color of the drawn ball, as shown in Figure 3.

Most individuals exhibit the preference rankings $f_1(\cdot) \succ f_2(\cdot)$ and $f_4(\cdot) \succ f_3(\cdot)$. When asked why, they explain that the probability of winning under $f_2(\cdot)$ could be anywhere from 0 to $\frac{2}{3}$ whereas the probability of winning under $f_1(\cdot)$ is known to be exactly $\frac{1}{3}$, and they prefer the act that offers the known probability. Similarly, the probability of winning under $f_3(\cdot)$ could be anywhere from $\frac{1}{3}$ to 1 whereas the probability of winning under $f_4(\cdot)$ is known to be exactly $\frac{2}{3}$, so it is preferred. However, these preferences are inconsistent with any assignment of

	30 balls	60 balls	
	Red	Black	Yellow
$f_1(\cdot)$	\$100	\$0	\$0
$f_2(\cdot)$	\$0	\$100	\$0
$f_3(\cdot)$	\$100	\$0	\$100
$f_4(\cdot)$	\$0	\$100	\$100

Figure 3. Four possible bets on the color of the drawn ball in the 'Ellsberg paradox'. The proportion of black to yellow balls is unknown. Most individuals prefer $f_1(\cdot)$ to $f_2(\cdot)$ and $f_4(\cdot)$ to $f_3(\cdot)$.

numerical subjective probabilities μ(red), μ(black), μ(yellow) to the three events: if the individual were choosing on the basis of such probabilistic beliefs, the ranking $f_1(\cdot) \succ f_2(\cdot)$ would reveal that $\mu(\text{red}) > \mu(\text{black})$, but the ranking $f_4(\cdot) \succ f_3(\cdot)$ would reveal that $\mu(\text{red}) < \mu(\text{black})$.

This phenomenon be cannot be accommodated by simply allowing the event to enter the utility function and working with the state-dependent form $\sum_{i=1}^{n} U(x_i|E_i)\cdot\mu(E_i)$, since this form still satisfies the Sure-Thing Principle, whereas the preferences $f_1(\cdot) \succ f_2(\cdot)$ and $f_4(\cdot) \succ f_3(\cdot)$ violate this axiom (for $E = \text{red} \cup \text{black}$ and $\sim E = \text{yellow}$). The Ellsberg paradox and related examples are attributed to the phenomenon of 'ambiguity aversion', whereby individuals exhibit a general preference for bets based on probabilistic partitions such as {red, black ∪ yellow} rather than on ambiguous partitions such as {black, red ∪ yellow}.

Just as the Allais paradox and related violations of the Independence axiom led to the development of non-expected utility models of risk preferences, the Ellsberg paradox and related examples have led to the development of nonprobabilistic models of beliefs. The most notable of these involves replacing the additive subjective probability measure $\mu(\cdot)$ over events by a 'capacity' $C(\cdot)$, which is similar to $\mu(\cdot)$ in that it satisfies the properties $C(\emptyset) = 0$, $C(S) = 1$, and $E \subseteq E^* \Rightarrow C(E) \leq C(E^*)$, but differs from $\mu(\cdot)$ in that it is not necessarily additive. By labeling the outcomes in any act so that $x_1 \preceq \ldots \preceq x_n$ and writing the subjective expected utility formula as $W_{\text{SEU}}(f(\cdot)) = \sum_{i=1}^{n} U(x_i)\cdot\mu(E_i) = \sum_{i=1}^{n} U(x_i)\cdot(\mu(\cup_{j=1}^{i} E_j) - \mu(\cup_{j=1}^{i-1} E_j))$, we can generalize from an additive $\mu(\cdot)$ to a non-additive $C(\cdot)$ to obtain the 'Choquet expected utility' preference function $W_{\text{Choquet}}(f(\cdot)) = \sum_{i=1}^{n} U(x_i)\cdot(C(\cup_{j=1}^{i} E_j) - C(\cup_{j=1}^{i-1} E_j))$ over subjective acts (Schmeidler, 1989). Selecting $U(\$100) = 1$, $U(\$0) = 0$, $C(\text{red}) = \frac{1}{3}$, $C(\text{black} \cup \text{yellow}) = \frac{2}{3}$, $C(\text{black}) = \frac{1}{2}$, $C(\text{red} \cup \text{yellow}) = \frac{3}{4}$ yields the values $W_{\text{Choquet}}(f_1(\cdot)) = \frac{1}{3}$, $W_{\text{Choquet}}(f_2(\cdot)) = \frac{1}{4}$, $W_{\text{Choquet}}(f_3(\cdot)) = \frac{1}{2}$, $W_{\text{Choquet}}(f_4(\cdot)) = \frac{2}{3}$, which correspond to the typical Ellsberg rankings.

Another alternative to the subjective expected utility model of act preferences, also capable of exhibiting ambiguity aversion, is the 'maxmin expected utility' form, which involves a family $\{\mu_\tau(\cdot)\,|\tau \in T\}$ of additive probability measures over the events, and the preference function $W_{\max\min}(f(\cdot)) = \min_{\tau\in T} \sum_{i=1}^{n} U(x_i)\cdot\mu_\tau(E_i)$.

DESCRIPTION AND PROCEDURE INVARIANCE

Although the alternative models described above drop or weaken many of the axioms of standard objective and subjective expected utility theory, they typically retain the primary implicit assumptions of the standard theory, namely that: the objects of choice (objective lotteries or subjective acts) can be unambiguously described; net changes in wealth are combined with any initial endowment and evaluated in terms of the final wealth levels they imply; and situations that imply the same set of final opportunities (the same set of objective lotteries or same set of subjective acts over final wealth levels) will lead to the same choice. They also assume that the individual is able to perform the mathematical operations necessary to determine this opportunity set, e.g. to calculate the probabilities of compound or conditional events and add net changes to initial endowments. However, psychologists have uncovered several systematic violations of these assumptions.

Framing Effects

Effects whereby alternative descriptions of the same decision problem lead to systematically different responses are called 'framing' effects. Some framing effects in choice under uncertainty involve alternative representations of the same likelihood. For example, contingency of a gain or loss on the joint occurrence of four independent events, each with probability p, is found to elicit a different response from contingency on the occurrence of a single event with probability p^4. In comparison with the single-event case, making a gain contingent on the joint occurrence of events seems to make it more attractive, and making a loss

contingent on the joint occurrence of events seems to make it more unattractive (Slovic, 1969).

Other framing effects in choice under uncertainty involve alternative representations of the same final wealth levels. Consider the following two proposals.

- 'In addition to whatever you own, you have been given 1,000 (Israeli pounds). You are now asked to choose between a ½:½ chance of a gain of 1,000 or 0 or a sure chance of a gain of 500.'
- 'In addition to whatever you own, you have been given 2,000. You are now asked to choose between a ½:½ chance of a loss of 1,000 or 0 or a sure loss of 500.'

These two problems involve identical distributions over final wealth. However, when put to two different groups of subjects, 84% chose the sure gain in the first problem but 69% chose the ½:½ gamble in the second (Kahneman and Tversky, 1979).

Response-Mode Effects and Preference Reversal

Effects whereby alternative response formats lead to systematically different inferences about underlying preferences are called 'response-mode' effects. For example, under the expected utility hypothesis, an individual's von Neumann–Morgenstern utility function can be assessed or elicited in a number of different ways, which typically involve a sequence of prespecified lotteries P_1, P_2, P_3,…, and ask for the individual's certainty equivalent $\mathrm{CE}(P_i)$ for each lottery P_i, or else the 'gain equivalent' G_i that would make the lottery $(G_i, \frac{1}{2};\ \$0, \frac{1}{2})$ of equal preference to P_i, or else the 'probability equivalent' $\wp_i$ that would make the lottery $(\$1000, \wp_i;\ \$0, 1-\wp_i)$ of equal preference to P_i. Although such procedures should be expected to generate equivalent assessed utility functions, they have been found to yield systematically different ones (Hershey and Schoemaker, 1985).

In an experiment that demonstrates what is now known as the 'preference reversal phenomenon', subjects were first presented with a number of pairs of bets and asked to choose one bet out of each pair. Each pair of bets took the form of a 'p-bet', which offered a p chance of $\$X^*$ and a $1-p$ chance of \$X, versus a '\$-bet', which offered a q chance of $\$Y^*$ and a $1-q$ chance of \$Y, where $X^* > X, Y^* > Y, p > q$ and $X^* < Y^*$. The names 'p-bet' and '\$-bet' derive from the greater probability of winning in the first bet, and greater possible gain in the second bet (in some cases, X and Y took on small negative values). Subjects were next asked for their certainty equivalents of each of these bets, via a number of standard elicitation techniques.

The expected utility model, and most of the aforementioned alternative models, predict that for each such pair, the bet that was selected in the direct-choice problem would also be the one assigned the higher certainty equivalent. However, subjects exhibit a systematic departure from this prediction in the direction of choosing the p-bet in a direct choice but assigning a higher certainty equivalent to the \$-bet (Lichtenstein and Slovic, 1971). Although this finding initially generated widespread scepticism, especially among economists, it has been widely replicated by both psychologists and economists in a variety of different settings involving real-money gambles, patrons of a Las Vegas casino, group decisions, and experimental market trading. By viewing it as an instance of intransitivity ($\$\text{-bet} \sim \mathrm{CE}(\$\text{-bet}) \succ \mathrm{CE}(p\text{-bet}) \sim p\text{-bet} \succ \$\text{-bet}$), some economists have explained the phenomenon in terms of the regret theory model. However, most psychologists and a growing number of economists regard it as a response-mode effect, specifically, that the psychological processes of valuation (which generates certainty equivalents) and choice are differentially influenced by the probabilities and pay-offs involved in a lottery, and that under certain conditions this can lead to choice and valuation that reveal opposite 'underlying' preference rankings over a pair of gambles.

SUMMARY

Since the work of Bernoulli, the theory of choice under uncertainty has seen both a tension and a scientific interplay between theoretical models of decision making and experimentally observed violations of these models. Current research in the field continues to reflect this tension, while the degree of interplay has increased, with theorists now more willing to model experimentally generated phenomena, and experimenters more willing to provide constructive feedback on these attempts.

References

Allais M (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica* **21**: 503–546.

Arrow K (1965) *Aspects of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnsson Säätiö.

Ellsberg D (1961) Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* **75**: 643–669.

Hershey J and Schoemaker P (1985) Probability versus certainty equivalence methods in utility measurement: are they equivalent? *Management Science* **31**: 1213–1231.

Kahneman D and Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* **47**: 263–291.
Karni E (1985) *Decision Making Under Uncertainty: The Case of State Dependent Preferences*. Cambridge, MA: Harvard University Press.
Lichtenstein S and Slovic P (1971) Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* **89**: 46–55.
Loomes G and Sugden R (1982) Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal* **92**: 805–824.
Machina M (1982) 'Expected utility' analysis without the Independence Axiom. *Econometrica* **50**: 277–323.
Marschak J (1950) Rational behavior, uncertain prospects, and measurable utility. *Econometrica* **18**: 111–141. [Errata: *Econometrica* **18**: 312.]
von Neumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press. [Second edition 1947; third edition 1953.]
Pratt J (1964) Risk aversion in the small and in the large. *Econometrica* **32**: 122–136.
Savage L (1954) *The Foundations of Statistics*. New York, NY: Wiley. [Revised and enlarged edition, 1972. New York, NY: Dover.]
Schmeidler D (1989) Subjective probability and expected utility without additivity. *Econometrica* **57**: 571–587.
Slovic P (1969) Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology* **79**: 139–145.

Further Reading

Camerer C and Weber M (1992) Recent developments in modeling preferences: uncertainty and ambiguity. *Journal of Risk and Uncertainty* **5**: 325–370.
Einhorn H and Hogarth R (1985) Ambiguity and uncertainty in probabilistic inference. *Psychological Review* **92**: 433–461.
Epstein L (1999) A definition of uncertainty aversion. *Review of Economic Studies* **66**: 579–608.
Fishburn P (1982) *The Foundations of Expected Utility*. Dordrecht: Reidel.
Heath C and Tversky A (1991) Preferences and belief: ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty* **4**: 5–28.
Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.
Kelsey D and Quiggin J (1992) Theories of choice under ignorance and uncertainty. *Journal of Economic Surveys* **6**: 133–153.
Starmer C (2000) Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* **38**: 332–382.
Tversky A and Fox C (1995) Weighing risk and uncertainty. *Psychological Review* **102**: 269–283.
Weber M and Camerer C (1987) Recent developments in modeling preferences under risk. *OR Spektrum* **9**: 129–151.

Decision-making, Intertemporal

Intermediate article

David Laibson, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

Intertemporal decisions imply trade-offs between current and future rewards. Intertemporal discounting models formalize these trade-offs by quantifying the values of delayed pay-offs.

INTRODUCTION

Decision-makers confront a wide range of critical choices that involve trade-offs between current and future rewards. For example, young workers save part of their paycheck to raise their quality of life in retirement. Habitual heroin users also make decisions with intertemporal consequences when they choose a short-term drug-induced pleasure that jeopardizes their long-term well-being.

To evaluate such trade-offs, decision-makers must compare the costs and benefits of activities that occur at different dates in time. The theory of discounted utility provides one framework for evaluating such delayed pay-offs. This theory has normative and positive content. It has been proposed as both a description of what people should do to maximize their well-being, and to describe what people actually do when faced with intertemporal decisions. Both applications of the model are controversial.

Discounted utility models typically assume that delayed rewards are not as desirable as current rewards or, similarly, that delayed costs are not as undesirable as current costs. This delay effect may reflect many possible factors. For example, delayed rewards are risky because the decision-maker may die before the rewards are experienced. Alternatively delayed rewards are more abstract than current rewards, and hence a decision-maker may not be able to appreciate or evaluate their full impact in advance. Some contributors have argued that delayed rewards should be valued no less than current rewards, with the sole exception of discounting effects that arise from mortality.

DISCOUNTED UTILITY

Formal discounting models assume that a consumer's welfare can be represented as a discounted sum of current and future utility. Specifically, the model assumes that at each point in time, t, the decision-maker consumes goods $c(t)$. These goods might be summarized by a single consumption index (say a consumption budget for period t), or these goods might be represented by a vector (say apples and oranges). The subjective value to the consumer is given by a utility function $u(c(t))$, which translates the consumption measure, $c(t)$, into a single summary measure of utility at period t.

To evaluate future consumption, the consumer discounts utility with a discount function $F(\tau)$, where τ is the delay between the current period and the future consumption. For example, if the current period is date t and a consumer evaluates consumption half a year from now, the consumer calculates the discounted utility value $F\left(\frac{1}{2}\right)u\left(c\left(t+\frac{1}{2}\right)\right)$.

Since future consumption is usually assumed to be worth less than current consumption, the discounted utility model posits that $F(\tau)$ is decreasing in τ. The more utility is delayed, the less it is worth. Since utility is not undesirable, $F(\tau) \geq 0$ for all values of τ. The model is normalized by assuming $F(0) = 1$. Combining these properties we have

$$1 = F(0) \geq F(\tau) \geq F(\tau') \geq 0, \qquad (1)$$

for $0 < \tau < \tau'$. For example if flows of utility a year from now are worth only $\frac{2}{3}$ of what they would be worth if they occurred immediately, then $F(1) = \frac{2}{3}F(0)$.

Continuous-time Discount Functions

Intertemporal choice models have been developed in both continuous-time and discrete-time settings.

Both approaches are summarized here. Readers without a calculus background may wish to skip directly to the discrete-time analysis.

In continuous-time, the welfare of the consumer at time t – sometimes called the objective function or utility function – is given by

$$\int_{\tau=0}^{\infty} F(\tau)u(c(t+\tau))d\tau, \tag{2}$$

where $F(\tau)$ is the discount function, $u(\cdot)$ is the utility function, and $c(\cdot)$ is consumption.

In both continuous-time and discrete-time models, discount functions are described by two characteristics: discount rates and discount factors. Discount rates and discount factors are normalized with respect to the unit of time, which is usually assumed to be a year.

A discount rate at horizon τ is the rate of decline in the discount function at horizon τ:

$$r(\tau) \equiv \frac{-F'(\tau)}{F(\tau)}. \tag{3}$$

Note that $F'(\tau)$ is the derivative of F with respect to time. Hence, $F'(\tau)$ is the change in F per unit time, so $r(\tau)$ is the rate of decline in F. The higher the discount rate, the more quickly value declines with the delay horizon.

A discount factor at horizon τ is the value of a util discounted with the continuously compounded discount rate at horizon τ:

$$f(\tau) \equiv \lim_{\Delta\to 0}\left(\frac{1}{1+r(\tau)\Delta}\right)^{1/\Delta} = \exp(-r(\tau)). \tag{4}$$

The lower the discount factor, the more quickly value declines with the delay horizon.

Discrete-time Discount Functions

Analogous definitions apply to discrete-time models. For this class of models the discount function, $F(\tau)$, need only be defined on a discrete grid of delay values: $\tau \in \{0, \Delta, 2\Delta, 3\Delta, \ldots\}$. For example, if the model were designed to reflect weekly observations, then $\Delta = \frac{1}{52}$ years.

Once the discrete-time grid is fixed, the discount function can then be written

$$\{F(0), F(\Delta), F(2\Delta), F(3\Delta), \ldots\}. \tag{5}$$

The welfare of the consumer at time t is given by

$$\sum_{\tau=0}^{\infty} F(\tau\Delta)u(c(t+\tau\Delta)). \tag{6}$$

At horizon τ, the discount function declines at rate

$$r(\tau) = -\frac{(F(\tau) - F(\tau-\Delta))/\Delta}{F(\tau)}. \tag{7}$$

The numerator of this expression represents the change per unit time.

The discount factor at horizon τ is the value of a util discounted with the discount rate at horizon τ compounded at frequency Δ.

$$f(\tau) = \left(\frac{1}{1+r(\tau)\Delta}\right)^{1/\Delta} = \left(\frac{F(\tau)}{F(\tau-\Delta)}\right)^{1/\Delta}. \tag{8}$$

As the time intervals in the discrete-time formulation become arbitrarily short (i.e. $\Delta \to 0$), the discrete-time discount rate and discount factor definitions converge to the continuous-time definitions.

Exponential Discounting

Almost all discounting applications use the exponential discount function: $F(\tau) = \exp(-\rho\tau)$. This discount function is often written $F(\tau) = \delta^{\tau}$, where $\delta \equiv \exp(-\rho)$. For the exponential discount function the discount rate is constant and does not depend on the horizon:

$$r(\tau) = \frac{-F'(\tau)}{F(\tau)} = \rho = -\ln\delta. \tag{9}$$

Likewise, the discount factor is also constant:

$$f(\tau) = \exp(-r(\tau)) = \exp(-\rho) = \delta. \tag{10}$$

Figure 1 plots three discount functions, including an exponential discount function. Note that the exponential discount function displays a constant rate of decline regardless of the length of the delay. Typical calibrations adopt an annual exponential discount rate of 5 percent.

Non-exponential Discounting

A growing body of experimental evidence suggests that decision-makers' valuations of delayed rewards are inconsistent with the constant discount rate implied by the exponential discount function. Instead, measured discount rates tend to be higher when the delay horizon is short than when the delay horizon is long. One class of functions that satisfies this property is generalized hyperbolas (Chung and Herrnstein, 1961; Ainslie, 1992; Loewenstein and Prelec, 1992). For example,

$$F(\tau) = (1+\alpha\tau)^{-\gamma/\alpha}. \tag{11}$$

For these functions, the rate of decline in the discount function decreases as τ increases:

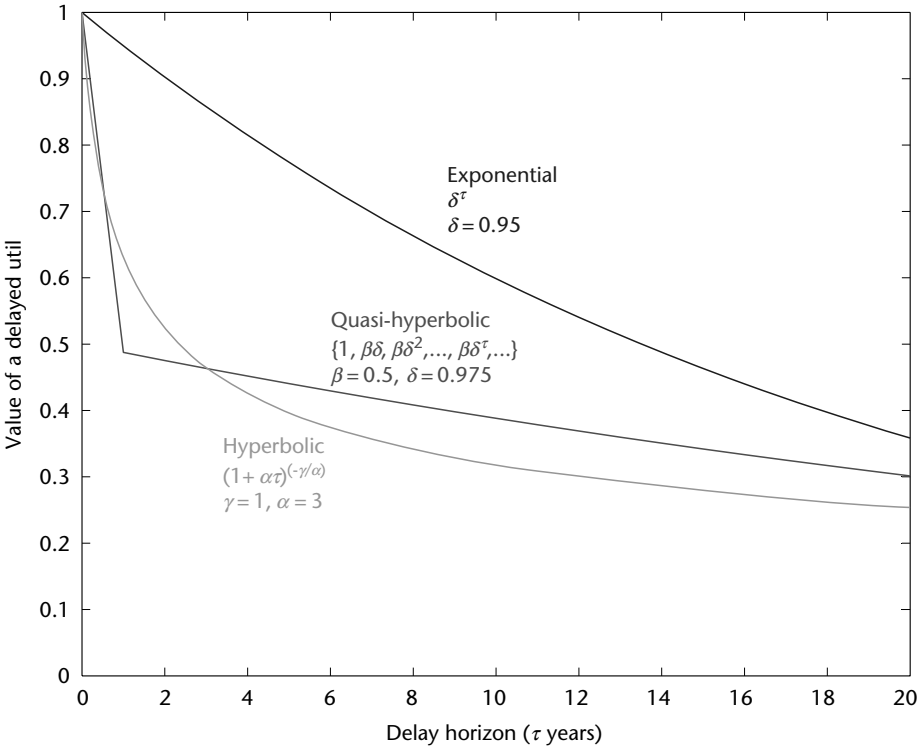


Figure 1. Three discount functions.

$$r(\tau) = \frac{-F'(\tau)}{F(\tau)} = \frac{\gamma}{1+\alpha\tau}. \tag{12}$$

When $\tau = 0$ the discount rate is γ. As τ increases, the discount rate converges to zero. Figure 1 also plots this generalized hyperbolic discount function. Note that the generalized hyperbolic discount function declines at a faster rate in the short run than in the long run, matching a key feature of the experimental data.

To capture this qualitative property, Laibson (1997) adopts a discrete-time discount function, $\{1, \beta\delta, \beta\delta^2, \beta\delta^3, \ldots.\}$, which Phelps and Pollak (1968) had previously used to model intergenerational time preferences. This 'quasi-hyperbolic function' reflects the sharp short-run drop in valuation measured in the experimental time preference data and has been adopted as a research tool because of its analytical tractability. The quasi-hyperbolic discount function is only hyperbolic in the sense that it captures the key qualitative property of the hyperbolic functions: a faster rate of decline in the short run than in the long run. This discrete-time discount function has also been called 'present-biased' and 'quasi-geometric'.

The quasi-hyperbolic discount function is typically calibrated with $\beta \simeq \frac{1}{2}$ and $\delta \simeq 1$. Under this calibration the short-run discount rate exceeds the long-run discount rate:

$$r(1) = 1 > 0 = r(\tau > 1). \tag{13}$$

More generally, any calibration with $0 < \beta < 1$, implies that the short-run discount rate exceeds the long-run discount rate.

Measuring Discount Functions

Measuring discount rates has proved to be controversial. Most discount rate studies give subjects choices among a wide range of delayed rewards. The researchers then impute the time preferences of the subjects based on the subjects' laboratory choices.

In a typical discount rate study, the researchers make three very strong assumptions. First, the researchers assume that rewards are consumed when they are received by the subjects. Second, the researchers assume that the utility function is linear in consumption. Third, the researchers assume that the subjects fully trust the researchers' promises to pay delayed rewards.

If these assumptions are satisfied, it is then possible to impute the discount function by offering subjects reward alternatives and asking the subjects

to pick their preferred option. For example, if a subject prefers \$x today to \$y in one year, then the experimenter concludes that

$$F(0) \cdot x > F(1) \cdot y. \tag{14}$$

Once the subject answers a wide range of questions of this general form, the researcher attempts to estimate the discount function that most closely matches the subject responses.

Such experiments usually reject the exponential discount function in favor of alternative discount functions characterized by discount rates that decline as the horizon is lengthened (Thaler, 1981). Related experiments also reject the implicit assumption that the utility function at date t is not affected by consumption at other dates. This 'separability' assumption is contradicted by the finding that subjects care both about the level *and* the slope of their consumption profiles (see Loewenstein and Thaler (1989) for findings that violate the discounted utility model).

DYNAMIC CONSISTENCY AND SELF-CONTROL

Exponential discount functions have the convenient property that preferences held at date t do not change with the passage of time. Consider the following illustration of this 'dynamic consistency' property. Suppose that at date 0 a consumer with an exponential discount function with discount factor δ, is asked to evaluate a project that requires investments that cost c utils at time 10 with resulting benefits of b utils at time 11. From the perspective of date zero, this project has utility value $\delta^{10}(-c) + \delta^{11}b$. Assume that this utility value is positive and that at date zero the consumer plans to execute the project.

Now imagine that 10 periods pass, and the consumer is asked whether she wishes to reconsider her decision to make the planned investment. From the perspective of period 10, the value of the project to the consumer is $-c + \delta b$. The costs are no longer discounted, since they need to be made in the current period. Likewise, the benefits are only discounted with factor $F(1) = \delta^1$ since they will now be available at only one period in the future.

Note that the consumer's original preference to pursue the project is unchanged by the passage of time, since $\delta^{10}(-c) + \delta^{11}b > 0$ implies that $-c + \delta b > 0$ (divide both sides of the original inequality by δ^{10}). This property of intertemporally consistent preferences is called 'dynamic consistency' and the property will always arise when the discount function is exponential. The passage of time will *never* cause the consumer to switch her preference regarding the investment project (unless new information arrives).

However, dynamic consistency is not a general property of intertemporal choice models. In fact, the only stationary discount function that generates this property is the exponential discount function. All other discount functions imply that preferences are dynamically inconsistent: preferences will sometimes switch with the passage of time. To see this, reconsider the investment project described above and evaluate it with the quasi-hyperbolic discount function (assume $\beta = \frac{1}{2}$ and $\delta = 1$). Suppose that the project requires an investment that costs 2 utils at time 10 and generates a pay-off of 3 utils at time 11. From the time 0 perspective,

$$\beta\delta^{10}(-c) + \beta\delta^{11}b = \frac{1}{2}(-2) + \frac{1}{2}3 = \frac{1}{2}, \tag{15}$$

so the project is worth pursuing. However, from the perspective of period 10, the project generates negative discounted utility

$$-c + \beta\delta b = -2 + \frac{1}{2}3 = -\frac{1}{2}. \tag{16}$$

Hence, the project that the consumer wished to pursue from the perspective of time 0 ceases to be appealing once the moment for investment actually arises in period 10.

This example captures a tension that many decision-makers experience. From a distance a project seems worth doing, but as the moment for sacrifice approaches the project becomes increasingly unappealing. For this reason, quasi-hyperbolic discount functions have been used to model a wide range of self-regulation problems, including procrastination, credit card spending, and drug addiction.

Sophistication, Commitment, and Naivité

The analysis above does not take a stand on whether consumers foresee these preference reversals. Strotz (1956) identifies two paradigms that can be used to analyze the question of consumer foresight: sophistication and naivité.

Sophisticated consumers will anticipate their own propensity to experience preference reversals. Such consumers will recognize the conflict between their early preference – i.e. the preference to undertake the investment project – and their later contradictory preferences. Such sophisticated consumers may look for ways to lock themselves into the investment activity. For example, consider a person

who forces himself to exercise by making an appointment with an expensive trainer.

At the other extreme, Strotz also considered consumers who exhibit naivité about their future preference reversals. Such consumers fail to foresee these reservals and expect to engage in investments that they will not actually carry out (e.g. quitting smoking or completing a project with no deadline). Akerlof (1991) discusses such procrastination problems and O'Donoghue and Rabin (2001) propose a framework that continuously indexes the degree of naivité.

SUMMARY

The discounted utility model provides a way of formally evaluating intertemporal trade-offs. The principal component of the model is a discount function that is used to calculate the discounted value of future utility flows. The key characteristics of the discount function are the discount rate and the discount factor. The discount rate measures the rate of decline of the discount function. The discount factor measures the value of a discounted util. Exponential discount functions are commonly used in most applications of the discounted utility model. Exponential discount functions have a constant discount rate. Exponential discount functions also have the convenient property that they do not generate preference reversals. However, the experimental evidence contradicts the constant discount rate property. Most experimental evidence suggests that the discount rate declines with the length of the delay horizon. Such discounting patterns may play a role in generating self-control problems.

References

Ainslie G (1992) *Picoeconomics*. Cambridge, UK: Cambridge University Press.

Akerlof GA (1991) Procrastination and obedience. *American Economic Review Papers and Proceedings* **81**: 1–19.

Chung SH and Herrnstein RJ (1961) Relative and absolute strengths of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Animal Behavior* **4**: 267–272.

Laibson DI (1997) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* **62**: 443–478.

Loewenstein G and Thaler RH (1989) Anomalies: intertemporal choice. *Journal of Economic Perspectives* **3**: 181–193.

Loewenstein G and Prelec D (1992) Anomalies in intertemporal choice. Evidence and an interpretation. *Quarterly Journal of Economics* **57**: 573–598.

O'Donoghue T and Rabin M (2001) Choice and procrastination. *Quarterly Journal of Economics* **66**: 121–160.

Phelps ES and Pollak RA (1968) On second-best national saving and game-equilibrium growth. *Review of Economic Studies* **35**: 185–199.

Strotz RH (1956) Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* **23**: 165–180.

Thaler RH (1981) Some empirical evidence on dynamic inconsistency. *Economic Letters* **8**: 201–207.

Further Reading

Angeletos G, Laibson DI, Repetto A, Tobacman J and Weinberg S (2001) The hyperbolic consumption model: calibration, simulation, and empirical evaluation. *Journal of Economic Perspectives* **15**(3): 47–68.

Frederick S, Loewenstein G and O'Donoghue T (in press) Time discounting and time preference: a critical review. *Journal of Economic Literature*.

Kirby KN (1997) Bidding on the future: evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology* **126**: 54–70.

Laibson DI (2001) A cue-theory of consumption. *Quarterly Journal of Economics* **66**: 81–120.

Loewenstein G (1996) Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes* **65**: 272–292.

Loewenstein G and Elster J (eds) (1992) *Choice Over Time*. New York, NY: Russel Sage Foundation Press.

Loewenstein G and Prelec D (1991) Negative time preference. *American Economic Review Papers and Proceedings* **82**: 347–352.

Loewenstein G, Read D and Kalyanaraman S (1999) Mixing virtue and vice: the combined effects of hyperbolic discounting and diversification. *Journal of Behavioral Decision Making* **12**: 257–273.

Loewenstein G, Read D and Baumeister R (eds) (in press) *Intertemporal Choice*. New York, NY: Russell Sage Foundation Press.

O'Donoghue T and Rabin M (1999) Doing it now or later. *American Economic Review* **89**: 103–124.

Thaler RH and Shefrin HM (1981) An economic theory of self-control. *Journal of Political Economy* **89**: 392–410.

Dynamic Decision Makers, Classification of Types of

Intermediate article

Daniel Houser, University of Arizona, Tucson, Arizona, USA

CONTENTS

A dynamic decision problem is one that requires a sequence of decisions in a setting where pay-offs and alternatives available for later decisions depend on earlier choices. Analysis of data from dynamic behavioral experiments can shed light on the nature of the different types of dynamic decision makers in the population.

INTRODUCTION

Types

Suppose a general orders his troops into battle against an approaching enemy and valiantly leads the charge himself. One of his soldiers takes advantage of the momentary confusion surrounding the order by slipping away from the battlefield to an area of relative safety. One reasonable explanation for the different decisions made by the general and soldier is straightforward: they are different 'types' just in that they face different incentives. Victory in battle means lasting glory and honor for the general, while the soldier might receive little but the chance to fight again another day. It is less straightforward, however, to explain why one soldier flees while his comrades, who are in the same situation, rush forward with the general into combat. As a casual description, we might also say that 'observationally' identical soldiers who make different decisions are different 'types'.

In order to decide whether to flee or fight, soldiers must solve a dynamic decision problem. It is dynamic because their decision affects in a nontrivial way the alternatives available to them at later times. For example, the soldier who fights might be in a position to save the life of a wounded comrade on the battlefield. The soldier who flees might save his own life, but at the risk of being punished for desertion. Their eventual decision rests on idiosyncratic characteristics such as preferences and subjective assessments of battlefield risks.

Game theory defines people who have different preferences as different 'types'. Unfortunately, preferences are not observable. In practice, it is more useful to define people as different types of dynamic decision makers if, like the soldiers, they make different decisions in observationally identical dynamic situations.

Interest in classifying and characterizing dynamic decision makers has grown as the importance of accounting for type heterogeneity in dynamic economic models has become apparent. This importance stems from the fact that most of dynamic economics has as its final goal policy analysis. That is, the goal is to predict how different sorts of incentive structures (e.g. the tax system) affect dynamic decisions (e.g. work and educational choices). As the example of the soldiers makes clear, not everybody responds to the same incentives in the same way. Economists increasingly recognize that analyses, which assume that firms and societies can be described as collectives and modeled as though they were single agents, can often lead to very misleading conclusions and policy recommendations (Furubotn and Richter, 2000). Models that take account of type heterogeneity have the potential to improve policy analysis substantially.

Decision Rules

Economists use so-called 'decision rules' to describe the way a person's actions depend on personal information. 'Information' here should be thought of broadly as everything a person knows (including demographic variables) that could be relevant to a decision. The soldiers above, for example, can be thought of as having the choice either to fight or flee. When presented with the information that the enemy is approaching, the decision rules of some soldiers generate the

decision to flee, while those of others generated the decision to fight. In general, people in observationally identical situations who make different decisions are viewed as using different decision rules, and a person's 'type' is defined by the decision rule they use. Increasingly, research is directed towards identifying, at least within narrow contexts, the number and nature of different decision rules, or types, that exist in the population.

Differences in decision rules can arise from differences in preferences. This is a good explanation in many situations, particularly when trying to explain idiosyncratic differences in tastes for, say, coffee and tea. On the other hand, differences along dimensions such as propensities to cooperate, which are well documented and involve higher-order cognitive processing, seem less naturally attributable to preferences. There is some evidence that differences in decision rules associated with higher-order functions are due to different cognitive algorithms employed to determine actions (e.g. McCabe *et al*., 2001). Such differences are analogous to the difference between human and computer decision making. When a human plays chess against a computer, both parties have the same objective and information, yet their decision rules differ because they use different algorithms to determine their moves.

Expectations

A higher-order task of particular importance to dynamic decisions is expectation formation, because all dynamic decision problems require some sort of forward-looking behavior. Economists have been using the 'rational expectations' assumption to model forward-looking behavior for decades. However, numerous studies in economics and psychology suggest that expectations are not formed rationally. Moreover, it is straightforward to show that different expectation formation mechanisms lead to different dynamic decision rules.

An important, and often-replicated, finding in the literature on static decision making is that, except in very simple cases, people do not assess objective probabilities correctly (e.g. Camerer, 1995). Since probability assessment is cognitively difficult, it is presumably accomplished with idiosyncratic heuristics. Moreover, because probability assessment is fundamental to expectation formation, it seems likely that expectations are formed with idiosyncratic and imperfect heuristics. Although research in this area is still in its early stages, it seems plausible that heterogeneity in expectation formation underlies much of the idiosyncratic variation in dynamic decision rules.

TYPE ELICITATION

Stopping Experiments

Dynamic decision problems (DDPs) faced by individuals provide perhaps the simplest interesting environment in which to study dynamic decision rule heterogeneity. In these environments an individual makes several decisions sequentially, and the decisions made early in the problem affect the nature of the decision task later in the problem. For example, a person might first decide whether to bicycle or walk to work, and then decide on the route to follow. This is a DDP, because the set of candidate routes depends on the outcome of the first decision. (Economists contrast DDPs with sequential 'static' decision problems, where one makes a series of unrelated decisions.) It is important to understand how people actually solve DDPs, particularly when the dynamic nature of the problem involves deciding between different pay-offs at different times (so-called 'intertemporal' decision problems), because many actual consumption, savings, and labor supply decisions must be made within this context.

'Stopping problems', a widely studied class of DDPs, have proved useful tools in classifying and characterizing dynamic decision rules. In a simple stopping experiment, subjects receive payment offers sequentially from the experimenter until they accept one, at which time the experiment ends. Many variants of this basic design have been studied. For example, subjects might have to pay for offers; they might not know the distribution from which offers are generated; they might be able to accept previously rejected offers; and they might not know the exact amount of the offer, but only whether it is higher or lower than other offers. An advantage of this framework is that theoretical predictions about behavior under various decision rules are straightforward to derive. Different decision rules often imply different stopping points. Hence, observing stopping times allows one to draw simple and compelling inferences about the sorts of decision rules that are used in the population.

Analysis of stopping experiment data, using techniques such as that of El-Gamal and Grether (1995) discussed below, show that there is great heterogeneity in the ways subjects solve experimental stopping problems. There is little evidence to suggest that people decide in ways that are

consistent with rational expectations. Instead, subjects seem to use sophisticated, nonstationary reservation pay-off heuristics. This means that subjects stop as soon as their pay-off is sufficiently high, where 'sufficiently' depends, for example, on the number of times they have had to search and on whether they are paying search costs.

There is a small set of reservation pay-off decision rules into which most subjects' behavior seems to fall. Moreover, there are two features that most of these rules share. Firstly, subjects who use them tend to stop searching somewhat earlier than a rational expectations searcher would. Secondly, these heuristics work well in the sense that subjects who use them earn only slightly less on average (often about one percent) than they would have if they had followed the rational expectations rule. Since the rational expectations rule is cognitively very complex to implement, there may be a sense in which using reservation pay-off heuristics is in fact optimal. For a survey of results from the experimental stopping experiment literature, see Cox and Oaxaca (1996).

The Voluntary Contribution Mechanism

Types can also be discerned in game environments where multiple subjects interact and make strategic decisions that affect each other's pay-off. The voluntary contribution mechanism (VCM) is an important example of such a game. There are N players, and player n has endowment w_n. Player n contributes g_n to the public good and leaves the remainder in a private account. The total contribution to the public good is $G = \sum_n g_n$. The interesting feature of the VCM is that the return on investment in the public account differs from that on investment in the private account. Without loss of generality we can suppose that the return to each player on the total investment in the public account is given by r while the return on the private account is set to unity. This means that the pay-off function for player n is

$$\Pi(g_n, \hat{g}_n) = (w_n - g_n) + rG \tag{1}$$

where $\hat{g}_n$ represents the vector of contributions of everyone except player n. Provided that $r < 1$ it is easy to see that, given any arrangement of contributions by the other subjects, each player maximizes his or her individual pay-off by 'free-riding', or contributing zero to the public good. Hence, free-riding is a dominant Nash equilibrium strategy. But if $rN > 1$ then it is Pareto optimal for each player to contribute everything to the public good, and this strategy Pareto-dominates free-riding. The parameter r is the marginal per-capital return (MPCR). When designing VCM experiments, the MPCR and the number of subjects are usually chosen to exploit the tension between free-riding and Pareto optimality.

Experimental research with the VCM, has generated many widely-replicated results, including clear evidence of decision rule heterogeneity (for a survey see Ledyard, 1995). In particular, there is usually a subset of subjects, 'free-riders', whose decision is to contribute very little to the public good in every round, and another subset, 'cooperators', who systematically contribute a large fraction of their endowment to the public good. A third subset uses 'reciprocal' rules, trying to match others' contributions.

The presence of reciprocal decision rules suggests that group dynamics might be influenced by the type composition of groups. For example, the presence of players who contribute little or nothing to the public good could lead to decreasing aggregate contributions over time if reciprocators attempt to match free-riders' small contributions. Hence, a feedback system that is sensitive to the proportion of each type within the group could be created, and could affect the extent to which a group is able to sustain cooperation.

Recent research has found that type composition seems to have important effects on group dynamics (e.g. Gunthorsdottir *et al.*, 2001). In particular, the number of free-riders in a group influences that group's path over the course of a game. Without free-riders, groups are capable of sustaining high levels of contribution to the public good; while the presence of free-riders often pushes groups towards successively lower levels of contributions.

TYPES AND PERSONALITY SURVEYS

The ability to learn about a person's behavioral type from a personality survey would be useful, since the dependence of group outcomes on type composition implies that knowledge of types could be used to design groups (such as school classes) efficiently. However, whether behavioral types broadly and systematically correlate with personality surveys is an open question, and experiments have generated widely conflicting results. Nevertheless, some personality variables seem to correlate with propensities to cooperate in experiments. Personality dimensions displaying this correlation include Machiavellianism, self-monitoring, and three of the 'big five' personality traits.

Machiavellianism

Inspired by the writings of Niccolo Machiavelli (1469–1527), and first developed by Christie and Geis (1970), the Machiavellianism (or Mach) scale measures the extent to which a person agrees that the end sanctifies the means. People who score highly on the Mach scale tend to be manipulative, opportunistic, and rational. Low Machs tend to be more emotional and less likely to depart from social norms in order to pursue their own self-interest. While high Machs tend to be competitive and exploitative, low Machs are usually more willing to cooperate (Gunnthorsdottir, McCabe and Smith, 2002).

Self-monitoring

The 'self-monitoring' scale is an measure of the dependence of an individual's behavior on the social context. High self-monitors work to create the impression needed to obtain their social goals, while low self-monitors are less concerned about the impression they make. High self-monitors have been found to be more likely to cooperate, particularly in experiments where repeated interactions with the same counterpart are possible.

The Big Five

The 'big five' personality traits are extraversion, agreeableness, conscientiousness, neuroticism, and openness. Among these, extraversion and agreeableness seem to be positively correlated with cooperativeness, while neuroticism seems to be negatively correlated. The relation of the other two traits with cooperativeness is not clear.

Many other personality variables, including self-esteem and locus of control, have been studied in relation to cooperation, but without clear results. For further discussion on the connection between types and personality surveys, see Kurzban and Houser (2002).

STATISTICAL METHODS FOR TYPE CLASSIFICATION

Sophisticated statistical procedures are not usually required to determine whether subjects in an experiment behave according to a particular decision rule. Intuitively, all that is required is to compare actual decisions with those that would arise under a hypothesized behavior. Although the details depend on the experimental design, formal procedures to accomplish this sort of comparison are typically straightforward. A more difficult task, and one that typically requires sophisticated statistics, is to determine how people actually make decisions in a given dynamic environment.

Attempting to characterize actual decision making requires, at least, allowing for multiple decision rules to be used in the population. The task is then to determine the number of decision rules, and to assign each subject to a decision rule. Broadly, there are two ways in which this can be done. We will briefly summarize the two approaches, and then discuss in greater detail an instance of each of them.

One approach, exemplified by a procedure suggested by El-Gamal and Grether (1995), requires one to specify in advance a set of candidate decision rules. A statistical procedure is then used to choose a 'best' subset of these rules. Finally, each subject is assigned to one of the rules in the subset. An advantage of this sort of procedure is that it is relatively straightforward to implement. However, unless it is feasible to include all of the rules that subjects might possibly use in the prespecified superset, a potential drawback is that the right rules might not be included. Misspecification could mask underlying commonalities in subjects' play.

An alternative approach, exemplified by a method suggested by Houser *et al*. (2001), requires no assumptions about the number of decision rules used in the population, the nature of each decision rule, or the assignment of subjects to decision rules. This approach requires cluster analysis: subjects are clustered according to commonalities in their behavior.

The goal of both these approaches is to put each subject into a behavioral category. El-Gamal and Grether require one to specify the categories in advance, while Houser *et al*. allow the categories to be determined by the data. Of course, it may not be easy to assign behavioral labels to groups that follow statistically similar decision rules.

The Classification Procedure of El-Gamal and Grether

Suppose one has data from a behavioral laboratory experiment where each of N subjects makes T decisions. Let C^K denote the prespecified set of K heuristics (i.e., decision rules) that subjects might use to make these decisions, and let $c \in C^K$ denote a particular heuristic. The idea is to determine, for each subject, the number of decisions consistent with each possible heuristic, and then assign that subject to the heuristic that best fits his or her behavior.

To implement the procedure one assumes that each subject follows exactly one of the heuristics in C^K. In practise, of course, a subject's behavior may not be perfectly consistent with any of the heuristics in C^K. El-Gamal and Grether (1995) circumvent this problem by assuming that subjects follow their heuristics with error.

Heuristics are chosen that specify a subject's decision uniquely from his or her state. A subject's 'state', which generally changes after each decision, is a vector that summarizes all of the person's decision-relevant information. Let x_t^c be an indicator variable that takes the value one if the subject's tth decision agrees with heuristic c and takes the value zero otherwise. Assume that the decisions are made independently with common error rate ε.

Let $(x_{n1}, \ldots, x_{nT})$ be a vector denoting subject n's actions, and let $(x_{n1}^c, \ldots, x_{nT}^c)$ be a vector of zeros and ones that summarizes the consistency of the subject's choices with c. That is, assume that $x_{nj}^c = 1$ if decision rule c predicts decision x_{nj}, and $x_{nj}^c = 0$ otherwise. Then set $X_n^c = \sum_t x_{nt}^c$. The likelihood function for the subjects' actions is then:

$$f^c(x_{n1}, \ldots, x_{nT}) = (1 - \varepsilon/2)^{X_n^c} (\varepsilon/2)^{T - X_n^c} \tag{2}$$

It is natural to assign each subject to the heuristic from the candidate set that maximizes his or her likelihood function.

This model can be 'overfit': including a large number of heuristics in C^K would allow the statistical model to fit a sample arbitrarily well. Overfitting usually leads to results with little external validity. To avoid overfitting, El-Gamal and Grether suggest penalizing the likelihood for each additional heuristic that is included in the set of candidate heuristics. Let C^k denote a subset of $k \leq K$ decision rules. El-Gamal and Grether argue that a reasonable penalized log-likelihood is obtained by forming the Bayesian posterior that arises under the following priors: (1) the probability that the population includes exactly k heuristics is $1/2^k$; (2) all possible k-tuples of heuristics in any C^k are equally likely (each with probability $1/K^k$); (3) all allocations of heuristics to subjects are equally likely (each with probability $1/k^N$); (4) all error rates (between zero and one) are equally likely; and do not depend on the number of rules used in the population or on the way those rules are assigned. This generates the following penalized log-likelihood function:

$$\log\left(\prod_n \max_{c_n \in C^k} f^{c_n}(x_{n1}, \ldots, x_{nT})\right) - k \log 2 - N \log k - k \log K \tag{3}$$

Determination of the population of heuristics as well as the assignment of subjects to heuristics is accomplished by simply maximizing the above expression over the set of all possible k-tuples that can be formed from the set of K decision rules.

The Classification Procedure of Houser, Keane, and McCabe

Suppose that subjects solve a 15-period DDP. At each period, subjects choose either A or B, each of which results in a nonnegative monetary reward. Pay-offs are stochastic. The realizations of the random variables for period t occur before the decision at t is made, and the realizations of the random variables for period $t+1$ occur after the decision at t. Each subject's total pay-off is the sum of the rewards earned over the 15 periods. Subjects have complete information regarding the stochastic link between their current choices and future pay-offs, but the link is complicated and it is difficult to determine the decision rule that maximizes expected total pay-offs.

The goal is to learn about the dynamic decision rules that subjects actually use when solving this difficult problem. Houser *et al*. (2001) begin by assuming that subjects are rational in a weak sense. In particular, a subject will choose alternative A in period t if and only if, in period t, the value the person places on choosing A is greater than the value he or she places on choosing B. Because the problem is dynamic, the values that subjects place on A and B depend both on the immediate reward to each choice and on the way subjects believe that choice would influence their future pay-offs. Houser *et al*. assume that alternative valuations are additively separable into a 'present' component, which captures immediate rewards (in this case the immediate monetary pay-off), and a 'future' component, which captures any benefits expected to accrue in subsequent periods as a result of that choice (in this case future monetary pay-offs).

Since the present pay-off structure is known for each agent, differences in behavior result only from differences in the future component. Hence, all differences in decision rules between subjects can be captured by differences in the future component. Houser *et al*. propose clustering subjects into groups that seem to have similar future components, while simultaneously drawing inferences about the future components' forms. In this way, they avoid the need to prespecify the nature of the decision rules used by the subjects.

Drawing on Geweke and Keane (1999), Houser *et al*. model the unobserved future component of each alternative's value as a parametric function of the subject's information set I_{nt}. The information set can include anything the researcher believes is relevant to the subject when making his or her decision, such as choice and pay-off histories. Then, the value that subject n assigns to alternative $j \in \{A, B\}$ in period t, V_{njt} (I_{nt}), assuming that the person uses decision rule k, can be written

$$V_{njt}(I_{nt}|k) = w_{njt} + F(I_{n,t+1}|I_{nt}, j, \pi_k, \varsigma_{njtk}) \quad (4)$$

$$I_{n,t+1} = H(I_{nt}, j) \quad (5)$$

Here, w_{njt} is the known immediate pay-off associated with alternative j. $F(\cdot)$ represents the future component. It depends on the alternative j and information set I_{nt} and is characterized by a finite vector of parameters π_k, whose values determine the nature of decision rule k, and a random variable ς_{njtk} that accounts for idiosyncratic errors subjects make when attempting to implement decision rule k. (The researcher must specify the distribution of the idiosyncratic errors.) The function $H(\cdot)$ is the information set's (possibly stochastic) law of motion. It provides the dynamic link between current information, actions and future information. Note that it does not vary with the decision rule.

We denote the choice in period t of subject n following decision rule k with information I_{nt} by

$$d_k(I_{nt}) = \begin{cases} A & \text{if } Z_{nt}(I_{nt}|k) > 0 \\ B & \text{otherwise} \end{cases} \text{ for all } k \in K, \quad (6)$$

where $Z_{nt}(I_{nt}|k) = V_{nAt}(I_{nt}|k) - V_{nBt}\ (I_{nt}|k)$.

The goal is to draw inferences about the parameters $\pi_k (k \in K)$, and about the probability with which each subject uses each decision rule. To this end, Houser *et al*. construct the likelihood function associated with this framework. This requires knowing the probability, conditional on a subject's information set, that he or she will choose A or B.

The probability that subject n using decision rule k chooses alternative A at period t, given that the person has information I_{nt}, is given by

$$\begin{aligned} P(d_k(I_{nt}) = A) &= P(V_{nAt}(I_{nt}) > V_{nBt}(I_{nt})) \\ &= P(w_{nAt} - w_{nBt} + f(I_{nt}|\pi_k) > 0) \end{aligned} \quad (7)$$

where $f(\cdot)$ is a stochastic function that represents the differenced future components $F(I_{n,t+1}|I_{nt}, A, \pi_k, \varsigma_{nAtk}) - F(I_{n,t+1}|I_{nt}, B, \pi_k, \varsigma_{nBtk})$. The conditional probability that B is chosen is one minus the conditional probability that A is chosen.

Knowing the conditional choice probabilities, it is straightforward to construct the likelihood function needed to draw inferences about the different decision rules used in the population, and the probability with which each subject uses each rule. Under the distributional assumptions made by Houser *et al*. the likelihood function corresponds to a mixture of normals probit model. Unfortunately, this likelihood can be computationally difficult to maximize. Further discussion of this point (and estimation strategies) can be found in Houser *et al*. (2001) and Geweke and Keane (1999).

CONCLUSION

Economists say that people who make different decisions in observationally identical situations are different 'types'. Decision rules form the link between a person's situation and decisions, and it is natural to define a person's type by the decision rule he or she uses. Investigating the nature of the various decision rules at use in the population is important, because the effects of incentives on behavior depend on the decision rules that incentives act upon.

Many economists are particularly interested in the decision rules people use in dynamic environments. Experimental studies have found that a small number of decision rules seem to explain most observed behavior in very narrow contexts, and that these rules do not usually include the rational expectations rule. Further research is needed to determine the nature and number of decision rules in the population, the relationship between decision rules used in different contexts, and consequent implications for individual and group outcomes and incentive structures. Such research may employ sophisticated statistical procedures that group people according to common behavioral patterns. These patterns may be either specified in advance or discerned directly from experimental data.

References

Camerer C (1995) Individual decision making. In: Kagel J and Roth A (eds) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Christie R and Geis F (1970) *Studies in Machiavellianism*. New York, NY: Academic Press.

Cox JC and Oaxaca R (1996) Testing job search models: the laboratory approach. *Research in Labour Economics* **15**: 171–207. Greenwich, CT and London: JAI Press.

El-Gamal M and Grether D (1995) Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association* **90**: 1137–1145.

Furubotn EG and Richter R (2000) *Institutions and Economic Theory: The Contribution of the New Institutional Economics*. Ann Arbor, MI: University of Michigan Press.

Geweke J and Keane M (1999) Bayesian inference for dynamic discrete choice models without the need for dynamic programming. In: Mariano, Schuermann and Weeks (eds) *Simulation Based Inference and Econometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Gunthorsdottir A, Houser D, McCabe K and Ameden H (2001) Disposition, history and contributions in public goods experiments. [Working paper, University of Arizona.]

Gunnthorsdottir A, McCabe K and Smith V (2002) Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology* **23**: 49–66.

Houser D, Keane M and McCabe K (2001) How do people actually solve dynamic decision problems? [Working paper, University of Arizona.]

Kurzban R and Houser D (2002) Individual differences in cooperation in a circular public goods game. *European Journal of Personality* **15**: 37–52.

Ledyard J (1995) Public goods: a survey of experimental research. In: Kagel J and Roth A (eds) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

McCabe K, Houser D, Ryan L, Smith V and Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Science* **98**: 11832–11835.

Further Reading

Andreoni J (1995) Cooperation in public goods experiments: kindness or confusion? *American Economic Review* **85**: 891–904.

Geweke J, Houser D and Keane M (2001) Simulation based inference for dynamic multinomial choice models. In: Baltagi B (ed.) *Companion to Theoretical Econometrics*. Oxford, UK: Blackwell.

Geweke J and Keane M (1997) Mixture of normals probit models. [Federal Reserve Bank of Minneapolis staff report 237.]

Geweke J and Keane M (2001) Computationally intensive methods for integration in econometrics. In: Heckman J and Leamer E (eds) *Handbook of Econometrics*, vol. V. North Holland.

Gilks WR, Richardson S and Spiegelhalter DJ (1996) *Markov Chain Monte Carlo in Practice*. London, UK: Chapman & Hall.

Krusell P and Smith AA (1995) Rules of thumb in macroeconomic equilibrium: a quantitative analysis. *Journal of Economic Dynamics and Control* **20**: 527–558.

Lettau M and Uhlig H (1999) Rules of thumb versus dynamic programming. *American Economic Review* **89**: 148–174.

McLachlan GJ and Basford KE (1988) *Mixture Models: Inference and Applications to Clustering*. New York, NY: Marcel Dekker.

Sargent TJ (1987) *Dynamic Macroeconomic Theory*. Cambridge, MA: Harvard University Press.

Stokey NL and Lucas RE (1989) *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

Economics Experiments, Learning in

Advanced article

Jacob K Goeree, University of Amsterdam, Amsterdam, The Netherlands
Charles A Holt, University of Virginia, Charlottesville, Virginia, USA

CONTENTS

Models of learning in economics are used to explain how people use information about past prices and other signals to make good decisions. These models can be tested with economics experiments in which decisions are made in a sequence of rounds or trading periods.

INTRODUCTION

The main focus of economic analysis is on equilibrium steady states, e.g. on prices determined by the intersection of supply and demand. The preoccupation with equilibrium is perhaps due to the fact that many markets operate for protracted periods of time under fairly stationary conditions. The awareness that there may be multiple equilibria, some of which are bad for all concerned, has raised interest in why behavior might converge to one equilibrium and not to another. As a result, there is renewed interest among economists in mathematical models of learning that were studied extensively by psychologists in the 1960s. This article will describe two of those models, 'reinforcement learning' and Bayesian 'belief learning'. These models and their generalizations will be discussed in the context of a binary prediction task, which may generate the kind of behavior that is known in the psychology literature as 'probability matching'.

We will then use these learning models to analyze behavior in an economic market where firms choose prices. Markets and games are more complex than individual decision tasks, in that people's choices affect others' beliefs. One role of learning models in such situations is to provide an explanation of the dynamic paths of prices, which can shed light on the nature of adjustment towards equilibrium. The equilibrium is characterized by an unchanging (steady-state) distribution of beliefs across individuals, which we call a 'stochastic learning equilibrium'.

TYPES OF LEARNING MODELS

We will introduce the basic learning models in the context of a binary prediction task that has been familiar to psychologists since the 1950s. This task is of special interest because humans are thought to be slow learners in this context. The typical setup involves two lights, each with a corresponding lever (or computer key). A signal light indicates that a decision can be made, and then one of the levers is pressed. Finally, one of the lights is illuminated. Animal subjects like rats and chicks are reinforced with food pellets when the prediction is correct. Human subjects are sometimes told to 'do your best' to predict accurately or to 'maximize the number of correct choices'. In other studies, humans are paid small amounts, typically pennies, for correct choices, and penalties may be deducted for incorrect choices.

The general result seems to be that humans are subject to 'probability matching', predicting each event with a frequency that approximately matches the frequency with which it actually occurs. For example, if the left light is illuminated three-fourths of the time, then subjects would come to learn this by experience and would tend to predict 'left' three-fourths of the time.

This behavior is not rational. A consistent prediction of the more likely event will be correct three-fourths of the time. Matching behavior will only generate a correct prediction with a probability of $\frac{3}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{4}$ (the first term corresponds to predicting the more likely event with probability $\frac{3}{4}$ and being correct with this prediction three-fourths of

the time, and similarly for the second term). Thus, the probability of being correct under probability matching is $\frac{5}{8} < \frac{3}{4}$.

In a recent summary of the probability-matching literature, the psychologist Fantino (1998, pp. 360–361) concludes: 'Human subjects do not behave optimally. Instead they match the proportion of their choices to the probability of reinforcement.… This behavior is perplexing given that non-humans are quite adept at optimal behavior in this situation.' For example, Mackintosh (1996) conducted probability-matching experiments with chicks and rats, and the choice frequencies were well above the probability-matching predictions in most treatments.

The resolution of this paradox may be found in the work of Sidney Siegel, who is perhaps the psychologist who has had the largest impact on experiments in economics. His early work from the 1960s exhibits a high standard of careful reporting and procedures, appropriate statistical techniques, and the use of financial incentives where appropriate. His work on probability matching is a good example. In one experiment, 36 male Penn State students were allowed to make predictions for 100 trials, and then 12 of these students were brought back on a later day to make predictions in 200 more trials (Siegel *et al*., 1964). The proportions of predictions for the more likely event are shown in Figure 1, in which each point is an average over 20 trials.

The 12 subjects in the 'no pay' treatment were simply told to 'do your best' to predict which light bulb would be illuminated. These prediction rates begin at about 0.5, as would be expected in early trials with no information about which event is more likely. Notice that the proportion of predictions for the more likely event converges to 0.75, as predicted by probability matching, with very close matching from about trial 100.

In the 'pay/loss' treatment, 12 participants received 5¢ for each correct prediction, and they lost 5¢ for each incorrect prediction. The rate seems to converge to about 0.9.

A third 'pay' treatment offered a 5¢ reward for a correct prediction but no loss for an incorrect prediction. The results (not shown) are in between those of the other two treatments, and clearly above 0.75.

Clearly, then, incentives matter. Probability matching is not observed with incentives in this context.

Siegel's findings suggest a resolution to the paradox that rats are smarter than humans in binary prediction tasks. You cannot tell a rat to 'do your best': incentives such as food pellets must be used. Consequently, the choice proportions are closer to those observed with financially motivated human subjects. In a recent survey of over 50 years of probability-matching experiments, Vulkan (2000) separates those studies that used real incentives from those that did not, and he concluded that probability matching is generally not observed with real pay-offs. However, humans can still be surprisingly slow learners in this simple setting. For this reason, probability-matching data are particularly interesting for valuating alternative learning theories.

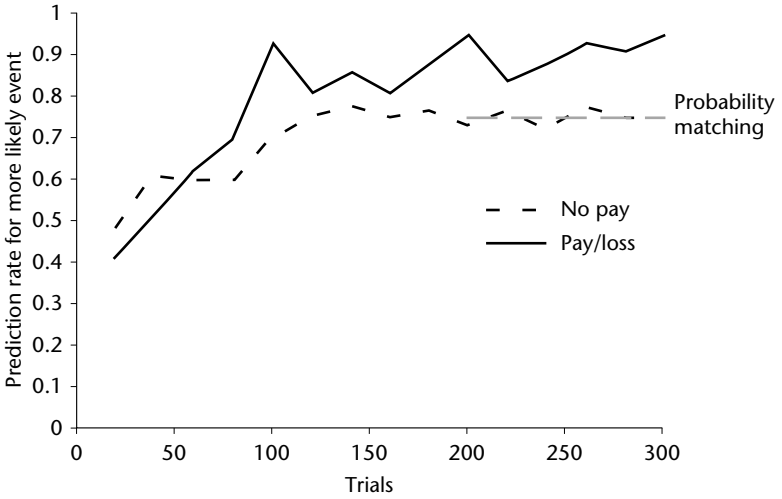


Figure 1. Prediction frequencies for an event that occurs with probability $\frac{3}{4}$ (Siegel *et al*., 1964). Each point plotted represents an average over 20 trials. The 'no pay' group were simply told to 'do your best'. The 'pay/loss' group received 5¢ for each correct prediction and forfeited 5¢ for each incorrect prediction.

Reinforcement Learning

One prominent theory of learning associates changes in behavior to the reinforcements actually received. Initially, when no reinforcements have been received, it is natural to assume that the choice probabilities for each decision are equal to one-half. Suppose that in the experiment there is a reinforcement of x for each correct prediction, and nothing otherwise. So if one predicts event L and is correct, then the probability of choosing L should increase. The extent of the behavioral change is assumed to depend on the size of the reinforcement. As the total earnings received for a particular decision increase, the probability of making that decision is assumed to increase. Suppose that event L has been predicted N_L times and that the predictions have sometimes been correct and sometimes not. Then the total earnings for predicting L, denoted by e_L, would be less than xN_L. Similarly, let e_R be the total earnings from the correct R predictions. One way to model the effect of total earnings for each decision on choice probabilities is to choose L or R with the following probabilities:

$$P(\text{choose L}) = \frac{\alpha + e_L}{2\alpha + e_L + e_R} \quad (1)$$

$$P(\text{choose R}) = \frac{\alpha + e_R}{2\alpha + e_L + e_R} \quad (2)$$

The parameter α determines how quickly learning responds to the reinforcements. Note that, as additional reinforcements are received, they are added into the relevant numerator, and to both denominators, to ensure that the probabilities sum to 1.

This kind of model might explain some aspects of behavior in probability-matching experiments with financial incentives. The choice probabilities would be equal initially, but a prediction of the more likely event will be correct 75% of the time, and the resulting asymmetries in reinforcement would tend to raise the prediction probability for that event, and the total earnings for this event would tend to be much larger than for the other event. If L is the more likely event, then e_L would be growing faster, so that e_R/e_L would tend to get smaller. Thus the probability of choosing L would tend to converge to 1.

This learning model can be simulated by using past accumulated earnings to compute choice probabilities. Then a random-number generator determines the actual choices. To make our data comparable with Siegel's data, we simulate decisions of a cohort of 1000 individuals for 300 periods, and calculate the 20-period choice averages for the more likely event.

The simulations were done for $\alpha = 5$ and $x = 1$. The value of α was chosen to create some initial inertia in behavior, which will tend to disappear after 40 or 50 periods. Setting α equal to 5 is analogous to having had each decision reinforced five times initially. Figure 2 shows simulated choice averages together with Siegel's original data. The simulated data are smoother, and start somewhat higher to start with, but the general pattern and final tendencies are similar. Erev and Roth (1998)

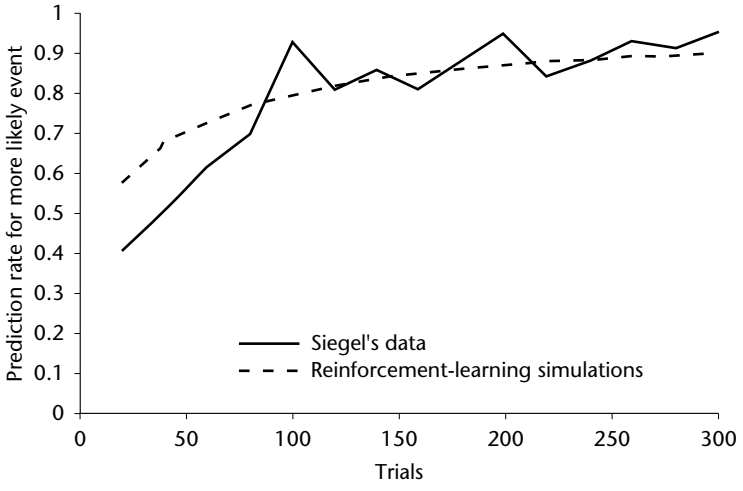


Figure 2. Data for Siegel's probability-matching experiment ('pay/loss' condition), with reinforcement-learning simulation data superimposed ($\alpha = 5$).

have used reinforcement learning to explain behavior in simple matrix games.

A Simple Model of Belief Learning

With reinforcement learning, beliefs are not explicitly modeled. An alternative approach that is more natural to economists is to model learning in terms of (Bayesian) updating of beliefs. Given the symmetry of Siegel's experimental setup, a person's initial beliefs ought to be that each event is equally likely, but the first observation should raise the probability associated with the event that was just observed. Moreover, the probability of event L should be an increasing function of the number N_L of times that this event has been observed, and a decreasing function of the number N_R of times that event R has been observed. Let N be the total number of observations to date. Then a standard belief-learning model is:

$$P(\mathrm{L}) = \frac{\beta + N_L}{2\beta + N} \tag{3}$$

$$P(\mathrm{R}) = \frac{\beta + N_R}{2\beta + N} \tag{4}$$

where $N = N_L + N_R$. Note that β determines how quickly the probabilities respond to the new information; a large value of β will keep these probabilities close to $\frac{1}{2}$. These formulae for calculating probabilities can be derived from Bayesian statistical principles (DeGroot, 1970, p. 160). In the early periods, the totals N_L and N_R might sometimes switch in terms of which one is higher, but the more likely event will soon dominate, and therefore $P(\mathrm{L})$ will be greater than $\frac{1}{2}$.

The beliefs determine the expected pay-offs (or utilities) for each decision, which in turn determine the decisions made. In theory, the decision with the highest expected pay-off should be selected with certainty. The prediction of the belief-learning model is, therefore, that all people will eventually predict the more likely event every time.

In an experiment, however, some randomness in decision-making might be observed if the expected pay-offs for the two decisions are similar. This randomness may be due to changes in emotions, calculation errors, selective forgetting of past experience, etc. Following Luce (1959), we introduce some 'noise' via a probabilistic choice model, where decision probabilities are positively but not perfectly related to expected pay-offs. Let π_L and π_R denote the expected pay-offs from choosing 'left' and 'right' respectively. Luce provided a set of axioms under which the choice probability is calculated as

$$P(\text{choose L}) = \frac{(\pi_L)^{1/\mu}}{(\pi_L)^{1/\mu} + (\pi_R)^{1/\mu}} \tag{5}$$

$$P(\text{choose R}) = \frac{(\pi_R)^{1/\mu}}{(\pi_L)^{1/\mu} + (\pi_R)^{1/\mu}} \tag{6}$$

The parameter μ is an 'error' parameter. It determines the sensitivity of choice probabilities to differences in expected pay-offs. In the limit when μ tends to zero, the decision with the higher expected pay-off is selected with probability 1. In the other extreme as μ gets large, behavior is random and independent of pay-offs.

In the probability-matching experiment, the expected pay-off of choosing 'left' is the reward (of 1, say) times the probability of 'left' that represents the person's beliefs. Thus the expected pay-off of 'left' is $P(\mathrm{L})$, and similarly the expected pay-off of 'right' is $P(\mathrm{R})$. $P(\mathrm{L})$ is greater than one-half if 'left' is more likely, and the error parameter μ determines how close the choice probability for the more likely event is to 1.

Figure 3 shows a simulation of the belief-learning model for $\beta = 20$. The thin solid line represents the average of the belief probabilities for the 12 simulated subjects. Notice that beliefs start close to one-half and converge to the true probability of the more likely event $\left(\frac{3}{4}\right)$. These beliefs determine expected pay-offs, and hence choice probabilities, via equations 5 and 6.

The dashed lines show the simulated average choice frequencies for three different levels of the error parameter. With high error ($\mu = 1$), the choice frequencies are close to the belief line, which would correspond to probability matching. This result is to be expected, since expected pay-offs are equal to belief probabilities. The denominator on the right-hand side of equations 5 and 6 is 1 when $\mu = 1$, and hence the probability of choosing 'left' equals π_L, which is equal to the belief probability.

As the error is reduced, the simulated choice frequencies move upwards towards the optimal level of 1. The top line, with $\mu = \frac{1}{3}$, converges to the level of about 0.9, which is close to the choice frequency observed by Siegel.

The simulations in Figure 3 were done for a cohort of size 12, to be comparable with Siegel's experiment. This allows us to see the degree of variation in the simulated data with a small group. In order to predict the average over a large number of individuals, we ran the simulation 1000 times, and the average proportions of choices for

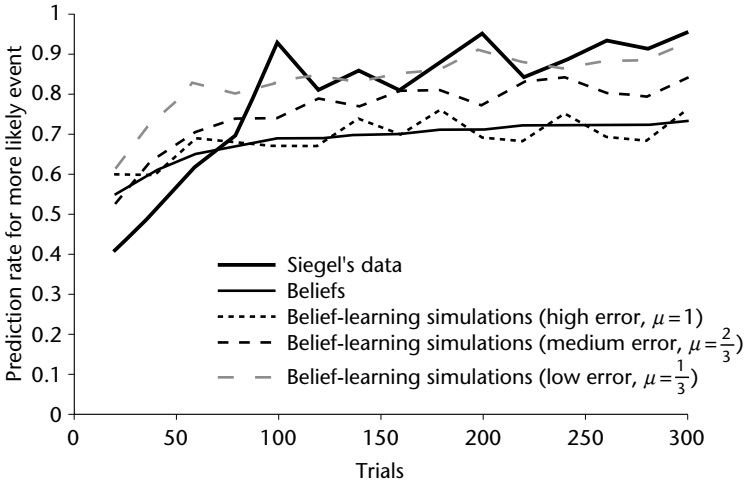


Figure 3. Data for Siegel's probability-matching experiment ('pay/loss' condition), with belief-learning simulation data superimposed ($\beta = 20$). The error parameter μ ranges from 1 (high error) to $\frac{1}{3}$ (low error).

the more likely event were: 0.76 for $\mu = 1$, 0.80 for $\mu = 0.67$, and 0.87 for $\mu = 0.33$.

Generalizations

Both of the learning models discussed above are somewhat simple, and this is part of their appeal. The reinforcement model builds in some randomness in behavior, and has the appealing feature that incentives matter. But it has less of a cognitive element. There is no reinforcement for decisions not made. For example, suppose that a person chooses L three times in a row (by chance) and is wrong each time. Since no reinforcement is received, the choice probabilities stay at 0.5 even after three incorrect predictions. This seems unreasonable. People do learn something in the absence of previous reinforcement, since they realize that making a good decision may result in higher earnings in the next round. Camerer and Ho (1999) have developed a generalization of reinforcement learning that contains some elements of belief learning. Roughly speaking, outcomes that are observed receive partial reinforcement even if nothing is earned.

These learning models can be enriched in other ways to obtain better predictions of behavior. For example, the sums of event observations in the belief-learning model weigh each observation equally. It may be reasonable to allow for 'forgetting' in some contexts, so that the observation of an event in the most recent trial may carry more weight than something observed a long time ago. This can be done by replacing sums with weighted sums. For example, if event L was observed three times, N_L in equation 3 would be 3, which can be thought of as $1 + 1 + 1$. If the most recent observation (listed on the right in this sum) is twice as prominent as the one before it, then the prior event would get a weight of one-half, and the one before that would get a weight of one-fourth, and so on. This type of 'recency' effect will be discussed below in the context of an interactive market game.

Finally, 'Luce's probabilistic-choice rule' (equations 5 and 6) is often replaced with the 'logit rule':

$$P(\text{choose L}) = \frac{\exp(\pi_L/\mu)}{\exp(\pi_L/\mu) + \exp(\pi_R/\mu)} \tag{7}$$

$$P(\text{choose R}) = \frac{\exp(\pi_R/\mu)}{\exp(\pi_L/\mu) + \exp(\pi_R/\mu)} \tag{8}$$

where μ is an error parameter as before. The Luce and logit rules are often similar in effect, and both are commonly used. The logit probabilities are unchanged when all pay-offs are increased by a constant, and the Luce probabilities are unchanged when all pay-offs are multiplied by a positive constant.

LEARNING AND PRICE DYNAMICS IN A MARKET GAME

We use a simple price competition example from Capra *et al*. (2002) to illustrate the effects of learning in an interactive setting. Consider a market game in which firms 1 and 2 simultaneously choose prices p_1 and p_2 in the range [60, 160] (units are cents). Demand is assumed to be a fixed total quantity ('perfectly inelastic'). The sales quantity of the firm with the lower price $p_{\min}$ is normalized to be

one, so the low-price firm earns an amount equal to its price. The high-price firm sells a 'residual' amount R, which is less than 1. The amount by which this residual is less than 1 indicates the degree of buyer responsiveness to price. The high-price firm has to match the lower price in order to make any sales, but some sales are lost because of the initially higher price. We assume that the high-price firm only earns $Rp_{\min}$, where $R<1$. In the event of a tie, the $1+R$ sales units are shared equally, so that each seller earns $(1+R)\,p_{\min}$.

As long as the high-price firm obtains less than half the market ($R<1$), the Nash equilibrium prediction is for both firms to set the lowest possible price of 60. To see this, note that at any common price, each firm has an incentive to undercut the other by a small amount to increase its market share. Therefore, the unique Nash equilibrium involves both firms charging the lowest possible price. The harsh competitive nature of the Nash prediction seems to go against the simple economic intuition that the degree of buyer inertia will affect the behavior of firms. When $R=0.8$, say, the loss from having the higher price is relatively small, and firms should be more likely to set prices above 60 when there is a small chance that rivals will do the same. Indeed, in the extreme case when $R=1$ it becomes a dominant strategy for both firms to choose the highest possible price of 160. While a standard Nash analysis predicts no change as long as $R<1$ (and then an abrupt change when $R=1$), it seems plausible that prices will gradually rise with R.

We ran an experiment based on this market game, using six cohorts of 10 subjects each. Each group of 10 subjects was randomly paired, with new partners in each of 10 periods. A period began with all subjects selecting a price in the interval [60, 160]. After these prices were recorded, subjects were matched, and each person was informed about the other's price choice. Pay-offs were calculated as described above: the low-price firm earned an amount equal to its price, and the high-price firm earned R times the lowest price. Three sessions were conducted with $R=0.2$ and three with $R=0.8$. Figure 4 shows the period-by-period average price choices. The upper solid line shows the average prices when buyers were relatively unresponsive ($R=0.8$), and the lower solid line shows average prices when buyers were relatively responsive ($R=0.2$). Recall that the Nash equilibrium was 60 for both treatments, as shown by the horizontal dashed line at 60. As intuition suggests, changes in the buyers' responsiveness has a large effect on price, even though the Nash equilibrium remains unchanged.

Notice that prices start high and stay high in the $R=0.8$ treatment, while prices decline before leveling off in the $R=0.2$ treatment. Standard economic models cannot explain either the levels or

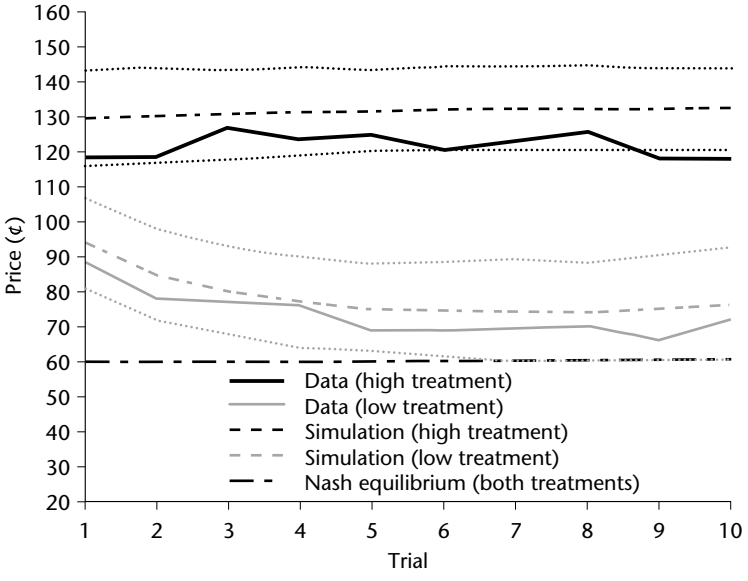


Figure 4. Data and simulations (plus or minus two standard deviations indicated by dotted lines) for a market game. In the 'high' treatment, buyers were relatively unresponsive to differences in price; in the 'low' treatment, buyers were relatively responsive. The simulations are based on a simple belief-learning model using a logit rule to determine probabilities.

the patterns of adjustment. Our approach is to consider a naive learning model in which players use observations of rivals' past prices to update their beliefs about others' future actions. In turn, the expected pay-offs based on these beliefs determine players' choice probabilities via a logit rule. This model was used to simulate behavior in the experiment.

To obtain a tractable model, the price range [60, 160] is divided into 101 one-cent categories. Players assign weights to each category and use observations of their rivals' choices to update these weights as follows: each period, all weights are 'discounted' by a factor ρ and the discounted weight of the observed category is increased by 1. In other words, the weight w of an observed category is updated as $w \to \rho\, w + 1$, whereas the other weights are updated as $w \to \rho w$. The belief probabilities in each period are obtained by dividing the weight of each category by the sum of all the weights. Hence, the learning parameter ρ determines the importance of new observations relative to previous information. Since the most recent observation gets a weight of 1, a lower value of ρ reduces the importance of prior history and increases recency effects.

Generally ρ will be between 0 and 1. When $\rho = 0$, the observations prior to the most recent one are ignored, and the model is one of best response to the previously observed price (Cournot dynamics). At the other extreme, when $\rho = 1$, the model reduces to 'fictitious play', in which each observation is given equal weight, regardless of the number of periods that have elapsed since. For intermediate values of ρ, the weight given to past observations declines geometrically over time.

The expected pay-off for player i choosing a price in category j is denoted by $\pi_i^{\mathrm{e}}(j|\rho)$. This determines player i's decision probabilities via the logit rule

$$P_i(j|p) = \frac{\exp\left(\pi_i^{\mathrm{e}}(j\,|\,\rho)/\mu\right)}{\sum_{k=1}^{101} \exp\left(\pi_i^{\mathrm{e}}(k\,|\,\rho)/\mu\right)},\ j = 1, \ldots, 101 \qquad (9)$$

Choice probabilities and expected pay-offs depend on the learning parameter. In this dynamic model, beliefs, and hence choices, depend on the history of what has been observed up to that point. Since individual histories are realizations of a stochastic process, the predictions of this model will be stochastic and can be analyzed with simulation techniques.

The structure of the computer simulation program matches that of the experiment reported below: for each session, or 'run', there are 10 simulated subjects who are randomly matched in a sequence of 10 periods. We specify initial prior beliefs for each subject to be uniform on the integers in the set [60, 160]. These priors determine expected pay-offs for each price, which in turn determine the choice probabilities via the logit rule in equation 9. The simulation begins by determining each simulated player's actual price choice for period 1 as a draw from the logit probabilistic response to the pay-offs for priors that are uniform on [60, 160]. The simulated players are randomly divided into five pairs, and each player 'sees' the other's actual price choice. These price observations are used to update players' beliefs using the naive learning rule explained above, with a learning parameter $\rho = 0.72$ (which was estimated from the data). The updated beliefs, which become the priors for period 2, will not all be the same if the simulated subjects encountered different price choices in period 1. The process is repeated, with the period-2 priors determining expected pay-offs, which in turn determine the logit choice probabilities, and hence the observed price realizations for that period. The whole process is repeated for 10 periods.

Figure 4 shows the sequences of average prices obtained from 1000 simulations, together with dotted lines indicating two standard deviations of the average. These simulation results predict that average prices decline in the $R = 0.2$ treatment and stay the same in the $R = 0.8$ treatment, as observed in the data. Thus, the learning model explains the salient features of the experimental data: both the directions of adjustment and the steady-state levels.

STOCHASTIC LEARNING EQUILIBRIUM

Next we consider what the learning model implies about the long-run steady-state distribution of price decisions. In particular, will learning generate a price distribution that corresponds to some equilibrium?

At any point in time, different people will have different experiences or histories. These differences may be due to the randomness in individuals' decisions or to randomness in the random matching. For each person, the history of what they have seen will determine a probability distribution over their decisions. This mapping of histories to decision probabilities may be direct, as in reinforcement learning. Alternatively, histories may generate beliefs, which in turn produce decisions via a probabilistic choice rule. The decisions made are then appended to the existing histories, forming new histories. Because of the randomness in

decision-making, there will be a probability distribution over all possible histories.

In a steady state of the learning model, the probability distribution over histories remains unchanged over time. The 'stochastic learning equilibrium' is defined as the steady-state probability distribution over histories. This formulation is general and includes many learning models as special cases. Goeree and Holt (2002) show that this equilibrium always exists when there is a finite number of decisions and players have finite (but possibly long) memories.

For example, consider the extreme case where a person can only remember the two most recent observations in the probability-matching experiment. There are four possible remembered histories: LL, LR, RL, and RR, with exogenously determined probabilities of $\frac{3}{4}\times\frac{3}{4}$, $\frac{3}{4}\times\frac{1}{4}$, $\frac{1}{4}\times\frac{3}{4}$, and $\frac{1}{4}\times\frac{1}{4}$, respectively. A stochastic learning equilibrium in this context would be a vector of transition probabilities between these states. The formulation of this model in terms of histories (instead of single-period choice distributions) allows the possibility of dynamic effects such as cycles and endogenous learning rules. The focus on histories (sequences of vectors of players' decisions) also facilitates the proof that a stochastic learning exists under fairly general conditions.

Given a specific learning rule, it is possible to determine the stochastic learning equilibrium. To illustrate, consider the market price game under two extreme conditions, fictitious play ($\rho = 1$) and Cournot best response ($\rho = 0$). Since there is no 'forgetfulness' in fictitious play, any steady-state distribution of decisions will eventually be fully learned by all players, i.e. the empirical frequencies of price draws from the distribution will converge to that distribution. In this case, each player is making a logit probabilistic best response to the empirical distribution, and these best responses match the empirical distribution. Notice that all players must have identical beliefs in this equilibrium. This is known as a 'quantal response equilibrium' as defined by McKelvey and Palfrey (1995.)

When $\rho = 0$, a player's history is simply the most recent observation, and beliefs are necessarily different across players. These differences in individuals' beliefs add extra randomness into the steady state. Figure 5 illustrates these observations for the high-R treatment of the price-choice game. The solid line represents the stochastic learning equilibrium with an infinite memory ($\rho = 1$), and the dashed line represents the price distribution for the case of one-period memory ($\rho = 0$). Both of these distributions are hump-shaped, with means near the price average observed in the experiment. The implied distribution of price choices is flatter and more dispersed for the case of one-period memory, since beliefs are being moved around by recent observations, which introduces extra randomness. Both cases, however, capture the salient feature of the prices observed in the high-R treatment of the market experiment. In particular, price averages are more than twice as high as the unique Nash equilibrium prediction.

When maximum-likelihood techniques are used to estimate the learning parameter from the choices made by the human subjects, the resulting estimate ($\rho = 0.72$) is intermediate between the extreme cases shown in Figure 5, and the resulting steady-state price distribution will also be intermediate. In fact, the weights determined by powers of 0.72 decline very quickly, and the equilibrium price distribution is rather close to the flatter ($\rho = 0$) case, as can be confirmed with computer simulations. Simulations of individual cohorts of 10 subjects (not shown) show the same up-and-down patterns exhibited by comparably sized cohorts of humans. The simulation averages shown in Figure 4 track the main features of the human data: prices start high and stay high in one treatment, and they start high and decline towards the Nash prediction in the other. Thus, computer simulations of learning models can explain data patterns that are not predicted with standard equilibrium techniques. In fact, we ran the computer simulations before we ran the experiments with human subjects, using the learning and error parameter estimates from a previous experiment (Capra *et al*., 1999). The simulations helped us select two values of the treatment parameter R that would ensure that there would be a strong treatment effect that is not predicted by the Nash equilibrium.

SUMMARY

The learning models presented here were pioneered by Bush and Mosteller (1955), and the stochastic choice models were introduced by the mathematical psychologist Luce (1959) and others. These techniques no longer receive much attention in the psychology literature, where the main interest is in theories of learning, biases, and heuristics that have a richer cognitive content. Yet they have yielded important insights in explaining economics experiments where the anonymity and repetitiveness of market interactions dominate. The incorporation of insights from the literature on heuristics and biases may also prove to be valuable in the future.

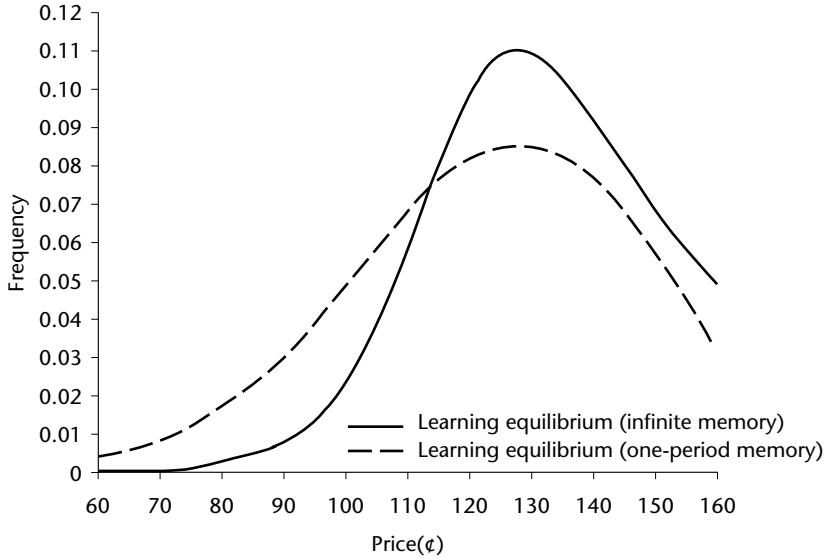


Figure 5. Stochastic learning equilibrium distribution of prices in the price-choice game for $R = 0.8$.

The belief- and reinforcement-based learning models depend on past history in a somewhat mechanical manner. In contrast, some laboratory experiments provide situations in which learning seems to proceed in response to cognitive insights. For example, if one subject observes that another has earned more money, the first person may decide to try to imitate the other. Consider a market in which subjects choose 'production quantities' simultaneously, with the advance knowledge that the price at which all production is sold will be a decreasing function of the total quantity produced. Since all output is sold at the same price, the person with the highest quantity will have the highest sales revenue. To the extent that high revenues translate into high profits, the sellers with low quantities may be tempted to imitate the high-output strategies of those who have higher earnings. In this context, the process of imitating high-production sellers can cause the total production to be high. The implications of imitation learning have been studied in a series of recent economics experiments. Offerman and Sonnemans (1998), for example, find evidence that learning is induced to some extent both by imitation of others and by one's own experience.

When some people are following predictable learning patterns, it may be advantageous to try to manipulate others' beliefs via 'strategic teaching'. For example, a dominant seller may punish new entrants by expanding production quantity and thereby driving prices down. This behavior may be intended to 'teach' potential rivals not to enter. This is an important area for future research, and it is complicated by the fact that the person doing the teaching should have a mental model of the others' learning processes.

In some economic situations, learning may occur as a sudden realization that some different decision will provide higher earnings or will avoid losses. A first step in the study of this type of learning may be to measure biological indicators of mental activity for economic tasks that may involve sharp changes in behavior or attempts to anticipate others' decisions (McCabe *et al*., 2000).

References

Bush R and Mosteller F (1955) *Stochastic Models for Learning*. New York, NY: Wiley.

Camerer C and Ho T-H (1999) Experience weighted attraction learning in normal-form games. *Econometrica* **67**: 827–874.

Capra CM, Goeree JK, Gomez R and Holt CA (1999) Anomalous behavior in a traveler's dilemma? *American Economic Review* **89**(3): 678–690.

Capra CM, Goeree JK, Gomez R and Holt CA (2002) Learning and noisy equilibrium behavior in an experimental study of imperfect price competition. *International Economic Review* **43**(3): 613–636.

DeGroot MH (1970) *Optimal Statistical Decisions*. New York, NY: McGraw-Hill.

Erev I and Roth AE (1998) Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88**(4): 848–881.

Fantino E (1998) Behavior analysis and decision making. *Journal of the Experimental Analysis of Behavior* **69**: 355–364.

Goeree JK and Holt CA (2002) Stochastic learning equilibrium. Working paper, University of Virginia.

[Presented at the Economic Science Association Meetings in New York City, June 2000.]
Luce D (1959) *Individual Choice Behavior*. New York, NY: Wiley.
Mackintosh NJ (1969) Comparative psychology of serial reversal and probability learning: rats, birds, and fish. In: Gilbert R and Sutherland NS (eds) *Animal Discrimination Learning*, pp. 137–167. London: Academic Press.
McCabe K, Coricelli G, Houser D, Ryan L and Smith VL (2000) Other minds in the brain: a functional imaging study of 'theory of mind' in two-person exchange. Working paper, University of Arizona.
McKelvey RD and Palfrey TR (1995) Quantal response equilibria for normal form games. *Games and Economic Behavior* **10**: 6–38.
Offerman T and Sonnemans J (1998) Learning by experience and learning by imitating successful others. *Journal of Economic Behavior and Organization* **34**(4): 559–575.
Siegel S, Siegel A and Andrews J (1964) *Choice, Strategy, and Utility*. New York, NY: McGraw-Hill.
Vulkan N (2000) An economist's perspective on probability matching. *Journal of Economic Surveys* **14**(1): 101–118.

Further Reading

Chen Y and Tang FF (1998) Learning and incentive compatible mechanisms for public goods provision: an experimental study. *Journal of Political Economy* **106**: 633–662.
Cooper DJ, Garvin S and Kagel JH (1994) Adaptive learning vs. equilibrium refinements in an entry limit pricing game. *RAND Journal of Economics* **106**(3): 662–683.
Fudenberg D and Levine DK (1998) *Learning in Games*. Cambridge, MA: MIT Press.
Goeree JK and Holt CA (1999) Stochastic game theory: for playing games, not just for doing theory. *Proceedings of the National Academy of Sciences* **96**: 10564–10567.

Economics, Experimental Methods in

Advanced article

Vernon L Smith, George Mason University, Arlington, Virginia, USA

CONTENTS

Experimental economics uses laboratory methodology to examine motivated human behavior and its interpretation in small group interactive games, and in bidding, auctioning, and market institutions. Subjects earn cash payments depending upon their joint interactive decisions, and the rules governing their interactions.

INTRODUCTION

It is useful to distinguish three complex self-ordering systems: the internal order of the mind (Hayek, 1952); the external order of social exchange between minds (McCabe and Smith, 2001); and the extended order of cooperation through markets and other cultural institutions (Hayek, 1988). Our concern here is with the first two.

We focus on social exchange because it was the cooperative behaviors registered in two-person anonymous interaction that first alerted experimental economists to a significant class of phenomena that violate certain static equilibrium concepts in game theory. Game theory is about strategic interdependent choice when the pay-off benefit to each of two or more people depends jointly on their decisions. These refutations generated alternative interpretations of that theory, and motivated questions directly concerned with the internal order of the mind. They are now leading to the study of the neural correlates of human decision-making in two-person strategic interactions.

Why do we study anonymous interactions? First, our theoretical model of single-play games assumes strangers without a history or a future, and anonymity provides the required control for testing this theory. Also, it is well documented that the effects of face-to-face interaction hide more subtle procedural effects in yielding cooperative outcomes (Hoffman and Spitzer, 1985). As will be illustrated in the experiments reported below, the anonymity condition provides great scope for exploring the natural human instinct for social exchange, and how it is affected by context, reward, and procedure.

Why Should Context Matter?

Context matters because all memory involves relationships and is associative. For example, priming experiments use cues to improve retrieval from memory because of associations between the cue and the stimulus. People perform better at completing word fragments (filling in missing letters) on words they have observed beforehand in lists, even if they are not told that the words appeared in the earlier list. Some priming effects are almost equally strong whether the interval between the original list and the test is a matter of hours or days. Furthermore, being able to state that a word was seen before does not correlate with improved completion performance. How one perceives a current task depends upon unconscious cues to past experience that are triggered by the context of the task. Two decision tasks with the same underlying logical structure may lead to different responses because they are embedded in different contexts and invoke different memory experiences (Gazzaniga *et al.*, 1998, pp. 258–261). This is because of the fundamental, though nonintuitive, nature of perception.

In the early 1950s Hayek articulated certain principles of perception, which are consistent with current neuroscientific understanding:

1. It is incorrect to suppose that experience is formed from the receipt of sensory impulses reflecting unchanging attributes of external objects in the physical environment. Rather, the process by which we learn about the external environment involves a relationship between current conditions and our past experience of similar conditions (Hayek, 1952, p. 165).
2. Categories are formed by the mind according to the relative frequency with which current perception and memory (past perceptions) concur (p. 64). What are

stored in memory are external stimuli modified by processing systems whose organization is conditioned by past experience of stimuli. All perception is produced by memory.

3. This leads to a 'constant dynamic interaction between perception and memory, which explains the … identity of processing and representational networks of the cortex that modern evidence indicates' (Fuster, 1999, p. 89). 'Although devoid of mathematical elaboration, Hayek's model clearly contains most of the elements of those later network models of associative memory … [and] comes closer, in some respects, to being neurophysiologically verifiable than those models developed 50 to 60 years after his.' (Fuster, 1999, pp. 88–89).

Hayek's model is incomplete, and did not influence the research it anticipated, but it captures the idea that perception is self-organized, created from abstract function combined with experience. This is relevant to the question of why context is important in the experiments reported below.

MENTAL MODULES AND EVOLUTIONARY PSYCHOLOGY

Evolutionary psychologists argue that the mind consists of circuitry, or interactive modules, that are specialized for vision, for language learning, for socialization, and for a host of other functions (Cosmides and Tooby, 1992). Language and socialization, which are of recent evolutionary origin, are hypothesized to have evolved in the two to three million years during which humans subsisted as hunter–gatherers. It is in this evolutionary environment of our ancestors (EEA) that humans developed mechanisms of social exchange in which assistance, meat, favors, information and other services and valuables were traded across time. This is evident in extant hunter–gatherer societies (e.g. the Ache of Paraguay) in which the product of the hunt is widely shared within the tribe as well as within the nuclear and extended family. In a world without refrigeration and only rare success in hunting, this made sense: if I am lucky in the hunt today, I share the meat with others; and tomorrow, when I fail to make a kill, you share your kill with me and with others. In contrast, the products of gathering – fruit, nuts, roots – depend more on effort than on luck, are much more predictable from day to day, and are shared only in the nuclear family where effort can be closely monitored. Traditions of sharing across time provide gains from exchange that support limited forms of specialization: women and children do the gathering; adult men do the hunting; older men make tools, advise in the hunt, and assist in gathering. Such patterns (subject to numerous variations) are common in tribal communities.

But delayed exchange across time based on reciprocity is hazardous. Favors cannot be retracted, and you might systematically fail to return mine. Without money – a recent invention not available in the EEA – it is adaptive to develop some skill in making judgments about who can or cannot be trusted. This puts a premium on 'mindreading', the ability to infer mental states from the words and actions of others. The minimal mental equipment required is a 'cheater-detector module' for social exchange. The results of experiments designed by Cosmides (1985) are consistent with the hypothesis that the human mind is attuned to detecting cheaters on perceived social contracts. With the development of language, our instincts for cheater detection were enhanced by gossip: comparing notes to determine those with good reputations for returning favors. Gossip, like language and reciprocity, is a human universal, an activity pursued in all human communities. None of this mental equipment was the product of our reason: rather, it was the unconscious product of the biological and cultural evolution that distinguished us from other primates.

Evolutionary psychologists see an inevitable tension between who we are (based on what we have inherited from the EEA) and the demands made on us by the world since the agricultural revolution 10,000 years ago. One account of this tension was articulated by Hayek: 'Part of our present difficulty is that we must constantly adjust our lives, our thoughts and our emotions, in order to live simultaneously within different kinds of orders according to different rules. If we were to apply the unmodified, uncurbed rules (a caring intervention to do visible good) … of the small band or troop, or of, say, our families, to … our wider civilization (the extended order of the market), as our instincts … often make us wish to do, *we would destroy it*. Yet if we were always to apply the rules of the extended order (action in the self-interest within competitive markets) to our more intimate groupings, *we would crush them*. So we must learn to live in two sorts of world at once.' (Hayek, 1988, p. 18).

This observation raises questions about game theory, which postulates that the players are strictly self-interested and that this condition is common knowledge to all the players. How is action in the self-interest affected by whether the anonymous players are in an *n*-person market or a two-person interactive game? How do the players come to have 'common knowledge'? Does the

procedural and instructional context of a two-person game affect cooperation by influencing how the players perceive the game? It is most natural to investigate such questions in experimental environments where monetary pay-offs, context, and interaction procedures can be controlled.

EXPERIMENTAL PROCEDURES

The experiments reported below show that context is important in the decision behavior we observe. This is to be expected, given what is known about the autobiographical character of memory and the interaction between current and past experience in creating memory. Below are reported behavioral results in two-person sequential-move game trees in which each pair plays once and only once through the move sequence defined by the tree, and the game is completely known to the subjects. However, the instructions for the experiments do not (except in systematic treatments) use words like 'game', 'play', 'player', 'opponent', or 'partner'; rather, reference is made to the 'decision tree', 'decision maker' 1 (DM1) or 2 (DM2), 'your counterpart', and other terms designed to provide a baseline context.

Your experience as a subject in a typical experiment might be as follows. You have been recruited to participate in an economics experiment for which you will be paid \$5 (or more, in some cases) for arriving on schedule, plus the amount in cash that you earn from your decisions, to be paid to you at the end. You arrive, sign in, receive \$5, and are assigned to a computer terminal in a large room with 40 stations. There are 11 other people, well spaced throughout the room. Each station is a partially enclosed booth, making it very easy to maintain your privacy. After everyone has arrived you log in to the experiment as directed on your screen. You read through the instructions for the experiment at your own pace, respond to the questions, and learn that in this experiment you are matched anonymously with another person in the room, whose identity you will never know, and vice versa. This does not mean that you know nothing about that person: it may seem evident that he or she is another 'like' person – for example, an undergraduate – with whom you may feel more or less of an in-group identity. Obviously, you bring with you a host of past experiences and impressions that you are likely to apply to the experiment.

Ultimatum Game Experiments

Consider the following simple two-stage two-person game. A fixed sum of money m is provided by the experimenter (e.g. m might be 10 one-dollar bills, or 10 ten-dollar bills). Player 1 moves first, proposing that a portion $x \leq m$ of the money be offered to player 2, player 1 retaining $m - x$. The offer is a 'take it or leave it' ultimatum. Player 2 then responds by either accepting the offer, in which case the experimenter pays $m - x$ to player 1 and x to player 2, or rejecting the offer, in which case each player receives 0.

Now consider four different instructional–procedural contexts in which an ultimatum game with this underlying abstract structure is played. In each case, imagine that you are the first mover (player 1 in the above abstract form). (See Hoffman *et al.* (2000) for instructional details, and for references to the literature and origins of the ultimatum game.)

Context 1: 'divide \$10'

In the first context, the instructions state that you and your anonymous counterpart have been 'provisionally allocated \$10'. Your task is to 'divide' the \$10 using the following procedure. You have been randomly assigned to the role of first mover. You (as person A) are asked to complete boxes (4) and (5) of the proposal form shown in Figure 1. The form then goes to your counterpart (person B) who checks 'Accept' or 'Reject'.

In this version, the \$10 consists of 10 one-dollar bills. In another version, there is \$100 (10 ten-dollar bills) to be divided.

Context 2: 'contest entitlement'

In the second context, each of the 12 people in the room takes the same general-knowledge quiz (10 questions). The results are used to determine the positions of persons A and B in each pairing. Your score is the number of questions answered

(1) Identification number..... #A

(2) Paired with..... #B

(3) Amount to divide..... \$10

(4) Person B receives .....

(5) Person A receives (3)–(4) .....

(6) Accept ☐ Reject ☐

Figure 1. Proposal form for an ultimatum game experiment using the 'divide \$10' context. You (as person A) are asked to complete boxes (4) and (5); the form then goes to your counterpart (person B) who checks 'Accept' or 'Reject'.

correctly, with ties broken in favor of the person who finished the quiz fastest. The scores are ranked from 1 (highest) to 12 (lowest). Those ranked from 1 to 6 will have 'earned' the right to be person A; the other six will be person B.

Context 3: 'exchange'

In the third context, person A is a 'seller' and B is a 'buyer'. A table lists the profit of the seller and of the buyer for each possible price ($0, $1, $2, …, $10) charged by the seller if the buyer chooses to buy. The profit of the seller is equal to the price chosen; the profit of the buyer is $10 minus the price. The profit of each is zero if the buyer refuses to buy at the price chosen by the seller.

Context 4: 'contest–exchange'

The fourth context combines the second and third: 'sellers' are selected by a general-knowledge quiz. In one version the total amount is 10 one-dollar bills; in another, it is 10 ten-dollar bills.

Results of Ultimatum Game Experiments

The game-theoretic concept of sub-game perfect equilibrium (SPE) yields the same prediction in all of these versions of the ultimatum game (Selten, 1975): player 1 offers the minimum positive unit of account ($1 if $m = \$10$, $10 if $m = \$100$), and player 2 accepts the offer. The analysis assumes that each player is self-interested in the sense of always choosing the largest of two pay-offs for himself or herself; that this condition is common knowledge for the two players; and that player 1 applies backward induction to the decision problem faced by player 2, conditional on player 1's offer. Thus player 1 should reason that any positive pay-off is better than zero for player 2, and therefore, player 1 need only offer the minimum positive amount.

But there are other models of decision for games like the ultimatum. A problem with the above analysis is that, perhaps depending on context, the ultimatum interaction may be interpreted as a social exchange between any two anonymously matched players who normally read intentions into the actions of others (Baron-Cohen, 1995). Suppose that the ultimatum game is perceived as a social contract in which player 2 has a ('fair claim') entitlement to more than the minimum unit of account; then an offer of less than the perceived entitlement (say, only $1, or perhaps even $2 or $3) may be rejected by some players 2. Player 1, reading this potential mental state of player 2 (e.g. by imagining what he or she would do in the same circumstance), might then offer substantially more than $1 to ensure acceptance.

Observe that in context 1, the original $10 is allocated imprecisely to both players, and does not clearly belong to either person A or B. Further, a common interpretation of the word 'divide' involves the separation of some divisible quantity into equal parts. Moreover, in western culture the use of a lottery or other random device is recognized as a standard mechanism for 'fair' or equal treatment. Hence, the instructions can be interpreted as suggesting that the experimenter is engaged in a 'fair' treatment of the subjects. This can serve as a strong, albeit unconscious, cue that the subjects ought to be 'fair' in their treatment of each other.

By contrast, context 2 deliberately introduces a contest procedure, before the game itself, in which those who score the highest earn the right to be person A, and those who score the lowest will be person B. In this treatment, nothing is said about who has been initially allocated the money, and the word 'divide' is not used. Rather, person A must choose how much person B is to receive, and person B must choose to accept or reject the proposal. Consequently, the instructions may cue some norm of 'just desert' based on test performance.

In context 3, the abstract ultimatum game is embedded in a transaction between a buyer and a seller. In such exchanges, buyers (in western culture) do not normally question the right of the seller to move first by quoting a price, nor that of the buyer to respond with a decision to buy or not to buy.

Context 4 combines the implicit 'property right' norm of a seller with an explicit mechanism whereby subjects 'earn' the privilege of being the seller in a contest whose outcome provides the same opportunity for all participants, depending on their general knowledge. This treatment introduces the 'equal opportunity' norm, as opposed to 'equality of outcome'.

Table 1 summarizes the results from two different studies of ultimatum game bargaining with stakes of either 10 one-dollar or 10 ten-dollar bills, where the number of pairs of players varies from 23 to 27. Note that 'divide' with random entitlement corresponds to context 1; 'divide' with earned entitlement to context 2; 'exchange' with random entitlement to context 3; and 'exchange' with earned entitlement to context 4.

Comparing 'divide $10' with 'divide $100' under random entitlement, we observe a trivial difference

Table 1. Mean percentage offered in ultimatum games, by context treatment. Data from Hoffman *et al*. (1996) and (1994)

		$10 stakes 'Divide'	*$100 stakes 'Exchange'*	*'Divide'*	*'Exchange'*
Random entitlement	Mean offer	43.7%	37.1%	44.4%	(n/a)
	N	24	24	27	(n/a)
	Rejection rate[a]	8.3%	8.3%	3.7%	(n/a)
Earned entitlement	Mean offer	36.2%	30.8%	(n/a)	27.8%
	N	24	24	(n/a)	23
	Rejection rate[a]	0	12.5%	(n/a)	21.7%

[a]Percentage of the *N* pairs in which the second player rejects the offer of the first.

in the amount offered between the low stakes (43.7%) and the high stakes (44.4%). There is no significant difference in the rate at which offers are rejected (8.3% and 3.7% respectively).

When 'exchange' is combined with an earned entitlement, the increase in stakes seems to lower the offer percentage from 30.8% for $10 stakes to 27.8% for $100 stakes, but this difference is within the normal range of sampling error using different groups of subjects and is not significant. Surprisingly, this minuscule decline in the mean offer correlates with an increase in the rejection rate from 12.5% to 21.7%. In the high-stake game, three out of four subject players 1 offering $10 are rejected, and one offer of $30 is rejected. As we shall see below in other games, this behavior is associated with a strong human propensity to incur personal cost to punish those who are perceived as cheaters, even under strict anonymity (as in Cosmides, 1985).

Comparing the 'divide' and 'exchange' conditions with random entitlement and $10 stakes, the offer percentage declines from 43.7% to 37.1%, and comparing the 'divide' conditions with random and earned entitlement and $10 stakes the offer percentage declines from 43.7% to 36.2%. Both reductions are statistically significant. Even more significant is the reduction from 43.7% to 30.8% with the 'exchange' condition and earned entitlement. Moreover, in all four of these contexts the rejection rate is null or modest (0 to 12.5%).

The small proportion of offers rejected (except when the stakes are $100 in the 'contest–exchange' context, where the mean offers decline to 27.8%) indicates that players 1 generally read their counterparts well and offer a sufficient amount to avoid being rejected. The one exception shows that trying to push back the boundary, even if it seems justified by the higher stakes, may provoke rejections.

One obvious conclusion from these data is that the effect of context on behavior cannot be ignored in the ultimatum game: the percentage offered varies by over a third between the highest (44%) to the lowest (28%) measured values. Studies of cross-cultural variation in ultimatum offers show a comparable variation. Thus, a comparison of two hunter–gatherer and five modern cultures reveals a variation from a maximum of 48% (Los Angeles subjects) to a minimum of 26% (Machiguenga subjects from Peru) (Heinrich, 2000). These comparisons attempted to control for instructional differences across different languages, but of course this is inherently problematic in that one cannot be sure that the translations, or the procedures for handling the subjects, completely control for context across cultures. Nor can it be assumed that the pay-offs are subjectively comparable across currencies.

The instructional comparisons also call into question the extent to which one can define what is meant by 'unbiased' instructions. Some results may be robust with respect to instructional changes, but this can only be established empirically, since we know little about the sources of behavioral variation due to context. Indeed, unless such robustness is established no claims can be made concerning the relative 'neutrality' of instructions, and the extent to which differences in behavior can be attributed to differences between cultures.

Because of the nature of perception and memory, we should expect context to be an important factor. In the ultimatum game, the variation of observed results with systematic instructional changes designed to alter context shows clearly that context can and does matter.

Trust Games

Ultimatum games have been studied extensively, but because of their simplicity they leave unanswered many questions about what underlies the behavior manifest in them. For example, one

cannot vary independently the cost of player 2's rejection of player 1's offer. The game is constant-sum, and is inherently confrontational: neither player can take action that increases the total gains from the transaction, and therefore the interpretation of the game as an exchange is limited.

We turn therefore to a somewhat richer class of two-person extensive-form trust games in which the return to equilibrium play, cooperation, defection, and the prospect of costly punishment of defection can be studied in a richer parameter space than that of the ultimatum game.

Figure 2(a) shows a trust game tree. Play starts at the top, node x_1, with player 1. Player 1 can move right; this stops the game, yielding \$7 to player 1 and \$14 to player 2. Alternatively, player 1 can move down, in which case player 2 has to choose a move at node x_2. If player 2 moves right, each player gets \$8. If player 2 moves down, player 1 can then move right at node x_3, yielding \$10 for each, or down, yielding \$12 for player 1 and \$6 for player 2.

The SPE is \$8 for each player. This is because at node x_1 player 1 can look ahead (use 'backward induction') to see that if play reaches node x_3 player 1 will want to move down. But player 2, also using backward induction, will see that at node x_2 player 2 should move right. Since this yields a higher pay-off to player 1, at node x_1 player 1 should move down.

The SPE outcome would prevail by the logic of self-interested players who always choose dominant strategies. There are other behavioral possibilities, however, depending on whether other preferences or perceptions of the interaction are applied.

If player 1 has other-regarding preferences (altruism, or utility from the other's pay-off), and is willing to incur some cost to greatly increase the pay-off to player 2, player 1 may move right at x_1. That way, at a cost of \$1, player 1 can increase player 2's pay-off by \$6, compared with the SPE. Thus, player 1 need have only a modest preference for an increase in player 2's welfare in order to move right.

At x_2, player 2 may move down, signaling to player 1 that such a move enables both to profit, provided that at x_3 player 1 cooperates by reciprocating player 2's favor. Alternatively, at x_3 player 1 can defect, by choosing the dominant strategy and moving down.

Figure 2(b) shows the tree for a punishment version of the trust game shown in Figure 2(a). The trees are identical except that at node x_3, player 1 chooses between the cooperation pay-off and passing back to player 2 at node x_4. Now player 2 decides whether to accept the defection or, at a cost, punish player 1 for the defection. By backward induction, the SPE is the same in the punishment version. The cooperation outcome can be justified (as a Nash equilibrium) only if the threat of punishment by player 2 at node x_4 is credible. But under the anonymity conditions, with no capacity to communicate, such a threat is not credible.

The outcome frequencies for the trust game ($N = 26$ pairs), and for the trust–punishment game ($N = 29$ pairs) are summarized in Figure 3.

In neither game is there a single case of a player 1 choosing the altruism outcome: all choose to pass to player 2, seeking a higher pay-off for themselves, and being content to give player 2 a much smaller pay-off than would be achieved by altruism.

The sub-game perfect equilibrium is chosen by 54% of the pairs in the trust game, and 55% in the trust–punishment game. Thus, there is no significant difference in behavior between the two games in terms of the frequency with which players 2 offer to cooperate by passing to players 1 at x_2.

There is, however, a considerable difference in the response of players 1 to the offer to cooperate: only 50% cooperate in the trust game, while 85% cooperate when facing the prospect of punishment for defection.

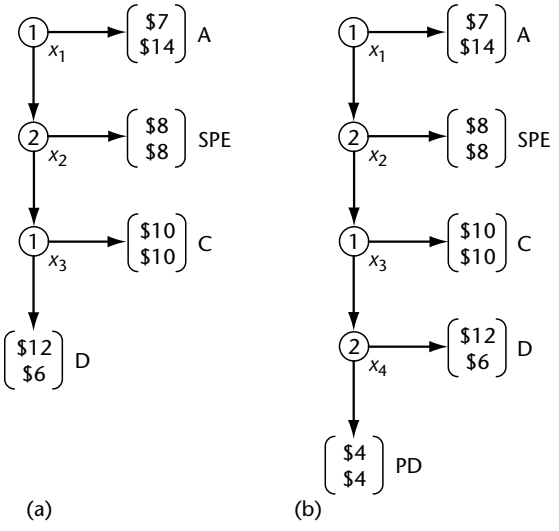
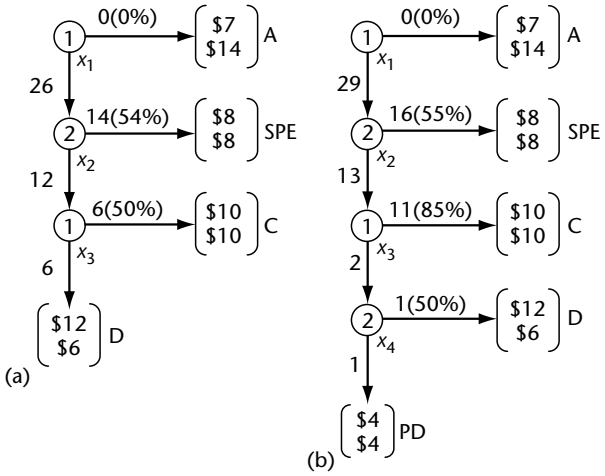


Figure 2. Trust game trees, (a) without punishment and (b) with punishment. At each terminal node the pay-off to player 1 is shown above the pay-off to player 2.



A: Altruism
SPE: Sub-game perfect equilibrium
D: Defection
C: Cooperation
PD: Punish defection

Figure 3. Experimental outcomes for the (a) trust and (b) trust–punishment games shown in Figure 2. A total of 26 subject pairs took part in the trust game and 29 in the trust–punishment game. The figures beside the arrows indicate the number of pairs following each route through the game tree and the percentage moving right at each decision node. The data are from McCabe *et al*. (2000). Source data for larger trees have been trimmed to eliminate rare outcomes, with commensurate reduction in sample size (from 30 to 29 in the punishment version).

Across the two games, why do nearly half of the players 1 forgo the sure SPE payoff in favor of the risky prospect of cooperation? McCabe and Smith (2001) argue that humans are eminently adapted for social exchange, or reciprocity among the individuals that constitute the small groups that form our primary networks of relationships. We constantly trade favors, services and assistance, with little conscious awareness of these trading relationships that are so much a part of our humanity. McCabe and Smith postulate an implicit mental accounting system for keeping track of trustworthy trading partners. This accounting system is part of the framework of our friendships and social connections.

Reciprocity is a human universal, characteristic of all cultures, as is the use of a spoken language. Like language, the form of reciprocity varies across cultures, but its common functionality is to produce gains from exchange. Smith (1998) argues that reciprocity in the family, extended family, and tribe is what ultimately led to the extended order of cooperation through market trade. He postulates that this proclivity for reciprocal social exchange is so natural and instinctive that it survives even in interactions between anonymously paired subjects in the two-person extensive-form games described above.

This interpretation has been reinforced by many other extensive-form game-tree experiments. Thus, in the game shown in Figure 4, player 1 chooses between the SPE, \$10 for each, and passing to player 2 who chooses between pay-offs of \$15 and \$25 (for players 1 and 2 respectively) and pay-offs of \$0 and \$40. The move frequencies for 24 pairs of undergraduates are shown on the tree. Very similar outcomes prevail with a group of graduate students trained in economics and game theory (McCabe and Smith, 1999).

Effects of Context, Repetition, and Opportunity Costs in Trust Games

The sensitivity of cooperative behavior in trust games to the procedural, instructional, and opportunity cost context of the experiment has been demonstrated by several treatment manipulations.

'Partners' versus 'opponents'

Consider two treatment variations on the trust game of Figure 2(a): wherever the word 'counterpart' is used in the instructions to refer to the other decision maker, substitute the word 'partner' in one treatment and 'opponent' in the other (Burnham *et al.*, 2000). Subjects (156 pairs in total) were recruited in either 'small' groups of 12 in a session or 'large' groups of 24 in a session. In all sessions, half of the subjects (6 or 12) were randomly assigned to each of the two instructional conditions;

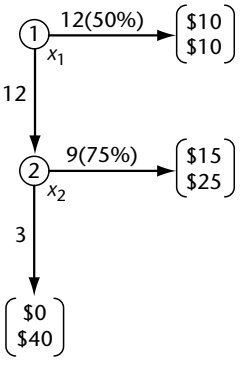


Figure 4. Experimental outcomes for a simple trust game (McCabe and Smith, 1999). A total of 24 subject pairs took part in the game. The figures beside the arrows indicate the number of pairs following each route though the game tree and the percentage moving right at each decision node.

each person was randomly paired with another and assigned to the player 1 or player 2 role; and the two experiments were run simultaneously in the same room. Neither group was informed that the other was reading slightly different instructions. Thus, the experimental design consisted of two group sizes, 12 or 24, and two instructional conditions, 'partner' and 'opponent'.

Each session began with a single play of the trust game. The subjects were then paid and informed that they would also participate in a second experiment. This second experiment used the same instructions except that the game was repeated for 10 periods of play. On each period of play, each person was matched with a new person, then randomly assigned the role of player 1 or 2. Each repetition was therefore between paired strangers. This is called 'repeat single' play. Repeat single play is like single play except that the subjects acquire experience under procedures that control for reputation formation across successive interactions.

It was found that 'partners' are more trusting (players 2 move down at node x_2) and more trustworthy (players 1 move right at x_3) than 'opponents'. (In the first single-play game, however, no difference was observed in the frequency of trust between the two treatments, but 68% of the 'partner' players 1 cooperated following an offer of cooperation, while only 33% of the 'opponent' players did.)

Over time (single play followed by 10 repeat single plays), with 'partners' trust increases through the first five plays then declines, while with 'opponents' trust steadily declines. Trustworthiness declines over time for 'partners', and remains low for 'opponents'. Hence, 'partners' learn to defect, but 'opponents' defect from the beginning.

Pairs who interact in groups of size 24 are less trusting than those in groups of size 12.

These results provide further support for the hypothesis, based on cortical memory theory, that context should matter. In this case, a simple two-level variation on the language used to describe the other person in each trial is sufficient to yield statistically significant differences in trust and trustworthiness.

Repeat single with and without punishment

The tendency for cooperation eventually to decline as play is repeated with distinct 'partners' is already suggested by Figure 3(a): of the 12 players 1 arriving at node x_3, half reciprocate and half defect. Hence, it is not profitable to offer cooperation in the trust game, and repetition with strangers is likely to cause a decline in cooperation, both offered and reciprocated, across time.

In the trust–punishment game in Figure 3(b), however, 85% reciprocate at node x_3, and only 15% defect, of which half are punished. Hence, it is profitable to offer cooperation, and it is not profitable to defect. This suggests that in repetition, using the repeat-single protocol, cooperation might not diminish. In fact, this is the case: when defection can be punished, the conditional probability of reciprocal cooperation by players 1 actually increases modestly across 15 periods of play (McCabe *et al.*, 2000).

Opportunity cost

An important implication of reciprocity theory (the value of option 'y' given up by choosing 'x') is that when person A chooses to forgo the SPE outcome and offer the cooperative option to person B, the pay-off alternatives should be such that person B sees clearly that person A is incurring an 'opportunity cost' – forgoing a smaller pay-off in an attempt to allow both persons to achieve larger pay-offs. There should be a cost incurred in order to gain from exchange. Failing this condition, the basis for an exchange, or reciprocation, is compromised: person B would be less likely to read clearly the intentions of person A, and person A will anticipate that an unclear message would be being conveyed.

Thus, in Figure 2, if instead of \$8 the SPE is \$10 for each player – identical to the 'cooperative' outcome – the outcome frequency results should change dramatically. This has been tested for the trust–punishment game tree in Figure 2(b) (McCabe *et al.*, 2002). The effect is to increase the frequency of the SPE outcome to about 95%. Thus, players have no difficulty concluding that the attempt to cooperate by player 2 at node x_2 is risky, and will not be chosen unless there is a compensating potential gain.

Another test of reciprocity is to contrast two versions of a game with the structure of Figure 4. Version 1 is like that in Figure 4, with different pay-offs but qualitatively the same outcomes. In version 2, player 1 has no option to move right. The prediction is that version 1 will yield more cooperative outcomes then version 2. In fact, defection is twice as frequent in version 2 as it is in version 1. The interpretation is that if nothing was given up by player 1 – the move did not deliberately forgo the pay-offs achievable at the SPE – then player 1's move does not constitute an 'offer'; so nothing need be reciprocated.

THEORY OF MIND AND ITS NEURAL CORRELATES

Experimental tests of non-cooperative equilibrium theory using anonymously paired subjects in two-person games consistently show that people do cooperate. Almost all subjects in the ultimatum game offer amounts in excess of the equilibrium predictions, and when they do offer equilibrium amounts their counterparts almost always reject the offer. Similarly, in trust games, up to half of subjects offer to cooperate at the risk of defection; and in varying degrees, depending on context, their counterparts cooperate at a cost to themselves. These data cannot be explained simply in terms of preferences – a utility for the other's payoff – nor can they be dismissed by the argument that the subjects are too unsophisticated or inadequately motivated.

A more satisfactory model is based on reciprocity and the human ability to communicate intentions through actions. This ability to invoke shared-attention, intention-detector, and 'mindreading' mechanisms in the brain is relevant to other observations of behavior in people impaired by frontal lobe damage and by autism.

Autism (whose genetic basis is indicated by its greater incidence in siblings and in identical twins) is characterized by 'mind blindness', a severe deficit in one's innate awareness of mental phenomena in other people. Children with autism fail developmentally to use pointing gestures to request objects or otherwise call the attention of others to items of joint interest. In contrast, blind children at age 3 are aware of what 'seeing' is in others, and will say 'see what I have'. At about age 3 or 4, normal children become aware of beliefs in others, and understand that others can hold false beliefs. Thus, shown a candy box, and asked what it contains, normal children will say that it contains candy. Upon opening the box, the child sees that pencils have replaced the candy. The child is then asked what the next child who comes in the room will think is in the box. Normal children will reply 'candy', whereas the majority of autistic children will say 'pencils' (Baron-Cohen, 1995).

Studies of autism, and of certain forms of brain damage from accidents or surgery, support hypotheses that particular regions of the brain have circuitry devoted to 'mindreading', an innate capacity for unconscious awareness of what others think or believe. Brain imaging studies of third-party false beliefs in story comprehension tasks have found activation in Broadman's area 8 (medial prefrontal cortex), and in other supporting regions such as the orbital frontal cortex (Fletcher *et al*., 1995). This role of Broadman's area 8 has been specifically corroborated by functional magnetic resonance imaging of subjects playing trust and trust–punishment games like those presented above (McCabe *et al*., 2001). These studies compare subjects' decision making when playing a human counterpart and when playing computer strategies with fixed known probabilities of moving 'left' or 'right'. Activation is significantly greater in the mindreading areas when playing a human than when playing a computer.

Thus, independent strands of research into the internal order of the mind and the external order of social exchange appear to be converging in support of the hypothesis that humans are so well adapted to personal exchange that reciprocity survives even in anonymous interactions.

References

Baron-Cohen S (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Burnham T, McCabe K and Smith VL (2000) Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization* **43**: 57–73.

Cosmides L (1985) The logic of social exchange. *Cognition* **31**: 187–276.

Cosmides L and Tooby J (1992) Cognitive adaptations for social exchange. In: Cosmides L and Tooby J (eds) *The Adapted Mind*, pp. 163–228. New York, NY: Oxford University Press.

Fletcher P, Happe F, Frith U *et al*. (1995) Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition* **57**: 109–128.

Fuster J (1999) *Memory in the Cerebral Cortex*. Cambridge, MA: MIT Press.

Gazzaniga M, Ivry R and Mangun G (1998) *Cognitive Neuroscience*. New York, NY: Norton.

Hayek F (1952) *The Sensory Order*. Chicago, IL: University of Chicago Press.

Hayek F (1988) *The Fatal Conceit*. Chicago, IL: University of Chicago Press.

Heinrich J (2000) Does culture matter in economic behavior? *American Economic Review* **90**(4): 973–979.

Hoffman E (2000). In: Smith (2000), pp. 79–90.

Hoffman E and Spitzer M (1985) Entitlements, rights and fairness. *Journal of Legal Studies* **14**: 259–297.

Hoffman E, McCabe K and Smith VL (1996) On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* **25**(3): 289–301. [Reprinted in Smith (2000).]

Hoffman E, McCabe K, Shachat K and Smith VL (1994) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* **7**: 346–380. [Reprinted in Smith (2000).]

McCabe K and Smith VL (1999) A comparison of naïve and sophisticated subject behavior with game theoretic predictions. *Proceedings of the National Academy of Sciences of the USA* **97**: 3777–3781.

McCabe K and Smith VL (2001) Goodwill accounting and the process of exchange. In: Gigerenzer G and Selten R (eds) *Bounded Rationality: the Adaptive Toolbox*, pp. 319–340. Cambridge, MA: MIT Press.

McCabe K, Houser D, Ran L, Smith VL and Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange in process. *Proceedings of the National Academy of Sciences of the USA* **98**: 11832–11835.

McCabe K, Rassenti SJ and Smith VL (1996) Game theory and reciprocity in some extensive form experimental games. *Proceedings of the National Academy of Sciences of the USA* **93**: 13421–13428. [Reprinted in Smith (2000).]

McCabe K, Rigdon M and Smith VL (2002) Positive reciprocity and intentions in trust games. *Journal of Economic Behaviour and Organization* (in press).

Selten R (1975) Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* **4**: 25–55.

Smith VL (1998) The two faces of Adam Smith. *Southern Economic Journal* **65**: 1–19.

Smith VL (2000) *Bargaining and Market Behavior*. Cambridge, UK: Cambridge, University Press.

Game Theory in Economics

Advanced article

Kevin A McCabe, George Mason University, Fairfax, Virginia, USA

CONTENTS

Game theory is an increasingly important tool to help economists understand the strategic interaction between groups of individual decision makers.

INTRODUCTION

One of the most frequent observations in experimental economics is that there is a great deal of variation in how subjects play. It seems reasonable to hypothesize that this variation exists because individuals differ in how they think about other players' rationality. While game theory provides us with a wealth of models that help provide retrospective explanations of why this variation is observed, these models have far less predictive capability. What seems to be missing is a unified treatment of cognition in strategic settings.

GAME THEORY AND DECISION THEORY

Game theory can be seen as the natural extension of rational-choice models of individual decision making as modeled by decision trees (Luce and Raiffa, 1957). Figure 1 gives a simple example of a decision tree. The first and second nodes of the decision tree are decision nodes, the first belonging to player 1, and the second to Nature. Player 1 must decide, before knowing Nature's move, whether or not to stop and make \$10 for sure, or go on and face a probability p of getting \$15 and $1-p$ of getting \$0. If player 1 desires to maximize the expected pay-off, then the decision whether or not to go on depends on whether p is sufficiently large that $p \times 15 + (1-p) \times 0 \geq 10$. If $p \geq \frac{2}{3}$ then player 1 should go on.

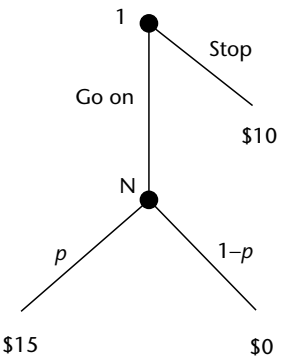


Figure 1. An individual decision problem for player 1 with perfect information.

We have assumed that the individual is interested in the monetary pay-offs associated with the outcome (or terminal) nodes. The usual economic assumption is that, everything else being equal, our player will prefer more money to less. However, in Figure 1 our player must choose between \$10 for sure, or a gamble g that pays \$15 with probability p and \$0 with probability $1-p$. We have assumed that our player will compute the expected value of g, that is, $15p$, in order to compare the value of the gamble to \$10. More generally, we could use expected-utility theory to model how our individual will choose between \$10 and the gamble. Let $U(x)$ be the subjective value that our player places on $\$x$. We assume U is an increasing, but not necessarily linear, function. In expected-utility theory, the value of g is calculated as $EU(g) = pU(15) + (1-p)U(0)$, and to make a choice our player would compare $U(10)$ to $EU(g)$. If the subject is trying to maximize expected utility, then the utility values should replace the pay-offs shown in Figure 1.

If the utility function for a player is nonlinear, an experimenter who wants to control individual values in an experiment faces a difficulty, since ultimately the experimenter pays off in dollars, not utilities. One approach is to risk-neutralize a subject by inducing a linear utility function. To do this, give the subject only two possible pay-offs, say \$5 and \$25. Now replace each dollar of pay-off with

a red lottery ticket. For example, \$0 is replaced by 0 red tickets, \$10 is replaced by 10 red tickets, and \$15 is replaced by 15 red tickets. When an outcome node is reached, the subject earns some number of red lottery tickets. Blue lottery tickets are then added until the total number of tickets is equal to 15, and all the tickets are then placed in an urn. One ticket (red or blue) is then randomly drawn from the urn. If the ticket is red, the subject earns \$25, valued at $U(25)$, and if it is blue, the subject earns \$5, valued at $U(5)$. Now, a subject who stops earns 10 tickets, valued at $\frac{1}{3}U(5)+\frac{2}{3}U(25)$. A subject who goes on earns either 15 tickets, valued at $U(25)$, with probability p, or 0 tickets, valued at $U(5)$, with probability $1-p$. We then know that the subject should go on just when $p \geq \frac{2}{3}$ since this is exactly the condition for $pU(25)+(1-p)U(5) \geq \frac{1}{3}U(5)+\frac{2}{3}U(25)$.

From here on we will simply use pay-off numbers, which can be thought of as utility values, but in experiments these numbers are either dollars or lottery tickets.

Consider the decision problem in Figure 2. Nature moves twice, but our player does not know which branch Nature chooses initially. If Nature has chosen the left branch, then Nature will be more likely ($p=\frac{3}{4}$) to choose 15 for the player, but if Nature has chosen the right branch, then Nature will be less likely ($p=\frac{1}{4}$) to choose 15 for the player. In the figure the decision nodes for the player are connected by a dashed line, indicating that they are in the same information set for the player. Nodes in the same information set must have the same number of branches, and they must all be played in the same way by the player, since the player does not know which of the decision nodes he or she is at. Suppose the player believes that Nature will choose the left branch 90% of the time, i.e., $q=0.9$. If the player decides to stop, the player gets 10, but if the player decides to go on, the expected pay-off is $0.9\times\frac{3}{4}\times 15+0.1\times\frac{1}{4}\times 15=10.5$. Since $10.5>10$, player 1 should go on.

EXTENSIVE-FORM GAMES

In Figure 1, Nature does not get a pay-off, or make choices, but instead is regarded as a probability distribution over possible actions outside the control of the individual. The use of types is suggested by Harsanyi (1995). If we replace Nature with another individual, we get the game tree shown in Figure 3. The pay-offs to player 2 are written below those to player 1. Thus, if player 1 stops, then player 2 also gets 10, but if player 1 goes on, then player 2 can either move left, in which case player 2 gets 25, or right, in which case player 2 gets 40. Given that player 2's actions are still outside the control of player 1, how should player 1 decide? To answer this question we can analyze the new decision problem as a two-person extensive-form game.

An extensive-form game with complete and perfect information can be defined as follows. Define a 'branch' as an arc that connects two nodes. A branch represents a move in the game. A node is an 'initial' node if it has no branches going into it, but has at least one branch leaving it. A node is a 'terminal' node if it has at least one branch going into it but no branches leaving it. A 'path' is a sequence of nodes that are connected by branches, starting at an initial node and ending at a terminal node. Nodes and branches must satisfy the following properties:

- All decision nodes are played by only one player.
- Each terminal node has a pay-off for every player in the game.

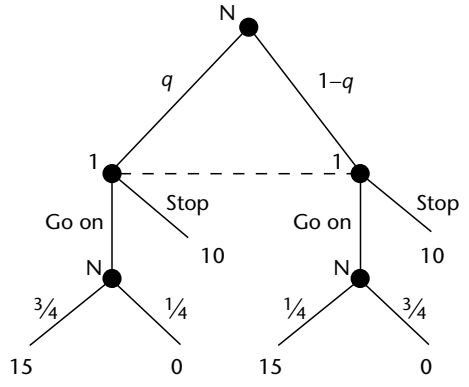


Figure 2. An individual decision problem for player 1 with imperfect information.

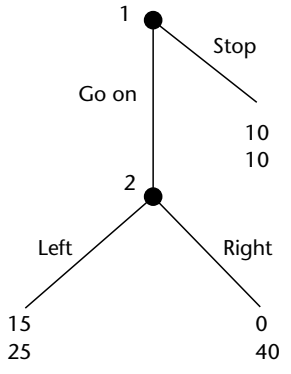


Figure 3. A two-person decision problem with perfect information.

- There is only one initial node.
- Every node, except the initial node, is connected to some earlier node (called its 'predecessor') by a single branch.
- Every terminal node is connected to the initial node by a unique path.

Kuhn (1953) showed that the principle of backwards induction can be used to guide player decisions. In backwards induction, each player reasons backwards through the tree, starting with the terminal nodes, and making the rational choice for each player at each decision node, assuming that decisions made after that node will also be rational.

The decision problem shown in Figure 3 is an extensive-form game with complete and perfect information. Using backwards induction, player 1 should reason as follows. If player 2 gets to move, player 2 will move right, since $40 > 25$. Therefore player 1 should stop.

However, in experiments with cash-motivated subjects, about half of the player 1s choose to go on, and of those player 2s who get to move, over two-thirds reciprocate the favor by moving left.

How can we explain this divergence from the theory? The value of game theory as a modeling tool becomes apparent as the experimenter attempts to reconcile the differences between theory and observation. One approach to explaining this divergence is to assume that Nature produces different types of player 2s.

Consider a simple model in which Nature creates two types of individuals who may end up being player 2. A 'trustworthy' type of player feels strongly compelled to reciprocate when believing she has received a favor. Moving left as player 2 may result in a pay-off of 40, but player 2 may 'feel' far worse off, like getting a pay-off of only 20. Trustworthy players will therefore move left, because $25 > 20$. The second type for player 2, the 'me' type of player, feels no such compunction, and plays right, since $40 > 25$.

This game can be described as in Figure 4, with Nature moving first. Notice the addition of an information set as a dashed line connecting player 1's two decision nodes. Player 1 does not know which node he is at, that is, which type player 2 is, when it is player 1's turn to move. On the other hand, player 2 knows his or her own type when player 2 gets to move.

Now suppose player 1 believes that more than two-thirds of player 2s provided by Nature are of the trustworthy type. (This frequency is shown as q in Figure 4.) Then player 1's expected pay-off is higher by going on, since $q \times 15 + (1 - q) \times 0 > 10$.

STRATEGIC-FORM GAMES

The extensive-form game shown in Figure 3 can be rewritten as the 'normal-form' or 'strategic-form' game shown in Figure 5. While in extensive-form games the order of moves is important, in a strategic-form game players choose their strategies simultaneously. All normal-form games can be written as n-dimensional arrays, with n being the number of players, making the game simpler to analyze. Kohlberg and Mertens (1986) show that the essential strategic features of all extensive-form games are retained when the games are converted to strategic form. When players move simultaneously it is less clear how they should think about the other player. Consequently, from a cognitive viewpoint it is likely that subjects think differently about how to play an extensive-form game and a strategically equivalent strategic-form game.

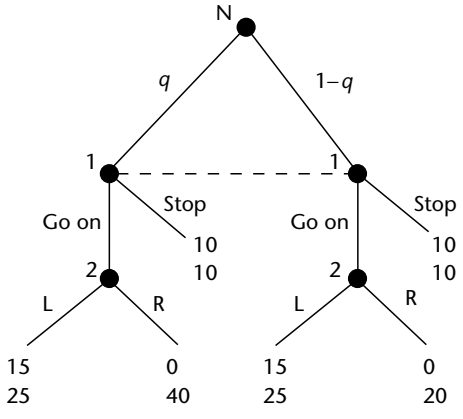


Figure 4. A two-person decision problem with incomplete information.

Player 1 \ Player 2	Left	Right
Stop	10 10	10 10
Go on	15 25	0 40

Figure 5. A two-person decision problem as a normal-form game.

In analyzing a strategic game, players must put themselves in each other's shoes and try to reason about how the other player will reason. For example, in Figure 5, player 1 can notice that player 2 is always no worse off by moving right. Such a choice is called a 'weakly dominant' strategy for player 2. Alternatively, we can say that moving right is a best response for player 2 independently of player 1's choice. For player 1, stopping is a best response to player 2's decision to move right. The pair of pure (deterministic) strategies (Stop, Right) is thus a 'Nash equilibrium', defined as a strategy for each player that is a best response to all the others.

Not every strategic-form game has a pure-strategy Nash equilibrium: consider, for example, the popular children's game of rock–scissors–paper, shown in Figure 6. If we extend our concept of a strategy to include 'mixed strategies' – choices of probability distributions over pure strategies – then, as Nash proved, every strategic game has a Nash equilibrium. In the case of rock–scissors–paper, each player should play each strategy with equal probability; i.e., the Nash equilibrium strategy is (⅓, ⅓, ⅓).

However, if player 2 knows that player 1 will play (⅓, ⅓, ⅓), then player 2 is just as well off playing (1, 0, 0), or rock for sure. Of course, player 2 does not want player 1 to know this, but how will player 1 find out? One approach to solving the 'incentive problem' associated with mixed strategies is to view a mixed strategy as the frequency distribution of pure strategies played in a frequently-replayed perturbed version of the original game (Harsanyi, 1977). Imagine that a game is frequently occurring and that a large population of potential players keeps experiencing minor fluctuations in their pay-offs. On average, they expect to play against someone playing (⅓, ⅓, ⅓). But suppose that if rock wins player 2's payoff is now $1+\varepsilon$, where $\varepsilon > 0$ is the small perturbation. Clearly player 2 now prefers (1, 0, 0), and will play it.

Figure 7 illustrates a game with two Nash equilibria, (Up, Left) and (Down, Right). What makes this game interesting is that even though (Up, Left) is preferred by both players, coordinating strategies on this equilibrium in simultaneous play is made difficult by the degree of risk associated with being wrong about guessing what the other player will do. Player 1 may reason as follows. 'If I play Up, and player 2 plays Right, then I get my worst pay-off, 0, and player 2 gets 9. Since getting 9 is not much worse than getting 10, player 2 has not given up much by playing Right, and if I play Down, then player 2 still gets 5. So player 2 has a lot of reasons to play Right. What is worse is that player 2 has probably reasoned the same way about me and is now convinced I will play Down, which I better now do.' Harsanyi and Selten (1988) appeal to the 'principle of insufficient reason' to define an equilibrium as 'risk-dominant' if each equilibrium strategy is a best response to the $n-1$ other-players mixed strategies that put equal weight on each of their possible pure strategies.

RATIONALIZING OTHER PLAYERS' DECISIONS

An important question in game theory is: what do people need to know about each other in order to make rational strategic choices? We have mentioned the 'default' reasoning, the principle of insufficient reason, which gives equal weight to other players' strategies. However, before falling back on this principle, subjects may do better to apply other methods such as backwards induction. Backwards induction can be generalized to strategic-form games as follows.

Player 1 \ Player 2	Rock	Scissors	Paper
Rock	0 0	1 0	0 1
Scissors	0 1	0 0	1 0
Paper	1 0	0 1	0 0

Figure 6. The game of rock–scissors–paper.

Player 1 \ Player 2	Left	Right
Up	10 10	0 9
Down	9 0	5 5

Figure 7. A risky coordination problem.

We say that a player's strategy s_i, is 'strictly dominated' by another strategy s_j if playing the strategy s_j will always return a higher pay-off than playing the strategy s_i. In this case, there is no set of beliefs for this player about the other players that would rationalize the use of the strategy s_i, and therefore we can eliminate this strategy from the player's choices. We can then repeat this procedure, eliminating any strictly-dominated strategies that remain, until all such strategies are gone. This process of iterative elimination of strictly-dominated strategies leaves only those strategies that are 'rationalizable'.

This principle can be generalized (Bicchieri, 1993; Osborne and Rubenstein, 1994). For example, we can extend this procedure to also eliminate weakly-dominated strategies. In Figure 5, Right weakly dominates Left for player 2 since it makes player 2 no worse off, no matter what player 1 does. Therefore we should eliminate Left. But then, for player 1 Stop strictly dominates Go-on, so we should eliminate Go-on. What remains is the dominance-solvable pair (Stop, Right).

Another principle that players may use is the principle of forward induction, as illustrated in Figure 8 – an extension of the game shown in Figure 4. In this game, after Nature moves by choosing player 2's type, player 2 gets to move. Player 2 can opt out and get 30, or player 2 can continue. The principle of forward induction asks: what should player 1 assume about player 2's strategy if player 2 chooses to continue? Obviously, player 2 is trying to win more than 30, which can occur only if Nature has chosen a 'me' type of player on the left branch. So, if player 2 chooses to continue, player 1 should realize that player 2 plans to move right, and therefore player 1 should stop. But this means that player 2 should always opt out, leaving player 1 with 5.

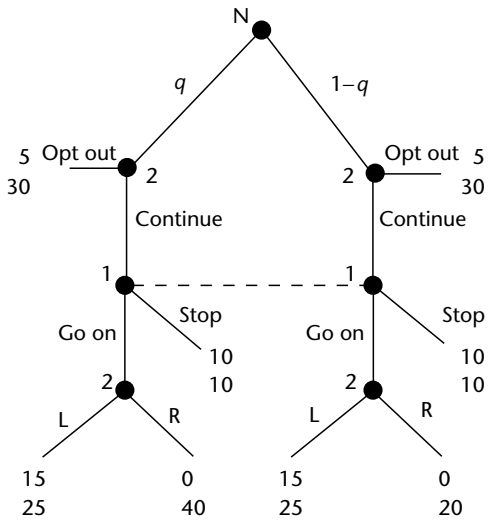


Figure 8. A two-person decision problem with incomplete information.

Going back to the risky coordination game in Figure 7, suppose we add the proviso that player 1 gets to choose whether or not to opt out and get 6 or continue and play the game. Again using the principle of forward induction, player 2 should assume that by choosing to continue player 1 plans to play Up. Furthermore, player 1 should assume that player 2, having made this analysis, will play Left. Thus player 1 should choose Continue, and then play Up with greater confidence that he or she will end up with 10.

REPEATED GAMES

Repeating a game allows for strategy spaces in which the play in future games depends on the play in past games. For example, in Figure 5, repeat game strategies (such as 'tit-for-tat' in the prisoner's-dilemma game) may be useful in achieving the more efficient (15, 25) or (0, 40) outcomes rather than the (10, 10) outcomes. Suppose, for example, player 1 adopts the following 'trigger' strategy: start by playing Go-on, and continue to play Go-on as long as player 2 continues to play Left; if player 2 chooses to play Right, then play Stop from then on. By playing Right, player 2 can get 40 (instead of 25), but from then on player 2 will get only 10.

If there is a fixed number of periods, this strategy is not an equilibrium strategy, for the following reason. In the last period, player 2 should play Right. Player 1 should realize this and play Stop. But then, player 2 should play Right in the second-to-last period. Again, player 1 should realize this and play Stop. By backwards induction, cooperation is unraveled all the way to the first period, with player 1 playing Stop. Kreps *et al.* (1982) show how adding a little bit of uncertainty about the other player's type is sufficient to achieve cooperation in the finitely-repeated prisoner's-dilemma game. Axelrod (1984) has shown that in practice people don't unravel the game using backwards induction, but rather play relatively myopically, while still incorporating elements of the repeat-game strategy. Given this, the trigger strategy can be replaced by the more forgiving strategy tit-for-tat, resulting in even larger efficiency gains.

McCabe and Smith (2002) report on a series of experiments with properties similar to the game shown in Figure 3, but where they vary the

likelihood that people will play each other again. They find that the likelihood of meeting again influences subjects' levels of cooperation. Both theoretical and empirical results point to the need for the experimenter to control for this likelihood. Different protocols are possible for an experiment with four people in which no pair will play each other twice. In the 'everyone meets once' condition, a total of six games will take place. However, there is still room for an indirect effect. In the first period, subjects 1 and 2 play and subjects 3 and 4 play. In the second period, subjects 1 and 3 play and subjects 2 and 4 play. In the third period, subjects 1 and 4 play and subjects 2 and 3 play. But subject 1 has already played subject 2 who has already played subject 4, so it is possible that subject 1 has already influenced subject 4's behavior before they play. To eliminate this contagion effect, the 'turnpike matching' condition is employed. This control reduces the number of observation pairs to four. In the first period, subjects 1 and 4 play. In the second period, subjects 1 and 3 play and subjects 2 and 4 play. In the third period, subjects 2 and 3 play. So subject 1 meets subject 4 and then subject 3, and subject 2 meets subject 4 and then subject 3.

References

Axelrod R (1984) *The Evolution of Cooperation*. New York, NY: Basic Books.

Bicchieri C (1993) *Rationality and Coordination*. Cambridge, UK: Cambridge University Press.

Harsanyi J (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge, UK: Cambridge University Press.

Harsanyi J (1995) Games with incomplete information. *American Economic Review* **85**: 291–303.

Harsanyi J and Selten R (1988) *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Kohlberg E and Mertens J-F (1986) On the strategic stability of equilibria. *Econometrica* **54**: 1003–1037.

Kreps D, Milgrom P, Roberts J and Wilson R (1982) Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* **27**: 245–252.

Kuhn H (1953) Extensive games and the problem of information. In: Kuhn HW and Tucker AW (eds) *Contributions to the Theory of Games*, pp. 193–216. Princeton, NJ: Princeton University Press.

Luce D and Raiffa H (1957) *Games and Decisions*. New York, NY: John Wiley.

McCabe K and Smith V (2002) Strategic analysis by players in some games: what information do they use? In: Ostrom E (ed.) *Trust and Reciprocity*. New York: Russell Sage Foundation.

Osborne M and Rubinstein A (1994) *A Course in Game Theory*. Cambridge, MA: MIT Press.

Further Reading

Gintis H (2000) *Game Theory Evolving*. Princeton, NJ: Princeton University Press. [An introduction to many of the recent advances in game theory along with numerous interesting examples.]

Nash J (1950) Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences of the USA* **36**: 48–49.

von Neumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Smith V (1992) Game theory and experimental economics: beginnings and early influences. In: Weintraub ER (ed.) *Toward a History of Game Theory*. Annual Supplement to *History of Political Economy* **24**: 241–282. [A historical account of the early influences of game theory on experimental economics.]

Games: Centipede

Intermediate article

Amnon Rapoport, University of Arizona, Tucson, Arizona, USA

CONTENTS

The centipede game is a finite N*-person extensive form game with perfect information for which the backward induction solution yields paradoxical results. It raises important questions concerning beliefs, knowledge, and rationality in interactive decision-making, which can be examined both theoretically and experimentally.*

INTRODUCTION

The discipline of game theory studies mathematical models of conflict and cooperation between rational decision-makers (called 'players') whose decisions affect each other. The centipede game is a game-theoretical model of a class of interactive decision situations that have challenged a basic solution concept of game theory called 'backward induction'. The centipede game has stimulated much theoretical research into the nature of beliefs, knowledge, and rationality in interactive decision situations, and experimental research on conflict, cooperation, and trust.

Formally, the centipede game is a finite N-person extensive form game with complete and perfect information and no chance moves. It consists of the following elements:

- A game tree, which is a graph with m nodes and $m-1$ links in which every two adjacent nodes are connected by exactly one link. The game tree includes a distinguished node, called the *origin* of the tree, which designates the first mover. Node k in the tree is said to *follow* node j if there is a path from the origin to node k through node j. A *terminal* node is one with no followers.
- A partition of the non-terminal (decision) nodes into player sets labeled $1, \ldots, n$. A node in the player set only i corresponds to a choice of move (a decision) by player i ($i = 1, \ldots, n$).
- Associated with every terminal (outcome) node is a vector of n real numbers. The ith component of this vector gives player i's utility ('pay-off') for the outcome represented by the terminal node.

For this class of games, perfect information means that when it is her turn to choose an immediate follower from any of her non-terminal nodes, a player knows perfectly all the choices made at the previous decision nodes. There is no possibility of secret moves. The assumption of complete information means that everything about the extensive form is known to all the players.

DESCRIPTION OF THE CENTIPEDE GAME

A two-person centipede game was introduced by Rosenthal (1981) and later studied theoretically by Aumann (1992, 1995, 1998), Ben-Porath (1997), Feinberg (2001), Ponti (2000), Stalnaker (1998), and many others. In a variant of this game discussed by Aumann (1992), there are two players, Alice and Bob, and a sum of \$10.50 lying on the table in front of them. Moving first, Alice has the option of taking \$10 and leaving \$0.50 to Bob. If she does so ('exit'), the game is over. If not ('continue'), the amount on the table is increased tenfold, to \$105, and it is Bob's turn to play next. Now Bob has the option of taking \$100, leaving \$5 to Alice. If he does so ('exit'), the game is over. If not ('continue'), the amount is increased tenfold, to \$1050. Continuing in this way, with players' roles being interchanged and payoff increasing tenfold on each move (stage), the game terminates after three full rounds of play (i.e. six stages). In the sixth and final stage, Bob has the opportunity of taking \$1 000 000, leaving \$50 000 to Alice. Otherwise, the game terminates with each player receiving zero pay-off.

Figure 1 displays this interactive situation as a finite extensive form game. There are six decision nodes, labeled alternately as 'Alice' and 'Bob' with Alice the first player to move. There are seven terminal nodes. Associated with each terminal node are two numbers. The top number is Alice's pay-off on this outcome, and the bottom number is Bob's pay-off on this outcome.

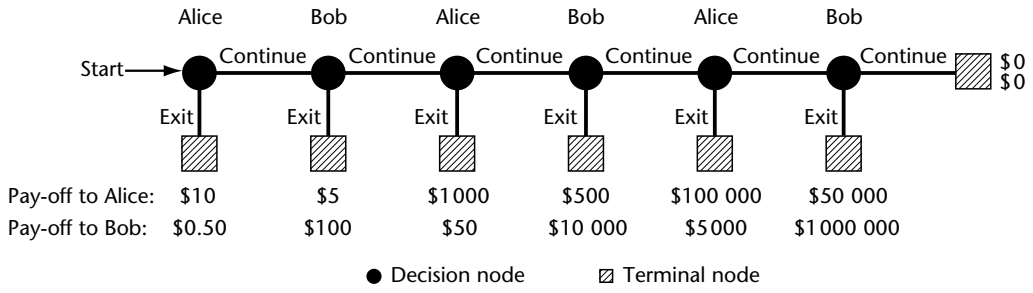


Figure 1. Two-person six-move centipede game (Aumann, 1992).

There are two major features that lend this game special interest. Firstly, the pay-offs are structured in such a way that both players are better off if play continues for at least two stages (one complete round). Secondly, if one of the players continues and the other exits on the immediately following stage, then the player who continues is worse off and the one who exits is considerably better off. Thus, there is an incentive to defect (exit) rather than cooperate (continue) and thereby risk a lower pay-off.

BACKWARD INDUCTION

How should rational players behave in this game if it is played only once? The standard way of answering this question for finite extensive form games with perfect information is by a process known as *backward induction*. Backward induction (also known as 'dynamic programming') is used to obtain the optimal decision in multistage individual decision-making problems described by decision trees. It is also used to derive the solution for the well-known finitely iterated two-person 'prisoner's dilemma' game. For the game in Figure 1, the recursive process of backward induction is straightforward. Begin with the sixth and final stage. At this node, Bob's option to exit (and receive $1 000 000) dominates his option to continue (and receive $0). Therefore, Bob will definitely exit if the game reaches stage 6. Move back to stage 5, where it is Alice's turn to play. Assuming that Bob will act rationally and exit at stage 6, Alice's option to exit at stage 5 (and receive $100 000) dominates her option to continue (and receive $50 000 at stage 6). Hence, Alice should exit if the game reaches stage 5. With the same reasoning applying to all decision nodes, an exit decision at each stage is prescribed. Therefore, by the logic of backward induction, Alice will exit at stage 1, with pay-offs of $10 and $0.50 to Alice and Bob respectively. Hence the paradox. Note that, provided the game is finite, the process of backward induction does not depend on the number of stages. The game in Figure 1 could be extended to any finite number of stages, with the pay-offs becoming enormous, without changing the solution.

Reactions of people to this solution vary. Some may feel that the solution for the centipede game is not really counterintuitive. Although Alice may feel frustrated, a careful analysis of the game may convince her in the end to exit immediately and collect the $10. As noted by Aumann, there is a difference between frustration and paradox. Thus, players in the one-shot 'prisoner's dilemma' game may feel frustrated by the prescription to always defect, but the logic of the game is compelling. However, most reasonable people are not willing to accept this solution, or at best believe that it represents an approach of little practical value. Aumann considers the centipede game to be one of the 'disturbing counterintuitive examples of rational interactive decision-making' (1992, p. 219). McKelvey and Palfrey, who have studied the game experimentally, claim that the solution is 'widely acknowledged to be intuitively unsatisfactory' (1992, p. 803). The experimental evidence is reviewed below.

THEORETICAL EXPLANATIONS

To get a better feel for the issues raised by the centipede game, consider a shorter, simpler, and slightly different version of the game with only three stages and considerably smaller pay-offs. In contrast to the game in Figure 1, if Alice 'continues' on her second (and last) decision node in the game in Figure 2, both players receive positive pay-offs ($2 and $3 for Alice and Bob respectively). By the same backward induction reasoning as before, Alice should exit at her first decision node. She should do so if she expects that she will behave rationally at her second decision node, and if she expects that Bob expects that she

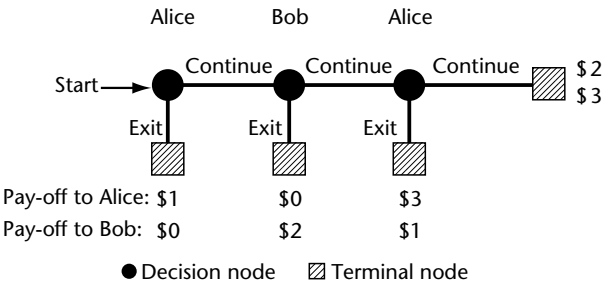


Figure 2. Two-person three-move centipede game.

will do so and that he will then act rationally (Feinberg, 2001).

The crucial question is what to make of her decision if she decides to continue on her first decision node. Bob may interpret this move in several possible ways. He may ascribe it to an error on Alice's part. He may believe that this move is intended to signal Alice's intention to continue on her second decision node, in which case he would receive his largest possible pay-off ($3). Or he may believe that Alice is attempting to trick him to continue so that she can exit on her second move and receive her largest possible pay-off ($3). In any case, Bob must realize that under the backward induction reasoning, continuing is not optimal for Alice, and that consequently this reasoning may be invalid. Bob's decision will clearly depend on his interpretation of Alice's decision to continue. If he believes that she intends to continue on her second move, the rational decision for him is to continue too. This, then, would imply that the rational decision for Alice on her first move is, indeed, to continue, since she would have been wrong in assuming that if she continues then Bob will exit.

The examples in Figures 1 and 2 raise two theoretical questions. The first is how to account for the paradox of backward induction; and the second is how to justify the process of backward induction in the centipede game. There has been considerable theoretical work on these questions, which has resulted in explications and clarifications of the elusive concepts of rationality, knowledge, and beliefs in extensive form games. This body of literature is too technical to be described in detail here. The basic ideas underlying some of the theoretical studies are briefly described below.

Common Knowledge of Rationality

Consider first the concept of knowledge in interactive decision-making. An event is called *common* knowledge if all players know it, all know that all know it, and so on *ad infinitum*. An event is called *mutual* knowledge if all players know it. It is called second-order mutual knowledge if it is mutual knowledge that it is mutual knowledge; and so on for any finite order. As illustrated above, the analysis of Alice's decision to exit on her first move in Figure 2 was based on iterated knowledge ('Alice knows that Bob knows that Alice knows…') of rationality up to a specified level, which depends on the number of decision nodes in the game tree, but no further.

Aumann noted that common knowledge of rationality and even the considerably weaker mutual knowledge of rationality are very strong assumptions and not likely to be true in practice. To resolve the paradox of backward induction in the centipede game, he proposed to relax this assumption by allowing a very small measure of doubt to enter at a high order of mutual knowledge of rationality. He showed that a small breakdown in the commonality of the knowledge of rationality is enough to justify the kind of behavior in the centipede game that most people would consider 'reasonable.' His approach is of particular relevance to cognitive scientists who are aware of the limitations of the human cognitive system and consider human rationality to be bounded. It reflects the fact that in real interactive situations there is typically uncertainty about what others will do, the extent of their rationality, and what others think about what you think and about your own rationality.

Aumann (1995, 1998) drew a distinction between *substantive* and *material* rationality. Substantive rationality refers to the condition where it is assumed common knowledge that each player i would act rationally at each of his decision nodes v, even when i knows that v will not be reached. The alternative condition of material rationality stipulates that player i acts rationally only at those of his decision nodes that are actually reached. The counterfactual component of the latter condition is smaller than in substantive rationality, as nothing is assumed about what actions players may take at unreached decision nodes. Aumann (1995) proved that if common knowledge of substantive rationality obtains in perfect information games, then backward induction is sustained. In a subsequent paper (1998) he proved that, in the particular case of the centipede game, the weaker condition of common knowledge of material rationality is sufficient to ensure backward induction.

Other Explanations

Additional theoretical work invoking other concepts has followed. Using evolutionary arguments

rather than assumptions about common knowledge and rationality, Ponti (2000) has shown that backward induction can still accurately predict players' behavior in the centipede game provided they are given enough time to study and appreciate the strategic environment in which they operate.

Stalnaker (1998) extended the notion of rationality by arguing that the rationality of choices in a game depends not only on what the players believe, but also on the way they revise their beliefs in response to some new or contradictory piece of information. In this framework for reasoning about hypothetical situations, he showed that the common knowledge of substantive rationality need not imply backward induction.

Ben-Porath (1997) has shown that common belief with probability one (instead of common knowledge) of rationality need not imply backward induction.

Feinberg (2001) argued that whereas a particular action can be called 'rational' in a well-specified context, ascribing rationality to a decision-maker is more problematic. This is because of the conflicting nature of objectivity and freedom of choice. If the player is 'objectively' rational, then his decisions are deterministic. But if the player is 'subjectively' rational, in the sense that his actions are perceived and evaluated as rational by an outside observer, this rationality becomes a descriptive and predictive tool for the observer. This interpretation implies that, from the subjective point of view of the observer, the player's rationality may be refuted. Rather than modeling the players' reasoning in the centipede game as an added structure separate from the game itself, Feinberg proposes a representation of dynamic games in which the game is described by the subjective knowledge of players at hypothetical situations. He further shows that common subjective knowledge ('confidence') of rationality contradicts the assumptions of the centipede game. However, rationality and common confidence of future rationality are shown to imply backward induction for perfect information games.

EXPERIMENTAL FINDINGS

Beliefs about the rationality and intentions of the other player, which may be revised during the course of the game, largely determine a player's decisions. Powerful and profound as they are, the theoretical analyses of the centipede game shed little light on how reasonable people, whose rationality is bounded, actually solve the conflict embedded in the centipede game. Only controlled experiments may provide some answers. The results of two experiments are described below.

McKelvey and Palfrey (1992)

Figures 3 and 4 show two examples of a variant of the two-person centipede game investigated experimentally by McKelvey and Palfrey (1992). These two games are similar to the game in Figure 2 in that if the final stage is reached, player 2 can reward player 1 for the trust placed in him by deciding to continue. If he does so, player 2 forgoes half of his pay-off in order to increase player 1's pay-off eightfold. The game in Figure 3 consists of two rounds of play, while the game in Figure 4 consists of three rounds.

McKelvey and Palfrey reported the results of multiple sessions of two carefully conducted centipede game experiments. The participants were undergraduate students who were paid in accordance with their performance. Each game started with a total pot ('pie') of 50 cents, divided into a 'large' pile of 40 cents and a 'small' pile of 10 cents. Each time a player 'continued', both piles were doubled in value and the roles of the two players were interchanged. Each experimental session included 20 or 18 participants who were divided into two groups (player 1 and player 2) at the beginning of the session. Player roles (types) remained fixed during the session. Each participant played one game with each of the participants assuming the opposite role. Thus, no participant was ever matched with another participant more than once. All this was common knowledge.

These two games were designed to study the descriptive power of the backward induction solution. Under this solution, all the games end at the first terminal node. On the other hand, if the two players fully cooperate by always continuing, all the games end in the final terminal node. The results support neither of these predictions (see Table 1). Out of a total of 281 plays of the four-

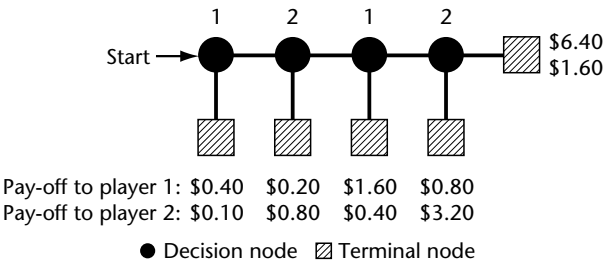


Figure 3. Two-person four-move game (McKelvey and Palfrey, 1992).

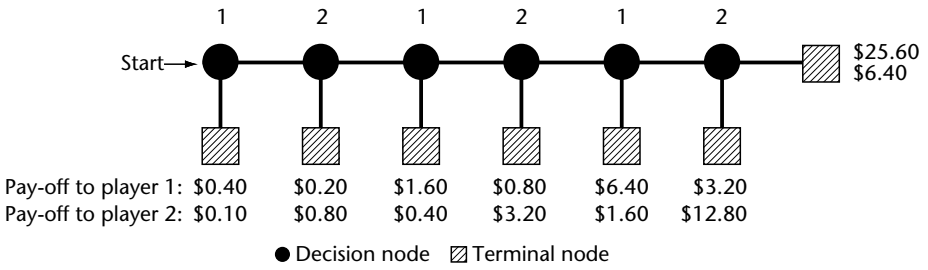


Figure 4. Two-person six-move centipede game (McKelvey and Palfrey, 1992).

Table 1. Proportion of plays ending at each terminal node

		Terminal node								
	Number of plays	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
4-move game (Figure 3)[a]										
Session 1	100	0.06	0.26	0.44	0.20	0.04				
Session 2	81	0.10	0.38	0.40	0.11	0.01				
Session 3	100	0.06	0.43	0.28	0.14	0.09				
All sessions	**281**	**0.071**	**0.356**	**0.370**	**0.153**	**0.049**				
6-move game (Figure 4)[a]										
Session 1	100	0.02	0.09	0.39	0.28	0.20	0.01	0.01		
Session 2	81	0.00	0.02	0.04	0.46	0.35	0.11	0.02		
Session 3	100	0.00	0.07	0.14	0.43	0.23	0.12	0.01		
All sessions	**281**	**0.007**	**0.064**	**0.199**	**0.384**	**0.253**	**0.078**	**0.014**		
9-move game (Figure 5)[b]										
Session 1	300	0.463	0.317	0.110	0.050	0.027	0.020	0.010	0.003	0.000
Session 2	300	0.393	0.277	0.157	0.087	0.030	0.017	0.023	0.013	0.003
Session 3	300	0.303	0.280	0.187	0.093	0.053	0.037	0.010	0.003	0.033
Session 4	300	0.407	0.257	0.183	0.077	0.037	0.017	0.013	0.003	0.007
All sessions	**1200**	**0.392**	**0.283**	**0.159**	**0.077**	**0.037**	**0.023**	**0.014**	**0.006**	**0.011**

[a]McKelvey and Palfrey (1992).
[b]Rapoport *et al.* (2000).

move game in Figure 3, the proportions of plays ending in the 1st, 2nd, 3rd, 4th, and 5th terminal nodes were 0.071, 0.356, 0.370, 0.153, and 0.049, respectively. Out of a total of 281 plays of the six-move game in Figure 4, the proportions of plays ending at the 1st, 2nd, 3rd, 4th, 5th, 6th, and 7th terminal nodes were 0.007, 0.064, 0.199, 0.384, 0.253, 0.078, and 0.014, respectively. With 7% of the four-move plays and less than 1% of the six-move plays ending in the first terminal node, there is clearly no support for backward induction. Moreover, the evidence suggests that whatever support this solution receives decreases with the number of rounds of play.

McKelvey and Palfrey also estimated the conditional probability of exit at each decision node j, denoted by $P(E|j)$; that is, the probability of exiting at node j if node j is actually reached. Table 2 shows that for each of the two games the values of $P(E|j)$ almost always increase with j ($j = 1, \ldots, 4$ in Figure 3 and $j = 1, \ldots, 6$ in Figure 4). Thus, as play progresses the tendency to cooperate decreases.

Rapoport, Stein, Parco, and Nicholas (2000)

Rapoport *et al*. (2000) have argued that the two-person centipede game constitutes a special case – albeit an important one – where reciprocity may assume a critical role. As the number of interacting players increases, the effects of reciprocity may diminish. Rapoport *et al*. proposed an extension of the centipede game from two to three players in an attempt to study the generality of the results reported by McKelvey and Palfrey. Figure 5 shows their game, which includes three players

Table 2. Inferred conditional probabilities of exit at each terminal node

		Terminal node								
	Number of plays	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
4-move game (Figure 3)[a]										
Session 1	100	0.06	0.28	0.65	0.83					
Session 2	81	0.10	0.42	0.76	0.90					
Session 3	100	0.06	0.46	0.55	0.61					
All sessions	**281**	**0.07**	**0.38**	**0.65**	**0.75**					
6-move game (Figure 4)[a]										
Session 1	100	0.02	0.09	0.44	0.56	0.91	0.50			
Session 2	81	0.00	0.02	0.04	0.49	0.72	0.82			
Session 3	100	0.00	0.07	0.15	0.54	0.64	0.92			
All sessions	**281**	**0.01**	**0.06**	**0.21**	**0.53**	**0.73**	**0.85**			
9-move game (Figure 5)[b]										
Session 1	300	0.46	0.59	0.50	0.46	0.44	0.60	0.75[c]	1.00	—[d]
Session 2	300	0.39	0.46	0.47	0.50	0.35	0.29	0.58	0.80[c]	1.00[c]
Session 3	300	0.30	0.40	0.45	0.41	0.39	0.44	0.21	0.09	1.00[c]
Session 4	300	0.41	0.43	0.55	0.50	0.48	0.42	0.57[c]	0.33[c]	1.00[c]
All sessions	**1200**	**0.39**	**0.47**	**0.49**	**0.46**	**0.42**	**0.44**	**0.52**	**0.56**	**1.00**

[a]McKelvey and Palfrey (1992).
[b]Rapoport *et al.* (2000).
[c]Based on fewer than 10 observations.
[d]Undefined.

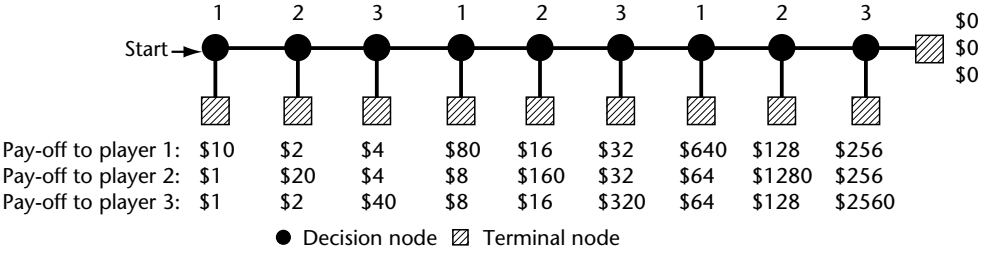


Figure 5. Three-person nine-move centipede game (Rapoport *et al.*, 2000).

and three full rounds of play (nine decision nodes). This game is similar to the one in Figure 1, in that the decision to continue on the last decision node results in zero pay-off to all players. If play continues for at least three stages (a single round), all three pay-offs increase eightfold. But if a player continues and either of the other two players exits at the next opportunity, the player who continues is worse off.

Although the six-move centipede game of McKelvey and Palfrey and the nine-move game of Rapoport *et al*. each includes three rounds of play, and the 'pie' doubles in size on each stage, the two studies are not directly comparable. Firstly, they differ in the number of players. Secondly, they differ in the assignment of pay-offs to the outcome resulting if all players always continue. Thirdly, they differ in the number of iterations of the game: 10 and 60 respectively. Fourthly, they differ in the method of assigning roles to players. Rapoport *et al*. conducted four sessions, each including a population of 15 participants who were divided on each iteration into five groups of three players each. Group membership was changed randomly from one iteration to another, as was the assignment of the group members to the three player roles. Although players received information only about their own pay-offs, their decisions could slowly and indirectly affect the behavior of the other members of the player population. Thus, population learning was possible. The fifth and possibly most important difference was in the size of the stakes. If player 3 exits on the last decision node in the study of Rapoport *et al*., his

pay-off is \$2560. Actual payment was half of the nominal earnings, so the maximum pay-off was a factor of 200 higher than in Figure 3 and 50 times higher than in Figure 4.

The bottom panel of Table 1 presents the proportions of games ending at each of the nine terminal nodes. The results are presented separately for each of the four sessions (five groups, with 60 iterations per group, for a total of 300 plays per session) and then combined across all sessions. There are three major findings. Firstly, except for the 9th decision node in session 1, all the nine decision nodes were reached in each session. Secondly, in contrast the results reported by McKelvey and Palfrey, the proportions of exit decisions tended to decrease across the nine decision nodes. Thirdly, and most importantly, backward induction was supported 46.3%, 39.3%, 30.3%, and 40.7% of the time across the four sessions. These results differ markedly from those reported by McKelvey and Palfrey. The bottom panel of Table 2 presents the values of $P(E|j)$ at each decision node j. If we exclude the inferred conditional probabilities $P(E|j)$ that are based on fewer than 10 observations (these were mostly contributed by a small subset of participants), the values of $P(E|j)$ are seen to be remarkably stable across the decision nodes 2 through 8. This finding, too, differs from the findings of McKelvey and Palfrey shown in the upper two panels of Table 2.

The most important result in the study of Rapoport *et al*. concerns the dynamics of play. Figure 6 shows (in three parts) the arithmetic moving average (in steps of 10) of the $P(E|j)$ values. The results are presented separately for the first three decision nodes in round 1 ($j=1, 2, 3$). For $j=1$ (top panel), the initial propensity in the first 10 or so iterations is to continue; the mean values of $P(E|j)$ in the first 10 iterations are about 0.10. However, all four functions in the top panel of Figure 6 tend to increase with further iterations. The values of $P(E|j=1)$ are seen to increase steadily in sessions 1 and 4. In session 2 they reach a plateau after 20 or so iterations but start increasing again after the 40th iteration. In session 3, the values of $P(E|j=1)$ are rather flat for the first 45 or so iterations and then start increasing. Although the trends in $P(E|j=2)$ (middle panel) and $P(E|j=3)$ (bottom panel) are also generally increasing, they are more difficult to characterize.

What we observe in each of the four sessions is a slow breakdown in mutual trust resulting in a trend towards the backward induction solution. No such trends were reported by McKelvey and Palfrey. Any of the differences in the design of the two experiments (particularly the difference in size of stakes), or a combination of them, may account for the differences in the results.

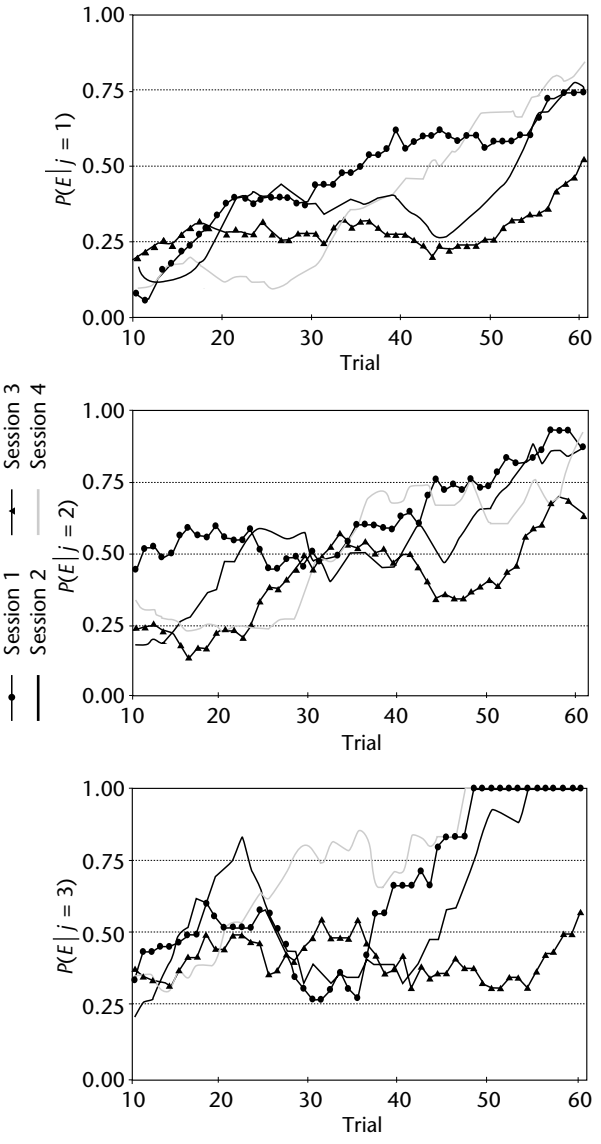


Figure 6. Arithmetic moving averages of the inferred conditional probabilities of exit on decision node j ($j = 1, 2, 3$) across players by session (Rapoport *et al*., 2000). The results for the first, second, and third decision nodes are displayed in the top, middle, and bottom panels, respectively. For $j=3$ (bottom panel) there were insufficient data in session 4 beyond trial 48.

CONCLUSION

The centipede game has extended and intensified the discussion of fundamental concepts of the evolving discipline of game theory. There is a need for further theoretical work on the concepts

of rationality, common knowledge, backward induction, and beliefs in interactive decision-making, as well as related experimental work on the effects of size of stakes, number of players, type of interaction, and rationality of the players on conflict, cooperation, and trust.

References

Aumann RJ (1992) Irrationality in game theory. In: Dasgupta P, Gale D and Maskin E (eds) *Economic Analysis of Markets and Games*, pp. 214–227. Cambridge, MA: MIT Press.

Aumann RJ (1995) Backward induction and common knowledge of rationality. *Games and Economic Behavior* **8**: 6–19.

Aumann RJ (1998) On the centipede game. *Games and Economic Behavior* **23**: 97–105.

Ben-Porath E (1997) Rationality, Nash equilibrium and backwards induction in perfect-information games. *Review of Economic Studies* **64**: 23–46.

Feinberg Y (2001) Subjective formulation and analysis of games and solutions. [Stanford Graduate School of Business, unpublished manuscript.]

McKelvey RD and Palfrey TR (1992) An experimental study of the centipede game. *Econometrica* **60**: 803–836.

Ponti G (2000) Cycles of learning in the centipede game. *Games and Economic Behavior* **30**: 115–141.

Rapoport A, Stein WE, Parco JE and Nicholas TE (2000) Equilibrium play and adaptive learning in a three-person centipede game. [University of Arizona, Department of Management and Policy, unpublished manuscript.]

Rosenthal RW (1981) Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* **25**: 92–100.

Stalnaker R (1998) Belief revision in games: forward and backward induction. *Mathematical Social Sciences* **36**: 31–56.

Further Reading

Aumann RJ and Brandenburger A (1995) Epistemic conditions for Nash equilibrium. *Econometrica* **63**: 1161–1180.

Balkenborg D and Winter E (1997) A necessary and sufficient epistemic condition for player backward induction. *Journal of Mathematical Economics* **27**: 325–345.

Binmore K (1992) *Fun and Games: A Text on Game Theory*. Lexington MA: DC Heath.

Brandenburger A (1992) Knowledge and equilibrium in games. *Journal of Economic Perspectives* **6**: 83–101.

Fey M, McKelvey RD and Palfrey TR (1996) An experimental study of constant-sum centipede games. *International Journal of Game Theory* **25**: 269–287.

McKelvey RD and Palfrey TR (1998) Quantal response equilibria for extensive form games. *Experimental Economics* **1**: 9–41.

Nagel R and Tang FF (1998) Experimental results on the centipede game in normal form: an investigation on learning. *Journal of Mathematical Psychology* **42**: 356–384.

Reny PJ (1993) Common belief and the theory of games with perfect information. *Journal of Economic Theory* **59**: 257–274.

Stalnaker R (1996) Knowledge, belief, and counterfactual reasoning in games. *Economics and Philosophy* **12**: 133–163.

Zauner KG (1999) A payoff uncertainty explanation of results in experimental centipede games. *Games and Economic Behavior* **26**: 157–185.

Games: Coordination

Intermediate article

Jacob K Goeree, University of Virginia, Charlottesville, Virginia, USA
Charles A Holt, University of Virginia, Charlottesville, Virginia, USA

CONTENTS

Coordination games typically possess multiple Nash equilibria, some of which are preferred by one or more players. Coordination on desired equilibria can be facilitated by communication, repetition, and introspection.

INTRODUCTION

Despite the well-known efficiency properties of competitive markets, economists have long been concerned that there may be multiple equilibria in complex social systems. The possibility of becoming mired in a bad equilibrium dates back at least to Thomas Malthus's notion of a 'general glut'. The problem is one of coordination: e.g. it may not make sense for me to go to the market to trade if you are going to stay home on the farm. This situation is one of common interest in that everyone prefers the equilibrium with high market activity to the one without any trade. Coordination problems can also arise when there is a conflicting interest, i.e. when each person prefers a different equilibrium outcome. The classic example is the 'battle-of-the-sexes' game, where the man prefers to attend a baseball game and the women prefers to attend an opera, but both would rather do something together than go to separate events. In all of these examples there are multiple outcomes that are equilibria in the sense that no single individual has an incentive to deviate if others are conforming to that outcome. For instance, in the battle-of-the-sexes, the man would attend the opera if he thinks the woman will be there even though he prefers the other equilibrium outcome in which both attend the baseball game.

Society has developed a number of ways to solve such coordination problems. The most obvious cases involve social norms and rules, e.g. driving on the left side of the road in the UK and Japan. In the absence of explicit rules, individuals may rely on focalness, e.g. Schelling (1980) points out that two people who are supposed to meet each other somewhere in New York city are likely to go to Times Square. Historical accidents and precedents can also be used as coordination devices, since past experience often affects beliefs about others' future behavior. For example, prior discrimination against workers of a particular racial background may produce a situation in which the workers do not invest in skills because they anticipate unfavorable job assignments. These beliefs can be self-fulfilling since the workers' choices make it rational for the employer to expect low performance from those workers. In this case, a period of 'reversed' discrimination, e.g. affirmative action, may be needed to change expectations and allow coordination on the preferable merit-based equilibrium. Coordination can also be achieved through explicit communication and reciprocity, e.g. in the battle-of-the-sexes game the man and the woman agree to attend the baseball game this weekend and opera next.

Economists and psychologists are also interested in strategic settings where opportunities for communication and reciprocity are limited, perhaps due to the large number of people involved. Rapoport *et al*. (1998), for example, consider a large group of sellers who must decide whether or not to enter a particular market that is profitable only in the absence of excessive entry. Even without communication, low profits will drive some sellers away while high profits will attract more sellers, and observed behavior in laboratory experiments shows that the costs of uncoordinated outcomes can be negligible. Laboratory experiments such as these have uncovered interesting patterns of behavior that have led to theoretical models of equilibrium selection. These models of learning and introspection provide natural extensions of classical equilibrium concepts, which are unable to predict which outcome has the strongest drawing power.

TYPES OF COORDINATION GAMES

A typical coordination game with conflicting interests is the battle-of-the-sexes game where the preferred equilibrium provides a player with a high payoff of 600, while the less preferred equilibrium only yields 200. This game, shown in Table 1, was used by Cooper *et al*. (1989) in a laboratory experiment that was presented in neutral terms. There are two obvious Nash equilibria of this game: (baseball, baseball) and (opera, opera). In addition, there is an equilibrium involving randomization, where both the man and woman choose their preferred activity with probability 0.75. In this manner, each person is indifferent between the two options since each yield a payoff of 150 on average ($0.75 \times 200 = 0.25 \times 600 = 150$).

Another coordination game with opposing interests is the market entry game shown in Table 2. Two firms have to decide whether or not to incur a cost of 50 to enter a market. The entrant's profits will be 150 if it is alone, but will be zero if both firms enter. The profits from staying out are zero independent of the rival's choice. Again there are two obvious Nash equilibria and each firm prefers the outcome in which it is the sole entrant. Notice that the case of excess entry produces losses for each firm, which makes this a 'game of chicken'. This terminology is inspired by the movie *Rebel Without a Cause*, where James Dean and his competitor drive toward a cliff and the first to stop is considered to be 'chicken'. Each prefers the outcome where the other one stops first, but if neither stops both incur a severe cost (possibly worse than -50).

Table 1. A battle-of-the-sexes game (woman's payoff, man's payoff)

		Man	
		Opera	Baseball
Woman	Baseball	0, 0	200, 600
	Opera	600, 200	0, 0

Table 2. A market entry game (Firm 1's payoff, Firm 2's payoff)

		Firm 2	
		Stay Out	Enter
Firm 1	Stay Out	0, 0	0, 100
	Enter	100, 0	-50, -50

Rousseau's classic 'stag hunt' game describes a situation in which two hunters decide to hunt for stag or hare. Each hunter alone could be sure of bagging a hare, but both hunters are needed to corner the stag, which is the preferred outcome. If only one hunts stag, that person is left empty handed, so there is a low payoff equilibrium in which both hunt hare and a high payoff equilibrium in which both hunt stag. Unlike the two previous examples, this game is one of common interests since both prefer the 'better' outcome. Rousseau's game has a 'weakest link' property since the stag will escape through any sector left unguarded by a hunter. This is analogous to a situation where workers provide different parts to be assembled, and the final product requires all parts. In this case, the total amount produced is determined by the slowest worker, i.e. the weakest link in the production chain.

One way to model a weakest-link, or minimum-effort coordination game is to let players choose effort levels, 1 or 2, where each unit of effort results in a cost of $c < 1$. The output per person is determined by the minimum of the effort levels chosen. For example, if both choose an effort level of 1, then each player receives a payoff of $1 - c$, as shown in the upper-left box of Table 3. Similarly, when both choose a high effort level of 2, they each obtain $2 - 2c$. But if they fail to coordinate with effort choices of 1 and 2, then the minimum is 1 and payoffs are $1 - c$ for the low effort individual and $1 - 2c$ for the high effort individual. Notice that the low effort outcome is an equilibrium since a costly increase in effort by only one person will not raise the amount produced. The high effort outcome is also an equilibrium, since a reduction in effort by only one person will lower the minimum by more than the cost savings c.

The game in Table 3 can be generalized to allow for more than two players and multiple effort levels, with payoffs determined by the minimum effort level chosen. If a player's effort is denoted by e_i, $i = 1, \ldots, n$, payoffs are:

$$\pi_i(e_1, \ldots, e_n) = \min\{e_1, \ldots, e_n\} - ce_i, \quad i = 1, \ldots, n, \tag{1}$$

Table 3. A 2×2 coordination game (Player 1's payoff, Player 2's payoff)

		Player 2's effort	
		1	2
Player 1's effort	1	$1 - c$, $1 - c$	$1 - c$, $1 - 2c$
	2	$1 - 2c$, $1 - c$	$2 - 2c$, $2 - 2c$

where c is the per-unit effort cost. As long as c is less than 1, payoffs are maximized when all players choose the highest possible effort. Note, however, that *any* common effort level constitutes a Nash equilibrium, since a costly unilateral increase in effort will not raise the minimum, and a unilateral decrease will reduce the minimum by more than the cost when $c < 1$. This argument does not depend on the number of players, so noncritical changes in c and n will not alter the set of Nash equilibria, despite the reasonable expectation that efforts should be high for sufficiently low effort costs and low numbers of participants.

EXPERIMENTAL EVIDENCE

Van Huyck *et al*. (1990) used a minimum-effort structure in one of the most widely cited and influential game theory experiments. In their experiment subjects could choose seven possible effort levels, 1 through 7, and payoffs were a linear function of the difference between the minimum effort and one's own divided by two ($c = ½$). Thus there are seven equilibria in which all players choose the same effort level, but the equilibrium with the highest payoff is the one where all players choose an effort of 7. The experiment involved 14–16 players who made independent effort choices. After choices were collected and the minimum was announced, payoffs were calculated. This whole process was repeated for 10 rounds. Van Huyck *et al*. report that efforts declined dramatically, with the final-period efforts clustered at the equilibrium that is worst for all. This result surprised many game theorists, who were comfortable assuming that rational individuals would be able to coordinate on an outcome that is best for all. Van Huyck *et al*. also find that an extreme reduction in the cost of effort ($c = 0$) results in an overwhelming number of high effort decisions, which is not surprising since raising effort is costless.

Goeree and Holt (2000) explored the effects of changes in effort cost and the number of players more systematically. They reported two- and three-player minimum-effort coordination game experiments with varying effort costs. Individuals could choose continuous effort levels in the range from 110 to 170 with payoffs determined as in (1). With two players, average effort levels are initially around the midpoint 140 in both the low-cost ($c = ¼$) and high-cost ($c = ¾$) treatments. By the third period, however, there is a strong separation, with higher efforts for the low cost treatment. In the final rounds, the average effort was 159 in the low cost treatment and 126 in the high cost treatment. Even though *any* common effort in the range from 110 to 170 is a Nash equilibrium independent of the effort cost, the observed behavior is affected by the magnitude of the effort cost in an intuitive manner. Similarly, Goeree and Holt (2000) found that an increase in the number of players tends to decrease the final-period effort levels.

A different line of experimentation involved factors that facilitate coordination. Cooper *et al*. (1989) investigated the effects of 'cheap talk' forms of communication in a battle-of-the-sexes game as shown in Table 1. Subjects first submitted a message about which choice they intended to make. These messages were nonbinding in the sense that a player could deviate from his or her reported intent after seeing the other's message. When both messages matched one of the equilibria, i.e. (baseball, baseball) or (opera, opera), then actual decisions corresponded to stated intentions more than 80 percent of the time. When messages differed, however, individuals tended to deviate from their original message and chose their preferred activity about 71 percent of the time. Communication can therefore facilitate coordination, but when communication fails the behavior corresponds more closely to the equilibrium in randomized strategies (choosing the preferred activity with probability 0.75).

The Cooper *et al*. (1989) experiments involved random pairings of subjects in each period. In contrast, fixed pairings permit coordination that is based on the history of past decisions. Prisbrey (1991) finds a common pattern of alternating choices, with outcomes (baseball, baseball) and (opera, opera) in successive rounds. (As in the other papers mentioned above, Prisbrey used neutral terminology in presenting the payoffs to subjects.) This alternation can be interpreted as a form of reciprocity, where 'nice' behavior in one round is rewarded by a nice response in the next. A number of other coordination experiments are surveyed in Ochs (1995).

THEORETICAL EXPLANATIONS

One of the most commonly suggested criteria for the analysis of games with multiple equilibria is to select the one with the highest payoffs for all, if such a 'Pareto-dominant' outcome exists. The Van Huyck *et al*. (1990) experiment showed that this method is incorrect, and the Goeree and Holt (2000) experiment showed that any explanation must take into account the effort cost and the number of players.

Harsanyi and Selten's (1988) notion of risk dominance is sensitive to the effort cost that determines

the losses associated with deviations from best responses to others' decisions. To illustrate the concept of risk dominance, consider the two-person minimum-effort game shown in Table 3. When both players are choosing efforts of 1, the cost of a unilateral deviation to 2 is only the cost of the extra effort, c, which will be referred to as the 'deviation loss'. Similarly, the deviation loss at the (2, 2) equilibrium is $1-c$, since a unilateral reduction in effort reduces the minimum by 1 but saves the marginal effort cost c. The deviation loss from the low effort equilibrium is greater than that for the high effort equilibrium if $c>1-c$, or equivalently, if $c>½$, in which case we say that the low effort equilibrium is risk dominant. Risk dominance, therefore, has the desirable property that it selects the low effort outcome if the cost of effort is sufficiently high.

There is, however, no consensus on how to generalize risk dominance for games with more players, a continuum of decisions, etc. A related concept that does generalize is the notion of maximization of a 'potential' of a game. Loosely speaking, the idea behind a potential is to find a function for a game that is maximized by a Nash equilibrium for that game. Stated differently, a potential function is a mathematical formula that is positively related to individual players' payoffs: when a change in a player's own decision raises that player's payoff, then this change necessarily raises the value of the potential function by the same amount, and vice versa for decreases. If such a potential function exists for the game, then each person trying to increase their own payoff may produce a group result that maximizes the potential function for the game as a whole. Think of two people holding adjacent sides of a treasure box, with one pulling uphill along the East–West direction and the other pulling uphill along the North–South axis. Even though each person is only pulling in one direction, the net effect will be to take the box to the top of the hill, where there is no tendency to change (a Nash equilibrium that maximizes potential).

For instance, for the n-player minimum effort game given in (1), the potential function is simply the common production function that determines a *single* player's payoff, minus the sum of *all* players' effort costs:

$$V(e_1,\ldots,e_n)=\min\{e_1,\ldots,e_n\}-c\sum_{i=1}^{n}e_i. \quad (2)$$

The maximization of potential will obviously require equal effort levels. At any common effort, e, the potential in (2) becomes: $V=e-nce$, which is maximized at the lowest effort when $nc>1$, and is maximized at the highest effort when $nc<1$. In two-person games, this condition reduces to the risk dominance comparison of c with ½.

The notion of potential can be used to evaluate results from previous laboratory experiments. Van Huyck *et al*. (1990) conducted games with 14–16 players and an effort cost of either 0 or ½, so nc was either zero or about seven. Compared to the critical nc value of 1, these parameter choices appear rather extreme which may explain why their data exhibit a huge shift in effort decisions. By the last round in the experiments in which $nc=0$, almost all (96 percent) participants chose the highest possible effort, while over three-quarters chose the lowest possible effort when nc was around seven. One purpose of Van Huyck *et al*. (1990) was to show that a Pareto-inferior outcome may arise in coordination games, presumably because it is harder for large numbers of participants to coordinate on good outcomes. Other experiments were conducted with two players, but the payoff parameters were such that nc exactly equaled the critical value 1, and, with a random matching protocol, the data showed a lot of variability. The Goeree and Holt (2000) experiment also implements two-person random pairings in order to avoid the serious possibility of tacit collusion in repeated games which may drive efforts to maximal levels in sufficiently long series of repeated two-person coordination games. Given the knife-edge properties of $c=½$ for two-person coordination games, we conducted one treatment with $nc=½$ and another with $nc=⅔$. As noted above, this change has a large effect on observed final-period effort choices even though the set of Nash equilibria is unaffected.

Risk dominance (or some generalization) predicts well in the long run, but the patterns of adjustment suggest that theories of learning and adaptation play an important role in the analysis of coordination. Each round of an experiment provides some information about others' decisions and the resulting payoffs, and individuals may adapt their behavior in the direction of a best response to past decisions. This is the approach taken by Crawford (1995), who specifies and estimates a model in which a person's decision is a weighted average of the person's past decision and the decision that would have been a best response to others' decisions in the past. This model incorporates some type of inertia, as well as an adaptive response. Another low cognitive model is that players respond to payoffs received. These reinforcement

learning models have been successfully applied by Erev and Roth (1997). In contrast, belief learning models assume that players use the history of past decisions to predict their opponents' next choice. Camerer and Ho (1999) have developed a hybrid model that combines elements of both reinforcement and belief learning models. Learning models such as these are especially useful in predicting how final-period behavior may be explained by the history of past decisions.

Finally, many games in real life are played only once: e.g. elections, military contests, and legal disputes. In such cases, there is no relevant past history and players must rely on introspection about what others might do. This problem is particularly interesting for coordination games with multiple equilibria:

> It should be emphasized that coordination is not a matter of guessing what the 'average man' will do. One is not, in tacit coordination, trying what another will do in an objective situation; one is trying to guess what the other will guess one's self to guess the other to guess, and so on ad infinitum. (Schelling, 1980, pp. 92–93)

Formal models of introspection often have such an iterative structure, incorporating responses to others' conjectured responses to others' conjectures, etc. Although these models are less well developed, they have shown some promise for explaining behavior in games played only once (Goeree and Holt, 2001).

References

Camerer C and Ho TH (1999) Experience weighted attraction learning in normal-form games. *Econometrica* **67**: 827–874.

Cooper R, DeJong DV, Forsythe R and Rust TW (1989) Communication in the battle of the sexes game: some experimental results. *Rand Journal of Economics* **20**: 568–587.

Crawford VP (1995) Adaptive dynamics in coordination games. *Econometrica* **63**: 103–144.

Erev I and Roth AE (1997) Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88**: 848–881.

Goeree JK and Holt CA (2000) An experimental study of costly coordination. Working Paper, University of Virginia.

Goeree JK and Holt CA (2001) Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* **91**(5): 1402–1422.

Harsanyi J and Selten R (1988) *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Ochs J (1995) Coordination problems. In: Kagel J and Roth A (eds) *Handbook of Experimental Economics*, pp. 195–249. Princeton: Princeton University Press.

Prisbrey J (1991) An experimental analysis of the two-person reciprocity game. Working Paper, California Institute of Technology.

Rapoport A, Seale DA, Erev I and Sundali JA (1998) Equilibrium play in large group market entry games. *Management Science* **44**: 129–141.

Schelling TC (1980) *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Van Huyck JB, Battalio RC and Beil RO (1990) Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review* **80**: 234–248.

Further Reading

Anderson SP, Goeree JK and Holt CA (2001) Minimum effort coordination games: stochastic potential and logit equilibrium. *Games and Economic Behavior* **34**: 177–199.

Cachon G and Camerer C (1996) Loss-avoidance and forward induction in experimental coordination games. *Quarterly Journal of Economics* **111**: 165–194.

Coate S and Loury G (1993) Will affirmative action eliminate negative stereotypes? *American Economic Review* **83**: 1220–1240.

Cooper R and John A (1988) Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* **103**: 441–464.

Goeree JK and Holt CA (1999) Stochastic game theory: for playing games, not just for doing theory. *Proceedings of the National Academy of Sciences* **96**: 10564–10567.

Sefton M (1999) A model of behavior in coordination game experiments. *Experimental Economics* **2**: 151–164.

Straub PG (1995) Risk dominance and coordination failures in static games. *Quarterly Review of Economics and Finance* **35**: 339–363.

Van Huyck JB, Battalio RC and Rankin FW (1997) On the origin of convention: evidence of coordination games. *Economic Journal* **107**: 576–596.

Games: Dictator

Introductory article

Mary Rigdon, George Mason University, Fairfax, Virginia, USA

CONTENTS

A dictator game is a game between two players, a 'dictator' and a 'serf'. The task is for the dictator to split some amount M *of money between the two of them.*

GAME-THEORETIC DESCRIPTION

A *dictator game* is a game between two players: a *dictator* (D), and a *serf* (S). The task is for the dictator to split some amount M of money (or whatever it is that the players have well-behaved utility functions for) between the two players. The serf cannot respond: the split decided upon by the dictator is the final allocation. Non-cooperative game theory makes a simple prediction about how dictators will behave in these games: a self-interested, non-satiated dictator will take M, leaving nothing for the serf. (*See* **Games: Ultimatum**)

Assume that M is a finite sum of money, and that it is discrete; i.e. there is a smallest unit of account c. We can think of the choice facing D as the choice among the different amounts of M that D can keep (since whatever D does not keep S gets). The dictator then has the pure strategy space $\mathfrak{S}_D = \{M - 0, M - c, \ldots, M - M\}$. S has only the empty strategy space, since S has no actions at all: $\mathfrak{S}_S = \emptyset$. The outcome of the game is solely determined by the action of D, and so the outcome is a reflection of D's preferences. A rational D prefers more money (or other utility) to less. Therefore, a rational D will take all of M and leave the serf with nothing. That is, $M - 0$ strictly dominates all other choices open to D, since, no matter what S does, D is better off doing $M - 0$ than anything else. The strategy profile in which D chooses $M - 0$ is a Nash equilibrium: no player in the game could achieve a better outcome by making a different choice (holding the actions of the other player constant).

BASIC EXPERIMENTAL RESULTS

Imagine you have been invited to participate in an experiment. You enter the laboratory and are paid $3 (or sometimes more) for arriving on time. There are 12 people in the room, spaced well apart. You begin by reading a set of instructions at your own pace, which are then read aloud by the experimenter. You are randomly assigned the role of person A (the dictator) or person B (the serf). Each experimental session involves only one pairing and only one decision by the randomly selected dictator. You will not be told who your counterpart is, either during or after the experiment, and they will not be told who you are, either during or after the experiment. The instructions state that 'a sum of $10 has been provisionally allocated to each pair, and person A in each pair can propose how much of this each person is to receive … The amount that person A is to receive is simply the amount to be divided, $10, minus the amount that person B is to receive'. If you are randomly chosen as a dictator, you indicate on a form how much of M you wish to keep for yourself and how much the serf will therefore receive. Once all of the dictators have made their decisions, the experiment is complete and you are paid your dollar earnings privately by the experimenter. (*See* **Economics, Experimental Methods in**)

Contrary to the game-theoretic prediction, many dictators offer more than $0 to their serf: 20% do give nothing, but 20% give away an equal share of the pie, $5. A large proportion of dictators transfer between 1 and 5 dollars: approximately 17% pass $1, 13% pass $2, 30% pass $3, and 0% pass $4. There are virtually no offers greater than half of the pie.

TREATMENT EFFECTS

Given the experimental results in the dictator game, which do not conform to the theoretical prediction, we need to investigate whether the results are robust to changes in the design of the experiment. Results of interpersonal experiments like the dictator game may be sensitive to the instructional and procedural settings of the experiment.

The instructions quoted above do not define a clear property right to the sum of $10. In fact, they suggest that it provisionally belongs to both of the participants. Will the dictator's offer be more self-regarding if players have to earn the right to move first, rather than being randomly given the role? One way to implement a contest treatment is to use subjects' scores on quizzes about current events to determine who will be the dictator in the pairings. Suppose subjects are ranked from highest to lowest using the number of correct answers given on the quiz, the top half earning the right of dictatorship and the bottom half becoming their serfs. Suppose this pairing rule is common knowledge to the subjects. Such a treatment lowers the offer distribution relative to the base-line. In fact, the fraction who offer $0 more than doubles, and there are now virtually no offers of equal division.

When role entitlement is combined with framing of the exchange as that between a seller and a buyer, where the seller sets the price for a precommitted buyer, the results are even more skewed towards self-regarding outcomes: 40% of the dictators now offer $0, and another 40% offer $1 or $2. Only 4% offer $4, and there are virtually no offers of equal division. This difference is significant at the 99% level of confidence.

In the above experiments, subjects are anonymous with respect to each other, but the experimenter still knows who made what decisions. The experimental protocol is 'single-blind'. Under such a protocol, subjects do not know the identities of their counterparts, and so fear of social reprisal by their peers cannot be an explanation of the results. However, there is another variable which affects social distance: the experimenter knows the identities of the subjects and actions each takes in the experiment. A natural question is whether a 'double-blind' protocol, in which neither other subjects, the experimenter, nor anyone else can map subjects to their actions, would yield different results.

A standard version of a double-blind dictator experiment runs as follows. Each (randomly selected) dictator is given an envelope which contains 10 one-dollar bills and 10 blank slips of paper. Each proceeds in turn to the back of the room, opens the envelope, and makes his or her decision privately. The task is to keep 0 to 10 of the one-dollar bills and 10 to 0 of the slips of paper, so that the number of bills plus the number of slips of paper in the serf's envelope equals the total size of the allocation ($10). The dictators leave the experiment and the envelopes are transferred to the room with the serfs. A paid monitor calls each person one at a time. The serf selects and opens an envelope, keeping the contents. The monitor records the amount on a sheet of paper without any names on it. The double-blind treatment has a very significant effect on dictator giving: more than two-thirds of the dictators offer $0, with 84% offering $0 or $1. A very small number (about 8%) offer $5. The results in this treatment more closely approach the game-theoretic prediction. These results from double-blind dictator games have been replicated by several researchers, suggesting they are indeed robust.

The results of experimental work on dictator games suggests three lessons: (1) the concept of Nash equilibrium neither adequately predicts nor explains all of the observed behavior in these games; (2) there is a significant 'property right' effect on dictator giving; and (3) the results are sensitive to procedural variation in the experiments themselves – varying the 'social distance' between dictator and serf, and between subject and experimenter – shifts the outcomes in systematic and interesting ways.

Further Reading

Berg J, Dickhaut J and McCabe K (1995) Trust, reciprocity, and social history. *Games and Economic Behavior* **10**: 122–142.

Camerer C and Thaler RH (1995) Anomalies: ultimatums, dictators, and manners. *Journal of Economic Perspectives* **9**: 209–219.

Croson RTA (1996) Information in ultimatum games: an experimental study. *Journal of Economic Behavior and Organization* **30**: 197–212.

Eckel C and Grossman P (1996) Altruism in anonymous dictator games. *Games and Economic Behavior* **16**(2): 181–191.

Eckel C and Grossman P (2000) Volunteers and pseudo-volunteers: the effect of recruitment method in dictator games. *Experimental Economics* **3**(2): 107–120.

Forsythe R, Horwitz J, Savin NE and Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic Behavior* **6**: 347–369.

Hoffman E, McCabe K, Shachat K and Smith V (1994) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* **7**: 346–380.

Hoffman E, McCabe K and Smith V (1998) Behavioral foundations of reciprocity: experimental economics and evolutionary psychology. *Economic Inquiry* **36**: 335–352.

Roth AE (1995) Bargaining experiments. In: Kagel JH and Roth AE (eds) *The Handbook of Experimental Economics*, pp. 253–348. Princeton, NJ: Princeton University Press.

Smith V (1994) Economics in the laboratory. *Journal of Economic Perspectives* **8**(1): 113–131.

Games: Principal–Agent

Intermediate article

Corinne Alexander, University of California, Davis, California, USA

CONTENTS

Principal–agent games address the situation where one party (the agent) can take an action unobserved by the other party (the principal), who bears the full consequences for that action. Anticipating that the agent may not act in the principal's best interest, the principal offers the agent incentives to influence the agent's actions.

INTRODUCTION

Consider the relationship between the shareholders or owners of a firm (the principal) who delegate the task of running the firm to the managers (the agent). The shareholders may be concerned that the managers are more interested in perks such as private jets than in working hard to benefit the shareholders. This problem of moral hazard (also called 'hidden action') arises when there is a conflict of interest between the agent who is taking an action for which the principal bears the consequences and the principal is unable to observe the agent's choice. The 2002 Enron scandal provides a dramatic example of moral hazard: the company executives created off-the-book partnerships to hide losses incurred from some poor decisions, and once these accounting practices were discovered, the shareholders lost their investment. Fundamentally, moral hazard also lowers the welfare of the agent, especially if the principal chooses not to hire the agent due to anticipated losses from moral hazard. Subsequent to the Enron scandal, investors have been reluctant to enter the market and companies face the challenge of convincing investors that they will act in their best interests.

Since both the principal and the agent profit from their relationship, both have the incentive to ensure that the agent delivers what the principal wants. The principal can alleviate the problem of moral hazard by giving the agent incentives to perform in the best interests of the principal. For example, since the shareholders of a firm want the managers to work hard to increase profits, they may make the managers' wages contingent on the firm's profit by offering stock options. Alternatively, the agent actively seeks the principal's trust by building a reputation or hiring monitoring services such as those offered by auditors.

Principal–agent theory predicts how rational individuals will behave in a complex decision-making environment, where one individual may rationally cheat another. Laboratory experiments have tested the predictions of principal–agent theory and tested other behavioral predictions based on reciprocity.

PRINCIPAL–AGENT MODEL

There are three key features to the principal–agent model: first, the principal delegates tasks to the agent, requiring the agent to exert effort; second, the agent's choice of effort is unobservable to the principal; and third, there is a conflict of interest since that effort is costly to the agent. The principal's solution to the moral hazard problem is to offer the agent an incentive compatible contract that aligns the agent's objectives with the principal's. A contract is incentive compatible when, given the contract specifications, the agent prefers the principal's desired action to all other possible actions. The principal uses a combination of punishments and rewards contingent on the outcome of the task which induce the agent to work hard.

The principal–agent relationship can be modeled as a two-stage game. In the first stage, the principal moves first and offers the agent a take-it-or-leave-it contract from the set of incentive compatible contracts. In the second stage, the agent accepts or rejects the contract. If the agent accepts the contract, then the agent chooses the optimal action or effort level in response to the contract specifications. The outcome or profit is a function of the agent's chosen effort. Ideally, the principal would like to appropriate all the gains from trade. However, the principal must offer the agent some of the profits in order

to incentivize the agent. (*See* **Game Theory in Economics**)

Full Information Benchmark

If effort is verifiable (either effort is directly observable or output is a perfect predictor of effort), then it is straightforward and costless for the principal to delegate a task with correct incentives. The principal can offer the agent a contract where payment is contingent on the desired action. One example of an optimal contract when output is a perfect predictor of effort is the piece-rate contract offered to strawberry pickers. The pickers are paid for the number of flats of strawberries picked which provides clear incentives. In contrast, according to principal–agent theory, an hourly wage would give the workers no incentive to work hard since they would receive the same pay even if they sat in the shade drinking lemonade.

Asymmetric Information

The principal's problem of designing a contract to reduce the problem of moral hazard becomes more complex when there is asymmetric information about the agent's effort. If the agent's action is hidden, the relationship between the unobservable action and the observable outcome must be probabilistic – where there is a higher probability of a good outcome if the agent exerts high effort. However, even if the agent takes the desired action, a bad outcome could occur. In the case of a bad outcome, the principal will be unable to distinguish whether it was caused by the agent's choice of action or events outside the agent's control.

In designing the incentive compatible contract, the principal has to take account of the following: the agent's risk preferences, any limits on the agent's liability, and additional sources of information on the agent's behavior.

Risk Preferences

Most individuals are 'risk averse', meaning that they dislike being faced with uncertain situations. For example, if the risk averse individual is offered a choice between two gambles with the same expected payoff, the individual will prefer the less risky gamble. Suppose the individual is offered the choice between a gamble with a 50 percent chance of earning \$40 and a 50 percent chance of earning \$0 and a second gamble where there is a 50 percent chance of earning \$30 and a 50 percent chance of earning \$10, both yielding expected winnings of \$20. The second gamble is less 'risky' because the bad outcome of \$10 is better than \$0. The risk averse individual prefers gambles where the bad outcome is less bad, even if it means forgoing the chance of a better outcome, in this case accepting a good outcome of \$30 over the good outcome of \$40. In order for the risk averse individual to be indifferent between the two gambles, the first gamble with a 50 percent chance of \$0 will have to be associated with a good outcome large enough so that the individual can expect to win more than the \$20 offered by the second, less risky gamble. In contrast, a risk neutral individual would be indifferent between the two gambles. (*See* **Choice under Uncertainty**)

In the context of the principal–agent model, when the principal conditions the agent's wage on the outcome, the agent shares the risk of the bad outcome faced by the principal. Risk sharing means dividing up the proceeds from a risky gamble among individuals.

Insurance provides the best example of risk sharing. Consider fire insurance for a community of a hundred homes. Suppose the probability of any one house getting burned down in any year is 0.01, so that on average one house is destroyed by fire each year. If the average cost of a home is \$50,000, then the community expects to lose \$50,000 each year in fire damage. While the risk of losing his or her home is large for each individual homeowner, the probability of fire is low. Suppose each homeowner pays a \$500 premium to an insurance company, so the community pays a total of \$50,000, and then each is fully reimbursed if an individual member's home is destroyed. In *ex ante* terms, everyone is better off and the community has shared the risk of a fire (Kanodia, 1991).

When a risk neutral individual and a risk averse individual maximize their joint expected utility over a risky gamble, it is optimal for the risk neutral individual to bear all the risk, fully insuring the risk averse individual (the risk averse individual receives the same payoff regardless of the outcome). In the principal–agent model, since the principal is risk neutral, when the agent is risk averse, optimal risk sharing calls for the principal to offer the agent a fixed wage contract. However, the agent's best response to a fixed wage contract is to shirk.

If the agent is risk neutral, then there is no cost associated with asymmetric information. The principal can achieve the same expected profit whether or not the agent's actions are observable by offering a contract that is contingent on the outcome. The optimal contract rewards the agent for the good

outcome and penalizes the agent in the bad outcome. Since agents bear the full consequences of their actions, they have the incentive to choose high effort. However, an incentive contract contingent on the outcome requires that the agents have sufficient wealth to forfeit in the case of the bad outcome. Furthermore, individuals are risk averse.

When the agent is risk averse, the optimal contract does not provide full insurance because the risk of a low payoff in the bad outcome provides a strong incentive for the risk averse agent to exert high effort. However, the use of performance incentives is costly to the principal because agents will demand a larger expected wage to compensate for the risk of a bad outcome outside their control. The principal faces a trade-off between inducing effort and providing insurance to agents.

Limited Liability

In the optimal contract, the principal uses punishment in the event of a bad outcome as an incentive for the agent to exert high effort. However, the agent's liability for the bad outcome may be limited by the agent's assets, or by legal institutions like bankruptcy. When the principal's optimal contract is constrained by the agent's limited liability, incentives that induce high effort are both more costly and less effective. In contrast, as the agent's liability increases, the agent is more likely to exert high effort.

Monitoring and Signals

The principal can hire supervisors to monitor the agent's behavior in order to generate additional information on the agent's behavior. Alternatively, the principal may be able to use publicly available information to aid in evaluating the agent's performance. The principal can use the informative signal to design a contract with more efficient risk sharing (more insurance for the agent). Consider the steel industry, where firms' profits are sensitive to the prices of raw materials such as iron and coal. If the owners or shareholders pay the executives based on the firm's absolute profits, they would be penalized for events beyond their control such as a spike in the price of coal. Instead, the owners could insure their executives against negative pay shocks due to unfavorable industry conditions by making executive compensation contingent on their firm's performance relative to the rest of the industry (information on the steel industry performance is publicly available).

Reputation

The agent's incentives change dramatically in a multiperiod game, where the principal can form expectations about the agent's future performance based on the agent's past actions. The agent will want to 'make a good impression', in order to receive favorable contracts in the future. In the quality-assuring price model, the principal is willing to pay reputable agents a price that includes a premium above their costs of delivering high quality. In addition, agents have an incentive to forgo the higher profits that can be attained by shirking in the current period in order to maintain their reputation and receive higher profits in the future. (*See* **Games: Signaling**; **Bayesian Learning in Games**; **Decision-making, Intertemporal**)

FAIR WAGE–EFFORT HYPOTHESIS

Akerlof and Yellen (1990) present an alternate model of the relationship between the firm and workers where their behavior is based on social norms rather than risk sharing, or reputation as in the quality-assuring price model. The fair wage–effort hypothesis asserts that workers have a perception of whether the wage they are offered is fair. Further, workers respond to wages that they perceive to be unfair by reducing their level of effort. Conversely, if wages are above the fair wage (also called an efficiency wage), workers will reciprocate by supplying more effort (also called voluntary cooperation). The fair wage–effort hypothesis and the model of quality assuring price predict observationally equivalent behavior; the only functional distinction is that a quality assuring price depends on the worker's reputation, while the fair wage–effort hypothesis holds in absence of reputation.

EXPERIMENTAL RESULTS

The experimental studies on principal–agent games have focused on three questions. First, is moral hazard a problem? The basic assumption of the principal–agent model is that in the absence of incentives the agent will always shirk. With the exception of Fehr and Gächter (2002), all the studies reported here find that a substantial portion of agents choose to shirk when incentive mechanisms are either weak or absent.

Second, can the moral hazard problem be eliminated by the incentive mechanisms described in the principal–agent literature? There is a large theoretical literature on the design of incentive

contracts and the trade-off between risk sharing and incentives. Only three experiments examine risk sharing in the context of a moral hazard problem; the rest of the experimental studies offer the agent a fixed wage contract. The other studies, which examine the role of accounting and auditing in mitigating moral hazard, focus on the effectiveness of incentives provided by reputation, monitoring, and liability.

Third, are there other incentive mechanisms beyond the scope of the principal–agent literature which can solve the moral hazard problem? Several studies examine a larger set of incentive mechanisms that include motivators such as trust, fairness, and reciprocity. These studies prevent the formation of reputations in order to test cleanly the fair wage–effort hypothesis. One study compares incentive compatible contracts, which are the traditional solution to moral hazard, with contracts enforced by reciprocity.

Risk Sharing

Berg *et al*. (1992) compare the principal's choice of contract, and the agent's subsequent choice of action in a no moral hazard environment, to a moral hazard environment. In each environment, the principal can offer one of three contracts: a flat wage contract which is the optimal risk sharing contract, or one of two incentive compatible contracts where the agent's wage is contingent on the outcome. In the no moral hazard environment, the principal offers the flat wage contract, and the agent chooses the optimal effort. In the moral hazard environment, the principal offers an incentive contract, and the agent chooses the optimal effort. They find that for the majority of the time principals and agents choose the predicted contract–action pair for the environment, demonstrating that the principals recognize the moral hazard problem, and that the agents respond to the incentives as expected.

Risk Sharing and Signals

Berg *et al*. (1990) examine whether the option for the agent to communicate additional information allows the principal to offer a contract with improved risk sharing (where the risk averse agent bears less risk) because the principal has the option to condition the contract both on the outcome and on the agent's report. In the experiment, the agent chooses an unobservable action (high or low) and subsequently also receives private, unverifiable information about the outcome (high or low). The principal chooses one of three contracts: a communication-independent contract where the agent's compensation depends only on the outcome, a truth-communication contract where the agent's compensation depends on both the outcome and the report and includes incentives for high action and truthful reporting, or a disincentives contract that encourages low action and misrepresentation of the signal. They compare the principal's choices in two different environments; first the principal chooses between the communication-independent contract and the truthful reporting contract and, second, the principal chooses between the communication-independent contract and the disincentives contract. They find that subjects respond to incentives for truthful reporting or for misrepresentation when the parameters provide large rewards for optimal decisions. In contrast, when the parameters offered smaller rewards, the principal's choice of contracts was statistically indistinguishable from random.

Reputation

DeJong *et al*. (1985a) examine whether moral hazard is a problem when the principal and the agent can develop individual reputations. They design a laboratory environment where the principal faces an uncertain payoff, and the principal can reduce the probability of a loss by purchasing services from an agent. The agents submit sealed offers to each principal, stating a price and level of service. The principals then choose among the offers. If an agent's offer is accepted, then the agent chooses the level of service to deliver, and both parties are informed about the outcome. Hence, the principal cannot directly observe the quality of the service provided by the agent and so cannot distinguish whether the loss was caused by bad luck or shirking. This design does not match all of the characteristics of the principal–agent model, because there are multiple principals and multiple agents, the agent moves first, and the agent offers a contract. But, once the principal chooses a contract, the agent's choice of effort has all the characteristics of a moral hazard problem. In each of the four markets, they find that agents tend to deliver a lower level of service than specified in their offer. However, there are fewer substandard deliveries than predicted by the principal–agent model. They find evidence that the lower frequency of substandard deliveries can be explained by reputation and quality-assuring price models.

Auditing and Monitoring

DeJong *et al*. (1985b) investigate whether liability rules and the possibility of costly investigation by the principal, and whether public disclosure of investigation findings, can mitigate moral hazard problems, using the same design as DeJong *et al*. (1985a) described earlier. In the costly investigation treatment, in the event of a loss, the principal has the option to pay to determine the level of service delivered, and the results are publicly disclosed. In the costly investigation and the negligence liability rule treatments, with either public or private disclosure of the findings, if the principal finds that the agent delivered a level of service below the 'due care' standard, then the agent bears the loss. They find that costly investigation partially deters substandard deliveries as long as the principal continues to investigate. When costly investigation is combined with the negligence liability rule, substandard deliveries are almost completely eliminated. They also find that public disclosure is a more effective deterrent to substandard deliveries than private disclosure, suggesting that agents perceive that there are incentives for them to develop a reputation for supplying high quality. In addition, they find some evidence to support an explanation of quality-assuring price, where the principals pay a premium to agents who have a reputation for delivering higher levels of service.

Mayhew (2001) examines reputation building by auditors, and the incentives provided by managers' decisions to hire auditors. In the experiment, there are two layers of moral hazard: both the manager and the auditor send reports to the investors about the quality of the asset. First, the manager chooses whether to invest in a higher quality product. If the manager invests, the product will be higher quality with probability 0.55. Prior to learning the investment outcome, the manager chooses which, if any, auditor to hire. Then the manager learns the investment outcome and makes a report to the investors. Once hired, the auditor chooses whether or not to engage in costly investigation without knowing the manager's investment decision. If the auditor investigates, he learns the true outcome and makes a truthful report. Otherwise, the auditor agrees with the manager's report. The investors receive both reports, and bid for the outcome in a first-price sealed-bid auction. All participants observe the winning bid, the manager's and auditor's reports, the auditor's identity, and the actual outcome. Mayhew finds that when investors reward managers for hiring reputable auditors, the managers respond by hiring reputable auditors, and auditors form reputations for high-quality audits. However, auditors do not form reputations in those sessions where the rewards to reputation do not manifest early in the session. Mayhew's findings support King's (1996) conjecture that the formation of reputations for truthful reporting is path dependent, and more likely to form when honest reporting is less costly early in the session.

Nagin *et al*. (2002) conducted a field experiment examining the role of monitoring on truthful reporting at a telephone solicitation company. The employees are paid a modified piece rate, where they receive a bonus for each reported successful solicitation. Some of their reported successful calls are monitored by 'call backs' and at the end of each week the employees are informed of the number of 'bad' calls, where the donor claims they did not make a pledge, which are deducted from their incentive pay. Since some donors renege on their pledges, bad calls are a noisy indicator of false reporting. At four of the company's 16 sites, the firm audited 25 percent of the successful calls, but they manipulated the audit rate observed by employees, so it appeared between 0 to 10 percent. In contrast to the theoretical prediction that all employees will shirk more when monitoring declines, they find that a substantial portion of the employees did not respond to manipulations in the monitoring rate. Consistent with predictions that employees are motivated by trust and reciprocity, they find that the employees who exploit the reduction in the monitoring rate also perceive the employer to be uncaring and unfair.

Reciprocity and Fairness

Fehr *et al*. (1997) examine whether reciprocity motives can act as a contract enforcement device when the firms offer a fixed wage contract. They compare worker and firm behavior in a no-reciprocity-treatment (NRT) where the firms post contracts and the workers' effort is exogenously determined by the experimenter, a weak-reciprocity-treatment (WRT) where the firm can fine the worker if it can verify that the worker shirked (they can verify the worker's choice 50 percent of the time), and, finally, a strong-reciprocity treatment (SRT) where, after the firm observes the worker's choice of effort, the firm can, for a cost, penalize or reward the worker. They prevent the formation of reputations by making the trading partners anonymous, so that any reciprocation is due to the trading partners' actions in the current period, rather than in previous periods. Comparing WRT to NRT, they find that firms are more

generous to the workers when the workers have the option to reciprocate, and the workers respond by shirking less frequently. Comparing SRT to WRT, their results suggest that if both parties in a transaction have the opportunity to reciprocate, then reciprocity motivations have a strong impact on the enforcement of contracts. Furthermore, the gains from trade are larger in SRT than WRT.

Fehr and Gächter (2002) find that incentive contracts which punish shirking may crowd out reciprocity-driven voluntary cooperation where the agent exerts excess effort. They compare a trust treatment where the principal offers a fixed wage and states a desired level of effort and an incentive treatment where the principal has the additional possibility of fining the agent if the principal can verify shirking (they can verify one-third of the time). If shirking cannot be verified, then the principal is committed to pay the fixed wage. Again, the principals and agents are not identified, in order to prevent the formation of individual reputations. They find that principals in the trust treatment offer higher wages and higher rents, and demand higher effort levels than in the incentive treatment. Principals in the incentive treatment frequently choose the maximum fine, suggesting that they are relying on the threat of punishment to induce agents to supply the desired effort. In the trust treatments, they find that agents cooperate voluntarily, on average supplying more effort than the principal's stated desired level when they are offered a larger rent. In the incentive treatments, they find almost no voluntary cooperation and, further, incentive compatible fines did not prevent shirking.

Rigdon (2002) examines how workers respond to fixed wage contracts, comparing the principal–agent model prediction of shirking and the fair wage–effort hypothesis prediction of voluntary cooperation in response to a generous contract. The experiment uses a variation of the Fehr and Gächter trust treatment with two significant changes: first, the experiment is double blind, eliminating all experimenter–subject interaction and, second, the agents face substantial costs associated with supplying high effort. In contrast to Fehr and Gächter, Rigdon finds that the workers chose to shirk 79 percent of the time, fulfilled the contract 18 percent of the time, and only cooperated voluntarily 3 percent of the time. In addition, though employers initially offered the workers generous contracts demonstrating trust when the workers did not reciprocate, the contracts converged to the competitive equilibrium of compensating workers for the lowest effort level. Rigdon postulates that the contradicting results can be explained by the elimination of interaction between the subject and experimenter; the experimental economics literature has demonstrated that subjects are more self-interested when their actions are anonymous. (*See* **Economics, Experimental Methods in**)

Incentive Compatibility versus Reciprocity

Guth *et al*. (1998) pair subjects in a principal–agent relationship and examine whether trust and reciprocity motivations are as effective at inducing high effort as incentive compatibility. In their experiment, the principal designs the contracts, specifying the composition of a fixed wage and a share of the dividend payment. The rational response to a fixed wage is for the agent to shirk, so that a high effort response is explained by reciprocity motives. In this setting, an incentive compatible contract would offer a relatively large profit share and no fixed wage. They find that principals initially offer contracts with large fixed wages and profit shares that are too small to be incentive compatible. The principals can offer up to three contracts, and most choose to offer at least two. They observe that the agents' contract choice appears to signal their intended work effort. In addition, the agents' work effort is positively influenced by more generous contracts, either higher profit shares or higher fixed wages. However, high effort in response to a generous fixed wage contract appeared in the initial period, while generous profit sharing induced high effort in later periods. These results suggest that incentive compatibility needs to be learned.

CONCLUSIONS

As predicted by the principal–agent model, and confirmed by experimental evidence, moral hazard is a problem. Berg *et al*. (1990, 1992) find that when faced with a moral hazard problem, the principal chooses a risk sharing contract, and the agents respond to the incentives as predicted. DeJong *et al*. (1985a) find that reputation may reduce the incidence of shirking but certainly does not eliminate moral hazard. DeJong *et al*. (1985b) and Nagin *et al*. confirm that monitoring and agent liability (ex post risk sharing) are effective in alleviating the moral hazard problem both in the laboratory and in the field. Mayhew finds that agents choose to hire reputable monitoring services when the principal is willing to pay more for assets verified by a reputable auditor. In contrast, Fehr and Gächter find that

increased monitoring and liability may reduce efficiency by destroying reciprocity. Fehr *et al*. and Guth *et al*. find that reciprocity motives reduce shirking, and even motivate the agent to deliver excess effort. Rigdon questions the results regarding the efficacy of reciprocity and comments that experimental procedures matter to the level of reciprocity across a wide variety of settings, including issues with moral hazard.

While principal–agent theory focuses on the role of risk sharing as the most effective incentive compatibility mechanism, most of the experiments have ignored risk sharing. Instead the majority of the experiments, including the experiments testing the effectiveness of reciprocity motives, offered the agent a fixed wage. In a risk sharing contract, depending on the outcome, an agent could be punished for high effort or rewarded for low effort. In contrast, the reciprocity experiments prevent the principal from punishing the agent when the agent has taken the correct action.

References

Akerlof GA and Yellen JL (1990) The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics* **105**: 255–283.

Berg J, Daley L, Dickhaut JW and O'Brien J (1992) Moral hazard and risk sharing: experimental evidence. In: Isaac MR (ed.) *Research in Experimental Economics*, vol. 5, pp. 1–34. Greenwich, CT: JAI Press.

Berg J, Daley L, Gigler F and Kanodia C (1990) The value of communication in agency contracts: theory and experimental evidence. *The Canadian Certified General Accountants' Research Foundation*. Research monograph number 16.

DeJong DV, Forsythe R and Lundholm RJ (1985a) Ripoffs, lemons, and reputation formation in agency relationships: a laboratory market study. *The Journal of Finance* **40**: 809–820.

DeJong DV, Forsythe R, Lundholm RJ and Uecker WC (1985b) A laboratory investigation of the moral hazard problem in an agency relationship. *Journal of Accounting Research* **23**: 81–120.

Fehr E and Gächter S (2002) Do incentive contracts crowd out voluntary cooperation? Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 34.

Fehr E, Gächter S and Kirchsteiger G (1997) Reciprocity as a contract enforcement device: experimental evidence. *Econometrica* **65**: 833–860.

Guth W, Klose W, Konigstein M and Schwalbach J (1998) Experimental study of a dynamic principal–agent relationship. *Managerial and Decision Economics* **19**: 327–341.

Kanodia C (1991) Lecture notes. Carlson School of Management, University of Minnesota.

King RR (1996) Reputation formation for reliable reporting: an experimental investigation. *The Accounting Review* **71**: 375–395.

Mayhew BW (2001) Auditor reputation building. *Journal of Accounting Research* **39**: 599–617.

Nagin D, Rebitzer J, Sanders S and Taylor L (2002) Monitoring, motivation and management: the determinants of opportunistic behavior in a field experiment. *NBER Working Paper Series*, 8811. [<http://www.nber.org/papers/w8811>]

Rigdon ML (2002) Efficiency wages in an experimental labor market. Interdisciplinary Center for Economic Science at George Mason University, working paper.

Further Reading

Davis DD and Holt CA (1993) *Experimental Economics*. Princeton: Princeton University Press. [Especially chapter 7.]

Fehr E and Camerer C (in press) Measuring social norms and preferences using experimental games: a guide for social scientists. In: Henrich J, Boyd R, Bowles S, Gintis H, Fehr E and McElreath R (eds) *Cooperation and Punishment in Simple Societies*.

Kreps DM (1990) *A Course in Microeconomic Theory*. Princeton: Princeton University Press.

Laffont JJ and Martimort D (2002) *The Theory of Incentives: The Principal–Agent Model*. Princeton: Princeton University Press.

Salanie B (1998) *The Economics of Contracts: A Primer*. Cambridge, MA: The MIT Press.

Watts RL and Zimmerman JL (1986) *Positive Accounting Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Games: Sequential Bargaining

Intermediate article

Glenn W Harrison, University of South Carolina, Columbia, South Carolina, USA
E Elisabet Rutström, University of South Carolina, Columbia, South Carolina, USA

CONTENTS

Sequential bargaining games involve one player responding to the offers from another player. They provide a structured opportunity to test game-theoretic hypotheses about subject motivation and cognitive processes.

INTRODUCTION

Sequential bargaining games are particularly interesting objects to analyze since they offer economists a rich context in which to study the way in which humans form expectations about the behavior of others. Traditional game theory provides a crisp series of predictions if one accepts that subjects have common expectations about the rationality of others, as well as about the game being played. Observed behavior in controlled experiments often deviates from those crisp predictions.

Three general hypotheses compete to explain these discrepancies between observed behavior and the prediction of theory. One popular view is that subjects have some desire for fairness with respect to the pay-offs to their opponents, and avoid the extreme outcomes predicted by game theory for this reason. Another popular view is that subjects do not know how to apply the idea of backward induction, or that their application of the idea of backward induction is flawed. A third view is that the subjects simply do not perceive the game in the way the experimenter does.

This article considers the nature of the process of formation of expectations found in these games. This process provides one cognitive window on the second and third explanations for observed behavior, without ruling out the first. Sequential bargaining games require subjects to undertake 'backward induction thought experiments' about the behavior of others, in which they look to the end of the game and figure out what is going to happen, then use that information to figure out what is going to happen in the penultimate stage, and so on. The outcomes of these thought experiments appear to be very sensitive to the context of the experimental design. This suggests that the cognitive process by which people apply backward induction is not 'hard-wired' in an unconditional manner, and that expectations depend on numerous variables in the bargaining environment.

NEGOTIATING A DISTRIBUTIVE OUTCOME

Bargaining occurs more often than many would expect. Whenever a contract is evaluated, an implicit bargaining session has been concluded. This is apparent when the contract price is a subject for negotiation, but even 'take it or leave it' settings entail bargaining in the broader sense in which the term is used by economists.

There are two popular formal approaches to the bargaining task. The first is to characterize the outcome of the bargaining process, remaining agnostic as to the bargaining process itself. A series of axioms are typically proposed as characterizing the outcome of a bargaining process, and the outcomes implied by those axioms derived (Davis and Holt, 1993; Kagel and Roth, 1995). The second approach is to write out the precise rules of the bargaining process, and characterize the game-theoretic equilibria of such games. (The first approach was introduced in Nash (1950), and the second in Nash (1953). See Binmore (1992) for a careful exposition of the relationship between the two.)

Sequential bargaining models provide a rich framework for the second approach, which depends on the expectation formation process. Although there is an infinite range of possible bargaining games that could be specified, extensive study has been generally limited to two broad

classes: 'alternating offer' and 'ultimatum' bargaining games.

SEQUENTIAL BARGAINING GAMES DESCRIBED

Alternating offer and ultimatum bargaining games are strategically similar. In an ultimatum bargaining game, one player makes an offer which the other player either accepts or rejects. In an alternating offer bargaining game, one player makes an offer and the other player then reacts to it by accepting it or making a counter-offer. The ultimatum bargaining game is a proper sub-game of the alternating offer bargaining game that occurs over a finite horizon. Most of the issues that arise in the more complex alternating offer bargaining game can be most directly studied through the ultimatum bargaining game. The latter is also worthy of study in itself since it arises in many other economic settings. (*See* **Games: Ultimatum**)

The Ultimatum Bargaining Game

The ultimatum game involves two agents in a seemingly transparent bargaining setting. One player, 'sender', decides on a proposed division of a pie. The other player, 'receiver', then decides whether to accept the proposal. If the proposal is accepted, the players receive pay-offs according to the proposal. If it is rejected, they each receive some disagreement outcome, typically zero.

Backward induction helps identify the sub-game perfect Nash equilibrium prediction. (For an introduction to game-theoretic terminology and concepts, see Fudenberg and Tirole (1991) or Binmore (1992).) If the receiver is not satiated in pay-offs, then any amount offered by the sender should be accepted providing it exceeds the pay-off the receiver would get in the disagreement outcome. Expecting this behavior in the sub-game, the sender can then ask for virtually all of the pie in the first stage. Backward induction is the thought process by which the sender uses the rules of the bargaining game, as well as his beliefs about the likely actions that the responder will then take, to identify his best strategy.

There have been many experimental studies of ultimatum bargaining: see Güth *et al*. (1982) for the earliest study, and Güth and Tietz (1990) for a review. The interesting behavioral outcome is that the sender does not ask for virtually all of the pie. Instead, we tend to observe average offers of just less than 50% to the receiver, with a common mode being precisely 50%. This result does not appear to be affected by standard treatment variations, such as the level of experience in the game, the level of pay-offs, or repetitions.

Figures 1 to 4 illustrate the main results from four series of ultimatum bargaining games. In each case the data have been normalized in terms of the percentage offer to the receiver. The graphs gloss over several treatments which do not appear to be crucial to the basic behavioral outcome.

Figure 1 shows the data from the experiments of Güth *et al*. (1982). It shows the data from the first series of 'inexperienced' play as well as the data from 'experienced' play with the same subject pool. The distributions of offers are similar, with means of 35% and 31% respectively. There is an apparent tendency to offer 50% or slightly less, with 23 of the 42 pooled observations being centered around 40% or 50%. Experience, in the sense of being able to think about the game for a week, does not make a great difference. Although there is a shift in the distribution towards the equilibrium with a new mode emerging at offers of 20%, the distribution is still skewed much further to the right than predicted by game theory.

Figure 2 displays the results of Prasnikar and Roth (1992), which allow us to observe the effect of repeated play against shuffled opponents. The individual period data of the experiment are pooled into two groups: rounds 1 to 5, and rounds 6 to 10. The distribution is again skewed well to the right of the prediction of game theory, with most of the offers again being over 35%. In this case experience seems to slightly worsen the results from the perspective of game theory, with the mean offer increasing from 39% over the first five rounds to 44% over the last five rounds.

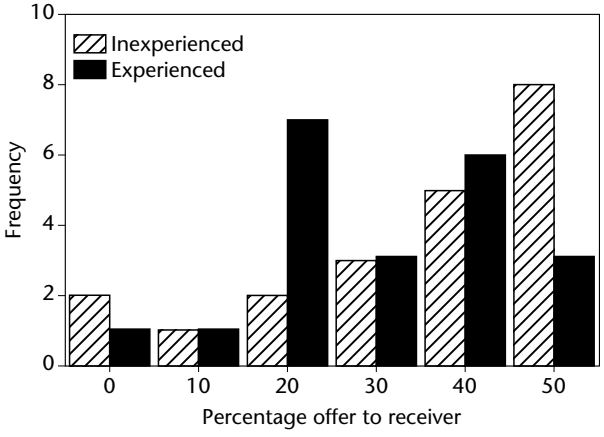


Figure 1. Ultimatum bargaining games: results from Güth *et al*. (1982).

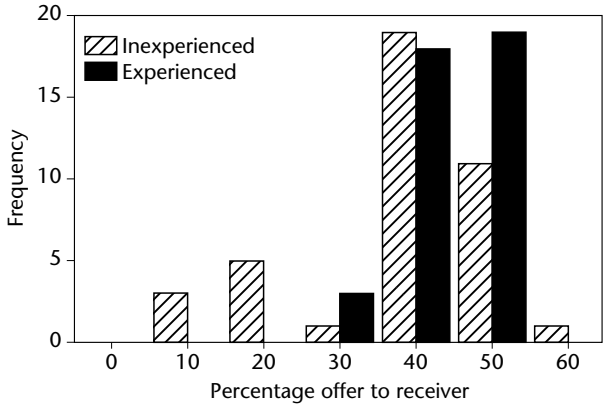


Figure 2. Ultimatum bargaining games: results from Prasnikar and Roth (1992). 'Inexperienced' data are from rounds 1–5, and 'experienced' data are from rounds 6–10.

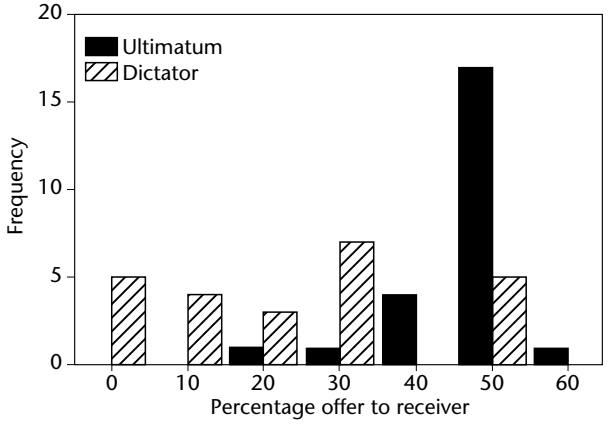


Figure 3. Ultimatum and dictator bargaining games: results from Forsythe *et al*. (1994). $10 games only are shown.

Figure 3 displays the results of Forsythe *et al*. (1994). Their ultimatum bargaining game results are comparable to those discussed already. But in order to study the role of the senders' expectations of responders' behavior, they also implemented a truncated version of the ultimatum game, known as the dictator game. In this game, the sender simply writes down his or her demand, with the receiver getting the residual. Unlike the ultimatum bargaining game, the dictator game does not require the receiver to agree or disagree.

Why conduct such a trivial game as the dictator game? The purpose was to see whether the results from ultimatum bargaining games can be attributed to senders having a 'taste for fairness', in the sense of a bias towards making a 50% offer. If they do have such a taste, then one should observe similar distributions of offers in the dictator and ultimatum games. If, however, the results of ultimatum games primarily reflect an expectation, on behalf of senders, that receivers will reject low offers, we would predict a difference in the distribution of offers between ultimatum and dictator games.

The results, displayed in Figure 3, strongly confirm the latter 'expectational' interpretation of the ultimatum game outcomes. There is little doubt that the distribution of offers in the dictator game is skewed much further to the left than in the ultimatum game. Thus one can conclude that at least some subjects in the ultimatum game are driven to offer large amounts to the receiver because they expect that the receiver will reject small amounts.

Other evidence of the role of expectations can be found in Hoffman *et al*. (1994), where offers are found to be significantly lower when market terminology is used instead of the terminology of splitting a pie. The terminology used may have an effect on expectations by encouraging a focus on common past experiences. When subjects had to earn the right to be the first mover through their performance in a task before the game, offers were found to be even lower. Modal offers were reduced from 50% to 30% of the $10 pie, and the proportion of equal-split offers dropped to 10%.

The Alternating Offer Bargaining Game

The alternating offer bargaining game is conducted over a sequence of periods, in each of which there is a proposer and a responder. If the responder in the first period rejects the offer, the game continues into the second period and the players switch roles. Players normally negotiate over the division of a pool of money, and the value of this pool gets discounted for each period that the game is extended.

Binmore *et al*. (1985) conducted a two-period alternating offer bargaining experiment where the money pool dropped from 100 pence in period one to 25 pence in period two. The sub-game perfect Nash equilibrium prediction is for the first mover to demand 75 pence in the first round, and for the second mover to accept the 25 pence offered. Binmore *et al*. found that, with some experience, subject behavior was consistent with the theory. Subsequent experiments (Neelin *et al*., 1988; Spiegel *et al*., 1994; Harrison and McCabe, 1992; Güth and Tietz, 1988; Ochs and Roth, 1989) have shown that behavior is fully consistent with theory only for some parameters; otherwise deviations from theory predictions are in the direction of equal splits.

Attempts have been made to test the explanatory power of both the fairness hypothesis and the expectations hypothesis. According to the fairness hypothesis, the reactions of subjects reflect a preference for fairness. According to the expectations hypothesis, senders react to responder behavior off equilibrium, and sender behavior is determined primarily by strategic considerations. Binmore *et al*. (1991) and Prasnikar and Roth (1992) find that strategic considerations are important, although they cannot rule out the existence of fairness preferences.

Prasnikar and Roth (1992) studied the 'best shot' game of public goods provision introduced by Harrison and Hirshleifer (1989). In this game the first mover makes an offer, after which the second mover makes an offer, and the pay-offs to each player are determined by the maximum of the two offers (multiplied by some redemption value) minus that player's own offer (multiplied by some cost parameter). The sub-game perfect Nash equilibrium prediction in this game is for the first mover to offer zero. In repetitions over 10 periods, Prasnikar and Roth found that by period 7 all first movers were making the equilibrium offers. They conclude that it is the off-equilibrium experiences, where first-mover offers above zero are countered by offers of zero by the second mover, which drive the adjustment of behavior in these games. Off-equilibrium offers by first movers clearly lead to lower earnings than equilibrium offers, contrary to findings in ultimatum games, where responders' propensity to reject declines with the amount offered so that proposers' earnings often increase with the amount offered. These observed behavioral differences are therefore consistent with the expectations hypothesis.

FORMING COMMON EXPECTATIONS

One way to appreciate the importance of common expectations for the interpretation of these results is by undertaking a simple arithmetic thought experiment with the observed experimental data. Pool all of the data from Figures 1, 2, and 3, generating a discrete distribution of 3, 5, 19, 12, 62, 82, and 6 observations of offers to the receiver of 0%, 10%, 20%, 30%, 40%, 50%, and 60%, respectively. The corresponding distribution of rejections is 2, 3, 9, 4, 16, 2, and 0, resulting in a distribution of the probability of an offer being accepted of 0.33, 0.40, 0.53, 0.67, 0.74, 0.98, and 1.00.

Now imagine that you were asked to play the ultimatum game knowing this history of acceptances. Also imagine that you have no feelings of altruism towards your opponent whatsoever. What is your best strategy as sender? Assuming local risk neutrality, your *expected* percentage of pie from each offer is 33, 36, 42, 47, 44, 49, and 40, respectively. Thus your best strategy (by a small margin) is to ask for only 50% of the pie, just as if you were playing fairly.

One problem with this numerical exercise is that the data are drawn from several experiments which differ in various ways, and this could contaminate the results. Figure 4 presents the data from Carter and Irons (1991). Their study matches our conceptual experiment much better. They asked each subject to submit strategies for both sender and receiver. The roles they played were determined after all strategies were submitted. The method used to allocate each subject to a role was to randomly pair subjects and then to let the winner of a word game be sender.

This is a particularly useful experiment for the purposes of our thought experiment. Firstly, we have data on the distributions of unconditional strategies over the whole subject pool. Secondly, the data are drawn from the same subject pool, supporting our interpretation of them as reflecting the distribution of expectations of the subject pool.

The experiment was not repeated; so subjects had no way to update their expectations. We assess the importance of this issue later in relation to the experimental design of Harrison and McCabe (1996), which addressed it directly.

Applying the same arithmetic as used earlier, the expected pay-off (percentage of pie) to a sender of demanding 100%, 90%, 80%, 70%, 60%, and 50% is 15, 36, 40, 42, 50, or 50, respectively, in this experiment. Again, assuming that the distribution of rejections is invariant, a sender seeking to maximize

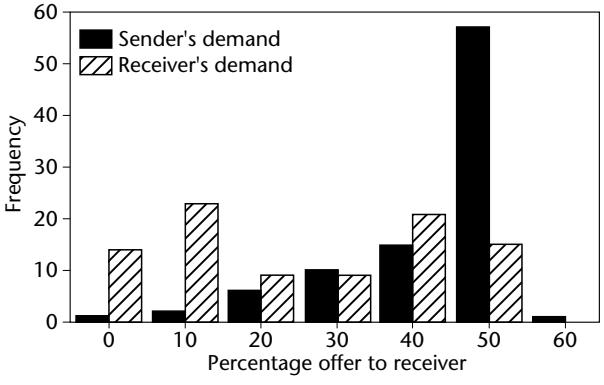


Figure 4. Ultimatum bargaining games: results from Carter and Irons (1991). The data are pooled over economists and non-economists and over first-year and fourth-year students.

expected pay-off would be led to ask for only 50%, just as if he or she had been motivated by fairness considerations.

The only conclusion one can draw from these simple numerical exercises is that one cannot reject out of hand the notion that subjects are bringing expectations into the experiment that are consistent with the distribution that is actually observed. To the extent that mere repetition of the game does not change behavior (and Prasnikar and Roth, 1992, have shown that it does not do so drastically), their posterior distributions during the experiment will continue to reflect their initial priors. They might then have no sense of playing equitably, but still look as if they do. On the other hand, without some control over subject expectations we cannot test the fairness and expectations hypotheses against each other in these experiments.

The critical experiment, then, is one that can control those expectations during the experiment and try to move them in one direction or another. In one sense, this is what was done with the dictator games, where it was found that expectations of rejections explained some, but not all, of sender behavior. Nevertheless, since receivers have no power at all over the outcome in dictator games, it is likely that other factors may cause the observed off-equilibrium sending behavior. Below we review two experimental games which were intended to shift players' expectations, but with the two players' relative positions of power unaffected. The natural direction in which to move expectations, given the priors that are reflected in the behavior exhibited in Figures 1 to 4, is towards the game-theoretic prediction. In this way one can try to generate data that test the fairness and expectations hypotheses against each other.

THE ROLE OF SUB-GAME EXPERIENCE

In the standard extensive form of the ultimatum bargaining game, the receiver gets to know what proposal the sender has chosen before selecting his strategy. A simple modification of this extensive form has the receiver selecting his strategy without knowing what proposal the sender has made. In this case the receiver would simply be asked at what threshold offer he would switch from rejecting to accepting. It is apparent that the equilibrium predictions of these two games are identical.

By eliciting strategies from the receiver, rather than just choices, subjects have an incentive to contemplate strategies that apply both on and off any equilibrium path. Considerations of appropriate actions off the equilibrium path are necessary for the application of backward induction. Essentially the same logic underlies the design in Harrison and McCabe (1992), in which subjects were given experience in the sub-games of an alternating offer bargaining game. Harrison and McCabe put subjects in three-period games in every even round, and in the final two-period sub-game in every odd round. Knowledge of the outcomes of these two-period sub-games was hypothesized to be needed for subjects to properly evaluate strategies for the full game, at least according to game theory, even though the equilibrium of the full game called for the sub-game to never be played. (This behavioral outcome is well known from experimental asset markets, and is referred to as the 'swingback hypothesis' by Forsythe *et al*. (1982) and Friedman *et al*. (1984).) As subjects gained knowledge of the sub-game outcomes in the experiment they indeed successfully managed to apply backward induction and select strategies for the full game consistent with game theory.

We can extend the idea of eliciting complete strategies one step further, and modify the extensive form ultimatum game again so that each subject does not know whether he will be sender or receiver in any given round. This determination is known by the subjects to be made by nature at random, with an equal chance of each subject being selected to be one or the other type. Again, it is apparent that the equilibrium strategies of this game are identical to those of the standard ultimatum bargaining game.

Harrison and McCabe (1996) undertook several experiments with this basic design. Their first experiment, called U1, had subjects submitting their strategies for 15 periods. No information on other players' strategy choices in any period was provided until the end of the session. At that point, each subject was told of the period that had been selected at random for payment purposes. Their second experiment, called U2, was identical to U1 except that subjects were given information on the strategy choices of all subjects after every period. The complete history of these period-specific strategy distributions was also displayed on a blackboard at the front of the room.

Their third experiment, called U3, was the same as U2 except that the human subjects were joined in the first five periods by an equal number of simulated opponents. A simulated player would ask for a percentage of the pie between 86% and 99% when playing sender, and between 1% and 14% when playing receiver. The random distribution of

simulated strategies was uniform within these intervals. The simulated strategies were included in the reported distribution of strategies during the 10 periods in which there were simulated players. For all subsequent periods the reported distribution included only the strategies of the human players, since no simulated players were present. This device for controlling the expectations of subjects is due to Roth and Schoumaker (1983).

Figures 5 and 6 present the results from the experiments of Harrison and McCabe (1996). The average percentage that human senders allocated to themselves is shown in Figure 5. The average percentage that human receivers requested as a minimum is shown in Figure 6.

The results are consistent with the hypothesis stated earlier: that allowing subjects to form common expectations will act as a behavioral surrogate for the common knowledge assumed in the game-theoretic prediction. The control experiment U1, in which there was no public information, produced behavior much like that observed in previous experiments, with the sender asking for slightly more than 50% of the pie. On the other hand, the minimum demands of the receiver in experiment U2 decay in the direction of the game-theoretic prediction. This indicates that expectations are affected by the provision of public information. However, in early periods they are also very close to 50% of the pie. Demands by the sender in experiment U2 are similar to those observed in the control experiment for the first five or so periods, but from period 9 onwards they diverge markedly, moving towards the game-theoretic prediction. By the end of the experiment, average demands by the sender were about 87% of the pie, compared with 56% in the control.

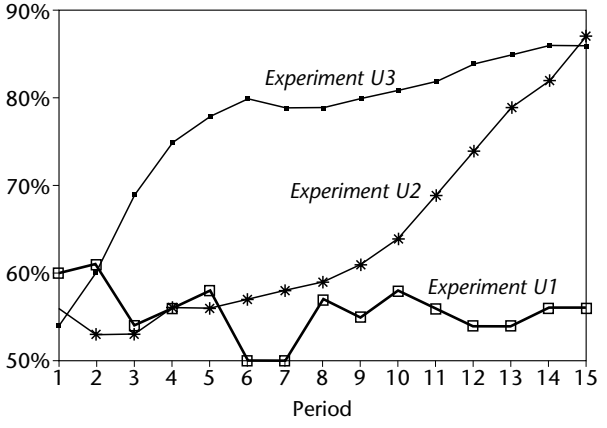


Figure 5. Demands by senders in the Harrison and McCabe (1996) experiments. The graphs show the average percentage that senders allocate to themselves.

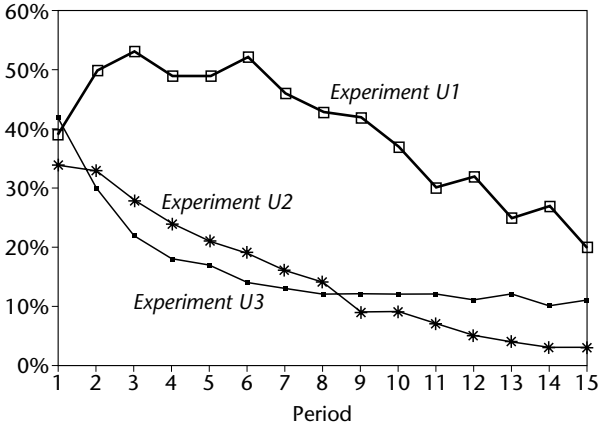


Figure 6. Demands by receivers in the Harrison and McCabe (1996) experiments. The graphs show the average minimum percentage required by receivers.

THEORETICAL EXPLANATIONS

What appears to be driving this behavior is the relatively rapid decline in minimum demands by the receiver in experiment U2. As early as period 2 they are well below those observed in the control, and they steadily decline to around 24% by period 4. Since all subjects are aware of this average requirement by the receiver when they come to enter their period 5 strategies, it is not surprising that subjects re-evaluate the wisdom of sticking with a period 4 demand of 56% on average when they need to give the receiver only 24% on average. A typical risk-neutral subject could increase his sender demand to 76%, assuming that receivers would again ask for only 24%. (A more complete calculation would account for the distribution of offers around the average of 24%.) Moreover, the steady decline in receiver demands is obvious (from 34% to 33% to 28% to 24% over periods 1 to 4), so assuming that receivers would continue with a 24% demand (or less) in period 5 looks like an even safer bet for our risk-neutral subject.

If sender behavior is primarily driven by expectations in this sense, we would expect to see changes in amounts offered as well. This is indeed what happened in experiment U3. Demands by the sender were much higher than observed in the appropriate control (experiment U2), at least for all periods except the first few and the last few. Expectation formation appears to be driven in periods 1 to 5 by the reported behavior of the simulated players, as indicated by the rapid increase in demands by senders. What is particularly

interesting about the results of U3 is the lack of a significant and persistent decline in demands by the sender once this treatment is removed after period 5. Although demands do not continue their rapid increase, they do remain quite stable and settle at an average demand in period 15 of 86% of the pie. Despite the differences in the demands of the sender in experiment U3, the demands of the receiver are virtually identical to those observed in the control experiment U2.

Additional evidence for the changes in receiver behavior over time can be gleaned from some ultimatum game experiments conducted in Japan, Israel, Yugoslavia, and the United States by Roth *et al*. (1991). Three sessions were conducted in each country. In Israel, Japan, and Yugoslavia only one experimenter was used for all sessions. Each of these three experimenters conducted one session in the United States. Thus it is possible to identify experimenter effects and country effects. No data were collected on the ages or sexes of the participants. However, some sessions did have identified differences in the subject pool. Two differences are noted by Roth *et al*.: in one of the Israeli sessions, and in one of the Yugoslavian sessions. Each subject participated in a session lasting 10 rounds, with very few exceptions where only 9 or 8 subjects turned up. Each subject faced a random opponent in each round. Thus the data consist of balanced panels of individuals responding over each of 10 rounds.

The response data in these experiments can be studied using a simple panel logit regression model, in which individual random effects for each subject are estimated. (The data and programs needed to replicate these calculations can be obtained from Harrison (2001). Alvin Roth kindly approved the use of the data, which were supplied by Miguel Costa-Gomes.) In addition to the individual random effects, explanatory variables for the response in period t include the offer received in period t, dummy variables for the country of the experiment, dummy variables for each of the experimenters, dummy variables for the two sessions identified as having unique subject recruitment procedures, and a variable keeping track of the period. The only statistically significant variables in this simple specification are the offer level, the country effect for Israel, and the variable keeping track of the period. A 1% increase in the amount offered in the current period increases the probability of acceptance in the current period by 2.2 percentage points; the Israeli subjects accepted with a probability that was 22 percentage points higher than those in the United States; and the probability of an acceptance increased by roughly 0.7 of a percentage point every period. These acceptance probabilities are conditional on the offer received, since one would expect unconditional acceptance probabilities to increase as the amount offered increased, possibly masking the effect of experience. On the sender side we find significant effects on the current offers from lagged accepted offers, and from just having had an offer accepted, using a panel regression estimated using feasible generalized least squares. We also find that these effects differ across countries.

These results, obtained across a wide sample of experimental data, confirm that there does appear to be a dynamic path to the behavior of receivers in the ultimatum game. The question of the exact form of that dynamic, and whether it corresponds to coherent learning behavior, awaits a more sophisticated econometric characterization.

CONCLUSION

These experimental results demonstrate that much of the observed behavior in sequential bargaining is consistent with an expectations-based explanation, and that it is possible to control these expectations to some extent. Moreover, by manipulating the expectations we can alter observed bargaining outcomes in predictable ways. This experimental control enables us to test the 'fairness' and 'common expectations' hypotheses. The results of Harrison and McCabe (1996) provide a strong refutation of the hypothesis that considerations of fairness drive bargaining outcomes independently of subject expectations, and this is consistent with the conclusions drawn in Binmore *et al*. (1991) and Prasnikar and Roth (1992). This is not to deny a possible role for fairness considerations when applied to subjects who do not have common knowledge of the rationality of opponents. In such environments the predictive power of noncooperative game theory is understandably weak.

Acknowledgment

Rutström thanks the US National Science Foundation for research support under grants NSF/IIS 9817518, NSF/MRI 9871019, and NSF/POWRE 9973669.

References

Binmore K (1992) *Fun and Games*. Lexington, MA: DC Heath.

Binmore K, Shaked A and Sutton J (1985) Testing noncooperative bargaining theory: a preliminary study. *American Economic Review* **75**(5): 1178–1180.

Binmore K, Morgan P, Shaked A and Sutton J (1991) Do people exploit their bargaining power? An experimental study. *Games and Economic Behavior* **3**(3): 295–322.

Carter JR and Irons MD (1991) Are economists different, and if so, why? *Journal of Economic Perspectives* **5**: 171–177.

Davis DD and Holt CA (1993) *Experimental Economics*. Princeton, NJ: Princeton University Press.

Forsythe R, Horowitz JL, Savin NE and Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic Behavior* **6**(3): 347–369.

Forsythe R, Palfrey TR and Plott CR (1982) Asset valuation in an experimental market. *Econometrica* **50**: 537–582.

Friedman D, Harrison GW and Salmon JW (1984) The informational efficiency of experimental asset markets. *Journal of Political Economy* **92**: 349–408.

Fudenberg D and Tirole J (1991) *Game Theory*. Boston, MA: MIT Press.

Güth W, Schmittberger R and Schwarze B (1982) An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* **3**: 367–388.

Güth W and Tietz R (1988) Ultimatum bargaining for a shrinking cake: an experimental analysis. In: Tietz R, Albers W and Selten R (eds) *Bounded Rational Behavior in Experimental Games and Markets*. Berlin, Germany: Springer.

Güth W and Tietz R (1990) Ultimatum bargaining behavior: a survey and comparison of experimental results. *Journal of Economic Psychology* **11**: 417–449.

Harrison GW (2001) http://dmsweb.moore.sc.edu/glenn/rpoz.zip. [Data from Roth *et al*. (1991).]

Harrison GW and Hirshleifer J (1989) An experimental evaluation of weakest-link/best-shot models of public goods. *Journal of Political Economy* **97**: 201–225.

Harrison GW and McCabe KA (1992) Testing noncooperative bargaining theory in experiments. In: Isaac RM (ed.) *Research in Experimental Economics*, vol. V. Greenwich, CT: JAI Press.

Harrison GW and McCabe KA (1996) Expectations and fairness in a simple bargaining experiment. *International Journal of Game Theory* **25**(3): 303–327.

Hoffman E, McCabe K, Shachat K and Smith VL (1994) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* **7**: 346–380.

Kagel JH and Roth AE (eds) (1995) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Nash JF (1950) The bargaining problem. *Econometrica* **18**: 155–162.

Nash JF (1953) Two-person cooperative games. *Econometrica* **21**: 128–140.

Neelin J, Sonnenschein H and Spiegel M (1988) A further test of noncooperative bargaining theory. *American Economic Review* **78**(4): 824–836.

Ochs J and Roth AE (1989) An experimental study of sequential bargaining. *American Economic Review* **79**(3): 355–384.

Prasnikar V and Roth AE (1992) Considerations of fairness and strategy: experimental data from sequential games. *Quarterly Journal of Economics* **107**(3): 865–888.

Roth AE, Prasnikar V, Okuno-Fujiware M and Zamir S (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* **81**: 1068–1095.

Roth AE and Schoumaker F (1983) Expectations and reputations in bargaining: an experimental study. *American Economic Review* **73**(3): 362–372.

Spiegel M, Currie JSH, Sonnenschein H and Sen A (1994) Understanding when agents are fairmen or gamesmen. *Games and Economic Behavior* **7**(1): 104–115.

Further Reading

Binmore K, Shaked A and Sutton J (1989) An outside option experiment. *Quarterly Journal of Economics* **104**(1): 753–770.

Cameron LA (1999) Raising the stakes in the ultimatum game: experimental evidence from Indonesia. *Economic Inquiry* **37**(1): 47–59.

Forsythe R, Kennan J and Sopher B (1991) An experimental analysis of bargaining and strikes with one sided private information. *American Economic Review* **81**(1): 253–278.

Harrison GW and McKee M (1985) Experimental evaluation of the coase theorem. *Journal of Law and Economics* **28**(3): 653–670.

List J and Cherry T (2000) Learning to accept in ultimatum games: evidence from an experimental design that generates low offers. *Experimental Economics* **3**(1): 11–29.

Schotter A, Snyder B and Zheng W (2000) Bargaining through agents: an experimental study of delegation and commitment. *Games and Economic Behavior* **30**(2): 248–292.

Games: Signaling

Intermediate article

Colin F Camerer, California Institute of Technology, Pasadena, California, USA

CONTENTS

In a signaling game, a player who is informed privately about an unobservable quality that he or she has chooses an observable signal, which an uninformed player observes and reacts to. Game theory describes how players' choices and beliefs match ('equilibrium'). Experiments show that behavior generally conforms to an equilibrium after players gain experience.

INTRODUCTION

A 'signal' is an observable action a person takes that can reveal something about the person. The idea was borrowed from the theory of international relations and formalized by Spence (1974), who shared a Nobel Prize in 2001. Spence's prime example was education. He theorized that employers cannot directly observe a person's intelligence, but can observe whether the person graduated from college. If college is too difficult for people who do not have the qualities the employer wants, finishing college 'signals' the presence of those qualities to the employer.

Many other examples share this underlying logic. A cheap warranty signals to consumers that a product isn't likely to break down often; if it was not so, the firm could not afford the costs of frequent product replacement. Giving flowers and small gifts (without the reminder of a splashy holiday like Valentine's Day) means that a person is thinking about his or her lover.

Signals are credible if they satisfy the following two properties. Firstly, signals must be affordable by certain types of people, for whom the cost of the signal is less than the benefit of the 'receiver' decoding the signal. (Note that signals may be free, but have a 'cost' because of the consequences of choosing one signal or another.) Education pays because it gets students jobs; credible warranties get firms eager customers; flowers are well worth it if she knows what they mean. Secondly, signals must be too expensive for players of the wrong type. Students who hate schoolwork cannot stand the pressure and boredom of a demanding university. People selling fake Rolexes from briefcases on the streets of New York cannot afford to offer warranties.

Combining these two properties, a logical observer can conclude that if one person buys the signal, and another does not, then the two people are of different types. This is called a 'separating' equilibrium. A Nash equilibrium is a pattern of mutual best responses, in which all players are making the best choices, knowing that all the others are doing so.

In an equilibrium, actions and beliefs match: signalers' beliefs about what will happen after they buy the signal turn out to be right; and the inferences that the observers of the signal draw about the 'type' of the person who chose the signal are correct. If actions and beliefs are not in equilibrium, then either somebody's belief is wrong or somebody is failing to extract information from what they see. In the education example, if intelligent students go to college but do not get better jobs as a result, then their guess about how employers would react was wrong and employers don't think that college graduates make better employees. Economists study equilibria in which these mistakes are not made, because they believe that players will adjust their behavior over time until consistency occurs.

Signaling is one way to explain actions that might otherwise seem irrational. Some armies build and test soldier loyalty with inefficient tasks, like digging a hole then filling it again. Digging and filling wastes valuable time and produces nothing. But it signals the soldier's willingness to obey orders. Another example is provided by labor strikes. The prevailing theory of why strikes occur is that a company's managers and workers can only credibly signal that they won't pay more, or won't work for less, by agreeing to stop paying and working. A strike is a way for both sides to signal their seriousness.

Signaling models have also begun to influence biology and anthropology. The theory of 'costly signaling' in biology (e.g., Zahavi, 1975) explains apparently fitness-reducing activities by asking what type of information they convey. Male animals often develop physical features that seem to be handicaps, such as a peacock's lush tail-feathers or an elk's heavy antlers, but these 'handicaps' may in fact signal fitness.

SENDER–RECEIVER GAMES

In a sender–receiver game, a sender observes her own 'type' and chooses a message. A receiver observes the message, but not the type, and chooses an action. The pay-offs of both players depend on the type, the message, and the action. (Types are features that the player knows about but are not perfectly evident to others.) All the games described above – signaling intelligence through education, signaling profits or wage requirements through strikes, signaling product quality through warranties or advertising – can be put into this general framework.

The first careful experiments on sender–receiver games were conducted by Brandts and Holt (1992) and Banks *et al*. (1994). Let us begin with game 1 from Brandts and Holt, shown in Table 1. Each cell in the table shows the pay-offs to a sender (the first number) and a receiver (the second number), depending on the sender's message and type (row), and the receiver's subsequent action (C or D). For example, if the sender is of type L and sends message S, and the receiver responds with action C, then the sender earns 140 and the receiver earns 75.

The pay-offs can be thought of as reflecting gains from education. Consider the following scenario: a worker draws her type, either L (low intelligence) or H (high intelligence). There is a common prior probability that her type is L with probability ⅓ and H with probability ⅔. (Everyone knows this probability distribution, and that the worker knows her own type.) After observing her type, the worker can either skip education (S) or invest in education (I). A prospective employer does not observe the worker's type (L or H), but does observe whether she invested in education (S or I). Then the employer assigns the worker to either a dull (D) job, which requires little skill, or a challenging (C) job, which requires more skill.

Table 1. Payoffs in sender–receiver game 1 (Brandts and Holt, 1992)

	Action after message S ('skip')		*Action after message I ('invest')*	
	C	*D*	*C*	*D*
Type L ('low')	140,75	60,125	100,75	20,125
Type H ('high')	120,125	20,75	140,125	60,75

The employer's pay-offs are simple: she wants to assign L employees to the D job, and H employees to the C job. Those assignments produce an employer pay-off of 125; the opposite assignments are mismatches and yield only 75.

The worker's pay-offs create a different incentive. Workers get pay-offs from both wages and 'psychic income'. Both types earn 100 from the challenging job C and only 20 from the dull job D. In addition, L types get an added pay-off of 40 if they skip college (S), and H types get an added pay-off of 40 if they invest in education and go to college (I). In addition, suppose the H types who skip college get an extra payoff of 20 from the challenging job C, perhaps reflecting on-the-job learning from the challenging job which is a substitute for what they would have learned in college. Adding up these pay-offs gives those in Table 1.

There is a conflict of interest between senders (workers) and receivers (employers) in this game. Employers would like to know the workers' types so that they can assign the Ls to job D and the Hs to job C. Since the probability that the worker's type is H is ⅔, if the employer doesn't learn anything about the worker's type from her choice of S or I, she will assign the worker to job C. (Assigning to C gives an expected pay-off of 108.3 and assigning to D gives 91.7.) Essentially, since H is more likely than L, it pays for the employer to take a chance and assign everyone to the more challenging job C.

Both types of workers prefer job C. Since the employer will assign a worker to job C unless she becomes fairly convinced (probability more than ½) that the worker is of type L, the type L workers have an incentive to 'pool' with the type H workers and mimic whatever the Hs do, so that they can get the lucrative C job assignment.

As in many signaling games, there is more than one equilibrium. In one 'pooling' equilibrium, both types choose S and employers respond with C. In addition, a complete 'Bayesian-Nash' equilibrium must specify how employers respond to unexpected messages ('off the equilibrium path'). In the equilibrium described above, to keep H types from breaking out of the pool and choosing I, employers must think that an unexpected choice of I is evidence that the worker is of type L, so that they assign a worker who sends message I to the dull job D. That is, in the S–C equilibrium, the

employer's belief about the worker's type, after the out-of-equilibrium choice I, is that the worker is more likely to be L than H. (That is, $P(L|I) > ½$.) Since the pay-offs from D after choosing I are lower for both types than the pay-offs from C after choosing S, neither type will deviate, and so the pattern of choosing S and getting job C is an equilibrium.

There is a problem with this equilibrium pattern. Imagine a country in which nobody gets advanced education for decades, and then one person does choose to get educated. Everybody knows that L types don't like education – they are bored and frustrated by college – but H types love it (they earn a higher psychic pay-off, reflected by the bonus of 40 in the game). Then why would an employer think that the person who went abroad to college is an L type? L types cannot possibly benefit, but H types might (if their education is taken as a clue that they are H types). The belief that L types are the ones to get educated first violates the 'intuitive criterion' proposed by Cho and Kreps (1987). Game theorists have proposed other criteria for when inferences are plausible or not. These criteria, called 'refinements', narrow the set of multiple equilibria by imposing further mathematical restrictions that only some equilibria satisfy. There is a second sequential equilibrium in which workers pool by choosing I, employers assign everybody who chooses I to job C, and a worker who deviates by choosing S is assigned job D. This equilibrium does satisfy the intuitive criterion because the H types cannot possibly earn a higher pay-off by deviating (by choosing I and getting job C they earn their maximum pay-off of 140), but L types might earn a higher pay-off.

Refinement concepts like the intuitive criterion were developed mathematically to resolve the question of which equilibrium is likely to occur if there are multiple equilibria (a difficult problem which has perplexed theorists). Whether refined equilibria are more likely to occur than unrefined ones is ultimately an empirical question. Experimental data show that only the simplest of these 'refinement' criteria match behavior closely.

Table 2 shows some experimental results along with equilibrium predictions. (The equilibrium predictions are the proportions of players predicted to choose the messages or actions shown in the column headings, according to the two different intuitive and unintuitive equilibrium criteria.) The relative frequencies of I messages and intuitive-equilibrium actions tend to support the intuitive equilibrium. The strongest evidence against it is a low frequency (25%) of I choices by L types in the early periods 1–4, but they quickly learn to pool with the Hs (who always choose I) by choosing I most of the time in later trials.

REPUTATION FORMATION

In game theory, a player's reputation is defined as the probability that he or she is a certain privately-observed type. Camerer and Weigelt (1988) explored these reputation models carefully, using a 'trust', or borrower–lender, game. In each eight-period repeated game, a single borrower drew a random type, either 'normal' (X) or 'nice' (Y). The type was the same for all eight periods. The same borrower played a series of eight-stage games, with a different lender player each time. The lender could either lend or not lend; if choosing to lend, the borrower then decided whether to default or repay. The pay-offs are shown in Table 3. Not lending gives the lender 10, and lending pays 40 if the loan is repaid or −100 if the borrower defaults. A normal borrower prefers to default, and then get 150, instead of getting 60 from repayment. A nice borrower gets 0 from defaulting and thus prefers to repay.

The equilibrium can be calculated by working backwards from the end. In the last period, normal borrowers will always default. Lenders who anticipate this will lend only if the chance that the borrower is nice is above some threshold (0.79). Moving back to the seventh period, normal borrowers know that if they default in period 7 they will reveal their type and lenders will be afraid to lend to them in period 8. It pays for them to *sometimes* repay (using a 'mixed strategy'). Working

Table 2. Results in sender–receiver game 1 (Brandts and Holt, 1992)

Periods	*Message given type*		*Action given message*		*Equilibrium predictions*	
	I|H	I|L	C|I	D|S	*Intuitive*	*Sequential (unintuitive)*
1–4	1.00	0.25	1.00	0.74	1	0
5–8	1.00	0.58	0.00	1.00	1	0
9–12	1.00	0.75	0.98	0.60	1	0

Table 3. Pay-offs in the borrower–lender (trust) game (Camerer and Weigelt, 1988)

Lender strategy	*Borrower strategy*	*Pay-off to lender*	*Pay-off to borrower*	
			normal (X)	*nice (Y)*
Lend	Default	−100	150	0
	Repay	40	60	60
Don't lend	No choice	10	10	10

Table 4. Lending and repayment rates (Camerer and Weigelt, 1988)

	Experiments	*Period (1 = start, 8 = end)*							
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Conditional frequency of lending	3–5	Predicted	1	1	1	1	0.64	0.64	0.64	0.64
		Actual	0.94	0.94	0.96	0.91	0.72	0.59	0.38[a]	0.67
	6–8	Predicted	1	1	1	0.64	0.64	0.64	0.64	0.64
		Actual	0.96	0.99	1.00	0.95[a]	0.85[a]	0.72	0.58	0.47
	9–10	Predicted	1	1	1	0.64	0.64	0.64	0.64	0.64
		Actual	0.93	0.92	0.83	0.70	0.63	0.72	0.77	0.33
Conditional frequency of repayment by normal (X) types	3–5	Predicted	1	1	1	0.81	0.65	0.59	0.44	0
		Actual	0.95	0.97	0.98	0.95[a]	0.86[a]	0.72	0.47	0.14
	6–8	Predicted	1	1	0.73	0.68	0.58	0.53	0.40	0
		Actual	0.97	0.95	0.97[a]	0.92[a]	0.85[a]	0.70[a]	0.48	0
	9–10	Predicted	1	1	0.73	0.67	0.63	0.56	0.42	0
		Actual	0.91	0.89	0.80	0.77	0.84[a]	0.79[a]	0.48	0.29

[a]Significant difference ($|z| > 2$) between predicted and actual frequencies.

further back, it pays to repay early in the game, because defaulting means that the borrower will never get a loan to repay ever again in that eight-period sequence. For example, defaulting immediately means that the normal borrower earns 150 plus a string of seven payoffs of 10, a total of 220. But repaying for the first three periods earns at least $3 \times 60 + 5 \times 10 = 230$, which is better than defaulting right away and getting no more loans. Table 4 shows the predicted and actual frequencies of lending and repaying in each of the eight periods (assuming there has not been a default earlier in the sequence). Note that these predictions are precise, and derived with no free parameters.

Two patterns in the data are of particular interest. Firstly, what is the rate of lending and default across periods? And secondly, does lending and repayment in each period of an eight-period sequence reflect prior history in that sequence, as the theory predicts?

The table shows the conditional frequencies of lending and repayment (by normal types), from the last two-thirds of all the sequences, pooled together. Actual frequencies significantly different from the predictions are marked.

One prediction is that lending should drop off sharply in period 5 (for experiments 3–5) or 4 (experiments 6–10). Lending does drop off, but not quite as sharply as predicted. For example, in experiments 3–5 the overall frequencies of lending in all early and late periods are 0.95 and 0.62 (predicted to be 1 and 0.64); in experiments 6–8 the corresponding figures are 0.98 and 0.80 (same predictions).

Repayment rates do fall, from close to 100% in the first couple of periods, to nearly zero in the last period, but repayment is generally more common than predicted. This deviation can be explained by the hypothesis that some subjects who earn normal pay-offs simply prefer to reciprocate by repaying.

Camerer and Weigelt concluded that 'sequential equilibrium predicts reasonably well, given its complexity'. That is, while the equilibrium predictions are not perfectly accurate, they vary across periods in the correct direction and are not far off in magnitude. Replications by Neral and Ochs (1992) and Brandts and Figueras (1997) support this basic conclusion.

Camerer *et al*. (2002a) propose a boundedly rational explanation of how players learn to

approximate the equilibrium. In their theory, some players learn what strategies work best from experience. Other players are 'teachers', who realize that others are learning, and choose strategies that change what the learners do in the future to benefit themselves (the teachers). In the trust game, teachers repay in early periods because they are trying to 'teach' the lenders that lending is profitable. But late in the game, the payback period from teaching is too short, so teachers eventually default. The teaching theory retains strategic foresight, but relaxes the requirement in equilibrium models that players all correctly guess what strategy others will follow.

CONCLUSION

In signaling games, one player observes private information and takes an action that may signal that information to a 'receiver' player (who knows the probability distribution of the private information, but does not know the information). The logic that underlies predictions about patterns of mutually-consistent equilibrium play is complex and depends subtly on players' own choices and beliefs and the choices and beliefs of others.

Experiments are particularly useful for testing signaling models, because the predictions depend delicately on variables that are difficult to observe directly. Early experimental results are supportive of equilibrium predictions, but equilibration seems to take time. Some recent experiments have been modeled on specific phenomena: political lobbying (Potters and van Winden, 1996, 2000), firms issuing shares of stock (Cadsby *et al*., 1998), and aggressive pricing to crush competition (Cooper *et al*., 1997a; Jung *et al*., 1994). In most games, players show limited strategic sophistication at first, followed by learning. That is, players often take actions that 'tip their hands', revealing their types to other players in a way that can be exploited.

Cooper *et al*. (1999) conducted an experiment on 'ratchet effects' in planned economies (Chaudhuri, 1998). (In a planned economy, firms are instructed to meet a quota, but do not earn more if they exceed the quota. Productive firms are effectively penalized for doing well because their quotas are raised in the future. Clever managers therefore deliberately underproduce to avoid having their quota ratcheted up.) Cooper *et al*. studied students and factory managers in China to see whether experience in a planned economy taught the managers to underproduce to avoid being ratcheted. When the games were described in the familiar language of production and quotas, the managers played closer to the game-theoretic prediction (that is, the high-productivity managers produced 'too little', to avoid being ratcheted) than the students.

While approximate equilibrium emerges over time, early behavior is constrained by cognitive limits: players do not initially anticipate how their choices reveal information; and players do not make immediate inferences from other players' choices, as equilibrium theories require. New models that incorporate limited cognition (Camerer *et al*., 2002b) and learning are a promising way to understand how players behave. Cognitive science could inform game theory further. As Van Damme (1999) concluded: 'At present our empirical knowledge is inadequate and it is an interesting question why game theorists have not turned more frequently to psychologists for information about the learning and information processing processes used by humans.'

References

Banks J, Camerer CF and Porter D (1994) An experimental analysis of Nash refinements in signaling games. *Games and Economic Behavior* **6**: 1–31.

Brandts J and Figueras N (1997) An exploration of reputation formation in experimental games. Institut d'Anàlisi Econòmica (CSIC) working paper.

Brandts J and Holt C (1993) Adjustment patterns and equilibrium selection in experimental signaling games. *International Journal of Game Theory* **22**: 279–302.

Cadsby CB, Frank M and Maksimovic V (1998) Equilibrium dominance in experimental financial markets. *Review of Financial Studies* **11**: 189–232.

Camerer CF and Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. *Econometrica* **56**: 1–36.

Camerer CF, Ho T and Chong JK (2002a) Sophisticated experience-weighted attraction and strategic teaching in repeated games. *Journal of Economic Theory* **104**: 137–188.

Camerer CF, Ho T and Chong JK (2002b) A cognitive hierarchy model of thinking in games. Caltech working paper.

Chaudhuri A (1998) The ratchet principle in a principal agent game with unknown costs: an experimental analysis. *Journal of Economic Behavior and Organization* **37**: 291–304.

Cho I-K and Kreps D (1987) Signaling games and stable equilibria. *Quarterly Journal of Economics* **102**: 179–221.

Cooper DJ, Garvin S and Kagel JH (1997) Signalling and adaptive learning in an entry limit pricing game. *RAND Journal of Economics* **28**: 662–683.

Cooper DJ, Kagel JH, Lo W and Gu QL (1999) Games against managers in incentive systems: experimental results with Chinese students and Chinese managers. *American Economic Review* **89**: 781–804.

Jung YJ, Kagel JH and Levin D (1994) On the existence of predatory pricing: an experimental study of reputation and entry deterrence in the chain-store game. *RAND Journal of Economics* **25**: 72–93.

Neral J and Ochs J (1992) The sequential equilibrium theory of reputation building: a further test. *Econometrica* **60**: 1151–1169.

Potters J and van Winden F (1996) Comparative statics of a signaling game: an experimental study. *International Journal of Game Theory* **25**: 329–353.

Potters J and van Winden F (2000) Professionals and students in a lobbying experiment. *Journal of Economic Behavior and Organization* **43**: 499–522.

Spence AM (1974) *Market Signaling*. Cambridge, MA: Harvard University Press.

Van Damme E (1999) Game theory: the next stage. In: Gerard-Varet LA, Kirman AP and Ruggiero M (eds) *Economics Beyond the Millenium*, pp. 184–214. Oxford, UK: Oxford University Press.

Zahavi A (1975) Mate selection: a selection for a handicap. *Journal of Theoretical Biology* **53**: 205–214.

Further Reading

Camerer CF (2002) *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton, NJ: Princeton University Press.

Cooper DJ, Garvin S and Kagel JH (1997) Adaptive learning vs. equilibrium refinements in an entry limit pricing game. *Economic Journal* **107**: 553–575.

Fudenberg D and Tirole J (1994) *Game Theory*. Cambridge, MA: MIT Press.

Gibbons R (1992) *Applied Game Theory for Economists*. Princeton, NJ: Princeton University Press.

Games: Trust and Investment

Advanced article

John Dickhaut, University of Minnesota, Minneapolis, Minnesota, USA
Aldo Rustichini, University of Minnesota, Minneapolis, Minnesota,USA

CONTENTS

Trust and investment games are experimental methods for assessing behaviorally the presence of trust and reciprocity.

THE NATURE OF INVESTMENT, POSSIBLE INEFFICIENCIES, AND REMEDIES

In economic environments, an investment is an advance payment to earn future productive returns and benefits. The agent who makes the investment must have the appropriate incentives. For instance, a company involved in scientific research is willing to pay for the costs of the research only if there is a reasonable probability that adequate benefits and profits may be reaped in the future. This might not be the case if competitors are allowed to copy a new design at no cost. Patents are a way to guarantee such a return.

When prospects of future benefits are not reasonably large and reasonably assured to the person making the advance payment, then the potential investor may decide that the investment is not convenient; then, the productive capability of the investment is lost. This outcome is inefficient, since the total amount of resources available reflects a forgone opportunity.

One way of overcoming this inefficiency is the institution of complete, binding, irrevocable contracts. The contract has to be binding, or it has no power; it has to be complete, or the uncertain outcome in some event might destroy the incentive to invest. And it must be irrevocable: the person who makes the advance payment in the form of investment is in an unfavorable bargaining position once the payment is made.

Clearly, contracts with such stringent requirements are hard to realize, and when they are realized they are typically costly and complex. In many informal or infrequent interactions, a contract is either not feasible or not convenient. A second way to insure the appropriate investment in repeated situations is building a reputation through the threat of future sanctions. An informal agreement to return something to the investor is credible if the situation is going to be repeated. The investor may threaten to stop the investment in future periods. The promise of future benefits, and the threat of their loss, may induce the trustee to honor his or her dues today.

When contracts are impossible, and the exchange is not repeated, investments are likely to be inefficient. The loss may be considerable. Yet often in real exchanges we observe investments even when these two solutions are unavailable. It is important to understand why investment still occurs. Recent research in experimental economics addresses this issue precisely.

The investment game is an effective tool in this research. It is a simple game, specifically designed for a laboratory environment. Its aim is to investigate whether observable investment occurs if agents cannot write binding contracts and cannot create reputations.

THE EFFECTIVENESS OF TRUST

The investment game is a tool to study how widespread and effective trust is in social interactions. Consider any situation in which one agent is giving another the power to make a decision that affects the utility of both. Trust occurs if, in pursuing an outcome that is potentially favorable to the two players, one of the two puts himself or herself willingly in a situation of vulnerability. It is far from clear whether trust has any effect in economic interactions between selfish agents. It is also far from clear how different factors may affect trust if it does

exist. The investment game allows a controlled study of the effectiveness and nature of trust in economic interactions.

THE STRUCTURE OF THE INVESTMENT GAME

In the basic investment game, or trust game, two players move in sequence. Player 1 has an initial monetary endowment A. Player 1 decides the amount x of A to transfer to player 2. The transfer x is a productive investment; it increases by a factor of R ($R > 1$) before reaching player 2. Then, player 2 decides how much, a, of Rx to return to player 1. Player 1 leaves with $[A-x+a]$ and player 2 leaves with $Rx-a$. For example, if the multiple is 3, and player 1 has an initial \$10 and decides to transfer \$5, player 2 would receive \$15. If player 2 decides to return \$8 out of the \$15, the final pay-offs for the players are \$13 and \$7 respectively.

Several features of the game deserve comment. Firstly, there is no possibility of writing binding contracts, so that player 2 cannot be penalized for not sending something back to player 1. Secondly, there is no opportunity for reputation formation, at least provided the game is played only once and the subjects are never going to meet again. Similarly, the threat of a zero transfer in future periods, which might be effective in situations where the game is repeated by the same subjects in the same roles, is ineffective. Given the assumption that each subject prefers more money to less, then the second player should not return any of the Rx that he was sent. Given that player 1 knows that player 2 follows the same principles of behavior as player 1, then player 1 knows he will receive nothing back, thus player 1 will send nothing. In summary, the investment game is a strategic situation in which, under standard economic modeling assumptions, no investment should occur.

This conclusion is valid even for very large R. Yet if the first player rationally decides to make no initial transfer, the two players completely forgo the potential benefits of the productive technology represented by R: the total amount they receive is A, rather than RA, which would be technically feasible. It is clear that if they could sign binding and irrevocable contracts before the beginning of the game, then the rational course of action would prescribe a transfer equal to the total amount A by the first player. This choice would maximize the total amount available to them, and they could bargain on its final allocation.

EXPERIMENTAL EVIDENCE

The original experiment implementing this game was run with very tight controls (Berg *et al.*, 1995). Two features were crucial. Firstly, the anonymity from the experimenter, as well as subjects' anonymity from each other, was guaranteed. Secondly, each subject was matched with only one other subject. The values chosen for the parameters were $A =$ \$10 and $R = 3$.

While the prediction of the sub-game perfect equilibrium is a zero transfer of the first player and a zero amount returned, the average amount sent was \$5.16, and the average amount returned was \$4.66. Very few subjects among the first movers sent zero (2 out of 31). The amount sent was highly variable, with a distribution statistically similar to a uniform distribution over the range of possible transfers. Many among the second movers who had received an amount of \$1 or more returned less than \$1 (12 out of 29). But 11 of the same 29 subjects returned more than they had received. It has become customary to call player 1's behavior 'trusting behavior' and player 2's 'positive reciprocity'.

This finding leads to the conclusion that new assumptions about individual behavior need to be incorporated in modeling activities in industrial organization, such as the modeling of asymmetric information, to better explain observed phenomena. Several related games, with similar designs, have produced the same experimental outcomes.

RELATED GAMES

In the investment game, the first mover has to make a decision that is costly to him or her, but potentially beneficial for both players. The benefits to the first mover, however, come only after a similar costly decision for player 2: so player 1 cannot insure that he will get some of the benefits. On the contrary, purely selfish motivations of the second player would imply that the first mover cannot derive any benefit from his payment; so he should not make any positive transfer. The same dilemma exists in several related games that have been tested. The general result, that an advance positive payment of player 1 is made, is confirmed (with some important exceptions). We now review some of these variations.

The Gift Exchange Game

As in the investment game described above, two players move in sequence. Player 1 has to propose

a wage w, in a given interval. Player 2 can accept or refuse. If player 2 refuses, both players get a zero pay-off. If player 2 accepts, he or she has to choose a costly effort e. Then the game is over, and the first player gets a pay-off of $Re-w$ while the second gets $w-c(e)$, where c is a function that is increasing at an increasing rate. The transfer, or investment, in the basic investment game is replaced by the commitment of the first player to pay a wage w, irrespective of the effort of the second player; the return payment is replaced by the effort e and its cost $c(e)$ to the second player, $c(0) = 0$. The opportunity for both players is represented by the productive factor R.

As in the investment game, a rational selfish first player should anticipate a minimum effort of the second player, and should promise a minimum wage. With wage, effort, and cost of effort all equal to zero, both players get zero equilibrium pay-off, even though $Re-c(e)$ is achievable (and hence divisible to each) with a feasible effort e.

Experimental evidence

This game was introduced and tested in Fehr *et al*. (1993, 1998); see also (Charness, 1996, 2000). In contrast to the equilibrium prediction of minimum effort and minimum wage (the market clearing levels of the competitive labor market), the observations in the experiment were consistent with a fair wage hypothesis. This hypothesis can be formulated as three statements: the effort level is increasing in the wage; the average wage is higher than the minimum necessary to assure employment; and the average effort is higher than the minimum. These outcomes persist after replications of the same (one-off) game.

The Moonlighting Game

In the investment game, player 1 can transfer only positive amounts of money to player 2. In the moonlighting game, both players begin with a positive endowment. Player 1 can give an amount g or take away an amount t from the second player. An amount given increases by a factor of R; an amount taken away is simply transferred to player 1. Player 2 can then return money or take money away from the first player: an amount b given back costs b to player 2, while the punishment in the form of an amount p penalizes player 1 $3p$. Player 1 is a moonlighter, his activity cannot be contracted upon. The game allows the study of positive and negative reciprocity.

Experimental evidence

The moonlighting game was introduced and tested in 'one-off' experiments by Abbink *et al*. (1999). They found that three-quarters of subjects transferred a positive amount g; on average, this transfer induced a positive return b. An amount taken away t is heavily punished with a positive p. Retribution is thus considered to be stronger and more effective than reciprocity. Player 2 was also allowed to propose a non-binding contract. These contracts increased trust but did not encourage reciprocity.

The Peasant–Dictator Game

In the peasant–dictator game, player 1 is a peasant and player 2 a dictator. The peasant decides the amount k he or she will invest out of an available amount W. The investment produces income $(1 + r)k$. Then the dictator moves: he or she can impose a tax on the income produced, and the amount $t(1 + r)k$ goes to him or her; $(1 - t)(1 + r)k$ returns to the peasant, and is added to $W - k$.

Two treatments are possible. In one, the dictator may commit to a level of taxes before the peasant decides. In another (slightly different but economically equivalent), the dictator announces the level of taxes before the investment, but is not committed to the announcement. When the announcement has no binding force, the economically rational choice for the dictator for any level of investment is to tax the entire amount. When the announcement is binding, the dictator knows that too high a tax will induce the peasant to make no investment.

The peasant–dictator game with no commitment is equivalent to the investment game: the amount invested by the peasant is equivalent to the transfer made by player 1 in the investment game, and the amount left after taxation corresponds to the amount returned by player 2 in the investment game. The specific parameters used by Berg *et al*. correspond to a rate of return $r = 2$ in the peasant–dictator game, with W equal to the initial endowment of the first mover. A possibly important difference in the description of the game is the nature of the amount decided by the first mover. In the instructions for the peasant–dictator game the amount is described as an investment, while in the investment game it is described as a transfer from the first to the second mover.

Experimental evidence

The game was introduced and tested by Van Huyck *et al*. (1995). The game with commitment played by rational selfish individuals would have an outcome where positive investment is made, while in the game without commitment the

peasant would not invest anything. Van Huyck *et al*. emphasized the difference in behavior between the commitment and no-commitment games. Their qualitative results correspond to the game-theoretical analysis: in the games where no commitment was possible, the average investment, and aggregate earnings, were lower.

No direct comparison was made between the results in Berg *et al*. (1995) and Van Huyck *et al*. (1995). From the evidence available it is not clear that the level of trust in the peasant–dictator game is lower than that in the investment game. In both games there is a positive level of investment without available commitments.

RECIPROCAL BEHAVIOR AND ITS REASONS

The results described above prompt certain questions. What is the reason for such behavior? Why, in particular, do we observe the deviation from predictions based on the assumption of selfish rational behavior – assumptions that are supported in different contexts? Are these deviations robust and meaningful? Important approaches to these questions can be found in various manipulations of the basic game that have been tested experimentally.

Manipulations

Social history

In a variation of their main experiment, Berg *et al*. (1995) showed subjects a report on the outcome of the same experiment run a few months earlier in the same university. The report showed, for each level of transfer, the number of subjects who had chosen that transfer, the average amount returned, and the average profit (average amount returned minus average amount sent). The behavior in the two treatments (with and without seeing a social history) was different, and statistically significant at the 6 percent level, suggesting an increase, among those subjects who had seen a social history, in the correlation between the amount sent and the amount sent back.

Induced strategic reasoning

Boeing *et al*. (2000) asked whether informing subjects of the theoretical logic of sub-game perfection makes them perform more consistently with the theory. They modified the way in which information was presented to participants and, through a questionnaire, prompted strategic reasoning. None of the various treatments led to a significant reduction in the amounts invested. The results of these experiments suggest that the findings of Berg *et al*. (1995) are robust to changes in information presentation and to cues about strategic reasoning.

Mood induction

A large literature in psychology has shown that mood and emotions can affect altruism, defined as the propensity to help unknown people with no material reward. The common finding is that good mood increases altruism. Reviews of these findings and their methodology can be found in Lewis and Haviland-Jones (2000) and Gilbert *et al*. (1998): in particular, see Isen (2000) and Zajonc (1998). An illuminating review of the main hypothesis and explanations is in Carlson *et al*. (1988).

The effect of mood on reciprocity, however, is not clear. An experimental analysis of this question was attempted by Kirchgeister *et al*. (2001). Their experiment was based on the investment game. Before the game, however, subjects in the role of player 2 underwent a mood induction phase. There were two treatments: to a first treatment group, a 'sad' film was shown (a five-minute sequence from *Schindler's List*); to a second treatment group, a 'happy' film was shown (a five-minute sequence from *City Lights*).

The mood induction was effective. The behavior of players in the two different moods was significantly different. The return transfer of 'happy' subjects was relatively insensitive to the original investment. 'Happy' subjects who received a low transfer x replied with a higher a than 'sad' subjects, while 'happy' subjects who received a high transfer x replied with a lower a than 'sad' subjects.

Personal and social interactions

The original investment game captures what can be called 'direct reciprocity'. However, in various social exchanges, reciprocity may not directly involve the subject who first makes the investment, but other members of a group. Subjects are called indirectly reciprocal if they are willing to reciprocate even if the originator of the transfer is not the target of the reciprocity, but belongs to the same group. Buchan *et al*. (2001a) devised and tested a variation of the original trust game to examine the extent to which subjects are willing to trust for the benefit of a group or a community.

The original design is called the 'direct condition'. In the 'group condition', a sender A transfers money to a responder B, and a sender C transfers money to a responder D. The money is multiplied

by 3, as in the original experiment of Berg *et al*. (1995). Responder *B* can then return some proportion of her wealth to sender *C*, while responder *D* can return some proportion of his wealth to sender *A*. In the 'society condition', sender *A* sends money to responder *B*, while responder *D* receives money from a different, randomly chosen, sender in the same room as sender *A*. Responder *D* then returns some proportion of his wealth to sender *A*, while responder *B* returns some proportion of her wealth to a randomly chosen sender in the same room as sender *A*. Subjects are said to engage in indirect trust by sending money even when they are not the recipients of potential reciprocation.

Out of a total amount normalized to 1000, the mean amounts sent were 782.14 units in the direct condition, 488.64 units in the group condition, and 505.46 units in the society condition. Thus, even in the society condition, where there is no possibility of personally benefiting from reciprocation, the amounts sent far exceed the equilibrium prediction.

Both the trust of the first mover and the reciprocity of the second mover, for both the group and the society conditions, are significantly reduced as the reciprocation becomes more indirect. But there is strong evidence of reciprocity even when the reciprocation is indirect. The mean proportions returned (fraction of the money returned compared with the total amount available to the two players after the productive investment) are 0.319 in the direct condition, 0.114 in the group condition, and 0.130 in the society condition. The situation for the amount sent is similar. For both the amounts sent and the amounts returned, there was no significant difference between the group and society conditions.

Choice of partner

Eckel and Wilson (2000) designed a variation of the basic investment game to study the partners people choose in investment enterprises. In their design, subjects were allowed to choose between two trust games. These games could be slightly different, or identical. The games had a marker represented by an icon, a stylized image of a face, which could appear to be friendly, unfriendly, or neutral. The first mover could choose the icon (hence the game), and then make the first move. The second mover could then observe the game chosen (and the icon) and the first move of the sender. (Eckel and Wilson interpret the choice of icon as the choice of the type of partner that the first mover would like to meet. This interpretation would not necessarily be made by subjects: for example, the choice of an angry icon might be interpreted by both players as a signal that the first mover would be angry at a lack of reciprocation of the second player.) Eckel and Wilson found that people chose partners who appeared to be most trustworthy. A choice of the neutral icon was usually followed by a more trusting choice of investment.

There were some counterintuitive results. For instance, a choice of angry icon was followed by a high level of trusting behavior; and reciprocation by the second player was higher when the icon chosen by the first was angry. These results might be explained on the basis of the interpretation of the choice of icon as a signal sent to the second mover.

Factors Affecting Behavior

Male and female behaviors

One dimension of analysis of the trust game is the study of the behavior of players of different sexes in the trust game. It is known that there are significant differences. For instance, in the peasant–dictator game, Bolton and Katok (1995) and Eckel and Grossman (1998) found that women gave more than men. The difference was significant in the second paper. Andreoni and Vesterlund (1998) compared differences in behavior between men and women in dictator games for different monetary values of the tokens to be divided. They found that women gave more overall and were more likely to divide tokens evenly despite changes in monetary value, while men became less generous as the value of their tokens increased.

For the trust game, Croson and Buchan (1999) examined this issue in a study extending over four different countries. The design of the experiment was the same as in Berg *et al*. (1995). Men and women sent approximately the same amount as first movers. But women were more reciprocal. The proportion of money returned grew by approximately one-third per dollar, and for women this factor was increased by approximately one-eighth. In Croson and Buchan (1999), the sex of the opponent was unknown. Buchan *et al*. (2001) varied the information that subjects had about the sex of the opponent. When subjects knew the sex of the opponent, the difference between men and women in observed trust and reciprocity became more apparent: men sent more than women (to subjects of either sex) when they were first movers; both men and women sent more to men than to women; and men sent significantly more to other men than to women.

Culture and environment

The trust game has been applied in different countries, providing a test of the possible importance of cultural and social factors. Bond and Hwang (1996) found evidence of face-saving behavior (investing to avoid appearing selfish) among subjects, a fact that makes even more desirable the use of double-blind procedures.

A more systematic study was conducted by Buchan *et al*. (2001b), who used the experiment designed to test direct and indirect trust and reciprocity to look for differences in behavior in four different countries. The study involved students from China (Beijing), Korea (Seoul), Japan (Osaka) and the United States (Madison, Wisconsin).

The differences in the levels of trust were significant and sizeable. Japanese and Korean subjects extended significantly less trust than did American subjects. Also, the amounts sent by American and Chinese subjects significantly differed from the amounts sent by Japanese and Korean subjects.

Expectations

Morrison and Rutstrom (2000) studied the effect that beliefs have on behavior. They used a discrete version of the trust game, in which the first mover could choose only between sending or not sending and the second could choose only between returning and not returning.

Morrison and Rutstrom focused on the prior belief of the second mover that the first mover will send money. A second mover is surprised if he or she was expecting no money to be sent yet receives a transfer. The main finding is that a surprised second mover is less likely to return the money.

The Need for New Theories

How can we give a unified explanation of these different phenomena? Probably the most widely used approach is to attempt to adjust standard theories of preferences to address the trust issue. Such an approach would introduce additional arguments to the utility function and then ask whether observed data can be accounted for. A common assumption is that individuals' preferences are defined in terms of their own and other people's pay-offs. This is systematically investigated in Fehr and Schmidt (1999). For example, reciprocation in the trust game occurs because player 2's preferences do not just depend on his or her own wealth but also on player 1's final wealth. If this approach is correct, then other traditional means of analysis remain valid. There is no doubt that this approach will continue to provide useful insights.

However, rather than viewing trust as something to be incorporated into existing models, it is possible that a more basic understanding of trust could arise from a more general attempt to trace behavioral regularities in strategic situations back to structures and functions of the brain. Neuronal structures are generally understood to be at the basis of all behavior; yet they may defy the application of the traditional mathematical structures used by decision scientists to examine choice. McCabe *et al*. (2001) have begun to examine the behavior of subjects in multiple-person settings using functional magnetic resonance imaging. They have found brain activation patterns for trusting players when they play against other humans that are different from when they play against computers. Their study is a first step in attempting to unravel the anatomical underpinnings of the trust phenomenon. It is not yet fully clear how such phenomena can easily be subsumed in any traditional account of choice. A major question still to be answered is how activation varies as a function of the level of observed trust. A hint at how this question might be addressed is provided by Smith *et al*. (2001). Their study finds that relatively more brain activation can be traced to ventromedial (dorsomedial) areas depending on the presence of ambiguous (risky) stimulus comparisons. The ventromedial area corresponds to areas that are evoked in the presence of emotional factors, while the dorsomedial area evokes areas that are associated with computation.

As we have noted, levels of trust are not independent of factors like social history, sex, and culture. The results in Kirchgeister *et al*. (2001) show that the prior viewing of a film sequence which contains no relevant information can significantly alter behavior in an investment game, suggesting that mood and emotion are involved in the way trusting behavior and reciprocity are regulated. This in turn suggests the following conjecture: the same ventromedial area that is associated with emotional factors in conjunction with ambiguous stimuli may also be associated with morally based emotional choice which is sensitized by such things as seeing *Schindler's List* or observing social history. Furthermore, this area would be activated in different degrees for men and women and for people of different cultures. Of course, it is possible to incorporate such factors as social history and sexual and cultural differences into a traditional functional representation, as in Fehr and Schmidt (1999), but such factors may be so far removed from human

deliberation that it may be more parsimonious to invoke notions of automaticity.

References

Abbink K, Irlenbusch I and Renner E (2000) The moonlighting game: an experimental study on reciprocity and retribution. *Journal of Economic Behavior and Organization* **42**: 265–277.

Andreoni J and Vesterlund L (2001) Which is the fair sex: gender differences in altruism. *Quarterly Journal of Economics* **116**: 293–312.

Berg J, Dickhaut J and McCabe K (1995) Trust, reciprocity and social history. *Games and Economic Behavior* **10**: 122–142.

Boeing C, Fitzgerald J and Ortman A (2000) Trust, reciprocity, and social history: a re-examination. *Experimental Economics* **3**(3): 81–100.

Bolton G and Katok E (1995) An experimental test for gender differences in beneficent behavior. *Economic Letters* **48**(3–4): 287–292.

Bond M and Hwang K (1996) The social psychology of the Chinese people. In: Bond M (ed.) *The Psychology of the Chinese People*, pp. 213–266. New York, NY: Oxford University Press.

Buchan N, Solnick S and Croson R (2001a) Gender and trust. [OPIM Working Paper, Wharton School of the University of Pennsylvania.]

Buchan N, Croson R and Dawes R (2001b) Direct and indirect trust and reciprocity. [OPIM Working Paper, Wharton School of the University of Pennsylvania.]

Carlson M, Charlin V and Miller N (1988) Positive mood and helping behavior: a test of six hypotheses. *Journal of Personality and Social Psychology* **55**: 211–229.

Charness G (1996) Attribution and reciprocity in a labor market: an experimental investigation. [Mimeograph, University of California at Berkeley.]

Charness G (2000) Responsibility and effort in an experimental labor market. *Journal of Economic Behavior and Organization* **42**: 375–384.

Croson R and Buchan N (1999) Gender and culture: international experimental evidence from trust games. *American Economic Review, Papers and Proceedings* **89**(2): 386–391.

Eckel C and Grossman P (1998) Are women less selfish than men: evidence from dictator experiments. *Economic Journal* **108**(448): 726–735.

Eckel C and Wilson (2000) Whom to trust? Choice of partner in a trust game. [Rice University discussion paper.]

Fehr E, Kirchgeister G and Riedl A (1993) Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* **108**: 437–460.

Fehr E, Kirchgeister G and Riedl A (1998) Gift exchange and reciprocity in competitive experimental markets. *European Economic Review* **42**: 1–34.

Fehr E and Schmidt M (1999) A theory of fairness, competition and co-operation. *Quarterly Journal of Economics* **114**: 817–868.

Gilbert D, Fiske S and Lindzey G (eds) (1998) *The Handbook of Social Psychology*, 4th edn. New York, NY and Oxford: Oxford University Press and McGraw-Hill.

Isen AM (2000) Positive affect and decision making. In: Lewis and Haviland-Jones (2000).

Kirchgeister, Rigotti and Rustichini (2001) Your morals are your moods. [Center discussion paper, Tilburg University.]

Lewis and Haviland-Jones (eds) (2000) *Handbook of Emotions*. New York, NY and London: Guilford Press.

Morrison W and Rutstrom E (2000) The role of beliefs in an investment game experiment. [Discussion paper, University of Calgary, Canada.]

McCabe K, Houser D, Smith V and Truord T (2001) A functional imaging study of cooperation in two persons reciprocal exchange. [Discussion paper, University of Arizona, USA.]

Smith K, Dickhaut J, McCabe K and Pardo J (in press) Neuronal substrates for choice under ambiguity, risk, gains, and losses.

Van Huyck JB, Battalio R and Walters M (1995) Commitment versus discretion in the peasant–dictator game. *Games and Economic Behavior* **10**: 143–170.

Zajonc RB (1998) Emotions. In: Gilbert *et al*. (1998), pp. 591–632.

Further Reading

Arrow K (1974) *The Limits of Organization*. New York, NY: Norton, York.

Dasgupta P (1988) Trust as a commodity. In: Gambetta D (ed.) *Trust, Making and Breaking Cooperative Relations*. New York, NY: Blackwell.

Fehr E and Gachter S (1998) Reciprocity and economics: the economic implications of *Homo reciprocans*. *European Economic Review* **42**(3–5): 845–859.

Glaeser E, Laibson E, Scheinkman J and Soutter C (2000) Measuring trust. *Quarterly Journal of Economics* **155**: 811–846.

Le Doux J (1996) The emotional brain: the mysterious underpinnings of emotional life. New York, NY: Simon & Schuster.

Pinker S (1997) *How the Mind Works*. New York, NY: WW Norton.

Tooby J and Cosmides L (1978) The psychological foundations of culture. In: Barkow J, Cosmedes L and Tooby J (eds) *The Adapted Mind: Evolutionary Psychology and Generation of Culture*, pp. 19–136. New York, NY: Oxford University Press.

Trivers R (1972) The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57.

Williamson O (1985) *The Economic Institutions of Capitalism*. New York, NY: Free Press.

Williamson O (1993) The calculativeness, trust and economic organizations. *Journal of Law and Economics* **36**: 453–486.

Games: Ultimatum

Advanced article

Terence C Burnham, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

The ultimatum game is a simple, two-person strategic situation. Human behavior in the ultimatum game and the economic prediction of behavior diverge in a robust manner with profound significance for economics.

INTRODUCTION

The ultimatum game is a simple strategic situation between two people. One person, called the proposer, divides a fixed amount of money into two parts. This division is presented to the second person, termed the responder, as a 'take it or leave it' offer (hence the name 'ultimatum'). The responder has two options: to either accept or reject the proposed division. If the responder accepts, then the two parties divide the money according to the proposal. If the responder rejects, then both parties receive nothing. In either case the game ends with the responder's decision. In most experimental settings, the proposer and responder communicate through the experimenter and never learn each other's identity.

Many of the experiments that are discussed below were played for cash stakes of the order of 10 US dollars. For example a proposer might offer a split of $7 for the proposer and $3 for the responder. If the responder accepts this division, the $10 is allocated as the proposer suggested: $7 to the proposer and $3 to the responder. If the responder rejects this division, then neither the responder nor the proposer receives any money. The game ends at this point; responders who reject the first offer (the 'ultimatum') are not allowed to make a counter-proposal.

THE IMPORTANCE OF THE ULTIMATUM GAME IN ECONOMICS

The ultimatum game has been the subject of hundreds of articles and fierce debate within economics. This prompts the question of why such a simple game should be so important. The answer is both empirical and theoretical. As we will see, there is no widely accepted theory to explain the way people behave in the ultimatum game, and this has broad and deep implications for economics. Two areas of research, discussed below, relate the ultimatum game structure and results to economics.

Game Theory

The use of game theory within the social and natural sciences is now widespread. Indeed, the 1994 Nobel Prize for Economics was awarded to John Nash, Reinhardt Selten, and John Harsanyi for their work in game theory.

Game theory deals with situations where the outcome of one party depends both on that party's own behavior and on that of another party. Thus, the province of game theory includes such diverse topics as employer and employee relations, predator and prey interactions, and global political situations. Whenever there are strategic interactions between parties, the analysis should include game theory.

The ultimatum game is one of the simplest game-theoretic situations. There is just a single proposal; the game has only two players; and each player has only one decision. Furthermore, the proposer and the responder can communicate only through the offer of the proposer. In contrast to the simple structure of the ultimatum game, most situations outside the laboratory have many parties, myriad terms, long histories, and undefined sets of choices for the parties.

For example, consider the relationship between a farmer and the people who pick the farmer's crops. The two parties negotiate over pay, hours, health insurance, unemployment compensation, and so on. The individuals in this strategic situation will often be involved with each other for decades. The

tactics range from cool-headed financial proposals to strikes, boycotts, lockouts, and violence.

The ultimatum game provides a check of the predictive and normative power of game theory in an extremely simplified setting. If game theory fails in this setting, its applicability to more complex situations is questionable.

Self-interest in Economics

In addition to game-theoretic interactions between parties, the ultimatum game involves the question of basic human motivation. Economic theory is predicated on the notion that individuals pursue self-interest. What is self-interest, and how does it relate to the ultimatum game? In answering these questions, there is a division between theoretical and applied economists.

When economic theorists define self-interest, the term has little predictive power. For example, a person who commits suicide could be maximizing utility and pursuing self-interest. Or a person may maximize utility by giving money away to others. Thus, an economic theorist may see self-interest in many acts that most lay people would call altruism or even self-destruction.

An example may clarify the economist's view of altruism. Consider a typical day of Mother Theresa, in which she tended to the needs of others. Was Mother Theresa an altruist? To most people the answer is 'yes', but to economists the answer is 'no'. Economics assumes that all people ruthlessly and efficiently seek happiness. Any behavior that a person performs is assumed to make that person happier than alternative, feasible actions. In this view, when Mother Theresa nursed a sick person she was just as self-interested as a landlord who evicts a tenant for not paying rent.

As this example demonstrates, the economic assumption of self-interest cannot be contradicted by behavior. Whatever someone does is assumed to be motivated by self-interest. The term 'revealed preference' is used to describe this view of human behavior. People have preferences, and seek to be as happy as possible given those preferences and the choices they face. An outsider can understand motivation only by observing behavior, which reveals the underlying preferences.

While economic theorists take a very general definition of self-interest, economic practitioners are more specific. For economists who make predictions in the world, and those who make normative judgments on policy matters, self-interest means consuming goods and leisure. In this materially self-interested version of economics, Mother Theresa would be considered an altruist, but altruists are assumed not to exist.

The ultimatum game results discussed below suggest that human motivations are more subtle than is assumed by this second, more commonly used, definition of economic self-interest. In a variety of settings, people show deep concern for the impact of their actions on others. Sometimes people act altruistically to help others, but they also show a spiteful willingness to damage others. Both altruism and spite are inconsistent with the standard economic model of human nature. The implications of the divergence between the economic theory of human nature and actual human nature are profound.

PREDICTED AND ACTUAL BEHAVIOR IN THE BASIC ULTIMATUM GAME

Predicted Behavior in the Ultimatum Game

Standard economic theory makes a precise prediction of behavior in the ultimatum game (Stahl, 1972; Rubinstein, 1982). The standard analysis is based on the game-theoretical concept of 'sub-game perfection' (Selten, 1975).

To understand the standard analysis, first consider the position of the responder deciding whether to accept the proposer's offer. If the responder rejects the offer, the responder will receive zero and the game will end. The standard assumption of economics is that a responder will care only about the money that he or she will receive. The choice of an ultimatum game responder is thus clear. Rejection results in zero money for the responder while acceptance means the responder will take home whatever amount has been offered. A responder who behaves as predicted by standard economic theory will accept any offer greater than zero.

Consider an ultimatum game played for $10. Standard economic theory predicts that the amount the proposer proposes to take will have no effect on the responder's behavior. So at an extreme in the $10 game, responders are predicted to accept one cent and allow the proposer to take home the rest.

Now that the prediction for the responder's behavior is clear, consider the behavior of the proposer. The outcome of any proposal depends on the responder's behavior; thus the proposer must have an implicit or explicit model of the responder. If the responder will behave as predicted by economic theory, the responder will accept any offer that is more than zero. If the proposer assumes that

the responder will act as predicted by economic theory, and if the proposer cares only about his or her own monetary pay-off, then the offer should be the smallest possible positive amount.

In other words, a selfish proposer who anticipates a selfish responder proposes to take the whole financial pie. The selfish responder faced with a crumb or nothing, accepts the crumb. Returning to our $10 ultimatum game, the standard economic prediction is that proposers will offer the smallest possible positive amount and all offers will be accepted by responders. If the proposals are constrained to whole numbers of dollars, the prediction for a $10 game is an offer of $1 to the responder and acceptance of the offer.

Actual Behavior in the Ultimatum Game

The first experimental ultimatum game study was conducted by Guth *et al*. (1982). They had people play the ultimatum game for 10 deutschmarks. The money was actually paid according to their choices.

Guth *et al*. found that actual human behavior differed from the economist's prediction of behavior in two significant ways. Firstly, 20 percent of the responders rejected offers. (Recall that self-interested responders are predicted to accept all offers greater than zero.) Secondly, the average amount offered to responders was 30 percent of the total. Economic theory predicts offers close to zero, so the human proposers in this study were significantly more generous than predicted.

These data present a challenge, for the reasons discussed above. Firstly, if economics generally assumes material self-interest, why are responders walking away from money? Secondly, if game theory fails to predict behavior in this setting, how applicable is it to more complex situations? Finally, game theory fails in both its predictive role and its normative role. The behavior of both proposers and responders differs from the prediction. Furthermore, proposers who follow the theoretical advice make less money than proposers who ignore this advice.

Because of these important implications, this landmark study led to a series of related experiments that continues to this day. We summarize below some of the experimental evidence.

Replication of Guth et al.

The original data have been widely replicated using subjects drawn from industrialized countries (often college students). Summarizing across many experiments: (1) the most common proposal is an equal split, (2) the average amount offered to responders is between 40 and 50 percent, and (3) responders are more likely to reject proposals that give them a smaller percentage of the money, and frequently reject proposals that give them less than 20 percent. (See Camerer and Thaler (1995) and Roth (1995) for excellent summaries.)

Economic theory predicts an extreme division, with proposers taking almost all of the money. As noted, the actual outcome is always closer to an equal split than this prediction. However, the proposer very rarely ends up with less than half the money, so there is a substantial advantage to being the proposer (Bolton, 1991).

The effect of stake size

One common reaction to the initial study of Guth *et al*. was that the results were due to the relatively small amount of money at stake. Defenders of the standard economic theory predicted that the deviations from theory would disappear as the stakes became larger. And in their paper, Guth *et al*. suggested that an important test of their results would be to raise the stakes substantially to the level of 100 deutschmarks.

Hoffman *et al*. (1996a) had 50 pairs of subjects play the ultimatum game for $100 (all pairs were paid). The behavior of people in this $100 game was compared with that of previous participants in a $10 game. There was no significant difference between the two games. In the $100 game, three out of four offers of $10 were rejected, and two out of five offers of $30 were rejected.

Hoffman *et al*. did find, however, that there was a difference between ultimatum games played for no money (where people are just asked for their decisions, without any monetary consequences) and games for $10 or $100. Specifically, they found that in games for no money the rejection rate was 17 percent, and this dropped, with statistical significance, to 4 percent for games with stakes of $10. So in this study, money did matter, and increasing the stakes from zero to $10 did induce behavior closer to the economic prediction. The original findings of Guth *et al*. were, however, replicated, and increasing the stakes from $10 to $100 did not change behavior towards that predicted by economic theory.

Another approach to raising the effective value of the stakes was taken by Cameron (1999), who travelled to Indonesia. Because of the lower average income there, this study was able to use stakes equal to three months' worth of an average person's salary. The result was that responders were more willing to accept smaller percentages of the

total as the stakes increased. However, proposer behavior did not change with the increase in stakes, and behavior did not converge to the prediction of economic theory.

The Indonesian proposer behavior highlights an important theme in ultimatum games. Recall the two most salient differences between actual and predicted behavior: 'generous' proposers offer more than predicted, and 'proud' responders reject low offers when theory predicts invariable acceptance.

The initial reaction to these results focused on the generosity of proposers. Given that responders reject small offers, however, proposers' generosity redounds to the benefit of both parties. Similarly, when the stakes are increased a proposer risks more financial loss, so proposers might temper their demands out of aversion to risk rather than altruism. The finding that proposers do not alter their behavior as the stakes increase, even though responders become more willing to accept low offers, is consistent with this assessment.

So the deviation from the economic prediction exists when the stakes are raised to $100 in the US and three months' salary in Indonesia. What about even higher stakes? Even in poor countries, it is difficult to raise the funds for experiments with extremely high stakes.

Some evidence from outside the laboratory bears on the question of extremely high stakes. Cohen *et al*. (1996) suggest that potentially violent arguments between strangers are characterized by high costs of disagreement and no material costs of concession and withdrawal. They write, 'approximately 20 000–25 000 Americans will die in homicides this year, and tens of thousands more will be injured in stabbings or gunfights that could have ended in death'. In these interactions, people are willing to bear even the chance of death rather than accept an outcome they deem unfair or humiliating.

Cultural variation

While many of the ultimatum game experiments have been conducted in the US and Western Europe, several studies have been performed to examine how play varies across cultures. The general conclusion is that actual human behavior in the ultimatum game does not conform to the standard economic prediction of play in any culture studied.

Roth *et al*. (1991) ran identically structured ultimatum games in Israel, Slovenia (in Ljubljana, which was then part of Yugoslavia), the US, and Japan. The experiment was carefully constructed to reveal cultural differences. For example, the instructions were all translated from the same script, the different experimenters all had the same training, and the stakes were set to have equivalent purchasing power (slightly above $10 in the US). In contrast to many other experiments, this study had each participant play ten games, each with a different counterpart.

In all four countries, the most common offer ranged between 40 and 50 percent of the stake. There were cultural differences in both responder and proposer behavior. Israeli responders were more willing to take small offers, and Israeli proposers did indeed make smaller offers. In the first round of play, the most common offer was 50 percent in all four countries. By the tenth round of play, the most common offer in Israel was 40 percent, the most common offer in the US and Slovenia remained at 50 percent, and the most common offers in Japan were 45 percent and 40 percent. These differences were statistically significant.

In summary, this first cross-cultural study found differences in ultimatum game play between cultures, but behavior in all cultures was similar to the original findings and different from that predicted by standard economic theory.

Henrich (2000) went further and took the ultimatum game to the Machiguenga of South America, who subsist on hunting, fishing, gathering, and small-scale horticulture. Then, together with other researchers, he studied ultimatum game behavior in 15 small-scale societies around the world (Henrich *et al*., 2001).

The ultimatum game results from these 15 small-scale societies confirm that actual human behavior in the ultimatum game deviates from the prediction of the standard economic model. The authors write that the standard model 'is not supported in *any* society studied' (emphasis added). The lowest average offer was among the Machiguenga, where it was 26 percent, and this is significantly higher than the standard economic prediction of zero.

The small-scale societies exhibited two behaviors not previously documented. Firstly, participants from the Achuar, Ache, and Tsimané accepted 100 percent of the offers. In the case of the Achuar, this may be because there was only one offer of under 20 percent. However, all five offers of under 20 percent were accepted among the Tsimané, as were all eight such offers among the Ache. Secondly, among the Au and Gnau cultures, offers were frequently for more than half, and these offers were frequently rejected. Among western populations, offers above 50 percent are very rare, and larger offers are more likely to be accepted. The authors explain that the Au and Gnau are

gift-giving cultures in which acceptance of gifts requires reciprocation.

Informational treatment

Kagel *et al*. (1996) investigated fairness by running ultimatum games and manipulating the stakes and what the proposer and the responder knew about the stakes. (For related experiments, see also Prasnikar and Roth (1992), Mitzkewitz and Nagel (1993), and Croson (1996).) Specifically, they used an ultimatum game played for 100 chips, each worth 30 cents or 10 cents, and varied which players know the value of the chips.

In the basic version, chips are worth the same to both players and the values are common knowledge. This is a standard ultimatum game, and Kagel *et al*. replicate other studies with the most common offer being 50 percent. In one variant, chips are worth 30 cents to proposers and 10 cents to responders, and this is known only to the proposers. In this setting, proposers can achieve equal monetary outcomes by offering 75 chips. What the proposers actually do, however, is to exploit their informational advantage by offering close to 50. Responders generally accept these 50–50 splits in chip terms and (unknowingly) give 75 percent of the money to the proposer.

The outcome differs, however, when the game is played with the same pay-off conditions (30 cents' value to the proposer and 10 cents' value to the responder), but where the responder knows about the asymmetry in value. In this case, 50–50 splits of chips tend to be rejected, and over time proposers increase their offers.

Context-dependent effects

Hoffman *et al*. (1994) take issue with the notion that 'fairness' is the source of the observed behavior in the ultimatum game. They designed experiments in which the right to be the proposer in an ultimatum game is earned by correctly answering questions on a trivia quiz. When the right to propose is earned – and the instructions reinforce this 'property right' aspect – ultimatum game behavior is significantly more self-regarding. Specifically, half the proposers offer equal splits when the right is assigned randomly, versus one in ten when the right is earned.

RELATED GAMES

Multistage Games

Researchers have examined the motivation for rejections in the ultimatum game. In particular, why do responders reject positive amounts and so end up with zero? Some insight comes from looking at slightly more complicated multistage games. Recall that the ultimatum game derives its name from the finality of the first offer. In other games, the responder is allowed to make a counter-proposal from a smaller financial 'pie'.

One interesting feature was discovered in these multistage games with shrinking stakes. Responders often make counter-proposals that earn them less money than the offer they have just rejected. For example, a responder may reject \$4 out of \$10, then turn around and propose to take home \$3 out of \$5. Roth (1995) summarizes four independent studies, and finds that between 65 percent and 88 percent of the counter-proposals would result in less money for the responder than the just-rejected offer. This fact is strong evidence that many subjects care about more than just their own material pay-off.

Dictator Games

To explore the role of altruism in the ultimatum game, Kahneman *et al*. (1986) had subjects play a different game in which the proposers had no strategic motive to be generous. The dictator game has the same structure as the ultimatum game except that the second player does not have the right to refuse the division. The first player's decision is unilaterally imposed (hence the name 'dictator'). Because of this difference, the second player is called the 'recipient' in dictator games, as opposed to the 'responder' in ultimatum games.

Kahneman *et al*. used 161 dictators, of which 122 divided \$20 evenly, consistent with an altruistic interpretation. In this study, however, only 8 of the dictators were paid. Forsythe *et al*. (1994) ran dictator games in which all subjects were paid for stakes of \$5 and \$10. They found that 64 percent of \$5 dictators and 79 percent of \$10 dictators give at least \$1. The average amount given was approximately 25 percent. So dictators in these experiments were more generous than predicted by standard economic theory. These allocations are, however, less generous than those observed in comparable ultimatum games, so there is a strategic component in the behavior of ultimatum game proposers.

THEORETICAL EXPLANATIONS

Roth and Erev (1995) looked at the evolution of play in ultimatum games that continued for several rounds, as well as data from other repeated games. In the repeated versions of the ultimatum game,

their model has proposers learning to make generous offers because their less generous offers get rejected. The model also captures an important asymmetry in the ultimatum game: rejections cause larger losses for proposers than for responders. In this model of adaptive learning, behavioral adjustments happen faster in the context of large losses. Thus, proposers learn to moderate their demands faster than responders learn to accept.

This learning model provides insight into the dynamics of adjustment, but it is not designed to explain why play starts where it does. For such explanations, other models are required. One important approach involves what are called 'other-regarding' preferences. Economists use the term 'preferences' to describe an individual's likes and dislikes. The standard assumption about preferences is that people derive satisfaction only from their own lives and not from the lives of others. The ultimatum game results are inconsistent with these standard, materially self-interested, preferences.

While human behavior deviates in a significant and robust manner from that predicted by self-interested preferences, economic theory has no difficulty constructing preferences with an interpersonal component. These other-regarding preferences can be defined in a manner that is consistent with the ultimatum game.

Since the late 1980s several models of other-regarding preferences have been proposed. These models have inspired empirical tests, which have in turn stimulated new models and more experiments.

Bolton (1991) posits other-regarding preferences that include a term for relative consumption. In the context of two-person games, Bolton's model suggests that people care about both the amount of money they receive and the percentage of the total that they receive. The ultimatum game data are consistent with a utility function of this form. Specifically, a responder's dislike for low relative pay-off may outweigh the value of an offer. Granted that the responder is willing to reject, the offers of the proposer are understandable: proposers anticipate the possible rejection and temper their behavior accordingly.

In the light of Bolton's model, consider two games. The first is the standard ultimatum game, and the second is identical except that a computer generates the proposed split. In the second game, with a computer-generated proposal, the responder is presented with a 'take it or leave it' offer exactly as in the ultimatum game. The material repercussions for both individuals are identical regardless of the origin of the ultimatum.

What is the prediction for the new game? Bolton's model predicts that players will react identically to an offer generated by a human or a computer. Blount (1995) performed this experiment and found that, in the computer-generated game, rejection rates approach zero. In this case, then, materially self-interested preferences more accurately predict behavior than does Bolton's model.

As with Bolton's model, economic theory has no difficulty incorporating Blount's results into different specifications of other-regarding preferences.

One second-generation model of other-regarding preferences modifies the utility function to include a term for 'fairness' (Rabin, 1993). If a player believes his or her counterpart is being fair (this is defined precisely), then the player's happiness increases as the counterpart's pay-off increases. Conversely, if a player believes his or her counterpart is being unfair, then the player's happiness decreases as the counterpart's pay-off increases.

In Rabin's model, rejections in the standard ultimatum game come from the responder perceiving the proposer as acting unfairly. In Blount's computer-generated game, there is no unfairness on the part of the proposer, hence no reason to reject.

The narrow self-interest model, Bolton's model, and Rabin's model share the feature that every person has preferences of the same form. There are no spiteful or altruistic people in these models. Rather, circumstances, and perceptions of circumstances, induce people to act in ways that can be viewed as selfish, spiteful, or altruistic.

In contrast to these universal preferences, Levine (1998) assumes that there are different types of people in the world. Some individuals derive satisfaction from their counterpart's pain and others from their counterpart's pleasure. Furthermore, Levine assumes that players' behavior is conditioned by their expectation of the counterpart's type. People who delight in the counterparts' consumption might, nevertheless, dislike rewarding selfish types. Although the preferences differ between individuals, Levine's model constrains the distribution of spiteful and altruistic types to be constant across all situations.

Levine's model is consistent with the ultimatum game results. In this model, two distinct groups of responders will reject unequal ultimatum splits. The first group contains individuals who derive satisfaction from others' pain. They reject because

the loss to them through their own pay-off is more than compensated for by the pleasure in hurting the proposers. The second group of rejecters is using the offer as a signal of the proposer's type. Proposers who make unequal splits reveal themselves to be people who ought to be punished. In Blount's computer-generated offer game, this second group does not exist, hence the lower rejection rates.

So both the Levine and Rabin models are consistent with both standard and computer-generated ultimatum game behavior. Preferences of these types can also be constructed that are consistent with the observed generosity in the dictator game. However, several variants of the dictator game have been performed experimentally with results that are not consistent with these models.

Hoffman *et al*. (1996b) ran variants of the dictator game that systematically varied the 'social isolation' of the dictators. In the most private treatment, neither the experimenter nor the recipient could identify the behavior of any individual dictator. Recall that dictators who make their allocations known to the experimenter tend to give over 20 percent. In the private treatment, almost two-thirds of dictators kept the entire stake, and the average allocation was $0.80 out of $10.

In summary, dictators' giving is affected by their social isolation, and dictators who make decisions under private conditions keep almost all the money. This result has been replicated by Burnham (2002). Burnham found further that dictators are more generous when their actions are less private. Specifically, dictators give more when they see a photograph of their recipient, or when the recipient sees a photograph of them.

Let us return to the models of other-regarding preferences. Neither the Levine nor the Rabin model predicts any change in dictator behavior as a function of social isolation or anonymity. A new generation of theories has been developed which are more consistent with the existing data (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Fehr and Gachter, 2000). Empirical tests of these theories are under way.

References

Blount S (1995) When social outcomes aren't fair: the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* **63**(2): 131–144.

Bolton G (1991) A comparative model of bargaining: theory and evidence. *American Economic Review* **81**: 1096–1136.

Bolton G and Ockenfels A (2000) ERC: a theory of equity, reciprocity and competition. *American Economic Review* **90**: 166–193.

Burnham T (2002) A theoretical and experimental investigation of anonymity and gift giving. *Journal of Economic Behavior and Organization* **50**: in press.

Camerer C and Thaler R (1995) Ultimatums, dictators and manners. *Journal of Economic Perspectives* **9**: 209.

Cameron L (1999) Raising the stakes in the ultimatum game: experimental evidence from Indonesia. *Economic Inquiry* **37**(1): 47–59.

Cohen D, Nisbett R, Bowdle B and Schwarz N (1996) Insult, aggression, and the Southern culture of honor: an 'experimental ethnography'. *Journal of Personality and Social Psychology* **70**: 945–960.

Croson R (1996) Information in ultimatum games: an experimental study. *Journal of Economic Behavior and Organization* **30**: 197–212.

Fehr E and Gachter S (2000) Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives* **14**(3): 159–181.

Fehr E and Schmidt K (1999) A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* **114**: 817–868.

Forsythe R, Horowitz J, Savin N and Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic Behavior* **6**: 347–369.

Guth W, Schmittberger R and Schwarze B (1982) An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* **3**(4): 367–388.

Henrich J (2000) Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* **90**: 973–979.

Henrich J, Boyd R, Bowles S *et al*. (2001) In search of *Homo economicus*: behavioral experiments in 15 small-scale societies. *American Economic Review* **91**: 73–78.

Hoffman E, McCabe K, Shachat K and Smith V (1994) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* **7**: 346–380.

Hoffman E, McCabe K and Smith V (1996a) On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* **25**: 289–301.

Hoffman E, McCabe K and Smith V (1996b) Social distance and other-regarding behavior in dictator games. *American Economic Review* **86**: 653.

Kagel J, Kim C and Moser D (1996) Ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior* **13**: 100–110.

Kahneman D, Knetsch J and Thaler R (1986) Fairness and the assumptions of economics. *Journal of Business* **59**(4): S285–S300.

Levine D (1998) Modelling altruism and spitefulness in experiments. *Review of Economic Dynamics* **1**: 593–622.

Mitzkewitz M and Nagel R (1993) Envy, greed and anticipation in ultimatum games with incomplete information: an experimental study. *International Journal of Game Theory* **22**: 171–198.

Prasnikar V and Roth A (1992) Considerations of fairness and strategy: experimental data from sequential games. *Quarterly Journal of Economics* **107**: 865–888.
Rabin M (1993) Incorporating fairness into game theory and econometrics. *American Economic Review* **83**: 1281–1303.
Roth A (1995) Bargaining experiments. In: Kagel J and Roth A (eds) *Handbook of Experimental Economics*, pp. 253–348. Princeton, NJ: Princeton University Press.
Roth A and Erev I (1995) Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* **8**: 164–212. [Special issue for the Nobel Symposium.]
Roth A, Prasnikar V, Okuno-Fujiwara M and Zamir S (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* **81**: 1068–1095.
Rubinstein A (1982) Perfect equilibrium in a bargaining model. *Econometrica* **50**: 97–109.
Selten R (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* **4**: 25–55.
Stahl I (1972) *Bargaining Theory*. Stockholm, Sweden: Economic Research Institute.

Further Reading

Cameron L (1999) Raising the stakes in the ultimatum game: experimental evidence from Indonesia. *Economic Inquiry* **37**(1): 47–59.
Forsythe R, Horowitz J, Savin N and Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic Behavior* **6**: 347–369.
Guth W, Schmittberger R and Schwarze B (1982) An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* **3**: 367–388.
Hoffman E, McCabe K and Smith V (1996) Social distance and other-regarding behavior in dictator games. *American Economic Review* **86**: 653.
Kahneman D, Knetsch J and Thaler R (1986) Fairness and the assumptions of economics. *Journal of Business* **59**: S285–S300.
Rabin M (1993) Incorporating fairness into game theory and econometrics. *American Economic Review* **83**: 1281–1303.

Games: Voluntary Contributions

Advanced article

R Mark Isaac, Florida State University, Tallahassee, Florida, USA

CONTENTS

Voluntary contributions are a common mechanism for individuals to provide public goods to an economy. Standard economic theories suggest problems with the use of voluntary contributions. However, both field data and data collected from controlled laboratory experiments suggest that this pessimism is only partially justified.

INTRODUCTION

A core proposition of neoclassical economic theory is that a competitive equilibrium of a market economy with individually maximizing, selfish agents achieves a desirable efficiency benchmark known as 'Pareto efficiency'. In other words, individually interested maximization is linked to social efficiency. (Briefly, an economy is Pareto-efficient if the only way to make one person better off is to make someone else worse off.) This proposition is often referred to as the 'first fundamental welfare theorem' in economics. However, the first welfare theorem does not hold in all cases. It may fail if there are spillover costs or benefits to economic agents not a part of the original transaction (pollution being the textbook example). And the link between equilibrium and optimality may fail if there are what economists call 'public goods' in the economy, particularly if the mechanism for producing the public goods is by voluntary contributions. In the voluntary contributions world, economic theory demonstrates that individually maximizing behavior is potentially decoupled from social efficiency. Furthermore, a long series of laboratory experiments in public goods decision-making has demonstrated the reality of the efficiency problems with the voluntary contributions mechanism for providing public goods. However, these same experiments have not only demonstrated but have also replicated behavior that some argue to be reflective of attempts by the subjects to substitute cooperative behavior for the inefficient individually maximizing social equilibrium.

MODELING THE PUBLIC GOODS PROBLEM

To set the stage more formally, we note that economists offer two boundary definitions of economic goods: purely public goods and purely private goods, with a wide variety of goods exhibiting intermediate characteristics. The purely private good is defined by the paired characteristics of ability to exclude nonpayers and nonrivalrous consumption. If X represents the total amount of a good available for consumption, and x_i represents the consumption of that good by a single person, then, at the extremes, a good in rivalrous consumption exhibits $X = \Sigma x_i$ while a good in nonrivalrous consumption exhibits $X = x_1 = x_2 = \ldots = x_N$. (Note that nonrivalrous consumption does not mean that each individual gives the same value to the good, but that each person can be represented as consuming the same amount.) Another way of looking at nonrivalry is to imagine a society that adds one new individual in a short run (no production) situation in which there is a fixed amount of the good. If that good has rivalrous consumption, the addition of a new person in society means that either that new person consumes nothing or a reallocation occurs, causing others to consume less. With nonrivalrous consumption, the new person consumes as much as everyone else, which is the same as everyone did before the newcomer's arrival. (Imagine adding one more person on a block to receive satellite TV; other current subscribers suffer no loss).

Obviously, not all goods are either perfectly public (perfectly exhibiting both characteristics) or perfectly private (exhibiting neither). Sometimes one of the characteristics can occur without the

other. For example, the Arizona-Sonora Desert Museum in Tucson (a private, nonprofit zoological institute) can exclude nonpayers but includes exhibits 'inside' that are, in some conditions, nonrivalrous. This is an example of a so-called club good (Buchanan, 1965). Similarly, a good may exhibit one or more of the characteristics of public goods under some circumstances but not under others. A university recreation center swimming pool may exhibit nonrivalry in consumption until the number of people wishing to swim exceeds the number of lanes. This is an example of a so-called crowding good.

The distinction between public and private goods is interesting to economists because of the issue of convergence of individual and group incentives. Competitive markets for private goods tend to coordinate individual incentives and aggregate efficiency. Public goods risk a disconnection between individual and group incentives. In general, several different mathematical representations describe some form of the following statement: in the presence of public goods, it is difficult or even impossible to find institutions and incentives such that each individual following his own interest will yield an outcome that is socially optimal (Samuelson, 1954; Walker, 1980).

One of the most common institutions for providing public goods is the voluntary contributions mechanism, in which N individuals each choose m_i amount of resources to contribute to a fund which will be used to purchase a given level Y of a public good. If the good is nonrivalrous in consumption, then a given level of Y will be consumed not just by one individual but by all N individuals. Increasing the provision of the public good by one unit will yield a vector of increases in well-being of the form $(v_1, v_2, v_3, \ldots, v_n)$. The problem, as modeled by economists, is that each individual bears the entire private cost of his marginal contribution, $c(m_i)$ but receives only his personal marginal benefit, v_i, ignoring the net benefits generated for others, v_{-i}. In addition, individuals who do not support the public good at an optimal level cannot be excluded for consumption. This inability of individual incentives to match social incentives, known as the 'free rider problem', results in a socially nonoptimal level of the provision of the public good. To embed this abstract problem in a more familiar framework, imagine a cul-de-sac of five houses being asked for contributions for mosquito spraying. Each individual homeowner, if making his decision as above, ignores the benefit to others. At the limit, any one homeowner can contribute nothing and yet not be excluded from the benefits of that spraying (which, due to the endemic free riding, cannot be funded).

Notice that this practical example is built on assumptions both about technology and about the legal system. A mosquito spraying device which allowed noncontributors' houses to be missed would dramatically change the public goods nature of the spraying, as would a legal system which provided contributors with a legal right to tax noncontributors (as might be found in many so-called homeowners' associations).

EARLY EXPERIMENTS IN PUBLIC GOODS

The public goods problem has been a part of classical microeconomics training for decades (see Samuelson, 1954 and 1955, citing and referring to Lindahl and Musgrave among others). However, it was in other social science disciplines with older traditions of experimentation that data on public goods environments were first collected. In this work, a pattern emerged of cooperation in providing the public good at levels greater than what might have been expected from standard economics models. Unfortunately, many of these experiments did not control for the incentive structures of most interest to economists, or had other design features that limited their acceptance by economists (this work includes Marwell and Ames, 1979; see Kim and Walker, 1984 and Isaac *et al*., 1985 for critiques and discussion).

In the late 1970s, economists began laboratory experimental tests of behavior in public goods situations. Smith (1980) worked on the design of a process that sought to achieve efficient outcomes in a public goods environment. Kim and Walker, and Isaac *et al*. (1985), each conducted new public goods experiments providing real economic incentives and altering design features to which economists had objected in earlier work by noneconomists. Both research groups created experimental environments in which there was a clear distinction between private incentives to underprovide the public goods (to free ride) and group optimal behavior. Both groups used a variant of the voluntary contributions mechanism, where the only resources available for the provision of the public good are those voluntarily transferred by the economic agents. The results were both surprising and not surprising. The results were not surprising in that conditions could be obtained in which the level of free riding was substantially closer to economic predictions than had been observed before. The results were surprising in that,

even when utilizing a design that addressed economists' concerns about previous work, unpredicted cooperation still occurred and, in some settings, was still economically significant.

Isaac *et al*. (1984) and Isaac and Walker (1988) began the first systematic investigation as to why there was this divergence in results. After almost 20 years, the twin pillars of the original results remain. The prediction that individuals by and large ignore individual incentives to free ride and instead adopt socially maximizing strategies is wrong. Just as we observe that communally provided day use bicycles seem to disappear, even in politically correct university towns, so, too, can we observe free riding behavior in the laboratory. How severe the free riding is, how far the outcome is from the social optimum, varies. This will be discussed in more detail below. Equally, a prediction that individuals by and large make free riding decisions based solely upon an individual maximization model is also wrong. Just as we observe significant charitable contributions of both money and time in our society, so, too, we observe levels of public goods provision in experiments that are significantly above the free-riding predictions of economists. The task for the last 20 years or so is still the task today: what conditions, environmental or institutional, lead people to behave more like individual maximizers, and which conditions lead people to behave more like social maximizers? In attempting this analysis, I must return a favor and acknowledge John Ledyard, whose early survey on public goods (1995) is still a standard reference.

Hence forth, I shall assume, unless otherwise stated, that we are considering a standard, linear, voluntary contributions public goods model of the following form: there are N individuals, each individual in possession of z_i units of wealth (tokens). The simultaneous decision facing each individual is how to 'invest' each token. A token may be invested in an individual exchange where it earns p_i (assume that $p_i = 1$ cent for all i) or it may be invested in a group exchange. Letting m_i represent individual i's investment in the group exchange, every individual, regardless of their personal m_i, earns the following from the group exchange: $1/NxGx\ m_i$ cents – this is the linear part of the model. (Changing the linear nature of the return from the public good will change some of the details of the discussion, but not the basic points.) For some examples, see Bagnoli and Mckee (1991), Harrison and Hirshleifer (1989), Isaac *et al*. (1989).

As long as $G/N < 1$, each individual has a dominant strategy to invest all tokens in the group exchange (to free ride), but if $G > 1$ the social optimum is to invest all the group tokens in the group exchange. Furthermore, if the allocation of tokens is sufficiently egalitarian, every individual is better off at the group optimum than at the free-riding equilibrium. In typical experiments, although not in every one, the task is repeated by the same group a known, finite number of times. (And, for those not familiar with the protocols of experimental economics, the earnings from the public good sessions are translated into salient economic rewards.)

DO INCENTIVES AND STRUCTURES MATTER?

A variable or treatment might matter in the sense that a change will alter the theoretical prediction; it might matter in the sense that we observe changes in actual individual behavior. These two may go together, but they need not. The role of individual versus group incentives is a factor that provides a perfect example of the difference in these two definitions. In our standard model, the dominant strategy prediction holds for any values of G, N, and p_i such that $G/N < p_i$. Isaac *et al*. (1984) defined $(G/N)/p_i$ (or G/N when $p_i = 1$ as I will continue to assume) to be the 'marginal per capita return' (MPCR hereafter). As long as the MPCR is less than 1 an individual's single period (noncooperative) dominant strategy is to free ride. Why? Because regardless of the behavior of others, at the margin reducing contributions by 1 token gains 1 cent in earnings in the individual exchange and costs G/N in the group exchange. (If G is not a constant then one looks at $G'(m)/N$). If MPCR > 1, an individual's dominant strategy is to contribute fully to the group good (at MPCR = 1, the individual is indifferent). Thus, there is no predicted difference when MPCR equals .3 as opposed to when it equals .75. A more complicated argument demonstrates that free riding is predicted as a Nash Equilibrium across multiple finite periods whenever the MPCR is less than 1.

The behavioral consequence of the change in MPCR is, however, much different. For groups of 4 and 10, increasing MPCR from .3 to .75 increases contributions to the group good. These effects have been replicated (Isaac *et al*., 1994). The observed MPCR effect in experimental public goods environments is one of the key behavioral anomalies with respect to standard microeconomic theory. It is not the only such anomaly.

There is a well-known conjecture in economics, notably associated with Olson (1965) and Buchanan (1968), that group size (the number of people in the

group) affects the ability of groups to provide public goods; specifically, large groups are thought to have a harder time providing public goods than smaller groups. What does our model say about that? In the basic model, N has an effect only when it lowers the marginal per capita return (G/N); there is no other effect of group size in the model. Isaac and Walker found that the behavioral effect of an increase in N acting to reduce MPCR is completely consistent with the standard conjecture. Reducing G/N by increasing N reduces contributions to the public good. However, when the independent effect of N is sought, there was a surprise in that data. There was no statistically significant independent effect of N (as consistent with the basic model), but the direction of the effect was, surprisingly, positive (that is, holding G/N constant; the qualitative effect of increasing N on contributions to the group good was positive). Isaac *et al*. (1994) were able to pursue these conjectures with more observations and with larger groups. Their data supported the following conclusions about group size. First, they found that, controlling for MPCR, large groups of 40 and 100 were more successful in providing public goods than small groups (4 or 10). In fact, these larger groups mean provision level was about 40 percent of optimum, typically not dipping below 30 percent even in the last period. Second, they found that for the larger groups the change from an MPCR of 0.30 to an MPCR of 0.75 made little difference, in stark contrast to smaller groups. However, even in a group of 40 the MPCR effect could be recaptured when the experimental design was altered to examine MPCR equal to 0.03

DOES LEARNING MATTER?

Perhaps the surprising level of individual decisions which involve 'overcontribution' to the public good (relative to the standard equilibrium theory) simply reflects individuals learning either the single-period (stage game) dominant strategy or the multiple-period Nash equilibrium version of free riding. In a classroom, teaching undergraduates the dominant strategy nature of the single-period decision is relatively straightforward. However, a common report of the same students (and of nonstudents confronted with the same task) is that in a multiple-period setting: 'I am trying to increase contributions by others in the future by my own positive contributions now.' In addition to the fact that this flies in the face of the Nash equilibrium built upon backward induction from the dominant strategy in the final period, it ignores other practical complications such as: 'Just because you raise your contribution in the period, how do you know that this will be reflected in an aggregate increase in contributions?' (It may not.) These arguments may not be transparent, especially to novices in the experimental task. However, this common self-description will be revisited in the discussion on reciprocity below.

From the very earliest experiments, there was evidence to support the idea that individuals were learning how to free ride, so to speak. Contributions to the public good generally decayed across time, and individuals returning for a second time seemed to free ride more than subjects making these decisions for the first time. There have been several attempts to disentangle how many of the equilibrium violations may be do to learning. Palfrey and Prisbrey (1997) pointed out that many of the early wave of public goods experiments used an experimental design in which the (dominant strategy and/or Nash) equilibria were at a boundary of the message space – namely, zero contribution to the public good. This meant that, potentially, what had been observed as theory violating contributions to the public good were merely the censored (one-sided) errors of convergence to the standard theory (errors in convergence had to be above the prediction by definition). They constructed a design in which MPCRs differed across individuals and across periods. They concluded that random error played a significant role in explaining positive contributions to public goods in standard experiments. Several other researchers addressed this same question by using an experimental design in which there was a suboptimal interior Nash equilibrium of contributions to the public good (see Andreoni (1993); Isaac and Walker (1998); Keser (1996); Sefton and Steinberg (1996), and others in the Laury and Holt (1998) review). Laury and Holt concluded that 'moving the equilibrium away from the boundary is not sufficient in itself to induce Nash behavior in public goods experiments.'

Finally, on a different front, Houser and Kurzban (2000) used computerized, nonreciprocal partners from which to measure a baseline in confusion in their experiments, finding that it explains about half of the observed 'cooperation'.

Pretty clearly there are at least two things going on here. First, there is some recoverable effect of censored errors in a learning process that may be mistaken for intentional contributions to the public good. Second, there remains a significant amount of that effect in standard experimental designs and in designs in which censoring is not an issue.

DO INDIVIDUAL PREFERENCES MATTER?

Perhaps the unpredicted tendency towards provision of public goods represents incorrect modeling of individuals' supposedly selfish preferences, a phenomenon whose general rubric is often called 'altruism'. Altruism would seem to be high on the list of topics of mutual interest for economists and cognitive scientists. However, this must be approached with some care. First, in the literature of the economics of public goods, altruism refers narrowly to a failure of the assumption that individuals are entirely self-interested. Mutually cooperative behavior based upon self-interested individuals' recognition of the benefits of such behavior is more commonly called 'reciprocity'.

An additional caveat with the concept of altruism is that it has different manifestations. 'Pure' altruism is usually characterized as the phenomenon in which John cares directly about Mary's well-being, variously defined as Mary's utility, income, wealth, or consumption of some commodity. 'Warm glow' altruism (Andreoni, 1995) refers to the idea that John gets utility (a 'warm glow') out of the act of contributing to Mary's well-being, or, perhaps, out of being seen by others to be a generous person, in this way also contributing to Mary's well-being. These are not observationally equivalent. Beyond these two lie numerous other concepts of preferences that, while clearly not selfish, may or may not be distinguishable from altruism; some examples are preferences for fairness, justice, or even duty.

There are other interesting paradoxes regarding altruism. For example, altruism may be significantly dependent upon the institutional context in which people find themselves. I certainly have not seen much evidence in market experiments, as buyers and sellers face off each other for a penny or two of trading profits, of interpersonal overlap in utility functions. Further, in markets with opportunities for collusive behavior by some participants, experimental economists typically observe at least some periods of successful collusion even though this collusion public good clearly makes some people in the room (those on the other side of the market) worse off. Where is the altruism? Likewise, altruistic behavior may be very dependent upon the social context. Would you be an income altruist if the other three people in your public goods experiment were Bill Gates, Rupert Murdoch, and Oprah Winfrey? What if the other three people were young residents of a shelter for battered teens?

There have been numerous attempts, using clever experimental designs, to disentangle how much of the replicable pattern of positive contributions to public goods is due to 'errors and learning' (as above) or to altruism, or to 'warm glow' effects, or to fairness or even to spite, and so forth. The reader is referred to the following representative papers: Anderson *et al*. (1998), Andreoni (1988), Carter *et al*. (1992), Goeree *et al*. (2000), Offerman *et al*. (1996), Palfrey and Prisbrey (1997), Saijo and Nakamura (1995), and with a particularly complete literature review, Croson (1998).

DOES THE WAY WE LOOK AT THE PROBLEM MATTER? OR WHAT IS THE ROLE OF RECIPROCITY?

I find each of the above families of explanations incomplete in order to explain why experimental subjects provide public goods at levels much higher than predicted by standard theories. (It would be worth noting at this point that while we cannot know for sure the equivalent parameters of these experiments in the field, it is clearly the case that the naturally occurring outcome is not zero contributions. Massive amounts of resources are contributed by individuals, using just the United States as an example, to goods that transparently have attributes of publicness and some inherent dangers for free riding.)

While error processes and altruism probably influence individual decisions about public goods provision, I believe that there is something else going on. I argue that this is that individuals are simply not looking at the public goods provision problem in entirely the same way that we economists model it. I use the word 'entirely' deliberately, because I conjecture that in the naturally occurring world, as in the laboratory, free riding does occur; outcomes are neither automatically nor uniformly socially optimal, but public goods are provided. Such a paradigm shift ought to be intriguing to cognitive scientists.

I believe that the vast majority of people who confront a public goods situation recognize, even if in some nontechnical sense, that everyone will be better off if the public good is provided. Furthermore, most believe that there is some individual benefit to be gained from everyone trying to make cooperation successful, even when direct communication is not possible. This concept has several different faces, but it goes by the generic name of 'reciprocity'.

There are a numerous different versions of reciprocity that have been considered. The following are presented in approximately chronological order. Auster (1983) based a theory of 'closed anarchy' upon the hypothesis that each individual in a public goods situation has some expectations about how his contributions will affect the decisions of others, and that this relationship is not the trivial null relationship. Rather, he posits that individuals typically have an expectation that they can lead others. Auster's formulation is very explicit that each person has an expectation belief $E = g(e)$, where e is one's own contribution decision and E is the total of others' contributions.

Sugden (1984) developed several versions of reciprocity. He refers to his 'principle of reciprocity' as Kantian, and it might be also be called 'do the right thing' reciprocity. Sugden states, in his introduction, that he is proposing a theory of the voluntary sector, 'based on the assumption that most people believe that free riding to be morally wrong.' His principle of reciprocity is that individuals should not free ride relative to the level of contributions others are making. Sugden does derive equilibria from this new principle, but for me the most important contribution is not the equilibrium mathematics but the idea that an external moral imperative changes the way we look at the problem. Presumably, if we are looking at the problem differently, it is possible that our calculation, learning, and judgment skills are being applied to the new problem. The cognitive question to which we may be responding may no longer be 'How do I noncooperatively make the most return' but 'How do I make sure that we are all made better off?', or 'How do I fulfill my moral obligations?'

A third concept of reciprocity is what Sugden calls his 'principle of unconditional commitment' and what others might call the 'golden rule': contribute to the public good as you would like others to do so. I believe the drift into the vocabulary of theology is both explicable and meaningful, but that is for another article.

A different definition is that of 'trust reciprocity' (Berg *et al*., 1994). Their concept of reciprocity is distinct, both in theory and in evidence, from what they call 'benevolence'. The key to their concept of reciprocity is trust. Individuals who seem to trust one another irrationally end up with higher payments than those who perfectly follow a backwards induction equilibrium. The shorthand application to public goods is that individuals ignore the backwards induction Nash equilibrium: I contribute today trusting that you will see my contribution and contribute tomorrow.

Isaac *et al*. (1994) propose a 'forward looking' model of individual decisions. Their model has three components. First, each individual believes that his decisions have some signaling content: that is, actions today will have an influence on the actions of others in the future. Second, each individual has a definition for success, say in achieving a cooperative outcome. Third, each individual has some expectations about his signals achieving that success. Isaac *et al*. relate this model to the outcomes of a series of public goods experiments which will be discussed below.

Finally, McCabe and Smith (1999), as elaborated specifically for public goods in Gunnthorsdottir *et al*. (2000), propose a 'goodwill accounting' framework. In their model, the standard public goods model is altered by multiplying the benefits side by a function $P_i(\alpha_{it})$, where the α_{it} is a measure at time t of individual i's expectations about contributions forthcoming from the rest of the group.

Although these models of reciprocity differ, what is striking is that they uniformly posit a decision problem that is neither approached nor solved in the economist's standard paradigm of backward induction equilibrium rationality, built upon assumed dominant strategy free riding in the final period. In fact, from my informal discussions with colleagues, students, and subjects who have participated in public goods experiments, I almost always find that their description of how they approach the problem is typically in terms similar to at least one of these reciprocity models. Very few people say that this is an exercise of backward induction from a dominant strategy. The problem for evaluating this self-reporting is that we do not know whether the way that people talk about a problem has a systematic relationship with the decisions they make. There is a great deal of more formal experimental research that ought to speak to this hypothesis. One of the first direct tests of whether reciprocity might be making a difference was Andreoni's (1988) innovative research in which he altered the traditional iteration scheme so that the groups of individuals rotated in composition. He called the new treatment 'strangers' and hypothesized that the strangers treatment would eliminate the role of reciprocity. Unfortunately, this design has proven problematic. First, Andreoni's own data gave the opposite of the hypothesized effect: strangers were more cooperative. While some experimenters have been able to reproduce Andreoni's results, others have recovered the expected result that strangers free ride more (Croson, 1996; Andreoni and Croson, 1998).

Croson (1998, 2000) also reports direct tests of reciprocity. A standard experimental design is altered so that subjects receive specific information on the pattern of individual contributions to the public goods, allowing subjects to calculate not only the mean (as in traditional designs) but also the minimum, medium, and the maximum contribution in any round. Statistical analysis finds significant positive correlations of current contributions with the contributions of others for 21 of 24 subjects, broadly consistent with reciprocity. The most powerful past effect turns out to be the median of the previous periods contributions. Cox *et al*. (2001) have developed a 'triadic' approach to disentangling behavior consistent with reciprocity from that better explained by altruism. Their paper looks primarily at other game-theoretic environments. Whether their conclusions, generally non-supportive of models based upon altruism, would transfer to something that looked more like the voluntary contributions system is open for further examination.

Next, consider some of the anomalous results discussed earlier:

- Holding other design variables constant, the level of MPCR matters over regions where it should not, according to the standard theory (Isaac *et al*., 1984; Isaac and Walker, 1988).
- The range over which MPCR makes a difference in contributions appears to vary according to group size. For example, changing MPCR from .3 to .75 makes a significant difference for $N = 4$ or 10. However, for $N = 40$, that treatment does not yield a significant difference. Changing MPCR from .03 to .75, however, reproduces the original result (Isaac *et al*., 1994).
- Holding MPCR constant, changing group size in some cases shows the unexpected result that larger groups free ride less than smaller groups (Isaac *et al*., 1994).

Isaac *et al*. (1994) argue that while traditional models cannot explain these phenomena, their 'forward looking' models can. Specifically, in their models, both MPCR and N influence the expected success of contributions, signaling in a manner consistent with the anomalous results just described.

Gunnthorsdottir *et al*. (2000) conducted a standard linear public goods series and then separated the subjects into smaller groups in a later round of experiments. The treatment was whether the later separation would be random or sorted, whether the individuals were grouped together based upon their initial contribution. In the sorted rule, high investors were grouped with high investors, and so forth. They found that the MPCR effect (as reported above) was larger and more significant for subjects classified as cooperators than for subjects with low initial contributions. Furthermore, at each level of MPCR studied, 'cooperators' contributions in the sorted condition become higher than contributions in the random conditions no later than the fourth round, and remain higher through round ten.'

Finally, there are a few other indirect indications. Contributions in public good processes typically decline across time (which, by itself, argues against altruism as a single-dimension explanation for these phenomena). However, contributions very rarely converge to full free riding, even in a known (dominant strategy) end period. The residual, end-period contribution would not be explained by those reciprocity models relying upon influence on future contributions (Auster's 'closed anarchy,' Isaac, Walker, and Williams's 'forward looking models', or McCabe and Smith's 'goodwill accounting'). Those reciprocity concepts that could explain the residual would be Berg *et al*.'s 'trust reciprocity' and Sugden's three concepts.

In summary, the term 'reciprocity' is broad and includes numerous different formulations about how one person will react to the actual or perceived past or future actions of others. This has led to numerous different theoretical formulations and a similar number of experiments which attempt to compare their theoretical usefulness. It would be outside the scope of this article to survey all the results. However, an excellent discussion of experiments that attempt to discern among candidate models of reciprocity is provided by Croson (1998). Despite the differences in the model, a key similarity is that they posit that individuals look at the public goods world differently than the standard non-cooperative equilibrium formulation. By altering the assumption of how individuals look at the world, reciprocity models can provide predictions that, like actual observations of human decision-making, differ from than the usual pessimistic predictions.

SUMMARY AND CONCLUSIONS

After more than 20 years of experimental research regarding the voluntary provision of public goods there is overwhelming support that three overarching conclusions can be drawn:

1. Free riding is not a figment of economists' imagination. It can be observed in the laboratory under controlled conditions.

2. The extent of free riding is much less than is predicted by standard economic models.
3. Where a particular group falls in the continuum between complete free riding and the full social optimum may have some random components but it is not completely random. We can observe replicable effects of the economic environment, economic institutions, and the role of information.

After these general statements and the general replicable effects began to be identified, attention shifted somewhat to the 'why' questions. The three most popular answers each have some implication for the brain, mind, or cognition. Some individuals whose outlook is purely selfish, non-cooperative, and self-maximizing may nevertheless take some time to learn the dominant strategy, or to find the Nash equilibrium. Follow-on questions would include: how do people learn? What cues can accelerate or retard their learning? Other individuals may exhibit preferences (or preferencelike attributes) that are altruistic. Economists typically end their questioning at this level; however, there are other disciplines whose research will continue to ask: 'Why are a person's preferences the way they are?'

The most productive avenue for interchange between economists interested in public goods provision and cognitive scientists may lie in a third category, typically called 'reciprocity'. Reciprocity models, to varying degrees, suggest that potential contributors are looking at the problem in a different way. This statement is intended to transcend the laboratory results and extend to the field. Significant numbers of people do free ride in many naturally occurring circumstances: roadway litter and the failure of 'honors system' payment schemes are two examples, but group efforts to clean up a local park might bring out a significant number of residents. Why does this happen? And why in some circumstances, and not others? Why with some people and not others? Twenty years of laboratory experiments have shown that economic variables suggest as individual return and group size matter but in unexpected ways.

It should not be assumed that cognitive research questions are mutually exclusive among these categories of models explaining public goods provision. For example, consider the 'learning' explanation. Across time, individuals in economics experiments may learn the most starkly functional components of the decision process: how to transmit decisions, how the mathematical relationships (described initially in experimental instructions) actually work, what an 'end period' actually means. On the other hand, the learning may be about key components of one or more of the reciprocity models, especially those that include a parametric representation of each person's expectations about the others. In this world, individuals learn about the state of reciprocity in their community. And, of course, there is the fact that most naturally occurring public goods problems occur on a continuum between a purely one-shot decision and perfect iteration and replication. Furthermore, a group facing a public goods provision problem may be composed of individuals best described by different referenced models. An altruist may interact with a reciprocator. This becomes very interesting very quickly, as cognition applied to the problem of public goods requires a process not simply about the world but about interaction with other potentially dissimilar and complex cognitive beings in the context of specific institutional and social rules.

References

Anderson SP, Goeree JK, and Holt CA (1998) A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics* **70**: 297–323.

Andreoni J (1988) Why free ride?: strategies and learning in public goods experiments. *Journal of Public Economics* **37**: 291–304.

Andreoni J (1993) An experimental test of the public goods crowding out hypothesis. *American Economic Review* **83**: 1317–1327.

Andreoni J (1995) Warm glow versus cold prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* **110**: 1–21.

Andreoni J and Croson R (1998) Partners versus strangers: random rematching in public goods experiments. In: Plott CR and Smith VL (eds) *The Handbook of Experimental Economics Results*. New York: Elsevier.

Auster RD (1983) Implicit unanimous consent: the level of group goods under closed anarchy. Mimeo, University of Arizona.

Bagnoli M and McKee M (1991) Voluntary contribution games: efficient private provision of public goods. *Economic Inquiry* **29**: 351–366.

Berg J, Dickhaut J, and McCabe K (1994) Trust, reciprocity, and social norms. Mimeo, University of Iowa.

Buchanan JM (1965) An economic theory of clubs. *Economica* **32**: 1–14.

Buchanan JM (1968) *The Demand and Supply of Public Goods*. Chicago: Rand McNally and Company.

Carter JR, Drainville BJ, and Poulin RP (1992) A test for rational altruism in a public goods experiment. Mimeo, Worcester, MA: College of the Holy Cross.

Cox JC, Sadiraj K, and Sadiraj V (2001) Trust, fear, reciprocity, and altruism. Mimeo, University of Arizona.

Croson R (1996) Partners and strangers revisited. *Economics Letters* **53**: 25–32.
Croson R (1998) Theories of altruism and reciprocity: evidence from linear public goods games, Mimeo, Wharton School, University of Pennsylvania.
Croson R (2000) Feedback in voluntary contributions mechanisms. In: Isaac RM (ed.) *Research in Experimental Economics*, vol. 8. Amsterdam: JAI.
Goeree JK, Holt CA, and Laury SK (2002) Private costs and public benefits: unraveling the effects of altruism and noisy behavior. *Journal of Public Economics* **83**: 255–276.
Gunnthorsdottir A, Houser D, McCabe K, and Ameden H (2000) Excluding free-riders improves reciprocity and promotes the private provision of public goods. Mimeo, University of Arizona.
Harrison GW and Hirshleifer J (1989) An experimental evaluation of weakest link/best shot models of public goods. *Journal of Political Economy* **97**: 201–225.
Houser D and Kurzban R (2000) Revisiting confusion in public goods experiments. Mimeo, University of Arizona.
Isaac RM, McCue K, and Plott CR (1985) Public goods provision in an experimental environment. *Journal of Public Economics* **26**: 51–74.
Isaac RM, Schmidtz D, and Walker JM (1989) The assurance problem in a laboratory market. *Public Choice* **62**: 217–236.
Isaac RM and Walker JM (1988) Group size effects in public goods provision: the voluntary contribution mechanism. *Quarterly Journal of Economics* **103**: 179–199.
Isaac RM and Walker JM (1998) Nash as an organizing principle in the voluntary provision of public goods: experimental evidence. *Experimental Economics* **1**: 191–206.
Isaac RM, Walker JM, and Thomas S (1984) Divergent evidence on free riding: an experimental examination of some possible explanations. *Public Choice* **43**: 113–149.
Isaac RM, Walker JM, and Williams AW (1994) Group size and the voluntary provision of public goods: experimental evidence utilizing very large groups. *Journal of Public Economics* **54**: 1–36.
Keser C (1996) Voluntary contributions when partial contribution is a dominant strategy. *Economics Letters* **50**: 359–366.
Kim O and Walker M (1984) The free rider problem: experimental evidence. *Public Choice* **43**: 3–24.
Laury S and Holt CA (1998) Voluntary provision of public goods: experimental results with interior Nash equilibria. In: Plott CR and Smith VL (eds) *Handbook of Experimental Economics Results*. New York: Elsevier.
Ledyard J (1995) Public goods: a survey of experimental research. In: Kagel J and Roth A (eds) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
Marwell G and Ames R (1979) Experiments on the provision of public goods: resources, interest, group size, and the free rider problem. *American Journal of Sociology* **84**: 1335–1360.
McCabe KA and Smith VL (1999) Goodwill accounting in economic exchange. In: Gigerenzer G and Selten R (eds) *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.
Offerman T, Sonnemans J, and Schram A (1996) Value orientations, expectations, and voluntary contributions in public goods. *Economic Journal* **106**: 817–845.
Olson M (1965) *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.
Palfrey TR and Prisbrey J (1997) Anomalous behavior in public goods experiments: how much and why? *American Economic Review* **87**: 829–846.
Saijo T and Nakamura H (1995) The 'spite' dilemma in voluntary contribution mechanism experiments. *Journal of Conflict Resolution* **39**: 535–560.
Samuelson PA (1954) The pure theory of public expenditure. *The Review of Economics and Statistics* **36**: 387–389.
Samuelson PA (1955) A diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* **37**: 550–556.
Sefton M and Steinberg R (1996) Reward structures in public goods experiments. *Journal of Public Economics* **61**: 263–287.
Smith VL (1980) Experiments with a decentralized mechanism for public good decisions. *American Economic Review* **70**: 584–599.
Sugden R (1984) Reciprocity: the supply of public goods through voluntary contributions. *Economic Journal* **94**: 772–787.
Walker M (1980) On the non-existence of a dominant-strategy mechanism for making optimal public decisions. *Econometrica* **48**: 1521–1540.

Further Reading

Attiyeh G, Franciosi R, and Isaac RM (2000) Experiments with the pivot process for providing public goods. *Public Choice* **102**: 95–114.
Davis DD and Holt CA (1992) *Experimental Economics*. Princeton, NJ: Princeton University Press.
Smith VL (1982) Microeconomics systems as an experimental science. *American Economic Review* **72**: 923–955.
Varian HL (1999) *Intermediate Microeconomics*, 5th edn. New York: W.W. Norton.

Inducing Risk Preferences

Intermediate article

John Dickhaut, University of Minnesota, Minneapolis, Minnesota, USA
Vesna Prasnikar, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

CONTENTS

Inducing preferences is a method of creating a laboratory commodity that enables tests of expected utility in both single-person and multi-person economic environments.

TESTING THE PREDICTIVE ABILITY OF ECONOMIC THEORIES

Economic theory attempts to explain the behavior of agents in an economy. There is no stipulation in the theory that these agents are economists, that these agents have been trained in mathematics, or that they have any conscious awareness of the forces that guide their behavior. Rather, agents are assumed to behave 'as if' the tenets of the theory guided their behavior, even though they may be unaware that such forces are at work. (For example, agents may not know their own utility functions.)

Since the 1950s, economics has moved towards being a formal science. Notable advances include the first laboratory demonstrations that the theory of competitive equilibrium, the theory of games, and theories of choice have predictive content (Smith, 1962; Roth and Malouf, 1979; Siegel and Goldstein, 1959). Since expected-utility theory remains a cornerstone of much theorizing, there is a continuing interest in how to test its exact predictions, specifically its predictions of individual choices and prices. The results, while contributing to our understanding of a theory when it performs well, also appear to bear on a related body of research on attention, according to which slight environmental changes (in this case levels of incentives) can alter subjects' attention and thus their decision-making performance.

EXPECTED-UTILITY THEORY

According to expected-utility theory, each economic agent is in a world of uncertainty. Any setting, no matter how simple or complex, can be viewed as consisting of lotteries from which the agent is choosing. The theory is especially useful for describing choices between lotteries with monetary pay-offs. In such cases, it is possible to make two important assessments. Firstly, it is possible to determine, for any two lotteries A and B, whether lottery A is preferred to lottery B; and secondly, it is possible to determine the maximum (or minimum) price at which the individual would buy (or sell) any lottery.

In theorizing about behavior, economists assume that the system of preferences over lotteries can be summarized by a function (the 'utility function') that describes an individual's attitudes towards pay-offs, and that the individual in effect maximizes the expectation of this function. Applying this assumption to knowledge about specific lotteries enables the economist in principle to assess choices and prices.

For example, suppose lottery A represents a 0.5 chance of winning \$5 and a 0.5 chance of winning \$15; and lottery B represents a 0.5 chance of winning \$1 and a 0.5 chance of winning \$20. Suppose that the subject's attitudes toward dollar pay-offs x can be summarized by the utility function $u(x) = \sqrt{x}$. For each gamble, the expected utility will be the sum of the probability-weighted utilities of the pay-offs. Specifically, the expected utility of A is $0.5 \times u(5) + 0.5 \times u(15) = 3.05$, and the expected utility of B is $0.5 \times u(1) + 0.5 \times u(20) = 2.74$. Given that the decision maker maximizes expected utility, lottery A would be chosen over lottery B.

The maximum (minimum) price at which the decision maker would buy (sell) the lottery would be the amount that yields the same utility as the lottery. For lottery A, this amount would be determined by solving the equation $u(x) = \sqrt{x} = 3.05$, i.e., \$9.33. Since the certainty equivalent to the decision maker of the lottery (\$9.33) is less than its expected pay-off ($\$10 = 0.5 \times \$5 + 0.5 \times \$15$), the individual is said to be 'risk-averse'. A certainty

equivalent greater than \$10 would indicate a risk-preferring individual.

HOW TO INDUCE PREFERENCES

The method of inducing preferences (Berg *et al*., 1986; Roth and Malouf, 1979) enables researchers to test in a laboratory virtually any individual choice or equilibrium prediction that derives from expected-utility theory. (In general, an economy is in equilibrium when no agent has an incentive to alter his or her action.) The particular advantage of the inducing technique is that, to predict any subject's choices and prices, the experimenter does not need to learn the subject's utility function.

We will show how, by constructing examples, inducing preferences solves the problem of predicting bids and prices of a subject. The technique requires no assumptions about the subject's utility function except that the subject prefers more money to less.

To predict subjects' choices and prices, we introduce a binary lottery to be played for cash (for definiteness, assume it pays \$0 or \$10), along with a new currency called the 'econo', and a function G, that converts econos to probabilities of winning the binary lottery.

Suppose the subject has to choose between a lottery with a 50 per cent chance of 5 econos and a 50 per cent chance of 15 econos, and a lottery with a 50 per cent chance of 1 econo and a 50 per cent chance of 20 econos. Also assume that the subject's conversion function is given by $G(x) = \sqrt{(x/50)}$, where x represents the number of econos that the subject has after the lottery is played. Figure 1 illustrates one way in which lotteries are presented to subjects using this technique.

Each lottery is presented as a wheel divided into two parts. The proportion of the wheel shaded in a given color is the probability of receiving the number of econos associated with that color.

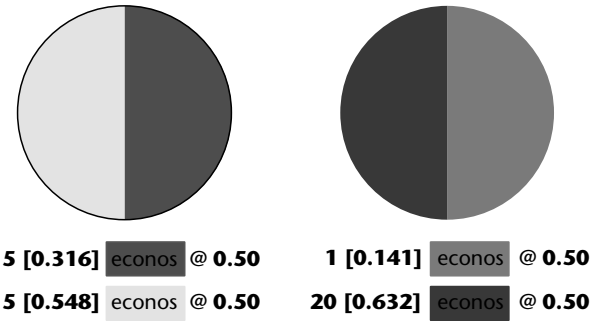


Figure 1. [*Figure is also reproduced in color section.*] A pair of lotteries as presented to subjects in the relative risk-averse condition.

Suppose the subject, with utility function u, considers lottery A. He or she would behave 'as if' going through the following computation. If red came up on the disk, the subject would have a G (5)probability of receiving \$10 and a $(1-G(5))$ probability of winning \$0. The expected utility would then be $G(5)u(\$10) + (1 - G(5))u(\$0)$. Similarly, the expected utility if yellow came up would be $G(15)u(\$10) + (1 - G(\$15))u(\$0)$. Therefore the expected utility of lottery A is $(0.5 \times G(5) + 0.5 \times G(15))u(\$10) + (1 - (0.5 \times G(5) + 0.5 \times G(15)))u(\$0) = 0.432 \times u(\$10) + 0.568 \times u(\$0)$.

A similar computation for lottery B yields $0.387 \times u(\$10) + 0.613 \times u(\$0)$.

Since more dollars are preferred to less, and since under lottery A there is a 0.432 chance of \$10 while in lottery B there is only a 0.387 chance, any individual who obeys expected-utility theory will prefer lottery A to lottery B. Note that this observation holds without the experimenter, or even the subject, knowing what u is.

PREDICTING BOUNDS ON ACCEPTABLE PRICES

Now consider the more difficult problem of predicting maximum (minimum) buying (selling) prices for a subject. (Again, for definiteness, assume that the binary lottery pays \$0 or \$10.) To price lottery A, the decision maker behaves 'as if' determining the price in econos (not dollars) of a lottery that pays off 5 econos with a 50 per cent chance and 15 econos with a 50 per cent chance.

Again, suppose the conversion function is $G(x) = \sqrt{(x/50)}$. The subject's expected utility is $0.432 \times u(\$10) + 0.568 \times u(\$0)$. The value x^* to the subject in econos of lottery A will be the number of econos that yields exactly the same probability of winning \$10 as when lottery A is played. Thus, $\sqrt{(x^*/50)} = 0.432$, or $x^* = 9.33$.

Without knowing the utility function of the subject, we are able to predict the price the subject will pay. Admittedly, the price is in econos; however, if our interest in studying behavior is the pricing process and not the price in dollars then it does not matter whether we are talking about dollars or econos. Since 9.33 is less than 10 (the expected number of econos), we can say that the subject is 'induced' to be risk-averse in econos.

EQUILIBRIUM BEHAVIOR

Much of economics is concerned with equilibrium behavior. By inducing preferences it becomes possible to unambiguously state what the equilibrium

predictions are in a laboratory experiment. By rewarding subjects with the probability of winning a binary lottery, Roth and Malouf (1979) and Cooper *et al*. (1990, 1993) have been able to test the predictive ability of the Nash equilibrium. A Nash equilibrium is one in which no individual agent has an incentive to alter behavior given the actions of the other agents. By mapping outcome–action pairs to probabilities of winning a binary lottery, Berg *et al*. (1992) study how well incentives (getting agents to take desirable actions) are traded off for risk sharing (getting agents to bear risk). By inducing concave functions on econos, Srivastava and O'Brien (1991) are able to ask if laboratory security markets lead to optimal risk sharing. The inducing technique has found application in accounting (Sprinkle, 2000) and marketing (John, 2001).

PERFORMANCE OF THE INDUCING TECHNIQUE

Since its inception, the inducing technique (and slight modifications of it) have been subjected to numerous examinations. One series of studies focused on its applicability in first-price auctions. (In a first-price auction, the highest bidder for an object gets the object at the price he or she bid.) Based on subjects' bids, Cox *et al*. (1984) estimated subjects' utility functions. Usually, subjects are estimated to have convex utility functions (i.e., they are risk-averse) when being paid in dollars. Cox *et al*. found that if subjects were induced to be risk-neutral, the corresponding estimated utility functions were not risk-neutral, but rather risk-averse in econos. These findings have been vigorously debated by Rietz (1993) and by Cox and Oaxaca (1995). Rietz argues that with proper modifications there is some support for inducing, while Cox and Oaxaca contest the ability of the technique to induce risk-neutrality in a majority of subjects.

Several points emerge from these two papers. First, it appears that in assessing predicted auction prices, the induction technique does reasonably well. Second, since to fit many subjects' data Rietz needs to incorporate a constant term, the inducing technique appears not to work for all values that come up in the auction.

It appears that both sets of results are consistent with the idea that the inducing technique works best when subjects are likely to win the auction, or alternatively when the potential gains are higher.

When lotteries are presented as in Figure 1, choices, on average, reflect risk-aversion (risk-preferring) when risk-aversion (risk-preferring) is induced. In pricing tasks, the effect of the induction technique is more variable than in the choice task. In a review paper, Berg *et al*. (2003) show that the variance in the bidding task is highest when penalties from erroneously bidding are lowest.

Selten *et al*. (1999) tried to induce risk neutrality in a nonstrategic setting. Their design involves four experimental treatments that vary in two dimensions: pay-offs are made in econos or in money, and subjects are allowed to request statistical summaries (expected value and mean absolute deviation from expected value). The tasks test the ability of the inducing procedure to mitigate traditional choice biases (the common-ratio effect, the reference-point effect, the preference-reversal effect, and violations of stochastic dominance). Selten *et al*. found that the inducing technique does not mitigate traditional choice biases.

Selten *et al*. do not assess differences in the performance of the lottery technique associated with the differences in expected rewards. Their raw data reveal that there were only three lotteries with high reward conditions. In only three lottery choices was the difference in winning the preferred prize greater than 14 per cent. For these choices, on average only 9 per cent of subjects' choices deviated from the predicted choice. In 70 per cent of all choices in the Selten *et al*. experiment, the expected difference in the probability of winning between the predicted and alternative choice was less than 6 per cent. Thus their data appears consistent with the notion that the level of incentive affects the performance of the technique.

Loomes (1998) further explores the Selten *et al*. result that individuals process probabilistic pay-offs in the same way that they process monetary pay-offs. In the Loomes task, a subject sees a lottery that has a 13/20 chance of outcome A and a 7/20 chance of outcome B. In the monetary task, the subject picks the pay-offs in pounds that he wishes to assign to each outcome (A and B), with the proviso that the sum of the pay-offs be 20 pounds. In the probability task, the subject assigns the probability of winning 20 pounds to outcome A and the probability of winning 20 pounds to outcome B, with the proviso that the probabilities sum to one. In the probability task, the subject will maximize expected utility by assigning probability 1 to the A outcome. However, in the money task, the subject will assign 20 pounds only if he or she is risk-neutral. Given a reasonable distribution of risk types, the distribution should be different in the two tasks. Loomes finds that he cannot identify any difference between performances in the

distribution of outcomes under this task. Like Selten *et al.*, Loomes did not attempt to measure whether the performance of the inducing procedure improved with the level of pay-offs.

THE PRASNIKAR STUDY

Induced Utility Functions

Prasnikar's (2001) study examines what produces variations in performance of the inducing procedure, and shows, as do Berg *et al.*, that conclusions about the inducing technique (paying off in probabilities) should be tempered by a consideration of the reward structure. Prasnikar attempted to induce one of three utility functions. One function reflected constant absolute risk-aversion, another reflected constant absolute risk-preferring, and the third, risk-neutrality. These functions were, respectively:

$$G(x) = \left(1 - e^{-0.07365x}\right)/\left(1 - e^{-0.35}\right) \quad (1)$$

$$G(x) = \left(-1 + e^{0.07365x}\right)/\left(-1 + e^{0.35}\right) \quad (2)$$

$$G(x) = x/50 \quad (3)$$

Estimation

For each induced utility function, Prasnikar asked if the observed responses deviated significantly from the coefficient of risk aversion she attempted to induce (the target). To assess the deviation, each subject's coefficient of risk aversion was estimated using probit analysis, assuming that the subject's utility function for econos was in the constant absolute risk-averse (preferring, relative risk) class. Then Prasnikar asked if the estimated coefficient was within a 95 per cent confidence interval around the target. The results are shown in Table 1. A total of 20 subjects were examined for each of the three utility functions. Estimated coefficients are stated, and the tests to determine if they are significantly different from prediction are presented, for each of the induced coefficients.

The Effect of Different Probabilities of Winning

Much of Prasnikar's study is concerned with determining the circumstances in which the induction technique does not predict. In the design, pairs of lotteries are selected so that the number of pairs of lotteries is evenly distributed from smaller to higher differences in expected rewards (i.e., absolute differences in expected probability between lottery A and lottery B). (In our example, the probability difference was $0.432 - 0.387 = 0.045$.) In Prasnikar's study, the differences in expected probabilities of winning for each choice pair are divided in increments of 5 per cent from 0 per cent to 35 per cent.

The data show a lower chance of making the predicted choice when the difference in expected probabilities between the choices is small. Even when the difference in expected probabilities is high, some subjects do make unpredicted choices. Figure 2 shows the percentage of correct predictions for each level of difference in expected probability of winning. The pattern, whereby predicted lotteries are selected more frequently at higher differences in expected probability than at lower differences, is observed in risk-neutral, risk-averse, and risk-preferring induction categories. For example, at 5 per cent difference in expected probability, subjects selected the predicted lottery in the risk-preferring sample only 37 per cent of the time, and 53 per cent of the time in the risk-averse and risk-neutral samples. The percentages did not change when Prasnikar compared the results of the lottery pair at 10 per cent minus 5 per cent and the lottery pair at 62 per cent minus 57 per cent. When the difference in expected probabilities is greater than or equal to 15 per cent, the chance of selecting the predicted lottery is always above 80 per cent for the risk-averse induced preferences; and it increases to 88 per cent at 30 per cent difference. For the risk-preferring induced behavior, at 15 per cent difference in expected probabilities subjects select the predicted lottery 71 per cent of the time, and if the difference is greater than 25 per cent, the predicted lottery is chosen 86 per cent of the time. Four subjects from the sample of 20 were making unpredicted choices when the difference was greater than or equal to 15 per cent, and when the difference was greater than or equal to 30 per cent there were two subjects who were making unpredicted choices for the risk-averse induced preferences. For the risk-preferring preferences, three subjects were making unpredicted choices for differences greater than 20 per cent.

The Effect of Different Knowledge Levels of Subjects

After completing all choices, each subject was tested to see whether they understood how to calculate the probability of winning the $10 prize in Prasnikar's experiment. (No instructions on computation of compound lotteries are given to subjects in the

Table 1. Parameter estimates of induced behavior for each individual (Prasnikar, 2001). Numbers in parentheses are standard errors

Subject	*Constant absolute risk-averse induced behavior*			*Constant absolute risk-preferring induced behavior*			*Risk-neutral induced behavior*	
	Estimated coefficient of risk-aversion	*LR test (proximity to target)*	*LR test (linearity)*	*Estimated coefficient of risk-aversion*	*LR test (proximity to target)*	*LR test (linearity)*	*Estimated coefficient of risk-aversion*	*LR test (proximity to target)*
1	0.086[a] (0.030)	0.84	66.40[c]	0.068[a] (0.011)	0.02	16.08[c]	0.482[a] (0.188)	8.44[c]
2	0.006 (0.181)	16.56[c]	2.80	0.017[a] (0.006)	1.52	5.98[c]	0.625[a] (0.144)	1.98
3	0.076[b] (0.044)	0.00	14.56[c]	−0.021[a] (0.006)	6.64[c]	19.22[c]	2.732 (1.528)	0.18
4	0.053[a] (0.007)	0.14	43.06[c]	0.052[a] (0.017)	0.08	35.46[c]	0.925 (0.902)	0.66
5	0.140 (0.220)	3.42	41.48[c]	0.028[a] (0.009)	1.14	9.03[c]	0.696[a] (0.123)	7.78[c]
6	0.039[a] (0.013)	2.08	6.72[c]	0.071[a] (0.012)	0.00	8.36[c]	0.781[a] (0.135)	2.2
7	0.084[a] (0.001)	0.28	31.73[c]	−0.012 (0.034)	7.96[c]	3.60[c]	0.543[a] (0.074)	5.98[c]
8	0.019 (0.007)	4.06[c]	1.80	0.056[a] (0.008)	0.28	5.40[c]	1.79[a] (0.53)	18.14[c]
9	0.023 (0.058)	0.80	16.08[c]	0.050[a] (0.009)	0.38	28.32[c]	0.971[a] (0.307)	2.34
10	0.084[a] (0.031)	0.24	28.67[c]	−0.065[a] (0.029)	8.78[c]	12.90[c]	0.816[a] (0.400)	2.06
11	0.059[a] (0.006)	2.52	50.82[c]	−0.016[a] (0.004)	5.98[c]	20.40[c]	0.751[a] (0.248)	0.42
12	0.113[a] (0.019)	1.30	24.82[c]	0.022[a] (0.005)	2.26	34.50[c]	1.105[a] (0.059)	3.12
13	0.068[a] (0.009)	0.16	40.08[c]	−0.024 (0.065)	12.18[c]	3.28	1.754[a] (0.526)	14.08[c]
14	0.018 (0.068)	4.76[c]	0.62	−0.037 (0.059)	14.10[c]	3.56	0.821 (0.737)	0.50
15	0.035[a] (0.007)	6.92[c]	14.32[c]	0.222[a] (0.051)	2.58	5.00[c]	0.989[a] (0.219)	0.02
16	0.087[a] (0.030)	0.84	46.40[c]	0.080[a] (0.014)	0.04	6.24[c]	0.834[a] (0.393)	0.14
17	0.058[a] (0.007)	2.06	36.18[c]	0.107[a] (0.019)	0.50	8.62[c]	0.818[a] (0.148)	1.72
18	0.060[a] (0.012)	0.58	25.74[c]	0.051[a] (0.014)	0.14	11.93[c]	0.935[a] (0.213)	0.10
19	0.006[a] (0.003)	4.28[c]	0.58	0.173[a] (0.038)	0.88	7.54[c]	1.508[a] (0.537)	3.80
20	0.087[a] (0.030)	0.84	19.82[c]	0.026[a] (0.008)	1.42	28.66[c]	2.315[a] (0.261)	4.24[c]

[a]Estimate significantly different from 0 at 5 per cent test level.
[b]Estimate significantly different from 0 at 10 per cent test level.
[c]Estimate significantly different from the null hypothesis.

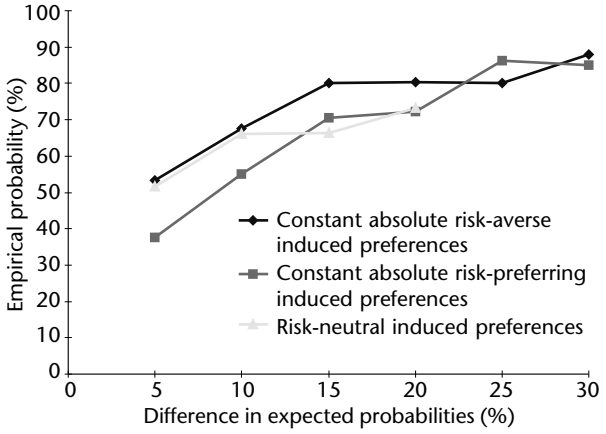


Figure 2. Empirical probability of making the predicted choice, as a function of the difference in expected probabilities (Prasnikar, 2001).

instructions before they finish the experiment.) This test question allows Prasnikar to distinguish subjects with good understanding of compound lotteries.

To identify whether unpredicted choices are related to the knowledge of how to compute the expected probability of winning, empirical frequencies were plotted separately for a sample of subjects who knew and for a sample who did not know how to calculate the expected probability. Figure 3 shows the empirical frequencies for each difference in expected probabilities.

Subjects who knew how to calculate the expected probability were making the predicted choices approximately 10 per cent more often than subjects who did not know how to calculate the expected probability, for risk-averse and risk-neutral induced preferences. At 30 per cent difference in expected probabilities, subjects who knew how to calculate the expected probability made the predicted choice 94 per cent of the time for the risk-averse induced preferences and 86 per cent of the time for risk-preferring induced preferences. Figures 2 and 3 show that the pattern of tendencies to make the predicted choice is similar for the induction of risk-aversion, risk-preference, and risk-neutrality.

The Effect of Natural Preferences

The design of the experiment also permits a test of the relationship of individuals' (natural) risk-aversion for money and the subjects' estimated risk coefficients (shown in Table 1). In theory, the natural risk-aversion for money should not affect the performance of the inducing technique. Prasnikar examines this hypothesis first by plotting the relationship between natural risk-aversion and the estimated coefficients of risk-aversion. Natural risk-aversion is assessed by asking a subject to answer a series of 19 questions in Table 2. Each question asks a subject to assess receiving \$5 for sure versus a p chance of \$10 and $1-p$ chance of \$0, where p varies between 5 per cent and 95 per cent. The minimum p for which the lottery is chosen is an indication of risk-aversion. Higher p is assumed to reflect higher risk-aversion.

Figure 4 shows this relationship for the three induced conditions (risk-averse, risk-preferring, and risk-neutral) with subjects divided into four categories, according to whether they knew how to calculate expected probability and whether their estimated coefficient of risk-aversion is significantly different from the induced coefficient. The horizontal axis represents the values of natural risk-aversion, measured by minimum selected probability p, and the vertical axis represents the estimated coefficients of risk-aversion obtained from Table 1.

In the risk-neutral sample, we find that a subset of players (i.e., subjects whose estimated coefficient is significantly different from the induced coefficient) with lower values of natural risk-aversion have lower values of the estimated coefficient of risk-aversion, while players with higher values of natural risk-aversion have higher values of the estimated coefficient of risk-aversion. A positive relationship between natural risk-aversion and estimated coefficient is also observed for the risk-averse sample.

To test formally for such a relationship, Prasnikar estimated a probit function with the coefficient of risk-aversion β replaced by $\beta_0 + \beta_1 p$, where β_0 measures the effect of induced preferences, β_1 measures the effect of natural risk-aversion, and p is the minimum probability selected by subjects. See Table 3.

Columns 6 and 9 suggest that subjects' decisions were influenced by their natural risk-aversion, and not only by their induced preferences. The estimates of natural risk-aversion – $\beta_1 = -0.078$ for risk-preferring induced preferences, and $\beta_1 = -1.012$ for risk-neutral induced preferences – are statistically significant. The effect of natural risk-aversion ($\beta_1 = 0.081$) for the risk-averse induced preferences

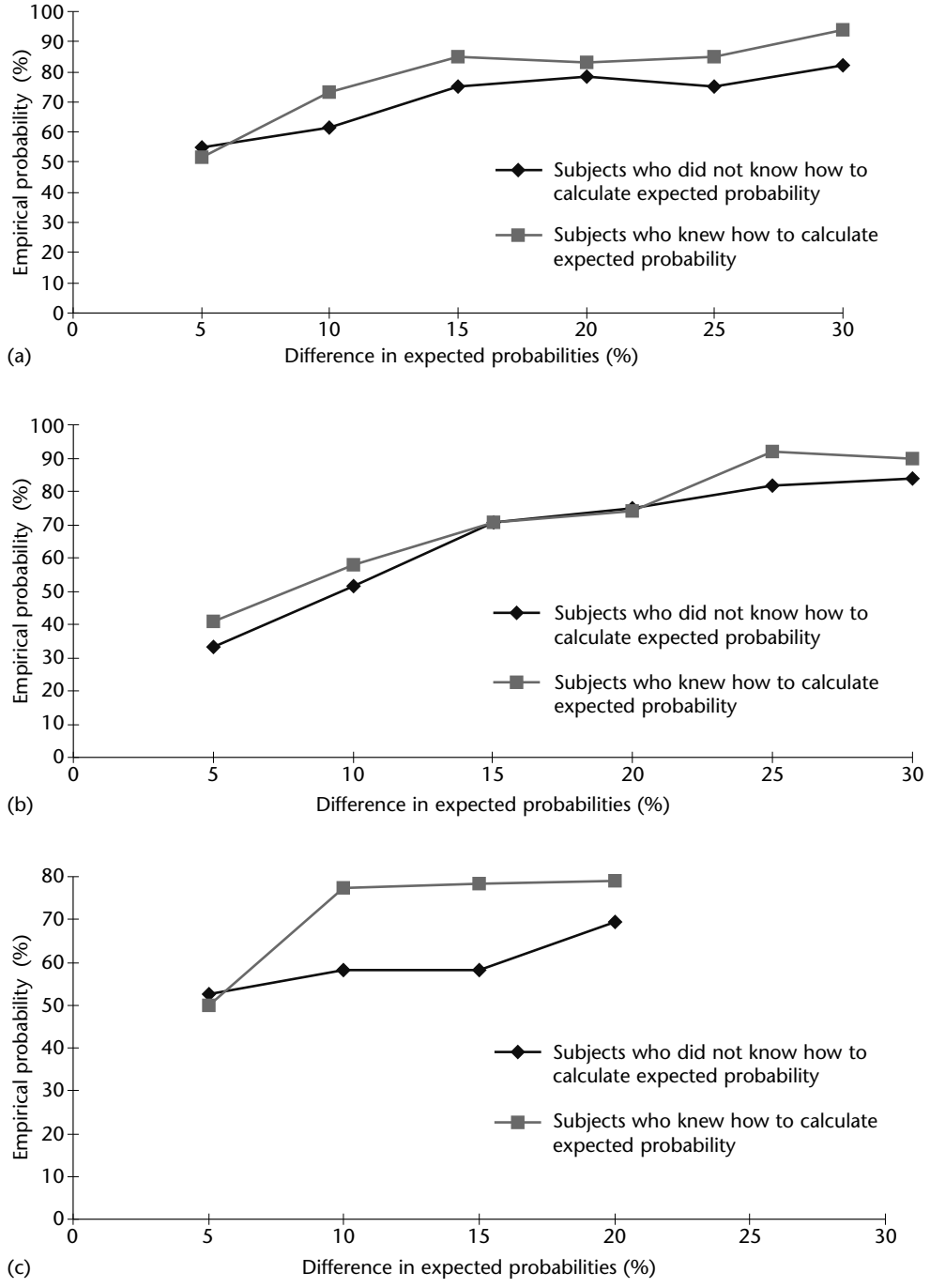


Figure 3. Empirical probability of making the predicted choice, as a function of the difference in expected probabilities for subjects who knew and subjects who did not know how to calculate the expected probability (Prasnikar, 2001). (a) Risk-averse induced preferences. (b) Risk-preferring induced preferences. (c) Risk-neutral induced preferences.

(column 3) is not statistically significant. Therefore the observed relationship between risk preferences and performance of the induction technique was positive; i.e., the higher a subject's natural risk-aversion, the higher was his or her estimated coefficient of risk-aversion for econos.

Interaction of Natural Risk Preferences and Knowledge Level

Next, Prasnikar addresses the question of whether this result represents the behavior of the whole sample, or only subjects who did not know how

Table 2. The sequence of choices given to subjects to test their natural risk-aversion for money

$5.00	for sure	< >	95 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	90 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	85 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	80 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	75 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	70 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	65 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	60 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	55 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	50 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	45 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	40 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	35 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	30 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	25 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	20 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	15 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	10 per cent	chance for	$10.00	or	$0.00
$5.00	for sure	< >	5 per cent	chance for	$10.00	or	$0.00
		Done					

use ↑ and ↓ to move up and down
use ← to select $5.00 for sure or → to select lottery

to calculate the expected probability (squares in Figure 4) or subjects who knew how to calculate the expected probability (circles in Figure 4). She examines these hypotheses by estimating the probit model of the form $\beta = \beta_0 + \beta_1 p$ separately for the subjects who did not know how to calculate the expected probability and the subjects who knew how to calculate the expected probability. The results are reported in Table 4.

For subjects who knew how to calculate the expected probability, Prasnikar finds no evidence that natural risk-aversion (as measured by the minimum probability p) influences the decision making. The estimated coefficient for the natural risk-aversion β_1 is never significantly different from zero for subjects who understand compound lotteries. For subjects who did not know how to calculate the expected probability, the estimates of natural risk-aversion become significant for the risk-preferring and risk-neutral samples. However, in the risk-averse sample the effect of natural risk-aversion is insignificant.

CONCLUSION

The research to date on inducing risk preferences is generally consistent with the hypothesis that the performance of the induction technique depends on the incentives. Simply stated, pay-offs matter. Knowledge of how to calculate joint probabilities affects the performance, but natural risk preferences seem to affect risk preferences only when knowledge of how to calculate joint probabilities is absent. When the differences between expected rewards are small, knowledge does not relate to performance (all subjects make a lot of mistakes). When the difference in expected rewards increases, everybody makes fewer mistakes, and subjects with computational knowledge perform better.

A related interpretation of the results can be based on research on attention. It is generally believed that there is no 'control homunculus', no central agent that allocates attention according to some rational principle; rather, attention can be easily diverted by slight environmental factors (Newell, 1980). Klien and Shore (2000) examined this issue in terms of 'covert' versus 'overt' orienting.

From a rational perspective, the induction technique is independent of the level of incentives; yet by manipulating the incentives we can produce observed results much more consistent with the induction model. It seems reasonable to postulate that the level of incentives changes the orientation and/or attention of the subject.

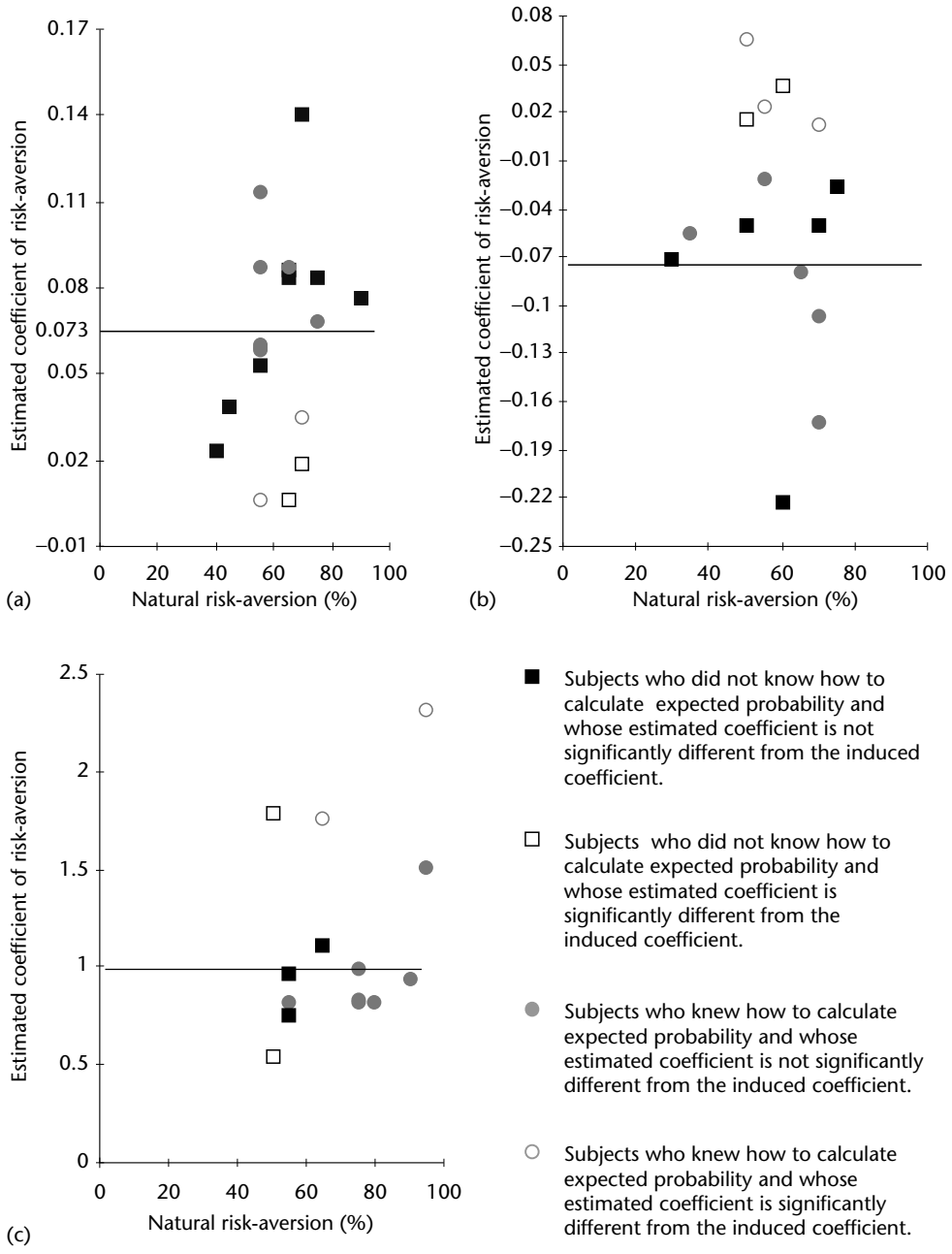


Figure 4. The relationship between the estimated coefficient of risk-aversion and natural risk-aversion for money (Prasnikar, 2001). (a) Risk-averse induced preferences. (b) Risk-preferring induced preferences. (c) Risk-neutral induced preferences.

From the standpoint of axiomatic choice, the critical assumption necessary to induce preferences is the compound-lottery axiom. As Luce (2000) points out, people's choices frequently violate this axiom. Camerer and Hogarth (1999) provide some understanding of incentive effects across studies, and suggest that incentives may very easily influence both mean and variance of performance levels. This may be the case with inducing preferences.

Table 3. Parameter estimates of the effect of induced preferences and natural risk-aversion (Prasnikar, 2001). Numbers in parentheses are standard errors

	Constant absolute risk-averse induced preferences			*Constant absolute risk-preferring induced preferences*			*Risk-neutral induced preferences*		
	(1)[c]	(2)[d]	(3)	(4)[c]	(5)[d]	(6)	(7)[c]	(8)[d]	(9)
α[e]	−1.072[a]	−0.157[a]	−0.172[a]	1.566[a]	1.566[a]	1.394[a]	1.204[a]	1.069[a]	1.082[a]
	(0.258)	(0.065)	(0.018)	(0.145)	(0.145)	(0.0383)	(0.198)	(0.227)	(0.221)
β_0	0.074[a]	0.175[a]	0.194[a]	0.039[a]	0.039[a]	0.092[a]	0.894[a]	0.876[a]	1.480[a]
	(0.011)	(0.042)	(0.024)	(0.002)	(0.002)	(0.097)	(0.064)	(0.084)	(0.231)
β_1	—	—	0.081	—	—	−0.078[a]	—	—	−1.012[a]
			(0.084)			(0.024)			(0.399)
Log-likelihood	−659.03	−471.19	−470.55	−680.87	−680.87	−672.78	−742.12	−559.96	−556.18
N	1155	770	770	1100	1100	1100	1100	825	825

[a]Estimate significantly different from zero at the 5 per cent test level.
[b]Estimate significantly different from zero at the 10 per cent test level.
[c]The model with $\beta_1 = 0$ for the whole sample.
[d]The model with $\beta_1 = 0$ and the sample that includes only subjects with identifiable natural risk-aversion.
[e]The intercept.

Table 4. Parameter estimates of the effect of induced preferences and natural risk-aversion (Prasnikar, 2001). Numbers in parentheses are standard errors

		Constant absolute risk-averse induced preferences		*Constant absolute risk-preferring induced preferences*		*Risk-neutral induced preferences*	
		(1)[c]	(2)	(3)[c]	(4)	(5)[c]	(6)
Subjects who knew how to calculate the expected probability	α[d]	−0.399[a]	−0.329[a]	1.692[a]	1.499	1.208[a]	1.215[a]
		(0.045)	(0.178)	(0.613)	(0.515)	(0.453)	(0.442)
	β_0	0.087[a]	0.160	0.047[a]	0.104[a]	1.004[a]	1.561[a]
		(0.023)	(0.087)	(0.010)	(0.028)	(0.145)	(0.413)
	β_1	—	0.009	—	−0.086	—	−0.891
			(0.074)		(0.067)		(0.636)
	Log-likelihood	−285.62	−285.54	−210.53	−209.87	−147.53	−146.06
	N	496	496	385	385	220	220
Subjects who did not know how to calculate the expected probability	α[d]	−0.054[a]	−0.275	1.142[a]	1.483[a]	1.031[a]	1.045[a]
		(0.012)	(0.362)	(0.517)	(0.600)	(0.264)	(0.253)
	β_0	0.054[a]	−0.102	0.041[a]	0.073[a]	0.817[a]	1.538[a]
		(0.022)	(0.144)	(0.012)	(0.024)	(0.107)	(0.268)
	β_1	—	0.018	—	−0.058[a]	—	−1.279[a]
			(0.151)		(0.029)		(0.480)
	Log-likelihood	−184.43	−182.32	−460.67	−457.48	−411.63	−408.81
	N	275	275	715	715	605	605

[a]Estimate significantly different from zero at the 5 per cent test level.
[b]Estimate significantly different from zero at the 10 per cent test level.
[c]The model with $\beta_1 = 0$.
[d]The intercept.

References

Berg JE, Daley LA, Dickhaut JW and O'Brien JR (1986) Controlling preferences for lotteries on units of experimental exchange. *Quarterly Journal of Economics* **101**: 281–306.

Berg JE, Daley LA, Dickhaut JW and O'Brien JR (1992) Moral hazard and risk sharing: experimental evidence. In: Isaac M (ed.) *Research in Experimental Economics*, vol. V, pp. 1–34. Greenwich, CT: JAI Press.

Berg JE, Dickhaut JW and Rietz TA (2003) On the performance of the lottery procedure for controlling risk preferences. In: Plott C and Smith V (eds) *Handbook of Experimental Economic Results*.

Camerer C and Hogarth R (1999) Incentives in experiments: a review and capital–labor production framework. *Journal of Risk and Uncertainty* **19**: 1–3, 47–48.

Cooper R, DeJong DV, Forsythe R and Ross TW (1990) Selection criteria in coordination games: some experimental results. *American Economic Review* **80**: 218–233.

Cooper R, DeJong DV, Forsythe R and Ross TW (1993) Forward induction in the battle-of-sexes games. *American Economic Review* **83**: 1303–1316.

Cox JC and Oaxaca RL (1995) Inducing risk neutral preferences: further analysis of the data. *Journal of Risk and Uncertainty* **11**: 65–79.

Cox JC, Smith VL and Walker JM (1984) Theory and behavior of multiple unit discriminative auctions. *Journal of Finance* **39**: 983–1010.

Klien RM and Shore DI (2000) Relationships among modes of visual orienting (commentary). In: Monsell S and Driver J (eds) *Attention and Performance*, vol. XVIII, pp. 195–208. Cambridge, MA: MIT Press.

Loomes G (1998) Probabilities vs money: a test of some fundamental assumptions about rational decision making. *Economic Journal* **108**: 477–489.

Luce RD (2000) *Utility of Gains and Losses*. Mahwah, NJ: Lawrence Erlbaum.

Newell A (1980) Reasoning, problem-solving, and decision processes. In: Nickerson R (ed.) *Attention and Performance*, vol. VIII, pp. 693–718. Hillsdale, NJ: Lawrence Erlbaum.

Prasnikar V (2001) How well does utility maximization approximate subjects' behavior? An experimental study. Working paper.

Rietz TA (1993) Implementing and testing risk preference induction mechanisms in experimental sealed bid auctions. *Journal of Risk and Uncertainty* **7**: 199–213.

Roth AE and Malouf MWK (1979) Game-theoretic models and the role of bargaining. *Psychological Review* **86**: 574–594.

Selten R, Sadrieh A and Abbink K (1999) Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision* **46**: 211–249.

Siegel S and Goldstein DA (1959) Decision making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology* **57**: 37–42.

Srivastava S and O'Brien J (1991) Dynamic stock markets with multiple assets: an experimental analysis. *Journal of Finance* **46**: 1811–1838.

Smith VL (1962) An experimental study of competitive market behavior. *Journal of Political Economy* **70**: 111–137.

Sprinkle G (2000) The effects of incentives on learning and performance. *The Accounting Review* **75** (3): 299–326.

Further Reading

Arrow KJ (1971) *Essays in the Theory of Risk Bearing*. Chicago, IL: Markham.

Becker GM, DeGroot MH and Marshak J (1964) Measuring utility by a single-response sequential method. *Behavioral Science* **9**: 226–232.

Breiter HC, Aharon I, Kahneman D, Dale A and Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**: 619–639.

Diamond P and Rothschild M (1978) *Uncertainty in Economics: Reading and Excercises*. New York, NY: Academic Press.

Kreps D (1987) *Notes on the Theory of Choice*. Boulder, CO: Westview Press.

Machina M (1987) Choice under uncertainty: problems solved and unsolved. *Journal of Perspectives* **1**: 121–154.

Pratt J (1964) Risk aversion in the small and in the large. *Econometrica* **32**: 122–136. [Reprinted in Diamond and Rothschild (1978).]

Savage L (1954) *The Foundations of Statistics*. New York, NY: John Wiley.

Information Cascades and Rational Conformity

Intermediate article

Lisa R Anderson, College of William and Mary, Williamsburg, Virginia, USA
Charles A Holt, University of Virginia, Charlottesville, Virginia, USA

CONTENTS

An information cascade is a pattern of matching decisions. A cascade can occur when people observe and follow 'the crowd', which can be rational if the information revealed in others' earlier decisions outweighs one's own private information.

INTRODUCTION

When individuals obtain private information and make publicly observed decisions in a sequence, the first decisions tend to act as 'signals'. If early decisions show a clear pattern, the information inferred from them may outweigh any one person's private information. This inference can cause people to 'follow the crowd', even when the group consensus conflicts with their own private information. This type of sequential conformity is termed an 'informational cascade' in a seminal paper by Bikhchandani *et al*. (1992).

For example, consider an employer who interviews a job applicant and forms a good impression. The employer, however, does not offer the job after discovering that the worker has been turned down previously by other employers. Even though this decision is in conflict with the employer's own information, it may be rational if the employer concludes that the other interviews went badly, and that the aggregate information implied by past interviews more than offsets the employer's own positive evaluation. Even a qualified applicant may make a bad impression on any given day, so may have difficulty finding a job after several unsuccessful attempts.

Bannerjee (1992) uses the term 'herd behavior' to describe similar patterns of conformity that arise in models where individuals must decide which type of financial asset to purchase. Indeed, much of the interest in cascade behavior arises from attempts to explain temporary patterns in investment behavior.

CONFORMITY INCENTIVES

There may, of course, be non-informational factors that produce conformity in social interactions. Sometimes people prefer to behave like the others in a group. Such behavior has even been recommended (Post, 1927, chap. XXXIII): 'to do *exactly as your neighbors* do is the only sensible rule'.

There may be social stigmas and punishments associated with nonconformity. For example, an economic forecaster may prefer the chance of being wrong with everybody else to the risk of providing a deviant forecast that turns out to be the only incorrect guess. In the words of John Maynard Keynes (1936, p. 158): 'worldly wisdom teaches that it is better for reputation to fail conventionally than to succeed unconventionally'.

Some research suggests that people prefer to maintain the status quo (Samuelson and Zeckhauser, 1988). For example, subjects in experiments were provided with a scenario in which they inherit a portfolio of cash and securities and are asked whether to leave the portfolio intact or to change it by investing in other securities. There was a strong tendency for individuals to retain portfolio A in preference to B when it was listed as the current portfolio, and to retain portfolio B in preference to A when it was listed as the current portfolio. A similar tendency was observed in response to other matched pairs of questions that alternated descriptions of the *status quo* choice.

Such a 'status quo bias' may explain herding behavior in sequential decision situations. But these decision patterns do not allow us to distinguish between behavior based on a preference for

conformity and behavior that is motivated by information inferred from prior decisions. Thus, someone inheriting a portfolio from a rich uncle might conclude that the uncle's wealth was due to wise portfolio choices that should be imitated. Laboratory experiments can be used to set up and control information flows in order to distinguish among alternative explanations of herding behavior.

RATIONAL INFORMATION CASCADES

We will use a numerical example to illustrate the concept of a rational information cascade (Anderson and Holt, 1997). In this example, there are two equally likely events, A and B, which might represent whether or not a particular patent is marketable. Decision-makers obtain private signals, *a* and *b*, which are correlated with the events. In particular, $P(a|\mathrm{A}) = P(b|\mathrm{B}) = \frac{2}{3}$, so the error rate is $\frac{1}{3}$ for each signal. (For example, the signal might be the result of a consultation with experts.) This model assumes that each person's private signal is correlated with the event but is independent of the other signals. After observing their signals, individuals are approached one by one in a sequence and are asked to make a prediction about which event has occurred. People find out the prior predictions, if any, made by others, but they cannot observe others' private signals. Thus the prediction made by the first person is based only on that person's signal, and hence will reveal that signal, since the signal is more likely to be correct than not.

Suppose the first person sees a *b* signal and publicly predicts event B. If the second person in the sequence sees a *b* signal, it is rational for this person to also predict B. If the second person sees an *a* signal, the observed and inferred signals essentially cancel each other, and each state is equally likely. We observe from laboratory experiments that individuals almost always use their own information in such cases, and therefore, the second decision will reveal the person's private signal, whether or not it conforms to the first prediction. When the first two individuals in the sequence observe the same signal, their decisions will also match. In this case, the information inferred from the matching decisions is greater than the information inferred from any one private signal. In particular, if the first two people choose B, then the third person should also choose B, even if that person's private signal is *a*. Information cascades form in this manner, and the effect is that all subsequent decision-makers will follow a pattern established by the first ones in the sequence.

This example was used by Anderson and Holt (1997) in a laboratory experiment in which subjects were paid a cash reward for each correct prediction. The events were referred to as 'urn A' and 'urn B'. Each urn was a cup with three colored balls, which we will refer to as *a* or *b* signals. There were two *a* balls and one *b* ball in urn A, and there were two *b* balls and one *a* ball in urn B.

A random device was used to select the urn, with each event being equally likely, and therefore, each of the six balls is equally likely to be drawn. Suppose that the draw is *b*. Since two of the three *b* balls are in urn B, the posterior probability of urn B given a draw of *b* is $\frac{2}{3}$. (This is an example of the application of Bayes' rule; or more precisely, the ball-counting heuristic used here corresponds to the conditional-probability calculations that are referred to as Bayes' rule. See Holt and Anderson (1996) for a discussion of how the simple counting heuristic can be used to make Bayesian calculations in more complicated settings, and an intuitive explanation of the mathematical expression of Bayes' rule.) Of course, it is intuitively obvious that the probability of urn B is greater than $\frac{1}{2}$ when the signal is *b*. Holt and Anderson (1996) show that the probability of urn B is still greater than $\frac{1}{2}$ when first two predictions are B and the third person's signal is *a*. Thus, a cascade can begin with two matching decisions, and all subsequent decision-makers should follow the pattern established in this manner.

Information cascades may not form immediately if there is not a bias in early predictions. Suppose, for example, that the first two predictions are A and B, so the third person would consider each urn to be equally likely, prior to seeing a private signal. If the third and fourth decision-makers both predict B, then this bias in favor of B would cause the fifth person to predict B, regardless of that person's private signal.

These calculations are based on a model of perfect rationality. The purpose of an experiment is to determine how people actually behave in such situations. Anderson and Holt (1997) used the 'ball and urn' scenario described above in order to remove any preference for conformity that is not based on informational considerations. Subjects were placed in small cubicles and were shown a single ball drawn from the relevant urn, but they could not see which urn was being used. Subjects were selected in a random order to make their predictions, which were announced by a neutral assistant who did not know the signals or which urn was being used. (Allowing subjects to announce their own predictions could have given them the chance

to convey additional uncontrolled information, by tone of voice, for example.) After all predictions had been announced, a non-decision-making subject serving as a 'monitor' announced which urn had actually been used. Those with correct predictions were paid $2, and others earned nothing for that trial. There were 15 trials and six decision-making subjects in each session. Altogether, there were 12 sessions in the experiment, each of which was conducted on a different day.

The sequences of draws made cascades possible in more than half of the trials, and cascades actually formed in about 70% of the trials in which they were possible. A particularly interesting trial is shown in Table 1. In this case, all six individuals earned $2, since urn B was actually used for the draws.

Generally, prior information is informative, and cascade behavior tends to increase the accuracy of predictions, and hence to increase earnings. It is possible, however, for initial predictions to be incorrect, which may create an incorrect cascade. This occurred in another trial, shown in Table 2. In this case the first person's prediction revealed their signal, but the second person made a serious error in predicting B after seeing an *a* signal and a prior A prediction. The third person predicted B, which was also an error, since previous predictions were balanced and their own signal suggested urn A. All of the remaining individuals predicted B, which turned out to be incorrect, since urn A was used.

The most common type of error occurred when a person saw a signal that was inconsistent with the implication of prior predictions. In this case, the optimal Bayesian prediction is to follow the established pattern, but subjects deviated in about a quarter of the cases in which their own private information was inconsistent with this pattern.

In summary, the general tendency was for subjects to use the information implied by previous decisions correctly, which produced rational information cascades. There were, however, deviations that could either break a cascade or result in an incorrect cascade. Incorrect cascades also occasionally resulted from 'unlucky' incorrect decisions observed by early decision-makers.

This pattern of results was replicated by Hung and Plott (2001), who added some interesting treatment variations. In one of their experiments, the incentive structure was altered so that subjects received a positive pay-off only if the majority of the group made the correct prediction. (This is somewhat like a jury whose decision is determined by a majority vote.) The effect was to reduce conformity for early decisions because individuals have an incentive to signal their information so that others can make better decisions. A second treatment rewarded conformity directly: subjects received a positive pay-off only if their decision matched that of the majority.

Table 1. A trial from Anderson and Holt's (1997) 'ball and urn' experiment. Three of the last four decision-makers follow the pattern established by the first two despite contradictory evidence

Subject	58	57	59	55	56	60
Observes	*b*	*b*	*a*	*b*	*a*	*a*
Predicts	B	B	B	B	B	B

Table 2. A trial from Anderson and Holt's (1997) 'ball and urn' experiment. The second and third subjects both made errors, which led to an incorrect cascade (urn A was used)

Subject	57	58	59	55	60	56
Observes	*a*	*a*	*a*	*a*	*a*	*a*
Predicts	A	B	B	B	B	B

APPLICATIONS TO MARKETS AND OTHER SOCIAL INSTITUTIONS

Strong movements in stock prices are sometimes attributed to herd-like behavior. Keynes (1936) noted the similarity between investment decisions and a guessing game in which participants have to predict which contestant in a beauty contest will get the most votes. In this game, each person has to think about whom the others think is attractive, and also about whom the others think others will find attractive, and so on. Similarly, when investing in stocks, one would like to guess which enterprises will become popular with other investors, since a strong demand will raise the prices of those stocks. A herd-like response may move asset prices out of line with market fundamentals, and thus set the stage for an equally strong 'stampede' in the other direction. Such behavior could be due to seemingly irrational 'animal spirits', to use Keynes's colorful term, or it could be due to a rational tendency to follow others' decisions when they are based on independent sources of information. Christie and Huang (1995), for example, argue that it can be rational to follow others' decisions during surges or declines in stock prices; i.e., to rely on inferences derived from information that is aggregated by market prices.

Other applications are suggested by the majority-voting treatment of Hung and Plott (2001). In a trial, for example, jurors may form independent judgments about the guilt or innocence of a defendant, but such judgments are often changed in the process of voting and deliberation. In this manner, herd behavior can create the consensus needed to avoid a 'hung jury'. Some experiments that simulate sequential jury voting have been conducted, and strong patterns of cascade-like conformity are observed in many cases.

Decisions may occur in sequence in these applications (e.g., as stock purchases appear on a ticker tape), but the order of decisions is not exogenously specified as it was in the experiments discussed above. The order of voting is not exogenous in jury voting unless the foreman chooses to take votes by going around the table. Similarly, stock purchases or responses to initial public offerings are not subject to order requirements. Plott *et al*. (1997) report some pari-mutuel betting experiments with an endogenously determined order of play. The incentive structure is like that of a horse race in which the purse is divided among those who bet on the winning horse, in proportion to their bets. The experiment was presented as a choice between six assets, with only one of them offering positive earnings, depending on the realized state of nature. Investors had private, noisy information about the assets, and they could observe others' bets as they were made. A considerable amount of information aggregation was observed in these experiments, with the asset prices accurately indicating the correct state in most cases. In some cases, however, heavy purchases of an asset triggered more purchases, even though the asset being purchased turned out not to be the one that paid off. This corresponds to an incorrect cascade. The application of information cascade theory to asset markets in richer and more realistic settings is a prime area for future research.

CONCLUSION

Theoretical models of 'herding' pertain to situations where individuals observe private signals that are correlated with some unknown event. Predictions about the event are made in sequence, with later decision-makers being able to base their predictions on their own signal and on information inferred from prior decisions. The first few decisions tend to reveal the private signals, which may establish a pattern of matching predictions that others follow, even if the conforming predictions are different from the prediction that would be best given only the person's own private signal. This type of 'information cascade' can produce conformity that is rational, because the information content in prior decisions may outweigh that in one's own private signal. There is some laboratory evidence to support such theories of rational cascades, although individuals do make mistakes, and behavior is sometimes influenced by biases and heuristics that may lead to non-Bayesian decisions.

References

Anderson LR and Holt CA (1997) Information cascades in the laboratory. *American Economic Review* **87**: 847–862.

Bannerjee AV (1992) A simple model of herd behavior. *Quarterly Journal of Economics* **107**: 797–817.

Bikhchandani S, Hirshleifer D and Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* **100**: 992–1026.

Christie WG and Huang RD (1995) Following the Pied Piper: do individual returns herd around the market? *Financial Analysts Journal* **51**: 31–37.

Holt CA and Anderson LR (1996) Classroom games: understanding bayes' rule. *Journal of Economic Perspectives* **10**: 179–187.

Hung AA and Plott CR (2001) Information cascades: replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review* **91**: 1508–1520.

Keynes JM (1936) *The General Theory of Employment, Interest, and Money*. London, UK: Macmillan.

Plott CR, Wit J and Yang WC (1997) Paramutuel betting markets as information aggregation devices: experimental results. Working Paper, California Institute of Technology.

Post E (1927) *Etiquette in Society, in Business, in Politics, and at Home*. New York, NY: Funk and Wagnalls.

Samuelson W and Zeckhauser R (1988) Status quo bias in decision making. *Journal of Risk and Uncertainty* **1**: 7–59.

Further Reading

Anderson LR (2001) Payoff effects in information cascade experiments. *Economic Inquiry* **39**: 609–615.

Anderson LR and Holt CA (1996) Classroom games: information cascades. *Journal of Economic Perspectives* **10**: 187–193.

Asch SE (1952) *Social Psychology*. New York, NY: Prentice-Hall.

Asch SE (1956) Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs* **70**(9).

Camerer C (1995) Individual decision making. In: Kagel JH and Roth AE (eds) *The Handbook of Experimental Economics*, pp. 587–703. Princeton, NJ: Princeton University Press.

Davis DD and Holt CA (1993) *Experimental Economics*. Princeton, NJ: Princeton University Press.

Devenow A and Welch I (1996) Rational herding in financial economics. *European Economic Review* **40**: 603–615.

Kahneman D and Tversky A (1973) On the psychology of prediction. *Psychological Review* **80**: 237–251.

Pound J and Shiller RJ (1989) Survey evidence on diffusion of interest and information among investors. *Journal of Economic Behavior and Organization* **12**: 47–66.

Thaler RH (ed.) (1993) *Advances in Behavioural Finance*. New York, NY: Russell Sage Foundation.

Welch I (1992) Sequential sales, learning, and cascades. *Journal of Finance* **47**: 695–732.

Markets, Institutions and Experiments

Intermediate article

Vernon L Smith, George Mason University, Arlington, Virginia, USA

CONTENTS

EXCHANGE, ECONOMIC THEORY, AND THE HAYEK CRITIQUE

Economics is about exchange. In modern economies this means that people trade tangible goods, services and rights, given and received for money, whose value is itself derived from the rights conveyed by the bearer. Money, however, is a recent social contrivance in the long dimly lit history of trade among early peoples. Such trade predates the state and even agriculture, which is only some 10 000 years old. The archaeological evidence for trade survives in the durables – weapons, tools, and ornaments – left behind in caves and campgrounds, such trade inferred largely from the fact that the geographical distribution of such artifacts is separated by great distances from the distribution of the raw materials from which they were manufactured.

Adam Smith saw that trade provided the foundation for specialization and a vast expansion in human productivity. Hence, the division of labor (among specialties) is limited by the extent of the market. It is the presence of market opportunities that permits one person to grow corn, another hogs, the baker to bake and the butcher to cut meat, as each specializes in that for which he is suited by temperament, experience, or natural skills, and then satisfies his general needs through markets. Smith understood that knowledge was dispersed in the market system, and that the individual, knowing his local situation, could better judge than the 'statesmen or the lawgiver' how to employ his capital to its greatest value. But it remained for F. Hayek (1984/1945) to articulate more fully the idea that the market order served to coordinate the utilization of this dispersed information through the price system and to see that it constituted an extended order of cooperation among thousands, indeed millions, of uncomprehending individuals. Markets are an example of a self-ordering system driven by cultural evolution and have important parallels with complex self-ordering systems in biology.

Economic theory traditionally has modeled the individual consumer in competitive markets as directed by the goal of maximizing his current period utility over a given commodity and work effort space, subject to a budget balance constraint requiring the income from sources, such as labor earnings, to equal expenditure on commodities, given commodity prices and wages. Similarly, producers maximize profit given wages, input and output prices, and their knowledge of technology. This way of representing the 'economic problem of society' leads to mathematical conditions defining a competitive equilibrium (CE) and the analysis of its existence and properties. The history of economic theory has suggested two different (equally unsatisfactory) answers to the question of how economic agents might be able to achieve a CE. One was to assert that a CE is simply an ideal state achievable only if the agents all have complete information on each other's individual preferences and production opportunities, i.e. on all data the theorist needs to calculate a CE. The second was that if all agents were 'price takers', having no power to control any price, then the CE would be attained.

The first begs the question of how an economy processes and utilizes information, while the second begs the question of how prices are formed.

ECONOMIC ENVIRONMENT

People trade because there are (expectations of) gains from exchange. What the seller vends is

worth more to the buyer than to the seller, and therefore the transfer can make each party better off. The term 'economic environment' will be used to describe the set of all individual circumstances in a market that defines the total potential gains from exchange. For simplicity most of our discussion will be confined to a single market where these statements can be defined unambiguously.

In an experimental market we need to motivate real people (the subjects) with real money (or other reward medium) to make consequential choices in trade. Subjects are recruited to the laboratory with the understanding that they will earn real money, depending largely upon their decisions, and that such earnings will be paid to them in US currency at the end of the experiment. Imagine that I have recruited you to an experiment. After being paid \$5 for reporting to the lab at the designed time, you and several other people are each assigned to a computer monitor and separated from one another so that you see only your own monitor screen. The instructions inform you that you will be a buyer (others in the room are buyers and sellers but you do not know who is which). You learn that you will have a capacity to buy up to three units of the identical items to be traded, the first with value to you of \$10, the second \$7, and the third \$4. You will have the opportunity to buy units against these values in each of a series of market trading periods, and you profit from selling below these values. Specifically, if in a given period you buy one unit for \$6, a second for \$5, and are unable to buy a third unit for less than \$4, then I (the experimenter) owe you \$6 ($10-6+7-5$). Hence, you are motivated to buy each unit in your assigned capacity at a low price, but if you attempt to buy at too low a price, you may not find a willing seller, and fail to earn a profit. The other buyers in the market have also been assigned values privately, but you know nothing of these and nothing of the costs assigned to sellers. Your values, \$10, \$7, and \$4 are a means of summarizing in concrete terms the fact that people have differing local circumstances, represented here by a maximum willingness-to-pay for each of three successive units. This is your little fragment of the dispersed information among all participants in the market.

Similarly, each seller is assigned values representing the costs incurred for transferring units owned by the seller to buyers. Sellers profit by selling at prices that are above these unit costs. Thus, a seller with unit costs of \$1 and \$3 for two units, if sold at the prices \$6 and \$5 respectively makes a profit of \$7 ($6-1+5-3$). Methodologically, this technique of using monetary rewards for inducing value on abstract items makes plain the fundamental truth that buyers as well as sellers profit from exchange.

The left panel of Figure 1 shows the value/cost environment for an early experiment consisting of four buyers and three sellers (Ketcham *et al*., 1984; reprinted in Smith, 1991, pp. 295–314). Note that the 'economic environment' (see Figure 2) becomes the effective demand and supply in the experimental market when we sort all buyer values from highest to lowest (demand) and all seller costs from lowest to highest (supply) regardless of 'ownership' identity. Thus 'demand' is the maximum willingness-to-pay schedule in the market, just as 'supply' is the minimum willingness-to-accept schedule. Observe for now that the sum total of this dispersed information in Figure 1 involves four buyers, each with a capacity to buy up to four units each (16 units total) and three sellers with a capacity to sell up to five units each (15 units total). Also note that the buyer's values are distinct for each unit demanded, while the sellers each incur a constant per unit cost up to their capacity, but these constant unit costs are distinct for each seller reflecting different individual circumstances, although each employs a constant unit cost technology.

INSTITUTIONS

All markets operate by rules, formal and explicit as in organized exchanges such as the Chicago Mercantile Exchange (Merc) for trading claims on assets or their derivatives, and the Automated Credit Exchange (ACE) for trading emission credits in Southern California, or by informal and implicit norms as in two-person social and economic exchange. Institutions define the language – the messages – of the market, such as bids, offers, and acceptances, the rules that govern the exchange of messages, and the rules that define the conditions under which messages lead to allocations and prices. If there are n agents, $i=1,2,\ldots,n$, and each i chooses a message m_i, then the allocation x_i to agent i is defined by the institution as a rule that we can express in the generic functional form

$$x_i = h(m_1, \ldots, m_i, \ldots, m_n).$$

Where the institution recognizes different agent classes subject to different rules, we would write $x_i = h_i(\cdot)$ indicating by means of the subscript i that the allocation rule also depends upon i's classification. Thus specialists on a stock exchange are subject to rules that differ from those of member traders.

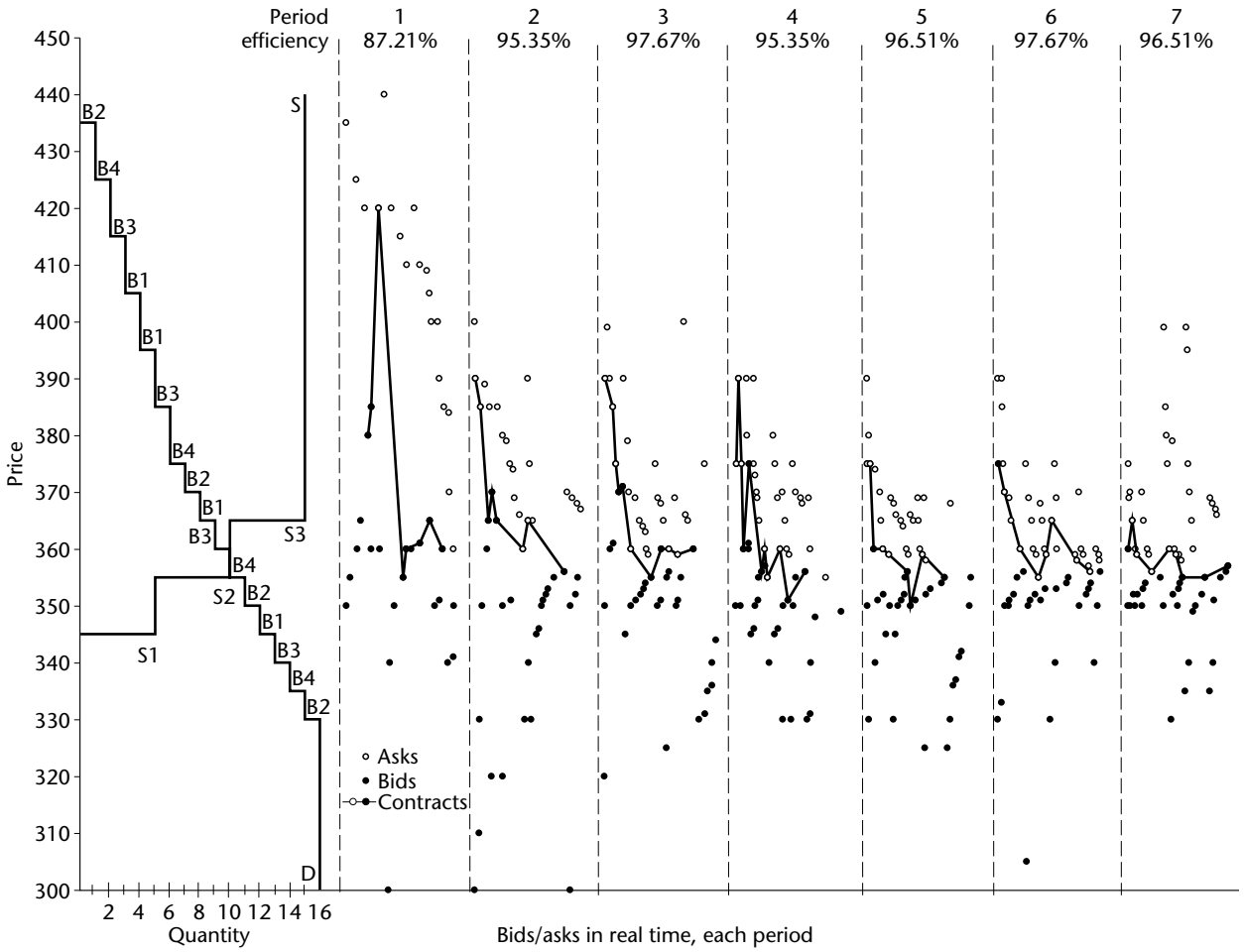


Figure 1. A double auction experiment. The economic environment for four buyers and three sellers is shown on the left. On the right is shown the sequence of bids, asks and contracts in each of the first seven periods of trading.

As an illustration of such institution-defining rules, in the ascending bid ('English') auction each new bid (message) must be higher than the standing bid, and the award is to the last bidder at a price equal to the last bid, when no new bids are forthcoming. That is, the English auction rules are

$$x_1 = h(m_1, \ldots, m_i, \ldots, m_n) = 1; x_k = 0 \quad \text{for all } k \neq 1,$$

where we have numbered the bidders so that $m_1 > m_2 > \ldots > m_n$, and $i = 1$ has the highest bid. Hence the single item for sale is awarded to Mr 1, and all others receive nothing. Note the important distinction between messages and awards: during the auction when Mr 1 announces the bid m_1, no one yet knows who will be awarded the item. Subsequently, all learn that no other agent, k, is willing to raise the bid, so that then and only then do the rules of the institution tell us that $x_1 = 1$, $x_k = 0$. Mr 1 does not choose to buy the item. He chooses to raise the standing bid. The institution subsequently declares him to be the buyer by virtue of the rules, under which it is discovered that no other bidder is willing to bid higher.

The advent of experimental economics in the mid-twentieth century has created a technology allowing the performance properties of institutions to be studied in controlled induced value/cost environments. In this article some simple but powerful cases will be used to illustrate what has been learned from the techniques of experiment. As we shall see, however, the institutions that survive are not born equal on any one measure of performance: efficiency, price volatility, demand responsivity, dependence on external information channels and the (transactions) cost of participation. Rather, each seems to be an adaptation to environmental wrinkles, or niches, that are not evident to the naked eye.

By the mid 1970s experimental auction market studies had become automated by computer/communication technology, i.e. participant messages were communicated by keyboard input and displayed on monitor screens; the institutional rules were encoded as algorithms applied to messages; and the data were time-stamped and recorded as specified. As noted in Rassenti *et al*. (2001) this had far reaching consequences, one of the more significant of which was the introduction of 'designer markets', of which ACE is a living, breathing, operating example (Ishikida *et al*., 2001).

BEHAVIOR

Behavior connects motivation in the environment with the institution to yield decisions and outcomes. Agents with differing circumstances have a differing urgency (maximum willingness to pay) to acquire goods/services and differing priorities (minimum willingness to accept) for relinquishing goods/services. The trading process is one in which people choose messages based on their circumstances, and knowledge of the language and rules of the market. Thus, if agent i's circumstances in the economic environment are represented by E_i and I is the set of institutional rules, behavior can be expressed by

$$m_i = \beta(E_i|I), i = 1, 2, \ldots, n.$$

Behavior is a mapping from individual circumstances, conditional on the market rules, into messages. The institution chooses outcomes by application of its rules to the messages, as indicated above by $x_i = h(m_1, \ldots, m_i, \ldots, m_n)$. This is illustrated in Figure 2. Each agent has an information state, preferences, costs, resources, knowledge, and, knowing the institutional rules, chooses messages. The institution processes the messages to determine allocations and prices. This pathway, from the economic environment down through choice and the institution, up to outcomes in Figure 2 represents the operation of the market. Across the top of Figure 2, the omniscient experimenter, whose information is not given to any one participant's mind, can use the information to compute the maximum gains from exchange and CE. This allows the observed outcomes to be used to compute performance measures: efficiency (percent of maximum gains realized by the agents), and the volatility or stability of observed prices relative to the CE. Because the rules of the market affect incentives, we expect, and experiments confirm, that institutions matter in the behavior we observe and in the outcomes that result.

This methodology allows different environments to be compared using the same institution. It also allows different institutions to be compared while holding constant the economic environment.

THE DOUBLE AUCTION INSTITUTION

This trading institution, used throughout the world in financial, commodity, and currency markets, is a two-sided multiple unit generalization of the ascending bid auction for unique items. Buyers submit bids to buy, while sellers submit offers or asks to sell, with a rich rule structure for defining priority based on price, quantity, and arrival time. We describe here a simple version used in the experiment shown in Figure 1, in which subjects trade single units in sequence. Each bid (ask) is understood to represent a buy (sell) order for a single unit. The moment that the first standing bid is entered by a subject it is displayed on all monitor screens. Any new bid is admissible only if it specifies a higher price than the standing bid, and so on in sequence as new bids are entered.

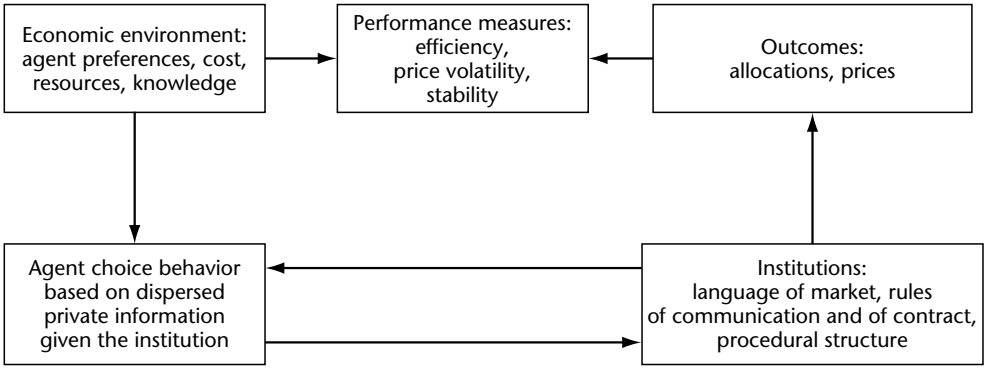


Figure 2. The components of every market: environment, institution and behavior.

Simultaneously, sellers are free to submit asks. When there is a standing ask, any new submission must specify a lower price. As soon as there is both a bid and an ask price, we have a bid/ask spread, say bid \$4, ask \$5.

The contract rule is simple: either a buyer accepts the standing ask price, or a seller accepts a standing bid. After each acceptance the auction ends and the computer waits for the submission of new bids and asks, as above. Note that the language of the market is bids, asks, and bid or ask acceptances. These are the only four messages that can be submitted by any subject agent to the trading system and the above filtering rules are applied to the messages as soon as they arrive.

A DOUBLE AUCTION EXPERIMENT: ENVIRONMENT AND BEHAVIOR

As we have seen, an economic environment is illustrated in the left panel of Figure 1. Notice that the demand crosses the supply at a range of market clearing prices, where demand = supply = 10 units, given by the interval (356, 360). Any whole number in this interval is a CE price. Only you and I know this; the subjects in this experiment know nothing of these facts. What Buyer 2 (B2) knows is that he or she can buy up to four units profitably at any prices below the values 435, 370, 350 and 330 respectively. Similarly, each of the other buyers knows only their own values and each of the sellers knows only their own costs. Units not purchased or sold incur no penalty.

The subjects were inexperienced, meaning that none had previously been in a double auction experiment. In Figure 1 we plot the displayed bids, asks, and contracts in a time sequence within each trading period. We show here the data for only the first seven of the 15 periods.

The behavior shown in the right panel of Figure 1 is typical. Efficiency in each period is the ratio of the total profits earned by all subjects in the period divided by the maximum possible profits (the area between the demand and supply schedules to the left of the intersection at 10 units). Note that after the first period, efficiency is always above 95 percent, implying that in each period the market participants realized at least 95 percent of the surplus due to exchange. Efficiency is 100 percent if and only if all buyer units valued at 360 and above are purchased and all seller units involving cost of 355 or below are sold. Efficiency is reduced if less than 10 units trade; it is also reduced if buyer valuation units at 355 or less are purchased and/or any seller cost units of 365 are sold. With experienced subjects efficiency tends to be 100 percent within the first few periods, even for subjects facing new and unfamiliar environments.

But Figure 1 is a static environment with demand and supply repeated in each period without any change throughout the experiment. Does the double auction institution perform well in tracking random shifts, up or down in the values (costs) assigned to buyers (sellers)? The answer is illustrated in Figure 3 which plots all contracts, and their mean price, over 15 periods with CE prices shifted for each period as shown. Observe that the mean price tracks the random shifts in the CE very closely, showing how this institution solves the problem of rapid adaptation to changes. Figure 3 also neatly illustrates that price volatility in a market has two components: variation due to exogenous fluctuations arising from changes in the economic environment, and endogenous variation within the market as traders search for the new equilibrium during each period. The former is represented by the random shifts in the CE, while the latter is represented by the dispersion of contract prices around their mean in each trading period.

The example in Figure 1 is for a single, isolated market. Do the strong equilibrating properties, demonstrated repeatedly in such examples, also hold for multiple markets? Yes, an example with two interdependent markets is reported in Smith (2000, pp. 245–247). Also see Williams *et al*. (2000). In these markets what a buyer is willing to pay for commodity A depends upon the price of B – the demand for A and B are opportunity cost demands. The equilibrium is defined by four simultaneous nonlinear equations in the price and quantity produced for A and B. In 10 out of 15 experiments prices are within 1 percent of their equilibrium predictions in period 10. With no knowledge of the equilibrium, or of the underlying equations defining it, six buyers and six sellers, each motivated by profit, unintentionally 'solve' the equations to reach the optimal equilibrium outcome. In effect subject behavior coordinated by the institution combines to provide algorithms that yield a competitive equilibrium.

All these examples show that it is not necessary for individuals to have complete information about the economic environment to achieve equilibrium outcomes. Some scholars may argue that the complete information condition was intended to provide only a strong sufficient, not necessary, condition for equilibrium to obtain. This has been tested. There are examples showing that when complete information on supply and demand is given to all individuals, the market performance

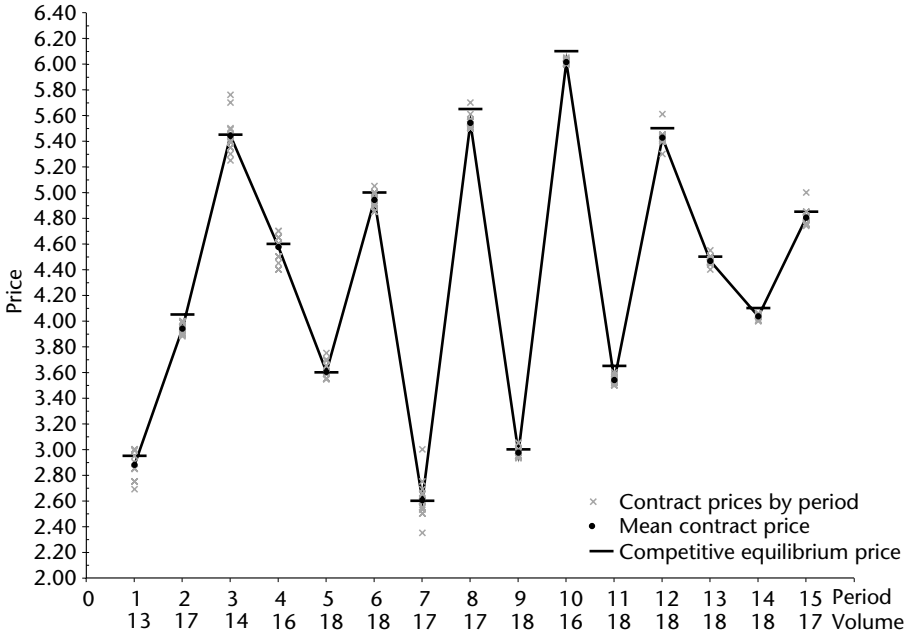


Figure 3. Effect of random shifts up or down in the supply and demand environment under double auction trading. Only the CE price was shifted, with CE volume always 18 units.

is worse than with private information (Smith, 1976/1991, pp. 103–105).

THE POSTED OFFER INSTITUTION

In ordinary retail trade the customer walks into the store (hardware, clothing, McDonalds) and observes a menu of take-it-or-leave-it price tags on each item offered for sale. Ketcham *et al*. (1984, reprinted in Smith, 1991) provide comparisons between the posted offer pricing and double auction institutions using identical environments. One of the environments held constant across the two institutions is the one exhibited in the left panel of Figure 1. Six experiments were run under each of the two institutions using this environment, using either an independent sample of eight subjects (five buyers and three sellers) or the same subjects for experienced sessions.

In Figure 4 we plot the mean double auction contract price for the experiment reported in Figure 1, and, on the same scale, the mean posted price (weighted by number of contracts at each price) for a matched comparison experiment, both using inexperienced subjects, and therefore strictly comparable; the institution is the only prominent treatment difference.

Observe that in every period in sequence the mean posted contract price strictly dominates the mean double auction contract price. Five of the six posted offer markets reported by KSW tend to converge to a price of 365, the cost of seller 3 in Figure 1, which is above the CE price range. For the double auction experiments it was the reverse: five converged to the CE price range, only one to 365. (The experimental environments all differed from each other by a random constant added to all values and costs, but all were equivalent when normalized on a standard CE price.)

In addition to prices being lower in the double auction, convergence to the CE price was more pronounced, the volume (number of units traded) was higher, and efficiency was higher. Thus as an exchange mechanism, the double auction dominates the posted offer in terms of efficiency and competitivity. Eventually, however, posted offer markets tend to converge to the CE, but the convergence is from initial prices that are above the CE price. In view of Figure 3 showing the good dynamic performance of the double auction, how does the posted offer compare? The answer is very poorly. Posted offer markets do not track shifts in demand well. Unlike the double auction such markets depend upon nonmarket sources of information concerning changes in the environment; the double auction does not because of its capacity for rapid adaptation. See Davis and Holt (1993, chap. 4) for further discussion and references.

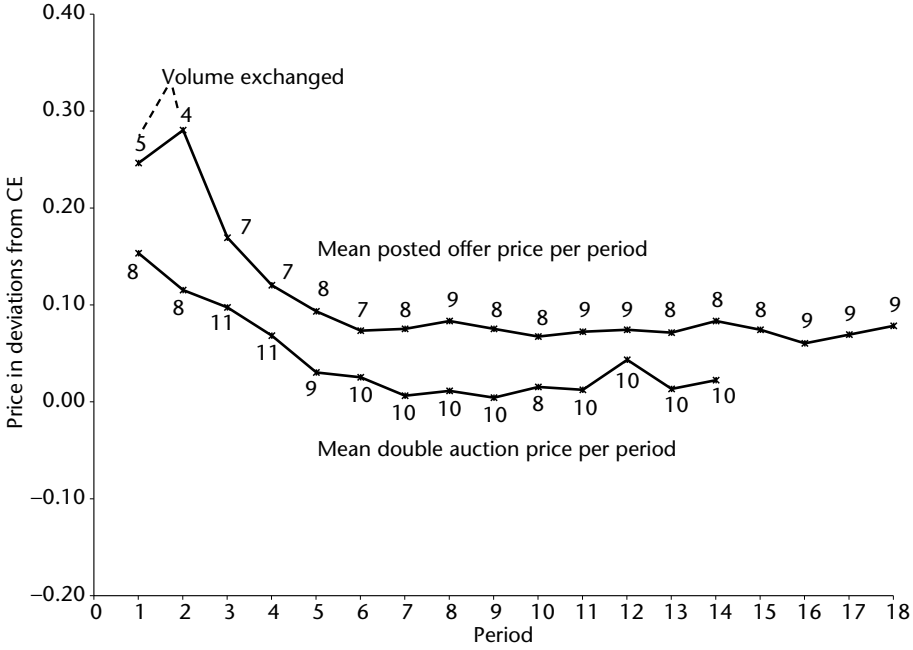


Figure 4. Plot of mean double auction price and volume (lower chart) for experiment shown in Figure 1, compared with posted offer mean price and volume (upper chart) with the same environment as in Figure 1.

THE CONCEPT OF A STRATEGY-PROOF EQUILIBRIUM

Although we can think of an allocation mechanism as an institutional procedure that allows the preferences of individuals to be mapped into final allocations, this abstract formulation does not take explicit account of the fact that preferences are private and unobservable, and institutions have to rely upon the messages reported by agents, not their true preferences. Consequently, the standard theoretical proposition is that it is possible for an agent to affect prices and outcomes in a market by strategically misreporting his or her preferences. Thus, in our example above of a buyer with a maximum willingness-to-pay of \$10, \$7, and \$4, respectively, for three units of a good – who believes sellers are willing to sell for less – might strategically bid for all three units at \$3 in an attempt to lower his or her purchase cost of the three units. Allocation mechanisms are actually mappings from preferences, and each agent's information or beliefs about other agents, into allocations. This state of affairs has motivated an intensive theoretical study of strategy-proof mechanisms designed to overcome the problem of strategic misrepresentation, but the results are negative and not encouraging. Thus, stated informally, 'an allocation mechanism is strategy-proof if every agent's utility-maximizing choice of what preferences to report depends only on his own preferences and not on his expectations concerning the preferences that other agents will report' (Satterthwaite, 1987, p. 519). This comes down to the strong requirement that each agent has a dominant strategy to report true preferences, and has led to impossibility theorems establishing the nonexistence of such a mechanism under certain minimal requirements.

Given these negative results, it is of particular interest to ask what people actually do in experimental environments in which the experimenter induces preferences on individual subjects so that the experimenter knows each agent's preferences, but the subjects know only their own preferences. Although it is possible that an agent can obtain an advantage by strategically underrevealing his/her demand or supply, whether or not such action is successful depends upon the actions – possibly countervailing – of others. In particular, has society stumbled upon institutions in which forms of behavior arise that approximate practical solutions to the problem of 'strategy-proofness' in economic environments with dispersed information?

The best-known example in which the answer to this question is 'yes' is the continuous double oral auction discussed above. Our theoretical understanding of why and how this is so is weak and

represents one of the outstanding unsolved problems in economic/game theory.

Are there other examples, offering good (if not perfect) solutions to the problem of achieving strategy-proof equilibria? If so, what are the strategic behavioral mechanisms that people adopt to solve it? A partial answer, based on what we have learned from experiments, is in the form of two versions of uniform price auctions: the uniform-price sealed bid-offer auction, and the uniform-price double auction (UPDA). For a more complete discussion see the chapters by Cason and Friedman, Friedman, and Wilson in Friedman and Rust (1991).

CONCLUSIONS

This article has briefly examined two historically common trading institutions – the continuous double auction and posted offer pricing. In each institution the experimental results show that order emerges out of an interaction between the choices of individuals with dispersed private information and the (property) rights to act specified by the institution. The institutions vary in terms of their exhaustion of the gains from trade, the speed and completeness of convergence to efficient, competitive outcomes, and the volatility of prices, but in both cases the participants are better off than if they were unable to trade.

Several propositions follow from these examples.

1. Many market institutions exist in the economy that are a complex product of cultural evolution, each invented by no one yet by everyone, and which exhibit the capacity to produce an exchange order from dispersed information.
2. What emerges is a form of 'social mind' that solves complex organization problems without conscious cognition. This 'social mind' is born of the interaction among all individuals through the rules of the institution.
3. In these institutions some are price takers, some price makers, some both. Hence, the idea that all must be price takers is neither necessary nor sufficient to yield an extended cooperative order of the market.
4. Participant knowledge of the circumstances of others is neither necessary nor sufficient to yield an extended order of cooperation.
5. The double auction, and other markets, provide rapid adaptation to random dynamic changes in individual circumstances.

Markets are rule-governed institutions representing algorithms that select, process, and order the exploratory messages of agents who are better informed about their personal circumstances than those of others. As precautionary probes yield to contracts, agents become more sure of what they must give in order to receive and the gains they can hope to capture. Out of this interaction between minds through the intermediary of rules the process tends to converge more-or-less rapidly to an equilibrium if one exists. The emergent order is invisible to the participants, unlike the visible gains they reap. They find out what they need to know to achieve optimal outcomes against the constraining limits imposed by the actions of others. The resulting order accommodates trade offs between the cost of transacting, attending, and monitoring and the efficiency of the allocations so that the institution itself generates an order that fits the problem it evolved to solve. Hence, the hundreds of variations on the fine structure of institutions, each designed without a designer to accommodate disparate conditions but all subservient to the reality of dispersed agent information.

References

Davis D and Holt C (1993) *Experimental Economics*. Princeton: Princeton University Press.

Friedman D and Rust J (eds) (1991) *The Double Auction Market Institutions, Theories, and Evidence*. Reading, MA: Addison-Wesley.

Hayek F (1984) The use of knowledge in society. In: *The Essence of Hayek*. Chicago, IL: The University of Chicago Press. Reprinted from the *American Economic Review* 1945.

Ishikida T, Ledyard J, Olson M and Porter D (2001) Experimental testbedding of a pollution trading system: Southern California's reclaim emissions market. In: Isaac M (ed.) *Research in Experimental Economics*. Amsterdam: JAI Press.

Rassenti S, Smith VL and Wilson B (2001) Using experiments to inform the privatization/deregulation movement in electricity. *Cato Journal* Spring/Summer.

Satterthwaite M (1987) Strategy-proof allocation mechanisms. In: Eatwell J, Milgate M and Newman P (eds) *The New Palgrave: A Dictionary of Economics*. London, UK: Macmillan Press.

Smith VL (1991) *Papers in Experimental Economics*. Cambridge, UK: Cambridge University Press.

Smith VL (2000) *Bargaining and Market Behavior*. Cambridge, UK: Cambridge University Press.

Williams A, Smith VL, Ledyard J and Gjerstad S (2000) Concurrent trading in two experimental markets with demand interdependence. *Journal of Economic Theory* **16**: 511–528.

Minimal Group Experiments

Intermediate article

Robert Kurzban, UCLA, Los Angeles, California, USA

CONTENTS

The 'minimal group paradigm' is an experimental method in which people are assigned to arbitrary groups, and then required to allocate rewards to members of their group or another group. The surprising result of these studies is that ingroup favoritism is elicited under these conditions, suggesting that merely placing people into different groups, even meaningless ones, leads to discrimination.

THE MINIMAL GROUP EXPERIMENTS

Henri Tajfel and colleagues were interested in the origins of discrimination, particularly when it was based on the group membership of the person against whom discrimination was being directed (Tajfel *et al*., 1971). At the time when they conducted their experiments, although it was well known that people often discriminated against members of other groups, the circumstances that were necessary and sufficient for such discrimination were less well understood. Tajfel *et al*. wanted to look more closely at what these conditions were.

Their experimental program was designed to begin with a situation in which discrimination was *not* expected to occur. Subsequently, changes could be made to the experimental context, and the experimenters could observe the features of the situation that elicited discrimination.

The first condition was to be minimal indeed. Participants were brought into the laboratory and asked to perform a relatively simple task: estimating the number of dots on a screen. They were then assigned to groups ostensibly based on their performance on this task. In one condition, participants were told that they were being grouped according to whether they were more or less accurate at estimating the number of dots. Because grouping was based on participants' skill at this task, which was presumed therefore to have a degree of meaning to the participants, it was hypothesized that this sort of grouping would lead to discrimination. In contrast, in a condition in which participants were grouped based on whether they overestimated or underestimated the number of dots, a division that had no obvious value either positive or negative attached to it, no discrimination based on this assignment of group membership was predicted. In fact, the assignment to groups in both conditions was made randomly, although participants were not aware of this. Because the assignment of participants to groups was based on these meaningless procedures, this experimental program came to be known as the 'minimal group paradigm'.

After the dot estimation task, participants were presented with a series of 'allocation tasks' in which they had to decide how to allocate money between two individuals. In some cases, both of these individuals were members of the same group. However, in the condition of interest to the experimenters, one of the two people was a member of the same group as the participant (an 'ingroup' member), and one was a member of the other group (the 'outgroup').

The allocation was done by a set of matrices. Each matrix consisted of two rows of 14 numbers corresponding to monetary rewards or punishments. Each row corresponded to the payoff to either an ingroup member or an outgroup member. Participants had to select one of the 14 columns for each matrix, and the corresponding values in the columns would be paid to the other participants. In no case could a participant allocate any money to himself or herself – it was always to another participant in the experiment.

The allocation decisions were structured so that the participant was required to make a series of tradeoffs. Consider Figure 1. In this case, a participant allocating the largest number of points possible (19) to the ingroup member would be forced to simultaneously allocate an even greater number of points (25) to the outgroup. This allocation would lead to the greatest number of points overall, but would lead to an outcome in which the outgroup member was favored over the ingroup member.

Ingroup	7	8	9	10	11	12	13	14	15	16	17	18	19
Outgroup	1	3	5	7	9	11	13	15	17	19	21	23	25

Figure 1. Matrix from the minimal group paradigm.

A series of matrices of this type generated a number of different possible tradeoffs. In each, the extent to which a participant favored the ingroup relative to the outgroup (or the reverse) could be assessed by looking at the column selected. In the above example, the number of columns to the left of the center of the matrix was a measure of how much the participant wanted to ensure the ingroup member received more than the outgroup member, sacrificing both total overall points allocated and the equity in allocation.

The surprising result was that when grouping was based on either accuracy or over- versus underestimation, participants allocated points with a bias toward their own group members. This contradicted the idea that categorization had to have some value (e.g. accuracy) in order for discrimination to occur.

In a second experiment, participants were asked to indicate their preference for paintings by two impressionist artists, Klee and Kandinsky. Participants were then told they were being assigned to one of two groups, each group consisting of people who preferred one of the artists. In this experiment, using a slightly different set of matrices for the allocation task, it was found that participants allocated less money to their ingroup member than they could have in order that the outgroup member received still less. Assignment to a group seems to have made the participants competitive, even when this competitiveness led to an overall drop in both the total amount of money allocated to a pair and the amount of money allocated to the ingroup member. It is important to point out, however, that ingroup favoritism and competitiveness were balanced by concerns about fairness – participants tended not to choose the extreme scores in the allocation task, but instead reached a kind of compromise in their motivations.

Since these original experiments, these basic results have been replicated a large number of times, and with a number of variations on the basic design. For example, the same type of ingroup bias occurs when participants are given the task of evaluating ingroup and outgroup members, rather than allocating money to them (Turner, 1978). A second important result is that these effects hold even when subjects are told that they are being randomly assigned to groups, instead of on the basis of the dot estimation task (Billig and Tajfel, 1973).

Perhaps most significantly, effects of the minimal group manipulation are observed even when the participants' own interests are at stake. Diverting from the procedures of most experiments of this type, Turner (1978) had his participants make allocation decisions in which the participant had to divide money between himself or herself and a member of either an ingroup or an outgroup. In this study, participants sacrificed their own monetary earnings to benefit the ingroup member significantly more than they did with the outgroup member. This showed that categorization into minimal groups affected distribution of rewards not only when two unknown, anonymous others were the recipients, but also when the participant's own outcomes were at stake.

EXPLANATIONS FOR MGP RESULTS: SOCIAL IDENTITY THEORY

Existing theory in psychology at the time could not easily explain these surprising results. A general view among many psychologists was that intergroup bias and discrimination was the result of genuine conflicts between members of groups for valued resources (Campbell, 1965; Sherif, 1967). However, in the minimal group experiments, because participants are allocating to other people, not themselves, this functional view of bias does not seem to apply. The people in the study could make themselves no better off, nor worse off, in making their allocation decisions. A new theory, one with more of an emphasis on the consequences of group membership *per se*, rather than functional relations between groups, was necessary.

The inadequacy of existing theories led to the development of social identity theory (SIT) (Tajfel and Turner, 1986). 'Social identity' was defined as 'those aspects of an individual's self-image that derive from the social categories to which he perceives himself as belonging' (Tajfel and Turner, 1986, p. 16). That is, social identity has to do with how the groups a person belongs to are part of the way that individual sees him or herself.

The focal element of SIT was the concept of self-esteem. The theory assumed that people want to have high self-esteem, and that an individual's social identity played an important role in determining self-esteem. So, belonging to a group that one viewed positively would increase self-esteem. In addition, evaluation of the individual's own

group was a comparative process. That is, people judged how good their group was by comparing it with other relevant groups.

According to this theory, then, when someone is a member of a group that they view unfavorably, this should diminish their self-esteem and the individual should be motivated to improve their social identity. This can be achieved by leaving the group that they view negatively, if it is possible. Alternatively, and more relevantly in the context of the minimal group experiments, the individual can try to improve their view of the group to which they belong, and, in particular, improve it relative to a group which they are comparing with their own group.

Social identity theory was able to account for the minimal group findings in a way that theories that depended on actual conflict of interest could not, because SIT implied that people are concerned with comparing favorably with other groups even when nothing except self-esteem was at stake. In these experiments, it is presumed that participants were using the monetary rewards to allow the allocators a way to compare favorably with the other group and to raise self-regard by discriminating against the outgroup.

An important additional conceptual element that has been incorporated by some researchers into SIT is the idea that categorizing an individual as in the MGP is not all that matters. For discrimination to occur, people must come to identify with the category in question. Further, the extent to which someone identifies with a particular category to which they belong can change with the situation. In some contexts, identity as a male or a female might be particularly relevant, while in others, national or religious identity might matter. Current construal of identity is therefore taken to be the critical determinant of discrimination, rather than the simple fact of category membership.

ADDITIONAL EXPERIMENTAL EVIDENCE

The minimal group paradigm has served both as the inspiration and as a testing ground for the ideas behind SIT. Michael Hogg of the University of Queensland in Australia and his colleagues (e.g. Hogg and Abrams, 1988; Hogg and Sunderland, 1991) have tried to clarify exactly what predictions SIT makes in the context of MGP studies and the extent to which existing evidence supports these predictions. In particular, they have argued that SIT can be separated into two corollaries. The first is that discrimination against the outgroup improves self-esteem. The second is the idea that when self-esteem is threatened, the individual is motivated to discriminate against the outgroup.

Unfortunately for exponents of the theory, evidence for these corollaries has been inconsistent and weak. In terms of the first, while some researchers have found that those who discriminate experience higher self-esteem (e.g. Lemyre and Smith, 1985), a careful experiment conducted by Hogg and Sunderland (1991) found that discriminating in a MGP allocation task did not raise participants' self-esteem. However, in the same experiment, participants who were made to feel low self-esteem by being told that they had performed poorly on an 'interpersonal empathy' task did indeed show more discriminatory behavior. While this finding tends to support the second corollary, other evidence contradicts it, including Crocker and McGraw's (1985) finding that, for certain groups, discrimination is greater among those with high self-esteem (cited in Hogg and Sunderland, 1991).

Ongoing research continues to examine the conditions under which the corollaries of SIT hold. There is growing consensus that whatever the relationship between self-esteem and discrimination turns out to be, it is not as simple and straightforward as initially conceived by social identity theorists. In particular, while it appears that being categorized as a member of a particular group as in the MGP can lead to discrimination, it does not necessarily do so, and the conditions under which it does are beginning to be better understood (Grieve and Hogg, 1999).

References

Billig M and Tajfel H (1973) Social categorization and similarity in intergroup behavior. *European Journal of Social Psychology* **3**: 27–52.

Campbell DT (1965) Ethnocentric and other altruistic motives. In: Levine D (ed.) *Nebraska Symposium on Motivation*, pp. 283–311. Lincoln: University of Nebraska Press.

Crocker J and McGraw KM (1985) Prejudice in campus sororities: the effects of self-esteem and ingroup status. Unpublished manuscript. Evanston, IL: Northwestern University.

Grieve PG and Hogg MA (1999) Subjective uncertainty and intergroup discrimination in the minimal group situation. *Personality and Social Psychology Bulletin* **25**: 926–940.

Hogg MA and Abrams D (1988) *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. London: Routledge.

Hogg MA and Sunderland J (1991) Self-esteem and intergroup discrimination in the minimal group

paradigm. *British Journal of Social Psychology* **30**: 51–62.

Lemyre L and Smith PM (1985) Intergroup discrimination and self-esteem in the minimal group paradigm. *Journal of Personality and Social Psychology* **49**: 660–670.

Sherif M (1967) *Group Conflict and Co-operation: Their Social Psychology*. London: Routledge & Kegan Paul.

Tajfel H, Billig M, Bundy R and Flament C (1971) Social categorization and intergroup behaviour. *European Journal of Social Psychology* **1**: 49–178.

Tajfel H and Turner JC (1986) The social identity theory of intergroup behavior. In: Worchel S and Austin W (eds) *Psychology of Intergroup Relations*, 2nd edn, pp. 7–24. Chicago: Nelson-Hall.

Turner JC (1978) Social categorization and social discrimination in the minimal group paradigm. In: Tajfel H (ed.) *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Behavior*, pp. 66–101. Oxford, UK: Blackwell.

Mixture Models of Individual Heterogeneity

Intermediate article

DO Stahl, University of Texas, Austin, Texas, USA

CONTENTS

Mixture models of individual heterogeneity are models of a population consisting of heterogeneous types of individuals, each type exhibiting a distinct probabilistic behavior, with the probability of each type equal to that type's proportion in the population.

INTRODUCTION

That all humans do not behave alike is obvious to even the most casual observer. How to represent this heterogeneity in a model that allows us to predict behavior is far less obvious. The simplist model of heterogeneity assumes that behavior is normally distributed about a mean for the population. An important feature of this model is that there is a single mode of behavior: i.e. if one graphed a histrogram of the behavior, one would find a single hump. We focus here on the countless cases where behavior is 'multimodal': a histogram of the behavior would have multiple humps. For example, when voting on the level of federal spending for education, Democrats might be a unimodal population, and Repulicans might be a unimodal population, but the mean desired spending level of these populations would be quite different, so Congress as a whole would be a bimodal population.

Formally, suppose there are $K+1$ unimodal subpopulations, which we will index $k=0,1,\ldots,K$. An individual member of subpopulation k will be refered to as a type-k individual. Suppose some member of the whole population faces a situation that calls for a behavioral response. Let s denote this situation and all the relevant data about the situation, and let $X(s)$ denote the set of possible behaviors in this situation. Let $P_k(x|s)$ denote the probability that a type-k individual will exhibit behavior x from the set $X(s)$. Note that x could be a whole dynamic sequence of behaviors. Finally, let α_k denote the proportion of the whole population comprising subpopulation k (or equivalently, the probability that a randomly drawn individual from the whole population is type k). Then, the probability that a random individual from the whole population will exhibit behavior x in situation s is given by

$$P(x|s) \equiv \sum_{k=0}^{K} \alpha_k P_k(x|s) \qquad (1)$$

Equation (1) is a canonical 'mixture model'.

When the probabilistic behavior of each subpopulation, $P_k(x|s)$, is predetermined, then only the population proportions, α_k, need be estimated from the behavioral data. Identifcation only requires that $P_k(x|s) \neq P_j(x|s)$, for $k \neq j$.

In other cases, the probabilistic behavior might be specified parametrically as $P_k(x|s, \beta_k)$, where β_k is a vector of parameters for the subpopulation k. For example, β_k might represent the mean and variance of a normal distribution. Since $\beta_j = \beta_k$ would yield identical probabilistic behavior, this parametric case generally requires identifying restrictions on the parameter space (such as $\beta_j < \beta_k$).

In the next section, we illustrate the application of mixture models with the level-n model of bounded rationality by Stahl and Wilson (1994, p. 5).

LEVEL-N MODEL OF BOUNDED RATIONALITY

Consider a two-player finite normal-form game in which the payoff to player i is U_{ijk} when player i chooses action j and the other player chooses action k. Let $A \equiv \{1,\ldots,J\}$ denote the set of actions available to both players. The Stahl–Wilson (1994) model begins with the assumption that a proportion, α_0 of the population, has no understanding of the game and by virtue of the principle of insufficient reason a type-0 individual is equally likely to

choose any action in A. Hence, $P_0(j|s)$ is the uniform distribution over A, and s in this context represents the data for the game (i.e. A and U). In the language of Stahl–Wilson, these players are called 'level-0' types.

A Bayesian rational player must have a belief about what the other player will do. The simplist noninformative model of other players is that they are level-0 types. A player who believes all others are level-0 types is called a 'level-1' type and chooses an error-prone best response to this belief. Specifically, define

$$y_{1ij} \equiv \sum_{k=1}^{J} U_{ijk} P_0(k|s), \tag{2}$$

which is the expected payoff to player i when choosing action j against a level-0 player. Then, the probabilistic choice function for a level-1 type is specified logistically as

$$P_{1i}(j|s,\beta_1) \equiv \exp(\beta_1 y_{1ij}) / \sum_{k=1}^{J} \exp(\beta_1 y_{1ik}), \tag{3}$$

where β_1 is the precision of a level-1 type; the higher the precision, the higher the likelihood the choice will be the best response to the belief, and the lower the precision, the more equally probable will be all the actions. Moreover, a logistic choice function has the property that the order of the choice probabilities corresponds to the order of the expected payoffs, y_{1ij}. The proportion of level-1 types in the population is denoted α_1, and for simplicity of presentation we implicitly assume the same proportion for both player 1 and player 2 (i.e. a single population model).

The level-n theory proposes a hierarchy of types in which a level-n type believes that all other players are level-k types with $k < n$. For simplicity of explanation and as an example, let us assume that a level-2 type believes that all other players are level-1 types. Then, the expected payoff to player i when choosing action j against a level-1 player is

$$y_{2ij} \equiv \sum_{k=1}^{J} U_{ijk} P_{-i1}(k|s,\beta_1), \tag{4}$$

where '$-i$' means 'the other player, not i'. The logistic probabilistic choice function is

$$P_{2i}(j|s,\beta) \equiv \exp(\beta_2 y_{2ij}) / \sum_{k=1}^{J} \exp(\beta_2 y_{2ik}), \tag{5}$$

where $\beta \equiv (\beta_1, \beta_2)$.

This simple three-type, level-n model yields a mixture model of the form:

$$P_i(j|s,\alpha,\beta) \equiv \alpha_0 P_0(j|s) + \alpha_1 P_{1i}(j|s,\beta_1) + \alpha_2 P_{2i}(j|s,\beta), \tag{6}$$

where $\alpha \equiv (\alpha_0, \alpha_1)$, and $\alpha_2 = 1 - \alpha_0 - \alpha_1$, leaving only four free parameters. This mixture model can be expanded in a straightforward manner to test for the presense of additional types in the population.

Identification Issues

Since all the types in equation (6) collapse to uniformly random choice when $\beta_n = 0$, we need to impose identifying restrictions so $P_0(j|s)$, $P_{1i}(j|s,\beta_1)$, and $P_{2i}(j|s,\beta_1)$ are distinguishable given the sample size of the observed data. A Monte Carlo simulation can be used to determine appropriate lower bounds for the β_n.

Even having imposed these parameter restrictions, for many games both level-1 and level-2 types will behave the same, so such games will be inadequate to identify the α_n parameters. The solution to this identification problem is to use a variety of games, so that each type predicts distinctly different patterns of behavior across all the games. Stahl and Wilson (1994) use 10 symmetric 3×3 games; which permits 3^{10} (59,049) distinct patterns of choice for an individual, and an underlying space of probabilistic behavior of dimension 2^{10}. While this approach creates more than enough possibilites to identify the parameters of the proposed mixture model, the curse of dimensionality renders it impossible in practice to use nonparametric methods to characterize the underlying distribution of behavior, and therefore specification tests vis-à-vis the true data generation process are hopeless.

Hypothesis Testing

On the other hand, likelihood ratio comparisons can be used to test alternative model specifications. For example, the hypothesis that there are no level-2 types in the population sampled is equivalent to restricting $\alpha_2 = 0$. However, there is a complication because 0 is on the boundary of the parameter space (see Self and Liang, 1987; Feng and McCulloch, 1996). Although the classical regularity conditions (as typically stated in advanced econometrics textbooks) are not met, the maximum likelihood estimators remain consistent. Self and Liang (1987) suggest that the true asymptotic distribution of the likelihood ratio statistic under the null hypothesis is a mixture of chi-squares with 0 and 1 degree of freedom. Since the right tail of the

density of any such mixture lies to the left of a chi-square (1) density, the conventional chi-square (1) test would be too conservative, increasing the p-value of the observed statistic and lowering the probability of rejection; hence, a rejection of the null hypothesis $(\alpha_2 = 0)$ using the conventional chi-square tests would hold under the true asymptotic distribution.

Alternative theories of behavior naturally suggest different types that can be easily added to the mixture model. For example, when each game has a unique pure-strategy Nash equilibrium, a 'Nash' type can be added. If the Nash type is specified as putting probability one on the unique Nash equilibrium, then it is highly likely that one will be able to reject the hypothesis that there are pure Nash types in the population, because just one non-Nash choice by an individual would imply a zero probability of being such a pure Nash type. In general, theories that make extreme predictions are easily rejected, and beg to be augmented with a theory of 'errors'.

There are two simple and reasonable theories of errors. The first entails uniform trembles: with probability $(1 - \varepsilon)$ the individual chooses the Nash equilibrium, and with probability ε any action is equally likely to be chosen. This specification introduces an additional parameter to be estimated. The second model of errors is prior-based: as in equation (2), we define the expected payoff of each action given the prior belief that all other players will choose the Nash equilibrium. Then, we define the logistic Nash choice function as in equation (3). Again, one parameter is introduced, but now the choice probabilities are positively correlated with the expected payoffs. Haruvy and Stahl (1999) find that the logistic theory fits laboratory data much better than the uniform tremble theory. When a game has multiple Nash equilibria, they find that the logistic theory using the prior belief that each Nash equilibrium action is equally likely fits the data much better than any equilibrium selection criteria (including payoff dominance, risk dominance, and security).

Given the huge variety of behavior that is possible, one might expect that adding virtually any type to a mixture model will improve the fit. However, we have often failed to reject the hypothesis that certain types are absent. For instance, Stahl and Wilson (1995) failed to reject the absense of a rational expectations type, Haruvy *et al*. (1999) failed to reject the absense of a maximin type, and Haruvy and Stahl (1999) failed to reject the absense of payoff-dominance and risk-dominance types. All of these tests had adequate power.

As pointed out, using a variety of games to identify the model parameters creates the potential for enormous behavioral diversity and highlights the poverty of typical laboratory sample sizes. Given this diversity, it is often possible to reject the hypothesis that the parameters that maximize the likelihood of one small sample (of 10–30 individuals) are the same as the parameters that maximize the likelihood of another similarly sized sample. However, we caution the reader that these rejections are artifacts of overfitting small samples. In the absense of any predetermined, observeable criteria for distinguishing among different groups of laboratory subjects, we learn nothing useful from separate parameter estimates for each group. Rather, it is better to pool all the data and obtain one estimate for the general population.

Posterior Probabilities of Individual Types

Given parameter estimates $(\hat{\alpha}, \hat{\beta})$, Bayes theorem allows us to compute the posterior probability that any given individual is type-n, denoted α^p_{ni}.

$$\alpha^p_{ni} = \hat{\alpha}_n P_{ni}(x_i|s, \hat{\beta}_n) / \sum_{k=0}^{K} \hat{\alpha}_k P_{ki}(x_i|s, \hat{\beta}_k), \tag{7}$$

where x_i stands for the choices of individual i over all the games. Stahl and Wilson (1995) show how to modify this formula to account for the uncertainty inherent in the parameter estimates. They also find that 38 of 48 participants in their experiments can be identified with one type (i.e. $\alpha^p_{ni} > 90$ percent for some n). We suggest that this would hardly have been the case if the level-n model were not capturing a significant part of the true data-generating process.

Constancy of Types

The mixture model discussed and estimated assumes that individuals are one type for all situations they face. The above results on the posterior probabilites supports that assumption. However, we need to consider alternatives to obtain a more rigorous conclusion. One alternative is the hypothesis that individuals draw their type from the population of types (charactertized by the α_k's) independently for each game. This implicitly entails that all individuals are ex ante alike. This ex ante homogeneity hypothesis is strongly rejected by our data.

Consider instead the hypothesis that a type-k individual is highly likely to behave in the typical

way but with some small probability can behave like another type. Specifically, suppose that with probability $(1-\varepsilon)$ the individual behaves like a type-k but with probability ε is equally likely to behave like any type. For the Stahl–Wilson (1995) data, we found a statistically significant improvement in the maximized log-likelihood with an estimate of $\hat{\varepsilon} = 0.05$. In other words, individuals appear to be true to one type of behavior 95 percent of the time, and otherwise tremble to other types of behavior. Since the level-0 type is part of the model, not surprisingly, allowing type trembles lowers the estimated proportion of the population that is type-0. Another attractive feature of this type-tremble mixture model is that it feeds easily into the rule learning framework of Stahl (2000, 2001).

EQUIVALENT MIXTURE MODELS

For any distribution of behavior, $P(x|s)$, there are obviously innumerable ways to represent that distribution as a mixture model, equation (1). All of these are equivalent mixture models, and as such none can be rejected in favor of any other. Does this fact mean that mixture models are nonfalsifiable and hence unworthy of scientific research? The answer is no for three reasons.

First, the true data-generating process may be a mixture of types, in which case clever ways of isolating subpopulations of types would lead to falsifiable predictions. Second, representing the data-generating process as a mixture is a constructive approach that starts with archetypes and aggregates to population behavior. This process does produce falsifiable hypotheses about the constituent types and also provides a practical means of making predictions for novel situations. Third, the curse of dimensionality prevents us from obtaining a full characertization of $P(x|s)$, while the mixture model allows us to construct an approximation from simple constituent types.

Furthermore, when faced with two competing mixture models, it is always possible to create an encompassing model and use nested hypothesis tesing to select the best model. See, for example, Haruvy *et al*. (1998).

INDIVIDUAL VERSUS POPULATION MODELS

While psychologists are primarily interested in the behavior of individuals (even in social settings), other social scientists such as economists and sociologists are more interested in aggregate population behavior. Indeed, in many applications, the economist or sociologist may have only aggregate data. Although theoretically models of individual behavior can be aggregated to produce models of population behavior, there are often informational and computational constraints to exact aggregation. For example, because each individual has a unique history at any point in a repeated game, exact aggregation must keep track of all these unique histories, which grow in number exponentially with time. Models which entail integration over possible histories can be computationally infeasible. There is also the information conservation principle which states that a given sample of data contains only so much information – if you use that data to find the best fit for models of individual behavior, the information gained will not give you the best fit for models of population behavior.

The mixture models discussed above can be applied in a straightforward manner to population data. The only difference is in the construction of the likelihood function for the data. In an individual model, one first computes the likelihood of the behavior of each individual for all the games by type (a product of probabilities over the games), then sums these type-conditional likelihoods, and finally takes the product of these summed likelihoods over all individuals. In a population model, one first computes the sum of the type-conditional probabilistic choice functions for each game, then computes the multinomial probability of the observed aggregate choices using the summed probabilities, and finally takes the product of these multinomial likelihoods over all the games. The latter model does not impose the restriction that an individual's behavior is of one type for all games. Since the likelihood functions are different, the maximum-likelihood estimates of the individual model will differ from the maximum-likelihood estimates of the population model.

CONCLUSION

Human behavior is often multimodal, so we need multimodal models to represent and predict such heterogeneous behavior. The mixture model is ideally suited for this purpose. The major challenge is the specification of constituent subpopulation types. In accordance with the scientific method, we advocate theory-based hypothesis generation and testing. Since the best mixture model at any point in time will be an approximation of the true data-generating process, there will always be the possibility of discovering a better approximation.

References

Feng Z and McCulloch C (1996) Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B* **58**: 609–617.

Haruvy E and Stahl D (1999) Empirical tests of equilibrium selection based on player heterogeneity, http://www.eco.utexas.edu/faculty/Stahl/experimental/HS99.pdf

Haruvy E, Stahl D and Wilson P (1998) Modeling and testing for heterogeneity in observed strategic behavior. *Review of Economic Studies*, forthcoming.

Haruvy E, Stahl D and Wilson P (1999) Evidence for optimistic and pessimistic behavior in normal-form games. *Economics Letters* **63**: 255–259.

Self S and Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**: 605–610.

Stahl D (2000) Rule learning in symmetric normal-form games: theory and evidence. *Games and Economic Behavior* **32**: 105–138.

Stahl D (2001) Population rule learning in symmetric normal-form games: theory and evidence. *Journal of Economic Behavior and Organization* **45**: 19–35.

Stahl D and Wilson P (1994) Experimental evidence of players' models of other players. *Journal of Economic Behavior and Organization* **25**: 309–327.

Stahl D and Wilson P (1995) On players' models of other players: theory and experimental evidence. *Games and Economic Behavior* **10**: 213–254.

Neuroeconomics

Intermediate article

Kevin McCabe, George Mason University, Fairfax, Virginia, USA

CONTENTS

Neuroeconomics is the study of how the embodied brain interacts with its external environment to produce economic behavior. Research in this field will allow social scientists to better understand individuals' decision making, and consequently to better predict economic behavior.

INTRODUCTION

Recent breakthroughs in neuroscience models and technologies allow us to study *in vivo* brain activity as individuals solve problems involving tasks such as making choices between alternative actions, forming expectations about the future, carrying out plans, and cooperating, producing, investing and trading with others. Knowledge of how the brain interacts with its environment to produce economic behavior will allow social scientists to better understand the variation both within and between individuals' decision making, and consequently to better predict economic behavior. In addition, understanding how the brain processes information can facilitate the building of economic institutions that better serve as extensions of our minds' capacities for social exchange.

A FRAMEWORK FOR DECISION MAKING

Figure 1 shows an abstract flowchart of the steps involved in decision making. Information inputs include internal states associated with the homeostasis of the decision maker, sensory information from the external world, and prior knowledge of the decision maker. This information is used to form subjective beliefs about the likelihood of different outcomes and an objective that the decision maker will try to attain. Choices interact with events outside the control of the decision maker to produce outcomes, which are then valued. Feedback allows the decision maker to improve his or her prior knowledge and leads to a return to homeostasis. Outcomes from choices may also affect future events and consequently influence new sensory data.

We can model the ultimate decision-making system as one that results in choices that maximize the genetic fitness of the organism as studied by sociobiology and evolutionary psychology. This leads to hypotheses about the neuronal encoding of decision-making variables that can be tested with a variety of neuroscientific methods.

DECISION MAKING BY NEUROLOGICAL PATIENTS

A famous early case illustrating the importance of prefrontal lobe lesions in decision making is that of Phineas Gage (Harlow, 1848). More recently, Bechara *et al*. (1997) have studied patients with similar ventromedial prefrontal damage in an individual-choice problem called the 'gambling task'. In this task, a subject is asked to choose cards from one of four decks starting with 40 cards each. Each deck has a fixed pay-off per card of $50 in decks C and D and $100 in decks A and B. Behind each card is a cost, ranging from $0 to $1250. The subject

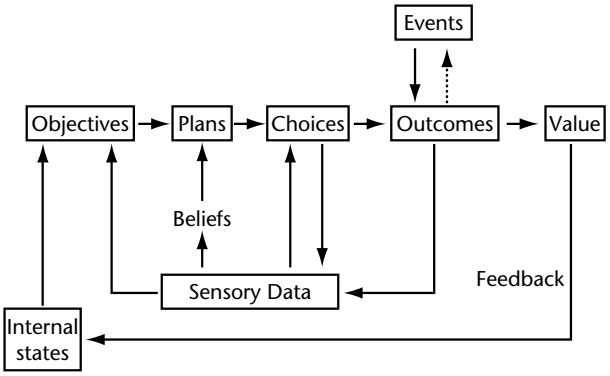


Figure 1. The steps involved in decision making.

must choose a card with a fixed pay-off, only then to learn the cost, if any, associated with the card. The subject starts with 2000 fictional dollars. On each draw, the fixed pay-off minus the cost is added to the initial amount. Although subjects did not get paid the final amount of dollars, it is assumed that they were motivated to do as well as possible in maximizing their fictional earnings.

While the subjects learn fairly quickly the location of the $50 and $100 decks, they must also learn which decks have positive net pay-offs. The decks are designed by the experimenter to have large punishments in the $100 decks, resulting in an overall loss if decks C and D are played for too long. The punishments in the $50 decks are much smaller resulting in an overall expected gain if decks A and B are played for long enough. Thus, the subject must learn to ignore the favorable signals of the $100 decks and instead play the $50 decks.

The typical behavior of a control subject is to shift play largely to the $50 (C and D) decks by period 60. By contrast, patients with ventromedial prefrontal damage do not shift away from the disadvantageous decks (A and B). Further research helped eliminate working-memory impairments as the reason for poor performance (Bechara *et al*., 1997). However, subjects with bilateral amygdala damage not only show similar behavioral impairment in the gambling task, but they also fail to show skin conductivity responses to rewards, punishments, or anticipation.

This research leads Bechara *et al*. to hypothesize that the amygdala couples a stimulus configuration with a 'somatic state' triggered by primary reward or punishment. The ventromedial prefrontal cortex is then responsible for coupling a strategy with a somatic state elicited by beliefs about how that strategy may produce outcomes.

To what extent do patients with similar ventromedial damage have difficulty with real-world decision making? In a series of articles, Jordan Grafman and colleagues have looked at the role of prefrontal cortex in allowing individuals to solve complex decision problems involving sophisticated planning and feedback. Goel *et al*. (1997) study the performance of ten patients with frontal lobe lesions in a financial planning task in which they were asked to prepare a budget, with planned projections, for a fictional couple's cash flow, in order to solve the following problems. First, get the couple out of the red. Second, allow them to buy a house in the next two years. Third, send their two children to college (in 15 to 20 years). Fourth, allow them to retire at the age of 65 (in 35 years). When compared to ten normal controls, the patients take much longer than normal subjects in structuring the problem, and in inferring abstract principles from particular instances, both of which are necessary in formulating a plan. Patients are also bad at processing feedback and judging performance, causing them to finish the task before the four problems have been adequately solved.

CHOICES BETWEEN COMPETING ALTERNATIVES

Economists have long studied decision making as the maximization of objective functions, such as utility or profit, subject to individual budgetary constraints. An important question is whether specified collections of neurons encode decision variables critical for optimization. For example, consider the simple utility-maximization problem of choosing x_1* and x_2* so as to maximize $U(x_1, x_2)$ subject to the budget constraint $p_1x_1 + p_2x_2 \leq m$, where x_1 and x_2 represent quantities of two different goods, p_1 and p_2 represent the prices of the goods, and m represents the money the person has at his or her disposal.

In Figure 1, the step marked 'choices' involves the ability to balance the relative gains of the two goods against the relative costs. The marginal utility of good 1 is the change in utility that the decision maker would receive for an additional unit of good, holding the amount of good 2 constant, or more formally, the partial derivative of the utility function, i.e. $U^1(x_1, x_2) = \partial U(x_1, x_2)/\partial x_1$, and similarly for good 2. Relative gains can then be measured as the ratio of marginal utilities $U^1(x_1, x_2)/U^2(x_1, x_2)$. A necessary condition for (x_1^*, x_2^*) to be a solution to this utility-maximization problem is that the ratio of marginal utilities should equal the ratio of the costs of acquiring the goods, i.e. $U^1(x_1^*, x_2^*)/\ U^2(x_1^*, x_2^*) = p_1/p_2$.

In making trade-offs between alternatives, we would expect the objective function to be sensitive to the relative reward value of the alternatives, independently of the alternatives in question. Is there any evidence that the brain encodes this kind of information? Tremblay and Schultz (1999) show that the firing rates of orbitofrontal neurons in two *Macaca fascicularis* monkeys were modulated by the relative reward values of different food (or, in separate trials, different drink) items. For example, the authors knew *a priori* that the monkeys preferred a piece of banana (B) to a piece of apple (A), which was in turn preferred to a piece of lettuce (L). Within a block of choice, symbols for two of the rewards were alternatively

shown on the left or right side of the screen. After a delay the monkey had to press a lever indicating where the picture was presented in order to get the reward symbolized. In different blocks, the monkeys were presented with all combinations of rewards.

Prior to choice, the same neurons in the orbitofrontal cortex would fire more frequently when the more desired food item symbol was displayed, compared with the less desired food item. Similarly, other neurons would fire more frequently when the symbol for the less desired item was presented. The authors conclude that such neurons encode relative preferences of food items, independently of the items themselves. However, it is still an open question how neurons encode the balancing of relative gains with relative costs.

MONETARY REWARD

Notice that money is not directly valuable to decision makers. But with fixed prices p_1 and p_2, we can always write $x_1^* = m_1/p_1$,where m_1 is the amount of money budgeted for good 1. In economics experiments human subjects are paid money as their salient reward. This raises the question: is there neuronal activity specifically associated with decision making over money? In general, neuroscientists predict that rewards are processed in the ventral striatum and orbitofrontal regions of the brain (Schultz, 2000). Using positronic emission tomography, Thut *et al*. (1997) studied brain activation in ten humans who received either a monetary reward or a simple 'OK' reinforcer for performance on a delayed 'go–no-go' task. The monetary reward resulted in significantly higher activation of the dorsolateral and orbitofrontal cortex, and also involved the midbrain and thalamus.

In an event-related functional magentic resonance imaging study, Knutson *et al*. (2000) studied brain activation in 12 subjects who engaged in a monetary-incentive delay task similar to a task orginally designed for monkeys (Schultz *et al*., 1997). In the human task, subjects were shown a cue indicating that they would receive some level of monetary reward ($0.20, $1.00, or $5.00), or nothing, or a punishments represented as a monetary loss ($0.20, $1.00, or $5.00). This was followed by a random delay of 2000–2500 ms, and then the appearance of a white target lasting 160–260 ms. Subjects received the relevant reward if they had got that reward cue and pressed a response button while the white target was visible. Subjects paid the relevant punishment if they were cued for that punishment and they failed to press the response button when the white target was visible. Levels of activation in the nucleus accumbens increased as anticipated rewards (but not punishments) increased.

CHOICES UNDER UNCERTAINTY

A simple example is the choice between two gambles g_1 and g_2. If the gamble g_1 is played, the decision maker gains $10 or $6, each with probability ½. Playing the gamble g_2 results in a gain of $14 or $2, each with probability ½. We assume that the decision maker prefers more money to less, so the utility function U satisfies $U(\$14) > U(\$10) > U(\$6) > U(\$2)$. In an experiment, these probabilities are likely to be shown to the subject in frequency terms, such as an urn containing 50 red balls and 50 blue balls. A subject is asked to pick a gamble. A ball is then randomly chosen. A red ball is worth $10 if g_1 is chosen and $14 if g_2 is chosen; and a blue ball is worth $6 if g_1 is chosen and $2 if g_2 is chosen.

Returning to our example, the expected value (calculated as $\frac{1}{2}(\$10) + \frac{1}{2}(\$6)$ for gamble g_1) is $8 for both gambles. However, if, in comparing the two gambles, the decision maker is more concerned about the potential loss of $4 ($6 − $2) when a blue ball is chosen, compared with the potential gain of $4 ($14 − $10) when a red ball is chosen, then the decision maker may choose to play the 'less risky' gamble g_1. This decision is consistent with an expected-utility calculation of the form $EU(g_1) = \frac{1}{2}U(\$10) + \frac{1}{2}U(\$6) > EU(g_2) = \frac{1}{2}U(\$14) + \frac{1}{2}U(\$2)$, which holds when the function U is concave, implying that $U(\$6) - U(\$2) \geq U(\$14) - U(\$10)$.

Suppose now that the dollar amounts in both gambles are replaced with negative amounts of the same magnitude: for example, $\bar{g}_1$ is a gamble whereby a subject loses $10 or $6, each with probability ½, giving an expected loss of $8. While $\bar{g}_1$ and $\bar{g}_2$ both have an expected loss of $8, our subject may prefer to play the gamble $\bar{g}_2$ in order possibly to lose only $2, and take the risk of losing $14. In this case we would call the person a 'risk-taker', since he or she prefers the possibility of avoiding a $4 loss (by going from $6 to $2) to the prospect of losing an additional $4 (by going from $10 to $14). In this case, the utility function over losses is convex to the origin, satisfying $U(-\$10) - U(-\$14) < U(-2\$) - U(-\$6)$. Behavioral data support the hypothesis that individuals are risk-taking over losses, and risk-averse over gains (Friedman and Savage, 1948; Tversky and Kahneman, 1986). (See Figure 2.)

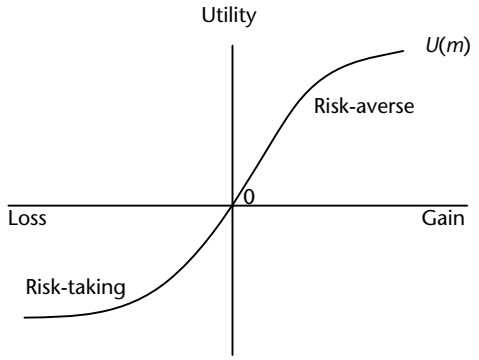


Figure 2. Utility $U(m)$ as a function of monetary gain ($m > 0$) or loss ($m < 0$). In experiments, subjects tend to apply a convex utility function for gains and a concave utility function for losses (see Friedman and Savage, 1948).

What objective function will the decision maker use to choose between the gambles? We would expect it to be sensitive to both probabilities and outcomes. Is there any evidence that the brain encodes this kind of information? Platt and Glimcher (1999) showed that the firing rates of lateral intraparietal neurons, in the posterior parietal cortex of rhesus monkeys, were modulated by the ratio of expected gains between two (juice reward) gambles, and consistent with the choices of the monkeys. In these cued saccade tasks, the experimenters varied both the gain associated with a particular response and the probability that a specific task would be required. As these variables were increased there was a commensurate increase in firing rates of specific lateral intraparietal neurons. The effects of both expected gain and outcome probability were strongest just before the monkey knew which response would be rewarded. Finally, in a free-choice task, where the monkey was not instructed, behavioral response frequency (towards the more favorable outcome) and the firing rate of posterior parietal neurons were correlated.

Using event-related functional magnetic resonance imaging, Breiter *et al*. (2001) studied brain activation in 12 humans who received monetary rewards or losses based on the outcome of a gamble chosen by the experimenter. Three different gambles were presented visually as spinners to subjects in a pseudorandom sequence. Each spinner had three equally likely outcomes. There was a 'good' spinner with outcomes of \$10, \$2.50 and \$0, an 'intermediate' spinner with outcomes of \$2.50, \$0 and −\$1.50, and a 'bad' spinner with outcomes of \$0, −\$1.50 and −\$6. Subjects were paid the cumulative results from the gambles they played. This allowed the experimenters to examine the activation associated with reward expectancy, prospect stage, and the actual outcome stage. They observed a broadly distributed set of brain regions active in both the prospect and outcome stages of the gamble. They also observed significant hemodyanmic response in the sublenticular extended amygdala and orbital gyrus, with both regions of interest rising monotonically with the expected value of the spinner, that is, from the bad spinner to the intermediate spinner to the good spinner. Finally, they observed some evidence for hemispherical specialization, with the right hemisphere predominantly active for gains, and the left hemisphere predominantly active for losses.

Using positron emission tomography, Smith *et al*. (2002) studied brain activation in nine humans who made choices between gambles resulting in monetary gains or losses based on the outcome of the gamble they had chosen. Within each block of choices, subjects chose several times between two different gambles. Each gamble had three possible outcomes, which depended on the random draw of a red, green, or blue ball from a container of balls at the end of the experiment. Each of the four blocks of choices consisted of either risky choices or ambiguous choices over either monetary gains or monetary losses. In the risky-choice treatment, subjects were shown a depiction of each gamble as a container with the number of balls of each color and the monetary payoff associated with each color, either positive (in the gain condition) or negative (in the loss condition). In the ambiguous treatment, the total number of balls was given, but not completely broken down by color, that is, the subjects could not be sure of the color of some of the balls in the container.

As expected from previous studies, subjects were risk-averse in the risky–gain condition and risk-taking in the risky–loss condition. Subjects showed ambiguity-aversion over both the loss and gain treatments. They showed a set of ventromedial activations consistent with the observations of Bechara *et al*. (1997) and Breiter *et al*. (2001) in the risky–gain condition. However, in the risky–loss condition, they showed a different set of dorsomedial activations.

STRATEGIC CHOICES WITH OTHERS

When the economic environment contains other individuals, their behavior must be anticipated in order to achieve good pay-offs. Figure 3 shows a

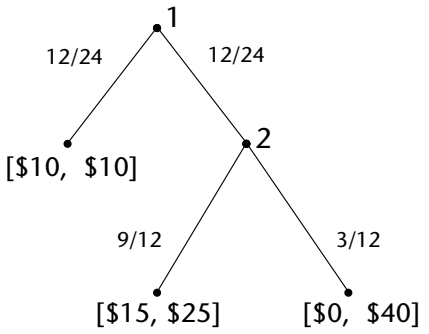


Figure 3. A simple game in which player 1 can either choose a pay-off vector of [$10, $10] or allow player 2 to choose between pay-off vectors of [$15, $25] and [$0, $40]. (The vectors represent the pay-offs to player 1 and player 2 in that order.) Out of 24 anonymously matched pairs, 12 player 1s chose to cooperate and 9 of their corresponding player 2s reciprocated.

simple example of a game in which cooperation can make both players better off. If player 1 moves right and player 2 moves left, then player 1 gets $15 and player 2 gets $25. However, player 2 may decide to move right in order to get $40. If player 1 believes that this is how player 2 will behave, then player 1 should move left, ending the game and resulting in $10 for each. In experiments where the game is played once between anonymously matched human subjects, half of player 1s moved right, and three-quarters of the player 2s who had the opportunity to move reciprocated by moving left. Therefore the expected gain for player 1 by moving right is greater than the $10 sure thing.

In a functional magnetic resonance imaging study, McCabe *et al*. (2001) studied brain activation in 12 humans who played sequential two-person games similar to that shown in Figure 3. Half the time they played as player 1, and half the time as player 2. Each time they played, their counterpart was either a computer playing a fixed probabilistic strategy, or a human who was recruited to play outside the scanner. Subjects were told before the game began whether they were playing the computer or the human.

On the basis of their individual plays, 7 of the 12 subjects were labeled as 'cooperators' while 5 were labeled as 'non-cooperators'. The cooperators all showed greater prefrontal activations in the computer condition than in the human condition, but with greater individual variation in and around BA-8. A conjunction analysis also suggested a common pattern of parietal, prefrontal, and frontal (BA-10) activations. By contrast, the non-cooperators did not generally display greater frontal activations in the human treatment than in the computer treatment. The observed activations in cooperators seem to be consistent with shared reciprocity intentions, resulting in both the inhibition of individual reward-seeking by player 2s and the inhibition of risk-avoiding behavior by player 1s.

References

Bechara A, Damasio H, Tranel D and Damascio AR (1997) Deciding advantageously before knowing the advantageous strategy. *Science* **275**: 1293–1295.

Breiter HC, Aharon I, Kahneman D, Dale A and Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**: 619–639.

Friedman M and Savage J (1948) The utility analysis of choices involving risk. *The Journal of Political Economy* **56**: 279–304.

Goel V, Grafman J, Tajik J, Gana S and Danto D (1997) A study of the performance of patients with frontal lobe lesions in a financial planning task. *Brain* **120**: 1805–1822.

Harlow JM (1848) Passage of an iron rod through the head. *Boston Medical Surgery Journal* **39**: 389–393.

Knutson B, Westdorp A, Kaiser E and Hommer D (2000) FMRI visualization of brain activity during a monetary incentive delay task. *Neuro Image* **12**: 20–27.

McCabe K, Houser D, Ryan L, Smith V and Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* **98**: 11832–11835.

Platt M and Glimcher P (1999) Neural correlates of decision variables in parietal cortex. *Nature* **400**: 233–239.

Schultz W (2000) Multiple reward signals in the brain. *Nature Reviews: Neuroscience* **1**: 199–207.

Schultz W, Dayan P and Montague R (1997) A neural substrate of prediction and reward. *Science* **275**: 1593–1599.

Smith K, Dickhaut J, McCabe K and Pardo J (2002) Neuronal substrates for choice under ambiguity, risk, gains, and losses. *Management Science* **48**: 711–718.

Thut G, Schultz W, Roelcke U *et al*. (1997) Activation of the human brain by monetary reward. *NeuroReport* **8**: 1225–1228.

Tremblay L and Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* **398**: 704–708.

Tversky A and Kahneman D (1986) Rational choice and the framing of decisions. *The Journal of Business* **59**: S251–S278.

Cognitive Assessment

Intermediate article

Jonna M Kulikowich, University of Connecticut, Storrs, Connecticut, USA
Patricia A Alexander, University of Maryland, College Park, Maryland, USA

CONTENTS

Cognitive assessment encompasses a wide variety of tests, tasks, and methods used to monitor and evaluate knowledge acquisition, strategic processing, and development of complex thinking.

INTRODUCTION

Measuring how we think and learn is not easy. Unlike behavior that can be observed directly, human cognitive processes are internal. Despite the difficulty in assessing cognitive processes, psychologists throughout history have designed many creative ways to represent them. In this article, we will introduce some of the assessments used by cognitive psychologists. We will consider assessment of domain-specific knowledge, strategic processing, and conceptual change. We will also discuss how researchers evaluate the effects of instruction, that is, academic achievement.

COGNITIVE PROCESSES

Anyone with an interest in science may wonder how experts like Albert Einstein or Marie Curie were able to make so many important contributions in physics and chemistry. Can we learn to think and solve problems like them?

It takes many years of hard work to become an expert. Experts possess vast amounts of knowledge related to their field of study, and this knowledge base has been acquired by working on complex problems, communicating with other experts, and making many mistakes. Fields of study, like physics and chemistry, are called 'domains'. Most researchers agree that domain-specific knowledge is of two primary kinds: 'declarative' and 'procedural' knowledge (Anderson, 1983). Declarative knowledge includes our knowledge of concepts, facts, and details. Experts possess an abundance of declarative knowledge, and this knowledge is structured in memory in representations, called 'schemata' in such a way that experts can quickly access the important principles of their domain. Procedural knowledge is knowing how to apply declarative knowledge. When we adjust the knob on a microscope to focus our view on a certain part of a cell, we use our procedural knowledge.

As well as possessing domain-specific knowledge, experts are also very strategic. 'Strategic learning' is effortful processing that helps experts link schemata and monitor and evaluate their progress while they are solving problems. Not all information can be immediately recalled or accessed. So, experts need a system of strategies that helps them put their existing knowledge to use. Some strategies are considered to be domain-specific because they are commonly used in one field of study while they may not be used to such extent in another field of study (mnemonics usually fall into this category). Other strategies can be used in any domain (for example, strategies for summarizing the main ideas in a passage of text).

We discuss how researchers have measured several important variables in the field of cognitive psychology. We begin with domain-specific knowledge and strategic learning and describe how these have been assessed. Then, we discuss how the study of expertise has influenced assessment of school students' performance. It is important to understand how students' cognitive processes develop over time. The topics of dynamic assessment and conceptual change pertain to developmental changes in students. Finally, we present some new developments in the analysis of student achievement. 'Achievement' relates specifically to what students know, and how they are able to solve problems, as a result of instruction. Performance-based assessment and portfolio

assessment are two testing techniques useful in the measurement of complex student achievement.

DOMAIN-SPECIFIC KNOWLEDGE

In studying how experts organize their knowledge and use it while solving problems, one of the simplest ways to assess domain-specific knowledge is by asking. Interview techniques are commonly used when studying experts. Psychologists develop lists of interview questions to ask experts, regarding what kinds of problems they like to solve, how long it takes them to solve problems, the challenges they face, and their evolving interests.

Two different types of interviews are popular: 'open-ended' and 'structured' interviews. In open-ended interviews, respondents are asked a general question that allows for a large variety of responses. Thus, we might ask Marie Curie: 'When did you realize that you were becoming an expert in chemistry?' Cognitive psychologists would hope to hear about how Curie realized that she knew a lot of information about chemistry that was more extensive than many of her peers. She might well mention that she could solve difficult problems quickly. She might mention that she recognized that her research was actually redefining the content of the field. In effect, Curie's discoveries, such as new chemical elements, became the declarative knowledge that others would have to know.

In structured interviews, cognitive psychologists seek very specific answers to detailed questions. These can be used in conjunction with observational records of the experts while they work. Thus, someone might observe an expert in chemistry like Curie while she is performing an experiment. After completion, a structured interview might be conducted, based on specific procedures that the expert used to solve a problem. Thus, we might ask her why she used one piece of equipment rather than another. Further, we might ask why and how she adjusted the Bunsen burner while heating a particular substance. We would compose these questions in an effort to assess the declarative and procedural knowledge of the expert based on our observations of her performance.

Expert–Novice Differences

Interview techniques are very useful when one has already identified an expert. But cognitive psychologists are also interested in assessing the differences between those who are experts and those who are not. Such work is usually referred to as the study of 'expert–novice differences' (Ericsson and Smith, 1991). Several methods of assessment have been very useful in studying these differences.

Concept maps

Concept maps allow cognitive psychologists to study how individuals organize their conceptual (primarily declarative) knowledge. Given a list of words that represent concepts in a domain, experts and novices are asked to create diagrams that show how these concepts are interrelated and how they lead to understanding of other concepts. Directional links show that concepts lead to one another, while nondirectional links simply show that concepts are related.

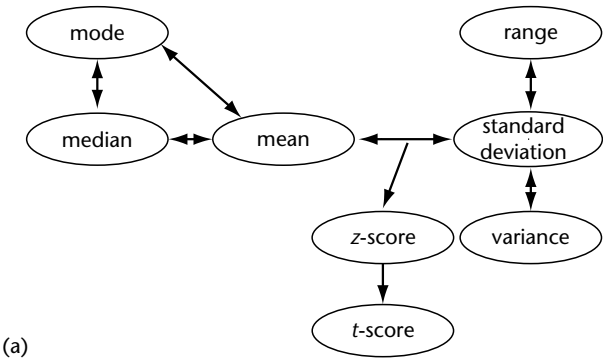
Expert statisticians, for example, know a lot about the normal distribution. Concepts related to the study of the normal distribution include central tendency, variability, and z-scores.

Figure 1 shows two concept maps. Figure 1(a) represents an expert's schematic understanding of the relations among a given list of statistical concepts, while Figure 1(b) represents a novice's understanding. The expert was also able to show how concepts lead to other concepts (for example, how means and standard deviations lead to computation of z-scores). The novice was not able to make these links. In general, cognitive maps demonstrate that experts organize their knowledge around important principles while novices rely on a rather fragmented organization of knowledge.

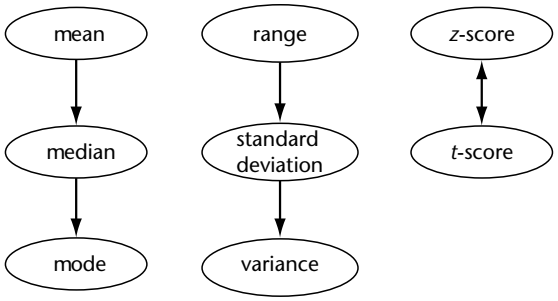
Think-aloud procedures

Think-aloud procedures can help in the assessment of procedural knowledge. Cognitive psychologists seeking to assess the processing that differentiates expert and novice performance might ask them to 'think aloud' while they are generating their solutions. Unlike interviews, which take place before or after the task, the think-aloud procedure takes place during the task. It can inform psychologists about such things as the information to which individuals attend as they solve problems (Ericsson and Simon, 1993). Further, it can provide hints as to the sequence of steps that problem solvers follow.

For example, if asked to think aloud while solving three essentially similar statistical problems all to do with examination results but expressed in different terms, experts in statistics are apt to study the problem structure in each task and compare those structures across tasks. Thus, an expert might say: 'All problems involve univariate distributions and require estimates of central tendency'. Novices, who have a more fragmented knowledge



(a)



(b)

Figure 1. Concept maps in the domain of statistics. The same list of eight concepts was given to an expert and to a novice. Directional links show that concepts lead to one another, while nondirectional links simply show that concepts are related. (a) The expert's concept map. (b) The novice's concept map.

base, do not tend to mention the principles that are common to all the problems as they think aloud. Instead, they tend to focus on surface features which are usually specific to a given problem. If we were to ask novices to think aloud while solving three statistics problems, they might mention that they all relate to experiences in school, particularly examinations; but they would probably not say that all involve the principle of central tendency.

Rating tasks

Analyzing interview and think-aloud protocols can take a long time. Studying them is like grading very long essay tests. A more simple way to assess cognitive processes is to ask individuals to rate the similarity between problems based on a specific characteristic. For example, experts and novices might be asked to rate the similarity between pairs of statistics problems as regards the principle of central tendency. Such ratings reveal that the experts see stronger association between pairs of problems that are essentially similar, compared with novices.

STRATEGIC KNOWLEDGE

Linking Domain-Specific Knowledge to Strategic Knowledge

Analogical reasoning relates the familiar to the unfamiliar. It is therefore relevant to the relationship between domain-specific knowledge and basic forms of strategic knowledge. Robert Sternberg (1977) has researched analogical reasoning and proposed that four component processes are used to solve problems: 'encoding', 'inferring', 'mapping', and 'applying'.

For example, consider a verbal analogy problem, such as: ' *boy* is to *man* as *girl* is to…?'. First, one has to define what the three terms *boy*, *man* and *girl* (in general, *A*, *B* and *C*) mean. This is the encoding step. Then, one has to infer the relationship between the *A* and *B* terms. The relationship between *boy* and *man* is that a boy is a male child while a man is a male adult. Then, one maps the relationship between the *A* and *C* terms. The common principle between boys and girls is that they are both children. The final component of analogical reasoning is to apply the rules learned in inferring and mapping to generate a solution, or construct a *D* term.

Essentially, the experts' ability to say that one problem is 'like' another problem because of the underlying theme of central tendency is a form of analogical reasoning.

Analogies and General Strategic Processing

Analogies are also relevant to general strategic processing, particularly when they are constructed in nonverbal, figural form. The same component processes (encoding, inferring, mapping and applying) seem to operate for these nonverbal analogies as for verbal analogies. Figure 2 is an example characteristic of those included on the 'Ravens Standard Progressive Matrices' tests (Raven, 1958). Items like this contain no content information, so they make for good measures of general strategic processing. Respondents have to track how the figures change in the matrix. Sometimes the strategic rules simply involve adding or deleting elements; sometimes the rules are very complex, involving rotations and transformations as well. Tracking complex rules requires considerable effort and concentration.

Psychologists have recently discovered that nonverbal analogy tests like the Ravens Standard Progressive Matrices assess not only general strategic

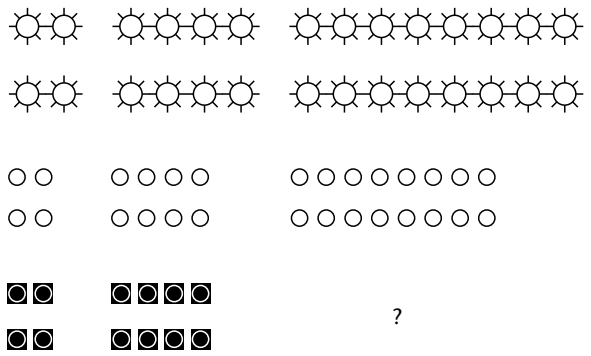


Figure 2. A nonverbal analogy matrix.

processing, but also metacognitive ability. Put simply, metacognition is thinking about our thinking. We use metacognition as we evaluate our progress, make plans, and consider possible new approaches to problem solving. As one works from matrix to matrix in solving the nonverbal analogies, one monitors how the rules change, from simple addition and deletion to more complex rotation and transformation problems.

The Link Between Metacognition and Complex Strategic Processing

For some, the analogy problems discussed above might seem easy to solve. Certainly, these types of problems are highly structured, and many items can be presented on one assessment tool. However, there are other types of tasks used to assess metacognition that are complex and less structured. These tasks reveal the degree to which project completion or problem solution require complex strategic processing. Consider the common situation where experts in many disciplines compose research papers to share their knowledge with others. Locating research material, taking good notes when reading research material, organizing notes into themes or summary statements, writing a draft of the paper so that main ideas are highlighted, and revising and editing the draft to produce a final version, are some of the many strategies that expert authors use to compose their research papers. Metacognition helps experts to select, monitor, and evaluate these strategies in an effort to produce papers of high quality that contribute new knowledge in their fields of study.

Cognitive psychologists have employed many methods to assess domain-specific and strategic knowledge. These methods have shed light on how expert processing differs from novice processing within and across domains. We will now turn our attention to assessment practices as they pertain to academic learning.

DYNAMIC ASSESSMENT

The goal of 'dynamic assessment' is to study cognitive development together with effectiveness of instruction. Dynamic assessment is therefore long-term: it is the study of one's potential to know more information in time. Dynamic assessment involves intermittent interventions of instruction. It is essential not only to provide feedback to a student, but also to instruct the student about concepts and strategies that contribute to proficient understanding. In their extensive review of dynamic testing, Grigorenko and Sternberg (1998) traced the roots of this assessment paradigm to the work of Lev Vygotsky (1962) in the 1930s. Vygotsky introduced the concept of the 'zone of proximal development', which reflects one's potential for growth in cognitive and social functioning facilitated through social interaction, such as learning from a teacher. The zone of proximal development is not to do with how one has developed, but with how one can become what one is not yet. Thus, rather than reflecting how an expert became proficient, it indicates how a novice might become proficient as a result of a synchronous interplay of evaluation and instruction.

CONCEPTUAL CHANGE

Like dynamic assessment, 'conceptual change' focuses on development. But while dynamic assessment attends to one's potential to learn, conceptual change attends to how knowledge is restructured as a result of learning and instruction (Pintrich *et al*., 1993). The study of human error patterns provides important insights as to how knowledge is restructured over time. From their study of the errors that participants produced given various types of tasks for various domains, Alexander *et al*. (1998) showed that the mistakes or errors we make are typically nonrandom; they occur in systematic patterns (e.g. we tend to make similar mistakes repeatedly); and they can provide the means for instructional intervention.

Multiple-Choice Questions and Error Patterns

Multiple-choice questions, for example, can be used to study how individuals choose options

that vary in their relatedness to the target domain. Consider the following example from the domain of statistics:

The standard deviation of a univariate distribution is:

1. The average distance from the mean. [correct statistics option]
2. The square root of the range. [incorrect statistics option]
3. A difference between a negative and a positive value. [incorrect mathematics option]
4. A common mistake made when spelling words. [incorrect non-mathematics option]

The four options have varying degree of correctness. All items on a test will employ the same hierarchy: there will always be a correct response in the domain of statistics; an incorrect response in the domain of statistics; an incorrect response in mathematics (not statistics) and an incorrect response outside the domain of mathematics.

Analysis of students' responses to such tests indicates that learners gravitate toward one error category rather than another, according to their familiarity with the domain. Such error patterns allow psychologists to profile learners according to their schematic knowledge. Thus, some students know a lot about the domain of statistics, and even when incorrect, they choose options that are representative of the correct domain. Other students do not have an initial sense of the domain: their errors indicate this. These error patterns are valuable to instructors, for they help them to modify the curriculum to meet the needs of learners with a wide range of abilities.

Error Patterns and Constructed-Response Tasks

Multiple-choice questions are referred to as choice tasks since the respondent is asked to check, circle, or mark an answer that is already presented on the test. Errors are also informative in so-called constructed-response tasks (Martinez, 1999). Alexander *et al*. (1998) demonstrated this in their study of domain-specific analogies. Using a categorization system to evaluate the types of response generated, they noted that even incorrect analogy solutions hinted at whether students lacked domain knowledge, strategic knowledge, or both. Table 1 presents this categorization scheme, using one analogy problem (in statistics) to illustrate the errors that inform researchers of the degree to which a student lacks knowledge. As with multiple-choice questions, error patterns help teachers to design intervention strategies. When such tests are applied at different times in studies of development, the transitions between error categories, leading toward correct responses, assist cognitive psychologists in their assessment of conceptual change: they can determine exactly when students' knowledge has been restructured to allow for principled use.

ACHIEVEMENT

'Achievement' refers to the assessment of knowledge acquisition, reading comprehension, or problem solving resulting from instruction. Traditionally, standardized assessments using multiple-choice formats have been used to evaluate student

Table 1. Categorization scheme for analysis of analogy problem responses

Category	*Example*	*Description*
No response	sample: statistic : : population:	No evidence of domain-specific or strategic knowledge.
Repetition	sample : statistic : : population : *statistic*	No strategic understanding of analogical relationships.
Non-domain response	sample : statistic : : population : *people*	Response is not related to target domain of statistics.
Structural dependency	sample : statistic : : population : *large statistic*	Response indicates use of component processes and some idea of the relationships among terms in statistics.
Domain response	sample : statistic : : population : *sigma*	Response indicates use of component processes and a response that is a term related to the correct response in statistics.
Target variant	sample : statistic : : population : *parametric*	Strong evidence of both domain-specific knowledge and strategic processing. The answer is a high-level error because it is merely a variation of the correct response. (The correct response should be a noun, not an adjective.)
Correct response	sample : statistic : : population : *parameter*	

achievement. However, poorly-constructed multiple-choice assessments encourage students to value rote memorization over more complex forms of thinking (Martinez, 1999). Constructed-response tasks in the forms of 'performance-based' and 'portfolio' assessments, are generally recommended by cognitive psychologists.

Performance-Based Assessment

Performance-based assessments require students to build objects, design plans, conduct experiments, deliver speeches and presentations, or stage plays and debates that are indicative of the kinds of tasks and procedures that experts perform. Psychologists develop 'rubrics' or scoring schemes, and train raters, or judges, to assign scores to student performance. Rubrics can take many forms, from simple checklists to complex rating scales. Imagine an instructor in physics who wants to assist students in their use of the internet to locate resources to conduct science experiments. The instructor finds a website that will be interesting to students. There are links at the site that allow students to study original scientific documents, visit virtual science museums, and communicate online with experts. Using the links on the site, the instructor hopes that students will learn to appreciate that finding information in the form of documents or dialogs with experts, and comprehending this information, is as important as working with equations or performing experiments in the laboratory. Knowing how to share this information with others in the form of a report is also important. Thus, the teacher creates the rubric presented in Figure 3 to evaluate student performance. This rubric is a simple rating scale that the teacher can use, the student can use for self-evaluation, or peers can use to provide feedback to one another.

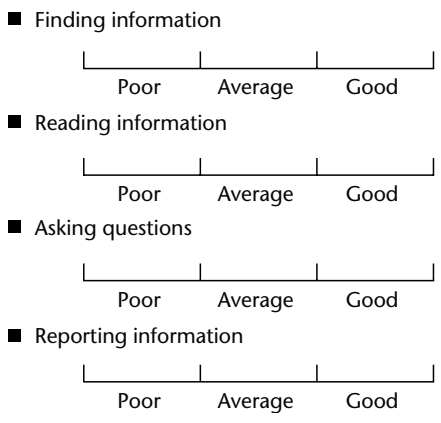


Figure 3. A possible rubric for a performance-based assessment of a research task.

Portfolio Assessment

Sometimes psychologists assess achievement over time using combinations of test formats that allow students to 'showcase' their best work. These 'portfolio assessments' not only provide information on the cognitive development of students, but also present examples of students' creative products and their interests. Interests and motivation are important to cognitive psychologists: according to models of academic development such as that proposed by Patricia Alexander (1997), one does not activate knowledge use and strategic processing without being motivated by one's domain of study. Indeed, being interested in the subject matter contributes as much to becoming an expert as does domain-specific knowledge, strategic processing, and metacognition. Portfolio assessments are a good way for students to express what is of interest to them, and they can show how these interests have changed over time in relation to changes in their knowledge.

FUTURE PROSPECTS

There are many more types of assessment than we have discussed in this article. With emergent technologies in areas like robotics and computer-aided and simulation design, and opportunities afforded by the internet, new ways to assess cognitive processes are emerging rapidly. Perhaps the only limitation of cognitive psychologists' ability to assess is their own creativity.

References

Alexander PA (1997) Mapping the multidimensional nature of domain learning: the interplay of cognitive, motivational, and strategic forces. *Advances in Motivation and Achievement* **10**: 213–250.

Alexander PA, Murphy PK and Kulikowich JM (1998) What responses to domain-specific analogy problems reveal about emerging competence. *Journal of Educational Psychology* **90**(3): 397–406.

Anderson JR (1983) *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Ericsson KA and Simon HA (1993) *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press. [Revised edition.]

Ericsson KA and Smith J (1991) *Toward a General Theory of Expertise: Prospects and Limits*. New York, NY: Cambridge University Press.

Grigorenko EL and Sternberg RJ (1998) Dynamic testing. *Psychological Bulletin* **124**(1): 75–111.

Martinez ME (1999) Cognition and the question of test item format. *Educational Psychologist* **34**(4): 207–218.
Pintrich PR, Marx RW and Boyle RA (1993) Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research* **63**: 167–199.
Raven JC (1958) *Standard Progressive Matrices: Sets A, B, C, D, and E*. London: H. K. Lewis.
Sternberg RJ (1977) *Intelligence, Information Processing and Analogical Reasoning: The Componential Analysis of Human Abilities*. Hillsdale, NJ: Erlbaum.
Vygotsky LS (1962) *Thought and Language*. Cambridge, MA: MIT Press. [First published 1934.]

Further Reading

Alexander PA (1992) Domain knowledge: evolving themes and emerging concerns. *Educational Psychologist* **27**: 33–51.
Chi MTH, Feltovich PJ and Glaser R (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science* **5**: 121–152.
Dole JA and Sinatra GM (1998) Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist* **33**(2–3): 109–128.
Embretson SE and Prenovost LK (2000) Dynamic cognitive testing: what kind of information is gained by measuring response time and modifiability? *Educational and Psychological Measurement* **60**(6): 837–863.
Ericsson KA, Patel V and Kintsch W (2000) How experts' adaptations to representative task demands account for the expertise effect in memory recall: comment on Vicente and Wang (1998) *Psychological Review* **107**(3): 578–592.
Feuerstein R, Rand Y, Jensen MR, Kaniel S and Tzuriel D (1987) Prerequisites for testing of learning potential: the LPAD model. In: Lidz CZ (ed.) *Dynamic Testing*, pp. 35–51. New York, NY: Guilford Press.
Hall BW and Hewitt-Gervais CM (2000) The application of student portfolios in primary-intermediate and self-contained-multiage team classroom environments: implications for instruction, learning, and assessment. *Applied Measurement in Education* **13**(2): 209–228.
Hunt E (1978) The mechanisms of verbal ability. *Psychological Review* **85**: 109–130.
Kelderman H (1996) Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement* **20**: 155–168.
Kulikowich JM and Alexander PA (1994) Evaluating students' errors on cognitive tasks: applications of polytomous item response theory and log-linear modeling. In: Reynolds CR (ed.) *Cognitive Assessment: A Multidisciplinary Perspective*, pp. 137–154. New York, NY: Plenum.
Shavelson RJ, Ruiz-Primo MA and Wiley EW (1999) Notes on sources of sampling variability in science performance assessments. *Journal of Educational Measurement* **36**(1): 61–71.

Instruction and Cognition

Introductory article

Richard E Mayer, University of California, Santa Barbara, California, USA

CONTENTS

During learning, instruction can guide cognitive processing, which in turn influences what is learned.

INTRODUCTION

Framework for Cognition and Instruction

What is the relation between cognition (in particular, cognitive processing during learning) and instruction? An important aspect of the relation concerns learning. Learning is a change in knowledge, and instruction is activity aimed at fostering this change. The premise of research on cognition and instruction is that if you want to improve instruction, it is useful to understand how people learn.

Figure 1 shows a framework for discussing cognition and instruction. Instruction (including instructional methods and materials) affects the cognitive processes of the learner during learning (including paying attention, mentally organizing the material, and mentally connecting the material with existing knowledge), which affects the knowledge changes in the learner (i.e., the learning outcome), which affects performance on tests. Instruction and performance are external and observable whereas cognitive processes and knowledge changes are internal and can only be inferred from performance. An important goal of research on cognition and instruction is to understand how instructional manipulations affect cognitive processing in the learner, and how this cognitive processing leads to changes in knowledge and, ultimately, performance.

Learning

'Learning' refers to enduring changes in a learner's knowledge that arise from the learner's cognitive processing of experiences. This definition contains: (1) learning results in a relatively permanent change (rather than a short-term change); (2) the change is caused by the learner's cognitive processing of his or her experiences (rather than by physical causes such as fatigue or drugs); and (3) the change occurs within the learner's mind (but must be inferred from changes in behavior). Examples include learning number facts, such as the fact that $5 \times 5 = 25$; learning to describe how a scientific system, such as the process of lightning formation, works; and learning to analyze a worldwide problem such as the consequences of global warming.

Instruction

'Instruction' refers to the teacher's manipulation of the learner's experiences that are intended to foster learning. This definition contains: (1) instruction involves the intentional creation of environments for learners to experience; and (2) instruction is intended to foster changes in the learner. Examples include drill and practice on arithmetic facts; multimedia presentations depicting how a lightning storm develops; and classroom discussions concerning the possible political ramifications of global warming.

THREE CONCEPTIONS OF LEARNING

The study of cognition and instruction is based on the search for an educationally relevant theory of learning – i.e. a theory of learning that links instructional manipulations with cognitive processing and knowledge change within the learner. The search for an educationally relevant theory of learning can be traced back to the work of E. L. Thorndike in the early twentieth century. During the past 100 years the conception of learning has been based on a progression of three metaphors: response

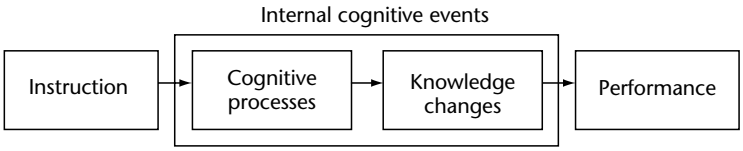


Figure 1. Four components in a model of cognition and instruction.

Table 1. Three conceptions of learning

	Learning as response strengthening	*Learning as information acquisition*	*Learning as knowledge construction*
Learning	Strengthening and weakening of stimulus–response associations	Adding information to memory	Building meaningful knowledge
Teacher	Dispenser of rewards and punishments	Dispenser of information	Cognitive guide
Learner	Recipient of rewards and punishments	Recipient of information	Sense maker
Instruction	Drill and practice	Textbooks, lectures and multimedia presentations	Guided exploration

strengthening, information acquisition, and knowledge construction. Table 1 lists the characteristics of each of these conceptions of learning.

Learning as Response Strengthening

The first conception listed in Table 1 is that learning involves the strengthening or weakening of the association between a stimulus and a response (S–R association), based on rewards and punishments. In this conception, the teacher becomes a dispenser of rewards and punishments, and the learner becomes a passive recipient of rewards and punishments. Following the four elements of Figure 1, the major instructional method is drill and practice, in which the learner is encouraged to make a short response to a stimulus (such as saying 'four' when the teacher says 'what is two plus two?'). The cognitive processing during learning is simply the strengthening or weakening of S–R associations. Thorndike's law of effect states that when a response is followed by a pleasurable state of affairs, the S–R link is strengthened; but when a response is followed by displeasure, its link with the stimulus situation is weakened. In its most strict form, this conception of learning holds that the change occurs without conscious mental effort, that is, the strengthening or weakening occurs automatically as a result of rewards and punishments. Thus, the link between the stimulus 'what is two plus two?' and the response 'four' gets stronger if the teacher says 'right', whereas the link between the stimulus 'what is two plus two?' and the response 'three' gets weaker if the teacher says 'wrong'. The learning outcome, or change in knowledge, involves an S–R link that is stronger or weaker than before learning. The resulting performance is that the learner is either more or less likely to make a certain response when confronted with a certain stimulus.

The conception of learning as response strengthening dominated psychology and education for most of the first half of the twentieth century, and was supported mainly by research with animals in contrived laboratory settings. It is still influential today, particularly where basic skills are taught, although there is mounting evidence that the learner's interpretation can influence the strengthening and weakening process.

Learning as Information Acquisition

The second conception listed in Table 1 is that learning involves the acquisition of information, in which presented information is added to the learner's memory bank. In this conception, the teacher's job is to present information as efficiently as possible and the learner's job is to add the presented information to memory. In this way, the learner is a passive recipient of information. Following the four elements listed in Figure 1, the major instructional methods involve techniques for presenting material to learners, such as textbooks, lectures, and multimedia presentations. The major

cognitive process during learning is encoding – that is, adding material to one's long-term memory. In computational terms, information from the outside world is input into the learner, held temporarily in a short-term memory store, and eventually transferred into a long-term memory store. Cognitive processing involves making mental manipulations (or computations) on the data that are input, and storing the results in a memory store. For example, if the teacher says 'Washington is capital of the United States' then this information may eventually be transferred for storage in the learner's memory. The outcome of learning is simply information that has been placed in long-term memory. Test performance is reflected in the learner's ability to answer questions based on the learned information.

The information acquisition conception became popular in the 1960s and 1970s, and was based largely on research with humans in contrived laboratory situations. It is still influential today, particularly concerning memorization of basic facts, although there is mounting evidence that the encoding process can be far more active than originally proposed.

Learning as Knowledge Construction

The third conception of learning listed in Table 1 is knowledge construction, in which learners actively build organized and meaningful representations in their minds. In this view, the teacher is a cognitive guide who influences the way the learner processes incoming material, and the learner is a sense maker who strives to understand the presented material. The major instructional methods involve forms of guided exploration, such as classroom discussion, or cognitive apprenticeship, in which students work with an expert on an authentic academic task. The cognitive processes during learning include paying attention to the relevant material, mentally organizing the material into a coherent structure, and mentally connecting the material with prior knowledge. F. C. Bartlett's research in the 1930s on people's memory for folk stories was among the first to portray learning as an active process in which the learner struggles to make sense of new material. In particular, Bartlett argued that active learning involved assimilating presented material with one's existing schemata in an attempt to make sense of the stories. The outcome of learning is a coherent mental representation that includes elements that were presented and elements from prior knowledge. The knowledge construction view rose in prominence during the 1980s and 1990s, and was based mainly on research with humans in realistic learning settings. It continues to influence psychology and education today, and has been expanded to include the social context of active learning.

HOW PEOPLE LEARN

Research in cognitive science has yielded three principles that are particularly useful in building an educationally relevant theory of learning: dual channels, limited capacity, and active learning. Figure 2 presents a cognitive theory of learning that incorporates these three principles. The boxes on the left represent material that is presented to the learner (in this case pictures and words). The next boxes show how the presented material is represented in sensory memory (through processing by the eyes and ears). The next boxes show how the presented material is represented in working memory (in pictorial and verbal codes). Next, the knowledge that is constructed in working memory is integrated with prior knowledge from long-term memory and ultimately stored in long-term memory. The final box represents performance of the learner based on the newly constructed knowledge.

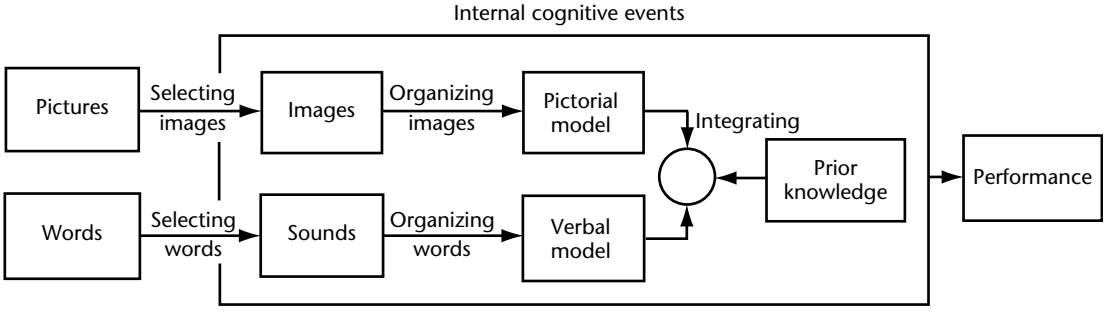


Figure 2. A cognitive theory of learning.

Dual Channels

The first principle is that humans possess separate information processing channels for verbal and pictorial material. The top row of Figure 2 shows the pictorial (or visual) channel, in which material is received by the eyes and is represented pictorially in working memory. The bottom row of Figure 2 shows the verbal (or auditory) channel, in which material is received by the ears (or initially by the eyes), and is represented verbally in working memory. Although verbal forms of presentation have dominated in most academic instruction for a long time, there is increasing evidence that learning can be enhanced when supporting visual material is also presented. Although humans possess channels for other modalities (such as haptic or vistibular), we focus on the visual and auditory modalities because they are the most relevant to classroom educational settings.

Limited Capacity

The second principle is that the capacity to process material in working memory is very limited. Humans are able to hold only a few items in visual working memory and only a few items in verbal working memory. These limitations make it inefficient to try to remember all of the presented material, so that learners have to focus on the most important aspects of the visual and verbal material.

Active Processing

The third principle is that meaningful learning occurs when learners engage in appropriate cognitive processing during learning. Three particularly important cognitive processes during learning are selecting, organizing, and integrating (see Figure 2). Selecting involves paying attention to relevant portions of the material in sensory memory for further processing in working memory. Organizing involves building coherent cognitive structures based on the selected material. Integrating involves building connections between the various verbal and pictorial models in working memory and relevant prior knowledge from long-term memory.

Meaningful learning depends on active cognitive processing during learning – whereby the learner selects, organizes, and integrates material – rather than on active behavior during learning, such as hands-on activities. While some hands-on activities (such as completing a laboratory project) may foster active cognitive processing, other hands-on activities may not; while some forms of passive instruction (such as lecturing) may foster active cognitive processing, others may not.

KINDS OF KNOWLEDGE

If learning involves the construction of knowledge, then it is useful to examine the kinds of knowledge that people build. Psychologists distinguish between declarative knowledge and procedural knowledge. Declarative knowledge is knowledge about the world that corresponds to questions about 'what', including factual and conceptual knowledge. Procedural knowledge is knowledge about how to do things that corresponds to questions about 'how to', including algorithmic and strategic knowledge. Educators distinguish between 'low-level' and 'high-level' knowledge. Low-level knowledge supports basic skills and consists of factual and procedural knowledge that forms the foundation of a subject. High-level knowledge supports higher-order skills and consists of conceptual and strategic knowledge that allow one to use a subject matter. Table 2 shows how the distinctions between declarative and procedural knowledge and between low-level and high-level knowledge give rise to four distinct kinds of knowledge.

Factual Knowledge

Factual knowledge is low-level declarative knowledge about the definitions or characteristics of elements. Factual knowledge includes knowing the definition of basic terms (e.g. 'factual knowledge is low-level declarative knowledge'), knowing the symbols used in a field (e.g. 'arrows refer to cognitive processes in Figure 2'), and knowing the features or properties of basic elements (e.g. 'knowledge exists only in the learner's mind').

Conceptual Knowledge

Conceptual knowledge is high-level declarative knowledge about the interrelations among elements within a meaningful structure. Conceptual knowledge includes knowing the classification relations among the main concepts or categories in a

Table 2. Four kinds of knowledge

	Declarative	*Procedural*
Low-level	Factual	Algorithmic
High-level	Conceptual	Strategic

field (e.g. the classification system shown in Table 2), knowing the principles or generalizations in a field (e.g. knowing what it means to say that working memory is limited in capacity), and knowing the theories or models in a field (e.g. the system of learning summarized in Figure 2).

Algorithmic Knowledge

Algorithmic knowledge is low-level procedural knowledge about how to carry out a precise procedure or skill. Algorithmic knowledge includes knowledge of specific procedures (e.g. how to compute a mean) or skills (e.g. how to change font style in a word processor).

Strategic Knowledge

Strategic knowledge is high-level procedural knowledge about how to use a general strategy or method, including knowing how to select an appropriate algorithm. Strategic knowledge includes knowing general methods (e.g. how to judge the scientific merits of a published research study), knowing general strategies (e.g. how to write an essay on a given topic), and knowing how to manage one's cognitive resources (e.g. how to determine whether a given paragraph conveys its intended meaning). This last example represents meta-cognitive strategic knowledge, that is, knowledge about one's own cognitive processing. In addition, the learner's attitudes and beliefs, such as the belief that 'I am good at learning about cognitive science', are important kinds of knowledge.

TYPES OF LEARNING PERFORMANCE

Learning is generally assessed by using two different kinds of tests: retention and transfer tests. This distinction is at the heart of Bloom's taxonomy of educational objectives, an important analysis originally published in the 1950s and recently updated.

Retention

Retention tests measure how much or how well the learner remembers the presented material. Retention is often the stated goal of training programs, and often focuses on the learner's low-level knowledge such as factual and procedural knowledge. Retention tests are the most straightforward way of measuring learning. They include recall and recognition tests. In a recall test, the learner is asked to produce some of the presented information (e.g. to answer the question 'what are the two kinds of tests used to assess learning?'). In a recognition test, the learner is asked to select or verify some of the presented information, as in a multiple-choice question (e.g. 'learning is generally measured by: (a) retention tests, (b) transfer tests, (c) both, (d) neither') or a true-or-false question (e.g. 'true or false: learning is generally assessed by retention and transfer tests').

Transfer

Transfer tests measure how well the learner can use what was learned in order to solve new problems or support new learning. Transfer is often the stated goal of schooling and often focuses on how well the learner can use high-level knowledge such as conceptual and strategic knowledge. Transfer tests are more complex than retention tests. They sometimes involve quantitative measurements (e.g. counting the number of correct answers to the question 'what are the advantages of transfer measures over retention measures?'), and sometimes qualitative measures (e.g. categorization of the mental model or strategy that the learner uses in describing his or her thought process in creating a transfer problem).

TYPES OF LEARNING OUTCOMES

According to the cognitive theory of learning, there are three broad classes of learning outcomes: no learning, rote learning, and meaningful learning.

No Learning

'No learning' is indicated by poor performance on tests of retention and transfer. In this case, the learner has failed to attend to the relevant material during learning (i.e. the cognitive process of selecting) or to retain it.

Rote Learning

Rote learning is indicated by good performance on tests of retention and poor performance on tests of transfer. In this case, the learner has attended to the relevant material during learning (i.e. the cognitive process of selecting) but has not mentally organized the material into a coherent structure or mentally integrated the material with existing knowledge. For example, after reading a lesson explaining how lightning storms develop, the learner remembers much of the factual material – such as the duration of a lightning strike – but is not

able to use the material to determine how to reduce the intensity of a lightning storm.

Meaningful Learning

Meaningful learning is indicated by good performance on retention and transfer tests. In this case, the learner has attended to the relevant material, and mentally organized and integrated it. For example, after reading a lesson on lightning formation, the learner remembers some of the basic facts and is able to answer transfer questions such as how to reduce the intensity of a lightning storm. According to cognitive theories of learning, meaningful learning is qualitatively different than rote learning. Rather than simply learning more than rote learners, meaningful learners construct knowledge that is structurally different from that of rote learners. In some cases, these differences can be described in terms of differences between the mental models – that is, representations of how a cause-and-effect system works – that rote and meaningful learners construct.

TEACHING FOR MEANINGFUL LEARNING

A major challenge is to understand how to create instructional methods that foster meaningful learning, as measured by superior retention and transfer performance. Such methods must guide appropriate cognitive processing during learning, including helping learners to select, organize, and integrate the presented material. Among the instructional methods that have been suggested to achieve this goal are: giving productive feedback, providing guided exploration, explaining worked examples, guiding cognitive processing of text, teaching learning strategies, creating cognitive apprenticeship, and priming learners' motivation to learn.

Giving Productive Feedback

Practice with feedback is a traditional instructional method, especially when teaching low-level knowledge such as algorithms and facts. The teacher encourages the learner to make some response and then provides feedback concerning it. For example, the learner may be asked to answer a question such as '$5 - (-3) = \ldots$'; if the learner replies '2', the teacher says 'wrong' and if the learner replies '8', the teacher says 'right'. This level of feedback – which may be called 'right–wrong feedback' – is consistent with the view of rewards and punishments as automatically increasing and decreasing response strengths, but inconsistent with the cognitive view that rewards and punishments are used as information that the learner interprets to adjust his or her knowledge. According to a cognitive view, for example, effective feedback might be to show a number line in which a bunny starting from 0 moves 5 steps to the right (positive direction), then turns backwards to face the left (negative) side, jumps backwards 3 steps, and lands on the number 8. This kind of feedback (given after wrong or right answers) is richer than right–wrong feedback, and is more likely to guide the learner in the process of knowledge building.

Providing Guided Exploration

Guided exploration is also a traditional instructional method, especially when teaching students how to understand algorithms, strategies, and concepts. The student is asked to solve a problem, or engage in a complex task, while the teacher provides some form of guidance. The guidance may involve giving hints (as in discovery methods), making the material more concrete (as in concrete methods), or linking the task to the learner's prior knowledge (as in inductive methods). For example, in teaching students how to find the area of a parallelogram, the teacher can allow learners to cut a triangle from one end of a cardboard parallelogram and place it on the other side to form a rectangle. The learners can then find the area because they know how to find the area of a rectangle. This procedure involves making the materials concrete (by using a cardboard parallelogram), giving hints (by suggesting that the learner use scissors), and linking the task to the learner's prior knowledge (by seeing how an unfamiliar shape – the parallelogram – is just a rectangle in disguise).

Explaining Worked Examples

When the goal is to help students learn strategies for solving problems, then it may be useful to teach by explaining how to solve example problems. When given worked examples, the learner is able to focus on realistic problems and reflect on successful ways of solving them. Two instructional approaches to teaching by example are example-based methods and case-based methods. In example-based methods, the learner is shown a step-by-step solution to an example problem. For example, when the goal is to teach learners how to troubleshoot a car that will not start, an example-based method would be to have an expert mechanic describe their thought process as they inspect

the car. In case-based methods, students learn to solve simulated cases – that is, realistic and authentic problems in a particular field. For example, when the goal is to teach learners how to diagnose symptoms in a patient, a case-based method would be to give the learners a full description of the patient and have them discuss possible diagnoses with each other and with a physician who explains his or her own thinking process.

Guiding Cognitive Processing of Text

When instructional material is presented in the form of prose – that is, printed or spoken text – then it may be useful to design the prose so that it guides the learner's cognitive processing. Guiding cognitive processing of text is most often used to teach factual and conceptual knowledge. For example, adjunct questions – questions placed before or after a passage – can help to guide the learner's attention. Signaling – whereby the passage begins with an outline, has headings keyed to the outline, and contains pointer words such as 'first', 'second' and 'third' – can help guide the learner's mental organizing of the material. Advance organizers – introductory material aimed at priming or providing a familiar context to which the material can be assimilated – can help the learner's mental integrating of the material with prior knowledge.

Teaching Learning Strategies

The process of learning depends both on the material that is presented and on how the material is processed by the learner. In teaching learning strategies, the teacher provides direct instruction in how to learn – that is, in helping learners to control their cognitive processing during learning. Learning strategies can range from mnemonic strategies for memorizing facts, to outlining strategies for organizing presented material, to summarizing strategies for integrating material with existing knowledge. In addition to building a collection of learning strategies, learners also need to know when to use them and how to monitor them; such skills may be called 'metacognitive strategies'.

Creating Cognitive Apprenticeship

Throughout human history, people have learned through various forms of apprenticeship in which they participate in authentic tasks under the supervision of more experienced mentors. In creating cognitive apprenticeship, teachers invite learners to join with them in tackling an authentic academic task such as making sense of a text passage or composing a persuasive essay. In working with learners, the teacher may provide scaffolding (taking away difficult portions of the task), coaching (giving advice about how to do part of the task), and modeling (showing how to complete parts of the task). Cognitive apprenticeship makes use of the social context of learning by creating a community of learners who help one another in the learning process.

Priming Learners' Motivation to Learn

Another instructional method focuses on the learner's motivation to learn, and is based on the idea that learners work harder to understand material when they are motivated to learn. Three ways to prime the learner's motivation to learn are: to teach topics that the student is interested in (motivation based on interest), to provide training aimed at bolstering the learner's perception of himself or herself as a competent learner (motivation based on self-efficacy), and to provide training aimed at creating the belief that successes and failures depend on the learner's effort rather than the learner's ability (motivation based on attributions).

ADVANCES IN COGNITION AND INSTRUCTION

Research on cognition and instruction can enrich both psychology and education. Education enriches psychology by challenging psychologists to develop theories of learning that apply in realistic settings beyond the laboratory. Psychology enriches education by searching for educationally relevant theories of learning. The interaction between cognition and instruction has allowed advances in several areas, including the psychology of subject matter, the teaching of cognitive strategies, and the analysis of cognitive abilities.

Psychology of Subject Matter

Psychologies of subject matter seek to create theories of how people learn and think within specific content areas (such as reading, writing, mathematics, science, or history). Instead of trying to build a general theory of learning applicable to all situations – an approach that has soundly failed in psychology in spite of decades of effort – researchers focus on understanding how people learn in specific content areas such as how they learn to read, write, or compute. Consistent with the larger

theme of domain-specificity of cognition, research in psychologies of subject matter demonstrates that ways of thinking differ between content areas.

Teaching of Cognitive Strategies

The teaching of cognitive strategies involves providing direct instruction in how to learn, remember, or reason. Instead of focusing solely on teaching of basic skills, researchers focus on understanding how to help learners develop strategies for using their knowledge and for processing new material to create new knowledge. Rather than trying to improve the mind in general – an approach that has failed despite decades of effort – researchers focus on identifying specific component strategies that are required for successful cognitive processing on various tasks.

Analysis of Cognitive Abilities

The analysis of cognitive abilities involves conceptualizing cognitive ability, (1) as a collection of component skills rather than as a single, monolithic trait, and (2) as a set of specific skills rather than as a general ability. An important advance in cognition and instruction is the attempt to analyze cognitive ability for any academic domain into smaller components that reflect differences in how people process information within that domain.

Overall, the interaction of the fields of cognition and instruction has generated useful advances for psychology and education. This process is likely to continue in the future.

CONCLUSION

Instruction (including instructional methods) affects the cognitive processes of the learner during learning (including paying attention, mentally organizing the material, and mentally connecting the material with existing knowledge), which affects knowledge changes in the learner (that is, the learning outcome), which affects performance on tests (such as retention and transfer). 'Learning' refers to enduring changes in a learner's knowledge that arise from the learner's cognitive processing of his or her experiences. 'Instruction' refers to the teacher's construction of experiences for learners that are intended to foster learning.

Three conceptions of learning are: learning as response strengthening, learning as information acquisition, and learning as knowledge construction.

Instructional practice should be based on a theory of how people learn. Three principles for building an educationally relevant theory of learning are: that people have separate channels for processing verbal and pictorial material (the dual channels principle); that processing capacity is severely limited within each channel (the limited capacity principle); and that active learning depends on the learner engaging in appropriate cognitive processing during learning (the active learning principle).

Four distinct kinds of knowledge are: factual, conceptual, algorithmic, and strategic. Two types of tests of learning are retention tests and transfer tests. Three kinds of learning outcomes are: no learning, rote learning, and meaningful learning.

Some instructional methods aimed at teaching for meaningful learning are: giving productive feedback, providing guided exploration, explaining worked examples, guiding cognitive processing of text, teaching learning strategies, creating cognitive apprenticeship, and priming learners' motivation to learn.

Three useful advances in research on cognition and instruction are: new conceptions of the psychology of subject matter, the teaching of cognitive strategies, and the analysis of cognitive abilities.

Further Reading

Anderson LW, Krathwohl DR, Airasian PW *et al*. (2001) *A Taxonomy for Learning, Teaching, and Assessing*. New York, NY: Longman.

Berliner DC and Calfee RC (eds) (1996) *Handbook of Educational Psychology*. New York, NY: Macmillan.

Bransford JD, Brown AL and Cocking RR (1999) *How People Learn*. Washington, DC: National Academy Press.

Bruer JT (1993) *Schools for Thought: A Science of Learning in the Classroom*. Cambridge, MA: MIT Press.

Glaser R (ed.) (2000) *Advances in Instructional Psychology*, vol. V. Mahwah, NJ: Lawrence Erlbaum.

Lambert NM and McCombs BL (1998) *How Students Learn*. Washington, DC: American Psychological Association.

Mayer RE (1999) *The Promise of Educational Psychology*. Upper Saddle River, NJ: Prentice-Hall.

Instructional Design

Advanced article

Michael Molenda, Indiana University, Bloomington, Indiana, USA
Charles M Reigeluth, Indiana University, Bloomington, Indiana, USA
Laurie Miller Nelson, Indiana University, Bloomington, Indiana, USA

CONTENTS

Instructional design is that branch of knowledge concerned with theory and practice, related to instructional strategies and systematic procedures for developing and implementing those strategies.

INTRODUCTION

Instructional design is a construct that refers to the principles and procedures by which instructional materials, lessons, and whole systems can be developed in a consistent and reliable fashion. The principles and procedures can be applied to guide designers to work more efficiently while producing more effective and appealing instruction suitable for a wide range of learning environments.

As well as being a construct, instructional design is also a field of theory and practice within the larger field of instructional technology. Instructional designers work in many settings, including schools, colleges and universities, corporations, and military and government agencies.

This article will address in turn the underlying instructional principles, the procedural guides by which these principles are put into application, and finally the construction of learning environments as an alternative way of putting the principles into action.

PRINCIPLES

The design of instruction can be informed by principles drawn from many disciplines, including educational psychology, cognitive science, systems theory, communications, philosophy, anthropology, and organizational theory.

Behaviorist Psychology Sources

The original momentum for the modern concept of instructional design came from B.F. Skinner's suggestions regarding the application of operant conditioning principles to education (Skinner, 1954). His vision became instantiated in programmed instruction, which was based on the following set of prescriptive principles:

> (1) an ordered sequence of stimulus items, (2) to each of which a student responds in some specified way, (3) his responses being reinforced by immediate knowledge of results, (4) so that he moves by small steps, (5) therefore making few errors and practicing mostly correct responses, (6) from what he knows, by a process of successively closer approximations, toward what he is supposed to learn from the program. (Schramm, 1962, p. 2)

As research and practical experience accumulated, the generality of many of these principles came into question. That is, the sequence of experiences, the nature of the response, the timing of feedback, and the size of steps all appeared to be contingent on various learner and learning conditions. The prescriptions of programmed instruction were broadened and reduced by Popham (1971) to four principles:

> (1) Provide relevant practice for the learner. (2) Provide knowledge of results. (3) Avoid the inclusion of irrelevancies. (4) Make the material interesting. (Popham, 1971, p. 171)

Cognitive Sources

Since the 1960s, instructional design has been increasingly informed by principles drawn from other sources, especially cognitive science and cognitive psychology. Cognitive models for instruction emphasize the importance of the learner's cognitive and affective processes in mediating the effects of instruction. From this perspective, learners use their memory and thought processes to generate strategies as well as to store and manipulate mental representations of images and ideas.

Robert Gagné was a leading interpreter of learning theory into instructional theory. Early editions of his influential book, *The Conditions of Learning* (Gagné 1965), proposed that the information-processing model of learning could be combined with behaviorist concepts to provide a more complete view of learning tasks. From descriptive theories of information processing, Gagné deduced prescriptive theories about instruction methods ('external conditions of learning'). His list of nine 'instructional events' became a robust and influential conceptual schema for the planning of lessons:

> (1) gaining attention; (2) informing learners of the objective; (3) stimulating recall of prior learning; (4) presenting the content; (5) providing 'learning guidance'; (6) eliciting performance; (7) providing feedback; (8) assessing performance; (9) enhancing retention and transfer. (Gagné and Medsker, 1996, p. 140)

More recently, other descriptive theories of learning that are derived from a cognitive perspective have influenced further prescriptive theories and principles for instruction. Schema theory, which emphasizes the schematic structure of knowledge, is one of the major sources of influence. Ausubel (1980) described schemata as providing ideational scaffolding, which contains slots that can be instantiated with particular cases. These schemata allow learners to organize information into meaningful units. This theory implies that the learner's cognitive structure at the time of learning is the most important factor in determining the likelihood of successful learning. One instructional design principle derived from this theory pertains to the advance organizer – a brief outline based on the learner's existing knowledge, which serves as 'ideational scaffolding' for new learning. Ausubel proposed that advance organizers could activate broader and more inclusive knowledge, providing a cognitive structure for new meaningful learning.

Constructivist Sources

Other educational theories emphasize the importance of the ideas generated by learners themselves. Wittrock (1974) described a view of learning and instruction in which the 'generations' (mental activities such as summaries, pictures, analogies, and discussions) performed by learners influence the success of instruction. This emphasis on learner generation characterizes constructivism, which assumes that 'knowledge is individually constructed and socially co-constructed by learners based on their interpretations of experiences in the world' (Jonassen, 1999, p. 217). Prescriptive principles from constructivism include the following:

> (1) Embed learning in complex, realistic and relevant environments. (2) Provide for social negotiation as an integral part of learning. (3) Support multiple perspectives and the use of multiple modes of representation. (4) Encourage ownership in learning. (5) Nurture self-awareness of the knowledge construction process. (Driscoll, 2000, pp. 382–383)

An alternative view of constructivism, known as the 'situated cognition' perspective, which is derived from anthropology (Lave and Wenger, 1991; Barab *et al.*, 1999), proposes that the need to understand the learning context supersedes the need to understand the mental processes going on inside individual learners – that is, the learner and the environment are always interacting. For example, what is understood as memorized algorithms in a mathematics classroom differs from what is understood through grappling with a real-world carpentry problem.

The field offers a wide variety of theories of instructional design that prescribe specific methods of instruction and the conditions under which they can best be used. A growing number of instructional design theories have been developed to address a wide range of learning situations in order to foster cognitive, psychomotor, or affective development. These include theories for such diverse types of learning as experiential, collaborative, and self-regulated learning, as well as emotional, social, and even spiritual development (Reigeluth, 1999).

Comprehensive Set of Design Principles

A recent synthesis by M. David Merrill (Merrill, 2001) provides a coherent and comprehensive overview of instructional design principles from an eclectic perspective, incorporating behaviorist, cognitivist, and constructivist conceptions:

- *Problem*. Learning is facilitated when the learner:
  - …is engaged in solving a real-world problem;
  - …is engaged at the problem or task level, not just the operation or action level;
  - …solves a progression of problems;
  - …is guided to an explicit comparison of problems.
- *Activation*. Learning is facilitated when the learner:
  - …is directed to recall, relate, describe, or apply knowledge from relevant past experience that can be used as a foundation for the new knowledge;
  - …is provided with relevant experience that can be used as a foundation for the new knowledge.

- *Demonstration*. Learning is facilitated when:
  - …the learner is shown rather than told;
  - …the demonstration is consistent with the learning goal;
  - …the learner is shown multiple representations;
  - …the learner is directed to explicitly compare alternative representations;
  - …the media play a relevant instructional role.
- *Application*. Learning is facilitated when:
  - …the learner is required to use his or her new knowledge to solve problems;
  - …the problem-solving activity is consistent with the learning goal;
  - …the learner is shown how to detect and correct errors;
  - …the learner is guided in his or her problem-solving by appropriate coaching that is gradually withdrawn.
- *Integration*. Learning is facilitated when the learner:
  - …can demonstrate his or her new knowledge or skill;
  - …can reflect on, discuss, and defend his or her new knowledge;
  - …can create, invent, and explore new and personal ways to use his or her new knowledge.

(Merrill, 2001, pp. 5–7)

APPLICATIONS

Instructional design theories and principles are put into practice by being embedded in procedural guides or protocols for instructional development. These often take the form of instructional systems development (ISD) process models.

ISD Process Models

Historically, instructional design can be seen as having two parents, namely the systems approach and behaviorist psychology. The relative contributions of each are difficult to assess because at the time when instructional design was conceived the two sources were quite intertwined. During the post-World War II period each of the US military services had developed doctrines for training development, all of which were based on the systems approach – a 'soft' version of systems analysis, itself an offshoot of operations research. Behaviorist learning theory was a pervasive influence in US military training, and was being enthusiastically explored in school and university instruction during this same time period. Many of those who had been involved in military training development were applying their craft in university research and development centers. Thus the systems approach and behaviorist concepts became increasingly intertwined, both in the military services and in academia.

During the 1960s, the systems approach began to appear in procedural models of instruction in US higher education. Barson's (1967) instructional systems development project produced an influential model and set of heuristic guidelines for developers. If one looks at the form and language of these early models, the influence of the systems approach paradigm is obvious. Early models instantiate the principles of gathering and analyzing data prior to making decisions, and using feedback to correct deficiencies in work completed. They include systems terminology such as 'mission objectives,' 'transmission vehicles,' 'error detection,' and so on. The 'soft' systems concept continued to evolve in terms of its application to complex problems in human organizations, since it was recognized that 'hard' mathematical systems concepts did not apply directly to complex clusters of human activities, which represented systems only in the loosest sense of the term. Thus the systems concept came to be seen more as an analogy or as a 'means of *structuring a debate*, rather than as a recipe for guaranteed efficient achievement' (Checkland, 1981, p. 150).

The largest group of models is derived from the 'soft' systems paradigm, commonly referred to as the ADDIE model (an acronym derived from the key steps in the model: Analysis, Design, Development, Implementation, and Evaluation). These steps identify a generic systems approach, similar to that applied in other fields such as software engineering and product design. The ADDIE approach is systematic in that it recommends using the decisions made at each step (the output) as the input for the next step. That is, the *analysis* stage begins by surveying the learners and learning environment in order to determine which learning problems are of high priority and should be chosen as objectives. In the *design* stage, those learning objectives are translated into lesson plans or blueprints. In the *development* stage, specific materials and procedures are created to give life to the blueprints. In the *implementation* stage, learners actually use the materials and procedures that were created. In the *evaluation* stage, the learners are assessed in order to determine the extent to which they mastered the objectives specified at the beginning, and revisions are made as necessary. The ADDIE family of models, represented by 13 different variations on the systems approach, has been analyzed by Gustafson and Branch (1997).

Instructional Theory-based ISD Models

In addition to generic ISD process models, a number of alternative models have been developed as guides to the application of particular instructional design theories. One of the earliest was *structural communication*, developed in the UK by Bennett and Hodgson (Hodgson, 1974). Originating as a reaction to the limitations of programmed instruction, structural communication involved a process of analysis and development contingent on different levels of thinking, namely creative, conscious, sensitive, and automatic levels (whereas programmed instruction lent itself only to the sensitive and automatic levels). The form of the instruction resembles a guided discussion, emphasizing the role of the learner as an active inquirer.

A more recent attempt to mold a process model around the constructivist view is the *reflective recursive design and development* model of Willis and Wright (2000). Their process revolves around three focal points, namely definition, design/development, and dissemination. It assumes

> that designers will work on all three aspects of the design process in an intermittent and recursive pattern that is neither predictable nor prescribable. The focal points are, in essence, a convenient way of organizing our thoughts about the work. (Willis and Wright, 2000, p. 5)

LEARNING ENVIRONMENTS

Some approaches to instructional design focus not on the procedural steps involved in creating specific lessons, but on the construction of whole learning environments that have special features conducive to efficient, effective learning. Such learning environments can themselves be viewed as large-scale methods – frameworks that are created in order to immerse learners in a consistent set of instructional conditions. Examples include the personalized system of instruction (Semb, 1997), goal-based scenarios (Schank *et al*., 1999), problem-based learning (Boud and Feletti, 1997), open learning environments (Hannafin *et al*., 1999), and constructivist learning environments (Jonassen, 1999).

References

Ausubel DP (1980) Schemata, cognitive structure and advance organizers: a reply to Anderson, Spiro and Anderson. *American Educational Research Journal* **17**: 400–404.

Barab SA, Cherkes-Julkowski M, Swenson R *et al*. (1999) Principles of self-organization: ecologizing the learner–facilitator system. *Journal of the Learning Sciences* **8**: 349–390.

Barson J (1967) *Instructional Systems Development: a Demonstration and Evaluation Project*. US Office of Education, Title II-B project OE 3-16-025. East Lansing, MI: Michigan State University.

Boud D and Feletti GI (eds) (1997) *The Challenge of Problem-Based Learning*, 2nd edn. London, UK: Kogan Page.

Checkland P (1981) *Systems Thinking, Systems Practice*. Chichester, UK: John Wiley.

Driscoll MP (2000) *Psychology of Learning for Instruction*, 2nd edn. Boston, MA: Allyn & Bacon.

Gagné RM (1965) *The Conditions of Learning*. New York, NY: Holt, Rinehart & Winston.

Gagné RM and Medsker KL (1996) *The Conditions of Learning: Training Applications*. Fort Worth, TX: Harcourt Brace College Publishers.

Gustafson KL and Branch RM (1997) *Survey of Instructional Development Models*, 3rd edn. Syracuse, NY: ERIC Clearinghouse on Information and Technology.

Hannafin M, Land S and Oliver K (1999) Open learning environments: foundations, methods and models. In: Reigeluth CM (ed.) *Instructional-Design Theories and Models: a New Paradigm of Instructional Theory*, vol. II, pp. 115–140. Mahwah, NJ: Lawrence Erlbaum.

Hodgson AM (1974) Structural communication in practice. In: *APLET Yearbook of Educational and Instructional Technology 1974/75*, pp. 139–153. London, UK: Kogan Page.

Jonassen D (1999) Designing constructivist learning environments. In: Reigeluth CM (ed.) *Instructional-Design Theories and Models: a New Paradigm of Instructional Theory*, vol. II, pp. 215–239. Mahwah, NJ: Lawrence Erlbaum.

Lave J and Wenger E (1991) Situated learning: legitimate peripheral participation. Cambridge, UK: Cambridge University Press.

Merrill MD (2001) *First Principles of Instruction*. Retrieved 18 December 2001 from Utah State University, ID2 Website [http://www.id2.usu.edu/Papers/5FirstPrinciples.PDF].

Popham WJ (1971) Preparing instructional products: four developmental principles. In: Baker RL and Schutz RE (eds) *Instructional Product Development*, pp. 169–207. New York, NY: Van Nostrand Reinhold.

Reigeluth CM (ed.) (1999) *Instructional-Design Theories and Models: a New Paradigm of Instructional Theory*, vol. II. Mahwah, NJ: Lawrence Erlbaum.

Schank RC, Berman TR and Macpherson KA (1999) Learning by doing. In: Reigeluth CM (ed.) *Instructional-Design Theories and Models: a New Paradigm of Instructional Theory*, vol. II, pp. 161–181. Mahwah, NJ: Lawrence Erlbaum Associates.

Schramm W (1962) *Programmed Instruction: Today and Tomorrow*. New York, NY: Fund for the Advancement of Education.

Semb GB (1997) The personalized system of instruction (PSI) and the three Rs: revolutions, revelations and

reflections. In: Dills CR and Romiszowski AJ (eds) *Instructional Development Paradigms*, pp. 353–370. Englewood Cliffs, NJ: Educational Technology Publications.

Skinner BF (1954) The science of learning and the art of teaching. *Harvard Educational Review* **24**: 86–97.

Willis J and Wright KE (2000) A general set of procedures for constructivist instructional design: the new R2D2 model. *Educational Technology* **40**: 5–20.

Wittrock MC (1974) Learning as a generative process. *Educational Psychology* **11**: 87–95.

Further Reading

Dick W, Carey L and Carey JO (2001) *The Systematic Design of Instruction*, 5th edn. New York, NY: Longman.

Dills CR and Romiszowski AJ (eds) *Instructional Development Paradigms*. Englewood Cliffs, NJ: Educational Technology Publications.

Gustafson KL and Branch RM (2002) What is instructional design? In: Reiser RA and Dempsey JV (eds) *Trends and Issues in Instructional Design and Technology*, pp. 16–25. Upper Saddle River, NJ: Merrill Prentice Hall.

Molenda M, Pershing JA and Reigeluth CM (1996) Designing instructional systems. In: Craig RL (ed.) *The ASTD Training and Development Handbook*, 4th edn, pp. 266–293. New York, NY: McGraw-Hill.

Reigeluth CM (ed.) (1987) *Instructional Theories in Action: Lessons Illustrating Selected Theories and Models*. Hillsdale, NJ: Lawrence Erlbaum.

Intellectual Ability

Introductory article

Robert J Sternberg, Yale University, New Haven, Connecticut, USA

CONTENTS

Intellectual abilities comprise those higher-order cognitive skills that are involved in coping with the environments in which we live, including but not limited to learning and thinking skills. Intelligence, creativity, and wisdom are three of the major intellectual abilities.

INTELLIGENCE

Defining Intelligence

Intelligence involves the ability to adapt to the environment. But what specifically is involved in intelligent thinking?

Two symposia have sought to ascertain the essential features of intelligence, one in 1921 and the other in 1986. Features that have been proposed include: adaptation in order to meet the demands of the environment effectively; elementary processes of perception and attention; higher-level processes of abstract reasoning, mental representation, problem-solving, and decision-making; ability to learn; and effective behavior in response to problem situations.

Some researchers, such as Boring in 1923, have been content to define intelligence operationally, simply as the 'intelligence quotient' (IQ). Originally, IQ was defined in terms of a ratio of one's mental-age level of performance to one's chronological-age level of performance, but today IQs are defined in terms of how much one differs from the average. An average IQ is 100. Slightly more than two-thirds of IQs fall between 85 and 115.

Scientific definitions rely on tests such as those invented by Binet and Simon in 1916 to measure judgmental abilities or by Wechsler in 1939 to measure verbal and performance abilities. Earlier tests proposed by Galton in 1883 measured psychophysical abilities (such as sensitivity of hearing or touch). They proved to be less valid, in that they correlated neither with each other nor with success in educational settings.

Laypeople also can be asked to define intelligence, and it turns out that their definitions differ from scientists' definitions in placing somewhat greater emphasis on social competence. In one study by Sternberg and his colleagues, for example, laypeople defined intelligence in terms of three broad classes of skills: practical problem-solving; verbal ability; and social competence. But how people define intelligence varies across occupations. For example, one study found that philosophy professors tend to stress critical and logical thinking very heavily, whereas physicists tend to place more value on precise mathematical thinking, the ability to relate physical phenomena to theoretical concepts of physics, and the ability to grasp quickly the laws of nature.

How people define intelligence also depends on the culture in which they live. For example, in 1974 Wober studied two tribes in Uganda. He found that the Baganda tended to associate intelligence with mental order, whereas the Batoro associated it with mental turmoil. Super analyzed concepts of intelligence among the Kokwet of western Kenya. He found that intelligence meant different things for children and adults. Children who were responsible as well as quick in comprehension and effective in their management of interpersonal relationships were viewed as intelligent; while adults were viewed as intelligent if they were inventive and, sometimes, if they were wise and unselfish. Yang and Sternberg found that Chinese people in Taiwan characterize intelligence in terms of cognitive abilities, but also interpersonal and intrapersonal (self-understanding) skills, as well as knowing when to show one's abilities and when not to show them.

Heritability and Modifiability of Intelligence

Whatever human intelligence may be, that aspect of it measured as IQ is both partially heritable – with a heritability coefficient estimated at about 0.5

(somewhat lower in childhood and higher in adulthood) – and modifiable to some degree. Thus, roughly half the variation in scores on intelligence tests is due to heritable factors, and half to environmental factors.

Flynn has discovered that intelligence as measured by IQ tests rose steadily throughout most of the twentieth century. Because of the rapidity of the increase, the increase could not be due to genetic factors. There may be unknown factors in the environment that have produced these increases; or it may be that IQ tests really, to some extent, measure something other than intelligence.

Theories of Intelligence

A theory, in contrast to a definition, must provide an explanatory framework and be testable. Proposed theories of intelligence have been of several different kinds.

Psychometric theories

The best-known theories are probably the psychometric theories. Such theories are based on, and often tested by, analysis of individual differences in scores among people who take tests. Most conventional intelligence tests have arisen from psychometric theories. These tests, mostly originating in the early twentieth century, became especially popular during the First World War as a means of screening soldiers. Then and now, the tests have tended, on average, to favor individuals of higher socioeconomic status and, in the United States, who are European–American or Asian–American rather than African–American or Hispanic–American. There are many alternative explanations for these differences, but most researchers regard them as environmental in origin.

The earliest major psychometric theory was that of Spearman, who proposed in 1904 that intelligence comprises a 'general factor' (*g*) common to all intellectual tasks, as well as 'specific factors' (*s*), each of which is unique to a given test of intelligence. His proposal was based on his finding of a 'positive manifold' among intelligence tests: all tests seemed to be positively intercorrelated, suggesting the existence of a general factor. Spearman's theory still has many proponents today, such as Jensen, whose analyses of factor-analytic and other data suggest what he believes to be a single factor underlying virtually all intellectual performances.

In 1938, Thurstone disagreed with Spearman, arguing that the general factor was an artifact of the way Spearman analyzed his data. Thurstone suggested that seven primary mental abilities underlie intelligence: verbal comprehension, verbal fluency, number skills, spatial visualization, inductive reasoning, memory, and perceptual speed. More recent theorists, such as Cattell and Carroll, have attempted to integrate these two kinds of views, suggesting that intelligence is best understood hierarchically, with a general factor at the top of the hierarchy (i.e. more central to intelligence) and narrower factors under it. Cattell proposed two such factors: fluid intelligence, which is involved in reasoning with novel kinds of stimuli; and crystallized intelligence, or stored knowledge.

The principal limitations of psychometric theories are that they rely heavily, often exclusively, for their validation on correlational methods; that they depend heavily on the existence and strength of individual differences; and that what one gets out of a psychometric analysis is a transformation of what one puts in, so that if one puts in questionable tests, one gets out questionable results.

Computational theories

Unlike psychometric models, which map the structure of human intelligence, computational models emphasize the processes underlying intelligent behavior. In particular, theorists using these models are interested in studying how people engage in information processing – that is, the operations by which people mentally manipulate what they learn and know about the world. Such studies differ primarily in terms of the complexity of the processes being studied. One way to study the relation between intelligence and information processing is to examine simple information processing such as occurs when one must make rapid judgments about which of two lines is longer.

Deary and Stough have proposed that a very low-level psychophysical measure, inspection time, may provide us with insights into the fundamental nature of intelligence. Their basic idea is that individual differences in intelligence may derive, in part, from differences in the rate of intake and processing of very simple stimulus information. In the inspection-time task, a person looks at two vertical lines of unequal length, and simply has to say which line is longer. Inspection time is the length of time of stimulus presentation an individual needs in order correctly to decide which of the two lines is longer. Pairs of lines are presented for different periods of time, and psychophysical methods are used to determine how quickly an individual can distinguish their lengths. Investigators have found that more intelligent individuals (as defined by IQ) can discriminate the lengths of

the lines with smaller stimulus durations (inspection times).

Another computational approach considers complex information processing such as occurs in analogies (problems of the form '*A* is to *B* as *C* is to what?'), series problems (e.g. completing a numerical or figural series), and syllogisms (problems such as 'John is taller than Peter; Peter is taller than Robert; who is tallest?'). The goal of this approach has been to find out just what it is that makes some people more intelligent processors of information than others. The idea is to take the kinds of tasks used in conventional intelligence tests and to isolate the components of intelligence – the mental processes used in performing these tasks. Examples of such processes include translating sensory input into a mental representation, transforming one conceptual representation into another, and translating a conceptual representation into a motor output. Typically, measurements are made both in terms of response time to answer test items and in terms of whether the response given is correct.

In general, according to Sternberg, more intelligent people take longer during global planning (encoding the problem and formulating a general strategy for attacking the problem or set of problems) but they take less time for local planning (forming and implementing strategies for the details of the task). The advantage of spending more time on global planning is the increased likelihood that the overall strategy will be correct. For example, the brighter person might spend more time researching and planning for writing a term paper, but less time actually writing it. This differential in time allocation has also been shown in other tasks, including solving physics problems.

Whereas information-processing investigators study such differences at the level of hypothesized mental processes, biological investigators seek to understand the origins of such differences in terms of the functioning of the brain.

Some related work looks at the role of working memory in intelligence. For example, a number of investigators, such as Kyllonen and Christal, have proposed that general intelligence actually is, in large part, working memory. This is the kind of ability that is measured when, say, a person is asked to remember a series of numbers (such as '3-2-6-1-7-4') backwards.

Computational theories have certain limitations. Reaction-time and related methodologies may make assumptions (for example, assumptions of serial information processing) that are not valid. Computational theories also sometimes ignore the contexts in which mental processes occur, and thus may make claims for generality that exceed their scope.

Biological theories

Most biological theories have been of aspects of intelligence rather than of the phenomenon as a whole. However, one early theory that has had considerable influence did try to deal with the phenomenon as a whole. Hebb proposed three kinds of intelligence: 'A', 'B', and 'C'. Intelligence A, he proposed, is largely genetic and 'hard-wired' in its origins. Intelligence B is the result of the interaction between genes and environment. Intelligence C is one's performance on a standardized test of intelligence. This theory, although heuristically useful, is largely speculative.

Systems theories

Some theories of intelligence have viewed intelligence as a system. By far the best-known theory of this kind is that of Piaget, according to which intelligence involves an equilibration between two processes: assimilation of new information to fit existing cognitive structures, and accommodation of existing cognitive structures to incorporate information that does not fit into existing cognitive structures. Sternberg has proposed that intelligence comprises three aspects: analytical abilities (used to analyze, evaluate, and criticize), creative abilities (used to create, discover, and invent), and practical abilities (used to apply, implement, and use). Intelligent people, according to this 'theory of successful intelligence', make the most of their strengths and either correct or compensate for their weaknesses. Gardner has suggested instead that there are multiple intelligences: linguistic, logical–mathematical, spatial, musical, bodily–kinesthetic, naturalist, intrapersonal, and interpersonal intelligence; and perhaps also existential intelligence. The recent theory of 'emotional intelligence' also describes intelligence in a broad way. (*See* **Intelligence**)

Gardner's theory has not, so far, generated any predictive scientific data that test the model as a whole, so that questions about its validity cannot yet be adequately answered.

CREATIVITY

Creativity involves producing ideas that are original (novel), of high quality, and appropriate to the task being considered. Creativity often involves seeing problems and their solutions in a variety of different ways, and acting upon these different

ways of seeing things. Several accounts of creativity have been offered, which shed light on different aspects of it. Sternberg and Lubart have distinguished the types of approach described below.

Psychodynamic Approaches

The psychodynamic approach can be considered the first of the major twentieth-century theoretical approaches to the study of creativity. Starting from the idea that creativity arises from the tension between conscious reality and unconscious drives, Freud proposed that writers and artists produce creative work as a way to express their unconscious wishes in a publicly acceptable fashion. These unconscious wishes may concern power, riches, fame, honor, or love. Case studies of eminent creators, such as Leonardo da Vinci, were used to support these ideas.

Later, the psychodynamic approach as used by Kris introduced the concepts of 'adaptive regression' and 'elaboration' for creativity. Adaptive regression, the primary process, is the intrusion of unmodulated thoughts in consciousness. It occurs when one returns to states of consciousness typical of one's earlier life. Such returns can be adaptive if they help one to generate useful ideas. Unmodulated thoughts can occur during active problem-solving; but they also often occur during sleep, intoxication from drugs, fantasies or daydreams, or psychoses. Elaboration, the secondary process, is the reworking and transformation of primary-process material through reality-oriented, ego-controlled thinking.

Other theorists such as Kubie emphasized that the preconscious, which falls between conscious reality and the encrypted unconscious, is the true source of creativity because preconscious thoughts are loose and vague but interpretable. In contrast to Freud, Kubie claimed that unconscious conflicts actually have a negative effect on creativity because they lead to fixated, repetitive thoughts.

In general, the evidence in favor of psychodynamic approaches tends to be more anecdotal than experimental.

Psychometric Approaches

When we think of creativity, eminent artists or scientists such as Michelangelo or Einstein immediately come to mind. However, these highly creative people are rare and difficult to study in the psychological laboratory. In 1950, Guilford noted that this difficulty had limited research on creativity. He proposed that creativity could be studied in ordinary subjects using paper-and-pencil tasks. One of these was the 'unusual uses test', in which an examinee thinks of as many uses for a common object (such as a brick) as possible. Many researchers adopted Guilford's suggestion, and 'divergent thinking' tasks quickly became the main instruments for measuring creative thinking. The tests were a convenient way of comparing people on a standard 'creativity' scale.

Building on Guilford's work, Torrance in 1974 developed a set of tests of creative thinking. These tests consist of several relatively simple verbal and figural tasks that involve divergent thinking and other problem-solving skills. They can be scored for fluency (total number of relevant responses), flexibility (number of different categories of relevant responses), originality (statistical rarity of the responses) and elaboration (amount of detail in the responses). The Torrance tests include: asking questions (the examinee writes out all the questions he or she can think of, based on a drawing of a scene); improving an object (the examinee lists ways to change a toy monkey so that children will have more fun playing with it); thinking of unusual uses (the examinee lists interesting and unusual uses of a cardboard box); and developing circles (the examinee expands empty circles into different drawings and gives them titles).

Two obvious disadvantages of such tests and assessments are the time and expense involved in administering them, and the subjective scoring of them. In 1962, Mednick produced a 30-item, objectively scored, 40-minute test of creative ability called the 'remote associates test'. The test is based on his theory that the creative thinking process is the forming of associative elements into new combinations that are in some way useful. The more mutually remote the elements of the new combination, the more creative the process or solution is alleged to be. Because the ability to make these combinations and arrive at a creative solution depends on the existence of the necessary elements in a person's knowledge base, and because the probability and speed of attainment of a creative solution are influenced by the organization of the person's associations, Mednick's theory suggests that creativity and intelligence are closely related.

The remote associates test requires the subject to supply a fourth word that is remotely associated with three given words. For example: 'rat', 'blue', and 'cottage' may yield 'cheese'; 'surprise', 'line', and 'birthday' may yield 'party'; 'out', 'dog', and 'cat' may yield 'house'.

Cognitive Approaches

The cognitive approach to creativity seeks understanding of the mental representations and processes underlying creative thought. By studying, say, perception, or memory, one would already be studying the bases of creativity; thus, the study of creativity would merely represent an extension, and perhaps not a very large one, of work that is already being done under another guise. For example, creativity has often been subsumed under the study of intelligence. Creativity and intelligence are certainly related; however, the subsumption has often been so strong as to suggest that they need not be regarded as distinct – a view which researchers such as Wallach and Kogan have argued against.

More recently, Weisberg has proposed that creativity involves essentially ordinary cognitive processes yielding extraordinary products. Weisberg attempted to show that insights depend on subjects applying ordinary cognitive processes (such as analogical transfer) to knowledge already stored in memory. He used case studies of eminent creators, as well as laboratory research, such as studies with Duncker's candle problem (proposed in 1945). This problem requires participants to attach a candle to a wall using only objects available in a picture (candle, box of thumbtacks, and book of matches). Langley and his colleagues have made a similar claim about the ordinary nature of creative thinking.

As a concrete example of this approach, Weisberg and Alba asked people to solve the nine-dot problem. The problem is to connect all of the dots, which are arranged in the shape of a square with three rows of three dots each, using no more than four straight line segments, never arriving at a given dot twice, and never lifting their pencil from the page. It can be solved only if the line segments go outside the square of dots. Weisberg and Alba showed that even when people were given this insight, they still had difficulty in solving the problem. In other words, whatever is required to solve the nine-dot problem, it is not just this one extraordinary insight.

Finke, Ward, and Smith have proposed what they call the 'Geneplore' model, according to which there are two main processing phases in creative thought: a generative phase and an exploratory phase. In the generative phase, an individual constructs mental representations ('preinventive structures'), which have properties promoting creative discoveries. In the exploratory phase, these properties are used to produce creative ideas. A number of mental processes may enter into these phases of creative invention, such as retrieval, association, synthesis, transformation, analogical transfer, and categorical reduction (i.e. mentally reducing objects or elements to more primitive categorical descriptions). In a typical experimental test based on this model, participants are shown parts of objects, such as circles, cubes, parallelograms, or cylinders. On a given trial, three parts will be named, and participants will be asked to imagine combining the parts to produce a practical object or device. For example, participants might imagine a tool, a weapon, or a piece of furniture. The objects thus produced are then rated by judges for their practicality and originality.

Computer-simulation approaches have as their goal the production of creative thought by a computer in a manner that simulates what people do. Langley, Simon, Bradshaw, and Zytgow, for example, have developed a set of programs that rediscover basic scientific laws. These computational models rely on heuristics – problem-solving guidelines – for searching a data set or conceptual space and finding hidden relationships between input variables. The program, called BACON (after the philosopher Francis Bacon), uses heuristics such as 'if the values of two numerical terms increase together, consider their ratio' to search data for patterns. One of BACON's accomplishments was to examine some of the observational data on the orbits of planets available to Kepler and rediscover Kepler's third law of planetary motion. This program is unlike creative functioning, however, in that problems are given to it in structured form: creative functioning is largely about discovering what the problems are.

Other programs have extended the search heuristics, the ability to transform data sets, and the ability to reason with qualitative data and scientific concepts. There are also models of artistic creation. For example, Johnson-Laird has developed a jazz improvisation program in which novel deviations from the basic jazz chord sequences are guided by harmonic constraints (or tacit principles of jazz), and random choice when several allowable directions for the improvisation exist.

Social–Personality Approaches

Developing in parallel with the cognitive approach, work in the social–personality approach has focused on personality variables, motivational variables, and the sociocultural environment as sources of creativity. Researchers such as Amabile, Barron, Eysenck, Gough, and MacKinnon have noted that

certain personality traits often characterize creative people. Through correlational studies and research using samples of both eminent and ordinary people at high and low levels of creativity, a large set of potentially relevant traits has been identified. These traits include independence of judgment, self-confidence, attraction to complexity, aesthetic orientation, and risk-taking.

Proposals regarding self-actualization and creativity can also be considered within the personality tradition. According to Maslow, boldness, courage, freedom, spontaneity, self-acceptance, and other traits of this kind lead a person to realize his or her full potential. Rogers described the tendency toward self-actualization as having motivational force and being promoted by a supportive, evaluation-free environment.

Focusing on motivation for creativity, a number of theorists, such as Amabile, Crutchfield, McClelland, and Golann, have hypothesized the relevance of intrinsic motivation, need for order, need for achievement, and other motives. Amabile and her colleagues have conducted important research on intrinsic and extrinsic motivation. Studies using motivational training and other techniques have manipulated these motivations and observed effects on creative performance tasks, such as writing poems and making collages.

Finally, the relevance of the social environment to creativity has also been an active area of research. At the societal level, Simonton has conducted numerous studies in which levels of creativity in eminent people over large spans of time in diverse cultures have been statistically linked to environmental variables. These variables include, among others, cultural diversity, war, availability of role models, availability of resources (such as financial support), and number of competitors in a domain. Cross-cultural comparisons by Lubart, and anthropological case studies, have demonstrated cultural variability in the expression of creativity. Moreover, they have shown that cultures differ in the value they place on the creative enterprise.

Evolutionary Approaches

The evolutionary approach to creativity was instigated by Campbell in 1960. Campbell suggested that the same kinds of mechanisms that have been applied to the study of the evolution of organisms – namely, selection and retention – could be applied to the evolution of ideas. This suggestion has been enthusiastically followed by a number of investigators, such as Simonton.

The basic idea underlying this approach is that there are two basic steps in the generation and propagation of creative ideas. The first is 'blind variation', by which the creator generates an idea without any real idea of whether the idea will be successful, in the sense of being selected for, in the world of ideas. Indeed, Simonton has argued that creators have no knowledge of which of their ideas will succeed. Therefore, their best strategy for producing lasting ideas is to produce a large quantity of them. Simonton argued that their 'hit' rate remains relatively constant throughout their professional life; in other words, a fixed proportion of their ideas will succeed. The more ideas they have in all, the more of their ideas will achieve success.

The second step is 'selective retention'. In this step, the field in which the creator works either retains the idea for the future or lets it die. Thus, some ideas a creator believes are creative may not be so viewed by the field, and they are not picked up by others. Those ideas that are selectively retained are the ones that are judged by the field to be novel and of value, that is, creative. Blind variation and selective retention are described further by Cziko.

Confluence Approaches

Much recent work on creativity, by Amabile, Csikszentmihalyi, Gardner, Gruber, Sternberg, Lubart, Mumford, Gustafson, and others, hypothesizes that multiple components must converge for creativity to occur. Sternberg, for example, examined laypeople's and experts' conceptions of the creative person. People's implicit theories contain a combination of cognitive and personality elements, such as 'connects ideas', 'sees similarities and differences', 'has flexibility', 'has aesthetic taste', 'is unorthodox', 'is motivated', 'is inquisitive', and 'questions societal norms'.

At the level of explicit theories, Amabile has described creativity as the confluence of intrinsic motivation, domain-relevant knowledge and abilities, and creativity-relevant skills. The creativity-relevant skills include: a cognitive style that involves coping with complexities and breaking one's mental set (way of looking at problems) during problem-solving; knowledge of heuristics for generating novel ideas, such as trying a counterintuitive approach; and a work style characterized by concentrated effort, an ability to set problems aside, and high energy.

Gruber and his colleagues have proposed a developmental 'evolving-systems' model for understanding creativity. A person's knowledge,

purpose, and affect grow over time, amplify deviations that the person encounters, and lead to creative products. Developmental changes in the knowledge system have been documented in cases such as Charles Darwin's thoughts on evolution. 'Purpose' refers to a set of interrelated goals, which also develop and guide an individual's behavior. Finally, the affect or mood system records the influence of joy or frustration on the projects undertaken.

Csikszentmihalyi has taken a different 'systems' approach, highlighting the interaction between the individual, the domain, and the field. The individual draws upon information in a domain and transforms or extends it via cognitive processes, personality traits, and motivation. The field, consisting of people who control or influence a domain (e.g. art critics and gallery owners), evaluates and selects new ideas. The domain, a culturally defined symbol system, preserves and transmits creative products to other individuals and future generations.

Gardner has conducted retrospective case studies that suggest that the development of creative projects may stem from an anomaly within a system (e.g. tension between competing critics in a field) or moderate asynchronies between the individual, the domain, and the field (e.g. unusual individual talent for a domain). In particular, Gardner has analyzed the lives of seven individuals who made highly creative contributions in the twentieth century, each specializing in a different one of the 'multiple intelligences': Sigmund Freud (intrapersonal), Albert Einstein (logical–mathematical), Pablo Picasso (spatial), Igor Stravinsky (musical), T. S. Eliot (linguistic), Martha Graham (bodily–kinesthetic), and Mohandas Gandhi (interpersonal). Charles Darwin would be an example of someone with extremely high naturalist intelligence. Gardner points out, however, that most of these individuals actually had strengths in more than one intelligence, and that they had notable weaknesses in others (e.g. Freud's weaknesses may have been in spatial and musical intelligences). Of course, retrospective case studies are subject to biases due to selection of information, and eventually need to be supplemented by prospective analyses (i.e. prediction of future behavior).

Sternberg and Lubart have proposed an 'investment' approach to creativity, according to which creative people are viewed as 'good investors'. They are people who 'buy low and sell high' in the world of ideas. They generate ideas that may be seen by others as 'a bit odd' and perhaps even as undesirable ('buying low'). They then try to persuade other people of the value of their ideas. Having done so, at least to some extent, they 'sell high', moving on to the next seemingly strange idea.

Sternberg and Lubart suggest, therefore, that creative people 'defy the crowd': they see things in their own way and try to persuade others to see things in this way. Sternberg and Lubert measured creativity by asking people to do a variety of different tasks – for example, to write short stories with titles such as 'The Octopus's Sneakers' or 'Trapped'; or to draw sketches with titles such as 'The Beginning of Time' or 'The Earth from an Insect's Point of View'; or to produce interesting advertisements for 'dull' products such as a new brand of doorknob or the Internal Revenue Service; or to indicate how we would know if there were extraterrestrial aliens among us seeking to escape detection. Creativity was assessed for each task in terms of novelty, quality, and task appropriateness. Using these kinds of measures, Sternberg and Lubart found that creativity is largely domain-specific: people who are creative in one domain are not necessarily creative in another. They also found creativity to be only weakly related to intelligence as it is traditionally defined.

WISDOM

Wisdom is called into play for tasks requiring good judgment and common sense. How is wisdom to be understood?

Implicit-Theoretical Approaches

Implicit-theoretical approaches to wisdom have in common the search for an understanding of people's folk conceptions of what wisdom is. Thus, the goal is not to provide a 'psychologically true' account of wisdom, but rather an account that is true with respect to people's beliefs, whether these beliefs are right or wrong.

Some of the earliest work of this kind was done by Clayton, who found three factors underlying people's conceptions of wisdom: experience; pragmatism; understanding; and knowledge.

Holliday and Chandler also used an implicit-theoretical approach to understanding wisdom. Analysis of one of their studies revealed five underlying factors: exceptional understanding; judgment and communication skills; general competence; interpersonal skills; and social unobtrusiveness.

Sternberg reported a series of studies investigating implicit theories of wisdom. Six components emerged: reasoning ability; sagacity; learning from

ideas and environment; judgment; expeditious use of information; and perspicacity.

Explicit-Theoretical Approaches

Explicit theories are constructions of theorists and researchers rather than of laypeople. In the study of wisdom, most explicit-theoretical approaches are based on constructs from the psychology of human development.

The most extensive program of research has been that conducted by Baltes and his colleagues. For example, Baltes and Smith gave adult participants life-management problems, such as 'A 14-year-old girl is pregnant. What should she, what should one, consider and do?' and 'A 15-year-old girl wants to marry soon. What should she, what should one, consider and do?' Baltes and Smith tested a five-component model on participants' protocols in answering these and other questions, based on a notion of wisdom as expert knowledge about fundamental life matters or as good judgment and advice in important but uncertain matters of life.

Three kinds of factors – general person factors, expertise-specific factors, and facilitative experiential contexts – are proposed to facilitate wise judgments. These factors are used in life planning, life management, and life review. Wisdom is then reflected in five components: rich factual knowledge (general and specific knowledge about the conditions of life and its variations); rich procedural knowledge (general and specific knowledge about strategies of judgment and advice concerning matters of life); lifespan contextualism (knowledge about the contexts of life and their temporal (developmental) relationships); relativism (knowledge about differences in values, goals, and priorities); and uncertainty (knowledge about the relative indeterminacy and unpredictability of life and ways to manage this). An expert answer should reflect more of these components, whereas a novice answer should reflect fewer of them. The data collected have generally been supportive of the model.

Over time, Baltes, Staudinger, and their colleagues have collected a wide range of data showing the empirical utility of the proposed theoretical and quantitative approaches to wisdom. For example, Staudinger, Lopez, and Baltes found that measures of intelligence and personality, and their interface, overlap with but are not identical to measures of wisdom in terms of constructs measured; and Staudinger, Smith, and Baltes showed that human-services professionals outperformed a control group on wisdom-related tasks. They also showed that older adults performed as well on such tasks as did younger adults, and that older adults did better on such tasks if there was a match between their age and the age of the fictitious characters about whom they made judgments. Baltes, Staudinger, Maercker, and Smith found that older individuals nominated for their wisdom performed as well as did clinical psychologists on wisdom-related tasks. They also showed that up to the age of 80, older adults performed as well on such tasks as did younger adults. In a further set of studies, Staudinger and Baltes found that performance settings that were ecologically relevant to the lives of their participants and that provided for actual or 'virtual' interaction of minds increased wisdom-related performance substantially.

Although most developmental approaches to wisdom are ontogenetic, Csikszentmihalyi and Rathunde have taken a philogenetic or evolutionary approach, arguing that wise ideas must have been selected for over time, at least in a cultural sense. In other words, wise ideas should survive better over time than unwise ideas in a culture. They define wisdom as having three basic dimensions of meaning: a cognitive process, or a particular way of obtaining and processing information; a virtue, or socially valued pattern of behavior; and a good, or personally desirable, state or condition.

A definition of wisdom proposed by Sternberg draws upon the notion of balance. Wisdom is defined as the application of practical intelligence, as mediated by values, towards the goal of achieving a common good through a balance among multiple interests (intrapersonal, interpersonal, and extrapersonal), responses to environmental contexts (adaptation to and shaping of existing environmental contexts, and selection of new environmental contexts), and time frames (long-term and short-term).

According to this definition, the essence of wisdom is balance. People can be intelligent, in the sense of being analytically keen, but not wise, if they fail to balance others' interests or the interests of institutions with their own. Wisdom, then, requires a use of intellectual abilities to benefit not only oneself, but also others and society. Current research on this paradigm is seeking to validate the theory and to show that it is possible to teach wisdom-related thinking to young children.

Further Reading

Baltes PB and Smith J (1990) Toward a psychology of wisdom and its ontogenesis. In: Sternberg (1990), pp. 87–120.

Carroll JB (1993) *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York, NY: Cambridge University Press.

Clayton V and Birren JE (1980) The development of wisdom across the life-span: a reexamination of an ancient topic. In: Baltes PB and Brim OG (eds) *Life-Span Development and Behavior*, vol. III, pp. 103–135. New York, NY: Academic Press.

Csikszentmihalyi M (1996) *Creativity*. New York, NY: HarperCollins.

Gardner H (1983) *Frames of Mind: The Theory of Multiple Intelligences*. New York, NY: Basic Books.

Gardner H (1993) *Creating Minds*. New York, NY: HarperCollins.

Jensen RB (1998) *The g Factor*. Greenwich, CT: Greenwood.

Neisser U (ed) (1998) *The Rising Curve*. Washington, DC: American Psychological Association.

Piaget J (1972) *The Psychology of Intelligence*. Totowa, NJ: Littlefield Adams.

Sternberg RJ (1985) *Beyond IQ: A Triarchic Theory of Human Intelligence*. New York, NY: Cambridge University Press.

Sternberg RJ (ed.) (1990) *Wisdom: Its Nature, Origins, and Development*. New York, NY: Cambridge University Press.

Sternberg RJ (1997) *Successful Intelligence*. New York, NY: Plume.

Sternberg RJ (1998) A balance theory of wisdom. *Review of General Psychology* **2**: 347–365.

Sternberg RJ, Forsythe GB, Hedlund J *et al*. (2000) *Practical Intelligence in Everyday Life*. New York, NY: Cambridge University Press.

Sternberg RJ and Grigorenko EL (eds) (1997) *Intelligence, Heredity, and Environment*. New York, NY: Cambridge University Press.

Learning Aids and Strategies

Intermediate article

John Sweller, University of New South Wales, Sydney, Australia

CONTENTS

Knowledge about human cognitive architecture has transformed instructional design. The fact that we have a limited working memory and an effectively unlimited long-term memory holding automated schemata can be used to derive principles that guide the manner in which information is presented and the activities learners should engage in.

INTRODUCTION

Cognitive science has transformed our understanding of learning. That transformation has provided, for the first time, a coherent theoretical basis from which to generate and test instructional hypotheses. As a consequence, the field of instructional design now has a unity and integrity that was previously missing, and learning aids and strategies have been devised that would have been difficult or impossible to devise without a knowledge of human cognitive architecture. This article discusses that architecture along with the instructional principles that derive from our deepening understanding of cognitive theory.

HUMAN COGNITIVE ARCHITECTURE

Described below are several structural features and processes, along with their general implications for instructional design. In particular, working memory, long-term memory, schema construction, and automation will be discussed. Other important aspects of cognitive architecture (e.g. sensory memory) will not be discussed because they have not had a comparable impact on instructional design principles.

Working Memory

In describing the characteristics of working memory we are describing the characteristics of consciousness. The severe limitations of working memory and its division into auditory and visual components will be discussed here.

Working memory is limited both in duration and in capacity. If people attempt to recall lists of nonsense syllables and are prevented from rehearsing, 50% of what was memorized is lost after 3 seconds and almost everything is lost after 18 seconds. These results indicate that working memory, the source of all conscious learning, thinking, and problem solving, is capable of holding new material for no more than a few seconds at a time. Capacity limitations are also severe. Working memory can hold no more than about four to seven items of new information at a time. Furthermore, it should be noted that working memory is not normally used to simply hold information. It is the processing engine of the human cognitive system, where processing involves combining, contrasting or manipulating information in some manner. The number of items of information that can be processed in working memory, as opposed to just held, is probably about two or three, depending on the nature of the manipulations required.

Working memory can be divided into separate auditory and visual components (Baddeley, 1992). Auditory working memory is used to deal with auditory material, primarily speech but also music and other sounds, while visual working memory deals with two- or three-dimensional objects. The two processors are partially independent. Total working memory capacity can be increased by providing different information to each. That the independence is partial rather than complete can be seen from the fact that the amount of information that can be handled by both processors simultaneously is substantially less than the sum of the amounts that can be handled by the processors in isolation.

Long-Term Memory

Over the last few decades, long-term memory has gradually and surprisingly emerged as central to human cognitive processing. Its importance may far exceed that of working memory. Several points concerning the characteristics of long-term memory need to be considered.

Firstly, all learning, from rote memorization to deep understanding, can be fully characterized by changes in long-term memory. If there are no changes in long-term memory, nothing has been understood or learned.

Secondly, since we are only conscious of the very limited material in working memory, at any given time we are unconscious of most of the material held in long-term memory.

Thirdly, long-term memory capacity is enormous and far beyond any current measurement techniques. It is effectively unlimited.

The importance of long-term memory in high-level cognitive processing was first revealed in the 1940s by de Groot's (1965) work using the game of chess. He found no differences between chess grandmasters and weekend players in depth or breadth of search. Both groups considered approximately the same number of alternative move sequences when faced with a board configuration taken from a real game. In contrast, there were dramatic differences in memory of configurations. If shown a board configuration taken from a real game for a few seconds, grandmasters could reproduce almost the entire configuration, while novices could reproduce very little of it. This difference was not due to working-memory differences because it disappeared when random board configurations were used. The ability of chess grandmasters derives not from a mysterious reasoning skill but from holding tens of thousands of board configurations, and the best move associated with each configuration, in long-term memory. Less able players must make constant use of limited working memory to attempt to derive appropriate moves for most configurations. Grandmasters can defeat them because they know which move is superior for many more configurations. All educated people learn similar skills in the variety of areas covered by their education.

Schemata

The structure of knowledge held in long-term memory is described by schemata, cognitive constructs that allow us to treat several elements of information as a single element classified according to the way it will be used (Chi *et al*., 1981). We have schemata for trees which allow us to treat the myriad of elements that go to make up a tree as a single element. Schemata for trees allow us to recognize something as a tree despite the fact that every tree has a unique shape, color, and context in time and place. Schemata for the letter *a* allow us to recognize this letter despite its having an infinite variety of forms when handwritten. Higher-order schemata allow us to recognize combinations of letters that make up words and combinations of words that make up phrases and sentences. If we have elementary algebraic knowledge a schema tells us instantly that the first move to solve the equation $(a+b)/c=d$ for a is to multiply by c. Chess grandmasters have schemata that allow them to recognize board configurations and the best moves associated with them.

Schemata, held in long-term memory, can be brought into working memory, where they may indicate appropriate actions and problem-solving procedures in everything from everyday life to the complex and sophisticated processes taught in educational institutions. Furthermore, because they can be treated as single elements in working memory, they massively reduce the load on working memory. Once learning through schema construction has occurred, the temporal and capacity limitations of working memory become irrelevant (Ericsson and Kintsch, 1995).

Automation

Information can be processed in working memory either consciously or automatically (Shiffrin and Schneider, 1977). Conscious processing is bound by the limitations of working memory discussed above. In contrast, a schema that has been automated can be processed in working memory with a minimal cognitive load. We can read text easily and rapidly because of the automated schemata we have developed for letters, words, and combinations of words. We can solve mathematical problems, including very complex problems, because a sufficient number of the required schemata have been constructed and automated to permit the additional processing in working memory needed to find a solution.

In summary, the learning mechanisms of schema acquisition and automation locate large numbers of automated schemata in long-term memory. Those schemata permit working memory to engage in the processes required to govern not only everyday life but also the more sophisticated activities required in educational contexts.

DIRECT INSTRUCTION

Our cognitive architecture has implications for the manner in which information should be presented to learners and the activities that they should engage in. If learning requires alterations to schematic structures in long-term memory then instruction should be primarily concerned with techniques that will facilitate schema construction and automation. Furthermore, if all new information must be processed in a limited working memory prior to incorporation into schemata held in long-term memory, then it is imperative that instructional techniques be organized in a manner that takes account of the limitations of working memory. Until the relatively recent incorporation of knowledge of cognitive architecture into instructional design principles, most instructional design recommendations ignored both the limitations of human working memory and the fact that sophisticated cognitive processing is dependent on the prior construction of innumerable schemata held in long-term memory.

The characteristics of human cognitive architecture demand the use of direct instruction rather than the indirect instruction associated with discovery learning, investigatory learning, learning through problem-solving or constructivist learning procedures. While the terminology has periodically changed, all of these techniques are essentially indistinguishable. They all require direct instruction to be reduced, or in extreme cases eliminated entirely, with learners discovering principles and procedures by constructing them with minimal direct input from external sources. On the 'investigatory learning' view, teachers should minimize direct presentation of information and maximize opportunities for students to elucidate information themselves. It is a view that is incompatible with our knowledge of the structures of human cognitive architecture.

Consider people learning through problem-solving (Sweller, 1988). They might be given a series of mathematical problems to solve. If they lack previously acquired schemata that indicate which moves should be made to solve the problems, they must derive solutions using problem-solving search. Learners must consider their current problem state, the goal, differences between the current problem state and the goal, and problem moves that could be used to reduce those differences. These activities must occur in the only structure available for such activities: working memory. The limitations of working memory will always ensure that the process is cognitively demanding. Furthermore, while the ultimate aim of the exercise is the acquisition of schemata, the problem-solving process bears no relation to schema construction.

One alternative, discussed in more detail below, is to directly demonstrate the solution procedure to students. Such direct instruction reduces cognitive load by reducing or eliminating search and substituting a procedure that shows students which problem states or configurations should be associated with which moves. Learning to recognize problem configurations and their associated moves is the essence of schema construction.

On this analysis, direct instruction is normally preferable to investigatory alternatives. Using limited working memory to discover procedures and relations that could easily and simply be demonstrated is an inefficient way to construct schemata. Not only should direct instruction be a preferred mode of teaching because it conforms with human cognitive architecture, but that instruction itself should be structured to reduce the load on working memory and facilitate schema construction. We will now look at procedures designed to accomplish these aims.

COGNITIVE LOAD THEORY

Cognitive load theory uses the cognitive architecture described above, along with some structural features of information, to design effective instruction (Sweller, 1999). If learning involves the construction of schemata held in long-term memory and processing of information occurs in a limited working memory, then instruction should be directed to assisting students to acquire the specific schemata required in a given subject area, and should be organized in a manner that reduces the load on working memory. The reduction of cognitive load on working memory is especially important if the subject matter imposes a heavy cognitive load because of its intrinsic nature. Some material imposes a heavy load on working memory irrespective of how it is presented. It is essential that the instructional procedures used to teach such material do not impose an additional, unnecessary load. Below, we shall be looking at several of the learning procedures directly based on cognitive load theory. Later we will consider other learning aids and strategies that, although not based on cognitive load theory, derive from the cognitive architecture discussed above.

Worked Examples

Problem-solving has been used as a learning and teaching device in many areas, especially mathematics and science, for a long time. As indicated above, discovery learning and investigatory and constructivist techniques are all based on, and essentially indistinguishable from, problem-solving. These techniques became increasingly popular teaching and learning devices in the decades preceding the early 1990s. Since that time, their popularity in the academic literature appears to have waned. They are incompatible with basic human cognitive architecture; and empirical evidence for their effectiveness is extremely sparse. If, as indicated above, problem-solving imposes a heavy cognitive load which interferes with schema acquisition and automation, an alternative is required. Worked examples provide such an alternative.

A worked example poses a problem and provides a solution to that problem. It provides a 'road map' to a problem solution. The analogy is a good one. Just as we would consider it foolish to expect someone to find their way around a new region without a road map or direct instruction, so it is foolish to expect a learner to find their way around a problem in mathematics, science or any other area without explicitly being shown the way. It can be done, but it can be a long, slow, and frustrating exercise.

We continue to use road maps until we are reasonably confident that we know our way. Normally, we will only attempt a route without a map once we have learned it. Similarly, it is of benefit to students to continue studying worked examples until they are confident that they understand the solution to the particular type of problem they are studying. Only then does it becomes useful to commence solving problems without the assistance of worked examples.

There are both theoretical reasons and data supporting a heavy use of worked examples as a learning aid. From a theoretical perspective, based on cognitive load theory, searching for a problem solution imposes a heavy load on working memory. Search normally involves simultaneously considering the current problem state and the goal state, extracting differences between the two states, and finding a problem-solving operator to reduce those differences. Furthermore, for learners confronted with problems in a new area, neither the various problem states encountered nor the problem-solving operators required to move from one state to another state will be familiar – automated schemata will not be available. The load on working memory may be far in excess of its capacity. The entire problem-solving process may collapse, reducing the chances of the construction of problem-solving schemata.

As an example, consider a student attempting to solve the equation $(a+b)/c=d$ for a. The initial state is $(a+b)/c=d$ and the goal state is a valid equation with a as the subject of the equation. In order to reduce differences between the initial state and the goal state, elimination of the addend or denominator on the left-hand side is required, and the rules of algebra provide the problem-solving operators to accomplish this task. To make the first move, each of these elements, including the nine elements in the initial equation and also including the many relations between the various elements, must be processed in working memory. That task is simple for anyone with a relevant problem-solving schema which conglomerates all of these elements into a single entity that can be handled readily in working memory. But it is virtually impossible for a learner who has just begun to solve problems in this area. The learner must make repeated failed attempts until a sufficient part of the procedure has been incorporated into schemata that do not exceed the limits of working memory. The process is likely to be long and slow.

In contrast, consider a learner who is given the following worked example to study:

$$(a+b)/c=d \Rightarrow a+b=dc \Rightarrow a=dc-b \qquad (1)$$

There is very limited problem-solving search, and the load on working memory is minimal. The example demonstrates the elements needed to construct a schema to solve problems of this type. It provides the start state, the goal state, and the moves required to transform the start state into the goal state. Working-memory resources are not required for problem-solving search.

The empirical evidence that studying worked examples is superior to solving the equivalent problems is conclusive. In controlled experiments conducted by many researchers around the world, it has been repeatedly demonstrated that if learners study worked examples rather than solve the equivalent problems, learning is enhanced and test problems are solved more accurately and more rapidly. Furthermore, there are extensive data indicating that the reason for this superiority is the reduction in the load on working memory.

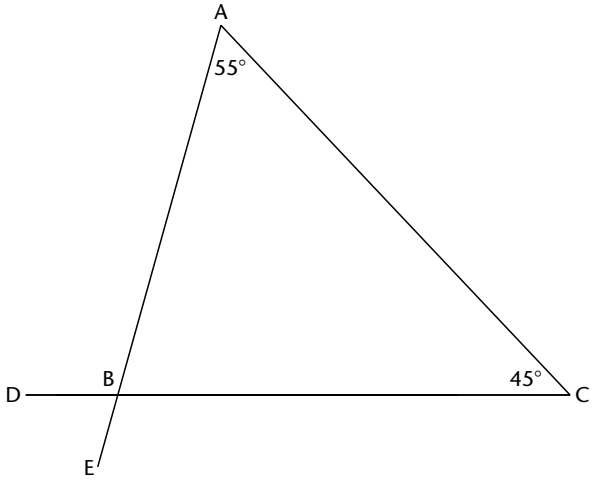
Designing Effective Instruction

Not all worked examples are effective. A worked example is only likely to be more effective than

solving the equivalent problem if it reduces unnecessary cognitive load. Unless they are appropriately structured, worked examples can also impose an unnecessary cognitive load. Cognitive load theory can be used to provide procedures for constructing effective worked examples. Furthermore, those procedures generalize to all instruction, not just worked examples.

Consider the geometry worked example of Figure 1. In isolation, the diagram does not indicate the solution; and the statements are unintelligible without the diagram. The example only becomes intelligible when the diagram and the statements have been mentally integrated. Integration requires working-memory resources, and those resources then become unavailable for schema acquisition. Geometry worked examples structured in this fashion are ineffective because they impose a heavy cognitive load.

In contrast, if they are restructured so that mental integration of disparate sources of information is not required, cognitive load is reduced and learning is enhanced. Figure 2 provides the same information in an integrated form, obviating the need for resource-demanding integration. Many experiments have demonstrated the advantages of physically integrated worked examples – and indeed of any other form of instruction consisting of multiple sources of information that cannot be understood in isolation. The advantage of integrated over split-attention instructions is known as the split-attention effect.



In the above figure, find angle DBE.

Solution:

Angle ABC = 180°– angle BAC – angle BCA (internal angles of a triangle sum to 180°)

= 180° – 55° – 45°

= 80°

Angle DBE = angle ABC (vertically opposite angles are equal)

= 80°

Figure 1. Geometry worked example in conventional split-source format.

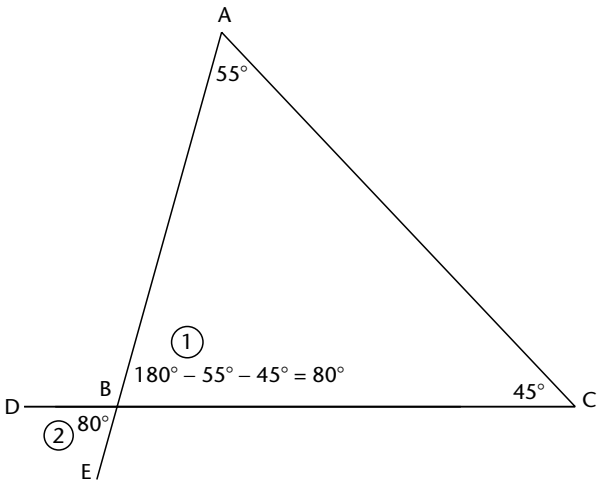


Figure 2. Geometry worked example in integrated format.

Physical integration of multiple sources of information is not always effective. It is required only when the two sources of information are unintelligible in isolation and so must be mentally or physically integrated. Frequently, multiple sources of information can be understood in isolation because one source is just a redescription of another source in a different mode or form. For example, textual material, rather than being essential to understanding a diagram (and vice versa), may simply describe the content of the diagram in verbal form. The text is redundant, and rather than being integrated with the diagram should be eliminated because processing it imposes an extra cognitive load. Because the diagram can be understood without the text, learning is enhanced by eliminating the text. This effect is called the redundancy effect, and has been demonstrated in many experiments.

There is an alternative to physical integration when faced with split-attention material. Working memory is divided into multiple channels or processors: at the least there are separate auditory and visual processors. While there is a large overlap between these processors, in the sense that using one processor does not leave the other completely free, they are sufficiently distinct to increase total working-memory capacity if used simultaneously. When faced with split-attention materials, such as a

diagram and text that must be processed together in order to be understood, rather than physically integrating them, the same advantage can be obtained by presenting the text in auditory rather than visual (written) form. Using both auditory and visual processors increases working-memory capacity, thus facilitating learning. This effect is known as the modality effect. It should be noted that dual-modality presentation should be used only under split-attention conditions. If used under redundancy conditions (where, for example, textual material is redundant), dual-modality presentation is ineffective. Redundant auditory text is no more useful than redundant written text. Accordingly, the common practice of presenting written material with an identical voice-over should be avoided.

All of the effects discussed above occur because different instructional designs impose different cognitive loads on working memory, and because the material being learned itself has a high intrinsic cognitive load. A high intrinsic cognitive load occurs when the structure of the elements being learned is such that they cannot be processed meaningfully in working memory serially, but instead must be processed simultaneously. For example, when considering the algebraic equation discussed above, we cannot conduct a manipulation and just consider the denominator on the left-hand side. The left-hand and right-hand sides must be considered simultaneously because any manipulation of one will have an effect on the other. The elements interact. Material with high element-interactivity imposes a high load on working memory, and it is only such material that requires instructional designs that reduce that load. Material with low element-interactivity, such as a second-language vocabulary, does not impose a heavy intrinsic cognitive load, and so instructional designs that derive their efficacy by reducing cognitive load will be ineffective. Other strategies are required for such material.

NOTE TAKING

Research on note taking has distinguished between the encoding that occurs when students take notes and the storage whereby students learn from notes (Kiewra, 1989). Encoding relies heavily on working memory, while storage occurs in long-term memory. Important findings from this area can be interpreted within the framework of the cognitive architecture discussed above.

With respect to encoding, one might expect that the act of encoding through taking notes might facilitate learning. In fact, the results are mixed, with some studies finding that students who take lecture notes learn more than students who merely listen to the lecture, others finding no effect, and a few finding listening to be superior to note taking. These mixed results may be a consequence of the attentional benefits of note taking being negated by the increased working-memory load associated with simultaneously taking notes and processing the lecture content.

Another finding is that students take better notes and learn more if they are provided with a framework for taking notes, such as incomplete outlines of a lecture that need to be filled in. Such a framework should reduce working-memory load by reducing the need for students to generate their own framework.

With respect to storage, reviewing notes has the expected effect of facilitating retention. Retention is improved if students are provided with notes rather than required to construct their own. This may be partly because the requirement to encode interferes with schema formation, and partly because provided notes tend to be better structured than student-constructed notes.

Notes presented to students need to be structured with human cognitive architecture in mind. In a clear example of the redundancy effect, experiments have indicated that students presented with a complete text of a lecture learn less than students presented with outlines or notes in tabular or matrix form. The redundant material of the complete set of notes is likely to overload working memory and interfere with schema construction. Another study has found that studying notes in an unrestricted manner results in superior learning to writing an essay based on the notes. In this context, essay writing is a redundant activity that interferes with learning. (In other contexts, such as learning to write, essay writing is, of course, a useful rather than a redundant activity.)

SIGNALING

Text signals (Lorch, 1989) are non-content devices intended to increase the intelligibility of text. They do not provide any substantive information, but rather direct the reader to process the information in a particular manner. The nature of text signals, and whether a particular text signal is effective, follow directly from the cognitive architecture discussed above.

The intelligibility of text will be increased if working-memory load is decreased. Text that reduces the extent to which inferences must be made and is organized to facilitate schema construction

will reduce cognitive load. Appropriate signals contribute to these aims.

Signals are important because of the structure of human cognitive architecture. If humans had unlimited working memory, the cognitive-load-reducing aspects of signaling would be unnecessary. If understanding and learning did not involve the acquisition of schemata held in long-term memory, text signals that assist in the construction of schemata would have no function. While not necessarily including essential content, signals help the reader to understand content. Because of the nature of human cognitive architecture, it is difficult to derive meaning from text consisting purely of content without signals.

There are many text signals used by effective writers. Several of the more important are discussed below.

Enumeration

Numbering a set of points or a list provides a signal that in some cases indicates a preferred or necessary order, but often signals that the content of the list is important and should be carefully noted. Readers will attend more closely to numbered points and will remember them better. While the numbers provide no content, they signal that the enumerated material should be given preference in limited working memory and that schema construction should be based on that material.

Print Style

While writers have schemata that tell them which aspects of a text are particularly important, readers of unfamiliar content do not, and in the absence of appropriate signals, must use working-memory resources to make inferences concerning levels of importance. Those memory resources then become unavailable for schema acquisition. Print styles such as italics, boldface text, or underlining reduce the need for readers to infer levels of importance, free working-memory capacity, and so can enhance schema acquisition.

Headings

Most text consists of a series of more or less connected ideas or topics. Where connectivity is relatively low, writers need to signal to readers that a new topic is beginning. Again, the reason is to reduce unnecessary inferencing on the part of the reader, reducing cognitive load. The reader is told by the heading that a new section is beginning, rather than having to infer it. Relevant signals can be included in the body of the text, but it is often easier and more effective to break the text by the use of appropriate headings. These not only indicate that a break has occurred, but can indicate the nature of the break.

Preview Sentences

Preview sentences signal the more important aspects of forthcoming information. Material signaled in preview sentences is more closely attended to and better remembered. This allows students to cluster together subsequent information, facilitating schema construction.

Recall Sentences

Recall sentences signal backwards, rather than forwards like preview sentences. Summary statements at the end of a section provide an example. Recall sentences signal to the reader the important aspects of the previous material, assisting in the reduction of cognitive load.

CONCEPTUAL MAPS

Schemata are not only essential structural features of human cognitive architecture; they can also be used as a learning aid. In this role, they have been called by a variety of names, including 'advance organisers', 'conceptual maps', and 'mental models'.

Pre-existing schemata can be used to organize and interpret new information. Learning is facilitated if learners are informed, before commencing learning, of how the new information relates to their previous knowledge. The previously acquired schemata can not only then organize the incoming information, but in the process can themselves be augmented by it. In effect, this form of learning consists of the construction of ever more sophisticated schemata.

As an example, consider a student beginning to learn algebra. The student is already familiar with arithmetic and knows that $3+4=7$. The schemata associated with addition and subtraction can be used to assist the student with the problem: '$\square+4=7$, fill in the box'. Once this procedure is established by the construction of a well-automated schema, the next step is to replace the box with a symbol, '$a+4=7$', indicating at the same time that calculating a value for a is identical to calculating the number that goes in the box. In this manner, schemata for basic algebra evolve from schemata

for addition and subtraction. The initial schemata for algebra can be considered to be alterations or derivations of the schemata for arithmetic. The precision and hierarchical nature of mathematics makes the need for connecting new material to previous schemata obvious. It is also usually obvious which previously-acquired schemata should be used. In other areas with a less well-defined knowledge base, the need to connect new material to previously acquired schemata is less obvious. Nevertheless, the advantages can be just as great.

The use of schemata to assimilate new information is more likely to be effective if the initial schemata are well established. Students are likely to find it difficult to use a newly-learned schema that is not yet automated to assist in the interpretation of new information. Working-memory resources are not likely to be available to both process the original schema and appropriately incorporate the new material into the recently-acquired schematic structures. A major reason for using previous schemata is to reduce working-memory load. The purpose is likely to be defeated if poorly-established initial schemata are used.

ELABORATION

New material that is to be learned can be processed and rehearsed at shallower or deeper levels. Material processed and rehearsed at a shallow level is simply repeated, as we might repeat a telephone number prior to telephoning. Elaborative processing and rehearsal require information to be related to other information, and in that sense, are deeper. The telephone number might be analyzed to see if it can be segmented in a manner that makes it easier to learn. For example, the number '306230' could be divided up as '306 230' or as '30 62 30'. The latter is likely to be easier to remember because '30' is repeated and can be treated as a single number for which there is likely to be a well-entrenched schema. Elaboration makes heavier demands on cognitive resources but provides a more effective form of learning.

In recent years, considerable work has been carried out on 'self-explanations' of information contained in instructional forms such as worked examples. Self-explanation is a form of elaboration that occurs when learners, rather than passively processing new information, attempt to explain it to themselves. When studying a worked example they might attempt to explain to themselves how and why each step is being made. Chi *et al*. (1989), who initiated this line of work, found that more able problem-solvers are more likely to elaborate worked examples in this manner than less able problem-solvers. That result has been replicated on several occasions and is now well established.

Why do better problem-solvers engage in greater self-explanation? One possibility is that a more able problem-solver has better-established schemata and greater working-memory capacity. Self-explanations are cognitively demanding and may require considerable working-memory capacity. A large working-memory capacity and established schemata will result in more capacity being available for self-explanations. On this interpretation, self-explanations do not improve problem-solving skill, but rather, working-memory capacity determines both problem-solving skill and self-explanation activity.

A more interesting possibility is that self-explanations assist learners to process information more deeply, resulting in increased problem-solving skill. Evidence for this explanation comes from results demonstrating improved problem-solving following instructions and training in self-explanations (e.g. Renkl *et al*., 1998). Results such as this indicate that training in self-explanations can be an effective learning aid.

ILLUSTRATIONS

The proverb 'a picture is worth a thousand words' expresses the truth that, for some types of material, a pictorial or diagrammatic representation is much easier to process than the equivalent text (Larkin and Simon, 1987). Diagrams can be a far more effective form of presenting information than text because diagrams make spatial relations explicit. Text does not clearly indicate the relations of elements to other elements, and so mental representations of those spatial relations must be constructed in working memory. In addition, text must be processed sequentially, while diagrams can be processed simultaneously, making it easier to detect multiple relations between elements. Levin and Mayer (1993) indicate that diagrams are more concise and perceptually clear than the equivalent textual statements. These factors combine to substantially reduce the cognitive load associated with diagrammatic, compared with textual, information. Illustrations work because they reduce working-memory load.

The different processing requirements of language-based and visually-based material have instructional design consequences. Whenever instruction deals with two- or three-dimensional space, it is likely to be more effective if it is

presented in pictorial form rather than in words. This effect is accentuated where the information is complex or difficult to understand (Levin and Mayer, 1993), because it is precisely such material that requires a reduction in cognitive load. By the same token, illustrations that do not reduce cognitive load appreciably are not likely to be effective. If cognitive load is low, or if illustrations have a purely decorative function, illustrations are not likely to act as an instructional aid.

MNEMONIC TECHNIQUES

Not only can the use of illustrations be an effective instructional device, but having learners generate their own images can, under appropriate circumstances, facilitate learning (Levin, 1986). Imagery, along with other mnemonic devices, can be a highly effective learning aid. Unlike the procedures described above, the effectiveness of mnemonic devices is restricted to learning low-element-interactivity material, which has very low levels of natural connectedness. Lists of items, such as a second-language or a scientific vocabulary, are examples. There is no evidence that mnemonic devices can assist in learning high-element-interactivity material that is difficult to understand because its natural structure imposes a heavy working-memory load.

Mnemonic devices rely on connecting new material to a well-established schema, often pictorial in nature. The 'peg' method requires learners to memorize a series of 'pegs' to which the new material must be attached using visual imagery. The pegs can consist of any well-memorized series of words to which the new material is attached by visualizing a connection between each of the words and each item of the new material. A grocery list might be remembered by imagining each grocery item in relation to one of the pegs in the previously-memorized list. Thus, if the memorized list consists of the common rhyme 'one is a bun, two is a shoe…', one might remember that milk is the second item on the grocery list by imagining milk poured into a shoe.

The method of loci is identical to the peg method except that, rather than a memorized list of words, a well-known location such as one's home is used. Items to be learned are attached to locations in one's home by imagining them in relation to those locations. A carton of milk might be imagined balancing on a door.

The 'link' method, use of stories, and initial-letter methods, are related techniques that do not require previously-memorized material. In the link method, an image of each item is constructed and then each image is linked to the next image. Milk being poured on bread enables one to remember to buy bread and milk. Stories link the items by inserting them into a story. The initial-letter method incorporates the first letter of each word into an acronym.

The 'keyword' method is a two-stage procedure that can be very effective in learning a language vocabulary. The first stage involves an acoustic link, the second a visual-imagery link. If one is learning that the Spanish word *caballo* means 'horse', the first step might be to connect *caballo* to the word 'cab' while the second might be to imagine a horse hailing a cab. There is a large body of evidence indicating that the keyword method can be very effective.

STUDY SKILLS

Many of the above descriptions have touched on study-skills instruction. The cognitive architecture that governs all instructional design principles equally governs the activities that learners should engage in when studying. Learners need to adopt procedures that reduce working-memory load and facilitate schema construction and automation. There are several study techniques that meet these requirements and that have empirical support (Pressley and Schneider, 1997).

Learners can be directly shown how to extract the main ideas from text by processes that resemble the use of worked examples. Demonstrating to students by underlining the main ideas of a text and having students practice underlining improves subsequent text comprehension. Modeling comprehension procedures such as connecting new material to previous knowledge, summarizing passages, thinking of questions associated with new material, and making inferences and predictions, can all be used to show students how to approach new material. Directly demonstrating these procedures rather than having students discover the procedures themselves reduces working-memory load and facilitates schema construction.

Having learners imagine procedures that need to be learned is a highly effective study technique (Cooper *et al*., 2001). If the material is already sufficiently well learned that it can be held in working memory, then instructing students to turn away from the material and imagine it is more effective than asking them to study it. If the material is not sufficiently well learned to be held in working memory, studying is more effective than imagining. If only very limited prior learning has occurred, instructions to imagine tend to fail

because insufficient working memory is available with which to imagine.

CONCLUSION

Until recently, instructional design principles and recommended learning aids and procedures tended to be more or less random. The lack of an effective theoretical base meant that there was no unifying set of principles that could be used to generate and test instructional hypotheses. Advances in cognitive science, and specifically a better appreciation of the relevance of our knowledge of cognitive architecture, have provided a secure base which can be used both to explain why some procedures are effective and to generate new procedures. As a consequence, cognitive science is now central to instructional design.

References

Baddeley A (1992) Working memory. *Science* **255**: 556–559.

Chi M, Bassok M, Lewis M, Reimann P and Glaser R (1989) Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* **13**: 145–182.

Chi MTH, Feltovich P and Glaser R (1981) Categorization and representation of physics problems by experts and novices. *Cognitive science* **5**: 121–152.

Cooper G, Tindall-Ford S, Chandler P and Sweller J (2001) Learning through imagining. *Journal of Experimental Psychology: Applied*.

De Groot A (1965) *Thought and Choice in Chess*. The Hague, Netherlands: Mouton. [First published 1946.]

Ericsson KA and Kintsch W (1995) Long-term working memory. *Psychological Review* **102**: 211–245.

Kiewra K (1989) A review of note-taking: the encoding–storage paradigm and beyond. *Educational Psychology Review* **1**: 147–172.

Larkin J and Simon H (1987) Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* **11**: 65–99.

Levin J (1986) Educational applications of mnemonic pictures: possibilities beyond your wildest imagination. In: Sheikh A (ed.) *Imagery in the Educational Process*, pp. 202–265. Farmingdale, NY: Baywood.

Levin J and Mayer R (1993) Understanding illustrations in text. In: Britten B, Woodward A and Binkley M (eds) *Learning from Textbooks: Theory and Practice*, pp. 95–113. Hillsdale, NJ: Lawrence Erlbaum.

Lorch R (1989) Text-signalling devices and their effects on reading and memory processes. *Educational Psychology Review* **1**: 209–234.

Pressley M and Schneider W (1997) *Introduction to Memory Development During Childhood and Adolescence*. Mahwah, NJ: Lawrence Erlbaum.

Renkl A, Stark R, Gruber H and Mandl H (1998) Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology* **23**: 90–108.

Shiffrin R and Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* **84**: 127–190.

Sweller J (1988) Cognitive load during problem solving: effects on learning. *Cognitive Science* **12**: 257–285.

Sweller J (1999) *Instructional Design in Technical Areas*. Melbourne, Australia: ACER Press.

Further Reading

Bruning R, Schraw G and Ronning R (1999) *Cognitive Psychology and Instruction*, 3rd edn. Upper Saddle River, NJ: Merrill.

Chi M, de Leeuw N, Chui M and LaVancher C (1994) Eliciting self-explanations improves understanding. *Cognitive Science* **18**: 439–477.

Ericsson KA, Krampe RT and Tesch-Romer C (1993) The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406.

Kotovsky K, Hayes JR and Simon HA (1985) Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology* **17**: 248–294.

Mayer R (1997) Multi-media learning: are we asking the right questions? *Educational Psychologist* **32**: 1–19.

Renkl A (1997) Learning from worked out examples: a study on individual differences. *Cognitive Science* **21**: 1–29.

Sweller J, van Merriënboer JJG and Paas FGWC (1998) Cognitive architecture and instructional design. *Educational Psychology Review* **10**: 251–296.

Van Merriënboer JJG (1997) *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training*. Englewood Cliffs, NJ: Educational Technology Publications.

Learning and Instruction, Cognitive and Situative Theories of

Intermediate article

Sharon J Derry, University of Wisconsin, Madison, Wisconsin, USA
Constance A Steinkuehler, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Cognitive theory, which views cognition as symbolic computation, and situativity theory, which views cognition as (inter)action in the social and material world, are two alternative theoretical perspectives on the nature of human learning. These theories are contrasted in terms of their implications for educational research and practice.

INTRODUCTION

During the past 10 years, social science researchers have engaged in heated debate over which of two theoretical viewpoints can best guide both the study of human thinking and the design of environments for productive learning and work.

One view, the symbolic processing perspective, represents the tradition upon which cognitive science was founded. Although this perspective does not represent a unified, homogenous theory, influential theorists from Carnegie-Mellon University (e.g. Anderson *et al*., 1996) have been important voices in this debate, and their position is widely known and accepted. This viewpoint has encouraged not only cognitive scientists, but also the public, to think and talk about the mind using a computational metaphor, a metaphor based on thinking about the mind in terms of digital computers (some, but not all, current thinking based on distributed network computation is more supportive of the alternative viewpoint to be described). The mind is configured as an information processing 'machine' that receives input from the environment through the senses, selectively and actively processes that information by constructing and reconstructing mental symbols and symbol systems representing knowledge and action, and stores some of those constructions as memories for later recall and use, in an indexed 'in-the-head' repository called 'memory'. Cognitive scientists combine experimental methods and computational modeling to advance and test hypotheses and complex theories about the basic computational structures of mind and the symbols and mechanisms of a variety of forms of human thought, such as language acquisition, remembering, problem-solving, decision-making, and emotional response. Many claim that this 'traditional' cognitive program continues to significantly advance our basic scientific knowledge about human thought processes, as well as our knowledge about how to design educational and work environments that facilitate learning and performance.

Here, the symbolic processing perspective is interpreted broadly to include modern socio-cognitive theory. Evolving from Jean Piaget's (e.g. 1952) genetic epistemology, current versions of this view hold that people learn through a process in which their existing conceptual knowledge is challenged and transformed through social and physical interaction with the environment. Although many socio-cognitive theorists do not explicitly advocate a computer metaphor of mind, their viewpoint is nevertheless a symbolic processing one that posits existence and transformation of symbolic structures (schemas, concepts) within the mind.

A view that challenges this position is 'the situativity perspective' (e.g. Greeno, 1997), which references a family of important social science theories including 'situated cognition' (e.g. Lave, 1988), 'sociocultural theory' (Wertsch, 1998), 'embodiment theory' (Glenberg, 1997), 'distributed cognition' (e.g. Hutchins, 1995), and 'activity theory' (Nardi, 1996). Researchers associated with this view are united in their rejection of viewpoints

that separate study of an individual's mind from the environment of which that mind is an integral part. This is not to say that situativity theorists reject cognitive science entirely. However, symbolic theories of learning and thinking are often viewed as special cases of cognition – those cases in which there is conscious reasoning. These theorists point out many ways in which mind is connected to, extended and shaped by the cultures, body, and the physical surroundings of which it is a part, and they do not believe that all cognition involves symbolic representation or, indeed, that all cognition takes place within the mind itself. For example, a researcher using a computer program and consulting an expert in performing a statistical analysis is extending cognitive capacity by relying on social support and employing a cultural artifact (the statistical package) to perform a task that the researcher need not fully understand in all its complexity and depth, given the availability of the social environment and its tools. From the situativity perspective, the cognition is 'stretched' over the entire activity (the problem-solving task) and partly resides in the social context, as well as in the individual's mind. Moreover, the individual need not symbolically represent all cognition within the activity system. For example, a well-designed statistical program would extend cognitive capacity by being so easy and obvious, the researcher would not need to think much in order to use the tool correctly. Intelligent interactions are thus 'afforded' by tool design features, of which the users may not be aware.

Taken together, symbolic processing and situativity theories provide the foundations for an important and growing field of study known as 'the learning sciences', which is currently having tremendous impact on educational research. This article will elaborate and explain these positions in greater detail, discussing their implications for educational research and design of learning environments. A summary of these implications is provided in Table 1, which highlights important differences in the theoretical schools.

SYMBOLIC PROCESSING THEORY

Before symbolic processing theory began to emerge in the 1960s, psychology was dominated by theories of behaviorism that treated human behavior as nothing more than direct response to environmental stimuli ($S \rightarrow R$). Symbolic processing theory rejected this assumption, concluding that human behavior could not be explained without positing an intermediate stratum of mental processes that occur *between* input (stimuli from the environment) and output (behavior). Human beings, it was argued, mentally represent information from the environment, process that information, then select behaviors accordingly. The mind, from this perspective, is an information processing system analogous to a computer: a physical symbol system. Input/output processes connect the symbol system with its environment, but it is between perception and action that 'thinking' (symbolic computation) occurs. Detailed computer models of 'in-the-head' processes have been constructed and successfully tested against specific forms of human performance, such as solving verbal analogies or algebra problems with two unknowns; however, they have been less successful in explaining complex, adaptive performance, such as expert medical practice.

This modeling approach to theory building is contingent, however, on a set of basic ontological assumptions. Such models assume cognition is a bounded, dependent but autonomous physical symbol system. The system is dependent in the sense that it is open to information from the environment, yet it is autonomous in the sense that its basic architecture is closed to reorganization. Information from the external world is the material the system operates on, but basic architectural changes are not determined by the outside; therefore, events 'in the head' can be factored from external events in the social and material environment and thereby accurately modeled by computer programs.

Behind this 'factoring assumption', in turn, is a fundamental division between the individual knower, knowledge, and the independent world based on the dualist ontological tradition of Kant and Descartes. From this perspective, the individual is a container with a sort of substance (albeit, symbol-based) called 'knowledge' inside. Learning is the acquisition, construction, and qualitative reorganization of this substance (knowledge), and the success of the learning process is measured by the transfer (application) of this substance from one place (the context in which the knowledge was acquired) to another (a different context in which that knowledge should be used). Therefore, useful knowledge is knowledge that is sufficiently abstract and general to allow for successful behavior across a wide range of relevant contexts (Anderson *et al.*, 1996).

SITUATIVITY THEORY

If symbolic processing theory unpacked the black box of mental representation and procedures

Table 1. Situative and symbolic theories: alternative lenses for educational research and practice

Educational issue	*Situative emphases*	*Symbolic processing emphases*
Major epistemic categories	People acting in context Interwoven systems of activity Discourse structures Design and use of artifacts and tools Nonsymbolic processes	Minds in the environment Memory stores Information Attention Mental representations Symbolic processes
Learning and development	*Individual:* Internalization of practice Trajectories of participation Identity development Becoming a leader *Community:* Development of norms Development of artifacts Tool development and use Community knowledge growth	*Individual:* Comprehension Knowledge growth Knowledge restructuring Knowledge transfer Development of expertise
Transfer	Enhanced ability to participate in new, complex, community activities Enhanced ability to participate in new communities	Analogical (schema-based) reasoning Flexible situational construction (cognitive flexibility)
Learning environment characteristics	Discourse communities Classrooms and schools extend broader social community Authentic activities Tools of authentic practice Varied levels of participation Mentoring and apprenticeship	Cognitive objectives Mastery of prerequisites Explanation-based learning Group and individual work Problem-solving tasks Skill practice Memorization Use of representational tools and manipulatives
Teacher roles	Provide community leadership Design community infrastructure Promote development of social and discourse norms Share problem-solving Provide for mentoring, apprenticeship	Transmit and explain Set cognitive objectives Perform task analyses Design/select activities and materials Individualize instruction Provide feedback/guidance Lead discussions Challenge misconceptions Promote metacognitive reflection
Assessment	Evaluation of student-designed artifacts Portfolio assessments Evaluation of authentic performance in context Dynamic (formative) assessment	Measures of facts recall Comprehension tests Problem-solving tests Standardized summative assessments Standardized diagnostic assessments Performance-based assessments Transfer tasks

between environmental stimuli and individual behavior that behavioral theory refused to open, then 'situativity theory' (Greeno, 1997) is attempting to unpack the black box of activity structures (structures of interactions of individuals within their material and social contexts) de-emphasized by symbolic processing theory. Growing out of work in ecological psychology, ethnography, ethnomethodology, and philosophical situation theory (Greeno, 1997), situativity theory focuses

on interactive systems of activity of which the individual is only one part. Cognition, from this perspective, cannot be explained by computational models of structures and processes 'in the head'; rather, one must look to the intact activity systems in which the individual participates. Such systems always necessarily include social relationships, physical and temporal contexts, symbolic and material resources (such as tools), and historical change. From this perspective, cognition is 'a complex social phenomenon … distributed – stretched over, not divided among – mind, body, activity, and culturally organized settings (which include other actors)' (Lave, 1988, p. 1). The structures of interest, then, are the interactional structures of such social and material systems, not structures in the individual mind.

Thus, cognition is (inter)action in the social and material world. According to Lave (1988), the basic organizing structures of this world are the social, cultural, and professional groups, or the 'communities of practice', in which people choose to participate. Through participation in a community of practice, individuals come to understand the world (and themselves) from the perspective of that community. In contrast to symbolic processing theory, which takes the meaning of a symbol as given, situativity theory focuses on how meaning evolves through enculturation. Here, semantic interpretation is taken as part of what people do in the lived-in world; it arises through interaction with social and material resources in the context of a community with its own participant structures, values, and goals. An individual becomes attuned to a particular object's meanings and uses through the regular pattern of interaction that individual has with it, but this regular pattern of interaction is shaped by the individual's membership in a particular community for whom the object has meaning, usefulness, and relevance for accomplishing tasks associated with individual or collective goals.

Such activities have direct import for the identity of the individual. Who one *is* determines, and is reflexively determined by, one's participations in various communities (Greeno, 1997). A community delineates practices for its members, provides means of forging identities, of possible ways to apprehend and understand the environment through social and physical activity with tools and symbols. Thus, changes in knowing become changes in being: through participation in communities of practice, an individual does more than merely acquire and reorganize symbolic knowledge about the world; she or he is ontologically transformed by it.

This conception of cognition as a culturally mediated, historically developing, and activity-based process assumes a nondualist ontology in sharp contrast to the assumptions of symbolic processing theory. This nondualist tradition can be traced back through the work of theorists as diverse as Marx, Heidegger, Vygotsky, and Dewey, to the work of Hegel (1807/1967), who argued that 'the individual self is in no sense an immediately given element of consciousness (as Descartes claims of his *cognito*) but a socially created concept … we are wholly social products and social participants' (p. 514). The mind, the individual, and the world with which we interact are not natural entities but historical and cultural products determined by human practices; their meaning – what they 'are' – is constituted through human activity. As such, activities and contexts are mutually constitutive of each other rather than one nested inside the other.

While symbolic processing theory focuses on epistemological processes and the mental architecture and functions that sustain them, situative theory focuses on the genesis of participation in communities of practice. Accounts of how an individual interacts with her material and social contexts, and how these interactions change over time, replace accounts of individual knowledge construction occurring 'in the head'. Learning, from this perspective, is progress along 'trajectories of participation' and growth of identity within a given community of practice (Greeno, 1997). It is the gradual transformation of an individual from peripheral participant to central member of a community, through apprenticeship and increased acceptance of community values and increased participation in community practices. Thus, a new teacher (or lawyer, or doctor, or researcher) acquires professionally relevant knowledge and skill as she increases in prestige and power within a professional community. Importantly, learning also takes place at the aggregate level of the community, a process that involves emergent reorganization in the patterns of member activities, coupled with a growth of shared knowledge through changing practices and the creating of artifacts and tools that facilitate work.

ALTERNATIVE LENSES FOR EDUCATIONAL RESEARCH AND PRACTICE

The situativity and symbolic processing theory families represent fundamentally different lenses through which to analyze, design, and conduct research on educational environments. Some typical differences in points of view are highlighted

in Table 1 and discussed in the following paragraphs.

With respect to epistemic categories, situativity theory views knowledge, not as individual mental representation, but as something that resides within communities and manifests itself through what members of the community do and create. Thus, designers and researchers working from this perspective view learning environments in terms of their observable activity structures, including the systems of interaction and discourse among learners and teachers, and the development and use of tools and artifacts within those systems. An analytical problem for situativity researchers is decomposition of the environment for study, since activities are constituted by other activity structures as contexts. Thus, situativity researchers always study activity structures qualitatively and ethnographically *in situ*. Most studies place little emphasis on specifically describing internal cognitions. Rather, situativity researchers may postulate nonsymbolic 'cognitive' processes and structures that are external to the individual mind and can be inferred from classroom observations.

In contrast, the major analytical structures of the cognitive program are the unobservable symbols and processes within minds that are inferred from performance on tests and tasks designed to allow inferences about individuals' possession of facts, concepts and skills. For researchers, such inferences support theories about human learning, knowledge, and performance, and about the connection between learning and features of instructional design. These theories may be specified in detail and translated into computer models that are tested for matches against human performance. In such analyses, individual students or teachers, not activity structures in the environment, are the units for analysis. Moreover, because learning and performance can legitimately be examined separately from the broader social and physical contexts in which they take place, laboratories, as well as classrooms, may be sites for educational research.

Table 1 highlights the two theoretical families' differing conceptualizations of learning and development, and how these imply different ways of supporting the learning and assessment of students within educational environments. Situativity theorists see learning and development in terms of individuals' growing capabilities as participants in multiple authentic communities of practice, communities that can, themselves, learn and grow organically as knowledge-building entities. This viewpoint has implications for conceptualizing and designing learning environments, for the roles for teachers within those environments, and for assessing student performance. To the greatest extent possible, classrooms become extensions of authentic communities of practice found in broader society. For example, students of teacher education are not (conceptually speaking) taught or assessed on the skills and facts of teaching, but are apprenticed into a culture of professional practice. Mathematics students are likewise apprenticed into the culture of mathematicians (when appropriate) or (at other times) the culture of a mathematically informed citizenry. Teachers are viewed, not just as transmitters and challengers of knowledge and designers of instructional tasks that help students construct individual understandings of subject domains, but as founders of classroom learning communities that extend and connect to authentic cultures of practice, and as leaders and mentors and fellow learners within those communities. The emphasis of assessment in such communities is on production and critique of authentic, socially valued products and performances with concern for how individuals interact in social context while working.

In contrast, classrooms designed from the symbolic processing perspective tend to engage students in a variety of instructional activities that are designed to help them, as individuals, acquire specific cognitive objectives. While these objectives may be derived from an analysis of what knowledge is required for desired, complex performances in the real world, and while effort is made to connect classroom learning to real-world issues and problems, from the symbolic perspective, the classroom activity in which students participate need not replicate authentic, real-world practice in social context. Through various activities, some authentic and some nonauthentic, students are expected to acquire fundamental symbol structures that will later be recalled, combined, and used as analogies to guide thinking and behavior in the world outside of class. Understanding and transfer of knowledge is promoted through reflection (rather than enactment), and by teaching explicit symbolic knowledge about the conditions under which one will use what is learned. The teacher is expected to help students engage in thinking about their thinking (metacognition) during learning – asking themselves what they do and do not understand; how they will use it in future, etc. From the symbolic perspective, assessment focuses on design of reliable, repeatable tasks and instrumentation that can be used to measure use of specific ideas and skills across many individuals and different forms of tasks that require those ideas and skills.

A SUBSIDING DEBATE

Much debate continues between members of these two communities. Yet, many educational researchers, theorists, and practitioners today are successfully fusing both points of view within their work (e.g. Bransford *et al*., 1999). Researcher–designers and teachers working in school environments are likely to view classrooms as communities and to attend to activity and discourse structures and the authenticity of required activities and assessments but may also regard these as contexts or vehicles for individual learning viewed as symbolic processing. They may chart trajectories of participation for particular students, and simultaneously examine development of symbolic knowledge using a wide range of cognitive tasks and instruments. In larger studies, educational researchers may study classroom communities with varying activity and discourse structures, developing coding systems to capture these differences and statistically examining the impact of these variations on individual performance measured using standardized assessment tools.

Thus, although the symbolic and situative viewpoints have not yet merged to form a well-defined theory of educational practice, we see much evidence that a dialectic process is moving the field toward rapprochement between them. We believe the emerging result may be a complex systems theory of cognition understood in its broadest ecological sense, and that the resulting methodological approach will be superior to either theoretical viewpoint standing alone, capable of providing more complete understanding of learning and education.

References

Anderson JR, Reder LM and Simon HA (1996) Situated learning and education. *Educational Researcher* **25**(4): 5–11.

Bransford J, Brown AL and Cocking RC (eds) (1999) *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academy Press.

Glenberg AM (1997) What memory is for. *Behavioral and Brain Sciences* **20**: 1–55.

Greeno JG (1997) On claims that answer the wrong questions. *Educational Researcher* **26**(1): 5–17.

Hegel GWF (1967) *The Phenomenology of Mind*, translated by JB Baillie. New York, NY: Harper & Row. [Original work published in 1807.]

Hutchins E (1995) *Cognition in the Wild*. Cambridge MA: MIT Press.

Lave J (1988) *Cognition in Practice: Mind Mathematics and Culture*. New York, NY: Cambridge University Press.

Nardi BA (ed.) (1996) *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge, MA: MIT Press.

Piaget J (1952) *The Origins of Intelligence*. New York, NY: International University Press.

Wertsch JV (1998) *Mind as Action*. New York, NY: Oxford University Press.

Further Reading

Derry S (1992) Beyond symbolic processing: expanding horizons for educational psychology. *Journal of Educational Psychology* **84**: 413–418.

Engestrom Y and Middleton D (eds) (1996) *Cognition and Communicaton at work*. Cambridge, UK: Cambridge University Press.

Gee JP (2000–2001) Identity as an analytic lens for research in education. *Review of Research in Education*. Washington, DC: American Educational Research Association.

Kirshner D and Whitson JA (eds) (1997) *Situated Cognition: Social, Semiotic and Psychological Perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Lave J and Wenger E (1991) *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

Resnick LB, Levine JM and Teasley SD (eds) *Perspectives on Socially Shared Cognition*. Washington, DC: American Psychological Association.

Salomon G (ed) (1993) *Distributed Cognitions: Psychological and Educational Considerations*. Cambridge, UK: Cambridge University Press.

Literacy: Reading (Early Stages)

Introductory article

Frank R Vellutino, State University of New York, Albany, New York, USA

CONTENTS

Literacy is the ability to read and write. Early literacy development consists of a prealphabetic period wherein the child begins to learn about the nature and purpose of print; an alphabetic period wherein the child becomes conversant with the alphabetic code and acquires increasingly functional word identification and text processing skills; and an advanced alphabetic/orthographic period wherein the child masters the alphabetic code and becomes increasingly fluent in word identification and text processing.

INTRODUCTION

Becoming fully literate is, perhaps, one of the greatest of the developing child's many achievements. It is also one of the most daunting. Whereas the acquisition of spoken language skills can be accomplished simply through immersion in a natural language environment, the acquisition of written language skills depends on formal instruction, deliberative intent, and cognitive effort. This is because learning to read and write entails the recoding of spoken language in the form of an arbitrary set of symbols, organized and implemented in accord with conventions that are unique to the orthography (writing system) that uses those symbols. Thus, whereas in an ideographic orthography, such as written Chinese, the individual characters represent meaning-bearing units of language, in alphabetic orthographies, such as written English, Spanish, Swedish, and German, the characters represent phonemes – that is, individual speech sounds such as the /c/, /a/, and /t/ in 'cat'. In the European orthographies the alphabetic characters representing written words are organized from left to right and the words themselves are read from left to right, whereas in written Hebrew the alphabetic characters are organized from right to left and the words are read from right to left. Accordingly, the first task of the novice reader is to learn about the nature and purpose of print, the structural properties and print conventions associated with the writing system to be studied, and, ultimately, the functional use of that system. This is a time-consuming and challenging enterprise which depends heavily on the normal development of relevant cognitive abilities, as well as on the enrichment and tuition provided by the child's home and school environments and on the child's motivation for and investment in becoming literate. The nature of the cognitive demands made on the child in acquiring early literacy skills will vary, depending on the properties of the orthography studied. The exposition below focuses on literacy acquisition in an alphabetically based orthography, exemplified by written English.

COGNITIVE FOUNDATIONS OF LITERACY

Written words are encoded (symbolized) representations of spoken words, which themselves are encoded representations of knowledge about concepts in the environment stored in permanent memory. Thus, the ability to learn to read and write depends heavily on linguistic coding processes that facilitate encoding (symbolizing) of concepts in the form of spoken words; linguistic and visual coding processes that facilitate encoding of linguistic units in the form of written words and word constituents; and visual-motor processes that, along with linguistic knowledge, facilitate writing. Given normal development of these processes, and adequate exposure to print and literacy instruction, early reading and writing development depends further on metalinguistic and metacognitive processes that allow the child to analyze and come to understand the structural characteristics of spoken and written words, in terms of the mapping relationships between them, as well as on memory processes that facilitate storage, retrieval, and integration of their visual, linguistic, and motor counterparts.

Figure 1 sets out the contribution made by each of these processes to the acquisition of literacy. The first steps depict the inception of literacy in terms of linguistic coding processes that facilitate the acquisition of lexical knowledge and language acquisition in general. These include phonological coding (the ability to use speech codes to represent information); semantic and morphological coding (the ability to store information about the meanings signified by words and word parts, such as 'ing', 'ed'); syntactic coding (the ability to store word order rules that govern how words are organized in sentences); and pragmatic coding. Pragmatic coding is the ability to store information about conventions governing the use of language as a medium for communication: for example, the relationship between a sentence and the broader context in which it is embedded (prior text, situations represented by the text, background knowledge, etc.) or the implicit contract between the speaker and hearer (or reader and writer) in terms of conventions that facilitate communication and understanding (choice of topic, turn-taking, etc.). Such processes not only facilitate the acquisition of lexical knowledge in the form of spoken and written words, but also the comprehension of spoken and written language during listening and reading respectively. (*See* **Language and Cognition**; **Language Comprehension**)

The linguistic coding processes are paralleled by the visual coding and visual-motor processes that contribute to spoken and written language acquisition. These include visualization ability (the ability to store and retrieve representations defining the visual characteristics of environmental stimuli, including the graphic symbols used to represent written words) and visual-motor ability (the ability to store and retrieve the motor programs required for learning to produce written letters and words). Both linguistic and visual coding processes contribute to the establishment of connective bonds between the spoken and written counterparts of printed words, in the interest of helping the child acquire 'sight words': printed words identified (named) on sight as unanalyzed wholes. However, given the heavy load on visual memory imposed by the high degree of visual similarity characteristic of printed words composed of alphabetic characters (e.g. 'pot' and 'top'; 'was' and 'saw'), linguistic abilities carry much greater weight as determinants of the ability to learn to read than do visual and visual-motor abilities, especially at the beginning of

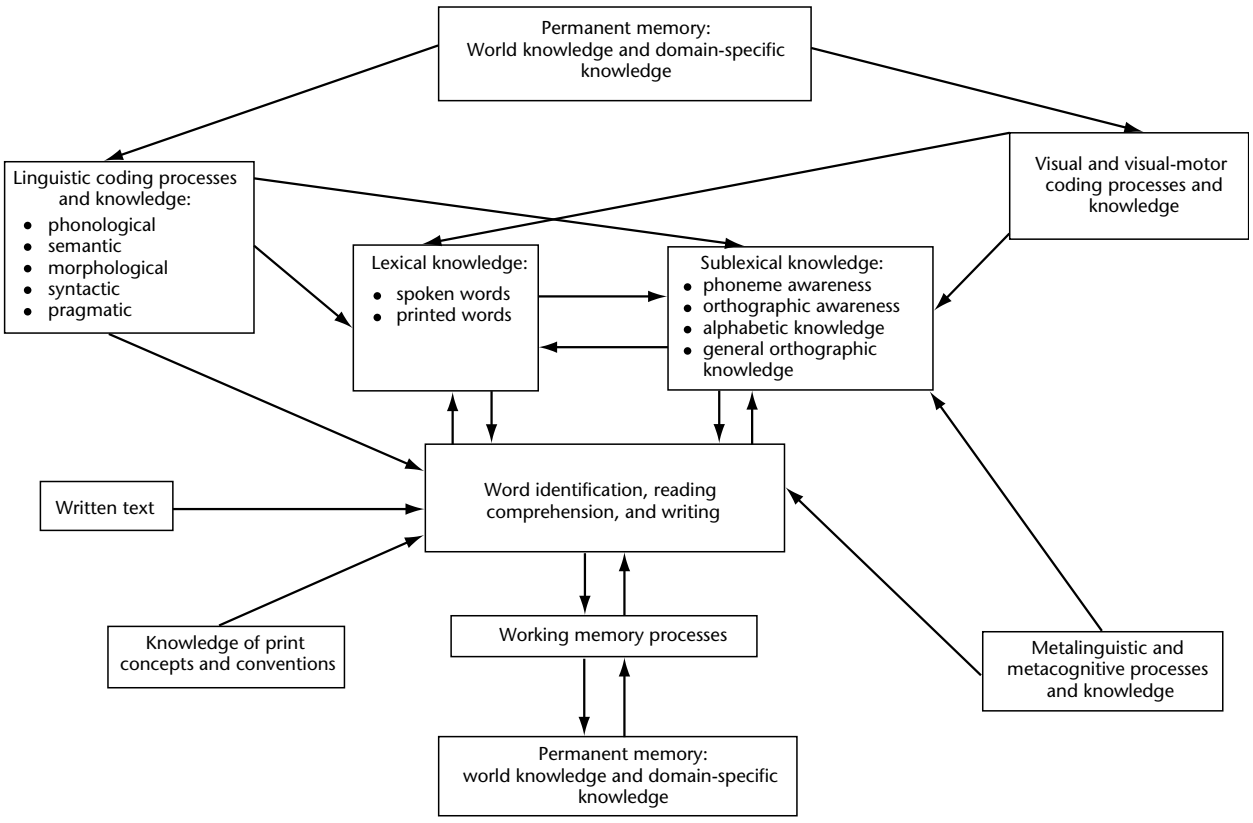


Figure 1. Cognitive processes and types of knowledge entailed in learning to read and write.

literacy development. Phonological coding ability is of special importance, not only because it helps the child learn sight words, but also because it facilitates acquisition of the types of sublexical knowledge that help the child acquire proficiency in letter-sound decoding, which is the primary vehicle that beginning readers use for reducing the load on visual memory imposed by an alphabetic writing system.

Sublexical knowledge is knowledge about the structural characteristics of spoken and written words. Such knowledge includes phoneme awareness – conceptual understanding and explicit awareness that spoken words consist of individual speech sounds (phonemes) – and orthographic awareness – sensitivity to the constraints on how letters in words are organized. Many scholars believe, and many studies have documented, that phoneme awareness is a prerequisite for acquiring alphabetic knowledge, that is, knowledge about letter-sound relationships. Similarly, orthographic awareness is believed to be important for distinguishing between 'legal' and 'illegal' combinations of letters in written English ('vid' is 'legal', 'xqr' is illegal), and both forms of conceptual knowledge are believed to have a mutually facilitative relationship to one another. Phoneme awareness facilitates orthographic awareness by virtue of the positive effect phoneme awareness has on letter-sound analysis, and, thereby, the proper sequencing of letters in both word identification and spelling. Orthographic awareness facilitates phoneme awareness, by virtue of the positive effect orthographic awareness has on letter organization and, thereby, letter-sound analysis in word identification and spelling. Together, these two types of knowledge help consolidate alphabetic knowledge. They also help the child detect and make functional use of orthographic regularities and redundancies (e.g. 'at' in 'cat' and 'rat'; 'ing' in 'running' and 'playing'), and successful acquisition of this more general and higher-level orthographic knowledge ultimately leads to mastery of the alphabetic code.

Two other types of conceptual knowledge are important components of early literacy development: knowledge of print concepts and conventions, and metalinguistic and metacognitive knowledge (Figure 1). Knowledge of print concepts and conventions reflects the child's understanding that printed words represent words and sentences in spoken language; that they are composed of letters carrying sound values; that they have spaces between them; that they are processed from left to right, and so forth. Metalinguistic knowledge is knowledge the child acquires through active analysis of the structural characteristics of spoken language – its sound structure, its syntax, and its conventional use. Metacognitive knowledge is a more general type of cognitive knowledge and is acquired through active analysis of the operations and strategies employed in any type of cognitive processing, not just language processing. Both types of knowledge help to lay the foundation for early literacy development by their contribution to the child's acquisition of basic word decoding and word identification skills, as well reading comprehension and text processing skills.

Finally, Figure 1 depicts both the permanent memory and the working memory processes involved in learning to read and write. It also depicts the interrelationships and reciprocity between the different coding and memory systems (double-direction arrows) involved in (a) establishing firm associations between lexical and sublexical components of spoken and printed words, and (b) encoding, storing, and retrieving information required for learning to read and write, and information processed during reading and writing. (*See* **Memory, Long-term**; **Working Memory**)

PERIODS OF LITERACY DEVELOPMENT

Prealphabetic Period

Literacy development begins long before the child enters school and is greatly dependent on the home environment. Such seminal experiences define what has often been called 'emergent literacy'. The child's home environment influences literacy development in three important ways: through language enrichment, through print enrichment, and through endorsement. Language enrichment is important because adequate language development is a prerequisite for adequate literacy development. The importance of phonological skills such as phoneme awareness and letter-sound decoding in learning to read has been well established by training, intervention, and classroom observation studies, which have shown that children who acquire these skills (among others) become better readers and spellers than children who do not acquire these skills. Preschool literacy studies have consistently found that children from language-enriched home environments that sensitize them to the phonological properties of spoken words tend to profit more from school literacy experiences and instruction than do children from less language-enriched environments. For example, preschool children who are exposed to nursery

rhymes (which typically involve much rhyme and alliterative language), rhyming games, and other activities that sensitize children to the similarities and differences in the sounds of words (e.g. word games such as 'Pig Latin') tend to be more successful in acquiring phoneme awareness, letter-sound decoding, and word identification skills in early reading development, than children who are not exposed to such activities.

Vocabulary enrichment can also positively influence early literacy development. It is easier for children to learn to read words that are already part of their speaking vocabulary than to learn to read words that do not have this standing. This has been established in research with children learning to read in a second language. It might also be expected that vocabulary knowledge would have a strong influence on language and reading comprehension; these expectations have been given some confirmation by prediction studies finding significant positive relationships between measures evaluating variability in the home language environments of preschool children and the acquisition of language and literacy skills in these children in later grades (kindergarten through middle school). The common finding among these studies is that children from homes where parents and siblings had well-developed speaking vocabularies and good language skills in general, and where dialog among family members was encouraged and reinforced, tended to perform better than children who came from less enriched home language environments on measures evaluating vocabulary knowledge, phoneme awareness, alphabetic knowledge, printed word identification, and reading comprehension.

Finally, a number of studies have evaluated other oral language skills in preschoolers such as comprehension and use of syntax and complexity of productive language, and have found significant positive relationships between measures of these abilities and measures of early and later literacy development.

The second way in which the child's home environment can influence early literacy development is by facilitating exposure to print. In the initial phases of the prealphabetic period (which some investigators have called the 'logographic' stage of literacy development), children do not have sufficient understanding that printed words represent functionally distinct and meaningful units of spoken language. In addition, they know little about letters of the alphabet or about the alphabetic principle, tend to treat printed words as logographs or picture-symbols, and often use incidental or even irrelevant features of these words (e.g. the two 'eyes' in the word 'look') as vehicles for retrieving the names of those words rather than their letters to do so. However, if the home environment provides the child with extensive contact with print, in the form of shared and guided reading and writing experiences (e.g. listening to and discussing stories, calling attention to the letters in words), availability of printed materials, and many opportunities to encounter environmental print (in street signs, supermarkets, etc.), the child will become increasingly conversant with print concepts and conventions and will begin to acquire seminal awareness of how the alphabet works. The child may also begin to make rudimentary attempts at writing, as manifested in random scribbles, alphanumeric characters, and perhaps even the letters of the child's name.

The third way in which the child's home environment facilitates literacy development is through endorsement: placing a high premium on learning to read and write, encouraging and reinforcing the child for actively engaging in these enterprises, and providing opportunities to do so. It will suffice to point out that there is abundant research evidence that children who come from highly enriching home environments that provide them with many opportunities for exposure to print and encouragement to actively engage print are better prepared to profit from beginning reading instruction than children who come from less enriching and less encouraging home environments.

Alphabetic Period

The acquisition of functional literacy skills occurs during what many scholars have called the 'alphabetic period' of early literacy development, which extends roughly from kindergarten through second grade for most children. During this period of literacy development, children become increasingly sensitized to the alphabetic nature of the writing system and become increasingly proficient at using letter sounds to help them read and spell printed words. In the initial phase of this period, which some have called the 'partial alphabetic' phase, children come to discriminate between and learn the names of letters of the alphabet. They also become increasingly sensitive to the structural characteristics of spoken words and begin to acquire phoneme segmentation and phoneme awareness skills. As a result, they begin to understand how the alphabet works and begin to use a few salient letter sounds to aid word identification. Although children rely heavily on a sight-word,

meaning-based approach to word identification during the partial alphabetic phase of literacy acquisition, and make generous use of rote memory as well as picture and sentence contexts to identify words they encounter in text reading, they also begin to use letter sounds to aid the identification process, especially initial and final consonant sounds. However, they are less proficient in using vowel sounds and redundant consonant clusters (e.g. 'ch', 'sh', 'th') to aid word identification during this phase of the alphabetic period, because these sounds are less salient, less reliable, and more complex than initial and final consonant sounds and require more extensive experience with print for full acquisition. Thus, children do not have word spellings fully represented in memory during the partial alphabetic period, and are often confused by words having similar spellings (e.g. 'pot/top', 'what/want', 'was/saw'). In addition, they are not yet able to use 'analogy' strategies for word identification; that is, using familiar words to help pronounce and spell unfamiliar words (e.g. using 'lamp' to help identify 'damp').

Children's writing during the initial phase of the alphabetic period is typified by what is commonly called 'invented spelling' – the use of novel, but principled and intuitively generated ways of using letter sounds to spell words the child has not yet learned to spell conventionally. Children's invented spellings during this period rely heavily on what is essentially a 'letter name' strategy for divining the sounds carried by given letters, whereby letter names are implicitly matched to salient phonemes in the words they attempt to spell (e.g. spelling 'hey' as 'ha') or to phonemes close in manner and place of articulation to given letter names (e.g. spelling 'jam' as 'gam' because 'j' and 'g' are pronounced alike). Some scholars believe that children's use of such intuitive spelling strategies constitutes the primary means by which they acquire phoneme awareness and come to understand the alphabetic principle. However, it is more likely to be the case that phoneme analysis and invented spelling are reciprocal and interdependent processes that both contribute to the child's understanding of the alphabetic principle.

With extensive experience in using letter sounds to decode and spell printed words, and given literacy instruction that facilitates alphabetic coding, children acquire a relatively complete command of grapheme–phoneme (letter–sound) correspondences. They also begin to learn about recurrent letter patterns within words that govern the way in which those words are pronounced, for example long vowel markers such as the 'silent e' (as in 'kite' and 'bike'), as well as digraphs such as 'ch' and 'sh', and diphthongs such as 'oi' that change the sounds of their component letters (as in 'chop' and 'join'). In addition, they increasingly discover words with similar spellings and pronunciations, and begin to use analogy strategies for decoding and encoding unfamiliar words. They also begin to discover words with similar spellings but dissimilar pronunciations (e.g. 'put' versus 'cut', 'nut' and 'but'), which helps make them aware of the need for diverse and flexible strategies for word decoding and spelling. Thus, as they expand their word level skills, children become less dependent on meaning-based strategies for word identification, and increasingly store new sight words in memory using both code-based and meaning-based strategies for word identification, spelling, and writing.

However, the child's alphabetic knowledge at this more mature phase of the alphabetic period (alternatively known as the 'full alphabetic' and 'within word pattern' phase) is limited to grapheme–phoneme correspondences, constraining the child's ability to decode and spell longer, multisyllabic words. This skill develops during the next period of early literacy development.

Advanced Alphabetic/Orthographic Period

During the advanced alphabetic/orthographic period of early literacy development (second grade and beyond), children increasingly discover and make functional use of the regularities and redundancies in the orthography, especially larger clusters of letters that occur in many words. These include letter clusters with regular spellings that can be sounded out letter by letter (e.g. 'est' in 'nest', 'best', and 'chest'), and irregularly spelled letter clusters that cannot be sounded out letter by letter but have relatively stable pronunciations ('ight' in 'night', 'sight', and 'light'). For example, whereas children having only grapheme–phoneme (letter–sound) correspondences stored in memory would have to use four grapheme–phoneme units to decode an unfamiliar sight word such as 'chest' ('ch', 'e', 's', 't'), children having both 'ch' and 'est' stored as units in memory would only need to use these two units to decode the word. As their inventory of redundant letter clusters grows, children become increasingly adroit at using analogy strategies for word identification and at combining newly acquired letter clusters with smaller alphabetic units already acquired (i.e. letter sounds, vowel patterns, digraphs, diphthongs) to learn new sight words. Such consolidation strategies

also help the child decode and spell multisyllabic words (e.g. 'in', 'ter', 'est', 'ing' – 'interesting'), including those governed by pattern–sound principles that cut across syllable boundaries (e.g. consonant doubling in words like 'rabbit'). Accordingly, this phase of the orthographic period has been called the 'consolidated alphabetic' or 'syllable juncture' phase of literacy development.

Some scholars also identify a more advanced phase of the orthographic period: the derivational constancy phase. During this phase, children learn about spellings and pronunciations derived from the semantic and syntactic properties of given words ('bomb/bombard', 'permit/permission'), along with other idiosyncratic features of the orthography, such as the different spellings associated with similar-sounding suffixes (e.g. 'ignorant', 'competent') and doubling conventions in words containing prefixes ('illiterate', 'immature').

The importance of alphabetic and orthographic knowledge in learning to read and spell in a writing system based on an alphabet has been well documented in empirical research, but the most compelling evidence comes from training, intervention, and classroom observation studies showing that children exposed to instructional programs that facilitate acquisition of phoneme awareness and alphabetic coding skills encounter fewer difficulties in learning to read, and perform at higher levels on measures evaluating word identification and phonological decoding skills, than children exposed to instructional programs that do not do so. Nevertheless, as the previous examples show, written English is a complex and abstract writing system that contains a large number of irregularly spelled and even many strange words ('epoch', 'aardvark') that cannot be identified or spelled accurately and fluently solely through the use of phonologically based strategies for word identification and spelling. Accordingly, several of the studies alluded to earlier have shown that the most effective classroom and remedial intervention programs are those that incorporate both code-based and meaning-based activities that promote text processing and comprehension skills along with word identification and alphabetic coding skills.

Fluency in Reading and Writing

Fluent reading is reading that is fast, accurate, effortless, and meaningful. During text reading, the meanings of the words in sentences in the text must be stored in working memory while the meanings of the sentences containing those words are 'computed'. Sentence meanings must also be stored in working memory, and because information in working memory is available only for a short period, successful computation of sentence meanings – and thereby comprehension of the text – is critically dependent on the speed and accuracy with which the words in the text are identified and understood. Fluent readers can identify and understand, with speed and accuracy, most or all of the words that might be encountered in a given text and do not have to spend much time using phonological decoding skills to identify and access the meanings of the words in the text. However, becoming a fluent reader takes a great deal of guided and appropriate practice in both word identification and text processing. Acquiring fluency in word identification depends not only on the ability to decode unfamiliar words using the alphabetic and orthographic strategies discussed above, but also on the frequency with which the child is exposed to (and successfully identifies) given words in reading and correctly spells those words in writing. It also depends on the amount and type of reading (and writing) the child does, and the availability of interesting reading materials at an appropriate level of difficulty.

Although fluency in word identification is a necessary condition for proficient reading comprehension, it is not a sufficient condition. Reading comprehension also depends on the acquisition of knowledge and skills in other areas: adequate language development in terms of both vocabulary knowledge and syntactic competence is critically important, as is the acquisition of metalinguistic and metacognitive text processing strategies such as those described above. Reading comprehension also depends on the child's exposure to different types of text and understanding of the features that define these texts (e.g. narrative versus exposition); on the child's general knowledge; and on the child's background knowledge for engaging and profiting from texts focusing on specific domains.

Reading and writing are reciprocally related skills that are mutually reinforcing. Thus, the acquisition of fluency in writing is highly dependent on the acquisition of fluency in reading, and on the amount of guided practice and feedback the child receives in spelling and writing. It also depends on whether the child's literacy experiences facilitate integration of reading and writing subskills. If children learn to spell and write the same words they are learning to read, and receive a great deal of practice in reading, spelling, and writing those words, then the words are more likely to become fully and strongly represented in memory and

be easily accessed in both reading and writing. Finally, the acquisition of fluency and proficiency in reading and writing is greatly dependent on the child's motivation, which in turn is influenced by the quality of the child's literacy experiences, both at home and at school.

That fluency in word identification is a prerequisite for reading comprehension has been amply documented in research demonstrating that speed and accuracy in word identification are the best predictors of performance on measures of reading comprehension in children whose language skills and background knowledge are sufficient for comprehending the text. The same research has demonstrated that when fluency in word identification is sufficient for text comprehension, the best predictors of performance on tests of reading comprehension are measures of language and text-processing skills and also measures of background knowledge and domain-specific knowledge. Similar results have been obtained in research evaluating creative writing skills. Finally, there is considerable evidence that motivated children have generally better reading skills than less motivated children, all other factors being equal. (*See* **Literacy: Reading (Later Stages)**; **Literacy: Writing**)

CAUSES OF EARLY READING AND WRITING DIFFICULTIES

Literacy researchers concerned with the question of why some children have difficulty becoming literate have distinguished between the immediate causes and ultimate causes of such difficulty. Immediate causes are observed deficiencies in component reading and writing subskills (e.g. word identification, letter-sound decoding, and visual-motor deficits), whereas ultimate causes are deficiencies in the cognitive abilities thought to underlie reading and writing.

Immediate Causes

A prerequisite for meaningful reading is fluent word identification, so it should not be surprising to find that a ubiquitous cause of reading comprehension difficulties is limited facility in word identification. However, this skill is itself highly dependent on the child's ability to acquire foundational skills such as knowledge of print concepts and conventions, phoneme awareness, alphabetic knowledge, higher-level orthographic knowledge, and flexible strategies for word identification. Accordingly, children who struggle with sight-word identification are generally found to have difficulty acquiring one or more of these skills – typically alphabetic and higher-level orthographic skills.

In addition to adequate facility in word identification, fluent and meaningful reading depends on vocabulary knowledge, syntactic competence, grade-appropriate knowledge of text structure, general knowledge, domain-specific knowledge, adequate development of comprehension strategies, and the habit of extensive and diverse reading. Deficiencies in one or more of these areas have been found to cause difficulties in reading comprehension, even in children who have adequate word identification skills.

Given the intrinsic and reciprocal relationships between word identification, spelling, and writing, children with significant difficulties acquiring facility in word identification inevitably have problems in spelling and writing. However, some children have reasonably well-developed word identification skills but less well-developed spelling skills. Such children are typically found to be phonetic spellers who 'overregularize' the spellings of words and do not pay sufficient attention to the structural irregularities characteristic of written English. This tendency is often associated with the failure to acquire a flexible set of strategies for acquiring conventional spellings, including rote memorization and extensive practice with irregularly spelled words.

Deficiencies in written expression have been found to be characteristic of children who have limited language and language-based skills, but such deficiencies may also be associated with gaps in general or domain-specific knowledge, perhaps caused to some extent by long-standing reading difficulties.

Ultimate Causes

The search for the ultimate causes of reading difficulties (and related spelling and writing difficulties) has a history that dates back to the latter part of the nineteenth century. Of special interest in this area of inquiry are children suffering from 'specific reading disability' (dyslexia): that is, children who have at least average intelligence, who do not have learning difficulties in other areas, and whose reading difficulties are not caused by sensory deficits, behavior problems, frequent absences from school, or socioeconomic disadvantage. The study of such children has produced a plethora of causal theories over the years, most of which have been discredited by empirical research. Of all of these theories, by far

the most popular and most ubiquitous have been those implicating deficits in the visual system.

The most prominent visual deficit theories propose that reading difficulties are caused by visual perceptual and visual-spatial processing deficits that affect letter orientation and sequencing in both reading and writing. According to such theories, confusing 'b' with 'd' or 'was' with 'saw', or writing them in mirror form – popular stereotypes associated with the term 'dyslexia' – are manifestations of optical reversibility or spatial confusion (i.e. 'seeing' and/or writing them backwards) that impair letter orientation and sequencing. However, there is now abundant evidence that impaired and normally developing readers do not differ significantly on visual processing tasks that do not depend on linguistic coding abilities. There is also evidence that the types of orientation and sequencing errors that spawned the visual perceptual and visual-spatial deficit theories of reading disability, leading to popular stereotypes of the types just exemplified, are actually caused by failure to master the alphabetic code. Consider, for example, the likely difference between children who have learned the sounds associated with the letters 'p', 'o' and 't', and children who are unaware that letters carry sound values, in distinguishing between and 'pot' and 'top' on word identification or writing tasks. Thus, the visual processing deficits implicated in optical reversibility and spatial confusion theories of reading disability are no longer considered viable explanations of difficulties in learning to read and write.

Other visual deficit theories of specific reading disability have been proffered over the years, including those implicating optical-motor deficits that (presumably) impair normal fusion and conversion of retinal images, and visual system deficits characterized by abnormally prolonged 'visual traces' that interfere with the processing of printed words from one eye fixation to the next. However, such theories are controversial, and none of the deficits implicated in these theories has been shown to be causally related to reading difficulties.

In recent years, language deficit theories of specific reading disability have gained in prominence, and there is a good deal of consensus that deficiencies in phonological skills such as phoneme analysis, letter-sound decoding, name retrieval, and verbal memory constitute basic causes of difficulties in learning to read. Correlational support for such relationships is provided by numerous studies finding robust and reliable differences between impaired and normally developing readers on measures of phonological skills such as those just mentioned. However, training and intervention studies have established causal relationships only in the case of phoneme analysis and letter-sound decoding. Some researchers have also suggested that deficiencies in vocabulary knowledge and syntactic competence may contribute significantly to difficulties in learning to read, but a causal relationship has yet to be established.

Specific reading disability has also been attributed to deficiencies in various cognitive abilities that are involved in all learning, not just learning to read: in particular, selective attention, associative learning, intersensory integration, serial-order processing, pattern analysis, and rule learning. Dysfunction in one or another of these basic learning abilities would seem to be logically ruled out as a significant cause of the disorder in a child who has at least average intelligence and who does not have general learning difficulties, given that all of these cognitive abilities are involved in virtually all tests of intelligence, and are most certainly involved in all academic learning. More importantly, however, each of these hypotheses has been discredited as a cause of specific reading disability by empirical research. (*See* **Dyslexia**)

Finally, estimates of the incidence of specific reading disability in school-age children have ranged between 10% and 20%, but results from intervention studies suggest that 1–5% is a more realistic estimate. These studies have shown that early reading difficulties in most children are caused by experiential and instructional deficits rather than basic constitutional deficits. For example, several of these studies have shown that many children who qualified for 'reading disabled' status in first grade, entered kindergarten lacking in foundational literacy skills such as knowledge of print concepts, knowledge of the alphabet, and phoneme awareness, and had been exposed to kindergarten language arts programs that did not facilitate acquisition of these skills. In contrast, children who had been exposed to balanced and comprehensive language arts programs in kindergarten, emphasizing both code-based and meaning-based literacy activities, experienced fewer difficulties in early literacy development. Thus, the key to reducing the incidence of early reading difficulties would seem to be to improve the quality of the language arts programs in the early grades.

CONCLUSION

Early literacy development can be usefully divided into three periods of acquisition: a prealphabetic period, an alphabetic period, and an advanced

alphabetic/orthographic period. During the prealphabetic period, children begin to become sensitized to the similarities and differences in words in spoken language; begin to learn about print concepts and conventions; begin to acquire rudimentary literacy skills such as knowledge of the alphabet and phoneme awareness, and begin to become acquainted with the structural features of written text. During the alphabetic period children expand their knowledge of print concepts and conventions as well as their sight-word vocabulary and text processing skills; most importantly, they acquire phoneme awareness, learn how the alphabet works, and develop proficiency in using letter sounds to aid word identification, spelling, writing, and text reading. They also become increasingly sensitive to frequently occurring and redundant spelling patterns within words, which facilitates initial use of analogy strategies among other generative strategies for word decoding and spelling.

During the advanced alphabetic/orthographic period children consolidate their alphabetic knowledge, become increasingly sensitive to larger letter clusters that have reasonably stable pronunciations across many words ('ight' in 'sight', 'night' and 'right'), increasingly store information about word-general properties of printed words, and learn to make functional use of such knowledge in decoding and spelling multisyllabic words. With more experience with print, children come to store high-quality representations of the unique spellings of words and build a large inventory of sight words that can be identified or written with speed and accuracy.

Difficulties in acquiring early literacy skills have been attributed to deficiencies in reading-related cognitive abilities such as visual perceptual and visual-spatial abilities, language and language-based abilities, and general learning abilities entailed in all learning (e.g. selective attention, associative learning, intersensory integration). However, research to date suggests that visual deficits and general learning deficits are not causally related to impaired literacy acquisition in otherwise normal children, whereas language-based deficits, especially phonological deficits, are so related. Intervention studies provide strong evidence that difficulties in acquiring early literacy skills are, in most cases, caused by experiential and instructional deficits rather than by basic cognitive deficits.

Further Reading

Adams MJ (1990) *Beginning to Read: Thinking and Learning About Print*. Cambridge, MA: MIT Press.

Bradley L and Bryant PE (1983) Categorizing sounds and learning to read: a causal connection. *Nature* **303**: 419–421.

Ehri LC (1991) Development of the ability to read words. In: Barr R, Kamil M, Mosenthal P and Pearson PD (eds) *Handbook of Reading Research*, vol. II, pp. 384–417. New York, NY: Longman.

Frith U (1985) Beneath the surface of developmental dyslexia. In: Patterson K, Marshall J and Coltheart M (eds) *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*, pp. 301–330. London, UK: Lawrence Erlbaum.

Henderson EH (1990) *Teaching Spelling*. Boston, MA: Houghton Mifflin.

Neuman SB and Dickinson DK (eds) (2001) *Handbook of Early Literacy Research*. Mahwah, NJ: Lawrence Erlbaum.

Perfetti CA (1985) *Reading Ability*. New York, NY: Oxford University Press.

Read C (1971) Preschool children's knowledge of English phonology. *Harvard Educational Review* **41**: 1–34.

Snow CE, Burns MS and Griffith P (1998) *Preventing Reading Difficulty in Young Children*. Washington, DC: National Academy Press.

Tunmer WE, Herriman ML and Nesdale AR (1988) Metalinguistic abilities and beginning reading. *Reading Research Quarterly* **23**: 134–158.

Vellutino FR (1987) Dyslexia. *Scientific American*, March: 34–41.

Vellutino FR and Scanlon DM (2001) Emergent literacy skills, early instruction, and individual differences as determinants of difficulties in learning to read: the case for early intervention. In: Neuman SB and Dickinson DK (eds) *Handbook of Early Literacy Research*, pp. 295–321. New York, NY: Guilford Press.

Vellutino FR, Scanlon DM, Sipay ER *et al*. (1996) Cognitive profiles of difficult to remediate and readily remediated poor readers: early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology* **88**: 601–638.

Vellutino FR, Scanlon DM and Tanzman MS (1994) Components of reading ability: issues and problems in operationalizing word identification, phonological coding, and orthographic coding. In: Lyon GR (ed.) *Frames of Reference for The Assessment of Learning Disabilities: New Views on Measurement Issues*, pp. 279–324. Baltimore, MD: Paul H. Brookes.

Whitehurst GR and Lonigan CJ (2001) Emergent literacy: development from prereaders to readers. In: Neuman SB and Dickinson DK (eds) *Handbook of Early Literacy Research*, pp. 11–29. New York, NY: Guilford Press.

Literacy: Reading (Later Stages)

Introductory article

Michael Pressley, University of Notre Dame, Notre Dame, Indiana, USA

CONTENTS

The later stages of reading are largely concerned with comprehension or the construction of meaning, which requires both conscious and unconscious cognitive activity.

INTRODUCTION

A reader in the later stages of development can comprehend many types of texts. Hence, the topic of reading comprehension has received substantial attention from researchers interested in advanced reading skills. As a result, much is known about how good readers construct meaning as they read; much of this article is concerned with the meaning construction that is skilled comprehension.

Although comprehension begins with processing individual words, it is much more than reading and understanding individual words. It is comprehending those words in relation to one another within sentences, then combining across sentences and paragraphs to construct higher-order meanings. Comprehension is a decidedly cognitive psychological set of activities.

After presenting a descriptive summary of skilled comprehension, there will be coverage of how comprehension can be developed in readers. It may seem backwards to consider development after presenting a conception of the developmental end product, but the components considered important to develop in young readers will make sense once skilled comprehension is understood. That is, the reading experiences and instruction provided to children should be done with an envisionment of the endpoint in mind, with the goal of comprehension development being adult readers who are skilled as described in the next section.

SKILLED COMPREHENSION

First, it is important to emphasize that comprehension is a constructive process. It was once believed that texts contained meaning, and that meaning was constructed by authors as they wrote, rather than by readers as they read. That is, when a writer wrote a text, she or he had a particular meaning in mind, and if the writer was successful, that meaning ended up in the text. According to this perspective, every good reader would find that particular meaning in the text. It is this conception of comprehension that drives many student publications that claim to provide the meaning of challenging texts to students (e.g. student notes for Shakespeare). What became increasingly apparent in the second half of the twentieth century to literary scholars and psychologists interested in comprehension was that the conception of texts having single meanings was not correct.

Readers construct their own understandings of text – albeit, processing the author's words, sentences, and paragraphs, but reacting to and reflecting on the author's input and in doing so, coming to unique understandings of the text. One proof of this comes when several intelligent people come together to discuss the same text. Often two highly intelligent readers can come to very different interpretations of the story, both of which make sense given the information that is in the text. What this situation reflects is that comprehension involves going beyond the information given in the text, with the good reader making inferences during reading. The good reader is very active during reading. Readers are said to respond to text, with responses varying from reader to reader. That is, although the interpretations of two readers can both include many of the same specific ideas and facts stated in a text, the two readers may emerge from the reading with very different conclusions about what was important in the text and how the ideas in this text relate to the world of ideas in general.

During the twentieth century, great progress was made in coming to understand how readers

construct meaning as they read. One research tool, in particular, proved to be very revealing about reading comprehension processes. Readers in a number of studies were asked to think aloud as they read, to report their thinking as they worked through a text. The resulting 'think-alouds' provided a great deal of information about what skilled readers do before they read a text, as they read it, and after reading. There is a complex articulation of a variety of processes as part of comprehending a text.

Before reading, the good reader has a goal about what she or he wants to get out of a text. That is, the good reader knows whether the text is being read for detailed or general understanding. The goal of reading is important, for it will influence greatly just how the reader proceeds to understand the text (e.g. processing will be more complete if the reader is attempting to understand the author's message in detail rather than simply wanting to be entertained by the story). Especially when reading expository texts (i.e. factual texts, in contrast to stories), good readers will often skim a text before reading it in detail. Skimming provides a great deal of information that permits planning of the upcoming reading, including information about the length and structure of the text and sections that might be particularly important or unimportant. As a result of skimming, the reader might decide to read only certain parts of the text carefully. Alternatively, the reader may decide not to read the entire text. If the skim reveals the text to be irrelevant to the reader's goal (e.g. the reader wants to know about the history of track and field sports and this article about the Olympics is only about the winter games), the reader may decide not to read the text. The initial skim can also result in the activation of prior knowledge (e.g. as the reader skims an article about track and field sports, the reader relates ideas encountered in the skim to her or his prior knowledge). Activation of prior knowledge is important, for a reader's construction of meaning very much depends on prior knowledge. Thus, based on an initial skim and reflection on the ideas encountered during the skim, the reader makes predictions about what ideas will be encountered in the text based on what she or he knows about the topic already. As reading proceeds, good readers are alert to whether their predictions and expectations about what might be in the text are confirmed or disconfirmed.

In general, once reading of a text begins, good readers go through that text from front to back, although sometimes they read selectively (i.e. skipping over sections or reading some sections more carefully than others). Good readers are aware of the characteristics of a text (e.g. whether it is difficult or easy to understand) and whether they are understanding what is read (i.e. they monitor their comprehension). If a text is challenging or there are particular points that the reader wants to remember, she or he may reread a relevant section of text or restate the section of text in her or his own words, make notes, or pause to reflect on the text. As reading proceeds, the reader is also alert as to whether expectations are being met, for example, whether predictions made during skimming based on prior knowledge are correct. In addition, as reading proceeds, the reader makes new predictions based on ideas found during reading of the text. This is, expectations and predictions continuously shift during reading, with the hypotheses that arise during reading evaluated as the reader continues going through the text.

Good readers pay attention to essential points in a text more than less essential points. Hence, when reading stories, good readers pay attention to scene-setting information, which is almost always near the beginning of the story. They also pay attention to introductions of main characters, which also occur early in the story. The skilled reader of stories expects that the main characters will encounter one or more problems in the story, with attempts to solve these problems following. One of the attempts will succeed, with this signaling that the story is starting to wind down. When reading an expository text, the reader is attentive to information that is particularly relevant to reading goals (e.g. looking specifically for information about track and field and ignoring information about other sports contained in a text). When processing expository text, the good reader looks for topic paragraphs, topic sentences, and key words. They are particularly attentive to parts of text that they want to remember. Sometimes the good reader will do more than simply pay more psychological attention to important points in a text. Good readers sometimes highlight text, underlining important points or making notes in the margin. Consistent with different readers making different interpretations as they read, readers differ in the portions of text they consider important.

Sometimes a reader's understanding changes dramatically while reading a text. Hence, perhaps in reading a text on track and field, the reader comes to the insight that her or his interest should only be running events. Hence, the reader might change the reading goal, now looking for information only pertaining to track. Perhaps the reader began the text with the belief that the very best

athletes were sprinters. If the author of the text makes a strong case that distance runners are in better condition and have greater natural skills than sprinters, the reader may change her or his mind, leaving the text with a different conception of the relative ability of sprinters and other athletes.

Inferences

Throughout the reading process, the reader participates in the meaning-making process, largely through making inferences. Sometimes these inferences are unconscious, and if they are, there may be little reporting of them as the individual thinks aloud. (Fortunately, there are other procedures for tapping such inferences.) Many inferences, however, are decidedly conscious, with readers readily reporting that they make them. For example, a reader might encounter the following bit of text: 'The ball went over Jack's head into the street. The car swerved to miss Jack.' Many readers would consciously infer that Jack must have run into the street.

What kinds of inferences do readers make as they read? The answer is, many: for example, they infer the referents of pronouns when referents are vague. Thus, if given the sentence 'Laura Bush then introduced the President. He came to the podium', most readers would infer that the 'he' is President Bush. Also, when new vocabulary words are encountered, readers often infer their meaning. Thus, a third-grade student reading the sentence 'There were blue, red, and chartreuse dresses', might infer that chartreuse is a color since the colors blue and red preceded it in the listing within the sentence. Readers also make inferences about connotations made in a text. Thus, if encountering 'Many members of this society believe there are visitors to Earth from other planets. That's pretty likely, isn't it?', many readers would infer from the sentence 'That's pretty likely, isn't it?' that the author did not think it really was likely that Earth is being visited by extraterrestrials.

Much inferencing involves relating ideas encountered in the text to the reader's prior knowledge. Thus, readers sometimes construct explanations for events reported in a text. For example, if a headline states 'New Yorkers Scoop Up World Series Tickets', the avid baseball fan who is just returning to the country from abroad might infer that the New York Yankees have won the American League playoff series. If reading about how certain cars do not live up to their advertised mileage, the Honda owner might infer that the author could not be talking about Hondas, since his Honda gets better fuel consumption than advertised. Alternatively, the reader might relate the author's point to a more general piece of knowledge, thinking to himself: 'The auto makers are engaging in false advertising, and a bad form of it, since fuel-inefficient cars undermine progress on the fossil fuel problem, which is huge.'

One common type of inference is about the author or the author's intention in writing the text being read. Thus, when reading a psychology article about recent important experiments, a reader might infer: 'The author really wants me to believe that psychology is a scientific enterprise.' Alternatively, another reader of the same text might infer: 'This article was written for people with a lot of background in psychology.' Sometimes the reader makes inferences about the author herself, perhaps concluding: 'She must be a psychologist. This could not have been written by a journalist.' In short, good readers have a great deal of prior knowledge based on their experiences in the world and previous reading, and they consistently relate what they know already to ideas encountered in the text, coming to conclusions not explicitly stated in the text.

Integration

Good readers relate across the ideas in a text, attempting to construct the larger messages in that text. Thus, when reading Charles Dickens's *A Christmas Carol*, the reader might reflect on the various scenes to construct the understanding that what happens in later life depends very much on what happens earlier in life, but it is always possible for people to change. The skilled reader of stories will explicitly make certain that she or he knew the setting, the characters, the problems encountered by the characters, and the problem resolution as part of coming to general conclusions about the story. For example, the reader might conclude that '*Christmas Carol* is much like Cinderella in that a life is transformed through the intervention of characters from another world'.

In reading an expository text, the skilled reader will be alert to the introduction of the article and to examples to illustrate general principles. The reader will also be careful to pay attention to logical structures in the text, such as cause-and-effect sequences and compare-and-contrast arguments. Such attention is often in the service of coming to higher-order conclusions, such as: 'This article is about ecological threats to sea-life, with the case being made that many species – from whales to

sharks – are being threatened by man's increasing use of the seas. Although the threats are different for whales and sharks, they are all threats due to humans invading the oceans with more sophisticated technologies.' The construction of such integrative understandings may require a lot of effort, including reviewing points made in the text, rereading sections of text to make certain the ideas were comprehended, and making written notes or outlines as reading proceeds.

Interpretations

The good reader is an interpretive reader, with interpretations obvious in a number of ways. First, good readers do summarize as just described. They also construct mental images reflecting story elements or ideas presented in a text. Thus, as a child reads L. M. Montgomery's *Anne of Green Gables*, she might develop in her mind's eye a detailed envisionment of Prince Edward Island. If a story has a scene in a McDonald's restaurant, the reader might envision that the story took place in the local McDonald's, which is a hangout for teenagers, and thus interpret that the story is about teenage culture. Sometimes the same reader can be multiply interpretive, perhaps trying to imagine alternative interpretations about a text. Thus, if reading a story about the 2000 US presidential election recount in Florida, the reader might reflect on how winning candidate George Bush would interpret the story (e.g. 'It's about public officials making hard but necessary decisions'). In contrast, losing candidate Al Gore might interpret the story differently (e.g. 'It's about how Republicans can twist the facts to fit their view of the world and steal an election in doing so').

Evaluations

Good readers are evaluative as they read. Hence, if ideas presented in a text clash with a reader's prior knowledge, she or he may devalue or dismiss a text (e.g. a patriotic American dismissing an article claiming that the former Soviet Union was the most advanced society in the mid-twentieth century). Alternatively, if the ideas in a text are consistent with a reader's perspective, she or he might be enthusiastic about the book. In addition to being aware of whether arguments in a text are valid or credible from the reader's perspective, good readers are critically aware as they read whether a text is well written (appropriate vocabulary, well-structured sentences and paragraphs, well-developed arguments) or interesting or novel. A reader's evaluations can often be quite pronounced and affective, including surprise, laughter, puzzlement, boredom, frustration, or anxiety (e.g. 'I hope this author isn't concluding that people like me might be at risk for cancer').

Post-reading Cognition

After a good reader has gone through a text, she or he often continues to think about it. Thus, sometimes readers will skim over a text just read as part of constructing the overall meaning of the text. They might be much more thorough, however, rereading parts of the text that seem especially pertinent or perhaps outlining the text. Sometimes they might imagine how the ideas in the text might be used later by them (e.g. thinking about how the ideas about threatened sea-life might make sense in a paper they are writing for a biology course).

Monitoring

Good readers are consciously aware as they read, with such awareness critical in directing their efforts to construct meaning. Such awareness is generally known as 'monitoring', with one specific example of comprehension monitoring provided earlier. More generally, good readers monitor a number of characteristics of a text as they read – whether it is relevant to the current reading goal, difficult or easy to read, written in a biased fashion, related to other texts the reader has read, ambiguous, or presenting ideas consistent or inconsistent with the reader's perspective.

Good readers also monitor their strategies while reading (i.e. whether they are skimming the text or reading it carefully, looking for particular types of information, making predictions and later confirming or disconfirming them, summarizing) and whether the strategies they are using are resulting in the level of comprehension they desire (e.g. whether they are achieving the detailed understanding they want). If the strategies currently being used are not working to produce the understanding the reader desires, the reader may change tactics. That is, as a reader monitors, she or he is especially sensitive to problems being experienced, with the detection of problems leading to actions to solve the problems. So, if the problem the reader detects is that she or he is spending a lot of time reading parts of text that are irrelevant to the reading goal, the reader might decide on a new strategy. The reader might begin to skim each paragraph before reading it, skipping over paragraphs that

do not appear to have much relevant information in them. Alternatively, if the reader monitors that she or he is not remembering some critical details, the decision may be to read much more carefully. If the reader monitors that some parts of a story are much more interesting than others, she or he might elect to read some sections of the story carefully, stopping to reflect on the scenes as developed by the author, and to skim other sections of the story. Thus, a reader who is reading *The Brothers Karamazov* to understand Dostoyevski's orientation to religion might read some sections of the book much more carefully than others. If a reader monitors that an unknown word is really critical to understanding a story, she or he might pay a great deal more attention to the word and the construction of its meaning from context clues than if the word seems less essential to understanding meaning. Thus, a modern child reader of Lewis Carroll's *Through the Looking Glass* might make certain that she or he knew the meaning of 'looking glass', recognizing that if the word is important enough to be in the title, it is important to make absolutely certain that the word is understood. In short, the awareness that is monitoring is a critical determinant of what readers do as they read.

The skilled reading described in this section is observed in some adults but rarely in children, or even high school students. More positively, even very young children can be taught to monitor and use comprehension strategies as they read. Most of the remainder of this article is dedicated to the development through instruction of reading skills that contribute to skilled comprehension.

DEVELOPMENT OF SKILLED COMPREHENSION

Many adult readers do not read as just described. Rather, they pick up a story or an article and begin at the beginning, reading every word, but doing little else. Often, at the end of the reading, little is remembered from the text. Such a reader's understanding of the ideas in the text is not nearly as complete as the understanding of a really skilled reader. One important reason that many people do not become skilled readers is that they are never taught how to read well. Many adults with reading problems can become better readers through the instruction considered in this section. Nonetheless, such instruction should not be put off until adulthood, for the evidence is quite strong that reading comprehension improves reliably when young readers are taught to carry out the processes that are used by good readers. In fact, the evidence is very strong that even struggling child readers (e.g. students with learning disabilities, specifically reading disabilities) benefit from the types of instruction considered in this section.

Word Recognition

One salient characteristic of good readers is that they have no trouble reading individual words. In fact, they are said to recognize words fluently. When word recognition is fluent, it requires hardly any cognitive effort for the reader, which is critical, since cognitive effort consumes short-term memory – short-term cognitive capacity. That is, human beings can do only so many things at once; they have only limited short-term memory capacity. Both word recognition and comprehension processes require cognitive capacity, and hence, compete with one another for the limited capacity that is available. When readers have to exert a lot of effort to recognize words, there is little capacity left over to understand the text being read. Conversely, when word recognition is effortless, there is more capacity left over for comprehension. Thus, it is very important that word recognition skills be developed in readers so that most words can be read fluently – that is, can be read without the reader thinking very hard about them.

In recent years, the research-based consensus is that fluent word recognition is most likely to develop if the beginning reader receives explicit instruction in how to sound out words – that is, explicit instruction in phonics. Although learning how to sound out words does not automatically lead to fluent word recognition, it seems to be a good start. With subsequent practice in reading words, fluency eventually comes for many students. Of course, even the most skilled reader sometimes has to exert effort to recognize a word. For example, when adults who are not physicians read a medical text, there are unfamiliar words that are read only by consciously sounding them out, with their meaning understood only by the reader consciously looking for root words and context clues. Thus, consider the sentence 'Mucous membranes sometimes develop adenocarcinomas'. The skilled nonphysician reader would have to sound out 'adenocarcinoma'. The reader might guess that it is a form of cancer because it contains the word 'carcinoma', which means cancer. In contrast, a cancer doctor would read the same word effortlessly and know what it meant automatically. Good readers eventually can read words they encounter often without effort. Early instruction in

phonics followed by lots of practice reading words can result in fluent reading of the words included in the types of text frequently read by the reader.

Much emphasis has been placed on the development of word recognition skills, almost to the neglect of other skills. Some of those who emphasize teaching of phonics seem to feel that the key skill in reading is word recognition, and that if word recognition is skilled, the young reader will eventually develop into a highly skilled reader. There is no scientific evidence to support such a position. Rather, the evidence is quite consistent that skilled reading more certainly develops from the acquisition of a variety of skills that comprise comprehension, one of which is word recognition. That word recognition skills are critical to skilled comprehension, however, makes clear that the development of readers with advanced reading skills depends on reading education efforts during early schooling. Other evidence also points to this conclusion.

Vocabulary

Skilled readers typically have good vocabulary, which is sensible, given that higher-order meanings in text cannot be constructed without knowing the meanings of individual words. Most vocabulary items are learned by individuals encountering them in context. Thus, young children learn the meanings of many words by hearing them in their worlds. They learn other meanings by encountering words in texts and making guesses about what a word means based on context clues. Even so, teaching students vocabulary that they do not know improves reading comprehension, especially of texts containing the words taught. The conclusion that follows from this set of observations is that encouraging vocabulary development is important as part of developing skilled readers.

There are multiple ways of encouraging vocabulary development. One is to make certain that the child experiences a rich language world, one in which the child interacts verbally with adults and other children. A second is to ensure that the child reads broadly, especially books that contain new vocabulary. A third is to teach vocabulary explicitly. Researchers have produced a variety of data confirming that the development of vocabulary competence has a striking effect on comprehension skills. Again, it is possible even during the early elementary school years to be educating children in ways that are consistent with the skills reflected in really skilled reading.

World Knowledge and Using World Knowledge

In addition to good vocabularies, good readers tend to have substantial and well-organized knowledge about the world. Again, this reflects rich interactions with the world, including a great deal of reading of high quality materials. This extensive world knowledge permits readers to go beyond the information stated in a text, to make appropriate inferences.

That said, a failure documented in recent years is that some readers possess world knowledge that could be used to understand ideas in a text (e.g. knowledge about football that could be used to understand ideas in an article about changes in the rules of football for the upcoming season), but they fail to use it as they read. Thus, there are now efforts to develop instruction that will encourage readers to use the knowledge they already have as they read texts. One powerful approach involves prompting readers to ask themselves why the ideas being presented in a text make sense. Thus, the person reading the article about changes in football rules would be encouraged to think about why the rule changes make sense, based on his knowledge of football.

Another problem involving world knowledge is that some readers will bring to a text world knowledge that seems to be irrelevant to the ideas in the text (e.g. a reader processing the article about football rules recalls how an umpire in a baseball game made a bad call that cost his team the game). Such irrelevant world knowledge can undermine comprehension of text, and hence, researchers are now working to identify ways to reduce the likelihood of readers making such distracting associations.

In summary, eventual skilled comprehension can be fostered by encouraging students to read materials filled with the information that literate people know. Skilled comprehension can be developed further by teaching students to relate the prior knowledge they have acquired through experience (including previous reading) to ideas encountered in texts they are now reading.

Comprehension Strategies

Relating ideas in a text is just one of the comprehension strategies used by good readers. Good readers also predict, ask questions while they read, construct mental images depicting ideas expressed in the text, seek clarification when confused as they read, and summarize the main ideas in the text. Use of such strategies does not develop reliably unless

they are taught. Unfortunately, these comprehension strategies often are not taught in school. More positively, instruction in these strategies can be profitable even during the early elementary grades.

The approach to comprehension strategies instruction that has the most support begins with teacher modeling and explanation of strategies. The strategies can be introduced one at a time, perhaps over a semester. After showing and explaining the strategies to students, the teacher encourages the students to use them when they read. Although the emphasis at first can be on single strategies, the teacher eventually emphasizes coordinated use of the strategies, with the students practicing prediction, questioning, clarification, constructing mental images, and summarizing as they read stories and articles, both in small groups and individually.

The explicit teaching of a repertoire of comprehension strategies through teacher modeling, teacher explanation, and teacher-supported practice is known as transactional strategies instruction. Although a year of *transactional strategies instruction* produces documented effects, several years of such instruction and practice is required for readers to use comprehension strategies automatically and well. In general, the thinking is that comprehension strategies instruction should begin in the early elementary years, with researchers believing that consistent teaching of such strategies is much more likely than traditional reading instruction to produce readers who are active in the ways skilled readers are active, as described in the first section of this article.

Monitoring

Such strategies instruction invariably includes teaching of monitoring as well. Readers are taught to be alert to when they are experiencing difficulties in reading and to respond to difficulties by applying new strategies – for example, looking for context clues, if the meaning of a word is unclear, or rereading, if the meaning of a sentence or paragraph is difficult to grasp. In short, student readers are being taught to be self-regulated readers, with self-regulation involving awareness about when the use of reading strategies is required, followed by the selection and use of appropriate strategies.

Cognitive Capacity Constraints

A reaction to the teaching recommendations in this section could be that young readers are being asked to do too much. In particular, sounding out words, figuring out the meaning of new vocabulary words, relating ideas in a text to world knowledge, using comprehension strategies, and monitoring all involve effortful processing, at least at first. Does the young child have sufficient cognitive capacity to carry out all of these processes? The answer is that the child does not have the capacity to do so consciously when all – or even some – of these processes can be accomplished only with effort. The key is to practice the various processes until they are carried out fluently. That is why word recognition and comprehension strategies instruction, in particular, are thought of as interventions that occur over years rather than over weeks or months. With years of practice, word recognition becomes fluent as does use of comprehension strategies. With increased fluency, there is decreased demand on the child's limited cognitive capacity. If the child is reading excellent material to practice word recognition and comprehension strategies, there are bonuses, with vocabulary and world knowledge expanding. Of course, the more vocabulary the reader knows and the more world knowledge she or he has acquired, the better comprehension is in the future.

MOTIVATION TO READ

Really skilled readers not only can read, but they do read. As they do so, their knowledge increases as do their reading skills, with the result that the richly skilled and knowledgeable get even richer in skills and knowledge. Because extensive reading is so essential to highly skilled reading, much attention has been given in recent years to what can be done to increase student motivation to read.

Schools regularly provide concrete incentives to students for reading extensively (e.g. a pizza gift certificate for reading so many books a month). Permitting students to read in topic areas that are interesting to them motivates reading. What undermines motivation to read, however, is difficulty in reading. Sadly, many students who do experience initial reading difficulties come to hate reading, with this passionate dislike of reading substantially reducing the likelihood that reading will be done voluntarily. Not reading in elementary and secondary school does not augur well with respect to future reading skills. Hence, one of the primary reasons that so much attention is now being given to increasing reading success through early instruction is to prevent students from experiencing the frustrations that undermine long-term commitment to reading. A goal of most reading educators is to develop life-long readers in their students,

since life-long practice of reading is absolutely required to become an effective reader.

BEYOND COMPREHENSION

Before a reading can be comprehended, the reader has to find the material that should be read. Searching for texts that contain particular information and searching for information within a text is a complicated skill. The latter task is understood more completely by psychologists than the former task.

To find material in a text, the reader has to know what she or he is looking for (e.g. information about Olympic track events). Then, the reader has to figure out which parts of a text might have that information (e.g. a chapter on the Olympics, a chapter on track and field). Once the relevant sections to search are identified, the reader has to go through the text sections and recognize the information that is crucial. None of this is easy.

For example, even if a college student knows what she or he wants to find and identifies some particular parts of source texts that might contain the information, there is no guarantee the reader will recognize pertinent information when it is being read. Why is this so? Because a great deal of prior knowledge about a topic is often necessary to be able to pick out sections of text that are directly relevant to a particular topic in the knowledge domain. Without extensive prior knowledge, the reader can read right over information that is exactly the information being sought without recognizing it as such. One of the things that skilled readers can do is to search text effectively, which depends very much on having extensive prior knowledge of the domain being searched.

Beyond finding information in a text, readers often want to use information read in a text for some purpose. A typical student purpose is to write a paper, which is analogous to research writing in general. Unfortunately, we know little about how people carry out research by reading multiple texts in anticipation of writing an article or a book. What we do know is that there is variability in this skill, with some people reading articles and making notes nonreflectively. Others are a bit more reflective, taking note only of ideas or facts in a text that seem true or important. It is a clear minority of college students who read several texts, compare the ideas in them, and attempt to understand the similarities and differences between authors of different articles or books. As far as using what is found by reading several sources, there are also individual differences among college students. Many write papers simply by listing out facts found in readings. It is a minority again that tries to write essays that go beyond the information given in source texts. People high in literacy skills are very reflective as they read multiple texts on a topic, doing much comparison across ideas in the texts, attempting to construct understandings that go beyond the understandings expressed in any one of the texts that have been read.

CONCLUSION

Advanced readers are active readers, constructing understandings of texts, making inferences as they do so, and integrating across ideas in a text to produce summary understandings. They can do so fluently because they have the many component skills of comprehension well honed. They recognize words easily, know the meanings of many words encountered in texts, have expansive world knowledge which they can relate appropriately to ideas encountered in new texts, and know and use comprehension strategies from before reading to during and after reading. All such competencies can be developed through instruction so as to promote comprehension skills. A complete education to develop the advanced reader requires years of instruction in the components of skilled reading as well as much practice doing actual reading. Excellent readers are motivated to read, can find information in texts that is relevant to topics they want to know about, and can combine information across texts to construct new understandings that go beyond the information given in any of the source texts read. When such reading is observed, it is among well-educated adults, individuals with years of reading experience.

Further Reading

Anderson RC and Pearson PD (1984) A schema-theoretic view of basic processes in reading. In: Pearson PD (ed.) *Handbook of Reading Research*. New York, NY: Longman.

Baker L and Brown AL (1984) Metacognitive skills and reading. In: Pearson PD, Barr R, Kamil M and Mosenthal P (eds) *Handbook of Reading Research*, pp. 353–394. New York, NY: Longman.

Block CC and Pressley M (eds) (in press) *Comprehension Instruction*. New York, NY: Guilford.

Flower L, Stein V, Ackerman J *et al*. (1990) *Reading to Write: Exploring a Cognitive and Social Process*. New York, NY: Oxford University Press.

Guthrie JT (1988) Locating information in documents: examination of a cognitive model. *Reading Research Quarterly* **23**: 178–199.

National Reading Panel (2000) *Final Report of the National Reading Panel*. Washington, DC: National Institute of Child Health and Development.

Pearson PD and Fielding L (1991) Comprehension instruction. In: Barr R, Kamil ML, Mosenthal PB and Pearson PD (eds) *Handbook of Reading Research*, vol. 2, pp. 815–860. New York, NY: Longman.

Pressley M (2000) What should comprehension instruction be the instruction of? In: Kamil ML, Mosenthal PB, Pearson PD and Barr R (eds) *Handbook of Reading Research*, vol. 3, pp. 545–561. Mahwah, NJ: Lawrence Erlbaum.

Pressley M and Afflerbach P (1995) *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. Hillsdale, NJ: Lawrence Erlbaum.

Pressley M, El-Dinary PB, Gaskins I *et al*. (1992) Beyond direct explanation: transactional instruction of reading comprehension strategies. *Elementary School Journal* **92**: 511–554.

Literacy: Writing

Advanced article

Steve Graham, University of Maryland, College Park, Maryland, USA
Karen R Harris, University of Maryland, College Park, Maryland, USA

CONTENTS

Writing is a self-initiated, self-directed, and self-sustaining activity of composition and inscription that requires the orchestration of a wide array of cognitive processes. Since the mid-1970s the cognitive science framework has dominated the psychological study of these mental operations.

INTRODUCTION

More than 5000 years ago, the Sumerians devised the first known system of writing, called cuneiform, to record goods using a wedge-shaped reed stylus to make impressions on a moist clay tablet. From this modest beginning, the application of writing as well as the means for producing it have undergone an incredible metamorphosis (Graham and Harris, 2000a). The act of producing writing, for instance, has progressed to the point where scientists at Germany's University of Tübingen have developed a machine, called the 'thought translation device', that allows a writer to amplify or dampen brain wave patterns to select letters from a video screen to spell out a message (Begley, 1999). Similarly, writing has evolved from its initial record-keeping debut to being one of humankind's most powerful tools. Among its many functions, it provides a flexible medium for artistic, political, spiritual, and self-expression. Just as important, writing provides an external memory, making it easier to remember, analyze, and share ideas. It can even have therapeutic effects; writing about one's feelings can lower blood pressure, reduce depression, and boost the immune system (Graham and Harris, 2000a).

In contrast, the scientific study of writing has a more recent genesis, totaling no more than 100 years. Since the 1980s, two basic approaches have dominated the scholarly and scientific study of writing (Scardamalia and Bereiter, 1986). One approach has been labeled as 'contextual', and focuses on the meaning that writing has for people in different contexts. The other is referred to as the 'cognitive science' approach. This approach has been particularly powerful in advancing our knowledge about the process of writing and its development. As two influential researchers noted: 'It provides pretty much the "only paradigm in town" for investigating complex mental processes, which all sides agree are of central concern in writing' (Scardamalia and Bereiter, 1986, p. 780).

COGNITIVE MODELS OF WRITING

Although some historians claim that they can identify the exact year (1956) or even date (11 September 1956) that contemporary cognitive psychology was born, such precision is not possible in the study of writing. A watershed event in this area, however, was an interdisciplinary conference at Carnegie Mellon University in 1978 designed to synthesize previous research on writing as well as showcase new research. Papers from this conference were assembled into a book entitled *Cognitive Processes in Writing* (Gregg and Steinberg, 1980). The publication of this book generated considerable interest in the cognitive nature of writing and the application of conceptual and methodological advances from the cognitive sciences.

A particularly influential chapter in this book presented a model of skilled writing (Hayes and Flower, 1980). This model was developed by asking adults to 'think aloud' while composing. Analyses of these verbal protocols provided a window into the cognitive as well as other psychological processes involved in writing. The resulting model comprised three major components.

The first component, *task environment*, included factors that were external to the writer, but influenced the writing task. These included both social factors, such as an assigned writing task, as well as physical ones, such as the text produced so far.

The second component, *cognitive processes*, provided a description of the mental operations involved in writing. This included three basic processes: planning what to say and how to say it, translating plans into written text, and reviewing to improve existing text. Planning, in turn, was composed of three ingredients – setting goals, generating ideas, and organizing ideas into a writing plan, whereas reviewing included reading and editing text. The execution of these cognitive processes was thought to be under the writer's direct control, and it was proposed that virtually any subprocess could interrupt or incorporate any other subprocess. For instance, planning might interrupt translation, if a writer identified the need to develop additional writing goals while producing a first draft. Another writer might combine translation and reviewing, generating a section and then revising it, then generating and revising a second section, and so on.

The third model component, *writer's long-term memory*, included the author's knowledge about the topic, the intended audience, and general plans or formulas for accomplishing various writing tasks.

The Hayes and Flower (1980) model emphasized the recursive nature of the cognitive processes involved in writing. In other words, writing did not necessarily proceed in a linear fashion from planning to translating to revising, as was commonly suggested in the composition textbooks of the time. Rather, skilled writing was viewed as a nonlinear activity, where major cognitive processes and subprocesses were typically interwoven or nested one within another. This property of recursion was a particularly powerful component of the model, because a small number of cognitive processes were now able to account for a diverse set of mental operations during composing. In addition, individual differences among writers could be explained by different specifications of the model. For instance, Scardamalia and Bereiter (1986) developed a description of immature writing, referred to as 'knowledge telling', that was essentially a radically reduced version of the model. Based on their study of school-age children, they proposed that novice or immature writers use a greatly simplified version of the idea-generation process in Hayes and Flower's model. Children converted the writing task into simply telling what one knows about the topic.

The Hayes and Flower (1980) model was also significant because skilled writing was cast as a self-directed activity, involving high levels of self-regulation. Analysis of the verbal protocols from the 'think-alouds' revealed that writing is a goal-directed process. Skilled writers set goals for what they wanted to do and say (e.g. be convincing, funny, and succinct), and frequently established subgoals for meeting their major goals (e.g. use strong arguments and refute counterarguments to convince the reader). Hayes and Flower's analysis further highlighted the cognitive complexity of writing. Skilled writers caught in the act looked very much like busy switchboard operators, trying to juggle simultaneously a number of demands on their attention (e.g. making plans, drawing ideas from memory, developing concepts, creating an image of the reader, testing ideas and text against that image, and so forth).

In 1996, Hayes revised the original model, reorganizing, expanding, and modifying the initial framework so that it captured the ensuing 16 years of writing research. In the new model, the task environment was modified so that it included both a social component (e.g. the audience, other texts read while writing, and collaborators) as well as a physical component (e.g. text read so far and the writing medium, such as a word processor). Hayes's overhaul was not limited to contextual factors, however, as he also reconceptualized internal factors involved in writing. First, he included motivation as a separate component and indicated how affective factors such as goals, predispositions, beliefs, and attitudes influence the writing process. Second, working memory was added to the model. This component provided a limited place for holding information and ideas for writing as well as carrying out cognitive activities that require the writer's conscious attention. Third, the long-term memory component was upgraded to include not only the writer's knowledge of the intended audience, the writing topic, and genre, but also linguistic knowledge as well as schemas for carrying out particular writing tasks, such as revising.

Hayes (1996) further reworked each of the cognitive processes identified in the original model. Planning was subsumed under a more general category, reflection, which encompassed problem-solving, decision-making, and inferencing. In the revised model, writers rely on general problem-solving (including planning) and decision-making skills to devise a sequence of steps to reach one or more writing goals. These reflective processes are abetted by inferencing, as writers make judgments or draw conclusions about their audience, possible writing content, and so forth.

Like planning, translation was also included under a more general category, entitled text production. Hayes (1996) indicated that cues from the

writing plan or the text produced so far guide the retrieval of semantic information, which is held in working memory. This information is then expressed as sentence parts that are produced vocally or subvocally and evaluated by the writer. The resulting production may be deleted, modified, or written down depending upon the subsequent evaluation.

Finally, the process of reviewing or revising was replaced by text interpretation. Hayes (1996) assigned reading a central role in text interpretation, indicating that the writer may read and evaluate text when revising, read source texts to obtain writing content, and read to define the writing task. With each of these tasks, the writer forms an internal representation of the text that can then be acted upon.

Although the original Hayes and Flower (1980) model and the revised model (Hayes, 1996) highlight the cognitive nature of writing, they remain incomplete. For instance, neither model provides an adequate description of all the major aspects of writing. The role of text transcription skills (handwriting, spelling, and so forth), for example, and their possible interaction with other cognitive skills, such as text generation and planning, is not addressed (Graham and Harris, 2000b). Even more importantly, little attention was directed at explaining how writers acquire the cognitive and noncognitive skills underlying skilled writing performance.

The issue of learning was addressed in a model of writing developed by Zimmerman and Risemberg (1997), at least in terms of the self-regulatory aspects of writing. They proposed that writers manage the composing process by bringing into play three general classes of strategies, strategies for controlling: their actions, the writing environment, and their internal thoughts and processes. Writers employ these strategies reciprocally when composing and monitor, evaluate, and react to their use. This allows them to learn from the consequences of their actions. Thus, self-regulatory strategies that are perceived to be successful are more likely to be retained, whereas those that are perceived as unsuccessful are more likely to be abandoned. It was further proposed that the writer's sense of efficacy may be enhanced or diminished depending upon the perceived success of the employed strategies. Self-efficacy, in turn, was hypothesized to influence intrinsic motivation for writing, the use of self-regulatory processes during writing, and eventual literary attainment.

Since the advent of the influential Hayes and Flower (1980) model, increasingly sophisticated descriptions of the mental operations involved in writing have emerged. The subsequent models, including those discussed here, however, are like a painting, where some parts have begun to take definite shape, other parts are being sketched in, and still other parts of the canvas remain blank (Hayes, 1996).

WRITING DEVELOPMENT

Although a great deal remains to be learned about the development of the cognitive aspects of writing, the road from novice to competent to expert writer is paved, at least in part, by changes in writers' strategic behavior (Graham and Harris, 2000b). Take, for instance, the development of planning in writing. As novice writers gain experience, there is an increase in the amount and conceptual complexity of their plans. Bereiter and Scardamalia (1987), for example, found that undergraduate students generated multiple and abbreviated lists of ideas when planning, and that these ideas were often connected by lines or arrows. Conceptual planning notes and evaluative statements (focusing on goals, structuring writing, and overcoming difficulties) were also quite common. In contrast, the planning notes developed by children in grades 4, 6, and 8 in their study showed that younger students simply generated complete sentences that were edited into a final draft when writing, while older students listed content ideas that were later worked into their compositions. Not surprisingly, strategic differences between skilled and novice writers are not limited just to planning, as the revising behavior of these two groups differs as well. For instance, skilled writers revise more for meaning and make more sentence- and theme-related changes than their developing counterparts who take a thesaurus approach to revising, focusing most of their efforts on making word substitutions and correcting mechanical errors (Fitzgerald, 1987).

As novice writers become more competent, their strategic behavior is likely to change in five ways (Alexander *et al.*, 1998). First, they become more *efficient*. Writing tasks that were once novel become more routine, and the strategies used to accomplish them become more practiced, requiring less cognitive energy and time. Second, their strategic behavior becomes more *effective*. This includes developing more powerful and elegant strategies for writing as well as using strategic knowledge more intelligently. Third, they apply strategic knowledge more *flexibly*. This includes modifying a given strategy or combining strategies in novel

ways to meet new purposes. Fourth, they become *less reliant* on strategic solutions for common writing tasks. Solutions for some writing tasks become so automatized for competent writers that they no longer require conscious or effortful processing. Finally, there is a *qualitative shift* in the strategies that writers most commonly rely on. As writers gain more competence, the importance of some strategies declines and the value of others increases.

For those writers who continue on to become experts, another shift in strategic behavior is likely to occur (Alexander *et al*., 1998). Professional writers, such as novelists or essayists, have progressed beyond merely solving writing tasks generated by others (e.g. a teacher) to formulating their own writing problems and tasks. These tasks are often novel or complex enough to require an increase in strategic behavior. Consider, for instance, Irving Wallace's description of writing a novel (Graham and Harris, 1994). He indicated that he used a variety of strategies to help him develop and manage this difficult task. This included making outlines, developing scenes and characters, working out the sequence of the story in his head and then roughly on paper, underlining story problems needing additional work, making many revisions in his plans and outlines, repeatedly rereading and revising his manuscript, constantly monitoring each step of the process, and keeping a detailed record of his progress.

The road from novice to competent to expert writer is also paved by changes in writers' knowledge (Scardamalia and Bereiter, 1986). As novice writers gain experience, their mental representations of writing become more complex and sophisticated. For example, when younger and older students are asked to describe the writing process, older students are more knowledgeable about the factors that constitute good writing, the strategies that writers employ, and the attributes of specific genres (Graham and Harris, 2000b). Advances in knowledge, however, are not just limited to the lore of composition, but involve knowledge of the topics of writing as well. For instance, writers who are more knowledgeable about a topic have to devote less cognitive effort to planning, translating, and revising than their less knowledgeable counterparts (Kellogg, 1987).

Initially, young writers have little knowledge about writing, and what knowledge they do have is fragmented and unorganized (Alexander *et al*., 1998). As they move towards competence, however, their knowledge of writing expands and becomes increasingly ordered. Those who continue on to become experts, such as professional writers, acquire exceptional levels of knowledge, and this knowledge is highly principled and integrated. A more articulated and richly interconnected knowledge base allows writers to behave more strategically and thoughtfully, as they are able to make better decisions about what they need to do to solve routine as well as novel writing problems.

Because writing is goal-directed and effortful, motivation also plays an important role in writing development. As with strategic processing and knowledge, motivational beliefs, such as self-efficacy and attributions for success, change as writers gain experience, and such beliefs are related to writing performance (Shell *et al*., 1995). Motivational factors may be especially important in determining level of effort as well as how a writer solves a particular writing problem. Consider, for instance, Robert, who was asked by his teacher to write a report (Many *et al*., 1996). Although he was familiar with how to conduct a systematic search for information and how to organize it, his search for writing content was driven by his personal interest, and not the topic of his report. As he encountered information that attracted his attention, he added it to his notes, changing the topic of his paper to maintain consistency with the newly added information.

Finally, development of the cognitive aspects of writing is influenced by a variety of contextual factors, especially schooling (Graham and Harris, 1994). Many teachers attempt to influence the course of this development in a relatively straightforward and direct fashion. They may model and explicitly teach the types of strategies and self-control procedures used by more skilled writers (Graham and Harris, 1996), or establish predictable routines where writing processes such as planning and revising are expected and reinforced. Teachers also use less direct methods to promote this development, employing instructional arrangements where children are encouraged to work together on writing projects as well as to rely on each other for input and support. This approach provides students with the opportunity to view and accommodate to other styles of composing, paving the way for each child to incorporate new influences into their own writing.

ASSESSMENT OF COGNITIVE PROCESSES

One general approach for assessing the cognitive aspects of composing is to observe writers directly

and record their behaviors in order to make direct inferences about the mental operations involved in writing. This can involve recording the amount of time that a writer devotes to a particular writing process, such as planning in advance (De La Paz and Graham, 1997). It can also involve using a computer to record the chronology, location, type, or duration of particular writing behaviors. For example, pauses that writers make while composing have been used to assess writers' cognitive planning and decision-making behavior (Matsuhashi, 1987). With this type of analysis, it is assumed that writers pause when they are planning or making decisions, and the location and duration of such pauses provides a window into these processes (e.g. longer pauses generally reflect more important decision episodes). Similarly, a record can be made of the keystrokes that occur when composing at a computer to provide a chronology of a writer's revising behavior (Eklundth and Kolberg, 1996). One advantage of direct observations and recordings is that they are generally nonreactive (i.e. they do not interact with the processes under investigation).

A second general assessment approach is to ask writers to describe what they do and what is happening while they write. For example, with the 'think-aloud' protocol used by Hayes and Flower (1980), writers are directed to say out loud everything they think and everything that occurs to them while completing a specific writing task. Think-alouds can be limited to specific processes, however, such as verbalizing thoughts just about planning. Writers can also be stopped during composing and asked to indicate what they are doing at that point. For instance, they can be stopped at specified intervals and asked to indicate if they are planning, translating, reviewing, or doing something else (Kellogg, 1987). Although these procedures are reactive and may directly influence the mental operations they are meant to assess, they provide a powerful means for accessing the writer's mental landscape during composing.

In addition to asking writers to describe what they do as they compose, they can also be asked to describe these processes retrospectively. One means for doing this is to query writers about their thoughts and actions after they have completed a specific writing task (Fitzgerald, 1987). For instance, a writer might be asked to remember what he or she was doing or thinking at particular points, while viewing a videotape of the writing session. Writers can also be asked simply to describe what they generally do when they write (Graham *et al*., 1993), such as how they go about the process of planning a paper. Although these approaches are not reactive, the resulting data must be interpreted cautiously, as writers may not remember exactly what they thought or did or they may embellish their descriptions.

A third general assessment approach is to examine writing artefacts, such as written plans or the composition itself, to draw conclusions about the mental operations employed by the writer. For example, the content and complexity of written plans can be analyzed, and the resulting written products can be examined to determine how these plans are modified and expanded as writers produce successive drafts of a composition (Bereiter and Scardamalia, 1987). Similarly, first and second drafts of a composition can be compared to ascertain not only the types of revisions made by writers, but their general approach to revising as well (Fitzgerald, 1987). Although analysis of artefacts is nonreactive, identification of intentions and the processes employed by the writers rests solely on inference.

A fourth assessment approach involves theory-embedded experimentation. With this approach, the theoretical contribution of a cognitive process to writing can be tested by explicitly teaching this process to writers who do not typically employ it when composing. For instance, novice writers who do little or no planning can be taught how to plan (De La Paz and Graham, 1997). A positive change in their planning and writing behavior following such instruction would support the contention that this process is an important component of skilled writing.

Simulation by experimentation provides a somewhat similar approach for assessing the role of cognitive processes in writing. With this approach, selected composing strategies or abilities are investigated by structuring writing tasks to simulate these processes. For example, a model of the mental operations underlying revising was tested in one study where researchers examined the effects of a procedure designed to ensure that participants used each of the hypothesized operations (Scardamalia and Bereiter, 1983).

A final approach involves assessing cognitive processing at a more global level. This may include examining performance on more general cognitive measures, such as tests of attention or working memory (Benton *et al*., 1984), or conducting magnetic resonance imaging to map regions of the brain associated with writing performance (Berninger *et al*., 2002).

WRITING DIFFICULTIES

According to data collected as part of the National Assessment of Educational Progress in the United States in 1998, one out of every five high school seniors' writing was so poor that they had not even obtained partial mastery of the writing skills and knowledge needed at that grade level. Although the origins of writing problems are not fully understood, poor writing performance is partially a consequence of difficulties with the cognitive aspects of writing (Graham and Harris, 2000b). This is reflected in findings that good writers are more strategic and knowledgeable than their poor writing counterparts. For example, good writers spend more time planning, and focus more of their attention on text-level concerns, than do struggling writers (Humes, 1983); better writers make more revisions than their less competent peers (Fitzgerald, 1987); and good writers are more knowledgeable than poor writers about the self-regulatory processes involved in composing (Englert *et al*., 1988).

These correlational findings are further buttressed by experimental studies showing that the performance of poor writers can be improved by providing them with procedural support designed to help them regulate specific writing processes. For instance, students with writing disabilities who were directed to use a simple procedure that ensured that each element of the revising process was activated and occurred in an orderly fashion made more and better revisions (De La Paz and Graham, 1997; Graham, 1997). Numerous studies have also demonstrated that the performance of poor writers can be bolstered by explicitly teaching them the types of planning, revising, and other self-regulatory strategies used by more skilled writers (Graham and Harris, 1994, 1996). Thus, an important consideration in preventing or overcoming writing difficulties is helping struggling writers master the cognitive aspects of composing.

References

Alexander P, Graham S and Harris K (1998) A perspective on strategy research: progress and prospects. *Educational Psychology Review* **10**: 129–154.

Begley S (1999, 5 April).Thinking will make it so. *Newsweek*, 64.

Benton S, Kraft R, Glover G and Plake B (1984) Cognitive capacity differences among writers. *Journal of Educational Psychology* **76**: 820–834.

Bereiter C and Scardamalia M (1987) *The Psychology of Written Composition*. Hillsdale, NJ: Lawrence Erlbaum.

Berninger V, Abbott R, Abbott S, Graham S and Richards T (2002) Writing and reading: connections between language by hand and language by eye. *Journal of Learning Disabilities* **35**: 39–56.

De La Paz S and Graham S (1997) Effects of dictation and advanced planning on the composing of students with writing and learning problems. *Journal of Educational Psychology* **89**: 203–222.

Eklundth K and Kolberg P (1996) A computer tool and framework for analyzing online revisions. In: Levy M and Ransdell S (eds) *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, pp. 163–187. Mahwah, NJ: Lawrence Erlbaum.

Englert S, Raphael T, Fear K and Anderson L (1988) Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly* **11**: 18–46.

Fitzgerald J (1987) Revision in writing. *Review of Educational Research* **57**: 481–506.

Graham S (1997) Executive control in the revising of students with learning and writing difficulties. *Journal of Educational Psychology* **89**: 781–791.

Graham S and Harris KR (1994) The role and development of self-regulation in the writing process. In: Schunk D and Zimmerman B (eds) *Self-regulation of Learning and Performance: Issues and Educational Applications*, pp. 203–228. Hillsdale, NJ: Lawrence Erlbaum.

Graham S and Harris K (1996) Self-regulation and strategy instruction for students who find writing and learning challenging. In: Levy M and Ransdell S (eds) *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, pp. 347–360. Mahwah, NJ: Lawrence Erlbaum.

Graham S and Harris K (2000a) Writing development. Introduction to the Special Issue, *Educational Psychologist* **35**: 1–62.

Graham S and Harris K (2000b) The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist* **35**: 3–12.

Graham S, Schwartz S and MacArthur C (1993) Knowledge of writing and the composing process, attitude toward writing, and self-efficacy for students with and without learning disabilities. *Journal of Learning Disabilities* **26**: 237–249.

Gregg L and Steinberg E (1980) *Cognitive Processes in Writing*. Hillsdale, NJ: Lawrence Erlbaum.

Hayes J (1996) A new framework for understanding cognition and affect in writing. In: Levy M and Ransdell S (eds) *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, pp. 1–27. Mahwah, NJ: Lawrence Erlbaum.

Hayes J and Flower L (1980) Identifying the organization of writing processes. In: Gregg L and Steinberg E (eds) *Cognitive Processes in writing*, pp. 3–30. Hillsdale, NJ: Lawrence Erlbaum.

Humes A (1983) Research on the composing process. *Review of Educational Research* **53**: 201–216.

Kellogg R (1987) Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory & Cognition* **15**: 256–266.

Many J, Fyfe R, Lewis G and Mitchell E (1996) Traversing the topical landscape: exploring students' self-directed reading-writing-research processes. *Reading Research Quarterly* **31**: 12–35.

Matsuhashi A (1987) *Writing in Real Time: Modeling Production and Process*. Norwood, NJ: Ablex.

Scardamalia M and Bereiter C (1983) The development of evaluative, diagnostic, and remedial capabilities in children's composing. In: Martlew M (ed.) *The Psychology of Written Language: A Developmental Approach*, pp. 67–95. London, UK: John Wiley.

Scardamalia M and Bereiter C (1986) Written composition. In: Wittrock M (ed.) *Handbook of Research on Teaching*, 3rd edn, pp. 778–803. New York, NY: Macmillan.

Shell D, Colvin C and Bruning R (1995) Self-efficacy, attribution, and outcome expectancy mechanisms in reading and writing achievement: grade-level and achievement-level differences. *Journal of Educational Psychology* **87**: 386–398.

Zimmerman B and Risemberg R (1997) Becoming a self-regulated writer: a social cognitive perspective. *Contemporary Educational Psychology* **22**: 73–101.

Further Reading

Berninger V, Fuller F and Whitaker D (1996) A process model of writing development: across the life span. *Educational Psychology Review* **8**: 193–210.

de Beaugrande R (1984) *Text Production: Toward a Science of Composition*. Norwood, NJ: Ablex.

Faigley L, Cherry R, Jolliffe D and Skinner A (1985) *Assessing Writers' Knowledge and Processes of Composing*. Norwood, NJ: Ablex.

Graham S and Harris K (2000) Writing development – the role of cognitive, motivational, and social/contextual factors. Special Issue, *Educational Psychologist* **35**(1): 1–62.

Harris KR and Graham S (1996) *Making the Writing Process Work: Strategies for Composition and Self-regulation*. Cambridge, MA: Brookline.

Kellogg R (1994) *The Psychology of Writing*. New York, NY: Oxford University Press.

Levy M and Ransdell S (1996) *The Science of Writing: Theories, Methods, Individual Differences, and Applications*. Mahwah, NJ: Lawrence Erlbaum.

McCutchen D (1995) Cognitive processes in children's writing: developmental and individual differences. *Issues in Education: Contributions from Educational Psychology* **1**: 123–160.

Scardamalia M and Bereiter C (1985) Fostering the development of self-regulation in children's knowledge processing. In: Chipman S, Segal J and Glaser R (eds) *Thinking and Learning Skills: Current Research and Open Questions*, vol. 2, pp. 563–577. Hillsdale, NJ: Lawrence Erlbaum.

Mathematical Processes

Introductory article

Martha W Alibali, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Mathematical thinking in the domains of number, arithmetic and algebra involves both core concepts and procedural skills. Understanding how mathematical knowledge is acquired and used can inform the design of mathematics instruction.

INTRODUCTION

Many concepts and skills are involved in mathematical thinking and reasoning. This article reviews a subset of these concepts and skills in the areas of number, arithmetic and algebra, and considers some general issues relating to mathematical thinking and how it develops. In the first section, core concepts and skills in these areas will be introduced and briefly described. The second section will address mathematical problem solving, focusing on arithmetic and algebra. The third section will address the relationships between conceptual knowledge and knowledge of problem-solving procedures. The final section will consider instructional approaches and how they can be informed by knowledge of mathematical processes.

Mathematical reasoning involves many different types of knowledge structures. These knowledge structures have been conceptualized in different ways within different psychological theories. Many theoretical frameworks distinguish between declarative and procedural knowledge. Declarative knowledge is typically defined as knowledge of facts. In mathematics, facts such as '$3+8=11$' or 'a triangle has three sides' are examples of declarative knowledge. Procedural knowledge is typically defined as knowledge of actions and how they relate to goals. In mathematics, procedural knowledge is applied when finding solutions to problems, such as performing long division or constructing a geometrical proof. Some theoretical frameworks also identify conceptual knowledge as a particular type of declarative knowledge. Conceptual knowledge is typically defined as knowledge of principles and knowledge of links among other knowledge structures. For example, a child who is learning to count may understand the principle that each object to be counted must receive a single number tag. A slightly older child might learn the principle of additive commutativity. Some theoretical frameworks also identify schemas, which are larger knowledge structures that involve many interrelated pieces. An example of a schema is understanding of the place value system.

Different individuals may draw on different knowledge structures when performing mathematical reasoning tasks. For example, when performing multi-column subtraction with borrowing, children may rely on their conceptual understanding of place value to guide their performance. They might remember that they can exchange one ten for ten ones, and they might realize that doing so will allow them to perform the necessary subtraction in the ones column. Other children might rely on their procedural knowledge of the steps involved in borrowing. They might recall that the first step in the borrowing procedure is to put a slash through the digit at the top of the tens column. Finally, both sets of children will certainly need to access their declarative knowledge of arithmetic facts (e.g. '$14-6=8$') in order to complete the task.

It is important to understand mathematical processes and how they are applied in problem-solving situations, because this understanding can guide the design of effective instruction.

KEY MATHEMATICAL CONCEPTS AND SKILLS

This section will delineate some of the central concepts and skills in the domains of number, arithmetic and algebra.

Number Concepts and the Number System

Two foundational concepts in mathematical thinking are numerical quantities and the number system. One of the earliest steps towards understanding these concepts is learning the counting string and the counting procedure. Young children often make errors in implementing the counting procedure, such as skipping items and double-counting items. The significance of such errors has been a matter of great debate in the literature on cognitive development.

The debate concerns whether there are innate knowledge structures that guide young children's learning about counting. Some theorists have argued that infants are born with knowledge of core numerical principles (e.g. one-to-one correspondence), and that mathematical development is a process of fleshing out this knowledge. According to this view, young children's counting errors stem from difficulties in performing the counting procedure, and not from lack of understanding of core numerical principles. This view has gained support from recent research on infants' mathematical understanding, which has shown that very young infants can distinguish small numerosities from one another (e.g. 1 versus 2).

Other theorists have argued against the nativist position, claiming that infants and toddlers abstract mathematical knowledge from their experiences with counting and enumeration. According to this view, young children's counting errors indicate that they have not yet acquired the principle of one-to-one correspondence.

Regardless of whether some core of quantitative knowledge is innate, it is clear that social and cultural factors are extremely important in the development of counting skills, as well as in further developments in understanding the number system. The counting string and the structure of the number system are cultural constructions and, as such, they must be learned through experience.

Another early step in mathematical thinking is understanding the principle of cardinality, which is the idea that the final count term in a sequence indicates the quantity of the counted set. With the achievement of cardinality, the number words are no longer simply part of the string of count words, but they are linked to and begin to represent real quantities.

Knowledge of counting and cardinality is an entry point into learning about the number system more generally. During the early school years, children learn about the base-10 system, about fractions and decimals and about negative numbers. Each of these domains presents significant challenges for students' understanding.

There are many factors that influence children's acquisition of these fundamental numerical concepts. Explicit instruction is of course an important factor. However, many other more subtle factors may also be involved. One not-so-obvious source of variation in mathematical development is the way in which numerical concepts are expressed in language. There is evidence that the structure of the words used to name multi-digit numbers in different languages can promote or hinder children's understanding of the base-10 structure of the number system. Children who speak languages such as English that use irregular forms in the teens decade (e.g. 'eleven') have more difficulty learning to count than children who speak languages such as Chinese that use regular forms (e.g. 'ten-one').

Arithmetic Operations

Another fundamental building block of mathematical competence is arithmetic. In the elementary years, children acquire skills in the four basic arithmetic operations of addition, subtraction, multiplication and division. Children first acquire knowledge of addition and subtraction, based on their prior knowledge of counting. Later, they acquire knowledge of multiplication and division, building on their prior knowledge of addition and subtraction.

As with the domain of counting, there has recently been debate in the developmental literature about innate knowledge structures that may guide children's acquisition of knowledge of mathematical operations. Recent studies suggest that even young infants understand that addition makes quantities larger and subtraction makes quantities smaller. Furthermore, some studies suggest that young infants have skills in the exact addition of small quantities. In one experimental paradigm, infant participants watch objects being placed behind a screen. The infants look longer when the screen is lowered to reveal an impossible outcome (e.g. one object plus one object makes one object) than when the screen is lowered to reveal a possible outcome (e.g. one object plus one object makes two objects). Based on such data, some investigators have inferred that infants are capable of arithmetic operations on small numbers.

There is also clear evidence that young children can reason non-verbally about operations such as addition and subtraction before they can perform

such operations on quantities that are represented symbolically. For example, if a 3-year-old child observes an experimenter place two chips and then three more chips under a cover, that child may be able to produce a set of five chips to show the final amount. However, most 3-year-olds cannot yet solve symbolically presented number statements such as '$2+3=?$'.

The bulk of the research on children's knowledge of mathematical operations has focused on procedural aspects of skill in arithmetic computation. Children are often explicitly taught strategies for performing such computations, and they also sometimes invent strategies based on their knowledge of counting and the number system. Recent studies have documented variability in strategy use in arithmetic across ages, within ages and within individual children. For example, when developing skills for single-digit addition, children typically begin by counting fingers for each of the addends. A child using this strategy for the problem $2+5$ might hold up two fingers on one hand and five fingers on the other hand, and count '2, 3, 4, 5, 6, 7'. Eventually children discover that they can solve addition problems by counting from the larger of the two addends, regardless of its position in the problem. A child using this strategy for the problem $2+5$ might simply count '5, 6, 7.' For many children, multiple strategies exist in their repertoires at some points in time, and they apply these strategies adaptively depending on the characteristics of the particular problem that they are attempting to solve.

Skills in complex arithmetical operations are usually learned in school over a period of several years, and these skills involve many specific procedures and notational conventions. Students must learn the carrying procedure, which is applied in multi-digit addition and multiplication, the borrowing procedure, which is used in multi-digit subtraction, and other specialized notations and procedures that are used in multi-digit multiplication and long division. Underpinning all of these skills is the concept of place value, which represents a significant challenge for many students.

Children's difficulties with place value are revealed by their frequent use of so-called 'buggy' procedures in multi-digit arithmetic. For example, in multi-digit addition, students sometimes write two-digit sums beneath each column (e.g. $85+46=1211$). In multi-digit subtraction, students often subtract the smaller number from the larger one within each column, regardless of position (e.g. $230-198=168$). In multi-digit multiplication, students often fail to line up the intermediate products correctly (e.g. $26\times15=130+26=156$). A well-developed understanding of place value can help students to avoid and eliminate such buggy procedures. However, a deep understanding of place value is often elusive, even for older students, and buggy procedures often persist in students' arithmetic for years.

Algebraic Reasoning

Arithmetic operations involve manipulations of numerical quantities. In contrast, algebraic reasoning involves operating on unknown quantities and on general statements, which are often expressed in symbolic form. In a sense, arithmetic operates in the realm of concrete quantities, whereas algebra operates in the realm of abstract quantities.

There are several core concepts that are essential to algebraic reasoning, such as variable, expression, equation and function. Students often have substantial difficulties with these concepts because they cannot straightforwardly generalize their prior knowledge of arithmetic to algebra. In algebra, the operations are not computational as they are in arithmetic. In arithmetic, operations are performed on numbers and yield other numbers, whereas in algebra, operations are performed on algebraic expressions and yield other algebraic expressions. When learning algebra, students must learn to view symbolic expressions (including those that include several terms, such as $2x+3y+14$) as mathematical objects in and of themselves.

Learning algebra largely involves learning a new form of symbolic notation and learning skills for operating on this symbolic notation in order to solve problems. In the view of many students, learning algebra boils down to learning the 'rules' of algebraic manipulation. As will be described in more detail in the section on problem solving, many students have difficulty in applying these rules correctly, and their errors often arise from an incomplete grasp of the symbolic notation used in algebra and a weak or inaccurate conceptual understanding of algebraic manipulations. Students' understanding of algebra often centers on the problem-solving procedures themselves, rather than on understanding what the procedures mean and why they work. Some recent reform efforts in algebra instruction have attempted to address this issue by emphasizing the conceptual underpinnings of the symbolic notation and of algebraic manipulations.

In addition to solving problems, competence in algebra involves the ability to generate, reason with

and translate among different ways of representing mathematical information, including symbolic equations, graphs, tables and words. In order to translate among different representations, students must have skills both for comprehending the given representation and for producing the target representation. For example, in order to generate an equation to represent a story problem, students must be able to comprehend the story sufficiently well to understand the mathematical relationships embodied in it, and they must also be able to produce an equation that represents those mathematical relationships in the symbol system of algebra. These representational skills are also emphasized in some recent reform efforts in algebra instruction.

Briefly, algebraic reasoning involves concepts, skills and notational systems that differ in important ways from those used in arithmetic. A better understanding of the cognitive processes involved in the transition from arithmetic to algebra will help to guide the design of effective instruction.

MATHEMATICAL PROBLEM SOLVING

Mathematical problem solving can be defined as the use of mathematical knowledge to answer questions and reach specific goals. This section will first introduce the distinction between informal and formal problem-solving strategies, and then review research on the development of skills for solving story problems and equations.

Informal and Formal Problem-Solving Strategies

Formal problem-solving strategies are approaches to solving problems that involve operating on mathematical symbols (e.g. numbers, variables, expressions or equations) and using procedures that were historically developed for operating on such symbolic representations (e.g. isolating the variable, factoring). Such strategies are specific to mathematics, and they tend to be learned in school or from other people.

Strategies that do not meet these criteria are termed informal strategies, and there are several different types of strategies that fall into this category. Some informal strategies are based on general approaches to problem solving that can be applied to mathematical content, but that are not specialized for mathematics (e.g. guess and test). Other informal strategies involve operating on representations that are not mathematical symbols (e.g. counting on one's fingers in order to add). Still other informal strategies involve operating on mathematical symbols in non-standard or 'shortcut' ways. Informal strategies often rely on an intuitive understanding of quantities and mathematical relationships. For example, when applying the guess and test strategy, people typically begin by estimating an approximate initial solution, and then iterate from that starting point. Many informal strategies are invented rather than taught in school.

The distinction between formal and informal strategies is not absolute, and blends of informal and formal approaches are often evident in people's attempts to solve mathematical problems.

Solving Story Problems

Both informal and formal strategies are evident in people's approaches to solving story problems. Story problems (sometimes called word problems) are defined as problems that present a story scenario in words and then pose a question about the mathematical content of the story. A simple example of an arithmetic story problem is the following. 'Jack had three cookies, and his mother gave him five more. How many cookies does he have altogether?'. A simple example of an algebraic story problem is the following. 'Margaret is saving her money for a bicycle that costs $400. She has already saved $140, and she can save $20 more each week from the money that she earns on her paper route. For how many more weeks will Margaret need to save before she can buy the bicycle?'. These two example stories will be used to illustrate students' approaches to story problems.

To solve a story problem successfully, the solver must comprehend the presented story and then use the mathematical information provided in the story to obtain a solution to the question that is posed. In the arithmetic example, this involves identifying the initial quantity (3), the quantity to be added (5) and the final or goal quantity (the total), and understanding that Jack's mother giving him cookies involves an additive relationship. In comprehending the problem, the solver may also access previously stored schemas that apply to particular types of story situations. The arithmetic problem presented above is an instance of a 'change' schema, which involves changing a set of objects from an initial state to a final state by applying some kind of transformation (in this case, adding more cookies).

Once the story is understood, a solver might attempt to reach a solution by using an informal strategy, such as counting out three fingers and then five fingers, and then enumerating the entire set of eight fingers. Alternatively, the solver might

attempt to reach the solution by using a more formal approach, namely retrieving the addition fact $3 + 5 = 8$.

In the algebraic example, comprehending the story is considerably more complex. It involves identifying the total quantity ($400), the starting quantity ($140) and the quantity to be added per unit of time ($20). Furthermore, it involves understanding that the total quantity is the sum of the starting quantity and the increments to be added, and understanding that the unknown quantity is the number of increments (i.e. the number of units of time) needed to reach the goal. Depending on the novelty and complexity of the story, the solver may either comprehend the problem with the aid of a previously stored schema, or construct a mental model of the problem situation 'from scratch'.

Once the story is understood, the solver might attempt to reach a solution by using an informal strategy, such as starting from $140 and repeatedly adding $20 until the goal of $400 is reached, while keeping track of the number of increments added. Alternatively, the solver might attempt to reach the solution by using a formal strategy, namely setting up an equation such as $140 + 20x = 400$, and solving for the unknown value x by using algebraic manipulation.

As these examples indicate, problem solvers can encounter difficulties at many different points in the process of solving story problems, and this contributes to the commonly held view that story problems are extremely difficult to solve successfully. Indeed, many people consider story problems to be more difficult to solve than corresponding symbolic problems that involve exactly the same mathematical relationships. For example, the above story problem about Margaret saving for a bicycle is thought by many people to be more difficult than the corresponding equation $140 + 20x = 400$. This is because story problems presumably involve the added difficulty of 'translating' the story scenario into a symbolic representation.

Indeed, if solving a story problem involves all of the same steps as solving the corresponding symbolic problem, plus the additional steps involved in translating to the symbolic representation, then solving the story problem must necessarily be more difficult than solving the corresponding symbolic problem. However, the above discussion should make it clear that solvers do not always solve story problems by translating to a symbolic representation and then using a formal strategy. Instead, people often use informal strategies to solve story problems, and such strategies tend to be highly successful. Story problems can actually be easier to solve than the corresponding equations, because they support the use of effective informal strategies. This is especially true of story problems that involve familiar content, because familiar content allows children to draw on their real-world knowledge of situations like those described in the stories.

Solving Equations

As noted above, students often successfully solve story problems using informal methods before they can solve the corresponding equations. Clearly, then, it is not the mathematical relationships *per se* that are challenging for students, but rather the formal symbol system in which these mathematical relationships are expressed. In particular, students often have great difficulty in generating symbolic expressions to model or represent situations mathematically. Learning the symbol system of mathematics seems to be akin to learning a foreign language and, by this analogy, expressing mathematical relationships in symbolic notation is similar to speaking or writing in a foreign language.

If students have difficulty in generating and understanding symbolic expressions, it should come as no surprise that they also experience difficulty in solving equations. One symbol that poses particular difficulties, both in the early grades and throughout middle school, is the 'equal' (=) sign. Students often interpret the equal sign as meaning 'the total' or 'get the answer', and they do not understand that it expresses a relationship between two quantities. Such students are said to have an 'operational' rather than 'relational' understanding of the equal sign. The roots of this misconception are likely to be in the types of contexts in which students encounter the equal sign during the early elementary years. Students most often see the equal sign in equations such as $7 + 4 = \square$, in which the equal sign comes at the end of the equation and precedes the symbol that is to be filled in with the solution. Hence students infer an inappropriate meaning for the equal sign, and consequently they tend to have difficulty in correctly solving nonstandard problems such as $4 + 8 + 6 = \square + 6$. It is easy to see how such a misconception could be adaptive (or at least not particularly problematic) in arithmetic contexts, but could lead to serious difficulties in the transition to algebra. Indeed, it is difficult if not impossible to integrate the idea of 'do the same things to both sides of the equation' with a conceptualization of the equal sign as meaning 'get the answer'.

A great deal of research has focused on students' acquisition of procedural skills for solving

algebraic equations. These studies have shown that structural features of equations have an important influence on students' performance and learning. Not surprisingly, equations that involve multiple operations (e.g. $4x - 20 = 43$) are more challenging than equations that involve a single operation (e.g. $4x = 63$). In addition, start-unknown equations (e.g. $2x + 30 = 44$) are more challenging than result-unknown equations (e.g. $2 * 7 + 30 = x$).

Students sometimes use rules for algebraic manipulation inappropriately, and they sometimes use incorrect rules. For example, when solving equations such as $3x + 15 = 36$, many students incorrectly apply the rule that states 'do the same thing to both sides' by doing the same thing to both sides of the plus sign. In this example, such a student might (inappropriately) divide both $3x$ and 15 by 3, yielding $x + 5 = 36$. Students also have difficulty with the order of operations rule, and they tend to have special difficulties with problems in formats such as $12 + 3x = 30$, which seem to 'invite' simplification to $15x = 30$. These types of procedural errors are often due to a poor conceptual understanding of the symbol system of algebra and of the meanings of algebraic manipulations. One strand of current research on algebraic reasoning focuses on understanding the conceptual basis for such procedural misapplications.

RELATIONSHIPS BETWEEN CONCEPTUAL AND PROCEDURAL KNOWLEDGE

Mathematical competence involves both conceptual and procedural components, and for optimal performance to be achieved, conceptual knowledge and procedural knowledge need to be well integrated. A great deal of the literature on mathematical development has focused on the relationships between these two types of knowledge. In many domains there has been significant controversy over whether conceptual knowledge precedes and sets the stage for procedural skill, or whether procedural knowledge provides the basis from which conceptual knowledge is induced or abstracted. This section will review the evidence for different temporal and causal relationships between conceptual and procedural knowledge.

Evidence that Procedural Knowledge Precedes Conceptual Knowledge

In some domains there is evidence that students acquire procedural skill before they develop conceptual knowledge. For example, in the domain of counting, several studies have shown that children can count accurately before they understand important principles that govern counting, such as (1) the one–one principle, which holds that each item must receive one and only one number tag, (2) the order-irrelevance principle, which holds that the order in which items are counted does not matter, and (3) the cardinality principle, which holds that the tag that is applied to the final item in a counted set represents the number of items in that set. In several studies, children's understanding of principles has been ascertained by asking them to evaluate the counting performance of a puppet. This technique minimizes performance demands, so that children can reveal the knowledge that they possess, even if they are unable to apply it to their own actions. These studies have shown that 3-year-old children who can execute the counting procedure properly do not always judge the puppet's violations of the one–one and order-irrelevance principles to be incorrect. Thus accurate counting precedes understanding of the one–one and order-irrelevance principles. Furthermore, most 3-year-old children are unable to predict the outcome of counting a row of objects from left to right, even if they have just counted it from right to left. Thus accurate counting also precedes understanding of the cardinality principle. In general, evidence from counting suggests that procedural skill precedes conceptual knowledge.

Students often apply procedural skills from one mathematical domain to solve problems in a related mathematical domain. For example, students often apply their procedural knowledge of whole number multiplication when learning to solve fraction multiplication problems. In such cases, children's procedural skill in the new domain can outstrip their conceptual knowledge in that domain. For this reason, many children can successfully multiply common fractions before they understand key fraction concepts. In general, children tend to display procedural skill before conceptual knowledge whenever they can induce the target procedures from an analogous procedure in a related domain. In such cases, procedural skill provides a basis for the development of conceptual knowledge.

Evidence that Conceptual Knowledge Precedes Procedural Knowledge

In some domains, students possess substantial conceptual knowledge before they acquire procedural skill. This conceptual knowledge is sometimes manifested in correct qualitative reasoning that

precedes the use of accurate problem-solving procedures. For example, there is evidence that young children understand the effects of arithmetic operations in qualitative terms (i.e. addition makes quantities larger and subtraction makes quantities smaller) before they are able to perform arithmetic computations successfully. Similar patterns have been found in older children with regard to proportional reasoning problems (e.g. story problems about mixing containers of water of different temperatures). Children understand the problems in qualitative terms (e.g. the final temperature must change in the direction of the water that was added) before they use correct procedures to actually solve the problems. This qualitative understanding may form the basis for their learning – or discovering – correct procedures.

There is also evidence that children understand certain mathematical principles before they capitalize on these principles in their problem-solving strategies. For example, children have been shown to understand additive commutativity (i.e. addend order does not matter in addition problems, such that $a + b = b + a$) before they reliably use the count-from-larger-addend procedure to solve single-digit addition problems. Some researchers have argued that children's understanding of such principles develops from their interactions with physical objects in the world, and that this understanding is gradually abstracted and applied to mathematical objects. In general, it seems likely that everyday experiences provide a basis for certain intuitive mathematical understandings, and that over historical time these intuitive understandings have been formalized in mathematical notations and procedures. Thus it is not surprising that in individual development, formal procedural skills often build on prior conceptual understanding.

Evidence that Procedural Knowledge and Conceptual Knowledge Develop Iteratively

Of course, the best answer to the question of 'which comes first' seems to be either 'both' or 'it depends'. In some cases, procedural skills are acquired first and form the basis for later achievements in conceptual understanding. In other cases, substantial conceptual knowledge appears to be present before procedural skills are learned. Recent studies have shown that the links between conceptual and procedural knowledge are bidirectional, in the sense that gains in either form of knowledge lead to corresponding gains in the other. Children who learn new problem-solving strategies often display gains in conceptual understanding, and children who learn new concepts often generate new problem-solving strategies. It seems likely that, in most cases, conceptual and procedural knowledge develop in a hand-over-hand fashion, with gains in one type of knowledge leading to gains in the other, which then feed back to influence the first.

It is clear that children's experiences at home and at school have an important role in shaping the developmental relationships between conceptual and procedural knowledge. If children's environments provide frequent experience of relevant concepts, their conceptual knowledge may outstrip their procedural skill. Alternatively, if their environments allow them the opportunity for extensive procedural practice, the reverse pattern may obtain. Among children's mathematical experiences, instruction is an especially powerful influence on mathematical thinking.

INSTRUCTIONAL APPROACHES

Mathematics instruction brings together the psychological study of mathematical thinking and learning (from both developmental and cognitive science perspectives), the study of classroom teaching and the discipline of mathematics itself. This section will briefly consider two broad aspects of mathematics instruction, namely cognitively based approaches to instruction and sociocultural influences on instruction.

Cognitively Based Approaches to Mathematics Instruction

Ideally, approaches to mathematics instruction should be informed by what is known about mathematical thinking and reasoning processes. However, the gap between psychological theories and classroom practices can be difficult to bridge. The scope of psychological analyses is sometimes too narrow and the grain size at which behavior is examined is sometimes too small for psychological findings to have immediate or direct applicability in instructional settings. For example, the study of buggy algorithms in children's multi-digit subtraction has led to a catalogue of common bugs, which can account for a large proportion of students' errors. However, is such information useful for teachers? Is it desirable or even practicable for teachers to attempt to diagnose and repair individual students' buggy procedures? Instruction that is aimed at remediating specific bugs might lead to improved student performance, but such instruction would seem to miss the target of

fostering deep, generalizable mathematical understanding.

It seems clear that a psychological perspective has the potential to inform mathematics instruction. However, for this potential to be realized, the broader implications of the psychological findings must be considered. To take the buggy subtraction example a step further, the detailed analysis of children's buggy procedures points to specific difficulties with the concept of place value that can themselves be addressed explicitly in instruction. Furthermore, if teachers know more about students' thinking, they may be better able to capitalize on students' errors as instructional opportunities. For example, a student's error in borrowing across zero might provide an excellent opportunity for a teacher to reiterate or unpack the concepts of place value and additive composition. In this way, teachers' actions may mediate the way in which knowledge of mathematical processes informs classroom practice.

Cognitive analyses of student thinking can also be a starting point and a source of ideas for curricular innovations. However, one cannot simply assume that the same processes that are involved in mathematical reasoning in experimental settings are also involved in reasoning in the classroom. The classroom is a unique sociocultural environment, and various aspects of the classroom setting may influence performance in unanticipated ways. Thus it is important that new instructional techniques be tested empirically. Evaluation is a key dimension of the process linking research on cognitive processes to mathematics instruction.

Sociocultural Influences on Mathematical Instruction

In recent years, increasing research attention has been focused on the sociocultural context of mathematics learning, both at the level of individual classrooms and at the level of the broader culture within which those classrooms are located.

The culture of each individual classroom has a profound effect on the types of mathematical learning and thinking that take place within it. Many aspects of classroom culture are significant, including the types of activities that are engaged in, the value that is attached to students' informal knowledge and invented strategies, the opportunities for social interaction and communication, and the nature of the mathematical discourse that takes place. These dimensions of classroom culture are in turn shaped by the goals, values and philosophies of the classroom teacher(s) and of the broader social structures within which the classrooms are situated (e.g. departments, schools and school districts). Textbooks also play an important role in shaping classroom culture, in part through their influence on teachers' goals and values.

At a broader level, teaching practices are also shaped by cultural norms. Recent cross-national comparisons of teaching practices have documented important differences in mathematics instruction between Japan, China, Germany and the USA. These differences cut across many aspects of lesson structure and content. For example, American lessons include more seatwork and less whole-class discussion than Japanese and Chinese lessons, and American teachers pose more factual questions and fewer conceptual questions than do Japanese teachers. These broader cultural norms shape the opportunities for learning and thinking that occur in individual classrooms within each culture.

At an even broader level, the presence or absence of a national curriculum in mathematics has an important influence on mathematics instruction. Most industrialized nations have a national curriculum. However, the USA is a notable exception. In countries with a national curriculum, there is relatively little variation across schools and across classrooms in the amount of time spent on mathematics, the specific topics covered and the level of difficulty of the implemented curriculum. In contrast, in countries without a national curriculum, there can be important differences in these dimensions across school districts and even within individual schools.

Mathematics learning is shaped by the sociocultural context at many levels, from the individual interactions of the student and the teacher, to the values and practices that constitute the classroom culture, to the national norms for the content and structure of mathematics instruction. A complete understanding of mathematical processes will require an understanding of how these layers of sociocultural context influence one another, and how they together influence mathematical thinking and learning.

SUMMARY

Understanding mathematical thinking involves identifying core concepts and skills and specifying how they develop and how they are interrelated. This article has focused on central concepts and skills in the domains of number, arithmetic and algebra. Studies of problem solving in these domains have revealed that people do not always solve mathematical problems using formal

mathematical procedures, which are typically learned in school. Instead, people often use informal strategies which rely on their intuitive mathematical knowledge.

Studies have also shown that knowledge of concepts and knowledge of problem-solving strategies can be related in different ways. In some cases, conceptual knowledge is acquired first and forms the basis for later learning of procedural skill. In other cases, procedural knowledge is acquired first, and conceptual knowledge is abstracted from procedures. Specific patterns depend on children's experiences with the relevant concepts and procedures, both at home and at school.

Instruction is a powerful force for shaping mathematical knowledge. The links between instructional practices and research on mathematical thinking are not always direct or obvious. However, there are many ways in which studies of mathematical thinking can inform the design of mathematics instruction for classroom settings.

Further Reading

De Corte E, Greer B and Verschaffel L (1996) Mathematics teaching and learning. In: Berliner DC and Calfee RC (eds) *Handbook of Educational Psychology*, pp. 491–549. New York: Macmillan.

Donlan C (1998) *The Development of Mathematical Skills*. East Sussex, UK: Psychology Press.

English LD (1997) *Mathematical Reasoning: Analogies, Metaphors and Images*. Mahwah, NJ: Erlbaum.

Geary D (1994) *Children's Mathematical Development: Research and Practical Applications*. Washington, DC: American Psychological Association.

Gelman R and Gallistel CR (1978) *The Child's Understanding of Number*. Cambridge, MA: Harvard University Press.

Grouws DA (ed.) (1992) *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan.

Reed S (1999) *Word Problems: Research and Curriculum Reform*. Mahwah, NJ: Erlbaum.

Memory

Intermediate article

David F Bjorklund, Florida Atlantic University, Boca Raton, Florida, USA
Wolfgang Schneider, University of Würzburg, Würzburg, Germany
Carlos Hernández Blasi, Universitat Jaume I, Castellón, Spain

CONTENTS

Memory can be defined as the mental storage of information and the processes involved in the acquisition, retention, and retrieval of that information.

INTRODUCTION

There are few things as pivotal to cognition as memory. Memory serves as the center stage around which both more elementary (e.g. encoding, categorization) and more elaborate (e.g. reasoning, problem-solving) forms of cognition revolve. Any application of cognitive science research to education must entail an examination of memory. This is particularly true when dealing with the education of children, who vary considerably in their memory abilities, depending on their developmental level. This article first provides a brief overview of several aspects of children's memory development. Second, it focuses on the role of two key components of memory development: strategies (middle section) and knowledge (final section). In the discussion of knowledge, we refer to how (1) expertise in a particular domain (knowledge base), and (2) a knowledge of the workings of one's own memory (metamemory) affect children's memory performance. However, as we will see, children's memory is influenced by a host of interacting factors and will vary depending on the context and the purpose to which 'memorizing' is put. (*See* **Memory, Development of**)

DEVELOPMENT

It's a fair question to ask 'When does memory begin?' Human infants at birth (and even before) are able to learn simple associations, indicating the presence of a functioning memory system (see DeCasper and Spence, 1986). By six months of age, babies can experience an event (the face of an unfamiliar person, for example), and remember that event later, as determined by the amount of time they look at that face versus a completely novel one (Fagan, 1974). Such memory performance increases over infancy, but is this the type of memory that typifies older children's cognition, especially in a school-related context? As we will see in the next section, memory is not a unified phenomenon. Rather, different types of memory or memory systems have been proposed that follow different developmental paths, and it seems that the type of memory displayed by infants is qualitatively different from that displayed by older children. One very basic, and primary, distinction in types of memory is that between implicit and explicit memory. (*See* **Memory: Implicit versus Explicit**)

The Development of Implicit and Explicit Memory

Cognitive scientists typically make a distinction between *implicit*, or *procedural memory* (sometimes known as *nondeclarative memory*) and *explicit*, or *declarative memory*. Implicit memory refers to memory without awareness, and is reflected in memory for routinized skills (i.e. procedural memory), priming, and classical and operant conditioning. Explicit memory refers to memory with awareness, or the conscious recollection of events. Explicit memory is composed of two different types of interacting systems, *episodic* and *semantic* memory. Episodic memory corresponds to a person's record of past experiences, specifically recollections that can be situated at a particular place or time. In contrast, semantic memory refers to our world knowledge, that is, knowledge of language, rules, and concepts (Tulving, 1985).

We believe that a majority of researchers would agree that the type of memory displayed by infants in conditioning and visual recognition tasks

involves implicit, and not explicit, processes. However, there is evidence for *deferred imitation* in infants and toddlers and a suggestion that this requires the long-term retention of explicit (self-aware) memories (Meltzoff, 1995). In these studies, infants watch as an experimenter demonstrates some novel behavior with an unfamiliar toy. At some later time, the infants are given the toy. If they display the novel behavior more than a control group of infants who had not previously been shown the toy, it implies they formed a long-term memory for the action they had only observed. The results of recent research reveal deferred imitation for simple actions in infants as young as nine months of age, with babies retaining these memories for as long as one year (Bauer and Wewerka, 1995). (*See* **Infant Cognition**)

But is such memory necessarily explicit? One source of evidence suggesting that it is comes from neuropsychological research examining deferred imitation in a sample of brain-damaged adults who were unable to acquire new explicit memories but could form new implicit memories (McDonough *et al*., 1995). McDonough and her colleagues (1995) administered declarative memory tasks and deferred-imitation tasks, similar to those passed by one-year-old toddlers, to a group of amnesics. The brain-damaged adults failed both sets of problems, suggesting that the same type of memory system that underlies explicit memory, as reflected by free recall and recognition memory tasks, also underlies deferred imitation and is functioning, at least in rudimentary form, by the beginning of the second year of life. (However, it should be noted that some researchers have argued that implicit and explicit memory do not actually represent different memory systems and are both available in the first year of life: see Howe, 2000; Rovee-Collier *et al*., 2001).

Although the explicit memory system may be functioning early in life, it goes through substantial development over the course of childhood. For example, three- and four-year-old children are able to recall many specific details of daily events, but they generally recall less information and require more prompts to do so than older children (Fivush, 1997). This is in contrast to implicit memory, which shows smaller age-related changes over time (e.g. Perez *et al*., 1998). For instance, in some studies children are asked to identify fragmented pictures (e.g. an umbrella). This is initially very difficult, but it becomes increasingly easier as more of the picture is provided. After completing a series of such picture-identification tasks, children are shown fragmented pictures of both previously seen and unseen objects. The typical finding is that children identify fragmented pictures of previously seen objects significantly faster than fragmented pictures of previously unseen objects, despite the fact that they may not remember (explicitly) seeing those pictures before (e.g. Hayes and Hennessy, 1996). Moreover, age differences on these tasks, and other implicit tasks (e.g. Anooshian, 1997), are typically small or nonexistent. That is, young children display memory without awareness that is superior to their explicit memory, with little age difference in levels of performance. (*See* **Working Memory**)

The Development of Working Memory

When we use the term 'memory' we often do so to refer to the long-term retention of information or to the recall of 'memories' from the distant past. But 'memory' can also refer to the relatively brief retention of small amounts of information, and developmental and individual differences in this ability, referred to as *working memory*, can greatly influence people's performance on a wide range of cognitive tasks. More specifically, working memory refers to the resource-limited mental space where active thinking occurs. In some tests of working memory, participants remember a series of items in exact order, but they are embedded in an additional task in which they must transform the target information. For instance, children may be given a set of sentences for which they must add the final word (e.g. 'In the winter it is very _____'). After hearing a series of such sentences, they are asked to recall the final word from each sentence in the order they were presented (Siegel and Ryan, 1989). In general, working-memory tasks show regular improvements with age. A similar pattern is also found for *memory span* tasks, in which children are asked only to recall in exact order a list of items without the additional cognitive 'work'. Performance on the less mentally arduous memory span tasks is typically about two items greater than on comparable working-memory span tasks.

Although these results give the impression of an increase in the absolute capacity of working memory with age, there is evidence that how much children know and how quickly they can access and articulate the stimuli plays a role in tests of working memory (Gathercole, 1998). For example, children who are experts at chess show greater memory for positions on a chess board than do nonexpert adults, despite the fact that adults have greater memory spans for digits (e.g. Chi, 1978). This and other research (e.g. Dempster, 1985) suggests that one reason for older children's

greater working memories relative to younger children's is their greater knowledge base – they know more than younger children do about most topics, which permits them to retain more information in working memory. This is critical, in that performance on working-memory span tasks is related to performance on a host of more complex cognitive tasks, including reading, mathematics, and IQ (e.g. Siegel and Ryan, 1989).

The speed with which children can articulate the words they must remember also influences memory span, with younger children generally requiring more time to articulate the stimulus words than older children, and as a result, having shorter memory spans (e.g. Hitch and Towse, 1995). One interesting implication of this is that digit words (e.g. 'one', 'two', 'three', and so on) in different languages are articulated at different rates, which contributes to cultural differences in digit span. For instance, digit words in Chinese can be articulated more rapidly than the corresponding words in English and this difference, which is found as early as four years of age, accounts for the greater digit spans of Chinese relative to American children (Geary *et al*., 1993). This suggests that one reason for the superior mathematics performance shown by young Chinese relative to American children may be the faster rate that number words can be articulated in the Chinese language.

Although factors such as knowledge base and speed of processing may influence performance on memory span tasks, there is evidence that age differences in how much children and adults can apprehend during a brief period serve as the foundation for performance on these tasks. For example, Cowan *et al*. (1999) evaluated age differences in the *span of apprehension*, which refers to the amount of information that people can attend to at a single time or the number of items that people can keep in mind at any one time. Span of apprehension can be assessed only when factors such as focused attention, knowledge of the target items, and encoding strategies are eliminated. Cowan and his colleagues (1999) had seven- and 10-year-old children and adults play a computer game while they heard a series of digits presented through earphones but which they were told to ignore. Participants were occasionally and without warning asked to recall, in exact order, the most recently presented set of digits they had heard. Because participants were not attending to the digits, it is unlikely they were using any strategies to remember them, making the task an appropriate one for assessing span of apprehension. Cowan and his colleagues reported that the average span of apprehension increased significantly with age, and they interpreted these results as reflecting a true developmental difference in the capacity of the short-term store. (*See* **Memory Mnemonics**)

STRATEGIES

Despite the limits of our working memory, people, including children, frequently retain large amounts of information for long periods of time. To do this often requires 'higher-order' cognitive processes, among them the use of memory strategies. *Strategies* refer to deliberate mental operations that are aimed at solving problems. The development of intentional cognitive control is an important question, both theoretically and practically, and the development of memory strategies lies at the center of this key issue in cognitive development.

Children are rarely fully strategic before six years of age. Although preschool children will behave strategically in some contexts (Bjorklund and Douglas, 1997), their strategies are rarely as complicated and effective as those used by older children. Effective memory strategies typically begin to develop between seven and 13 years of age, often under the tutelage of parents, teachers, or older peers. Rehearsal is an early-used strategy, although children usually adopt a 'passive' style (i.e. repeating one word at a time), and only later use the more 'active' and effective cumulative style of rehearsal (i.e. rehearsing several different words at one time: Ornstein *et al*., 1975). Young children can be easily induced to use a simple organizational strategy (e.g. to recall categorically related items together), but typically do not do so on their own until eight or nine years of age or older (Bjorklund and Douglas, 1997).

Mediational, Production, and Utilization Deficiencies

Strategies, unlike other aspects of memory development, do not always develop 'spontaneously', but sometimes require explicit instruction. The extent to which children can and do benefit from strategy instruction led to the discovery of two types of 'deficiencies': (1) *mediational deficiency*, in which children are not able to use a strategy even when one is demonstrated to them, and (2) *production deficiency*, in which children who do not use a strategy spontaneously do so when instructed and display improved task performance as a result. Most strategy development research through the 1980s was concerned with production deficiencies (mediational deficiencies have actually received

relatively little research attention), with the *training study* being the workhorse of researchers. Basically, the logic went, if we can train children to use and benefit from a strategy, we not only demonstrate that children have the underlying conceptual ability to execute the strategy, but we can learn important things about development by carefully controlling the factors we manipulate in training.

A third type of deficiency has only recently been identified (Miller, 1990). A *utilization deficiency* occurs when children use a strategy but garner little or no benefit from it. The existence of utilization deficiencies questioned the canonical interpretation of the strategy development literature – that strategy use was directly responsible for increasing memory performance. And utilization deficiencies were not rare occurrences: Miller and Seier (1994), in a review of 30 years of research, reported strong or partial evidence of utilization deficiencies in more than 90 percent of all experiments examining children's spontaneous use of memory strategies; evidence of utilization deficiencies was similarly found in more than 50 percent of memory training studies conducted over a 30-year period (Bjorklund *et al*., 1997).

Explanations for utilization deficiencies center around the fact that implementing strategies consumes too much of children's limited mental capacity, and, as a result, they do not have sufficient resources remaining to devote to the actual retrieval of items from memory. Other possible reasons for utilization deficiencies are that children are not aware that the strategies are not improving their performance, and that new strategies are used for the sake of novelty, without consideration for their consequences. The end result may be adaptive. As the novelty wears off, the strategy becomes less effortful to execute (as a result of repeated use), resulting in eventual memory benefits (Miller and Seier, 1994).

Multiple and Variable Strategy Use

Despite the presence of the various strategy deficiencies, children's memory strategies do increase in frequency and effectiveness with age. However, recent research has shown that children do not necessarily replace simple and inefficient strategies with more complicated and efficient ones in a stage-like fashion. Rather, children of all ages have a variety of different strategies available to them, with these strategies 'competing' for use (Siegler, 1996). Over time, children's modal strategies increase in sophistication, but they exist alongside less sophisticated strategies that are occasionally used when the more advanced strategy doesn't quite do the job. From this perspective, there is much variability in a child's strategy use, with this variability representing an important component of strategy development and not simply random 'noise'.

Children's multiple and variable strategy use has been demonstrated for a wide variety of tasks (see Siegler, 1996, for a review), including memory. For example, Coyle and Bjorklund (1997) gave second-, third-, and fourth-grade children a series of five sort–recall trials, using different sets of categorically related words on each trial. Children were allowed to study the words during a two-minute study period prior to each trial, during which several study strategies were observed: sorting cards into groups, category naming, and rehearsing. Above-chance levels of clustering (remembering categorically related words together) during recall were also assessed as a fourth strategy. Coyle and Bjorklund reported that children of all ages used multiple strategies on most trials, although mean number of strategies used per trial increased with age. Within an age level, children who used more strategies on a trial generally recalled more words. Children of all ages also showed substantial (and comparable) variability in strategy use, frequently switching strategies between trials. When looking at the relation between variability and recall, no consistent pattern was noted for the second- and third-grade children. However, fourth graders showed significant, although moderate, negative correlations between recall and strategy variability. That is, the more stable the fourth-grade children were in their strategy use, the higher their levels of recall tended to be. In related research, gifted second- through fourth-grade children were found to use multiple strategies as frequently as nongifted children, but recalled more words per trial and displayed lower levels of strategy variability than their nongifted peers (Coyle *et al*., 1999). Moreover, strategy variability was more consistently (and negatively) related to recall for the gifted than the nongifted children. This research points to consistent relations between multiple strategy use, strategy variability, and memory performance, making it clear that variability on cognitive tasks is more than just 'noise'.

KNOWLEDGE

Domain Knowledge and Expertise

Since the late 1970s, there has been increasing evidence for the striking effects of domain knowledge on performance in many memory tasks.

For example, several researchers have proposed that age and individual differences in children's *knowledge base* account for more of the variance in strategic memory performance than any other factor (e.g. Hasselhorn, 1995). Bjorklund (1987) proposed that having an elaborated knowledge base can affect memory in at least three ways: (1) by increasing the accessibility of specific items; (2) by making relatively effortless the activation of relations among sets of items; and (3) by facilitating the use of deliberate strategies. For example, in one line of research children were asked to recall the current members of their school class (Bjorklund and Zeman, 1982). Levels of recall were high and most children recalled their classmates' names systematically by seating arrangement, reading groups, or sex of child. Yet, there was little relation between degree of strategy use and how much children remembered. And in a later experiment when children were required to recall their classmates by a specified category (sex of child or seating arrangement), six-, eight-, and 10-year-old children did so almost perfectly, but recalled no more names than children who organized their recall less well (Bjorklund and Bjorklund, 1985). For these children, it seems, most of their elevated level of recall was attributed to nonstrategic factors associated with an elaborated knowledge base. (*See* **Knowledge Representation, Psychology of**)

The effects of knowledge on memory performance are perhaps best illustrated when children are 'experts' in a given domain such as sports, music, or science (for a review, see Schneider, 2001). For instance, Schneider, Körkel, and Weinert (1989) presented about 500 third-, fifth-, and seventh-grade children with a story about a soccer game, with the participants classified as soccer experts or novices. The expected differences between experts and novices were especially evident in the recall and comprehension of the soccer-related passage. A reversal of the typical developmental trend was observed in that third-grade experts recalled significantly more text units than both fifth- and seventh-grade novices. Moreover, there were also qualitative differences in the way experts and novices recalled the text. Whereas experts of all ages recalled more important than unimportant aspects, the soccer novices recalled as much important as unimportant text information, regardless of age. Thus performance differences between experts and novices can be attributed to both quantitative and qualitative differences in information processing. (*See* **Expertise**)

Interestingly, rich domain knowledge further compensated for low overall aptitude: when the samples of experts and novices were further subdivided into subgroups of high- and low-aptitude children, there were no significant effects found for psychometric intelligence. This finding is in accord with basic assumptions of the 'deliberate practice' model of expertise acquisition developed by Ericsson and colleagues (e.g. Ericsson, 1996). According to this model, the extent and intensity of practice during childhood and adolescence determines the level of expertise eventually reached later in adulthood. Thus levels of expertise seem to depend more on noncognitive factors such as motivation and endurance than on cognitive variables such as IQ.

The developmental studies on expertise discussed above have all demonstrated the rapid development of domain-specific knowledge in child experts and its close relationship to performance in the domain of interest. Domain knowledge also contributes to the development of other competencies that have been proposed as sources of memory development, such as basic capacities, memory strategies, and metacognitive knowledge (Schneider, 2001). Accordingly, it seems evident that changes in domain knowledge play a large role in memory development.

Metacognitive Knowledge

Although differences in content knowledge, such as those that distinguish experts and novices within a particular domain, account for substantial differences in memory performance, so also does knowledge of the workings of one's memory. Memory research using a variety of materials (e.g. categorized words, texts) has consistently demonstrated that frequent experience with tasks and materials increases children's knowledge about how to proceed in order to be effective. For instance, when children are repeatedly asked to do the kind of sort–recall task described earlier, they slowly develop an understanding of the importance of strategies such as sorting, rehearsal, and self-testing for successful performance (Bjorklund and Douglas, 1997). Children's knowledge about memory has been labeled *metamemory* (Flavell and Wellman, 1977). There are two general types of metamemory, *declarative* and *procedural*. Declarative metamemory refers to the explicit, conscious, and factual knowledge concerning person and task characteristics and memory strategies. Procedural metamemory refers to *memory monitoring* and *self-regulation* of memory behavior during ongoing memory activities. For instance, realizing that one does not learn a list of items fast enough to

reach the goal of perfect recall within a reasonable time limit constitutes an example of memory monitoring. Replacing a suboptimal learning strategy by a more effective one could be a self-regulation measure well suited to solving this problem. (*See* **Metacognition**)

Empirical research exploring the development of declarative metamemory has shown that children's knowledge of facts about memory increases considerably over the primary-grade years but is not complete by the end of childhood. Improvements in declarative metamemory are typically correlated with age-related increases in memory behavior (see Schneider, 1999, for a recent review). The average correlation between metamemory and memory performance is of moderate size (about 0.4) and increases with age.

On the other hand, developmental trends in procedural metamemory are not similarly clear. Although findings are not consistent, there is evidence that basic monitoring skills are already well developed in kindergarteners and young schoolchildren and do not improve much later on. It appears that developmental trends in procedural metamemory are mainly due to age-related increases in the interplay between monitoring and self-regulation skills. That is, whereas young schoolchildren may well be able to monitor a memory problem during a given task, only the older schoolchildren act accordingly and use self-regulation strategies in order to overcome the problem. For instance, when the task is to learn easy and difficult word pairs, both younger and older schoolchildren are able to identify the difficult pairs. However, only the older children will use this information to allocate their study time accordingly; that is, to devote more time to the study of difficult items (cf. Schneider, 1998).

CONCLUSION

Memory is not a single phenomenon, but reflects a host of related mechanisms that develop over time and are used to navigate the social and physical worlds humans inhabit. Memory is at the center of cognition for all people in all cultures. But an understanding of memory and its development becomes especially important in technological societies, where formal school requires the retention of information and techniques of remembering that would be foreign to our ancient ancestors.

References

Anooshian LJ (1997) Distinctions between implicit and explicit memory: significance for understanding cognitive development. *International Journal of Behavioral Development* **21**: 453–478.

Bauer PJ and Wewerka SS (1995) One- and two-year olds recall events: factors facilitating immediate and long-term memory in 13.5 and 16.5-month-old children. *Child Development* **64**: 1204–1223.

Bjorklund DF (1987) How age changes in knowledge base contribute to the development of children's memory: an interpretive review. *Developmental Review* **7**: 93–130.

Bjorklund DF and Bjorklund BR (1985) Organization versus item effects of an elaborated knowledge base on children's memory. *Developmental Psychology* **21**: 1120–1131.

Bjorklund DF and Douglas RN (1997) The development of memory strategies. In: Cowan N (ed.) *The Development of Memory in Childhood*, pp. 201–246. Hove, UK: Psychology Press.

Bjorklund DF, Miller PH, Coyle TR and Slawinski JL (1997) Instructing children to use memory strategies: evidence of utilization deficiencies in memory training studies. *Developmental Review* **17**: 411–442.

Bjorklund DF and Zeman BR (1982) Children's organization and metamemory awareness in the recall of familiar information. *Child Development* **53**: 799–810.

Chi MTH (1978) Knowledge structure and memory development. In: Siegler R (ed.) *Children's Thinking: What Develops?* pp. 73–96. Hillsdale, NJ: Lawrence Erlbaum.

Cowan N, Nugent LD, Elliott EM, Ponomarev I and Saults JS (1999) The role of attention in the development of short-term memory: age differences in the verbal span of apprehension. *Child Development* **70**: 1082–1097.

Coyle TR and Bjorklund DF (1997) Age differences in, and consequences of, multiple- and variable strategy use on a multitrial sort–recall task. *Developmental Psychology* **33**: 372–380.

Coyle TR, Read LE, Gaultney JF and Bjorklund DF (1999) Giftedness and variability in strategic processing on a multitrial memory task: evidence for stability in gifted cognition. *Learning and Individual Differences* **10**: 273–290.

DeCasper AJ and Spence MJ (1986) Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development* **9**: 133–150.

Dempster FN (1985) Short-term memory development in childhood and adolescence. In: Brainerd CJ and Pressley M (eds) *Basic Processes in Memory Development: Progress in Cognitive Development Research*, pp. 209–248. New York, NY: Springer-Verlag.

Ericsson KA (1996) The acquisition of expert performance: an introduction to some of the issues. In: Ericsson KA (ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Science, Sports and Games*, pp. 1–50. Mahwah, NJ: Lawrence Erlbaum.

Fagan JF III (1974) Infant recognition memory: the effects of length of familiarization and type of discrimination task. *Child Development* **45**: 351–356.

Fivush R (1997) Event memory in early childhood. In: Cowan N (ed.) *The Development of Memory in Childhood*, pp. 139–161. Hove, UK: Psychology Press.

Flavell JH and Wellman HM (1977) Metamemory. In: Kail RV and Hagen JW (eds) *Perspectives on the Development of Memory and Cognition*, pp. 3–33. Hillsdale, NJ: Lawrence Erlbaum.

Gathercole SE (1998) The development of memory. *Journal of Child Psychology and Psychiatry* **39**: 3–27.

Geary DC, Bow-Thomas CC, Fan L and Siegler RS (1993) Even before formal instruction, Chinese children outperform American children in mental arithmetic. *Cognitive Development* **8**: 517–529.

Hasselhorn M (1995) Beyond production deficiency and utilization inefficiency: mechanisms of the emergence of strategic categorization in episodic memory tasks. In: Weinert FE and Schneider W (eds) *Memory Performance and Competencies: Issues in Growth and Development*, pp. 141–159. Hillsdale, NJ: Lawrence Erlbaum.

Hayes BK and Hennessy R (1996) The nature and development of nonverbal implicit memory. *Journal of Experimental Child Psychology* **63**: 22–43.

Hitch GJ and Towse J (1995) Working memory: what develops? In: Weinert FE and Schneider W (eds) *Research on Memory Development: State-of-the-Art and Future Directions*, pp. 3–21. Hillsdale, NJ: Lawrence Erlbaum.

Howe ML (2000) *The Fate of Early Memories: Developmental Science and the Retention of Childhood Experiences*. Washington, DC: American Psychological Association.

McDonough L, Mandler JM, McKee RD and Squire LR (1995) The deferred imitation task as a nonverbal measure of declarative memory. *Proceedings of the National Academy of Sciences of the USA* **92**: 7580–7584.

Meltzoff AN (1995) What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. *Journal of Experimental Child Psychology* **59**: 497–515.

Miller PH (1990) The development of strategies of selective attention. In: Bjorklund DF (ed.) *Children's Strategies: Contemporary Views of Cognitive Development*, pp. 157–184. Hillsdale, NJ: Lawrence Erlbaum.

Miller PH and Seier WL (1994) Strategy utilization deficiencies in children: when, where, and why. In: Reese HW (ed.) *Advances in Child Development and Behavior*, vol. 25, pp. 105–156. New York, NY: Academic Press.

Ornstein PA, Naus MJ and Liberty C (1975) Rehearsal and organizational processes in children's memory. *Child Development* **46**: 818–830.

Perez LA, Peynircioglu ZF and Blaxton TA (1998) Developmental differences in implicit and explicit memory performance. *Journal of Experimental Child Psychology* **70**: 167–185.

Rovee-Collier C, Hayne H and Colombo M (2001) *The Development of Implicit and Explicit Memory*. Philadelphia, PA: John Benjamins.

Schneider W (1998) The development of procedural metamemory in childhood and adolescence. In: Mazzoni G and Nelson TO (eds) *Metacognition and Cognitive Neuropsychology*, pp. 1–22. Mahwah, NJ: Lawrence Erlbaum.

Schneider W (1999) The development of metamemory in children. In: Gopher D and Koriat A (eds) *Attention and Performance XII: Cognitive Regulation of Performance – Interaction of Theory and Application*, pp. 487–514. Cambridge, MA: MIT Press.

Schneider W (2001) Giftedness, expertise, and exceptional performance: a developmental perspective. In: Heller KA, Mönks FJ, Sternberg RJ and Subotnik RF (eds) *International Handbook of Research and Development of Giftedness and Talent*, 2nd edn. pp. 165–177. London, UK: Elsevier Science.

Schneider W, Körkel J and Weinert FE (1989) Domain-specific knowledge and memory performance: a comparison of high- and low-aptitude children. *Journal of Educational Psychology* **81**: 306–312.

Siegel LS and Ryan EB (1989) The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development* **60**: 973–980.

Siegler RS (1996) *Emerging Minds: The Process of Change in Children's Thinking*. New York, NY: Oxford University Press.

Tulving E (1985) Memory and consciousness. *Canadian Psychology* **26**: 1–12.

Further Reading

Bauer PJ (1997) Development of memory in early childhood. In: Cowan N (ed.) *The Development of Memory in Childhood*, pp. 83–111. Hove, UK: Psychology Press.

Cowan N (ed.) (1997) *The Development of Memory in Childhood*. Hove, UK: Psychology Press.

Ericsson KA, Krampe RTh and Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406.

Harnishfeger KK and Bjorklund DF (1990) Children's strategies: a brief history. In: Bjorklund DF (ed.) *Children's Strategies: Contemporary Views of Cognitive Development*, pp. 1–22. Hillsdale, NJ: Lawrence Erlbaum.

Hernández Blasi C and Bjorklund DF (2001) El desarrollo de la memoria: avances significativos y nuevos desafíos (Memory development: accomplishments of the past and directions for the future). *Infancia y Aprendizaje* **24**: 233–254.

Kuhn D (2000) Does memory development belong on an endangered topic list? *Child Development* **71**: 21–25.

Schneider W (2000) Research on memory development: historical trends and current themes. *International Journal of Behavioral Development* **24**: 407–420.

Schneider W and Bjorklund DF (1998) *Memory*. In: Kuhn D and Siegler RS (eds) *Cognitive, Language, and Perceptual Development*, vol. 2, pp. 467–521. Damon W (General Editor), *Handbook of Child Psychology*, 5th edn. New York, NY: John Wiley.

Schneider W and Pressley M (1997) *Memory Development between 2 and 20*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Siegler RS (2000) The rebirth of children's learning. *Child Development* **71**: 26–35.

Scientific Thinking

Intermediate article

Rosemary A Rosser, University of Arizona, Tucson, Arizona, USA

CONTENTS

Scientific thinking – both domain-general and domain-specific aspects of it – is a model of rationality, everyday thinking, and cognitive development.

INTRODUCTION

Why would cognitive scientists be interested in studying scientific thinking? For a number of reasons, some of them more obvious than others.

First, the reasoning behind the scientific enterprise represents an impressively high level of complexity, involving discovery, innovation, problem-solving, and intellectual creativity. Additionally, the products of scientific thought have had such great value to humankind that the cognitive processes responsible for those products are important as well as inherently interesting (Klahr, 2000). In fact, scientific thinking has become the standard for rational thought, and seeking to understand rationality is what cognitive scientists do (Stanovich, 1999).

Second, perhaps less intuitively, the processes that underlie scientific thinking may resemble those responsible for everyday thinking, the sort on which human beings ordinarily rely to understand their environments. If scientific thinking and everyday thinking is the same sort of mental 'stuff', then investigations of the former offer the promise of elucidating the phenomenon of thought more generally.

And yet a third reason: one goal of those who study cognitive development is to explain the origins of mind (Spelke, 1994; Spelke *et al*., 1992). In pursuit of that goal, they have sometimes adopted the 'child-as-scientist' metaphor for the developing intellect (Rosser, 1994), equating the naive child facing the task of figuring out the world with the scientist on a quest of discovery. The child, from this perspective, is also a 'discoverer' – acquiring data, accumulating knowledge, and constructing theories. The study of scientific thinking, then, provides information about the culmination of the developmental process, a standard with which to compare the evolving mind, and a model to represent children's thought.

In short, we study scientific thinking in order to discover: (1) the nature of a form of innovative, complex reasoning so fundamental to the Western intellectual tradition, (2) the nature of the everyday reasoning that characterizes our species, and (3) the emerging nature of the child's ability to think. In this article, research spurred by each of these pursuits will be evaluated for what it contributes to our understanding of scientific thought.

THE NATURE OF SCIENTIFIC THINKING

The scientific enterprise entails adopting a theoretical position and even-handedly evaluating evidence *vis-à-vis* the position taken. Scientific training, regardless of discipline, is about learning the procedures required to (1) gather that evidence, data, or observations – i.e. accumulate knowledge, and (2) provide coherent explanations of the knowledge accumulated. Thus, there are two critical aspects of doing science. There is the *empirical* side of science, which is tied to experimentation, observation, and data collection; and there is the *theoretical* side of science, which goes beyond empirical observation to supposition about the underlying nature of things. Rather than being mere summaries of facts, theories are *explanatory systems*, constructions from the facts that present hypotheses about causal mechanisms to account for those facts. To hold a theory of some phenomenon is to have an explanation for how the phenomenon works. Data serve as the raw material and inform theory; theory interprets data and points the way to new observations. Thus, empirical and theoretical knowledge have an important relationship with one another. The former is the sum of the descriptions, facts, and data acquired through experimentation and observation; the latter explains empirical

knowledge, augments it where facts are scarce, predicts what new facts may yet be found, and provides a cohesive, interpretive structure within which the facts make sense.

When a theory can account for the accumulating facts, there is synchrony between empirical and theoretical knowledge. In one account of the history of science (T. Kuhn, 1962), these are periods of *normal science*. During such a period, a single theory or *paradigm* predominates. Sometimes, however, incoming facts will be at odds with the theoretical account. As contradictory data accumulate, there comes pressure to re-evaluate the theory in relation to the changing data base. Theoretical reformulations that serve to bring theory into alignment with observation represent *paradigm shifts*. Scientific development can be envisioned as a succession of dominant paradigms, each prevailing for a time until empirical pressure mounts, a new replacement system evolves, and equilibrium is returned to the balance between fact and the framework for fact. Paradigm shifts entail *conceptual change* (Carey and Spelke, 1994, 1996), as new concepts better able to account for new facts are added to a theory, old concepts are adjusted, and the relations among those theoretical concepts are altered.

The scientist's first task is to collect facts and make inferences from those facts. Scientists adopt an *objective* perspective towards the task of observation and data collection: objectivity means that the procedure for recording observations and obtaining facts is conducted in such a way that the products are independent of the individual doing the observing and collecting. Generalization of principles from observations of phenomena are judged to be only as sound and supportable as the care with which scientific procedure is followed. Method and the rules of reasoning associated with that method are the *hows* of science, independent of the subject matter investigated. These 'hows' are *domain-general*, meaning that the same thinking, the same rationality, would apply to any phenomenon in any domain.

The scientist's second task is to explain the facts collected. Principles of observation and explanation may be domain-general, but the outcome of scientific endeavors, the *whats* of science, are not. Observations collected are *about* something in particular, not something in general; and the sum of the observations accumulated constitute a body of facts characterizing some specific thing(s) or event(s). The accompanying theory that proposes an explanatory 'story' for a phenomenon is also *about* something. The products of science are *domain-specific*. There are biological facts and biological theories, physical facts and physical theories, psychological facts and psychological theories. Principles from one theoretical domain, such as biology, are not used to explain phenomena from another domain, such as physics. When we examine how scientists might think, we differentiate domain-general thinking from the domain-specific variety.

DOMAIN-GENERAL THINKING

Domain-general reasoning abilities are those abstract intellectual capacities that the individual can recruit in a variety of contexts, including contexts the scientist would face when designing and conducting experiments, drawing inferences from data, and evaluating the validity of conclusions. This type of cognitive activity is typically what we construe as rationality and is a broadly applicable, consciously deployed, cognitive resource. While we might contend that this kind of intellectual activity is characteristic of what scientists do, it is characteristic of everyday thinking too (Kuhn, 1993).

In experiments intended to tap domain-general thinking, the individual is presented with a problem and given time to apply a formal solution strategy, algorithm, or rule in order to achieve a correct solution. This sort of reasoning is conscious, deliberate, demanding of cognitive resources, and context- and task-independent. That means, the mental operation deployed is a generalized abstraction applicable across tasks and circumstances. To judge the validity of an argument or solve a presented problem, the specifics of the premises or the specifics of the problem context are of less interest than the cognitive procedures and operations recruited.

Investigators whose research is focused on domain-general thinking describe these cognitive activities in terms of *strategies, heuristics, algorithms*, or *logical structures*, all of which are abstract solution devices or programs applicable across a wide variety of situations. Problems well suited for the study of this aspect of cognition include deductive reasoning tasks, analogical reasoning tasks, and problems from intelligence tests. Puzzles present an appropriate problem-solving context too, as do problems that specifically elicit scientific reasoning (Chi *et al*., 1982; Kuhn, 1989). Common features of these tasks are: the premises of the problem are specified, the way it is framed can be manipulated, and the correct solution is known. Investigators can observe how the solution is achieved; and the cognitive operations inferred from protocols may permit formal modeling.

Domain-general Thinking and Cognitive Development

The 'child-as-scientist' metaphor is a popular one for cognitive development. References to children *discovering* the workings of the world, *testing hypotheses* about that operation, and *formulating principles* of explanation link children's thinking with the scientific enterprise. Jean Piaget (Inhelder and Piaget, 1958) represents one who adopted this metaphor in a domain-general sense. He depicted the infant as an experimenter deciphering the physical laws governing the behavior of objects. He extended the metaphor to the toddler and the child acting upon the environment in order to construct representations of how things work, as scientists do. And he envisioned the important achievement of the adolescent as the ability to test hypotheses in a systematic and exhaustive manner, again much as a scientist would. The hallmark of mature thinking for Piaget was the construction of highly generalized abstractions as tools for thinking. And he envisioned development as the progressive movement towards a greater and greater *objective* understanding of the workings of the universe. The formal operational thinker behaves much like a scientist, *forming* hypotheses, *deducing* conclusions, *predicting* outcomes, and *evaluating* data. This kind of thinking is deliberate, intentional, conscious, and explicit, whether engaged in by the formally tutored scientist or by the adolescent who has reached the formal operational stage of development. Like the ideal scientist, the formal operational thinker should systematically examine all possibilities, consider the empirical evidence, and objectively evaluate each particular hypothesis in accord with that evidence.

Piaget's theory is one of the major treatments of the development of domain-general reasoning and logical thinking. Individuals' abilities to generalize cognitive routines across contexts, content, and situations was what most interested him. He construed cognitive structures as highly abstract, symbolic formalisms, and depicted development as the sequential evolution of thought from very situation-bound forms, the thought of the infant and toddler, to the symbolic forms displayed by the adolescent and adult. Essentially, Piaget argued that development involved a succession of increasingly powerful logical systems, where *powerful* is a synonym for *generality* of application. The formal reasoner can access these systems and apply a mental logic that is fully abstract, efficient, and disembedded from specific content. As a consequence, they are able to conceive of possibilities, to engage in hypothetical-deductive reasoning, and to speculate about a myriad empirical outcomes (i.e. evidence), including those not yet encountered (i.e. contradictory evidence). In short, the model for the culmination of cognitive development is scientific thinking.

How Well Do Individuals Do?

Deliberate, conscious, broadly applicable thinking routines are not likely to surface early in development. Such accomplishments require an understanding of mental operations, a cognitive sophistication not expected of the young child. Deliberate use of mental operations presumes explicit control of those operations. Again, this is unlikely with young children. And the ability to apply a strategy, rule, or algorithm across contexts is related to experiences in application; amount of experience is developmentally constrained as well. Intentional reasoning that the individual can reflect upon and explicitly describe is *not* characteristic of early emerging mental abilities. In fact, preadolescent children rarely do well on the sorts of tasks Piaget used to assess formal reasoning (Rosser, 1994). The young child does not look like a scientific thinker in the domain-general sense.

But another question is whether older individuals demonstrate the capacity to reason as a scientist would. Some evidence supports the existence of these abilities in adults, but the evidence is mixed (Johnson-Laird, 1993). For example, consider this conditional reasoning problem: 'if p then q; not q therefore not p'. With a task like this one, investigators determine whether a subject will evaluate the validity of a proposition by seeking falsifying counterexamples or *disconfirmation*. Both adolescents and adults will experience difficulty. Most subjects will ignore the possibility of counterexamples and just restrict their evaluation to the *confirmation* of 'if p, then q'. In science, one must seek to disconfirm contentions rather than confirm them; that is why the ability to search for counterexamples is so important for scientific thinking. But it is not characteristic of everyday thinking.

The factors that affect success rates on reasoning problems are related to the specific content of a problem. Indeed, psychologically realistic reasoning is often better captured by models based on the *semantics* of the problem presented (i.e. what the problem is about) than on the underlying logical *syntax* of that problem (i.e. the abstract form) (Bell and Johnson-Laird, 1998). So, even adults do not exhibit generalized domain-specific thinking all of the time.

One way of explaining this discrepancy between formal models of reasoning and what people actually do is to differentiate between *competence* and *performance*. Competence is the ideal, what the human mind is capable of doing; it is the *potential* for employing abstract reasoning strategies. Performance is what people tend to do on actual problems presented to them; it varies from the ideal as it is affected by the vagaries of the situation, by randomly occurring departures from rationality. Characterizing performance may bring us closer to modeling actual thinking. Discrepancies from the ideal are often systematic (Stanovich, 1999). For example, when individuals must evaluate evidence, they often display *bias*, as opposed to scientific objectivity (Klaczynski, 2000). One kind of bias comes from personal beliefs about content. People tend to evaluate evidence consistent with their beliefs differently than they would evidence inconsistent with those beliefs (Klaczynski and Narasimham, 1998). In fact, they are more apt to bring the cognitive resources of formal reasoning and logic to bear when evaluating inconsistent evidence, while taking logical short cuts when evaluating consistent evidence, which they are predisposed to accept anyway. Objectivity, then, is situational.

In short, belief-relevant evidence is interpreted, evaluated, and judged as a function of existing belief rather than uniformly evaluated with domain-general cognitive schemes. Scientific thinking of the modal sort is a possible cognitive achievement we use some of the time, but it may not be a probable stance we generally assume all of the time.

DOMAIN-SPECIFIC THINKING

There is another way, however, that everyday thinking may overlap with scientific thinking. Scientists collect data and also construct theories. Their theories account for the empirical phenomena. Not just scientists, but people generally may consult theories of a similar sort for a similar reason – to make sense of what is experienced and observed in the world. Like scientists, all of us need frameworks to provide conceptual glue for the bits and pieces of knowledge we accumulate. These bits and pieces form domain-specific knowledge systems (Keil, 1991), systems better characterized as *intuitive*, rather than rational; *constrained* to specific problems, rather than generally applicable; *effortless* and easily accessed, rather than demanding of cognitive resources and energy; and *pre-conscious*, rather than conscious (Klaczynski, 2000). The connection between this kind of knowledge and formal science is that untutored, domain-specific knowledge may also be organized into naive 'theories' (Gopnik and Meltzoff, 1997). A 'theory' of this sort serves as a causal explanatory system for entities which fall into a common category. So, for example, to have a naive theory of biology means one can (1) identify a set of entities that qualify as biological, (2) group those entities into a shared conceptual category, (3) apply unique causal principles to that set, and as a consequence, (4) draw inferences, make predictions, and explain outcomes, specific to biological phenomena. For both the trained biologist and her common-sense counterpart, the theory would apply only to biological phenomena, not phenomena in general. But the naive version is rooted in fundamental ways we as human beings construe the world, rather than in formal training.

What are some of these naive theories? Research attests to the existence of three commonly held, naive theories parallel to those of a scientist: (1) a physics, (2) a biology, and (3) a theory of mind, or intuitive psychology (Gopnik and Meltzoff, 1997; Keil, 1992; Spelke, 1988). These common-sense systems are like formal theories in that they (1) explain existing facts, (2) help us interpret new facts, (3) predict what facts may yet be forthcoming, and (4) change in response to the accumulation of knowledge.

Domain-specific Thinking and Development

Like its domain-general cousin, domain-specific cognition reveals both individual and developmental variation (Rosser, 1994). Expertise and maturity in one domain can be independent of another and not predictable from general indicators of intellectual functioning. Thus, a person could display precocious knowledge and reasoning in one domain but not necessarily in others, partly as a consequence of the survival relevance of a domain. The basic contention is: some information is so relevant to our survival that specialized mechanisms have evolved to help us deal with that information in very efficient ways. The prediction is: abilities mediated by such mechanisms will surface early in development.

A case in point is our understanding of other people as a special category of objects whose behavior is uniquely motivated by mental events such as intentions, beliefs, and desires. Predicting the behavior of other people by consulting a 'theory of mind' has survival relevance for a social animal. And, indeed, evidence indicates that such intuitions are available to the child early in

development (Wellman, 1990). Infants distinguish people from other objects and respond differently to people, suggesting that the domain of human behavior is construed as special. By their second year of life toddlers understand deceit, pretense, desire, and the role desire plays in motivating behavior. By their third year, preschoolers distinguish appearance from reality, reason about belief and intention, and differentiate their own perceptions and cognitions from those of other people. In short, early in development, when domain-general rationality is still limited, children are able to deal quite effectively with psychological 'stuff' (Rosser, 1994).

A second domain-specific knowledge system is associated with understanding biology (Keil, 1992). Again, children can access an intuitive theory, a 'folk biology' that is dissociable from other sorts of thinking (Hatano and Inagaki, 1994). This biology can support causal reasoning about such varied phenomena as kinship, inheritance, and biological origins, concepts of life and death, mind–body relationships, and the causes of disease and contamination. Early emerging biological notions overlap somewhat with adult conceptions, but the two versions are not identical. Common-sense biological knowledge continues to modify with increased exposure to biological information and the accumulation of facts.

As for an intuitive physics: rudimentary knowledge of physical objects, their properties, and their behavior is displayed very early in life (Baillargeon, 1994; Spelke, 1994). Infants reveal an appreciation of the constraints which operate in the physical world when they behave as if they expect objects to react to forces applied to them, move on connected paths, follow a continuous trajectory between locations, and not pass through spaces occupied by other objects (Carey and Spelke, 1996; Spelke, 1988). From this beginning, naive conceptions of physics build on core knowledge (Spelke *et al*., 1992) as children come to appreciate more subtle object properties and to understand the nature of more influences on object behavior (e.g. gravity). But the origins of a naive physics, like the basics of an intuitive biology and theory of mind, emerge very early in development.

How Well Do Individuals Do?

With such an early start to scientific thinking, can we expect the cognitively mature to exhibit sophisticated, domain-specific reasoning and problem-solving? To some extent they do, but demonstrations of expertise are primarily restricted to everyday contexts. In addition, naive theories lack the explicit conceptual richness of the formal versions. For example, the untutored can draw on a semblance of 'germ-theory' that permits them to avoid contagion, but few understand what a virus is. They can employ a 'physics' that facilitates the successful management of object-filled space and a 'psychology' that enables complex social negotiation; but these naive theories do not match formal science either. In fact, there is ample evidence, particularly in the case of physics, for pervasive misconception.

One relevant literature attests to the problem that high school and college students exhibit when faced with acquiring principles in physics (Chi, 1992; Halloun and Hestenes, 1985). Both adults and older children demonstrate misconceptions about the natural motions of objects and make erroneous predictions about trajectories (Kaiser *et al*., 1986; Proffitt *et al*., 1990). They also err, systematically not randomly, when anticipating the expected collision of a moving object and a target (Rosser and Chandler, 1995); and they misconstrue certain kinds of physical concepts, such as sound and light, and reason about them as if they were matter (Chi, 1992; Chi and Slotta, 1993). Why might they make such errors? McCloskey (1983) contends that naive theories of physics exhibit characteristics of pre-Newtonian impetus theory, a predominate paradigm during medieval times. In impetus theory the object is the center of the explanatory system, rather than the forces acting on objects. Our intuition leads us to the *bias* to be 'object-centric', and as a consequence of that bias, we perceive physical information differently from how the scientist would. Hence, our informal solutions will be incorrect some of the time. In short, naive theories, while providing us with *sufficient* explanatory systems for everyday life, do not match the conceptual systems scientists have so far devised to explain natural phenomena.

What might the source of the mismatch be? For one, informal reasoning and formal reasoning may differ in how problems are represented. In physics, for example, scientists rely on mathematical representations of events. In everyday reasoning, perception-informed representations are a more likely mediator. Also, formal problem-solving occurs in circumstances where there is time to manipulate the mathematical representations in rule-governed ways: not so for everyday problem-solving. There is rarely sufficient time, and such attempts would be much too demanding of cognitive resources. Instead, when reasoning 'online', we fall back on heuristics that are less cognitively

demanding, less effortful, but provide solutions close enough to still be adaptive (Rosser and Chandler, 1995). When a person is driving a car in traffic and approaches another car which has stopped, there is insufficient time to calibrate mathematically the dynamics of the impending collision. Intuitive reasoning, supported by perceptual representations, results in reasonable action anyway. Our intuition provides us with obligatory cognitive 'short cuts' that suffice.

A second reason intuitive and formal theories diverge has to do with differences in conceptual structure. Formal theories contain concepts absent from the informal versions. The rich data base scientists have gathered in a domain will have fueled a conceptual elaboration that need not occur in folk systems. Everyday theory accounts for everyday data (e.g. the behavior of medium-sized objects in a friction-filled environment), not all possible data (e.g. the behavior of particles in a vacuum, in space, within an atom) as a formal system must. Thus the two 'sciences' will differ conceptually. The intuitive scientist will not share the knowledge structure of the formally trained one and may not have a conceptual 'place' to put novel entities. How we categorize an entity affects what properties we assume that entity possesses and what causal explanations we will use to account for its characteristics. When the naive encounter ontologically ambiguous entities, such as sound, or light, or a virus, they may very well misassign them. So, for example, the child lacking a category for sound might assign it to the MATTER category and infer accordingly that sound takes up space, travels quickest through a vacuum, cannot pass through space occupied by other matter, and so forth. Everyday domain-specific reasoning of both children and adults shares some of the features of scientific thinking. But naive systems, though still 'theoretical', are not as informationally rich nor as conceptually elaborate.

CONCLUSIONS: SCIENTIFIC THINKING REVISITED

Scientific thinking is an inherently interesting cognitive phenomenon as a model of rationality and possibly as a model of everyday thinking. We have seen that there are two aspects of it: (1) a *domain-general* style of reasoning equally applicable to all phenomena, and (2) a *domain-specific* style, constrained to particular phenomena. We have also seen that the scientist must assume two roles: (1) the *empirical role* of objective data-collector, the accumulator of fact, and (2) the *theoretical role*, the constructor of explanatory systems to account for fact. The trained scientist understands the relationship between those roles, understands the connection between data and theory, and appreciates how evidence must be taken into account when evaluating the adequacy of theoretical contention.

The performances of individuals, including children, reveal both similarities and differences compared to the scientific model. In everyday reasoning, we collect facts and evaluate evidence too, but the objectivity we adopt will vary as a function of development, individual skill, the particulars of the task, and existing belief. Generally, ideal versions of domain-general rationality are not achievable until cognitive maturity, but even then they are not uniformly applied. While adults may be able to achieve the ideal in some circumstances, they will also exhibit systematic bias as well as random departures from rationality. Modal domain-general reasoning is cognitively effortful; apparently human beings have acquired cognitive 'short cuts' that are adaptive, conserve cognitive resources, and suffice most of the time.

Both adults and children are also naive theorists consulting intuitive explanatory knowledge systems to account for experience and observation. Unlike domain-general aspects of thought, domain-specific knowledge emerges early in development. These rudimentary systems enrich with development and the accumulation of information. Naive theories serve the same function as formal ones, but they are less elaborate, less sophisticated, less adequate, and account for fewer phenomena than the formal versions. In sum, we can reason in everyday life in ways similar to the trained scientist some of the time. We often do not; and we are simply not as good at it all of the time.

References

Baillargeon R (1994) Physical reasoning in young infants: seeking explanations for impossible events. *British Journal of Developmental Psychology* **12**: 9–33.

Bell VA and Johnson-Laird PN (1998) A model theory of modal reasoning. *Cognitive Science* **22**: 25–51.

Carey S and Spelke E (1994) Domain-specific knowledge and conceptual change. In: Hirschfeld L and Gelman L (eds) *Mapping the Mind: Domain-specificity in Cognition and Culture*. Cambridge, UK: Cambridge University Press.

Carey S and Spelke E (1996) Science and core knowledge. *Philosophy of Science* **63**: 515–533.

Chi MTH (1992) Conceptual change within and across ontological categories: examples from learning and discovery in science. In: Giere RN (ed.) *Cognitive Models of Science*. Minneapolis, MN: University of Minnesota Press.

Chi MTH, Feltovich PJ and Glaser R (1982) Categorization and the representation of physics problems by experts and novices. *Cognitive Science* **5**: 121–152.
Chi MTH and Slotta JD (1993) The ontological coherence of intuitive physics. *Cognition and Instruction* **10**: 249–260.
Gopnik A and Meltzoff A (1997) *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
Halloun IA and Hestenes D (1985) Common sense concepts about motion. *American Journal of Physics* **53**: 1056–1065.
Hatano G and Inagaki K (1994) Young children's naive theory of biology. *Cognition* **50**: 171–188.
Inhelder B and Piaget J (1958) *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: John Wiley.
Johnson-Laird PN (1993) *Human and Machine Thinking*. Hillsdale, NJ: Lawrence Erlbaum.
Kaiser MK, McCloskey M and Proffitt DR (1986) Development of intuitive theories of motion: curvilinear motion in the absence of external forces. *Developmental Psychology* **22**: 67–71.
Keil FC (1991) The emergence of theoretical beliefs as constraints on concepts. In: Carey S and Gelman R (eds) *The Epigenesis of Mind: Essays on Biology and Cognition*. Hillsdale, NJ: Lawrence Erlbaum.
Keil FC (1992) The origins of an autonomous biology. In: Gunnar MR and Maratsos M (eds) *Modularity and Constraints in Language and Cognition: Minnesota Symposia on Child Psychology*, vol. XXV. Hillsdale, NJ: Lawrence Erlbaum.
Klaczynski PA (2000) Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: a two-process approach to adolescent cognition. *Child Development* **71**: 1347–1366.
Klaczynski PA and Narasimham G (1998) The development of self-serving biases: ego-protective versus cognitive explanations. *Developmental Psychology* **34**: 175–187.
Klahr D (2000) *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.
Kuhn D (1989) Children and adults as intuitive scientists. *Psychological Review* **96**: 674–689.
Kuhn D (1993) Connecting scientific and informal reasoning. *Merrill Palmer Quarterly* **39**: 74–103.
Kuhn T (1962) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
McCloskey M (1983) Naive theories of motion. In: Gentner D and Stevens AL (eds) *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum.
Proffitt DR, Kaiser MK and Whelan SM (1990) Understanding wheel dynamics. *Cognitive Psychology* **22**: 342–373.
Rosser RA (1994) *Cognitive Development: Psychological and Biological Perspectives*. Boston, MA: Allyn & Bacon.
Rosser RA and Chandler K (1995) The influence of object conceptions on the mechanical intuitions of children and adults. *Cognitive Development* **10**: 599–620.
Spelke ES (1988) The origins of physical knowledge. In: Weiskrantz L (ed.) *Thought without Language*. Oxford, UK: Clarendon Press.
Spelke ES (1994) Initial knowledge: Six suggestions. *Cognition* **50**: 431–445.
Spelke ES, Breinlinger K, McComber J and Jacobson K (1992) Origins of knowledge. *Psychological Review* **99**: 605–632.
Stanovich KE (1999) *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.
Wellman HM (1990) *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Further Reading

Carey S and Gelman R (eds) (1991) *The Epigenesis of Mind: Essays on Biology and Cognition*. Hillsdale, NJ: Lawrence Erlbaum.
Hirschfeld L and Gelman S (eds) (1994) *Mapping the Mind: Domain-specificity in Cognition and Culture*. Cambridge, UK: Cambridge University Press.
Klahr D (2000) *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.
Stanovich KE (1999) Who is rational? *Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Self-regulated Learning

Intermediate article

Gregory Schraw, University of Nevada, Las Vegas, Nevada, USA
Douglas F Kauffman, University of Oklahoma, Norman, Oklahoma, USA
Stephen Lehman, Utah State University, Logan, Utah, USA

CONTENTS

Self-regulated learning relates to our ability to understand and control our learning environments.

DEFINITIONS AND BACKGROUND

Self-regulated learning refers to our ability to understand and control our learning environments. To do so, we must set goals, select strategies that help us achieve those goals, implement those strategies, and monitor our progress towards our goals (Schunk, 1996). Few students are fully self-regulated; however, those with better self-regulatory skills typically learn more with less effort and report higher levels of academic satisfaction (Pintrich, 2000; Zimmerman, 2000).

Self-regulated learning theory is a relatively recent development in cognitive psychology, with its origins dating back to the social-cognitive learning theory of Albert Bandura (1997). At the heart of Bandura's theory is the idea of *reciprocal determinism*, which suggests that learning is the result of personal, environmental, and behavioral factors. Personal factors include a learner's beliefs and attitudes that affect learning and behavior. Environmental factors include things such as the quality of instruction, teacher feedback, access to information, and help from peers and parents. Behavioral factors include the effects of prior performance. Reciprocal determinism states that each of these three factors affects the other two factors.

Since the 1980s, researchers have applied Bandura's (1997) social-cognitive theory to many settings, including school learning. These attempts led to the development of self-regulated learning theory, which states that learning is governed by a variety of interacting cognitive, meta-cognitive, and motivational components (Butler and Winne, 1995; Zimmerman, 2000). Social-cognitive approaches to self-regulated learning postulate that individuals learn to become self-regulated by advancing through four levels of development, including observational, imitative, self-controlled, and self-regulated levels (Schunk, 1996; Zimmerman, 2000). Learning at the observational level focuses on modeling, whereas learning at the imitative level focuses on social guidance and feedback. Both of these levels emphasize a reliance on external social factors. In contrast, as students develop, they rely more and more on internal, self-regulatory skills. Thus, at the self-controlled level, students construct internal standards for acceptable performance and become self-reinforcing via positive self-talk and feedback. At the self-regulatory level, individuals possess strong self-efficacy beliefs, as well as a large repertoire of cognitive strategies, that enable them to self-regulate their learning.

Contemporary self-regulated learning theory differs in two important ways from Bandura's social-cognitive learning theory. One is the emphasis that self-regulated learning theory places on the construction and management of cognitive strategies to control one's academic learning, otherwise known as meta-cognitive control. A second change is the inclusion of broad motivational constructs, such as causal attributions, goal orientations, and intrinsic motivation, that extend far beyond the boundaries of social-cognitive motivational constructs such as self-efficacy.

The remainder of this article is divided into five sections. The first of these describes three main components of self-regulated learning, including cognitive, meta-cognitive, and motivational components. Section two summarizes three different process models of self-regulated learning based on the work of Michael Pressley, Philip Winne, and Barry Zimmerman. Section three summarizes recent research on strategy instruction and

provides a seven-step guide to implementing strategy instruction. Section four reviews recent research on comprehension monitoring. Section five summarizes our main themes.

COMPONENTS OF SELF-REGULATED LEARNING

Experts agree that self-regulated learning includes three main components, including *cognition*, *meta-cognition*, and *motivation*. Cognition includes skills necessary to encode, memorize, and recall information. Meta-cognition includes skills that enable learners to understand and monitor cognitive processes. Motivation includes beliefs and attitudes that affect the use and development of cognitive and meta-cognitive skills. Each of these three main components is necessary for self-regulation. Those who possess cognitive skills, but are unmotivated to use them, for example, do not achieve at the same level of performance as individuals who possess skills and are motivated to use them (Zimmerman, 2000). Similarly, those who are motivated, but do not possess the necessary cognitive and meta-cognitive skills, often fail to achieve high levels of self-regulation.

The three main components of self-regulation can be further subdivided into the subcomponents shown in Figure 1. We describe each of these components below, as well as several finer-grained subcomponents.

Cognition

The cognitive component includes *encoding*, *organization*, *elaboration*, and *inferencing* subcomponents.

Encoding refers to our ability to process information currently in working memory in order to store it in long-term memory. Working memory is thought to be temporary and limited in capacity, whereas long-term memory is thought to be permanent and unlimited in capacity (Neath, 1998).

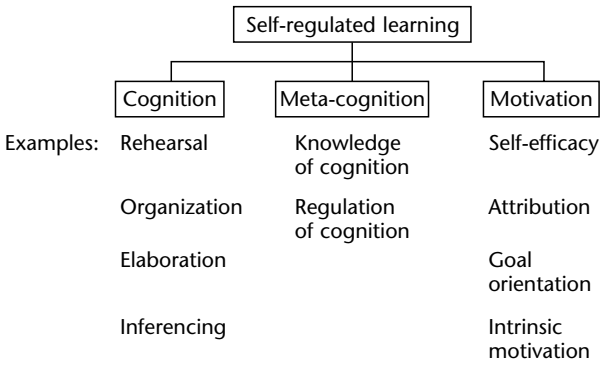


Figure 1. Components of self-regulated learning.

Organization refers to how information is sorted and arranged in long-term memory. Most experts agree that information in long-term memory is organized into knowledge structures called *schemata* and *scripts* (Neath, 1998). A schema is an organized body of declarative knowledge (i.e. stateable factual or conceptual knowledge). One example is a politics schema in which government is divided into three branches (i.e. executive, judicial, and legislative) headed by the President, Supreme Court, and Congress respectively. Schemata are crucial to self-regulated learning because they enable us to organize, store, and recall large amounts of information very quickly. Indeed, without schemata, the automated processing on which expertise depends would be impossible. Similarly, a script is an organized body of procedural knowledge that enables us to perform a task or skill automatically. Examples include scripts for getting dressed, ordering food at restaurants, riding a bicycle, and most other routine procedural activities. Collectively, schemata and scripts enable us to organize and access huge amounts of information in memory quickly.

Elaboration refers to our ability to embellish new information by linking it to information in long-term memory. Elaboration can occur at a shallow or deep level. Shallow elaboration is often referred to as *maintenance rehearsal*. For example, students may memorize the great lakes by repeating them over and over. In contrast, students could engage in several types of *elaborative rehearsal* such as creating an acronym (e.g. HOMES: Huron, Ontario, Michigan, Erie, and Superior), or constructing a mental image based on a map. Students could encode this information even more deeply by associating different lakes with different colors. For example, Lake Ontario could be remembered as orange, while Lake Michigan is remembered as magenta.

Inferencing refers to our ability to infer new information from existing knowledge and information. Inferencing is crucial to self-regulated learning because it enables us to go beyond what we know, constructing what we need to know to perform at a higher level of proficiency. The role of inference generation has been studied extensively by cognitive psychologists, particularly as it pertains to reading. Researchers know that self-regulated readers combine the facts and main ideas in a text into themes that are never stated explicitly, but are essential for comprehension.

Meta-cognition

Meta-cognition includes two main subcomponents generally referred to as *knowledge of cognition* and *regulation of cognition* (Schraw and Moshman, 1995).

Knowledge of cognition refers to what we know about our cognition, and usually includes three subcomponents. The first, *declarative knowledge*, includes knowledge about ourselves as learners and what factors influence our performance. For example, most adult learners know the limitations of their memory system and can plan accordingly based on this knowledge. *Procedural knowledge*, in contrast, refers to knowledge about strategies and other procedures. For instance, most adults possess a basic repertoire of useful strategies such as note-taking, slowing down for important information, skimming unimportant information, using mnemonics, summarizing main ideas, and periodic self-testing. Finally, *conditional knowledge* includes knowledge of why and when to use a particular strategy. Individuals with a high degree of conditional knowledge are better able to assess the demands of a specific learning situation and, in turn, select strategies that are most appropriate for that situation.

Research suggests that knowledge of cognition is late developing and explicit (Alexander *et al*., 1995). Adults tend to have more knowledge about their own cognition and are better able to describe that knowledge than children and adolescents. However, many adults cannot explain their expert knowledge and performance, and often fail to spontaneously transfer domain-specific knowledge to a new setting. This suggests that meta-cognitive knowledge need not be explicit to be useful and, in fact, may be implicit in some situations (Butler and Winne, 1995).

Regulation of cognition typically includes at least three components, planning, monitoring, and evaluation (Schraw and Moshman, 1995). *Planning* involves the selection of appropriate strategies and the allocation of resources. Planning includes goal setting, activating relevant background knowledge, and budgeting time. Previous research suggests that experts are more self-regulated compared to novices largely due to effective planning, particularly global planning that occurs prior to beginning a task. *Monitoring* includes self-testing skills necessary to control learning. Research indicates that adults monitor at both the local (i.e. an individual test item) and global levels (i.e. all items on a test). Research also suggests that even skilled adult learners are poor monitors under certain conditions (Pressley and Ghatala, 1990). *Evaluation* refers to appraising the products and regulatory processes of one's learning. Typical examples include re-evaluating one's goals, revising predictions, and consolidating intellectual gains.

Experts believe that self-regulatory processes, including planning, monitoring, and evaluation, may not be conscious or explicit in many learning situations. One reason is that many of these processes are highly automated, at least among adults. A second reason is that some of these processes may develop without any conscious reflection and therefore are difficult to report to others.

Motivation

The motivation component shown in Figure 1 includes four important subcomponents, consisting of *self-efficacy*, *attributions*, *goal orientations*, and *intrinsic motivation*.

Self-efficacy: refers to the degree to which an individual is confident that he or she can perform a specific task or accomplish a specific goal (Bandura, 1997). Self-efficacy is extremely important for self-regulated learning because it affects the extent to which learners engage and persist at challenging tasks. Previous research indicates that students with higher self-efficacy are more likely to engage in a difficult task and more likely to persist at a task even in the face of initial failures compared to low-efficacy students (Pajares, 1996). Higher levels of self-efficacy are related positively to school achievement and self-esteem. The trends observed with respect to student self-efficacy also generalize to teachers and even schools. Teachers with higher levels of teaching self-efficacy, for example, set higher goals and standards, give more autonomy to students, and help students reach higher levels of achievement than do teachers with lower levels of self-efficacy (Goddard *et al*., 2000).

Self-efficacy is affected by a number of variables, but especially vicarious learning and modeling. *Vicarious learning* occurs when individuals learn by observing others perform a skill or discuss a topic. Vicarious learning is advantageous to learners because they are not expected to perform the task, and therefore experience less anxiety, and can also focus all of their resources on observing experts. *Modeling* occurs when learners learn intentionally from other individuals such as teachers and students. Modeling typically includes the teacher breaking a complex task into manageable parts and asking students to demonstrate each part separately in sequence. Bandura (1997) proposed that modeling is effective because it raises

expectations that a new skill can be acquired, in addition to providing a great deal of knowledge about the skill. Peer models are usually the most effective because they are most similar to the learner. Indeed, students are most likely to increases their own self-efficacy when observing a model of similar ability level perform the skill (Schunk, 1996).

There are two main ways to increase students' self-efficacy. One is to use both expert (e.g. teacher) and non-expert (e.g. student peers) models. Research demonstrates that models improve cognitive skills and self-efficacy. The second is to provide as much informational feedback to students as possible. Feedback should indicate not only whether the skill was performed acceptably, but provide as much information as possible about how to improve subsequent performance. Given detailed informational feedback, performance and self-efficacy can increase even after students experience initial difficulty performing a skill.

Attributions: refer to causal explanations of events that happen in our lives. For example, two students may do poorly on a test. One student may attribute her poor performance to bad luck, while the other student attributes her poor performance to lack of effort. These attributions provide very different explanations of the same event. Attribution theory states that it is not an event *per se* that affects us, but our interpretation of that event (Graham and Weiner, 1996; Weiner, 1986).

Weiner (1986) proposed that attributions vary along three dimensions. The first is *locus of control*, which defines the cause of an outcome as either internal or external to the individual. Mood and emotions are examples of internal causes, whereas teachers are external causes. A second dimension is *stability*, which pertains to whether an attributional cause is permanent or temporary. Ability is stable, whereas effort tends to be less stable. A third dimension is *controllability*, which refers to whether an event is under the student's control or is uncontrollable. Controllable causes of academic success include effort and strategy use, whereas uncontrollable causes include luck and task difficulty.

Researchers have considered the separate effects of locus of control, stability, and controllability; however, of greater importance is how the three dimensions contribute simultaneously. Internal, controllable, stable causes such as effort promote positive academic responses, whereas external, uncontrollable, unstable causes such as luck produce frustration or undermine academic confidence. Weiner (1986) reported that internal, controllable causes, such as strategy use, promote positive affective responses, whereas internal, uncontrollable causes such as ability may create negative emotions such as shame and guilt.

Fortunately, students may be helped to change negative attributional responses through observation and training. A review of the attributional retraining literature found that the majority of attribution retraining programs are successful. Successful programs included the following three components: (1) individuals are taught to identify desirable behaviors such as effort and strategy use, (2) attributions that support positive behaviors are evaluated, (3) favorable attributional responses are rewarded. Overall, the attributional retraining literature provides evidence that individuals can learn to make more adaptive attributional responses that improve motivation and achievement-related behaviors such as effort, help seeking, and persistence.

Goal orientations: refer to beliefs about ability and how those beliefs affect learning. Dweck and Leggett (1988) proposed that learners adopt either *performance goals* or *learning goals* based on personal beliefs about the stability of intelligence. Students who believe that intelligence is fixed and unchangeable adopt performance goals, in which they seek to *prove* their competence in academic settings. Those who believe that intelligence is malleable and changeable adopt learning goals, in which they seek to *improve* their competence. A number of studies suggest that students who adopt learning goals are more adaptive and satisfied than students who adopt performance goals. Learning-oriented students typically achieve more because they seek challenge, persist, use strategies, attribute success to effort, and demonstrate positive responses to periodic failure. In contrast, performance-oriented students often adopt maladaptive response patterns characterized by avoidance of challenge, quitting after initial failure, use of inappropriate strategies, helplessness, and attributing success to uncontrollable causes such as ability and luck (Ames and Archer, 1988).

Learning- and performance-oriented students differ with respect to academic self-efficacy and self-regulated learning (Midgley *et al.*, 1995). Schunk (1996) reported that students with learning goals report higher levels of self-efficacy which, in turn, is related to higher levels of academic achievement. Bouffard *et al*. (1995) found that college students who reported strong learning goals also attained the highest level of academic self-regulation.

Learning-oriented students also appear to have better relationships with their teachers. In fact, in a study by Ames and Archer (1988),

learning-oriented students considered teachers to be more important than effort, ability, or strategy use. Surprisingly, however, learning-oriented students did not attribute their failure to teachers, whereas performance-oriented students did!

Bruning *et al*. (1999) have suggested a number of ways to foster adaptive goals. One is to promote a flexible attitude about the role of ability. Students should be encouraged to make the most of their existing ability rather than focus on how much ability they have compared to other students. Second, teachers and parents should concentrate on rewarding effort. Third, teachers should stress that mistakes are a normal part of learning and are best dealt with by persistence, help seeking, and strategy use.

Intrinsic motivation: refers to behaviors that are engaged in for their own sake (Deci and Ryan, 2000). When an individual is intrinsically motivated, tasks are performed for internal reasons such as joy and satisfaction, rather than for external reasons such as rewards, obligation, or threat. Extrinsic motivation refers to behaviors that are performed to achieve some externally prized consequence, not out of interest or personal desire for mastery. Studies reveal that performing a task because of intrinsic motivation results in satisfaction and a desire to perform the task again. In contrast, performing a task due to extrinsic motivation may lead to indifference or displeasure, and may decrease the desire to perform the task again.

Summary

Self-regulated learning refers to learners' abilities to understand and control their learning environments. Self-regulated learning involves a combination of cognitive strategy use, meta-cognitive processing, and motivational beliefs. Cognitive strategies take the form of encoding, organization, elaboration, and inference-making. Meta-cognitive processing refers to knowledge and control of cognitive skills, and usually involves planning, monitoring, and evaluating. Finally, the motivational component refers to students' beliefs in their capacity to learn. Motivation takes many forms including self-efficacy, attributions, goal orientation, and intrinsic motivation.

PROCESSES AND MODELS

Self-regulated learning is a relatively new field of study in cognitive psychology. Different theorists view self-regulated learning in different ways. Most experts agree that self-regulated learning includes the three main components described above (i.e. cognition, meta-cognition, and motivation). However, experts differ in terms of the relative contribution of each of these three components. In this section, we describe three overlapping yet distinct models of self-regulated learning. The first model is based on the work of Michael Pressley and colleagues and is known as the *Good Information Processor* model (Pressley *et al*., 1989). This model places special emphasis on the role of cognitive strategies. The second model is based on the work of Philip Winne and colleagues and is known as the *Self-regulated Learning* model. This model emphasizes the interactive relationships among cognitive, meta-cognitive, and motivational components, although it differs from the good information processor model in that it more strongly emphasizes the role of meta-cognition, and especially the role of monitoring and feedback. The third model is based on the work of Barry Zimmerman and is often referred to as the *Phases of Self-regulation* model. Though quite similar to Winne's self-regulated learning model, the phases of self-regulation model differs in that it has fewer phases and emphasizes the role of personal volition.

Pressley's Good Information Processor Model

The good information processor model was developed initially to explain effective strategy use. The model includes five main characteristics: (1) a broad repertoire of strategies, (2) meta-cognitive knowledge about why, when, and where to use strategies, (3) a broad knowledge base that is relevant to the task at hand, (4) the ability to eliminate unwanted distractions, (5) automaticity in the four components mentioned previously.

Regarding the first of these characteristics, Pressley *et al*. (1989) distinguished between two different types of strategies. The first of these include domain-specific strategies that are appropriate only for a specific task such as solving a quadratic equation. The second type is a higher-order strategy, which is used to control other lower-level strategies. One example of a higher-order strategy is sequencing the use of several domain-specific strategies while reading; tactics such as skimming before reading, drawing conclusions, then reviewing. Using higher-order strategies to orchestrate lower-level strategies is crucial to one's ability to self-regulate.

The second characteristic corresponds closely to what experts call conditional knowledge; that is,

knowledge about when, why, and where to use strategies in an optimal fashion. Conditional knowledge is important because knowing how to do something is of little practical use unless one also knows when to do it. For example, one can study the correct information for a test but do so quite poorly. Knowing how to study is at least as important as knowing what to study.

The third characteristic is a broad knowledge base. Researchers agree that learning is extremely difficult and time-consuming without supporting knowledge already in long-term memory. Indeed, a number of studies report that average-ability students with high levels of knowledge about a topic generally outperform higher-ability students with low levels of background knowledge (Bruning *et al*., 1999). In addition, background knowledge is related to the effective use of memory resources and the ability to construct integrated internal representations of a task.

The fourth characteristic is the ability to eliminate unwanted distractions. Pressley *et al*. (1989) refer to this as *action control*. Students with action control are able to motivate themselves in several ways. One is to allocate effort to the task and persist when the task is difficult. A second is to attribute their success to controllable causes such as effort and strategy use. A third is to tune out unwanted distractions.

The fifth characteristic is automaticity of the four previous characteristics. *Automaticity* refers to being able to perform a task or retrieve information from memory with little conscious effort. Automaticity is important for two interrelated reasons. First, automating lower-level cognitive skills conserves our resources for higher-level cognitive tasks that are less likely to be fully automated. Second, because fewer resources are consumed, students have more resources available to engage in complex information processing. Automaticity is one of the key components of self-regulation because our effort can be devoted to planning and monitoring the outcome of our performance, rather than performing the task.

Collectively, the five skills characteristics of good information processors enable students to self-regulate learning with a great deal of efficiency. Needless to say, each of the five components is necessary and must work in synchrony with the others. For this reason, Pressley and colleagues (Pressley and Wharton-McDonald, 1997) suggest teaching the five components described above in an integrated fashion in which all components are addressed simultaneously.

Winne's Self-Regulated Learning Model

Winne's Self-Regulated Learning model makes three broad assumptions about self-regulated learning. First, unlike the Good Information Processor model, it emphasizes the sequence of self-regulated learning over the five individual components described by Pressley and colleagues. Second, Winne conceptualizes self-regulated learning as the ability to bridge the gap between setting and achieving learning goals (Winne and Perry, 2000). Third, he highlights the importance of self-generated feedback as a mechanism that supports self-regulated learning. Winne and colleagues emphasize four stages in the self-regulated learning process, including (1) defining the task, (2) planning and goal setting, (3) enacting tactics, (4) adapting meta-cognition (Butler and Winne, 1995; Winne and Perry, 2000).

Phase one of the self-regulated learning model consists of defining the task. This phase can be broken down into two major subcomponents, task conditions and cognitive conditions that support self-regulated learning. *Task conditions* refer to factors external to the learner such as time, instructional cues, and availability of resources that affect the learner's ability to perform a task successfully. *Cognitive conditions* refer to factors internal to the learner that affect performance. Examples include personal beliefs such as self-efficacy and attributions, domain knowledge, knowledge of the task, conditional knowledge, and personal motivational factors such as goals and intrinsic motivation. According to Winne, both task and cognitive conditions conjointly influence the learner's ability to evaluate the task and formulate outcome expectations. After defining the task, the self-regulated learner must plan and set goals.

Phase two consists of planning and goal setting. In this phase, learners evaluate task and cognitive conditions information in order to establish their main goals. Learners may have multiple goals, each with its own standard for performance. Learners set these standards based on their knowledge about the task domain, automaticity performing the task, and an assessment of how well they can monitor their performance.

Phase three consists of selecting and coordinating a wide variety of cognitive learning strategies based on the goals and standards that have been set previously. These include information search strategies such as retrieving information from long-term memory, information management strategies such as identifying important information and

summarizing, and help-seeking strategies such as working in groups or asking peers or teachers for help. The purpose of enacting strategies is to produce cognitive products such as organized information in memory or written reports. In turn, products can be evaluated against the goals and standards set in phase two.

Phase four consists of using meta-cognitive knowledge, particularly monitoring, to evaluate one's performance. Winne and Perry (2000) distinguished between meta-cognitive knowledge and meta-cognitive monitoring. Meta-cognitive knowledge includes conditional knowledge about cognitive strategies, knowledge about the task, knowledge about one's current knowledge base, and one's own interests. Meta-cognitive monitoring includes judgments of one's available resources, assessing the relative difficulty of the task, evaluating current performance, and generating feedback to correct comprehension errors. Meta-cognitive knowledge and monitoring are used to assess the fit between students' initial goals and final performance. The extent to which a disparity exists necessitates a return to phase one of the cycle to eliminate this disparity. This self-regulated learning process continues until performance matches one's learning goals.

Zimmerman's Phases of Self-Regulation Model

Zimmerman and colleagues proposed a cyclical model of self-regulated learning that consists of three distinct phases: (1) forethought, (2) performance control, (3) self-reflection (Zimmerman and Kitsantas, 1999; Zimmerman, 2000). Zimmerman's model is similar to Winne's sequential self-regulated learning model in many respects, yet differs in three ways. First, motivational factors such as self-efficacy play a more influential role during the forethought phase. Second, self-affect plays a more influential role during the self-reflection phase. Third monitoring plays a smaller role in Zimmerman's model than in the Pressley and Winne models.

The forethought phase of Zimmerman's model includes two main components consisting of *task analysis* and *motivational beliefs*. Task analysis variables are similar to those described by Winne and Perry (2000), and include planning and goal setting. However, Zimmerman includes more motivational variables, emphasizing the role of self-efficacy, goal orientations, and intrinsic motivation.

The performance phase likewise includes two components consisting of *self-control* and *self-observation*. Self-control includes attention focusing strategies that enable students to tune out unwanted distractions. This component also includes the use of a wide variety of study strategies to control learning. In addition, students may also use what Zimmerman refers to as *self-instruction*, in which individuals vocalize instructions to themselves explaining how to perform a task or monitor their comprehension. The self-observation component includes a variety of record keeping activities in which learners keep track of their cognitive progress and emotional reactions.

The self-reflection phase includes both *self-judgments* and *self-reactions*. Self-judgments include monitoring one's cognitive performance, evaluating affective reactions to performance, and making appropriate causal attributions. During this phase, learners monitor whether they have met their previously established learning goals. Progress is evaluated on a number of dimensions, including whether basic goals have been mastered, how well they have performed relative to others, and how well they have performed compared to their previous performance. Self-reactions pertain mainly to judgments of their affective engagement. Zimmerman (2000) argues that task satisfaction depends on one's ability to meet previously established goals and standards.

Summary

The three models of self-regulation described above agree on the need for cognitive, meta-cognitive and motivational components. They differ in the extent to which these main components interact. Pressley's Good Information Processor Model focuses more on the componential makeup of self-regulated learning, while the other models emphasize the sequential (Winne) or the cyclical (Zimmerman) nature of the self-regulated learning process. In Winne's model, the process of bridging the gap between setting and achieving goals, as well as the importance of self-generated feedback is crucial. In contrast to Winne, Zimmerman emphasizes the importance of motivational variables, such as self-efficacy as well as suggesting that self-reactions play an integral part in the learner's continued regulatory processes.

SELF-REGULATION AND STRATEGIES

Research on Strategy Instruction

Research on strategy instruction has boomed since the early 1980s. Unfortunately, implementing a

strategy training program is time-consuming and expensive; thus, most interventions have focused on the effectiveness of only one or two strategies. Researchers have conducted *meta-analyses* to better understand the effectiveness of strategy instruction. A meta-analysis is a procedure that aggregates similar studies to determine their overall effectiveness. Two analyses by Hattie *et al*. (1996) and Rosenshine *et al*. (1996) supported the following four claims:

1. Strategy instruction typically is moderately to highly successful.
2. Strategy instruction appears to be most helpful for younger and under-achieving students.
3. Programs that combine several interrelated strategies are more effective than those that include only one strategy. An interrelated repertoire of four or five strategies seems optimal (Pressley and Wharton-McDonald, 1997).
4. Strategy interventions are more effective when they teach conditional knowledge.

Strategy research has also addressed whether strategy instruction is more effective in teacher-centered versus student-centered classrooms. Neither type of setting appears to increase the effectiveness of interventions, although it should be noted that few studies have compared different instructional approaches directly.

Another important question addressed by strategy instruction research is what strategies should be taught to students. Experts generally agree that a limited number of general strategies are most effective (e.g. 4–8). Based on their analysis, Hattie *et al*. (1996) suggested the following set of strategies: self-checking, creating a good study environment, planning and goal setting, reviewing, summarizing, and seeking teacher and peer assistance. Similarly, in a comprehensive review of the strategy instruction literature, Dole *et al*. (1991) recommended a similar set of five core learning strategies that included determining what is important to learn, summarizing, drawing inferences, generating questions before and during studying, and monitoring one's comprehension.

How to Teach Strategies

Strategy instruction should be an integral part of every class (Pressley and Wharton-McDonald, 1997). Prior to strategy instruction, teachers should help students understand the value of strategies and decide which strategies to teach their students. Most experts recommend an instructional sequence that stretches from 10 to 20 weeks in which strategies are introduced, modeled, practiced, and finally automated. The following seven-step sequence is typical of effective strategy instruction programs:

1. Discuss and explain the value of strategies. Strategies increase efficiency, save time, and enhance deeper processing.
2. Introduce a limited number of strategies. Most programs recommend four or five general strategies such as summarizing and comprehension monitoring.
3. Practice each strategy over an extended period of time until it becomes automated.
4. Model strategies extensively so students acquire not only the strategy, but conditional knowledge about how, when, and where to use the strategy.
5. Provide feedback to students about strategy use. This information helps them evaluate the effectiveness of strategies and monitor their comprehension more effectively.
6. Promote transfer by encouraging students to use strategies in new settings or by adapting them to new tasks. Previous research suggests that strategies learned in one setting do not transfer unless students are instructed to use them in different settings.
7. Encourage reflection on strategy use. Students who reflect on strategy use acquire more meta-cognitive knowledge and are more apt to use strategies in a flexible way to self-regulate. One way students can self-reflect is through journals. A second way is by comparing the advantages and disadvantages of different strategies with peers.

Overall, research suggests that effective strategy use is critical for self-regulated learning. Effective strategy use is accomplished most efficiently through the extended instruction, modeling and practice of a small repertoire of general strategies such as planning, inferencing, and monitoring. In addition to a repertoire of strategies, increasing self-regulated learning is dependent on the learner's ability to monitor comprehension.

COMPREHENSION MONITORING

Comprehension monitoring refers to evaluating the ongoing state of one's understanding. Monitoring takes place during or after a learning activity and provides information about the effectiveness of that activity. Monitoring is important because it provides self-generated feedback to the learner. Without accurate monitoring, efficient control of one's performance is impossible.

Monitoring studies typically require individuals to make subjective judgments of learning or test performance during or after an initial study phase. Four types of judgments have been used in the adult monitoring literature, including *ease of learning* (i.e. judgments of encoding difficulty),

judgments of learning (i.e. the degree to which information was learned during the study phase), *feeling of knowing* (i.e. the degree to which one has access to previously learned information in memory), and *performance judgments* (i.e. assessments of performance accuracy).

Studies measuring these four types of judgments indicate that adults monitor their learning and performance with a moderate degree of success, although results vary from study to study. Surprisingly, monitoring proficiency does not appear to be related strongly to relevant domain knowledge or academic achievement (Pressley and Ghatala, 1990). These conclusions have been supported in the children's monitoring literature as well, although there is considerable debate regarding whether children monitor as accurately as adults (Alexander *et al*., 1995).

These studies indicate that there are three specific factors that affect monitoring proficiency. First, situational constraints affect estimates of monitoring proficiency. One constraint is the point in the learning-test sequence in which monitoring judgments are made. A number of studies indicate that calibration of comprehension (i.e. the correlation between pre-test judgments and actual test performance) is often quite poor, with most studies reporting correlations in the 0.00 to 0.25 range (Pressley and Ghatala, 1990). In contrast, calibration of performance (i.e. the correlation between post-test judgments and actual test performance) appears to be much better in both children and adults, often ranging from 0.30 to 0.50 (Pressley and Ghatala, 1990).

Second, specific testing conditions affect monitoring proficiency. For example, calibration of comprehension can be improved under the following circumstances: (1) when adjunct questions similar to post-test questions are provided during study, (2) when periodic feedback is provided to test takers, (3) when expert knowledge about the to-be-learned material is *minimized*, (4) when test takers generated missing text information. Surprisingly, calibration of comprehension does not appear to improve when learners are specifically requested to monitor their comprehension or when they are given the opportunity to re-study the to-be-learned materials, or when they are given practice questions prior to study.

Like calibration of comprehension, calibration of performance improved under a number of testing conditions, especially when adjunct questions were provided during the study phase, when test takers received external incentives to improve monitoring accuracy, and when test takers received recall rather than recognition tests. Calibration of performance was also related to level of test performance. In addition, individuals monitored with less bias when judging their performance on easy rather than more difficult items.

Third, feedback, incentives, practice, and training positively affect monitoring proficiency. Schraw (1994) reported that pre-experimental estimates of monitoring proficiency were related to both local (i.e. the accuracy of item-specific performance judgments made during testing) and global (i.e. judgments of overall performance made after testing) monitoring accuracy. The accuracy of local monitoring was correlated positively to the accuracy of global monitoring. In addition, the change in monitoring accuracy between local and global monitoring improved significantly among good monitors, but did not improve among poor monitors.

Monitoring training also improves performance. Delclos and Harrington (1991) examined fifth- and sixth-graders' ability to solve computer problems after assignment to one of three conditions. The first group received specific problem-solving training, the second received problem-solving plus self-monitoring training and practice, while the third received no training. The monitored problem-solving group solved more of the difficult problems than either of the remaining groups and took less time to do so. The group receiving problem-solving and monitoring training also solved complex problems faster than the control group.

The monitoring research summarized above leads to a number of conclusions. Overall, adults monitor their performance with a moderate degree of accuracy. Monitoring accuracy improves as tests become easier and more factual. Second, monitoring proficiency appears to be independent of intellectual ability (Alexander *et al*., 1995) and academic achievement (Pressley and Ghatala, 1990). Third, monitoring proficiency may be independent or even negatively related to domain knowledge, independent of ease of comprehension judgments, but correlated with other types of meta-cognitive knowledge. Fourth, one's ability to monitor one's performance may improve with practice (Delclos and Harrington, 1991).

SUMMARY

Self-regulated learning theory evolved from Bandura's (1997) social-cognitive learning theory. In 2002 self-regulated learning theory focuses on the transition from social to self-directed learning processes. Several main themes emerge from this

research. The first is that self-regulated learners rely on an integrated repertoire of cognitive, meta-cognitive, and motivational skills. Second, self-regulated learners use these skills to plan, set goals, implement and monitor strategy use, and evaluate their learning goals. Third, self-regulated learners use a wide variety of strategies in flexible ways, augmenting these strategies with a variety of adaptive motivational beliefs such as high self-efficacy, attributions to internal controllable causes of academic success, learning goals, and intrinsic motivation.

A review of the strategy instruction literature suggests that a repertoire of four or five general strategies such as summarizing and comprehension monitoring skills can be taught to students of all ages. Strategy instruction is most effective when it extends over a 10- to 20-week period, includes extensive modeling and feedback, teaches conditional knowledge necessary for effective strategy use, and helps students recognize the cognitive and motivational benefits of strategy use. In addition to strategies, self-regulated learners monitor their comprehension and debug learning problems when they occur. Research suggests that monitoring is unrelated to ability and improves with practice.

Overall, there is strong agreement that self-regulated learning is necessary for academic success and is attributable to meta-cognitive knowledge and a repertoire of learning strategies, rather than ability *per se*. These skills are learned through observation and modeling, feedback from others, and consistent practice. Models also provide important motivational support.

References

Alexander JM, Carr M and Schwanenflugel PJ (1995) Development of metacognition in gifted children: directions for future research. *Developmental Review* **15**: 1–37.

Ames C and Archer J (1988) Achievement in the classroom: student learning strategies and motivational processes. *Journal of Educational Psychology* **80**: 260–267.

Bandura A (1997) *Self-Efficacy: The Exercise of Control*. New York, NY: Freeman.

Bouffard T, Boisvert J, Vezeau C and Larouche C (1995) The impact of goal orientation on self-regulation and performance among college students. *British Journal of Educational Psychology* **65**: 317–329.

Bruning R, Schraw G and Ronning R (1999) *Cognitive Psychology and Instruction*, 3rd edn. Upper Saddle River, NJ: Merrill.

Butler DL and Winne PH (1995) Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research* **65**: 245–281.

Deci EL and Ryan RM (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well being.

Delclos VR and Harrington C (1991) Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology* **83**: 35–42.

Dole JA, Duffy GG, Roehler LR and Pearson PD (1991) Moving from the old to the new: research on reading comprehension instruction. *Review of Educational Research* **61**: 239–264.

Dweck CS and Leggett ES (1988) A social-cognitive approach to motivation and personality. *Psychological Review* **95**: 256–273.

Goddard RD, Hoy WK and Woolfolk Hoy A (2000) Collective teacher efficacy: its meaning, measure and impact on student achievement. *American Educational Research Journal* **37**: 479–508.

Graham S and Weiner B (1996) Theories and principles of motivation. In: Berliner DC and Calfee RC (eds) *The Handbook of Educational Psychology*, pp. 63–84. New York, NY: Macmillan.

Hattie J, Biggs J and Purdie N (1996) Effects of learning skills interventions on student learning: a meta-analysis. *Review of Educational Research* **66**: 99–136.

Midgley C, Anderman EM and Hicks L (1995) Differences between elementary and middle school teachers and students: a goal theory approach. *Journal of Early Adolescence* **15**: 90–113.

Neath I (1998) *Human Memory: An Introduction to Research, Data, and Theory*. Pacific Grove, CA: Brooks-Cole Publishing.

Pajares F (1996) Self-efficacy beliefs in academic settings. *Review of Educational Research* **66**: 543–578.

Pintrich P (2000) The role of goal orientation in self-regulated learning. In: Boekaerts M, Pintrich P and Zeidner M (eds) *Handbook of Self-Regulation*, pp. 452–501. San Diego, CA: Academic Press.

Pressley M, Borkowski J and Schneider W (1989) Good information processing: what is it and what education can do to promote it. *Journal of Experimental Child Psychology* **43**: 194–211.

Pressley M and Ghatala ES (1990) Self-regulated learning: monitoring learning from text. *Educational Psychologist* **25**: 19–33.

Pressley M and Wharton-McDonald R (1997) Skilled comprehension and its development through instruction. *School Psychology Review* **26**: 448–466.

Rosenshine B, Meister C and Chapman S (1996) Teaching students to generate questions: a review of the intervention studies. *Review of Educational Research* **66**: 181–221.

Schraw G (1994) The effect of meta-cognitive knowledge on local and global monitoring. *Contemporary Educational Psychology* **19**: 143–154.

Schraw G and Moshman D (1995) Metacognitive theories. *Educational Psychology Review* **7**: 351–371.

Schunk D (1996) Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal* **33**: 359–382.

Weiner B (1986) *An Attributional Theory of Motivation and Emotion*. New York, NY: Springer-Verlag.

Winne P and Perry N (2000) Measuring self-regulated learning. In: Boekaerts M, Pintrich P and Zeidner M (eds) *Handbook of Self-Regulation*, pp. 531–566. San Diego, CA: Academic Press.

Zimmerman B (2000) Attaining self-regulated learning: a social-cognitive perspective. In: Boekaerts M, Pintrich P and Zeidner M (eds) *Handbook of Self-Regulation*, pp. 13–39. San Diego, CA: Academic Press.

Zimmerman B and Kitsantas A (1999) Acquiring writing revision skills: shifting from process to outcome self-regulatory goals. *Journal of Educational Psychology* **91**: 241–250.

Further Reading

Brown R and Pressley M (1994) Self-regulated reading and getting meaning from text: the transactional strategies instruction model and its ongoing validation. In: Schunk DH and Zimmermann BJ (eds) *Self-regulation of Learning and Performance: Issues and Educational Implications*, pp. 155–180. Hillsdale, NJ: Lawrence Erlbaum.

Ericsson KA (1996) The acquisition of expert performance. In: Ericsson KA (ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts, Sciences, Sports, and Games*, pp. 1–50. Mahwah, NJ: Lawrence Erlbaum.

Ericsson KA and Kintsch W (1995) Long-term working memory. *Psychological Review* **102**: 211–245.

Hofer B, Yu S and Pintrich P (1998) Teaching college students to be self-regulated. In: Schunk DH and Zimmerman BJ (eds) *Self-regulated Learning: From Teaching to Self-reflective Practice*, pp. 57–85. New York, NY: Guilford.

Kintsch W (1998) *Comprehension: A Paradigm for Cognition*. Cambridge, UK: Cambridge University Press.

Schunk DH and Ertmer PA (2000) Self-regulation and academic learning: self-efficacy enhancing interventions. In: Boekaerts M, Pintrich P and Zeidner M (eds) *Handbook of Self-regulation*, pp. 631–649. San Diego, CA: Academic Press.

Schunk DH and Zimmerman BJ (1998) Conclusions and future directions for academic interventions. In: Schunk DH and Zimmerman BJ (eds) *Self-regulated Learning: From Teaching to Self-reflective Practice*, pp. 225–235. New York, NY: Guilford.

Zeidner M, Boekaerts M and Pintrich P (2000) Self-regulation: directions and challenges for future research. In: Boekaerts M, Pintrich P and Zeidner M (eds) *Handbook of Self-regulation*, pp. 750–768. San Diego, CA: Academic Press.

Zimmerman B (1998) Academic studying and the development of personal skill: a self-regulatory perspective. *Educational Psychologist* **33**: 73–86.

Thinking Skills

Intermediate article

Diane F Halpern, Claremont McKenna College, Claremont, California, USA

CONTENTS

There are many different types of thinking skills and frameworks for categorizing them. Good or clear thinking is called 'critical thinking'. Critical thinking skills are those skills that increase the probability of a desirable outcome, and are essential to problem-solving and decision-making.

INTRODUCTION

Thinking is commonly regarded as the highest form of human activity. It is a universal, private activity, inferred from behavior. Thinking is a multivariate construct, and people vary widely in the skill and ease with which they think. Some psychologists define 'thinking' as the mental manipulation of symbolic representations of objects and of relations among objects. Consider, for example, a chess player who is contemplating the next move; in deciding which move to make, she needs to mentally envision or rehearse a limited subset of all possible moves and countermoves by her opponent. Skillful chess players use strategies or plans for deciding how to move. Similarly, an architect who is designing a building will mentally manipulate a symbolic representation of the building that is being planned in deciding how to design the structure. 'Thinking skills' are the plans, strategies, or tactics for good or clear thinking. The terms used to define thinking skills – plans, strategies, tactics – are also used in military science to refer to military maneuvers designed to achieve an objective, which is an apt analogy for the abstract concept of skilled thinking. Good thinkers use higher-order cognitive skills, and, consequently, have better outcomes than poor thinkers.

There has been a growing interest in understanding and enhancing thinking skills because contemporary society is more complex than that at any previous time in history. Advances in technology have created an information glut, and workplace skills are increasingly technical at all but the very lowest levels of employment. As a result of these changes, a greater proportion of the general population needs more advanced thinking skills. Thus, there is increased concern about the identification, development, and transfer of those skills that characterize clear, precise, purposeful thinking.

Good or clear thinking is called 'critical thinking'. 'Critical thinking is the use of those cognitive skills or strategies that increase the probability of a desirable outcome. It is used to describe thinking that is purposeful, reasoned, and goal directed – the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions' (Halpern, 2003). Critical thinkers are predisposed to use these skills; they are willing to engage in the hard work of critical thinking. Thus, any consideration of thinking skills needs to be considered along with the disposition or willingness to use those skills.

A Taxonomy of Thinking Skills

There are several schemes for categorizing critical thinking skills (Halpern, 1998). A five-part taxonomy that encompasses many of the skills that have been identified in reviews of the literature (e.g. Jones *et al.*, 1995) uses the following organizing categories: (a) *Verbal reasoning skills*: the skills listed under this rubric include those that are needed to comprehend and defend against the persuasive techniques that are embedded in everyday language (also known as natural language). (b) *Argument analysis skills*: an argument is a set of statements with at least one conclusion and one reason that supports the conclusion. Argument analysis requires the skills of identifying conclusions, rating the quality of reasons, and determining the overall strength of an argument. (c) *Skills in thinking as hypothesis testing*: the rationale for this category is that much of our day-to-day thinking is like the scientific method of hypothesis-testing. In many

of their everyday interactions, people function like intuitive scientists in order to explain, predict, and control the events in their lives. (d) *Using likelihood and uncertainty*: because very few events in life can be known with certainty, the correct use of probability and likelihood plays a critical role in almost every decision. (e) *Decision-making and problem-solving skills*: in some sense, all of the critical thinking skills are used to make decisions and solve problems, but the ones that are included here involve the use of multiple problem statements to define the problem and identify possible goals, the generation and selection of alternatives, and judging among the alternatives.

Taken together, these five categories define an organizational rubric for a 'thinking skills' approach to critical thinking. They have face validity and can be easily communicated to the general public and to students, and offer one possible answer to the question of what to teach when the educational goal is improvement in thinking. Thinking skills are often difficult to articulate because thinking is an abstract concept. Some common distinctions among types of thinking are presented in the following sections.

REASONING

Reasoning skills are one category of thinking skills. The formal discipline of reasoning is closely tied to logic, the branch of philosophy that explicitly states the rules for deriving valid conclusions. According to logic, a conclusion is valid if it necessarily follows from premises or statements that we accept as though they were true. A conclusion that does not follow from its premises is illogical. A distinction is often made between deductive and inductive reasoning.

Deduction and Induction

Deductive reasoning applies to a class of thinking tasks in which the thinking process begins with statements that are known or believed to be true. The task is to determine if a conclusion is valid. There are several types of deductive reasoning tasks, each characterized by the nature of the relationship among the premises. Linear reasoning tasks include concepts that vary along a single dimension, such as height (e.g. 'A is taller than B, and B is taller than C,' therefore A is taller than C'); spatial location (e.g. 'A is to the left of B, B is to the left of C'), and intelligence (e.g. 'A is smarter than B, B is smarter than C'). Syllogistic reasoning tasks include concepts that are related by the inclusiveness of their group membership (e.g. 'All A are B, some B are C'). Conditional reasoning tasks include 'if–then' relationships (e.g. 'If A, then B. B occurred'). In all deductive reasoning tasks, conclusions are always valid if they must be true given that the premises are true.

By contrast, inductive reasoning applies to tasks where observations are collected and generalized to suggest conclusions about the population from which the observations were drawn. For example, if every secretary you have ever met was female, you might inductively conclude that all secretaries are female. Unlike conclusions from deductive reasoning tasks, conclusions from inductive reasoning tasks can never be proven to be true because, in this example, it would take only one secretary who is not female for the conclusion to be false, and unless you have observed every secretary in the population, you cannot be sure that an unobserved secretary would not make the conclusion invalid. Hypothesis-testing skills are conceptually similar to inductive reasoning skills. Conclusions are always probabilistic with inductive reasoning; better conclusions will result from large and representative samples of the population, but as long as the conclusion is based on a sample, there is always some probability that it will be incorrect.

Although it is common to differentiate between deductive and inductive reasoning tasks, the distinction may not be a particularly useful description of how people reason in real life. In everyday contexts, most people switch back and forth from inductive to deductive reasoning in the course of thinking. Our hypotheses and beliefs guide the observations we make, while our observations, in turn, modify our hypotheses and beliefs. Often, this process will involve a continuous interplay of inductive and deductive reasoning. Thinking in real-world contexts almost always involves the use of multiple types of thinking skills.

Are Humans Logical?

When people use the formal rules of logic in formulating and testing conclusions, we label their thinking 'rational'. Do most people use these rules, or are we (mostly) irrational (Stanovich, 1999)? This question can be addressed from both an experimental and a philosophical perspective.

On the one hand, there is an abundance of data that shows that under most everyday circumstances, few people use the rules of formal logic. It seems that in the everyday use of reasoning, few people determine whether a conclusion is valid

solely on the basis of the statements that are given. Instead, they alter the statements according to personal beliefs and then decide whether a conclusion follows from the altered statements. People function under a kind of personal logic in which they utilize their personal beliefs about the world to formulate conclusions about related issues. People tend to use *pragmatic* reasoning rather than the rules of logic: the word 'pragmatic' refers to anything that is practical. In real life, people have a *reason for reasoning*, and sometimes the laws of logic are at odds with the setting, consequences, and commonly agreed upon reasons and rules for deriving conclusions. For example, according to the rules of logic, the sentence 'If you wash the dishes, I will give you $5' also implies that I may give you $5 even if you do not wash the dishes. In most everyday settings, 'If you wash the dishes, I will give you $5' means '*if and only if* you wash the dishes, I will give you $5'. Thus, people use everyday or pragmatic logic, which is not the same as the formal laws of logic.

In real-world settings, most people fail to accept the logical task. In other words, even when they are told to base their conclusions on the premises that are given, they use their prior knowledge and personal belief systems instead of reasoning from only the given premises. In most real-life settings, we would expect that people would use their beliefs and knowledge about the world when reasoning. Often, people are able to use the rules of formal logic, but reject these rules as irrelevant in most real-life settings. Thus, there is a distinction between the ability to reason logically and the practice of reasoning logically.

PROBLEM-SOLVING

Components and Stages

For many psychologists and other scientists, the term 'problem-solving' is synonymous with thinking. A problem exists whenever there is a gap between where problem-solvers are and where they want to be. In understanding how people solve problems, it is useful to think of all problems as composed of different parts – a start state, a goal state, solution paths that lead from the start state to the goal state, blind paths that may appear to connect the start and goal states but, in reality, do not, and the entire problem space, which encompasses all of these parts. The goal of any problem-solving task is to move from the start state to the goal state in an efficient manner. Problems occur in virtually every aspect of life – making one's money last until the next pay check, completing a crossword puzzle, discovering a cure for a dread disease, negotiating a peace settlement, finding a mate, etc. With a thinking skills approach, guidelines or heuristics that apply to the type of problem being solved are tried out as a means of achieving the goal.

Problem-solving often proceeds in a series of stages. In the first stage, the problem-solver becomes familiar with the nature of the problem – understanding the start and goal states and the constraints on reaching the goal. In the second stage, the problem-solver produces several possible solutions. Finally, there is a judgment or evaluation stage when the various solutions are assessed in order to select the best one. Sometimes, when there do not seem to be any readily evident solutions, there is an incubation stage. During this time, the problem-solver seems to forget about the problem and goes on to other activities. The term 'incubation' suggests the image of a mother hen sitting on eggs that will hatch into chicks; similarly, good ideas can 'hatch' from an incubation stage, during which the problem is not being actively considered. There is a large literature of anecdotal evidence suggesting that time away from a difficult problem can sometimes lead to good solutions, perhaps because a 'time out' allows for ineffective mental sets to dissipate. It is difficult to demonstrate incubation effects in the laboratory, which necessarily weakens causal statements about the validity of incubation.

Well-defined and Ill-defined Problems

In real life, most problems are ill-defined. Ill-defined problems have multiple possible goals and the problem itself has to be recognized in the hectic array of the natural environment. For example, someone with a 'money problem' would need to recognize that there is not enough money to pay all the bills and then would have to decide which of many possible goals should be pursued – for example, ask for a raise, get a second job, find a less expensive apartment, move in with a friend, rob a bank, etc. Some of these solutions would cause other problems (e.g. jail time for bank robbery), so each possible solution and goal state needs to be assessed in terms of the new problems that would arise in solving the money-shortage problem. One thinking skill for ill-defined problems is to restate the problem and several different goal states as a way of generating multiple possible solutions. Redefinitions of the desired goals should suggest additional solutions for the problem.

By contrast, the problems that are typically used in formal learning environments and in textbooks are well-defined. Students are given all the information they need to solve the problem, and it is easy to determine whether the problem has been solved. Usually there is only one correct answer. For example, problems in mathematics classes and physics classes are usually well-defined. Well-defined problems are often artificial and have been criticized for giving students a false picture of how problems are identified and solved in out-of-school settings.

MENTAL MODELS

How do we understand complex, abstract topics such as human nutrition, the national debt, or how electricity works? The information that we have about topics like these is organized into interconnected knowledge structures that specify causal, predictive, and covariate relationships, hierarchies of categories, and other relationships that allow us to understand these topics. For example, we may believe that eating vegetables and fruit will make someone healthy (causal relationship) and that a healthy person has energy (covariate relationship) and will live a long life (predictive relationship).

An interconnected web of concepts that we use to make sense of the world is called a *mental model*. We use the mental models that we create to make decisions and to assess claims. Mental models are altered when we encounter new information that has to fit into an existing model (e.g. eggs are part of a healthy diet), and the model has to change when new information exposes an error in the existing model (e.g. the amount of protein eaten needs to be limited). Individuals with expertise in these areas have extensive and accurate mental models. The mental models for complex topics are for most people incomplete, a mix of solid, factual information, some half-truths, and some incorrect information. When we make decisions about what and how much to eat, whether it is safe to eat vegetables that have been irradiated, or whether it is worth the additional cost to buy produce that has been grown without pesticides, we rely on the unique model of human nutrition that we built from our partly solid, partly incomplete, and partly wrong information.

It might seem that mental models would continually be improving as we encounter additional relevant information, but, in fact, beliefs about the world that make sense to the individual are highly resistant to change. Instead, we tend to interpret new information in ways that make the new information fit into our existing models, and to reject information that runs counter to our beliefs. This is why it is so difficult to get people to reassess their positions on controversial topics, such as capital punishment or gun control. We tend to concentrate on information that is consistent with prior beliefs and downplay or ignore contradictory data (Nickerson, 1998).

ANALOGICAL PROCESSES

Analogies are pervasive in human thought. Whenever we focus on the similarity between two objects or ideas or other phenomena, we are using an analogical process. Analogies are always imperfect because they focus on the similarity between objects and events that are not identical. Analogies are frequently used in problem-solving by noticing the way in which two problems that differ in surface features (e.g. radiating a tumor and capturing a fortress) are similar in their deep structure or underlying features (use multiple lines of attack when radiating a tumor or capturing a fortress). When using analogical processes in thinking, we are mapping attributes from a known domain onto a novel domain. Research has shown that the spontaneous use of analogical processes in problem-solving depends on the thinker's ability to recognize when two seemingly dissimilar topics are similar in their underlying structure.

Unlike the sciences, where the rules of evidence require causal experimental designs (with random assignment of participants to groups), the legal profession relies most heavily on analogies to persuade a judge and jury that the conclusion the lawyers favor is the best one. The basic premise of a case law system is that each situation is unique, so no general statements of law can be counted on to settle complex individual cases. Instead, similarities are noted between a case being tried and the facts and circumstances of earlier cases. Thus, lawyers and others in the legal system base their arguments on the degree to which two or more cases can be thought of as similar. Of course, the opposing side in a legal case will attempt to argue that the dissimilarities are more important than the similarities when the earlier case had an outcome that was not desired by the opposing lawyers.

Creativity and Analogies

Creative thinking is a type of thinking in which the outcome is unusual and appropriate (or meaningful or especially good; Halpern, 2003). Creative

outcomes often involve 'noticing' the similarities between two different domains and then borrowing a concept from one domain to apply in the other. For example, in an analysis of the creative genius of Alexander Calder (1898–1976), one author noted that Calder combined the bright colors and abstract art of Piet Mondrian with the jewelry-making skills he learned from his mother to create colorful mobiles that hang from wire, much like a very large earring (Weisberg, 1993). Similarly, a new method for bonding dental material onto teeth was based on an analogy from another wet environment: the suction used by barnacles to attach themselves to piers served as the inspiration for suction-based attachments for teeth. These are two of a very large number of creative outcomes that can be traced directly to a situation that is highly dissimilar in surface information, but essentially similar in its deep structure.

When solving a problem, one heuristic is to think of a problem in nature that is essentially similar to the one you are solving. The underlying idea of this technique is that there are many ingenious solutions in nature that we can use for other, more mundane problems. The idea for Velcro, those two-sided sticky closure tabs that can be found on shoes and many clothing items for children and the disabled, is analogous in design to the way the sticky balls that drop from trees adhere to shoes and clothing when we walk through a forest. The closing for many bottles of glue is similar to a flap-like mechanism that closes the human bowel. When these common natural solutions are applied to problems with similar underlying structure, they are often judged to be creative.

THINK-ALOUD PROTOCOLS

How do we study something as unobservable and illusive as thought? Thought leaves no directly observable traces and can be known or suspected only by inference. One way that researchers study thinking skills is by having thinkers 'think aloud', which means that thinkers are asked to say out loud everything they are thinking as they work through a problem or comprehend a text or oral message. The idea of think-aloud protocols is that they make private activities overt so that thinking can be studied and modified. Unfortunately, it turns out to be very difficult to articulate the thought process. It is as though the act of speaking aloud about the thinking process alters that process. By directing attention to something we usually do without any conscious thought, we change the very activity that we want to study. Very often people cannot say what they are doing when they are thinking.

Verbatim records of think-aloud protocols have shown that many people use images when they think, especially when they are thinking about concrete nouns, which are easy to image. When think-aloud protocols are analyzed, each portion is broken into separate segments and scored as to content, relationship to earlier and later segments, number of hypotheses considered, and the order in which thoughts occur. There has been some attempt to use the protocols of experts as a way of training novices how to think about topics in the expert's domain of knowledge. By making individuals think aloud, they are being forced to evaluate their thinking strategies, with the goal of enhancing communication about the thinking process. Think-aloud protocols are also used in the design of artificial intelligence programs that mimic the strategies used by real humans.

CONCLUSION

There are many varieties of thinking skills, ranging from the formal rules of logic used to assess the validity of a conclusion to the heuristics that are designed to increase the probability of a creative response. There is much interest in understanding and teaching the skills of critical thinking because enhanced thinking skills are needed in our increasingly complex world. Studies have shown that often people can use the rules of formal logic, but choose not to because it would require that they disregard everything they know about the world. Thinking can be studied by having people verbalize everything they are doing when they are working on a problem. Although results obtained with this technique of 'think-aloud protocols' have provided a glimpse into some of the processes used in thinking, it cannot capture the full complexity of human thought.

References

Halpern DF (1998) Teaching critical thinking for transfer across domains: dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist* **53**: 449–455.

Halpern DF (2003) *Thought and Knowledge: An Introduction to Critical Thinking*, 4th edn. Mahwah, NJ: Lawrence Erlbaum.

Jones EA, Hoffman S, Moore LM *et al.* (1995) *National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking. (NCES 95–001.)* Washington, DC: US Government Printing Office.

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology* **2**(2): 175–220.
Stanovich KE (1999) *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.
Weisberg RW (1993) *Creativity: Beyond the Myth of Genius*. New York, NY: Freeman.

Further Reading

Halpern DF (1996) *Thinking Critically about Critical Thinking*. Mahwah, NJ: Lawrence Erlbaum.
Shermer M (2001) *The Borderlands of Science*. New York, NY: Oxford University Press.

Cooperation and Conflict, Large-scale Human

Advanced article

Francisco J Gil-White, University of Pennsylvania, Philadelphia, Pennsylvania, USA
Peter J Richerson, University of California, Davis, California, USA

CONTENTS

Along with a few other animals such as bees, termites, ants and wasps, humans live in societies that cooperate for a common goal. Serving the greater good seems to go against basic Darwinian theory, so the question is, how and why have complex, cooperative societies developed in some species? In spite of so much cooperation, why do human societies remain so conflict ridden?

INTRODUCTION

Suppose you stroll to the corner restaurant for breakfast: eggs, bacon, and a glass of orange juice. A simple activity? No. Mind-numbing complexity is more like it. A farmer in Virginia produced your egg, another in Florida your orange juice, and yet another in the Midwest your bacon. Different truckers brought each of these to a supermarket. The restaurateur then bought them there and had them prepared for you. Seven people are involved in your 'simple' activity? Well, no. This is a caricature. Just for starters, the egg farmer/capitalist hires several workers to operate considerable equipment, all of which was purchased from other companies, made up of capitalists and workers, who in turn bought their parts from yet other companies, which…(the mind reels). Your day has barely begun, and a few dollars' worth of breakfast has already brought an army of considerable size to your service.

Only a select few animal species have societies with extensive cooperation, fine coordination, and massive division of labor: the social insects (bees, termites, ants, and wasps), possibly naked mole rats (but their level of complexity is far below that of the advanced social insects, as is the scale of their societies) and us. This form of social organization is clearly evolutionarily successful; social insects are diverse and abundant, especially in the tropics, and human populations grow so fast that our rapid and energy-expensive expansion into every conceivable niche is a considerable threat to other species and the climate. Given this, one might be naturally inclined to ask: why is this adaptation not more common among the species of the world? After all, beneficial adaptations should proliferate, shouldn't they?

Actually, for modern students of evolutionary theory – trained as they are in the framework of what has been called the 'modern synthesis' and 'neo-Darwinism' – the puzzle is rather different. Modern biologists are trained to be surprised not by the rarity of this dramatic adaptation but by the fact that it is possible at all. How could something this strange evolve? Darwin himself worried that the self-sacrificial altruism of the social insects might be fatal to his theory. To see the problem his way, we must briefly develop the theoretical instincts of a modern evolutionary theorist.

DIFFICULTIES ACCOUNTING FOR HUMAN ULTRA-SOCIALITY IN DARWINIAN TERMS

Individuals attempt to reproduce before they die, and some do better than others. If the features of an individual are passed on to the offspring, then good reproducers will beget good reproducers, who in turn will beget good reproducers once again. And so on. Each time, good reproducers leave more descendants than other types, so after a number of generations the entire population will become of the type that reproduces best (with the

exception of frequency-dependent effects, when selection will maintain several types at equilibrium). This is the basic Darwinian insight of 'natural selection'.

The mechanism responsible for stable similarities between parent and offspring is genetic inheritance. Mere individuals live and die, but genes can potentially keep going forever. Modern Darwinism focuses on changing genes in order to understand the processes responsible for historical change in organic populations.

The analytical focus of a modern Darwinian is the 'gene's eye view' heuristic, which relegates individual organisms to the status of temporary 'vehicles' conveying the potentially immortal genes from one generation to the next. Genes that have a better chance of proliferating are those that increase the reproductive success of their vehicles in competition with other vehicles. Finding 'unselfish' genes that cause their vehicles to suffer sacrifices to benefit another vehicle's reproduction is thus a major puzzle. But nothing in these arguments really depends upon there being a single gene for altruism; this is just a convenient way to strip the problem to its bare essentials. Darwin was right to worry.

This brings us to the social insects. Massive division of labor is impressive, but the reason it is possible in the first place is the truly big puzzle. Although many ants in a colony will famously give up their lives protecting it, for example, this is only because they have already given up their reproduction. The latter is, to a Darwinian, the really dramatic fact. How could they give up their reproduction? In human ultra-sociality, on the other hand, defense is the most dramatic puzzle because those who risk and often give their lives to defend their society are indeed capable of reproducing and by fighting give up some or all of this capacity.

For non-human altruism, twentieth-century evolutionary theory has provided two elegant and very successful explanations: kin selection and reciprocity. Before examining them, notice that the fundamental issue for any explanation in this domain is the problem of assortment. The gene's eye view allows us to state the obvious: since the gene is trapped inside its vehicle, the vehicle must reproduce if the gene is to proliferate. So how can a gene proliferate more than competing genes if it makes its vehicle transfer reproductively useful resources to other vehicles? At first glance this would seem impossible, and for most kinds of resource transfers it will be. But if the vehicle is making resource transfers to other vehicles also containing copies of that same gene, then the gene promotes its own proliferation at one remove.

The question therefore is: what could cause vehicles with altruistic genes to assort with one another?

KIN SELECTION

Green Beards

Imagine a gene – 'G' – producing two effects: (1) it gives you a green beard, and (2) it makes you help those with green beards (Dawkins, 1989, pp. 88–89). G's twin effects solve the problem of assortment: if you help those with green beards, then, because those individuals also have G (hence their beard), G is making you help other copies of itself. Copies of G can 'find each other' thanks to the beards, and therefore when G causes its vehicle to transfer resources to other vehicles it is nevertheless promoting the spread of G.

It is virtually impossible that the same gene will cause a discriminatorily altruistic behavior and also the cue used to discriminate, unless altruism itself is the cue. Theoretical considerations suggest that it is also highly improbably that "green beard" genes can arise as a result of two tightly linked loci where the gene at one locus would code for the green beard, and the gene at the other for the altruistic behavior. But the thought experiment brings the problem of assortment into sharp focus: if an 'altruistic' gene is to prosper, its vehicle must confer benefits disproportionately on other vehicles containing copies of the same gene. Something like a green beard must facilitate this nonrandom assortment for altruistic genes to evolve.

In one proposal (Hirshleifer, 1987; Frank, 1988), if altruism is mediated by emotions, and if emotions result inevitably in facial expressions and other bodily manifestations, and if such manifestations are hard to fake, then altruists can assort with each other by examining each other's expressions of emotion. In other words, those who 'look' altruistic probably are, so altruists can find and prefer each other for mutual benefit. Genes coding for altruistic emotion/displays will be favored.

But the problem with this kind of 'green beard' argument is that, once the signal is common, selection will favor selfish individuals who pretend to be altruists but don't help. Actors and confidence artists can fake emotions well enough to fool us. Darwinians indeed expect that the evolution of clever, green-beard-exploiting sociopaths will undermine the evolution of emotional signals. This theoretical embarrassment to the green beard

argument is accompanied by an empirical one: emotions that appear very similar to ours occur in other mammals (as Darwin himself wrote in his book *The Expression of Emotions in Animals and Man*). Thus, if nonhumans can produce emotions and signal them, why can't they use this to assort for altruism and build ultra-social communities? Human emotions are no doubt involved in motivating and signaling cooperation in humans, but this is likely to be a secondary effect of other evolutionary processes, not something that can be shaped directly by natural selection to favor the original emergence of altruism. If it could, many nonhuman mammals should have it.

Kinship as a 'Green Beard' Substitute

If not emotional green beards, then what? Suppose that if you have the altruistic gene, then you can use an observable cue X to guess with some probability p that somebody else also has the gene. If so, then the altruistic gene – call it gene 'K' – will be helping itself so long as it specifies 'to individuals with cue X give a benefit size b, where b satisfies the following:

$$bp > c,$$

where c is the cost to the altruist of transferring the benefit. In other words, out of a large population of individuals bearing cue X and therefore receiving my help, only a proportion p will actually carry gene K. Thus – on average – the benefit that K's vehicle (me) confers on other copies of K is not b but the scaled down benefit bp. If this weighted payoff is greater than what it costs me to help, then K is giving itself a net benefit.

In 1964 William Hamilton argued persuasively that kinship can play the role of cue X. Consider two siblings, Higley and Bob. Bob carries a gene K that makes him an altruist. What is the probability that Higley also has gene K? Well, Bob's father passed down half of his genes to each sibling, who get the other half from their mother. These samples are subject to independent random assortment, so that Bob and Higley share a quarter of their father's genes and a quarter of their mother's. Thus, the probability that Bob and Higley share gene K is at least $p = \frac{1}{2}$. The probability may be higher if the gene is common in the population, but the critical value is the chance that siblings share the identical gene by common descent. This is the same as the probability of sharing the gene when it is rare. So suppose that K specifies a behavior that makes Bob give 5 units of benefit to siblings like Higley, for a cost to the actor of 2 units. Will K spread? Yes.

$$\frac{5}{2} = 2.5 > 2$$

On the other hand, if at the cost of 2 units K confers only 3 units of benefit to these recipients, then K will not spread.

What have we shown? That if a gene makes its vehicle assist its close kin, then it has found a way for its vehicle to assort (a fair amount of the time) with other vehicles carrying copies of the same gene. This assortment is what makes it possible for altruistic genes – within benefit/cost limitations – to evolve. This is, of course, far short of the perfect assortment that green beards would make possible, but it is what nature uses because green beards or their equivalents are usually impossible. This 'kin selection' argument explains the widespread observation of nepotistic altruism in humans and many other species. In particular, it explains the ultra-sociality of the social insects, for in, say, an ant colony, everybody is a close relative due to the fact that everybody is a child of the queen.

Washburn's Fallacy

The above insight is usually expressed as Hamilton's famous rule: $br > c$. Here r replaces p, and stands for 'coefficient of relatedness': the probability that two individuals have identical copies of the same gene, descended from the same, recent ancestor gene. Thus, recall that for Higley we calculated the probability that he has an identical gene to Bob's that is in fact descended from their father's or mother's copy.

The r in Hamilton's rule is often misinterpreted as 'the probability or proportion of genes shared in common between two individuals'. This is commonly referred to as 'Washburn's fallacy' because the anthropologist Sherwood Washburn used to argue – in critical fashion – that Hamilton's rule would imply altruism towards everybody and only slightly more altruism towards kin. Why? Because any of us shares about 80 percent of our genetic alleles with any other randomly chosen member of the human species, and 80 percent is a lot. If true, this argument would appear to solve the puzzle of human ultra-sociality, but it would create an even bigger puzzle: why aren't many more species ultra-social?

But Washburn's argument follows only if the r is interpreted as the proportion of genes shared in common, rather than as the probability of sharing identical copies of a gene descended from the same, recent ancestor.

Why is Washburn wrong? Even if 80 percent of the people in the population have the altruistic gene (and the others have a selfish alternative), since Washburn's altruistic gene says 'help anybody', having an altruistic gene will not make a vehicle disproportionately likely to get help. The 20 percent of people not sharing the altruistic gene will get the same benefit as the 80 percent that do, and since they don't pay the costs of helping others, they have higher fitness. Selfish genes will increase in frequency and drive out the altruistic genes. An altruistic gene that said 'help close kin', on the other hand, would make altruistic genes disproportionately likely to get help. Individuals with the altruistic gene are more likely than randomly chosen members of the population to have close relatives with copies of this gene, and are therefore more likely to get helped, than individuals with the selfish gene.

Washburn could have avoided his fallacy simply by imagining how his 'help anybody' gene could have become common in the first place. Here things become crystal clear: unless a gene codes for a behavior promoting its spread when it is a new and therefore rare mutation, the gene will wink out of existence as quickly as it appeared. When the 'help anybody' gene first appears virtually no other vehicles have copies of it, so 'helping anybody' confers no benefits on the gene's spread and the gene quickly goes extinct. A new mutant gene is, by definition, rare, and thus only close kin of its vehicle are likely to carry copies. An altruistic gene therefore has a chance of spreading from low frequency only if it discriminates in favor of close kin. Why not distant relatives? When the gene is rare, distant relatives are about as unlikely to have copies of the gene as a randomly chosen member of the population – in fact, at the limit, these are the same, because all members of a population are (very) distant relatives.

Kin selection can explain nepotism in many species, most spectacularly in the case of eusocial ants, bees, and termites, where huge numbers of close relatives cooperate. But in human social systems, even at their most simple, average r is so low that we may well say members are not, in fact, related. As Campbell (1983) rightly observed, human societies, unlike the social insects, exhibit cooperation among reproductive competitors. If kin selection can cause ultra-sociality with human levels of average r, then many more animal species should have such complex societies. Humans are probably a special case requiring a special explanation.

RECIPROCITY

In the logic of reciprocity (first explored by Robert Trivers (1971)) an 'actor' suffers a cost to benefit a 'recipient', expecting a return benefit at some other time (I'll scratch your back if you scratch mine). The time delay distinguishes this from 'trade' as commonly understood, for trade lacks the risk of no payback. Perhaps this explains the unfortunate popularity of Trivers's coinage 'reciprocal altruism', which has caused much confusion. If we stick to the gene's eye view, however, the terminological tangles quickly evaporate. When will a gene specify a transfer of reproductively relevant resources from its own to other vehicles? Kin selection can lead to this, as we have seen. Reciprocity can too, but it differs from kin selection in that, so long as the recipient pays back the favor, it matters little whether the recipient's motivation arises from a gene identical by recent descent (or indeed from some entirely different gene). Reciprocity may even occur between species, as in mutualisms. What matters is that there be some reasonable probability that the favor will be returned and a method for assessing this probability. If favors are made when they are relatively cheap for the actor but beneficial for the recipient, and if they are returned, then a gene making its vehicle do such favors will prosper.

How well will a rare reciprocity gene do? When it is rare, a vehicle carrying the gene is very unlikely to meet another such vehicle that will reciprocate its good turns. Thus, the evolution of reciprocity requires some initial assistance from kin selection. For example, since the individuals carrying a new and rare mutation will be close relatives, vehicles carrying the reciprocity gene will be likely to meet other such vehicles – even when rare – if individuals are organized in local kin groups. Once the gene for reciprocity becomes a little more common, such kin-biased population structure is unnecessary for the success of the reciprocity gene.

Even when reciprocators are common, it is important to ensure assortment to prevent 'cooperators' from being exploited by 'defectors', and this brings us to the question of the cognitive mechanisms involved. Theoretical considerations suggest that nice-but-not-gullible strategies like 'tit for tat' (if you cooperated with me last time, I will cooperate this time; if you didn't, I won't) are at the heart of our reciprocating psychology (Axelrod and Hamilton, 1981; Axelrod, 1984), but the actual mechanisms are complex and subtle.

The logic of reciprocity can easily explain cooperation in very small groups, especially dyads. However, reciprocity cannot so easily explain cooperation in larger groups (Boyd and Richerson, 1988). In a dyad, my help is a private benefit directed to one individual; if the partner does not reciprocate, I can ignore this individual in the future and direct my help towards another who *will* pay back my assistance. But when my benefit is consumed not by one partner but by two or more simultaneously (say, for example, that I build a wall which protects everybody who lives inside of it), the structure of the problem changes. (Notice, by definition, if the benefit is being consumed by a group, this means I cannot selectively withdraw the benefit from nonreciprocators, and am therefore producing a 'public good'. If I can discriminate, then we don't really have a 'group', but are back to dyadic interactions.) When everybody in the group returns my favor we all benefit, but if some don't return my favor, they create a dilemma for me: either (1) I can cooperate, and reward the defector (who gets the benefit without paying the cost of returning my favor); or (2) I can defect, giving up the benefits of reciprocity with those in the group who *are* reciprocators. The larger a group is, the less likely that just by chance it will have disproportionately large numbers of cooperators, so genes supporting (2) will do better with increasing group sizes. In particular, when the gene for reciprocity is new and therefore rare, the chances of having many reciprocators in a large group are vanishingly small. As groups get larger, then, kin selection is less and less effective at helping group-based reciprocity get started. For groups as small as 10, the potential to get group-based reciprocity off the ground becomes very small.

Some have considered the indirect benefits of reciprocity as a possible explanation for human ultra-sociality that sidesteps the public goods problem. Trivers (1971) speculated that given widespread dyadic reciprocity, selection would favor a strategy that used altruism towards third parties as a gauge of trustworthiness. Richard Alexander (1987) argued that the resulting structured webs could solve the problem of reciprocity in large groups. The argument is that humans are smart enough that each individual can keep track of who reciprocates with third parties; a strategy that prefers such reciprocators as partners will do well because it is better at picking low-risk partners. The resulting large webs of 'indirect reciprocity' can build much more complex societies of nonrelatives than in other species.

More recent models (Nowak and Sigmund, 1998a, b) challenge Boyd and Richerson's (1988) conclusion that large-group reciprocity cannot evolve. However, as Leimar and Hammerstein (2001) argue, the Nowak and Sigmund model makes a very unrealistic assumption: interactants never make mistakes. (see also Panchanathan, 2001). They show that when mistakes are allowed to occur, indirect reciprocity does not easily evolve because one needs information about people's intentions, not just their behavior (e.g., did Bob not reciprocate because he was punishing a nonreciprocator or because he himself is a nonreciprocator?). Indeed this is true even of dyadic reciprocity: if people make mistakes, we need to distinguish between honest mistakes and defections, and for that we need a gauge of people's intentions (Sugden, 1986; Boyd, 1989; Boerlijst *et al*., 1997). Panchanathan (2001) concludes that language (in the form of gossip) can furnish people with very good information about the reputations of others, where reputation (based on the person's known record of interactions) works as a gauge of someone's probability of defection. Indirect reciprocity may thus help explain why a language-endowed social mammal was capable of organization on the scale of hunter–gatherer bands, which are larger and considerably more complex than other mammalian societies but small enough that people can keep track of reputation through gossip. Whether indirect reciprocity is a sufficient explanation for organization on the level of tribes, chiefdoms, and states is unclear.

Undoubtedly, dyadic and indirect reciprocity are importantly involved in the evolution of cognitive mechanisms such as guilt and shame, and their associated signals. For example, Fessler (1999) provides a detailed analysis of the situations that elicit shame. The purpose of the emotion/display is apparently to signal one's recognition of having made a 'mistake', with the implication that one is not really challenging the social norms. The importance of signaling contrition is evidence that people care about intentions, not merely behaviors.

Signaling

If large-scale organization does depend on generating public goods altruism, perhaps such behaviors can emerge through signaling. If I benefit from advertising my qualities to others, I will want a signal that cannot be faked by lower-quality competitors. This may explain the provision of expensive public goods as a form of signaling the quality

of one's genes (Smith and Bliege-Bird, 2000). Male hunters, for example, may share difficult to catch prey items with everybody because they index the hunter's skill. Attention-getting sharing thus might ensure a strong broadcast of the 'hunting quality' signal. The benefits to such hunters would be things such as being preferred in the market for mates and greater political leverage.

The first benefit is obvious, as those who make themselves known as good hunters will be perceived, on average, as better providers, and their popularity in the marriage market will allow them to choose the most desirable (e.g., rich, healthy, hardworking, fertile) partners. This translates into healthier and more abundant progeny. The second benefit requires that there be a reason for other people to defer to the political interests of the hunter (and thus entails a form of trade or reciprocity, even if not a straightforward one). Since the prey is being shared collectively, one will not get more meat by deferring to the hunter, so why do it? Hawkes (1990) argues: in order to keep the hunter in the group (although she refers to the benefit that the hunter gets as 'social attention'). But this explanation does not solve the problem of selfishness, it merely places it elsewhere, as Smith and Bliege-Bird (2000) argue. Henrich and Gil-White (2001) suggest a reciprocal altruism hypothesis to explain deference to good hunters: sycophants who defer to the political interests of a hunter are buying access in order better to acquire the very skills the hunter has advertised.

The signaling hypothesis probably explains some altruism. However, it suffers from the same general problems as 'green beard' explanations. Why can't the selfish use the signals of altruists as a cue for whom to exploit selfishly? Why doesn't the signaling of qualities support complex societies in other species? Costly displays of good genes occur in many species, yet in no other species is aid to the group used to signal value as a mate. Emotional commitments to an altruistic moral order no doubt are a proximate explanation for such behaviors, but such emotions in turn have to be explained. The real puzzle is explaining how we came to be equipped with such emotional attachments to norms, and for that we probably need an explanation in terms of group selection generating the emergence of punishment for deviance, as argued below.

CULTURAL GROUP SELECTION

The Problem of Genetic Group Selection

Suppose we have two groups of the same species, one full of individuals with generalized altruism genes, and the other full of individuals with selfish genes. Which gene will do better evolutionarily? The fitness of a gene is equal to the average fitness of the vehicles carrying it, so here an average altruistic vehicle has higher fitness because it is surrounded by other such vehicles (which results in profitable mutual assistance). A selfish vehicle, on the other hand, has relatively lower fitness because it is surrounded by other selfish vehicles.

So the altruistic gene will win? The problem is maintaining sufficient variation between groups for group selection to be a potent force. Two forces erode variation in altruism between groups: the relative success of selfish individuals within groups and the migration of selfish individuals from group to group. Group selection can favor altruistic genes so long as (1) migration is sufficiently low; and (2) the fitness benefits of being in a group of mostly altruists is so large that new groups of altruists which competitively displace selfish groups are generated at a pace fast enough to more than compensate for the dilution of altruists by within-group processes and the arrival of selfish migrants.

Some students of altruism (Sober and Wilson, 1998) like to think of kin selection as a form of group selection in which relatedness creates sufficient variation between groups for group selection to operate. Terminological disputes aside, the kin selection view of groups illustrates the problem with large-scale group selection; if kin groups are reasonably outbred, relatedness falls dramatically with genealogical distance and the evolution of altruism is restricted to close kin. Outbreeding is equivalent to migration into the kin group. Observed rates of migration are generally too large to allow relatedness to build up in large groups, hence making group selection in them implausible. Ever since Williams's (1966) criticism of early attempts to explain adaptations as group selected, many evolutionists reject group selection as a plausible explanation almost as a matter of principle.

A Cultural Solution

Despite the problems with large-scale group selection explanations in outbred organisms, many, starting with Darwin, have speculated that some form of group selection is important in the special case of humans (Sober and Wilson, 1998). Humans certainly do compete as groups, and organized warfare is a spectacular example. But our groups are so porous (e.g., successful groups often induce a flow of mates from less successful ones) that one is brought back to the problem of migration. If some process could minimize the effects of migration – something quintessentially human – this would give us an elegant explanation simultaneously accounting for human ultra-sociality and also for the fact that other animal societies are restricted to forms of altruism derived from kin selection. That something might be culture, defined here as the intergenerationally stable, high fidelity, social transmission of information (socially transmissible packets of information are often referred to as 'memes', after Dawkins (1989, chap. 11)).

Theoretical models show that, given a capacity for acquiring information directly from others (which appears to be uniquely hypertrophied in humans), a bias for conformity will evolve. Conformity is adaptive because it helps individuals pick up useful memes that others have already converged upon (Boyd and Richerson, 1985; Henrich and Boyd, 1998). It is also advantageous to the degree that human societies often involve games of coordination in which direct advantages stem from doing what others do, such as driving on the agreed-upon side of the road (Gil-White, 2001). When in Rome, do as the Romans do. Many psychological studies have documented this cognitive bias (Miller and McFarland, 1991; Kuran, 1995; Asch, 1956, 1963). Conformity reduces the problem of migration (Boyd and Richerson, 1985; Henrich and Boyd, 1998) because when migrants absorb the memes in their host community they tend not to affect the local equilibrium. Rather the local equilibrium tends to convert *them*. Thus, selfish migrants arriving in an altruistic group will – if they are conformists – absorb the local altruistic norms even as their own are discriminated against, thus preserving rather than diluting the altruistic character of the group. This allows cultural group selection to generate new altruistic groups fast enough to overcome the rate at which spontaneous (cultural) mutations of individuals from altruistic to selfish erode altruism within groups (cf. Soltis, Boyd, Richerson, 1995). If cultural group selection operated over sufficiently long periods of time in the late Pleistocene, gene-culture coevolution might have resulted in the evolution of innate predispositions and skills adapted to participation in group selected social units (Richerson and Boyd, 2001).

WITHIN-GROUP COOPERATION AND BETWEEN-GROUP CONFLICTS

A complementary explanation maintains that if a norm for punishing deviations is adhered to by most members of a group, it can stabilize anything, including a norm for altruism (Boyd and Richerson, 1992). If much group competition is active rather than passive (e.g., violent combat for land), then within-group altruistic norms maintained by punishment will confer dramatic advantages. This could make the production of new altruistic groups faster than the processes which dilute altruism within the group (Boyd *et al*., unpublished). The result would be a panhuman selection pressure for cognitive adaptations reducing the likelihood of 'mistakes' in order to avoid costly punishment (prosocial emotions such as duty, patriotism, moral outrage, etc. that commit us to predominant social norms even in the absence of coercion). These could easily form the basis for large-scale ultra-social organization, including dramatic cultural adaptations for collective defense. Such emotions could help explain why humans often engage in altruistic acts even in the absence of monitoring or reputational benefits and why they die anonymously in battlefields.

Clearly, the other side of the coin of group cooperation is group conflict. Groups that develop norms that channel their within-group cooperation towards outward bellicosity will force other groups to develop the same (or better) or become extinct. This process selects for ever stronger forms of within-group cooperation and outward aggression and is likely to be an important force responsible for the creation of ever larger and more complex social human groups.

IDEOLOGY, SYMBOLS, AND INGROUP MARKING

No society can exist without the acquiescence of its members to the roles they must play in the maintenance and reproduction of the social whole. Historically, anthropology and sociology were both centrally interested in the question of the functional organization of individuals into such roles (both disciplines owe much to the pioneering sociology

of Emile Durkheim and pioneering anthropology of Bronislaw Malinowski), but these days the topic itself has fallen out of favor with the rise of 'methodological individualism' and 'rational choice' perspectives that insist on a picture of human nature as driven by selfish, individualistic considerations. Rational choice theorists, however, can account for high-cost altruism, such as soldiers being willing to die in battle, only by including in the concept of self-interest rewards and punishments that are in turn hard to explain on individual selection grounds. A soldier may not fight out of altruistic feelings (though at least a few undoubtedly do). But whatever the personal motives (glory, duty, shame, need for recognition from others, blind respect for authority), his behavior is more likely the result of adhering to a particular ideology, and the emotions which are inculcated as part of it, than a narrow calculation of the relative material costs and benefits to himself in the evolutionist's reproductive fitness sense. (The reader should note that group selected altruism is not saintly self-sacrifice. When the final tally is completed, altruists must do better at reproducing their genes or their culture than those adopting the selfish alternatives. One target of group selection may be systems of reward and punishment, especially culturally transmitted social institutions in the human case, that indeed motivate even the highly self-interested individual to cooperate. A relatively low frequency of altruistic moralistic punishers may be all that is necessary to keep reluctant cooperators cooperating.) If so, this means we must understand the cognitive processes by means of which ideas are acquired through social learning and emotions are attached to them. We must also understand why and how rendering ideas in the forms of reified symbols makes these ideas so attractive. Such work has barely begun.

In the domain of ethnic-group cognition, some first steps are being taken. It appears that the human brain is predisposed to essentialize ethnic and racial groups. One approach argues that essentialized 'human kinds' can be created out of any social category (Hirschfeld, 1996), depending on local cultural and historical circumstances.

Another approach argues that only those categories – such as, say, ethnic groups and castes – that superficially resemble biological species will tend to be essentialized (Gil-White, 2001). The salient resemblances to species categories are (1) normative endogamy; (2) descent-based membership; (3) characteristic marking (in ethnic groups this is outward marking in the form of dress, scarification, etc.); (4) a distinctive local social adaptation (in ethnic groups this is a local norm equilibrium). These surface resemblances fool the brain into thinking that it is looking at a species category, and the essentialism normally applied to biological kinds is activated. Features 1 and 2 are caused by 4 because interaction – especially in marriage – with outsiders who have different coordination norms is costly. A recent model shows that feature 3 also follows from 4 (McElreath *et al.*, in press). The model shows that everybody benefits from broadcasting the community of origin – if such communities differ in their norms – because in this way costly interactions between partners who will likely fail to coordinate properly will be avoided.

It is important to note that the above is insufficient for an explanation of, say, racial conflict. For most of history the antagonistic political units have often not been maximal ethnic units but smaller (e.g., subethnic tribes, clans) or larger (e.g., multiethnic chiefdoms, empires) units. Only with the recent advent of ethnonationalism – an ideology maintaining that political and ethnic boundaries should coincide – do we get a proliferation of conflicts where the antagonistic units are maximal ethnic groups. These conflicts appear especially difficult to contain and negotiate precisely because the groups in conflict perceive themselves as unalterably 'natural' groups. Smaller groups often recognize their 'inherent' similarities with coethnics, and larger ones usually find it impractical to motivate emotional adherence based on belonging to the same imperial system. We are still very far from understanding how and why ideologies such as ethnonationalism spread and remain stable and why they are so easily exportable into vastly different cultures. An understanding of the cognitive processes that make certain ideologies attractive in particular circumstances (i.e., become cultural selection pressures), and which commit us emotionally to such ideologies, is sorely needed.

CONCLUSION

Explanations that don't go beyond the mechanisms responsible for cooperation in nonhuman species fail to account in a satisfactory manner for the vast aggregations of cooperating nonrelatives that constitute human societies. Kin selection and reciprocity arguably need to be complemented by cultural group selection as the main driving force. While some work has been done to elucidate the formal properties of cultural group selection, the task of understanding the cognitive mechanisms that such processes have shaped, and their

interactions, have only begun. As a result, we don't yet have a good theoretical handle on how the social brain creates selection pressures that affect the distribution and maintenance of ideologies central to large-scale human cooperation and conflict. However, we can now at least begin to ask the questions in a Darwinian framework, applied to culture as a system of inheritance in its own right.

References

Asch SE (1956) Studies of independence and conformity: I. Minority of one against a unanimous majority. *Psychological Monographs* **70**: (Whole No. 416).

Asch SE (1963 [1951]) Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow H (ed.) *Groups, Leadership, and Men*. New York, NY: Russell & Russell.

Axelrod R (1984) *The Evolution of Cooperation*. London, UK: Basic Books/HarperCollins.

Axelrod R and Hamilton WD (1981) The evolution of cooperation. *Science* **211**: 1390–1396.

Boerlijst MC, Nowak MA and Sigmund K (1997) The logic of contrition. *Journal of Theoretical Biology* **185**(3): 281–293.

Boyd R (1989) Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology* **136**: 47–56.

Boyd R and Richerson PJ (1985) *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Boyd R and Richerson PJ (1988) The evolution of reciprocity in sizeable groups. *Journal of Theoretical Biology* **132**: 337–356.

Boyd R and Richerson PJ (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* **13**: 171–195.

Campbell DT (1983) Two distinct routes beyond kin selection to ultra-sociality: implications for the humanities and social sciences. In: Bridgeman D (ed.) *The Nature of Prosocial Development: Theories and Strategies*, pp. 11–39. New York, NY: Academic Press.

Dawkins R (1989 [1976]) *The Selfish Gene*, 2nd edn. Oxford and New York, NY: Oxford University Press.

Fessler DMT (1999) Toward an understanding of the universality of second order emotions. In: Hinton AL (ed.) *Biocultural Approaches to the Emotions*. New York, NY: Cambridge University Press.

Frank RH (1988) *Passions Within Reason: the Strategic Role of the Emotions*. New York, NY: WW Norton.

Gil-White FJ (2001) Are ethnic groups biological 'species' to the human brain?: essentialism in our cognition of some social categories. *Current Anthropology* **42**(4): 515–554.

Hawkes K (1990) Why do men hunt?: benefits for risky choices. In: Cashdan E (ed.) *Risk and Uncertainty in Tribal and Peasant Economies*, pp. 145–166. Boulder, CO: Westview Press.

Henrich J and Boyd R (1998) The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior* **19**(4): 215–241.

Henrich J and Gil-White FJ (2001) The evolution of prestige: freely conferred status as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior* **22**: 165–196.

Hirschfeld L (1996) *Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds*. Cambridge, MA: MIT Press.

Hirshleifer J (1987) On the emotions as guarantors of threats. In: Dupré J (ed.) *The Latest on the Best: Essays in Evolution and Optimality*. Cambridge, MA: MIT Press.

Kuran T (1995) *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

McElreath R, Boyd R and Richerson P (In press) Shared norms can lead to the evolution of ethnic markers. *Current Anthropology*.

Leimar O and Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B* **268**: 745–753.

Miller DT and McFarland C (1991) Why social comparison goes awry: the case of pluralistic ignorance. In: Suls J and Ashby T (eds) *Social Comparison: Contemporary Theory and Research*. Hillsdale, NJ: Lawrence Erlbaum.

Nowak MA and Sigmund K (1998a) The dynamics of indirect reciprocity. *Journal of Theoretical Biology* **194**: 561–574.

Nowak MA and Sigmund K (1998b) Evolution of indirect reciprocity by image scoring. *Nature* **393**: 573–577.

Panchanathan K (2001) *The Role of Reputation in the Evolution of Indirect Reciprocity*. Unpublished Master's Thesis, University of California, Los Angeles.

Richerson PJ and Boyd R (2001) The evolution of subjective commitment to groups: a tribal social instincts hypothesis. In: Nesse R (ed.) *Evolution and the Capacity for Commitment*, pp. 186–220. New York, NY: Russell Sage Foundation.

Smith EA and Bliege Bird RL (2000) Turtle hunting and tombstone opening: public generosity as costly signaling. *Evolution and Human Behavior* **21**(4): 245–261.

Sober E and Wilson DS (1998) *Unto Others: the Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sugden R (1986) *The Economics of Rights, Co-operation, and Welfare*. Oxford and New York: Basil Blackwell.

Trivers R (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57.

Williams GC (1966) *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.

Further Reading

Richerson PJ (2001) Built for speed, not for comfort: Darwinian theory and human culture. *History and Philosophy of the Life Sciences* **23**: 423–463.

Nesse RM (2001) *The Evolution of the Capacity for Commitment*. New York, NY: Russell Sage.

Weingart P, Mitchell SD, Richerson PJ and Maasen S (1997) *Human by Nature: Between Biology and the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Wilson DS (2002) *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago, IL: University of Chicago Press.

Cultural Processes: The Latest Major Transition in Evolution

Introductory article

Eörs Szathmáry, Collegium Budapest (Institute for Advanced Study), Budapest, Hungary

CONTENTS

The origin of humans and human society is linked to the appearance of a unique mechanism for cultural inheritance, namely language. Earlier major evolutionary transitions can shed light on the latest transition.

A REVIEW OF THE MAJOR TRANSITIONS IN EVOLUTION

What Are the Major Transitions?

Table 1 lists what have recently been termed the 'major transitions in evolution'. A few remarkable features can be seen. Some major transitions in evolution (e.g., the origin of multicellular organisms or that of social animals) occurred a number of times, whereas others (e.g., the origin of the genetic code, or language) appear to have been unique events. However, one must be cautious about using the word 'unique'. Owing to the lack of 'true' phylogeny of all extinct and extant organisms, one can give it only an operational definition. If all of the extant and fossil species which possess traits arising from a particular transition share a last common ancestor after that transition, then the transition is said to be unique. Obviously it is quite possible that there had been independent 'trials', as it were, but we do not have comparative or fossil evidence for them.

Important Common Features

There are a number of sufficiently common features of the major transitions for them to require special attention, namely the emergence of an evolutionary unit at a higher level from lower-level ones, an increase in complexity, the appearance of a novel inheritance system, and the 'freezing in' of the transition (often there is no way back). The means whereby these features are achieved include local interaction, synergy, contingent irreversibility, and central control. We will look at each of these in turn, but will first briefly consider evolutionary units.

Units of Evolution

A unit of evolution must be capable of *multiplication*, *heredity*, and *variation* (Figure 1). If some hereditary traits influence the likelihood of survival and/or reproduction of the unit, then in *a population* of such units, evolution by natural selection can take place. Note that this definition does not refer to living systems. Many consider that viruses are not alive (e.g., they lack metabolism), but they do evolve. Some computer programs evolve in the electronic environment, and they are not regarded as alive either. In addition, there are items of culture, which are passed on from individual to individual, that also behave as evolutionary units (referred to as 'memes', by analogy with genes; see below). It is important to bear the generality of this definition in mind, as it enables us to apply Darwinian reasoning to nontrivial cases as well. The reference to population is also crucial, as individuals metabolize, reproduce, run, behave, etc., but they do not evolve. Evolution takes place in populations through the generations of evolutionary units.

Emergence of higher-level evolutionary units

The origin of the eukaryotic cell is a good example. Eukaryotic cells are much more complex than bacteria. Humans are also built of such cells. Mitochondria are the cell organelles that serve as the power plant of the cell. They are very simple, tiny structures that look like bacteria. It turns out that this resemblance is not casual, as they are indeed descended from once free-living bacteria. They became captured by an ancestor of our cells

Table 1. The major transitions in evolution

Before	*After*
Replicating molecules	Populations of molecules in protocells
Independently replicating genes	Chromosomes
RNA as gene and enzyme	DNA genes, protein enzymes
Bacterial cells (prokaryotes)	Cells with nuclei and organelles (eukaryotes)
Asexual clones	Sexual populations
Single-celled organisms	Animals, plants, and fungi
Solitary individuals	Colonies with non-reproductive castes
Prelinguistic societies	Human societies with language

Reproduced from Maynard Smith J and Szathmáry E (1999) *The Origins of Life. From the Birth of Life to the Origin of Language*. Oxford: Oxford University Press.

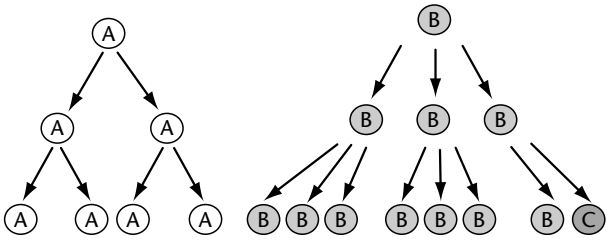


Figure 1. Units of evolution. A and B are reproducing entities of different types. Their average fecundity is shown to be different. Heredity is not exact (there is variability), as one B gives rise to a novel type C. Such units are not necessarily alive.

(around 2 billion years ago) and became enslaved for the production of ATP (the energy-storage molecule of all cells). Obviously, before this transition, the proto-eukaryote and the proto-mitochondrion were two types of unrelated, independently reproducing cell, but now they are integrated into one functional unit.

If such a transition is successful, then *adaptations*, which are discernible at the higher-level unit, evolve that suppress the competitive tendencies of the integrated lower-level units. Essentially, viewed from the lower level, a 'super-organism' is created. However, viewed from the higher level, it is just an organism.

Increase in complexity

Although natural selection does not guarantee that organisms will increase in complexity as they evolve, it is apparent that complexity along certain lineages, such as the one leading to humans, has increased during evolution. Is the number of genes in an organism's genome an appropriate measure of biological complexity? The recent flurry of completed genome sequences, including our own, suggests that this is not necessarily the case. There must be more sensible genomic measures of complexity than the mere number of genes. It is the *regulatory gene interactions* that seem to play a crucial role. In fact, *Drosophila* has more regulatory interactions than *Caenorhabditis elegans*, although mere gene numbers give the reverse order.

Figure 2 illustrates the ways in which genetic complexity can increase in evolution.

Novel inheritance systems

DNA is commonly referred to as the genetic material, with ample justification. However, there are other hereditary mechanisms. A good example is epigenetic inheritance in the cells of multicellular organisms. It is easy to see that something like that must operate in our bodies. Most animals start their lives as a fertilized egg. Cells divide and undergo differentiation in embryogenesis, so muscle, liver, nerve cells, etc. arise that look different and function differently. Some of them remain capable of proliferation. When a healthy liver cell divides, it gives rise to two liver cells – 'liver-cell-ness', as it were, is passed on. Note that the state of being a liver cell was not present in the fertilized egg, but rather it is generated in development. Thus it seems that a characteristic has been acquired and can be inherited at the level of the cell. Just so – this is a Lamarckian dimension of multicellular organisms. However, it extends very rarely from organism to organism during reproduction. To conclude, if a dual inheritance system were not active in us, we simply would not exist.

Obviously language, which is so central to our concern, is also a radically new method of information storage and retrieval. As we shall see, it has a Lamarckian component.

Local interactions

Whenever one develops a theory for a certain transition, one finds that some type of local interaction in the dynamics of the population is important. This can take several forms, and all of them are known to be important for the evolution of *cooperation*. *Reciprocal altruism* can lead to cooperation between unrelated individuals. Because of limited dispersal, cooperating individuals may remain close to each other, or they may remember past interactions with particular individuals. *Kin selection* is a mechanism in which genetic relatedness

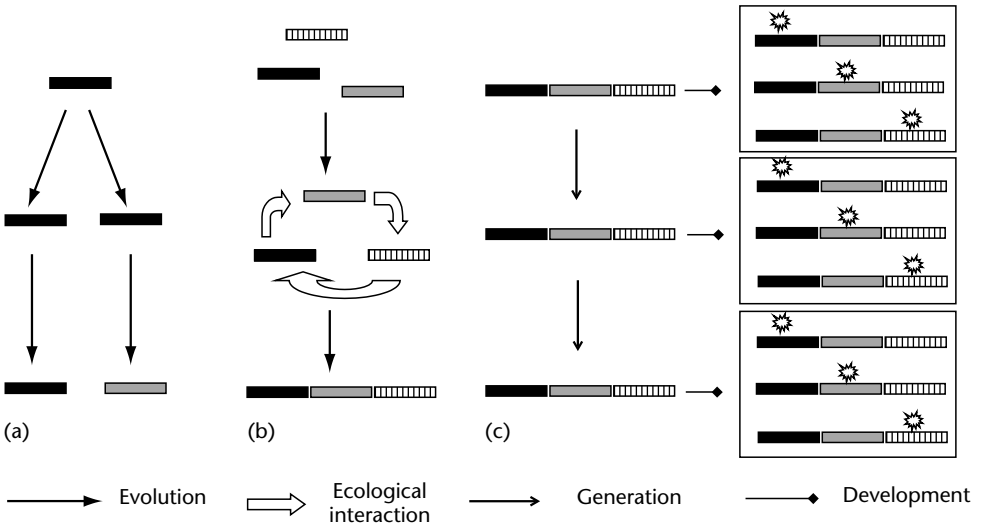


Figure 2. Means of increasing genomic complexity. (a) Duplication and divergence. (b) Symbiosis. First, independently reproducing units engage in an ecological interaction. Finally, the units cannot reproduce alone, and a new evolutionary unit has been formed. (c) Epigenesis. The genes remain the same, but they are differentially activated in the different cells. Modified from Maynard Smith J and Szathmáry E (1995) *The Major Transitions in Evolution*. Oxford: Freeman.

plays a crucial role. Hamilton's rule states that it pays to be an altruist if $br > c$, where b is the benefit received by the helped relative, c is the cost paid for by the helper, and r is the degree of genetic relatedness between them. Finally, *group selection* is a mechanism that applies when not only the individuals, but also groups formed of them, multiply and have heredity and variability. This criterion is readily satisfied when (1) the number of groups is much higher than the number of individuals per group, (2) each group is formed from one parental group only, and (3) there is no migration between groups.

The relative importance of the above mechanisms in accounting for the transitions varies from one case to another, but there is little doubt that all of them have been influential.

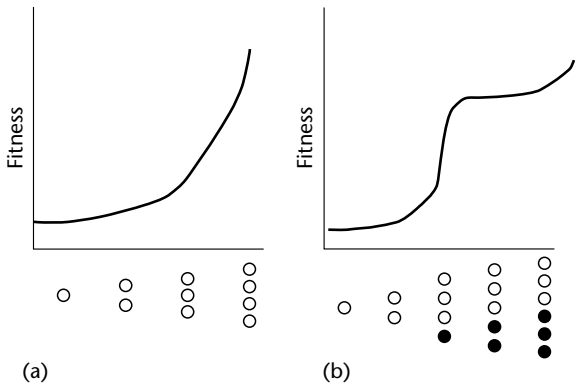


Figure 3. Types of synergy. (a) Quantitative synergy. (b) Qualitative synergy. In the latter case, combination of functions causes a very steep rise in performance, and hence fitness.

Synergy

Synergy can be both quantitative and qualitative (Figure 3). In both cases the performance (efficiency) of the interacting units increases non-linearly. In evolution, this translates into non-additive fitness interactions. In the case of qualitative synergy there are at least two types of interacting unit – they cannot substitute for each other. Cooperative guarding of the young is a good example of quantitative synergy. The interaction between different cell organelles, such as the mitochondrion and the plastid in a plant cell, is an example of qualitative synergy. *Economy of scale* and *combination of functions* are other terms used to refer to quantitative and qualitative synergy, respectively.

Contingent irreversibility

In many cases, once a transition has occurred there seems to be no way back. However, there are exceptions. For example, there are insects whose solitary state is secondary – all of their living relatives are highly social. Yet, in contrast, there is no mitochondrial cancer. This can be understood, given the fact that most mitochondrial genes had been lost in evolution, a fraction had been moved to the cell

nucleus, and very few genes remain in the organelle. Emphatically, all of the genes that are necessary for the division of this organelle have moved to the nucleus, and therefore the latter is in complete control of mitochondrial division. We can thus appreciate contingent irreversibility as a key mechanism for 'locking in' the result of a transition. It is not the case that a reversal would be *logically* impossible, rather it is just far too demanding on the side of the requisite heritable variation – the number of simultaneous, chance genetic changes enabling the reversal is so large (and their joint probability is so small) that *for all practical purposes* we can assume that they will not occur.

Comparative 'Transitionology'

It is a striking feature that some transitions involved related individuals, whereas others involved unrelated ones (Table 2). This gives rise to the important distinction between 'fraternal' and 'egalitarian' types of transition. Kin selection does not work for the latter, but local interactions are crucial.

THE ORIGINS OF ANIMAL SOCIETIES

Animal societies with a complex division of labor between their members have evolved by different routes. Apart from the division of labor, and the economic advantages that result from it, the various types of society have one other feature in common. The existence of non-reproductive castes (the so-called workers) in the social insects and in some other social animals poses a formidable problem to the theory of evolution, as Darwin had already recognized. Why should worker bees give up reproduction? In what sense would this increase their fitness?

Relatedness

Haldane once stated that he was willing to lay down his life to save two brothers or ten cousins. His reason was that these relatives shared, on average, ½ and ⅛ of the genes possessed by him. Why should the proportion of shared genes matter? To answer this question, we have to take a '*gene's eye view*'. A gene that would cause Haldane to die but ten of his cousins to survive would cause more genes identical to itself to survive than would a gene that let Haldane live and his cousins die (in fact, $^{10}/_{8}$ copies of the gene, on average, would survive, compared with only one). In the same way, genes present in worker bees that cause their bearers to give up reproduction in order to rear their sisters can spread, provided that the advantage of cooperative breeding over individual reproduction is great enough. This consideration is expressed elegantly in Hamilton's inequality.

Eusociality

The degree of sociality in different species can be placed on a gradient. Most biologists are interested in what is called eusociality – 'real sociality'. By definition, eusocial animals must satisfy three criteria:

1. reproductive division of labor – that is, only some individuals reproduce;
2. an overlap of generations within the colony;
3. cooperative care of the young produced by the breeding individuals.

Eusociality is well known in ants, bees, wasps, and termites. It is less well known that a similar degree of eusociality can be observed in naked mole rats, spotted hyenas, African wild dogs, and some social spiders.

Table 2. Egalitarian and fraternal major transitions

	Egalitarian	*Fraternal*
Examples	Different molecules in compartments, chromosomes, nucleus and organelles, sex	Same molecules in compartments, organelles in the same cell, cells in individuals, individuals in colonies
Units	Unlike, nonfungible	Like, fungible
Reproductive division of labor	No	Yes
Control of conflicts	Fairness in reproduction, mutual dependence	Kinship
Initial advantage	Combination of functions	Economies of scale
Means of increase in complexity	Symbiosis	Epigenesis
Greatest hurdle	Control of conflicts	Initial advantage

Reproduced from Queller DC (1997) Cooperators since life began. *Quarterly Review of Biology* **72**: 184–188.

Super-organisms

Colonies of social animals can with some justification be regarded as 'super-organisms', in the sense that they display adaptations (traits that increase fitness) at the colony level. For example, the mound built by termites has a system of air channels that function as an air-conditioning system. By this analogy, the queen and the reproductive males are analogous to the germ line of multicellular organisms, and the non-reproductive individuals would be analogous to the soma of the super-organism.

THE ORIGINS OF HUMANS AND HUMAN SOCIETIES

Characteristics of Human Societies

Relatedness and individual recognition

Despite the obvious similarities between a termite mound and a human city, there are profound differences between the mechanisms that lead to cooperation in the two cases. One important feature of human societies, namely the recognition of individuals, already exists in some social mammals and birds. Although insects may recognize group membership, they do not recognize individuals. In contrast, a monkey recognizes other members of its troop as individuals, and behaves differently towards them. As the phrase 'pecking order' implies, the members of a flock of chickens sort themselves into a linear dominance hierarchy, and this probably requires individual recognition. Those who have studied baboons and other monkeys have observed the formation of alliances in which two or more individuals support one another in conflicts with other members of the group. Such alliances may be based on genetic relatedness, but this is not always so. The essential points are (a) that, in higher animals, social interactions within a group depend on individual recognition, and (b) that one individual's behavior towards another depends both on genetic relatedness and on a memory of previous interactions with that individual.

Cultural inheritance

It is often said that the defining characteristic of human societies is cultural inheritance – that is, individuals in a society acquire their beliefs and behavior, and their knowledge and skills, by learning from previous generations, and not by genetic inheritance. There is obviously much truth in this idea, particularly with regard to the differences between one individual and another, or between one society and another. At the level of the individual, differences in political opinions are not caused by differences between genes. Having said this, however, there are some reservations that need to be stated. First, there is some cultural inheritance in animals, a fact that is important when considering the origins of human culture (Table 3). Secondly, the ability of humans to learn, and to build societies that are dependent on cultural transmission, is genetic – human societies differ from chimpanzee societies because humans and chimps differ genetically. Thirdly, humans learn some things more readily than others – the human mind is not a blank slate upon which experience can write what it will.

Social learning

Different types of social learning are at the heart of cultural transmission (Table 4). Young rats can acquire a preference for a new food by smelling it on the coat of other rats. This is a type of cultural inheritance – two groups of rats feed on different foods, and the difference is transmitted by learning. This mechanism has been called *stimulus enhancement*, whereby the adults create an environment in which it is easier for the young to learn. This contrasts with *observational learning*, in which one animal watches what another is doing, and then copies it. It is difficult to believe that all culturally inherited traits in animals depend only on local enhancement. For example, in some areas of

Table 3. Criteria for culture

Invention of a new pattern, or modification of an existing pattern
Transmission from innovator to another
Consistent copying of pattern (often stylized)
Long-term persistence of pattern in the acquirer
Spread of pattern across social units
Pattern enduring across generations

Table 4. Examples of social learning

Type of learning	*Description*
Stimulus enhancement	B learns from A where to orient behavior
Observational learning	B learns to what circumstances a behavior should be a response
Imitation	B learns from A some part of the form of a behavior
Goal emulation	B learns from A the goal of an action

Greece, golden eagles feed mainly on tortoises. The bird is unable to break open the shell with its beak, so it picks up a tortoise, flies up to a considerable height, and then drops the tortoise on to the rocks below, thus breaking the shell. It would be absurd to suggest that in Greece, but nowhere else, this behavior in eagles is genetically programmed. A young bird could learn by local enhancement that tortoise shells contain meat that is good to eat, but how – other than by copying – could they learn to fly upward carrying a tortoise, and then drop it? A second example, involving chimpanzees, is given below.

The distinction between stimulus enhancement and observational learning is important, because only observational learning can lead to cumulative cultural change, which is the characteristic feature of human history. By observational learning, young individuals can learn from adults, but also, if one individual stumbles upon an improved way of doing something, that improvement can be copied. The result is that change can be continuing rather than occasional, and that an individual can learn, by copying, a skill that it could never have learned on its own.

Chimpanzee culture

It is clear that humans depend on observational learning, reinforced by teaching, including verbal instruction. As the above example of golden eagles shows, observational learning is not unknown in animals. There are examples in chimpanzees. Some (but not all) populations of chimps dip sticks into the nests of driver ants, and feed on the ants that crawl up the sticks. The chimps in Gombe use a different technique to those in Tai, and catch about four times as many ants per minute. Local enhancement could explain why some individuals in one population dip for ants, whereas others do not. However, it cannot explain why Tai chimps continue to use an inefficient technique, when there is no reason why they should not adopt a more efficient one. Yet continued use of an inefficient technique is what we would expect if young chimps copy their elders.

Language and culture

If higher animals are at least sometimes able to copy their elders, why is it that continuous cultural change does not occur among them, as it does in humans? Thus one population of chimps may differ from another for cultural reasons, but a given population is not continuously acquiring new habits. The likely explanation is that, in humans, the main mechanism whereby culture is transmitted is language. The nature and origin of language are discussed elsewhere. At the risk of repetition, two conclusions will be reiterated here, namely the close analogy between genetic and linguistic methods of transmitting information, and the implications of linguistics for the modular nature of the human mind.

From Ape to Human

Social groups

All Old World monkeys and apes live in social groups, with the single exception of the orangutan. Figure 4 shows a reconstructed phylogeny, or ancestral tree, of these animals. In Old World monkeys, females remain in the social group in which they were born, whereas males leave their natal group before sexual maturity, and must enter another group in order to breed. They are said to be 'female kin-bonded'. In chimpanzees, the situation is reversed – that is, males remain in their natal groups and females move. Other hominoids vary in their social systems, but none is female kin-bonded. The most parsimonious assumption is that male kin-bonding originated in the common

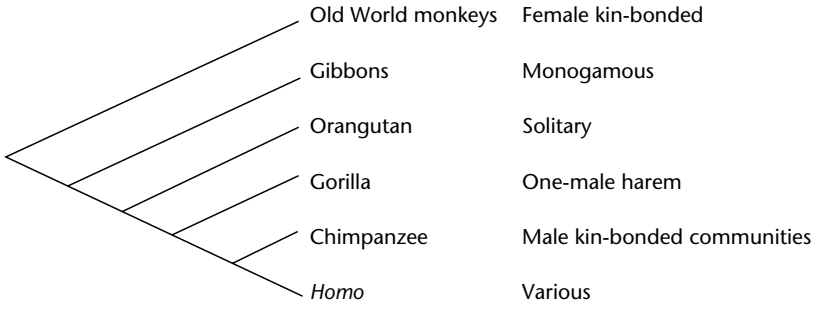


Figure 4. Primate phylogeny and social structure. Reproduced from Foley RA (1996) An evolutionary and chronological framework for human social behaviour. *Proceedings of the British Academy* **88**: 95–117.

ancestor of humans and chimps, since they are more closely related than either is to the gorilla. If this was so, then male kin-bonding is the ancestral condition for hominids. The social systems of modern humans are so varied that it is difficult to be sure whether this conclusion is correct, but it is the best we can do on the basis of the comparative evidence available.

Fossil records

The fossil record provides a second source of information about human origins (Figure 5). It is illuminating to compare this record with what is known of human technical achievements, if only because of the puzzles that the comparison raises. The australopithecines were bipedal and lived in open country. Their relative brain size was only slightly larger than that of the apes, and their tool kit was limited and uninventive. In the lineage from *Australopithecus* through *Homo habilis* and *Homo erectus* there was a gradual increase in brain size, but relatively little technical innovation. The most advanced tool used by *H. erectus* was the handaxe, made from a single block of stone worked on both surfaces, and symmetrical in shape. Such handaxes first appeared around 1.4 million years ago, and persisted almost unchanged for over a million years – hardly an example of cumulative cultural change.

Brain size

Although it seems that there was substantial brain evolution in *H. erectus*, the most rapid increase in relative brain size has occurred during the last

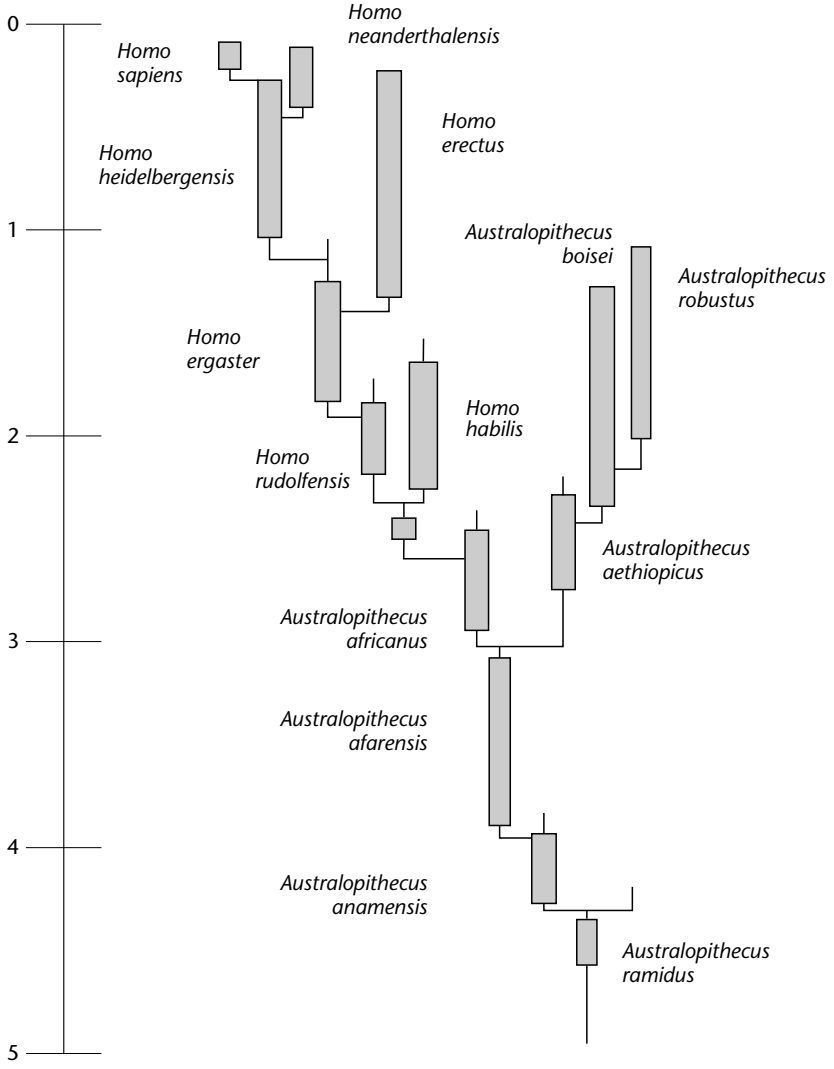


Figure 5. Fossil record of human origins.

300 000 years, culminating in the appearance of effectively modern humans about 100 000 years ago. Yet the acceleration in human technical inventiveness, with the appearance of a varied range of tools made of stone, bone and antler, dates back only 40 000–50 000 years. Burial of the dead, art in the form of cave paintings and musical instruments, personal adornment, and trade originated at much the same time. From around 40 000 years ago we are faced with evidence of continuing cultural innovation. This raises several questions. Why was there a delay of 50 000 years between the appearance of the first anatomically modern humans and the technological revolution? What selective force was responsible for the accelerated increase in brain size 300 000 years ago? When and why did language as we know it originate?

Modules

The problems are difficult, because a fossil skull can tell us relatively little about the brain that was once inside it, and stone tools tell us little about the society that made them. We may try to combine palaeontology, archaeology and psychology in order to obtain a tentative answer to these questions. The essence of this argument is as follows. The human mind does indeed contain modules that are adapted to particular tasks, as suggested by studies of linguistic competence. During much of human evolution these modules increased in efficiency, but they remained to a large degree isolated from one another. Language evolved in the first instance to serve social functions, but once grammatical competence had developed, it provided a means whereby the barriers between modules could be broken down. The burst of creativity during the last 50 000 years resulted from the breaking of these barriers.

Some authors have suggested the existence of three mental modules, concerned with social intelligence, technical intelligence, and natural history, respectively – that is, with the knowledge of animals and plants that is necessary for efficient foraging. We now look at each of these in turn.

Social intelligence

Social intelligence is a common characteristic of the primates. It has been argued that it is the main reason for the increase in brain size in monkeys and apes, as there is a striking correlation between brain size in a species, and the size of social groups in that species.

A crucial question concerns the degree to which apes and monkeys have what has been called a 'theory of mind'. To have a theory of mind is to be able to ascribe to others the possession of a mind like one's own, with similar desires and powers of reasoning. There is no convincing evidence that monkeys have such an ability. For example, a vervet monkey gives a different alarm call if it sees an eagle, a snake or a leopard, but it seems that the monkey does not have in mind the knowledge that another monkey may hear its call and respond appropriately. For example, a monkey may continue to call after all of the others have responded. However, many who have studied the social behavior of chimpanzees, and their skill in manipulation and deceit, are convinced that they do indeed have a theory of mind.

We can conclude that selection for social intelligence was a major cause of the increase in brain size in monkeys, apes, and humans, and that a theory of mind was likely to be present in the common ancestor of chimps and humans, around 5 million years ago.

Technical intelligence

Chimpanzees do use tools in the wild. For example, some populations use stones to crack nuts. However, even in captivity their ability to make tools is very rudimentary. Australopithecines used tools, but there is no convincing evidence for deliberate tool-making, which first appears to be associated with the remains of *H. habilis*, although the tools are little more than irregular chipped stones. *H. erectus* marks a clear advance, with the manufacture of symmetrical handaxes, indicating that the toolmaker had an image of the desired result in mind, and the skill to realize it. However, as was mentioned earlier, there remains an astonishing degree of conservatism. Thus there is evidence of a limited increase in technical intelligence, combined with a lack of inventiveness.

There is also evidence for a degree of independence of social and technical intelligence, even in modern humans. For example, researchers studying autism have suggested that autistic children have impaired understanding of the behavior of other humans (what has been called 'folk psychology') but better than average understanding of the behavior of inanimate objects ('folk physics').

Understanding of natural history

There was obviously selection for improved foraging skills, and hence for knowledge of the distribution and behavior of animals and plants. But was this achieved by an increase in general-purpose intelligence, or by the evolution of a specialized natural history module?

In favor of the latter, it has been argued that all human societies share certain ideas about the living world. First, all living things belong to one – and only one – 'natural kind'. An animal is a dog, or a cat, or a badger, and so on – it must belong to a particular 'species', it cannot belong to two, or to none, and it cannot change its species. Secondly, all human societies share the idea that natural kinds can be classified hierarchically into higher taxa. For example, a dog is a flesh-eater, a mammal (as opposed to a fish, reptile, etc.) and an animal (i.e., not a plant). These universal human attitudes to living things may reflect an innate predisposition.

The alternative is that they could be universally believed because they are true, or almost so, and would be learned by any human society to which knowledge of the living world was important. A second argument in favor of a special natural history module is the speed with which children acquire these beliefs. However, as yet the case for a special module is not decisive.

Coupling

The argument, then, is that the increase in human brain size prior to the emergence of modern humans around 100 000 years ago was associated with an increase in social, technical, and natural history skills, but that these abilities were to a large degree independent. Perhaps the competence in language, including grammar, also evolved during this period, although precise dating is obviously difficult. We may ascribe the cultural explosion that began around 50 000 years ago, and which has led to continuous and cumulative cultural change, to a breakdown of the isolation between mental modules as a result of the emergence of language. The essential point is that once words exist for social, technical, and living things, the same grammar can be used to say things about them.

THE ORIGINS OF LANGUAGE

Writing is much more recent than language. This is really unfortunate, because it is only by writing that language could have become 'fossilized'. We must therefore resort to comparative analyses in biology and linguistics, as well as to building theoretical models. Comparative analysis is also limited, because language is a uniquely human phenomenon. It is undoubtedly an adaptation, and a very complex one at that. Thus it is very unlikely to have arisen as a mere by-product of anything else, without considerable evolutionary fine-tuning by natural selection.

Generative Mechanisms

It is not the aim here to give a thorough description of language with all of its known components. Rather, this section will highlight some important features. First, the number of sounds, words, and grammatical rules that we use in any human language is finite. With these finite means we can potentially cover an indefinitely large domain of grammatically correct, possible sentences.

Symbols

Our vocabulary is finite but open-ended. Without the latter feature, cultural evolution would be impossible. Words are usually highly symbolic signs – they stand at the abstract end of the object–concept–sign triple. Because there is no *immediate* link between object and word, we can have words for purely imaginative concepts, such as *unicorn*.

The capacity for symbolic communication in overt or covert forms seems to be present in some other species as well as humans. For example, bottle-nosed dolphins, chimpanzees, and gray parrots are able to master protolanguage, defined bluntly as word use without grammar. Children under 2 years of age are also roughly at this level when they speak. This already suggests that the hardest nut to crack when contemplating the evolution of the language faculty is syntax.

Grammar

The syntax of every known human language can be characterized by a finite set of grammatical rules. By the application of these rules, all possible grammatically correct sentences can be generated – hence the term *generative grammar*. We are usually unaware of these rules, but we learn our language surprisingly fast, despite the fact that randomly assembled words hardly ever result in grammatical utterances. Our brain seems to be specially tuned, and genetically predisposed, to fast language learning – in short, we have an 'instinct' for language.

Biological Foundations

Genetics

Chimpanzees do not share the language instinct. This difference must ultimately be traced back to genetics. However, genetics of humans is notoriously difficult, because you cannot 'breed' humans as you can, say, breed fruitflies for genetic experiments. Instead, one must identify familial linguistic problems in the existing medical record. Such a

syndrome, or rather a collection of them, has been described and is called *specific language impairment* (SLI). It is specific because, at least for certain individuals, other cognitive deficits are not apparent. Some authors claim that a subgroup of affected individuals has something even more special, known as *grammatical SLI*. It is indeed true that in some such people only aspects of grammar are affected.

In some cases we have the genetic description of this syndrome. In an English-speaking family the problems are associated with a single autosomal dominant allele. It is very unlikely that we shall find many such genes. In contrast, most of the genes that affect the language faculty are likely to be *liability genes* – that is, other things being equal, these genes increase the probability that we will have a normal language competence.

Our remarkable ability to learn a very complex system so fast, and the hints at a specific genetic predisposition, appear to be consistent with the idea of a *language organ* – that is, a genetically determined module in the brain that processes linguistic information. Obviously this organ must be absent from chimpanzees, and must have evolved somehow during the last 5 million years.

Quest for a 'language organ'

However, there is a snag. From what we know of the brain, the language organ cannot be a macroscopically distinct anatomical unit. The fact that neural localization of language can be plastic is now widely known. Studies of brain injury have revealed that damage to the left hemisphere which occurs before a critical period is not lifelong, as the right hemisphere can take over the necessary functions. This does not contradict the finding that in normal individuals Broca's area does appear to be specialized for syntax. It thus seems that the common left-hemisphere localization of language is just the *most likely* outcome when there is no genetic or epigenetic disturbance. Non-invasive brain studies have revealed a truly shocking feature of language development, namely the localization of linguistic processing shifts during normal ontogenesis. The outcome in 'normal' individuals is also highly variable.

The following conclusions can be drawn from brain studies.

- Localization of language is not entirely genetically determined, as even large injuries can be tolerated before a critical period.
- Language localization to certain brain areas is a highly plastic process, both in its development and in its end result.
- It does seem that a surprisingly large part of the brain can sustain language. There are (traditionally recognized) areas that seem to be most commonly associated with language, but they are by no means exclusive, either at the individual level or at the population level, during either normal or impaired ontogenesis.
- Whereas a large part of the human brain can sustain language, no such region exists in apes.

Evolution

It simply cannot be the case that the language faculty originated by a single crucial mutation. This also suggests that linguistic performance itself has evolved. Unfortunately, we can be only hypothetical at the moment. Figure 6 illustrates a possible scenario. Note that, according to this flowchart, components of language have co-evolved and – by inference – the underlying neural networks must also have co-evolved. A much deeper understanding of the brain during processing of linguistic information will be needed before we can establish whether this scenario – or any alternative – is nearly correct or not.

THE USES OF LANGUAGE

Language and Society

It is impossible to imagine our society without language. The society in which we live, day and night, depends on it. Even as we sleep, information about us is being stored and maybe processed. Imagine that we apply for a job on the other side of the world. We are confident, or at least we hope, that our application will be fairly treated, and that the country to which we would like to move is running properly – that is, that social contracts are observed. Our lives depend on the social division of labor and on detailed social contracts which could not exist without language. No ape or dolphin could comprehend, even in spoken form, a contract for a job.

Both the genetic and linguistic systems are able to transmit an infinitely large number of messages by the linear sequence of a small number of distinct units. In genetics, the sequence of four bases enables the specification of a large number of proteins, which in turn, by their interactions, can specify an indefinitely large number of morphologies. In language, the sequence of some 30 to 40 distinct unit sounds, or phonemes, specifies a large number of words, and the arrangement of these words in grammatical sentences can convey an infinitely large number of meanings.

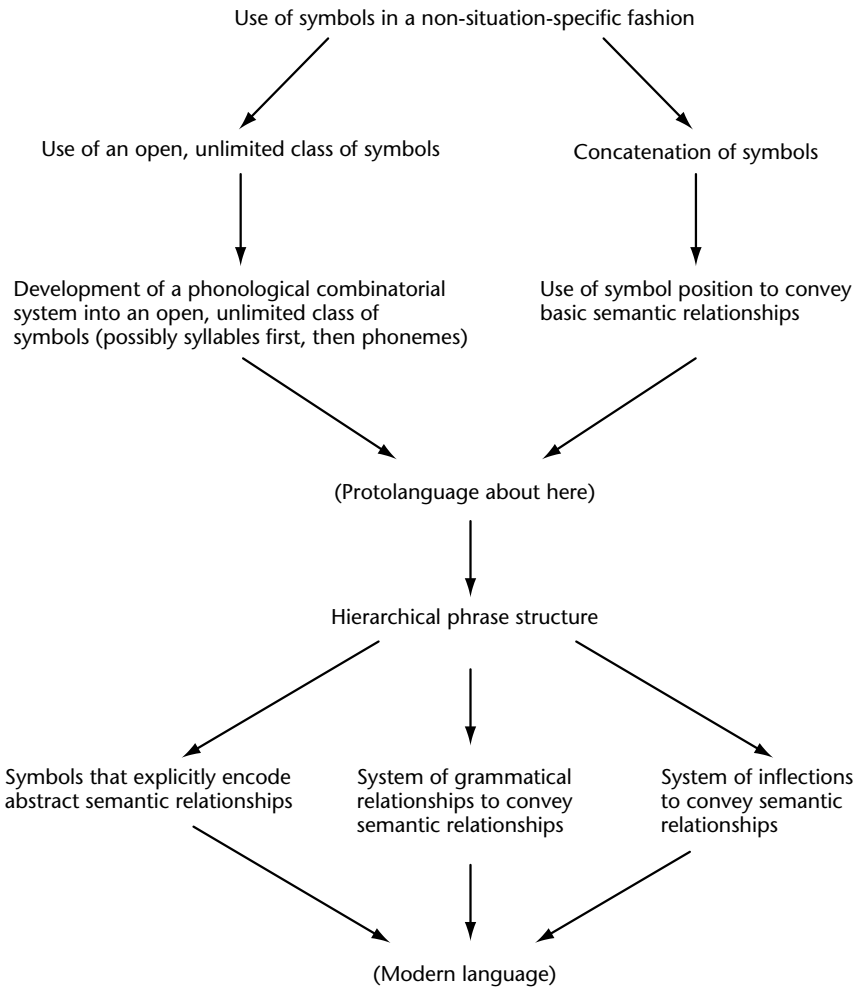


Figure 6. Scenario for the evolution of language. Reproduced from Jackendoff R (1999) Possible stages in the evolution of the human language capacity. *Trends in Cognitive Sciences* **3**: 272–279.

Language for Internal Representation

Language is not only for communication – it is also a means for powerful internal representation. If a person says to him- or herself 'I saw two leopards climbing up that tree; only one has come down, so it's better for me to stay away from that tree until the other one comes down, too', then this reasoning requires syntax that is functionally equivalent to syntax of natural language, even if it is not explicitly 'spoken' in one's head (although frequently it is). Some authors call such an internal language, that serves internal representation, *mentalese*. This emphasizes the idea that language, but not necessarily the spoken one, is a tool for thought as well.

Memes and Cultural Evolution

Richard Dawkins has emphasized the analogy between genetic and linguistic transmission, introducing the concept of a 'meme' – that is, the unit of cultural inheritance analogous to a gene. A meme, he argues, is a replicator. If we invent and tell you a limerick, you may tell it to your friends, and they may tell it to theirs. In this way a single original entity – the representation of the limerick in my brain – has replicated, as a gene might replicate. Clearly, there is room for selection. For example, if we invent a funny limerick, it is more likely to replicate than if we invent a boring one. Of course, whether a meme will replicate or fail depends on the nature of the human mind, and on the cultural milieu (i.e., on the other memes present in the population). However, the same is true of a gene – its increase depends on the environment and on what other genes are present.

There are of course differences. Genes are transmitted from parent to offspring, whereas memes can be transmitted horizontally, or even from offspring to parent. Yet there is a deeper difference

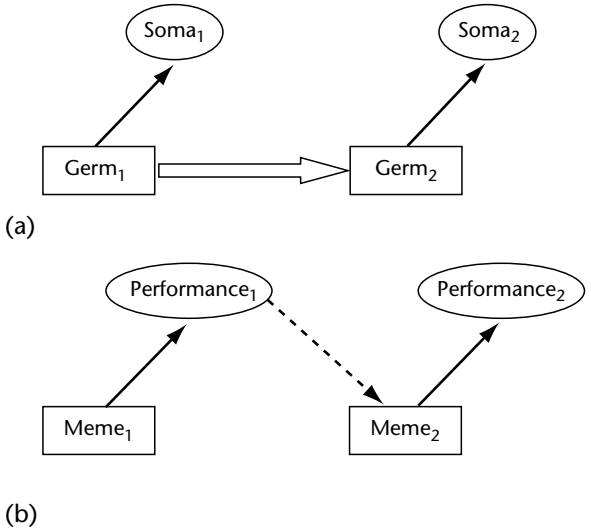


Figure 7. Transmission of genes and memes. (a) A so-called Weismann diagram, named after the famous late-nineteenth-century theoretical biologist. It expresses the idea that only information in the gametes is passed on, whereas the body (soma) cannot transmit its information by genetic means. Weismann thought that this also precludes the inheritance of acquired characteristics (Lamarckism). Although not strictly true, other formulations analogous to this diagram are more accurate and convey the message originally intended by Weismann. (b) Memes are passed on from the performance level ('phenotype'). Language itself is built up analogously from generation to generation. Reproduced from Szathmáry E (2002) Units of evolution and units of life. In: Pályi G, Zucchi L and Caglioti (eds) *Fundamentals of Life*, pp. 181–195. Paris, France: Elsevier.

between genes and memes. Genes specify structures or behaviors (i.e. phenotypes) during development. In inheritance, the phenotype dies and only the genotype is transmitted. The transmission of memes is quite different. A meme is in effect a phenotype – the analogue of the genotype is the neural structure in the brain that specifies that meme. When I tell you a limerick, it is the phenotype that is transmitted – I do not pass you a piece of my brain (Figure 7). It follows that, in the inheritance of memes but not that of genes, acquired characters can be inherited. If I tell you a limerick and you think of an improvement to it, you can incorporate it before you pass it on. In this sense, cultural inheritance is Lamarckian.

For these reasons, population genetic theory cannot readily be applied to cultural inheritance. However, the analogy between memes and genes can be suggestive in a qualitative sense if not in a quantitative one. Furthermore, although for the sake of simplicity we have illustrated the idea of a meme by the example of a limerick, it can refer to more important examples, such as a belief in the Trinity, or a knowledge of how to manufacture gunpowder.

Further Reading

Jackendoff R (1994) *Patterns in the Mind: Language and Human Nature*. New York, NY: Basic Books.

Jackendoff R (1999) Possible stages in the evolution of the human language capacity. *Trends in Cognitive Sciences* **3**: 272–279.

Leakey R (1995) *The Origin of Humankind*. New York, NY: Basic Books.

McGrew WC (1998) Culture in nonhuman primates? *Annual Review of Anthropology* **27**: 301–328.

Maynard Smith J and Szathmáry E (1995) *The Major Transitions in Evolution*. Oxford, UK: Freeman.

Maynard Smith J and Szathmáry E (1999) *The Origins of Life. From the Birth of Life to the Origin of Language*. Oxford, UK: Oxford University Press.

Mithen S (1996) *The Prehistory of the Mind: A Search for the Origins of Art, Religion and Science*. London, UK: Thames & Hudson.

Pinker S (1994) *The Language Instinct*. New York, UK: William Morrow.

Runciman WG, Maynard Smith J and Dunbar RIM (eds) (1996) *Evolution of Social Behaviour Patterns in Primates and Man*. Oxford, UK: Oxford University Press.

Cultural Transmission and Diffusion

Advanced article

Robert Aunger, University of Cambridge, Cambridge, UK

CONTENTS

Cultural diffusion is the process by which information is disseminated through a population, typically by information exchanges among population members (cultural transmission), presumably with the involvement of social learning mechanisms.

INTRODUCTION

People rely heavily on shared beliefs and values to coordinate their social activities. Indeed, phenomena from ritualized greetings to the elaborate ceremonies that mark social events such as marriage can partly be explained by referring to the need to fulfill shared behavioral expectations. These expectations derive from norms and standards specific to each social group. The set of behavioral practices specific to a group constitute a pool of information which is typically what people mean by 'culture'. In anthropology textbooks, for example, culture is typically defined as a 'system of shared beliefs, values, customs, behaviours, and artifacts that the members of society use to cope with their world and with one another, and that are transmitted from generation to generation through learning' (Bates and Plog, 1990, p. 7). How these kinds of information and practices come to be shared by different people in the first place remains to be explained, however, even though it is arguably a central question in the social sciences.

CULTURAL TRANSMISSION PROCESSES

One explanation for shared culture is the exchange of information through social learning. However, considerable controversy attends this apparently common-sense notion. In particular, the recently burgeoning field of evolutionary psychology has claimed that much of what appears to be learned from others is in fact information already in place in people's brains – put there by a long history of natural selection for the retention of that information, which remains ready to be elicited by circumstances. In effect, some ecological trigger sets off an appropriate innate response, making so-called 'culture' a store of mental records – just like the way a jukebox stores musical records, any one of which can be chosen by punching the appropriate button (Tooby and Cosmides, 1992). The fact that people in different places do different things therefore cannot be taken as evidence that culture is transmitted, since they may be simply responding as individuals to subtle environmental differences. The crucial issue, then, is to distinguish the social transmission of information from individual phenotypic plasticity.

Advocates of the position that human beings and other social animals in fact depend significantly on social learning typically suppose that the need to acquire up-to-date information derives from the need to deal with quickly changing ecological conditions – including, in particular, the sometimes ephemeral nature of relationships with other organisms. Further, large brains had to evolve to support the ability to engage in such learning. Evolutionary psychologists counter, however, that big brains are necessary simply to store the many 'cultural' rules that might be elicited by varying circumstances. How can this deadlock be broken?

It seems likely that brains are not big enough to contain all the information required by the jukebox analogy. Many of the problems that contemporary urban life throws up, for example, could not yet have been incorporated into a genetic response: such fundamental experiences as living in large groups of unrelated people are simply too novel. Evolutionary psychologists acknowledge that

modern conditions are likely to spawn maladaptive responses from 'Stone Age' minds. But the fact is that modern culture on the whole seems to be highly adaptive because it has massively extended the niche in which humans can live, and tremendously increased the total population of our species. It seems unlikely that brain mechanisms selected when humans lived under different circumstances would lead to the adoption of behavioral traits that currently contribute so greatly to genetic fitness.

Further, there are reasons to expect that evolution would naturally settle on social learning as an optimizing strategy for the acquisition of information relevant to changing environmental circumstances (Boyd and Richerson, 1985). So it seems reasonable to suppose that at least some of our behaviors are informed by rules acquired from other agents. Social learning should particularly be favored when the lessons from individual trial and error are expensive (either cognitively or in fitness terms), such as determining what is edible when a species can live in a variety of habitats.

Biology therefore determines our general capacity for cultural learning and is responsible for universal abilities like language. However, cultural variations among peoples are attributable to learnt traditions and not to innate or genetic propensities. At the same time, specific psychological adaptations have probably evolved to foster the selective but accurate acquisition of rules through social learning. Language itself can be seen as such an adaptation; it generates signals that allow the reliable transmission of complex, highly contextualized rules for behavior between people.

However, questions remain as to the nature of cultural transmission. In particular, is it like a copying process, or is it a reconstructive one? Some argue that even though cultural information must make its way from person to person in the coded form of messages, this process nevertheless results in a high-fidelity duplicate being produced in the receiver's brain – much as if it had been faxed or photocopied – thanks to evolved mechanisms for social communication. On the other hand, studies of social interactions suggest that the receiver – given the relative paucity of information that actually passes between people – must reconstruct a considerable proportion of a message's content. The nature of the process that underlies human communication has implications for the nature of human psychology. However, if reconstruction and fax-like transmission are equally reliable, then they may exhibit the same population-level dynamics.

CULTURAL SELECTION PROCESSES

Not all social messages are equally attended to or adopted by their receivers. In effect, selection among messages occurs. A selection process requires a population of entities whose frequencies increase or decrease according to their relative fitnesses. Cultural selection can be distinguished from natural selection by the kinds of units on which it operates (Cavalli-Sforza and Feldman, 1981). Just as natural selection is supposed to influence the evolutionary fate of genes, cultural selection works on the prevalence of cultural traits over time.

Cultural selection can occur at each point in the process of communicating information from one individual to another. At the source, there can be psychological selection among potential messages. Once a message has been sent into its channel, physical selection pressures can also affect the chances of that signal reaching its destination. For example, sometimes the message's code does not match well with the modality in which it is sent. Then, after the signal has been detected by the receiver, further psychological biases can exist for attending to and adopting the idea expressed by the signal. Some kinds of information acquired through social learning might not be consistent with other beliefs that an individual holds, for example, and would be rejected for that reason. Later, for further transmission to occur, performances of the related behaviors must also be motivated. Thus, certain traits are favored in effect by the social or physical circumstances in which they find themselves.

Analysts of cultural evolution must be careful to distinguish cultural selection from natural selection, since both can affect the frequency of cultural traits. Some beliefs, for example, cause people to engage in behavior that is detrimental to their health, survival, or likelihood of reproducing. Belonging to a religious group that forbids engaging in sex is only the most obvious case of such an effect. In this form of natural selection, the frequency of a culturally acquired belief is reduced not by changes in belief but by the culling of hosts with such beliefs from the population. This makes cultural evolution a process of 'dual inheritance', in which cultural selection and natural selection operate in parallel or in opposition on cultural traits (Boyd and Richerson, 1985).

Where do the mechanisms of cultural selection come from? How do they in turn evolve? The values used to discriminate between incoming messages can themselves be the product of earlier social learning, or constitute biases produced by a

history of natural selection for discriminating between useful and harmful stimuli. Thus, some cultural traits can 'feather the nest' for later-arriving ones, suggesting that cultural evolution can engage in positive or negative feedback processes.

CULTURAL TRAITS

What is the unit of analysis in studies of culture? Traits – segregating particles of culture – or cultures themselves, taken as a whole?

Throughout its long history, anthropology has attempted to deal with the problem of identifying cultural traits. More recently, however, this attempt has largely been abandoned as impossible – and in any event unnecessary. According to this view, cultures are to be considered as unified wholes, or at least complexes, which are not necessarily divisible. Many ethnographers would argue that there is no 'atomic level' to culture, no way to uncover mutually exclusive entities with stable properties from which cultural compounds are formed. The tendency is now to describe culture as an ideal type, an artificial conglomeration of knowledge compiled from different people occupying varying social roles. However, this metaphysical Platonism – seeing culture as an integrated whole that transcends the minds of individuals – is analytically barren, since there is no contesting a representation built up by the imagination of the ethnographer.

Many would say that we have no legitimate basis for postulating that cultural transmission is intrinsically particulate. There may be no discrete variant that is reliably learned from others through observation or any other form of social interaction. Categories are imposed on a blended reality. You cannot count up cultural values in people's heads like votes for political parties. Traits are 'clumps' of culture content, not well-bounded entities (Gatewood, 2000). However, if culture is learned, and by individuals one at a time, then no one learns everything; culture winds up being distributed. Since learning a culture takes place through social interactions, only parts of culture can be acquired at any given moment: in effect, there must be units of transmission. Admittedly, after acquisition, these units may become amalgamated into complex mental representations which become difficult to tease apart. Nevertheless, these units of acquired information – the equivalent of the atom in physics, the molecule in chemistry, or the phoneme in linguistics – are the smallest possible meaningful unit of cultural information.

So culture must be analyzed as a set of traits, but these need not correlate with natural categories of things in the minds of those living in that culture. At the empirical level, an important question concerns the way in which one is supposed to pick out cultural traits. Where do the division points for defining categories come from – from 'outside' (researchers) or 'inside' (from the group members themselves)? Either is possible. Traits can be just some item of cultural content that the analyst finds it convenient to label – in effect, they become ethnographic conventions. Alternatively, much hard work can be devoted to ascertaining how expert informants themselves classify things.

Once the categories have been established, the question of how these traits are related to one another then arises. Computer scientists, who have thought hard about this problem, have come up with a variety of frameworks for representing knowledge bases, from nested hierarchies to network structures of nodes for concepts with links between them identifying kinds of relationships. What sort of representation is best? This may again be an empirical question, the answer to which depends on the analytical questions being addressed by a particular study.

What, then, is the locus of culture? What is the culture-bearing unit? The traditional solution of simply using the classification people apply to themselves retains considerable appeal, since it has the subjective authority of the participants in the study. However, the actual group influenced by some cultural process or knowledgeable about some domain of belief or practice may vary from domain to domain. This suggests that there is a fluid boundary to the social group to be identified as sharing cultural traits. At best, one might be able to identify a network of people who tend to be linked together for a wide range of cultural traits; but in the end, no easy solution to this problem has appeared.

ARTIFACTS

Human beings, along with many other species, engage in behaviors that significantly alter the environments in which they find themselves. This activity has been called 'niche construction' (Laland *et al*., 2000). Its importance from an evolutionary viewpoint is that such alterations can subsequently influence the kinds of selection pressures that any species interacting with that modified feature of the environment will experience. As a result, population dynamics and the course of evolution can change. Niche construction introduces a feedback loop between behavior and its physical products, artifacts.

The production of artifacts requires 'technique', or knowledge and skills specific to that production process, typically acquired through social learning as well as individual practice. Technology is then the combination of artifacts and technique in a manufacturing context. A technology plus its supporting procedures and institutions, as well as environmental and sociopolitical conditions, can be considered a technological system.

Artifacts themselves can be divided into two general classes: tools and machines. Tools such as hammers or rulers can originate as ideas which are then turned into physical objects by people (or machines) as the expression of that representation. Alternatively, the first exemplar of a tool type might be produced by accident while manipulating some object, or through modification of an existing tool. In any case, it is an object that skilled people can use to extend their physical or mental capacities. Machines, on the other hand, can be considered artifacts with multiple component parts that together perform a novel function which any of the parts, taken individually, would not be able to accomplish. Only humans produce machines, probably because the planning involved in the execution of the multistep process of machine construction is beyond the cognitive abilities of other species.

How do such complex objects evolve? By playing an important part in cultural evolution. The existence of artifacts in the environment can have a significant influence on the course of cultural evolution, because they can constitute a store of information or a channel for information transmission between people (e.g. telephone wires). Artifacts, as transformed objects, should be distinguished from signals and tokens. A signal can be considered to be a patterned particle stream flowing through a channel. A token consists of a physical substrate with information inscribed as pattern in or on it; it is a template for the generation of signals. A signal is 'natural' if it is directly produced by the body (e.g. speech), or 'artificial' if produced by machines (e.g. laser beams or internet 'packets'). It is worth distinguishing between signals and tokens because they have different evolutionary roles: signals are short-lived, dynamic and energetic, designed for the transport of information. Tokens, on the other hand, are static and inert, because they are meant to be secure stores of information. Communicative artifacts contain a token, and hence are able to facilitate the communication of ideas. This means that artifacts can act as signal templates because signals can, after contact with some token, exhibit a new pattern (e.g. amplitude or frequency), to reflect the fact that it now carries information about that artifact.

This kind of communication – mediated by artifacts – has many benefits over traditional face-to-face communication: people separated by distance and time can exchange information (e.g. through email); the effective population reached by a given message can be increased (e.g. through mass media); new communication channels can have novel effects on message receivers (e.g. 'spamming' in email); and greater control over the distribution of messages can be achieved (e.g. through 'gatekeepers' such as the media corporations). The net effect is that this progress in the design of more and more complex artifacts allows the information available to human groups to accumulate beyond that available to any other species (Tomasello, 1999). It is the highly constructed nature of the human niche that arguably makes our cultural adaptation unique.

MEMES, MEMETICS, AND ASSOCIATED CONTROVERSIES

'Meme' is a word coined by Richard Dawkins (1976) to identify cultural traits with the ability to replicate themselves. The word has gained sufficient currency to be included in recent editions of the *Oxford English Dictionary*, where it is defined as 'an element of a culture that may be considered to be passed on by non-genetic means, especially imitation.' This reliance on imitation as a special form of social learning has proved crucial to the definition of memes, since many assert that only the relatively quick copying of instructions (not just behavior, but the directives for instigating that behavior) can support the chains of replication events necessary to sustain an evolutionary lineage, and particularly the curiously cumulative quality of human culture.

What distinguishes memetics from other evolutionary approaches to the understanding of cultural evolution (such as evolutionary psychology or sociobiology) is the insistence that what is transmitted during social learning is a replicator. Although the concept of replication is itself in need of some theoretical work, it generally is a process in which a copy is made of some source. This process must involve causation – that is, the source must be causally involved in the production of the copy (a causation condition). Second, the copy must be like its source in relevant respects (a similarity condition). Third, the process that generates the copy must obtain the information that makes the copy similar to its source from that same source

(an information transfer condition). Finally, during the process, one entity must give rise to two or more (a duplication condition).

Thus, for a memeticist, not only are cultural traits identifiable as individualized units, but they have the power to cause their own duplication. Further, memes are cultural traits that behave as if they were interested in preserving themselves, using individual minds as hosts. Memes, in this view, are responsible for the persistence of certain cultural traits, including those that do not directly favor the biological fitness of the group in which those traits spread themselves.

It was originally thought that memes could encompass behaviors, artifacts or mental contents as varied as 'tunes, ideas, catch-phrases, clothes fashions, ways of making pots or building arches' (Dawkins, 1976). It seems unlikely, however, that replication could occur in a similar fashion in all of these contexts. In fact, behaviors generally do not replicate in the sense described above. It is true that one person can mimic the speech and even the accent of another, reproducing spoken phrases perfectly. Why is this not an example of cultural replication? The answer is that the spoken signal either dies 'in the air' prior to a second signal being constructed in response by the imitator; or it is converted into another form which circulates through the receiving brain before emerging again as the same spoken phrase. In either case, the signal is not duplicated; it either lives through a complex cycle of exchange between people, or a second example is produced after some lag. Thus, despite the facile appearance of a signal being replicated, behavioral mimicry does not hold up to our criteria for replication.

Similarly, the processes through which most artifacts are produced fail the information transfer condition. For example, on the factory floor, it is seldom the case that features of one car on the assembly line are determined by reference to those same characteristics in the previous exemplar. Instead, the assembler (a person or robot) relies on instructions from a centralized database such as a computer-derived datasheet to tell it what to do next. If one agrees with the majority of memeticists that behaviors and artifacts should be eliminated from consideration, then memes are restricted to minds, where the conditions for replication may hold.

As noted above, this means that imitation is the replication mechanism identified by memeticists. It has even been suggested that the human ability to imitate makes the existence of memes self-evident. The argument is that if imitation is defined as the ability to copy behavior or ideas by observing them, then surely the product of that observation must be replicated information. Although transmission is a process in which information is duplicated, it does not follow that the information itself is responsible. Other factors – such as common mental mechanisms for inferring mental or cultural content from sensory stimuli – could lead to the same result. The existence of memes therefore remains unproven (Aunger, 2002).

Neither is there any theoretical reason to suppose that just because culture evolves it must be founded on an independent replicator. Even though biological evolution is grounded in the replication of a biological entity (genes), this does not mean that every evolutionary process must be. In fact, evolution is a more general process than that. Why this is so requires some explanation.

The central concept in evolutionary theory, arguably, is that of a population. It is populations that evolve, and this evolution consists of shifts in the relative frequencies of various types of traits within the population. This population can consist of any collection of things. Simply divide these things up according to type. If these types have tendencies to increase or decrease in the local environment owing to the presence of instances of the same type, this is their relative fitness. If these types reproduce, that makes the process more 'biological', but it is not necessary. We only need to be able to establish objective criteria for determining what type of object something is, and ascertain it has fitness, to declare it subject to an evolutionary process. The replicator approach, however, makes additional demands – in particular, that changes in frequency result from the essential differentiating features of the type being copied from old tokens into the new ones.

Fitness, from this perspective, is just the tendency of a type to increase or decrease, given a specifiable suite of conditions in the local environment. Selection is change in the frequency of types owing to their relative fitness, and variation is change in a type's frequency independent of the type's fitness. Conspicuously missing from these definitions are the notions of replication, heredity (or copying fidelity) and the ancestor–descendant relation. This implies that we can understand evolution without invoking the concepts of either an evolutionary lineage, or replicators. Neither concept is a prerequisite for evolutionary theory to be fruitfully applied to some phenomena (Harms, 1996).

The natural conclusion, therefore, is that culture can evolve without memes. Support for memetics

must come in the form of empirical studies that identify replicators at work in cultural reproduction. Until that happens, we can remain convinced that culture evolves, despite having no firm idea about the mechanisms underlying the reappearance of cultural traits in each generation.

CONCLUSION

How people come to share similar mental contents remains an essential problem for social science to solve. Does culture provide big-brained creatures with a system of informational inheritance which operates independently of, but in parallel with, genes? Alternatively, are supposedly cultural responses induced by mental algorithms that simply reflect a cumulative history of natural selection for genetically directed behaviors? Unfortunately, it will require additional research to find a conclusive answer to this question. However, the fact that cultural change exhibits the basic characteristics of an evolutionary process – inheritance, variation, and selection – seems less open to debate. Whether there will be overarching rules for describing cultural transmission processes remains to be determined, and it seems likely that who learns what from whom will turn out to be specific not only to particular groups, but also to particular periods within the life history of each group – perhaps even to each kind of trait being considered. The goal of establishing general rules of cultural transmission and selection applicable to everyone everywhere remains elusive, making the study of cultural evolution intrinsically complex.

References

Aunger R (2002) *The Electric Meme: A New Theory of How we Think and Communicate*. New York, NY: Free Press.

Bates DG and Plog F (1990) *Cultural Anthropology*. New York, NY: McGraw-Hill.

Boyd R and Richerson PJ (1985) *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Cavalli-Sforza LL and Feldman MW (1981) *Cultural Transmission and Evolution*. Princeton, NJ: Princeton University Press.

Dawkins R (1976) *The Selfish Gene*. Oxford, UK: Oxford University Press.

Gatewood JB (2000) Reflections on the nature of cultural distributions and the units of culture problem. *Ethnology* **35**: 293–303.

Harms W (1996) Cultural evolution and the variable phenotype. *Biology and Philosophy* **11**: 357–375.

Laland KN, Odling-Smee J and Feldman MW (2000) Niche construction, biological evolution and cultural change. *Behavioural and Brain Sciences* **23**: 131–75.

Tomasello M (1999) *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tooby J and Cosmides L (1992) The psychological foundations of culture. In: Barkow JH, Cosmides L and Tooby J (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 19–136. Oxford, UK: Oxford University Press.

Further Reading

Aunger R (ed.) (2001) *Darwinizing Culture: The Status of Memetics as a Science*. Oxford, UK: Oxford University Press.

Blackmore S (1999) *The Meme Machine*. Oxford, UK: Oxford University Press.

Durham W (1991) *Coevolution: Genes, Culture and Human Diversity*. Stanford, CA: Stanford University Press.

Lumsden CJ and Wilson EO (1981) *Genes, Mind and Culture*. Cambridge, MA: Harvard University Press.

Plotkin HC (1993) *Darwin Machines and the Nature of Knowledge*. London, UK: Penguin.

Evolutionary Psychology: Applications and Criticisms

Introductory article

Aaron Sell, University of California, Santa Barbara, California, USA
Edward H Hagen, Humboldt University, Berlin, Germany
Leda Cosmides, University of California, Santa Barbara, California, USA
John Tooby, University of California, Santa Barbara, California, USA

CONTENTS

Theories from evolutionary biology have many implications for research in the cognitive sciences. Evolutionary psychologists have been using these theories to guide their research, the goal of which is to map the evolved, species-typical cognitive and neural architecture of humans (and other species).

INTRODUCTION

Evolutionary psychology (EP) is a paradigm that can be applied to any issue in psychology, rather than a subfield built around the study of a single topic such as vision, social psychology, or child development. As a result, evolutionary theories have opened up many previously unexplored research areas to investigation. Moreover, topics that have already been explored from other perspectives can be advanced empirically and theoretically by adding an evolutionary perspective to the mix. This is because few (if any) of the mechanisms that make up the human mind/brain have been completely mapped. Evolutionary analyses of the adaptive functions that a mechanism evolved to perform usually provide specific hypotheses about its as yet unmapped and undetected design features, prompting further discoveries. Even if the mechanism under study were fully mapped, a correct theory of its adaptive function would still be needed to explain how it came to exist, and why it has the computational design that it does; and to identify which of its components are design features (i.e., functional components), which are incidental byproducts of the mechanism's functional design, and which are evolutionary accidents. In this way, EP is playing the central role in transforming psychology from a largely atheoretical collection of findings to a discipline with principled explanations for why the components of the mind/brain have the designs that they do.

APPLICATIONS OF EVOLUTIONARY PSYCHOLOGY

Heuristic Role of Evolutionarily Derived Predictions

A feature that distinguishes evolutionary psychology from other approaches is that researchers have principled theoretical reasons for their hypotheses derived from evolutionary biology, paleoanthropology, game theory, and hunter–gatherer studies. Such theoretically derived hypotheses allow researchers to devise experiments that make possible the detection and mapping of mechanisms that no one would otherwise have thought to test for in the absence of such theories. Evolutionary psychologists argue that this practice allows a far more efficient research strategy than experiments designed and conducted without reference to the likely functions of the brain. A key insight that emerges from integrating psychology with evolutionary biology is that the mechanisms that psychologists study are adaptations – mechanisms that acquired their organization because that arrangement solved adaptive problems for our ancestors (i.e., had an adaptive function). This insight links the study of psychological mechanisms to theories of adaptive function developed in evolutionary biology. This, in turn, allows a large number of predictions to be derived about the design of human information-processing mechanisms from the large pre-existing body of theories already developed and empirically tested within modern evolutionary biology.

Using this new research program, many theoretically motivated discoveries have been made concerning, for example, internal representations of trajectories; social reasoning specializations; the frequency format of probabilistic reasoning representations; the decision rules governing risk aversion and its absence; universal mate selection criteria and standards of beauty; eye direction detection and its relationship to understanding others' mental states; principles of generalization; life history shifts in aggression and parenting decisions; social memory; reasoning about groups and coalitions; the organization of jealousy, and scores of other topics. Several examples are discussed in more detail below.

Social Exchange and Cheater Detection

Sometimes known as reciprocal altruism, social exchange is an 'I'll scratch your back if you scratch mine' principle: *X* provides a benefit to *Y* conditional on *Y* doing something that *X* wants. This mutual provisioning of benefits, each conditional on the other's compliance, is rare in the animal kingdom: some species (e.g., humans, vampire bats, chimpanzees, baboons) have the cognitive machinery necessary to engage in this behavior, whereas others do not. Robert Trivers, W. D. Hamilton, Robert Axelrod and other evolutionary researchers used game theory to understand the conditions under which social exchange can and cannot evolve. For adaptations causing this form of cooperation to evolve and persist, cooperators must have mechanisms that perform certain specific tasks. For example, reciprocation cannot evolve if the organism lacks reasoning procedures that can effectively detect cheaters (those who take conditionally offered benefits without providing the promised return). Such individuals would be open to exploitation, and hence selected out. Based on such analyses, Leda Cosmides and John Tooby hypothesized that the human neurocognitive architecture includes social contract algorithms: a set of circuits that were specialized by natural selection for solving the intricate computational problems inherent in adaptively engaging in social exchange behavior, including a subroutine for cheater detection.

Because conditionally delivered behavior requires conditional reasoning for its regulation, Cosmides and Tooby used the Wason selection task, an experimental protocol developed to study conditional reasoning, in order to test for the presence of social contract algorithms and their predicted properties. The Wason selection task asks subjects to look for violations of a conditional rule (*If P then Q*), such as 'If a person eats hot chili peppers, then he will drink a cold beer'.

A conditional rule is violated whenever *P* happens but *Q* does not happen (in this case, whenever someone ate hot chili but did not drink cold beer). In the Wason task, the subject is given incomplete information about four people (in this case, one ate hot chili peppers (*P*), one ate broccoli (*not-P*), one drank cold beer (*Q*), one drank hot tea (*not-Q*)). To respond correctly, the subject would need to investigate the person who ate chili and the person who drank hot tea (i.e., *P* and *not-Q*). Yet studies in many nations have shown that reasoning performance on descriptive rules like this is low: only 5–30 percent of people give the logically correct answer, even when the rule involves familiar terms drawn from everyday life.

To show that people who ordinarily cannot detect violations of conditional rules can do so easily when the rule expresses a social contract and a violation represents cheating would be (initial) evidence that the mind has reasoning procedures specialized for detecting cheaters.

Evolutionary psychologists found just that pattern: people who ordinarily cannot detect violations of if-then rules can do so easily and accurately when that violation represents cheating in a situation of social exchange. Given a rule of the general form, 'If you take benefit B, then you must satisfy requirement R' (e.g., 'If you borrow my car, then fill up the tank with gas'), people will point to the person who accepted the benefit and the person who did not satisfy the requirement – the individuals who represent potential cheaters. The adaptively correct answer is immediately obvious to almost all subjects, who commonly experience a pop-out effect. No formal training is needed. Whenever the content of a problem asks one to look for cheaters in a social exchange, subjects experience the problem as simple to solve, and their performance jumps dramatically. In general, 65–80 percent of subjects get it right, the highest performance found for a task of this kind. Further experiments showed that this does not occur because social contracts activate logical reasoning, but because they activate a differently patterned, specialized logic of social exchange. On social exchange problems when formal logic (i.e., the propositional calculus) and social exchange logic predict different answers, subjects overwhelmingly follow the evolved logic of social exchange.

Many cognitive scientists have now investigated social contract reasoning, and many of the predicted design features have been tested for and

found. For example, the mind's automatically deployed definition of cheating is tied to the perspective one has adopted, for the reasoning enhancement to occur, the violations must potentially reveal cheaters; if detecting violations of social contracts reveals only innocent mistakes, enhancement does not occur. Perhaps the strongest evidence that there is a neural specialization designed for cheater detection is the discovery that cheater detection can be selectively impaired by brain damage, without impairing other reasoning abilities. If social contract reasoning were a byproduct of a more general ability to reason, one could not lose the specific ability without also suffering damage to the general ability supposedly responsible for it. Consistent with its being a species-typical ability, social contract reasoning effects are found across cultures, from industrial democracies to hunter–horticulturalist groups in the Ecuadorian Amazon. Most surprisingly, people are just as good at detecting cheaters on culturally unfamiliar or imaginary social contracts as they are for ones that are completely familiar, providing a challenge for any counter-hypothesis resting on a general-learning skill acquisition account. (*See* **Reasoning**)

Foraging and Sex Differences in Spatial Ability

Before 1992 there were hundreds of studies published on sex differences in spatial abilities. Not one showed a replicable female advantage in spatial abilities, and many showed a male advantage.

While evolutionary psychologists do not assume that male and female psychological architectures must be identical in all respects, they do think that whatever differences exist will reflect the different distributions of tasks faced ancestrally by men and women (when they do not reflect differential treatment during development). If there was stronger selection for ancestral women for some spatial tasks, and stronger selection for ancestral men for others, this would produce female superiority in some tasks, and male superiority in others. Starting from an adaptationist perspective, Irwin Silverman and Marion Eals asked a question no spatial researcher had ever asked before: what kind of spatial cognition is required to be good at gathering plant foods? Finding plant foods is a predominantly female activity in foraging populations, and it poses different spatial problems than hunting. Unlike animals, plants do not change location. They do, however, develop over time: a vine, herb, bush, or tree yielding nothing edible now will bear ripe and edible fruit, nuts, tubers, or leaves later in the year. To be an efficient forager, therefore, one must be good at encoding and remembering the locations of thousands of different plants within a complex spatial array. Ideally, this information should be learned incidentally, as one goes about other activities.

Silverman and Eals designed spatial tests that could assess this ability. Some of these tests involved pictures of objects in a complex array, others involved objects in a room. Regardless of format, women consistently recalled more objects than men did. More critically, however, women were more accurate than men at recalling the locations of these objects. This held even controlling for the fact that women recalled more objects: given that an object was recalled, women's location memory was more accurate. This female advantage was large – about as large as the male advantage in tests of mental rotation – and it was found even in incidental learning paradigms.

A century of evolutionarily agnostic approaches to spatial cognition had failed to find a female advantage for any spatial task. But the first time evolutionary psychologists asked what kind of spatial problems ancestral women would have had to solve to forage efficiently, they were able to discover a new spatial ability that shows a large female advantage. As a paradigm, EP allows one to ask more specific questions about the computational design of cognitive sex differences without these being euphemisms for the age-old question of which sex is 'better'. With an adaptationist focus, questions of better or worse quickly disappear and are replaced with questions about the possible functions of a trait, design features predicted by those functions, and empirical data that confirm or falsify confirm those design features. (*See* **Spatial Cognition, Psychology of**)

Coalitional Psychology: Is Race Encoding a Reversible Byproduct of Coalition Encoding?

Ingroup favoritism paired with outgroup indifference or hostility exists in all human cultures. Field and laboratory studies have shown that this behavior is easy to elicit: the simple act of categorizing individuals into two social groups predisposes humans to discriminate in favor of their ingroup and against the outgroup in both allocation of resources and evaluation of conduct. This occurs even when subjects are assigned to groups temporarily and anonymously by an experimenter who used dimensions that are trivial, previously

without social significance, and random with respect to any real characteristics of the individuals assigned. Given that categorizing people into groups along nearly any dimension elicits discrimination, it would be discouraging to find that the human mind cannot help but categorize people on the basis of their race.

Yet, social psychologists had reluctantly concluded that the human mind has circuits that automatically encode (notice and remember) the race of each individual we encounter, as a normal part of impression formation. This conclusion was based on years of experiments in which researchers had searched in vain for ways to weaken the tendency for subjects to categorize others by race. However, the idea that automatic racial categorization is an evolved feature of the human mind is implausible from an evolutionary point of view. Our hunter–gatherer ancestors would rarely – if ever – have encountered a person of a different race, so natural selection could not have favored brain mechanisms designed to notice and remember a non-existent dimension of ancestral social life. This line of reasoning implies that race encoding is a side-effect of a mechanism designed to detect something else that was important for our ancestors.

Accordingly, Robert Kurzban and colleagues proposed and tested an alternative hypothesis: that the (apparently) automatic and mandatory encoding of race is instead a byproduct of brain mechanisms that evolved for an alternative function that was a regular part of the lives of our foraging ancestors: detecting coalitions and alliances. Hunter–gatherers lived in bands, and neighboring bands frequently came into conflict with one another. Similarly, there were coalitions and alliances within bands, a pattern found in related primate species and likely to be more ancient than the hominid line. Mechanisms designed to track these shifting alliances would have benefited our ancestors by helping them to predict the likely social consequences of alternative courses of action.

Brain mechanisms for detecting coalitions should identify patterns of coordinated action, cooperation, and competition. But behaviors that reveal who is allied with whom are rare; to allow judgments even when such events are not in process, coalition encoding mechanisms should note and boost the saliency of any perceptually available marker that is correlated with coalitional alliance. Otherwise arbitrary cues – such as accent, skin color, or manner of dress – should pick up significance only insofar as they acquire predictive validity for coalitional membership. In societies that are not completely racially integrated, shared appearance – a highly visible and always present cue – may be correlated with patterns of association, cooperation, and competition. Under these conditions, coalition detectors may perceive (or misperceive) race-based social alliances, and the mind will map 'race' onto the cognitive variable *coalition*.

Using the same unobtrusive measures that had led social psychologists to believe race encoding is intractable, Kurzban *et al*. showed that race encoding is instead a reversible byproduct of coalition encoding. By creating a social context in which race was not predictive of a cooperative alliance, evolutionary psychologists were able to drastically decrease the extent to which subjects encoded race. Even a few minutes' exposure to a world in which race no longer predicted alliance was enough to substantially deflate the tendency to notice and remember another's race. The experiments also confirmed predictions about the design features of the coalition detection system. Without being asked to do so, subjects nevertheless spontaneously grouped persons into coalitions, noticing and remembering who was affiliated with whom. They did so even in the absence of common appearance, simply on the basis of expressions of mutual ingroup support and outgroup emnity. Other conditions with appearance cues confirmed that the mind appears designed to pick up any perceptual marker, however arbitrary, that is correlated with patterns of cooperation and alliance. The same marker is ignored when it is not correlated with coalition. Humans appear to have an evolved coalition detection system.

Relevance of Evolutionary Psychology to Psychiatric Syndromes

Evolutionary psychologists argue that the paradigm provides fresh insight into psychiatric syndromes. According to Jerome Wakefield, the (implicit) consensus view in medicine is that disorders are conceptually defined as harmful dysfunctions. That is, to qualify as a disorder, a syndrome must consist of damage to one or more adaptations, and people must make the value judgment that these effects are harmful. (For example someone may suffer damage to their adaptations for incidentally learning plant locations or coalitional affiliations but neither notice nor care.)

This places the study of adaptations – computational ones in the case of mental illnesses, noncomputational ones in the case of other illnesses – at the center of medicine. Autism, for example, appears to

involve a breakdown in the psychological mechanisms that allow one individual to interpret the mental states of others. Some syndromes, such as postpartum depression, gestational diabetes, and the food aversions of pregnancy sickness, may not be disorders at all, but the expression of adaptations. Others – such as phobias to snakes and spiders – may result from a dysfunctional overactivation of an adaptive system designed to reduce poisonous bites among our foraging ancestors. Yet others may result from a mismatch between modern environments and the ancestral world our mechanisms evolved to operate in. For example, fat storage as a buffer against occasional famine – adaptive for a hunter–gatherer – may lead to obesity in calorie-abundant environments.

Example: postpartum depression

One fact about the past that researchers know with certainty is that raising offspring was, for both parents, very costly in terms of time and energy. These costs are particularly high when a baby is born. Nursing alone requires approximately an additional 500 calories per day above and beyond what the mother needs for herself. Our ancestors had to forage for all their food, and studies of hunter–gatherers and other primates indicate that food would often have been scarce. A mother in an ancestral foraging society would have benefited greatly from food and other assistance provided by the father and other family members, especially if she also had other children.

Robert Trivers pointed out that investment of scarce resources in one offspring decreases the investment that can be made in other offspring (or future offspring). In the unforgiving world of our ancestors, attempting to raise a newborn without sufficient food and help from others would sometimes have led to the death of the mother or her other children, or might have damaged her ability to have any future children. In the environment of evolutionary adaptedness, mothers who reduced or eliminated investment in offspring that were unlikely to survive and reproduce due to poor health or insufficient parental resources would have been able to increase their investment in other offspring that were more likely to survive and reproduce, and thus would have had a greater number of total descendents than mothers who invested in all offspring equally.

Martin Daly, Margo Wilson, Janet Mann, and Edward Hagen have argued that our ancestors would have been under heavy selection pressure to evolve a motivational system that evaluated the availability of resources, social support, and the health of the newborn to regulate the sentiments in a new mother that decide whether or not to expend her scarce resources on the newborn. Because in humans, evolved decision-making processes are often experienced as emotions, evolutionary psychologists predict that mothers with sufficient resources, social support, and a healthy baby will feel positive emotions towards the newborn, and will therefore invest in it. In contrast, mothers with low levels of resources, social support, or a very unhealthy baby may experience negative emotions towards the newborn (and a corresponding lack of positive emotions), and will therefore reduce their investment in it, saving this investment for other (or future) offspring that are more likely to survive and reproduce.

Postpartum depression (PPD), which is characterized by sustained low mood, sadness, loss of interest, and other symptoms, is suffered by about 10 percent of all mothers, and has been universally regarded as a mental illness, one probably caused as a byproduct of abnormal hormone fluctuations. A close look at PPD from an evolutionary perspective, however, suggests that it is not a disorder, because it exhibits a series of design features that would have been too functional in the world of our ancestors to have appeared by accident. PPD operates in a manner consistent with what would be expected in an adaptation whose function is to reduce investment in the newborn when there are insufficient resources and social support, or when the infant has serious health problems suggesting that it would not survive and reproduce in a foraging world. Studies by Hagen and others have confirmed that PPD is associated with lack of social support, lack of resources, and infant health problems, and that mothers with PPD reduce their investment in their newborns. Many studies have failed to find any association between PPD and abnormal hormone fluctuations, and the conclusion that PPD is not caused by abnormal hormonal fluctuations is strengthened by the fact that fathers (who, unlike mothers, are not experiencing dramatic hormonal changes) also suffer PPD. Even though mothers in industrialized countries typically have enough food, they nevertheless inherit adaptations designed to be sensitive to the availability of social support, a resource that would have been critical to successfully raising a child in ancestral environments. When this essential resource is in short supply, it would have been evolutionarily adaptive for mothers to lower their interest in their newborn.

Because this is a new theory, it cannot form the basis for clinical interventions in PPD cases without

considerable further testing. If this theory turns out to be true, however, it suggests that treatment of PPD should not consist of antidepressants alone, but should instead involve the judicious use of antidepressants in concert with providing the mother with what she needs: more social support and resources. (*See* **Depression**)

CRITICISMS OF EVOLUTIONARY PSYCHOLOGY

Storytelling

Some critics, such as Stephen Jay Gould, have argued that the field consists of post-hoc storytelling ('just-so stories'). It is difficult to reconcile such claims with the actual practice of EP, since in evolutionary psychology the evolutionary model or prediction typically precedes and causes the discovery of new facts, rather than being constructed post hoc to fit some known fact. Of course, all scientific fields from physics to geology also entertain new explanations for already known facts, and evolutionary psychology is no exception. However, new theories for already known facts almost always make new predictions that other theories do not, allowing them to be tested against each other. For example, race encoding was a known phenomenon, but the theory that it was a byproduct of coalitional specializations led to new predictions. Indeed, the value of any new scientific paradigm is measured by its ability to explain findings that other paradigms could not explain, by its ability to explain large sets of facts economically, and by its ability to make novel and significant predictions that are subsequently confirmed by observation.

Though evolutionary psychology is a young science with relatively few practitioners, its researchers have (1) predicted, tested for, and found, a large number of previously unknown mechanisms, whose existence was predicted in advance by evolutionary theory; (2) advanced theories that explain previously unexplained phenomena; (3) advanced theories that provide economical and unified explanations for previously unconnected findings.

Falsifiability

Another common assertion has been that evolutionary hypotheses are unfalsifiable, a claim that is sometimes justified by arguing that the evolutionary past cannot be observed. Evolutionary psychologists reply that there are two components to their theories: (1) hypotheses about the psychological architecture of modern humans; and (2) hypotheses about the ancestral selection pressures that designed the architecture.

Saying a mechanism evolved by natural selection to solve a particular function makes specific and testable predictions about the present design features of that mechanism. (For example humans will exhibit an enhanced ability to reason about cheaters, which can be selectively impaired while other reasoning abilities stay intact. When tested, women will exhibit an advantage in incidentally learning the locations of things.)

Only the second component (ancestral selection pressures) involves phenomena that are difficult to observe directly. In analyzing ancestral selection pressures, EP draws heavily on knowledge about ancestral environments, and critics have plausibly maintained that reconstructions of the past (which routinely go on in astronomy, geology, physics, and biology) are inherently speculative. This would be a problem if researchers had to have complete information about all aspects of the ancestral world to make progress. However, while some features of the ancestral world are difficult to investigate, researchers know with certainty or high confidence thousands of important things about our ancestors, many of which can be used to derive falsifiable predictions about our psychological architecture: our ancestors had two sexes; contracted infections by contact; collected plant foods; inhabited a world where the motions of objects conformed to the principles of kinematic geometry; chose mates; had color vision; were predated upon; had faces; lived in a biotic environment with a hierarchical taxonomic structure, etc. To the extent that reconstructions are uncertain or erroneous, they will simply lead to experiments that are no more or less likely to be productive than the blind guessing of evolutionarily agnostic empiricism, the alternative research strategy.

How Good are Adaptations?

Critics have argued that adaptationist analyses are misconceived, because (they assert) natural selection is a weak force in evolution, making adaptations poor, and rendering functional predictions irrelevant. However, researchers point out that the empirical record shows that selection regularly produces very well-engineered adaptations to long-enduring adaptive problems, providing a solid empirical foundation for analyzing the human psychological architecture in functional terms. Whenever engineers have attempted to duplicate any natural human competence (color

vision, object recognition, grammar acquisition, texture perception, object manipulation, language comprehension, etc.), even when using huge budgets, large research teams, and decades of effort, they are unable to engineer artificial systems that can come close to competing with naturally engineered systems. It seems illogical to assert as a theoretical precept that evolved systems are poorly designed when human engineers cannot produce anything nearly as good.

Why is Function Emphasized?

Because EP emphasizes function, critics have argued that proponents think that all traits are adaptive. This is untrue, either in theory or in practice (for example Kurzban *et al*.'s argument above is that race encoding is a byproduct, not an adaptation). Although evolutionary psychologists do not think all – or even most – traits are adaptations, they do emphasize the study of adaptations and their byproducts for four reasons.

1. Adaptationist theories of function provide clear and useful prior predictions about cognitive organization.
2. The functional elements of an adaptation are far more likely to be species-typical, making them easier to discover through experimentation.
3. Very few constrained or falsifiable predictions about cognitive architecture follow from analyses of the random or contingent components of evolution.
4. As yet, there are few, if any, useful or well-developed theories of non-adaptive constraint.

Finally, evolutionary psychologists do not maintain that the developed architecture of the human mind is immune to modification, that genes or biology are deterministic, that culture is unimportant, or that existing human social arrangements are fair or inevitable. Indeed, they provide testable theories about the developmental processes that build (and can change) the mechanisms that generate human behavior.

Further Reading

Baron-Cohen S (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Barkow J, Cosmides L and Tooby J (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Buss DM (1994) *The Evolution of Desire*. New York, NY: Basic Books.

Cosmides L and Tooby J (1992) Cognitive adaptations for social exchange. In: Barkow J, Cosmides L and Tooby J (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Cosmides L and Tooby J (2000) The cognitive neuroscience of social reasoning. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 1259–1270. Cambridge, MA: MIT Press.

Cosmides L and Tooby J (2002) *What is Evolutionary Psychology? Explaining the New Science of the Mind*. London, UK: Weidenfeld & Nicolson.

Daly M and Wilson M (1988) *Homicide*. New York, NY: Aldine DeGruyter.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Hagen EH (1999) The functions of postpartum depression. *Evolution and Human Behavior* **20**: 325–359.

Kurzban R, Tooby J and Cosmides L (2001) Can race be erased?: coalitional computation and social categorization. *Proceedings of the National Academy of Sciences of the USA* **98**(26): 15387–15392.

Pinker S (1994) *The Language Instinct*. New York, NY: Morrow.

Silverman I and Eals M (1992) Sex differences in spatial abilities: evolutionary theory and data. In: Barkow J, Cosmides L and Tooby J (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Symons D (1979) *The Evolution of Human Sexuality*. New York, NY: Oxford University Press.

Tooby J and DeVore I (1987) The reconstruction of hominid behavioral evolution through strategic modeling. In: Kinzey W (ed.), *Primate Models of Hominid Behavior*. New York, NY: SUNY Press.

Wakefield J (1999) Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology* **108**: 374–399.

Evolutionary Psychology: Theoretical Foundations

Introductory article

Leda Cosmides, University of California, Santa Barbara, USA
John Tooby, University of California, Santa Barbara, USA

CONTENTS

Evolutionary psychology is an approach to the cognitive sciences whose goal is to map the evolved, species-typical cognitive and neural architecture of humans (and other species). Its focus is on integrating what is known about evolution into the research process, allowing evolutionary psychologists to derive hypotheses about the design of human information-processing mechanisms from the large pre-existing body of theories already developed and empirically tested within modern evolutionary biology.

INTRODUCTION

Evolutionary psychologists view the human mind as a set of computational machines that were designed by natural selection to solve adaptive problems faced by our hunter–gatherer ancestors. They argue that this basic Darwinian insight, when properly applied, can be uniquely informative for anyone who seeks to discover and understand the design of the human mind – that is for anyone who wishes to discover which programs reliably develop in the brains of all normal human beings, the conditions that activate these programs, and how each program processes information. In their view, attention to adaptive function will allow psychologists to (1) explain why the human mind contains those programs that are already known, (2) discover new programs that no one had thought to look for before, and (3) together with the analytical tools of the cognitive sciences, address previously intractable or neglected topics, such as emotion and motivation.

FOUNDATIONS

Although evolutionary psychology is an inclusive discipline that draws on many fields, its core ideas emerged from the intersection of three scientific research traditions: (1) work by David Marr, Noam Chomsky, and other cognitive scientists, that showed that the mind contains a number of different cognitive programs, many of which are specialized for performing a particular function (such as seeing or learning a language), (2) hunter–gatherer and primate studies, and (3) the revolution in evolutionary biology led by George C. Williams, W. D. Hamilton, John Maynard Smith, and Richard Dawkins, that replaced vague notions of function with a theoretically and empirically rigorous modern adaptationism, based on theories of natural selection that were formalized using game theory and replicator dynamics.

Out of replicator dynamics, evolutionary researchers derived a series of theories about how natural selection designs mechanisms that deal with parenting, mating, cooperation, kinship, communication, conflict, and dozens of other adaptive problems. These theories were then validated on thousands of animal and plant species. Evolutionary anthropologists have enriched this body of knowledge by studying primate and hunter–gatherer behavior and ecology, investigating hominid evolution and ancestral environments, and by extending evolutionary theory to cover novel features of the human species. This allows increasingly refined models of the adaptive problems our ancestors faced and how selection acted on them.

The understanding that natural selection is the only anti-entropic force known to scientists that builds functional machinery into organisms led to the third major insight: mechanisms studied by cognitive scientists necessarily had to be adaptations. This connected evolutionary research to cognitive science in the most direct possible way: cognitive science is the study of adaptations – computational adaptations. This allowed evolutionary psychologists to widen cognitive science

into a comprehensive mapping of the computational mechanisms that underlie all human behavior, not just traditional cognitive topics such as attention, learning, and memory. The fact that no single information-processing design can efficiently solve a diversity of adaptive problems means that specialized cognitive mechanisms are likely to have evolved to regulate human parenting, social interaction, mating, foraging, incest avoidance, sexual jealousy, coalitions, and so on. The goal of evolutionary psychology is the construction of a high resolution map of the whole species-typical computational architecture of humans, including motivational and emotional mechanisms.

Engineering and Reverse Engineering

Evolutionary psychologists approach their field conceptually as if it were a form of reverse engineering. Engineers start with a problem, and then design machines that are capable of solving that problem in an efficient manner. As a result, the machine's structure reflects its function: it has certain properties and components rather than others because those structures solve a problem better than alternative ones.

Engineers can also work in reverse: given a strange machine, they can figure out what its design features are – i.e., which of its components are functional and how their arrangement accomplishes the machine's function. Doing this is relatively simple if one knows what problem the machine was designed to solve, because one can then look for structures capable of accomplishing that function. But, as any engineer will confirm, reverse engineering is exceedingly difficult when one has no idea what the machine was designed to do. Without a theory of function, how does one determine which parts are functional.

Cognitive psychologists are engineers working in reverse: the brain is a strange machine, and cognitive psychologists are attempting to figure out how it works, i.e. which of its components are functional and how their arrangement accomplishes various functions. Doing this is difficult, however, without knowing what problems this organic machine was designed to solve.

Evolutionary biology is helpful because it provides theories about what problems the brain was designed to solve, that is, theories about the functions of its constituent programs. This is done using (1) knowledge about basic problems any organism must solve if it is to survive and reproduce (e.g., finding food efficiently, choosing a fertile mate), (2) knowledge about ancestral environments for the species in question, and (3) evolutionary game theory to model which of an array of possible solutions would have replicated fastest under ancestral conditions (and therefore have been favored by natural selection). From these elements, one can develop a task analysis for an adaptive problem, the first step in developing a design specification: an answer to the question, 'What would a machine capable of solving this problem well under ancestral conditions look like?' The answer(s) to this question then guide one's empirical investigations.

For example, certain species (including our own) trade goods and favors (cooperation for mutual benefit). But results from evolutionary game theory showed that natural selection will not favor cognitive machinery enabling this somewhat unusual form of cooperation unless the individuals involved are able to detect cheaters (those who do not reciprocate favors). This led Cosmides and Tooby to look for, and find, reasoning programs specialized for cheater detection. Baron-Cohen's research on 'mindreading' – programs that allow people to infer the intentions, beliefs, and desires of others – was guided by theories about coevolutionary arms races, as well as by knowledge about what information was available in ancestral environments for inferring mental states. In the evolutionary past (as now) eye direction provided reliable and useful information about the intentions of other people and of predators. Noting this, Baron-Cohen hypothesized that specialized eye direction detectors may have evolved as a component of social cognition and predator detection, and he designed experiments testing for their existence and design.

Despite all the obvious differences between living beings and human-made machines, reverse engineering is a successful strategy for studying organisms because the two resemble each other in one crucial respect. Like human-made machines, organisms are comprised of structures that reflect their function. This is an inevitable consequence of how natural selection works, and fundamental to the logic of evolutionary psychology (see below).

Evolutionary Restrictions on the Concept of 'Function'

George Williams's 1966 book, *Adaptation and Natural Selection*, played a key role in the development of evolutionary psychology. Williams elucidated the levels at which natural selection can operate

(genes and individuals, yes; groups, species, and ecosystems, weakly or not at all); he demonstrated why it will operate most powerfully in constructing adaptations at the level of the gene and individual; he clarified the logic of adaptationism; and he established standards of evidence that must be met before any trait can be considered an adaptation.

Before Williams, vague, panglossian functionalist thinking permeated evolutionary biology (and such thinking continues, implicitly, to saturate other fields even today). Evolutionary accounts explain the existence of traits by reference to their function, but many biologists (and psychologists) were attributing functionality merely by identifying a beneficial consequence to some entity, whether this was an individual, social group, species, or ecosystem. They did not focus on establishing whether the design systematically caused the propagation of its genetic basis reliably under ancestral conditions, as the theory of natural selection requires. Williams showed why looser notions of function were deficient, demonstrated how tightly constrained any adaptationist (i.e., functionalist) or byproduct claim had to be to be consistent with neo-Darwinism, and outlined the strict criteria such claims had to meet. Evolutionary psychologists attempt to apply these stringent adaptationist constraints on functionalism to limit the looser and less formalized ideas of function commonly employed in the cognitive, neural, and social sciences. They maintain that cognitive scientists should be aware that cognitive theories typically assume complex functional organization of types that are inconsistent with what evolution is likely to have produced.

Perhaps more importantly, when cognitive scientists do not understand what legitimately counts as a function in an evolved system, they fail to look for forms of complex functional organization that natural selection is likely to have produced. For example, although cognitive neuroscientists look for brain systems designed to cause fear in response to physical threats, they do not look for systems designed to cause sexual jealousy in response to threats to a mating relationship. This failure to investigate stems from an erroneous belief that 'beneficial' refers to survival rather than gene replication. A system that causes sexual jealousy jeopardizes the individual's survival by triggering aggressive conflicts, yet it would have promoted its own reproduction in the past in relation to the design alternative: indifference to a mate's extrapair sexual behavior. Unsurprisingly, adaptations to prevent others from having sexual access to one's mate have evolved in a large variety of animal species, including humans (as Buss, Symons and Daly & Wilson have shown).

THE DESIGN OF ORGANISMS

The goal of Darwin's theory was to explain the designs of organisms. Darwin asked, for example, why the beaks of finches differ from one species to the next, and have the forms that they do. Why do animals expend energy attracting mates, energy that could be spent on survival? Why are human facial expressions of emotion similar to those found in other primates?

Two Principles: Common Descent and Adaptation

One of the most important evolutionary principles accounting for the characteristics of organisms is common descent. An increasing body of evidence indicates that all organisms alive today are the descendents of a single originating organism. Over the course of evolutionary time, new species originate because one breeding population sometimes becomes subdivided into two or more populations, and stops interbreeding. Although they start out with the same set of genes, they subsequently can evolve independently because the different populations no longer exchange genes through matings. This process of species splitting gives a hierarchical tree structure of similarity to all species on Earth. Offspring inherit their parents' genes and design features, which stay the same across the generations unless selection or chance modifies genes. So, the more recently two species were descended from the same ancestral species, the more design features they will share in common. Hence, we expect to find many similarities between humans and our closest primate relatives. For example, humans and chimpanzees both use exactly the same set of muscles in making parallel facial expressions.

This is the phylogenetic approach, and it consists of the search for features (called 'homologous features') that are similar because both species inherited them from the species that was their common ancestor. This approach has a long and productive history in psychology. But as valuable as it is for many questions, this approach cannot adequately address features that evolved uniquely in only one lineage, because there are then no similarities to compare. Because there is the widespread misimpression that evolutionary psychology consists solely or primarily of applying a phylogenetic

approach, many take it as a given that evolutionary psychology cannot address the large set of properties that make us uniquely human. They think it is limited to the study of characteristics that we share with other animal species.

However, the second principle that accounts for the characteristics of organisms is that natural selection builds adaptations into their designs. Indeed, natural selection can cause the designs of different species to diverge from one another, sometimes producing characteristics that are unique to a given species, such as the elephant's trunk or the cognitive mechanisms that allow humans to learn language. *Adaptationism* is the name for the research program that gives a central role to exploring how natural selection functionally organizes the designs of organisms. It can be applied to analyze features that are unique to humans, because the theory of natural selection illuminates equally well features that are shared and features that are unique to a single species.

Although evolutionary psychologists certainly appreciate and invoke phylogenetic explanations where they are appropriate (as well as other relevant theories and analytic tools), it is the application of adaptationist logic that has provided the brightest illumination to formerly murky issues in human psychology.

Organization in Evolved Systems

Organisms, like watches or automobile engines, exhibit a multitude of parts and subassemblies that are arranged in precise and highly ordered ways so that they operate to achieve the functional ends they were designed to perform. The eyes, immune system, umbilical cord, cell nucleus, and lungs, to pick a handful of examples, all display a very advanced technology, built out of organic molecules. The more that chance events, such as accidents or violence, act to randomize this internal order, the more the watch, automobile, or organism is damaged. In a world where everything is bombarded by chance forces, where did all this functional order in animals and plants come from?

The evolutionary process has only two components, chance and natural selection, that govern how the genes in a species change over time. Chance processes act to randomize relationships within the organism, and so cannot account for the accumulation of the highly ordered arrangements of functional parts that permeate organisms. For this reason, modern researchers now understand that natural selection is the only component of the evolutionary process that can build complex functional organization into a species' structure. This means that all complex *functional* design in organisms was created by natural selection. Consequently, we know that all functional organization in humans must be built in a way that is consistent with the principles of natural selection. This recognition is what makes evolutionary biology the foundation of psychology and neuroscience, not to mention anatomy, physiology, the medical sciences, and the social sciences. Our functional order originally comes from evolution.

To be sure, there is much that is not functional in organisms as well, introduced by chance evolutionary and non-evolutionary processes. But the functional architecture of organisms is central to their organization, and they would not exist without it. Indeed, in evolved systems there is a sense in which function determines structure, and that is the key to understanding the design of the human cognitive architecture.

Natural Selection: How (and Why) Function Determines Structure

The notion that species evolve – that their design changes over time – had been proposed and hotly debated before Darwin was born. But the early evolutionists lacked a clear and convincing account of how or why this happens. That is what Darwin and Wallace provided. They discovered a materialist mechanism – natural selection – that explains how organisms acquire their design, as well as why that design changes over time. The revolution that ensued bears Darwin's name because he is the one who elaborated the theory, provided the most extensive evidence for it, and was willing to pursue its implications wherever they led – even when they led to the human mind.

Many breakthroughs in science happen not because of new data, but because of a new way of looking at things. This was true for Darwin. Everyone already knew that organisms reproduce, and that when they do, they give rise to similar organisms: rabbits give birth to rabbits, not to ducks. They also knew that, while offspring closely resemble their parents, they are not perfect replicas of them. They vary a bit, and some of these variants are able to perform certain tasks, such as producing milk, better than others. This was common knowledge based on centuries of animal husbandry in which people selectively bred individual animals with special abilities – cows that produced more milk, sheep that grew softer wool. And Darwin, like Descartes, Harvey, and many others before

him, knew that an organism can be thought of as a machine: a system whose parts are designed to perform certain functions.

All of these facts fall into place, Darwin realized, if you think of an organism as a self-reproducing machine. What distinguishes living from nonliving machines is reproduction: the presence in a machine of devices (organized components) that cause it to produce new and similarly reproducing machines. Given a population of living machines, this property – self-reproduction – will drive a system of positive and negative feedback that can explain the remarkable fit between organisms and their environment.

In contrast to human-made machines, which are designed by inventors, living machines acquire their intricate functional design over deep time, as a downstream consequence of the fact that they reproduce themselves. Indeed, modern Darwinism has an elegant deductive structure that logically follows from Darwin's initial insight that reproduction is the defining property of life, the driving force that causes species to change over time. That logic is as follows. When an organism reproduces, replicas of its design features are introduced into its offspring. But the replication of the design of the parental machine is not always error-free. As a result, randomly modified designs (mutants) are introduced into populations of reproducers. Because living machines are already exactingly organized so that they cause the otherwise improbable outcome of constructing offspring machines, random modifications will usually introduce disruptions into the complex sequence of actions necessary for self-reproduction. Consequently, most newly modified but now defective designs will remove themselves from the population – a case of negative feedback.

However, a small number of these random design modifications will, by chance, improve the system's machinery for causing its own reproduction. Such improved designs (by definition) cause their own increasing frequency in the population – a case of positive feedback.

This increase continues until (usually) such modified designs outreproduce and thereby replace all alternative designs in the population, leading to a new species-standard design. After such an event, the population of reproducing machines is different from the ancestral population: the population- or species-standard design has taken a step 'uphill' toward a greater degree of functional organization for reproduction than it had previously. Over the long run, down chains of descent, this feedback cycle pushes designs through state–space towards increasingly well-engineered – and otherwise improbable – functional arrangements. These arrangements are functional in a specific sense: the elements are well-organized to cause their own reproduction in the environment in which the species evolved.

For example, if a mutation appears that causes individuals to find family members sexually repugnant, then they will be less likely to conceive children incestuously. They will produce children with fewer genetic diseases, more of these children will mature and reproduce than will the children of those who are not averse to incest. Such an incest-avoiding design will produce a larger set of healthy children every generation, down the generations. By promoting the reproduction of its bearers, the incest-avoiding circuit thereby promotes its own spread over the generations, until it eventually replaces the earlier-model sexual circuitry and becomes a universal feature of that species' design. This spontaneous feedback process – natural selection – causes functional organization to emerge naturally and inevitably, without the intervention of an intelligent designer or supernatural forces.

Genes are simply the means by which design features replicate themselves from parent to offspring. They can be thought of as particles of design: elements that can be transmitted from parent to offspring, and that, together with stable features of an environment, cause the organism to develop some design features and not others. Genes have two primary ways they can propagate themselves: by increasing the probability that offspring will be produced by the organism in which they are situated, or by that organism's kin.

An individual's genetic relatives carry some of the same genes, by virtue of having received some of the same genes from a recent common ancestor. This means that a gene in an individual that causes an increase in the reproductive rate of that individual's kin will, by so doing, tend to increase its own frequency in the population. A circuit that motivates an individual to help feed her sisters and brothers, if they are starving, is an example of a program that increases kin reproduction. As W. D. Hamilton pointed out, design features that promote both direct reproduction and kin reproduction, and that make efficient trade-offs between the two, will replace those that do not (a process called 'kin selection').

How well a design feature systematically promotes direct and kin reproduction is the bizarre but real engineering criterion determining whether a specific design feature will be added to or discarded from a species' design. Therefore, we can

understand why our brains are constructed in the way they are, rather than in other perfectly possible ways, when we see how its circuits were designed to cause behavior that, in the world of our ancestors, led to direct reproduction or kin reproduction.

The concept of *adaptive behavior* can now be defined with precision. Adaptive behavior, in the evolutionary sense, is behavior that tends to promote the net lifetime reproduction of the individual or that individual's genetic relatives. By promoting the replication of the genes that built them, circuits that – systematically and over many generations – cause adaptive behavior become incorporated into a species' neural design. In contrast, behavior that undermines the reproduction of the individual or his or her blood relatives removes the circuits causing those behaviors from the species, by removing the genes that built those circuits. Such behavior is *maladaptive*, in the evolutionary sense.

So, evolutionists analyze how design features are organized to contribute to lifetime reproduction not because of a warped and biasing obsession with sexuality, but because reproduction was the final causal pathway through which a functionally improved design feature caused itself to become more numerous with each passing generation, until it became standard equipment in all ordinary members of the species.

Adaptive problems create adaptations

Enduring conditions in the world that create reproductive opportunities or obstacles, such as the presence of predators, the ability to pool risk through food sharing, or the vulnerability of infants, constitute adaptive problems. Adaptive problems have two defining characteristics. First, they are conditions or cause-and-effect relationships that many or most individual ancestors encountered, reappearing again and again during the evolutionary history of the species. Second, they are that subset of enduring relationships that could, in principle, be exploited by some property of an organism to increase its reproduction or the reproduction of its relatives. Enduring relationships of this kind constitute reproductive opportunities or obstacles in the following sense: if the organism had a property that interacted with these conditions in just the right way, then this property would cause an increase in its own reproductive rate.

One can think of these reproductive opportunities and obstacles as problems. A property is a solution to such a problem when it allows organisms with this property to take advantage of prevailing conditions, where 'advantage' means a reproductive advantage. If a bird would realize a reproductive advantage by being able to travel at night, and stars are prevailing conditions that – given the right brain mechanism – would make this possible, then a brain mechanism that uses stars for navigation would be a solution to the problem of traveling at night. Egg-eating predators pose an obstacle to a bird's reproduction. A property that circumvents this obstacle – such as a program that causes the bird to remove from its nest broken eggshells whose bright white interiors are easily spotted by predators – would be a partial solution to this problem. A property is a solution to an adaptive problem if it had the systematic effect, over generations, of increasing the reproduction of individual organisms or their relatives. The causal chain linking that property to reproduction may be indirect, and the effect on the organism's own offspring or the offspring of kin may be relatively small. As long as its consequences on relative reproduction are the cause of its spreading through the population, that property is a solution to an adaptive problem. All solutions are, of course, temporary and subject to improvement over time. But each modification that spread because it improved reproduction – however stop-gap or impermanent it may turn out to have been – counts as a solution to an adaptive problem.

Most adaptive problems have to do with relatively mundane aspects of how an organism lives from day to day: what it eats, what eats it, who it mates with, who it socializes with, how it communicates, and so on. Adaptive problems for our hunter–gatherer ancestors included such recurrent tasks as giving birth, winning social support from band members, remembering the locations of edible plants, hitting game animals with projectiles, breast-feeding, breathing, identifying objects, recognizing emotional expressions, protecting family members, maintaining mating relationships, self-defense, heart regulation, assessing the character of self and others, causing impregnation, acquiring language, maintaining friendships, thwarting antagonists, and so on.

An enduring adaptive problem constantly selects for design features that promote the solution to the problem. Over evolutionary time, more and more design features accumulate that fit together to form an integrated structure or device that is well engineered to solve its particular adaptive problem. Such a structure or device is called an 'adaptation'. Indeed, an organism can be thought of as largely a collection of adaptations, such as the functional subcomponents of the eye, liver, hand, uterus, or circulatory system. Each of these adaptations exists in the human design now because it contributed

to the process of self and kin reproduction in the past.

Recognizing adaptations

Natural selection is a hill-climbing feedback process that chooses among alternative designs on the basis of how well they function. It has produced exquisitely engineered biological machines – the vertebrate eye, photosynthetic pigments, efficient foraging algorithms, color constancy systems – whose performance is unrivaled by any machine yet designed by humans.

Because adaptations are problem-solving machines, they can be identified using the same standards of evidence that one would use to recognize human-made machines (e.g., TV versus stove): design evidence. One can identify an aspect of the phenotype as an adaptation by showing that (1) it has many design features that are complexly specialized for solving an adaptive problem, (2) these phenotypic properties are unlikely to have arisen by chance alone, and (3) they are not better explained as the byproduct of mechanisms designed to solve some alternative adaptive problem.

Adaptations, byproducts, and noise

The features of a species' cognitive or neural architecture can be partitioned into: adaptations, which are present because they were selected for (e.g., the enhanced recognition system for snakes coupled with a decision-rule to acquire a motivation to avoid them); byproducts, which are present because they are causally coupled to traits that were selected for (e.g., the avoidance of harmless snakes); and noise, which was injected by the stochastic components of evolution (e.g., the fact that a small percentage of humans sneeze when exposed to sunlight). The standards for recognizing adaptations also allow one to recognize byproducts and noise.

One payoff of integrating adaptationist analysis with cognitive science was the realization that, in long-lived, sexually recombining species (like humans), complex functional structures will be overwhelmingly species-typical. That is, the complex adaptations that compose the human cognitive architecture must be human universals (at least at the genetic level), whereas variation caused by genetic differences is predominantly noise: minor random perturbations around the species-typical design. This principle allows cross-cultural triangulation of the species-typical design, which is why many evolutionary psychologists include cross-cultural components in their research.

WHAT IS A COMPUTATIONAL ADAPTATION?

Organisms are composed of many parts, but only some of these parts are computational. By *computational* we mean that they are designed to (1) monitor the environment for specific changes, and (2) regulate the operation of other parts of the system functionally on the basis of the changes detected. For example, the diaphragm muscle, which causes the lungs to contract and expand, is not computational. But the system that measures carbon dioxide in the blood and regulates the contraction and extension of the diaphragm muscle is. The plastic cover on a thermostat is not computational, nor are the parts of a furnace that generate heat. But the thermocouple that responds to ambient temperature by toggling the switch on the furnace, and the connections between them, form a computational system. Muscles are not computational, but the visual system that detects the presence of a hungry-looking lion, the inference mechanisms that judge whether that lion has seen you or not, and the circuits that cause your muscles either to run to a nearby tree (if the lion has seen you) or freeze (if it hasn't seen you) do compose a computational system. The language of information-processing can be used to express the same distinction: one can identify the computational components of a system by isolating those aspects that were designed to regulate the operation of other parts of the system on the basis of information from the internal and external environment.

By 'monitoring the environment for specific changes', we mean the system is designed to detect a change in the world. That change can be internal to the organism (such as fluctuations in carbon dioxide levels in the blood or the activation of a memory trace) or external to the organism (such as the onset of a rainstorm or the arrival of a potential mate). Changes in the world become *information* when (1) they interact with a physical device that is designed to change its state in response to variations in the world (i.e., a transducer), and (2) the changes that are registered then participate in a causal chain that was designed to regulate the operation of other parts of the system. A photon, for example, does not become information until it causes a chemical reaction in a retinal cell, which was designed for this purpose and is part of a causal system that was itself designed to regulate an organism's behavior on the basis of inferences about what objects exist in the world and where they are.

A set of features is not computational unless it was *designed* to exhibit these properties. For example, the outer cells of a dead tree stump expand in the rain, and as this happens, the inner portions of the stump might become compressed. But these dead cells were not designed for detecting changes in weather. More importantly, although their swelling does cause a change in the inner part of the stump, it is not *regulating* the operation of the stump. Regulation means more than merely influencing or changing something. It means systematically modifying the operation of a system so that a *functional* outcome is achieved. In the case of a thermostat, that function was determined by the intentions of the engineer who designed it. In the case of an organism, that function was determined by natural selection, which acted to organize the properties of the organism.

The Relationship between Brains, Computation, and Selection

Neurons do not perform any significant metabolic function for an organism. They exist because of the computational relationships they create. Natural selection retains neural mechanisms on the basis of their ability to create functionally organized relationships between information and behavior (e.g., the sight of a predator activates inference procedures that cause the organism to hide or flee) or between information and physiology (e.g., the sight of a predator increases the organism's heart rate in preparation for flight). Each neural program was selected for because it created the correct information–behavior or information–physiology relationship, and, so long as a physical implementation produces this relationship, its particular form is free to vary according to other factors. (Indeed, when people recover function after brain damage, repair processes often restore the original information-processing relationship – but using a different set of physical connections.)

In other words, the brain was designed by natural selection to be an information-processing device. The brain has the physical structure that it does *because* this structure embodies a particular set of programs, and each program has the computational structure that it does *because* that structure solved a particular problem in the past. This is the causal chain that licenses inferences from function to program structure to physical structure. If one knows what problems our ancestors faced, then one can make educated guesses about what programs evolved to solve them, including what computational procedures they would have required. Once the existence of these programs has been experimentally confirmed, one can search for their neural basis. Having a theory of adaptive function is useful to psychologists and neuroscientists because it allows one to look for programs and neural systems that otherwise one would not look for. It also allows one to understand why programs already known have the computational design that they do.

Function Determines Computational Structure

In principle, a computer or neural circuit could be designed so that any given stimulus in the environment (e.g., feces) could cause any kind of resulting behavior (avoid it, eat it, dance around it, meditate, declaim, sculpt, etc.). Which behavior a stimulus gives rise to is a function of the neural circuitry of the organism. This means that if you were a superhuman designer of brains, you could have engineered the human brain to respond in any way that you wanted, to link any environmental input to any behavioral output. You could have made a human being who frowns when pleased, is erotically transported by tree bark, or howls and devotedly incubates chicken eggs with her body heat whenever the days grow short enough. To explain behavior, therefore, we need a theory of brain organization that describes how circuits are designed to respond to environmental inputs throughout the lifecycle, and why they have the form they do. We will call this organization 'the design of the mind'. Because how the brain is organized to respond to the environment, prior to experience, cannot itself be supplied by the environment, it is easy to see that the design of the mind – including its learning circuits – must be built in to the developing brain. This means that the design was created by evolution.

Adaptive problems that required information-processing for their solutions selected for neural adaptations organized to compute these solutions: function determined computational structure. Over evolutionary time, neural circuits were cumulatively added to the design of the human brain because they reasoned or processed information in a way that enhanced the adaptive regulation of behavior and physiology for these enduring adaptive problems. Such cognitive adaptations include emotion programs, such as fear of falling, fear of snakes, or parental love; motivational programs, such as sexual attraction or revenge; reasoning instincts such as cheater detection algorithms; and learning programs such as the language acquisition device or the food aversion system.

Even 'learned' behaviors, such as speaking English, are the product of evolved learning programs. Evolutionarily novel skills, such as reading and writing, are learned via programs that evolved for learning other, evolutionarily important skills, such as language acquisition – reading and writing are byproducts of adaptations designed for learning other things. Consequently, the mental and neural organization that results from learning is simply another example of the operation of our evolved adaptations, not an exception.

NATURE AND NURTURE

At a certain level of abstraction, every species has a universal, species-typical evolved architecture. For example, we all have a heart, two lungs, a stomach, and so on. This is not to say there is no biochemical individuality, especially in quantitative features. Stomachs vary in size, shape, amount of HCl produced, but all stomachs have the same basic functional design. They are attached at one end to an esophagus and at the other to the small intestine, secrete the same chemicals necessary for digestion, etc. Presumably, the same is true of the brain and, hence, of the evolved architecture of our cognitive programs – of the information-processing mechanisms that generate behavior. Evolutionary psychology seeks to characterize the universal, species-typical architecture of these mechanisms.

Our evolved cognitive architecture, like all aspects of the phenotype from molars to memory circuits, is the joint product of genes and environment. But the development of architecture is buffered against both genetic and environmental insults, such that it reliably develops across the (ancestrally) normal range of human environments. Adaptations are not impervious to environmental conditions: a certain envelope of environmental conditions must be present for any adaptation to develop properly. Moreover, the evolutionary function of computational adaptations is to make behavior (and physiology) sensitively contingent upon information from the environment.

A mechanism – computational or otherwise – need not be present at birth to be considered an adaptation or part of the human evolved architecture (consider teeth and breasts). The development of an adaptation may be triggered at any point in life-history; the trigger can be an internal, physiological event or a set of external conditions (including social conditions).

Evolutionary psychology is not behavior genetics. Behavior geneticists are interested in the extent to which *differences* between people can be accounted for by *differences* in their genes. Evolutionary psychologists are interested in individual differences primarily insofar as these are the manifestation of an underlying architecture shared by all human beings. Because their genetic basis is universal and species-typical, the heritability of complex adaptations (e.g., the eye) is usually low, not high. Moreover, sexual recombination constrains the design of genetic systems, such that the genetic basis of any complex adaptation (such as a cognitive mechanism) *must* be universal and species-typical. This means the genetic basis for the human cognitive architecture is universal, creating what is sometimes called the 'psychic unity of humankind'.

Evolutionary psychologists do not assume that genes play a more important role in development than the environment does, or that 'innate factors' are more important than 'learning'. Instead, they reject the traditional nature/nurture dichotomies as ill-conceived. In their view, there is not a zero-sum relationship between 'nature' and 'nurture'. For them, 'learning' is not an explanation; it is a phenomenon that requires explanation. Learning is caused by cognitive mechanisms, and to understand how it occurs one needs to know the computational structure of the mechanisms that cause it. The richer the architecture of these mechanisms, the more an organism will be capable of learning: toddlers can learn English while the family dog cannot because the cognitive architecture of humans contains mechanisms that are not present in that of dogs.

DOMAIN-SPECIFICITY AND FUNCTIONAL SPECIALIZATION

Evolutionary psychologists do not assume that 'learning', reasoning, or memory are unitary phenomena. The learning mechanisms that cause the acquisition of grammar, for example, are different from those that cause the acquisition of snake phobias. Corkscrews and cups have different properties because they are solutions to different problems; similarly, machinery that causes predator fears to be reliably and efficiently acquired lacks properties that cause the reliable and efficient acquisition of grammar, and vice versa. This applies to choice as well as learning: in many cases, the computational requirements for producing adaptive behavior in one domain are incompatible with those for another. Consider, for example, the domains of food and sex. The computational structure of programs designed for choosing nutritious foods will fail to produce adaptive behavior unless

they generate different preferences and trade-offs than programs designed for choosing fertile sexual partners.

Because natural selection tends to produce mechanisms that are well designed for solving adaptive problems, evolutionary psychologists expect the human mind will be found to contain a large number of information- processing devices that are functionally specialized and therefore domain-specific. Most think the multipurpose flexibility of human thought and action is possible precisely because our cognitive architecture contains a large number of these expert systems.

The proposed domain-specificity of many of these computational devices separates evolutionary psychology from those approaches to the cognitive sciences that assume the mind to be composed of a small number of domain-general, content-independent, general-purpose mechanisms.

Relevance to Modularity Debate in Cognitive Science

In cognitive science, computational systems that are functionally specialized and domain-specific are sometimes called 'modules'. The criteria for calling a device a module are inconsistent and vague (some view information encapsulation as criterial; others emphasize specialization, etc.), especially when compared to the crisp criteria for calling a device an 'adaptation'. As a result, evolutionary psychologists are more comfortable discussing functional specialization rather than modularity. That said, it is fair to say that most take a more modular view of cognition than do many cognitive scientists. (*See* **Modularity)**

Examples of evolved computational devices that show evidence of being specialized in function include: face recognition systems, a language acquisition device, mindreading systems, navigation specializations, animate motion recognition, cheater detection mechanisms, and mechanisms that govern sexual attraction. Most evolutionary psychologists are skeptical that an architecture consisting predominantly of content-independent cognitive processes, such as general-purpose pattern associators, could solve the diverse array of adaptive problems efficiently enough to reproduce themselves reliably in complex, unforgiving natural environments that include, for example, antagonistically coevolving biotic adversaries, such as parasites, prey, predators, competitors, and incompletely harmonious social partners. Such systems may be able to detect some patterns in the environment, but they are value-free: that is, they contain no criteria for deciding between alternative courses of action in a way that would have tracked fitness in ancestral environments. Indeed, evolutionary psychologists have argued that there is no single criterion for adaptive behavior that could be applied across domains yet still track fitness and, for this reason, evolution could not have produced a completely domain-general cognitive architecture.

Some cognitive scientists have argued in favor of domain-general computational processes (usually of an unspecified nature) on the grounds that they can solve a wider array of problems, including evolutionarily novel ones (such as learning to read or write). Even if this were true (and there are strong reasons to believe it is false, having to do with combinatorial explosion and the greater inferential power of a knowledge-rich over a knowledge-poor reasoning system), it would provide no basis for assuming that the human cognitive architecture is composed primarily of domain-general, knowledge-free (i.e., content-independent) computational systems.

This is because selection drives design features to become incorporated into architectures in proportion to the actual distribution of adaptive problems encountered by a species over evolutionary time. There is no selection to generalize the scope of problem-solving to include never or rarely encountered problems at the cost of efficiency in solving frequently encountered problems. To the extent that problems cluster into types (domains) with statistically recurrent properties and structures (e.g., facial expression statistically cues emotional state), it will often be more efficient to include computational specializations tailored to inferentially exploit the recurrent features of the domain (objects always have locations, are bounded by surfaces, cannot pass through each other without deformation, can be used to move each other, etc.). Because the effects of selection depend on iteration over evolutionary time, evolutionary psychologists expect the detailed design features of domain-specific inference engines to intricately reflect the enduring features of domains. Consequently, they are very interested in careful studies of enduring environmental and task regularities, because these predict details of functional design. Adaptationist predictions of domain-specificity have gained support from many sources (e.g., from cognitive neuroscience), demonstrating that many dissociable cognitive deficits show surprising content-specificity, and from developmental research indicating that infants come equipped with evolved domain-specific inference engines (an intuitive physics, a mindreading system, a folk biology, etc.).

The Environment of Evolutionary Adaptedness (EEA)

Evolutionary psychologists do not study behavior *per se*; they study the cognitive machinery that generates behavior, using the theory of evolution by natural selection to develop hypotheses about its design and function. According to this view, behavior in the present is generated by information-processing mechanisms that exist because they solved adaptive problems in the past – in the ancestral environments in which the human line evolved.

As a result, evolutionary psychology is relentlessly past-oriented. Cognitive mechanisms that exist because they solved problems efficiently in the past will not necessarily generate adaptive behavior in the present (e.g., a taste for fat, adaptive in fat-poor ancestral environments, can generate maladaptive behavior in a modern environment flush with fast-food restaurants). Indeed, evolutionary psychologists reject the notion that one has 'explained' a behavior pattern by showing that it promotes fitness under modern conditions.

Although the hominid line is thought to have evolved on the African savannas, the environment of evolutionary adaptedness, or EEA, is not a place or time. It is the statistical composite of selection pressures that caused the design of an adaptation.

Further Reading

Baron-Cohen S (1995) *Mindblindness: An essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Barkow J, Cosmides L and Tooby J (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Buss DM (1994) *The Evolution of Desire*. New York, NY: Basic Books.

Cosmides L and Tooby J (2002) *Universal Minds: Explaining the New Science of the Mind*. London, UK: Weidenfeld & Nicolson.

Daly M and Wilson M (1988) *Homicide*. New York, NY: Aldine.

Dawkins R (1976) *The Selfish Gene*. New York, NY: Oxford University Press.

Dawkins R (1982) *The Extended Phenotype*. San Francisco, CA: W.H. Freeman.

Dawkins R (1986) *The Blind Watchmaker*. New York, NY: Norton.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Kelly R (1995) *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways*. Washington, DC: Smithsonian Institution Press.

Pinker S (1997) *How the Mind Works*. New York, NY: Norton.

Symons D (1979) *The Evolution of Human Sexuality*. New York, NY: Oxford University Press.

Tooby J and DeVore I (1987) The reconstruction of hominid behavioral evolution through strategic modeling. In: Kinzey W (ed.) *Primate Models of Hominid Behavior*, pp. 183–237. New York, NY: SUNY Press.

Wakefield J (1999) Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology* **108**: 374–399.

Williams GC (1966) *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press.

Gene–Culture Coevolution

Intermediate article

Kevin N Laland, University of Cambridge, Cambridge, UK

CONTENTS

Evolution in species with a dynamic, socially transmitted culture may be different from evolution in other species. Population geneticists have proposed the gene–culture coevolutionary approach to describe the way in which cultural change may drive a population's biological evolution.

INTRODUCTION

Many researchers have noted analogies between the processes of biological evolution and cultural change. For instance, both genes and culture are informational entities that are differentially transmitted from one generation to the next. These similarities have led to the idea that culture evolves, and prompted the development of mathematical models of cultural evolution.

The main scientific approach to the study of how culture evolves is a branch of theoretical population genetics, known variously as 'cultural evolution', 'gene–culture coevolution', or 'dual inheritance' theory. This intellectual tradition has nothing in common with the nineteenth-century 'cultural evolution' schools, which, based on an erroneous view of evolution as progressive, set out to model stages of societal development. Rather, the population genetics approach regards culture as an evolving pool of ideas, beliefs, values, and knowledge that is learned and socially transmitted between individuals. Researchers focus on a single trait, such as a preference for drinking milk, or for sons over daughters, and employ a rigorous mathematical approach to describe how the cultural trait changes over time, sometimes coevolving with genetic variation. Where the cultural entity is a discrete package, it has much in common with Richard Dawkins' idea of the 'meme', defined as a cultural analogue of the gene (Dawkins, 1976).

Stone tools appear in the archeological record approximately two and a half million years ago. If, as is widely believed, lithic technologies and skills were transmitted from one generation to the next, these simple artifacts represent the earliest evidence for culture. In fact, comparative evidence for social learning in a variety of vertebrate species suggests that cultural transmission almost certainly preceded *Homo habilis* by a considerable length of time. However, social learning in other animals is rarely stable enough to support traditions in which information accumulates from one generation to the next. For at least 2 million years our ancestors have reliably inherited two kinds of information, one encoded by genes, the other by culture.

There is only one evolutionary approach to the study of human behavior that takes up the challenge of understanding genetic and cultural evolution simultaneously by focusing directly on their interaction. Gene–culture coevolutionary theory (or dual inheritance theory), together with evolutionary psychology and human behavioral ecology, is one of three principal evolutionary approaches that emerged in the aftermath of the human sociobiology debate (Smith, 2000).

Conceptually, gene–culture coevolution is like a hybrid between memetics and evolutionary psychology, although its methods are quite different, relying as they do on rigorous mathematical theory. Like memeticists, gene–culture coevolution enthusiasts treat culture as evolving learned knowledge. Like evolutionary psychologists, these researchers believe that the cultural knowledge individuals adopt may sometimes – although certainly not always – depend on their genetic constitution. Moreover, selection acting on the genetic system is commonly generated or modified by the spread of cultural information. For gene–culture coevolutionary theorists, the 'leash' that ties culture to genes tugs both ways. The advent of culture was a precipitating evolutionary milestone, generating selection that favored a reorganization of the human brain, and leaving it specialized to acquire, store, and use cultural information. It was culture,

loosely guided by genes, that allowed humans the adaptive flexibility to colonize the world.

The quantitative study of gene–culture coevolution began in 1976, when two population geneticists, Luca Cavalli-Sforza and Marc Feldman of Stanford University, published the first simple dynamic models with both genetic and cultural inheritance. The fundamental innovation that Cavalli-Sforza and Feldman instigated was that, in addition to modeling the differential transmission of genes from one generation to the next, they incorporated cultural information into the analysis, allowing the evolution of the two systems to be mutually dependent. However, one curious feature of the history of gene–culture coevolution is that both archetypal sociobiologists and some of their most severe critics almost simultaneously recognized the importance of gene–culture interactions, with each starting to develop methods to address the problem. By the late 1970s, Charles Lumsden and Edward Wilson at Harvard University were engaged in a race with Cavalli-Sforza and Feldman to produced the first book on this topic. While Lumsden's and Wilson's *Genes, Mind and Culture* was published first (Lumsden and Wilson, 1981), it was not well regarded (Maynard-Smith and Warren, 1981). In contrast, Cavalli-Sforza's and Feldman's more cautious tome *Cultural Transmission and Evolution* (Cavalli-Sforza and Feldman, 1981) was much better received.

Together with many co-workers, Cavalli-Sforza and Feldman gradually built up an impressive body of mathematical theory exploring the processes of cultural change and interaction between genes and culture. Frequently they took advantage of the parallels between the spread of a gene and the diffusion of a cultural innovation to borrow or adapt established models from population genetics. Drawn by the ongoing sociobiology debate, other mathematically minded researchers joined the fray, most notably anthropologists Rob Boyd and Peter Richerson, whose book *Culture and the Evolutionary Process* introduced a variety of novel theoretical methods and stimulating ideas (Boyd and Richerson, 1985). Gradually a consensus as to the most appropriate methods for tackling gene–culture interactions began to emerge, which today forms the basis of modern coevolutionary theory.

The technical and explicitly mathematical nature of modern gene–culture coevolution is one of several features that distinguishes this perspective from alternatives such as evolutionary psychology. A second is the incorporation into analyses of a variety of genetic and cultural processes in addition to the natural selection of genes. Gene–culture coevolution exhibits a concern for nonadaptive and even maladaptive outcomes of the evolutionary process. This stance continues both to surprise and confuse outside observers used to characterizing all these evolutionary approaches as 'sociobiology'. However, the rigorous theoretical approach has led to little experimentation or other forms of empirical work, and this school remains the prerogative of a comparatively small band of workers.

The emerging body of theory has developed in a variety of ways. One class of models investigates the inheritance of behavioral and personality traits, extending traditional models by incorporating a transmitted cultural component into the analysis. Other models address general questions about the adaptive advantages of learning and culture. More recently, these methods have been applied to address specific cases in which there is an interaction between cultural knowledge and genetic variation that influences its prevalence. These include the evolution of language and of handedness, an analysis of changes in the genetic sex ratio in the face of sex-biased parental investment, the spread of agriculture, the coevolution of hereditary deafness and sign language, the emergence of incest taboos, and an exploration of how cultural niche construction affected human evolution (see Feldman and Laland, 1996).

As the rules of cultural transmission are usually different from those of genetic transmission, similar selective regimes may result in very different equilibria. A good example of this is provided by the hypothesis of Boyd and Richerson (1985) that group selection can act on cultural variation. The theoretical argument against group selection is based on models which assume genetic inheritance, and the criticisms may not hold for culturally transmitted traits. When individuals adopt the behavior of the majority a conformist transmission is generated (Boyd and Richerson, 1985). As a result of its frequency dependence, conformist transmission can act to amplify differences in the frequency of cultural traits in different subpopulations, but reduce variance within groups. Boyd and Richerson showed that one of the by-products of a conformist bias is an increase in the strength of the group selection of cultural variation so that it may be a strong force relative to forces acting within groups, such as natural selection. Since selection between groups may favor beliefs and attitudes that benefit the group at the expense of the individual, Boyd and Richerson's theory provides a new explanation for human cooperation.

EVIDENCE FOR TRANSMITTED CULTURE

For most social scientists 'culture' is a given. The notion that much of the variation in the behavior of humans is brought about by their being exposed to divergent cultures is so widespread and intuitive that is beyond dispute. While it used to be fashionable to define culture as the interwoven complex of behavior, ideas, and artifacts that characterize a particular people (e.g. Tylor, 1871), among social scientists this view has been superseded by a more cognitive perspective that restricts culture to information stored in the brain. In contrast, biological approaches to culture (including those of most sociobiologists, human behavioral ecologists, and evolutionary psychologists) tend to regard the transmitted elements of culture as either exerting a comparatively trivial influence on human behavior, or that whatever influence they have is strictly circumscribed by genes.

For advocates of gene–culture coevolution these biological perspectives underemphasize one critical factor: socially transmitted culture. Too much culture changes too quickly to be feasibly explained by genetic variation, while the fact that different behavioral traditions can be found in similar environments would appear to render environmental explanations of behavior impotent much of the time. To give an example, Guglielmino *et al*. (1995) analyzed variation in cultural traits among 277 contemporary African societies, and found that most traits examined correlated with cultural history rather than with ecology. Such findings suggest that most human behavioral traits are maintained in populations as distinct cultural traditions, rather than evoked by the natural environment. Genes and environment undoubtedly account for some variation in human behavior, but the socially transmitted component of culture is also important.

A capacity for culture is an unusual adaptation. It allows humans to learn about their world rapidly and efficiently. We do not have to scour our environment for sources of food and water, devise our own means of communication, or reinvent technological advances from first principles. Our capacity to acquire valuable skills and information from more knowledgeable others, such as parents, teachers or friends, as well as indirectly through artifacts such as books and computers, furnishes us with a short cut to adaptive (and sometimes maladaptive) behavior. Advocates of gene–culture coevolution share with the vast majority of social scientists the view that what makes culture different from other aspects of the environment is the knowledge passed between individuals. Culture is transmitted and inherited in an endless chain, frequently adapted and modified to produce cumulative evolutionary change. This infectious, information-based property of transmission is what allows culture to change rapidly, to propagate a novel behavior through a population, to modify the selection pressures acting on genes, and to exert such a powerful influence on our behavioral development.

Gene–culture enthusiasts point to countless studies that have found that the attitudes of parents and offspring are similar. They maintain that the most obvious explanation for this is that children learn social attitudes in the family. For instance, a study of Stanford University students revealed that the religious and political attitudes were strongly consistent between parents and offspring (Cavalli-Sforza *et al*., 1982). The same applies to nonindustrial societies. For instance among Aka pygmies, an African group of hunter-gatherers, there was evidence for parent to child transmission of many customs (Hewlett and Cavalli-Sforza, 1986). Such correlations do not prove cultural transmission to be prevalent: for instance, there could be heritable genetic effects. However, the weight of evidence supports the notion of a transmitted culture; see Boyd and Richerson (1985) for a more extensive collation of evidence for cultural transmission.

TYPES OF CULTURAL SELECTION

Researchers in the gene–culture coevolution tradition have described a number of processes that underpin cultural change. In order to distinguish cultural from biological evolution, Cavalli-Sforza and Feldman (1981) defined 'cultural selection' as a process by which particular socially learned beliefs, or pieces of knowledge, increase or decrease in frequency owing to their adoption by other individuals at different rates. In contrast, natural selection can change the frequency of a cultural preference through the differential survival of individuals expressing different types of preference. For instance, in developed countries fertility control (contraception) is at a clear disadvantage in natural selection as users typically have fewer offspring, but has spread by virtue of its advantage in cultural selection since fertility control is a popular choice. Working on these two interacting subsystems simultaneously helps us to understand how nonadaptive cultural traditions could evolve (Cavalli-Sforza and Feldman, 1981). When it has

sufficiently high cultural fitness, cultural information can increase in frequency despite decreasing genetic fitness. Cavalli-Sforza and Feldman's framework also considers cases in which cultural selection operates without affecting Darwinian fitness (e.g. a preference for a particular soft drink).

Researchers in gene–culture coevolution are also interested in how information spreads within populations. The mode of transmission is the route by which cultural knowledge spread among individuals (Cavalli-Sforza and Feldman, 1981), and different models are required for alternative modes of information transmission. Social transmission can occur vertically (that is, from parents to offspring), obliquely (from the parental to the offspring generation; for instance, learning from teachers or religious elders) or horizontally (that is, within-generation transmission such as learning from friends or siblings). Genetic inheritance is exclusively vertical, and as social transmission frequently occurs through some combination of these modes of information transmission, cultural evolution and gene–culture coevolution will exhibit quite different properties from those of biological evolution.

Boyd and Richerson (1985) extended the taxonomy of cultural processes by splitting the general concept of 'cultural selection' into subtypes. One such subtype is 'guided variation', which refers to a process by which individuals acquire information about a behavior culturally, and then modify the behavior on the basis of their personal experience. Here cultural variation is guided by individual experience which, as human behavioral ecologists envisage, allows behavioral traditions to evolve gradually towards the adaptive optimal behavior for that environment.

Another set of processes that Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985) consider is known as 'biased cultural transmission'. Biased transmission occurs when, given a choice between two alternative behavior patterns, individuals are more likely to adopt one variant than another. Various types of bias exist. In direct bias, individuals choose which of two or more alternative behavior patterns to adopt. A direct bias might result from a genetic predisposition to favor certain types of information. Stanford University anthropologist Bill Durham has argued that the individual choices underpinning these cultural processes are guided, but not determined, by predispositions and prior knowledge (Durham, 1991).

In the case of frequency-dependent bias, however, the commonness or rarity of a behavior affects the probability of information transmission. When, as frequently seems to be the case, individuals are predisposed to adopt the behavior of the majority, this frequency-dependent bias generates conformity. People may also use cues about one trait, for example, wealth, to choose which individuals to observe in order to acquire information about another trait, such as clothes fashions. Boyd and Richerson call this 'indirect bias'.

A Case Study: Coevolution of Dairy Farming and Genes for Processing Milk

The evolution of the ability of adult humans to consume dairy products represents a good example of gene–culture coevolution. Unlike that of human infants, virtually all of whom can all drink milk without problems, the milk digestive physiology of adult humans varies considerably. In fact, if the entire world's population is considered, consuming dairy products makes the majority of adult humans ill. This is because the activity level of the enzyme lactase in their bodies is insufficient to break down the energy-rich sugar lactose in dairy products, and milk consumption typically leads to sickness and diarrhea. The ability of adult humans to digest lactose largely depends on whether they possess the appropriate variants of a single gene. A correlation exists between the incidence of the genes for lactose absorption and a history of dairy farming in populations, with absorbers reaching frequencies of over 90% in dairy-farming populations, but typically less than 20% in populations without dairy traditions (Durham, 1991). The correlation is extremely suggestive. Milk and milk products have been a component of the diets of some human populations for over 6000 years, roughly 300 generations. Is it conceivable that dairy farming might have created the selective regime under which the allele for absorption was favored?

Feldman and Cavalli-Sforza (1989) used gene–culture coevolutionary models to investigate the evolution of lactose absorption. They assumed that the capacity to absorb lactose was affected by alleles (variants) of a single gene, with one particular allele allowing adults to digest milk. In addition, they modeled the cultural transmission of milk usage. The analysis suggested that whether or not the allele allowing adult lactose absorption and milk digestion achieves a high frequency depends critically on the probability that the children of dairy-product users themselves become milk consumers. If this probability is high, then a significant fitness advantage to the genetic capacity for lactose absorption will generally result in the selection of

the absorption allele to high frequency within 300 generations. However, if a significant proportion of the offspring of milk users do not exploit dairy products, then unrealistically strong selection favoring absorbers would be required for the gene for absorption to spread. In other words, differences in the strength of cultural transmission between cultures may account for genetic variability in lactose absorption. The analysis is able to account for both the spread of lactose absorption, and the culturally related variability in its incidence. Moreover, there is a broad range of conditions under which the absorption allele does not spread despite a significant fitness advantage, indicating that traditional genetic models would frequently give the wrong answer. Cultural processes complicate the selection process to the extent that the outcome may differ from that expected under purely genetic transmission.

The traditional view among the scientific community was that adult lactose tolerance in humans is an adaptation to reduced exposure to the sun, as both the sun and the enzyme lactase promote calcium absorption (Durham, 1991). However, an analysis by Holden and Mace (1997) of a phylogeny of human cultural groups using sophisticated statistical techniques found no evidence for the latitudinal theory, but strong support for the dairy-farming hypothesis. Moreover, their analysis revealed that dairy farming evolved first, which then favored tolerance to lactose, and not the other way around. Holden and Mace's analysis provides compelling confirmation of the findings of this gene–culture coevolutionary analyses.

DO GENES AND CULTURE COEVOLVE?

It is frequently suggested that genetic evolution is too slow and cultural change too fast for the latter to drive the former. In fact, selection experiments and observations of natural selection in the wild reveal that biological evolution may be extremely fast, with significant genetic and phenotypic change sometimes observed in a small number of generations. At the same time, observations of hominid stone tool technologies reveal that cultural change can be extraordinarily slow. Acheulian and Oldowan stone tools traditions remained very similar for hundreds of thousands – even millions – of years. Even cultural institutions such as labor markets can be extremely persistent, albeit on a shorter time scale. Furthermore, theoretical analyses have revealed that cultural transmission may change selection pressures to generate unusually fast genetic responses to selection in humans (Feldman and Laland, 1996). It is thus entirely feasible that genetic and cultural evolution could operate at similar rates. In fact, the past 2 million years of human evolution may even have been dominated by gene–culture coevolution. Culture can, of course, cause rates of environmental change that really are too fast for human genetic evolution to track, and it is probably doing so increasingly. In fact, in the last 25 000–40 000 years the dominant mode of human evolution has probably been exclusively cultural.

Gene–Culture Coevolution and Niche Construction

Organisms frequently choose, regulate, construct, and destroy important components of their environments, such as nests and burrows, in the process changing the selection pressures to which they and other organisms are exposed. These processes are known as niche construction (Odling-Smee, 1988). Niche-constructing traits are more than just adaptations, because they have the additional role of modifying natural selection pressures. Like natural selection, niche construction can be regarded as an evolutionary process potentially capable of generating a complementary match between organism and environment. Organisms may adapt to environments, and environments may be shaped by organisms. In humans, culture has greatly amplified our capacity for niche construction and our ability to modify selection pressures, and cultural niche construction may frequently instigate further biological or cultural change (Laland *et al*., 2000).

Standard gene–culture analyses incorporate niche construction implicitly, by assuming that some human cultural activities feed back to modify some selection pressures in human environments, and thus cultural transmission may affect the fate of some selected human genes. Generally, the relevant aspect of human selective environments is defined as cultural. For example, the trait that affected human genetic evolution in the lactose tolerance case was milk usage (Durham, 1991). Here, gene–culture theory is applicable because the link between milk usage and its genetic consequences are sufficiently simple to allow it to be modeled without bringing in any intermediate variables (Feldman and Cavalli-Sforza, 1989).

However, standard gene–culture coevolutionary models are less appropriate in more complicated situations. Take, for example, the case of Kwa-speaking yam cultivators in West Africa, who increased the frequency of a gene for sickle cell

anemia in their own population as a result of the indirect effects of yam cultivation. These people traditionally cut clearings in the rainforest, creating more standing water and increasing the breeding grounds for malaria-carrying mosquitoes. This, in turn, intensified selection for the sickle cell allele, because of the protection offered by this allele against malaria in the heterozygotic condition (Durham, 1991). Here the causal chain is so long that simply plotting the cultural trait of yam cultivation against the frequency of the sickle cell allele would be insufficient to yield a clear relationship between the cultural trait and allele frequencies (Durham, 1991). The crucial variable is probably the amount of standing water in the environment caused by the yam cultivation, but standing water is an ecological variable, not a cultural variable, and it partly depends on factors (rainfall) that are beyond the control of the population. So here the simplifying assumption of a direct link between cultural and genetic inheritance distorts reality too much to allow their interaction to be modeled in the standard way. This time the two human inheritance systems can interact only via an intermediate, abiotic, ecological variable subject to niche construction, which should be included to complete the model.

This shortcoming led Laland *et al*. (2000) to propose an extended version of gene–culture coevolutionary theory, which incorporated niche construction as a general evolutionary process. Culturally modified selection pressures are regarded as a part of a more general legacy of modified natural selection pressures bequeathed by humans to their descendants. Laland *et al*. (2001) used extended gene–culture coevolutionary models to explore the evolutionary consequences of culturally generated niche construction through human evolution. The analysis revealed circumstances under which cultural transmission can overwhelm natural selection, accelerate the rate at which a favored gene spreads, initiate novel evolutionary events, and trigger hominid speciation. Because cultural processes typically operate faster than natural selection, cultural niche construction is likely to have more profound consequences than gene-based niche construction, and is likely to have played an important role in human evolution.

Empirical Studies of Gene–Culture Coevolution

While gene–culture coevolution has a strong and rigorous theoretical foundation, it is vulnerable to the charge that it has not spawned a vigorous empirical science. Where gene–culture analyses have been applied to specific case studies they do make a variety of testable predictions; for instance, Soltis *et al*. (1995) used data on rates of population extinction in New Guinea to test Boyd and Richerson's group selection hypothesis. Yet no general empirical method has been established. The closest to a general approach is that advocated by anthropologist Bill Durham, who illustrates with compelling examples, each backed by considerable data, how variability in human behavior and society may be interpreted as resulting from interactions between genetic and cultural processes (Durham, 1991). Durham identifies five categories of interaction:

- genetic mediation, where genetic differences underlie cultural variation, as may be the case for the terms used by humans to describe color which reflect features of the human nervous system;
- cultural mediation, where culture drives genetic change, such as with the evolution of adult lactose absorption in populations that consume dairy products;
- enhancement, where culture reinforces genetic predispositions, as with the emergence of incest taboos that guard against the deleterious effects of inbreeding;
- neutrality, where memes are adopted independently of an individual's genotype, as is the case for learning different languages;
- opposition, where culture leads to maladaptive traditions, for instance the cannibalism of the Fore, a New Guinea community, which spread the deadly nerve disease kuru.

CONCLUSION

Gene–culture coevolutionary theory has rarely been subject to the same level of criticism as human sociobiology or evolutionary psychology. In fact, in the debates over human sociobiology and its progeny, it has been almost completely ignored, perhaps because of its technical nature. However, some social scientists have objected to the idea that culture can be modeled as if composed of discrete psychological or behavioral characteristics, while others have questioned the legitimacy of 'borrowing' population genetics processes to model culture, or criticized the analyses as promoting a false gene–culture dichotomy. However, for gene–culture researchers these assumptions do not represent ideological stances but are made purely for pragmatic reasons. Culture is difficult to analyze unless it is broken down into manageable units. Building on population genetic models is a reasonable place to start developing models of cultural evolution (providing the differences between

biological and cultural processes are accommodated). Gene–culture methods represent comparatively simple descriptions of the interactions between genetic and cultural processes. However, approaches that focus on a single process (be it exclusively cultural or exclusively genetic) have made the fundamental and sweeping assumption that the processes do not interact, or that there is only one process that matters.

Gene–culture coevolutionary analyses suggest that evolution in species with a dynamic, socially transmitted culture may be different from evolution in other species. Culture is a particularly effective means of modifying natural selection pressures and driving the population's biological evolution, as was the case for lactose absorption. Culture may generate new evolutionary processes, for instance cultural group selection. Moreover, cultural transmission may strongly affect evolutionary rates, sometimes speeding them up and sometimes slowing them down. Such findings suggest that traditional evolutionary approaches to the study of human behavior may not always be adequate.

References

Boyd R and Richerson PJ (1985) *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Cavalli-Sforza LL and Feldman MW (1981) *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Cavalli-Sforza LL, Feldman MW, Chen KH and Dornbusch SM (1982) Theory and observation in cultural transmission. *Science* **218**: 19–27.

Dawkins R (1976) *The Selfish Gene*. Oxford, UK: Oxford University Press.

Durham WH (1991) *Coevolution: Genes, Culture and Human Diversity*. New York, NY: Stanford University Press.

Feldman MW and Cavalli-Sforza LL (1976) Cultural and biological evolutionary processes, selection for a trait under complex transmission. *Theoretical Population Biology* **9**: 238–259.

Feldman MW and Cavalli-Sforza LL (1989) On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In: Feldman MW (ed.) *Mathematical Evolutionary Theory*. Princeton, NJ: Princeton University Press.

Feldman MW and Laland KN (1996) Gene–culture coevolutionary theory. *Trends in Ecology and Evolution* **11**: 453–457.

Guglielmino CR, Viganotti C, Hewlett B and Cavalli-Sforza LL (1995) Cultural variation in Africa: role of mechanisms of transmission and adaptation. *Proceedings of the National Academy of Science of the USA* **92**: 7585–7589.

Hewlett BS and Cavalli-Sforza LL (1986) Cultural transmission among Aka pygmies. *American Anthropologist* **88**: 922–934.

Holden C and Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* **5**: 605–628.

Laland KN, Odling-Smee FJ and Feldman MW (2000) Niche construction, biological evolution and cultural change. *Behavioral and Brain Sciences* **23**: 131–146.

Laland KN, Odling-Smee FJ and Feldman MW (2001) Cultural niche construction and human evolution. *Journal of Evolutionary Biology* **14**: 22–33.

Lumsden CJ and Wilson EO (1981) *Genes, Mind and Culture*. Cambridge, MA: Harvard University Press.

Maynard Smith J and Warren N (1982) Models of cultural and genetic change. *Evolution* **36**: 620–627.

Odling-Smee FJ (1988) Niche constructing phenotypes. In: Plotkin HC (ed.) *The Role of Behavior in Evolution*. Cambridge, MA: MIT Press.

Smith EA (2000) Three styles in the evolutionary analysis of human behavior. In: Cronk L, Chagnon N and Irons W (eds) *Adaptation and Human Behavior*. New York, NY: Aldine de Gruyter.

Soltis J, Boyd R and Richerson PJ (1995) Can group-functional behaviors evolve by cultural group selection? An empirical test. *Current Anthropology* **36**: 473–494.

Tylor EB (1871) *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom*. London, UK: John Murray.

Human Behavioral Ecology

Intermediate article

Eric Alden Smith, University of Washington, Seattle, Washington, USA
Bruce Winterhalder, University of North Carolina, Chapel Hill, North Carolina, USA

CONTENTS

Human behavioral ecology applies theory and method developed in evolutionary biology, anthropology and economics to elucidate adaptive variation in human behavior, particularly social behavior. Hypotheses about resource use, mating and parenting strategies, cooperation and competition, and life history are derived from models using a selectionist logic, and empirically tested to increase our understanding of how humans adapt to their natural and social environments.

INTRODUCTION

Human behavioral ecology (HBE) is one of several approaches in the evolutionary social sciences (Smith *et al*., 2001), other prominent ones being evolutionary psychology and cultural evolution (or meme) theory. Like these other approaches, HBE combines theory and methods from a number of different academic disciplines. From evolutionary biology it draws mathematical or graphical models anchored in basic principles of evolution by neo-Darwinian natural selection. From neoclassical economics it adopts concepts and analytical techniques such as optimization, marginal value analysis, and game theory. From anthropology, it borrows ethnographic research methods: the extended recording of behavioral observations in their immediate socioenvironmental context, often in small communities, supplemented with data collected by survey, interview or archival research. Human behavioral ecology emphasizes quantitative methods such as those characteristic of ethology (naturalistic observation of animal populations): focal follows, scan sampling, and the like.

The topics analyzed in HBE research can be grouped into three main categories (Winterhalder and Smith, 2000): production (resource acquisition and related topics), reproduction (mating and parenting), and distribution (exchange, sharing, and coercive transfers).

Behavioral ecology has developed as the behavioral branch of the larger field of evolutionary ecology, the study of evolution and adaptive design in ecological context. Evolutionary ecology emerged as a distinct field in the 1960s, and includes topics ranging from the structural and behavioral traits of organisms to the organization of ecological communities. Behavioral analyses have been an integral element of evolutionary ecology from the beginning, treating topics such as foraging strategies, mating systems, spatial organization, and competition. The first textbooks on behavioural ecology appeared in the late 1970s and early 1980s, and there is now a voluminous literature, including monograph series, dedicated journals (e.g. *Behavioural Ecology* and *Behavioural Ecology and Sociobiology*) and a widely read series of books edited by Krebs and Davies.

ECOLOGICAL SELECTIONISM

The adaptationist program in contemporary evolutionary biology assumes that natural selection has designed organisms to respond to local socioenvironmental conditions in fitness-enhancing ways. With this as a starting point, behavioral ecologists formulate and test formal models incorporating specific optimization goals, currencies, and constraints. Because their subject is behavior, and particularly social behavior with a strong cultural component, human behavioral ecologists must analyze a much more labile and causally complex set of phenomena than an evolutionist studying, for example, skeletal morphology or even avian courtship behavior.

Human behavioral ecology generally attempts to explain such complex patterns of cultural and behavioral variation as forms of phenotypic adaptation to varying social and ecological conditions. The field is less concerned with genetic variation

on evolutionary time scales than with variation in behavior that occurs within an individual's lifetime (often in minutes or hours), or that accumulates over a few generations through cultural change. The link between such phenotypic or cultural adaptation and genetic evolution is provided by positing that the former is guided by 'decision rules' – cognitive or problem-solving propensities that themselves evolved genetically through natural selection. However, the genetic basis of these phenotypic capacities is not addressed directly in HBE, which takes an agnostic view of the underlying causal mechanisms that might shape adaptive variation in behavior. Rather, the focus is on testing predictions about the match between environmental conditions or payoffs and behavioral variation, without worrying too much about developmental or learning mechanisms that create or maintain this match.

MODELS AND DECISION RULES

In common with many scientific fields, including general behavioral ecology, HBE research is strongly theory-driven. The research strategy is built around mathematical models of particular phenomena. Any given model is designed to answer a particular set of questions: for example, what is the optimal set of prey to harvest? How much should a parent invest in male versus female offspring? Models are used to generate hypotheses that can then be tested empirically, and the results of these tests indicate whether the model appears to capture correctly essential features of the phenomenon being investigated, needs significant modification, or should be discarded. As an area of research develops, sets of related models are linked together to form a body of theory (e.g. optimal foraging theory, or parental investment theory) covering a relatively broad empirical domain.

A complete HBE explanation combines models of circumstance and models of mechanism (Winterhalder, 1997). Models of circumstance ask how socioecological factors shape the costs and benefits associated with alternative behavioral strategies in a given domain. Models of mechanism attempt to specify how natural selection, or a variant such as sexual, kin, or cultural selection, will act on these costs and benefits. By combining these two elements, the HBE approach avoids some of the problems associated with functionalist explanation in the social sciences. In particular, neo-Darwinian theory identifies a restricted set of units, costs, and benefits that have a significant role in evolutionary processes (for example, ruling out strategies that increase longevity without increasing number of surviving descendants or other genetic kin).

All HBE models incorporate a goal (which the strategy under consideration is designed to optimize), a currency (for measuring the relevant costs and benefits), a set of constraints (characterizing the social and environmental context in simplified form) and a decision set (the range of behavioral options considered). Different evolutionary goals may require different optimization methods: deterministic, stochastic, or dynamic optimization, as well as game-theoretic analysis.

Human behavioral ecology usually frames the study of adaptive design in terms of decision rules, which are presumed to be panhuman adaptations that have evolved by natural selection in order to generate behavioral variation that is sensitive to environmental context. These decision rules are often conditional strategies that take the general form 'In context *X*, adopt one behavioral tactic; in context *Y*, switch to the other tactic' (and more complex variants for strategies with more than two tactics). For example, the polygyny threshold model (Figure 1) assumes that female mate choice follows the evolved decision rule 'If the bachelor suitor has at least half the resources of an already married suitor, accept his offer; otherwise, become the second wife of the married suitor.' Behavioral variation arises as individuals match their conditional strategies to their particular current socioecological settings.

THE PHENOTYPIC GAMBIT

Emphasizing generality, most HBE models strive to be as simple as possible. They seek to capture the essential features of an adaptive problem, and thus analyze complex socioecological phenomena in a relatively reductionistic fashion. Such models are thus caricatures of reality intended to be heuristic tools, rather than realistic descriptions of the cognitive or ontogenetic processes that produce human behavior. This sacrifice of realism is made in order to obtain compensating benefits (increased generality and analytical tractability).

More generally, HBE research often assumes that the details of genetic, phylogenetic, and cognitive mechanisms do not, to a first approximation, seriously constrain human adaptive responses to ecological variation. This strategic short cut, known as the 'phenotypic gambit', is taken because it makes it much easier to build and test general (widely applicable) models, focused on ultimate adaptive design. The phenotypic gambit may of course

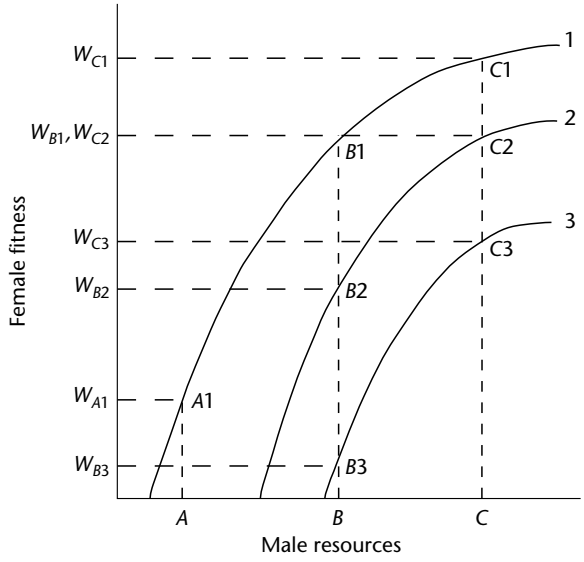


Figure 1. The polygyny threshold model. When female fitness is at least partially a function of the resources controlled by her mate, females may benefit reproductively by mating polygynously with males controlling higher amounts of resources. In this example, a female who became the third wife of a male controlling *C* amount of resources would obtain higher fitness (W_{C3}) than if she were to be married monogamously to a male controlling *A* resources, or the second wife of a male controlling *B* resources. She would have higher fitness if married monogamously to a poor male controlling *A* (W_{A1}) than being the third wife of a male controlling *B*.

be wrong in any particular case: humans may, as some evolutionary psychologists claim, be easily addicted to sugar and fats because in our ancestral environments these nutrients were rare and of high adaptive value. Ignoring such an evolved bias might lead to erroneous predictions about the adaptive value of diets in modern populations. On the other hand, humans seem to have the cognitive and cultural machinery needed to produce adaptive responses quickly to novel environmental conditions (including, in the present example, dieting regimens, gymnasiums filled with exercise equipment, and nutritional and medical knowledge for dealing with the threats posed by overeating). Hence, it is probably premature to draw any firm conclusions about the overall validity of the phenotypic gambit.

EMPIRICAL RESEARCH

Production

Research into HBE can be grouped into three broad topical areas: production, reproduction, and distribution. Analyses of production – resource acquisition behavior – draws on optimal foraging theory (OFT), a family of models initially developed by biologists borrowing heavily from neoclassical economics. These models address resource choice, time allocation, and movement between different habitat sectors or 'patches'. By far the most popular has been the prey choice model (sometimes termed the diet breadth model). It is used here to exemplify the HBE research strategy.

As with all HBE models, the prey choice model (PCM) incorporates a goal, a currency, a set of constraints, and a decision set. The PCM predictions test our assumption that foragers have the goal of choosing the set of available prey types that, under given environmental conditions, yield the maximum value per unit foraging time. Because it can be readily measured and is quite general, the currency used in the PCM is usually the net energy acquisition rate (e.g. kilojoules per hour). Net acquisition rate is appropriate if foragers are time-limited (i.e. gain more from freeing time for other activities than from harvesting additional resources), energy-limited (i.e. gain more from additional units of harvest than from reduced foraging time) or face foraging conditions that expose them to hazard levels greater than those they experience when not foraging (e.g. predation, higher risk of injury, or climate stress). Thus, contrary to common intuition, energy return rate may be adaptively important even if food energy is not strictly limiting.

The constraints of the PCM include those endogenous to the forager, such as available information, cognitive capacities, and technology, as well as exogenous factors such as the availability, behavior and nutritional value of the potential prey resources. In any given application all constraints but one are considered to be relatively fixed, and the remaining constraint becomes the independent variable that predicts choices among the decision set. For example, the independent variable might be the encounter rate with various resource types, the foraging technology for pursuing them, or information processing capabilities, depending on the researcher's interest. The decision set specific to the PCM consists of all the possible combinations achieved by stepwise addition of resources which have been ranked by their pursuit and handling profitability.

The model predicts that diet breadth will shrink as high-ranking prey become more abundant, and that increased abundance of a resource outside of the optimal set will not cause it to be harvested. These and other predictions, as well as ones

derived from other OFT models, have been tested among a variety of hunter-gatherer populations (Kaplan and Hill, 1992). The theory is relatively successful in explaining observed patterns of prey choice and patch use, as well as shifts in subsistence patterns in response to such factors as changes in technology, climatic fluctuations, anthropogenic prey depletion, and human population growth. The OFT framework has illuminated why children harvest different resources from adults in the same society, why hunters often fail to conserve prey species, why some resources are processed at the harvest or kill site and others transported whole back to camp (of great significance for interpreting archaeological data), and even why foragers in various parts of the world have independently engaged in a process of plant and animal domestication leading to agricultural production systems (Winterhalder and Smith, 2000).

Given the universal and recurrent short-term need for metabolic energy, it is reasonable to assume that foraging strategies that maximize the net acquisition rate of energy while foraging have higher fitness, at least within broad limits. We should expect selection to favor cognitive mechanisms and culturally inherited rules of thumb that produce behaviors keyed to this goal. However, most optimal foraging models are general enough that the currency could be any rate measure of resource value – protein capture, raw material value, monetary return, or prestige. For instance, application of the PCM has been used to examine the circumstances under which sexual selection might favor different currencies for males and females (Bliege Bird, 1999).

Distribution

Foraging models concern themselves with the short-term production decisions of individuals. However, for humans and their hominid ancestors, the harvesting and consumption of resources generally occur in a social group, a context that adds a host of theoretical and empirical challenges.

Cooperative subsistence efforts may offer several advantages: increased per capita resource harvest rate, reduced variation in harvest rates, reduced losses to competitors, and increased vigilance and predator detection. However, group foraging can also increase resource depletion and competition; and even where cooperation is beneficial, modeling has shown that optimal group size itself may be unstable owing to conflicts of interest between existing members and potential joiners (Figure 2). Once groups form they provide the context for complex social dynamics, including competition and conflict.

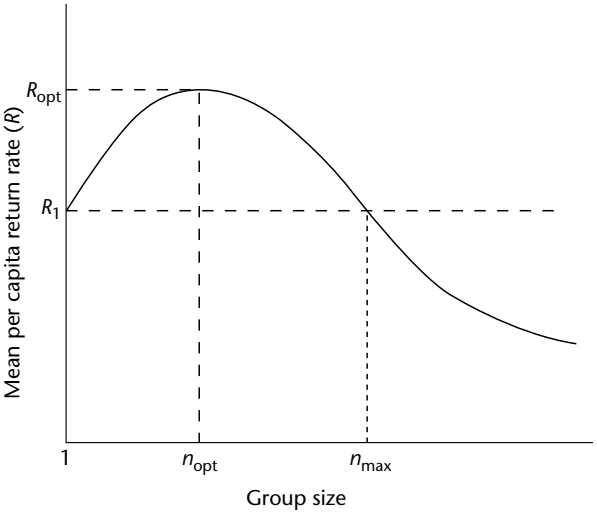


Figure 2. Optimal group size and member–joiner conflict. When the per capita return rate R (e.g. kilo joules per forager per hour) reaches a maximum (R_{opt}) at a group size of $n_{opt} > 1$, then members of a group will suffer a decline in their share of group production if additional individuals join, but potential joiners have an incentive to join as long as their share will be greater than what they can obtain through solitary production (R_1), up to the equilibrium group size n_{max}.

The conditions favoring different kinds of resource transfers (sharing, scrounging, and so on) have been the focus of considerable research in HBE. Unlike most other primates, human foragers and agriculturalists often harvest resources of sufficient 'package size' (e.g. large game) or in sufficient bulk (e.g. an agricultural crop) that some combination of transfer to those without the resource or storage for later use is likely. There are a variety of models to study this, each making somewhat different assumptions about the socioecological circumstances and the evolutionary mechanisms that may shape resource transfers.

All resource transfer models address a common circumstance, the unsynchronized acquisition of valuable resource packets by individuals within a group. The models differ primarily in the additional circumstances specified (e.g. group size, information flow, frequency of interactions among the individuals involved, the nature of the resource) and in the evolutionary mechanism they invoke (e.g. individual, kin, sexual, group or cultural selection) (Winterhalder, 1997). Simple individual-level selection will generate transfer by scrounging (also known as tolerated theft) when those not possessing a resource packet benefit

more by taking portions than the holder can benefit by defending them. Voluntary resource sharing is usually modeled in terms of the delayed, cost–benefit calculus of reciprocal altruism. For example, if resource harvest is unpredictable and relatively unsynchronized, harvesters might benefit by pooling the catch and thereby minimizing subsistence risk (variance in resource consumption).

In by-product mutualism the individual discovering or possessing a resource obtains a net gain as a result of encouraging others to participate in its capture, defense or consumption. In this case short-term cooperation is mutually beneficial, so defection or cheating, a potential threat to reciprocal altruism, is not an issue. One form of by-product mutualism is costly signaling: by successfully harvesting and then distributing difficult-to-capture resources, individuals reliably signal their prowess, benefiting themselves as well as potential allies, mates or competitors, who gain both food and useful information about the provider (Smith and Bliege Bird, 2000). In trade, individuals swap unlike resources or services because both will gain by doing so. Finally, inclusive fitness selection should lead to transfers through kin-provisioning that balance costs and benefits against degrees of relatedness.

There are several empirical studies assessing the relative importance of one or more of the proposed resource transfer models. Collectively, these studies indicate that transfer behaviors are much more diverse and context-specific than has been appreciated in the standard ethnographic literature. These studies also suggest that transfer behaviors are probably multicausal in origin, the result of several selective pressures whose relative importance depends on the situation.

Reproduction

While classical sociobiology analyzed reproductive behavior in terms of factors inherent in sexual reproduction, such as genetic relatedness and gamete asymmetry, HBE analyzes variation in reproductive behavior as a function of local ecological context. In contrast to evolutionary psychology, HBE posits that this variation involves phenotypic tracking of current circumstances, rather than the playback of relatively fixed behavioral routines specific to species, sex or age that were adaptive in our remote evolutionary history. Nevertheless, HBE approaches overlap considerably with these other two evolutionary traditions and with certain versions of cultural evolution, as well as with less explicitly neo-Darwinian fields such as demography and reproductive ecology.

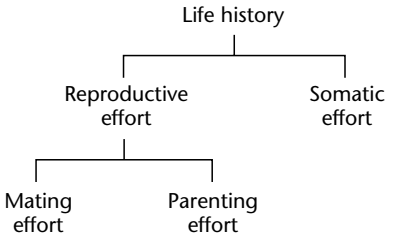


Figure 3. The domains of adaptive effort defining major life history trade-offs.

In HBE, analyses of reproductive behavior can be divided into three topics: life history, mating, and parenting (Borgerhoff Mulder, 1992). Life history is the broadest category, subsuming in principle the entire range of activities involved in survival and reproduction (Figure 3). The central concept in life history theory is the principle of allocation (Hill and Hurtado, 1996): any effort (time, energy, resources) allocated to one domain (for example, enhancing one's own survival and maintenance) cannot be allocated to another domain (for example, reproduction). Mating and parenting together constitute reproductive effort, and models often assume that effort allocated to mating cannot be allocated to parenting (and vice versa). Thus, the principle of allocation can be used to define a set of key trade-offs that are amenable to optimization models.

Mating strategies

The distribution of key resources strongly shapes the behavior of males and females, generally through different routes. If some males can monopolize resources necessary for female survival and reproduction, they can use this control to attract mates, or to compete with other males for social dominance. Polygyny and increased variance in male mating success is the predicted result. Male resource control coupled with female mate choice is the basis for the polygyny threshold model mentioned earlier (see Figure 1). The outcome predicted by the simplest versions of this model is an 'ideal free distribution', in which the number of mates per male will match the resources each can offer, and female fitness will be equal across mateships.

The polygyny threshold model has received broad support in empirical tests among a variety of human societies, though with various qualifications (Borgerhoff Mulder, 1992; Winterhalder and Smith, 2000). For instance, the male-controlled

resources may be political rather than economic. Male coercion (especially by agnatic kin groups) may severely constrain female choice. Females mated polygynously may face reduced reproductive success due to competition with co-wives, though this may be compensated for in the next generation if the sons of polygynously married women have increased chances of inheriting wealth and mating polygynously themselves.

Polygyny has been the preferred marriage form for the great majority of societies in the ethnographic record. Even in putatively monogamous societies, extramarital mating and remarriage biased towards wealthier or more powerful males creates a situation of effective polygyny. Human behavioral ecologists have also analyzed monogamous systems, especially those involving social stratification and dowry, as well as the rare but intriguing polyandrous case (though serial polyandry, in which women remarry to find better mates, is presumably much more common, but has only recently begun to be studied).

Parenting

Whatever form the mating system takes, human offspring require extensive and extended parental care. This parental investment begins with gestation and, among humans, can continue beyond the parent's death (through the bestowing of land, wealth, and other forms of property inheritance). Human behavioral ecology analyses ask how the amount and timing of such investment might vary according to social and environmental constraints. Most research falls into one of three categories: birth spacing, differential investment in offspring (by sex or expected reproductive value), and interactions between mating and parenting.

Birth spacing

If parental time and resources are finite, higher fertility rates should result in less parental investment per offspring and may eventually reduce total reproductive success. This insight provided the basis of the optimal clutch-size model first developed by behavioral ecologists to study avian reproduction, but easily generalized to apply to any species with parental investment, including humans (Figure 4(a)). This model predicts that beyond a certain point, increased fertility (larger clutches, or shorter interbirth intervals) will result in lowered overall parental reproductive success.

Blurton Jones (1986) used this approach to show that among the !Kung San hunter-gatherers of southern Africa, interbirth intervals much shorter than the actual mode of 4 years resulted in increased offspring mortality, sufficient to cause a net loss in expected reproductive success. At least one careful attempt to replicate the !Kung results among Ache foragers of Paraguay failed, possibly because Ache offspring mortality is less sensitive to variation in interbirth interval (Hill and Hurtado, 1996). Alternatively, it may be that in many cases the relationship between fertility and offspring survival is confounded by phenotypic correlation,

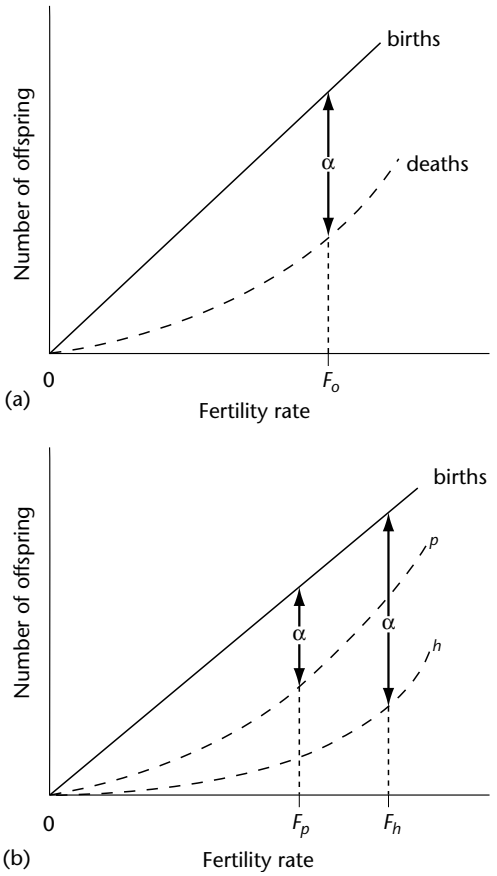


Figure 4. A graphical model of optimal fertility rate. Solid lines represent fertility rate (births per unit time), while dashed lines represent mortality as a function of fertility rate, the latter curving upwards to reflect the effect of reduced parental investment per offspring. The net difference between these is the number of surviving offspring, with a local maximum of α. The model assumes that selection favors maximizing α given the constraints that a parent faces, and hence favors an optimal fertility rate F_o. (a) Fertility rate of a single parent. (b) Comparison of two parents with different constraints and hence different offspring mortality curves as a function of fertility (parental investment). A poorly endowed parent suffers higher offspring mortality (P) at a given fertility rate, and hence a lower optimal fertility rate F_o than a higher-quality parent (h) with optimal fertility rate F_o.

which generally will mask the predicted functional relationship of the optimal clutch-size model (Figure 4(b)).

Phenotypic correlation occurs when hidden heterogeneity in uncontrolled variables confounds the effect of the causal variable under investigation. For example, wealthy individuals might tend to have more expensive houses and more expensive cars (a phenotypic correlation), even though we have good reason to expect a negative correlation between investment in houses and investment in cars due to the fact that the same dollars cannot be spent on both. If selection has designed the reproductive system to adjust facultatively to situational constraints, then interbirth intervals will be short when the parent's resources are relatively abundant and long when their condition is poor; there is abundant evidence for such adaptive variability in human reproductive ecology. Solutions to this problem include multivariate analysis, use of historical data to track functional links between birth spacing and resources, and experimental manipulation (the latter being unlikely in the human case).

Differential investment

Parental investment affects a child's health, survival and future mating success, and thus the parents' inclusive fitness. Parental fitness payoffs depend on three sets of variables: (1) the genealogical relatedness between parent (or other caregiver) and offspring, (2) the effect of investment on the expected reproductive value of the offspring (as well as present and future siblings), and (3) the effect of investment on the caregiver's own reproductive value. Sets (2) and (3) are more directly affected by ecological variables, and hence are at the center of HBE analyses.

Postpartum parental investment decisions range chronologically from whether or not to keep the child – the alternatives being infanticide, abandonment, or adopting out – to delegation of care to others, to investments in nurturing and education, to wealth transfers that often accompany marriage (bride price or dowry), to any legacy the child may receive upon death of the parents. Fitness payoffs for nearly all of these decisions may differ according to the sex of the offspring. A variety of parental investment hypotheses have been the subject of ethnographic and historical research in HBE (Table 1).

Parenting-mating interactions

Following the lead of primate and avian behavioral ecologists, HBE researchers have begun to consider when paternal care and resource provisioning, rather than simply being forms of parental investment, may be designed to attract or maintain a relationship with a mate. For example, men in Albuquerque, New Mexico, as well as Xhosa men in South Africa, invest more time and resources in stepchildren who are offspring of their current mates than they do in stepchildren from former relationships (though less than they invest in genetic offspring under comparable circumstances). This pattern does not match predictions from parental investment theory, but it does make adaptive sense if viewed as investment in maintaining a current mating relationship (i.e. as mating effort). Unravelling interactions such as these is a promising area for future HBE research.

Life History Strategies

Life history is a broad topic, subsuming in principle the entire range of activities involved in survival and reproduction. In practice, life history analyses center on a few key decision categories: the timing of growth and maturation, subadult and adult dispersal strategies, the onset of reproduction, the timing of reproductive events (e.g. birth spacing, weaning, menopause), mortality patterns, and senescence. Most HBE work on life history thus far has focused on four topics: links between production and reproduction; reproductive effort and maturation; menopause and extended human life span; and evolutionary analysis of the so-called 'demographic transition' (reduction in fertility and family size with modernization). The first three of these topics are given exemplary treatment in an extended case study by Hill and Hurtado (1996), while the fourth is reviewed by Borgerhoff Mulder (1998).

CONCLUSION

Human behavioral ecology applies neo-Darwinian theory and ethnographic and ethological methods to elucidate adaptive variation in human behavior, particularly social behavior. It combines elements from a number of different academic disciplines, including anthropology, economics, ethology, and evolutionary biology. Mathematical or graphical models anchored in basic principles of evolution by natural selection are used to derive hypotheses concerning how humans adapt to their natural and social environments. Predictions are then tested with ethnographic, historical, and archeological data. Human behavioral ecology assumes that human decision-making is guided by evolved 'decision rules' or conditional strategies, but focuses

Table 1. Human behavioral ecology analyses of parental investment (PI)

Predictions tested	*Representative studies (locales and authors)*[a]
Reduced genetic relatedness or reproductive conflicts of interest lead to lower PI	Cross-cultural (Daly and Wilson); USA and South Africa (Anderson *et al*.); Mali (Strassman)
Offspring with reduced prospects of survival receive lower PI	Cross-cultural (Daly and Wilson)
Parents reduce PI per offspring as number of offspring increases	Tanzania (Borgerhoff Mulder); Paraguay (Hill and Hurtado)
Parental investment per child is increased when marginal benefit of PI is higher	Hungary (Bereczkei); USA (Kaplan *et al*.)
Adoption and fostering are adaptively modulated according to parental circumstances	Cross-cultural (Silk); Botswana (Pennington)
Mothers able to delegate nursing and infant care gain increased reproductive success	Hungary (Bereczkei); Europe (Hrdy); Ifaluk, Micronesia (Turke)
Postmenopausal women allocate resources and care to grandchildren or other close relatives	Paraguay (Hill and Hurtado); Tanzania (Hawkes *et al*.)
Parents with less resources preferentially invest in offspring (usually daughters) with lower variance in expected RS	Historic Portugal (Boone); Tanzania (Borgerhoff Mulder); North America (Gaulin and Robbins); USA (Judge and Hrdy); North America (Mealey and Mackey)
Offspring of the sex that faces better adult economic opportunities receive higher PI	Cross-cultural (Hewlett); historic Sweden (Low *et al*.); historic Germany (Voland *et al*.)
Offspring of the sex that has greater probability of contributing to future support of siblings receive higher PI	Tanzania (Borgerhoff Mulder); cross-cultural (Hewlett); Paraguay (Hill and Hurtado); Canada (Smith and Smith)
Offspring of the sex that has greater probability of competing with siblings for resources or mates receive lower PI	Tanzania (Borgerhoff Mulder); East Africa (Mace); historic Germany (Voland *et al*.)
If daughters (or sons) have greater future mating opportunities, they receive higher PI	Historic Asia (Dickemann); historic Portugal (Boone); Hungary (Bereczkei and Dunbar); Kenya (Cronk)
If daughters (or sons) are better able to claim and hold political power, they receive higher PI (especially inheritances)	North America (Hrdy and Judge); historic Portugal (Boone); cross-cultural (Hewlett)
Parental resources with increasing marginal benefits to offspring characterized by unigeniture (single heirs)	Historic Portugal (Boone); historic Germany (Voland *et al*.)
Increasing marginal benefit of biparental care leads to increased pair-bond stability	Paraguay and Venezuela (Hurtado and Hill); Tanzania (Blurton Jones *et al*.)

[a]Full references provided in Winterhalder and Smith (2000). PI, parental investment; RS, reproductive success.

on facultative behavioral outcomes and adaptive consequences rather than on the underlying cognitive mechanisms or ontogenetic processes. The topics analyzed in HBE research fall into three main categories: production (resource acquisition and related topics), reproduction (mating and parenting), and distribution (exchange, sharing, and coercive transfers).

References

Bliege Bird RL (1999) Cooperation and conflict: the behavioural ecology of the sexual division of labor. *Evolutionary Anthropology* **8**: 65–75.

Blurton Jones N (1986) Bushman birth spacing: a test for optimal interbirth intervals. *Ethology and Sociobiology* **7**: 91–105.

Borgerhoff Mulder M (1992) Reproductive decisions. In: Smith EA and Winterhalder B (eds) *Evolutionary Ecology and Human Behaviour*, pp. 339–374. Hawthorne, NY: Aldine de Gruyter.

Borgerhoff Mulder M (1998) The demographic transition: are we any closer to an evolutionary explanation? *Trends in Ecology and Evolution* **13**: 266–270.

Hill K and Hurtado AM (1996) *Ache Life History: The Ecology and Demography of a Foraging People*. Hawthorne, NY: Aldine de Gruyter.

Kaplan H and Hill K (1992) The evolutionary ecology of food acquisition. In: Smith EA and Winterhalder B (eds) *Evolutionary Ecology and Human Behaviour*, pp. 167–201. Hawthorne, NY: Aldine de Gruyter.

Smith EA and Bliege Bird RL (2000) Turtle hunting and tombstone opening: public generosity as costly signaling. *Evolution and Human Behaviour* **21**(4): 245–261.

Smith EA, Borgerhoff Mulder M and Hill K (2001) Controversies in the evolutionary social sciences: a guide for the perplexed. *Trends in Ecology and Evolution* **16**: 128–135.

Winterhalder B (1997) Gifts given, gifts taken: the behavioural ecology of nonmarket, intragroup exchange. *Journal of Archaeological Research* **5**: 121–168.
Winterhalder B and Smith EA (2000) Analyzing adaptive strategies: human behavioural ecology at twenty-five. *Evolutionary Anthropology* **9**: 51–72.

Further Reading

Betzig LL (ed.) (1997) *Human Nature: A Critical Reader*. New York, NY: Oxford University Press.
Blaffer Hrdy S (1999) *Mother Nature: A History of Mothers, Infants and Natural Selection*. New York, NY: Pantheon Books.
Cronk L, Chagnon N and Irons W (eds) (2000) *Adaptation and Human Behaviour: An Anthropological Perspective*. Hawthorne, NY: Aldine de Gruyter.
Hill K (1993) Life history theory and evolutionary anthropology. *Evolutionary Anthropology* **2**: 78–88.
Krebs JR and Davies NB (eds) (1997) *Behavioural Ecology: An Evolutionary Approach*, 4th edn. Oxford, UK: Blackwell.
Low BS (2000) *Why Sex Matters: A Darwinian Look at Human Behaviour*. Princeton, NJ: Princeton University Press.
Smith EA and Winterhalder B (eds) (1992) *Evolutionary Ecology and Human Behavior*. Hawthorne, NY: Aldine de Gruyter.
Voland E (1998) Evolutionary ecology of human reproduction. *Annual Review of Anthropology* **27**: 347–374.

Human Brain, Evolution of the

Introductory article

Michael C Corballis, University of Auckland, Auckland, New Zealand

CONTENTS

The human brain has evolved to be about three times as large as predicted in a primate of our body size. This may explain such uniquely human attributes as language and theory of mind.

INTRODUCTION

The brain is the seat of cognition. Without it we would have no consciousness, no perception and no memory. As humans, we are blessed with extraordinarily large brains, containing around 10 000 million neurons, each of which may connect to 100 other neurons. The ways in which these neurons receive information from the environment, interact with each other, and generate action provide the basis for our uniquely human intelligence.

BODIES, BRAINS, AND ENERGY

The body, including the brain, requires energy, which is why we must eat. This energy requirement can be measured in terms of the *basal metabolic rate* (BMR), which is the rate at which energy obtained from food is used by an animal at rest in a temperature-neutral environment – that is, an environment which is neither too cold nor too hot, so the animal does not have to expend energy trying to warm up or cool down. The BMR depends on the body weight of the animal, but not in a straightforward linear fashion. Rather, the BMR in mammals is proportional to body weight raised to the power of 0.75. Since this power is less than one, it means that energy requirements per unit body weight are lower for larger animals than for smaller ones. Figure 1 illustrates the extent of variation in brain size in a sample of mammals.

Brain size also tends to increase with body size, and interestingly the relationship is the same as that relating BMR to body weight – that is, the brain size of mammals is roughly proportional to body weight raised to the power of 0.75. This reflects the fact that in most animals the brain is largely concerned with routine bodily functions. However, the human brain is exceptionally large for an animal of our size, as we shall see below. The excess of brain volume over that required for basic bodily functions is presumably dedicated to functions that we might think of as 'intelligent'.

Yet having a large brain comes at a cost. The brain is especially demanding of energy, and the problem is compounded in our own species, where the brain is out of proportion to our body weight. In an adult human, the brain weighs only about 2% of the total body weight, but takes up around 20% of the BMR, and in infancy the latter proportion is over 60%. In the adult chimpanzee, the proportion is only about 8%. Moreover, the brain has no energy reserves, and is therefore continually dependent on blood supply. If the supply is interrupted for only a few minutes, brain tissue can die and can bring about permanent dysfunction.

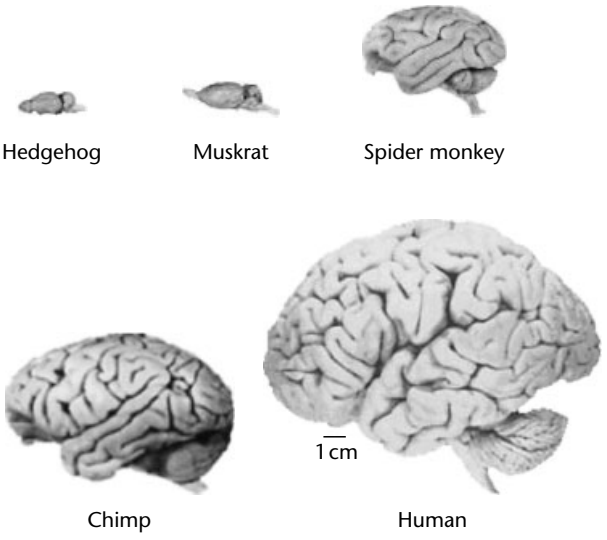


Figure 1. A sample of mammalian brains, equivalently scaled.

Cerebral bloodflow accounts for around 12–15% of the total output from the heart, and the brain consumes around 20–25% of the oxygen intake, and 70% of the total consumption of glucose (which neurons require in order to function). The supply remains largely constant, but is distributed to different parts of the brain according to demand. Measurement of relative bloodflow to different parts of the brain forms the basis of brain-imaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), which show which parts of the brain are active during different types of mental processing.

EVOLUTION OF LARGE BRAINS

Although we are large-brained creatures, there are animals with larger brains. Whales lead the way, with brains weighing a massive 6800 g, followed by elephants (about 4700 g) and then dolphins (at around 1700 g). In humans, the male brain typically weighs about 1440 g and the female brain weighs about 1230 g. This does not mean that men are more intelligent than women. Rather, it is a reflection of the fact that men are typically slightly larger than women, and simply require more brain in order to maintain their basic metabolism. In any case there is no evidence that intelligence in humans varies with brain size, even though brain size itself varies quite markedly from one person to another. Nevertheless, it is fairly generally accepted that variations in intelligence between species do depend, at least in part, on brain size.

However, to compare species in terms of the brain volume available for intelligent behavior, we need to correct for variations in body size. One way to do this is in terms of what has been called the *encephalization quotient*, which is the ratio of actual brain size to that expected of an animal of equivalent body size. In these terms, humans do indeed top the list, with a quotient of about 7.44. The dolphin, interestingly, comes second (with a quotient of about 5.31), followed by our closest relative, the chimpanzee (at about 2.49). On this scale, the elephant is well down in the league, with a ratio of about 1.87, and the rat, which once formed the basis for much psychological theory, has a quotient of only about 0.40. Nevertheless, the phrase 'rat-like cunning' suggests that this figure might well underestimate that resourceful creature, and the quotients probably provide only an approximate ranking at best. Even so, they provide some reassurance that we humans are the most intelligent of earthly species.

The Brain in Primate Evolution

Humans belong to an order of animals known as primates, which includes monkeys and apes. More specifically, we belong to a group of large-bodied primates known as great apes, whose present-day members include orangutans, gorillas, chimpanzees, bonobos (or pygmy chimpanzees) and humans. The great apes have all emerged relatively recently in primate evolution, and the most recent common ancestor of ourselves and the chimpanzee and bonobo probably lived around 5 or 6 million years ago.

Primates are generally intelligent, curious, manipulative creatures, and by mammalian standards they are relatively large brained. However, even by primate standards our own brains are exceptionally large. It is estimated that the human brain is about three times as large as one would expect of a primate of the same body size, and no other primate even comes close. For example, a typical human adult may weigh around 70 kg, but our brains have a volume of around 1200 to 1400 mL. The orangutan is around the same size, but its brain volume is only around 400 mL while our nearest relative, the chimpanzee, has a body weight of around 55 kg but has a brain volume of some 340 mL. It is fairly likely that the added brain size has something to do with such uniquely human capacities as language and what is known as 'theory of mind' (i.e. the ability to take the mental perspective of others).

Another way of calibrating the evolution of the brain focuses on the neocortex, which makes up the outer layers of the brain, and which emerged in mammalian species and represents the bulk of the brain in humans. It is especially critical to higher mental processes such as language, perception, imagination, and memory. It has been suggested that the ratio of the size of the neocortex to that of the rest of the brain might provide a more useful index of brain evolution than the overall size of the brain. This ratio has been termed the *neocortical ratio*. Humans have the largest neocortical ratio, at 4.1, clearly ahead of our close relative the chimpanzee, with a ratio of 3.2, although the gap seems smaller than that implied by the encephalization quotient, and might perhaps restore a little humility to our evaluation of ourselves. In gorillas the neocortical ratio is 2.65, and in orangutans it is 2.99. Across different primate species, the neocortical ratio is roughly proportional to the size of the social groups to which animals belong, which suggests that intelligence may be partly driven by the complexities of social interactions. This idea underlies

what has been termed 'Machiavellian intelligence,' and is discussed more fully below.

The Brain in Hominin Evolution

The family of species that split from the branch leading to chimpanzees and bonobos around 5 or 6 million years ago is known as the *hominins*. Around 20 different hominin species have now been tentatively identified from their fossil remains, but with the exception of *Homo sapiens* all of them have become extinct. These species have been classified into several genera, including *Ardipithecus* (possibly the first of the hominins to emerge), *Australopithecus*, *Paranthropus*, *Praeanthropus*, and *Homo*. Although there is considerable debate about the naming and classification of these species, the common characteristic that distinguished them from their great ape cousins is that they were bipedal. In other words, whereas chimpanzees and bonobos move around on all fours, in a style known as knuckle-walking, the hominins walked upright on two legs.

However, in most other respects the early hominins were like great apes. When corrected for body size, their brains were of about the same size as that of a chimpanzee. The dramatic increase in brain size did not begin until the emergence of the genus *Homo* around 2.5 million years ago. Larger brains posed a special problem because the bipedal posture imposed limits on the size of the birth canal. This meant that the infants were effectively born prematurely, before their brains had grown too large to pass through the birth canal, and as every mother knows it is still a difficult passage – which is why it is called labor. It is estimated that human infants are born around nine months prematurely relative to what would be expected of a primate of our size. Whereas the brain of a newborn chimpanzee is about 60% of its adult size, that of a newborn infant is only about 24% of its adult dimensions. It is likely that this premature birth goes back at least 1.6 million years to *Homo erectus*, whose brain size was already more than twice that of the chimpanzee.

However, premature birth proved to be an unexpected bonus, since it means that most of the growth of the brain takes place outside the womb, where it is exposed to environmental input. Since the brain is most receptive to learning while it is still growing, it can be shaped by the environment to a larger extent than is possible in other great apes, such as the chimpanzee. This may partly explain why humans have developed skills such as language to such a high level of complexity. As with many questions about human evolution, it is not clear whether premature birth was merely a by-product of increasing brain size, or whether it was one of the adaptive features that drove the selection of larger brains.

Some of the characteristics of hominin behavior that can be inferred from fossil evidence and other artefacts probably reflect this increase in brain size and the concomitant extension of postnatal growth. Manufactured stone tools date from around 2.5 million years ago, which is also the earliest known date of *Homo rudolfensis*, the first member of our genus so far identified, although stone tools are more firmly associated with a slightly later species, *Homo habilis*. Moreover, up until this time the hominins were confined to Africa, mostly to the east of the Great Rift Valley, but around 2 million years ago they began to migrate out of Africa, starting with the migrations of *Homo erectus* to Asia. Migrations and the gradual development of manufacture were no doubt associated with the increase in brain size, and perhaps the emergence of more complex thought, such as language and the ability to plan complex activities in advance.

Later waves of migrants reached Europe, probably via Asia, and were to form the Neanderthal population who survived in Europe until around 30 000 years ago. Our own species, *Homo sapiens*, probably emerged in Africa around 150 000 years ago. There is recent evidence that the ancestors of present-day *Homo sapiens* remained in Africa until as recently as about 52 000 years ago, when they began to migrate. These and subsequent migrants eventually replaced all of those who had migrated earlier, including the Neanderthals in Europe, *Homo erectus* in South-East Asia, and probably members of their own species in Europe and Asia. They were eventually to populate the entire globe.

It is not clear why these late-migrating members of our species were able to achieve such dominance. It is sometimes suggested that there was some biological change that gave them an edge, but this was unlikely to have been a further increase in brain size. The evidence suggests that the Neanderthals actually had slightly larger brains, although they also had slightly larger bodies. It is perhaps more likely that *Homo sapiens* had developed superior technology and representational abilities, as reflected in cave drawings in Europe that date from around 35 000 years ago, as well as other evidence of advanced technology. These advances need not imply larger brains, but they may have been due to a combination of cultural influences and perhaps subtle changes in brain organization.

THE BRAIN OF *HOMO SAPIENS*

It is wrong to think that brains became larger or more complex simply as a natural consequence of evolution itself. Natural selection does not inevitably lead to greater size or complexity. After all, the dinosaurs did not survive, and many small-brained creatures, such as ants, have adapted successfully to existence on the planet and may well out-survive *Homo sapiens*. Moreover, we are the only surviving species of the large-brained genus *Homo*, and even the Neanderthals, who had larger brains than our own, became extinct. The increase in brain size and the intellectual capacities that go with it were almost certainly adaptations to specific and perhaps unusual environmental challenges, and not an inevitable consequence of evolving systems. What might those challenges have been?

To answer this question, we need to go back to our primate heritage and consider what changes have occurred since that time, and during our hominin past. For most of their existence, primates were adapted to life in the trees. We still exhibit characteristics from our ancient tree-dwelling past, such as the use of the hands for grasping and the ability to raise the arm directly above the head, still enabling us to swing from branches. Around 2.5 million years ago there was a global shift to cooler climates and a shrinkage of forested areas in eastern Africa, as a consequence of which our forebears were increasingly thrust into more open, savannah-like areas. Having been preadapted to the relative safety of the forest, they were relatively ill-equipped physically for life on the savannah, a landscape populated by dangerous killers such as saber-toothed cats and hyenas. It was the change from one type of environment to another that may have shaped the unique characteristics of the human mind.

The life of our forebears on the savannah largely coincides with the epoch known as the Pleistocene, dating from about 1.64 million years ago until about 10 000 years ago. Evolutionary psychologists have argued that selective pressures operating during this epoch formed most of the distinctive characteristics of the human mind. Our hominin ancestors could not compete physically with professional killers such as saber-toothed cats, nor could they easily flee like fleet-footed antelopes from danger. In order to survive they had to live on their wits, occupying what has been called a 'cognitive niche'. At first they probably scavenged for food, but they then became increasingly proficient at hunting for live prey. They evolved a hunter–gatherer mode of existence, and indeed a few hunter–gatherer societies persist to this day. This type of existence would have depended on cooperation and communication, favoring the selection of larger brains.

Compared with the brains of other primates, the human brain is not only larger, but it is also structured differently. For example, it has been estimated that the prefrontal cortex is about double the size one would expect of a typical primate with a human-sized brain. This is why our foreheads are higher and more protruding than those of the other great apes. The prefrontal cortex is involved in complex planning of action, and includes areas concerned with language and so-called 'theory of mind' (see below). Other parts of the brain, such as the visual areas in the occipital lobes at the rear of the brain, the primary auditory cortex in the temporal lobe, and the motor and premotor cortices just anterior to the central fissure, may be proportionally reduced. In short, the areas of brain that are devoted to basic sensorimotor functions may not have changed much at all, with the major increase occurring in the prefrontal cortex and perhaps in other areas, such as the parietal lobes, that are concerned with higher-level thought processes which are relatively independent of the immediate input or output.

The Social Brain Hypothesis

The structure of the brain may have changed in much more specific ways as a result of natural selection during the Pleistocene epoch. Evolutionary psychologists have proposed what might be termed the 'social brain hypothesis', in which the human brain is likened to a society in which different members each develop their own different trades and professions. By analogy, the mind is conceptualized as consisting of independent processing units known as 'modules'. This is also sometimes called the Swiss-army-knife theory, in which the mind is likened to a knife with a collection of blades, each of which serves a different purpose. Although it is often supposed that these mental modules map on to specific brain regions, this is undoubtedly an over-simplification, since individual modules may involve different brain areas, and specific brain areas may contribute to different modules. Nevertheless, it has generally proved useful to identify different brain areas in terms of their specific psychological functions, while at the same time bearing in mind that mental processes typically involve multiple areas.

For example, language seems to consist of a specialized module that is more or less independent of

other aspects of thought, yet which involves a number of dispersed areas of the brain, mostly in the left cerebral hemisphere. There are also submodules within language. One of these is concerned with grammar. Damage to an area in the left prefrontal cortex can produce difficulties in both producing and understanding grammatical aspects of language, a condition that is known as *agrammatism*. This area is called Broca's area, after the nineteenth-century French physician who first discovered its role in speech. Damage to another area around the junction of the left temporal and parietal lobes produces difficulties in comprehension, but may leave grammar more or less intact, so that people with damage to this area of the brain produce meaningless but grammatically correct speech. This area is known as Wernicke's area, and it may be part of a submodule concerned with the comprehension of language (see Figure 2).

There may also be a face-recognition module that gives us our remarkable ability to recognize the hundreds if not thousands of people whom we know by sight. This is all the more extraordinary when we consider that human faces are all broadly similar to one another, unlike most of the other objects in the world that we recognize. Brain damage can sometimes result in a very specific loss of this ability, while the ability to recognize other objects remains intact. This condition is called *prosopagnosia*.

There may also be a 'theory of mind' module, probably involving the frontal lobes of the brain. For example, it has been argued (somewhat controversially) that the disorder known as autism is well described in terms of an inability to empathize with others, or to understand what is going on in other people's minds. Autistic children often have relatively normal language, including grammatical competence, but they are deficient in that aspect of language, known as *pragmatics*, which is concerned with understanding the speaker's intentions or tailoring one's own speech to the expectations and understanding of listeners. In short, the apparent deficit in theory of mind of autistic children appears to affect their language as well as their general social behavior.

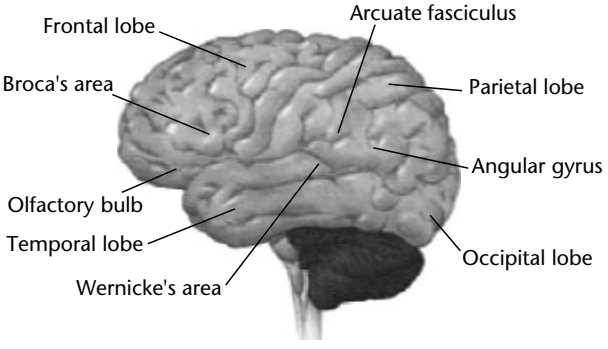


Figure 2. Major areas of the brain, including portions specialized for language.

It has been proposed that social pressures during the Pleistocene epoch led to what has been called 'Machiavellian intelligence'. People sometimes deliberately set out to deceive, and human interactions often involve a subtle interplay between trust and deception, or between coalition and conspiracy. Instances of apparently spontaneous tactical deception have also been recorded in the great apes, although rarely if ever in other primates, and humans undoubtedly surpass all other primates in the frequency and complexity of ways in which they deceive one another, as well as in the ability to detect deceptive practices.

However, there is a caveat to the social brain hypothesis. In its extreme form, it lends itself too readily to the postulation of new modules to account for every neurological deficit, and the often over-facile account of how each such module might have been selected during the hunter–gatherer phase in the Pleistocene epoch. At least some modules may depend on the environmental shaping of the mind that occurs during the long period of postnatal growth, rather than on specific 'hard-wired' biological adaptations. One example of this is reading, which has modular characteristics but cannot have evolved during the Pleistocene epoch. Evolution more often involves tinkering with existing characteristics than the emergence of completely new ones.

Moreover, studies of intelligence show us that although people differ with regard to specific abilities (e.g., verbal ability, mechanical ability, spatial ability), these specific abilities are all positively correlated with one another, implying that there is also a general intelligence underlying them. This general intelligence, which is sometimes simply called 'g', can be expressed as an 'intelligence quotient' or IQ, although psychologists have generally been unable to come up with a good definition of intelligence, other than to suggest that it is what intelligence tests measure!

The Asymmetrical Brain

There is one respect in which human brains have clearly divided functions. In around 90% of us, speech is controlled by the left side of our brains,

while the right side is essentially mute. The left side of the brain also has a more complete ability to understand language than the right, and in addition it is dominant for written language and for the sign languages of the deaf. Conversely, the right side of the brain seems to have the dominant role in spatial abilities, especially in spatial attention. For example, damage to the right side of the brain can produce a striking phenomenon called hemineglect, in which the sufferer is unable to attend to events on the left side of space, even though the sensory systems themselves are unimpaired. People with this condition may dress only the right side of the body, eat from only the right side of a plate, and ignore objects on the left or voices that address them from the left side. Damage to the left side of the brain seldom causes comparable inattention to the right side, and any right-sided neglect is typically temporary, whereas left-sided neglect may be permanent.

The functions of the two sides of the brain have been strikingly revealed in studies of 'split-brained' individuals who have undergone section of the corpus callosum (the main fiber tract connecting the two sides of the brain) for the relief of intractable epilepsy. They are unable to read words or name objects shown to the left side of visual fixation, since the left side of visual space projects to the mute right side of the brain. They can easily name things shown to the right of fixation, since the right half of visual space projects to the left side of the brain, which is capable of articulate speech. Yet split-brained people can point to the names of objects in left space, or point to the objects represented by names, which indicates that the right side of the brain can identify objects and has at least some limited verbal understanding. There are some perceptual tasks, such as imagining shapes rotated into different orientations, or seeing how incomplete pictures would look if they were completed, that the right brain seems to accomplish more efficiently than the left brain.

A more obvious manifestation of brain asymmetry is handedness. Around 90% of humans are right-handed. Most of the remainder are left-handed, although a small minority are ambidextrous. The asymmetry resides in the brain processes that control the hands, rather than in the hands themselves. With the exception of superficial signs such as the wearing of rings or a watch, it is generally not possible to tell whether people are left- or right-handed simply by inspecting their hands.

It is not entirely clear when these asymmetries evolved. Although other individual primates, and indeed mammals, may prefer to use one hand or paw over the other for various activities, there are generally as many left-handed animals as right-handed ones. There is some evidence for a slight overall tendency for monkeys to be left-handed when reaching, and for captive chimpanzees to prefer to use the right hand when gesturing or feeding, but these asymmetries are neither as extreme nor as general as human right-handedness. The pattern of wear on stone tools suggests that there may have been a preponderance of right-handed individuals as far back as 2 million years ago, but it is not clear whether the proportion matched that of modern humans.

There is better documentation of other types of cerebral asymmetry in animals. A wide range of species, including birds and rodents, appear to show a left-brained dominance for aggression, and a right-brained dominance for spatial processing. There is also evidence that the left side of the brain is dominant for vocalization in a wide range of species, including frogs, birds, mice, rats, and monkeys. Furthermore, an area of the brain known as the *temporal planum*, which roughly corresponds to part of Wernicke's area, is larger on the left than on the right in the majority of chimpanzees, as it is in humans. Cerebral asymmetry in humans may have emerged from these earlier asymmetries, although it also seems to exhibit unique characteristics, perhaps partly because the two activities for which human cerebral dominance is most marked, namely language and tool use, themselves have distinctively human properties.

Some insight into the evolution of cerebral asymmetry comes from studies of the brain area corresponding to Broca's area. In monkeys, this area seems to be involved in the planning and perception of reaching and grasping movements of the hands. There is now evidence that Broca's area is involved in the orchestration of hand movements in humans, too. However, in monkeys both sides of the brain seem to be equally involved, whereas in humans there is a strong left-sided dominance – although it has recently been suggested that the equivalent area on the right may be involved in musical syntax. At some stage in the evolution of our own species, the functions of Broca's area became lateralized to the left side of the brain, and began to incorporate speech as well as gesture. This transition has been taken by some authors to imply that language originated in manual gestures, and was perhaps at one stage similar to the present-day sign languages of the deaf. It may have been the final switch to vocal language that resulted in the eventual dominance of *Homo sapiens* over other large-brained hominins.

CONCLUSION

Humans are descendants of hominin species that split from the other great apes in Africa around 5 or 6 million years ago. However, the enlargement of the brain that characterizes our species began only around 2.5 million years ago with the emergence of the genus *Homo*, and may be associated with the emergence of stone tool manufacture and migrations out of Africa. Many of the distinctive organizational characteristics of the human brain were probably selected for during the Pleistocene epoch, when our forebears adapted to a hunter–gatherer existence on the African savannah. These adaptations were fundamentally social, including the emergence of complex language and the development of a sophisticated theory of mind. It is less clear when the distinctive asymmetry of the brain that underlies human right-handedness and left cerebral dominance for language evolved.

Further Reading

Byrne RW (1995) *The Thinking Ape*. Oxford, UK: Oxford University Press.

Byrne RW, Whiten A (1988) *Machiavellian Intelligence*. Oxford, UK: Oxford University Press.

Cartwright J (2000) *Evolution and Human Behaviour*. London, UK: Macmillan.

Corballis MC (1991) *The Lopsided Ape*. New York, NY: Oxford University Press.

Corballis MC (2002) *From Hand to Mouth: the Origins of Language*. Princeton, NJ: Princeton University Press.

Corballis MC and Lea SEG (eds) (1999) *The Descent of Mind*. Oxford, UK: Oxford University Press.

Deacon T (1997) *The Symbolic Species*. Harmondsworth, UK: Penguin.

Jones S, Martin R and Pilbeam D (eds) (1992) *The Cambridge Encyclopedia of Human Evolution*. Cambridge, UK: Cambridge University Press.

Lieberman P (1998) *Eve Spoke: Human Language and Human Evolution*. New York, NY: WW Norton.

Mithen S (1996) *The Prehistory of the Mind*. London, UK: Thames and Hudson.

Pinker S (1997) *How the Mind Works*. Harmondsworth, UK: Penguin.

Tattersall I (1997) Out of Africa again … and again? *Scientific American* **276**: 60–70.

Human Evolution

Introductory article

Marta Mirazón Lahr, University of Cambridge, Cambridge, UK

CONTENTS

The origin of human species and their relationships with the apes have been reassessed in the light of genetic studies. Hominins are now believed to have emerged in Africa about 7--5 million years ago.

INTRODUCTION

Human evolution is more than the evolution of humans. More than 15 species have existed since the time of our last common ancestor with the apes. The features that make humans different from apes did not evolve at once, or in the same ancestor, and no single trait made the difference. Humans are unique among the apes in the way they move, their global distribution, the size of their brains, their dependency on technology, their linguistic abilities, their slow growth, and their demography. These distinguishing features evolved at different times, each giving an advantage to some individuals at the time they were competing with those lacking them, only to then become part of the heritage of subsequent generations. Progress in understanding human evolution has come through the amazing discoveries of new fossils and archeological sites, the development of sound chronological and paleoecological frameworks, and the application of new techniques such as molecular genetics to the subject. Paleoanthropology is an interdisciplinary subject by its very definition.

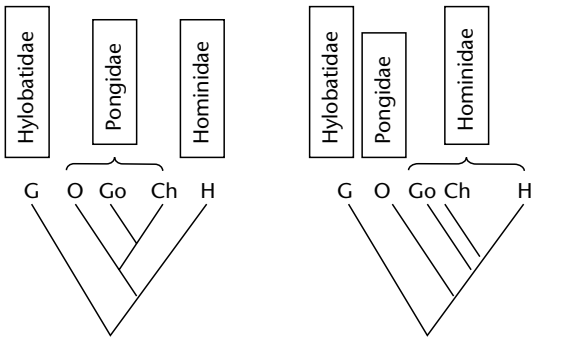
HOMININ ORIGINS AND EARLY HOMININ DIVERSITY

When and Where Hominins Evolved

Fossil and genetic evidence indicates that hominins emerged in Africa between 7 million and 5 million years ago (Mya) in the late Miocene. The earliest fossils that have been clearly identified as hominin date to 4.4 Mya. Earlier, seemingly hominin remains have been found dating from 6.0--5.8 Mya, although their relationship to the hominin or chimpanzee lines remains controversial given the lack of fossil remains of the apes that would have given rise to them. Between the earliest hominin and fossil apes dated to about 9.0 Mya, the African hominoid fossil record does not exist. However, genetic evidence allows us to establish who, among living apes, are our closest relatives, as well as to confirm the dates established from fossils. In fact, these very early hominin fossils have been discovered only in the last few years, and it has been on the basis of genetic evidence that the age of divergence between apes and hominins has been established.

The genetic comparison of apes and humans, with a view towards establishing their pattern of relationships, was pioneered by Sarich and Wilson in the late 1970s. These comparisons, now based on the study of a large number of different genes and gene products, established three key facts: first, that the African apes form a clade within the great apes; second, that within that clade, chimpanzees and humans are more closely related to each other than either is to the gorillas; and third, that on the basis of a molecular clock, hominins and chimpanzees diverged from each other between 7 Mya and 5 Mya. These results have had a major impact on hominoid systematics (Figure 1), as well as establishing a chronological and ancestral framework for the study of hominin origins.

However, controversy exists as to who were the ancestors of the living African apes and where they lived. The fossil record of Miocene apes is a relatively rich one, and yet two very contrasting hypotheses exist regarding the biogeography of hominin origins. The Miocene is a long period (25--5 Mya), which can be divided into two phases. The first of these is characterized by climatic warming and expansion of African forests, in which early apes diversified. The second phase is



(a) Early classification and inferred phylogenetic relationships of living hominoids

(b) Accepted correct classification and phylogenetic relationships of living hominoids

Figure 1. Hominoid systematics. Traditionally the superfamily Hominoidea has been seen as consisting of three families, the Hylobatidae, the Pongidae, and the Hominidae. These groupings emphasized the similarities of the great apes and the distinctiveness of humans. Molecular approaches have radically changed this, and shown a close relationship between humans and African apes, especially the chimpanzees. As one of the aims of taxonomy is to reflect phylogeny, the traditional classification has changed. Orang›utans are the only living representatives of the family Pongidae, while gorillas, chimpanzees and humans form the family Hominidae. This reclassification also changes the vernacular of these scientific names, so that instead of using the term hominid to describe humans and their ancestors, the term hominin (reflecting the subfamily status Homininae) is used. Ch, chimpanzees, genus *Pan*; H, humans, genus *Homo*; G, gibbons, genus *Hylobates*; Go, gorillas, genus *Gorilla*; O, orang›utans, genus *Pongo*.

characterized by global cooling, contraction of forests, lowering of sea levels, and major geotectonic changes. During this period the continents and oceans acquired the general shape by which we know them, the Rift Valley and the Himalayas were formed, the Antarctic partially froze, and the Arabian peninsula emerged forming a permanent land bridge between Asia and Africa. These two continents exchanged fauna through this land bridge, and some apes dispersed to Eurasia for the first time. Among those Eurasian apes we find fossils that represent forms ancestral to orang›utans.

For many years the ancestors of African apes were sought among those apes who in the middle Miocene remained in Africa. In this context, the origin of hominins has been interpreted as part of the process by which East and West African faunas differentiated as the Rift Valley was formed. Hominins would represent an African ape who adapted to the drier and more open environments of East Africa from an arboreal forest ancestor. However, the genetic evidence implies that Asian and African great apes share a common ancestor who lived after apes had dispersed into Eurasia, suggesting that the ancestors of the African apes (and hominins) should be found among late Miocene Eurasian forms. In this context, the differentiation of African apes, including the origin of hominins, would result from the geographical separation of populations of an ancestral species as it dispersed into Africa from the north›east. Therefore, although we know hominins evolved in Africa from a common ancestor with chimpanzees around 6--5 Mya, fascinating issues relating to the origin of that common ancestor will be resolved only by new fossil finds from the crucial period between 8 Mya and 5 Mya in Africa and possibly Arabia.

What Made Early Hominins Different?

Early anthropologists had two main expectations in relation to the origins of humans and their ancestors: that they had diverged from apes a very long time ago, and that this divergence was based upon the development of a large brain. Both were proved wrong. Hominins have a relatively short history, and we now know that brain expansion began in the last 2 million years. What set hominins apart from other apes was their locomotion.

Bipedalism has long been recognized as the key adaptation of the hominin lineage. In terms of morphology, it requires the modification of most of the skeletal system, although such modifications were not all in place at once. The selective advantages of bipedalism are much debated. As a means of terrestrial locomotion, it clearly provides an efficient way of moving in open environments for long periods, especially when compared with knuckle›walking animals, such as gorillas and chimpanzees, rather than true quadrupeds. It also results in more efficient thermoregulation during the middle of the day, as less body surface is exposed to the sun in comparison with quadrupeds. Other advantages may relate to predator vigilance and potential for carrying. Most researchers would argue that a combination of these factors gave bipedal apes a competitive advantage in relation to their quadrupedal relatives at the time.

However, when considering the selective pressures that led to the evolution of bipedal apes, two factors should be taken into account. First, early hominin bipedalism was significantly different from ours. The fragmentary fossil record of the earliest hominins shows clear evidence that these animals had already changed key aspects of their

anatomy towards a bipedal stance. However, the earliest associated fossil skeleton -- that of the famous Ethiopian *Australopithecus afarensis* fossil nicknamed Lucy -- shows that several of the changes that make bipedalism more efficient, such as the shape of the ribcage, the curvature of fingers and toes, and the elongation of the legs, had yet to take place. Therefore, although it sets early hominins apart from other apes, bipedalis itself took millions of years to evolve. The second factor is related to the first. It has become clear that early hominins still spent significant amounts of time in trees. Models of energy expenditure stress that the advantages of bipedalism would become meaningful only when hominin ancestors had to spend a large percentage of their time on the ground.

The Origin of the Lineage That Gave Rise to Modern Humans

Early hominins were clearly successful in the niche they colonized -- that of a terrestrial bipedal ape, with a body and brain size similar to present›day chimpanzees and a broadly similar diet based on fruits, young leaves, and probably some meat. This success led to demographic growth and a greater distribution in East and South Africa, and eventually, competition among themselves for the resources offered in this new niche. Competition can lead to progressive specialization, in which animals become better and better at processing the resources on which they depend, and to differentiation, in which animals try to exploit new resources to avoid their competitors. Hominin evolution in the Pliocene (the period between 5 Mya and 2 Mya) is characterized by both.

Among these bipedal apes, two adaptive trends became established. One of these was towards comparatively specialized animals, who adapted their teeth, masticatory muscles and the shape of the face to maximize their ability to crush large quantities of rough food. They are known as the robust australopithecines, and they consist of several very successful species, usually grouped under the genus *Paranthropus*. This group of hominins was well established before 2 Mya, and survived until approximately 1 Mya in Africa. The other adaptive trend was towards animals who became encephalized, had larger body sizes, slower growth, and appeared behaviorally more flexible, as suggested by their association with stone tools, their wider distribution, and the apparent consumption of larger quantities of meat. In this suite of characteristics, researchers recognize the appearance of many features that we identify with humans, and indeed, this group of animals represent the origins of the genus *Homo* to which we belong.

Establishing the precise point of beginning of new adaptive trajectories in evolution is always difficult, and accordingly much controversy exists as to which fossils mark the threshold between *Australopithecus* and *Homo*. The hominin fossil record in East Africa around 2 Mya shows, besides the big›toothed *Paranthropus*, a very variable collection of animals, grouped under the species name *habilis*, with some but not all of the *Homo* characteristics listed above. Nevertheless, by about 1.8 Mya, a new group of hominin can be identified -- *Homo erectus/ergaster*, who shows the complete combination of incipient *Homo* traits, including the ability to disperse widely. *Homo erectus/ergaster* is the first hominin to leave Africa, and fossils of this group from 2--1.5 Mya are found not only throughout East and South Africa, but also in North Africa, the Caucasus, and Java.

How can we interpret the combination of features that differentiates the genus *Homo* from earlier hominins and represents our own distinct evolutionary heritage? Some of these traits can be seen as fundamental biological changes in behavior, anatomy and growth, while others represent the consequences of these changes given energetic and morphological constraints. Among those fundamental changes it is impossible to determine which came first, as their evolutionary success depended on their concerted modification. However, it could be argued that new adaptive trajectories become established when new behaviors resulting in greater reproductive success lead to selection of individuals whose genetic make›up facilitates those behaviors. In this context, the fundamental behavioral shift associated with early *Homo* was probably the shift to a more carnivorous diet in conjunction with the manufacture of tools that allowed its better exploitation. A more carnivorous diet, rich in protein and fats, would have both released the energetic constraints to increasing the size of the body and the brain (the most energetically expensive organ), and favored more intelligent individuals with greater memory, and capable of more complex cooperation and planning. Furthermore, greater carnivory is associated in mammals with larger geographical ranges and less dietary exclusivity, and thus the underlying condition for dispersal. Encephalization, in turn, would have posed new pressures upon these animals, both in terms of the need for greater maternal energy intake during pregnancy and lactation (most of the period during which human

brains grow), and in terms of maintaining metabolic rate relative to body size.

These animals seem to have pursued two separate strategies for meeting these pressures, strategies that have a feedback effect. On the one hand, researchers have argued that the increasing energetic costs to mothers imposed by encephalized infants were met by budgeting these costs over a longer period, thus leading to slower growth rates. On the other hand, research has shown that while humans have brains significantly larger than expected for their body size, their guts show the reverse condition. On this basis, researchers have argued that there was a trade›off between the energy channeled towards brains and guts so as to maintain stable metabolic rates, and that this was possible through the smaller absorptive gut surface necessary in carnivores in relation to herbivorous animals. These correlated changes and their feedback effects (Figure 2) represent the underlying adaptive trend of the genus *Homo*.

HUMAN EVOLUTION IN THE PLEISTOCENE

Paleoenvironmental Context for the Evolution of Neanderthals and Humans

One of the distinguishing features of the genus *Homo* is its wide geographical distribution. For more than 3 million years, hominins existed and diversified within sub›Saharan Africa. In the subsequent period, the Pleistocene epoch (1.8--0.01 Mya), hominins colonized first the northern and southern (temperate) extremes of Africa (by 1.8 Mya), south›western Asia (by 1.7 Mya), tropical south›east Asia (by 1.8 Mya), then east Asia (by 1.0 Mya) and Europe (by 0.8 Mya). Human evolution in the Pleistocene is no longer an African affair, and allopatry becomes a key issue in interpreting the patterns of differentiation among species and populations. Under a simple allopatric model, Neanderthals and modern humans represent,

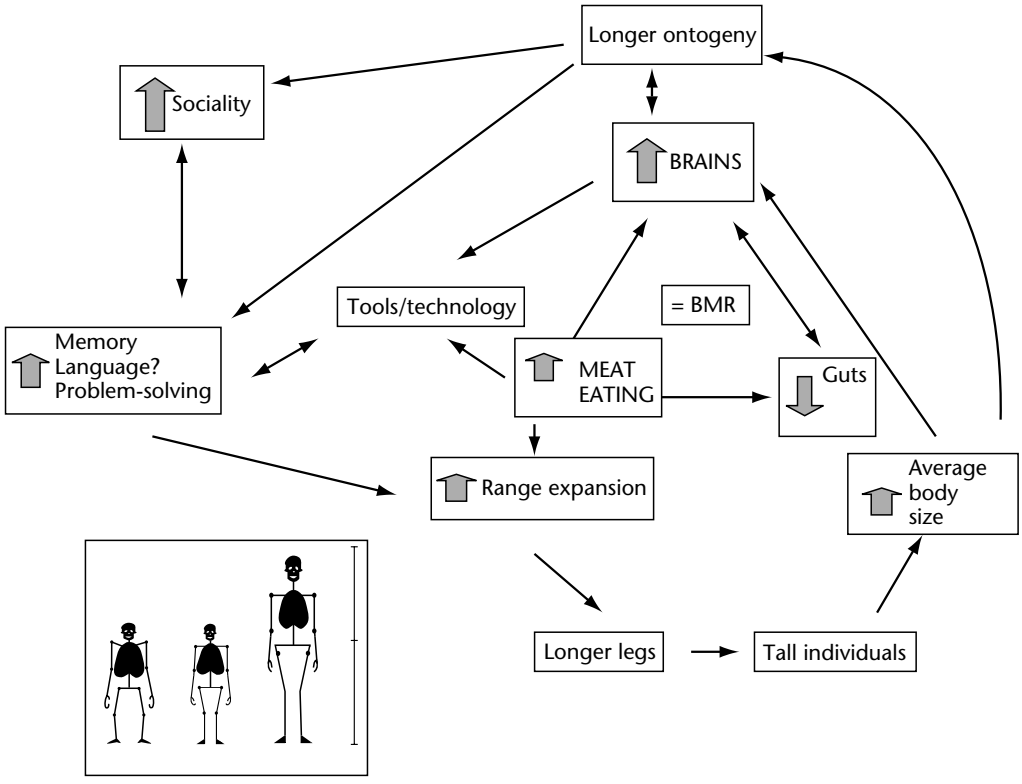


Figure 2. The complex adaptive system that underlies the evolution of the genus *Homo*. BMR, basal metabolic rate.

respectively, the contemporaneous European and African descendants of a Pleistocene hominin species (Figure 3).

However, in order to understand why Neanderthals and modern humans evolved, and why they had such different fates, it is necessary to understand the overall conditions that established their selective environments. Evolution is the result of changes in the competitive environment of species that lead to new patterns of mortality. These establish new parameters of individual fitness, which underlie the process of natural selection and consequently adaptation. The competitive environment of any species has both biotic and abiotic components, which determine the conditions in which the evolutionary process takes place. Understanding climatic change, with its potential for promoting expansion or contraction of ecological niches, is essential for interpreting the diversification of species. The Pleistocene climate was marked by alternation between relatively cold glacial periods, and warmer interglacial ones. The duration and amplitude of the glacial periods both appear to have increased around 800 000 years ago. These climatic changes had effects on faunal distributions and survivorship, some of which are predictable and some of which are not.

A glacial cycle lasts approximately 100 000 years, although each cycle varies in extent and severity. Most of the cycle is characterized by a long period of low, fluctuating temperatures, during which continental ice accumulates at high latitudes. This accumulation of water in the form of ice sheets leads to a drop in sea levels, with the consequent exposure of continental shelves, and to increasing aridity in the tropics. As temperatures increase, the ice sheets which had taken tens of thousands of years to build start to melt, and interglacial phases begin. That excess of circulating water leads to a short period of extreme precipitation in the tropics, creating extended freshwater networks and high lake›level stands. When this water eventually drains to the seas, conditions similar to those of the present day set in. The overall effect for hominins, depending on where they were, was one of expanding and contracting ranges, of alternating rich and scarce resources, and of highly variable thermal conditions.

By 1 Mya hominin populations were established throughout Africa, and in tropical and temperate Asia as far north as northern China. Soon afterwards, hominins occupied Europe for the first time. This broad spatial distribution, in the context of small population sizes, sets the conditions for regional differentiation among the various groups of encephalized hominins. That differentiation becomes very apparent in the Middle Pleistocene (0.8--1.25 Mya). The main distinction in both morphology and archeology is between east and west, along what is known as the Movius line. In the east, hominins have not changed from their *Homo erectus* ancestry. Both in Java and China, fossils in the time›span between 1.0 Mya and 300 000 years ago are *Homo erectus*, associated with relatively simple tools. *Homo erectus* continued to inhabit south›east Asia until about 50 000 years ago, suggesting long›term isolation from other regions of the world. The situation in China becomes increasingly more complex after 300 000 years ago, when fossils displaying a quite distinct morphology make their appearance alongside *Homo erectus* ones, and their affinities are still a matter of debate. In the west, hominins change, losing many of the *erectus* features, as well as showing an overall increase in body size and brain size. This morphology was for a long time referred to as archaic *Homo sapiens*. Most researchers today attribute it to a new species -- *Homo heidelbergensis*. The main specimens of *H. heidelbergensis* are the fossils from Bodo, Lake Ndutu, Kabwe, and Elandsfontein in Africa, and from Arago, Boxgrove, and Petralona (among many others) in Europe.

What makes *H. heidelbergensis* really interesting is not only the strong similarities between the early African and European forms (around 0.5--0.4 Mya), but how they differentiate afterwards. In Europe, a number of fossils of late Middle Pleistocene date (0.3--0.125 Mya) show characteristics later to be found only among Neanderthals. In Africa, differentiation from *H. heidelbergensis* also takes place, with the appearance of features that we identify as modern human traits. Thus we have in *H. heidelbergensis* an even larger›brained hominin species that in the middle of the Pleistocene epoch extended from Africa to Europe. It is this species, or its possible descendant, *Homo helmei*, which in the opinion of many researchers represents the last common ancestor of Neanderthals and modern humans.

The Origins of Modern Humans and Neanderthals

The climatic cycles of the last half›million years promoted periods of recurrent contact between African and European hominins (as shown by the strong similarities between *Homo heidelbergensis* in both continents), while the timing of such cycles suggests that the direction of contact was from Africa to Europe. It is in these contacts that the origin of the lineages leading to modern humans

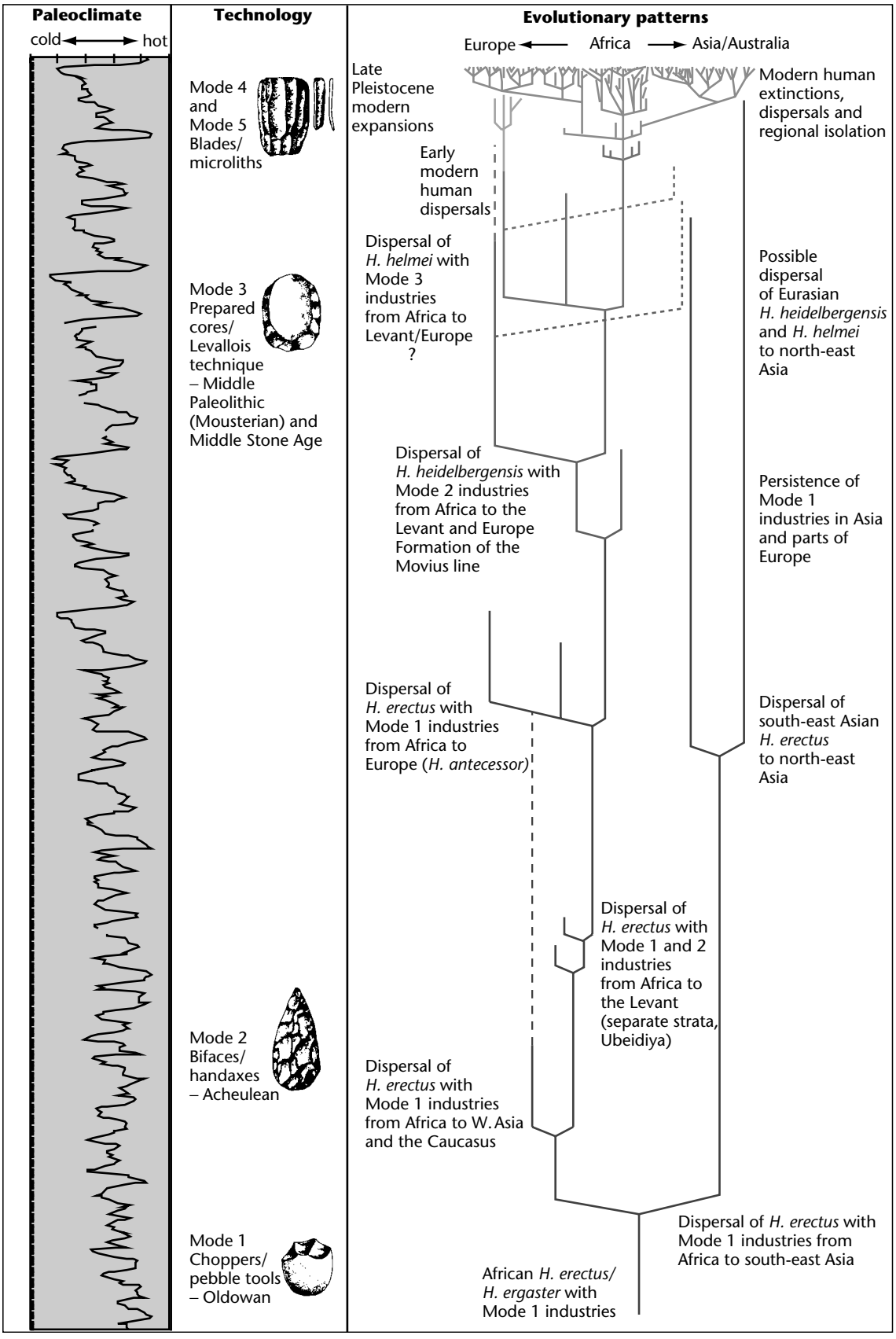
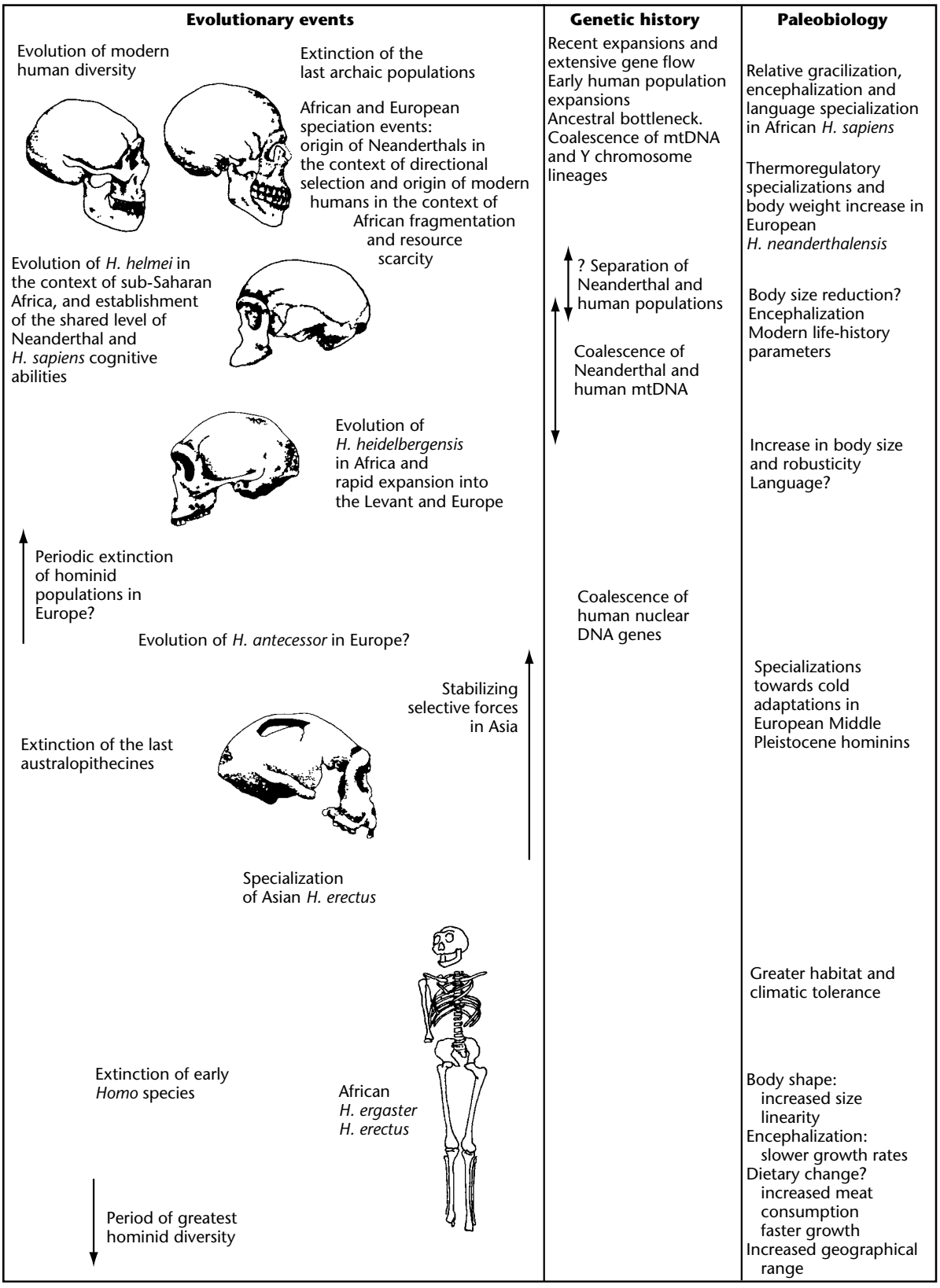


Figure 3. The record of human evolution in the Pleistocene.

Figure 3. (*cont.*)



on the one hand, and Neanderthals on the other, are to be found. As mentioned above, different opinions exist as to when the African and European lineages exchanged significant genes for the last time, whether it was around half a million years ago in the form of *H. heidelbergensis*, or whether it was more recently, around 300 000 years ago, in the form of *H. helmei*. Independent of this, from around 130 000 years ago African and European fossils show modern human and Neanderthal features respectively.

The period preceding the appearance of Neanderthals and modern humans, 200 000--130 000 years ago, was one of extensive glaciation. These extreme climatic conditions would have generated different pressures on the populations of both continents. In Europe, thermoregulation would have been the main factor affecting survivorship, while in Africa the immediate effect would have been scarcity of resources because of aridity and ecological fragmentation. It is in response to these different selective pressures that researchers interpret the adaptive differences between Neanderthals and modern humans.

Many of the features of *H. sapiens* and *H. neanderthalensis* were inherited from their large›brained and large›bodied ancestor, who had a relatively long ontogeny, who manufactured complex tools, used a degree of planning and social organization in its strategy to exploit the environment, and most importantly, probably had some linguistic abilities. The differences between the two species thus reflect their modification of this ancestral heritage in order to meet the pressures posed by their respective evolutionary environments. In Neanderthals, these adaptations are represented by changes to the face, in particular how much it protruded and how large the nasal aperture was, together with a broad chest, and an overall dense and robust skeleton, with relatively short distal limb proportions. These changes are thought to reflect thermoregulatory adaptations to the warming of inhaled air and overall heat retention, as well as to a tough lifestyle that required extensive nomadism. In modern humans these differentiating features are more difficult to specify, largely because the history of the species is one of early geographical differentiation, thus making it difficult to identify the universal traits. However, a few can be described -- these are represented in the small size of a nonprotruding face with a chin in the jaw, in a differentiated cranial shape that resulted in a rounded head, and an overall gracile body. The increased aridity throughout Africa that led to increasingly scarce resources would also have led to demographic contractions, which might in turn have increased the role of genetic drift, as well as set new selective pressures. The changes observed in the evolution of a modern morphology, to a large extent associated with the gracilization of the skeleton, are consistent with selective pressures on resources, by which a cheaper skeleton would have been advantageous. Recent studies on the ontogeny of Neanderthal and modern human cranial growth suggest that the differences characteristic of both species were present at birth.

Genetic evidence has significantly increased our understanding of the evolution of these two lineages. Genetic studies have made a major contribution towards establishing a recent African origin for modern humans, and revealing their relationship to Neanderthals through the retrieval of ancient DNA. The genetics of living populations can be used to draw inferences about their past, in a similar way to the use of molecular clocks to understand the phylogenetic relationship between living apes and humans, and to infer, through genetic distances, the time lapsed since their divergence. The study of the genetic origins of modern humans is based on several genes and a range of different techniques.

For a long time geneticists used classical markers (polymorphic gene products) to study the variation among human populations. These are measured as percentages of different alleles in different groups, and their study has been championed by the Italian geneticist Luca Cavalli›Sforza. Classical markers are interesting, particularly because a large number of gene systems can be analyzed together to provide a tree of relationships of human populations. These are population trees, in the sense that they track the multivariate history of populations through the genetic expression of many systems. The results of studies of classical markers point clearly to an African origin of all modern humans, with the first branches of a human genetic tree taking place among African populations.

Since the 1980s molecular markers have been used to track the history of human populations. These studies were first developed using the mitochondrial deoxyribonucleic acid (mtDNA) genome, but are today based on a number of genes, including mutations on the Y chromosome. The advantages of the mtDNA and Y chromosome genomes are that they are inherited through only the maternal or the paternal line, and therefore do not undergo recombination during meiosis. Furthermore, most of these markers are believed to be selectively neutral, which means that not only does their distribution reflect history rather than

selection, but also they accumulate mutations faster than functional genes. Together, they provide the most powerful tool for the investigation of the evolutionary history of living human populations (of course, they cannot throw light on any group that has since become extinct, with the rare exception of the extraction of DNA from actual ancient remains). Their disadvantage is that they provide gene trees rather than population trees, and the history of any population is made up by the sum of each genes history. This is an important caution, as, for example, mtDNA and Y chromosome histories are not the same because men and women have had different histories determined by their patterns of mating and dispersal behavior. These single gene systems are analyzed in a variety of ways, the most important of which are based on coalescence theory and network analysis (Figure 4).

The results of these studies all point to an African origin of modern humans. Among the thousands of individuals sampled for either mtDNA or Y chromosome markers, Africans are the most diverse in having the greatest number of mutations. Furthermore, all the earliest branching events occur among African individuals. In other words, these data show that the last common ancestor of all living humans was part of a population from which Africans derive, and that only much later, other populations branched off this group. The application of a molecular clock to these various systems provides a range of dates for the last common ancestor of all humanity -- between 200 000 and 130 000 years ago, thus in full agreement with the paleoanthropological record.

Molecular data can also provide information on the demographic parameters of the ancestral population of any given species. The basis for these is the estimation of the effective population size of a given system, derived from the equation

$$F_{ST} = \mu 2N_e \quad (1)$$

Both F_{ST} (a measure of diversity in a gene) and μ (the mutation rate of a gene) are empirical observations, allowing the calculation of N_e, the effective ancestral population size. From such calculations, geneticists have suggested that the ancestral population of *Homo sapiens* underwent a severe demographic bottleneck before its diversification. Estimates of the magnitude of the ancestral bottleneck vary, ranging from 1000 to 10 000 individuals. Transforming these values into census values (i.e. the whole population), it is possible to estimate that the ancestral population of modern humans was at one point composed of 3000 to 30 000 people. The scar of this bottleneck is one of the main features of *Homo sapiens*. Although our species contains more than 6 billion individuals, we have a tenth of the diversity of any of our close primate relatives. From their African origin, modern humans underwent various periods of demographic expansion that led to multiple dispersals into Europe and Asia. A large proportion of the diversity among human populations can be related to this early fragmentation as small groups of hunter›gatherers dispersed and colonized the world. One such dispersal took modern humans into Europe around 40 000 years ago, into the Neanderthals homeland.

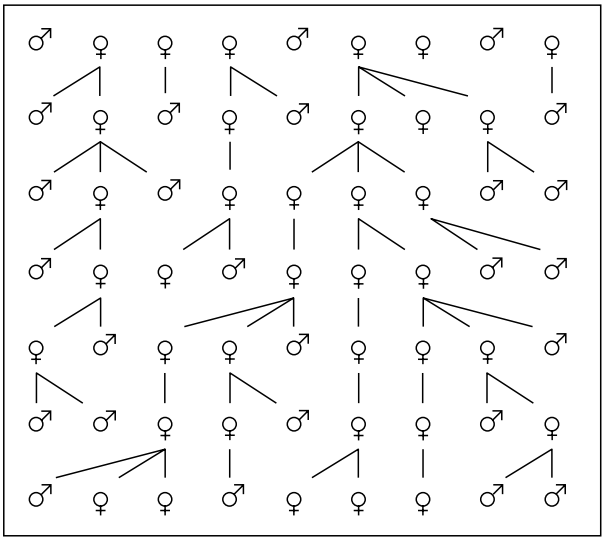


Figure 4. Basic principles of coalescence theory, showing as an example the coalescence of a maternally inherited gene.

The nature of the relationship between modern humans and Neanderthals has been much debated. After modern humans dispersed into Europe, Neanderthal fossil remains and their associated Mousterian technology disappeared from most of the continent. However, in certain areas these two hominins appear to have coexisted for some 10 000 years. Studies of Neanderthal mtDNA, obtained through the extraction of ancient DNA from three fossils, show that 600 000--500 000 years ago there lived a population from which both Neanderthals and modern humans descend, although not when this population separated into its European and African descendants. Furthermore, the absence of the particular mutations observed in the Neanderthal samples among the thousands of recent people studied indicate that Neanderthals did not contribute to the living mtDNA gene pool, offering further evidence for the extinction of Neanderthals, although whether small amounts of interbreeding took place 40 000 years ago cannot be ruled out. The last 30 000 years have been probably the first time in the last 5 million years in which only one species of hominin exists in the world.

HUMAN DISPERSALS

Dispersal is an essential element of the evolutionary process for all organisms. Most species start as small localized populations, and if successful spread more widely to fill up the available habitat. The rate and direction of any such dispersal are functions of the ability of the species to survive and thrive in the environments encountered, and of its own characteristics. Looking broadly across the animal kingdom, it is clear that not all species disperse to the same extent: some are much more widely distributed than others. Lions, for example, existed across the whole of Africa and Eurasia, whereas most monkeys are very localized.

Human evolution fits into this biogeographic pattern. From their localized African origins, humans have come to live all over the world, and that can only have happened as a result of dispersals. However, that pattern of colonization is a complex one, occurring irregularly over long periods. The first part of human evolution, that of the period 5--2 Mya, was dominated by australopithecines, and they never managed to disperse beyond Africa. In this sense they were much like the other anthropoid apes. With the emergence of *Homo* more extensive dispersals occurred. From 2--1.5 Mya *Homo ergaster/erectus* spread into the southern parts of Asia, becoming the first non›African hominin. After 1 Mya they gradually dispersed into the colder, more northerly latitudes of northern Eurasia. However, rather than this being a single continuous dispersal, it is probably best seen as a series of repeated events, probably interspersed with extinction of populations. The primary driving force of this process would have been the climate, particularly after the onset of glacial conditions.

The dispersal of modern humans is another of these events, beginning during the warmer phase of the last interglacial period, around 120 000 years ago. Again, the dispersal of modern humans was not a continuous single event, but is best described as a series of multiple dispersals. The major events in this process were the first dispersals out of Africa and along the southern rim of Asia, through to Australia, around 60 000 years ago; the dispersals into northern Eurasia from around 40 000 years ago; recolonization following the extremes of the last glaciation around 15 000 years ago, resulting in the first peoples in the New World; and, perhaps most important in terms of the current human population, the spread of the first farming peoples in the last 10 000 years. While archeologists and anthropologists have argued about the relative importance of migration compared with local development, it is becoming increasingly evident that the ability to migrate and colonize has been an important element of human evolution.

HUMAN TECHNOLOGIES

The archeological record represents the appearance and development of hominin technological abilities, and it extends to the last 2.5 million years. Stone tools are thus associated with the second half of hominin evolution, most markedly with species of the genus *Homo*, although tools in possible association with late australopithecines have been found. What is the importance of the archeological record?

- It provides very strong evidence of behavior in the past, ranging from use of the landscape, to subsistence strategies, to defense organization.
- It provides unique evidence of cognitive processes in extinct taxa, through the inferred thought processes required to produce either the tools themselves or their accumulation pattern.
- Owing to its durability, it provides the best evidence for the spatial distribution of populations in the past, allowing, in certain circumstances, paleodemographic inferences to be made.

The range of tools possible is unlimited to our minds, with new inventions appearing daily, but this was not so in the past. Several factors acted to constrain the range of tools earlier hominins made: manual dexterity, cognitive ability, and availability of raw materials. This has allowed archeologists to categorize stone tool traditions according to the shape of the tools produced or the methods by which they were made. Several such categorizations with different levels of detail exist (Figure 5). One of the most useful classifications of the archeological record is that developed by Graham Clark. He organized stone tools according to what he called modes, and the timing and distribution of these modes, as well as trends within the technocomplexes they define, have important consequences for our interpretation of hominin cognition through time (see Figure 3).

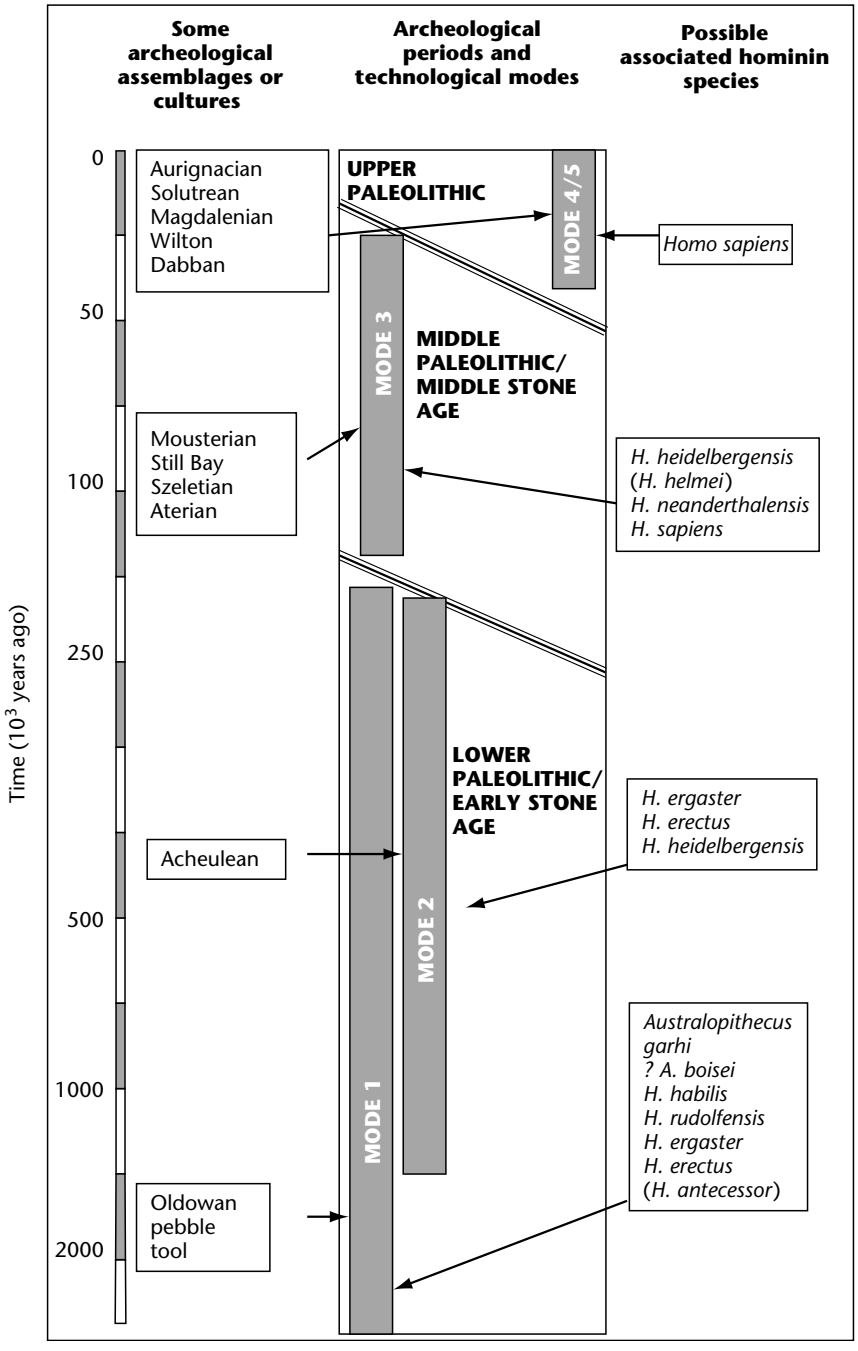


Figure 5. Temporal distribution of archeological assemblages and modes.

For those working within the framework of a recent African origin of modern humans, *Homo heidelbergensis* has been considered as the last common ancestor of Neanderthals and modern humans. This view argues that, some time between 600 000 and 500 000 years ago, during an interglacial phase, hominins dispersed from Africa to Europe, taking Mode 2 or Acheulean technologies to Eurasia for the first time. However, around 300 000 years ago, a new technology appears in Africa and soon afterwards in Europe. This technology, known as Mode 3, or Middle Stone Age (Africa) or Middle Paleolithic (Europe), has some striking features. Hominins were no longer turning stone cores into tools, but rather flaking them to obtain thinner, preshaped, and more abundant objects. In order to do this, hominins needed to prepare the cores prior to flaking, a step that required a level of mental abstraction. This technique of preparing the core before flaking, first observed within late Mode 2/Acheulean assemblages, is the technological background of both Neanderthals and modern humans.

Neanderthals and modern humans may have inherited this ability from a more recent common ancestor, known as *Homo helmei*. This would be consistent not only with the archeological record, but also with the little evidence there is for the evolution of language. Language is considered one of the key traits that define and to some extent structure human cognition and human behavior. However, the ability to speak does not affect the skeleton, and thus its origins cannot be traced in the fossil record. However, in recent years two indirect lines of evidence for linguistic ability among hominins have been explored -- one related to the expansion of the vertebral canal at the point of innervation of the diaphragm, and the other related to the size and shape of the hyoid bone. Both suggest that some of the anatomy that allows human speech evolved within the last half›million years, and that Neanderthals probably had some language ability. This would further reinforce the view that the two lineages share a recent common ancestor in the late Middle Pleistocene who had increased cognitive abilities.

HOMININ EVOLUTION: THE EVIDENCE

Paleoanthropology has recovered a rich record of the evolutionary history of humans. This record is made up of fossils and of the material culture left behind by various hominin species, integrated with the information derived from geology, paleontology, and paleoclimatology. It is further enriched by the evolutionary inferences possible from our own genetic diversity and that of our closest living relatives, the apes. Together, this information allows the reconstruction of a complex process, in which we recognize some broad patterns:

- Hominin origins, the divergence of our lineage from that leading to chimpanzees, date to 7--5 million years ago, and are based on changes in locomotive strategy.
- The first 3 million years of this history, during the Pliocene, are characterized by the diversification of bipedal apes within eastern and southern Africa, leading to the evolution of two separate lineages: one in which megadontic adaptations are associated with the consumption of large quantities of low›quality foods, the other in which encephalization, increased carnivory, tool use, and a longer ontogeny are associated with a more opportunistic exploitation of resources.
- The last 2 million years, the Pleistocene, are characterized by dramatic geographical expansions in the context of deteriorating climatic conditions, and the establishment of distinct regional trajectories expressed in both morphology and material culture.
- Neanderthals and modern humans represent the contemporaneous evolution, in Europe and Africa respectively, of large›brained and behaviorally complex hominins from a common ancestor who lived 500 000--300 000 years ago, and thus uniquely share a number of traits.
- From their African ancestral homeland, modern humans dispersed throughout the world by means of multiple events. These, together with the different selective pressures in the areas colonized and the effects of drift on small hunter›gatherer groups, shaped the diversity of the species.
- The extinction of the Neanderthals was probably associated with the migration of modern humans into Europe and the changes in the competitive environment caused by this influx of people.
- In recent millennia, and in association with the shift to an agricultural and sedentary lifestyle, humans have undergone a demographic revolution that took the species from a few thousand to over 6 billion individuals.
- When compared with other species, *Homo sapiens* shows a distinct lack of genetic diversity in spite of its current population size (the heritage from a recent ancestry as a very small population in Africa), and most of that diversity is to be found within, rather than between, populations.

In summary, humans are the result of the same evolutionary mechanisms that acted on other animal lineages. The history of their lineage is made up of the evolution, differentiation, and extinction of a large number of species during the last 7--5 million years. The history of the species is relatively recent and characterized by unpreced-

ented geographic and demographic success, one in which our unique cultural complexity and diversity played a major part.

Further Reading

Boyd R and Silk JB (2000) *How Humans Evolved*, 2nd edn.
Conroy G (1997) *Reconstructing Human Origins: A Modern Synthesis*. New York, NY: WW Norton.
Donnelly P and Foley RA (eds) (2001) *Genes, Fossils and Behaviour: An Integrated Approach to Human Evolution*. Brussels, Belgium: NATO.
Foley RA (1995) *Humans Before Humanity: An Evolutionary Perspective*. Oxford, UK: Blackwell.
Gamble C (1994) *Timewalkers: The Prehistory of Global Colonization*. Cambridge, MA: Harvard University Press.
Johanson D and Edgar B (1996) *From Lucy to Language*. New York, NY: Simon & Schuster.
Klein RG (1999) *The Human Career: Human Biological and Cultural Origins*. Chicago, IL: Chicago University Press.
Stringer C and Gamble C (1993) *In Search of the Neanderthals*. New York, NY: Thames & Hudson.
Tattersall I (1995) *The Fossil Trail: How We Know What We Think We Know About Human Evolution*. Oxford, UK: Oxford University Press.

Human Reproduction and Life Histories

Intermediate article

Gillian R Bentley, University College, London, UK
Ruth Mace, University College, London, UK

CONTENTS

Life history theory can explain varying patterns of human reproduction in different human societies given that organisms must allocate energy between the competing demands of growth, maintenance, and reproduction. One enigma to evolutionary theory is the demographic transition when humans began voluntarily to reduce fertility to what appear to be maladaptive levels.

THE USE OF LIFE HISTORY THEORY IN STUDIES OF HUMAN REPRODUCTION

To understand the reproductive strategies of species and even individuals within species, evolutionary biologists routinely use life history theory as an analytical tool. It is a way of understanding how organisms allocate available energy to growth, maintenance, and reproduction in order to maximize their fitness. Life history analysis therefore examines factors such as maturation rates, fertility, interbirth intervals, age at reproductive cessation, and maximum lifespan. The differential allocation of energy among growth, maintenance, and reproduction necessarily represents trade-offs – for example, allocating more energy to growth may mean postponing reproduction; allocating energy to produce a greater number of offspring may compromise the survivorship of these offspring or of the individual that produces them. Such trade-offs are usually dependent on the kinds of environments in which individuals live and the availability of sufficient resources.

Biologists also refer to species as having specific life history characteristics which can be used on a comparative basis and to acquire useful comparative information about adaptive strategies. For example, humans have long life spans compared to other similar-sized mammals; they usually give birth to one altricial infant; human offspring have a comparatively long period of juvenile dependency before being able to reproduce and become independent; adolescents in preindustrial societies become fertile between about 14 and 16 years of age; lactational periods average between 2 and 4 years; interbirth intervals in noncontracepting societies are about 3–4 years long; and the total fertility rates (TFRs) in such societies averages six. Understanding some aspects of the life history of species can also generate predictions about others. For example, large-bodied mammals generally have longer life spans and produce fewer offspring at slower rates compared to smaller mammals.

THE LIFE HISTORIES OF HUMAN FORAGERS COMPARED TO OTHER SUBSISTENCE GROUPS

Recently, anthropological demographers as well as evolutionary psychologists have suggested that we can only fully understand the adaptive significance of life history patterns in our species by studying those of foraging groups (Hill and Hurtado, 1996). The rationale behind this is that human physiological and cognitive evolution occurred under the constraints of a hunting and gathering way of life. In this case, studying the life history patterns of humans in modern, industrialized nations who are exposed to novel and sometimes unprecedented stresses may not be informative about potential adaptive responses. In addition, understanding the life histories of foraging groups may be helpful in attempting to reconstruct similar parameters for our hominid ancestors who, in turn, evolved from the common ancestor to modern apes. There is, however, no single characteristic suite of traits

that can be used to describe foraging societies, precisely because such traits are mediated by environmental conditions even within the same population. This is amply demonstrated by Hill and Hurtado (1996) for the Ache foragers of Paraguay. There are certainly generalities that one can make – for example, the maximum average number of offspring recorded for women in foraging societies is much lower than that of women in subsistence agricultural societies – but the patterns outlined below show considerable variation (Table 1).

Table 1 contains comparative data from three foraging groups for whom there exist probably the most reliable information on life history traits; even so, information for many aspects of interest is still missing. The !Kung san of Botswana are somewhat problematic since it is highly probable that many women in this population had been exposed to sexually transmitted diseases at the time of study. This would explain their low TFRs and very early average age at last birth. Despite this problem, the TFRs of the three foraging groups here range from five to eight children, the Ache producing almost twice as many offspring as the !Kung. That resource availability may be a factor explaining differential fertility is suggested by the average energy intake of the three groups (as well as energy spent in obtaining these resources) with the Ache having the greatest intake and also the highest TFR.

Many human life history characteristics, particularly those relating to fertility among noncontracepting societies, are shared with our closest primate relatives – the apes. They are similarly long-lived animals that invest several years in their single offspring. Chimpanzees and gorillas produce a total of about five offspring during their lifespans, females breastfeed their infants for up to 4 years, and have long interbirth intervals of approximately 5 and 4 years respectively. Chimpanzees and gorillas also have similar reproductive cycles and gestational lengths to humans (Bentley, 1999). The main features that distinguish us from other apes is a longer maximum life span, a longer period of juvenile dependency, and the potential for very short interbirth intervals. Human females are also distinguished by having a menopause.

Table 1. Comparison of life history traits among foraging groups

	!Kung[a]	*Agta*[b]	*Ache*[c]
Age at menarche	16.6	17.1	15.3
Period of adolescent subfecundity	2.19	3.01	3.2
Age at first birth	18.8	20.1	18.5
Interbirth interval	3.7	2.9	3.2
Age at last birth	34.4	NA	NA
Age at menopause	NA[d]	43.9	late 40s
Years of reproductive capacity	15.6	18.7	NA
Total fertility rate	4.7	6.5	7.8
Daily caloric consumption (kcal)	2,355	NA	3,827
Daily time spent in food acquisition (mins)	108	NA	79

[a]Howell, N (2000)
[b]Goodman *et al*. (1985)
[c]Hill and Hurtado (1996)
[d]Not available.

As mentioned above, the maximum TFR recorded for foraging societies is lower than for subsistence agriculturalists. The reasons for this disparity are linked to technological and environmental changes brought about by the development of agriculture and urbanization. These changes appeared to reduce interbirth intervals probably by improving maternal nutritional status rather than permitting earlier weaning. There were also significant socioeconomic changes that accompanied agricultural innovation and that permitted, for example, the permanent employment of wet nurses for individuals who could afford them.

In contrast, the average TFR of modern industrialized nations averages less than two. This translates into a rate of growth below replacement levels and, if it persists (as seems likely) population numbers can only be maintained in such countries through immigration. Humans in such societies are able to achieve and maintain this low fertility level through the extensive use of effective contraceptives. Tied to this are the economic and psychological pressures exerted on parents that motivate low fertility. The differences between the life history of reproduction between humans in noncontracepting, foraging societies, modern industrialized societies, and our closest primate relative, the chimpanzee, are exemplified in Figure 1.

ALTERNATIVE APPROACHES TO LIFE HISTORY THEORY

In many respects, life history characteristics are also studied by demographers who are generally trained in social sciences with little or no biological background. Demographers are interested in factors such as reproductive rates, patterns of mortality, interbirth intervals, and comparison of these rates between different groups. What demographers lack compared to biological practitioners is a higher-level body of theory with which to organize and understand their data. Integrating the two areas of research has become a goal for a number of

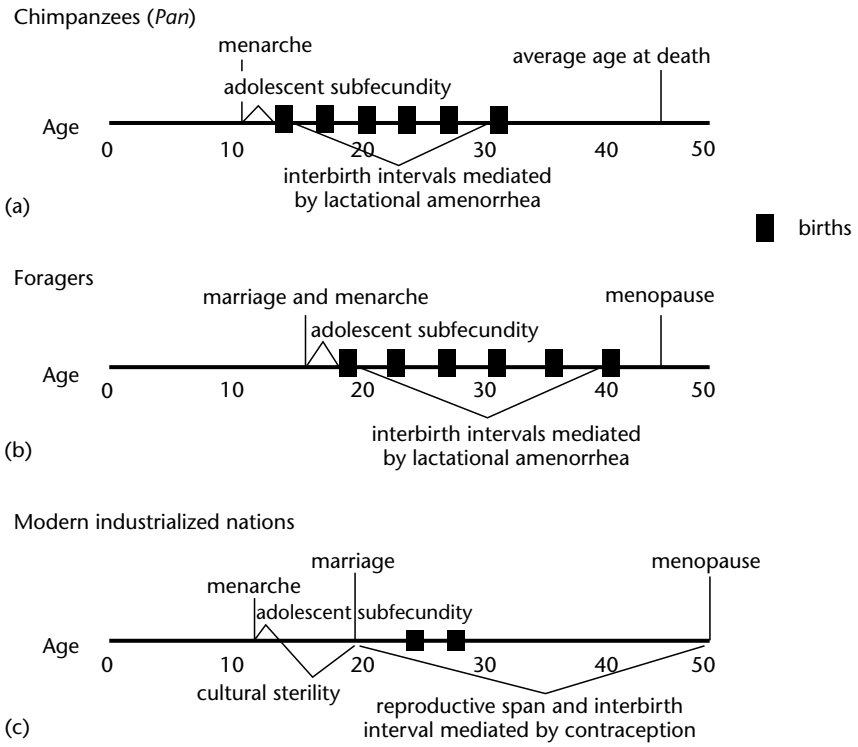


Figure 1. Life history traits of chimpanzees, human foragers, and humans in modern industrialized societies. Adapted from Lancaster and Hamburg (1986).

biological anthropologists and evolutionary ecologists interested in population issues and life history theory.

In addition, traditional demographers typically do not study populations such as foragers who live in remote and inaccessible parts of the world. Data from such groups have derived almost exclusively from anthropologists. In explaining factors such as the comparatively low fertility of many foraging groups, anthropologists have drawn on models from the medical literature, particularly those dealing with nutrition and energy expenditure (sports medicine). For example, the low fertility of the !Kung San of Botswana (with a TFR of 4.7) may be explained by the high levels of energy expenditure necessitated by their foraging lifestyle where women are responsible for most of the energy intake of the group. On the other hand, the !Kung San have also been exposed to sexually transmitted diseases which may have lowered their fertility, as outlined above. Other models refer to the importance of postpartum lactational amenorrhea and extensive birth spacing to explain low fertility among subsistence populations. Many anthropologists have studied the poor nutritional intake of foragers and horticulturalists who may live in marginal environments, and the impact of low energetic status on ultimate fertility. Reproductive ecologists are also using hormonal analyses to understand the mechanisms that govern the reproductive system and to explain patterns of low fertility among subsistence groups.

At the same time, all of these approaches within anthropology – in contrast to the demographic perspective – take a fundamental biological approach: namely, that achieved fertility conforms to life history expectations and is likely to be optimal for the individuals involved.

THE DEMOGRAPHIC TRANSITION

Despite the exhortation to study foragers, traditional demographers have spent much more time studying a phenomenon in contemporary human population history known as the 'demographic transition'. This refers to a historical (and still ongoing) process first described in the 1950s by demographer F. W. Notestein, whereby countries undergoing industrialization, beginning with Europe and North America about two centuries ago, gradually transform from high fertility and high mortality patterns to low fertility and low mortality. During this transition, countries generally experience very high growth rates when

fertility rates outstrip mortality rates. The drop in mortality rates is associated with improved nutritional and health conditions that accompany industrialization. The larger family sizes that result from a reduction in mortality eventually leads families to respond by decreasing family sizes. Populations after the demographic transition have many fewer individuals in the lower age classes than those that have not undergone fertility decline.

Although frequently called demographic transition *theory*, the traditional description of population events in Western Europe referred to a *process*, as opposed to an explanation, for what was observed and documented. In fact, demographic transition theory has been criticized on a number of grounds by behavioral ecologists including its lack of a theoretical framework. In addition, while demographers have frequently predicted a demographic transition for countries in Africa, South America, and Asia, it is by no means clear that similar processes are, or will, be operating there. Research into historical demography also paints a far more complex picture of events in preindustrial Western Europe than was previously thought. In particular, many areas had comparatively low fertility caused by voluntary abstinence and late ages at marriage well before the period of time that was traditionally assigned to demographic transitions. Yet there are sufficient similarities around the world to define a phenomenon in need of an explanation.

THEORETICAL DIFFICULTIES EXPLAINING LOW FERTILITY

As already stated, human behavioral and evolutionary ecologists have for some time argued that life history theory could provide an excellent and proven model with which to explain human demographic processes and behaviors, just as it does for other species (Hill and Hurtado, 1996). But, from this perspective, the human demographic transition to low fertility regimes remains a theoretical enigma. A primary tenet in evolutionary theory is that, where resources are abundant, organisms should opt to maximize their reproductive success by leaving as many offspring as possible in the expectation that more will survive to reproduce. This prediction appears to hold true for preindustrial human groups.

Humans have evidently evolved to enjoy and seek out sexual opportunities, and also to acquire resources to attract mates. Children would have been the inevitable consequence of these preferences in ancestral environments. Indeed, we still have such preferences, and there is evidence that wealthy men in modern societies do have a higher number of sexual partners. But, sexual activity does not necessarily translate into progeny if contraceptives are used. This raises the question of why people choose to use contraception, and why fertility started to decline in some European countries long before any modern methods of contraception were available. Reproductive decision-making is plastic in traditional and modern societies, and it is therefore interesting to examine how an evolved human psychology that favored reproduction could have produced such a consistent, modern trend towards small family sizes. An explanation for the demographic transition is therefore of particular interest to evolutionary anthropologists, because the reduction in family size generally occurs at a time of improving living conditions.

Various schools of thought have arrived at different levels of explanation for this apparently maladaptive fertility decline. One possible explanation is put forward by a particular school of theorists in evolutionary psychology who argue that many behaviors in our modern industrialized world are, in fact, maladaptive given the disjunction between this kind of environment and that of our human foraging ancestors in which we evolved. This is specifically referred to as the 'environment of evolutionary adaptation'. In this case, we need look no further for other explanations: low fertility rates are essentially maladaptive since the kind of environment in which we live is unprecedented and produces aberrant behaviors. This form of evolutionary psychology, however, has been criticized on several grounds. Of particular relevance here are critiques against the idea that humans have been unable to adapt to their environments since the Pleistocene period (e.g. Irons, 1998).

In fact, no simple biological, economic, or ecological correlate can predict the onset of the demographic transition across countries. Historical demographers tend to consider the demographic transition as an idea that is spreading around the world, using language rather similar to the meme concept, introduced by Richard Dawkins. France experienced the first and probably best documented demographic transition, starting in the eighteenth century. The prevalence of smaller families, presumably associated with higher parental investment in each child, spread from centers of sophistication over time. In remote areas fertility decline came later, but once it started, it proceeded more rapidly. One of the best predictors of the onset of the trend was simply distance from these epicenters.

Again, such findings describe rather than explain the demographic transition. More formal models of cultural evolution, in which units of culture are considered to evolve in ways similar to genes (but without the need to necessarily enhance biological success), have also been used to generate some explanations. Boyd and Richerson (1985) show that it is theoretically possible for a 'meme' for low fertility to spread in a scenario where we copy successful individuals. Since success is defined through wealth, when child rearing tends to compete directly with the ability to earn money, those with small families are more likely to be wealthy and to become role models.

GENETIC FITNESS VERSUS OTHER GOALS IN CONTEMPORARY SOCIETIES

All these explanations described above imply that the demographic transition is maladaptive. This could be taken to support the claims of some scholars that humans are different from other animals and that evolutionary theory has little utility when applied to our own species. Evolutionary ecologists, however, are familiar with the notion that fertility might be reduced by some quantity/quality trade-off (Lack, 1968), and hence have explored the possibility that very low fertility could indeed be adaptive in some circumstances. The best evidence for this comes from recent models that combine themes drawn both from economics and evolutionary ecology. For example, economists like the Nobel prize winner Gary Becker have long argued that it is the economic costs and benefits of children that influence fertility. Evolutionary anthropologists maintain that wealth actually always flows down the generations, even in traditional societies (i.e. children are nearly always net costs) and that, in any case, Darwinian fitness is the currency being maximized, not parental income.

There are empirical examples of almost instant shifts in fertility occurring when the economic costs and benefits relevant to raising children change. In societies such as hunter-gatherers where juvenile mortality rates cannot always be predicted but can assumed to be high, an optimal fertility strategy is to produce large numbers of offspring with the hope that some will survive to maturity. In such societies, the costs of offspring are relatively low, particularly once children are old enough to begin foraging for themselves (this will itself vary from society to society depending on the environment and difficulties of acquiring food resources). In agricultural societies, where children can perform agricultural labor and become producers relatively early, as well as become caretakers of the elderly at older ages, the costs of offspring can be even lower. Where resources and especially land is both heritable and in short supply, however, having too many children may not be optimal in the long term. In fact, women in such societies appear to approach an optimal and intermediate number of children compared to their theoretical maximum.

The costs of raising children in modern industrialized societies is, by almost any standard, relatively high, and estimated as approximately £350,000 in the UK including the cost of education, food, and clothing (McAllister and Clark, 2000). The returns on this investment are also relatively low compared to foraging groups – most offspring in modern industrialized societies eschew care of the elderly, who are often housed instead in specialized and segregated facilities. In addition, the chances of survival of offspring given the quality of modern medical care are excellent, reducing the need to produce surplus offspring. Producing low numbers of children in a highly competitive atmosphere may in fact be optimal for most parents. The evidence thus shows that parents have large families if children can generate income for their family. However, the logical consequence of this argument is that when children become a net cost, population fertility will decline to zero, which has never been observed. One argument advanced to explain this is that children fulfill an innate biological desire for nurturing that will tend to override other economic considerations (Foster, 2000).

Economic demographers have recently focused on the link between relative cohort size and fertility decline – the argument is that those individuals born in 'baby booms' face stiff competition, particularly when trying to find jobs and build families; economic hardship relative to expectations precipitates fertility decline. The political state of a population can sometimes have a surprisingly large influence on fertility. While government edicts often have no effect – as in Romania during the Ceaucescu regime of the 1980s – pro-natalist policies do seem to work when they occur during warfare or territorial disputes. One such example is Palestinians living in the Occupied Territories who have one of the highest TFRs anywhere. A political *volte face* in the government of Iran from pro-natalist policies to those encouraging family planning correlated with an unprecedented and dramatic decline in fertility in just 10 years during the 1990s.

Parental investment is the key to understanding fertility decline. In humans, parental investment

can include inherited wealth and education, and anything else that might enhance the long-term chances of gaining wealth or status. In hunter-gatherer populations, children still require some level of parental support until their late teens to survive, so a psychology of high parental investment is likely to extend far back in our evolutionary history. In farming populations, heritable resources like land and livestock become strong determinants of reproductive success.

A state-dependent, dynamic model that optimizes reproductive success over two generations has been used to predict both optimal family size and the optimal amount of wealth to allocate to each child at the end of the parents' reproductive lives (Mace, 1998). Wealth inherited from parents was assumed to be an important determinant of future reproductive success. When wealth is modeled as having the potential to generate more wealth, it can be shown that very small families can be optimal if the costs of raising children to adulthood are high. The models shows that the higher the cost of raising children, the more wealth it is optimal to allocate to each child rises, making the average family in such a population wealthier. In populations where the costs of raising children were low, higher fertility and lower levels of wealth inheritance were optimal, meaning the average family is poorer.

Such a model offers some insight into the paradox of the negative correlation between wealth and reproductive success that is commonly observed. Within each model population, wealth is positively associated with reproductive success, but those model populations that had the poorest families also had the largest family sizes. Homogeneous populations, such as those from traditional cultures usually studied by anthropologists, frequently show a positive correlation between wealth and reproductive success, as predicted from evolutionary ecological theory. However, the large, heterogeneous populations usually studied by demographers typically show a negative correlation between wealth and reproductive success. If heterogeneous populations represent a group of subpopulations, each experiencing (or believing they experience) different costs and benefits associated with investing in children, then a decoupling or even a negative relationship between wealth and family size might be observed.

Empirical evidence is mounting that many human populations are structured in this heterogeneous way. However, this does not explain directly why different levels of parental investment across subpopulations have arisen. Formal population genetic models, where the quality of offspring is assumed to have long-term consequences for their competition against others, and hence for their future reproductive success, are needed. It is from these models that an understanding of the evolutionary basis of high parental investment might emerge, although as yet this branch of life history theory is relatively underdeveloped.

Even if a plausible evolutionary model predicting the origin and maintenance of very high parental investment and/or low fertility can be built, the existing empirical evidence from modern populations suggests that, at least across two generations, contemporary low fertility is not adaptive. In a survey of New Mexican men conducted in the early 1990s, it was found that those born into smaller families had greater educational achievements and earning power, but this did not translate into higher Darwinian fitness (Kaplan *et al.*, 1995). However, these individuals lived in a period of economic and population expansion – extra parental investment might pay better fitness returns in a stable population where competition is more intense. It is possible that low fertility combined with high parental investment may yet prove a successful reproductive strategy over the very long term, but there is necessarily no reason to believe this will be so. From an evolutionary perspective, we appear to be overinvesting in children's education. Competition between children, and between children's parents, appears to be driving a form of 'runaway parental investment' that pays good financial returns to children, but not to their fitness.

Our psychology may simply not be adapted to the range of choices that we now have with regard to our fertility, and hence this new range of preferences is being exposed to strong selection for the first time. There is some evidence that number of own siblings is a good predictor of number of own children in postindustrial populations, and there is also evidence that this heritability has some genetic basis (Rodgers *et al.*, 2001). If this is true, it is possible that a low desire for children, which until the advent of contraception would not have been an achievable desire, will now be subjected to strong negative selection. Over the very long term, fertility may start to rise again in countries that have undergone fertility decline as a low desire for children is selected out.

SUMMARY

Life history theory provides a useful theoretical tool with which to explore human reproduction.

Its use is being advocated by a number of behavioral ecologists and reproductive ecologists, but is only just being realized by traditional demographers. Despite its utility, practitioners have difficulties in explaining why humans have recently opted for low fertility beginning with the demographic transition in industrialized Europe. However, models that take into account the trade-offs inherent to human reproduction are beginning to explain this enigma.

References

Bentley GR (1999) Aping our ancestors: comparative aspects of reproductive ecology. *Evolutionary Anthropology* **7**: 175–185.

Boyd R and Richerson PJ (1985) *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Foster C (2000) The limits to low fertility: A biosocial approach. *Population and Development Review* **26**: 209–234.

Goodman MJ, Estioko-Griffin AA and Grove JS (1985) Menarche, pregnancy, birth spacing and menopause among the Agta women foragers of Northeastern Luzon, the Philippines. In: Bion Griffin P and Estioko-Griffin AA (eds) *The Agta of Northeastern Luzon: Recent Studies*, pp. 147–157. Cebu City, Philippines: University of San Carlos Publications.

Hill K and Hurtado AM (1996) *Ache Life History: The Ecology and Demography of a Foraging People*. Hawthorne, New York: Aldine de Gruyter.

Howell N (2000) *The Demography of the Dobe !Kung: Evolutionary Foundations of Human Behavior*, 2nd edn. New York: Academic Press.

Irons W (1998) Adaptively relevant environments versus the environment of evolutionary adaptedness. *Evolutionary Anthropology* **6**: 194–204.

Kaplan H, Lancaster J, Bock JA and Johnson SE (1995) Does observed fertility maximize fitness among New Mexican men? A test of an optimality model and a new theory of parental investment in the embodied capital of offspring. *Human Nature* **6**: 325–360.

Lack D (1968) *Ecological Adaptations for Breeding in Birds*. London: Methuen.

Lancaster JB and Hamburg BA (eds) (1986) *School-Age Pregnancy and Parenthood*. New York: Aldine de Gruyter.

McAllister F and Clarke L (2000) Voluntary childlessness. In: Bentley GR and Mascie-Taylor CGN (eds) *Infertility in the Modern World: Present and Future Prospects*, pp. 189–237. Cambridge, UK: Cambridge University Press.

Mace R (1998) The co-evolution of human fertility and wealth inheritance. *Philosophical Transactions of the Royal Society of London B*: **353**: 389–397.

Rodgers JL, Kohler HP, Kyvik OK and Christensen K (2001) Behavior genetic modeling of human fertility: findings from a contemporary Danish twin study. *Demography* **38**: 29–42.

Further Reading

Barkow JH, Cosmides L and Tooby J (eds) (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Borgerhoff-Mulder M (1987) On cultural and reproductive success: Kipsigis evidence. *American Anthropologist* **89**: 617–634.

Cronk L, Chagnon N and Irons W (eds) (2000) *Adaptation and Human Behavior: An Anthropological Perspective*. New York: Aldine de Gruyter.

Dawkins R (1976) *The Selfish Gene*. New York: Oxford University Press.

Ellison PT (2001) *On Fertile Ground: A Natural History of Human Reproduction*. Cambridge, MA: Harvard University Press.

Kaplan H (1996) A theory of fertility and parental investment in traditional and modern human societies. *Yearbook of Physical Anthropology* **39**: 91–136.

Stearns SC (1992) *The Evolution of Life Histories*. New York: Oxford University Press.

Material Culture

Intermediate article

Ofer Bar-Yosef, University of Harvard, Cambridge, Massachusetts, USA

CONTENTS

The evolution of human material culture over the past 2.5 Ma, as seen through the archaeological windows of tool use, the controlled use of fire, and prehistoric art, provides interesting clues about the parallel evolution of human cognition as both a cause and an effect of these cultural changes.

USE OF TOOLS

Since research into human prehistory began in the nineteenth century, stone tools have been considered the marker of human activity. During the twentieth century, investigations of primate behavior, and particularly of chimpanzee behavior, modified this view. Observations of chimpanzees and orang-utans in their natural habitats revealed how they make simple vegetal artifacts for instant use. An exception is West African chimpanzees, which employ naturally-shaped hammer-stones and large cobbles to crack nuts.

Experimental study of the bonobo (pygmy chimpanzee) Kanzi demonstrated that apes are not capable of the same degree of precision as humans in tasks requiring hand dexterity. The physical limitations in hand movements explain why non-human primates do not systematically chip stones as humans do. This difference between humans and non-human primates has been expressed since about 2.5 Ma ago in the archeological record by the consistency of humans in making and shaping stone tools. Owing to the uneven preservation of objects in most Pliocene and Pleistocene sites, wooden items, which were probably employed by humans, have only rarely been discovered in contexts dated to the last 0.4 Ma.

The basic method of making a stone tool is to break a nodule ('core') of rock by hitting it with another rock ('hammer'), and then to use the sharp edges of the detached piece ('flake'). When the action is repeated, additional flakes are chipped off the core. During the last 2.5 Ma various techniques for obtaining blanks with sharp edges were developed. Some of these are named after the sites where they were first discovered.

Owing to the poor preservation of human-made or human-used wooden objects, research into prehistoric material elements concentrates, for most of the earlier periods, on the study of stone tools. This research includes the reconstruction of the process of the actions required for detaching blanks from a nodule, known as the 'operational sequence' (chaîne opératoire). Effectively, this means the sequence of events involved in making (knapping) stone tools from a particular raw material in the form of a nodule or cobble, through the initial phases of decortication (the removal of the natural cortex of the nodule), and the detachment of the first blanks obtained by direct (hammer) or indirect (hammer and chisel) percussion. Subsequent flaking often requires the knapper to reshape the volume of the core. Hence, the detachment of the blanks is interspersed with chipping aimed at maintaining the desired form of the core. The products of core reshaping are known as 'core-rejuvenation pieces' or 'core-trimming elements'. The optimal situation for the prehistorian who is attempting to trace the decisions of a knapper is when all or most of the pieces can be fitted together to recreate the original nodule. Of course, the final products that were taken from the site to be used cannot always be found in the excavated site, but the general sequence of the core reduction can be analyzed.

Chronology of Human Tool-making

Early prehistoric times are subdivided into three major periods within the Old Stone Age (Paleolithic): Lower, Middle and Upper Paleolithic; or Early, Middle and Late Stone Age in Africa. These periods were in common use before the introduction of radiometric dating techniques. Geological and archeological research of the Late Pliocene and the Quaternary resulted in the establishment of a simplified sequence of stone industries. Thus,

the view that stone tools are chronological markers for each Stone Age period is now obsolete.

The first type of stone tools, uncovered in the geological contexts dated to about 2.5 Ma near Lake Turkana in Kenya and the Afar basin in Ethiopia, are classified as a 'core-and-flake industry'. The debate over whether these early stone tools were produced by the hominins known as *Homo habilis* or by *Australopithecus robustus* remains unresolved, but most scholars regard the early toolkits in stratified sites of the Late Pliocene and Early Pleistocene as manufactured by members of the *Homo* evolutionary line.

Oldowan assemblages

Cores and detached flakes from the early sites have been classified and given names such as 'choppers' and 'core-scrapers', and are all incorporated under an entity named the 'Oldowan' culture, after the extensive excavations in Olduvai Gorge in Tanzania. The Oldowan assemblages also contain other tool types, such as the spheroids, which resemble stone balls. Except for those that attained a rounded shape by serving as hammer-stones, their function is generally unknown. Other types include flakes whose sharp edges were reshaped by direct percussion (a process known as 'retouching') to create a desired form. Certain kinds of retouch could have been the results of use.

The Acheulian Industrial Complex

The next improvement, which did not require a cognitive change, was the modification of the shape of a cobble or a large flake by chipping both faces to create what we call 'hand-axes' or bifaces. Bifaces are associated with an entity known as the 'Acheulian Industrial Complex'. They exist in a variety of forms, including some that are almond-shaped and pointed and some that are oval and rounded. U-shaped bifaces, with what is considered a straight cutting edge opposite the rounded base, are called 'cleavers'. The production of Acheulian bifaces continued for at least 1.5 Ma. Most of the later ones are highly symmetric, indicating increased attention to and investment in chipping both faces. The appearance of symmetrical forms may have marked a greater efficiency of what is considered an 'all-task tool' and an evolving level of awareness of esthetics.

Acheulian tools are found in Africa, Europe (except for its eastern part), western Asia, and India. Only a few assemblages of bifaces have been found in eastern Asia. These tools are often considered to be the product of *Homo erectus*, with their spatial distribution supposedly reflecting the dispersal of *Homo erectus* from Africa into Eurasia. The discovery of early hominins in Dmanisi, in Georgia, with an Oldowan-type ('core-and-flake') stone-tool assemblage, may indicate that early hominins were able to adapt to new environments without bifaces. Furthermore, this tool form was probably invented more than once in the course of the Pleistocene period.

Fossil hominins from the Middle Pleistocene (780 000 to 130 000 years ago) display a physical and morphological variability across Africa and Eurasia. In Africa, Acheulian assemblages were the dominant kind, while in Eurasia both core-and-flake industries (such as the Clactonian, or the assemblages of Zhoukoudian in northern China) and Acheulian toolkits were spread in an uneven geographical distribution. In the well-preserved contexts of lignite deposits currently being quarried in Schöningen in Germany, a series of wooden spears have been uncovered in a site that dates to about 400 000 years ago. Similar isolated finds from the latter part of the Middle Pleistocene are already known from Lheringen in Germany and from Clacton on Sea in the UK.

Among the notable populations that emerged in the second part of the Middle Pleistocene, but are better known from the Upper Pleistocene (130 000 to 11 000 years ago) are the Neanderthals and the Modern Humans. Molecular and nuclear genetic evidence acquired from today's world populations indicate that Modern Humans emerged in sub-Saharan Africa between 300 000 and 100 000 years ago. Some Modern Humans colonized Australia, beginning about 60 000 years ago. Others, known as Cro-Magnons, later moved across Europe – perhaps originating in the Nile valley or the Levant – splitting the territories of the Neanderthals, in a process that lasted from about 40 000 to 30 000 years ago. The Neanderthals, a native population of Europe, emerged at least 300 000 years ago, and expanded from the western provinces to beyond the Caspian Sea. They were well adapted to a variety of periglacial, temperate, and Mediterranean environments of Europe and western Asia.

The Levallois technique

All late Middle Pleistocene humans, from about 300 000 years ago, employed a variety of operational sequences for producing stone tools. The best known is the Levallois technique, which required a new cognitive model. In this stone-knapping technique, the volumetric concept of the nodule and the desired design were predetermined. The clear goal of obtaining a series of useable

blanks (whether we define these as flakes, blades or points) was required before commencing the process of blank detachment. In the Levallois technique the form of the core was usually relatively flat. A series of different preparatory flaking stages was required before the desired blanks were produced. This could be from one primary knapping direction, or from others (identified as 'convergent', 'bidirectional', and 'radial'). So the blanks, and the scars left after the removal of the cortex (the natural face of the nodule), disclose the method preferred by the knapper.

Some tool types, such as Levallois points (often hafted as spear points), were manufactured without the need for secondary retouch. The edges of other blanks were shaped by retouch, and their most common form is classified as side-scrapers. Subsequent resharpening altered the form of the cutting or scraping edges.

Among the Middle Palaeolithic industries, several geographical varieties of stone tools should be mentioned. In northern Africa, the 'Aterian' entity is characterized by tanged Levallois points and scrapers, as well as bifacial leaf-shaped points. Similar bifacial points have been found across Africa, without evidence for continuous geographic spread in Europe in archeological entities such as the Szeletian.

Middle Palaeolithic humans were also able to produce blades, which are defined as flakes whose length is more than twice their maximum width. Blades are considered the cultural hallmark of the Upper Paleolithic cultures, and *Homo sapiens sapiens* the manufacturer. Research in the last two decades has demonstrated that blades, which were generally produced from cores with a semicylindrical volume, were manufactured in different regions at different times. Given the constraints on how an artisan can knapp stone cobbles, it is not surprising to find that blades were made in eastern Africa and the Levant some 250 000 ago, in north-western Europe some 100 000 years ago (by Neanderthals), in southern African humans some 80 000 to 60 000 years ago (in an archaeological entity known as the Howieson's Poort), and then in proliferation since about 50 000 years ago in western Asia, northern Asia, the Indian subcontinent, and Europe. In most of southern Asia, the simple core-and-flake industries predominated, and were produced by Modern Humans.

Stylistic variability

The Upper Paleolithic or Late Stone Age is traditionally considered the time when Cro-Magnons, or Modern Humans, expanded across Eurasia and Africa. In Africa, the change in material culture appears to have been gradual. However, in Eurasia, except in south-eastern Asia, the acceleration in lithic techniques, art, and the use of body decorations during the Upper Paleolithic is clearly reflected in the cultural remains. One of the traditional characteristics of the Upper Paleolithic is the production of blades from nodules where, following the process of decortication, the volumetric concept motivates the knappers to achieve cores with a semicylindrical to fully cylindrical shape. Blades were detached from these cores either from one striking platform or from two opposing ones.

One distinctive attribute of the Upper Paleolithic (and later) stone tools is the regional and relatively rapid shifts (within a hundred or a few thousand years) in artifact design, interpreted as reflecting stylistic variability. There is evidence of long-distance exchange networks in lithics and other raw materials ranging over several hundred kilometers. Improved hunting tools were invented, such as spear-throwers, and eventually bows and arrows and boomerangs. There was clearer spatial organisation within natural habitations and hunting stations. Pit-houses were constructed in open-air sites. The design of natural shelters changed: careful excavations at well-preserved sites uncovered the locations of human activities (kitchen areas, sleeping grounds, knapping floors, butchering stations, and the like). Storage facilities, often underground, were constructed. Bone and antler artifacts that served for daily and ritual uses were systematically produced. Body decorations made from marine shells, animal teeth, and ivory were made. There emerged also mobile imagery (human and animal figurines), decorated and carved bone, antler, ivory and stone objects, and representational and abstract images and signs, either painted, engraved or both, in caves and rock-shelters, as well as on exposed outdoor rocky surfaces.

During the several millennia of the Terminal Pleistocene, when the retreat of the glaciers led to environmental improvements, human populations expanded over the entire globe. This is the time when the Americas were colonized by groups coming from Asia, carrying their original lithic technologies. Subsequent human dispersals into these continents, and adaptation to the new environments, led to the emergence of particular local methods for making stone tools, characterized by the production of bifacial points such as the Clovis, Folsom, and others.

In Eurasia and Africa, the variability of lithic types, and the dominance of microliths, are well recorded, except for south-eastern Asia, where the old core-and-flake industries were maintained until adzes, bone and shell tools began to be produced.

The Terminal Pleistocene and Early Holocene are also characterized by the establishment of societies of complex hunter–gatherers, forming a social arena for the emergence of agriculture in certain areas (south-western Asia and China).

Tools and Human Anatomy

The relationship between human anatomy and the production and use of stone tools is difficult to trace. The size and shape of thumb joints in robust australopithecines and early *Homo* suggest that these hominins had the ability to grip objects both firmly and precisely. These abilities may be adaptations that permit the systematic detachment of flakes from a cobble, which characterizes the oldest stone tools. Thus, the earliest stone tools may have been fabricated by more than one species. Experiments with the bonobo Kanzi have demonstrated that chimpanzees lack these skills: in nature, chimpanzees sometimes use unmodified stones for nut cracking, but they do not detach flakes from a nodule.

Later in human evolution there is more abundant evidence for tool use. Middle Paleolithic hominins, especially Neanderthals, have upper-arm asymmetries indicating a 'dominant arm', and heavily-muscled shoulder joints that indicate extensive tool use, possibly throwing heavy spears. In addition, some differences in hand grip have been observed between remains of roughly contemporary Neanderthals and early Modern Humans (the Qafzeh-Skhul group), in spite of the archeological evidence that both taxa fabricated and used similar tool assemblages.

Finally, technological changes in food production and processing technologies may be related to decreases in both facial size and tooth size, especially since the end of the Pleistocene.

Tools and Human Cognition

The formal classification of prehistoric stone objects expressing variability was originally intended to identify the relative chronology within the Late Pliocene and the Quaternary. When radiometric dating techniques were introduced, tool forms and how they were obtained became the subject of a search for meaning. Assuming that stone tools were used for purposes similar to those of metal tools in hunting, fishing, trapping, food preparation, carpentry, basketry, and the like, it was natural to suppose that the different forms were determined by their functions. However, numerous replications and microscopic examinations of stone tools by researchers in several countries suggest that, while there is a certain correlation between form and function, many forms can serve for the same or similar tasks. Hence, the study of style, as reflecting the mental process of the human actor, was reintroduced into prehistoric research. Style is thought to reflect formal variation and communication. Because of the lack of organic remains (except for bones, antlers, and horn-cores) prior to the later historical periods (except for rare cases of waterlogged sites in which organics are better preserved), most of the debate concerning the meaning of style focuses on stone tools.

Early hominins produced simple stone tools, and although they were regularly obtained from nodules of pebbles, they did not exhibit the kind of symmetry found in the bifaces dated to the Middle Pleistocene. Cognitive models seem to have first appeared, as mentioned above, with the Levallois technique, which required a predetermined design. Blanks were selected for use, and in particular for secondary retouch, which shapes the active edge. However, continuous resharpening often creates variable forms, not necessarily predicted by the original design. Although not a global practice, during the Upper Paleolithic, knapping blades produced blanks that were shaped by retouch into various forms, such as points, burins (used for a variety of tasks), scrapers, and borers.

Interestingly, during the closing millennia of the Pleistocene, people in many regions of the world began to manufacture small tools, known as 'microliths'. These were shaped from small blades and flakes, and the pointed forms among them were hafted as arrowheads and barbs.

Hence, style reflects the particular design preferred by a social entity. Style is therefore an important social marker. The shift in design, when suitability for function overrides style, is better expressed in the following Neolithic period. At that time, when farming communities established themselves in a few areas and dispersed across the world, tool types such as sickle blades, perforators, adzes, and axes expressed the specific functions of the human actors. Thus, the style of stone tools in the last 50 000 years carries the social message more clearly than in any preceding period.

CONTROLLED USE OF FIRE

In terms of human evolution, the intentional use of fire is regarded as improving the nutritional values of vegetal and animal foods, as well as having implications for social organization (e.g., a place for focal gatherings). A few archeologists believe that the evidence for the use of fire begins with *Homo erectus*, some 1.5 Ma ago; but archeological contexts do not demonstrate the presence of hearths before 500 000 years ago, even though we find burned bones.

A certain set of visible features is considered direct evidence for the intentional use of fire. These features are (from lower to upper): the reddened soil produced by heat from the fire; the black layer of partially-burned substances; and ashes, which are white in color if well preserved. Frequently, however, post-depositional processes, such as leaching by water and burrowing by animals and insects, destroy hearths. Laboratory techniques, including mineralogical analysis of thin sections of artificially-consolidated blocks of the prehistoric deposits, have proved useful in recognizing the use of fire by humans. This technique, known as micromorphological analysis, can also identify the mineralogical elements of the ashes long after they change color and contextual appearance, and identify those blackened bones that have been next to a fire, rather than stained by manganese.

Caves located in the Mediterranean climate areas (including southern Africa) contain ample evidence for the use of fire by humans. Hearths for heating, cooking, and parching were built by humans everywhere even before the Middle Paleolithic, and around most of the well-preserved hearths there is evidence for collecting firewood, represented by the remaining phytoliths. The consumption of cooked food influenced the masticatory system of humans, and it has been proposed that the reduction in tooth size is related to this.

ART

Works of prehistoric art are seen today as variable expressions of the minds of their producers, users, and viewers. The symbolism of the images, whether painted, engraved, incised or carved, conveys various meanings depending on the social unit and the time of their creation. Art objects may indicate the evolving self-awareness of humans and their intricate relationships with their own selves and the outside world. Therefore the search for the earliest art is considered crucial for our reconstruction of the evolution of human cognition, and is a controversial issue.

The earliest manifestations include objects like the incised Late Acheulian figurine from Berekhat Ram in the Levant, and a few objects from Mousterian contexts. There is no doubt that the full bloom of imagery expressions began, as with the systematic use of body decorations, during the early millennia of the Upper Paleolithic period, although not in every region of the world. The Franco-Cantabrian region is the most famous and the richest, but somewhat similar phenomena are found in other areas such as Australia.

European cave and rock imagery was discovered in the nineteenth century, and the history of its study is an eventful story which is still unfolding, and often triggers rival interpretations. Scholars agree that the variability of the subjects (from portable human figurines to animals and schematic signs) conveys a set of symbols that reflect the beliefs of ancient populations. The topics under debate essentially concern the identity of the makers or wall painters, their position within their own society, and the occasions of the production and use of objects and rock art as daily, seasonal, or annual events.

A schematic classification suggests that all symbolic expressions can be divided in terms of material and size into: elements of personal jewelry (made from bone, teeth, ivory, etc.); portable objects (made from bone, antler, ivory, terracotta, and hard rocks); engraved and incised slabs ('plaquettes'); and finally parietal images carved, incised and painted on cave walls and open-air rocky surfaces, which could be in a living area or rock-shelter, or in deep, dark caves and narrow passages.

Portable objects – including functional tools such as spear throwers, shaft straighteners, animal and human figurines, decorated pendants, and other small objects – are geographically spread across all regions. These objects probably served the people on a domestic, family or band level, reflected social alliances, and were probably passed on through generations. The images reflect beliefs in human and natural powers and spirits, including those of fertility, recorded information, ownership, and hunting magic.

The dominant subjects of the painted, engraved open-air rock surfaces and cave walls are the animals. The most common ones are those that inhabited temperate and Mediterranean Europe. They include deer, horse, bison, aurochs, mammoth, rhinoceros, and carnivores such as cave bear, brown bear, and felines, as well as fish (mainly on portable objects), tortoise, birds (swan,

duck and heron), rare plants, and humans (commonly represented by hands).

Since the first discovery of rock art in western Europe, attempts to establish the relationship between the dated archeological cultures and the images have been connected with various interpretations. These include: the uncommon view of 'art for art's sake'; totemism and sympathetic magic as derived from ethnographic parallels; and the binary-structuralist analysis, opposing the most common two animals in each cave site (e.g., bison and horse), as male and female symbols.

In recent years, the empiricist approach, based on correlations between ethnographic reports of the San in southern Africa, from the nineteenth and twentieth centuries, and the rich recorded images from local rock-shelters, have demonstrated the association between the rock art and the cosmology, symbols and rituals of San shamanism. Shamans performed their spiritual 'duties', sometimes in altered states of consciousness. The responsibility of prehistoric shamans, and probably of other prominent leaders of hunter–gatherer bands, in creating and modifying parietal art, is generally accepted.

The role of shamans today includes facilitating connections with the spiritual world among foragers. It is therefore suggested that comparisons between the activities of recent shamans (and other members of their society) and the preserved historical 'art' may help investigators to achieve a better understanding of Upper Paleolithic imagery and of later prehistory.

CONCLUSION

Studies of the development of human material culture produce a wealth of evidence concerning the evolution of human cognition, although direct interpretation of this evidence is fraught with difficulty and likely to remain so.

Further Reading

Bahn PG and Vertut J (1988) *Images of the Ice Age*. New York: Facts on File.

Belfer-Cohen A and Goren-Inbar N (1994) Cognition and communication in the Levantine Lower Paleolithic. *World Archaeology* **26**(2): 144–157.

Deacon T (1997) *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York: Norton.

Dunbar RIM (1996) *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard University Press.

Gibson K and Ingold T (eds) (1993) *Tools, Language and Cognition in Human Evolution*. Cambridge, UK: Cambridge University Press.

Klein RG (1999) *The Human Career: Human, Biological and Cultural Origins*. Chicago: University of Chicago Press.

Kuhn SL *et al*. (2001) Ornaments of the earliest Upper Paleolithic: new insights from the Levant. *Proceedings of the National Academy of Sciences of the USA* **98**(13): 7641–7646.

Marshack A (1972) *The Roots of Civilization: The Cognitive Beginnings of Man's First Art, Symbol, and Notation*. New York: McGraw-Hill.

McGrew WC (1992) *Chimpanzee Material Culture: Implications for Human Evolution*. Cambridge, UK: Cambridge University Press.

Social Learning in Animals

Intermediate article

Bennett G Galef Jr, McMaster University, Hamilton, Ontario, Canada

CONTENTS

'Social learning' is a general term referring to several behavioral processes that allow social interactions to bias what individuals learn. Processes involved in social learning include 'local enhancement', when the normal activities of one individual simply focus attention of others on a particular part of the environment with which they then interact, and 'teaching', when a model changes its own behavior to facilitate learning by naive individuals.

INTRODUCTION

Understanding the behavioral processes that promote diffusion of behavioral traditions through animal populations, and comparing these processes with those that support culture in human societies, suggests important similarities as well as important differences in the processes that support social learning in humans and animals. This article discusses a variety of such social learning processes that are common to both humans and animals, and two of these processes, namely imitation and teaching, that may be used only by humans and their closest relatives.

WHAT IS SOCIAL LEARNING?

Understanding the role of behavior in promoting survival and reproduction is a goal of life scientists working in a variety of disciplines. Students of animal social learning are particularly interested in the question of how interactions among members of a species affect development of their behavioral repertoires. Social learning is only one of several factors that interact to influence behavioral development. For example, ethologists studied instinctive patterns of behavior produced by natural selection acting on heritable variation, whereas students of animal learning were and are interested in how individual experience of events in the physical environment shapes behavior.

EXAMPLES OF SOCIAL LEARNING

Contemporary interest in social learning arose from the observation that members of one free-living population of a species often behaved quite differently to members of other populations of the same species. For example, chimpanzees living to the west, but not to the east, of the Sassandra-N'Zo river on the Ivory Coast use stones to crack nuts (Whiten *et al*., 1999). Of course, population-specific behaviors such as nut cracking could reflect differences in either the genetic substrate of populations (i.e., different subspecies of chimpanzee might inhabit the two banks of the Sassandra-N'Zo river) or differences in the physical environments in which the two populations live (e.g., there might be no nuts on the east bank of the river). In fact, there are no known genetic differences between chimpanzees on the two banks of the Sassandra-N'Zo, and similar nuts and stones are found in both places.

The nut-cracking example is not unique. There are often systematic differences in the behavior of populations of a species even when there is no evidence of genetic or environmental differences between populations (Whiten *et al*., 1999). For example, chimpanzees living in Gombe National Park in Tanzania use twigs or blades of grass to feed on ants and termites. These implements are used to probe the passageways of insect mounds, and when the residents attack an intruding probe, the chimpanzees extract the probe and eat any termites that are clinging to it.

There is variation among chimpanzee populations not only in the species of insect preyed upon and the materials used as probes, but also in how probes are prepared for fishing and how insects are captured. For example, chimps at Gombe in Tanzania hold a long stick in one hand and use the other hand to wipe a ball of ants into their mouths. Chimps living in the Tai Forest on the Ivory Coast use a short stick to collect ants, and they place the

stick directly into their mouths, removing the ants with their lips and tongue. Chimpanzees at Assirik in Senegal usually peel bark from twigs before using them as probes, whereas chimpanzees at Gombe do not peel the twigs before using them. Such observations have led many scientists to conclude that at least some population-specific behaviors of chimpanzees are traditions learned by one individual as a result of observing the behavior of another individual (McGrew, 1992; Whiten *et al*., 1999).

It has been known for many years that some animals can learn complex patterns of behavior by imitating others of their species. 'Imitation' is a special type of social learning that has been defined in different ways by different scientists. For example, imitation has been defined both as 'copying of a novel or otherwise improbable act' (Thorpe, 1956) and as 'learning to do an act from seeing it done' (Thorndike, 1898). An example of the former is to be found in the songs that adult male songbirds produce both to attract conspecific females and to repel conspecific males. Males from different populations of many species produce different variants of a basic, species-typical song, and laboratory studies have shown that a 'song dialect' is produced only by males that, as juveniles, heard adults of their species sing that dialect (Marler, 1970).

Thus we know both that even birds (with their relatively small brains) can learn song dialects by imitation, and that chimpanzees from different populations engage in different feeding behaviors. Why should we not simply conclude that if a difference in the behavior of two groups of chimpanzees cannot be explained easily by reference to differences in either their genes or their ecology, the different behaviors have been transmitted by imitation among chimpanzees within each group?

Although birds learn their songs by imitation, as we shall see below, this may be a special case. More often, when a possible instance of learning by imitation observed in natural circumstances is brought into the behavioral laboratory for analysis, it is found that the learning depends on some social learning process that is not truly imitative. For example, rats living in the pine forests of Israel are unique in that they, and no other rats, survive by stripping the scales from pine cones and eating the pine seeds that the scales protect (Aisner and Terkel, 1992). It is difficult for rats to strip pine-cones in a way that allows them to get more energy from the pine seeds than is consumed in obtaining them. The rats must start by removing the scales from the base of a cone, and then take advantage of the cone's architecture by removing the remaining scales in a spiral from the base of the cone to its apex. A more direct method, which involves gnawing through individual scales to access seeds, is used by rats from populations that do not typically feed on pine-cones if given access to them. However, this method consumes more energy than it produces.

In the laboratory, rats born to pine-cone-stripping mothers grew to strip pine-cones in the efficient manner only if they were reared by adult rats that demonstrated pine-cone stripping, and not if they were foster-reared by adult rats that did not know how to open cones efficiently. Furthermore, even young rats reared by adults that did not know how to strip pine-cones became efficient strippers of seeds when provided with cones started appropriately by either an adult rat, or a human using a pair of pliers to remove scales from the base of the cones. Apparently, interaction with cones started in the right way guides the development of behavior in young black rats, making them efficient strippers of seeds from pine-cones. In nature, young rats probably snatch partially opened cones from adults, and by interacting with the cones learn how to finish the job.

The study of social learning in Israeli rats has yielded two important findings. First, complex motor patterns can be transmitted from one generation of animals to another. Second, the existence of a complex tradition of behavior in a population of animals cannot be used to infer that a complex social learning process, such as imitation or teaching, was involved in its transmission.

CLASSIFYING SOCIAL LEARNING

Some authors have suggested that describing a behavior as traditional is totally uninformative unless the spread of the behavior through a population can be attributed to a particular social learning process. This is almost certainly an overstatement. Referring to a behavior as traditional implies that the spread of that behavior through a population was facilitated by social interactions (i.e., that each group member did not learn the behavior independently). Still, the assertion that describing a behavior as traditional is uninformative raises an important issue. As we have already seen in the cases of bird-song learning and pine-cone stripping, animals may influence one another's behavior in quite different ways. Consequently, for those with an interest in understanding how behaviors develop, describing a population-specific behavior as traditional answers relatively few questions and raises many.

While almost everyone seems to find instances of traditional behavior in animals intrinsically interesting, many outsiders to the field of social learning find attempts to categorize the many ways in which social interactions affect the acquisition of behavior either boring or impenetrable. Admittedly, attempts to define various social learning processes have produced quite complex and not altogether satisfactory vocabularies describing the many ways in which social interactions can bias behavioral development (Whiten and Ham, 1992). However, such attempts at categorization are important because they make explicit the fact that social learning can occur in many different ways in both humans and other animals.

One rather simple way in which one animal can bias development of the behavior of others, thereby facilitating spread of the behavior through a population, is by focusing the attention of others on particular parts of the environment. For example, if one Norway rat sees others eating in a particular place, or smells rat odors left at a potential feeding site by other rats that have eaten there, that rat is much more likely to begin eating whatever food is at the socially marked site than to start eating at unmarked sites where different foods might be available. Because wild Norway rats are extremely reluctant to eat unfamiliar foods, anything that causes a rat to begin eating one food rather than another has a profound effect on the rat's subsequent food choices.

Such social biasing of learning by other individuals has been termed 'local enhancement', defined as 'apparent imitation resulting from directing the animal's attention to a particular object or to a particular part of the environment' (Thorpe, 1956). Local enhancement can be contrasted with imitation, defined as 'copying of a novel or otherwise improbable act' (Thorpe, 1956) or 'learning to do an act through seeing it done' (Thorndike, 1898).

In local enhancement, an animal learns only that it should interact with one part of the environment rather than with another. In true imitation, an animal learns directly about the behavior in which it should engage. The distinction is fundamental to all academic discussions of social learning in animals.

Are two terms – local enhancement and imitation – sufficient to enable discussion of all animal social learning? Unfortunately this is not the case. Humans and other animals can learn from the behavior of others, either directly or indirectly, in more than one way. For example, if a chimpanzee watches another of its species use a rake to pull in food items that would otherwise be out of reach, the observing chimp learns to use the rake to pull in food faster than it would if it had never seen another chimp use a rake (Tomasello *et al*., 1987). However, observer chimps failed to imitate in the sense of copying the particular actions used by the demonstrator to obtain food. Rather, observers developed their own techniques for using the rake. Observers seemed to learn that a rake was useful for acquiring food, but did not learn much about the actual behavior that a model uses when raking and, as noted above, what is meant by imitation is 'learning to do an act'. Tomasello *et al*. (1987) proposed that the observers were not so much copying the model's behavior as attempting to create the results of the model's efforts, a process that they termed 'emulation.'

Finally, consider teaching, an activity which contributes to social learning and that is common in our own species. In local enhancement, imitation or emulation, the model is essentially passive. The observer extracts information from a model engaged in activities it performs without reference to the observer. In teaching, according to the most widely used current definition (Caro and Hauser, 1992), a model modifies its behavior in the presence of a pupil, often leading to some reduction in the efficiency of the model's performance. Furthermore, the model either encourages or punishes the pupil, or provides the pupil with examples of behavior or experiences so that the pupil acquires information or skill more rapidly than it otherwise would.

The important elements of this somewhat complex definition, distinguishing teaching from other activities that play a part in social learning, involve potentially costly modification of the teacher's normal behavior, resulting in accelerated learning by pupils. Surprisingly, given the importance of teaching (at least in modern Western societies), there are few plausible examples of teaching in the animal world, and even those few instances are hotly debated. Feline mothers may meet the definition by delaying their killing and eating of prey and providing their young with incapacitated, live prey on which to practice predatory behavior. However, there is contradictory evidence as to whether experience with incapacitated prey facilitates the development of hunting behavior in young cats. Some authors have suggested that killer whales teach their young to hunt seals, although this view is not generally accepted (Rendell and Whitehead, 2001). Chimpanzees may teach juveniles how to crack open nuts using a stone hammer and anvil (Boesch, 1991).

It is worth quoting verbatim Boesch's (1991) descriptions of one such instance, as they provide

a fine indication of both the strengths and the weaknesses of unusual field observations. Ricci's 5-year-old daughter Nina is trying without success to crack open nuts using an irregularly shaped hammer. Ricci joins Nina, and Nina gives the hammer to Ricci.

> Then with Nina seated in front of her, Ricci, in a very deliberate manner, slowly rotated the hammer into the best position with which to pound the nut effectively. As if to emphasize the meaning of this movement, it took a full minute to perform this simple rotation. (Boesch, 1991, p. 532)

Ricci then cracks 10 nuts with the hammer in the correct position, and leaves. Nina then picks up the hammer, adopts the same grip that her mother used to crack nuts successfully, and she opens four nuts in 15 minutes. According to Boesch:

> In this example, the mother corrected an error in her daughter's behaviour and Nina seemingly understood this perfectly, since she continued to maintain the grip demonstrated to her. (Boesch, 1991, p. 532)

Some authors accept Boesch's interpretation of his observations, while others do not. The fact that such complex interactions between an apparent pupil and teacher have been seen only twice in many years of field study surely makes them difficult to interpret with certainty. Indeed the issue of whether animals teach is only one of a number of questions that are being actively debated by researchers who are interested in social learning in animals.

DISPUTES ABOUT SOCIAL LEARNING

For more than a century, two central questions have motivated much of the research on social learning in animals.

1. Which non-human animals, if any, imitate?
2. What is imitation and how can it best be distinguished empirically from other forms of social learning, such as local enhancement?

Can Animals Imitate?

The discovery over decades that many different types of social learning can bias behavioral development has made it increasingly difficult to determine whether any given case of social learning is a product of true imitation or of some other less cognitively demanding process. Why should anyone care enough about the types of social learning that animals employ when using one another as sources of information about the environment to actually argue about it? Many believe that understanding similarities and differences in social learning processes in humans and other animals will provide insight into similarities and differences in their mental processes. Indeed, the first laboratory investigations of social learning in animals were undertaken in order to determine whether, like humans, non-human animals had access to mental representations that they could manipulate in order to solve problems (Thorndike, 1898).

To imitate the behavior of a model, an imitator must store a visual image of the model's behavior and then match its motor output to that stored representation. Earlier we considered the learning by birds of song dialect as a result of hearing the song of adults belonging to their social group. Such learning might be described as 'learning to sing a song from hearing it sung'. This is somewhat different from 'learning to do an act from seeing it done', which some hold to be the definition of imitation.

Why should a distinction be made between seeing and hearing when defining imitation? The task of learning to sing a song by listening may be far simpler than that of learning to do an act by seeing it. In order to learn to sing like another, all the singer needs to do is to match the sounds it produces when singing with a stored representation of the sound of the song of another. Thus song imitation can occur within a single modality, namely audition. On the other hand, when learning to perform an act by seeing it done, a match has to be made across modalities, between a stored representation of a visual stimulus (the sight of another performing a behavior) and kinesthetic feedback from a motor act. Such cross-modality matching is necessary because even if an observer perfectly imitates an act performed by a model, the visual input to the observer while imitating is quite different to the visual input that the observer received when observing the model perform the act. For example, when I see someone bow, what I see is very different from what I see when I bow myself. On the other hand, when I whistle a song, what I hear is very similar to what I heard when someone else whistled the same song (Heyes, 2001). So when we are discussing imitation in animals, should we be limited to instances where overt motor patterns are copied, or should we include bird-song learning? Most authors think that we should not include bird-song learning.

There is even controversy about whether, in order for a behavior to be considered a result of imitation, the imitated behavior has to be new to the imitator and, if so, how one knows whether a motor pattern is really new. With regard to the first

point, how can you tell if an individual has learned to perform an act by seeing it done, if it has previously performed the act? With regard to the second, an individual may never have used a rake to pull in food before, but it has grasped objects, used objects to move other objects about, sought food, etc. An apparently novel act may be nothing more than a combination of acts that are already in an individual's repertoire. Perhaps all a subject does when it imitates a familiar act is to use the behavior of another as a cue as to which elements of its own behavioral repertoire it should try in the situation that it now faces. If you are interested in imitation as a tool for exploring the cognitive capacities of animals, then cases in which observers simply use others' actions to cue their own behavior are not very informative.

What is the Best Empirical Method for Discovering Imitation?

For many decades the predominant strategy for demonstrating imitation learning was first to determine whether watching the performance of a task allowed observers to learn to perform that task faster than animals that did not have the observational experience, and then by conducting additional experiments, to attempt to exclude all explanations of the accelerated performance other than imitation. As the number of alternative social learning processes described by scientists increased, the strategy of excluding alternatives became increasingly cumbersome. For example, if kittens are allowed to watch a demonstrator cat press a lever to obtain food, they subsequently learn to press the lever faster than kittens that do not see a cat receiving rewards for pressing the lever. To determine whether this accelerated acquisition of lever pressing was due to imitation of the cat by the kittens, one would have to exclude, among other things, the possibility that lever pressing is facilitated by either local enhancement of the lever or observing others eat. In general, such attempts have produced no convincing evidence of imitation learning in any animal.

At present, the 'two-action method' is the preferred laboratory procedure for discovering imitation in animals. In the two-action method, each subject sees a demonstrator receive rewards for manipulating an object in one of two ways. The observer is then given access to the object, and it is determined whether observers tend to manipulate the object in the same way as the demonstrator did. For example, chimpanzees first watch a human demonstrator either pull or twist bolts in order to open a closed box containing fruit, and they are then given a closed box. The most recent evidence suggests that chimpanzees tend to use the same action that they saw demonstrated to open the box (Whiten, 1998). Surprisingly, the two-action method has also produced evidence of apparent imitation in Japanese quail, starlings, pigeons, grackles, and budgies (Heyes, 2001), although other methods for demonstrating imitation have not provided evidence of the latter, even in animals as sophisticated as monkeys (Visalberghi and Fragaszy, 2002).

Perhaps more fundamentally, there is even a question as to whether evidence gathered in the field or evidence collected in the laboratory is more suitable for determining whether members of a species can imitate. Laboratory workers feel that the social learning mechanisms responsible for apparent imitation of one animal by another can be analyzed satisfactorily only in the controlled environment of the laboratory. However, some field workers suggest that the sterile environment and abnormal social conditions of the laboratory result in systematic underestimation of the imitative abilities of complex animals such as chimpanzees (McGrew, 1992).

Just How Important is Social Learning to Animals?

Recently, several students of animal behavior have argued in book-length monographs that social learning is central to the development of adaptive behavioral repertoires in a variety of animals (Avital and Jablonka, 2000; Dugatkin, 2000). These authors argue that social learning and natural selection are practically co-equals in producing adaptive behavior in animals, and that social learning of one type or another is necessary for animals to maximize everything from selection of a mate to avoidance of potential predators. Although social learning has been demonstrated to play some role in the mate choices of guppies and quail, and in the avoidance of predators by blackbirds and monkeys, claims of a major role for social learning in the evolution of behavior are relatively recent, and have yet to be evaluated. Still, the issue of just how important social learning is to understanding animal behavior is an open one, and is sure to be contentious.

THE ROOTS OF CULTURE

An intense area of debate concerns the degree of similarity between 'traditions' of animals and

'cultures' of humans. As is often the case in discussions of the relationship between human and non-human animals, some authors emphasize apparent similarities and others highlight apparent differences. The former group suggests, for example, that if an anthropologist were to describe populations of humans whose technologies and social customs varied as much as do those of well-studied populations of free-living chimpanzees (Whiten *et al*., 1999), the anthropologist would surely refer to the human populations as having different cultures. Therefore it is foolish to deny culture to chimpanzees (McGrew, 1992).

Those in the opposing camp look at the same data and focus on apparent differences between chimpanzee and human social learning. They argue that differences in the behavioral repertoires of various groups of chimpanzees tell us nothing about the processes responsible for the development of those differences. Moreover, if you are interested in discovering true precursors of human culture, you should look for behaviors that are transmitted from one generation to the next by the same processes that support human cultures (Galef, 1992). Humans can teach one another; they do learn by imitation. As we have seen, generally animals do not teach, and they rarely learn by imitation. Consequently, most instances of human culture and animal tradition may depend on rather different behavioral processes.

The type of culture that animals which do not teach or imitate could produce would be severely restricted. In human cultures, techniques or behaviors often become increasingly complex over generations. Youngsters learn from elders what the elders know. The young can then spend a lifetime improving on what they have learned socially, and then transmit those improvements as the starting point for the next generation. Such 'ratcheting' can occur only if social transmission involves learning directly about behavior, as in imitation or teaching, not when social transmission involves changes in attention to environmental stimuli, as in local enhancement or emulation.

A local enhancer only increases attention to some aspect of the environment. A model for emulation merely indicates that a tool can be used to achieve a goal. Consequently, in each generation, naive individuals whose learning was shaped by local enhancement or emulation must learn for themselves how to behave with respect to the part of the environment to which their attention was directed, and no ratcheting across generations can occur.

The argument goes on. Many, perhaps the majority, of those who study apes in their natural environments are convinced that these animals exhibit something quite similar to human culture. Many, perhaps the majority, of those who study social learning processes in animals in the laboratory believe that the differences between human culture and animal tradition are sufficiently great to require different terms to describe them. They often want to restrict use of the term 'culture' to humans, and to refer to 'animal traditions' when discussing population-specific behaviors of non-human animals.

Is There Anything Worth Arguing About?

At first glance, such arguments over terminology may seem arcane or even useless. However, our language both reflects and influences the way in which we think about the natural world. Those who believe that it is appropriate to discuss 'culture' in animals generally differ from those who prefer to talk about 'animal traditions' in their views of how similar the behavioral and mental capacities of human and non-human animals may be.

Understanding how we humans both resemble and differ from other animals is neither arcane nor useless. It is a basic part of our effort to discover what we are as a species, and to define the ways (if any) in which we are unique among animals. Thus unraveling the processes involved in social learning in animals, whether they are living free in their natural habitat or maintained in the more restricted laboratory environment, is but one of many ways available to us to increase our understanding both of ourselves and of our place in nature.

References

Aisner R and Terkel J (1992) Cultural transmission of pine-cone-opening behaviour in the black rat (*Rattus rattus*). *Animal Behaviour* **44**: 327–336.

Avital E and Jablonka E (2000) *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge, UK: Cambridge University Press.

Boesch C (1991) Teaching among wild chimpanzees. *Animal Behaviour* **41**: 530–532.

Caro TM and Hauser MD (1992) Is there teaching in nonhuman animals? *Quarterly Review of Biology* **67**: 151–174.

Dugatkin LA (2000) *The Imitation Factor: Beyond the Gene*. New York, NY: Free Press.

Galef BG Jr (1992) The question of animal culture. *Human Nature* **3**: 157–178.

Heyes CM (2001) Causes and consequences of imitation. *Trends in Cognitive Science* **5**: 253–261.

McGrew WC (1992) *Chimpanzee Material Culture: Implications for Human Evolution*. Cambridge, UK: Cambridge University Press.
Marler P (1970) A comparative approach to vocal learning: song development in white-crowned sparrows. *Journal of Comparative and Physiological Psychology* **71**: 1–25.
Rendell L and Whitehead H (2001) Culture in whales and dolphins. *Behavioral and Brain Sciences* **24**: 309–382.
Thorndike EL (1898) Animal intelligence: an experimental study of the associative process in animals. *Psychological Review Monographs* **2**(Supplement 4): 1–109.
Thorpe WH (1956) *Learning and Instinct in Animals*. London, UK: Methuen.
Tomasello M, Davis-Dasilva M, Camak L and Bard K (1987) Observational learning of tool use by young chimpanzees. *Human Evolution* **2**: 175–183.
Visalberghi E and Fragaszy D (2002) Do monkeys ape? Ten years after. In: Dautenhahn K and Nehaniv C (eds) *Imitation in Animals and Artifacts*, pp. 471–500. Cambridge, MA: MIT Press.
Whiten A (1998) Imitation of sequential structure of actions by chimpanzees. *Journal of Comparative Psychology* **112**: 270–281.
Whiten A and Ham R (1992) On the nature and evolution of imitation in the animal kingdom: reappraisal of a century of research. *Advances in the Study of Behavior* **21**: 239–283.
Whiten A, Goodall J, McGrew WC *et al*. (1999) Cultures in chimpanzees. *Nature* **399**: 682–685.

Further Reading

Avital E and Jablonka E (2000) *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge, UK: Cambridge University Press.
Box HO and Gibson KR (1999) *Mammalian Social Learning: Comparative and Ecological Perspectives*. Cambridge, UK: Cambridge University Press.
Boyd R and Richerson PJ (1985) *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.
Dugatkin LA (2000) *The Imitation Factor: Beyond the Gene*. New York, NY: Free Press.
Galef BG Jr (1976) Social transmission of acquired behavior: a discussion of tradition and social learning in vertebrates. *Advances in the Study of Behavior* **6**: 77–100.
Heyes CM and Galef BG Jr (1996) *Social Learning in Animals: the Roots of Culture*. San Diego, CA: Academic Press.
King BJ (1994) *The Information Continuum*. Santa Fe, NM: S & R Press.
Tomasello M (1990) Cultural transmission in the tool use and communicatory signaling of chimpanzees? In: Parker S and Gibson K (eds) *Language and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, pp. 274–311. Cambridge, UK: Cambridge University Press.
Wrangham RW, McGrew WC, deWaal FB and Heltne PG (1994) *Chimpanzee Cultures*. Cambridge, MA: Harvard University Press.
Zentall TR and Galef BG Jr (1988) *Social Learning: Psychological and Biological Perspectives*. Hillsdale, NJ: Lawrence Erlbaum.

Sociobiology

Intermediate article

Ullica Segerstråle, Illinois Institute of Technology, Chicago, Illinois, USA

CONTENTS

Sociobiology, the study of animal social behavior, is represented by a number of approaches, and scientists who practice sociobiology may call themselves by other names. Central sociobiological ideas include inclusive fitness, kin selection, the gene's eye view, reciprocal altruism and mutualism, and game theory.

INTRODUCTION

Sociobiology is part of modern evolutionary biology. It is concerned with explaining such things as the evolution of altruistic behavior, the regulation of animal conflict, and the differences in male and female 'strategies' (unconscious behavior patterns). The reasoning in sociobiology can often be expressed exactly in the mathematical language of population genetics and tends to be of a game-theoretical nature.

A central concept is that of 'inclusive fitness' associated with William D. Hamilton (1964). This concept explains why from a 'gene's eye' point of view it can make sense for an animal to behave altruistically towards others (who are typically relatives). Altruism is here defined in a behavioristic way, strictly focusing on the outcome of a behavior, not ascribing conscious motives. An altruistic act is an act where the donor incurs a considerable cost while helping a beneficiary. The ideas behind sociobiology were popularized by Edward O. Wilson (1975) in *Sociobiology: The New Synthesis* and Richard Dawkins (1976) in *The Selfish Gene*.

The Varieties of Sociobiology

There are important differences between the types of sociobiology introduced by Wilson and Dawkins. In his book Wilson defines sociobiology as 'the scientific study of the biological basis of social behavior in all kinds of organisms including man' and presents a broad synthesis of existing theories and research, including ecological considerations. His aim is to establish a new comprehensive discipline of comparative study of behavior. Dawkins's focus in *The Selfish Gene* is much narrower. He is firmly grounded in the population genetic tradition and limits himself to explicating the mechanisms of evolution in the light of the new ideas of Hamilton, George Williams, Robert Trivers, and John Maynard Smith, using a novel 'gene's eye' view of evolution. Wilson presents altruism as the central problem of sociobiology, but Hamilton's theory is introduced together with a number of group selectionist models. Dawkins limits himself to animals, but Wilson extends sociobiology to include the human species.

Currently the term 'sociobiology' is used to refer to a number of different things: (1) Wilson's view of sociobiology as a broad, comparative discipline, (2) the more narrow focus on a gene selectionist explanation of evolutionary mechanisms associated with the theorists presented by Dawkins, (3) a new version of sociobiology seen as a co-evolutionary process between genes, mind, and culture introduced by Wilson and Charles Lumsden in 1981, and (4) popular theorizing about human behavior. Scientific research in sociobiology often goes under the name 'behavioral ecology' or 'functional ethology'. Human sociobiology is often called 'evolutionary psychology' or 'Darwinian anthropology'.

SOCIOBIOLOGY'S ORIGINS AND ANTECEDENTS

At the time of Wilson's synthesis, the term 'sociobiology' was already in use in a division of the Animal Behavior Society led by John Paul Scott and others. Important earlier synthesizers in the field of animal ecology were Warder C. Allee and

Alfred E. Emerson (1949). The idea of comparative study of animal societies, again, goes back at least to the early nineteenth-century Swiss entomologist Pierre Huber. The most important antecedents of sociobiology are to be found in ethology on the one hand and the Neo-Darwinist or Modern Synthesis on the other.

Sociobiology and Ethology

Ethology, the study of animal behavior in natural environments, comes from the tradition of Konrad Lorenz and Niko Tinbergen. Ethology asks questions not only about function (adaptive value) but also phylogeny (evolutionary origin), ontogeny (development), and mechanism (proximate cause) of a behavior (Tinbergen, 1963). Sociobiology focuses on ultimate, genetic explanations. Many see sociobiology as 'functional ethology'. Wilson, in contrast, regards ethology as a subfield of the more encompassing discipline of sociobiology.

The Modern Synthesis

The Neo-Darwinian or Modern Synthesis, taking place between 1920 and 1950, aimed at integrating the various fields of biology within a common framework of mathematical population genetics, where evolution was expressed as the change of gene frequencies in a population. This international endeavor happened in two major steps: first the unification of Mendelian genetics with Darwinism, and second the extension of the synthesis to the rest of biology.

The problem of the seeming incompatibility between Darwinism and Mendelism was solved when it could be shown that continuous Darwinian variation could be expressed mathematically in particulate form, and that mutations were typically small and could serve as needed 'raw material' for evolution. The scientists making this first step, who worked out the translation of the various evolutionary forces into the new language of population genetics in the 1920s and 1930s, were Ronald A. Fisher and J. B. S. Haldane in the United Kingdom and Sewall Wright in the United States. Prominent architects of the second step of the synthesis were Theodosius Dobzhansky, Ernst Mayr, Julian Huxley, George Gaylord Simpson, Ledyard Stebbins, Bernard Rensch, and S. S. Chetverikov, who demonstrated that fields such as systematics, paleontology, and botany were compatible with Neo-Darwinist assumptions. Embryology (developmental biology), physiology, and morphology, however, still remained outside the synthesis. The extension of the synthesis to behavior came in the 1960s with Hamilton.

There were already conflicting views of the meaning of the Modern Synthesis among the original architects: there was a population genetic and a more holistic view. The former regarded the synthesis as completed following the work of the population geneticists. This opposition was later reflected in Dawkins's and Wilson's different conceptions of sociobiology. Harvard's Ernst Mayr (1963, chap. 10) set the tone for the 'holists' by dismissing what he called 'beanbag genetics' in favor of 'the unity of the genotype'. The resistance to talking about genes as if they were freely moving beans in a bag was to continue in the next generation with holistically oriented critics of sociobiology, such as Stephen J. Gould, Richard C. Lewontin, and Niles Eldredge.

INCLUSIVE FITNESS

Inclusive fitness is the key concept in sociobiology. The idea of inclusive fitness solves the long-standing mystery of how it is possible for altruistic behavior to evolve. It was this concept that Hamilton developed mathematically in his trailblazing 1964 paper, 'The genetical evolution of social behaviour'. Inclusive fitness has been typically used to formalize the reasoning about natural selection in kin groups (kin selection), but Hamilton himself intended that the concept have a broader application (see below).

Hamilton's Rule

Altruistic behavior can be favored by natural selection if the donor and recipient share a sufficient number of genes. Therefore, by helping relatives, an individual is actually helping the survival and propagation of its own genes (including the gene that promotes altruism). The coefficient of relatedness between relatives can be exactly calculated for various relationships. It is, for instance, ½ between siblings, ½ between parent and offspring, and ⅛ between first cousins.

'Hamilton's rule' (Hamilton, 1963) postulates that altruism can evolve if the reduction in fitness (number of offspring) of the donor is more than made up for by the increased fitness (number of offspring) of the recipient. It is simply expressed in the formula $k > 1/r$, where k is the cost–benefit ratio of the behavior, and r is the coefficient of relatedness between two individuals.

Technically, inclusive fitness of an organism is rather complicated to measure. It is typically

measured as the organism's own fitness plus the effect of its behavior on its relatives' reproduction minus its relatives' effect on its own reproduction, multiplied by its genetic relatedness to each related organism (Dawkins, 1982, chap. 10; Grafen, 1982)

The Case of the Social Insects

An illustration of the principle of inclusive fitness is the evolution of social behavior in *Hymenoptera* (bees, ants, and wasps). Here female workers are sterile and help raise their sister's (the queen's) offspring instead of their own. This makes evolutionary sense because of the unusual coefficients of relationship in a haplodiploid species like *Hymenoptera*. Here the female workers are related more closely to their sister, the queen, than to their own (potential) daughters.

The females are diploid, having two sets of chromosomes (they come from the queen's fertilized eggs). One of them will later become a queen. The males (drones) are haploid, having only one set of chromosomes (they come from the queen's unfertilized eggs). In diploid species for two sisters on average half of the mother's genes and half of the father's are the same. In haplodiploid *Hymenoptera* the haploid father (a drone) has only one set of genes to transmit. As a result, *Hymenoptera* sisters end up sharing three quarters of their genes. Meanwhile each sister's relationship to any potential daughter would be only the standard one-half. It therefore 'pays' for a female worker to forego reproduction and help raise her queen sister's offspring instead.

Expanding Inclusive Fitness

Hamilton intended 'inclusive fitness' to be a broader concept than 'kin selection', 'group selection' or even 'reciprocal altruism'. The basic requirement for inclusive fitness is that the benefits of altruism fall on individuals who are likely to be altruist rather than on random members of the population. These may be relatives (kin selection) but do not have to be. Altruists may live with altruists simply because they have never parted, or because they are able to recognize fellow altruists (what Dawkins has called 'the green beard effect'), or even because of some pleiotropic effect of the gene on habitat preference. This is why Hamilton later recommended a flexible use of terms, depending on the actual case: kin selection, group selection, kin-group selection, low migration, assortation, or other alternatives, as appropriate (Hamilton, 1975).

KIN SELECTION

Kin selection is often the way in which inclusive fitness has been understood in practice: as natural selection acting on the family or kin group instead of the individual. The case of *Hymenoptera* is a good example of kin selection. Another example is the assistance that group-living animals (flocks, herds) give one another.

Kin Selection and Group Selection

'Kin selection', Maynard Smith's (1964) term for the process of inclusive fitness maximization, is often contrasted with the idea of 'group selection', natural selection working on groups rather than individuals. Group selection was an earlier explanation for animal altruism, where individuals were seen as sacrificing themselves for 'the good of the species' or group. The Scottish zoologist Vero Wynne-Edwards (1962) was the first to explicate the possible mechanisms involved. Many evolutionists believe group selection is unlikely in nature. The problem is how altruism might initially arise in a group of individuals with regular, selfish genes, without individuals possessing mutant, altruistic genes getting outreproduced. Some evolutionists, however, consider group selection a real phenomenon in nature, among others E. O. Wilson. Elliott Sober, and David Sloan Wilson (1998) have reintroduced group selection as 'trait group' selection, invoking Price (1972) and Hamilton (1975).

Kin Selection in Practice

Early critics of kin selection wondered how a gene can calculate coefficients of relationships and make cost–benefit analyses (e.g. Sahlins, 1976). The answer lies in the existence of (subconscious) behavioral rules. Individuals act as if they are following such rules (Dawkins, 1979). A rule of this type for a bird might be: feed gaping mouths in your nest! This rule may not be specific enough to identify a cuckoo chick. Another perceived problem was kin recognition. Later research has shown that various kin recognition mechanisms do exist, such as smell (Fletcher and Michener, 1987).

THE GENE'S EYE VIEW

Inclusive fitness and other theories in sociobiology are often best understood by taking a 'gene's eye' view. This means regarding a gene causing an individual to behave in a particular way as a strategist interested in its own survival and

propagation. This view is associated with William Hamilton (1963), George Williams (1966), and Richard Dawkins (1976). According to Hamilton (1963), if there is a gene G that causes its carrier to act in an altruistic way, 'the ultimate criterion which determines whether G will spread is not whether the behaviour is to the benefit of the behaver, but whether it is to the benefit of the gene G'. Williams (1966) introduced the gene-selectionist approach as a methodological device, famously suggesting that in evolutionary explanation no higher level than necessary be invoked. Dawkins showed the fruitfulness of the new approach with imaginative explications of 'the selfish gene' causing its carrier to behave in ways that promote its own interest, including acting altruistically towards relatives carrying copies of itself.

The Gene's Eye View as a Tool for Thinking

The gene's eye view is a particularly useful tool when it comes to thinking about behavioral strategies (that is, behavioral patterns) adopted by interacting animals. An example is mating strategies. Depending on what type of body the gene is in (male or female), a gene will cause its vehicle to adopt a different strategy depending on the particular investment its carrier has in its offspring. A male typically invests little beyond the act of mating itself, while the female invests a lot. For a male it is often advantageous to mate with as many females as possible to maximize reproductive success. For a female, in contrast, it pays to be choosy, finding the best father for her offspring and making him help rear it. This argument was developed in detail by Robert Trivers (1972) in his paper on parental investment, drawing on Darwin's theory of sexual selection.

Replicators and Vehicles

Some early critics of the selfish gene idea believed that the gene was now claimed to be the true unit of selection (e.g. Gould, 1977; Wright, 1980) and protested loudly that other levels of selection existed. Dawkins, in response, made an important distinction between 'replicators' and 'vehicles' (Dawkins, 1978). A replicator needs to last long enough to produce additional replicators which retain their structure largely intact; it needs to possess the characteristics of 'longevity, fidelity, and fecundity'. A gene fulfills these criteria, making it a replicator. In contrast, individual organisms, groups, and so on are not replicators but vehicles.

Dawkins's argument was a logical one about the workings of evolution, not a factual statement about the vehicles that actually exist in nature. This misunderstanding was to persist throughout the so-called sociobiology debate. Dawkins emphasized (1982) that a gene could physically be any piece of DNA – as long as it was small or stable enough to have a 'frequency' that could be changed by natural selection. He also noted that genes are selected for their capacity to cooperate and that we call the result of this cooperation 'an organism'. However, the cooperation may not be perfect, as indicated by free rider effects and other phenomena.

RECIPROCAL ALTRUISM

Reciprocal altruism is altruism that occurs between unrelated individuals when there will be repayment (or at least promise of repayment) of the altruistic act in the future, and where the cost incurred by the altruist is smaller than the benefit of the beneficiary (Trivers, 1971). This principle of 'I'll scratch your back, you'll scratch mine' can be found in many bird and mammal species, where individuals clean one another of pests. Another example is blood-sharing in vampire bats. A successful vampire bat will share the blood it has swallowed during the night with less fortunate fellow bats. Later on, the recipients repay the favor. Further, animals (e.g. chimpanzees) may form alliances and coalitions with nonrelatives. An animal offering help can expect assistance itself in the future. Reciprocal altruism is an example of game theory (see below) where two (all) parties cooperate and both (all) can win.

Mutual assistance between members of different species is called 'mutualism' or 'symbiosis'. Here both parties derive an immediate benefit. An example is the lichen, a symbiosis between a fungus and green algae. Another is the aphids (greenfly) kept as 'milk cows' by ants. The aphids process plant juice, which the ants then 'milk' from them, while protecting the aphids and their offspring in return. Finally, there are the cleaner fish which get food by cleaning the mouths and gills of larger fish, which in turn refrain from eating them.

One perceived problem is that of individual animals recognizing each other later. This in fact often happens. Another problem is cheating: taking advantage of a benefit offered but not paying back. Trivers (1971) suggested that many human psychological characteristics may have evolved for us to be able to cheat, to detect cheaters, and, if cheating, to avoid being detected. The idea of cheater

detection has become one of the cornerstones of evolutionary psychology.

GAME THEORY

The basis for much sociobiological reasoning is in game theory, an approach originally developed in the 1940s and 1950s by the mathematician John von Neumann and the economist Oskar Morgenstern. Game theory captures much of social life, because here often the best course of action for one actor is dependent on what others do. Note that one individual's gain does not necessarily mean another's loss. In zero-sum games, one's gain is the other's loss. In non-zero sum games, both (all) parties can win.

The Prisoner's Dilemma

The prototypical two-person game is the famous Prisoner's Dilemma. This model illustrates how each individual's perceived self-interest typically wins out over a solution which would have been in the best interest of both. Two prisoners are arrested for a crime for which there is no good evidence. Each one is now invited to confess ('defect') against a greatly reduced prison sentence. If both keep quiet ('cooperate'), there is little evidence to keep them in prison. If both confess, both will get a severe penalty. The maximum penalty, however, will result from one keeping quiet while the other one confesses. Assessing the options, each typically reasons that he is better off defecting – and so both end up worse off than if they had cooperated. This outcome is typical in laboratory-type Prisoner's Dilemma games played only once with participants who do not know each other. In so-called iterated Prisoner's Dilemma games individuals meet again and again (which is often the case in social life). Under this condition reciprocal altruism may develop.

Evolutionarily Stable Strategy (ESS)

An important concept in game theory applied to animal social life is that of an evolutionarily stable strategy (ESS). This is a pattern of behavior (strategy) which when it is the dominant one in a population will prevail against any alternative 'mutant' strategy. The idea of ESS was developed by John Maynard Smith and George Price in their work on animal signals (1973), and later by Maynard Smith (e.g. 1982). An early version was Hamilton's 'unbeatable strategy' (1967). ESS was initially developed to explain the regulation of aggressive behavior; it was sociobiology's answer to the ethologists' question of why animals which are in principle able to kill one another do not actually fight to the death. In natural populations ESS often takes the form of stable polymorphisms (an evolutionarily stable ratio of 'hawk' to 'dove' genes, say, where 'hawk' and 'dove' represent different patterns of behavior).

Winning Strategies

Using iterated Prisoner's Dilemma games Hamilton together with political scientist Robert Axelrod (1981) explored the possible evolution of cooperation between unrelated individuals. An early winning strategy was Tit for Tat. Here an individual starts off by cooperating, is 'nice' as long as the other cooperates but responds to defections with immediate retaliation. Examples of later successful strategies are Generous Tit for Tat, which occasionally reponds to defection with cooperation, and Pavlov, a 'nasty' strategy whose 'win–stay, lose–shift' approach makes it defect even against a cooperator. Note that 'winning strategy' here means that it won in computer tournaments against other strategies: Tit for Tat, for instance, won because a number of 'nasty' strategies all eliminated one another.

Applications of Game Theory

Game theory can be employed to explain a variety of phenomena, for instance why animal formations are often small and circular. Every individual in a 'selfish herd' wants to protect itself against a predator by having other animals around itself (Hamilton, 1971). Game theory has also been used to explain why sex ratios are typically 50:50 and not otherwise (Hamilton, 1967). Trivers (1974) used game theory to develop his theory of parent–offspring conflict. A mother may wish to wean her offspring in order to continue reproducing, while her present offspring wants to enjoy continued feeding. Moreover, game theory can be used to model intragenomic conflicts, for instance genomic imprinting, in which a gene is expressed differently depending on whether it derives from the father or the mother. Still, the existence of seemingly stable genomes indicates that evolutionarily stable strategies may also exist at the genomic level (Haig, 1997).

THE SOCIOBIOLOGY DEBATE

The sociobiology controversy started in 1975 around E. O. Wilson and his book *Sociobiology*: *The New*

Synthesis, whose last chapter applied sociobiology to humans. Among the most active critics were Wilson's own Harvard colleagues in evolutionary biology, Stephen J. Gould and Richard Lewontin, members of a Boston area group who accused sociobiology of supporting a conservative political agenda and legitimizing racism and sexism. After the initial upheaval, including an incident in 1978 where a pitcher of ice-water was poured into Wilson's neck at the American Association for the Advancement of Science conference in Washington, DC, followed a more scientifically oriented critique of the adaptationist program (Gould and Lewontin, 1979). Human sociobiology meanwhile continued being attacked for biological determinism and reductionism (e.g. Gould, 1981; Lewontin *et al*., 1984).

While Dawkins (e.g. 1982) tried to clear up apparent misunderstandings, Wilson, with the help of physicist Charles Lumsden (1981), responded with a new model of sociobiology, now based on gene, mind, and culture co-evolution. Gould continued working on alternatives to adaptation, such as 'punctuated equilibria' (Eldredge and Gould, 1972; Gould, 1980), 'exaptation' (Gould and Vrba, 1982), and contingency (Gould, 1989).

The controversy soon became one chiefly between Dawkins and Gould, who for over two decades in popular books continued attacking each other while promoting their own views. (Wilson had moved on to biodiversity.) Dawkins showed how adaptation can give rise to complex design while Gould argued for chance and contingency in evolution. While Dawkins explained the mechanisms of evolution, Gould treated these as statements about the real world.

Scientifically, the sociobiology debate can be seen as the next round in the ongoing conflict about the Modern Synthesis and the true meaning of Neo-Darwinism. It was a continuation of the protest against reducing everything to 'beanbag genetics' and an attempt to reopen the debate about the relationship between micro- and macro-evolution (e.g. by bringing in punctuated equilibria), and about phylogeny and ontogeny (e.g. by emphasizing the role of developmental constraints and interaction effects).

An indication of this continuing conflict is Niles Eldredge's (1995) and Gould's (1997a,b) division of the Darwinians into 'ultra-Darwinians' (who try to explain everything in terms of relative gene frequencies) and 'naturalists' or 'pluralists' (who accept the existence of multiple forces in evolution and multilevel selection). While Eldredge admits that both sides 'agree on the rudiments of evolutionary change: adaptative modification through natural selection', Gould is less conciliatory. His magnum opus (2002) is an anti-adaptationist manifesto against those who claim the triumph of sociobiology (e.g. Alcock, 2001).

But the sociobiology controversy has also had a moral/political dimension. Learning from the sociobiology controversy, the new field of evolutionary psychology has tried to avoid political traps. It self-consciously deals with the evolved architecture of the human mind rather than with 'genes for behavior' and emphasizes human universals rather than individual differences. Still, because the field employs adaptive explanation, it has inherited the adaptationist critique of sociobiology. The critics seem convinced that claims about a biologically constrained human nature will inevitably be used as justification for social inequality or immoral behavior, while an anti-adaptationist stance is connected to progressive politics and human liberation.

Although often presented that way, the sociobiology debate was not a feud between the political left and right. It was rather a conflict between a new brand of academic activists who wanted to weed out seemingly dangerous ideas from science before these could do social harm and traditional scientists who believed in the benefit of knowledge and trusted the democratic process (Segerstrale, 2000).

References

Alcock J (2001) *The Triumph of Sociobiology*. Oxford and New York: Oxford University Press.

Axelrod R and Hamilton WD (1981) The evolution of cooperation. *Science* **211**: 1390–1396.

Dawkins R (1976) *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins R (1978) Replicator selection and the extended phenotype. *Zeitschrift für Tierpsychologie* **47**: 61–76.

Dawkins R (1979) Twelve misunderstandings of kin selection. *Zeitschrift für Tierpsychologie* **51**: 184–200.

Dawkins R (1982) *The Extended Phenotype. The Gene as Unit of Selection*. Oxford and San Francisco: Freeman.

Dawkins R (1987) *The Blind Watchmaker*. New York: W. W. Norton & Co.

Eldredge N (1995) *Reinventing Darwin. The Great Debate at the High Table*. New York: John Wiley & Sons.

Eldredge N and Gould SJ (1972) Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf TJM (ed.) *Models in Paleobiology*. San Francisco: Freeman Cooper.

Fletcher DJC and Michener CD (eds) (1987) *Kin Recognition in Animals*. New York: John Wiley & Sons.

Gould SJ (1977) Caring groups and selfish genes. *Natural History* **86**(12): 20–24.

Gould SJ (1980) Is a new and general theory of evolution emerging? *Paleobiology* **6**: 119–130.
Gould SJ (1989) *Wonderful Life*. New York: W. W. Norton & Co.
Gould SJ (1996) *The Mismeasure of Man*, 2nd edn. New York: W. W. Norton & Co.
Gould SJ (1997a) Darwinian fundamentalism. *The New York Review of Books*, 12 June: 34–37.
Gould SJ (1997b) Evolution: the pleasures of pluralism. *The New York Review of Books*, 26 June: 47–52.
Gould SJ (2002) *The Structure of Evolutionary Theory*. Cambridge, MA: Alfred Knopf.
Gould SJ and Lewontin RD (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B* **205**: 581–598.
Gould SJ and Vrba E (1982) Exaptation: a missing term in the science of form. *Paleobiology* **8**: 4–15.
Grafen A (1982) How not to measure inclusive fitness. *Nature* **298**: 425–426.
Haig D (1997) The social gene. In: Krebs JR and Davies NB (eds) *Behavioural Ecology*, 4th edn, pp. 284–304. Oxford: Blackwell Scientific Publications.
Hamilton WD (1963) The evolution of altruistic behavior. *The American Naturalist* **97**: 354–356.
Hamilton WD (1964) The genetical theory of social behavior. I and II. *Journal of Theoretical Biology* **7**: 1–16; 17–32.
Hamilton WD (1967) Extraordinary sex ratios. *Science* **156**: 477–488.
Hamilton WD (1972) Geometry for the selfish herd. *Journal of Theoretical Biology* **31**: 295–311.
Hamilton WD (1975) Innate social aptitudes of man: an approach from evolutionary genetics. In: Fox R (ed.) *Biosocial Anthropology*, pp. 133–157. New York: John Wiley & Sons.
Lewontin RC, Rose S and Kamin L (1984) *Not in Our Genes*. New York: Pantheon Books.
Lumsden CL and Wilson EO (1981) *Genes, Mind and Culture: The Coevolutionary Process*. Cambridge, MA and London: Harvard University Press.
Maynard Smith J (1964) Group selection and kin selection. *Nature* **201**: 1145–1147.
Maynard Smith J (1982) *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
Maynard Smith J and Price G (1973) The logic of animal conflict. *Nature* **246**: 15–18.
Mayr E (1963) *Animal Species and Evolution*. Cambridge, MA: Harvard University Press.
Price GR (1972) Extension of covariance selection mathematics. *Annals of Human Genetics* **35**: 485–490.
Sahlins M (1976) *The Use and Abuse of Biology*. Ann Arbor: University of Michigan Press.
Segerstrale U (2002) *Neo-Darwinism. Encyclopedia of Evolution*. Oxford and New York: Oxford University Press.
Sober E and Wilson DS (1998) *Unto Others*. Cambridge, MA: Harvard University Press.
Tinbergen N (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* **20**: 410–433.
Trivers RL (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57.
Trivers RL (1972) Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man*, pp. 136–179. Chicago: Aldine.
Trivers RL (1974) Parent-offspring conflict. *American Zoologist* **14**: 249–264.
Williams GC (1966) *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.
Wilson EO (1975) *Sociobiology: The New Synthesis*. Cambridge, MA: Harvard University Press.
Wright S (1980) Genic and organismic selection. *Evolution* **34**(5): 825–843.
Wynne-Edwards VC (1962) *Animal Dispersion in Relation to Social Behavior*. Edinburgh: Oliver & Boyd.

Further Reading

Brandon R and Burian R (eds) (1984) *Genes, Organisms, Populations*, pp. 239–248. Cambridge, MA: MIT Press.
Caplan AL (1978) *The Sociobiology Debate*. New York: Harper & Row.
Dennett D (1995) *Darwin's Dangerous Idea*. New York: Simon and Schuster.
Hamilton WD (1996) *Narrow Roads of Gene Land. The Collected Papers of W. D. Hamilton. I. Evolution of Social Behaviour*. Oxford and New York: W. H. Freedman.
Hamilton WD (2001) *Narrow Roads of Gene Land. The Collected Papers of W. D. Hamilton. II. Evolution of Sex*. Oxford and New York: Oxford University Press.
King's College Sociobiology Group (1982) *Current Problems in Sociobiology*. Cambridge: Cambridge University Press.
Kitcher P (1985) *Vaulting Ambition*. Cambridge, MA: MIT Press.
Krebs JR and Davies NB (1997) *Behavioural Ecology*, 4th edn. Oxford: Blackwell Scientific Publications.
Ruse M (1979, 1985) *Sociobiology: Sense or Nonsense*? 1st and 2nd edns. Dordrecht: Reidel.
Segerstrale U (2001) *Neo-Darwinism*. Oxford Encyclopedia of Evolution.

Development

Introductory article

Nora S Newcombe, Temple University, Philadelphia, Pennsylvania, USA

CONTENTS

Development refers to the acquisition of mature knowledge states and cognitive capabilities. Cognitive scientists seek to characterize initial knowledge states and cognitive capabilities, and the processes involved in their transformation into adult competence.

BACKGROUND

Babies do not appear to think or act like adults. Thus, there seems to be a self-evident problem of development. How does the helpless infant change into the competent adult? Answers to this question have both practical and theoretical implications. On the practical side, knowledge about when and how children change can help parents to be aware of their children's capacities and sensitive to their needs. Understanding cognitive development can also help educators to work with children to learn optimally, and can help clinicians asked to deal with a variety of challenges to learning. On the theoretical side, the study of development is one of the crucial aspects of cognitive science. No theory of knowledge or skill would be complete without an approach to how the knowledge or skill is acquired. Indeed, thinking about acquisition can place important constraints on the theoretical enterprise and strengthen or rule out specific approaches.

Consideration of the nature of cognitive development predates psychology. Philosophers pondering the matter generally argued for one of two answers tothe question of how the baby becomes the adult. The first kind of answer was to suggest that infants are molded into adults by the influence of the world in which they live. There are various versions of this kind of theory, generally known as empiricism and most famously associated with the name of the English philosopher John Locke. Empiricists postulate a baby born with very little knowledge or capability – Locke spoke of a 'tabula rasa', or blank slate – for whom responses are shaped by the associations of certain sensations in time and space with other sensations, and by reward and punishment.

The second kind of answer to the problem of development is to argue that infants know more than they may seem to know at first sight. Although they appear helpless, they may be endowed with capabilities and categories of thought that allow them to apprehend a world governed by immutable laws of space, time, and causality. Such a theory, emphasizing the importance of the human mind in creating the reality of the world around us, is most famously associated with the German philosopher Immanuel Kant. Kant's theory of knowledge has not one but two modern descendants. It can, in fact, be seen as leading to two quite different approaches to development within cognitive science: constructivism, as originally proposed by Jean Piaget, and nativism, as originally proposed for the case of language development by Noam Chomsky.

Philosophers discussed the nature of infants', children's, and adults' knowledge without actually observing any infants, children, or adults in the process of thinking. When the science of psychology began in the second half of the nineteenth century, it gave rise almost immediately to the empirical study of babies and young children and to theorizing about the nature of development based on observation. Some investigators, working in the empiricist tradition, emphasized the role of the environment in development. For example, John S. Watson made studies of conditioning that he argued showed the origins of certain thoughts and fears. B. F. Skinner analyzed language development as the product of contingencies in the world. Other investigators, working in a biological tradition that can be seen as an early version of nativism, portrayed development as the maturational unfolding of a preprogrammed biological being. For instance, Arnold Gesell and G. Stanley Hall created tables and charts showing just what children at typical ages could be expected to do and think, implying that functioning would be in

a constant relation to chronological age. Some thinkers emphasized the Kantian theme of an active mind that structured understanding of the world. James Mark Baldwin and Heinz Werner are early examples of this type of constructivist thinking.

During this early period, it is interesting to note, writing about development already had a property that is still true today: it was seen to have relevance for thinking about applied problems of child rearing and education. Hence, for example, Edward Thorndike was simultaneously a learning theorist and a founder of the field of educational psychology.

CLASSIC APPROACHES

Prior to the 'cognitive revolution' of the 1960s, comparatively little was known about cognitive development, in large part because comparatively little was known about the nature of mature cognitive functioning. Psychological science, especially in the United States, was dominated by the view that there were universal laws of learning that applied equally to all animal species including humans. Therefore the problem of development tended to be conceptualized as the simple accumulation of learning achieved by means of the operation of the universal laws of learning (B. F. Skinner's cumulative record), or, slightly more developmentally, as the problem of learning to learn in an adult way. An example of research in the latter tradition would be experiments designed to examine exactly what children learned when they learned that a response to a particular stimulus was rewarded. For instance, children might learn that a reward would follow pressing a square of a particular absolute size that was also the middle-sized square of three. There was interest in determining whether there was a developmental change in children seeing the absolute size as controlling the reward as opposed to seeing the 'middle' relation as controlling the reward. Work by Jean Piaget and Lev Vygotsky was published in Europe in the first half of the twentieth century, but their writing was not widely appreciated when it first appeared.

The cognitive revolution led to a radical change in this situation. There were several events central to this revolution, many of them with developmental aspects. Chomsky's writings on language in the 1950s and 1960s led to a rethinking of the psychology of language by figures such as George Miller, and also to work on children's acquisition of language led by the pioneering investigations of Roger Brown. This research emphasized children as active constructors of a child grammar, in the Kantian tradition. Simultaneously, building on his earlier work on the constructive nature of perception and on the nature of adult concepts and thinking, Jerome Bruner started a research program on children's cognition that emphasized children as theorizers. Bruner's work, along with John Flavell's introductory writing on Jean Piaget, led to widespread interest in the United States in Piagetian research. By the mid-1970s, research on cognitive development was dominated by a focus on Piaget's theory and hypotheses. (*See* **Piagetian Theory, Development of Conceptual Structure**)

The cognitive revolution put an end to simple empiricism as an approach to development. Almost from the start, however, there were controversies between developmental theorists who emphasized the Kantian idea of the developing mind as an active constructor of knowledge and those who focused on the equally Kantian idea of innate capabilities. Piaget was the central example of the constructivist approach. But his theory quickly came under fire for several reasons. One criticism was that his thinking seriously underestimated the initial capabilities of infants. Piaget postulated that infants are born with only simple sensorimotor capabilities – they can see, hear, feel, taste, and smell the world, and they are also equipped with stimulus-response reflexes and other, less patterned motor exploration patterns. A large amount of research during the 1980s and 1990s showed, however, that infants and young children had far more substantial abilities than Piaget had seen. A second criticism was that various capabilities did not cohere in the way that Piaget's stage theory had envisioned. Piaget postulated that development consists of qualitative transformation through four distinct stages of thinking ability: sensorimotor, preoperational, concrete operational, and formal operational. Research on the abilities said to characterize these four stages showed, however, that children frequently achieved competence in some of the relevant tasks while lagging behind on others said to be characteristic of the stage, often for protracted periods of time. A third problem was that Piaget talked in general terms about the process of developmental change as being the operation of equilibration, a process of achieving cognitive equilibrium either by seeing the environment in terms of existing cognitive structures (assimilation), or, if needs be, changing cognitive structures in order to make sense of observations of the environment (accommodation). Many investigators argued that a more

detailed specification of the mechanisms of change was needed.

The waning of support for Piagetian theory led to an increased interest in nativism as a solution to the problem of development. Nativism had long been espoused as a theory of language development, because Noam Chomsky had argued strongly that there must be an innate language acquisition device. Otherwise, he said, it would be impossible for children to learn language from the impoverished and error-laden speech sample that they would obtain from listening to adults. Beginning in the late 1970s, nativism also became popular when considering domains of cognitive development in addition to language, including understanding of the Kantian aspects of the world, such as causality, space, time, and number. Nativism was also applied to other aspects of cognitive development, such as the acquisition of categories, or the possibility of imitation of the actions of adults. In addition, many investigators who were not strict nativists became very intrigued by the demonstration of the surprising early competence of infants, such as their memory abilities.

A different response than nativism to the waning of support for Piagetian theory was to champion approaches that more strongly emphasized the role of the environment in development. Vygotsky's writings underwent a revival because their focus on cultural transmission through language and teaching seemed to emphasize the role of the environment in development, without representing a retreat to pure empiricism. Information processing approaches to development also stressed the role of the environment in development but conceptualized it somewhat differently than Vygotskyan theory, as a provider of informational feedback. Information processing theory melded a focus on environmental feedback with a developing part of the emerging field of cognitive science, namely computational modeling. Many information processing theorists also incorporated biological thinking in their models by using parameters in the models that could be conceptualized as changing maturationally, notably short-term memory and processing capacity. (*See* **Categorization, Development of**; **Memory, Development of**; **Culture and Cognitive Development**)

RECENT INFLUENCES

As the field of cognitive science developed during the 1980s and 1990s, in conjunction with the development of cognitive neuroscience during roughly the same period, the study of cognitive development changed concomitantly. The use of computational models expanded. In addition, the style of modeling changed from the predominant use of production systems, written line by line by modelers with certain assumptions about relations between input and processing, to an increasing use of parallel distributed processing (PDP) models, set up with fewer content-related assumptions. PDP modelers postulated a cognitive architecture and made assumptions about the nature of environmental input and feedback, but they then needed to run their systems as experiments to see what they actually did. Modeling in this tradition has led, in some hands, to a view of development as the self-modification of systems that can be identified as constructivist. However, in other models, there are *tabula rasa* assumptions and decisions about the nature of environmental feedback that bring the models close to old-fashioned empiricism. (*See* **Cognitive Development, Computational Models of**)

A second change in thinking about cognitive development in recent times has been the increasing attention given to the role of the neural substrate in thinking. The advent of developmental cognitive neuroscience has added new techniques and dependent variables to the armamentarium of investigators, as well as renewing interest in developmental disorders. In addition, neuroscience has been the source of certain broad insights about the probable nature of development. One important example is the attention received by Peter Huttenlocher's research on synaptogenesis in different areas of human cortex. Huttenlocher has shown that increases in synaptic connections occur during the early part of life, probably in a maturational way. When connections reach a critical level, one often sees the advent of new forms of functioning. From a peak far above the levels typical for adults, synaptic connections seem to be eliminated, or pruned, in a way that is dependent on environmental input. (*See* **Neural Development**; **Neuropsychological Development**; **Developmental Disorders of Language**)

Newer approaches to cognitive development are not only the products of modeling and the use of neuroscience – they have also been driven in traditional ways by the rethinking of existing data and the acquisition of new data. There are three theories that attracted much attention in the 1990s. Dynamic systems theory was developed in other areas of investigation, such as characterization of motor behavior and it has been productively applied in thinking about motor development. However, Esther Thelen and Linda Smith have argued that

it can also be very useful in thinking about development in many other domains as well. In brief, this approach suggests that behavior is complexly determined by contextual variables, so that thinking about development as the acquisition of specific static competences is liable to be misleading and unproductive. Variation-and-selection theory is an updated version of information processing theory proposed by Robert Siegler. Siegler argues that various strategies for dealing with cognitive problems co-exist, with development consisting of the relative predominance of these theories at different points in time. Domain-specific interactionism is a constructivist approach to development that recognizes the fact that the initial starting points for cognitive development are often quite strong and specific, and that development proceeds in a way that is not strongly stagelike. Many key domains have been investigated from this point of view. (*See* **Motor Development**; **Lexical Development**; **Causal Perception, Development of**; **Intermodal Perception, Development of**; **Naive Theories, Development of**; **Object Concept, Development of**; **Space Perception, Development of**; **Speech Perception, Development of**; **Object Perception, Development of**)

SUMMARY

The central problem addressed by research on cognitive development is the issue posed philosophically as the opposition between nativism and empiricism. A wide variety of theories and research has addressed the question, and we have moved beyond the most extreme versions of the debate to consider more integrated and balanced solutions. In the course of this research, we have also learned a great deal about the nature of development in various domains and at various ages. Much of this knowledge is now mature enough to be useful in a variety of applied areas, such as designing intervention and remediation for children with developmental disabilities, or improving the effectiveness of education for normally developing children. In addition, from a theoretical point of view, the field of cognitive development is a vital part of cognitive science. Because no cognitive model can be considered complete if it cannot be acquired by human children, given sensible assumptions about the nature of their learning, developmental thinking offers both constraints on theories and opportunities for theory testing. (*See* **Early Experience and Cognitive Organization**; **Language Development, Critical Periods in**)

Further Reading

Chapman M (1988) *Constructive Evolution: Origins and Development of Piaget's Thought*. New York: Cambridge University Press.

Elman J, Bates E, Johnson M, Karmiloff-Smith A, Parisi D and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: The MIT Press.

Hirsh-Pasek K and Golinkoff RM (1996) *The Origins of Grammar: Evidence from Early Language Comprehension*. Cambridge, MA: The MIT Press.

Karmiloff-Smith A (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: The MIT Press.

Mix KS, Huttenlocher J and Levine SC (2002) *Quantitative Development in Infancy and Early Childhood*. New York: Oxford University Press.

Newcombe NS and Huttenlocher J (2000) *Making Space: The Development of Spatial Representation and Reasoning*. Cambridge, MA: The MIT Press.

Rakison DH and Poulin-Dubois D (2001) Developmental origin of the animate-inanimate distinction. *Psychological Bulletin* **127**: 209–228.

Rogoff B (1990) *Apprenticeship in Thinking: Cognitive Development in Social Context*. New York: Oxford University Press.

Siegler RS (1996) *Emerging Minds: The Process of Change in Children's Thinking*. New York: Oxford University Press.

Thelen E and Smith L (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: The MIT Press.

Agreement

Intermediate article

Marcel den Dikken, City University of New York, New York, USA

CONTENTS

Agreement in linguistics is a relationship of matching or systematic covariation of the features of constituents of a syntactic construct. All major syntactic categories and many minor categories can entertain agreement relationships of a variety of different kinds, typically involving subject–verb or modifier–head configurations.

AGREEMENT: GENERAL PROPERTIES

Agreement (or concord) is a relationship of matching or systematic covariation of the features of constituents of a syntactic construct. The constituents are said to agree in features: ϕ-features (where 'ϕ' is a cover for person, number, gender); case (e.g., Latin *ill<u>arum</u> du<u>arum</u> bon<u>arum</u> femin<u>arum</u>* – 'of those two good women', with genitive feminine plural marked throughout); noun class (e.g., Bantu); or some other properties (e.g., categorial features, as in Chamorro complementizer agreement; or tense).

All major syntactic categories can entertain agreement relationships with other constituents. In many languages, finite verbs agree with their subjects ('subject agreement'), and there are also languages in which finite verbs can agree with one or more of their objects ('object agreement'), or with *wh*-extracted constituents ('*wh*-agreement' as in Bantu, Palauan, Chamorro); nonfinite verbs can also show agreement with their dependents (past participle agreement in Romance languages; inflected infinitives in Portuguese, Hungarian). Predicative adjectives can agree with their subjects, attributive adjectives with the head noun. Predicate nominals often agree with their subjects as well; and possessed nouns may show agreement with their possessors. Finally, prepositions can agree with their objects (e.g., Celtic, Hungarian, Abkhaz).

Minor (or closed-class) syntactic categories may also show agreement. Determiners (articles, demonstratives) typically agree with the head noun of a complex noun phrase. Complementizers (subordinating conjunctions) may bear agreement inflection for either the subject of the clause they introduce (as in varieties of the Germanic languages) or a *wh*-phrase extracted out of that clause (as in Irish; cf. also relative clause constructions featuring inflected relative complementizers).

HEAD AGREEMENT AND DEPENDENT AGREEMENT

An important typological distinction between languages is that between head-marking and dependent-marking languages (Nichols, 1986). Head-marking languages are rich agreement languages; dependent-marking ones express relationships between heads and their dependents in other ways (typically with case-marking on the dependent). Abkhaz, a typical head-marking language, shows agreement inflection on the verb for several of its arguments, and possessive inflection on nouns and prepositions. Languages may combine a rich head-marking agreement system with a system of morphological case-marking on dependents (e.g., Hungarian); languages showing neither type of marking also exist (e.g., Chinese).

Agreement can be classified along head/dependent lines as well. We define head agreement as involving agreement marking realized on the head, not on the dependent. Interpretively, in a sentence with an overt subject and an inflected verb, the expression of ϕ-features on the subject is meaningful (the difference between singular and plural noun phrases is semantically significant) while that on the finite verb is not (it is merely the reflection of the agreement relationship with the subject). Thus, we may call the ϕ-features on the subject interpretable and those on the verb uninterpretable (cf. Chomsky (1995) for a theory in which this distinction plays a major role). In languages in which nominal arguments may remain unexpressed in the presence of ϕ-feature inflection

on the head (so-called pro-drop languages), the inflection on the head may itself be taken to be meaningful – the inflection would be the argument of the head ('pronominal argument languages'; Jelinek, 1984). The pronominal agreement approach bears a strong relationship to analyses of clitic constructions; indeed, the dividing line between clitics and agreement is often difficult to draw and remains a contentious issue.

Agreement may also be marked on the dependent. The quintessential example of dependent agreement is the inflection of attributive adjectival modifiers of noun phrases, where the head noun determines the form of the modifying adjective. Another possible case of dependent agreement is that between anaphoric expressions and their antecedents (*They like themselves*); agreement here is not always strictly grammatical agreement, though (*Everybody thinks they are smart*). Both these dependent agreement cases can be reanalyzed as involving head agreement (Abney, 1986; Kayne, 2001); whether the theory needs to recognize two separate agreement types is not immediately obvious, therefore.

In addition to the above agreement patterns, in which one member is the dependent of the other member of the pair, we find agreement relationships between items where neither is a direct dependent of the other. We will encounter some of these in the next section; they may be assimilable to the head/dependent pattern in ways sketched in the last section.

INTRICACIES OF AGREEMENT

Agreement in ϕ-features exhibits a complicated distribution when it comes to the subset of features picked out. Number shows up cross-linguistically in all types of head agreement; person is frequently marked in finite verb agreement but not all languages having past participle or adjective agreement express person there (cf. Romance); gender, on the other hand, is much less commonly marked in finite verb agreement than in adjective agreement. Animacy and definiteness are two other major agreement features.

The question of whether we find agreement or not may be influenced by complicated syntactic factors, especially in the context of subject agreement and extraction. The position of a noun phrase *vis-à-vis* the agreeing verb may affect agreement possibilities: thus, in Arabic, prenominal subjects agree in all relevant features while postnominal subjects trigger person agreement only. In Berber and varieties of Celtic, *wh*-extracted subjects fail to agree with the verb except if the clause is negated, in which case subject agreement does show up (Ouhalla, 1993). And regular subject agreement can be suspended in sentences in which the finite verb agrees with a *wh*-constituent – as in Bantu (Kinyalolo, 1991) and varieties of American English (Kimball and Aissen, 1971: *the people who John think are in the garden*) – or with a subconstituent of the subject ('agreement attraction': *The identity of the participants are to remain a secret*).

Such 'overruling' tends to be skewed with respect to number: plurals can supplant regular singular agreement, but the opposite is much less common (cf. the frequently occurring *the key to the doors are missing* versus the much rarer *the keys to the door is missing*). This points towards number involving a privative opposition, with plural as the marked member. 'Overruling' of regular subject agreement also tends not to occur when the subject is pronominal. There is a robust cross-linguistic tendency for agreement between a head and a pronoun to be richer and more persistent than that between a head and a full noun phrase. Thus, in Welsh VSO sentences the verb does not show number agreement with full-nominal subjects, but subject pronouns must agree for number; *mutatis mutandis*, the same is found in Hungarian possessed noun phrases.

Within the realm of pronouns, first and second person pronouns often behave differently when it comes to agreement-related phenomena than do third person noun phrases (whether full-nominal or pronominal). Hungarian definiteness agreement between finite verbs and their objects yields straightforward results with third person objects, but first and second person object pronouns surprisingly trigger indefinite agreement on the verb. Splits between first and second person on the one hand, and third person on the other, characterize many so-called split ergative languages as well. The Mayan language Mocho, for instance, exhibits such a split. Other Mayan languages show split ergativity conditioned by tense, aspect, or clause type (main versus subordinate; Dixon, 1994, p. 201 and sections 4.3–4.4). Many morphologically ergative languages (e.g., Warlpiri) exhibit a nominative–accusative verb agreement pattern in tandem with an ergative–absolutive case system, showing that case and agreement patterns need not coincide.

Hybrid agreement patterns manifest themselves in a variety of forms. In French *Vous êtes loyal* [you-2PL are-2PL loyal-M.SG], the second person plural pronoun *vous* is used as a polite form with a

singular referent, in which case it triggers second person plural agreement on the verb but singular agreement on the predicate; similarly, in Spanish *Su Majestad suprema está contento* [your supreme-F.SG Majesty-F.SG is happy-M.SG], *Majestad* triggers femine agreement on the attributive adjective regardless of the referent but has predicate agreement determined by the gender of the referent (here, masculine).

These kinds of hybrid agreement may also be classified as semantic agreement, with the ϕ-feature composition of the head being determined by the referent of the dependent rather than by the morphosyntactic features of the dependent *per se*. Semantic agreement seems to be confined to head agreement.

AGREEMENT IN CURRENT GRAMMATICAL THEORIES

Semantic agreement is the cornerstone of Dowty and Jacobson's (1988) theory of agreement, in which agreement relationships are given a semantic explanation. In Reed's (1991) functional approach to verb number in English, meaning is also the epicenter: for Reed, what is generally referred to as an agreement relationship between the subject and the finite verb is not a case of agreement at all; instead, the number specification of each is chosen independently of that of the other, with each contributing independently to the message the speaker seeks to convey. Naturally, the emphasis in this work is on lack of agreement.

A semantic theory of agreement faces difficulties wherever semantic factors fail to have the final say. Thus, while *the dog* can be pronominalized with either *it* or *he*, in a sentence like *That dog is so ferocious, it/he even tried to bite itself/himself*, the assignment of gender to the subject pronoun and the object anaphor has to be uniform: the combinations *it* + *himself* and *he* + *itself* are impossible. This uniformity is not semantically determined; instead, it cues the need for a morphosyntactic theory of agreement.

Pollard and Sag (1994), who contributed this argument against semantic approaches, offer a theory of agreement built on the feature-based formalism of head-driven phrase structure grammar but allowing agreement access to semantic and pragmatic information as well. In this theory, ϕ-features are taken to be part of the internal structure of referential indices, the latter being the key notion of their theory. Indices in this theory make both a semantic and a syntactic contribution; they are vital in the analysis of agreement phenomena and referential dependencies.

A unified approach to agreement and referential dependencies in terms of indices is found also in the early principles-and-parameters literature (Chomsky, 1981, 1986; Borer 1986). In more recent principles-and-parameters work (Chomsky 1995), however, indices are assumed not to play a theoretical role. Instead, agreement is represented in terms of a local structural configuration ('specifier–head agreement') or is established under a (potentially long-distance) Agree relationship (Chomsky, 2001). Of these two options, the former represents theories in which agreement is a combination of feature matching and a specific structural configuration under which such matching is 'checked' – the spec–head structure (see the next section for more discussion), or Chomsky's (1995) 'checking domain' (see Chung (1998) for a different approach, cast in terms of the Associate relationship). The more recent Agree approach reduces agreement strictly to feature matching, with specific structural configurations resulting not from the need to establish agreement but from other, unrelated requirements.

Agreement as feature checking is essentially a symmetrical relationship; the Agree approach, by contrast, conceives of agreement as an asymmetrical relationship between a 'probe' and a 'goal'. Early generative approaches to agreement were asymmetrical as well, with transformations copying ϕ-feature specifications from one member of the agreeing pair to the other. Asymmetrical agreement also characterizes Keenan's (1979) approach to agreement in terms of the function–argument relationship. In the framework of Generalized Phrase Structure Grammar (Gazdar *et al*., 1985), agreement relations are likewise encoded asymmetrically, but in the more recent Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994) agreement is treated in symmetrical terms.

AGREEMENT AS EVIDENCE FOR STRUCTURE

Agreement relationships are severely restricted: though there may be a variety of noun phrases present in the domain of a head, this head establishes agreement with only a narrow subset of those noun phrases. Thus, in *John ate Bill's cereal this morning*, there are four noun phrases surrounding the verb, but in no language will the verb agree with all four at the same time; at most, the verb agrees with the subject and the object.

While the possessor can agree with the verb under special circumstances ('possessor ascension to direct object'), bare NP adverbs never agree.

There are structural reasons why agreement relationships are so restricted. Agreement can be established in specific structural configurations only, of which the subject or specifier relationship seems to be the canonical case. If all agreement relationships are taken to involve such a structure, the occurrence of agreement between any two constituents is evidence for a structure in which these constituents are in a specifier–head relationship. In the principles-and-parameters theory of generative grammar, this hypothesis has led to the introduction of agreement phrases (AgrPs) for objects of verbs and prepositions and in the complementizers system. Agreement thus plays a pivotal role in the establishment of syntactic structures in some theories.

Agreement between complementizer and *wh*-extracted constituents and between possessors and possessed nouns is readily recast in these structural terms. When the possessed object itself incorporates into the verb, the possessor in a sense becomes a derived specifier of the verb; similarly, when the finite verb is incorporated into the complementizer position, the subject becomes a derived specifier of the complementizers (Zwart, 1997). In this way 'possessor ascension' and complementizer–subject agreement may be captured. Kayne (1995) shows that a similar treatment is available for cases of agreement in which the finite verb of an embedded clause agrees not with the subject but with an extracted nonsubject (*the people who John think* are *in the garden*). More 'exotic' cases of agreement (like that between a subconstituent of a complex subject noun phrase and the finite verb in English 'agreement attraction' constructions like *The identity of the participants are to remain a secret*) may, when assimilated to specifier–head agreement, provide evidence for syntactic constituency or derivation as well (cf. Kayne, 1998; Den Dikken, 2000).

Often harder to recast as a specifier–head relationship, long-distance agreement between the matrix verb and an argument of the clause it embeds (as found in Daghestanian, Indic, and Finno-Ugric languages) is restricted in ways which likewise provide highly specific evidence for syntactic structure (see, e.g., Polinksy and Potsdam, 2001). Throughout, agreement is a key diagnostic in the syntactician's toolkit.

References

Abney S (1986) *The English Noun Phrase in Its Sentential Aspect*. Unpublished PhD dissertation, MIT.

Borer H (1986) I-subjects. *Linguistic Inquiry* **17**: 375–416.

Chomsky NA (1981) *Lectures on Government and Binding*. Dordrecht, Netherlands: Foris.

Chomsky NA (1986) *Barriers*. Cambridge, MA: MIT Press.

Chomsky NA (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky NA (2001) Derivation by phase. In: Kenstowicz M (ed.) *Ken Hale: A Life in Language*. Cambridge, MA: MIT Press.

Chung S (1998) *The Design of Agreement: Evidence from Chamorro*. Chicago, IL: University of Chicago Press.

Dikken M den (2000) 'Pluringulars', pronouns and quirky agreement. *The Linguistic Review* **18**: 19–41.

Dixon RMW (1994) *Ergativity*. Cambridge, UK: Cambridge University Press.

Dowty D and Jacobson P (1988) Agreement as a semantic phenomenon. In: *Proceedings of the 5th Eastern States Conference on Linguistics*, pp. 1–17.

Gazdar G, Klein E, Pullum GK and Sag IA (1985) *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Jelinek E (1984) Empty categories, case and configurationality. *Natural Language and Linguistic Theory* **2**: 39–72.

Kayne RS (1998) A note on prepositions and complementizers. Posted on the MIT Press website celebrating Noam Chomsky's 70th birthday (http://addendum.mit.edu/chomskydisc/Kayne.html).

Kayne RS (1995) Agreement and verb morphology in three varieties of English. In Haider H, Olsen S and Vikner S (eds) *Studies in Comparative Germanic Syntax*, pp. 159–165. Dordrecht, Netherlands: Kluwer.

Kayne RS (2001) Pronouns and their antecedents. Ms. New York: New York University.

Keenan EL (1979) On surface form and logical form. *Studies in the Linguistic Sciences* **8**(2).

Kimball J and Aissen J (1971) I think, you think, he think. *Linguistic Inquiry* **2**: 241–246.

Kinyalolo K (1991) *Syntactic Dependencies and the Spec–Head Agreement Hypothesis in KiLega*. Unpublished PhD dissertation, UCLA.

Ouhalla J (1993) Subject-extraction, negation, and the anti-agreement effect. *Natural Language and Linguistic Theory* **11**: 477–518.

Pollard C and Sag IA (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Polinsky M and Potsdam E (2001) Long-distance greement and topic in Tsez. *Natural Language and Linguistic Theory* **19**: 583–646.

Reed W (1991) *Verb and Noun Number in English: A Functional Explanation*. London, UK: Longman.

Zwart CJW (1997) *Morphosyntax of Verb Movement. A Minimalist Approach to the Syntax of Dutch*. Dordrecht, Netherlands: Kluwer.

Further Reading

Barlow M and Ferguson CA (eds) (1988) *Agreement in Natural Language*. Stanford, CA: Center for the Study of Language and Information.

Corbett G (1991) *Gender*. Cambridge, UK: Cambridge University Press.

Corbett G (2000) *Number*. Cambridge, UK: Cambridge University Press.

Kathol A (1999) Agreement and the syntax–morphology interface in HPSG. In: Levine R and Green G (eds) *Studies in Contemporary Phrase Structure Grammar*, pp. 223–274, Cambridge, UK: Cambridge University Press.

Steele S (1990) *Agreement and Anti-Agreement: A Syntax of Luiseño*. Dordrecht, Netherlands: Kluwer.

Anaphora, Processing of

Introductory article

Andrew Barss, University of Arizona, Tucson, Arizona, USA
Janet L Nicol, University of Arizona, Tucson, Arizona, USA

CONTENTS

INTRODUCTION: REFERENCE AND REFERENTS

Whenever a speaker utters a sentence, he or she has a particular message in mind, which the utterance conveys to the listener. An ongoing aspect of communication is the act of pointing out to the listener what objects in the world the speaker is trying to talk about. The speaker chooses certain words and phrases, and uses them to indicate to the listener what thing or things are being described. When a *phrase* of language is used to pick out a *thing* or *entity* in the world (more precisely, either the real world or the inner world of our thoughts), we say that the phrase *refers* to the thing or entity, and that the thing or entity is the *referent* of the phrase. Reference to objects in the world is one of the most fundamental, and common, things we do with language. Virtually every sentence, in any language, contains phrases that refer to things in the world or in our thoughts.

There is a fundamental distinction to be made between *directly referring* and *indirectly referring* phrases. The first type consists of a phrase whose lexical content – the meanings of the word(s) of the phrase – contains sufficient information to allow the listener to understand what entity is referred to. The phrase by itself carries enough information to refer to, or 'pick out', the entity. Typically, a directly referring phrase contains a noun (which describes the kind of entity being referred to) together with various modifiers the speaker might use to further specify the referent. (In this discussion, we use 'speaker' to mean whoever is producing the utterance, and 'listener' to mean whoever is trying to comprehend the utterance, independent of the medium of expression.) Because the phrase is built around the noun, we call it a Noun Phrase, or NP. For example, each of the italicized expressions below is an NP, and in each case it refers to an entity or group of entities:

The fat dog is barking. (1)

The fat dog that lives next door is barking. (2)

Dogs are barking. (3)

These phrases refer to their referents directly, independently of any other words in the sentence (thus the term *independent reference*). If you know what 'the', 'fat', and 'dog' all mean (i.e. if you are a speaker of English), you will know what is being referred to in sentence (1).

DEPENDENT PHRASES AND INDIRECT REFERENCE

The second type of reference, *indirect reference*, works differently. It always involves a pair of NPs. One will belong to a small class of semantically incomplete NPs, called the *proforms*, which includes the personal *pronouns* (English pronouns are listed in (4)); *reflexives* (listed in (5), which always consist (in English) of a personal pronoun plus the suffix '-self' (in the singular) or '-selves' (in the plural)); and *reciprocals*, seen in (6).

English personal pronouns (4)

	First person	*Second person*	*Third person*
Singular	*I, me*	*you*	*he, she, it, him, her*
Plural	*we, us*	*you*	*they, them*

English reflexives (5)

	First person	*Second person*	*Third person*
Singular	*myself*	*yourself*	*himself, herself, itself*
Plural	*ourselves*	*yourselves*	*themselves*

English reciprocals: *each other; one another* (6)

The other NP in the pair is the *antecedent* of the proform, the phrase that the dependent element

'looks back to' to determine its reference. Antecedents will typically be independently referring NPs. Proforms are nouns that are scaled down in meaning and, in English, convey information about person and about number, gender, and humanness. In other languages, proforms may convey additional information, including the familiarity of the speaker and listener, whether the listener is included or excluded, and information about fundamental aspects of the antecedent, such as its orientation in a plane, or whether it is flat. Regardless of the internal information expressed by the proforms, they are typically used in a sentence to 'echo', or repeat, the reference established by using the antecedent. For example, in (7), the reflexive 'themselves' and reciprocal 'each other' refer again to the dogs down the street:

a. The dogs down the street are looking at *themselves*.
b. The dogs down the street are looking at *each other*. (7)

The term *anaphora* refers to this pairing-up of a proform and its antecedent. As an aside on terminology, we note that the ancient Greek grammarians distinguished between cases where the antecedent preceded the dependent element, as in (7), which they termed 'anaphora', from cases where the antecedent followed the proform, which they termed 'cataphora', as in, for example, 'His mother talked about John'. In most modern literature the term 'anaphora' is used to cover both cases, and we follow that convention here. (Cases of the latter type are statistically very infrequent in written and spoken English, a fact which might follow from the increased processing burden placed on the language comprehension system by receiving the proform in advance of its antecedent.)

For the discussion below, it is crucial to keep in mind that sentences such as (7) contain two separate references to objects: the antecedent (e.g. 'the dogs down the street') refers directly to the referent; and then the proform connects back to the antecedent, and refers to whatever the antecedent did. Thus, in (7a), there is no way to determine what 'themselves' refers to without first figuring out what the antecedent is; the reference by the word 'themselves' to the dogs is indirect, mediated by the connection of the reflexive to its antecedent phrase. Proforms are also called *dependent expressions*. The two NPs are said to co-refer when they have such a dependent connection. Anaphora, then, is the link between a proform and its antecedent, which establishes co-reference. Figure 1 illustrates this graphically.

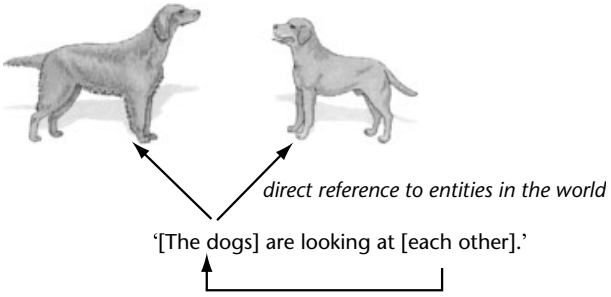


Figure 1. The antecedent–proform relation, creating indirect reference to the dogs.

CONSTRAINTS ON ANAPHORA

There is a major contrast between the pronouns in (4) and the reflexives and reciprocals in (5, 6). Typically, reflexives and reciprocals appear only in very limited positions within a sentence, and pronouns typically appear in cases of anaphora where a reflexive would be unacceptable.

The antecedent of a reflexive or a reciprocal will usually be structurally close to the proform, typically occurring as the subject of the sentence in which the reflexive occurs (square brackets are used here to mark the boundaries of propositions or clauses; '*' indicates that the sentence with the intended interpretation is not grammatically acceptable):

a. [John thinks that [Mary admires herself]].

b. *[John thinks that [Mary admires himself]]. (8)

By contrast, the antecedent of a personal pronoun has a different distribution, and cannot be in the same position as the antecedent of a reflexive or reciprocal:

a. *[John thinks that [Mary admires her]].

b. [John thinks that [Mary admires him]]. (9)

Notice the complementariness of the distribution of proform–antecedent pairs. (8b) is ungrammatical because the antecedent, 'John', is structurally too far away from the reflexive. This then permits the use of a pronoun to co-refer with 'John', as in (9b). By contrast, in (8a), the antecedent is close enough to the proform to permit use of a reflexive. This consistently blocks the use of a pronoun: (9a) cannot be understood to mean co-reference between 'her' and 'Mary'.

Consequently, in the course of understanding language, the listener must keep track of (1) the type of proform being used (pronoun versus reflexive/reciprocal), (2) where it occurs in the sentence, (3) the position(s) in which other NPs, which are candidate antecedents, occur, and (4) the person, number, and gender features of each NP (including the proform), which further constrain the pairing up of a proform with an antecedent.

Above we noted that a proform and its antecedent must match in person, number, and gender. Notice that (10a) is ambiguous (since both NPs match the pronoun in features), while (10b) and (10c) are unambiguous, since only one NP matches:

a. The boy told the man that [he was lucky].
b. The girl told the man that [he was lucky].
c. The boy told the woman that [he was lucky]. (10)

There are some cases in which feature mismatch is permitted. One is the use of the genderless, normally plural 'they/them' as an indefinite singular pronoun, when the speaker does not wish to seem gender-biased in using the singular (English contains no singular, gender-unmarked pronoun), as in (11a):

a. Every student must make sure they check in with the nurse.
b. Every student must make sure he checks in with the nurse. (11)

A second case is the use of the normally second-person pronoun to mean 'one', 'anyone', in casual speech:

a. You shouldn't do that.
b. One shouldn't do that. (12)

(12a) is ambiguous. It may be understood as directed at the listener(s), in which case the second-person feature is required. But it may also be taken as a casual variant of (12b), in which case the reference is generic, and not restricted to the listener(s).

In general, these mismatches can be seen as a type of compromise: a way of avoiding gender commitment or formal register.

THE PROCESSING OF ANAPHORA

The term *processing* refers to how the human mind registers and makes sense of sensory information over time. Cognitive processes, such as language production and comprehension, are largely unconscious and automatic. When listeners hear a sentence, they must first recognize the words contained in the sentence. For instance, for the sentence 'The dogs are looking at the mail carrier', the listener first must determine that the words 'the' and 'dogs' occurred, then figure out what the words mean, and then combine the meanings. Word meanings seem to just pop into mind: when we hear the words 'the dogs', we automatically think of the concept DOG. This process – of identifying words, thinking about their meanings, and combining meanings together – occurs over and over in the course of communication. One might wonder what happens to all this information: if the DOG concept is in mind, what happens to it when 'mail carrier' is uttered? Clearly, a listener would have to keep both concepts (THE DOG and THE MAIL CARRIER) in mind simultaneously (along with the LOOKING concept) in order to understand the sentence. And, of course, the concepts must be kept separate, so that we do not think of the dogs as being mail carriers, or the carriers carrying both mail and dogs. There is some evidence that words and their meanings are kept in mind until the end of a proposition (usually the end of a sentence), and then they are converted into a less detailed memory representation.

The presence of anaphora complicates matters considerably. In a sentence such as 'The dogs are looking at each other', what meaning 'pops into mind' when 'each other' appears? Do listeners automatically think about 'the dogs' again? Or does it require time and effort to make the connection between 'each other' and 'the dogs'? Before we address this question, we will describe how research on the processing of anaphora has been conducted.

Methods

There have been three basic approaches to the study of how proforms are interpreted. (1) Ask people. The most direct approach is to have people read sentences and then ask them about their

interpretations (e.g. presenting a sentence like 'John told Bill that Frank admired him' and asking 'Who does *him* refer to?'). The trouble with this approach is that it does not reveal anything about the process over time; it probes only the final interpretation of the sentence, and so does not answer any of the questions raised just above. There are two types of methods that do provide information about how a listener or reader understands a sentence as it unfolds.

(2) Probing word meanings. One variant of this method probes the activation of word meanings by having people listen to sentences and simultaneously make judgments about items they see on a computer screen. Here is how it works. Listeners hear the phrase 'The dog' at the beginning of the sentence. Then immediately they see (on a computer monitor) either a related word such as 'cat' or a completely unrelated word such as 'pen'. They are asked to simply decide whether the word they see is a real word of English, and press one button if it is and another button if it is not. The logic is this: if they see the word 'cat' while they are thinking DOG, then listeners will be quite fast to respond, faster than if they see 'pen' while thinking DOG, because the concepts CAT and DOG are connected in our word system, and PEN and DOG are not. By measuring the time it takes listeners to respond, experimenters can get some idea of what people are thinking about. Now suppose that listeners hear the sentence 'The dog down the street hurt itself yesterday'. If we probe what listeners are thinking about when they hear the word 'itself', we should be able to tell whether they are thinking about the concept DOG. If they are, this means that the connection between the reflexive 'itself' and the antecedent 'the dog' is established very quickly indeed. Another variant on this technique simply repeats the antecedent, and listeners or readers must decide if the word appeared in the sentence or not. The logic is similar: if people are thinking about DOG after the reflexive 'itself', they should be relatively fast to indicate that the word 'dog' was in the sentence.

(3) Measuring reading. Another approach has been to have people read sentences and to determine at what points in the sentences they slow down. In one variant of this approach, researchers construct sentences that have a 'surprise' ending IF the reader has adopted a particular interpretation. For example, given the sentence 'The actress liked the queen because she thought the queen did a good job', readers could initially assume at the point where 'she' occurs ('The actress liked the queen because she…') that 'she' co-refers with 'the queen'. Such an assumption is reasonable because an explanation for the proposition that 'the actress liked the queen' is likely to involve something that the queen did. If readers do make this assumption when they read the pronoun 'she', then they will be surprised when material after the pronoun reveals that 'she' is really 'the actress'. What happens when readers are surprised by how a sentence unfolds is a slowdown in reading (compared to the reading of a sentence that does not contain a surprise ending, for example 'The actress admired the queen because she thought the actress did a good job'). Reading time can be measured with the use of a device called an *eye tracker*, which, when it is linked to a computer, determines where on a computer screen the eye fixates and for how long (and whether the reader backtracks in order to re-read a section of text). A simpler way to measure reading time is to (have a computer) present pieces of a written sentence in sequence, in such a way that only one section is visible at a time. The reader controls the presentation by pressing a designated key on a computer keyboard. The key press makes the current sentence fragment disappear and the next fragment appear. The time between key presses provides a measure of reading time.

The Experimental Findings

Experiments that have focused on the questions raised above suggest that listeners do automatically think about the antecedent of the proform – immediately, or very soon after the proform appears – if they are reasonably attentive to what they are hearing. Hence, it appears to be automatic that proforms serve to reinstantiate their antecedents. Such findings raise a number of questions: does the information about sentence position have an effect on what people think about when they encounter a proform? What happens when there is more than one possible antecedent for a proform? Must a single antecedent be selected right away, or does a whole set of possible antecedents get computed, awaiting final narrowing down? What sources of information enter into the process? We will consider each of these questions in what follows.

First, recall the facts about the complementarity of the distribution of antecedents of pronouns on one hand and reflexives/reciprocals on the other. Examples are shown in (13).

a. The actress thinks that the queen admires her.
b. The actress thinks that the queen admires herself. (13)

In (13a), 'the queen' cannot be the antecedent of the pronoun 'her', but 'the actress' can be. In (13b), 'the queen' must be the antecedent of the reflexive 'herself' and 'the actress' cannot be. Some research suggests that knowledge about these restrictions constrains the process: listeners and readers think about only those antecedents that are 'allowed'. Further, they think about only those antecedents that agree in number and gender (this point is discussed in greater detail below).

Now consider the case in which there is more than one possible antecedent for a proform, as in the following examples.

The actress told the queen that the director wanted to meet her after dinner. (14)

The actress told the queen that she would be sitting next to the director. (15)

In (14), both 'actress' and 'queen' are possible antecedents of the pronoun 'her'. It might be slightly more likely that the director would want to sit with a queen rather than an actress (this is information about how the world works, or information about *plausibility*), or people might tend to think that since 'her' is the object of a verb (the verb 'meet'), an object ('the queen' in the first clause) is a better antecedent (this is called the *parallel function* interpretation). These kinds of information do matter, but evidence from word-probe experiments suggests that they do not come into play immediately. That is, listeners think about both 'actress' and 'queen' after hearing 'her'. Then they might use a variety of cues to settle on one antecedent or the other. (Typically, listeners do attempt to link proforms with antecedents right away; it would be difficult to follow a conversation without doing so.)

In (15), again, both 'actress' and 'queen' are potential antecedents of the pronoun, and both would come to mind after the occurrence of 'she'. Then other cues would be used to narrow down the set: parallel function would dictate that since 'she' is a subject, another subject ('the actress') should be the antecedent. Another factor that appears to affect pronoun interpretation is the order of potential antecedents: the first NP in the sentence has a sort of privileged status, and, given a choice, people are more likely to fix on that NP as the antecedent. A third factor has to do with plausibility: the likelihood that an actress would inform the queen about where the queen would be seated at dinner; the likelihood of the queen being seated next to the director, and so on. Again, it appears that such information acts to eliminate antecedents from a candidate set: the proform makes a listener think about a number of antecedents, selected via inspecting just the sentence structure, and other information acts later on to eliminate potential antecedents from the set.

It should be emphasized here that such instances of antecedent ambiguity are far from uncommon. This is because speakers are unaware when they say a pronoun that it may not be clear who the antecedent is; after all, for the speaker, there is no confusion at all. (Sometimes they do catch themselves, and identify the antecedent: e.g. '… and then she… that is, Susan, disappeared…'. Other times a listener might ask 'she, who?'.)

Now consider example (16).

The actresses told the queen that they were going to be late. (16)

Here there are also multiple antecedents for the pronoun 'they': just the 'actresses' or both 'the actresses' and 'the queen'. Research shows that listeners consider just 'the actresses', which matches in number with 'they'. When a matching antecedent is available, as it is here, listeners tend not to consider a joint referent consisting of both the actresses and the queen. (Note that, in contrast, a sentence such as 'The actress told the queen that they were going to be late' requires the listener to infer that 'the actress' and 'the queen' together constitute the antecedent of 'they'. It is not known at present whether or not this inference takes additional time.)

The examples above contain nouns that are gender-specific: 'actress' and 'queen' are both feminine. What happens when a gender-neutral noun appears, as in the following?

The supervisor warned the actress that she was going to be late. (17)

Only a handful of English nouns are specified for gender. These include kinship terms (e.g. 'aunt, grandfather, daughter'), royal titles (e.g. 'king, princess, duchess'), and miscellaneous others, many of which are disappearing from English (e.g. 'aviatrix, murderess'). Of the miscellaneous set of feminine-marked forms, the male counterparts are not clearly marked for gender: although 'aviatrix', 'murderess', and 'actress' must refer to women, 'aviator', 'murderer', 'actor' can be applied to both men and women. In general, most occupation nouns are gender-neutral: e.g. 'doctor, nurse, lawyer, director, president, manager, supervisor, teacher, professor, student, programmer, architect, gardener, coach'. Of this set, there is considerable variation with respect to the probability with which a term is likely to refer to a man or a woman. While 'student' seems to be equally likely

to refer to a male or a female, some nouns are more likely to refer to men (e.g. 'astronaut') and some to refer to women (e.g. 'nurse'). How does this type of information affect the processing of anaphora? There is not much research in this area, but the research to date suggests that listeners initially consider as a potential antecedent for a proform any noun that does not actually clash in terms of gender. Given a sentence like (17), and no additional information about the gender of the referent of 'supervisor', both 'actress' and 'supervisor' would initially be considered potential antecedents. Then the gender-probability information, along with information about real-world plausibility, parallel function, and order of mention, are all likely to come into play to guide the selection of a single antecedent. In a sentence such as (17), there will be conflicting information: parallel function and first-mention weight 'the supervisor' more heavily. But gender marking favors 'the actress'. Pragmatics arguably favors both equally, since the plausibility of 'being late' applies equally to both. The processing of sentences such as this should be slower than for sentences in which all the information favors the same antecedent.

Overall, the research suggests that the processing of anaphora involves a number of stages. First, the occurrence of a proform evokes a candidate set of antecedents. This set includes only those antecedents that are grammatically allowed, and only those that are not incongruent with respect to gender and number. Next, other types of information (about order of mention, sentence function, gender-probability, and plausibility) act to eliminate nouns from the candidate set.

Actually, the picture is more complicated than this because sometimes potentially helpful information about the identity of the antecedent is not available until later in the sentence. Take a sentence such as 'The director told the actress that she couldn't recommend her for the role'. At the point in the sentence at which 'she' occurs, it really is not clear which is the correct antecedent. Competition of the other types of information could lead the listener to settle on 'actress' (possibly because there are many more male directors than female ones). But the appearance of 'her' throws this interpretation into question: if 'she' co-refers with 'actress', and if 'her' cannot co-refer with 'she', then 'her' co-refers with 'director', and here's where the sentence goes awry. It is unlikely (but not impossible) that someone would tell someone else whom he or she could recommend. So, 'she' ought to co-refer with 'director', and 'her' with 'actress'. And when it is revealed that the recommendation is for a role, then this confirms the revised interpretation.

Given that late-occurring material can inform the interpretation of proforms, it might make more sense for listeners to postpone co-reference until the end of the sentence. Listeners and readers could simply take the information conveyed by a proform (e.g. third person, female, plural, subject) and plug that into the ongoing meaning of the sentence. But it appears to be the case that, as in language comprehension in general, a 'wait-and-see' approach carries too high a cost. The cost is that we might forget some of the critical details of an utterance and lose track of what the sentence is conveying. After all, for the most part, human beings talk about people and things. During a discourse, a speaker and listener both construct a *mental model* of the participants and events being discussed. In order to follow a discourse listeners and readers must figure out who is doing what to whom, and it appears that they make these calculations quickly, lest they forget.

Further Reading

Cloitre C and Bever T (1988) Linguistic anaphors, levels of representation, and discourse. *Language and Cognitive Processes* **3**(4): 293–322.

Ehrlich K and Rayner K (1983) Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior* **22**: 75–87.

Garnham A, Oakhill J and Cruttenden H (1992) The roles of implicit causality and gender cue in the interpretation of pronouns. *Language and Cognitive Processes* **7**: 231–255.

Gernsbacher M, Hargreaves D and Beeman M (1989) Building and accessing clausal representations: the advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language* **28**: 735–755.

MacDonald MC and MacWhinney B (1990) Measuring inhibition and facilitation from pronouns. *Journal of Memory and Language* **29**: 469–492.

Matthews A and Chodrow M (1988) Pronoun resolution in two-clause sentences: effects of ambiguity, antecedent location, and depth of embedding. *Journal of Memory and Language* **27**: 245–260.

McDonald J and MacWhinney B (1995) The time course of anaphor resolution: effects of implicit verb causality and gender. *Journal of Memory and Language* **34**: 543–566.

McKoon G, Greene S and Ratcliff R (1993) Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory & Cognition* **19**(5): 1040–1052.

Nicol J and Swinney D (2002) The psycholinguistics of anaphora. In: Barss A (ed.) *Anaphora*. Oxford, UK: Blackwell.

Shillcock R (1982) The on-line resolution of pronominal anaphora. *Language and Speech* **25**: 385–401.

Stevenson RJ, Nelson AWR and Stenning K (1995) The role of parallelism in strategies of pronoun comprehension. *Language and Speech* **38**(4): 393–418.

Anaphora

Advanced article

Gregory Carlson, University of Rochester, Rochester, New York, USA

CONTENTS

Anaphora refers to referentially dependent expressions in natural language which contribute their meaning by identifying another expression to give them their semantic value.

INTRODUCTION

Anaphora, in its primary instances, is the establishment of a referential dependency between two (or more) expressions. The pronoun 'him' in (1) below is one such instance of anaphora:

Mark felt that there was someone watching *him*. (1)

On the understanding that 'him' refers to Mark, the pronoun is the *anaphor* and the expression 'Mark' is the *antecedent*. Both expressions refer to the same individual. The relationship between these expressions is not an equal one, however, since the reference of the pronoun is dependent upon the reference of its antecedent, whereas the reference of the antecedent is established by virtue of its meaning alone. The term 'co-reference' is often used to describe this referential connection between anaphor and antecedent. But anaphor–antecedent relations must be distinguished from the phenomenon of *accidental co-reference*. This occurs when two independently referring expressions happen to refer to the same individual. So, for instance, in (2) the two italicized expressions will be co-referential, 'accidentally', only when the president of the company is also the company's best employee:

The president of the company rewarded *the best employee*. (2)

This requires an understanding where the company has a self-rewarding president, but there is no anaphoric connection established between the expressions. Thus, anaphora is a matter of co-reference, and something more.

ANAPHOR–ANTECEDENT RELATIONS

Anaphors depend upon their antecedents to determine their referential content. One reflection of this referential dependency is that in many instances an anaphor cannot be interpreted as co-referential with another noun phrase (NP). For instance, in the following examples, the pronouns cannot be construed as non-accidentally having the same reference as the italicized NPs:

Bob was nominated by him. (him $\neq$ Bob) (3)

She hoped that *Mary* would win the contest. (she $\neq$ Mary) (4)

This is because an anaphor cannot receive its reference from another NP if that NP does not have an appropriate syntactically defined relationship to the anaphor. This relationship is not simply one of linear precedence, as in many instances an anaphor may precede its antecedent (a phenomenon which is occasionally called *cataphora*, though more commonly *backward anaphora*):

Near *her*, *Jill* saw a snake. (5)

If *he* wins the race today, *Bret* will be a hero. (6)

Much research has focused on the question of the precise nature of this syntactic relationship. The research is detailed and extensive (for example, research on *Binding Theory*). Most agree that the notion of *c-command* is crucial (Langacker, 1966; Lasnik, 1976). In general, an anaphor cannot c-command its antecedent, and in examples such as (3) and (4) above where the two designated NPs cannot be interpreted co-referentially, the pronoun would c-command its antecedent, and a referential connection cannot be established. The reference for the pronoun in these instances needs to be determined by other means, such as finding another, appropriate antecedent for it, or by providing it

with a *deictic* interpretation (discussed further below). (*See* **Binding Theory**)

One class of pronouns that has also received extensive attention is that of *reflexive pronouns*, exemplified below:

We found *ourselves* with too much to do. (7)

The professor taught *herself* French. (8)

These differ from the other personal pronouns in important respects. Primarily, the syntactic relations to their antecedents are much more limited. In general, reflexive pronouns may only have antecedents within the same clause, though the precise conditions remain a topic of detailed investigation. In the following examples, the reflexive pronoun may not be construed as co-referential with the italicized NPs:

We thought that [$_S$Jim liked ourselves] (9)

The professor remembered when [$_S$herself lived in Paris] (10)

As there is no appropriate antecedent for the reflexive pronoun within the same clause in these instances, the sentences are not grammatical.

Pronouns not only may find their reference by identifying an antecedent and using the reference of the antecedent as its own value, but they may function as *bound variables* as well. In such instances, the 'reference' of the pronoun is not determined by the reference of its antecedent NP, but rather by the assignment of values to variables that is determined by the quantifier, as in first-order logic. A representation of a sentence such as (11), with 'Every man' construed as the antecedent of 'he', would be as indicated:

Every man thinks that [*he* deserves a raise]
$\forall x$ [man(x) $\Rightarrow$ x thinks that [x deserves a raise]] (11)

Bound variable pronouns and their antecedents are syntactically more constrained than identity of reference pronouns and theirs. In the following examples, (12a) precludes any bound variable reading; this is despite the fact that a natural identity of reference reading is available when the antecedent NP has a clear referential value, as with proper names (12b):

a. The dean who placed *no student* on probation told *her* to check back in the fall.
b. The dean who placed *Hillary* on probation told *her* to check back in the fall. (12)

(12a) has no bound variable interpretation, because the antecedent is in a syntactic position which precludes this possibility. The relation that must hold, in the case of bound variable readings, is for the antecedent NP to c-command the pronoun.

The phenomenon of anaphora is much broader than the personal pronouns discussed thus far. One form of anaphora that has received much attention is *temporal anaphora* (Partee, 1984). This applies not only to pronouns referring back to time NPs,

The mail arrived *this morning*. I was at home *then* (= this morning) (13)

but also to the time introduced by the *tense* of a sentence:

Ali woke up. It was cold *then*. (14)

The study of temporal anaphora includes a wide variety of forms which are used to coordinate the time in one sentence with that of another. Beyond 'then', expressions such as 'when, before/after, until, as, while, since, immediately thereafter', and many others fall within this domain. Perhaps most significantly, tenses themselves appear to function anaphorically. In the examples below, the tense in the second sentence is understood as coordinated with the time reference in the first:

Samantha opened the door. She *had been* repairing the car. (15)

Daryl fell down. He *was* drunk. (16)

Our understanding that the repairing occurred prior to the door opening, or that the falling occurred while Daryl was drunk, is often attributed to temporal anaphora.

A wide variety of other anaphoric forms, beyond personal pronouns and temporal anaphora, make reference to an extensive array of other types of things. These include the demonstratives 'this' and 'that' (with or without a following noun), and epithetics such as 'the fool' or 'the bastard'. Other forms take as antecedents phrases that are not NPs. 'So' may take a verb phrase as an antecedent; 'such' takes a modifier; 'there' may take a locative prepositional phrase; 'one' may take a noun:

a. Sam tried to *win the race* before Al could do *so*.
b. If *intelligent* students attend college, *such* students usually do very well.
c. Everyone who was *at the party* had a good time *there*.
d. I own a big *car*, and you own a small *one*. (17)

In many cases the anaphor is expressed as null: that is, the anaphor is indicated by having some

constituent missing from the sentence. The following sentence is missing a noun plus its modifying adjective at the point indicated by the blank:

Jack owns three *large dogs*, and I own two____. (18)

In this case, the 'blank' takes as its antecedent the portion of the NP italicized. It functions as an anaphor in the same way as a pronoun.

Null anaphora extends well beyond nouns and NPs. Verb phrases (VPs) can function as antecedents for null VPs (known as *VP ellipsis*):

If you want to____, we can *take a break.* (19)

Verbs, and verb complexes, can serve as antecedents in the *gapping* construction:

Joseph *ate* a bagel, and Samuel, ____ a grapefruit. (20)

Null complement anaphora takes complement sentences as antecedents:

Kevin claimed *that our television was broken, but I'm not so sure*____. (21)

This by no means exhausts the range of anaphora expressed by null expressions.

DISCOURSE ANAPHORA

'Discourse' is the normal mode of communication: the use of more than one independent sentence or utterance put together in a way that 'makes sense'. The discussion above was limited to those instances of anaphora that take place within the boundaries of a sentence. Anaphora takes place across sentence boundaries as well. Many instances of anaphora that appear within sentence boundaries take place as well in discourse:

Several team members were suspended. Reportedly, *they* had missed a practice. (22)

Most people want to *win a million dollars.* Doris doesn't____. (23)

Certain cases of anaphora that occur within the boundaries of a sentence do not function as discourse anaphors. For instance, the phenomena of reflexive pronouns, gapping, relative pronouns, and bound variable anaphora do not appear to be able to function this way.

One treatment of discourse anaphora is to treat all such pronouns as *free variables*, which are assigned a reference independently by an *assignment function*, which designates a referential value for any free variables within its domain (e.g. Cooper and Parsons, 1976). It becomes co-referential with a NP in a previous sentence by virtue of being assigned the same reference. (*See* **Semantics, Dynamic**)

So, in (24), if a function assigns the same referential value as the proper name 'Leonard' has, to the pronoun 'he' in the following sentence, then a co-referential reading arises:

Leonard is a famous composer. *He* writes operas. (24)

On the other hand, if 'he' is assigned a different value (e.g. Fred), then the discourse will be understood as saying that Fred writes operas, and no co-referential reading will occur. All phrases with which the pronoun is co-referential must have a reference value in the first place, if this is to be the appropriate analysis. The case of indefinite NPs in discourse raises questions, though. Indefinite NPs are those which appear with a number of different determiners, most prominently the indefinite article 'a(n)'. Such NPs can be 'referred back to' by anaphors in discourse:

A man walked into the room. *He* sat down. (25)

Most researchers, however, question whether an indefinite NP should be properly assigned a reference value (Kamp, 1981). This is because the reference value determines the truth-conditions of the sentence, and if one assigns a certain individual as the reference of 'a man' in a sentence such as (25), then it would be true if that *particular* man walked into the room, and false if he did not (regardless of whether any other man walked in). However, these are not the truth-conditions for such a sentence, since (an utterance of) the sentence will be true if any man whatsoever walked into the room (and false only if no man at all did). If one assigned a reference for 'a man' as some particular man, one could not characterize these truth-conditions. It appears that the truth-conditions of the sentence are best represented quantificationally, with an existential quantifier binding a variable:

$\exists x$ [man(x) & x walked into the room] (26)

This treatment raises some problems, however, when we turn to discourse anaphora. The representation of the two-sentence discourse in (25) would have the existential quantifier binding the instance of 'he' in the subsequent sentence:

$\exists x$ [man(x) & x walked into the room & x sat down] (27)

Since anaphors referring back to indefinites are both very common and natural, unlike those functioning as bound variables with their antecedent quantified expressions in another sentence, one would need to make a separation between classes of quantifiers, some of which could bind variables in other sentences, and others which could not. This has proven an unsatisfactory analysis of the phenomenon, however, for both syntactic and semantic reasons. One of the main issues has centered around the treatment of *donkey sentences* (or, *donkey anaphora*). Such examples are so called because of the example below, commonly cited in the literature:

Every farmer who owns *a donkey* beats *it*. (28)

These sentences pose a problem of logical representation that has been known since medieval times. The problem is this. If one were to take 'it' in this sentence to be a free variable assigned the same reference as 'a donkey', there is, very clearly, no particular donkey which this sentence is in any way 'about'. The other, more attractive, possibility is that the pronoun is functioning as a bound variable, bound by an existential quantifier that is taken to be the meaning of the indefinite article. However, the only consistent representation available is essentially as follows:

$\exists x$ [donkey(x) & Every farmer who owns x beats x] (29)

The truth-conditions of this (which are directly reflective of the meaning), however, are very different from the truth-conditions of the sentence itself. This formula is true only when there is some donkey or other that every owner of it beats, which is far from the meaning of the sentence itself. There is no consistent way of representing the meaning by treating the indefinite article as an existential quantifier which binds the pronoun in the predicate of the sentence.

Replacing the quantificational analysis is one where indefinites are treated as contributing no existential meaning on their own, but only a free variable and a property ascription; so indefinites have neither inherent reference nor inherent quantificational force. The free variable implicitly contributed by the indefinite is bound by an operation of *text closure*. This is where the existential force arises. So, a single sentence such as the following has the representation given immediately below it, and then text closure operates to bind the variable as indicated:

A man walked into the room.
man(x) & x walked into the room
$\Rightarrow$ Text closure
$\exists x$ [man(x) & x walked into the room] (30)

Text closure is formulated in such a way that it operates over stretches of discourse, as more sentences are added. So a two-sentence discourse would be represented and operated on by text closure as follows:

A man walked into the room. He sat down.
man(x) & x walked into the room & x sat down
$\Rightarrow$ Text closure
$\exists x$ [man(x) & x walked into the room & x sat down] (31)

This analysis allows us to distinguish quantified NPs from indefinites, on the one hand, and allows us to treat the pronouns as free variables at the same time. Also, though not presented here, it offers a solution to the donkey sentences problem.

This approach raises issues of its own, as illustrated in the following sentence:

There is a man in the garden. The dog is barking at *him*. (32)

The 'There is …' construction in English quite plausibly introduces an existential quantifier of its own, rendering the variable contributed by 'a man' unavailable for binding by text closure. But the pronoun in the second sentence could be bound by text closure. If this is so, then the text would have the meaning 'Some man is in the garden. The dog is barking at someone'. Another problem with text closure is that the representation

$\exists x$ [man(x) & x is in the garden & the dog is barking at x] (33)

will be true also in cases where there are more men in the garden than just one. However, the original text means – or possibly strongly implies – that there is one and only one man in the garden.

Evans (1980) has argued that there is a need for still another category of pronoun, which he calls *E-type pronouns*. These pronouns, like the bound variable and co-referential pronouns, share all the same forms, but function differently: they are disguised definite descriptions, picking out a unique individual given the information present in the context. Informally, the analysis of the pronoun 'him' in (32) would be:

There is a man in the garden. The dog is barking at him (= *the man that is in the garden*). (34)

Since these descriptions can contain pronouns, or variables of their own, one can obtain solutions to problems like the following (an example that is often called a *pronoun of laziness*, a term coined by Peter Geach since it was a 'lazy' way to avoid repeating an entire NP):

The woman who deposited *her paycheck* in the bank was wiser than the woman who gave *it* to her teenage son. (35)

In this case, analyzing 'it' as meaning 'her paycheck', with 'her' in this instance assigned the same value as the second woman rather than the first, will give the right reading. However, on a co-referential reading (or a bound variable reading) the paychecks would have to be one and the same.

IDENTITY OF SENSE AND IDENTITY OF REFERENCE ANAPHORA

A traditional distinction is made between what are called 'identity of sense' and 'identity of reference' anaphora. The distinction between sense and reference goes back to the writings of the philosopher Gottlob Frege. In the case of NP meanings, this distinction concerns whether the individuals designated by the antecedent and the anaphor must be interpreted as identical. So, in (36a) the cars driven by Lyle and Maria must have been the same; however, in (36b) they need not:

a. Lyle drove *a car*. Maria drove *it*, too.
b. Lyle drove *a car*. Maria drove *one*, too. (36)

The difference between the anaphors 'it' and 'one' (the latter taking a noun meaning as its antecedent) would seem to suggest that anaphors themselves fall into these two classes. While this is so to a certain extent, many instances of anaphora can be identified in which the same form can play both roles. Consider the following:

a. *The President* (of the United States) walked off the plane. *He* waved to the crowd.
b. *The President* is elected every four years. *He* has been from a southern state ten times. (37)

The *reference* of the phrase 'The President' is whoever happens to be in that office at the time – currently it is George W. Bush; the *sense*, on the other hand, is that which picks out the president at the time, whoever it may be. In (37a), 'he' refers to a certain individual – Bush, for instance. The sentence 'he' appears in would be synonymous with saying 'Bush waved at the crowd'. In (37b), 'he' is anaphoric to the sense of the term, not its reference. It does not follow that any particular president, such as the current one, has been from a southern state ten times. Rather, this instance of the pronoun is talking about the presidents of the USA across time, regardless of who that individual may currently be. That is, it refers to the sense of the antecedent, not its reference.

When we turn to other types of anaphora, it is more difficult to make this sense/reference distinction. Consider null VP anaphora:

Zelda will *get up early* if Harry does____. (38)

The question that arises in this case is whether VPs themselves have a sense/reference distinction in their meanings to begin with. If, for instance, VPs have individual events as their reference, and have classes of events as their sense, then VP anaphora would be sense anaphora (as presumably Harry waking and Zelda waking would be distinct events). One can make reference to individual events by using the pronoun 'it', as exemplified below, but VP anaphora does not appear to ever make reference to events in this particular way:

The train blew its whistle. *It* (= the blowing of the whistle) was heard for miles. (39)

A similar situation holds for anaphora to sentence meanings. In a Fregean analysis, the sense of a sentence is a proposition, and its reference is a truth-value (T or F). However, with Null Complement Anaphora, which takes sentences as antecedents, the proposition rather than the truth-value is clearly the value assigned to the anaphor, as any other proposition with the same truth value (e.g. 'Grass is green') does not yield a synonymous sentence:

Bruno was selling drugs. When the FBI found out____, he was arrested. (40)

Thus, the sense/reference distinction is most useful in describing anaphora to NP meanings.

PRAGMATIC ANAPHORA

Pragmatics, that is knowledge of how the world works and what it contains, the circumstances under which a sentence is uttered, and of how language is used, is crucial for the study of anaphora. In most instances, a pronoun or other anaphor could, in principle, find more than one unique antecedent in the sentence or discourse, as in the following:

Mary told the *woman* talking to *her sister* that *Lesley* was sick today. *She* then turned and walked away. (41)

While 'she' could, in principle, find any of the previous NPs as its antecedent, in practice only one 'makes sense' and so is the one that is readily understood (in this instance, Mary). *Centering Theory* is one proposal that attempts to deal with this phenomenon (Grosz *et al*., 1995).

Another area requiring pragmatic knowledge to resolve reference of anaphora is *bridging inferences* (Clark, 1975). The listener or reader must make use of real-world knowledge to interpret a definite NP appropriately. For example:

John bought *a new car*. *The engine* was painted bright red. (42)

Here, one knows that the engine is the engine in the car that John bought, making use of real-world knowledge that cars have engines.

Much work in pragmatic anaphora focuses not on the process of selecting an appropriate antecedent from candidates given in the text or discourse, but on instances where the sentence or discourse itself provides no possible antecedents for an anaphor. For instance, imagine I was standing on the street with someone else when a man walks by very abruptly. Under these circumstances I can say:

He appears very upset. (43)

In so doing, I refer to the man that just walked by even though there is no expression within the previous discourse to serve as an antecedent for the pronoun. The man himself, in some sense, provides the 'antecedent' for the pronoun. When elements themselves in the real-world context of use provide the values for anaphoric expressions, they are said to be *pragmatically controlled*.

The example above might suggest that perceptual evidence establishes possible antecedents for *deictic* uses of pronouns. However, having the referent perceptually available is not always necessary. Consider the case where I am walking down the hallway at work, and a student is knocking on the door of the office of another faculty member. I can say, under the circumstances:

I haven't seen *her* today. (44)

even though the professor is not perceptually available for reference at the time. Thus, some contexts are 'rich' enough to support pragmatic control even in the absence of who or what is being referred to.

Most (but not all) instances of anaphora may be pragmatically controlled, including certain instances of reflexive pronouns and *logophoric* pronouns. These are pronouns, indicated by specialized forms in some languages, which are canonically used in indirect discourse to make reference to the person whose speech is reported (e.g. 'Ariel said that *he*[logophoric] was going to write a paper'). Below are instances of other types of anaphora that may be controlled pragmatically:

a. [Picking up a coat from the coat-check attendant] '*This* is torn!'
b. [Pointing through the glass at the candy counter] 'A green *one* and a red *one*, please.'
c. [Sally hides cigarettes in her room. Her sister, seeing this, says:] 'Better hope our parents don't find out____.'
d. [Trying on suits, and the salesman says:] '*Which* appeals to you most?' (45)

Certain instances of anaphora cannot be pragmatically controlled (Hankamer and Sag, 1976). These are the instances of *surface anaphora*, which, unlike *deep anaphora*, require a specifically linguistic antecedent. The distinction between deep and surface anaphors hinges not only on whether they may be pragmatically controlled, but also on whether, when there is a linguistic antecedent, the syntactic details of the antecedent determine the possibility of it serving as an antecedent, regardless of what meaning is expressed (which applies to surface anaphora but not deep).

Gapping, null anaphora to a verb or verb complex, requires an explicitly (surface) linguistic antecedent, even in very clear contexts, and cannot be pragmatically controlled; the example below is not felicitous:

[Bob throws a baseball] '. . . and Cary, a basketball.'
(Contrast with: 'Bob threw a basketball, and Cary, a baseball.') (46)

Sluicing likewise requires a linguistically introduced 'surface' antecedent, so the following too sounds strange:

[From outside, a scream; a shot is fired; and a thud] 'I wonder who____?' (47)

Similarly, the phenomenon of bound variable pronouns is not amenable to pragmatic control.

References

Clark H (1975) Bridging. In: Schank R and Nash-Webber B (eds) *Theoretical Issues in Natural Language Processing*. Cambridge, MA: MIT Press.

Cooper R and Parsons T (1976) Montague Grammar, generative semantics, and interpretive semantics. In:

Partee B (ed.) *Montague Grammar*, pp. 311–362. New York, NY: Academic Press.
Evans G (1980) Pronouns. *Linguistic Inquiry* **11**: 337–362.
Grosz B, Joshi A and Weinstein S (1995) Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* **21**: 2.
Hankamer J and Sag I (1976) Deep and surface anaphora. *Linguistic Inquiry* **7**: 391–428.
Kamp H (1981) A theory of truth and representation. In: Groenendijk J, Janssen T and Stokhof M (eds) *Formal Methods in the Study of Language*, pp. 277–322. Amsterdam, Netherlands: Mathematisch Centrum.
Langacker R (1966) On pronominalization and the chain of command. In: Reibel W and Schane S (eds) *Modern Studies in English*, pp. 160–186. Englewood Cliffs, NJ: Prentice-Hall.
Lasnik H (1976) Remarks on coreference. *Linguistic Analysis* **2**: 1–22.
Partee B (1984) Nominal and temporal anaphora. *Linguistics and Philosophy* **7**: 243–286.

Further Reading

Chierchia G (1992) Anaphora and dynamic binding. *Linguistics and Philosophy* **15**(2): 111–183.
Chomsky N (1981) *Lectures on Government and Binding*. Dordrecht, Netherlands: Foris.
Jacobson P (2000) Paycheck pronouns, Bach–Peters sentences, and variable-free semantics. *Natural Language Semantics* **8**: 77–155.
Lewis D (1979) Scorekeeping in a language game. *Journal of Philosophical Logic* **8**: 339–359.
Reinhart T (1983) *Anaphora and Semantic Interpretation*. Chicago, IL: University of Chicago Press.
Roberts C (1989) Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy* **12**: 683–721.
Van der Sandt R (1992) Presupposition projection as anaphora resolution. *Journal of Semantics* **9**: 333–377.

Articulation: Dynamic Models of Motor Planning

Advanced article

Kevin G Munhall, Queen's University, Kingston, Ontario, Canada

CONTENTS

The complex muscular system of the human vocal tract changes states rapidly and flexibly during fluent speech. The planning and control system responsible for this remarkable behavior must coordinate the spatial and temporal aspects of movement for a large number of independent articulators. Current models suggest that detailed representation of vocal tract movements and the acoustic consequences of those movements is required.

INTRODUCTION

All human cultures communicate primarily by using spoken language, and the children of these cultures acquire a remarkable level of articulatory skill by their second birthday. Although talking may thus be the most common and natural of human activities, it is still very much a scientific mystery. The linguistic, cognitive, motoric and neural mechanisms responsible for talking have remained beyond the grasp of a multidisciplinary research effort. As Gallistel (1999) has stated, there appears to be an inverse relationship between the apparent ease of performing an activity and the computational apparatus required to account for the behavior.

It is now clear that a complete analytic understanding of the production of speech will require the solution of a number of general motor control problems, as well as some problems that are relatively unique to oral language production. These problems include accounting for the mapping between a symbolic linguistic representation and a biomechanical system, the learning and production of sound categories, the coordination of a large number of movement systems so that the temporal and spatial requirements of articulation are met, and the accomplishment of this complex movement in real time. This article briefly summarizes articulatory phonetics and then considers current views of the way in which the nervous system solves the speech motor control problem.

ARTICULATORY PHONETICS

Articulatory phonetics is traditionally defined as the study of the way in which the sounds of a language are physically produced. This includes a description of vocal tract acoustics and the key aspects of sound production for consonants and vowels. Briefly, speech production involves the generation of sound (e.g., by the vocal folds) and its filtering by the acoustic properties of the vocal tract (size, shape, wall characteristics, etc.). This separation of source and filter has been the standard model of speech research since the middle of the twentieth century (Chiba and Kajiyama, 1941/1958; Fant, 1960), but the precise details of the process are still being elucidated.

SOURCE CHARACTERISTICS

During speech there are a number of different sources of sound. The primary source of sound is the vibration of the vocal folds. Through the activity of a complex muscle framework, the frequency, amplitude and mode of vocal fold vibration can be controlled. The muscular control of voicing has been reviewed elsewhere (Honda, 1995), as has the laryngeal anatomy and physiology (Titze, 1994). These attributes of laryngeal behavior produce the perceived pitch, loudness, and quality of voicing, and are used to signal segmental, prosodic, and paralinguistic information.

Our understanding of this subsystem of speech has depended on converging evidence from different experimental techniques and approaches. Electromyographic studies of humans during voicing (Honda, 1995), animal and cadaver studies of vibration (Kakita *et al*., 1981), photographic and

imaging research during speech (Hirose, 1988), and sophisticated biomechanical models (Titze, 1973, 1974) have revealed a vocal articulation system the behavior of which results from tuning the patterns of muscular tension to control the tissue biomechanics of the vocal folds themselves.

In addition to vocal fold vibration, sound is produced at a number of locations in the vocal tract above the larynx. This mainly occurs during consonant production (e.g., stop consonants and fricatives), and it involves sound being created through the manipulation of the air flowing through the oral cavity. The sudden release of air during the production of stop consonants produces an acoustic burst that varies depending on the air pressure, position of constriction in the vocal tract, and release movements. Frication is generally produced by air rushing though a small opening.

The characteristics of fricative sounds are similarly determined by the air pressure and position of constriction, but also by the details of the small channel through which air rushes, and whether or not the air jet strikes an obstacle such as the teeth. The articulatory conditions that are required for producing these sounds have been studied using a range of different techniques, such as electropalatography (Hardcastle and Hewlett, 1999), physical and computational models (Shaddle, 1985), electromyography, and kinematic measurements of articulator movement (Löfqvist and Gracco, 1997).

FILTER CHARACTERISTICS

The sounds that are produced in the vocal tract are modified or filtered by the acoustic properties of that structure. These filtering characteristics are known as the transfer function of the vocal tract, and include the cross-sectional area, length, and wall characteristics. Although there has been a general understanding of both vowel and consonant production for more than 100 years, the details of the filtering aspect of articulation are still being revealed. This slow progress is partly due to the relative inaccessibility of the speech articulators, as most of the determinants of the vocal transfer function are internal and not directly accessible to measurement without invasive recording techniques.

According to the simple textbook view, vowels are distinguished by a small number of articulation parameters (e.g., tongue height, front–back location of the constriction, lip roundedness) and consonants by a different but similarly small number of articulation parameters (e.g., place of articulation, manner of articulation, voicing). However, this textbook view significantly under-represents the full complexity of articulation. At least two aspects of this simplification should be noted. First, vowel and consonant acoustics are determined by the shape of the entire vocal tract. The gross details of this have been known since the first X-ray studies of speech (for a summary of these studies see Moll, 1960). Figure 1 shows two-dimensional magnetic resonance imaging (MRI) views of the three English vowels /i/, /u/, and /a/. As can be seen, the surface of the tongue, lips, jaw, pharynx, velum, and larynx changes position for each vowel to produce a unique transfer function. New developments in imaging of the vocal tract have revealed intricate three-dimensional shape changes associated with different phonemes (for a review of progress in imaging the shape of the vocal tract, see Munhall, 2001). Secondly, speech sound production involves movements, not a sequence of vocal tract configurations. According to the textbook

Figure 1. Mid-sagittal view of the vocal tract during vowel production. The vocal tract shapes for the vowels (a) /i/, (b) /u/, and (c) /a/. Figure provided by M. Tiede.

view, the position of a few key articulators at one point in time is used to describe the sound. In contrast, what really characterizes fluent speech is the seamless flow of gestures of the entire vocal tract.

Both of these factors contribute to the difficulty in mapping between traditional phonetic descriptions of speech and the reality of articulation. During speech the vocal tract is continuously changing in many dimensions, with the articulations for a given sound varying according to the context. However, the traditional phonetic description depicts speech as a series of static, invariant postures where only a few key dimensions are essential. An account of articulatory phonetics that recognizes its dynamic nature must depict speech as a motor skill.

Timing and Coordination

In order to control the vocal tract during speech, the nervous system must pattern the muscle activations of a large number of articulators so that the resultant movement paths shape the vocal tract properly. Because of the large number of degrees of freedom of the vocal tract, the spatial paths in articulation can be complicated. Figure 2 shows a mid-sagittal view of movement paths of the jaw and three points on the tongue. The speaker is uttering the following phrase: 'the oleander is a flower'. As can be inferred from the changes in articulator positions, the vocal tract shape is changing continuously.

The timing of muscle activations required to produce speech such as this also requires precise

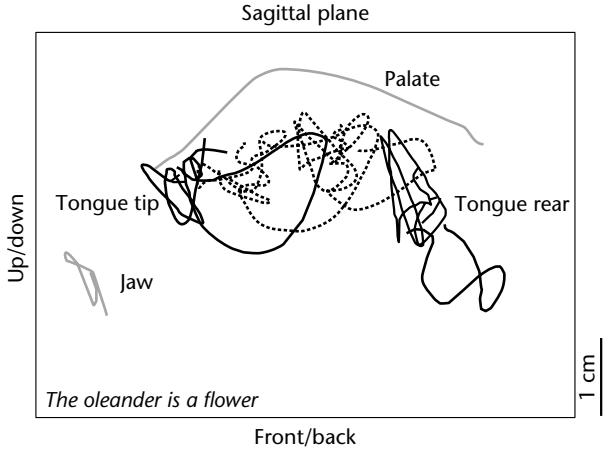


Figure 2. The movement paths of the jaw and three positions on the tongue during production of the utterance 'the oleander is a flower'. Figure provided by V. Gracco.

control, and the muscle activation generates patterns at many temporal scales. During the production of consonants and consonant clusters, coordinated movements between articulators can occur within tens of milliseconds, and even within a single articulator such as the tongue tip, successive movements can be coordinated to occur within 50 ms (Kent and Moll, 1975). At the syllable level, there is evidence of temporal organization that spans hundreds of milliseconds (for a recent example see Shaiman, 2001). At longer temporal spans there is a tradition of characterizing languages based on their rhythm class. For example, Japanese is classified as a mora-timed language in which strings of equal-duration mora are considered to be produced. However, Warren and Arai (2001) have recently published a critical review of mora timing.

Explanations for this complex behavior have been hindered by high levels of variance in the measured data. For both temporal and spatial descriptions of speech, the consistent impression is one of coordination but also of variability from one trial to another. The potential sources of such variance are manifold, including biomechanical, social, linguistic, and environmental contexts. Speaking rate, lexical and emphatic stress, and syllabification can vary from one repetition to another. In natural conversation the talker's mood, intent and social attention, as well as the conceptual and emotional meaning of a message, are all transmitted in parallel with phonetic information by subtle differences in the way in which words are spoken. These subtle differences are produced by systematic modifications to the movement timing and paths of the oral articulators. This variability that characterizes natural speech suggests that the control system must be programming articulation in real time and balancing many simultaneous requirements.

MODELING SPEECH PRODUCTION

In order to understand such complex behavioral organization as that of the speech motor control system, both experiments and sophisticated models are required. A common strategy is to identify the significant behavioral patterns and then decide how best to account for those patterns. Part of the difficulty that faces researchers is knowing what parts of behavior should be attributable to what parts of the planning and production process. Language planning involves many different stages and processes (e.g., semantic, syntactic, lexical, phonological; for one account of these stages, see Levelt, 1989), and speech motor control involves

additional stages of planning (Munhall *et al*., 2000; Perkell *et al*., 2000). Any single observation about articulation could thus be attributed to many parts of this complex chain of events. For example, context sensitivity or coarticulation in speech has been attributed to serial planning processes that take into account upcoming phonemes in order to plan smooth trajectories (Henke, 1966). However, David Ostry and his colleagues (Ostry *et al*., 1996; Perrier *et al*., 1996) have demonstrated that kinematic patterns due to speaking rate changes and coarticulatory context could, in principle, fall out from the physiology of the articulators. In their simulations, apparent context effects can be reproduced by modeling realistic physiological structures such as the jaw. The input to their jaw model does not take context into account, but the output is a smooth trajectory in which the articulator and muscle dynamics significantly shape the movement form to reproduce coarticulation effects.

The success of the research by Ostry and his colleagues depended on their dynamic modeling of the jaw and its muscle system. The term 'dynamics' is used here in a number of senses. First, it is used to refer to the general study of movement. Secondly, the term is used to refer to a subcomponent of the general study of movement that deals with forces (as opposed to kinematics, which deals with spatial descriptions of movement). Finally, it is used to refer to the behavior of complex nonlinear systems (so-called dynamical systems). A dynamic model of the vocal tract in all of these senses is necessary to test linguistic models of timing, phonological organization, etc. This model would represent the 'plant' or the physiology and biomechanics that are actually being controlled. It would also represent the complex interactions of control signals, biomechanics, and reflex coupling that are a part of stable movement. Unfortunately, such a complete model is currently beyond the scope of the field. To model the plant alone accurately is a daunting task. Histological, kinematic, dynamic (force), and myoelectric parameters are required for all of the musculoskeletal systems in the vocal tract, as well as good estimates of the three-dimensional morphology of the articulators. Many of these parameters are not available at present, and some are difficult if not impossible to acquire using current technology.

Models of individual articulators such as the vocal folds (Titze, 1973, 1974, 1994), the tongue (Wilhelms-Tricarico, 1995), the jaw (Laboissièrre *et al*., 1996), the tongue–jaw complex (Sanguineti *et al*., 1998; Dang and Honda, in press) and the face (Lucero and Munhall, 1999; Pitermann and Munhall, 2001) have been constructed with simulations of the tissue biomechanics (e.g., mass, stress/strain characteristics), muscle physiology (e.g., length tension characteristics of force development) and morphology. Although these simulations all involve some simplification of the articulator being studied (e.g., simplified geometry or finite element approximation of tissue structure), the models have played an important role in advancing the study of speech motor control. They have revealed the extent to which articulatory phonetics can be influenced by the physiological and biomechanical substrate of the vocal tract (Shiller *et al*., 1999). Furthermore, they have indicated that issues such as force, stability, and feedback processing are central problems for articulatory phonetics.

More global models of articulation have been proposed, which involve even more extensive simplifications of the oral motor mechanisms than the individual articulator models. However, these models address more complex phonetic output by controlling the vocal tract as a whole over time.

Saltzman's task dynamic model (Saltzman, 1986; Saltzman and Munhall, 1989) controls a set of stylized model articulators in a mid-sagittal vocal tract. The model describes the behavior of a set of 'tract variables' that refer to constrictions along the longitudinal axis of the vocal tract. The behavior of these tract variables is determined by a number of influences on the model articulators associated with the constrictions. For example, the behavior of the lip aperture tract variable is determined by the manner in which the lips and jaw are coordinated, as well as by the sequence of movements involving those articulators. The strengths of this model include its explanation of the coupling between articulators during articulation, and its focus on a task-level frame of reference in motor planning. The model can account for a number of important articulatory phenomena, such as some aspects of coarticulation and compensation following unexpected perturbations. When loads are applied to the lips or jaw, compensation from other articulators to achieve a goal at the task level is observed. For example, loads applied to the lips result in compensatory changes in the lips and jaw (Abbs and Gracco, 1984; Gracco and Abbs, 1985, 1988) as well as in the larynx (Munhall *et al*., 1994). The task dynamics model reproduces these findings and others. However, the model's vocal tract and articulator degrees of freedom are greatly simplified, and the articulators themselves are massless, with no physiology or biomechanics represented.

Another global model has been proposed by Guenther (Guenther, 1994; Guenther *et al*., 1998). The DIrections in auditory space to Velocities in Articulator space (DIVA) model represents a series of mappings between different frames of reference in speech motor control (e.g., auditory to articulatory mapping, articulatory to auditory mapping, phoneme to auditory mapping). These mappings are learned using neural networks trained in a 'babbling' phase in which random settings are applied to the articulator system and the acoustic consequences are used as feedback (for a similar approach, see Bailly *et al*., 1997). Like the task dynamics model, a series of key speech motor control phenomena are accurately replicated by the model. For example, when subjects are asked to speak with a bite block between their teeth, they are quickly able to produce normal vowel acoustics, even though the configuration of individual articulators may be unique. The DIVA model can reproduce this finding.

The DIVA model has a number of important features. It includes mappings between perceptual and motor systems as well as kinesthetic and acoustic feedback representations of speech. It involves learning, and it has been extended into a developmental model (Callan *et al*., 2000). Finally, it is consistent with current views on the detailed representations required for motor control (Wolpert and Kawato, 1998; Wolpert *et al*., 1998). Perkell *et al*. (2000) have described these detailed representations in DIVA. However, like the task dynamics model, the DIVA model is not controlling a realistic vocal tract. The work to date has used Maeda's articulatory synthesizer (Maeda, 1990), which is a two-dimensional mid-sagittal representation with stylized articulators that have no biomechanics or physiology.

An ambitious statistical approach has been developed by researchers at Advance Telecommunications Research (ATR) laboratories (Kawato, 1989; Hirayama *et al*., 1993, 1994; Vatikiotis-Bateson *et al*., 1993, 1994; Wada *et al*., 1995; Yehia *et al*., 1998, 2000). This model differs in many ways from existing models of speech motor control (Saltzman, 1986; Saltzman and Munhall, 1989; Guenther, 1994) in that it attempts to take into account the dynamics of the actual articulatory system and the fact that all parameters are derived from observed data. Briefly, the current model attempts to simulate the complete process from linguistic primitives to the dynamics and kinematics of the articulatory system, and finally to the output acoustics. The computation of motor commands (and thus movement trajectories) is constrained by global settings of the speech apparatus that last for a number of segments (e.g., for speaking rate). Linguistic units are provided as input to the speech-planning system and converted to a sequence of targets in vocal tract, acoustic, and articulator planning spaces.

The actual computation of the trajectories requires a knowledge of the physiological and biomechanical structures of the vocal tract in order to deal with the complex and nonlinear mappings between muscle activity and force, muscle force and movement, movement and speech acoustics, etc. It has been postulated that the control system builds up this knowledge though experience during speech development. In other words, the nervous system learns dynamic models of its own motor system as an aid to controlling the complex dynamics of articulation. The ATR group was among the first to recognize the importance of these 'internal models' of the articulator system as a major component of the control system in speech. An internal model is a neural representation of the spatial (kinematic), force (dynamic), and feedback (e.g., acoustic, proprioceptive) characteristics of movements that could be used by the nervous system to predict movement outcome (Miall and Wolpert, 1996; Kawato, 1999).

Like Guenther's model, the structure of the ATR model is composed of a series of learned mappings, and the plausibility of this statistical mapping has been demonstrated for a number of different dimensions and articulator systems (Yehia *et al*., 1998). This is a complex modeling program, but one that is addressing the full complexity of the articulation process (for an overview of the approach, see Munhall *et al*., 2000).

CONCLUSION

The speech motor system is one of the most complex human movement systems. The study of its control requires detailed models of the articulators in action, including simulations of their physiology and biomechanics. Surprisingly, recent studies suggest that the nervous system itself requires models of its own articulator system to accomplish trajectory planning. Research progress in this area will depend on the integration of detailed individual articulator models into global models of the full articulation process.

References

Abbs JH and Gracco VL (1984) Control of complex motor gestures: orofacial muscle responses to load

perturbations of the lip during speech. *Journal of Neurophysiology* **51**: 705–723.

Bailly G, Laboissière R and Galván A (1997) Learning to speak: speech production and sensori-motor representations. In: Morasso P and Sanguineti V (eds) *Self-Organization, Computational Maps and Motor Control*, pp. 593–615. Amsterdam, Netherlands: Elsevier.

Callan DE, Kent RD, Guenther FH and Vorperian HK (2000) An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language and Hearing Research* **43**: 721–736.

Chiba T and Kajiyama M (1941/1958) *The Vowel: its Nature and Structure*. Tokyo, Japan: Phonetic Society of Japan.

Dang J and Honda K (in press) A physiological model of a dynamic vocal tract for speech production. *Journal of the Acoustical Society of Japan*.

Fant G (1960) *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Gallistel CR (1999) Coordinate transformations in the genesis of directed action. In: Bly BM and Rumelhart DE (eds) *Cognitive Science*, pp. 1–42. New York, NY: Academic Press.

Gracco VL and Abbs JH (1985) Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes. *Journal of Neurophysiology* **54**: 418–432.

Gracco VL and Abbs JH (1988) Central patterning of speech movements. *Experimental Brain Research* **71**: 515–526.

Guenther F (1994) A neural network model of speech acquisition and motor equivalent production. *Biological Cybernetics* **72**: 43–53.

Guenther FH, Hampson M and Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* **105**: 611–633.

Hardcastle WJ and Hewlett N (eds) (1999) *Coarticulation: Theory, Data and Techniques*. Cambridge, UK: Cambridge University Press.

Henke W (1966) *Dynamic Articulatory Models of Speech Production Using Computer Simulation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Hirayama M, Vatikiotis-Bateson E and Kawato M (1993) Physiologically based speech synthesis using neural networks. *Institute of Electronics, Information and Communication Engineers (IEICE) Transactions* **E76-A**: 1898–1910.

Hirayama M, Vatikiotis-Bateson E and Kawato M (1994) Inverse dynamics of speech motor control. In: Hanson SJ, Cowan JD and Giles CL (eds) *Advances in Neural Information Processing Systems*, vol. 6, pp. 1043–1050. San Mateo, CA: Morgan Kaufmann.

Hirose H (1988) High-speed digital imaging of vocal fold vibration. *Acta Oto-Laryngologica* **458**: 151–153.

Honda K (1995) Laryngeal and extra-laryngeal mechanisms of F0 control. In: Bell-Berti F and Raphael LJ (eds) *Producing Speech: Contemporary Issues*, pp. 215–232. New York, NY: American Institute of Physics.

Kakita Y, Hirano M and Ohmaru K (1981) Physical properties of the vocal fold tissue: measurements on excised larynges. In: Stevens K and Hirano M (eds) *Vocal Fold Physiology*, pp. 377–397. Tokyo, Japan: University of Tokyo Press.

Kawato M (1989) Motor theory of speech perception revisited from the minimum torque-change neural network model. In: *Eighth Symposium on Future Electron Devices*, pp. 141–150. Tokyo.

Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinions in Neurobiology* **9**: 718–727.

Kent R and Moll K (1975) Articulatory timing in selected consonant sequences. *Brain and Language* **2**: 304–323.

Laboissière R, Ostry D and Feldman A (1996) Control of multi-muscle systems: human jaw and hyoid movements. *Biological Cybernetics* **74**: 373–384.

Levelt WJM (1989) *Speaking: from Intention to Articulation*. Cambridge, MA: MIT Press.

Löfqvist A and Gracco VL (1997) Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language and Hearing Research* **40**: 877–893.

Lucero JC and Munhall KG (1999) A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America* **106**: 2834–2842.

Maeda S (1990) Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: Hardcastle WJ and Marchal A (eds) *Speech Production and Speech Modeling*, pp. 131–149. Boston, MA: Kluwer Academic.

Miall RC and Wolpert DM (1996) Forward models for physiological motor control. *Neural Networks* **9**: 1265–1279.

Moll K (1960) Cinefluorographic techniques in speech research. *Journal of Speech and Hearing Research* **3**: 227–241.

Munhall KG (2001) Functional imaging during speech production. *Acta Psychologica* **107**: 95–117.

Munhall K, Löfqvist A and Kelso JAS (1994) Lip–larynx coordination in speech: effects of mechanical perturbations to the lower lip. *Journal of the Acoustical Society of America* **96**: 3605–3616.

Munhall KG, Kawato M and Vatikiotis-Bateson E (2000) Coarticulation and physical models of speech production. In: Broe M and Pierrehumbert J (eds) *Papers in Laboratory Phonology. V. Acquisition and the Lexicon*, pp. 9–28. Cambridge, UK: Cambridge University Press.

Ostry D, Gribble P and Gracco V (1996) Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? *Journal of Neuroscience* **16**: 1570–1579.

Perkell JS, Guenther FH, Lane H *et al.* (2000) A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics* **28**: 233–272.

Perrier P, Ostry DJ and Laboissière R (1996) The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech and Hearing Research* **39**: 365–378.

Pitermann M and Munhall KG (2001) An inverse dynamics approach to face animation. *Journal of the Acoustical Society of America* **110**: 1570–1580.

Saltzman EL (1986) Task dynamic coordination of the speech articulators: a preliminary model. Generation and modulation of action patterns. In: Heuer H and Fromm C (eds) *Experimental Brain Research*, series 15, pp. 129–144. New York, NY: Springer-Verlag.

Saltzman EL and Munhall KG (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology* **1**: 333–382.

Sanguineti V, Laboissière R and Ostry DJ (1998) A dynamic biomechanical model for neural control of speech production. *Journal of the Acoustical Society of America* **103**: 1615–1627.

Shadle C (1985) *The acoustics of fricative consonants*. Research Laboratory of Electronics, Technical Report 506. Cambridge, MA: Massachusetts Institute of Technology.

Shaiman S (2001) Kinematics of compensatory vowel shortening: the effect of speaking rate and coda composition on intra- and inter-articulatory timing. *Journal of Phonetics* **20**: 89–107.

Shiller DM, Ostry DJ and Gribble PL (1999) Effects of gravitational load on jaw movement in speech. *Journal of Neuroscience* **19**: 9073–9080.

Titze IR (1973) The human vocal chords: a mathematical model. Part I. *Phonetica* **28**: 129–170.

Titze IR (1974) The human vocal cords: a mathematical model. Part II. *Phonetica* **29**: 1–21.

Titze IR (1994) *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

Vatikiotis-Bateson E, Hirayama M, Wada Y and Kawato M (1993) Generating articulator motion from muscle activity using artificial neural networks. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics (RILP)* **27**: 67–77.

Vatikiotis-Bateson E, Tiede M, Wada Y, Gracco V and Kawato M (1994) Phoneme extraction using via point estimation of real speech. In: *Proceedings of 3rd International Conference on Spoken Language Processing, ICSLP-94*, pp. 531–534. Yokohama, Japan.

Wada Y, Koike Y, Vatikiotis-Bateson E and Kawato M (1995) A computational theory for movement pattern recognition based on optimal movement pattern generation. *Biological Cybernetics* **73**: 15–25.

Warren N and Arai T (2001) Japanese mora-timing: a review. *Phonetica* **58**: 1–25.

Wilhelms-Tricarico R (1995) Physiological modeling of speech production: methods for modeling of soft-tissue articulators. *Journal of the Acoustical Society of America* **97**: 3085–3098.

Wolpert DM and Kawato M (1998) Multiple paired forward and inverse models for motor control. *Neural Networks* **11**: 1317–1329.

Wolpert DM, Miall C and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Sciences* **2**: 338–347.

Yehia HC, Rubin PE and Vatikiotis-Bateson E (1998) Quantitative association of vocal-tract and facial behavior. *Speech Communication* **26**: 23–44.

Yehia H, Kuratate T and Vatikiotis-Bateson E (2000) Facial animation and head motion driven by speech acoustics. In: Hoole P (ed.) *Fifth Seminar on Speech Production: Models and Data*, pp. 265–268. Munich, Germany.

Binding Theory

Intermediate article

Eric Reuland, Utrecht University, Utrecht, The Netherlands

CONTENTS

Languages contain elements such as pronouns and anaphors or reflexives, which lack lexical/semantic content and may or must depend on another element for their interpretation. These dependencies cannot always be established freely, but are subject to structural conditions.

INTRODUCTION

In any language, distinct expressions may be used to refer to the same object. For instance, English 'morning star' and 'evening star' both refer to the planet Venus. Such expressions are said to *co-refer*. Here, co-reference holds on the basis of an empirical fact discovered by astronomers. In other cases speakers' intentions suffice to establish co-reference. The pronominal 'he' can be used to refer to any object that is linguistically classified as masculine and singular, as in 'John's mother thought he was infallible'. Here, 'he' may refer to John but also to some other masculine individual. However, in 'no one believes he is infallible' there is no individual such that both 'no one' and 'he' refer to that individual. Yet under the most salient interpretation 'he' does depend on 'no one' for its interpretation: the interpretive dependency is linguistically encoded, and instantiates *binding*. Binding is subject to constraints which cannot be explained on the basis of logic alone. Rather, they provide us with a window into the computational principles underlying language.

THE BINDING CONDITIONS

The theory of A(rgument)-binding explains the interpretive dependencies between phrases in *argument positions*, or *A-positions*, briefly *arguments*. A-positions are positions for phrases with grammatical functions, such as subject, object, etc., to which a predicate assigns a semantic role (agent, patient, beneficiary, etc.), or of which a predicate governs the case (nominative, objective, etc.) In (1), for instance, all the nominal expressions are arguments in A-position:

The old baron was crossing the bridge at dusk with a ramshackle carriage. *The driver* was visibly tired. Suddenly, the carriage tipped over and *the man* fell into the swamp. When *he* had pulled *him/himself* out there came no end to *his* tall tales. (1)

Arguments can be dislocated, ending up in a non-A-position (by topicalization, question formation, etc.), as in (2); *t* indicates their canonical position:

a. *Him*, I never believed the baron to have pulled out *t*
b. *Which man* did he think *t* fell from the bridge
c. *Himself*, the driver pulled *t* out immediately (2)

Binding theory treats the italicized phrases as if they are in the position of *t* (see also below).

Arguments are classified as R-expressions, pronominals, or anaphors. If the head of a phrase has lexical features it is an *R-expression*. Thus 'the old baron', 'no one', 'everyone', 'which man', etc., are all R-expressions. R-expressions are interpretively independent. *Pronominals* ('I', 'you', 'he', etc.) are only specified for *person*, *gender*, and *number* (the ϕ-features). They may, but need not, depend on another argument for their interpretation and they can be accompanied by a pointing gesture, that is, used deictically. *Anaphors* are referentially defective nominal elements: they cannot be used deictically. (A word of caution: some of the literature uses *anaphor* for a class including pronominals, and *reflexive* for anaphor in the present sense.) Anaphors are generally interpreted by binding. Under certain conditions anaphors can, nevertheless, remain unbound (see below). Anaphors come in two types: *simplex anaphors* and *complex anaphors*.

Also reciprocals, such as *each other*, behave as anaphors, although their semantics is more complex (see Heim *et al*., 1991). Also, elements such as '(his/her) own', '(the) other', '(the) same' are in some sense anaphoric. Here, we discuss only simplex and complex anaphors.

Simplex anaphors are like pronominals that are underspecified for certain features. Quite generally, specifications for *number* and *gender* are lacking; a *person* specification may be lacking as well (as in Russian *s'eb'a*, (Mandarin) Chinese *ziji,* or Japanese *zibun*). English lacks simplex anaphors, but cross-linguistically they abound. Some well-studied examples are Dutch *zich*, Icelandic *sig*, Chinese *ziji*, and Japanese *zibun*. Their interpretation often corresponds to English 'himself'.

Complex anaphors generally consist of a pronominal or a simplex anaphor and some other element. These other elements may be of varied provenance. Some are historically intensifiers, and currently semantically virtually empty, such as English *self* in 'himself', Dutch *zelf* (*zichzelf*), and Icelandic *sjalfan* (*sjalfan sig*). Many languages use body-part reflexives. These are based on an element that occurs independently as a nominal head designating a part of the body such as 'head', 'bones'; but also, designations such as 'soul, spirit, reflection' are found (see Schladt, 2000). For instance, in Basque 'the father killed himself' is literally expressed as 'the father killed his head'. The form *bere burua* 'his head', which in this sentence means 'himself', is also used in 'he put the cap on his head'. Other languages double a pronominal form, as in Cachur (spoken in Daghestan) and Old Syriac (a Semitic language), or put a special marker on the verb as in Kannada, a Dravidian language (see Lust *et al*., 2000).

Anaphor binding and pronominal binding differ with respect to allowing 'split antecedents', as in 'John talked to the girls about themselves' and 'John talked to the girls about them'. The pronominal 'them' can refer to John (subject) and the girls (indirect object) together (a 'split' antecedent); the anaphor 'themselves' only allows 'the girls' as an antecedent. (Like many other linguistic tests, this one must be used with caution.)

If *a* binds *b*, *a* is the *antecedent* of *b*. Since binding relations cannot be directly read off linguistic expressions they are indicated by a system of *indexing*. Each argument is assigned an integer as its index: in practice, one uses subscripts such as *i*, *j*, *k*, etc. If *a* and *b* have identical subscripts they are *co-indexed*. Thus, in (a_i … b_i), *a* and *b* are co-indexed. Since indices are just linguistic markers, two expressions can still be assigned the same object in some outside world if they are not co-indexed ('morning star' and 'evening star' are not necessarily co-indexed). Binding without co-indexing is not possible, though. In order for *a* and *b* to be co-indexed, (3) must hold:

> *a* and *b* are nondistinct in features for person, number, and gender (3)

Nondistinctness, rather than identity of features, is required for co-indexing, since in many languages one anaphoric element can have masculine or feminine, singular or plural antecedents. Dutch *zich* and Icelandic *sig* are like that; but they are specified as third person, since they cannot have first- or second-person antecedents. In other languages (for instance, Slavic languages such as Russian) a person specification is also lacking, and we find one anaphoric form for all persons.

Binding relations can also be represented in a *logical syntax notation*. In such a notation, pronouns and anaphors are represented as variables; R-expressions, such as 'every old man', are analyzed containing a *determiner* ('every') and a *set expression* which consists of a *variable* (not overtly represented) and a *restriction* ('old man'). A sentence such as 'Every old man sleeps' can then be represented as 'for all individuals *x*, provided you pick your individuals from *x*'s that are old and that are men, it is also the case that *x* sleeps', or, in formula, $\forall x$ *((old (x) & man (x)) → sleeps (x))*. Often an informal notation is used, in which determiner and restriction stay together, as in $Every\ old\ man_x$ *(x sleeps)*, or $John_y$ *(y sleeps)* (= John sleeps). Here, 'Every old man' and 'John' bind their respective variables. Using this notation, A-binding can also be represented by variable binding. 'Translating' 'himself' in 'John saw himself' as the variable *x*, yields $John_x$ *(x saw x)* as a logical syntax representation. Similarly, in 'Every boy expected Mary to see him', 'him' can be translated as a variable bound by 'Every boy', as in the informal structure $Every\ boy_x$ *(x expected Mary to see x)*.

For binding to obtain, the binder must *c-command* the bindee in the syntactic structure. The standard definition is (4):

> *a* c-commands *b* if and only if *a* does not contain *b* and the first branching (or maximal) projection dominating *a* also dominates *b* (4)

Binding by a non c-commanding antecedent is impossible, as illustrated by the ungrammaticality of *$John_i$'s mother loves himself$_i$*. Putting both conditions together yields (5) as the standard condition on

binding ('iff' = 'if and only if'):

a binds *b* iff *a* and *b* are co-indexed and *a* c-commands *b* (5)

Outline of the Canonical Binding Theory

Over the last few decades, the English system has served as a standard model of the binding theory. Its canonical form is presented in Chomsky (1981), and elaborated in Chomsky (1986).

Anaphors and pronominals impose specific conditions on their binders. An anaphor must be locally bound: its antecedent must be sufficiently 'nearby'. A pronominal may be bound, but not locally: its antecedent must be sufficiently 'far away'. An R-expression cannot be bound at all; that is, it must be *free*. The notion of a *governing category* provides a measure for the relevant distance. As a first approximation, pronominals and anaphors are in complementary distribution. This is captured by the binding conditions in (6):

Binding Conditions
(A) An anaphor is bound in its governing category
(B) A pronominal is free in its governing category
(C) An R-expression is free (6)

Governing category is defined in (7):

γ is a governing category for α if and only if γ is the minimal category containing α, a governor of α, and a SUBJECT accessible to α (7)

The SUBJECT of a category is its most prominent nominal element (including the agreement features in finite clauses). This is illustrated in (8). Binding is indicated by italics; $[_{\text{GC-}\alpha}]$ stands for the *governing category* of α.

a. *John* expected $[_{\text{GC-himself/him}}$ the queen to invite him/*himself for a drink]
b. $[_{\text{GC}-\text{himself/him}}$ *John* expected $[_{\text{Clause}}$*himself*/ **him* to be able to invite the queen]] (8)

(8) exemplifies the *Specified Subject Condition* (SSC); the governing category is the domain of the nearest subject to α. For 'him/himself' this is 'the queen' in (8a) and *John* in (8b).

Finite clauses are governing categories for their major arguments, including their subjects. Noun phrases with possessive phrases are governing categories for any other object they contain. There is a class of exceptions to this generalization, such as (9):

a. They expected that pictures of themselves would be on sale.
b. Max expected the queen to invite Lucie and *himself* for a drink. (9)

Sentences of the type *John$_i$ saw a snake behind him$_i$/?? himself$_i$* are also difficult to fit in with the binding theory of (6) (Chomsky, 1981, 1986). (See below for further discussion and references.)

Binding and Co-reference

A pronominal can be bound by an antecedent, but also be *co-referential* with it. Hence, there is a potential ambiguity when the antecedent is referential (if the antecedent is not referential no ambiguity can arise). The ambiguity surfaces in the two interpretations of (10):

a. Only Lucie loves her husband
b. Binding: (Of all the women) Only Lucie has the property *x* loves [husband of (*x*)]
c. Co-reference: (Of all the women) Only Lucie has the property (*x* loves *a*) & *a* = *the individual* (*for instance*, *Jack*) *who happens to be Lucie's husband* (10)

Readings as in (10b) are also called *sloppy readings*, readings as in (10c) *strict readings*. Consider next:

a. **John* saw *him*
b. We all know what's wrong with Oscar. Everyone hates him. $^{\text{ok}}$Even *Oscar* hates *him* (11)

Binding Condition B rules out (11a), but what about (11b)? Suppose 'Oscar' and 'him' are *co-referential*, being assigned identical values directly, what prevents this in (11a)? Rule I (Reinhart, 1983) regulates binding and co-reference:

Rule I: Noun phrase (NP) A cannot co-refer with NP B if replacing A with C, C a variable A-bound by B, yields an indistinguishable interpretation. (12)

In (11b), under co-reference Oscar is ascribed the property of *Oscar*-hatred: (*x hates him* & him = Oscar). Under binding the property *(x hates x)* = *self*-hatred is ascribed. *Oscar*-hatred and *self*-hatred are different properties. The fragment is only felicitous for *Oscar*-hatred. So, Rule I allows the co-reference interpretation. In (11a) the two interpretations are indistinguishable, hence Rule I rules out co-reference, leaving no well-formed interpretation

since binding is ruled out by Condition B (as it is in (11b)).

Applying (12) requires comparing two different derivations. The processing difficulties this entails have been proposed to explain the fact that children master Condition B at a substantially later age than Condition A (the 'delayed Condition B effect').

Binding and Reflexivity

Many languages have been investigated with binding systems outside the scope of the canonical binding theory. Results indicate that binding uses a modular system, in which the canonical binding conditions arise as special cases. Only a selection of results can be discussed here.

Many languages have a three-way distinction between pronominals, simplex anaphors, and complex anaphors, instead of the two-way distinction covered by the canonical binding theory; some (the Scandinavian languages) even have a four-way system, not counting reflexive possessives.

These systems show that binding interacts with reflexive properties of predicates. A predicate is *reflexive* iff two of its arguments (e.g. subject and object) are co-indexed (Reinhart and Reuland, 1993). A predicate can be *marked* as reflexive by its intrinsic lexical properties (exemplified in English by 'behave' or 'wash'). In English such predicates allow the direct object to be absent; Dutch has a simplex anaphor (*zich*) in such cases. If a predicate is not intrinsically reflexive, argument co-indexing is not sufficient for a licit reflexive interpretation. Reflexivity must be licensed by an extrinsic reflexive marker operating on the lexical entry of the predicate. This is the effect of SELF-anaphors, that is, elements with English *self*, or Dutch *zelf*. For instance, nonreflexive *bewonderen* 'admire' requires the complex anaphor *zichzelf* 'himself', as in *George*$_i$ *bewondert zichzelf*$_i$/**zich*$_i$ 'George admires himself'. When the anaphor and its antecedent are not co-arguments, a complex anaphor is not required, since no reflexive predicate is formed, as in *Jan*$_i$ *voelde* [*zich*$_i$ *wegglijden*] 'John felt [himself slide away]', where the anaphor is a small clause subject.

Cross-linguistically, licensing may also involve clitics, pronoun doubling, body parts, verbal affixes, etc., with varying further syntactic and semantic effects.

Whether a SELF-anaphor must be locally bound is in part determined by its content, in part by its relation to a predicate. In Dutch, *zichzelf* must always be locally bound. In English, a SELF-anaphor is exempted from this requirement if it does not exhaustively occupy the argument position of a predicate. Compare the ungrammatical (8a) where it does, to (9b) where it does not. A discussion of bound versus exempt anaphors based on grammatical functions can be found in Pollard and Sag (1992).

In other languages, for instance Malayalam, the element licensing reflexivity need not be locally bound at all (Jayaseelan, 1997).

Local Binding of Pronominals

In many languages (including all Romance and Germanic languages except English), first- and second-person pronominals can be locally bound (as in French *Je me lave* 'I wash myself' or German *Du sahst dich im Spiegel* 'You saw yourself in the mirror'). An important factor allowing local binding of pronominals is underspecification. Benvéniste (1966) has shown that first- and second-person pronouns are not grammatically, but lexically, marked for number ('we' is not a plurality of 'I's). So, these pronominals are grammatically underspecified. This allows them to be locally bound, despite being true pronominals in all other respects.

Frisian (spoken in a northern province of The Netherlands) also fits this generalization. It has a two-way distinction: (1) a complex anaphor *himsels*, when the predicate is nonreflexive, i.e. where Dutch has *zichzelf*; (2) a pronominal (*him*/etc.) in all environments where Dutch has *zich*, violating Condition B for third person as well. However, Frisian *him*/etc. is also underspecified; not for number, but for Case. This is the reason local binding is allowed in all persons.

LONG-DISTANCE ANAPHORA

Many languages allow anaphoric elements with an antecedent beyond their governing category as defined in (7), or without any linguistic antecedent whatsoever. Icelandic has become a classical case (see Thráinsson, 1979), but there is also long-distance anaphora in English (see Reinhart and Reuland, 1993 for an overview and references). It is often claimed that long-distance anaphors are simplex (i.e. they consist of only one morpheme) and require a subject as their antecedent, but this is a rather rough characterization.

Icelandic *sig* requires an antecedent within an indicative clause, but if *sig* is in an infinitival clause its antecedent may be outside it. The same holds for the cognate forms of *sig* in other Scandinavian languages. Also, *sig* in a subjunctive clause may have a long-distance antecedent. Yet, there are differences

between subjunctives and infinitives. If the antecedency relation crosses a subjunctive, binding is not required: the antecedent need not c-command the anaphor, and the existence of a discourse antecedent (not linguistically expressed) may suffice:

> María var alltaf svo andstyggileg. ψegar Olafur$_j$ kaemi segδi hún sér$_{i/*j}$ áreidanlega aδ fara . . . (Thráinsson, 1991)
> Mary was always so nasty. When Olaf would come, she would certainly tell himself [the person whose thoughts are being presented – not Olaf] to leave [NB. (13) could not begin a story] (13)

Many languages admit anaphor binding which violates the SSC (see (8)). For instance, Russian allows binding across infinitival boundaries; Dutch allows binding into perception verb complements. Yet, in all these cases c-command must be respected.

There is evidence that *sig* in the subjunctive domain behaves like a pronominal instead of an anaphor. Such pronominal use of an anaphoric form is often called *logophoric*.

The term *logophor* was introduced by Hagège (1974) to characterize a paradigm of specialized pronominal elements in languages of the Niger-Congo group. Subsequently, this term has been generalized to all elements with the following characteristics (Clements, 1975, pp. 171–172):

1. logophoric pronouns are restricted to *reportive contexts* transmitting the words or thoughts of an individual or individuals other than the speaker/narrator;
2. the antecedent does not occur in the same reportive context as the logophoric pronoun;
3. the antecedent designates the individual or individuals whose words or thoughts are transmitted in the reported context in which the logophoric pronoun occurs.

Conditions (1) and (3) are not structural, but involve the discourse status of the antecedent. In the following Icelandic example these conditions are met and c-command is not necessary (in all other cases c-command and subject orientation are strictly enforced in Icelandic):

> Skoδun Jóns$_i$ er [aδ ζδ hafir svikiδ sig$_i$] . . . (Thráinsson (1991) opinion John's is that you have betrayed self (14)

Here, *Jón* holds the opinion expressed. (In (13) above, *ser* refers to the person whose thoughts are being presented.) If these conditions are not met, logophoric elements are infelicitous. The situation is more complex than the above quote indicates. In some languages logophoricity is restricted to verbs of saying, excluding thoughts; there may be special logophoric forms with respect to the adressee instead of the speaker, etc. (For more discussion of logophoricity the reader is referred to Sells (1987), and the extensive literature on Icelandic – see the Further Reading list for references.) Well-known further cases of elements that vary between a bound and a referential, logophoric, use are Japanese *zibun* and Chinese *ziji* (see Huang and Tang (1991) and Cole *et al*. (2000) for discussion and references).

Also, English allows a logophoric use of 'himself' (which is not mono-morphemic). Its sensitivity to the discourse status of the antecedent is illustrated by the contrast in (15). (15a) is presented from John's perspective, (15b) from Mary's:

> a. John$_i$ was going to get even with Mary. That picture of himself$_i$ in the paper would really annoy her, as would the other stunts he had planned
> b. *Mary was quite taken aback by the publicity John$_i$ was receiving. That picture of himself$_i$ in the paper had really annoyed her, and there was not much she could do about it (Pollard and Sag, 1992) (15)

It is an important result of these investigations that a systematic distinction exists between true long-distance binding and a logophoric use. A language may allow long-distance binding without admitting a logophoric use of anaphors; the converse holds as well.

CROSSOVER PHENOMENA AND PARASITIC GAPS

Crossover

Since binding is defined in terms of c-command and co-indexing, movement is expected to feed binding. (16) shows this with anaphor binding:

> a. — seemed to himself [John to have been incompetent]
> b. John$_i$ seemed to himself$_i$ [t_i to have been incompetent] (16)

The R-expression 'John', which cannot bind the anaphor 'himself' from its base position, can do so after *A-movement*. (16) involves an antecedent moving from one A-position to another.

Wh-movement, *topicalization*, *relativization*, etc., move a phrase to an *non-A- (= A′-) position*. The process of assigning scope to expressions such as 'everyone' has properties of covert A′-movement.

Pronominals can be bound by *wh*-phrases and by 'everyone':

a. Who$_i$ t_i complained that Mary damaged his$_i$ car?
b. Everyone$_i$ complained that Mary damaged his$_i$ car (17)

When A′-movement crosses a pronominal, two cases are to be considered. In (18) the pronominal c-commands the trace of the moved element:

a. He saw who?
b. *Who$_i$ did he$_i$ see t_i? (18)

(18) instantiates *strong crossover*. The *wh*-trace is an R-expression. 'He' binds the trace; this is a Condition C violation, which is strongly ungrammatical. (See Chomsky, 1982, p. 35 for an alternative account.)

In (19b) the pronoun does not c-command the base position of the *wh*-phrase that crossed over it:

a. His$_i$ mother loves who$_i$
b. $^{??}$who$_i$ does his$_i$ mother love t_i (19)

In (20) the object quantifier 'everyone' is assigned scope over the subject, resulting in a similar configuration:

a. His mother loves everyone
b. $^{??}$Every x_i [his$_i$ mother loves x_i] (20)

Both (19b) and (20b) are not felicitous, but also not as ungrammatical as (18b). Hence this phenomenon is called *weak crossover*. Reinhart (1983) argues that weak crossover violates the requirement that at surface structure the antecedent c-command the bound element from an A-position. It is easily seen that this requirement is violated in both (19) and (20). An alternative is based on the *bijection principle* (Koopman and Sportiche, 1982). This principle requires that a quantifier bind precisely one variable and that a variable be bound by precisely one quantifier. In (19b) and (20b) the quantifier ('who' or 'every') has to bind both the pronominal and its trace, which violates the biuniqueness requirement of the bijection principle.

Parasitic Gaps

A configuration that is reminiscent of weak crossover is found in so-called *parasitic gap* constructions (Chomsky, 1982 and references cited there). The construction is illustrated in (21). There are gaps in the object positions of both 'file' and 'reading' in (21a), and, similarly, in the object position of 'cook' and 'eat' in (21b), yet only one phrase has been extracted:

a. ?Which article$_i$ did John file t_i without reading e_i
b. ?This is the kind of food$_i$ you must cook t_i before you eat e_i (21)

Furthermore, only the direct object gaps of 'file' and 'cook' are in a position from which an element could have been moved. No elements can be moved out of a *without*-clause or a *before*-clause (these are so-called *islands for extraction*). This is brought out by the contrast resulting from filling one or the other position by a pronominal:

a. *Which article did John file it without reading *e*
b. Which article did John file *t* without reading it (22)

Filling the object position of 'file' with a pronominal, as in (22a), makes the other gap the only possible source for the *wh*-phrase, which results in full ungrammaticality. Hence, only one of the gaps can be a trace; the other one has a different status.

The licensing of the gaps in (23a, b), which correspond to the rightmost gaps in (21a, b), is parasitic on the existence of an extraction site resulting from A′-movement. If no such extraction takes place the structure is degenerate:

a. *John filed those articles$_i$ without reading e_i
b. *You must cook this food$_i$ before you eat e_i (23)

The main factors in the distribution of parasitic gaps are given in (24):

In the construction (A), where order is irrelevant and α, *t*, *e*, are co-indexed, in order for the parasitic gap *e* to be licensed (B) must hold :

(A) ... α ... *t* ... *e* ...
(B) i. α c-commands *t* and *e*
ii. *t* does not c-command *e* and conversely
iii. *t* is a variable (24)

From (24Bii) it follows, for instance, that a trace in subject position will not be able to license a

parasitic gap, as illustrated in (25):

*Which articles$_i$ t_i were filed without reading e_i (25)

The reason for the non c-command requirement in (24Bii) is that the 'parasitic gap' must be interpreted as a variable; i.e. it must be A′-bound. If it is bound by an operator such as the *wh*-phrase, the latter can assign it the content necessary for its interpretation. It must be licensed by an operator, since English does not generally allow pure null-pronominals. If the extraction site c-commands the parasitic gap the latter is bound by the trace, not by the operator. Hence, it will not be licensed. In (23) no phrase is able to bind the gap. In (25), on the other hand, the gap is bound by the trace. In neither case can the gap be interpreted. If the gap is assigned content by an operator, and interpreted as a variable, this violates the bijection principle, hence the relative marginality of (21).

CONNECTIVITY AND RECONSTRUCTION PHENOMENA

Can phrases satisfy or violate certain conditions via their traces, or not? The issue is called *connectivity*, and arises in binding, but also in the licensing of so-called *polarity* items. (26) illustrates polarity:

a. Nobody could see anything
b. *Anything$_i$, nobody could see t_i (26)

'Anything' must be in the domain of a negative (or, more generally, downward-entailing) quantifier such as 'nobody'. However, when 'anything' is fronted, this requirement cannot be satisfied via its trace.

For binding, however, (2) showed that pronominals, anaphors, and R-expressions behave as if they are in the position of their traces, when they are moved to an A′-position. That is, they *reconstruct*. Reconstruction is limited to overt A′-movement. A-movement does not show connectivity effects for binding, as illustrated in (27). Despite the fact that 'the claim …' originates from the trace position there is no Condition C violation in (27a), although there is in (27b):

a. The claim that *John* was asleep seems to *him* [t to be correct] (ok. *him* = *John*)
b. It seems to *him* that the claim that *John* was asleep is correct (*him* ≠ *John*) (27)

Where the moved phrase is complex and contains elements that are themselves in A-position more than just reconstruction may seem to be involved. (28a–c) are often taken to show that movement enlarges the domain of anaphor binding:

a. John said [that Bill liked that picture of himself best]
b. John wondered [[which picture of himself] Bill liked t best]
c. [which picture of himself] did John say [t' that Bill liked t best] (28)

In (28a) 'Bill' is the antecedent of 'himself', in accordance with Condition A. In (28b) and (28c) 'John' is also possible as antecedent of 'himself'. Moving 'which picture of himself' causes 'John' to be added as a possible antecedent. This follows if the governing category of 'himself' can be calculated from the source position of the moved phrase (t), its derived position, and also the intermediate t' position. (It is assumed that in (28c) 'which picture of himself' moves to its derived position via the C-projection of the embedded clause.)

Yet, sentences modeled on (28a) do not consistently disallow the matrix subject as an antecedent. Moreover, the latitude found here does not extend to pronominals and R-expressions, which can only be interpreted in their base position. Alternatively, then, reconstruction applies in all cases equally. The additional interpretations for 'himself' would still follow since, being the object of a picture noun, it is in an exempt position and can be interpreted logophorically.

References

Benveniste E (1966) *Problèmes de linguistique générale*. Paris: Gallimard.

Chomsky N (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky N (1982) *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

Chomsky N (1986) *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Clements GN (1975) The logophoric pronoun in Ewe: its role in discourse. *Journal of West African Languages* **10**: 141–177.

Cole P, Hermon G and Huang J (2000) *Long Distance Reflexives*. Syntax and Semantics, vol. 33. San Diego, CA: Academic Press.

Evans G (1980) Pronouns. *Linguistic Inquiry* **11**(2): 337–362.

Hagège C (1974) Les Pronoms logophoriques. *Bulletin de la Société de Linguistique de Paris* **69**: 287–310.

Heim I, Lasnik H and May R (1991) Reciprocity and plurality. *Linguistic Inquiry* **22**: 63–101.

Huang J and Tang J (1991) The local nature of long-distance reflexives in Chinese. In: Koster J and Reuland E (eds) *Long-distance Anaphora*, pp. 263–283. Cambridge, UK: Cambridge University Press.
Jayaseelan KA (1997) Anaphors as pronouns. *Studia Linguistica* **51**(2): 186–234.
Koopman H and Sportiche D (1982) Variables and the bijection principle. *Linguistic Review* **2**: 139–160.
Lust B, Wali K, Gair JW and Subbarao KV (2000) *Lexical Anaphors and Pronouns in Selected South Asian Languages*. Berlin: Mouton de Gruyter.
Pollard C and Sag I (1992) Anaphors in English and the scope of the binding theory. *Linguistic Inquiry* **23**: 261–305.
Reinhart T (1983) *Anaphora and Semantic Interpretation*. London: Croom Helm.
Reinhart T and Reuland E (1993) Reflexivity. *Linguistic Inquiry* **24**(4): 657–720.
Schladt M (2000) The typology and grammaticalization of reflexives. In: Frajzyngier Z and Curl T (eds) *Reflexives: Forms and Functions*, pp. 103–124. Amsterdam: Benjamins.
Sells P (1987) Aspects of logophoricity. *Linguistic Inquiry* **18**: 445–479.
Thráinsson H (1979) *On Complementation in Icelandic*. New York: Garland.
Thráinsson H (1991) Long-distance reflexives and the typology of NPs. In: Koster J and Reuland E (eds) *Long-distance Anaphora*, pp. 49–76. Cambridge, UK: Cambridge University Press.

Further Reading

Baltin M and Collins C (eds) (2000) *The Handbook of Contemporary Syntactic Theory*. Oxford, UK: Blackwell.
Barss A (1986) *Chains and Anaphoric Dependence*. Doctoral dissertation, MIT.
Bennis H, Pica P and Rooryck J (eds) (1997) *Atomism and Binding*. Dordrecht: Foris.
Burzio L (1991) The morphological basis of anaphora. *Journal of Linguistics* **27**: 81–105.
Chien Y-C and Wexler K (1991) Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition* **1**: 225–295.
Cole P, Hermon G and Sung L-M (1990) Principles and parameters of long-distance reflexives. *Linguistic Inquiry* **21**: 1–22.
Everaert M (1986) *The Syntax of Reflexivization*. Dordrecht: Foris.
Faltz LM (1977) *Reflexivization: A study in Universal Syntax*. Doctoral dissertation, University of California at Berkeley. Distributed by University Microfilm International, Ann Arbor, MI and London.
Fiengo R and May R (1994) *Indices and Identity*. Cambridge, MA: MIT Press.
Gelderen E van (2000) Bound pronouns and non-local anaphors: the case of earlier English. In: Frajzyngier Z and Curl T (eds) *Reflexives: Forms and Functions*, pp. 187–225. Amsterdam: Benjamins.
Gelderen E van (2000) *A History of English Reflexive Pronouns*. Amsterdam: Benjamins.
Grodzinsky Y and Reinhart T (1993) The innateness of binding and coreference. *Linguistic Inquiry* **24**: 69–101.
Heim I (1998) Anaphora and semantic interpretation: a reinterpretation of Reinhart's approach. In: Sauerland U and Percus O (eds) *The Interpretive Tract*, pp. 205–246. MIT Working Papers in Linguistics, vol. 25. Cambridge, MA: MIT.
Hornstein N (2000) *Move! A Minimalist Theory of Construal*. Oxford, UK: Blackwell.
Keller RE (1961) *German Dialects: Phonology and Morphology*. Manchester, UK: Manchester University Press.
Koopman H and Sportiche D (1989) Pronouns, logical variables, and logophoricity in Abe. *Linguistic Inquiry* **20**: 555–589.
Lasnik H (1989) *Essays on Anaphora*. Dordrecht: Reidel.
Lebeaux D (1988) *Language Acquisition and the Form of Grammar*. Doctoral dissertation, University of Massachusetts at Amherst.
Lidz J (1995) Morphological reflexive marking: evidence from Kannada. *Linguistic Inquiry* **26**(4): 705–710.
Pollard CJ and Sag IA (1994) *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
Reinhart T (2000) Strategies of anaphora resolution. In: Bennis H, Everaert M and Reuland E (eds) *Interface Strategies*, pp. 295–324. Amsterdam: Royal Academy of Arts and Sciences.
Reuland E (2001) Primitives of binding. *Linguistic Inquiry* **32**: 439–492.
Reuland E and Koster J (1991) Long-distance anaphora: an overview. In: Koster J and Reuland E (eds) *Long-distance Anaphora*, pp. 1–27. Cambridge, UK: Cambridge University Press.
Reuland E and Reinhart T (1995) Pronouns, anaphors and case. In: Haider H, Olsen S and Vikner S (eds) *Studies in Comparative Germanic Syntax*, pp. 241–269. Dordrecht: Kluwer.
Reuland E and Sigurjónsdóttir S (1997) Long-distance 'binding' in Icelandic: syntax or discourse? In: Bennis H, Pica P and Rooryck J (eds) *Atomism and Binding*, pp. 323–340. Dordrecht: Foris.
Safir K (1996) Semantic atoms of anaphora. *Natural Language and Linguistic Theory* **14**: 545–589.
Wexler K and Chien Y-C (1985) The development of lexical anaphors and pronouns. *Papers and Reports on Child Language Development* **24**. Stanford, CA: Stanford University.

Categorial Grammar and Formal Semantics

Intermediate article

Michael Moortgat, Utrecht Institute of Linguistics, Utrecht University, Utrecht, Netherlands

CONTENTS

Categorial grammar is a lexicalized grammar formalism based on logical type theory. A categorial lexicon assigns one or more types to the atomic elements of a language; the assembly of form and meaning is accounted for in terms of the rules of inference for these types, seen as formulae of a grammar logic. Cross-linguistic variation results from extending the invariant core of the grammar logic with facilities for structural reasoning.

INTRODUCTION

Categorial grammar, a linguistic framework with firm roots in type theory and constructive logic, is well represented in the logical and mathematical literature. This article puts the emphasis more on the categorial modeling of the cognitive abilities underlying the acquisition, use, and understanding of natural language. The sections below address two central questions. First of all, what are the invariants of grammatical composition, and how do they capture the uniformities of the form–meaning correspondence across languages? Secondly, how can we reconcile the idea of grammatical invariants with structural variation in the realization of the form–meaning correspondence?

The slogan 'parsing as deduction' concisely expresses the categorial perspective on these questions. A grammar, essentially, is given by an assignment of types to the elementary units in the lexicon. The type-forming operations have the status of logical connectives: determining whether an expression is well formed amounts to presenting a derivation, or proof, in the logic for these connectives. Natural-language expressions are signs, with a form and a meaning dimension. The categorial type language, consequently, is model-theoretically interpreted with respect to these two dimensions, and a derivation encodes an effective procedure for building up the structural organization of an expression, and for associating this structure with a recipe for meaning assembly.

The article is organized as follows. First, we focus on the form dimension of expressions. We identify the logical constants of the computational system, and study how the base logic for these constants can be extended with facilities for structural reasoning. Then, we see how the logical rules of inference for the type-forming operations can be read as instructions for meaning assembly, and how the structural rules determine which components of an expression can enter into the assembly process.

FORM

The Base Logic

Natural-language expressions are structured objects with a linear order and a hierarchical grouping. In categorial grammar, the traditional parts of speech assume the form of type formulae. The structure of these types mirrors the composition of the expressions they categorize. The set *Type* of type formulae is obtained as the closure of a small set *Atom* of *basic* types under a number of type-forming operations. Individual categorial grammars will differ with respect to the type-forming operations they employ. For our present purposes, the following clauses will be representative:

1. *Atom* is a subset of *Type*.
2. If A is a formula in *Type*, then $\Diamond A$ and $\Box A$ are too.
3. If A and B are formulae in *Type*, then $A \bullet B$, A/B, and $A\backslash B$ are too.

Basic types play a role similar to that of major constituents in phrase-structure grammar: they categorize expressions one can think of as

'complete'. Examples could be the types *np* for proper names, *s* for sentences, and *n* for common noun phrases. Languages can differ as to which basic type distinctions they make. The unary and binary operations provide a vocabulary to categorize expressions in terms of their constituent parts. Informally, a formula $A \bullet B$ categorizes an expression that can be decomposed into a constituent of type A followed by a constituent of type B. An expression with a fraction type A/B (or $B\backslash A$) is incomplete: it combines with an expression of type B on its right (or left, respectively) into an expression of type A. The unary type-forming operations are more recent additions to the categorial vocabulary. They can be thought of as features: an expression of type $\Box A$ issues a request for a feature to be checked; such an expression can be used as a regular A as soon as the $\Box$ feature is eliminated. The operation $\Diamond$ provides the means to perform the required feature-checking.

Frame semantics

To make this informal description precise, Došen (1992) and Kurtonina (1995) make use of *frame-based* models familiar from possible-world semantics for modal logics. For the categorial type language, a *frame* is a tuple $(W, R_\Diamond, R_\bullet)$. W is a nonempty set, the set of expressions. $R_\Diamond$ and $R_\bullet$ are binary and ternary relations over W, interpreting the unary and binary type-forming operations, respectively. One can think of $R_\bullet$ as the 'merge' relation: $R_\bullet xyz$ holds in case x is the composition of the parts y and z. Similarly, $R_\Diamond xy$ holds if the feature-checking relation connects y to x. One obtains a *model* by adding a *valuation* V assigning subsets of W to the atomic formulae. For complex types, the valuation respects the conditions below:

$x \in V(\Diamond A)$ iff there exists a y such that $R_\Diamond xy$ and $y \in V(A)$.
$x \in V(\Box A)$ iff for all y, $R_\Diamond yx$ implies $y \in V(A)$.
$x \in V(A \bullet B)$ iff there are y and z such that $y \in V(A)$, $z \in V(B)$ and $R_\bullet xyz$.
$x \in V(C/B)$ iff for all y and z, if $y \in V(B)$ and $R_\bullet zxy$, then $z \in V(C)$.
$x \in V(A\backslash C)$ iff for all y and z, if $y \in V(A)$ and $R_\bullet zyx$, then $z \in V(C)$.

Type computations, soundness, and completeness

On the proof-theoretic level, we are interested in a deductive system to perform type computations $A \to B$ ('type B is derivable from type A'). We want this system to be faithful to the interpretation of the type-forming operations, in the sense that $A \to B$ is provable iff $V(A) \subseteq V(B)$, for every frame F and valuation V. Such a system is 'sound' and 'complete'.

An axiomatization satisfying the soundness and completeness requirements starts with an identity axiom $A \to A$, and an inference rule allowing one to conclude $A \to C$ from premises $A \to B$ and $B \to C$. Semantically, these express the reflexivity and transitivity of the derivability relation. In addition, one has the inference rules below, establishing the relationship between the interpretation of $\Diamond$ and $\Box$, and between $\bullet$ and left and right division $\backslash$ and $/$. These patterns turn $(\Diamond, \Box)$, $(\bullet, /)$, and $(\bullet, \backslash)$ into what are known as 'residuated pairs' in algebra, or 'adjoint functors' in category theory:

(R0) $\Diamond A \to B$ if and only if $A \to \Box B$
(R1) $A \bullet B \to C$ if and only if $A \to C/B$
(R2) $A \bullet B \to C$ if and only if $B \to A\backslash C$

Elementary theorems

Let us look at some elementary theorems of the grammatical base logic. From the identity axioms of line 1 below, one obtains the Application schemata of line 2 in one step, using the residuation inferences in the 'if' direction; from Application, one derives the Lifting schemata of line 3, this time reasoning in the 'only if' direction:

1.	$A\backslash B \to A\backslash B$ (Ax)	$B/A \to B/A$ (Ax)
2.	$A \bullet (A\backslash B) \to B$ (R2 $\Leftarrow$)	$(B/A) \bullet A \to B$ (R1 $\Leftarrow$)
3.	$A \to B/(A\backslash B)$ (R1 $\Rightarrow$)	$A \to (B/A)\backslash B$ (R2 $\Rightarrow$)

The Application schemata are no doubt the most familiar laws of categorial combinatorics. In fact, the original categorial grammars of Ajdukiewicz and Bar-Hillel were restricted to Application. Using the Application schemata, one can 'lexicalize' the rules of a context-free phrase-structure grammar. Take the productions '$S \to NP\ VP$' and '$VP \to TV\ NP$' for the derivation of a Subject–Transitive-Verb–Object (SVO) pattern. In categorial terms, one types the Transitive Verb as $(np\backslash s)/np$, thus projecting the SVO pattern in two Application steps: rightward application consumes the Object, leftward application the Subject. The auxiliary label *VP* disappears; the complex type $np\backslash s$ expresses its combinatory role.

Instances of Lifting would be type transitions from *np* (the type assigned to simple proper names) to $s/(np\backslash s)$ or $((np\backslash s)/np)\backslash(np\backslash s)$. These lifted types are appropriate for noun phrases with a distribution restricted to the subject position, in the case of $s/(np\backslash s)$, or the direct-object position, in the case of $((np\backslash s)/np)\backslash(np\backslash s)$. What the derivability arrow says here is that any expression that is assigned the type *np* will be able to occur in the subject or object position, but that there can be

expressions with a restricted subject or object distribution, expressed through the higher-order types. One can think of case-marked pronouns, as Lambek (1958) pointed out. With $s/(np\backslash s)$ as the lexical type assignment for 'he'/'she', but $((np\backslash s)/np)/(np\backslash s)$ for 'him'/'her', we correctly rule out 'him irritates she' while allowing 'he irritates her'.

Elementary theorems for the unary type-forming operations are established below:

1. $\Box A \to \Box A$ (Ax) and $\Diamond A \to \Diamond A$ (Ax)
2. $\Diamond\Box A \to A$ (R0 $\Leftarrow$) and $A \to \Box\Diamond A$ (R0 $\Rightarrow$)

An illustration of the added expressivity of the unary operators can be found in Bernardi (2002), where they are used to control the distribution of polarity-sensitive items. Consider the contrast between 'nobody left yet', with the negative-polarity item 'yet', and '*somebody left yet'. In a type language with just the binary type-forming operations, both 'somebody' and 'nobody' would receive the subject type $s/(np\backslash s)$, and 'yet' the modifier type $(np\backslash s)\backslash(np\backslash s)$. Such type assignment is too crude to block the ungrammatical '*somebody left yet'. In the extended type language, the negative-polarity trigger 'nobody' can be assigned the type $s/\Box\Diamond(np\backslash s)$, whereas 'somebody' keeps the undecorated type $s/(np\backslash s)$. By typing the negative-polarity item 'yet' as $(np\backslash s)\backslash\Box\Diamond(np\backslash s)$ one expresses the fact that it requires a trigger such as 'nobody' to check the $\Box\Diamond$ decoration in its numerator subtype. For the derivation of the simple sentence 'nobody left' (with no polarity item to be checked), we rely on the fact that in the base logic, we have $s/\Box\Diamond(np\backslash s) \to s/(np\backslash s)$; i.e., the $\Box\Diamond$ decoration on argument subtypes can be simplified away, allowing the combination (in terms of the Application schema) of 'nobody' with a simple verb phrase 'left' of type $np\backslash s$.

Monotonicity properties

Apart from these theorems, the base logic has several derived rules of inference. With respect to the derivability relation, the operations $\Diamond$ and $\Box$ are order-preserving (isotone). The $\bullet$ operation is order-preserving in its two arguments; the division operations/and $\backslash$ are order-preserving in their numerator, and order-reversing (antitone) in their denominator argument.

$A \to B$ implies:

$\Diamond A \to \Diamond B$ and $\Box A \to \Box B$
$A/C \to B/C$ and $C\backslash A \to C\backslash B$
$C/B \to C/A$ and $B\backslash C \to A\backslash C$
$A \bullet C \to B \bullet C$ and $C \bullet A \to C \bullet B$

From a combinatorial point of view, these rules produce an infinite number of type transformations from some small inventory of 'primitive' ones. Consider the Lifting schema. From it, one obtains the transformations known as Value Raising (for example, lifting a determiner type np/n to $(s/(np\backslash s))/n$) and Argument Lowering (for example, lowering a third-order verb phrase type $(s/(np\backslash s))\backslash s$ to first-order $np\backslash s$).

Alternative presentations, and natural deduction

The categorial base logic allows many alternative axiomatizations, each serving its own function. The essential point is that the different presentations must find their justification in the model-theoretic interpretation of the connectives; i.e., one has to prove that they are equivalent syntaxes for performing valid type computations. In the Gentzen sequent calculus, one replaces the arrows $A \to B$ by statements $\Gamma \Rightarrow B$ ('structure Γ is of type B'). The antecedent Γ is built out of formulae by means of the structure-building operations $\langle\cdot\rangle$ and $(\cdot\circ\cdot)$, counterparts of the logical connectives $\Diamond$ and $\bullet$. The purpose of this presentation is to show that the transitivity rule (the Cut rule) can be eliminated. Every logical rule of inference in the Gentzen calculus introduces a connective either in the antecedent or in the succedent, so that backward-chaining, cut-free proof search immediately yields a decision procedure for categorial derivability, as shown in Lambek (1958) for the binary and in Moortgat (1996) for the unary connectives.

The derivational format of 'combinatory categorial grammar' (CCG) (e.g., Steedman, 2000b) is a Hilbert-style presentation. Functional Application here is taken as the basic, primitive schema for type combination. To the Application schema are added extra schemata, such as Lifting, also known as combinator T. The CCG format of derivations is related to the Gentzen style as the combinator presentation of intuitionistic logic is to its Gentzen presentation. The recursive generalization of the primitive type transformations under monotonicity is important for such 'combinatory' presentations of categorial derivability: without this generalization, one loses completeness.

In a third format, 'natural deduction' (ND), every type-forming connective has an introduction and an elimination rule. As a result, ND doesn't have the pleasant proof search properties of the Gentzen calculus, but it provides a perspicuous presentation of a derivation once it has been found. For this reason, ND is often used in linguistic discussion of categorial analyses. Also, ND is the most

$$\frac{\Gamma \vdash \Box A}{\langle \Gamma \rangle \vdash A}\ (\Box \mathrm{E}) \qquad \frac{\langle \Gamma \rangle \vdash A}{\Gamma \vdash \Box A}\ (\Box \mathrm{I})$$

$$\frac{\Gamma \vdash A}{\langle \Gamma \rangle \vdash \Diamond A}\ (\Diamond \mathrm{I}) \qquad \frac{\Delta \vdash \Diamond A \quad \Gamma[\langle A \rangle] \vdash B}{\Gamma[\Delta] \vdash B}\ (\Diamond \mathrm{E})$$

$$[/\mathrm{I}]\frac{\Gamma \circ B \vdash A}{\Gamma \vdash A/B} \qquad \frac{\Gamma \vdash A/B \quad \Delta \vdash B}{\Gamma \circ \Delta \vdash A}[/\mathrm{E}]$$

$$[\backslash \mathrm{I}]\frac{B \circ \Gamma \vdash A}{\Gamma \vdash B \backslash A} \qquad \frac{\Gamma \vdash B \quad \Delta \vdash B \backslash A}{\Gamma \circ \Delta \vdash A}[\backslash \mathrm{E}]$$

$$[\bullet \mathrm{I}]\frac{\Gamma \vdash A \quad \Delta \vdash B}{\Gamma \circ \Delta \vdash A \bullet B} \qquad \frac{\Delta \vdash A \bullet B \quad \Gamma[A \circ B] \vdash C}{\Gamma[\Delta] \vdash C}[\bullet \mathrm{E}]$$

Figure 1. Natural deduction. $\Gamma \vdash A$ stands for the deduction of a conclusion A from a configuration of assumptions Γ. Axioms are of the form $A \vdash A$. Antecedent structures are built from formulae with the structure-building operations $\langle \cdot \rangle$ and $(\cdot \circ \cdot)$. These are the structural counterparts of $\Diamond$ and $\bullet$, respectively, as the $\Diamond$ and $\bullet$ Introduction rules show.

transparant format to associate meaning assembly with a derivation, as we will see. Figure 1 shows the ND rules for the base logic, using the Gentzen sequent style, which is explicit about the structural configuration of the antecedent assumptions.

Multimodal generalization

One can straightforwardly generalize the base logic to a system where one has not just one single merge and feature-checking relation, but families of them. In terms of modal logic, this means moving from a unimodal to a multimodal system, with frames $(W, \{R^2_i\}_{i \in I}, \{R^3_j\}_{j \in J})$ where the different relations are kept apart by indexing them with a composition mode label. Similarly, in the formula language, we index the connectives for these composition modes. The concept of multiple composition modes is not unfamiliar. For the binary operations, one can think of a distinction between the structure of words (morphology) and the structure of phrases (syntax): one can give a categorial analysis of morphology and syntax in terms of $/$, $\bullet$, and $\backslash$, but still one will want to keep these grammatical levels distinct, say as $\bullet_w$ versus $\bullet_\phi$. For the unary connectives $\Diamond$ and $\Box$, multimodality makes it possible to distinguish a number of named features in the grammar, so that they can play different roles in controlling composition.

The multimodal perspective turns out to be particularly useful once we move beyond the base logic and consider its structural extensions, where one can then have interaction between different binary composition modes (between morphology and syntax, in the case of complement inheritance, for example), and between specific unary control features and binary composition operations. Such interaction principles are discussed below.

The Structural Module

The laws of the base logic do not depend on specific structural properties of the merge and feature-checking relations: the completeness condition does not impose any restrictions on the interpretation of $R_\bullet$ and $R_\Diamond$. In this sense, the base logic can be said to capture the invariants of grammatical composition. Although the base logic already has a rich deductive structure, the system also has its limitations. If an expression can occur in different structural configurations, one would like to relate these configurations. In the base logic, this cannot be done: type assignment is structurally rigid, in the sense that different structural environments will lead to different type assignments. To overcome the problem of structural rigidity, one extends the base logic with facilities for structural reasoning. Technically, such facilities have the status of non-logical axioms, or postulates. They can be introduced in a global or in a controlled fashion. We look at these in turn.

Global structural rules

The postulates below create a hierarchy of categorial systems: with the addition of structural options, the flexibility of type combination increases, but structural discrimination deteriorates.

($\mathrm{A_l}$) $(A \bullet B) \bullet C \rightarrow A \bullet (B \bullet C)$
($\mathrm{A_r}$) $A \bullet (B \bullet C) \rightarrow (A \bullet B) \bullet C$
(C) $A \bullet B \rightarrow B \bullet A$

The rebracketing postulates $\mathrm{A_l}$ and $\mathrm{A_r}$, added to the fragment of the base logic formed by $\bullet$, $/$, and $\backslash$, produce the system known as L, the associative calculus of Lambek (1958). The fragment of the base logic formed by $\bullet$, $/$, and $\backslash$ is known as NL: in Lambek (1961) this system was obtained by dropping the associativity postulates from L. Characteristic theorems of L are the type transitions below: the Geach laws $\mathrm{G_r}$ and $\mathrm{G_l}$, and the functional composition schemata (known as combinator **B** in CCG) of which $\mathrm{B_r}$ and $\mathrm{B_l}$ are the simplest forms.

($\mathrm{G_r}$) $A/B \rightarrow (A/C)/(B/C)$
($\mathrm{G_l}$) $B\backslash A \rightarrow (C\backslash B)\backslash(C\backslash A)$
($\mathrm{B_r}$) $(A/B) \bullet (B/C) \rightarrow A/C$
($\mathrm{B_l}$) $(C\backslash B) \bullet (B\backslash A) \rightarrow C\backslash A$

Adding the commutativity postulate to L produces LP (Lambek calculus with permutation), a

system coinciding with the multiplicative fragment of linear logic, which has a commutative product operation matched by a single linear implication. The distinction between left-incompleteness and right-incompleteness collapses in the presence of the commutativity postulate C.

Extending the base logic with facilities for structural reasoning has consequences for the interpretation of the type-forming operations (Došen, 1992; Kurtonina, 1995). An interpretation with respect to arbitrary frames is obviously not available anymore. Instead, each postulate introduces a corresponding frame constraint restricting the interpretation of the merge relation $R_\bullet$, and completeness is stated with respect to frames respecting the relevant constraints. A Commutativity postulate, for example, would impose the semantic constraint that for all $x, y, z \in W$, $R_\bullet xyz$ implies $R_\bullet xzy$. Similarly for the other postulates discussed. In the presence of such semantic constraints, it will often be the case that one can specialize the abstract relational interpretation to more concrete models. A good example is the system L with its associative composition relation $R_\bullet$. In this case, one can read $R_\bullet xyz$ as concatenation; i.e., $x = y \cdot z$. Pentus (1994) proves that L is indeed complete with respect to this concatenation interpretation.

Controlled structural reasoning

There are many natural-language phenomena that seem to require some of the flexibility offered by the postulates A_l, A_r, and C. Cases of nonconstituent coordination can be naturally handled with the possibilities for type combination that follow from the rebracketing postulates. Displacement phenomena are ubiquitous in natural language, and seem to require some form of commutativity. At the same time, it is clear that in a global form, these structural options overgenerate. Commutativity would entail that well-formedness is preserved under arbitrary changes in word order; free rebracketing makes constituent structure irrelevant for determining grammaticality.

To obtain controlled structural extensions of the base logic, various strategies have been pursued. In the rule-based approach of combinatory categorial grammar, one augments the Application–Lifting basis with structural combinators which, in an unconstrained form, would be overgenerating. One then imposes type restrictions on these extra combinators. In addition, the set of rule schemata (combinators) is kept finite, so that one can avoid the consequences of the recursive generalization of rules under monotonicity. The alternative is to exploit the intrinsic logical instruments for structural resource management offered by richer type systems with unary control features and multimodal interaction principles. To compare these two strategies, consider the following cases of extraction: 'what Alice found' and 'what Alice found there' (see Figure 2).

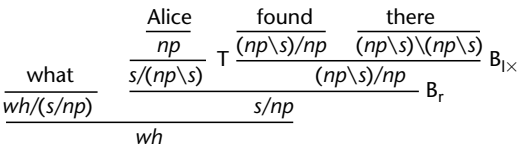


Figure 2. *wh*-extraction: combinator-style derivation. The clause body 'Alice found there' is assigned type s/np by means of the backwards crossed composition combinator $B_{l\times}$. The rule can apply because the cancelled $(np\backslash s)$ satisfies the type restriction on $B_{l\times}$.

In CCG, the peripheral case of extraction ('what Alice found') is derived from an assignment $wh/(s/np)$ to the *wh*-pronoun by lifting the type for 'Alice' to $s/(np\backslash s)$ which is then composed with the transitive verb type $(np\backslash s)/np$ for 'found' by means of B_r. To obtain the nonperipheral case of extraction ('what Alice found there'), one needs the combinator $B_{l\times}$, a form of composition which depends on the commutativity postulate. To avoid collapse into LP, one imposes a side condition on the rule, restricting the middle term B to certain verbal categories, in this case $(np\backslash s)$:

($B_{l\times}$) $(B/C) \bullet (B\backslash A) \rightarrow A/C$ where B is a predicate category

The $\Diamond$ and $\Box$ connectives make it possible to avoid extralogical type restrictions. The postulates P1 and P2 below implement a controlled form of rebracketing and reordering for formulae carrying the $\Diamond$ control feature (Moortgat, 1999). With a lexical type assignment $wh/(s/\Diamond\Box np)$ to the *wh*-pronoun, one obtains peripheral and medial extraction from right branches. Under this analysis, one does not attribute any associativity or commutativity to the $\bullet$ operation itself; displacement effects arise through the interaction of the merge operation with a gap hypothesis carrying the licensing $\Diamond$ feature. A derivation is given in Figure 3.

(P1) $(A \bullet B) \bullet \Diamond C \rightarrow (A \bullet \Diamond C) \bullet B$
(P2) $(A \bullet B) \bullet \Diamond C \rightarrow A \bullet (B \bullet \Diamond C)$

Generative Capacity and Computational Complexity

The modular view on grammatical invariants and structural variation invites a comparison between the categorial landscape and the Chomsky hierarchy. For a recent survey, see Buszkowski (1997).

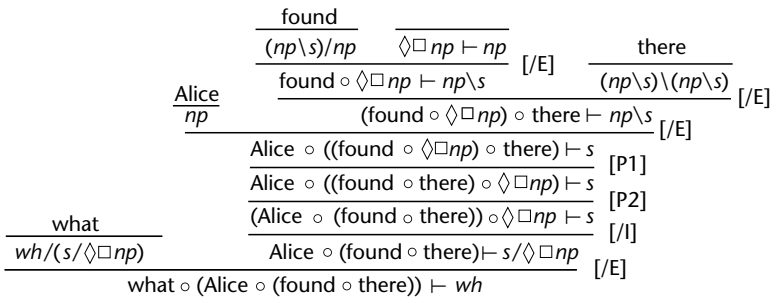


Figure 3. *wh*-extraction: ◊ control. The type assignment to the relativizer 'what' expresses the fact that the relative clause body is a sentence built with the help of a 'gap' hypothesis of type ◊□ *np*. The feature-marked hypothesis has to be withdrawn at the right periphery, but it is not selected in that position. It is related to the nonperipheral direct-object position within the relative clause body by virtue of the postulates P1 and P2. Once it has found the direct-object position, the licensing feature ◊ has done its work and can be cleaned up by the law ◊□ *np* ⊢ *np*. The 'gap' hypothesis is then used as a regular direct object with respect to the selecting verb 'found'.

The discovery in the 1980s of dependency patterns that cannot be adequately captured by context-free grammars has led to an interest in 'mildly context-sensitive' formalisms; i.e., systems with an expressivity beyond the context-free, but sufficiently restricted to have polynomial parsing algorithms. The Ajdukiewicz–Bar-Hillel grammars have long been known to be weakly equivalent to context-free grammars, hence to be too poor to serve as models of universal grammar. The same is true for the base logic described above. The correctness of Chomsky's conjecture that context-free equivalence extends to the Lambek calculus L was finally established in Pentus (1993). This result does not have a direct corollary for polynomial parsability, because the construction of a context-free grammar from an L grammar is of exponential complexity.

For the structural extensions of the base logic discussed above, the challenge is to identify appropriate constraints: it is clear that arbitrary combinator extensions, or structural rule packages, lead to excessive expressivity. But Vijay-Shanker and Weir (1994) show that an appropriately restricted version of CCG is weakly equivalent to the linear indexed grammars, hence polynomially parsable. In a similar spirit, Moot (2002) shows how, with appropriate restrictions on lexical assignments and structural postulates, one can carve out a class of multimodal categorial grammars equivalent to lexicalized tree adjoining grammars and inheriting the polynomial parsability of these systems. The general theory of ◊ and □ as control operators has been investigated in Kurtonina and Moortgat (1997). These authors establish a number of embedding theorems showing that the full logical space between the base logic and LP can be navigated in terms of the control connectives, both in the 'licensing' direction illustrated above (allowing structural inferences that would be unavailable without the control features) and in the 'constraining' sense (blocking structural options that would be licit in the absence of the control features).

More important than weak generative capacity are issues of strong capacity, which in the categorial tradition would mean the proof structures (or their lambda terms, discussed below) that produce a certain string. In this area, Tiede (2001) has obtained interesting results, showing that while the Lambek systems L and NL are weakly context-free, their expressivity in terms of strong capacity goes beyond that of context-free grammars.

Language Learning

Kanazawa (1998) has studied formal learning theory for categorial grammar within Gold's paradigm of identification in the limit on the basis of positive data. The focus is on classical categorial grammars, using only the Application rules, and on combinatory extensions with extra rule schemata. On the input side, Kanazawa considers learning both from strings and from function-argument structures. On the output side, the class of 'rigid' grammars (where the grammar assigns a unique type to each word) is compared to the class of *k*-valued grammars (where at most *k* types are assigned to a lexical item). It is a matter of dispute whether Gold's very abstract formulation of the learning problem is directly relevant for first-language acquisition. An alternative, purely inductive, approach – learning a subclass of the shallow context-free languages – is presented in Adriaans (2001).

The discussion above suggests some directions for further research in this area. First of all, one would like to obtain learnability results for classes of Lambek-style categorial grammars, where the learner has access to both the Elimination rules and the Introduction rules for the type-forming operators. Secondly, one would like to go beyond systems with a 'hard-wired' structural component, in order to investigate the learnability effects of different choices of structural packages, in combination with an invariant base logic. The work of Foret (2001) is promising in this respect: she mixes unification and substitution with Lambek-style deduction, so suggesting modulation of learnability questions in terms of different structural postulates.

Finally, the role of semantic information in learning needs further investigation. The challenge here is to find a level of informativity that would be realistic in the setting of first-language acquisition.

MEANING ASSEMBLY

Categorial grammar adheres to the truth-conditional theory of semantics: the interpretation process establishes a systematic relationship between linguistic expressions and states of affairs in the world in such a way that specifying the meaning of a sentence comes down to giving its truth conditions. Model theory provides the tools to carry out this program. For semantic interpretation this involves the construction of a set-theoretic model of 'the world' in terms of objects and configurations of objects; these set-theoretic constructs then serve as the semantic values of natural-language expressions.

The integrated treatment of syntax and semantics, which is now seen as the most attractive aspect of categorial grammar, is of relatively recent origin. The original Lambek systems (Lambek, 1958, 1961) were presented as syntactic type calculi. At about the same time, Curry (1961) was advocating the use of purely semantic types in natural-language analysis. Curry in fact criticized Lambek for the admixture of syntactic considerations in his category concept, coining the famous distinction between 'tectogrammatic' and 'phenogrammatic' organization. The tectogrammatic level, in Curry's view, provides the appropriate information for meaning composition; the phenogrammatic pertains to the way this abstract grammatical structure is represented in terms of surface expressions. About the actual mapping between the two levels, Curry provides no specific information.

The design of the syntax–semantics interface becomes of central importance in Richard Montague's work. The cornerstone of his 'universal grammar' program is a precise implementation of Frege's *compositionality principle*. Informally, this fundamental principle of natural-language semantics requires that the meaning of a complex expression be given as a function of the meanings of its constituent parts, and the way they are put together. In Montague's algebraic system, compositionality takes the form of a homomorphism, that is, a structure-preserving mapping, between a syntactic and a semantic algebra. Ironically, when van Benthem (1987) reintroduced semantic interpretation in the discussion of Lambek's syntactic calculi, it was by establishing the connection between categorial derivations and Curry's own 'formulae-as-types' program which we describe below. (The discussion below is restricted to functional types; the full Curry–Howard interpretation involves extension to the other type-forming operations.)

Model-theoretic Semantics, Type Theory, and the Lambda Calculus

For semantic interpretation, we associate every type A with a semantic domain D_A. Expressions of type A find their denotations in D_A. Semantic domains can be set up in two ways: directly on the basis of the syntactic types discussed in the previous section, or indirectly, via a mapping from syntactic to semantic types. The indirect option is attractive for a number of reasons. On the level of atomic types, one may want to make different basic distinctions depending on whether one uses syntactic or semantic criteria. For complex types, a map from syntactic to semantic types makes it possible to forget information that is relevant only for the way expressions are to be configured in the form dimension. Finally, the semantic type system naturally fits the language of the typed lambda calculus, which we can then use, together with its standard interpretation, to specify the instructions for meaning assembly.

Semantic and syntactic types

For a simple extensional interpretation, the set of atomic semantic types *SemAtom* could consist of types e and t, with D_e the domain of discourse (a nonempty set of entities, objects), and $D_t = \{0, 1\}$, the set of truth values. The full set of semantic types *SemType* is then obtained by closing *SemAtom* under the rule that if A and B are in *SemType* then $A \rightarrow B$ is also. $D_{A \rightarrow B}$, the semantic domain for a functional type $A \rightarrow B$, is the set of functions from

D_A to D_B. The mapping from syntactic to semantic types $(\cdot)^*$ could now stipulate for basic syntactic types that $np^* = e$, $s^* = t$, and $n^* = (e \to t)$. Sentences, in this way, denote truth values; (proper) noun phrases individuals; common nouns functions from individuals to truth values. For complex syntactic types, we set $(A/B)^* = (B \backslash A)^* = B^* \to A^*$. On the level of semantic types, the directionality of the 'slash' connective is no longer taken into account. The distinction between numerator and denominator – domain and range of the interpreting functions – is kept. Notice that both verb phrases with syntactic type $np \backslash s$ and common nouns are mapped to the semantic type $e \to t$.

The language of the simply-typed lambda calculus

Below, a procedure is presented to associate a derivation $A_1,\ldots,A_n \vdash B$, with a term t of type B representing a recipe for meaning assembly, with parameters $x_1,\ldots,x_n$ for the lexical assumptions $A_1,\ldots,A_n$. To prepare the ground, we build up the set of meaningful expressions (terms) of semantic type A, starting from a denumerably infinite set of variables for each type. For each expression t of type A, we specify its interpretation $[\![t]\!]^g$ relative to an assignment function g which assigns to each variable of type A a member of D_A.

1. *Variables*. Let x be a variable of type A. Then x is a term of type A. Interpretation: $[\![x]\!]^g = g(x)$.
2. *Application*. Let t and u be terms of type $A \to B$ and A respectively. Then $(t\ u)$ is a term of type B. Interpretation: $[\![(t\ u)]\!]^g = [\![t]\!]^g\,([\![u]\!]^g)$; i.e., the result of applying the function $[\![t]\!]^g$ to $[\![u]\!]^g$.
3. *Abstraction*. Let x be a variable of type A and t a term of type B. Then $\lambda x.t$ is a term of type $A \to B$. Interpretation: $[\![\lambda x.t]\!]^g$ is that function h from D_A into D_B such that for all objects $k \in D_A$, $h(k) = [\![t]\!]^{g'}$, where g' is the assignment that is exactly like g except for the possible difference that it assigns the object k to the variable x.

Given this interpretation, certain equalities hold between terms. One can see them as syntactic simplifications, replacing a more complex term (the *redex*) by a simpler one with the same interpretation (the *contractum*):

$(\lambda x.t)\ u \rightsquigarrow_\beta t[u/x]$ provided u is free for x in t.
$\lambda x.(t\ x) \rightsquigarrow_\eta t$ provided x is not free in t.

Formulae as Types, Proofs as Programs

Curry's basic insight was that one can see the functional types of type theory as logical implications, giving rise to a one-to-one correspondence between typed lambda terms and natural deduction proofs in positive intuitionistic logic. A natural deduction presentation for $\to$ starts from identity axioms $A \vdash A$ and has the introduction and elimination rules below, where Γ and Δ represent finite lists of formulae, and where $\Gamma - A$ results from dropping some or all occurrences of A from Γ.

$$\frac{\Gamma \vdash A \to B \quad \Delta \vdash B}{\Gamma, \Delta \vdash B} \to \text{Elim}$$

$$\frac{\Gamma \vdash B}{\Gamma - A \vdash A \to B} \to \text{Intro}$$

Let us write $\Gamma(t)$ for the string of types of free occurrences of variables in a term t. Each term t of type A now encodes a natural deduction proof of the sequent $\Gamma(t) \vdash A$. The Variable clause in the definition of well-formed terms corresponds to the axiom sequent, the Application clause to $\to$ Elimination, and the Abstraction clause to $\to$ Introduction, where the dropped A assumption corresponds to the variable bound by the lambda abstractor. In the opposite direction, every natural-deduction proof is encoded by a lambda term. The normalization of natural-deduction proofs corresponds to the β–η reductions of terms.

Translating Curry's 'formulae-as-types' idea to the categorial type logics we are discussing, we have to take the differences between intuitionistic logic and the grammatical resource logic into account. Below we repeat the natural-deduction presentation of the base logic, now taking term-decorated formulae as basic declarative units. Judgments take the form of sequents $\Gamma \vdash t : A$. The antecedent Γ is a structure with leaves $x_1 : A_1,\ldots,x_n : A_n$. The x_i are unique variables of type A_i^*, where $(\cdot)^*$ is the mapping from syntactic to semantic types. The succedent is a term t of type A^* with exactly the free variables $x_1,\ldots,x_n$, representing a program which given inputs $k_1,\ldots,k_n$ produces $[\![t]\!]^g$ under the assignment that maps the variables x_i to the objects k_i. The x_i, in other words, are the parameters of the meaning-assembly procedure. A derivation starts from axioms $x : A \vdash x : A$. The Elimination and Introduction rules have versions for the right and the left implications. On the meaning-assembly level, this syntactic difference is ironed out, as we already know that $(A/B)^* = (B \backslash A)^*$. As a consequence, we don't have the isomorphic (one-to-one) correspondence between terms and proofs of Curry's original program. But we do read off meaning assembly from the categorial derivation. (See Figure 4.)

A second difference between the programs and computations that can be obtained in intuitionistic implicational logic and the recipes for meaning assembly associated with categorial derivations has to do with the resource management of

$$[/\mathrm{I}]\ \frac{\Gamma \circ x : B \vdash t : A}{\Gamma \vdash \lambda x.t : A/B} \qquad \frac{\Gamma \vdash t : A/B \quad \Delta \vdash u : B}{\Gamma \circ \Delta \vdash (t\ u) : A}\ [/\mathrm{E}]$$

$$[\backslash\mathrm{I}]\ \frac{x : B \circ \Gamma \vdash t : A}{\Gamma \vdash \lambda x.t : B\backslash A} \qquad \frac{\Gamma \vdash u : B \quad \Delta \vdash t : B\backslash A}{\Gamma \circ \Delta \vdash (t\ u) : A}\ [\backslash\mathrm{E}]$$

Figure 4. Natural-deduction rules: term labeling.

assumptions in a derivation. The formulation of the $\rightarrow$ Introduction rule makes it clear that in intuitionistic logic the number of occurrences of assumptions (the 'multiplicity' of the logical resources) is not critical. One can make this style of resource management explicit in the form of structural rules of Contraction and Weakening, allowing for the duplication and waste of resources:

$$\frac{\Gamma, A, A \vdash B}{\Gamma, A \vdash B}\ [C]$$

$$\frac{\Gamma \vdash B}{\Gamma, A \vdash B}\ [W]$$

In contrast, the categorial type logics are resource-sensitive systems where each assumption has to be used exactly once. At the level of LP, we have the following correspondence between resource constraints and restrictions on the lambda terms coding derivations:

No empty antecedents: each subterm contains a free variable.
No Weakening: each λ operator binds a variable free in its scope.
No Contraction: each λ operator binds at most one occurrence of a variable in its scope.

Moving from LP to the grammatical base logic imposes even tighter restrictions on binding: in the absence of Associativity and Commutativity, the 'slash' introduction rules responsible for the λ operator can only reach the immediate children of a structural domain.

The syntax–semantics interface

Applied to the composition of natural-language meaning, the 'proofs-as-programs' approach has some interesting consequences for the syntax–semantics interface.

A first point to notice is the strictly modular treatment of derivational versus lexical semantics. The proof term that is read off a derivation is a *uniform instruction* for meaning assembly that fully abstracts from the contribution of the particular lexical items on which it is built. As a result, no assumptions about lexical semantics can be built into the meaning assembly process as represented by a derivation. The interplay between lexical and derivational semantics is illustrated in Figures 5 and 6. Whereas the proof term in Figure 5 is a faithful encoding of the derivation (modulo directionality and structural operations), the term one obtains in Figure 6 after substitution of lexical meaning programs and β-simplification has lost transparency with respect to the derivation.

The second feature is the limited semantic expressivity of a structure-sensitive type logic: many forms of meaning assembly that can be straightforwardly expressed in the language of the lambda calculus cannot be obtained as Curry–Howard images of the Introduction–Elimination inferences of the categorial base logic.

To resolve the tension between structure sensitivity and semantic expressivity, categorial grammars can exploit a combination of two strategies. Structural reasoning (in terms of combinators or structural postulates) makes it possible to explicitly determine which positions are accessible for semantic manipulation (binding). The example of controlled *wh*-extraction in Figure 3 is an illustration. Secondly, lexical-meaning programs do not have to obey the resource constraints of the derivational semantics. Specifically, we do not impose the single-bind condition on lexical meanings (although the ban on vacuous abstraction does make sense, also in the lexicon.) An example of multiple binding is the lexical lambda term for the relative pronoun 'that' in Figure 6, a program which computes property intersection. Another example would be a reflexive pronoun like 'himself'. With a type $((np\backslash s)/np)\backslash(np\backslash s)$, it consumes its transitive-verb argument in a resource-sensitive way. The identification of subject and object arguments of the verb is realized through its lexical lambda term $\lambda x \lambda y.((x\ y)\ y)$.

The interplay between these two strategies in current research is nicely illustrated by the construal of quantifier scope ambiguities and antecedent–anaphor dependencies. Generalized quantifier expressions, like 'everyone', 'someone', and 'nobody', require an interpretation as sets of properties; i.e., they find a denotation in $D_{(e \rightarrow t) \rightarrow t}$. A syntactic type compatible with such denotations would be $s/(np\backslash s)$. But there are two problems with such a type. First of all, it is restricted to subject position, and one wouldn't like to resort to multiple type assignments for non-subject occurrences. Secondly, it doesn't allow nonlocal scope readings, as in line 3 below, where the embedded quantifier takes scope at the main clause level.

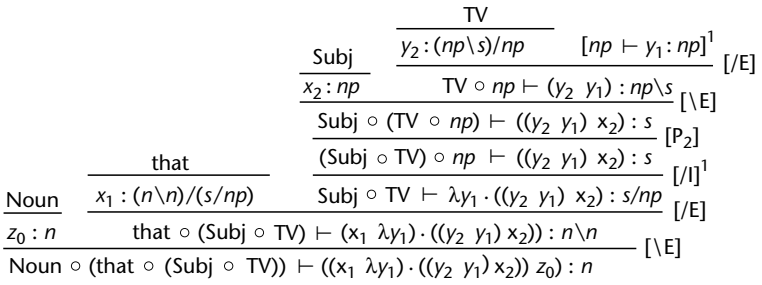


Figure 5. Computation of the proof term for the pattern 'Noun that Subj Transitive-Verb'. Leaves are labeled with variables. The derivation produces a meaning recipe with parameters for the lexical meaning programs. The recipe can be applied to any choice of lexical items fitting the type requirements: 'biscuit that Alice ate', 'book that Carroll wrote', and so on.

1.	biscuit : n – **biscuit**	Lex
2.	that : $(n\backslash n)/(s/np)$ – $\lambda z_{15}.\lambda x_{16}.\lambda y_{16}.((z_{15}\ y_{16}) \wedge (x_{16}\ y_{16}))$	Lex
3.	alice : np – **a**	Lex
4.	ate : $(np\backslash s)/np$ – **eat**	Lex
5.	np : np – y_1	Hyp
6.	ate $\circ$ np : $np\backslash s$ – (**eat** y_1)	/E (4, 5)
7.	alice $\circ$ (ate $\circ$ np) : s – ((**eat** y_1) **a**)	\E (3, 6)
8.	(alice $\circ$ ate) $\circ$ np : s – ((**eat** y_1) **a**)	P2 (7)
9.	alice $\circ$ ate : s/np – $\lambda y_1.$((**eat** y_1) **a**)	/I (5, 8)
10.	that $\circ$ (alice $\circ$ ate) : $n\backslash n$ – $\lambda x_{16}.\lambda y_{16}.$(((**eat** y_{16}) **a**) $\wedge$ (x_{16} y_{16}))	/E (2, 9)
11.	biscuit $\circ$ (that $\circ$ (alice $\circ$ ate)) : n – $\lambda y_{16}.$(((**eat** y_{16}) **a**) $\wedge$ (**biscuit** y_{16}))	\E (1, 10)

Figure 6. Substitution of lexical semantics in the pattern 'Noun that Subj Transitive-Verb'. Bold-face is used for nonlogical constants. In steps 10 and 11, β-conversion is applied on the fly to the application terms obtained from the 'slash' elimination rules. The derivation is presented in the linear or Fitch-style natural-deduction format.

1. Alice thinks someone left.
2. ((**think** ($\exists\ \lambda x.$(**leave** x))) **a**)
3. ($\exists\ \lambda x.$((**think** (**leave** x)) **a**))
4. *Alice* thinks *she* dreams.
5. ((**think** (**dream a**)) **a**)

The construal of antecedent–anaphora relations, like that of quantifier scope, involves nonlocal dependencies beyond the reach of the grammatical base logic, as in line 4 above, where the anaphor in the subordinate clause can pick up its antecedent in the main clause. In addition, meaning composition for anaphora resolution involves a duplication of resources, in the sense that one would like to make the pronoun 'she' in the example above responsible for the copying of the antecedent meaning.

Proposals for dealing with these problems rely either on combinator-style type-shifting rule schemata or on structural extensions of the Lambek calculus. For quantifier scope construal, these options are discussed in depth in Carpenter (1998). For anaphora resolution, Jäger (2001) offers a comparison of the CCG approach of Jacobson (1999) with a type-logical treatment based on identity semantics for anaphora, in combination with a restricted copying rule in syntax, in the form of a controlled structural rule of Contraction. An alternative perspective on scope and anaphora, more in the spirit of Curry's tectogrammatic program, simplifies the categorial type theory to a nondirectional LP system, and enforces structural control by introducing lambda term labeling also for the form dimension of grammatical signs. Oehrle (1994) provides an early formulation of this approach, which has recently found new advocates.

Processing Issues

The interpretation procedure discussed above is essentially dynamic: interpretations are assembled 'online' in the course of the derivation process, rather than being computed *post hoc* from a given static structure. This has led to a distinctly 'categorial' view on processing issues.

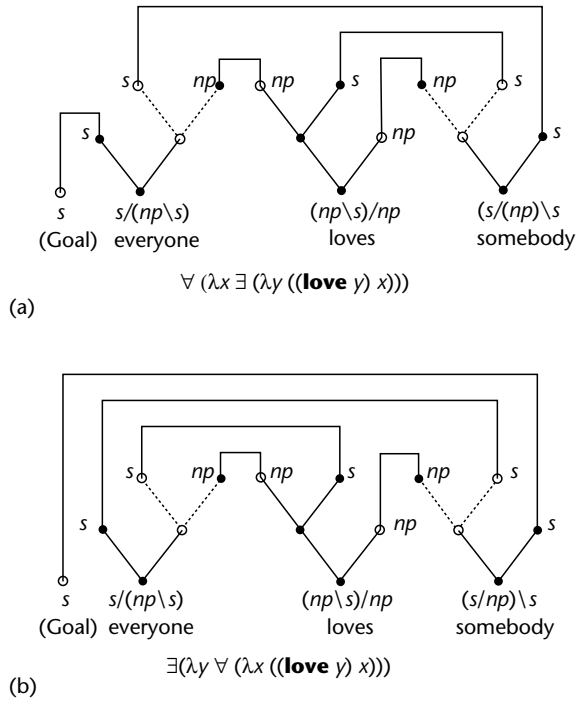


Figure 7. Incremental proof-net construction. The diagrams show a proof net for the sentence 'everyone loves somebody'. Formula decomposition trees have polarized vertices (black for input, white for output). Solid edges represent positive slashes; dotted edges represent negative slashes. A linking of leaves with opposite polarities is well-formed if it produces a graph that is connected, acyclic (for each removal of a dotted edge from a pair), and planar. The net is constructed in a left-to-right incremental fashion. Processing complexity is measured in terms of the number of unresolved dependencies. (a) The subject-wide-scope reading for 'everyone loves somebody' (maximum of 3 unresolved dependencies). This is preferred over (b), the object-wide-scope reading (maximum of 4 unresolved dependencies). Adapted from Johnson (1998) and Morrill (2000).

Incrementality and information structure

The flexible notion of derivational constituency engendered by type-changing principles makes left-to-right parsing directly compatible with incremental interpretation. The resulting categorial modeling of natural-language processing has been worked out in Steedman (2000a). This work shows that derivational constituency is guided by 'prosodic' articulation (intonation contour). To do justice to this dimension of grammatical organization, one needs a richer notion of semantic interpretation, accommodating notions of focus and information structure. Steedman's proposals are formulated in the CCG style; Hendriks (1999) analyzes information packaging and intonation contour in multimodal type-logical terms.

Proof nets

A novel computational view on natural-language processing derives from the 'proof net' approach. Proof nets were originally developed in the context of linear logic, where they elegantly capture the essence of resource-sensitive derivations in graph-theoretical terms. Moot and Puite (2002) refine the proof-net techniques for use with the grammatical type logics discussed in this article, where, apart from resource multiplicity, structural patterns also have to be taken into account.

Johnson (1998) and Morrill (2000) have pointed out that proof nets offer an attractive perspective on performance phenomena. A net can be built in a left-to-right incremental fashion by establishing possible linkings between the input–output connectors of lexical items as they are presented in real time. This suggests a simple complexity measure on a traversal, given by the number of unresolved dependencies between literals. This complexity measure on incremental proof-net construction makes the right predictions about a number of well-known processing issues, such as the difficulty of center embedding, garden-path effects, attachment preferences, and preferred scope construals in ambiguous constructions. An illustration is presented in Figure 7.

CONCLUSION

In this article we have presented the core part of the categorial vocabulary: a set of type forming connectives controlling the structure-building operations of language, and the way in which a derivation is associated with an instruction for meaning assembly.

References

Adriaans P (2002) Learning shallow context-free languages under simple distributions. In: Vermeulen K and Copestake A (eds) *Algebras, Diagrams and Decisions in Language, Logic and Computation*, Stanford, CA: CSLI.

van Benthem J (1987) Categorial grammar and lambda calculus. In: Skordev D (ed.) *Mathematical Logic and Its Applications*, pp. 39–60. New York, NY: Plenum Press.

Bernardi R (2002) *Reasoning with Polarities in Categorial Type Logic*. PhD thesis, Utrecht University.

Buszkowski W (1997) Mathematical linguistics and proof theory. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, chap. XII, pp. 683–736. Amsterdam: Elsevier/Cambridge, MA: MIT Press.

Carpenter B (1998) *Type-Logical Semantics*. Cambridge, MA: MIT Press.

Curry HB (1961) Some logical aspects of grammatical structure. In: Jacobson R (ed.) *Structure of Language*

and Its Mathematical Aspects, pp. 56–68. Providence, RI: American Mathematical Society.
Došen K (1992) A brief survey of frames for the Lambek calculus. *Zeitschrift für mathematischen Logik und Grundlagen der Mathematik* **38**: 179–187.
Foret A (2001) On mixing deduction and substitution in Lambek categorial grammars. In: de Groote P, Morrill G and Retoré C (eds) *Logical Aspects of Computational Linguistics*, pp. 158–174. Berlin, Germany: Springer.
Hendriks H (1999) The logic of tune: a proof-theoretic analysis of intonation. In: Lecomte A, Lamarche F and Perrier G (eds) *Logical Aspects of Computational Linguistics*, pp. 132–159. Berlin, Germany: Springer.
Jacobson P (1999) Towards a variable-free semantics. *Linguistics and Philosophy* **22**(2): 117–184.
Jäger G (2001) Anaphora and quantification in categorial grammar. In: Moortgat M (ed.) *Logical Aspects of Computational Linguistics*, pp. 70–90. Berlin, Germany: Springer.
Johnson M (1998) Proof nets and the complexity of processing center-embedded constructions. *Journal of Logic, Language and Information* **7**(4): 443–447.
Kanazawa M (1998) *Learnable Classes of Categorial Grammars*. Stanford, CA: CSLI.
Kurtonina N (1995) *Frames and Labels: A Modal Analysis of Categorial Inference*. PhD thesis, OTS Utrecht, ILLC Amsterdam.
Kurtonina N and Moortgat M (1997) Structural control. In: Blackburn P and de Rijke M (eds) *Specifying Syntactic Structures*, pp. 75–113. Stanford, CA: CSLI.
Lambek J (1958) The mathematics of sentence structure. *American Mathematical Monthly* **65**: 154–170.
Lambek J (1961) On the calculus of syntactic types. In: Jacobson R (ed.) *Structure of Language and its Mathematical Aspects*, pp. 166–178. Providence, RI: American Mathematical Society.
Moortgat M (1996) Multimodal linguistic inference. *Journal of Logic, Language and Information* **5**(3–4): 349–385.
Moortgat M (1999) Constants of grammatical reasoning. In: Bouma G, Hinrichs E, Kruijff GJ and Oehrle RT (eds) *Constraints and Resources in Natural Language Syntax and Semantics*, pp. 195–219. Stanford, CA: CSLI.
Moot R (2002) *Proof Nets for Linguistic Analysis*. PhD thesis, Utrecht University.
Moot R and Puite Q (2002) Proof nets for the multimodal Lambek calculus. *Studia Logica* **71**. [Special issue on the occasion of Lambek's eightieth birthday, edited by Wojciech Buszkowski and Michael Moortgat.]
Morrill G (2000) Incremental processing and acceptability. *Computational linguistics* **26**(3): 319–338.
Oehrle RT (1994) Term-labeled categorial type systems. *Linguistics and Philosophy* **17**(6): 633–678.
Pentus M (1993) Lambek grammars are context free. In: *Proceedings of the 8th Annual IEEE Symposium on Logic in Computer Science*, pp. 429–433. IEEE Computer Society Press.
Pentus M (1994) Language completeness of the Lambek calculus. In: *Proceedings of the 9th Annual IEEE Symposium on Logic in Computer Science*, pp. 487–496. IEEE Computer Society Press.
Steedman M (2000a) Information structure and the syntax–phonology interface. *Linguistic Inquiry* **31**(4): 649–689.
Steedman M (2000b) *The Syntactic Process*. Cambridge, MA: MIT Press.
Tiede HJ (2001) Lambek calculus proofs and tree automata. In: Moortgat M (ed.) *Logical Aspects of Computational Linguistics*, pp. 251–265. Berlin, Germany: Springer.
Vijay-Shanker K and Weir D (1994) The equivalence of four extensions of context free grammars. *Mathematical Systems Theory* **27**(6): 511–546.

Further Reading

Ajdukiewicz K (1935) Die syntaktische Konnexität. *Studia Philosophica* **1**: 1–27. [The seminal work on categorial grammar. English translation in: McCall S (ed.) (1996) *Polish Logic, 1920–1939* pp. 207–231. Oxford, UK: Oxford University Press.]
Bar-Hillel Y (1953) A quasi-arithmetical notation for syntactic description. *Language* **29**: 47–58. [Important continuation of Ajdukiewicz' work.]
Buszkowski W, Marciszewski W and van Benthem J (eds) (1998) *Categorial Grammar*. Amsterdam, Netherlands: Benjamins. [Includes several early papers, including Lambek's seminal paper of 1958.]
Dowty D, Wall R and Peters S (1981) *Introduction to Montague Semantics*. Dordrecht, Netherlands: Reidel.
Girard J-Y (1987) Linear logic. *Theoretical Computer Science* **50**: 1–102.
Girard J-Y, Lafont Y and Taylor P (1988) *Proofs and Types*. Cambridge, UK: Cambridge University Press. [A good source for the Curry–Howard interpretation.]
Kruijff G-J and Oehrle R (2002) *Resource Sensitivity in Binding and Anaphora*. Dordrecht, Netherlands: Reidel. [A reflection of current categorial views on anaphora and binding.]
Lambek J (1999) Type grammar revisited. In: Lecomte A, Lamarche F and Perrier G (eds) *Logical Aspects of Computational Linguistics*, pp. 1–27. Berlin, Germany: Springer.
Montague R (1974) *Formal Philosophy: Selected papers of Richard Montague*. Yale, CT: Yale University Press.
Moortgat M (1997) Categorial type logics. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, chap. II, pp. 93–177. Amsterdam: Elsevier / Cambridge, MA: MIT Press. [The primary source for this article.]
Moot R (2002) *Grail*. http://www.let.uu.nl/~Richard.Moot/personal/grail.html. [A grammar development environment that provides a versatile computational tool for categorial exploration. The user interacts with the kernel via a graphical user interface, which provides control over the lexicon and the structural module, and which gives access to a fully-fledged proof-net-based

debugger. A number of sample fragments can be accessed online at http://www.grail.let.uu.nl/tour.pdf.]

Morrill G (1994) *Type Logical Grammar: Categorial Logic of Signs*. Dordrecht, Netherlands: Kluwer. [A rich fragment of syntactic and semantic phenomena in the grammar of English, using a variety of type-forming operations (Boolean and quantificational) in addition to the composition operators.]

Oehrle R, Bach E and Wheeler D (eds) (1988) *Categorial Grammars and Natural Language Structures*. Dordrecht, Netherlands: Reidel. [A good picture of categorial research in the 1980s, in both the rule-based and the logical traditions.]

Restall G (2000) *An Introduction to Substructural Logics*. New York, NY: Routledge. [An accessible textbook.]

Retoré C and Stabler E (eds) (2002) *Resource Logics and Minimalist Grammars*. *Proceedings ESSLLI'99 workshop*. [Special issue on language and computation. An exploration of the connections between linear logic, categorial grammar, and computational formulations of minimalist grammars.]

van Benthem J (1995) *Language in Action: Categories, Lambdas and Dynamic Logic*. Cambridge, MA: MIT Press. [A detailed study of the relations between categorial derivations, type theory and lambda calculus, and of the place of categorial grammars within the general landscape of resource-sensitive logics.]

van Benthem J and ter Meulen A (eds) (1997) *Handbook of Logic and Language*. Amsterdam: Elsevier/Cambridge, MA: MIT Press. [Includes chapters on the connections between categorial type systems and mathematical linguistics and proof theory, formal learning theory, type theory, and Montague grammar.]

Cognitive Linguistics

Intermediate article

Gilles Fauconnier, University of California, San Diego, California, USA

CONTENTS

Cognitive linguistics is a theoretical and empirical programme that goes beyond the visible structure of language to investigate the complex background operations of cognition that create grammar, conceptualization, discourse, and thought itself.

INTRODUCTION

Cognitive linguistics is a powerful approach to the study of language, conceptual systems, and human cognition. It addresses the structuring of basic conceptual categories such as space and time, scenes and events, entities and processes, motion and location, force and causation, and also the structuring of ideational and affective categories, such as attention and perspective, volition and intention (Talmy, 2000). In doing so, it develops a rich conception of grammar that reflects fundamental cognitive abilities: the ability to form structured conceptualizations with multiple levels of organization, to conceive of a situation at varying levels of abstraction, to establish correspondences between facets of different structures, and to construe a situation in different ways (Langacker, 1987, 1991).

Much of traditional linguistics (including structural and generative linguistics) focuses on the ways in which words are combined into sentences. For example, in looking at the four English sentences *the plane flies over the city*, *the post office is over the hill*, *the war is over the oil wells*, and *this topic is over my head*, a grammarian might see a single structure, and leave it at that. For cognitive linguistics, these four sentences would present a far greater challenge: to explain how the same structure, and the same word *over*, gives rise to such different kinds of conceptualizations: a plane moving 'above' a city, a post office located 'at the end of' a path on the hill, a war 'caused by' something to do with the oil wells, and a topic 'high' on some scale of difficulty. Hidden behind simple words and everyday language are vast conceptual networks manipulated unconsciously through the activation of powerful neural circuits.

Cognitive linguistics recognizes that the study of language is the study of language use, and that when we engage in any language activity, we draw unconsciously on vast cognitive and cultural resources, call up models and frames, set up multiple connections, coordinate large arrays of information, and engage in creative mappings, transfers, and elaborations. Language does not 'represent' meaning: it prompts for its imaginative construction. Very sparse grammar guides us along rich mental paths, by prompting us to perform complex cognitive operations. A large part of cognitive linguistics centers on the creative 'online' construction of meaning as discourse unfolds in context (Fauconnier and Sweetser, 1996; Sweetser, 2000).

Aspects of language and expression that have often been consigned in formal work to the rhetorical periphery of language, such as metaphor (Lakoff and Johnson, 1980, 1999; Sweetser, 1990) and metonymy (Panther and Radden, 1999; Barcelona, 2000) are central within cognitive linguistics. They are understood to be powerful conceptual mappings essential to human thought, important for the understanding not just of poetry, but also of science, mathematics, religion, philosophy, and everyday speaking and thinking.

Thought and language are embodied. Conceptual structure arises from our sensorimotor experience and the neural structures that give rise to it. Reason is embodied and imaginative. A grammar is ultimately a neural system, and general cognitive capacities drive language (Fauconnier and Turner, 1998, 2002).

The stage was set for cognitive linguistics in the 1970s and early 1980s with Len Talmy's work on figure and ground, Ronald Langacker's cognitive grammar framework, George Lakoff's research on metaphor, gestalts, categories and

prototypes (Lakoff, 1987), Fillmore's frame semantics (Fillmore, 1982) and Fauconnier's mental spaces (Fauconnier, 1994). A wealth of discoveries, empirical studies, and applications have since emerged (Janssen and Redeker, 2000; Tomasello, 1998; Cuyckens and Geeraerts, forthcoming).

GRAMMAR AND COGNITION

The relation of grammar to cognition is studied in fine detail in the foundational work of Talmy (2000) and Langacker (1987, 1991). Talmy shows that there are great restrictions on the conceptual categories that grammatical systems actually specifiy. For example, languages often require that number be marked with forms like singular, plural, or dual, but no language has a system for marking the color of nouns. Topological configurations are marked grammatically, but not Euclidean metrics. (Thus, prepositions like *across* indicate the same configuration regardless of the size of the landmarks (*across the sky*, *across the table*), but absolute size or range of size is not marked or differentiated grammatically.) Talmy singles out multiplexing, and states of boundedness and dividedness, as strongly markable in grammatical systems. He also singles out perspective and sequentializing. Thus, *the door slowly opened and two men walked in* and *two men slowly opened the door and walked in* describe the same event from two different perspectives; *there are some houses in the valley* and *there is a house now and then in the valley* describe the same objective state of affairs, but the second sentence takes us along a sequential path where we 'see' the houses one at a time.

Langacker shows how grammar imposes 'trajector–landmark' organization on scenes and events. In *the table is below the lamp*, the table (the 'trajector') is located with respect to the lamp (the 'landmark'). In *the lamp is above the table* this relationship is reversed: the table serves as landmark. But both sentences reflect the same spatial configuration. 'Profiling' is another important construct of Langacker's cognitive grammar. The word *hypotenuse* evokes a right-angled triangle and 'profiles' a particular part of it: the same line segment without the rest of the triangle is no longer a hypotenuse. In *I melted it*, the word *melted* profiles an entire action chain with causation and change leading to a liquid state. In *it melted easily*, only the change is profiled, although the causation is still evoked. In *it is finally melted*, only the resultant state is profiled, but the unprofiled change is evoked. Langacker analyzes in considerable detail the ways in which component structures are integrated through correspondences and elaboration to form composite structures: for example, how the phonological integration *jar lid* symbolizes the semantic integration of 'jar' and 'lid' (Langacker, 1987, 2000; Van Hoek, 1997).

Other basic aspects of conceptual structuring, as reflected by grammar and found in many languages (Talmy, 2000), include 'fictive motion' (*the blackboard goes all the way to the wall*), 'event integration' (*the ball rolled in*, *the candle blew out*, *I kicked the door shut*), and 'force dynamics' (*the ball kept rolling*, *he refrained from closing the door*). (Force dynamics also applies to social organization and abstract reasoning (Sweetser, 1990).) Fictive motion allows stationary scenes (the position of the blackboard) to be construed in terms of the motion of a fictive trajector (*goes all the way to*). Event integration compresses complex chains of events. Force dynamics is based on our experience of physical forces that cause or prevent motion: *kept rolling* indicates an absence of force to impede the ball's movement; *refrained* indicates a clash of counteracting forces. *Max must be home by ten* indicates a social force (e.g. parental rules, curfew) directed towards an outcome (Max being home by ten). *Max must be home by now* uses the same modal *must* to indicate a logical force directed towards the conclusion that Max is now home.

The way in which language structures space is interesting both linguistically and psychologically. No two languages are quite alike in this respect, although there are some general principles. Deceptively simple-seeming prepositions like *in*, *out* or *over* define elaborate networks of spatial meaning with hundreds of linked schemata, some of which are prototypical and central. Compare the very different senses of *over* in *the plane flew over the field*, *the post office is over the hill*, *the log rolled over*, *the party is over*, *he had to do it over again*, *he overlooked it*, *he looked it over*, *he oversaw it*. A native speaker of English unconsciously masters a vast network of related schemata linked to the single, simple word *over*. Remarkable work on this topic has been done by cognitive linguists (Lindner, 1982; Brugman, 1981; Herskovits, 1986; Vandeloise, 1991; Talmy, 2000). Explicit computational models (Regier, 1996) reflect the cognitive complexity of the human capacity to structure space linguistically.

METAPHOR THEORY

A second strand of work in cognitive linguistics since the 1980s has been the development of metaphor theory. Launched by George Lakoff and Mark Johnson, this line of research begins with the insight

that metaphor is basic and constitutive for all the thinking that we do. Metaphor theory is based on source domains of human experience and on neural connections to our embodied sensations, actions, and emotions. Metaphors create the possibility of 'abstract' reasoning, scientific and mathematical thought, and language and culture generally.

Source domains seem to be used systematically to structure target domains by means of metaphorical mappings. For example, our general way to talk and think about event structure is in terms of motion. In this metaphorical mapping, states are locations, change of state is change of location, causes are forces, purposes are destinations, means are paths to destination, guided action is guided motion, etc. This metaphor is reflected in the language we use to express event structure: *he went crazy, she entered a state of euphoria, the clothes are somewhere between wet and dry, the home run threw the crowd into a frenzy, she walked him through the problem, I've hit a brick wall, we're moving ahead, we're at a standstill* (Lakoff and Johnson, 1999). The structure and inferences of the source domain of motion are projected to the target domain of events and action to create a rich emergent conceptualization.

Time itself is conceptualized in terms of space and motion. In English, times can be represented as objects moving towards and then past a stationary observer, or as objects that are stationary with respect to a moving observer: *the time will come; Christmas is approaching*; *the summer just zoomed by*; *we're getting close to Christmas*; *we've reached the end of May already*.

Conventional metaphors such as these can be extended to enrich conceptual understanding. Time can *fly* and *crawl* and *disappear*. In a line by Shakespeare (*Troilus and Cressida*, IV. v. 202–203; cited in Gibbs (1994)), where Hector greets Nestor, Time becomes a moving person, who holds the hand of the venerable Nestor:

> *Let me embrace thee, good old chronicle,*
> *That hast so long walk'd hand in hand with time.*

MENTAL SPACES AND CONCEPTUAL INTEGRATION

Mental spaces are small conceptual packets, constructed as we think and talk, for purposes of local understanding and action. They are partial assemblies of elements, structured by frames and cognitive models. They are interconnected and can be modified as thought and discourse unfold.

Mental spaces proliferate in the unfolding of discourse, map onto each other, and provide abstract mental structure for shifting viewpoint and focus, allowing us to direct our attention at any time onto partial and simple structures while maintaining an elaborate web of connections in working and long-term memory.

For example, if we say *in reality, Richard Burton loves Elizabeth Taylor, but in the movie, he kills her*, we set up two mental spaces, one for reality and one for the movie. Richard Burton in reality has a counterpart (say Mark Antony) in the movie, and Elizabeth Taylor in reality has a counterpart (say Cleopatra) in the movie. Connections between mental spaces allow access to elements in one mental space via counterparts in other mental spaces (e.g. Mark Antony via Burton). Mental spaces offer a general and elegant means of dealing with opacity, presupposition, counterfactuals, and tense and mood in language.

Take for example the sentence *in 1957, the president was a baby*, appearing in a discourse where a base mental space with G. W. Bush as current US president has been set up. The phrase *in 1957* sets up a new '1957' space. If we take *the president* to describe Bush in the base, its counterpart 'Bush in 1957' will be accessed, and the sentence will mean that Bush was a baby back in 1957. If, on the other hand, we take *the president* to describe someone in the new mental space of '1957', then that person will be both a baby and a president in 1957, and the sentence will mean that a baby was president in 1957. Multiple access possibilities of this kind allow the same sentence to prompt for different connection paths in different situations. A wide range of puzzling reference phenomena can be explained in terms of this general underspecification of connecting paths.

Behind the idiosyncrasies of language, cognitive linguistics has uncovered much evidence for the operation of more general cognitive processes. 'Conceptual integration' is an example of a general cognitive operation on mental spaces that is reflected universally in the way we think.

Conceptual integration consists in setting up networks of mental spaces that map onto each other and blend into new mental spaces in various ways. Some of the integrations are novel, others are more entrenched, and we rarely pay conscious attention to the process, because it is so pervasive. In a conceptual integration network, partial structure from input mental spaces is projected to a new blended mental space which develops dynamic (imaginative) structure of its own.

For example, the counterfactual *in France, Watergate would not have done Nixon any harm* is intended to prompt inferences on the difference between the

American and French political systems. It requires the listener to construct input spaces for American politics and for French politics and to project selectively into a blended space in which Nixon and Watergate are embedded into French politics. The imaginative emergent structure of that mental space (why Nixon is not harmed, etc.) provides insight into the political realities of the two countries.

Most aspects of human life evoke conceptual integration networks. This remarkable cognitive capacity has been studied in a variety of domains, including mathematics, music, action and design, distributed cognition, magic and religion, anthropology and political science (Zbikowski, 2002; Hutchins, in preparation; Sorensen, 2000; Lakoff and Núñez, 2000; Liddell, 1998; Turner, 2001). It has been suggested that the capacity of conceptual integration evolved biologically to reach a threshhold, 'double-scope creativity', that constitutes a necessary condition for the cognitively modern human singularities of art, creative tool-making, religious thought, and grammar (Fauconnier and Turner, 2002).

SUMMARY

Cognitive linguistics goes beyond the visible structure of language and investigates the complex background operations of cognition that create grammar, conceptualization, discourse, and thought itself. The theoretical insights of cognitive linguistics are based on extensive empirical observation in multiple contexts, and on experimental work in psychology and neuroscience (Gibbs, 1994; McNeill, 2000; Coulson, 2001; Mandler, 1992; Gentner, 2001). Results of cognitive linguistics, especially from metaphor theory and conceptual integration theory, have been applied to wide ranges of nonlinguistic phenomena.

References

Barcelona A (ed.) (2000) *Metaphor and Metonymy at the Crossroads*. Berlin: Mouton de Gruyter.

Brugman C (1981) *Story of Over: Polysemy, Semantics, and the Structure of the Lexicon*. New York, NY: Garland.

Coulson S (2001) *Semantic Leaps*. New York, NY and Cambridge, UK: Cambridge University Press.

Cuyckens H and Geeraerts D (forthcoming) *Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.

Fauconnier G (1994) *Mental Spaces*. New York, NY: Cambridge University Press. [First published 1985. Cambridge, MA: MIT Press.]

Fauconnier G and Sweetser E (eds) (1996) *Spaces, Worlds, and Grammar*. Chicago, IL: University of Chicago Press.

Fauconnier G and Turner M (1998) Conceptual integration networks. *Cognitive Science* **22**(2): 133–187.

Fauconnier G and Turner M (2002) *The Way We Think*. New York, NY: Basic Books.

Fillmore C (1982) Frame semantics. In: Linguistic Society of Korea (eds) *Linguistics in the Morning Calm*, pp. 111–137. Seoul: Hanshin.

Gentner D (2001) Spatial metaphors in temporal reasoning. In: Gattis M (ed.) *Spatial Schemas in Abstract Thought*, pp. 203–222. Cambridge, MA: MIT Press.

Gibbs R (1994) *The Poetics of Mind*. Cambridge, UK: Cambridge University Press.

Herskovits A (1986) *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Cambridge, UK: Cambridge University Press.

Hutchins E (in preparation) Material anchors for conceptual blends.

Janssen T and Redeker G (eds) (2000) *Scope and Foundations of Cognitive Linguistics*. The Hague: Mouton de Gruyter.

Lakoff G (1987) *Women, Fire, and Dangerous Things*. Chicago, IL: University of Chicago Press.

Lakoff G and Johnson M (1980) *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lakoff G and Johnson M (1999) *Philosophy in the Flesh*. New York, NY: Basic Books.

Lakoff G and Núñez R (2000) *Where Mathematics Comes From*. New York, NY: Basic Books.

Langacker R (1987) *Foundations of Cognitive Grammar*, vol. I. Stanford, CA: Stanford University Press.

Langacker R (1991) *Foundations of Cognitive Grammar*, vol. II. Stanford, CA: Stanford University Press.

Langacker R (2000) Assessing the cognitive linguistic enterprise. In: Janssen and Redeker (2000), pp. 13–60.

Liddell S (1998) Grounded blends, gestures, and conceptual shifts. *Cognitive Linguistics* **9**(3): 283–314.

Lindner S (1982) What goes up doesn't necessarily come down: the Ins and Outs of opposities. *Chicago Linguistic Society* **18**: 305–323.

Mandler JM (1992) How to build a baby II: Conceptual primitives. *Psychological Review* **99**: 587–604.

McNeill D (2000) *Language and Gesture*. Cambridge, UK: Cambridge University Press.

Panther KU and Radden G (eds) (1999) *Metonymy in Language and Thought*. Amsterdam: John Benjamins.

Regier T (1996) *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press.

Sorensen J (2000) *Essence, Schema, and Ritual Action: Towards a Cognitive Theory of Magic*. PhD thesis, University of Aarhus.

Sweetser E (1990) *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge, UK: Cambridge University Press.

Sweetser E (2000) Compositionality and blending: working towards a fuller understanding of semantic composition in a cognitively realistic

framework. In: Janssen and Redeker (2000), pp. 129–162.
Talmy L (2000) *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.
Tomasello M (ed.) (1998) *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Erlbaum.
Turner M (2001) *Cognitive Dimensions of Social Science*. New York, NY: Oxford University Press.
Vandeloise C (1991) *Spatial Prepositions: A Case Study from French*. Chicago, IL: University of Chicago Press.
Van Hoek K (1997) *Anaphora and Conceptual Structure*. Chicago, IL: University of Chicago Press.
Zbikowski L (2002) *The Conceptualizing Music: Cognitive structure, theory and analysis*. New York, NY: Oxford University Press.

Further Reading

Fauconnier G (1997) *Mappings in Thought and Language*. Cambridge, UK: Cambridge University Press.
Israel M (forthcoming) *The Rhetoric of Grammar*. Cambridge, UK: Cambridge University Press.
Jackendoff R (1983) *Semantics and Cognition*. Cambridge, MA: MIT Press.
Kemmer S and Verhagen A (1994) The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* **5**: 115–156.
Mandelblit N (1997) *Grammatical Blending: Creative and Schematic Aspects in Sentence Processing and Translation*. PhD thesis, University of California, San Diego.
Moore T and Carling C (1982) *Language Understanding: Towards a Post-Chomskyan Linguistics*. New York, NY: St Martin's Press.
Robert A (1998) Blending in the interpretation of mathematical proofs. In: Koenig J-P (ed.) *Discourse and Cognition*. Stanford, CA: CSLI.
Turner M (1996) *The Literary Mind*. New York, NY: Oxford University Press.

Constraint-based Processing

Intermediate article

Bob Carpenter, SpeechWorks International, New York, NY, USA

CONTENTS

In a constraint-based grammar formalism, a class of constraints is used to reduce a class of potential representations to the representations that are well formed, or grammatical.

INTRODUCTION

Linguistic theories have primarily been concerned with the structures of languages and utterances within those languages. Extensionally, all of the mainstream linguistic theories concentrate on picking out the set of grammatical, or well-formed, structures (sounds, words or phrases, for instance) in a language. Intensionally, linguists debate the right way to pick out such structures, often arguing from a cognitive perspective based on evidence of human language development, comprehension or production, or typological variation. Computer scientists are typically concerned with the efficiency of processing in terms of time and space, often sidestepping cognitive issues in the interest of building effective software.

CONSTRAINT-BASED GRAMMAR FORMALISMS

Two basic approaches to characterizing well-formed structures can be usefully compared and contrasted. The first, more traditional, approach is based on the notion of inductive definitions. To use a logical example that motivated linguists, consider an inductive definition of the well-formed formulae (WFF) of propositional logic. First, there is the base case of propositional symbols, which are assumed to be WFFs. Then, if P and Q are WFFs then so is the conjunction (P and Q) and the negation (not P). Iterating these constructions generates the full set of WFFs. Any formula that can be built up from propositional symbols by applying conjunction or negation is taken to be well formed.

The second approach to picking out the well-formed structures is based on constraints. Rather than starting from a small set of structures known to be well formed, a large set of possibilities is considered and the ill-formed ones are removed. For instance, the prime numbers are naturally characterized in this way. Start with all of the natural numbers from 2 onwards, and eliminate the compound numbers, that is numbers that are the product of two smaller numbers greater than one. The remaining numbers are the prime numbers.

Bar-Hillel (1950), working from the tradition of logical languages in mathematics and philosophy, presented one of the first mathematically rigorous definitions of a system for generating the expressions of a natural language. He worked bottom-up from a lexicon associating words and categories, such as *dog* and 'singular common noun', and introduced rules to build larger phrases. For instance, by concatenating a determiner such as *the* with a common noun such as *dog*, the noun phrase *the dog* is generated. This was essentially an early instance of a phrase structure grammar. (*See* **Phrase Structure Grammar, Head-driven**)

Although there is no universally agreed definition, a generative theory is typically taken as one in which a mechanism is proposed for generating (usually by some formal mathematical means) the set of grammatical sentences of a language. Chomsky (1957) proposed perhaps the best-known example of an early generative theory of language. Chomsky introduced the notion of transformation, which he used to relate constructions in a language, such as interrogative and declarative forms of sentences, through derivation. He employed phrase-structure techniques similar to those of Bar-Hillel to generate 'base forms', and then applied various structure-permuting transformations to generate grammatical 'surface forms'. For instance, Chomsky produced the interrogative *will John run* from the delcarative *John will run* by means of a transformation that moves the matrix finite auxiliary to the front of the sentence.

Over the next 20 years, generative systems acquired all the accoutrements of a maturing science. Constraints were introduced to limit the

applicability of some operations. For instance, noun compounding combines two nouns into a new noun, such as *coal oven*, but a constraint would be imposed that the first noun not be plural. By the late 1970s, in the 'government-binding theory' (Chomsky, 1980), the role of constraints on transformations had grown to the point where the transformations themselves were of the general form 'move anything anywhere', requiring constraints to control specifically what moved where. This resulting framework is thus explicitly constraint-based. (*See* **Government–Binding Theory**)

MATHEMATICAL FOUNDATIONS

The mathematics of constraint processing in general involves the distinction between inductive and co-inductive definitions (Barwise and Etchemendy, 1987). An inductive definition provides a base set, and then operations that expand the set. The set defined is then taken to be the smallest set containing the base cases and closed under the operations. A co-inductive definition, on the other hand, begins with a set of structures and provides constraints on the forms of elements. The set defined is the largest subset of the original set satisfying all of the constraints.

The standard general formulation of constraint problems is in terms of logical satisfiability. Consider the simple language of propositional formulae given above, under the standard interpretation. It is natural to ask, for a given formula, whether there is an assignment of truth values to its propositional symbols under which it is true. The formula is thus a kind of constraint on assignments of truth values. Unfortunately, even this simple constraint problem for this simple logical language is very complex computationally. The problem of determining propositional satisfiability (known as 'SAT') is NP-complete (Cormen *et al.*, 1990): every known algorithm to solve such problems can encounter cases whose processing time grows faster than any polynomial. (*See* **Computability and Computational Complexity**)

Robinson (1965) introduced the general technique of 'resolution' for proving theorems in first-order logic. Using the notion of resolution and the simple facts of linguistic agreement, Colmerauer (1993) introduced the notion of logic programming, motivated primarily by the search and constraint problems introduced by linguistic grammars. Colmerauer used logical rules such as the following to code the fact that the noun phrase and verb phrase must agree in number and that the sentence carries the verb form of its matrix verb:

```
s(Words1 + Words2, VerbForm)
  if np(Words1, Number) and
     vp(Words2, VerbForm,
        Number)                    (1)
```

The symbol `s` is taken to be a two-place predicate, whose first argument is a sequence of words and whose second argument is a verb form such as `finite` or `infinitive`. The symbols for noun phrases and verb phrases, `np` and `vp`, are to be read similarly, with an additional argument for number. Such rules are implicitly universally quantified, and thus the above rule can be read as saying that if `Words1` is a sequence of words that forms a noun phrase of number `Number` and `Words2` is a sequence of words that forms a verb phrase of the same number and a verb form `VerbForm`, then the concatenation of `Words1` and `Words2`, indicated as `Words1 + Words2`, forms a sentence with the given verb form `VerbForm`. (*See* **Resolution Theorem Proving**)

The notion of constrained phrase-structure rules was introduced, along with a very general notion of feature, into mainstream linguistics by Harman (1963). Since then, every branch of the field, from phonetics to pragmatics, has adopted some kind of 'feature-based' analysis. In these theories, rather than relying on a positional encoding, as in first-order terms, the values are named.

In theories such as lexical–functional grammar and head-driven phrase structure grammar, feature structures are used to represent partial information. This more general data structure is particularly adept at representing the kind of sparse information found in linguistic constraints, which often indicate the behavior of a handful of linguistic features among hundreds. Feature structures of the form employed in head-driven phrase structure grammar also allow a natural form of knowledge representation through inheritance. Processing with such structures has essentially followed that of first-order terms, with systems being based on constraint resolution and, in particular, unification. (Unification is an operation that takes two terms or feature structures and produces a new term that contains all of the information in both, or returns failure if they are inconsistent.) There are also strong similarities to production systems operating over frames. (*See* **Production Systems and Rule-based Inference**; **Knowledge Representation**; **Lexical-Functional Grammar**)

CONSTRAINT-BASED PARSING

Natural language parsing, as a stage in more general natural language processing, involves generating some or all of the structures compatible with a given sequence of words. Given a context-free grammar and a sequence of words, a representation of all parses can be generated in polynomial time (slightly subcubic). The 'universal recognition problem' involves taking a grammar and a sequence of words and determining whether there is some parse for a sequence of words. Barton *et al*. (1987) showed that the introduction of agreement constraints like those of equation 1 yields a universal recognition problem that is NP-complete; the problem for any fixed grammar, though, remains polynomial. Similarly, word-order constraints added to a general context-free grammar result in NP-complete parsing problems. (*See* **Natural Language Processing**; **Parsing: Overview**)

Most constraint-based parsing systems are hybrids that involve a standard search-based parsing algorithm with side constraints. At any stage where constituents are combined or rules are expanded, a check is made to ensure that the set of constraints remains consistent. If not, the search path is abandoned. At each stage, the information known about variables and their values is represented by means of an assignment to variables. For instance, in parsing *the papers falls* from left to right and bottom up, the determiner *the* is encountered first. This is unspecified for number. When it is combined with the plural noun *papers*, the value of the noun phrase's agreement feature becomes 'plural'. Finally, the attempt to combine the plural noun phrase with the singular verb phrase fails due to a violation of an agreement constraint. Exponential growth in simple parsing arises because the number of possible assignments to variables can grow exponentially. (*See* **Agreement**)

In the field of artificial intelligence, the desire for richer, more structured systems of knowledge representation and reasoning has led to a generalized notion of features and values known as a *frame*. Frames generalize simple feature systems with a finite set of features and a finite set of values by allowing some features to take frames themselves as values. On top of this frame-based representational system, many forms of constraints can be introduced. These lead to a variety of reasoning systems for resolving the constraints. Of particular interest for natural language is the pure constraint-based representation of natural language grammars, which was first introduced in the system KL-ONE (Brachman and Schmolze, 1985).

Of particular interest is the notion of constraint resolution, which is often known as 'classification'. Early work in classification was applied to parsing, most notably in the system KL-ONE. In this purely constraint-based representation of parsing, features are used to represent phrase-structure trees in the same way in which trees are usually coded in computational data structures. For instance, the simple sentence *the kids laughed*, represented as [S [NP [Det the] [N kids]] [VP laughed]], would be represented in a frame as:

```
CAT: S
DTR1: CAT: NP
  DTR1: CAT: Det
    WORD: the
  DTR2: CAT: N
    WORD: kids
DTR2: CAT: VP
  WORD: laughed                    (2)
```

Note that features like CAT (for syntactic category) take simple atomic values, whereas features like DTR1 and DTR2 (for daughter constituents) take values that are themselves frames. A set of grammatical 'rules' can then be naturally modeled as a disjunctive constraint: every local frame must satisfy one of the phrase-structure rules, including lexical rules. In a logical language, this could be expressed as follows:

```
(CAT:S & DTR1:CAT:NP & DTR2:
  CAT:VP) |
(CAT:NP & DTR1:CAT:Det & DTR2:
  CAT:N) |
(CAT:Det & WORD:(the|a|every|
  some|...)) |
(CAT:N & WORD:(kid|boy|dog|...))   (3)
```

Representation of agreement is straightforward:

```
CAT: MAJOR: NP
     AGR: singular
DTR1: CAT: MAJOR: Det
           AGR: singular
      WORD: every
DTR2: CAT: MAJOR: N
           AGR: singular
      WORD: dog                    (4)
```

The constraint on agreement can be represented using equations between paths of features:

```
CAT:NP & DTR1:CAT:Det & DTR2:
  CAT:N & (CAT:AGR = DTR1:CAT:
  AGR) & (CAT:AGR = DTR2:CAT:AGR)
                                    (5)
```

Pollard and Sag's (1994) head-driven phrase structure grammar (HPSG) has taken this representational strategy to the limit, representing all grammatical structure by means of constraints. But rather than being a naive encoding of phrase structure grammars, HPSG followed the linguistic trend of factoring the rule-specific equations such as equation 5 into general theories of agreement, semantics, unbounded dependencies, etc. A typical constraint in HPSG would enforce agreement between a mother and a head daughter, and would apply to every phrase structure configuration. Another constraint might say that the unresolved dependencies in a phrase must be the union of all unresolved dependencies in the daughters. Thus rather than a disjunction of rules like

$$\text{Rule}_1|\text{Rule}_2|\ldots|\text{Rule}_n \qquad (6)$$

HPSG would have

$$\text{Principle}_1 \;\&\; \text{Principle}_2 \;\&\ldots\&\; \text{Principle}_n \qquad (7)$$

Of course, the principles themselves might involve disjunctions of cases. (*See* **Constraint Satisfaction**)

In general, constraint systems such as the one employed in HPSG are Turing-equivalent computational devices, and as such do not even have decidable parsing problems, much less tractable ones. However, the properties of actual models expressed in feature formalisms are much more tractable – the theory actually picked out by HPSG, for instance, is decidable, as are the other major feature-based linguistic theories. Even so, most parsers for HPSG and related constraint-based grammar formalisms are built on top of efficient parsers for phrase structure, with constraints being maintained on the side to filter search.

More recently, the trend in constraint-based theories of language has been to relax the all-or-nothing nature of constraints. In theoretical linguistics, optimality theory, for instance, relies on defeasible constraints that are ordered by strength (Archangeli and Langendoen, 1997). Within computational linguistics, statistical models have largely supplanted logical ones, with the advantage that constraints are not 'hard', but rather 'soft'. That is, like optimality theory, statistical algorithms search for the 'best' structure. In statistical models, that is assumed to be the one with the highest likelihood. In optimality theory, it is the one violating the fewest high-ranking constraints.

References

Archangeli D and Langendoen DT (eds) (1997) *Optimality Theory: An Overview*. Oxford, UK: Blackwell.

Bar-Hillel Y (1950) On syntactical categories. *Journal of Symbolic Logic* **15**: 1–16.

Barton GE, Berwick RC and Ristad ES (1987) *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.

Barwise J and Etchemendy J (1987) *The Liar: An Essay in Truth and Circularity*. Oxford, UK: Oxford University Press.

Brachman RJ and Schmolze JG (1985) An overview of the KL-ONE knowledge representation system. *Cognitive Science* **9**: 171–216.

Chomsky N (1957) *Syntactic Structures*. The Hague, Netherlands: Mouton.

Chomsky N (1980) *Rules and Representations*. New York, NY: Columbia University Press.

Colmerauer A (1993) Les systèmes-q ou un formalisme pour analyser et synthétier des phrases sur ordinateur. *Traitement Automatique des Langues* **33**: 105–148. [Written in 1970 as a technical report in the University of Montréal.]

Cormen TH, Leiserson CE and Rivest RL (1990) *Introduction to Algorithms*. Cambridge, MA: MIT Press.

Harman G (1963) Generative grammars without transformation rules: a defense of phrase-structure. *Language* **39**: 597–616.

Pollard C and Sag I (1994) *Head-Driven Phrase Structure Grammar*. Stanford, CA: CSLI.

Robinson JA (1965) A machine-oriented logic based on the resolution principle. *Journal of the ACM* **12**: 23–41.

Further Reading

Bird S (1995) *Computational Phonology: A Constraint-Based Approach*. Cambridge, UK: Cambridge University Press. [A purely constraint-based theory of morphophonology.]

Carpenter B (1992) *The Logic of Typed Feature Structures*. Cambridge, UK: Cambridge University Press. [A comprehensive analysis of logical constraints and constraint resolution.]

Manning CD and Schuetze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press. [A thorough overview of statistical approaches to language.]

Saraswat VA (1993) *Concurrent Constraint Programming*. Cambridge, MA: MIT Press. [Discusses a general form of constraint logic programming.]

Sowa J (ed.) (1991) *Principles of Semantic Networks*. San Mateo, CA: Morgan Kaufmann. [An excellent collection of papers on frame-based knowledge representation systems.]

Constraints on Movement

Advanced article

Gert Webelhuth, University of North Carolina, Chapel Hill, North Carolina, USA

CONTENTS

Within a sentence, constituents may appear in the local context of the items they stand in a grammatical relation with or they may move to other positions. Such movement phenomena are subject to a number of grammatical constraints.

INTRODUCTION

Theories of syntactic movement try to capture the systematic relationship between groups of sentences such as

[$_S$ *Sandy* wants to show *those pictures* to Jill] (1)

[$_S$ Those pictures Sandy wants to show to Jill] (2)

[$_S$ *Those pictures* I am told [$_S$ Sandy wants to show to Jill]] (3)

In all three sentences the noun phrase *those pictures* is understood as the thing that Sandy wants to show to Jill, but only in sentence 1 does the noun phrase appear in the position immediately following the verb *show*, which is the canonical position for a noun phrase being interpreted as the thing whose state changes as a result of the event denoted by a verb like *show*. In sentences 2 and 3, the noun phrase appears intuitively too far to the left, at the left edge of a clause, a position frequently occupied by a characteristic set of elements in many languages of the world, namely question phrases (sentence 4), relative pronouns (sentence 5), and expressions receiving particular emphasis (sentence 6):

Which pictures does Sandy want to show to Jill? (4)

The pictures *which* Sandy wants to show to Jill are lying on the floor. (5)

The pictures on the left I like but the ones on the right I find ugly. (6)

In transformational grammar, it is assumed that sentences containing dislocated phrases are generated in two steps by a formal grammar. The first step occurs within the phrase structure component and puts each semantic dependent of a verb (or other part of speech) into a local relationship with that verb. This is often referred to as a 'base-generated' structure. For example, the base-generated structure of sentence 2 would be sentence 7, which is structurally identical to sentence 1:

[$_S$ Sandy wants to show *those pictures* to Jill] (7)

The second step, a 'movement' transformation, moves the NP *those pictures* to the beginning of the clause and thereby creates the visible surface order of sentence 2.

Sentence 1 would also be derived from the basic structure in sentence 7, the difference between sentences 1 and 2 being that the NP *those pictures* stays in its base-generated position in sentence 1 but moves to the left sentence periphery in the process of deriving sentence 2.

Not every potential movement of an expression like the movement of *those pictures* in sentence 2 leads to a grammatical sentence. This is illustrated by examples 8 and 9, and 10 and 11:

Sandy met the photographer who showed *those pictures to Jill*. (8)

**Those pictures* Sandy met the photographer who showed to Jill. (9)

Sandy questioned the assumption that the photographer showed *those pictures* to Jill. (10)

Those pictures Sandy questioned the assumption that the photographer showed to Jill. (11)

There must thus exist constraints on movement that allow sentence 2 to be derived from sentence 7 by movement of *those pictures* but which prevent the NP from moving to the left in sentences 8 and 10. This article will discuss the nature of these constraints.

THEMATIC RELATIONS AND ARGUMENT STRUCTURE

In order to address the question of what constraints on movement exist, one needs a way to tell which expressions have moved in a given sentence and where the movement originated. Rather than postulating sentence 7 as the common base structure of sentences 1 and 2 and deriving sentence 2 from sentence 7 by moving *those pictures* to the left, one might alternatively take the left-peripheral occurrence of the NP in sentence 2 as its base position and derive its position to the right of the verb in sentence 1 by a rightward movement operation. One consideration that favors the original derivation over this hypothetical alternative is that sentence 7 base-generates the moved NP next to the verb that it is semantically related to, in the following sense.

Verbs typically refer to events that contain a certain number of participants that stand in a relation specified by the verb. In the case at hand, the verb *show* refers to an event of showing that has three participants: somebody shows something to somebody. Other verbs may refer to events that contain fewer participants: thus, the verb *sneeze* refers to an event with just a single participant, and the verb *peel* refers to a two-place event where somebody peels something. However, across events, participants may share similar properties. For instance, the participant who peels potatoes is more actively involved in the peeling event (in the sense that he or she brings the event into existence) and as a result of this action changes the state of the potatoes. Likewise, whoever shows pictures to somebody else is typically more actively involved in the action and changes the state of the other participants of the event. There has been considerable effort in linguistics to develop a classification of participant roles.

Table 1 contains a number of participant roles that are widely used in the description of word meanings. (The 'theme' role assumed a prominent position when the concept of participant roles came into wide use, and therefore participant roles are often referred to as 'thematic' or 'theta' roles.)

Verbs (and other parts of speech) can be classified according to how many participants their events involve and which participant roles must be present. This information is often referred to as the 'argument structure' of a verb. Table 2 lists some verbs and the argument structures in which they are typically used.

Most verbs have more than one argument structure, i.e. they are compatible with more than one combination of arguments. For instance, *open* can be used with a single patient argument as in *the door opened* or with both an agent and a patient as in *Sue opened the door*.

LINKING BETWEEN LEXICAL AND SYNTACTIC STRUCTURE

The argument structure of a verb strongly influences the syntactic configuration in which its arguments are realized, as can be seen by examining the example sentences in Table 2. In each case, the subject precedes the verb. The intransitive verbs in (a), (b) and (c), i.e. those with a single participant role in their argument structure, make their single argument into a subject. In verbs with multiple participant roles the choice of subject is determined by the following two principles. (Argument-linking principles only apply to active verbs. The linking of arguments in passive verbs is determined by the

Table 1. Some common participant roles

Name	*Abbreviation*	*Approximate definition*
Agent (Actor)	Ag	A participant most actively involved in the event or a participant who causes a change of state in other participants.
Patient (Theme)	Pt	A participant who undergoes a change of state in the event.
Experiencer	Exp	A participant who undergoes a mental or perceptual change of state.
Source	So	A participant who is the source of a transfer or movement.
Goal	Go	A participant who is the goal of a transfer or movement.

Table 2. Typical argument structures of verbs

Verb	*Argument structure*	*Example sentence*
(a) *bark*	(Ag)	The dog *barked*.
(b) *die*	(Pt)	The dog *died*.
(c) *hallucinate*	(Exp)	The witness *hallucinated*.
(d) *kick*	(Ag, Pt)	The horse *kicked* the dog.
(e) *see*	(Exp, So)	The horse *saw* the dog.
(f) *show*	(Ag, Pt, Exp)	The owner *showed* the dog *to* her neighbor.
(g) *give*	(Ag, Pt, Go)	The owner *gave* the key *to* her neighbor.

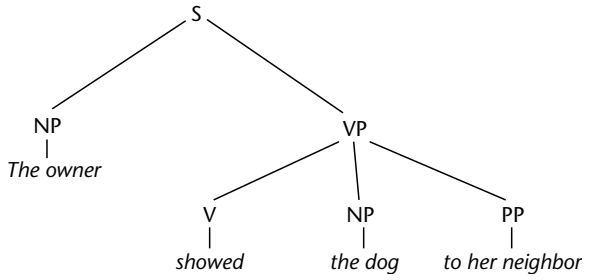


Figure 1. A syntactic structure determined by the argument structure (Ag, Pt, Exp).

linking in the active verbs from which the passive verbs are derived. See Dowty (1991), Levin (1993), and Davis (2001) for further discussions of argument linking.)

1. The participant that is most agent-like is realized as the subject in transitive verbs.
2. The participant that is most patient-like is realized as the direct object of transitive verbs.

In 'X-bar theory', subjects are analyzed as specifiers and direct objects as the NP immediately following the verb within the verb phrase. The argument structure (Ag, Pt, Exp) for the verb *show* thus determines the syntactic structure shown in Figure 1 for sentence (f) in Table 2.

The canonical position of the agent is before the verb, whereas the patient canonically follows the verb. This makes the order in sentence 1 a more natural candidate for the base order than the order in sentence 2. Indeed, sentence 1 is felt as more neutral by native speakers of English. The word order in sentence 2 obtains only under certain constructional conditions, as illustrated in sentences 4, 5 and 6.

ARGUMENTS FOR POSITING MOVEMENT

The motivation for analyzing some sentences in terms of base-generating an expression in one place and subsequently moving it to another place is that it elegantly captures certain grammatical generalizations which would otherwise be hard to account for. One class of such generalizations relates to the issue of argument structure which we have already discussed. For example, sentences based on the verb *show* are ungrammatical if the theme participant remains unexpressed:

*The owner showed to her neighbor. (12)

The ungrammaticality of example 12 can be captured elegantly by requiring that in the base structure of a sentence, the argument structure of every word must be fully realized by respecting linking constraints like the two principles given above. The verb *show* has a patient argument, which according to the second of those principles should be expressed as the first postverbal NP; this condition is not met in example 12, which is therefore ungrammatical. By deriving sentence 14 from 13 and sentence 16 from 15, we have an immediate explanation of why they are grammatical, in contrast to example 12:

Which dog did the owner show to her neighbor? (13)

The owner showed which dog to her neighbor. (14)

Which dog did Sandy say the owner showed to her neighbor? (15)

Sandy said that the owner showed *which dog* to her neighbor. (16)

In sentences 14 and 16, the structures from which sentences 13 and 15 are derived, the patient argument of *show* appears immediately behind the verb in base structure. In example 12, however, no patient NP appears in the sentence at all, either in postverbal position or in a possible moved position, and hence this syntactic structure does not fully realize the argument structure of the verb *show*.

Another powerful argument for movement operations comes from the locality of certain grammatical relations, such as agreement. For instance, present-tense verbs in English systematically agree with their subjects in person and number:

The *owner shows*/**show* the dog to her neighbor. (17)

But they do not systematically agree with any expression outside their clause, because subject–verb agreement is a local relationship:

*The *onlookers* think that the owner *show* the dog to the neighbors. (18)

Yet, in sentences like 19 where the subject of the subordinate clause headed by *show* appears at the beginning of the main clause, the verb *shows* still agrees with the dislocated constituent in person and number, just as it does in sentence 20 where *the owner* stays within the subordinate clause:

Which owner do the onlookers think [$_S$ *shows*/**show* the dog to the neighbors]? (19)

The onlookers think [$_S$ *the owner shows*/ **show* the dog to the neighbors]. (20)

If subject agreement relations are determined before *wh*-extraction, then the verb *show* should agree with its subject in both sentences 19 and 20, no matter whether that subject is subsequently moved to a position in the main clause from which agreement with the embedded verb is otherwise impossible.

The co-occurrence restrictions on anaphoric pronouns provide similar motivation for movement. These expressions must find an antecedent within the immediate clause containing them:

The journalists said [$_S$ *the youngest participants* only paid attention to *themselves*]. (21)

In this sentence, *themselves* can only be understood as referring to the subject of the subordinate clause *the youngest participants*. If the reading were intended whereby the youngest participants only paid attention to the journalists, a personal pronoun would have to be substituted for *themselves*:

The journalists said [$_S$ the youngest participants only paid attention to *them*]. (22)

Compare sentence 23 to sentences 21 and 22:

Who did the journalists say [$_S$ only paid attention to *themselves*]? (23)

In this sentence, the NP *who*, moved from the subordinate clause to the initial position of the main clause, must still be interpreted as the antecedent of the anaphoric pronoun *themselves*, even though it no longer occurs within the immediate clause containing the anaphor. In fact, *who* must be chosen as the antecedent even though it is linearly further away from *themselves* than the subject of the main clause *the journalists*. All of these facts fall into place when we examine the base structure that underlies sentence 23:

The journalists said [$_S$ *who* only paid attention to *themselves*].

Of the two potentially available antecedents of *themselves* in sentence 23, only *who* is contained within the immediate clause containing the anaphor in base structure. If the co-occurrence restriction on anaphoric pronouns applies before the kinds of movement operations discussed here, then the empirical facts about sentence 23 immediately follow, because *who* has been moved out of the subordinate clause whereas *the journalists* was never contained in the same clause as *themselves* and therefore is not capable of anteceding it. The co-occurrence options of moved expressions in their base position should carry over to the new positions they occupy following movement operations.

ROSS'S CONSTRAINTS

Ross (1967), reacting to earlier proposals by Chomsky (1964), presented the first systematic investigation of movement constraints on a large scale. One movement rule that is systematically constrained in many languages is *wh*-movement, the rule that moves a *wh*-expression to the left periphery of a sentence in the formation of an interrogative clause. As Ross showed, this rule may operate in an unbounded fashion in English, i.e. the *wh*-expression can move an arbitrary number of clauses to the left:

What did Bill buy _? (25)

What did you force Bill to buy _? (26)

What did Harry say you had forced Bill to buy _? (27)

What was it obvious that Harry said you had forced Bill to buy_? (28)

And so on. Yet, we find that the following examples are all ungrammatical:

*What did Bill buy potatoes and _? (29)

*What did that Bill wore _ surprise everyone? (30)

*What did Cindy believe the claim that Otto was wearing _? (31)

*Whose did you find _ book? (32)

*What$_1$ did Jill wonder where$_2$ Sandy put $_{-1}$ $_{-2}$? (33)

The underscores in these examples mark the base positions of the moved *wh*-expressions and point to the intended interpretations. For instance, the intended interpretation of example 30 is similar to that of the echo-question in example 34:

That Bill wore WHAT surprised everyone? (34)

Clearly, the paradigm in examples 25 to 33 cannot be captured by putting an upper limit on the number of words or sentences that a *wh*-expression can move across on its way to the left periphery of the sentence. The movement path of the *wh*-word in the grammatical sentence 28 is much longer than the one we find in the ungrammatical sentence 32.

Ross proposed that what matters is the structural organization of the string that *wh*-movement crosses. For instance, to distinguish between examples 25 and 29, Ross formulates the 'coordinate structure constraint' namely, that in a coordinate structure, no conjunct may be moved, nor may any element contained in a conjunct be moved out of that conjunct.

In example 29, *what* is contained in the coordinate noun phrase *potatoes and what*:

Bill bought [$_{NP\text{-}1}$ [$_{NP\text{-}2}$ potatoes] and [$_{NP\text{-}3}$ what]] (35)

The coordinate structure constraint prevents the *wh*-movement transformation from pulling NP-3 out of the coordinate structure NP-1, and so example 29 is ungrammatical.

Example 30 may be contrasted with sentences like:

What did Cindy say that Bill wore _? (36)

These examples are structurally different in that *what* is extracted from a subject clause in example 30 but from an object clause in example 36. Ross formulates the 'sentential subject constraint' to prohibit any movement operation from extracting an element from a sentential subject. Example 31 involves a violation of yet another constraint, the 'complex NP constraint' which blocks extraction from complement clauses to nouns. Compare example 31 with the similar sentence:

What did Cindy believe that Otto was wearing _? (37)

What makes sentence 37 grammatical in contrast to 31 is that the clause *that Otto was wearing what* from which the interrogative element is extracted is a complement to a verb in the grammatical context in sentence 37 but a complement to the noun *claim* in 31, violating the complex NP constraint.

Ross formulates a substantial number of constraints on transformations, as well as empirical insights which have been elaborated on in later work. Ross's doctoral thesis is generally considered one of the most insightful and influential works in the history of generative grammar.

THE CONDITIONS FRAMEWORK

Chomsky (1973) sought to systematize a number of constraints on transformations, including constraints on movement and constraints on semantic interpretation. This effort to postulate general structural constraints that can replace groups of language-particular or construction-specific constraints marked the beginning of a research program that was highly influential and led to the theories of 'government' and 'binding' and the 'minimalist program' which dominated syntactic theorizing in the 1980s and 1990s. The three most far-reaching constraints that Chomsky proposes are the 'tensed sentence condition', the 'specified subject condition', and the 'subjacency condition'.

The tensed sentence condition captures the differences between examples 38, 39 and 40, under the assumption that these are derived from 41, 42 and 43 respectively by movement of the quantifier *each* to the right:

The candidates hated *each* other. (38)

The candidates expected [$_S$ *each* other to win] (39)

*The candidates expected [$_S$ that *each* other would win] (40)

The candidates *each* hated the other(s). (41)

The candidates *each* expected [$_S$ the other(s) to win] (42)

The candidates *each* expected [$_S$ that the other(s) would win] (43)

In example 40, unlike 38 and 39, the quantifier must move into a tensed sentence. In 39, the complement

of *expected* is analyzed as a sentence in order to give this verb a uniform subcategorization frame. Yet, in 39 the quantifier moves into a nonfinite sentence and in 38 it moves into the noun phrase *the other*.

The tensed sentence condition prohibits a rule from applying to two constituents if one of the constituents is contained within a tensed sentence and the other one occurs outside that sentence.

The tensed sentence condition needs to be supplemented by the specified subject condition, since subjects of complement clauses behave differently from non-subjects, as a comparison between examples 38, 39 and 40 and the examples below illustrates:

The candidates expected [$_S$ PRO to defeat *each* other] (44)

*The candidates expected [$_S$ the soldier to shoot *each* other] (45)

The candidates$_1$ *each* expected [$_S$ PRO$_1$ to defeat the other] (46)

The candidates *each* expected[$_S$ the soldier to shoot the other] (47)

Both 44 and 45 are nonfinite, so the tensed sentence condition is unable to account for the grammaticality contrast. What differentiates them is that, in its rightward movement, *each* must cross the lexicalized subject *the soldier* in 45, whereas in 44 it must only cross the abstract anaphoric subject PRO. The specified subject condition prohibits a transformation from relating two constituents if they are separated by a specified subject, where lexical subjects always count as specified.

Finally, the subjacency condition generalizes a number of earlier conditions on movement out of sentences and noun phrases. This will be discussed below.

BARRIERS

Chomsky (1986) attempts to unify all movement constraints in terms of a phrase-structurally defined notion of 'barrier', which is also meant to play a role in the theory of government. Chomsky starts from the assumption that complements generally play a more active role in extraction constructions than either subjects or adjuncts. Consider the following examples:

Who did [$_{IP}$ Mary find [$_{NP}$ a picture of _]]? (48)

*Who did [$_{IP}$ [$_{NP}$ a picture of _] scare the children]? (49)

*Who did [$_{IP}$ Mary laugh [$_{PP}$ when she found a picture of _]]? (50)

??Which car did Mary sleep [$_{PP}$ while Jill fixed _]? (51)

*Which mechanic did Mary sleep [$_{PP}$ while _ fixed the car]? (52)

*How quickly did Mary sleep [$_{PP}$ while Jill fixed the car _]? (53)

(Note the intended interpretation in example 53: *how quickly* modifes *fixed*.)

Examples 48, 49 and 50 show that it is easier to extract a constituent from a complement than from a subject or an adjunct. In 49 *who* has been extracted from the subject of the sentence, and in 50 from an adjunct, and both sentences are worse than 48. Examples 51, 52 and 53 show that under certain conditions it is also easier to extract complements than to extract subjects or adjuncts. Each of these examples involves extraction from an adjunct, but the violation incurred by the subject in 52 and by the adjunct in 53 is felt to be stronger than the violation incurred by the complement in 51.

Complements are thus both easier to extract and easier to extract from, whereas subjects and adjuncts are harder to extract and harder to extract from. Chomsky develops a complex web of constraints and definitions in an attempt to capture these generalizations. Crucial to his approach are certain assumptions about the X-bar theory of phrase structure, including the tree structure shown in Figure 2.

One problem with these assumptions is that Chomsky requires this full abstract structure to be

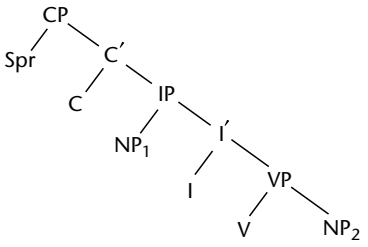


Figure 2. Chomsky's proposed uniform structure for sentences. 'C' stands for complementizers like *that* and *whether*; 'Spr' is a 'landing site' for *wh*-operators of any part of speech; 'I' stands for inflectional elements and auxiliaries; 'NP$_1$' is the clausal subject position; and 'NP$_2$' is the clausal direct object position.

present even in sentences where the structure is not supported by observable morphological or lexical material. Thus, even simple sentences like *Mary worked* are given abstract C and I nodes, even though there is no empirical evidence for the existence of a complementizer or a self-standing inflectional element separate from the verb in these kinds of sentences.

In order to account for the contrasts in examples 48 to 53, which show that it is possible to extract from a complement but impossible to extract from subjects or adjuncts, Chomsky formulates several assumptions. Firstly, he defines 'blocking categories' as maximal projections that are not theta-governed by a lexical head. Here, lexical heads include the major parts of speech (verb, adjective, noun, etc.) but specifically exclude the minor parts of speech (complementizer and inflection).

IP is governed by C, while VPs and subjects are governed by the I. Adjuncts may in principle be governed by lexical heads, but they are, by definition, not theta-marked. Thus, IP, VP, subjects and adjuncts are blocking categories. Complements are the only class of expressions that are always theta-marked (this is a requirement of Chomsky's (1986) 'projection principle') and they are governed by a lexical head. Thus, complements are not blocking categories.

According to this definition, then, VPs should be barriers to movement, but this is empirically incorrect. Chomsky therefore proposes a mechanism that allows the 'barrierhood' of VP to become voided. He defines barriers to movement as follows: IP is a barrier for everything within its subject or a sentence-level adjunct; subjects and adjuncts are barriers for everything contained in them; and CP is a barrier for everything contained within IP.

The ungrammaticality of examples 49 and 50 now follows, given the subjacency condition, which stipulates that no movement step is allowed to cross more than one barrier. In 49, the subject NP is a barrier and so is the IP. The extraction of *who* from [$_{IP}$ [$_{NP}$ *a picture of* _] …] thus crosses two barriers, implying that the sentence is ungrammatical. Example 50 is similar, except that in this case the second barrier is the adjunct PP. In contrast, example 48 is grammatical, since the complement NP is not a barrier, given that it is theta-marked by a lexical head.

Chomsky also addresses the data in examples 51, 52 and 53. While adjuncts are generally hard to extract from, the judgment of any given extraction depends on the kind of element that is extracted. In English, extraction of complements from movement islands often yields a slightly more favorable judgment than extraction of subjects or adjuncts from the same kind of island. Chomsky uses the 'empty category principle' (Chomsky, 1981) in an attempt to solve this problem. According to this principle, traces must be properly governed, i.e. they must be either theta-governed by an X^0 head or antecedent-governed. To be antecedent-governed, no barrier is allowed to intervene between the trace and its antecedent (i.e. the moved expression or another one of its traces).

From this definition of proper government it follows that the trace of a complement is always properly governed and hence need not be antecedent-governed. Subject and adjunct traces, in contrast, are not governed by X^0 heads that theta-mark them and thus need to be antecedent-governed in order to be properly governed. Extraction of a complement from an adjunct thus yields a violation only of the subjacency condition, as shown earlier, whereas extraction of a subject or an adjunct under the same conditions yields a violation of the empty category principle in addition to the subjacency violation incurred by complements. Examples 52 and 53 should thus be stronger violations than example 51.

The barriers framework has a number of conceptual and empirical weaknesses. Ultimately, Chomsky abandoned it to pursue even more abstract phrase-structural theories. Other researchers turned their back on the principles-and-parameters framework, believing that the syntactocentric, derivation-driven design of the theory as a whole, rather than individual definitions or principles, were to blame for its chronic empirical problems. The proliferation of abstract phrase-structure categories has already been mentioned. Furthermore, there are large amounts of data that are incompatible with the main ideas of the Barriers framework. There are languages in which the subject–object and object–adjunct asymmetries that Chomsky's theory is specifically designed to derive do not obtain. For instance, in Swedish subjects can be extracted much more easily than Chomsky's theory predicts. Moreover, with respect to adjuncts, von Stechow and Sternefeld (1988, p. 371) present constructions in German that show the precise opposite of what Chomsky's theory predicts: adjuncts can be extracted from certain phrases that do not allow complements to escape. In fact, even English is much less uniform than Chomsky's theory would lead one to expect. Santorini (2001) has collected a list of attested examples that violate a number of prominent constraints on movement that have been postulated, including the following

example which contains an extraction from an adjunct:

a scenario that government agencies are spending billions of dollars preparing for (54)

Other data confirm that principles-and-parameters theories are an oversimplification of the data. Culicover (1999) presents the following contrast, illustrating that whereas prepositions generally allow their complement to move away in English, the preposition *since* is an idiosyncratic lexical exception to this generalization which cannot be captured by setting a category-wide parameter:

*Which party hasn't John called since _? (55)

As successive reforms of Chomsky's parametric framework have been unable to address these sorts of detailed empirical problems, constraint-based lexicalist theories of grammar have gained increasing acceptance in linguistics, psycholinguistics, language acquisition, and computational linguistics. We will now discuss such theories.

NON-TRANSFORMATIONAL APPROACHES

Non-transformational approaches to grammar are generally motivated by a belief that, as Chomsky has sought more elegant answers to a restricted set of problems, his theories have become less empirically realistic and the theoretical constructs he invokes have become less open to theory-neutral verification by other researchers. In the 1980s and 1990s, a number of Chomskyan linguists moved on to other theories, in particular constraint-based lexicalist theories.

One problem that systematically arises with Chomsky's theory is that the broad generalizations he attempts to derive are counterexemplified by small classes of items or individual words. In order to circumvent this problem, researchers have developed tools that are capable of deriving grammatical generalizations of variable scope, including the fully productive generalizations that Chomsky's theory can handle, but also semi-productive patterns and completely idiosyncratic phenomena.

Currently the most credible alternatives to Chomsky's syntactocentric principles-and-parameters approach invoke a sophisticated set of lexical and constructional tools in the analysis of grammatical phenomena, including movement phenomena. Most prominent among these theories are 'Head Driven Phrase Structure Grammar' (Pollard and Sag, 1994), 'Lexical Functional Grammar' (Bresnan, 2000), and 'Construction Grammar'. The remainder of this article will focus on the approach to movement proposed in the first of these theories.

Head Driven Phrase Structure Grammar developed from 'Generalized Phrase Structure Grammar' (Gazdar *et al*., 1985). The desire for a linguistic theory that is simultaneously formally precise, empirically accurate and broad, and computationally implementable in an efficient manner, led Gazdar to avoid all transformations in his theory, including those that move constituents from a base-generated position to another position in a tree-altering fashion. Instead, he proposed that 'movement' is a metaphor for a featural dependency in a tree, encoding the information that a constituent that in principle can be expressed in one part of a tree can also be expressed elsewhere in that tree.

Sag and Fodor (1994) and Bouma *et al*. (2001) have combined Gazdar's theory of extraction with a theory of argument realization that gives words fine control over the syntactic realization options of their arguments. Together with constructional constraints on the flow of argument information in a syntactic tree, these assumptions yield an empirically powerful theory of extraction phenomena which its adherents believe to be capable of deriving the same broad generalizations that Chomsky's syntactocentric theory is able to capture, without sacrificing coverage of semi-productive and idiosyncratic phenomena.

Bouma *et al*. make use of the idea that a word can be represented by a data structure having an argument structure and a valence. The argument structure of a word is involved in interpretative properties, including the binding theory, while the valence determines the surface-syntactic context that the word must appear in (for principled reasons, the theory does not countenance any unobservable syntactic levels, such as Chomsky's D-structure or LF, nor any unobservable signs such as traces and phonologically empty heads). Part of the individual grammar of a language is a set of principles that determine how each of a word's arguments may be realized on the surface. Thus, if the language permits arguments to be realized morphologically, then they will not be projected into the syntax, which makes the dependence on phonologically empty subjects superfluous and keeps syntactic structures concrete and open to theory-independent verification. Extraction phenomena are handled in terms of permissible mappings between argument structure and valence. Just as words have the ability to determine

whether or not their arguments can be spelled out morphologically, they have the ability to constrain the syntactic realization options of their arguments: arguments that appear in a word's valence are spelled out in the word's local syntactic domain, whereas the descriptions of arguments that appear in the word's GAPS specification are percolated upward in the syntactic tree until this information appears in a syntactic configuration where a filler sign is found which is featurally compatible with the argument description. For the verb *show* in examples 1 and 2, an argument structure like the following would be postulated:

$$\text{ARG-ST} \ < \text{NP}_{\text{agent}}, \ \text{NP}_{\text{patient}}, \ \text{PP}_{\text{goal}} > \qquad (56)$$

Since *show* permits each of its arguments to be realized either locally or nonlocally, its valence and gaps specification may be any of the following, among others:

$$\begin{array}{l}\text{SUBJ} \ < \text{NP}_{\text{agent}} > \\ \text{COMPS} \ < \text{NP}_{\text{patient}}, \text{PP}_{\text{goal}} > \\ \text{GAPS} \ \{\}\end{array} \qquad (57)$$

$$\begin{array}{l}\text{SUBJ} \ < \text{NP}_{\text{agent}} > \\ \text{COMPS} \ < \text{PP}_{\text{goal}} > \\ \text{GAPS} \ \{\text{NP}_{\text{patient}}\}\end{array} \qquad (58)$$

$$\begin{array}{l}\text{SUBJ} \ < \text{NP}_{\text{agent}} > \\ \text{COMPS} \ < \text{NP}_{\text{patient}} > \\ \text{GAPS} \ \{\text{PP}_{\text{goal}}\}\end{array} \qquad (59)$$

Whereas subjects and complements must be realized within the valence domain of the verb, the information in GAPS is percolated up the tree in a fashion that allows fillers to be found for these gaps outside the verb's local valence domain. This is how long-distance dependencies arise. Specification 57 is used to generate surface forms like sentence 1, where each of the three arguments of *show* is generated within its local valence domain. Specification 58 would underlie a sentence like sentence 2, where the subject and the PP object remain within the valence domain, but the direct object NP is realized as a filler at the beginning of the clause. Specification 59 would underlie such sentences as *To Jill I want to show these pictures* or *to whom do you want to show these pictures*, where the PP argument is realized at a distance.

This theory gives words full control over whether an argument may or must be realized within the local valence domain of the word. It is thus predicted that, as in other grammatical domains (e.g. inflection), some languages will be very homogenous in their constraints, whereas others will be less so. It is even predicted that different head types within a language may behave differently. There is impressive evidence to support this prediction. For instance, whereas verbs in English permit their specifier (their subject) to be extracted, nouns never do, as example 32 demonstrates. This generalization can be captured by imposing featural well-formedness conditions on the permissible relationship between members of the parts of speech noun and verb and their respective specifier arguments. The kind of lexical idiosyncrasy that is represented by example 55 in the extraction domain (i.e. that *since* does not allow its NP argument to be realized in GAPS) is parallel to idiosyncratic lexical requirements in other domains (e.g. that a verb exceptionally requires a different case on its direct object than most other transitive verbs in the language).

Other differences between languages and constructions can be captured elegantly by constraining the values of the GAPS attribute in different parts of the syntactic tree or in different constructions. For instance, in languages where extraction from subjects is impossible, it suffices for the grammar to contain a constraint that the GAPS value of any subject be the empty set. An analogous constraint will prevent extraction from adjuncts or *wh*-islands where these constraints are empirically called for. Finer distinctions can be drawn by imposing constraints on the content of GAPS in specific constructions and languages. Thus, it would be easy to derive the generalization that arguments are easier to extract from some constructions whereas adjuncts are easier to extract from others. Universal tendencies that are not the result of parsing or memory preferences can be encoded in terms of the grammatical archetypes proposed in Ackerman and Webelhuth (1998), which are conceived of as the set of grammatical concepts that guide the language learner's acquisition process.

References

Ackerman F and Webelhuth G (1998) *A Theory of Predicates*. Stanford, CA: CSLI Publications.

Bouma G, Malouf R and Sag IA (2001) Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory* **19**: 1–65.

Bresnan J (2000) *Lexical-Functional Syntax*. Oxford: Blackwell.

Chomsky N (1964) *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky N (1973) Conditions on transformations. In: Andersons S and Kiparsky P (eds) *Festschrift for Morris Halle*, pp. 232–286. New York, NY: Holt, Rinehart and Winston.

Chomsky N (1981) *Lectures on Government and Binding*. Dordrecht: Foris.
Chomsky N (1986) *Barriers*. Cambridge, MA: MIT Press.
Culicover P (1999) *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford: Oxford University Press.
Davis T (2001) *Linking by Types in the Hierarchical Lexicon*. Stanford, CA: CSLI Publications.
Dowty D (1991) Thematic proto-roles and argument selection. *Language* **67**: 547–619.
Gazdar G, Klein E, Pullum GK and Sag IA (1985) *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press/Oxford: Blackwell.
Levin B (1993) *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.
Pollard C and Sag IA (1994) *Head Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.
Ross JR (1967) *Constraints on Variables in Syntax*. PhD thesis, MIT. [Reproduced by the Linguistics Club of Indiana University.]
Sag IA and Fodor JD (1994) Extraction without traces. In: *Proceedings of the Thirteenth Annual Meeting of the West Coast Conference on Formal Linguistics*, pp. 365–384. Stanford, CA: CSLI Publications.
Santorini B (2001) *(Un)expected Movement*. http://www.ling.upenn.edu/∼ beatrice/examples/movement.html
von Stechow A and Sternefeld W (1988) *Bausteine Syntaktischen Wissens*. Opladen, Germany: Westdeutscher, Verlag.

Construction Grammar

Introductory article

Adele E Goldberg, University of Illinois, Urbana, Illinois, USA

CONTENTS

Construction Grammar is a linguistic theory concerned with the nature of speakers' knowledge of language. Like traditional grammars, Construction Grammar takes the basic units of language to be form–meaning pairings, or constructions.

CONSTRUCTIONS

A *construction* is defined as a pairing of form with meaning/use, such that some aspect of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established as existing in the language. On this view, phrasal patterns, including the constructions of traditional grammarians, such as relative clauses, questions, locative inversion, and so on, are given theoretical status. Words (or really, *morphemes*) are also constructions, according to this definition, since their form is not predictable from their meaning or use. Given this, it follows that the mental dictionary or *lexicon* is not neatly delimited from the rest of grammar, although phrasal constructions differ from lexical items in their internal complexity.

Both phrasal patterns and lexical items are stored in an extended 'constructicon'. Elements within the constructicon vary in degrees of idiomaticity. At one end of the idiomaticity continuum, we find very general, abstract constructions such as the subject–predicate construction; at the other end, we find simple lexical items and constructions with all of their lexical fillers specified but with noncompositional meanings (e.g. 'kick the bucket'). In between, we find the full range of possibilities: for example, idioms which have freely fillable positions (e.g. 'keep/lose X's cool'), compositional collocations with fixed word order (e.g. 'up and down'), phrasal patterns that are only partially productive (e.g. the English ditransitive), and phrasal patterns which are partially morphologically specified (e.g. 'The *X*er, the *Y*er', as in 'The less it rains, the better the potatoes').

Construction Grammar shares with several other current theories, including Head-Driven Phrase Structure Grammar, Cognitive Grammar, and Montague Grammar, the basic and fundamental idea that the construction (or *sign*) is central to an account of language. This view of grammar can be contrasted with the claim made by Principles and Parameters theories that constructions are entirely epiphenomenal, a mere by-product of the interaction of the principles of Universal Grammar, once the values of the parameters are fixed. Although most aspects of language are highly motivated, in the sense that they are related to other aspects of the grammar and are non-arbitrary, Construction Grammar holds the view that much of language is idiosyncratic to varying degrees and must therefore be learned.

Declarative, Monostratal Representation

A given sentence is licensed by the grammar if and only if there exists in the language a set of constructions which can be combined (or superimposed) to produce an accurate representation of the surface structure and semantics of that sentence. An ambiguous sentence is a sentence for which there exists more than one set of constructions that can be assembled to produce a possible representation. Constructions are represented declaratively, and any constructions which do not conflict may be combined to give rise to grammatical expressions. Thus Construction Grammar is monostratal: no derivations are posited.

Typically, particular sentences (or *constructs*) instantiate several constructions simultaneously. For example, sentence (1) below instantiates the subject–predicate construction, the ditransitive construction, the determiner construction ('the letter'), the past tense morphological construction ('fax-*ed*'), and five simple morphological constructions, corresponding to each word in the sentence:

Elena faxed Ken the letter. (1)

Integrated Information

Conventionalized aspects of both meaning and use are directly related to particular syntactic patterns within individual constructions. Thus, Construction Grammar does not assume that syntax is generally isolated or isolatable from semantics or conditions of use. Construction Grammar also eschews a strict division between the pragmatic and the semantic. 'Frame-semantic' (encyclopedic) meaning is considered fundamental to an adequate understanding of linguistic entities, and as such is integrated with more traditional definitional characterizations. Generalizations about particular arguments being topical, focused, inferable, and so on, are also stated as part of the constructional representation. Facts about the use of entire constructions, including register, dialect variation, etc., are stated as part of the construction as well. Thus a construction may be posited because of something not strictly predictable about its frame-semantics, its packaging of information structure, or its context of use.

Relations among Constructions within a Language

Constructions do not form an unstructured set, but rather a highly integrated system, based on general principles of categorization. Constructions are typically closely related to other constructions, and are, in that sense, not arbitrary. Generalizations across constructions are captured within the theory via an inheritance hierarchy, which allows shared structure to be represented.

For example, an abstract *Left Isolate* construction is inherited by several different constructions, exemplified by the following:

a. the woman who she met yesterday (restrictive relative clause)
b. Abby, who she met yesterday (nonrestrictive relative clause)
c. Bagels, I like. (topicalization)
d. What do you think she did? (main clause nonsubject *wh* − question) (2)

Each of these patterns – restrictive and nonrestrictive relative clauses, topicalization, and *wh*-questions – requires a distinct construction of its own, owing to its particular formal and pragmatic properties. But each inherits from the more general Left Isolate construction, which specifies the properties that are shared. In particular, this construction has two 'sisters', with the specification that the left sister satisfies the valence requirement of some predicator at an undefined depth in the right sister; the right sister is a maximal verb phrase, with or without a subject. Thus the Left Isolate construction serves to capture the generalizations across these various patterns.

Formalization

Many practitioners of this theory have adopted the use of a unification-based formalism in order to rigorously detail the specifications of particular constructions. Thus each construction is represented by an Attribute–Value Matrix (AVM). Each attribute can have at most one value. Attributes may be *n*-ary, or may be feature structures themselves.

Any pair of AVMs can be combined to license a particular expression, as long as there is no value conflict on any attribute. When two AVMs unify, they map onto a new AVM, which has the union of attributes and values of the two original AVMs.

RESEARCH FOCUS

Data

Research in Construction Grammar has emphasized the importance of attested data, gathered from discourse or corpora. At the same time, Construction Grammarians routinely supplement corpus data with data gained from introspection, one obvious reason being that corpora do not contain sentences marked as unacceptable. Another source of data comes from psycholinguistic experimentation.

Full Coverage: Lexical Semantics and Marked Constructions

There has been a focus on the semantics and distribution of particular lexical items within the framework, owing to the belief that the rich semantic/pragmatic constraints on individual words or idiomatic phrases reveals much about our knowledge of language. There has been a great deal of attention paid to marked constructions within the theory. For example, consider the Covariational Conditional construction, exemplified by 'the more you think about it, the less you understand'. Independent knowledge of 'the' and grammatical comparison will not directly predict that this relevant

class of expressions will exist or have exactly the form and meaning they have; therefore a distinct construction is posited. Other examples of marked constructions include the 'What's *X* doing *Y*?' construction, exemplified by sentences such as 'What's that fly doing in my soup?', and the Nominal Extraposition construction, e.g. 'It's amazing the difference!'.

As these examples indicate, Construction Grammar aims to account for the full range of facts of any language, without assuming that a particular subset of the data is part of a privileged 'core'. Researchers argue that marked constructions shed light on more general issues, and serve to illuminate what is required for a complete account of the grammar of a language. Construction Grammarians takes the point of view that the ordinary patterns of grammar do not differ qualitatively from these sorts of quantitatively more complex constructions.

Argument Structure Constructions

In many current linguistic theories, the form and general interpretation of basic sentence patterns of a language are taken to be determined by semantic and/or syntactic information specified by the main verb in the sentence. The sentence patterns given in (3) and (4) indeed appear to be determined by the specifications of 'give' and 'put' respectively:

Chris gave Pat a ball. (3)

Pat put the ball on the table. (4)

'Give' is a three-argument verb and is expected to appear with three complements corresponding to agent, recipient, and theme. 'Put', another three-argument verb, requires an agent, a theme, and a location, and appears with the corresponding three complements in (4). However, while (3) and (4) represent perhaps the prototypical case, the interpretation and form of sentence patterns of a language are not reliably determined by independent specifications of the main verb. For example, it is implausible to claim that 'sneeze' has a three-argument sense, and yet it can appear as in (5):

She sneezed her tooth across the yard. (5)

The following attested examples similarly involve sentential patterns that do not seem to be determined by independent specifications of the main verbs:

'She smiled herself an upgrade.' (Douglas Adams, *Hitchhiker's Guide to the Galaxy*; Harmony Books) (6)

'We laughed our conversation to an end.' (J. Hart, *Sin*; 1992, Ivy Books) (7)

Moreover, verbs typically appear with a wide array of complement configurations. Consider the verb 'sew' and the various constructions in which it can appear (labeled in parentheses):

a. Pat sewed all afternoon. (intransitive)
b. Chris sewed a shirt. (transitive)
c. Pat sewed Chris a shirt. (ditransitive)
d. Pat sewed the sleeve shut. (resultative)
e. Pat sewed a button onto the jacket. (caused − motion)
f. Chris sewed her way to fame and fortune. (way − construction) (8)

In Construction Grammar, instead of predicting the surface form and interpretation solely on the basis of the verb's independent specifications, the lexical verb is understood to combine with an argument structure construction (e.g. the ditransitive, resultative, the caused-motion construction, etc.). Verbs constrain the type of argument structure constructions with which they can combine by their frame-specific semantics and particular obligatory roles, but they typically can combine with constructions in several ways.

It is the argument structure constructions that provide the direct link between surface form and general aspects of the interpretation such as something causing something else to move, someone causing someone to receive something, something moving somewhere, someone causing something to change state, and so on. The argument structure constructions, which provide the basic sentence patterns of a language, directly reflect these types of basic frames of experience. That is, the skeletal patterns, independently of the main verb, designate such patterns of experience. Thus constructions are invoked both for marked or especially complex pairings of form and meaning and for many of the basic, unmarked patterns of language.

FUTURE PROSPECTS

Cross-linguistic Work

Constructions that are sometimes labeled as the 'same' in two languages typically differ subtly in their form, their meaning, and/or their use. Thus

Construction Grammarians have generally been cautious about trying to explain generalizations that may not be exceptionless. There is, however, a growing body of work on constructions in various languages, and a growing focus on accounting for cross-linguistic tendencies, similarities, and implicational hierarchies.

Psycholinguistics: Processing and Acquisition

A central claim made by Construction Grammar is that words and phrases are the same basic type of entity: learned pairings of form and meaning/use. A good deal of interest in the theory has been generated within psycholinguistics, by researchers in both processing and acquisition, because they see in Construction Grammar the possibility of a psychologically plausible and testable linguistic theory.

Further Reading

Bates E and Goodman JC (1997) On the inseparability of grammar and the lexicon: evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes* **12**(5–6): 507–584.

Bencini G and Goldberg A (2000) The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language* **43**: 640–651.

Fillmore CJ, Kay P and O'Connor MC (1988) Regularity and idiomaticity in grammatical constructions: the case of LET ALONE. *Language* **64**: 501–538.

Goldberg AE (1995) *Constructions: A Construction Grammar Approach to Argument Structure Constructions*. Chicago, IL: University of Chicago Press.

Jackendoff R (1997) Twistin' the night away. *Language* **73**(3): 534–559.

Jurafsky D (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* **20**: 137–194.

Kay P and Fillmore CJ (1999) Grammatical constructions and linguistic generalizations: the *What's X doing Y?* construction. *Language* **75**(1): 1–33.

Koenig J-P and Jurafsky D (1994) Type underspecification and on-line construction in the lexicon. *Proceedings of the Thirteenth West Coast Conference on Formal Linguistics*.

Lakoff G (1987) *There*-constructions. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.

Michaelis L and Lambrecht K (1996) Toward a construction-based theory of language function: the case of nominal extraposition. *Language* **72**(2): 215–247.

Tomasello M (1998) The return of constructions. *Journal of Child Language* **25**(2): 431–443.

Zwicky A (1994) Dealing out meaning: fundamentals of syntactic constructions. *Berkeley Linguistics Society* **20**: 611–625.

Conversation, Structure of

Introductory article

Herbert H Clark, Stanford University, Stanford, California, USA

CONTENTS

Conversations emerge as people use dialogue to coordinate on joint activities they engage in. People proceed turn by turn as they reach local agreements on the course of each section and subsection, including the opening and closing of the conversation itself.

INTRODUCTION

Conversations are the product of people engaged in joint activities. A joint activity is one in which two or more people have to coordinate with each other to succeed. When two people waltz, play a duet, or wrestle, they coordinate their individual actions largely by gesture, touch, and other techniques. When two people gossip, plan a vacation, or negotiate a contract, they coordinate largely through dialogue. The structure of these conversations emerges as the participants jointly manage their way through the gossip, the planning, or the negotiation.

Conversations, therefore, reflect the joint activities they coordinate. Every joint activity has participants who are distinct from bystanders, onlookers, or overhearers. In most joint activities, each participant has a role, such as clerk or customer, teacher or student, friend calling or friend called, and the roles help determine what the participants do and say. Most joint activities have mutually recognized goals such as exchanging gossip, planning a vacation, or negotiating a contract, and these have subgoals. Some goals are set from the start, but others get established in the course of the conversation. The participants also have private agendas – such as being polite, or finishing quickly – and these, too, constrain what they do and say. Often, people alternate between two or more joint activities – such as gossiping and eating dinner – and the structure of their conversation reflects the alternations.

ACTIONS OF DIALOGUE

It takes coordination to carry out a joint activity. Joint activities have boundaries – distinct beginnings and ends, and transitions from one part to the next – but these boundaries don't exist until the participants agree to them. To enter a planning session, for example, two people must agree on (1) what the joint activity is to be, (2) who is to take part, and (3) in what roles. They must also maintain or change these agreements at each transition point. People accomplish all this with dialogue, locally, turn by turn.

One basic method for reaching these agreements is the *adjacency pair*, as in this spontaneous example from Svartvik and Quirk (see Further Reading):

Ann	where is your office,
Burton	in the Strand,
Ann	oh well, yes,

Adjacency pairs consist of two parts, by different speakers, where part 2 is conditionally relevant given part 1. Part 1 is a *proposal*, and part 2 is expected to be the *uptake* of that proposal. Here, in turn 1, Ann proposes that Burton tell her where his office is, and in turn 2, he takes up the proposal by saying that it is in the Strand. Ann and Burton use the two turns to agree on the content, participants, and roles of Ann's projected joint action. They would have failed to reach that agreement if, for example, Burton had replied 'What do you mean?' (failing to coordinate content) or 'You mean me?' (failing to coordinate participants). Turns 2 and 3 constitute a second adjacency pair, an assertion plus its uptake.

People in conversation use adjacency pairs for many types of joint actions. They use them for exchanges of information (as in Ann and Burton's question plus answer), greetings ('Hi,' 'Hi'), farewells ('Bye,' 'Bye'), offers ('Have a beer,' 'Thanks'),

orders ('Sit down,' 'Yes, sir'), and apologies ('Sorry,' 'Oh, that's okay'), among others. They use them for even the simplest exchanges of information ('In the Strand,' 'Oh well, yes').

Adjacency pairs can also be used to *project* larger sections, as in this spontaneous example:

B I like tuh ask you something.
A Shoot.
B Y'know *I*'ad my license suspended fur six munts.
A Uh huh.
B Y'know for a reason which, I rathuh not, mention tuh you, in othuh words, – a *ser*ious reason, en I like tuh know if I w'd talk tuh my senator, or – somebuddy, could *they* help me get it back.

B's first turn is a *pre-question*. With it he proposes to ask A a question, and A agrees. B now has the freedom to take up preliminaries to his question, and it takes the two of them several turns to do that. Only then does he ask his question proper, 'Could they help me get it back?' Pre-questions project not only the eventual question but preliminaries to that question.

Pre-questions and their responses belong to a large family of so-called *pre-sequences*. Here are a few more examples:

Pre-request	Customer	Do you have hot chocolate?
	Waitress	Yes, we do.
Pre-invitation	Man	What are you doin?
	Woman	Nothin. What's up?
Pre-narrative	June	Did I tell you I was going to Scotland?
	Kenneth	No.
Pre-conversation	Caller	(rings telephone)
	Recipient	Miss Pink's office.

Each pre-sequence prepares the way for another joint action. The pre-request sets up a request ('I'll have one'); the pre-invitation sets up an invitation ('Would you like…'); the pre-narrative sets up a narrative; and the pre-conversation sets up an entire telephone conversation.

SECTIONS OF CONVERSATIONS

Conversations tend to emerge as a sequence of topics, or sections. Each section reflects a different phase in the overall joint activity – the next bit of gossip, the next segment of the vacation being planned, the next issue of the contract being negotiated. The participants must agree on the opening and closing of each section, and that is where pre-sequences are useful.

Sections that consist of narratives (jokes, anecdotes, recountings of events), for example, are often introduced by a pre-narrative and its response. The following is an instance from Svartvik and Quirk (see Further Reading):

Nancy: I acquired an absolutely magnificent sewing-machine, by foul means, did I tell you about that?
Kate: no,
Nancy: well when I was. doing freelance advertising – (proceeds to give a five minute narrative)

Nancy proposes to tell Kate a story ('Did I tell you about that'), and Kate accepts ('No'). That allows Kate to embark on her narrative – an extended section of the conversation. It takes both parties to agree, because the recipient can always decline, as in this example, also from Svartvik and Quirk:

Connie: did I tell you, when we were in this African village, and (- they were all out in the fields, - the)
Irene: (yes you did, yes, - yes)
Connie: babies left alone, -
Irene: yes.

Irene interrupts Connie (the speech in brackets is overlapping) to say that she *has* heard the story, and the two of them then go down a different path. So conversations are opportunistic: the paths people take depend on the opportunities that become available with each agreement. Nancy and Connie use their pre-narratives to find the best way to proceed and, receiving different replies, go in different directions.

People help signal which opportunities they are taking by using *discourse markers*. For example, Nancy used 'well' to signal that she was introducing a change in perspective as she began her story. Other discourse markers indicate such boundaries as the start of a new topic (e.g., 'so', 'then', 'speaking of that'), the start of a digression ('incidentally', 'by the way'), or the return from a digression ('anyway', 'so'). All help in coordinating what happens next.

Opening a conversation takes special coordination as two or more people move from not being in a conversation to being in one. The following is the opening of a conversation between acquaintances, again from Svartvik and Quirk:

Karen: (rings Charlie's telephone)
Charlie: Wintermere speaking? -
Karen: hello?
Charlie: h*el*lo
Karen: Charlie

Charlie: Yes
Karen: actually it's
Charlie: hello Karen
Karen: it's me
Charlie: M
Karen: I (- laughs) I couldn't get back last night, (continues)

First, Karen and Charlie coordinate contact through a proposal to have a conversation (the telephone ring) and its uptake ('Wintermere speaking?'). Next, they mutually establish their identities. Karen tells Charlie that she recognizes him in turn 5, but Karen has to say 'hello?' 'Charlie', and 'actually it's' before he identifies her in turn 8. Only then does Karen introduce the first topic. It took 10 turns for them to coordinate on the participants, roles, and content of the conversation.

Conversations are no easier to close, as illustrated in this ending to a telephone conversation:

June and I'll. I'll ring again, as soon as I can on the tenth, uhh to definite confirm it,
Kay right,
Kay okay,
June right,
June thanks a lot,
Kay r. right,
June bye bye,
Kay bye

Although June and Kay finish a topic in turns 1 and 2, they cannot hang up without agreeing to hang up. So in turn 3, Kay proposes to close the conversation ('Okay'), and although June could introduce a new topic, she agrees to Kay's proposal ('Right'). That opens up the closing in which the two exchange thanks ('Thanks a lot' 'Right') and then good-byes. The two must *agree* to close the conversation before they actually close it.

GROUNDING WHAT IS SAID

People carry out joint activities against their *common ground* – their mutual knowledge, mutual beliefs, and mutual assumptions. They infer their common ground from past conversation, joint perceptual experiences, and joint membership in cultural communities. When Ann asks Burton 'Where is your office?' she *presupposes* certain common ground – for example, that Burton works on computers and has an office in London, but that she doesn't know where. And with the question itself, she *adds to* their common ground that she wants to know. Conversations proceed by orderly increments to common ground – especially to the common ground relevant to the current joint activities.

So if conversations are to succeed, the participants must *ground* what they say. To ground what is said is to establish the mutual belief that the addressees have understood the speakers well enough for current purposes. One technique for grounding is the adjacency pair itself. When Burton said 'In the Strand', he displayed to Ann how he had interpreted her question. If Ann hadn't been satisfied with that interpretation, she could have corrected it, for example by replying 'No, I meant …'. By following up Burton's reply with 'Oh well, yes,' she displayed her acceptance of his interpretation. Another technique is the *back-channel response*, *acknowledgment*, or *continuer*. In two-party conversations, addressees are expected to add 'uh huh' or 'mhm' or 'yeah' at or near the ends of certain phrases. With these, they signal that they understand well enough for the speaker to continue.

Grounding is sometimes achieved through *side sequences*, as in this spontaneous example, once more from Svartvik and Quirk:

Roger well there's no general agreement on it I should think,
Sam on what?
Roger on uhm - - on the uhm – the mixed up bits in the play, the
Sam yes

When Sam didn't understand Roger's 'it', he initiated an embedded adjacency pair in turns 2 and 3, a side sequence, to clear up the problem. Only when he had cleared it up did he acknowledge or agree with 'Yes'. Side sequences are initiated to clear up not only mishearings and misunderstandings but other preconditions to taking up the first part ('Why do you want to know?'). Grounding is sometimes accomplished by overlapping speech. When Irene interrupted Connie's offer 'Did I tell you …' to say, 'Yes you did, yes, – yes', she was signaling to Connie that she already understood and Connie didn't need to go on.

CONCLUSION

The structure of conversations emerges step by step as people coordinate on each new move in their joint activities. People need to coordinate on the content, participants, and roles of each joint action, and they do that in a sequence of local, opportunistic agreements. It is these techniques that give conversations their structure.

Further Reading

Atkinson JM and Heritage J (eds) (1984) *Structures of Social Action: Studies in Conversation Analysis*. Cambridge, UK: Cambridge University Press.

Clark HH (1996) *Using Language*. Cambridge, UK: Cambridge University Press.

Drew P and Heritage J (eds) (1992) *Talk at Work : Interaction in Institutional Settings*. Cambridge, UK: Cambridge University Press.

Duncan S and Fiske DW (1977) *Face-to-Face Interaction*. Hillsdale, NJ: Lawrence Erlbaum.

Goffman E (1981) *Forms of Talk*. Philadelphia, PA: University of Pennsylvania Press.

Goodwin C (1981) *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Kendon A (1990) *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge, UK: Cambridge University Press.

Levinson SC (1983) *Pragmatics*. Cambridge, UK: Cambridge University Press.

Ochs E, Schegloff EA and Thompson SA (1996) *Interaction and Grammar*. New York, NY: Cambridge University Press.

Sacks H, Schegloff EA and Jefferson G (1974) A simplest systematics for the organization of turn-taking in conversation. *Language* **50**: 696–735.

Schegloff EA, Jefferson G and Sacks H (1977) The preference for self-correction in the organization of repair in conversation. *Language* **53**: 361–382.

Searle JR, Parret H and Verschueren J (eds) (1992) *(On) Searle on Conversation*. Amsterdam, Netherlands: J. Benjamins Pub. Co.

Stenström A-B (1994) *An Introduction to Spoken Interaction*. London, UK: Longman.

Svartvik J and Quirk R (eds) (1980) *A Corpus of English Conversation*. Lund, Sweden: Gleerup.

Discourse Processing

Intermediate article

Barbara Di Eugenio, University of Illinois, Chicago, Illinois, USA

CONTENTS

The theory of discourse processing concerns the computational processes underlying the interpretation and production of text encompassing more than one sentence – i.e., discourse. Discourse is generally taken to be written, and often, but not always, monologic.

INTRODUCTION

Two phenomena are considered intrinsic to discourse processing: the interpretation and production of phrases and utterances whose meaning depends on the discourse context; and the fact that a sequence of two or more utterances almost always conveys a meaning that is more than the sum of the meanings of the individual utterances. Consider the following example:

> As soon as they got to the beach, Karin jumped into the water. She was so hot from the long drive. (1)

Example 1 illustrates the issues most closely associated with the above phenomena: respectively, reference resolution and production, and text coherence.

Reference resolution concerns the interpretation of those noun phrases speakers use to refer to what are called discourse entities; i.e. entities in the model of the discourse – for example, *Karin*, *she*, *the long drive*. Reference resolution is closely related to the notion of processing of anaphora. This article will discuss the converse problem of referential expression generation, namely, how to choose a specific referential expression among all those that can potentially be used to refer to a discourse entity. (*See* **Anaphora, Processing of**)

It is difficult to define text coherence exactly. We could define it as the quality of a text that is 'tied' together in just the right way. It is text that can be readily comprehended by the hearer, apparently without effort; at the same time, it is text where relations between different sentences are not so explicit as to make it uninteresting. Example 1 is coherent; however, consider the following:

> As soon as they got to the beach, Karin jumped into the water. She hates ice cream. (2)

Example 2 sounds incoherent: it is likely that the hearer will wonder about the connection between hating ice cream and jumping into the water. The hearer may proceed to invent scenarios in which the text makes sense: for example, Karin had ice cream on the way to the beach, it gave her a stomach ache, and her way to deal with stomach aches is to swim. Scenario building of this sort is an exercise precisely in accounting for text coherence; i.e., in explaining text in terms of sentence connections to one another. However, this does not mean that every possible link between the sentences should be explicit. Expanding Example 1 results in a more tedious text, not a clearer one:

> As soon as they got to the beach, Karin jumped into the water. She was so hot from the long drive, so she wanted to cool down. Because the temperature of the sea is generally much lower than that of the air, going for a swim accomplished her goal. (3)

Thus, text coherence appears to obey Grice's maxims of quantity. (*See* **Implicature**)

Text coherence encompasses more than appropriate connections between individual sentences. Discourse appears to have a hierarchical structure: sentences are part of segments, which in turn are part of superordinate segments. Informally, a segment can be defined as a group of locally coherent utterances (see below for a more formal definition).

Consider the following discourse:

(a) Georgia called Jeffrey on the phone.
(b) She wanted to wish him happy birthday.
(c) She also asked him if she could borrow his tent.
(d) She had bought a tent herself a few months back.
(e) However, it got torn on her summer hikes.
(f) After the phone call, she went out for a jog. (4)

Discourse 4 is about Georgia's activities. Intuitively, we recognize that sentences (a) through (e) form a subsegment S_{a-e} of the whole discourse, as they pertain to Georgia's phone call to Jeffrey. In turn, sentences (c), (d), and (e) form the subsegment S_{c-e} of S_{a-e} that concerns the request for the tent; and sentences (d) and (e) form subsegment S_{d-e} of $S_{a-e,}$ because together they provide a justification for the request in sentence (c).

We will now look at how different researchers account for text coherence in both its manifestations, connections between sentences and discourse segmentation, and how coherent discourses can be interpreted and generated.

THEORIES OF DISCOURSE STRUCTURE

Two theories of discourse structure came to the fore in the mid-1980s, and these are still the most prominent: Grosz and Sidner's (1986) and Mann and Thompson's (1988).

Grosz and Sidner's theory (GST) sees discourse structure as the surface manifestation of the relationships among elements of the intentional structure underlying the discourse. The intentional structure is comprised of the intentions that a speaker brings to the discourse. There will be a primary intention, the discourse purpose (DP), which is, the intention that underlies engaging in that particular discourse. Further, a discourse segment purpose (DSP) is associated to each discourse segment, which is fully individuated by the corresponding DSP. Each DSP specifies how the specific segment contributes to achieving the overall DP. A plausible DP underlying the whole of discourse 4 is *Tell hearer about Georgia's activities*. The DSP associated to the subsegment S_{d-e} could be something like *Explain to hearer why Georgia needs to borrow Jeffrey's tent*. GST does not specify which intentions can count as DPs or DSPs, except by noting that they are meant to be recognized (cf. Grice's notion of utterance-level intentions). DSPs can be related to one another only via two relationships: dominance and satisfaction-precedence. DSP_1 dominates DSP_2 if DSP_2 is intended to provide part of the satisfaction of DSP_1. DSP_1 satisfaction-precedes DSP_2 if DSP_1 must be satisfied before DSP_2. Grosz and Sidner argue that the intentional structure of the discourse is also intertwined with the attentional state, the set of entities that are salient at any point in the discourse. Attentional state is modeled by a set of focus spaces, which are associated to discourse segments, and contain the entities salient within the corresponding discourse segments. The processing of focus spaces is modeled via a stack. Shifts of attentional state that are local to a discourse segment are outside the scope of GST, but are accounted for by centering theory (Grosz *et al*., 1995; Walker *et al*., 1998).

Mann and Thompson (1988) propose 'rhetorical structure theory' (RST) as a descriptive framework that identifies hierarchical structure in text. RST is based on relations between two non-overlapping text spans, the 'nucleus' and the 'satellite'. The nucleus is the central member of the pair; the satellite is peripheral. Relations include an effect and constraints on nucleus and satellites. The relations defined by Mann and Thompson include 'elaboration', 'enablement', 'evidence', and 'contrast'; comparable inventories of discourse relations have been proposed by a number of other researchers (e.g. Hobbs, 1979; Lascarides and Asher, 1993). For example, the evidence relation has as effect that the belief of the hearer in the nucleus is increased, and a constraint that the hearer will find the satellite believable. In Mann and Thompson's view, an analyst will first identify the minimal units of the analysis, which they assume to be clauses. Then, the analyst will start applying relation schemas to adjacent text spans, which are minimal units or, recursively, constituents of relations. In the end there will be one relation schema encompassing text spans that cover the whole text.

From these brief descriptions, we can see that GST mainly accounts for the segmentation aspect of discourse coherence, but does not address how individual sentences are linked to one another by domain or rhetorical relations. As Grosz and Sidner believe that the intentions underlying discourse are too diverse, they argue that it would impossible to enumerate the intentions that can serve as DSPs; hence, they conclude that enumerating a fixed set of relations, as in RST, is wrong. On the other hand, RST accounts both for individual relations between individual sentences and for hierarchical segmentation of the discourse. The latter is a side effect of

how the RST analysis is conducted. Note that an analysis of a discourse according to GST generally results in fewer and shallower segments than an RST analysis.

Grosz and Sidner present their theory as a computational account of discourse processing, but they do not provide much insight into the underlying computational processes, except by proposing that the attentional state is modeled as a stack. Mann and Thompson do not make any computational claims; however, their theory has been widely used in computational linguistics. The question thus arises, which processing paradigm is most appropriate for each theory. GST lends itself to a top-down model of discourse processing: the hearer recognizes the DP, and then recursively the subordinate DSPs. RST, on the other hand, lends itself more directly to a bottom-up interpretation of discourse.

A synthesis of GST and RST has been proposed (Moser and Moore, 1996). The synthesis is based on the observation that the dominance relation between intentions in GST closely corresponds to the 'nucleus versus satellite' distinction between text spans in RST.

INTERPRETATION OF DISCOURSE

Discourse interpretation consists of the computational inferences that compute the extended meaning of discourse. We can divide the approaches into two main groups: logical approaches that compute domain and rhetorical relations between sentences in written texts (Lascarides and Asher, 1993), and plan inference approaches that compute the speech acts performed by participants in a dialogue (Perrault and Allen, 1980; Litman and Allen, 1990; Carberry and Lambert, 1999). Plan inference approaches have been applied mainly to dialogue; nevertheless, they are considered part of discourse processing. Because of its inherent difficulty, not many researchers have tried to compute discourse segmentation as proposed in GST, but see Lochbaum (1998) for such an attempt.

Traditionally, approaches to inferring relations between sentences make use of some type of logical inferencing, such as a variant on non-monotonic logic or abduction. We will briefly look at approaches based on abduction (Hobbs *et al*., 1993). Abduction is an unsound inference rule that reasons from an effect to a potential cause: for example, 'the alarm went off, so there is a burglar in the house'. Clearly, there may be other reasons why the alarm went off (e.g. the landlady forgot to switch it off). Abduction is a useful reasoning mechanism because it tries to find the best explanation for a fact. As far as discourse coherence is concerned, an abductive approach tries to find the most plausible coherence relation linking two utterances, on the basis of rhetorical, domain, and world knowledge. An abductive approach will build a full explanation that supports a specific coherence relation. For example, to establish a cause relation between the two sentences in example 2, an abductive approach would build an explanation, expressed in first-order predicate logic, akin to the one in example 3. The problem abduction has to face is how to choose the most plausible explanation among many possible ones. One can adopt heuristics, such as choosing the explanation that uses the fewest assumptions, or compute the probability of each explanation and choose the most likely one. Both approaches have serious flaws: the former, that even plausible heuristics can fail fairly often; the latter, that it is unclear over which space of events to compute those probabilities.

The computational approaches just discussed are not explicitly based on cognitive findings on text comprehension. Nevertheless, questions addressed by cognitive scientists and psycholinguists have affected computational models. Relevant issues include: inference control (i.e. which of the many possible inferences are made at comprehension, and which later, during recall); and how the connectedness of sentences affects reading times and the accuracy of recall. For example, it has been found that sentences that have a close causal connection are read faster and result in better content recall (Myers *et al*., 1987).

Plan Inference

The plan inference approach to discourse has been applied mainly to dialogues, although applications to monologic discourse that describes one or more agents' actions have also been attempted. It originated at the end of the 1970s (Perrault and Allen, 1980), with the goal of providing an interpretation for indirect requests such as *I need to be in Boston on the 20th in the afternoon* (directed to a travel agent), or *The next train to Brighton* (directed to a clerk at the ticket booth). It rests on three components: the notion of speech acts from pragmatics; a theory of belief, desire, and intentions from computational linguistics, which in turn owes much to the philosophy of action; and planning models from artificial intelligence.

Every utterance counts as an action performed by the speaker (a speech act), such as asking or

promising (Austin, 1962). (We are oversimplifying here. In reality there are three acts associated with each utterance: locutionary, illocutionary, and perlocutionary. The term 'speech act' has come to refer mostly to the illocutionary act.) Utterances can perform speech acts directly, as in example (a) below, or indirectly, as in example (b):

(a) Please find me a flight that arrives in Boston on the 20th in the afternoon.
(b) I need to be in Boston on the 20th in the afternoon. (5)

To explain how a statement such as example 5(b) can count as a request, proponents of the 'inferential' approach (Searle, 1975) contend that indirect speech acts concern felicity conditions on the corresponding direct act. For example, a request such as example 5(a) is felicitous under the assumption that the speaker wants to fly to Boston on that specific date and time. Example 5(b) then works because it explicitly states the speaker's mental attitude, once the hearer has recognized that the literal meaning of it is inappropriate and must be 'repaired' by some inference. (*See* **Pragmatics, Formal**)

Planning is a computational technique from artificial intelligence that, given a goal s_g to achieve, builds a plan, a partially ordered sequence of actions whose execution will bring the agent from the initial state s_0 to s_g. Often the plan is built as a tree, whose leaves are the actions to be executed; the internal nodes represent actions at a higher level, which decompose into lower-level actions. For example, if an agent has the goal *Attend conference in Washington* and lives in Chicago, the agent may build a plan that includes taking a flight from Chicago to Washington; in turn, to achieve taking the flight, the agent will need to buy a ticket, drive to the airport, and board the plane. Planners build plans on the basis of action operators, which must include: preconditions, the conditions that need to hold for the action to be executable; effects, what becomes true after performing the action; and body, a decomposition into a partially ordered set of subactions whose execution will result in the execution of the action.

In the plan inference approach, speech acts are modeled as action operators from planning. However, the logical language in which the operators are expressed is augmented with mental attitudes such as knowledge, beliefs, and desire. For instance, a formalization of `Request` (S, H, α) will include as a precondition that the speaker S wants the hearer H to perform action α (one of the felicity conditions on requests), and as an effect that H wants to perform α. Such a formalization can be used to build the interpretation of an indirect speech act via plan inference rules that work backwards from the utterance to its interpretation. One such rule is: if γ is a precondition of action α and H believes S to want γ, then it is plausible that H believes S to want α. Note that the representation can also be used by a regular planner to produce a speech act, starting from a communicative goal to be achieved.

The plan inference approach has been extensively used in dialogue modeling. The original model has been extended in various ways, such as by introducing different levels of inferred plans (e.g., the discourse plan and the domain plan) that the speaker is pursuing (Litman and Allen, 1990; Carberry and Lambert, 1999).

Approaches to discourse based on abduction, non-monotonic logic or plan inferencing, while elegant, suffer from brittleness. One missing domain axiom may cause the model to fail as it is not able to find a complete explanation. Thus, many implemented systems, instead of just using a logical approach, use information that can easily be derived from the surface form of the utterance, such as intonation, connectives, idiomatic expressions, and lexical associations between words (Reithinger and Maier, 1995; Qu *et al*., 1997; Samuel *et al*., 1998). These cues to the phenomenon of interest are derived from linguistic and corpus analysis (see below).

GENERATION OF DISCOURSE

From a computational point of view, discourse generation concerns the production of coherent, extended text. Whereas discourse processing is seen as the last stage in language interpretation, after parsing and semantic analysis, it is the first stage for language production. Computationally, language generation starts from a nonlinguistic representation of information that we can consider parceled into messages to be conveyed. The first task to be performed is discourse planning; i.e. imposing ordering and structure over the set of messages to be conveyed. This is followed by sentence planning and linearization, including sentence aggregation (grouping the elements of the discourse plan together into sentences) and the choice of referential terms to individuate the entities of interest. The final step is linguistic realization proper, namely, applying the rules of grammar in order to produce a text that is syntactically and morphologically correct.

Here, we concentrate on discourse planning, and on referring expression generation.

Planning and Linearization

There are two main approaches used to generate a coherent discourse: planning, and schemata.

The discourse planner is given a communicative goal to achieve such as `Intend` S (`Intend` H α). Communicative goals represent the speaker's intentions to affect the beliefs or goals of the hearer. The planner will build a plan consisting of rhetorical actions to achieve the given communicative goal. For example, to achieve `Intend` S (`Intend` H α), S may look for a β such that S expects H to want β, and then utter β as motivation for α (motivation is an RST relation):

> Come to the party on Saturday. I will make your favorite deviled eggs. (6)

The connection between discourse planning and the theories of discourse structure discussed earlier has mainly been achieved through RST. RST relations are recast in the terms of planning operators (Moore and Paris, 1993). The planner posts a high-level communicative goal such as `Intend` S (`Intend` H α) in terms of the effect ε the text should have on the reader. The planner will then search for an RST operator whose effect unifies with ε, and post the subgoals that correspond to constraints on the nucleus and satellite of that rhetorical relation. These subgoals are recursively expanded until the planner reaches the leaves of the rhetorical structure tree, which are expressible as simple clauses.

Schemata are an alternative approach to using a planner. Schemata represent common patterns that texts in a specific domain or genre follow (McKeown, 1985). A schema specifies how a particular discourse plan should be built using other schemata or messages, and the discourse relations that hold between different components of the discourse plan. Although schemata are not generally developed following a planning model, they can be considered as compilations of discourse plans produced by a planning system. As a mechanism for generation, schemata are less flexible, but easier to develop, than a fully-fledged discourse planner. For example, because schemata lack information on the intentions of the speaker, they cannot be used if the system needs to replan, for example if the explanation of a certain p is not understood by the hearer and the discourse planner needs to build a different explanation for p (Moore and Paris, 1993).

Note that a discourse plan, whether built by a planner or as a schema instantiation, does not encode decisions regarding how the leaves should be parceled into individual sentences, or how these sentences should be connected. For instance, the two sentences in example 6 could be linked in a variety of different ways, both paratactically and hypotactically. For example:

> Come to the party on Saturday if you don't want to miss your favorite deviled eggs. (7)

There are also more subtle decisions that need to be made. In example 6 the adjective *favorite* in the second clause may well be derived from a full proposition in the discourse plan, such as *You like the deviled eggs I make a lot*. This is why many researchers consider lexicalization as part of sentence planning. Lexicalization is concerned with choosing words to express concepts and relations.

The solutions proposed in the literature for sentence planning and linearization are diverse, but some general paradigms are beginning to emerge.

Establishment of Referential Terms

The task of generating referring expressions is to select words or phrases to identify discourse entities. The choice of referring expressions greatly affects the readability of a text. Compare the two texts below, the second of which always uses the nominal expression *Bill Gates*.

> When a Stanford University professor asked for volunteers to have their heads scanned, Bill Gates was the first to volunteer.
> The billionaire CEO of Microsoft Corporation sat patiently while a laser scanner orbited his head several times. A short time later, a 3-D image of Gates's head floated on a screen. (8)

> When a Stanford University professor asked for volunteers to have their heads scanned, Bill Gates was the first to volunteer.
> Bill Gates sat patiently while a laser scanner orbited Bill Gates's head several times. A short time later, a 3-D image of Bill Gates's head floated on a screen. (9)

In text 9, the repeated use of the proper name *Bill Gates* makes the text sound clumsy. The much more fluent text 8 makes use of different forms of proper names (*Bill Gates* or simply *Gates*), pronouns (*his*), and complex definite referring expressions (*the billionaire CEO of Microsoft Corporation*).

The problem of generating referring expressions can be subdivided into two tasks:

- Initial introduction, or how to perform the initial reference to a discourse entity.
- Subsequent references. This includes choosing between a pronoun and a definite description: if the latter is chosen, then the question is which features of the entity in question to include in the description.

The initial introduction and the choice of pronoun or definite description are generally performed by algorithms based on the 'given–new' distinction (Prince, 1981) or on centering (Grosz *et al*., 1995; Walker *et al*., 1998) or on a combination of the two. (*See* **Computability and Computational Complexity**)

Regarding the choice of appropriate definite descriptions, early approaches (Appelt, 1985) took a full planning approach to generating referring expressions. This means that in principle they could generate any description that satisfies a given communicative goal. As this approach was computationally inefficient, later approaches, most notably Dale's (1992), focused on the restricted problem of building a 'distinguishing description'. A distinguishing description is true only of the entity being described and of no others among the currently salient discourse entities. These algorithms generally aim at finding the minimal distinguishing description. However, even computing a minimal distinguishing description is an inherently hard computational task. Dale and Reiter (1995) showed that it is NP-hard by reducing it to a set cover problem. Moreover, humans do not produce minimal distinguishing descriptions, either because humans also face computational limitations, or because they intend to achieve other goals besides identification (Jordan, 2000). In text 8, the complex noun phrase *the billionaire CEO of Microsoft Corporation* may be used to introduce information that the hearer is not expected to know, or, more likely in this case, to remind the hearer of Gates's position. Algorithms used today try to strike a balance between conciseness of the definite description and reproducing human behavior, as observed in corpus analysis.

EMPIRICAL APPROACHES TO DISCOURSE

In the 1990s there was a shift in focus towards a rigorous empirical foundation for discourse processing work. The general methodology that has emerged (Walker and Moore, 1997) comprises the following components:

- Development and evaluation of coding schemes. Coding schemes are used to annotate language corpora for features deemed likely to affect the phenomena under study (e.g. correlates of discourse segments, minimality of referential expressions with respect to providing distinguishing descriptions). A necessary condition for a coding scheme to be useful is that it be reliable, namely, that two or more independent coders can use that coding scheme to annotate the same text in a 'similar enough' way. Much research has thus been devoted to measures of inter-coder agreement (Carletta, 1996; Di Eugenio, 2000).
- Extraction of information from the annotated corpus. Researchers use either statistical measures or data mining tools on the annotated features (Di Eugenio *et al*., 1997; Poesio and Vieira, 1998; Samuel *et al*., 1998; Jordan, 2000). The purpose is to verify hypotheses (e.g. the hypothesis that in real-world situations speakers use minimal distinguishing descriptions), and to find linguistic correlates of higher-level phenomena, such as intonation patterns and adverbs for discourse segmentation.
- Development of computational frameworks based on the information extracted from the corpus. For example, the result of an annotation for referring expressions is used to inform algorithms to generate referring expressions (Poesio and Vieira, 1998; Jordan, 2000).
- Evaluation. The computational models developed either theoretically or on the basis of corpus analysis need to be evaluated. This has motivated much interest in evaluation methodologies for computational models and implemented systems (e.g. Walker *et al*., 1997). However, it is still too early to report specific results that determine which techniques, models, or systems are the most promising. Systematic evaluations have only recently become the norm, and there is no standard test-bed of problems and phenomena that can be used to make comparisons across systems and techniques.

Acknowledgment

This work was partially supported by grant N00014-00-1-0640 from the Office of Naval Research, Cognitive, Neural and Biomolecular S&T Division.

References

Appelt D (1985) Planning English referring expressions. *Artificial Intelligence* **26**(1): 1–33. [Reprinted in: Grosz, Sparck Jones and Webber (eds) (1986) *Readings in Natural Language Processing*. Santa Monica, CA: Morgan Kaufmann.]

Austin JL (1962) *How to Do Things With Words*. Oxford, UK: Oxford University Press.

Carberry S and Lambert L (1999) A process model for recognizing communicative acts and modeling

negotiation subdialogues. *Computational Linguistics* **25**: 1–53.
Carletta J (1996) Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics* **22**: 249–254.
Dale R (1992) *Generating Referring Expressions*. Cambridge, MA: MIT Press.
Dale R and Reiter E (1995) Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* **18**: 233–263.
Di Eugenio B (2000) On the usage of Kappa to evaluate agreement on coding tasks. In: Gavrilidou M, Carayannis G, Markantonatou S *et al*. (eds) *LREC2000, Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 441–444. Athens, Greece: National Technical University of Athens Press.
Di Eugenio B, Moore JD and Paolucci M (1997) Learning features that predict cue usage. In: *ACL-EACL97, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 80–87. San Francisco, CA: Morgan Kaufmann.
Grosz B, Joshi A and Weinstein S (1995) Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* **21**: 203–225.
Grosz B and Sidner C (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics* **12**: 175–204.
Hobbs JR (1979) Coherence and co-reference. *Cognitive Science* **3**(1): 67–82.
Hobbs J, Stickel M, Appelt D and Martin P (1993) Interpretation as abduction. *Artificial Intelligence* **63**(1–2): 69–142.
Jordan PW (2000) *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. PhD thesis, University of Pittsburgh.
Lascarides A and Asher N (1993) Temporal interpretation, discourse relations, and commonsense entailment. *Linguistics and Philosophy* **16**: 437–493.
Litman D and Allen J (1990) Discourse processing and commonsense plans. In: Cohen P, Morgan J and Pollack M (eds) *Intentions in Communication*, pp. 365–388. Cambridge, MA: MIT Press.
Lochbaum KE (1998) A collaborative planning model of intentional structure. *Computational Linguistics* **24**: 525–572.
Mann WC and Thompson S (1988) Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**(3): 243–281.
McKeown KR (1985) *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge, UK: Cambridge University Press.
Moore JD and Paris CL (1993) Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics* **19**: 651–695.
Moser M and Moore JD (1996) Towards a synthesis of two accounts of discourse structure. *Computational Linguistics* **22**: 409–419.
Myers JL, Shinjo M and Duffy SA (1987) Degree of causal relatedness and memory. *Journal of Verbal Learning and Verbal Behavior* **26**: 453–465.
Perrault R and Allen J (1980) A plan-based analysis of indirect speech-acts. *American Journal of Computational Linguistics* **6**: 167–182.
Poesio M and Vieira R (1998) A corpus-based investigation of definite description use. *Computational Linguistics* **24**: 183–216.
Prince E (1981) Toward a taxonomy of given–new information. In: Cole P (ed.) *Radical Pragmatics*, pp. 223–255. New York, NY: Academic Press.
Qu Y, Di Eugenio B, Lavie A, Levin L and Rosé CP (1997) Minimizing cumulative error in discourse context. In: Maier E, Mast M and LuperFoy S (eds) *Dialogue Processing in Spoken Language Systems*, pp. 171–182. Heidelberg, Germany: Springer-Verlag.
Reithinger N and Maier E (1995) Utilizing statistical dialogue act processing in Verbmobil. In: *ACL95, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 116–121. Cambridge, MA: MIT Press.
Samuel K, Carberry S and Vijay-Shanker K (1998) Dialogue act tagging with transformation-based learning. In: *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1150–1156. San Francisco, CA: Morgan Kaufmann.
Searle JR (1975) Indirect speech acts. In: Cole P and Morgan JL (eds) *Syntax and Semantics*, vol. III, *Speech Acts*, pp. 59–82. San Diego, CA: Academic Press. [Reprinted in: Davis S (ed.) (1991) *Pragmatics: A Reader*. New York, NY: Oxford University Press.]
Walker M, Joshi A and Prince E (eds) (1998) *Centering Theory in Discourse*. Oxford, UK: Oxford University Press.
Walker MA, Litman DJ, Kamm CA and Abella A (1997) PARADISE: a framework for evaluating spoken dialogue agents. In: *ACL-EACL97, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 271–280. San Francisco, CA: Morgan Kaufmann.
Walker MA and Moore JD (1997) Empirical studies in discourse. *Computational Linguistics* **23**: 1–12.

Further Reading

Allen J (1995) *Natural Language Understanding*, 2nd edn. Menlo Park, CA: Benjamin/Cummings.
Cohen PR, Morgan J and Pollack ME (eds) (1990) *Intentions in Communication*. Cambridge, MA: MIT Press.
Grosz B and Sidner C (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics* **12**: 175–204.
Jurafsky D and Martin JH (2000) *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall.
Levinson S (1983) *Pragmatics*. Cambridge, UK: Cambridge University Press.

Mann WC and Thompson S (1988) Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**(3): 243–281.

Reiter E and Dale R (2000) *Building Applied Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.

Walker MA, Joshi A and Prince E (eds) (1998) *Centering Theory in Discourse*. Oxford, UK: Oxford University Press.

Walker MA and Moore JD (1997) Empirical studies in discourse. *Computational Linguistics*: **23**: 1–12. [Special issue on empirical studies in discourse.]

Webber BL (1999) Computational aspects of discourse and dialogue. In: Schiffrin D, Tannen D and Hamilton H (eds) *The Handbook of Discourse Analysis*. Oxford, UK: Blackwell.

Disfluencies in Spoken Language

Intermediate article

Jean E Fox Tree, University of California, Santa Cruz, California, USA

CONTENTS

Speech disfluencies are phenomena that cause a break in the smooth flow of talk, affecting speaking and understanding. Everyone who talks produces disfluencies. Several decades of research have provided a good understanding of the types of disfluencies people produce, and a beginning understanding of why disfluencies occur and how they impact addressees.

CATEGORIES OF DISFLUENCIES

Disfluencies are any of a group of phenomena that cause a break in the smooth flow of spoken talk. Three main categories of disfluencies are: (1) pauses; (2) *um*s and *uh*s; and (3) repetitions, replacements, and restarts, although there are others. Disfluencies are normal nonfluent speech, as opposed to dysfluencies, or abnormal nonfluent speech, such as stuttering (Wingate, 1984).

A pause is a stretch of speech that is heard as silence. Pauses are usually identified as times when the speaker is not saying anything, but they may also be heard when there is no actual silence in the speech, but rather a slowing down of tempo. Traditionally, only silences over a quarter of a second long were considered meaningful breaks in talk (Goldman-Eisler, 1968), but some have argued that pauses over a tenth of a second long should form the lower cutoff point (Hieke *et al*., 1983). Most pauses in spontaneous talk are under 1 second long (Jefferson, 1989); in fact, pauses over 3 seconds have defined conversational lapses (McLaughlin and Cody, 1982). Pauses have been further categorized by their position in the clause, such as whether they are within or between sentences, or their purpose, such as whether or not they are produced for rhetorical effect. Although only a subset of pauses may be disfluent pauses, it can be hard to determine whether a pause is a disfluency or not.

*Um*s and *uh*s describe a group of sounds that sound, in English, like /um/ and /uh/, with some variation in the shape of the vowel (/em/ and /eh/, for example). They are sometimes referred to as fillers or filled pauses. These labels descend from historical comparisons between *unfilled pauses* (pauses) and *filled pauses* (*um*s and *uh*s; Maclay and Osgood, 1959); *um*s and *uh*s were thought to be ways to fill silence to show that a speaker intends to continue speaking (Cook, 1971). But this theoretical position has not held up over time. To avoid the implication that *um*s and *uh*s are equivalent to or alternative versions of silent pauses, some researchers identify them only as *um*s and *uh*s, or use the label *interjection*. Instead of being equivalent to pauses, *um*s and *uh*s indicate the lengths of upcoming pauses, with *um*s indicating major pauses and *uh*s minor pauses (Clark, 1994; Clark and Wasow, 1998; Smith and Clark, 1993).

Repetitions, replacements, and restarts are stretches of speech where people have: (1) stopped the smooth flow of speech; (2) possibly uttered a pause, *um* or *uh*, or other words such as *I mean* or *you know*; and (3) resumed their talk (Clark, 1996). In repetitions, words are repeated exactly in the resumption, as in 'of her of her daughter'. In replacements, some words are repeated but some are changed, as in 'there were a lot of tricks that the um tricks and toys that the ant could play with'. In restarts, no words are repeated, as in 'what would you-can I help you?'. Although people can detect problems that need correcting after hearing themselves start to say something wrong (Levelt, 1989), people can also choose in advance when they will suspend their speech, detecting problems while speaking and suspending their speech when they have the continuation ready (Blackmer and Mitton, 1991; Fox Tree and Clark, 1997). People also choose how to resume fluent talk, resulting in various resumptions.

The part of speech that is stopped is sometimes called the *reparandum*, and the part of speech that is resumed is sometimes called the *repair* (Levelt,

1983). The term 'repair' has also been used to refer to repetitions, replacements, and restarts as a group along with similar phenomena (Fox Tree and Clark, 1997). But the term 'repair' can imply the revising of something said earlier, which is not always the case for these disfluencies. For example, repetitions can be viewed as early commitments to speaking at particular moments with subsequent restorations of continuity in the resumption (Clark and Wasow, 1998), instead of as second occurrences' revising first occurrences, without changes.

The three types of disfluencies discussed here – pauses, *um*s and *uh*s, and repetitions, replacements, and restarts – are interrelated. Pauses can predict upcoming repetitions, replacements, and restarts. *Um*s and *uh*s indicate the lengths of upcoming pauses. And information between the suspension and resumption of repetitions, replacements, and restarts can contain pauses or *um*s and *uh*s.

DISFLUENCIES IN LANGUAGE PRODUCTION

Since at least the 1960s, disfluencies have been seen as windows into the speech production process. They could be the auditory remains of a problem in turning thoughts into words, or they could be the normal result of speakers' planning their talk. Hypotheses about disfluencies' aetiologies or purposes were arrived at by analyzing when they occurred.

Hypotheses about pauses were that they were epiphenomena of a general need for more processing time to produce talk (Levelt, 1989), or the result of more specific effort at lexical access (Goldman-Eisler, 1968; Maclay and Osgood, 1959; Martin and Strange, 1968), syntactic formulation (Brotherton, 1979; Clark and Wasow, 1998; Duez, 1982; Ferreira, 1991; Maclay and Osgood, 1959), or phonological encoding (Ferreira, 1991). Pauses were also thought to be used more purposefully for rhetorical effect (Duez, 1982; Kowal *et al*., 1985), such as making people appear sincere (Maclay and Osgood, 1959). Pause placement also influenced hypotheses about the order of speech production processes, such as that syntactic formulation precedes lexical access (Maclay and Osgood, 1959).

Similar hypotheses were made about *um*s and *uh*s. Without distinguishing between proposals that they are symptoms or signals, *um*s and *uh*s have been thought to foreshadow: (1) general speech production difficulty (Brotherton, 1979; Reynolds and Paivio, 1968) or specific difficulty, such as upcoming delays (Clark, 1994) or error avoidance (Jefferson, 1974); (2) particular kinds of words, such as difficult to produce or unpredictable words (Brotherton, 1979; Tannenbaum *et al*., 1965) or words with more competitors (Schachter *et al*., 1991); (3) the major chunks of talk in a discourse (Swerts, 1998); (4) difficulty in planning what one wants to say and how to say it syntactically (Maclay and Osgood, 1959; Martin and Strange, 1968; Reynolds and Paivio, 1968); (5) speakers' desires to maintain control of the floor in a conversation (Maclay and Osgood, 1959); and (6) speakers' desires to show awareness of upcoming delays, to avoid being cast in a negative light by a silent pause (Smith and Clark, 1993).

Repetitions, replacements, and restarts in speech production come about because of a variety of problems, including conceptualizing ideas, formulating sentences, selecting words, or articulating utterances (Levelt, 1989). Different types of production trouble may yield different kinds of recovery (Tannenbaum *et al*., 1965). One explanation for the reason repetitions, replacements, and restarts look the way they do is that speakers follow rules for making them well formed (Levelt, 1983, 1989). They are well formed if there is a way of converting the suspended talk and the resumption into a coordination; for example, because 'there were a lot of tricks that the- um tricks and toys that the ant could play with' could be filled out to the well-formed sentence 'there were a lot of tricks that the [ant could do, and] um tricks and toys that the ant could play with', the replacement without the bracketed talk is well-formed.

DISFLUENCIES IN LANGUAGE COMPREHENSION

Fewer researchers have explored the role of disfluencies in speech comprehension.

Disfluencies can be difficult to detect in talk, although listeners can detect them with effort (Martin, 1967; Martin and Strange, 1968). Detection of pauses may depend on where in the clause they fall; one study found that within-clause pauses can be detected if they are over 200 ms, but between-clause pauses need 500 ms to 1000 ms for detection (Boomer and Dittmann, 1962). Detection of *um*s and *uh*s varies depending on whether listeners are paying attention to what speakers are saying or how they are saying it (Christenfeld, 1995). Detection of repetitions and restarts takes place after the smooth flow of speech has stopped (Lickley and Bard, 1998).

None the less, effects of disfluencies on comprehension have been measured. There are generally two different measurement techniques, those that

involve offline tasks (measuring comprehension after speech has been heard) and those that involve online tasks (measuring comprehension while speech is being heard).

In offline tasks, disfluencies have influenced what listeners think about a speaker's personality. For example, pauses can make a conversationalist appear less adept (McLaughlin and Cody, 1982), and also have implications for interpretations of what the speaker does or does not know (Brennan and Williams, 1995). Saying *um* can make people who know the answer to a question appear less sure of their answer, or give the appearance that someone who doesn't know the answer really does know it. *Um*s can also make people appear more relaxed compared to speech with the *um*s replaced by pauses, although pauses make people appear less relaxed than no pauses (Christenfeld, 1995). Offline tasks have also demonstrated that pauses can aid in syntactic disambiguation (Price *et al*., 1991), and, if placed at syntactic boundaries, can improve recall of the gist of sentences (Reich, 1980). *Um*s and *uh*s can provide turn-ending or turn-continuation information, depending on whether they fall at grammatical or ungrammatical points (Cook and Lalljee, 1970).

In online tasks, disfluencies have been shown to produce a variety of effects. *Uh*s speed up the recognition of upcoming words in sentences but *um*s don't, a result that can be attributed to their differing roles in anticipating the lengths of upcoming pauses (Fox Tree, 2001). Attention may be heightened after hearing an *uh* in anticipation of the short upcoming pause and continuation, but it may not be after *um* because of the indeterminacy of the upcoming delay. Repetitions do not negatively affect recognition of subsequent words (Fox Tree, 1995), as would be expected if repetitions are a solution to a fluency problem as opposed to an error (Clark and Wasow, 1998). But certain kinds of restarts, those altering information mid-sentence, do slow recognition (Fox Tree, 1995). Restarts are only costly when listeners need to store information about one part of the discourse record while making the correction. Information between the suspension and resumption can help listeners follow the speaker successfully (Fox Tree and Schrock, 1999).

References

Blackmer ER and Mitton JL (1991) Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition* **39**: 173–194.

Boomer DS and Dittmann AT (1962) Hesitation pauses and juncture pauses in speech. *Language and Speech* **5**: 215–220.

Brennan SE and Williams W (1995) The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* **34**: 383–398.

Brotherton P (1979) Speaking and not speaking: processes for translating ideas into speech. In: Siegman AW and Feldstein S (eds) *Of Speech and Time*. New York: Wiley.

Christenfeld N (1995) Does it hurt to say um? *Journal of Nonverbal Behavior* **19**(3): 171–186.

Clark HH (1994) Managing problems in speaking. *Speech Communication* **15**: 243–250.

Clark HH (1996) *Using Language*. New York: Cambridge University Press.

Clark HH and Wasow T (1998) Repeating words in spontaneous speech. *Cognitive Psychology* **37**: 201–242.

Cook M (1971) The incidence of filled pauses in relation to part of speech. *Language and Speech* **14**(2): 135–139.

Cook M and Lalljee M (1970) The interpretation of pauses by the listener. *British Journal of Social and Clinical Psychology* **9**: 375–376.

Duez D (1982) Silent and non-silent pauses in three speech styles. *Language and Speech* **25**(1): 11–28.

Ferreira F (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* **30**: 210–233.

Fox Tree JE (1995) The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* **34**: 709–738.

Fox Tree JE (2001) Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* **29**(2): 320–326.

Fox Tree JE and Clark HH (1997) Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* **62**(2): 151–167.

Fox Tree JE and Schrock JC (1999) Discourse markers in spontaneous speech: oh what a difference an oh makes. *Journal of Memory and Language* **40**: 280–295.

Goldman-Eisler F (1968) *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.

Hieke A, Kowal S and O'Connell DC (1983) The trouble with 'articulatory' pauses. *Language and Speech* **26**(3): 203–214.

Jefferson G (1974) Error correction as an interactional resource. *Language in Society* **2**: 181–199.

Jefferson G (1989) Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In: Roger D and Bull P (eds) *Conversation: An Interdisciplinary Perspective*. Philadelphia, PA: Multilingual Matters.

Kowal S, Bassett MR and O'Connell DC (1985) The spontaneity of media interviews. *Journal of Psycholinguistic Research* **14**(1): 1–18.

Levelt WJM (1983) Monitoring and self-repair in speech. *Cognition* **14**(1): 41–104.
Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Lickley RJ and Bard EG (1998) When can listeners detect disfluency in spontaneous speech? *Language and Speech* **41**(2): 203–226.
Maclay H and Osgood CE (1959) Hesitation phenomena in spontaneous English speech. *Word* **75**: 19–44.
Martin JG (1967) Hesitations in the speaker's production and listener's reproduction of utterances. *Journal of Verbal Learning and Verbal Behavior* **6**(6): 903–909.
Martin JG and Strange W (1968) The perception of hesitation in spontaneous speech. *Perception & Psychophysics* **3**(6): 427–438.
McLaughlin ML and Cody MJ (1982) Awkward silences: Behavioral antecedents and consequences of the conversational lapse. *Human Communication Research* **8**(4): 299–316.
Price PJ, Ostendorf M, Shattuck-Hufragel S and Fong C (1991) The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America* **90**(6): 2956–2970.
Reich SS (1980) Significance of pauses for speech perception. *Journal of Psycholinguistic Research* **9**(4): 379–389.
Reynolds A and Paivio A (1968) Cognitive and emotional determinants of speech. *Canadian Journal of Psychology* **22**(3): 164–175.
Schachter S, Christenfeld N, Ravina B and Bilous F (1991) Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* **60**(3): 362–367.
Smith VL and Clark HH (1993) On the course of answering questions. *Journal of Memory and Language* **32**: 25–38.
Swerts M (1998) Filled pauses as markers of discourse structure. *Journal of Pragmatics* **30**: 485–496.
Tannenbaum PH, Williams F and Hillier CS (1965) Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior* **4**: 134–140.
Wingate ME (1984) Fluency, disfluency, dysfluency, and stuttering. *Journal of Fluency Disorders* **17**: 163–168.

Further Reading

Clark HH (1996) *Using Language*. New York, NY: Cambridge University Press.
Fox Tree JE (2000) Coordinating spontaneous talk. In: Wheeldon LR (ed.) *Aspects of Language Production*, pp. 375–406. Philadelphia, PA: Psychology Press.
Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Nofsinger RE (1991) *Everyday Conversation*. Prospect Heights, IL: Waveland.

Distinctive Feature Theory

Introductory article

Elizabeth Hume-O'Haire, Ohio State University, Columbus, Ohio, USA
Stephen Winters, Ohio State University, Columbus, Ohio, USA

CONTENTS

The term 'distinctive features' is used in phonology to refer to the minimal units of sound that serve to distinguish the meaning of one word from another within a language. Distinctive features generally correspond to a specific articulatory or acoustic property of sound.

INTRODUCTION

Distinctive features are the universal set of cognitive properties associated with the speech sounds that are used in language. Distinctive features determine the contrasts which may exist between speech sounds, account for the ways in which these sounds may change, or *alternate*, and define the sets of sounds – known as natural classes – which may behave similarly in language. This limited set of features enables linguists to make powerful predictions about the kinds of sound structures that are expected to exist in the languages of the world. As universal properties of speech sounds, distinctive features also provide key insights into the cognitive organization of sound in language. The fact that distinctive features have such explanatory power makes their discovery one of the most important advances in linguistic science during the twentieth century.

THE FUNCTIONS OF DISTINCTIVE FEATURES

Contrast

The notion of contrast in language has been important for as long as linguists have been aware that the relationship between sound and meaning is arbitrary. For example, there is no logical connection between the sounds used to express the word 'tree' and the actual meaning of the word 'tree'. It does not matter, therefore, what particular sounds a language chooses to express the meaning of a word like 'tree'; what matters is that the sounds combined to form this word are distinct from those that combine to denote a word with a different meaning, such as 'true'. In other words, the sounds of language are meaningful only inasmuch as they *contrast* with one another.

Linguists originally conceived of contrast as a relationship that held between two entire sounds, such as between the [b] and [p] in a pair of words like 'bit' and 'pit'. Further investigation revealed, however, that such contrasts were always based on certain aspects of the articulation (i.e. the pronunciation) or the acoustics (i.e. the physical characteristics of sound waves) of the individual sounds. A [b] and a [p] differ, for instance, only in that the articulation of [b] involves the low-frequency vibration of the vocal cords (denoted by the distinctive feature [voice]) while the articulation of [p] does not. Likewise, the sound [s] contrasts with the sound [θ] (which is represented in English by *th*, as in 'with') in that, acoustically, [s] is louder and has more energy concentrated in higher frequencies, while [θ] is quieter and has acoustic energy spread across a broad range of frequencies. This acoustic distinction is represented with the feature [strident]. Any two sounds can differ meaningfully only in terms of such 'distinctive features' of their articulations or acoustics. Furthermore, linguists have observed that there are only a limited number of such features that can make meaningful distinctions. This insight has the important implication that the number and kind of meaningful sound distinctions in any language is limited by the set of available distinctive features. The fact that language is limited in this way – and is not completely arbitrary – provides significant evidence for the universal cognitive structures the mind imposes on language.

Natural Classes

Distinctive features further organize language by defining groups of sounds which may exhibit similar sound patterns. A *natural class* of sounds in a language consists of those sounds which share certain distinctive features to the exclusion of all other sounds in the language. Such natural classes of sounds often pattern together in similar ways. For example, the labio-velar sound [w], as in 'wit', cannot follow a specific group of sounds in English; [w] may follow [d] or [k] sounds, as in 'dwell' or 'quell', but it may never follow sounds like [b], [f], or [m]. That is, there are no words in English like 'bwell', 'fwell', or 'mwell'. The group of sounds which [w] cannot follow in English is collectively known as the natural class of *labial* consonants. The distinctive feature [labial] characterizes sounds that are articulated with the lips; as such, [labial] generalizes over the more specific phonetic categories of *bilabial* (pronounced with two lips, as in [b] or [p]) and *labio-dental* (pronounced with both the lips and the teeth, as in [f] or [v]). Though such finer distinctions could be made, phonological systems ignore them. The fact that distinctive features may draw broader distinctions than objectively exist in speech production or transmission provides further evidence for their essentially cognitive status. The observation that it is the sound [w] that fails to occur before a [labial] consonant also follows from the view that sounds defined by a common feature are predicted to pattern together, such as in the [labial] cooccurrence restriction; labio-velar [w] also belongs to the natural class of [labial] sounds.

Alternations

Speech sounds may also *alternate* with one another under certain conditions. The voiceless [s] sound, which may designate plural formations in English, alternates with a voiced [z] sound whenever it follows any of the natural class of voiced sounds at the end of a word which it pluralizes. Voiced [z] is always found after the voiced [b] in 'cabs', for instance, while the voiceless [s] always follows the voiceless [p] in 'caps'. The alternation here is not simply between the sounds [s] and [z] but between the distinctive feature of voicing which distinguishes the two. In fact, all sound alternations in language may be defined in terms of distinctive features. Furthermore, since distinctive features define natural classes of sounds, an entire group of sounds (as opposed to just one) is predicted to condition or undergo the same alternation. Thus, the pronunciation of the English plural ending as voiceless [s] occurs after all words ending in a voiceless sound, for example, [p, t, k, f]. Characterizing alternations in terms of distinctive features therefore not only captures the observation that speech sounds do not alternate arbitrarily but also provides linguists with a powerful predictive device.

THE NATURE OF DISTINCTIVE FEATURES

Phonetic Grounding

Although the primary motivation for the existence of a feature comes from considerations of language sound patterning as noted above, features are defined in the phonetic terms of articulations and acoustics. For example, the common articulatory property of sounds defined by the feature [coronal] for both consonants and vowels is a constriction involving the front or mid-portion of the tongue. Speakers use such articulations to produce sounds like [ʃ] (usually represented in English with the spelling *sh*), as in 'shirt', and [i], as in 'magazine'. Acoustically, the vocal tract configuration that this articulation creates results in a sound produced with a greater concentration of energy among the higher frequencies used in speech sounds.

Table 1 provides a list of some commonly used distinctive features along with their articulatory definitions.

Feature Values

A feature's ability to minimally contrast two sounds with each other is generally represented by listing that feature with either of two opposite values, indicated before a given feature by a plus or minus symbol. For example, the property of voicing which distinguishes [b] from [p] in English is defined by the feature values [+voice] and [−voice], respectively. The characterization of a sound as [+voice] indicates that it bears the property of vocal cord vibration, while the value [−voice] defines sounds which lack vocal cord vibration. Thus, it is possible to refer to natural classes defined by each of these two feature values.

There are certain features, however, for which the negative value never defines a natural class of sounds. In such cases, these features contrast by virtue of either their presence or absence in a sound; there is no negative feature value, as there is for voicing, which can be referred to in the formal description of the sound system. Such features are considered to be unary-valued, or privative. Place

Table 1. Some distinctive features and their articulatory definitions

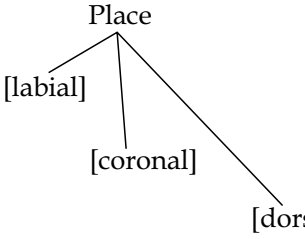
[sonorant]	Sounds which are produced without an obstruction which would impede the flow of air that has passed through the vocal cords in voicing. Sonorant sounds include vowels, nasals (e.g. [n, m]), liquids (e.g. [l, r]), and glides (e.g. [j, w]).
[voice]	Sounds produced with vocal cord vibration, e.g. [b, z, n, r, i, w].
[nasal]	Sounds produced with a lowered velum, which allows air to pass through the nose, as in nasal consonants and nasalized vowels.
[continuant]	Sounds produced in such a way that air flows continuously through the vocal tract as in, among others, vowels and fricatives, e.g. [f, s].
[lateral]	Sounds produced by lowering the mid-section of the tongue at the sides, allowing air to flow out near the molars, e.g. [l].
[strident]	Sounds produced in such a way as to create a turbulent noise, e.g. [f, v, s, z].
[labial]	Sounds produced with a constriction formed by the lower lip, e.g. [p, b, f, v, m, o, u].
[coronal]	Sounds produced with a constriction formed by the front or mid-portion of the tongue, e.g. [t, z, ʃ, i, e].
[dorsal]	Sounds produced with a constriction formed by the back of the tongue, or dorsum, e.g. [k, g, u, o].
[high]	Sounds produced by raising the body of the tongue towards the roof of the mouth, e.g. [i, u].
[low]	Sounds produced by lowering the body of the tongue away from the roof of the mouth, e.g. [a].

of articulation features ([labial], [coronal], [dorsal]) are typically considered to be unary-valued, for example. Since it is impossible to refer to sounds defined by the lack of a given property, this approach makes the strong prediction that sounds defined by the absence of a given feature, for example nonlabial sounds, will not function as a natural class in sound systems.

FEATURE GEOMETRY

In the earliest versions of distinctive feature theory, all of the features that comprised a speech sound were considered to have the same status – no features were more closely related to each other than to any others. In more recent years, however, linguists have begun to recognize that there is internal organization of the features within a speech sound. The fact that certain features commonly pattern together motivates the idea that there are 'natural groupings' of features into higher-level functional units. For example, in many languages the features [labial], [coronal], and [dorsal] function together as a unit. In English, for instance, the nasal sound in a prefix such as 'en' typically takes on the place of articulation value of a following consonant, a process commonly referred to as *place assimilation*. Thus, the nasal is pronounced as coronal [n] before coronal sounds (as in 'encircle', 'endanger'), as labial [m] before labial sounds ('empower', 'embitter'), and as a dorsal nasal [ŋ] ('ng') before dorsal sounds ('encamp', 'encourage'). Notice that all three place features are involved and, furthermore, it is *only* the place of articulation value that is adopted from the following consonant. The three place features are thus patterning together as a functional unit.

An influential model of feature representation known as *feature geometry* incorporates the insight that certain features regularly function together by grouping these sets of features into constituents. Evidence from processes such as nasal place assimilation suggests that the place features [labial, coronal, dorsal] are grouped together into a single Place unit, as shown in (1). Since sounds are defined in terms of more than just place features, the representation in (1) constitutes only one of potentially many feature groupings that, combined together, define the universal geometry of speech sounds.



(1)

A crucial assumption embodied in this type of model is that phonological processes are described most simply by reference to a single element – either a feature (e.g. [labial], [voice]) or the head of a constituent (e.g. Place). As a result, the model

makes strong predictions about the way in which features (and hence, speech sounds) function in language. For example, a phonological process may be described by reference to an entire constituent, such as Place, as would be the case for the process of nasal place assimilation in English described above. Or, a phonological process may be described in terms of any of the features individually. As seen above in the discussion of English plural formation, for example, the phonological process resulting in the alternation between voiceless [s] and voiced [z] can be described simply by reference to the feature [voice]. However, since processes are defined by reference to a single element, the model rules out phonological processes that would be described by a subset of features headed by a constituent. Given the feature grouping in (1), for instance, the sounds characterized by the features [coronal] and [labial] should never pattern together to the exclusion of those with the feature [dorsal]. Thus, no language should have a nasal assimilation rule in which the nasal consonant acquired the place feature of a following [coronal] or [labial] consonant, but not that of a [dorsal] consonant.

FEATURES AND MARKEDNESS

Linguists have long observed that, across languages, certain speech sounds are more common than others. The most common sounds are considered *unmarked* while progressively rarer sounds are considered more and more *marked*. Interestingly, there are certain groups of marked segments which never exist in a language unless another, less marked group of segments does as well. For instance, a language will never have voiced stops unless it also has voiceless stops, or nasal vowels without also having oral vowels. Since the existence of these marked natural classes always implies the existence of the less marked class, the relationship that holds between them is called an *implicational law*. Distinctive feature theory elegantly captures such cross-linguistic relationships between natural classes; for example, the implicational law that holds between voiced and voiceless stops can be characterized as: [+voice, −continuant] → [−voice, −continuant]. Informally stated, this rule claims that any language with voiced stops such as [b, d, g] must also have voiceless stops such as [p, t, k].

Universal patterns of markedness have also been claimed to correspond to the order in which children acquire contrasts (and thereby features) during language acquisition. Children generally acquire less marked sounds before acquiring the contrasts that define distinct, but more marked sounds in the system; extremely marked sounds may sometimes never be acquired at all. An apt example of the latter is the English *r* sound whose markedness stems in part from its articulatory complexity: it is commonly produced with multiple articulations, including lip rounding, retroflexion (curling or bunching of the front of the tongue), and pharyngeal constriction (a narrowing of the passageway between the root of the tongue and the back of the pharynx). A possible featural description of the sound would thus need to include three place features: [labial], [coronal], and [pharyngeal]. Conversely, the less marked [t] sound is specified for only a single place feature, [coronal]. The inherent difficulty of acquiring more marked sounds, such as English *r*, means that these sounds are not as likely to be passed on to succeeding generations as less marked sounds are. It only follows, therefore, that less marked sounds will appear in more languages than marked ones will.

CONCLUSION

Analyzing speech sounds in terms of fundamental properties known as distinctive features accounts for a wide variety of phenomena in language sound systems, including contrast, the grouping of sounds together into natural classes, and the alternations of sounds in various contexts. Along with this evidence, regular patterns in markedness, language acquisition, and historical change strongly attest to the psychological reality of features and the contrasts they define in speech sound systems. Understanding the nature of distinctive features and the various functions they have therefore provides crucial insights into the cognitive organization of sound in language. (*See* **Phonology**)

Further Reading

Anderson SR (1985) Roman Jakobson and the theory of distinctive features. *Phonology in the Twentieth Century,* chap. 5. Chicago and London: University of Chicago Press.

Chomsky N and Halle M (1968) The phonetic framework. *The Sound Pattern of English,* chap. 7. New York: Harper & Row.

Clements GN and Hume E (1995) The internal organization of speech sounds. In: Goldsmith J (ed.) *Handbook of Phonological Theory*, pp. 245–306. Oxford, UK and Cambridge, MA: Blackwell.

Halle M (1983) On distinctive features and their articulatory implementation. *Natural Language and Linguistic Theory* **1**: 91–105.

Jakobson R, Fant G and Halle M (1952) *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

Kenstowicz M (1994) *Phonology in Generative Grammar*, chaps 1 (The sounds of speech), 4 (The phonetic foundations of phonology), and 9 (Feature geometry, underspecification and constraints). Oxford, UK and Cambridge, MA: Blackwell.

McCarthy JJ (1994) The phonetics and phonology of Semitic pharyngeals. In: Keating P (ed.) *Phonological Structure and Phonetic Form: Papers from Laboratory Phonology III*, pp. 191–233. Cambridge, UK: Cambridge University Press.

Ellipsis

Intermediate article

Robert C May, University of California, Irvine, California, USA

CONTENTS

A linguistic ellipsis, most generally expressed, is a truncated or partial linguistic form. It is a linguistic form in which constituents normally occurring in a sentence are superficially absent, licenced by structurally present prior antecedents.

INTRODUCTION

A linguistic ellipsis, most generally expressed, is a truncated or partial linguistic form. This partiality is measured relative to a complete sentence; an elliptical sentence is one in which some of the constituent parts of a 'full' sentence are missing. For example, in answer to the question 'Who went to the store', one may answer 'Max went to the store'. In most contexts, however, speakers would avoid such prolixity and instead would employ an elliptical form: 'Max went', 'Max did', or even just 'Max' would suffice as answers to the question. The reason that they would suffice appears to be quite obvious; it is because they mean just what 'Max went to the store' means, except they express this more economically by leaving off at least part of what can otherwise be gleaned from the initial question. From such simple examples we can already observe a fundamental property of ellipsis that we would expect to be captured under any account of the relation between elliptical and non-elliptical forms: meaning is constant under ellipsis. Specifying this constancy is not as straightforward as it may initially appear, however, and trying to capture what it amounts to has been an on-going issue in linguistic theory. It has been an issue of particular importance because the accounts of elliptical phenomena have been used to bring empirical weight to fundamental claims about linguistic theory, including the extension of the notion of linguistic identity, the relation of syntax and semantics, and the abstractness of grammar.

To frame our discussion, we note the most well-known types of elliptical constructions that have been studied in the literature:

VP-ellipsis: Max went to the store, and Oscar did, too
gapping: Max went to the store, and Oscar to the arcade
sluicing: Max went to the store, but Oscar wondered why
stripping: Max saw Sally at the store, and Oscar, too
pseudo-gapping: Max loves Jane, and Harry does Sally
N'-ellipsis: Max's father went to the store, but Oscar's went to the arcade

Each of the these types of ellipsis have idiosyncratic properties. For example, with gapping there is a well-known correlation of the direction of gapping and word order, and with sluicing the restriction that the complementizer of a sluiced clause must be interrogative; these are among observations originating with Ross (1969, 1970). Our purpose here is not, however, to survey the differences between these various elliptical phenomena but to explore two fundamental properties that they all have in common: (1) ellipsis is of a syntactic constituent; (2) an antecedent occurrence of the elided constituent governs the ellipsis. For example, in VP-ellipsis the elided material is a verb phrase, and the ellipsis is licensed in the presence of a fully lexicalized antecedent: in the case above, 'went to the store'. The primary issue for ellipsis is how to properly characterize (1) and (2); what sorts of linguistic description are called for to capture these fundamentally grammatical aspects of ellipsis in their full generality? A number of subsidiary issues are implied by answers to the primary questions, including the nature of the grammatical mechanisms that account for the absence of the lexical material: is it a deletion of underlyingly present syntactic elements; or are these elements absent even at the underlying level? In what follows we will outline some of the main approaches that have been developed to these issues, focusing primarily on the case of VP-ellipsis, largely because the relevant

theoretical issues have been most clearly and widely discussed in this context.

VP-ELLIPSIS

Syntactic Reduction: The Transformational Theory

The initial approaches to ellipsis within contemporary linguistic theory attempted to account for the semantic constancy alluded to above by reducing it to syntactic constancy. The idea here is quite intuitive: if an elliptical form can be seen as a syntactic repetition of a corresponding non-elliptical form, i.e. as simply two occurrences of the same syntactic form, then it would follow immediately that they have the same meaning. The first systematic investigations along these lines is found in the work of J. R. Ross in the late 1960s (Ross, 1967, 1969). The approach plays itself out very naturally with respect to the deep structure/surface structure distinction: elliptical and non-elliptical forms have the same deep structure but different surface structures, the difference arising from the application of transformations that delete syntactic structure to give the elliptical forms. That (1) and (2) fall together was taken as a natural consequence of assuming that the transformational rules involved delete syntactic constituents and do so under identity with an antecedent constituent. So, for example, because there is an antecedent occurrence of the verb phrase 'went to the store', 'Oscar did, too' can be derived from 'Max went to the store' by VP-deletion.

During this period in the development of generative grammar, it was generally assumed that deletion transformations were constrained by a general theoretical condition that required the 'recoverability' of deletions; cf. Peters and Ritchie (1973) for discussion of the formal importance of this condition for the theory of transformations. Informally put, a deletion transformation satisfies this condition if it is possible to reconstruct the deleted material from within the structural context that the rule applies. Deletions that apply under identity obviously meet this criterion. For example, VP-deletion would appear to meet the condition because within the overall sentence 'Max went to the store, and Oscar did, too' we can reconstruct the deleted VP as a copy of the antecedent VP. However, two problems were noticed that indicated quite clearly that such deletion rules fail to meet the recoverability criterion.

The first problem is that the domains in which transformational rules apply are not the same as the domains in which deleted constituents can match up with their antecedents. While the domain of transformational rules is the sentence, the antecedent/deletion pairing is not restricted to this domain but transcends sentential boundaries. So, in our example of VP-deletion, we could replace the conjunction with a full stop, turning one sentence into two: (1) 'Max went to the store' and (2) 'Oscar did, too'. Moreover, we could imagine discourses in which other sentences would be interpolated between them or, even more telling, discourses in which the two sentences were uttered by different speakers. VP-deletion, it would appear, is not a rule of *sentence* grammar, but of *discourse* grammar.

Notice that broadening the applicability of syntactic deletion operations to discourse does not impact on what would seem a more fundamental aspect of recoverability, the reconstructive aspect of deletion given by the syntactic identity of the deleted constituent and its antecedent. The following observations, initially made by Ross (1967) do, however, and herein lies the second problem. Consider the following sentence:

Max saw his mother, and Oscar did, too (1)

Clearly, there is a reading of (1) that entails that Max and Oscar saw one and the same person; they each saw Max's mother. This interpretation could easily be obtained on the transformational view by taking the deep structure of (1) to be roughly (1′):

Max saw Max's mother, and Oscar saw Max's mother (1′)

(1) can derived from (1′) first by applying VP-deletion, applicable given the identity of the verb phrases, to be followed by pronominalization in the first clause. The construal of (1) characterized in this way is not, however, the only construal available of this sentence. It can also be understood in a manner comparable to (1″):

Max saw Max's mother, and Oscar saw Oscar's mother (1″)

The problem is that although (1″) ought to be a possible underlying source for (1), it is not a structure to which VP-deletion can apply, since the verb phrases are not identical. Insofar as (1) can be derived from an underlying form like (1″), it will be a nonrecoverable deletion.

Semantic Non-Reduction: The Property Theory

The ambiguity shown by (1) between a 'strict identity' reading (1′) and a 'sloppy identity' reading (1″)

thus scotches a transformational account of the sort envisaged as deletion under identity. But what went wrong? The answer that emerged is that the underlying problem lies with the initial presumption, that semantic constancy can be reduced to syntactic constancy. Deletions, in some sense, must still be recoverable, not least because knowing what has been deleted is necessary for understanding an elliptical sentence. But what is to be recovered is not syntactic information *per se*, but information about the logico/semantic roles played by syntactic expressions; it is in the identity conditions applicable to these roles that we are to look for the conditions that govern deletion. The semantic constancy of ellipsis, in this view, is not to be reduced to something else, but is to have a direct semantic analysis.

The nonreductionist approach found its fullest hearing in the work of Ivan Sag (1976) and Edwin Williams (1977) in the mid-1970s. The central observation animating this approach is that a sentence such as 'Max saw his mother' is ambiguous, depending upon whether the verb phrase 'saw his mother' expresses the property of seeing Max's mother or the property of seeing one's own mother. These two properties are not unrelated; the latter is an abstraction of the former, being general where the former is particular. This ambiguity, however, is masked in simple sentences because 'Max saw his mother' has the same truth conditions under either interpretation of the verb phrase. It becomes unmasked in elliptical contexts: depending on which property the ellipsis is taken as being identical with, different interpretations are obtained. Thus, if in 'Max saw his mother, and Oscar did, too', the ellipsis is understood as the property 'saw Max's mother' the strict reading ensues; if the ellipsis is understood as 'saw his own mother', the sloppy reading follows. In either case note that what is elided is identical with the antecedent; they each express the same property. In this view, the constancy of ellipsis is thus a matter of property identity, and this is an inherently semantic notion.

These ideas found a natural representation by assuming that the logical representation of natural language incorporate aspects of a λ-calculus. These sorts of logistic systems (first introduced by Church), incorporate an operation that abstracts properties from propositions. This operation, known as λ-abstraction, derives from the proposition expressed by 'Max read *Moby Dick*' the property: λx (x read Moby Dick). This is to be parsed as a λ-operator binding a variable in the following open sentence; it is interpreted as a characteristic function, taking an individual as argument and returning a truth value. If we supply an argument for this function, represented by placing it before the λ-expression: Max, λx (x read Moby Dick), we can return to our original proposition by the inverse operation, λ-conversion; we effect this by placing the argument in the place of the variable in the open sentence, and erasing the λ-operator.

In their analysis of ellipsis, Sag and Williams make two basic assumptions about the semantics of natural language. First, following a suggestion of Barbara Partee (1975), they assume that verb phrases are interpreted as λ-expressions, so that the logical form of 'Max read Moby Dick' would be represented as immediately above. Second, they assume that anaphoric pronouns can be interpreted as either constants or variables. Taking these together, it follows that 'Max saw his mother' is representationally ambiguous between the following logical forms: 'Max, λx (x saw Max's mother)' and 'Max, λx (x read x's mother)'. In the first form, the pronoun is represented as a constant specifying its anaphoric reference; in the latter, the pronoun occurs as a bound variable. Thus far these representations are only distinguished formally; via λ-conversion both convert to the same proposition. The distinction becomes more than this, however, when a third assumption comes into play: VP-ellipsis requires identity of λ-expressions. This gives two representations for 'Max saw his mother, and Oscar did, too':

Max, λx (x saw Max's mother) & Oscar, λy (y saw Max's mother) (2)

Max, λx (x saw x's mother) & Oscar, λy (y saw y's mother) (3)

The first representation is of the *strict* reading; the second clause means Oscar saw Max's mother. The second representation is of the *sloppy* reading; in it, the second clause means that Oscar saw his own mother; that is, Oscar saw Oscar's mother.

The part of the identity conditions in λ-expressions relevant to strict and sloppy identity is known as the alphabetic variance condition. This condition breaks down into two subcases. The first applies to λ-expressions if there are only bound occurrences of variables; in this case, the λ-expressions must be exactly the same up to alphabetic values of the variables. Thus, in (3) for example, the λ-expression on the left is nondistinct from that on the right because they are alphabetic variants, differing only in that where 'x' occurs on the left, 'y' occurs on the right. That is, it matters not that we have 'x' on the right and 'y' on the left so long as the

pattern of binding remains unaltered. When this is changed, the condition is not satisfied; a consequence of this is that a sloppy reading is unavailable in 'Max saw his mother, and Oscar believes Jane did, too'; i.e. the right-hand clause cannot mean that Oscar believes that Jane saw Oscar's mother. This is because in the following representation, which would represent this reading, there are no λ-expressions that are alphabetic variants: 'Max, λx (x saw x's mother)' and 'Oscar, λz (z believes Jane, λy (y saw z's mother))'. In particular, 'λx (x saw x's mother)' and 'λy (y saw z's mother)' are not alphabetic variants because 'z' is free within the λ-expression while the corresponding occurrence of 'x' is bound. Not all free occurrences of variables within λ-expressions are illicit however; this is the effect of the second case of the condition that permits λ-expressions to be alphabetic variants only if the free variables are all bound by the same operator. This is what we find in the representation of 'Max saw everyone before Oscar did':

$$\forall x \text{ (Max, } \lambda y \text{ (} y \text{ saw } x\text{) before Oscar, } \lambda z \text{ (} z \text{ saw } x\text{))}$$

Although within each λ-expression there is a parallel occurrence of 'x' free, they are both bound by the universal quantifier, and hence are alphabetic variants.

The success of this account in overcoming the problems that plagued the prior transformational approach extends beyond the account of strict and sloppy identity. Because the notion of property identity on which the account depends is semantic, unlike in the syntactic account, which was limited by the structural extent of structural descriptions of transformational rules, there is nothing comparable that inherently restricts the context in which the identical properties may occur. Therefore, in the absence of some external constraint, the antecedent of an ellipsis may occur in positions quite detached in discourse from the ellipsis itself; the sentence in which the antecedent is expressed neither needs to be adjacent to the elliptical sentence, nor need it be uttered by the same speaker. All that is required is that the antecedent of the ellipsis be sufficiently salient in the surrounding context.

Property theory and the syntax of ellipsis

The property theory approach initiated by Sag and Williams has been highly influential in the study of ellipsis, and there have been any number of variations of this view. One important source of these variations arises from a changed perspective regarding what is the main syntactic issue raised by elliptical constructions. In the transformational deletion analysis, the concern was over what we may call the 'generation' problem: what are the syntactic operations that produce elliptical structures? But on the property theory, which assigns the explanatory role to semantics for the matters that were so troublesome *vis-à-vis* recoverability for the prior account, the focus is shifted to the mapping problem: how are syntactic structures translated into semantic structures? In particular, how are verb phrases translated into λ-expressions that satisfy the identity conditions (alphabetic variance)? Understanding the mapping problem in this way places a constraint on its solutions; however syntactic derivation is to be effected (i.e. whatever the solution to the generation problem is), it must be such that it allows for systematic translation. There are two broad approaches to the mapping problem falling within this constraint that differ in their views of the need to attribute syntactic structure to the ellipsis in order to generate the property it expresses.

The first, the rich syntax view, assumes that elliptical structures retain a syntactic relation to forms in which all constituents are structurally present. Thus, whatever procedures translate 'Oscar read *Moby Dick*' also translate 'Oscar did, too', because at the input to the translation, the latter has the same structure as the former. The rich syntax view thus calls for the 'reconstruction' of syntactic information in order to obtain a property that then serves as the basis of comparison for identifying salient antecedent properties in the context of an ellipsis. This is the view Sag and Williams take, although they differ on the derivational direction of this relation. Sag takes the elliptical structure to be derived from the non-elliptical by deletion, while Williams reverses the direction of the derivation, the non-elliptical arising from the elliptical by syntactic copying. The alternative, the poor syntax view, sees no need for such a syntactic relation; in this view, at all stages of derivation the elided phrase is missing, or if not actually missing is structurally noncomplex; i.e. an empty category with no internal constituents (cf. Hardt, 1993). Since there is no verb phrase, there is also no translation to a property, however, and hence, in contrast to the rich theory, there is no basis of comparison for determining the antecedent of an ellipsis. But, according to this view, there is in fact no need to derive this, for there is an independent model to draw upon, the anaphoric resolution of pronouns. In both the ellipsis/antecedent and the pronoun/antecedent relations, a possibly complex antecedent for a syntactically simple element is determined by the conditions on salience in

context. Thus, the relation of ellipsis to antecedent in 'Max read *Moby Dick*, and Oscar did, too' is comparable to the relation of pronoun to antecedent in discourse (e.g. 'Herman Melville wrote *Billy Budd*. He is more famous, however, for *Moby Dick*') or even more directly to VP-anaphors like 'so' and 'it' (e.g. 'Max read *Moby Dick*, and so did Oscar') for which elided verb phrases are, in this view, the covert analogues. In the poor syntax view, then, there is no reconstruction as in the rich view; information relevant to the resolution of the ellipsis is only that which is found in the antecedent.

Problems with the property theory

In subsequent research a number of problems have emerged with the property theory. One case, initially observed by Shalom Lappin (1984) calls into question the validity of the second clause of the identity condition in λ-expressions on the basis of examples such as (4):

I know which book Max read, and which book Oscar didn't (4)

Recall that the restriction that λ-expressions are alphabetic variants only if the free variables are all bound by the same operator is what accounted for 'Max saw everyone before Oscar did' as discussed above. But in allowing for this case, (4) should be disallowed, for here the λ-expression corresponding to the elided phrase – λz (z read x) – contains a free variable bound by a different operator (i.e. the *wh*-phrase in the second clause) than that which binds the free variable in the λ-expression corresponding to the antecedent phrase.

A second sort of case, derived from initial observations of Schiebe (1971) and Dahl (1974), is among a series of cases most extensively discussed in joint research by Robert Fiengo and Robert May (Fiengo and May, 1994), who label them the 'eliminative puzzles of ellipsis'. It does not pertain directly to the identity condition but to a systematic overgeneration problem in the translation of pronouns. This case, which Fiengo and May called the 'many-pronouns puzzle', arises when the number of pronouns is increased beyond the one found in the standard examples used to illustrate strict and sloppy identity:

Max said he saw his mother, and Oscar did, too (5)

Here, the expectation is that there should be a four-way ambiguity. This is because whether a pronoun is translated as a constant or a variable is independent of how any other pronoun is translated, predicting, for n-many pronouns, 2^n readings. Thus in (5) the pronouns could be (i) both variables, (ii) both constants, (iii) the first one a variable, the second a constant, or (iv), vice versa, the first a constant, the second a variable. Given the correspondence of variables with the sloppy reading, and constants with the strict reading, (i) and (ii) will result in 'across-the-board' sloppy and strict readings, respectively, while (iii) and (iv) will give readings mixed between sloppy and strict. The problem is that in (5) we observe only three, not four, readings: precluded is the reading corresponding to (iv). The ellipsis in (5) cannot be glossed as '… and Oscar said Max saw Oscar's mother'. What we actually observe in this case, as well as in those of increasing complexity, are only $n+1$ readings; readings do not grow exponentially.

Syntactic Reduction Redux: The Dependency Theory

What the 'many-pronouns puzzle' indicates is that the assumption of translational independence embedded in the property theory's account of anaphora is incorrect, suggesting instead that what is involved is some dependence relation between the pronouns. The most highly developed theory that seeks to capture these dependencies is the dependency theory developed by Fiengo and May (1994). Central to the dependency theory picture is that dependencies must have a syntactic characterization, and in establishing this result, Fiengo and May turn away from the nonreductive account of ellipsis embedded in the property theory, and return to a reductive syntactic approach. The dependency theory assumes that a sentence such as 'Max saw his mother' is structurally ambiguous, the ambiguity being attributed to a distinction in the representation of anaphoric pronouns that indicates whether the pronoun is formally dependent on its antecedent or not. By Fiengo and May's conventions, grammatically anaphoric pronouns are represented by co-indexing, with those that are formally dependent on their antecedent marked by 'β', those that are not by 'α'. 'Max saw his mother' thus has the following pair of representations: 'Max_1 saw his_1^{α} mother' and 'Max_1 saw his_1^{β} mother'. The dependency that a β-marked pronoun enters into is specified via a structural description of the sequence of categories that lies between the pronoun and its antecedent. So for the latter structure, this would be the dependency: $\langle(Max, his), 1, \langle NP, V, NP\rangle\rangle$, where the first member of the triple is the elements of the dependency, the (unique) antecedent and the dependent pronouns, the second

the index of the elements in the dependency, and the third the string of categories that links the co-indexed elements together.

The dependency theory is applied to ellipsis by holding that the identity conditions that allow for ellipsis are satisfied in the following two sorts of structure, where, given the co-indexings, the first represents the strict reading and the second represents the sloppy reading.

Max_1 saw his_1^{α} mother, and Oscar_2
saw $\textit{his}_1^{\alpha}$ *mother*

Max_1 saw his_1^{β} mother, and Oscar_2
saw $\textit{his}_2^{\beta}$ *mother*

While it is apparent that in the first structure the antecedent and elided verb phrases are simple syntactic copies, it is not in the second, for the pronouns are different in the two verb phrases. This discrepancy is reconciled by allowing for an identity condition such that dependencies are the same so long as there is the same sequence of categories, regardless of the index; dependencies that stand in this relation – same pattern, different index – Fiengo and May call '*i*-copies'. The dependencies in which the pronouns in the latter structure occur meet this criterion: since they are *i*-copies, they are sufficiently alike to allow for ellipsis even though the pronouns are syntactically distinct. On the other hand, where there are no dependencies to be calculated – where the pronouns are marked α, not β – there is no alternative but for the index of the pronoun to be unchanged.

Sloppy identity on the dependency theory view is thus the re-creation of an antecedent structural pattern of anaphora; strict identity, on the other hand, is the re-creation of the anaphora itself. Either may be extended to more complex structures, but when they are certain limitations arise. Thus, recall the 'many-pronouns puzzle' that swirled around examples such as (5) above – 'Max said he saw his mother, and Oscar did, too'. For this case, there are four possible combinations of indices for the antecedent; of these, only three give rise to well-formed elliptical structures:

Max_1 said he_1^{α} saw his_1^{α} mother, and Oscar_2
said $\textit{he}_1^{\alpha}$ *saw* $\textit{his}_1^{\alpha}$ *mother*

Max_1 said he_1^{β} saw his_1^{β} mother, and Oscar_2
said $\textit{he}_2^{\beta}$ *saw* $\textit{his}_2^{\beta}$ *mother*

Max_1 said he_1^{β} saw his_1^{α} mother, and Oscar_2
said $\textit{he}_2^{\beta}$ *saw* $\textit{his}_1^{\alpha}$ *mother*

$^{*}\text{Max}_1$ said he_1^{α} saw his_1^{β} mother, and Oscar_2
said $\textit{he}_1^{\alpha}$ *saw* $\textit{his}_2^{\beta}$ *mother*

The first case is the across-the-board strict reading; since both pronouns are α, they have the same index in the antecedent and the ellipsis. The second case is across-the-board sloppy: the two pronouns are in a dependency, with 'Oscar' as the antecedent in the same way structurally as the pronouns in the prior clause are in a dependency with 'Max'. The third structure represents the mixed reading: the first pronoun is β, and thus may be in a dependency. The second pronoun, however, is α so it must be strict, i.e. co-indexed with 'Max', not 'Oscar'. The final case is the one that is excluded. This is because the dependency that reaches from the pronoun 'his' to 'Oscar' as antecedent is not structurally identical to any dependency in the prior clause. Insofar as there is a dependency in that clause, it must be to the closer possible antecedent, the pronoun 'he', but this is not structurally parallel to the dependency in the clause with the ellipsis that has a greater syntactic extent.

Antecedent-Contained Deletion

Thus far, the structural context of ellipsis we have considered has been that of a discourse; i.e. the ellipsis and its antecedent have each occurred in independent sentences. While the examples we have examined are ones in which these sentences are conjoined, this is not essential, for in all the examples 'and' could be replaced by a full stop. The one exception was 'Max saw everyone before Oscar did', but notice here that the ellipsis is contained in an adjunct, not a subordinate, clause, so that even this case falls under the generalization that the ellipsis and its antecedent are syntactically independent. The generalization, however, appears to clearly fail in the following case:

Dulles suspected everyone that Angleton did (6)

Sentence (6) is well formed, and is naturally understood to mean the same thing as its unelided counterpart, 'Dulles suspected everyone that Angleton suspected'. But in this case the elided VP is not independent of its antecedent – rather it is contained within it – and hence the name for this construction, 'antecedent-contained deletion'. Notice that not only does antecedent-contained deletion appear to run counter to the generalization, but it does so in a particularly curious way. We understand the antecedent of the elided VP to be that VP headed by the verb 'suspected'. But if we plug that VP into the ellipsis, it will again contain the ellipsis: 'Dulles suspected everyone that Angleton suspected everyone that Angleton did'.

Additional iterations of the process will continually give structures that still contain an ellipsis. Given this vicious regression, it is unclear how we are to establish the relation of ellipsis and antecedent.

The critical insight here was provided by Sag (1976). Sag observed that if we attend to the logical form of (6), that at the appropriate level of syntactic description it falls under the antecedence generalization. This is because (6) contains a quantified phrase, and this must be scoped out; if we assume that the logical structure of (6) is roughly as follows: '$\forall x$: Angleton *suspected* x (Dulles suspected x)', we then need only observe that here the ellipsis (filled in and indicated by italics) is no longer contained within the antecedent. (This logical form is comparable to the logical form that would be assigned to the fully lexicalized counterpart of (6).) Thus, the significance of antecedent-contained deletion, given the generalization regarding the relation of ellipsis and antecedent, is that the notion of structure relevant to this generalization must be sufficiently abstract so as to represent the logical form of sentences. In May (1985) it is argued that this structure is a form of syntactic structure, a result of the syntactic rule QR, which gives (7) as a representation of (6):

[everyone that Angleton did [Dulles suspected t]] (7)

In this structure the ellipsis is no longer contained within the antecedent, and so we can now plug in the antecedent VP without any regress:

[everyone that Angleton *suspected* t [Dulles suspected t]] (7′)

May's account of antecedent-contained deletion has been one of the main arguments that has been cited in support of the view that there is a syntactic level – *LF* – that represents the logical structure of natural language; see May (1985, chap. 1), as well as Hornstein (1995), for a contrary view.

VP-ANAPHORA

Previously, we briefly mentioned the phenomena of VP-anaphora, citing examples like 'Max hit Oscar, and Harry did it, too' or 'Max hit Oscar, and Harry did so, too'. These cases are closely akin to VP-ellipsis except that they have an anaphoric element – 'so' or 'it' – rather than an ellipsis. VP-anaphora is distributionally more restricted than VP-ellipsis. For example, with stative verbs, VP-ellipsis is possible, but not the 'it' form of VP-anaphora, and the 'so' form is marginal; compare 'Max knows French, and Oscar does, too' with 'Max knows French, and Oscar does it, too' and 'Max knows French, and Oscar does so, too'. However, what is more relevant to the present discussion is that with VP-anaphora we can find the same ambiguity of strict and sloppy identity that we observed with VP-ellipsis; compare 'Max hit his mother, and Oscar did, too' with (8):

Max hit his mother, and Oscar did it, too (8)

As before, the second clause can be taken to mean that Oscar hit Max's mother (strict) or that Oscar hit Oscar's mother (sloppy).

Examples like (8) pose issues as to how we are to understand ellipsis. If VP-anaphora and VP-ellipsis display uniform behavior, it would seem natural to posit a uniform analysis. One way to do this would be to take ellipses as anaphoric elements, silent counterparts, if you will, of 'it' or 'so'. (For one account along these lines, couched in a variant of the property theory, see Hardt (1993).) In this view, in which VP-ellipsis is reduced to VP-anaphora, that ellipsis involves some sort of syntactic reconstruction is effectively denied. The alternative would be to run the reduction in the opposite direction by maintaining that at an appropriately abstract syntactic level VP-anaphora, like VP-ellipsis, is syntactically complex, and that the overt pronominal elements are but superficial syntactic reflexes. In either account, the goal would be to isolate a level of representation in which VP-anaphora and VP-ellipsis are structurally nondistinct in order to account for their common behavior.

This search for a common analysis, however, needs to be weighed against ways in which VP-anaphora and VP-ellipsis do not cluster but diverge in properties. One sort of divergence has already been mentioned: namely, the distributional differences. Another can be gleaned from examples of antecedent-contained deletion, as in (9):

Dulles talked to everyone that Angleton did (9)

If VP-ellipsis were just a variant of VP-anaphora, then we would expect that the VP-anaphora version of (9) would also be grammatical. But, as was observed by Fiengo and May (1994), it is not:

*Dulles talked to everyone that Angleton did it (9′)

This case indicates (along with the distributional facts cited above) that there are substantial differences between VP-anaphora and VP-ellipsis. It remains an open research question how observations like these can be integrated with those about

strict and sloppy identity; until then it remains equivocal whether VP-ellipsis and VP-anaphora are a unified phenomenon.

SUMMARY

Returning to our initial observation, we have seen that capturing the basic intuition of semantic constancy under ellipsis devolves to the issue of the proper way to state the identity conditions that govern ellipsis. That ellipsis in general requires a notion of identity was the initial insight of the transformational account; its flaw, in a sense, was that the notion of identity it had available was not sufficiently abstract. What followed were attempts to find the right locus of abstractness for ellipsis. The property theory argues that the appropriate identity conditions are semantic and does not fundamentally question the relative lack of abstractness of the underlying syntax. The dependency theory, in contrast, argues for a more abstract notion of syntax by refining the criteria for identity of occurrences of syntactic categories. But regardless of how matters turn out, it is clear that in seeking to fix the identity conditions for ellipsis, issues fundamental to our conceptions of linguistic description have been raised. Phenomena relevant to these issues extend beyond what we have been able to consider here. Among them are further interactions with anaphora, such as the 'vehicle change' effect noticed by Fiengo and May (1994): the observation that 'Mary loves John, and he thinks that Sally does, too' is interpreted as comparable to 'Mary loves John, and he thinks that Sally loves him, too', not 'Mary loves John, and he thinks that Sally loves John, too'. Moreover, we have left aside discussion of the conditions on discourse that allow an ellipsis to be resolved, including conditions on discourse coherence with respect to antecedents that may occur in sentences at some degree of removal in the discourse (cf. Kehler, 2000), and on the abstractness of discourse: e.g. consider when Butch says to Sundance at the edge of the cliff before jumping: 'I will if you will', indicating that the antecedent need not even be uttered (cf. Chao, 1987). We have tried, however, to highlight some of the core issues that have animated the discussion and that have made understanding elliptical phenomena of continued interest within linguistic theory.

References

Chao W (1987) *On Ellipsis*. PhD dissertation. University of Massachusetts, Amherst, MA.

Dahl Ö (1974) How to open a sentence: abstraction in natural language, *Logical Grammar Reports*, no. 12. Göteberg, Sweden: University of Göteberg.

Fiengo R and May R (1994) *Indices and Identity*. Cambridge, MA: MIT Press.

Hardt D (1993) *VP Ellipsis: Form, Meaning, and Processing*. PhD dissertation, University of Pennsylvania.

Hornstein N (1995) *Logical Form*. Oxford, UK: Blackwell.

Kehler A (2000) Coherence and the resolution of ellipsis. *Linguistics and Philosophy* **23**: 533–575.

Lappin S (1984) VP-anaphora, quantifier scope, and logical form. *Linguistic Analysis* **13**: 273–315.

May R (1985) *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.

Partee BH (1975) Deletion and variable binding. In: Keenan E (ed.) *Formal Semantics of Natural Language*, pp. 16–34. Cambridge, UK: Cambridge University Press.

Peters PS and Ritchie RW (1973) On the generative capacity of transformational grammars. *Information Sciences* **6**: 49–83

Ross JR (1967) *Constraints on Variables in Syntax*. PhD dissertation. Massachusetts Institute of Technology, Cambridge, MA.

Ross JR (1969) Guess who? In: Binnick RI, Davison A, Green GM and Morgan JL (eds) *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pp. 252–286. Chicago, IL: Department of Linguistics, University of Chicago.

Ross JR (1970) Gapping and the order of constituents. In: Bierwisch M and Heidolph K (eds) *Progress in Linguistics*. The Hague, Netherlands: Mouton.

Sag I (1976) *Deletion and Logical Form*. PhD dissertation. Massachusetts Institute of Technology, Cambridge, MA.

Schiebe T (1971) Zum Problem der grammatisch relevanten Identität. In: Kiefer F and Ruwet N (eds) *Generative Grammar in Europe*. Dordrecht, Netherlands: Reidel.

Williams E (1977) Discourse and logical form. *Linguistic Inquiry* **8**: 101–139.

Further Reading

Berman S and Hestvik A (eds) (1992) *Proceedings of the Stuttgart Ellipsis Workshop*. Heidelberg: Arbeitspapiere des Sonderforschungsbereichs 340, Bericht Nr. 29, IBM Germany.

Chung S, Ladusaw WA and McCloskey J (1995) Sluicing and logical form. *Natural Language Semantics* **3**: 239–282.

Dahl Ö (1973) On so-called 'Sloppy' identity. *Synthese* **26**: 81–112.

Dalrymple M, Shieber S and Perreira F (1991) Ellipsis and higher-order unification. *Linguistics and Philosophy* **14**: 399–452.

Hankamer J and Sag I (1984) Deep and surface anaphora. *Linguistic Inquiry* **7**: 391–426.

Kennedy C (1997) Antecedent contained deletion and the syntax of quantification. *Linguistic Inquiry* **28**: 662–688.

Lappin S (1996) The interpretation of ellipsis. In: Lappin S (ed.) *The Handbook of Contemporary Semantic Theory*, pp. 145–175. Oxford, UK: Blackwell.

Lappin S and Benmamoun EA (eds) (1999) *Fragments: Studies in Ellipsis and Gapping*. Oxford, UK: Oxford University Press.

Lobeck A (1995) *Ellipsis: Functional Heads, Licensing, and Identification*. New York, NY: Oxford University Press.

Merchant J (2001) *The Syntax of Silence – Sluicing, Islands, and the Theory of Ellipsis*. Oxford, UK: Oxford University Press.

Rooth M (1992) Ellipsis redundancy and reduction redundancy. In: Berman S and Hestvik A (eds) *Proceedings of the Stuttgart Ellipsis Workshop*. Heidelberg: Arbeitspapiere des Sonderforschungsbereichs 340, Bericht Nr. 29, IBM Germany.

Schwabe K and Zhang N (2000) *Ellipsis in Conjunction*. Tübingen, Germany: Max Niemeyer Verlag.

Steedman M (1990) Gapping and constituent coordination. *Linguistics and Philosophy* **13**: 207–264.

Zoerner CE (1995) *Coordination: The Syntax of &P*. PhD dissertation. University of California, Irvine, CA.

Finite State Processing

Intermediate article

Gertjan van Noord, University of Groningen, Groningen, Netherlands

CONTENTS

Finite state processing is the analysis of (natural) language by means of finite state automata. Finite state automata are computational devices with limited capabilities.

FINITE STATE AUTOMATA

A finite state automaton (Kleene, 1956) is a formal computational device which defines a language in the mathematical sense of a set of strings (where a string is a sequence of symbols). Finite state automata have finite memory; therefore they are only capable of defining languages of a particularly simple kind: the 'regular' languages. A finite state automaton is often represented by a transition diagram; some examples are shown in Figure 1.

Finite state automata provide a well-understood mechanism for the definition of simple languages: the theory of finite state automata is at the heart of theoretical computer science and mathematical linguistics.

Finite state automata are efficient: the problem of deciding whether a given string is a member of a language defined by a finite state automaton can be solved within a time linear in the size of the input string. Moreover, any finite state automaton can be converted into an equivalent 'deterministic' automaton. For deterministic automata, the time taken to solve the decision problem is independent of the size of the automaton (nondeterministic automata may require time exponential in the size of the automaton). Two finite state automata are equivalent if they define the same language.

Finite state automata can be combined using a variety of operations. If M_1 and M_2 are finite state automata defining the languages L_1 and L_2 respectively, then there is a very simple algorithm to construct a finite state automaton which defines $L_1 \cup L_2$. In a similar way, finite state automata can be constructed for intersection, complementation and difference. An example of intersection is given in Figure 1.

The 'concatenation' of two languages L_1 and L_2 is defined as $L_1 \cdot L_2 = \{xy | x \in L_1, y \in L_2\}$; the 'Kleene closure' of a language L is defined as $L^* = \{x_1 \ldots x_n | x_1 \in L, \ldots, x_n \in L\}$. There are simple algorithms to construct a finite state automaton for the concatenation of two given finite state automata, and for the Kleene closure of a given finite state automaton.

The availability of such operations means that a complicated language can be defined in terms of mechanical combinations of smaller, more manageable finite state automata. This is exploited in the concept of 'regular expressions': a declarative notation for regular languages. Regular expressions are used, for instance, in a variety of software tools to define string patterns. In typical implementations, such regular expressions are compiled into equivalent finite state automata.

A formal definition of finite state automata can be given as follows. A finite state automaton M is a tuple (Q, Σ, E, S, F) where Q is a finite set of states, Σ is a set of symbols, $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times Q$ is a finite set of transitions, $S \subseteq Q$ is a set of start states, and $F \subseteq Q$ is a set of final states. (Here, ϵ denotes the empty string.) The relation $\hat{E} \subseteq Q \times \Sigma^* \times Q$ is defined inductively as the minimal relation such that:

$\hat{E} \supseteq E \cup \{(q, \varepsilon, q) \mid q \in Q\}$
If (q_0, x_1, q_1) and (q_1, x_2, q_2) are both in $\hat{E}$
then $(q_0, x_1x_2, q_2) \in \hat{E}$.

The language $L(M)$ 'accepted' by M is $\{w | q_s \in S, q_f \in F, (q_s, w, q_f) \in \hat{E}\}$.

A finite state automaton is 'epsilon-free' if none of its transitions contains ϵ. A given finite state automaton can always be converted into an equivalent epsilon-free automaton. An epsilon-free finite state automaton is 'deterministic' if it has at most one start state and the outgoing transitions of all its states have different labels. A given nondeterministic automaton can always be converted into an equivalent deterministic automaton (potentially at the cost of increasing the size of the automaton considerably). An efficient minimization algorithm

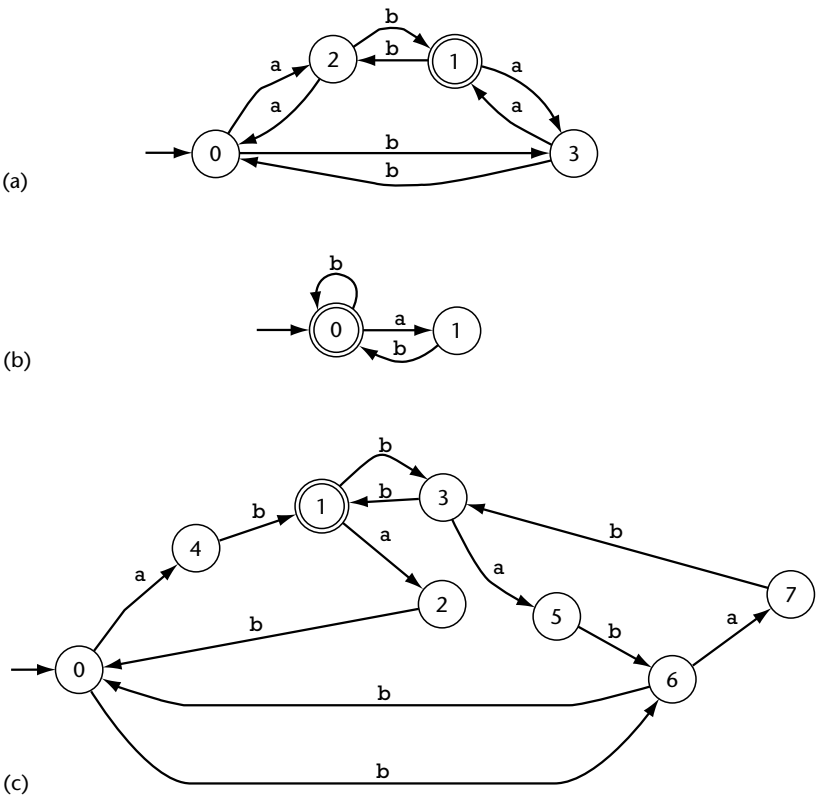


Figure 1. Examples of finite state automata. (a) In the first example, state 0 is a start state, and state 1 is a final state. The string `abbaab` is 'accepted' because there is a path $0 \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow 0 \rightarrow 2 \rightarrow 1$. Note that the path must start in a start state and end in a final state. The language L_1 defined by the automaton is the set of all strings over the alphabet {`a`, `b`} such that the number of occurences of `a` is odd and the number of occurrences of `b` is odd. (b) The second example defines the language L_2 consisting of all strings in which every `a` is immediately followed by a `b`. (c) The third automaton defines the language $L_1 \cap L_2$.

exists to convert a given deterministic automaton into an equivalent automaton with the fewest possible number of states (Hopcroft, 1971). This automaton is unique, ignoring the labeling of the states.

A language that can be defined by means of a finite state automaton is called a regular language. The set of regular languages is a proper subset of the set of context-free languages. There are many languages that are not regular. For example, the set of strings over the alphabet {`a`, `b`} where the number of occurrences of `a` equals the number of occurences of `b` is not a regular language. Nor is the language over the alphabet {`(`, `)`} such that the brackets are balanced. Nor is the 'copy' language $\{ww | w \in \{$ `a`, `b`$\}^* \}$.

Finite state automata are used in a variety of applications, such as the compilation of programming languages, pattern matching (not only for text, but also in other domains such as bioinformatics), and computational linguistics.

FINITE STATE TRANSDUCERS

Finite state transducers are a generalization of finite state automata, in which transitions are 'labeled' with pairs of symbols. A finite state transducer defines a mapping between two regular languages. Examples of finite state transducers are given in Figures 2 and 3.

A relation that can be defined by a finite state transducer is called 'regular' (Kaplan and Kay, 1994). Finite state transducers inherit many of the useful properties of finite state automata. The application of a finite state transducer is linear in the size of the input, although it can be exponential in the size of the transducer (it is not always possible to convert a given transducer into an equivalent deterministic transducer). Moreover, regular relations are closed under union, concatenation and Kleene closure. Same-length regular relations are also closed under intersection, complementation and difference.

The 'composition' of two binary relations R_1 and R_2 is $R_1 \circ R_2 = \{ (x_1, x_3) | (x_1, x_2) \in R_1, (x_2, x_3) \in R_2 \}$. Regular relations are closed under composition: if M_1 and M_2 are transducers for R_1 and R_2 respectively, then there is a simple algorithm to construct a transducer for the relation $R_1 \circ R_2$.

A formal definition of finite state transducers can be given as follows. A finite state transducer M is a tuple $(Q, \Sigma_d, \Sigma_r, E, S, F)$ where Q is a finite set of states, Σ_d and Σ_r are sets of symbols, E is a finite subset of $Q \times (\Sigma_d \cup \{\epsilon\}) \times (\Sigma_r \cup \{\epsilon\}) \times Q$, and S and F are sets of start and final states respectively. The relation $\hat{E} \subseteq Q \times \Sigma_d^* \times \Sigma_r^* \times Q$ is defined inductively as the minimal relation such that:

$\hat{E} \supseteq E \cup \{ (q, \epsilon, \epsilon, q) | q \in Q \}$

If (q_0, x_1, y_1, q_1) and (q_1, x_2, y_2, q_2) are both in $\hat{E}$ then $(q_0, x_1 x_2, y_1 y_2, q_2) \in \hat{E}$.

The relation $R(M)$ 'accepted' by M is $\{(w_d, w_r) | q_s \in S, q_f \in F, (q_s, w_d, w_r, q_f) \in \hat{E}\}$.

Finite state transducers can be extended with weights. Weighted transducers not only map a given string to one or more outputs, but also provide a score indicating, for instance, the likelihood of each mapping.

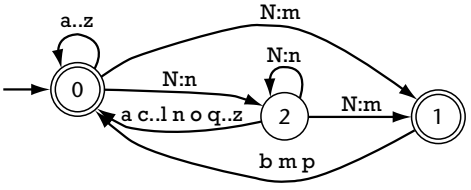


Figure 2. Example of a finite state transducer in phonology (adapted from Kaplan and Kay (1994)). N is an abstract nasal segment, which must be realized as either m (if it is followed by a labial) or n (otherwise). We use a colon to separate the labels of a label pair. We abbreviate a symbol pair s:s by writing s; and we represent intervals of symbols using ..

FINITE STATE PARSING AND HUMAN SENTENCE PROCESSING

Finite state techniques have been used widely in language processing, in particular in phonology and morphology. For instance, it has been shown that phonological models consisting of sets of context-sensitive rewrite rules are finite-state (Johnson, 1972; Kaplan and Kay, 1994). A similar result has been obtained for phonological models expressed in optimality theory (Frank and Satta, 1998).

The syntax of natural languages is generally assumed not to be finite-state (Chomsky, 1957). For instance, the recursive nature of syntactic analysis trees suggests that natural language is at least as complex as the language of balanced brackets, which is not regular. Furthermore, certain syntactic phenomena appear to require an unbounded amount of memory. An interesting example is the

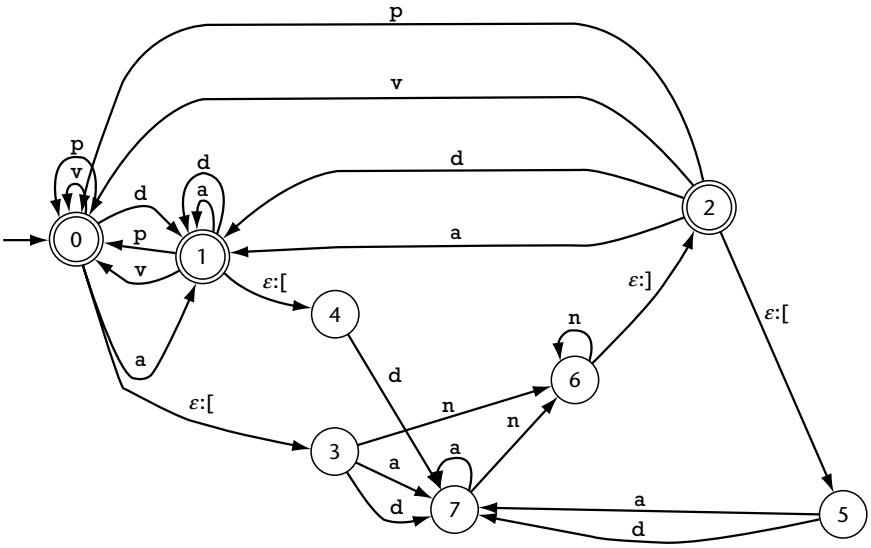


Figure 3. Example of a finite state transducer. The transducer constitutes a very simple NP-chunker. It assumes its input consists of strings over the part-of-speech labels a, d, n, p, and v, for adjective, determiner, noun, preposition, and verb respectively. The transducer places square brackets around sequences of tags in the input corresponding to noun phrases. For instance, the sentence *the quick brown fox jumped over the lazy dog* corresponds to the part-of-speech sequence daanvpdan. The transducer maps this string to [daan] vp [dan].

occurrence of crossing dependencies between verbs and their direct objects found in Swiss German (Huybregts, 1984).

However, the assumption that finite state automata are not suited to model the syntax of natural language raises the question of how the observed efficiency of language understanding by humans can be explained, because more complex language models cannot in general be processed so efficiently. Furthermore, human memory is finite; and only a fraction of it is used for language processing. Chomsky introduced the distinction between 'competence' – the speaker's and hearer's knowledge of language – and 'performance' – the actual use of this knowledge of language, including language processing. It may be that competence is modeled by more powerful means, while an efficient finite state processing model only approximates this competence model.

Interestingly, humans have problems with certain grammatical constructions, such as center-embedding, which are impossible to describe by finite state means; but which are supposed to be grammatical (Miller and Chomsky, 1963). For example, the following sentences are grammatical, but hard to understand:

I called the man who put the book that you told me about down up. (1)

The man whom the boy whom the students recognised pointed out is a friend of mine. (2)

The man the boy the students recognized pointed out is a friend of mine. (3)

The rat the cat the dog chased bit ate the cheese. (4)

Such observations suggest an underlying finite state approach, since finite state devices are incapable of describing center-embedded constructions (of arbitrary depth).

Some researchers have proposed parsing algorithms which interpret a given (nonregular) grammar, but which are restricted to operate in a small, fixed amount of memory. For instance, Resnik (1992) shows how a 'left-corner' parser, augmented with an operation of 'eager' composition, and formalized as a push-down automaton, utilizes its stack only for center-embedding constructions. Left-recursive and right-recursive structures, on the other hand, can be processed without exploiting the stack. If a limit to the depth of the stack is imposed, then the resulting model explains why center embedding cannot be processed beyond a certain complexity. In contrast, a purely bottom-up (or top-down) parser formalized in a similar fashion needs to use its stack for right-recursive (respectively, left-recursive) structures: if a limit to the stack were imposed then the model would wrongly predict that left-recursive and right-recursive structures are equally difficult. A finite state automaton can be explicitly constructed from a given grammar, on the basis of similar restrictions (Nederhof, 2000).

The distinction between competence and performance has been challenged (e.g. Abney, 1996). Rather than distinguishing between a powerful competence grammar and a finite state processing component, an alternative is to create finite state models directly. In some of these approaches (e.g. Roche and Schabes, 1997), syntactic analysis is divided into a number of levels. Each level consists of a transducer recognizing some amount of structure. For example, the first level might recognize base noun phrases, the next level prepositional phrases, the next level complex noun phrases, and so on. Such transducers are not unlike our simple example in Figure 3. They are applied in sequence; or equivalently, a single transducer may be constructed by composing them.

References

Abney S (1996) Statistical methods and linguistics. In: Klavans JL and Resnik P (eds) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pp. 1–26. Cambridge, MA: MIT Press.

Chomsky N (1957) *Syntactic Structures*. The Hague, Netherlands: Mouton.

Frank R and Satta G (1998) Optimality theory and the computational complexity of constraint violability. *Computational Linguistics* **24**: 307–315.

Hopcroft JE (1971) An *n* log *n* algorithm for minimizing the states in a finite automaton. In: Kohavi Z (ed.) *The Theory of Machines and Computations*, pp. 189–196. New York, NY: Academic Press.

Huybregts R (1984) The weak inadequacy of context-free phrase structure grammars. In: de Haan G, Trommelen M and Zonneveld W (eds) *Van Periferie naar Kern*, pp. 81–99. Dordrecht, Netherlands: Foris.

Johnson CD (1972) *Formal Aspects of Phonological Description*. The Hague, Netherlands: Mouton.

Kaplan RM and Kay M (1994) Regular models of phonological rule systems. *Computational Linguistics* **20**(3): 331–378.

Kleene SC (1956) Representation of events in nerve nets and finite automata. In: Shannon CE and McCarthy J (eds) *Automata Studies*, pp. 3–42. Princeton, NJ: Princeton University Press.

Miller G and Chomsky N (1963) Finitary models of language users. In: Luce R, Bush R and Galanter E

(eds) *Handbook of Mathematical Psychology*, vol. II, pp. 419–491. New York, NY: Wiley.
Nederhof M-J (2000) Practical experiments with regular approximation of context-free languages. *Computational Linguistics* **26**(1): 17–44.
Resnik P (1992) Left-corner parsing and psychological plausability. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, Nantes, vol. I, pp. 191–197.
Roche E and Schabes Y (eds) (1997) *Finite-State Language Processing*. Cambridge, MA: MIT Press.

Further Reading

Hopcroft JE, Motwani R and Ullman JD (2001) *Introduction to Automata Theory, Languages and Computation*, 2nd edn. Boston, MA: Addison-Wesley.
Johnson-Laird PN (1983) *Mental Models*. Boston, MA: Harvard University Press.
Kaplan RM and Kay M (1994) Regular models of phonological rule systems. *Computational Linguistics* **20**(3): 331–378.
Karttunen L (2000) Applications of finite-state transducers in natural language processing. In: Yu S and Paun A (eds) *Implementation and Application of Automata*, pp. 34–46. Berlin, Germany: Springer.
Karttunen L, Chanod J-P, Grefenstette G and Schiller A (1996) Regular expressions for language engineering. *Natural Language Engineering* **2**(4): 305–328.
Karttunen L and Oflazer K (eds) (2000) *Computational Linguistics* **26**(1). [Special issue on finite-state methods in NLP.]
Kornai A (ed.) (1999) *Extended Finite State Models of Language*. Cambridge, UK: Cambridge University Press.
Mohri M (1997) Finite-state transducers in language and speech processing. *Computational Linguistics* **23**(2): 269–312.
Perrin D (1990) Finite automata. In: van Leeuwen J (ed.) *Handbook of Theoretical Computer Science*, vol. B 'Formal Models and Semantics', pp. 1–57. Amsterdam, Netherlands: Elsevier/Cambridge, MA: MIT Press.
Pulman S (1986) Grammars, parsers and memory limitations. *Language and Cognitive Processes* **1**(3): 197–225.
Roche E and Schabes Y (eds) (1997) *Finite-State Language Processing*. Cambridge, MA: MIT Press.

Generalized Quantifiers

Advanced article

Fritz Hamm, University of Tübingen, Tübingen, Germany

CONTENTS

Generalized quantifiers in logic are those quantifiers which go beyond the traditional first order quantifiers ∃, ∀. Quantifiers in natural language semantics are the denotations of determiners; generalized quantifiers are the denotations of noun phrases (NPs).

GENERALIZED QUANTIFIERS IN LOGIC AND NATURAL LANGUAGE

Both the modern logical concept of quantifier and the natural-language-semantic notion of quantifier have their historical roots in philosophical and mathematical logic.

The founding father of logic, Aristotle, introduced in his theory of syllogisms the notion of quantifier as a relation between universal terms (variables) or properties. The familiar quantifiers in the square of opposition consisting of *all* or *every, some, no*, and *not every* are binary relations Q between properties A and B; formally we write QAB. For instance, the sentence *all men are mortal* is true if the property of being a man is part of the property of being mortal.

If we presume that the extension of a one-place property, like the property of being a man, is a set, we can rephrase Aristotle's insight in modern set-theoretic terminology. A quantifier in the above sense is then just a relation between two sets. The quantifier *all*, for example, is the subset relation, and the sentence *all men are mortal* is therefore considered true if and only if the set of men is a subset of the set of individuals that are mortal.

The notion of quantifier as a variable binding operation, familiar from modern first-order logic, was first introduced much later by G. Frege (1960) and reached its full generality in the work of A. Mostowski (1957) and P. Lindström (1966). The Lindström quantifier covers nearly every instance of the notion of quantification both in logic and in natural-language semantics. (An exception will be mentioned below.)

Quantifiers in logic and most natural-language quantifiers share an important characteristic: they are both permutation-invariant (PERM), that is, only the number of elements of their argument sets determine their behavior. For example, given a universe of discourse M, a permutation π of M (i.e. a bijective map of M onto itself), and subsets A and B of M, assume $every_M AB$. This means that $A \subseteq B$. Permutation-invariance says that $\pi(A)$, the image of A under π, is a subset of $\pi(B)$. In other words, $every_M \pi(A)\pi(B)$.

There are only a few natural-language quantifiers that are potential exceptions to PERM. This contrasts sharply with the semantics of other major syntactic categories. For example, Westerståhl (1985) shows that the requirement of permutation-invariance for extensional adjectives or adverbs rules out practically all of them, leaving only the trivial ones *existent* and *non-existent*.

This suggests that natural-language quantifiers, which are usually denoted by determiners, form the logical backbone of natural-language semantics, since permutation-invariance is a characteristic logical property.

What, then, is the difference between quantifiers in logic and quantifiers in natural languages? A first, superficial, difference is in the kinds of quantifiers that are of interest to logicians and semanticists.

If L is standard first-order logic with quantifiers $\forall$ and $\exists$, then let $L(Q)$ be first-order logic augmented with an additional quantifier symbol Q. Let $Q_\alpha = \{X \subseteq M: |X| \geq \aleph_\alpha\}$ ($|X|$ stands for the cardinality of the set X).

The formula $Q_0\psi$ therefore says that there is an infinite number of elements satisfying formula ψ, and $Q_1\psi$ says that there is an uncountably infinite number of elements satisfying ψ. It is easy to show that the logic $L(Q_0)$ admits of no complete

axiomatization, but J. Keisler has shown that the logic $L(Q_1)$ does admit of a complete axiomatization. Proofs of these results can be found in Kaufmann (1985).

The quantifiers Q_α are not very interesting from a linguistic point of view. A more interesting linguistic question was posed by Keenan (1992). Keenan argues that the semantic analysis of the sentence *every student read a different book* involves a complex quantifier *every … a different*, which relates the sets of *students* and *books* and the binary relation *read* to truth values. Keenan shows that this complex quantifier is not reducible to any combination of simpler quantifiers such as *every* or *a*.

Keenan's investigations therefore raise the general question of which types of quantifiers are realized in natural languages and how their interpretation is arrived at in a compositional way. Such investigations led to new research in linguistic typology (e.g. Bach *et al*., 1995).

Another difference between quantifiers in logic and in natural languages is that the latter have a restriction set built into their syntactic realization. Thus *every* applies to, say, the denotation of the count noun *book* to produce the noun phrase (NP)-denotation *every*(*book*). In the pioneering work of Barwise and Cooper (1981) these NP-denotations were called 'generalized quantifiers'.

Even more importantly, natural-language quantifiers are contextually restricted. The sentence *every student reads Ulysses* does not mean that every student in the world reads *Ulysses*. For the sentence to be considered true, it may suffice that every student in my class reads *Ulysses*.

A still more fundamental difference between quantifiers in logic and in natural language concerns the role of semantic universals.

GENERALIZED QUANTIFIERS AND SEMANTIC UNIVERSALS

The most important semantic universal concerning the denotation of natural language determiners is 'conservativity' (CONS).

According to Montague's extensional type theory, determiners denote in $D_{<<e,t>,\ <<e,t>,\ t>>}$. Now, given a universe of discourse M with two elements, a simple calculation shows that there are 65 536 quantifiers on M. This seems to be an absurdly large number of quantifiers in a universe with two elements. Therefore, the majority of quantifiers in $D_{<<e,t>,\ <<e,t>,\ t>>}$ cannot be denoted by natural-language determiners. (*See* **Semantics and Pragmatics: Formal Approaches**)

Semantic universals ought to determine those quantifiers that are in principle denotable by natural-language determiners. This does not mean that there is no language-specific variation: some languages may have a richer system of determiners than others.

Conservativity says that the left argument of a quantifier is more prominent than the right one: formally, $QAB \leftrightarrow QAA \cap B$. Thus the sentence *a student sleeps* is true just in case the sentence *a student is a student and sleeps* is true. Neglecting intensional determiner structures, conservativity appears to be an empirically trivial requirement, in the sense that there seem to be no counterexamples. A potential counterexample that comes to mind is *only*, defined as follows: *only AB* iff $B \subseteq A$. This determiner denotation is not conservative, but it is highly questionable whether *only* is a determiner at all.

Although CONS is empirically trivial in the above sense, it has major systematic effects. Given our universe M with two elements, CONS reduces the number of quantifiers from 65 536 to 512. Of course, this result means that there are a lot of quantifiers that are not conservative. Examples from mathematical logic abound here. For instance consider the quantifier H (for Härtig), which compares the cardinalities of two sets: HAB iff $|A| = |B|$. It is clear that H is not conservative. CONS therefore not only serves as a principle that rules out the majority of quantifiers in $D_{<<e,t>,\ <<e,t>,\ t>>}$, but also helps to distinguish logical quantifiers from many quantifiers in natural language.

A further important semantic universal for determiners is 'extension' (EXT). This principle formalizes the intuition that determiners are constants, in contrast to common nouns like *farmer*. For example, we do not want to suppose that *most* can mean *every* on a universe M, *two* on a universe M' and *most* on a third universe M''. EXT excludes such situations. The principle says that for $A, B \subseteq M \subseteq M'$, $Q_M AB$ iff $Q_{M'} AB$.

Given the semantic universals CONS, EXT and PERM, many semantic facts concerning determiners can be explained.

For example, we do not find quantifiers Q satisfying the relational scheme 'asymmetry', i.e. the principle $QAB \rightarrow \neg QBA$. By contrast, there are many quantifiers Q satisfying symmetry, i.e. $QAB \rightarrow QBA$: for instance, the quantifier *some*.

Assuming CONS, EXT and PERM, it is possible to prove that no nontrivial quantifiers exist that satisfy asymmetry. Such results help to explain certain lexical gaps.

Another important class of properties of quantifiers, which are not universal but specific, are various forms of 'monotonicity'. For example, the so called Ladusaw–Fauconnier generalization (Keenan and Westerståhl, 1997) attributes the distribution of negative-polarity items such as *any* or *ever* to monotonicity properties of quantifiers. A negative-polarity item is licensed in the scope of a monotone-decreasing operator like *at most four*, but not in the scope of *at least four*, which is not monotone-decreasing. This explains why *at most four students have ever been to Munich* is acceptable but *at least four students have ever been to Munich* is not. A quantifier Q is monotone-decreasing iff whenever QAB and $B' \subseteq B$ then QAB'. For example, *no*, defined as $noAB$ iff $A \cap B = \emptyset$, is monotone-decreasing.

The most important semantic universal, CONS, has an interesting characterization, which may even be relevant to learnability issues. Keenan and Stavi (1986) show that on a given universe M the conservative quantifiers are exactly those that can be generated from a relatively small set of monotone-increasing quantifiers by applying the Boolean operators of conjunction, disjunction and negation. Therefore a person learning the possible denotations of natural-language determiners has to learn the denotations of a relatively small class of determiners, and can derive the rest by applying operations, which have to be learned anyway.

GENERALIZED QUANTIFIERS AND CONCEPTUAL SEMANTICS

In order to study in more detail the relationship between quantifier theory and conceptual semantics, it is useful to concentrate for a while on the grammatical category of (spatial) prepositions, which is a central topic in the field of conceptual semantics. (*See* **Categorial Grammar and Formal Semantics**)

Landau and Jackendoff (1993) observe a significant asymmetry between the vocabulary of object-denoting nouns and that of prepositions. An average adult native speaker of English uses about 10 000 count nouns denoting objects, but at most 100 prepositions that locate these objects in space. This seems to be a fact not only about English but about natural languages in general.

With prepositions, as with quantifiers, one observes characteristic lexical gaps. For example there seems to be no preposition *sprough* meaning 'reaching from end to end of a cigar-shaped object', which would distinguish a sentence like *the rug extended sprough the airplane* from the unacceptable **the rug extended sprough my dining room*.

The explanation Landau and Jackendoff offer for this phenomenon is that although the semantics of common nouns rests on abstractions from the objects denoted by them, the degree of abstraction is far greater in the case of prepositions. The preposition *on* in *the cat is sitting on the mat* relates the 'figure' object *the cat* to the 'reference' object *the mat*, which defines the region in which the figure object is located. Only very abstract representations of the figure and ground objects serve as arguments of the relations that propositions express. Landau and Jackendoff entertain the hypothesis that two conceptual systems, the 'what' and the 'where' systems, which determine geometries of different granularity, are responsible for the difference between object-denoting and object-locating expressions. Although the details of their proposal have been criticized (e.g. Herskovits, 1997), their basic insight seems to be correct.

If we consider the type of prepositions $D_{<<<e,t>,\, t>,\, <<e,t>,\, t>>}$ in (extensional) Montagovian type theory, we immediately get the prediction that there should be many more prepositions than common nouns, which are of type $D_{<e,t>}$. This is the opposite of what is observed. Therefore, as in quantifier theory, one would like to find restrictions on the set of possible natural-language prepositions. This is partly achieved in Zwarts and Winter (2000), where a version of conservativity adapted for prepositions plays a major role. However, this work does not account for the abstraction process inherent in the semantics of prepositions.

A device which helps to formalize abstraction has been introduced by van der Does and van Lambalgen (2000).

Conditional quantification $\exists(\varphi|G)$ is a new form of quantification. Given a set of information states G and a formula φ this form of quantification expresses the maximum information about φ on the basis of G. Therefore the meaning of $\exists(\varphi|G)$ depends on G. For a specific choice of G, $\exists(\varphi|G)$ is just the usual existentially quantified $\exists x\varphi$, but for other choices of G, $\exists(\varphi|G)$ is not even first-order-definable. Depending again on G, $\exists(\varphi|G)$ does not always bind variables in φ. This means that in general $\exists(\varphi|G)$ is not a Lindström quantifier.

Van der Does and van Lambalgen give a rough proposal for a semantic analysis of prepositions. The sentence *the book is on the table* is formalized as $\exists(B(x) \wedge \mathrm{on}(x, y) \wedge \mathrm{T}(y)|G)$, where G erases all properties of the table y except its surface and all structure of the book x except its internal axis.

(The precise description of G is a rather complex matter.)

The approach taken by van der Does and van Lambalgen therefore promises significant insights from conceptual semantics formulated in terms of model-theoretic quantification.

References

Bach E, Jelinek E, Kratzer A and Partee B (eds) (1995) *Quantification in Natural Languages*, vols. I–II. Dordrecht: Kluwer.

Barwise J and Cooper R (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy* **4**: 159–219.

van der Does J and van Lambalgen M (2000) A logic of vision. *Linguistics and Philosophy* **23**: 1–92.

Frege G (1960) On function and concept. In: Geach P and Black M (eds) *Translations from the Philosophical Writings of Gottlob Frege*, pp. 21–41. Oxford, UK: Blackwell [First published 1892.]

Herskovits A (1997) Language, spatial cognition and vision. In: Stock O (ed.) *Spatial and Temporal Reasoning*, pp. 155–202. Dordrecht: Reidel.

Kaufmann M (1985) The quantifier 'there exist uncountably many' and some of its relatives. In: Barwise J and Feferman S (eds) *Model-Theoretic Logics*, pp. 123–176. New York, NY: Springer.

Keenan E (1992) Beyond the Frege boundary. *Linguistics and Philosophy* **15**: 199–222.

Keenan E and Stavi J (1986) Semantic characterization of natural language determiners. *Linguistics and Philosophy* **9**: 253–326.

Keenan E and Westerståhl D (1997) Generalized quantifiers in linguistics and logic. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 837–893. Amsterdam: Elsevier.

Landau B and Jackendoff R (1993) 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences* **16**: 217–265.

Lindström P (1966) First-order predicate logic with generalized quantifiers. *Theoria* **32**: 186–195.

Mostowski A (1957) On a generalization of quantifiers. *Fundamenta Mathematicae* **44**: 12–36.

Westerståhl D (1985) Logical constants in quantifier languages. *Linguistics and Philosophy* **8**: 387–429.

Zwarts J and Winter Y (2000) Vector space semantics: a model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information* **9**: 483–511.

Further Reading

Bach E, Jelinek E, Kratzer A and Partee B (eds) (1995) *Quantification in Natural Languages*, vols. I–II. Dordrecht: Kluwer.

Barwise J and Cooper R (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy* **4**: 159–219.

Barwise J and Feferman S (eds) (1985) *Model-Theoretic Logics*. New York, NY: Springer.

van Benthem J (1986) *Essays in Logical Semantics*. Dordrecht: Reidel.

van der Does J and van Lambalgen M (2000) A logic of vision. *Linguistics and Philosophy* **23**: 1–92.

van Eijck J (1991) Quantification. In: Wunderlich D and von Stechow A (eds) *Semantics: An International Handbook of Contemporary Research*, pp. 459–487. Berlin: De Gruyter.

Frege G (1960) On function and concept. In: Geach P and Black M (eds) *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Blackwell. [First published 1892.]

Keenan E and Stavi J (1986) Semantic characterization of natural language determiners. *Linguistics and Philosophy* **9**: 253–326.

Keenan E and Westerståhl D (1997) Generalized quantifiers in linguistics and logic. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 837–893. Amsterdam: Elsevier.

Landau B and Jackendoff R (1993) 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences* **16**: 217–265.

Partee B, ter Meulen A and Wall R (1990) *Mathematical Methods in Linguistics*. Dordrecht: Kluwer.

Westerståhl D (1989) Quantifiers in formal and natural languages. In: Gabbay D and Guenthner F (eds) *Handbook of Philosophical Logic*, vol. IV, pp. 1–131. Dordrecht: Reidel.

Government–Binding Theory

Introductory article

Howard Lasnik, University of Maryland, College Park, Maryland, USA

CONTENTS

Government–Binding theory is an approach to the study of the syntax of human languages based on an abstract underlying representation and transformations successively altering that structure. The approach posits universal principles innately represented in the mind, and simple parameters, fixed by the language learner from simple evidence, determining how languages can differ.

ORIGINS OF GOVERNMENT AND BINDING

Government–Binding (GB) Theory is an approach to the study of human language that developed out of Noam Chomsky's work in the 1970's. Like all of Chomsky's earlier work, it centered on two fundamental questions:

> What is the correct characterization of someone who speaks a language? What kind of capacity is 'knowledge of language'? (1)

> How does this capacity arise in the individual? What aspects of it are acquired by exposure to relevant information ('learned'), and what aspects are present in advance of any experience ('wired in')? (2)

Chomsky's earliest work, in the 1950's, particularly concentrated on question (1), since explicit and comprehensive answers to that question had never been provided before. Chomsky's answer posited a computational system in the human mind that provided statements of the basic phrase structure patterns of languages (phrase structure rules) and operations for manipulating these basic phrase structures (transformations). This framework, and all its descendants, fall under the general title Transformational Generative Grammar ('generative' meaning explicit, in the sense of mathematics).

In the 1960's, the focus began to shift more towards question (2). Chomsky coined the term 'explanatory adequacy' for answers to that question. A theory of language, regarded as one component of a theory of the human mind, must make available grammars for all possible human languages. To attain a high degree of explanatory adequacy, the theory must in addition show how the learner selects the correct grammar from among all the available ones, based on quite limited data. The theories of the 1950's and early 1960's made a very large number of grammars available, so the explanatory problem was severe.

Through the late 1960's and 1970's, to address this problem of explanation, more and more constraints were proposed on the notion 'possible human grammar'. For example, Chomsky's 'standard theory' of the mid- to late-1960's proposed a limitation on the kinds of transformations previously assumed to exist. Chomsky's earliest syntactic theory postulated phrase structure rules that create simple structures, 'generalized transformations' that combine separate simple structures into more complex ones, and 'singulary transformations' that alter the structures created by phrase structure rules and generalized transformations. Chomsky argued that the system should be constrained by the elimination of generalized transformations, with their work taken over by phrase structure rules themselves. He argued that since the resulting expansion of descriptive power of phrase structure rules is so slight, the resulting system is both simpler and more explanatory, in that it makes fewer options overall available to the learner. The next major move in the direction of explanatory adequacy came in the late 1960's in the form of the 'X-bar theory' of phrase structure, which proposed limitations on phrase structure rules, and further limitations on transformations, so that they no longer were responsible for derivational morphology of the *destroy-destruction* type. Such idiosyncratic relations are better captured in

the lexicon, which is, after all, the repository of all that is idiosyncratic about particular lexical items.

A human language is a systematic way of relating sound to meaning, with syntax mediating between the two. At the point in the development of the theory just summarized (around 1970), the model can be graphically represented as follows, with deep structure, the initial phrase structure representation created in conformity with the requirements of X-bar theory, connected to meaning, and surface structure, the final result of the whole syntactic derivation, connected to sound:

D(eep)-Structure – Semantic interpretation
|
Transformations
|
S(urface)-Structure – Phonetic interpretation (3)

While this was the basic architecture, it was known from the earliest work in generative grammar that some aspects of meaning depend on surface structure. In particular, while grammatical relations (subject of, object of, etc.) are most directly related to D-structure (and are opaque in S-structure), virtually all other aspects of meaning (including scope of quantifiers, anaphora, focus) relate to S-structure. For example, in his earliest work Chomsky already had pointed out that transformations often alter scope possibilities, while leaving understood grammatical relations intact, as in (4) versus (5).

Everyone in the room knows three languages (4)

Three languages are known by everyone in the room (5)

This led to a revised model (the 'extended standard theory') in which both D-structure and S-structure are inputs to semantic interpretation.

Within this model, through the 1970's, more and more restrictions were proposed on the phrase structure and transformational options assumed to be available to the child learning a language. These moves were explicitly motivated by considerations of explanatory adequacy, though general considerations of simplicity also played a role. One small simplification in the extended standard theory model was the result of a technical revision concerning how movement transformations operate. Trace theory proposed that when an item moves, it leaves behind a 'trace', a silent placeholder marking the position from which movement took place. Under trace theory, the importance of D-structure for semantic interpretation is further reduced, and ultimately eliminated. Once S-structure is enriched with traces, even grammatical relations can be determined at that level of representation. Using the term LF ('Logical Form') for the syntactic representation that relates most directly to semantics and PF ('Phonetic Form') for the one relating most directly to phonetics, we have the so-called T-model in (6), which was at the core of GB theorizing.

D-structure
|
Transformations
|
S-structure
/ \
PF LF (6)

Modularity

On first examination, human languages appear to be almost overwhelmingly complex systems, and the problems, for the linguist, of successfully analyzing them, and for the learner, of correctly acquiring them, seem virtually intractable. But if the system is broken down into smaller parts, the problem might likewise be decomposed into manageable components. In fact, with this divide and conquer ('modular') approach, by the GB era of the 1980's, languages began to seem much simpler. Apparently complex phenomena were seen as the result of interactions of simple modules. The phrase structure module was virtually reduced to the X-bar schema, with specific instantiations following from properties of particular lexical items. For example, the verb *prove* must be specified in the lexicon as taking a direct object, a kind of requirement called 'subcategorization' in the standard theory, since such requirements divide big categories (like verb) into smaller subcategories (like transitive verb). Given the subcategorization properties of *prove*, a specific phrase structure rule saying that a Verb Phrase (VP) can consist of a V and a Noun Phrase (NP) would be completely redundant. Further, the X-bar schema itself was extended from just 'lexical' categories (noun, verb, adjective, etc.) to 'functional' categories. It was an irony of the original formulation of X-bar theory that it excluded the most fundamental unit of syntactic analysis – the sentence. All other phrasal units were analyzed as projections of a head, but sentence was *sui generis*. GB theorizing brought sentence into the fold, by analyzing it as the projection of an inflectional head, Infl, containing tense and agreement information. In certain

instantiations of the general framework, Infl was divided into two separate functional heads, Tense (T) and Agreement (Agr).

The transformational module is also dramatically simplified in comparison with its predecessors. In the 1950's and 1960's, the transformational component of the grammar of a particular language was thought to be a long ordered list of very detailed transformations, some marked optional and others marked obligatory, specific to the language in question. In such a framework, explanatory adequacy is a very distant goal. The GB framework replaced these transformations with very general optional operations, Move α (displace any item anywhere), or even Affect α (do anything to anything). There is thus very little transformational syntax that the child has to learn. A grammar this simple and general would seem to massively overgenerate, producing countless numbers of unacceptable sentences. To deal with this overgeneration problem, GB theorists, further developing a line of research begun in the 1960's, posited general constraints on the operation of transformations (locality constraints in particular), and also conditions on the output of the transformational component (including 'filters').

PARAMETERS

The postulated universal ('wired-in') parts of the computational system are called principles. The (limited) ways in which languages can differ syntactically are called parameters. The system is fundamentally based on principles and parameters. In fact, Chomsky came to prefer the term 'principles and parameters' to 'government and binding' for the approach being outlined here, as 'government' and 'binding' are just two technical notions among many in the theory, as will be seen below. The child learning a language is pre-equipped with the principles, and needs only to set the 'values' of the parameters. The standard assumption is that there are few parameters, they are very simple, and their values can be determined by the child based on readily available primary linguistic data. One of the major parameters, the head parameter, is responsible for a significant word order difference among languages. In languages like English, X-bar heads invariably precede their complements in D-structure. For example, verbs precede their direct objects, and English has prepositions rather than postpositions. English is head-initial. Languages like Japanese are just the reverse. They are head-final. Another parameter, the null subject parameter, concerns the need for overt subjects. In languages like Spanish, a subject of a simple indicative clause need not be expressed, while in languages like English, subjects must be expressed. For all of the hypothesized parameters, simple data is expected to be readily available to the child.

θ-THEORY AND THE LEXICON

As indicated above, GB posits a highly modular organization of human linguistic ability. The X-bar schema for phrase structure is one module. The lexicon is another. These modules determine D-structure configurations via the regulation of a third module 'θ-theory'. Subcategorization properties follow, in large measure, from semantic properties. Thus, in a sentence with the verb *solve*, there is a semantic function for a direct object to fulfill, while there is no such function in the case of *sleep*. These semantic functions that arguments (direct objects, subjects, indirect objects, etc.) fulfill are called 'thematic (θ-)roles'. The verb *prove* demands a direct object since the object would fulfill a necessary θ-role determined by the meaning of the verb. Conversely, an intransitive verb like *sleep* does not take a direct object since there would be no θ-role for it to fulfill. These paired requirements on assigners and recipients of theta roles are called the 'θ-Criterion': Every θ-role must be assigned to one and only one argument, and every argument must receive one and only one θ-role.

CASE THEORY

S-structures result from the transformational component operating on D-structures. Given the generality of Move α, derivations often seem to yield ungrammatical sentences. One module reining in this overgeneration by regulating S-structure is 'Case theory'. Consider the following pair of sentences:

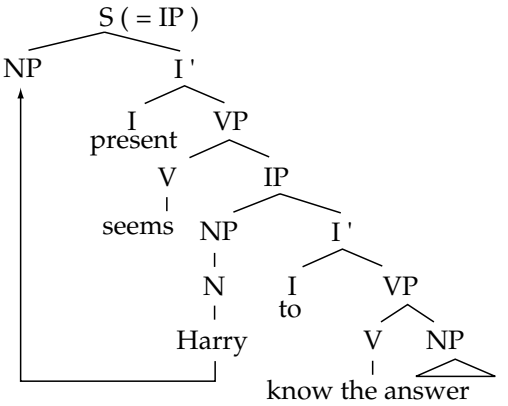
It seems Harry knows the answer (7)

Harry seems to know the answer (8)

Harry in (8) is in subject position at S-structure, but given the close semantic parallelism between (7) and (8), especially with respect to θ-relations, their D-structures (and therefore the resulting trace-augmented S-structures) are assumed to be similar. In particular, since *Harry* is the understood subject of *know the answer* in both examples, it must be in the subject position of that predicate in the D-structures of both examples. The D-structure of (7) is identical to the S-structure in relevant respects, with the subject of the main sentence *it* not

an argument, but rather, an 'expletive', a placeholder lacking semantic import (unlike the referential *it* in 'It is on the table.'). The D-structure of (8) is similar, the only differences being that the main subject position is completely empty, and the embedded sentence is an infinitive:

*seems [Harry to know the answer] (9)



(10)

To derive the correct S-structure, the NP *Harry* raises into the empty higher subject position, as indicated by the arrow in (10). The movement is obligatory in this instance. If *Harry* remains in the lower subject position, the result is ungrammatical, whether or not the expletive *it* occurs as the higher subject:

*Seems Harry to know the answer (11)

*It seems Harry to know the answer (12)

This obligatoriness of movement of the subject has been analyzed as a requirement of case. There are characteristic structural positions that 'license' particular cases, as in the following table:

Position	**Case**	**Example**
Subject of finite sentence	Nominative	*He* left
Direct object of transitive verb	Accusative	I saw *him*
'Subject' of NP	Genitive	*John's* belief that Mary solved the problem
Object of preposition	Oblique	near *him*

(13)

In many languages (such as Latin, Russian, German), these case distinctions are overtly manifested. In English, only pronouns show an overt distinction between nominative and accusative, but Case Theory posits that all NPs have abstract case (henceforth, Case), even when it is not phonologically visible. The requirement that all NPs occur in appropriate Case positions is the 'Case Filter', a well-formedness condition on the S-structure level of representation. The NP *Harry* in its D-structure position in (10) lacks a legitimate Case, since that position does not license nominative, accusative, genitive, or oblique. Movement of *Harry* to the higher subject position salvages the example, since that position licenses nominative Case. Such 'subject raising' is not possible for structure (14), where the embedded sentence is finite:

*Harry seems [*t* knows the answer] (14)

This, too, has been analyzed in terms of Case: movement of an NP from one Case position to another is prohibited. Metaphorically speaking, movement to a Case position is a 'last resort', an option taken only if necessitated by the Case Filter.

Government

The GB approach always sought regularities and generalizations. The notion 'government' is itself a generalization of the X-bar theoretic head-complement relation. The basic definition is as follows:

> A head H governs Y if and only if every maximal projection dominating H also dominates Y and conversely. (15)

By (15), a head governs its complement and also its specifier. Case licensing then is under government, with the governor licensing the governee. A transitive verb govern its direct object NP; a preposition governs its complement NP; Infl governs its specifier (the subject of the clause); and N (or D, for determiner, in a development of the theory, the 'DP hypothesis') governs its specifier, the 'subject' of the nominal expression. Thus, a Case-licensing head (transitive verb; preposition; finite Infl; possessive determiner) licenses Case on a nominal expression (DP) that it governs.

TYPES OF MOMENT

The transformational module of the theory recognizes three major subtypes of movement. 'A-movement' is movement to an argument-type position (especially subject position), as exemplified in the subject raising seen in (8), and in passive sentences, where an understood object surfaces in subject position ('Mary was elected'). In both of

these instances, the movement is to specifier of IP. 'A-movement' is movement of an XP to a non-A position. The movement of an interrogative expression as in (16) (WH-movement) is a central exemplar:

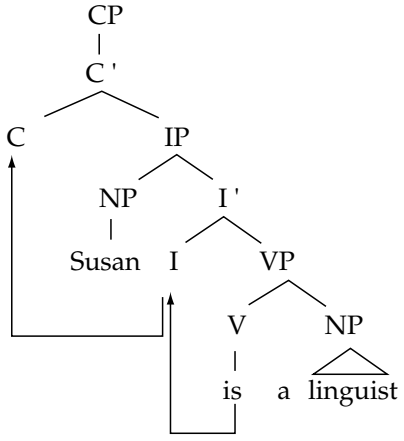
Who will they hire *t* (16)

WH-movement is standardly analyzed as movement to the specifier of CP (Complementizer Phrase), a functional projection above IP. All three types of movement are regarded as instantiations of one very general operation: Move α. The differences follow from independent properties of the items moved and the positions moved to.

Head movement and Inflection

The third major type of movement is head movement, where an X° adjoins to a higher head (the very next higher head by the Head Movement Constraint). One of the classic analyses of generative grammar was restated in the GB framework in terms of head movement. Pairs of sentences like those in (17) are related via movement of the verb *be/is* to Infl followed by movement of Infl To C, schematized in (18).

a. Susan is a linguist
b. Is Susan a linguist (17)



(18)

Similar head movement, along with WH-movement, is involved in the derivation of (16).

A major GB research topic of the late 1980's and early to mid-1990's was the parametric differences among languages with respect to the kind of head movement just illustrated. In many languages, all types of verbs can raise, but in English, only 'auxiliary' verbs (*be*, *have*, modals) can raise; 'main' verbs cannot:

a. Susan saw a linguist
b. *Saw Susan a linguist (19)

This difference shows up in negative sentences as well, as in the following contrast, under the standard assumption that the base position of negation is higher than VP and lower than Infl:

*John likes not Mary (20)

Jean (n')aime pas Marie (21)

A parametric property of Infl (or one of its components T and Agr) was argued to be responsible for the observed divergence. In languages, such as French, that allow all verbs to raise, Infl is 'strong', while in English it is 'weak'.

BINDING

The 'Binding' part of Government–Binding theory has as its core anaphoric relations, circumstances under which one expression can or cannot take another as its antecedent, that is, pick up its reference from the other. Among the imaginable anaphoric relations (relations of referential dependence) among NPs, some are possible, some are necessary, and still others are proscribed, depending on the nature of the NPs involved and the syntactic configurations in which they occur. For example, in (22), *him* can take *John* as its antecedent, while in (23), it cannot.

John said Mary criticized him (22)

John criticized him (23)

That is, (23) has no reading corresponding to that of (24), with the pronoun *him* replaced by the 'anaphor' *himself*.

John criticized himself (24)

A pronoun cannot have an antecedent that is 'too close' to it. This is Condition B of the binding theory. Conversely, an anaphor requires an antecedent quite close to it (Condition A). Compare (24) with (25).

*John said Mary criticized himself (25)

The pertinent locality is, roughly, being in the same clause (though in certain instances a more complicated notion involving government is implicated, hence Chomsky's name 'governing category' for the relevant domain).

A third binding condition (Condition C) excludes an anaphoric connection between *She* and *Mary* in (26), as contrasted with (27).

*She thinks Mary will solve the problem
[with *She* intended to refer to Mary] (26)

Mary thinks she will solve the problem (27)

THE ROLE OF LOGICAL FORM

In the core GB model schematized above in (6), LF is not distinct from S-structure. However, more and more arguments were put forward that transformational operations of the sort successively modifying D-structure, ultimately creating S-structure, also apply to S-structure, creating a distinct LF. One such operation, Quantifier Raising (QR), moves quantifiers from their surface positions to positions more transparently representing their scope, with the traces of the moved quantifiers ultimately interpreted as variables bound by those quantifiers. The ambiguities of sentences with multiple quantifiers are captured by the multiple LF representations QR makes available. For example, a sentence like (28) has the two LF representations in (29), depending on the order in which the two quantifiers are raised. Subscripts mark the association of quantifier with trace.

Some student solved every problem (28)

a. some student$_1$ [every problem$_2$ [t_1 solved t_2]]
b. every problem$_2$ [some student$_1$ [t_1 solvedt_2]]
(29)

(29a) represents the reading of (28) in which *some student* has wider scope than *every problem* (there is a particular student that solved every problem), and (29b) represents the reading in which *every problem* has wider scope (every problem was solved, but not necessarily by the same student). Unlike the transformational operations mentioned earlier, applications of QR exhibit no phonological displacement. All the expressions in (28) are pronounced in their surface structure position, on either reading. This follows from the organization of the grammar. When a transformation operates between D-structure and S-structure, it will have an effect on the phonetic output, since S-structure feeds into PF. On the other hand, a transformational application between S-structure and LF will have no phonetic effect, since LF does not feed into PF. QR in (29) is an example of the latter type of 'covert' transformational operation.

Another covert operation is the analogue of overt WH-movement. Under the assumption that overt WH-movement positions an interrogative operator in its natural position for interpretation (with the trace it leaves behind in the natural position for a variable bound by the operator), in sentences with multiple interrogatives, such as (30), at the level of LF all are in sentence initial operator position, as illustrated in (31).

Where should we put what (30)

what$_1$ [where$_2$ [we should put t_1 t_2]] (31)

(31) is then rather transparently interpreted as:

For which object x and which place y, we should put x at y (32)

One of the most powerful arguments for covert WH-movement involves constraints on movement. For example, it is difficult to move an interrogative expression out of an embedded question (a question inside another sentence):

*Why$_1$ do you wonder [what$_2$ [John bought t_2 t_1]] (33)

If (33) were acceptable, it would mean 'What is the reason such that you wonder what John bought for that reason'. In languages where WH-phrases are *in situ* (unmoved) at S-structure, such as Chinese, their interpretation apparently obeys the same constraints that the overt movement in a language like English obeys. So, in Chinese an example like (34) is possible but one like (35) is impossible on the relevant reading (the one where *weisheme* is understood as modifying the action of the lower clause):

ni renwei [ta weisheme bu lai]
you think he why not come
'Why do you think he didn't come?' (34)

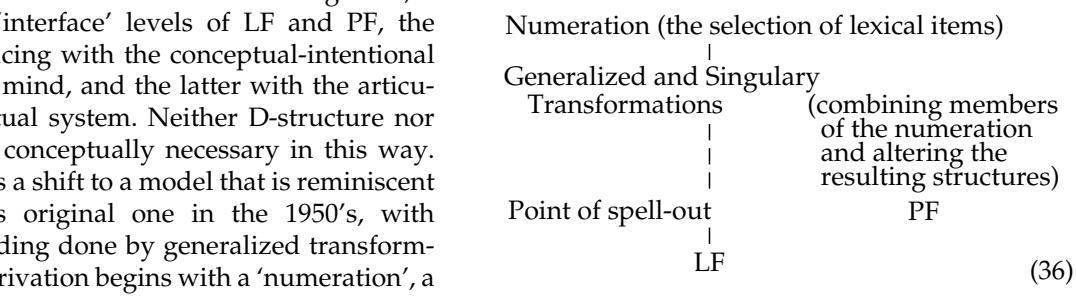
(*) ni xiang-zhidao [Lisi weisheme mai-le sheme]
you wonder Lisi why bought what
*'What is the reason such that you wonder what Lisi bought, where the purchase was for that reason?'
*'*Whatiscausingyoutowonderwhat* John bought?' (35)

This argues that even though the 'why' is not phonetically displaced, it really is moving: that is why it is obeying movement constraints. But this movement is 'covert', occurring in the mapping from S-structure to LF, hence not contributing to pronunciation.

ECONOMY AND MINIMALISM

The diminishing role of D- and S-structure in the theory suggests that neither is actually a significant level of representation. The 'Minimalist Program' for linguistics carries still further the simplifications in the theory that marked the GB framework.

If a language is to relate sound to meaning at all, it requires the 'interface' levels of LF and PF, the former interfacing with the conceptual-intentional system of the mind, and the latter with the articulatory-perceptual system. Neither D-structure nor S-structure is conceptually necessary in this way. This motivates a shift to a model that is reminiscent of Chomsky's original one in the 1950's, with structure building done by generalized transformations. The derivation begins with a 'numeration', a set of lexical items selected from the lexicon. The lexical items are inserted 'on-line' in the course of the syntactic derivation. The derivation proceeds 'bottom-up' with the most deeply embedded structural unit created first, then combined with another lexical item to create a larger phrasal unit, and so on. The 'extension condition' requires that a transformational operation 'extends' the tree upwards. Decades earlier, Chomsky had argued that eliminating generalized transformations yields a simplified theory, with one class of complex operations jettisoned in favor of an expanded role for a component that was independently necessary, the phrase structure rule component. Further, that simplification was a substantial step towards answering the fundamental question of how the child selects the correct grammar from a seemingly bewildering array of choices. Eliminating one large class of transformations, generalized transformations, was a step towards addressing this puzzle. This was a very good argument. But since then, numerous discoveries and analyses have indicated that the transformational component can be dramatically restricted in its descriptive power. In place of the virtually unlimited number of available highly specific transformations of the theories of the 1950's and early 1960's, we can have instead a tiny number of very general operations: Merge (the generalized transformation, expanded in its role so that it creates even simple clausal structures), Move, Delete. The complex apparent results come not from complex transformations, but from the interactions of very simple ones with each other, and with very general constraints on the operation of transformations and on the ultimate derived outputs. The argument can then be reversed on itself: eliminate phrase structure rules. This model, similar in some respects to the original one in the 1950's, is graphically represented in the following diagram, where the point of 'spell-out' is where the derivation splits off on one branch towards PF, ultimately phonetics, while the transformational derivation itself (the 'syntactic' portion of the derivation) continues on towards LF, ultimately semantics:



(36)

A major technical goal of the minimalist program is to reduce all constraints on representation to 'bare output conditions', determined by the properties of the systems external to the language faculty (but still internal to the mind) that PF and LF must interface with. Internal to the computational system, the desideratum is that constraints on transformational derivations will be reduced to general principles of economy. Derivations beginning from the same lexical choices (the numeration, in Chomsky's term) are compared in terms of number of steps, length of movements, etc., with the less economical ones being rejected. Lexical items are assumed to be composed of 'features', some of which need to be 'checked' in particular configurations. This is what drives movement, since, all else equal, a derivation with an instance of movement is less economical than one without. A related idea ('procrastinate') is that covert movement is less costly than overt movement. Thus, movement will be overt only if it is forced to be by some special requirement. Given that movement is driven by the need for features to be checked, this special requirement is stated in terms of special kinds of 'strong' features, features that must be checked in overt syntax. WH-movement in a language like English is then driven by a strong feature in C, while the corresponding feature in a language like Chinese is weak (under the assumption that the latter has 'covert' WH-movement). Similarly, verb raising in French is driven by a strong feature in Infl, while the corresponding feature in English is weak (although something special must be said about the fact that auxiliary verbs in English do raise overtly).

The simplifying developments in the theory leading towards the minimalist approach have generally led to greater breadth and depth of understanding of both how human languages are organized and how they are acquired by children. This success has led Chomsky to put forward the audaciously minimalist conjecture that the human language faculty might be a 'perfect' solution to the problem of relating sound and meaning, the

minimal computational system given the boundary conditions provided by other modules of the mind.

Further Reading

Chomsky N (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky N (1982) *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky N (2000) Minimalist inquiries: the framework. In: Martin R, Michaels D and Uriagereka J (eds) *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, pp. 89–155. Cambridge, MA: MIT Press.

Haegeman L (1994) *An Introduction to Government and Binding Theory*, 2nd edn. Oxford: Blackwell.

Lasnik H (1999) *Minimalist Analysis*. Oxford: Blackwell.

Lasnik H and Juan U (1988) *A Course in GB Syntax: Lectures on Binding and Empty Categories*. Cambridge, MA: MIT Press.

Uriagereka J (1998) *Rhyme and Reason: An Introduction to Minimalist Syntax*. Cambridge, MA: MIT Press.

Iconicity

Introductory article

John Haiman, Macalester College, St Paul, Minnesota, USA

CONTENTS

In spite of the relative autonomy of grammar, there are many ways in which the structure of language directly reflects the structure of thought.

FUNCTIONAL EXPLANATIONS

> George Bush and his driver are out for a drive in the country. Suddenly a pig crosses their path and they run over it. 'Oh, my gosh!' says Bush. 'Go tell the farmer what's happened. I'll wait here.' An hour later, the chauffeur staggers out of the farmhouse, carrying a bottle of wine and a cigar, his clothing ripped. 'What happened?' asks Bush. 'Well, I got the wine from the farmer, the cigar from the farmer's wife, and for the last hour, their daughter has been making love to me', replies the chauffeur. 'What did you tell them?' asks Bush, astonished. 'I just said "I'm George Bush's driver, and I've killed the pig."'

Any speaker of English will realize that this scenario is linguistically impossible. The chauffeur must have uttered sentence 1, but the farmer's family could have responded only to sentence 2:

I'm George Bush's driver, and I've killed the PIG. (1)

I'm George Bush's driver, and I've KILLED the pig. (2)

In sentence 2, *the pig* is acting like a pronoun whose antecedent is *George Bush*. Speakers of English know that such pronouns are uttered more softly than the names that refer to their referents for the first time. In this respect pronouns are like copies, fainter than their originals. The minimal contrast in the two sentences above reappears in the grammars of all the languages we know: personal pronouns are generally short and unstressed, and often affixes on other words, while common and proper nouns may be either long or short, and are invariably words. The ease with which we can distinguish between sentences like 1 and 2 demonstrates that this contrast is not accidental, but systematic, and reflects the following principle:

Give reduced expression to what is already familiar. (3)

We should note that principle 3 violates a principle of grammar widely accepted in post-Saussurean linguistics, the principle of the autonomy of grammar: it invokes a system-external motivation ('familiarity') for a system-internal linguistic fact ('reduced expression').

Such 'functional' principles may be of various types. One of the most popular is the 'principle of least effort', made famous by Passy, von der Gabelentz, and Zipf. Since speakers, like all creatures, will exert themselves as little as possible to achieve their ends, they will resort to abbreviations when referring to something their hearers can already identify. Therefore, the bulk of a referring expression will tend to correlate with the novelty of its referent.

This principle seems to apply to contrasts between descriptions like *a one-eyed, one-horned flying purple people-eater* (on first mention) and *it* or zero (on subsequent mentions): it does not apply to the contrast between sentences 1 and 2, since the speaker's total energy output in these contrasting sentences is identical. One word in each sentence is uttered louder than the others: that word is *killed* when *pig* is a pronoun, and *pig* when *pig* is a noun. The stress difference directly signals a semantic contrast, rather than emerging from the speaker's tendency to minimize effort. The name is loud, as befits a novelty; the pronoun is soft, as befits a 'copy'.

External motivation of this sort (according to which 'the sound fits the sense') is known as iconicity. All external motivations violate systematicity, but iconicity in particular would seem to violate an even more fundamental grammatical assumption, the 'arbitrariness of the linguistic sign'. This assumption has been accepted by linguists since Plato's *Cratylus*, and at the word level, it is largely true. Cases of onomatopoeia, which seem to

contravene the principle, are marginal, and often language-specific.

Because system-internal explanations account so well for many linguistic facts, and because onomatopoeia is such a marginal phenomenon, iconicity has until recently received little attention as a structural grammatical principle. Yet some of the most widely recognized operative principles of language structure are iconic.

The minimal units in an utterance may be purely conventional. But the ways in which they are put together may imitate the ways in which their referents are related in our minds, as in a diagram (whether of a football sequence or of a radio wiring). Minimal elements in languages can be related to other elements both 'paradigmatically' and 'syntagmatically'. Different words that 'compete' for the same position in a message constitute a paradigm, while words that co-occur constitute a syntagm.

PARADIGMATIC ICONICITY

In the minimally contrasting pair of sentences 1 and 2, there is a paradigmatic contrast between the mutually exclusive words *PIG* and *pig*. Paradigmatic contrasts are frequently iconic.

Benveniste pointed out that the expression of the third person singular (whether as a pronoun or as a verbal affix) is often zero. The iconic motivation for this is that while the first person represents the speaker, and the second person represents the hearer, the third person represents someone who is absent. The zero morpheme represents the non-person. This is not a mere historical accident. As Watkins and others have shown, where paradigms are restructured, the third person singular is frequently reinterpreted as if it were zero.

Arguments for the autonomy of grammar often invoke the arbitrariness of grammatical categories in general. As Bloomfield asked (rhetorically): why is *wheat* a mass noun, while *oats* is a count noun? The answer is that this is an externally unmotivated, purely system-internal, fact about English. But there are a few nouns in English, like *hair*, that can belong to both categories: *nice hair* is mass, but *five gray hairs* is count. And, as any speaker of English will realize, the count form is used in exactly those cases where the hairs are thought of as individual countable objects.

Another example of an iconically exploited paradigmatic contrast is that between middle and reflexive verbs, for verbs of position, motion, and grooming, as in *I got up* (middle) and *I got myself up* (reflexive). The middle verb names only one participant. The reflexive, by naming two, suggests a mind–body duality.

Principle (3) is therefore subsumed in:

> Given a paradigmatic contrast between nearly synonymous forms, the formal difference will mirror the semantic difference. (4)

Sometimes paradigmatic contrasts with the same cognitive motivation will be frozen in the vocabulary or the grammar, like the contrast between pronouns and common nouns. In English, and in many other languages, there are minimally contrasting expressions like *kill* and *cause to die*. Where the two events of causation and result are expressed by separate words, there is the possibility that the two events are also separated in real time and space. To 'raise' a cup is usually to lift it via physical contact; to 'cause the cup to rise' is usually to effect the change of state without such contact. Lexicalization patterns will iconically reflect conceptual distinctions.

A similar contrast between levels of juncture is manifested by minimal contrast pairs like *lighthouse keeper* and *light housekeeper*: the conceptual contrast is reflected in the orthography (a 'lighthouse' is one thing; a 'housekeeper' is one person), but, as Bolinger and Gerstman demonstrated, it is also reflected in the spoken language: a pause corresponding to the word space is what makes these examples unambiguous to native listeners.

A formal-linguistic near-universal is that derivational affixes will tend to occur closer to the root than inflectional affixes, as for example in *king-dom-s*. Bybee has proposed that this principle may reflect a deeper cognitive principle, that the closeness of an affix to a root will reflect its relevance to that root. This principle may also account for the relative order of inflectional affixes, as in *child-ren-'s* (number closer to the root than case).

In English, the conceptual distinction between *my purse* and *my good name* is familiar enough, but the nature of the distinction is not clear in the grammar. But in many languages, a formal distinction is made between alienable possession (*my purse*) and inalienable possession (*my good name*). In all of those cases, the formal-linguistic distance between the possessor and the thing possessed reflects the conceptual distance between the two.

All of these examples reflect, more or less precisely, a principle which, far from being marginal, has been called (by Behaghel and Bolinger) the 'first law of syntax':

> Conceptual closeness is mirrored in formal closeness. (5)

More specifically, given two minimally contrasting expressions that differ in the relative distance between constituents *X* and *Y*, the relative formal distance between *X* and *Y* will correspond to the relative conceptual distance between their referents.

Still within the realm of paradigmatic relations is the fact of analogy as a motivation for language change. The fundamental principle of linguistic analogy is:

> The same meaning should be expressed via the same form. (6)

This principle is responsible for such familiar changes as extension (for example, almost all plurals in English are now marked with the same suffix *-s*; almost all verbs are now weak and mark the past tense with the same suffix *-ed*), paradigm coherence, and back formation. It is also responsible for the typological recurrence of motivated polysemy (for example, 'if' clauses in many unrelated languages are marked in the same way as topics – they may exhibit the same word order or occur with the same grammatical particles – because in some ways that is what they are), and for the fact that to a considerable extent grammatical categories and parts of speech correspond with conceptually homogeneous sets of things or relations in the world. The set of nouns in a language may seem somewhat arbitrary to speakers of another language, but it will include the names of most time-stable things.

SYNTAGMATIC ICONICITY

Words and linguistic expressions within a language relate to other words with which they compete, but also to those with which they co-occur in a message. Because words are uttered in order, syntagms are fundamentally asymmetrical; and it is in the asymmetrical syntagm that some of the clearest examples of iconicity occur.

Other things being equal, the order of clauses in a narrative corresponds to the order of events. Within a clause, a great deal of variety is possible, but in the majority of cases, subjects precede objects, definite noun phrases are placed earlier than the corresponding indefinites, topics are placed before comments, causes are placed before results, and in general:

> Old information tends to be placed before new information. (7)

Behaghel called this iconic principle the 'second law of syntax'.

The tendency towards systematization (driven ironically by analogy, since it is widely accepted that the drive to create analogies is iconically motivated) is undoubtedly responsible for the continual loss of iconicity in grammars. So too is the tendency to ritualization (manifested in grammaticalization), and erosion (manifested in sound change). But iconicity must be more than just what remains from 'the first language', or it would have disappeared by now. For iconicity to be as prevalent as it is, it must be a creative force contending partly against grammaticalization and erosion, each of which tends to create arbitrary structures.

Iconicity may even manifest itself in 'the dog that didn't bark in the night'. For example, many languages manifest a kind of aversion to 'accidental repetition'. This tendency is neatly illustrated by the sporadic process of haplology whereby an identical sequence of sounds is avoided: thus *haplology* might be replaced by *haplogy*. But repetition is rife in language, and is used to iconically signal repetition, plurality, and intensity at the levels of grammar and discourse. In reducing accidental repetition, haplology clears the decks: repetition is so important as an iconic signal that no iconically unmotivated repetition is tolerated.

Further Reading

Anttila R (1972) *Introduction to Comparative and Historical Linguistics*. New York, NY: Macmillan.

Behaghel O (1932) *Deutsche Syntax*, vol. IV. Heidelberg, Germany: Carl Winter.

Benveniste E (1946) Relations de personne dans le verbe. *Bulletin de la Société de Linguistique* **43**: 1–12.

Bloomfield L (1933) *Language*. Chicago, IL: University of Chicago Press.

Bolinger D and Gerstman L (1957) Disjuncture as a clue to constructs. *Word* **13**: 246–255.

Fischer O and Naenny M (eds) (1999) *Form Miming Meaning*. Amsterdam, Netherlands: Benjamins.

von der Gabelentz G (1891) *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse*. Leipzig, Germany: Weigel.

Greenberg J (1966) Some universals of language with particular reference to the order of meaningful elements. In: Greenberg J (ed.) *Universals of Language*, 2nd edn, pp. 73–113. Cambridge, MA: MIT Press.

Haiman J (1985a) *Natural syntax*. Cambridge, UK and New York, NY: Cambridge University Press.

Haiman J (ed.) (1985b) *Iconicity in Syntax*. Amsterdam, Netherlands: Benjamins.

Jakobson R (1965) Quest for the essence of language. *Diogenes* **21**: 21–37.

Menn L and MacWhinney B (1984) The repeated morph constraint: toward an explanation. *Language* **60**: 519–541.

Peirce C (1932) *Philosophical Writings*, vol. II. Cambridge, MA: Harvard University Press.
de Saussure F (1969) *Cours de Linguistique Générale*. Paris: Payot. [First published 1916.]
Watkins C (1962) *Indo-European origins of the Celtic verb. Part One: The Sigmatic Aorist*. Dublin, Ireland: Institute for Advanced Studies.
Zipf GK (1935) *The Psychobiology of Language*. Boston, MA: Houghton-Mifflin.

Implicature

Introductory article

Laurence R Horn, Yale University, New Haven, Connecticut, USA

CONTENTS

Implicature is a non-truth-conditional aspect of speaker meaning that represents (part of) what is meant in a speaker's utterance without being part of what is said. What the speaker intends to convey is almost always far richer than what he or she directly expresses, as linguistic meaning radically underdetermines the communicated message. The speaker tacitly exploits pragmatic principles to bridge this gap.

THE GRICEAN MODEL

The Gricean model proposes a bridge connecting what is said (largely computed directly from the grammatical structure of the uttered sentence, but with reference resolved) with what is implicated (indirectly expressed). Implicated meaning encompasses two distinct phenomena: 'conventional' implicature, involving non-truth-conditional aspects of meaning constituting appropriateness conditions on the use of a given word or construction, and 'conventional' implicature, a relation based on a computation of speaker's intention operating from the presumption that the speaker is interacting rationally and cooperatively with the addressee.

Different types of implicature are illustrated by the pairs of sentences below, in each of which the second (primed) member is (sometimes) deducible from its unprimed counterpart by virtue of a nonlogical inferential relation:

(a) Even Bill knows it's unethical.
  (a′) Bill is the least likely [of a contextually invoked set] to know it's unethical.
(b) She has a good personality.
  (b′) She's not that attractive.
(c) The cat is either in the hamper or under the bed.
  (c′) I don't know for a fact that the cat is under the bed. (1)

Unlike an entailment or logical presupposition, the inference induced by *even* in sentence 1(a) is irrelevant to the truth conditions of the proposition: sentence 1(a) is true if and only if Bill knows it's unethical. The inference is not cancelable (one cannot say 'even Bill knows it's unethical, but that's not surprising'), but it is detachable, in the sense that the same truth-conditional content is expressible in a way that removes (detaches) the inference: 'Bill knows it's unethical too'. Such detachable but non-cancelable inferences, which are neither part of what is said (or of truth-conditional content) nor calculable from what is said, are conventional implicata, akin to pragmatic presuppositions. Indeed, along with connectives like *but*, the standard examples of conventional implicature involve precisely those particles traditionally analyzed as instances of pragmatic presupposition: adverbial particles like *even* and *too*, truth-conditionally transparent verbs like *manage to* and *bother to*, and focus constructions like clefts.

But whereas these inferences are non-truth-conditional components of an expression's conventional lexical meaning, the inferences associated with sentences 1(b) and 1(c) are non-conventional in that they are calculable from the utterance of such sentences in a particular context, given the nature of conversation as a shared goal-oriented enterprise. In each case, the inference of the corresponding primed proposition is cancelable (either explicitly by appending material inconsistent with it ('but I don't mean to suggest that …') or by altering the context of utterance) but nondetachable (given that any other way of expressing the literal content of sentences 1(b) and 1(c) in the same context would license the same inference).

Sentence 1(b) differs from sentence 1(c) in the generality of the circumstances in which the inference is ordinarily licensed. Only when the speaker of sentence 1(b) is responding to a query about the attractiveness of the referent will the addressee

normally infer that the speaker had intended to convey the content of sentence 1(b′): this is an instance of 'particularized' conversational implicature. In sentence 1(c), on the other hand, the inference – that the speaker does not know in which of the two disjoined locations the cat can be found – is induced in the absence of a special or marked context. Thus sentence 1(c′) represents an instance of the linguistically significant concept of 'generalized' conversational implicature. But in both cases, as with conventional implicature, it is not the proposition or sentence, but the speaker or utterance, that induces the relevant implicatum.

Participants in a conversational exchange systematically compute what was meant (by the speaker's utterance at a given point in the interaction) from what was said, operating from the assumption that both the speaker and the hearer are rational agents. The conversation is governed by the 'cooperative principle', which stipulates that the speaker is expected to make his or her contribution appropriate for the current purposes of the exchange. This general principle is analyzed into the four general and arguably universal maxims of conversation on which all rational interchange is grounded:

- Quality (try to make your contribution true):
  - Do not say what you believe to be false.
  - Do not say what you lack evidence for.
- Quantity:
  - Make your contribution as informative as is required (for the current purposes of the exchange).
  - Do not make your contribution more informative than is required.
- Relation (be relevant).
- Manner (be perspicuous):
  - Avoid obscurity of expression.
  - Avoid ambiguity.
  - Be brief.
  - Be orderly.

Note, however, that not all these maxims are equally important. Following Grice himself, many have posited a privileged status to 'quality', on the grounds that unless quality – or a convention of truthfulness – is observed, none of the other maxims can easily be satisfied: false 'information' is no information at all. Other researchers, working within the framework of 'relevance theory', have challenged the primacy of the principle of quality and advocated its subsumption within an all-encompassing revised principle of relevance.

SCALAR IMPLICATURE AND CONSTRAINTS ON LEXICALIZATION

Within neo-Gricean pragmatics, the starting point is the first quantity maxim, which is systematically exploited to yield upper-bounding generalized conversational implicatures associated with scalar operators. Quantity-based scalar implicature (e.g., my inviting you to infer from my use of *some* … that for all I know *not all* …) is driven by our mutual knowledge that I expressed a weaker proposition in lieu of an equally unmarked utterance that would have expressed a stronger proposition. Thus, what is said in the use of a weak scalar value like those in bold face in the sentences of Table 1 is the lower bound ('at least …'), with the upper bound ('at most …') implicated as a cancelable inference generated by (some version of) the first maxim of quantity. What is communicated in the default case is the 'two-sided reading', which combines what is said with what is implicated.

Negating such predications denies the lower bound: to say that something is not possible is to say that it's impossible; i.e., less than possible. When the upper bound appears to be negated ('It's not possible, it's necessary'), a range of syntactic, semantic, and intonational evidence indicates that we are dealing with an instance of the 'metalinguistic' or echoic use of negation, in which the negative particle is used to object to any aspect of a mentioned utterance, including its conventional and conversational implicata, register, morphosyntactic form, or pronunciation. If it's hot, it's (*a fortiori*) warm, but if I know it's hot, the assertion that it's warm can be echoed and rejected as (not false but) insufficiently informative:

Table 1. What is said vs. what is communicated in examples with scalar operators

Sentence	*One-sided reading*	*Two-sided reading*
Pat has **three** children.	'… at least three …'	'… exactly three …'
You ate **some** of the cake.	'… some if not all …'	'… some but not all …'
It's **possible** she'll win.	'… at least …'	'… but not certain …'
He's a knave **or** a fool.	'… and perhaps both'	'… but not both'
It's **warm**.	'… at least warm'	'… but not hot'

(a) It's not warm, it's hot!
(b) You're right, it's not warm – it's hot! (2)

As seen in sentence 2(b), the metalinguistic understanding typically requires a second pass, and the effect is typically that of an ironic unsaying or retroactive accommodation.

The central role played by scalar implicature in natural language is illustrated by a systematic pattern of lexical gaps and asymmetries. Consider the post-Aristotelian square of opposition defined by the logical relations among quantified expressions (ranging over non-empty sets), shown in Figure 1.

Note in particular that the assertion of either of the two subcontraries quantity-implicates the negation of the other. While what is said in 'some men are bald' and 'some men are not bald' is distinct, what is communicated is typically identical: 'some men are bald and some aren't'. Given that languages tend not to lexicalize complex values that need not be lexicalized, we might predict that *some … not* tends not to be lexicalized, and this is precisely what we find.

In a wide variety of languages, values mapping onto the **O** corner of the square are systematically restricted in their potential for lexicalization. Thus alongside the quantificational determiners *all*, *some*, and *no*, we never find an **O** determiner **nall* ('not all'); corresponding to the quantificational adverbs *always*, *sometimes*, and *never*, we have no **nalways* ('not always', 'sometimes not'). We may have

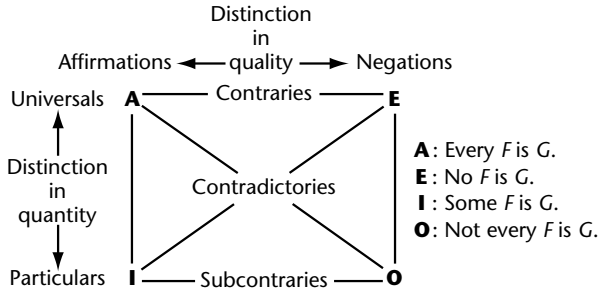


Figure 1. The square of opposition. Corresponding **A** and **E** statements are 'contraries': they cannot be simultaneously true (though they may be simultaneously false). Corresponding **A** and **O**, and **I** and **E**, statements are 'contradictories': they cannot be simultaneously true or simultaneously false. An **I** statement is the 'subaltern' of its corresponding **A** statement, and an **O** statement of its corresponding **E** statement: a subaltern is unilaterally entailed by its corresponding superaltern. Corresponding **I** and **O** statements are 'subcontraries': they cannot be simultaneously false (though they may be simultaneously true).

equivalents for *both (of them)*, *one (of them)*, and *neither (of them)*, but never for **noth (of them)* ('not both', 'at least one … not'); we find connectives corresponding to *and*, *or*, and sometimes *nor* ('and not'), but never to **nand* ('or not', 'not … and'). The missing-**O** phenomenon extends to the modals and deontics, as illustrated by the fact that the inflected negative in 'he can't go', or the orthographic lexicalization in 'he cannot go', only allows wide-scope ('**E**' vertex) negation, while the unlexicalized counterpart 'he can not go' is ambiguous. The relation of mutual quantity implicature holding between positive and negative subcontraries results in the superfluity of one of the two subcontraries for lexical realization, while the functional markedness of negation assures that the unlexicalized subcontrary will always be **O**.

Q-BASED AND R-BASED IMPLICATURE

The significance of the first quantity maxim for the form and function of natural language reflects its status as one of two cardinal principles regulating the economy of linguistic information. Setting quality aside as irreducible, we can collapse the remaining maxims and submaxims into two fundamental principles. The 'Q principle' is a lower-bounding hearer-based guarantee of the sufficiency of informative content ('say as much as you can, modulo quality and R'); it collects the first quantity maxim along with the first two 'clarity' submaxims of manner, and is systematically exploited (as in the scalar cases) to generate upper-bounding implicata. The 'R principle', by contrast, is an upper-bounding correlate of the 'law of least effort' dictating minimization of form ('say no more than you must, modulo Q'); it collects the relation maxim, the second quantity maxim, and the last two submaxims of manner, and is exploited to induce strengthening or lower-bounding implicata.

Q-based implicature is typically negative in that its calculation refers crucially to what could have been said but wasn't: the hearer infers from the speaker's failure to use a more informative or briefer form that the speaker was not in a position to do so. R-based implicature typically involves social rather than purely linguistic motivation. It is exemplified by indirect speech acts and negative strengthening (including so-called 'neg-raising', the tendency for 'I don't think that …' to implicate 'I think that not…').

R-based implicata, while calculable (as are all conversational implicata), are often not calculated

'online'; a specific form of expression may be associated with a given pragmatic effect while an apparently synonymous form is not. Thus, the question 'Can you close the window?' is generally used to convey an indirect request, while 'Are you able to close the window?' is not; 'I don't guess that …' allows a strengthened 'neg-raised' understanding in only a subset of the dialects for which 'I don't think that …' does. These are instances of 'standardized nonliterality', or 'short-circuited conversational implicature'.

The two antinomic Q and R forces interact dialectically, each appealing to and constraining the other. Thus Grice incorporates R in defining the primary Q maxim ('make your contribution as informative as is required'), while the second quantity maxim is constrained by the first and essentially incorporates relation: what could make a contribution more informative than is required, except the inclusion of contextually irrelevant material?

The opposition between the two forces may result in maxim clash. Thus an utterance of 'I broke a finger yesterday' R-implicates that it was one of my own fingers I broke, unless the common ground entails or accommodates the proposition that I am enforcer for the mob, in which case the opposite, Q-based implicature is derived.

Notice too that *finger* here conveys 'non-thumb'. In such Q-based narrowing, the existence of a specific hyponym *H* of a general term licenses the use of the general term for the complement of the extension of *H*. This is further illustrated by the development of a specific use or sense of *dog* (excluding bitches), *cow* (excluding bulls), or *animal* (excluding humans, birds, and fish). This is distinct from R-based narrowing, in which the restriction of a more general lexical item, such as *poison*, *liquor*, *drink, man*, or *undertaker*, to a particularly salient subset or exemplar of the original denotation is not prompted by the existence of a specific word preempting that portion of semantic space.

Related to Q-based narrowing is the 'division of pragmatic labor': given two coextensive expressions, a relatively unmarked form – briefer or more lexicalized – will tend to become R-associated with a particular unmarked, stereotypical meaning, use, or situation, while the use of the periphrastic or less lexicalized expression, typically more complex or prolix, will tend to be Q-restricted to those situations outside the stereotype, for which the unmarked expression could not have been used appropriately. Consider the following pairs of sentences:

(a) He got the machine to stop.
(a′) He stopped the machine.
(b) Her blouse was pale red.
(b′) Her blouse was pink.
(c) She wants her to win.
(c′) She wants to win.
(d) I am going to marry you.
(d′) I will marry you.
(e) My brother went to the church (the jail, the school).
(e′) My brother went to church (jail, school).
(f) It's not impossible that you will solve the problem.
(f′) It's possible that you will solve the problem.
(g) That's my father's wife.
(g′) That's my mother. (3)

The use of the periphrastic causative in sentence 3(a) implicates that the agent achieved the effect in a marked way (perhaps by pulling the plug or throwing a shoe into the machine); *pale red* in sentence 3(b) implicates a tint not preempted by *pink*; the selection of a full pronoun over a null pronoun in sentence 3(c) signals the absence of the coreferential reading associated with the reduced syntax; in sentence 3(d) the periphrastic form blocks the indirect speech act function of promising conveyed by the modal; the full + noun versions of sentence 3(e) imply literal motion to the specified location without the socially stereotypic connection R-associated with the corresponding institution on the anarthrous version; the double (contradictory) negation in sentence 3(f) signals a rhetorical effect absent from the direct positive; and the more complex description in sentence 3(g) suggests that the more basic and lexicalized alternative could not have been used appropriately (the referent is probably the speaker's stepmother). When a speaker opts for a more complex or less fully lexicalized expression over a simpler alternative, there is a sufficient reason, but which reason depends on the particular context.

In addition to maxim clash and the division of pragmatic labor, another area of conflict between the Q and R principles is in implicature derivation itself. If I tell you that my wife is either in the kitchen or the bedroom, I will Q-implicate that I don't know that she's in the kitchen; but I can tell you 'the kitchen is a mess' without implicating that the bedroom isn't. If you tell me that something is 'possibly' true, I will assume you don't know it's true; but if you tell me that something is true (e.g., that all bachelors are unmarried), I will not assume you don't know it's necessarily true. The use of the

weak **I** or **O** proposition licenses the inference that the speaker was not in a position to use the basic unquantified, unmodalized proposition that unilaterally entails it, as the Q principle predicts; but the use of the basic propositional form does not Q-implicate the negation of its strong counterpart, **A** or **E** respectively. Since there is no quantity- or information-based distinction between these subalternations, we must seek the source of the asymmetry elsewhere. The crucial distinction here relates not to the content (what is said) but to the form (how what is said is said): it is because the intermediate values are not only more informative but briefer than their **I** or **O** counterparts that the use of the latter will strongly implicate against the former. But the strong values, while more informative than their unmodified counterparts, are also more prolix, so quantity here is offset by manner and potentially by relation: the Q principle of informative sufficiency yields to the R principle of least effort. The richness of the pragmatic framework makes it possible to begin to develop a theory of not just what can be implicated but what will be implicated in a given context.

When degree of lexicalization is not a factor, scalar implicature normally goes through. Thus, each of the following ordered lists constitutes a Q-relevant scale in that the affirmation of any weak or intermediate value will implicate (*ceteris paribus*) that – for all the speaker knows – the value on its left could not have been substituted:

- *always*, *usually*, *often*, *sometimes*
- *and, or*
- *certain*, *likely*, *possible*
- *cold*, *cool*, *lukewarm*
- *excellent*, *good*, *OK*

But when the stronger value is less economical than the weaker one, no Q-implicature is triggered. This extends to non-quantitative 'scales' of items differing in informative strength. Thus, while the use of *finger* typically conveys 'non-thumb', it does not convey 'non-pinky (finger)', nor does the use of *toe* convey 'toe other than the big toe', although the big toe is analogous to the thumb. What is crucial is the status of *thumb* (as opposed to *pinky*) as a viable lexicalized alternative to *finger*. In the same way, *rectangle* conveys 'non-square' (i.e., 'non-equilateral rectangle') because of the availability of the lexicalized alternative *square*, while *triangle* will not convey 'non-equilateral triangle' – indeed, the prototype triangle is equilateral – because of the non-existence of a lexicalized term for a general triangle.

The model described above retains two antinomic principles along with an unreduced maxim of quality. A more radical simplification has been proposed in the framework of relevance theory, in which a redefined 'principle of relevance' is taken to be the only source of pragmatic inference required.

IMPLICATURE, EXPLICATURE, AND PRAGMATIC INTRUSION

Even for Grice, propositional content is not fully fleshed out until reference, tense, and other indexical elements are fixed. But proponents of relevance theory have pointed out that the pragmatic reasoning used to compute implicated meaning must also be invoked to fill out underspecified propositions where the semantic meaning contributed by the linguistic expression itself is insufficient to yield a proper accounting of truth-conditional content. Thus, when a news reporter observed as the jury retired to consider their verdict in the O. J. Simpson murder trial that 'It will take them some time to reach a verdict', the proposition he communicated (that it will take a long time) was, in the event, false, a fact inconsistent with a strict Gricean analysis on which the time communicated by the speaker is just an implicatum read from the underspecified content contributed by linguistic meaning alone – a trivially true existential proposition. Instead, the pragmatically recoverable strengthened communication comprises the 'explicature' or truth-conditional content. Thus, pragmatically derived aspects of meaning are not necessarily implicatures; indeed, there is substantial pragmatic intrusion into propositional content.

Such apparent intrusion is also illustrated by the temporal and causal asymmetry of conjoined event-denoting verb phrases and sentences. On the standard Gricean account, the conjunction in sentence 4(a) will typically implicate sentence 4(b) by virtue of the 'be orderly' submaxim of manner:

(a) They had a baby and they got married.
(b) They had a baby and then they got married. (4)

Arguments against a lexical ambiguity for *and* ('and also' versus 'and then') include the following.

- On the two-*and* theory, conjunction in virtually every language would just happen to be ambiguous in the same way.
- No natural language contains a conjunction *C* that would be ambiguous between 'and also' and 'and earlier' readings, so that 'they had a baby [*C*] they got

married' would be interpreted either atemporally or as 'they had a baby and, before that, they got married'.
- The same 'ambiguity' exhibited by *and* arises when two clauses describing related events are juxtaposed without an overt connective. ('They had a baby. They got married.')

If conjunctions are semantically univocal, while manner-implicating that the events occurred in the order in which they were described, the impossibility of the conjunction *C* can be attributed to the absence of any maxim enjoining the speaker to 'be disorderly'. As with scalar implicature, the asymmetric implicatum may be canceled or suspended: 'they had a baby and got married, but not necessarily in that order'. But if the 'and then' reading comes in only as an implicature, it is hard to explain its apparent contribution to truth-conditional meaning in embedded contexts, and in particular the non-contradictory nature of the three sentences below, as pointed out by L. Jonathan Cohen and Deirdre Wilson:

(a) If they got married and had a baby, their parents will be pleased; but if they had a baby and got married, their parents will not be pleased.
(b) They didn't get married and have a baby, they had a baby and got married.
(c) It's more acceptable to get married and have a baby than to have a baby and get married. (5)

One conclusion is that, while pragmatically derived, the strengthened or enriched meaning is an explicature, corresponding to what is said rather than to what is (merely) implicated.

The explicature view yields a re-evaluation of the traditional view of scalar predications, on which both one-sided and two-sided understandings of the sentences in Table 1 will be directly represented at the level of logical content. While such scalar predications are now all taken to be ambiguous, the ambiguity is no longer situated at the lexical level but has been relocated to the propositional level: what is said in an utterance is systematically underdetermined by the linguistic content of what is uttered.

Other work has challenged some of these results. Thus, while a strong case can be made for an enrichment analysis of the meaning contribution of the cardinals, it does not generalize straightforwardly to 'inexact' scalar values. Evidence for this conclusion comes from the contextual reversibility of cardinal scales and the non-implicating ('exactly *n*') reading of cardinals in mathematical, collective, and elliptical contexts, none of which applies to the scalar operators in Table 1. Note also the contrast in the exchanges below:

A : Do you have two children?
B_1 : No, three.
B_2 : Yes, (in fact) three. (6)

A : Did many of the guests leave?
B_1 : No, all of them.
B_2 : Yes, (in fact) all of them. (7)

In exchange 6, a bare 'no' answer is compatible with an 'exactly *n*' reading in an appropriate context; while in exchange 7, an unadorned negative response can only be understood as conveying 'fewer than many'. Similarly, if 'it's warm' were truly propositionally ambiguous, there is no obvious reason why a 'no' response to the question 'Is it warm?' should not be interpretable as a denial of the enriched, two-sided content and thus as asserting that it's either chilly or hot; or why the comparative in 'it's getting warmer' cannot denote 'less hot' instead of 'less cold'. This suggests the need for a mixed theory in which cardinal values may demand a relevance-theoretic pragmatic enrichment analysis, while other scalar predications continue to submit to a standard neo-Gricean treatment on which they are lower-bounded by their literal content and upper-bounded, in default contexts, by Q-implicature.

IMPLICATURE VERSUS IMPLICITURE

The arguments we have been reviewing rest on the tacit assumption that whatever is communicated but not said must be implicated. Some have argued from this assumption that implicatures can affect truth-conditional meaning after all, given cases like the asymmetric conjunction in sentences 4(a) and 4(b); others have argued instead for the notion of explicature; i.e., pragmatically determined content. But if there can be implicit components of communicated meaning that are not implicatures, these aspects of meaning – such as the bracketed expansions below – need not be considered part of what is said:

(a) I haven't had breakfast {today}.
(b) John and Mary are engaged {to each other}.
(c) They had a baby and they got married {in that order}.
(d) Robin ate the shrimp and {as a result} got food poisoning.
(e) Everybody {in our pragmatics class} solved the riddle. (8)

In each case, the bracketed material cannot be derived as an implicature, but neither can it be part of what is said, since it is felicitously cancelable ('John and Mary are engaged, but not to each other'). It may be regarded instead as an 'impliciture', an implicit weakening, strengthening, or specification of what is said. This approach, urged by Kent Bach, permits an intuitive characterization of propositional content, a conservative mapping from syntactic structure to what is said, and an orthodox Gricean conception of implicature, albeit as a more limited construct than in much neo-Gricean work. At the same time, the standard view that every sentence expresses one and only one proposition must be abandoned, as it is typically, and in some cases only, the impliciture – the expanded proposition conveyed but not directly expressed – that is assessed for truth or falsity.

With a fuller understanding of the interaction between pragmatics and propositional content, we see that while the explanatory scope of conversational implicature may have been reduced from that of the original Gricean program, this program and the pragmatic principles motivating it – rationality, common ground, and the distinction between implicit and explicit components of utterance meaning – continue to play an essential role in the elaboration of dynamic models of context. As recent work on language acquisition and on lexical change has further demonstrated, a suitably sharpened notion of implicature remains at the heart of communicated meaning.

Further Reading

Bach K (1994a) Conversational impliciture. *Mind and Language* **9**: 124–162.

Bach K (1994b) The myth of conventional implicature. *Linguistics and Philosophy* **22**: 327–366.

Carston R (1988) Implicature, explicature, and truth-theoretical semantics. In: Kempson (ed.) *Mental Representations*, pp. 155–181. Cambridge, UK: Cambridge University Press.

Carston R (1995) Quantity maxims and generalized implicature. *Lingua* **96**: 213–244.

Green G (1990) The universality of Gricean interpretation. *Proceedings of the Berkeley Linguistics Society* **16**: 411–428.

Grice HP (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hirschberg J (1991) *A Theory of Scalar Implicature*. New York, NY: Garland.

Horn LR (1990) Hamburgers and truth: why Gricean inference is Gricean. *Proceedings of the Berkeley Linguistics Society* **16**: 454–471.

Horn LR (1993) Economy and redundancy in a dualistic model of natural language. In: Shore and Vilkuna (eds) *SKY 1993: 1993 Yearbook of the Linguistic Association of Finland*, pp. 33–72.

Karttunen L and Peters S (1979) Conventional implicature. In: Oh CK and Dinneen D (eds) *Syntax and Semantics*, vol. XI, *Presupposition*, pp. 1–56.

Levinson SC (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

Récanati F (1989) The pragmatics of what is said. *Mind and Language* **4**: 295–329.

Sperber D and Wilson D (1986) Relevance. Cambridge, MA: Harvard University Press.

Innateness and Universal Grammar

Introductory article

Stephen Crain, University of Maryland at College Park, Maryland, USA
Paul Pietroski, University of Maryland at College Park, Maryland, USA

CONTENTS

In linguistics, the Innateness Hypothesis is the claim that all children have, by virtue of a common biology, a 'Universal Grammar' that defines a space of possible human languages. Children explore this space, influenced by the environment, until they stabilize on grammars that are equivalent to those of adult speakers in the linguistic community.

INTRODUCTION

Human beings speak languages, so children must have the capacity to speak languages. In order to understand this capacity, theorists often focus on a cluster of observations about how language develops in children. (1) While languages differ in certain respects, the capacity to speak a language is universal in our species, such that any normal child can acquire any human language. (2) This capacity is manifested at an early age, and it appears to be independent of other cognitive abilities. (3) Despite considerable variation of experience within the same linguistic community, all normal children in that community converge on (more or less) the same grammar. (4) There is a dramatic gap between the linguistic experience available to children and the kinds of linguistic competence that children achieve as a matter of course.

These observations are noteworthy because they are hallmarks of innateness: facts which suggest that speaking a language is a property human beings have, at least in large part, by virtue of their biological endowment – much like the property of having a liver or being able to see. Of course, it is always possible that a trait bearing the hallmarks of innateness will turn out to be acquired (by all children) through experience. And it is a truism that each individual's properties stem from interactions of her particular genetic heritage with her environment. But our common biology presumably plays the major role in explanations for why human beings have livers, even though different people have different livers (which are affected by diet).

There has, however, been resistance to the very idea that our common biology plays the major role in explanations for why human beings speak languages. So it bears emphasis, first, that no one proposes that the capacity to speak a particular language, say English, is innate. Children who acquire English have obviously been influenced by their environment, in a way that differs from how children who acquire Japanese have been influenced by their environment. The Innateness Hypothesis is the proposal that all children have, by virtue of their common biology, a property one might call 'having Human Language'; and this 'Universal Grammar' (UG) interacts with each child's linguistic experience to determine the specific language(s) that the child will acquire. From this perspective, linguists try to find the common core of human grammars, that is, the universal principles that underlie any language that human children can acquire.

PRINCIPLES AND PARAMETERS

Linguists are also in the business of explaining how and why specific human languages differ. According to one active line of research, called the Principles and Parameters approach, even the possible dimensions of variation among human languages are innately specified, at least to a large extent. To take a simple example, direct objects precede verbs in many languages, so many children will encounter strings of words like 'Chris rice eats'. This differs from the corresponding experiences of American children, who encounter verbs preceding direct objects, as in 'Chris eats

rice'. This word-order variation is explained by a parameter. The idea is that subject–object–verb experiences interact with Universal Grammar to produce languages of one sort, while subject–verb–object experiences interact with Universal Grammar to produce languages of another sort. So while the environment has an effect, the number of options available – and thus the range of possible languages the child can acquire – is severely constrained in advance of any experience. To the extent that the relevant parameter can be 'set' by simply detecting features of linguistic experience (without significant computation or extrapolation), talk of 'learning' will be unhelpful.

Linguists who argue for the innate specification of linguistic principles and parameters typically have nothing to say about how our genes give rise, in the normal course of events, to the biological structures that underlie human linguistic capacities. Rather, linguists argue that certain features of human languages are determined by innate aspects of the human mind/brain, as opposed to being determined by aspects of human environments. The argument rests on two main observations. First, children encounter a finite and haphazard collection of expressions, each of which presumably conveys a single meaning in its conversational context. Yet children internalize grammars that go beyond these experiences along several dimensions. Children can produce and comprehend an unlimited number of novel expressions. They can discern paraphrase and ambiguity relations among linguistic expressions. Second, children attain knowledge of linguistic structures for which they have no corresponding experience. For example, children are not systematically informed that certain linguistic expressions are unacceptable in the local language, but this knowledge emerges in the course of language development.

POVERTY-OF-THE-STIMULUS PROBLEMS

Examples of knowledge not drawn from experience – illustrations of the 'Poverty of the Stimulus' – are offered in (1) and (2). Examples (1a,b) indicate that the verbal elements *want* and *to* may sometimes be contracted to form *wanna*. In fact, contraction of *want* and *to* is licensed in general; the deviance of (1d), which is the exceptional case, is indicated by '#'. If children were to generalize on the basis of examples like (1a,b) using standard principles of induction, children would be inclined to produce linguistic expressions like (1d), which are unacceptable for adults.

a. Who do you want to beat?
b. Who do you wanna beat?
c. Who do you want to win?
d. #Who do you wanna win? (1)

Similarly, examples (2a,b) show that contraction of the copula, *is*, is optional in certain linguistic expressions. Again, contraction is licensed in general; (2d) is the exception. So, learners who acquired language by means of general induction principles, or analogy, would be tempted to allow *is*-contraction in expressions like (2d). If human children do not produce expressions like (2d), they evidently must know (unconsciously) that contraction is held in check by some linguistic principle.

a. What is that in the tree?
b. What's that in the tree?
c. Do you know what that is in the tree?
d. #Do you know what that's in the tree? (2)

A single linguistic principle has been advanced to explain both the *wanna*-contraction and the *is*-contraction facts. On this account, *want* and *to* cannot contract in (1d) because there is an unpronounced element between them. This 'empty' element, labeled 'e' in (1d′), is associated with the question word *who*.

who do you want **e** to win (1d′)

Turning to (2), *is*-contraction is prohibited whenever an empty element follows *is*, as in (2d′), where the empty element *e* is associated with the question word *what*.

do you know **what** that is **e** in the tree (2d′)

This means that *is* contracts to the right, as in (2b′), orthographic conventions notwithstanding:

what is that **e** in the tree ⇒ **what** s′ that **e** in the tree (2b)′

But the linguistic constraint is the same for *wanna*-contraction and *is*-contraction: contraction is blocked by an empty element. Contraction is permitted in (1b) and (2b) because in these examples the empty elements associated with *who* and *what* do not interfere.

No one tells children that expressions like (1d) or (2d) are unacceptable; yet all English-speakers somehow know this. This illustrates the poverty-of-the-stimulus problem: linguistic knowledge that is demonstrably not acquired by attending to input data. The emergence of a property in the absence of environmental input is one hallmark of genetic

prespecification. This invites us to ask about other hallmarks: universality and early emergence. The prohibition against contraction across an empty element is a universal phenomenon, as far as we know. Using an experimental technique known as elicited production, it has also been shown that as soon as children can be tested, they manifest adult competence with regard to *wanna*-contraction and *is*-contraction. That is, children prefer to contract *want* and *to* when this option is licensed in the local language, but children never contract these verbal elements when this is not tolerated by adults, as in (1d). Similarly, children prefer to contract *is* (to the right), but they never produce reduced expressions like (2d), where linguistic theory posits an empty element to the right of *is*. In short, children converge on generalizations that govern the local language, but do not generalize in many other (perfectly coherent) ways; and this is so, even though children's linguistic experience would lead them to generalize in ways that speakers of the local language do not, if children were to adopt general principles of induction.

The linguistic principle blocking contraction emerges in the absence of decisive evidence in the primary linguistic data available to children, and children respect this constraint from the earliest stages of language development. Because the relevant linguistic principle bears all the hallmarks of innate specification, it is a likely candidate to be part of Universal Grammar. We noted earlier that linguistic parameters, such as word-order parameters, are also candidates for inclusion in Universal Grammar. If so, we would expect such parameters to be set by children early in the course of language development. Indeed, this seems to be the case for word-order parameters, and also for parameters involving the inflectional system. We would not expect the same acquisition scenario for more peripheral phenomena, however, which do not closely track the natural seams of human languages. Phenomena on the periphery of human languages, such as irregular forms of rare verbs, are more likely to emerge late and may be subject to individual differences.

THE CONTINUITY ASSUMPTION

In debates about the Innateness Hypothesis, the *details* of linguistic theory matter: linguists first identify principles that characterize human grammars, *then* they ask which aspects of these grammars are plausibly learned from experience, and which are more likely to be innately specified. The nativist claim is not that language must be innate since children say things they do not hear. The argument is, rather, that many aspects of language are innate, since human languages seem to exhibit features (such as the constraint on contraction) that children could not 'pick up on' given their linguistic experience. Correspondingly, although poverty-of-the-stimulus problems are superficially similar to 'induction problems' that arise in any domain where knowledge is logically underdetermined by data, there are important differences between the study of human languages and the study of other aspects of human cognition. One difference is the expectation that children may try out various linguistic options that are available in human languages, but are not attested in the primary linguistic data available in the local language. This is the 'continuity assumption'.

According to the continuity assumption, while children will not entertain linguistic hypotheses that extend beyond the boundary conditions imposed by Universal Grammar (UG), children might well entertain linguistic hypotheses with features found in languages elsewhere on the globe, despite the absence of any evidence for such hypotheses in their primary linguistic data. An example of a non-adult (but UG-compatible) production is the 'medial-*wh*' phenomenon observed in many 3- and 4-year-old children of English-speaking parents. Such children are observed consistently to insert an 'extra' *wh*-word in questions like (3) and (4):

What do you think what pigs eat? (3)

Who did he say who is in the box? (4)

These expressions are presumably not responses by children to their environment, since medial-*wh* expressions are not produced by adults in English-speaking environments. However, structures similar to (3) and (4) are attested in many languages, such as Irish, Chamorro, and in dialects of German. Given a Principles and Parameters theory that includes the continuity assumption, this is noteworthy but unsurprising: children are not born knowing if the local language allows medial-*wh* questions; so until the relevant parameters are set, children may speak a dialect of 'Human Language' that differs from the local language. This kind of mismatch between child and adult language may constitute the strongest argument for an innate universal grammar, and against models according to which children construct linguistic hypotheses based on experience.

THE NATURE VERSUS NURTURE DEBATE

The 'nature versus nurture' debate continues, as various poverty-of-the-stimulus arguments are challenged or supported by developments in linguistic theory and empirical investigations of child language. Opponents of nativism view language acquisition as an induction problem of the same sort that arises in other domains where the knowledge achieved is underdetermined by the learner's experience. This perspective highlights the 'cues' that are available in the input to children, as well as children's skills in extracting relevant information and forming generalizations on the basis of the data they receive. Nativists counter with evidence showing that children project beyond their primary linguistic data in ways that the input does not even suggest, as illustrated by the facts about *wanna*-contraction, *is*-contraction, and the medial-*wh* phenomenon. Instead of viewing language acquisition as a special case of theory induction, nativists posit a universal grammar, with innately specified principles of grammar formation. From this perspective, innate linguistic principles define a space of possible human languages – a space that children explore, influenced by the environment, until they stabilize on a grammar equivalent to the grammar(s) used by adults in the linguistic community.

Further Reading

Chomsky N (1975) *Reflections on Language*. New York, NY: Pantheon Books.

Chomsky N (1986) *Knowledge of Language: Its Nature, Origin and Use*. New York, NY: Praeger.

Cowie F (1999) *What's Within: Nativism Reconsidered*. New York, NY: Oxford University Press.

Crain S and Thornton R (1998) *Investigations in Universal Grammar: A Guide to Experiments in the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.

Elman JL, Bates E, Johnson MH, Karmiloff-Smith A, Parisi D and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Hornstein N and Lightfoot D (eds) (1981) *Explanations in Linguistics: The Logical Problem of Language Acquisition*, pp. 9–31. London, UK: Longman.

Lightfoot DW (1991) *How to Set Parameters: Arguments from Language Change*. Cambridge, MA: MIT Press.

Mind and Language (1998) vol. 13. Oxford, UK: Blackwell Publishers Ltd.

Pinker S (1994) *The Language Instinct*. London, UK: Lane.

Smith N and Tsimpli I-M (1995) *The Mind of a Savant*. London, UK: Blackwell.

Intonation

Introductory article

Sun-Ah Jun, UCLA, Los Angeles, California, USA

CONTENTS

Intonation is the melody of a sentence or a phrase. It marks a grouping of words or the prominence relations among words, and reflects the syntactic structure and the semantic and pragmatic meaning of a sentence.

WHAT IS INTONATION?

Traditionally, intonation has been defined as the melody of a sentence or a phrase. A melody is, in a general sense, a pattern of pitch changes. Some parts of a sentence are produced with higher or lower pitches than other parts of a sentence, and the overall pitch pattern of a sentence delivers the speaker's intention as well as certain aspects of the meaning of the sentence or sentence type. However, intonation involves more than just the global changes in pitch over the course of a sentence. The overall pitch pattern has an internal structure. Intonation marks groupings of words and defines a hierarchy for these groupings. This means that, while the acoustic realization of pitch is continuous, intonation is not in fact an undividable whole – instead it is decomposable into subcomponent pitch events. Some pitch events mark the boundaries between groupings of words, either small boundaries or large, and others mark the prominence relations within a group of words.

The boundary of a group of words, marked by pitch patterns called boundary tones, is sometimes also marked by duration. For example, syllables at the end of a group of words are often substantially lengthened, and this is known as preboundary lengthening. Further, prominent words that are marked by pitch are also often characterized by higher intensity and longer duration. In this way, pitch patterns carry the tune or melody of a sentence or a phrase and, together with intensity and duration, they also convey the prominence relations among the words in a phrase. Thus, intonation, whether it marks a phrase boundary or prominence relations, is comprised not only of pitch changes but also of other prosodic features such as intensity and duration. For this reason, the terms 'intonation' and 'prosody' are often used interchangeably. In sum, intonation has an internal structure, and the structure is cued by pitch, intensity, and duration. (*See* **Prosody**)

Intonation can change the meaning of a sentence but not the meaning of a word. In this sense, it is different from (lexical) pitch accent or tone, both of which can change the meaning of a word. In a pitch accent language, such as Japanese or Swedish, each word has a distinctive pitch pattern. That is, words with different pitch patterns do not have the same meaning, even when the words have the same sequence of consonants and vowels. For example, in Japanese, 'kami' means 'God' when high pitch occurs on the syllable 'ka', but the word means 'paper' when high pitch occurs on the syllable 'mi'. Similarly, in a tone language such as Mandarin, each syllable of a word has a distinctive pitch pattern. For example, one syllable 'ma' can mean 'mother' when it is produced with a high pitch, 'horse' when it is produced with a low rising pitch, 'hemp' when it is produced with a high rising pitch, and 'scold' when it is produced with a falling pitch. In languages such as these, where pitch is distinctive at the word level, pitch variations at the phrasal level, i.e. intonation, are not as easily observable as they are in nontonal and/or nonpitch accent languages such as English, French, or Korean. Still, all languages have intonation, and intonation conveys the structure and meaning of a sentence as well as the speaker's intention.

Though there are certain features of intonation that are similar across languages, many attributes of the intonation patterns found within a given language are determined on a language-specific basis. For example, in most languages pitch rises at the end of a yes–no question and falls at the end of a statement. However, in Chickasaw, an American Indian language of the Western Muskogean

language family spoken in south-central Oklahoma, the reverse pattern is found. That is, pitch falls at the end of a yes–no question and rises at the end of a statement. Languages also differ in terms of sentence-medial pitch patterns. This depends on the intonational groupings within a sentence and the types of pitch movement permitted within a word in the language. For instance, in English, the stressed syllable of a word can be realized with several different pitch patterns (e.g. high, mid, low, rising), and the edge of an intonational grouping can be marked by either a high or a low pitch. On the other hand, in Seoul Korean, pitch patterns are not linked to specific syllables within a word. Instead, they are linked to certain locations within a phrase. The phrase initial syllable can be either high or low, but the phrase final syllable is almost always high. This suggests that the pitch patterns of sentences within a given language reflect the prosodic system of that language, i.e. the intonational structure of that language and whether the language has stress, tone, pitch accent, or none of these.

Finally, the pitch categories that appear in descriptions of intonation, such as high tone or low tone, are not absolute but are in fact relative. They are relative to adjacent pitch values within the sentence or phrase and are relative to the speaker's pitch range. In a given speaker's production, a syllable can be said to have a high tone intonationally if its pitch is higher than that of the immediately preceding syllable within the same prosodic phrase or if the pitch is realized at the top of the speaker's current pitch range. On the other hand, the interpretation of pitch categories in intonation is not gradient but categorical. The categorical nature of intonation is determined by the categorical nature of its component pitch events. If a statement ending in a low pitch is interpreted as a declarative, the absolute lowness of the pitch at the end of the sentence does not influence the degree of 'declarativeness' of that statement. It simply is or is not declarative. This contrasts with the paralinguistic features of speech which convey a speaker's emotional state (e.g. angry, fearful, joyous) and his or her attitude towards the listener (e.g. friendly, appeasing, aggressive). The interpretation of paralinguistic features is gradient. So, for instance, the higher the absolute pitch or intensity in a word or a phrase, the angrier or more aggressive the speaker is likely to be. In this article, we will discuss the categorical properties of intonation, and in the following sections we will show the role of intonation in grammar, especially with respect to its role in prosodic grouping, its relation to syntactic structure, and its semantic and pragmatic functions. (*See* **Emotion**)

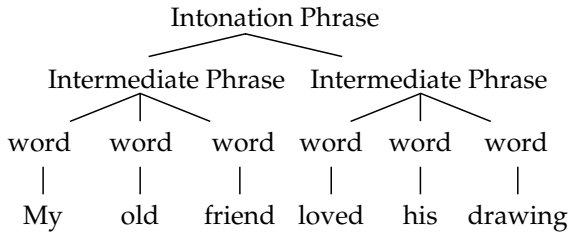
PROSODIC CONSTITUENCY

When we produce a sentence, we tend to form subgroups of words within the sentence. This grouping reflects how close adjacent words are in terms of their meaning and function relative to one another. The degree of juncture or distance between each word is not always equal. Words belonging to the same group are closer together, i.e. have smaller juncture between them, than words belonging to different groups which have a larger juncture between them. The degree of juncture between words and groups of words is abstract and subjective, but studies have found that native speakers of the same language generally agree on the degree of juncture between the words in a given utterance. This is because the abstract entity termed 'juncture' reflects the organizational structure of the spoken utterance and is realized by consistent acoustic features such as pauses, preboundary lengthening, and other intonational attributes. (*See* **Prosody**)

The grouping of words within an utterance is called prosodic grouping or phrasing. There is typically more than one level of grouping within an utterance, reflecting the different degrees of juncture. That is, there are subgroupings within various groups within an utterance. For example, a sentence such as 'The lady with flowers loves John's brother' could be divided into two large groups, (The lady with flowers) and (loves John's brother), but the degree of juncture among the four words in the first group is not equal. The juncture between the first two words, 'the lady', or the next two words, 'with flowers', is smaller than the juncture between the second and the third word, 'lady with'. This shows that a whole utterance can be divided into groups which in turn can be divided into smaller groups, and so on. The hierarchical structure of groupings within an utterance is called the prosodic hierarchy, and each group in the prosodic hierarchy is called a prosodic constituent.

In English, investigators disagree on exactly how many levels of prosodic constituents there are above the word. We will not discuss the different proposals or the differences among the proposed levels here except to say that there are in general two prosodic levels above the word. The higher of the two is often called an intonation phrase and the lower one is called an intermediate phrase (or a

phonological phrase), and these are marked by intonation. The prosodic grouping for an example sentence is shown in (1). The whole sentence forms one intonation phrase, and is further divided into two intermediate phrases. Each intermediate phrase includes three words. This is one of many possible phrasings for this sentence.



(1)

Words can also be subdivided into prosodic constituents such as the foot (a sequence of strong and weak syllables), the syllable, and the mora. However, in this article we will focus on the prosodic constituents above the word level, since prosodic constituents below the word level are not directly related to intonation.

Experiments have shown that prosodic constituency and the hierarchical structure are psychologically real. This means that these abstract entities are part of native speakers' tacit knowledge about their language. Phonetic studies on speech production have shown that prosodic constituents are phonetically cued so as to mark the prosodic hierarchy. For example, the duration of a single sound segment is longer in the initial position of a higher prosodic phrase than it is in the initial position of a lower prosodic phrase, which in turn is longer than that in the initial position of a word occurring somewhere in the middle of a lower prosodic phrase. The same pattern is found for the duration of phrase final sounds. That is, the higher the prosodic constituent is in the prosodic hierarchy, the greater the degree of lengthening is at the edges of this prosodic constituent. (*See* **Phonetics**)

Psycholinguistic experiments have also shown that native speakers of a language are sensitive to the boundaries of prosodic constituents when they are processing sentences of their language. For example, when English speakers listen to sentences such as the one in (2) below, the processing time, i.e. the time taken to understand the meaning of the sentence, is shorter if the prosodic phrasing (marked by acoustic features such as preboundary lengthening, boundary tones, and pause) matches the syntactic or semantic grouping of the sentence (see next section for more detail). Thus, English speakers process (2a) more quickly than (2b) because there is a match between the prosodic phrasing and the syntactic/semantic grouping in (2a), while there is a mismatch between the prosodic phrasing and syntactic/semantic grouping in (2b). (*See* **Sentence Processing: Mechanisms**; **Sentence Processing**)

Sentence: When George left the house, it was dark.
a. expected prosodic phrasing: (When George left the house) (it was dark)
b. unexpected prosodic phrasing: (When George left) (the house it was dark) (2)

In sum, utterances are prosodically divided into large and small phrases, and these prosodic constituents are hierarchically structured. Furthermore, native speakers phonetically cue the prosodic hierarchy in speech production by manipulating pitch, duration, and intensity, and they are also able to perceive these cues when processing a sentence.

RELATION TO SYNTACTIC STRUCTURE

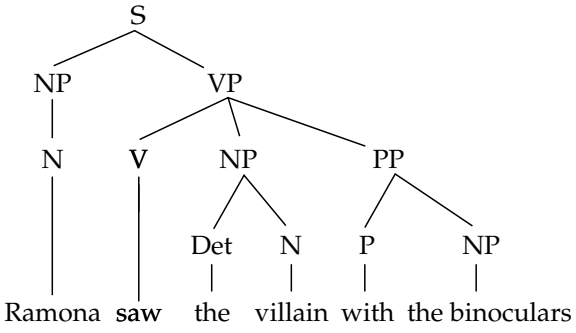
We have seen that sentences, when produced, are divided into intonational phrasings, or prosodic groupings, marked by intonation. (Thus, the term 'prosodic phrase' is used interchangeably with 'intonational phrase' below.) However, a sentence can be phrased in various ways depending on multiple factors such as syntax, the location of a focused word within the sentence, speech rate, and/or the length of the words and phrases. Therefore, prosodic phrasing does not always match the syntactic phrase of a sentence but is heavily influenced by syntax. In this section, we will show how the intonational phrasing is influenced by the syntactic structure of the sentence. (*See* **Syntax**)

Though there are many ways to phrase a sentence prosodically, prosodic phrase boundaries cannot be placed at random but are constrained by syntactic constituents. Specifically, there are certain syntactic constituents that tend to trigger phrase boundaries and others that do not. For example, speakers generally put a prosodic phrase boundary at a sentence-internal clause boundary (e.g. the sentence 'When you are tired, it's better to rest' is phrased as '(When you are tired) (it's better to rest)'), at the boundary of a parenthetical phrase (e.g. 'Disneyland is, Jane said, the most popular place to visit in LA' is phrased as '(Disneyland is) (Jane said) (the most popular place to visit

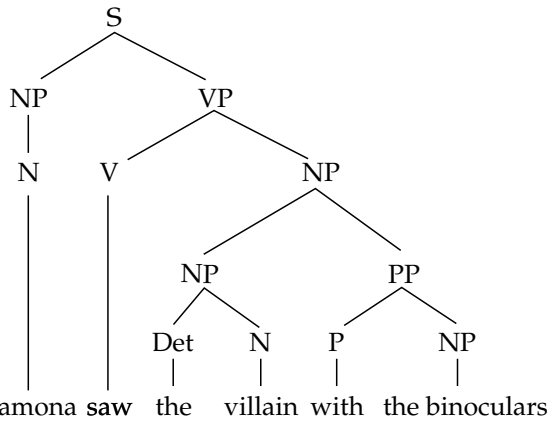
in LA)'), and before a tag question (e.g. 'He will win, won't he?' is phrased as '(He will win)(won't he)'). Similarly, speakers generally put a prosodic boundary between two words if the two words belong to two different syntactic phrases unless each syntactic phrase has only one word (e.g. between subject noun phrase and verb phrase).

On the other hand, speakers do not in general put a prosodic phrase boundary between words if they belong to the same syntactic phrase, unless the phrase is too long or each word is emphasized or produced very slowly. For example, it is very unlikely that a determiner and the following noun (e.g. 'a book', 'the teacher'), or an adjective and a noun (e.g. 'pretty girl', 'last year'), or a possessive pronoun and a noun (e.g. 'my book') would belong to two different prosodic phrases. For the same reason, it is unlikely for speakers not to put a prosodic boundary between two words if each belongs to a different syntactic phrase and the second syntactic phrase has other component words. For example, for the sentence 'The child with asthma outgrew the condition last year', combining 'asthma' and 'outgrew' together in the same prosodic phrase while putting a phrase boundary after 'outgrew' is not likely (i.e. *(The child) (with asthma outgrew) (the condition) (last year)). In sum, this suggests that prosodic phrasing reflects the syntactic grouping of the words.

Psycholinguistic and phonetic studies have shown that speakers disambiguate a syntactically ambiguous sentence by producing different prosodic phrasings for each syntactic structure. For example, the sentence 'Ramona saw the villain with the binoculars' can mean two things. Namely, Ramona could have used the binoculars to see the man, or Ramona could have seen the man who possessed the binoculars. For the first meaning the prepositional phrase 'with the binoculars' modifies the verb 'saw', as shown in (3a), and for the second meaning the prepositional phrase modifies the preceding noun phrase, 'the villain', as in (3b).



(3a)



(3b)

It has been found that speakers may produce the sentence differently depending upon which of the two syntactic structures is intended. For the structure in (3a), a prosodic phrase boundary is placed between 'the villain' and 'with the binoculars', i.e. (Ramona saw the villain) (with the binoculars), while for the structure in (3b), no such boundary is produced. Figure 1a shows a pitch track for the utterance with the first meaning, i.e. (3a), and Figure 1b shows a pitch track for the utterance with the second meaning, i.e. (3b). The horizontal axes in the figures show the time in milliseconds and the vertical axes show the pitch in hertz (Hz). A higher value in Hz indicates higher pitch. Above the pitch tracks, waveforms are presented which indicate the amplitude (vertical axes) over time (horizontal axes). Each word of the sentence is written over the waveforms, and a vertical line marks the end of each word or phrase. (The same format is used in Figure 2.) The word 'villain' in Figure 1a is longer than the same word in Figure 1b. This is because the word in Figure 1a is the final word in the prosodic phrase, but it is not in Figure 1b. The word in Figure 1a also shows a falling and slightly rising pitch, indicating a phrase boundary after the word.

Psycholinguistic studies have also found that native speakers perceive these disambiguating intonational and timing cues in ambiguous sentences and that these acoustic cues help native speakers process the sentences to arrive at the intended meanings. This shows that syntax imposes some constraints on the production of prosodic structure and that the prosodic structure is parsed or accessed by native speakers together with the syntactic structure when processing a sentence.

In sum, intonation can help cue the syntactic structures of a sentence, although not all syntactic structures are cued by intonation. Additionally, intonation plays a further role in conveying the

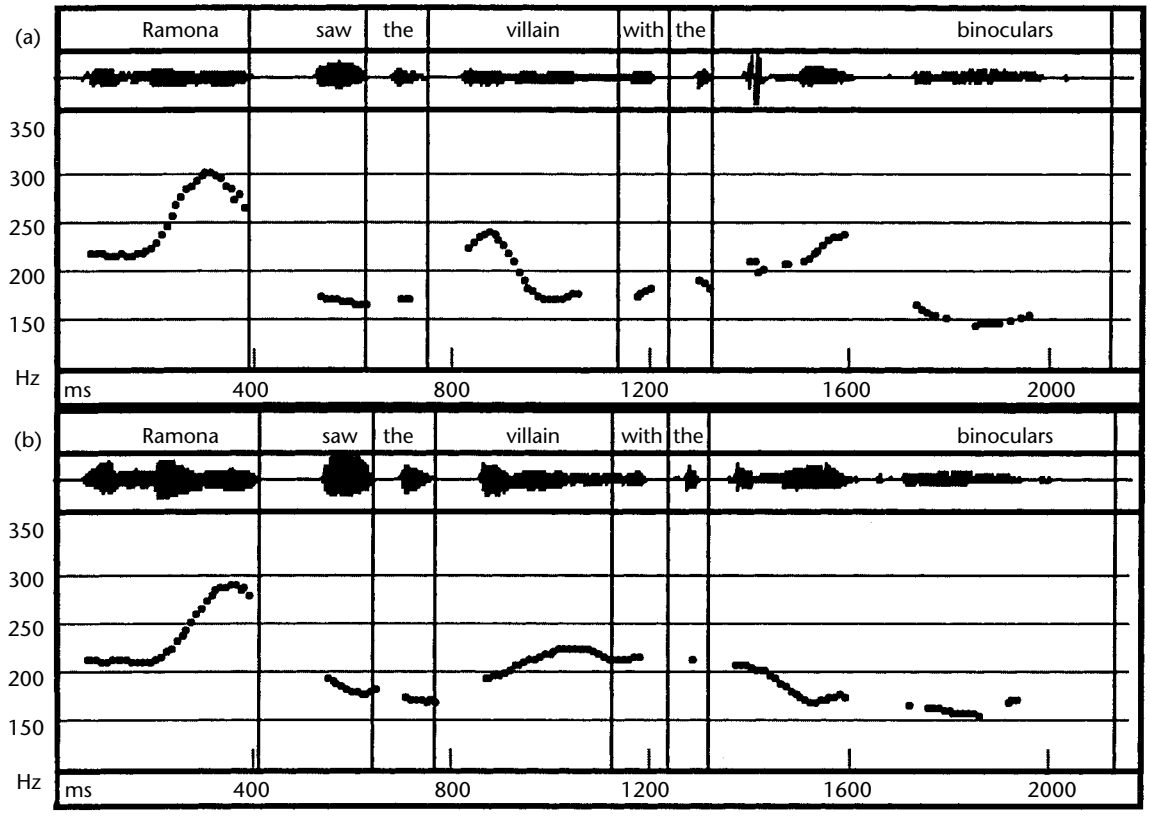


Figure 1. (a) A pitch track of a sentence with (3a) structure, and (b) the same sentence with (3b) structure.

semantic and pragmatic meaning of a sentence, and this will be discussed in the next section.

SEMANTIC AND PRAGMATIC FUNCTIONS

It is well known that native speakers emphasize or focus an important word within a sentence by means of pitch, intensity, and duration. For example, in English a focused word is realized with a sharply rising pitch and has longer duration and higher intensity than a nonfocused word. Listeners interpret the focused word by comparing it with other items belonging to the same category as the focused word. For instance, the sentence 'John only introduced Bill to Mary' can mean different things depending upon which word is focused. If 'Bill' is focused ('John only introduced BILL to Mary'), the sentence is true only when John introduced exactly one person to Mary and that person was Bill. On the other hand, if 'Mary' is focused ('John only introduced Bill to MARY'), the sentence is true only when John introduced Bill to exactly one person and that person was Mary. Thus, it can be seen that focus intonation changes the truth conditions of the sentence, suggesting that the semantic interpretation of a sentence is influenced by intonation. (*See* **Semantics and Cognition**; **Pragmatics, Formal**)

Intonation can also influence the relationship between words within a sentence. For example, the sentence 'Mary didn't approach them because they were yelling' can have two meanings depending upon how the sentence is produced. It can mean either that Mary did not approach them and the reason was because they were yelling, or it can mean that Mary approached them not because they were yelling but for some other reason. For the former meaning, 'not' modifies the verb 'approach', while for the latter meaning, 'not' modifies the word 'because'. Example pitch tracks for this sentence conveying these two different meanings are presented in Figure 2. (These are not the only possible intonation patterns that would convey these two meanings.) The utterance in Figure 2a conveys the former meaning and that in Figure 2b conveys the latter meaning. There is a phrase boundary after 'them' in Figure 2a, but not in Figure 2b. (Note the lengthening on 'them' in Figure 2a.) Further, the sentence ends with a falling boundary tone in Figure 2a, but with a rising boundary tone in Figure 2b. These

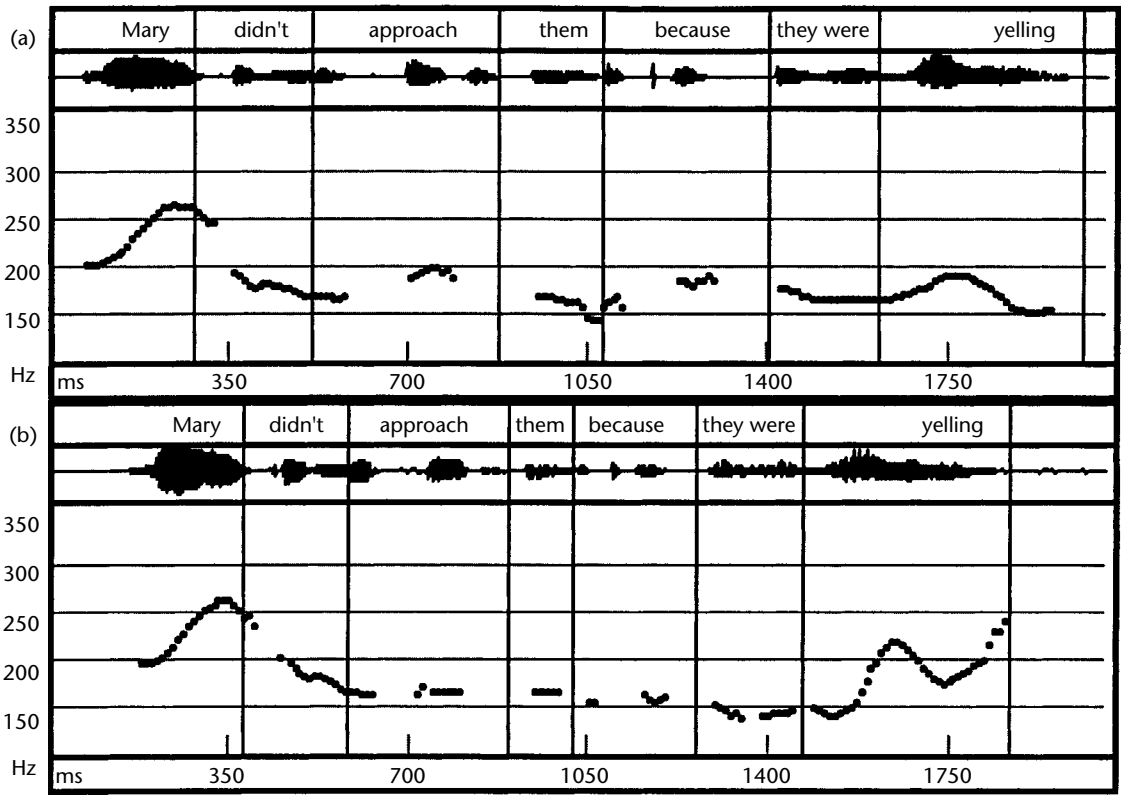


Figure 2. A pitch track of the sentence meaning (a) 'Mary did not approach them and the reason was because they were yelling' and (b) 'Mary approached them not because they were yelling, but for some other reason'.

examples illustrate that intonation can change the modification relations between words.

Furthermore, intonation reflects the structure of a discourse and the relative informativeness of words or constituents in a sentence (e.g. new versus old information). Words or constituents of a new topic or new information become prominent by pitch range manipulation or by getting pitch accent. In order to cue a discourse structure, a pitch range is manipulated in such a way that sentences with a new topic are produced with an expanded pitch range while those with an old topic (or a nonmajor topic as in parenthetical phrase) is produced in a reduced pitch range. That is, the pitch peak is higher at the beginning of a paragraph introducing a new topic than in the middle of the paragraph, and the pitch valley is lower at the end of the paragraph finishing a major topic than in the middle of the same paragraph.

Similarly, in English, a word with new information is produced with pitch accent, and a word that is repeated or 'given' in the discourse does not get pitch accent. When the sentence 'I don't read German' is produced with broad focus, the sentence final noun, i.e. 'German', would receive pitch accent. However, when this word is repeated and becomes old information, it does not get any accent, as in (4B). Here, the verb 'read' is accented and the object noun 'German' is deaccented. Deaccenting is also found in word(s) after contrastive focus. For example, when the word 'John' is contrastively focused (for example, against 'I') as in (5), it will get pitch accent and the rest of the sentence will get deaccented. Thus, accentedness reflects the relative semantic weight and the informativeness within a constituent.

A: I found an article for you in a German journal.
B: I don't READ German. (4)

JOHN reads German. (5)

This relationship of semantic weight versus accent, however, is not universal. In Italian and Romanian, for example, a word with given information still receives accent, and in French and Korean, which do not have lexical stress, new

versus given information is delivered by phrasing, not by accenting. A word with new information begins a new prosodic phrase, while a word with old information is dephrased, i.e. does not begin a new phrase but belongs to the preceding phrase. Thus, beginning a new phrase in French and Korean is equal to receiving pitch accent in English.

To conclude, we have seen how intonation is related to prosodic structure, syntactic structure, and the semantic and pragmatic meaning of a sentence. Intonation marks a structure as well as prominence relations among words. It also conveys semantic and pragmatic meaning. It is related to all subareas of grammar and is psychologically real. Thus, intonation is a crucial part of a native speaker's knowledge of his or her grammar, and it is also essential for effective communication between the members of a speech community.

Further Reading

Beckman M (1996) The parsing of prosody. *Language and Cognitive Processes* **11**(1/2): 17–67.

Beckman ME and Pierrehumbert J (1986) Intonational structure in Japanese and English. *Phonology Yearbook* **3**: 255–309.

Cutler A, Dahan D and Donselaar W. van (1997) Prosody in the comprehension of spoken language: a literature review. *Language and Speech* **40**: 141–201.

Ferreira F (1993) The creation of prosody during sentence production. *Psychological Review* **100**: 233–253.

Gee TP and Grosjean F (1983) Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology* **15**: 411–458.

Jun S-A (1996) *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*. New York: Garland Publishing.

Jun S-A and Fougeron C (2000) A phonological model of French intonation. In: Botinis A (ed.) *Intonation: Analysis, Modeling and Technology*, pp. 209–242. Dordrecht: Kluwer Academic Publishers.

Kjelgaard MM and Speer SR (1999) Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language* **40**: 153–194.

Ladd DR (1996) *Intonational Phonology*. Cambridge, UK: Cambridge University Press.

Nespor M and Vogel I (1986) *Prosodic Phonology*. Dordrecht: Foris.

Price P, Ostendorf M, Shattuck-Hufnagel S and Fong C (1991) The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America* **90**(6): 2956–2970.

Rooth M (1996) Focus. In: Lappin S (ed.) *Handbook of Contemporary Semantic Theory*. London: Blackwell.

Selkirk EO (1984) *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA : MIT Press.

Shattuck-Hufnagel S and Turk A (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* **25**(2): 193–247.

Ward G and Hirschberg J (1985) Implicating uncertainty: the pragmatics of fall-rise intonation. *Language* **61**(4): 747–776.

Language Acquisition and Language Change

Intermediate article

David W Lightfoot, Georgetown University, Washington, DC, USA

CONTENTS

Under a cognitive view of language, a mental system emerges in the mind/brain of an individual child; linguists study the acquisition of that system, a 'grammar', which characterizes the individual's mature linguistic capacity. If a different system develops in another child in the next generation of speakers, then we have grammatical change, and that change reflects differences in the acquisition process.

INTRODUCTION

Language acquisition and change are intimately related: ideas about acquisition influence models of change, and vice versa. For example, children are sometimes viewed as input matchers: they acquire a grammar, which generates the set of primary data to which they are exposed in the first years of life, matching the input. This view of acquisition was promulgated in Chomsky (1965) and persists in recent work (Gibson and Wexler, 1994; Clark and Roberts, 1993). It entails stability and one would not expect languages or grammars to change except through major disruptions such as invasions or transfer of populations. Indeed, Keenan (1994) and Longobardi (2001) have argued for an 'inertia', under which languages do not change structurally. On the other hand, if children are not input matchers but converge on a grammar by identifying abstract 'cues', then minor shifts in experience may entail identifying a different distribution of cues, hence the development of a different grammar (Dresher, 1999; Lightfoot, 1999). This view leads one to expect to find catastrophic changes when a new grammar emerges, changing substantially the kinds of expressions found in historical texts.

The explanatory schema adopted by generative grammarians who study natural language with a view to learning about human cognition is shown in (1), where (1a) has general biological terminology and (1b) the corresponding linguistic terms.

a. triggering experience (genotype → phenotype)
b. Primary Linguistic Data (Universal Grammar → grammar) (1)

A child's triggering experience causes the genotype to develop into a phenotype; exposure to a range of utterances in, say, French allows the genetically prescribed Universal Grammar (UG) to develop into a particular mature grammar. A child develops a grammar by setting the open parameters of UG in the light of her particular experience.

CHANGE: SOCIAL AND COGNITIVE

Changes take place in two ways. First, there are changes in the primary linguistic data, some of which have no effect on the grammars which emerge. No two children have identical triggering experiences; children do not hear the same set of expressions in the same order and in the same context, but they may none the less converge on the same abstract system. Consequently, variation or change in the primary linguistic data does not necessarily entail a different grammar. One may study variation and change at this level ('change' is just variation taking place over time) and it may be minor, piecemeal, chaotic, and gradual. This variation, what one may call 'language change', is influenced by social factors and may spread through population groups in the course of time. Briscoe (2000) offers a computer simulation of such variation and the way it may affect acquisition.

Second, change may also take place in grammars, resulting from prior changes in the primary linguistic data, and here particular theories of UG become relevant. If one works with a theory of UG with a significant degree of abstraction, a single change at the grammatical level may proliferate through the system and have 'catastrophic' effects, that is, simultaneous change in several superficially

unrelated phenomena. Grammatical change is often bumpy: the set of expressions occurring in texts change significantly in a short period. By examining the clusters of simultaneous changes and by taking them to be related by properties of UG, we discover something about the nature of grammars; the time and conditions of grammatical change may illuminate the relevant trigger experiences. In this perspective, work on grammar change informs work on variation and acquisition: the set of parameters postulated as part of UG explains the unity of the changes, why superficially unrelated properties cluster in the way that they do.

AN EXAMPLE OF CHANGE

English modal auxiliaries such as 'can', 'could', 'may', 'might', 'will', 'would', 'shall', 'should', and 'must' differ from verbs in their distribution. A modal auxiliary is fronted in a question, but a verb like 'understand' is not, as in (2); a modal occurs to the left of a negative particle, unlike a verb (3); a modal does not occur with a perfective (4) or progressive (5) marker, unlike a verb; a modal does not occur in the infinitival complement to another verb (6), nor as the complement of another modal (7); and no modal occurs with a direct object, unlike some verbs (8). (Asterisks indicate non-occurring forms.)

a. Can he understand chapter 4?
b. *Understands he chapter 4? (2)

a. He cannot understand chapter 4.
b. *He understands not chapter 4. (3)

a. *He has could understand chapter 4.
b. He has understood chapter 4. (4)

a. *Canning understand chapter 4, . . .
b. Understanding chapter 4, . . . (5)

a. He wanted to understand.
b. *He wanted to can understand. (6)

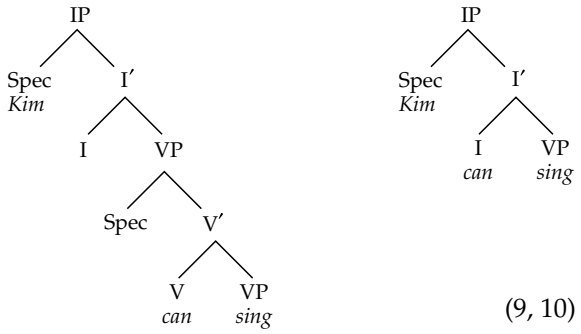
a. He will understand.
b. *He will can understand (7)

a. *He can music.
b. He understands music. (8)

The distribution of these modal auxiliaries is peculiar to Modern English. Sentences along the lines of the nonexistent utterances of (2)–(8) were well formed in earlier English. If the differences between Old and Modern English were a function of separate features with no unifying factor, we would expect these features to come into the language at different times and in different ways. On the other hand, if the difference between Old and Modern English reflected a single property, a categorical distinction, then we would expect the trajectory of the change to be different. If the differences between 'can' and 'understand' in (2)–(8) were a function of the single fact that 'understand' is a verb while 'can' is a member of a different category, Inflection (I), then we are not surprised to find that: (2b); (3b); (4a); (5a); (6b); (7b); and (8a) dropped out of people's language in parallel, at the same time.

If we attend just to changing phenomena, the change consists of the *loss* of various forms, not the development of new forms; people ceased to say things which had been said in earlier times. Before the change, all of the utterances in (2)–(8) might have occurred in a person's speech, but later only those forms not marked with an asterisk. That fact alone suggests that there was a change in some abstract system. People might start to use some new expression because of the social demands of fashion or because of the influence of speakers from a different community, but people do not cease to say things for that sort of reason. Changes involving only the loss and obsolescence of forms need to be explained as a consequence of a change in an abstract cognitive system.

If one focuses on the disappearance of the relevant forms, one sees that they were lost at the same time. The most conservative writer in this regard was Sir Thomas More, writing in the early sixteenth century. He used many of the asterisked forms in (2)–(8) and had the last attested uses of several constructions. His grammar treated 'can', etc. as verbs in the old fashion (see (9)), and the fact that he used *all* the relevant forms, and his heirs none, suggests that his grammar differed from theirs in one way, not that the grammars differed in unrelated features. The uniformity of the change suggests uniformity in the analysis. There was a single change, a change in category membership: 'can', etc., formerly verbs which moved to I in the course of a derivation, came to be analyzed as I elements, as in (10). The fact that there was a single change in grammars accounts for the bumpiness: several phenomena changed simultaneously. The notion of change in a grammar is a way of unifying disparate phenomena, taking them to be surface manifestations of a single change at the abstract level. The change in category membership of the English modals explains the catastrophic nature of change, in that phenomena changed together but not in the sense that the change spread through the population rapidly.



(9, 10)

Linguists have also studied the spread of changes through populations. Anthony Kroch and his associates (e.g. Kroch, 1989) have done interesting work on the replacement of one grammar by another, enriching grammatical analyses by describing the variability of individual texts and the spread of grammatical change through a population. Niyogi and Berwick (1997) have offered a dynamical systems model for describing the spread of new grammars; a key idea is that a new grammar acquired by one speaker entails new primary linguistic data for the community, which in turn is more likely to trigger the new grammar in other children. Yang (2002) also has an interesting model in which grammars may coexist, with one eventually winning out.

The cause of the grammatical change in English was two earlier changes in primary linguistic data (here we discuss a result of these earlier changes; *their* cause is another story). First, the modal auxiliaries became distinct morphologically, the sole surviving members of the preterite-present class of verbs. There were many verb classes in early English and the antecedents of the modern modals were preterite-presents. The preterite-presents were distinct in that they never had any inflection for the third person singular, although they were inflected elsewhere: 'þu cannst', 'we cunnan', 'we cuþon'. They did not change in this regard, but a great mass of inflectional distinctions disappeared and, as a result, the preterite-presents were isolated; they looked different from other verbs in lacking their one morphological feature, that *-s* ending. That property, insignificant when there were many classes of verb inflection, took on a new importance when present tense inflection consisted of only one property.

Second, the morphological distinctiveness of the surviving preterite-presents, the new modals, was complemented by a new opacity in their past tense forms. The past tense forms of the preterite-present verbs were phonetically identical in many instances to the subjunctive forms and, when the subjunctive forms were lost (part of the general loss of inflections just alluded to), past tense forms survived with subjunctive-type meanings rather than indicating time reference. While 'loved' is related to 'love' in terms of time reference in present-day English, the relationship between 'can' and 'could' sometimes has nothing to do with time; 'might' is never related to 'may' in terms of time in Modern English. So 'might', 'could', 'should', etc. came to take on new meanings which had nothing to do with past time, residues of the old subjunctive uses.

The preterite-present verbs came to look different from all other verbs in the language. UG provides a small inventory of grammatical categories, and elements are assigned to a category on the basis of their morphological and distributional properties. Morphological changes entail new primary linguistic data, which may trigger new category distinctions. In this case, we know that, following the morphological changes, the surviving verbs of the preterite-present class were assigned to a new grammatical category, and that change was complete by the early sixteenth century.

There were two stages to the history of English modal auxiliaries (Lightfoot, 1999, chap. 6). First came a change in category membership, whereby 'can', etc. ceased to be treated as verbs and came to be taken as manifestations of the Inflection category; this change affected some verbs before others, but it was complete by the sixteenth century. Second, English lost the operation moving verbs to a higher Inflection position (e.g. in (9)). This change was completed only in the eighteenth century. At this point, sentences with a finite verb which had moved to some initial position (2b) or to the left of a negative (3b) became obsolete and were replaced by equivalent forms with the periphrastic 'do': 'Does Kim understand this chapter?' 'Kim does not understand this chapter', etc.

Gradual, piecemeal changes in primary linguistic data (such as loss of verbal inflections) sometimes cross a threshold such that they entail changes in people's grammars (e.g. category membership), part of their cognitive make-up. Such grammatical changes need to be explained through the mechanisms of acquisition and consequently they may illuminate the properties of Universal Grammar, including the nature of grammatical categories. Taking grammars to be elements of cognition has been productive for work on language change and helps to inform theories of grammar and of acquisition.

References

Briscoe E (2000) Grammatical acquisition: inductive bias and co-evolution of language and the language acquisition device. *Language* **76**(2): 245–96.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clark R and Roberts IG (1993) A computational model of language learnability and language change. *Linguistic Inquiry* **24**: 299–345.

Dresher BE (1999) Charting the learning path: cues to parameter setting. *Linguistic Inquiry* **30**: 27–67.

Gibson E and Wexler K (1994) Triggers. *Linguistic Inquiry* **25**: 407–54.

Keenan E (1994) Creating anaphors: an historical study of the English reflexive pronouns. MS, University of California, Los Angeles.

Kroch A (1989) Reflexes of grammar in patterns of language change. *Journal of Language Variation and Change* **1**: 199–244.

Lightfoot DW (1999) *The development of language: Acquisition, change, and evolution*. Oxford, UK: Blackwell.

Longobardi G (2001) Formal syntax, diachronic minimalism and etymology: The history of French *chez*. *Linguistic Inquiry* **32**(2): 275–302.

Niyogi P and Berwick R (1997) A dynamical systems model for language change. *Complex Systems* **11**: 161–204.

Yang CD (2002) Grammar competition and language change. In: Lightfoot DW (ed.) *Syntactic Effects of Morphological Change*. Oxford, UK: Oxford University Press.

Further Reading

DeGraff M (1999) *Language Creation and Language Change: Creolization, Diachrony and Development*. Cambridge, MA: MIT Press.

Lass R (1997) *Historical linguistics and language change*. Cambridge, UK: Cambridge University Press.

Lightfoot DW (1979) *Principles of Diachronic Syntax*. Cambridge, UK: Cambridge University Press.

Paul H (1880) *Prinzipien der Sprachgeschichte*. Tübingen: Niemeyer.

Roberts IG (1993) *Verbs and Diachronic Syntax*. Dordrecht: Kluwer.

Warner A (1993) *English Auxiliaries: Structure and History*. Cambridge, UK: Cambridge University Press.

Language Acquisition by Animals

Intermediate article

Duane M Rumbaugh, Georgia State University, Atlanta, Georgia, USA
Michael J Beran, Georgia State University, Atlanta, Georgia, USA

CONTENTS

Language acquisition by animals is concerned with how the acquisition of linguistic skills by nonhuman animals compares to language acquisition in humans. Although language acquisition by nonhuman animals is not claimed to be identical, or even approximate, to language acquisition by humans, researchers in this area are concerned with the specific cognitive skills of nonhuman animals that are necessary in the productive use and understanding of symbols.

INTRODUCTION

Animal language acquisition research has produced important data and, at times, led to highly contentious debate. The importance of animal language research is more than just to better understand how human language may have evolved. It is also important to demonstrate exactly what language does, and does not, involve. Language research with nonhuman animals allows us to identify processes in humans that may derive from general rather than specialized cognitive structures. Additionally, animal language research sheds important light on the role of the environment, the role of early experience, and the impact of cultural variables on cognitive factors influencing the exchange of information. This article outlines briefly the recent history of attempts to 'speak' with the animals, some of the more important issues in animal language research, and the major findings in this research area.

APE LANGUAGE ACQUISITION

Early Attempts

The earliest recorded attempt to teach an ape language involved training an orangutan to produce speech sounds. Furness (1916) observed an orangutan that acquired a number of different speech sounds that were produced voluntarily and apparently in appropriate contexts. The attempt to teach language to apes was the continued focus of research by both Kellogg and Kellogg (1933) and Hayes and Hayes (1951). The Kelloggs observed a chimpanzee named Gua. The chimpanzee was raised with the Kelloggs' own son, named Donald, for the first nine months (from the time the chimpanzee was 7 months to 16 months), and Gua proved the more capable student in the realm of locomotor behavior and navigation. However, although both the chimpanzee and the child began responding to spoken requests appropriately at around 12–14 months of age, the human soon exceeded the chimpanzee in comprehension of these spoken requests. When Hayes and Hayes began their research with a chimpanzee, the emphasis was returned to speech production, but the chimpanzee, Vicki, also provided much insight into the psychology of the ape mind. Vicki was given formal speech training, in which she was required to produce voluntary sounds in order to be rewarded. Eventually, Vicki produced a small number of different speech sounds that sounded like the words 'Papa', 'cup', 'up', and 'Ma Ma'. However, there was no evidence that Vicki used these words symbolically to refer to those items.

The next major scientific studies examining language-related skills in apes began in the 1960s and 1970s (Gardner and Gardner, 1969; Patterson and Linden, 1981; Premack and Premack, 1983; Rumbaugh, 1977; Terrace, 1979), and as the field evolved from that time period, its emphasis and methods changed considerably. Initially, the emphasis in ape-language research during the 1960s

and 1970s was on production of ordered sequences of symbols as well as on vocabulary acquisition and the meaningfulness of symbols, reflecting the issues of productivity and rule-generation emphasized by Chomsky in the then-emerging field of linguistics. The chimpanzees Washoe, Sarah, and Lana were conditioned to produce single symbols or strings of symbols through extensive training, and their combinations of the symbols in novel ways was examined.

Washoe was taught American Sign Language by the Gardners. Because chimpanzees had anatomical limitations that prevented production of the sounds of human speech, the Gardners decided that speech production was not a useful method for examining language in chimpanzees. Thus, only signs were used around Washoe, and signing was a part of every daily activity. Additionally, Washoe was trained to make and use various signs, and by three years of age she had become proficient in making many signs. This research program was continued by Fouts with Washoe and other chimpanzees. Additionally, Patterson studied sign language with gorillas, and Miles studied sign language acquisition by an orangutan. All of these studies confirmed that apes learned to use signs more readily than they had been able to produce speech.

Premack's work with a chimpanzee, Sarah, was with a different symbol system. Plastic tokens were used as representations for various items, actions, and concepts. Sarah was presented with various 'problems' that she had to solve through recourse to her symbol system. Rumbaugh also began working with a female chimpanzee, Lana, using special symbols called lexigrams. These lexigrams were geometric symbols representing English words or phrases, and they were located on a large computer-monitored keyboard in Lana's housing area. Lana communicated with her human caretakers and they communicated with her exclusively through use of this computerized system. This computer was important because it helped to standardize the interactions with Lana and to control for many of the cuing problems that had been evident in earlier projects with apes. Lana was taught to string symbols together to form various stock sentences (e.g. 'Please machine make window open'), and she came to exhibit flexible and appropriate use of her symbols in novel situations.

These early ape-language studies produced data indicative of rudimentary language or language-like skills in these apes. The initial conclusion made by those involved in the research was that the chimpanzees had learned many skills associated with language and that the language use of these chimpanzees was similar to that of human children.

Associative Language versus Representational Language Learning in Apes

Terrace had conducted research with a chimpanzee, Nim, in an attempt to replicate the work of the Gardners. However, Terrace analyzed the signing of Nim and concluded that Nim's utterances were simply imitations of the sign language utterances of his human caretakers. Analyses of the signing of Washoe produced the same conclusion: Washoe was simply imitating the signs of those around her. Lana's sentences also were attacked by Terrace and his associates who claimed that her system was not syntactic, but rather was the result of a series of chained associative responses that could be interchanged at various points. Additionally, it was claimed that Lana had learned only to associate certain things with certain lexigrams, and this associative language did not lead to true understanding of the meanings of lexigrams. However, Lana was able to complete sentences started by others with appropriate lexigrams, and she even 'erased' sentence beginnings that were intentionally produced in incorrect sequences by an experimenter. This indicated that Lana was not simply performing a chained sequence of associative responses. Later reports from Rumbaugh's team also showed that Lana only infrequently imitated her human companions' lexigram strings, and her productions were not to be accounted for by an appeal to her memory of recent experimenter-initiated utterances of lexigram symbols.

As issues of competence became apparent, ape language researchers placed emphasis on understanding whether the 'words' apes used really contained the same set of underlying representations for them as they did for humans. Psychologists came to recognize a clear difference between associative naming, such as that learned by Lana, and representational language, which indicates a true understanding of what words mean. Child language studies in the 1970s and 1980s were guided by the 'semantic revolution'. Emphasis was placed on acquisition of single words and the meaning of symbols that precedes grammatical competence. 'Reference' (the notion that symbols such as lexigrams, manual signs, or printed words represented real-world items, actions, and ideas) became an important concept in the nonhuman language

debate. Production was still the primary focus of inquiry. However, the tide was turning towards recognition of comprehension as a vital part of language acquisition.

At the Language Research Center in Atlanta, Georgia, the learning environment of a second generation of apes, Sherman and Austin, was substantially richer than that of Lana in that they experienced a greater variety of daily activities in a more open environment (e.g. they were allowed to go for walks outdoors), and their communications were directed to humans or to each other rather than exclusively to a computer (Savage-Rumbaugh, 1986). Human caretakers communicated with Sherman and Austin using the lexigram keyboard as well as through spoken English statements and various context-specific gestures. Sherman and Austin were trained to use lexigrams under a paradigm that relied on discrete-trial operant conditioning. Training sessions were distinct from other activities, and symbols were, initially, restricted to the training context. The use of spoken English was of only secondary importance in the interactions of the caretakers and the chimpanzees during these training sessions, as demonstrating comprehension of spoken English by the chimpanzees was not the goal of the project. There was no indication that these chimpanzees understood spoken English. Savage-Rumbaugh later determined through work with bonobos that exposure to human speech at an early age was critical in the subsequent comprehension of human speech by apes.

The processes by which the apes learned were different in many ways from those experienced by children; in fact, children learn more than these apes did without any formal training at all. This left the work open to criticism that there was no homology between what the apes were learning and children's early language skills. However, through their training, Sherman and Austin did master referential functions that are considered an important part of symbolic competence, allowing at least an argument of analogy to be made with regard to some aspects of human language competency. The two chimpanzees made statements about their future actions and then carried out those actions, they requested items from each other so as to accomplish tasks, and they responded correctly to each other's appropriate requests whereas they refused to respond to inappropriate requests. None of these behaviors were evident in other language-trained apes. Sherman and Austin (but not Lana) demonstrated metalinguistic ability when they categorized lexigrams into functional categories. The apes were first taught to place items into either the category 'food' or the category 'tool'. The apes then learned to categorize photographs of those same items into the two categories, and finally the lexigrams for those items were categorized correctly. Most importantly, when novel lexigrams (i.e. lexigrams that the apes had never before categorized but had learned as symbols for real-world items) were presented, Sherman and Austin correctly categorized those lexigrams as 'foods' or 'tools' on the very first trial. This clearly demonstrated that these lexigrams functioned as symbols for the apes, and that these lexigrams had semantic meaning for them. Another way of demonstrating that these symbols had meaning for Sherman and Austin was through showing that they could perform symbol-based cross-modal matching. When presented with the lexigram for an object, the chimpanzees could reach into a box and find the object which that lexigram represented without ever having seen the object itself. As in the categorization study, the lexigrams represented things that were not present, and this is the essence of semantics: it is word meaning.

Language Learning, Speech Comprehension, and Early Rearing

In the early 1980s, the developing knowledge base relevant to ape and child language radically changed the approach to teaching language skills to apes as well as to the evaluations of what the apes were capable of learning (Savage-Rumbaugh and Lewin, 1994). Savage-Rumbaugh began work with an adult female bonobo (*Pan paniscus*) in an attempt to replicate the findings with Sherman and Austin. During these training sessions, the bonobo's son, Kanzi, remained in the area but was not taught to use the lexigrams or even to attend to them (he was thought too young to learn the symbols). The adult female proved unsuccessful at learning the symbols, and after some time she was separated from Kanzi so that she could breed. At this time, it was discovered that the infant had learned not only to appropriately use and respond to others' use of lexigrams but also to comprehend English speech. Apparently, being immersed in the daily routines of lexigram use and speech use by the humans around him, Kanzi had gleaned the meanings of these symbols and spoken words. This occurred with no training, but through observation of a structured series of daily situations.

After the young bonobo Kanzi surprised his caregivers with untrained symbol use, subsequent apes participating in language studies have been

immersed in a language-structured environment or culture throughout much of their infancies – much as human children are. The apes in these studies were not trained in formal repeated-trial training sessions. Rather, they participated in daily life and observed the behavior and communications of others as they did so. For example, the apes traveled through a forest each day during which time there was much discussion by human caretakers (through both speech and lexigram use) about what was going to occur or what had occurred already. These structured interactions and routines provided a framework from which the apes learned the meanings of lexigrams and large numbers of spoken words. And most importantly, this lexigram use and speech by caretakers was important to the apes as it informed them about events and items of interest to them.

The initial route to language competence in these apes was through comprehension of speech and lexigram use starting very early in infancy, and this generalized to production of symbols without explicit training by one to two years of age. These studies coincided with improvements in measuring early comprehension in children and the realization of its importance to natural language acquisition. Also, at this time there emerged a new emphasis on the process and context of language acquisition as a social phenomenon, with examination of the social support systems that facilitate communicative competence in children. These apes came not only to understand the meanings of lexigrams, but also to comprehend spoken English in a variety of contexts. For example, strict experimental testing showed that the bonobo Kanzi comprehended spoken requests and statements in highly controlled situations with hundreds of novel sentences. This comprehension included a sensitivity to word order, such that Kanzi responded appropriately to both 'Pour the Coke in the lemonade' and 'Pour the lemonade in the Coke' (Savage-Rumbaugh *et al*., 1993). Kanzi's productive competence was comparable to that of an 18-month old human child, and analyses of Kanzi's use of lexigrams indicated that he used grammatical rules of his own invention and rules modeled after those of his human caretakers (Greenfield and Savage-Rumbaugh, 1991). For example, Kanzi always used lexigrams before gestures within the context of combining an agent gesture with an action lexigram. When Kanzi produced two-element utterances with action lexigrams, regularities that reflected preferred action orders in social play were evident (Rumbaugh and Savage-Rumbaugh, 1994).

Three of the four apes who have been reared in this way and who comprehended spoken English have been bonobos. The one chimpanzee, Panzee, who also was raised in this enriched environment, demonstrated that *Pan troglodytes* is capable of learning to understand spoken English as well as use a large number of lexigrams communicatively without explicit training during the first few years of life. This chimpanzee and her bonobo companions have learned to use lexigrams to fulfil many communicative functions, to understand another's communicative intent when that other individual touched lexigrams, and to understand spoken English sentences and words as well. At age three, Panzee was able to select the correct lexigram when presented with 79 different spoken English words, and her comprehension of English today stands at more than 120 words for which she has lexigrams. Today, these apes have a functional understanding of even more spoken English words than are represented with lexigrams, as the lexigram keyboard has trailed behind the spoken word comprehension vocabularies of these apes as they are exposed to new objects, foods, people, and activities.

PARROT LANGUAGE ACQUISITION

At nearly the same time that Savage-Rumbaugh was investigating Kanzi's language competence, Pepperberg was training the African gray parrot, Alex, using what she called the 'model–rival' approach (Pepperberg, 1999). Alex observed linguistic interactions between two people, who served alternatively as 'models' and 'rivals' for Alex. The model/rivals talked about and exchanged objects that they hoped would interest Alex. When Alex entered the conversation through emitting vocalizations, he was rewarded by having the humans respond as if he 'meant' what he had said. For example, if he had made a request, it was honored. Alex could observe and enter the conversation at will, and he eventually learned to ask for, choose, and describe the color, shape, and materials of the objects with which he was familiar. As with the animals at the Language Research Center, it was assumed that Alex really did want whichever item he requested. Further, Alex had the opportunity to observe the way in which his human model/rivals interacted with each other. This ideal situation in which to learn relationships was important in the language acquisition of Alex, and it resembles the contextual way in which human infants learn human language both by interacting with parents and by observing parents and others interacting

with each other. Note also that these contexts are designed so that Alex can respond spontaneously. A broad range of different responses is available, rather than the single responses that have for so long been the standard in the typical classical or operant conditioning experiment.

From this approach, it was found that the parrot learned many names for items, and he used those names appropriately in a variety of contexts. Additionally, Alex learned to respond to the spoken requests of his caretakers. Alex could be queried on the names of items, the quantity of items in an array, the color of items, and even the substance of which an item was made. Even complex questions of multi-item, multicolor, multisubstance arrays could be answered correctly, indicating that Alex understood the entire sentence and could decompose the request into relevant subcomponents so that the correct category and answer within that category could be determined. To do this, the parrot must not only understand word meanings but also perform tasks involving recursive processing in many instances. Alex can interpret spoken requests and statements and respond appropriately based on conceptual understanding of those requests. Thus, linguistic abilities in nonhuman animals are not limited to apes. The research with Alex further demonstrates that rearing and teaching methods are important to the competencies that emerge in nonhuman animals.

DOLPHIN LANGUAGE ACQUISITION

Language acquisition research with dolphins centers around the work of Herman who taught two dolphins two different artificial language systems (Herman *et al*., 1984). One dolphin, Akeakamai (Ake), was taught to respond to the hand and arm gestures of humans, whereas the other dolphin, Phoenix, was taught computer-generated vocal sounds. The words in each of these artificial language systems referred to objects, actions, properties, and relationships. These words could be combined into meaningful sentences that the dolphins could interpret. Some sentences were relatively simple and required an action to be done to an object; other sentences, however, were more complex and involved relations between words within the sentence as well as multiple relevant item words. Using the same words, multiple sentences with multiple meanings could be constructed, and this required that the dolphins learn not only the semantic meaning of the symbols, but also the syntactic rules governing the combinations of those symbols.

Herman believed that comprehension of this language system was a more important route to investigate than production (as was used in the early ape-language studies) because he claimed that investigations of productive use of language in apes had revealed no convincing evidence for the understanding or use of syntax (although, as already noted, Rumbaugh believed that Lana did show evidence of grammatical understanding). Additionally, Herman (and Savage-Rumbaugh) noted that language comprehension precedes language production in human children, and thus is a more likely place to start an investigation of language in nonhuman animals. Therefore, Herman was interested in the ways in which his dolphins responded to different types of linguistic statements by humans.

In addition to responding appropriately to commands in this linguistic situation, the dolphins also answered questions pertaining to the status of named objects (such as HOOP QUESTION to ask whether a hoop was present in the tank). Herman conducted numerous studies of this language system with the dolphins, including investigations into how Ake would respond to anomalous gestural sequences (the dolphin discriminated anomalous sequences from normal sequences and rejected the anomalous ones or completed only the part of the gesture sequence that 'made sense'). Additionally, the dolphin responded appropriately to gestures that were videotaped and even substantially degraded in the clarity of the image. In both symbol systems, the dolphins demonstrated an understanding of both the syntactic rules and the semantic content of the symbols in their system; they responded appropriately not only to the meaning of individual gestures and signals, but also to the relation of those gestures and signals to each other within a sentence. Syntactic understanding was shown through responses to semantic contrasts in sentences in which word order was reversed, to syntactically anomalous sentences, to novel sentences, to sentences in which one word modified the meaning of another, to interrogative sentences, and to sentences with variations in the placement of modifiers.

Despite the data indicating language comprehension by these dolphins, there is a lack of any productive language data. Unlike with the work with bonobos, in which both comprehension and production of lexigram use, as well as comprehension of spoken English (including for syntactic structure) was investigated and demonstrated, the work with dolphins has remained tied to the realm of comprehension. Therefore, little is

known about what the dolphins would 'say' if they had the means to produce linguistic output that was comprehensible to humans (dolphins use a seemingly complex communicative system of whistles among conspecifics, and this points even more to a need to look to productive language use in the dolphin).

THE FUTURE OF ANIMAL LANGUAGE RESEARCH

With the advances made since the late 1970s, nonhuman animal language studies have become established within the fields of biology, psychology, and anthropology. Nonhuman animal language acquisition is now an established phenomenon that has been documented in numerous species and through numerous symbol systems. That said, what is (and should be) the future of this research area?

First, it is necessary to expand the scope of such studies to include other species that have the requisite skills and abilities needed for this kind of research. These species may include elephants, whales, dogs, and even other nonhuman primate species. However, we believe the most interesting and informative data will continue to come from the great apes. In particular, recent studies show that the chimpanzee brain and the human brain are much more closely related in morphology than previously thought (Gannon *et al*., 1998), and it will be important to look for homology in brain processes during language-related activities to understand further the specific evolutionary course of language-like skills in our human ancestors. With the advent of brain imaging technologies such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), it will be important to determine what role the chimpanzee brain plays in comprehension of human speech, in symbol use, and in species-specific vocal communicative utterances. We also hope such studies can be conducted with other species and their respective symbol systems or natural communicative capacities.

Additionally, we believe it will be important to define further the role of the environment in the language acquisition of nonhuman animals. In particular with chimpanzees and bonobos, we must examine the extent to which language competence in these animals affects, and is affected by, the cultural variables that make up the environment in which these animals live and are raised. One cannot learn language in a vacuum, at least not in the sense of using it for its intended purpose of communication. Language is necessary only within the social and cultural context in which it is learned and employed. Therefore, studies of language acquisition and use are intricately tied to studies of culture and social interactions with nonhuman animals.

Finally, we would note that language acquisition is an example of an *emergent* (Rumbaugh *et al*., 1996). Emergents are forms of silent learning that, in many cases, are acquired through social observation. They are not reinforced via training regimens, but are established through induction by the organism, and they are noted for their appropriateness to novel situations. We believe that the language acquisition of Kanzi (and the other apes raised in a manner similar to Kanzi) demonstrates perhaps the clearest example of this form of learning. Further research into nonhuman animal language will continue to provide data that force us to reconsider our views of early rearing, the role of culture, the uniqueness of human language, and the principles of learning as they are manifest in intelligent nonhuman animals.

References

Furness W (1916) Observations on the mentality of chimpanzees and orangutans. *Proceedings of the American Philosophical Society* **45**: 281–290.

Gannon PJ, Holloway RL, Broadfield DC and Braun AR (1998) Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog. *Science* **279**: 220–222.

Gardner RA and Gardner BT (1969) Teaching sign language to a chimpanzee. *Science* **165**: 664–672.

Greenfield PM and Savage-Rumbaugh ES (1991) Imitation, grammatical development, and the invention of a protogrammar by an ape. In: Krasnegor NA, Rumbaugh DM, Schiefelbusch RL and Studdert-Kennedy M (eds) *Biological and Behavioral Determinants of Language Development*, pp. 235–258. Hillsdale, NJ: Lawrence Erlbaum.

Hayes KJ and Hayes C (1951) The intellectual development of a home-raised chimpanzee. *Proceedings of the American Philosophical Society* **95**: 105–109.

Herman LM, Richards DG and Wolz JP (1984) Comprehension of sentences by bottlenosed dolphins. *Cognition* **16**: 129–219.

Kellogg WN and Kellogg LA (1933) *The Ape and the Child*. New York, NY: McGraw-Hill.

Patterson FL and Linden E (1981) *The Education of Koko*. New York, NY: Holt, Rinehart, & Winston.

Pepperberg IM (1999) *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Cambridge, MA: Harvard University Press.

Premack D and Premack AJ (1983) *The Mind of an Ape*. New York, NY: Norton.

Rumbaugh DM (1977) *Language Learning by a Chimpanzee: The LANA Project*. New York, NY: Academic Press.
Rumbaugh DM and Savage-Rumbaugh ES (1994) Language in a comparative perspective. In: Mackintosh NJ (ed.) *Animal Learning and Cognition*, pp. 307–333. San Diego, CA: Academic Press.
Rumbaugh DM, Washburn DA and Hillix WA (1996) Respondents, operants, and emergents: toward an integrated perspective on behavior. In: Pribram K and King J (eds) *Learning as a Self-organizing Process*, pp. 57–73. Hillsdale, NJ: Lawrence Erlbaum.
Savage-Rumbaugh ES (1986) *Ape Language: From Conditioned Response to Symbol*. New York, NY: Columbia University Press.
Savage-Rumbaugh ES and Lewin R (1994) *Kanzi: The Ape at the Brink of the Human Mind*. New York, NY: John Wiley.
Savage-Rumbaugh ES, Murphy J, Sevcik RA *et al*. (1993) Language comprehension in ape and child. *Monographs of the Society for Research in Child Development* **1**: 1–221.
Terrace HS (1979) *Nim*. New York, NY: Knopf.

Further Reading

Brakke KE and Savage-Rumbaugh ES (1995) The development of language skills in bonobo and chimpanzee – I. Comprehension. *Language and Communication* **15**: 121–148.
Brakke KE and Savage-Rumbaugh ES (1996) The development of language skills in Pan – II. Production. *Language and Communication* **16**: 361–380.
Herman LM (1988) The language of animal language research: reply to Schusterman and Gisiner. *Psychological Record* **38**: 349–362.
Parker ST and Gibson KR (1990) *'Language' and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*. Cambridge, UK: Cambridge University Press.
Pepperberg IM (1990) Cognition in an African Gray parrot (*Psittacus erithacus*): further evidence for comprehension of categories and labels. *Journal of Comparative Psychology* **104**: 41–52.
Roitblat HL, Herman LM and Nachtigall PE (1993) *Language and Communication: Comparative Perspectives*. Hillsdale, NJ: Lawrence Erlbaum.
Rumbaugh DM and Gill TV (1976) The mastery of language-type skills by the chimpanzee (Pan). *Annals of the New York Academy of Sciences* **280**: 562–578.
Savage-Rumbaugh ES (1991) Language learning in the bonobo: how and why they learn. In: Krasnegor NA, Rumbaugh DM, Schiefelbusch RL and Studdert-Kennedy M (eds) *Biological and Behavioral Determinants of Language Development*, pp. 209–233. Hillsdale, NJ: Lawrence Erlbaum.
Savage-Rumbaugh ES, McDonald K, Sevcik RA, Hopkins WD and Rubert E (1986) Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*). *Journal of Experimental Psychology: General* **115**: 211–235.
Savage-Rumbaugh ES, Rumbaugh DM, Smith ST and Lawson J (1980) *Reference: The Linguistic Essential. Science* **210**: 922–924.
Schusterman RJ and Gisiner R (1988) Artificial language comprehension in dolphins and sea lions: the essential cognitive skills. *Psychological Record* **38**: 311–348.
Seidenberg MS and Petitto LA (1979) Signing behavior in apes: a critical review. *Cognition* **7**: 177–215.
Terrace HS, Petitto LA, Sanders RJ and Bever TG (1979) Can an ape create a sentence? *Science* **206**: 891–900.

Language Acquisition

Introductory article

Michael P Maratsos, University of Minnesota, Minneapolis, Minnesota, USA

CONTENTS

Language acquisition comprises the developmental processes by which children develop into competent users of words and sentences in their native language, to express themselves, and to influence and cooperate with others.

INTRODUCTION

Children do not need to be taught language; they formulate it themselves. In school, children and adolescents are taught the underlying rules and principles of complex systems, and given examples to flesh out the general rules. In learning a language, the child only hears the examples – particular utterances used by others – and is expected to deduce the underlying rules and principles independently.

BEHAVIORIST VIEWS OF ACQUISITION

A half-century ago it was largely taken for granted that in acquiring language, children showed no distinctive talent. Rather, it was thought that learning a language, like learning everything else, was a relatively simple process of experience 'stamping in' specific behaviors. Skinnerian psychologists believed that children spontaneously give off sound productions. Those that are appropriate are rewarded, by mechanisms such as parental approval, and are thus shaped and strengthened by experience. Inappropriate responses are suppressed by mechanisms such as parental disapproval. Learning theorists who emphasized imitation added that children could acquire responses by imitation, without contingent reward or punishment.

Inadequacies of Behaviorist Theories

Behaviorist views proved inadequate on a number of grounds, some empirical, some theoretical. Empirically, it was shown that in most known cultures, parents do not approve and disapprove on the basis of how well-formed sentences are. A child's sentence like 'her curl my hair', though ill-formed, would be approved of if it were factually accurate, disapproved of if were factually inaccurate, in most cultures that have been studied (although in one culture, the Kaluli of New Guinea, adults do in fact overtly disapprove of poorly formed sentences). Also, many children simply do not imitate very much in general, at least overtly.

Theoretically, the behaviorist view of language as a set of particular memorized or previously reinforced responses is wrong. A competent speaker of a language can and does produce new utterances that the speaker has never heard or given forth before, that are nevertheless appropriate and communicative. That is because in acquiring a language, a speaker formulates a set of rules and principles that underly the particular utterances that are heard. These rules and principles can then be used to generate new appropriate sequences of words and morphemes from these underlying rules and principles.

What is meant by this? To illustrate with a non-linguistic example, suppose one heard and learned a set of number sequences: 1–2–4, 8–16–32, 121–242–484. A reasonably intelligent person, having learned these particular sequences, would soon figure out a more general underlying pattern or rule that generated the sequences: any first number x is followed by $2x$, which is then followed by 2 times $2x$. One could then generate new sequences that follow the same underlying patterns, e.g. 3–6–12, 14–28–56, 3000–6000–12 000, and so on.

As simple as this case is, what is done here is indescribable in strict behaviorist terms. One has not just acquired a set of imitated (or reinforced) particular responses. One has gone beyond this to generate a set of underlying mental rules, which can then generate new examples. Notions such as

'rule' or 'underlying mental rule' were unacceptable in behaviorism, which forbade theorization of mental entities like mental rules, feelings, plans, goals or other things inside the organism. In fact, we might better say the problem-solver above formulated a set of rules from the input, rather than 'learned the rules', which implies that experience presented the rules directly, which it does not – experience only provided data for the formulation of the rules.

Acquiring a natural language, of course, does require a good deal of acquisition of specific knowledge, some of it basically imitative, such as acquiring the sound sequences of particular words. However, it also requires the formulation of a heterogeneous set of underlying rules and principles of many kinds. Experience can supply particular data points, examples of speech, for doing this, but it cannot directly supply the underlying rules and principles themselves. Even if parents did approve well-formed utterances and disapprove badly formed ones, this would only tell a child about the validity of specific examples; it could not supply the underlying system for the child.

COGNITIVIST AND NATIVIST VIEWS OF LANGUAGE ACQUISITION

Noam Chomsky, a linguist, used these and other arguments about the nature of language and its acquisition in an influential set of criticisms of the behaviorist view. In general, in the 1960s, most of those who believed humans formulate underlying rules and principles, like Jean Piaget, believed that they construct these structures by the use of a general, intelligent set of constructive and inductive mechanisms which could be used to analyze causality, number and space, as well as language. We can call this a 'cognitivist' approach.

Chomsky, however, proposed a far more radical view. He believed that the underlying structures and rules of natural languages, especially those of syntax (sentence structure) are too particular in nature, and too far removed from the input data, to be derived by a general cognitive device. Rather, he proposed that to acquire language, children must know a good deal of its specifically linguistic nature in advance: they must have considerable innate knowledge. In Chomsky's view, we should try to postulate mechanisms that make language acquisition more like instantaneous perception, rather than slow, cumulative acquisition and construction. This innate knowledge would not be particular knowledge of particular languages. Rather, it would be like a child who innately knew the general principles of architecture, who was to analyze the structure of particular buildings encountered later. Chomsky's hypothesis was both radical and skilfully argued, and set the central issues for language acquisition in the decades to come. Indeed, nearly all work in the field, even if not strongly relevant to his hypothesis, is referred to the basic nativist–antinativist controversy.

In reality, most of the work in language acquisition is not strictly relevant to Chomsky's particular claims, though investigators generally try to make references to issues of nativism and nonnativism. Whatever the explanation, empirical work does show surprisingly competent acquisition of complex language systems and knowledge, even if some of this complexity in fact has a strong non-Chomskyan cast. While acquisition of a language covers many domains, the central areas remain phonology (sound systems), semantics (word meaning) and grammar (sentence structure and structurally realized meaning). The accumulated empirical findings of some decades, from many languages, resist adequate summary; a selective survey of key findings and arguments is given instead.

EMPIRICAL FINDINGS AND KEY ARGUMENTS

The First Year and Phonology

Phonological acquisition means acquiring the meaningfully distinguished set of sounds used in one's language (ranging from about 30 to 60 in various languages), the sets of rules for how these sounds can be combined into words, intoned in phrases, varied across related forms of words, and other related acquisitions. Currently the best-known results have arisen in studies of infants. Generally, we can say that infant babbling creates many different sounds. Then, around the age of 1 year, a smaller set of sounds is used to make initial words. This initial set expands at different rates in different children, with highly variable rates of error in pronunciation.

More exciting than the study of what children produce, however, has been the experimental study of how they perceive sounds. At birth, children can discriminate – apparently – all or virtually all the sound distinctions employed by the world's languages. At first it was believed these innate auditory abilities might be specific to humans acquiring language, but it was found that minks, chinchillas and monkeys can discriminate the same sounds (at least when tested).

Even at birth, from womb experience, children prefer to hear characteristic sound sequences of their own language. Over the next months, further language-specific adjustments take place. For example, the /o/ of English and Swedish is somewhat differently centered in sound. By 6 months, infants have a properly centered set of preferences for the 'o-center' of their own language. Rhesus monkeys do not seem to revise their basic sound perception system similarly with similar experience, so this sound-center learning ability might be specific to humans. Infants also lose the ability to discriminate sounds not characteristic of, or not distinguished meaningfully, in their own languages. For example, English-language infants can discriminate various sounds used in Hindi at 6–8 months, but lose this ability by 10–12 months of age, probably because the relevant sound differences do not make meaningful differences in English although they do in Hindi.

Perhaps most surprising are infants' abilities to analyze sequences of sounds. By the age of 6 months, infants prefer sequences containing whole-phrase intonation patterns, over recordings of broken-up phrase patterns chained together. Ten-month-old infants hearing new words read to them in a passage can differentially recognize these new words in a later passage. Surprisingly, if one plays to them sound sequences in which some sound transition sequences are more likely (e.g. /r/ followed by /a/ followed by /l/) than others, it appears that they learn something about these probabilistic transitions, an ability thought relevant to isolating word-sequence patterns in the sound input. This is startlingly well-developed sequential analytic ability; in fact, it is learning from experience, the analysis of large amounts of data to find statistical sequential probabilities.

Word Meaning Acquisition

Before first words

The first recognizable word appears around the age of 10–12 months. This initial utterance rests on much previous preparation; by the time it appears, an average child probably comprehends some meaning in around 50 words. Some evidence indicates that children can manually sign words earlier, so it is likely that the difficulty of reproducing sounds itself presents an obstacle. Also, it is likely that children gather information about the use of a word for some time before using it, rather than quickly venturing its use.

Fast mapping

Many competences probably feed into word meaning: the understanding that others have minds and intentions, and the ability to tell where the attention of others falls. Beyond this, there is the problem that a word is just a sound sequence used in a situation, and in any situation there are many possible meanings to assign. Yet once 50 words have entered the spoken vocabulary, acquisition proceeds at the rate of six or seven words a day. Furthermore, on the whole, children's word meanings are accurate, though some errors of use occur.

One theoretical response to this combination of quickness and accuracy has been to posit various 'fast mapping' strategies and heuristics children might use to deduce a word meaning – or a great deal of it – quickly from a single exposure. For example, it has been proposed children are biased to guess that a new word refers to an unnamed object, that each object has only one name, and that words refer to categories rather than individuals. Furthermore, it is proposed that there are natural 'basic categories' for children to guess as the meaning of an object-word, a category defined roughly by the highest distinctive shape common to groups of objects: that is, certain categories of objects, like chairs, cups and birds, fall together naturally in the mind, and so their meanings are easily assigned to words. Indeed, across languages, word meanings for basic object-word categories tend to remain constant, which supports a natural conceptual-semantic analysis.

Slow mapping

So far, it is fast-acting strategies for basic object-words that have been most successfully proposed. Other kinds of meanings often vary across language communities, so a 'fast guess' good in one community would be a poor guess in another. For example, suppose a child hears a new word, *gib*, while someone is eating a tortilla, and guesses that it refers to the action. We might think it natural for the child to encode this in the action-category English speakers encode with 'eat', meaning to 'ingest solid food'. Yet in Tzetal, there is one verb for eating very solid (crunchy) foods, another verb for eating softer solid foods (like fruit), and another for eating tortilla and bread-like foods. So if the child guessed that 'gib' meant the same as the English 'eat', the child would apply it too broadly, and make many errors. In Korean or Navajo, as another illustration, various verbs of handling or moving objects bundle meaning differently from English. One Korean verb, for example, applies to

removing objects from a tightly contained or supported place, while another applies to removing objects from loose containment or support. Another commonly used verb refers to placing an object long in dimension (long or tall) on a support surface. Surely the child learning Korean would not luckily guess just these right meanings immediately. Yet young children's uses of these Tzetal and Korean terms are accurate. Such acquisition requires, it seems, 'slow mapping' – that is, the child must collect detailed information across time on the situations for which a word is used, and from this information analyze the correct situational elements. 'Slow mapping' is a form of learning from experience.

Furthermore, the meanings discussed above are all relatively concrete objects or actions. Still other mechanisms must obtain for the learning of internal emotion words, or words referring to past, present, future, possible, and many other kinds of nonconcrete meanings. Word meanings cover the entire human conceptual sphere, and no single set of procedures can be expected to be adequate for all.

The Acquisition of Grammar: Morphology

Grammar (competence in sentence structure) is usually divided into two major parts: syntax and morphology. Morphology relates to the composition of words from morphemes, defined as the smallest sound sequences that have independent meaning. In English, morphology is used to express notions such as number on nouns ('dog+s') or tense on verbs ('push+ed') or possession on persons ('John's', 'the boy's'). Morphology is used in many languages to express not just these meanings, but also notions expressed by syntax in English. In Turkish, for example, any of 'John-u helped Ann', 'Ann helped John-u', 'helped John-u Ann' and so on constitute an acceptable way of saying, 'Ann helped John.' The accusative marker '-u' on 'John' marks John as the direct object of 'help', the one helped, and thus Ann as the helper. In morphologically rich languages like Turkish, morphological combinations in a single word may represent the first grammatical utterances.

The 'wug' test and productive grammatical knowledge

Morphological knowledge has most simply shown the productive, nonimitative nature of finished grammatical competence. For example, a preschool child is shown a picture of a man doing something novel, and is told, 'This man is nissing.' The experimenter continues, 'Yesterday the man did the same thing. Yesterday he …', leaving a pause in which the child can fill in 'nissed'. 'Nissed' follows from the general English morphological pattern of adding '-ed' to verbs to express past tense. It could not have been previously learned by imitation or reinforcement, because 'niss' has only just been introduced. Only the speaker's previous formulation of general rules, including the appropriate formulation of categories such as verb, makes it available. English-speaking children also produce overregularized past forms like 'runned' and 'breaked', not heard from parents, which show their analysis of a general 'ed-past' pattern.

Morphological acquisition

Children's skill in acquiring complicated morphological rules provides one of the most striking findings of current acquisitional work. For example, in Turkish, the accusative case marks direct objects such as 'John' in English 'Ann helped John', or 'the dog' in 'the cat chased the dog.' Yet only definite direct objects are so marked; indefinites corresponding to English 'a cat' or 'something' are not so marked. Furthermore, there are four accusative markers, rounded and unrounded '-oo' and '-ee' (rounded versus nonrounded lips during pronunciation), which are used according to the back-front and rounded-unrounded qualities of the stem vowel of the noun. None of this, of course, is told to the child, who just hears markers used on nouns sometimes and not others. One might expect long-drawn-out, error-filled acquisition of this complex set of contingencies. Yet instead, competence in production and listening are generally complete by the age of 2 years, with no errors of applying a marker wrongly. Acquisitional findings from other languages offer similarly impressive early, error-free pictures.

Slow mapping again

It should be obvious here that a child could not just hear such a marker used once and luckily guess just the right rule for its use. The child must store a great deal of information about the use and nonuse of each of these morphological markers, across many speech situations, to ferret out the right controlling properties. Since some hundreds of properties may affect morphological marking in different languages, sifting through hundreds of such candidate properties appears necessary to achieve the right analysis. This kind of hugely impressive grammatical acquisition, like nothing else currently known in general cognitive development,

supports some kind of special acquisitional systems for language. Nevertheless, like much discussed above, it does not correspond to the 'quick perception' mode of Chomskyan nativism, even if some kind of special adaptations for language learning may be present.

Syntactic Acquisition and Constraints

Syntax is the appropriate ordering of whole words to say what one intends. 'Helped John Mary' is simply not a possible syntactic ordering of English. 'John helped Ann' is possible, but is wrong if one intended to say Mary was the agent (i.e. helper). Syntax is partly controlled through meaning, and partly through nonmeaningful formal properties of word categories. In syntax-dominated languages such as English, children's initial grammatical combinations appear to be controlled by meanings such as agency, possession, location, negation and others. For example, an English-speaking child would say 'mommy go' (actor followed by action) or 'daddy chair' (possessor before possessed object).

While there is a fairly good map of how syntactic constructions are added over time, their underlying descriptions remain controversial. For simple-looking early speech, for example, theoretical descriptions range from highly specific word-centred descriptions, to general semantic-structural formulas, to essentially adult-like semantic-formal rules which are described as being reduced in production by children's memory restrictions. Indeed, the most striking aspects of normal syntactic acquisition lie in certain complex constructions which are rarely actually found in children's speech or even in that of adults, yet have central places in linguists' theoretical analyses, and play central roles in Chomskyan nativism. The major paradox lies in the interaction between the highly general rules that languages use in conjunction with highly specific, abstract constraints that restrict their generality. For example, in English, a highly general pattern is that where there is a form in which a wh-expression is part of a sentence, like 'You will talk to who(m)?' there is a corresponding form in which the wh-expression is placed at the front of a clause, e.g. 'Whom will you talk to —?'. Various evidence indicates there is indeed a gap understood where the wh-expression would 'normally' go.

The initial wh-expression may be very far from its 'gap', which may be hierarchically embedded as well, e.g. 'Which theory does John say that Mary claims that the current government will force you to talk about—?' The simplest rule for the input children hear would be that a wh-expression can leave a gap anywhere later in the sentence. Yet this is not true. One cannot say, corresponding to 'John really admires doctors who can cure *which diseases*?' the wh-question *'*Which diseases* does John really admire doctors who can cure—?'

Similarly, a noun expression like 'John' and a coreferential pronoun like 'he' or 'his' can appear in many different relative positions in sentences, e.g. the pronoun can follow the noun expression, as in 'John-i thinks he-i is right' or precede it, as in 'Mary says he-i would flee because John-i is a coward.' Yet the pronoun cannot precede its noun expression sometimes, as in *'He-i thinks that people don't like John-i', where 'He' and 'John' cannot refer to the same person (in English and many other languages).

In all these cases, something constrains these apparently general patterns from applying with complete generality. One cannot say people have not heard 'sentences just like these' before, because people can generally produce and accept an unlimited number of sentences which are new in some of their properties. So what makes the particular 'new properties' of these unacceptable uses so conspicuous – especially when speakers have no conscious knowledge of the relevant properties? Chomskyans would say that speakers have a limited range of settings for certain types of properties, and watch carefully to see where the input sets them for their own language. The properties are quite abstruse, such as how many types of syntactic boundaries of what type, the 'path' between a wh-expression and its 'gap' can cross, or abstruse structural relations between pronouns and noun expressions. Speakers would never know to pay attention to just these unless they were innately set to do so. Yet knowledge of what they are, and how they can work, in turn would require innate knowledge of large amounts of other grammatical equipment.

What is stated here briefly is a set of arguments that in reality, some months of linguistic training would probably make fully sensible. Yet the basic argument is simple: people have strange quirks in how they interpret the input data, and it is hard to get these quirks out of general, nonlinguistic cognition. Unless it can be shown that such quirks can be acquired naturally from more general cognition (as some linguists are attempting to do), the Chomskyan case, though based on uncommon (indeed unheard) sentences, has considerable general strength.

Yet little is known about children's acquisitional patterns in dealing with such constructions, and

what is known is inconsistent. Sometimes they appear precocious in dealing with such constraints, sometimes they do not, and for the most part, we do not know. It is not clear Chomskyans in the end would care about empirical acquisitional patterns, however. They might well say that the ability to formulate such constraints properly, without overt instruction, is the main point.

FINDINGS FROM 'NON-USUAL' POPULATIONS

Unclear Findings from Brain Localization

In attempting to test the innateness hypothesis, acquisition of language under 'non-usual' conditions has been prominent. Much of this work, especially work on children with particular retardation syndromes, or work on children and adults with damage in various parts of the brain, was initially used to argue for a biologically specific 'language faculty' housed in particular parts of the brain. However, the work has turned out to be ambiguous or puzzling at best. Indeed, some of the findings fit into no straightforward theory at all. For example, children with hydrocephalus have an excess of spinal fluid in their brains. Current treatment is to insert a shunt into the spine to drain the liquid immediately. Decades ago this was not done, however, and the result was extensive, near-total destruction of cerebral cortex tissue, the tissue believed critical for higher thinking. Yet a good many of these children turned out to be average or even well above average in developing normal language and cognition. Absolutely no current theory of brain localization or recovery can accommodate such data. In general, the brain injury data show surprising robustness of language, without giving straightforward answers to questions of language innateness.

'Home Sign'

From the nativist point of view, perhaps the most exciting 'non-usual' acquisitional work has concerned deaf children raised with little language input who seem nevertheless to devise and develop their own sets of signing symbols, which they may combine into simple propositions. Unless very generous analyses are made, these propositions seem so far to have simple structure far from the complexities at the heart of Chomskyan analysis, but the phenomena still point to a basic symbolic and propositional 'set' in human children.

Chomsky's particular nativist hypotheses remain unproved. However, when he proposed nativist explanations some forty years ago they were radical proposals, whereas now they are generally viewed as at least plausible. Beyond this, Chomsky's work pointed out the complexity of language under any reasonable analysis, making language acquisition the most impressive mental accomplishment that is widespread and sturdy throughout the human species.

Further Reading

Brown R (ed.) (1970) *Psycholinguistics*. New York: Free Press.

Brown P (1998) The acquisition of Tzeltal verbs. *Linguistics* **36**: 675–694.

Choi S and Bowerman M (1992) Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. *Cognition* **41**: 283–322.

Chomsky N (1959) Review of *Verbal Behavior* by B. F. Skinner. *Language* **35**: 26–58.

Goldin-Meadow S and Mylander C (1984) Gestural communications in deaf children: the effects and non-effects of parental input on language development. *Monographs of the Society for Research in Child Development* **49**(3–4, serial no. 207).

Ingram D (1986) *First Language Acquisition*. Cambridge, UK: Cambridge University Press.

Juscyzk P (1997) *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Maratsos M (1998) The acquisition of grammar. In: Damon W (series ed.) *Handbook of Child Psychology*, 5th edn, vol. 2, Kuhn D and Siegler R (vol. eds) Cognition, perception, and language, pp. 421–466. New York: John Wiley.

Pinker S (1984) *Language Learnability and Language Acquisition*. Cambridge, MA: Harvard University Press.

Slobin DI (ed.) (1985, 1992) *The Crosslinguistic Study of Language Acquisition*, vols. 1, 2. Hillsdale, NJ: Erlbaum.

Language and Brain

Introductory article

Merrill F Garrett, University of Arizona, Tucson, Arizona, USA

CONTENTS

The major structural features of language systems are systematically reflected in specialized neural systems of the human brain. Evidence for this conclusion derives from patterns of language breakdown and from physical measures of brain activity made during normal language use.

INTRODUCTION

What questions are addressed in studying brain and language relations? And why are they of scientific interest? A broader context suggests some answers. The objective in all brain and behavior studies is to understand how specific abilities can be physically realized, given the capacities and limitations of nervous systems. Beautifully detailed accounts of such relations exist for neural representations of sensory and motor capacities in human and nonhuman systems. The step to higher order mental capacities (cognitive systems), however, presents difficult additional problems. Cognitive systems involve abstract levels of representation that do not have direct links to physical features of the world. The principles that link abstract representations to neural activity are not well understood as compared to, for example, sensory transduction and early stages of perception. Language provides a superb means to attack this issue. Although it is a very specialized behavior, it supports communication across the range of human thought. It therefore contains components that are language specific and those that are not. Language is a truly central capacity that ties together many different mental systems. Studying its brain organization can indicate how in the general case the complex, multicomponent processes of human cognition are realized in neural systems.

Language and brain study is historically tied to language deficits (aphasias and related disorders) that result from brain injury and disease, or from surgical interventions to treat neurological problems. Those data have been enriched in recent years by brain imaging measures. These assess neural activity during behavior, showing which brain regions support different mental processes. Imaging is complemented by electrophysiology, which monitors changes in the brain's electrical activity to distinguish among types of mental processes and to indicate when they occur. Combining these sources provides an intricate mosaic of brain data that can be linked to the complex structure of language processing.

This work requires careful distinctions to be made among types of language processes. Everyone who has undertaken the study of a foreign language becomes acutely aware that many different skills must be acquired – 'proper' pronunciation, rapid word retrieval, syntactic details absent or different from one's native language, and different ways of thinking about the meanings of individual words. All these (and more) must be fluidly combined in order to use a language as normal communicative exchange demands. The intuitions developed from such experiences are an informal expression of ideas that drive scientific investigations of language in linguistics and psycholinguistics. The studies reviewed here draw on those sciences. Figure 1 gives a summary of relations among language structures. Broadly speaking, language is studied at the levels of sounds *(phonetics, phonology, and prosody)*, words *(morphology and lexicon)*, sentences *(syntax)*, and meanings *(semantics)*. Theories of language model the mental processes that generate and recover such structures in real time (i.e. while listening or speaking).

Figure 1 displays major classes of relations among language structures: those within a structural type

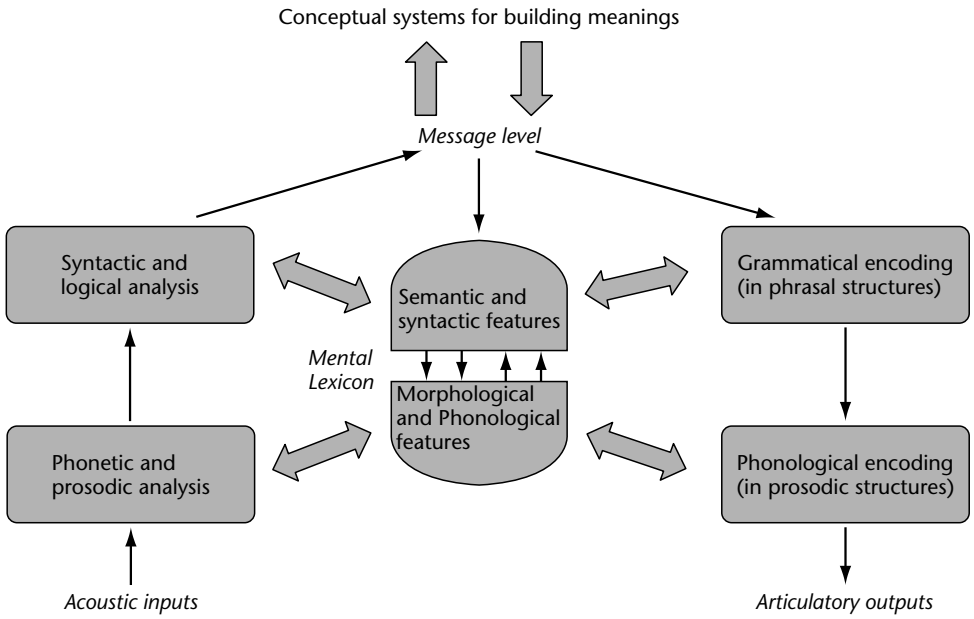


Figure 1. The major structural systems of language and their relations in processing. Sound systems include phonetic, phonological and prosodoc structure: phonetic categories of language connect to sensory and motor data by primary speech perception/production processes. In turn, the information represented by phonetic distinctions is linked to the more abstract categories of phonological and morphological structure in systematic ways. The phonological and morphological structures are systematically associated with the inventory of words in a language (the lexicon). In the lexicon, the form of every word must represented so that it can be identified or pronounced. Lexical representations are also associated with sets of syntactic and semantic features. These reflect the syntactic and semantic regularities of a given language. Systems for syntax and meaning use lexical codes to develop phrasal structures and coordinate them with semantic interpretations.

and those between types. These define the set of functions needed for a comprehensive theory of language. The study of language and brain targets the neural processes that are correlates of mental systems that assign such structures to sensory inputs (comprehension), or generate such in response to conceptual inputs (production).

LOCALIZATION OF LANGUAGE FUNCTIONS

What is the localization question? At first glance, it seems fairly easy. Are distinct neural systems associated with the processing of the major classes of linguistic structure? Broadly speaking, the answer is 'yes'. Processing for sounds, words, syntax, and semantics activates distinguishable brain areas. But the fuller story requires additional distinctions.

Some General Issues in Localization

Several related ideas animate localization claims. First, localization need not mean there is just one place where all activity related to the processing of, for example, sounds, or syntax, or semantics is done. Keep in mind the organization of Figure 1. In particular, different effects may appear for *system internal* processes (e.g. relations of sounds to sounds, or syntactic categories to syntactic categories, etc.) as compared to interface processes that relate sounds to words, or word forms to meanings or syntax, etc. System internal processes do pose the question of whether brain regions select for one type of structure. But brain responses that combine two or more structural types are also to be expected, and identifying which combinations of structure are associated with specialized brain activity may reflect system interfaces. Still another question is this: if there is a brain region for function X, is it specific to language – i.e. is it dedicated to language *per se*, or does it also serve other mental systems? It is perfectly possible to find a positive outcome for association of brain regions with specific structural types, and a negative outcome for specificity to language *per se*. Current evidence does not provide definitive answers for all variants of localization questions. There is good reason to claim brain specialization for language processing, with localization for syntactic aspects

most sharply defined. The possibility that some regions of specialization for language may also serve other mental functions remains open.

An Overview for Brain and Language Relations

As preliminary orientation, the regions of brain implicated by discussion in later sections are outlined in Figure 2. Localization questions arise at two levels. One is lateralization (hemispheric specialization). Many observations show differences between left and right hemispheres for the support of language. The second level of localization claims is finer grained. It associates focal areas of the left and right hemispheres with different language functions. A word about variability regarding localization is in order before proceeding further, however.

Discussion here assumes the dominant patterns of localization, but two exceptions should be kept in mind. First, some left-handers (about a third) have reversed lateralization. This is complementary to the pattern in almost all right-handers and the majority of left-handers. In reversed dominance, right hemisphere regions homologous to left hemisphere regions support language processing. This systematic variation poses no great theoretical problem. A second exception is more challenging. There is evidence for significant individual variability in language areas. Such variation in expression does not itself invalidate the idea of functional localization. It does highlight the need for precise accounts of mechanisms that produce localization and how they account for variability. This issue is not well understood and is a matter of active research. With these reservations in mind, significant generalizations can be summarized. Despite the variability just discussed, the association of major language functions with particular brain regions is robust.

Figure 2 shows the left hemisphere as the major focus of language activity. The notion that left and right hemispheres support different cognitive and perceptual functions is widespread in popular culture. Left brain thinking versus right brain thinking is a cliché. But, clichés sometimes reflect important truths, and that is so here. Many specializations of the hemispheres exist and that for language is among the most compelling, but here too there is indication of the brain's dissection of language: not all of language is strongly lateralized. Interpretive functions and lexical representation are more distributed between the hemispheres, and aspects of speech perception are bilaterally represented as well. By contrast, the structural core of morphology and syntax supports the clearest claims for lateralization. Aphasia patterns and split-brain patients' performance reflect this. That evidence is complemented by brain imaging and electrophysiology studies. Historically, perisylvian areas of the left hemisphere (Broca's area and Wernicke's area) were identified as the primary language areas on the basis of aphasic disturbances. More recent evidence using better tools for the identification of lesion site and effects, coupled with imaging data from normals, has broadened the regions of language-responsible tissue in the left hemisphere and implicated some right hemisphere activity.

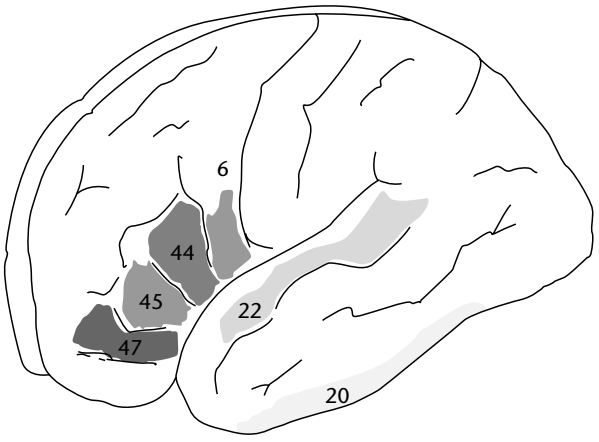


Figure 2. Left lateral view of the brain, labeled for cytoarchitectonic areas (viz., Broadman's areas: BA 20, BA 22, BA 44, etc) referred to in the text. Contrasting colors indicate areas responsive to distinguishable language processing activities as discussed in the text.

- BA 20: inferior temporal gyrus
- BA 22 Anterior portion; posterior portion of area 22 is in the region called Wernicke's area
- BA 44: Pars opercularis; in the region called Broca's area
- BA 45: Pars triangularis; in the region called Broca's area
- Area 6: anterior Rolandic operculum

THE APHASIAS: PATTERNS OF BREAKDOWN IN LANGUAGE FUNCTION

Aphasia is the general term for language-specific deficits caused by brain injury. The plural form 'aphasias' signals the fact that many subvarieties of language compromise exist. Brain injury usually causes multiple cognitive problems, but some deficits can be quite selective – features of sound structure, word structure, sentence structure, and meaning all seem isolable. Brain damage can pick

out a particular piece of the language repertoire and leave the rest largely intact.

Some History

How can such a variety of language impairments arise? The relevant history begins with two pioneers whose names survive as labels for the constellations of symptoms they reported: a French physician, Paul Broca ('Broca's aphasia') and a German physician, Carl Wernicke ('Wernicke's aphasia'). In 1861, Broca identified patients with damage to fronto-temporal areas (Broca's area, Figure 2) who showed hesitant speech ('nonfluent') marked by omission or reduced use of certain grammatical classes (verbs and some classes of function words and inflections); comprehension was relatively well preserved. A few years later, in 1874, Wernicke reported a contrasting pattern for patients with damage to posterior regions of the temporal lobe (Wernicke's area, Figure 2). Their speech remained fluent and superficially grammatical but had frequent errors for phonological and semantic features of words; comprehension was seriously impaired.

These patient groups differ in almost every way. Nonfluent Broca patients, with relatively good comprehension, have impaired production of syntactic structures (described as agrammatism). By contrast, the fluent Wernicke's patients, with poor comprehension, produce syntactically complete structures but make semantic errors in word selection (described as paragrammatism). The losses seem complementary and arise from very different lesion sites. These contrasts were interpreted for a long period in terms of a model that identified posterior language zones with comprehension and the anterior language zone with production. Other aphasic patterns were linked to damage to underlying fibers (the *arcuate fasiculus*) linking the two areas. The resulting model (the Wernicke–Lichtheim Model) and its variants provided the classic account of aphasia until about three decades ago. Several findings then began to erode confidence in the presumed functions of Broca's and Wernicke's areas.

Problems for the Classic Account: Behavioral Inconsistencies

An influential series of studies in the 1970s demonstrated that many agrammatic aphasics have comprehension limitations that mirror their production problems. When such patients must rely on syntactic cues (of the sort missing from their speech), comprehension is impaired. Their successful comprehension relies on lexical semantic cues. So, they understand sentences like 'The flowers were watered by the girl' based on (preserved) knowledge of plausible relations between objects and events in the world; but such knowledge is not decisive in sentences like 'The boy was pushed by the girl' – syntactic detail is essential. In these circumstances, agrammatics perform poorly. This fostered the idea that basic syntactic capacity resides in Broca's area.

This idea did not survive long. Very shortly, patients were reported who displayed agrammatic speech but had no comprehension problem. Moreover, still other patients showed comprehension deficits but no agrammatic symptoms. Even more perplexing, other tests of agrammatics showed they could correctly judge the grammaticality of the very sentences that they could not interpret. They showed a syntactic loss in one domain, but not in the other. All this rules out an exclusive association of Broca's area with syntactic capacity – and it begins to suggest problems in processes that integrate different types of structures as opposed to loss of a specific structural type. Other complications exist, but these suffice to illustrate the difficulties in explaining language deficits with the components of the classic aphasia theory. A richer set of assumptions about brain and language is required.

Other Problems for the Classic Model: Lesion Data

The perisylvian areas of the left hemisphere are undoubtedly central to language function, but saying precisely how has been a vexed matter for aphasia. Careful examination of large numbers of patients with contemporary methods of observation has revealed a poor fit between Broca's aphasia and classic lesion sites. One finds patients with agrammatism but no lesion in Broca's area and patients without Broca's aphasia with a lesion in Broca's area. Moreover, other areas (area 22, anterior portion Figure 2) are statistically better linked to Broca's aphasia than Broca's area itself.

The evidence conflicts in some respects. Why? Does it invalidate the idea that focal areas for language exist? Not necessarily. There are other good reasons for variability in the lesion data. First, lesion site is only a part of the story. Lesions also vary in size and character. Moreover, reduced activation in adjacent tissues is known to be important, but has been hard to evaluate in patient populations. Recent imaging studies have verified

and clarified this factor: the observable lesioned area does not fully identify the regions of compromised neural function. When this fact is coupled with variability of localization across individuals noted earlier, it helps explain why the lesion landscape is so hard to interpret.

Variable lesion data complicates inferences about specific brain areas and language functions. But, a salient fact remains: the existence of many highly selective language deficits is not open to serious question. Thus, although one may have difficulty in saying exactly which region of brain is the locus of an effect, the inference that there are specific brain regions for language functions seems correct. The isolability of language processes regarding brain activity is attested even if the physical base is uncertain on lesion evidence.

Other Important Performance Issues: The Modes of Language Exercise

Attention has so far focused on types of structure, but modality is also important. Modalities include speaking and listening, reading and writing, and spoken versus signed languages. These modes of language may be independently affected in aphasia. Production and comprehension may be separately affected, with loss of a structurally defined capacity in production and its preservation in comprehension (and vice versa). Written language (at several levels) can be affected independently of spoken expression (and vice versa). Moreover, some aphasic disorders are mirrored in acquired reading problems ('dyslexias') – structural contrasts for language compromise appear across modality.

Visual/manual-based languages, such as American Sign Language (ASL), are of special significance. The basic grammatical structure of ASL is like that of spoken language, and aphasias in ASL show similar structural bases to those of spoken languages. In ASL, gestural systems can be selectively impaired in ways that mirror spoken language deficits (while preserving nonlinguistic gestural capacity). This is a striking result, given that important visual processes are specialized for right hemisphere function, but careful study of ASL users with brain damage indicates that language functions are left lateralized in ways comparable to spoken language users.

Moral: the many subsystems for implementing language are represented in the brain in ways that allow selective interference from physical damage, and do so in similar ways across domains of performance. On any account of language disorders, the affinity of phonological, morphological, and syntactic language functions for a relatively circumscribed set of areas in the left hemisphere remains a central explanatory challenge for brain and language theories.

DISORDERS OF SEMANTIC MEMORY AND LEXICAL ACCESS

Lexical structures are an essential resource for sentence level processes. Knowledge of the structures peculiar to individual words guides language generation and comprehension. Is X a noun, a verb, a quantifier? Does it have morphological parts? How is it pronounced? What does it mean? Such information must be accessible in different ways. To talk, the words that express a thought must be picked from all the words a speaker knows. Experimental study of normal language processing indicates that this is a two-stage process, one meaning based and one form based. The initial step links a lexical concept with a word representation that provides only its semantic and syntactic characteristics. A second step connects this abstract representation to sound structures for pronunciation. Similar ideas apply in access for comprehension (See Figure 1). How is such lexical processing represented in the brain?

Dissociations of Meaning, Syntax, and Form

Various errors of meaning (e.g. saying 'sword' for 'arrow') and form (e.g. saying 'sympathy' for 'symphony') are frequent in language pathologies (and not uncommon in normal speech and reading). One finds well-attested cases of patients whose errors primarily involve meaning relations or primarily word form failures, and this dissociation can arise either for production or for comprehension. Aphasic patterns demonstrate that meaning-based and form-based systems of lexical recognition and retrieval can be separately affected by brain injury.

Outright retrieval failures are also common in aphasia. 'Anomia' is the cover term for word finding difficulties – well-known words are not available for utterance. A similar phenomenon in normal language users is the 'tip of the tongue' (TOT) state. The sharp pang of failure to recover a well-known word is a universal subjective experience. Indeed, naming failures in anomic patients have parallels to those of normal speakers in TOT states: dissociations of semantic, syntactic, and phonological levels of word description are observable in both groups. To illustrate: anomics usually

have detailed semantic and conceptual information about a target picture they cannot name. But syntactic features of blocked words may also be available. Tests of anomic patients in languages with gender marking systems (e.g. Italian, German, Spanish) show this. Grammatical gender is tested because it is a feature of syntax that can be distinguished from conceptual and semantic covariates. So, for example, in Italian, different varieties of pastas may be masculine, feminine, or neuter, as are rocks of varying size (pebble, stone, bolder), etc. Italian anomics can report the grammatical gender of a target word with high accuracy even when unable to provide any information about its pronunciation. Similar dissociations arise for normal speakers in TOT states. All this demonstrates that sound structures for words are distinguished from their meaning and syntax in both brain and behavior.

Syntactic Lexical Disorders

Lexical deficits can also pick out syntactic categories. Nouns can be selectively impaired in anomic patients. Striking cases of this are seen in patients who speak in fluent, syntactically well-organized and communicatively appropriate sentences but do so with greatly diminished ability to recover contentful nouns. These are replaced by pronouns or nonspecific nouns ('thing', 'place', 'stuff', etc.). Verbs, by comparison, are specific and appropriate. Interestingly, a contrast with this pattern is typical for many Broca's aphasics for whom verbs are more compromised than nouns.

A second kind of syntactic loss was earlier remarked in discussion of agrammatism. It involves failure to recover function words and some word affixes (e.g. inflections for number, tense, and aspect). This set is sometimes labeled the 'closed class vocabulary' to call attention to its stability – new nouns and verbs are invented as the need arises, but new function words and affixes are rare in language change. This stability makes sense. Closed class words are tied to enduring structural properties of a language. They convey grammatical relations and functions rather than referring to objects, events, and properties. Broca's aphasics show a diminished capacity for such words. Certain reading disorders (deep dyslexia) are also marked by much poorer ability to read closed class words as compared with major category words.

These patterns are robust, but their explanation in terms of storage or processing theories of the lexicon, or relations to brain structure, is not well understood. Localization accounts based on lexical representations seem unlikely for such broad ranging category effects. General computational mechanisms for syntactic analysis and composition are more likely candidates.

Category-Specific Semantic Deficits

Selective semantic losses are also found. Aphasia research has yielded several reports of word losses restricted to particular meaning classes: such as inability to name pictures of animate objects while succeeding with pictures of inanimates, or relative success with all word classes except, for example those referring to fruits. These category-specific lexical losses are not the consequence of conceptual level impairments. Such patients may make accurate non-linguistic judgments about a problematic word and provide detailed information about its reference ('it's small, you use it for cutting paper; it's made of metal…') while unable to name it if it is an artifact. The same patient may succeed with tests on animate objects, naming even rather exotic animals. The reverse pattern has also been reported. Deficits may be quite circumscribed: body parts, foods, flowers, vegetables, musical instruments, and tools are among other reported examples.

What about brain areas linked to these deficits? The lesion data are variable – some cases arise for diffuse damage, and others are more localized. Multiple brain mechanisms are likely to be implicated across the affected categories. Moreover, the numbers of patients showing well-defined losses is not large. Not surprisingly, controversy surrounds accounts of category-specific losses, but that discussion is beyond the scope of this article. However, studies that combine aphasia information with brain imaging do give some helpful insights. A good example compares brain areas for patients with losses in one or more of these sets: famous faces, animals, and tools. In one such study of naming by aphasics, impairments for these three categories were associated with lesions in different parts of the left temporal lobe but outside the usual perisylvian language areas (area 20; see Figure 2). The results were complemented by a positron emission tomography (PET) study with normals showing activation of these areas during picture naming for the categories. Do these selective losses reflect localization driven by meaning features for words? Analysis of the deficits for faces, animals, and tools suggests not. Some patients showed mixed losses – two categories impaired, but not the third. Not all combinations occurred. Faces and animals did coincide, but tools and faces did not. Why? One might

first suppose the similarity of the semantic representations is at work. People have faces and people are animals. Tools and faces should contrast, and indeed, they did not co-occur. But the third pairing does not fit this picture: animals and tools did co-occur. Furthermore, lesion data and the PET study showed the tools area separated from faces by the animal area, so two isolated infarcts would be required to get tools and faces while preserving animals. The effect seems to reflect spatial distribution, not meaning similarity.

HEMISPHERECTOMY AND SPLIT-BRAIN RESEARCH: DETAILS OF LATERALIZATION

Surgical procedures undertaken to alleviate otherwise untreatable neurological conditions turn the spotlight on right hemisphere (RH) language functions. 'Split-brain' patients reveal RH language capacities in adult brains with established left lateralized language prior to surgery. By contrast left hemispherectomized patients reveal the potential for language development by infants who must rely only on RH. The implications are complementary: the split-brain research shows significant RH language capacity even in left lateralized brains, while hemispherectomy research shows sophisticated language processes can be acquired in RH. The flip side is that both classes of data also suggest that some language capacities are specific to the LH and that these involve aspects of the syntactic organization of sentences.

Split-Brain Patients

Certain medical conditions require a surgical procedure that severs the channels of communication between left and right hemispheres (the corpus callosum). When the callosal fibers are cut, each hemisphere operates independently. The hemifield structure of the visual system allows selective stimulation of LH or RH (inputs to the left hemifield of each eye project only to RH and vice versa for the right hemifield of each eye). Thus, procedures that project words to one or other hemifield can test individual hemispheric capacity, to judge, for example the suitability of lexical targets for given syntactic or semantic environments. In such tasks, the RH supports accurate semantic decisions about words and can process the syntactic structure of simple sentences. Complex syntactic structures are beyond its apparent capacity.

What interpretation should be put on this? Should LH/RH contrasts be seen as exclusive specialization and the limited success of RH a residual but irrelevant capacity (low-level redundancy) with regard to normal language use – perhaps left over from early stages of development during which left dominance is established? Or does the RH provide useful language support that is normally invisible because the LH controls overt language expression? Some hints favor the latter perspective. Speech processing illustrates this. Both hemispheres are involved in the translation of acoustic inputs to phonetic structures, but recent research indicates different functions for the hemispheres. Speech perception maps a complex set of acoustic cues onto phonetic categories. RH and LH appear sensitive to different classes of these cues (steady state versus transient). There is hemispheric asymmetry, but it is complementary rather than competitive. The extent to which this perspective should be generalized is not clear, but there is other suggestive evidence. For example, reaction time studies of word priming (e.g. word 1 of a successively presented pair of words facilitates decisions for word 2) show a complementary sensitivity of RH and LH to associative relations (e.g. fire/hydrant, fossil/fuel) versus semantic relations (e.g. categorial relations: plant/flower, horse/dog). RH performance emphasizes the former, and LH the latter. A potentially related effect arises in studies of lexical ambiguities: RH may support recovery of nondominant senses of ambiguous words.

Hemispherectomy Patients

Surgical intervention for a congenital neurological disorder (Sturge–Weber syndrome) requires removal of the entire LH ('hemispherectomy'). Infants who undergo this procedure (in the first year of life) develop apparently normal communicative language capability. The RH can, if the process begins early in life, assume major LH language functions. A similar conclusion derives from observations of childhood aphasia. Effects of brain damage are dramatically different in very young children compared to adults. Children rapidly recover normal language function after injuries that would leave an adult with an enduring deficit. Language functions are apparently taken over by nearby LH brain areas or by RH regions homologous to the LH language areas.

These observations suggest early equipotentiality for language in left and right hemispheres, with a commitment of left hemisphere tissue that becomes increasingly fixed with age. But, there is an important caveat: adult syntactic performance in hemispherectomy patients is not fully comparable to normal performance. Complex syntactic structures

are more difficult for such patients to comprehend or produce. This limitation indicates a LH priority in language development. Evidence from a quite different source supports the same conclusion.

Another Learning Perspective

Developmental research supports the idea of sensitive periods in language acquisition. Success in attaining nativelike performance by second language learners depends on when exposure to the second language begins. Experimental tests have shown detection of grammatical violations (e.g. word order or agreement errors) to be less accurate for those who learn a second language 'late' (after about 5 years old) compared to those who learn earlier. The effects are graded (systematically poorer performance is associated with increasing age of first exposure). Is this sensitive period linked to hemispheric specialization? Brain imaging effects suggest so: early learners displayed sharply lateralized LH brain activation when they heard syntactic violations. But learners whose exposure to the target language was delayed by a few years had RH activation as well as left. By contrast, tests of semantic violations in the same groups did not show a sensitive period. Early and late acquiring groups showed similar brain responses for words that did not fit the meaning of test sentences. These and related findings indicate localization depends on the stage of brain development at which processing is acquired. Further, the consequences of lateralization seem most sharply focused on the syntactic devices of language. This developmental picture thus fits neatly into the framework that emerges from aphasia, hemispherectomy, and split-brain research.

BRAIN IMAGING: MEASURING ACTIVATION DURING LANGUAGE PROCESSING

Imaging approaches compare neural activation in different brain areas during mental activity. A prominent example is PET, which measures variation in blood flow to different brain regions as indexed by the concentrations of a weak (harmless) radioisotope injected prior to observation. Another example is fMRI (functional magnetic resonance imaging), which measures blood flow based on changes in the magnetic properties of red blood cells. These and related measures have been the focus of intense activity over the past two decades, and the 1990s saw extraordinary advances in their reliability, precision, and availability.

Imaging Studies Contrasting Semantic and Syntactic Processes

Many imaging studies have used the strategy of comparing semantics and syntax, and distinct patterns for syntactically and semantically based judgments have been repeatedly observed. In one such study, listeners had to judge whether two sentences had the same meaning. One experimental condition varied content words, and so judgment required a lexical semantic process. A second condition varied ordering of phrases, and so required a syntactic process. Activation areas differed: syntactically driven judgments produced activation in a *subpart* of Broca's area (pars opercularis), while semantically driven judgments activated an area below Broca's area (Figure 2, area 47). Several other imaging studies requiring semantic processing have also showed activation in this latter area and area 45. From study to study, regions in or near Broca's area appear as a focus (see Figure 2), but there is some variation in the detailed activation patterns. The variation may reflect differences in tasks and materials, along with possible individual subject differences.

The preceding example highlights a semantics and syntax contrast in anterior language regions. Semantic responses occur elsewhere in other imaging work, including more posterior areas. Earlier discussion of semantically specific lexical losses also cited areas outside the classic language zones associated with identification of persons, animals, and tools. Moreover, split-brain research indicates significant RH lexical competence. Imaging outcomes support this as well. In general, imaging work shows lexical representations and meaning-based processing broadly distributed in the left hemisphere, with both anterior and posterior regions represented and right hemisphere activation commonly observed.

Syntactic Responses in Brain Imaging

A preoccupation of aphasia research has been the relation of syntactic deficits to Broca's area. Imaging studies that vary syntactic complexity amplify the aphasia work. As syntactic processing demands increase, activation levels in responsible areas should increase. A contrast in relative clause structures illustrates this. Compare: 'The dog the boy teased bit him', versus 'The dog bit the boy who teased him'. These two make the same set of claims about the world – they mean the same thing – but the organization of information differs. Behavioral measures show the first (with a

center-embedded relative clause) is more difficult for normal listeners than the second (a right-branching relative clause). Moreover, agrammatic aphasics have difficulty with center-embedded structures when they lack semantic cues that tell who did what to whom. When syntax load varies but semantic content does not, what is the imaging result? PET studies showed selective activation in Broca's area. In some studies, a subpart of the area called 'pars opercularis' (Figure 2, area 44) was again singled out. Other imaging experiments, using fMRI measures, have also manipulated syntactic complexity. These too reveal activation in Broca's area, though with some variation for locations within the region. A related observation that does not rely on complexity *per se* is that differential responses to open and closed class vocabulary also implicate a part of Broca's area. Given the association of this vocabulary contrast with syntactic processes, the result makes sense, but it is not clear how it should be linked to the work involving complexity. Overall, this work affirms a syntactic role for Broca's area, though it leaves open questions about what aspects of processing are involved and how they relate to each other.

To this mix, a modality factor must be added. A few imaging studies have tested syntactic processing in production rather than comprehension. Recall that the initial characterization of Broca's area in aphasia assigned production processing as its primary function. That claim is not supportable for the full range of aphasia observations. What does imaging research contribute? Recent imaging studies show selective activation for syntactically structured as compared to unstructured word strings in a production task. These were equated for the conceptual input that drives encoding, but activation was not in Broca's area. It was immediately adjacent (Figure 2; left anterior Rolandic operculum). There is, thus, good reason to locate multiple language functions in or near Broca's area and these include several structural types and modes of processing.

ELECTROPHYSIOLOGY: LANGUAGE PROCESSING SYSTEMS

Measurements recorded from scalp electrodes (electroencephalography) have been used to study perceptual processes for many decades, with wide applications in health-related areas. An adaptation of those techniques (event-related potentials: ERP) can be used to study cognitive processes. Recordings are time-locked to presentations of a specific class of stimuli and the signal averaged across many examples to extract regularities associated with a stimulus type. The onset and duration of a brain response to different language structure types can thus be estimated and compared. Characteristic response patterns are defined in terms of scalp location of recording electrodes, time course, and polarity of the electrical signals. The sources of these electrical changes (generators) are difficult to pinpoint precisely from scalp recording, but different patterns at quite separate electrode sites indicate different underlying brain processes. Moreover, there is substantial correspondence between variation in ERP recording sites and imaging results regarding the general locales for different classes of processes.

Semantic and Syntactic Contrasts in Electrophysiology

A classic study of language using ERP methods revolves around responses to semantically incongruous words in sentences. A characteristic waveform (N400) is robustly triggered by the appearance of an unexpected word, as, for example 'socks' in 'She spread the warm bread with *socks*'. The N400 label reflects the characteristic negative polarity and latency of this response to reach peak amplitude. The pattern of activity is similar across several electrode sites – it is not tied to a narrow range of recording locations. It is virtually absent at anterior sites, bilaterally represented at most posterior sites, and often more prominent at right hemisphere locations. Discovery of this regularity was a milestone in language study and generated many studies that explore semantic processing, most specifically aspects of integrating the meanings of individual words into the overall interpretation of a sentence.

Note, however, that an N400 is not elicited by syntactic deviations (as agreement errors or grammatical category violations). Thus, the imaging results mentioned above, showing distinct regions of activation for semantic and syntactic tasks, have a corresponding contrast in electrophysiological measures.

Syntactic Contrasts in Electrophysiology

Work in several languages (English, Dutch, German, Italian, and others) shows two classes of syntactically driven ERP patterns. These are the P600 and various anterior negativities in the 200–400 millisecond range. Though all have syntactic triggers, there are important differences in the conditions that elicit them.

Many studies have demonstrated a sustained positive polarity shift that normally peaks around 600 ms following a syntactic violation site (P600). This response is broadly distributed, as is the N400, but it is very different in polarity and latency from that associated with semantic violations. P600 shows effects both of syntactic acceptability and of relative probability of syntactic analysis. When two analyses are syntactically possible partway through a sentence, a P600 can be triggered by a word that requires the less preferred (less frequent) of the two.

A rather different complex can also be triggered by syntactic violations, but as compared to P600 the conditions that elicit it are more restricted. These responses involve negative polarity shifts at anterior recording sites. They range in latency from 200 ms to 400 ms and contrast with semantically driven responses in recording sites and latency. They are strongly lateralized to the left in some studies and bilateral in others. Sentences with memory requirements tied to moved elements (e.g. question forms and center branching relatives) have been shown to produce a bilateral anterior negativity around 400 ms after the element that signals the movement (e.g. the relative pronoun or question word). More strongly lateralized response patterns have been observed for syntactic violations involving phrase structure and morphological errors of inflection. Latencies vary from 200 to 400 ms for these responses across studies from different languages and task variables. Moreover, among the strongly lateralized left anterior responses, several studies indicate an ERP signature for closed class words (N280). This latter finding has a rough parallel in the imaging work showing a left anterior locus for contrasts of open and closed class vocabulary.

Comparison of these two classes of syntactically driven ERP responses suggests interesting processing interpretations. The strongly left lateralized anterior negativities arise only for syntactic violations rather than syntactic preferences or interpretive strategy. By contrast, P600 is also modulated by multiple nonsyntactic influences (task strategies and syntactic preferences both syntactic and semantically driven). The differences are thus suggestive of early and late stages of sentence integration, with early reliance on relatively inflexible automatic processing routines, while later stages show accommodation to multiple constraints on the final analysis and interpretation of a sentence.

SUMMARY

Several conclusions supported by evidence from aphasia and brain injury are refined and extended by contemporary measures of brain activity. In the process the early aphasia-based picture of localization has been extensively redrawn. Several findings are instructive.

First, imaging studies confirm the involvement in language processing of the major regions of perisylvian cortex identified on the basis of lesion data from aphasia study. But, regions associated with various lexical and interpretive processes are widely distributed. Imaging and ERP data have made it clear that there is more RH activity during language processing in left lateralized brains than was supposed. Regions associated with syntactic integration are more circumscribed but nevertheless represented at more than a single site. More particularly, however, imaging studies show multiple regions within Broca's area and immediately adjacent areas that are associated with distinguishable processing tasks. These include both syntactic and semantically based processes and include comprehension and production modes. In short, the language-responsible tissue in the perisylvian areas of the frontal, temporal, and parietal cortex shows a structure more in keeping with a co-ordinated multicomponent system than a monolithic segregation into meaning-based comprehension processes and form-based production processes.

Electrophysiology shows distinct semantic and syntactic processing systems but also evidences of semantic and syntactic integration systems. Within the systems triggered by syntax, there is a contrast between processors sensitive to a limited range of morphosyntactic and phrasal variables and processors responsive to a broader range of syntactic processes. These latter reflect interactions between syntactic structures and semantic constraints. Though localization issues for electrophysiological signals measured at the scalp are problematic with regard to underlying brain regions, the general distributional features of the scalp recording sites for the several different ERP patterns are broadly compatible with the distribution of language-responsible areas revealed in brain imaging.

Sound systems, syntactic systems, and semantic systems are all complex. None of them are 'one thing'. And they interact with each other in diverse ways. The newer information sources emphasize the need to provide accounts of how different information classes in language systems interact to yield the rapid and highly accurate performance profiles of human language use. The specificity of response measurable with contemporary tools of observation now has the potential to discriminate the details of sophisticated accounts of structure and

processing provided by linguistic and psycholinguistic theory.

Further Reading

Brown CM and Hagoort P (eds) (1999) *Neurocognition of Language*. Oxford: Oxford University Press.

Caplan D (1992) *Language: Structure, Processing, and Disorders*. Cambridge, MA: MIT Press.

Caramazza A and Shapiro K (2001) Language categories in the brain: evidence from aphasia. In: Rizzi L and Belletti A (eds) *Structures and Beyond*. Oxford: Oxford University Press.

Corina DP (1998) The processing of sign language: evidence from aphasia. In: Whitaker H and Stemmer B (eds) *Handbook of Neurology*. San Diego, CA: Academic Press.

Friederici AD (2001) The neuronal dynamics of language comprehension. In: Miyashita Y, Marantz AP and O'Neil W (eds) *Image, Language, Brain*. Cambridge, MA: MIT Press.

Gazzaniga MS (1998) The split brain revisited. *Scientific American* **279**(1): 35–39.

Hickok G and Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* **4**: 131–138.

Indefrey P and Levelt WJM (2000) The neural correlates of language production. In: Gazzaniga M (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 845–865. Cambridge, MA: MIT Press.

Language Comprehension, Methodologies for Studying

Advanced article

Matthew S Starr, University of Massachusetts, Amherst, Massachusetts, USA
Keith Rayner, University of Massachusetts, Amherst, Massachusetts, USA

CONTENTS

Language comprehension processes are those by which people extract meaning from texts, whether written or spoken. Researchers have developed a variety of experimental methodologies in order to study these processes.

INTRODUCTION

The term 'cognitive science' has come to represent a wide variety of disciplines, each with its own methods and objectives. However, cognitive scientists as a whole are working towards the same ultimate goal – to discover how the complex processes of the mind operate. Given the somewhat enigmatic nature of human mental processes, one problem faced by researchers has been how to measure the activity of the mind. An early solution was to use introspection, wherein subjects were presented with a stimulus and were simply asked to relate their subsequent thoughts and feelings to the experimenter. Unfortunately, although some facts regarding the human mind may be ascertained through introspection, introspection alone proved to be largely uninformative. Introspective data tended to be: (1) highly variable across subjects; (2) frequently invalid; and (3) often strongly biased by individuals' preconceptions. What's more, many mental processes occur too quickly to be measured by conscious report, and some processes occur without any conscious awareness at all.

Given the early failure of introspection, cognitive scientists began to develop a number of tasks and paradigms – both generalized and specialized – in order to observe, record, interpret, and predict the activity of the mind. In this article we will discuss a number of such tasks and paradigms that have been utilized by researchers in order to study language comprehension processes. Specifically, we will delineate how each has been used to examine how people extract meaning from both written and spoken language. Since most of these techniques have been shown to have both strengths and weaknesses, we will also discuss some of the limitations inherent in various tasks.

REACTION TIME MEASURES

Reaction time measures are arguably the most common procedure for tapping into comprehension processes, and cognitive scientists generally use such measures to examine the relative time course of a process. In general, reaction time (RT) is defined as the interval between the presentation of a stimulus and the onset of the subject's subsequent response. This interval is typically measured with a high degree of precision (e.g. in milliseconds), and response types may vary from simply naming the stimulus to making a more complex decision, such as deciding whether two words are related in meaning to one another.

Naming and Lexical Decision

In the naming task, subjects are simply asked to articulate a word (or a pronounceable nonword) and reaction times are measured from the presentation of the stimuli to the onset of the named response. By contrast, in the lexical decision task, subjects are asked to decide whether a letter string is a word (e.g. DESK) or a nonword (e.g. DOSK), with reaction times measured from the presentation of the letter string to the onset of the word/nonword response. In the past, the most popular

usage of these tasks has been to determine the time course of visual word identification. For example, when extraneous factors such as word length and syntactic class are controlled, naming and lexical decision times for high frequency (more common) words are shorter than those for low frequency (less common) words. However, one problem with such tasks is that overall response time is not simply a measure of word identification, since both naming and lexical decision times also include the time it takes a subject to formulate and initiate the appropriate response for the task (i.e. an articulation or a manual button-press). Furthermore, it is not clear whether such responses even require the subject to identify the stimulus. The naming task, for example, simply requires the subject to pronounce a string of letters based upon grapheme-to-phoneme conversion rules without necessarily requiring that word meaning be accessed (e.g. most people can formulate a pronunciation for BLICKET, although no corresponding meaning exists in the lexicon). Similarly, in the lexical decision task, subjects may be able to judge whether a letter string is a word by simply basing their decision on the familiarity of the letter string rather than on the actual identification of the word. This does not necessarily mean that these tasks are insensitive to semantic properties of words – quite the contrary, naming and lexical decision tasks have been shown to exhibit effects of word frequency and familiarity, which would be unlikely unless some aspect of word meaning was accessed. Despite these limitations, response times in these tasks may be used to classify the upper limits of the time course for word recognition. However, many researchers have used naming and lexical decision tasks in conjunction with other tasks to determine whether the patterns of reaction times converge (see Taft, 1991).

Priming and Masking

Two methodologies have emerged which are often used in conjunction with naming and lexical decision. One paradigm, priming (Meyer and Schvaneveldt, 1971), typically involves the presentation of a sequence of two words: a prime followed by a target. Subjects are then asked to make a decision regarding the target word, and the researchers measure how quickly they are able to respond. One early finding that emerged from priming studies is that when the prime is semantically related to the target (e.g. DOG followed by CAT), subjects respond more quickly than when the prime is not semantically related to the target (e.g. PEN followed by CAT). This indicates that the relationship between prime and target words influences processing time on the target.

A second paradigm, masking, also examines word identification time by limiting the exposure of a stimulus. For example, the word DOG may be presented for 60 ms, at which time DOG disappears and is replaced by a pattern mask consisting of either a series of X's or of random letters. As with many reaction time measures, the masking paradigm has been useful in allowing researchers to examine the time course of lexical processing. Studies utilizing masking, for example, suggest that subjects may extract information from words which are presented for even very brief durations (e.g. 30 ms).

More recently, a paradigm which combines priming and masking procedures, masked priming, has been used to shed light on the early stages of word comprehension. In this paradigm, subjects look at a fixation target and a mask is presented followed by the brief presentation of a prime word which is then followed by a target word (presented for about 200 ms) which is in turn followed by another mask. Although subjects are generally unable to identify the prime word, it still has an effect on their report of the target word.

Dual Task

The dual task paradigm is often used to study attentional processes, and it follows two basic assumptions: (1) that subjects have a limited processing capacity, and (2) that different cognitive activities may make use of different processing resources. For example, subjects may be asked to read sentences while listening for a tone. If response times are slower when performing two tasks simultaneously as compared to performing a single task (e.g. simply reading sentences), this would be evidence that both tasks are drawing upon the same cognitive resources. Further, the rate of slowdown may also indicate the degree of resource utilization.

One example of the dual task paradigm is phoneme monitoring. Most often, phoneme monitoring is used to study speech comprehension, and it involves listening to auditorily presented sentences while monitoring for a particular phoneme (e.g. to detect the /ba/ sound while listening to the sentence 'The emperor went to the royal baths'). Thus the two tasks are to comprehend the sentence and to press a button when the target phoneme is detected. The idea is that if contextual or lexical processing prior to the target word (*baths* in this

example) is difficult, it should take subjects longer to detect a target phoneme. For example, subjects are slower to detect a phoneme when the target word is preceded by an ambiguous word.

Researchers utilizing this task are often interested in determining the basic units of speech perception or in studying the processing complexity of sentence contexts, lexical ambiguity, and attentional issues. However, the data emerging from phoneme monitoring tasks are often affected by a number of extraneous variables such as the frequency of targets across sentence stimuli, the discriminability of the phoneme, target word length, and the frequency of the target word in which the to-be-detected phoneme is located.

Speed–Accuracy Tradeoff

In addition to the dual task paradigm, another technique puts subjects under various types of speed constraint. For example, one variation of the technique involves training subjects to respond immediately upon the presentation of a signal that occurs at various times after the end of a sentence (which the subject reads and presses a button to indicate completion of reading). Accuracy of a decision made about a sentence increases as the response deadline increases. The manner in which reaction times and errors trade off against each other as the response deadline increases (or decreases) can therefore reveal information about processing activities. The major concern with this technique is that it may induce strategies that are specific to the demands of the task.

PROCESSING TIME AND OTHER MEASURES TO ASSESS COMPREHENSION

The reaction time measures discussed above are most commonly used when the unit of interest is a single stimulus (e.g. a word). Researchers interested in examining readers' comprehension of larger units, such as sentences or sentence phrases, may use one of a variety of processing time methodologies. By manipulating characteristics of the text, researchers can infer selected attributes of comprehension processes. In these tasks, subjects may be asked to read a paragraph/sentence while elapsed reading times are recorded (so that paragraph reading time or sentence reading time is measured). Similarly, subjects may be given a limited amount of time to read a portion of text while error rates are recorded.

Self-paced Reading

Sometimes, if an experimenter is interested in how long it takes a subject to read a segment of text, measuring overall reading times for sentences or paragraphs may be too imprecise. In the self-paced reading task, the experimenter controls the amount of text that the subject can see at any one time, and the size of the segment (e.g. a word or a clause) available to the subject is generally a function of the topic under investigation. When the subject has finished reading one segment, they push a button and the next segment of text is presented. When only one word at a time is presented, this procedure yields a processing time measure for each word in the text. A variation of this task is called the *makes sense* task, in which subjects advance word-by-word through a sentence as long as it makes sense. However, when the sentence no longer makes sense, or becomes ungrammatical, subjects push a different button.

One problem with the self-paced reading task is that it does not mimic natural reading. Reading times in the self-paced reading paradigm are slower (about half as fast) than those in more natural reading tasks since subjects must press a button to read subsequent segments of text. Since it takes longer to manually press a button than it does to move the eyes, words stay on the screen for about 400 ms in this task, as compared to average eye fixation times of approximately 250 ms in natural reading. Given that reading in the self-paced paradigm is slower in general, one possibility is that subjects may develop different comprehension strategies. It should also be noted that self-paced listening paradigms have been used to study speech perception. Similar concerns regarding strategic effects also apply to this paradigm.

RSVP

Natural silent reading involves moving the eyes to successive segments of text – hence the reader controls how quickly text is read. By contrast, in the rapid serial visual presentation (RSVP) task, the experimenter controls the rate at which text is presented. In this paradigm, the subject sits in front of a computer screen while new words are presented one at a time for various durations (e.g. 50 to 400 ms). Studies utilizing this technique have found that readers can comprehend short passages of text which are presented at rates of up to 1200 words per minute, with a new word being presented every 50 ms. Interestingly, when each word

is presented for 250 ms, reading comprehension in the RSVP task is often better than in natural reading.

This paradigm has several limitations. First, although comprehension performance is high for short passages of text, as the amount of text increases, comprehension begins to suffer. This is partially because RSVP reading disallows regressions back to previously read text, thus preventing readers from looking back at 'misunderstood' portions of text (during normal reading, readers make regressions on approximately 10 per cent of all fixations). Moreover, RSVP reading is also mentally taxing for subjects, as it requires their constant attention to text.

Phoneme Restoration

The self-paced reading and RSVP tasks described above are generally utilized to measure higher-order cognitive comprehension processes in reading. In contrast, the phoneme restoration effect has most commonly been used to measure lower-order, perceptual processing. The phoneme restoration effect is an auditory illusion that arises when part of an utterance is either deleted or replaced by an extraneous sound such as a cough or white noise. In such instances, listeners often perceptually fill in (restore) the missing phoneme and report that they heard the complete utterance. When utilized as a dependant measure in cognitive science, the phoneme restoration paradigm is used to examine the activation of sublexical phonemic structures (see Samuel, 1996 for a brief review of the literature). Early studies using this method found that psychoacoustic factors related to the nature of the replacement sound (e.g. amplitude, frequency) affected the probability of detecting the missing phoneme. Subsequent studies have shown that the restoration effect is also influenced by word length, lexical neighborhood size, and sentence context. More generally, subjects' ability to discriminate between an intact utterance and one in which a phoneme has been replaced by an extraneous noise is a function of perceptual processing difficulty – as the demand for processing resources increases, discriminability decreases.

Effects emerging from the phoneme restoration paradigm are typically replicable and valid, but the paradigm is sensitive to small changes in methodology. For example, differences in effects may arise between studies when experimenters utilize different forms of extraneous noise, even if the same independent variable is manipulated in both studies. Hence, experimenters need to take care in selecting the nature (e.g. cough, noise), phonemic class (e.g. vowels, stops), and position (e.g. initial, medial, final) of the replacement sound.

EYE MOVEMENTS

With the continuing development of technological innovations, some researchers have begun to replace (or supplant) reaction time and processing time paradigms with eye movement measures. In a typical eye-tracking experiment, subjects read sentences presented on a computer monitor while their eye movements are recorded. Researchers then look at patterns of readers' eye movements noting, for example, how long readers' eyes remain fixated on words or phrases within sentences, how far readers' eyes move from fixation to fixation, or how frequently readers' eyes regress back to reread text.

Eye movements have been utilized to study a variety of language comprehension processes, and data gleaned from eye-tracking studies have been found to reflect moment-to-moment cognitive processes. One early finding was that where readers look and how long they look there is directly related to the ease or difficulty of cognitive processing. For example, when extraneous factors are controlled, fixation times are longer for lower frequency words, which are less likely to be encountered during reading, as compared to higher frequency words, which are more likely to be encountered during reading. Eye movements have also been used to examine the effects of lexical ambiguity, morphological complexity, discourse processing, semantic relatedness, phonological processing, syntactic disambiguation, and the perceptual span (see Rayner, 1998 for an extensive review of the eye movement literature).

A number of methods have emerged within the eye-tracking paradigm including the development of the *eye-movement contingent display change* paradigm. In this paradigm, text displayed on a computer screen is manipulated as a function of where the eyes are fixated. As readers' eyes move across a line of text, letters or words may be modified in foveal, parafoveal, or peripheral locations, thus allowing the experimenter to control the nature and amount of information available to the reader. One variation of the eye-movement contingent paradigm is the moving window paradigm. In this paradigm, as readers move their eyes across the text, upon each fixation, text is exposed within an experimenter-defined 'window' while all text outside the window is altered in some way (e.g. all letters might be replaced by X's). Wherever the

reader looks, the text within the window is available. The logic of the paradigm is that when the window is as large as the region from which information can normally be obtained, reading will proceed as smoothly as when there is no window (normal text). Using this technique, the size of the perceptual span in reading has been determined.

Another variation is the boundary paradigm, in which characteristics of a target word in a particular location within a sentence may be manipulated. For example, in the sentence 'The man picked up an old map from the chart in the bedroom', when readers' eyes move past the space between *the* and *chart*, the target word *chart* would change to *chest*. In this manner, researchers can examine the types of information (e.g. orthographic, phonological, semantic) that readers obtained from the target word prior to fixating upon it.

A final variation is the fast-priming paradigm, in which a prime word is briefly presented for a very short duration (i.e. less than 50 ms) and is immediately replaced by a target word. Primes may be related in meaning to target words, but they may also be phonologically related (e.g. BAT-CAT) or orthographically related (e.g. BENCH-BEACH). This paradigm has been used to examine the time course of word processing.

One advantage of using eye movements over reaction time and processing time measures is that it allows researchers to study comprehension processes in a more natural setting. As mentioned previously, one disadvantage of reaction time and processing time measures is that they may result in the formulation of task-specific strategies or may simply slow the reading process. In the eye-movement paradigm, readers are free to read text as they would during normal reading. Moreover, eye-movement measures are flexible, allowing researchers to examine both fine-grain and coarse-grain language comprehension processes.

Finally eye-movement recording techniques have been utilized in the context of speech understanding. It has been demonstrated that when subjects listen to a narrative while a scene is presented in front of them which depicts objects in the narrative, their eyes tend to move to those objects that are mentioned in the narrative. This technique allows researchers to make inferences about on-line speech perception (see Tanenhaus and Spivey-Knowlton, 1996).

PHYSIOLOGICAL MEASURES

Since the 1980s, a number of physiological measures in addition to eye movements have been developed to study cognitive processes. These measures range from simply recording heart rate to recording more complex physiological activity, such as measuring changes in the magnetic activity of atoms within the brain. It is hoped that by using such measures, scientists will be able to accomplish two major goals: (1) to locate language comprehension regions (or pathways) in the brain, and (2) to more closely examine the time course of cognitive activity within the brain. Although there are a large number of physiological measures, in this section we will focus only on those measures which involve examining activity in the brain (see Gazzaniga, 2000 for a more complete review of physiological measures).

ERP

Among the most common physiological measures used today is event-related potentials (ERPs), which involves placing electrodes on the scalp. By averaging electrical potentials on the scalp over a number of trials, researchers hope to time-lock brain activity to a particular sensory event (e.g. the presentation of a word stimulus). The voltages associated with brain activity vary in both polarity and magnitude over time, resulting in a series of electrical 'peaks and valleys'. For example, when subjects are presented with a semantic incongruity, a relative large negative potential (i.e. a valley) occurs about 400 ms after the presentation of the stimulus (this is termed an N400 wave).

One advantage of using ERPs over other methodologies is that they may allow experimenters to more directly examine the time course of language comprehension processes within the brain itself. On the other hand, there is no guarantee that the ERP activity being measured is the direct result of a particular cognitive process, as opposed to being the result of later (e.g. memory) processing.

PET

Positron-emission tomography (PET) scans are based on a somewhat different framework than ERPs. This method involves the ingestion of a small amount of radioactive material which may be traced and used to measure blood flow in the brain; cognitive activity is represented by changes in blood flow to specific parts of the brain. Although the results of PET scans often involve complicated patterns of metabolic activity, studies using them have found that many different parts of the brain are involved in language comprehension (including

parts of the left temporal, parietal, and frontal cortex).

This complexity is perhaps the greatest disadvantage to the PET methodology. It is perhaps not surprising that language comprehension involves the coordination of a number of brain systems, but the metabolic activity measured by PET scans may also reflect additional processing not directly related to language. For example, researchers have found increased metabolic activity in brain systems which are not specific to language processing *per se*. Specifically, studies examining reading processes have found increased metabolic activity in the anterior cingulate cortex, which is normally associated with sustained attentional processing, as well as in the contralateral cerebellum, which is thought to be involved in the rapid shifting of attention.

MRI/*f*MRI

Magnetic resonance imaging (MRI) and its newest counterpart, functional magnetic resonance imaging (*f*MRI), are based on a framework similar to that of a PET scan – namely, that sensory, motor, and cognitive tasks produce a localized increase in neural activity which gives rise to subsequent increases in blood flow. MRI is very generally based upon the detection of electromagnetic signals which emanate from such increases in blood flow. Researchers utilizing MRI technology to examine language processes are typically interested in localizing language comprehension functions in the brain. For example, in a baseline condition an experimenter may present subjects with a word and simply require the subject to look at the word. In another condition, subjects may be asked to decide whether the word represents a living thing. Differences in neural activity between the two conditions can then be used to determine the region in the brain used in processing aspects of word meaning.

One advantage of the MRI/*f*MRI paradigm is that it is relatively non-invasive and represents little health risk to subjects (as opposed to PET scans which involve the ingestion of potentially harmful radioactive materials). In addition, they permit the experimenter to collect hundreds (or even thousands) of images from a single subject, with highly accurate spatial resolution. MRI technology is also becoming increasingly available to cognitive scientists, as many hospitals have MRI facilities.

The MRI/*f*MRI paradigm also suffers from several disadvantages. The most significant limitation is that temporal resolution is relatively poor (e.g. although it may only take about 250 ms to recognize a word, an *f*MRI can only acquire data in about two to three seconds), thus disallowing any clear examination of the time course of language processing. However, some scientists have also begun to combine the temporal resolution of ERPs with the spatial resolution of *f*MRIs. In addition, as mentioned earlier in reference to PET scans, a great deal of activity in the brain occurs which is only indirectly related to language functions, resulting in some degree of difficulty in localizing areas in the brain specific to language comprehension.

Finally, in addition to the recent *f*MRI methodology, cognitive scientists have begun to explore the potential of a relatively new physiological measure. The magnetoencephalogram (MEG) measures the magnetic fields which result from the electrical currents inherent in neurons. The advantage to MEG is that the sources of magnetic currents are more easily localized than are the sources of electrical currents (as measured by a more simple electroencephalogram), and the hope is that MEG measures will allow a more accurate topographic mapping of brain function.

CONCLUDING REMARKS

Given the rapid, complex, and sometimes unconscious nature of human mental processes, cognitive scientists have been forced to develop a growing number of empirical methods in order to study the activity of the mind. Analogous to the telescope of the astronomer or the electron microscope of the chemist, the tools and methods of cognitive science serve to enhance our limited physical senses, and they allow us to observe cognitive processes which were previously beyond our grasp.

References

Gazzaniga MS (2000) *Cognitive Neuroscience: A Reader*. Oxford, UK: Blackwell Publishers.

Meyer DE and Schvaneveldt RW (1971) Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* **90**: 227–234.

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* **124**: 372–422.

Samuel A (1996) Phoneme restoration. *Language and Cognitive Processes* **11**: 647–653.

Taft M (1991) *Reading and the Mental Lexicon*. London: Lawrence Erlbaum Associates.

Tanenhaus MK and Spivey-Knowlton MJ (1996) Eye-tracking. *Language and Cognitive Processes* **11**: 583–588.

Further Reading

Just MA and Carpenter PA (1987) *The Psychology of Reading and Language Comprehension*. Boston, MA: Allyn and Bacon.

Pollatsek A and Rayner K (1998) Behavioral experimentation. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*. Oxford, UK: Blackwell Publishers.

Posner MI (1989) *Foundations of Cognitive Science*. London: MIT Press.

Rayner K and Pollatsek A (1989) *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice Hall.

Schilling HEH, Rayner K and Chumbley JI (1998) Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition* **26**(6): 1270–1281.

Language Comprehension

Introductory article

Morton Ann Gernsbacher, University of Wisconsin, Madison, Wisconsin, USA
Michael P Kaschak, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Language comprehension is the process whereby linguistic input (spoken or written) is transformed into meaning.

INTRODUCTION

Language comprehension is the process whereby information is extracted from speech or written text, resulting in the apprehension of the meaning of the language. Most research on this topic has focused on the comprehension of written text, rather than conversations. Despite the complexity of the processing involved, most humans are able to comprehend language quickly and efficiently. This efficiency is, in part, due to two types of knowledge that can be brought to bear on the task of determining the meaning of an utterance. The first is linguistic knowledge: knowledge of the syntax and lexical items of the language. Syntax provides the comprehender with cues as to which words go together, and as to who is doing what to whom in the sentence. The lexical items refer to the specific objects, actions, states, or relations between things that are being discussed in the sentence. The second source of information is what has been called 'world knowledge', or knowledge about how the world works. This knowledge is useful in filling in details of events and situations that are underspecified in the linguistic input. (*See* **Syntax**; **Lexicon**)

INTEGRATING MEANING AND GRAMMATICAL INFORMATION

The process of comprehending language might be described as follows. First, words must be recognized from the perceptual input (spoken or written) to the comprehender. Then, the comprehender must determine how these words are arranged into sentences. One step in this process is to try to analyze the string of words into a series of chunks, called phrases. This analysis is performed according to the comprehender's knowledge of the syntax of his or her language. (*See* **Parsing**)

Syntax is a set of rules that constrains the ways in which the words in a language can be fitted together to form a sentence. For example, the syntax of English provides the following rough stipulations about how words can be combined:

- Noun phrases (NP) are phrases in which a noun may be preceded by a determiner (*a*, *an*, *the*) or adjective (e.g., *the big red truck*, *the truck*, *red trucks*).
- Verb phrases (VP) are phrases in which a verb may be followed by a noun phrase or prepositional phrase (e.g., *kicked the ball*, *threw the ball over the fence*).
- Sentences contain noun phrases and verb phrases, and typically take the form NP–VP (e.g., *the big red truck bumped into the small white car*).

On the basis of one's knowledge of syntax and one's knowledge about the words in the sentence, a sentence like *the big red truck bumped into the small white car* can be broken down into the phrases *the big red truck* and *bumped into the small white car*. This is an important step in language comprehension, because this grouping of the input into phrases ('parsing') tells the comprehender who is doing what to whom in the sentence. In this example, the syntax of the sentence specifies that the truck bumped into the car. If the phrases are changed around (*the small white car bumped into the big red truck*), the relationship of the objects mentioned in the sentence changes. The parsing of a sentence can sometimes be difficult, but once the comprehender has determined how the linguistic input can be broken down into phrases, he or she can apply knowledge of the syntax of the language to get an initial sense of the action described in the sentence.

CHOOSING THE CORRECT SENSE OF A WORD

Parsing is but one component of the initial processing of speech or written language. Indeed, the comprehender cannot group a string of words into phrases until he or she knows what the words are, whether the words are nouns, verbs, adjectives, and so on. Beyond knowing what part of speech each word is, the comprehender also needs to retrieve information about what the word means. Determining what a word means can be difficult in some cases. Some words, such as *bug*, have many meanings ('insect', 'surveillance device', 'bother'). Given a sentence that begins *there was a bug …*, which meaning of *bug* should be chosen?

Research has shown that multiple meanings of *bug* are often initially activated by the language comprehension system. Then, within milliseconds, the appropriate meaning is selected, and the other meanings are discarded. The inappropriate meanings are deactivated through a process known as *suppression*. When information is suppressed, it is inhibited from being accessed. By dampening the activation of the inappropriate meanings, the language comprehension system ensures that the correct meaning will be available for generating the meaning of the sentence.

Sometimes, information is chosen through the opposite process, known as *enhancement*. Information that is enhanced is given a boost in activation relative to other available information. For instance, a story might be written that includes several characters interacting: Joan, Mary, Simon, and Jane. If the story repeatedly mentions Mary and the things that Mary is doing, her representation will be enhanced relative to those of Joan, Simon, and Jane. At some point in the story, the pronoun *she* might appear. To whom does *she* refer: Joan, Mary, or Jane? Because Mary's representation is enhanced relative to those of Joan or Jane, the comprehender will assume that *she* refers to Mary.

Thus, by selectively suppressing and enhancing information, comprehenders can select the appropriate meanings of the words they encounter. These word meanings, in conjunction with the comprehender's syntactic analysis, can be used to build the meaning of the sentence in which they appear.

BRIDGING INFERENCES

The language that we process is typically not in the form of isolated sentences like *the big red truck bumped into the small white car*. Rather, we encounter language within a rich context, such as a story, a conversation, a description, or a teaching session. We have considered some ways in which a single sentence could be understood; but how is a series of sentences understood as a whole? (*See* **Discourse Processing**)

Consider the following pair of sentences: *It was a cold winter morning. As Harriet made her way down the front steps, she fell and broke her leg*. These sentences are syntactically independent (it was cold outside; a person named Harriet broke her leg); however, most comprehenders will assume that they relate to the same situation. In fact, most readers will interpret the two sentences as implying that there may have been ice or snow on the front steps, which caused Harriet to fall. This information is nowhere to be found in the two sentences; so readers must use their background knowledge of winter weather to infer it from the information that is explicitly stated. This inference is called a *bridging inference*. It is generated in order to connect the information presented in different sentences.

Bridging inferences arise when readers make use of their world knowledge in interpreting language. World knowledge refers to a general store of information that humans possess about the way the world operates. For example, we know that birds fly to get from one place to another, but dogs walk or run. We know that people ordinarily make their way down front steps easily, but may slip and fall if the steps are icy. When readers encounter language about a particular situation, such as walking down steps on a cold winter morning, they are able to use their knowledge about such situations to go beyond the information literally presented in the sentences in order to build an understanding of what the language is communicating.

AUTOMATICITY OF INFERENCE GENERATION

We know that inferences are generated during the reading process, but there is some debate as to when inferences are generated automatically and when they are generated through some effort. In the example above, one can ask if the inference that the front steps were icy is automatically generated by readers, or if it is generated only in certain conditions, such as when the reader is being particularly attentive to the details of a story.

There are two basic positions in this debate. According to the 'minimalist' position, a comprehender draws an inference only when either the information leading to that inference is highly available, or the inference is necessary to maintain

the coherence of the text. According to the 'constructivist' position, a wide range of inferences are generated as a matter of course during reading, not just the minimum set of inferences needed for understanding. It is likely that the correct resolution of this debate is an intermediate position. Readers will make the inferences necessary to meet their goals in processing the text. Thus, if they are only skimming the text to get the gist of what happened, they will make few inferences. If they are reading the text in detail, they will make a more elaborate set of inferences.

BUILDING DISCOURSE REPRESENTATIONS

As a series of sentences is processed, the language comprehension system accumulates the information presented across the sentences into a discourse representation. This representation, often called a 'mental model', is a cumulative record of the information and events presented in the text. Mental models are representations of what the language is about, rather than representations of the structure (e.g. syntax) and lexical information in the linguistic input. (*See* **Mental Models**)

Mental models are built by creating a representation in memory into which information from the present situation can be inserted. This representation may be described metaphorically as a mental 'stage'. At the beginning of a story, a new mental stage is created. As characters are introduced in the story, they are placed on the mental stage. The actions of the characters are performed on the stage. By updating the mental stage with information from successive sentences in the story, the reader can keep track of who has done what to whom and where the characters are at the present time. In this way, the reader is prepared to receive information about what happens next. The same general processes operate in building mental models of texts that are not stories.

Mental models appear to be particularly sensitive to spatial and temporal information. Thus, in keeping a record of what happens on the mental stage, readers can keep track of which things (or people) are close to each other, and which are not. They can also keep track of which events belong to a single temporal unit (e.g. what happened when two characters went to the movies), and which events belong to different units (e.g. what happened on Thursday and what happened on Sunday). The selection of what information is tracked and what is not appears to be influenced by the goals of the reader.

INDIVIDUAL DIFFERENCES IN COMPREHENSION

Most humans learn to process language quickly and efficiently. But there are noticeable differences in how quickly and efficiently different people comprehend language. There are two basic explanations of this.

According to the 'capacity' account, people differ in the capacity of their working memory. Working memory is important in language comprehension because words and syntactic structures need to be held in memory until a given sentence has been processed fully. In addition, working memory is needed to manipulate information in mental models. Readers who have a large working memory capacity will be able to handle a lot of information when processing sentences or building mental models, and will thus be good at language comprehension. Readers with smaller capacity will lose some information from working memory during comprehension, and will consequently be less efficient in language processing. (*See* **Language Comprehension and Verbal Working Memory**)

According to the 'skill' account, better readers are those who have more experience reading, and thus have more skill in performing the mental computations necessary for comprehension. Although these views are often presented as opposing theories, it is likely that they are both correct in some sense. Language comprehension undoubtedly relies on keeping information in working memory, and so readers with better or bigger working memories will be better comprehenders. At the same time, it is likely that the reason why some readers can store more information is that they are more skilled at manipulating the information that is maintained in working memory.

Further Reading

Gernsbacher MA (1995) The Structure Building Framework: what it is, what it might also be, and why. In: Britton BK and Graesser AC (eds) *Models of Text Understanding*, pp. 289–311. Hillsdale, NJ: Lawrence Erlbaum.

Glenberg AM, Meyer M and Lindem K (1987) Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language* **26**: 69–83.

Graesser AC and Kreuz RJ (1993) A theory of inference generation during text comprehension. *Discourse Processes* **16**: 145–160.

Johnson-Laird PN (1983) *Mental Models*. Cambridge, MA: Harvard University Press.

Just MA and Carpenter PA (1992) A capacity theory of comprehension: individual differences in working memory. *Psychological Review* **99**: 122–149.

Kintsch W (1988) The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review* **95**: 163–182.

MacDonald MC, Pearlmutter NJ and Seidenberg MS (1994) Lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703.

McKoon G and Ratcliff R (1992) Inferences during reading. *Psychological Review* **99**: 440–466.

Singer M, Andrusiak P, Reisdorf P and Black NL (1992) Individual differences in bridging inference processes. *Memory and Cognition* **20**: 539–548.

Language Disorders

Intermediate article

Yosef Grodzinsky, Tel Aviv University, Tel Aviv, Israel; McGill University, Montréal, Canada

CONTENTS

Language disorders (as distinct from communicative deficits) provide an important tool for research on brain–language relations.

INTRODUCTION

The neurosciences seek to understand how the healthy brain works. What, then, is the reason to study it in disease? Language as a cognitive capacity is investigated intensively; yet under what guise do disturbances to language and speech become a topic for scientific inquiry? One undisputed goal is remedial: we study pathologies in order to cure them. Yet there is another, less obvious goal: We would like to study language diseases as a vehicle for biologically-based componential analyses of the human language faculty. Deficit analyses have been an extremely valuable research tool for over a century, providing a critical testing ground for theories of brain–language relations.

This article begins with an attempt to situate the language faculty within the human communicative system. This allows us to distinguish between linguistic disorders and other communicative pathologies. A selective review of exemplary pathological phenomena follows. We conclude with a discussion of the relevance of these phenomena to neurobiological and (psycho)linguistic approaches to natural language.

COMPONENTS OF COMMUNICATION

Communication involves interlocutors, communicative means, and media. For a successful communicative act, at least two interlocutors, a common code, a medium in which signals in this code can travel, and a common universe of discourse, are prerequisite. In humans, the language faculty constitutes a part of this scheme. Communication is fragile, as each part of the systems is liable to problems: a deaf ear, a broken wire in a telephone conversation, or a gap in the knowledge that the interlocutors share, are among the many possible disruptions that might bring about communicative breakdown. Note that communicative disorders need not be consequences of deficits in the language faculty. While most public aspects of our mental life are conveyed through language, a linguistically manifested disorder is not necessarily a language disorder. It is important, therefore, to be explicit about what language is, so that a delineation of its disorders would follow. We begin with a brief sketch of the language faculty as conceived by some major theoretical frameworks.

The Structure of Linguistically Represented Knowledge

The human language faculty consists of a system of knowledge, and neural mechanisms that implement it using dedicated algorithms. Briefly, the knowledge consists of a finite set of elements (symbols), and a finite set of combinatorial rules, which concatenate the elements into strings (phonetic units into phonemes, morphemes, words, phrases, sentences, and so on up to meaningful expressions and stories). This knowledge is multifaceted: it is divided into distinct levels in which formal rule systems operate over structured inventories of elements – phonetics, phonology, morphology, syntax, and semantics. Thus a speaker possesses a universal phonetics, a system of phonological elements from which sound sequences in his or her language can be created (and analyzed), a system of morphological elements and rules for words, a syntax, and a system for interpretation. This knowledge is formal, specific to the language capacity (as distinct from other cognitive capacities), and

cerebrally represented. It constitutes a biological module, putatively distinct from other knowledge bases and mental processes. This is the 'mental organ for language', for which Chomsky has argued so vigorously (e.g. Chomsky, 1995). (*See* **Government–Binding Theory**; **Morphology**; **Syntax**; **Phonology**; **Learnability Theory**; **Phonetics**)

During linguistic communication, the speaker puts to use mechanisms that regulate the production of utterances, and mechanisms involved in their reception. Yet these do not operate in isolation: a communicative act cannot be successful unless a host of nonlinguistic communicative skills are also invoked. Although the main vehicle for communication is language, other abilities are required: the abilities to integrate a sequence of sentences into a story, to make eye contact and take turns in conversation, to make inferences, to gesture, and to grasp irony, are all highly relevant to an individual's communicative functioning; and, while they interact with linguistic principles, they are neither governed by them nor mediated by linguistic processes in the head. Knowledge of the world is also linguistically communicated, but is not part of language.

This distinction is critical for a proper understanding of the nature of communication. In the present context, the specificity of the language domain demonstrates the communicative, rather than linguistic, nature of a number of linguistically manifested disorders. The deep involvement of language in cognition sometimes blurs the distinction between the two types, but we have ample scientific as well as clinical reasons to make it. Before we focus on true language disorders, we will provide a short survey of communicative disorders that are not linguistic.

LINGUISTICALLY MANIFESTED DISORDERS THAT ARE NOT LANGUAGE DISORDERS

Among the disorders that affect linguistic communication there are several pathologies for which a true language disorder can be ruled out with relative confidence. In some cases, direct language testing has determined the intactness of the language faculty; in others, indirect test results, or clinical descriptions, lead to a similar conclusion.

Autosomal Dominant Speech Disorder

Linguistic geneticists believe that they found a 'language gene'. This is mostly due to the discovery of the KE family in Scotland, whose members' phenotypes manifest a severe disturbance of speech and language. The disorder seems to follow from an autosomal dominant trait: half the family members in three generations (15 out of a total of 31 individuals, of both sexes) are affected; the offspring of an unaffected member are unaffected; and half the offspring of an affected member are affected. A series of sophisticated molecular and bioinformatic techniques were used to locate the disorder in a region on chromosome 7. Researchers have administered a variety of speech, language, and other cognitive tests in an attempt to distinguish impaired from healthy members. Yet among the many tests (e.g. of memory, lexical decision, naming, grammar, and meaning), only three established a 'characterizing phenotype', i.e. a deficit shown by every affected member but not by any who are unaffected. These deficits related to: word repetition; non-word repetition; and simultaneous and sequential orofacial movement to command (Lai *et al*., 2001; Vargha-Khadem *et al*., 1998). None of these tests pertains to an essentially linguistic rule system, or to components thereof. For all we know, then, this genetic disorder is not linguistic, although it is has a very adverse effect on linguistic communication.

Williams Syndrome

Chromosome 7 appears to be a locus related to communication. Williams syndrome also results from mutations in it, in this case spontaneous deletions. This syndrome is often mentioned in the context of language disorders. But it may instead be an extreme case of linguistic modularity: that is, outstanding linguistic ability accompanies moderate mental retardation. Consider these utterances, recorded from Williams syndrome patients by Ursula Bellugi and her team (Bellugi *et al*., 1998, p. 183):

> 'When I got up the next morning, I talked but I couldn't say anything so my mom had to rush me to the hospital.'

> 'There is a huge magnetic machine. It took a picture inside the brain. You could talk but not move your head because that would ruin the whole thing and they would have to start all over again. After it's all done they show you your brain on a computer and they see how large it is. And the machine on the other side of the room takes pictures from the computer. They can take pictures instantly. Oh, and it was very exciting!'

An impaired language faculty could hardly produce sentences like these (note that they are

phonetically well formed – the speech of Williams syndrome children is not garbled). Upon formal testing of their receptive and productive abilities, these patients show a linguistic ability on a par with normal children of their mental age, suggesting that their impairment is not linguistic (Zukowski, 2000). (*See* **Williams Syndrome**)

Autism

Autistic children are usually diagnosed on the basis of impaired development in communication, social interaction, and action. Given this broad range of problems, is language a primary source of their deficit? The available evidence suggests that the problem manifests, at least in part, as an inability to attribute beliefs to others, or to see things from a perspective other than one's own. This is the well-known 'theory of mind' disorder. The evidence concerning language appears contradictory. In at least one major study, autistic children were indistinguishable from controls in their language; another major study, however, documented many differences in language skills: children (who were normal in the domain of articulation) varied greatly in their ability in the various grammatical domains (Kjelgaard and Tager-Flusberg, 2000). The conflict among these studies, and the degree of variation, suggest that language is sometimes affected in autism and sometimes not. Even when it is, the nature and degree of the impairment are not clear. It is difficult to reach a definite conclusion on the basis of such evidence (which is a result of sound research on large numbers of cases). It may be that the linguistic disturbance is a consequence of a more general cognitive impairment that is projected onto communicative skills and hampers aspects of language use. To classify autism as a language disorder would thus be misleading.

Schizophrenia

Standard descriptions of schizophrenia explicitly mention a language disorder as part of the symptom complex. For example, Andreasen (2000, p. 107) states:

> Schizophrenia is characterized by symptoms that reflect multiple mental processes: hallucinations, or abnormalities in perception; delusions, or abnormalities in inferential thinking; disorganized speech, or abnormalities in language; disorganized behavior, or abnormalities in behavioral monitoring and control.

Upon closer examination, the relevant aspect of the disturbance looks different: while productive abilities of schizophrenic patients appear impoverished, they do not appear to be linguistically ill-formed (either grammatically or semantically). On the view of language described above, the cognitive disorder from which these patients suffer is not linguistic. They may use fewer words than normal, but no phonological or morphological errors are reported; their sentences may be short and low on content, but incorrect syntax or incoherent meaning are not central features of the disease. Some descriptions refer to an inability to make proper inferences, a 'negative thought disorder', failures to integrate texts, and 'discourse coherence' disturbances (somewhat reminiscent of right-hemisphere-damaged patients), but there is nothing in their communicative problem that suggests a disturbance to language mechanisms (Barsch and Berenbaum, 1997).

CONGENITAL AND DEVELOPMENTAL PATHOLOGIES THAT ARE LINGUISTIC

Dyslexia and Developmental Reading Disorder

Some reading disorders have been explained in terms of deficits in phonological awareness. The argument is that because alphabetic writing systems require that the reader and writer have awareness that words are made up of individual speech sounds (i.e. phonemes), insufficient awareness of the speech segments in spoken words would impede learning to read and write. Young children and poor readers have been shown to be lacking in awareness of the sound structure of words, as demonstrated by failures on metaphonological tasks. It is now known that very young children lack phonological awareness; that they acquire it over several years; and that many children when they begin school are deficient in phoneme awareness and find the writing system difficult to acquire. This problem has been related to observed deficiencies in auditory processing, which have been increasingly documented in reading-disabled children.

It is not clear what the causes of this failure are. One influential approach has attributed developmental reading (and some writing) disorders to a processing limitation. On this view, there is an informational 'bottleneck' in the course of processing from form to meaning. This bottleneck can manifest in phonological disorder in spoken language comprehension, or in a reading disorder when an orthographic code is involved. Poor readers have been found to fail in tasks that require auditory comprehension of syntactically complex

sentences (e.g. relative clauses such as *point to the butterfly that the kangaroo is flying over*). More generally, it is claimed that limited processing resources (a bottleneck) may affect all levels of linguistic analysis, reading being a special, more tangible case (Shankweiler and Crain, 1986).

The Neurobiology of Developmental Dyslexia

The co-occurrence of reading disorder, possible auditory processing deficiencies, reduced right–left asymmetries in a cerebral region (planum temporale), and cortical malformations at the cytoarchitectonic level (microgyria), has inspired research into the neural basis of dyslexia. A group of researchers led by Norman Geschwind and Albert Galaburda have developed methods to model such neural defects in laboratory animals (although obviously the animals cannot read). The modeling focuses on certain properties of brain cells in animals and humans, similar in health and disease, in ways that may lead to a deeper understanding of the neural mechanisms that underlie congenital dyslexia. These models currently point to a neural antecedent of an auditory processing deficit that presumably causes dyslexia in humans (Rosen *et al*., 2001). This line of research is promising, and has attracted a lot of interest. It remains to be shown how this defect is projected further up the language processing system, to cause problems located at higher levels of language processing, most notably problems of sentence comprehension. An explanation of this would represent a major advance towards a truely biological model of language processing.

Specific Language Impairment

There is a fairly large group of children whose language abilities in all modalities remain immature while their peers master language fully. This is called 'specific language impairment' (SLI) or child dysphasia. SLI children are generally a rather heterogeneous group, but certain interesting subgroups have been identified, and investigated extensively. One such subgroup tends to omit, or substitute, certain grammatical morphemes, despite having normal intelligence and relatively few deficits in the phonological or semantic domains. This is observed in many languages, in forms that, by and large, fall under a small set of generalizations. The fact that many of these children are otherwise normal, and suffer from a highly specific, grammatical deficit, implies that this is a true language disorder. Moreover, there is evidence that certain aspects of the disturbance are in fact delays in normal grammatical (as opposed to cognitive) development. (*See* **Developmental Disorders of Language**)

Three questions remain: (a) what is the reason for this disorder, (b) what is its correct characterization, and (c) how can it be fixed. Regarding (a), while no biological element (brain area, or a gene) has been tied to SLI directly, there are some reasons to believe that there is a genetic component in SLI. As for (b), one influential approach has proposed that SLI children are delayed in that a stage that is fairly short for non-SLI children – a stage in which certain grammatical rules are optional – is extended in SLI for a long period of time. This has led to (c) attempts to help these children overcome the difficulty by helping them, through remedial teaching, to get beyond optionality (Leonard, 1998).

ACQUIRED LINGUISTIC PATHOLOGIES

Aphasia

Acquired aphasia is a collection of linguistic impairments, mostly subsequent to organic brain disease. Central to this extensively studied group of disorders are impairments due to focal brain lesion. Focal lesions (common etiologies of which are stroke, intracranial hemorrhage, protrusion wound, and brain tumor) are contrasted with diffuse lesions, due to neurodegenerative disease (e.g. Alzheimer's disease), anoxia, and other circumstances that affect brain tissue in a nonfocal fashion. Scientifically, the most important aspect of the aphasias is their selectivity: a well-delineated brain region is affected, bringing about a selective functional impairment. Localization, accompanied by a precise characterization of the functional damage, might improve our understanding of the brain mechanisms of language, as well as our ability to restore them.

Injury to several different cerebral locations brings about language problems of different kinds. Here we will focus on limited aspects of a narrow range of pathologies, mostly those interpretable within current theories of linguistic representation and processing: namely, the Broca's and Wernicke's aphasias, which are subsequent to lesions in the frontal and temporal lobes, respectively, of the left cerebral hemisphere, and are traditionally characterized as problems of language production and reception, respectively.

The language 'centers' were initially thought of as governing receptive or expressive modalities.

But as research progressed it became clear that the emphasis should be placed on grammatical rule systems. It was shown that despite the importance of the channels through which language is practiced, the correct (and most telling) unit of analysis for the interpretation of lesion data is the rule type. This applies to phonology, syntax, and semantics. Activities and tasks no doubt play a mediating role in linguistic communication, yet the defining characteristic of the language faculty is the rule systems it possesses. How these rule systems are instantiated in neural tissue then became the central question in neurolinguistics. (*See* **Aphasia**)

In the domain of language production, the brain makes fine distinctions among rule types: Broca's aphasics are deficient in producing tense inflection, but intact in agreement inflection. This has been demonstrated for a wide variety of languages (Friedmann and Grodzinsky, 2000). Cross-linguistic studies further indicate that not only the inflection type of a verb, but also its position in the sentence, determines whether it is produced subsequent to a lesion in Broca's region. (*See* **Tense and Aspect**; **Agreement**)

In receptive language, the distinction between transformational and nontransformational sentences yields a big performance contrast: aphasics with lesions in Broca's region understand active sentences, subject relatives, subject questions, and the like, normally, yet fail on their transformational counterparts: passives, object relatives and questions, and so on (Zurif, 1995). This has led some to the more general hypothesis that in receptive language, Broca's aphasics are unable to compute transformational relations. According to this hypothesis (known as the trace-deletion hypothesis (Grodzinsky, 1986, 2000)), critical components of the mechanisms that underlie transformational analysis are localized in Broca's region. Furthermore, the highly selective character of this deficit has major implications for linguistic theory and the theory of sentence processing. (*See* **Constraints on Movement**)

A particularly compelling argument that supports the localization of transformations in Broca's region comes from cross-linguistic comparisons: Chinese, Japanese, German, Dutch, Spanish and Hebrew have different properties, and the performance of Broca's aphasics is determined by the trace-deletion hypothesis according to the particular grammar of the language. In English, aphasics comprehend active sentence correctly. Yet the finding are different in Japanese, which has two types of actives: *Taro-ga Hanako-o nagutta* ('Taro hit Hanako' – subject, object, verb), and *Hanako-o Taro-ga nagutta* (object, subject, verb). These constructions are simple; they mean the same thing and are identical except that the latter is derived transformationally, with the object moved to the left edge of the sentence. Broca's aphasics handle the subject–object–verb type correctly, but are at chance level on the object–subject–verb type. (*See* **Local Dependencies and Word-order Variation**)

Chinese is generally a subject–verb–object language like English, but heads of relative clauses follow the relative, unlike English in which they precede it. This correlates perfectly with the cross-linguistic results in aphasia: subject relatives are comprehended at chance level in Chinese (sentence 1a) and above chance level in English (sentence 1b), whereas object relatives yield the opposite pattern (sentences 2a and 2b).

a. [_zhuei gou] de **mau** hen da — ***chance***
   chase dog that cat very big
b. **The cat** that [_chased the — ***above chance***
   dog] was very big (1)

a. [mau zhuei_] de **gou** hen xiao — ***above chance***
   cat chased that dog very small
b. **The dog** that [the cat chased_] — ***chance***
   was very small (2)

English and Chinese relative clauses thus give rise to a fascinating mirror-image result pattern, which correlates with a similar syntactic contrast between the two languages. Other interesting cross-linguistic contrasts have also been documented, providing further evidence that Broca's region is critically involved in transformational analysis. Moreover, reflections of the same disruption are also found in the domain of real-time processing (Zurif, 1995). There are further results regarding grammatical aspects of the mental lexicon (Shapiro *et al*., 1993), which are also localizable, as they appear to be retained in Broca's aphasia but severely disrupted after a lesion in Wernicke's area (Grodzinsky, 2000).

While Broca's region appears to govern syntactic knowledge and algorithms that implement it in use, Wernicke's region governs some combinatorial semantic operations. This has been demonstrated by tests of semantic composition that have been administered to both Broca's and Wernicke's aphasic patients. Phonological impairments subsequent to focal brain damage also exist. Thus, parts of the left hemisphere are now believed to govern aspects of combinatorial linguistic knowledge at all levels.

Other Acquired Disorders

Brain disease – whether focal or diffuse – has many possible causes, manifestations and locations. It can produce a multitude of cognitive and affective disorders, many of which have linguistic manifestations. These include the alexias and agraphias, and a variety of pragmatic disorders due to damage to the right hemisphere of the human brain. There are also diseases that bring about diffuse lesions, such as the various dementias. Because these affect multiple cognitive functions, our ability to discern linguistic from other impairments in this class of diseases is severely limited. (*See* **Language and Brain**)

RELATING BRAIN AND LANGUAGE

We have discussed language disorders, but said little about ways to study them. A central and traditional method of investigation is based on error analysis. Most of the results discussed in this article are of this type. The idea is to compare error rates on a given task for deficient and normal speakers, and then try to infer the site of the functional damage.

There are other methods: differences in reaction time, in electroencephalographic (EEG, MEG) and electromagnetic (fMRI) measurements, in measures that involve neuroradiation (PET), and in other kinds of precise measures of brain activity, are used in the effort to understand the brain mechanisms for language and communication. Studies of disordered speakers are conducted in parallel with investigations of intact language users, in the hope of providing not only remedy to the former, but also a comprehensive theory about the latter.

Such a theory is not easy to reach. Any theoretical consideration of brain–behavior relations begins with a complex equation: on one side there are properties of neural tissue, and on the other side, behavior. The relationship between the two sides is highly complex, and understanding it requires, as a first step, a clear set of ideas regarding the proper description of each side. Thus, in the case of language, a comprehensive and coherent picture of what language is – i.e. a general linguistic theory – is a prerequisite.

In this context, the disorders discussed in this article have a dual role. Firstly, they point to natural biological classes among linguistic types. Thus, if brain damage undermines linguistic performance in a particular way (say, it impairs the ability to process inflections), then the line separating the impaired from the preserved must feature in any biologically feasible account of language. Secondly, the patterns of impairment and sparing observed in the various disorders support the claim that a precise neurological account of the language regions in the brain must be based on linguistic concepts. A theory of brain–language relations, then, must be both linguistically and neurologically adequate. This makes the investigation difficult, but interesting and rewarding.

References

Andreasen NC (2000) Schizophrenia: the fundamental questions. *Brain Research Reviews* **31**: 106–112.

Barsch D and Berenbaum H (1997) The effect of language production manipulations on thought disorders and discourse coherence disturbances in schizophrenia. *Psychiatry Research* **71**: 115–127.

Bellugi U, Marks S, Bihrle A and Sabo H (1988) Dissociation between language and cognitive functions in Williams Syndrome. In: Bishop D and Mogford K (eds) *Language Development in Exceptional Circumstances*, pp. 177–189. Hillsdale, NJ: Erlbaum.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Friedmann N and Grodzinsky Y (2000) Split inflection in neurolinguistics. In: Friedemann M and Rizzi L (eds) *The Acquisition of Syntax*, pp. 84–104. London: Longman.

Grodzinsky Y (1986) Language deficits and the theory of syntax. *Brain and Language* **27**: 135–159.

Grodzinsky Y (2000) The neurology of syntax. *Behavioral and Brain Sciences* **23**: 1–71.

Kjelgaard M and Tager-Flusberg H (2000) An investigation of language impairment in autism: implications for genetic subgroups. *Language and Cognitive Processes* **15**: 287–308.

Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F and Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**: 519–523.

Leonard L (1998) *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.

Rosen G, Fitch RM, Clark MG *et al*. (2001) Animal models of developmental dyslexia: is there a link between neocortical malformations and defects in rapid auditory processing? In: Wolf M (ed.) *Time, Fluency, and Dyslexia*, Timmonium, MD: York Press.

Shankweiler D and Crain S (1986) Language mechanisms and reading disorder: a modular approach. *Cognition* **24**: 139–168.

Shapiro LP, Gordon B, Hack N and Killackey J (1993) Verb-argument structure processing in complex sentences in Broca's and Wernicke's aphasia. *Brain and Language* **45**: 423–447.

Vargha-Khadem F, Watkins KE, Price CJ *et al*. (1998) Neural basis of an inherited speech and language disorder. *Proceedings of the National Academy of Sciences* **95**: 12695–12700.

Zukowski A (2000) *Uncovering Grammatical Competence in Children with Williams Syndrome*. PhD thesis, Boston University.
Zurif EB (1995) Brain regions of relevance to syntactic processing. In: Gleitman L and Liberman M (eds) *An Invitation to Cognitive Science*, 2nd edn, vol. I, pp. 381–397. Cambridge, MA: MIT Press.

Further Reading

Avrutin S (2001) Linguistics and agrammatism. *GLOT International* **5**: 3–11.
Bellugi U (ed.) (2001) *Journey from Cognition to Brain to Gene*. Cambridge, MA: MIT Press.
Brown C and Hagoort P (eds) (2000) *The Neurocognition of Language*. New York, NY: Oxford University Press.
Crain S, Shankweiler D and Ni W (2001) Grammatism. *Brain and Language* **77**: 294–304.
Friedmann N and Grodzinsky Y (1997) Tense and agreement in agrammatic production: pruning the syntactic tree. *Brain and Language* **56**: 397–425.
Galaburda A (ed.) (1993) *Dyslexia and Development: Neurobiological Aspects of Extraordinary Brains*. Cambridge, MA: Harvard University Press.
Geschwind N (1979) Specializations of the human brain. *Scientific American* **241**(3): 180–199.
Grodzinsky Y, Shapiro LP and Swinney DA (eds) (2000) *Brain and Language: Representation and Processing*. San Diego, CA: Academic Press.
Jarrold C, Boucher J and Russell J (1997) Language profiles in children with autism: theoretical and methodological implications. *Autism* **1**: 57–76.
Marantz A, Miyashita Y and O'Neil W (eds) (2000) *Image, Language, Brain*. Cambridge, MA: MIT Press.
Rice ML (1999) Specific grammatical limitations in children with specific language impairment. In: Tager-Flusberg H (ed.) *Neurodevelopmental Disorders*, pp. 331–359. Cambridge, MA: MIT Press.
Shankweiler D, Crain S, Katz L *et al*. (1995) Cognitive profiles of reading-disabled children: comparison of language skills in phonology, morphology, and syntax. *Psychological Science* **6**: 149–156.
Thompson CK, Shapiro LP, Ballard KJ *et al*. (1997) Training and generalized production of wh- and NP-movement structures in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research* **40**: 228–244.
Uriagereka J (2000) *Rhyme and Reason*. Cambridge, MA: MIT Press.
Zurif EB, Swinney D, Prather P, Solomon J and Bushell C (1993) An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language* **45**: 448–464.

Language Production, Incremental

Advanced article

Linda R Wheeldon, University of Birmingham, Birmingham, UK
Antje S Meyer, University of Birmingham, Birmingham, UK
Mark Smith, University of Birmingham, Birmingham, UK

CONTENTS

Incrementality concerns the processing stages of linguistic output: conceptual structure, lexical selection, and generation of grammatical and phonological structure, and suggests that these processes can to an extent run in parallel to aid the speed and fluency of speech production.

INTRODUCTION

In order to express ourselves in speech we must translate some nonlinguistic idea into a fluently articulated utterance. All models of speech production postulate that this translation process occurs in a number of successive steps. The starting point is a conceptual structure for an utterance that details the information or message to be conveyed. Based on this information, the appropriate lexical items are selected and a grammatical structure is generated that places these lexical items in the correct relation to each other. Finally, the phonological structure of the utterance must be generated before articulation can commence.

The question to be addressed in this article concerns the way in which processing at each level proceeds over time and, in particular, how processes at different levels are coordinated with each other. The speed and fluency with which we can articulate complex sentences makes it unlikely that processing at each level is completed before processing at the next begins. Following Kempen and Hoenkamp (1987), Levelt (1989, 1992) proposed that processing at all levels occurs in an incremental fashion with a processor being triggered by any piece of characteristic input from the processors that feed into it. This means that some processing must have occurred at a particular level before processing at the next level can begin, but that processing at all levels can run in parallel on different pieces of the utterance to be produced.

An incremental production system is attractive for a number of reasons. It can explain the fluency of speech because it allows the fast release of formulated chunks of the utterance for phonological processing and articulation. This in turn removes the need for temporary storage of completed chunks of the utterance, thereby reducing processing costs. Finally, incremental processing spreads the costs of production evenly across the period of time in which a sentence is articulated, thus reducing the likelihood of dysfluencies that can result from excessive processing loads. In order to work efficiently, an incremental system requires that processing can occur from left to right in an utterance with minimal look ahead. Thus problems for an incremental system would be caused by evidence of large scope dependencies where processing of a particular fragment of an utterance is dependent on information available in much later fragments. A detailed incremental model of language production must clearly specify for each component process the minimal chunk of information it processes and delivers as output. The sections below discuss the relevant data for the processes involved in lexical access and the generation of grammatical and phonological structure.

ACCESSING LEXICAL INFORMATION

The mental lexicon is a store of words that can be retrieved independently of each other and in any order. The minimal input unit for lexical access is a lexical concept (or, in some theories, a set of semantic features) that maps onto one lexical entry. Such one-to-one mapping occurs, for instance, when a speaker names a single object in one word.

All models of lexical access assume that lexical access is essentially the same for words spoken in isolation and in context. When speakers produce

sentences, they simply retrieve many words. Theories widely agree that lexical selection is a sequential process. Thus, while several lemmas or several word forms may simultaneously receive some activation, their selection as part of the grammatical or phonological representation is a sequential process (e.g. Levelt *et al*., 1999).

In a recent series of experiments testing the sequentiality assumption, the speakers' eye movements were registered while they named picture pairs in noun phrases such as 'the dog and the bike' (see Levelt and Meyer, 2000, for a review). On most trials, the speakers fixated upon each of the objects in the order of mention. Importantly, the viewing time for an object depended on the time required for lexical access to its name. For instance, objects with frequent or short names were looked at for shorter periods than objects with less frequent or longer names. In addition, phonological priming reduced the viewing times. In one study speakers referred to objects in two- or four-word phrases (e.g. 'the dog' versus 'the little brown dog'). Measurements of the coordination of speech and eye movements showed that for both types of phrases the shift of gaze from the first to the second object usually occurred just before the onset of the last word of the phrase referring to the first object ('dog', in the example).

These results support one aspect of the incrementality assumption, namely that speech planning and speaking co-occur in time. However, they do not support another aspect of the assumption, which is that speakers plan different fragments simultaneously. Instead the speakers first focused on one object until they had planned the corresponding utterance fragment down to the level of phonological form, and then turned to the next object.

One may argue that the two-objects naming paradigm is particularly conducive to the use of a sequential planning strategy, and that speakers may often be forced to plan several utterance fragments in parallel. There are, for instance, idioms, such as 'to fall into disuse' versus 'to sink into oblivion', in which the choice of verb and noun seem to be interdependent. However, idioms must be stored in the mental lexicon. Sprenger *et al*. (1999) have proposed to represent them as complex structures including a super-lemma representing the idiom as a whole plus subordinate lemmas representing the individual words. A speaker first selects a super-lemma corresponding to the idiom and then sequentially selects the subordinate lemmas. Thus, idioms do not require simultaneous selection of several lemmas.

Obviously, speakers must often engage in considerable advance planning to decide what they wish to say. For example, when they describe an event, they must first seek to understand who does what to whom. However, these processes are conceptual planning processes. As soon as a coherent message has been generated, it can be used to select lexical items in a strict sequential fashion.

GENERATING GRAMMATICAL STRUCTURE

In an influential theory of speech production, Garrett (1982) suggested that grammatical processes such as lemma access and grammatical role assignment are not incremental but employ a more holistic, clausal scope. Recent empirical work, however, suggests otherwise. In a study by Schriefers *et al*. (1998) participants were presented with incomplete German sentences and required to complete them. As participants completed the sentences they heard distractor words which were either semantically related to the words they were producing, or unrelated. When the distractor word was similar semantically to the first word of the clause produced by the participant, speech onset was delayed relative to the unrelated word, indicating that the first word of the clause had been accessed prior to speech onset. In contrast, when the distractor word was related to the final word of the clause, no delay was found, indicating that the final word had not been accessed prior to speech onset. Similarly, Smith and Wheeldon (1999) found that speech onset latencies were longer to sentences featuring an initial coordinate noun phrase (as in 1) than to sentences featuring a final coordinate noun phrase (as in 2):

The dog and the boot move above the car. (1)

The car moves above the dog and the boot. (2)

Clearly, if participants had planned the entire sentence prior to speech onset we should expect latencies to (1) and (2) to be similar since they feature equivalent numbers of words. In contrast, we should expect longer latencies to (1) than (2) if only the lemmas within the first phrase of the sentence had been accessed prior to speech onset, since the first phrase of (1) contains more lemmas than that of (2).

Crucially, such evidence is supported by observations of naturally occurring language. Using a corpus of English utterances, Clark and Wasow (1998) investigated the frequency of word repetitions such as:

I uh I wouldn't be surprised at that. (3)

Such repetitions were more frequent for the leftmost phrases in a clause than for the rightmost phrases and more frequent for words at the left edge of a phrase than at the right edge. On the assumption that dysfluency frequency is an index of planning load, Clark and Wasow concluded that speakers employ both a clausal and a phrasal planning scope during speech. The phrasal scope fits with experimental evidence of incremental grammatical encoding (Schriefers *et al*., 1998; Smith and Wheeldon, 1999), whilst the clausal scope fits with experimental evidence of holistic conceptual planning (Ford and Holmes, 1978; Smith and Wheeldon, 1999). Despite such evidence of incremental grammatical encoding, there remains debate over whether incremental processing is parallel, with grammatical and phonological planning stages simultaneously processing distinct chunks (De Smedt, 1996), or whether it is serial so that only a single chunk is processed at a time (Ward, 1992).

Finally, there is the question of the minimum size of the chunks employed in grammatical encoding (Schriefers *et al*., 1999). There must be such a lower limit since, in many languages, there is grammatical agreement between elements. Moreover, this lower limit will differ depending on the scope of the dependencies in a given language, for example a noun and a determiner (German) or a verb and a direct object (Swahili). Models of grammatical encoding often assume that the controlling lemma (e.g. the noun in German determiner–noun phrases) transmits features to the controlled lemma, thereby specifying the morphological and phonological form of the controlled element. Since the controlling element can follow the controlled one in the surface structure of the utterance, the controlled element cannot be phonologically encoded or articulated before the controlling lemma has been accessed.

GENERATING PHONOLOGICAL STRUCTURE

Producing the sound form of the selected words in a sentence is a generative process. We do not simply retrieve and concatenate the stored phonological representation for each word. This is because the exact sound structure of a word is dependent on the context in which it is produced.

Levelt (1989, 1992) argued that the unit of phonological encoding is the phonological word – a prosodic unit comprising minimally a stressed foot and maximally a single lexical word combined with any associated unstressed function words (see Wheeldon, 2000, for a review). Levelt claims that the phonological representations of the selected words are combined to form phonological word frames. The phonological segments for each word are made available separately and then associated to the newly constructed phonological word frames in a left-to-right manner. For example, in the utterance, 'I gave it to him', the five lexical items resyllabify to form one phonological word $[aɪ\text{-}geɪ\text{-}vɪ\text{-}tɪm]_ω$ with one main stress. As the segments for each syllable are associated to their prosodic frame they are used to retrieve stored, syllable-sized, articulatory routines (Levelt *et al*., 1999; Levelt and Wheeldon, 1994). When the articulatory routines for the entire phonological word have been retrieved, the phonetic plan can be articulated. Thus, during the production of connected speech, a whole phonological word is constructed before articulation commences.

A number of experimental results provide support for this hypothesis. Using a picture–word interference task, Meyer and Schriefers (1991) found significant priming effects from spoken distractors that overlapped with either the first or the second syllable of the target picture name (e.g. picture: ha-mer, 'hammer'; distractors: *ha-vik*, 'hawk' or *zo-mer*, 'summer'). Moreover, latency to produce a spoken word has been shown to increase with its number of syllables (see Wheeldon, 2000, for a review). These findings suggest that all the syllables of a word are encoded prior to the onset of articulation.

However, these experiments do not distinguish between phonological and lexical words as the minimal unit of phonological encoding. Wheeldon and Lahiri (1997) tested the production of phonological words comprising a lexical word plus unstressed function words. They demonstrated that sentence production latencies are a function of the size of the initial phonological word in the utterance. Latencies for the sentence $[\text{Ik zoek het}]_ω$ $[\text{water}]_ω$ 'I seek the water' were longer than latencies for sentences like $[\text{Ik zoek}]_ω$ $[\text{vers}]_ω$ $[\text{water}]_ω$ 'I seek fresh water'.

However, the syllable latency effect has proved difficult to replicate. Moreover, Schriefers and Teruel (1999) have demonstrated that priming may be limited to the initial syllable of disyllabic words in some two-word utterances. It is possible therefore that while the phonological word may be the preferred unit of phonological encoding it may not constitute the *minimal* unit of phonological

encoding. Indeed it is possible that the unit chosen may differ for different utterances (e.g. single word or sentences) and in different speaking contexts. Finally, there remains the issue of cross-linguistic differences in the scope across which dependencies may operate during the generation of prosodic structure. Lahiri (2000) discusses several tonal and intonation processes in languages that require a processing scope greater than a phonological word.

CONCLUSION

All current models of speech production incorporate some form of incrementality of processing in order to explain both the speed and fluency of speech production. Nevertheless, it is clear that the minimal size of the units that can be processed at each level is constrained by the scope of the linguistic dependencies operating in a given language – and perhaps by differences in the demand for speed in different speaking situations.

References

Clark HH and Wasow T (1998) Repeating words in spontaneous speech. *Cognitive Psychology* **37**: 201–242.

De Smedt K (1996) Computational models of incremental grammatical encoding. In: Dijkstra T and de Smedt K (eds) *Computational Psycholinguistics*, pp. 279–307. London, UK: Taylor & Francis.

Ford M and Holmes VM (1978) Planning units in sentence production. *Cognition* **6**: 35–53.

Garrett MF (1982) Production of speech: observations from normal and pathological language use. In: Ellis AW (ed.) *Normality and Pathology in Cognitive Functions*, pp. 19–76. London, UK: Academic Press.

Kempen G and Hoenkamp E (1987) An incremental procedural grammar for sentence formulation. *Cognitive Science* **11**: 201–258.

Lahiri A (2000) Phonology: structure representation and process. In: Wheeldon LR (ed.) *Aspects of Sentence Production*, pp. 165–226. Hove, UK: Psychology Press.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt WJM (1992) Accessing words in speech production: stages, processes and representations. *Cognition* **42**: 1–22.

Levelt WJM and Meyer AS (2000) Word for word: multiple lexical access in speech production. *The European Journal of Cognitive Psychology* **12**: 433–452.

Levelt WJM, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 59–60.

Levelt WJM and Wheeldon LR (1994) Do speakers have access to a mental syllabary? *Cognition* **50**: 239–269.

Meyer AS and Schriefers H (1991) Phonological facilitation in picture–word interference experiments: effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition* **17**: 1146–1160.

Schriefers H, de Ruiter JP and Steigerwald M (1999) Parallelism in the production of noun phrases: experiments and reaction time models. *Journal of Experimental Psychology: Learning, Memory and Cognition* **25**: 702–720.

Schriefers H and Teruel E (1999) Phonological facilitation in the production of two-word utterances. *European Journal of Cognitive Psychology* **11**: 17–50.

Schriefers H, Teruel E and Meinshausen RM (1998) Producing simple sentences: results from picture–word interference experiments. *Journal of Memory and Language* **39**: 609–632.

Smith MC and Wheeldon LR (1999) High level processing scope in spoken sentence production. *Cognition* **73**: 205–246.

Sprenger S, Levelt WJM and Kempen G (1999) Producing idiomatic expressions: idiom representation and access. Poster AMLaP-99, Edinburgh.

Ward N (1992) A parallel approach to syntax for generation. *Artificial Intelligence* **57**: 183–225.

Wheeldon LR (2000) Generating prosodic structure. In: Wheeldon LR (ed.) *Aspects of Sentence Production*, pp. 249–274. Hove, UK: Psychology Press.

Wheeldon LR and Lahiri A (1997) Prosodic units in speech production. *Journal of Memory and Language* **37**: 356–381.

Further Reading

Bock JK, Loebell H and Morey R (1992) From conceptual roles to structural relations: bridging the syntactic cleft. *Psychological Review* **99**: 150–171.

Clark HH (1996) *Using Language*. Cambridge, UK: Cambridge University Press.

Ferreira F (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* **30**: 210–233.

Ferreira F and Swets B (2002) How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language* **46**: 57–84.

Kello CT, Plaut DC and MacWhinney B (2000) The task dependence of staged versus cascade processing: an empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology, General* **129**: 340–360.

Kempen G and Hoenkamp E (1987) An incremental procedural grammar for sentence formulation. *Cognitive Science* **11**: 201–258.

Kuiper K (1996) *Smooth Talkers*. Mahwah, NJ: LEA.

Wheeldon LR (2000) *Aspects of Language Production*. Hove, UK: Psychology Press.

Language, Connectionist and Symbolic Representations of

Advanced article

Mark Steedman, University of Edinburgh, Edinburgh, UK

CONTENTS

Connectionist or neurocomputational representations are based on sets of similar, multiple connected neuron-like parallel processing units, the connections bearing modifiable weights adjusted on the basis of locally available information using a learning algorithm. Symbolic representations, by contrast, are defined in terms of rules relating expressions in a formal language. Among the claims that have been made for connectionist models is the 'emergence' of generalizations that had been thought to require the mediation of rule-based grammars and modular symbolic processing architectures. A more convincing linguistic role for such networks lies in their potential for inducing grounded conceptual structure and statistical models as infrastructure for acquisition and processing of standard symbolist representations of lexicalized syntax and semantics.

INTRODUCTION

Connectionist theories of the parallel distributed processing (PDP) variety (Rumelhart *et al*., 1986) begin from the reasonable belief that phenomena of mind are the result of computation in richly interconnected networks of neurons or neuron-like units. They embody the hypothesis that the nature of this computational device critically determines the nature of the computation itself. Symbolic theories begin from the equally reasonable belief that such phenomena of mind as language use, understanding, and reasoning are symbolic in nature, in much the same sense as mathematical and logical inference are, and that the computation involved can therefore be characterized independently of the specific device, whether parallel-distributed or not, that implements it. The connectionist approach is by nature reductionist (in the best sense of a much-abused term): it attempts to generalize from low-level mechanisms to higher-order phenomena. The symbolic approach is phenomenological (again in a positive sense): it tries to work in the opposite direction, from a high-level description to the implicit underlying mechanism. These approaches are clearly compatible, and can coexist. Both are necessary: the interesting question to ask is to which particular problems each is best suited, and under what circumstances they can be unified.

In our present state of knowledge in this rapidly evolving field, the answers to these questions are far from clear. Neurocomputational mechanisms have proved their worth in the field of pattern recognition and classification, where they can extract structure latent in inputs such as images of faces, handwritten letters, and speech, and embody that structure in recognizers that would be impossible to specify by hand, or that are orders of magnitude more efficient than rule-based mechanisms, even when these are statistically optimized. On the other hand, except where they have been used to explicitly simulate the structure of a symbolic parser and associated devices such as the push-down automaton, these mechanisms have been much less successful in demonstrating the kind of recursive productivity that rule-based symbolic systems are good at, or in supporting semantic interpretation and inference.

In contrast, rule-based discrete symbolic systems express productivity or systematicity, semantic interpretation, and processes of inference immediately. However, it has proven very difficult to build rule-based linguistic or computational-linguistic systems with coverage on the scale characteristic of human linguistic and reasoning abilities, and in recent years such systems have increasingly relied upon machine learning and statistical optimization techniques, of a kind closely related in mathematical terms to neurocomputational techniques. It seems likely that there is much to be gained from combining these approaches.

In order to compare the approaches, it is helpful to recall that symbolist theories distinguish a number of distinct components of language processors. One fairly generally applicable architecture distinguishes: (1) a *grammar*, characterized by syntactic and semantic rules of certain classes and a related characteristic automaton; (2) a nondeterministic *algorithm*, characterized by properties such as the order in which rules are applied to strings, whether bottom-up or top-down, the order in which the words of a string are examined, whether first-to-last or otherwise, and certain memory resources, such as those used in building structure and the charts or tables used in parsers based on dynamic programming; and (3) an oracle, or decision criterion for rendering the algorithm deterministic and deciding which rule to apply in cases where there is more than one possibility.

In any given theoretical presentation or implementation, these modules may be combined, but in rule-based theories they can usually be distinguished in functional terms. The fact that they are in that sense distinct modules does not of course imply that the corresponding computations must be carried out in series, in chronologically separate phases: for example, it is quite possible to construct systems in which the oracle can call on the results of semantic interpretation and contextual reference in mid-sentence, while syntactic analysis is still under way.

It is important to be clear on this last point, because the fact that phenomena like 'garden path' effects in parsing can be affected by semantics and even by extrasentential context is often adduced as evidence in specific support of PDP models and against symbolic ones. However, parallel constraint satisfaction drawing on multiple knowledge sources is commonplace among rule-based models of sentence processing, and is acknowledged by Fodor (1983, p. 78, pp 134–135) to be entirely modular.

It follows by the same token that connectionism is not intrinsically less modular than any other approach. Nevertheless, there has been a considerable emphasis in the connectionist literature on the idea that the appearance of rule-like behavior is 'emergent' in such systems, and that language processing and language acquisition can be modeled in monolithic nonmodular PDP machines or algorithms without the explicit involvement of grammars.

In assessing the connectionist claims it is important to ask what it is that neurocomputational machinery actually learns, and whether it can in principle do the jobs that language does for us. In particular, we must ask whether it will deliver meaning in a form that will in turn support logical inference. In practice, inference systems of any generality have generally depended on explicit representations of structure of some kind.

In investigating these questions, two claims are of particular interest. The first is the claim that grammars, in the sense used by symbolists, are an emergent phenomenon of the learning of sequential dependencies by recurrent neural networks with 'contextual' units. The second is the claim, and originating with Niklasson and van Gelder (1994), that structure can be represented in distributed memory and manipulated systematically (in the sense of that term that Fodor and Pylyshyn (1988) claimed to intrinsically require symbolic representation), without explicitly representing the pointer structure of the symbolic representation. Clearly if both of these claims are correct then the neurocomputational account has gone a long way towards delivering a distinctively nonsymbolic account of language.

NETWORKS AND GRAMMARS

Simple Recurrent Networks

The 'simple recurrent network' (SRN) (Elman, 1990) approximates the more costly but exact 'back propagation through time' algorithm of Rumelhart *et al*. (1986) for learning sequential dependencies. It does so by using a single set of context units which store the activations of the hidden units at time $t-1$ as an input to the hidden units at time t, along with the activations of the normal input units corresponding to the current item, item_t, as in Figure 1. Since the activations of the hidden units at time $t-1$ were themselves partly determined by the activations on the hidden units at time $t-2$, which were in turn determined by those at time $t-3$, and so on, the context units

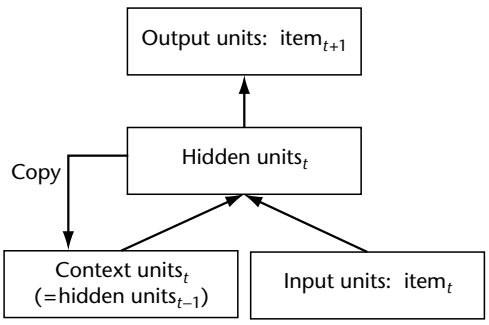


Figure 1. The simple recurrent network.

carry ever-diminishing echoes of ever more distant items in the preceding sequence.

Because there is no clear bound to the extent of the preceding sequence about which information can be captured in the context units, it is not clear precisely what is the automata-theoretic power of such 'graded-state machines' (Cleeremans *et al*., 1995). However, the signal-to-noise ratio for information concerning distant items falls off very rapidly with this mechanism, and it is fairly clear that in practice SRNs of the kind that can actually be built and trained approximate the class of finite-state Markov machines that can be learned using the exact technique to a degree of accuracy that depends on the maximum number of timesteps required.

Finite-state machines are interesting devices, and it is often surprising to see the extent to which they can approximate the output of devices that are intrinsically higher in the automata-theoretic hierarchy. It is interesting to ask whether similar mechanisms can play any part in natural-language processing.

This is possible, because some neural-network algorithms are capable of inducing extremely efficient – and correspondingly opaque – representations, when compared with standard hidden Markov models. However, as SRNs are actually used by psychologists and linguists they appear to approximate something much closer to a familiar standard symbolist finite-state device, namely the *n*-gram part-of-speech (POS) tagger. (This also seems to be their role in 'hybrid architecture' connectionist parsers, which use networks to implement a push-down stack, and structure-building modules in a more standard parser architecture.)

Finite-state Part-of-speech Tagging

n-gram POS tagging – that is, the determination of the form-class of ambiguous lexical items like *bear* on the basis of sequential probabilities at the word level – can be remarkably successful in reducing the degree of nondeterminism that practical parsing algorithms must cope with (accuracies over 97% are standard). Moreover, there is growing evidence that if the standard Brown-corpus POS categories like VB are replaced with the more informative lexical categories that are used in lexicalized grammars – such as lexical functional grammar, tree adjoining grammar, head-driven phrase structure grammar, and the various forms of categorial grammar – and if different senses of the same syntactic type are also distinguished as different lexical items, lexical disambiguation may do a great deal of the work of parsing itself, leaving only structural or 'attachment' ambiguities to be resolved by parsing proper.

Why do SRNs and Part-of-speech Taggers Work?

Finite-state POS taggers, and by assumption SRNs, work reasonably well on tasks like category and sense disambiguation and prediction of succeeding category, because the implicit Markov processes encode a lot of the redundancy (in the information-theoretic sense of the term) that is implicit in grammar, interpretation, and world knowledge. This means that, like standard Markov processes, SRNs can be made the basis of good predictors of processing difficulty. For example, the SVO word order of English and our knowledge of the world between them determine the fact that the transitive category for the word *arrested* is more likely to follow the noun *cop* than the past participial category, while these preferences are reversed following the word *crook*:

The cop arrested by the detective was left. (1)

The crook arrested by the detective was left. (2)

These likelihoods are reflected in increased reading times for human subjects at the word *by* in sentence 1 as compared with sentence 2, for example.

Are Grammars Emergent from SRNs?

While claims exist in the literature to the effect that recognizers for string sets of kinds that in general require grammars of higher than finite-state power have been acquired by SRNs (Christiansen and Chater, 1994), none of these results suggests that the grammars in question are therefore 'emergent' properties of mechanisms like SRN, any more than they are of *n*-gram POS taggers. Although the context defined by the context units is in a limited sense unbounded, and SRNs can in theory be used to model long-distance agreement dependencies, because of already-noted properties of the context unit representation, reliability falls off with distance, and these dependencies cannot be regarded as unbounded in the technical grammatical sense. They should instead be regarded as a finite-state cover of the higher-power grammar to some limited maximum string length. This means that claims to model human language acquisition using SRNs must be treated with some caution, although such models raise developmentally

interesting questions. (Interestingly, there are conflicting claims by Elman *et al*. (1996) and by Rohde and Plaut (1999) as to whether SRN learning depends on 'starting small', ordering presentation of simple examples before complex ones, or is rather inherently biased towards acquisition of local dependencies before long-range dependencies, and therefore independent of the order in which training examples are presented.)

Even within these limits, error-free sequences of grammatical categories fall short of semantic interpretability, as can be seen from the fact that the following word sequence has two interpretations based on identical Brown-corpus categories:

Put the block in the box on the table. (3)

Although SRNs can be regarded as disambiguating lexical items, this other kind of ambiguity – structural or attachment ambiguity – remains, as in the case of POS taggers.

For the same reason, it does not seem legitimate to regard 'trajectories' through the high-dimensional space defined by the hidden units as the equivalent of parses or interpretation (Tabor *et al*., 1997). Many other defining properties of interpretations – such as the ability to support the kind of structure-dependent transformations characteristic of inference – seem to be lacking in trajectories or category sequences of this kind. To find such properties in neurocomputational representations, we need to look to devices other than SRN.

INTERPRETABLE STRUCTURE AND ASSOCIATIVE MEMORY

Much neurocomputational work has explicitly or implicitly taken on board the need for the equivalent of trees or pointer structures to represent syntactic or semantic analyses (Hinton, 1990a) using associative memories of various kinds.

Such devices are of interest for (at least) two reasons. Firstly, they inherit some psychologically desirable properties of distributed representations, such as content-addressability and smooth degradation under noise and damage. Secondly, they offer a way to think about the interface between neurally embedded map-like sensorimotor inputs and outputs, and symbolic knowledge representation.

Recursive Auto-associative Memory

Recursive auto-associative memory (RAAM) (Pollack, 1990) is a device that uses hidden unit activation patterns to store associations between input and output patterns. It is called 'auto-associative' because it uses the same patterns as input and output.

An *n*-ary recursive structure can be stored bottom-up in the RAAM starting with the leaf elements by recursively auto-associating vectors comprising up to *n* hidden-unit activation patterns corresponding to either leaves or previously encoded structures. The activation pattern that results from each auto-association of the children can then be treated as the address of the parent (see Figure 2).

Since by including finitely many further units on the input and output layers we can associate node-label or content information with the nodes, a modification sometimes referred to as 'labeled RAAM', this device can store recursive parse structures, thematic representations, or other varieties of logical form of sentences.

The device should not be confused with a parser: it is trained with fully articulated structures which it merely stores efficiently. The hidden units can be regarded as encoding to some approximation the context-free productions that defined those structures, in a fashion similar to the way Hinton (1990b) encoded part–whole relations, enabling recognition of pattern instances that have not been encountered before. In that sense the device has been claimed to be capable of inducing the corresponding grammar from the trees (Pollack, 1990), contrary to the claims of Fodor and Pylyshyn (1988) concerning the systematicity or generalizing capacities of neural networks.

This claim was challenged by Hadley (1994a), who extended Fodor and Pylyshyn's critique to distinguish a number of levels of systematicity in the induced classifier, contingent on the relation of the test examples to the training set. The system of levels of systematicity was further refined by Niklasson and van Gelder (1994) who among other levels distinguished: level 3, generalization to all and only legal seen structures with novel constant–position pairings; level 4, generalization

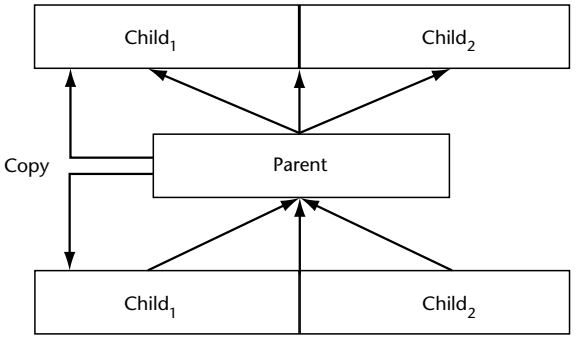


Figure 2. Recursive auto-associative memory.

to all and only legal novel embeddings of seen structures with seen constant–position pairings; and level 5, generalization to all and only legal novel embeddings of seen structures with novel constant–position pairings – where 'legal' means permitted by the corresponding rule-based system.

Scope and Limits of Systematicity

Niklasson and van Gelder's (1994) experiment had the novel feature of representing both formula-like structures and (via a separate network) operations over those representations resembling logical equivalence transformations, such as the rule that replaces formulae of the form $P \rightarrow Q$ with corresponding formulae $\neg P \vee Q$, or syntactic transformations, such as the passive rule that relates psuedo-English sentences of the form *cat chase dog* to those of the form *dog chased by cat*. Niklasson and van Gelder were able to demonstrate level-3 systematicity on this task (although Hadley (1994b) expresses some reservations about statistical reliability). Remarkably, these operations worked without explicitly unpicking the representation, via pointer dereferencing of the kind standard in symbolic representations of logical formulae and their transformation using list-processing computer programming languages like LISP and PROLOG.

This is probably a more appropriate use for RAAM than building parse trees, since RAAM is slow to train, and inherits poor scaling properties from its use of back propagation. Devices such as the 'holographic reduced representations' (HRR) proposed by Plate (1994) are an interesting alternative. Their properties for the storage and holistic transformation of such structures have been investigated by Neumann (2001), who reports replication of Niklasson and van Gelder (1994) using RAAM and an extension to their level-5 systematicity (generalization to novel embeddings of seen structures with novel constant–position pairings) for a related system using HRR.

Neumann shows that these networks can learn superimposed collections of linear-transformational rules that depend on the relation of isotopy which holds between the distribution of their inputs and that of their outputs in the hyperspace defined by the weights. (This space can be examined using principal components analysis and other clustering techniques. This representation of whole disambiguated structures as points in a high-dimensional space is unrelated to the SRN representation of sequences as trajectories in such a space.) Among the rules that Neumann shows can be learned in this way is one mapping $P \wedge (P \rightarrow Q)$ to Q. This rule is related to the rule of modus ponens, a fundamental rule of inference. It therefore seems possible in principle that structures represented in this way could support inference.

However, some caution is needed in interpreting these results, as Neumann points out. To build a practical inference engine to exploit this property requires a number of further steps, including the identification of suitable pairs of formulae of the form P and $P \rightarrow Q$ from a larger set to form the input to the rule, and a process of search for proofs through a potentially exponentially growing space of sequences of inference steps. It is unclear whether such processes can be helpfully thought of in distinctly neurocomputational terms, or whether this is the level at which at least part of cognition becomes distinctively symbolic.

USING CLASSIFIERS TO LEARN GROUNDED CONCEPTUAL CATEGORIES

One promising use of associative memory models like RAAM and HRR might be to learn the bounded syntactic–semantic structures that are associated with lexical items, particularly verbs, which provide the input to lexicalized grammars and parsers of the kinds referred to earlier.

We might assume that a subset of such structures is available prelinguistically, resulting relatively directly from the evolved or learned structure of connections to the sensorium, short-term memory, and the like. At higher levels, such structures may arise from nonlinguistic network concept learning along lines set out by Hinton (1990b), without mediating symbolic forms.

Part of the interest of this proposal lies in the possibility that the interaction of such structured sensorimotor manifolds and this novel form of concept learning might give rise to 'grounded' conceptual categories within a standard symbolist approach. Grammar acquisition would then mainly reduce to the association of lexical items to concepts, and decisions such as whether the syntactic type corresponding to the concept of walking looks forwards, backwards, or either way, for its subject in the particular language that the child is faced with, and how the multiple arguments of transitives, ditransitives, and the like map to the underlying universal logical form, as reflected for example in the possibilities for reflexivization. Since directionality can be represented as a value on an input unit, and since an individual category can be defined as a finite-state machine, and its

mapping to universal logical form can be captured as a finite-state transduction, such categories are good candidates for learning with neurocomputational devices.

A similar tendency towards lexical involvement is evident in current statistical computational linguistic research. Much work in probabilistic parsing moves away from autonomous Markovian POS tagging and prefiltering, and towards a greater integration of probabilities with grammar – see Manning and Schütze (1999) for a review.

Part of the interest of this proposal lies in the possibility that such learning might capture word-order generalizations over the lexical categories. Certain constraints that have been discussed within optimality theory (Prince and Smolensky, 1997), such as the tendency for semantically related categories (such as tensed transitive verbs) to have the same directionality (such as SVO order) are 'soft,' in the sense that they can have exceptions (such as English auxiliary verbs). It seems likely that the lexical acquisition device based on associative memory, sketched above, might also be suited to acquiring such lexicons based on soft constraints. If so, then the claim that the form of possible human lexicons is 'emergent' from the neural mechanism would have real force.

In this connection it is interesting to note that monolithic PDP-based connectionist models have been successful in modeling the acquisition and processing of systems of morphologically inflected lexical items that mix regular and irregular forms, such as the English past-tense system (first approached by Rumelhart and McClelland (1986)), competing successfully if not conclusively with models based on rules and exceptions (Pinker, 1999) in accounting for the course of acquisition in children, including the 'U-shaped curve' in frequency of correct versus incorrect use of irregulars, whereby children initially use forms like 'ran', then drop them in favor of over-regulations like 'runned', before returning to using 'ran'.

IMPLICATIONS FOR NATIVIST THEORIES OF THE LANGUAGE FACULTY

Given the origins of the term 'connectionism' in the behaviorist theories of Thorndike and others as modified by Hebb, it is perhaps not surprising that its advocates see themselves as in conflict with the nativist position associated with Noam Chomsky, the opponent of behaviorist theories of language and founder of modern symbolic approaches to language.

The conflict, which has been most eloquently pursued by Elman *et al*. (1996) and in the response by Marcus (2001), is more apparent than real. Chomsky's point has always been that attempts to explain the universal form of language and the course of language acquisition in terms of more general-purpose cognitive faculties or psychological laws have in practice not been notably effective. While constantly deriding the inability of currently available theories of learning, cognitive development, or semantics to make any significant contribution to linguistic explanation, he has consistently advocated the study of innate universal principles of language in isolation from other aspects of cognition as a matter of methodological expediency. The fact remains that the only plausible source for the innate component lies in the conceptual structure with which the child comes to language learning, and which either evolved or was learned for more general cognitive purposes. This observation is implicit at least as early as Chomsky (1965, section 1.5) and is explicit in Pinker's early work on modeling acquisition.

Elman *et al*. (1996) do a good job of demolishing certain dubious claims for evidence of specifically linguistic genetic components. (In particular, their review of the widespread exaggeration of the significance of the heritable disorder of the KE family is telling.) However, the contribution of neurocomputational theories to our understanding of language seems unlikely to be to demonstrate the emergence of grammar from monolithic neural computation. Machines that are emergent in this sense are intrinsically implausible as psychological models, because they offer the same kind of obstacles to further evolutionary development that unstructured programs offer to software developers, as Holland (1998) points out in different terms. The distinctive contribution of neurocomputational models is more likely to lie in explaining how the structure implicit in the sensory manifold and our interactions with the world can be extracted by classifiers based on algorithms like back propagation and the restricted Boltzmann machine learning algorithm, to provide the grounded conceptual structure upon which both reasoning about the world and the development of language depend. The question of how much of this conceptual structure is actively learned by the individual prelinguistic child, and how much of it has been compiled into heritable 'hard-wired' components during the process of evolution of humans and their animal ancestors, and the question of what further apparatus is needed for the development of language and whether its origins can also be

traced to more generally useful cognitive abilities, remain open. (One promising but in formal terms under-investigated source for the latter apparatus that is suggested by both developmental and neuroanatomical evidence lies in the system for planning action in the world.) These will be questions of considerable empirical interest to both symbolic and neurocomputational theorists.

References

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Christiansen M and Chater N (1994) Generalization and connectionist language learning. *Mind and Language* **9**: 273–287.

Cleeremans A, Servan-Schreiber D and McClelland J (1995) Graded state machines: the representation of temporal contingencies in feedback. In: Chauvin Y and Rumelhart DE (eds) *Backpropagation: Theory, Architectures, and Applications*, pp. 274–312. Hillsdale, NJ: Lawrence Erlbaum.

Elman J (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Elman J, Bates E, Johnson MH *et al*. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **35**: 183–204.

Hadley R (1994a) Systematicity in connectionist language learning. *Mind and Language* **9**: 247–272.

Hadley R (1994b) Systematicity revised: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind and Language* **9**: 431–444.

Hinton G (ed.) (1990a) *Connectionist Symbol Processing*. Cambridge, MA: MIT Press/Elsevier. [Reprint of *Artificial Intelligence* **46**(1–2).]

Hinton G (1990b) Mapping part–whole hierarchies into connectionist networks. *Artificial Intelligence* **46**: 47–75. [Reprinted in Hinton (1990a).]

Holland J (1998) *Emergence: From Chaos to Order*. Reading, MA: Perseus.

Manning C and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus G (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science (Learning, Development, and Conceptual Change)*. Cambridge, MA: MIT Press.

Neumann J (2001) Holistic processing of hierarchical structures in connectionist networks. PhD thesis, University of Edinburgh.

Niklasson L and van Gelder T (1994) On being systematically connectionist. *Mind and Language* **9**: 288–302.

Pinker S (1999) *Words and Rules: The Ingredients of Language*. New York, NY: Basic Books.

Plate T (1994) Distributed representations and nested compositional structure. PhD thesis, University of Toronto.

Pollack J (1990) Recursive distributed representations. *Artificial Intelligence* **46**: 77–105. [Reprinted in Hinton (1990a).]

Prince A and Smolensky P (1997) Optimality: from neural networks to universal grammar. *Science* **275**: 1604–1610.

Rohde D and Plaut D (1999) Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition* **72**: 67–109.

Rumelhart D and McClelland J (1986) On learning the past tenses of English verbs. In: McClelland J, Rumelhart D and the PDP Research Group (eds) *Parallel Distributed Processing*, vol. II: *Psychological and Biological Models*, pp. 216–271. Cambridge, MA: MIT Press.

Rumelhart D, McClelland J and the PDP Research Group (eds) (1986) *Parallel Distributed Processing*, vol. I. Cambridge, MA: MIT Press.

Tabor W, Cornell J and Tanenhaus M (1997) Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* **12**: 211–271.

Further Reading

Cleeremans A (1993) *Mechanisms of Implicit Learning*. Cambridge, MA: MIT Press.

Gluck M and Myers C (2000) *Gateway to Memory*. Cambridge, MA: MIT Press.

Hebb D (1949) *The Organization of Behavior*. New York: Wiley.

Mikkulainen R (1993) *Subsymbolic Natural Language Processing*. Cambridge, MA: MIT Press.

Piattelli-Palmarini M (ed.) (1980) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Pinker S and Mehler J (eds) (1988) *Connections and Symbols*. Cambridge, MA: MIT Press. Reprint from *Cognition* **28**.

Plaut D and Shallice T (1994) *Connectionist Modeling in Cognitive Neuropsychology: a Case Study*. Hillsdale, NJ: Erlbaum.

Steedman M (1999) Connectionist sentence processing in perspective. *Cognitive Science* **23**: 615–634.

Steedman M (2000) *The Syntactic Process*. Cambridge, MA: MIT Press.

Vaina L (ed) (1991) *From Retina to Neocortex: Selected Papers of David Marr*. Boston, MA: Birkhauser.

Learnability Theory

Advanced article

Ken Wexler, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

The issue of learnability has dominated the study of generative grammar for most of its history. The problem of learnability is how to account for the fact that any normal child can learn any natural language.

INTRODUCTION

There is a theoretical tension between the fact that natural language is learnable and the fact that there exists (so it appears) more than one natural language. If there were exactly one natural language that all humans shared, then it would be possible (though not necessary) that every feature of this language was encoded by the genome. The fact that there is some variability in natural languages means that there is at least some learnability problem.

Empirical and theoretical discoveries in linguistic theory have led to the conclusion that most of the problem of learnability can be solved via the assumption that variability in the bases of natural languages is severely restricted. Let us use the term 'grammar' as it is used in linguistic theory: to mean the underlying computational system of the language (e.g. Chomsky, 1965, 1995), the part that is encoded in the human genome. Then the fundamental axiom of linguistic theory is that there is exactly one grammar. This grammar is called 'universal grammar' (UG). We take UG to mean this one computational system of language that is encoded in the human genome.

This assumption (to the extent that it is true) solves the problem of learnability. However, it is an idealization, as there is at least some variation in human grammar. The grammar of French is different from the grammar of English, which is different from the grammar of Chinese, and so on. However, the results of linguistic theory imply that the differences between these grammars are minor, compared to the fundamental similarities.

We will concentrate on syntax (though similar arguments might apply to phonology, morphology, and semantics).

The fundamental framework of linguistic theory is the 'principles-and-parameters' framework. Principles are the universal parts of grammar, the properties that govern the computational system of language in all languages. Parameters represent those parts of syntax that vary from language to language. Thus, French sets a particular parameter one way, while English sets it another way. Parameters account for the differences in syntactic properties between languages.

PARAMETER SETTING

There are two possibilities for how parameters are determined. The first possibility is that parametric variation is encoded in the genome: for example, a particular individual has a particular value of a parameter as a function of her genetic inheritance. If, say, she comes from a long line of Italian speakers, then a parameter setting for Italian is encoded in her genome.

There is no evidence that there is any genetic variation that predisposes individuals to have particular parameter settings. That is, normal experience (as well as research in language acquisition) shows that any normal child, of whatever genetic background, can learn any natural language. (We ignore questions of language impairment here.)

Thus, we must assume the second possibility, that children set their parameters via experience, as a process of learning.

Of course, the process of learning that allows children to set their parameters interacts with children's UG, the genetically inherited principles that account for the basic properties of language. But the differences between languages result from the child's being exposed to a particular language and learning what the correct parameter-settings are for that language.

For example, many languages (like English and Chinese) have the basic constituents of their sentences in the order SVO (subject–verb–object), as in *Mary ate the cake*. On the other hand, many languages (probably a greater number, in fact) have the order SOV, as in *Mary the cake ate*. These languages include Korean and Japanese (and German, although the order in simple sentences in German is complicated by additional processes).

To account for these differences, one might postulate an 'object' parameter that has two values: V before O, and O before V. The basic idea of parameter setting is that children listen to sentences. If they hear sentences like *Mary ate the cake*, then they set the object parameter as VO; if they hear *Mary the cake ate*, then they set the parameter as OV.

Research indicates that most (perhaps all) parameters are binary: they can take one of two values, like VO and OV for the object parameter. It would thus seem to be a simple matter for a child to set parameters – in principle, experience should tell a child, who has the knowledge of UG from genetic inheritance including a genetically inherited list of parameters, what the value of each parameter is. This is the basic model of parameter setting that is the basis for the theory of variation and learning in syntactic theory.

A theory that has become especially popular in phonology is 'optimality theory'. Here, variation is treated somewhat differently, though in the same spirit. Namely, what distinguishes languages, and what the child has to learn from experience, is the ordering of a set of innately (genetically) specified constraints. There is at least as much need for an innate, genetically specified basis for language in optimality theory, in many respects even more minutely specified than in a principles-and-parameters framework.

EMPIRICAL FINDINGS

Before discussing how well such a theory explains parameter setting, we should ask what the basic empirical properties of parameter setting are. As we have already indicated, children learn the parameters of the language to which they are exposed. But how difficult is this learning? Is it fast and simple and relatively error-free, or is it subject to many errors?

Answering this question of how easily and well children learn parameters is important in order to contrast two types of theory. The first is the theory of language that comes from linguistic theory, whose framework we have just sketched. This theory says that there is a large genetic component to language, and fairly simple learning of a few variable parameters. It predicts that learning of language-particular properties should be simple, and that even very young children should learn their parameters correctly, since the parameters are simply stated given the biological basis for UG.

However, many theories emanating from empiricist psychology hold that the child has little or no knowledge of the computational system of language passed down through the genome, and that all learning, including language learning, starts with a 'blank slate'. If these theories are right, we would expect that language learning would be slow and tedious, that – given the enormous complexity of human language – there would be many errors in learning the properties of any particular language that a child was exposed to.

The empirical answer is quite clear. Extensive research on first-language acquisition has shown that many basic parameters of sentence structure are acquired at an early age. The principle of very early parameter setting (VEPS) (Wexler, 1998) states that basic parameters of clause structure are set correctly at the earliest observable age, that is, the beginning of the two-word stage of language production, at around 18 months of age.

VEPS has been demonstrated by extensive quantitative studies of children's production of utterances. For most syntactic parameters, in order to tell whether a production agrees with a parameter setting or not, at least two words must exist in the utterance (for example, we must observe a verb together with a grammatical object in order to determine whether the object precedes or follows the verb). Thus we cannot determine whether a parameter has been set correctly until a child begins to utter two-word sentences, and this typically happens at around 18 months of age. By this time, children's productions agree with the parameter setting of the ambient language – they vary, for example, according to whether the language is Japanese or English.

Before this age, the child may already have set the parameters correctly: we just cannot tell given the evidence currently available, though experimental studies on infants (e.g. using the 'selective looking paradigm') may eventually tell us that children set parameters before they are 18 months old. The important point is that for crucial parameters for which evidence is available, children apparently never go through a period in which a parameter is set incorrectly or uncertainly. This includes not only relatively straightforward parameters such as the 'verb–object' parameter, but more

subtle parameters such as the 'verb-second' parameter (which says that a finite verb in a simple clause moves to second position) or the 'verb-raising' parameter (which says that a finite verb is raised to a particular position).

Given this empirical result – one of many results from the study of linguistic development in the 1990s – it must be the case that parameters are set relatively easily. There is no trial-and-error period during which parameters are changed by correction or similar processes. Children listen to sentences and produce sentences in accordance with the correct parameter.

Therefore, children must use genetically endowed abilities to help them to set their parameters quickly and well. The drawn-out, error-filled process proposed by empiricist psychologists simply does not occur.

POSITIVE AND NEGATIVE EVIDENCE

Wexler and Hamburger (1973) introduced the term 'negative evidence' to indicate the kind of input that would tell the child that certain utterances were ungrammatical. They argued that empirical research by Brown and Hanlon (1970) meant that negative evidence did not systematically exist. Namely, many (probably most) children at the very young age when grammatical development was taking place were not corrected for speaking ungrammatically. However, Wexler and Hamburger proved that linguistic systems of the 1970s (the 'Standard Theory') could not be learned by any specifiable procedure if the only information available was positive evidence. They therefore argued that even more innate constraints had to be specified, or information had to be available from 'semantic' sources available to the child.

Since these early results, researchers in the field of language acquisition influenced by generative grammar have for the most part assumed that negative evidence does not exist. Both its proponents (e.g. Hoekstra and Schwartz, 1994) and its opponents (e.g. Tomasello, 2000) agree that the generative approach to the empirical study of language acquisition has become dominant. Thus the assumption that negative evidence does not exist has become dominant in the field of language acquisition.

Note that the VEPS result confirms that there is no necessity for negative evidence in order for a child to set their parameters correctly. This is because correct parameter setting has taken place before the child produces utterances that can even potentially be corrected, or misunderstood. The child does not produce any sentence that shows an incorrectly set parameter, so that no matter what an adult's intentions, the adult cannot provide negative evidence. In other words, parameters are set by the psychological process of 'perceptual learning' (Wexler, 2002). Of course, to make this learning possible, there has to be a large amount of genetically specified knowledge of the computational system of language.

THEORETICAL RESULTS

Wexler and Hamburger's learnability result was proved by mathematical methods. They showed that no specifiable learning device (in the accepted sense of 'specifiable device' in cognitive science, an effective system) could learn every natural language specified by linguistic theory. The mathematical framework for that result was the abstract framework of 'text-identifiability' devised by Gold (1967). No effective device could eventually converge on every natural language given sentences from that language. For an extensive mathematical study of learnability in the Gold framework (though one with very few linguistic applications to date), see Osherson *et al*. (1986).

DETAILED MECHANISMS OF PARAMETER SETTING

Since Wexler and Hamburger's result, there have been a large number of learnability results with direct relevance to linguistic theory and the problem of language acquisition, although these have had to diverge from the Gold framework. Thus, Wexler and Culicover (1980) were concerned with the problem of 'boundedness', in essence showing that all grammars specifiable by a theory could be learned from data of bounded complexity, in particular data of degree 2 (no more than an embedding of a sentence within a sentence within a sentence). Their work was motivated by the goal of demonstrating that learning procedures that were psychologically real (i.e. agreeing with a child's abilities) could learn the specified grammars. Wexler and Culicover's 'degree-2 theorem' showed that any grammar of the standard kind could be learned from input of degree no more than 2. To achieve this result (rather complex in its proof), many aspects of UG had to be assumed to be genetically specified. Thus, learnability theory agreed with the results of linguistic theory.

It still was possible in the degree-2 theory for the learner to make a very large number of mistakes, to not be quickly led to the correct learning. The

principles-and-parameters framework was introduced into linguistic theory in an attempt to understand how children can learn language quickly and well, without a large number of errant paths. Theories of learnability are now carried out in this framework. The problem is to find a theory of learning that shows how parameters can be set correctly given bounded evidence (in the sense of Wexler and Culicover (1980)).

This problem has stimulated a number of studies, none of which has been wholly successful in explaining how parameters can be set easily and well. Thus, Gibson and Wexler (1994) showed that even in a very simple system of three linguistically plausible parameters, the learner could end up in incorrect states and not be able to recover. The assumptions that went into this theory were the traditional ones of learnability theory – that learning (parameter setting) took place by the learner's noting that the sentence he or she produced was the wrong sentence, and that the learner had a large amount of grammatical knowledge specified genetically, but only limited processing and memory capacities. With greatly expanded processing and memory capacities, at least in principle, correct learning might result. However, the most natural and plausible theory is probably one in which the child has limited memory capacities while setting parameters, so that most of the work should be done by the genetically specified UG. Remember that the VEPS result indicates that parameters are set before 18 months of age. The learner cannot be assumed to have memory or processing capacities that go beyond those of an 18-month-old child.

Some researchers (e.g. Dresher and Kaye, 1990) attempt to solve the learnability problem by assuming that the child's genetically inherited knowledge of UG is supplemented by a genetically inherited set of 'cues': patterns of input that would be used to set a parameter a particular way. Thus learning does not take place by the child's noting that he or she makes an error and has to correct it. Rather, the child has to note that a (possibly quite complex) pattern (possibly using large memory resources) exists in the input.

No such cue theory has been successful in showing that a range of linguistically natural parameters can be set correctly. The problem is that there is a great deal of interaction between parameters, so that if n parameters have to be set, 2^n cues may have to be built into the learner through inheritance, one for each combination of parameter values. This would seem very unlikely.

On the other hand, no theory of the 'learning on errors' type has been shown to be successful for a large set of parameters either. There may be missing concepts in the theory of learnability. Even given a large amount of innately specified linguistic knowledge, the detailed theory of parameter setting is not fully understood.

However, a great deal of empirical evidence on language acquisition, as well as linguistic theory, converges to show that the general thrust of the theory is correct. Language learning turns out to be continuous with other biological abilities, as suggested by Lenneberg (1967). There is a large core of genetically specified pieces of knowledge, together with a learning system that accounts for the effects of experience. Unification of the study of language and the study of biology is in principle possible.

References

Brown R and Hanlon C (1970) Derivational complexity and order of acquisition in child speech. In: Hayes JR (ed.) *Cognition and the Development of Language*, chap. 1, pp. 11–54. New York: John Wiley and Sons.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Dresher E and Kaye M (1990) A computational learning model for metrical phonology. *Cognition* **34**: 137–195.

Gibson E and Wexler K (1994) Triggers. *Linguistic Inquiry* **25**: 407–454.

Gold EM (1967) Language identification in the limit. *Information and Control* **10**: 447–457.

Hoekstra T and Schwartz BD (eds) (1994) *Language Acquisition Studies in Generative Grammar: Papers in Honor of Kenneth Wexler from the GLOW 1991 Workshops*. Amsterdam: John Benjamins. [Introduction.]

Lenneberg E (1967) *Biological Foundations of Language*. New York: John Wiley and Sons.

Osherson D, Stob M and Weinstein S (1986) *Systems That Learn*. Cambridge, MA: MIT Press.

Tomasello M (2000) Do young children have adult syntactic competence? *Cognition* **74**: 209–253.

Wexler K (1998) Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. *Lingua* **106**: 23–79.

Wexler K (2002) Lenneberg's dream: learning normal language development and specific language impairment. In: Schaeffer J and Levy Y (eds) *Language Competence Across Populations: Towards a Definition of Specific Language Impairment*, pp. 1–103. Mahwah, NJ: Erlbaum.

Wexler K and Culicover (1980) *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

Wexler K and Hamburger H (1973) On the insufficiency of surface data for the learning of transformational languages. In: Hintikka KJJ, Moravcsik JME and Suppes P (eds) *Approaches to Natural Language*, pp. 167–179. Dordrecht: Reidel.

Lexical Access

Intermediate article

Anne Cutler, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

CONTENTS

Lexical access occurs in speaking when the pronounceable word forms are found which encode the concepts the speaker wishes to communicate; and it occurs in listening when the meaning is found which is expressed in the word forms the listener has heard.

THE LEXICON

Communication of a message from a speaker to a hearer can take the form of word sequences which the speaker has never before produced and the listener never heard before. *This article tries to give a plain introduction to lexical access* – most people have probably not heard this sentence, but it is no problem for listeners presented with such a new utterance to understand it. This is because utterances, though they may be uniquely constructed, are built up of discrete units (such as *article*, *give*, *plain*, *access*) which speakers assume their listeners will already know. These discrete units we call 'words', and the stock of words which speakers of a given language use in speaking and listening constitutes the vocabulary. The mental lexicon is the mental representation of the vocabulary.

Entries in the mental lexicon may correspond to words such as *give* and so on, but they may also be other forms which speakers store as discrete units: fixed phrases such as *bon appetit*, manipulable idiomatic phrases such as *let the cat out of the bag*, productive derivational affixes such as *re-* or *un-* or *-ish*, inflections for pluralization, tense and so on, stems which occur in multiple words. That is, the forms in the mental lexicon are those which language users store as discrete entities, and they may or may not coincide with forms which are written as discrete words. (Nevertheless, in this text 'words' will serve as shorthand for lexical representations.)

Speakers begin with the intention to communicate a message, and to achieve this they must encode the message in words and articulate the resulting string of words. Listeners hear the string of words, which they must decompose into its word parts, and their first step in decoding the message is identification of the meaning associated with each word. Lexical access in speaking is the process of finding the lexical representations to express the desired meaning. Lexical access in understanding is the process of finding the lexical representations which correspond to the heard sounds. Both speaking and understanding thus require lexical access, but the two processes are the reverse of one another. It has been a matter of dispute whether there is a unitary mental lexicon that is drawn upon both in speaking and understanding.

LEXICAL ACCESS IN SPEAKING

The speaker begins with an intended message and must convert this into spoken sound patterns. Models of word production (Dell, 1986; Levelt *et al*., 1999) agree that the conversion process consists of multiple stages, and in particular that retrieval of meaning and retrieval of sound are separate processes.

Retrieval of Meaning

In the model proposed by Levelt *et al*. (1999) it is assumed that the speaker effectively translates the intended message into individual lexically represented concepts. Each such concept has a unique lexical representation (the lemma) associated with it, and activation of the concepts expressed in the message causes activation of the lemmas to which they are connected. The lemmas contain the syntactic constraints associated with the concepts.

For example, Figure 1 depicts a single part of the lexical network which could be involved in expressing the message that this article is trying to give – *a plain introduction*. The lemma for *plain* is assumed to be activated by a corresponding conceptual node selected to express an attribute; the

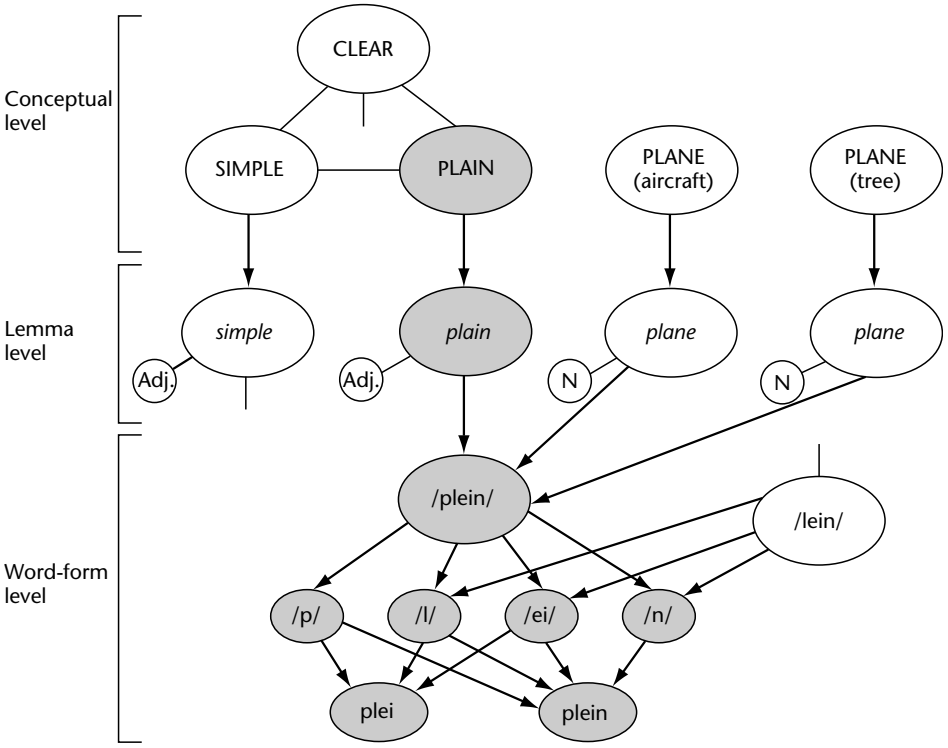


Figure 1. Fragment of a lexical network for word production.

lemma *plain* accordingly includes the syntactic attribute of being an adjective. Other syntactic information which can form part of a word's lemma representation includes the expression of, for instance, tense and number (in verb lemmas), case (in noun and adjective lemmas for languages which mark case), or grammatical gender (in noun lemmas for languages with gender).

Related conceptual nodes are connected to one another, but each conceptual node is connected to only one lemma. Lemmas are not connected to one another at all; they receive activation from concepts, and pass activation to the phonological forms (the word-forms) which express them as sound patterns, but they do not pass activation to other lemmas.

Activation from the conceptual node may spread to other conceptual nodes to which it is connected, and these nodes may in turn send activation to their associated lemmas. At the lemma level, however, a selection takes place; the most-activated lemma is chosen to pass activation on to the word-form level.

Retrieval of Sound

The lexicon includes separate representations of phonological form, the word-forms. Each lemma is connected to just one word- form. (An exception may perhaps occur in the special case of a single word for which a speaker knows two pronunciations; for example, some British speakers can pronounce the word *garage* to rhyme with either *marriage* or *mirage*.)

Word-forms, of course, may be connected to more than one lemma, because word-forms are very often homophonous – two words can sound the same. Thus the word-form to which the adjective lemma *plain* sends activation has many other connections – from the noun *plain* (a level tract of land), and from the various lemmas *plane* (an aircraft; a kind of tree; a surface; to make level; and so on). Figure 1 shows just two of these. None of these lemmas are connected to one another, nor to the adjective *plain*.

Evidence for the connection of word-forms to more than one lemma, and in general for separation of the lemma and word-form levels, appeared in some studies of frequency effects in production. Speakers can produce common words more rapidly than uncommon words; but if an uncommon word happens to be homophonous with a common one (e.g. *plane* the name of the tree and *plain* the adjective), then the uncommon word is produced as rapidly as the common one. Apparently it is the frequency of the word-form which matters.

The word-form representation contains information about how the word is pronounced, in the form of connections to the appropriate phonemic representations, together with instructions for the compilation of these phonemes into possible syllables. In actual utterances, syllable boundaries are determined by the string of words as a whole rather than only by the lexical representations of the individual words. A syllable-final phoneme may become syllable-initial if the word following it begins with a vowel (and, in English and similar languages, especially if that second word is a function word). In *plane that wood!*, the [n] is syllable-final, but in *plane it!*, the [n] may be syllable-initial, and the precise phonetic realization of the [p] can differ in the two positions. Therefore the phonemic representations activated at the word-form level are not fully compiled, because their final compilation depends on the rest of the utterance. In Figure 1, the word-form of *plain* is connected to four phoneme nodes (three of which, for example, it shares with the word-form of *lane*), and can activate two potential syllable nodes.

Slips of the Tongue

Errors can arise at all points in the process of producing an utterance. Slips of the tongue (see Fromkin, 1973) fall into two major classes: mis-selections and mis-orderings. Among the former are errors of lexical access, in which the wrong word is selected. These in turn are of three main kinds. Most common are mis-selections in which the meaning is similar but not identical – *next* instead of *last*. Mis-selections also occur in which the erroneously selected word is similar to the target word in sound though not in meaning – *single* instead of *signal*. And finally, blends of two words can occur – *bookstop* as a blend of *bookstore/bookshop*.

How do such errors arise? Semantic errors, understandably, arise in the process of accessing word meaning. They can be explained in terms of activation spreading among related conceptual nodes, leading to a lemma related in some way to the target being activated to the same or a greater extent than the target lemma. The nonintended lemma (*next*) is selected and from that point on word production proceeds just as if *next* had been the intended word.

Blends could also arise in the same way during the access of meaning representations. The difference between semantic errors and blends is that in the latter case the problem is carried beyond the lemma level: two equally activated lemmas could simultaneously pass activation to the word-form level, the two word-forms would then activate phonetic components at the articulatory level, and at this point the potential problem of competing candidates would perforce be resolved since the output would allow utterance of only one of the available syllable onsets, one of the available syllable nuclei, and so on.

Finally, form-based selection errors (malapropisms) could arise at the word-form level; activation spreading to the phonemic segments could be misassigned, leading to the syllabic packets appropriate for a similar form (*single*) being articulated, instead of those required by the target form (*signal*). The existence of both purely form-based and purely meaning-based slips lends further support to the two-stage model of lexical access in speaking. The 'tip-of-the-tongue' phenomenon also supports this division; a speaker who has a word on the tip of the tongue has accessed the word's meaning (and, in languages with grammatical gender, the speaker has also accessed the gender), but the pronunciation, i.e. the word-form representation, is for some reason temporarily inaccessible.

LEXICAL ACCESS IN UNDERSTANDING

The listener's task is to identify the words making up the speaker's message. Languages do not make life easy for listeners. The vocabulary of any language consists of tens or hundreds of thousands of words on average made up of only about 30 speech sounds, or phonemes (English has – depending on dialect – 40 or more phonemes, and is thus a relatively phoneme-rich language). Thus words inevitably resemble one another, and short words may fortuitously occur within longer ones; in consequence, in any spoken utterance the words uttered by the speaker are not the only words present in the speech signal. This would pose the listener little problem if words in speech were reliably demarcated with signals indicating where each word ends and the next begins. In many written texts (such as this one in English) such helpful signals are indeed available – white spaces between the words. That is not true of speech; spoken utterances reach the listener's ear as a continuous stream.

Segmentation

Figure 2 is a spectrogram of the phrase *A plain introduction to lexical access*, uttered by an American speaker. The gaps in the spectrogram correspond to

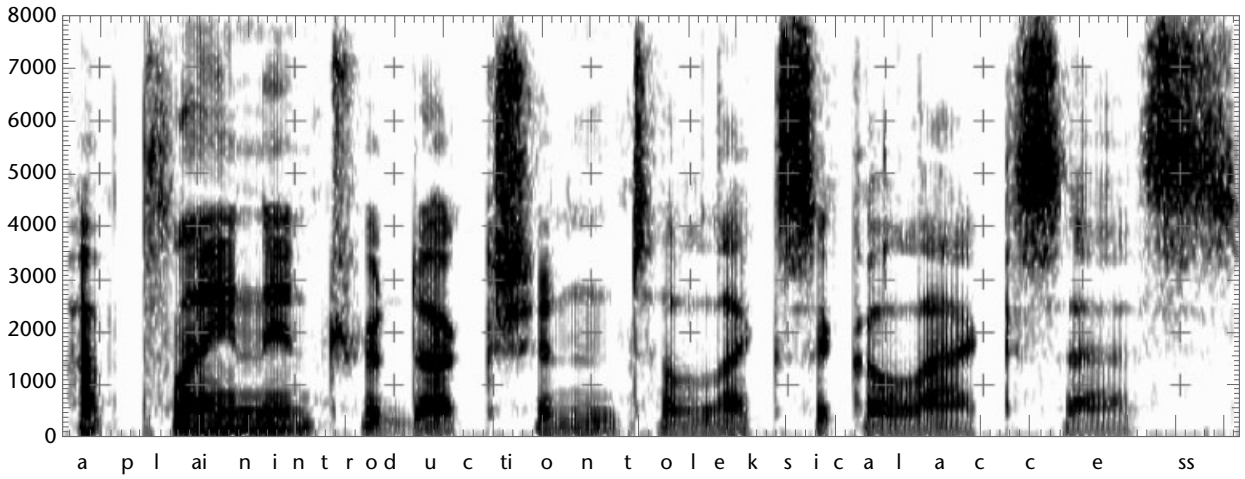


Figure 2. Spectrogram of the utterance 'A plain introduction to lexical access'. The display represents frequency on the vertical axis against time on the horizontal axis, with greater energy represented by darker shading. The transcription is aligned as closely as possible with the corresponding sounds on the spectrogram. There are clear breaks in the speech signal, but these do not correspond to breaks between words. The breaks occur whenever a speech gesture actually stops the flow of air through the speaker's vocal tract for a brief period – for instance the sounds 't' and 'd' in 'introduction'. In contrast, the individual words adjoin to one another continuously, without a break.

speech sounds, not to pauses between words. Certain speech sounds cause a momentary obstruction of the vocal tract, resulting in a brief period of silence – the consonants [p], [t], [k] and [d] are among such stop consonants and they occur in this utterance – [p] in *plain*, [k] in *introduction*, *lexical* and *access*, and so on. In contrast, there is no pause at word boundaries – *plain* runs continuously into *introduction*, *lexical* into *access*. The listener must find the words despite the absence of clear word boundary signals. This utterance in Figure 2 also shows how easy it is for shorter words to appear by accident within longer ones – thus *plain* contains *play*, *lay*, and *lane*; *introduction* contains *duck*; *access* contains *axe*. Words may even occur across word boundaries, because of the continuous flow of one word into another: thus across the boundary of *lexical* and *access* there is a string of sounds consistent with *lack* and *lax*.

Listeners can use knowledge about their language to help them segment continuous speech into its component words. For instance, certain phoneme sequences cannot occur within syllables (e.g. [nl] in English), or even within words (e.g. [mg] in English). Listeners can use this information to find words; the beginning of *lexical* would thus be clearer in the phrase *in lexical access* than it is in *to lexical access*. In English, words more often than not begin with a stressed syllable, and this information is also useful in segmentation; *plain*, *lexical* and *access* all begin with a primary stressed syllable, and *introduction* has secondary stress on its initial syllable. Only the function words *a* and *to* are unstressed; this also is typical of English, and also used by listeners to assign grammatical category. Of course, all these types of information differ across languages, so that the way they are exploited in listening will also be language-specific.

Activation and Competition

Models of spoken-word recognition (e.g. Marslen-Wilson and Welsh, 1978; McClelland and Elman, 1986; Luce *et al*., 1990; Norris, 1994) agree that words which are present in the speech signal are automatically activated in the listener's mind. Even words which are accidentally present (e.g. *lane* in *plain*) can be activated.

Simultaneous activation of all candidate words which are supported by information in the input means that a unique selection within the lexicon is not immediately possible; selection is achieved by allowing the activated words to compete with one another until one or more winners emerge. This process of competition provides a solution to the problems arising from the similarities between words and the embedding of words within other words; *plain* and *lane* compete for one portion of the input, *axe* and *access* for another, and so on.

Concurrent activation has been a feature of all models of spoken-word recognition since Marslen-Wilson and Welsh (1978). Competition was first

proposed in the TRACE model (McClelland and Elman, 1986), and in the same form – competition via lateral inhibition between competitors – it is the central mechanism of the Shortlist model (Norris, 1994). In other forms it is also found in the other main models currently available, such as the Neighborhood Activation Model (Luce *et al*., 1990) and the latest Cohort model (Gaskell and Marslen-Wilson, 1997).

There is substantial evidence of activation of words embedded within other words, and of simultaneous activation of partially overlapping words. This sort of evidence is gathered in laboratory studies in which listeners hear words or parts of words, and also perform another task such as deciding whether a visually presented letter string (e.g. GIVE or FLERK) is a real word or not. Hearing a word facilitates acceptance of a related word – it is easier to decide that GIVE is indeed a word after just hearing *take*, for instance. Activation of embedded words has been proposed because it has been observed that they too provide such facilitation – e.g. the recognition of GIVE could be facilitated by hearing *mistake*. And simultaneous activation of partially overlapping words has been supported by experiments which showed that hearing a fragment such as *lec-*, which could be the beginning of several words (e.g. *lexicon*, *lecture*) facilitated words related to all of them (such as DEFINITION, COLLEGE).

The activation process is continuous, and can effectively use early co-articulatory information, as is shown by experiments in which words are cross-spliced with other words and with nonwords. It is hard to decide that *shrud* is indeed a nonword if its spoken form includes a 'shru-' taken from an utterance of *shrub* or *shrug*. Thus even nonwords which resemble words will activate lexical information, and the more similar the nonword is to a real word, the more effective it will be in activating word candidates. Activation of a lexical representation thus does not obligatorily require full presentation of the corresponding word form; partial information (in partial words, for instance, or in nonwords which in part overlap with real words) suffices to produce partial activation.

Simultaneous multiple activation of words does not, of course, necessarily entail competition between those words. Models such as TRACE and Shortlist predict that simultaneously activated words will compete by passing inhibition to one another. This too has been demonstrated in the laboratory. One way to show such effects is in experiments in which listeners hear nonsense strings, some of which have a real word in them; the task is to find any such real word. In the strings *obzel crivlish lakfid*, for example, only the last string contains a real word – *lack*. When such nonsense strings activate competing words, listeners find it harder to find the embedded words. Thus, *mess* is harder to find in *domess* (which could partially activate *domestic*) than in *nemess* (which activates no competitor). The more competing words may be activated, the more the recognition of embedded words will be inhibited. This kind of finding is direct evidence for competition between words.

Figure 3 shows simulations of the input *lexical access* in the Shortlist model (Norris, 1994). The input activates temporarily words such as *lecture* and *excema* which are eventually defeated by competition from *lexical*; *collapse* and *lax*, which straddle the word boundary, are also briefly active, but joint competition from *lexical* and *axe* defeats them in turn; initially, *axe* enjoys stronger activation than *access*, but by the end of the input the two most activated words are, as intended, *lexical* and *access*. Competition between candidate words which are not aligned in the signal helps to achieve segmentation of the speech stream into individual words. Thus although the recognition of *lexical access* involves competition from other words, this is eventually overcome by joint inhibition from the actually spoken words. The competition process, and its concomitant constraints, can so efficiently result in victory for words which are fully present in the signal that concurrent activation of partially present words, or of words embedded within other words, is simply a low-cost by-product of the efficiency with which the earliest hints of a word's presence can be translated into activation.

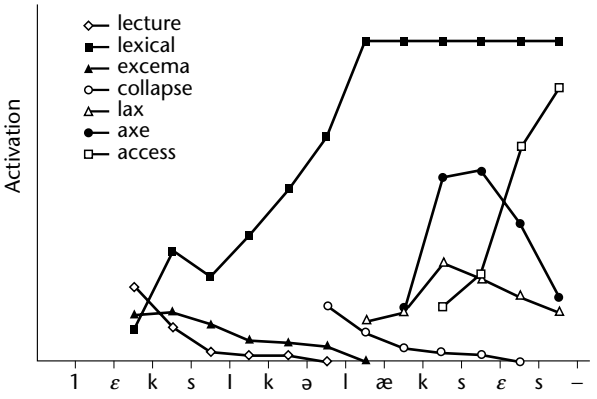


Figure 3. Shortlist simulations of word activations given the input 'lexical access'.

An activated word form passes activation on to the meaning associated with it. In word recognition models activation is assumed to cascade continuously through the levels, rather than to wait for a clear winner at each level. Thus activation of the word form *plain* can pass activation on – in this case, of course, to several meanings rather than just one. The string *a plain* is ambiguous as to whether the word-form *plain* is noun or adjective, and which of several interpretations of either syntactic form might be intended. Empirical studies (e.g. with the priming tasks described above) have indeed shown that multiple meanings of a homophone can be simultaneously accessed. However, the integration of the word's meaning with the rest of the syntactic and semantic context is also assumed to be a continuous process, as a result of which the intended meaning can very rapidly be selected. In the context of a story about flight, the message interpretation component might very quickly select the noun *plane* in the sense of aircraft, on semantic grounds. Syntactic considerations would suggest that *A plain* followed by *introduction* is best interpreted with the adjectival meaning of *plain*.

Slips of the Ear

Errors can, again, arise at any point in the process of comprehending an utterance. Errors specifically in the process of lexical access imply that the outcome of the competition process is a word or sequence of words which does not correspond to what the speaker actually said. Such slips of the ear are particularly likely when the input is unclear – due to background noise, for example, so that weak acoustic support is provided for a number of possible candidate words. Thus an old popsong contained the words *she's a must to avoid*; very many listeners independently interpreted this as *she's a muscular boy*.

Slips of the ear often (as in this example) produce implausible results, and may produce reports of low frequency words or even nonwords (see Bond, 1999). This suggests that frequency and plausibility do not play a strong role in the activation and competition process.

The operation of the strategies constraining the competition process can be seen in the patterns of slips of the ear. In English, for instance, slips are much more likely to produce interpretations in which strong syllables are word-initial and weak syllables are not word-initial – as in the example, in which *must to avoid* (strong, weak, weak-strong) is thought to be *muscular boy* (strong-weak-weak, strong).

THE INDEPENDENCE OF LEXICAL ACCESS

A recurring question in research on language processing is whether the different components of the process are independent, or whether they can be influenced by processing decisions made by other components which are logically later in the processing chain. In models which preserve independence ('autonomous models'; e.g. Levelt *et al*., 1999 for word production, or Norris, 1994 for word recognition), information feeds forward only – in speaking, from conceptual formulation to articulation, and in listening, from acoustic-phonetic processing to message interpretation. In models which allow non-independence ('interactive models'; e.g. Dell, 1986 for word production, or McClelland and Elman, 1986 for word recognition), there is feedback, e.g. from phonological encoding to lemma selection in speaking, or from word activation to phonetic processing in listening.

This is one of the most controversial issues in cognitive science, and it is a subject of continuing debate. For lexical processing in both speaking and listening, however, strictly feedforward models have been proposed which adequately account for the known empirical evidence (Levelt *et al*., 1999; Norris *et al*., 2000). Considerations of parsimony suggest that the simplest model consistent with the evidence should be preferred; the proponents of these models therefore argue that lexical processes should not be held to involve feedback connections.

In speaking, considerations of added efficiency have not been invoked in favor of feedback; the strongest evidence which has been used to argue for feedback in fact comes from performance failure, i.e. from slips of the tongue. First, semantic errors are more likely if the error is similar in sound to the intended word; that is, *plane* is more likely as a semantic error for *train* than for *boat*. And second, single-phoneme errors tend to produce another real word rather than a nonword: *plane* is more likely to be mispronounced *plate* than *plake*. However, both of these tendencies also arise in Levelt *et al*.'s (1999) feedforward model. Errors which result in real words are more likely to be missed by the monitoring component in that model. The split-stage architecture in the model further provides a natural explanation for phonological similarity in semantic errors. If two equally active lemmas are passed on to the word-form level instead of one, the error word-form is more likely to be selected at that level if it is similar to (and its phoneme connections hence receive activation from) the intended word-form.

In listening, similarly, the performance of the lexical processor cannot actually be improved by feedback to prelexical processing. Suppose the input *plai-*, with co-articulatory information in the vowel, is enough to activate *plain* to such an extent that it is the most well-supported word in the ongoing lexical competition process. Feedback from the lexicon can then determine that phonetic processing identifies an [n]. But this offers no assistance to the recognition of *plain*, since *plain* is already the most activated word. It would also not assist word recognition to pass down information from a number of equally supported competing words; if *plain*, *place*, *plate* and *plague* all passed down activation to their separate terminal phonemes, the phonetic processing level would simply receive more incorrect than correct information. The information in the signal would ultimately have to determine the decision at the phonetic level, so that no useful role would be played by the feedback.

References

Bond ZS (1999) *Slips of the Ear: Errors in the Perception of Casual Conversation*. New York: Academic Press.

Dell GS (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–321.

Fromkin VA (1973) *Speech Errors as Linguistic Evidence*. The Hague: Mouton.

Gaskell MG and Marslen-Wilson WD (1997) Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes* **12**: 613–656.

Levelt WJM, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 1–38.

Luce PA, Pisoni DB and Goldinger SD (1990) Similarity neighborhoods of spoken words. In: Altmann GTM (ed.) *Cognitive Models of Speech Processing*, vol. 7, pp. 122–14. Cambridge, MA: MIT Press.

Marslen-Wilson WD and Welsh A (1978) Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* **10**: 29–63.

McClelland JL and Elman JL (1986) The TRACE model of speech perception. *Cognitive Psychology* **18**: 1–86.

Norris D (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition* **52**: 189–234.

Norris DG, McQueen JM and Cutler A (2000) Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences* **23**: 299–370.

Further Reading

Brown C and Hagoort P (eds) (1999) *Neurocognition of Language*, chaps 4 (Levelt), 5 (Cutler and Clifton), and 7 (Price, Indefrey and van Turennout). Oxford: Oxford University Press.

Friederici A (ed.) (1998) *Language Comprehension: A Biological Perspective*, chaps 1 (Frauenfelder and Floccia), 2 (Cutler), and 3 (Zwitserlood). Heidelberg: Springer.

Grosjean F and Frauenfelder UH (eds) (1996) Spoken word recognition paradigms. Special issue of *Language and Cognitive Processes* **11**: 553–699.

Lexical Development

Advanced article

Cecile Maxine McKee, University of Arizona, Tucson, Arizona, USA

CONTENTS

Lexical development refers to children's acquisition of meaningful and grammatical elements of the lexicon.

INTRODUCTION

What comes first? Many people think that words are children's first linguistic elements. This may be because words are the first meaningful units that we recognize in children's observable behavior. With the exception of scientists who study this topic, people rarely consider infants' production of well-formed but meaningless syllables like 'bababa' as evidence of language knowledge. However, words are not first in any scientifically interesting sense. Children develop multiple aspects of linguistic knowledge simultaneously, and they do so well before producing their first words. (*See* **Phonology and Phonetics, Acquisition of**)

One theme in this encyclopedia is information; another is information processing. The study of lexical development relates to both. First, lexical development refers to children's mastery of a kind of information. However, the evidence for such mastery is affected by information processing. This distinction between information and the processing of information is comparable to the distinction between competence and performance. Developmentalists face subjects who are clearly gaining both target competence (i.e. representation of information) and proficient performance (i.e. processing of that information). Limitations in either or both can be sources of error for learners. For example, the holophrastic stage when children emphasize one-word utterances may reflect processing limitations as much as or more than informational limitations. Language processing systems, including those governing production, are developing at the same time as children are learning the very language(s) that these systems manipulate. Crosslinguistic considerations bring home the point. Although some aspects of language processing may be the same across languages and therefore at least a candidate for early development or even hard-wiring, some are language-specific and therefore not determinable until the learner's target language is – at least partly – in place (for example, see Levelt, 1989). (*See* **Performance and Competence**; **Speech Error Models of Language Production**)

An example of the complex relation between competence and performance comes from Smith's study of his son Amahl (Smith, 1989), which shows how the study of lexical development relates to both the information theme and the information processing theme. There was a time in Amahl's linguistic development when he pronounced 'mouth' and 'mouse' the same. Having an interest in competence–performance questions, Amahl's father devised a test. He put pictures of a mouth and a mouse in one room. From another room, he asked Amahl to bring him one or the other picture. The child repeatedly performed perfectly, showing that his representation of these words was closer to target than his utterances suggested. Were that not true, Amahl could not have perceived the distinction that his father made. That Amahl's production error was systematic requires an explanation in terms of both listed knowledge (e.g. the arbitrary information about the pronunciation or form of these words) and procedural knowledge (e.g. the processes that relate the sounds in 'mouth' and 'mouse' and the production mechanisms that apply those processes). Interestingly, this example is a case in which the child was still learning relevant procedural knowledge. It also shows how research on lexical development can be done.

The study of learning also informs this area of cognitive science. Whether the relevant learning rests on general mechanisms or on mechanisms specific to language, lexical development forces us to consider questions such as what a learner can do with partial information and/or less-than-smooth processing routines. In other words, the lexical system – both in terms of the information it

represents and the processing of that information – changes over time. Children begin to use language before it is fully in place. Research on lexical development thus aims at a moving target. Description of the adult lexicon and explanations of lexical development in children must recognize that. (*See* **Learning, Psychology of**)

WHAT IS A WORD?

If words are the focus of research on lexical development, we need to know what a word is. Languages vary in how they package lexical information: in languages such as Navajo, American Sign Language and Spanish, a 'word' may represent all of a sentence's major elements (e.g. subject, verb and object, as in Spanish *commelo*, 'you eat it'). Such examples reflect a language's syntax and morphology, e.g. whether the language permits null subjects or objects and, if so, whether the verb marks such arguments. Let us consider English. Literate English users might define a word as a meaningful group of letters with space on either side. But that definition is of no use to the preliterate, word-learning machine inside your typical 2-year-old human. It runs into trouble elsewhere as well. Consider 'greenhouse' and 'green house'; 'the' and 'of'; the ambiguous word 'bank'; or the idiom 'pushing up daisies' (meaning 'dead'). Consider 'hood', as in 'neighborhood': in some dialects of English, 'hood' meaning 'area of residence' can stand alone without 'neighbor'. Is it then a word? Perhaps we will agree that it is a word only when it appears in our favorite dictionary. In a decade then, 'hood' may be a word. In the meantime, are such elements targets of lexical development? In other words, do children learn them? Yes! Indeed, children's ability to learn words of dubious official status may even contribute to language change. These and related puzzlers are why this article is called 'lexical development' rather than 'word development'. The latter would focus on an object of doubtful definition and force out too much of great interest (especially if cognitive science aims to understand cognition across languages and cultures). (*See* **Idioms, Comprehension of**; **Syntax**; **Morphology**)

The term 'word' is just shorthand then – often useful, but not very precise. Distinctions from linguistic theory provide greater precision. The distinction between lexical and functional categories is emphasized because these categories are probably learned in different ways (Pinker, 1991). The lexical categories include elements such as nouns (e.g. 'house'), verbs (e.g. 'slap'), and adjectives (e.g. 'green'). The functional categories include elements such as complementizers (e.g. 'that'), determiners (e.g. 'the') and inflections (e.g. '-ing'). In English, lexical categories are usually free morphemes (words, if you will), and their semantic content is relatively clear. Functional categories are either free or bound, and their grammatical function is relatively clear while their semantic content is less obvious. If you think children can only learn meaningful elements, then you must think they start with words; but if you think children can pick up any linguistic pattern, then you must allow that they start with whatever occurs often enough to be a pattern. Questions about what comes first do interest researchers, so we will return to this issue below.

To a large extent, the fundamentally syntactic distinction between lexical and functional categories rests on the notions of selection and phrase structure. That is, one way to define a verbal inflection or a noun is by its associations with other elements. This approach contrasts with the kind of referent-based definition many of us learned in school, e.g. 'a noun is a person, place, or thing'. Some such associations are very strong. For example, all verbs can host '-ing'. The relatively new nouns 'email' and 'fax' easily became 'emailing' and 'faxing' as soon as English speakers began to use them as verbs. This example shows an association of two categories, one lexical and the other functional. As we shall see below, linguistic elements select for or prefer association with a variety of kinds of syntactic or semantic information. For now, consider the significance of such associations for learners. For one thing, if cross-category relations characterize the target system, learners must do much more than link the form of each word with its meaning. They must also acquire each element's selectional restrictions. If an element is at least partly defined by such associations, then learning the element includes learning its associations. (*See* **Phrase Structure and X-bar Theory**)

Interestingly, most research on lexical development has emphasized children's acquisition of the lexical categories and links to semantic and conceptual development. See Bloom (2000) for a review of this emphasis. There is less research on children's acquisition of the functional categories. What there is links to syntactic and morphological development. A number of interesting questions that cannot be addressed here should nevertheless be noted. For example, one might ask whether all children map the same initial set of concepts onto words or whether all languages draw from the same full set of category options. For the purpose

of this article, it is important just to appreciate that the target lexicon contains subclasses of elements (i.e. several different types of information). It should be clear that the business of lexical development is challenging. We proceed now to theoretical explanations of that challenge. (*See* **Semantics, Acquisition of**; **Concept Learning**)

TARGET LEXICON

At a general level of description, a lexicon is a person's mental dictionary. It lists the meaningful and/or grammatical elements known by that individual: some of these elements are well-behaved words like 'green' and 'the', some are word parts like '-ing' or perhaps '-hood', and some are collections of both, such as 'pushing up daisies'. If lexical and functional elements are represented separately and processed differently, then it is reasonable to presume that such elements are also learned differently. To put it another way, these classes of linguistic elements may be influenced by distinct learning mechanisms. (*See* Pinker, 1991 for a comparison of associationist learning to rule-and-representation learning.) A particularly striking finding from this research area is the speed of word learning. Many researchers have concluded that young children go through a period of unusual lexical development when they average a new word every waking hour (sometimes called the vocabulary spurt or the naming explosion). However, the existence of such a stage has been challenged, for example by Bloom (2000). For one thing, unlike some other areas of linguistic development, word learning is lifelong. Let us thus focus on the feat itself rather than the question of whether children differ from adults in this domain. Lexical acquisition is undoubtedly fast. How does it work? (*See* **Lexical Acquisition**; **Lexicon**; **Word Learning**)

Lexical Categories and Meaning

The question of how children learn the meanings of words has been in the scientific spotlight for a long time. The seventeenth-century philosopher John Locke, for example, described children's language learning as a consequence of adults pointing out and naming things like 'white' and 'dog'. As Gleitman (1990) emphasized, this view of language acquisition persists both in common sense and in scientific theory. Regarding the latter, she characterized the theory of semantic bootstrapping as reflecting a modern version of Locke's view (e.g. Grimshaw, 1981; Pinker, 1984). While semantic bootstrapping is primarily a theory about children's acquisition of syntax, not about their lexical development, it begins with word learning. During semantic bootstrapping, words link to syntactic categories through their meanings, which are hypothesized to be observed in the world and associated with lexical categories. This hypothesis narrows the learner's initial focus to meaningful elements, perhaps even to concrete and imageable objects. In some sense then, semantic bootstrapping assumes that words are first (i.e. lexicon precedes syntax). Syntactic bootstrapping, on the other hand, is a theory about children's acquisition of word meaning. It does not address children's very first linguistic steps. Because its central hypothesis is that word learning is informed by observation of a word's syntactic context (i.e. the range of its associations with other linguistic elements), it applies only after some acquisition of syntax has already occurred. So semantic and syntactic bootstrapping explain different aspects of language acquisition. Interestingly, they both focus on lexical categories. While Locke's and Pinker's paradigm examples involve nouns – perhaps more properly noun phrases; see Bloom (1994) – where the link to observable referents is plausible, Gleitman's paradigm examples are verbs. It is less obvious that word-to-world mappings will work for verbs (or any other category that is complexly related to the observable world). We return to this point in the section on argument structure below. (*See* **Syntax, Acquisition of**)

Developmentalists' focus on words, especially nouns, has shaped many theoretical questions. For example, the semantics suggested to mediate the learner's progress towards the target lexicon has emphasized features of objects (e.g. animate versus inanimate, countable versus substance). This emphasis is most obvious in hypotheses aimed at explaining the speed of early word learning.

Constraints on Word Learning

There are many proposals for principles or biases that restrict the learner's initial hypotheses about words. Note that these are not general restrictions. Rather, they are specific to word learning. They include – among others – the whole object constraint (Markman, 1989), the shape bias (Landau *et al.*, 1988), and the principle of contrast (Clark, 1993). To clarify the motivation for such constraints, let us examine Clark's principle of contrast. The idea here is that children will learn more if they assume that difference in form entails difference in meaning. One source of evidence for

children's use of this principle comes from how they retreat from overextension errors. In an overextension of the word 'dog', the speaker uses it to refer to dogs as well as to animals that share features with dogs (e.g. other four-legged, fuzzy animals like cats and sheep). When children who overextend 'dog' observe the word 'cat' referring to what they previously called 'dog', they stop referring to cats as 'dogs'. The principle of contrast forces the learner to adjust the meaning of 'dog' when a word with a different form overlaps in reference. This brief mention of how developmental errors can motivate constraints on word learning relates to an earlier point. It is likely that some such errors reflect performance failures rather than being direct reflections of lexical competence.

Argument Structure

We turn now to a topic that might at first seem minimally related to word learning. Recall the hypothesis that lexical elements encode information about other elements. These associations might affect linear relations (as in the strict word-order rules found in English) or morphological relations (as in the rich inflections found in Spanish). That information is part of what learners must master – in addition to more obvious information about a word, such as its form (which could be instantiated as sounds in a spoken language or signs in a signed language). To further appreciate the complexity of lexical development, let us turn to more detailed consideration of selectional restrictions. I will exemplify the notion with some verbs. Recall that these restrictions or associations can involve different kinds of linguistic information. To exemplify syntactic restrictions, compare intransitive verbs (e.g. 'sleep'), transitive verbs (e.g. 'devour') and ditransitive verbs (e.g. 'give'). The examples 1 to 3 show these three subcategories (an asterisk precedes example sentences that are considered to be ungrammatical or ill-formed). A subcategory can be defined in terms of the types and functions of the phrases that are associated with it. These phrases are a verb's arguments, so the argument structure of a verb helps define it. In 3a, for example, we see that a ditransitive verb can have two noun phrase arguments. Words can also encode semantic restrictions. For example, the verb 'drink' prefers its direct object to be liquid, while 'break' prefers it to be rigid. Such preferences help define these words. Poetic use of words, however, can involve fiddling with such restrictions, as in 'drinking a fragrance' or 'breaking a silence'.

1. sleep
   a. Carol is sleeping.
   b. *Carol is sleeping the bed.
2. devour
   a. Carol devoured the casserole.
   b. *Carol devoured.
3. give
   a. Carol will give me a book.
   b. *Carol will give me.

This brief illustration of selectional restrictions re-emphasizes the complexity of lexical development. It also illustrates that syntax and semantics are related. An opportunistic learner should use such relations, which idea drives both bootstrapping hypotheses. Philosophers and linguists think that the meaning of 'sleep' is to some extent reflected by its lexical type and its argument structure. Syntactic bootstrapping exploits this idea by maintaining that the range of argument structures that a verb participates in narrows the hypothesis space that learners consider for its meaning. Read example 4 and guess at the meaning of 'VERB'.

4. Zonk VERBs that the blick is floopy.

Many readers will guess a translation like 'think' (i.e. some mental verb). Perhaps this guess reflects what we know about an argument structure that is typical of such verbs. Interestingly, getting the syntax of the sentence in 4 requires mastery of functional elements like '-s' and 'the'. Let us consider such elements in more detail.

Functional Categories and Grammar

The functional categories are better described by their roles than their referents: that is, a determiner such as 'the' functions to mark and modify noun phrases (e.g. 'the white dog') and an inflection such as '-ing' functions to mark and modify verbs (e.g. 'sleeping'). It is also important to note that the functional categories of any particular language encode its grammar: that is, sentence-level phenomena such as ordering of phrases and agreement between phrases is affected by the language's choice of functional elements. Many researchers maintain that children's development of functional category knowledge lags behind their knowledge of lexical categories. This claim is based, to a large extent, on children's errors or the differences between their behavior and that of adults (who we assume to be finished products of development). Children's development of functional categories illustrates two types of errors: omission of obligatory elements, such as 'is' and '-ing' in 5a, and overt but misused inflectional morphology, such as '-ed' and '-s' in 5b.

5. a. Dog sleep now (to refer to a dog that is sleeping)
   b. breaked, sheeps

The study of lexical development has to some extent ignored functional categories, and this may be because developmental psycholinguists tend to emphasize production data and children's early errors. A wonderful exception to this trend is Gerken *et al*. (1989), whose research is especially instructive with respect to our interpretation of children's errors of omission. In this study, young children responded to commands such as 'Find the bird for me' by pointing to pictures. Phrases like 'the bird' were compared with 'was bird', 'gub bird', and '—bird' (the wrong functional element, an invented functional element and an omitted functional element, respectively). Interestingly, children whose own utterances lacked the target element 'the' were most accurate when responding to commands containing it. The other error type illustrated in 5 also invites interesting conclusions. The child who starts with 'breaked' appears to be applying the regular '-ed' rule to a lexical element requiring a different inflection, and will eventually replace that word with 'broke'. As an aside, it is interesting to speculate about what happens if such replacements do not occur. Could persistent developmental errors drive language change? Might, for example, the shifts in American English from 'learnt' and 'dove' to 'learned' and 'dived' be explained in developmental terms? Returning to the question of what overregularization errors tell us about linguistic development, many have taken children's production of such errors as evidence that they do not know the irregular or exception morphology while they do know the regular morphology. Perhaps regular patterns are learned before irregular patterns.

Thus, study of the development of both lexical and functional elements reveals a busy and sophisticated learner. The learner may produce errors on the way to the target lexicon (e.g. meaning-based overextension of words, omission or overuse of grammatical elements), but when we probe past the obvious, even the errors point to early mastery of extraordinarily complex information and information processing. Returning to the question raised at the beginning of this section, there are many reasons to think that children's knowledge of functional elements develops alongside and in synchrony with their knowledge of lexical elements. For one thing, children's language comprehension precedes their language production. This fact lacks explanation if we attribute to children grammar that is as primitive as their utterances suggest. In addition, crosslinguistic comparisons reveal that children of the same age who are learning different languages produce varying indications of mastery of functional elements (McKee, 1994). English is a relatively isolating language. The omission of a functional element in English often leaves well-formed words. That is not true in all languages (e.g. compare *eats/eat* in English with *mangia/*mang* in Italian). Study of what English learners omit may reveal more about English than it does about lexical development, and children's initial grammatical competence may be richer than their utterances suggest.

OTHER ASPECTS OF COGNITIVE DEVELOPMENT

Finally, children's lexical development raises many philosophical questions. One of these is the extent to which language depends on or even just interacts with extralinguistic cognition. For example, how might the learning of a linguistic label (e.g. 'yak' or plural '-s') affect one's mental representations and processing? In a different direction, consideration of children's developing theory of mind also bears on their learning of linguistic labels. We must also ask where linguistic categories come from, a question that should be aimed both at children and at the evolution of language in our species. If linguistic categories are not selected from an innate list, how do learners discover them? How does the frequency of an element affect its learning? What about the frequency of crosslinguistic associations between the meanings of individual elements and classes of argument structure? How does language in the individual learner differ from the external target language? Can lexical development bring on language change over decades and centuries? Relations between language and other aspects of cognition also bring to mind questions about how language itself is affected by lexical development. Study of this exciting topic thus bears on many domains of cognitive science.

References

Bloom P (1994) Possible names: the role of syntax-semantics mappings in the acquisition of nominals. *Lingua* **92**: 297–329.

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Clark EV (1993) *The Lexicon in Acquisition*. Cambridge, UK: Cambridge University Press.

Gerken LA, Landau B and Remez RE (1989) Function morphemes in young children's speech perception and production. *Developmental Psychology* **27**: 204–216.

Gleitman LR (1990) The structural sources of verb meanings. *Language Acquisition* **1**: 3–55.
Grimshaw J (1981) Form, function, and the language acquisition device. In: Baker CL and McCarthy JJ (eds) *The Logical Problem of Language Acquisition*, pp. 165–182. Cambridge, MA: MIT Press.
Landau B, Smith LB and Jones SS (1988) The importance of shape in early lexical learning. *Cognitive Development* **3**: 299–321.
Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Markman EM (1989) *Categorization and Naming in Children*. Cambridge, MA: MIT Press.
McKee C (1994) What you see isn't always what you get. In: Lust B, Suñer M, and Whitman J (eds) *Syntactic Theory and First Language Acquisition: Cross-Linguistic Perspectives*, vol. 1, Heads, Projections, and Learnability, pp. 201–212. Hillsdale, NJ: Lawrence Erlbaum.
Pinker S (1984) *Language Learnability and Language Development*. Cambridge, MA: MIT Press.
Pinker S (1991) Rules of language. *Science* **253**: 530–535.
Smith NV (1989) *The Twitter Machine*. New York: Basil Blackwell.

Further Reading

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
Clark EV (1993) *The Lexicon in Acquisition*. Cambridge, UK: Cambridge University Press.
Gleitman LR (1990) The structural sources of verb meanings. *Language Acquisition* **1**: 3–55.
Pinker S (1991) Rules of language. *Science* **253**: 530–535.

Lexical Semantics

Introductory article

Chris Barker, University of California, San Diego, California, USA

CONTENTS

Word meaning mediates between conceptualization and language: simply put, words name concepts. Studying which concepts can have names – and how the fine-grained structure of those concepts interacts with the linguistic structures that contain them – reveals something important about the nature of language and cognition.

WHAT CAN WORDS MEAN?

To a first approximation, lexemes are words, so lexical semantics is the study of word meaning. The main reason why word-level semantics is especially interesting from a cognitive point of view is that words are names for individual concepts. Thus lexical semantics is the study of those concepts that have names. The question 'What can words mean?', then, amounts to the question 'What concepts can have names?'

There are many more or less familiar concepts that can be expressed by language but for which there is no corresponding word. There is no single word in English that specifically names the smell of a peach, or the region of soft skin on the underside of the forearm, though presumably there could be. Furthermore, it is common for one language to choose to lexicalize a slightly different set of concepts from another. American speakers do not have a noun that is exactly equivalent to the British *toff*, an 'upper-class person', nor does every language have a verb equivalent to the American *bean*, 'to hit on the head with a baseball'.

The situation is quite different for other concepts, however. To adapt a famous example from philosophy, there is no word in any language that refers specifically to objects that were green before 1 January 2000 but blue afterwards. Of course, we can artificially agree to assign this concept to a made-up word; in fact, let's call it *grue*. But doing so doesn't make *grue* a legitimate word, let alone a genuine one. Could there be such a word? That is, is naming the grue concept truly unnatural, or merely unlikely? (*See* **Meaning**)

Identifying systematic patterns governing the meanings of related words can provide convincing answers to such questions. Furthermore, the insights go in two directions simultaneously: on the one hand, natural classes of words that resemble each other syntactically (for instance, the class of nouns) have similar kinds of meaning. This places bounds on the set of possible word meanings. On the other hand, natural classes of words that resemble each other semantically (for instance, the class of words whose meanings involve the notion of causation) behave similarly from a syntactic point of view. Thus studying how the fine-grained conceptual structure of a word's meaning interacts with the syntactic structures that contain it provides a unique and revealing window into the nature of language, conceptualization, and cognition.

POLYSEMY

Lexemes are linguistic items whose meaning cannot be fully predicted based on the meaning of their parts. Idiomatic expressions by definition, then, are lexemes: the expression *kicked the bucket* (as in the idiomatic interpretation of 'My goldfish kicked the bucket last night') means essentially the same thing as the word *died*. (Although it is convenient to draw examples from English in this article, it gives an unfortunate English-centric perspective to the discussion. Those readers familiar with agglutinative languages such as Turkish or West Greenlandic should consider whether 'morpheme' might be a more accurate term than 'word' in the general case.) Other articles in this encyclopedia discuss cases in which lexemes do not correspond exactly to words, but this article will adopt the simplifying assumption that lexemes are words. (*See* **Construction Grammar**; **Lexicon**)

But how many distinct words are there? We can feel comfortable deciding that homonyms (words that sound the same but have completely different meanings) should be counted as different lexical items: *bat* the flying mammal versus *bat* the wooden stick used in baseball certainly ought to be considered as two different words. It is more difficult to be confident of such judgments, however, when the two meanings seem to be closely related. Is *wave* in reference to a beach a different concept than *wave* in 'He experienced a wave of anger'? What about verbal uses such as 'The infant waved' or 'The policeman waved us through'? Words that have multiple related but distinct meanings are 'polysemous'. Identifying and dealing with polysemy is a hotly debated and difficult problem in lexical semantics.

Some polysemy seems to be at least partly systematic. Pustejovsky proposes that language users rely on *qualia*, which are privileged aspects in the way they conceptualize objects. The four main qualia are the object's constituent parts, its form (including shape), its intended function, and its origin. Thus a fast car is fast with respect to performing its intended function of transporting people, but fast food is fast with respect to its origin. We can recognize the polysemy of *fast* as indeterminacy in selection of the most relevant qualia, without needing to postulate two lexical entries for *fast*, let alone two distinct kinds of fastness.

In addition, qualia provide a way to understand how to coerce a conventional meaning into a variety of extended meanings. In 'The ham sandwich at table 7 is getting restless' (spoken by one waiter addressing another in a restaurant), *the ham sandwich* refers not to the material parts of the sandwich, but to the person involved in specifying the sandwich's intended function (i.e. to be eaten by someone).

LEXICAL RELATIONS (MEANING IN RELATION TO OTHER WORDS)

There are two main modes for exploring word meaning: in relation to other words (this section), and in relation to the world (see below). The traditional method used in dictionaries is to define a word in terms of other words. Ultimately, this strategy is circular, since we must then define the words we use in the definition, and in their definitions, until finally we must either run out of words or reuse one of the words we are trying to define.

One strategy is to try to find a small set of *semantic primes* or basic constituent elements: Wierzbicka identifies on the order of 50 or so concepts (such as *good*, *bad*, *before*, *after*, *I*, *you*, *part*, *kind* …) that allegedly suffice to express the meaning of all words (in any language). Whether this research program succeeds or not has important implications for the nature of linguistic conceptualization.

In any case, speakers clearly have intuitions about meaning relations among words. The most familiar relations are synonymy and antinomy. Two words are synonyms if they mean the same thing, for example *filbert* and *hazelnut*, *board* and *plank*, etc. Two words are antonyms if they mean opposite things: *black* and *white*, *rise* and *fall*, *ascent* and *descent*, etc. One reason to think that it is necessary to recognize antinomy as an indispensable component of grammatical descriptions is because at most one member of each antonymous pair is allowed to occur with measure phrases: *tall* is the opposite of *short*, and we can say 'Bill is 6 feet tall', but not '*Tom is 5 feet short'. (Asterisks denote a deviant construction.)

Other major semantic lexical relations include hyponymy (names for subclasses: *terrier* is a hyponym of *dog*, since a terrier is a type of dog), and meronymy (names for parts: *finger* is a meronym of *hand*).

Words whose meanings are sufficiently similar in some respect are often said to constitute a 'semantic field', though this term is rarely if ever given a precise definition. Terms such as *red, blue, green*, etc., are members of a semantic field having to do with color.

Miller and associates have developed a lexical database called WordNet, a kind of multidimensional thesaurus, in which these types of lexical relations are explicitly encoded. Thus WordNet is an attempt to model the way in which a speaker conceptualizes one kind of word meaning.

DENOTATION (MEANING IN RELATION TO THE WORLD)

Defining words in terms of other words or concepts will take us only so far. Language talks about the world, and ultimately, in order to know what a word such as *horse* or *despair* means, it is necessary to know something about the world. One way of getting at the connection between meaning and the world is to consider 'truth conditions': what must the world be like in order for a given sentence to be true? In general, words participate in most of the kinds of truth-conditional meaning that sentences do, plus one or two more of their own.

'Entailment' is the most concrete aspect of meaning: one sentence entails another just when the

truth of the first sentence is sufficient to guarantee the truth of the second. For instance, the sentence 'I have four children' entails the sentence 'I have at least two children'. We can extend this notion to words by assuming that one word entails another in cases where substituting one word for the other in a sentence produces an entailment relation. Thus *assassinate* entails *kill*, since 'Jones assassinated the President' entails 'Jones killed the President'. Similarly, *whisper* entails *spoke*, *devour* entails *ate*, and so on.

'Presuppositions' are what must already be assumed to be true in order for a use of a sentence to be appropriate in the first place. For instance, *and* and *but* mean almost the same thing, and differ only in the presence of a presupposition associated with *but*. If I assert that 'Tom is rich and kind', it means exactly the same thing as 'Tom is rich but kind' as far as what Tom is like; the only difference is that the use of the sentence containing *but* presupposes that it is unlikely for someone who is rich also to be kind.

'Selectional restrictions' are a kind of presupposition associated exclusively with words. The verb *sleep* presupposes that its subject is capable of sleeping; that is, it selects a restricted range of possible subjects. One reason why the sentence '*Green ideas sleep furiously' is deviant, then, is because ideas are not capable of sleeping, in violation of the selectional restriction of *sleep*.

'Implicature' is weaker than entailment. If you ask me how a student is doing in a course, and I reply 'She's doing fine', I imply that she is not doing great (otherwise, I would presumably have said so). Yet the sentence 'She's doing fine' does not entail the sentence 'She's not doing great', since it can be true that a student is doing fine work even if she is actually doing spectacular work. The implicature arises in this case through contrast with other words that could have been used. Interestingly, such lexical implicatures favor the formation of lexical gaps: because *fine* implicates *not good*, there is no word that means exactly *not good* (for instance, *mediocre* entails both *not good* and *not bad*). Similarly, because *some* implicates *not all*, there is no single word (perhaps not in any language) that means exactly *not all*.

'Connotation' is the part of the meaning of a word that adds a rhetorical spin to what is said. More specifically, connotation signals the attitude of the speaker towards the object or event described. If I describe someone as *garrulous* rather than *talkative*, the described behavior may be exactly the same, but *garrulous* conveys in addition the information that I disapprove of or dislike the behavior in question.

'Vagueness' is perhaps the quintessential problem for scientific theories of word meaning. Imagine a person with a full head of hair; clearly, this person is not bald. Now imagine pulling out one hair at a time. Eventually, if this process is continued long enough, the person will qualify as bald. But at what point precisely does the switch from not-bald to bald occur? It's impossible to say for sure. But if we can't answer this simple question, then how can we possibly claim to truly understand the meaning of the word *bald*? A little reflection will reveal that virtually every word that refers to objects or events in the real world contains vagueness: exactly how tuneless must a vocal performance get before the claim 'She is singing' is no longer true? Is that short sofa really a chair? The pervasiveness and intractability of vagueness gives rise to profound philosophical issues. (*See* **Vagueness**)

KINDS OF WORDS, KINDS OF MEANING

One of the most important principles in lexical semantics (part of the heritage of Richard Montague) is that words in the same syntactic class (e.g. noun, verb, adjective, adverb, determiner, preposition, etc.) have similar meanings. That is, the meaning of a determiner like *every* differs from the meaning of a verb like *buy* in a way that is qualitatively different from the difference in meaning between two verbs like *buy* and *sell*.

The overarching division in word classes distinguishes function words from content words. Function words (determiners, prepositions, conjunctions, etc.) are shorter, higher-frequency words that serve as syntactic glue to combine words into sentences. Content words (nouns, verbs, adjectives, adverbs) carry most of the descriptive payload of a sentence. In the sentence 'He asked me to give her a haircut', the content words are *asked*, *give*, and *haircut*.

Semantically, function words are much closer to being able to be expressed as a mathematical or logical function. We can write a rule that gives a good approximation of the contribution of *and* to truth conditions in the following way: 'A sentence of the form [S1 and S2] will be true just in the case where S1 is true and S2 is also true.' Thus the sentence 'It was raining and it was dark' will be true just in the case where the sentence 'It was raining' is true and the sentence 'It was dark' is also true.

Among function words, the meaning of determiners such as *every*, *no*, *most*, *few*, *some*, etc. (but crucially excluding noun phrase modifiers such as *only*) have been studied in exquisite detail, especially from the point of view of their logical properties. One of the most celebrated results is the principle of 'conservativity': given any determiner D, noun N, and verb phrase VP, sentences of the form [D N VP] are guaranteed to be semantically equivalent to sentences of the form [D N be N *and* VP]. For instance, if the sentence 'Most cars are red' is true, then the sentence 'Most cars are cars and are red' is also true, and vice versa; similarly, 'Some cars are red' is equivalent to 'Some cars are cars and are red', and so on. It is conjectured that every determiner in every language obeys conservativity. These equivalences may seem so obvious that it is hard to appreciate just how significant this generalization is; it may help, therefore, to imagine a new determiner *antievery* as having a meaning such that a sentence 'Antievery car is red' means the same thing as 'Everything that is not a car is red'. The claim is that there could be no such determiner in any natural language; that is, *antievery* (or any determiner that violated conservativity) is an impossible thing for a word to mean.

There is a class of lexemes whose behavior is especially fascinating from a cognitive point of view. In English this class includes expressions such *ever*, *anymore, budge an inch*, and many others. These items can only be used in sentences that exhibit certain properties related to the presence of negation. For instance, it is possible to say 'She doesn't ever go to the movies anymore', but it is impossible to say '*She ever goes to the movies anymore'. It is the presence of the negation in the first sentence that licenses the words in question, and it is because of this affinity for negation that they are called 'negative polarity items'. Not only are NPIs interesting in their own right, they provide clues about the internal structure of the meaning of other words. Consider the following contrast: 'I doubt that she ever goes to the movies anymore' is fine, but '*I think that she ever goes to the movies anymore' is deviant. Because negative polarity items are licensed by *doubt* but not by *think*, we can conclude that the meaning of *doubt* contains negation hidden within it, so that *doubt* means roughly *think that not* (i.e. *think that it is not the case that*).

EVENTS

The main linguistically important divisions among event types are as shown in Figure 1. Stage-level

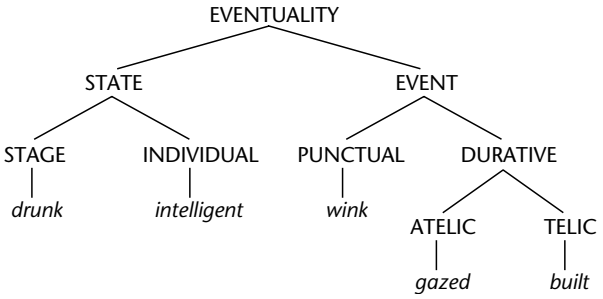


Figure 1.

adjectives such as *drunk* correspond to temporary stages of an individual rather than properties that are (conceived of as being) permanent properties of individuals, such as *intelligent*. Telic activities differ from atelic ones in having a natural end point. Crucially, these semantic distinctions correspond to systematic differences in syntactic behavior. For instance, assume that the temporal use of the preposition *for* requires the absence of an end point, but *in* requires a definite end point. If *gaze* is naturally atelic but *built* is telic, this explains why it is possible to say 'He gazed out to sea for an hour' but not '*He gazed out to sea in an hour'; and conversely, why it is impossible to say '*He built the raft for an hour' at the same time that 'He built the raft in an hour' is perfectly fine.

Although many lexical predicates are naturally atelic or telic, the telicity of the sentence in which they occur can be determined by other elements in the sentence. For instance, some verbs (*drink*, *write*, *draw, erase, pour*, etc.) have meanings that establish a systematic correspondence between the physical subparts of one of their arguments and the described event. Consider the verb *mow*: the subparts of an event of mowing the lawn correspond to the subparts of the lawn. If the event of mowing the lawn is half over, then roughly half the lawn will have been mown. This contrasts with a verb like *interview*: if an interview with the President is half over, that does not mean that half of the President has been interviewed. Because of the systematic correspondence between the internal structure of the noun phrase (NP) denotation and the internal structure of the described event, semantic properties of this special type of verbal argument-NP (which is called an 'incremental theme') can project to determine semantic properties of the described event. In particular, the telicity of sentences containing incremental theme verbs depends on whether the incremental-theme argument is conceived of as quantized (*a beer* is quantized,

compared with just *beer*): 'John drank a beer in/*for an hour' versus 'John drank beer for/*in an hour'.

THEMATIC ROLES

Decomposition

Sentences have internal structure in which the basic unit is the word. For instance, the sentence 'The girl hit the boy' has for its coarse-grained structure [[the girl] [hit [the boy]]]. Words have internal structure as well, in which the basic unit is the morpheme. For instance, *undeniable* = [un [deni [able]]]. Just as the meaning of a sentence must be expressed as a combination of the meaning of its parts and their syntactic configuration, we expect the meaning of a word (in the normal case) to be a combination of the meanings of its parts, and indeed *undeniable* means roughly *not able to be denied*. The internal structure of words can be quite complex, as in agglutinative languages like Turkish, in which there is no principled limit to the number of morphemes in a word. (*See* **Morphology**; **Syntax**)

Yet in general, word structure is qualitatively less complex than syntactic structure in a variety of measurable ways. If the structure of word meaning mirrors the way in which words are built up from morphemes, then we might also expect that word meaning will be simpler than sentence meaning in some respects. However, as discussed in the next section, many scholars believe that the meaning of words has internal structure that does not correspond to any detectable morphological structure.

Argument Structure and Conceptual Structure

One of the main ways in which words in the same main syntactic class differ in their syntax and in their meaning is in terms of 'argument structure'. *Fall*, *hit*, and *give* are all verbs, but differ in the number and arrangement of noun-phrase arguments they occur with: intransitive verbs have one argument ('John fell'); transitive verbs have two arguments ('John hit the ball'), and ditransitive verbs have three arguments ('John gave the book to Mary'). Thus it is necessary to annotate each word with its argument structure, and the meaning of the verb will refer to elements in the argument structure.

'Conceptual structure' defines the connection between argument structure and meaning. According to Jackendoff, for example, the transitive verb *drink* corresponds to the following (simplified) lexical entry:

Word:	*drink*, verb (transitive)
Argument structure:	$\langle NP_i, NP_j \rangle$
Conceptual structure:	CAUSE (NP_i, GO (NP_j, TO (IN (MOUTH-OF (NP_i)))))

In the sentence 'John drank the soda', then, NP_i = *John*, NP_j = *the soda*, and the sentence as a whole means 'John caused the soda to go into John's mouth'.

Note that the conceptual structure is considerably more complex than the syntactic structure in which the verb participates: the concept named by *drink* has been decomposed into five separate predicates (CAUSE, GO, TO, IN, and MOUTH-OF), and there are three arguments at the conceptual level (John, the soda, and John's mouth) instead of two, as at the overt syntactic level.

Levin and others show how separating conceptual structure from argument structure provides a framework for explaining a type of polysemy that gives rise to various syntactic 'alternations': cases in which a single word can be used with a variety of argument-structure configurations. The basic idea is that when a verb like *break* occurs in a transitive sentence such as 'The girl broke the window', it has a conceptual structure that explicitly involves causation – perhaps (CAUSE (THE-GIRL, BECOME (BROKEN, THE-WINDOW))). When the same verb occurs with a single argument, as in 'The window broke', it takes for its conceptual meaning a proper subpart of the transitive structure – (BECOME (BROKEN, THE-WINDOW)). Crucially, the substructure omits mention of the causing entity.

A particularly interesting kind of alternation is known as an 'implicit argument'. For instance, when comparing the meaning of 'John ate something' versus 'John ate', the omitted argument in the shorter version is intuitively still present conceptually. In this case the two senses of *eat* differ only in argument structure but not in conceptual structure.

Thematic Roles and Linking

It is not an accident given the meaning of the word *drink* that the entity that causes the drinking event is expressed as the subject of the sentence: in 'John drank the soda', we know that John is consuming the soda, and not vice versa. More generally, to some degree at least, the meaning of verbs clearly

interacts with the syntactic form of sentences containing those verbs. The various theories developed to characterize the connection between verb meaning and syntactic structure are called 'linking theories', and in many linking theories the connection depends on 'thematic roles'. A verb like *kiss*, for instance, describes events that have two main conceptual participants, which we can call the kisser (the entity that does the kissing) and the kissee (the entity that gets kissed). When *kiss* occurs as the main verb in a sentence of English, the kisser is always syntactically realized as the subject and the kissee is realized as the direct object:

Syntactic structure:	[The girl]	kissed	[the boy]
Syntactic function:	Subject		Direct object
Verb-specific roles:	kisser		kissee
Thematic roles:	Agent		Patient

Thus the linking rules for English must guarantee that the verb-specific kisser role always comes to be associated with the grammatical subject of the sentence. (Passive sentences such as 'The boy was kissed by the girl' are a systematic variation on the basic pattern in which the normal basic linking is deliberately reversed for discourse purposes.)

Building on ideas of Davidson that emphasize the importance of events in the semantics of verbs, Parsons accounts for linking regularities by providing thematic roles as primitive notions, and stipulating that when a verb assigns a noun phrase the role of Agent, that NP must appear in subject position. The sentence 'Brutus stabbed Caesar with a knife' would be rendered as [STABBING(e) & AGENT (BRUTUS, e) & PATIENT (CAESAR, e) & INSTRUMENT (KNIFE, e)], which asserts that e is a stabbing event, Brutus is the agent of the stabbing event, Caesar is the patient, and the knife is the instrument.

Other linking theories attempt to derive thematic roles from independently motivated aspects of meaning. In a theory that recognizes conceptual structure, we can hypothesize that for any verb that contains as part of its meaning the conceptual primitive CAUSE, the first conceptual argument of the CAUSE predicate will be an agent. Put another way, the hypothesis is that the syntactic subject is the participant that the speaker conceptualizes as taking the more active causal role in the event.

Both the neo-Davidsonian approach and the conceptual-structure approach rely on defining the meaning of a word in terms of other more basic concepts. Dowty approaches the issue from the point of view of the relation between word meaning and the world. In particular, he suggests that it suffices to consider the entailments of the verb in question. The verb *kiss* entails that the kisser participant must have lips, that those lips must come in contact with the kissee participant, and the kisser must intentionally cause the contact to occur. In contrast, it is not an entailment that the kissee has lips, since it is perfectly possible to kiss a clam, or a rock, nor is there an entailment that the kissee intentionally causes the kissing event to occur (imagine kissing a sleeping child). On entailment-based theories of thematic roles, if a verb entails that a participant intends for the event described by the verb to come about, or causes it to come about, then that participant will (be more likely to) be expressed as the subject rather than as the direct object.

All theories that recognize the existence of thematic roles provide at least two basic thematic roles: Agent and Patient (the term 'Patient' is often collapsed with the term 'Theme', whence the cover term 'thematic roles'). Agents tend to be subjects; Patients, which are conceived of as undergoing an action (alternatively, are entailed to undergo an action, or which are incremental themes) make good direct objects. Among the other thematic roles that play a prominent part in many linking theories are Experiencer, Source, Goal, and Instrument.

Linking theories make strong predictions about possible and impossible words. Assume we are presented with a verb *glarf* that is allegedly a previously unknown transitive verb of English. We are told the following concerning its meaning: it describes situations in which one participant comes into violent contact with the foot of another participant. Furthermore, we know that it is the participant that possesses the foot who initiates or causes the touching event to occur. Now we encounter the following sentence: 'The ball glarfed the player', which is used as a description of an event in which a football player kicked a football for a goal. (In other words, *glarf* seems to be the verb *kick* but with subject and direct object reversed.) At this point, we can conclude that *glarf* is not a legitimate word of English (and in fact, is not a legitimate basic verb in any human language that recognizes the grammatical relations of subject and object), because it associates the volitional participant with the direct object and the more passive undergoer participant with the subject, contrary to the predictions of linking theories.

In fact, thematic roles can provide insight even into the syntactic behavior of verbs that have only one argument. The 'unaccusative hypothesis', due largely to Perlmutter, was a breakthrough in the use of lexical semantics to explain syntactic patterns. The hypothesis claims that intransitive verbs whose sole argument is semantically an Agent (predicates like *sing*, *smile*, or *walk*) behave differently cross-linguistically from so-called 'unaccusative' verbs, whose one argument is semantically a Patient (e.g. *fall*, *bleed*, or *die* – crucially, these activities are conceived of as not being done intentionally). For instance, unaccusative verbs in Italian generally take *essere* ('to be') as a past auxiliary rather than *avere* ('to have').

Like the correspondence between thematic roles and argument linking, the unaccusativity pattern is systematic enough to be quite compelling. However, as with virtually every attempt to derive syntactic behavior directly from semantic distinctions, examination of a wider range of data has revealed a considerable number of complications and exceptions. Nevertheless, there can be no doubt that meaning and form are deeply and inextricably connected, and that lexical semantics is the place at which they come into contact with one another. Therefore understanding lexical semantics is critical to understanding the yin and yang of language and conceptualization.

Further Reading

Dowty D (1991) Thematic proto-roles and argument selection. *Language* **67**(3): 547–619.

Fellbaum C (ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Grimshaw J (1990) *Argument Structure*. Cambridge, MA: MIT Press.

Jackendoff R (1990) *Semantic Structures*. Cambridge, MA: MIT Press.

Kamp H and Partee B (1995) Prototype theory and compositionality. *Cognition* **57**: 129–191.

Levin B (1993) *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Levin B and Pinker S (eds) (1992) *Lexical and Conceptual Semantics*. Oxford, UK: Blackwell.

Parsons T (1990) *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, MA: MIT Press.

Pustejovsky J (1995) *The Generative Lexicon*. Cambridge, MA: MIT Press.

Wierzbicka A (1996) *Semantics, Primes and Universals*. Oxford, UK: Oxford University Press.

Lexical-Functional Grammar

Introductory article

Mary Dalrymple, Xerox PARC, Palo Alto, California, USA

CONTENTS

Lexical-functional grammar is a linguistic theory that explores the nature of the various aspects of the structure of language and the relations between them.

INTRODUCTION

Lexical-functional grammar (LFG) is a theory of the structure of language. The theory views the overall structure of language in terms of substructures representing different aspects of linguistic organization, each with its own organizing rules and principles. For example, the ordering and grouping of words into phrases, and of phrases into larger phrases, is represented by a 'constituent structure tree'. Traditional abstract syntactic roles like 'subject' and 'object' are represented by a 'functional structure', different from but related to the constituent structure. Other subsystems of linguistic organization have also been explored, including semantic structure, argument structure, and information structure. LFG provides an explicit theory of the structure and function of each subsystem as well as a detailed specification of the relations among the different systems, enabling a clear, well-grounded and well-motivated theory of the organization of human language. (*See* **Semantics and Pragmatics: Formal Approaches**; **Categorial Grammar and Formal Semantics**)

CONSTITUENT STRUCTURE AND FUNCTIONAL STRUCTURE

LFG had its beginnings in the 1970s, when linguists became dissatisfied with the then-current transformational theory of syntactic structure. According to transformational theory, complex sentences result from the application of a series of transformations that change or combine simple sentences into more complex sentences. In contrast, early LFG research showed that a psychologically more realistic, conceptually simpler, computationally more tractable, and theoretically more satisfactory theory results from abandoning transformations and adopting a lexical view: different sentences with similar meaning are related because the words in the sentences are related. LFG also proposed a finer-grained view of syntactic structure than was assumed in transformational theories. In the LFG view, the syntactic organization of every well-formed sentence is represented in terms of two separate but closely related syntactic structures. The constituent structure tree is like the tree structures commonly used in transformational theories: it represents the organization of words into phrases, and of phrases into sentences. The functional structure represents more abstract functional roles like 'subject' and 'object'.

Well-formed sentences are organized hierarchically: the words in the sentence form phrases, which in turn form larger phrases. For the sentence *David is eating an apple*, the result is the constituent structure tree shown in Figure 1. This tree represents a great deal of information about the phrasal structure of the sentence. The two most important aspects of the tree are the categories of the nodes in the tree and the phrasal groupings represented by the hierarchical structure: phrasal categories are represented as node labels like VP, NP, N, I′, and so on, and phrasal groups are sequences dominated by a single tree node.

In this tree, the word *eating* is dominated by a V node, indicating that it is a verb. Categories like V (for verb), N (for noun), and I (for inflection) are heads of phrases, and are dominated by their corresponding 'single-bar-level' category: V′ (pronounced 'V-bar'), N′ ('N-bar'), and I′ ('I-bar'). The single-bar categories are distinguished in that they can also dominate complements of the head: in this example, the V′ node dominates the V head as well as the NP complement *an apple*; similarly, the I′ category dominates the inflectional auxiliary *is*

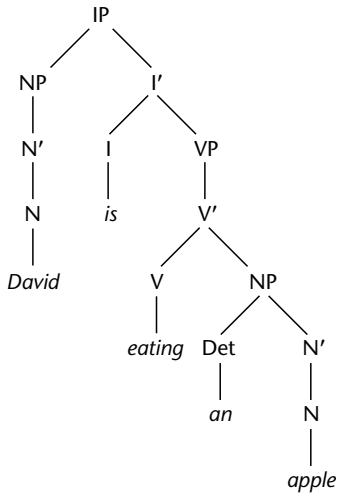


Figure 1. The constituent structure tree for the sentence *David is eating an apple.*

and its VP complement *eating an apple*. Categories like VP ('verb phrase') and NP ('noun phrase') are full phrases, containing both modifiers and complements of the head, and can appear in various places in the sentence according to their syntactic role; their appearance in Figure 1 is governed by basic phrase-structure principles as well as by the particular constraints relevant for English phrase structure. (*See* **Phrase Structure and X-bar Theory**; **Phrase Structure Grammar, Head-driven**; **Parsing: Overview**)

The tree in Figure 1 also indicates that several sequences of words are phrasal units, or 'constituents', dominated by a single node in the tree: *an apple* is an NP constituent, *eating an apple* is a VP constituent, and *is eating an apple* is an I′ constituent. These units also play an important role in other English sentences, since the phrasal rules of English tell us that an NP constituent can appear in any one of a number of phrase-structure positions: the NP *an apple* can appear at the beginning of the sentence, just after the verb, or as the object of a preposition:

[An apple] makes a good snack.
David gave [an apple] to Ken.
He nibbled at [an apple]. (1)

As we would expect, other NPs can also appear in these positions. The phrase *a delicious piece of fruit* is also an NP:

[A delicious piece of fruit] makes a good snack.
David gave [a delicious piece of fruit] to Ken.
He nibbled at [a delicious piece of fruit]. (2)

Rules of constituent structure in English govern where such phrasal units can appear. (*See* **Syntax**)

Some other languages also have consitutent structures like those described above. For example, Chinese, a language unrelated to English, has a constituent structure that is similar in some ways to that of English. However, many languages are unlike English in constituent structure. The constituent structure of Warlpiri, an Australian language, is very different from that of English: there is no VP category, for example, and a noun and its modifier need not form a single phrase. The sentence in Figure 2 means 'the two small children are chasing the dog' in Warlpiri. The noun *kurdu-jarra-rlu* ('two children') and its modifier *wita-jarra-rlu* ('small') do not form a phrasal unit in the constituent structure: *kurdu-jarra-rlu* appears as the first word of the sentence, and *wita-jarra-rlu* is the last word. This is not possible in English (an asterisk in front of a sentence means that it is not an acceptable sentence of standard English):

*[Fruit] David is eating [a piece of delicious]. (3)

In fact, in Warlpiri the words of the sentence can appear in almost any order, with no change to the basic meaning of the sentence. In the two sentences below, the words appear in different orders, reflecting differences in information structure – what is emphasized, assumed to be common knowledge, or presented as new information – but with no change to the basic meaning ('the two small children are chasing the dog'):

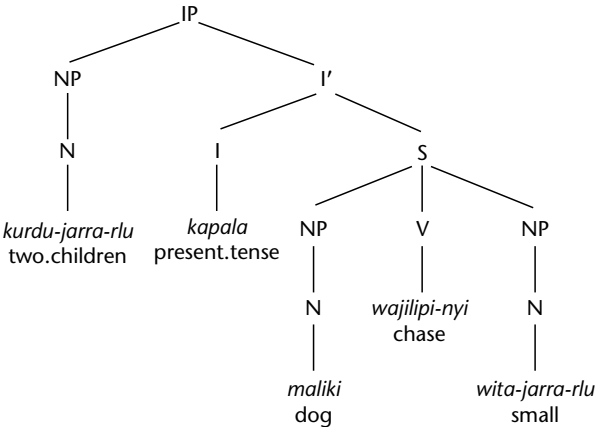


Figure 2. The constituent structure tree for the Warlpiri sentence *kurdu-jarra-rlu kapala maliki wajilipi-nyi wita-jarra-rlu*, which means 'the two small children are chasing the dog'.

maliki	kapala	kurdu-jarra-rlu	wita-jarra-rlu	wajilipinyi
dog	present tense	two.children	small	chase

wajilipinyi	kapala	kurdu-jarra-rlu	maliki	wita-jarra-rlu
chase	present tense	two.children	dog	small

In contrast, changing the order of words in a English sentence does change the meaning, and some orders are impossible:

A delicious piece of fruit ate David.
*A fruit delicious David piece of ate. (5)

Languages can vary a good deal in constituent structure: some, like English, have a more or less rigid word order, so that changing the order of the words changes the meaning of the sentence, while other languages are more like Warlpiri in allowing different word orders to correspond to the same meaning.

However, languages vary much less in functional structure, which represents more abstract syntactic information. Every well-formed sentence in a language has a functional structure that indicates the grammatical roles of the parts of the sentence: what is the subject (SUBJ), what is the object (OBJ), and so on. In LFG, grammatical relations like SUBJ and OBJ are not assumed to be defined in terms of particular constituent-structure-tree configurations, or in terms of particular participant roles in an event such as 'agent' or 'undergoer'; instead, these relations are primitives of the theory, manifested in terms of a cluster of properties that tend to go with particular grammatical roles.

English and Warlpiri are much more alike in functional structure than their very different constituent structures might lead us to expect. The English sentence *David eats apples* has a subject (*David*), a main predicate (*eats*), and an object (*apples*). These traditional grammatical roles are represented in LFG by the functional structure. The functional structure for the sentence *David eats apples* is given in Figure 3. The diagram shows that the sentence has a main PRED, 'eat', which requires a SUBJ and an OBJ. The SUBJ functional structure has a PRED, 'David'; it also contains the information that 'David' has singular number, represented by the number feature NUM with value 'singular'. The OBJ functional structure has a PRED, 'apple', and a NUM feature whose value is 'plural', since 'apples' is a plural phrase referring to more than one apple. A list of functional-structure features is given in Table 1. For simplicity, the functional structures we display here are abbreviated:

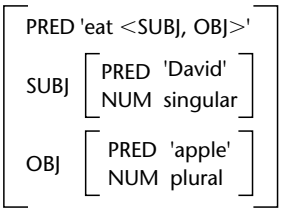


Figure 3. The functional structure for the sentence *David eats apples*.

Table 1. Functional-structure features

Concept	*Feature*	*Values*
Person	PERS	first, second, third
Gender	GEND	masculine, feminine, neuter, …
Number	NUM	singular, plural, …
Case	CASE	nominative, accusative, …
Surface form	FORM	[surface word form]
Verb form	VFORM	pastpart, prespart, …
Tense	TENSE	present, past, …

for example, we have displayed only the number of the subject 'David', omitting features like person and gender.

Several pieces of evidence show that *David* is the subject and *apples* is the object in the sentence *David eats apples*. In English the subject comes before the verb, and the object comes after the verb. If we switch *David* and *apples*, we get a sentence with a different meaning, where *apples* is the subject and *David* is the object:

Apples eat David. (6)

Verb agreement in English provides another way of identifying the subject of a sentence. An English third-person verb ends with 's' if its subject is singular, but not if it is plural:

David eats, apples.
*David eat, apples.
*Apples is eaten by David.
Apples are eaten by David. (7)

English tag questions such as *doesn't he* or *isn't he* appear after the main sentence and show agreement with the subject. If the subject is singular (and masculine), we use a tag question like *doesn't he* or *isn't he*, whereas if it is plural we use *don't they* or *aren't they*:

David eats apples, doesn't he?
*David eats apples, don't they?
*Apples are eaten by David, isn't he?
Apples are eaten by David, aren't they? (8)

This and other evidence shows that *David* is the subject of the sentence *David eats apples* and *apples* is the object. Although the sentence *Apples are eaten by David* has the same meaning, its syntactic structure is different: *apples* is the subject of the sentence, and *David* is the object of the preposition *by*. (*See* **Agreement**; **Local Dependencies and Word-order Variation**; **Linguistic Evidence, Status of**)

Some of these tests for subjecthood are also relevant in other languages. Verb agreement is an indicator of subjecthood in many (though not all) languages; likewise, in many (though not all) languages it is possible to determine the grammatical role of a phrase from its position in the sentence. Although evidence for grammatical functions can differ from language to language, LFG research has shown that functional structure is relevant in every language, and that functional roles and features can be identified in every well-formed sentence in every language. Conversely, if a sentence is ill-formed or ungrammatical, the explanation for its ill-formedness often lies in a violation of functional requirements, as in the examples above.

Now let us consider a sentence meaning 'David eats an apple (or 'apples')' in Japanese, a language that indicates subjecthood and objecthood in a different way from English:

David ga ringo o taberu
David nominative apple accusative eats
'David eats apples.' (9)

The subject of this sentence is *David* and the object is *ringo*; therefore, the functional structure of this sentence is almost the same as the functional structure of the English sentence *David eats apples*. The only difference is that Japanese does not indicate the number of noun phrases like *ringo*, which can mean either 'apple' or 'apples', so that the number feature NUM and its value are not present in the OBJ functional structure.

As in Warlpiri, word order does not help us identify the subject and object in a Japanese sentence; the subject and object can be reordered without affecting the functional structure or the meaning of the sentence. The sentence below has the same functional structure and meaning as the sentence above, even though the object comes first, followed by the subject:

ringo o David ga taberu
apple accusative David nominative eats
'David eats apples.' (10)

In Japanese, unlike English, grammatical roles are identified by the case marker they appear with: the subject is usually marked with *ga*, the nominative case marker, and the object is usually marked with the accusative marker *o*. This case-marking appears regardless of whether the subject is before or after the object, and helps to determine the grammatical role of each part of the sentence and the appropriate functional structure for the sentence.

English and Japanese exemplify two different ways that languages can encode information about grammatical roles: by position and by case-marking. Although sentences in the two languages have very different constituent structure, they exhibit an underlying unity in functional structure.

SYNTACTIC RELATIONS

We have seen that an English sentence like *David eats apples* and a Japanese sentence like *David ga ringo o taberu* have a constituent structure representing phrasal organization and a functional structure representing abstract functional roles and features. We now turn to an examination of the relation between the constituent structure and the functional structure.

In English, grammatical roles are correlated with phrase-structure positions: for example, the phrase appearing as the first daughter of the IP phrase is the subject. There is a mathematical function that defines the relation between each node of the constituent structure tree and the functional structure it corresponds to. We call this function ϕ ('phi'), and we represent it by means of a line pointing from a constituent-structure node to the functional structure it corresponds to, as in Figure 4. In other languages, the relation in different. We have seen that in Japanese and Warlpiri, word order does not correlate with grammatical role: the words in a sentence can appear in different orders without affecting the functional structure. In these languages, the subject can appear in various places in the constituent structure tree. The constituent structure and functional structure for the Warlpiri sentence meaning 'the man is spearing the kangaroo' are shown below. The subject *ngarrka-ngku* ('man') is the third word in the sentence, appearing as the first daughter of the S node in the constituent structure tree (see Figure 5), while the object appears in initial position in the sentence:

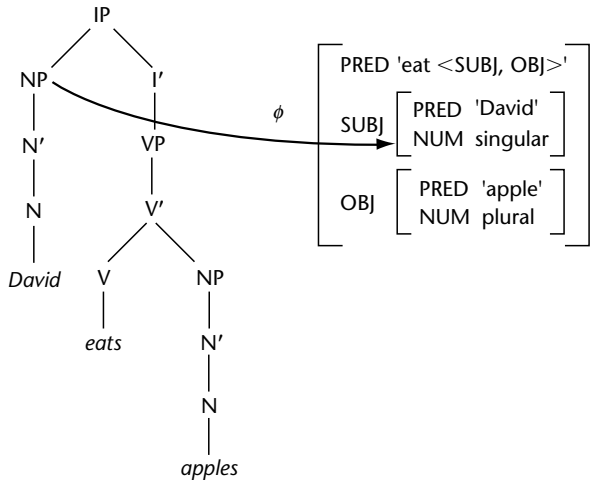


Figure 4. The relation ϕ between the nodes of the constituent structure tree and the corresponding functional structures, for the sentence *David eats apples*.

wawirri ka ngarrka-ngku panti-rni
kangaroo present.tense man spear
'The man is spearing the kangaroo.' (11)

LFG research has explored the relation between constituent structure and functional structure in a number of languages, establishing universally valid constraints on the function ϕ from constituent structure nodes to functional structures, as well as exploring language-particular constraints on the ϕ function. These constraints fully determine the ϕ function. A full specification of the relation between the constituent structure and the functional structure for the sentence *David eats apples* is given in Figure 6.

CONSTRAINTS ON SYNTACTIC STRUCTURES

So far, we have discussed the syntactic structures that LFG assumes – the constituent structure and the functional structure – and we have seen how they are related by the ϕ function from nodes of the constituent structure tree to functional structures. We will now show how constraints on these structures are specified in a linguistic description of languages like English, Japanese, or Warlpiri.

The phrase-structure rule below states that a node labeled V′ can dominate two nodes labeled V and NP:

$$V' \rightarrow V \quad NP \tag{12}$$

LFG allows more complex characterizations of well-formed constituent structure configurations by the use of 'regular expressions' on the right-hand side of a phrase-structure rule. The following more complete V′ rule uses the Kleene star operator ('*') to state that a node labeled V′ can dominate a V node, an optional NP node, and any number (including zero) of prepositional phrase nodes labeled PP:

$$V' \rightarrow V \ (NP) \ PP^* \tag{13}$$

We must also specify the relations between the functional structures corresponding to these nodes. We can do this by using the 'up arrow' $\uparrow$ and the 'down arrow' $\downarrow$ in annotations on the daughter nodes in a phrase-structure rule to refer to the functional structures of the mother node and the daughter node. In the annotated rule below, the up arrow $\uparrow$ refers to the functional structure of the mother node V′, and the down arrow $\downarrow$

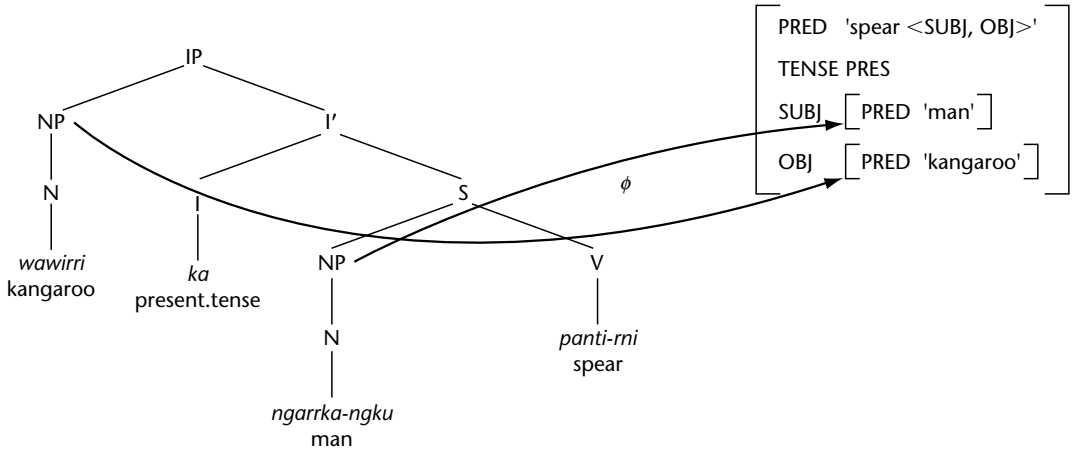


Figure 5. The relation ϕ between the nodes of the constituent structure tree and the corresponding functional structures, for the Warlpiri sentence *wawirri ka ngarrka-ngku panti-rni*, meaning 'the man is spearing the kangaroo'.

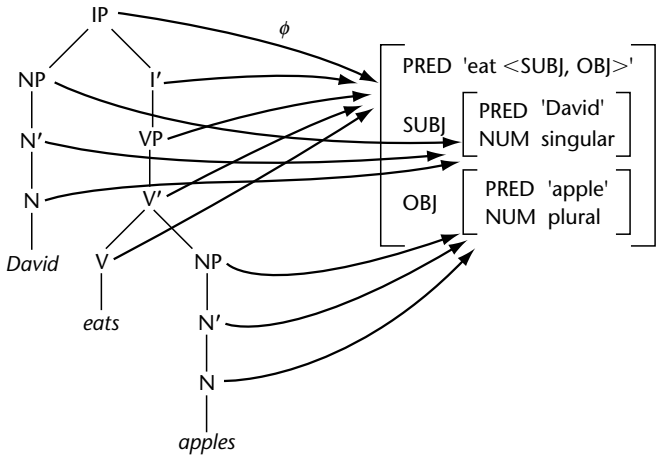


Figure 6. Complete specification of the relation ϕ between the constituent structure tree and the functional structure for the sentence *David eats apples*.

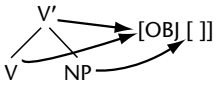


Figure 7. A simple mapping from constituent structure to functional structure, which obeys the constraint $\begin{matrix} V' \rightarrow & V & & NP \\ & \uparrow = \downarrow & & (\uparrow \text{OBJ}) = \downarrow \end{matrix}$.

refers to the functional structure of the V. The equation $\uparrow = \downarrow$ means that the functional structure for the V′ must be the same as the functional structure for the V:

$$\begin{matrix} V' \rightarrow & V \\ & \uparrow = \downarrow \end{matrix} \qquad (14)$$

The more complete annotated rule below tells us that a V′ node dominates a V node and an NP node, that the functional structure for the V is the same as the functional structure for the V′, and that the functional structure for the NP bears the OBJ function in the functional structure for the V′:

$$\begin{matrix} V' \rightarrow & V & & NP \\ & \uparrow = \downarrow & & (\uparrow \text{OBJ}) = \downarrow \end{matrix} \qquad (15)$$

The pairing between constituent and functional structures (constituent structure–functional structure pair) in Figure 7 obeys this constraint. The annotation $\uparrow = \downarrow$ on the V node is satisfied, since the functional structure of the V′ node $\uparrow$ is the same as that of the V node $\downarrow$; the annotation $(\uparrow \text{OBJ}) = \downarrow$ on the NP node is satisfied, since the functional structure of the V′ node $\uparrow$ has an OBJ attribute whose value is the functional structure of the NP node $\downarrow$. Similar rules for IP, I′, NP, and so on can be written, to describe and constrain a complete constituent structure–functional structure pair like Figure 6.

Words also contribute information about their constituent-structure and functional-structure properties. The lexical entry below contains information about the word *apples*:

$$\begin{matrix} \textit{apples}\ \text{N} & (\uparrow \text{PRED}) = \text{'apple'} \\ & (\uparrow \text{NUM}) = \text{plural} \end{matrix} \qquad (16)$$

According to this lexical entry, the word 'apples' is of phrase-structure category N; it contributes the feature PRED with value 'apple', and the number feature NUM with value 'plural'. This constraint is satisfied by the constituent structure–functional structure pair in Figure 6. All constituent-structure configurations, all functional-structure features and values, and the ϕ function from constituent structure nodes to functional structures, must be justified by rules like rule 15 or by information in lexical entries like entry 16. (*See* **Lexicon, Computational Models of**)

Besides basic equations like '$\uparrow = \downarrow$' or '$(\uparrow \text{NUM}) = \text{plural}$', LFG allows a range of other constraints on functional structures. Existential constraints are used to require the presence of a feature without constraining its value; negative constraints are used to forbid the appearance of a feature with any value; set-membership constraints require a set of functional structures to contain a member with particular characteristics; and constraining equations require that some other word or rule in the sentence contribute a particular basic equation in order for the overall functional structure to be well-formed. (*See* **Constraint-based Processing**)

OTHER LINGUISTIC STRUCTURES

'Morphosyntactic structure' represents morphosyntactic dependencies between words that are not reflected in the functional structure. For example, the form and sequence of auxiliary verbs in English is fixed:

David has been eating apples. (17)

Other orders or forms are not possible:

*David been has eating apples.
*David have being eat apples. (18)

Such dependencies are represented and constrained in the morphosyntactic structure. (*See* **Syntax and Semantics: Formal Approaches**)

'Semantic structure' is important in determining the meaning of an utterance and the conditions under which it is true. Meanings of sentences are obtained by logical deduction from a set of semantic premises contributed by the words and syntactic structures in the sentence. These semantic premises are stated in a resource logic, 'linear logic', and they specify how the meanings of the parts of a sentence can be put together to produce the meaning of the full sentence. Different deductions from the same premises can correspond to different meanings: the phrase *alleged criminal from London* is ambiguous, referring either to someone who is alleged to be a criminal from London, or to someone who is actually from London and who is alleged to be a criminal. These two meanings correspond to different orders in which the modifiers *alleged* and *from London* are combined.

'Argument structure' represents the semantic information that is relevant for determining which participant in an event is the subject, which is the object, and so on. Consider a predicate that requires two arguments, one an active participant or instigator (an 'agent') and one that is acted upon (a 'patient'). In an active English sentence with such a predicate, the agent is always the subject and the patient is always the object. In the following English sentences, the kicker or thrower is an agent, and is the subject; the thing that is kicked or thrown is a patient, and is the object:

David kicked the ball.
David threw the javelin. (19)

There are no active sentences in English in which the kicker or thrower is the object and the thing that is kicked or thrown is the subject. The sentence below cannot mean the same thing as the sentence *David kicked the ball*:

The ball kicked David. (20)

However, in the related passive sentence, the thing that is kicked can in fact be the SUBJ:

The ball was kicked by David. (21)

The role of argument structure in determining functional syntactic roles like SUBJ and OBJ, and the nature of relation alternations as exhibited by active–passive pairs (like *David kicked the ball* and *the ball was kicked by David*), has been extensively explored in a number of languages.

'Information structure' represents how the information that the speaker intends to convey is presented and structured according to the topic of conversation, the focus of attention of the utterance, the common assumptions of the speaker and hearer, and what part of the information the speaker intends to emphasize to the hearer. Information structure has been shown to be closely related both to functional structure and to constituent structure; and so constraints on information structure are stated in terms of properties of both syntactic levels of structure. (*See* **Natural Language Processing**)

FUTURE PROSPECTS

Two recent developments in LFG deserve particular attention: a new view of language acquisition and processing in an LFG setting, and the combination of optimality theory with an LFG syntactic base.

'Data-oriented parsing' (DOP) views language acquisition as the analysis of a pool of linguistic structures that are presented to the language learner. The learner breaks up these structures into their component pieces, and new utterances are assembled from these pieces. The likelihood of assigning a particular analysis to a new sentence depends on the frequency of occurrence of its component parts in the original pool of structures. (*See* **Natural Language Processing, Statistical Approaches to**,)

LFG–DOP specializes DOP theory to LFG assumptions about linguistic structures and the relations between them: the body of linguistic evidence that a language learner is presented with consists of constituent structure–functional structure pairs, and language acquisition consists in determining the relevant component parts of these structures and then combining these parts in different ways to produce constituent structure–functional structure pairs for novel sentences.

Much recent research in phonology, morphology, syntax, and semantics has been conducted in the framework of optimality theory (OT). OT-based analyses assume that the grammar of a language consists of a 'generator' component, which proposes candidate linguistic structures for an input, and an 'evaluation' component, which selects the optimal structure from these candidates. The evaluation component consists of a ranked set of universally valid constraints that the optimal analysis must meet.

In OT–LFG, the input is taken to be an underspecified functional structure, and the generator component is a 'universal' LFG grammar that generates all well-formed constituent structure–functional structure pairs that are compatible with the input. The evaluation component determines the optimal candidate in a particular language from this set.

Further Reading

Alsina A (1996) *The Role of Argument Structure in Grammar: Evidence from Romance*. Stanford, CA: CSLI.

Andrews A and Manning CD (1999) *Complex Predicates and Information Spreading in LFG*. Stanford, CA: CSLI.

Bresnan J (ed.) (1982) *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Bresnan J (2001) *Lexical-Functional Syntax*. Oxford: Blackwell.

Butt M (1996) *The Structure of Complex Predicates in Urdu*. Stanford, CA: CSLI. [Revised and corrected version of 1993 Stanford University dissertation.]

Butt M and King TC (eds) (2000) *Argument Realization*. Stanford, CA: CSLI.

Butt M and King TC (eds) (2001) *Time over Matter: Diachronic Perspectives on Morphosyntax*. Stanford, CA: CSLI.

Butt M and King TC (eds) (2002) *On-Line Proceedings of the LFG Conferences*. http://csli-publications.stanford.edu/hand/miscpubsonline.html.

Choi H-W (1999) *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI. [Revised and corrected version of 1996 Stanford University dissertation.]

Dalrymple M (ed.) (1999) *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. Cambridge, MA: MIT Press.

Dalrymple M (2001) *Lexical Functional Grammar*. New York, NY: Academic Press.

Dalrymple M, Kaplan RM, Maxwell JT and Zaenen A (eds) (1995) *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI.

Falk YN (2001) *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, CA: CSLI.

King TH (1995) *Configuring Topic and Focus in Russian*. Stanford, CA: CSLI. [Revised and corrected version of 1993 Stanford University dissertation.]

Kroeger P (1993) *Phrase Structure and Grammatical Relations in Tagalog*. Stanford, CA: CSLI. [Revised and corrected version of 1991 Stanford University dissertation.]

Manning CD (1996) *Ergativity: Argument Structure and Grammatical Relations*. Stanford, CA: CSLI. [Revised and corrected version of 1994 Stanford University dissertation.]

Nordlinger R (1998) *Constructive Case: Evidence from Australian Languages*. Stanford, CA: CSLI. [Revised version of 1997 Stanford University dissertation.]

Simpson J (1991) *Warlpiri Morpho-Syntax: A Lexicalist Approach*. Dordrecht: Kluwer.

Lexicon, Computational Models of

Intermediate article

Ann Copestake, University of Cambridge Computer Laboratory, Cambridge, UK

CONTENTS

Natural language processing systems require information about individual words which is stored in the lexicon. Computational lexical representation techniques have been developed to model lexical knowledge.

INTRODUCTION

At the most basic level, a computational lexicon is a mapping from orthography or phonology to some combination of syntactic, semantic and pragmatic information. Its ultimate role is to deliver information to a system which analyzes or generates text or speech. This is in contrast to an electronic dictionary, which provides information to a human user (although there are some resources which combine the functions of lexicon and dictionary). Exactly what information is in the computational lexicon depends on the application: the lexicon for a part of speech tagger is likely to be considerably simpler than one for a natural language interface for instance.

A quite detailed lexicon is required by most computational systems, especially those based on linguistically motivated frameworks such as tree-adjoining grammar (TAG), lexical-functional grammar (LFG), head-driven phrase structure grammar (HPSG), dependency grammar (DG) or categorial grammar (CG). Although there are considerable differences between the various approaches, the main issues that concern lexical representation are common to all of them. In particular, they all require some way of representing generalizations about lexical behavior and some way of dealing with lexical ambiguity.

LEXICAL REPRESENTATIONS, LEXICAL HIERARCHIES AND INHERITANCE

Since the role of a lexicon is to provide information to an analyzer or generator, the precise nature of lexical representation depends on the requirements of those components. However, it is possible to think of all the commonly used approaches to lexical representation as forms of attribute-value representations. In such a formalism, information is expressed as pairings of attributes (or paths, consisting of lists of attributes) and associated values (which may be unspecified). Within this very broad characterization there is considerable variation in detail, but there are common organizational principles. Figure 1 shows a simplified example of an attribute-value representation for the lexeme (i.e. uninflected) transitive verb *tolerate*. The path-value notation is used explicitly at the top of the figure, while the bottom shows the same structure in attribute-value matrix (AVM) notation. In the examples in this article, an attribute, SUBCAT, will be used to represent the subcategorization properties as a whole, with attributes such as SUBJ, COMP1, COMP2 indicating the properties of the subject and individual complements. Such a representation could be used directly in a grammar expressed in a unification-based formalism, such as the PATR formalism (Shieber, 1986), which can be used to encode most of the grammar frameworks mentioned in the introduction. Although some systems maintain hard dividing lines between orthography/phonology, syntax and semantics, it is now common to assume a sign-based approach, in

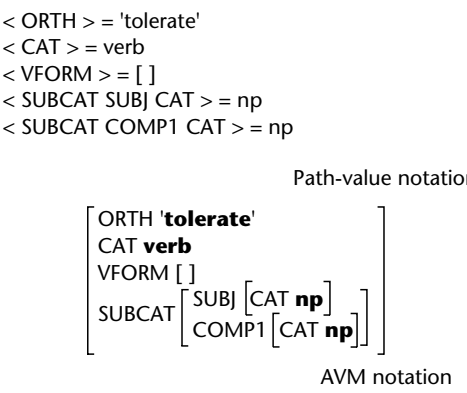


Figure 1. A simplified attribute-value representation for *tolerate* shown in two alternative notations. [ ] expresses an underspecified value.

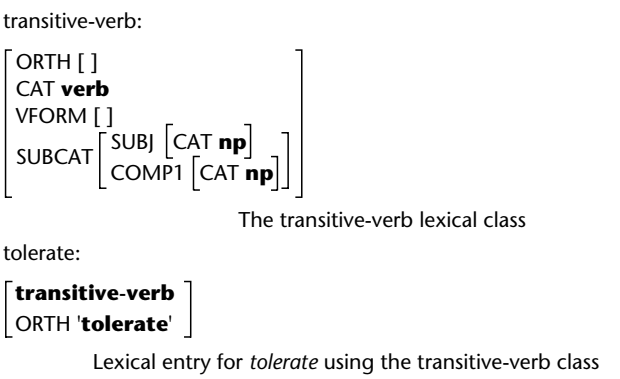


Figure 2. Using lexical classes to simplify lexical entries.

which a single representation combines information about all three levels (see, for instance, Jurafsky (1996), for discussion of why this is relevant to psycholinguistically plausible models). Pragmatic information might also be included, but in current practice pragmatic processing is nearly always application-specific and pragmatic information is omitted from the lexicon. For simplicity, Figure 1 omits semantics, and the discussion of lexical organization in the next sections concentrates on syntactic information, but the reader should bear in mind that the principles also apply to semantics.

Lexical Entries and Lexical Classes

The most basic form of organization is the lexical class. Notice that Figure 1 contains several individual pieces of structure. Rather than repeat all this for each transitive verb, the lexical entry (that is, the information stipulated for *tolerate*) should only contain the information that is idiosyncratic: generally the phonology/orthography and the semantic relation. It is therefore desirable to define a class for all simple transitive verbs, shown at the top of Figure 2, and used by the revised lexical entry for *tolerate* shown at the bottom of the figure. Note the contrast between the lexical entry and the structure to which it expands (i.e. Figure 1), which is what is accessed by the grammar rules when parsing or generating.

From an engineering viewpoint, the use of a class avoids many errors that might arise if a more complex structure had to be created directly (manually or automatically) and allows changes to be made in the representation without affecting the individual entry. Because the class expresses generalizations, it can add extra information. For instance, most entries in the COMLEX lexicon (Grishman *et al*., 1994) distributed by the Linguistic Data Consortium can be automatically converted into an HPSG lexicon by mapping them into predefined classes. This involves adding considerable extra information when considered in terms of the expanded lexical structures, but this is not problematic, because the extra information is predicted by the class. From a theoretical viewpoint, the class captures generalizations about lexical structure, for instance that simple transitive verbs have a subject and object position, and the inventory of lexical classes delimits the possible structures of the linguistic theory.

It is necessary to have a formal specification of the function that takes as input the individual entry together with its class and yields the expanded structure. Many systems which use unification in parsing also use unification for this operation, although this is not a necessary assumption.

Inheritance Hierarchies

The use of lexical classes can be seen as one-step inheritance. However, it is also desirable to capture relationships between the classes themselves by relating them in a hierarchy. The underlying idea is the same as inheritance in object-oriented programming: common information should be expressed in one place and inherited by subclasses. There is an indirect relationship to the use of IS-A links in semantic networks, but there are also differences: most importantly, lexical inheritance in modern representation languages is formally specified in such a way that the number of links in the hierarchy has no effect on behavior.

A portion of an inheritance hierarchy is sketched in Figure 3. Notice that some classes inherit from more than one parent. In a monotonic inheritance hierarchy, all information from all parents is combined with the information on the node itself:

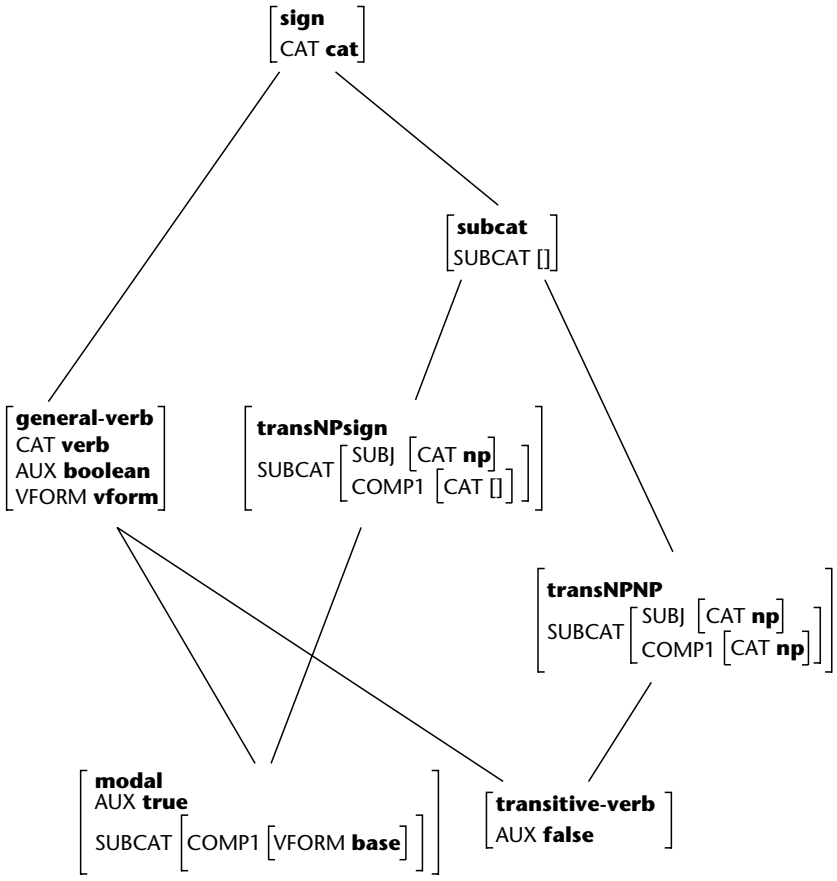


Figure 3. A fragment of a lexical inheritance hierarchy. Only the information local to each node is shown.

conflicts mean there has been an error in defining the hierarchy.

HPSG has made extensive use of inheritance in lexical description, following Flickinger *et al*. (1985), but all the frameworks mentioned in the introduction adopt it to some extent. HPSGs are often encoded using typed feature structures: in such formalisms, a type hierarchy is used for inheritance and to constrain the allowable structures.

Realistic lexical hierarchies are much bigger than Figure 3 would suggest and may easily involve many hundreds or thousands of classes. Cross-classification can be very helpful because it allows different parts of the hierarchy to be developed semi-independently of each other. It is worth noting that inheritance can be useful in the description of grammar rules and lexical rules as well as lexical structures, because the same principles of capturing generalizations are relevant.

Default inheritance

Many systems assume monotonic inheritance, which means that all information has to be mutually consistent. But there are some phenomena for which a purely monotonic account is problematic. The hierarchy in Figure 3 includes a class **modal-verb**, which specifies a single verbal complement which must have VFORM **base**. This is valid for most modal verbs, but *ought* is an exception (in most dialects of English): it patterns like a modal verb in most respects (it inverts, it can be negated without *do*, it takes the contracted negation *oughtn't* and it does not have distinct inflected forms) but it takes a *to*-infinitive complement. The distinction is demonstrated in the following examples:

1. I ought to go to bed.
2. *I ought go to bed.
3. *I should to go to bed.
4. I should go to bed.

In a formalism where inheritance is purely monotonic, the only option would be to split the class of modal verbs into two classes, one of which specifies VFORM **base** and another which specifies the *to*-complement. But this complicates the hierarchy and does not capture the intuition that *ought*

is the exceptional case. An alternative is the use of defaults, which is possible in the DATR representation language (Evans and Gazdar, 1996) and in various extensions of unification-based formalisms. For this particular example, the class for modal verbs specifies by default that the complement is a base form, but this is overridden by the entry for *ought*. This is shown in Figure 4. In this figure, only information after the '/' is default, so the only information about the modal verb class that can be overridden is the value of VFORM on its complement. In other variants of default inheritance, all information is potentially overridable.

Morphological Relationships and Lexical Rules

Lexical representation and morphology interact when representing inflected and derived forms of words. The syntactic and semantic effects of inflection may be captured in a simple inheritance hierarchy, on the assumption that inflection involves a fixed number of slots for number, gender, case etc., which take simple values, and that the base form is unmarked. Alternatively, lexical rules may be used, see below. It is also possible to specify slots for the stem and affix(es), with default inheritance being used for irregular verbs, for instance. This requires separate spelling rules to combine stem and affix to capture effects such as loss of *e* (e.g. *love* plus *ed* combine to give *loved*, not *loveed*). However in many computational systems, the phonological/morphological effects of affixation are captured with a finite-state approach or other device which involves a different formalism from the rest of the lexical representation. One exception is DATR, which can be used to define a finite-state transducer (Evans and Gazdar, 1996). Bird and Klein (1994) discuss the representation of phonology in a typed feature structure formalism.

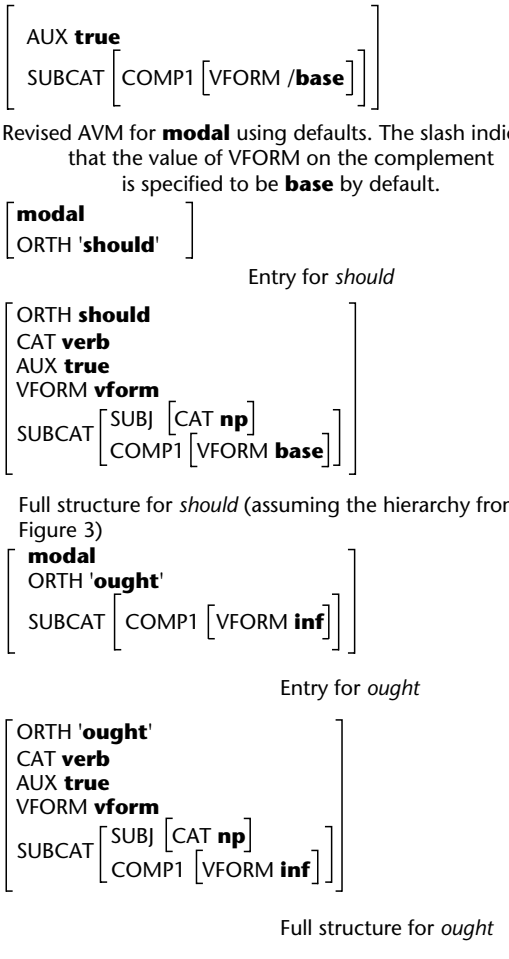


Figure 4. Default inheritance for *ought*.

Representation of derivational morphology is more complex, because it cannot be seen as instantiating a fixed number of simple attributes. For instance, the ending *ize/ise*, can convert an adjective X into a verb meaning (roughly) 'cause to become X' (e.g. *modern/modernize*). The number of derivational affixes does not in principle have an upper bound (e.g. *modernize/modernization/overmodernization/antiovermodernization*…), affixes do not have a fixed order of application (*godli-ness-less, god-less-ness*) and there is potential for multiple application of affixes (e.g. *antiantimissile*). This means that derivational morphology cannot in general be adequately represented using a simple inheritance hierarchy but requires some device capable of supporting recursive application.

Most commonly, derived forms are assumed to be related by some variety of *lexical rule*. Lexical rules are most straightforwardly thought of as having an input and an output which correspond to lexical classes, with some information being shared between the two parts. The rules are instantiated by individual lexemes: for derivational morphology, the output will be the affixed form. A sketch of a lexical rule for *+er* affixation of verbs to form agentive nominals is shown in Figure 5. The rule itself is very simple, because it makes use of the same lexical classes as the lexical entries do. The boxed integer tag indicates that the stem orthography is the same in the input and the output.

A complicating factor is that derivational morphology is semi-productive. This shows itself in two ways: a derived form may not have exactly the behavior predicted by the rule, or it may not occur at all. To avoid generating incorrect forms, lexical rules may be used as redundancy rules: that is, derived forms are explicitly listed in the lexicon

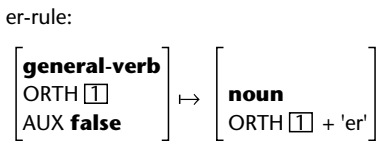


Figure 5. A lexical rule for affixation with *+er*.

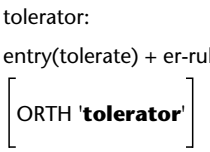


Figure 6. Using a lexical redundancy rule in a lexical entry.

but the lexical rule is used as part of the description. See Figure 6, for example. If default inheritance is available, the entry may specify that some part of the rule output is overridden. The disadvantage of this approach is that it does not predict novel forms: thus, faced with a previously unseen word *cloner*, the system has no way of relating it to *clone*. Theoretically, it is clear that productive derivational morphology is required at some level, although it might be seen as similar to acquisition of completely novel lexical entries rather than as analogous to inflection. Practically, no system can ever include all words in its lexicon: even after hundreds of millions of words of text have been processed, previously unseen forms still arise. But whether to treat some cases of derivational morphology as productive or not depends on the application. For a language generator, it is generally better to avoid making derivational morphology productive, since nonce formations are often regarded with some degree of distaste by hearers. On the other hand, for an analyzer, accepting previously unseen forms is essential for robustness.

There is a close relationship between derivational morphology and some cases of lexical ambiguity, discussed later in this article.

Multi-word Expressions

The lexical entries considered so far concern single (orthographic) words of a language, though possibly with internal morphological structure. There will also be entries that can be considered as corresponding to words which happen to be spelled with spaces (e.g. *ad hoc*). However, the lexicon must also contain entries for genuine multi-word expressions, including compounds (e.g. *house boat*), verb particle constructions (e.g. *tidy up*), idioms and collocations. The challenge to lexical representation is that many multi-word expressions are partly but not wholly compositional. For instance, the use of *tidy* in *tidy up* is clearly related to the verb *tidy* and *up* has a comparable use in *clear up, wash up* etc. In the extreme case of idioms such as *pull strings*, parts of the idiom may occur in different clauses: *The diplomatic strings that the ambassador pulled got the company the contract*. The representation of such idioms must therefore involve at least some aspect of the representation of compositionally constructed phrases. Several approaches have been proposed, most notably within TAG (Abeillè and Schabes, 1989), but this aspect of lexical representation is still very much an open issue. This is also true of the treatment of collocations such as *heavy sea* (note that *heavy swell* and *strong swell* are also collocations, but *strong sea* is unidiomatic). Collocations are the focus of much work within Meaning-Text Theory (MTT), some of which is computational in orientation (e.g. Mel'čuk and Polguère, 1987).

Semantic Representation

Unfortunately there is little agreement about the amount of semantic information that should be included in the computational lexicon, and even less about the way it should be represented. There has been much discussion about the distinction between lexical and real-world knowledge, but no consensus as to where the dividing line might be drawn or even if there is a theoretically motivated dividing line. At one extreme, semantic information in the lexicon is limited to predicate-argument structure. For instance, Figure 7 shows an entry for *tolerate* containing a semantic representation which links the subject and object to argument positions in the semantic representation (the linking is indicated by the boxed integers in the AVM). But by itself this says very little about the meaning of *tolerate* and it misses generalizations about the syntax–semantics interface. The representation in Figure 7 uses the semantic roles EXPERIENCER and EXPERIENCED but there is no agreed inventory of such roles. However, the FrameNet project (Baker *et al*., 1998) is developing frames that encode basic conceptual structures for a variety of semantic domains.

Unfortunately, attempting to fully represent word meaning leads into complex knowledge representation issues which are theoretically and computationally intractable. Even apparently simple information, such as selectional restrictions, can be problematic. For instance, it would seem

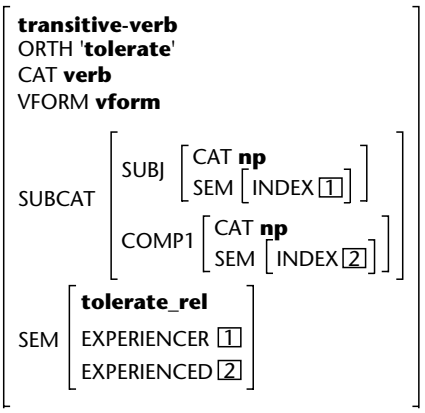


Figure 7. Lexical entry for *tolerate* including a very simple predicate-argument representation.

that the subject of *tolerate* (i.e. the experiencer) should be sentient, but, like many other similar verbs, *tolerate* is also used of plants, for instance: *Azaleas cannot tolerate alkaline soil*. If selectional restrictions are encoded as hard constraints, this sentence cannot be parsed. Selectional restrictions are thus generally used to guide processing by means of preferences rather than to completely rule out parses.

Practically, much depends on the application and on the system architecture: for instance, it is often possible to make semantic relationships domain-specific. There is considerable scope for utilizing statistical techniques to provide an approximation of (some aspects of) meaning: this is discussed further below, in the context of ambiguity and disambiguation.

One compromise position is to restrict semantic information in the lexicon itself to that which is necessary to capture generalizations about syntax or morphology. The representation tools described so far are generally adequate for this, since the need is for inheritance rather than full logical inference. For instance, in typed feature structure formalisms, a hierarchy of semantic types which includes predicates such as **tolerate_rel** can be defined.

Taxonomic information of a more general sort is also of proven utility in some applications, such as retrieval from document databases, and other uses are being very actively investigated. Much of this work uses domain-specific taxonomies but researchers requiring general-purpose taxonomies often rely on WordNet (Fellbaum, 1998). Taxonomic links are generally not formally defined: WordNet's hyponymy links, for instance, cover traditional taxonomic information, such as the relationship between a species and a family in biology, but are also used for relationships between verbs, which raises the question of argument mapping. Both WordNet and FrameNet are resources which can be utilized by a range of systems, rather than as computational models in their own right.

LEXICAL AMBIGUITY

Discussions of ambiguity are frequently motivated by examples such as *bank*: financial institution versus *bank*: geographical feature. It is clear that there are two quite distinct senses of *bank* and there is nothing interesting to say about the synchronic relationship between them, even if there is a historical relationship. A corresponding computational representation would be to have two lexical entries for the noun *bank* which are unrelated apart from having the same spelling.

There are several reasons why computational models of ambiguity are less straightforward than this would suggest. The first is that most ambiguity actually concerns related senses: for instance *bank*: financial institution versus the verb *bank*: to carry out transactions at a bank. Here the senses are clearly related, though they have different parts of speech. As illustrated below, different types of lexical ambiguity require different computational models (Copestake and Briscoe, 1995). Furthermore, the distinctions made between related senses in conventional dictionaries are not mutually consistent and do not offer a particularly good guide to a computational lexicographer. The second issue is more related to the engineering issues of computational lexicography: representing lexical ambiguity with multiple entries is computationally expensive, since it expands the search space required by parsing. This leads to the use of various techniques to pack ambiguity into one lexical entry. Many computational lexicons do not in fact have completely distinct lexical entries for the two distinct nominal senses of *bank*, either because the representations for the senses are packed together or because the semantics is so coarse-grained that the distinction is not representable. Finally, it is important both theoretically and practically to consider the relative frequencies of different senses.

Classes of Ambiguity

The classical *bank* ambiguity is an example of homonymy, which technically refers to totally unrelated senses. From the computational perspective, words which intuitively have some relationship but where the meaning difference is unpredictable are treated with distinct entries in the same way as homonyms proper. An example is *bank*: geographical feature

and *bank*: to turn an airplane. Many people intuitively feel there is a relationship between these senses, since both involve a slope, and there is a clear historical connection, but there is no generalization to be made about the relationship.

More interesting cases involve some regularity in the polysemy. For example, the nouns *heap*, *pile*, *mound* etc., referring to a raised mass of some type, have corresponding verbs, which (roughly) mean to construct such an object. This class of polysemy is very similar to derivational morphology (compare this example with the suffixes *-ize*, *-ify*, for instance) and can be represented computationally by a lexical rule in a comparable way. Issues of semi-productivity arise here, much as they do with derivational morphology.

A subtle form of polysemy is exemplified by the noun *cloud*, which can be used not only for the weather phenomenon, but also for similar amorphous masses made up of dust, flies, etc. Usually the non-weather use is made explicit by compounding (*dust cloud*) or an *of*-PP (*cloud of flies*). A similar pattern is found in many other nouns, for example *belt/belt of snow*, *ribbon/ribbon of sauce*. Effectively there is one relatively specific sense and a range of other more general uses which contain only part of its meaning. Such examples may be classed as figurative but they are conventionalized and quite common in standard corpora. Furthermore, the syntactic properties sometimes differ, since the general uses subcategorize for the of-PP. It is therefore necessary to account for these senses in a computational lexicon. One approach is to assume a general (underspecified) structure for *cloud*, which by default refers to weather clouds, but which can be contextually specialized to get the other uses, overriding default information about composition. Copestake and Briscoe (1995) discuss such examples and relate them to the examples of *logical metonymy* which have been discussed in detail by Pustejovsky (1995).

There are many cases where some conventional dictionaries split senses where there may well be no reason to make any sense distinction in a computational lexicon. For instance, many dictionaries distinguish between *bank*: mound and *bank*: river bank as separate senses but most computational lexicographers would not split the uses. In this particular case, that decision could be justified by examples such as the following: *Removal of vegetation may destabilize banks*. If *banks* here refers to both mounds and river banks, then there must be a single lexical structure which encompasses them both (although this could be an underspecified structure, with specializations corresponding to the two dictionary senses). But, in general, if the computational lexicon gives the same structure for two senses, there is little point in having two entries.

Practically speaking, the sense distinctions made in many systems are driven by the application. For instance, a natural language interface may use the ontology of the underlying system for guidance in distinguishing senses while a machine translation lexicon might use distinctions in the target language. But this can produce problems if the systems are extended. For instance, if another target language was added that shared the ambiguity found in the source language, the strategy of multiplying lexical entries would mean that an unnecessary disambiguation step was required for the new language pair. On the whole it is best to arrange the system architecture so that choices are made at the latest possible processing stage.

Lexical Probabilities

Lexical representation is not necessarily entirely symbolic – many systems include some probabilistic or frequency information. This is most apparent in statistical approaches. A lexical entry for a statistical part-of-speech tagger contains all possible tags for the string plus associated probabilities, which are acquired automatically while training the tagger on text. For instance, the entry for *bank* might look as follows (based on data adapted from the tagged British National Corpus): bank: V 0.002 NP 0.214 N 0.784. Lexical probabilities are important in statistical tagging and statistical parsing, since, as the example illustrates, the distribution of parts of speech is very uneven. This is true of senses more generally: usually some are much more frequent than others, although the particular frequencies will depend on the domain of the text. Relative lexical frequency metrics are also important in some analyzers based on manually constructed grammars, although these are generally implemented as weights rather than genuine probabilities because of the difficulty of producing probabilistic grammars for non-context-free formalisms. Developing lexicons which have accurate estimates of probabilities for finer-grained information than a simple part-of-speech is an ongoing research goal: constructing such lexicons manually is unrealistically time-consuming, but automatic acquisition techniques run into difficulties because of data sparseness and ambiguity. Lexicons with accurate frequency information can be expected to lead to more psychologically plausible models of processing, in particular with respect to ambiguity and the treatment of multi-word expressions.

Disambiguation and Lexical Representation

Disambiguation is dependent on lexical representation in a very basic way since it is only required to the extent that the lexicon makes a sense distinction. However, because of the need to compare results, much research has involved disambiguation with respect to some external standard, such as a dictionary or WordNet. Disambiguation algorithms in complete systems can take advantage of the lexical information required for other aspects of processing, but many researchers also use some external information source to guide disambiguation. For example, information about word co-occurrences (acquired either manually or automatically) is an important source of information for many disambiguation algorithms. Arguably, at least, this is lexical information, but it is of a somewhat different type from the sort of information needed by a conventional parser. Lexical disambiguation is inevitably imperfect: it is easy to invent examples which require arbitrary amounts of world knowledge and reasoning. But human hearers sometimes make errors due to lexical ambiguity too! What is important is whether the disambiguation rate that can be achieved with the available knowledge sources is acceptable for a given application and whether it is possible to recover from errors.

Some research on disambiguation uses meaning representations automatically derived from corpora. Given the discussion of sense distinctions above, it seems plausible to assume a model of meaning where individual uses of words have different degrees of relatedness, and senses are clusters of uses, which naturally leads to a notion of a sense distinction as a cline, rather than a hard boundary. This fits in quite naturally with models of meaning based on local context of individual uses acquired automatically from corpora, which have become much more attractive in recent years because of the availability of large amounts of text in machine-readable form and increases in computing power. Schütze (1996) describes one approach that uses a relatively simple notion of co-occurrence to give representations which consist of very high dimension vectors, but there are many ways in which this basic idea might be refined. Work along these lines is promising because it may address some of the weaknesses of purely qualitative approaches to lexical representation. Specifically, it may provide an approach to obtaining a fine-grained meaning representation automatically in a way that allows for different sense frequencies and for collocations. But this is unlikely to entirely supersede symbolic attribute-value representations and the challenge will be to combine the two in a way that realizes the strengths of both.

References

Abeillè A and Schabes Y (1989) *Parsing Idioms in Lexicalized Tree Adjoining Grammars*. Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL-89), Manchester, England, pp. 1–9.

Baker CF, Fillmore CJ and Lowe JB (1998) *The Berkeley FrameNet Project*. Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, Montreal, Canada, pp. 86–90.

Bird S and Klein E (1994) Phonological analysis in typed feature systems. *Computational Linguistics* **20**(3): 455–491.

Copestake A and Briscoe EJ (1995) Semi-productive polysemy and sense extension. *Journal of Semantics* **12**: 15–67.

Evans R and Gazdar G (1996) DATR: a language for lexical knowledge representation. *Computational Linguistics* **22**(2): 167–216.

Fellbaum C (ed.) (1998) *WordNet, An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Flickinger DP, Pollard C and Wasow T (1985) *Structure Sharing in Lexical Representation*. Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL-85), University of Chicago, pp. 262–268.

Grishman R, Macleod C and Meyers A (1994) *Complex Syntax: Building a Computational Lexicon*. Proceedings of the 15th International Conference on Computational Linguistics. *COLING-94*, Kyoto, Japan, pp. 268–272.

Jurafsky D (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* **20**: 137–194.

Mel'čuk I and Polguère M (1987) A formal lexicon in Meaning-Text Theory (or how to do lexica with words). *Computational Linguistics* **13**(3–4): 261–275.

Pustejovsky J (1995) *The Generative Lexicon*. Cambridge, MA: MIT Press.

Schütze H (1996) *Ambiguity in Language Learning: Computational and Cognitive Models*. Stanford, CA: CSLI Publications.

Shieber SM (1986) *An Introduction to Unification-based Approaches to Grammar*. CSLI Publications.

Further Reading

Briscoe EJ (1991) Lexical issues in natural language processing. In: Klein E and Veltman F (eds) *Natural Language and Speech*, pp. 39–68. Berlin: Springer-Verlag.

Briscoe EJ, Copestake A and de Paiva V (eds) (1993) *Inheritance, Defaults and the Lexicon*. Cambridge, UK: Cambridge University Press.

Carpenter B (1992) *The Logic of Typed Feature Structures*. Cambridge, UK: Cambridge University Press.

Daelemans W, de Smedt K and Gazdar G (1992) Inheritance in natural language processing. *Computational Linguistics* **18**(2): 205–218.

van Eynde F and Gibbon D (eds) (2000) *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic.

Ide N and Véronis J (1998) Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics* **24**(1): 1–40.

Levin B (1993) *English Verb Classes and Alternations*. Chicago, IL: University of Chicago Press.

Pustejovsky J and Bergler S (eds) (1992) *Lexical Semantics and Knowledge Representation*. Berlin: Springer-Verlag.

Saint-Dizier P and Viegas E (eds) (1995) *Computational Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Walker D, Zampolli A and Calzolari N (eds) (1995) *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Oxford, UK: Oxford University Press.

Zernik U (ed.) (1992) *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum.

Lexicon

Advanced article

James Pustejovsky, Brandeis University, Waltham, Massachusetts, USA

CONTENTS

The lexicon of a grammar must provide a systematic and efficient way of encoding information about the words in a language.

WHAT IS A WORD?

Methodological Preliminaries

Linguistic computation requires access to knowledge about words. In this article, the requirements on the lexicon are examined in terms of the information required for syntax, and in terms of the knowledge needed for semantic interpretation.

Although there is little agreement on the exact nature of what should go into a lexicon, there are some important common assumptions held by researchers in linguistics on the form of a lexical entry. It is generally assumed that there are four necessary components for the structure of a lexical item. These are: syntactic information (e.g. what part of speech the word is); semantic information (i.e. how the word is interpreted); orthographic and morphological information (i.e. how the word is spelled and what form it takes); and phonetic and phonological information (e.g. how the word is pronounced and stressed).

But the lexicon is not merely a collection of words with their associated phonetic and orthographic forms. Words are structured objects that participate in larger operations and compositions, acting as signatures to semantic entailments and implicatures in the context of discourse and text. So we need to address the following four questions when designing a computational lexicon or modeling our mental lexicon:

- What information goes into a single lexical entry?
- How do lexical entries relate to one another?
- How is this information exploited by the grammar?
- How is this information available to general reasoning?

As regards all of these questions views on lexical knowledge vary in the linguistic sciences. To a large extent, the nature of the grammar itself is determined by what information the lexicon contains for the other grammatical components. Nevertheless, the lexicon is often viewed as the most passive module of grammar. Since the 1970s, the amount of information that is associated with lexical items, as well as what is taken to be a lexical item in the first place has changed considerably. What started as a passive repository of pure stems, with active word-formation rules generating everything from inflected forms to derivational compounds, has developed into a distributed library of lexically based syntactic encodings, generative operations, and object-oriented processes.

There are essentially two reasons for the changing perspective on the lexicon in recent years. Firstly, there is now a tighter integration of compositional operations of syntax and semantics with the lexical information structures that bear them. Secondly, lexical researchers have become more concerned with how lexical types reflect the underlying ontological commitments of the grammar. The field has moved towards addressing more encompassing problems in linguistic theory, such as polysemy, global lexical structure, syntactic linking, and the relation between lexical and world knowledge. In this article, we review briefly the conventional view of the lexicon and then contrast this with newly emerging theories of lexical information.

The conventional view of the lexicon is of a file of words, acting in the service of the more dynamic components of the grammar. This view has its origins in the generative tradition (Chomsky, 1975) and has been an integral part of the notion of the lexicon ever since. While the *Aspects*-model of selectional features (Chomsky, 1965) restricted the relation of selection to that between lexical items, work by Jackendoff (1972) and McCawley (1968) showed that selectional restrictions must be available to computations at the level of derived semantic representation rather than at deep structure. Subsequent work by Pollard and Sag (1994) and others extends the range of phenomena that can

be handled by the projection and exploitation of lexically derived information in the grammar. Recently, with the convergence of several areas in linguistics (lexical semantics, computational lexicons, type theories), several models for the determination of selection have emerged that put even more compositional power in the lexicon, making explicit reference to the paradigmatic systems that allow for grammatical constructions to be partially determined by selection. Examples of this approach are 'generative lexicon theory' (Pustejovsky, 1995), and to a certain extent 'construction grammar', combinatory categorial grammar (CCG), and Jackendoff's recent work. These developments have helped to characterize approaches to lexical design in terms of a hierarchy of semantic expressiveness. There are at least three such classes of lexical description (see Pustejovsky, 1995): 'sense-enumerative' lexicons, where lexical items have a single type and meaning, and ambiguity is treated by multiple listings of words; 'polymorphic' lexicons, where lexical items are active objects, contributing to the determination of meaning in context, under well-defined constraints; and 'unrestricted sense' lexicons, where the meanings of lexical items are determined mostly by context and conventional use. The most promising direction of study seems to be a careful and formal elucidation of the polymorphic lexicons, and this will form the basis of our discussion below of both the structure and the content of lexical entries.

Multi-Word Expressions

If we assume that a word can be viewed as a data structure associating a specific orthography and phonology with syntactic behavior, meaning, and conventions of usage, then arguably the lexicon is populated by not just single morphemes but multiword expressions as well. There are two major classes of multi-word expressions that we will consider here: collocations and idioms.

It is often merely an historical accident that a language (or dialect) lexicalizes a concept as a single morpheme rather than a multi-word expression: compare the British English *hoover* with American English *vacuum cleaner*, for example. The latter type of compound word is commonly referred to as a 'collocation', of which there are two kinds:

- A collocation that makes up a single, non-phrasal constituent. Examples include proper names such as *George Miller* and *Arlington Heights*, and grammatical formatives such as *in addition to* and *such that*. There is a single, non-phrasal category associated with the entire expression.
- A collocation that makes up more than one non-phrasal category. This happens in the case of lexical items that must co-occur with a particular word, rather than an entire category. Examples include so-called 'verb–particle' combinations, such as *blow up* and *write down*, and verbs that select a prepositional phrase headed by a particular preposition, such as *look for* and *wait for*. Note that, in the verb–particle case, material may intervene between the verb and the particle (*blow the building up*). In both cases, inflectional material appears on the first word of the expression (*blew up, looked for*).

Idioms might be considered a special case of collocations, having many constituents. They are, however, distinguished from collocations by a number of criteria. Firstly, they have a noncompositional semantics. Typically, the interpretation of an idiom is not derivable from its component parts: *kick the bucket*, *the cat's got his tongue*. Secondly, they exhibit 'fixedness'. Often an idiom, unlike a collocation, is a fixed expression, specifying an entire complement structure, rather than some subsidiary element. Contrast *kick the bucket* (idiom) with *look for him* (collocation). This is not universally true of idioms, since some do allow unfixed elements to appear as part of the complement structure: *the cat's got his tongue*, *take advantage of him*. Note, however, that even in these cases, much more of the complement structure is specified than in cases of collocation. Also, lexical items, and not merely grammatical formatives (such as determiners), are specified in these examples. Finally, they violate grammatical constraints. The idiom *take advantage of* presents a good example of this. Firstly, it violates the normal subcategorization frame of *take*, which does not permit *of* as the preposition following the object NP (*John took books from/to/*of Bill*). Secondly, *advantage* appears without a determiner in this idiom, though it normally requires one (*I can see *(the) advantage in that*).

If collocations and idioms are stored as units in our mental lexicon, then there are several related issues that linguistic theories must deal with. How are multi-words indexed? And how much internal morphological or syntactic structure do multiwords have or allow access to?

THE PROBLEM OF POLYSEMY

Given the compactness of a lexicon, relative to the number of objects and relations in the world and the concepts we have for them, lexical ambiguity is inevitable. Add to this the cultural, historical, and linguistic blending that contributes to the meanings

of our lexical items, and ambiguity can appear arbitrary as well. Hence, homonymy – where one lexical form has many meanings – is to be expected in a language. Examples of homonyms are illustrated in the sentences below.

(a) Mary walked along the *bank* of the river.
(b) FleetBank is the largest *bank* in the city. (1)

(a) Drop me a *line* when you are in Boston.
(b) We built a fence along the property *line*. (2)

(a) First we leave the gate, then we *taxi* down the runway.
(b) John saw the *taxi* on the street. (3)

(a) The discussion *turned* on the feasibility of the scheme.
(b) The bull *turned* on the matador. (4)

(a) The judge asked the defendant to approach the *bar*.
(b) The defendant was at the *bar* in the pub. (5)

Weinreich calls such lexical distinctions 'contrastive ambiguity', where the senses are distinct and generally unrelated. For this reason, it is generally assumed that homonyms are represented as separate lexical entries within the organization of the lexicon. This accords with a view of lexical organization that has been termed a 'sense-enumeration' lexicon (Pustejovsky, 1995). A lexicon L is a sense-enumeration lexicon if and only if for every word w in L, having multiple senses $s_1,\ldots,s_n$ associated with it the lexical entries expressing these senses are stored as $\{w_{s_1},\ldots,w_{s_n}\}$. Words with multiple senses are simply listed separatedly in the lexicon. This does not seem to compromise or complicate the compositional process (how words combine in the interpretation of a sentence).

This model becomes difficult to maintain, however, when we consider the phenomenon of polysemy. Polysemy is the relationship that exists between different senses of a word that are related in some logical manner, rather than arbitrarily, as in the above examples. We can distinguish four broad types of polysemy, each presenting a distinct set of challenges to linguistic theory:

- Deep semantic typing: single-argument polymorphism.
- Syntactic alternations: multiple-argument polymorphism.
- Dot objects: lexical reference to objects that have multiple facets.
- Terms of generalization: light verbs and general predicates.

The first class refers mainly to functors allowing a range of syntactic variation in a single argument. For example, aspectual verbs (*begin, finish*), perception verbs (*see, hear*), and most propositional attitude verbs (*know, believe*) subcategorize for multiple syntactic forms in complement position, as illustrated in example 6 below.

(a) Mary began to read the novel.
(b) Mary began reading the novel.
(c) Mary began the novel. (6)

(a) Bill saw John leave.
(b) Bill saw John leaving.
(c) Bill saw John. (7)

(a) Mary believes that John told the truth.
(b) Mary believes what John said.
(c) Mary believes John's story. (8)

What these and many other cases of multiple selection share is this: the underlying relation between the verb and each of its complements is essentially identical. Thus, in example 8, the complement to the verb *believe* in all three sentences is a proposition; in example 6, what is begun in each sentence is an event of some sort; and in example 7, the object of the perception is (arguably) an event in each case. This observation has led some linguists to argue for semantic selection and others to argue for structured selectional inheritance. In fact, these perspectives are not very distant from one another (Pustejovsky, 1995): on either view, there is an explicit lexical association between syntactic forms which is formally modeled by the grammar.

The second type of polysemy mentioned above (syntactic alternations) involves verbal forms taking arguments in alternating constructions, the so called 'verbal alternations'. These are true instances of polysemy because there is a logical (typically causal) relation between the two senses of the verb. As a result, the lexicon must either relate the senses through lexical rules, as in head-driven phrase structure grammar, or assume that there is one lexical form that has multiple syntactic realizations.

(a) The window opened suddenly.
(b) Mary opened the window suddenly. (9)

(a) Bill began his lecture on time.
(b) The lecture began on time. (10)

(a) The milk spilled onto the table.
(b) Marry spilled the milk onto the table. (11)

The final form of polysemy we will review here is encountered mostly in nominals, and has been termed 'regular polysemy' (Apresjan, 1973) and 'logical polysemy' (Pustejovsky, 1991) in the literature. It is illustrated in the sentences below.

(a) Mary carried the book home.
(b) Mary doesn't agree with the book. (12)

(a) Mary has her lunch in her backpack.
(b) Lunch was longer today than it was yesterday. (13)

(a) The flight lasted three hours.
(b) The flight landed on time in Los Angeles. (14)

Notice that in each of the pairs above, the same nominal form is assuming different semantic interpretations relative to its selective context. For example, in sentence 12(a) the noun *book* refers to a physical object, while in sentence 12(b) it refers to the informational content. In sentence 13(a), *lunch* refers to the physical manifestation of the food, while in sentence 13(b) it refers to the eating event. In sentence 14(a), *flight* refers to the flying event, while in sentence 14(b) it refers to the plane. This phenomenon of polysemy is one of the most challenging in the area, and has stimulated much research recently. In order to understand how each of these cases of polysemy can be handled, we must first familiarize ourselves with the structure of individual lexical entries.

BASIC PROPERTIES OF A LEXICAL ENTRY

As stated above, it is generally assumed that there are four components to a lexical item: phonological, orthographic, syntactic, and semantic information. We now focus on what syntactic and semantic information must be encoded in an individual lexical entry.

There are two types of syntactic knowledge associated with a lexical item: its 'category' and its 'subcategory'. The former includes traditional classifications of both the major categories (such as nouns, verbs, adjectives, adverbs, and prepositions) and the minor categories (such as adverbs, conjunctions, quantifier elements, and determiners). Knowledge of the subcategory of a lexical item is typically information that differentiates categories into distinct, distributional classes. This sort of information may be usefully separated into two types: 'contextual features' and 'inherent features'. The former are features that may be defined in terms of the contexts in which a given lexical entry may occur. Subcategorization information marks the local syntactic context for a word. It is this information that ensures that the verb *devour*, for example, is always transitive in English, requiring a direct object; the lexical entry encodes this requirement with a subcategorization feature specifying that an NP appear to its right. Another type of context encoding is collocational information, where patterns that are not fully productive in the grammar can be tagged. For example, the adjective *heavy* as applied to *drinker* and *smoker* is collocational and not freely productive in the language (Mel'čuk, 1988). 'Inherent features', on the other hand, are properties of lexical entries that are not easily reduced to a contextual definition, but rather refer to the ontological typing of an entity. These include such features as count versus mass (e.g. *pebble* versus *water*), abstract, animate, human, physical, and so on.

Semantic, information can also be separated into two categories: 'base semantic typing' and 'selectional typing'. While the former identifies the semantic class that a lexical item belongs to (such as entity, event, or property), the latter specifies the semantic features of arguments and adjuncts to the lexical item. We turn first to semantic typing. (*See* **Lexical Semantics**)

Word Classes and Types

Lexical items can be systematically grouped according to their syntactic or semantic behavior in the language. There have thus been two major traditions of word clustering. Broadly speaking, for those concerned mainly with grammatical behavior, the most salient aspect of a lexical item is its argument structure; for those concerned with a word's entailment properties, the most important aspect is its semantic class. We will examine these two approaches and see how their concerns can be integrated into a common lexical representation.

Conventional approaches to lexicon design and lexicography are relatively informal regarding the taxonomic structures for the word senses in the language. For example, the top concepts in Word Net illustrate how words are characterized by local clusterings of semantic properties. As with many ontologies, however, it is hard to discern a coherent

global structure for the resulting classification beyond a weak descriptive labeling of words into extensionally defined sets.

One of the most common ways to organize lexical knowledge is by means of type or feature inheritance mechanisms (Pollard and Sag, 1994). Furthermore, Briscoe *et al*. (1993) describe a rich system of types for allowing default mechanisms into lexical type descriptions. Similarly, type structures can express the inheritance of syntactic and semantic features, as well as the relationship between syntactic classes and alternations (Sanfilippo, 1993; Davis, 1996; Alsina, 1992) and other relations (Pustejovsky and Boguraev, 1993). (*See* Figure 1.)

Next, we will examine the approach to characterizing the weak constraints imposed on a lexical item according to its arguments. Then, we will examine attempts to model lexical behavior by means of internal constraints imposed on the predicate itself. Finally, we will show how, in some respects, these are very similar enterprises, and how both sets of constraints may be necessary to model lexical behavior.

Argument Structure

Once the base syntactic and semantic typing for a lexical item has been specified, its subcategorization and selectional information must be encoded in some form. There are two major techniques for representing this type of knowledge:

- Associate 'named roles' with the arguments to the lexical item (Jackendoff, 1972).

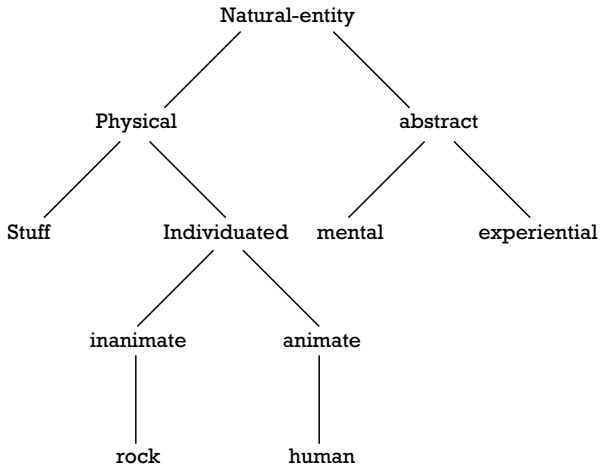


Figure 1. Inheritance of features between types of natural entity.

- Associate a logical decomposition with the lexical item. Meanings of arguments are determined by how the structural properties of the representation are interpreted (Hale and Keyser, 1993; Levin and Rappaport Hovav, 1995).

One influential way of encoding selectional behavior has been the theory of 'thematic relations' (Gruber, 1976; Jackendoff, 1972). Thematic relations are now generally defined as partial semantic functions of the event being denoted by the verb or noun, and behave according to a predefined calculus of role relations. For example, semantic roles such as agent, theme, and goal, can be used to partially determine the meaning of a predicate, when they are associated with the grammatical arguments to a verb.

(a) put < AGENT, THEME, LOCATION >
(b) borrow < RECIPIENT, THEME, SOURCE >
(15)

Thematic roles can be ordered relative to each other in terms of an implicational hierarchy. Universal subject hierarchies such as shown below are often used.

AGENT > RECIPIENT/BENEFACTIVE > THEME/PATIENT > INSTRUMENT > LOCATION (16)

Many linguists have questioned the general explanatory coverage of thematic roles, however, and have have chosen alternative methods for capturing the generalizations they promised. Dowty suggests that theta-role generalizations are best captured by entailments associated with the predicate itself. A theta-role can then be seen as the set of predicate entailments that are properties of a particular argument to the verb. Characteristic entailments might be thought of as prototype roles, or '*proto-roles*'; this allows for degrees or shades of meaning associated with the arguments to a predicate. Others have opted for a more semantically neutral set of labels to assign to the parameters of a relation, whether it is realized as a verb, noun, or adjective. For example, the theory of argument structure as developed by Williams (1981), Grimshaw (1990), and others can be seen as a move towards a more minimalist description of semantic differentiation in the verb's list of parameters. The argument structure for a word can be seen as the simplest specification of its semantics, indicating the number and type of parameters associated with the lexical item as a predicate. For example, the verb *die* can be represented as a predicate taking one argument, *kill* as taking two arguments, and *give* as taking three arguments:

(a) die($\underline{x}$)
(b) kill($\underline{x}, y$)
(c) give($\underline{x}, y, z$) (17)

What began as the simple listing of the parameters or arguments associated with a predicate has developed into a sophisticated view of the way arguments are mapped onto syntactic expressions. Williams's (1981) distinction between 'external' arguments (the underlined arguments above) and 'internal' arguments, and Grimshaw's (1990) proposal for a hierarchically structured representation, provide us with the basic syntax for one aspect of a word's meaning. Similar remarks hold for the argument list structure in head-driven phrase structure grammar (Pollard and Sag, 1994) and lexical-functional grammar.

The interaction of a structured argument list and a rich system of types, such as that presented above, provides a mechanism for semantic selection through inheritance. Consider, for instance the sentence pairs below:

(a) The man/the rock fell.
(b) The man/*the rock died. (18)

Now consider how the selectional distinction for a feature such as animacy is modeled so as to explain the selectional constraints of predicates. For the purpose of illustration, the arguments of a verb will be identified as being typed from the system shown above.

(a) λx[fall($\underline{x}$: physical)]
(b) λx[die($\underline{x}$: animate)] (19)

In example 18, it is clear how rocks cannot die and men can, but it is not obvious how this judgment is computed, given what we would assume are the types associated with the nouns *rock* and *man*, respectively. What accomplishes this computation is a rule of subtyping, Θ, that allows the type associated with the noun *man* (i.e. human) to also be accepted as the type animate, which is what the predicate *die* requires of its argument as stated in equation 19(b).

$$\Theta[\texttt{human} \sqsubseteq \texttt{animate}] : \texttt{human} \rightarrow \texttt{animate} \quad (20)$$

The rule Θ applies since the concept human is subtyped under animate in the type hierarchy. Parallel considerations rule out the noun *rock* as a legitimate argument to *die* since it is not subtyped under animate. One of the concerns mentioned above about how syntactic processes can systematically keep track of which 'selectional features' are entailed and which are not is partially addressed by such lattice-traversal rules as the one presented here.

Event Structure and Lexical Decomposition

The second approach to lexical specification mentioned above is to define constraints internally to the predicate itself. This is commonly called 'lexical decomposition'. We will now review the motivations for decomposition in linguistic theory, and the proposals for encoding lexical knowledge as structured objects. We will then relate this to the way in which verbs can be decomposed in terms of eventualities.

Since the 1960s, lexical semanticists have attempted to formally model the semantic relations between lexical items, such as those between the adjective *dead* and the verbs *die* and *kill* in the sentences below (McCawley, 1968).

(a) John killed Bill.
(b) Bill died.
(c) Bill is dead. (21)

Assuming that the underlying form for a verb like *kill* directly encodes the stative predicate in sentence 21(c) and the relation of causation, generative semanticists have posited representations such as:

$$\text{CAUSE } (x, (\text{BECOME } (\text{NOT } (\text{ALIVE } y)))) \quad (22)$$

Here the predicate CAUSE is represented as a relation between an individual causer x and an expression involving a change of state in the argument y. Carter proposes a similar representation shown here for the causative verb *darken*:

$$x \text{ CAUSE } (\ (y \text{ BE DARK}) \text{ CHANGE}) \quad (23)$$

Although there is an intuition that the cause relation involves a causer and an event, neither Lakoff nor Carter make this commitment explicitly. In fact, it has taken several decades for Davidson's observations regarding the role of events in the determination of verb meaning to find their way convincingly into the major linguistic frameworks. Recently, a new synthesis has emerged which attempts to model verb meanings as complex predicative structures with rich event structures (Pustejovsky, 1991; Hale and Keyser, 1993). This research has developed the idea that the meaning of a verb can be analyzed into a structured representation of the event that the verb designates, and

has furthermore contributed to the realization that verbs may have complex internal event structures. Recent work has converged on the view that any complex event is structured into an inner and an outer event, where the outer event is associated with causation and agency, and the inner event is associated with telicity (completion) and change of state.

Jackendoff (1990) develops an extensive system of what he calls 'conceptual representations', which parallel the syntactic representations of sentences of natural language. These employ a set of canonical predicates, including CAUSE, GO, TO, and ON, and canonical elements, including Thing, Path and Event. Verb meaning is represented by decomposing the predicate into more basic predicates. This work owes an obvious debt to the innovative work within generative semantics illustrated by McCawley's (1968) analysis of the verb *kill*. Recent versions of lexical representations inspired by generative semantics can be seen in the 'lexical relational structures' of Hale and Keyser (1993), where syntactic tree structures are employed to capture the same elements of causation and change of state as in the representations of Carter, Levin and Rapoport Hovav, Jackendoff, and Dowty. The work of Levin and Rappaport Hovav (1995), building on Jackendoff's lexical conceptual structures, has been influential in further elucidating the internal structure of verb meanings.

Pustejovsky (1991) extends the decompositional approach presented in Dowty (1979) by explicitly reifying the events and sub-events in the predicative expressions. Unlike Dowty's treatment of lexical semantics, where the decompositional calculus builds on propositional or predicative units (as discussed above), a 'syntax of event structure' makes explicit reference to quantified events as part of the word meaning. Pustejovsky further introduces a tree structure to represent the temporal ordering and dominance constraints on an event and its sub-events. For example, a predicate such as *build* is associated with a complex event such as that shown below.

$$[_{\text{TRANSITION}}\ [e_1\text{:PROCESS}]\ [e_2\text{:STATE}]\ ] \qquad (24)$$

The process consists of the building activity itself, while the state represents the result of the object being built. Grimshaw (1990) adopts this theory in her work on argument structure, where complex events such as *break* are given a similar representation. In such structures, the process consists of what *x* does to cause the breaking, and the state is the resultant state of the broken item. The process corresponds to the outer causing event, as discussed above, and the state corresponds in part to the inner change-of-state event. Both Pustejovsky and Grimshaw differ from other authors in assuming a specific level of representation for event structure, distinct from the representation of other lexical properties. Furthermore, they follow Higginbotham in adopting an explicit reference to the event place in the verbal semantics.

Recently, Levin and Rappaport Hovov have adopted a large component of the event-structure model for their analysis of the resultative construction in English. Event decomposition has also been applied to properties of adjectival selection, the interpretation of compounds, and stage-and individual-level predication.

Qualia Structure

Thus far, we have focused on the lexical information associated with verb entries. All of the major categories, however, are encoded with syntactic and semantic feature structures that determine their constructional behavior and subsequent meaning at logical form. In 'generative lexicon theory' (Pustejovsky, 1995), it is assumed that word meaning is structured on the basis of four generative factors, or 'qualia roles', that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behavior of lexical items. (These are inspired in large part by Moravcsik's interpretation of Aristotelian *aitia*.) These are: the FORMAL role (the basic category that distinguishes the object within a larger domain); the CONSTITUTIVE role (the relation between an object and its constitutent parts); the TELIC role (its purpose and function); and the AGENTIVE role (factors involved in the object's origin or 'coming into being'). Qualia structure is at the core of the generative properties of the lexicon, since it provides a general strategy for creating new types. For example, consider the properties of nouns such as *rock* and *chair*. These nouns can be distinguished on the basis of semantic criteria, which classify them in terms of general categories such as `natural_kind` and `artifact_object`. Although very useful, this is not sufficient to discriminate semantic types in a way that also accounts for their grammatical behavior. A crucial distinction between *rock* and *chair* concerns the properties which differentiate *natural kinds* from artifacts: functionality plays a crucial role in the process of individuation of artifacts, but not of natural kinds. This is reflected in grammatical behavior, whereby *a good chair* and *enjoy the chair* are well-formed expressions reflecting the

specific purpose for which an artifact is designed, but *good rock* or *enjoy a rock* are semantically ill-formed since for *rock* the functionality (i.e. TELIC) is undefined. Exceptions arise when new concepts are referred to, for example when the object is construed relative to a specific activity, as in '*the climber enjoyed that rock*': *rock* itself takes on a new meaning, by virtue of having telicity associated with it, and this is accomplished by integration with the semantics of the subject NP. Although *chair* and *rock* are both physical_object, they differ in their mode of coming into being (i.e. AGENTIVE): artifacts are man-made, while rocks develop in nature. Similarly, a concept such as 'food' or 'cookie' has a physical manifestation or denotation, but also a functional grounding, pertaining to the relation of 'eating'. These apparently contradictory aspects of a category are orthogonally represented by the qualia structure for that concept, which provides a coherent structuring for different dimensions of meaning.

GLOBAL LEXICAL DESIGN

The view of what computational resources are available to the mental lexicon has changed significantly in recent years. It is now believed that lexical storage and access capabilities are considerably higher than originally imagined by early grammatical theorists. It now seems that combinatorial aspects of grammar that could once only be accounted for by phrase-structure rules and transformations can be handled (in part) by the lexicon directly. This has been a general trend in the generative tradition. The various approaches to lexical encoding can be analyzed using two strategies: precompiling the information into lexical items or forms; and computing or generating new forms or senses during the compositional process. As discussed above, idioms are examples of multiple words of possibly diverse lexical categories behaving as one lexical unit. They thus lend themselves to the first strategy above. Perhaps an even better example of precompiled lexical entries might be noun–noun compounds in English: *pencil sharpener* and *coffee cup* are arguably compositional in nature, but there is enough semantic drift in each form to merit consideration as precompiled lexical entries (Choueka, 1988).

Some linguistic frameworks have adopted this idea as fundamental for the entire compositional operation. The best example of this is CCG. CCG has recently been articulated in enough detail to handle most of the major linguistic phenomena using a library of precompiled lexical types, together with the combinatoric rules of categorial syntax. If the grammar utilizes representations with such nonlocal dependencies, then there must be additional mechanisms for unifying these representations; these are provided in the form of function composition rules and lexical rules.

Lexical rules have also been invoked in head-driven phrase structure grammar to explain the relationship between the various senses for lexical items, from grinding and packaging operations (such as that relating the animal and food senses of *chicken* and *lamb*), to the relation between logically polysemous items, such as *book* and *lecture*, as discussed above. In generative lexicon theory such relations are represented explicitly in the type itself, by means of a typing construction, called 'dot objects', and are disambiguated in context (Johnston, 1995). It is very likely, however, that language makes use of both types of devices: namely, complex types, such as dot objects, and lexical rules. In any case, both types of devices must be seriously constrained by the grammar in order not to overgenerate forms and interpretations.

Independently of the issue of precompiled versus generated forms and senses, there is no question that the mental lexicon is large, containing arguably as many as 350 000 lexical entries. This figure is based on fairly conservative estimates of speaker competence with active and passive vocabularies. For example, an average speaker lexicon might contain 5 000 distinct verbs, 30 000 distinct nominal forms, and 5 000 adjectives. Combine this with an additional 10 000 compound forms and 300 000 distinct proper names. Obviously, the psychological (and hence computational) demands of these classes are quite distinct. There are two parameters that can help us distinguish these classes: the degree of combinatoric (functional) complexity of the lexical item; and whether the lexical item is part of active or passive lexical knowledge. Most closed-class items, for example, are functionally complex, as are many open-class verbs and relational nouns. The majority of the open-class items will also involve a fair amount of information regarding combinatoric possibilities. However, the class of names, although it is by far the largest class of lexical items, is the least demanding in terms of computational resources.

We should view the lexicon neither as a listing of the morphemes in the language, nor as a passive database of items used in other grammatical processes. The lexicon is a dynamic module of grammar, incorporating as well as dictating essential components of syntactic and semantic composition and interpretation. Furthermore, it acts as the

interface to deeper and wider aspects of inference and reasoning.

References

Alsina A (1992) On the argument structure of causatives. *Linguistic Inquiry* **23**(4): 517–555.

Apresjan JD (1973) Regular polysemy. *Linguistics* **142**: 5–32.

Briscoe T, de Paiva V and Copestake A (eds) (1993) *Inheritance, Defaults, and the Lexicon*. Cambridge, UK: Cambridge University Press.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky N (1975) *The Logical Structure of Linguistic Theory*. Chicago, IL: University of Chicago Press. [First published 1955.]

Choueka Y (1988) Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In: *Proceedings of the RAIO*, pp. 609–623. Cambridge, MA: MIT Press.

Davis A (1996) *Lexical Semantics and Linking and the Hierarchical Lexicon*. PhD thesis, Standford University.

Dowty DR (1979) *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Grimshaw J (1990) *Argument Structure*. Cambridge, MA: MIT Press.

Gruber JS (1976) *Lexical Structures in Syntax and Semantics*. Amsterdam: North-Holland.

Hale K and Keyser J (1993) On argument structure and the lexical expression of syntactic relations. In: Hale K and Keyser J (eds) *The View from Building 20*. Cambridge, MA: MIT Press.

Jackendoff R (1972) *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Jackendoff R (1990) *Semantic Structures*. Cambridge, MA: MIT Press.

Johnston M (1995) Semantic underspecification and lexical types: capturing polysemy without lexical rules. In: *Proceedings of ACQUILEX Workshop on Lexical Rules*, August 9–11, 1995, Cambridgeshire.

Levin B and Rappaport Hovov M (1995) Unaccusativity: at the syntax–semantics interface. Cambridge, MA: MIT Press.

Mel'čuk IA (1988) Semantic description of lexical units in an explanatory combinatorial dictionary: basic principles and heuristic criteria. *International Journal of Lexicography* **1**: 165–188.

McCawley J (1968) Lexical insertion in a transformational grammar without deep structure. *Proceedings of the Chicago Linguistic Society* **4**: 71–80.

Pollard C and Sag I (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press, Stanford, CA: CSLI.

Pustejovsky J (1991) The syntax of event structure. *Cognition* **41**: 47–81.

Pustejovsky J (1995) *The Generative Lexicon*. Cambridge, MA: MIT Press.

Pustejovsky J and Boguraev P (1993) Lexical knowledge representation and natural language processing. *Artificial Intelligence* **63**: 193–223.

Sanfilippo A (1993) LKB encoding of lexical knowledge. In: Briscoe T, de Paiva V and Copestake A (eds) *Inheritance, Defaults, and the Lexicon*. Cambridge, UK: Cambridge University Press.

Williams E (1981) Argument structure and morphology. *Linguistic Review* **1**: 81–114.

Further Reading

Baker M (1988) *Incorporation: A Theory of Grammatical Function Changing*. Chicago, IL: University of Chicago Press.

Boguraev B and Briscoe E (1989) *Computational Lexicography for Natural Language Processing*. London: Longman.

Boguraev B and Pustejovsky J (1996) *Corpus Processing for Lexical Acquisition*. Cambridge, MA: Bradford Books/MIT Press.

Bresnan J (1994) Locative inversion and the architecture of universal grammar. *Language* **70**(1): 2–31.

Copestake A and Briscoe E (1992) Lexical operations in a unification-based framework. In: Pustejovsky J and Bergler S (eds) *Lexical Semantics and Knowledge Representation*, pp. 1–4. New York, NY: Springer.

Dowty D (1991) Thematic proto-roles and argument selection. *Language* **67**: 547–619.

Goldberg AE (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press.

Grimshaw J (1979) Complement selection and the lexicon. *Linguistic Inquiry* **10**: 279–326.

Gunter C (1992) *Semantics of Programming Languages*. Cambridge, MA: MIT Press.

Higginbotham J (1985) On semantics. *Linguistic Inquiry* **16**: 547–593.

Higginbotham J (1989) Elucidations of meaning. *Linguistics and Philosophy* **12**: 465–517.

Ingria R (1986) Lexical information for parsing systems: points of convergence and divergence. In: *Proceedings of Workshop on Automating the Lexicon*, Marina di Grosseto, Italy.

Ingria R, Boguraev B and Pustejovsky J (1992) Dictionary/lexicon. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 341–365. New York, NY: Wiley.

Levin B (1993) *Towards a Lexical Organization of English Verbs*. Chicago, IL: University of Chicago Press.

Lyons J (1968) *Introduction to Theoretical Linguistics*. Cambridge, UK: Cambridge University Press.

McCawley JD (1968) The role of semantics in a grammar. In: Bach E and Harms RT (eds) *Universals in Linguistic Theory*, pp. 124–169. New York, NY: Holt, Rinehart, and Winston.

Miller G (ed.) (1990) WordNet: an on-line lexical database. *International Journal of Lexicography* **3**(4): 235–312.

Miller G (1991) *The Science of Words*. New York, NY: Scientific American Library.

Pustejovsky J (1992) Lexical semantics. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 812–819. New York, NY: Wiley.

Pustejovsky J (1998) The semantics of lexical underspecification. *Folia Linguistica* **32**:

Schabes Y, Abeille A and Joshi A (1988) Parsing strategies with leixcalized grammars. In: *Proceedings of the 12th International Conference on Computational Linguistics*. Budapest.

Steedman M (1997) *Surface Structure Interpretation* Cambridge, MA: MIT Press.

Tenny C and Pustejovsky J (2000) *Events as Grammatical Objects*. Stanford, CA: CSLI/Chicago, IL: University of Chicago Press.

Weinreich U (1972) *Explorations in Semantic Theory*. The Hague: Mouton.

Linguistic Evidence, Status of

Intermediate article

Carson T Schütze, University of California, Los Angeles, California, USA

CONTENTS

Linguistics concerns itself in principle with any kind of evidence that seems capable of informing us about the human language faculty, though individual researchers may find certain kinds of evidence more fruitful for their goals than others. A survey of the major sources of evidence used by linguists leads to an assessment of their relative merits and consideration of the kinds of explanations of linguistic phenomena this evidence can lead to.

INTRODUCTION

The evidence used by linguists can be roughly divided into three categories: language data that exist independent of the linguist's desire to study then (corpora); language behavior induced with the specific intent of studying knowledge of language (behavioral experiments); and unconscious reflections of linguistic knowledge or processes (brain measures). Within the second category there is in practice a dichotomy between grammaticality judgments, which constitute the everyday staple of data, and controlled experimentation in the psychological tradition. Since no single kind of evidence is ideal in all respects, an efficacious approach to linguistic investigation is to seek convergence from a wide array of techniques. The discussion that follows chiefly concerns evidence relevant to theories of linguistic competence rather than, for example, evidence concerning how people parse sentences. Points are illustrated mainly with examples from syntax, but most can be easily translated to other levels (e.g. phonology, morphology, semantics, pragmatics). (*See* **Syntax**; **Phonology**; **Morphology**; **Categorial Grammar and Formal Semantics**; **Pragmatics, Formal**)

Any discussion on the status of linguistic evidence would be incomplete without noting that much of it is disappearing: at least 20 per cent and perhaps as many as 50 per cent of the languages spoken in the world as of the year 2000 are not being learned by children and will thus cease to be spoken at all by 2100, unless dramatic changes occur (Krauss, 1996). This state of affairs has led to the suggestion that the field of linguistics should devote greater attention to documenting and helping to preserve endangered languages. This undertaking typically includes fieldwork with native speakers, for instance by elicitation, a specialized interview-style technique (e.g. Munro, 2000).

CORPUS DATA

Corpus data can be characterized as (a collection of) pre-existing samples of language, that is, language that was (or could have been) originally produced for some purpose other than directly answering linguists' questions.

Benefits

Corpus evidence has the following benefits. The contents of corpora cannot have been biased by the researcher. They may bring to our attention phenomena that would not otherwise have come to mind as relevant. Because they typically contain extended passages of continuous language, they allow for study of phenomena not testable in sentences in isolation, for example, how referents are identified throughout a narrative or discourse. They include records of languages no longer spoken. The ability to search and calculate statistics by computer makes detailed quantitative analysis possible. (Ideally, one would like to search both for strings of words and for particular structural configurations of phrase types; this is beginning to be possible.) Finally, they may provide samples of language as produced when the speaker/writer was not consciously reflecting on the form of the utterance. This last point is nontrivial: Labov (1975) has documented that speakers who denied that they would ever produce a given construction

went on to do so spontaneously during the very same interview.

Increasingly, corpus data are being used to show that constructions that tend to sound artificial and awkward when presented as examples in linguistics articles nonetheless occur and sound natural in everyday language situations. A further advantage to corpora of spontaneous language use is reflected particularly in studies of acquisition and speech production: one can find 'errors' in such corpora, whether they be systematic productions by children that are not possible for adult speakers, or spontaneous speech errors (slips of the tongue). (*See* **Discourse Processing**; **Statistical Methods**; **Syntactic Form Frequency: Assessing**; **Language Acquisition**; **Speech Production**)

Drawbacks

There are concomitantly certain limitations to corpus data. Most importantly, the absence of some phenomenon from a corpus is hard to interpret – its nonoccurrence may have been accidental or systematic. There is thus an asymmetry in corpus findings. The presence of some phenomenon indicates that it is part of the language, at least for one speaker/writer (although there is no standard for how frequently a structure must occur before it is considered attested), and this can be a valuable tool in theorizing: if one finds in real text systematic violations of, say, island constraints, this could demand some change in the theory. On the other hand, if one wishes to confirm the impossibility of some structure, failure to find that structure in a corpus constitutes only weak evidence for ungrammaticality, even if one can argue that sufficient opportunities arose for it to appear. Additional difficulties arise in interpreting written corpora. In the composing or editing of text it is possible to make piecemeal alterations to a sentence without being forced to re-read it to verify that it still hangs together as intended (as well as, of course, making typos). Editing may also introduce structures that would not be producible in real time, such as complex embeddings (cf. sentence (4) below), and especially in old texts we must allow for the possibility that systematic 'corrections' and random errors may have been introduced during reproduction. (*See* **Constraints on Movement**)

GRAMMATICALITY JUDGMENTS

So-called grammaticality judgments involve explicitly asking speakers whether some particular string of words is a possible utterance of their language, with an intended interpretation either implied or explicitly stated. The term is misleading in at least two respects (Schütze, 1996). First, since a grammar is a mental construct not accessible to conscious awareness (or a formal model of that construct), speakers cannot have any intuitions about the status of a sentence with respect to that grammar; rather, in Chomsky's (1965) terms one might say their intuitions are about acceptability. Second, they are not judgments, inasmuch as they need not involve the speaker reasoning consciously about an appropriate answer (as opposed to, say, judging whether a defendant is guilty of some crime). Thus, a more accurate and less loaded term might be 'acceptability reactions', but for familiarity's sake I continue to use the traditional label. Grammaticality judgments are potentially the richest sort of information about knowledge of language, but at the same time they are the kind most susceptible to distortion.

A simple judgment that a string is a possible sentence, with no mention of what it would mean, is virtually worthless. For this reason, whenever there could be confusion, the presentation of judgment data typically includes an indication of crucial interpretive features, for example the use of subscripts to indicate intended co-reference relationships in (1):

$$\text{John}_i \text{ believes himself}_i \text{ to be in danger.} \qquad (1)$$

In the general case it can also be important to know what prosody the utterance is meant to have, and possibly the linguistic or situational context; both of these considerations are especially important in judging topic and focus structures, for example. (It is sometimes claimed that syntactic well-formedness can be judged without one being able to identify what a sentence means, as in Chomsky's celebrated example 'Colorless green ideas sleep furiously', or Lewis Carroll's poem *Jabberwocky*. This claim is misleading, however. One actually knows perfectly well what the sentences *should* mean; in the former case, some nonliteral interpretations of content words are required for the meaning to be coherent, and in the latter case one would be able to determine the meaning of the sentence upon receiving the meanings of the nonsense content words.) (*See* **Anaphora**; **Prosody**)

Benefits

Grammaticality judgments have the following benefits. They provide evidence about the status of phenomena that occur so rarely in spontaneous

language use that we could not learn about them from studying corpora; in particular, they distinguish possible from impossible utterances among those that have never been naturally produced. Establishing which utterances are impossible has become increasingly important as the development of restrictive theories of possible human languages has become a priority, along with the observation that language learners apparently do not receive systematic negative evidence, so that their knowledge of what is impossible demands nontrivial explanation. The opposite situation also arises: rare and scattered attestations of some structure could be artefacts of the production situation and might not be felt to be appropriate upon reflection, even by the very speaker who produced them. Grammaticality judgments sometimes demonstrate knowledge of language in speakers whose behavior on other tasks does not evince the same degree of knowledge (e.g. Linebarger *et al*., 1983); this has also been observed anecdotally in the fact that children will sometimes reject an adult's faithful repetition of an utterance they have just produced, apparently realizing that their own productions are not hitting the intended target (e.g. Berko and Brown, 1960).

Drawbacks

Turning to drawbacks of grammaticality judgment data, the most systematic source of confusion is the semantic content of a sentence – it can be difficult to separate a well-formed expression of incoherent content from an ungrammatical sentence. Young children often respond to truth-value when asked to judge well-formedness (e.g. Hakes, 1980). Another common source is parseability: it can be hard to distinguish parsing difficulty from grammatical ill-formedness, for example, for clausal embeddings like that in (2), claimed to be strictly ungrammatical by Koster (1978) but possible by Delahunty (1983):

If [that John likes Mary] surprises you, you obviously haven't been paying much attention. (2)

An example of the opposite kind, where the parser seems to make a structure spuriously sound grammatical, is the comparison between (3) and (4):

The patient that the nurse the clinic had hired met Jack. (3)

The patient that the nurse the clinic had hired admitted met Jack. (4)

Sentence (3) often seems grammatical when considered next to (4), although in fact only (4) is well formed (Frazier, 1985, attributed to Janet Fodor). An example whose explanation is less obvious is (5):

More people have been to Russia than I have. (5)

Sentence (5) often sounds fine on initial presentation, a feeling that changes after it is pointed out that this sentence has no possible meaning.

Although there are strings of words that seem unequivocally completely well formed (e.g. 'Mary saw John') and strings that seem completely ill formed (e.g. 'Fred the the this'), most strings of theoretical interest lie in a vast gray area in between these two extremes. This means that theories are necessarily concerned with distinctions among strings in the gray area; for this reason, simply eliciting yes/no grammaticality judgments is of limited use. Two solutions are to elicit responses on a multi-valued scale (e.g. 1–7, as is typical in psychology), or to elicit relative judgments that compare minimal pairs of sentences. The notation used to indicate the status of sentences has evolved to allow numerous gradations of well-formedness to be indicated, but it should be kept in mind that even so, the *relative* status of examples is primarily what matters; the same notation does not always convey the same absolute level of well-formedness. The most common symbolization, marked at the beginning of a sentence, is as follows, from most to least well formed: unmarked, (?), ?, ??, ?*, *, **. These symbols are combined in a nonobvious way with the use of parentheses to indicate optionality: symbols inside parentheses mark the status of a sentence with the optional element present, as in (6), while symbols immediately preceding a parenthesis indicate the status of omitting the parenthesized element, as in (7):

This sentence (*might) has a problem with verbs. (6)

This sentence requires *(a) determiner. (7)

A possible objection to the use of grammaticality judgments that has often been raised is that their metalinguistic nature makes them artificial and hence undermines their external validity; they may, the objection goes, have little to do with how people actually use language. One response is to suggest that similar metalinguistic behaviors are part of everyday language use, for instance, in forming opinions about people based on how they

talk. Another response is simply to note that to the extent that grammaticality judgment data are systematic, it is highly unlikely that their underlying cognitive source would be something unrelated to grammar – the existence of a separate system that duplicates many aspects of language use but is invoked only for metalinguistic judgment seems highly improbable.

Because there are many ways in which one's judgments can be misled, the question has arisen whether it is possible to obtain useful judgments on subtle issues from naive speakers with no linguistic training. In some cases it is difficult to replicate relatively agreed-upon judgments of linguists while testing naive subjects (e.g. Gordon and Hendrick, 1997). Partly for this reason, and partly for sheer convenience, linguists rely increasingly on other linguists for judgment data on some languages. It is not clear whether data obtained from linguists should be taken as in any way more desirable (cf. Spencer, 1973). Valian (1982) makes a case in favor of using such 'expert' judgments, based on analogy to another domain: she notes that wine tasting, for example, relies on the acquired ability to detect subtle distinctions that inexperienced wine drinkers simply cannot make. One practice that is clearly undesirable, however, is for investigators to use their own judgments as the primary basis for theorizing. There is no reason to believe that people with a stake in the outcome of such judgments can remain unaffected by their theoretical stance.

Interpretation of Judgment Data

Some researchers have raised the possibility that gradience in our judgments of sentences implies that grammar itself does not make categorical distinctions between the possible and the impossible. This does not follow, however. For one thing, we know that humans can give systematic gradient judgments about virtually any phenomenon, including ones such as *even number* whose only meaning is formally defined as categorical (Armstrong *et al*., 1983; Barsalou, 1987). This observation does not entail that our knowledge of mathematics fails to make a perfectly sharp distinction between even and odd numbers. Rather, our judgments can evidently be sensitive to factors other than our underlying competence. Moreover, even within the grammar it is possible that numerous separate components are implicated in the structuring and interpretation of a sentence (e.g. the modules of Government–Binding theory); one might expect that a violation in any single component should lead to a lesser degree of overall deviance than violations at multiple levels. (*See* **Government–Binding Theory**)

The nature of grammaticality judgments has sometimes been characterized with reference to traditions in psychology that might lead one to worry about the wisdom of using them as evidence in linguistics: they have been described as 'introspective' judgments or 'intuitions', but neither of these terms applies accurately (Carr, 1990; Schütze, 1996). They have more in common with the responses in psychophysics experiments, having the character of sensations, reactions to stimuli, or reports about mental states. There is also no basis for the occasionally encountered view that grammaticality judgments are somehow related to competence while spontaneous production and comprehension data are related to performance; all are behaviors, hence performance data, but all can in principle bear on competence. (*See* **Introspection**; **Performance and Competence**)

EXPERIMENTAL DATA

It has sometimes been suggested that claims made exclusively on the basis of grammaticality judgment data do not necessarily bear on how the human language faculty is actually constructed unless their 'psychological reality' has been tested by means of some other sort of data, typically a formal experiment. This view belies a misunderstanding: grammaticality judgments are, as noted above, themselves data about human behavior that need to be accounted for; they are not intrinsically less informative than, say, reaction-time measures (one might argue that they are more informative). The elicitation of grammaticality judgments is itself a behavioral experiment on a speaker, albeit one whose generally casual nature may leave it susceptible to certain kinds of misinterpretation. (These can be due, for example, to the limited number of types and tokens of items tested, the small number of participants questioned, lack of randomized order of presentation, etc.)

Behavioral Measures

The most direct behavioral experiment one can conduct to assess the status of sentences is essentially a more rigorous version of grammaticality judgment collection, typically carried out using a written questionnaire with a multi-valued response scale. A special variant is the speeded grammaticality judgment task, which requires participants to

respond with a yes/no button-press decision within a very short time, and is thereby intended to get at their earliest reactions. (The appropriateness of this task for assessing grammar is debatable; however, it can be useful in the study of sentence processing.) Beyond this and similar types of experiments that ask participants for their reactions to sentences directly, a large variety of experimental paradigms used in psychology can be employed to assess participants' knowledge of language in more indirect ways. These generally involve a trade-off between reducing the influence of conscious reflection about language and making response measures more difficult to interpret *vis-à-vis* linguistic competence. (*See* **Sentence Processing**)

Brain Measures

Recent years have shown the promise that we might learn about details of linguistic knowledge by direct measurements from the brain, in some cases bypassing the need for participants to perform any task other than simply comprehending or producing language. Techniques used in this way have included event-related potentials (ERP), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG). The trade-off for not requiring conscious attention to an additional task is, of course, the difficulty of interpreting results of such experiments. Even a relatively fine-grained characterization of which parts of the brain are involved in processing certain aspects of linguistic stimuli is not terribly helpful for uncovering the particulars of linguistic competence. Dissociations of the sort obtained from lesion studies can be informative with respect to the architecture of the grammar, but are not especially so with respect to its contents. More provocative have been suggestions (mostly in the ERP literature) that the grammaticality status of certain classes of utterance types might be reflected in characteristic patterns of activity (e.g. Neville *et al*., 1991). For example, it has been claimed that syntactic versus semantic anomalies have distinct signatures, and that violations of different grammatical constraints pattern differently. Interpretation of these claims is still open to debate, however, in that it is not established whether a full range of superficially diverse sentence types that by hypothesis share a common grammatical property (e.g. violating Subjacency) have a detectable commonality in brain signals; if it turns out that each kind of ungrammaticality produces a different ERP pattern, the value of these methods for theoretical purposes may be limited. (*See* **Neuroimaging**; **Syntax and Semantics, Neural Basis of**; **Constraints on Movement**)

NATURE OF EXPLANATION IN LINGUISTIC THEORY

A central goal in linguistic theory is to explain numerous disparate language phenomena using as few grammatical principles as possible, and in turn to seek to ground the grammar in terms of properties of the mind/brain.

Phenomena To Be Explained

Linguistics as a whole seeks to explain much more than simply adult linguistic competence, though this remains the focus of a large amount of research. Increasingly, additional phenomena are part of the enterprise, as potential sources of evidence about the nature of linguistic knowledge and as phenomena that theories of grammar could in turn be used to help explain. The aforementioned three kinds of data sources (corpora, judgments, and experiments) largely cross-cut these areas of investigation. The kinds of behaviors studied include first and second language acquisition, the typology of existing and nonexisting languages, patterns of historical change and creolization, multilingualism and intrasentential code-switching, language disorders in adults and children, and speech errors. Each of these has been hypothesized to be characterizable to some degree by grammar. (*See* **Language Acquisition**; **Second Language Acquisition**; **Typology**; **Language Disorders**; **Speech Error Models of Language Production**)

Kinds of Explanations

There are essentially two kinds of answers to the question of why human languages are the way they are: (1) given other facts about the nature of human beings, they could not be any other way; and (2) they are accidents of history that could just as easily have come out some other way. Most likely, each of these answers is correct for some aspects of language, but we have little idea which ones are which.

Learning- and processing-based explanations

The possible human languages may be constrained by limits on the procedure by which language is acquired – that is, some hypothetically possible languages fail ever to come to be spoken because

a child would never reach them as the target of acquisition (even if such a language were being spoken around the child), because the finite amount of evidence available to the child would be consistent with other, less marked grammars, or because the acquisition system affords no learning path that would reach the target. The nonoccurrence of other types of languages may be explained because those languages would not be processable to a degree sufficient to express common ideas. Lack of processability might in turn be explained either by the inherent architecture of the parsing mechanism or by other (not purely linguistic) cognitive limits such as the capacity of short-term memory. Results from theories of computation and learnability are thus relevant to determining the form of human linguistic knowledge. If we can formally prove that a system lacking property X would be unlearnable, not parseable in reasonable time, etc., then human language must have X (e.g. Wexler and Culicover, 1980; Berwick and Weinberg, 1984). (*See* **Sentence Processing: Mechanisms**; **Sentence Comprehension, Linguistic Complexity in**; **Computability and Computational Complexity**)

Historical and biological accidents

There are certainly properties of particular languages whose only explanation lies in the course of human history; for example, the fact that English has a huge Latinate vocabulary with distinctive properties is traceable to the Norman conquest. But many linguists believe that there are limits on what a language *could be* like, regardless of the course of human events, limits not of the sort discussed in the previous paragraph, but rather, strictly properties of the language competence system, imposed by human biology in its determination of brain structure, often referred to as Universal Grammar (UG). In what sense can we explain facts about languages by hypothesizing that they are properties of UG? The utility of appeals to UG would seem to depend on a notion that some kinds of properties are of the right character (say, being constructed from sufficiently simple primitives) to plausibly be part of the human innate genetic endowment. Unfortunately, we have almost no idea yet what this character really is, so for now, successful explanations in linguistics mostly involve reduction of numerous, superficially unrelated facts within and across languages to fewer principles that are consistent with what is known about all languages. (*See* **Innateness and Universal Grammar**)

Even more speculatively, we might seek to explain why the genetic underpinning of UG is the way it is. While having a system of communication obviously has evolutionary benefits, many of the details of how human language works do not seem to carry adaptive advantages. For this reason some researchers suspect that UG has 'piggy-backed' on other systems, where the counterparts of these properties might have made a difference. There remains the possibility that some aspects of UG are the result of random evolutionary events that played no role in natural selection, in which case they might have no explanation in any interesting sense.

Individual Differences

In addition to explaining what is apparently common to every human's language faculty, a problem linguists face every day is how to explain apparently conflicting evidence about grammars, most commonly, situations where different speakers of 'the same language' or even 'the same dialect' report different grammaticality judgments for identical stimuli. On the one hand, from a cognitive perspective we are investigating what is in the mind/brain of each individual (I-language), rather than some purported object out in the world called a language (E-language), so there would be no contradiction in finding that no two speakers have identical grammars. In this sense, *idiolects* are expected to exist.

On the other hand, it is worrying if too many such differences are found, because it might be a sign that the judgments are reflecting, in heterogeneous ways, individual differences *outside* the grammar. The conundrum arises because any data collected from a single individual are liable to random interference or misinterpretation; the best way to eliminate such complications is to consider multiple speakers and hope for convergence. Still, an important emerging finding stressed by Cowart (1997) is that while individuals may differ widely in how they use any given response scale, so that a particular sentence might be rated 5/7 by one participant and 3/7 by another, for instance, there is much less individual variation in the *patterns* of responses, that is, how sentences rate relative to one another. This is the reason why linguists are primarily concerned with contrasts, say between a sentence marked '??' and one marked '*', but not about how good or bad '??' really is in any absolute sense. The more we learn based on patterns agreed upon by substantial numbers of speakers, the more

we may hope to understand what is and is not plausibly a locus of cross-speaker grammatical variation, and therefore to become more secure in our interpretation of individual differences. (*See* **Individual Differences**)

References

Armstrong SL, Gleitman LR and Gleitman H (1983) What some concepts might not be. *Cognition* **13**: 263–308.

Barsalou LW (1987) The instability of graded structure: implications for the nature of concepts. In: Neisser U (ed.) *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pp. 101–140. Cambridge, UK: Cambridge University Press.

Berko J and Brown R (1960) Psycholinguistic research methods. In: Mussen PH (ed.) *Handbook of Research Methods in Child Development*, pp. 517–557. New York, NY: Wiley.

Berwick RC and Weinberg AS (1984) *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. Cambridge, MA: MIT Press.

Carr P (1990) *Linguistic Realities: An Autonomist Metatheory for the Generative Enterprise*. Cambridge, UK: Cambridge University Press.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cowart W (1997) *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage.

Delahunty GP (1983) But sentential subjects do exist. *Linguistic Analysis* **12**: 379–398.

Frazier L (1985) Syntactic complexity. In: Dowty DR, Karttunen L and Zwicky AM (eds) *Natural Language Processing: Psychological, Computational, and Theoretical Perspectives*, pp. 129–189. Cambridge, UK: Cambridge University Press.

Gordon PC and Hendrick R (1997) Intuitive knowledge of linguistic co-reference. *Cognition* **62**: 325–370.

Hakes DT (1980) *The Development of Metalinguistic Abilities in Children*. New York, NY: Springer.

Koster J (1978) Why subject sentences don't exist. In: Keyser SJ (ed.) *Recent Transformational Studies in European Languages*, pp. 53–64. Cambridge, MA: MIT Press.

Krauss M (1996) Linguistics and biology: threatened linguistic and biological diversity compared. In: McNair L, Singer K, Dobrin LM and AuCoin MM (eds) *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*, pp. 69–75. Chicago, IL: Chicago Linguistic Society.

Labov W (1975) *What Is a Linguistic Fact?* Lisse: Peter de Ridder. [Also published as (1975) Empirical foundations of linguistic theory. In: Austerlitz R (ed.) *The Scope of American Linguistics*, pp. 77–133. Lisse, Netherlands: Peter de Ridder.]

Linebarger MC, Schwartz MF and Saffran EM (1983) Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* **13**: 361–392.

Munro P (2000) Field linguistics. In: Aronoff M and Rees-Miller J (eds) *The Handbook of Linguistics*, pp. 130–149. Malden, MA: Blackwell.

Neville H, Nicol JL, Barss A, Forster KI and Garrett MF (1991) Syntactically-based sentence processing classes: evidence from event-related brain potentials. *Journal of Cognitive Neuroscience* **3**: 151–165.

Schütze CT (1996) *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago, IL: University of Chicago Press.

Spencer NJ (1973) Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* **2**: 83–98.

Valian V (1982) Psycholinguistic experiment and linguistic intuition. In: Simon TW and Scholes RJ (eds) *Language, Mind, and Brain*, pp. 179–188. Hillsdale, NJ: Lawrence Erlbaum.

Wexler K and Culicover PW (1980) *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

Further Reading

Birdsong D (1989) *Metalinguistic Performance and Interlinguistic Competence*. New York, NY: Springer.

Botha RP (1981) *The Conduct of Linguistic Inquiry: A Systematic Introduction to the Methodology of Generative Grammar*. The Hague, Netherlands: Mouton.

Chomsky N (1961) Some methodological remarks on generative grammar. *Word* **17**: 219–239.

Fillmore CJ, Kempler D and Wang WS-Y (eds) (1979) *Individual Differences in Language Ability and Language Behavior*. New York, NY: Academic Press.

Gerken LA and Bever TG (1986) Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science* **10**: 457–476.

Greenbaum S (1988) *Good English and the Grammarian*. London, UK: Longman.

Levelt WJM (1974) *Formal Grammars in Linguistics and Psycholinguistics*, 3 vols. The Hague, Netherlands: Mouton.

McNair L, Singer K, Dobrin LM and AuCoin MM (eds) (1996) *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*. Chicago, IL: Chicago Linguistic Society.

Newmeyer FJ (1983) *Grammatical Theory, its Limits and its Possibilities*. Chicago, IL: University of Chicago Press.

Perry TA (ed.) (1979) *Evidence and Argumentation in Linguistics*. Berlin, Germany: de Gruyter.

Local Dependencies and Word-order Variation

Intermediate article

Gereon Müller, Institut für Deutsche Sprache, Mannheim, Germany

CONTENTS

Certain syntactic displacement operations target clause-internal positions. These movement types are typically clause-bound, and they are responsible for word order variation within and across languages.

INTRODUCTION

Certain syntactic displacement operations target clause-internal (IP-internal, where IP = inflection phrase) positions; among them are noun phrase (NP)-movement, scrambling, pronoun movement, and extraposition. A conspicuous common property of these movement types is that they are clause-bound: they cannot cross a complementizer phrase (CP) and target an IP-internal position in a higher clause. Because of this strict locality property, it is more difficult to establish the existence of a syntactic movement operation (and a trace) than in the case of movement types that target IP-external positions and are not clause-bound (such as *wh*-movement to SpecC). Accordingly, analyses that do without syntactic displacement have been proposed for all local dependencies addressed below (see, e.g., Williams (1994) on NP-movement, Fanselow (2001) on scrambling, and Culicover and Rochemont (1990) on extraposition). In what follows, a movement analysis will nevertheless be presupposed throughout – first, because there is a growing body of empirical evidence in support of this view (see the Further Reading); second, because a displacement analysis is forced under a general, conceptually attractive assumption, the *Uniformity of Theta Assignment Hypothesis* (UTAH; Baker, 1988) according to which identical thematic relationships between items must be represented by similar structural relationships between those items at deep structure (D-structure). (*See* **Phrase Structure and X-bar Theory**; **Constraints on Movement**; **Government–Binding Theory**)

NOUN PHRASE MOVEMENT

Passive

The core property of passive constructions is that the external argument of a verb cannot be realized as an NP in the subject position SpecI (specifier of the IP). This argument reduction effect typically goes hand in hand with a morphological reflex (e.g. special passive morphology on the verb, presence of a passive auxiliary). In some languages (e.g. in Ukrainian), this is all there is to say; a remaining internal argument receives objective Case, and SpecI can remain empty. However, in many languages (e.g. in English), argument reduction is accompanied by Case absorption – a passivized verb cannot assign objective Case any more. An internal argument that receives objective Case in active sentences (see (1a)) is moved to the subject position SpecI in passive sentences (see (1b)), where it is assigned nominative Case. This operation is called NP-movement. Chomsky (1981) argues that NP-movement is possible in (1b) because SpecI is not a Theta-position (so that a general ban on movement into Theta-positions is respected); and it is necessary because NP_2 would otherwise violate the Case Filter (that demands that every NP is assigned Case).

a. $[_{IP}$ John$_1$ I$[_{VP}$ kissed Mary$_2]]$
b. $[_{IP}$ Mary$_2$ was$[_{VP}$ kissed $t_2]$ (by John)$]$ (1)

The thematic relations between the two arguments are identical in (1a, b). Hence, the UTAH not only implies that the argument bearing the Theta-role Theme (NP_2) is base-generated in VP in (1b); it also requires a syntactic representation of the argument bearing the Theta-role Agent in this sentence. Proposals as to what acts as the external argument in (1b) include the passive morphology itself, the

by-phrase, and various kinds of empty categories (pro, PRO).

Some languages (e.g. German) behave like Ukrainian in that NP-movement is not required, and like English in that an internal argument is assigned nominative rather than objective Case. The question arises of which Case is absorbed by passivization in double object constructions. In English, it is normally the object Case assigned to the first NP in double object constructions that is absorbed by passivization (see (2a, b)); but there is considerable variation in this domain, and even closely related languages (like Norwegian and German) may behave differently.

a. $[_{IP}$ Mary$_1$ was $[_{VP}$ given t_1 a book$_2$]]
b. $^*[_{IP}$ A book$_2$ was $[_{VP}$ given Mary$_1$ t_2]] (2)

Raising to Subject, Exceptional Case Marking, and Control

NP-movement is also involved in raising constructions like (3).

$[_{IP}$ John$_1$ $[_{VP}$ seems $[_{IP}$ t_1 to be a fool]]] (3)

The matrix predicate *seem* shares two properties with passivized verbs: it does not take an external argument NP, and it does not assign objective Case. Given that the SpecI position of an infinitive is not assigned nominative Case by nonfinite I in English, NP$_1$ in (3) can and must move to the matrix SpecI position, where it is assigned nominative Case by finite I. Raising to subject must be distinguished from two related constructions. *Exceptional Case Marking* (ECM) differs from raising in that the matrix verb takes an external argument (hence, raising is not possible), and in that the matrix verb 'exceptionally' assigns objective Case to an embedded subject that it does not Theta-mark (hence, raising is not required); see (4):

$[_{IP}$ Mary$_2$ $[_{VP}$ believes $[_{IP}$ John$_1$ to be a fool]]] (4)

It seems that a verb's ability to assign objective Case and the presence of an external NP argument go hand in hand; this observation is known as 'Burzio's generalization'.

Raising and ECM constructions have in common that the infinitive is transparent (for movement and Case assignment, respectively). This is often accounted for by assuming that raising and ECM infinitives possess less structure than other clauses: they are bare IPs, not CPs. A CP destroys the transparency of an infinitive, whereas an IP does not. Accordingly, exceptional Case assignment is impossible in *control* constructions like (5) (where the matrix verb takes an external NP argument) if we assume that a CP projection with a phonologically empty complementizer C is present.

$[_{IP}$ Mary$_1$ $[_{VP}$ tries $[_{CP}$ C $[_{IP}$ PRO$_1$ to work hard]]]] (5)

The embedded CP is a *barrier* for exceptional Case assignment. Hence, the subject of the embedded infinitive cannot be realized as an overt NP (which would violate the Case Filter). A possibility that is often entertained is that the infinitive's external argument is nevertheless realized syntactically, albeit as an empty category PRO; PRO is confined to ungoverned positions. (Chomsky (1981) suggests that this can be derived from independent assumptions; hence, the restriction on PRO is sometimes referred to as the *PRO theorem*.)

More on Control

PRO is co-indexed with the matrix subject (which acts as its controller) in (5) and (6a), but depending on lexical and structural factors, PRO may also be co-indexed with a matrix object (as in (6b)), or may receive an arbitrary, generic, or discourse-based interpretation (as in (6c, d)).

a. $[_{IP}$ Mary$_1$ $[_{VP}$ promised John$_2$ $[_{CP}$ C $[_{IP}$ PRO$_{1/*2}$ to leave]]]]
b. $[_{IP}$ Mary$_1$ $[_{VP}$ persuaded John$_2$ $[_{CP}$ C $[_{IP}$ PRO$_{*1/2}$ to leave]]]]
c. $[_{IP}$ $[_{CP}$ C $[_{IP}$ PRO$_x$ To behave oneself in public]] would help John]
d. $[_{IP}$ John$_1$ does not know $[_{CP}$ C $[_{IP}$ how PRO$_x$ to prove the theorem]]] (6)

Examples like (5) and (6a, b) are often said to involve *obligatory control* (OC); examples like (6c, d) show *non-obligatory control* (NOC). It has been argued that the theory of control can be based on the assumption that PRO is an anaphor that obeys an appropriately revised version of principle A of the binding theory (Manzini, 1983; Koster, 1987): PRO must be bound within its binding domain if it has one. The next-higher clause is the binding domain in OC contexts; however, there is no binding domain for a PRO in subject clauses and *wh*-clauses, which are therefore NOC contexts. A complication is that control tends to be even more local than predicted by the binding theory: with the exception of certain verbs like 'promise' (see (6a)), the interpretation of PRO follows a *Minimum Distance Principle* (MDP) according to which PRO picks the minimally c-commanding NP as its antecedent in OC contexts. (*See* **Anaphora**; **Binding Theory**)

Another kind of approach is developed by Hornstein (2001). He argues that the locality of control in OC contexts results from the fact that movement rather than anaphoric binding is involved. Under this view, PRO is a trace of NP-movement, and control and raising are similar after all, the main difference being that movement is to a Theta-position in OC, but not in raising contexts. (Note that this analysis is incompatible with the ban on movement into Theta-positions mentioned above.) MDP effects are traced back to the *Minimal Link Condition* (MLC), which is independently known to restrict movement (Chomsky, 1995). Furthermore, NOC contexts are assumed to involve an empty pronominal (pro) that is inserted as a last resort operation in syntactic environments that block movement.

At present, it is an open question to what extent such structural explanations (based on binding or movement) can succeed in accounting for the varieties of control, and to what extent non-structural properties (such as the thematic properties of matrix predicates) play a role (Culicover and Jackendoff, 2001).

General Properties

NP-movement has several characteristic properties, some of which distinguish it from movement types that target an IP-external position, such as *wh*-movement. First, like *wh*-movement, NP-movement obeys the *Relativized Minimality* constraint (Rizzi, 1990), according to which α-movement must not cross an intervening α-position. Thus, NP-movement must be successive-cyclic if a SpecI position intervenes between the base position and the target SpecI position (see (7a, b)); note that a passivized ECM verb behaves in every respect like an ordinary raising verb).

a. *A man$_1$ seems/is believed [$_{IP}$ there to have been kissed t$_1$]
b. A man$_1$ seems/is believed [$_{IP}$ t$_1'$ to have been kissed t$_1$] (7)

However, in contrast to *wh*-movement, NP-movement is clause-bound: it cannot cross a CP. Thus, NP-movement from a finite clause (*super-raising*) is impossible in English (see (8a)), whereas *wh*-movement is not (see (8b)):

a. *[$_{IP}$ Mary$_1$ seems [$_{CP}$ (t$_1'$) that [$_{IP}$ t$_1$ likes John]]]
b. [$_{CP}$ Who$_1$ do you think [$_{CP}$ t$_1'$ that [$_{IP}$ Mary likes t$_1$]]]? (8)

A second difference concerns *binding of anaphors* (i.e. reflexive and reciprocal pronouns). NP-movement creates new possibilities for anaphoric binding (see (9a)), whereas *wh*-movement does not (see (9b)).

a. [$_{IP}$ The students$_1$ seem [$_{PP}$ to each other$_1$] [$_{IP}$ t$_1$ to be intelligent]]
b. *[$_{CP}$ Which man$_1$ does himself$_1$ think [$_{CP}$ t$_1'$ that Mary likes t$_1$]]? (9)

Third, there is a difference with respect to *weak crossover*. In a weak crossover configuration, a quantified NP has been moved across a co-indexed personal pronoun that does not c-command the trace of NP. This configuration leads to ungrammaticality with *wh*-movement, but not with NP-movement; see (10a, b):

a. [$_{IP}$ Every girl$_1$ seems [$_{PP}$ to her$_1$ mother] [$_{IP}$ t$_1$ to be intelligent]]
b. *[$_{CP}$ Which girl$_1$ does [$_{NP}$ her$_1$ mother] think [$_{CP}$ t$_1'$ that John likes t$_1$]]? (10)

Fourth, NP- and *wh*-movement diverge with respect to *parasitic gaps*. A parasitic gap is a trace (noted here as 'e') in a position that is typically not accessible to regular movement (because of locality constraints). However, e is permitted if it is in a sufficiently local relation with a well-formed movement chain with the same index. As shown in (11a, b), NP-movement does not license parasitic gaps in English, whereas *wh*-movement does (Chomsky (1982), who also notes that (11a) is grammatical with *them* in place of e, which implies that the construction is not ill formed because of a control failure with the PRO subject of the *without* phrase).

a. *[$_{IP}$ The books$_1$ can be sold t$_1$ [$_{CP}$ without reading e$_1$]]
b. [$_{CP}$ Which books$_1$ did they sell t$_1$ [$_{CP}$ without reading e$_1$]]? (11)

Further Dissociation of Case-positions and Theta-positions

In NP-movement constructions, the first member of the movement chain is in a position to which Case is assigned, and the last member occupies a Theta-position. NP-movement has so far been motivated on the basis of passive and raising (also recall the above remarks on an NP-movement approach to control), but it has also been argued to underlie other constructions.

The VP-internal subject hypothesis

The assumption that an external argument of a verb is base-generated in the verb phrase (VP)-external SpecI position has been called into question in recent years. The alternative (suggested by Sportiche (1988) and many others, and now widely adopted) is that all arguments of V are base-generated in a VP-internal position; this is known as the *VP-Internal Subject Hypothesis* (VISH). (Since this assumption is usually taken to hold for all predicates, a more adequate term is *Predicate-Internal Subject Hypothesis*.) Under this view, the 'external' argument of V is base-generated VP-internally, in SpecV. Given that SpecI is the position to which nominative Case is assigned in English, NP-movement to SpecI must then take place in active sentences as it does in passive sentences; see (12):

$[_{IP}$ John$_1$ I $[_{VP}$ t$_1$ wrote the book$_2$$]]$ (12)

The VISH raises a question concerning the structure of double object constructions, where there are three argument NPs but, it seems, only two VP-internal positions (specifier and complement). The most widely adopted approach to this problem is one that relies on a *shell* (Larson, 1988; Chomsky, 1995): on top of the lexical VP, there is a vP shell with an empty head v. Two internal arguments can now be base-generated in VP, and the remaining external argument is base-generated in Specv. Obligatory V-to-v movement and Case-driven NP-movement to SpecI yield the correct surface string. (13) is a possible analysis of an English double object construction.

$[_{IP}$ John$_1$ I $[_{vP}$ t$_1$ gave$_2$ $[_{VP}$ Mary $[_{V'}$ t$_2$ a book$]]]]$ (13)

The vP shell analysis is usually extended to simple transitive and intransitive verbs, such that an external argument is base-generated in Specv throughout; it thereby qualifies as 'external' in the literal sense again.

Raising to object

There is an alternative analysis of ECM constructions that goes back to Postal (1974). Under this view, (4) does not involve exceptional Case assignment of the matrix verb to the subject position of the infinitive, but raising of the embedded subject to the object position of the matrix clause, as in (14).

$[_{IP}$ Mary$_1$ I $[_{vP}$ t$_1$ believes John$_2$ $[_{IP}$ t$_2$ to be a fool$]]]$ (14)

As before, the construction would then depend on the transparency of IP, but it would be transparency for movement, not for Case assignment.

The empirical evidence that might decide between the two possibilities is not decisive. On the one hand, adverbials that belong to the matrix clause may usually not intervene between the lower external argument and the rest of the infinitive; see (15). This is an argument for the ECM analysis.

*John believed Mary sincerely to have left (15)

On the other hand, the external argument of the infinitive c-commands a matrix adverbial that follows the infinitive, as shown by binding of the reciprocal in (16). This piece of evidence supports the raising to object analysis.

The DA proved the defendants$_1$ to be guilty during each other's$_1$ trials (16)

The raising to object analysis has been refined and generalized in such a way that all NPs that bear objective Case must undergo raising to a Case-related position provided by a functional head (e.g., the specifier of an 'object agreement' phrase AGR_OP). In such an approach, there is a complete dissociation of Theta-positions and positions to which structural Case is assigned; see Johnson (1991) and Chomsky (1995).

SCRAMBLING

Many languages exhibit a considerable amount of clause-internal free constituent order. For instance, all permutations of three argument NPs in a double object construction can result in well-formedness in German (on which the following discussion will focus). Given the UTAH, only one of the orders can be base-generated; the remaining orders are derived by *scrambling*, a movement type introduced by Ross (1967). For present purposes, we can assume (following what is arguably the standard view) that subject → indirect (dative) object → direct (accusative) object is the base order in German. Thus, (17b–f) are derived from (17a) by scrambling.

a. dass $[_{IP}$ die Frau$_1$ dem Mann$_2$ das Buch$_3$ gegeben hat$]$
   that the woman$_{nom}$ the man$_{dat}$ the book$_{acc}$ given has
b. dass $[_{IP}$ die Frau$_1$ das Buch$_3$ dem Mann$_2$ t$_3$ gegeben hat$]$
c. dass $[_{IP}$ das Buch$_3$ die Frau$_1$ dem Mann$_2$ t$_3$ gegeben hat$]$

d. dass [$_{IP}$ dem Mann$_2$ die Frau$_1$ t$_2$ das Buch$_3$ gegeben hat]
e. dass [$_{IP}$ dem Mann$_2$ das Buch$_3$ die Frau$_1$ t$_2$ t$_3$ gegeben hat]
f. dass [$_{IP}$ das Buch$_3$ dem Mann$_2$ die Frau$_1$ t$_2$ t$_3$ gegeben hat] (17)

An independent argument for a scrambling approach is provided by the fact that clause-internal word order variation is not confined to co-arguments of a predicate in German. Scrambling can also move an item that is base-generated within an object NP to a clause-internal position. This is shown for extraction of a pronoun *da* (lit. 'there', here 'it') from a prepositional phrase (PP) embedded in an object NP in (18).

dass [$_{NP}$ da]$_1$ keiner [$_{NP}$ eine Ahnung [$_{PP}$ t$_1$ von] gehabt hat]
that it no one$_{nom}$ a notion$_{acc}$ of had has (18)

There are different views as to what the landing site of scrambling is. Data such as (17e, f) show that scrambling can be iterated. This requires either additional functional categories that provide a unique landing site for each scrambling operation, or a general mechanism that produces as many landing sites as are needed in one domain. Assuming the latter, we can postulate that scrambling in German (where NP-movement to SpecI is not obligatory) is movement within the vP domain – adjunction to vP (as was standardly assumed), or substitution in (outer) specifiers of v (see Chomsky (1995) on the possibility of multiple specifiers).

Scrambling (in German) differs from NP-movement (in English) in being optional. It has so far proven difficult to find a trigger for scrambling, which is required in syntactic theories that adopt economy principles and thereby require each movement operation to be forced in some way. Triggers that have been suggested include abstract features and information-structural requirements.

Much recent work has focused on how scrambling fits into the typology of movement operations that is based on the distinction between NP-movement and *wh*-movement. (This distinction is often generalized to a dichotomy of *A*- versus *A-bar movement*, with NP-movement an instance of the former, and *wh*-movement an instance of the latter.) In some respects, it does not seem to fit at all. For instance, since scrambling can cross an intervening scrambled item (see (17e, f)), it appears to be exempt from Relativized Minimality effects. In most cases, however, scrambling patterns with either NP-movement or *wh*-movement.

First, in many languages, scrambling is clause-bound. Thus, scrambling in German can never leave a finite CP (see (19a)), like NP-movement and unlike *wh*-movement. Depending on the nature of the matrix verb, scrambling may or may not take place from a nonfinite clause. This is often taken to show that so-called *restructuring* verbs select a bare IP infinitive, whereas *non-restructuring* verbs select a CP infinitive (see (19b, c)).

a. *dass den Fritz$_1$ keiner sagt [$_{CP}$ (t$_1'$) dass Maria t$_1$ mag]
that ART Fritz$_{acc}$ no one$_{nom}$ says that Maria$_{nom}$ likes
b. *dass den Fritz$_1$ keiner [$_{CP}$ (t$_1'$) C t$_1$ einzuladen] abgelehnt hat
that ART Fritz$_{acc}$ no one$_{nom}$ to invite rejected has
c. dass den Fritz$_1$ keiner [$_{IP}$ t$_1$ einzuladen] versucht hat
that ART Fritz$_{acc}$ no one$_{nom}$ to invite tried has (19)

However, in some languages (among them Russian, Persian, Korean, and Japanese), long-distance scrambling across a CP is possible, like *wh*-movement in English.

Second, scrambling at first glance seems to pattern with NP-movement as regards the binding of anaphors. (20a) shows that a direct object can bind an indirect object that follows it; given that the reverse order is base-generated, this implies that scrambling creates new possibilities for anaphoric binding. However, it is then unclear why an indirect object can never bind a direct object; see (20b). Furthermore, scrambling in front of a subject cannot license a nominative anaphor, like *wh*-movement; see (20c). All this might be taken to show that anaphoric binding is not solely regulated by structural factors, but relies on linear precedence and dominance on a thematic hierarchy, in which case the evidence is neutral between an NP-movement and a *wh*-movement analysis (see Jackendoff, 1990; Williams, 1994).

a. dass der Fritz die Gäste$_1$ einander$_1$ t$_1$ vorstellte
that ART Fritz$_{nom}$ the guests$_{acc}$ each other$_{dat}$ introduced
b. *dass der Fritz den Gästen$_1$ einander$_1$ vorstellte
that ART Fritz$_{nom}$ the guests$_{dat}$ each other$_{acc}$ introduced
c. *dass den Fritz$_1$ sich$_1$ t$_1$ mag
that ART Fritz$_{acc}$ self$_{nom}$ likes (20)

Third, German scrambling does not give rise to (clear) weak crossover effects, like NP-movement; see (21)

dass jeden Jungen$_1$ [$_{NP}$ seine$_1$ Mutter] t_1 liebt
that every boy$_{acc}$ his mother$_{nom}$ loves (21)

Fourth, scrambling licenses parasitic gaps, like *wh*-movement; see (22):

dass das Buch$_1$ jeder [$_{CP}$ ohne e$_1$ zu lesen] ins Regal t$_1$ zurückgestellt hat
that the book$_{acc}$ everyone$_{nom}$ without to read into the shelf put has (22)

Thus, scrambling in German seems to share some properties with NP-movement, and some with *wh*-movement. To preserve the strict A-/A-bar dichotomy mentioned above, scrambling must be assimilated with either NP-movement or *wh*-movement, and conflicting pieces of evidence must be explained away. Alternatively, the non-homogeneous evidence can be taken to indicate that the A-/A-bar distinction should be dispensed with, and be replaced by a finer-grained system according to which, for example, scrambling forms a natural class with NP-movement insofar as it targets an IP-internal position, and with *wh*-movement insofar as it is not Case-driven. Syntactic constraints can then refer to these distinctions. (*See* **Anaphora**, **Binding Theory**)

PRONOUN MOVEMENT

Object Shift

Unstressed object pronouns move out of the vP to a clause-internal position in Scandinavian languages, thereby crossing vP-external material such as adverbs and negation; this operation is known as *object shift*. The landing site follows the canonical subject position SpecI; see the contrast between (23a) and (23b) in Danish.

a. Hvorfor købte$_3$ [$_{IP}$ Peter$_2$ den$_1$ ikke [$_{vP}$ t$_2$[$_{v'}$ t$_3$ t$_1$]]]?
why bought Peter it not
b. *Hvorfor købte$_3$ [$_{IP}$ Peter$_2$ — ikke [$_{vP}$ t$_2$[$_{v'}$ t$_3$ den$_1$]]]?
why bought Peter not it (23)

The nature of the landing site is not generally agreed on; a possibility is that it is the specifier of a functional projection like AGR$_O$P that intervenes between IP and vP. Similarly, it is unclear whether pronominal object shift is phrasal (XP) movement or head (X^0) movement. (*See* **Phrase Structure and X-bar Theory**)

While being confined to unstressed pronouns in the Mainland Scandinavian languages, object shift can also affect non-pronominal NPs in Icelandic, the main difference being that movement of the latter is optional. Furthermore, object shift can (and, in the case of two unstressed object pronouns, must) be iterated, as we have seen with scrambling. However, Scandinavian object shift differs from scrambling in German (and other languages) in a number of respects. First, in contrast to scrambling, object shift requires movement of the main verb to a higher position ('Holmberg's generalization'; see Holmberg, 1999; Chomsky, 2001). This is ensured by V-to-C movement in (23a), but not in a minimally different sentence where *købte* ('brought') is replaced by the perfect form *har købt* ('has bought'), such that *har* is in C and *købt* stays within vP. Consequently, object shift is impossible in the latter case. (Dependence on main verb movement also explains why object shift is strictly local and cannot even leave restructuring infinitives.) Second, unlike scrambling, object shift does not seem to pattern with *wh*-movement in any respect (e.g. it does not license parasitic gaps). Third, while scrambling typically reverses the D-structure order of arguments, object shift is strictly order-preserving: a shifted direct object can never show up in front of an indirect object.

Pronoun Fronting

German also exhibits obligatory fronting of unstressed object pronouns to a clause-internal position; see (24a, b). This position precedes the scrambling domain vP. However, it can in turn optionally be preceded by the subject; see (24c):

a. dass es$_1$ [$_{vP}$ der Fritz der Maria t_1 gegeben hat]
that it$_{acc}$ ART Fritz$_{nom}$ ART Maria$_{dat}$ given has
b. *dass —[$_{vP}$ der Fritz der Maria es$_1$ gegeben hat]
that ART Fritz$_{nom}$ ART Maria$_{dat}$ it$_{acc}$ given has
c. dass der Fritz$_2$ es$_1$ [$_{vP}$ t$_2$ der Maria t$_1$ gegeben hat]
that ART Fritz$_{nom}$ it$_{acc}$ ART Maria$_{dat}$ given has (24)

Pronoun fronting in German is local in the same way that scrambling is: a finite CP can never be crossed, and movement from an infinitive is possible with restructuring verbs, but not with others:

a. *dass es_1 keiner $[_{CP}$ (t'_1) C t_1 zu lesen]
abgelehnt hat
that it_{acc} no one_{nom} to read
rejected has
b. dass es_1 keiner $[_{IP}$ t_1 zu lesen] versucht
hat
that it_{acc} no one_{nom} to read tried
has (25)

It seems desirable to analyse German pronoun fronting and Danish object shift in the same way. Under this view, the fact that subject NPs precede fronted pronouns optionally in German, and obligatorily in Danish, would result from an independently motivated difference with respect to optional versus obligatory NP-movement to SpecI. Furthermore, multiple pronoun fronting results in a fixed order that is reminiscent of the order-preservation effect with object shift. Still, there are many differences. First, the fixed-order effect is not the same: the order is indirect object → direct object with object shift, but direct object → indirect object with pronoun fronting. Second, pronoun fronting can cross an intervening non-pronominal NP, which object shift cannot. Third, main verb movement does not seem to be required with pronoun fronting. Finally, pronoun fronting in German shares some properties with *wh*-movement (e.g. it licenses parasitic gaps).

Cliticization

Unstressed pronouns may be (pro- or en-) *clitic* in the sense that they must attach to the left or to the right of a suitable phonological host, usually V. Pronominal cliticization is widespread in the Romance and Slavic languages. (26a) is an example from French.

Jean les_1 mange t_1
Jean them eats (26)

The trigger for cliticization is arguably phonological, given the clitic pronoun's need to form a phonological word with an appropriate host. Since nothing can intervene between a clitic pronoun and its host, cliticization is often analyzed as involving head movement (*incorporation*) into the X^0 position in which the verb shows up. However, whereas head movement is extremely local (Baker, 1988), clitic movement seems to obey roughly the same locality constraints as scrambling and pronoun movement in German: a finite CP cannot normally be crossed, but cliticization from an infinitive (*clitic climbing*) is permitted in some languages, where a restructuring verb occurs in the matrix clause. Clitic climbing is impossible in French, but applies optionally in Italian: see (27a, b) (alternatively, in (27a), *lo* can attach as an enclitic to *leggere*):

a. Mario lo_1 vuole $[_{IP}$ leggere $t_1]$
Mario it wants to read
b. *Mario lo_1 odia $[_{CP}$ (t'_1) C $[_{IP}$ leggere $t_1]]$
Mario it hates to read (27)

In view of this dual nature, it has been argued that a moved clitic pronoun simultaneously acts as an X^0 category and as an XP.

EXTRAPOSITION

Whereas NP-movement, scrambling, and pronoun movement involve leftward displacement, extraposition is rightward movement. This movement type is often optional and seems to be motivated at least in part by parsing requirements. Extraposition has been argued to underlie instances of optional rightward PP- and CP-displacement from NP (see (28a, b)), so-called *heavy NP shift* (see (28c)), and argument CP-displacement in languages with subject–object–verb word order (see (28d) from German, which is strongly preferred over the preverbal option).

a. $[_{NP}$ A review $t_1]$ will appear shortly
$[_{PP}$ of his new book$]_1$
b. $[_{NP}$ A woman $t_1]$ came into the room
$[_{CP}$ that no one knew$]_1$
c. She threw t_1 into the wastebasket
$[_{NP}$ the letter which she had not decoded$]_1$
d. dass er t_1 dachte $[_{CP}$ dass sie schläft]
that he thought that she sleeps (28)

Extraposition is clause-bound (Ross, 1967), like NP-movement; see (29a). On the other hand, heavy NP shift has been claimed to license parasitic gaps, like *wh*-movement; see (29b):

a. *John always maintains $[_{CP}$ that $[_{NP}$ a review $t_1]$ will appear shortly] whenever he is asked about it $[_{PP}$ of his new book$]_1$
b. John offended t_1 $[_{CP}$ by not recognizing e_1 immediately] $[_{NP}$ his favourite uncle from Cleveland$]_1$ (29)

Extraposition has a number of peculiar properties that set it apart from other instances of movement. To name just one, subject NPs as in (28a, b)

are barriers for leftward movement; so the well-formedness of these examples is initially surprising. Moreover, rightward movement has been argued to be dubious on purely conceptual grounds (Kayne, 1994).

LOCALITY

There is one fundamental property that all movement types that target an IP-internal position share, and that sets them apart from movement types that target a position in the CP domain: displacement is clause-bound, that is, long-distance movement across a CP is impossible; see (30). (There is evidence that the CP domain can be targeted by scrambling in languages with long-distance scrambling.)

$$^{*} \ldots [_{IP} \ldots \alpha_1 \ldots [_{CP} (t_1') \ldots t_1 \ldots] \ldots \quad (30)$$

Thus, there is a correlation between the position targeted by a movement type (low versus high) and the distance over which it can apply (short versus long). This generalization seems hardly accidental; it can be explained by a conspiracy of two constraints. First, there is a locality constraint that permits extraction from a CP only via SpecC (see, e.g., the *Phase Impenetrability Condition* in Chomsky, 2001). This precludes one-step movement without t_1' in (30). Second, there is a constraint on improper movement according to which movement to an IP-internal position may precede movement to SpecC, but not vice versa; this asymmetry can be taken to reflect the hierarchy of the target positions in the tree. See Williams (1974) for the basic observation, and May (1979), Chomsky (1981), and Müller and Sternefeld (1993) for specific proposals. (*See* **Constraints on Movement**)

References

Baker M (1988) *Incorporation*. Chicago, IL: Chicago University Press.

Chomsky N (1981) *Lectures on Government and Binding*. Dordrecht, Netherlands: Foris.

Chomsky N (1982) *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky N (2001) Derivation by phase. In: Kenstowicz M (ed.) *Ken Hale. A Life in Language*, pp. 1–52. Cambridge, MA: MIT Press.

Culicover P and Jackendoff R (2001) Control is not movement. *Linguistic Inquiry* **32**: 493–512.

Culicover P and Rochemont M (1990) Extraposition and the complement principle. *Linguistic Inquiry* **21**: 23–47.

Fanselow G (2001) Features, Theta-roles, and free constituent order. *Linguistic Inquiry* **32**: 405–436.

Holmberg A (1999) Remarks on Holmberg's generalization. *Studia Linguistica* **53**: 1–39.

Hornstein N (2001) *Move! A Minimalist Theory of Construal*. Oxford, UK: Blackwell.

Jackendoff R (1990) On Larson's treatment of the double object construction. *Linguistic Inquiry* **21**: 427–456.

Johnson K (1991) Object positions. *Natural Language and Linguistic Theory* **9**: 577–636.

Kayne R (1994) *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Koster J (1987) *Domains and Dynasties*. Dordrecht, Netherlands: Kluwer.

Larson R (1988) On the double object construction. *Linguistic Inquiry* **19**: 335–391.

Manzini MR (1983) On control and control theory. *Linguistic Inquiry* **14**: 421–446.

May R (1979) Must Comp-to-Comp movement be stipulated? *Linguistic Inquiry* **10**: 719–725.

Müller G and Sternefeld W (1993) Improper movement and unambiguous binding. *Linguistic Inquiry* **24**: 461–507.

Postal P (1974) *On Raising*. Cambridge, MA: MIT Press.

Rizzi L (1990) *Relativized Minimality*. Cambridge, MA: MIT Press.

Ross, JR (1967) *Constraints on Variables in Syntax*. Doctoral dissertation, MIT, Cambridge, MA. [Appeared 1986 as *Infinite Syntax*. Norwood, NJ: Ablex Publishing Corporation.]

Sportiche D (1988) A theory of floating quantifiers and its corollaries for constituent structure. *Linguistic Inquiry* **19**: 33–60.

Williams E (1974) *Rule Ordering in Syntax*. Doctoral dissertation, MIT, Cambridge, MA.

Williams E (1994) *Thematic Structure in Syntax*. Cambridge, MA: MIT Press.

Further Reading

Baltin M (2001) A-movements. In: Baltin M and Collins C (eds) *The Handbook of Contemporary Syntactic Theory*, pp. 226–254. Oxford, UK: Blackwell.

Beerman D *et al*. (eds) (1997) *Rightward Movement*. Amsterdam, Netherlands: Benjamins.

Borer H (ed.) (1986) *The Syntax of Pronominal Clitics*. Orlando, FL: Academic Press.

Bošković Ž and Takahashi D (1998) Scrambling and last resort. *Linguistic Inquiry* **29**: 347–366.

Chomsky N (2000) *Minimalist inquiries: the framework*. In: Martin R, Michaels D and Uriagereka J (eds) *Step by Step*, pp. 89–155. Cambridge, MA: MIT Press.

Corver N and Riemsdijk H van (eds) (1994) *Studies on Scrambling*. Berlin, Germany: Mouton de Gruyter.

Frank R, Lee Y-S and Rambow O (1995) Scrambling, reconstruction and subject binding. *Rivista di Grammatica Generativa* **21**: 67–106.

Haider H (1997) Scrambling – locality, economy, and directionality. In: Tonoike S (ed.) *Scrambling*, pp. 61–91.

Tokyo, Japan: Kurosio Publishers (Linguistics Workshop Series 5).
Jaeggli O (1986) Passive. *Linguistic Inquiry* **17**: 587–622.
Riemsdijk H van (ed.) (1999) *Clitics in the Languages of Europe*. Berlin, Germany: Mouton de Gruyter.
Thráinsson H (2001) Object shift and scrambling. In: Baltin M and Collins C (eds) *The Handbook of Contemporary Syntactic Theory*, pp. 148–202. Oxford, UK: Blackwell.
Ura H (2000) *Checking Theory and Grammatical Functions in Universal Grammar*. New York and Oxford, UK: Oxford University Press.
Webelhuth G (1992) *Principles and Parameters of Syntactic Saturation*. New York and Oxford, UK: Oxford University Press.

Metaphor

Intermediate article

Matthew S McGlone, Lafayette College, Easton, Pennsylvania, USA

CONTENTS

Metaphor is a figure of speech in which a word or phrase is used to describe something it does not literally denote. How we are able to go 'beyond the literal' to understand metaphors is a central question in the study of language and thought.

INTRODUCTION

Metaphor, from the Greek *metapherein* ('transference'), is a figure of speech in which a word or phrase is used to describe something it does not literally denote, for example, 'This encyclopedia is a *feast*'. Whether or not you agree with this characterization of the encyclopedia, you probably had no difficulty understanding it. Furthermore, your understanding did not hinge on a literal reading of the sentence – in other words, at no point in your reading did you contemplate tasting (let alone feasting upon) the book. The meaning of a metaphorical expression does not coincide with the literal meanings of the words comprising it. How then do we go 'beyond the literal' to understand metaphors? There are scholarly contemplations of this question dating back to Aristotle, but only in the twentieth century has it been regarded as an important problem in the study of language and thought.

PHILOSOPHICAL PERSPECTIVES ON METAPHOR

Aristotle's analysis of metaphor in the *Poetics* is generally considered the starting point of the topic's intellectual history. He characterized metaphor as the sign of language mastery and genius. However, he also believed that it was largely ornamental, appropriate for poetry but too enigmatic to use in philosophical or scientific discourse. Few contemporary language scholars agree with his limited view of metaphor's function, but many still endorse his account of metaphor understanding. According to what has become known as the Aristotelian 'comparison view', metaphors of the form *X is a Y* (e.g. 'This encyclopedia is a feast') are understood by converting them into simile form, *X is like a Y* ('This encyclopedia is *like* a feast'). This conversion serves the dual purpose of affording the proposition literal truth (in that any two things, even an encyclopedia and a feast, are literally alike in some respects) and making explicit the analogical comparison Aristotle presumed to be the crux of metaphor. Once converted to a simile, the metaphor is then interpreted by determining in what respects the two things being compared are similar. The comparison view thus (a) treats metaphor as a species of analogy, and (b) asserts that the perception of similarity is the basis of metaphor use and comprehension. These proposals have recently been challenged on theoretical and empirical grounds, but none the less have historically dominated scholarly discussions of metaphor.

Aristotle's relegation of metaphor to ornamentation had the unfortunate effect of leading many subsequent generations of language scholars to ignore the topic altogether. Up until the late nineteenth century, the study of metaphor was primarily the province of rhetoricians who focused almost exclusively on the interpretation of particular metaphors in literary texts. Around the turn of the twentieth century, an English translation of Michel Breal's *Essay de semantique* sparked new interest in the topic among American linguists and philosophers. Breal persuasively argued that metaphor was not mere ornament, but a ubiquitous feature of language and a principal device of linguistic change. I. A. Richards took up this cause in the 1930s and introduced a terminology of metaphor that has become fairly standard: the term used metaphorically is the 'vehicle' (e.g. feast), the term to which it is applied is the 'topic' or 'tenor' (e.g. encyclopedia), and the meaning of the metaphor is the 'ground'.

In his 1962 book *Models and Metaphors*, Max Black articulated an influential alternative to traditional

views of metaphor understanding. He rejected Aristotle's 'comparison view' of metaphor as elliptical comparison, and criticized what he called the 'substitution view' wherein metaphor is assumed to be a fancy substitute for literal language. Building on Richards' work, he argued that the product of metaphor comprehension is a complex 'interaction' of the topic and vehicle concepts. According to this 'interaction view', metaphors are understood by perceiving the topic concept 'in terms of' the vehicle to produce a ground that transcends their literal meanings. The imprecision of the notion of 'interaction' has limited the influence of Black's view on subsequent metaphor theory; however, his claim that the topic and vehicle play asymmetric roles in metaphor understanding (contrary to the traditional comparison view) has endured.

John Searle's 1979 analysis of metaphor in terms of speech act theory focused attention on the question of how a metaphorical interpretation of an utterance is initiated. He assumed that a defective literal meaning was a necessary cue for the hearer to interpret an utterance metaphorically. Searle thus assumed that before one interprets 'This encyclopedia is a feast' metaphorically, you must first contemplate and reject the possibility that the speaker wanted you to think that the book is edible and tasty. This assumption implies a model of metaphor comprehension comprising three stages. First, a literal interpretation of the utterance is derived. Second, the appropriateness of this interpretation is assessed against the context of the utterance. Third, if the literal meaning is deemed defective in context, then *and only then* is an alternative interpretation derived. For Searle (as for Aristotle), this alternative interpretation involved the conversion of metaphor into an analogical comparison.

PSYCHOLOGICAL MODELS OF METAPHOR COMPREHENSION

By the mid-1970s, metaphor had become a topic of interest among cognitive psychologists. Initial research on this topic focused on the question of whether metaphors and other nonliteral expressions (idioms, indirect requests, irony, hyperbole, etc.) required comprehension strategies different from those of literal language. Although Searle's speech act analysis appeared to motivate the postulation of a distinct nonliteral interpretation strategy, this analysis is based on two dubious psychological assumptions. The first is that the literal meaning of an utterance is derived before any possible nonliteral meaning it might have. This claim is contradicted by empirical demonstrations that when people understand nonliteral expressions such as indirect requests (e.g. '*Could* you pass the salt?') and idioms (e.g. 'The old man *kicked the bucket*'), literal meanings are not (and need not be) derived at all. Furthermore, experimental evidence indicating that metaphors are no more difficult to understand than comparable literal expressions suggests that literal meaning does not have interpretational priority.

Second, metaphor comprehension is optional on Searle's analysis, in that a nonliteral meaning of an utterance should not be considered when its literal meaning makes sense in context. This claim is also unfounded. Sometimes a metaphor can be recognized because it is literally false. For example, when a wife asserts that 'My husband is a pig', no one would interpret her assertion as meaning that her husband is a broad-snouted barnyard animal. But a metaphor need not be literally false. The opposite assertion – 'My husband is not a pig' – is literally true; the husband is not a barnyard animal. However, this is not likely to be the speaker's intended meaning or the hearer's interpretation. Thus even when the literal meaning is not defective in context, people go beyond the literal meaning of the utterance to arrive at the speaker's intention. This example is consistent with empirical demonstrations of people's inability to ignore the metaphorical meanings of utterances, even when their literal meanings are contextually appropriate.

Subsequent psychological research on metaphor has focused on how the topic and vehicle concepts 'interact' to produce the metaphoric ground. Two general types of model have been proposed to describe this interaction: similarity models and attribution models. Following Aristotle's lead, similarity models assume that the first step in metaphor comprehension is recognition of the elliptical comparison that the metaphor implies. Once the comparison is recognized, it is interpreted in much the same manner as a literal comparison, by searching for the properties and relations common to the topic and vehicle concepts. The chief difference between literal and metaphoric comparisons is that the metaphoric ground is composed of common properties that are of low salience in the topic and high salience in the vehicle. For example, the ground of 'Religion is a drug' might include properties such as 'soothing' and 'euphoria-inducing', which are highly salient properties of drugs but less salient for religions.

Two important assumptions underlie similarity models of metaphor. The first is that the referential

scope of the topic and vehicle terms is restricted to entities that conform to their conventional dictionary definitions. This assumption is what motivated Aristotle to assert that metaphors are covert comparisons in the first place. If it were possible that a vehicle term such as 'drug' could be understood as referring to a category that can include 'religion' as a member, then it would be unnecessary for people to transform 'Religion is a drug' into a comparison statement to understand it. There are good reasons to question this assumption, which we will consider below. The second assumption is that people are aware of the relevant properties that the topic and vehicle concepts have in common. This assumption is clearly violated when people are able to interpret informative metaphors. For example, consider 'That film was a sermon'. People not familiar with the film in question do not, prior to hearing this utterance, have a representation of the film that includes properties such as 'preachy' or 'moralistic'. Yet these are exactly the sorts of properties that come to mind upon reading the statement, even when the film is not familiar to the hearer. This argument applies with equal force to literal comparisons. If a person knows nothing about Saab sedans, then telling her 'A Saab is like a BMW' will introduce new properties into her mental representation of Saabs (e.g. hand-built, expensive), rather than produce a match between Saab and BMW properties.

Motivated in part by the case of informative metaphor, attribution models characterize metaphors as exactly what they appear to be: class-inclusion assertions of the form *X is a Y*. This claim implies that the referential scope of the terms comprising a metaphor is not as limited as Aristotle and other similarity theorists have assumed. According to attribution models, the vehicle term of a metaphor (e.g. 'sermon') is extended to name a class of things that its literal referent exemplifies (e.g. 'preachy, moralistic discourses'). To say that a film is a sermon is to attribute (i.e. transfer, in the sense of its Greek root) salient properties of the metaphoric category exemplified by sermons and named 'sermon' to this particular film. By virtue of the metaphoric assertion, this particular film is now included in the 'sermon' category and as a consequence of this categorization is now similar in relevant respects to literal sermons. Predicative metaphors, which employ verbs metaphorically, function the same way. For example, the verb 'to fly' literally entails movement through the air. Because flight through the air exemplifies fast movement, expressions such as 'He jumped on his bike and flew home' are as readily understood as nominal metaphors such as 'His bike was an arrow'. Arrows exemplify the category of objects that move at fast speeds; flying exemplifies the category of fast movement. In this manner are nominal and predicative metaphors used to attribute properties to particular topics.

WHY ARE METAPHORS USED?

Metaphors are used because there are often no comparable literal expressions to convey the same proposition. This is frequently the case when metaphorical expressions reflect a complex analogy between domains. For example, there are a variety of metaphorical expressions that reflect an analogy between love and journeys, as in 'Love is a two-way street', 'We've come a long way since we first dated', or 'Our relationship has come to a crossroads'. Once a target domain (e.g. love) has been described in terms of a source domain (e.g. journeys), new correspondences can be introduced, as in 'Let's walk off into the sunset together'. Whether the analogical coherence of metaphorical expressions constitutes conceptual knowledge *per se* (as argued by the linguist George Lakoff and colleagues) or simply an efficient means of communicating about complex domains remains largely an unresolved issue.

Further Reading

Aristotle (1996) *Poetics*, translated by M Heath. New York, NY: Penguin Books.

Black M (1962) *Models and Metaphors*. Ithaca, NY: Cornell University Press.

Breal MJA (1964) *Semantics: Studies in the Science of Meaning*, translated by H Crust. Toronto: University of Toronto Press.

Gentner D and Wolff P (1997) Alignment in the processing of metaphor. *Journal of Memory and Language* **37**: 331–355.

Glucksberg S (2001) *Understanding Figurative Language*. New York, NY: Oxford University Press.

Lakoff G and Johnson M (1980) *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

McGlone M (1996) Conceptual metaphors and figurative language: food for thought? *Journal of Memory and Language* **35**: 544–565.

Ortony A (1993) *Metaphor and Thought*. New York, NY: Cambridge University Press.

Richards IA (1936) *The Philosophy of Rhetoric*. New York, NY: Oxford University Press.

Searle J (1979) *Expression and Meaning*. New York, NY: Cambridge University Press.

Morphological Processing

Intermediate article

R Harald Baayen, University of Nijmegen/Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Robert Schreuder, University of Nijmegen, Nijmegen, The Netherlands

CONTENTS

Words with internal morphological structure are processed differently in the mental lexicon than monomorphemic words, both in language comprehension and in speech production. How morphological structure is represented in the brain is hotly disputed. Dissociations between the processing of regular and irregular complex words on the one hand, and gradient morphological phenomena on the other, are key issues in the debate between single and dual route approaches, and between symbolic and connectionist approaches.

INTRODUCTION

Many words have internal morphological structure (e.g. 'walk-ed', 'good-ness', 'key-pad'). Such internal structure is evident from systematic correspondences between the form and meanings of series of words. Some patterns of form–meaning correspondences are pervasive in the language and are easily and readily extended by language users to form new words. Such productive patterns are typically found in inflectional morphology (e.g. person and number marking on verbs, number marking on nouns). Even young children have already internalized such patterns and readily apply them to nonsense words, producing 'memped' as the past tense of 'memp' without difficulty. In linguistics, rules (e.g. add *ed* to form the past tense in English) are used to account for such productive form–meaning patterns. By contrast, other patterns of form–meaning correspondences occur in only a few words; for example, the derivational suffix *-th* that forms nouns such as 'length' and 'strength' from the adjectives 'long' and 'strong'. Unproductive patterns appear to be exceptional instead of governed by an active rule, and linguists generally assume that the words exemplifying such patterns are listed in the lexicon as wholes.

There are two main views in linguistics as to the nature of morphological rules. Linguists in the generative tradition (e.g. Pinker, 1997) regard rules as existing in the mind independently of the data from which they were abstracted. In this approach, any complex word that is predictable by rule (e.g. 'linked') is not stored in the lexicon; only words with unpredictable properties (e.g. 'sang') are listed there. Formally, morphological rules are described in this tradition using the mechanisms of formal grammars.

Others (e.g. Skousen, 1989) argue that rules are analogical in nature. They assume that a great many complex words, both irregular and regular, are available in lexical memory. For a past tense form such as 'linked', assuming that it is not already available in memory, an analogical rule would first look for all verbs in memory that are most similar to the verb 'link' (think, sink, blink, wink). It would then look at the corresponding past tense forms, and select the one that occurs most often with the greatest probability. Every once in a while, a less likely exemplar would be selected, leading to deviant forms such as 'lank'. Explicit computational algorithms have been developed for modeling analogy.

While linguistic theories account for morphological structure using techniques borrowed from computer science, it is not self-evident that these theories provide good descriptions of the computational mechanisms actually used by the human brain.

MORPHOLOGY IN THE MENTAL LEXICON

When considering the question how morphologically complex words are processed in the mental lexicon, it is important to distinguish between speech production and language comprehension. The speaker has to produce the correct morphological form given conceptual input

('walk' + PAST → 'walked'), while readers and listeners receive the correct morphological form and have to unravel its meaning. The different tasks of the speaker and the listener lead to an asymmetry between comprehension and production. For instance, second language learners of German master the comprehension of the complex system of German noun plurals fairly quickly, while getting the plural forms right for production takes considerably more experience.

The most striking evidence of the role of morphology comes from studies of errors in speech production (see, e.g., Stemberger, 1998), which show that during production things can go wrong both with the lining up of morphemes, and with the selection of morphemes. This shows that complex words are constructed online on the basis of their constituent morphemes.

Turning to language comprehension, one key question is whether morphological relationships are represented in the mental lexicon independently of orthographic and phonological relationships. Priming studies (see, e.g., McQueen and Cutler, 1998) have shown that words preceded by morphologically related words are processed faster than words preceded by orthographically, phonologically, or semantically matched control words. Moreover, morphological priming effects persist over many intervening trials, while orthographic and semantic repetition effects fade away rapidly. Another phenomenon pointing to the relevance of morphological structure in the mental lexicon is the morphological family size effect (De Jong *et al*., 2000). Words that occur as a constituent in many other words are responded to faster in visual lexical decision than words occurring as a constituent in just a few other words. This effect is a semantic effect involving the co-activation of semantically related words along lines of morphological similarity. Thus, there is strong evidence that morphological structure is represented in the mental lexicon.

Another key question is whether complex words are necessarily decomposed into their constituent morphemes during recognition, or whether complex words are recognized on the basis of stored forms of the complex words as wholes, without decomposition into the constituent morphemes. It appears that decomposition may take place, depending on a wide range of factors, including the semantic transparency of the complex word (Marslen-Wilson *et al*., 1994), frequency of use, the specific affix under consideration, its possible functional ambiguity, and its productivity (Bertram *et al*., 2000). Decomposition is more likely to occur for transparent, lower-frequency words with productive and functionally unambiguous affixes. However, recognition on the basis of stored forms of complex words seems to be fairly pervasive, not only for compounding and derivation, but also for regular inflection, especially in languages with simple morphological systems such as English.

COGNITIVE MODELS FOR MORPHOLOGICAL PROCESSING

There are four main approaches to the modeling of morphological processing.

Cascaded Dual Route Models

The Cascaded Dual Route approach is similar in spirit to the generative linguistics view. Regular complex words are assumed to be processed by symbolic rules; irregular complex words are supposed to be retrieved from an associative memory. The two routes, the route using morphological rules and the route retrieving irregular complex words from an associative memory, are taken to be subserved by different parts of the brain. The two routes are cascaded in the sense that rules are used only upon failure of memory retrieval (Pinker, 1997; Clahsen, 1999).

This model provides an elegant and simple account for a great many differences that have been reported in the literature with respect to the processing of regular versus irregular inflected words. However, many of the differences that supposedly support it are open to multiple interpretations and some appear, on closer scrutiny, not to exist. For instance, inflected words have been argued to differ with respect to the frequency effect for the inflected form as a whole, with only irregular inflected words showing full form frequency effects. However, various studies (see, e.g., Taft, 1979; Sereno and Jongman, 1997) have reported such full form frequency effects even for fully regular inflected words. This casts doubt upon the linguistic idea that it is only irregular words that are in the lexicon.

Parallel Dual Route Models

In Parallel Dual Route models (see, e.g., Caramazza *et al*., 1988), morphological processing also proceeds on the basis of two routes, a direct route retrieving whole complex words from a (not necessarily associative) memory, and an indirect route

using rules (not necessarily symbolic and potentially analogical in nature). Crucially, the two routes are assumed to run in parallel. Because they may use shared or even competing representations, the two routes are not necessarily independent.

Perhaps the greatest weakness of these models is that they are difficult to falsify without computational implementations that embody explicit mechanisms for lexical look-up and parsing. On the positive side, the Parallel Dual Route approach seems promising for dealing with the resolution of parsing ambiguity. Long complex words can often be segmented into strings of morphemes in a great many ways. In Dutch, for instance, the word 'belangstellende', 'interested party', has some 80 mostly spurious segmentations. Simulation studies suggest that storage protects the parser against such spurious segmentations. In the Parallel Dual Route framework, the relevant constituents of a longer complex word (belang, stellen, de: 'interest', 'to place', INFLECTIONS) may become available without sequences of possible but irrelevant constituents (bel, angst, ellende: 'to ring', 'fear', 'misery') becoming available as well (Baayen and Schreuder, 2000).

Symbolic Single Route Models

In Symbolic Single Route approaches, morphological processing is explained in terms of activation spreading through lexical networks. Standard linguistic insights are translated into activation models driven by general incremental procedures sequentially activating symbolic representations at various cascaded levels. In Levelt's production model (Levelt *et al*., 1999), for instance, a past tense form such as 'walk-ed' is produced by activating the lemma of 'walk' and the diacritical feature PAST TENSE. As the lemma of 'walk' is prespecified for combining with the morpheme *-ed*, the morphemes 'walk' and *-ed* are selected for further processing. The lemma for 'see', when combined with the diacritic PAST TENSE, directly passes on activation to the word form 'saw'.

The strength of this model is that it accounts for a wide range of chronometric data on speech production. With respect to the production of complex words, the model explains in detail how morphemes might be strung together, beginning with the first morpheme to be pronounced and ending with the last morpheme. But it has nothing to say about the production of new complex words. Its main weakness, which it shares with the dual route models, is that in its present form it cannot handle gradient morphological phenomena.

Connectionist Single Route Models

In models that make use of artificial neural networks (ANNs), words no longer have independent representations. Instead, they receive implicit representations in terms of activation weights and patterns of activation distributed in the ANN. Currently, pools of orthographic, phonological, and semantic units are combined in one ANN, an architecture generally referred to as the triangle model. In this approach, morphological effects in lexical processing are considered to be epiphenomena of statistical regularities in the lexicon.

The ability to handle gradient effects is the main strength of distributed connectionist single route models. Various ANN models predict the right kind of regular and irregular past tense forms in English without incorporating explicit rules (Plunket and Juola, 1999). ANN models are not without their own share of problems, however. First, reduplication is widespread in the languages of the world, but ANN cannot be trained to handle reduplicated forms without training individual forms into the network (Sproat, 1992). Second, it is a matter of dispute whether the kind of errors made by ANN models are really similar to the kind of errors made by humans (but see Stemberger, 1998).

CONCLUSION

Morphologically complex words are at the juncture of form and meaning. The complexity of this juncture arises from the absence of a clear and unambiguous mapping of the two domains.

In production, relatively little is known about the underlying conceptual processes that lead to the selection of a sequence of morphemes to be combined into a complex word, while detailed knowledge is available on how these morphemes are sequenced for articulation. For comprehension, it is becoming increasingly apparent that a wide range of linguistic and statistical factors are all jointly exploited in order to extract the intended interpretation from the auditory or visual input. The details of semantic processing are as yet unknown both for comprehension and for production.

The challenge for the cognitive modeling of morphological processing is twofold. On the one hand, cognitive models should capture the linguistic functions of morphology in all its complexity in

a plausible process model. On the other hand, cognitive models should also be sensitive to the many nondeterministic, statistical patterns that are present in the language and the way it is used. Current computational models tend to meet only one of these challenges.

References

Baayen RH and Schreuder R (2000) Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society, Series A* **358**: 1–13.

Bertram R, Schreuder R and Baayen RH (2000) The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Memory, Learning, and Cognition* **26**: 419–511.

Caramazza A, Laudanna A and Romani C (1988) Lexical access and inflectional morphology. *Cognition* **28**: 297–332.

Clahsen H (1999) Lexical entries and rules of language: a multi-disciplinary study of German inflection. *Behavioral and Brain Sciences* **22**: 991–1060.

De Jong NH, Schreuder R and Baayen RH (2000) The morphological family size effect and morphology. *Language and Cognitive Processes* **15**: 329–365.

Levelt WJM, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 1–38.

Marslen-Wilson W, Tyler LK, Waksler R and Older L (1994) Morphology and meaning in the English mental lexicon. *Psychological Review* **101**: 3–33.

McQueen J and Cutler A (1998) Morphology in word recognition. In: Zwicky AM and Spencer A (eds) *The Handbook of Morphology*, pp. 406–427. Oxford, UK: Blackwell.

Pinker S (1997) Words and rules in the human brain. *Nature* **387**: 547–548.

Plunkett K and Juola P (1999) A connectionist model of English past tense and plural morphology. *Cognitive Science* **23**: 463–490.

Sereno J and Jongman A (1997) Processing of English inflectional morphology. *Memory and Cognition* **25**: 425–437.

Skousen R (1989) *Analogical Modeling of Language*. Dordrecht, Netherlands: Kluwer.

Sproat R (1992) *Morphology and Computation*. Cambridge, MA: MIT Press.

Stemberger J (1998) Morphology in language production with special reference to connectionism. In: Zwicky AM and Spencer A (eds) *The Handbook of Morphology*, pp. 428–452. Oxford, UK: Blackwell.

Taft M (1979) Recognition of affixed words and the word frequency effect. *Memory and Cognition* **7**: 263–272.

Further Reading

Baayen RH (2002) Probability in morphology. In: Bod R, Jannedy S and Hay J (eds) *Probability in Linguistics*. Cambridge, MA: MIT Press.

Bybee JL (2001) *Phonology and Language Use*. Cambridge, UK: Cambridge University Press.

Marcus G (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Pinker S (1999) *Words and Rules: The Ingredients of Language*. London, UK: Weidenfeld & Nicolson.

Seidenberg M and Gonnerman L (2000) Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences* **4**: 353–361.

Morphology, Computational

Intermediate article

Kemal Oflazer, Sabanci University, Istanbul, Turkey

CONTENTS

Computational morphology can be defined as the study of computational analysis and synthesis of word forms in the context of natural language processing with the computer.

INTRODUCTION

Morphology is the study of the structure of words and how words are formed by combining smaller units of linguistic information called *morphemes*. Morphemes can be classified into two groups: free morphemes (e.g. roots) can occur by themselves as a word, while bound morphemes are not words in their own right but have to be attached in some way to a free morpheme. The way in which morphemes are combined, and the information conveyed by morphemes and their combination, differ from language to language. Sproat (1992) details how languages are classified with respect to their morphology (*isolating*, *inflecting*, *agglutinative*, *polysynthetic* languages); discusses morphological processes such as *inflection* (which is used to generate forms of a word as demanded by the syntactic context), *derivation* (which is used to generate new words semantically related to the original word but possibly with a different part of speech), and *compounding* (which is used to generate new words by concatenating two or more word stems); and describes types of morphological combinations, such as *suffixation, prefixation, circumfixation, templatic combination, reduplication,* etc. (*See* **Morphology**)

Computational morphology aims at developing formalisms and algorithms for the computational analysis and synthesis of word forms for use in language processing applications. Applications such as spelling checking and correction, stemming in document indexing, and so on, also rely on techniques in computational morphology, especially for languages with rich morphology.

ASPECTS OF COMPUTATIONAL MORPHOLOGY

Computational morphology has two main computational tasks: *morphological analysis* analyses a word token to extract all linguistically relevant information for later use; *morphological generation* synthesizes a word from a set of features. Morphological analysis has problems analogous to all those in full-blown syntactic parsing, albeit usually on a smaller scale. Words may be ambiguous with regard to their part of speech, or may be divided up in a number of ways, each giving rise to possibly different morphological interpretations. For instance, in English, the word 'books' can be interpreted as:

book + Noun + Plural
  (plural of the noun 'book')
book + Verb + Present + 3PS
  (3rd person singular present of the verb 'to book')

In a morphologically complex language like Turkish, a simple word like 'oyun' can be interpreted in a number of ways, as:

oyun + Noun + Singular + NoPossessive + Nominative	(*game*)
oy + Noun + Singular + Possessive2SG + Nominative	(*your vote*)
oy + Noun + Singular + NoPossessive + Genitive	(*of the vote*)
oy + Verb + Imperative + 2SG	(*carve!*)

Morphological generation deals with the reverse process of producing the correct word form from a set of features. For example, in a machine translation application, say from French to English, one may need to produce the past tense of the verb 'stop' and the morphological generation system should produce the form 'stopped', dealing with the gemination (or doubling) of the consonant *p*.

Computational morphology attempts to model and capture two main aspects of word formation: morphophonology or morphographemics, and morphotactics. Morphophonology and its counterpart for words in written form, morphographemics, refer to the changes in pronunciation and orthography that occur when morphemes are put together. For instance in English, when the derivational suffix *-ness* is affixed to the adjective stem 'happy' to derive a noun, we get 'happiness'. The word-final *-y* in the spelling of 'happy' changes to an *-i*. Similarly, in the present continuous form of the verb 'stop', we need to geminate the last consonant of the root to get 'stopping'. Turkish, for instance, has a process known as *vowel harmony*, which requires that vowels in affixed morphemes agree in various phonological features with the vowels in the root or the preceding morphemes. For instance, *-lar* in 'pullar' ('stamps') and *-ler* in 'güller' ('roses') both indicate plurality; the vowel *-u* in the first word's root forces the vowel in the suffix to be an *-a*, and the *-ü* in the second word's root forces the vowel in the suffix to be an *-e*. Words where such agreement is missing are considered to be ill-formed. Computational morphology develops formalisms for describing such changes, the contexts they occur in, and whether they are obligatory or optional (e.g. 'modeling' and 'modelling' are both valid forms).

MORPHOTACTICS

Morphotactics describes the structure of words, that is, how morphemes are combined to form words as demanded by the syntactic context and with the correct semantics (in the case of derivational morphology). The root words of a language are grouped into *lexicons* based on their part of speech and other criteria that determine their morphotactical behavior. Similarly, the bound morpheme inventory of the language is also grouped into lexicons. If morphemes are combined using prefixation or suffixation, then the morphotactics of the language describes the proper ordering of the lexicons from which morphemes are chosen. Morphotactics in languages like Arabic require more elaborate combinations where roots consisting of just consonants are combined with a vocalization template that describes how vowels and consonants are interdigitated to form the word with the right set of features.

The state-of-the art formalisms for describing morphographemic phenomena are based on the mathematically well developed and understood theory of regular languages and relations, and their computational models of finite state recognizers and transducers (Kaplan and Kay, 1994). Such morphophonological and morphographemic changes are described as either a cascade of *replace rules* or a set of parallel *two-level rules* (Koskenniemi, 1983), both of which are compiled into finite state transducers. The morphotactics for most languages can also be handled by finite state machinery: the ordering of morphemes can be described by relatively simple finite state grammars and can be compiled into finite state transducers. Even for languages like Arabic with complex morpheme combinations, finite state methods can be used (Beesley, 1996; Kiraz, 2000). The use of finite state machinery enables the combination of the finite state transducers for morphographemics and morphotactics into a single finite state transducer, which can process words very fast (Karttunen, 1994). There are, however, certain morphotactical phenomena, notably partial or full duplication (for example, in Indonesian, the plural of 'orang' ('man') is 'orangorang'), which cannot be elegantly described by finite state machinery, apart from full listing of the duplicated forms in a lexicon.

Finite state tools for building wide-coverage morphological analyzers typically let (computational) linguists describe the morphographemics and morphotactics using high-level notational tools which are then compiled into finite state transducers. This compilation may, however, be quite memory-intensive as algorithms for combining finite state transducers rely on operations such as composition that may require large amounts of memory. Another source of complexity which (at least theoretically) comes up in the context of two-level morphology is that the generate-and-test approach employed in some of the earlier two-level morphology implementations (e.g. Antworth, 1990) may be inherently computationally inefficient (Barton *et al*., 1988, chap. 5). However, empirical evaluations have shown that this complexity is not observed in practice (Koskenniemi and Church, 1988).

Another recent thread of research in computational morphology is the use of machine learning techniques for inducing morphophonological transducers (Gildea and Jurafsky, 1996), for two-level rules (Theron and Cloete, 1997), and for bootstrapping full morphological analyzers (Oflazer *et al*., 2001).

References

Antworth E (1990) *PC-KIMMO: A Two-level Processor for Morphological Analysis*. Dallas, TX: Summer Institute of Linguistics.

Barton GE, Berwick RC and Ristad ES (1988) *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.

Beesley KR (1996) Arabic finite-state morphological analysis and generation. In: *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

Gildea D and Jurafsky D (1996) Learning bias and phonological-rule induction. *Computational Linguistics* **22**(4): 497–530.

Kaplan RM and Kay M (1994) Regular models of phonological rule systems. *Computational Linguistics* **20**(3): 331–378.

Karttunen L (1994) Constructing lexical transducers. In: *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Kiraz GA (2000) Multi-tiered nonlinear morphology using multi-tape finite automata: a case study on Syriac and Arabic. *Computational Linguistics* **26**(1): 77–105.

Koskenniemi K (1983) *Two-level Morphology: A General Computational Model for Word Form Recognition and Production*. Publication no. 11, Department of General Linguistics, University of Helsinki, Finland.

Koskenniemi K and Church K (1988) Complexity, two-level morphology and Finnish. In: *Proceedings of COLING-88*, pp. 335–339. Budapest, Hungary.

Oflazer K, Nirenburg S and McShane M (2001) Bootstrapping morphological analysers by combining human elicitation and machine learning. *Computational Linguistics* **26**(1): 59–85.

Sproat R (1992) *Morphology and Computation*. Cambridge, MA: MIT Press.

Theron P and Cloete I (1997) Automatic acquisition of two-level rules. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, USA.

Further Reading

Oflazer K (1999) Morphological analysis. In: van Halteren H (ed.) *Syntactic Wordclass Tagging*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Ritchie GD, Russell GJ, Black AW and Pulman SG (1992) *Computational Morphology*. Cambridge, MA: MIT Press.

Morphology

Introductory article

Stephen R Anderson, Yale University, New Haven, Connecticut, USA

CONTENTS

Morphology, in linguistics, is the study of the forms of words, and the ways in which words are related to other words of the same language. Formal differences among words serve a variety of purposes, from the creation of new lexical items to the indication of grammatical structure.

INTRODUCTION

If you ask most nonlinguists what primary thing has to be learned in order to *know* a language, the answer is likely to be 'the words of the language'. Learning vocabulary is a major focus of language instruction, and while everyone knows that there is a certain amount of 'grammar' that characterizes a language as well, even this is often treated as a kind of annotation to the set of words – the 'uses of the accusative', etc. But what is it that is involved in knowing the words of a language?

Obviously, a good deal of this is a matter of learning that *cat*, pronounced [k^hæt], is a word of English, a noun that refers to a 'feline mammal usually having thick soft fur and being unable to roar'. The notion that the word is a combination of sound and meaning – indeed, *the* unit in which the two are united – was the basis of the theory of the linguistic *sign* developed by Ferdinand de Saussure at the beginning of the twentieth century. But if words like *cat* were all there were in language, the only thing that would matter about the form of a word would be that it differs from the forms of other words (i.e., cat is pronounced differently from *mat, cap, dog*, etc.). Clearly there is no more specific connection between the parts of the sound of cat and the parts of its meaning: the initial [k^h], for example, does not refer to the fur. The connection between sound and meaning is irreducible here.

But of course *cat* and words like it are not the end of the story. Another English word is *cats*, a single word in pronunciation but one that can be seen to be composed of a first part *cat* followed by another part *–s*, with the meaning of the whole made up of the meaning of *cat* and the meaning of *–s* ('plural'). *Cattish* behavior is that which is similar to that of a cat; and while a *catbird* is not itself a kind of cat, its name comes from the fact that it sometimes sounds like one. All of these words are clearly connected with *cat*, but they are also all words in their own right.

We might, of course, simply have memorized *cats, cattish* and *catbird* along with *cat*, even though the words seem to have some sort of relation to one another. But suppose we learn about a new animal, a *wug*, say 'a large, hairy bovine mammal known for being aggressive and braying'. We do not need to learn independently that two of these are *wugs*, or that *wuggish* behaviour is likely to involve attacking one's fellows, or that a *wugbird* (if there were such a thing) might be a bird with a braying call. All of these things follow from the knowledge we have not just of the specific words of our language, but of their relations to one another, in form and meaning. The latter is our knowledge of the morphology of our language.

In some languages, the use of morphology to pack complex meanings into a single word is much more elaborate than in English. In West Greenlandic, for example, *tusaanngitsuusaartuaannarsiinnaanngivipputit* is a single word meaning 'you simply cannot pretend not to be hearing all the time'. Other languages do much less of this sort of thing: Chinese and Vietnamese are often cited in this connection, though Chinese does have rather exuberant use of compounding (structures like *catbird* made up of two existing items). Despite this variation, however, morphology is an aspect of the grammar of all languages, and in some it rivals syntax in the expressive power it permits.

INFLECTION

Traditionally, morphology is divided into several types, depending on the role played in grammar by

a given formation. The most basic division is between inflection and word formation: the latter is easy enough to characterize as 'morphology that creates new words' (*wuggish, wug-like, wugbird*), but inflection (e.g., *wugs*) is rather harder to define. Often, inflection is defined by example: categories like number (e.g., 'plural'), gender (e.g., masculine, feminine and neuter in Latin), tense ('past'), aspect (e.g., the difference between the *imparfait* and the *passé simple* in French), case ('accusative'), person (first *vs* second *vs* third) and perhaps a few others are inflectional while everything else is word formation. But this approach is inadequate, because the same category may be inflectional in some languages, and not in others. In Fula (a West Atlantic language), for example, the category 'diminutive' is fully integrated into the grammar of agreement in the language, just as much as person, number, and gender. Verbs whose subjects are diminutive indicate this with an agreement marker, as do adjectives modifying diminutive nouns etc. In English, in contrast, diminutives appear in forms like *piglet*, but these are clearly cases of word formation. On the other hand, while number is clearly involved in important parts of English grammar (verbs agree with their subjects in number), other languages, like K^{w}ak^{w}'ala (or 'Kwakiutl') treat the category of plural as something that can optionally be added to nouns, or to verbs, as an elaboration of meaning that has no further grammatical consequence.

Despite the intuitively clear nature of the category of inflection, other efforts to define it explicitly do no better. Inflection is generally more *productive* than other sorts of morphology, for instance: virtually every German noun has an accusative, a plural etc., while only a few English nouns have a diminutive formation like *piglet*. But in some languages, categories that we would certainly like to call inflectional are quite limited: in Basque, for example, only a few dozen verbs (the number varying from one dialect to another) have forms that show agreement. In English, on the other hand, the process of forming nouns in *–ing* from verbs (as in *Fred's lonely musings about love*) can take virtually any verb as its basis, despite being intuitively a means of creating new words, not of inflecting old ones. A variety of other attempts that have been cited also fail, either because of ready counterexamples, or because they are insufficiently general: inflectional material is generally found at the word's periphery, while word formational markers are closer to the stem (cf. *piglets* but not **pigslet*), but this property is only useful in words that contain material of both types, and even then, it does not help us to find the boundary in a word like French *im-mort-al-is-er-ait* 'would immortalize'.

In fact, the intuition underlying the notion of 'inflection' seems to be the following: inflectional categories are those that provide information about grammatical structure (such as the fact that a noun in the accusative is likely to be a direct object), or which are referred to by a grammatical rule operating across words (such as the agreement of verbs with their subjects). The validity of other correlates with inflectional status, then, follows not from the nature of the categories themselves, but rather from the existence of grammatical rules in particular languages that refer to them, and to the freedom with which items of particular word classes can appear in positions where they can serve as the targets of such rules.

For any given word, we can organize a complete set of its inflectional variants into a *paradigm* of the word. Thus, a German noun has a particular gender, and a paradigm consisting of forms for two numbers (singular and plural) and four cases (nominative, genitive, dative, and accusative). German adjectives have paradigms that distinguish not only case and number but also gender (since they can agree with nouns of any of the three genders), plus another category that distinguishes between 'strong' and 'weak' declensions (depending on the presence of certain demonstrative words within the same phrase).

All of the word forms that make up a single inflectional paradigm have the same basic meaning. In general, they are all constructed on the basis of a basic shape, or stem, though in many languages with complex inflection, the paradigm of a given word may be built from more than one stem. In French, for example, the verb *pouvoir* 'to be able to' shows different stems in (*je*) *peux* 'I can' and (*je*) *pourrais* 'I would be able to'.

Certain terminology has become more or less accepted in describing facts of these sorts. We refer to a particular sound shape (e.g. [fawnd]) as a specific *word form*; all of the inflectional forms in a single paradigm are said to make up a single *lexeme* (e.g., *find*). A specific *morphosyntactic form* of a particular lexeme (e.g., the past tense of *find*) is realized by a corresponding word form [fawnd]). These terms are all distinct, in their way: thus, the same morphosyntactic form of a given lexeme may correspond to more than one word form (e.g., the past tense of *dive* can be either [daivd] or [dowv]), while the same word form can realize more than one morphosyntactic form (e.g., [hɪt] can be either the past tense of *hit*, the non-third-person present tense of *hit*, or the singular of the noun *hit*).

WORD FORMATION

Inflection, then, is the morphology that distinguishes the various forms within the paradigm of a single lexeme. Some languages, like ancient Greek or Georgian, have a great deal of inflectional morphology, while others (like English) have much less, and some (like Vietnamese) have hardly any at all. Regardless of this, however, essentially all languages have ways of constructing new lexemes from existing ones, or patterns of word formation. These fall into two broad classes: *compounding* is the process of combining two or more independently existing lexemes (perhaps with some additional material as 'glue') into a single new lexeme (as in *catbird*). *Derivation*, in contrast, is the formation of a new lexeme from an existing one by means of material that does not appear by itself as a word. It is common to refer to such non-independent content as *bound*, in contrast with independently occurring or *free* elements.

Derivation

A typical derivational relation among lexemes is the formation of adjectives like *inflatable* from verbs (*inflate*). In this case, the meaning of the adjective is quite systematically related to that of the verb: VERB-*able* means 'capable of being VERB-ed'. It is therefore tempting to say that English contains an element *–able* with that meaning which can simply be added to verbs to yield adjectives. The facts are a bit more complex that that, though.

For one thing, the related adjective may not always be just what we would get by putting the two pieces together. For instance, *navigate* yields *navigable, formulate* yields *formulable*, etc. These are instances of *truncation*, where a part of the base is removed as an aspect of the word formation process. Then there are cases such as *applicable* from *apply*, where we see the same variation (or *allomorphy*) in the shape of the stem as in *application*. These patterns show us that the derivational whole may be more than the simple sum of its parts.

When we consider the class of adjectives in *–able* (or its spelling variant *–ible*), we find a number of forms like *credible, eligible, potable, probable* … which seem to have the right meaning for the class (they all mean roughly 'capable of being [SOMETHING]-ed'), but the language does not happen to contain any verb with right form and meaning to serve as their base. This suggests that derivational patterns have a sort of independent existence: they can serve as (at least partial) motivation for the shape and sense of a given lexeme, even in the absence of the possibility of deriving that lexeme from some other existing lexeme. In some instances, the force of this analysis is so strong that it leads to what is called *back-formation*: thus, the word *editor* was originally derived from Latin *e:dere* 'to bring forth' plus *–itor*, but it fit so well into the pattern of English agent nouns in *–er* (e.g., *baker, driver*) that a hypothetical underlying verb *edit* actually became part of the language.

We may also notice that some *–able* forms do not mean precisely what we might predict. Thus, *comparable* means 'roughly equal', not just 'able to be compared'. In the world of wine, *drinkable* comes to mean 'rather good', not just 'able to be drunk', etc. This shows us that even though these words may originally arise through the invocation of derivational patterns, the results are in fact full-fledged words of the language; and as such, they can undergo semantic change independent of the words from which they were derived. This is the same phenomenon we see when the word *transmission*, originally referring to the act or process of transmitting (e.g., energy from the engine to the wheels of a car) comes to refer to a somewhat mysterious apparatus which makes strange noises and costs quite a bit to replace.

Finally, we can note that in some cases it is not at all evident how to establish a 'direction' of derivation. In Maasai, for example, there are two main noun classes ('masculine' and 'feminine'), and one derivational pattern consists in taking a noun which is 'basically' of one class and treating it as a member of the other. Thus, *en-kérái* is a feminine noun that refers to any child, of either gender; while *ol-kérái* is a corresponding masculine noun meaning 'large male child'. Here it looks plausible to take the feminine form as the basis for the derivational relationship; but when we consider *ol-abáánì* (masculine) 'doctor' vs. *enk-abáánì* 'small or female doctor, quack' it looks as if the direction of derivation goes the other way. In fact, it looks as if what we have here is a case of a relation between two distinct patterns, where membership in the feminine class may (but need not) imply femaleness and/or relatively small size, as opposed to the masculine class which may imply maleness and/or relatively large size. When a word in either class is used in the other, the result is to bring out the additional meaning associated with the class, but there is no inherent directionality to this relationship. The possibility of back-formation discussed above suggests that this interpretation of derivational relationships as fundamentally symmetrical may be applicable even to cases where the formal direction of derivation seems obvious.

Compounding

The other variety of word formation, compounding, seems fairly straightforward, even if the actual facts can be quite complex at times. Compounds are built of two (or more) independent words, and have (at least in their original form) a meaning that involves those of their components. Thus, a *catfish* is a kind of fish sharing some property with a cat (in this case, the whiskers). Like derived forms, compounds are independent lexemes in their own right, and as such quickly take on specialized meanings that are not transparently derived from those of their parts. We need to tell a story to explain why a *hotdog* is called that, why a *blackboard* can be green, etc.

Where it is possible to relate the meaning of a compound to those of its parts, it is often possible to establish a privileged relationship between the semantic 'type' of the whole compound and that of one of its pieces. Thus, a *doghouse* is a kind of *house* (and certainly not a kind of *dog*), *outdoing* is a kind of *doing*, etc. When such a relation can be discerned, we refer to the 'privileged' member of the compound as its head, and speak of the compound itself as *endocentric*.

By no means all compounds would appear to be endocentric, however: a *pickpocket* is neither a kind of pocket nor a kind of picking, and a *sabre-tooth* is a kind of tiger, not a kind of tooth. Traditional grammar provides a variety of names for different types of such *exocentric* compounds, some deriving from the Sanskrit grammatical tradition in which these were of particular interest. A *bahuvrihi* compound is one whose elements describe a characteristic property or attribute possessed by the referent (e.g., *sabre-tooth, flatfoot*); a *dvandva* compound is built of two (or more) parts, each of which contributes equally to the sense (e.g., an *Arab-Israeli* peace treaty).

In some languages, the decision as to which compounds are endocentric and which are not depends on the importance we give to different possible criteria. For instance, in German, *Blauhemd* '(soldier wearing a) blue shirt' is on the face of it a bahuvrihi compound, exocentric because it does not denote a kind of shirt. On the other hand, the gender of the compound (neuter, in this case) is determined by that of its rightmost element (here, [*das*] *hemd* '[the] shirt'). Semantically, *blauhemd* is exocentric; while grammatically, it could be regarded as endocentric with its head on the right.

Languages can vary quite a bit in the kinds of compound patterns they employ. Thus, English compounds of a verb and its object (like *scarecrow*) are rather rare and unproductive, while this constitutes a basic and quite general pattern in French and other Romance languages. English and German tend to have the head, when there is one, on the right (*dollhouse*), while Italian and other romance languages more often have the head on the left (e.g., *caffelatte* 'coffee with milk'). Most English compounds consist of two elements (though one of these may itself be a compound, as in [[*high school*] *teacher*], leading to structures of great complexity such as German [[[[Leben]s-versicherung]s-gesellschaft]s-angestellter], 'life insurance company employee', but many dvandva compounds in Chinese consist of three or four components, as in *ting-tai-lou-ge*, '(pavilions-terraces-upper stories-raised alcoves) elaborate architecture'.

Finally, although we have defined compounds as built from free elements or independent lexemes, this leaves us with no good way of describing structures such as the names of many chemical compounds and drugs (*dichlorobenzene*, *erythromycin*) and words such as *Italo-American*. On the one hand, we surely do not want to say that there is a process that affects a base such as *American* by prefixing *Italo*. On the other hand, *Italo-*, *erythro-*, *chloro-* etc. do not occur on their own, but only in this class of compounds. Even more striking examples occur in other languages: the Mandarin root *yi*, 'ant', freely forms compounds such as *yiwang*, 'queen ant' (literally ant-king); *gongyi*, 'worker ant', *baiyi*, 'white ant, termite'. But *yi* is clearly not a word: the free word for 'ant' in Mandarin is *mayi*. While English *erythro* etc. are always prefixes, in Mandarin, the roots in question occur in both head and non-head position and are therefore like normal compound components in every respect except that they are not free forms. It appears that the very definition of compounding needs more thought than was initially evident.

REPRESENTATION OF MORPHOLOGICAL KNOWLEDGE

So far, we have talked of morphological relationships as existing between whole lexemes (in the case of word formation), or between word forms (in the case of inflection). Much traditional thought about morphology, however, regards these matters somewhat differently. Earlier, it was pointed out that the model of the Saussurean sign as the minimal unit where sound and meaning are connected could not serve as a description of the word, since (proper) parts of words often display their own connection between sound and meaning. It was

this observation, in fact, that led us to explore the varieties of morphology displayed in natural language. But many have felt that the proper place for the sign relation is not the word, but rather a constituent part of words: the *morpheme*. In that view, morphology is the study of these units, the morphemes: how they may vary in shape (the *allomorphy* they exhibit) and how they can be combined (*morphotactics*).

Morphemes and Words

The notion that words can be regarded as (exhaustively) composed of smaller signlike units, or morphemes, is extremely appealing. It leads to a simple and uniform theory of morphology, one based on elementary units that can be regarded as making up a sort of lexicon at a finer level of granularity than that of words. Nonetheless, it seems that this picture of word structure as based on a uniform relation of morpheme concatenation is literally too good to be true.

If morphemes are to serve the purpose for which they were intended, they ought to have some rather specific properties. It ought to be possible, for any given word, to divide its meaning into some small number of subparts, to divide its form into a corresponding number of continuous substrings of phonetic material, and then to establish a correspondence between the parts of meaning and the parts of form. Of course, it is possible to do exactly that in a great many cases (e.g., *inflatable*): hence the intuitive appeal of this notion. But in many other instances, such a division of the form is much more labored or even impossible.

One fairly minor problem is posed by parts of the form that are not continuous. When we analyze words containing circumfixes (e.g., *ke—an* in Indonesian *kebisaan*, 'capability', from *bisa*, 'be able') or infixes (e.g., *-al-* in Sundanese *ngadalahar*, 'to eat several', from *ngadahar*, 'to eat') one or the other of the component morphemes is not a continuous string of material.

Other cases are more serious. For instance, we may find no component of meaning to correspond to a given piece of form (an 'empty morph' such as the *th* in English *lengthen*, 'make long(er)') or no component of form that relates to some clear aspect of a word's meaning (English *hit* 'past tense of *hit*'). Sometimes two or more components of meaning are indissolubly linked in a single element of form, as in French *au* ([o]), 'to the (masc.)' or the ending *–o*: of Latin *amo:* which represents all of 'first person singular present indicative', various categories that are indicated separately in other forms. When we look beyond the simple cases, it appears that the relation between form and meaning in the general case is not one-to-one at the level of the morpheme, but rather many-to-many.

In fact, it seems that even though both the forms and the meanings of words can be divided into components, the relation is still best regarded as holding at the level of the entire word, rather than localized exclusively in the morpheme. We have also seen support for this notion in the fact that entire words, presumably composed of multiple morphemes, develop idiosyncratic aspects of meaning that cannot be attributed to any of their component morphemes individually (e.g., *appreciable* and *considerable* coming to mean not 'capable of being appreciated/considered' but 'substantial, relatively large'). On this basis, many linguists have come to believe that morphological relations are based on the word rather than the morpheme. Actually, we need to take into account the fact that in highly inflected languages like Latin or Sanskrit, no existing surface word form may supply just the level of detail we need, since all such words have specific inflectional material added. For such a case, we need to say that it is *stems* (full words minus any inflectional affixation) that serve as the basis of morphological generalizations, in the sense of representing the phonological component of a lexeme.

Items and Processes

A further difficulty for the notion that morphemes are the basis of all morphology comes from the fact that in many cases some of the information carried by the form of a word is represented in a way that does not lend itself to segmentation. One large group of examples of this sort is supplied by instances in which it is the replacement of one part of the form by another, rather than the addition of a new piece, that carries meaning. Such relations of *apophony* include *umlaut* (*goose/geese, mouse/mice*), *ablaut* (*sing/sang/sung*) and such miscellaneous relations as those found in *food/feed*, *sell/sale, sing/song*, *breath/breathe*, and many others. Terms for these relations often refer to their historical origins and do not reflect any particularly natural category in the modern language (e.g., the categories of *umlaut* as opposed to *ablaut* in modern English).

Sometimes some information is carried in a word's form not by the addition of some material (a morpheme), but by the deletion of something that we might expect. In the Uto-Aztecan language Tohono O'odham ('Papago') for example, the perfective form of a verb can in most instances be

found by dropping the last consonant of the imperfective form (whatever that may be): thus, *gatwid* 'shooting' yields perfective *gatwi* 'shot'; *hikck* 'cutting' yields *hikc* 'cut', etc.

Examples like these (and several others which we cannot consider here for reasons of space) suggest that the relations between words that constitute a language's morphology are best construed as a collection of *processes* relating one class of words to another rather than as a collection of constituent morphemic *items* that can be concatenated with one another to yield complex words. Of course, the simplest and most straightforward instance of such a process is one that adds material to the form (a prefix at the beginning, a suffix at the end or an infix within the basic stem), but this is only one of the formal relations we find in the morphologies of natural languages. Others include changes, permutations, deletions and the like. Linguists set on treating all morphological relations as involving the addition of morphemes have proposed analyses of many of these apparent processes in such terms, but it is possible to ask whether the extensions required in the notion of what constitutes an 'affix' do not in the end empty it of its original theoretical significance.

CONCLUSION

We have seen above that the forms of words can carry complex and highly structured information. Words do not serve merely as minimal signs, arbitrary chunks of sound that bear meaning simply by virtue of being distinct from one another. Some aspects of a word's form may indicate the relation of its underlying lexeme to others (markers of derivational morphology or of compound structure), while other aspects indicate properties of the grammatical structure within which it is found (markers of inflectional properties). All of these relations seem to be best construed as knowledge about the relations between *words*: That is, relations between whole lexemes, even when these can be regarded as containing markers of their relations to still other lexemes; and relations between word forms that realize paradigmatic alternatives built on a single lexeme's basic stem(s) in the case of inflection. These relations connect substantively defined classes in a way that is only partially directional in its essential nature, and the formal connections among these classes are signaled in ways that are best represented as processes relating one shape to another.

Further Reading

Anderson SR (1992) *A-Morphous Morphology*. Cambridge, UK: Cambridge University Press.

Aronoff M (1976) *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Bybee JL (1985) *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam, Netherlands: Benjamins.

Carstairs-McCarthy A *Current Morphology*. London, UK: Routledge.

Halle M and Marantz A (1993) Distributed morphology and the pieces of inflection. In Hale K and Keyser SJ (eds) *The View from Building 20* pp. 111–176. Cambridge, MA: MIT Press.

Marchand H (1969) *The Categories and Types of Present-Day English Word-Formation*. Munich, Germany: CH Beck.

Matthews PH (1991) *Morphology*, 2nd edn. Cambridge, UK: Cambridge University Press.

Pinker S (1999) *Words and Rules*. New York, NY: Basic Books.

Spencer A (1991) *Morphological Theory*. Oxford, UK: Blackwell.

Spencer A and Zwicky AM (eds). *The Handbook of Morphology*. Cambridge, UK: Blackwell.

Optimality Theory

Intermediate article

Diana Archangeli, University of Arizona, Tucson, Arizona, USA

CONTENTS

Optimality theory proposes that universal grammar consists of a set of constraints, and that language-specific grammars consist of different rankings of these constraints.

INTRODUCTION

Optimality theory was introduced in the early 1990s (Prince and Smolensky, 1993; McCarthy and Prince, 1993a). These early works recognized a class of cases that continued to challenge 'classical' generative phonological theory, and proposed a dramatically different model – optimality theory – with which to address those challenges. Rather than a series of rules for rewriting that change an underlying form into a surface form, the core of optimality theory (OT) is to determine the best pairing between input ('underlying') and output ('surface'), given the requirements of the language in question. This is accomplished through a device known as 'constraint satisfaction'. A very appealing aspect of OT is the nature of these constraints that need to be satisfied. They are proposed as universal statements about language, thereby providing a means for encoding universal language properties directly in the grammar of specific languages. In addition to addressing standard issues of the synchronic nature of language, OT also is proving to be a valuable tool for a deeper understanding of how language works, giving insights into areas such as language acquisition and language change. OT is not without its own challenges, addressed at the close of this article.

CHALLENGES FOR CLASSICAL PHONOLOGICAL THEORY

Classical rule-based phonological theory faces a number of challenges. One class of challenges is that in which the rule-based analysis is mechanical rather than explanatory. For example, Kisseberth (1970) pointed out that in order to account for certain patterns under a rule-based model of phonology, it was necessary that a variety of rules 'conspire' to create a particular type of surface form. The rules participating in such a conspiracy need not have any formal relation to each other, and in fact can be virtual opposites. For example, rules of vowel insertion and vowel deletion work together in Yawelmani to ensure that syllables have one of three forms: CV, CVC, or CVV. Thus, the conspiracy emerges only when the surface forms are compared. Although it is possible to express the phenomena through a complex set of rules, this strategy fails to provide an explanation: it is simply coincidental that the mental representations are such that a particular set of rules produces surface forms of a uniform type.

The second class of challenges to the rule-based system is conceptual. Rule-based models typically include constraints as well as rules, such as constraints on underlying representations (defining the feature inventories from which underlying representations are created, requiring that certain features be unassociated, etc.). If rules alone are inadequate, the challenge, then, is to test whether constraints alone are adequate. This is the goal of OT.

Another important conceptual difference is in the way markedness is characterized in the grammar. Markedness refers to powerful cross-linguistic tendencies, such as the tendency for syllables to have onsets, for heavy syllables to be stressed, etc. Under a rule-based model, markedness is peripheral to the formal model. There are efforts to include markedness somehow, such as the marking conventions introduced in *The Sound Pattern of English* (Chomsky and Halle, 1968). However, the coherence of the rule-based model is not affected by whether or not markedness is incorporated formally. Thus, significant cross-linguistic generalizations are peripheral, not integral, to classical phonological theory. OT, as a theory of constraint satisfaction, places markedness constraints at the core of individual language phonologies.

CONSTRAINT SATISFACTION

The architecture of OT is simple. There are universal primitives of two types, primitive elements and primitive constraints. Phonological elements (or units, such as features, moras, and syllables) are the basis from which representations are constructed. Constraints govern both markedness and faithfulness. Markedness constraints prefer unmarked configurations. Faithfulness constraints prefer a perfect match between input and output.

In addition, there are two universal operations, generate (GEN) and evaluate (EVAL). GEN takes a specific input and randomly adds and deletes elements to construct an (in principle) infinite set of candidates for the output. EVAL examines the candidates and the relations between the candidates and the input in order to select the optimal output for that input. EVAL is composed of a language-particular ranking of the universal constraint set. The mechanism by which EVAL makes this selection is constraint satisfaction.

Candidates are evaluated with respect to the highest ranked constraint first; the candidates which survive for evaluation by the next constraint are those that best satisfy the top ranked constraint (either by fully satisfying the constraint or by violating it less than any other candidates). This procedure is repeated with the next constraint until a single candidate remains.

For example, in numerous languages, sequences of three or more consonants may result from morphological concatenation. Languages vary in their response to such sequences. OT claims that the variation is due to different rankings of the constraints in EVAL. For this case, three markedness constraints are necessary: that syllables have vowels (PEAK), that they not have codas (NOCODA), that at most one consonant appears in the onset or the coda (*CC). In addition, two faithfulness constraints are necessary, one preferring that input and output vowels match (FAITHV) and a similar one for consonants (FAITHC).

Consider, for instance, a language like English, which tolerates codas and tolerates consonant clusters within syllables. Formally, under OT, NOCODA and *CC are relatively unimportant and so are low ranked, indicated by the heavy black bar in the display. Each violation is indicated by an asterisk in the appropriate cell. The exclamation points are added to show the violation that eliminates a particular candidate from consideration. For example, (b) [lɪm.p.nɛs] is eliminated by a violation of PEAK. The ☞ indicates the optimal candidate. The shaded-in cells indicate constraints whose violations are irrelevant in the selection of the optimal candidate.

	/lɪmp-nɛs/	PEAK	FAITHV	FAITHC	NOCODA	*CC
☞a.	lɪmp.nɛs				**/	*
b.	lɪm.p.nɛs	*!			**/	
c.	lɪm.nɛs			*!	**/	
d.	lɪ.mip.nɛs		*!		*	
e.	lɪ.mi.pi.nɛ.si		*!*.*			
f.	lɪ.nɛ			*!**		

A display like this is known as a 'tableau', and is used to demonstrate constraint satisfaction given a particular constraint ranking. The heavy vertical bar indicates the most critical constraint ranking: constraints to the left are all ranked above constraints to the right.

Different constraint rankings produce different languages. For example, should *CC and FAITHC outrank FAITHV, the form with an inserted vowel (d) would be selected: this is the pattern chosen in Yawelmani. Languages like Hawaiian, which tolerate only open syllables, would rank NOCODA above all other constraints. Whether the NOCODA effect were achieved by vowel insertion or consonant deletion would depend on the relative rankings of FAITHC and FAITHV.

More abstractly, given a set of *n* constraints, there are *n*-factorial possible rankings, predicting up to *n*-factorial different languages, known as the factorial typology. As shown in the tableau, where the heavy black bar shows the only critical constraint ranking, many constraint rankings are irrelevant and so the full range of the factorial typology cannot be exploited.

THE UNIVERSAL BASIS FOR CONSTRAINTS

Constraints fall into two classes, markedness constraints and faithfulness constraints, both of which are universal. High-ranked markedness constraints indicate ways in which a language is unmarked. Low-ranked markedness constraints not only indicate ways in which a language is marked (because of being violated) but also the unmarked configurations emerge when higher ranked constraints fail to decide among candidates.

The interplay of two premises of the model, that constraints are violable and that there are markedness constraints, immediately explains the puzzle

that has frustrated earlier attempts to deal with markedness: namely, why many universals are tendencies, not absolutes.

Markedness constraints include constraints such as those noted above which define preferred configurations. In addition to constraints governing syllable structure there are constraints over feature co-occurrence, such as *if [+ round] then [+ back]*, the universal tendency for round vowels also to be back vowels. These constraints all fall into the class defining preferred combinations of primitive elements. There is, however, another large class of markedness constraints: those that prefer anchoring or aligning units with each other. Anchoring constraints may be straightforward phonological constraints, such as aligning the edge of a heavy syllable to the edge of a foot. They also may anchor some phonological unit to some morphological unit. For example, vowel harmony might be characterized by an anchor or align constraint, preferring that the harmonic feature align to the edges of the word. Align and anchor constraints differ from other markedness constraints in that their form is universal, but particular instantiations of the forms vary from language to language (McCarthy and Prince, 1993b).

Faithfulness constraints are also universal under OT. Faithfulness constraints vary along different parameters. On the formal side, faithfulness may either refer to input corresponding to output ('MAX', prohibits deletion) or to output corresponding to input ('DEP', prohibits insertion). (The FAITHV and FAITHC constraints above conflate these two types of faithfulness.) Substantively, faithfulness constraints may refer to phonological units (e.g., MAX[Voice], DEP[High]), or to their organization, such as the order or number of elements in a string (preventing metathesis and merger or breaking, respectively) (McCarthy and Prince, 1995). One advantage of this perspective is that the issue of universality arises with every constraint posited for an analysis, driving us to a broader and deeper understanding of the claims made by an analysis.

EXTENSIONS OF OT

OT, originally proposed as a model of phonology, is being extended in a variety of ways. OT proposes a means of characterizing adult grammars. As such, it is also being explored as a model of syntax (Speas, 1997, and Pesetsky, 1997, in Archangeli and Langendoen, 1997; Grimshaw, 1997; Bresnan, 1999; Barbosa *et al*., 1998 and others), of semantics (Hendriks and de Hoop, 2001), and even of the semantics–pragmatics interface (Blutner, 2000). However, most extensions explore phonological effects in language variation, change, and acquisition. These three classes of phenomena may be closely related, under the view that language variation and change occur, at least in part, due to learning a grammar that is slightly different from the one learned by the previous generation.

The optimality theoretic analysis of both variation and change hinges heavily on the concept of constraint reranking. The idea is that different versions of a grammar (whether synchronic or diachronic) arise due to minimal changes in the ranking of constraints. Research efforts are underway to determine what principles, if any, restrict the types of rerankings that occur. For example, is there a particular distance that constraints must adhere to when reranking? Are there principled limits on the types of constraints that move higher? Lower? How about the types of constraint that remain stationary (Holt, 1997)?

Under OT, language acquisition consists of the problem of determining the correct ranking of the universal constraints. The most convincing models are those which take two things as basic assumptions: (i) anything proposed to be universal under the model (i.e., constraints and primitives) and (ii) anything to which a child has physical access (i.e., adult pronunciations) (Smolensky, 1996).

CHALLENGES FOR OT

Despite the advantages and interest of OT pointed out above, there are challenging issues faced by the model. One of these is its computational and psychological impossibility: if GEN truly creates an infinite set, then in real time we should all be stuck in a GEN state.

There are also empirical challenges: for example a phenomenon known as 'opacity'. Opacity arises in a rule-based system when the environment necessary for an earlier rule to apply is eliminated by application of a later rule: both rules apply, but at the surface the environment for the earlier rule is not present – hence opacity. These cases present a serious challenge to OT because of its reliance on the surface form. Opacity has led to a variety of proposals in efforts to account for these patterns. The most successful of these approaches is known as 'sympathy', wherein the optimal candidate bears a similarity to a (failed) sympathetic candidate (McCarthy, 1998.)

More broadly, the role of phonetics with respect to phonology is being eagerly explored, with some suggesting that the constraint hierarchies should

include phonetic constraints determining exact pronunciations as well as phonological constraints determining a phonological form (see, for example, Hayes, 1996).

Another broad class of empirical challenges arises from the ways in which phonology and morphology interface. The later stages of classical rule-based theory included the theory of lexical phonology (e.g., Kiparsky, 1982) which separated the phonological rules and the morphological operations of a language into distinct strata, pairing certain sets of morphological operations with specific sets of phonological rules. The OT hypothesis might accept this approach, and posit separate constraint hierarchies. However this is counter to the overall effort to simplify the formal mechanisms, for it would require that languages be able to refer to morphological constituents both in constraints and in defining strata. A new challenge introduced by OT is to determine whether it is possible to characterize the different phonological behaviors of specific morphemes through a single constraint hierarchy. Note that if both these efforts are successful, phonetics, phonology, and morphology will all be included in a single constraint hierarchy.

References

Archangeli D and Langendoen DT (eds) (1997) *Optimality Theory: An Overview*. Oxford, UK: Blackwell.

Barbosa P, Fox D, Hagstrom P, McGinnis M and Pesetsky D (eds) (1998) *Is the Best Good Enough?* Cambridge, MA: MIT Press.

Blutner R (2000) Some aspects of optimality in natural language interpretation. *Journal of Semantics* **17**: 189–216.

Bresnan J (1999) Explaining morphosyntactic competition. *Rutgers Optimality Archive* 299–0299.

Chomsky N and Halle M (1968) *The Sound Pattern of English*. New York, NY: Harper & Row.

Grimshaw J (1997) Projection, heads, and optimality. *Linguistic Inquiry* **28**: 373–422.

Hayes B (1996) Phonetically driven phonology: the role of optimality theory and inductive grounding. *Rutgers Optimality Archive* 158–1196.

Hendriks P and de Hoop H (2001) Optimality theoretic semantics. *Linguistics and Philosophy* **24**: 1–32.

Holt D (1997) *The Role of the Listener in the Historical Phonology of Spanish and Portuguese: an Optimality-Theoretic Account*. Doctoral dissertation, Georgetown University, Washington, DC.

Kiparsky P (1982) From cyclic phonology to lexical phonology. In: van der Hulst H and Smith N (eds) *The Structure of Phonological Representations*, part 2, pp. 131–176. Dordrecht, Netherlands: Foris.

Kisseberth C (1970) On the functional unity of phonological rules. *Linguistic Inquiry* **1**: 291–306.

McCarthy J (1998) Sympathy and phonological opacity. *Rutgers Optimality Archive* 252–398.

McCarthy J and Prince A (1993a) Prosodic Morphology I: Constraint Interaction and Satisfaction. *Rutgers Optimality Archive* 482–1201.

McCarthy J and Prince A (1993b) Generalized Alignment. *Rutgers Optimality Archive*-7.

McCarthy J and Prince A (1995) Faithfulness and reduplicative identity. In: Beckman J, Dickey L and Urbanczyk S (eds) *Papers in Optimality Theory*, pp. 249–384. University of Massachusetts Occasional Papers 18. Also, *Rutgers Optimality Archive*-60.

Pesetsky D (1997) Optimality theory and syntax: movement and pronunciation. In: Archangeli D and Langendoen DT (eds) *Optimality Theory: An Overview*, pp. 171–199. Oxford: Blackwell.

Prince A and Smolensky P (1993) Optimality Theory: Constraint Interaction in Generative Grammar. *Rutgers Optimality Archive*.

Smolensky P (1996) On the comprehension/production dilemma in child language. *Linguistic Inquiry* **27**: 720–731.

Speas M (1997) Optimality theory and syntax: null pronouns and control. In: Archangeli D and Langendoen DT (eds) *Optimality Theory: An Overview*, pp. 171–199. Oxford: Blackwell.

Further Reading

Alderete J (1997) Dissimilation as local conjunction. *Proceedings of North East Linguistics Society* **27**: 17–31. Also *Rutgers Optimality Archive*-175.

Archangeli D (1999) Introducing optimality theory. *Annual Review of Anthropology* **28**: 531–552.

Ito J, Mester A and Padgett J (1995) Licensing and underspecification in OT. *Linguistic Inquiry* **26**: 571–613.

Kager R (1999) *Optimality Theory*, Cambridge, UK: Cambridge University Press.

McCarthy J (1999) Sympathy and phonological opacity. *Phonology* **16**: 331–399.

Rutgers Optimality Archive [<http://ruccs.edu/roa-nog.html>]

Smolensky P (1995) On the internal structure of the constraint component CON of UG. *Rutgers Optimality Archive*-86.

Parsing: Overview

Introductory article

Florian Wolf, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
Edward Gibson, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

Parsing is the task of determining how the words in a sentence combine to yield a structured representation of the sentence meaning, given a grammar.

INTRODUCTION

To understand a sentence, one has to determine how the words in the sentence are combined to arrive at a meaning for the sentence. Consider the following examples:

a. The dog bites the man.
b. The man bites the dog. (1)

Sentences (1a) and (1b) have the same words. However, the words are combined differently, resulting in two very different sentence meanings. The set of rules governing how words are combined for a given language is called a *grammar* of that language. Sentences (1a) and (1b) are acceptable English sentences, whereas the asterisk preceding (2) indicates that this sentence is unacceptable. Sentences (1a) and (1b) follow the rules of English, whereas sentence (2) does not.

*The dog man bites the. (2)

The task of *parsing* is to determine how the words in a sentence combine to yield a structured representation of the sentence meaning, given a grammar. In contrast to a *parser*, a *recognizer* merely determines whether a sentence is grammatically correct or not, without producing a structured representation for the input sentence. A further contrast holds between *parsers* and *generators*. Whereas a parser takes a sentence as input and provides a representation of the meaning of the sentence as output, a *generator* takes some representation of meaning as input and provides a sentence as output.

This article discusses the process of parsing. The process of parsing is not just retrieving a representation of the sentence's meaning that is already stored in memory. If that were so, we would not be able to understand sentences that we have not heard before. Furthermore, there are an infinite number of possible sentences. Storing a representation of the meaning of each sentence in memory would require infinite memory resources.

This article will give a basic introduction to some parsing strategies that were developed in computational linguistics. It will also describe a method that deals efficiently with *local* (or *temporary*) and *global* ambiguity. Human languages are highly ambiguous. For example, sentence (3) contains both local and global ambiguity:

The man saw the woman on the hill with the telescope. (3)

The word 'saw' is locally ambiguous. That is, without disambiguating context, it could be either a tool or the past tense form of the verb 'see'. Sentence (3) is also globally ambiguous. It has five different readings, because the *prepositional phrases* (*PP*s) in (3), 'on the hill' and 'with the telescope', can modify either the *noun phrase* (*NP*) 'the woman' or the *verb phrase* (*VP*) 'saw the woman'. 'With the telescope' can also modify the *NP* 'the hill'. One of these readings would be equivalent to 'The man used the telescope in order to see the woman who was on the hill'. Another possible reading would be equivalent to 'The man saw the woman who was on the hill and who had a telescope'.

The number of readings of such ambiguities grows exponentially with the number of modifying phrases, PPs in this case. Such ambiguities are very frequent in human languages, but they usually do not present a problem to humans. This is in large part because humans use their knowledge of the world in order to rule out less likely possibilities as they are processing sentences word by word. For instance, humans usually do not get a reading of (3) in which 'with the telescope' modifies 'the hill', because without further context or assumptions we assume that it is unlikely that hills have telescopes.

PARSING STRATEGIES

The task in parsing is to discover how the words of a sentence can combine, using the rules in the grammar. A very simple grammar is presented in Figure 1 (where S = sentence; NP = noun phrase; VP = verb phrase; Det = determiner). The grammar indicates that a sentence (S) expands to a noun phrase (NP) and a verb phrase (VP), and that an NP expands to a determiner (Det) and a Noun, etc. We will refer to the symbol to the left of the arrow as the *left-hand side* (*LHS*) of a rule, and the right side of the arrow as the *right-hand side* (*RHS*) of a rule. The first symbol of the RHS is called the *left corner* of an RHS. The categories on the LHS of rules are sometimes called *nodes* in the grammar. Nodes that expand directly to a word (e.g. Det, Noun and Verb) are called *pre-terminals*. Nodes that do not expand directly to a word (e.g. S, NP and VP) are called *non-terminals*.

Using a grammar like this (much more complex in real applications), a parsing algorithm then establishes a syntactic structure for an input sentence. One possible parsing strategy starts by looking at the rules and seeing what input one can find that is compatible with the rules. Such a strategy is called *top-down*. Alternatively, one might start by looking at the input, and seeing which rules in the grammar apply to that input. This is a *bottom-up* strategy. Still another possibility would be some combination of top-down and bottom-up. The following sections describe these different parsing strategies in more detail.

The parsing algorithms described below process a sentence one word at a time, from left to right, similar to when people read or listen to language. A *stack* data structure (last in, first out) is used by the parser to keep track of the categories that the parser still needs to process to obtain a complete sentence structure. The parser also keeps a record of the structure that it has built so far.

Notice that the grammar in Figure 1 is unambiguous. That is, each left-hand side node has exactly one right-hand side. However, in more realistic grammars, this is not the case. Consider the possible expansions of the category VP in Figure 2. The first rule would be used for intransitive verbs such as 'walk', as in 'I walk'. The second rule would be used for transitive verbs such as 'like', as in 'I like the apple'. The third rule would be used for VPs that are modified by a prepositional phrase, for example, 'I see you with the telescope'. A parsing algorithm can handle ambiguity either by trying one choice at a time – a serial approach – or by following multiple alternatives at once – a parallel approach. Under a serial approach, the parser tries one of the rules first and pursues the resulting structure further, and backtracks and tries the other rule(s) only if the first rule fails, or if the goal is to find all possible parses. The ordering of the rules in the grammar determines which rule is applied first. Under a parallel approach, the parser works on all alternative rules and further pursues the resulting structures in parallel threads. A parallel approach requires some data structure that contains a set of parse trees, one tree for each combination of rules.

S	→ NP VP	Det	→ the
NP	→ Det Noun	Noun	→ man
VP	→ Verb NP	Noun	→ woman
		Verb	→ likes
		Verb	→ meets

Figure 1. A very simple grammar.

VP	→	Verb
VP	→	Verb NP
VP	→	VP PP

Figure 2. Some possible expansions of a *VP*.

Top-down Parsing

In *top-down* parsing, one starts with the assumption that the input will eventually form a sentence. This involves initially positing an S-node and all of the extensions of the S-node specified by the grammar (in our case, only $S \rightarrow NP\ VP$). One keeps expanding nodes, following the rules of the grammar, until one finds some matching input. The stack in a top-down parser keeps track of what still needs to be found in order to get a sentence that is well-formed according to the grammar. The pseudocode in Figure 3 shows a top-down algorithm in a general form. Notice that in the case of ambiguity, that is, if a left-hand side of a rule has more than one possible right-hand side (cf. Figure 2), the parser does not specify an order in which these rules are applied. This is called *non-deterministic* parsing.

In the following, we provide a step-by-step example of how a top-down parser would parse the sentence 'The man likes the woman', assuming the grammar from Figure 1 (cf. Figure 4):

```
FUNCTION top-down-parse (SENTENCE)
    initialize STACK = [S]
    DO
      IF (top-element of STACK = non-terminal N) THEN
        Select a rule N → A B
        Pop N from STACK
        Push A B onto STACK
        Add A B below N in the structure for the input
      ELSE IF (top-element of STACK = pre-terminal P) THEN
        Find next word W in SENTENCE
        IF (there is a rule P → W) THEN
          Pop P from STACK
          Add W below P in the structure for the input
        ELSE
          Fail.
        ENDIF
      ENDIF
    UNTIL STACK = [] AND end of SENTENCE is reached.
    RETURN PARSE-TREES
END top-down-parse
```

Figure 3. Pseudocode for top-down parser.

- Step 1: Push *S* onto the stack. Stack = [*S*]
- Step 2: Apply the rules that have *S* on their LHS. There is only one such rule in our grammar, $S \rightarrow NP\ VP$. Pop *S* from the stack. Push *NP* and *VP* onto the stack. Stack = [*NP VP*]
- Step 3: Apply the rules that have *NP* on their LHS. Here, this is only $NP \rightarrow Det\ Noun$. Pop *NP* from the stack. Stack = [*Det Noun VP*]
- Step 4: Find 'the' in the input and incorporate it into the parse tree. Pop *Det* from the stack. Stack = [*Noun VP*]
- Step 5: Find 'man' in the input and incorporate it into the parse tree. Pop *Noun* from the stack. Stack = [*VP*]
- Step 6: Apply the rules that have *VP* on their LHS. Here, this is only $VP \rightarrow Verb\ NP$. Pop *VP* from the stack. Push *Verb* and *NP* onto the stack. Stack = [*Verb NP*]
- Step 7: Find 'likes' in the input and incorporate it into the parse tree. Pop *Verb* from the stack. Stack = [*NP*]
- Step 8: Apply the rules that have *NP* on their LHS. Here, this is only $NP \rightarrow Det\ Noun$. Pop *NP* from the stack. Push *Det* and *Noun* onto the stack. Stack = [*Det Noun*]
- Step 9: Find 'the' and incorporate it into the parse tree. Pop *Det* from the stack. Stack = [*Noun*]
- Step 10: Find 'woman' in the input and incorporate it into the parse tree. Pop *Noun* from the stack. Stack = [ ]

One of the advantages of a top-down parser is that it never tries to form a structure that will never end up being an *S*, because it starts from *S*. One of the disadvantages of a top-down parser is that it can try to build trees that are inconsistent with the input. This did not happen with our extremely simplified grammar. However, if we also had a rule like $NP \rightarrow Det\ Adj\ Noun$ (*Adj* = Adjective), the parser could have predicted a structure that is inconsistent with the input, one that contains an Adjective.

Another problem with top-down algorithms is left-recursion. A grammar is left-recursive if it has some LHS that can be expanded through a series of rules such that the left corner of one of these expansions is the same LHS category. For example, if a grammar has a rule $VP \rightarrow VP\ NP$, the parser could keep extending the left corner of its RHS, *VP*, and get caught in an endless loop, as shown in Figure 5.

Bottom-up Parsing

Whereas top-down parsing starts with the rules, *bottom-up* parsing first looks at the input and then tries to find rules in the grammar that apply to the input. The stack in a bottom-up parser keeps track of what has been found so far and what still has to be integrated in a parse tree. The bottom-up parser considered here consists of two basic steps – pushing categories on the stack that still need to be integrated into the input structure (*shift*), and applying rules in the grammar to the categories on the stack (*reduce*). This algorithm is therefore also called *shift-reduce* parsing. The pseudocode in Figure 6 shows a bottom-up shift-reduce algorithm in a general form.

In the following, we provide a step-by-step example of how a bottom-up shift-reduce parser would parse the sentence 'The man likes the woman', assuming the grammar from Figure 1 (cf. Figure 7):

- Step 1: Find 'the' and its lexical category, *Det*. Push *Det* onto the stack. Stack = [*Det*]
- Step 2: Find 'man' and its lexical category, *Noun*. Push *Noun* onto the stack. Stack = [*Noun Det*]
- Step 3: Apply the rule $NP \rightarrow Det\ Noun$ to top of stack. Pop *Det* and *Noun* from the stack. Push *NP* onto the stack. Stack = [*NP*]
- Step 4: Find 'likes' and its lexical category, *Verb*. Push *Verb* onto the stack. Stack = [*Verb NP*]
- Step 5: Find 'the' and its lexical category, *Det*. Push *Det* onto the stack. Stack = [*Det Verb NP*]
- Step 6: Find 'woman' and its lexical category, *Noun*. Push *Noun* onto the stack. Stack = [*Noun Det Verb NP*]

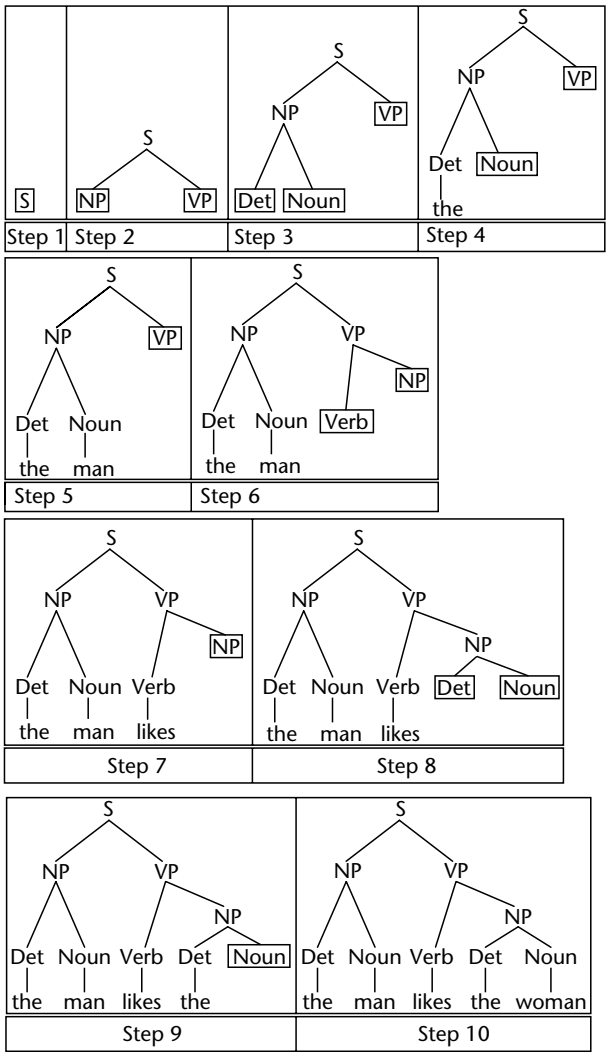


Figure 4. Top-down parsing step-by-step.

- Step 7: Apply the rule $NP \rightarrow Det\ Noun$ to top of stack. Pop *Det* and *Noun* from the stack. Push *NP* onto the stack. Stack = [*NP Verb NP*]
- Step 8: Apply the rule $VP \rightarrow Verb\ NP$ to top of stack. Pop *Verb* and *NP* from the stack. Push *VP* onto the stack. Stack = [*VP NP*]
- Step 9: Apply the rule $S \rightarrow NP\ VP$ to top of stack. Pop *NP* and *VP* from the stack. Push *S* onto the stack. Stack = [*S*]

An advantage of bottom-up parsers is that they do not predict parse trees that are inconsistent with the input. Furthermore, unlike top-down parsers, bottom-up parsers cannot get caught in an endless loop if the grammar contains left-recursive rules. A disadvantage of bottom-up parsers is that they can generate structures that never result in an *S*.

Left-corner Parsing

Left-corner parsing combines some elements from both top-down and bottom-up parsing. In left-corner parsing, only rules consistent with the input are predicted. That is, a rule is only predicted (top-down) if the current input (bottom-up) matches the leftmost node (hence left-corner) of the RHS of a rule. In left-corner parsing, a stack is used to keep track of what input is still needed to complete a predicted rule.

Left-corner parsing is particularly interesting from a cognitive science point of view, because it mirrors human performance in parsing more closely than pure top-down or bottom-up parsers. Consider the following sentences:

John's brother's dog's tail fell off. (4)

The dog chased the cat that caught the mouse that squeaked. (5)

The mouse that the cat that the dog chased caught squeaked. (6)

Sentence (4) has a *left-branching* structure, (5) a *right-branching* structure, and (6) a *center-embedded* structure. Humans do not have difficulty understanding sentences with left-branching structures like (4) or with right-branching structures like (5). However, sentences with center-embedded structures like (6) are hard for humans to process, in spite of the fact that (6) has virtually the same meaning as (5). It is plausible that the difficulty in center-embedded structures like (6) has to do with the quantity of storage space that is required to parse them. Table 1 shows a comparison between the processing complexity, for humans, of different

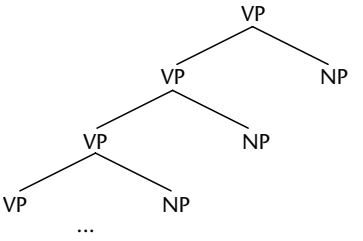


Figure 5. Endless loop due to left-recursion in a top-down algorithm.

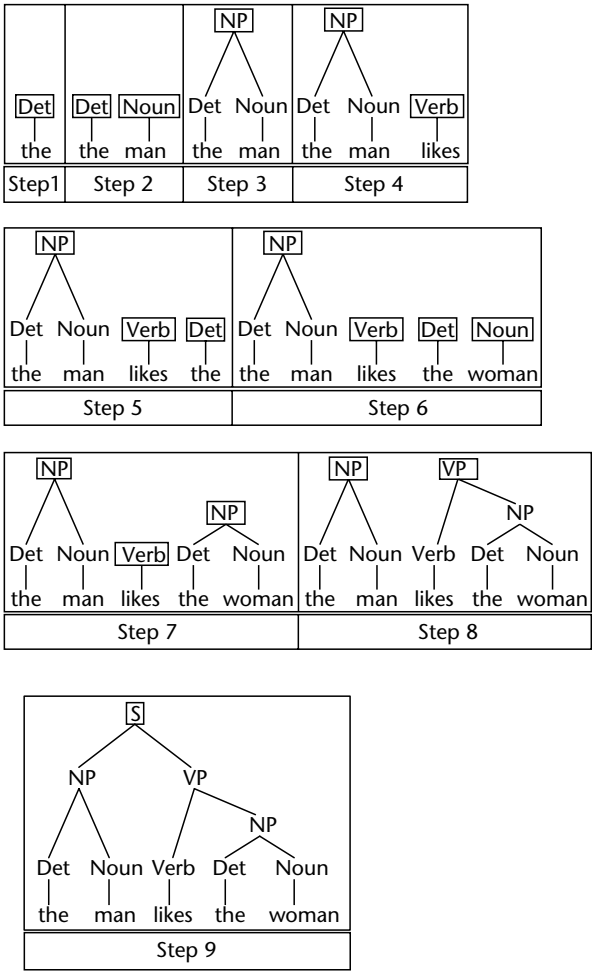


Figure 7. Bottom-up parsing step-by-step.

```
FUNCTION bottom-up-parse (SENTENCE)
  Initialize STACK = []
  DO
    Find next word W in SENTENCE
    IF (there is a rule P ⟶ W) THEN      /* shift */
      Push P onto STACK
    ENDIF
    IF (nodes on STACK = right-hand side of rule N ⟶ A B) THEN    /* reduce */
      Pop from STACK
      Push N onto STACK
      Add N above A B in the structure for the input
    ENDIF
  UNTIL STACK = [S] AND end of SENTENCE is reached
  RETURN PARSE-TREES
END bottom-up-parse
```

Figure 6. Pseudocode for bottom-up parser.

structural types, and the storage space for different parsing algorithms.

Table 1 shows that the stack size in a left-corner parsing algorithm mirrors human performance, but not in a top-down or a bottom-up algorithm. Thus, left-corner parsing algorithms are more psychologically plausible than top-down or bottom-up parsing algorithms.

The pseudocode in Figure 8 shows a left-corner algorithm in a general form.

Here is a step-by-step example parse of the sentence 'The man likes the woman' (cf. Figure 9):

- Step 1: Find 'the' and its lexical category, *Det*. Stack = [*S*]
- Step 2: Apply the rule $NP \rightarrow Det\ Noun$. Push *Noun* onto the stack. Stack = [*S Noun*]
- Step 3: Find 'man' and its lexical category, *Noun*. Pop *Noun* from the stack. Stack = [*S*]
- Step 4: Apply the rule $S \rightarrow NP\ VP$. Push *VP* onto the stack. Stack = [*VP*]
- Step 5: Find 'likes' and its lexical category, *Verb*. Pop *VP* from the stack. Stack = [ ]
- Step 6: Apply the rule $VP \rightarrow Verb\ NP$. Push *NP* onto the stack. Stack = [*NP*]
- Step 7: Find 'the' and its lexical category, *Det*. Pop *NP* from the stack. Stack = [ ]
- Step 8: Apply the rule $NP \rightarrow Det\ Noun$. Push *Noun* onto the stack. Stack = [*Noun*]
- Step 9: Find 'woman' and its lexical category, *Noun*. Pop *Noun* from the stack. Stack = [ ]

Head-corner Parsing

Head-corner parsing is a generalization of left-corner parsing. In left-corner parsing, one looks for the left-corner of the RHS of a rule to make top-down predictions. In head-corner parsing, one looks for the head of a rule. Roughly speaking, the head of a rule is the word at the semantic core of that rule. For example, in the rule $NP \rightarrow Det\ Noun$, *Noun* is at the semantic core. Thus, Noun is the head of NP. The idea behind looking at the head of a rule first is that it contains a lot of useful information (such as subcategorization, thematic roles,

Table 1. Performance of parsing algorithms and humans compared

	Sentence structure		
	Left-branching	*Center-embedded*	*Right-branching*
Stack size, top-down	unbounded	unbounded	bounded
Stack size, bottom-up	bounded	unbounded	unbounded
Stack size, left-corner	bounded	unbounded	bounded
Processing for humans	easy	hard	easy

```
FUNCTION left-corner-parse (SENTENCE)
  Initialize STACK = [S]
  DO
    Find next word W in SENTENCE
    IF (there is a rule P ⟶ W) THEN
      IF ((top-element of STACK = non-terminal N) AND ( P = left-corner
      of rule R ⟶ A B)) THEN
        Pop N from STACK
        Push B onto STACK
      ELSE IF ((STACK = []) AND (P = left-corner of rule R ⟶ A B)) THEN
        Push B onto STACK
      ENDIF
    ELSE
      Fail.
    ENDIF
  UNTIL STACK = [] AND end of SENTENCE is reached
  RETURN PARSE-TREES
END left-corner-parse
```

Figure 8. Pseudocode for left-corner parser.

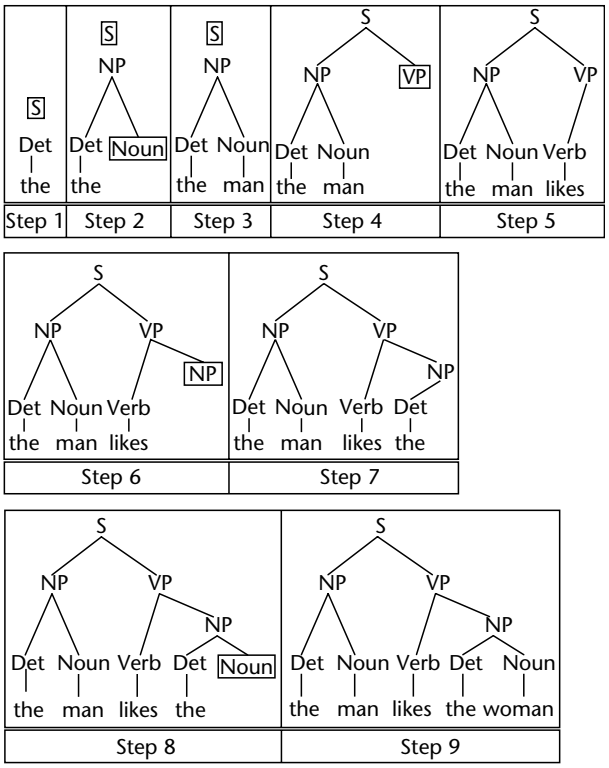


Figure 9. Left-corner parsing step-by-step.

etc.) which can allow for better top-down predictions. For instance, the lexical entry for a verb, which is the head of a VP, contains information about what NPs or PPs are needed to form a complete VP. As an example, consider the verb 'like'. Its lexical entry would contain the information that in order to form a complete VP, it needs an NP that denotes someone or something that is liked. These additional items are called *arguments*.

Having such information about head–argument relations available can improve the efficiency of parsing languages such as Japanese or German, whose word order is less constrained than that of English. Consider sentence (7):

The man likes the woman. (7)

In German, this sentence might also be written as

The woman likes the man. (8)

meaning, 'the man likes the woman'. The subject and object NPs are distinguished by morphological case marking in German. To be able to parse (7) as well as (8) correctly, a left-corner parsing algorithm would need two sets of rules. There would have to be one set of rules for (7) that includes $S \rightarrow NP\ VP$ and $VP \rightarrow Verb\ NP$. Another set of rules for (8) would have to include $S \rightarrow VP\ NP$ and $VP \rightarrow NP\ Verb$. A head-corner parsing algorithm does not require such a complex grammar. It is enough to have one set of rules that includes $S \rightarrow NP\ VP$ and $VP \rightarrow Verb\ NP$. Once a head-corner parser has found the head of the VP, 'likes', it can look for the argument of this head. This can be done bidirectionally – left-to-right or right-to-left – so that word order does not matter. The same applies to the rule $S \rightarrow NP\ VP$.

CHART PARSING

As mentioned in the introduction, human languages are extremely ambiguous. The parsing strategies discussed so far are very inefficient at handling ambiguity, whether they use a serial approach or a parallel approach. Consider the following sentence:

The tall man with brown hair saw the short woman with blond hair. (9)

Remember that 'saw' is ambiguous between a *Noun*-reading (a tool) and a *Verb*-reading (past tense of 'see'). This means that 'the tall man with brown hair' would have to be processed twice – once for the *Noun*- and once for the *Verb*-reading – although this segment gets the same structure in both parses. In realistic large-scale applications with big grammars, this can lead to efficiency problems that paralyse the whole application.

A way out of this situation is to give the parser a memory. With a memory, the parser can keep a record of all the parses it has attempted so far, and look them up instead of reparsing them. In parsing, such a memory is called *chart*. The chart keeps a record of partially as well as completely parsed rules from the grammar. In a chart, these rules are called *edges*. Partially parsed edges are called *incomplete* or *active*. Completely parsed edges are called *complete* or *inactive*. Once an edge is entered into the chart, it stays there. This is because we want to keep a record of all possible parses of a sentence, to facilitate backtracking. The following information about all edges (both active and inactive) is contained in a chart:

- the syntactic category of the edge, e.g. *NP*
- where in the sentence the edge begins (the *left end*)
- where in the sentence the edge ends (the *right end*)
- pointers to further inactive edges, e.g. to *Det Noun* for *NP*
- The following information is also included for active edges: a list of what categories are still needed to complete the active edge (to make it inactive).

A chart parser functions by combining active edges with inactive edges via the *Fundamental*

Rule. The Fundamental Rule looks for matches between categories that are needed in an active edge and the set of inactive (complete) edges. In order to start a parse, one has to specify a top-down or bottom-up strategy. This is accomplished by initializing the chart by adding either an empty active edge for an S (top-down) or inactive edges for the words in the sentence (bottom-up). The algorithm presented in Figure 10 uses a bottom-up strategy.

During the parse, when a match is found, a new edge is constructed by adding the inactive edge to the contents of the active edge. The original edges are left in the chart to allow reanalysis. The edges in the chart are labeled as follows:

```
STRUCTURE EDGE {
      CATEGORY
      CHILDREN
      LEFT-END
      RIGHT-END
      REQUIRED-CATEGORIES
}
FUNCTION chart-parse (SENTENCE)              /* the top-level function */
      initialize-chart SENTENCE, returning CHART-ACTIVES, CHART-INACTIVES,
      PARSE-TREE
      IF ((CHART-INACTIVES contains node S) AND (distance (left-corner (node S)
      AND to right-corner (node S)) = length (SENTENCE))) THEN
          RETURN PARSE-TREES
      ENDIF
END chart-parse
FUNCTION initialize-chart (SENTENCE)
      FOR (all words W in SENTENCE) DO
          add-new-edge LEXICAL-ENTRY (W)
      END
      RETURN (PARSE-TREE, CHART-ACTIVES, CHART-INACTIVES)
END initialize-chart

FUNCTION add-new-edge (EDGE)
   IF (EDGE = inactive) THEN
     FOR (all CHART-ACTIVES) DO
       fundamental-rule EDGE, CHART-ACTIVES
       add EDGE to CHART-INACTIVES
       add-null-active-edges EDGE
     END
   ELSE IF (EDGE = active) THEN
     FOR (all CHART-INACTIVES) DO
       fundamental rule EDGE, CHART-INACTIVES
       add EDGE to CHART-ACTIVES
     END
   ENDIF
END add-new-edge
```

```
FUNCTION fundamental-rule (ACTIVE-EDGE, INACTIVE-EDGE)
  IF ((left-end of INACTIVE-EDGE matches right-end of ACTIVE-EDGE) AND
  (INACTIVE-EDGE satisfies first category requirement of ACTIVE-EDGE))
  THEN
    NEW-EDGE = INACTIVE-EDGE incorporated in ACTIVE-EDGE
    IF (NEW-EDGE = active) THEN
      add NEW-EDGE to CHART-ACTIVES
    ELSE IF (NEW-EDGE = inactive) THEN
        add NEW-EDGE to CHART-INACTIVES
    ENDIF
  ENDIF
END fundamental-rule

FUNCTION add-null-active-edges (INACTIVE-EDGE)
   FOR (each rule from GRAMMAR whose rhs is initiated by INACTIVE-EDGE) DO
     NEW-ACTIVE-EDGE = rule in GRAMMAR with rhs initiated by INACTIVE-
     EDGE
     add-new-edge NEW-ACTIVE-EDGE
   END
END add-null-active-edges
```

Figure 10. Pseudocode for chart parser.

[category]/[what is already there]. [what is still needed]

Examples:

- *S/NP. VP*: category is S, an NP has been processed already, a VP is still needed, so the edge is active.
- *NP/Det Noun.* : category is NP, a Det and a Noun have already been processed, so the edge is inactive.

Active edges are printed above the words, inactive edges below. Furthermore, there are solid circles that denote stages of the parser.

Here is a step-by-step example, parsing the sentence 'The man likes the woman' (cf. Figures 11–21; to keep the figures legible, only the more important edges are shown):

- Step 1: Find 'the' and its lexical category, *Det*. Push an inactive edge, [*Det* / the .], onto chart-inactives. Chart-actives = [ ], chart-inactives = [*Det* / the .]
- Step 2: Make null-active-edge, [*NP*/. *Det Noun*]. Chart-actives = [*NP*/. *Det Noun*], chart-inactives = [*Det* / the .]
- Step 3: Apply Fundamental Rule to inactive edge, [*Det* / the .], and incorporate it into null-active-edge, [*NP*/. *Det Noun*]. Push active edge, [*NP* / the . *Noun*], onto chart-actives. Chart-actives = [[*NP* / the . *Noun*] [*NP*/. *Det Noun*]], chart-inactives = [*Det* / the .]
- Step 4: Find 'man' and its lexical category, *Noun*. Push an inactive edge, [*Noun* / man .], onto chart-inactives. Apply Fundamental Rule to inactive edge, [*Noun* / man .], and incorporate it into active edge, [*NP*/*Det* . *Noun*]. Push inactive edge, [*NP*/*Det Noun*.], onto chart-inactives. Chart-actives = [[*NP* / the . *Noun*] [*NP*/. *Det Noun*]], chart-inactives = [[*NP* / the man .] [*Noun* / man .] [*Det* / the .]]
- Step 5: Make null-active-edge, [*S*/. *NP VP*]. Chart-actives = [[*S*/. *NP VP*] [*NP* / the . *Noun*] [*NP*/. *Det Noun*]], chart-inactives = [[*NP* / the man .] [*Noun* / man .] [*Det* / the .]]
- Step 6: Apply Fundamental Rule to inactive edge, [*NP* / the man .], and incorporate it into active edge, [*S*/. *NP VP*]. Push active edge, [*S* / the man . *VP*], onto chart-actives. Chart-actives = [[*S* / the man. *VP*] [*S*/. *NP VP*] [*NP* / the . *Noun*] [*NP*/. *Det Noun*]], chart-inactives = [[*NP* / the man .] [*Noun* / man .] [*Det* / the .]]
- Step 7: Find 'likes' and its lexical category, *Verb*. Push an inactive edge, [*Verb* / likes .], onto chart-inactives. Chart-actives = [[*S* / the man . *VP*] [*S*/. *NP VP*] [*NP* / the . *Noun*] [*NP*/. *Det Noun*]], chart-inactives = [[*Verb* / likes .] [*NP* / the man .] [*Noun* / man .] [*Det* / the .]]
- Step 8: Make null-active-edge, [*VP*/. *Verb NP*]. Chart-actives = [[*VP*/. *Verb NP*] [*S* / the man . *VP*] [*S*/. *NP*

Figure 11. Chart parsing, step 1.

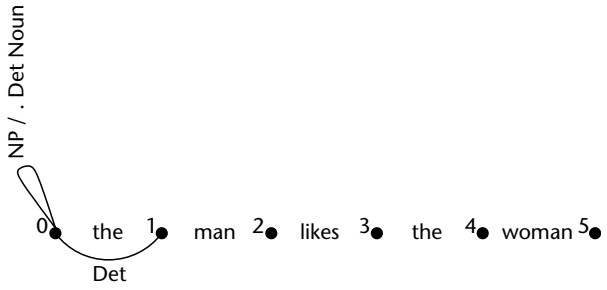


Figure 12. Chart parsing, step 2.

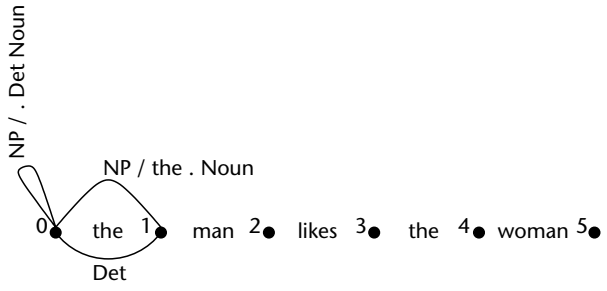


Figure 13. Chart parsing, step 3.

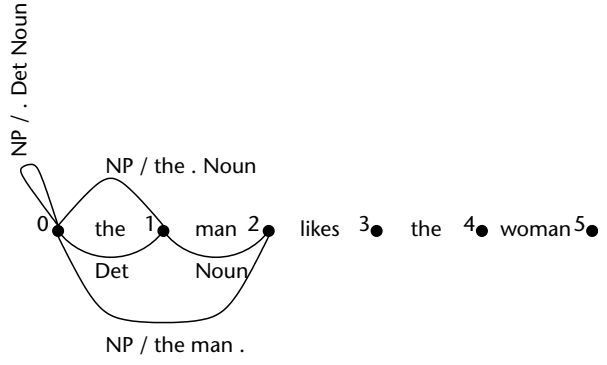


Figure 14. Chart parsing, step 4.

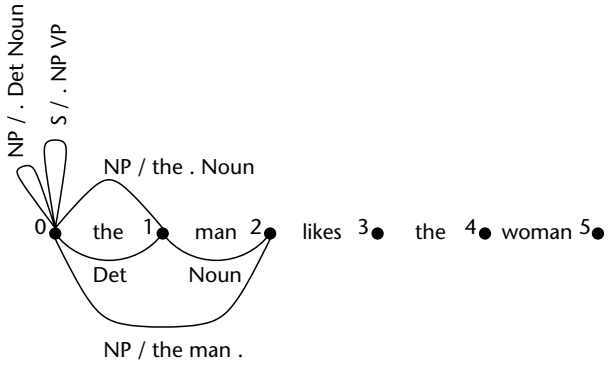


Figure 15. Chart parsing, step 5.

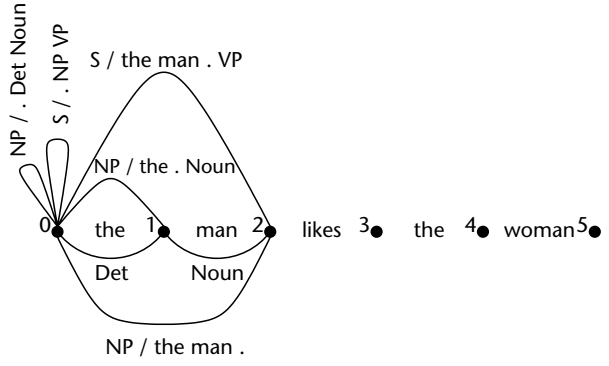


Figure 16. Chart parsing, step 6.

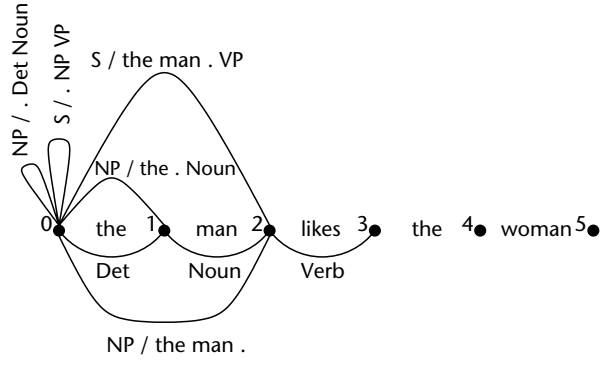


Figure 17. Chart parsing, step 7.

VP] [*NP* / the . *Noun*] [*NP* /. *Det Noun*]], chart-inactives = [[*Verb* / likes .] [*NP* / the man .] [*Noun* / man .] [*Det* / the .]]

- Step 9: Apply Fundamental Rule to inactive edge, [*Verb* / likes .], and incorporate it into active edge, [*VP* /. *Verb NP*]. Push active edge, [*VP* / likes . *NP*], onto chart-actives. Chart-actives = [[*VP* / likes . *NP*] [*VP* /. *Verb NP*] [*S* / the man . *VP*] [*S* /. *NP VP*] [*NP* / the . *Noun*] [*NP* /. *Det Noun*]], chart-inactives = [[*Verb* / likes .] [*NP* / the man .] [*Noun* / man .] [*Det* / the .]]
- Step 10: Parse *NP* 'the woman', similar to *NP* 'the man' (cf. steps 1–4). Chart-actives = [[*NP* / the . *Noun*] [*NP* /. *Det Noun*] [*VP* / likes . *NP*] [*VP* /. *Verb NP*] [*S* / the man . *VP*] [*S* /. *NP VP*] [*NP* / the . *Noun*] [*NP* /. *Det*

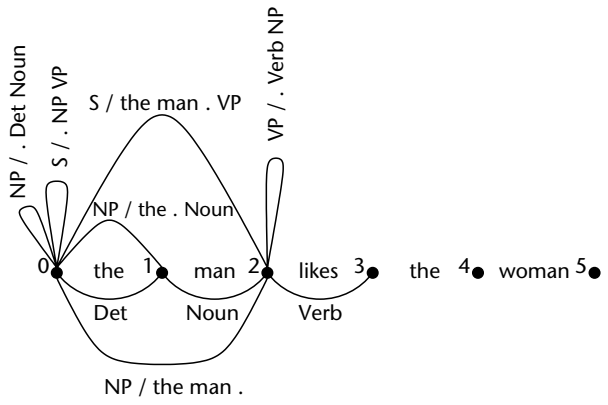


Figure 18. Chart parsing, step 8.

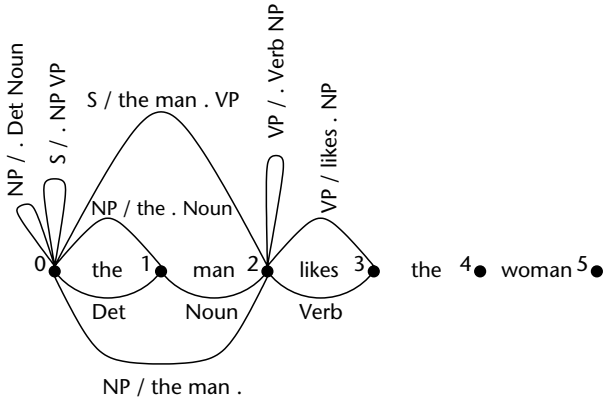


Figure 19. Chart parsing, step 9.

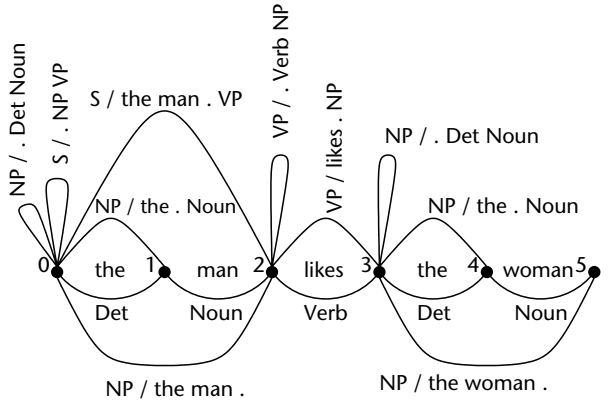


Figure 20. Chart parsing, step 10.

Noun]], chart-inactives = [[*NP* / the woman .] [*Noun* / woman .] [*Det* / the .] [*Verb* / likes .] [*NP* / the man .] [*Noun* / man .] [*Det* / the .]]

- Step 11: Apply Fundamental Rule to inactive edge, [*NP* / the woman .], and incorporate it into active edge, [*VP* / likes . *NP*]. Push inactive edge, [*VP* / likes the woman .], onto chart-inactives. Apply Fundamental Rule to inactive edge, [*VP* / likes the woman .], and incorporate it into active edge, [*S* / the man . *VP*]. Push inactive edge, [*S* / the man likes the woman .], onto chart-inactives. Chart-actives = [[*NP* / the . *Noun*] [*NP* /. *Det Noun*] [*VP* / likes . *NP*] [*VP* /. *Verb NP*] [*S* / the man . *VP*] [*S* /. *NP VP*] [*NP* / the . *Noun*] [*NP* /. *Det Noun*]], chart-inactives = [[*S* / the man likes the woman .] [*VP* / likes the woman .] [*NP* / the woman .] [*Noun* / woman .] [*Det* / the .] [*Verb* / likes .] [*NP* / the man .] [*Noun* / man .] [*Det* / the .]]

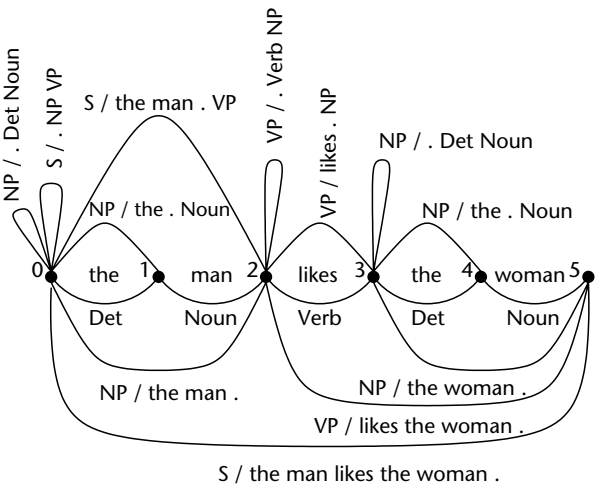


Figure 21. Chart parsing, step 11.

SUMMARY

Parsing – determining the structure of a sentence, given a grammar – is a crucial aspect of establishing meaning in language processing. There are two basic sets of constraints in parsing. One set of constraints is top-down: the rules in the grammar constrain which structures sentences can have in a given language. The other set of constraints is bottom-up: the input sentence to the algorithm constrains which rules from the grammar can apply. Parsing algorithms differ with respect to the constraints they use more prominently: bottom-up, top-down, or a combination of both (left- and head-corner).

An important issue in parsing is efficiently dealing with structural and lexical ambiguities. In an ambiguity, if the later disconfirmed alternative reading is pursued initially, a standard parser has to backtrack and re-parse parts of the input sentence. Re-parsing can be avoided by using a chart, a data structure that stores partially parsed input, so that it can be looked up during backtracking.

Further Reading

Abney SP and Johnson M (1991) Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research* **20**(30): 233–250.

Aho AV and Ullman JD (1972) *The Theory of Parsing, Translation, and Compiling*, vol. 1: Parsing. Englewood Cliffs, NJ: Prentice-Hall.

Crocker M (1999) Mechanisms for sentence processing. In: Garrod SC and Pickering M (eds) *Language Processing*. London, UK: Psychology Press.

Dowty D, Karttunen L and Zwicky A (eds) (1985) *Natural Language Processing: Psychological, Computational and Theoretical Perspectives*. Cambridge, UK: Cambridge University Press.

Grosz BJ, Jones KS and Webber BL (eds) (1986) *Readings in Natural Language Processing*. Los Altos, CA: Morgan Kaufmann.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Upper Saddle River, NJ: Prentice-Hall.

Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Pereira F and Shieber SM (1987) *Prolog and Natural Language Analysis*. Cambridge, UK: Cambridge University Press.

Parsing

Introductory article

Martin J Pickering, University of Edinburgh, Edinburgh, UK

CONTENTS

Parsing refers to the way that people assign a syntactic (grammatical) analysis to a sentence as they hear or read it.

INTRODUCTION

Our understanding of language is normally extremely efficient and often appears effortless. However, the underlying processes are extremely complex and sophisticated. They clearly make use of very specific information, as is apparent from the inability we have to perform any kind of analysis on an utterance in a language we are unfamiliar with. Researchers in the area assume that comprehension can be broken into several components, including word recognition and comprehension of entire discourses. Between such 'early' and 'late' stages lies a component that is concerned with processing combinations of words. Psycholinguists normally refer to this component as the 'parser'.

As a first approximation, parsing is concerned with assigning a syntactic analysis and a semantic interpretation to a sentence (or complete utterance). Hence, it is sometimes referred to as 'sentence processing'. The most important source of linguistic information that it employs is knowledge of syntax and semantics – in other words, the information that is discussed in most linguistic textbooks concerned with grammar. However, instead of simply describing analyses that can be assigned to sentences, a theory of parsing is concerned with understanding how people actually convert an uninterpreted string of words into a syntactic analysis and associated meaning.

Many studies of parsing involve reading or listening to sentences containing some syntactic ambiguity. Consider, for example, *The farmer hit the tramp with the stick*. According to linguistic theory, this sentence is ambiguous, not because a word has two meanings, but because the phrase *with the stick* can modify either the verb phrase *hit the tramp* or the noun phrase *the tramp*. In the former case, the sentence means that the farmer used the stick to hit the tramp; in the latter, it means that the farmer hit the tramp who had the stick. Theories of parsing attempt to determine how people analyze sentences, and they often use evidence about the processing of syntactically ambiguous sentences to do this. In order to do this, they examine what happens while (and sometimes after) people are reading or listening to such sentences.

One thing that we can now be certain of is that sentence comprehension begins almost immediately. We know this because a 'unified' interpretation is assigned to each sentence fragment as every word is encountered. For instance, people are disrupted by a word whose interpretation is incompatible with preceding context as soon as that word is encountered. But can information such as semantic plausibility be used to 'guide' parsing in cases of ambiguity? This has been the dominant question in parsing research and is motivated, ultimately, by the question of whether parsing constitutes an autonomous process that occurs before true interpretation takes place. In other words, answering this question may help resolve one aspect of the 'modularity' debate that has played such an important role in cognitive science. This article first considers the evidence for and against an autonomous parser, and then briefly considers ways in which parsing research is currently reaching beyond this question.

AUTONOMOUS ACCOUNTS

The 'classic' autonomous account is the 'garden-path' theory. On this account, the parser makes initial parsing decisions using some syntactic information and very little else. It uses a core principle known as 'minimal attachment' to decide which analysis to pursue in cases of ambiguity. This says adopt the simplest analysis, which is defined as the one involving fewest nodes in a phrase structure tree. According to Frazier's syntactic assumptions, the verb-phrase analysis of *The farmer hit the tramp with the stick* is simpler than the noun-phrase analysis, and so that analysis is initially adopted. In this

case the sentence is globally ambiguous (because the sentence has two interpretations). But the parser applies the principle as soon as an ambiguity emerges, and so the principle affects the processing of sentences that are locally ambiguous but which are eventually disambiguated. An example of a local ambiguity occurs in *The defendant examined by the lawyer turned out to be unreliable*, where *examined by the lawyer* is a 'reduced relative' that modifies *the defendant*; however, after *examined*, it might continue as a simple 'main-clause' sentence (e.g. *The defendant examined the jurors*). Because the main-clause analysis involves fewer nodes, the garden-path theory predicts that readers initially adopt this analysis. This analysis turns out to be incompatible with the words after *examined*, and so readers are 'led up the garden path': they experience processing difficulty and are forced to re-analyze (or fail to parse the sentence at all) (e.g. Frazier and Rayner, 1982).

If the two analyses have the same number of nodes, another principle, known as 'late closure', is applied. This says, when possible, attach new words into the phrase that is currently being processed. So for instance, in *John said Fred died yesterday*, the word *yesterday* is preferentially attached to *died* (i.e. Fred died yesterday). Neither minimal attachment nor late closure draws upon any information apart from phrase-structure geometry. According to garden-path theory, initial parsing pays no attention either to the relative frequency of the analyses or to whether the interpretation of one analysis is more compatible with background knowledge than the others (in other words, to plausibility). According to garden-path theory, information like frequency or plausibility can be used during subsequent processing. In other words, they can affect re-analysis, but not initial analysis. Thus, the model is called a 'two-stage' account, because the stages of initial analysis and re-analysis employ different sources of information. Additionally, notice that this theory assumes that the processor always adopts a single analysis at a time. Because alternative analyses are only ever considered if the initial analysis is abandoned, the garden-path theory is often called a serial account (i.e. one analysis at a time). Other autonomous accounts exist (including ones that make reference to thematic role assignment), but the garden-path theory is by far the best known.

Early experimental evidence provided strong support for the garden-path theory. For example, some work found that people had difficulty with sentences like *The spy saw the cop with a revolver*, where the verb-phrase analysis is implausible. This suggested that they initially adopted the verb-phrase analysis and then re-analyzed. Similarly, people appeared to have difficulty with reduced-relative sentences. But reduced-relatives are generally fairly rare types of sentence, so an alternative explanation for this result is surely that people have adopted the more frequent analysis.

INTERACTIVE ACCOUNTS

More recently, a great deal of experimental work has critiqued autonomous parsing accounts. In general, such research has attempted to show that a source of information, such as frequency or plausibility, plays a role in initial parsing decisions. This work supports 'interactive' accounts, in which all potentially relevant sources of information can be drawn upon at the same time. Although interactive accounts have a long history in language comprehension, the recent resurgence in their popularity has been associated with so-called constraint-based theories. These theories are broadly based on connectionist models and assume that language comprehension involves the integration of a large number of 'soft' constraints and that alternative analyses compete for activation. Importantly, such models assume that many analyses can be considered in parallel.

To explicate such accounts, consider a study that contrasted sentences like (a) *The defendant examined by the lawyer turned out to be unreliable* with (b) *The evidence examined by the lawyer turned out to be unreliable*. As in most good parsing research, the sentences are very closely related (only one critical word differs between them, and between other pairs of sentences used in the study). According to garden-path theory, people should initially adopt the reduced-relative analysis in both (a) and (b), because they are syntactically identical. But notice that while a defendant can examine things, evidence cannot. This means that the main-clause analysis of (b) is implausible at the word *examined*. Constraint-based accounts therefore predict that people will find it much easier to adopt the reduced relative analysis in (b) than (a). In accord with this, Trueswell *et al*. (1994) found that people were disrupted reading *by the lawyer* in (a) but not in (b). Early effects of 'referential' semantic processing have also been demonstrated, though alternative interpretations of these data are possible.

Constraint-based accounts also assume that the relative frequency of different analyses affects initial parsing preferences rather than (for example) minimal attachment. In other words, people initially prefer frequent analyses, not simple ones.

The sentence *The criminal confessed his sins harmed too many people* is locally ambiguous, in that confessed can either take a noun-phrase object (e.g. *his sins*) or a sentential complement (as turns out to be the case). If people experience difficulty reading *harmed*, this suggests that they initially interpreted *his sins* as the object, and then had to re-analyze. As the object analysis is simpler, garden-path theory predicts that it will always be adopted initially, and hence that re-analysis will always be necessary. While some recent work has supported this, other work has found that people re-analyze if the main verb more frequently takes an object than a sentential complement, but not if it more frequently takes a sentential complement.

Overall, there is good evidence that people use many relevant sources of information rapidly, but it is unclear precisely which model of comprehension is correct. Although it is by now unlikely that the garden-path model is correct, other autonomous parsing accounts may well be able to explain current data. Constraint-based accounts face challenges too (for instance to their claims that ambiguity resolution involves competition between alternatives and that lexical ambiguity resolution is a form of syntactic ambiguity resolution).

PARSING AND COMPREHENSION

The traditional focus on the nature of ambiguity resolution and its implications for the initial stages of parsing is, to some extent, being replaced by an interest in a series of broader questions about parsing and its relation to other aspects of language comprehension. One critical issue is what people do when they realize they have adopted the wrong analysis (or have preferred it, within a parallel architecture). Some recent evidence in this area suggests that some kinds of re-analysis are easier than others and that people adopt particular re-analysis strategies (e.g. when more than one alternative analysis is available). Another interesting question is whether people only re-analyze as a last resort or whether they sometimes 'bail out' of an analysis before being absolutely certain that it is incorrect.

Questions of re-analysis, however, still relate entirely to syntactic processing, that is how people adopt a syntactic analysis. However, the ultimate goal of parsing is, of course, to determine the appropriate interpretation that should be assigned to a string of words and to integrate that interpretation with discourse context and general knowledge. Thus, researchers have asked how people use context to decide on the appropriate interpretation for expressions – for example how people interpret elliptical phrases, how they should interpret a pronoun or other referring expression, and so on. It is clear that theories of parsing need to be fully integrated into more general accounts of language comprehension. This is likely to be a major focus of future research.

Further Reading

Altmann GTM (1998) Ambiguity in sentence processing. *Trends in Cognitive Sciences* **2**: 146–152.

Crocker MW (1999) In: Garrod S and Pickering M (eds) *Language Processing*. Brighton and Cambridge, MA: Psychology Press/MIT Press. [Covers basic computational issues.]

Crocker MW, Pickering MJ and Clifton C Jr (eds) *Architectures and Mechanisms for Language Processing*. Cambridge, UK: Cambridge University Press.

Frazier L and Rayner K (1982) Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* **14**: 178–210.

Gibson E and Pearlmutter NJ (1998) Constraints on sentence processing. *Trends in Cognitive Sciences* **2**: 262–268.

Harley TA (2001) *The Psychology of Language*, 2nd edn, chap 9. Hove, UK: Psychology Press.

Haberlandt K (1994) Methods in reading research. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Mitchell DC (1994) Sentence parsing. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Pickering MJ (1999) Sentence comprehension. In: Garrod S and Pickering MJ (eds) *Language Processing*. Brighton and Cambridge, MA: Psychology Press/MIT Press.

Tanenhaus MK and Trueswell JC (1995) Sentence comprehension. In: Miller J and Eimas P (eds) *Speech, Language, and Communication*, vol. 11, pp. 217–262. San Diego, CA: Academic Press.

Trueswell JC, Tanenhaus MK and Garnsey S (1994) Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language* **33**: 283–318.

Performance and Competence

Introductory article

Lyn Frazier, University of Massachusetts, Amherst, Massachusetts, USA

There is an important distinction between knowledge of the general principles of grammar (competence) and the way that knowledge is applied in actual language production or comprehension (performance).

Knowing the rules of addition is one thing. Actually adding a column of numbers is another. Likewise, knowing the rules of grammar is one thing. Actually using that knowledge to assign a structure and interpretation to a sentence is quite another. A theory of a native speaker's implicit knowledge of language is what is known as a 'grammar' or, equivalently, a theory of language competence. A theory of how that knowledge is applied to produce sentences, or to comprehend sentences, is a theory of language performance.

If one were to collapse the theory of competence and the theory of performance, a variety of problems would emerge. First, one would need to choose: should the theory characterize language production, mapping from an intended 'message' to a set of motor instructions, or should it characterize language comprehension, mapping from a perceptual input to an intended meaning? Given a theory of language competence, the well-formedness rule of grammar (the subject precedes the verb in English, the object follows the verb, proper names aren't preceded by determiners, e.g. **the Mary*) can be stated once and for all, even though this may be used for purposes of speaking, writing, listening, or reading.

Further, the theory of grammar can abstract away from issues concerning step-by-step processing in real time. The subject of a sentence must precede the verb independent of the moment-by-moment processing operations required to utter an actual sentence token or to understand one. To return to our addition example, imagine that the rules of addition were stated only in a theory of mathematical performance. In this case, someone who was capable of adding a column of numbers from top to bottom might be unable to add from bottom to top, or in groups with the even numbers added first and the odd numbers later. The step-by-step performance mechanisms would be inextricably bound up with the characterization of the mathematical principles themselves.

The grammar of English or any other natural language will define the well-formed sentences of the language (syntax), how they are pronounced (phonology), and what they mean (semantics). To account for actual language behavior, it must be determined what computation is performed at each step of processing, what information is consulted, what options are pursued at choice points, and the like. Without a detailed theory of language performance, it would be difficult to explain why some sentences are 'unacceptable' even though they are grammatical. Consider the sentences in (1):

a. The woman died.
b. The woman the man loved died.
c. The woman the man the child liked loved died. (1)

Sentence (1a) is a simple clause. In (1b), the subject 'the woman' is modified by a relative clause, producing the intelligible phrase 'the woman the man loved'. In (1c) the phrase 'the man' is also modified by a relative clause, producing 'the man the child liked'. But (1c) is usually considered unacceptable by speakers when they first encounter it. The problem is that (1c) is difficult to process even though it abides by all the same rules as the acceptable (1b). Given the competence–performance distinction, the account of (1) is straightforward: all the sentences in (1) are grammatical, but only (1a) and (1b) are acceptable. The grammatical (1c) has a structure that is difficult for perceivers to recover.

Examples like (1) have been used to argue that it would be arbitrary to rule out sentences like (1c) using grammatical mechanisms. In fact this argument can be strengthened. Consider (2):

The vase the man you met bought broke. (2)

Sentence (2) has the same grammatical structure as (1c) but it seems acceptable. Perceivers can recover the correct structure for (2) because the initial noun phrases are dissimilar, discouraging any attempt at a conjoined noun phrase structure. Further, constraints from the meaning of 'met' and 'bought' aid the perceiver. What the acceptability of (2)

shows is that limiting the number of self-embedded relative clauses (say to two) would not only be arbitrary, but it would be empirically wrong. In (1c) two is too many, but in (2) it is not. What differs between (1c) and (2) are not the grammatical properties of the structure, but factors which make the structure easier or harder for the perceiver to recover.

Examples like (3), due to Tom Bever, make a similar point. When native speakers first encounter (3), they cannot correctly identify its structure:

The horse raced past the barn fell. (3)

Typically they take 'The horse raced past the barn' as a simple sentence and then don't know what to do with 'fell'. But (4) does not present a similar problem:

The horse ridden past the barn fell. (4)

The structure in (3) and in (4) is the same. The problem in (3) is simply the ambiguity of 'raced' which, unlike 'ridden', could be a simple past tense verb. Once perceivers have encountered (4) and they realize that (3) is equivalent to 'The horse (which was) raced past the barn fell' they usually accept the sentence. Their grammar (theory of competence) has not changed, they have simply managed to recover the structure of (3).

Performance mechanisms and actual language comprehension behavior may be influenced by numerous nongrammatical factors. For example, perceivers are less likely to accept a sentence as being grammatical/acceptable if they are asked to look into a mirror while they are making grammaticality judgments. Mirrors make people self-conscious which in turn makes judgments more conservative. But this is most likely not a fact about linguistic knowledge *per se*. It is probably best explained by a theory of social psychology, not grammar.

Distinguishing between competence and performance allows us to capture the fact that grammatical knowledge is atemporal and is the same whether it is exploited for purposes of production or comprehension. The theory of competence will characterize the grammatical sentences of the language. The performance theory will account for which of these structures is easy, difficult, or impossible to produce or to understand under various circumstances, relieving the grammar of the need to make arbitrary stipulations on self-embedding, as in (1) or (2), or *ad hoc* restrictions that would be needed to capture grammatically the effects of ambiguities in particular contexts such as the 'garden-path' sentence in (3). Realizing that grammaticality/acceptability judgments are themselves a type of linguistic and cognitive behavior also eliminates what would otherwise be inconsistencies in the data (judgments) that inform theories of grammar.

Further Reading

Bever TG (1994) The ascent of the specious; or, There's a lot we don't know about mirrors. In: Cohen D (ed.) *Explaining Linguistic Phenomena*. Washington, DC: Hemisphere.

Bever TG, Katz J and Langendoen T (eds) (1976) *An Integrated Theory of Linguistic Ability*. New York, NY: Thomas Crowell.

Gibson E and Thomas J (1999) Memory limitations and structural forgetting. *Language & Cognitive Processes* **14**: 225–245.

Miller GA and Chomsky N (1963) Finitary models of language users. In: Luce R, Bush R and Galanter E (eds) *Handbook of Mathematical Psychology*, vol. II. New York, NY: John Wiley.

Schütze CT (1996) *The Empirical Base of Linguistics*. Chicago, IL: University of Chicago Press.

Phonetics

Introductory article

IM Roca, University of Essex, Colchester, UK

CONTENTS

Phonetics is the discipline concerned with the study of the articulation, acoustics, and perception of the sounds of speech.

INTRODUCTION

We all know from experience (unless we happen to be profoundly deaf) that language is manifested as sound: the speaker *speaks* (produces sound), and the hearer *hears* (receives and perceives sound). (The sign languages of the deaf have a visual manifestation, rather than a sonorous one, for obvious reasons; we will not be concerned with these languages here.) It is sound that allows us to differentiate, when we speak, 'cat' from 'kit' from 'cot' from 'cut' from 'coot', and 'lad' from 'dad' from 'mad' from 'sad' from 'pad' from 'cad' from 'fad', or, more dramatically, 'table' from 'anticonstitutionalism': we *hear* the differences between these words spontaneously, without any overt training, and we hear such differences because the speaker has *made* them in the first place. All speakers (that is, all of us when we speak) produce sound, and hearers (all of us when we hear) receive and perceive sound. It is in fact rather like music: some people play it, and others hear (and enjoy) it. *Phonetics* is the discipline that investigates speech sound in its three material stages: production by the speaker, transmission through the air, and reception by the hearer.

ARTICULATION OF SPEECH SOUND

Speech sound is invariably made in the mouth (also in the nasal and laryngeal spaces, concomitantly), not, for instance, with the fingers (unlike finger clicking). The reason for this is purely biological: we do not think about it or decide on it, nor does Parliament legislate on the matter. Indeed, babies show an irrepressible urge to use their vocal apparatus and make speech sounds, gradually shaping their output in the direction of the language of their surroundings. (Noam Chomsky has been arguing for close to half a century that language is a biological, not a cultural, phenomenon: more like laughter or hair growth than like car manufacturing.) The organs with which both babies and grown-ups make the sounds of language can be likened to the instruments with which musicians play music. Like a musical instrument, the vocal apparatus has its own physical characteristics, which can be studied. Also, the sound(s) produced by a musical instrument or by the vocal apparatus can be dissected and analyzed. With regard to language, matters like these fall within the remit of *phonetics*.

Among the principal organs of speech contained in the mouth, the tongue stands out as the chief active articulator (the articulator that moves). It produces a variety of consonantal sounds by approaching a passive articulator (an inert articulator) inside the mouth: *dental* consonants when it approaches the upper teeth, *alveolar* consonants when it approaches the ridge behind the upper teeth, *palatal* consonants when it approaches the hard palate (the bony section of the roof of the mouth, at the front), *velar* consonants when it approaches the soft palate (the fleshy section of the roof of the mouth, at the back), *uvular* consonants when it approaches the uvula (the appendage that dangles at the back of the soft palate), and *pharyngeal* consonants when it approaches the back wall of the pharynx (the very back part of the oral cavity, which connects it to the larynx, in the throat: see below). In addition, the lips can produce *bilabial* consonants (upper lip with lower lip) and *labiodental* consonants (lower lip with upper teeth). Examples of all these sounds are given below.

Not surprisingly, different parts of the tongue are involved: for any one sound, the tongue area

involved in its production faces the passive articulator. Three main tongue areas can be distinguished: the *blade* of the tongue at the front (involved in the production of dental and alveolar consonants), the *body* of the tongue in the middle (involved in the production of palatal, velar, and uvular consonants), and the *root* of the tongue bending down right at the back (involved in the production of pharyngeal consonants).

Figure 1 provides a visual representation of the vocal tract and the speech articulators.

CLASSIFICATION OF SPEECH SOUNDS

The first criterion for the classification of consonants concerns the place where the consonant is articulated: its *place of articulation*. According to this criterion, consonants are classified into bilabial, labiodental, dental, alveolar, palatal, velar, uvular, and pharyngeal, as defined above. Narrower criteria are possible: for instance, dentals can be divided into dental proper and interdental, palatals into prepalatal, palatal, and postpalatal, and so on. Notice that the labels derive from the Latin names of the corresponding (usually passive) articulators: 'dental', for instance, refers to the teeth, 'velar' to the soft palate.

So far we have been considering the action of the active articulator on the passive articulator as one of 'approaching'. We must now be more precise. There are two main ways in which the articulators can produce consonants. The first involves the active articulator making contact with the passive articulator and stopping the airflow from the lungs: the corresponding consonants are known as *stops* (e.g. '**b**ay'). The alternative manner of articulation involves the creation of a narrow channel between the two articulators (crucially, without stopping the air): such action results in the production of

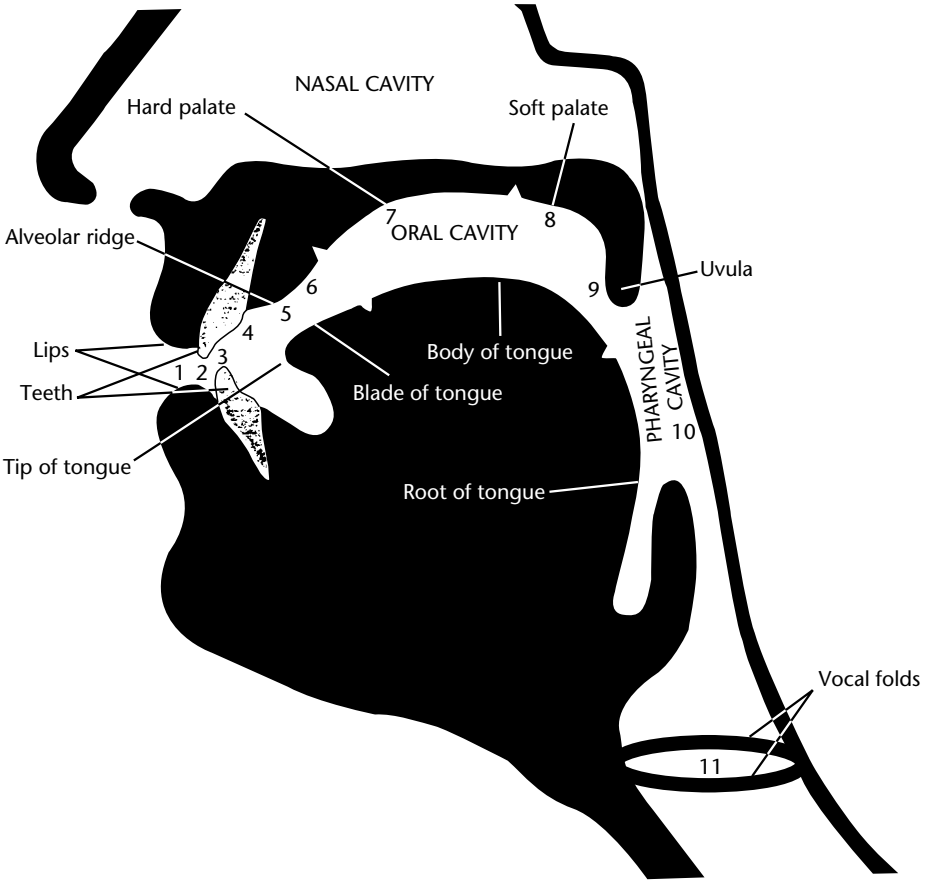


Figure 1. Outline of the vocal tract, with the speech articulators and the places of articulation: (1) bilabial, (2) labiodental, (3) interdental, (4) dental, (5) alveolar, (6) palatoalveolar, (7) palatal, (8) velar, (9) uvular, (10) pharyngeal, (11) glottal. Adapted from Roca and Johnson, *A Course in Phonology*, first reprint, 2000, p. 726, by kind permission of Blackwell Publishers.

fricative noise, and the corresponding consonants are known as *fricatives* (e.g. '**s**ay'). A handful of consonants, known as *affricates*, involve the production of both a stop and a fricative in rapid succession, in this order ('**ch**air'). The term *approximant* is sometimes applied to sounds with a wider channel than fricatives and therefore no fricative noise ('**w**ay'). These sounds resemble vowels, and may indeed be simple instances of them. *Manner of articulation* is the second criterion for the classification of consonants.

The tongue is, of course, also involved in the production of vowels. The difference between vowels and consonants is in essence one of manner of articulation: while, as we have seen, the production of consonants involves the two articulators either blocking the air (for stops) or restricting it (for fricatives), the production of vowels involves neither action. In particular, while vowels are also produced by the movement of an active articulator (more specifically, a section of the tongue) towards a passive articulator, the distance between the articulators remains sizeable, such that no noise is generated. From this perspective, vowels can be construed as 'sounds' with a wider articulatory channel than consonants (width of articulatory channel is what the traditional label 'manner of articulation' essentially refers to), in particular a channel wide enough to prevent the production of noise. According to degree of channel opening, vowels are classified into *close* (or *high*) with the narrowest channel, *open* (or *low*) with the widest channel, and *mid* with intermediate channel width.

The nose and the larynx can also participate in the production of speech sounds. A raised soft palate (or velum) keeps the nasal cavity shut off from the oral cavity. If the velum is lowered, however, the air coming out of the lungs will also flow through the nasal cavity, where it will resonate and add a particular nasal quality to the sound: it will make it a *nasal* sound. Nasality (or its converse, *orality*) constitutes the third criterion for the classification of consonants, and can also be a component in the classification of vowels.

The last criterion for the classification of speech sounds involves the state of the vocal folds, in the larynx. The larynx is the voice box housed in the throat (it can protrude as the Adam's apple, particularly in males). It constitutes the upper section of the trachea (the windpipe by which the mouth cavity connects to the lungs). Across the larynx lies a membrane with a slit-shaped back-to-front opening in the middle controlled by two small lips, known as the vocal folds (or vocal cords): they close up automatically to prevent food falling into the trachea and causing death by choking, among other physiological functions. When the vocal folds are relaxed and held relatively close to each other, a fast airflow will cause them to open and close in rapid succession, in an action reminiscent of a flag flapping in the wind. The effect of this vibration is known as 'voice' (in effect, a form of humming). Consonants can be *voiced* (i.e. have voice superimposed on their articulation) or *voiceless* (with no voice). Vowels are usually voiced, but can also be voiceless (in that case they sound as if they were whispered).

The criteria we need to classify speech sounds from the viewpoint of their articulation are summarized in Table 1.

PHONETIC TRANSCRIPTION

So far, we have looked at examples in ordinary spelling, which (particularly in English) is a very poor index of the actual sound produced by a speaker. The following anonymous rhyme provides a convenient illustration of the problem:

Hints on Pronunciation for Foreigners

I take it you already know
Of tough and bough and cough and dough?
Others may stumble, but not you,
On hiccough, thorough, lough and through?
Well done! And now you wish, perhaps,
To learn of less familiar traps?

Beware of heard, a dreadful word
That looks like beard and sounds like bird,
And dead: it's said like bed, not bead –
For goodness sake don't call it 'deed'!
Watch out for meat and great and threat
(They rhyme with suite and straight and debt).

A moth is not a moth in mother,
Nor both in bother, broth in brother,
And here is not a match for there
Nor dear and fear for bear and pear;
And then there's dose and rose and lose –
Just look them up – and goose and choose,

And cork and work and card and ward,
And font and front and word and sword,
And do and go and thwart and cart –
Come, come, I've hardly made a start!
A dreadful language? Man alive!
I'd mastered it when I was five!

It is clear that if we are to keep an accurate written record of sound we must avail ourselves of a different system.

Table 1. Main classification criteria for the sounds of speech

Place of articulation							
Bilabials **p**ea	Labiodentals **f**ee	Dentals **th**ee	Alveolars **t**ea	Palatals **h**ue, **y**es	Velars **c**ow	Uvulars **r**iz 'rice' (French)	Pharyngeals ħuut 'whale' (Arabic)
Manner of articulation							
Stop consonants **p**ea	Fricative cons. **f**ee		Close/high vowels k**ey**			Mid vowels c**au**ght = c**our**t	Open/low vowels c**a**r
Nasal activity							
Nasals **m**e	Orals **b**ee						
Laryngeal activity							
Voiced **z**oo	Voiceless **S**ue						

The most widely used system is known as the alphabet of the International Phonetic Association (IPA). It is similar to conventional English spelling in as much as it is made up of letters (not, for example, ideographic symbols, along the lines of traditional Chinese writing: ideographic symbols correspond to ideas, rather than to sounds). Moreover, the symbols of the IPA alphabet are mostly taken from the Roman alphabet (like the letters used in English spelling), although some are taken from the Greek alphabet, and some have been made up specifically for the purpose. The advantage of the IPA alphabet (like any other phonetic alphabet) is that each symbol corresponds by definition to only one sound, and, conversely, each sound is represented by only one symbol. So, for instance, every time we see the symbol [h] (phonetic symbols are conventionally enclosed in square brackets) we know that we are referring to the first sound in words like '**h**ot', '**h**eat', or '**h**at' in standard English accents (not in accents that characteristically drop this sound, like London Cockney or Yorkshire, which are accordingly described as *h*-dropping). Therefore, the phonetic representation [ph] will correspond to the sequence [p] (as in 's**p**ace' or 'sus**p**ect') + [h] (as in '**h**ot' or '**h**at'): crucially, it does not correspond to the sound sometimes represented as *ph* in English spelling, which is in fact usually [f] (cf. '**ph**iloso**ph**y' or '**ph**onetics'; but [v] in Ste**ph**en!). Indeed, the phonetic sequence [ph] corresponds to the sound represented by the first letter in words like '**p**ie', '**p**et', or '**P**eter', an aspirated *p*; i.e. a [p] followed by an [h] (as against the plain [p] of 'sus**p**ect').

By way of illustration, consider the phonetic transcription in 'Received Pronunciation' (RP, the Southern English prestige accent of Britain) of some of the words with spelling–sound mismatches contained in the rhyme reproduced above (':' represents lengthening of the preceding sound, and the underscript '̥' devoicing of the sound represented by the symbol above it):

tough : [thɐf] bough : [b̥a͡ʊ̯]
cough : [khɒf] dough : [d̥ə͡ʊ̯]
hiccough : [hɪkɐp] thorough : [θɐɹə]
lough : [lɒx] through : [θɹuː]

We are now in a position to tabulate the main phonetic symbols, exemplified for English sounds in conventional spelling to facilitate recognition (Tables 2 and 3). The spelling correspondent is shown in bold when the sound it represents is sufficiently close in quality (in RP or some other main accent) to the sound the phonetic symbol is meant to stand for, and underlined when they only exhibit some resemblance.

UNIVERSAL INVENTORY OF SOUNDS

The sounds of English obviously do not exhaust the inventory of the sounds found in the world's languages. Indeed, some speech sounds are strikingly exotic to the ear of the English speaker. For instance, some languages spoken in Southern Africa contain clicks in their consonant inventories; that is, sounds similar to the ones we use to gee up horses (*tlk-tlk*, impressionistically) or to express

Table 2. Phonetic symbols for some common consonants[a]

	Bilabial	*Labiodental*	*Dental*	*Alveolar*	*Post-alveolar*	*Palatal*	*Velar*	*Uvular*	*Pharyngeal*	*Glottal*
Stop	p: s**p**y b: ab**b**ey			t: s**t**y d: ad**d**er		c ɟ	k: s**k**y g: ea**g**er	q ɢ		ʔ: pi**t**y[b]
Fricative	ɸ β	f: **f**ame v: **v**ain	θ: **th**igh ð: **th**y	s: **s**ea z: **z**oo	ʃ: bu**sh** ʒ: rouge	ç: h**u**e ʝ	x ɣ	χ ʁ	ħ ʕ	h: **h**ead ɦ: a**h**ead
Affricate (central)				ʦ ʣ	ʧ: **ch**in ʤ: **g**in					
Affricate (lateral)				ɬ ɮ						
Nasal	m: so**m**e	ɱ: ny**m**ph		n: so**n**		ɲ: o**ni**on[c]	ŋ: su**ng**			
Lateral				l: **l**ung		ʎ: mil**li**on[c]				
Trill				r: **r**ay[d]						
Tap				ɾ: pi**t**y[e]						
Approximant	w: **w**ell			ɹ: **r**ay		j: **y**es	w: **w**ell			

[a]Where the box has two lines, the consonant in the top line is voiceless, and that in the bottom line voiced; where the box only has one line, the consonant is voiced.
[b]In London Cockney, for instance.
[c]In fast, unguarded speech, in some accents.
[d]In traditional Scottish pronunciation.
[e]Typically in American English.

Table 3. Main phonetic symbols for vowels[a]

	Front	*Central*	*Back*
High/close	i: b**ea**d y	ɨ ʉ	ɯ u: b**oo**m
	ɪ: b**i**d ʏ		 ʊ: p**u**t
Mid-close/high	e: b**a**de ø	ɘ ɵ	ɤ o: b**o**de[b]
		ə: **a**bout	
Mid-open/low	ɛ: b**e**d œ	ɜ: b**i**rd ɞ	ʌ ɔ: p**o**rt
	æ: b**a**d[c]	ɐ: b**u**d[d]	
Low/open	a: b**a**d[c] Œ	(a: b**i**te)[e]	ɑ: p**a**rt ɒ: p**o**t

[a]Where the box has two lines the vowel in the top line is unrounded, and that in the bottom line rounded; where the box only has one line, the vowel is unrounded.
[b]In American English, rather than RP.
[c]Older RP [æ] now tends to be pronounced [a].
[d]Current RP English [ɐ] in words like 'bud' is often (mis) transcribed [ʌ].
[e]The IPA alphabet does not include a special symbol for the low central vowel, inconveniently (this is probably the most common vowel in the world's languages); in its absence, the symbol for the low front vowel, [a], is often made to do double work, and sometimes marked diacritically as [ä] when meant as central.

disapproval to children or even grown-ups (*tut-tut*). Some languages contain what is known as *implosive* consonants; that is, stop consonants made with air going *into* the mouth, rather than out, as is the case in English and generally. Some languages have *ejective* consonants; that is, consonants which incorporate a glottal stop, a stop sound made with the vocal folds (glottal stops are a common pronunciation of intervocalic (between-vowel) *t* in several accents of English, London Cockney, for instance: *pity* = [phɪʔi]). French has nasal vowels (vowels pronounced with a lowered velum), as in *main* 'hand' [mɛ̃] (making up a minimally contrasting pair with *mai* 'May' [mɛ]), and also front vowels with rounded lips, as in *vue* 'sight' [vy] (contrasting with *vie* 'life' [vi]), rendered [vi̯uː] by English speakers in a name like 'Bellevue'. And so on.

ACOUSTIC AND AUDITORY PHONETICS

Our focus so far has been on articulatory phonetics for at least two reasons. First, the articulation of sounds is accessible to plain observation (particularly self-observation, but to some extent also observation in others). Second, articulation is usually thought to provide the bridge with the sister discipline of phonology. There are, however, two other areas of phonetics which must be mentioned here, if only cursorily. (*See* **Phonology**)

The first is *acoustic phonetics*, which deals with the intrinsic physical properties of sound. In particular, what we call 'sound' in general corresponds to the

vibration of molecules (their to-and-fro movement) in the surrounding air. In turn, each speech sound corresponds to a specific pattern of such vibration, brought about by the corresponding articulatory gestures. The patterns of air movement can be captured instrumentally and displayed visually, a technique that greatly facilitates investigation into acoustic phonetics.

The other branch of phonetics, *auditory phonetics*, deals with the mechanisms responsible for the perception of sound in the hearer's brain. First, the soundwave hits the ear-drum, a membrane stretching across the inner entrance of the ear. Second, the vibration moves on through a chain of three tiny ear bones. Third, this mechanical vibration is converted into pressure differentials in a special fluid that fills the cochlea, a snailshell-like structure connected on the outside to the last of the three ear bones and divided lengthways inside by the basilar membrane, from which stem a number of 'hair cells' set in motion by the pressure differentials. Finally, the movement of the hair cells converts to electrochemical signals which are perceived by the brain. (*See* **Phonology and Phonetics, Acquisition of**)

SPELLING AND SOUND

Before ending, we must return to the relationship between sound and spelling, to dispose of any remaining misconception. Indeed, some may think that words like 'cat' and 'cot', or 'pad' and 'fad', are different, not so much because they *sound* different, but because they are *spelled* differently: their difference in sound would simply be a consequence of their difference in spelling. That is not correct.

First, each word of a given language usually has a constant spelling (the differences in spelling between America and the UK are absolutely minimal, and in any event unconnected with the pronunciation): 'cat' is always spelled *cat* in English, and 'tonsillitis' *tonsillitis*. And yet, when each of these or other words is spoken there are obvious differences in the way it sounds, not just subtle voice-related differences from one individual to the next (and, more systematically, from a child to a grown-up, from a woman to a man, from someone with a cold to someone without, etc.), but rather gross differences related to accent. For instance, the *t* of the English word 'writing' can sound similar (if not identical) to the *t* of 'tingle', or be weakened and sound very similar to a *d* (in a typical North American accent, among others), or be realized as a glottal stop (in a London Cockney accent, for instance). And besides the usual weak sound for English *r*, there is a stronger one in Scottish English, and one similar to *w* in the pronunciation of some individuals (most commonly children). Next, the sequence *ng* in *-ing* can sound as a velar *n* [ŋ] (remember, an *n* pronounced with the body of the tongue on the soft palate), or as an alveolar *n* [n] (the *n* of words like 'tin', or, for that matter, 'knit'), or as a sequence *n* + *g* [ŋg] (this would be the typical Birmingham or Liverpool pronunciation, for instance). Finally, the first *i* of 'writing' typically sounds like 'eye', but can be less open (in Scottish or Canadian English, for instance), or sound more like the *a* of 'spa' (e.g. in the proverbial Southern US states pronunciation). So, the only letter in 'writing' that seems to have a more or less uniform phonetic correspondent throughout the English-speaking world is the *i* of *-ing* (the *w*, of course, is silent everywhere).

What these simple examples confirm is that the spelling is not really a mirror of the pronunciation, or conversely. Rather, the spelling of a word is a conventional symbolic representation that we associate with the pronunciations of the word (notice the plural!), similar to the way we associate the letter 'tee' with its multiple materializations on paper: small roman 't', capital roman 'T', small italic '*t*', capital italic '*T*', courier 'ｔ', monotype corsiva *t*, century gothic 't', and so on, not to mention its multiple individually based implementations in longhand writing.

There are many other facts that incontrovertibly show that pronunciation and spelling are not just two sides of one coin. Many English words that sound the same (at least in most accents) have a different spelling ('homophones'): 'road' versus 'rode', 'flour' versus 'flower', 'peer' versus 'pier', 'tail' versus 'tale', 'earn' versus 'urn', 'course' versus 'coarse', 'loan' versus 'lone', etc. Conversely, some words with the same spelling sound different ('homographs'): 'row' (a line or a quarrel), 'bass' (a fish or a singer), 'wind' (air in motion or to bend). Some words are spelled the same in English and in other languages, but the sound is substantially different, and conversely: 'sublime', 'admirable', and 'suave' are [səblá͡im, syblím, sublíme], [ædmɪɹəbl, admɪʁábl, admiɾáble], [suάːv, syáv, suábe], in English, French, and Spanish, respectively, while [bo͡i] and [dan] are spelled *voy* 'I go' and *dan* 'they give' in Spanish, but 'boy' and 'done' in English (on a broad phonetic interpretation: a more rigorous English transcription would be [b̥ɔ͡ɪ] and [d̥ʊn]), and [kʰiː] can correspond to English 'key' or 'quay', or to French *qui* 'who' (again, broadly interpreted). Some languages, like Chinese, do not use letters to represent their words,

but, rather, stylized pictures which originally bore a (more or less abstract) relation to the meaning, rather than to the sound. Last, and decisively, speech invariably precedes writing in both children and mankind as a whole. Indeed, some people remain illiterate for the whole of their lives: how could such individuals pronounce anything (and therefore speak) if sound were only a manifestation of spelling? The conclusion we must draw is that, interesting and important as it is for cultural and educational reasons, spelling ultimately need not have much to do with the sound of language.

Further Reading

Ball MJ and Rahilly J (1999) *Phonetics: The Science of Speech*. London, UK: Arnold.

Carney E (1994) *A Survey of English Spelling*. London, UK: Routledge.

Catford JC (1977) *Fundamental Problems in Phonetics*. Edinburgh, UK: Edinburgh University Press.

Catford JC (1988) *A Practical Introduction to Phonetics*. Oxford, UK: Oxford University Press.

Clark J and Yallop C (1995) *An Introduction to Phonetics and Phonology*. Oxford, UK: Blackwell.

Gimson A (1994) *An Introduction to the Pronunciation of English*. London, UK: Edward Arnold.

International Phonetic Association (1999) *Handbook of the International Phonetic Association*. Cambridge, UK: Cambridge University Press.

Ladefoged P (1993) *A Course in Phonetics*. Fort Worth, TX: Harcourt Brace Jovanovich.

Ladefoged P (2000) *Consonants and Vowels*. Oxford, UK: Blackwell.

Ladefoged P and Halle M (1988) Some major features of the International Phonetic Alphabet. *Language* **68**: 577–582.

Ladefoged P and Maddieson I (1996) *The Sounds of the World's Languages*. Oxford, UK: Blackwell.

Laver J (1994) *Principles of Phonetics*. Cambridge, UK: Cambridge University Press.

Pullum G and Ladusaw W (1996) *Phonetic Symbol Guide*. Chicago, IL: Chicago University Press.

Roca I and Johnson W (1999) *A Course in Phonology*. Oxford, UK: Blackwell. [Chapters 1, 3, 5, 7.]

Sampson G (1985) *Writing Systems*. London, UK: Hutchinson.

Smith NV (1999) *Chomsky*. Cambridge, UK: Cambridge University Press.

Phonological Encoding of Words

Intermediate article

Antje S Meyer, University of Birmingham, Birmingham, UK
Linda R Wheeldon, University of Birmingham, Birmingham, UK

CONTENTS

During phonological encoding of a word the speaker retrieves the word's sound form. It consists of two representational tiers: an ordered set of segments and a metrical representation capturing the word's syllable structure and stress pattern.

INTRODUCTION

In this article, we will first explain how phonological encoding is related to other processes involved in the production of words (for details see Levelt *et al*., 1999). We then turn to the description of the two main layers of representation generated during phonological encoding – the segmental representation, which consists of phonological segments, and the metrical representation, which captures the word's syllable structure and stress pattern.

WORD PRODUCTION: AN OVERVIEW

In order to produce a word, a speaker must decide what to say, retrieve the appropriate word from the mental lexicon, and pronounce the word. The retrieval of a word from the mental lexicon is often thought to involve two steps: the selection of a lemma, which specifies the syntactic properties of the word (e.g. whether it is a noun or a verb), and the retrieval of the phonological form. An important argument for the distinction between lemma and word-form retrieval is the occurrence of tip-of-the-tongue states, in which speakers can retrieve grammatical information about a word but not the corresponding word form (e.g. Vigliocco *et al*., 1997). Some models assume that lemma and word-form retrieval are serial processing stages, which implies that first one lemma is selected and then the corresponding phonological form is retrieved. Other models assume cascading processing, which implies that several lemmas can be simultaneously activated and can retrieve their word forms in parallel. Some cascading models assume feedback from lower to higher processing levels. In such models, the availability of word-form information can affect the selection of lemmas. The available experimental evidence concerning the flow of information between lemmas and word forms is inconclusive (for a comparison of the models see Rapp and Goldrick, 2000).

Word-form retrieval consists of two steps: the retrieval of morphological and of phonological information. For many words (e.g. 'hand', 'umbrella', 'generate') only one morpheme is to be retrieved, but for others (e.g. 'handbag', 'generated') two or more morphemes are to be retrieved and combined. Reaction time experiments have revealed two important facts about morpheme retrieval. First, it is frequency sensitive, that is, speakers can retrieve frequently occurring morphemes faster than less frequently occurring ones. Second, it is sequential, that is, speakers retrieve morphemes in sequence, as they appear in the utterance (for a review of the evidence see Levelt *et al*., 1999).

The second step during word-form retrieval is the generation of the phonological code of the word. During this process segmental and metrical information are retrieved and combined. The output of phonological encoding is an abstract representation of the sound form of the word. This representation is the input for the next set of processes, phonetic encoding. They generate a more detailed context-specific representation that governs the articulatory movements carried out during speaking. Below we consider the process of phonological encoding in some detail. It is important to keep in mind that most of the available empirical work stems from studies of Indo-European languages. Further research will have to show whether the conclusions that seem warranted for these languages will generalize to languages with fundamentally different phonological systems (e.g. tone languages or languages with less variable syllable structures).

THE SEGMENTAL REPRESENTATION

The retrieval of a morpheme is followed by a process often called segmental spell-out. During this process, the morpheme is decomposed into individual phonological segments and perhaps certain clusters (e.g. /st/, /sp/). The most compelling evidence for phonological decomposition stems from analyses of speech errors (for further discussion of the speech error evidence described here see Fromkin, 1971; Shattuck-Hufnagel, 1983). Speakers often make sound errors, such as (1) and (2) below, in which the intended word and the word that is actually produced differ by a single segment. These errors show that at some point morphemes are decomposed into segments – if they were treated as units throughout the word-planning process, segmental errors could not arise.

it's a real mystery → ... a meal mystery (1)

gone to seed → god to seen (2)

the heater switch → the sweeter hitch (3)

big and fat → pig and vat (4)

infantry men → intry men (5)

at an early period → at a pearly period (6)

(all errors from Fromkin, 1971)

To understand the significance of segmental errors, two other results of speech error analyses must be considered. First, about 90 per cent of all sublexical errors involve single segments or clusters of two adjacent segments, usually syllable onsets, as in (3) above. Errors like (4) involving single features (such as voicing or place of articulation) or complete syllables (like (5), where a syllable has been deleted) do arise, but, compared to segmental errors, they are extremely rare. Thus, the most important processing units at the phonological level appear to be segments, not syllables or features. It should be noted, however, that phonological features must be represented in some way. This is because the segments interacting in errors tend to be phonologically very similar. In most cases, they differ by only one feature. In addition, syllabification (to be discussed below) and some phonological rules refer to phonological features. If features were not represented, these processes could not apply.

The second important result of speech error analyses is that the error outcomes are usually phonotactically well formed, that is, they rarely result in sound sequences that are pronounceable but are not permitted in the language. Moreover, sometimes the preceding context is changed after an error. This is illustrated in (6) above, where the error consists of the addition of the onset consonant, and the preceding determiner is changed from 'an' to 'a', as appropriate for the new consonantal onset. The well-formedness of sound errors is important because it shows that the errors do not arise during phonetic encoding or articulation, but must arise earlier, during phonological encoding, such that phonetic rules still have a chance to apply after the error has occurred. Thus, the errors can be taken as evidence for decomposition during phonological processing rather than during later processes.

Further evidence for the decomposition of morphemes into segments stems from reaction time experiments. These experiments have also shown that the segments of a morpheme are likely to be activated simultaneously. However, their association to metrical frames (to be discussed below) is a sequential process, proceeding from the beginning of the word to the end (e.g. O'Seaghdha and Marin, 2000).

THE METRICAL REPRESENTATION

Word forms are not just strings of segments. Instead segments are grouped into syllables, which are stressed or unstressed. In line with current linguistic theory, most theories of word production represent the syllabic structure and stress pattern of words on a separate tier from the segmental information. Syllables are often viewed as frames with slots, corresponding to syllable constituents. The main constituents of a syllable are the consonants preceding the vowel, which form the onset (e.g. /b/ in 'bin', /sp/ in 'spin'), and the rhyme. The rhyme can be further divided into a vocalic nucleus (/I/ in the examples) and a postvocalic coda (/n/ in the examples). During word-form encoding, segments and syllable frames are independently retrieved and then the segments are associated to positions in the syllable frames. This view is supported by another important result of speech error analyses, often called the syllable position constraint. This is the observation that misplaced segments almost always move from their target positions to corresponding positions in other syllables. Most commonly, consonants that were intended to be onset consonants assume new onset positions (as in (1) and (3) above). They rarely move to a coda position as in (7):

fish → shiff (7)

Theories of word-form retrieval differ in the types of frames they postulate. The options include frames with the syllable positions onset and rhyme, or the positions onset, nucleus, and coda (Dell, 1986), and so-called CV-frames with separate positions for each consonant and vowel (Stemberger, 1990). It is difficult to discriminate between these options on the basis of results of speech error analyses because most errors involve word onset consonants, which move to new word onset positions. To determine the nature of the frames one would have to inspect word-internal and syllable-internal errors, which are very rare. The available experimental evidence on the nature of the frames is likewise inconclusive (see Levelt *et al*., 1999, for further discussion).

According to many models, metrical information is stored in the mental lexicon for all lexical entries. However, Levelt *et al*. (1999) argued that the syllable structure of all words can be derived by rule from segmental information. (Essentially, each vowel or diphthong is assigned to a different syllable, and consonants are treated as syllable onsets whenever permitted by the phonotactic rules of the language.) In addition, many words are stressed according to simple default rules. For example, there is only one way to stress monosyllabic words, and most disyllabic English words are stressed on the first syllable with a full vowel. Thus, Levelt *et al*. proposed that the lexical entries for most words consist only of ordered sets of segments. During phonological encoding, the segments are syllabified and assigned stress according to the rules of the language. Only for words with exceptional stress patterns, metrical information is stored and combined with segmental information during phonological encoding. As Levelt *et al*. pointed out, the syllable position constraint on sound errors can easily be accounted for without assuming stored syllable frames. It could arise because segmental errors usually involve word onsets and because misplaced segments tend to replace phonetically similar rather than dissimilar segments.

In sum, all models of phonological encoding assume that morphemes are decomposed into segments, and that the string of segments is syllabified and assigned stress. The result of phonological encoding is an abstract phonological representation consisting of stressed and unstressed syllables. This representation is likely to be the representation speakers attend to in inner speech and when they monitor their planned speech for errors. It is the input to the next set of processes, phonetic encoding, which compute the articulatory gestures to be carried out. For frequent syllables there may exist pre-assembled packages of articulatory gestures (e.g. Levelt and Wheeldon, 1994).

One may ask why morphemes are first decomposed into segments and subsequently reassembled into syllables. The likely reason is that morphemes can be pronounced in many different ways, depending on the context in which they appear. For instance, 'hand' may lose its final consonant in 'put your hand down', and it may be pronounced with a final /m/ in 'handbag'. The word 'hand' corresponds to a syllable in 'I will hand Kate the book', but not in 'I am handing you the book'. More generally, when words are produced in context, phonological information is retrieved from the mental lexicon as described above. But the retrieved word forms are not just concatenated but are combined into new phonological units. The smallest unit is the syllable, already discussed above. The next larger unit is the phonological word. Phonological words often correspond to lexical words, that is, the forms stored in the speaker's mental lexicon. However, morphologically complex words (such as 'handbag') may comprise several phonological words, and unstressed closed class items (e.g. conjunctions such as 'and', pronouns such as 'it', and bound morphemes such as '-s' in 'shoes' or '-ed' in 'played') combine with a neighboring content word (i.e. a noun, verb, or adjective) into a single phonological word. Importantly, phonological, rather than lexical, words are the domain of syllabification and of the application of many phonological rules. Thus, when a speaker says 'find it', two morphemes are retrieved and combined to form one phonological word. The string of segments stemming from both morphemes is syllabified, yielding [fein] [dit]. Thus, syllables can, and often do, straddle the boundaries of lexical words. In many cases, the decomposition of morphemes and the re-assembly into phonological forms are not vacuous processes but yield phonological forms that differ from those stored in the mental lexicon. A likely reason why these 'connected speech forms' are generated is that they are easier to articulate than concatenations of stored phonological forms would be.

This article concerned phonological encoding in speech production. Phonological representations are, of course, also generated during spoken language comprehension and they play an important role in reading. The structural properties of the phonological representations involved in these different tasks are likely to be similar, but the representations are generated in fundamentally different ways. In speaking, the immediate input to phonological encoding is a morphological representation,

which in turn is generated on the basis of syntactic and semantic information. By contrast, in reading, phonological representations are derived on the basis of the orthographic input, and during auditory speech processing, they are accessed on the basis of the continuous speech signal. (*See* **Reading and Writing**; **Word Recognition**)

References

Dell GS (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–321.

Fromkin VA (1971) The non-anomalous nature of anomalous utterances. *Language* **47**: 27–52.

Levelt WJM, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 1–75.

Levelt WJM and Wheeldon L (1994) Do speakers have access to a mental syllabary? *Cognition* **50**: 239–269.

O'Seaghdha P and Marin JW (2000) Phonological competition and cooperation in form-related priming: sequential and nonsequential processes in word production. *Journal of Experimental Psychology: Human Perception and Performance* **26**: 57–73.

Rapp B and Goldrick M (2000) Discreteness and interactivity in spoken word production. *Psychological Review* **107**: 460–499.

Shattuck-Hufnagel S (1983) Sublexical units and suprasegmental structure in speech production planning. In: MacNeilage PF (ed.) *The Production of Speech*, pp. 109–136. New York: Springer.

Stemberger JP (1990) Wordshape errors in language production. *Cognition* **35**: 123–157.

Vigliocco G, Antonini T and Garrett MF (1997) Grammatical gender is on the tip of Italian tongues. *Psychological Science* **8**: 314–317.

Further Reading

Dell GS, Burger LK and Svec WR (1997) Language production and serial order: a functional analysis and a model. *Psychological Review* **104**: 123–147.

Garrett MF (1975) The analysis of sentence production. In: Bower GH (ed.) *The Psychology of Learning and Motivation*, vol. 9, pp. 133–177. New York: Academic Press.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt WJM (1999) Models of word production. *Trends in Cognitive Sciences* **3**: 223–232.

Roelofs A (1997) The Weaver model of word-form encoding in speech production. *Cognition* **64**: 249–284.

Santiago J, Mackay D, Palma A and Rho C (2000) Sequential activation processes in producing words and syllables: evidence from picture naming. *Language and Cognitive Processes* **15**: 1–44.

Wheeldon L (ed.) (2000) *Aspects of Language Production*. Hove, UK: Psychology Press.

Phonological Processes

Advanced article

Diana Archangeli, University of Arizona, Tucson, Arizona, USA

CONTENTS

Phonological processes are the patterns that are found in the sound sequences that occur in individual human languages. Languages limit which sounds are used; they also limit the permissible sequences of these sounds. These phonological processes, or regular restrictions on sound sequences, generalize both within and across languages.

INTRODUCTION

Phonological processes are the patterns that the sounds of our languages form. Broadly speaking, languages manipulate both the syllable structure and other prosodic properties of words as well as the detailed properties of individual sounds. This article is limited to a discussion of those phonological processes that affect characteristics of individual sounds, that is, the featural processes. Prosodic phenomena such as epenthesis, syncope, and stress assignment are not addressed.

This discussion assumes some familiarity with the terms 'segment', 'feature', and 'lexical representation', and with the contrasts between 'autosegmental' and 'linear' representation of features, as well as a nodding acquaintance with certain features. For those readers who find these terms unfamiliar, they are explained briefly here. Others are invited to skip ahead to the section on harmony.

Lexical representation. A discussion of phonological processes presupposes that such processes do, indeed, exist. Phonologists view these processes as the link between the abstract representation in the speaker's head, called the 'lexical representation', and the 'surface representation', the form that connects to the phonetic interpretation, that is, the sounds coming out of the speaker's mouth.

Segment refers to individual sounds within a word. For example in the English word 'late', there are three sounds, a 'l' sound, a 'long-A', and a 't' sound. Note that the number of sounds does not necessarily correspond to the number of letters. Alphabets such as the International Phonetic Alphabet have been created to provide a one-to-one mapping between sound and symbol. Such symbols are used here, inside square brackets: *late* [leit].

Feature refers to individual properties of sounds. These features are manipulated phonologically, but they are not purely abstract: they roughly correlate to certain acoustic and/or articulatory properties. For example, the sounds [m] and [n] share a feature [nasal], for both have air passing through the nose. The sounds [o], [u] share the feature [round] because both involve rounding of the lips. Any sound involving constriction of the lips is viewed as having the feature [labial]: in addition to [o], [u], this would include [m, p, b, f, v]. Some features are less easily sensed: one dimension that some languages use widely is the position of the root of the tongue. Vowels with advanced tongue root (ATR) sound more or less like the English 'long' vowels, i.e. those in *seed* [i], *staid* [e], *soup* [u], and *soap* [o]. Those with retracted tongue root sound more like the English 'short' vowels, such as *sip* [ɪ], *set* [ɛ], *soot* [ʊ], and *sought* [ɔ].

Segments, then, are characterized as composites of features: a [nasal, labial] consonant is [m] while a [round, labial, advanced] vowel is [u].

The *autosegmental/linear distinction* refers to how these features are represented with respect to the segment. In the linear model, all features for a given segment are bundled into a single unit (a *matrix*). Lexical representations consist of sounds lined up like beads on a string. Manipulation affects the features of each sound individually. The autosegmental representation separates features from the segment, so that a single feature may be a property of several segments simultaneously, as in harmony, and multiple features may be properties of a single segment, as in one type of apparent metathesis.

With this brief introduction as background, the next topic is a description of a number of different phonological processes. The goal here is to identify

the clusters of phenomena typically associated with each of the types of process under discussion. It is important to remember, however, that these terms have been used by numerous different individuals in a variety of different ways. The descriptions here are intended to provide a generic overview, not to account for every variant that has occurred in the literature.

HARMONY

The term 'harmony' refers to a pattern in which some feature(s) inherent to one sound in a word propagate across a domain larger than the one sound. For example, in Chukchee, vowels lower if there is even one low vowel in the word: the [e] vowels of *jeɣvel* 'orphan' are lowered to [a] under the influence of the [aa] of *aacek* 'youth': *jajvaɣ-aacek* 'orphaned youth'.

The variations on this one theme reveal much about the ways in which humans organize the sounds of their languages. Languages with harmony differ in how they identify the harmonic feature(s), the source of that feature (the 'trigger'), the elements undergoing harmony (the 'targets'), which elements are irrelevant to the harmony ('transparency'), which elements prevent harmony from propagating ('opacity'), the direction of harmony, and the morphological domain of harmony.

The Harmonic Feature(s)

Many harmonies involve only a single harmonic feature (e.g. tongue root harmonies such as Yoruba ATR harmony); others involve a complex of features propagating in a single pattern (e.g. Yokuts round/back harmony); still others involve complexes of features, each of which propagates in its own fashion, for example in Barra Gaelic.

A second consideration is the provenance of the harmonic feature(s). In some cases, these are necessarily a part of the lexical representation; nonlexical features do not harmonize. In other cases, both lexical and nonlexical instances of the feature harmonize (e.g. [−ATR] in Yoruba tongue root harmony).

A very entertaining type of harmony is one in which the harmonic feature is introduced as a morpheme in and of itself. One such instance is found in Coatzospan Mixtec: the feature [+nasal] marks the difference between a verb (oral) and verb with a familiar second-person subject (nasal): *kaʔu* 'write' versus *kãʔũ* 'you (fam.) will write', *βiðe* 'wet' versus *βĩðẽ* 'you (fam.) are wet'.

Triggers

In some cases, to trigger a harmonic pattern, a segment need only have the relevant feature. In other cases, triggers are defined as a subset of those segments bearing the feature. In Menomini, for example, although both high and nonhigh vowels may have the feature advanced tongue root, [+ATR], only the high vowels serve as triggers for the harmony process. The retracted [ɪ] in the first syllable of *sɪ̱ːpɪ̱ːw* 'river' raises to [i] when followed by a suffix with the advanced [i], as in *siːpiah* 'river-locative'. By contrast, when the suffix contains an advanced nonhigh vowel, no harmony takes place: *maskʊ̱ːtə̱w* 'prairie' with a retracted [ʊ:] despite the advanced [ə] versus *maskʊ̱ːtiah* 'prairie-locative' with an advanced [u:] due to the advanced high [i]. Only the high advanced vowels trigger harmony.

Targets

Targets are identified in a number of ways. In some cases, any segment that might bear the harmonic feature is a potential target. In other cases, targets are restricted. Common restrictions are 'only vowels', or 'only some vowels'. For instance, in Menomini, long high vowels are targeted; in Standard Yoruba, only nonhigh vowels are targets; in Akan, nonlow vowels are always targets while low vowels are not: advanced *o-fiti-i* 'he/she pierced (it)', retracted *ɔ-cɪrɛ-ɪ* 'he/she showed (it)', versus the disharmonic *o-bisa-ɪ* 'he/she asked for (it)' – the advanced [i] does not affect the low [a].

In other types of harmony, the target may be consonants rather than vowels, for example laryngeal harmony in Salish and in Indo-European, and place harmony of Chumash, Navajo, and other languages. In Chumash, for example, [s/š] alternate to match the rightmost sibilant: *k-s̱unon-us̱* 'I obey him' versus *k-šunon-š* 'I am obedient'.

Finally, some harmony processes affect both vowels and consonants, for instance nasal harmony (e.g. Orejon) and emphasis spread in Arabic.

Transparency and Opacity

Although there are cases in which essentially all segments are targets, such as the Arabic emphasis harmony, there are also cases in which only some segments are acceptable targets. The unacceptable targets may exhibit either transparency or opacity.

Transparency

Typically some segments do not undergo harmony, nor do they hinder harmony. In a sense, they behave as if they are not present at all. These are the transparent segments. For instance, when a feature such as [round] passes from vowel to vowel in Turkish, the consonants remain unaffected by the harmony. In the Menomini examples above, long vowels advanced to [i:] and [u:] despite being separated from the triggering vowel by several segments, including other vowels: [nɪsu̲:pɒma:hki̲m] 'sugar-maker' (compare this to the retracted high long vowels in [su̲:pɒma:hkʊw] 'he makes sugar').

Opacity

In many cases, there are segments which neither undergo harmony nor do they trigger harmony, nor do they permit harmony to pass them. These are opaque segments. For instance, in Yoruba tongue root harmony, high vowels are ineligible targets; at the same time, they also prevent further propagation of the tongue root feature.

A particularly striking example illustrating both transparency and opacity is found in translaryngeal harmony (Steriade, 1987). In translaryngeal harmony, all vowel features spread across only the laryngeal consonants ([h] and [ʔ]). For example, in Yapese, the suffixes *-o* and *-i* cause a root vowel to harmonize when the final consonant is a laryngeal. Note the alternation between *boh-* and *bih-* in the root *bah-* 'go' in *ma bo̲h-o̲* 'he is not going out' versus *ma bi̲h-i̲* 'he did not go out'.

Direction

Harmonies orient either towards the left edge of the word or towards the right edge of the word. A left-edge harmony is one in which the targets are to the left of the trigger; a right-edge harmony is one in which the targets are to the right of the trigger. Some languages orient their harmonies in both directions, but restrict the two directions in different ways. For instance, in Maasai, the left-edge harmony targets nonlow vowels but low vowels are opaque: note complete harmony in *kidotuñe* 'we shall pull it out with something' versus the opaque effect of *ta* in *kɪta̲dotuñe* 'we pulled it out with something'. By contrast, in Maasai, right-edge harmony targets all vowels, turning the final retracted *-ta* into advanced *-to*: *atɛpɛ̲ta̲* 'I kept close to it' versus *atapɐ̲to̲* 'I smeared it'.

Morphological Domains

The distance that the harmonic feature propagates varies from harmony pattern to harmony pattern. Harmony may be limited to a single morpheme, as in the case of Tiv verbs. Harmonies may also be limited phonologically within a particular morphological domain, for example within the foot. Harmony may occur both within morphemes and across morpheme boundaries, as in the Chukchee example above. Harmony may even jump across word boundaries, as in Kinande: the retracted vowels in *è-mí:-tì* 'trees' are advanced in a phrasal context: *è-mí:-tí míkù:hì* 'short trees'.

Analysis of Harmony

Within the generative tradition, there have been two general representational strategies to analyzing harmony, linear and autosegmental. Whichever method is adopted, it is necessary to account for all of the general types of phenomena noted above.

The linear approach changes feature values depending on the context; it requires recursive application of rules in order to force feature-changing across the entire domain (feature propagation). The autosegmental approach involves feature-spreading and/or feature copy, subject to limitations on trigger, target, and domain. (See Clements, 1981; Archangeli and Pulleyblank, 1994.)

There are analyses of both types within Optimality Theory (Prince and Smolensky, 1993; McCarthy and Prince, 1993a). The Alignment view (McCarthy and Prince, 1993b) assumes an autosegmental representation while the Optimal Domains approach (Cole and Kisseberth, 1994) adopts a more linear segmental representation.

ASSIMILATION AND DISSIMILATION

Assimilation refers to two segments becoming more like each other. A familiar example of assimilation is that of nasals assimilating in place to a following consonant. For example, in English, the negative prefix *in-* appears in such words as 'inopportune' and 'inapplicable' as well as 'intolerable' and 'indecisive'. However, when added to a word such as 'balance' or 'possible', we find a labial nasal, not a coronal nasal: 'imbalance', 'impossible'. The nasal has assimilated place features from the following consonant, thereby becoming more like that consonant.

Assimilation is a fairly common process. There is also evidence suggesting that there are phonetic

processes of harmony and assimilation, and again there is a high degree of similarity. Phonetic harmony is also termed a *cline* effect and assimilation is termed *co-articulation*.

Dissimilation refers to two segments becoming less like each other. For example, certain Latin suffixes (such as *-alis/-aris*) alternate between a form with [l] and a form with [r], the latter surfacing when the affix is attached to a root already containing an [l]: *naval̲is* 'naval' but *sol̲ar̲is* 'solar'. The [l] of the suffix has become an [r], and therefore less like the [l] of the root. Dissimilation is not rare, but it does seem to be less common than assimilation.

Assimilation is not restricted to place features alone. For example, in Malayalam, a nasal–stop sequence may surface as a geminate nasal in colloquial speech: formal *can̲d̲ʾanam* 'sandalwood' is colloquially *can̲n̲anam*. Note also that in this example, it is the stop that assimilates to the nasal, rather than the nasal assimilating to the stop.

Vowels assimilate to neighboring vowels, even vowels separated by a consonant. For instance, in Basque, vowels raise when following a high vowel. The 'professional' suffix alternates between *-ari* and *-eri* depending on the vowel that precedes: *eskel̲ar̲i* 'beggar' but *trat̲ul̲eri* 'dealer'. (This type of effect of one vowel on another is also termed *umlaut*.)

Finally, there are cases of consonants assimilating to neighboring vowels and vowels assimilating to neighboring consonants, for example the palatalization of coronal consonants [t, n, l] following the vowel [i] in Barrow Inupiaq: this is illustrated with [n/ñ] in *sisu-n̲iaq* 'will slide' versus *niRi-ñ̲iaq* 'will eat'. Labialization, pharyngealization, and spirantization are other effects that may result from the influence of a vowel on a neighboring consonant.

When comparing characteristics of harmony described above and those of assimilation given here, we see great similarities. In both cases, only some feature(s) may harmonize/assimilate. Only some segments are potential targets; there may be transparent and/or opaque elements. Direction varies: in an AB sequence, one of the two is the target and the other is the trigger. Finally, each type of process may be sensitive to certain types of morphological boundaries.

These many phenomenological similarities suggest that assimilation and harmony are formally extremely similar phonological processes, and that the formal devices necessary to account for one set of patterns will also account for the other set. The primary difference is that assimilation is local, involving only adjacent or almost adjacent segments, while harmony is long-distance, involving segments that might be separated from each other by several other segments.

There is a second and more subtle difference, however, in the use of these terms. Typically, 'harmony' is used for the assimilation (long-distance or local) of any feature(s) that may be involved in a long-distance process, while 'assimilation' is used for the assimilation of any feature(s) that typically are involved only in local processes. Thus, we speak of place assimilation, since place features typically have only a local effect. And we speak of tongue root harmony, even in a language like Lango in which tongue root features affect only vowels in adjacent syllables, because tongue root features frequently have a long-distance effect.

An important area for investigation is to understand which features typically (or necessarily) pair with local effects and which with long-distance effects, and then to understand why this is so. Once the role of the features is understood, we will be in a better position to determine whether these two effects must be characterized in formally distinct manners.

The term *dissimilation* suggests in some sense the antithesis of assimilation. Broadly speaking, this is true in descriptive terms, for dissimilation refers to two sounds becoming less similar.

Dissimilation typically identifies some feature(s) that cannot occur too close to each other. Thus, in the Latin example above, two instances of the [lateral] feature cannot occur in close proximity.

In some cases, it is necessary to define the nature of the trigger more specifically than simply any segment bearing the dissimilatory feature. For example, in Akkadian, a prefixal [m] alternates with [n] when attached to a root containing a labial consonant: *m̲a-zuukt* 'mortar' versus *n̲a-raam̲u-m* 'favorite'. Labial vowels do not cause this effect (*m̲a-zu̲u̲kt*, not **n̲a-zu̲u̲kt*). Similarly, suffix [m] has no effect (*m̲a-śʔanu-m̲* 'place'). Thus the trigger must be identified as [+ labial] borne by a root consonant – both phonological and morphological restrictions.

Targets of dissimilation may be restricted as well. A particularly interesting case of restricted targets is found in Woleaian. The low vowel [a] dissimilates to a mid [e] when the next syllable contains another low vowel, long or short [a, aa, ɔɔ]: *g̲a-b̲o̲sǫ̊* 'causative-him to show off' versus *g̲e-m̲aarǫ̊* 'causative-make him starve'. However, this occurs only if the target vowel is short: *f̲aar̲ag̲i̥* 'to walk', not **f̲eer̲ag̲i̥*. Thus, the element targeted for [low] dissimilation is restricted to unround, short vowels.

As with harmony, both transparency and opacity play a role in dissimilation too. Transparent

elements are ignored by the dissimilatory effect. In the Latin example already cited, for instance, consonants and vowels other than [l] are irrelevant: dissimilation occurs in both *lun-aris* 'lunar' and *sol-aris* 'solar', despite the greater distance between trigger and target in *lunaris*. Latin also gives evidence of opaque elements: an [r] intervening between the two [l]s prevents dissimilation: *flor-arlis* 'floral', not **flor-aris*.

Latin exhibits a further important point, that the domain is critical. Latin lateral dissimilation takes place only across morpheme boundaries, not within morphemes: *calculus* 'pebble, stone', not **calcurus*.

Dissimilatory effects also have a left or right orientation. In Latin, for instance, it is the righthand [l] which must change. In Russian, it is the leftmost of two nonhigh vowels which must change, not the rightmost: *d'is'átka* 'tenfold' versus **d'es'átki*.

Finally, the key point that distinguishes 'harmony' from 'assimilation' is the potential distance of trigger from target. In the cases considered thus far, the dissimilatory trigger–target pair can be quite distant from each other. However, this is not always the case. In Kera, for instance, two [a] vowels are separated by no more than one consonant in order for the leftmost [a] to dissimilate to [ə]: *bəla* 'want me' (not **bala*) but *balla* 'you must want', not **bəlla*.

Intriguingly, then, although the basic effect is quite different (segments becoming less like each other rather than more like each other), the specific properties which define a particular instance of dissimilation find clear counterparts with the properties defining both harmony and assimilation.

METATHESIS

The classic definition of metathesis is that two adjacent segments trade places, as in the difference between the Old English root *acs* (and the current colloquial pronunciation *aks*) and present day *ask*. Metathesis also occurs in language acquisition, for example the child's *aminal* for *animal*. Metathesis is perhaps more familiar from studies of language change than it is from synchronic processes. Another class of diachronic cases of metathesis is sketched in Sagey (1986), in which cognates in related languages reveal the [AB] versus [BA] ordering. For example, two mutually intelligible dialects of Yatye reverse the order of place:

Alifokpa dialect	Ijiegu dialect	
icwɛndɛ	ipyɛndɛ	'pot'
ecwu	epyu	'head'
jwu	byu	'drink'

Sagey suggests that in fact there is no reordering of segments, but rather that the labial and palatal features are both properties of a single consonant segment, but the order in which they are phonetically realized varies. That is, in both dialects, there are palatal-labial stops; in Alifokpa, the palatality is timed to begin before the labialization while in Ijiegu, the timing is the opposite. Thus, the formal account involves no reversal of the order of segments.

There are a few cases where the analysis does suggest the reversal of features. In Rotuman, words have two versions, one which is vowel-final and another which tends to be consonant-final: *tokiri* versus *tokir* 'to roll', *rako* versus *rak* 'to imitate'. Depending on the qualities of the two vowels, other effects may occur, such as umlaut: *mose* versus *mös* 'to sleep'. McCarthy (1995) argues that the shorter form is prosodically reduced; in particular the position for the final vowel is lost. The features of that vowel, however, do their best to surface. Further support for this general analysis is that words ending with two vowels, such as *keu* 'to push', alternate with a form ending with a short diphthong, *ke͡u*, that is, the final vowel position is lost, but the features survive by crowding into the position for the preceding vowel. In certain cases, where umlaut and diphthongization are impossible, there is apparent metathesis: *pure* versus *puer* 'to rule'. Again, this can be viewed as the final vowel (here *e*) losing its position, yet finding a place for the features to survive by crowding into the remaining vowel position.

Metathesis seems to be relatively uncommon. Many of the known cases are amenable to a single-segment analysis along the lines of Sagey (1986). Others are similar to the McCarthy (1995) analysis of Rotuman, wherein dislodged features seek an anchor so that they may surface.

However, not all examples argued to show metathesis obviously resolve themselves in one of these fashions. For example, an analysis of Kasem claiming metathesis is presented in Chomsky and Halle (1968), yet there are no simple cases of input /AB/ being realized as [BA]. Metathesis here is argued to reorder two vowels when followed by a third vowel (there are further restrictions). However, other aspects of the analysis include the merger of two vowels into one and the deletion of one vowel next to another. Consequently, trivocalic inputs, such as /pia-i/ 'sheep-plural', are realized as monovocalic, e.g. [pæ], with no obvious surface reflex of the metathesis. Unfortunately, in many cases of apparent metathesis, including Kasem,

the data are incomplete, making it difficult to be confident of any analysis.

WEAKENING AND STRENGTHENING

Weakening and strengthening (also termed *lenition* and *fortition*, respectively) are another pair of descriptive terms that do not necessarily make a smooth transition to analysis. Descriptively, the phenomena are relatively simple: weakening refers to any change that makes a sound weaker while strengthening refers to any change that makes a sound stronger. Yet these definitions are circular, since 'weaker' and 'stronger' are not defined.

Weakening includes phenomena such as word-final devoicing (e.g. in Woleaian) or high vowel devoicing (e.g. in Japanese); it also includes intervocalic voicing of obstruents (e.g. in some dialects of Japanese) and intervocalic spirantization. Gosiute Shoshone provides a case in which obstruents undergo both voicing and spirantization when intervocalic, becoming doubly weakened. Voiced fricatives only occur between vowels while voiceless stops only occur initially (or in a strengthened form as geminates, [moppo] 'mosquito'): there are words such as [tɨβa] 'pine nut' and [(peðɨ)] 'daughter, niece', but no words such as *[βɨta] or *[ðepɨ].

Strengthening includes intervocalic gemination (which occurs in certain morphological environments in Gosiute). Strengthening might also be viewed as including lengthening or stressing phenomena.

At this point, weakening and strengthening are useful to suggest particular types of phenomenon. However, in many cases, such as the Gosiute examples, it is not clear whether the phenomena are phonetic (as in the weakening example of voicing and spirantization cases) or morphological (as in the strengthening gemination example). There are also cases where weakening and strengthening are complementary. For example, Spanish voiced obstruents are 'strong' stops when following a homorganic sonorant, but otherwise are 'weak' fricatives between sonorants. Thus, as with metathesis, there is currently a poor understanding of the dimensions along which variation can be found with these types of processes.

CONCLUSION

A large class of phonological processes involving features vary along a small set of dimensions: which features are critical, which segments undergo the process, which cause the process, the relevant domain, and the leftward or rightward orientation. A fine-tuning of the process comes in identifying transparent elements – those which are ignored by the process – and opaque elements – those which end a domain which otherwise would not end at that point. There remains a residue of ill-understood processes which may lead us to enlarge this set of dimensions but which may, under closer examination, fall neatly within this set.

References

Archangeli D and Pulleyblank D (1994) *Grounded Phonology*. Cambridge, MA: MIT Press.

Chomsky N and Halle M (1968) *The Sound Pattern of English*. New York: Harper & Row.

Clements GN (1981) Akan vowel harmony: a nonlinear analysis. In: Clements GN (ed.) *Harvard Studies in Phonology 2*, pp. 108–177. Department of Linguistics, Harvard University.

Cole J and Kisseberth C (1994) *An Optimal Domains Theory of Harmony*. Cognitive Science Technical Report UIUC-BI-CS-94-02, University of Illinois, Urbana-Champaign.

McCarthy J (1995) *Faithfulness in Prosodic Morphology and Phonology: Rotuman Revisited*. MS, University of Massachusetts, Amherst.

McCarthy J and Prince A (1993a) *Prosodic Morphology I: Constraint Interaction and Satisfaction*. MS, University of Massachusetts, Amherst, and Rutgers University.

McCarthy J and Prince A (1993b) Generalized alignment. In: Booij G and van Marle J (eds) *Yearbook of Morphology 1993*, pp. 79–153. Dordrecht: Kluwer.

Prince A and Smolensky P (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*. TR-2, Rutgers University Cognitive Science Center.

Steriade D (1987) Locality conditions and feature geometry. In: McDonough J and Plunkette B (eds) *Proceedings of NELS 17*, pp. 595–617. Distributed by the Graduate Linguistic Student Association, University of Massachusetts at Amherst, Amherst, MA.

Walli-Sagey E (1986) On the representation of complex segments and their formation in Kinyarwanda. In: Wetzels L and Sezer E (eds) *Studies in Compensatory Lengthening*. pp. 251–295. Dordrecht: Foris.

Further Reading

Clements GN and Sezer E (1982) Vowel and consonant disharmony in Turkish. In: van der Hulst H and Smith N (eds) *The Structure of Phonological Representations II*, pp. 213–255. Dordrecht: Foris.

Davis S (1995) Emphasis spread in Arabic and grounded phonology. *Linguistic Inquiry* **26**: 465–498.

Elzinga D (1999) *The Consonants of Gosiute*. Doctoral dissertation, University of Arizona.

Gerfen H (1996) *Phonology and Phonetics in Coatzospan Mixtec*. Studies in Natural Language and Linguistic Theory. Dordrecht: Kluwer.

Kenstowicz M (1994) *Phonology in Generative Grammar*. Cambridge, MA: Blackwell.

Kenstowicz M and Kisseberth C (1979) *Generative Phonology*. New York: Academic Press.
Kent RG (1936) Assimilation and dissimilation. *Language* **12**: 245–258.
Shaw P (1991) Consonant harmony systems: the special status of coronal harmony. In: Paradis C and Prunet JF (eds) *Phonetics and Phonology 2: The Special Status of Coronals*. San Diego, CA: Academic Press.
Suzuki K (1998) *A Typological Investigation of Dissimilation*. Doctoral dissertation, University of Arizona.

Phonology and Phonetics, Acquisition of

Introductory article

Peter W Jusczyk,† Johns Hopkins University, Baltimore, Maryland, USA

CONTENTS

During their first year infants discover the elementary sound units that are used to form words in their native language. During their second year they begin to produce these sound units and acquire the structure of their native language.

INTRODUCTION

Much like the way that letters are the building blocks of written words, phonetic segments are the elements out of which spoken words are built. There is a finite set of such elements that are used in creating words in any natural language. However, any particular language, such as English, uses only a subset of the set of phonetic segments that are available for forming words. For example, English does not make use of click sounds that are used in some African languages, such as Zulu. By the same token, some sounds used in English, such as the initial sounds in the words 'then' and 'thick', are not found in French. So, one of the problems facing language learners is to discover which elementary sounds are used in forming words in the language they are trying to acquire. In addition to learning which elementary sounds are used, learners also have to understand the ways in which the elementary units can be combined to form words in their language. For instance, someone learning Polish will find that it is permissible to begin words with consonant sequences, such as 'kt' and 'db'. By comparison, someone learning English will never encounter these particular consonant sequences at the beginning of words.

ACQUISITION OF PERCEPTUAL CATEGORIES IN INFANCY

When do infants begin to learn about the elementary sound units of their language? For many years, researchers could not begin to investigate this issue until infants actually began to produce speech sounds themselves. This is because they lacked suitable methods for studying whether infants could perceive differences in speech sounds, even before they are capable of producing these distinctions in their own utterances. However, during the 1960s a number of different methods were developed to study the visual and auditory capacities of young infants.

One method that became widely adopted for studying the speech perception capacities of infants was called the high-amplitude sucking procedure (often abbreviated as HAS). In this procedure, infants are given a pacifier to suck on. The pacifier is connected to a pressure transducer, which allows an experimenter to record the frequency and intensity of infants' sucking responses. After obtaining an estimate of how often an infant sucks on the pacifier in the absence of any auditory stimulation, the experimenter makes the presentation of an auditory stimulus, such as the syllable 'ba', contingent on the infant's sucking. The more often the infant sucks on the pacifier, the more often the syllable 'ba' is played. Even one-month-olds take only a few minutes to learn the contingency between their sucking responses and the presentation of the speech sound. Consequently, they increase their sucking rates considerably over their baseline rates. After listening to the same sound for several minutes, infants lose interest in it (i.e. they habituate to the stimulus) and their sucking rates begin to decline. At this point, if a new sound, such as the syllable 'pa', is substituted for the original one, and the infant is able to perceive the change, then their sucking rates will typically increase in response to the novel sound. By observing whether infants react to a change from one speech sound to another, investigators can determine which types of phonetic contrasts infants can perceive.

Investigations using methods such as HAS showed that even one-month-olds are able to perceive a distinction, such as the one between 'ba' and 'pa'. This is remarkable considering that the sounds are alike in all respects, except for one that linguists refer to as *voicing*. For a voiced sound like 'ba', the vocal cords begin vibrating when air is released from the lips, but for a voiceless sound like 'pa', air is released from the lips about 50 milliseconds (ms) before the vocal cords begin vibrating. Whenever two sounds are alike in all respects except for one phonetic dimension, such as voicing, linguists refer to them as *minimal pairs*. In addition to the voicing distinction used in the first investigation of infants' speech discrimination abilities, infants in the first few months of life have been shown to discriminate a number of different kinds of minimal pairs, such as the stop consonants [b] versus [d] which differ in place of articulation, [b] versus [w] which contrast between a stop consonant and a glide, the liquids [r] versus [l], the nasals [m] versus [n], etc.

What role does experience play in infants' abilities to discriminate minimal pairs? Do infants have to have some period of prior exposure to these different speech sounds before they are able to discriminate them? The available evidence suggests that such experience is not required. First, even infants only a few days old have been shown to discriminate phonetic contrasts involving minimal pairs. Second, studies show that during the first six months of life, infants are able to perceive contrasts between phonetic segments not present in the native language that they are exposed to. For example, Kikuyu (a language spoken in Kenya) does not have a distinction between [b] and [p], yet infants from Kikuyu-speaking homes are able to perceive this distinction. The picture that emerges from these studies is that infants are born with the ability to discriminate any kind of phonetic contrast that occurs in the world's languages. However, this does not necessarily mean that they respond to any acoustic difference that they can detect between two different speech sounds. For example, even infants a few days old display some ability to distinguish one speaking voice from another. Yet six-month-olds can ignore such talker differences in performing a task that relies on their ability to distinguish [a] from [i]; they are able to do so, even when listening to many different talkers' productions of these vowels. The latter ability is important because in order to learn which phonetic segments are used to form words in their language, infants must be able to ignore individual differences in how different talkers produce these segments.

These findings indicate that, in the first months of life, infants possess the perceptual capacities that would allow them to learn the phonetic segments in any native language. However, infants still have to identify which segments are present in their own native language. There are some indications that they learn about these segments during the second half of their first year. During this period of time, infants appear to become much more focused on the particulars of the sound organization of their native language.

One indication that infants are becoming more focused on their native language comes from studies of their speech discrimination abilities. Between six and eight months, infants still display an ability to discriminate non-native language speech contrasts. Thus, English learners are able to discriminate contrasts from Hindi (the retroflex versus dental distinction that varies the precise positioning of the tongue behind the teeth) that do not appear in English. By eight to 10 months, though, their abilities to discriminate these contrasts begin to decline, and by 10 to 12 months, they no longer show any ability to discriminate them. Interestingly enough, it is during this same period in which a similar decline has been noted in Japanese infants' ability to discriminate [r] from [l], which is a distinction that is not used in Japanese.

Why does a decline occur in the ability to discriminate many non-native speech contrasts towards the end of the first year, such as the Hindi /ṭa/–/ta/ and the English /r/–/l/? One theory was that lack of exposure to the relevant speech sounds was the cause of the decline. However, it has been shown that at least some non-native contrasts continue to be perceived well, even into adulthood. An alternative view is that towards the end of the first year, infants learn about the organization of the elementary sound categories of their native language. An important part of this process involves learning when two apparently different sounds are actually variants (or allophones) of a particular phoneme of the language. For example, the English phoneme /t/ is realized as any one of a number of different phonetic segments, depending on the context in which it occurs. When /t/ occurs at the beginning of a word, it is produced with a large puff of air (i.e. it is aspirated), yet when /t/ occurs as the last segment of a word, it is usually unaspirated. When /t/ occurs immediately before /r/ as in 'intricate', it is produced as a retroflexed segment (i.e. the tongue is curled when it contacts the roof of the mouth). In a language such as Hindi, the distinction between a retroflexed *t* and an unaspirated *t* might convey a

meaningful distinction between words. In English it does not, so the English-learner has to learn that both are variants of the same English segment. Learning which kinds of distinctions that occur in speech input are meaningful and which are to be ignored appears to contribute to the decline in discriminating non-native contrasts.

Other findings indicate that infants are learning about other aspects of the sound organization of their language during the latter half of the first year. For instance, infants appear to learn about the permissible orderings of phonetic segments in words between six and nine months. This was shown in a study in which infants were played lists of unfamiliar English words and lists of unfamiliar Dutch words. Dutch and English words have the same rhythmic and pitch structure; they differ chiefly in which segments and segment sequences they use. The words used in the experiments from each of these languages contained phonetic sequences that were not permissible in words in the other language. At six months, English-learners are as content to listen to the lists of Dutch words, which violate English word structure, as they are to listen to the lists of English words. However, by nine months, English-learners listen longer to the English lists, whereas Dutch learners listen longer to the Dutch lists. The fact that nine-month-olds display these preferences, even when phonetic segments distinctive to each language were removed, suggests that they have learned about which kinds of phonetic sequences are permissible in words in their language. This phonetic sensitivity presumably arises from exposure to sounds in the infant's native language environment that vary in their relative frequency of occurrence.

In summary, the well-developed perceptual capacities that infants display in the first months of life enable them to make considerable progress in learning about the organization of the sound structure of their native language by the end of their first year. In the second year, infants further refine this sound structure as they begin to produce words.

EARLY SPEECH PRODUCTION

Although infants produce some vocalizations during the first four months of life, the characteristics of these differ considerably from the sounds that are typically used in producing words. At best, infants might produce some sounds that resemble drawn-out vowel sounds and occasional 'gooing' sounds. Only after four months do they begin to produce sounds that resemble some consonant–vowel sequences. These sounds are often produced in the contexts of other sounds, such as 'raspberries', growls, and squeals. Moreover, productions of sounds that resemble consonant–vowel sequences during this period lack the temporal characteristics of syllables used in speech production. At some point, around six or seven months of age, infants begin to produce syllables and strings of syllables with the temporal characteristics of ones used in language. This marks the onset of what linguists refer to as *canonical babbling*. Infants now produce well-formed syllables that exhibit the kind of vocal cord vibration, articulatory movements, resonance patterns, and temporal changes that are typical in language production. However, the syllables which infants produce during the first few months of the babbling period lack a clear segmental organization. Although these syllables may approximate consonant–vowel sequences, they do not show the range of permutations that would be expected if there were segments that could be freely recombined. Thus, the syllables that infants initially produce seem to be unanalyzed wholes, rather than combinations of individual phonetic segments.

Over the years, there has been considerable discussion about the relation that babbling bears to the development of linguistic production. One early view held that babbling is prelinguistic, and that a silent period intervened between babbling and true speech production. It was suggested that during the babbling period, much as young infants show the ability to discriminate phonetic contrasts from any of the world's languages, they also showed the ability to produce all the sounds of these languages. However, infants were thought to emit these sounds in a random fashion, rather than in any systematic attempt to achieve a targeted articulation. Moreover, a natural consequence of this view that babbling is a prelinguistic activity, not systematically controlled by the infant, was the claim that the babbling of deaf infants is the same as that of normally hearing infants. It was claimed that the chief difference between the two groups was a decrease in the amount of vocalizing that deaf infants do as they get older.

The view just described was based largely on anecdotal observations and diary descriptions of infants' babbling. More carefully documented analyses and recordings of babbling have largely refuted this view. For example, observers who have systematically recorded the babbling behavior of infants over many months have found no evidence of a silent period intervening between babbling and the production of words. In fact, the available evidence suggests that there is

considerable continuity between babbling and the production of infants' first words. For example, phonetic segments that occur frequently in later stages of canonical babbling are more likely to appear in the words that infants choose to produce. Similarly, although a range of different sounds does occur in babbling, there is no evidence that infants produce all the sounds of the world's languages. Indeed, the physical state of infants' developing vocal tracts during the first year precludes the production of certain types of speech sounds. Finally, the development of better clinical screening methods for detecting deafness in early infancy has allowed researchers to make more direct comparisons of babbling of deaf and hearing infants. Recent investigations reveal several abnormalities in the vocalizations of deaf infants. In particular, there is little evidence of canonical babbling in deaf infants during their first year, well after the onset of canonical babbling in normal infants.

The kinds of sounds that are likely to appear initially in canonical babbling are usually stop consonants such as [p], [b], [t], [d], and less frequently [g] and [k]. Nasal consonants such as [m] and [n] also appear early. These consonant sounds are most often produced with the vowels [a] (as in *hot*), [æ] (as in *bat*), and [ʌ] (as in *nut*). The production of these types of consonants and vowels early is a natural consequence of the opening and closing of the jaw. The consonants that appear are ones that result from completely closing the vocal tract at some point, whereas vowels such as [a] and [æ] result from a large opening of the mouth. Many of the repeated sequences of syllables that appear during canonical babbling (e.g. *dadada*, *mamama*) are a natural consequence of repeated openings and closings of the jaw.

For many years, it has been claimed that babbling drifts in the direction of the native language sound patterns, so that the babbling of English-learners comes to resemble English, whereas the babbling of Cantonese-learners comes to resemble Cantonese. Initially, these claims were based on the subjective judgments of adults listening to babbling produced by infants from different language-learning environments. One report suggested that influences of the native language on babbling are evident as early as eight months, in that French adults could distinguish the babbling of French-learners at this age from that of Arabic-learners. More recently, careful acoustic analyses of the vowel sounds produced by French, Arabic, English, and Cantonese 10-month-olds documented differences in infants' productions that reflect the vowels produced by native speakers of these languages. Changes reflecting the structure of the native language have also been noted in infants' production of consonants. For instance, a cross-linguistic study of English, Swedish, Japanese, and French infants revealed that the proportion of stop consonant sounds produced by 10-month-olds reflected the frequency and distribution of these sounds by adult speakers of these languages. Similarly, the intonation patterns and rhythmic patterns that occur in babbling have been shown, during the course of the latter half of the first year, to increasingly resemble those of the native language. Thus, just as influences of native language sound organization are evident in the perceptual capacities of infants towards the end of the first year, so too are they evident in the speech sounds that infants produce.

Infants' earliest productions of words are not usually accurate reproductions of words used by adults. Rather, they often involve some simplification of the segments present in the adult form of the word. For instance, a child attempting to say 'duck' may pronounce it as 'guck', so that the initial and final consonants have their vocal tract closure in the same place – namely, at the velum. Clusters of consonants may be reduced to a single consonant so that 'blue' may be produced as 'boo' and 'spoon' may be produced as 'poon'. Syllables of longer words are often dropped in a systematic fashion. Typically, unstressed syllables are more often omitted, especially word-initial ones: so, a child would be more likely to omit the first syllable of 'giraffe' than the second syllable of 'monkey'. The source of such differences between the child's productions and the adult word forms does not appear to be a failure of perception on the child's part. Rather, such differences appear to be governed more by difficulties that children have in planning, sequencing, and remembering the articulatory gestures required to produce the adult target word correctly. Children often appear to notice the discrepancy between their productions and the adult word forms, as evidenced by the fact that they often object when an adult reproduces the child's form rather than the correct form. Also, there is some evidence suggesting that children will avoid producing words that contain segments that they are unable to produce accurately.

OVERGENERALIZATION IN MORPHOPHONOLOGY

Linguists refer to the smallest unit of linguistic meaning as a *morpheme*. A word such as 'dog' is considered to be a morpheme, but so is the '-s' that

is added to make the plural of this word. Consequently, the word 'dogs' is considered to have two morphemes. Similarly, the '-ly' ending that we add to 'pretty' to make the adverbial form 'prettily' is another morpheme of English, as are the endings such as '-ing' and '-ed' that we add to the ends of verb forms.

One interesting property of some of the morphemes that we add to the ends of words in English is that these particular morphemes often have more than one phonetic realization. For example, consider the '-s' that is added to form the plural of nouns in English. In some cases, '-s' is pronounced as [s], such as when it is added to words such as 'cat, bike, rope'. However, in other cases, the '-s' is pronounced as [z] when added to the end of words such as 'dog, tab, pig', or even as [ɪz] when added to 'dress, bush, church'. Which phonetic realization of '-s' is used to form the plural is not arbitrary. Rather, it is governed by the nature of the phonetic segment at the end of the morpheme to which it is to be attached. With a few exceptions, words that end in voiced consonants or in vowels typically take the [z] form, whereas those that end in voiceless consonants take the [s] form. The exceptions are words that end in the fricatives [s], [ʃ] (the last sound in 'fish'), [ʒ] (the last sound in 'barrage'), and [z], or affricates (such as the final sounds in 'bench' and 'judge'). Similarly, it is the phonetic properties of the final segments of verbs that governs whether the past tense '-ed' will be phonetically realized as [d] (as in words such as 'fibbed'), [t] (as in words such as 'looked'), or [əd] (as in words such as 'fitted').

There has been a longstanding interest in when language learners master these apparently rule-governed alternations in the phonetic realization of these types of morphemes. A classic experiment in the study of language development presented pre-school-aged children with a picture of an object described as a 'wug'. The experimenter named the object, then pointed to a picture of two such objects and asked the child for the correct completion of the sentence, 'There are two ____'. In this case, the children correctly pronounced the plural ending as [z]. They also correctly pronounced the plural ending as [s] when shown comparable pictures of another object referred to as a 'bik'.

Although many nouns in English form plurals by adding '-s' and many verbs form their past tense by adding '-d', there are some exceptions. For example, the plural form of 'child' is 'children', the plural form of 'mouse' is 'mice', and the plural form of 'foot' is 'feet'. Similarly, the past tense form of 'is' is 'was', the past tense form of 'go' is 'went', and the past tense form of 'break' is 'broke'. These kinds of nouns and verbs are said to be irregular in that they do not form plurals or past tense in the way that most of the other words in the language do. Interestingly, many of these irregular nouns and verbs have a higher frequency of occurrence in the linguist input than do particular words that take regular forms. Moreover, children often produce the past tense of irregular verbs such as 'bring' and 'go', before producing the past tense of regular verbs such as 'kiss' and 'hug'. However, when they do begin to produce regular plural and past tense forms of words, they often overgeneralize these forms by applying them to irregular nouns and verbs. Thus, children who begin to pluralize 'ring' as 'rings' and 'cat' as 'cats' may also begin to say 'foots' and 'mouses' for the plural forms of 'foot' and 'mouse'. Similarly, children who correctly produce the regular past tense forms 'kissed' and 'hugged' may also begin to say 'goed' and 'breaked', even though they may never have heard any adults produce these forms. The existence of this kind of overgeneralization is an indication that children do not simply imitate what they hear adults say. Instead, these types of errors are an indication that children look for general kinds of patterns in the input and attempt to apply such patterns systematically in their productions. In time, children learn not to apply the rules for the regular forms to these irregular nouns and verbs.

Acknowledgement

The publishers would like to thank Richard Aslin, Professor of Brain and Cognitive Sciences, University of Rochester, for his help in proofreading and revising the text of this article.

Further Reading

Berko J (1958) The child's learning of English morphology. *Word* **14**: 150–177.

Boysson-Bardies B (1999) *How Language Comes to Children*, translated by M DeBevoise. Cambridge, MA: MIT Press.

Eimas PD, Siqueland ER, Jusczyk PW and Vigorito J (1971) Speech perception in infants. *Science* **171**: 303–306.

Fromkin V and Rodman R (1998) *An Introduction to Language*, 6th edn. Fort Worth, TX: Harcourt Brace.

Gerken LA (1994) Child phonology: past research, present questions, future directions. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 781–820. New York, NY: Academic Press.

Jusczyk PW (1997) *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Oller DK (2000) *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum.

MacNeilage PF (1998) The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* **21**: 499–546.

Vihman MM (1996) *Phonological Development: The Origins of Language in the Child*. Cambridge, MA: Blackwell.

Werker JF and Pegg JE (1992) Infant speech perception and phonological acquisition. In: Ferguson CA, Menn L and Stoel-Gammon C (eds) *Phonological Development: Models, Research, Implications*, pp. 285–311. Timonium, MD: York Press.

Phonology, Computational

Intermediate article

John Coleman, University of Oxford, Oxford, UK

CONTENTS

Various methods of computational modeling of phonological processing have been developed, especially in speech and language technology but also for cognitive models.

INTRODUCTION

Computer programs for various phonological tasks have been developed in several areas. Speech synthesis and recognition devices, for example, relate symbolic representations (e.g. phonetic transcriptions) to speech signals and to normal spelling. For speech synthesis, prosodic properties of a text must be computed, such as stress, intonation, and the division of words into syllables. All these computations can be performed by programs that employ phonological rules. These rules must be: (1) explicit (they must not need any human intervention to interpret them), (2) complete (they should deal with all inputs that they could ever be given), and (3) computable (a program that follows the rules must stop after a finite number of processing steps, not go into an endless loop). In real-world applications, they must also be computable in practice, i.e. reasonably quickly, given available hardware. In cognitive models of phonological processing, rule-based models are sometimes proposed, such as Coltheart's model of reading (Coltheart *et al*., 1993), which is similar to the grapheme-to-phoneme translation programs used in speech synthesis. But in cognitive science, connectionist methods now enjoy great popularity. (*See* **Prosody**; **Stress**; **Intonation**; **Reading, Psychology of**)

APPLICATIONS

A division can be drawn between applications intended to be useful tools for working linguists and 'pure' computational phonology, the development and evaluation of implementations of theoretical proposals. An example of the former is Lowe and Mazaudon's (1994) 'reconstruction engine', which helps linguists to compile a lexicon of reconstructed words in a proto-language, given lexicons of the daughter languages and a list of sound change rules. This helps researchers to manage their data and to explore the consequences of altering the proposed sound changes or the reconstructed words. An example of the latter is Williams's (1994) 'LexPhon' system, an implementation of the kind of phonological rule system proposed by theoretical phonologists. Computational implementations of most areas of theoretical phonology have been developed. But most research in 'pure' computational phonology is concerned with the development of new frameworks for phonological processing, such as two-level phonology, using finite-state transducers (Kaplan and Kay, 1994), constraint-based phonology (e.g. Bird and Klein, 1994), and probabilistic approaches. (*See* **Phonology**)

FINITE-STATE APPROACHES

The rules employed in traditional generative phonology are all of the form 'A → B/C — D', read as 'symbol A is rewritten as symbol B when it follows C and precedes D'. In many phonological analyses, B may be empty, meaning that A is to be deleted. It has been proved that such grammars are more powerful than is necessary for the definition of human languages: they are computationally expensive and psychologically implausible. Phonological analysis using such rules may be inefficient or even uncomputable (Bear, 1990). This processing problem can be ameliorated, to some degree (Maxwell, 1994). However, the most important advance in dealing with such rules was the proof by Johnson (1972) that if phonological rules are restricted in two ways, only regular languages can be generated. Such languages can be recognized, generated, and processed in other ways very efficiently. The two restrictions are (1) the distinctive features of the

symbols in the language must have a finite number of values (e.g. + and −, but not an open-ended set of integers); and (2) rules may not reapply to their own output. These restrictions are actually met in most phonological analyses. (*See* **Distinctive Feature Theory**)

FINITE-STATE AUTOMATA

The words or expressions of any regular language can be generated or recognized by a type of abstract computing device, a finite-state automaton. These have a mathematical specification, described in most formal language theory textbooks, and they can be implemented as working software. An automaton can be depicted as a network of nodes, each representing a state of the automaton, and arrows, representing changes from one state to the next. The arrows are labeled with sets of symbols, one of which must be read or written by the machine in order to move from one state to the next. Figure 1 shows an automaton which can be used to generate or recognize many of the syllables of English. (*See* **Finite State Processing**)

The automaton can be used to generate a syllable, as follows. Node 1, dashed, is the start state of the machine. From here we can move to state 2, 3, or 4, by following an arrow. In doing so, one of the phoneme symbols on the arrow should be written down. For example, to move from state 1 to 2, the letter 's' must be written. From state 2 moves to states 3 or 4 are permitted. To move to state 3, a 'p', 't' or 'k' must be written. A path can be followed from state 1 to state 8, 9, or 10, writing out a sequence of symbols as we proceed, for example, 's, t, r, a, j, f' (representing the pronunciation of *strife*), or 'p, j, u, w' (*pew*). States 8, 9, and 10 are end states, shown by a double-ringed circle. One way of traversing states 1, 2, 4, 5, and 8 writes out 's, m, a, r', a word that does not occur in English, though it is a possible word. Continuing to state 10 would generate the word *smart*.

A partial encoding of Figure 1 that is amenable to computational implementation is Table 1, a symbol–state table. Starting in state 1 (the second row), select any of the numbered cells in the row, write out the symbol at the top of the column, and move to one of the states (rows) given in the cell. Dashed cells show which symbols cannot be written in that state. Cells containing more than one number have several possible next states, making this automaton nondeterministic: a mechanism

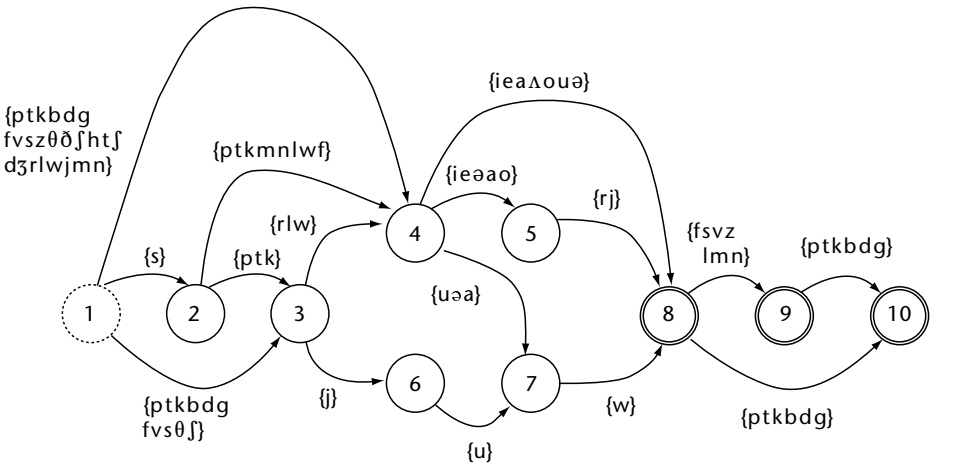


Figure 1. An automaton which defines many of the syllables of English.

Table 1. Symbol–state table encoding part of the automaton in Figure 1

Symbols	b	d	ð	ʤ	f	g	h	j	k	l	m	n	p	r	s	ʃ	t	tʃ	θ	v	w	z
States																						
1	3, 4	3, 4	4	4	3, 4	3, 4	4	4	3, 4	4	4	4	3, 4	4	2, 3, 4	3, 4	3, 4	4	3, 4	3, 4	4	4
2	–	–	–	–	4	–	–	–	3, 4	4	4	4	3, 4	–	–	–	3, 4	–	–	–	4	–
3	–	–	–	–	–	–	–	6	–	4	–	–	–	4	–	–	–	–	–	–	4	–
4	…	…																				
…																						

would have to guess at this point which state to go to next. Nondeterministic automata can be made deterministic, if necessary, but we might prefer to keep the nondeterminism: for example, we could associate probabilities with different state transitions.

Automata can also recognize strings, if the symbols on each arrow are read in and checked rather than written out. For instance, the string 'splajn' (*spline*) is acceptable: 's' takes the automaton from state 1 to 2, 'p' from 2 to 3, and so on. But 'sflajn' is unacceptable: 's, f' (as in *sphere*) takes the automaton from state 1 to state 4, but further moves are impossible as none of the transitions out of state 4 are labeled with 'l'. Thus, *sfline* is not a well-formed English word.

Finite-State Transducers

Most phonological computations involve relationships between two levels of representation; for example, converting English spelling to a phonemic transcription, or vice versa. We can modify Figure 1 accordingly: instead of single symbols, the arrows can be labeled with *pairs* of symbols. For instance, (th, /θ/) means that the pronunciation of 'th' can be /θ/, as in 'thin'. It can also be /ð/, as in 'this'. We could replace the label on the arrow from state 1 to 4 by a set of paired symbols, including {(c, /k/) (ph, /f/) (th, /θ/) (th, /ð/) (sh, /ʃ/) (ch, /tʃ/) (j, /ʤ/)}. Now transitions can be interpreted in four ways: (1) given the normal spelling, write out the corresponding phoneme; (2) given a phoneme, write out the spelling; (3) given a string in normal spelling and a transcription, check whether they correspond to one another; (4) given no inputs, generate paired spellings and transcriptions. Of these modes of use, the first might be used in speech synthesis, and the second in a speech recognition device (for spelling unknown words, perhaps). The fourth will generate a pronouncing dictionary of possible words. (*See* **Speech Recognition, Automatic**)

Similarly, paired phonemic and phonetic symbols enable us to generate or recognize the specific details of pronunciation at different positions in a word. Thus, some aspects of pronunciation variation according to context may be modeled. If instead of alphabetic phonetic symbols we use acoustic representations of slices of the speech signal, such as linear prediction coefficients, we can relate speech signals to their phonemic transcriptions. Because of the large number of distinct states of a signal, the construction of such a device has to be automatized. Hidden Markov Models (HMMs), in which a probability distribution is associated with each state transition, are an important extension of this approach. The probability distributions are determined empirically, by training the automaton on pairings of known signals with their transcription. Such devices are usually used for automatic speech recognition (Rabiner and Juang, 1993), though HMMs can also be used to generate a signal from a transcription (Donovan, 1996).

Finite-state techniques yield richer fruit which we do not have the space to savor here. Automata may be combined in various ways to yield more complex automata. Importantly, most phonological rules may be expressed as finite-state transducers. For example, the Japanese rule that /s/ is pronounced as [ʃ] when it occurs before /i/ is encoded as a transducer in Figure 2. An entire rule system can be translated into finite-state transducers and then automatically combined into a single, large automaton for efficient processing (Kaplan and Kay, 1994). Because of these discoveries, speech technology makes extensive use of finite-state transducers.

CONSTRAINT-BASED APPROACHES

Traditional phonological rules are applied in a certain order, rather like a simple program of the form 'first, change A into B; then, if C, change B into D, etc.' Computer science no longer relies on this *imperative* approach to programing. Techniques such as structured programing and object-oriented programing have broadened the repertoire of computational methods. In constraint-based approaches, the specification of languages is kept apart from questions of how computations such as generation, recognition, and translation are performed. Linguistic properties can be specified by declarative constraints, such as 'a word consists of one or more syllables', 'a syllable consists of an onset, a nucleus, and a coda', '/b/ can be a coda', '/a/ is a nucleus', or prohibitions such as '/b/ is not a nucleus'. Such constraints define a set of 'words', such as /bab/,

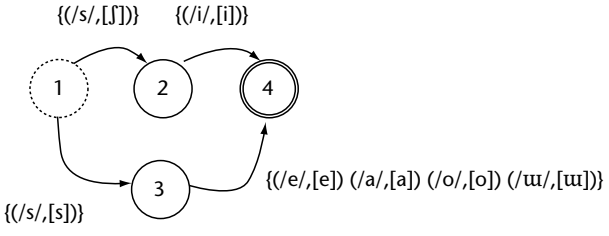


Figure 2. A finite-state transducer which encodes the phonological rule of Japanese that /s/ is pronounced as [ʃ] when it occurs before /i/, usually formalized as /s/ → [ʃ] / — /i/.

/babbab/, /babbabbab/, etc., and a set of nonwords, such as /bbbb/. If constraints are expressed as propositions of logic, questions such as 'Is /babbabab/ a word?' and 'What is the set of words that end with /k/?' may be determined using automated deduction. Because of the similarities between this approach to computation and current thinking in phonology, constraint-based methods are arguably the most popular approach to pure computational phonology (Bird and Klein, 1994). Most phonological rules can be expressed as constraints. For example, a feature-filling rule such as [+nasal] → [+voice] can be interpreted as the proposition: 'if x is [+nasal] then x is also [+voice]'. (*See* **Constraint-based Processing**; **Inference using Formal Logics**)

Constraints are usually required to be consistent, without contradictions. Theoretical phonologists have recently developed a constraint-based approach, optimality theory, in which constraints may conflict with one another. For instance, the constraints that 'a word consists of one or more syllables' and 'a syllable begins with a consonant' conflict with the statement that '*astronaut* is a word', because *astronaut* begins with a vowel. Optimality theory permits conflicts between constraints which express defaults or tendencies rather than exceptionless regularities. Where two constraints conflict, it is necessary to state which one has priority. Statements of prioritization define a ranking of the set of constraints, expressed 'A >> B >> C, D >> … X', meaning that constraint A has the highest priority and X the lowest; C and D have equal rank. A lower-ranked constraint does not need to be upheld if to do so would lead to the violation of a more highly ranked constraint. Various approaches to computation using this kind of system have been explored, usually employing finite-state methods yet again (e.g. Ellison, 1994). (*See* **Optimality Theory**)

CONNECTIONIST APPROACHES

Connectionist models are well-suited to the computation of relationships between distinct levels of representation, especially when the correspondence between the two levels is unclear or ill-formalized, because (like HMMs) they can be 'taught' correspondences between representations by presenting them with numerous examples of the relation in question. This makes it unnecessary to discover and debug a list of phonological rules, a task which often yields unforeseen errors. Faced with the difficulty of finding a complete and correct set of rules for e.g. grapheme to phoneme translation, training of a model is an appealing prospect. Connectionist models have been employed with some success for grapheme to phoneme translation (Sejnowski and Rosenberg, 1987), recognition of phonemes from speech acoustics (Waibel *et al*., 1989), learning syllabification and stress patterns (Larson, 1992; Gupta and Touretzky, 1994), predicting the next phoneme in a string of phonemes (Elman, 1990), and even acquiring phonological representations, given semantic, acoustic, and articulatory representations (Plaut and Kello, 1999). From the perspective of speech technology, connectionist approaches leave much to be desired, because of the degree of error in such models. However, from a cognitive science perspective, the similarities between the performance of these models and human behavior is most interesting.

References

Bear J (1990) Backwards phonology. In: Karlgren H (ed.) *COLING-90. Papers presented to the Thirteenth International Conference on Computational Linguistics*, vol. III, pp. 13–20. Helsinki, Finland.

Bird S (ed.) (1994) *Computational Phonology: First Meeting of the ACL Special Interest Group in Computational Phonology. Proceedings of the Workshop*. Association for Computational Linguistics. Bernardsville, NJ.

Bird S and Klein E (1994) Phonological analysis in typed feature systems. *Computational Linguistics* **20**: 455–491.

Coltheart M, Curtis B, Atkins P, and Haller M (1993) Models of reading aloud – dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608.

Donovan RE (1996) *Trainable Speech Synthesis*. PhD dissertation, University of Cambridge.

Ellison TM (1994) Phonological derivation in optimality theory. *International Committee on Computational Linguistics*, pp. 1007–1013. Kyoto, Japan.

Elman J (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Gupta P and Touretzky DS (1994) Connectionist models and linguistic theory: investigations of stress systems in language. *Cognitive Science* **18**: 1–15.

Johnson CD (1972) *Formal Aspects of Phonological Description*. The Hague: Mouton.

Kaplan RM and Kay M (1994) Regular models of phonological rule systems. *Computational Linguistics* **20**: 331–378.

Larson GN (1992) *Dynamic Computational Networks and the Representation of Phonological Information*. PhD dissertation, University of Chicago.

Lowe JB and Mazaudon M (1994) The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics* **20**: 381–417.

Maxwell M (1994) Parsing using linearly ordered phonological rules. In: Bird S (ed.) *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology*. Las Cruces, 1994, pp. 59–70.

Plaut DC and Kello CT (1999) The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In: MacWhinney B (ed.) *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rabiner L and Juang B-H (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: PTR Prentice Hall.

Sejnowski TJ and Rosenberg CR (1987) NETtalk: a parallel network that learns to pronounce English text. *Complex Systems* **1**: 145–168.

Waibel A, Hanazawa T, Hinton GE, Shikano K and Lang KJ (1989) Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**: 1888–1898.

Williams SM (1994) Lexical phonology and speech style: using a model to test a theory. In: Bird S (ed.) *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology*. Las Cruces, 1994, pp. 43–57.

Further Reading

Bird S (1995) *Computational Phonology: A Constraint-Based Approach*. Cambridge, UK: Cambridge University Press.

Jurafsky D and Martin JH (2000) *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

Phonology

Introductory article

IM Roca, University of Essex, Colchester, UK

CONTENTS

Phonology is the branch of linguistics that studies the patternings of sound in language(s), and formalizes them by means of networks of distinctive features and higher abstract structure organized into levels of representation related by rules.

INTRODUCTION

The sounds musicians make are, typically, not random, but organized into tunes and melodies. Similarly, the sounds of language obey certain principles and are patterned in certain ways. Put differently, what hits our ear may be the blast of the trumpet or the beat of the drum, but what we perceive (and enjoy) is the melody of which this blast and this beat are a part. The melody (which can be written down as a score) is an abstract pattern: abstract if nothing else because we cannot hear it all at once, but must keep mental track of it as the music plays along. Without the score (written or mental) there would be no melody, and therefore no music to speak of, just random sound. Similarly in language, the sounds we produce follow an abstract score. It is the job of *phonology* to discover the substance of such a score and the properties that govern it (the entry on *phonetics* provides the necessary background on speech sound, and should be consulted first). (*See* **Phonetics**)

PHONEMES AND ALLOPHONES

The idea of the *phoneme* underpins the English spelling system (among others). Consider the sounds that correspond to the letters *p*, *t*, *c* in words like 'pie', 'tie', 'corn', and compare them with those in the words 'spy', 'sty', 'scorn': the sounds are different, even if they are represented with the same letter. In particular, in the first set of words, but not in the second, the sound corresponding to each of our letters is followed by a distinct puff of air ('aspiration') which, if exaggerated, would be powerful enough to blow out a candle. The sounds represented by *p* in 'pie' and 'spy' are therefore objectively different from each other. Does this mean that we are wrong in spelling them both with the same letter? Well, although different, the two sounds are still pretty close to each other: they both involve retention of air behind the lips, and its subsequent sudden release, with no vocal fold vibration (i.e. no concurrent humming, or 'voice'): technically, they are both voiceless oral labial stops. Moreover, were we to spell the *p* of 'pie' as, say, *ph*, with the *h* representing the aspiration (in conventional English spelling *ph* is of course usually equivalent to *f*, [f]), to keep it distinct from the *p* of 'spy', which has no aspiration, what exactly would we be gaining? Phonetic accuracy, to be sure: the phonetic symbol for the *p* of 'pie' is indeed [ph] (or [p^h]). But will such phonetic accuracy help us in any way? The answer is, clearly, no, at least for ordinary purposes (it would obviously help us if we were carrying out a phonetic investigation). The reason is that the presence of aspiration in English is totally predictable: it occurs, for instance, at the beginning of words, but not after *s*. Therefore, there is no point in writing in such information: it does not carry any functional load and therefore it is redundant. Conventional English spelling is, accordingly, quite right in representing the *p*s of 'pie' and 'spy' identically.

Compare now a language like Thai. Thai also has the two types of *p* that English has: with and without the aspiration. Consider such minimally contrasting Thai word pairs as *phet* 'spicy' versus *pet* 'duck', *phraang* 'to disguise' versus *praang* 'face', *phaw* 'to barbeque' versus *paw* 'to blow air', and many others, where the *h* following the *p* indeed stands for aspiration. What do these words show

us? They show us that aspiration does carry a functional load in Thai, since it keeps words apart: [ph]*et* from [p]*et*, etc. In English, by contrast, there simply cannot be a word *su*[ph]*er* distinct from *su*[p]*er*, for instance. This is where the 'phoneme' comes in: /ph/ and /p/ are distinct *phonemes* in Thai, but not in English, even though the two sounds [ph] and [p] also exist in English, as realizations, or *allophones*, of a single phoneme /p/ (phonemes are conventionally enclosed in slant bars /…/, and allophones in square brackets […]).

The advent of the phoneme marks the birth of modern phonology: besides the sounds we hear, we now have an abstract sound level underlying them (abstract because it is structural, not phonetic). Sometimes, the contents of the level of abstract sound and those of the level of concrete sound are nondistinct (cf. the Thai correspondences /ph/ ↔ [ph], /p/ ↔ [p]), but sometimes they are distinct (cf. the allophone [ph] of the phoneme /p/ in English). When the contents of the two levels are distinct, we can (somewhat anachronistically) relate them formally by means of context-sensitive rules, each connecting a phoneme to an allophone. For instance, we can 'derive' the English allophone [ph] from the phoneme /p/ by means of the rule 'p → ph /#__', where '→' relates input and output, '/' announces the context, '#' symbolizes the edge of the word, and '__' stands for the position where the conversion, or mapping, takes place. In the absence of one such allophonic rule, the identity mapping is assumed: simply, the phonemic form manifests itself unchanged phonetically.

At this juncture a question arises, why should we assume /p/, or /ph/, at all? (Obviously, if we assume /ph/ we will have to reverse our rule of allophony.) The point is that both [p] and [ph] are *surface* realizations: there is no *a priori* reason to think of either of them as basic. Instead, we could assume that the basic form is something in between [p] and [ph], which we will formalize as /P/. But what do we mean by something in between? In order to answer this question we must first work out what exactly we mean by [p], or [ph].

DISTINCTIVE FEATURES

By [p], of course, we mean the *p* of 'spy', and by [ph] the *p* of 'pie'. But what are these? If you think of it, the pronunciation of a *p* (whether it is the *p* of 'spy' or the *p* of 'pie'), or of any other sound, is rather like a tone an orchestra plays: the product of *several* instruments playing together in certain *ways*. What are the instruments for *p*? Well, mainly the lips, which 'play' the *p* by *stopping* the airflow on its way out of the mouth (if they were to *restrict* it, rather than stop it, we would get a sound similar to *f*; *f* is actually produced with the lower lip and the upper set of teeth). Other available instruments are: the vocal folds in the larynx, which are wide open during the production of *p* (not vibrating, as they are for the *b* of 'rubber'), and the nose cavity, which is shut off for both *p* and *b* (air would be going through if we were pronouncing *m*). The point, therefore, is that a sound like *p* is a composite of several gestures. It takes only a short step to designate these gestures, rather than the end product, as the primitive elements of linguistic sound, that is, the *phonological primes*. This step (foreshadowed at several points in history) was taken in the 1950s by Roman Jakobson, developing work by Nikolai Trubetzkoy. This means that, in some way, the layman is mistaken in thinking that 'pie' is made up of two sounds, respectively spelled out *p* and *ie*: the sound spelled *p* (and similarly for the sound spelled *ie*) is not a unitary entity, but a conglomeration of more basic elements, as we have just seen, rather like molecules are made up of atoms, or atoms of subatomic particles.

Such primitive elements of phonology are known as *distinctive features*. The original set of distinctive features set up by Jakobson (eventually in collaboration with Morris Halle) was deliberately rather small (12 features in all), and was orientated towards acoustics, the physics of sound. With the advent of 'Generative Phonology' in the 1960s, this stance underwent a radical shift, and in 1968 Chomsky and Halle declared the set an open one, actually listing over 20 features. Moreover, they based the description of the features on their articulatory, rather than acoustic, make-up, that is, on the mechanics of their production in the mouth and allied articulators. For example, they differentiated the vowels in words like 'bead', 'bed' and 'bad' by means of features named 'high' and 'low', which refer to the raising and lowering of the body of the tongue, respectively.

Whichever way we approach distinctive features, the strongest phonological justification for them lies in the fact that they define 'natural classes', that is, classes of sounds that behave similarly. We shall illustrate with data from English plural formation. Before doing so, however, we shall answer our pending question and state that indeed nothing prevents us in principle from postulating an *archiphoneme* /P/ as underlying the allophones [ph] and [p]. Assuming that the difference between [ph] and [p] lies in their opposing specification for a feature [±aspiration] (distinctive features are usually formalized as binary, that is, as

Table 1. Definitions of distinctive features[a]

+consonantal	obstruction to the airflow along the median plane of the mouth: [p, b, f, m, t, d, n, l, r, θ, s]
+sonorant	equal air pressure in the mouth and outside: [m, n, ɲ, ŋ, l, ʎ, r, ɹ; i, e, ɛ, a, ɔ, o, u, y, ø, œ]
+continuant	no airflow stoppage along the median plane of the mouth: [f, v, θ, ð, s, z, ʃ, ʒ, x, ɣ; i, e, ɛ, a, ɔ, o, u]
+strident	a hissing sound: [s, z, ʃ, ʒ, ɬ, ɮ]
+nasal	airflow through the nasal cavity: [m, n, ɲ, ŋ]
+lateral	unimpeded airflow over the sides of the tongue: [l, ʎ, ɬ, ɮ]
+voice	concomitant vocal fold vibration: [b, v, m, d, ð, z, n, l, ɹ, g, ŋ; i, e, ɛ, a, ɔ, o, u, y, ø, œ]
+aspiration	puff of air following release of blockage: [ph, th, kh]
labial	involving the lips: [p, b, f, v; u, ʊ, o, ɔ, ɒ, y, ø, œ]
+round	with rounding of the lips: [u, ʊ, o, ɔ, ɒ, y, ø, œ]
coronal	involving the blade of the tongue: [t, d, θ, ð, s, z, ʃ, ʒ, n, l, r, ɹ]
−anterior	coronal sound articulated behind alveolar ridge: [ʃ, ʒ, ɹ]
+distributed	coronal sound articulated with long constriction: [θ, ð, ʃ, ʒ]
dorsal	involving the body of the tongue: [c, ɟ, ɲ, k, g, ŋ; i, e, ɛ, a, ɔ, o, u, y, ø, œ]
+high	raising of the body of the tongue: [c, ɟ, ɲ, k, g, ŋ; i, ɪ, u, ʊ, y]
+low	lowering of the body of the tongue: [æ, ɑ, ɒ]
+back	retraction of the body of the tongue: [k, g, x, ɣ, ŋ; ɑ, ɒ, ɔ, o, u]
radical	involving the root of the tongue: [i, e, ɛ, a, ɔ, o, u, y, ø, œ]
+ATR	advancement of the root of the tongue: [e, i, u, o]

[a]Where a value (±) has been specified, the opposite value has the opposite definition; where there is no value specified, the feature is assumed to be monovalent, i.e. either present or absent.

carrying one of the two values + or −), i.e. [+aspiration] and [− aspiration], respectively, /P/ will be *underspecified* for this feature, which will accordingly be absent from its representation.

We make things more concrete in Table 1, with a succinct definition of each distinctive feature and a selection of sounds (drawn mainly from English) that contain the feature value defined in the table.

ENGLISH PLURAL ALLOMORPHY

We can now turn to the English plural. Orthographically, regular English plurals are formed by the addition of *-s* or *-es* to the singular: 'book**s**', from 'book', 'field**s**' from 'field', 'pie**s**' from 'pie', 'window**s**' from 'window', on the one hand, and 'ash**es**' from 'ash', 'church**es**' from 'church', or 'class**es**' from 'class', on the other (plurals like

'men' from 'man', or 'mice' from 'mouse', are of course irregular). English spelling is, however, a very poor reflection of actual sound. Indeed, there are three alternative realizations ('allomorphs') of the English regular plural: [z], [s], and [ɪz], as in *field*[z], *book*[s], and *ash*[ɪz], respectively ([ɪz] can be [əz] in some accents). The question is – why doesn't the English regular plural have a uniform manifestation?

Let us consider the *alternation* between [z] and [s] first: why does the plural show up as [z] in *field*[z], but as [s] in *book*[s]? The answer is that the plural suffix comes out voiced if the preceding sound is voiced, and voiceless if the preceding sound is voiceless: [d] is voiced, hence *fiel*[dz], but [k] is voiceless, hence *boo*[ks]. Now, how are we going to formalize this alternation between [z] and [s]? First, of course, we have to decide which is the underlying form of the suffix. An argument for an underlying form /z/ (rather than /s/) can be made on the basis that the surface manifestation of the plural suffix is [z] when the preceding segment is not a consonant, but a vowel (*spa*[z] from 'spa', *pea*[z] from 'pea', *zoo*[z] from 'zoo', etc.): it is unusual for a vowel to affect the voice status of the following consonant. Assuming thus /z/, we could formulate a rule to change it to [s] after /k/, to get *book*[s], as follows (the '+' in the environment stands for the boundary between two 'morphemes', that is, the lexical or grammatical elements that make up a word: here, the base morpheme and the plural morpheme, the latter a 'suffix' because it *follows* the base):

$$\mathrm{z} \rightarrow \mathrm{s} / \mathrm{k} + \_\_ \qquad (1)$$

This rule will obviously implement the change we are seeking, as can be seen in the corresponding *derivation* in (2) (for simplicity, we provide phonetic symbols only for the segments under examination):

underlying representation	*boo*/k + z/	
rule (1)	s	
surface representation	*boo*[ks]	(2)

Now, because rule (1) requires /k/ in the left environment, it will not apply in forms like *soaps*, *mats*, *cliffs*, etc. If the underlying representation is /z/ and the rule does not apply, the plural of these forms is predicted to be **soap*[z], **mat*[z], and **cliff*[z], contrary to fact (forms which do not correspond to fact are conventionally represented starred; the correct plurals are of course *soap*[s], *mat*[s], *cliff*[s]). How shall we proceed? One alternative would be to say that for these words the underlying representation of the plural morpheme is /s/, rather than /z/. This is only a pseudosolution, however, for at least two reasons. First, generative phonology is grounded in the assumption that each family of alternants derives from a single underlying form: this is indeed the way their relationship (the fact that they constitute a family) is formalized in this theory. Second, the solution we are exploring is totally *ad hoc*: it handles the facts, but it does not explain why singulars ending in /p/, /t/, /f/ should not combine with the same underlying plural form as those ending in /k/.

Let us accordingly try a different tack. A better solution would be to postulate a similar rule for each of the consonants in question. For example, we can account for *soap*[s] by means of the following rule:

$$\mathrm{z} \rightarrow \mathrm{s} / \mathrm{p} + \_\_ \qquad (3)$$

The effectiveness of this rule is demonstrated in the following derivation:

underlying representation	*soa*/p + z/	
rule (3)	s	
surface representation	*soa*[ps]	(4)

Likewise, of course, for *t*, *f*, etc. The problem with this solution, however, is that it requires a different rule for each different contextual consonant. Moreover, it is quite unexplanatory, because it does not tell us why it is the consonants in question (/p/, /t/, /k/, /f/, etc.) that trigger the change of /z/ into [s], let alone *why* they trigger it. Indeed, on this approach the fact that it is precisely these consonants that trigger precisely this change is quite arbitrary.

NATURAL CLASSES

We shall now bring distinctive features into the picture. Assuming features, the only difference between [z] and [s] concerns their respective value for [±voice]: [z] is [+voice], because it is voiced, whereas [s] is [−voice], because it is voiceless. Now, notice that *all* the consonants that trigger the change z → s are voiceless, i.e. [−voice]. If so, we can compress our rules above (and the remaining rules we have not formulated) into one simple rule:

$$\mathrm{z} \rightarrow \mathrm{s} / [-\mathrm{voice}] + \_\_ \qquad (5)$$

Rule (5) says that plural /z/ becomes [s] if the preceding sound is voiceless ([−voice]), as are all the consonants in question. If the sound that precedes the /z/ is voiced, the rule will not apply, because the contextual conditions it imposes are not met. This is indeed correct (cf. *fiel*[dz], *sp*[ɑz], etc.). Therefore, our approach (which makes crucial use of features) successfully passes the empirical

test. Moreover, we are now seeing that distinctive features define *natural classes* of sounds: in this case, a class of voiced sounds (formalized as [+voice]) and a class of voiceless sounds (formalized as [−voice]).

There is actually more to it. Why should the presence of a voiceless ([−voice]) consonant on the right edge of the base trigger the change from /z/ to [s] in the suffix? The answer becomes obvious when we remember that the only difference between these two consonants concerns the feature [±voice]: the value for [±voice] of the plural suffix therefore agrees with the value for [±voice] of the segment that precedes it. In particular, the underlying + value of this suffix changes to − when the preceding sound is [−voice]:

$$\text{z} \rightarrow [-\text{voice}] / [-\text{voice}] + \_\_ \qquad (6)$$

Agreement in the value of a feature in adjacent segments ('assimilation') is a common occurrence in the phonology of languages.

We now turn to the third alternant of the English plural suffix, [ɪz], illustrated in words like *ash*[ɪz], *massag*[ɪz], *atlas*[ɪz], *quizz*[ɪz], *finch*[ɪz], *cabbag*[ɪz]. Where does this alternant come from? Notice that the rule we have currently available is unable to derive it (e.g. for *ash*[ɪz]):

underlying representation	*a*/ʃ + z/	
rule (6)	s	
surface representation	**a*[ʃs]	(7)

We obviously need an additional rule to derive the correct output. On the basis of *ash*[ɪz], we can formulate this rule as follows:

$$\text{z} \rightarrow \text{ɪz} / \text{ʃ} \_\_ \qquad (8)$$

The incorporation of this rule has the desired effect on the derivation (we shall explain shortly why rule (6) fails to apply in this derivation):

underlying representation	*a*/ʃ + z/	
rule (8)	ɪz	
surface representation	*a*[ʃɪz]	(9)

Our success, however, is tempered by the fact that the approach does not work for forms which do not have a base ending in /ʃ/ and which nonetheless also select the plural allomorph [ɪz]: *massag*[ɪz], *atlas*[ɪz], *quizz*[ɪz], *finch*[ɪz], *cabbag*[ɪz]. The solution, once more, lies in the incorporation of distinctive features. In particular, the consonants that take [ɪz], viz. /ʃ/, /ʒ/, /s/, /z/, /ʧ/, and /ʤ/, share the specifications [+coronal, +strident], where [+coronal] designates a segment articulated with the blade of the tongue (the front part of the tongue, that we can stick out without discomfort), and [+strident] a fricative segment that sounds 'rough', like [ʃ] or [s], for instance (as against [ɵ], which sounds 'smooth'). We can accordingly give rule (8) a more general scope, as follows:

$$\text{z} \rightarrow \text{ɪz} / \begin{matrix}[+\text{coronal}] \\ [+\text{strident}]\end{matrix} + \_\_ \qquad (10)$$

All the plurals *ash*[ɪz], *massag*[ɪz], *atlas*[ɪz], *quizz*[ɪz], *finch*[ɪz], *cabbag*[ɪz] are taken care of by this rule.

As in the previous case, the question arises of why the process should occur at all. Also as in the previous case, the answer requires decomposing into distinctive features. In particular, /z/ is also specified as [+coronal, +strident]: evidently, thus, a sequence of two [+coronal, +strident] consonants is not allowed in English (nor in many other languages). In order to circumvent this prohibition, the phonology of the language provides a rule inserting ('epenthesizing') a vowel between the two [+coronal, +strident] consonants, rather as the umpire stands between the two boxers when they need to be separated. Strictly speaking, therefore, the rule must be formulated as follows ('∅' stands for the null symbol, or zero; 'z' informally stands for the remainder of the features that make up [z]):

$$\emptyset \rightarrow \text{ɪ} / \begin{matrix}[+\text{coronal}] \\ [+\text{strident}]\end{matrix} + \_\_ \begin{matrix}[+\text{coronal}] \\ [+\text{strident}]\end{matrix} \qquad \text{z} \qquad (11)$$

The motivation for the process is now transparent. The formal linchpins are, of course, the distinctive features which make possible direct reference to natural classes of sounds.

RULE ORDERING

We must now explain why the application of rule (11) excludes the application of its competitor, rule (6). If it did not, an incorrect output would, of course, be derived:

underlying representation	*a*/ʃ + z/	
rule (6)	s	
rule (11)	ɪ	
surface representation	**a*[ʃɪs]	(12)

The output **a*[ʃɪs] does not match the correct form *a*[ʃɪz].

The solution traditionally adopted consists in imposing on the sequence of rules the specific ordering that yields the targeted output. In the present case, all we have to do is invert the order of the two rules in derivation (12):

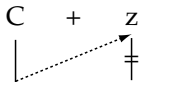
underlying representation	*a*/ʃ + z/
rule (11)	ɪ
rule (6)	not applicable
surface representation	*a*[ʃɪz] (13)

As can be seen, rule (6) is now inapplicable, for the simple reason that the environment it requires has been destroyed by the previous application of rule (11). In technical parlance, the order rule (11) > rule (6) that we are imposing on these two rules is a 'bleeding order', because the application of rule (11) removes (bleeds away) the environment required by rule (6), that follows it in the proposed ordering.

AUTOSEGMENTAL PHONOLOGY

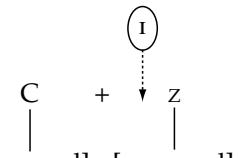
As a matter of fact, the formalization we have been using for our rules is now a little old-fashioned. In particular, our two rules of Voice Assimilation (6) and Epenthesis (11) would nowadays tend to be formalized as follows:

Voice Assimilation:



(14)

Epenthesis:



(15)

The new formalization is visually more transparent than the previous one. Consider the rule of voice assimilation (14) first. We can immediately see what is going on: given a sequence C (= any consonant), + (= morpheme boundary), and /z/, such that C is associated to the valued feature [−voice] (notice the continuous line linking the two), this minus-valued feature takes over the neighboring segment /z/ (notice the arrow-headed broken line), which concomitantly loses its previous association to [+voice] (notice the crossing out of the corresponding association line). The end result, of course, is the devoicing of /z/, which becomes [s] (implicit in the figure in the new association of /z/ to [−voice]). With regard to epenthesis (15), the ringed ɪ is shown being inserted between the two [+coronal, +strident] segments.

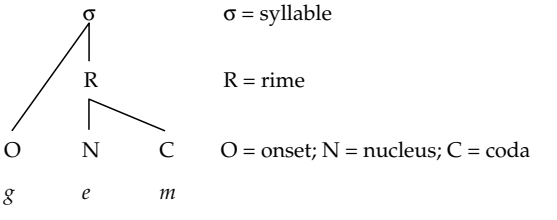
Visual clarity aside, the main advantage of the new formalization (commonly referred to as *autosegmental* phonology, perhaps not very transparently) is that it affords each feature true formal autonomy in a three-dimensional space. We do not have room to expand on the details of this autonomy here, nor to justify its superiority. Suffice it to say in the present context that there is practically unopposed consensus in the field that phonological representations indeed are autosegmental.

SYLLABLES

The autosegmental modeling of segments represented a significant step forward in our understanding of phonology. However, we will now see that it needs to be supplemented.

Consider the two very similar English words 'attractive' and 'atlantic' (similar in particular in their initial sequences *attrá*…, *atlá*…). Despite this similarity, the respective pronunciations are quite different. First, the first *a* of 'attractive' reduces to 'schwa' (it sounds [ə], like the article *a* in the phrase 'a tractor'), but not so the first *a* of 'atlantic', which sounds [æ], like the *a* in 'at (Lancaster)'. Second, the *t* of 'attractive' always comes out as a full *t*, indeed an aspirated *t*: [th]. By contrast, the *t* of 'atlantic' is weaker: definitely not aspirated, and possibly even reduced to a glottal stop [ʔ] again as in 'at Lancaster'. Last, the *l* of 'atlantic' is fully voiced, similarly to the *l* of 'land'. The *r* of 'attractive', by contrast, usually loses its voice, totally or partially: it does not quite sound like the *r* of 'rack'. These differences pose an obvious puzzle.

The solution involves recognizing that segments are not like carriages in trains, each linked to the next. Rather, segments cluster in ways comparable to planetary systems, each centered around a sun, surrounded by planets. In particular, segments organize themselves into *syllables*, typically with a vowel as the *nucleus*, and possibly with consonants on either side: in the *onset* when preceding the vowel, and/or in the *coda* when following the vowel. Such internally structured syllables can be graphically represented as in the following diagram, where the nucleus and the coda are grouped under a common node *rime*, responsible for rhyme in verse, among other things:



Each syllabic subconstituent onset, nucleus, and coda can either be simple, when it dominates only one segment, or complex, when it hosts more than one segment: compare 'pan', 'plan', 'land', 'planned', etc., all monosyllabic (= 1 syllable), 'planner', 'landing', 'landless', etc., bisyllabic (= 2 syllables), 'landlessness', etc., trisyllabic (= 3 syllables), and so on. Speakers usually have very clear intuitions about the number of syllables in the words of their language, and even about the boundaries between syllables, although in English the location of syllable boundaries may be somewhat blurred (would you syllabify 'planning' as *plann.ing* or as *pla.nning*, where the dot signals the syllable boundary?). Poets make privileged use of intuitions like these by applying them to poetic metre, which obviously involves syllable counting. In some languages (Spanish, for instance, although not English), a convention dictates that in ordinary writing lines can only be broken between syllables.

We can now return to the contrasts between 'attractive' and 'atlantic'. Suppose we syllabify the two sequences differently: *a.ttrac*… versus *at.lan*… Two questions ensue: first, what is the point of doing so, and second what is the empirical justification (we should not do things in an *ad hoc* manner just to suit our analytic convenience, but, rather, ground them in true fact). Let us answer these two questions in turn. The point of postulating these two different syllabifications is that they allow us to understand the differences in pronunciation we pointed at. In particular, *a* reduces to schwa in 'attractive' because it is not protected by a coda consonant (in 'atlantic' it is); on the other hand, *t* is always strong in 'attractive' because it occupies the syllable onset, whereas in 'atlantic' it occupies the syllable coda, and therefore it can undergo weakening; finally, the *r* of 'attractive' can lose its voicing because of the influence of the preceding *t*, with which it shares the onset (the [h] of [th] invades the *r*, so to speak): there is no such influence of the *t* of 'atlantic' on the *l* because the two segments belong in different constituents, and therefore do not interact (at least to a degree that makes devoicing possible). So, what seem to be three independent (and puzzling) phenomena are unified and made sense of when we refer them to syllable structure, the case for which is accordingly strengthened. Moreover, in the case at hand there is additional independent evidence for the different syllabic allocation of the *t* in the two words: while there are many English words beginning with the sequence *tr* ('tractor', 'try', 'troglodite', etc.), there is *none* beginning with *tl*. Crucially, this is not an accidental gap, but a principled one, probably attributable to the fact that *t* (and *d*) and *l* share their place of articulation: [l] is like [d] with the sides of the tongue lowered; in turn, [t] is like [d] without the voicing. Languages do tend to eschew similar adjacent elements within the same structural unit (a bit like close blood relations eschew marriage!).

SYLLABLE STRUCTURE AND STRESS

Syllables (and their internal structure) manifest themselves in many other ways. We shall mention just a couple of them here. In all English words, one of the syllables carries special prominence, or *stress*. For instance, we know whether 'torment' corresponds to a noun ('a horrible torment') or a verb ('I don't wish to torment you') because of the location of the stress (*tórment* and *tormént*, respectively). Native speakers know intuitively where to put the stress in each word, and phonologists have discovered the laws that govern such distribution in many languages, English among them. One reason syllables are relevant to stress is, therefore, that stress is carried by syllables (here *tor* and *ment*, respectively).

There is an additional, perhaps less obvious, aspect of the relevance of syllables to stress: in some languages, the location of stress is influenced by the structure of syllables. Consider English words like 'algebra', 'asparagus', 'hippopotamus', 'aluminium', 'aluminum', on the one hand, and 'agenda', 'amalgam', 'memento', 'asbestos', 'amanuensis', on the other, with the stress falling on different syllables: antepenultimate (*álgebra*) and penultimate (*agénda*), respectively. Assuming (abstracting away much complexity) that English nouns target antepenultimate stress, the question arises why stress is penultimate in *agénda*, rather than antepenultimate. One answer (on the right track, as we will see) is that it is drawn in, in some way, by the consonant cluster *nd*. The problem with this explanation is that *álgebra* also has a consonant cluster in the same position: *br*. However, reflection will reveal that the two clusters are syllabified differently: *n.d* and *.br*, respectively. In other words, the two consonants of the cluster belong in different syllables in *agen.da*, but in the same syllable in *alge.bra*. As a result of these syllabifications, the syllable *gen* in *agenda* includes a coda consonant (*n*), and is therefore 'closed', whereas the corresponding syllable *ge* in *algebra* does not, and is 'open'. If we now assume that closed syllables (as well as syllables containing a long nucleus) are 'heavy', and open syllables 'light', we can say that a heavy syllable in penultimate position stops the leftward journey of stress to the antepenultimate

syllable, the desired target in English nouns, as we said (final syllable heaviness is obviously irrelevant here). Syllable structure, correspondingly, has a direct effect on the positioning of stress.

SONORITY

We still have not explained the reason for the syllabifications *alge.bra*, *agen.da*, rather than *algeb.ra*, *age.nda*, which would yield the opposite results with regard to stress (**algébra*, **ágenda*). The reason is twofold. First, each syllable is a mountain-shaped cluster of sonority, where by sonority we mean the amount of actual sound carried by each segment. It is not difficult to see that vowels carry more sonority (i.e. project more sound) than consonants: for instance, we use vowels, rather than consonants, to call someone's attention (*oy*, or *eh*, rather than *ppp*, or even *fff* or *lll*). The point about syllables is that the nucleus must constitute the sonority peak, with the sonority of the onset and the coda sloping down to the respective syllable boundaries, a profile commonly known as *sonority sequencing*. A few sets of sonority relations are now provided, with examples in brackets ('>' = 'more than'):

a. vowels ([i]) > consonants ([s])
b. low vowels ([ɑ]) > mid vowels ([ɛ]) > high vowels ([i])
c. obstruent consonants ([b]) > sonorant consonants ([m])
d. stop obstruents ([b]) > fricative obstruents ([v])
e. voiced stops ([b]) > voiceless stops ([p])
f. voiced fricatives ([v]) > voiceless fricatives ([f])
g. liquid consonants ([l]) > nasal consonants ([n])
h. rhotic liquids ([r]) > lateral liquids ([l])

These partial sets can be combined into a global *sonority scale*, as follows:

low vowels > mid vowels > high vowels >
rhotics > laterals > nasals > voiced fricatives >
voiceless fricatives > voiced stops >
voiceless stops

The problem with a syllabification *age.nda* is that a syllable *nda* would not be mountain-shaped with regard to sonority: instead of the sonority of its segments sloping up to the nucleus (such that $S_1 < S_2 <$ Nucleus), it contains a sonority trough between the *n* and the *a* (*n* > *d* < *a*). The presence of this sonority trough rules out *nda* as a possible syllable, and the *n* and the *d* must instead be assigned to different syllables: *agen.da*.

A syllabification *algeb.ra* would not violate intrasyllabic sonority sequencing, since both *geb* and *ra* would be mountain-shaped with regard to sonority. However, such a syllabification would infringe another syllabification principle adhered to by many languages, English among them: the principle that onsets must be maximized, that is, made as large as is compatible with the other syllabification principles (the principle of sonority sequencing paramount among them). Therefore, on onset maximization 'algebra' is syllabified *alge.bra* (not **algeb.ra*), like 'attractive' is syllabified *a.ttractive* (not **att.ractive*), although 'atlantic' does need to be syllabified *at.lantic* (not **a.tlantic*), as we explained above.

The model of phonology we have developed up to this point contains strings of autosegmentalized clusters of distinctive features making up segments. In turn, segments are organized into syllables, each with an obligatory nucleus (usually a vowel) and with an optional onset and/or coda (typically made up of consonants). One of the word's syllables carries special prominence, or stress. The composition of the syllable can have an effect on the location of stress: heavy rimes arrest the train of stress in some languages, like English, as we saw.

OPTIMALITY THEORY

Before ending our presentation, we shall refer briefly to a recent approach ('Optimality Theory') which substitutes surface constraints for rules and derivations. Remember that we have accounted for the surface alternation of the English plural morpheme ([z], [s], [ɪz]) by postulating a common underlying representation /z/ and a set of two ordered rules (Epenthesis (15) > Voice Assimilation (14)) mapping this representation onto the appropriate surface allomorphs. Suppose instead that we allow a general, universal device (the 'generator', GEN) to generate (by means of a rather general universal set of rules) an indefinite number of surface forms, not just the correct ones, but any number of them. The phonological grammar of English would now have the task of *selecting* the correct form in each case, discarding the rest.

Let us see how this would be done in the case of plural allomorphy. Suppose that the phonology of English includes the two surface constraints in (16) and (17), conveniently streamlined to keep the presentation simple ('α', a 'Greek letter variable' = + or −):

NO VOICE CLASH (NVC): *[αvoice][−αvoice] (16)

NO STRIDENT CORONAL SEQUENCE (NSCS): *[+coronal][+coronal] [+strident][+strident] (17)

It is clear that NVC will rule out **boo*[kz], and NSCS **a*[ʃz] or **a*[ʃs] (**a*[ʃz] will also be ruled out by NVC, of course). This means that these forms will not be allowed to occur, exactly as sought.

It is not obvious yet how the *correct* forms are obtained, however. For instance, a form **boo*[z], from *boo*/kz/, would also comply with our two constraints, as would **boo*[k]. Similarly, **a*[ʃ], **a*[z] or **a*[s] would all be legitimate plurals of *ash* (also other possible forms, too numerous to be listed here). How can we ensure that the procedure selects precisely *boo*[ks] and *ash*[ɪz], respectively? The answer is that constraints (16) and (17) are not the only constraints that control the surface plural form (or, indeed, other surface forms). Of direct relevance here is a family of *faithfulness* constraints that aim at ensuring the identity of the surface form with the underlying form (its 'faithfulness' to it). Clearly, in the case at hand the faithfulness constraints will strive for the surface forms **boo*[kz] and **a*[ʃz]. How can we get the correct forms *boo*[ks] and *a*[ʃɪz], then?

The answer is that, in any given language, not all the constraints have the same *power*: some are more powerful than others. In English, NVC and NSCS overpower the relevant faithfulness constraints, whichever these are. These power relations are formalized by means of constraint 'ranking', here as follows (>> signifies 'ranked higher than'):

NVC, NSCS >> faithfulness

It is now clear why the plural surface forms cannot always replicate the underlying forms: faithfulness must give way to prevent violations of NVC or NSCS. The question is – why does it give way the way it does? That is, why do we get *boo*[ks], rather than, for example, **boo*[gz], or *a*[ʃɪz], rather than, for example, **a*[z]? The answer is implicit in our discussion: There must be some constraints ranked higher than the constraints discussed so far which therefore rule out these forms. In particular, the faithfulness constraints affecting the base are ranked higher than the faithfulness constraints affecting the suffix, and the constraints prohibiting deletion are ranked higher than the constraints prohibiting insertion, such that *a*[ʃɪz] is favored over **a*[ʃ], for instance:

some faithfulness constraints >> NVC,
NSCS >> other faithfulness constraints

For reasons of space we cannot go into further technical details here. However, we feel reasonably confident that the basic mechanics of Optimality Theory are satisfactorily summarized in the outline we have managed to offer.

COGNITIVE RELEVANCE

The cognitive relevance of phonology, as of linguistics in general, must be sought at the level of *competence*, the knowledge of the language its speakers need to have in order to engage in actual *performance* (this dichotomy originates in the work of Noam Chomsky). It is perhaps helpful to relate competence in language to programs in the world of computers: hidden from the user's view, but obviously a prerequisite to what we see on the screen. However, formal models of phonology are not supposed literally to represent actual operations in the mind, although a mental correlate of some kind is usually assumed, at least implicitly.

A central issue which arises out of formal models like the ones we have presented here concerns learnability: how can the learner of the language arrive at the specific analyses (underlying representations included) from the surface data? In particular, certain rule orderings render certain surface forms *opaque*, in as much as the generalizations embodied in some rules are irrecoverable directly from the surface. If so, how can the learner induce such rules and, *a fortiori*, the underlying representation? The problem obviously disappears if the power to order rules externally (i.e. by the analyst's fiat) is removed, and indeed such power is commonly, although not universally, disfavored by practitioners. Among the advantages claimed for Optimality Theory is precisely its flat structure, with no derivations: this architecture obviates rule ordering. Optimality theorists also postulate a procedure ('lexicon optimization') by which learners work their way up from the surface form to the underlying representations, making use of the same set of constraints and constraint rankings that are necessary for the selection of the winning candidates in the language. Notwithstanding these advances, opacity seemingly remains an irreducible empirical fact of the phonology of languages, and is motivating important modifications in the theoretical apparatus of Optimality Theory, in an attempt to empower it to overcome the challenge.

Various types of experiments aimed at establishing the psychological reality of the various formal constructs have been carried out through the years, often with mixed results, although clearly questioning the more abstract analyses. More recently, access to actual neural reality is being attempted through neuroimaging techniques that register activity in areas of the cerebrum. The main results obtained reveal that productive regular alternations (e.g. regular plurals of the 'book' to 'books' kind examined above) are associated with different

brain potentials and involve different brain areas than do irregular alternations (e.g. plurals like 'children' or 'mice'). The former, thus, parallel the neurological correlates of syntactic processing, while the latter are reminiscent of those associated with the handling of single words. A reasonable inference to draw from these equivalences is that productive regular processes are neurologically real, whereas irregular alternants are stored piecemeal.

Further Reading

Anderson S (1985) *Phonology in the Twentieth Century*. Chicago, IL: University of Chicago Press.

Archangeli D and Langendoen T (1997) *Optimality Theory: An Overview*. Oxford, UK: Blackwell. [Chapters 1, 2, 3, 4.]

Goldsmith J (ed.) (1995) *The Handbook of Phonological Theory*. Oxford, UK: Blackwell.

Gussenhoven C and Jacobs H (1999) *Understanding Phonology*. London: Arnold.

Halle M (1991) Phonological features. In: Bright W (ed.) *International Encyclopedia of Linguistics*, pp. 207–212. Oxford, UK: Oxford University Press.

Kager R (1999) *Optimality Theory: A Textbook*. Cambridge, UK: Cambridge University Press.

Kenstowicz M (1994) *Phonology in Generative Grammar*. Oxford, UK: Blackwell.

Ohala J (1995) Experimental phonology. In: Goldsmith J (ed.) *The Handbook of Phonological Theory*, pp. 713–722. Oxford, UK: Blackwell.

Pinker S (1999) *Words and Rules*. London: Weidenfeld & Nicolson.

Roca I (1994) *Generative Phonology*. London: Routledge.

Roca I and Johnson W (1999) *A Course in Phonology*. Oxford, UK: Blackwell.

Roca I and Johnson W (1999) *A Workbook in Phonology*. Oxford, UK: Blackwell.

Smith NV (1999) *Chomsky*. Cambridge, UK: Cambridge University Press.

Phrase Structure and X-bar Theory

Intermediate article

Geoffrey K Pullum, University of California, Santa Cruz, California, USA

CONTENTS

The phrase structure of an expression is its structure in terms of the subexpressions of which it is composed and the categories to which they belong. X-bar theory is a family of related theories of phrase structure.

INTRODUCTION

Most modern syntactic theories for natural languages are based on three fundamental insights. The first is that expressions are classified into a restricted range of types, traditionally called parts of speech and referred to here as syntactic categories: *enjoyed* is a verb, *movie* is a noun, and so on. (The categories may also have subcategories: *enjoy* is a transitive verb, for example. In some languages words belonging to some of the categories also have forms that belong to morphosyntactic (inflectional) subcategories; for example, *enjoyed* is a verb inflected in the preterite tense.) Every syntactic theory recognizes at least some distinct categories for different kinds of word.

The second insight is that there are phrases. Although some expressions (e.g. words) are syntactically atomic (i.e. have no subparts), there are other expressions that have subparts that are also expressions. Thus *enjoyed the movie* is a verb phrase that is made up of *enjoyed* and *the movie*. It is not a necessary assumption that some expressions have parts that are expressions; dependency grammar essentially denies this, treating syntax as a matter of relations of dependency holding between words. But it is fundamental to all American structural analysis, and to the transformational syntax that succeeded it, that expressions have this kind of structure, which is known as 'constituent structure' or 'phrase structure'.

The third insight is that the subparts of an expression have specific functions within the expressions they make up. For example, the two parts of the expression *enjoyed the movie* have different functions: *enjoyed* is the head of the expression *enjoyed the movie*, and *the movie* is a complement (specifically, a direct object) of the expression. Neither part could serve the function that the other serves (*the movie* could not be the head of a verb phrase, for example). In some theories of syntax such functions (or grammatical relations) are very important; in relational grammar they are the basis for nearly all the content of the theory, and in dependency grammar they completely replace phrase structure. Within transformational grammar since 1970 the function 'head' has become particularly important, and the notion of 'X-bar theory' described below is founded on it.

SYNTACTIC CATEGORIES

Traditionally there are eight to ten word-level categories in English. The most significant, seldom absent from any English grammar, are listed in Table 1.

Table 1. Major syntactic categories

Symbol	*Category name*	*English examples*
N	Noun	*tree*, *cat*, *spoon*, *absence*, *sugar*, *Michael*, *Japan*, *April*
V	Verb	*weep*, *eat*, *feel*, *escape*, *greet*, *give*, *furnish*, *be*, *have*, *do*
Adj	Adjective	*good*, *long*, *nice*, *heavy*, *wide*, *large*, *predatory*, *continual*
Adv	Adverb	*soon*, *quite*, *so*, *too*, *nicely*, *heavily*, *beautifully*, *continually*
P	Preposition	*at*, *by*, *from*, *of*, *in*, *out*, *into*, *between*, *throughout*, *despite*

N is universally the category containing the most basic words for naming persons, places, things, substances, natural kinds, and abstract notions referred to as if they were things or substances (*idea*, *knowledge*). It is an open category: new nouns are added constantly as names are invented or borrowed to name new entities and kinds of entity. V, the category of verbs, traditionally a class of labels for actions, events, and occasionally states, is also open to a very considerable extent: new verbs are added to the language frequently (often by conversion of nouns). The categories Adj and Adv are not always distinct from one another in all theories, and indeed, are not always present in every grammar (there are a few languages that appear to do without adjectives, and it is not entirely clear that adverbs can be identified in all of those languages either). In languages that have both, adverbs are often transparently formed from adjectives by derivation (this is very largely the case in English, where most adverbs are formed by suffixing *-ly* to adjective roots). Both kinds of word are often coined and borrowed, though perhaps not as frequently as new verbs.

The category P contrasts with the other categories in that it changes membership relatively little, though new members have been added frequently during the history of English (the chief sources are recategorization from verbs and adjectives, and occasional borrowing from languages like Latin and French). It also contains a number of items that have very specific grammatical functions and in many of their occurrences have little lexical meaning. P is listed here as a major category, however, because there can be little doubt that P can be the head of full internally complex phrases (Jackendoff, 1973; Huddleston and Pullum, 2002, Ch. 7).

In addition to these major categories there are minor ones on which there has been more divergence in nomenclature. Within earlier generative grammar many *ad hoc* categories were employed: 'T' for tense morphemes, 'M' for modal verbs, 'Aux' for a constituent containing various auxiliary verbal items, 'Q' for a class of determinatives known as quantifiers, 'Deg' for degree particles like *too*, and other minor categories that have been posited in recent generative grammar (these are discussed further below under 'functional categories').

For the major categories, it is accepted by most grammarians that there are corresponding phrase types. The standardly accepted ones are listed in Table 2. (See below for discussion of the current view that a phrase like *the tree* is not an NP, but has *the*, belonging to the category D of determinatives, as its head.)

GENERALIZATIONS ACROSS CATEGORIES: THE X-BAR THEORY

The correspondence between Table 1 and Table 2, and names like noun phrase (NP) and noun (N), might seem to suggest that noun phrases should be expected to contain nouns, verb phrases should be expected to contain verbs, and so on. It was noted by Lyons (1968, p. 330ff) that this does not follow from phrase structure rules saying things like 'VP → V NP' (a verb phrase may consist of a verb and a noun phrase), because 'VP → P NP' would have been just as simple a rule to write.

Another thing that is clear from inspection is that the five phrasal categories listed in Table 2 have a strikingly similar internal structure. The NP *his teacher's absence from class* contains an obligatory N (*absence*) optionally followed by a phrase of some category with an appropriate meaning (here, the PP *from class*) and preceded by some word or phrase adding specificatory detail (here, *his teacher's*). The obligatory item that gives the phrase its specific syntactic character is called the 'head' of the phrase; the phrase that may follow it if its meaning requires

Table 2. Phrasal categories

Symbol	*Category name*	*English examples*
NP	Noun phrase	*the tree*; *eight cats*; *this spoon*; *every trick I knew*; *her teacher's absence from class*; *many nice pictures of us*
VP	Verb phrase	*wept bitterly*; *eat lots of fresh vegetables*; *usually feel pretty dreadful*; *was able to see clearly yesterday*
AdjP	Adjective phrase	*very good*; *stronger than I had expected them to be*; *proud of the result*; *quite obviously better than the rest*
AdvP	Adverb phrase	*really soon*; *just then*; *so much more nicely than the others*; *quite separately from all the other species*
PP	Preposition phrase	*at home*; *by candlelight*; *right out over the edge*; *away from such temptations*; *because you have been so kind*

Table 3. Structural parallelism across phrasal categories

	Specifier	*Head*	*Complement*
NP	*her teacher's*	*absence*	*from class*
VP	*sometimes*	*acted*	*rather stupidly*
AdjP	*very obviously*	*better*	*than the rest*
AdvP	*quite*	*separately*	*from all the others*
PP	*right*	*out*	*of the picture*

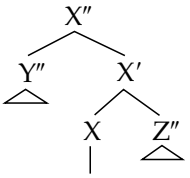
or permits it is called a 'complement'; and the item that precedes the head and adds specificatory detail is generally called the 'specifier'. The same structure is found in the case of all of the other phrasal categories, VP, AdjP, AdvP, and PP, as can be seen from the parallel examples in Table 3.

X-bar theory has its origins in attempts to express both of these generalizations – that the categories of phrases are related to the categories of their heads, and that phrases have similar internal structure across categories. The idea of defining successively more inclusive phrases of noun type as N, N^1, N^2, etc., is due to Harris (1951, Ch. 16). It was revived by Chomsky (1970, p. 210ff). (Chomsky writes N^k as N with *k* overbars, which gave rise to the term 'X-bar theory' and the term 'bar level'. Later works took to writing N^k as N with *k* primes: N, N′, N″, etc. Here we use either the prime notation or the numerical superscript notation, as appropriate.)

The idea was that phrase structure rules would be written in forms using variables X, Y, etc., over the lexical categories, and adding bar levels: $X^2 \rightarrow Y^2\ X^1$, $X^1 \rightarrow X\ Y^2$, etc., where X and Y are chosen from among N, V, P, etc. Chomsky employed this idea in an attempt to account for structural similarities between NPs and clauses which was not very successful: despite the semantic parallelism between a clause like *Microsoft dominates the software industry* and its nominalized correspondent, *Microsoft's domination of the software industry*, the internal structures of NPs and clauses differ in many ways (see McCawley, 1975, for criticism). The idea is much more successful in bringing out the similarities seen in Table 3.

The most thorough attempt to work out the implications of X-bar theory was that of Jackendoff (1977), which diverged considerably from Chomsky's initial presentation. Jackendoff allowed no nonhead-dependent grammatical items at all; he proposed that every word of the language, grammatically restricted in function or not, belonged to a category that projected a full phrase. Thus *the captain* for Jackendoff had not one projection (of the noun *captain*) but two: the first constituent was a nonhead Article Phrase (for him, Art^3), and the second was the phrasal head of NP (i.e. of an N^3), labeled N^2. For Jackendoff, clauses were V^3, their heads were V^2, the head of a V^2 was a V^1 (the traditional verb phrase), and the head of a V^1 was V^0(=V). (A similar proposal about clauses with 2 as the maximum bar level was proposed in Gazdar *et al*., 1985.)

Assuming 2 rather than Jackendoff's 3 as the maximum bar level, the structures of the phrases in Table 3 can now all be seen as instantiating the general pattern seen in the following tree, where Y″ is in what is known as 'specifier position' and Z″ is the complement of the head X:



Linear order is not relevant here; X-bar theory says nothing about it. In most current generative analyses the specifier position is the position for external arguments such as subjects. The 'VP-internal subject hypothesis' holds that the VP category contains subjects as well as objects at the pretransformational level: in the diagram above, X would be V, Y″ would be the subject position, and Z″ would be the direct object position.

Six principles of X-bar theory can be identified in the literature. Not all authors respect all of them, but all six are widely accepted as default assumptions, and are sometimes made explicit. Kornai and Pullum (1990) state the principles in fairly precise terms. They may be summarized informally as follows:

1. Lexicality: the only phrasal categories are those constructed from categories to which words belong.
2. Succession: each phrase has a unique constituent called its 'head' that has a bar level one lower than the phrase itself.
3. Uniformity: for every category X of bar level 0 there is a corresponding one-bar category (X′) and a two-bar category (X″).
4. Centrality: the root node label is X″ for some X. (Usually this is assumed to be C, so the root node label is C″ = CP.)
5. Maximality: nonhead constituents are X″ for some X.
6. Optionality: nonhead constituents are optional.

Lexicality expresses the fundamental principle that all phrasal categories are projections of some word class: there are no phrases that are not founded on categories to which words are assigned. Succession implements Zellig Harris's idea of labeling phrases

using numbers to indicate increasing levels of inclusiveness, i.e. increasing remoteness from the head. Uniformity claims that phrases have a maximum bar level that is the same for every lexical category (the adherence to this principle even in the face of little supporting argument is very clear in Jackendoff, 1977). Centrality guarantees that the label of root nodes (the category corresponding to 'sentence') is of maximum bar level, and maximality requires the same of all nonhead daughter categories.

The last principle, optionality, states that no nonhead is a strictly obligatory constituent. The plausibility of this condition can be appreciated by examining the results of omitting constituents from the examples in Table 3. In every case, either specifier, or complement, or both can be omitted, and the result is grammatical (while omission of the head does not lead to ungrammatical phrases of the relevant category in any case, as succession predicts). None the less, optionality is the principle that syntacticians have shown the least inclination to respect in descriptive practice. A key problem is that both constituents are obligatory in a simple clause like *You lied*, yet if heads are unique and obligatory, one of the two words has to be a nonhead. There are generally assumed to be additional principles, external to X-bar theory (for example, a principle stipulating that clauses must have subjects) that guarantee the necessary obligatoriness, in effect vacating the optionality claim.

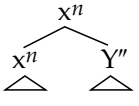
Various proposed revisions of X-bar theory have attempted to get rid of bar levels as such, or at least to eliminate bar levels intermediate between 0 and the maximum m. In a sense this is trivial: it was proved by Kornai (see Kornai and Pullum, 1990, for a proof) that for every language generated by an X-bar grammar with maximum bar level k, for any $k \geq 2$, there is also an X-bar grammar with maximum bar level 1. But Chomsky (1995) advances a theory of 'bare phrase structure' that sets out to revise the theory of phrase structure to get rid of bar levels completely, and even eliminates categories, nonterminal nodes being labeled with structured objects built up out of the lexical items that label terminal nodes. For an attempted clarification of the content of this theory using techniques from logic, see Kracht (1997).

HEADS, COMPLEMENTS, AND ADJUNCTS

The notion 'head of a phrase' is reconstructed by X-bar theory via the principle of succession: the head of an X^n is its only daughter labeled X^{n-1} (and also, according to optionality, its only obligatory daughter). A head labeled X (with bar level 0, therefore necessarily the head daughter of a node labeled X′ with bar level 1) is a lexical head. Lexical heads are divided up into subcategories according to which constituents must or may accompany them within their immediately containing X′. Verbs are thus spoken of as subcategorized into (i) intransitive verbs (which require nothing else in the V′), (ii) transitive verbs (which require a noun phrase in the V′), and so on – there are perhaps a dozen or more other types of verb. (The notion of subcategorization in this sense, sometimes known as 'strict subcategorization', is introduced in Chomsky, 1965.)

The phrases that occur in X′ are determined by the subcategory of the head X and are known as 'complements'. (Not all of the literature uses the term 'complement' in the same way; Huddleston and Pullum (2002) draw a distinction between internal complements such as objects and external complements such as the subject. In X-bar theory descriptions it is standard to call the subject an 'external argument', but not a complement; only nonhead constituents of X′ are called complements.) Some complements appear to be syntactically obligatory. For example, the verb *keep* must have a complement (either an NP, as in *I decided to keep the money*, or a clause, as in *They kept laughing at me*); likewise the preposition *at* must have a N″ complement. Such cases provide part of the evidence that the optionality condition cannot be interpreted as absolute: nonhead constituents are optional unless the subcategorization restriction of the head of their immediately containing X′ dictates their obligatory appearance.

Some constituents are not selected according to the subcategorization restrictions of lexical heads but instead function as optional modifiers of X′ or X″ phrases. These constituents are known as 'adjuncts'. Under strict versions of X-bar theory they must be daughters of X″ and thus sisters to X′, because phrases are not permitted to occur anywhere else. But many syntacticians assume a weaker X-bar theory in which adjunction (often called 'Chomsky-adjunction' after the tacit use made of it in early works of Chomsky's) is permitted; that is, configurations like the following (where the n stands for some specific bar level) are allowed:



(As noted above, linear order is not relevant; the Y″ could precede or follow the X^n.) Here the head of an X^n is another X^n, in contravention of succession. Jackendoff (1977) maintains a strict X-bar theory with succession and thus permits no configurations in which an X^n immediately dominates an X^n, but this is not compatible with a 'stacked' structure for adjuncts: e.g. the bracketed phrases of category P″ in *We met* [*on Tuesday*] [*in the board room*] [*in the presence of our lawyers*] or the bracketed relative clauses in *Anyone* [*who did not sign up*] [*who still wants to go on the trip*] *should see me*. Adjunction structures are commonly assumed by syntacticians to be permissible, and this means that they weaken the succession condition.

FUNCTIONAL CATEGORIES

In recent generative grammar (increasingly since the mid-1980s), full X-bar projections have been assumed for many categories beyond N, V, A (the category embracing the traditional Adj and Adv), and P.

The idea that 'complementizer' (C) is a zero-bar-level category that is the head of a subordinate clause (CP = C″) actually originates with Langendoen (1975, p. 540) but gained no currency until it was reinvented, apparently independently, in the 1980s (see e.g. Chomsky, 1986). After that it was rapidly adopted by almost all transformational grammarians.

The idea that the definite and indefinite articles of English belong to a category that is the basis of an X-bar projection is due to Jackendoff (1977), but for Jackendoff the 'article phrases' (ArtP) that the idea gives rise to were nonhead adjuncts in the structure of noun phrases (thus in *only these books* the phrase *only these* would be an ArtP). The idea re-emerged in a different form when Abney (1987) suggested that determinatives (D) were actually the heads of what had until then been called 'noun phrases', so that the phrase *those boxes of books* would be a DP with *those* (D) as its lexical head, *boxes of books* being a complement to D. The label NP now took on a new meaning as the label for this sort of complement, of which *boxes* (N) would be the head, *of books* (PP) being a complement to N.

Both CP and DP introduce problems relating to selection. Regarding CP, when verbs (or other lexical heads) select particular clause types, they select properties reflected in the internal structure of the clause (the IP) rather than in its complementizer: *know* takes a *that*-clause in which the clausal part is finite and indicative (*know that this was not done*), while *demand* takes a *that*-clause in which the clausal part is subjunctive (*demand that this not be done*). Instead of verbs selecting properties of the head of their complement constituent, we find here that a V is selecting the T head of a TP which is complement to the C head of its complement CP: selection is treating C as if it were not a head. This observation is taken by Huddleston and Pullum (2002) to justify rejection of the view that complementizers ('subordinators') are heads of subordinate clauses. Transformationalists generally take it as indicative of some kind of special syntactic relationship between C and its closest subordinate I or T.

And with regard to the DP analysis, there is no case of a lexical head syntactically or semantically selecting a complement with a particular type of D, such as a verb that demands a DP with definite determiner. The DP hypothesis predicts them, but none has ever been identified. Verbs that are only appropriate with certain types of complement PPs select the head preposition (*approve of it* but not **approve from it*; *rely on it* but not **rely to it*); but instead of selecting D in phrases like *the bishop* they select appropriate nouns: *frighten* needs a direct object with a noun denoting an animate entity, *merge* needs a direct object denoting a plurality. This is just what would be expected if (as in more traditional views) the noun were the head of a direct object phrase rather than the determinative. (Again, Huddleston and Pullum, 2002, accept this consequence.)

A development of the late 1980s was the hypothesis that clauses actually had as their heads instances of the tense/agreement inflectional element, generally called 'Infl', or simply 'I'. Clauses would therefore be labeled IP (I″, assuming a maximal bar level of 2). Pollock (1989) proposed revising this by breaking tense and agreement apart and treating each as an element founding a separate projection. He also treated the negation element as founding a projection. Given three bar levels (0, 1, and 2), this introduced nine new category labels: T, T′, T″, Agr, Agr′, etc., two of them present in every finite clause. In subsequent work many additional functional categories and functional projections have been posited.

STRUCTURAL RELATIONS DEFINED ON TREES

Constituent structure trees can be formalized as graphs. A graph is a set of elements (linguists call them 'nodes') on which is defined a relation (linguists call it 'dominance'). Diagrammatic representations indicate nodes by points and represent two

nodes as being in the edge relation by drawing a line between them. Constituent structure trees meet several conditions. First, the nodes are labeled with category names. Second, the edges are directed: a line from a to b (representing the fact that a dominates b) is not the same as a line from b to a (representing the reverse situation). Third, another basic relation is defined on certain node pairs, a strict partial ordering left-to-right relation called 'linear precedence': each pair of nodes is related either by dominance (in one direction or the other) or precedence (in one direction or the other), and never by both. Fourth, when a node precedes (or follows) another it is also required to precede (or follow) all of the nodes dominated thereby (i.e. precedence is recursively inherited by daughters). And fifth, there is a node called the 'root' that, uniquely, dominates every node in the tree (the single root condition). It is customary to show trees as diagrams with the root at the top and precedence represented by the left-to-right direction across the page, as is done above.

Additional derivative relations between nodes can be defined on trees in terms of dominance and/or precedence. Subjacency, for example, is definable in terms of these two relations and certain references to categories. Among the other relations that are commonly referred to in the syntactic literature are these:

- Linearly adjacent: x and y are linearly adjacent if and only if x precedes y and there is no intervening node i such that x precedes i and i precedes y.
- Immediately dominates (is the mother of): x immediately dominates y if and only if x dominates y and there is no intervening node i such that x dominates i and i dominates y.
- Daughter of: is the inverse of 'immediately dominates'.
- Sister of: x is the sister of y if and only if there is a node m that immediately dominates both x and y.
- Commands: x commands y if and only if every clausal node that dominates x also dominates y.
- C-command: x c-commands y if and only if every node with two or more daughters that dominates x also dominates y.
- M-command: x m-commands y if and only if every node labeled with a maximal bar-level category that dominates x also dominates y.

The definitions of the last three relations are clearly very similar. Abstracting their similarities, Barker and Pullum (1990) defined a more general notion of 'command relation': given any property of nodes ϕ, a node x is said to ϕ-command a node y if and only if every node with property ϕ that dominates x also dominates y. Barker and Pullum show that there is a minimal (smallest) command relation, which holds between a pair of nodes x and y if (and only if) any command relation holds between x and y. This relation is IDC-command, and it is generated by letting ϕ be a trivial property possessed by every node. That is, x IDC-commands y if and only if every node that dominates x also dominates y.

Kayne (1994) proposed that asymmetric c-command should be regarded as the fundamental relation in syntax, precedence being parasitic on it in the sense of being predictable from it (and proposed structures should be revised where this is not the case). In fact it appears that Kayne's intent is best captured if the relation appealed to is IDC-command rather than c-command under its classical definition.

CONCLUSION

Phrase structure (or constituent structure) analysis involves analyzing expressions into subparts that are themselves expressions. Subparts are of specific categories, and have specific functions such as 'head of'. X-bar theory is based on the idea that every phrase has a head subconstituent, and phrases can only be founded on specific lexical categories, i.e. categories that specific words belong to. Lexical categories have bar level 0, and project phrases of higher bar levels up to some specified maximum. A phrase X^m with the maximum bar level m is known as a maximal projection, and corresponds to a maximally inclusive phrase of the type founded on X. Most current theories of grammar maintain some version of X-bar theory (Chomsky's 'bare phrase structure' replacement for it cannot be said to have become standard). X-bar theory comes in differing versions (what is most typical is to set the maximum bar level at 2 and have X ranging over the whole vocabulary of lexical and functional categories), but the general framework is constant across a wide range of transformational-generative syntactic theories.

References

Abney SP (1987) *The Noun Phrase in its Sentential Aspect*. Doctoral dissertation, MIT.

Barker C and Pullum GK (1990) A theory of command relations. *Linguistics and Philosophy* **13**: 1–34.

Chomsky N (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky N (1970) Remarks on nominalization. In: Jacobs R and Rosenbaum PS (eds) *Readings in Transformational Grammar*, pp. 184–221. Waltham, MA: Ginn & Co.

Chomsky N (1986) *Barriers*. Cambridge, MA: MIT Press.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.
Gazdar G, Klein E, Pullum GK and Sag IA (1985) *Generalized Phrase Structure Grammar*. Oxford: Basil Blackwell and Cambridge, MA: Harvard University Press.
Harris Z (1951) *Methods in Structural Linguistics*. Chicago: University of Chicago Press.
Huddleston R and Pullum GK (2002) *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.
Jackendoff RS (1973) The base rules for prepositional phrases. In: Anderson SR and Kiparsky P (eds) *A Festschrift for Morris Halle*, pp. 345–356. New York: Holt Rinehart and Winston.
Jackendoff RS (1977) *$\bar{X}$ Syntax*. Cambridge, MA: MIT Press.
Kayne RS (1994) *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.
Kornai A and Pullum GK (1990) The X-bar theory of phrase structure. *Language* **66**: 24–50.
Kracht M (1997) On reducing principles to rules. In: Blackburn P and de Rijke M (eds) *Specifying Syntactic Structures*, pp. 43–73. Stanford: CSLI Publications.
Langendoen DT (1975) Finite-state parsing of phrase-structure languages and the status of readjustment rules in grammar. *Linguistic Inquiry* **6**: 533–554.
Lyons J (1968) *Introduction to Theoretical Linguistics*. Cambridge, UK: Cambridge University Press.
McCawley JD (1975) Review article on Noam A. Chomsky, *Studies on Semantics in Generative Grammar. Studies in English Linguistics* **3**: 209–311. Reprinted in McCawley JD (1982) *Thirty Million Theories of Grammar*, pp. 10–127. Chicago, IL: University of Chicago Press.
Pollock JY (1989) Verb movement, universal grammar and the structure of IP. *Linguistic Inquiry* **20**: 365–424.

Further Reading

Corbett GG, Fraser NM and McGlashan S (eds) (1993) *Heads in Syntactic Theory*. Cambridge, UK: Cambridge University Press.
Culicover PW (1997) *Principles and Parameters*. Oxford: Oxford University Press.
Emonds JE (1976) *A Transformational Approach to English Syntax*. New York: Academic Press.
Hellan L (1977) *$\bar{X}$* syntax, categorial syntax, and logical form. In: Fretheim T and Hellan L (eds) *Papers from the Trondheim Syntax Symposium*, pp. 83–135. Trondheim, Norway: University of Trondheim.
Webelhuth G (1995) X-bar theory and Case theory. In: Webelhuth G (ed.) *Government and Binding Theory and the Minimalist Program*, pp. 15–95. Cambridge, MA: Basil Blackwell.

Phrase Structure Grammar, Head-driven

Intermediate article

Robert D Levine, Ohio State University, Columbus, Ohio, USA

CONTENTS

Head-driven phrase structure grammar is a monostratal theory of natural language grammar, based on richly specified lexical descriptions which combine according to a small set of abstract combinatory principles stated as formulae in a constraint logic regulating, for the most part, the satisfaction of valence and other properties of syntactic heads. These constraints, applying locally, determine the flow of information, encoded as feature specifications, through arbitrarily complex syntactic representations, and capture all syntactic dependencies – both local and non-local – in elegant and compact form requiring no derivational apparatus.

BASIC CONCEPTS

Generative theories of linguistic structure can be categorized in various ways, most fundamentally, into theories that license sentences by constructing a set of structures and then deriving further structures from these, until some point where no further structures are derivable; and theories that admit only a single structural object corresponding to each sentence (or a set of such objects, corresponding to syntactic ambiguity), where admission entails simultaneous satisfaction of a set of applicable constraints. The first approach, generally referred to as transformational grammar, typically uses the successive stages in the derivation of the sentence to establish appropriate linkages in information content among parts of the representation; while the second, often called monostratal, establishes such linkage by directly constraining information distribution, typically encoded as the values assigned to certain attributes, or 'grammatical features'.

Grammatical frameworks may also be divided into those whose admissible objects are characterized by phrase structure, and those in which no persistent configurational properties of representations are posited. Categorial grammars, for example, specify functor–argument possibilities for individual lexical items: such possibilities determine whether a given sequence of words can combine to give rise to an object corresponding to the type associated with a clause, but the order of combination does not impose structural configurations on the the string or its substrings which are subsequently available for stating rules or generalizations.

A third important division distinguishes theories that require a one-to-one match between syntactic form and propositional structure from those that do not. For example, in the sentence *Robin seemed to understand Dana's arguments*, it is evident that the individual named Robin has a semantic relation to the predicate denoted by the linguistic form *understand Dana's argument*, but has no semantic relation to the verb *seemed*. That this is generally true for the subject of such verbs (called 'raising' verbs) is made clear by a variety of syntactic diagnostics (preservation of idiomaticity, preservation of truth under passivization, and so on). Yet in the syntax of this sentence, *Robin* is the subject of *seemed*, with no evident local structural relationship to *understand Dana's arguments*. There is thus a mismatch between syntactic form and semantic interpretation in such examples. On one standard approach to this problem, the mismatch is the result of a derivational history in which a structure with *Robin* appearing in the subject position of *understand Dana's arguments* in some early structural representation is displaced to the subject position of *seemed*. On another approach, *Robin* is syntactically always and only a subject of *seemed*, but is linked through various constraints to the semantic content of *understand Dana's arguments*, without any comparable linkage to the semantics of *seemed*. The former approach will naturally be restricted to derivational theories, but not all derivational theories follow this approach (e.g. Brame, 1976; Culicover and Wilkins, 1984). On the other hand, nonderivational theories have no choice: they must take as

basic the syntax–semantic mismatch exemplified here and motivate the discrepancy in their own formal structure – via meaning postulates, lexical entailments, representational properties, or some other means.

Against this background, phrase structure grammar (PSG) theories can be concisely characterized as a family of theories which assume only a single level of representation, and take this level to consist in a set of hierarchically structured objects, in which syntactic and semantic properties need not mirror each other in any respect.

Within the family of PSG theories, there is considerable variety. In this article we will illustrate the possibilities of the overall PSG framework with reference to one particular version, head-driven phrase structure grammar (HPSG).

HPSG

Ontological Assumptions

HPSG was developed by Carl Pollard and Ivan Sag (Pollard and Sag, 1987, 1994), initially as a refinement and extension of generalized phrase structure grammar (Gazdar, 1981; Gazdar *et al*., 1985). It belongs to a family of phrase-structure-theoretic approaches in which a rich set of lexical specifications, coupled with a few very general combinatorial constraints and restrictions on information-sharing, interact monotonically to give rise to sets of complex objects called feature structures, which model the properties of linguistic signs (though certain recent developments of HPSG have abandoned strict monotonicity by allowing defeasible default characterizations of feature-sharing principle to interact with 'hard' lexical entries or type declarations). Each feature structure admitted by an HPS grammar for some language licenses an expression in that language. Considerable effort has been made to provide an explicit model theory for HPS grammars, and to formulate the constraint systems which HPS grammars themselves consist of as statements in certain formal logics. Such logics are comparable to familiar systems such as the first-order predicate calculus, with the crucial difference that constraint satisfaction takes the place of model-theoretic satisfaction of, for example, a set of conventional formulae by some interpretation given for the latter. This approach, which has its origin in the computational linguistics of the mid-1980s, greatly expedites implementation of HPS grammars, but from the theoretical perspective its main significance is that the denotation of particular analyses and claims stated in the framework is always mathematically explicit: a necessary but often neglected condition for generativity.

An HPS grammar is a description, i.e. a set of formulae in some feature logic (see, e.g. Richter *et al*. (1999) for an example of such a logic that is becoming widely employed in HPSG), where only feature structures that satisfy the formulae of the grammar are taken to be complete representations of the grammatical properties of some linguistic expression. Feature structures themselves are sorted unary partial subalgebras in which feature names are functions from the set of nodes to the set of nodes, with each node assigned a 'type' or 'sort' label. Each node corresponds to some object with certain properties, and the node a given feature maps this object to corresponds to the value of that object for the property named by the feature.

Consider, for example, auxiliary verbs in English. Such verbs select a certain class of subjects and appear either in canonically ordered sentences (*Robin has left*) or inverted sentences (*Has Robin left?*). In HPSG, such a linguistic expression will be modeled by a feature structure including a specification for the feature `CAT`, providing relevant syntactic information, including a specification of the `HEAD` properties of that expression – those which are invariably shared between mother and head daughter. The feature `HEAD` is thus taken to be a function which maps nodes labeled by the sort `cat` ('category') to nodes of sort `noun`, `verb` and so on; for verbs, this latter node itself is mapped by a function `VFORM` to a node labeled by one of a set of sorts (`fin` ('finite'), `inf` ('infinitive'), `bse` ('base'), etc.) by a function `AUX` to one of the sorts `plus` or `minus`, and so on. In the same way, the feature `SUBJ` is a function from nodes of sort `cat` to nodes whose type corresponds to a (singleton) list identifying the kind of subject the verb can combine with, and the feature `COMPS` is a function from `cat` nodes to a different node of the same sort, identifying the morphosyntactic properties of the complements that the verb selects.

More generally, various versions of HPSG model theory essentially converge on the characterization of the model universe for an HPS grammar along the following lines:

1. Each grammar specifies a 'signature': a set of feature names, a set of sort names, and an assignment function which pairs with each feature name a mapping from the set of sort names to the power set of sort names. The signature in effect defines what kinds of objects there are, and what kinds can be mapped to (i.e. specified for) what other kinds of objects under the operation denoted by some feature name.

2. An 'interpretation' for an HPS grammar is a set of partial algebras, each of which specifies a set of sorted nodes and a set of unary operations, corresponding to the feature names. Each unary operator takes some subset of nodes as its domain and some subset of nodes as its range. For all mappings from subset to subset, the sort labels of the domain and range elements must adhere to the constraints imposed by the assignment function.
3. Given the objects and functions so far defined, it is possible to construct a formal notion of pathway – sequences of nodes – in terms of successive applications of appropriate feature functions. Thus, the pathway `CAT|HEAD|VFORM` specifies a function which maps an input node of a certain sort to a node of a sort such as `fin`. Feature logics can be defined on such pathways, and an HPS grammar specifies a number of constraints, stated in terms of such pathways. Constraints are descriptions that are satisfied for just those sets of partial algebras in which all pathways specify the values that these descriptions require them to. The constraints, as noted earlier, in effect play the role of meaning postulates in traditional model theories of natural language semantics: they restrict the set of admissible interpretations of the grammar – in this case, sets of partial algebras – to just those which reflect the property of the natural language modeled by the grammar.

These components of the HPSG model theory express standard HPSG assumptions: respectively, that the object that is the value of any given feature belongs to a particular sort, that only objects of certain sorts can be the value of any given feature, and that objects of different sorts are specified for different sets of features. The interaction of these assumptions constitutes an overall constraint on the possible set of feature structures that the grammar admits, corresponding to just the set of linguistic expressions taken to be well-formed by speakers.

It should be noted that this algebraic characterization of feature structures is similar to the formalization of category specifications of generalized phrase structure grammar (e.g. Gazdar *et al*., 1985, 1988). Feature structures, however, are typically represented not as partial algebras, but graphically – specifically, as pointed directed graphs in which nodes, each annotated with a sort label, are linked by arcs labeled with the names of grammatical features, where for some linguistic expression *C*, each feature *f* is taken to correspond to a property of *C* whose value is identified as the subgraph originating at the node on which the arc labeled by *f* terminates. A graph corresponding to the example discussed above is given in Figure 1. While such graphical representations seem to be more accessible than the algebraic structures which essentially define feature structures, they are still rather cumbersome to manipulate and far from perspicuous. Furthermore, under the HPSG assumption that feature structures are complete objects – that is, total specifications of the grammatical properties of linguistic expressions – the object in Figure 1 is not a feature structure, but rather a

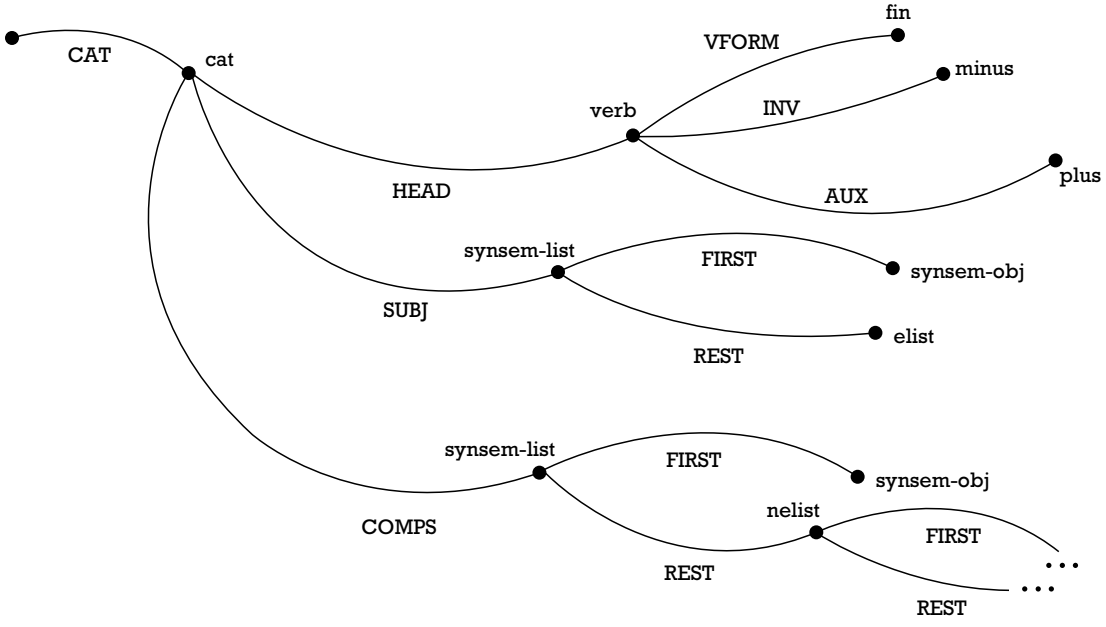


Figure 1.

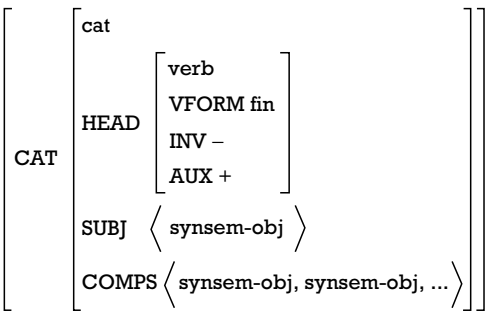


Figure 2.

small substructure; a full representation would involve an enormous number of nodes and arcs. Hence, a different and far more perspicuous format is used to describe feature structures. The feature substructure in Figure 1 satisfies the description given in Figure 2. Objects such as that in Figure 2 are called 'attribute–value matrices', and are used both to provide descriptions of feature structures and to state constraints on them.

It often proves convenient to assume that sorts themselves are organized hierarchically, into networks of subsort–supersort relations. Every object that the grammar licenses must satisfy the signature declarations on all the sorts of which that object is a subsort, in addition to whatever other constraints apply to it beyond those in the signature (e.g. the 'head feature principle' discussed below). Schematically, one can represent such relationships among sorts as a branching hierarchy. Each subsort is subject to a particular set of restrictions, imposed on it outside the signature properties themselves, in the constraint logic constituting the description language of the grammar. Elements lower (i.e. more specific) in the sort hierarchy inherit all restrictions imposed on the higher (i.e. more general) sorts dominating them. The objects in the HPSG model theory, the feature structures themselves, only contain maximally specific sort labels; but descriptions of (classes of) feature structures, including the constraints of the grammar, may refer to sorts of any level of inclusiveness.

For example, the very top of the sort hierarchy is, by convention, taken to be `object`, one of whose immediate subsorts is `sign`. Signs contain complete specifications for properties of classes of linguistic expressions. The signature will identify one of the immediate subsorts of `sign` as `word` and the other as `phrase`; a further constraint will restrict the feature 'daughters' (DTRS) to specification only for objects of the latter sort. The value of DTRS in turn must be of sort 'constituent structure' (`cons-struc`), which has two subsorts: `coordinate-structure` and `headed-structure` (`headed-struc`). The latter is subject to the restriction that it contain a specification for a feature 'HEAD DAUGHTER' (HEAD-DTR). The sort `headed-struc` has the further subsorts `head-subject-structure`, `head-adjunct-structure`, `head-filler-structure`, and so on. Each of these subsorts is of sort `headed-struc` and so necessarily contains a specification HEAD-DTR; in addition, each may be subject to idiosyncratic restrictions. All phrases, for example, of sort `head-filler-structure`, where a filler is linked to a gap site, require a specification FILLER-DTR which is related to the feature specifications on the HEAD-DTR category in a specific fashion allowing the filler and gap site to share relevant properties.

It is often suggested informally by researchers within HPSG that such type hierarchies are cognitively realistic, insofar as they are crucial to characterizing psycholinguistic processes such as language acquisition. The constraint inheritance property of the type hierarchy sketched above allows the lexicon itself to be relatively spartan, given a sufficiently rich hierarchy of sorts with associated constraints. This will be especially true if multiple lines of inheritance are permitted in the hierarchy. A given lexical item will have to bear only the feature information that is absolutely idiosyncratic; any information that can be predicted from the fact that it belongs to a certain class of lexical items can be built into a constraint on a sort to which those items belong, and if that class of items shares a certain property with a separate class of items, there will be a supersort under which the sorts of both classes appear, with the shared property identified as a constraint on the supersort. The sort hierarchy thus allows generalizations over the lexicon to be stated in a compact fashion, with fully specified lexical descriptions deducible in a straightforward monotonic fashion. It is not yet evident to what degree the type hierarchy captures significant properties of grammars that are otherwise unstatable. In other words, whether the type hierarchy has more than an abbreviatory status is still an open question. To answer this question, well-developed accounts of the interaction between formal grammar, on the one hand, and other linguistic phenomena, such as acquisition, processing, or language change, on the other, are needed.

The Architecture of Categories

The inventory of attributes or features in HPSG descriptions expressing properties of linguistic signs does not in itself determine the detailed structure of such signs. The latter is an empirical question, and since work in HPSG began, there has been continuous rethinking of the specific feature geometry required to capture grammatical properties of natural language in a sufficiently restrictive way. Currently, for example, there is general agreement among HPSG theorists that lexical signs require specification for phonology (PHON), syntax–semantics (SYNSEM), and morphology (MORPH), while phrasal signs require PHON, SYNSEM, and DTRS attributes. The general organization of phrasal signs is given in Figure 3. Such diagrams are to be understood as informal summaries of properties of the class of admitted feature structures. The value of SYNSEM is itself a set of specifications for 'local' (LOC) features, reflecting locally relevant properties of the sign, and for 'nonlocal' (NONLOC) features, encoding information that propagates over unbounded syntactic domains, such as information connecting fillers to gap sites, or information about *wh* formatives introducing information into relative or interrogative structures. LOC specifications identify syntactic properties of lexical or phrasal categories via the feature CAT, information relevant to logical aspects of interpretation via the feature CONT ('content'), and contextual information via the feature CONX. LOC is also specified for a feature QSTORE which plays a significant role in tracking the scope of quantifiers. For nominal categories, an INDEX feature is defined as part of the CONT specification, which (ignoring a very small class of exceptions) encodes reference to individuals or events. Values for CAT are particularly highly structured: this feature specifies valence information and also HEAD values, where the latter feature encodes properties of lexical items that are shared with all phrasal projections of those items, including part of speech, as well as other properties that depend on the part of speech of the item (case values for signs whose HEAD is of sort noun; auxiliary or non-auxiliary status for items of sort verb, and so on). Valence requirements are encoded in subspecifications of the feature VAL. The particular features that VAL is in turn specified for depend on the syntactic category of the head; thus, nominal lexical items select both complements (via a feature COMPS) and a specifier, identified by a feature SPR and manifest as a possessor, a determiner or nothing, while the VAL specification for verbs requires specification for a subject (encoded as a feature SUBJ) and COMPS, but not SPR.

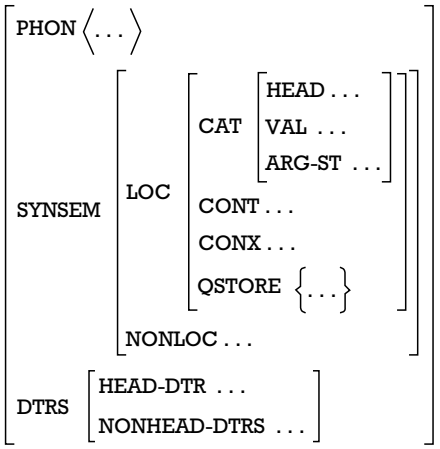


Figure 3.

In addition to the valence features SUBJ, COMPS, and SPR, HPSG posits a feature ARG-ST which typically takes the form of a list-append operation carried out on the valence lists. In the case of verbs, for example, the COMPS list is appended to the singleton SUBJ list, so that the ARG-ST list is identical to the COMPS list with one extra element corresponding to the subject added at the top of the list. The ordering of the elements on the ARG-ST list is the same as that on the valence list, and expresses the notion of 'relative obliqueness', discussed below. The ARG-ST list is not a valence list, but rather a record of the verb's argument structure; it plays a major role in the nonconfigurational HPSG theory of coreference, outlined below.

This organization of features expresses certain well-motivated assumptions about linguistic modularity. For example, the value of each of the valence features is required by appropriate sort declarations to be a list of synsem specifications, where the feature geometry of the framework packages syntactic and semantic information about categories, but not phonological or subconstituent information, within synsem objects. The result is the exclusion of access by a selecting head to information about either the phonological form or the syntactic subconstituency of its arguments.

Capturing Syntactic Dependencies

The critical task facing any grammar is capturing systematic linkages in form between (possibly quite distant) substructures of a larger syntactic structure. Were it not for the existence of such linkages,

or 'syntactic dependencies', natural language grammars would be far simpler than they actually are, since there would be little or no need to account for the sharing of often quite detailed information between different subcomponents of a sentence. Given the existence of such dependencies, a major goal of grammatical theory is the discovery of the optimal mechanism for expressing them.

The explicit specification of grammatical properties by means of highly structured descriptions, as sketched above, allows HPSG to capture syntactic dependencies economically. Agreement and case government are common examples of local dependencies, which typically hold between elements in a single clause, but other dependencies are more intricate, including several which have been taken as canonical demonstrations of the need for derivational relationships among syntactic objects. Such derivations typically serve to link the grammatical information shared among (possibly arbitrarily distant) structural positions. In HPSG, the linkage among the elements of the dependency is specified directly via the interaction of lexical properties with a small number of highly general declarative constraints, without the invocation of any mediating device such as movement.

Local dependencies

Local syntactic dependencies can in most cases be expressed by the correlation of properties of a selecting head with properties of a selected constituent, corresponding to some valence element. The simplest cases of this mechanism are syntactic selection and agreement. For example, the fact that the verb *give* must appear with an NP and a PP whose head is the preposition *to* is captured by identification of the verb's COMPS value as ⟨[1]NP, [2]PP[*to*]⟩, where [*to*] abbreviates [PFORM *to*], identifying the morphological form of the preposition via a feature PFORM, and where the angled brackets conventionally denote a list whose elements are ordered from left to right according to their relative obliqueness, corresponding to a traditional hierarchy of grammatical relations. Direct objects, for example, are the least oblique elements on a verb's COMPS list. At the same time, the CONT value of *give* specifies a particular relationship among an agent, a recipient and an object given. The HPSG approach to selection is illustrated in Figure 4.

Boxed alphanumeric symbols, called 'tags', are variables over nodes in the feature structures to which the descriptions containing those tags apply. Thus, when two features in a description appear with identical tags as their values, they have precisely the same node (and its associated subgraph) in the modeling object for those values. Sorts, where relevant, are indicated in lower case at the top of the description. The notation $X_{[n]}$ indicates that the INDEX value of X is n.

With these conventions, Figure 4 represents a partial lexical entry for *give*.

Two general principles of the grammar are assumed here. The 'valence principle' ensures that a phrase can be projected from a head just in case there is exactly one sister for that head corresponding to every synsem object on the head's COMPS list. The 'head feature principle' requires the HEAD specifications of a mother and its head daughter to be structure-shared (i.e. the arc from the mother category labeled HEAD and the arc from the head daughter category labeled HEAD terminate on the same node, in the graphic representation of admissible feature structures described above).

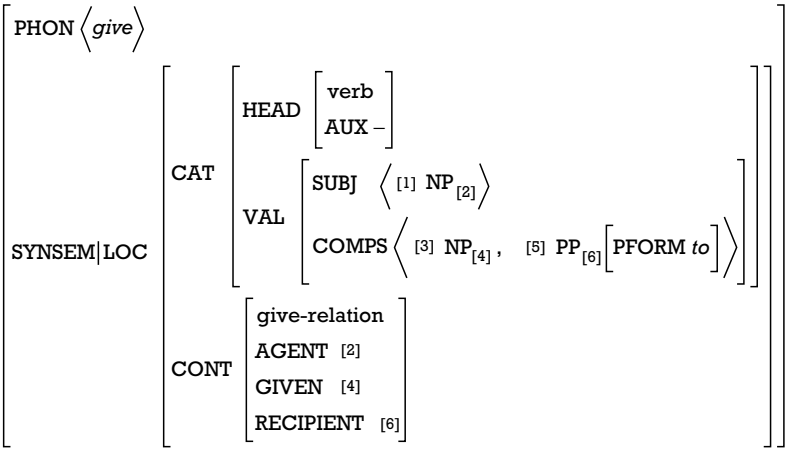


Figure 4.

The interaction of these principles with the lexical entry given in Figure 4 entails that the feature structures that will be admitted are just those in which the valence of the head is satisfied. Thus, if we represent the various DTRS feature values in terms of familiar branching tree structure, then the feature structure meeting the description in Figure 5(a) will be licensed, but those described by Figures 5(b), 5(c) and 5(d) will not be. All four descriptions in Figure 5 satisfy the head feature principle, but only Figure 5(a) the valence principle. Hence, only the VP depicted in Figure 5(a) is licensed by the grammar, yielding the relevant aspects of *give*'s subcategorization.

This same category selection mechanism can also enforce a correct match between the form of a verb and properties of its subject. Note, in the first place, that constraints on lexical types capture morphological generalizations; thus, it is unnecessary to separately stipulate that the third person singular form of *give* is *gives*, because a general constraint on the `third-sing` subsort of the `verb` sort automatically yields this form for a lexeme whose phonology is identified as [*g*ɪv*z*], and simultaneously imposes the specification [SUBJ ⟨NP[1][PERS 3, NUM sg]⟩] on all verbs of this subsort. It then follows from the valence principle that the only NPs that will be able to combine with VPs projected from *gives* and morphologically parallel verbs are third-person and singular.

Given the possibility of constraining the subject selection properties of classes of verbal heads with reference to some other grammatical property of those heads, the standard local dependencies of English can be captured simply along the following lines:

- Consider the consequence of positing a lexical relationship which guarantees that for every verb with a [SUBJ⟨ *S* ⟩, COMPS⟨ α_1, …, α_n ⟩] specification, there is a matching lexical entry with an identical phonology with the specification [SUBJ⟨NP$_{it}$⟩, COMPS⟨ α_1, …, α_n , S⟩], where `it` is a reserved INDEX specification uniquely associated with the dummy pronoun *it*. Nominals must in any case be subclassified to make expletives available for reference, so that, for example, ambient weather predicates such as *rain* take only *it* as subjects. But the lexical relationship just sketched will interact with the valence principle to ensure that for every clause projected from verbs such as *bother* or *astonish*, as in *That Robin would even consider spying for an unfriendly power astonished us* and *For Dana to act like that constantly would bother most people*, there is a matching clause with an *it* subject and an identical set of complement daughters, along with an additional clausal daughter (e.g. *It astonished us that Robin would even consider spying for an unfriendly power, It would bother most people for Dana to act like that constantly*). Thus, extraposition is accounted for without having to posit a derivational origin for this construction.
- The active–passive correspondence can be treated as a lexical relationship in which, either by a lexical rule or by a type constraint, an active verb is related to its passive counterpart in such a way that the COMPS list of the latter is that of the former minus the highest-ranking complement specification, where this 'lost' element reappears as the passive's SUBJ specification.
- The existential *there* construction can be captured by means of a partial lexical entry for the lexeme *be*, as in Figure 6, where the subscript `there` notates a separate `index` subsort like `it`, and both are distinct from the subsort `ref` ('referential'). Under this lexical analysis, *There are two lions in the closet* will denote exactly what the predicate *in the closet* applied to *two lions* denotes, while the verb morphology is (indirectly) correlated with the person and number specifications of the postcopular NP.
- Properties of raising constructions (e.g. *Robin seems to be having a difficult time*) can be straightforwardly captured by allowing raising verbs, such as *seem*, to bear specifications for the attribute SUBJ identical to those of their infinitival VP subjects; at the same time, the CONT specifications of raising verbs do not identify a semantic role for their subjects, only for their infinitival VP complements. This simple lexical property immediately ensures that raising verbs preserve the subject-selection properties of their complements. It also yields other familiar properties of raising verbs, such as their preservation of meaning invariance and idiomaticity under passivization. Similar observations follow for auxiliaries, if these are also taken to structure-share their subject values with their complements but not to associate these values with semantic roles.

Nonlocal dependencies

The selectional mechanism sketched above is not sufficient to guarantee the observed linkage between elements in the syntactic relationships usually referred to as unbounded dependencies. The difficulty can be illustrated in an example such as Figure 7. The examples in Figure 7 show that a significant amount of grammatical information is shared between the fronted constituent and the place from which some element is clearly missing, notated with an underscore. The apparently displaced material in brackets is an NP, and the reader can verify that replacing it with an AP or a PP, or indeed any other phrasal category, gives rise to an ill-formed result. Furthermore, it is not only the gross categorial description that is shared between this filler and the gap site: information about the number required for subject–verb agreement is also part of the linkage. Case information is also evidently preserved; e.g. the requirement that

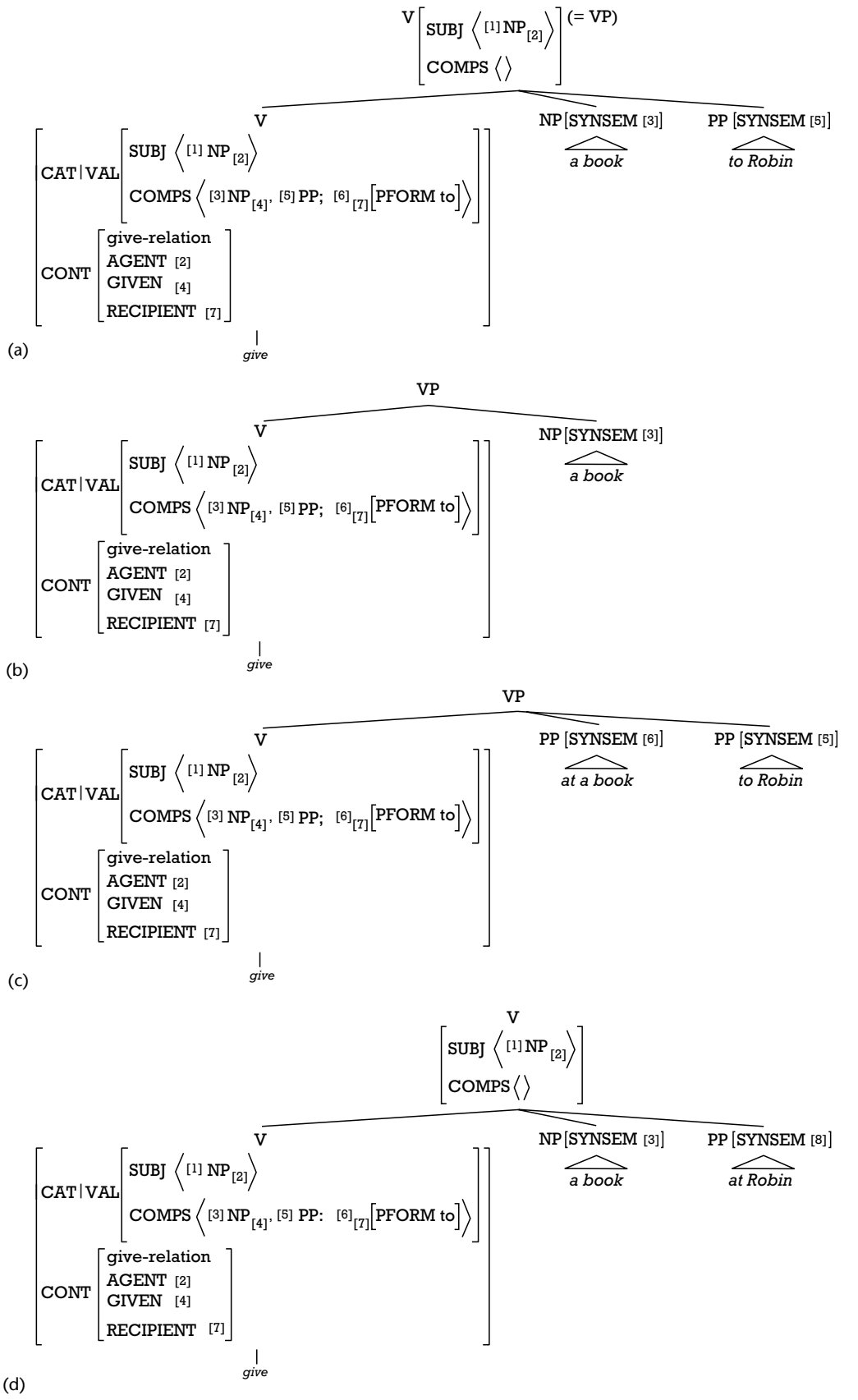

V [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ ⟩] (= VP)
V
NP [SYNSEM [3]]
PP [SYNSEM [5]]
a book
to Robin
CAT | VAL [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ [3] NP[4], [5] PP; [6][7][PFORM to] ⟩]
CONT [give-relation AGENT [2] GIVEN [4] RECIPIENT [7]]
give
(a)
VP
V
NP [SYNSEM [3]]
a book
CAT | VAL [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ [3] NP[4], [5] PP; [6][7][PFORM to] ⟩]
CONT [give-relation AGENT [2] GIVEN [4] RECIPIENT [7]]
give
(b)
VP
V
PP [SYNSEM [6]]
PP [SYNSEM [5]]
at a book
to Robin
CAT | VAL [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ [3] NP[4], [5] PP; [6][7][PFORM to] ⟩]
CONT [give-relation AGENT [2] GIVEN [4] RECIPIENT [7]]
give
(c)
V [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ ⟩]
V
NP [SYNSEM [3]]
PP [SYNSEM [8]]
a book
at Robin
CAT | VAL [SUBJ ⟨ [1] NP[2] ⟩ COMPS ⟨ [3] NP[4], [5] PP: [6][7][PFORM to] ⟩]
CONT [give-relation AGENT [2] GIVEN [4] RECIPIENT [7]]
give
(d)

Figure 5.

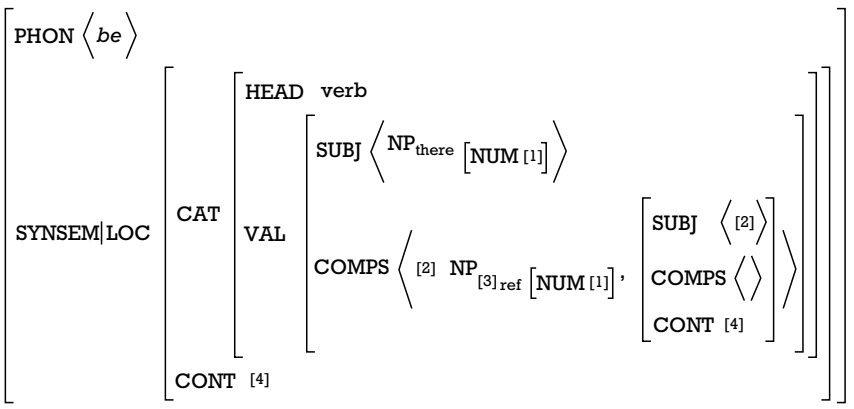


Figure 6.

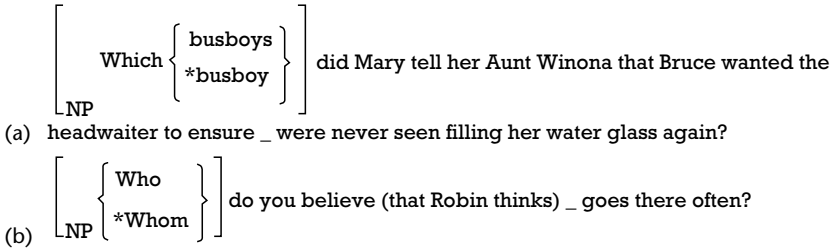


Figure 7.

subjects of a finite verb must be nominative is apparently imposed on arbitrarily distant fillers associated with the gap. Such data were for decades taken as the strongest possible evidence that derivational machinery – by allowing the filler to satisfy contextual requirements at a level of structure prior to its displacement in subsequent stages – is essential for an adequate account of syntactic phenomena.

It is evident that the local mechanisms introduced above cannot mediate such relationships. These mechanisms require the components of the dependency to be visible to a single lexical head, allowing properties of subjects and VPs to be correlated in appropriate fashion, giving rise to the dependencies observed, whereas in a nonlocal dependency, there is no single lexical head in whose VAL list specifications the properties of the filler and of the gap site may both be specified. Indeed, the filler in such constructions typically appears in a configurational relationship with a clause which, by definition, is saturated for all valence features, so that there is no way for information about the filler to be linked by means of such valence features. The solution, inherited in HPSG from work by Gazdar (1981) and Gazdar *et al*. (1985), is to introduce a separate feature SLASH, whose values are a set of LOC specifications – all the information contained in SYNSEM values except for information about nonlocal features, including SLASH itself – which is shared between a mother and at least one daughter. The propagation pathway for SLASH terminates in a gap under specific conditions ensuring that the crucial information about the filler is preserved at the gap site. In HPSG, following the approach in generalized phrase structure grammar, this is accomplished by defining a three-part mechanism for filler–gap linkages:

1. The 'top' of such linkages is licensed by identifying a clause whose DTRS value is of sort `head-filler-phrase`. This sort is constrained to specify a non-head daughter with a LOC value that is structure-shared with the SLASH value of the clausal head daughter. The effect is to guarantee the existence of structures of the form shown in Figure 8.

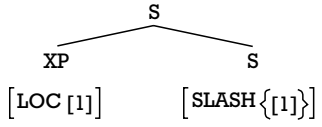


Figure 8.

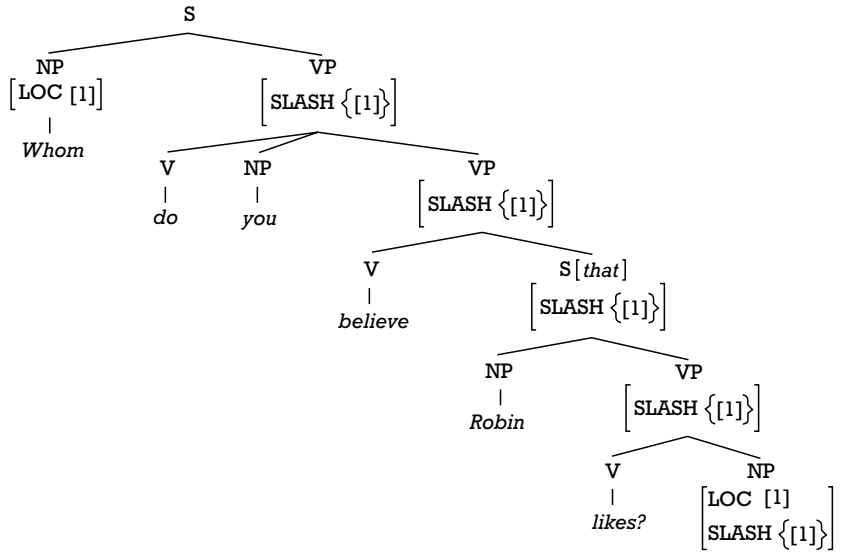


Figure 9.

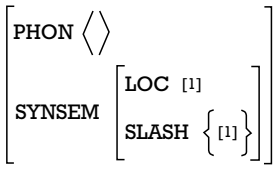


Figure 10.

2. The 'middle' of the linkage arises as the effect of a constraint, the 'nonlocal feature principle', which forces SLASH specifications on mothers to be shared with at least one daughter and vice versa, subject to conditions which ensure that the SLASH value at the top of the dependency is bound off at that point. Thus, once SLASH has been introduced into a structure, it must appear at every hierarchical level below the point of its introduction until it is terminated.
3. HPSG theorists are divided on how the SLASH pathway terminates. On one account, the SLASH specification is 'cashed out' by means of an empty category. An alternative account terminates the percolation of SLASH at the bottom of its path on a lexical head whose valence is reduced by an element corresponding exactly to the SLASH value of that head. Figure 9 illustrates the form of a filler–gap dependency using the first approach.

Given the existence of the lexical entry in Figure 10, the SLASH path will terminate in a phonologically null category which – since it has a LOC value identical to that of the SLASH (and of the filler) – preserves all the categorial and content information of the filler, and therefore must be compatible with any restriction imposed at the gap site.

Syntactic conditions on anaphora

The obliqueness ordering on valence elements provides an essential component in the HPSG account of syntactic factors that enter into the determination of whether two given constituents can be interpreted as denoting the same entity. A very concise set of principles – the so-called 'binding theory' – can provide an account of this relationship of coreference over a wide range of linguistic phenomena. The binding theory defines, for each of a small number of types of referring expression, the conditions under which it may be coreferential with some other component of the linguistic sign in which it appears. The HPSG framework is perhaps unique among phrase-structure-theoretic approaches in appealing to conditions on coreference that make no appeal to the configurational relationship between coreferring elements. Rather, the basis of the binding theory is the obliqueness relationship defined on the ARG−ST list, in which all the elements selected by some lexical head are displayed in order of ascending obliqueness, with subjects taken to be the least oblique elements on such lists.

An argument of a given head is said to 'locally o-command' any more oblique coargument. More generally, it 'o-commands' any more oblique coargument, the head of that coargument, any arguments of that head, and any other elements that the arguments of that head themselves o-command. For any lexical head, this recursive definition in effect picks out a path from some argument of that head through arbitrarily deep levels of structure linked by headship and selection. Its

application to the statement of coreference possibilities is illustrated below directly.

The HPSG binding theory can then be succinctly stated as follows:

1. Reflexive and reciprocal forms with referential local o-commanders must be coindexed by one such o-commander.
2. Pronouns that are not reflexive or reciprocal forms must not be coindexed with a local o-commander.
3. Non-pronominals cannot be coindexed with any other elements.

This theory of coreference provides a satisfactory account of a wide variety of coreference facts in English, illustrated in the sentences below:

Robin_i admires $\text{him}^*(\text{self})_i$. (1)

Pictures of $\text{him}(\text{self})_i$ intrigue Robin_i. (2)

$\text{It was only herself}_i$ that she_i had to blame for what happened. (3)

Robin_i believes that people_j admire $\text{her}(^*\text{self})_i$. (4)

*She_i believes that people_j admire Robin_i. (5)

*Who_i does she_i believe that people admire t_i? (6)

In sentence 1, the subject appears on the same ARG-ST list as the object, hence locally o-commands it. If the object is a reflexive, then by principle 1 above it must be coindexed with the local o-commanding NP, and by the same token, if the object is a personal pronoun, it must not be bound by a locally o-commanding antecedent. In sentence 2, on the other hand, there is no locally o-commanding antecedent, since the nominal head *pictures* has no other arguments; hence the example vacuously satisfies principle 1 and reference is syntactically free, though constrained by pragmatic factors. By the same token, the absence of any local o-commanding antecedent in this example allows the personal pronoun to appear. Thus, the often-noted failure of reflexives and personal pronouns to exhibit completely complementary distribution follows straightforwardly from the HPSG binding theory.

Sentence 3 again illustrates how reflexives can appear freely in the absence of a local coindexed o-commander; here, *it*, though it o-commands the reflexive, does not have an index of sort ref, hence, by principle 1 above, it does not require a coindexed local o-commander. In sentence 4, a local, referential o-commander is present with which a reflexive fails to be coindexed, yielding ill-formedness. By the same token, however, a pronoun contraindexed with the local o-commander can appear, in accordance with principle 2 above.

Sentences 5 and 6 illustrate the ill-formedness resulting when non-pronominals are coindexed. In both cases, *she* as matrix clause subject is less oblique than the clausal complement of the matrix verb, and therefore (locally) o-commands this clause. In view of the preceding discussion, it therefore o-commands the head of this clause, *admire*, and all the elements that *admire* itself selects, including *Robin* in sentence 5 and the trace of *who* (which is taken to be non-pronominal) in sentence 6. Hence these non-pronominal elements are o-commanded by, and coindexed with, *she*, in violation of principle 3 above, and sentences 5 and 6 are therefore ruled out as required.

Preliminaries to Semantic Interpretation

Every sign in an HPS grammar is specified for a CONT value which provides the input to semantic interpretation. Nominal content values, of type nom-obj, identify an index and a set of restrictions on that index; thus, the noun *gift* optionally combines with two PPs, *of* $\text{NP}_{[2]}$ and *to* $\text{NP}_{[3]}$, whose own CONT values are identified with those of their complement daughters, yielding nominals such as *gift of a book to Robin*, whose CONT has the form shown in Figure 11. This partially saturated nominal will then combine with some determiner, such as *the* or *every*, whose own content identifies its quantificational properties, and whose lexical specifications allow it to take the content of a nominal head such as Figure 11 as the restriction on the relevant quantifier. The semantic content of the quantifier is identified as the NP's QSTORE value, an implementation in a feature-based framework of earlier proposals made by Robin Cooper (e.g. Cooper, 1983). QSTORE values of daughters are added to those of their mothers progressively up the tree, to the point where a quantifier is retrieved from the QSTORE of some category and

$$\begin{bmatrix} \text{INDEX} & [1] & \\ \text{RESTR} & \begin{bmatrix} \text{gift-relation} \\ \text{INSTANCE} & [1] \\ \text{GIVEN} & [2] \\ \text{RECIPIENT} & [3] \end{bmatrix} \end{bmatrix}$$

Figure 11.

added to the QUANTS value at that same node, thus fixing its scope. When the top of the tree is reached, all elements have been retrieved from QSTORE, with their left-to-right order identifying their relative scoping.

Thus, essentially two kinds of information must be tracked under this architecture: on the one hand, there is a kind of quantifier-free 'nuclear' content, and on the other the retrieval and ordering of scoping among quantifiers over structures of arbitrary depth. The former corresponds to the successive embedding of argument structures as parts of increasingly complex linguistic expressions, as required by the valence principle and other constraints of the grammar. The latter may be thought of as the gathering of quantifier contents into the QSTORE set at successively higher levels in phrasal structure and their removal at the point in the structure where they are required to take scope in order for a given reading to be represented. This dual percolation of information is regulated by the constraints already mentioned, and a separate condition, the 'semantics principle'. The retrieval of quantifiers from QSTORE is nondeterministic, so that all possible scopings are realized, subject to whatever empirically motivated restrictions prove necessary.

ADVANTAGES AND LIMITATIONS

Among the various competing grammatical frameworks currently under investigation, HPSG enjoys a number important strengths:

- Like other monostratal frameworks, HPSG can compactly express analyses of natural language phenomena, incorporating significant generalizations, without the assumption of derivational histories to encode the relevant properties of the phenomena described.
- The assumption of phrase structure configurations in natural language provides a straightforward basis for defining the percolation of information in syntactic objects. In particular, syntactic heads can serve as the locus of feature sharing between mother and daughter categories. This has been used in recent HPSG treatments of the distribution of nonlocal features and of QSTORE.
- The formal foundations of HPSG are mathematically explicit and well understood, taking the form of a logic of constraint satisfaction which ensures a complete interpretation for all posited constraints. This guarantees that any inconsistent components of an HPSG analysis are in principle detectable.
- The monotonicity of HPSG's mechanisms for regulating the distribution of grammatical information in syntactic structures, combined with the formal explicitness already noted, makes computational implementation of HPSG analyses particularly straightforward. Indeed, much of HPSG's theoretical apparatus has been developed in tandem with computational practice (e.g. the existence of efficient unification algorithms). Although defaults have sometimes been introduced, the nature of these defaults is essentially abbreviatory, in that every such grammar can be compiled out to one in which no defaults appear, using more restricted formulations of the relevant constraints. The essentially abbreviatory status of such defaults ensures that, from a computational point of view, all current HPSG implementations can be regarded as monotonic.
- The use of rich modeling objects such as typed feature structures allows the theory to combine a high degree of descriptive flexibility (conspicuous, for example, in its ability to model various kinds of unbounded dependency construction) with a feature-geometric organization that captures important modularity properties of natural language grammars.

The downside of HPSG's phrase structure basis emerges in contexts where it is less clear that the elements involved in the relevant constructions are definable in highly configurational terms. Coordination, for example, manifests in a number of phenomena in which it appears that string parallelism, rather than constituency, is the basis of a satisfactory account. HPSG currently lacks a detailed theory of coordination, and nonconstituent coordination, in particular, is problematic for phrase-structure approaches. The greater combinatorial flexibility of categorial grammar probably compares most favorably to phrase-structure approaches in this empirical domain. Currently, the lack of a well-developed approach to coordination is one of the most serious deficits in HPSG's theoretical coverage.

A second difficulty faced by current HPSG is a consequence of defining the relevant command relation for binding theory strictly in terms of valence. Since, under conventional assumptions about adjunct elements, adverbs and other modifiers are not lexically selected by heads, the binding theory sketched above should be inapplicable to elements within adjuncts. But it is not clear that this prediction is empirically warranted. Assuming that adjuncts are indeed lexically selected brings adjuncts into line with clearly selected elements, but again in a way that clashes with significant data. It is not clear, therefore, whether a strictly valence-based account of coreference possibilities is tenable, or if a configurational component must be reintroduced in order to bring the binding theory into line with the full range of relevant facts.

FUTURE PROSPECTS

Important discoveries tend to deflect established lines of research onto unforeseen pathways, but certain trends already evident in HPSG are very likely to become increasingly significant. In particular, future work in HPSG will almost certainly focus on a number of issues where the current theory is notably incomplete:

- New approaches to problems posed by word order, in particular cases of so-called 'discontinuous constituency', are likely be be developed based on what has been called the 'linearization framework' within HPSG, in which word order facts are associated with a descriptive domain related to, but formally separate from, that which defines constituency.
- The 'constructional' approach to syntactic phenomena (Sag, 1997), relying on elaboration of the phrasal sort hierarchy to capture both regular and idiosyncratic syntactic properties, will become increasingly prominent. Exceptional behavior, such as the fact that relative clause modifiers with neither *wh* nor *that* formatives cannot be attached to nominal heads corresponding to their subjects (*the guy Robin criticized is here* / **the guy criticized Robin is here*), or the fact that PPs containing *wh* can link infinitival relative clauses to nominal heads but *wh* NPs themselves cannot (*something about which to complain/*something which to complain about*) will increasingly be handled through sort declarations that encode them as construction-specific constraints. It must be stressed, however, that there is far from unanimous acceptance of this 'constructional turn'.
- The semantic component of HPSG is still seriously underdeveloped, following its dissociation from the situation-theoretic perspective which guided much earlier work. The CONT and QSTORE features provide information which constitutes a kind of logical form representation, but there is currently no generally accepted constraining relationship between such representations and empirically motivated formal models. A particularly stubborn obstacle to the formulation of such relationships is the difficulty posed by major outstanding empirical problems, such as the semantic nonequivalence in opaque contexts of intensionally equivalent objects. Research currently in progress in HPSG is aimed at eliminating this deficit of the theory, and will probably intensify.
- Finally, there has already been work on constraint-based phonology which extends the HPSG approach of using sorted feature structures to state mutually constraining representations to phonology, with the ultimate objective of providing a completely nonderivational theory. An overview of work along these lines, with pointers to relevant sources, is given in Bird and Klein (1994). The application of sorted feature structures and other aspects of the HPSG formalism will probably become increasingly important for foundational work in phonology.

References

Bird S and Klein E (1994) Phonological features in typed feature systems. *Computational Linguistics* **20**: 455–491.

Brame M (1976) *Conjectures and Refutations in Syntax and Semantics*. New York, NY: North-Holland.

Cooper R (1983) *Quantification and Syntactic Theory*. Dordrecht: Reidel.

Culicover PW and Wilkins W (1984) *Locality in Linguistic Theory*. New York, NY: Academic Press.

Gazdar G (1981) Unbounded dependencies and coordinate structure. *Linguistic Inquiry* **12**: 155–184.

Gazdar G, Klein E, Pullum GK and Sag I (1985) *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Gazdar G, Pullum GK, Carpenter R *et al*. (1988) Category structures. *Computational Linguistics* **14**: 1–19.

Pollard CJ and Sag I (1987) *Information-Based Syntax and Semantics*. Stanford, CA: CSLI.

Pollard CJ and Sag I (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Richter F, Sailer M and Penn G (1999) A formal interpretation of relations and quantification in HPSG. In: Bouma G, Hinrichs E, Krijff GJM and Oehrle R (eds) *Constraints and Resources in Natural Language Syntax and Semantics*, pp. 281–298. Stanford, CA: CSLI.

Sag I (1997) English relative clause constructions. *Journal of Linguistics* **33**: 431–484.

Further Reading

Borsley R (1996) *Modern Phrase Structure Grammar*. Oxford: Blackwell.

Borsley R (1999) *Syntactic Theory*, 2nd edn. London: Arnold.

Green G and Levine RD (1999) Introduction. In: Levine RD and Green G (eds) *Studies in Contemporary Phrase Structure Grammar*. Cambridge, UK: Cambridge University Press.

Hukari T and Levine RD (1996) Phrase structure grammar: the next generation. *Journal of Linguistics* **32**: 465–496.

Kathol A (2000) *Linear Syntax*. Oxford: Oxford University Press.

Pollard CJ and Sag I (1987) *Information-Based Syntax and Semantics*. Stanford, CA: CSLI.

Pollard CJ and Sag I (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Sag I and Wasow T (1999) *Syntactic Theory: A Formal Introduction*. Stanford, CA: CSLI.

Pragmatics, Formal

Intermediate article

Craige Roberts, Ohio State University, Columbus, Ohio, USA

CONTENTS

Formal pragmatics is a subfield of linguistics, with a focus on the influence of context on interpretation. It has close ties to the study of formal semantics, as well as to philosophy of language and to language processing in the field of artificial intelligence.

PRAGMATIC PHENOMENA

Within linguistics and the philosophy of language, the study of pragmatics over the past century has encompassed a number of phenomena pertaining to the way that context influences interpretation, including at least the following:

- what a given utterance may be taken to presuppose, especially via conventional presupposition triggers (conventional implicature);
- the various ways in which speakers may be taken to mean more than they say (conversational implicatures);
- how we understand indexical expressions (including first- and second-person pronouns, and expressions such as 'this' and 'that') and other context-sensitive expressions (e.g. adjectives such as 'local', 'former', 'relevant', 'good', among many others);
- sentence mood (e.g. when is 'you will like broccoli' an assertion, when a question, when a command?) and the relations between moods (what is an acceptable answer to a given question? – does it mean the same thing to reply to a question with 'yes', as to an assertion, or to a command?);
- speech acts (e.g. when is 'I will eat my broccoli' a prediction? a promise? a threat? – and how do we differentiate between these (and other) things we can do in uttering this sentence?);
- the ways in which speakers negotiate the construction of their dialogue, including conversational turn-taking;
- rhetorical structure, the ways in which different utterances, even by the same speaker, may be related to build an argument, make an explanation, etc.

More recently, there has been considerable interest in how contextual factors play a role in these facets of semantic interpretation, as well:

- anaphora resolution: how we find the intended antecedents for various anaphoric elements in discourse, including third-person pronouns, but also null arguments (as in Japanese or Spanish, where subjects are optional) and other forms of ellipsis, and non-nominal proforms (pro-VerbPhrases, pro-clauses, etc.);
- the semantic role(s) of intonation, and the related phenomena of Focus and Topic (which very often involve prosodic factors such as intonation, but may also involve word order, special syntactic constructions, or morphological markings);
- domain restriction: the ways in which the interpretation of quantificational operators is influenced by context (e.g. when we say 'Everyone seems cold today', we generally mean some relevant subset of the set of all people; when we use a modal auxiliary such as the deontic 'should' we generally assume that the assertion in question is true relative to certain background assumptions about what is right and reasonable; when we say 'John only likes BAGELS', we generally mean that the only relevant property John has is liking bagels, ruling out perhaps liking toast or sausage, but not breathing or liking dogs).

GOALS OF A FORMAL PRAGMATIC THEORY

Until the 1970s, most of the theories proposed to explain such pragmatic phenomena were informal. But since that time, more and more work on pragmatics has been carried out in the spirit of generative linguistics. In a generative theory, for example of the syntactic structure of a particular language, one attempts to characterize all and only those utterances which are grammatical in the language in question. We can say that a generative grammar predicts whether a given sentential structure is grammatical or ungrammatical (in the language). In order to be fully explicit and clear about these predictions, such theories are relatively formal, formulating hypotheses (or *rules* in the grammar) in such a way that they are both clear as to their conditions of application and unambiguous about the results. Formal linguistic theories hence provide more predictive power than

informal theories. Instead of *post hoc* explanations, we have clear predictions about what will be acceptable to native speakers of the language in question.

Where pragmatics is concerned, the notion of acceptability is not that of grammaticality, as in syntax, but of meaning and felicity in context. The data derive from the behavior of native speakers as this reflects their intuitions about whether a given utterance seems felicitous in a particular context, and, if so, about what they take the utterance to mean in that context. For example, in (1) and (2), the same sentence is uttered by a speaker B in two different contexts:

[Context: A and B are perfect strangers, standing at a bus stop. B turns to A and says:] 'I got splashed by a bus this morning, too.' (1)

[Context: A and B are perfect strangers, standing at a bus stop. A bus pulling up to the curb splashes A with water. B turns to A and says :] 'I got splashed by a bus this morning, too.' (2)

In neither context has there been prior dialogue. But in (2), the utterance seems perfectly natural, while in (1) it does not. We say that the utterance of this sentence is *felicitous* in the context given in (2), while it is *infelicitous* in (1); we often mark infelicity with a '#' at the beginning of the sentence, a pragmatic parallel to the use of '*' to mark ungrammaticality in syntax.

Just as context may determine felicity, it may license certain types of variation in understood meaning. B's utterance in (3) takes place in a different context from that of the same sentence in (4), and because of this difference in context of utterance, the two utterances are likely to convey different information:

A: Do you have any idea who might have three chairs they could loan me for the party tonight?
B: I'm pretty sure Mary has three chairs. (3)

A: I need to know how many chairs each of these people on the planning committee owns, so I know how many chairs we already have for the party.
B: I'm pretty sure Mary has three chairs. (4)

We are more likely to understand B in (3) to mean that Mary has three or more chairs, while in (4), we are more likely to understand B to mean that she has exactly three.

As illustrated by these examples, in explaining what determines felicity and meaning in pragmatic theory we cannot simply consider the linguistic structure(s) of the sentence uttered and the semantic rules for interpreting these structures. Rather, pragmatic phenomena generally involve interaction between the utterance (as characterized by its various linguistic structures – syntactic, prosodic, morphological, etc.) and some aspect or aspects of the context of utterance. So the fundamental requirements for a formal pragmatic theory are (1) a theory of the linguistic structures of an utterance, (2) a theory of linguistic context, and (3) a theory of how these interact to yield the attested felicity and interpretations. We look to the other subfields of linguistics – syntax, phonology, morphology – for (1). Then the development of (2) and (3) constitute the principal goals of formal pragmatic theory.

CONTEXT AND CONTEXT CHANGE

A theory of linguistic context should make clear the kinds of information to which interlocutors in a conversation typically have access that may bear on the interpretations they give to utterances in that conversation. And it should organize that information in such a way as to optimally facilitate the characterization of the dynamics of information change over the course of the conversation. One question about the information influencing interpretation is *whose* information this is. Stalnaker (1972) argued that the relevant information is that in the common ground of the interlocutors – that information which they commonly (purport to) hold true, characterized as a set of propositions (each proposition a set of possible worlds). Although many authors since have assumed that the common ground is the correct information set for pragmatic interpretation, others have argued that it is, instead, the speaker's assumptions about the hearer's information which are critical, while others argue for a more complex notion than the common ground, for example, including a number of alternative possible common grounds to reflect the fact that we are not always able to determine with confidence exactly what our interlocutors believe. As these different conceptions reflect, it seems clear that successful communication requires that a speaker attempt to keep track of the information shared by all the interlocutors. The question, then, is whether the theory should idealize to the case where this attempt is relatively successful, or instead attempt to capture directly

the fact that keeping track of shared information is fraught with difficulty in the actual case.

Early work in formal pragmatics came out of formal semantics. Context in the theories of semanticists who followed the general approach of Montague was captured as a set of indices, or contextual parameters. These were pointers to specified sorts of contextual information, utilized to feed the relevant information into the process of compositional interpretation which yielded the proposition expressed by the sentence in the specified context. This limited set of indices typically included the world and time of utterance (a way of capturing the facts about the utterance situation, as well as interpreting tenses and utterances of such words as 'now'), the speaker and sometimes the addressee (for 'I', 'we', 'you', etc.), the location of the utterance (e.g. for the interpretation of 'here' or 'local'), and a function assigning values to free variables (the logical form counterparts of pronouns and other proforms). Additional indices were sometimes posited for elements such as indicated objects (for deixis accompanying 'this' and 'that') or even the relative status of the interlocutors (e.g. for Japanese honorifics or French *tu* versus *vous*) and the level of formality of the discourse. However, it is not clear that one could in principle specify a finite set of indices of this type which would be adequate to all the types of information relevant for capturing pragmatic influences on interpretation. Moreover, the values given by these indices for a given sentence were arbitrarily selected, without any mechanism for keeping track across a given discourse of what was being talked about and how this might bear on the interpretation of that sentence in that discourse.

Following Stalnaker (1972), Gazdar (1979) worked with a rather different notion of the context of utterance, conceived of as a set of propositions, and treated the meaning of an utterance as a function from contexts into propositions. Gazdar used this notion to develop a formal theory of implicature and presupposition in which the felicity of an utterance depended on whether the context of utterance could be consistently updated with the (propositional) information presupposed or implicated.

Theories of dynamic interpretation like those in the pioneering work of Heim (1982) and Kamp (see Kamp and Reyle, 1993) take the notion of context-incrementation yet further, marking a turning point in the development of pragmatics. In this work the meaning of an utterance is neither a proposition nor a function into propositions, but a function from contexts (of utterance) to (updated) contexts; Heim called this the utterance's *context change potential*. This approach offers a new dimension to Bar-Hillel's earlier characterization of an *utterance* as a pair of a sentence and a context. On the dynamic view of interpretation, we might consider an utterance to be a pair of the sentence under a linguistic analysis, or its logical form, and an input context, the context just prior to utterance. Given that the logical form is conventionally (perhaps even compositionally) correlated with a context change potential, this implies as well an output context, that is, the value of the context change potential given the input context as argument. For example, utterance of (5) in the context suggested will result in the updated context indicated:

Input context, IC: Information characterized as a set of propositions, including the proposition that B is speaking.
B: I am hungry.
Output context: IC plus the proposition that B said *I am hungry* in IC and (assuming no one questions B's trustworthiness) the proposition that B is hungry. (5)

The dynamic theories cited take different views of what context is. Heim takes it to be an elaboration of Stalnaker's common ground to include not only the set of propositions that the interlocutors hold in common to be true, but also a set of *discourse referents*. A discourse referent is an abstract entity-under-discussion. Such an entity may or may not actually exist – we can talk about hypothetical entities, even nonexistent ones – but we keep track of the information about each such entity across discourse. Heim characterizes a discourse referent informally as a file card; technically, it is an index, corresponding to the referential index on the noun phrases (NPs) used to refer to this entity in the discourse. Keeping track of discourse referents permits a theory of definite NPs such as 'the sun' in which such an NP carries a presupposition of familiarity to the interlocutors; that is, its utterance presupposes that there is a corresponding discourse referent in the input context of interpretation. Indefinites such as 'a cloud' are said to carry novelty presuppositions, requiring that in an input context there be no pre-existing corresponding discourse referent. Heim's context, then, is an abstract notion, a set with two kinds of information.

Kamp's Discourse Representation Theory (DRT) is very similar in spirit to Heim's Context Change Semantics and obtains many of the same results (as well as encountering many of the same problems). For example, discourse representations contain elements which behave very much like Heim's

discourse referents, as well as formulae that play much the same role as Heim's propositional component of the common ground. But instead of Heim's sets of information, DRT offers instead a representational characterization of context, as a particular type of logical form. And Kamp seems to conceive of the context in more psychologistic terms than Heim, as a mental representation or mental model, a tendency continued in much of the subsequent work in DRT and its descendants. But differences aside, in both these theories as well as in other subsequent work on dynamic interpretation, most contextual information can be characterized, directly or indirectly, in propositional terms, with a proposition viewed as a set of possible worlds (or situations).

Dynamic theories offer a number of advantages over the earlier index-based theories of context. Since contextual information is basically propositional in the dynamic theories, it is no longer necessary to attempt to characterize that information as a set of indices, with all the awkwardness of attempting to determine just how many indices, and of what character, are required. With no loss of theoretical elegance, there may be any number of different types of propositions in the context, bearing on the interpretation of an utterance in as many different ways. Moreover, the context now contains information about prior discourse and the situation of discourse, and it is this information which plays a central role in constraining the interpretation of any anaphoric or deictic elements which may be used. Heim treats such elements as presuppositional, and in subsequent work proposes an important extension of Context Change Semantics to include a full theory of utterance presuppositions and of presuppositional felicity in context. In this extension, an utterance presupposition is taken to be a constraint on contexts of utterance; technically, the context change potential corresponding to the utterance's logical form is undefined for any context which does not satisfy the presupposition in question. For example, in (2) above, B's utterance presupposes that someone other than B has been splashed by a bus, a presupposition conventionally triggered by the adverbial 'too'. The utterance is felicitous in this context because the utterance's presupposition is satisfied by the context, which makes it clear to both interlocutors that A (someone other than B) has just been splashed by a bus. But nothing in the context of (1) gives us this type of information. Hence, the (same) utterance presupposition fails in (1), so that the utterance is infelicitous; that is, context update is undefined for utterance of this sentence in this particular context.

Thus, dynamic theories of interpretation avoid the arbitrariness and disconnectedness of the earlier, index-based theories; each utterance looks to the preceding context for presupposition satisfaction, and in turn updates it with the information it contains. Still, these theories are primarily designed to capture logical relations between utterances. And because of this, they offer insights into only a few of the types of pragmatic phenomena listed at the outset of this article, presupposition and (in part) anaphora resolution and indexicality. What of the other phenomena which a full-fledged theory of formal pragmatics ought to say something about?

TOWARDS A UNIFIED PRAGMATIC THEORY: QUESTIONS, INTENTIONS, RHETORICAL RELATIONS

Recently, several theorists have attempted to develop more ambitious, inclusive formal theories of pragmatics. There seem to be two main trends in this work, one focusing on speakers' intentions and the role of these intentions in organizing discourse, the other on the role of rhetorical structures in discourse. Some of the most interesting work in each of these directions builds on the successes of the dynamic theories by extending one of those theories to allow for other dimensions of the information shared in conversation.

The first trend, focusing on speakers' intentions and the plans which these intentions constitute, comes as much out of work in artificial intelligence as out of linguistics or philosophy. Grice's (1967) conversational principles acknowledged the central role of speakers' intentions, what he called 'the purposes of the exchange', in conveying conversational implicatures in discourse. For example, we can understand what someone means only if we understand the purposes of the conversation and assume that their utterance pertains to those purposes. In artificial intelligence, Grosz and Sidner (1986) argue that speakers' intentions are the central organizing features of discourse, determining what information in the discourse is salient at any given point. They develop a theory based on this intentional structure to help explain anaphora resolution, offering insight into facets of that problem not addressed by the dynamic theories, but do not consider the role of compositional semantics in determining interlocutors' plans. Carlson (1983) develops a theory of discourse as structured by questions, and the relationships between questions and between questions and answers, as an extension of the game-theoretic semantics of Hintikka.

However, the theory is couched in syntactic terms, offering a grammar for discourse without making reference to the interpretations of the structures in question, and so again, fails to connect with the literature in formal semantics.

The work of Roberts (1996) exemplifies the use of intentions and plans in the analysis of discourse pragmatics in such a way as to preserve the results of formal semantics, borrowing on earlier work on the semantics of questions and the (formal) pragmatics of answers due to Groenendijk and Stokhof (1984). In Roberts' theory, discourse is organized by questions and the relations among them. The basic idea is that discourse is essentially a game of intentional inquiry. In this game, questions establish the local goals in a discourse – answering these questions, and thereby constrain what a speaker may say at a given point. Interlocutors are required to address the immediate question under discussion; anything else would be irrelevant. Knowing this, and knowing the question under discussion, a hearer can make certain reasonable inferences about what information the speaker intended to convey with a particular utterance. For example, A's question in (3) makes it clear that she is only interested in knowing who has at least three chairs, since that's all that matters to her for being able to borrow three. But in (4), A's question is about determining the total number of chairs owned, so that she needs exact figures to the extent possible. Someone responding cooperatively to one of these questions will only give the type of information requested, leading to the quantity implicature in (4) ('exactly three chairs'), but not in (3). Welker (1994) argues that under this type of view, conversational implicatures like that in (4) become contextual entailments, and that so-called implicature cancellation really amounts to context revision. That is, if context is sufficiently well defined, one can predict which implicatures will arise as a consequence of a given utterance, and even offer minimal pairs of contexts which differ in which implicatures they trigger for a given utterance. This approach can be extended to yield the intended understanding of ellipses (see Ginzburg, 1996), and the intended domain restriction for any operators in an utterance.

In addition, Roberts argues that Focus can be adequately analyzed as a presupposition about the question which an utterance is intended to address, whether that question is explicit or merely implicit; this would represent a simplification of the ground-breaking formal work on Focus by Rooth (see Rooth, 1992), which posits multiple functions for Focus. Finally, as Grice had predicted, the conversational maxims (e.g. Relevance) may be seen as theorems which follow from the basic goals and structure of the language game, rather than independent axioms of the theory. For example, given interlocutors' commitment to the established discourse goals of the language game, the questions under discussion, an utterance is Relevant if and only if it addresses the immediate question under discussion. Roberts' work is couched as an extension of the dynamic semantics of Heim, in which the discourse context is treated as a tuple, including both the common ground, or propositional information, and a set of questions under discussion, as well as other types of information. Each utterance is interpreted as a context change potential, updating the complex tuple of the input context by adding either a new question for discussion (if the utterance is interrogative in mood) or a new proposition to the common ground (if the utterance is indicative in mood).

The second trend in contemporary work on formal pragmatics is exemplified by the work of Asher and Lascarides (1998), who borrow from an earlier tradition of rhetorical analysis to develop a set of primitive rhetorical relations between utterances in a discourse, including Explanation, Elaboration, and Narrative, among others. They use an extension of Kamp's DRT in which the discourse representations for various clauses, determined on the basis of syntactic structure, are 'glued' together with these rhetorical relations. They offer detailed hypotheses about how the particular rhetorical relation intended for a given pair of clauses is determined by the hearer, using a default logic, on the basis of the compositional semantic interpretation of the clauses involved, representations of the assumed cognitive states of different interlocutors, a set of pragmatic principles, and various default assumptions. They argue that sentence mood and the speech act which a speaker intends to perform can be inferred on the basis of this information.

However, besides arguing for rhetorical relations as primitives of their theory, Asher and Lascarides also consider the crucial role of speakers' plans and intentions, as represented by particular types of questions, in drawing pragmatic inferences. One issue which subsequent work must address is whether rhetorical relations should be taken as primitives in a theory of discourse, or whether they can instead be defined in terms of the interlocutors' independently inferred intentions and plans. In the meantime, it seems clear that plans, intentions, and the intended relations between utterances in discourse need to be captured by one's

theory of discourse, and hence by pragmatic theory. In order to do this, contexts must be more than simply a set of indices or propositions (possibly with discourse referents). They must also include the questions under discussion in the discourse at that point, the interlocutors' intentions, and information about the relationships between the propositions and questions proffered in the discourse. And eventually, of course, contexts must also take into account the third major type of mood, the imperative.

PRAGMATICS AND LINGUISTICS

Current work in formal pragmatics raises interesting questions about the place of pragmatics in linguistic theory. For one thing, we might wonder about the extent to which the pragmatic theories and principles considered appear to hold across different languages (and different cultures). We know, for example, that rules of politeness do differ from culture to culture, and hence that while some languages include certain types of honorifics, and even special sublanguages, for addressing interlocutors of particular status, other languages do not. How deeply do such differences go? The work discussed so briefly here seems to assume implicitly that the pragmatic principles posited are universal in the use of human language. If questions or rhetorical relations underlie the organization of discourse, they do so regardless of the language in use (and the culture of the people who use it). If this is true – and it is an empirical question – why would it be so? Both the philosophical underpinnings of pragmatics and the particular theories under development suggest an answer: pragmatic principles are not part of the grammar of a language. Rather, as Grice (1967) argued, they reflect essential features of what it is to communicate with human language, rational principles which guide its optimal use. Hence, exploring pragmatics cross-linguistically may help to make more precise what exactly the human language faculty does, and how it is related to other cognitive capacities, including reason.

References

Asher N and Lascarides A (1998) Questions in dialogue. *Linguistics and Philosophy* **21**(3): 237–309.

Carlson L (1983) *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht, Netherlands: Reidel.

Gazdar G (1979) *Pragmatics: Implicature, Presupposition, and Logical Form*. New York, NY: Academic Press.

Ginzburg J (1996) Dynamics and the semantics of dialogue. In: Seligman J and Westerstähl D (eds) *Language, Logic and Computation*. Stanford, CA: CSLI Publications.

Grice P (1967) Logic and conversation. William James Lectures, Harvard University. Published in Grice P (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Groenendijk J and Stokhof M (1984) *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD dissertation, University of Amsterdam, The Netherlands.

Grosz B and Sidner C (1986) Attention, intentions and the structure of discourse. *Computational Linguistics* **12**: 175–204.

Heim I (1982) *The Semantics of Definite and Indefinite Noun Phrases*. PhD dissertation, University of Massachusetts, Amherst.

Kamp H and Reyle U (1993) *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Roberts C (1996) Information structure: towards an integrated theory of formal pragmatics. In: Yoon J-H and Kathol A (eds) *OSU Working Papers in Linguistics*, vol. 49: *Papers in Semantics*. The Ohio State University Department of Linguistics.

Rooth M (1992) A theory of focus interpretation. *Natural Language Semantics* **1**(1): 75–116.

Stalnaker RC (1972) Pragmatics. In: Davidson D and Harman G (eds) *Semantics of Natural Language*. Dordrecht, Netherlands: Reidel.

Welker K (1994) *Plans in the Common Ground: Toward a Generative Account of Implicature*. PhD dissertation, The Ohio State University.

Further Reading

Beaver D (1997) Presupposition. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 939–1008. Amsterdam: Elsevier Science, and Cambridge, MA: MIT Press.

Bosch P and Van der Sandt R (eds) (1999) *Focus: Linguistic, Cognitive, and Computational Perspectives*. New York, NY: Cambridge University Press.

Cohen P, Morgan J and Pollack M (eds) (1990) *Intentions in Communication*. Cambridge, MA: MIT Press.

Culicover P and McNally L (eds) (1998) *Syntax and Semantics*, vol. 29: *The Limits of Syntax*. San Diego, CA: Academic Press.

Halliday MAK and Hassan R (1976) *Cohesion in English*. London, UK: Longman.

Kadmon N (2001) *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus*. Oxford, UK: Blackwell.

Levinson SC (1983) *Pragmatics*. Cambridge, UK: Cambridge University Press.

Stalnaker RC (1979) Assertion. In: Cole P (ed.) *Syntax and Semantics* 9.

Van Deemter K and Kibble R (eds) (2002) *Information Sharing*. Stanford, CA: CSLI Publications.

Yourgrau P (ed.) (1990) *Demonstratives*. Oxford, UK: Oxford University Press.

Production–Comprehension Interface

Intermediate article

Victor S Ferreira, University of California, San Diego, California, USA

CONTENTS

Many features of spoken language, from the nature of child-directed speech to audience design effects, reflect a sensitivity in language production to the needs and strategies of language comprehension. These adjustments, sometimes deliberate, sometimes automatic, ensure successful communication.

INTRODUCTION

The primary reason that speakers speak is so that their listeners can understand them. It is therefore unsurprising that unlike many cognitive tasks (such as perceiving, remembering, or decision-making), the language production performance of an individual speaker must take into account the processing capabilities and the knowledge states of another, namely, that speaker's intended listener. Research into this topic of the nature of the production–comprehension interface generally explores how production accommodates its processing to the needs of comprehension.

This article discusses the production–comprehension interface by exploring two related issues. The first concerns how speakers cater specific details of their utterances to take into account the knowledge and the comprehension capabilities and strategies of their listeners. The second narrows in on how speakers use common ground in their utterances – information that the speaker and listener believe to be mutually known in their present conversation or discourse.

INCLUSION OF INFORMATION BASED ON LISTENER KNOWLEDGE

The ways that speakers and listeners try to accommodate one another in conversations was described by Grice (1975). A Gricean approach to language specifies that interlocutors tacitly agree to the cooperative principle when they participate in a conversation, namely, that they agree to be mutually productive and efficient participants in their conversation. The cooperative principle is instantiated by four specific maxims. The maxim of *quality* states that interlocutors mutually assume that contributions to a conversation are legitimately provided as statements of truth. Thus, if a speaker says, 'it's raining', she or he has some evidentiary basis upon which to make that statement, and she or he does not intend the listener to understand that it is snowing. The maxims of *relation* and of *quantity* are relevant to the issue of common ground discussed below. The maxim of relation states that interlocutors are assumed to make their current utterances relevant by ensuring that current references relate to already known information. The maxim of quantity states that interlocutors are assumed to contribute an appropriate amount of information with each utterance, providing neither too little nor too much information (which cannot be done unless the speaker knows what the listener already knows). Finally, the maxim of *manner* states that interlocutors should endeavor to make their utterances unambiguous, brief, orderly, and that they should avoid obscurity. The maxim of manner can thus be taken to claim that speakers should make the form of their utterances easy for a listener to understand. Note that these maxims apply both to speakers and to listeners, in that speakers use the principle to guide utterance formation, and listeners use the principle to guide utterance interpretation. Furthermore, the cooperative principle can be flouted, or seemingly violated for the sake of communication (creating an implicature). For example, when asked how a tennis match was, if

a speaker says 'well, it was a tennis match', she or he seems to violate the maxim of quantity (by providing too little information), but thereby communicates that not much is to be said about the tennis match.

Grice's maxims are of two importantly distinct types. The maxims of quantity, quality, and relation are about what a speaker chooses to mention, whereas the maxim of manner is about how a speaker describes what they have decided to mention. This latter issue is discussed next, describing how speakers modify the form of their utterances to make those utterances easy to understand. Then, the issue of what a speaker chooses to mention is described, leading to the issue of the use of common ground.

MODIFICATIONS OF THE FORM OF SPOKEN UTTERANCES

Speakers cater the form of their utterances to their listeners in two general ways. One way involves explicit recognition by the speaker that their listener may encounter comprehension difficulty, typically because the listener may have limited linguistic abilities. Here, speakers are consciously aware of the potential for communicative difficulty, and they thereby make general modifications to their speech to circumvent that difficulty. The second way involves implicit production strategies that speakers use to create easy-to-understand sentences. Here, speakers are typically unaware of the difficulty their listener may have, but nevertheless tacitly make very specific modifications of their speech to enhance comprehensibility. Each of these kinds of modifications are discussed in turn.

When a speaker addresses a listener, the general style or register of his or her speech will be partially based on the nature of the relationship between the speaker and listener. Thus, for example, a speaker generally adopts a different register when addressing parents or grandparents, compared to when speaking to siblings. Students say 'yes, professor' to a university professor, but talk about 'my prof' to their friends.

While many of the features of a speech style reflect or communicate something about the social relationship between speaker and listener, other register effects may reflect a speaker's estimation of the linguistic capabilities of his or her listener. This is most strongly revealed by analyses of speech directed to populations with limited linguistic capabilities, including speech directed to children (child-directed speech or 'motherese'), to foreign-language speakers (for the purpose of communication or for language instruction), and to cognitively impaired listeners (e.g. the mentally retarded). When addressing populations like these, speakers (intentionally) include a host of features in their utterances that (intentionally or unintentionally) serve two related purposes: speakers modify their utterances so that they are more easily understood, and so that they are more likely to capture and hold their listener's attention. The features of utterances that accomplish these goals are listed in Table 1.

In addition to these general modifications that occur in response to perceived characteristics of a speaker's current listener, speakers make other, more specific modifications that occur automatically. Unlike the register effects just described, these more specific changes do not depend on the characteristics of a speaker's listener, and indeed occur in the absence of any listener at all.

For example, one bit of information that is useful for a listener to know is whether a mentioned entity has previously been mentioned in a discourse, or whether that entity is new to a discourse (this is part of the information structure of a discourse). Languages have a number of devices to mark this difference between given information and new information, including lexical, syntactic, and acoustic devices. Lexically, speakers can refer to given

Table 1. Features of a speech register that make it easier to understand and more likely to capture and maintain listeners' attention

Features that make speech easier to understand	*Features that make speech more attention-getting*
Shorter utterances	Greater use of names
More repetition of words and phrases	More questions
More rephrasals	Higher and more modulated pitch
More common words	Longer vowels
Simpler and more transparent syntactic structures	Longer pauses
Hyperarticulation	Greater rhythm

information with pronouns ('she', 'he', etc.) and by using definite articles ('the', 'this', etc.). Syntactically, speakers mark the difference between given and new information typically by word order, where speakers tend to mention given entities in their utterances prior to new entities. For example, if the sentence 'I saw Bill yesterday' sets 'Bill' up as given information, it is more natural to say 'he was disappointed by the election', with the given information first, compared to 'the election disappointed him', with the given information second. Finally, acoustically, speakers tend to diminish the auditory characteristics of words that refer to given information (affecting the prosody of those words), by reducing the duration, the loudness, and the pitch of those words. These cues to the given versus new status of information are used redundantly, in that the use of one kind of cue does not preclude the use of another (e.g. given information might be produced early in a sentence and in reduced fashion).

Speakers use other subtle signals to communicate useful information to their listeners. Some of these signals arise when speakers anticipate upcoming or recognize past production difficulty. For example, speakers use the familiar fillers 'uh' and 'um' differently, in that 'uh' is used when speakers anticipate a relatively short disruption to their own speech, whereas 'um' is used when speakers anticipate a longer disruption. Similarly, speakers tend to use 'thee' instead of 'the' when they are about to have difficulty retrieving the following word (Fox Tree and Clark, 1997). When speakers repair an erroneously produced part of an utterance, they typically say 'er' or 'I mean' to mark that a repair is to follow, and they reproduce enough of the original utterance so that the utterance plus repair are easily understood. Other constraints (termed *perceptual constraints*; see Cutler, 1987) include a tendency to speak with a consistent rhythm (which assists listeners in finding word boundaries), and a tendency to use easily recognizable forms when creating novel words (e.g. speakers prefer to say 'ambiguize' over 'ambigwify', because the former sounds more similar to the root form, 'ambiguous').

It is important to recognize that these specific modifications occur unintentionally and automatically. For example, the tendency to manipulate word order so that given information occurs before new information occurs in the complete absence of any listener, in a manner that appears to be sensitive to how easily retrieved an entity is from the speaker's memory. The tendency to diminish the acoustic characteristics of words referring to given information has been found when speakers address 14-month-old children (who, at this largely preconversational stage of linguistic development, are unlikely to usefully distinguish given from new information), and when speakers produce the first and the second mention of a word to different listeners. This is different from what occurs with the register effects discussed above, which are generally dependent upon the presence of a linguistically challenged interlocutor for the appropriate register features to appear correctly.

Different mechanisms are used by the language production system to automatically modify the forms of utterances to ease comprehension. In some cases, variation emerges naturally from the manner in which production operates, and then comprehension processes learn to infer the validity of the systematic cue that production provides. For example, there is nothing about positioning a word early in a sentence that inherently marks it as referring to given information; this systematic variation is likely to be a consequence of normal production that comprehension processes become sensitive to. In other cases, an easily comprehended utterance comes from a pressure to create easily produced sentences. For example, speakers will tend to produce a more easily understood common word instead of a less easily understood uncommon word, not because of the difference in comprehension ease *per se*, but because production of the more common word is easier than production of the less common word. Interestingly, a different kind of mechanism seems to underlie the adjustments speakers make to the content of their utterances based on listener knowledge. This is described next.

MODIFICATIONS OF THE CONTENT OF SPOKEN UTTERANCES

Speakers cater the content of their utterances to the comprehension strategies and the knowledge states of their listeners. Interestingly, the role that the actual listener plays in the nature of these listener-sensitive modifications varies; sometimes, the modifications are quite independent of the actual knowledge states of listeners, whereas at other times, speakers specifically take into account what their listeners know. This variability in how listeners' knowledge is accounted for during production seems to reflect a general mechanism that people use when they estimate the knowledge states of others.

For example, many factors affect what kinds of labels speakers use for reference (e.g. referring to a tree outside as 'the tree on the left'). One factor is

that when possible, a speaker will collaborate with their listener to arrive at a suitable label for successful reference, by proposing a label that the listener can approve of, or suggest modifications to. Another important factor is the prior discourse history between a speaker and listener (Brennan and Clark, 1996). If, for example, a speaker has previously referred to 'the creepy tree' with a particular listener, they can use that label again with that listener without renegotiating from scratch. Interestingly, an egocentric bias arises when speakers use their prior discourse history. Whereas it is the case that with a new listener, speakers are relatively less likely to use a label that was used previously with a different listener, the labels that speakers do use with new listeners are nevertheless more similar to the labels that they've previously used (with different listeners), compared to labels that speakers use for the first time. The egocentric bias is revealed by the fact that a speaker's own discourse history, independent of his or her listener's history, influences the nature of the labels that speakers use.

Another example demonstrates how the egocentric bias operates. It has been shown that in story retellings, speakers are more likely to mention atypical instruments of actions (e.g. an icepick for a stabbing event) than typical instruments of actions (a knife for a stabbing event; see Dell and Brown, 1991). This tendency is an elegant reflection of Grice's maxim of quantity: if an idea to be conveyed includes an action that is performed with a highly typical instrument, then the instrument need not be explicitly described, as the listener is likely to infer that particular instrument if necessary. On the other hand, if the action is performed with an atypical instrument, the instrument should be explicitly described, because a listener cannot infer the existence of an atypical instrument. Interestingly, an egocentric bias still arises, in that it has been shown that under some circumstances, whether the speaker knows the listener to already possess knowledge of the instrument does not affect the tendency to especially mention atypical instruments. This is only true, however, with sentences that describe the action itself. So, for example, speakers are as likely to say 'The robber stabbed the victim with the icepick' whether or not their listeners know about the icepick. On the other hand, in subsequent sentences (after the action has been described), speakers are sensitive to listener knowledge. So, a speaker is less likely to say 'Oh, and there was an icepick' when a listener already knows about the icepick, compared to when the listener does not.

These observations hint at a mechanism that turns out to be common in social-cognitive functioning generally (see Nickerson, 1999), as well as in processing related to the production–comprehension interface specifically. The fact that speakers' initial references to an action include mention of atypical instruments equally often regardless of listener knowledge suggests that speakers' initial production decisions are based on egocentric information. That is, utterances are first formulated based on what the speaker knows, relatively independently of what listeners know. This can account for the general egocentric bias that arises in production. Then, the observation that subsequent references to the instrument are sensitive to listener knowledge suggests that after the initial production, speakers are better able to accommodate the differences between what the speaker knows and what their listeners know. Overall, it appears that speakers accommodate the content of their utterances to the knowledge states of listeners with a general two-stage strategy: speaker-centered, egocentric knowledge guides initial utterance formulation, and then listener-specific knowledge is taken into account in subsequent processing.

Why might the production system adopt such a two-stage strategy? For reasons of processing efficiency. In terms of processing resources, it is costly to keep track of the knowledge states of all listeners, with respect to all possible facts that a speaker could express. A reasonable heuristic is that if the speaker knows some fact, his or her listeners are also likely to know that fact. Based on this egocentric strategy, production processes can efficiently formulate an initial utterance that is likely to be good enough for successful communication. However, once the initial utterance is formulated, the specific information expressed in the utterance can more easily be evaluated with respect to listener knowledge. So, the initial formulation can be followed by elaborations and modifications that take into account listener knowledge. Interestingly, if production occurs with relatively little pressure, it is possible that this formulation-then-adjustment strategy could occur prior to articulation. Indeed, this kind of mechanism seems to reflect how common ground information is processed, which is described next.

COMMON GROUND EFFECTS

For a speaker to make definite reference – for example, as occurs with the use of a noun phrase with the definite determiner 'the' – she or he must assume the truth of a remarkable number of facts.

First, the speaker must know about the entity to which she or he wishes to refer. Second, the speaker must know that the listener knows about this entity. And third, the speaker must think that both the speaker him- or herself as well as the listener know about the other's knowledge of this entity. For example, Mary cannot say to Bill 'the guy you voted for is a clown' unless: (a) Bill knows who Bill voted for; (b) Mary knows who Bill voted for; and (c) Mary thinks that both of them know of the other's knowledge of who Bill voted for. Shared knowledge is not sufficient for definite reference, in that it cannot merely be the case that each independently knows of Bill's vote. There must be mutual knowledge, in that both knows of the other's knowledge. If Bill doesn't know that Mary knows of his voting choice, then Mary's statement can fairly be responded to with 'how do you know who I voted for?' It is this mutual knowledge that a set of interlocutors possesses that is termed *common ground*.

THE NATURE OF COMMON GROUND

Without some semblance of the common ground between speaker and listener already in place, a conversation can hardly get off the ground. Many a science fiction story portrays two members of mutually unknown species, struggling futilely to communicate without any shared basis of communication, until that shared basis is discovered, at which point the journey of information exchange begins. The same occurs any time two strangers meet, if less dramatically. In fact, 'current weather conditions' as the reliable kindling of any conversation can be understood in this light: the current weather conditions are encompassing enough that any co-present interlocutors can assume them as a shared basis of conversation (i.e. that it's likely to be common ground), though they are changing and important enough that they are (more or less) worth talking about (thus permitting speakers to respect Grice's maxim of quantity).

Though common ground begins as a necessary starting point for a conversation, it then continues to build across that conversation, and indeed across the continuing interactions between any given interlocutors. Common ground refers not only to what interlocutors can talk about, but how they can talk about it. For example, based on common ground, a speaker will decide whether to speak in English, French, or American Sign Language, and in those languages, what register to use. Research has identified two different bases that jointly contribute to these many different aspects of the common ground that is shared between interlocutors (Clark, 1996). Communal common ground refers to the abilities and knowledge that interlocutors can assume to be possessed by virtue of the community membership of the participating interlocutors. Thus, for example, if a speaker is aware that his or her interlocutor is a French Canadian, an accountant, a bridge player, or a parent, she or he will assume mutual knowledge of a set of facts that members of each of those communities will typically know. Personal common ground refers to the set of facts that interlocutors come to know as a consequence of their joint experiences, including things that happen within a conversation (the entities that speakers refer to) as well as those outside the conversation (events that interlocutors can be assumed to have noticed). All facts learned in past conversations or mutual experiences among a set of interlocutors become parts of those interlocutors' common ground.

PROCESSING CONSIDERATIONS

Speakers clearly recognize information that is in common ground when they speak. This is evident not only from the fact that speakers' definite references overwhelmingly succeed in conversation, but also by more subtle audience design effects (e.g. Krauss and Fussell, 1991), where speakers seem to cater their utterances to account for what their listeners are likely to know already (i.e. for what is in common ground). For example, speakers will refer to local landmarks or provide navigation directions differently if they think their listeners are local residents, compared to if they think the listeners are not local. Speakers will also produce utterances differently if they think their speech is for themselves (i.e. for the purpose of future reference) compared to if the speech is for others, or if speech is intended for a personal friend compared to if it is intended for a stranger. Furthermore, research has shown that communication is more successful when speakers correctly know who their intended audience is.

How speakers recognize common ground is not a trivial issue, however. Because language users do not have direct access to the cognitive states of others, each must infer whether a particular fact is mutually known. There are many specific strategies that interlocutors can use to infer whether a fact is in common ground. For example, if interlocutors notice that each has made eye contact with a particular object in the environment, they may assume that object to be in common ground. Or, if one interlocutor verbally refers to an entity, all

interlocutors who could reasonably be assumed to have heard the verbal reference can have that entity ascribed to their common ground. Interestingly, the egocentric bias described above arises too when speakers attempt to account for common ground. With respect to communal common ground, it has been found that while speakers can quite accurately estimate what other members of specified communities do and do not know, they still exhibit a systematic bias to overestimate the likelihood that others know things that they themselves know. This effect, a version of the 'false consensus effect' from social psychology, suggests that with respect to common ground, an interlocutor who idiosyncratically knows a fact is relatively likely to ascribe knowledge of that fact to others, and therefore assign that fact to common ground.

Research investigating the use of personal common ground has found evidence not only of an egocentric bias, but also for the two-stage model described above. If speakers formulate utterances freely, without any time pressure, they can easily provide enough information to refer successfully to an intended concept in its context (see Deutsch and Pechmann, 1982), and can distinguish information in common ground from privileged information (information that only the speaker knows about). However, if speakers formulate utterances under time pressure, the distinction between common ground and privileged information erodes, as speakers are more likely to make (infelicitous) references to privileged information in utterances that are directed to their listeners. Thus, initial formulations – ones that are used when time pressure does not permit subsequent revision – are based on egocentric information, until subsequent modifications of the initial formulation can correct for the different knowledge states that speakers and listeners might have (see Horton and Keysar, 1996).

Overall, then, for communication to succeed, speakers must take into account the linguistic capabilities and the knowledge states of their listeners. Grice's cooperative principle provides an ideal to which speakers and listeners can appeal with each contribution to a conversation. Given this ideal, the extent to which speakers follow the cooperative principle is impressive. Although speakers show some general deviations from producing language in an entirely cooperative manner, these can largely be understood as the result of speakers' use of fallible heuristics – heuristics that are necessary given the pressures involved in timely production, as well as the intractability of the demands involved in estimating mutual knowledge. Despite these difficulties, speakers modify the form and the content of their utterances so that their listeners can more easily understand those utterances. The listener-sensitive accommodations that speakers make during language production are a revealing compromise between what speakers know, and what they guess others know.

References

Brennan SE and Clark HH (1996) Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition* **22**: 1482–1493.

Clark HH (1996) *Using Language*. Cambridge, UK: Cambridge University Press.

Cutler A (1987) Speaking for listening. In: Allport A, MacKay DG, Prinz W and Scheerer E (eds) *Language Perception and Production: Relationships between Listening, Speaking, Reading, and Writing*, pp. 23–40. London: Academic Press.

Dell GS and Brown PM (1991) Mechanisms for listener-adaptation in language production: limiting the role of the 'model of the listener'. In: Napoli DJ and Kegl JA (eds) *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman*, pp. 105–129. Hillsdale, NJ: Lawrence Erlbaum Associates.

Deutsch W and Pechmann T (1982) Social interaction and the development of definite descriptions. *Cognition* **11**: 159–184.

Fox Tree JE and Clark HH (1997) Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* **62**: 151–167.

Grice HP (1975) Logic and conversation. In: Cole P and Morgan JL (eds) *Syntax and Semantics*, vol. 3: *Speech Acts*, pp. 41–58. New York, NY: Academic Press.

Horton WS and Keysar B (1996) When do speakers take into account common ground? *Cognition* **59**: 91–117.

Krauss RM and Fussell SR (1991) Perspective-taking in communication: representations of others' knowledge in reference. *Social Cognition* **9**: 2–24.

Nickerson RS (1999) How we know – and sometimes misjudge – what others know: imputing one's own knowledge to others. *Psychological Bulletin* **125**: 737–759.

Further Reading

Bock K (1995) Sentence production: from mind to mouth. In: Miller JL and Eimas PD (eds) *Handbook of Perception and Cognition*, vol. II: *Speech, Language, and Communication*, pp. 181–216. San Diego, CA: Academic Press.

Brown PM and Dell GS (1987) Adapting production to comprehension: the explicit mention of instruments. *Cognitive Psychology* **19**: 441–472.

Clark HH and Clark EV (1977) *Psychology and Language: An Introduction to Psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.

Clark HH and Marshall CR (1981) Definite reference and mutual knowledge. In: Joshki AK, Webber BL and Sag IA (eds) *Elements of Discourse Understanding*,

pp. 10–63. Cambridge, UK: Cambridge University Press.

Ferreira VS and Dell GS (2000) Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* **40**(4): 296–340.

Levelt WJM (1983) Monitoring and self-repair in speech. *Cognition* **14**: 41–104.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

MacKay DG (1987) *The Organization of Perception and Action: A Theory for Language and other Cognitive Skills*. New York, NY: Springer-Verlag.

Newport EL, Gleitman H and Gleitman LR (1977) 'Mother, I'd rather do it myself': some effects and non-effects of maternal speech style. In: Snow CE and Ferguson CA (eds) *Talking to Children: Language Input and Language Acquisition*. Cambridge, UK: Cambridge University Press.

Prosody

Introductory article

Fernanda Ferreira, Michigan State University, East Lansing, Michigan, USA

CONTENTS

Every utterance is produced with variations in word duration, pitch, and stress. These features constitute the prosody of language and provide important information about meaning and structure.

INTRODUCTION

Any spoken utterance is produced with a particular sound pattern. For example, in English declarative sentences, pitch gradually falls, whereas pitch rises at the end of an interrogative or question. Words such as 'of' and 'the' tend to be short in duration, but words at the ends of clauses are usually stretched. In the default case, the word at the end of a clause also receives the greatest stress, but emphatic stress can be shifted to some other word, indicating that it is semantically prominent. These variations in pitch, duration, and stress provide an utterance with its prosody. Regardless of whether the utterance is a list of grocery items or a complex multiclause sentence uttered as part of a conversation, it will be assigned a prosodic pattern. This generalization holds for all human languages. Indeed, the prosodic patterns of a language are among the first linguistic abilities infants acquire. As anyone who has listened to the meaningless babble of a baby can appreciate, the utterances of even six-month-olds are gibberish but to a remarkable degree the prosody is similar to what would be appropriate for a real sentence of the child's language.

Prosody, then, is a general term that refers to the aspects of an utterance's sound that are not specific to the words themselves. For example, consider the sentences 'Bill likes to run' and 'Jane quit her job'. The two are made up of entirely different words, and so they are completely different at the level of their *phonemic content* (the individual sounds that make up words are called phonemes). And yet the sentences would be spoken with an almost identical prosody: pitch would gradually fall, giving both an intonational pattern that is sometimes called the 'declarative contour'; 'to' and 'her' would both be clipped, but 'run' and 'job' would be lengthened considerably; and the major stress of the sentence would fall at the end, on 'run' and 'job'. Prosody can be divided into three main aspects: intonation, which has to do with variations in pitch; timing, which concerns the durations of words and the locations and durations of any pauses between them; and stress, which roughly has to do with the loudness of the various words (more precisely, particular syllables within the words).

Prosody is an important tool that speakers use to define the informational structure of their utterances – the separation between information that links up with the ongoing discourse and information that is new and therefore being added to interlocutors' common ground. The latter is often referred to as the 'focus' of the utterance, and focused elements tend to be those that are prosodically prominent.

An important question in cognitive science is: how do speakers produce these prosodic patterns? That is a question about the creation of prosody during language production, a topic that will be discussed at the end of this article.

INTONATION

An 'intonational phrase' can be thought of as a series of words that fall into the same pitch group. For example, the sentence 'Bill wants to walk but Jane prefers to take a taxi' would typically be produced as two intonational phrases: [Bill wants to walk] and [but Jane prefers to take a taxi]. A single intonational phrase normally consists of one accented syllable ('Bill', for example) and a series of unaccented syllables ('wants to walk'). The accented word is said to receive a 'pitch accent', and the overall pitch movement of the phrase is

said to create its 'tune'. The tunes of the two phrases in the above example are not the same, however. The first one falls but ends with a rising tone that indicates that the utterance will continue; the second one ends with a falling tone, which is characteristic of pitch at the ends of complete utterances. The last syllable of an intonational phrase will tend to be stretched and might also be followed by a pause. In addition, the first syllable of the new intonational phrase will begin at a higher pitch. This pitch resetting allows the speaker to produce the gradually falling intonation pattern which is characteristic of declarative utterances without falling outside his or her pitch range.

Different tunes are associated with particular sentence meanings. The distinction between a statement and a question has already been mentioned. Even statements of various types may have distinct intonational patterns. If someone asks, 'Why are you going to Montreal?', one might reply 'My brother goes to McGill'. But if that person asked, 'Are you excited about visiting Concordia when you're in Montreal?', the same reply might be given but the tune would be different – pitch would be flatter until the word 'McGill', at which point pitch would rise quite sharply. The latter tune would signal that the speaker is negating the other conversational participant's presupposition about the university in Montreal with which she has an association.

The formal system that is used to characterize intonation is referred to as *ToBI*. This term stands for 'tones and break indices', and it is a formalism used to represent different tunes. The fundamental idea that underlies the system is that all tunes comprise just two basic tones: High (H) and Low (L). The most accented syllable in an intonational phrase – the syllable that receives a pitch accent – is symbolized as H*. The syllable that ends the intonational phrase is represented with a %, as in L%, and is referred to as the boundary tone. Figure 1 shows the pitch contour for a declarative sentence ('It was the terrorist that kidnapped the ambassador'), along with simplified ToBI transcription and corresponding waveform.

Intonation, then, concerns the tune or melody of an utterance and is the basis for its division into prosodic phrases. The next section describes another aspect of prosody: the durations of words and pauses, which make up an utterance's timing.

TIMING

Consider the utterance 'Bill wants to walk but Mary wants to drive' and compare it to 'Bill wants to walk to the store'. *Walk* would have a longer duration in the first example than in the second. This difference is based on the phrasal position of the word 'walk' in the two cases: when it occurs at the boundary of a phrase or clause, it tends to be

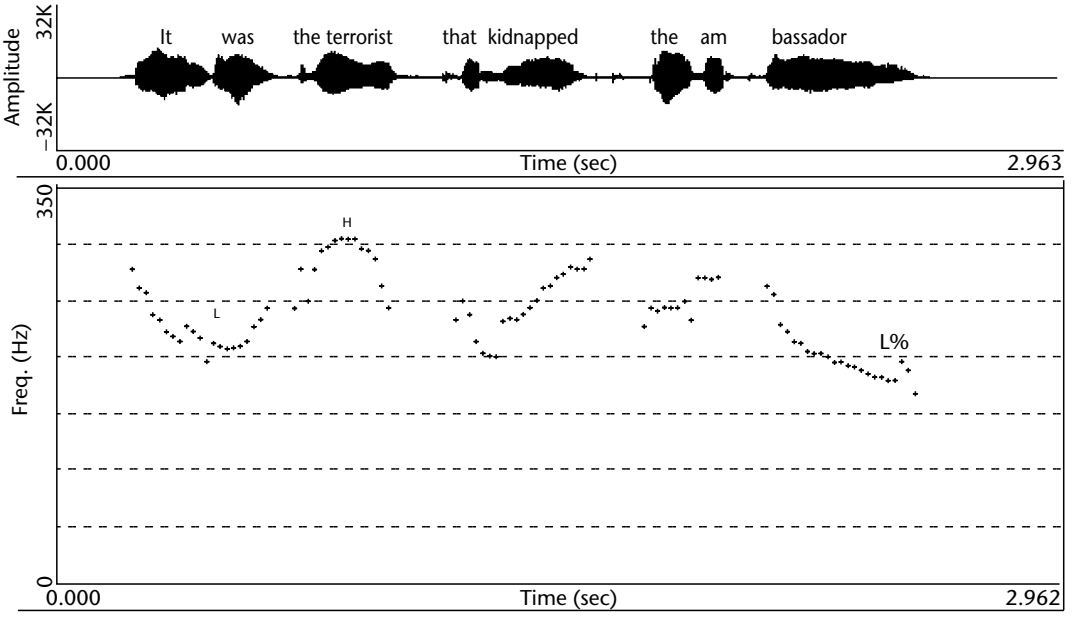


Figure 1. Acoustic representation of a spoken sentence. The top portion is a waveform, which displays amplitude variations as a function of time. The bottom portion is a representation of changes in fundamental frequency (pitch) as a function of time. The bottom portion also includes ToBI annotations.

lengthened to a greater extent. This phenomenon is known as 'phrase-final lengthening'.

An important consequence of phrase-final lengthening is that important syntactic boundaries will tend to be marked in normal speech. This cue might be used by young children as they learn about the grammatical properties of sentences. This cue is also useful for helping the listener resolve certain ambiguities in normal language. For instance, the sentence 'Mary saw the man with binoculars' could mean either that Mary used binoculars as a seeing instrument, or that the man she saw was in possession of binoculars and she saw that particular man. The spoken version of this sentence might distinguish between these two readings prosodically – if the speaker intends to convey that Mary has the binoculars, then 'man' will be at the end of a phrase (the object of the verb) and so the word would be lengthened. If the speaker intended the other meaning, then 'man' would not be lengthened, because in that case it does not occur at the end of a phrasal boundary. Lengthening, then, serves as a cue that can be used both by children acquiring language and by competent adults trying to understand utterances.

Another important feature of the timing of utterances is that so-called 'function words' are almost always short. Function words are grammatical words such as 'of', 'the', 'and', 'on', and so on. The term 'function word' is meant to highlight the fact that such words tend to be important for the grammatical properties of a sentence. In contrast, a 'content word' conveys most of the meaning of a sentence, and these overall tend to be longer than function words. For example, we can directly compare 'I want two leaves' and 'I want to leave'. The words 'two' and 'to' are homophones, but 'two' is a content word and 'to' is a function word. The latter has a noticeably shorter duration than the former in a normal sentential context, and that difference is directly attributable to the word's status as a function word.

Of course, function words *can* be made long under the right discourse circumstances. One factor that may cause a function word that would normally have a short duration to be lengthened is emphatic stress. For example, one would normally say 'Put the book on the table' so that the word 'on' was quite clipped. But if one were in a situation in which one wanted to emphasize that the book should go on and not under the table, then the word 'on' would be stressed and, consequently, also lengthened. A second factor that can cause a function word to be lengthened is its position within a phrase. Recall the phenomenon of phrase-final lengthening. In a sentence such as 'Who do you want to talk to?', the final word 'to' would be much longer than it would be in 'Do you want to talk to Tom?' To execute this lengthening, it is necessary to produce the word 'to' in its citation form; that is, the vowel in 'to' that would normally be reduced to a schwa (the default lax, unstressed vowel sound found, for example, in the second syllable of 'other') instead has its full, long pronunciation.

Pauses also support utterances' characteristic timing patterns. People often pause because they are having trouble formulating their utterances, but these are not the sorts of pauses that are relevant to timing. Consider that even in flawless renditions of text, it is normal to pause at phrase and especially at clause boundaries. These pauses that tend to cluster at syntactic boundaries are correlated with phrase-final lengthening; words that are lengthened often tend to be followed by a pause. The phenomenon is similar to that of a rest in music: in speech, people insert a brief pause in order to preserve the overall rhythm of the utterance's timing pattern, just as a musician uses a rest to maintain appropriate metre. These pauses are generally quite short – between a quarter-second and a half-second in duration.

STRESS

Words differ in the amount of stress (corresponding closely to loudness) that is normally assigned to their constituent syllables. Each of the syllables of a word such as 'chimpanzee' receives a different amount of stress: *-zee* gets the greatest stress, *-pan-* the smallest, and *chimp*-is assigned a stress level somewhere in between. In addition, when words occur in multiword contexts, their stress patterns are affected. For example, the word at the end of a phrase tends to be stressed the most, as in 'Mary wants a chimpanzee'.

One formalism for representing the different levels of stress that each syllable receives is to use a *metrical grid* that symbolizes the metre of the utterance. The sentence above would be represented as shown in Figure 2. A column of Xs occurs above each syllable, and the height of the column indicates the level of stress. The function word 'a', for example, gets only level 1 stress; the final syllable of 'chimpanzee', in contrast, goes all the way up to level 5, making it the most prominent syllable of the utterance (the most prominent syllable of the word is also the one that becomes the most prominent syllable of the utterance). This syllable's standing is a direct consequence of the placement

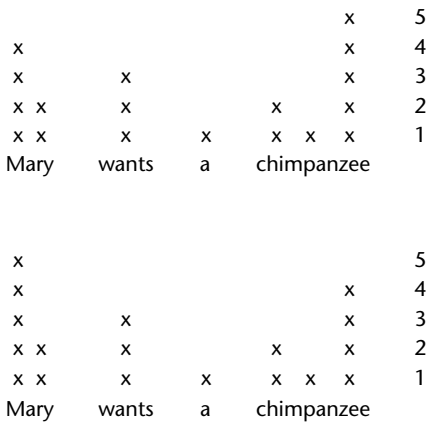


Figure 2. Metrical grids for two prosodic versions of the same utterance. The first would be a neutral rendition, and the second would be a rendition in which 'Mary' was focused.

of the word 'chimpanzee' as the last word of the utterance.

It is possible, of course, to make a different word in the utterance the most prominent – in our example, 'Mary' could be the informational focus. This rendition might be appropriate in a context in which someone mistakenly attributed the desire for a chimpanzee to Susan rather than Mary. The grid that would result for an utterance in which 'Mary' is most prominent is also shown in Figure 2. Notice that it is now 'Mary' that receives level 5 stress rather than 'chimpanzee'.

There is a tight link between stress and duration; the two features together create an utterance's timing pattern. Words that are stressed tend to be longer; words that receive little stress tend to be short. Indeed, notice that the default rule of utterance stress which reserves prominence for the last word operates on the same word that is usually the target of phrase-final lengthening. This correlation is not accidental; as mentioned earlier, stressed words are louder, and it is almost inevitable that a word which is produced so that it is louder will also tend to be made longer. This correlation can be verified through acoustic measurements. Any spoken sentence can be *digitized* so that, for example, amplitude values are sampled at various intervals. The result is a waveform – a representation that plots amplitude against time (see Figure 1). Waveform editors permit speech scientists to measure the durations of syllables exactly as well as their corresponding amplitude levels (the psychological experience of amplitude is loudness). If the word 'Mary' in 'Mary wants a chimpanzee' were the informational focus of the utterance, then it would not only be the site of higher amplitude values, but it would also tend to be longer than it would be in an utterance in which it was not the focus.

FOCUS

Up to now, the concept of focus has been invoked in various contexts but it has not been discussed in any detail. Focus is a somewhat complex concept, because it has both semantic and phonological aspects. Fundamentally, focus is a semantic notion; but the way focus is implemented in utterances often involves the manipulation of sound features such as stress, timing, and intonation.

To begin to appreciate what focus is, it is necessary to consider utterances in discourse contexts. One very common type of discourse is a conversation between two participants consisting of a question and corresponding answer. For example, consider the question 'Who does Mary admire?' In reply someone might say 'Mary admires the man who won the marathon'. The reply has a particular *information structure*. The presupposition that Mary admires someone is part of the background information shared by the speaker and hearer. The utterance then adds to the ongoing discourse the information that Mary admires the man who won the marathon. Another way to view this division is as follows: the question sets up the structure 'Mary admires x'; what the other participant in the conversation then does is to provide a value for x. The contents of x are the focus of the utterance.

So far we have established that sentences have information structures – they can be divided into background and focus. This issue pertains to the meaning or semantics of language. The next question that can be asked is how speakers signal this informational structure.

The example above makes clear that, in principle and for some types of discourses, nothing beyond the content of the question–answer pair is necessary, because a question may set up the background and the remainder would then be the focus. But often discourses do not have such a tidy organization. For example, person A might say 'Mary admires body-builders.' Person B might know this statement to be false, and furthermore, he might know that Mary actually admires endurance runners. Therefore, person B might reply, 'No, Mary admires MARATHONERS.' Following standard conventions, the focused element has been capitalized to indicate that the word is spoken with a pitch accent (recall that pitch-accented words receive heavy stress and tend to be lengthened).

Speakers also have other devices at their disposal for indicating focus; for instance, the speaker could have used a so-called 'cleft construction', a particular type of syntactic form, to focus the appropriate element. Thus, person A might say 'Mary admires body-builders,' and person B might reply 'No, it's marathoners that Mary admires.' Focus can be signaled by the use of a particular type of grammatical structure, or it can be indicated solely with sound, via a pitch accent. But interestingly, even when a syntactic device for focusing is used, pitch accenting seems to occur as well. This point can easily be appreciated by saying the cleft sentence above out loud: a speaker almost invariably ends up placing phonological emphasis on the word 'marathoners' in the structure 'It's marathoners that Mary admires.' Thus, it appears that even when a speaker in principle has already taken care of the need to signal the informational structure of her utterance by choosing a syntactic form tailor-made for that purpose, she still adds the apparently redundant phonological cues associated with pitch accents as well.

Why would speakers behave this way? Why would they signal focus with a pitch accent when the syntactic form of an utterance adequately conveys the information structure? Indeed, as was observed above, in some question–answer pairs simply the content of the question is sufficient to parse the utterance that follows into background and focus – neither a specialized syntactic form nor a pitch accent is necessary. Moreover, experimental studies of people's understanding of sentences have shown that the same sentence is perceived as having a different informational structure depending on the question that precedes it, even when the sentence itself is identical in the different contexts. Thus, if someone hears the question 'Who is wearing a blue hat?' they perceive the sentence 'The man on the corner is wearing a blue hat' as having the following focus–background structure: the background is 'X is wearing a blue hat', and the focus is X = the man on the corner. But when the exact same sentence spoken identically is preceded by the question 'What is the man on the corner wearing?' then the sentence is informationally parsed as follows: the background is 'The man on the corner is wearing X', and the focus is X = a blue hat.

But even though it has been shown that people understanding sentences *can* divide a sentence into background and focus without the aid of prosody, it is also clear that this task is easier if the focus is signaled with a pitch accent. It is possible, then, to design a psycholinguistic experiment to show that a sentence whose focus–background structure is not explicitly signaled is still understood as having one; but the reality of natural language production is that speakers tend to provide the phonological cues anyway, and listeners take advantage of those cues to locate the focus more efficiently. Recent research has even shown that focus can help to clear up ambiguity within sentences. For example, consider 'Everyone smiled at the son of the driver who had the beautiful, bushy moustache.' It is either the son or the driver who has a striking moustache; the syntactic structure of the sentence is not sufficient to specify unambiguously which one. But if the speaker places a pitch accent on the word 'son', then listeners tend to think he is the one with the moustache; if the pitch accent instead occurs on 'driver', then people interpret the sentence to mean that the driver has a moustache. These results demonstrate that not only is phonological signaling of focus helpful, it can even help the comprehender to resolve ambiguities in interpretation.

CREATION IN PRODUCTION

Most of the topics that have been discussed to this point concern linguistic representations having to do with prosody. Linguists have proposed various formal systems for symbolizing aspects of prosody, including intonation, timing, and stress. The issue that will be considered in this final section can be described as the psychological question: how do people give their utterances the prosodic features that have been detailed?

This question must be considered in the context of the entire enterprise in which the language production system is engaged. Briefly, to produce an utterance, the speaker must make a number of decisions. First, she must decide on the idea to be conveyed. This stage of 'message-level processing' requires the speaker to determine what it is she wishes to talk about and what she intends to say about the entity or entities. The speaker must choose words that appropriately express that meaning. The words must be organized into a grammatical form that places the words in an order that other speakers of the language will recognize as legitimate and that allow her semantic intention to be appreciated. For instance, if the speaker wishes to convey that it is raining outside at the moment, she could say 'It is raining.' The rules of English commit her to organizing the sentence in a particular manner; she could not express the same idea with the utterance 'Raining it is,' nor could she choose to omit the apparently

meaningless word 'it' (other languages, for example Romance languages, do allow such elements to be omitted). The sounds that make up the words must also be selected. Eventually, orders must be sent to the speech articulators to move the mouth, lips, tongue, and so on in the right ways so as to make the appropriate speech sounds.

One may ask, then, at what point in this general sequence does the prosodic form of a sentence get decided? Recent work has suggested that prosody is not spelled out in just a single stage during processing; instead, decisions about the prosodic features of sentences are made almost throughout the entire information-processing sequence, and then get implemented when the utterance is actually pronounced. The decision to focus a word or phrase is a high-level decision that typically is made during message-level processing. The speaker focuses a particular concept because semantically he takes certain information to be shared background between him and his interlocutor, and he wishes to add to their common ground – that is, he wishes to add new information. Decisions about intonation are also, to some extent, semantically based. For example, normally one would utter the sentence 'Billy slammed the ball' as a single intonational unit; but a sportscaster who wishes to convey something about the magnitude of the event might choose to divide the utterance into more than one intonational phrase as follows: '(Billy) (slammed) (the ball).' The effect the speaker wishes to create determines the choice, and so this decision too is largely based on semantic considerations.

The decisions that concern timing are not so much semantic as they are a function of the grammatical form of the sentence. If a speaker chooses to say 'Billy slammed the ball', then when a metrical grid for this utterance is set up, it will have to accept the words in those particular positions. As a result (in the simplest case), the sentence would be spoken as a single intonational phrase and the word 'ball' would be lengthened. In contrast, if the speaker had decided that the passive voice was required to convey his meaning properly, then the sentence would be 'The ball was slammed by Billy.' The processes that create the metrical grid would create a representation in which 'ball' had a much shorter duration, and 'Billy' would be the locus of phrase-final lengthening.

Another question related to the creation of prosody is what determines whether a speaker simply extends a word's duration or also chooses to insert a pause. The basis for this choice is largely not under a speaker's conscious control. A word's position in the metrical grid determines the amount of time that each word (more technically, each syllable of each word) should occupy in the utterance. If a word that is subject to phrase-final lengthening is made up of phonemes that are fairly stretchable, then a pause might not be necessary. But if the grid mandates an amount of time that cannot be filled out because the word is less stretchable, then a pause will be used to fill up the time that remains. Thus, consider the sentences 'The table that is black goes in the hall' versus 'The table that is green goes in the hall.' The words 'black' and 'green' both occur in a phrase-final position (each is the last word of the sentence's subject) and so they will be lengthened (compared to the durations they might have in 'The green table goes in the hall' and 'The black table goes in the hall'). The word 'green' is much more stretchable than is the word 'black', and so 'green' is much less likely to be followed by a pause. The logic is that if a word is not stretchable then the interval mandated by the grid cannot be filled with syllable lengthening. A pause must then be used to fill up the time that remains.

CONCLUSION

Prosody refers to the regular sound characteristics associated with spoken, multiword utterances. The primary components of prosody are variations in stress, duration, and pitch. The focus of a sentence is the new information that a speaker adds to an ongoing conversation with his or her linguistic contribution. When people speak, they make prosodic choices at virtually every stage of processing: semantic, syntactic, and word-level. These prosodic decisions then get implemented at the stage at which the speaker sends motor commands to the speech articulators.

Further Reading

Bing JM (1985) *Aspects of English Prosody*. New York, NY: Garland.

Bolinger D (1986) *Intonation and its Parts: Melody in Spoken English*. Stanford, CA: Stanford University Press.

Cutler A, Dahan D and van Donselaar W (1997) Prosody in the comprehension of spoken language: a literature review. *Language and Speech* **40**: 141–201.

Ferreira F and Anes MD (1994) Why study spoken language processing? In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Ladd RD (1996) *Intonational Phonology*. New York, NY: Cambridge University Press.

Morgan JL and Demuth K (1996) *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Pierrehumbert J and Hirschberg J (1990) The meaning of intonational contours and the interpretation of discourse. In: Cohen PR, Morgan J and Pollack ME (eds) *Intentions in Communication*, pp. 271–312. Cambridge, MA: MIT Press.

Selkirk EO (1984) *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Warren P (1996) *Prosody and Parsing*. Hove, UK: Psychology Press.

Warren P (1999) Prosody and language processing. In: Garrod S and Pickering M (eds) *Language Processing*, pp. 155–188. Hove, UK: Psychology Press.

Zubizarreta ML (1998) *Prosody, Focus, and Word Order*. Cambridge, MA: MIT Press.

Psycholinguistics, Computational

Advanced article

Richard L Lewis, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

Computational psycholinguistics seeks to build theories of human linguistic processes that take the form of working computational models. These models address processes ranging from word recognition to discourse comprehension, and produce behavior that constitutes predictions to be compared to human data.

INTRODUCTION

Computational psycholinguistics seeks to build theories of human linguistic processes that take the form of implemented computational models. These models are intended to explain how some psycholinguistic function is accomplished by a set of primitive computational processes. The models perform a psycholinguistic task and produce behavior that can be interpreted as a set of predictions to be compared to human data. As such, computational psycholinguistics is a paradigmatic example of cognitive modeling more generally. One problem with the label *computational psycholinguistics* is the implication that there is something that can be identified as *noncomputational* psycholinguistics. This is not presently the case: all psycholinguistic theories are, at some level, assertions about computational processes. Computational psycholinguistics is distinguished from other forms of cognitive modeling by its domain (not its techniques), and it is distinguished from other forms of psycholinguistic theorizing by its focus on producing functioning computational mechanisms that embody an explicit process model. The remainder of this article is devoted to reviewing the state of computational modeling in several of the major subfields of psycholinguistics.

MODELS OF LEXICAL PROCESSING

The most influential computational models in psycholinguistics have been those focused on word-level processes, in particular, spoken and visual word recognition. In fact, there are currently no major psycholinguistic theories of word recognition that do not take the form of a computational model. Competing theories are routinely tested by running the corresponding computational models to determine how well the models' behavior fits human data. At some level, there is significant theoretical convergence. All of the models of lexical processing are activation-based: lexical access is modeled as a dynamic process of modulating the activation of patterns of representation that encode information associated with specific lexical (or morphological) items. However, the models differ dramatically along many important architectural dimensions, such as the degree of top-down feedback and the nature of the computational principles determining the dynamic activation patterns.

Spoken Word Recognition

Models of spoken word recognition must satisfy a number of challenging functional and empirical constraints. These include: speech occurs in time, with no clear boundaries between words or phonemes, which may in fact overlap; there are effects of both left and right context on word recognition; lower-level phoneme identification may depend on higher-level lexical information; and there may be considerable noise in the environment (McClelland and Elman, 1986).

Current computational models of word recognition are extensions of ideas first put forward explicitly in the COHORT theory of speech perception (Marslen-Wilson and Tyler, 1980). The key principles in COHORT are that the initial sound of a word establishes a *cohort* or candidate set of possible words beginning with that sound, and this candidate set is incrementally narrowed down in real time as subsequent acoustic input arrives. Word recognition is achieved when the candidate set is narrowed to one, which may occur before the end of the word.

The TRACE model of McClelland and Elman (1986) provides an explicit computational realization of these basic ideas in COHORT, while addressing some of its most critical shortcomings. In particular, COHORT had no clear account of how word boundaries were identified in the continuous speech stream, and it assumed accurate bottom-up identification of phonemes. TRACE is an interactive-activation architecture with bidirectional excitatory connections between nodes representing acoustic features, phonemes, and words. Each time slice of input occupies a separate part of the input vector, and there are multiple copies of phoneme and word detectors centered over every three time slices. There are also inhibitory links within levels between mutually incompatible words or phonemes; thus, word and phoneme recognition is a competitive process. This competition and the distribution of multiple detectors across the network permits the model to recognize words without clear boundaries known in advance. The bidirectional nature of the within-level connections provides a way for the lexicon to directly influence the perception of lower-level phonemic and acoustic features.

TRACE has been used to account for a wide range of psycholinguistic data on word recognition, including the signature data originally used to motivate COHORT. Among these phenomena are: the effect of lexical context on phoneme recognition and its modulation by factors such as ambiguity; phonotactic rule effects on phoneme recognition, and their modulation by specific lexical items (phonotactic rules determine what sequences of phonemes are possible in a language); and the categorical nature of phoneme perception. TRACE was one of the prominent early successes of the PDP (parallel distributed processing) approach to modeling cognition and perception, and played a significant role in establishing the viability of the PDP paradigm.

TRACE has been challenged on both empirical and theoretical grounds, most notably by the Shortlist model of Norris (1994). A number of empirical studies have directly tested the assumption of top-down feedback in TRACE and yielded results more consistent with a purely bottom-up architecture in which phoneme recognition is autonomous and receives no feedback from lexical recognizers. For example, certain top-down lexical influences are dependent on using degraded stimuli, though TRACE should predict the effects in undegraded stimuli as well. Norris also argued that the TRACE architecture is implausible because it assumes the duplication of the entire network of lexical recognizers across multiple time slices. Shortlist is a purely bottom-up model that avoids the duplication of lexical recognizers by separating the process of generating candidate words (the 'shortlist') and the process of resolving identification via lexical competition.

Visual Word Recognition: Lexical Naming and Decision

Current prominent models of visual word recognition also take the form of computational models. One of the most influential of these models, the connectionist model of Seidenberg and McClelland (1989) (henceforth SM89), is a descendant of the McClelland and Rumelhart (1981) interactive activation model of word perception, which used localist word, letter, and feature units with hand-coded connections. SM89 builds on this earlier model but adopts distributed representations of both orthographic and phonological information. The model is a feedforward network with one hidden layer interposed between orthographic and phonological units. The connections between units were trained by back propagation on a word-naming task. The model accounts for several phenomena in word-naming, including differences among regular and exception words and differences in word-naming and lexical decision tasks. Because the model exhibits a gradual learning curve, it was also used to simulate the behavior of children acquiring word recognition skills.

One of the major debates in theories of word recognition is whether or not there is a single processing route from print to speech, or dual processing routes – separate lexical and nonlexical routes. The SM89 model is a clear example of a single-route architecture, and has come under sharp criticism from proponents of dual-route architectures. For example, Coltheart *et al*. (1993) note that the SM89 model actually performs more poorly on nonwords than humans do. Dual-route architectures are well suited to handing nonwords because the nonlexical route implements a general rule-based system that converts letter strings to strings of phonemes. Coltheart *et al*. also criticize the SM89 model for its inability to account for the dissociations evident in pure developmental surface dyslexia: normal nonword reading accuracy accompanied by gross impairments in reading exception words. Coltheart *et al*. offer a modular dual-route computational model, the Dual-Route Cascaded Model, which incorporates a learning algorithm for inducing the general pronunciation rules from examples (it was tested on the same letter-string/phone-string pairs

used by SM89). Although Coltheart *et al*. did not commit to the details of the lexical route, they suggest that something like the original McClelland and Rumelhart (1981) model may be an appropriate realization of that part of the word-naming system.

The debate surrounding dual-route and single-route architectures continues, with data from various forms of dyslexia playing an increasingly important role. The dual-route models have evolved to include explicit accounts of both reading aloud and lexical decision (Coltheart *et al*., 2001), and the connectionist models have evolved away from feedforward networks towards recurrent attractor networks that better handle generalization (Plaut *et al*., 1996).

Lexical Ambiguity Resolution: Processing Words in Context

One of the key lessons learned from several decades of attempting to program computers to process natural language is that massive local ambiguity is pervasive at all levels of linguistic representation. This is clearly evident in lexical processing, in which individual words are often associated with multiple syntactic and semantic senses, some mutually inconsistent, some partially inconsistent. Many of the theoretical themes noted above in word recognition are important in ambiguity resolution as well, in particular, the degree of autonomy or interaction present in initial lexical access. Differing positions on this issue distinguish the major theories of ambiguity resolution: *selective access models*, most closely associated with interactive theories, assume that contextual information provides direct top-down influence on initial sense activation; *ordered access models* assume that different senses are accessed in order of frequency of use; *exhaustive access* models, most closely associated with modular theories, assume that all senses are autonomously and exhaustively accessed in parallel; and *hybrid* models assume some combined effects of context and frequency.

In contrast to word recognition, the major theories of lexical ambiguity resolution are not strongly identified with specific implemented computational models (for reasons discussed below). However, there have been attempts to build detailed comprehensive computational models. One of the most successful is Kawamoto's (1993) recurrent connectionist model of ambiguity resolution. In this model, each lexical entry is represented by a pattern of activity over a 216-bit vector divided into separate subvectors representing a word's spelling, pronunciation, part of speech, and meaning. The network is trained with a simple error-correction algorithm by presenting it with the lexical patterns to be learned. The result is that these patterns become *attractors* in the 216-dimensional representational space. The network is tested by presenting it with just *part* of a lexical entry (e.g. its spelling pattern) and noting how long various parts of the network take to settle into a coherent pattern corresponding to a particular lexical entry. Kawomoto used these settling times to predict reading times, lexical decision times, and semantic access times. The model accounts for a wide range of phenomena, including frequency effects on processing of unambiguous and ambiguous words, context interactions with frequency, and the effect of task on the relative difficulty of processing ambiguous versus unambiguous words.

MODELS OF COMPREHENSION

Language comprehension involves more than the identification and disambiguation of words; the meanings of these parts must be pieced together in real time to yield the meanings of the sentences and the discourse. The state-of-the-art in computational linguistics and artificial intelligence places an upper bound on the field's ability to develop functional theories of comprehension processes. The best understood of these processes computationally and psychologically is syntactic parsing, the incremental assignment of grammatical structure to a string of words. Syntactic parsing is often assumed (though not universally) to be a necessary precursor to assigning a semantic interpretation.

Parsing

The major computational problem in parsing is how to handle local ambiguity. In fact, the prominent theories of sentence processing are actually theories of ambiguity resolution, and are distinguished by the positions they take on the key architectural questions surrounding ambiguity resolution. These include: are multiple structures computed and maintained in parallel at ambiguous points, or does the parser commit to a single structure immediately? What determines what structures the parser prefers when faced with ambiguity (e.g. referential discourse context, structural complexity, frequency of usage)? How do syntactic and lexical ambiguity resolution interact?

Two of the most influential models of sentence processing take opposing positions on most of these issues (though many of the issues are orthogonal).

Frazier's (1987) Garden Path Model asserts that the parser computes and pursues a single structure at ambiguous points, and that this initial structure is computed on the basis of general phrase structure rules without appeal to frequency, context, or detailed lexical information. Instead, structural simplicity is the principle that determines which structure is pursued in the case of local ambiguity. In contrast, the Constraint-based Lexicalist approach (MacDonald *et al*., 1994) claims that parsing is a constraint-satisfaction process that uses multiple information sources (or constraints), including context and detailed lexical information, without special architectural priority given to any particular constraint.

In sharp contrast to theories of word recognition, the dominant theories of sentence processing have not been strongly identified with specific computational models. (For example, the Garden Path Model was not implemented until 17 years after it was introduced (Spivey and Tanenhaus, 1998).) Among the earliest influential computational models were Marcus's (1980) wait-and-see parser, and the Wanner and Maratsos (1978) augmented transition network (ATN) grammar, which briefly contended with the Garden Path Model as a framework for understanding ambiguity resolution. Nevertheless, implemented computational models of sentence processing largely dropped from the scene in the 1980s.

Understanding why this happened will help place current parsing models in context. First, the early success of the Garden Path Model and the rise of modularity as a central theoretical theme in cognitive science jointly led the field to focus on modularity as the key architectural issue in sentence processing, and on ambiguity resolution as the key phenomenon providing insight into that issue. Second, Minimal Attachment is an extremely simple and practical theory – it can be stated in a few sentences and easily used to derive predictions cross-linguistically (once the underlying syntactic structures have been agreed upon). Computational models offered little advantage over such a theory, given this relatively narrow empirical and theoretical focus.

Two developments in the field are now leading researchers to develop more computational models. One is the need to provide more comprehensive, integrated accounts of sentence processing. Modularity is but one of several important architectural issues (Lewis, 2000), and computational modeling provides a way to develop and test interactions among components in a more functionally complete architecture. For example, computational models figure prominently among recent attempts to provide integrated accounts of both garden-path effects and working memory complexity effects in unambiguous constructions (Gibson, 1998; Lewis, 2000; Vosse and Kempen, 2000). Computational modeling also provides a way to import theoretical constraints from other areas of cognitive psychology, as in the Just and Carpenter (1992) working memory-constrained model.

A second development leading to more computational models is the rise of the constraint-based theories of sentence processing noted above. While these theories were initially proposed without associated computational models, it has become clear that the nature of these theories demands that they be formulated and tested as precise computational models. Several activation-based/connectionist models (e.g. Spivey and Tanenhaus, 1998) have been developed in the constraint-based framework.

Unlike computational models of word-level processes, which are almost exclusively the domain of connectionism, current computational theories of sentence processing are a mix of symbolic, connectionist, probabilistic, and hybrid models. As a class, the symbolic models tend to account for more complex cross-linguistic data, such as phenomena in head-final languages (e.g. Konieczny *et al*., 1997; Sturt and Crocker, 1996). However, recent models based on recurrent networks are attempting to push connectionist models in the direction of handling more complex syntactic structures, including difficult center-embeddings (Christiansen and Chater, 1999; Tabor *et al*., 1998). Several hybrid models are also under development, which have the promise of combining some of the strengths of both approaches (Jurafsky, 1996; Just and Carpenter, 1992; Lewis, forthcoming; Stevenson, 1994; Vosse and Kempen, 2000).

Discourse Processing

Processing running discourses of sentences in a text or verbal exchanges between interlocutors requires keeping track of multiple related levels of information (including, at least, the linguistic structure of the utterances, the goals and intentions of the participants, and the content of what is being discussed). Several major discourse processing theories have long been associated with implemented computational models. These include the Centering theory of Grosz and colleagues (Grosz *et al*., 1995), which provides an explicit algorithm for keeping track of attentional shifts among discourse entities and binding referring expressions to

these entities. The theory makes predictions about preferential patterns of pronominal reference that have been tested in reading time experiments (Gordon *et al*., 1993).

Another influential model is the Construction-Integration (CI) architecture of Kintsch and colleagues (Kintsch, 1998). Comprehension in the CI architecture is an activation-based process that proceeds in two phases. The *construction* phase produces local sentence-level propositions using simple, context-independent rules. The *integration* phase uses a constraint satisfaction process to integrate the possibly incoherent set of local propositions into a coherent whole organized by higher-level macropropositions. Many of the CI model's predictions about anaphora resolution, word identification, and the generation and retrieval of macropropositions have been empirically confirmed (Kintsch, 1998).

MODELS OF PRODUCTION

The dominant psycholinguistic theories of production are now associated with implemented computational models. Most psycholinguistic theories of production focus on the final stages of production: producing an ordered set of phonemes corresponding to some (given) intended utterance. (In contrast, much work on production in computational linguistics and artificial intelligence is focused on the functionally more difficult processes of higher-order discourse and speech act planning.) The theoretical landscape is quite similar to theories of lexical processing: all the models are activation-based, but differ in their assumptions about the nature of interaction between independent levels of representation. Among the best-known models are those of Dell (Dell *et al*., 1997) and Levelt (Levelt *et al*., 1999), which take opposing positions along this dimension. The Dell model is an interactive-activation-based theory that takes an ordered set of word units as input and generates a string of phonemes. Most of the important phenomena accounted for by the model are speech errors, including perseverations (e.g. *beef needle soup*) and anticipations (e.g. *cuff of coffee*). Dell's model consists of a network of word units (lemmas) and phoneme units and bidirectional links between word units and their constituent phonemes. The signature phenomenon accounted for by the feedback from phonemes to words is the statistical overrepresentation of mixed errors, such as saying *rat* when the intention is *cat*. When the word node for *cat* is active, the phoneme segments /k/, /æ/, and /t/ are activated. The latter two segments then feed activation to *rat*, which may already be above baseline due to a semantic association.

The WEAVER++ model (Levelt *et al*., 1999) is also activation-based, but eliminates bidirectional connections. Processing is staged in strictly feedforward fashion, starting with conceptual preparation (not implemented), and proceeding to lexical selection, morphological and phonological encoding, phonetic encoding, and finally articulation. Unlike most other production theories, the WEAVER++ model accounts primarily for reaction time (RT) data, and was developed exclusively on the basis of RT data from simple production paradigms such as picture naming. However, Levelt and colleagues have also shown that the model can account for some speech errors as well, including those used to motivate the bidirectional connectivity in the strongly interactionist models.

MODELS OF ACQUISITION

With one prominent exception noted below, computational models have only recently begun to play an important role in theorizing about language acquisition. A fundamental difficulty facing the development of serious computational models of acquisition is that the input to such models must generally be a large corpus of utterances *in context*. Although large computer databases of naturally occurring text and speech are now readily available, such databases currently lack a component that nearly all acquisition theories assume is necessary: some representation of the context in which the utterance occurs. For this reason, much computational modeling of grammar acquisition is currently done using small-scale, artificially created grammars or lexicons, in small-scale, artificial domains (Feldman *et al*., 1996).

However, current speech and text databases are well suited to exploring *distributional* theories of acquisition. For example, certain kinds of lexical and syntactic information can be determined from purely distributional analyses (Cartwright and Brent, 1997). One important example is specific verb subcategorization frames, which play a critical role in all modern syntactic theories and sentence comprehension theories. Computational models of speech segmentation have also been developed that learn to identify word boundaries from exposure to continuous speech (Christiansen *et al*., 1998).

By far the most controversial and influential computational acquisition model is the Rumelhart and McClelland (1986) (henceforth RM86) connectionist model of the acquisition of the past tense form of English verbs. Past tense inflection

acquisition has served as a kind of *Drosophila* for research on the mechanisms underlying apparently rule-governed linguistic behavior, and lies at the center of a much broader debate on connectionism and language. The RM86 model was proposed as an alternative account to the traditional view that the past tense form of English verbs is formed by dual routes: an abstract rule that handles all regular forms by adding *–ed* to a stem, and a memory that contains a list of irregular exception words (such as *ran*). The connectionist model instead proposed a single processing route, implemented as a feedforward network with a single hidden layer, and no explicit representation of a rule. The network was trained on 460 pairs of root and inflected forms. The network reproduced the well-known U-shaped performance curve often taken as prima facie evidence for the formation of a general *–ed* rule: children initially do not make overgeneralization errors (e.g. saying *runned* for *ran*), but then go through a period of apparently over-applying the general rule, and finally recover to adult levels of performance. Crucially, the network also generalized and transferred appropriately to novel low-frequency verbs (e.g. the network correctly produced *wept* as the past tense of *weep*), capturing subregularities among the irregular words in the corpus.

Every aspect of this work has come under sharp criticism, including the content of the artificial database on which RM86 trained their original network, the empirical robustness of the U-shaped curve itself, and the use of connectionist architectures more generally as accounts of human linguistic and cognitive performance (Marcus, 1996; Pinker and Prince, 1988). Some of these criticisms have been addressed in revisions to the model (MacWhinney and Leinbach, 1991), but new empirical evidence from adult processing has also accumulated in favor of the dual-route view (Marslen-Wilson and Tyler, 1998).

CURRENT DIRECTIONS

A number of short-term and long-term theoretical directions are evident in this review. One overarching trend is clear: computational modeling is playing an increasingly important role in theorizing in all subfields of psycholinguistics. There are several reasons for this, all related to theoretical trends in psycholinguistics more generally. There are four trends in particular that are likely to continue in the near term. First, there is a gradual move towards providing more *integrated accounts* of multiple components of linguistic processing. For example, several computational models now combine theories of lexical ambiguity resolution and sentence processing, or ambiguity resolution and working memory (e.g. Kintsch, 1998). Second, there is an increasing move towards developing theories that are *jointly constrained by processing and acquisition data* (e.g. Seidenberg and McClelland, 1989). Accompanying this trend is a growing reliance on large machine-readable corpora to test models that have some role for linguistic experience. Third, theories of normal linguistic performance are increasingly constrained by *neuropsychological data* from patients with linguistic deficits due to brain damage. Computational models of intact performance can be 'lesioned' and tested against both normal and patient data (e.g. Plaut *et al*., 1996). Fourth, there is increasing convergence in all subfields of psycholinguistics towards *continuous activation-based* models of processing. These include parallel distributed processing approaches, but also many activation-based symbolic models.

There are also some emerging trends that will most likely play out over the longer term. These include increasing attempts to integrate psycholinguistic models with other process theories in cognitive psychology, such as detailed models of memory and skill, and increasing convergence with efforts in computational linguistics as both fields attempt to tackle functionally difficult areas such as word sense disambiguation and robust parsing. These latter efforts will naturally result in greater contact with linguistic theory. In particular, linguistic theories which prove to be important in the development of scalable and robust speech and natural language systems will be incorporated in psycholinguistic models that place a premium on functionality and scalability.

References

Cartwright TA and Brent MR (1997) Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition* **63**(2): 121–170.

Christiansen MH, Allen J and Seidenberg MS (1998) Learning to segment speech using multiple cues: a connectionist model. *Language and Cognitive Processes* **13**(2–3): 221–268.

Christiansen MH and Chater N (1999) Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* **23**(2): 157–205.

Coltheart M, Curtis B, Atkins P and Haller M (1993) Models of reading aloud: dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608.

Coltheart M, Rastle K and Perry C (2001) DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* **108**(1): 204–256.

Dell GS, Burger LK and Svec WR (1997) Language production and serial order: A functional analysis and a model. *Psychological Review* **104**(1): 123–147.
Feldman J, Lakoff G, Bailey D, Narayanan S and Regier T (1996) L0: the first five years. *Artificial Intelligence Review* **10**: 103–129.
Frazier L (1987) Sentence processing: a tutorial review. In: Coltheart M (ed.) *Attention and Performance XII: The Psychology of Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Gibson EA (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**: 1–76.
Gordon PC, Grosz BJ and Gilliom LA (1993) Pronouns, name, and the centering of attention in discourse. *Cognitive Science* **17**(3): 311–347.
Grosz BJ, Weinstein S and Joshi AK (1995) Centering: a Framework for modeling the local coherence of discourse. *Computational Linguistics* **21**(2): 203–225.
Jurafsky D (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* **20**(2): 137–194.
Just MA and Carpenter PA (1992) A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* **99**(1): 122–149.
Kawamoto AH (1993) Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language* **32**: 474–516.
Kintsch W (1998) *Comprehension: A Paradigm for Cognition*. Cambridge, UK: Cambridge University Press.
Konieczny L, Hemforth B, Scheepers C and Strube G (1997) The role of lexical heads in parsing: evidence from German. *Language and Cognitive Processes* **12**(2/3): 307–348.
Levelt WJM, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**(1): 1–38.
Lewis RL (2000) Specifying architectures for language processing: process, control, and memory in parsing and interpretation. In: Crocker M, Pickering M and Clifton C (eds) *Architectures and Mechanisms for Language Processing*. Cambridge, UK: Cambridge University Press.
Lewis RL (forthcoming) *Cognitive and Computational Foundations of Sentence Processing*. Oxford, UK: Oxford University Press.
MacDonald MC, Pearlmutter NJ and Seidenberg MS (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703.
MacWhinney B and Leinbach J (1991) Implementations are not conceptualizations: revising the verb learning model. *Cognition* **40**: 121–157.
Marcus GF (1996) Why do children say 'breaked'? *Current Directions in Psychological Science* **5**: 81–85.
Marcus MP (1980) *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.
Marslen-Wilson W and Tyler LK (1980) The temporal structure of spoken language understanding. *Cognition* **8**: 1–71.
Marslen-Wilson W and Tyler LK (1998) Rules, representations, and the English past tense. *Trends in Cognitive Sciences* **2**(11): 428–435.
McClelland JL and Elman JL (1986) The TRACE model of speech perception. *Cognitive Psychology* **18**: 1–86.
McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception: 1. An account of the basic findings. *Psychological Review* **88**: 375–407.
Norris D (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition* **52**: 189–234.
Pinker S and Prince A (1988) On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**(1–2): 73–193.
Plaut DC, McClelland JL, Seidenberg M and Patterson KE (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115.
Rumelhart DE and McClelland JL (1986) On learning the past tense of English verbs. In: McClelland JL and Rumelhart DE (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.
Seidenberg M and McClelland JL (1989) A distributed developmental model of word recognition and naming. *Psychological Review* **96**: 523–568.
Spivey MJ and Tanenhaus MK (1998) Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 1521–1543.
Stevenson S (1994) Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research* **23**(4): 295–322.
Sturt P and Crocker M (1996) Monotonic syntactic processing: a cross-linguistic study of attachment and reanalysis. *Language and Cognitive Processes* **11**: 449–494.
Tabor W, Juliano C and Tanenhaus MK (1998) Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* **12**: 211–272.
Vosse T and Kempen G (2000) Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* **75**: 105–143.
Wanner E and Maratsos M (1978) An ATN approach to comprehension. In: Halle M, Bresnan J and Miller GA (eds) *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

Further Reading

Clifton C and Duffy SA (2001) Sentence and text comprehension: roles of linguistic structure. *Annual Review of Psychology* **52**: 167–196.
Crocker MW (1996) *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Dordrecht: Kluwer Academic.

Elman JL (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Gernsbacher MA (ed.) (1994) *Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Lewis RL (1999) Cognitive modeling, symbolic. In: Wilson RA and Keil FC (eds) *The MIT Encyclopedia of Cognitive Science*. Cambridge, MA: MIT Press.

McClelland JL (1999) Cognitive modeling, connectionist. In: Wilson RA and Keil FC (eds) *The MIT Encyclopedia of Cognitive Science*. Cambridge, MA: MIT Press.

Norris D (1999) Computational psycholinguistics. In: Wilson RA and Keil FC (eds) *The MIT Encylopedia of Cognitive Science*. Cambridge, MA: MIT Press.

Reading and Writing

Introductory article

Alexander Pollatsek, University of Massachusetts, Amherst, Massachusetts, USA
Brett Miller, University of Massachusetts, Amherst, Massachusetts, USA

CONTENTS

Reading and writing are two important human activities in which language transmission is carried out through skilled visual processing and motor acts. Evidence suggests that reading, even for skilled readers, is essentially a word-by-word process, involving accessing the spoken language.

INTRODUCTION

In a speech on discoveries and inventions, Abraham Lincoln characterized writing as the greatest invention of mankind, as it allowed people to communicate over great distances in space and time – and, as a result, he thought that civilization and institutions such as democracy were not possible without reading. The focus in this article is on the 'lower-level' issues in reading and writing: how the transmission of language, which is biologically programmed to be in the form of speaking and listening, has been transformed into skilled visual processing and motor acts. Indeed, it's remarkable that the average skilled reader can easily read text at about 300–350 words a minute, or about twice the speed at which language is normally spoken, even though we are genetically programmed to hear language rather than read it.

WRITING SYSTEMS

The production of writing systems has been a very recent development in human evolution. Of course, when 'writing' first occurred depends to some extent on one's definition of a writing system. According to the Merriam and Webster dictionary, writing is 'the act or art of forming visible letters or characters'. This definition is ambiguous, in that it's not clear what the characters are supposed to represent. If we demand that 'writing' not only has to convey the essence of a spoken utterance, but convey it word by word, then the first unambiguous examples of a writing system (found in Mesopotamia) are from about 7,000 years ago.

Written languages have usually developed in a systematic fashion: from a pictorial representation of words (logography), to a system in which the characters represent syllables, and then to an alphabetic representational system which represents even smaller units of speech. In a logographic system, a visual symbol is used to represent a word (e.g. a drawing of a duck might represent a 'duck'). There are several problems with such a system. First is the problem of representing abstract nouns, articles, prepositions, or pronouns. Second is the difficulty of drawing recognizable characters, even if the word is a concrete noun. In fact, although such systems started out with many fairly recognizable 'characters', in most modern logographic systems the majority of the characters are quite abstract. The most widely used logographic writing system today is Chinese, and most non-Chinese readers would be quite unlikely to guess the meaning of all but a few characters. Actually, the characters in Chinese represent the smallest units of meaning (*morphemes*) rather than words. In English, words with affixes and compounds have two or more morphemes (e.g. de/code, ring/ing, re/view/er, cow/boy). A character in Chinese also reliably represents a spoken syllable. Japanese also employs logographic characters (*kanji*), that are derived from the Chinese system, but also has phonologically based characters (*kana*) as well.

Most modern writing systems are alphabetic, and have evolved from the Phoenician and Hebrew writing systems, which in turn evolved from systems that were logographic and syllabic. In alphabetic systems, an individual character does not convey meaning; instead it is a code that ideally represents the smallest unit of sound, the *phoneme.* In some languages, such as Spanish, the alphabetic code is quite regular (i.e. there are rules allowing

one to convert the letters to phonemes): once one knows the rules any written word can be correctly pronounced. English has probably the most irregular alphabetic writing system, having both strangely spelled words like 'island' and letter combinations that have no rules governing how they are pronounced (e.g. 'bough, dough, rough, through'). However, the relation between letters and phonemes is not a simple one even in regular alphabetic systems, because a letter doesn't always represent a single phoneme. For example, in Spanish, *ch* or *qu* represents a single consonant phoneme, and in English, *x* usually represents two phonemes (a 'ks' sound). Moreover, a letter can represent different phonemes depending on the context (e.g. in Latin American Spanish *c* has an 's' sound when it is before either *e* or *i*, but has a 'k' sound before the other letters). It is worth noting that in most Near Eastern languages that use some variant of the ancient Hebrew orthography (e.g. Hebrew and Arabic), most vowels are not represented at all. One has to figure most of them out from context. However, there is a 'training' form of the orthography in which the vowels are represented by little marks above the letters.

The direction of text and relevance of spacing and punctuation varies across languages. In all European languages, words (and sentences) are written from left to right. In contrast, most Near Eastern languages (e.g. Hebrew and Arabic) go from right to left on the page. Chinese, historically, has used several orders. The current order that is most prevalent is that the script is organized in columns that go from the top of the page to the bottom, and the columns are read from the right to the left of the page. The use of spacing also varies by language. In most alphabetic languages spacing indicates a word boundary, whereas there is no spacing in Chinese to represent words.

THE PROCESS OF EXTRACTING VISUAL INFORMATION FROM WRITTEN LANGUAGE

Obviously, a major component of reading is moving one's eyes through the text. Perhaps a bit less obvious is the fact that these eye movements are not continuous. In general, unless one's eyes are following a moving target, they move in discrete and very rapid movements (*saccades*) and then stay still for relatively long periods (*fixations*). Saccades in reading take on the order of 15–30 milliseconds (ms) (depending only on the length of the saccade) and fixations in reading are usually about 150–400 ms, although they can be shorter or longer than that. No meaningful information is picked up during the saccades in reading (or any normal viewing situation), so that reading is essentially a 'slide show' with the slides (fixations) coming four to five times a second. This pattern is similar to what the eyes do when one is viewing a photograph or a static scene. About 10–15 percent of all eye movements in typical reading situations are *regressions* (i.e. going back in the text). Many are short (just going back a word or two). (See Figure 1 for an illustration of a typical eye-movement pattern in reading a sentence.) As the text becomes more difficult, this difficulty is reflected in many eye-movement measures: the average duration of fixations increases, the average size of a forward saccade gets smaller, and the number of regressions increases.

Why are the eyes moving so continually during reading? The answer is that in order for the reader to be able to identify the words, they have to keep on moving across the text. As Figure 1 illustrates, not all words are fixated, but the field of vision in which detail can be extracted is fairly small, so that one has to be quite near a word in order to be able to read it. We know that the field of vision in reading is small from several perspectives. First, the density of photoreceptors in the retina of the eye gets dramatically sparser as one gets further away from the center of vision (the *fovea*) making it hard to see detail away from the center of vision. Second, a series of studies have shown conclusively that the area of the page in which readers extract useful information from the text is quite limited, extending at most two words from the word being fixated. This means that claims that there are 'speedreaders' – people who can read whole sentences or paragraphs in a single glance – are false.

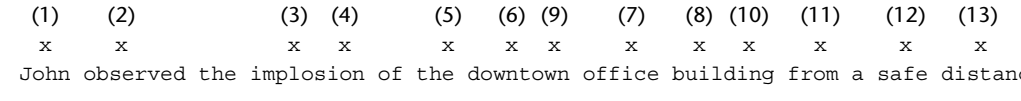


Figure 1. A typical eye-movement pattern when reading a sentence. Each *x* above the line indicates the location of a fixation on the line of text directly below the *x*, and the numbers in parentheses represent the order in which the fixations occurred. Note that although most words were fixated, not all were, and that some words were fixated twice. Also note that fixation 9 on 'downtown' was a *regression*, as the prior fixation had been on 'building'.

These studies used the *moving window* technique. Although we will mainly discuss studies in English, similar studies done with other writing systems indicate that the results are similar for all writing systems and languages. In these studies, people view text on a computer monitor while their eye movements are monitored by a precise system that knows what letter they are fixating (i.e. what letter their eyes are pointing to) and can measure the eye position with about 1 millisecond accuracy, so that the text can be changed very rapidly during each saccade. The reader sees normal text in a certain region around the fixation point (the *window*), but all the text outside the window is changed to something like random letters or 'x's. Regardless of whether the eyes move forward or backward through the text, the window will follow them and there will be normal text only in a restricted region around the fixation point. (See Figure 2 for an illustration of a moving window experiment.)

The major finding is that the window only has to extend from four letters to the left of fixation to 14 characters to the right of fixation in order for reading to be perfectly normal (i.e. both normal comprehension and normal speed). Conversely, if one presents random letters in a region that extends 14 characters to either side of fixation and normal text outside that region, reading is virtually impossible. Thus, not only do readers normally extract information in a relatively narrow window around where they are fixating, they have to. Moreover, cutting the window of normal text down somewhat from 14 letters doesn't prevent skilled reading. If it contains only the fixated word, reading speed is roughly 60 percent of normal reading speed for skilled readers, and if it contains the fixated word and the word to the right, reading speed is 90 percent of normal!

In summary, the acuity of the eyes limits the region of text from which meaningful information can be extracted to 14 characters, or about three words. Attentional patterns limit the region even more: virtually no information is extracted from the left of the fixated word or from other lines of text. There's nothing magic about right versus left, however. For Hebrew readers (Hebrew is read right to left) the asymmetry of the window is reversed, as their window went from 14 characters to the left of fixation to four characters to the right of fixation. (As Hebrew–English bilingual readers, when reading English, had windows that looked like those of other readers of English, the asymmetry of the window is clearly due to attentional factors.)

The data thus indicate that readers are moving their eyes through the text to read individual words. However, the agenda on each fixation is often a bit more than reading a single word:

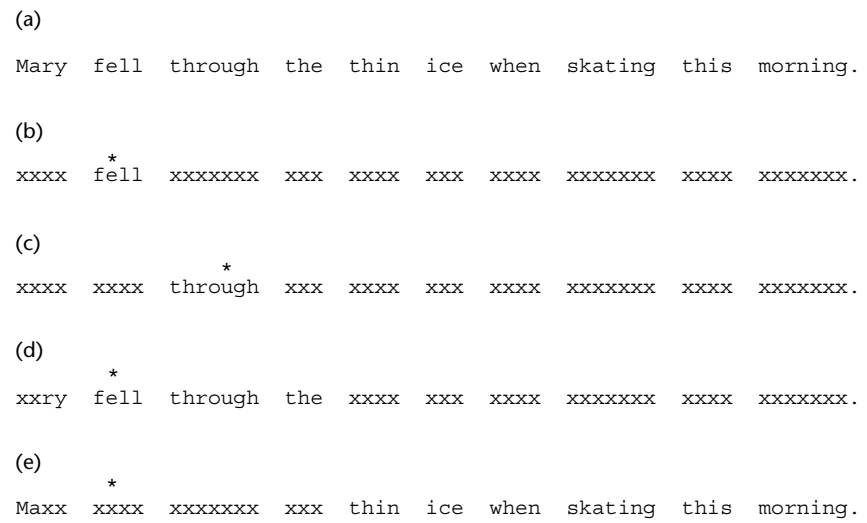


Figure 2. Illustration of the moving window technique. (a) Shows how a sentence would appear under normal viewing conditions. In (b) and (c), two successive fixations (indicated by the asterisks) of a *one-word* window condition are illustrated. (d) Illustrates a single fixation in a different window condition, where the reader sees four characters to the left of fixation and 14 characters to the right. (Here, reading is normal.) (e) Illustrates a single fixation in a condition where the area around the fovea is occluded (and reading is virtually impossible). The usual measure of the effect of the window is reading rate (comprehension is usually near perfect) along with finer measures such as average fixation time and number of fixations.

sometimes two words, and perhaps rarely three short words, can be read. The statistics in reading bear this out: *content words* (nouns, verbs, and adjectives) are fixated about 90 percent of the time and *function words* (articles, prepositions, and conjunctions) are fixated about 30 percent of the time in normal text. However, the processing on each fixation is even more complex than this. On many fixations, partial information about a word seen to the right of fixation is processed and used to help identify that word on the next fixation.

This use of partial information has been uncovered using the *boundary* paradigm, where only a single word in the text is changed as the eyes move. (See Figure 3 for an illustration of the boundary paradigm.) As in the figure, if 'steak' is the *target* word that belongs in the sentence, one of a number of possible *previews* (e.g. 'stone', 'stald', 'roast', 'rozcd', 'clasp', or 'dncbo') might appear in the target word location before that location is fixated. By the time the target location is fixated, the preview is replaced by the target and the target remains visible for the rest of the time. (Readers are totally unaware of the existence of the preview and are unaware that there are any text changes in these experiments.) A major finding is that when the preview (either a word or a nonword) shares

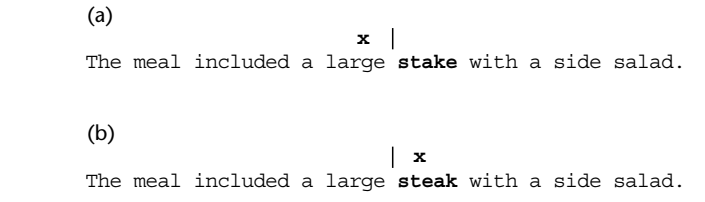


Figure 3. Illustration of the boundary technique. The critical word that is changed is marked in bold in the figure for illustrative purposes, but it is not so marked during the experiments. Before the reader's eyes cross the imaginary boundary (the boundary is represented here by a '|' and the fixation point by an **'x'**), a preview – another word or a nonword – appears in the target word location (see (a)). After the imaginary boundary has been crossed, the preview changes to the target word and remains as the target word for the rest of the time that the reader spends reading the sentence (see (b)). Other possible preview stimuli could be totally different words (e.g. 'clasp'), a visually similar word or nonword (e.g. 'stone', 'stald'), or a semantically similar word (e.g. 'roast'). Note that the preview and target always have the same number of letters. The usual measure of the effect of the preview is fixation time on the target word (and possibly fixation time on the following word or two as well). Also note that people are virtually never aware of the preview or the change.

several letters with the target, such as 'stone' or 'stald', the fixation time on the target word 'steak' is less than if the preview were something like 'round' or 'rozcd'. This suggests that people are extracting visual features from words before they fixate them. Yet other studies indicate that the features aren't really 'visual': one gets just as big a benefit from 'STONE' as a preview for 'steak' as when 'stone' is a preview for 'steak'. Thus, it appears that abstract information about letters (especially beginning letters) is extracted on one fixation that helps identify the word on the next fixation.

This raises the question of what other partial information about words is extracted from one fixation to help identify the word on the following fixation. One of the most surprising findings is that the meaning of the preview word is irrelevant. In the above example, a preview of 'roast' for 'steak' provides no benefit. In fact, for English–Spanish bilinguals, a preview that was a translation of the word fixated provided no benefit beyond what could be explained by letter overlap. Moreover, components of meaning (morphemes) are also largely irrelevant; 'cowxxx' is no better a preview for 'cowboy' than 'corxxx' is for 'corner'. However, the sound of the preview is relevant: a homophone (word that sounds the same) is a better preview for a target word than a word that is just as visually similar to the target but is not a homophone of it (e.g. 'stake' is a better preview of 'steak' than is 'stoke').

To summarize, although skilled readers read quite fast (over 300 words per minute), they do so by moving through the text fairly methodically. Reading isn't quite 'word by word' as not every word is fixated; some words are identified before they are fixated and are skipped. Moreover, the phenomenon of preview benefit indicates that words are often partially identified before they are fixated. However, the data indicate that people extract the information from virtually all words. Almost no words are merely 'guessed'.

The above discussion of previews indicates that the fixation time on a word is sensitive to the type of preview. In addition, the fixation time on a word is quite sensitive to many aspects of the text. Long words are fixated longer than shorter words, and less common words are fixated longer than more common words (even when they are equal in length). Moreover, when a word is highly predictable from the prior sentence or discourse context, it is both more likely to be skipped and will be fixated for a shorter period of time. As a result of these findings and others, it appears that the forward

movement of the eyes through the text is largely governed by identifying words. The signal to move the eyes forward is likely to be a decision that the fixated word has been identified sufficiently to be able to move on.

As indicated earlier, however, many eye movements in text are *regressions* or movements backwards to earlier parts of the text. These often signal processing difficulty. Regressions can appear because a word is misidentified, and as a result, the text becomes nonsensical. They have also been shown to occur in many situations where processing the sentence or discourse becomes difficult. They occur, for example, when the syntax is difficult or ambiguous (especially when one chooses the wrong syntactic interpretation of an ambiguous construction), or when it is unclear what concept a pronoun is referring to. Consider the sentence 'While the manager ate the man robbed the store.' This sentence is difficult because the reader may initially think the 'man' is the object that the manager ate, realize that this is implausible, and have to reanalyze the sentence. In fact, readers often make regressions to reread the earlier parts of such sentences. Although many of these regression effects are quite rapid (occurring half a second or so after the text that produces the difficulty has been read), they rarely occur immediately, in contrast to word frequency effects. It thus appears that most discourse processing lags a little behind word identification, and that the primary 'engine' driving the eyes forward is the identification of individual words. This makes sense: if we always kept fixating on a word until we were sure it made sense in context, it would slow reading down appreciably. Instead, a better strategy appears to be to move ahead throught the text by successfully identifying individual words (with the sentence processing lagging behind a little) and then regress when something doesn't make sense at the sentence or discourse level, especially because most such errors are caught quickly (after a word or two). Thus fixing them doesn't incur a high cost.

USE OF PHONOLOGICAL INFORMATION DURING READING

Perhaps the most contentious issue in silent reading is how the spoken language enters into the process. Many people believe that this happens only for beginning readers, and that skilled readers go directly from print to meaning. However, there is now a large body of literature that indicates that this is not the case and that even skilled readers rely quite a bit on getting to a *phonological* or sound-based code. It's important to realize that phonological coding (i.e. having an auditory image) and *subvocalizing* (actually using the speech apparatus to form the sounds very quietly) are different. It's easy to say the digits 1 to 9 aloud at a rapid rate while thinking of the sound of a word. (Try it.)

The above illustration makes plausible that sounds can be accessed without subvocalization, but is there evidence that people do so? There are several lines of evidence that indicate that sound coding is important. First are experiments with accessing individual words, which ask people to make a decision about whether a word is in a particular semantic category. For example, people will first see a category label, such as FOOD, followed by a word, such as MEAT, that they have to judge as being a member of the category or not. The finding is that skilled readers are not only slower at responding *no* to homophones of category members (e.g. MEET) than to visually similar controls (e.g. MELT), they also make 10–15 percent more errors on them! A smaller, but similar, effect has also been observed with words that could be homophones but are not. For example, people have difficulty in rejecting PILLOW–BEAD as being unrelated (where 'bead' could be a homophone of 'bed' if it were pronounced analogously to 'head'). The evidence for sound coding comes not only from such experiments with individual words, but in silent reading experiments as well. As indicated earlier, in a boundary experiment (see Figure 3), a homophone of the target word served as a better preview than a visually similar word. Thus, phonological coding enters into reading even before a word is fixated.

Moreover, there is little evidence, developmentally, that reading becomes less dependent on sound coding as readers become more proficient at reading. It has been argued that less skilled readers are more dependent on phonological coding because they tend to show bigger effects produced by sound coding. For example, when asked to judge whether two words are synonyms, people are slower in responding *no* to homonyms than to unrelated words, and furthermore, this interference effect is bigger, in absolute terms, for beginning readers. However, from another perspective, the size of the interference effect doesn't change developmentally, as the size of the interference effect is proportional to the absolute times in the task. This latter observation (and some related experiments) suggest that as readers become more skilled, the process becomes more automated and less accessible to consciousness rather than changing qualitatively.

Another indication of the importance of sound coding is that poor performance on two different kinds of tasks that tap phonological coding are quite reliable predictors of *dyslexia* (i.e. substantial difficulty in reading). One reliable predictor of reading difficulty is being slow in naming things. People who have difficulty are not only slow at naming 'verbal materials', such as numbers and letters, but they are also slow at naming nonverbal materials, such as line drawings of common objects or color patches. A second predictor of reading difficulty is difficulty in analyzing a word for its constituent *phonemes*. One example is that people having trouble reading will have difficulty with the phoneme deletion task (e.g. responding /at/ after hearing /bat/). Young children, who are pre-readers, or people who have no experience with an alphabetic language also have severe problems doing this task. However, the children who have difficulty with this task usually have no difficulty deleting syllables in an analogous 'language game'. Thus, their problem in the phoneme deletion task is not being able to 'hear' the units that the letters represent rather than some general inability to play 'linguistic games'. (Incidentally, it is largely a myth that dyslexics are people who see upside-down or backwards and confuse 'dog' with 'god'. Only a few people with reading difficulty seem to have these kinds of problems.)

STUDIES OF WRITING

Writing research has proven much more difficult to conduct than reading research. The major reason is probably that, in writing, the text is produced by the participant, whereas in reading, the researcher can generate the materials and thus can more easily manipulate the variables of interest. Although the sheer amount of research conducted in reading may be larger than for writing, there have been some interesting findings from the research on writing.

Similar to reading difficulties (*dyslexia*), there are people who have writing and/or spelling difficulties (*dysgraphia*). Dysgraphia can either occur developmentally or be the result of later brain injury. There are a tremendous variety of symptoms that dysgraphic individuals display. These symptoms frequently co-occur with other language-processing difficulties such as dyslexia and aphasia. Some of the problems displayed by dysgraphic individuals include, but are not limited to: misspelling text; slow, painstaking writing; and producing illegible text despite formal writing training. Individuals may also show lexical intrusions (using inappropriate words) when writing (e.g. spelling 'quell' as *kwell*). It is important to note that not all individuals display the same pattern of deficits; however, there are patterns of deficits that tend to occur together, yielding different classifications under the umbrella of dysgraphia.

One interesting type of dysgraphia is phonological dysgraphia. Individuals with this impairment do not have much difficulty writing words that they know, but when asked to spell nonwords or novel words, these individuals are unable to do so correctly. For example, if a teacher said *agog* in class, someone with phonological dysgraphia would often either spell it incorrectly or be completely 'stumped' and not write anything. The normal writer would have the sound code available for the word he or she had just heard, and should be able to generate a plausible spelling. However, individuals with phonological dysgraphia often either have difficulty in breaking this word into phonemes or in translating known phonemes into appropriate letters or letter combinations. This syndrome indicates that phonological processing is important for writing as well as for reading.

There is also interesting evidence from dysgraphia about the involvement of abstract (i.e. case- and font-independent) letter representations in writing as well as reading. On some level, it is obvious that such a level exists in normal writers – otherwise, it would be impossible to copy strings of random letters where one has to change them all from upper to lower case, or vice versa. However, the work with dysgraphia indicates that there is structure to this level of representation. In general, patients who have writing problems as a result of strokes or other brain injuries do not make random errors when they attempt to write from dictation. For example, these patients make many substitution errors (e.g. writing *bord* instead of *bird*), but these errors are far from random. They typically substitute a vowel for a vowel and a consonant for a consonant over 80 percent of the time, which is clearly well above chance. Thus, it seems clear that their abstract letter level of representation includes the vowel–consonant distinction, and is likely to include other letter properties as well, but falls short of an adequate representation for generating 'correct' (i.e. possible) spellings for the words.

SUMMARY

To conclude, we've learned a lot about reading and writing throughout the last century or so. Reading is not a magical process. Contrary to some popular beliefs, we have to encode words, more or less one

at a time, by looking quite close to them, if not directly at them. Moreover, even though it is often believed that skilled readers go directly from print to meaning, this is not so. Even skilled readers use internal sound codes when they read and people who find it difficult to access the sounds of words have difficulty reading. To some extent, this may partly explain why alphabetic systems are so common. That is, they code for the sounds of words in an efficient way, which makes them a natural coding system.

Further Reading

Allport A, MacKay D and Prinz W (eds) (1987) *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing*. San Diego, CA: Academic Press.

Coltheart M, Patterson K and Marshall J (eds) (1980) *Deep Dyslexia*. London, UK: Routledge & Kegan Paul.

Gelb IJ (1963) *A Study of Writing*, 2nd edn. Chicago, IL: University of Chicago Press.

Henderson J, Singer M and Ferreira F (eds) (1995) *Reading and Language Processing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Henderson L (ed.) (1984) *Orthographies and Reading*. Mahwah, NJ: Lawrence Erlbaum Associates.

Huey E (1908) *The Psychology and Pedagogy of Reading*. New York, NY: Macmillan.

Hulme C and Snowling M (eds) (1994) *Reading Development and Dyslexia*. London, UK: Whurr Publishers.

Miles TR (1983) *Dyslexia*. London, UK: Granada Publishing.

Perfetti C, Rieben L. *et al*. (eds) (1997) *Learning to Spell: Research, Theory, and Practice across Languages*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rayner K (ed.) (1983) *Eye Movements in Reading*. New York, NY: Academic Press.

Rayner K and Pollatsek A (1989) *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice-Hall.

Snowling M (1987) *Dyslexia: A Cognitive Developmental Perspective*. Oxford, UK: Basil Blackwell.

Thomson M (1984) *Developmental Dyslexia: Its Nature, Assessment, and Remediation*. Baltimore, MD: Edward Arnold.

Von Euler C, Lundberg I and Lennerstrand G (eds) (1989) *Brain and Reading*. London, UK: Macmillan.

Second Language Acquisition

Introductory article

Alan Juffs, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
Robert M DeKeyser, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Second language acquisition (SLA) involves the development of the knowledge of a complex system of sound, word, sentence structure, and meaning of any non-native language by children as well as adults; it also involves the acquisition of the ability to use that system appropriately in different social settings.

THE RELATION OF SECOND LANGUAGE ACQUISITION TO FIRST LANGUAGE ACQUISITION

The acquisition of first language (L1) phonology (sound structure) and morphosyntax (the rules of word formation and sentence formation) normally occurs rapidly, effortlessly, and without explicit instruction. By about age five, research suggests that acquisition of the basic linguistic system is largely complete. Many studies that are based on abstract theories of language structure have also shown that children possess knowledge of subtle properties of their first language that they could not have acquired by simply listening to the speech of their care-givers or by repeating only what they hear. Although children are sometimes told when they make mistakes with vocabulary, they rarely receive useful corrections about mistakes in word order and morphology.

These facts about L1 acquisition have led to a widely accepted theory that children's acquisition of their mother tongue is guided by innate principles, which linguists have called Universal Grammar (UG). It remains a matter of debate whether these innate principles are specific only to humans' capacity for language or whether they are related to more general learning mechanisms. In either case, UG can be thought of as a blueprint of – or a set of constraints on – what a possible human language is; in other words, UG is a theory of properties that all human languages share and which vary only slightly from one language to another. The task that children face in acquiring their native language is to match the language spoken around them (the 'input' in acquisition) with the blueprint they have been born with. Using this blueprint to filter and understand the input, they construct the grammar of the specific language they are acquiring without conscious reflection.

Other views of first language acquisition downplay the role of innate mechanisms and seek to account for acquisition by studying the nature of the input that children receive, as well as their abilities to derive generalizations from this input. Hence, while UG is widely accepted within the field of linguistics, in the wider field of cognitive science the existence of a UG separate from general cognition remains highly disputed.

Many researchers believe that the acquisition of a second language by adults involves some of the same abilities that are involved in L1 acquisition. Even those researchers who question whether UG is still available after puberty (i.e. researchers who believe in a critical period for UG) accept that some knowledge of universal constraints on human languages may be accessible through knowledge of the first language. However, second language acquisition (SLA) is also thought to involve many other factors and processes that are not evident in the acquisition of a mother tongue.

One of the most important of these factors is that most adults have some classroom instruction or at least some opportunity to reflect on the process of learning a second language (L2). (Some adults may acquire a second language simply by contact with speakers of another language, but even in this case it is unlikely that they never think about their learning experiences at all.) Second, the role of affect, attitude to learning in general, and to the culture of L2 speakers can have an influence on learning outcomes. Third, the role of language aptitude and general intelligence may play a more important role in SLA than in first language acquisition.

Finally, one of the most important differences between L1 and L2 acquisition is that the learners already have a complete linguistic system in place when they begin learning a second language.

THE ROLE OF FIRST LANGUAGE KNOWLEDGE

Background

The effects of the phonology and morphosyntax of the first language (L1) in SLA is known as 'transfer'. Transfer can both help and hinder second language development. Positive transfer occurs when principles that apply to the L1 can be readily applied to the second language, whereas negative transfer occurs when a structure or principle that applies in the L1 is incorrectly used as part of the L2 grammar. Transfer can affect all levels of the linguistic system from basic sound patterns to morphology, syntax, semantics, and social conventions of language use. Here we will focus on morphosyntax, the lexicon, and phonology.

Morphosyntax and the Lexicon

Researchers disagree substantially on the role of transfer in the acquisition of morphosyntax and the lexicon; two major positions will be briefly discussed.

The first position is that the L1 plays only a limited role in SLA. The basis for this claim is that speakers of different L1s sometimes show the same developmental patterns when they are learning a second language; in addition, it has been claimed that these patterns show some similarities to stages that are found in first language acquisition. For example, during the early stages of acquisition children learning English, whether as a first or a second language, omit the subjects of sentences, they acquire verbal agreement morphology later than other morphemes, and they learn how to ask questions and negate sentences in somewhat similar developmental stages.

Furthermore, learners make some errors that cannot be traced back to the first language influence at all, but to other kinds of learning 'strategies'. These strategies include deliberate avoidance of a difficult structure or the use of a simple L2 structure to communicate an idea that is correctly expressed with a more complex structure. Finally, some learners do not use L1 forms to express certain meanings because they believe they are unlikely to have the same meaning or use in the L2. One example is the transfer of idiomatic meanings of words from L1 to L2. For instance, in addition to its basic meaning, the word 'eye' can also refer to the hole at the top of a needle or the sprout on a potato in both English and Dutch. Nevertheless, it has been reported that Dutch-speaking learners of English are reluctant to use the word 'eye' to refer to these non-basic concepts when speaking English.

In contrast, other researchers have claimed that L1 transfer has a very important role to play in SLA. Some researchers claim that the beginning point of second language acquisition is in fact the whole of first language grammar, minus the sound representation of vocabulary items. The evidence for this claim is that at the very early stages, learners appear to attach words from the second language (L2) to what seem to be L1 sentence patterns. Only later do they begin to adopt the syntactic patterns of the L2.

In addition, when learners fail to learn a property of the second language the cause can often be traced to the fact that the L1 allows consistently more options than the L2 where that property is concerned. For example, in English one can say 'John gave a book to Mary' or 'John gave Mary a book'; in French, the second word order is not possible, and so one can say that French is a subset of English in this case. As a result of this wider range of structures, English-speaking learners of French have difficulty in learning that French does not allow the equivalent of 'John gave Mary a book'. This relationship between languages has implications for instruction that are discussed below.

Phonology

A 'foreign accent' is perhaps the most easily identifiable characteristic of the adult second language learner; indeed, it is sometimes possible to identify the first language of the learner based on his or her accent. It is therefore clear that the sound system of the first language has an effect on that of the developing second language system. L1 influence can be detected in the pronunciation of individual sounds, in syllable structure, and in intonation. However, L1 phonology does not transfer wholesale and the severity of phonological transfer may be predicted on the basis of subtle differences in the underlying sound systems, not just on the basis of whether the first language has the same sounds as the second language.

For example, it is well known that Japanese-speaking learners of English as a second language have difficulty in distinguishing between the sounds [l] and [r] in English, as in the words

'lock' and 'rock'. It was thought that the source of this problem is that Japanese lacks the sounds /l/ and /r/, but instead has one single sound which is close to both, but not the same as either one.

While this explanation may be a simple one, it may not in fact be the best or correct one. Modern theories of phonology propose that sounds can be broken down into subparts that represent the separate, but coordinated and simultaneous, movements of the vocal cords, tongue, and other parts of the mouth. Research suggests that the source of difficulty is the construction of the abstract sound system based on these articulatory movements, rather than the presence or absence of individual sounds in different languages. For example, English only has lip-rounding with vowels that are produced at the back of the mouth; in contrast, French allows lip-rounding to occur with front, central, and back vowels. English-speaking learners of French must therefore learn to produce and distinguish vowels that have lip-rounding in more environments than English does. It is important for acquirers to incorporate the feature of 'lip-rounding' as a general distinguishing feature of *the whole vowel system* in French, not just individual sounds.

Second language learners also show influence from their L1 when they combine individual speech sounds into words and sentences. For example, some languages do not allow more than one consonant at the beginning of a word. In English the words 'plate' and 'floor' both begin with a sequence of two consonants ([p] and [f] followed by [l]), but the Korean language does not allow such sequences. As a result, Korean-speaking learners of English make syllabification errors in words with word-initial consonant sequences; they insert a weak vowel [ə] between the consonants, creating two syllables. Hence, they pronounce 'floor' as 'feloor' [fəlOr], and plate as 'pelate' [pəlejt]).

Summary

In general, the L1 grammar can be shown to influence the development of the L2 grammar, but it is clear that a direct relationship between a sound or syntactic structure in the L1 and the sound or syntactic structure in the L2 does not always exist.

PSYCHOLINGUISTIC ISSUES

Input Processing, Comprehension, and Acquisition

In order to understand a sentence when we read or listen to language, it is necessary to put the words into the structure of a sentence. This structure does not only reflect a strict word-by-word order, but it has a hierarchical as well as a linear organization. For example, in the sentence 'The man who was running tripped over', the verb 'tripped' is more important to the whole clause than the verb 'running'. This importance is represented in an abstract syntactic representation of the sentence. The process of putting words into a hierarchical structure is called *parsing*.

It is important to understand parsing because it may be an essential part of the acquisition mechanism and it may help develop better ways of helping students understand the L2. Unfortunately, we still have a poor grasp of how learners process the form and meaning of input in real time and how they use the results of that processing in constructing the developing grammar.

Methods used in first language sentence-processing research have only just begun to inform the issue of input processing in second language development. Computers and eye-tracking devices are being used to measure where learners pause in reading; from these pauses researchers can infer what the process of comprehension might be, or at what point in the sentence the learners are having problems. From results of studies like these we can gain insights into second language processing.

Results of preliminary L2 research suggest that second language learners, like native speakers, construct the sentence structure of a sentence that they are reading one word at a time, revising the structure they are building as they go along. They do not read seven or eight words and then decide on a possible structure for the group of words. It is becoming clear that the influence from first language parsing mechanisms, as well as the structure of the first language, can explain why even some advanced learners fail to process the second language efficiently.

Language Production in SLA

Speech production is a highly complex process that begins with conceptualization and ends with actually making the muscular movements to pronounce words in a sentence. Research suggests that conceptualization is not specific to a given language, but that all other components of the production process have some elements unique to that language. For instance, our conceptualization of properties of objects, such as 'wetness' and knowledge of the world (e.g. rain is wet, the sun is hot) may be separate from the individual words in each

language that represent concepts and properties. Evidence for the independence of concepts from language derives from at least two sources. Where nouns are concerned it seems that humans classify things into natural kinds, which have conceptual 'prototypes' for membership. For example, when asked whether *a penguin*, *a robin*, or *an ostrich* are good examples of 'BIRD', participants in studies tend to choose *robin*, even though penguins, robins, and ostriches all have wings. Hence, the word 'bird' appears to be a collection of properties rather than a list of discrete definitions which can be analyzed in a purely linguistic way. With verbs, all humans can presumably conceptualize the placement of things on a flat surface or the placement of something tightly over another item, for example a thimble on a finger, or a top on a pen. However, individual languages tend to encode these two concepts of placement in verbs in quite specific ways that are different from one another. Hence, linguistic systems are separate from conceptual systems, but they may of course influence one another.

It has been proposed that language learners are subject to processing constraints in their second language production and that these processing constraints dictate different stages in the acquisition of second language syntax. In this theory, language structures are available, but limited processing capacity constrains production. At the initial stages, the claim is that learners have a basic word order strategy and that they are unable to reorder words during the production process. Subsequently, they are able to add material to the beginning or end of a sentence, and later are able to insert material into the basic sentence patterns of their second language. Current revisions of this theory propose that processing constraints limit the ability to match the syntactic properties of one part of a sentence with another part of the sentence. Sentences that require minimal matching of parts, with only a small number of words between the parts, will be easiest to produce. For example, the agreement between a subject and a verb is easiest when the subject is right next to the verb. As language develops, learners acquire abstract properties of words (e.g. grammatical gender, tense) and become capable of matching these abstract properties with those of another word at increasing 'distances'. These distances are defined as the boundaries between phrases and clauses in a sentence.

Summary

Second language processing and production research is still very much in its infancy. A great deal more work remains to be carried out on the interaction of memory constraints, processing, and developing grammars.

IMPLICIT AND EXPLICIT LEARNING

Psychological Background

It was noted above that second language learners may still have access to some principles that guide first language acquisition, but that they also use conscious strategies. This distinction is relevant to the difference between implicit and explicit learning. While many definitions of implicit learning (IL) exist, the one that is used most commonly in the field of SLA and in related work in cognitive psychology is the following: a process whereby complex information about a set of stimuli is acquired independently of the learner's awareness of either the acquisition process or the resulting knowledge. Explicit learning is the opposite: learning with awareness of the learning that is taking place and of the resulting knowledge. This distinction is the focus of a large, complex, and controversial body of research in cognitive psychology. It is also central to much contemporary thinking about SLA, because one's views on this distinction determine in large part both how one understands the psychology of SLA and how one thinks second/foreign languages should be learned and taught.

The psychological research on IL that is most directly relevant to SLA deals with artificial grammars. In such experiments people are shown large numbers of strings of letters; these strings follow certain patterns ('rules') that constitute a system ('grammar'). The participants in these experiments are never told that there is a system of rules underlying the strings they see, yet they are able, to some extent, to classify new strings into two groups: those that follow the pattern of the previous strings and those that do not. This seems to be evidence of implicit learning: even though people have not thought about the patterns, and even though they cannot say what the patterns are, their classification of new strings shows that, at some level, they have knowledge of the underlying patterns. It is controversial, however, to what extent learning can be both implicit and abstract at the same time. Many researchers claim that learning in adults can be implicit, but that it will then be limited to fairly concrete patterns. They also believe that learning can be abstract, but that it will then be explicit, involving awareness/consciousness of what is learned.

Empirical Research on SLA

While young children clearly learn virtually all of their native language implicitly, the situation in adults is less obvious. Clearly, many adult immigrants eventually learn to communicate well in the language of their new country, without taking classes in the language or even reflecting much on its structure. On the other hand, many people have acquired much explicit knowledge about an L2 in school (in the sense that they can tell you how verbs are conjugated, or how gender agreement between nouns and adjectives or articles work), but they cannot speak the language fluently at all. These two facts together clearly illustrate the big difference between implicit and explicit knowledge. Some educators have concluded from these facts that successful learning of a language, even in adults, *requires* implicit rather than explicit learning. Recent research of various kinds, however, casts doubt on that reasoning.

First, research on naturalistic language acquisition (outside a school context) has shown that untutored adult immigrants, while learning to communicate well, often do so through an extremely rudimentary grammatical system, even after many years in their new country, and that even those who seem to speak perfectly grammatically have intuitions about grammaticality that rarely if ever match those of native speakers.

Second, research in L2 classrooms has shown that, at least for certain elements of grammar, students need explicit teaching of grammar and error correction. This is especially the case when the second language offers fewer options than the first language (e.g. about where in the sentence to put an adverb or about how to form a question). While it is easy to notice that the second language has forms or structures that do not exist in the L1, it is impossible to notice the absence in the L2 of one of the possibilities that exist in the L1. Thus explicit teaching is required to make students aware that certain L1 options do not exist in the L2.

Third, research under laboratory conditions (with miniature linguistic systems or very restricted parts of an existing language), often through computerized instruction, has shown that explicit teaching and practising of rules is far superior to mere exposure to relevant examples without explanation of structure, even when that exposure is very extensive (thousands of examples per rule). Only when the rules are very complex or fuzzy are the rules of no help; there might even be a slight advantage for IL in such a case.

In summary, while IL of many elements of an L2 by adults is possible, and while many adults can learn the whole set of abstract grammar rules of an L2 well, it is very doubtful that any adult can successfully learn the whole set of L2 grammar rules implicitly. This conclusion is very reminiscent, of course, of the research in cognitive psychology which suggests that adults can learn implicitly and can learn abstract patterns, but not both at the same time. It also explains the traces of nonnativeness in the syntax of untutored immigrants, even those who are able to get their ideas across perfectly.

Implications for Learning and Teaching of an L2

The history of L2 teaching has shown a repeated pendulum movement between emphasis on implicit and on explicit learning. Even over the last century alone, various teaching methods have embodied one extreme or the other.

Many people remember foreign language classes from their high school or college days in which they spent substantial amounts of time learning grammar rules and vocabulary items, and translating sentences, without ever practising the language for communication (the grammar-translation method), let alone learning anything implicitly through exposure to spoken or written text in the language. On the other hand, in more recent decades, some language teachers have tried to largely leave out grammar teaching and error correction, and to make their students absorb the structure of the L2 from large amounts of exposure to (usually adapted) input (the natural approach).

Most teaching practice, however, is somewhere in between. Audiolingual methodology, especially popular in the 1950s and 1960s, while emphasizing implicit absorption through endless drills, certainly did provide a highly systematic presentation of structures for practice, and not just adapted input. In many cases, in fact, grammar rules were explicitly presented before the extensive practice. Communicative language teaching, probably the most widespread methodology world-wide from the 1970s onwards, stresses communicative practice in comprehension and production, and varies in the extent to which rules are taught explicitly and practised systematically.

Variants of all these methods are still widely used in educational settings throughout the world. To some extent this variation reflects continuing debate in the language teaching profession

about the roles of implicit and explicit knowledge, but other, and often more important, factors are the training that the teacher received (sometimes decades earlier), the methodology endorsed by widely available textbooks, time constraints, and individual characteristics of the students. The last two factors deserve some further explanation. Time constraints often work against teaching methodologies that make sense from the point of view of research on second language acquisition because these methodologies are only efficient in the long run. Even those educators who are strong advocates of explicit grammar teaching and systematic practice will admit that this approach cannot lead to full proficiency after just a couple of years of study in high school or college, and that a less rigid approach with more time for authentic materials and fun communicative activities can seem more attractive under such conditions. Individual characteristics play a role in the sense that only students with a certain level of language learning aptitude (strongly related to verbal intelligence) can be expected to benefit maximally from a method that requires them to learn and use explicit rules. Therefore, in cases where all students of a certain age (are required to) take L2 classes, a teaching approach focused on the absorption of vocabulary and common phrases through extensive communicative use may be more efficient than an approach based on the systematic learning of explicitly taught rules. One should keep in mind, however, that while the more implicit methodologies may be the most *efficient* under those circumstances, they can by no means be maximally *effective* in bringing about high levels of both fluency and accuracy in spontaneous communication. Adults do need explicit learning to reach high levels of grammatical accuracy, as was argued above.

The need for explicit learning does not automatically entail the need for explicit teaching. Besides taking classes, adult learners have two other options for explicit learning: formal self-study by means of textbooks, phrasebooks, grammars, and dictionaries, and informal self-study by figuring out structural patterns in the input available from interaction with native speakers, recordings, or texts. From a psycholinguistic point of view, this institutional versus naturalistic distinction does not matter; in all these cases the learning process is explicit, because the learner is aware of both the learning process and the resulting knowledge of L2 structure.

Also, explicit learning does not require the use of extensive grammatical terminology or a traditional syllabus organized structure by structure. The crucial factor is whether the learner somehow pays attention to the structure of the language; in current professional literature this is often called 'focus on form'.

AUTOMATIZATION

Psychological Background

A behavior is said to be automatic when it requires little effort or attention, is carried out with high speed and a low error rate, and is hard to suppress consciously. In many areas of everyday life we engage in behaviors that have become automatic as a result of extensive practice: driving a car, typing a text, using our favorite piece of software. In all these cases our initial behavior may have been slow, effortful, prone to error, and very attention-demanding, but as a result of years of practice these behaviors have become 'second nature'. This gradual process of making a behavior more automatic through extensive practice is called automatization.

Some psychologists conceive of this process as the creation and fine-tuning of highly specialized rules for behavior. These are called production rules; they take the form 'if X and Y is the case, then do Z', and are created when people repeatedly engage in the relevant behavior while all the relevant knowledge is easily accessible in long-term memory. This stage of rule creation (often called proceduralization or production compilation) is followed by a long period of automatization in a more narrow sense. The latter involves the gradual decrease in reaction time and error rate that follows from extensive practice of these production rules. The reason why automatic behavior is so efficient is that these highly specific behavioral rules do not require interaction with elements stored in various parts of long-term memory; they are ready-made 'chunks' of conditions and commands, which can be retrieved and carried out in a single step. Other psychologists see automatic behavior as the retrieval from long-term memory of complex instances of past experience, which serve as the perfect model for action in the present. They all agree, however, that the gradual process of automatization follows a specific learning curve, which is often referred to as 'the power law of practise', and which has been shown to apply to the acquisition of a wide variety of cognitive skills, from making cigars to writing computer programs.

Automatization in SLA

There is little detailed empirical research on automatization processes in SLA, but the available

research suggests that automatization of second language grammar rules follows the same patterns as in other cognitive domains and that this process must, at least in part, involve the increasingly efficient use of specialized production rules. There is also evidence from reaction time measurements that even highly proficient bilinguals tend to be somewhat faster at vocabulary retrieval in one language than in the other, which reflects the current view of psychologists that automatization always remains a matter of degree.

Even though limited in size and scope, this SLA research and the related work in cognitive psychology have two implications for language teaching. When practising an individual grammar point, learners should be repeatedly engaged in a communicative behavior that draws on that grammar point, while the relevant grammatical knowledge is easily accessible in their minds. This is a precondition for the formation of production rules, which can then be gradually automatized through further practice. From the point of view of the curriculum as a whole, the challenge for language teaching is to integrate systematic practice of linguistic structures with a gradual sequencing of communicative tasks that will ensure increased accuracy, fluency, and complexity of expression.

Further Reading

Archibald J (ed.) (2000) *Second Language Acquisition and Linguistic Theory*. Oxford, UK: Blackwell.

Cleeremans A, Destrebecqz A and Boyer M (1998) Implicit learning: news from the front. *Trends in Cognitive Sciences* **2**(10): 406–416.

DeKeyser RM (2001) Automaticity and automatization. In: Robinson P (ed.) *Cognition and Second Language Instruction*. New York, NY: Cambridge University Press.

Gass SM and Selinker L (2000) *Second Language Acquisition: An Introductory Course*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Levelt W (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Lightbown P and Spada N (1999) *How Languages are Learned*, 2nd edn. Oxford, UK: Oxford University Press.

Pienemann M (1998) *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Segalowitz N (2001) Automaticity and second language acquisition. In: Doughty C and Long M (eds) *Handbook of Second Language Acquisition*. Oxford, UK: Blackwell.

Skehan P (1998) *A Cognitive Approach to Language Learning*. Oxford, UK: Oxford University Press.

Stadler MA and Frensch PA (1998) *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage.

Second Language Learning and Instruction

Advanced article

Barbara Graves, University of Ottawa, Ottawa, Ontario, Canada

CONTENTS

Our understanding of language development and second-language learning has been narrowly conceived and has underestimated the interrelated and dialectical processes involved.

INTRODUCTION

Despite general agreement that reading, writing and speaking a second language involve much more than the mastery of vocabulary and syntax, little attention has been directed towards understanding the sociocultural contexts of learning and the discursive practices that occur in classrooms and communities. In particular researchers have neglected the agency, identity, and voice of second-language learners. In part, this has been the result of a research tradition that adheres to cognitively oriented theories to explain second-language learning. In order to investigate the engagement with a second language as contextual, and to explore connections among knowledge, agency and identity construction from the perspectives of the learners themselves, researchers have argued for a reformulation of theory to investigate what it means to become literate in a second language (Cummins, 1996; Hall, 1997; Rampton, 1997; Lantolf, 2000).

While research in second-language acquisition has been informed from many disciplinary traditions including linguistics, psychology, anthropology, sociology and education, it has also experienced several distinct research phases, each of which can be linked to a specific psychological learning theory. These theoretical orientations can be characterized as behavioral, cognitive and (most recently) sociocultural. Each has had specific consequences for the research and teaching of second languages.

THREE THEORETICAL ORIENTATIONS

The Behavioral Orientation

Beginning in the 1950s, the dominant learning model informing second-language research was based on behavioral learning theory, which understood learning as a series of stimulus–response associations. An important tenet of this approach was that stimulus–response connections that are repeated become strengthened, and so learning became the formation of learned habits. The teaching method associated with this perspective for second language was the audiolingual method which, in adherence to the theory, focused on drill, practice and rote memorization of grammar and syntactic structures. Theories of learning framed by behavioral perspectives supported a teaching model based on the transmission of information from the more knowledgeable person (the teacher) to the less knowledgeable (the learner). Under this model knowledge was acquired incrementally from the simple to the complex.

Although influential, the behavioral approach was challenged by cognitive theorists who argued that such approaches could not account for the generative properties of human learning. In linguistics, Chomsky successfully argued that repetition and imitation could not account for the generative and problem-solving capabilities of language learning in particular.

The Cognitive Orientation

Cognitive psychologists sought to understand cognitive functioning and to that end focused on the mind and mental processes, not simply on observable behaviors. Concepts of importance within this

approach relate to learning as information processing. In the 1960s cognitive research began to reveal the mind as an information processing device actively engaged in the generation, storage and retrieval of knowledge. With the focus on knowledge, its acquisition and use, much attention was paid to how knowledge was organized in memory. One of the most robust findings from this research pertains to the role of prior knowledge in the learning process. Repeatedly it was found that learning was more effective when it could be meaningfully associated by the learner with information previously learned. The mental processes studied included perception, attention, memory, reasoning and problem-solving. These processes were studied without considering context, and the goal was to uncover universal principles of human performance. Instructional applications associated with cognitive research provided learning materials and activities with which learners could actively construct their understanding through exploration and problem-solving.

The cognitive approach conceptualizes learning a second language as an individual, psychological process, in which the means to produce and understand a second language are to be found in the learner. Thus much of the research investigated psychological properties such as motivation, ability, attitude, and cognitive learning style in relation to language learning. Measures of language proficiency were correlated with identified learner traits, as successful language learners were compared with those less successful in order to identify the qualities that would account for their success. From this perspective language learners were often seen as 'deficient' in relation to their second-language knowledge. Language itself was viewed as a formal system organized as a hierarchy of processing levels including the phonological, lexicomorphological, syntactic, and pragmatic.

While the contribution from cognitive science research has been substantial, this approach is wanting since it does not account for the context of mental activities. In addition, much of the cognitive research agenda rested on assumptions entrenched in binary oppositions such as mind–body, cognition–emotion, individual–social, and recognition–interpretation, which restricted an examination of the dynamic interactions in human activities.

The Sociocultural Orientation

In order to investigate how second-language learners make meaning, build knowledge through discursive interactions, and develop identities in the process, researchers are turning increasingly to theoretical perspectives located within a constructivist, sociocultural framework. The sociocultural orientation challenges traditional assumptions about cognition and human nature, and replaces the binary oppositions that have structured our thinking and activities for so long with a theory of cognition that emphasizes the mutual constitutions of persons and the experienced world. The sociocultural orientation which embraces the concept of the mind mediated by language and the role of discourse in knowledge-building and identity construction, draws on the theoretical contributions of Vygotsky's sociocultural theory of human development (Vygotsky, 1978, 1986), Halliday's view of language as social semiotic (Halliday, 1993), and Bakhtin's theory of discourse and the self (Bakhtin, 1986). Within the cognitive science tradition, researchers interested in exploring the situated nature of cognitive activities have developed a theory of situated cognition that integrates aspects of cognitive science with anthropological and cultural traditions.

Sociocultural theory views learning as a situated activity acquired from meaningful participation in specific communities. From this perspective, learners construct knowledge in relational networks which emerge from the interactions of people and activity contexts. Knowledge and inquiry processes are thus considered to be social activities, and learning of all kinds is mediated by cultural artifacts and resources, both symbolic (e.g. language, and numeracy systems) and material (e.g. computers). This shifts the traditional view of learning from the acquisition of something to a formulation of learning as participatory. In this way, learning in classrooms is viewed as a set of social practices which are learned by participants and understood through participation in language and multiple activities in specific contexts. This has led to classroom practices that provide learners with more opportunities for communicative interaction and discussion.

While Vygotsky writes of mediational means including both material and symbolic resources (Vygotsky, 1986), he focused much of his empirical research specifically on the examination of the role of language as a central mechanism of learning. Similarly to Vygotsky Bakhtin (1986) wrote about the individual's appropriation of language experienced in a world mediated by social texts, and his theory – which examines how the self is formed in dialogic response to a social world – bridges the divide between the individual and the social self.

Bakhtin maintained that all spoken or written language is dialogical since it is always addressed to someone. In addition, it is always delivered from a particular viewpoint. Taken together, these theoretical perspectives create a valuable explanatory framework for examining how learning takes place, how learners make meaning, what counts as knowledge, and whose knowledge and voices are recognized and valued.

DEVELOPMENT AND LEARNING

Traditionally, development and learning have been treated as occurring independently of one another, but sociocultural theorists see them as integral parts of a dynamic process. In particular, Vygotsky's theory of human development maintains that complex psychological actions such as reasoning and memory develop first in social interaction with others and only later become part of an individual's psychology. Learning and development both occur as a function of participating in meaningful activities. The sources of knowledge are from other participants as well as the mediational means which are part of the activities. Vygotsky formulated these views in terms of the zone of proximal development (ZPD), which he understood as

> the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers (Vygotsky, 1978, p. 86).

Simply stated, what a learner can do today with the help of another, a learner may do tomorrow independently. Because the individual and the social worlds are mutually constitutive of each other, transformation of the learner also involves transformation of the communities and of the joint activities. Activities situated in time and place may share features, but each activity remains unique owing to the participation of particular individuals with their own lived experiences and values. Consequently, outcomes of an activity cannot be completely specified in advance, nor can we predict with any certainty the the extent of the knowing as this ultimately depends on the characteristics of the learner as well as on the characteristics of the interaction.

NEW DIRECTIONS FOR SECOND-LANGUAGE RESEARCH

Since the 1980s there has been a substantial increase in second-language research focusing on pedagogical questions and the situated contexts in which they occur. Many learners first encounter a second language in the classroom, and it is important to understand the ways in which the discursive practices of teachers and learners create a view of what it means to learn a second language. The sociocultural perspective creates a valuable explanatory framework for examining the activities, voices, and multiple discourses in second-language classroom practice, as well as investigating how this community of practice is understood and valued by teachers and learners.

Agency and Biliteracy

A 3-year ethnographic research study of multilingual children as they learned English in a culturally diverse primary school in urban Montreal examined children's writing from an emic perspective (Maguire and Graves, 2002). This research challenged the epistemologies and research practices underlying many classroom-based studies which have ignored the possibility that bilingual children can be agents of cultural knowledge and self-monitoring in discursive activities. Beginning with pupils in the first grade, the researchers collected the journal writing entries of the children along with interview data. In the analysis of these data collected over 3 years, the children revealed competencies far beyond those normally assessed in many second-language classrooms. The children in these classrooms were given ample opportunity to express their views and to connect school to home, and the contribution that they brought to the classroom was highly valued. By grade three these multilingual writers were able to express their opinions, give reasons, explain, joke, and adopt a fictitious persona while writing in a third language. It appears that we have a great deal to learn about how much young learners are capable of when given this type of learning opportunity and validation.

The Good Language Learner

A qualitative study investigating the concept of the 'good' language learner from a sociocultural perspective (Norton and Toohey, 2000) concluded that the learners' successful second-language performance was not simply the result of personal abilities and strategies, but was related to the way they successfully negotiated access to the social networks in their respective learning communities and how this access transformed their identities as second-language learners. This research

suggests that it is necessary to understand what possibilities communities offer participants and how learners are able to partake in those activities. The researchers conclude that to understand what can be learned by whom, we need to understand what discursive practices and values are present in the culture and to explore how these interact with the lived experiences of the learners.

Creating Communities of Learners

A sociocultural investigation of two first-year high-school Spanish language classrooms (Hall, 1997) presents a comparative analysis of the communicative practices of the two classrooms. While the researcher found apparently similar resources in both contexts, the way in which these resources were understood and deployed created substantially different communities of learners. On one level the practices of 'reviewing homework' and 'practicing vocabulary' looked the same, but after detailed analysis of the turn-taking patterns and the ways in which students paid attention in the two settings, it became apparent that a different learner emerged in the two contexts. In addition, this was accompanied by a different understanding of what constitutes the appropriate use of Spanish in the classroom.

The findings from these studies in combination with the theoretical assumptions of sociocultural perspectives suggest that our understanding of language, development and second-language learning has been narrowly conceived and has seriously underestimated the interrelated and dialectical processes that connect many of these dimensions.

References

Bakhtin MM (1986) *Speech Genres and Other Late Essays*, translated by Y. McGee (Emerson C and Holquist M, eds). Austin: University of Texas Press.

Cummins J (1996) *Negotiating Identities: Education for Empowerment in a Diverse Society*. Ontario, CA: California Association for Bilingual Education.

Hall JK (1997) A consideration of SLA as a theory of practice: a response to Firth and Wagner. *Modern Language Journal* **81**(3): 301–306.

Halliday MAK (1993) Towards a language-based theory of learning. *Linguistics and Education* **5**: 93–116.

Lantolf J (ed.) (2000) *Sociocultural Theory and Second Language Learning: Recent Advances*. Oxford: Oxford University Press.

Maguire MH and Graves B (2002) Speaking personalities in primary school children's second language writing. *TESOL Quarterly* **35**(4), 561–593.

Norton B and Toohey K (2000) Changing perspectives on good language learners. *TESOL Quarterly* **35**(2): 307–322.

Rampton B (1997) A sociolinguistic perspective on L2 communication strategies. In: Kasper G and Kellerman E (eds) *Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives*. London: Longman.

Vygotsky LS (1978) *Mind in Society: The Development of Higher Psychological Processes* (Cole M, John-Steiner V, Scribner S and Souberman E, eds). Cambridge, MA: Harvard University Press.

Vygotsky LS (1986) *Thought and Language*. Cambridge, MA: MIT Press.

Further Reading

Bialystok E (2001) *Bilingualism in Development: Language, Literacy, and Cognition*. New York: Cambridge University Press.

Duff P and Uchida Y (1997) The negotiation of teachers' sociocultural identities and practices in post secondary EFL classrooms. *TESOL Quarterly* **31**(3): 451–487.

Firth A and Wagner J (1977) On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal* **81**(3): 285–300.

Hakuta K and McLaughlin B (1996) Bilingualism and second language learning: seven tensions that define the research. In: Berliner DC and Calfee RC (eds) *Handbook of Educational Psychology*, pp. 603–621. New York: Prentice-Hall.

Kramsch C (2000) Social discursive constructions of self in L2 learning. In: Lantolf J (ed.) *Sociocultural Theory and Second Language Learning: Recent Advances*, pp. 133–153. Oxford: Oxford University Press.

Lantolf JP (1996) SLA theory building: letting all the flowers bloom. *Language Learning* **46**(4): 713–749.

Lantolf JP and Appel G (eds) (1994) *Vygotskian Approaches to Second Language Research*. Norwood, NJ: Ablex.

Lightbown PM (2000) Anniversary article: classroom SLA research and second language teaching. *Applied Linguistics* **21**(4): 431–462.

McKay SL and Wong SC (1996) Multiple discourses, multiple identities: investment and agency in second-language learning among Chinese adolescent immigrant students. *Harvard Educational Review* **66**(3): 577–608.

Moll LC and Dworin J (1996) Biliteracy development in classrooms: social dynamics and cultural possibilities. In: Hicks D (ed.) *Discourse, Learning, and Schooling*, pp. 221–245. Cambridge, UK: Cambridge University Press.

Peirce BN (1995) Social identity, investment, and language learning. *TESOL Quarterly* **29**(1): 9–18.

Semantics and Cognition

Intermediate article

Cliff Goddard, University of New England, Armidale, New South Wales, Australia
Anna Wierzbicka, Australian National University, Canberra, Australia

CONTENTS

The words and grammar of any language encode a vast array of prepackaged concepts, most of them complex and culture-related. Since language plays an important role in normal human cognition, the nature and extent of semantic variation across languages is a key research question for cognitive science.

INTRODUCTION

There are, broadly speaking, two traditions of semantics: the linguistic or conceptual tradition, which sees meaning as a cognitive phenomenon, and the logical or formal tradition, which sees meaning in terms of correspondences (truth-conditions) with an objective reality. This article adopts the linguistic or conceptual perspective.

Many aspects of human cognition, such as basic perception, attention, and visual processing, are substantially shared with other primates. Language is primarily relevant to higher-order cognitive processes which are largely, if not entirely, species-unique. Importantly, human cognition not only includes reasoning and information processing about physical reality, it also includes so-called social cognition (Tomasello, 1999), that is, assessing and reasoning about intentions, mental states, and social situations, and it is in this arena that language has some of its clearest cognitive effects. (*See* **Categorial Grammar and Formal Semantics**; **Social Cognition**)

The words and grammar of any language express 'prepackaged' concepts – concepts which are largely acquired in infancy and which are intersubjectively shared among members of the speech community. While higher-order thinking need not be conducted exclusively in terms of linguistic concepts, there can be no doubt that language plays a substantial role in normal cognition (including in categorization, planning, problem-solving, and memory). It is, therefore, a crucial fact that languages differ considerably in their conceptual semantics, in both lexicon and grammar. Of equal interest to cognitive science is the possibility that there may be language-universal conceptual categories embedded in human languages.

The relationship between semantics, culture, and cognition, often discussed under the rubric of 'linguistic relativity', has a long history in Western thought. Beginning in classical times it proceeds through John Locke, Johann Gottfried Herder, and Wilhelm von Humboldt into the early twentieth century, in which period Franz Boas, Edward Sapir, and Benjamin Lee Whorf are key names. The topic experienced a malaise from the 1950s until a revival of interest in the 1990s, coinciding with advances in contrastive psycholinguistics, cognitive linguistics, and cultural psychology. (*See* **Linguistic Relativity**)

MEANING AND MENTAL REPRESENTATION

Issues

The basic arguments for the involvement of language-specific semantics in cognitive processes are simple and compelling. In order to speak using the lexical categories and observing the grammatical distinctions of any language, speakers must attend to and manipulate a large number of language-specific conceptual distinctions. Thus, at the very least there must be a mode of cognitive processing (Dan Slobin's 'thinking for speaking'), which dovetails with linguistic concepts (Slobin, 1996). Viewed from another angle, any language can be thought of as a 'tool for thinking', inasmuch as its lexical and grammatical categories provide speakers with a vast array of ready-made concepts. The packaging of complex concepts into words enables intricate manipulations to be undertaken which would be impossible without conceptual 'chunking'. From the developmental perspective,

there is mounting evidence that language acquisition does not merely reflect conceptual development but contributes to it in complex ways. As a shared system of cultural representations, language is one of the main instruments by which children are socialized into the values, beliefs, and practices – including thinking practices – of their culture. (*See* **Semantics, Acquisition of**; **Language and Cognition**)

The intimate connection between language and cognition poses special conceptual and methodological problems. If one takes the view that language and cognition are in principle separable, then it seems obvious that linguistic and cognitive processes must be studied independently if one is to explore the relationship between them. Whorf's failure to observe this requirement is a standard critique of his celebrated work on the conceptual semantics of the Hopi language. An alternative, and perhaps ascendant, view rejects the assumption that higher-order cognition can be separated from language and studied independently, maintaining that language is both partially constitutive of higher-order cognition and epistemologically essential to any inquiry into it (Enfield, 2000). In any case, in order to describe and compare language-specific conceptual systems, no matter how divergent they may be, one needs a common measure, a *tertium comparationis*. The nature of any such conceptual *tertium comparationis* is a fundamental methodological issue.

Leading Current Approaches

Current approaches to semantic representation differ on several dimensions. At a theoretical level, one may distinguish propositional systems, whose fundamental units are discrete and word-like, from systems with scalar or image-like representations. Degree of abstraction is another theoretically interesting characteristic. One may also distinguish between systems which are partial or domain-specific, as opposed to those which are intended to be comprehensive, and between systems which have been devised from and tested only on English, as opposed to those which have been subject to significant cross-linguistic testing.

One point of agreement is that new models of mental representation are called for, beyond the binary feature analysis of traditional structural semantics and cognitive anthropology. There is wide agreement that many concepts incorporate 'script-like' structures which include stages arranged in a temporal sequence, and prototypical scenarios or schemas, also known as idealized cognitive models. Some approaches see metaphor and metonymy as fundamental to normal linguistic thinking. (*See* **Distinctive Feature Theory**; **Metaphor**)

Key features of two leading approaches can be summarized as follows. Both are propositional systems. From the point of view of their relevance to cognition, it is useful to highlight different conceptions about the fundamental semantic units – for these embody claims about fundamental elements of cognition. (*See* **Lexical Semantics**; **Cognitive Linguistics**)

The *Natural Semantic Metalanguage* (NSM) model has been developed over several decades by Anna Wierzbicka and colleagues (Wierzbicka, 1996; Goddard and Wierzbicka, 2002). It is based on the Leibnizian idea that a set of universal and indefinable concepts (semantic primes) can be discovered by intensive analysis of natural language. After numerous descriptive and analytical studies across a wide range of languages, the current NSM model includes about 60 concepts (see Table 1) which, it is claimed, constitute the semantic bedrock of human cognition and communication. Along with certain combinatorial properties, also held to be universal, semantic primes constitute a kind of mini-language of simple concepts in terms of which all other linguistic concepts can be explicated. A further key claim is that all languages can be expected to have concrete exponents – words, bound morphemes, or fixed expressions – for expressing the precise meanings of all the proposed semantic primes.

The NSM primes represent a rich set of hypotheses about many areas of conceptual representation. Although the NSM theory remains controversial, it has the longest and strongest track record in cross-linguistic semantics on the contemporary scene.

Conceptual Semantics (CS) has been developed primarily by Ray Jackendoff (1990), the major semantic theorist in Chomskyan generative grammar. Conceptual Semantics sees all possible word meanings (lexical concepts) as built up from a finite set of conceptual primitives and principles of combination, but unlike the NSM system, these elements and principles are held to be highly abstract. Doubting that any conceptual primitive could correspond to the meaning of an ordinary word, Jackendoff's system necessarily consists of an extensive formalism.

It is not feasible to list exhaustively CS semantic functions, but some impression of its character can be gained from example (1) below. This is intended to represent a rule of inference governing people's

thinking about rights and obligations (Jackendoff, 1999), namely, that if one person (X) has an obligation to another person (Z) and fails to fulfil it in a reasonable time, then Z has an 'existential right' to impose a roughly proportionate punishment. The primitive operators RT and OB are roughly equivalent to the English modals 'may' and 'must'. ACT stands for Action. VALUE maps two arguments, a Stimulus and an Experiencer, onto a Value – either positive or negative. The EXCH operator says that the action is in exchange for X's nonperformance.

$$\begin{aligned}&\text{HAVE } (X\alpha, \text{OB } (\text{ACT}_1\ (\alpha), \text{TO } Z)) \text{ at } t_1 \text{ and}\\&\text{NOT } \text{ACT}_1\ (X) \text{ in period from at } t_1 \text{ to } t_2\\&\quad \text{entails}\\&\text{HAVE}\\&\left(Z\beta, \text{RT}_{\text{Ex}}\left(\left[\begin{array}{l}\text{ACT}_2\ (\beta)\\ \lambda a\ (\text{VALUE } (a, X) = -)\\ \text{EXCH } (\text{NOT } \text{ACT}_1(X))\end{array}\right]\right)\right) \text{at } t_2\end{aligned} \tag{1}$$

CS has shown a tendency to evolve in the direction of increasing abstractness. For example, Jackendoff's (1990) treatment of motion employed conceptual functions such as [$_{\text{Path}}$TO (PLACE)], but within a few years the TO-function was further decomposed into a more abstract representation of features and functions. Relatively little cross-linguistic work has so far been conducted within the CS framework.

CONCEPTUAL CATEGORIES IN THE LEXICON

The cognitive influence of the lexicon is sometimes downplayed on the grounds that speakers may exercise a conscious choice over a range of lexical options. It is also said that although many lexical items, for example terms for kinship, material culture, and social institutions, designate culture-specific concepts, this fact is obvious and unlikely to influence sophisticated thinking. In reality, however, many culture-specific words are nonobvious, including words for emotions and sensations, for virtues and vices, for speech acts, and for social categories. Examples will be given shortly, but before that it is useful to review some other lexical matters of relevance to cognition.

Lexical elaboration designates the situation of a language having an impressively large number of words within a particular semantic domain, thereby providing the terminological scaffolding for fine conceptual discrimination within the domain. Lexical elaboration is often reflective of sociocultural facts. For example, compared with most indigenous cultures, European languages have a large stock of expressions to do with measuring and reckoning time (words such as *clock*, *calendar*, *date*, *second*, *minute*, *hour*, *week*, *Monday*, *Tuesday*, etc., *January*, *February*, etc.). Australian Aboriginal languages are celebrated for their profusion of kinship terms, amounting to a so-called algebra or calculus of kinship. Lexical elaboration is not necessarily apparent to native speakers.

Lexicalization pattern refers to a characteristic pattern of semantic packaging within an area of the lexicon. A well-known example concerns verbs of motion. Typically, English lexically encodes manner-of-motion (*run*, *walk*, *creep*, *crawl*, *limp*, *jump*, *fly*, *dash*, *climb*, *clamber*, etc.), but to further encode the path (direction, etc.) one needs to add

Table 1. Proposed NSM semantic primes

Substantives and relational substantives	I, YOU, SOMEONE, PEOPLE, SOMETHING(THING), BODY
	KIND OF, PART OF
Determiners	THIS, THE SAME, OTHER
Quantifiers	ONE, TWO, SOME, ALL, MANY/MUCH
Attributes	GOOD, BAD, BIG, SMALL
Intensifier	VERY
Mental predicates	THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech	SAY, WORD, TRUE
Actions and events	DO, HAPPEN, MOVE
Existence and possession	THERE IS, HAVE
Life and death	LIVE, DIE
Logical concepts	NOT, MAYBE, BECAUSE, CAN, IF
Time	WHEN(TIME), NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME
Space	WHERE(PLACE), HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE, TOUCH(CONTACT)
Augmentor	MORE
Similarity	LIKE

Source: after Goddard and Wierzbicka (2002).

an additional adverb (*away*, *out*, etc.) or prepositional phrase (*into the room*, etc.). Many languages follow the English pattern while many others – including Japanese, Turkish, Hebrew, and most of the Romance languages of Europe – follow the converse pattern. They have many verbs coding a path specification (akin to English words such as *enter*, *exit*, *ascend*, which have been borrowed from Romance languages), but to further express manner-of-motion one needs to add additional description. Evidence from a variety of measures – description tasks, composition, recall with visualization – suggests that preferred lexicalization pattern predisposes speakers to conceptualize differently motion events (Slobin, 2000). (*See* **Spatial Language**)

Cultural key words designate salient concepts which serve as conceptual focal points for an entire complex of culture-specific values, attitudes, and beliefs. Such terms are usually resistant to translation, but under close semantic study they can be extremely revealing. Two examples will suffice. The Russian word *sud'ba* (roughly 'fate', 'destiny') is used very widely, in very different registers, from colloquial speech to scholarly discourse to literary works. Corpus counts show that it is far more frequent than words such as 'fate' and 'destiny' are in English, being involved in numerous collocations, set phrases, and proverbs. It can be argued that the *sud'ba* concept encapsulates a characteristically Russian way of looking at human life – as an incomprehensible experience which, in the words of one classic source (Vladimir Solov'ev), is at the mercy of 'some overwhelming necessity to which we must submit'.

Scholars of Japan agree that *wa*, usually translated as 'social harmony', 'concord', 'peace', or 'unity', is one of the key Japanese social ideals. As one authority (Rohlen, 1974) has written: 'To achieve *wa* is certainly a major goal for any Japanese group, and it also is an essential ingredient in the attainment of other goals. In this regard, it is something like "love" in American popular culture, for it is both a major means to social improvement and an end in itself.' Thus, *wa* has clear implications of the value placed in Japanese culture on 'groupism' and, conversely, on something like 'anti-individualism'. Forged by collective effort, *wa* offers not only the comforting feeling of group unity and absence of conflict, but also the promise of group achievement.

Culture-specific concepts like *sud'ba* and *wa*, which are everyday 'common-sense' concepts in their home cultures, enable and promote patterns of thinking which would be otherwise impossible to carry out in a routine fashion.

Ethnopsychological categories hold particular interest for cognitive science, both because of their role within indigenous folk psychologies and because of their possible relevance to the agenda of cognitive science itself. This can be demonstrated with the aid of a thumbnail study in contrastive semantics – English *mind* versus Russian *duša*. The English word *mind* lacks precise equivalents in most of the world's languages, even European languages such as French and German. When Descartes argued for 'mind–body' dualism, for example, he was opposing the word *corps* 'body' to *âme*, a word with a significantly richer meaning than modern English *mind*. Similarly, Freud's primary concept was *die Seele* (roughly, 'soul') and to translate *Seele* as 'mind' is significantly to distort Freud's thinking. Arguably, the modern English concept of *mind* can be analyzed as in (2a) below (Wierzbicka, 1992). This explication reflects, firstly, the bifurcation of the person into two parts, the *mind* being the invisible and immaterial part, and secondly, the fact that *mind* is focused on thinking and knowing, rather than on feeling, wanting, or any other non-bodily processes. (To say that someone has 'a good mind' suggests that a person can *think* well, rather than, say, the emotional and moral qualities suggested by the phrase 'a good heart'.) (*See* **Cultural Psychology**)

Russian *duša*, variously rendered in English as 'soul', 'heart', and 'spirit', is a much broader concept. It is one of the leitmotifs of Russian literature and of everyday Russian conversation. Whereas *mind* is linked with rational functions, *duša* is linked, above all, with feelings, and especially those with a certain profundity or spiritual quality: the *duša* endows a person with moral capabilities. The *duša*, furthermore, is viewed as a kind of internal spiritual theater – a place where events happen which are in principle unknowable to outsiders. But though, as the proverb says, *čužaja duša potëmki* 'another person's *duša* is unfathomable', in Russian culture readiness to open and to 'pour out' one's *duša* is seen as important and good. Cognitive functions are not altogether excluded: one can know or say things *v duše* 'in one's heart' so long as these things are linked with values and feelings. The human will, too, is included in the domain of *duša*, as can be seen in the expression *duševnaja sila* 'spiritual strength, strength of character'. These and other considerations suggest a semantic structure as in (2b).

a. *mind* =
one of two parts of a person
people cannot see it, people cannot touch it
because of this part, a person can think
because of this part, a person can know things

b. *duša* (Russian) =
one of two parts of a person
people cannot see it, people cannot touch it
because of this part, many things can happen inside a person
these things can be good, these things can be bad
because of this part, a person can feel many things
other people cannot know what these things are if the person does not say it
it is good if a person wants someone else to know what these things are
because of this part, a person can be a good person (2)

A complex ethnopsychological concept such as *mind* or *duša* embodies an 'ethnotheory of the person'. Both ethnotheories are basically dualistic, but whereas the Russian lexicon opposes the body to a psychological entity which is unpredictable, emotional, spiritual, expressive, and moral, in English the basic dualism is focused on intellectual and rational aspects.

An important body of cross-linguistic semantic research (e.g. Harkins and Wierzbicka, 2001) concerns emotion terminology. One major finding is that despite the existence of cross-culturally recurrent themes, emotion lexemes differ significantly in their semantic structure across languages. Even salient basic-level categories of the English emotion lexicon, such as 'anger' and 'sadness', lack precise equivalents in many languages. For example, Yankunytjatjara (Central Australia) has three main translation equivalents for English *angry*, none of which is sufficiently general to serve as a precise equivalent: *pikaringanyi* is associated with active hostility, *mirpa<u>n</u>arinyi* with a sense of grievance, and *kuyaringanyi* with resentfulness. Another language, Malay, has a single basic word – *marah* – that is usually translated as 'angry' but really designates a more specific concept, closer in some ways to English 'offended'. It can be shown, furthermore, that many aspects of emotion terminology are significantly culture-related.

Clearly it would be invalid to assume that English lexical categories (be they ethnopsychological entities such as 'mind', or emotion categories such as 'anger' and 'sadness') transparently reflect psychological reality whereas those of other languages are misleading. One technique for minimizing the danger of such terminological ethnocentrism is to frame hypotheses about mental states and processes in terms of semantically prime meanings such as 'think', 'know', 'want', and 'feel', which do appear to have precise equivalents in all or most languages.

CONCEPTUAL CATEGORIES IN GRAMMAR

Theoretical linguistics in the second half of the twentieth century was largely preoccupied with the pursuit of abstract, innate universals of syntactic structure. Semantic aspects of grammar were of marginal interest until the last 15 years or so of the century when interest revived, largely on account of increasing evidence that lexical semantics plays an important explanatory role in syntax. Only a relatively few scholars in psycholinguistics and contrastive semantics have concentrated on investigating the conceptual content of grammar in a cross-linguistic and cultural perspective. Psycholinguistic research into correlations between grammar and cognitive processing has concentrated on domains which lend themselves to objective reference-based testing, such as categorization of objects and spatial orientation. Contrastive semantic analysis has concentrated on social and interpersonal domains. Some highly abbreviated examples follow.

Categorization in Yucatec Maya

John Lucy (1992) found significant differences in the way in which adult speakers of English and of Yucatec (a Mayan language of southeastern Mexico) process information about concrete objects, on a variety of sorting, similarity judgment, memory, and grouping tasks. English speakers show greater attention to number than Yucatec speakers and tend to classify by shape, while Yucatec speakers tend to classify by material composition. These differences correspond to what could be predicted on the basis of grammatical differences between the two languages. In English, number marking (via singular versus plural suffixes) is obligatory for most nouns, whereas in Yucatec it is often optional (*yàan pèek' té'elo'* 'there-is dog over-there' can be used regardless of the number of dogs). On the other hand, numeral classifier constructions, for example *ká'a-túul 'úulum* 'two-classifier turkey', are obligatory whenever a noun referent is quantified and force attention to referential categories.

English Interpersonal Causatives

Compared even with other European languages, English has a wealth of analytic causatives, that is, grammatical constructions which encode causative scenarios by means of an auxiliary verb. In many languages, the subtle differences between 'She had him do it', 'She made him do it', 'She got him to do it', and 'She let him do it' cannot be expressed in a compact grammatical form but would require several sentences of explanation. The *have*-causative, for example, implies some kind of hierarchical relationship such that the causee's readiness to take directions can be assumed; these directions, furthermore, do not have to be expressed directly but can be conveyed via another person. The *make*-causative implies that the causee does something unwillingly, in response to some kind of pressure (threats, parental authority, nagging, etc.). The causee's will is not completely overridden (compare 'She made him do it' and 'She forced him to do it'); rather, the implicit scenario is that the causee comes to realize that he or she has no choice but to act as the causer wants. The *get*-causative implies that the causee carries out the desired action willingly but only because the causer has done something to bring this about. Though these constructions are sometimes given labels such as 'coercive', 'manipulative', and so on, detailed analysis shows that the scenarios they express are too subtle to be accurately summed up in a single word.

'Fatalism' in Russian Grammar

The Russian language has a large family of impersonal dative-infinitive constructions which refer to inexplicable things that happen to people against their will or irrespective of their will (Wierzbicka, 1992). For example, the sentence *Ne budet tebe pasporta* 'There'll be no passport for you', which involves a dative pronoun, a negated existential verb, and a noun in the genitive case, conveys a subtle message that is not fully evident in the English translation. Essentially, a construction of this kind rules out the possibility of someone being able to obtain something beneficial and desired, implying at the same time that this is due to the fact that someone in a superior position does not want it to happen.

A second construction combines a human noun in dative case with a mental verb in the third person singular reflexive form; for example, *Segodnja mne vspomnilas' Praga – sady* 'Today I was reminded of Prague – of its gardens'. This depicts a mental event simply happening inside a person inexplicably, and, in a sense, irresistibly. The most important Russian expression of this sort is *xočetsja* lit. 'it wants itself to me', which suggests a spontaneous and involuntary desire. Also worthy of mention are impersonal modal predicates with dative subjects – extremely common in colloquial Russian – such as *neobxodimo* 'it is indispensable', *nel'zja* 'one may not', *nado* 'it is necessary', *nužno* 'it is necessary/required', *sleduet* 'one ought to', and *dolžno* 'one has to'; and the sundry infinitive and reflexive constructions conveying meanings related to helplessness, obligation, and necessity. Russian furnishes an excellent example of the grammatical elaboration of a semantic theme; in this case, the theme of *sud'ba* 'fate' (mentioned above) – of not being in control, of living in a world which is unknowable and which cannot be rationally controlled. Taken together, these constructions give the Russian language a cognitive profile which enables and facilitates habitual thinking in these terms.

CONCLUSION

This article has summarized evidence and arguments in support of the view that language-specific lexical and grammatical semantics significantly influences higher-order cognition. The issues remain controversial, however, largely due to lack of consensus about appropriate methodology for linguistic semantics, a field which remains underdeveloped in comparison with formal syntax. There are also differences of opinion as to what counts as acceptable evidence in conceptual analysis: intuition and appeal to usage, corpus studies of texts, controlled experiments. There is a pressing need for many more cross-linguistic studies, especially combining methods from linguistics and psycholinguistics, but a relative shortage of scholars with appropriately broad training who can carry these through. With the revival of interest in cognitive aspects of language and promising findings emerging from a number of research groups, it seems reasonable to expect significant progress in the near future.

References

Enfield NJ (2000) On linguocentrism. In: Pütz M and Verspoor M (eds) *Explorations in Linguistic Relativity*, pp. 125–157. Amsterdam: John Benjamins.

Goddard C and Wierzbicka A (eds) (2002) *Meaning and Universal Grammar – Theory and Empirical Findings*. Amsterdam: John Benjamins.

Harkins J and Wierzbicka A (eds) (2001) *Emotions in Crosslinguistic Perspective*. Berlin: Mouton de Gruyter.

Jackendoff R (1990) *Semantic Structures*. Cambridge, MA: MIT Press.

Jackendoff R (1999) The natural logic of rights and obligations. In: Jackendoff R, Bloom P and Wynn K (eds) *Language, Logic, and Concepts*, pp. 67–95. Cambridge, MA: MIT Press.
Lucy JA (1992) *Grammatical Categories and Cognition. A Case Study of the Linguistic Relativity Hypothesis*. Cambridge, UK and New York, NY: Cambridge University Press.
Rohlen TP (1974) *For Harmony and Strength: Japanese White-collar Organisation in Anthropological Perspective*. Berkeley, CA: University of California Press.
Slobin DI (1996) From 'thought and language' to 'thinking for speaking'. In: Gumperz JJ and Levinson SC (eds) *Rethinking Linguistic Relativity*, pp. 70–96. Cambridge, UK: Cambridge University Press.
Slobin DI (2000) Verbalized events: a dynamic approach to linguistic relativity and determinism. In: Niemeier S and Dirven R (eds) *Evidence for Linguistic Relativity*, pp. 107–138. Amsterdam, Netherlands: John Benjamins.
Tomasello M (1999) *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
Wierzbicka A (1992) *Semantics, Culture, and Cognition: Universal Human Concepts in Culture-specific Configurations*. New York, NY: Oxford University Press.
Wierzbicka A (1996) *Semantics: Primes and Universals*. Oxford, UK: Oxford University Press.

Further Reading

Bowerman M and Levinson SC (eds) (2001) *Language Acquisition and Conceptual Development*. Cambridge, UK: Cambridge University Press.
Enfield NJ (ed.) (2002) *Ethnosyntax: Explorations in Grammar and Culture*. Oxford, UK: Oxford University Press.
Goddard C (1998) *Semantic Analysis: A Practical Introduction*. Oxford, UK: Oxford University Press.
Goddard C and Wierzbicka A (eds) (1994) *Semantic and Lexical Universals – Theory and Empirical Findings*. Amsterdam: John Benjamins.
Gumperz JJ and Levinson SC (eds) (1996) *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.
Lakoff G (1987) *Women, Fire and Dangerous Things*. Chicago, IL: Chicago University Press.
Lee P (1996) *The Whorf Theory Complex*. Amsterdam: John Benjamins.
Shweder RA (1993) Cultural psychology: who needs it? *Annual Review of Psychology* **4**: 487–523.
Whorf BL (1956) *Language, Thought and Reality*, edited and with an introduction by JB Carroll. Cambridge, MA: MIT Press.
Wierzbicka A (1999) *Emotions across Languages and Cultures*. Cambridge, UK: Cambridge University Press.

Semantics and Pragmatics: Formal Approaches

Advanced article

Allan Ramsay, University of Manchester Institute of Science and Technology, Manchester, UK

CONTENTS

Formal languages can be used to throw light on how natural language relates to the world.

FORMAL APPROACHES

Language is used for getting other people to do things for you. You tell them what you believe and what your goals are, and if they are feeling helpful they do what they think you want. If it did not have this key function, it would be no more (and no less) than a highly complex set of rituals aimed at establishing social relationships – rather like the courtship displays of various bird species, or the cues that indicate dominance relationships within groups of primates. It would still be interesting, but it would hardly be the crucial facet of human behavior that it is. Just how interested would we be in the relationship between the active and passive voice, or in the effect of moving a phrase from its canonical position in a sentence to the left or right boundary, if these structural relationships did not convey *ideas*?

The fact that language can convey ideas (that utterances have meanings), then, is one of its most important properties. It is also one of the slipperiest to capture. Structural properties such as the syntactic and morphological characteristics of a language can be observed and described more or less objectively, by looking at corpora to see the patterns that occur and to explore hypotheses about patterns that might occur and don't, and by asking for native speaker judgments about what is and is not acceptable. Social consequences of particular linguistic behaviors can be judged, at least partly, by observing the correlation between linguistic choice and social interaction. But there is very little to *see* when someone assimilates an utterance: if I say '*I think Darwin's claim that people are descended from apes is nonsense*' and you say '*I've been saying that for years*', no observer can actually see what is going on in each of our heads.

To make matters worse, the only tool you have for describing meanings is language itself. There seems to be an unacceptable degree of circularity in Tarski's attempt to describe meaning by saying that the meaning of '*Snow is white*' is that snow is white. Saying that '*Snow is white*' means that the crystalline form of water uniformly reflects all wavelengths of visible light seems a bit better, as does saying that '*La neige est blanche*' means that snow is white. Nonetheless, it seems rather unsatisfactory to rely on natural language as the means by which we describe natural language meanings.

One way out of this impasse is to use a language for which we do have a clear independent notion of meaning. Giving the meaning of an English sentence by translating it into English seems rather unsatisfactory, whereas giving the meaning of a French sentence by translating it into English does appear to be helpful, at least for someone who understands English but not French. So perhaps translating into some other *kind* of language will enable us to make even further progress.

What other kinds of language are there? One family of languages that we might use is the family of mathematical languages, and especially the languages of formal logic. These are languages for which a tight relationship between the *form* of a sentence and the states of affairs that make it true has been very carefully mapped out. It is this relationship between form and content that earns these languages the title 'formal'. A formal logic specifies a relationship between what a sentence looks like, how it relates to the world, and what you can do with it.

There are many varieties of formal logic. In some cases this relationship between what a sentence

looks like, how it relates to the world (its semantics) and what you can do with it (its proof theory) is rather complex or unintuitive. So be it. Natural language is a remarkably expressive medium, allowing us to capture the most delicate nuances of human thought in very concise and elegant ways. A formal language with the same expressive power is therefore likely to be a complex object, since it will include devices for expressing the same nuances as the original natural language. The current article will consider some of these complexities, and will sketch the most prominent approaches to them.

SEMANTICS

A distinction is often drawn between sentence meaning and utterance meaning. Consider the sentence

The door is open. (1)

The meaning of this *as a sentence* is that some door that the speaker and hearer are familiar with is in a particular geometrical relation to the wall in which it is mounted. Its meaning as an utterance might be that the speaker wants the hearer to come into the room, it might be that the speaker wants the hearer to shut the door, it might be a warning that the tiger in the next room is about to come in and eat the hearer, it might be … The study of sentence meaning – semantics – deals with the relationship between the sequence of words the speaker chose and possible ways for the world to be. The study of utterance meaning – pragmatics – deals with the speaker's reasons for evoking a particular view of the world. The two are clearly interwoven, but it is generally assumed that any very detailed analysis of pragmatics will depend on a prior understanding of semantics. To put it bluntly, I will have difficulty understanding how '*The door's open*' can function as a request to close it if I don't understand the geometry of the situation that it conveys. The first task, then, is to extract an expression in some appropriate formal language which 'means the same as' the target natural language sentence.

Logical form

Such a paraphrase is often referred to as a logical form. If we can paraphrase natural language sentences into logical forms that have the same meaning, then we are in a position to see how they can be used to change other people's beliefs and affect their behavior. There are, however, two problems. The first is that we have to agree that the logical form does mean the same as the original, and the second is that we have to find some systematic way of finding logical forms.

The first of these seems to be the same circular problem as the one we had before. How can I be sure that $\forall X(man(X) \rightarrow mortal(X))$ means the same as '*All men are mortal*' when I don't know how to specify the meaning of '*All men are mortal*'?

Fortunately the languages of formal logic do provide a test which will at least tell us when our descriptions are *wrong*, namely we can see whether our logical forms enter into the right entailment relations. If someone says that all men are mortal and that Socrates is a man then they are committed to the claim that Socrates is mortal, and if they also say that no men are elephants then they are further committed to the claim that Socrates is not an elephant. Words and sentences enter into a web of entailment relationships: any theory of meaning must support the right entailments. A theory of meaning that allows me to say that Socrates is a man and that no men are elephants without committing me to the claim that Socrates is not an elephant is a poor theory of meaning. We can therefore exploit the fact that the languages of formal logic are designed to support entailment relations, so that if you produce a formal paraphrase in some logic you can see whether it supports the entailments that you believe the original natural language utterance supports. Different logics embody different notions of entailment, and it may well be that some of these are more appropriate when you are thinking about natural language semantics than others. But all of them support some such notion, and hence all of them make it possible to test semantic theories.

It seems, then, that it may be worth trying to obtain formal paraphrases of natural language sentences. The next task is to show how this can be done in a systematic way. The key to this lies in the principle of compositionality:

> The meaning of the whole is a function of the meaning of the parts and their mode of combination (Dowty *et al.*, 1981)

This principle underpins the entire tradition of formal semantics, from Frege's pioneering work through the crucial work of Montague to the current range of theories such as situation semantics (Barwise and Perry, 1983), discourse representation theory (Kamp, 1984; Kamp and Reyle, 1993), dynamic predicate logic (Groenendijk and Stokhof, 1991) and so on. The aim is to develop a *systematic* way of obtaining logical forms from natural

language sentences by assigning meanings to words and assigning rules for combining meanings to syntactic rules. If we can do this then we have an account of how sentences come to carry meanings, and we can test our account by seeing whether the formal paraphrases of related sentences have the same entailment relations as the original natural language forms.

The first substantial attempt to develop a compositional semantics of this kind was carried out by Richard Montague (1974), in a widely referenced (but seldom read) paper entitled *The proper treatment of quantification in ordinary English*, or PTQ for short. This paper, and the description of it given by Dowty *et al*. (1981), led to an explosion of interest, since it showed that for at least some parts of at least one language it was possible to obtain testable logical forms directly and systematically from the natural language text.

The key to Montague's work, and to most of what followed, was that if you have a hypothesis about what *might* be a good logical form for a given sentence, you can take it apart and identify those parts of the logical form that arise from the various syntactic constituents of the sentence. The notion of abstraction is crucial here. Consider sentences (2)–(4) and the corresponding (very simple) logical forms.

John loves Mary: *love*(*john*, *mary*) (2)

Bill loves Mary: *love*(*bill*, *mary*) (3)

Peter loves Mary: *love*(*peter*, *mary*) (4)

The English sentences share the property that they are each made up of a subject (which varies from case to case – '*John, Bill, Peter*') and a common VP – '*loves Mary*'. The logical forms share the property that they all fit the pattern *love* (-, *mary*), with the 'hole' being filled in by different individuals in each case. So if we knew how to 'fill holes' and we knew how to get '*loves Mary*' to mean *love* (-, *mary*) then we would be able to obtain *love* (*John, mary*) from '*John loves Mary*' and *love* (*ronald, mary*) from '*Ronald loves Mary*' and …

How can we make '*loves Mary*' mean *love* (-, *mary*)? Consider the VPs in (5)–(7), where each of them has a paraphrase which is analogous to the one for '*loves Mary*'.

loves Mary: *love*(-, *mary*) (5)

loves Susan: *love*(-, *susan*) (6)

loves Jane: *love*(-, *jane*) (7)

Again the English forms have something in common, namely that they are each made of an object, which varies from case to case, and a common verb: and the logical forms have something in common, namely that they fit the pattern *love* (-, -) with the *second* hole being filled by different items.

So what we have to do is to give lexical items meanings that may have holes in them, and then specify how we will fill in those holes. There are two standard approaches to this task, using function application or unification.

Function application

Montague suggested that the λ-calculus, which is a logic that was designed for talking about sets and properties, would provide a suitable framework for assembling meaning fragments into well-formed semantic expressions. The particular version of the λ-calculus that he used was carefully formulated to avoid the paradoxes of self-reference, such as the Liar Paradox – '*What I am now saying is untrue*', or '*I am lying*', which is true if it is false and false if it is true – and Russell's 'set of all sets that are not members of themselves', which is a member of itself if it isn't, and isn't a member of itself if it is. A number of alternative formulations have been developed and may also be used for this task (Turner, 1987; Bealer, 1982; Ramsay, 2001).

The λ-calculus provides a notation for talking about 'holes' and for specifying the order in which they will be filled. To take the example given previously, we could use the expression $\lambda A \lambda B love(B, A)$ to say that the word '*love*' denotes the relationship between two individuals where one loves the other, and that we are going to supply those individuals by supplying the person who is loved first and the person doing the loving second. The λ-variables indicate the holes that are to be filled. Since A is the outermost such variable, its hole is the one that will be filled first; that is, we are expecting to be told about the object before we get the subject (which is the order in which most grammars of English incorporate the arguments of the verb '*love*').

The process of supplying fillers for holes is then done by function application. If you have a λ-expression of the form $\lambda x P(x)$ and you apply it to some term t, written $(\lambda x\ P(x)).t$, then you can turn it into $P(t)$. The rule that licenses this, namely $(\lambda x P(x)).t \equiv P(t)$, is known as the Tarski biconditional (TB). You have to be careful how you use it when confronted with things like the Liar Paradox, but it does just what is needed for composing small fragments of meanings into larger ones. You just annotate your grammatical rules with appropriate expressions and apply the simplifications allowed

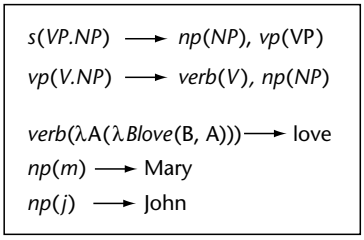


Figure 1. Simple grammar annotated with λ-calculus expressions.

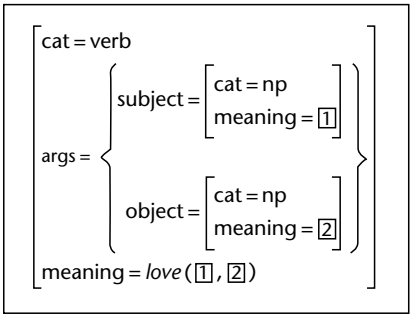


Figure 2. Lexical entry for '*love*'.

by TB. Figure 1 shows a tiny annotated grammar. If you apply this grammar to the sentence '*John loves Mary*' you will obtain the logical form $(\lambda A(\lambda Blove(B, A)).m).j$. The subformula $\lambda A(\lambda Blove(B, A)).m$ simplifies to $\lambda Blove(B, m)$, so the original simplifies to $\lambda Blove(B, m).j$, which in turn becomes $love(j, m)$ as required.

This works fine for very simple examples of the kind shown in Figure 1. It gets considerably more complicated as you try to cover more substantial fragments of the language, but it has remained popular because the rigorous work by logicians on the properties of the λ-calculus means that the interpretations that emerge are well-formed and well-understood, with even fragments and incomplete phrases being given sound interpretations.

Unification

An alternative way of filling in holes is to use unification for copying the meaning of one constituent into a hole in the meaning of another (Fenstad *et al*., 1987). Unification provides a way of saying that two items should be the same – if you refer to them using the same name they must be identical. It is widely used in grammar for imposing constraints such as number agreement between subject and verb, but it can also be used for specifying that the meaning of some fragment of an utterance plays a specified role in the meaning of the whole thing. This technique is often used with highly lexical grammars such as HPSG, (Pollard and Sag, 1994; Sag and Wasow, 1999), where a lexical entry will contain a complete description of its arguments and the way their semantics contributes to the meaning of the whole. Figure 2, for instance, contains a description of the word '*love*', showing that it requires a subject and an object and indicating how their meanings are utilized in the meaning of any sentence built around this word.

Using unification as the way you build meanings of larger units from the meanings of their constituents is well suited to highly lexical grammars, and provides a slightly easier starting position. It is less flexible than using the λ-calculus, and less well-suited to providing interpretations of fragmentary items. Recent work on glue languages attempts to provide some of the advantages of both (van Genabith and Crouch, 1997; Dalrymple *et al*., 1996), but it remains the case that building large scale semantic descriptions remains almost as hard as it is important.

Dynamic semantics

Montague's original work concentrated on a fragment of English which contained names, indefinite NPs, and universal NPs – sentences like (8)–(10).

All men are mortal. (8)

Some lecturer owns a red car. (9)

John wants a new bike. (10)

It's hard to get satisfactory logical forms for sentences like these, particularly when you realise that (10) has an unexpected ambiguity, since there may be some specific bike that John wants – '*John wants a new bike. He's put down a deposit on it, but he has to find the remainder by Saturday*' or it might just be that he wants to be in a situation where he owns a bike that is better than his current one – '*John wants a new bike. He hasn't chosen one yet, but he needs something with 18 gears*'. It quickly became apparent that none of the available logics could cope with definite NPs or pronouns. Attempts by philosophers such as Russell, Strawson, and Kripke to capture the meaning of sentences with definite descriptions using the tools of ordinary predicate logic all ran into problems. The tools for turning natural language sentences into logical forms existed, but it seemed that the available logical languages were somehow inadequate.

The development of a suitable logic for dealing with such expressions, and with other

'presupposition inducing' forms, became a priority, and a number of dynamic logics were developed (Barwise and Perry, 1983; Kamp, 1984; Groenendijk and Stokhof, 1991; Heim, 1983), with some, at least, of these growing out of work in dynamic logic as a foundation for computer science (Pratt, 1974). Other researchers in the area such as Gazdar (1979) and Karttunen (1973) tried to draw on non-monotonic logics and epistemic logics for the same purposes. The goal was to find some formal language which could specify a relationship between the situation in which an utterance was produced and the effect it was intended to evoke. The assumption was that natural language sentences were objects that *changed* the situation in which they were uttered, and that the way a sentence changes a situation depends on what that situation is like. In other words, natural language is dynamic – it has effects, and those effects depend on the context.

The range of logics that were developed or adapted for this task gives some indication of the complexities and subtleties that emerge when you try to give a concrete account of the various phenomena. What they all have in common, however, is that there is a two-stage process of relating the utterance to the current situation – of 'anchoring' it, in Barwise and Perry's terminology – and then using the resulting expression to 'update' the situation. What situations are like, how you do anchoring, how you update a situation, varies from theory to theory, but the key steps are pretty constant. So for a sentence like (11) you would get something like the logical form in Figure 3.

Mary saw the boy who she fancied. (11)

The conditions in Figure 3 have to be satisfiable in the current situation for this sentence to make any sense. In other words, you can only interpret (11) in a situation where there is some X called Mary, and some Y who is a female, and some Z who is a boy who Y fancies. Verifying that these conditions hold is likely to give you a handle on the identities of these individuals (with Mary and the female person probably being the same individual), though there are cases where you can verify the conditions without actually being able to identify the individuals in question. Once you have identified X, Y, and Z you can determine the 'content', namely that X saw Z, and use this to change your view of the world.

conditions: *name* (X, Mary), *female* (Y), *boy* (Z) & *fancy* (Y, Z)
content: *saw* (X, Z)

Figure 3. Logical form for (11).

This basic notion that natural language sentences express conditions that have to be verified by the hearer, with the process of verification fleshing out the content, is shared by a number of theories. The details can get rather complex, not least because people will often produce utterances in situations where they know that the hearer has no chance of verifying the conditions. I can say (12) to someone who has no idea how I get to work and confidently expect that although they cannot verify the condition introduced by the NP '*my train to work this morning*' they will be prepared to 'accommodate' it. But if conditions don't actually have to hold for a sentence to make sense, just what status *do* they have?

My train to work this morning was late. (12)

There are plenty of other problems associated with the general notion. Sometimes, for instance, conditions can be hypothetically satisfied, as in 'donkey sentences' such as (13) (Geach, 1962) and in certain opaque contexts, as in (14) (Karttunen, 1973).

If a former owns a donkey he beats it. (13)

John believes that the king of Ruritania is in the kitchen. (14)

In (13) the referents for '*he*' and '*it*' are only present in the *hypothetical* situation described by the antecedent '*a farmer owns a donkey*', so it is clear that the conditions do not have to be verified with respect to the *actual* situation. Similarly neither the speaker nor the hearer has to believe in the existence of a king of Ruritania (or even a country called Ruritania!) for (14) to make sense. Statement (14) is, in fact, open to at least two readings, one where the speaker believes that there is someone who is the king of Ruritania and that John believes that that person is in the kitchen (on this reading John need *not* believe that there is such a king) and one where the speaker doesn't believe there is a king of Ruritania but she does think that John believes there is.

There are plenty of other such problems, with different theories providing good solutions to some and not to others. The key to all such dynamic theories of semantics, however, is that something like the context, or the speaker's view of the context, or the speaker and hearer's mutual knowledge plays an important role.

PRAGMATICS

Even theories which emphasize the role of the context for instantiating the content of an utterance generally fail to explain why people say things or why they say them the way they do. Thinking about dynamic properties of language does raise some of these questions – why, for instance, would someone force their hearer to accommodate a presupposition, as in (12) above, rather than spelling out all the details as in (15) – but semantic theories tend to restrict themselves to considering the content of an utterance rather than its function.

I came to work on a train this morning. It was late. (15)

If it is true, as I suggested earlier on, that the most important role for language is as a means of influencing other people's behavior then we have to move on from the idea that language is used for describing the way the world is. We have to move on from semantics – what does this utterance say about the world? – to pragmatics – what does it say about me and my goals?

Speech acts

The form of an utterance gives the hearer some hint about what they are expected to do. Consider (16)–(18):

You are going to London tomorrow. (16)

Are you going to London tomorrow? (17)

Go to London tomorrow. (18)

The form of (16) marks it as being a statement – the kind of thing that you might say to someone who is known to be absent-minded and is likely to have lost their diary. The form of (17) marks it as a query. You would say (17) if you didn't know the plans of the person you were talking to. The form of (18) marks it as a command or an instruction, as something that the speaker wants or advises the hearer to do.

We could therefore wrap the content of the sentence up inside some operator that told us what to do. If I tell you something I want you to add it to your model of the world; if I ask you something then I must want you to tell me whether it is true; if I command you to do something then I must want you to do it. So we could try to map the form of the utterance onto something which is recognisable as some kind of action specification or state change description, obtaining logical forms like that in Figure 4 (which says that anyone who hears (16) is expected to accept it). The problem here is that what the speaker wants the hearer to do is often only indirectly related to the form of their utterance. Consider (1) again:

accept (∃*C agent* (*C*, *ref* (λ*Dhearer* (*D*))))
&*go*(*C*)
&*C is event*
&*to* (*C*, *ref* (λ*F* (*name* (*F*, *London*))))
&*tomorrrow* (*C*)

Figure 4. Logical form for (16).

The door is open. (1)

This sentence looks like a statement. Indeed it *is* a statement. But what the speaker wants when he utters it will vary greatly from situation to situation – he may want the hearers to note this new fact, he may want them to note it and take advantage of it by entering the room, he may want them to note it and make it untrue, …

Austin (1962) and Searle (1969) addressed this problem by introducing the notion of speech acts. A speech act is something that a speaker can do by uttering a sentence in an appropriate context. To take a simple example, I can successfully inform you that I am going out by uttering the words '*I am going out*' in a situation where I correctly believe that you are not already aware of my plans; and I can successfully *remind* you of this fact by uttering the same words in a situation where I correctly believe that you *are* already aware of them. I can perform different acts by uttering the same words in different contexts, with part of Austin and Searle's goal being to describe the effects that a given form of words will have in a range of contexts.

In some cases, however, there seems to be a considerable gap between the form of words and the intended effect. Although (1) has the form of a statement, it can clearly function as a command, and with a rising intonation it can function as a query, but since the intonation is part of the form this seems less problematic. Although '*Do you know the time?*' has the form of a yes/no query, it often has the same function as the WH-query '*What's the time?*'. The intended effect of an utterance, then, is often at odds with its apparent form.

Searle characterizes speech acts by describing the conditions under which it is possible to perform the act in question (so you can *assert* some proposition *P* only if you have grounds for believing *P* and it is not obvious that your hearer already knows *P*), and by specifying their intended effects (though he only

discusses the effects in any detail for the rather unusual act of promising). The idea that speech acts can be analyzed in terms of preconditions and effects maps very neatly onto work in artificial intelligence (AI) on planning (Fikes and Nilsson, 1971), and there is a large body of work showing how to explain the 'indirect' effects of utterances like (1) by linking AI planning theory and epistemic reasoning (Cohen and Perrault, 1979; Cohen *et al*., 1990; Bunt and Black, 2000). It is clear that the form of an utterance underspecifies its intended effects, and that some mechanism for connecting the consequences of the literal meaning to the speaker's and hearer's beliefs and goals is required. Whether this can be achieved by producing a table of speech acts, delineated by their preconditions and effects, remains open to question.

Discourse structure

Sentences have structure, which is used for encoding meanings. Extended discourses also have structure. Compare (19) and (20):

> Beach regarded him solicitously, but did not develop the theme. He had a nice sense of the proprieties. Between himself and this young man there had existed for eighteen years a warm friendship.
> *Heavy Weather*, P.G. Wodehouse (19)

> Beach regarded him solicitously, but did not develop the theme. Between himself and this young man there had existed for eighteen years a warm friendship. He had a nice sense of the proprieties. (20)

(19) makes far more sense than (20) even though the sentences in the two passages are identical. Part of the problem with (20) lies in the difficulty of dereferencing the pronoun '*he*', which clearly points to Beach in (19) but which might just as well pick out the young man in (20). Part of it lies in the *implicit* relationships between the propositions encoded in the three sentences. In (19) the second sentence explains why Beach did not develop the theme, and the third is the start of an extended description of the social positions of these two people. It is much harder to give a similar account of (20). It would be possible to link the first and second sentences, with their long-standing warm relationship explaining Beach's reluctance to discuss a difficult subject, but then it is very hard to see where the third sentence might fit in.

Extended discourses are used for conveying complex ideas – for telling stories, for providing explanations, for specifying courses of action. Complex ideas generally have a natural structure. A story is usually worth telling only if it concerns some individuals who are placed in a situation where one sequence of events would normally be expected to occur but some other sequence actually takes place. To tell a story, then, you have to introduce the individuals and the situation, describe how the actual course of events differs from what might have been expected, and then say what the individuals did in response to this deviation: a story has to have a beginning, a middle, and an end (Schank and Abelson, 1977; Rumelhart, 1975). A description, on the other hand, might proceed by introducing some entity which is quite like the thing you are trying to describe and then enumerating the differences (McKeown, 1985). Whatever a discourse is about, the subject matter is likely to be decomposable into smaller parts, and it is a good idea to make the structure of the discourse match this decomposition.

Language provides numerous devices for organizing extended discourses. There are lexical items that explicitly express connections between sentences – '*therefore*', '*however*', '*before that*', '*anyway*', … There are textual conventions such as the breaks that occur before paragraphs, and the chapter and section headings that are used to indicate the structure of very large texts such as books and articles: and there are a number of sentence-internal clues that indicate the speaker's attitude to what they are saying, and hence help to impose a structure.

The first of these sentence-internal clues relates to the fact that, in English, the first phrasal constituent of a sentence seems to carry more weight than the remainder. Halliday (1985) talks about the *theme* and *rheme* where the theme is the thing the sentence is *about* and the rheme is what is being said about it. Other authors use slightly different terminology – Hajicova and Sgall (1984), for instance, talk of the *topic* and *comment*, and the terms *given* and *new* are often used – and the syntactic device that marks the theme may vary from language to language, but the basic idea is clear: in a sentence like (2) I am telling you something about John, namely that he loves Mary, whereas in (21) I am telling you something about Mary, namely that John loves her.

> John loves Mary. (2)

> Mary is loved by John. (21)

The problem is that although the basic idea is clear at an intuitive level, it is exceptionally difficult to

pin it down. What does it mean to say that a sentence is 'about' John?

If we want to carry out our program of obtaining logical forms that mean the same as natural language sentences, we have first of all to find some way of including the division of this sentence into two parts in our logical form. To do this we have to decide what the parts are, which is quite easy in English since the theme is just the leftmost phrasal constituent; and then note that the meaning of one of these parts is 'about' the other – something like the logical form given in Figure 5, which was obtained as described by Seville (1999). Note that using the λ-calculus as the language in which we produce logical forms makes it possible to produce a logical form in which parts of the overall meaning appear independently.

Once we have such a logical form, however, it immediately becomes apparent that we have to say what *about* means. Intuitively it is fairly clear, but the motivation behind producing formal models is to make them precise and testable – to go beyond intuition. This turns out to be extremely hard. We need to know what follows from the fact that someone said (2) that would not have followed if they had said (21) and vice versa, and this is really very unclear.

Similar remarks apply to other local indicators of discourse structure, for instance to (22)–(24).

I married *Rosie*. (22)

I *married* Rosie. (23)

I married Rosie. (24)

As with (2) and (21), the state of affairs being reported is the same in each case. (2) and (21) both describe a state of affairs where John loves Mary, and (22)–(24) all say that the speaker married Rosie. The difference between them is that (22) would be uttered in a context where it had been assumed that the speaker married someone else, (23) in a context where it had been assumed that the speaker did something else to Rosie, and (24) in one where it had been assumed that someone else married Rosie. As with the theme:rheme distinction, it is possible to obtain logical forms which indicate that some part of the meaning is in *focus* (Ramsay, 1994; Pulman, 1993), but drawing out the consequences of choosing one of (22)–(24) rather than another remains extremely difficult.

Marking things as being interesting, or contrastive, or thematic, or … does, however, help to constrain the discourse structure. The tighter the connection between the themes of two consecutive sentences, the more closely those sentences are likely to be linked (so if S1 and S2 have the same theme, then S2 is probably a continuation of S1, if the theme of S2 was mentioned in S1 then S2 is likely to be some kind of explanation or elaboration of S1, and if S2's theme wasn't even mentioned in S1 then it may be that S2 is actually connected to some earlier sentence). Unfortunately, trying to pin down what is meant by saying that S2 is an explanation of S1, or an elaboration, or a justification is as hard as trying to formalize the claim that (2) is 'about' John. Mann and Thompson (1988) and Mann (1999) try to provide a systematic attempt to enumerate and describe the range of possible discourse relations, but there is certainly no consensus about how the components of a discourse relate to one another, even when explicit lexical cues such as '*anyway*' and '*moreover*' are present.

Nonetheless, although it is hard to be clear about how to relate elements of a discourse, the fact that they *are* related does seem worth pursuing. Centering theory (Grosz *et al*., 1995; Brennan *et al*., 1987; Joshi and Weinstein, 1998) extracts the structure of a discourse by following the way the theme changes. The devices used for spotting which item is the center are slightly different from Halliday's very simple decision to use the first phrasal daughter, but the ideas are very similar. Centering theory has been widely exploited for helping with pronoun dereferencing, partly to help pick out where the referent was likely to be introduced (e.g. in the last sentence which mentioned the theme of the current one (Hitzeman and Poesio, 1998; Seville, 1999)) and partly to control the order in which potential referents are tried (roughly speaking, you can't use a pronoun to refer to some item that was mentioned in the previous sentence if you have referred to its theme with a full NP). We may not quite know everything about how the form of an extended utterance encodes the message, but we do know that it plays an important role in making that message manageable, and that a system of mutual

$$about(ref(\lambda F(name(F, John))),\ \lambda G \exists D(agent(D,G)) \& love(D) \& object(D, ref(\lambda H(name(H, Mary)))))$$

Figure 5. Logical form for (2).

constraints between nominal form, especially the choice of whether or not to use a pronoun, and discourse structure plays an important role in specifying this structure.

References

Austin J (1962) *How to Do Things with Words*. Oxford, UK: Oxford University Press.

Barwise J and Perry J (1983) *Situations and Attitudes*. Cambridge, MA: Bradford Books.

Bealer G (1982) *Quality and Concept*. Oxford, UK: Clarendon Press.

Brennan SE, Friedman MW and Pollard CJ (1987) *A centering approach to pronouns*. Proceedings of 25th meeting of the ACL.

Bunt H and Black WJ (eds) (2000) *Abduction, Beliefs and Context: Studies in Computational Pragmatics*. Amsterdam, PA: John Benjamins.

Cohen PR, Morgan J and Pollack ME (1990) *Intentions in Communication*. Cambridge, MA: Bradford Books.

Cohen PR and Perrault CR (1979) Elements of a plan-based theory of speech acts. *Cognitive Science* **7**(2): 171–190.

Dalrymple M, Lamping J, Pereira FCN and Saraswat V (1996) A deductive account of quantification in LFG. In: M Kanazawa, C Piñón and de Swart H (eds) *Quantifiers, Deduction and Context*, pp. 33–58.

Dowty DR, Wall RE and Peters S (1981) *Introduction to Montague Semantics*. Dordrecht, Netherlands: Reidel.

Fenstad JE, Halvorsen P-K, Langholm T and van Bentham J (1987) *Situations, Language and Logic*. Amsterdam, Netherlands: Kluwer Academic.

Fikes RE and Nilsson NJ (1971) Strips: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence* **3**(4): 251–288.

Gazdar G (1979) *Pragmatics: Implicature, Presupposition and Logical Form*. New York, NY: Academic Press.

Geach PT (1962) *Reference and Generality: An Examination of Some Medieval and Modern Theories*. Ithaca, NY: Cornell University Press.

van Genabith J and Crouch R (1997) How to glue a donkey to an f-structure. In: H Bunt, L Kievit, R Muskens and M Verlinden (eds) *2nd International Workshop on Computational Semantics*, pp. 52–65, Tilburg.

Groenendijk J and Stokhof M (1991) Dynamic predicate logic. *Linguistics and Philosophy* **14**: 39–100.

Grosz BJ, Joshi A and Weinstein S (1995) Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, pp. 175–204.

Hajicova E and Sgall P (1984) From topic and focus of a sentence to linking in a text. In: Bara BG and Guida G (eds) *Computational Models of Natural Language Processing*, pp. 151–163. Oxford, UK: North-Holland.

Halliday MAK (1985) *An Introduction to Functional Grammar*. London, UK: Arnold.

Heim I (1983) File change semantics and the familiarity theory of definiteness. In: Baüerle R, Schwarze C and von Stechow A (eds) *Meaning, Use and Interpretation of Language*, pp. 164–189, Berlin, Germany: Walter de Gruyter.

Hitzeman J and Poesio M (1998) Long distance pronominalisation and global focus. COLING/ACL 98, Montreal.

Joshi A and Weinstein S (1998) Formal systems for complexity and control of inference: a reprise and some hints. In: Walker MA, Joshi AK and Prince EF (eds) *Centering Theory in Discourse*, pp. 31–38. Oxford, UK: Oxford University Press.

Kamp H (1984) A theory of truth and semantic representation. In: Groenendijk J, Janssen J and Stokhof M (eds) *Formal Methods in the Study of Language*, pp. 277–322. Dordrecht, Netherlands: Foris Publications.

Kamp H and Reyle U (1993) *From Discourse to Logic: Introduction to Model Theoretic Semantics of Natural Language*. Dordrecht, Netherlands: Kluwer Academic Press.

Karttunen L (1973) Presuppositions of compound sentences. *Linguistic Inquiry* **4**: 169–193.

Mann WC (1999) An introduction to rhetorical structure theory. [http://www.sil.org/linguistics/RST.]

Mann WC and Thompson SA (1988) Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**(3): 243–281.

McKeown K (1985) *Generating English Text*. Cambridge, UK: Cambridge University Press.

Montague R (1974) The proper treatment of quantification in ordinary English. In Thomason RH (ed.) *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, CT: Yale University Press.

Pollard CJ and Sag IA (1994) *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press.

Pratt VR (1974) *Dynamic Logic*. Proceedings of the 6th International Congress on Logic, Philosophy and Methodology of Science.

Pulman SG (1993) *Higher Order Unification and the Semantics of Focus*. Technical report, University of Cambridge.

Ramsay AM (1994) Focus on 'Only', and 'Not'. In: Wilks Y (ed.) *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-94), pp. 881–885, Kyoto.

Ramsay AM (2001) Theorem proving for untyped constructive λ-calculus: implementation and application. *Logic Journal of the Interest Group in Pure and Applied Logics* **9**(1): 89–106.

Rumelhart DE (1975) Notes on a schema for stories. In: Bobrow DG and Collins A (eds) *Representation and Understanding: Studies in Cognitive Science*, pp. 211–236. New York, NY: Academic Press.

Sag IA and Wasow T (1999) *Syntactic Theory: a Formal Introduction*. Stanford, CA: CSLI.

Schank RC and Abelson R (1977) *Scripts, Goals Plans and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Searle JR (1969) *Speech Acts: an Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Seville H (1999) Experiments with discourse structure. In: Bunt HC and Thijsse EGC (eds) *3rd International Workshop on Computational Semantics*, pp. 233–247, Tilburg.

Turner R (1987) A theory of properties. *Journal of Symbolic Logic* **52**(2): 455–472.

Semantics, Acquisition of

Introductory article

Robin Clark, University of Pennsylvania, Philadelphia, Pennsylvania, USA
Laura Wagner, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

Semantic acquisition concerns the problems of learning word meanings, of learning how language marks time and temporal relations, and of learning the interpretation of quantified expressions.

INTRODUCTION

A fundamental part of learning a language is learning how to mean things with that language. The most straightforward approach to the problem of meaning in natural language has been to adopt some form of the principle of compositionality:

> *Compositionality*: The meaning of a syntactically complex expression is a function of the meanings of its parts.

For example, the meaning of a complex expression like 'John walked to the store' would be computed by combining the meaning of the subject, 'John', with the meaning of the predicate, 'walked to the store'. The meaning of the latter expressions would ultimately be a function of the meanings of the words that make them up. On this view, words (or their constituent morphemes – that is, the stems and affixes that combine together to form words) would constitute the atomic elements of meaning. An alternative view would be that words (or morphemes) only have meanings in the context of sentences. Although this view is internally consistent, it would require an abandonment of compositionality in the sense outlined here, with the result that acquisition of semantics could not be done on the basis of acquisition of word (or morpheme) meanings.

A reasonable first approach to a theory of semantic acquisition would use the principle of compositionality to work backwards to the meanings of lexical items. The problem for the learner would be to recover meaning from the pairing of the observed situation in the nonlinguistic environment along with the linguistic behavior of a fluent speaker, presumably an adult care-giver. There have been two broad approaches to the problem of semantic acquisition so construed. The empiricist approach assumes that the learner has access only to this 'utterance–world' pairing. The nativist approach assumes that the learner inherits a particular structure that biases its learning in certain predetermined ways. Both approaches have certain advantages and disadvantages.

According to the empiricist approach, learners apply some simple general inductive rule to utterance–world pairings to arrive at the meanings of the component elements. While the empiricist approach is appealing in its simplicity, it is subject to a number of difficulties. First, the utterance–world pairings massively underdetermine the generalizations that the learner can make. Second, even if the utterance–world pairing is unambiguous, the meanings of the atomic elements cannot be uniquely determined from the meaning of the utterance. To illustrate the first point, we take the famous example introduced by the philosopher W. V. O. Quine; given a speaker's utterance of 'Gavagai!' in the presence of a rabbit, a hearer cannot infer from the utterance–world pairing whether the speaker meant to refer to the rabbit or to a collection of undetached rabbit parts. Both possibilities are compatible both with the speaker's linguistic behavior and with the state of the world. The second point, the impossibility of inferring the atomic elements of meaning from the meaning of the entire utterance, follows immediately from the previous point. We will examine the problem of lexical meaning in the next section.

Quine's example uses the reference of a simple concrete noun to pose the basic problem that faces a theory of semantic acquisition. The problem is worse if we observe that linguistic meanings include such esoterica as the relationship between events and time. The following sentences all involve different mappings between the basic event

referred to in the sentence and time:

a. John has walked to the store.
b. John is walking to the store.
c. John will walk to the store. (1)

We will return to this problem below in the section on tense and aspect.

Finally, the elements of a sentence may combine in ways that differ from their surface word order. The following sentence, for example, is ambiguous with respect to how the quantified expressions 'every student' and 'some language' should be interpreted with respect to each other:

Every student in the class speaks some language. (2)

This is the problem of scope; learners must somehow discover that a single word order can sometimes encode different scoping relations between the quantified expressions in it. We will turn to this problem in the last section.

Nativist approaches have taken the problems that confront empiricism as evidence that learners have an innate structure that eases the learning problem by constraining the learner to make certain kinds of generalizations. Thus, learners have definite expectations about possible word meanings and can use linguistic evidence to sort words into semantic classes. Equally, learners may simply assume that certain elements can enter into scope ambiguities without needing direct perceptual evidence that they do so. Nativist theories have the problems that they are more complex than empiricist theories and they are more difficult to falsify.

LEXICAL ACQUISITION

Word-learning happens very rapidly in development. Children begin acquiring words around their first birthday, and know thousands of them by the time they are three years old. This happens despite the fact, noted above, that the world grossly underdetermines the possible referents for any given word. One part of the solution to this problem is cross-situational observation. Children must keep track of what's going on in the world when they hear a new word, and then compare across the situations in which the word was used to find out what they have in common. The common-sense idea here is simply that the word 'dog' (or French *chien*, or Japanese *inu*) is uttered more frequently in the presence of dogs than the word *cat* (or French *chat*, or Japanese *neko*) is; the regularity between the word's referent in the world and the word's use is information the child can use. This technique is useful, but it cannot be the whole story: it cannot solve the 'Gavagai' problem, since every time a rabbit is observed, so too are temporal stages of the rabbit. How then can the child learn the meanings of so many words so fast, when the world does not provide unambiguous evidence?

Empiricist approaches stress the importance of general cognition and social cognition in solving the word-learning problem. For example, it has been shown that children are sensitive to their interlocuter's intent during labeling. In a study by Dare Baldwin, children as young as 16 months old learned that a new label is the name for a novel object only if the labeler made eye contact with the object during the labeling. If the labeler was looking away (e.g. into a bucket), the child would not match the label with the novel object, even if the child herself had been looking at the object.

Nativists believe that word-learning also requires more language-specific constraints. For example, researchers have shown that children expect a new noun to refer to a whole object and to a taxonomic category; and they expect new words to contrast in meaning with other words in their vocabulary. When a child is shown an object whose name they know (e.g. a spoon) and a novel object (e.g. a honey-dipper), and given a novel name (e.g. *dax*), they assume that the new word labels the new object. In this way, the more children know, the easier the word-learning process is.

An additional language-specific cue that children might use to help them learn words is syntax. Since words are used as parts of sentences, the learner can use the structure of the sentence and morphological elements in the sentence to help them narrow down the meaning of the word. This process has been called *syntactic bootstrapping*, since children are using their syntactic knowledge to help them learn the meanings of words. For example, only verbs of cognition, perception, and saying can appear with a sentence complement structure, as can be seen in the examples below (the asterisk denotes an unacceptable form):

a. Suzy thought/saw/said that the world was round.
b. *Suzy ran/built/slept that the world was round. (3)

When a child hears a novel verb in a sentence complement structure, she can assume that its meaning has something to do with cognition, perception, or saying. In this way she can narrow

down the hypothesis space offered by the world and alleviate the 'Gavagai' problem. Research has shown that children are sensitive to this sort of information and can use it to help establish a novel verb's referent. For example, Naigles and colleagues showed that by the age of two years, children can correctly match a transitive sentence ('Bigbird is gorping Cookie Monster') to a causative scene in which Bigbird is pushing down on Cookie Monster and making him squat, and also match an intransitive sentence ('Bigbird and Cookie Monster are gorping') to the noncausative version of the scene in which both Bigbird and Cookie Monster squat beside each other.

TENSE, MOOD, AND ASPECT

Tense, mood, and aspect (TMA) are central features of grammar that serve to situate the meaning of a sentence. Tense situates an event in time (principally, past versus present), Mood with respect to reality (principally, realis versus irrealis), and Aspect with respect to a particular interval (principally, perfective versus imperfective). The acquisition of TMA poses problems for the learner on two main fronts. The first difficulty concerns the mapping of the TMA concepts onto the right morphology in the target language. Languages vary widely in how to mark TMA information. At the extremes, Mandarin doesn't mark tense grammatically at all, while Modern Hebrew doesn't mark aspect grammatically. The missing category can be expressed explicitly (e.g. *yesterday* [past tense], *right now* [present tense], *finished* [perfective], *in the middle of* [imperfective]) but typically it is conveyed only as a matter of conversational implicature: for example, events marked as perfective are probably in the past and events marked as present tense are probably also imperfective. Between these two extremes, languages also vary in how they combine TMA information into individual morphemes. For example, French combines the past tense and imperfective aspect into the imparfait form, while English combines the past tense and perfective aspect in the simple past form. Thus, even once children can correctly identify TMA reference, they must still determine which elements their language marks explicitly and how it does so, which it pairs together in a single morpheme, and which it conveys only through implication.

The second difficulty concerns how children identify TMA reference in the first place. TMA features refer to abstract properties of an event which are not readily observable. For example, both past tense and irrealis modality are only true of events that are not currently happening at the time of utterance – but they are true of different kinds of non-occurring events. Moreover, there is a natural tendency for some TMA dimensions to group together; for example, all things being equal, events happening in the present time are usually ongoing and hence warrant both present tense and imperfective marking. These natural groupings, however, pose a learning problem: how is the child to know whether the speaker intended to mark present tense or imperfective aspect, or even whether the two concepts can be expressed separately in the language? In their own speech, children appear to take a very conservative strategy, maintaining not only these natural groupings of tense and aspect, but extending them to inherent properties of lexical predicates. Thus, in their early production of TMA morphology, children tend to restrict past tense and perfective aspect marking to telic predicates (i.e. predicates describing result-oriented events such as 'find' and 'break') and they restrict present tense and imperfective aspect marking to atelic predicates (i.e. predicates describing general activities such as 'ride' and 'laugh').

In the course of development, children generally acquire the modality distinction (realis–irrealis) first, followed by aspect (perfective–imperfective) and finally tense (past–present). It is unclear, however, whether this pattern of acquisition is the result of language-specific biases children have (as predicted by the nativist approach) or is simply the result of the kind of evidence available for identifying each category (more in line with the empiricist approach).

QUANTIFIERS AND SCOPE AMBIGUITIES

Quantified noun phrases, like those underlined in (4), pose a particular problem for learners on two counts. First, they do not denote objects straightforwardly in the way 'Mary' denotes an individual, Mary. Second, they seem to be a counterexample to the simple approach to compositionality given in the introduction:

a. Every student studies.
b. Some girl eats apples.
c. At least two politicians accept bribes. (4)

Let us turn, first, to the problem of the denotation of quantified noun phrases. It is widely accepted that quantified expressions like 'every student' do not refer to an abstract individual but, rather, to a

function. A similar analysis can be given to any of the quantified expressions in (4); the problem for the learner is to discover the function that any given quantified expression refers to. For example, the learner must discover that 'every' denotes a function that returns true if the set of individuals with the property described by the subject is contained in the set of individuals with the property described by the predicate. Any individual pairing of the world with an utterance will be insufficient to determine the function that a quantified expression denotes. Instead, the learner must be presented with a sufficiently broad sample of such pairings in order to infer the correct meaning. This problem remains very much open, with some work done in artificial intelligence on the formal problem of inferring quantifier denotations but, as yet, little work on the problem in developmental psychology.

Turning to the second problem, compositionality, recall sentence (2) presented in the introduction. It contains two quantified noun phrases, 'every student' and 'some language' and is ambiguous, meaning that it can have two very different interpretations:

a. Each student in the class speaks at least one language, although the students may differ as to which language they speak.
b. There is a language which has the property that every student in the class speaks it. (5)

If (5a) is the intended interpretation, then 'every student' is said to have wide scope over 'some language'; if (5b) is the intended interpretation, then 'some language' has wide scope over 'every student'.

The ambiguity here should be contrasted with a case of lexical ambiguity, as in (6a), and with a case of structural ambiguity, as in (6b):

a. John went to the bank.
b. John saw the man with the telescope. (6)

The examples in (6a) and (6b) do not pose a particular problem for the principle of compositionality. In (6a), the word 'bank' is simply lexically ambiguous. Example (6b) is structurally ambiguous because the syntax of English allows the prepositional phrase 'with the telescope' to combine either with the noun phrase 'the man' or with the verb phrase 'saw the man', resulting in two distinct interpretations from normal composition.

The sentence in (2) is neither lexically ambiguous as was (6a), nor do the phrases appear to have the option of combining in distinct ways to generate the ambiguity, as was the case with (6b). In this sense, sentences like (2) seem to defy the simplest interpretation of compositionality. Instead, the learner must discover that quantified expressions can be given a variable scope within a sentence, resulting in the interpretations noted in (5b) and (5a). It has been well documented in the literature, by Crain among many others, that young children prefer to assign scope to quantified expressions in the order in which they appear in the sentence. That is, the leftmost quantified expression is given widest scope. In the earliest stage, children, when presented with an example like 'some horse jumped over every dog', interpret the sentence as unambiguous, with a single horse jumping over all the dogs. Indeed, this seems to be the most readily accessible reading for adults. Adults, unlike children, admit that the sentence can also be true when different horses jump over each dog. Children before the age of about four years will deny that this reading is possible. As they acquire more experience with the language, children appear to be less tied to the surface syntactic interpretation of such sentences and admit that the quantified expressions can take a variety of different scopes. While the stages that children go through in discovering how to interpret sentences containing several quantified expressions have been thoroughly documented, it is as yet mysterious how children discover that these alternative scopings are possible; that is, it is still unknown what it is about the child and her linguistic experience that forces the child to accept the adult interpretation of such sentences.

Language stands in a complex relationship to the world. The problems that the learner faces during the course of semantic acquisition are surely daunting at first glance. Word meanings cannot be fixed just by looking at the world, and some elements, TMA elements in particular, involve dynamic interpretations that involve highly abstract computations. The learner must untangle lexical and structural ambiguities. Structural ambiguities highlight the fact that linguistic interpretation involves computation. That children are able to fix the semantic atoms of their target languages with such apparent ease is, without doubt, a wonder of the natural world, one that presents a variety of interesting and subtle scientific problems.

Further Reading

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Davies M (1981) *Meaning, Quantification, Necessity: Themes in Philosophical Logic*. London: Routledge & Kegan Paul.

Naigles L (1990) Children use syntax to learn verb meanings. *Journal of Child Language* **17**: 357–374.

Pinker S (1994) *The Language Instinct*. New York: W Morrow and Co.

Platts M (1997) *Ways of Meaning: An Introduction to a Philosophy of Language*, 2nd edn. Cambridge, MA: MIT Press.

Putnam H (1981) *Reason, Truth and History*. Cambridge, UK: Cambridge University Press.

Semantics, Dynamic

Intermediate article

Carl Vogel, University of Dublin, Dublin, Ireland

CONTENTS

Dynamic semantics is concerned with modeling the processes that change the information held by participants in a discourse or dialogue where those changes are brought about by the sentences contained in the discussion.

INTRODUCTION

The term 'dynamic semantics' refers both to a set of tools for analyzing the semantics of sentences in natural language and to ways of relating those tools to other areas of logic and computation. Muskens *et al.* (1997) provide a comprehensive survey of research along both of these lines (see also van Benthem, 1986, 1996). This article concentrates on the analysis of meaning of natural-language expressions as processes, using formal models of those processes that are logic-based dynamical systems.

Early work on natural-language processing began with processing models of syntax, with the aim of verifying the cognitive viability of process models (Winograd, 1983), but even identifying input–output correspondence between computational and cognitive models proved to be very difficult. Exact cognitive-processing mechanisms have still not been established. Nonetheless, current formal work is setting the stage for this program, both from the perspective of the syntax–semantics interface (e.g., Kempson *et al.*, 2000), and from the perspective of semantics alone.

This article describes varieties of dynamical formal systems for analyzing the semantics of natural language. This work is grounded in an understanding of how static systems work. Then various ways of achieving dynamics are described. In all cases, the basic intuition being modeled is that interpreting an utterance can lead to a change in the information state of the interpreter. Initially, this is modeled in systems in which propositions are Boolean combinations of propositional atoms, and subsequently in systems that admit more representational expressivity, with predicate–argument structure and truth depending on what names and pronouns refer to. The basic contrast between static and dynamical systems among models of formal semantics is a contrast between the meaning of a sentence being characterized by the conditions in which it is true (i.e., truth-conditional semantics) and the meaning of a sentence being given by the change of state induced by interpreting it in some context.

STATIC SEMANTICS

In formal semantics, there are two complementary traditions: the model-theoretic approach, and the use of intervening representation languages. Models can be viewed as representations, but in the model-theoretic paradigm, semantic phenomena are analyzed in terms of the constraints they impose on models into which sentences are directly interpreted. When indirect interpretation into an intervening representation language is adopted, formal semanticists typically rely on a logic. Many logics, with varying degrees of expressivity, exist: propositional logic, first-order (predicate) logic, and higher-order logics, with modal variations of each. Semanticists who use logics as representation languages are concerned with the translation of natural-language utterances into (as natural as possible) formal representations, while logicians clarify the mathematical properties of the corresponding languages. The advantages of logics include the fact that at least their semantics are well understood, even if the object language of study, natural language, is not. Thus, identifying alternative candidate translations from natural-language expressions into formal logics clarifies what semantic properties the natural expressions have. A logic is a formal language with a clear syntax and a clear semantics, as well as a deduction relation, which expresses what sentences must be true given the truth of a base set of sentences.

Certain forms of dynamic semantics follow from assuming that the initial base set of sentences is only partial (i.e. a sentence may be neither true nor false, but undefined), and that dynamics ensue from maintaining consistent sets of sentences following from the deduction relation while augmenting the base set. Other models of dynamics are insensitive to whether the base set is partial or not, because they effectively allow updating (i.e. addition of new sentences) as well as 'downdating' (i.e. deletion of old sentences), still maintaining deductive consistency everywhere else. The theory of mental models (e.g. Johnson-Laird, 2001) is an example of an approach based on fleshing out partiality in a system that is fitted to actual cognitive behavior.

To place the relatively large space of dynamical systems on comparable terms, it is helpful to consider a standard static semantics for a basic logical language, and to then indicate the various parameters under which the system might be dynamized.

A Propositional Language

An infinite propositional language can be constructed from a basic set of atomic propositions $\{p^0,\ldots,p^n\}$ and the following syntax rules for composite sentences:

- If p^i is a well-formed sentence, so is $\neg p^i$.
- If p^i and p^j are well-formed sentences, so are:
  - $p^i \wedge p^j$
  - $p^i \vee p^j$
  - $p^i \rightarrow p^j$

These syntax rules are more than are necessary under the standard interpretation of the connectives, but they illustrate how finite means can be used to provide an infinite number of sentences. While the rules indicate the sentences that must be grammatical, they do not stipulate what sentences must be true. That is the role of the semantics for a system, and the semantics is always relative to possible valuations. One way to consider the semantics is to say that relative to a particular valuation function, certain atomic propositions are true, and others false (and composite propositions' truth or falsity is determined compositionally). Another way of thinking about the meaning of sentences in the language is to say that the sentences all denote the space of valuation functions that make them true. These two possibilities are illustrated below. A great deal of work in dynamic semantics takes the latter view; but it is really just an abstraction from the former view.

When providing a semantics for a propositional language, it is common to outline a valuation function v which tells whether each atomic proposition is true or false. Another function $[\![\cdot]\!]$ is defined to deterministically identify the meaning of a complex expression from the meanings of its components, relative to the choice of valuation function for the basic expressions. Thus, at the outset v is used to map each p^i unambiguously into the set {1, 0}, where 1 represents truth and 0 represents falsity. The meaning function for arbitrary sentences is commonly defined as follows:

- $[\![p^i]\!]^v = v(p^i)$ iff p^i is atomic.
- $[\![\neg p^i]\!]^v = 1$ iff $[\![p^i]\!]^v = 0$.
- $[\![p^i \wedge p^j]\!]^v = 1$ iff $[\![p^i]\!]^v = 1$ and $[\![p^j]\!]^v = 1$.
- $[\![p^i \vee p^j]\!]^v = 1$ iff $[\![p^i]\!]^v = 1$ or $[\![p^j]\!]^v = 1$.
- $[\![p^i \rightarrow p^j]\!]^v = 1$ iff $[\![p^i]\!]^v = 0$ or $[\![p^i]\!]^v = 1$.

Consider the meaning of the proposition $p^1 \vee p^2$. Given a complex sentence made up from n distinct atomic propositions, there are 2^n distinct valuation functions to consider – in this example, four possible valuation functions. The value of $[\![p^1 \vee p^2]\!]$ depends on the particular valuation function chosen (see Table 1). Using v^1 it is false, but using v^2, v^3, or v^4 it is true.

The generalization of this view is to let $[\![p^i]\!]$ denote, not a truth value, but the set of valuation functions that make the proposition true. We write v below for the class of possible valuation functions:

- $[\![p^i]\!] = \{v^j | v^j(p^i) = 1\}$
- $[\![\neg p^i]\!] = v - [\![p^i]\!]$
- $[\![p^i \wedge p^j]\!] = [\![p^i]\!] \cap [\![p^j]\!]$
- $[\![p^i \vee p^j]\!] = [\![p^i]\!] \cup [\![p^j]\!]$
- $[\![p^i \rightarrow p^j]\!] = (v - [\![p^i]\!]) \cup [\![p^j]\!]$

Using the example above, $[\![p^2]\!] = \{v^2, v^4\}$ and $[\![p^1 \vee p^2]\!] = \{v^2, v^3, v^4\}$.

A few immediate observations are relevant to subsequent dynamic reinterpretations. A contradiction denotes the empty set: no valuation will make it true. A tautology denotes v: it is true no matter what valuation function is selected. A

Table 1. How $[\![p^1 \vee p^2]\!]^v$ depends on the valuation function v.

v	$v(p^1)$	$v(p^2)$	$[\![p^1 \vee p^2]\!]^v$
v^1	0	0	0
v^2	0	1	1
v^3	1	0	1
v^4	1	1	1

conjunction denotes an intersection: a subset of the set of valuations denoted by either conjunct individually. If each valuation provides a truth-value for each atomic proposition, then each valuation represents a total specification of how things might be: a possibility. A conjunction then involves an intersection of sets of possibilities. Therefore, a conjunction involves eliminating those possibilities not in the intersection. Intuitively, if a proposition represents the meaning of a sentence, then a conjunction of propositions represents a discourse, and interpreting the discourse as true eliminates possible ways the world might be.

A First-order Language

Generalizations to first-order languages are necessary because propositional languages do not capture the predicate–argument structure inherent in natural-language syntax. Arguments (apart from those of attitude verbs like 'knows') are typically individual entities syntactically marked as nominals, and predicates are typically relations, often marked as verbs. In addition to predicate–argument structure, it is desirable to have variables to capture anaphoric relations among sentences. Along with variables come possibilities for binding them, such as with existential or universal quantification. Here we will ignore issues associated with dynamics and generalized quantifiers such as 'most' and 'few'; however, see Kanazawa (1994). In dealing with a predicate logic, it is necessary to augment the syntax of the language with a set of variables (Var) and a set of individual constants (Cons); these represent the arguments of predicates. Additionally, propositions are regarded as predicates with no arguments (zero-arity) and predicate–argument structures are well formed only if an n-ary predicate combines with a sequence of exactly n constants or variables.

In the semantics, the valuation functions require more structure. In addition to the notion of truth and falsity, a domain of individuals is required, for variables and individual constants to refer to. Each $v^i \in v$ will be defined in terms of three parts: a function N that maps each constant to an element of the domain D, a function g that maps variables in an expression into D (on analogy with pronouns, which don't always refer to the same individual in discourse), and an interpretation function I for relation names that takes over the propositional role of v and generalizes it to predicates of arbitrary arity. Each of these relations is defined solely for its intended sort (i.e., N for names of individuals, I for predicate names, and g for pronouns (variables)); and for each v^i, N is injective (onto) and I and g are bijective (one-to-one and onto).

$$v^i(t) = \begin{cases} N(t) \in D & \text{iff } t \text{ is a constant} \\ g(t) \in D & \text{iff } t \text{ is a variable} \\ I(t) \subseteq (\overset{n}{\text{x}}D) & \text{iff } t \text{ is a predicate and} \\ & \overset{n}{\text{x}}D \text{ is the } n\text{-place} \\ & \text{cartesian product over} \\ & \text{the domain } D \end{cases} \tag{1}$$

Thus, many v^i agree in their specification of what names (N) denote and what relations are indicated by predicates (I), but differ on variable assignments (g), as many assignments will agree on all of N, all of I, and all but one variable specified by g. Moreover, variation is possible on each of the other parameters as well. Let $v^i \overset{N}{=} v^j$ indicate that the two valuations v^i and v^j have identical denotations for constants, and similarly let $v^i \overset{I}{=} v^j$ mean that the two valuations have identical interpretations of predicates. With such an arrangement, there are only three clauses to add to the propositional semantics (changing the semantics for atomic propositions, but retaining the clauses for composite sentences):

- $[\![p^i_n(t^1,\ldots,t^n)]\!] = \{v^j | \langle v^j(t^1),\ldots,v^j(t^n)\rangle \in v^j(p^i_n)\}$. A predicate of arity n denotes the set of valuations that make it true, allowing for the possibility of variance in the interpretation of constants, predicates, and variable assignments, where a valuation makes it true just if the sequence of elements denoted by the arguments (according to the valuation) is an element of the interpretation of the predicate name (according to the same valuation).
- $[\![\exists x P]\!] =$

$$\left\{ v^i \;\middle|\; \left(\begin{array}{ll} \{v^j | v^i \overset{I}{=} v^j\} & \cap \\ \{v^j | v^i \overset{N}{=} v^j\} & \cap \\ \{v^j | [\![P]\!] \neq \emptyset\} & \cap \\ \{v^j | v^i(y) = v^j(y) \text{for all } y \in \text{VAR except possibly } x\} & \cap \\ \{v^j | \text{for some } d \in D \text{ there is a } v^j{:}x^j(x) = d\} & \end{array} \right) \neq \emptyset \right\}$$

  An existentially quantified formula denotes the set of valuations in which the matrix formula (P) is true, where also the interpretation of all but the quantified variable (x) is held constant, and such that the quantified variable can point to some entity (randomly chosen) that makes the predication true, holding the interpretations of other variables, names and predicates constant.
  The formal representation above separates these issues ($\{y|q(y)\}$ refers to the set of things that have the property q). The property that each v^j in the denotation of $[\![\exists x P]\!]$ has when interpreted relative to a particular valuation v^i is that the intersection of a number of properties associated with the v^j are nonempty: the

valuations all agree on interpretations of basic predicates with v^i; the valuations all agree on the interpretations of constants; the valuations make the matrix formula P true; the valuations agree on the assignment of all variables except for possibly x, the quantified variable; the valuations make x point to some element of the domain such that the intersection with the set of valuations that make P true is nonempty (along with the intersections of the other properties).

- $[\![\forall x P]\!] =$

$$\left\{ v^j \;\middle|\; \left(\begin{array}{cc} \{v^j | v^i \stackrel{I}{=} v^j\} & \cap \\ \{v^j | v^i \stackrel{N}{=} v^j\} & \cap \\ \{v^i | [\![P]\!] \neq \emptyset\} & \cap \\ \{v^j | v^i(y) = v^j(y) \text{for all } y \in \text{VAR except possibly } x\} & \cap \\ \{v^j | \text{for all } d \in D \text{ there is a } v^i{:}v^j(x) = d\} & \end{array} \right) \neq \emptyset \right\}$$

A universally quantified formula denotes the subset of all those valuations in which the matrix formula (P) is true, where also the interpretation of all but the quantified variable (x) is held constant, and such that the quantified variable can point to every entity in the domain, making the predication true while still holding the interpretations of names and predicates constant.

It is common in the presentation of first-order languages to separate out the relativization of interpretation to the domain, the assignment of interpretations to terms and predicate names, and variable assignments, just as in the initial presentation of the semantics for a propositional language above (e.g., Gamut, 1991; Chierchia and McConnell-Ginet, 1992). In the preceding presentation the choice of domain is a background parameter, and the other three components are all taken as part of the basic valuation functions, so that sentences may uniformly denote sets of valuations containing the parameters most likely for dynamic analysis to attend to. Presentation using somewhat monolithic valuation functions is equivalent to the alternatives, but facilitates discussion of dynamics.

DISCOURSE DYNAMICS

In the static semantics presented so far, a formula denotes the set of valuation functions that make the formula true. The essential idea of dynamic semantics, whether in the propositional or first-order setting, is to associate with a formula two sets of valuation functions: the set of valuation functions that made all preceding formulae true, and the related set of valuation functions that make the current formula true. Thus, there is an input and an output set of valuation functions. A natural parameter for a dynamic theory to stipulate is the set of constraints that must hold between input and output valuations (e.g., one candidate is that the output valuations should be a subset of the input valuations).

As before for propositional systems, formulae may be evaluated as true or false relative to input valuations against which subsequent formulae are evaluated, or expressions may simply denote pairs of sets of input and output evaluations. In an irreducibly dynamic system, this can mean that the intepretation of some connectives is no longer commutative. Conjunction, for example, is in traditional systems taken to be commutative: if $p^1 \wedge p^2$ is true, then so is $p^2 \wedge p^1$. However, under a dynamic analysis, even if the two composite formulae have the same input and output valuations, within the formulae the situation is different – in the first expression p^2 is interpreted relative to the meaning of p^1, and in the second one it is not. Thus, internally the formula p^2 has a different meaning depending on where it is interpreted and what has been interpreted before it.

Recalling that valuation functions essentially encode possibilities, noncommutativity based on input and output valuations has been used in update semantics to explain the contrast in the pairs of discourses below (Veltman, 1991):

(a) It might be cold. It isn't cold.
(b) It isn't cold. It might be cold.
(c) It isn't cold. It might have been cold. (2)

Discourse 2(b) is usually perceived to be rather more odd than 2(a). The explanation is that in 2(b), the possibility of its being cold is eliminated before it is asserted to exist. Discourse 2(c) provides a paraphrase of a possible exegesis of 2(b) that makes it less strange; notice that it introduces an additional parameter of varying tense. In a propositional setting, when all that exists are truth values and valuations of proposition names, all that can be dynamized is which proposition names are true or false. Schütze (1996) provides a useful discussion of the foibles associated with grammaticality and acceptability judgments. Statements about ill-formedness made in this article are given with his caveats in mind.

In a first-order system, the domain is richer, and so are the valuations. As presented so far, the first-order system is intended to capture the spirit of 'discourse representation theory' (DRT) (Kamp, 1981); Kamp and Reyle, 1993) or 'file change semantics' (Heim, 1982), formalizing constraints on anaphoric relations across discourse, based on the 'dynamic predicate logic' of Groenendijk and Stokhof (1991). The basic ideas include the notions

that indefinite noun phrases introduce entities that may be referred to again, subject to accessibility constraints, that proper noun phrases are universally accessible, and that pronouns and definite noun phrases are anaphoric to accessible antecedents.

(Cataphora (e.g. *He$_i$ asked for a martini. Bond$_i$ further specified, 'stirred, not shaken'*) and generic noun phrases (e.g. *the dodo is extinct*) are not outside the scope of DRT, but are outside the scope of this article. Generiticy has received much important attention in a range of semantic frameworks, for example from Carlson and Pelletier (1995).)

Dynamics associated with anaphora is intertwined with Accessibility, correlates with the syntactic constructions (logical connectives) in which a discourse referent is introduced. These are all tied to the variable-binding components of the valuations (g), as the difference between a new and old discourse referent is in the existence of assignments for the variable corresponding to the discourse referent.

A variable introduced inside the scope of a basic sentence is accessible to a subsequent discourse-level conjunct. A variable introduced inside the scope of a negation is inaccessible to a conjunct of that negated sentence. For conditionals, a variable introduced in the antecedent is accessible to the consequent, but not outside the conditional as a whole. Data motivating these constraints include the following:

(a) A person is walking. The person is smiling.
(b) The person is walking. A person is smiling.
(c) The person is smiling. A person is walking. (3)

Neither 3(b) nor 3(c) induces coreference between the walker and the smiler: the antecedent to anaphoric requirements of the definite noun phrase is in both of these cases inaccessible. However, in 3(a) the variable is accessible to subsequent discourse.

(The notion of a 'basic sentence' is difficult. Keenan (1975) provides an operational definition as a sentence that doesn't require the understanding of other sentences for its interpretability for the purposes of identifying a family of properties cross-linguistically associated with subjects. For example, a sentence containing a verb that embeds a sentence is not basic. In contrast, in this literature, if a sentence is not marked by peculiar constraints on accessibility related to embedding, then it can be taken as basic.)

The following examples demonstrate constraints imposed by negation:

(a) Jan owns a bike.
A bike is in the stairwell.
The bike is rusty.
(b) Jan owns a bike.
A bike isn't in the stairwell.
The bike is rusty. (4)

While 4(a) allows the rusty bike to be identical with Jan's bike or the bike in the stairwell (or both), there is a strong preference for the discourse in 4(b) to disallow identity of reference among the three mentions of 'bike'.

The following examples show that conditionals also influence the information available for the interpretation of subsequent sentences:

(a) If Lee owns a bike it is rusty and it is in the stairwell.
(b) If Lee owns a bike it is rusty.
It is in the stairwell. (5)

In both cases an indefinite is introduced inside the antecedent of the conditional, the 'if' part. In 5(a) two pronominal references are made back to the entity introduced, both in the consequent of the conditional, the 'then' part. There is a contrast in 5(b): a pronoun in the consequent of the conditional may refer back to the entity introduced in the antecedent, but an entity introduced inside the conditional is not accessible to pronouns in separate sentences. While interpreters regularly accommodate discourses, finding antecedents where none explicitly exists (for a DRT treatment of presupposition accommodation, see van der Sandt, 1992), there is a difference in ready interpretability between 5(b) and 5(a). Structurally, the difference is in whether the anaphor occurs in the consequent of a conditional whose antecedent introduces the referent or whether the anaphor occurs outside the conditional in a discourse sister.

Dynamics, applied here, are tied to the way in which valuations are passed among constituents in various clause types and among discourse connectives. Importantly, the DRT tradition is aimed at stating which referents are accessible, not in resolving anaphora. Other dynamic frameworks have been developed to provide theories of which accessible antecedents are most likely to be anchors to anaphors. A well-known approach uses the notion of 'focus' within the discourse to explain the changes through the discourse in what anaphors may refer back to (Grosz and Sidner, 1986). Essentially these are theories that interface syntax, semantics, and pragmatics in suggesting what constrains the accessible referents for definites and

pronouns. Grosz (1977) emphasizes the importance of a mechanism for defining the relevance of the various things that might be referred to for the reference resolution at hand. Another parameter is an encoding of world knowledge that allows explicit mention in discourse of some entity or event to give access to other relevant entities, such as relevant parts of a complex object or sub-events of heterogeneous activities. The final parameter is a mechanism for changing the focus on potential referents through the course of dialogue. One mechanism is to model potential referents as existing in focus sets determined by relevance at their introduction in a stack of available focus sets until, for example, a discourse relation (e.g., 'on a new topic …') allows their accessibility to be eliminated.

In general, dynamics can apply to more than variable assignments – they can apply also to names of constants and interpretations of predicates, the other two components of valuations as set out above. The domain parameter can also be dynamized, as well as the formulae themselves (van Benthem, 1986). Vogel (2001), for example, analyzes first uses of metaphors, the introductions of new senses for terms and predicate names, via a dynamic view on the N and I components of valuations.

DIALOGUE DYNAMICS

The discussion so far has been strongly influenced by a discourse view of the phenomenological domain of semantics: that is, sequences of sentences in monologues. Chiefly, this results in eliminations of possibilities (successively smaller sets of valuations) across sentences conjoined in discourses. Of course, dialectic discourses exist, but these have more in common with dialogue than with narrative monologue. In dialogue settings, more than narrowing of possibilities is required, and even more than propositions are necessary as sentential contents: also necessary as potential contents are questions as bona fide semantic objects (e.g. that which is embedded by 'wonder' and 'know', but not by 'believe' (Ginzburg, 1995a, b)), acceptances of informational contributions (e.g., 'OK', 'yeah') and rejections (e.g. 'no way', 'you're wrong'). To the extent that negotiation enters and individuals revise beliefs, it is necessary for dynamic semantics to attend to the flow of information and beliefs among interloctuors. This area has been explored extensively at the propositional level (e.g., Alchourrón *et al*., 1985; Gärdenfors, 1986) and at the first-order level (Lemon, 1998). While much work in dynamics has been followed in discourse settings, some of the original motivations actually follow from dialogue situations.

Presuppositions of an utterance are the background assumptions that must be true (or accepted as true by speaker and hearer) in order for an utterance to make sense, independently of whether the utterance itself is true. Beaver (1997) provides a comprehensive review of both static and dynamic approaches to presuppositions; the literature is enormous. A brief overview of the sort of problems involved follows.

Consider the following utterances:

(a) The train is late.
(b) The train is not late.
(c) If there is a train, the train is late.
(d) There is a train. (6)

Both 6(a) and 6(b) presuppose that there is a train (6(d)). However, 6(c) does not have this presupposition. The difference in presuppositions between 6(a) uttered in an empty context (empty apart from background information about the meanings of words, etc.) and 6(c) uttered in an empty context is that 6(c) makes sense, but 6(a) does not. The utterance of 6(a) requires at least 6(d) in the context prior to interpreting the same sequence of words, 'the train is late' that also occurs in 6(c). Thus, the presuppositions of an utterance depend on the utterances that precede it. Such phenomena include a great deal more than just definite noun phrases (e.g., questions presuppose that answers exist). It is typical to model these phenomena by taking contexts as sets of valuations, and requiring that the contexts also make true the material presupposed by a sentence. Context must be sensitive to common knowledge. In a dialogue setting, it makes sense to use a definite noun phrase like 'the train' only if there is a unique entity that is salient to speaker and hearer as a train. Thus, the speaker must have a model of what the hearer knows, even if that model is not accurate. In practice, it has been demonstrated that speakers are not always sensitive to hearer's perspectives: Schober (1993) found that speakers with interactive partners were more egocentric in making references than speakers who spoke to imaginary partners, effectively in monologues. Nonetheless, in dialogue semantics it is clear that each speaker must maintain some sort of model of information that is shared, and that the course of dialogue induces updates to those models.

As indicated above, dialogue involves semantic contributions in addition to declarations of

propositions. There are also acceptances of propositional content (and acceptances or denials of presuppositional content) and questions. There are also other things, like exclamations, that have received more attention in pragmatics than in dynamic semantics. Obviously, the flow of information included in questions, responses, and the acceptance of answers is a dynamic process. This is perhaps less obvious in human–machine dialogue systems, in which one interlocutor (the machine) is supposed to be effectively omniscient about its domain and able to answer questions for the user about the domain. However, even in these systems the machine must have a model of dialogue interactions if it is to be easily used, even if only so that the human can make use of pronouns rather than making fully explicit all of the intended background restrictions that identify a noun phrase for each intended referent. Models of the syntax and semantics of dialogue contributions are very active areas of research (e.g. Ginzburg and Sag, 2000). The model of dialogue explored by Ginzburg and Sag involves tracking the 'resolvedness' of questions under discussion with stack-based mechanisms like those for resolving nominal references in the focus-tracking system of Grosz and Sidner (1986) mentioned earlier.

OPEN QUESTIONS

Dynamic semantics, in all its manifestations in formal semantics, involves data structures that model basic sorts of information, and processes that increment or decrement those information stores in accordance with the flow of sentences interpreted. Some representations of information are more expressive than others, and some uses of language require more sorts of semantic objects than others. Many models of dynamic semantics exist, differing in the ontology modeled, the technical details of the modeling, and the phenomena addressed. Primary open questions concern the unification of dynamical approaches to ranges of phenomena in comprehensive systems, and evaluation of their predictions with respect to the robustness of actual human behavior, and not simply introspectively circumscribed subsets of the phenomena.

References

Alchourrón CE, Gärdenfors P and Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic* **50**: 510–530.

Beaver D (1997) Presupposition. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 939–1008. Amsterdam, Netherlands: Elsevier.

van Benthem J (1986) *Essays in Logical Semantics*. Dordrecht, Netherlands: Reidel.

van Benthem J (1996) *Exploring Logical Dynamics*. Stanford, CA: CSLI.

Carlson G and Pelletier J (eds) (1995) *The Generic Book*. Chicago, IL: University of Chicago Press.

Chierchia G and McConnell-Ginet S (1992) *Meaning and Grammar: An Introduction to Semantics*. Cambridge, MA: MIT Press.

Gamut L (1991) *Language, Logic and Meaning, Part 1: Introduction to Logic*. Chicago, IL: Chicago University Press.

Gärdenfors P (1986) *Knowledge in Flux*. Cambridge, MA: MIT Press.

Ginzburg J (1995a) Resolving questions, I. *Linguistics and Philosophy* **18**(5): 459–527.

Ginzburg J (1995b) Resolving questions, II. *Linguistics and Philosophy* **18**(6): 567–609.

Ginzburg J and Sag I (2000) *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. Stanford, CA: CSLI.

Groenendijk J and Stokhof M (1991) Dynamic predicate logic. *Linguistics and Philosophy* **14**: 39–100.

Grosz B (1977) The representation and use of focus in a system for understanding dialog. In: *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pp. 67–76.

Grosz B and Sidner C (1986) Attention, intention and the structure of discourse. *Computational Linguistics* **12**: 175–204.

Heim I (1982) *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts.

Johnson-Laird P (2001) Mental models and deduction. *Trends in Cognitive Science* **4**(10): 434–442.

Kamp H (1981) A theory of truth and semantic representation. In: Groenendijk J, Janssen T and Stokhof M (eds) *Formal Methods in the Study of Language*, pp. 277–322. Amsterdam, Netherlands: Mathematical Centre Tracts.

Kamp H and Reyle U (1993) *From Discourse to Logic*. Dordrecht, Netherlands: Kluwer.

Kanazawa M (1994) Dynamic generalized quantifiers and monotonicity. In: Kanazawa M and Piñón CJ (eds) *Dynamics, Polarity and Quantification*, pp. 213–249. Stanford, CA: CSLI.

Keenan E (1975) Towards a universal definition of 'subject'. In: Li C (ed.) *Subject and Topic*, pp. 304–333. London, UK: Academic Press.

Kempson R, Meyer-Viol W and Gabbay D (2000) *Dynamic Syntax: The Flow of Language Understanding*. Oxford, UK: Blackwell.

Lemon O (1998) First-order theory change systems and their dynamic semantics. In: Ginzburg J, Khasidashvili Z, Vogel C, Levy J-J and Vallduvi E (eds) *The Tbilisi Symposium on Logic, Language and*

Computation: Selected Papers, chap. 8, pp. 85–99. Stanford, CA: CSLI.

Muskens R, van Benthem J and Visser A (1997) Dynamics. In: van Benthem J and ter Meulen A (eds) *Handbook of Logic and Language*, pp. 587–648. Amsterdam, Netherlands: Elsevier.

van der Sandt R (1992) Presupposition projection as anaphora resolution. *Journal of Semantics* **9**: 333–377.

Schober MF (1993) Spatial perspective-taking in conversation. *Cognition* **47**: 1–24.

Schütze C (1996) *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*. Chicago, IL: University of Chicago Press.

Veltman F (1991) *Defaults in Update Semantics*. Technical Report LP-91-02, Institute for Language, Logic and Information, University of Amsterdam.

Vogel C (2001) Dynamic semantics for metaphor. *Metaphor and Symbol* **16**(1–2): 59–74.

Winograd T (1983) *Language as a Cognitive Process*, vol. I 'Syntax'. Reading, MA: Addison Wesley.

Sentence Comprehension, Linguistic Complexity in

Introductory article

Edward Gibson, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

The process of comprehending a sentence involves the moment-by-moment evaluation of a variety of informational and computational constraints. The combination of these constraints makes a sentence easy or difficult to process.

INTRODUCTION

Most sentences that we encounter in everyday life are easy to understand. But occasionally a sentence can be confusing or complex. For example, consider sentence (1):

The dog walked to the park had been chewing the bone. (1)

While reading (1), people generally get confused when the words 'had been chewing …' are encountered. People experience difficulty like this (a so-called 'garden-path effect') when there is an ambiguity earlier in the sentence for which they follow a likely interpretation, but which turns out to be incorrect. The temporary ambiguity in (1) is initiated at the word 'walked', which is ambiguous between a past tense verb and a passive participle verb. The past tense reading – the reading that is initially preferred – results in a main clause interpretation, in which 'the dog' is the actor (i.e. the agent carrying out the action) in a past-tense walking event. This interpretation is no longer viable when the words 'had been …' are input. The passive participle reading – the reading that is not usually noticed initially – involves 'the dog' as a patient (i.e. to whom the action is done) of a walking event, in which an unmentioned individual walked the dog.

Now consider sentence (2), which presents its readers with a second kind of processing difficulty:

The reporter that the senator that John met at the party attacked admitted the error. (2)

Sentence (2) consists of a main clause 'the reporter admitted the error', whose subject is modified by a relative clause (RC) 'that the senator … attacked'. The subject of the RC is then itself modified by another RC, 'that John met at the party'. This sentence is extremely difficult to understand, independent of any temporary ambiguities that it contains. In particular, sentence (2) is difficult to understand even when reading it for the second or third time. This is not so for (1): once it is known that the passive participle interpretation is the target structure, (1) becomes understandable.

Complicated sentences like (1) and (2) can be highly informative for telling us how the process of sentence comprehension occurs. In ambiguous sentences like (1), the combination of a variety of factors results in preferences for some structures over others, because not all structures can be pursued in parallel. In sentences like (2), without confusing temporary ambiguities, the combination of the same factors results in easier or harder processing, depending on the difficulty of the target interpretation. The analysis of a wide range of sentence comprehension studies involving many different structures across languages suggests that a variety of information sources are used in constructing an interpretation for a sentence, and that the process of building the target representation is constrained by the available computational resources. Some of the informational and resource factors are discussed below.

MEASURING LINGUISTIC COMPLEXITY EXPERIMENTALLY

Studies of sentence comprehension involve comparisons between target sentences and appropriate control sentences. For example, a control for sentence (1) above is a disambiguated version, as in (3):

The dog that was walked to the park had been chewing the bone. (3)

This sentence has the same structure as (1), but is disambiguated by the words 'that was' towards the passive participle interpretation of 'walked'. This sentence is correspondingly easier to understand than (1). A control sentence structure for sentence (2) is given in (4):

At the party, John met the senator that attacked the reporter that admitted the error. (4)

Sentence (4) contains all the same words as (2), with the same thematic relations among them, but (4) is much easier to understand. Thus, it is something about the structure of (2) that makes it difficult to understand.

A single individual's reactions to sentences like (1) and (2) on the one hand and their controls in (3) and (4) are not very informative by themselves. Because of individual variation, it is necessary to test hypotheses regarding sentence comprehension on a range of experimental participants and items. Perhaps the simplest method for gathering data on sentence comprehension is by means of acceptability (or grammaticality) judgments. This method consists of having experimental participants answer a questionnaire in which sentences are rated for their understandability, according to the participants' intuitions. A more objective method involves measuring reaction times to full sentences (presented either visually or auditorily) and accuracy to questions about the content of the presented sentences.

Often we want to know about the time course of processing load in a sentence: where in the sentence does the difficulty begin and end? End-of-sentence (offline) measures do not address such questions. Online measurements are necessary. In reading, measuring participants' eye movements to visually presented sentences is one online technique. At points of high complexity, participants slow down and/or regress to previous regions. A commonly used online alternative to tracking eye movements is self-paced reading with a moving window display. In this method, a sentence is initially presented on a computer screen as a series of dashes marking the length and position of the words in the sentence. Participants press a key (usually the spacebar) to reveal each region of the sentence, often one word at a time. As each new region appears, the preceding region disappears. The amount of time the participant spends reading each region is recorded as the time between key-presses. As in eye-tracking, participants tend to slow down at points of high complexity. Unlike eye-tracking, however, there is no way in self-paced reading for a participant to back up and re-read a region that was confusing.

It is more difficult in auditory language presentation to obtain an online dependent measure such as reading time. One way to obtain an online measure of difficulty in auditory presentation of language is via cross-modal priming. In a cross-modal priming paradigm, participants listen to sentences over headphones, and then perform a different task – such as deciding whether a string of letters is a word – at predetermined locations in the target sentences. People are faster at the interrupting task at points of lower complexity, such as when a word in the linguistic context is semantically related to the target stimulus (e.g. 'doctor' is related to 'nurse'). A second online auditory method involves tracking participants' eye movements with respect to a visually presented scene while the participants listen to instructions about the scene. This method is particularly informative with respect to the question of how the context of the presented scene influences sentence understanding. A third method involves monitoring the scalp with electrodes in order to measure event-related potentials (ERPs): minute voltage changes due to differences in neural activity while participants listen to sentences. This method, which is also used in reading research, requires an understanding of what the voltage changes mean. Some plausible interpretations are currently being proposed in this area.

RESOURCE CONSTRAINTS THAT AFFECT PROCESSING LOAD

In the process of understanding a sentence, it is necessary to integrate structures for incoming words into the structure(s) that have been built thus far, such that the potential integrations for an incoming word are determined by the syntactic rules for the language. According to one current theory – the dependency locality theory (DLT) – the processing cost of integrating a new word w is proportional to the distance between w and the syntactic item with which w is being integrated. Structural integration cost has been shown to be an important factor in accounting for online processing load. For example, consider the RC structures in (5) and (6):

The reporter that attacked the senator admitted the error. (5)

The reporter that the senator attacked admitted the error. (6)

In (5), the RC pronoun 'that' is interpreted as the subject of the verb 'attacked', whereas in (6), the same pronoun is interpreted as the object of the verb 'attacked'. People read the verb 'attacked' more slowly in a sentence like (6) than in a sentence like (5). This difference can be explained by integration distances. In (5), there is one local integration when processing 'attacked': this verb is integrated with the preceding RC pronoun as its subject. In contrast, there are two integrations at the point of processing 'attacked' in (6): this verb must be integrated as the verb for the subject 'the senator' (a local integration) and the object position of 'attacked' must be integrated with the RC pronoun 'that', a nonlocal integration. As a result of the extra nonlocal integration, the processing load at 'attacked' is larger in (6) than in (5), resulting in longer reading times at this word. Furthermore, reading times are slow in both sentence types for the verb 'admitted', a point of long-distance integration with the subject 'the reporter' in both sentence types. Reading times are relatively faster for the other words in the sentences, because integrations at all other positions are local.

Interestingly, aphasic (speech loss) stroke patients understand the subject-RCs in sentences like (5), but they do not understand the object-RCs in sentences like (6), as evidenced by their inability to reliably answer questions about the object-RCs. This suggests that the brain damage suffered by these aphasic patients has reduced the available resources for processing sentences.

An interesting question raised by a distance-based theory of integration cost is how distance is quantified. It appears that complexity may depend not only on the number of words or syllables between two integration points but also on the complexity of the intervening discourse structures. In particular, the ease or difficulty of constructing and/or accessing the referents in the intervening material affects the complexity of integrations across these referents. For example, people read the verb 'attacked' in the object-RC in (6) more quickly when the subject of the RC 'the senator' is replaced with a pronoun such as 'I' or 'you'. This decrease in complexity is arguably due to the fact that the pronouns 'I' or 'you' indicate highly accessible referents in the discourse – the speaker/writer and the hearer/reader – whereas 'the senator' refers to an individual who is not part of the current discourse, in a single sentence paradigm. Reading times for the embedded verb 'attacked' also decrease substantially when there is a referent for the subject noun-phrase (NP) 'the senator' in the current context. This result provides more support for the discourse-based integration cost metric, assuming that it is easier to access a structure for a referent that has just been built than it is to build a structure for a new referent.

Integration distances provide a partial explanation for the extreme complexity of (2). The integrations at the words 'attacked' and 'admitted' are all substantially longer in (2) than in (6), with the inclusion of the RC 'that John met at the party' modifying the NP 'the senator'. Such long-distance integrations, in combination with storage costs at these processing states, may be too complex for the limited capacity of sentence-processing resources. In contrast, all integrations are local in the control for (2) in (4).

The processing of ambiguous structures provides further evidence that longer-distance integrations are more costly. Consider (7):

The bartender told the detective that the suspect left the country yesterday. (7)

The adverbial 'yesterday' can be associated with either the most local verb 'left' or the earlier verb 'told'. The more local integration is strongly preferred. This preference is consistent with the hypothesis that people attempt to minimize integration costs when faced with ambiguity.

At the same time that the process of integration is going on in understanding a sentence, it is also necessary to store the partially processed structures. Both structural integration and storage consume computational resources. According to the DLT, the resources required for storing a partially processed structure are proportional to the number of incomplete syntactic dependencies at that point in processing the structure. As a result, processing load increases when the number of incomplete syntactic dependencies increases. Results from processing ambiguous structures further support the hypothesis that the sentence comprehension mechanism is sensitive to the number of incomplete dependencies in a structure: at choice points, structures with fewer incomplete dependencies are preferred over structures with more incomplete dependencies.

It should be noted that the DLT is just one theory of computational resource use in sentence comprehension. According to an earlier theory, ambiguity resolution is guided by two principles: a locality principle (like the integration component of the DLT), and a principle known as Minimal

Attachment, which prefers phrase structures involving fewer phrase structure rule applications. This theory and the DLT make largely the same predictions with respect to processing ambiguous inputs. It remains an open question precisely how resources constrain sentence interpretation.

INFORMATIONAL CONSTRAINTS THAT AFFECT PROCESSING LOAD

The difficulty of an integration depends not only upon its resource use, but perhaps even more importantly upon the informational complexity of the resulting structure. Recent work has demonstrated that people's preferred interpretations of (temporary) ambiguities in sentences are affected by factors such as (1) the frequency of the different lexical entries for the word being integrated; (2) the plausibility of the meaning of the resultant structure in the world; and (3) the context that the sentence is uttered in. Consider sentence (1) once again:

The dog walked to the park had been chewing the bone. (1)

The preference for initially analyzing the word 'walked' as a past tense verb rather than a passive participle verb is driven by a number of factors. First, the past tense lexical entry for 'walked' is much more frequently used than the passive participle lexical entry for 'walked', which is used in the passive interpretation. Second, although it is plausible for a dog to be walked, it is even more plausible for a dog to be walking, because there is a general bias to treat animate beings as agents or experiencers of events. Third, there is more syntactic storage cost associated with the passive participle structure at the point of processing 'walked': the passive participle structure requires at least a verb to make a complete sentence (and possibly also a modifier for the verb 'walked'), whereas the past tense structure requires no further words to make a grammatical sentence. Note that integration costs do not have a bearing on the preference, because both potential integrations of the word 'walked' are local.

Now consider (8), a sentence with the same structural ambiguity as in (1), but which does not cause processing difficulty:

The evidence examined by the lawyer turned out to be unreliable. (8)

Like the verb 'walked', the verb 'examined' is ambiguous between a past tense and a passive participle. There are three differences between this ambiguity and the ambiguity in (1) that make the ambiguity in (8) much easier to resolve as a passive participle. First, the relative frequencies of past tense and passive participle lexical entries for the verb 'examined' are less biased towards the past tense reading. Second, and most importantly, plausibility information is highly biased towards the passive participle structure: it is plausible for evidence to be examined, as in the passive participle interpretation, but it is not plausible for evidence to examine something. Third, syntactic storage costs are less biased in favor of the past tense reading, because unlike the past tense entry of 'walked', which is optionally intransitive, the verb 'examined' obligatorily requires a noun phrase object. Thus there is a smaller difference in storage costs between the past tense and passive participle structures than for 'walked'.

Referential context also strongly affects people's initial interpretations of ambiguous structures, as evidenced by monitoring people's eye movements while they listen to commands spoken to them over headphones. Consider the command in (9):

Put the apple on the towel in the box. (9)

In a context with only one apple, there is a strong preference to interpret 'on the towel' as the goal for the verb 'put', even if the apple is already on a towel. But in a context with two apples, one of which is already on a towel, the phrase 'on the towel' is initially analyzed as a modifier of the apple, specifying the apple's source location.

Another informational factor which affects people's initial interpretations of an ambiguous input is the intonation of the speech signal. Intonational (or prosodic) properties include variations in the pitch, amplitude, and duration of individual speech sounds and larger segments, as well as the place of pauses. Intonational phrasing which is consistent with the target syntactic structure helps sentence understanding, and intonational phrasing which conflicts with the target syntactic structure makes sentence understanding more difficult.

IMPLICATIONS AND OPEN QUESTIONS

The details of the relative timing and strengths of the resource and informational constraints are currently not known. An influential early hypothesis in the sentence processing literature was the modularity hypothesis: that syntactic preference constraints (constraints related to the resource constraints described here) apply first, followed by informational constraints such as context and plausibility. This hypothesis predicts that there should be reanalysis effects (small garden-path

effects) in instances where syntactic preferences favor one interpretation, and plausibility and/or contextual constraints favor another interpretation. Recent work using eye-tracking methods has failed to observe such a reanalysis effect. For example, there is no measurable initial preference for the goal interpretation of the phrase 'on the towel' in sentence (9) when the visual context favors the source location, despite the fact that syntactic preferences favor the goal interpretation. As a result, an interactive position has gained in popularity, in which resource and information constraints apply immediately to the available alternatives.

The interacting constraints have been described here as if they are all independent of one another. However, this may not be the case. Many researchers argue that language is implemented in a highly interactive architecture, such as a connectionist neural network of some kind. What may superficially look like independent constraints may be emergent properties of such an architecture. Much further research is needed to investigate this issue.

Further Reading

Crain S and Steedman M (1985) On not being led up the garden path: the use of context by the psychological parser. In: Dowty D, Karttunen L and Zwicky A (eds) *Natural Language Processing: Psychological, Computational and Theoretical Perspectives*, pp. 320–358. Cambridge, UK: Cambridge University Press.

Elman JL (1991) Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning* **7**: 195–225.

Frazier L and Clifton C Jr (1996) *Construal*. Cambridge, MA: MIT Press.

Gibson E (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**: 1–76.

Gibson E (2000) The dependency locality theory: a distance-based theory of linguistic complexity. In: Miyashita Y, Marantz A and O'Neil W (eds) *Image, Language, Brain*, pp. 95–126. Cambridge, MA: MIT Press.

Gibson E and Pearlmutter N (1998) Constraints on sentence comprehension. *Trends in Cognitive Science* **2**: 262–268.

Grodner D, Gibson E and Tunstall S (2002) Syntactic complexity in ambiguity resolution. *Journal of Memory and Language* **46**: 267–295.

Kjelgaard MM and Speer SR (1999) Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguities. *Journal of Memory and Language* **40**: 153–194.

MacDonald M, Pearlmutter N and Seidenberg M (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703.

McElree B and Griffith T (1998) Structural and lexical constraints on filling gaps during sentence comprehension: a time-course analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 432–460.

Tabor W, Juliano C and Tanenhaus MK (1997) Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* **12**: 211–272.

Tanenhaus MK, Spivey MJ, Eberhard KM and Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* **268**: 1632–1634.

Tanenhaus MK and Trueswell JC (1995) Sentence comprehension. In: Miller J and Eimas P (eds) *Speech, Language, and Communication*. San Diego, CA: Academic Press.

Trueswell JC, Tanenhaus MK and Garnsey SM (1994) Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language* **33**: 285–318.

Sentence Processing: Mechanisms

Advanced article

Matthias Schlesewsky, University of Potsdam, Potsdam, Germany
Angela D Friederici, Max Planck Institute of Cognitive Neuroscience, Leipzig, Germany

CONTENTS

Mechanisms of sentence processing during language comprehension are described on the basis of behavioral and electrophysiological evidence. The major principle underlying human parsing mechanisms is the principle of simplicity, which leads the parser to construct the simplest syntactic structure whenever possible.

INTRODUCTION

Humans communicate primarily by means of language, mostly in the form of sentences. Sentences allow the relation and qualification between different entities in the real world to be expressed. In most languages, sentences contain at least a subject (animate or inanimate) and a verb (qualifying an action or state). In principle a sentence can be infinitely complex (multiple embedding of sentences) and still be grammatical; however, the human parser has been shown to be able to deal with only two or three embeddings at most (Bach *et al*., 1986). Although human parsers do have processing limitations, their ability to process incoming information on-line is remarkable, as is their capability to produce semantically coherent and syntactically correct sentences on-the-fly.

CREATION AND ATTACHMENT OF SYNTACTIC CONSTITUENTS

The building of sentential structure, which serves the purpose of creating an interpretable form, must be considered a process in which a number of mechanisms interact. These include the integration of previously predicted or unpredictable, and thereby unexpected, constituents. This may be illustrated by imagining the extent of the parser's knowledge when the processing of a sentence input begins. While the subject and the verb are obligatory components of every sentence and can therefore be anticipated, complements of the verb become predictable only with the processing of the verb. A determiner or a preposition allows the prediction of nominal complements and thereby the on-line construction of a noun phrase or a prepositional phrase. By contrast, modifying elements important for the interpretation of a sentence, such as prepositions, adverbs, or relative clauses, are generally not predictable and therefore become relevant to processing only at the point in time when they are encountered.

Prediction and integration of obligatory constituents

Most sentence processing models assume in a more or less explicit fashion that the prediction of sentential components plays a significant role during language processing (Gibson, 1998). It is generally assumed that sentence structure is built up in accordance with principles of simplicity that state that only those syntactic constituents necessary for a minimal structure are postulated. Given a language that displays a restricted degree of freedom with regard to the positioning of sentential arguments but does not mark their grammatical function, there are two possibilities of interpreting an ambiguous initial argument. Take for example the relative pronoun *who* in English relative clause constructions such as *the man who visited the ambassador* versus *the man who the ambassador visited*. Following a principle of minimal structure building, a subject interpretation of *who* is preferred as it allows the prediction of only one further constituent, whereas an object interpretation requires the prediction of both a verb and a subject.

Why language processing follows such criteria of simplicity, which are possibly of a general

cognitive nature as they appear to be related to working memory capacities, can be shown when comparing the processing of subject-first and object-first structures in a case marking language such as German. In German, the grammatical function of sentential arguments is overtly marked morphologically and their ordering is relatively free. With unambiguously marked elements in sentence initial position it is immediately clear whether the structure consists of at least three constituents, namely an object, a subject and the verb (if the initial argument is marked for accusative) or, of at least two constituents (if the initial argument is marked for nominative). Results from reading time studies and event-related brain potential (ERP) studies show that the identification of an object in sentence initial position leads to an increase of processing effort. Interestingly, the integration of the dislocated object at its base position induces further processing costs.

For structures containing an ambiguous initial constituent, such processing costs may be avoided by assuming that the structure contains only a single argument; i.e. an initial subject. Thus, the subject interpretation of an initial argument as well as the preference for a subject-first order in transitive structures results exclusively from the endeavor of the human language processing system to create simple syntactic structures. The validity of such a principle, which is generally known as minimal attachment, is also evident when there is a choice with respect to the syntactic complexity of the complement selected by the verb (e.g. *Ruben knows the answer* versus *Ruben knows the answer is correct*). In this case also, the component with the least number of nonterminal nodes is selected – in the present case, the reading with a nominal object complement (Frazier, 1987).

Integration of modifiers

The integration or attachment of modifiers constitutes a further crucial aspect of sentence structure building. As mentioned above, modifier attachment is not a matter of processes that affect the foundations of a sentence but of mechanisms that serve to integrate additional components (Frazier and Clifton, 1996). For example, the prepositional phrase *with the tiara* in the sentence *the king surprised the maid with the tiara* is an interpretationally relevant constituent, yet its presence or absence has no effect on the grammaticality of the sentence. In contrast to obligatory constituents (e.g., subject or verb), modifiers can generally not be predicted and the processes determining their integration are therefore not affected by the complexity-related criteria discussed in the last section.

From a linguistic perspective, an interesting aspect of modifier integration is the difference between various types of modifier. While there are no syntactic dependencies between a prepositional phrase and the noun it modifies, such dependencies are present between a noun and a relative pronoun coindexed with it. Assuming that the human language processing system endeavors to unify the features of head noun and relative pronoun, this unification should encompass not only obligatory syntactic features, such as gender and number, but also facultative ones, such as case and thereby grammatical function. While these unification mechanisms can be observed in the resolution of ambiguous contexts (in German, for example, the case of a marked relative pronoun may be transferred to the ambiguous head noun for the purpose of ambiguity resolution), it is not yet clear whether and how such mechanisms influence attachment preferences. In any case, it cannot be excluded that they are at least partially responsible for the differences observable between the attachment preferences for various modifiers (Konieczny and Hemforth, 2000).

Nonsyntactic influences

A still open question concerns the influence of extrasyntactic factors such as context or plausibility on the building of sentential structure. It appears to be uncontroversial that these factors influence the interpretation of a sentence, for example in the attachment of modifiers. However, it is not clear whether they have an effect on the initial structural analysis of a given sentence or only later during processing. Examples from English indicate an immediate influence of factors such as animacy (compare, for example, *the defendant examined by the lawyer* versus *the evidence examined by the lawyer*, where the prepositional phrase *by the lawyer* gives rise to processing difficulties only in the first but not in the second case). However, data from German, a language allowing both subject-initial and object-initial structures, show that syntactic preferences dominate even when they induce an implausible interpretation. Thus, in the sentence *die Blinde beobachtete die Ärztin* [*the blind woman watched the doctor*] the ambiguous initial noun phrase is interpreted as the subject, even though this is actually implausible, excluded by the semantics of the verb. Further evidence against assuming that all information is initially accessible is

provided by studies by McElree and Griffith (1995) and Friederici (1999, 2002).

PROCESSING OF FILLER-GAP RELATIONS

The perspective on the building of syntactic structures that was presented in the last section presupposes certain basic assumptions, one of which will be discussed in more detail in the following. Both in theoretical and in experimental linguistics there is a debate with regard to the question of whether noncanonical word orders are derived by means of movement operations and whether these are psychologically real in the sense that they may be experimentally captured.

The examination of so-called filler-gap dependencies – the relation between a constituent in a noncanonical position (filler) and its base position (gap or trace) – has a long tradition within psycholinguistics. A starting point for such studies was provided by the attempt to explain experimentally observed preferences with regard to grammatical function (especially the subject preference for an ambiguous initial argument, for example in Dutch, German, or Italian). The observed preference for subject-initiality was derived by means of structurally motivated differences in the distance between the filler ($_i$) and its gap ($_{-i}$). The distance is smaller in subject-initial structures (e.g. *the wombat$_i$ that$_i$ $_{-i}$ saw the hunter*) than in object-initial structures (e.g. *the wombat$_i$ that$_i$ the hunter saw$_{-i}$*). The perspective argued for here is that such preferences may be derived by criteria of simplicity with respect to syntactic structure building, thus speaking against the assumption that preferences result from the filler-gap relation in and of itself. This, however, excludes neither the existence of a filler-gap relation nor its relevance for language processing.

Under the assumption that the gap associated with a dislocated constituent is psychologically real, one should expect this trace position to be experimentally detectable during on-line processing in a manner similar to arguments in a transitive structure. Nicol *et al*. (1994) investigated sentences such as *the policeman saw the boy$_i$ that the crowd at the party accused$_{-i}$ of the crime*, and were able to show a re-activation of the dislocated argument *the boy$_i$* at the anticipated gap position. However, results such as this have been criticized, in as much as in an SVO (subject–verb–object) language such as English, in which the verb and the trace position are adjacent, an anticipation of the gap/trace cannot easily be dissociated from a re-activation of the arguments at the position of the verb. Thus, the re-activation at a position immediately following the verb *accused* need not be the result of anticipating the trace of the dislocated element but could just as well reflect the integration of the argument into the verbal frame and therefore be the result of the processing of the verb.

Data from SOV languages such as German can help to clarify this issue. They provide evidence against the 're-activation at the position of the verb' and therefore in favor of 'trace anticipation'. For example, in sentences with a topicalized object such as *dem Richter$_i$ ist ein Krug$_{-i}$ zerbrochen* [*the judge broke a jug*], the initial dative-marked object *the judge* has been shown to be re-activated before the verb is reached, at the actual trace position (Muckel and Pechmann, 2000). The view that traces are anticipated is further supported by studies examining the processing of *wh*-clauses. Indirect (embedded) object-initial questions such as *Peter fragte, wen am Sonntag nach dem Unfall der Gärtner besuchte* [*Peter asked whom the gardener visited on Sunday after the accident*] show a higher processing load than their subject-initial counterparts. In an event-related brain potential (ERP) study, a sustained negativity set in immediately after the identified filler *wen* [*whom*] had been processed and continued until the subject *der Gärtner* [the gardener] was reached (Fiebach *et al*., in press).

Several pieces of evidence indicate that the dislocated argument is integrated before the verb is processed. First, the additional processing costs for object-initial structures (in the form of the sustained negativity) disappear once the subject is encountered. Second, the processing of the subject elicits a particular ERP component that is associated with increased integration cost (integrative P600), which in this case may be seen as an index of the initial object's integration. These data are a clear argument in favor of the psychological reality of traces.

Thus, filler-gap dependencies occupy a special place within sentence processing and they are associated with an increase of processing costs. Following the principle of simple structure building, it appears that a sentence without dislocated constituents is always preferred over one which entails the postulation of such dependencies.

SYNTACTIC RE-ANALYSIS

Structure building of the human sentence processing system also includes the need to analyze unexpected events and, in the case of a problem, to integrate the new input into the existing structure either by means of a local correction or through a

re-analysis of the previously built structure. In this regard, one may dissociate two separate processes, diagnosis and re-analysis (Friederici, 1998). During diagnosis, the parser must recognize the cause of the local processing problem. During the re-analysis that follows, the language processing system must solve the identified problem by re-interpreting the current input or altering the structure already built. The relation of both processes to one another has been the focus of many studies. One major question is how the processing costs observed experimentally are distributed over both operations. Are the costs only due to diagnosis (Fodor and Inoue, 1994), or is it that re-analysis (i.e. the extent of the corrections required) induces such costs? Or, alternatively, do both play a part in determining the processing problems observed in behavioral studies?

Diagnosis

As stated, a necessary prerequisite for syntactic re-analysis occurring is that the human language processing system recognizes a violation, be it a violation of syntactic structuring principles or an incompatibility with a structural preference. Thus, re-analysis may already fail because the parser does not even recognize the violation. This is, for example, observable in … *and what a performance by the man who some of us thought that maybe the pressure of being the winner of Wimbledon might not let him win* or in the German example *Welcher Jäger aus dem Schwarzwald beobachtete der Gärtner [which-nominative hunter from the black forest observed the-nominative gardener*]. Both the resumptive pronoun *him* in the English sentence and the second nominative NP der *Gärtner* [*the gardener*] in the German sentence induce an ungrammaticality which, however, is not perceived during processing. The ungrammaticality in the first example even makes the structure easier to process. Just as problematic are the structures in which parsers falsely recognize a violation of a preference-based expectation (illusion of ungrammaticality) the source of which, however, they cannot locate. This occurs, for example, in grammatical but not processable structures such as the English sentence *the daughter of the pharaoh's son admires himself* or the German sentence *dass Caroline die Studentin hilft* [*that the student helps Caroline*]. In the first case, the processing problem occurs because the reflexive pronoun *himself* cannot be bound to the head of the phrase (*the son of the daughter of the pharaoh*) owing to a wrong interpretation of the complex noun phrase (*the daughter of the son of the pharaoh*). In the second case the interpretation fails because it is not possible to interpret *Caroline* as a dative object in a noncanonical position owing to a lack of case marking. Thus, in both cases it is the inability to categorize the problem that leads to an abortion of processing. Whether this may be taken as an argument that diagnosis alone is responsible for the cost of so-called garden path effects will be discussed in more detail below.

Before turning to this question, however, we will consider a further aspect of diagnosis that touches upon fundamental questions of human language processing. With regard to the question of whether the initial integration of a word into the existing structure is based not only on syntactic, but also semantic, contextual, or world-knowledge information, we have already cited the papers of McElree and Griffith (1995) and Friederici (1999, 2002). Both argue on the basis of experimental findings that different violations such as (a) phrase-structure and (b) thematic/semantic violations are detected at different points in time (e.g. (a) **some people rarely books* versus (b) *some people alarm books* or (a) **der Honig wurde im geschleudert* [**the honey was in the centrifuged*] versus (b) *der Honig wurde im Keller ermordet* [*the honey was murdered in the cellar*]. The results show that phrase structure information is accessible to the parser earlier than thematic/semantic information. Furthermore, examples such as **der Honig wurde im ermordet* [**the honey was in the murdered*] which violate both the syntactic and the thematic/semantic requirements show that the diagnosis of a phrase structure violation may block thematic/semantic processes following it. These findings suggest that both the structure building and the analysis of a sentence proceed in a hierarchical manner with the parser first analyzing word category information before taking non-structural factors into account.

Re-analysis

While behavioral experimental techniques (such as reading- or reaction-time studies or eye-tracking experiments) cannot always provide an insight into the fine-grained temporal structure of experimentally observable processing problems and thereby do not allow conclusions as to whether these stem from diagnosis or re-analysis, the measurement of ERPs provides the possibility of unraveling these processes. This is again illustrated by an example from German. Assuming that an ambiguous initial constituent is preferentially interpreted as a subject, the re-analysis required for the revision of such a preference can be shown to differ depending on the type of clause in which the

re-analysis is initiated. Thus, the re-analysis required in an object relative clause such as *das ist die Botschafterin, die die Minister eingeladen haben* [*this is the ambassador who the minister invited*] is apparently much less extensive and thereby less costly than that in an analogous object-initial complement clause such as *Peter hörte, dass die Botschafterin die Minister eingeladen haben* […*that the minister invited the ambassador*]. In terms of ERP effects, this is shown by the fact that both the latency and the duration of the positivity are functionally dependent on the ease/difficulty of re-analysis processes (P345 versus P600/syntactic positive shift) (Friederici, 1998). This difference in the ease of re-analysis between relative clauses and complement clauses is derivable from a structural perspective. In relative clauses, only the position of the trace that is coindexed with the initial constituent must be 'relocated' (from after *die Minister* to after the relative pronoun *die*), whereas in complement clauses an additional position must be created for the clause initial object (after *die Minister*). Thus, it appears that the cost of processing is influenced by the type of re-analysis required.

To summarize, there is evidence that processing costs are caused by both diagnosis and re-analysis processes. While the ERP reflection of diagnosis cost differs in polarity (negativity/positivity) as a function of violation type (outright syntactic violation versus incompatibility with a structural preference), the costs of re-analysis are always expressed in a positivity. Outright syntactic violations like in **der Honig wurde im geschleudert* [*The honey was in the centrifuged*] induce an early left anterior negativity (ELAN) followed by a late positivity (P600). Translating this into processing costs, both the diagnosis of the phrase structure violation (in form of the ELAN) and the attempt at a structural re-analysis (visible in the P600 component) contribute to these. Violations of structural preferences such as object-first structures, in contrast, only elicit positivities which, however, are shown to house two factors, one reflecting processes of diagnosis (a positivity with an occipital distribution) and one reflecting the actual processes of recomputation (a positivity with a centro-parietal distribution) (Friederici *et al*., 2001).

SERIAL VERSUS PARALLEL PROCESSING

A further question that is central to psycholinguistic research is concerned with whether language processing proceeds in a serial fashion (Frazier, 1987), with only one analysis carried out at a time, or in a parallel fashion (at least for subjects with good working memory capacity – Just and Carpenter, 1992). This question is often equated with the debate as to whether language processing proceeds in a modular (Frazier, 1987) or an interactive (MacDonald, 1994; McClelland *et al*., 1989; Marslen-Wilson and Tyler, 1980) fashion, but it must be stated clearly that both questions are actually orthogonal to one another.

With respect to modularity/interactivity – whether all types of information are available to the parser at the same time (interactive) or the parser accesses these in a hierarchical manner (modular) – the above cited articles by Friederici (1999, 2002) and McElree and Griffith (1995) show that the latter (hierarchical access to information) is a basic property of the human language processing architecture. These studies used techniques which are highly sensitive to the temporal aspects of online processing.

By contrast, the data in favor of an interactive model – one claiming that syntactic and nonsyntactic information interact during sentence processing – are generally taken from studies employing experimental techniques that are clearly less sensitive to temporal processing aspects than event-related brain potentials or the speed–accuracy tradeoff method. Thus, the question of whether these two information types interact may be reducible to the question of *when* syntactic and nonsyntactic information interact.

One ERP study systematically crossing lexical-semantic incongruity with syntactic incongruity suggests that both information types are processed independently during the first 400 milliseconds (in the form of an N400 for lexical-semantic aspects and in the form of a left anterior negativity (LAN) for morpho-syntactic aspects), but interact about 600 milliseconds after the critical lexical item is encountered (Gunter *et al*., 2000). Further research must reveal to what extent these findings generalize over different linguistic material with respect to the question of whether language processing proceeds in a modular or in an interactive manner. There are a large number of different theories in this regard, ranging from the assumption that the parser only has access to a single (syntactically based) analysis to the diametrically opposed assumption that the parser calculates all alternatives at all times. The latter appears implausible, especially in view of the existence of clear parsing preferences. With regard to the models that are situated between the two extremes, several theses are particularly worth mentioning: first, the assumption that the parser locally calculates several

alternatives but only continues with one (local or momentary parallelism; e.g. Altmann and Steedman, 1988); second, the idea that the parser calculates various alternatives, which are then pursued hierarchically, with a preference for one analysis (ranked parallel) (Hickok, 1993). Let us consider some of the structures discussed above in order to examine this controversy more closely.

One variant of the serial approach assumes that modifiers, in contrast to obligatory constituents, are first associated with the structure and only integrated when sufficient information of other types (e.g. verb and world-knowledge information) is available. Such an approach can explain, for example, why so-called NP–V ambiguities exhibit no measurable reaction time difference on the prepositional phrase between the sentences *he saw the butcher with the binoculars* and *he ran over the hamster with the birthmark*. These findings, however, cannot *per se* be taken as evidence for or against serial or local/ranked parallel parsing, as one could argue that secondary processing mechanisms, though active, may not be relevant with regard to the basic structure of a sentence (Frazier and Clifton, 1996).

Evidence that the parser does not always make a local decision with respect to structural analysis, and that it is not principally restricted to a single analysis, stems from the domain of object–object ambiguities. While both intuitively and experimentally there is no evidence between a final verb that selects for an accusative and a verb requiring a dative object (e.g. *der Professor wollte Studenten unterstützen/helfen* [*the professor wanted to support/help students*]), a preference for an accusative interpretation of the ambiguous object *Studenten* manifests itself when the argument is no longer adjacent to the verb (e.g. *der Professor wollte Studenten, die die erste Prüfung nicht bestanden hatten, unterstützen/helfen* [*the professor wanted to support/help the students who had not passed the first exam*]). This example clearly shows that the immediate disambiguation on the following element does not reveal the underlying, and perhaps still developing, accusative preference for a locally ambiguous object.

There is, however, also uncontroversial evidence that the parser tends to make decisions immediately, even if they could theoretically be postponed until the next element has been encountered (Crocker, 1994). This is the case, for example, in English NP–S ambiguities (*Stefan knows the paper/the paper shows some problems*) and in German subject–object ambiguities (*welche Radfahrerin beobachteten die Polizisten* [*which bike rider did the policemen watch*]), in which the disambiguating element is the verb adjacent to the ambiguous NP.

In summary, it can be said – and thus we return to the discussion of the first section – that the human language processing system follows economy criteria insofar as the parser always attempts to attain an interpretation of a sentence using the simplest structure or the simplest analysis, drawing upon one structural analysis only. Taking the data presented here into account, however, such a perspective is plausible only under the assumption that the basic architecture of the parser is ranked parallel. Only the assumption of such an architecture can explain the immediate assignment in ambiguous contexts, the strength with which different alternatives appear, the problem of increased processing costs in disambiguating regions, as well as the postponement of an analysis. While it appears clear that structural or syntactic factors are initially responsible for the hierarchical ordering of the alternatives stemming from an ambiguity, nonstructural parameters appear to influence this process at some point.

References

Altmann G and Steedman M (1988) Interaction with context during human sentence processing. *Cognition* **30**: 191–238.

Bach E, Brown CM and Marslen-Wilson WD (1986) Crossed and nested dependencies in German and Dutch: a psycholinguistic study. *Language and Cognitive Processes* **1**: 249–262.

Crocker MW (1994) On the nature of the principle-based sentence processor. In: Frazier L and Rayner K (eds) *Perspectives on Sentence Processing*, pp. 245–266. Hillsdale, NJ: Lawrence Erlbaum.

Fiebach CJ, Schlesewsky M and Friederici AD (in press) Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions. *Journal of Memory and Language*.

Fodor JD and Inoue A (1994) The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research* **25**(5): 407–434.

Frazier L (1987) Sentence processing: a tutorial review. In: Coltheart M (ed.) *Attention and Performance*, vol. 12, pp. 559–586. Hillsdale, NJ: Lawrence Erlbaum.

Frazier L and Clifton C (1996) *Construal*. Cambridge, MA: MIT Press.

Friederici AD (1998) Diagnosis and reanalysis: two processing aspects the brain may differentiate. In: Fodor J and Ferreira F (eds) *Reanalysis in Sentence Processing*, pp. 177–200. Dordrecht, Netherlands: Kluwer.

Friederici AD (1999) The neurobiology of language comprehension. In: Friederici AD (ed.) *Language Comprehension: A Biological Perspective*, 2nd edn, pp. 265–304. Berlin/Heidelberg/ New York: Springer.

Friederici AD (2002) Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences* **6**: 78–84.

Friederici AD, Mecklinger A, Spencer KM, Steinhauer K and Donchin E (2001) Syntactic parsing preferences and their on-line revisions: a spatio-temporal analysis of event-related brain potentials. *Cognitive Brain Research* **11**: 305–323.

Gibson E (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**: 1–76.

Gunter TC, Friederici AD and Schriefers H (2000) Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience* **12**: 556–568.

Hickok G (1993) Parallel parsing: evidence from reactivation in garden-path sentences. *Journal of Psycholinguistic Research* **22**: 239–250.

Just MA and Carpenter PA (1992) A capacity theory of comprehension: individual differences in working memory. *Psychological Review* **99**: 122–149.

Konieczny L and Hemforth B (2000) Modifier attachment in German: relative clauses and prepositional phrases. In: Kennedy A, Radach R, Heller D and Pynte J (eds) *Reading as a Perceptual Process*, pp. 517–527. Amsterdam, Netherlands: Elsevier.

MacDonald MC (1994) Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* **9**: 157–202.

Marslen-Wilson WD and Tyler LK (1980). The temporal structure of spoken language understanding. *Cognition* **8**: 1–71.

McClelland JL, St John M and Taraban R (1989) Sentence comprehension: a parallel distributed processing approach. *Language and Cognitive Processes* **4**: 287–335.

McElree B and Griffith T (1995) Syntactic and thematic processing in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition* **21**: 134–157.

Muckel S and Pechmann T (2000) Does the parser search for traces? Paper presented at the 13th Annual CUNY Conference on Human Sentence Processing, La Jolla, California, March 30–April 1.

Nicol JL, Fodor JD and Swinney D (1994) Using cross-modal lexical decision tasks to investigate sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**: 1229–1238.

Further Reading

Friederici AD (ed.) (1999) *Language Comprehension: A Biological Perspective, 2nd edn*. Berlin/Heidelberg/New York: Springer.

Gleitman LR and Liberman M (eds) (1995) *Language. An Invitation to Cognitive Science*, 2nd edn. Cambridge, MA: MIT Press.

Gorrell P (1995) *Syntax and Parsing*. Cambridge, UK: Cambridge University Press.

Altmann GTM (ed.) (1990) *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.

Sign Language

Introductory article

Diane C Lillo-Martin, University of Connecticut, Storrs, Connecticut, USA

CONTENTS

Practitioners of sign languages use movements of their hands, face, and body for communicating what language speakers convey using their mouth, tongue, and vocal chords. The natural sign languages of Deaf communities have the properties of human language, including grammatical structure at the phonological, morphological, and syntactic levels.

INTRODUCTION

Communities of Deaf people all over the world have developed sign languages, conveying by movements of the hands, face, and body what spoken languages convey using the mouth, tongue, and vocal chords. These sign languages are distinct from the spoken languages used around them, and from each other, although they share properties with both. Research over the past 40 years has established that sign languages satisfy the criteria for natural human languages, both in form and in function. Researchers have worked to determine the grammatical properties of sign languages, and how they are acquired and represented in the brain.

By far the greatest attention has been given to American Sign Language (ASL), the sign language used in the United States and parts of Canada, and this language will be the focus here. However, the sign languages of Europe have been studied in some detail, as well as Israeli Sign Language, Japanese Sign Language, Brazilian Sign Language, and others. Much further research is needed on all sign languages, but current studies have begun to reveal similarities and differences between the world's sign languages, an area of especial interest for its potential to reveal those aspects of language which are modality-dependent, those which are universal, and those which vary by language regardless of modality.

One of the fundamental goals of the study of language as a cognitive phenomenon is the determination of its essential properties. The search for *universal grammar* involves going beyond the surface differences between languages to find their deeper similarities of structure and organization. The study of sign language is an important part of cognitive science because it enables linguists to further separate out effects of the modality of expression from true universals of language. Those properties of spoken and signed languages which hold are the true linguistic universals, while those properties specific to spoken or signed languages are probably due to particular aspects of the modalities.

PHONOLOGICAL AND MORPHOLOGICAL FEATURES OF SIGN LANGUAGE

Sign Language Phonology

In 1965, William Stokoe published his *Dictionary of American Sign Language on Linguistic Principles*. For the first time, the signs of ASL were presented in terms of their formational components – handshape, location, and movement. Prior to the publication of Stokoe's dictionary, signs were thought of holistically, like mimes or gestures. What Stokoe showed was that signs have parts: that signs can be decomposed into smaller, meaningless units which combine in different ways to produce different words – just as words do in spoken languages. Stokoe began to show that there is a phonology of sign language. Figure 1 shows pairs of ASL signs which differ only in their location, or in their handshape.

In Stokoe's dictionary, the parts of a sign were represented in the order location–handshape–movement, but this sequence does not represent the linear order of a sign. Instead, Stokoe assumed that these components were produced simultaneously – unlike the component pieces of a spoken word which are produced sequentially. This

Figure 1. Pairs of ASL signs differing only in (top) location; (bottom) handshape.

feature was considered a major difference between spoken and signed words. Other researchers highlighted this modality effect, and assumed that theories of the phonology of sign languages would have to accommodate this difference.

However, in the 1980s, researchers brought forth evidence that the phonology of ASL crucially involves sequential segmentation, just as in spoken languages. Actually, spoken language researchers were beginning to notice important effects of simultaneity in spoken language phonology, captured in theories of autosegmental phonology and feature geometry. These theories provided the tools for capturing the nature of both signed and spoken languages as having both simultaneous and sequential aspects.

Thus were born models of sign language phonology which separated out the beginning, middle, and end of a sign. For example, Sandler's theory identifies for many signs the beginning location as the initial segment, the movement as the middle segment, and the ending location as the final segment. The hand configuration, which may remain constant or make very limited changes across the duration of a sign, is represented on a separate tier, and autosegmentally spreads across the segments. A simplified representation of a typical monomorphemic sign is provided in Table 1.

Table 1. Schematic diagrams of phonological structure

Single sign	*Combination of two signs to form a compound*			
HC	HC_1	+	HC_2	→ HC_2
L M L	L_1 M L_2		L_3 M L_4	L_2 M L_4

The representation of signs as in Table 1 allows the sign phonologist to explain how signs change when combined into compound words. For example, when the compound sign FAINT is formed by combining the signs for MIND and DROP, the compound is reduced practically to the form of a single sign. The compound begins where the sign MIND ends: at the forehead. It moves to the location where the sign DROP ends using one movement, thus taking the overall shape of a single

sign. The handshape used in MIND is lost completely; only the handshape of DROP is used. This combination of the feature specifications for two different signs merging into one sign involves assimilation, a phonological process whereby one part of a word becomes more similar to an adjacent word.

Rules like assimilation are an important part of the arsenal of phonologists attempting to explain the patterning of units within a word. Sign language phonologists have worked most extensively on developing the most parsimonious representations of signs, refining the hierarchies used to describe handshapes, or the models which best allow for the depiction of how signs change in different contexts. (*See* **Phonology**; **Phonology and Phonetics, Acquisition of**)

Sign Language Morphology

In some ways, ASL has more robust and complex morphology than English does. It is like Swahili in that verbs may show agreement with both their subjects and their objects. It is like Diegueño in that other verbs, expressing location and movement, may combine information about the type of entity with detailed spatial information. It is like Greek in that verbs may be marked to show the intensity and duration of the actions they depict.

As with the earlier studies of ASL phonology, researchers have been impressed by the 'simultaneity' of ASL morphology. Most morphological processes in ASL do not involve prefixation or suffixation, the most common forms of morphological processes in spoken languages. Rather, in many cases the movement of a sign may change, or the handshape or location may change, to indicate the addition of morphemes. Because of the way morphemes combine in sign languages, many morphemes can be expressed using a single syllable (a sequence of a location, one movement, and another location). While such nonconcatenative processes are present, they are relatively rare in spoken languages; however, they are the norm in sign languages.

There are two areas of ASL morphology which have received much attention, and they will be summarized here. They are the use of classifiers with predicates of motion and location, and the marking of subject and object agreement on verbs.

Classifiers in sign languages are handshapes which represent the entities being described by predicates indicating existence, location, or movement. The classifier may indicate a semantic group, such as land and air vehicles, or two-legged beings; or it may indicate the size and shape of the entity, such as thin, round, and flat, or deep and boxy. The movement of the sign represents the core meaning, such as travel in a particular direction. Modifications of the movement may indicate a bumpy terrain, or slow, deliberate motion. Using classifiers, sign language can lay out a detailed specification of the spatial arrangement in a story or description. It is difficult to find English translations for classifier sequences, and skillful manipulation of classifiers is considered one of the hallmarks of excellent sign language story-telling.

The area of space in front of a signer is used by verb agreement as well as classifier verbs. To mark agreement, a verb (usually) moves from the location of the subject to the location of the object. For example, suppose the signer wishes to say 'I teach her', referring to a woman on her right. The sign for TEACH will move from a location close to the signer's body, towards a location close to the woman. If the signer wished to say 'She teaches me', the sign for TEACH would move in the opposite direction: from the woman to the signer. The same process is used with abstract locations representing referents not present in the situation. In this way, the sign marks agreement with the subject and object by indicating the referent using their location. An illustration of this agreement system is given in Figure 2.

One interesting fact about agreement in sign languages is that not every verb marks agreement. Classifier verbs and some others indicate spatial locations but not necessarily subject/object agreement. A third class of verbs fails completely to mark agreement. 'Plain' verbs are produced the same way regardless of the locations of their subjects or objects. Both semantic and phonological factors help to determine which verbs mark agreement. For example, verbs which take inanimate objects do not mark agreement. Likewise, verbs which require continued contact with the body do not indicate agreement.

SYNTACTIC FEATURES OF SIGN LANGUAGE

Notation

SIGN Signs are indicated using upper-case English glosses with approximately the same meanings.

$_{a}$**SIGN**$_{b}$ Subscripts before or after a gloss indicate the use of morphological spatial locations.

nm
SIGN Nonmanual markers are indicated using a

I-TEACH-HER

SHE-TEACHES-ME

Figure 2. Verb agreement in ASL. The photos represent the beginning and ending locations of the movement, which is repeated.

line above the glosses for the signs with which they co-occur.

American Sign Language

One of the most interesting aspects of the syntax of ASL is the large variation in word order found. This is a dimension on which languages may differ greatly; some languages have a relatively strict order of elements, while others permit considerable flexibility. English is among the stricter languages; most sentences comply with the 'basic' word order subject–verb–object, although there are processes for changing word order, such as topicalization and passivization. Japanese allows variability for placement of the noun phrases of a sentence, although the verb must always be final. Serbo-Croatian has a great deal of variability: the subject, verb, and object may appear in any ordering with respect to each other. However, the various orders are used in different contexts – they have different pragmatic effects, such as focusing or topicalizing elements.

It is generally agreed that the basic word order of ASL is subject–verb–object (SVO). However, this order may vary, but in different kinds of context. Like Serbo-Croatian, much of the word order variation found in ASL is for different discourse or pragmatic effects. The topic of a discourse is often expressed first. An element which is focused is often expressed last. To complicate matters, the subject or the object in an ASL sentence may frequently be unexpressed. When there is sufficient information in the context for the referent to be understood (such as from verb agreement), it may be left out. So, all of the examples given in Table 2 (and more) would be acceptable ways to say 'the boy ate an ice cream', in different contexts.

The examples given in Table 2 have some indications for nonmanual markers (the line above certain signs), but not all relevant nonmanual markers have been indicated. Although most signs are

Table 2. Word order variation in ASL

Example sentence	*Order of elements*
BOY EAT ICE-CREAM t	S V O
ICE-CREAM, BOY EAT t t hn	O S V
BOY, ICE-CREAM, EAT t hn	S O V
ICE-CREAM, EAT, IX(boy)	O V S
BOY EAT hn	S V
EAT, IX(boy)	V S
EAT ICE-CREAM t	V O
ICE-CREAM, EAT	O V

produced by the hands, movements of the body and head and specific expressions of the face are also used in the grammar of sign languages. These nonmanual markers are used to mark questions, conditionals, and relative clauses, among other things. In the examples of Table 2, the 't' nonmanual indicates the topic of a sentence (signaled mainly by raised eyebrows and a prosodic break), while 'hn' indicates a head nod, which often accompanies material at the end of a sentence.

Nonmanual markers can be compared to intonation in many ways. Like intonation, the nonmanual marker of a particular construction (such as a question) 'spreads' across a certain domain. The marker for a yes/no question (illustrated in Figure 3), which consists of raised brows and a forward head tilt, co-occurs with the whole question, as the examples in Table 3 indicate.

The examples we have seen so far are all simple sentences with only one clause. Of course, ASL permits syntactic recursion, a process which

Figure 3. Nonmanual marker indicating a yes/no question.

Table 3. Spread of the yes/no question nonmanual marker

	ynq STUDENT ALL PASS TEST 'Did all the students pass the test?'
Unacceptable	ynq *STUDENT ALL PASS TEST

embeds one phrase inside another. An example of a sentence with three embeddings (from the work of Carol Padden) is given below. (The notation IX refers to an indexical pointing sign – a pronoun.)

IX(me) $_1$TELL$_a$ WOMAN $_a$TELL$_b$ MAN
$_b$PERSUADE$_2$ GO THERE PARTY IX(me)
'I told the woman to tell the man to
persuade you to go to the party.'

Cross-linguistic comparison

It seems that sign languages around the world share many grammatical properties, such as a preponderance of nonconcatenative morphology, classifiers, verb agreement, and nonmanual grammatical markers. However, there are some interesting ways in which sign languages differ, and others will probably emerge from future research.

One way in which sign languages differ from each other is in basic word order. While ASL and Brazilian Sign Language have SVO as their basic order, German Sign Language and Japanese Sign Language have the basic order SOV. Moreover, Brazilian Sign Language and German Sign Language (and others) have an element often glossed AUX, which marks agreement in sentences with plain verbs. ASL has no such sign. The presence or absence of AUX in a sign language may be correlated with other syntactic properties. This is one of the topics sign language syntacticians are now researching.

THE ACQUISITION OF SIGN LANGUAGE

Only 5–10 percent of Deaf children are born in Deaf, signing families. These children, like hearing children, receive linguistic input from birth, by watching their parents signing to and around them. Since this population is most comparable to the majority of children studied in spoken language acquisition research, it has received the bulk of the attention of sign language research. We will focus on this group, leaving aside the issue of sign

language acquisition by deaf children born in hearing families.

In general terms, Deaf children acquire ASL in much the same way that hearing children acquire their native languages. They acquire the language by exposure to it; they are not taught it by parents or educators. They do so along a timeline much like that for spoken languages, hitting major milestones in the same range of ages as hearing children do. This should be emphasized since, given the difference in modality, things could have turned out quite differently. The similarities are striking evidence of the biological nature of language.

In fact, there seems to be one difference, which is extremely relevant to the study of modality effects on language. For some years, researchers claimed that the first words produced by signing children (appearing at an average age of about 8 months) came some months before the first spoken words (an average age of about 10–11 months). This difference at such an early stage could indicate that the onset of language is not biologically determined. However, closer examination determined that both children exposed to sign and children exposed to speech used similar communicative gestures at an early age. When the same criteria for distinguishing a word from a gesture were used, the onset of words in speech and sign became much closer. Also, the first truly symbolic words in signed and spoken language both come at around 12 months. If there is a difference between signed and spoken language development in the appearance of first words, it seems this difference is actually rather small, and it can easily be explained by considering the different physical development of the articulators of spoken and signed languages.

It is also clear that beyond the first words, other linguistic milestones emerge at comparable ages. For example, in both sign and speech the first word combinations into simple sentences come at around 18 months. Early development of word order, verb agreement, and more complex constructions seems to be entirely comparable to that of spoken languages.

Like speaking children, signing children make relatively few errors in language development. One common type of error involves the substitution of one handshape or location for another (as illustrated in Figure 4). Two-year-old children tend to use short sentences, but they follow the adult language in the way that they are formed, merely missing some elements of elaboration. The core properties of the language are acquired by children before they are 5 years of age.

Figure 4. Child's production of the sign CRACKER uses the wrong location; the adult sign is made at the elbow.

Unlike other aspects of the grammar, some nonmanual markers in ASL seem to be acquired relatively late. Nonmanual markers like that for yes/no questions, where no manual sign also indicates the grammatical function, are acquired early, but markers like the *wh*-question marker are acquired later – and even when children begin to use the *wh*-question nonmanual they may not do so consistently. However, when they do use such markers they rarely, if ever, make mistakes in its spreading. (*See* **Syntax**; **Language Acquisition**; **Pragmatics, Formal**)

APHASIA IN SIGN LANGUAGE

It has long been observed that damage to specific areas of the left cerebral hemisphere results in impairments of language; while damage to the right hemisphere may result in impairments of spatial cognition. Since sign languages use spatial information so intricately, an obvious question is whether language deficits following brain damage would be as strongly left lateralized for sign language as for spoken language.

A series of studies, mostly conducted at the Salk Institute in La Jolla, California, indicates that sign

language is left lateralized, just as spoken language is, and that damage to the right hemisphere does not significantly impair even the spatial aspects of sign language. Deaf (and occasionally hearing) ASL signers, who experience brain damage in the temporal and parietal regions of the left hemisphere, may experience significant disruptions to language. In fact, different types of disruptions are caused by damage to different areas, in ways very similar to the differences between Broca's and Wernicke's aphasias for spoken languages. The signing of some patients may be slow and effortful, mostly devoid of functional elements; while others may sign fluently but nonsensically, with frequent paraphasias.

Right hemisphere damaged signers may experience severe impairments in spatial cognition, showing classic symptoms such as left neglect. However, their signing is not impaired overall, and even their use of signing space is much like that of age-matched controls. It seems clear that the brain's lateralization for language is based on the cognitive function rather than the modality of expression.

However, this is not to say that there is no evidence for differences in the neural control of signed and spoken languages. Some recent studies using new technologies for studying the neural processing of intact subjects have indicated that, while the left hemisphere's involvement is definite, there may be more right hemisphere involvement in the processing of sign language as compared to spoken languages. Some of these studies have included as one subject group hearing individuals who, growing up in a Deaf household, are native signers. The inclusion of this group allows the researchers to determine whether potential differences between the neural control of signed and spoken languages are due to the nature of the language or to the user's sensory abilities. While some differences are clearly due to sensory deprivation (Deaf signers may use areas of the brain which normally serve audition for visual–spatial processing), others seem to be related to early language use: those who grow up signing use more of the right hemisphere to process sign language than those who grow up speaking do for spoken language.

These recent findings are still preliminary, and much more work remains to be done. However, it is important to search for the ways in which sign languages may be controlled differently from spoken languages, and to inquire into the causes of such differences. Eventually, a full understanding of the nature of signed and spoken languages, including structure, acquisition, and neural control, will help researchers to construct the optimal model of the language faculty. (*See* **Aphasia**)

IMPLICATIONS OF SIGN LANGUAGE FOR UNIVERSAL GRAMMAR

To the extent that sign languages adhere to the principles of universal grammar (UG), they provide support for the hypothesis that a set of properties hold for language qua language, with limited variation possible in the abstract parameters which define linguistic options. Presumably, the principles and parameters say nothing about modality, so sign languages should not differ radically from spoken languages. In particular, if the language faculty is autonomous, then it might be expected that modality effects in sign language would be restricted to the interface between language and input/output. That is, sign languages, having visual input, would have different phonological features, and potentially different types of phonological realizations from spoken languages. However, the languages would not differ in areas of syntax, whose computations would be blind to details of the surface form. Such a model of the architecture of the language faculty is presented in Figure 5.

Although much more research needs to be done, many sign linguists believe that ASL and other sign languages do not violate the principles of UG. The grammatical properties shown by sign languages

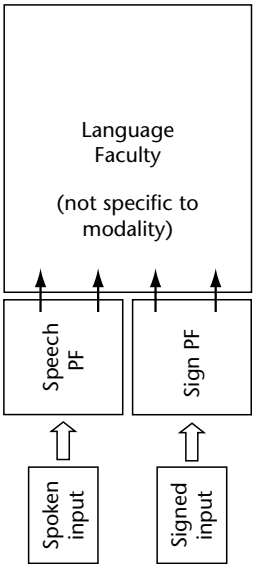


Figure 5. Modular representation of the language faculty. Differences between signed and spoken languages are restricted to the interface. PF, phonological form.

are remarkably similar to those found in the range of spoken languages studied. Even in the area of phonology there are striking similarities between sign and spoken language, although of course there are important differences as well.

There are, however, some possible challenges for the conclusion that sign languages fall perfectly within the range of variation permitted by UG. This is not to say that the hypothesis that sign languages are governed by the same UG as spoken languages is invalid, but that certain potential modality effects need to be more deeply considered and fully explained.

One such difference concerns the use of space. As described above, sign languages use spatial distinctions to represent abstract referents or spatial locations. As one consequence, pronominal signs must pick out particular referents unambiguously. Pronouns may be directed toward any location with which a referent has been associated – thus it is not possible to list all the possible surface pronoun forms of a sign language. Does this mean the lexicon of sign languages must be indefinitely long?

One resolution of this problem is to consider all the various pronoun signs as realizations of two lexical forms: the first person form, and the non-first person form. Many syntacticians use referential indices for pronouns to keep track of reference, co-reference, and disjoint reference. Under the present hypothesis, the difference between sign and spoken languages is that the referential indices for pronouns are overtly realized in sign. That is, in a particular context, the 'she$_1$' which refers to Sally is produced differently from the 'she$_2$' which refers to Jane, in signed languages but not in spoken languages. This kind of solution places the locus of the difference between signed and spoken languages within the domain of what is realized phonologically, consistent with the autonomy hypothesis.

A second type of problem comes from considering the nature of sign languages as a group. Although sign languages do not seem to go outside the range of variation permitted by UG, it has been noticed recently by sign researchers that they display characteristic properties as of a language family – and this fact has not been captured by linguistic theory. Since the properties shared by all sign languages, such as the referential use of spatial locations in verb agreement and pronouns, are clearly related to the sign modality, it is important for linguistic theory to address these similarities and generalizations.

Thus, linguists in general are interested in the study of sign languages. Research on sign language is vital to a complete understanding of the nature of universal grammar. Sign languages allow linguists to test their proposed principles and parameters, and they provide the advantages both of understudied languages, and more importantly, of testing the proposals for their application across modalities. The study of sign languages reveals true linguistic universals.

Further Reading

Emmorey K (2001) *Language, Cognition, and the Brain: Insights from Sign Language Research*. Mahwah, NJ: Lawrence Erlbaum.

Klima ES and Bellugi U (1979) *The Signs of Language*. Cambridge, MA: Harvard University Press.

Marschark M, Siple P, Lillo-Martin D, Campbell R and Everhart V (1997) *Relations of Language and Thought: The View from Sign Language and Deaf Children*. New York, NY: Oxford University Press.

Newport E and Meier R (1985) The acquisition of American Sign Language. In: Slobin DI (ed.) *The Acquisition of American Sign Language*, pp. 881–938. Hillsdale, NJ: Lawrence Erlbaum.

Poizner H, Klima ES and Bellugi U (1987) *What the Hands Reveal about the Brain*. Cambridge, MA: MIT Press.

Sandler W and Lillo-Martin D (2001) Natural Sign Languages. In: Aronoff M and Rees-Miller J (eds) *The Handbook of Linguistics*, pp. 533–562. Malden, MA: Blackwell.

Sandler W and Lillo-Martin D (in press) *Sign Language and Linguistic Universals*. Cambridge, UK: Cambridge University Press.

Stokoe WC, Casterline D and Croneberg C (1965) *A Dictionary of American Sign Language on Linguistic Principles*. Washington, DC: Gallaudet College Press. [Reprinted in 1976 by Linstok Press.]

Valli C and Lucas C (1992) *Linguistics of American Sign Language*. Washington, DC: Gallaudet University Press.

Wilbur RB (1987) *American Sign Language: Linguistic and Applied Dimensions*. Boston, MA: College-Hill Press.

Spatial Language

Intermediate article

Stephen C Levinson, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

CONTENTS

The study of spatial language is concerned with the systematic properties of spatial descriptions in natural languages, that is, the description of where things are, or the description of moving bodies.

SPATIAL COGNITION AND LANGUAGE

There are two reasons why the study of spatial description in language might be of special interest to cognitive science. First, space is a central cognitive domain for any roving animal, and human thinking is deeply spatial, reflecting no doubt this ancient phylogeny. The role of gestures, figures and diagrams, geometry, and maps in our thinking all attest to this fundamental role that spatial thinking plays in our cognition. In linguistics, this idea that spatial notions form the foundation for much of our nonspatial concepts is known as 'localism', much evident in cognitive linguistics. Second, language seems to offer a window on the inner world of spatial concepts – in the case of the honey bee for example, we know more about spatial cognition from the communication system than from direct observation of other behavior. However, optimism that we might find a uniform core of human spatial communication may be misplaced: due to the complexity of human cognition, the conceptual systems that underlie language display little of the metric precision of our perceptual systems and are quite variable, being deeply interlocked with cultural concepts (see Bloom *et al*., 1996). They must therefore be studied in their own right, even though it turns out that there are close interrelations between nonlinguistic spatial cognition and linguistic concepts. (*See* **Spatial Representation and Reasoning**; **Spatial Cognition, Psychology of**; **Spatial Disorders**; **Cognitive Linguistics**)

In a tradition that goes back to Kant and beyond, an orthodoxy has grown up that holds that naive human spatial cognition, as reflected in language, is universally egocentric and anthropomorphic in character, and characterized by such universal primitives as ON, IN, AT, and so forth (see e.g. Miller and Johnson-Laird, 1976; Lyons, 1977; Herskovits, 1986). Such spatial concepts are often put forward as good candidates for innate concepts, reflected universally in language (see e.g. Landau and Jackendoff, 1993). Readers are left with the impression that such notions as 'to the left of', 'in front of', 'on', etc., are universally expressed, and moreover that they are coded in limited parts of speech, especially adpositions (prepositions and postpositions). This impression is deeply misleading – recent cross-linguistic work shows that there is no such uniformity in either the semantics or the formal expression of spatial distinctions across languages. This article details these more recent findings, and their consequences for the language–cognition interface.

THE SEMANTICS OF SPACE

Fundamental Concepts

The details of the semantics of spatial description are quite complex and vary considerably across languages, but the general outlines tend to follow rather simple functional principles, as follows (see Levinson, 1996 for details). The Newtonian concept of space as an infinite, abstract three-dimensional envelope plays relatively little role in naive spatial conception – rather, the Leibnizian view of space as a system of relations between things is predominant (some exceptions are noted below). In particular, one object, the *figure* (or theme or trajector, in alternative terminologies), is located by reference to another, the *ground* (or the landmark or relatum). When figure F and ground G are contiguous, it is often sufficient to say in effect 'F is at G', where 'at' glosses some kind of contiguity relation. However, languages may subdivide contiguity into different kinds of relation, such as superadjacency,

subadjacency, containment, and so forth (as in 'The ball is on the table/under the cloth/in the bowl'). This kind of relation is called *topological*, following Piaget. Where G is relatively large compared to F, it may be helpful to subdivide G into parts, and say in effect 'F is at the X-part of G' (as in 'The book is in the back of the car'). Place-names or *toponyms* may be thought about as an elaborate subdivision of a territory for just this purpose. (*See* **Piagetian Theory, Development of Conceptual Structure**)

When F and G are displaced in space, a more complex solution to spatial location is required: we now need an indication of the direction from landmark G in which to search for F. To specify a direction, we need an angular specification, and natural languages provide this in *polar* (rather than Cartesian) *coordinates* mostly based on G. Such coordinate systems are called *frames of reference* in the psychological literature, and it turns out that languages use just three main types. One type uses the system for partitioning objects into parts already mentioned, and projects an axis from the centre of G through a named part to determine an angle or direction, specifying in effect 'F is to the X-side of G', as in 'The ball is at the rear of the truck'. This kind of coordinate system is called the *intrinsic* frame of reference because it relies on reference to the inherent or intrinsic parts of objects (although this terminology is misleading – different languages have quite different ways of assigning parts to objects). Another type uses the bodily axes of the viewer, front and back, left and right, and maps this coordinate system onto the landmark object G, so that we can talk, for example, of F as to the left of G (the mappings are subject to different transformations in different languages, as will be explained). This kind of system is called the *relative* frame of reference (it is often also called the *deictic* frame of reference, but this is misleading, as the viewer whose bodily coordinates are used need not be the speaker, and all frames of reference can have the speaker as the ground object). The third and final kind of system uses abstract, antecedently fixed bearings, a bit like our North or East. Such systems are called *absolute* or 'geocentric' or 'environmental' frames of reference, and all languages use such a system in the vertical dimension, as in 'The lamp is above the table'. But it has only recently been documented that some languages use such systems on the horizontal plane to the exclusion of the relative frame of reference (see Levinson, 1996; Pederson *et al*., 1998).

The properties of motion description are somewhat different. Motion can, of course, be described as located in a place, in which case the systems already mentioned are relevant, but normally the interest is in the direction of motion, or at least in where it is originating or terminating. Again we talk about the figure (the object in motion), but we may need to distinguish multiple grounds, especially the *source* and *goal* of the motion. The specification of source or goal alone does not give us an angle or vector of motion (a fully specified direction) – it only tells us that the motion progressively increased or decreased the figure's distance from the landmark or ground object (a radial trajectory towards or away). Source and goal together do fix a vector, as in 'He went from London to Birmingham', but many languages do not permit source and goal specification in one clause. Languages which use the absolute frame of reference often use fixed bearings to determine a vector, roughly as in 'He left the village northwards'.

One omnirelevant location is the place of speaking, the *deictic centre*. This location may be the ground in any location or motion description, and in principle it is not different from any other landmark, but in practice languages tend to code it specially, for example in demonstratives ('this' versus 'that'), deictic adverbs ('here versus there'), and deictic verbs of motion ('come' versus 'go', 'bring' versus 'take'). (*See* **Indexicals and Demonstratives**)

Languages tend to treat space as a single large semantic field; for example, most languages have a single shared root meaning 'Where?' used across topology, frames of reference and motion description. But this large field tends to be systematically subdivided, so that motion versus location, and within location topology versus frames of reference, are distinguished as important subdomains organized distinctively in both form and meaning, as in Figure 1, and linguistic descriptions should treat these subdomains individually.

Cross-linguistic Variation and Underlying Universals in Spatial Language

Recent investigations have shown that there is much more cross-linguistic variation in spatial language than had been supposed. The semantic parameters involved can be quite various and differently interconnected. Consider first the distinction between location and motion – Talmy (1983) argued that these are deeply interlocked, and indeed location can be thought of as a special case of motion (consider e.g. the parallelism of English 'He is out of the room' versus 'He went out of the room'). However, many languages use entirely

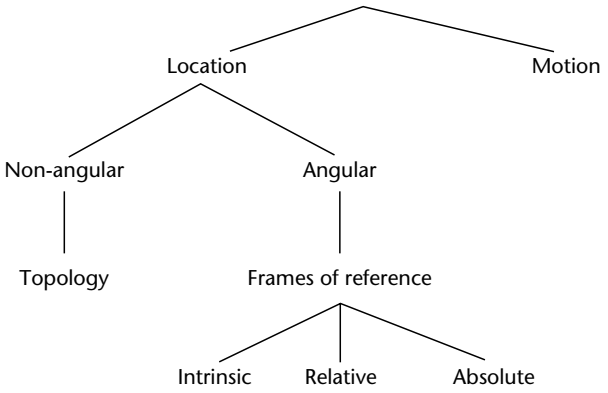


Figure 1. Some important subdomains of spatial language.

different semantic and formal resources in these two domains, so that no such parallelism can be presumed.

Consider the topological subdomain. As mentioned, many authors have presumed that notions like ON, IN, and AT would be universally coded in adpositions (capitals here denote supposed semantic primitives). But in fact there are no easy generalizations of this kind – for example, central Australian languages often conflate IN and UNDER, Japanese conflates ON and OVER, and they do so in different parts of speech, spatial nominals and postpositions respectively. However, variation is not random. Detailed comparison of many languages using standardized stimuli shows that the topological subdomain seems to form a single, universal multidimensional similarity space – the space is very variably subdivided by different languages, but in doing so they conflate into single lexical concepts only neighboring spatial relations. The semantic parameters in this space are at a much more abstract or componential level than ON; if that is conceived of as unattached contact with the vertical support provided by a horizontal surface, then it is notions like *contact*, *adhesion*, *superposition*, *horizontal supporting surface* that form the dimensions of the space. Moreover, these notions are encoded in various parts of speech in different languages: in grammatical case, adpositions, spatial nominals (special minor form classes of nouns), and frequently in locative verbs. Many languages have a small form class of locative verbs, the choice of which makes distinctions concerning the shape and orientation of the figure and its relation to the ground, and some languages have large sets of such verbs that code detailed figure–ground configurations, such as containment within a bowl-shaped ground, or wedged between two supports, as in Tzeltal.

Turning now to the ways of indicating angular direction between figure and ground, the intrinsic frame of reference is by far the most widespread of coordinate systems – indeed a case can be made for the intrinsic frame being universal, at least in vestigial form. However, the way in which objects are partitioned and assigned named sides or facets is very variable. Some languages (like Zapotec) use a fixed armature, assigning a 'top', 'bottom', and designated 'sides' (one can think of this as a superimposed box, as it were, where 'top' is always the vertically uppermost surface). Some languages (like Tzeltal) use an orientation-free system of internal object geometry: the longest axis has names associated with the end faces according to their shape, and similarly for the secondary axis, and so on. Such a system is intriguingly like the system David Marr (q.v.) imagined must be involved in visual object recognition. Yet other languages like English involve a complex mix of orientational and functional criteria, so that the 'top' of a bottle remains the 'top' whichever way up the bottle is (unlike in Zapotec), but the notion is tied to canonical orientation of the artifact. Miller and Johnson-Laird (1976, p. 403) sketch an algorithm for such part-assignment in English, involving such factors as the leading facet in typical motion (the 'front' of a truck), the facet with perceptual apparatus (the 'front' of a camera), the characteristic orientation of the user to the object (as in the 'front' of a desk). Despite the evident complexities, children learn the application of these terms in English earlier than other spatial relations.

The relative frame of reference involves, as mentioned, mapping the body axes, front/back, and left/right, onto the ground. Despite the fact that Kant and many other theorists have assumed the primacy of these axes in our naive spatial conception, many languages make no systematic use of them in this way – that is, they have no locutions of the kind 'The boy is behind the tree' or 'The boy isleft of the tree' (note that such a language may have a term for 'left hand' but makes no generalization of this concept to spatial regions). Many Australian and other languages around the world are of this kind. When languages do provide such locutions, the interpretations can be very various. Note that in English, the 'front' of the tree is the side facing the speaker or observer, thus rotating the speaker's front and back, while 'left' and 'right' are not rotated. One interpretation of this is that the observer's body axes are mapped under reflection onto the ground. In contrast, in

Hausa and many other languages, the observer's axes are translated without reflection or rotation, so 'left' and 'right' remain as in English, but 'front' and 'back' are reversed – 'The boy is behind the tree' now means the boy is between the speaker or observer and the tree (yet the term 'behind' applied to myself means just what it does in English). Hausa thus adopts the convention that the speaker and the tree are in single file, as it were, while English acts as if speaker and tree were confronting each other. Finally, in a few languages, full rotation of the axes occurs under mapping onto the ground: now the 'front' and 'back' of the tree are as in English, but 'left' and 'right' are reversed.

These relative systems may originate as generalizations of the intrinsic system onto ground objects such as trees which are not easily assigned 'fronts', 'backs', or other facets. This would account for the fact that relative systems always occur with associated intrinsic systems, and like intrinsic systems are largely coded in a series of nominal expressions. Thus in English 'The boy is in front of the tree' is unambiguously relative, but 'The boy is in front of the truck' is ambiguous between the intrinsic reading (at the front end of the truck) and the relative reading (the boy is between the observer and the truck). But many languages do not allow this ambiguity – if there is a possible intrinsic interpretation, then that pre-empts a relative one. The ambiguity in English has prompted theorists to assume that the terms themselves are not semantically specified one way or the other, and that frames of reference are psychological in character, not linguistic (Miller and Johnson-Laird, 1976, p. 404). But in many languages the intrinsic and relative systems are clearly distinct in construction, and in English the distinctions can be made linguistically (as in 'at the truck's front').

The absolute frame of reference used on the horizontal plane will be unfamiliar to most readers except in discourse about geography, where *North* may really be thought about as 'up' on a map, that is, intrinsically. However, many languages use an absolute frame of reference for nearly all spatial discriminations for objects separated on the horizontal plane, speaking thus of 'the northern knife' or 'your western knee' and so forth. There are again many different types. Some languages, like most Australian Aboriginal ones, have fully abstract cardinal direction systems, like our North, South, East, and West, except that they may be skewed in different directions and are likely to have precise quadrants or arcs associated with each. Orthogonal axes are normal, but not invariable, and quadrants of application may be equal (of 90 degrees) or not. Another common kind of system (found e.g. in Nepal and MesoAmerica) uses major inclines in local geography, with an uphill versus downhill major axis, and an 'across' minor axis (the directions further specified by landmarks). A third common type of system (found e.g. in Arnhem Land and Alaska) uses the major axis of river drainage to provide 'upstream'/'downstream' and 'across' axes. Prevailing winds are also a common source of inspiration, as in Eskimo wind direction systems, which in naming up to sixteen directions around the compass card allow precise subdivisions down to 22.5 degrees. It should be stressed that although these latter systems may seem hooked to local environmental conditions, these systems are mostly abstracted off this ecological background, and have become fully abstract fixed bearings, which do not vary when the landscape varies or when used outside traditional territories. Such systems often have considerable linguistic importance, forming a systematic underlying set of oppositions, a grammatical category, which shows up in different lexical and morphological sets – for example, such languages are likely to have, in addition to nouns denoting the directions, motion verbs meaning 'to go north', etc., and demonstratives meaning 'that northern one', etc. Incidentally, absolute systems of spatial description are the only naive spatial concepts that seem to surpass in abstraction the Leibnizian view of space as consisting only of relations between things. A description such as 'the southern edge' or 'going east' does not rely on a figure–ground relation – instead of the relation to the ground an abstract spatial vector is specified in a Newtonian space. In addition to this abstract quality, these systems are of considerable interest to the cognitive scientist because they require speakers of such languages to constantly and correctly reckon their orientation with respect to these fixed bearings, a point that we will return to.

Clearly not all languages use all three of these frames of reference (intrinsic, relative, and absolute), but some do. Some languages use the intrinsic system alone, and others use the absolute system alone, or with only traces of the intrinsic system. The only constraint appears to be that the relative frame is dependent on the intrinsic one. This relative freedom of occurrence is intriguing, since the different frames have rather different spatial and logical properties. For example, unlike the intrinsic frame, both the relative frame and the absolute frame map axes from a larger space onto the figure–ground relation – hence when the

figure–ground configuration is rotated, the intrinsic description may remain constant, but not the relative or absolute descriptions. Relative and absolute frames thus support logical transitivity and converseness (e.g. if A is north of B, and B is north of C, A is north of C), unlike the intrinsic frame. However, if one rotates the viewpoint (e.g. by walking around to the other side of the array), the figure–ground relation changes in a relative description (what was to the left becomes to the right), but not in an absolute description. Thus it is only absolute descriptions that sustain full logical inferences under different viewpoints – they are clearly the logically superior systems, but require a significant cognitive overhead, namely constant mental orientation.

In summary, then, the semantic distinctions made in spatial descriptions vary quite widely across the world's languages, and there are no high-level concepts of the order of IN or FRONT OF or LEFT OF that turn up universally in languages. Nevertheless, there seem to be underlying constraints on the semantic spaces involved in each subdomain, such that, for example, the topological space seems universally specified as a single similarity space, languages can draw from only three frames of reference, each based on polar coordinates, and so on. Another area of surprising diversity is the way in which such distinctions are coded – there are no universal tendencies for spatial relations to be coded in just one or two parts of speech, but rather the tendency is for spatial information to be distributed through the clause, in nominals, case, adpositions, and spatial predicates.

THE COGNITIVE CONSEQUENCES OF LINGUISTIC DIVERSITY

Many species (ants, bees, fish, birds, and bats) have quite extraordinary spatial and navigational skills, often based on such exotic senses as polarized light detectors, echo-location, and magnetoreceptors. All the evidence points to native human spatial perception as being indifferent to poor, as generally in the primate order. Western subjects, for example, displaced to an unfamiliar location, can rarely point to home-base, or even a recent waypoint, at much better than chance levels.

The diversity of linguistic systems for spatial location points to the special role that culture plays in human spatial thinking. The same point is suggested by the elaboration of different navigational traditions and by the technological development of a prosthetic sense of direction through maps, compasses, and satellite systems. The acquisition of spatial language by children also suggests that most spatial concepts in language are anything but 'natural', being learnt relatively late. The facts of acquisition, as far as we now know them, are as follows. Western children clearly learn topological spatial terms first, starting at about age two, proceed to intrinsic uses, and about the age of four have relative usages of 'front' and 'back'; but 'left' and 'right' terms lag far behind, with relative 'left' and 'right' often not being fully mastered before 11. This development is in line with predictions from Piagetian Theory, where topological concepts are held to be conceptually simpler than the Euclidean geometry underlying frames of reference. But children whose native languages have fundamentally different spatial systems from European languages do not seem to start from a common universal notional core, and then gradually diverge – rather they seem to adapt to the local system of categories from the beginning (Bowerman and Choi, 2000). Thus some children learning languages with intrinsic and absolute frames of reference, but no relative frame, do not seem to learn the intrinsic system before the absolute one, but rather at the same time or even partially in reverse order (Brown and Levinson, 2000). All of this suggests that in this domain the child must construct the relevant categories – they are not given by innate endowment as some authors have supposed. (*See* **Navigation**; **Piaget, Jean**)

Nevertheless, cultural and linguistic concepts of space can be shown to have profound cognitive consequences. For example, speakers of languages where the absolute frame of reference is predominant must run a constant background mental computation of absolute direction, reproducing in cultural 'software' what ants and bees do in 'hardware' through specializations for solar compass estimation. Pointing experiments show that such peoples, in contrast to speakers of relative languages, are capable of great accuracy in direction estimations without special attention during motion. Further, experiments on nonverbal memory and inference on these same subjects show that they code spatial scenes in memory in terms of fixed bearings, and not in terms of, for example, left and right, as Western subjects do. This can be shown using the distinct properties of relative and absolute frames of reference under rotation – for example, if subjects are shown an arrow facing left and south, are asked to memorize it, and are then rotated 180 degrees, and asked to pick the similar stimulus from a pair of arrows, one facing right and south, and one left and north, speakers of relative languages will pick the

north-facing arrow because it preserves leftness, but speakers of absolute languages will pick the right-facing arrow because it preserves southness. When embedded in a reasoning task, such manipulations are good tests for unreflective coding strategy, and the results show systematic effects of the semantics of the native language on the subjects' mental representations for general reasoning (see Levinson, 1996; in press). These are among the strongest Whorfian effects of language on cognition that have been demonstrated. (*See* **Whorf, Benjamin Lee**; **Linguistic Relativity**)

As mentioned at the beginning, spatial thinking seems to play a special role in human thinking. Further evidence of this comes from a pervasive phenomenon associated with speaking, namely gesture. Although the functions of gesture are not fully understood, it is clear that gesture co-occurs especially with talk about space, and although part of the motivation is communicational (especially in the case of pointing), part is conceptual – gesture seems to help the formulation of spatial messages (see McNeill, 2000). Interestingly, absolute speakers gesture while retaining the correct bearings of events, while relative speakers tend not to, indicating shared frames of reference in language and gesture. Finally, a matter of special interest is the encoding of spatial information in languages coded in a spatial medium, namely sign languages, where it turns out that there is no single, 'natural' solution to the depiction of space. (*See* **Sign Language**)

References

Bloom P, Peterson M, Nadel L and Garrett M (eds) (1996) *Language and Space*. Cambridge, MA: MIT Press.

Bowerman M and Choi S (2000) Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In: Bowerman M and Levinson SC (eds) *Language Acquisition and Conceptual Development*, pp. 475–511. Cambridge, UK: Cambridge University Press.

Brown P and Levinson SC (2000) Frames of spatial reference and their acquisition in Tenejapan Tzeltal. In: Nucci L, Saxe G and Turiel E (eds) *Culture, Thought, and Development*, pp. 167–197. Hillsdale, NJ: Lawrence Erlbaum Associates.

Herskovits A (1986) *Language and Spatial Cognition*. Cambridge, UK: Cambridge University Press.

Landau B and Jackendoff R (1993) 'What' and 'Where' in spatial language and spatial cognition. *Behavioral and Brain Sciences* **16**(2): 217–265.

Levinson SC (1996) Frames of reference and Molyneux's question: crosslinguistic evidence. In: Bloom P *et al*. (eds) *Language and Space*, pp. 109–170. Cambridge, MA: MIT Press.

Levinson SC (2000) *Presumptive Meanings*. Cambridge, MA: MIT Press.

Levinson SC (2002) *Space in Language and Cognition: Explorations in Linguistic Diversity*. Cambridge, UK: Cambridge University Press.

Lyons J (1977) *Semantics*. Cambridge, UK: Cambridge University Press.

McNeill D (ed.) (2000) *Language and Gesture*. Cambridge, UK: Cambridge University Press.

Miller G and Johnson-Laird P (1976) *Language and Perception*. Cambridge, UK: Cambridge University Press.

Pederson E, Danziger E, Wilkins D *et al*. (1998) Semantic typology and spatial conceptualization. *Language* **74**: 557–589.

Talmy L (1983) How language structures space. In: Pick H and Acredolo L (eds) *Spatial Orientation: Theory, Research and Application*, pp. 225–282. New York, NY: Plenum Press.

Further Reading

Carlson-Radvansky LA and Irwin D (1993) Frames of reference in vision and language: where is above? *Cognition* **46**: 223–244.

Eilan N, McCarthy R and Brewer B (eds) (1993) *Spatial Representation: Problems in Philosophy and Psychology*. Oxford, UK: Blackwell.

Haviland J and Levinson SC (eds) (1994) Spatial conceptualization in Mayan languages. *Special issue of Linguistics* **32**(4/5). Berlin: Walter de Gruyter.

Jammer M (1954) *Concepts of Space: The History of Theories of Space in Physics*. Cambridge, MA: Harvard University Press.

Jarvella R and Klein W (eds) (1982) *Speech, Place and Action*. New York: Wiley.

Johnston R and Slobin D (1978) The development of locative expressions in English, Italian, Serb-Croatian, and Turkish. *Journal of Child Language* **6**: 529–545.

Piaget J and Inhelder B (1956) *The child's conception of space*. London: Routledge.

Svorou S (1993) *The Grammar of Space*. Amsterdam: Benjamins.

Talmy L (2000) *Toward a Cognitive Semantics*, vols 1 and 2. Cambridge, MA: MIT Press.

Vandeloise C (1991) *Spatial Prepositions*. Chicago, IL: University of Chicago Press.

Speaker Recognition

Advanced article

Diana Van Lancker, New York University, New York, New York, USA
Jody Kreiman, University of California, Los Angeles, California, USA

CONTENTS

Voices, like faces, communicate personal and emotional information, and figure importantly in nearly every human encounter. From puberty on, barring disease or accident, personal voice quality is a robust, distinctive signature of each individual human existence. Much information about the speaker is revealed in the speaker's voice, including age, sex, emotional and physical state, personality, socioeconomic status, geographic history, and attitude towards the listener and the topic under discussion.

INTRODUCTION

The terms 'voice' or 'voice quality' are often used (especially by speech pathologists or otolaryngologists) to refer to those aspects of sound that are due solely to vocal fold vibration. Terms like 'breathiness', 'roughness', 'hoarseness', and 'clarity' are often used. Changes in the rate of vocal fold vibration contribute to alterations in perceived voice pitch and intonation, and changes in the manner of vibration produce differences in voice timbre (corresponding to the characteristic amplitudes of the series of overtones of the fundamental frequency). However, the notion of voice quality is also construed more broadly when discussing talker identity. 'Speaker recognition' refers to any attribution of identity on the basis of a voice. Speech is produced when the acoustic laryngeal source (the sound energy arising when pulses of air pass through the vibrating vocal folds) is shaped acoustically by the vocal tract. Vocal tract filtering produces resonance peaks and valleys in the acoustic voice signal, which contribute significantly to perceived speaker identity. As the jaw, soft palate, tongue, and lips move, the resonance characteristics of the vocal tract change, giving rise to particular patterns of variation that characterize individual speakers. Other parameters, including intensity (loudness) and its range and variation, rhythm and speaking rate (e.g. words per minute, word or syllable durations, number and duration of pauses), and use of lip rounding or pharyngeal tension or other pronunciation habits, may also cue speaker identity. Although these characteristics can be described and labeled (Laver, 1980), no stable weighting or hierarchical arrangement of all of these components serves to specify an individual human voice. Nor do we understand how listeners select cues or elements from the acoustic voice pattern to achieve speaker recognition.

HUMAN RECOGNITION OF VOICES

Voice recognition has been studied in such diverse disciplines as forensics, biology, evolution, human development, phonetics, psychology, sociolinguistics, neuropsychology, and engineering. Forensic concerns, including the reliability of 'earwitness' testimony and expert reading of spectrograms ('voiceprints'), have motivated many studies of voice (Hammersley and Read, 1996). Much of this research has proceeded pragmatically to establish appropriate guidelines for using voice records in court.

The results of different studies often appear contradictory (Kreiman, 1997). For example, seeing a face while hearing a voice may aid recognition, but may also distract the listener and result in poor learning of the voice. Both male and female voices have been found by different researchers to be easier to recognize overall, and listeners of both the same and different sex have been found to show an advantage. Other factors, such as length and content of the speech sample, quality and context of the speech, time since hearing the voice, and psychological state of the listener, in various combinations, may influence listeners' abilities to accurately identify voices.

Listeners are better at identifying voices speaking a language they know, and they apparently remember voices more easily after active conversation than after passive listening. However, speakers differ in how memorable their voices are, and recognition rates across similarly designed studies vary. Experts generally conclude that neither eyewitness nor earwitness testimonies should be held as incontrovertible evidence in court.

Vocal patterns allow many animals to identify and locate kin without visual contact (Cheney and Seyfarth, 1980). Bats, vervet monkeys, pigtail macaques, chimpanzees, certain species of birds, and timber wolves recognize the individual voices of their offspring, and some species of monkeys can associate the cries of an infant with its mother. For humans, depending on the voice sample and listening conditions, recognizing a familiar speaker requires at most three seconds of speech. The voices of famous individuals can be recognized at well above chance levels when listeners can choose responses from a list of names, but recognition is less certain when listeners must freely identify the speaker. The most rudimentary acoustic material may suffice to cue the identity of the voice (Remez *et al*., 1997).

Studies examining the effects of acoustic manipulations on voice recognition indicate that very varied cues can specify individual voice patterns. For example, for one speaker, precise pronunciation and large pitch variations may be salient, but another voice may exhibit breathiness and long vowels (Van Lancker *et al*., 1985). Converging evidence from acoustic manipulations, mimicry, and neuropsychology suggests that a familiar voice is stored and processed cognitively as a pattern, or 'Gestalt'. In contrast, in dealing with unfamiliar voices, listeners rely on auditory–acoustic features in a comparative and detailed manner, perhaps comparing the perceived features to a template for an ideal or 'average' speaker (Kreiman and Papcun, 1991).

Voices play an important role from the earliest stage of child development. Newborn infants can pick out their own mother's voice from a set of maternal voices (DeCasper and Fifer, 1980), suggesting that voices are differentiated very early in development. In the first year after birth, infants show more interest in words spoken in a familiar voice. A study of school-age children reported that developmental schedules are different for recognizing familiar and unfamiliar voices, in that children recognize classmates' voices as well as adults do, but have relative difficulty distinguishing unfamiliar voices.

For many years, speech perception studies avoided the issue of voice differences, instead seeking consistent cues to speech understanding that were believed to be independent of the personal voice information. However, it has become increasingly clear that the speaker's voice forms an integral part of words stored in auditory memory (Palmeri *et al*., 1993). Meanings remembered from speech samples are affected by the listeners' perceptions of the speakers' abilities, intentions, and opinions as derived from the speakers' voice patterns (Geiselman and Bellezza, 1977). This 'incidental memory' for voice information is nearly as substantial as when subjects are asked to intentionally remember voice information.

VOICE RECOGNITION ABILITY AND DISABILITY

There are large differences in normal abilities to discriminate among voices, recognize voices, and learn new voices. Callers by telephone expect to recognized by voice: when the person they are calling hesitates, due to an initial person-identification failure, there may be brief 'trouble' in the interaction (Schegloff, 1979). In everyday life, listeners mistake one voice for another. In 'tip of the ear' slips, a listener fails to identify a voice that seems familiar. In laboratory testing, people may be able to describe the auditory details of a voice sample that should be familiar to them, or distinguish it from other voice samples, but still not recognize the identity of the voice. The reverse may also occur (Van Lancker, 1991).

CEREBRAL LOCALIZATION OF SPEAKER RECOGNITION

For over a century, it was believed that the left hemisphere was the sole arbiter of communication through the medium of speech. However, early studies of hemispheric specialization for processing of voice quality demonstrated that patients with right temporal and parietal damage performed worse than those with left hemisphere damage when asked to discriminate among voices spoken in different dialects (Assal *et al*., 1981). Later studies of phonagnosia – voice perception deficits – used the voices of famous male entertainers played to subjects who had suffered a stroke to one cerebral hemisphere (Van Lancker and Canter, 1982). Again, right hemisphere damage significantly interfered with recognizing familiar voices, whereas patients with left hemisphere damage

recognized the voices, in many cases as well as normal listeners. Even severely aphasic patients, who could not understand what was being said, knew who was saying it.

Further studies in neurological patients demonstrated that voice recognition (perception of familiar voices) and voice discrimination (perception of unknown voices) are separate abilities (Van Lancker *et al*., 1988). Brain lesions can interfere with one ability while leaving the other unaffected. Brain imaging studies have shown that most patients with deficits in their recognition of familiar voices have right parietal lobe lesions. In these studies, difficulty discriminating among unfamiliar voices occurred in cases of lesions involving the temporal lobe of either the right or the left hemisphere. Thus, voice recognition and discrimination are not dependent on each other and have different neuroanatomic substrates (Kreiman, 1997).

MACHINE RECOGNITION OF VOICES

Individuals cannot 'forget' or 'misplace' their voices, as they can their identification badges. This fact means that reliable automatic voice recognition protocols are convenient, powerful additions to many security systems. Commercial speaker recognition systems perform well in many applications, with false recognition and false elimination rates as low as 1% (Kunzel, 1994).

Two tasks should be distinguished here. The most common ('speaker verification') involves deciding if a speaker (the 'client') is in fact who he or she claims to be. In 'speaker recognition', on the other hand, no identity claim is made by the client. Instead, the system either decides who the speaker is from a database of known speakers or determines that the client is unknown.Verification thus involves matching a client's voice to a single target, while recognition involves comparing the client to a large number of targets: a much more difficult task. Although the precise analysis techniques and statistical matching algorithms vary (Campbell, 1997), the general procedure is as follows. A cooperative speaker who wishes to be recognized is recorded under ideal conditions with good-quality equipment. On subsequent occasions, the speaker enunciates clearly and makes no attempt at vocal disguise. The utterance is analyzed acoustically and compared to the previously created template for the speaker the client claims to be. Matches are based on low-level, mathematically derived, computationally efficient acoustic parameters, and not on high-level, psychologically derived parameters like those that listeners appear to use (Campbell, 1997). A statistical index of the match is created. If the index exceeds some threshold, the person's identity is verified or the speaker is recognized.

Voice recognition by machine differs substantially from how human listeners recognize voices, and the design and performance of automatic systems are not very informative about human behavior. The unlimited-set task performed by listeners is much more difficult than the highly specific matching task performed by most machine algorithms. In human speaker recognition by listening, much less control is possible than in automatic systems. Speakers say whatever they want, and may shout, stutter, mumble, whisper, disguise their voices, or imitate another speaker. Environmental noise – traffic, the voices of other speakers, telephone static – is often present, or the voice sample may be poorly recorded, for example on a low-quality answering machine. Often, the speaker could be virtually anyone – whereas an automatic system usually need only compare an unknown voice sample to the stored template for the one person the speaker claims to be.

Finally, in machine recognition algorithms the threshold for 'recognition' can be set explicitly. Such thresholds are usually rather strict, to minimize incorrect identification of imposters. This necessarily means sometimes excluding clients who should be recognized. However, a level of appeal is possible in machine recognition protocols. For example, workers denied access to a secure area (perhaps because a cold has altered their voice qualities) can normally contact a human security guard who will check their identification. On the other hand, human listeners' thresholds for recognizing a voice – the amount of evidence they require to associate a voice with an identity – cannot be consistently manipulated, and listeners' confidence in their responses has no statistically significant relationship to the accuracy of their identification. Further, no level of appeal is available in forensic applications, where an incorrect identification may result in conviction of an innocent person.

Statistically, given comparably constrained listening conditions, humans do not recognize voices as well as machines do, but in real life they face a much harder task. They separate the target voice from noise, possibly involving a background of other voices; they map this uncontrolled, degraded signal onto some decayed memory trace. Then, from an apparently limitless repertory of potential targets accumulated over many years, listeners somehow recognize the talker. We still have much

to learn about the human ability to recognize speakers, and it continues to amaze.

References

Assal G, Aubert C and Buttet J (1981) Asymmétrie cérébrale et reconnaissance de la voix. *Revue Neurologique* **137**: 255–268.

Campbell JP (1997) Speaker recognition: a tutorial. *Proceedings of the IEEE* **85**: 1437–1462.

Cheney DL and Seyfarth R (1980) Vocal recognition in free-ranging vervet monkeys. *Animal Behavior* **28**: 362–367.

DeCasper AJ and Fifer WP (1980) Of human bonding: newborns prefer their mothers' voice. *Science* **208**: 1174–1176.

Geiselman RE and Bellezza FS (1977) Incidental retention of speaker's voice. *Memory and Cognition* **5**: 658–665.

Hammersley R and Read JD (1996) Voice identification by humans and computers. In: Sporer SL, Malpass RS and Koehnken G (eds) *Psychological Issues in Eyewitness Identification*, pp. 117–152. Hillsdale, NJ: Erlbaum.

Kreiman J (1997) Listening to voices: theory and practice in voice perception research. In: Johnson K and Mullennix JW (eds) *Talker Variability in Speech Processing*, pp. 85–108. New York, NY: Academic Press.

Kreiman J and Papcun G (1991) Comparing discrimination and recognition of unfamiliar voices. *Speech Communication* **10**: 265–275.

Kunzel HJ (1994) Current approaches to forensic speaker recognition. In: Chollet G, Paoloni A and Bimbot F (eds) *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, pp. 135–141. Martigny, Switzerland: ESCA.

Laver J (1980) *The Phonetic Description of Voice Quality*. Cambridge, UK: Cambridge University Press.

Palmeri TJ, Goldinger SD and Pisoni DB (1993) Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 309–328.

Remez RE, Fellowes JM and Rubin PE (1997) Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* **23**: 651–666.

Schegloff EA (1979) Identification and recognition in telephone conversation openings. In: Psathas G (ed.) *Everyday Language Studies in Ethnomethodology*, pp. 23–78. New York, NY: Irvington.

Van Lancker D (1991) Personal relevance and the human right hemisphere. *Brain and Cognition* **17**: 64–92.

Van Lancker D and Canter J (1982) Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition* **1**: 185–195.

Van Lancker D, Cummings J, Kreiman J and Dobkin BH (1988) Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex* **24**: 195–209.

Van Lancker D, Kreiman J and Wickens T (1985) Familiar voice recognition: parameters and patterns. Part II: recognition of rate-altered voices. *Journal of Phonetics* **13**: 39–52.

Further Reading

Bricker PD and Pruzansky S (1976) Speaker recognition. In: Lass NJ (ed.) *Contemporary Issues in Experimental Phonetics*, pp. 295–326. New York, NY: Academic Press.

Hecker MHL (1971) Speaker recognition: an interpretive survey of the literature. *ASHA Monographs* **16**.

Laver J (1981) The analysis of vocal quality: from the classical period to the 20th century. In: Asher R and Henderson E (eds) *Toward a History of Phonetics*, pp. 79–99. Edinburgh: Edinburgh University Press.

McGehee F (1944) An experimental study of voice recognition. *Journal of General Psychology* **31**: 53–65.

Murray IR and Arnott JL (1993) Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* **93**: 1097–1108.

Nolan F (1997) Speaker recognition and forensic phonetics. In: Hardcastle WJ and Laver J (eds) *The Handbook of Phonetic Sciences*, pp. 744–767. Oxford, UK: Blackwell.

Scherer KR (1986) Voice affect expression: a review and a model for future research. *Psychological Bulletin* **99**: 143–165.

Titze IR (1994) *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

Van Lancker D (1997) Rags to riches: our increasing appreciation of cognitive and communicative abilities of the human right cerebral hemisphere. *Brain and Language* **57**: 1–11.

Williams CE and Stevens KN (1981) Vocal correlates of emotional states. In: Darby J (ed.) *Speech Evaluation in Psychiatry*, pp. 221–240. New York, NY: Grune and Stratton.

Speech Error Models of Language Production

Intermediate article

Joseph P Stemberger, University of British Columbia, Vancouver, British Columbia, Canada

CONTENTS

Some models of language production use speech errors (malfunctions of language processing) as evidence for the organization of the language system.

INTRODUCTION

Some models of language production have been developed to account for the involuntary errors that occur during speaking. The ways in which a system malfunctions is assumed to reveal details about how information is represented, about how the system is structured, and about the nature of processing. It should be noted that no model has been developed solely on the basis of information about errors. All models are also based partly on information about correct production and are of types that have been developed for other aspects of language. Most models are restricted to adult language systems, but a few network models additionally are concerned with learning (though none of these address actual developmental data). After a brief review of the levels of language and of the basic kinds of speech errors that occur, I address models that are heavily based on error phenomena and evaluate how well they account for what is known about speech errors. (*See* **Aphasia**; **Language Acquisition**)

LEVELS WITHIN THE LANGUAGE PRODUCTION SYSTEM

Before we can address errors, it is necessary to review the types of information and processes that are basic to language, as well as the degree to which these are organized into separate modules. Table 1 presents a typical picture of language production. Some models have separate modules for each function, while others merge all the intermediate levels into one. (*See* **Speech Production**; **Modularity**)

Models vary in the degree of seriality between levels in the system (serial modularity). In a fully serial system (which no model uses), processing at one level goes to completion before any processing begins at the next level down. Hesitations during speech suggest at least a cascading serial system: as soon as partial information is available at a given level, it is passed on to the next level down, so that processing may begin there. However, once information has been passed to the next level, the speaker does not return to the higher level to process it further. In a cascading parallel system, in contrast, as soon as partial information is available at the lower level, it is passed back to the higher level, and may affect the final outcome of processing at the higher level. Processing at higher levels nevertheless goes to completion earlier than processing at lower levels. In a fully parallel system, all levels are processed simultaneously; the levels need not be discrete, and information from all levels can be intermixed. It should be noted that, in order to check the accuracy of the message at all levels, the monitoring function reverses the order of processing from bottom to top; a speaker may decide, on the basis of information at a lower level, to make changes at a higher level, leading to apparent effects of, for example, phonological information on lexical processing.

TYPOLOGY OF SPEECH ERRORS

All levels of the language system can be involved in error. However, errors in message formulation and in motor processing are difficult to detect and are rarely studied. The bulk of studies (and models) have focused on phonological errors, followed by

Table 1. Basic functions in language production

Message-formulation	Deciding what information to encode for speech production
Lexicon	Encoding meaning into individual words
Syntax	Ordering of words (including inflectional morphology)
Phonology	Constructing the pronunciation of a word at an abstract level
Phonetics/motor processing	Constructing a more detailed and more concrete representation of the pronunciation of a word
Monitoring	Making a final decision about the appropriateness of the content; if an error is detected, production is halted and the error is repaired

work on lexical errors, morphological errors, and (a distant fourth) syntactic errors.

Errors bear up to seven different relations to the intended target (Table 2). Rule application errors are possible only in systems that contain rules, and can be treated as nonbasic. For example, the morphological regularization error *choosed* for 'chose' can be viewed either as the overapplication of a morphological rule (the past tense *-ed* rule), or as the substitution of a regular past tense pattern (*base+ed*) for an irregular (lexically idiosyncratic) past tense pattern. There is also ambiguity relative to other error types. Consider the following error:

target bite block
error bLite block

This can be analyzed as an addition error, because the element /l/ (anticipated from the next word) was added where no element appeared in the target word. However, it can also be analyzed as a substitution error involving a higher-level unit: the onset /bl/ (anticipated from the next word) was substituted for the target onset /b/. It is possible to create a theory-neutral typology of errors, but the actual mechanisms underlying errors may differ substantially in different theories.

Another important dimension is the relation between the target and source words (Table 3); different relations have different statistical properties. Contextual errors result from interference between two elements that are both appropriate for a given utterance. For example, in *why was see shurprised* instead of 'why was she surprised', there are two consonants that are both a part of the utterance, but they have been produced in reversed order. In this example, the two interacting elements are part of the same unit at one level (the same sentence), but are part of different units at a lower level (two different words); in other errors, the two interacting consonants may be in different sentences, or in the same word. Noncontextual errors do not involve interference from another element in the context. Examples include use of the wrong word or phoneme (e.g. *hair* for 'feathers'), blending two competing words into one (e.g. *flaste* for 'flavor' plus 'taste'), regularization, and (apparently) random loss or addition of words or phonemes.

Table 2. Relation of an error to the target

Substitution	Wrong element
Deletion	No element where an element is required
Addition	An element where no element should be present
Blend	One element with characteristics of two competing elements
Rule overapplication	A rule applied that should not have
Rule underapplication	A rule should have applied, but did not
Combinations of the above	

MODELS OF LEXICAL REPRESENTATION

There are three main approaches to the representation of lexical items, plus an orthogonal issue.

Lexical Item as Information Store

A lexical item is a location in long-term memory wherein is stored all relevant information about a word: its meaning, syntactic properties, pronunciation, spelling, etc. Gaining access to the lexical entry gains access to all information within the entry. This is possible only within symbolic models (e.g. Garrett, 1975). (*See* **Symbolic versus Subsymbolic**)

Lexical Item as Node with Pointers

A lexical item contains no information. It is a node that is connected to different types of information. Connections come in from semantic and syntactic units that are relevant to the item, and go out to

Table 3. Relation between target and source words

Contextual	Location: Directionality:	Within-unit Anticipation	Between-unit Perseveration	Exchange
Noncontextual				
Combined contextual and noncontextual				

phonological and syntactic units. The function of a lexical item is to map semantic activations onto phonological activations. This is characteristic of local connectionist and similar models (Dell, 1986; MacKay, 1987; Stemberger, 1985; Berg, 1988). (*See* **Spreading-activation Networks**)

No Lexical Items

Semantic activations are mapped directly onto phonological activations with no intervening levels. This is possible only within distributed connectionist and similar models (Dell *et al*., 1993; Vousden *et al*., 2000). (*See* **Connectionism**; **Language, Connectionist and Symbolic Representations of**)

Meaning versus Phonological Form

One orthogonal issue is whether the lexical units are semantically based or phonologically based. Following Levelt (1989), many models have both. Lemmas are accessed on the basis of meaning, which in turn lead to the access of form-based lexemes. For example, homophones like *need* and *knead* correspond to two different lemmas, but to a single lexeme. From lexemes, the phonological representation of the word is then constructed.

MECHANISMS OF RETRIEVAL AND PRODUCTION

We may distinguish between mechanisms to retrieve information, and mechanisms that combine different items once they have been retrieved (though many connectionist models do not make this distinction). We can also distinguish the mechanisms in terms of search versus spreading activation. The next section presents details about models and evaluates them. (*See* **Lexicon, Computational Models of**; **Lexicon**)

Symbolic Models

Motley *et al*. (1983) proposed that all errors arise through competition, either between closely related items (in noncontextual errors) or between two elements slated to be produced within some chunk of speech. Competition could be between lexical items, phonological elements, or syntactic structures. However, detailed proposals about processing have not been made. The spirit of their proposals has been instantiated in network models.

Garrett (1975) proposed a serial cascading model that relied on search mechanisms. The result of message formulation is a semantically based pattern, which is compared to lexical entries. Words are accessed in several stages. (1) A lexical item (=lemma) is located in the lexicon via search (of an unclear nature but presumably involving pattern matching). Errors at this stage result in a wrong word that is semantically related to the target; any additional phonological resemblance to the target is due to chance. (2) The word's form (=lexeme) is accessed from a phonologically organized list. Errors at this stage result in a word that is phonologically related to the target but not semantically related. (3) Occasionally, two contextually synonymous words are both accessed; they compete for copying into the same syntactic position, leading to word blends.

Other types of error occur when lexical items are inserted into syntactic structures (which are accessed in parallel; the exact interplay between lexical items and syntactic structures has not been spelled out). (4) The syntactic structure includes inflectional affixes and closed class words as a part of the structure. Errors can occur in which the wrong structure is selected, leading, for example, to use of an uninflected form or of the wrong inflected form. Errors can also occur in which a word is inserted into a structure in which it is not supposed to occur, leading, for example, to violations of subcategorization and selectional restrictions. (5) The phonological form is then inserted into the sentence. Two kinds of errors can arise at this stage. In the first, a target word is inserted into the wrong location in the sentence structure, displacing some other word. The displaced word may be inserted in the slot that had been intended for the first word (leading to the transposition of words), or the first word may be

inserted in both slots (leading to the anticipation or perseveration of the word). Second, sound errors may occur. Because inflectional affixes and closed class words are already a part of the structure, they are never involved in phonological errors. Further, two inflectional affixes may never be transposed. (Stemberger (1985) shows that closed class lexical items and affixes do take part in morphological exchange errors and phonological errors, but at rates lower than those of open class items. He argues that this may be a frequency effect, since high frequency words in general show low rates of phonological errors.)

Shattuck-Hufnagel (1979) provides further detail about the insertion process. There is a scan-copier that copies the words from the lexicon to the syntactic slots, one phoneme at a time. To keep track of where it is in the copying process, the scan-copier notes what has been copied (by checking off the segments as they are copied) as well as the surrounding segments and the syllable structure (so it can return to the right word to copy the next segment). Errors may occur when the scan-copier fails to check off a copied segment. The unchecked segment remains available for copying, and may be erroneously copied into a later word, leading to a perseveration. Errors may also occur when the scan-copier (after copying a segment) returns to an incorrect location and copies whatever segment is there; this is an anticipation or perseveration. Because the scan-copier keeps track of the segments to be copied and their location, these errors tend to involve segments that share many features with the target segment, occur in a similar location in the word and in the syllable, and occur in the vicinity of similar consonants and vowels. The erroneously copied segment is checked off. When the scan-copier reaches the location where the erroneously copied segment should have been, it cannot copy the segment because it is checked off. The only free consonant is the original target segment that was not copied when it should have been, so that consonant is now copied. The result is that the two consonants have exchanged places. Exchanges are predicted to be the most frequent type of error. In a minority of instances, the first miscopied segment is not checked off, and so is additionally copied into its rightful place, resulting in an anticipation error. (No mechanisms were provided for noncontextual phonological errors, but perhaps these are the result of miscopying individual features.)

Shattuck-Hufnagel (1979) proposed that the monitor searches for errors, and can correct them. This monitor gives rise to 'incomplete' errors, in which the speaker anticipates a later phoneme, but stops speaking and corrects the error before reaching the word that is the source of the miscopied phoneme. (Such errors are ambiguous between anticipation and exchange errors. Dispute over the nature of these errors prevents us from evaluating the fact that Shattuck-Hufnagel's model predicts that exchange errors predominate.) Levelt *et al*. (1991) point out that the monitor can lead to 'mixed' errors involving both semantic and phonological similarity to the target word. Semantically related word substitutions can be spotted by the monitor as being phonologically dissimilar to the speaker's intended target. Phonologically related word substitutions can be spotted as being semantically dissimilar to the speaker's intent. Word substitutions that are related to the target along both dimensions are less likely to be detected by the monitor, and so mixed errors are more likely than mere chance, even though there is no direct influence of the phonological level on the lexical level.

Network Models: Local Models

A very different alternative is presented by work within an interactive activation model (also known as a local connectionist model: Dell, 1986; Stemberger, 1985; Berg, 1988; with a similar approach by MacKay, 1987). In these models, elements are accessed via spreading activation rather than via search. Initially, a meaning-based representation for a sentence is activated. Activation spreads from active semantic units over connections to words and syntactic structures that express those meanings. Word units such as {DOG} and {CAT} sum the activation that arrives over many connections. The two word units are in competition, and inhibit each other (i.e. lower each other's activation levels). The most active unit is the one that has the most activation from semantic units, and it inhibits all competitors. As those competitors decrease in activation levels, the winner is 'disinhibited', and increases in activation level, until some upper bound is reached; this is known as the 'rich-get-richer' principle. If the wrong word has the greatest activation level, it wins, leading to a semantically related word substitution. If two words are equally activated, both win; competition between them is passed down to lower levels in the system, where some phones from each word win, leading to a blend error. Activated words spread activation to their component phonemes. Activated phonemes spread activation back to words. If a non-target word gains too much activation, it may inhibit the

target word, resulting in a phonologically related word substitution. The model correctly predicts that mixed errors will be statistically overrepresented, since the erroneous word sums activation from both semantics and phonology. Noncontextual phonological errors result when some non-target phone erroneously has the most activation, and inhibits the target phone. This is especially likely if the non-target phone shares features with the target phone, and thus sums activation from target feature units. Different words in the same sentence are partially active at the same time, and pass activation to the phonological units, resulting in interference between words. If the interference is sufficient, a contextual phonological error results. This is especially likely if the two interfering phonemes share a lot of features, and if they occur in similar parts of the word, and if they occur in words that contain the same phonemes. It should be noted that apparent ordering errors (such as anticipations or exchanges) occur during the accessing phase of processing. These models easily derive perseveration and anticipation errors because they are a single accessing error, but exchange errors, which involve accessing errors at two locations, are predicted to make up a small proportion of total errors; because of the frequency of incomplete errors, this prediction is difficult to test.

Competition between two elements need not end up with a single accessed element. Structures may be altered to accommodate both competitors. If two nouns compete, a compound noun can be produced (e.g. *in the ancient stick-time* for 'in ancient times' plus 'in the sticks'). If /p/ and /l/ compete, a /pl/ cluster can be produced. Stemberger (1990) demonstrates that competition between /p/ and /l/ is more likely to be resolved as a cluster than as a single consonant. He also demonstrates that syllabic /r/ (as in the North American English pronunciation of *bird*) is less likely to replace a vowel than to be added to an onset or coda cluster (e.g. for 'Berkeley bus' *Berkeley brus*, not **Berkeley berse*).

Stemberger (1985) emphasizes that there may be 'gang effects'. Many competitors with low levels of activation can reinforce the same output, if they are all very similar phonologically. For example, a nonce past tense form such as *yecked* can be produced in the absence of morphological rules as a blend of the base form *yeck* (which is the most activated form but not a clear winner, because it is not semantically appropriate to the past tense) and many low-activated past-tense forms, especially if they are phonologically similar to *yeck* (such as *pecked, checked, wrecked*, etc.). High-frequency phonemes that occur in large numbers of words are reinforced in a similar way and are less likely to undergo errors.

MacKay (1987) provides a position intermediate between symbolic and connectionist models. Noncontextual errors are accessing errors, with similar mechanisms and properties as in interactive activation models. Contextual errors, in contrast, are not accessing errors but relate to incorrect execution of correctly assembled structures.

Stemberger (1985) also assumes that every word in a sentence is accessed and held active. A command is then given to execute the sentence, and the words and their sounds are then produced in the target order. Execution errors may occur at this point: a word or a segment shows up one or two units too early. Unlike accessing errors, these result in grossly non-English word and consonant sequences (such as *rpinciple* for 'principle' and *she was the driving bus* for 'she was driving the bus'). Such errors are rare, but most models do not account for them.

Network Models: Distributed Models

These models do not distinguish different intermediate levels (though they could be modified to do so). In principle, information is mapped from semantic input to phonological output in one step (via an intermediate layer of hidden units). These models involve spreading activation, but there is no interaction between input and output except in coarse time steps. The focus of these models is on learning, rather than on structure or attention to the full range of error types.

Dell *et al*. (1993) present a recurrent network model. Inputs could in theory correspond to the meaning of whole sentences but in practice were either random patterns or the orthography of the target word, and the system learned to output a single word at a time. However, if proposition-sized (or larger) meaning patterns were to be used, the system could in principle learn to produce whole sentences. The system outputs a single segment-sized unit at a time, as a set of phonological features. Because the input does not change during the entire word (or sentence), context units are added to keep track of progress through the word. The activation of the hidden units is copied onto internal context units, and the activation of the output units is copied onto external context units. Both sets of context units serve as input to the hidden units on the next segment-sized iteration. The external context units tell the system to

produce, for example, the segment in *cat* that is first (/k/, first iteration), that follows /k/ (/æ/, second iteration), or that follows /œ/ (/t/, third iteration), or that follows /t/ (nothing, fourth iteration). The internal context units give more complex information, so that in a word with a repeated phoneme (e.g. *perplexed*) the system can produce different segments after the first and second tokens of /p/.

Because of the limitations on the nature of the input, the model could not deal with syntactic errors, word ordering errors, or between-word contextual phonological errors. However, noncontextual errors occurred, at both the lexical and phonological levels. Both semantically related and phonologically related word substitutions occurred, with a predicted overrepresentation of mixed errors. Word blends also occurred, especially at points where the two synonyms shared a phoneme. It was found that errors were most likely to occur at the beginning of the word. It was found that sharing phonemes with other words in the lexicon (high neighborhood density) decreased error rates. This model can account for frequency effects (lexical and phoneme), effects of position in a word, and similarity between phonemes (features and syllable position).

Vousden *et al*. (2000) present another distributed model of a very different sort: OSCAR (Oscillator-based Associative Recall). The system takes meaning-based input, and has a simultaneous representation of all phonemes in the output unit. Potentially, an entire sentence can be encoded in the output, but the implementation uses the word as the output, and does not address lexical retrieval or syntactic processing issues. The system outputs phonemes in a way that encodes serial ordering. There is a series of oscillators of various duration. Durations can be as short as a syllable or as long as a long sentence (or even a higher unit). A phoneme is associated with a given value for all the oscillators. Two phonemes are similar if they have a similar set of feature vectors or they are associated with similar states in the oscillators. For example, in *big cat*, the /b/ and /k/ are both at a certain point in cycle of a syllable-sized oscillator; the /ɪ/ and /æ/ are both at the same later point; and the /g/ and the /t/ are both at an even later point. But the /b/ and the /k/ are at a different point for an oscillator with a long cycle (and so they are produced in different syllables). The model was tested against the following error findings and did well. It can account for

- the proportion of errors that are contextual versus noncontextual,
- the proportion of errors that are anticipations versus perseverations versus exchanges (under the assumption that anticipations are more common than exchanges),
- the fact that errors involve phonemes with similar features,
- the syllable position effect (as a strong tendency, not an absolute), and
- the effect of distance (contextual errors decrease as a function of distance between the two sounds).

It cannot, however, handle cross-position effects, where /p/ and /l/ are more likely to be blended into the onset /pl/, or syllabic /r/ is more likely to become part of an onset than a nucleus. Nor can it account for the fact that two phonemes are more likely to interact if they are preceded or followed by identical phonemes. It does a mixed job of dealing with other effects as well.

CONCLUSIONS

Models of speech production based on errors show the same range of variation as other processing models, and have been used to address similar issues (such as symbolic versus subsymbolic representations and processes, search versus spreading activation, etc.). Computer-implemented models tend to deal with a subset of error types, with much success, but it is unclear whether they will be successful when a fuller range of phenomena are addressed. Serial models have an inability to deal with interactions between levels (such as mixed errors) in an integral fashion but can add additional components (such as a monitor) to derive such effects. No model has attempted to deal with the full range of error phenomena that are attested.

References

Berg T (1988) *Die Abbildung des Sprachproduktionsprozesses in einem Aktivationsflussmodell: Untersuchungen an deutschen und englischen Versprechern*. Tübingen, Germany: Niemeyer.

Dell GS (1986) A spreading activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–321.

Dell GS, Juliano C and Govindjee A (1993) Structure and content in language production: a theory of frame constraints in phonological speech errors. *Cognitive Science* **17**: 149–195.

Garrett M (1975) The analysis of sentence production. In: Bower G (ed.) *Psychology of Learning and Motivation*, vol. 9, pp. 133–177. New York, NY: Academic Press.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt WJM, Schriefers H, Vorberg D, Meyer AS, Pechmann T and Havinga J (1991) Normal and deviant lexical processing. Reply to Dell and O'Seaghdha. *Psychological Review* **98**: 615–618.

MacKay D (1987) *The Organization of Perception and Action: A Theory for Language and Other Cognitive Skills*. New York, NY: Springer.

Motley MT, Baars BJ and Camden CT (1983) Experimental verbal slips studies: a review and an editing model of language encoding. *Communication Monographs* **50**: 79–101.

Shattuck-Hufnagel S (1979) Speech errors as evidence for a serial ordering mechanism in sentence production. In: Cooper WE and Walker ECT (eds) *Sentence Processing*, pp. 295–342. New York, NY: Halsted Press.

Stemberger JP (1985) An interactive activation model of language production. In Ellis A (ed.) *Progress in the Psychology of Language*, vol. 1, pp. 143–186. London, UK: Lawrence Erlbaum.

Stemberger JP (1990) Wordshape errors in language production. *Cognition* **35**: 123–157.

Vousden JI, Brown GDA and Harley TA (2000) Serial control of phonology in speech production: a hierarchical model. *Cognitive Psychology* **41**: 101–175.

Further Reading

Baars BJ (1993) *Experimental Slips and Human Error: Exploring the Architecture of Volition*. New York, NY: Plenum Press.

Cutler A (1982) *Slips of the Tongue and Language Production*. Amsterdam, Netherlands: Mouton.

Dell GS, Burger LK and Svec W (1997) Language production and serial order: a functional analysis and a model. *Psychological Review* **104**: 123–147.

Fromkin VA (1973) *Speech Errors as Linguistic Evidence*. The Hague, Netherlands: Mouton.

Fromkin VA (1980) *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*. New York, NY: Academic Press.

Rossi M and Peter-Defare É (1998) *Les lapsus: ou comment notre fourche a langué*. Paris, France: Presses Universitaires de France.

Stemberger JP (1989) Speech errors in early child language production. *Journal of Memory and Language* **28**: 164–188.

Speech Perception

Introductory article

Miranda Cleary, Speech Research Laboratory, Indiana University, Bloomington, Indiana, USA
David B Pisoni, Speech Research Laboratory, Indiana University, Bloomington, Indiana, USA and DeVault Otologic Research Laboratory, Indiana University School of Medicine, Indianapolis, Indiana, USA

CONTENTS

Speech perception is the process of recognizing the words and speech sounds of a spoken language. Specific acoustic cues in the speech signal contribute to the identification of phonemes, minimal linguistic contrasts that distinguish different words from each other.

INTRODUCTION

Speech perception is the process of recognizing the words and speech sounds of a spoken language. As in most perceptual research, the fundamental issue in speech perception is explaining perceptual constancy, specifically, in this case, how we recognize certain sound patterns as being 'linguistically equivalent' despite large differences in the physical properties of the acoustic signals that reach our ears. For example, how does a listener recognize the word 'dog' spoken rapidly in connected speech by a small child as somehow equivalent to 'dog' uttered slowly in isolation by a large man?

In order to understand perceptual constancy, we also need to investigate the complementary problem of how we perceive the meaningful linguistic differences that differentiate one word from another – how we tell 'bit' from 'bet' or 'right' from 'light', for example. It may be useful here to emphasize that perceiving a meaningful linguistic difference is not the same as perceiving 'any old difference' in what you hear. Although the study of auditory psychophysics has revealed much about the basic limits of our ability to tell one acoustic signal apart from another, these findings do not necessarily help us understand the perceptual equivalence of spoken words. Unlike psychophysical 'sensory equivalence', where beyond a certain point listeners may be *unable* to distinguish between two physically different stimuli, the 'perceptual equivalence' of 'dog_1' and 'dog_2' as spoken by the child versus the grown man exists regardless of a listener's awareness that the two utterances differ acoustically.

Speech perception is nevertheless closely linked to our more general-purpose auditory sense. To a lesser extent speech perception is also tied to how we visually process human faces in motion: viewing the motions of a talker's lips and jaw is known to help people identify speech sounds and recognize words. In some unusual instances in which sound and image do not coincide (like a badly dubbed movie), a view of the talker's lips can actually partially 'override' the acoustic signal, changing perceivers' judgements about what they are hearing.

THE SPEECH SIGNAL

The study of speech perception is closely related to the study of phonetics, that is, the measurement and description of the physical speech signal in terms of its acoustic characteristics, and the physiological/anatomical properties and actions that give rise to this signal. The human vocal tract includes those parts of the mouth and throat that can be used to modify the air stream issuing from a talker's lungs. A partial list of these speech articulators includes the vocal folds (the valve-like structure atop the trachea), the oral cavity of the inner mouth, the nasal cavities, the tongue, and the lips. The articulators constrain the airflow in particular ways, resulting in a pattern of pressure changes at the listener's ear. A speech waveform (see Figure 1) is a graphical display of how this displacement of

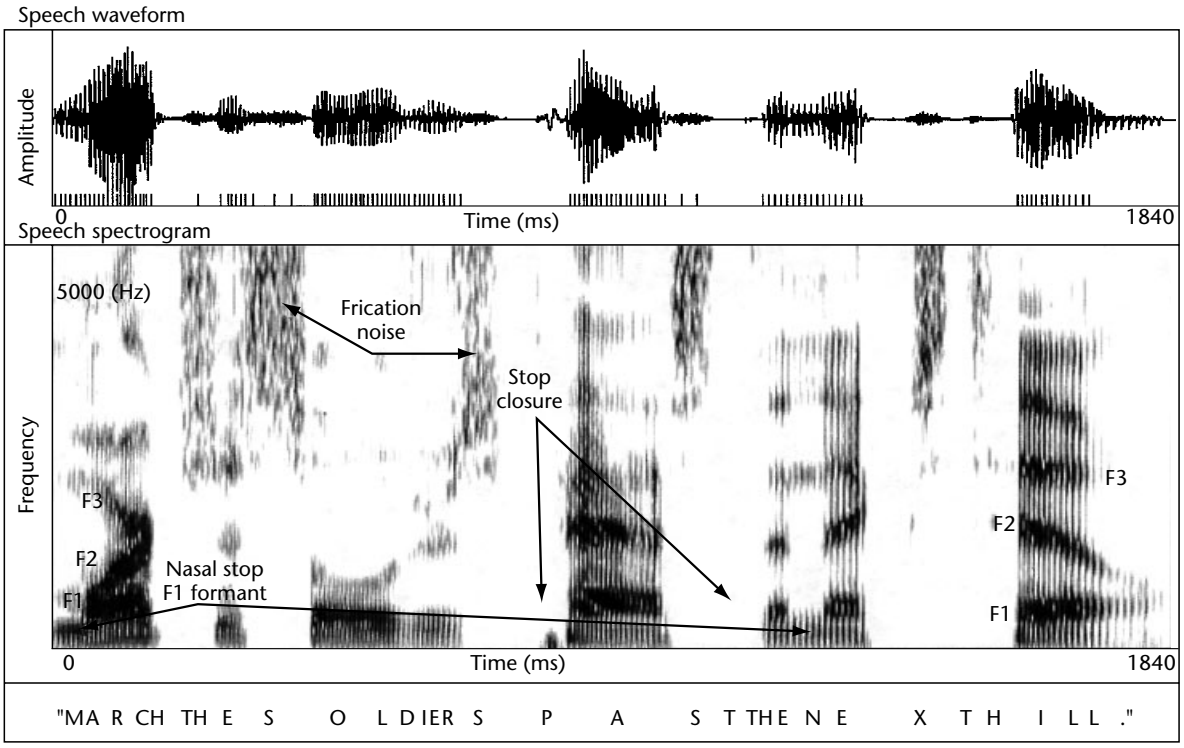


Figure 1. Wide-band speech spectrogram with examples of well-studied acoustic cues to phoneme identity as labeled. (Higher frequencies have been made more intense for purposes of illustration.)

air particles at a particular location in space changes as a function of time. The greater this displacement, the louder the sound, typically.

Speech sounds are complex signals in that not only does the overall pitch of a voice rise and fall during an utterance, but the speech waveform itself can be analyzed as the simultaneous combination of sounds of different frequencies, where the relative loudness of these contributing frequencies varies from moment to moment. Some of the basic acoustic properties of human speech can be represented using a visual display called a spectrogram (see Figure 1). A spectrogram shows how the intensities of the different component frequencies of the speech signal change relative to each other, over time. Typically, the x-axis of a spectrogram represents the passage of time in milliseconds, running from left to right, while the y-axis represents frequency in units called hertz (Hz), with low frequencies plotted near the origin and higher frequencies above. Shaded areas indicate the presence of acoustic energy at particular frequencies, over time.

A cochleagram resembles a spectrogram except that a cochleagram incorporates what is currently known about the sensitivity of the human ear to different frequencies at different intensities. For example, although it is true that at low intensities, the mid-frequencies (around 1000–5000 Hz) are actually perceived as louder than either very low frequencies or very high frequencies, it can generally be said that at a given intensity, the ear is more sensitive to small frequency changes (measured in terms of Hz) at the lower end of the frequency range (e.g. $< \sim$6000 Hz) than equal-sized changes in terms of Hz at very high frequencies. Listeners are also more likely to notice a change in loudness at already low intensities than when exposed to an equal-sized air pressure change at the high end of the intensity range. Thus, frequency in a cochleagram is often graphed on a logarithmic frequency scale, and relative loudness (darkness) plotted according to an analogously adjusted intensity scale. Other auditory effects on the signal, such as the perceptual enhancement of sudden onsets of intensity, can also be incorporated. Although these auditory-model-based methods of representing the speech signal at the level of the peripheral nervous system are still relatively new and undergoing revision, they potentially offer a more accurate picture of the speech signal as it enters the central nervous system.

In a wide-band spectrogram, each of the semi-regularly spaced vertical striations corresponds to

one opening and closing cycle of the vocal folds. During speech, as air is pushed out of the lungs, the flapping of the vocal folds creates a quasi-periodic 'buzz'. This 'buzz' is called the fundamental frequency, or 'F0', and is defined as the rate at which the vocal folds flap open and shut. F0 is among the most basic characteristics of a human voice and its average value (e.g. ~120 Hz for an adult male) contributes to the perceived pitch of a speaker's voice. During fluent speech, the vocal folds vibrate about two-thirds of the time. Talker-controlled adjustments of F0 play a major role in perception of sentence intonation and word stress.

An important acoustic property related to F0 are its harmonics. Speech, like other complex sounds based on a periodic noise source, tends to contain energy not just at the fundamental frequency, but also at frequencies called harmonics, that are integer multiples of F0. So, for example, if the F0 value is 100 Hz, its harmonics would include 200 Hz, 300 Hz, 400 Hz, etc. The actual intensity of each harmonic is determined by other aspects of the vocal tract configuration, so that some harmonics might be particularly intense while others are weak or inaudible.

In wide-band spectrograms of carefully articulated speech, vowel sounds are visually quite prominent, usually characterized by vertical F0 striations and dark bands of horizontally running formants. Formants are concentrated regions of acoustic energy created (usually) through the enhanced intensity of certain harmonics and the attenuation of other harmonics due to the natural resonance characteristics of the vocal tract (that is, factors such as its shape, size, and composition). The lowest frequency formant is typically referred to as 'F1', the next highest formant as 'F2', and so on. F1 typically ranges between 270 and 850 Hz, F2 between 850 and 2700 Hz. In a spectrogram, steeply sloping formants indicate rapid changes in the resonant frequencies due to articulator movement and are typically referred to as formant transitions.

The relative positioning along the frequency dimension of F1, F2, and to some extent, F3 helps to perceptually distinguish the different vowel sounds. Low-frequency first formants, as in vowel portions of the words 'wh*o*' or 'h*e*', result from the central body of the tongue arching up towards the roof of the mouth. The further back in the mouth this arching occurs, the lower the F2 frequency. The F2 of 'wh*o*' is much lower than the F2 in 'h*e*', for example. Relative duration is also used to distinguish some vowels: for example, the vowel in 'b*e*t' is typically shorter than the vowel in 'b*a*t', thereby helping to distinguish the two sounds even though their formant values are quite similar. Some information about vowel identity is also present in the rapid formant transitions that initiate and conclude a vowel sound flanked by consonants.

Consonant sounds can be roughly sorted along three major articulatory dimensions: (1) manner of articulation (the way and degree to which the articulators constrict), (2) place of articulation (where the constriction occurs in the vocal tract), and (3) voicing (whether the vocal folds are vibrating during articulation).

Fricative consonants are produced via a narrow but incomplete closure of the articulators and are primarily characterized by intervals of aperiodic noise at particular frequencies – that is, energy not structured into formants, but instead 'smeared' across a band of frequencies, without emphasis of particular harmonics. In American English, the fricatives [s] as in *s*ue, [ʃ] as in *sh*oe, [z] as in *z*oo, and [ʒ] as in a*z*ure, have noise energy concentrated at the higher frequencies (higher for [s] and [z] than for [ʃ] and [ʒ]) and are of relatively long duration, while [f] as in *f*in, [ɵ] as in *th*in, [v] as in *v*at, and [ð] as in *th*at, have shorter durations and consist of weaker energy spread over a wider frequency range.

Stop consonants involve making a complete vocal tract closure which is then abruptly released, allowing pressurized air to escape (as in [b], [d], and [g], for example). Abrupt intervals of near-silence in a spectrogram can therefore indicate the presence of stop consonants. The release of a stop closure is necessarily marked by a very brief high-frequency noise burst, the intensity and frequency of which tends to vary predictably for the different stop consonants. This release burst is sometimes followed by additional aspiration noise generated at the open vocal folds just prior to the onset of vibration. The nasal stops, [n] as in ra*n*, [m] as in ra*m*, and [ŋ] as in ra*ng*, differ from other stops in that they involve, in addition to the oral cavity closure, the opening of a passageway into the nasal cavities through which air escapes during the oral closure. Due to this alternative routing, nasal stops have a low (< 300 Hz) but relatively intense first formant (also called the nasal murmur) during the oral constriction and some higher formant structure, weakly visible on a spectrogram above about 2000 Hz. Additional classes of sounds in American English include the liquids ([l] as in *l*ip, [ɹ] as in *r*ip) and glides ([j] as in *y*ou, [w] as in *w*oo), both of which involve only partial constriction, permitting strong 'vowel-like' formant structure.

Each of these 'manner' classes (fricative versus stop versus nasal, etc.) contains sounds

distinguished from each other only by place of constriction in the vocal tract. Invariant acoustic cues for distinguishing the various places of articulation have proved elusive, making perception of place contrasts a long-standing area of study. In the case of stop consonants, acoustic information known to contribute to perception of place contrasts includes the frequency and rate of change of F2 and F3 transitions into (or out of) adjacent vowels. The distribution of frequency components present in the brief release burst, in the case of nonnasal stops, and within the prolonged frication noise, in the case of fricatives, can also serve to help cue place of articulation.

UNITS OF RECOGNITION

The 'speech sound categories' we have been referring to, [p] versus [b] for example, are more commonly referred to as *phonemes*. If you remember doing 'phonics' exercises as a child just beginning to learn to read the English language, you are already somewhat familiar with the notion of a phoneme. Phonemes are basically those sound contrasts that a language community uses to distinguish one word from another – the type of contrast which distinguishes the words 'pig' from 'big' or 'pit' from 'pat', for example. Do not confuse the printed alphabetic letters of English with phonemes – English spelling only partially reflects phonemic information. The letter sequence 'sh', for example, represents just one phoneme, [ʃ], in the International Phonetic Alphabet (IPA), a system of notation similar to the pronunciation key used in dictionaries.

It is important to realize that different languages use different sets ('inventories') of phonemes. To illustrate, although the difference between [l] and [ɹ] may seem obvious to an American English speaker, there are languages, such as Japanese, which do not utilize [l] and [ɹ], but use instead a phoneme [ɾ], usually described as a 'flap'. Monolingual speakers of Japanese perceive instances of American [l] and [ɹ] as both highly similar to the phoneme used in their own language, and tend to have difficulty discriminating American [l] from [ɹ]. Similarly, although speakers of Hindi distinguish between a [t]-like sound produced with the tongue-tip up against the front teeth and a [t]-like sound produced with the tongue 'retroflexed' (curled up and slightly back, hitting against the alveolar ridge, the hard ridge behind the front teeth), American English listeners listening to the Hindi phonemes will tend to label both as representing instances of the English phoneme [t].

Despite its seemingly simple functional definition, there are serious drawbacks to trying to use the phoneme as the basis for a perceptual unit of analysis. In particular, during fluent speech, the articulators can be shown to move into position for an upcoming sound (or sounds) even while the current sound is still being produced. This causes the acoustic cues to particular phonemes to be 'overlapped' in time with each other, interacting in complex ways for different combinations of phonemes. This interaction between successive phonemes is usually referred to as co-articulation. Co-articulation across adjacent phonemes has motivated researchers to propose larger units of analysis that subsume more of the surrounding acoustic context. One promising candidate is the syllable. A syllable is typically defined as consisting minimally of at least a vowel, with, optionally, one or more immediately preceding or following consonants. Adopting the syllable as the fundamental unit of analysis does not, however, fully eliminate the contextual problems faced by the phoneme since co-articulation often extends across multiple syllables and even across multiple words in an utterance. The prevalence of co-articulation is one of the most formidable challenges faced by speech researchers.

THE MOTOR THEORY OF SPEECH PERCEPTION

In the early 1960s, after working extensively on trying to develop a fluent and naturalistic-sounding reading machine for the blind, experimental psychologist Alvin Liberman (1918–2000) and his colleagues at Haskins Laboratories began to seriously consider the inherent difficulties in building a complementary machine able to recognize speech from just a representation of acoustic information. Would the machine lack crucial information that a real human listener possesses? The relative speed and accuracy with which humans recognize speech, despite the lack of invariant cues in the acoustic signal, led Liberman to speculate that listeners achieve perceptual constancy in speech through reference to their own internalized knowledge about how speech sounds are articulated. For example, Liberman and colleagues pointed out that the acoustic cues for a [d] at the beginning of a syllable differ depending on the vowel that follows due to co-articulation, yet a description of [d] as 'a constriction at the alveolar ridge' is applicable in all vowel contexts. This 'motor theory of speech perception' hypothesized that listeners use knowledge about the effects of

co-articulation on their own productions to recognize other people's attempts to produce the various phonemes.

Electromyographic techniques revealed, however, that patterns of motor activity at the level of the articulator muscles and joints map no less variably to perception than do acoustic patterns. Both within and across individual speakers, articulatory variability was shown to be quite high even when the perceptual result was relatively stable. Evidence of the development of normal speech perception skills by humans unable since birth to control their speech articulators was also problematic. Motor theory therefore underwent revision, such that listeners were assumed to extract abstract 'higher-level' motor commands for intended phonetic 'gestures' rather than information about low-level, externally observable motor activity. With this change, however, motor theory's primary hypothesis became extremely difficult, if not impossible, to test. Researchers, however, continue to explore the idea that the role of the speech perceiver as also evolutionarily developed to be a speech producer should contribute a potentially valuable source of information.

Motor theory is historically important for having generated a wealth of experiments focused on acoustic cues. Numerous studies were conducted showing that many different patterns of acoustic information could yield the same phonemic percept, and also that the same acoustic pattern could be perceived differently given different surrounding contexts. Even though these studies were successful at debunking the notion of invariant acoustic cues to phoneme identity, they failed to yield truly convincing positive evidence that speech perception is based instead on recognition of articulatory gestures.

Motor theory also gained prominence as a theory of speech perception which argued for a strongly modular organization of cortical speech-processing mechanisms. In a number of ingenious experiments, Liberman and his colleagues found evidence that the human brain is predisposed to interpret speech and speech-like stimuli differently from other auditory stimuli. A variety of experimental situations were constructed to demonstrate the 'special' processing of speech signals, 'duplex perception' and 'categorical perception' being the most well-known examples.

THE DIRECT REALIST APPROACH

The direct realist approach to perception has a long history in the sub-area of philosophy known as epistemology. Epistemology is concerned with how humans acquire knowledge and arrive at beliefs about the world. Philosophers define 'direct realism' as the claim that the world has properties that exist regardless of our ability to perceive them, and that what we do perceive is objective – that we perceive the world more or less as it exists. Direct realism contrasts with other epistemological claims such as 'phenomenalism' (which argues that we cannot access the physical world directly since all information about our environment is mediated via our sensory and cognitive systems), and 'representationalism' (which says that physical objects cannot meaningfully be said to exist outside of how they are perceived).

The direct realist approach to speech perception was strongly influenced by the writings of the sensory psychologist James J. Gibson (1904–1979). Although Gibson discussed direct perception largely in terms of vision and haptic (kinesthesis touch) perception, Carol Fowler and her colleagues at Haskins Laboratories have been widely recognized since the 1980s for their efforts in applying Gibson's ideas to the study of speech perception.

In his work, Gibson argued that objects in the world have affordances, which he defined as characteristics that provide the perceiver the opportunity to interact with the object in a specific way. For example, a flat door panel affords pushing but does not afford pulling in order to open it. Furthermore, Gibson suggested that stimulus information arrives at the sensory receptors already in a form that permits 'immediate' recognition of these affordances – that is, that no further processing of this information is required for perception, particularly, no inference-making, either conscious or unconscious.

Direct realists propose that speech is directly perceived like any other meaningful sound in one's environment, in terms of identifying the physical object(s) that imparted structure to the intervening medium (usually air). Listeners do not consciously perceive the intermediate acoustic structure, but rather the source responsible for it. Perceiving speech is therefore the act of identifying vocal tract configurations that could have given rise to particular air pressure patterns, that is, recognizing affordances such as 'could have been articulated by constricting the lips'. Resembling motor theory, the perceived events in direct realism are therefore vocal tract gestures rather than acoustic cues. According to the direct realist approach, these abstract gestures are not tied exclusively to the auditory modality, but can be perceived via a variety of intervening media, for example,

cutaneous sensation in the case of 'Tadoma', a method of identifying words from touching a talker's moving lips and jaw.

The most radical implication of a direct realist's view of speech perception is the idea that investigating stages of information processing during perception is irrelevant to the scientific endeavor of speech scientists. Although it seems possible that knowledge of some physical affordances may be innate or instinctual, recognizing other affordances would seem to involve some degree of prior experience or learning. The role of learning and memory in the development and function of direct perception was not well explicated by Gibson. If these cognitive factors can be shown to play a role in perception, it is not clear if the 'directness' of perception is preserved. The time course over which spoken language emerges as a skill in young children makes it difficult to understand how perception of speech can be purely direct.

What types of empirical research are suggested by the direct realist approach? One area of work has examined the speed and accuracy with which listeners can repeat back spoken syllables; results suggest that speech can be repeated more quickly than a model with many intermediate steps of processing between perception and production would predict. Other studies have examined how listeners are influenced by tactile information about a speaker's articulations – especially when the two sources of information (auditory and tactile) are deliberately designed to conflict with each other. Fowler interprets the finding that phoneme identification can be influenced by stimulation via the relatively unfamiliar intervening medium of skin deformation as evidence of speech perception taking place in terms of identifying articulatory gestures, not recognition of patterns in the intervening medium.

EXEMPLAR MODELS OF SPEECH PERCEPTION AND SPOKEN WORD RECOGNITION

An active area of current research addresses the question of the degree to which memories for speech experiences are stored *intact* in memory. The two extreme positions on this issue are usually referred to as the prototype hypothesis and the exemplar hypothesis. Prototype theory claims that a representation of an average instance of a given lexical item emerges over multiple exposures and that the details of each acoustic signal (which vary from instance to instance) are *averaged out* of the stored representation and are difficult or impossible to retrieve. That is, as a listener experiences situations in which the word 'dog' is uttered multiple times or by different talkers, a representation that is an 'average' instance of 'dog' is created in the listener's brain. The exemplar hypothesis, in contrast, argues that a detailed representation is stored for each and every instance of a word that has been encountered, with no information loss occuring during the storage process.

For many years, the assumption that storage space in the brain was relatively limited favored the notion that memory could be used most efficiently by storing a prototype (perhaps formed by averaging each new instance with the existing prototype), instead of exemplars. Recent developments suggest, however, that the memory capacity of the brain was significantly underestimated, and that storage of acoustic details may actually make listeners' speech perception skills more robust and flexible. In order to test the exemplar hypothesis, experiments have been conducted to see if listeners show evidence of storing exemplars. A number of studies have demonstrated retention of instance-specific information over time, even though listeners may not necessarily be able to consciously report memory for exemplar details. Evidence for instance-specific memory is often indirect (for example, in terms of relative reaction times in word identification tasks). Note that a purely 'exemplar' view is probably not correct either since there is also evidence that people tend to store or generate information about prototypes. Exemplar models make interesting and specific predictions about how spoken word recognition ability unfolds as a function of experience. Two areas in which these claims are actively being tested are child language acquisition and second/foreign language learning.

CONCLUDING REMARKS

Further efforts are needed to integrate what is known about speech perception at the level of the phoneme into accounts of spoken word recognition and sentence processing. Researchers also need to reevaluate carefully the role of acoustic variability in light of the insights offered by exemplar models of speech perception. Continued empirical study of how acoustic cues are used in different languages, the development of these perceptual skills in infants, and the effects of different forms of hearing impairment on speech sound discrimination, will be crucial to advancing our current understanding of speech perception and spoken language processing.

Further Reading

Borden GJ, Harris KS and Raphael LJ (1994) *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, 3rd edn. Baltimore, MD: Williams & Wilkins.

Denes PD and Pinson EN (1993) *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd edn. New York, NY: WH Freeman.

Fowler CA (1986) An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics* **14**: 3–28.

Fowler CA (1996) Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America* **99**: 1730–1741.

Johnson K (1997) *Acoustic and Auditory Phonetics*. Cambridge, MA: Blackwell.

Kent RD (1997) *The Speech Sciences*. San Diego, CA: Singular Publishing Group.

Ladefoged P (1993) *A Course in Phonetics*, 3rd edn. Fort Worth, TX: Harcourt Brace.

Lass NJ (1996) *Principles of Experimental Phonetics*. St Louis, MO: Mosby.

Liberman AM (1996) *Speech: A Special Code*. Cambridge, MA: MIT Press.

Pisoni DB (1997) Some thoughts on 'normalization' in speech perception. In: Johnson K and Mullennix JW (eds) *Talker Variability in Speech Processing*. San Diego, CA: Academic Press.

Speech Recognition, Automatic

Advanced article

Hynek Hermansky, Oregon Graduate Institute of Science and Technology, Beaverton, Oregon, USA
Nelson Morgan, International Computer Science Institute and University of California, Berkeley, California, USA

CONTENTS

Automatic speech recognition is the attempt to use a machine to derive the linguistic message from a speech signal.

INTRODUCTION

The linguistic message of a speech signal originates in a speaker's mind. The speaker converts it to speech sounds, considering the intended recipient of the message and any feedback the recipient might be providing. Human listeners can generally determine a functional equivalent to the speaker's intended message quite effortlessly. However, this task represents a difficult challenge for a machine. (*See* **Speech Perception and Recognition, Theories and Models of**)

Automatic speech recognition (ASR) in its unrestricted sense would probably require emulation of human intelligence. While some believe this can be achieved in the foreseeable future, many others believe that it is still far away, despite apparent advances since the criticism of this field by John Pierce (1969).

The general problem of ASR can be seen as that of finding a mapping from the space of all allowable acoustic signals to that of all meaningful messages. Each space grows with the length of the message. Even when the length of the message is bounded, the mapping is many-to-one, since the speech signal carries information from many sources and many different speech utterances may carry the same linguistic message. The message may be produced by different speakers; a speaker may speak slower or faster; the acoustic environment may change; the health or emotions of the speaker may be reflected in changes of voice; and so on. Thus, ASR must deal with significant nonlinguistic variability in the data. We have limited understanding of how this is done by human listeners.

ASR typically incorporates a model of the speech production process. Such a model should generate all possible acoustic sequences for all legal linguistic messages. The model is trained using a large number of different acoustic signals, which are associated with their corresponding word sequences. Recognition may then be done by testing all possible word-sequence models to find the one that would produce the acoustic signal best matching the actual signal being recognized.

The difficulty of ASR varies greatly with the complexity of the task, which may range from recognition of a few anticipated voice commands uttered by a single speaker, to transcribing conversational speech of an arbitrary speaker in realistic noisy environments.

BASIC PRINCIPLES

The goal of ASR is to find the most likely sequence of symbols representing the linguistic message in the speech signal, given the acoustic data. In essentially all current ASR systems, an intermediate goal for the recognizer is to determine the likelihood of possible sequences of speech-sound states, where these states typically correspond to parts of phonemes. (*See* **Phonology**; **Phonetics**)

The input to the recognition process is a digitized version of the electric signal from the microphone that represents changes in acoustic pressure at some distance from the mouth of the speaker of the message. The signal typically undergoes further information-reducing transformations before being used in the evaluation of the likelihood of each

possible sub-phonetic state sequence, that in turn determines the likelihood of each possible word sequence.

In general, the outcome of the recognition process is a list of hypothesized word sequences, where 'words' here mean some tokens from a predefined recognition 'lexicon'. The hypothesis is generated by a decision process that aims to produce a 'most likely' word sequence w given the input signal x. This can be expressed mathematically by the so-called 'maximum *a posteriori*' decision rule (using the concept of conditional probability) as

$$w = \arg\max_i P(w_i|x) \tag{1}$$

where w_i represents the i^{th} word sequence from the lexicon and the conditional probability is evaluated over all sequences from the lexicon.

The word sequences w_i are not represented by the acoustic signals but are actually represented by their models $M(w_i)$:

$$w = \arg\max_i (P(M(w_i)|x)) \tag{2}$$

where w is a sequence of symbols representing the linguistic message in speech, $M(w_i)$ is a model of the sequence w_i, P is the posterior probability of the model given the acoustic input, and the maximum is evaluated over all models.

Bayes' rule:

$$P(M(w_i)|x) = \frac{P(x|M(w_i))\,P(M(w_i))}{P(x)} \tag{3}$$

can be applied. Since the probability $P(x)$ is the same for all choices of i, it can be ignored during recognition, and equation 1 then simplifies:

$$\begin{aligned} w &= \arg\max_i (P(M(w_i)|x)) \\ &= \arg\max_i (P(x|M(w_i))P(M(w_i)) \end{aligned} \tag{4}$$

The probability distributions are in practice unknown, and must be estimated, using parameters that are trained on speech samples, in order to solve equation 4. So strictly speaking, each probability is also conditioned on the dataset from which its estimators were trained. Usually, the acoustic estimators of $P(x|M(w_i))$ are trained from speech data and the estimators for language probabilities $P(M(w_i))$ are trained primarily from written text. (*See* **Machine Learning**)

The acoustic modeling is done as follows. Acoustic input is represented by a series of vectors, each of which represents a short segment of the acoustic signal (typically 10 to 20 milliseconds). The feature vectors usually represent spectral properties of these short segments of the speech signal, typically transformed further into some other set of variables (most typically 'cepstra', which are the discrete fourier transform of the log spectra). The analytic methods that generate these features are designed to preserve relevant linguistic information while alleviating unwanted variability due to nonlinguistic sources (such as microphone sensitivity). The data x are then represented by a sequence of vectors v.

If all word sequences w_i in the lexicon were of the same physical length, and enough examples of each sentence were available for estimating the distribution of acoustic data for the sequence, then models $M(w_i)$ could be derived by estimating probability distributions of the training data for all anticipated word sequences. Then the probability $P(v|M(w_i))$ of the unknown sequence v could be directly estimated using pattern classification techniques. However, such models would take no advantage of the temporal structure of the sentence.

In practice, word sequences differ in length because of the different speaking styles of individual talkers. Furthermore, there are usually an insufficient number of examples of complete utterances to train the models in this way. Therefore, the acoustic model is broken into subsets of smaller submodels m representing phrases, words, phonemes, and most typically smaller units such as parts of phonemes. These smaller units (which we have already referred to as 'states') may be shared among many different words.

Assuming the statistical (conditional) independence of the submodels, the overall probability $P(x|M(w_i))$ can then be expressed in terms of local probabilities $P(v|m)$. Efficient heuristic search techniques for solving equation 4 from local probabilities $P(v|m)$ have been developed. Temporal variability is typically handled by dynamic time alignment of the incoming speech representation with the stored models. This time alignment attempts to find the path of the best match by local comparisons between possible model state sequences to find the most probable representation for the incoming speech vectors.

The sub-word units (e.g. phonemes) do not need to be recognized categorically before the decisions are made. They merely represent units that are shared between different words in the lexicon and whose choice is dictated by considering the required lexicon, its distinctive symbolic representation by the chosen phonemes, and expected variations in pronunciations of the words in the lexicon. The hypothesized word sequences are

formed by a heuristic search that in principle considers the probabilities of the states in all the phonemes in the vocabulary. The hypothesized word sequence is chosen via a competition between various paths that are evaluated at the end of the acoustic input. The evaluation is derived from equation 4, consisting of the product of all local acoustic likelihoods corresponding to a state sequence, combined with appropriate language probabilities for each word sequence. The combination of local scores into scores for the entire utterance assumes conditional independence (independence when conditioned on the state variable) and the Markov property (that for both words and states, only a limited history need be considered to represent the probability of moving from one to the next).

In practice, the least viable hypotheses are pruned from the search to save on computation, so not all possible hypotheses are actually considered. 'Search errors' (caused by dropping seemingly unlikely hypotheses) are fairly uncommon, however.

The difficulty of an ASR task derives partly from the acoustic confusability of the words (for instance, the digits in English are much less similar to one another than the spoken alphabet, which contains pairs such as 'E' and 'B'), but also from the linguistic unpredictability of words in context, since a strongly predictable word permits use of a language model that can counteract acoustic ambiguity. For instance, if a language model only permits the words 'yes' or 'no' at a particular point in a dialog, the similarity between 'no' and 'go' will not be a concern. The difficulty of an ASR task in terms of this linguistic unpredictability is typically measured by the 'perplexity', which is equal to the number of equiprobable words for which the prediction is as difficult as for the original task (Jelinek, 1997).

Acoustic inputs are typically assumed to be limited to speech utterances that are anticipated during the current ASR task (closed-set recognition). This of course limits the difficulty of the task. On the other hand, this makes ASR vulnerable to inputs from outside the closed set (such as words outside the vocabulary, silences between words, breath noises, or noise in general). Words outside the vocabulary are sometimes handled by a so-called 'garbage model', which may allow for matching of these unanticipated acoustic inputs.

The language model generates a probability $P(M(w_i))$ for each given sequence of words w_i. As mentioned above, this probability is constructed by finding the product of individual conditional probabilities under the model of words given 'word histories', where the 'word history' typically refers to a few words prior to the current word. The probability of a group of n words can be approximated by the product of terms that are each dependent on $n-1$ previous words, using the Markov assumption. The resulting estimate is called an 'n-gram probability'. A simple example of this approach is the approximation of the word-sequence probability by the product of bigram probabilities:

$$\begin{aligned} &P(w(n),\ w(n-1), \ldots w(1)) \\ &\sim P(w(n)|w(n-1))\, P(w(n-1)|w(n-2)) \\ &\ldots P(w(1)) \end{aligned} \tag{5}$$

In large vocabulary-recognition tasks, the language model can be somewhat more complicated, including longer n-grams for common sequences, word-class probabilities, etc., and can incorporate either interpolations or switches between the different submodels. In smaller tasks, the language model can simply consist of lists of words that are allowed to follow each other.

The basic plan of an ASR system is summarized in Figure 1.

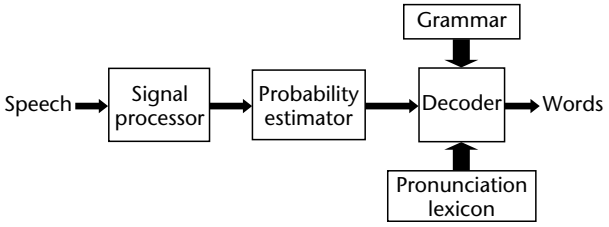


Figure 1. Overall plan of a speech recognition system.

ACOUSTIC PROCESSING

Speech is first converted to an electric signal that represents acoustic pressure changes at the microphone. It is then amplified, filtered, sampled, and digitized to facilitate further processing. 'Short-term analysis' converts consecutive overlapping short segments of the signal (about 20 ms long) into feature vectors (e.g. in steps of 10 ms). Some information-rate reduction typically occurs at this stage, with the primary goal of reducing the non-linguistic variability from the speech signal while retaining characteristics that are useful for distinguishing between different speech sounds.

Feature design may incorporate some of the characteristics of human auditory perception, since the effect of speech on the listener may be a primary cause of the linguistic message being reconstructed by the listener (Hermansky, 1998).

While this view is not universally accepted, most current ASR systems incorporate methods that emulate some of the basic properties of human hearing.

PATTERN CLASSIFICATION

At this stage, the feature vectors are compared (typically in a probabilistic sense) to models that represent precomputed representations of speech. The best match of the sequence of incoming feature vectors with the acoustic and language models provides an estimate of the linguistic content of the speech segment. The models may be in the form of templates of feature-vector sequences in the case of the simplest ASR systems (more common in the 1970s and 1980s), while in modern systems they may be stochastic models for words or phonemes. As already mentioned, the designer typically decides on the architecture of the models. The lexicon of the application is also formed at this stage, incorporating common pronunciation variations of the words that are modeled. Similarly, the language model is chosen. In both cases, the final forms are determined by some combination of designer decisions and training from data. The model parameters are learned by training from speech data for which the word sequences are known. The phonetic sequences are typically not known, even for the training data, so having good pronunciation models is essential for associating phonetic sequences with the data. During training, the model parameters are learned, taking into account the uncertainty in these sequence labels. (*See* **Pattern Recognition, Statistical**)

HIDDEN MARKOV MODELS OF SPEECH

Most ASR systems incorporate the kind of speech representations known as 'hidden Markov models' (HMM). These are stochastic models with unknown (hidden) variables for which probability density functions are learned from training data, and for which the Markov property (ignoring all but recent history in estimating probabilities) is assumed. Each model consists of a sequence of states, the identities of which are hidden from the observer. Each state has certain stochastic properties for transition to other states and for the emission of feature vectors. The HMM associates identical stochastic properties to all feature vectors emitted by a given state, in principle with no statistical dependence between the vectors; however, the learning process modifies parameters based on all of the states, so that in some sense there is a dependence on all of the speech and on all models within the statistics for each frame. (*See* **Markov Decision Processes, Learning of**)

SUB-WORD UNITS

Small-vocabulary ASR systems may use models for whole words or even complete utterances. However, it is hard, if not impossible, to train models for all words in large-vocabulary systems, since many examples of each are required in order for the system to effectively learn the model parameters. Consequently, there is usually a need for sub-word-based models, which can be dynamically assigned to different words in the vocabulary.

The most often-used sub-word units are phonemes or phoneme-like units. A sequence of these units may be compared to 'beads on a string': the feature emission statistics are derived from a single probability density function. In order to better approximate the dynamics of speech production, HMMs for ASR are usually composed of multiple (typically two to five) states for each phonemic sound unit. Furthermore, the phoneme classes typically are further subdivided into different contextual classes (e.g. with particular neighbors to the left and right). This is done in order to better represent the wide variability in sounds that could be associated with any particular phoneme, for which a primary cause is the coarticulation between neighboring speech sounds due to the inertia in the vocal apparatus. Since using all allowable contexts would lead to a large number of context-dependent phoneme classes, the context phonemes are often clustered into larger classes, thus keeping the number of sub-word units reasonably small. In large systems, acoustic probabilities from differing contexts and degrees of clustering are integrated together, much as in the language models. (*See* **Word Recognition**)

Distributions of the feature vectors associated with each state of the acoustic HMM are usually assumed to be emitted according to a multivariate Gaussian mixture model (GMM), which consists of a weighted sum of Gaussians. All the means, covariances (typically variances only) and weights are estimated from the training data using iterative procedures that are based on the learning method called 'expectation maximization'.

During training, each GMM is presented with examples of feature vectors along with class membership (or class distribution) for each vector. This allows for estimation of the distribution of feature vectors within each class. In simpler systems, there

is one GMM for each sub-word class; in more complex systems, there is some parameter sharing between classes. During the classification, an unknown feature vector is compared to each state of the GMMs. This process generates a probability for each state of the GMM to have produced the feature vector. The resulting set of probabilities is used in the search to find the most likely sequence of models that explains the observed sequence of feature vectors.

To simplify the estimation and to reduce the number of parameters that need to be estimated from the data, the elements of the feature vector are assumed to be uncorrelated with one another. This assumption allows the use of diagonal covariance matrices in the model (i.e. variances only), which reduces the number of parameters of the model significantly.

GMMs are not the only way to represent distributions for feature vectors. Some systems use 'codebooks' that hold tables of probabilities for each quantized feature vector and each phoneme type. A very different approach is based on a system that computes the posterior probability $P(m|v)$, which can then be transformed via Bayes' rule to a scaled version of the more typical likelihood $P(v|m)$. One such structure is the so-called 'multi-layer perceptron' (MLP) neural network, which can be hybridized with an HMM system for ASR (Morgan and Bourlard, 1995). The MLP is an inherently nonlinear classifier, and it often appears to be more suited to unusual feature distributions. The hybrid system is also inherently discriminative, in that the learning process has the goal of choosing parameters to distinguish between target and non-target classes. Discriminative approaches can also be applied to GMM-based systems, but this greatly increases the complexity, and typically yields only moderate gains. Discriminative training potentially allows the system to focus on class boundaries rather than on centers of distributions. (*See* **Connectionism**; **Connectionist Implementationalism and Hybrid Systems**; **Connectionist Systems, Learning in**)

The phoneme is an idealized unit that does not always correspond to real acoustic events observed in the data. Efforts continue to derive alternative units directly from the data.

PROBLEMS WITH THE STOCHASTIC APPROACH

Speech is used for communication. A fundamental principle of communication theory is that the most likely events carry the least information. It is therefore unavoidable that the distributions of events both in the acoustic space and in the language model are long-tailed, and that the most information-rich events occur in the tails of distributions of the training data. In order to get reliable estimates of these rare but important events using stochastic training, very large training databases would be required. But in practical applications of the technology, the amount of training data is always limited, so it seems unavoidable that the most important rare events are represented in the stochastic model least accurately. For this reason, much of the effort in the field has focused on ways to smooth statistics from rare but pertinent data with statistics from larger sources of data that are less specific to the ultimate task.

In general, the stochastic approach to ASR favors mean trends in the training data. For instance, a database that is 80% male and 20% female will tend to generate parameters that will yield much better recognition for males than for females. ASR applied in conditions that were not well represented in the training data is likely to perform poorly, unless the test set variability introduced by the varying conditions is handled by some other means. The required amount of training data also increases with the increasing complexity of the model.

On the other hand, models of the speech communication process that are both simple and accurate do not yet exist. It is an open question whether it is only a matter of time until such a powerful model is developed, or whether there are strong reasons for looking for alternatives to the current stochastic style of approach to ASR.

CURRENT RESEARCH

Feature Extraction

Features typically represent some aspects of the short-term spectral envelope of the speech segments, processed by very simple models of human auditory perception that emulate unequal spectral resolution of human hearing (Bridle and Brown, 1974; Mermelstein, 1976; Hermansky, 1990). The spectral envelopes are easily corrupted by external factors, and various schemes are applied during feature extraction to alleviate these effects. Among the most often-used normalization techniques are: subtracting the mean from the time trajectory of each feature and normalizing the variances of utterances; and filtering the time trajectories of features to suppress slow and fast

changes that may be outside the typical rates of change for speech components (Hermansky and Morgan, 1994). These so-called static features, which represent the instantaneous shape of the spectral envelope, are often augmented with dynamic features such as time derivatives, representing the time evolution of the feature trajectory in the vicinity of the static vector (Furui, 1986). A promising approach is to optimize the feature-extraction module by learning discriminant transformations from the speech data (Hermansky, 1998). Interestingly, such an optimization appears to yield feature extraction techniques with properties that are consistent with those of human hearing. In general, feature-extraction methods are now beginning to incorporate time contexts significantly larger than 20 ms.

Adaptation

Some ASR systems that can take advantage of learning from the target speaker (particularly for offline processing) address the anticipated variability in the data by adaptation of the statistical parameters (Woodland *et al*., 1996). The adaptation can be either supervised or unsupervised. In supervised adaptation, the speaker is asked to produce speech with known linguistic content and the ASR system parameters are adjusted to best match the appropriate sequence of models. In unsupervised adaptation, the learning is done during recognition of speech with an unknown message, using the unadapted recognizer itself to give a first estimate of what was said. Since there are many applications for which a single speaker may provide much of the input, the adaptation may also deal with predictable changes in the pronunciation of a particular speaker.

Incorporating Multiple Subsystems

Requirements for the amount of training data grow with the size and the complexity of the stochastic model. One way to simplify the model (or, alternatively, to improve performance for a model of a given size) is to break it into several parts. If these different partial models have different properties, they can potentially complement one another to improve the system's robustness to variability during testing. Different parts of the spectrum may be used to develop separate probability estimators. Alternatively, feature vectors with differing temporal properties may be used together. Discriminant systems can be used to generate feature vectors for standard GMM-based systems classification (Hermansky *et al*., 2000).

Finally, many researchers are working to combine word-sequence hypotheses from several complete recognizers.

In each of these approaches, the performance seems to improve if the subsystems make, at least to some extent, different errors.

Building More Prior Knowledge into an ASR System

Current HMM-based systems appear to make only minimal assumptions about the nature of speech. In fact, researchers have shown that such a system can be used for handwriting recognition essentially by changing only the feature-extraction module. Certainly the feature-extraction module is an obvious place to incorporate more information about the speech signal, and this has usually been done by emulating dominant properties of human speech perception. Another approach involves applying dynamic stochastic models (e.g. Ghitza and Sondhi, 1993). Segmental models have also been developed, in which statistics associated with a complete phonetic segment are incorporated, as opposed to only those associated with a single 20 ms feature vector (e.g. Ostendorf *et al*., 1992).

Finally, representations of semantic structure, long-distance dependencies, and dialog state models are likely in the future to be important sources of predictive information for language models. (*See* **Natural Language Processing**; **Natural Language Processing, Statistical Approaches to**; **Prosody**)

References

Bridle JS and Brown MD (1974) An experimental automatic word recognition system. JSRU Report No. 1003. Ruislip, UK: Joint Speech Research Unit.

Furui S (1986) Speaker independent isolated word recognizer using dynamic features of speech spectrum. *IEEE Transactions on Audio, Speech and Signal Processing* **34**: 52–59.

Ghitza O and Sondhi MM (1993) Hidden Markov models with templates as non-stationary states: an application to speech recognition. *Computer Speech and Language* **2**: 101–119.

Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* **87**(4): 1738–1752.

Hermansky H (1998) Should recognizers have ears? *Speech Communication* **25**: 3–28.

Hermansky H, Ellis DPW and Sharma S (2000) Connectionist feature extraction for conventional HMM

systems. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing 2000*, vol. III, pp. 1635–1638. Istanbul, Turkey.
Hermansky H and Morgan N (1994) RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* **2**(4): 587–589.
Jelinek F (1997) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.
Mermelstein P (1976) Distance measures for speech recognition, psychological and instrument. In: Chen RCH (ed.) *Pattern Recognition and Artificial Intelligence*, pp. 374–388. New York, NY: Academic Press.
Morgan N and Bourlard H (1995) Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach. *Signal Processing Magazine* **12**(3): 25–42.
Ostendorf M, Bechwati I and Kimball O (1992) Context modeling with the stochastic segment model. In: *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, San Francisco, CA, pp. 389–392.
Pierce J (1969) Whither speech recognition? *Journal of the Acoustical Society of America* **46**: 1049–1051.
Woodland PC, Pye D and Gales MJF (1996) Iterative unsupervised adaptation using maximum likelihood linear regression. In: *Proceedings of the International Conference on Speech and Language Processing*, pp. 1133–1136.

Further Reading

Furui S (2001) *Digital Speech Processing, Synthesis and Recognition*. New York, NY: Marcel Dekker.
Gold B and Morgan N (1999) *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York, NY: John Wiley & Sons.
Huang X, Acero A, Hon HW and Reddy R (2001) *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Englewood Cliffs, NJ: Prentice Hall.
Jelinek F (1998) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.
Rabiner L and Juang BH (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

Statistical Methods

Advanced article

Steven Abney, AT & T Laboratories – Research, Florham Park, New Jersey, USA

CONTENTS

'Statistical methods' refers here specifically to statistical methods in computational linguistics. This represents a new body of practice in computational linguistics that has become standard since the 1990s.

INTRODUCTION

Since the early 1990s a new body of practice has become standard in computational linguistics. It is known variously as *corpus-based*, *empirical*, or *statistical methods of language analysis*, most common being the simple rubric *statistical methods*. Present-day computational linguistics differs from 'traditional' computational linguistics in the pervasiveness of probabilities in its theoretical models, the centrality of large data collections, including *text corpora* and *treebanks*, and the emphasis on rigorous empirical evaluation. The change in computational linguistics is part of a larger shift to statistical methods in computer science, particularly in artificial intelligence, pattern recognition, speech recognition, and machine learning. (*See* **Information Theory**)

THE RE-EMERGENCE OF EMPIRICAL LINGUISTICS

American Structuralism

The statistical paradigm has strong precedents in American structural linguistics. One of the major aims in structuralism was the development of procedures for taking a representative corpus of raw language data and determining the elements of which it is composed (the sounds and words of the language) and the conditions on their distribution. The motivation was methodological and philosophical: one wished to report only what was evident in the data, thereby avoiding the speculation and subjectivity that had been characteristic of earlier 'philosophical' linguistics. Bloomfield famously wrote that 'the only useful generalizations about language are inductive generalizations' (Bloomfield, 1933, p. 20). Only the observable regularities in the data were of concern. Unobservables – in particular, meaning – had no place in structuralist descriptions.

Two important classes of linguistic elements are phonemes and morphemes. In the structuralist view, they are not abstract postulates, but rather, features of the data. Structuralist procedures for identifying phonemes and morphemes (and other aspects of structure) are generally known as *discovery procedures*, but in a real sense they are not so much concerned with *discovering* elements of structure as with *defining* them. A phoneme (for example) is defined to be what a given procedure returns when applied to the data. Faced with two alternative phonemicizations, a structuralist does not ask which one is correct. Definitions cannot be right or wrong. In the words of Harris, 'they differ not in validity but in their usefulness for one purpose or another (e.g. for teaching the language, for describing its structure, for comparing it with genetically related languages)' (Harris, 1951, p. 9, n. 8).

Current computational linguistics takes a similar stance on the question of truth. Unlike structuralism, it does not eschew 'deep' or 'hidden' models – witness the discussion of stochastic grammars below – but its concern is utility rather than Platonic truth, and it insists on rigorous evaluation of model utility against a quantifiable measure of success at some task.

Current computational linguistics also follows structuralism in its interest in mechanically inducing from a corpus the elements of the language and the conditions on their arrangement. There are obvious parallels between some of the structuralist procedures and newer statistical methods. For example, one of the procedures that Harris used to segment a corpus into morphemes is the following. Consider an utterance, for example,

/hiyzklev^{e}r/ 'he's clever'. Count the number of phonemes that could have occurred after /h/. That is, among all utterances in the corpus beginning with /h/, how many distinct segments appear in the second position? In Harris's estimate, there are nine (the English vowels and semi-vowels). After /hi/, there are 14 possibilities; after /hiy/, 29; after /hiyz/, 29; and after /hiyzk/, 11. Morpheme boundaries occur where the number of possibilities is highest, the intuition being that the constraints on succession are more stringent within a word than between words. Hence morpheme boundaries are defined to occur after /hiy/ and /hiyz/, but not after /h/ or /hiyzk/.

If we consider *probabilities* of phonemes instead of just *possibilities*, a natural analogue of Harris's proposal is to measure how much phoneme probabilities are affected by the context:

$$p(\text{phoneme} \mid \text{context}) \,/\, p(\text{phoneme}) \qquad (1)$$

The measure (1) is precisely the measure proposed by Stolz (1965) in an experiment to induce phrase boundaries. In information theory, the average of the logarithm of (1) is called *mutual information*; it is a key measure of coherence in modern statistical methods, and is commonly used to induce phrases.

In a similar way, the structuralist use of substitutability to define natural classes of elements has modern parallels. In structuralism, the class of nouns is defined as a class of elements that appear in similar contexts. Information theory provides a mathematically well-founded measure of substitutability: two elements are intersubstitutable if the *divergence* of their distributions is small. The divergence between distributions p and q is the average (with respect to p) of the log of $p(x)/q(x)$. It has been used for constructing classes of similarly distributed elements (Finch, 1993).

Language Models

As the previous discussion suggests, information theory is an important tool for putting structural induction procedures on a firmer mathematical foundation. Information theory was motivated by the problem of transmitting information over a *noisy channel*. To transmit text (for example) over a telegraph wire with maximum efficiency and minimum error, it is necessary to identify the redundancies, which is to say, the regularities, in the text. This is precisely the task that the structuralists had set for themselves (though with a very different motivation).

In his seminal work on information theory, Shannon introduced the following problem, which has become known as the *Shannon Game* (Shannon, 1951). Take a random text and uncover it one element (e.g. one letter) at a time. At each point, the task is to predict the next element; it is not revealed until you guess correctly. The quantity used to measure difficulty in guessing is effective vocabulary size or *perplexity*: this is the vocabulary size that would cause a random guesser to make the same number of mistakes as you make. Your estimate of the text's perplexity is a measure of how good you are as a guesser. But Shannon also showed that there is a limit on how good any guesser can be. This *Shannon limit* is the inherent perplexity of the text (Shannon, 1948).

Algorithms that play the Shannon Game are known as *language models*. (To be precise, a language model is a probability distribution over all possible sequences of elements, and its ability to play the Shannon Game, as measured by its perplexity, is the measure of its quality as a language model.) A simple family of language models is the family of n-th order *Markov models*, which approximate the conditional probability $p(x_t|x_1 \ldots x_{t-1})$ of an element x_t given the corpus up to time t as $p(x_t \mid x_{t-(n-1)}, \ldots, x_{t-1})$, the probability of x_t given the preceding $n-1$ elements. Shannon showed that Markov models of increasing order converge to the Shannon limit: by choosing n large enough, the true distribution over language elements can be approximated arbitrarily closely.

However, Chomsky criticized Markov models (Chomsky, 1956). First, he emphasized that any given Markov model accounts for dependencies only up to a certain distance, leaving a residue of longer-distance dependencies not captured. Second, he pointed out that simple frequency counts are not adequate for estimating any but the lowest-order Markov models.

Concerning the first criticism, Markov models are mathematically useful because of their extreme simplicity, and they are surprisingly effective in practice, but they are obviously inadequate for many purposes, particularly for representing language semantics. (Unlike structuralism, modern computational linguistics is very much concerned with language meaning.) Soon after Chomsky defined context-free grammars, stochastic versions were explored, and have since been well developed; they are discussed below.

The second of Chomsky's criticisms is in part addressed by moving to more expressive grammars, and in part it is a technical issue concerning

estimation of model parameters in the face of *sparse data*. This has been a major topic in speech recognition, and very sophisticated *smoothing* techniques have been developed to address it.

Causes of the Revival of Statistical Methods

In the late 1970s, Shannon's noisy channel model, and Markov models in particular, were applied to speech recognition by Jelinek and his colleagues. The speaker is approximated by a Markov model, and the channel includes both the conversion of words into sounds and the transmission of the sounds to the hearer. The setting as a whole is approximated by a Hidden Markov Model (HMM), a generalization of Markov models to the case in which the elements of the text are not directly observable (Baum *et al*., 1970). The result was a dramatic improvement in speech recognizer performance, and by the mid-1980s virtually all work on speech recognition was based on HMMs. (*See* **Speech Recognition, Automatic**)

The state of the art in speech recognition systems is the trigram Markov language model, which predicts each word on the basis of the preceding two words. A trigram model is obviously a poor model of language – for example, if one generates text by random sampling from the distribution it defines, the results cannot be mistaken for real English text. However, it has proven remarkably difficult to improve on trigrams. Only since about 2000 have more sophisticated models been developed that significantly outperform trigrams.

Early on, HMMs were applied to the problem of assigning parts of speech to words (the *tagging* problem). That work, when it became known in the computational linguistics community, was the proximate cause of the surge of interest in statistical methods.

The impressive performance of statistical taggers attracted the attention of computational linguists, but the reason the statistical approach so quickly became the dominant paradigm is because it directly addressed several issues that had frustrated computational linguists immensely.

First was the desire for *robustness*. Real user input is noisy: it is full of misspellings, unanticipated syntactic constructions, and so on; and computational linguistics to that time had failed to develop genuinely noise-tolerant systems. A hallmark of statistical techniques is their noise tolerance.

Second was the desire for *portability*. Applying a manually constructed system to a new subject domain or a new language requires a prohibitive effort. By contrast, algorithms based on statistical methods can be adapted to new domains or new languages by training them on language corpora, and collecting and annotating corpora is usually easier than adapting systems by hand.

A third issue that had frustrated at least some computational linguists was the lack of measurable progress in the field. The statistical approach provides objective measures of success. One desires models that *generalize*: models that capture genuine regularities, not idiosyncrasies of a given data set. A model that captures idiosyncrasies is said to *overfit* the training data. Generalization is measured by a model's performance on a *test set* that is representative of the universe of data of interest, but never seen during construction or training of the model.

Kinds of Statistical Methods

For expository purposes, statistical methods can be divided into three broad classes, which we consider in the following sections: corpus statistics, generative models, and classification and clustering.

CORPUS STATISTICS

Corpus statistics are descriptive statistics that operationalize linguistic concepts. They are closest in spirit to the structuralist procedures, though with an important difference: the operationalizations are not taken to *define* linguistic concepts, but to *approximate* them. Examples are the use of mutual information as a measure of coherence, and divergence as a measure of distributional (dis)similarity. As mentioned above, they can be used to induce grammars; they can also be used to induce lexical information. Mutual information, for example, is used to identify multiword terms such as 'stock market'. Other targets of lexical acquisition include subcategorization frames and selectional restrictions.

GENERATIVE MODELS

A second class of statistical methods involves probability distributions over families of structures. They can be classified by the complexity of the structures involved. Stochastic finite-state automata define distributions over strings, stochastic context-free grammars define distributions over trees, and stochastic attribute–value grammars define distributions over attribute–value structures.

For each class of grammars, there are three main questions of interest: how probabilities are attached

to a grammar in such a way as to give a well-behaved probability distribution to the structures generated by the grammars; how the most likely structure can be computed for an arbitrary input; and how the probabilities in the stochastic grammar can be estimated from a sample.

Finite-state Automata (Hidden Markov Models)

A finite-state automaton (FSA) consists of a set of states, and a set of arcs leading from one state to another. Finite-state transducers of the most familiar sort (called 'Mealey machines') associate output symbols with arcs. In stochastic FSAs, however, it is more common to use automata that associate output symbols with states ('Moore machines'). Machine computations consist in alternately producing an output symbol from the current state ('emission'), then following an arc to a new state ('transition'). (*See* **Finite State Processing**)

In a stochastic FSA, a probability distribution is associated with the outputs from a given state, and a probability distribution is placed on the outgoing arcs from a given state. 'Hidden Markov Model' is another name for a stochastic FSA of this type. The probability of a computation is the product of probabilities of the individual emissions and transitions constituting the computation. The string generated by a computation is the concatenation of the strings generated in each emission step. The *derivation* of a string is the computation – that is, the sequence of states – by which it was generated.

A Hidden Markov Model is hidden in the sense that the derivation of a string cannot in general be uniquely determined. Nothing prevents there being more than one state that emits a given output symbol. To determine the most likely derivation, one can in principle enumerate all derivations of the string and compute their probabilities. This is impractical for strings of any length, inasmuch as the number of derivations increases exponentially with the length of the string. (*See* **Natural Language Processing, Disambiguation in**)

Fortunately, there is an algorithm (the *Viterbi algorithm*) for computing the most likely derivation in time linear in the length of the string. The Viterbi algorithm is a special case of dynamic programming. The key observation is that the most likely partial derivation leading to state q at string position t consists of the most likely partial derivation leading to some state q' at the previous position $t-1$, followed by the transition from q' to q. Instead of keeping track of all (exponentially many) partial derivations at position t, we need only keep track of the most likely partial derivation for each state q at t.

The discussion up to now has assumed that transition and emission probabilities are given. In practise, however, they are not given, but must be estimated from a *training corpus*. A *labeled corpus* is one containing not only natural-language text, but also the sequence of states that the HMM passed through to generate the text. With a labeled corpus, we can essentially estimate HMM probability parameters by counting. For example, the probability of a transition from state q_1 to state q_2 is estimated as the relative frequency of transitions from q_1 to q_2 among transitions out of q_1 in the labeled corpus.

With an *unlabeled* training corpus, in which only the text is available, the sequence of actions taken by the HMM is unknown, and its probabilities obviously cannot be estimated by simple counting. In this case, the standard algorithm is the *forward-backward algorithm*, which is a special case of the *Expectation-Maximization (EM)* algorithm (Dempster *et al*., 1977). One begins with arbitrary parameter estimates – for example, uniform probabilities. The probability of every possible derivation is computed, and one pretends that a derivation occurs a fractional number of times, in proportion to its probability. This gives one a labeled corpus, from which new probabilities can be estimated by relative frequency. One then repeats the process with the new probability estimates. It can be shown that this method improves the probability estimates at each iteration, measuring goodness by the standard *maximum likelihood* criterion.

Even counting relative frequencies involves some subtleties. It is complicated by the *sparse data problem*: the fact that many unobserved actions fail to occur only because the training corpus is not large enough. Indeed, in most cases of practical interest, the majority of possible actions fail to occur even in the largest available corpora. Methods to address the sparse data problem are known as *smoothing methods*. A large variety of them have been studied (Chen, 1996). The easiest is simply to pretend that every possible action occurred at some low frequency. That is, one adds a small count (usually 1 or 1/2) to every count, including the zero counts, before taking relative frequencies. Much better smoothing methods are available; the most commonly used are *Katz back-off* and *deleted interpolation*.

In addition to speech recognition and part-of-speech tagging, stochastic FSAs have applications in *entity recognition* (that is, detecting references in

text to people, companies, dates, times, monetary amounts, and so on) and *partial parsing* (recognizing the major phrases and clauses of a text without completely parsing it).

Stochastic Context-free Grammars

The next more complex grammar class comprises the context-free grammars. A context-free grammar consists in a set of rules of form $A \rightarrow Y_1 \ldots Y_n$, in which A is a nonterminal symbol and $Y_1 \ldots Y_n$ is a (possibly empty) sequence of mixed terminal and nonterminal symbols. $Y_1 \ldots Y_n$ is said to be an *expansion* of A.

A derivation begins with a single, distinguished, nonterminal symbol S. At each point in the derivation, the leftmost nonterminal symbol is replaced with one of its expansions. The derivation continues until no nonterminal symbols remain. A derivation is equivalent to a parse-tree. Each node in the tree represents an expansion: the parent node is labeled with the nonterminal A that is being expanded, and its child nodes are labeled with the expansion symbols $Y_1 \ldots Y_n$.

In a stochastic context-free grammar (SCFG), probabilities are associated with expansions. For any given nonterminal symbol, the probabilities of all its expansions sum to one. The probability of a derivation is the product of the probabilities of the expansions constituting the derivation.

One can modify practically any context-free parsing algorithm to recover the most likely parse. For example, the *CKY parsing algorithm* proceeds as follows. The grammar is assumed to be in Chomsky-normal form, meaning that all expansions are of the form $A \rightarrow B\ C$ (two nonterminals in the expanded form) or $A \rightarrow a$ (one terminal in the expanded form). This assumption involves no loss of generality, as any CFG can be converted to an equivalent grammar in Chomsky-normal form. All possible parse-tree nodes are constructed, in order of increasing width, where the width of a parse-tree node is the number of words of input it covers. For each triple (X,i,w), where X is a nonterminal category, i is a start position, and w is a width, only the most probable subtree is recorded. Since (X,i,w) is constructed out of subtrees of smaller width, all its possible components are guaranteed to exist. (*See* **Parsing**)

There is also a specialization of the EM algorithm, known as the *inside-outside algorithm*, that can be used to estimate the probabilities of an SCFG from unlabeled data. Unfortunately, grammars trained using the inside-outside algorithm produce parse-trees that are not useful for sentence interpretation. This is attributed to the fact that the inside-outside algorithm is designed to minimize sentence perplexity; it has no source of information concerning sentence meaning. As a practical matter, stochastic parsers are trained using labeled data, known as *treebanks*.

Stochastic Attribute–Value Grammars

For our purposes, attribute–value structures can be thought of as directed acyclic graphs (DAGs) with labeled edges. DAGs differ from trees in that DAGs contain *re-entrancies*: nodes that have multiple parents. DAG nodes represent either parse-tree nodes or their values for given attributes. For example, a singular noun phrase can be represented as a node labeled 'noun phrase' linked by an edge labeled 'number' (an attribute) to a node labeled 'singular' (a value). The constituents of the noun phrase are distinguished by edges with labels 'child 1', 'child 2', etc.

Attribute–value structures are generated by attribute–value grammars. Rules in an attribute–value grammar are context-free rules equipped with constraints. Constraints determine the re-entrancies in the DAG. For example, the rule

$$S \rightarrow NP\ VP;\ NP.number = VP.number$$

specifies that the node representing the noun phrase's value for attribute 'number' is the very same node as the one representing the verb phrase's value for 'number'.

Attribute–value grammars are stochasticized by attaching weights to their rules. The probability of a DAG is the product of weights of the rules that define it. Unlike in the finite-state and context-free cases, however, the rule weights cannot be called probabilities. In the finite-state and context-free cases, structures are built up of a number of independent stochastic decisions, and because of the independence of local decisions, the probability of the structure as a whole is the product of local decision probabilities. In the attribute–value case, re-entrancies introduce dependencies among local decisions. Global probabilities are defined as products of local weights, but because of the dependencies among local decisions, the weights are not local probabilities.

Stochastic attribute–value grammars are essentially a variant of *Markov random fields* or *graphical models*. There is a considerable literature on estimation of graphical models (Lauritzen, 1996). For stochastic attribute–value grammars, estimation methods that have been considered include varieties of *Iterative Scaling*.

No tractable exact parsing algorithm is known for stochastic attribute–value grammars. Because of the dependencies among substructures, dynamic programming is not possible. For practical purposes, a common technique is to use stochastic context-free parsing to obtain a small set of candidate structures, which are then evaluated using the full attribute–value grammar.

The unavailability of a dynamic programming algorithm also means that there is no advantage in attaching weights solely to local rules. Typically, weights are attached to features that span much more of the structure than a local expansion.

Specialized Generative Models

Specialized generative models are often developed for specific tasks. Prominent examples are *corpus alignment* and *machine translation*. Corpus alignment involves lining up sentences or smaller phrases between corpora in two different languages, one of which is a translation of the other. Machine translation can be thought of as a similar task, but one in which the problem is to generate the source-language text that would align best with the (observed) target-language text. There have been efforts to estimate fairly direct transfer models, as well as efforts to equip more traditional 'deep' translation models with probabilities.

CLASSIFICATION AND CLUSTERING

There is a lively interchange between computational linguistics and machine learning. New machine-learning techniques are continually introduced into computational linguistics, and the unique challenges of language learning stimulate new directions of research in machine learning.

A central topic in machine learning is classification. The aim of classification is to determine which of a fixed number of classes a given item belongs to. Classification is a supervised learning method – the learning algorithm is given a training corpus of correctly classified examples. A simple example of a classification problem in computational linguistics is prepositional phrase attachment. One widely used data set is constructed from verb phrases of the form 'verb – noun phrase – prepositional phrase', for example 'was selling machine parts from Dresden'. The task is to classify each such example as 'noun attachment' (parts that are from Dresden) or 'verb attachment' (they were sold from Dresden). Generative models can be used in service of classification, but classification can also be done without generative models. A wide variety of classification techniques have been applied to computational linguistic problems, including classification and regression trees (CART), decision lists, Naive Bayes, likelihood ratios, and margin-based methods such as support vector machines (SVMs) and boosting.

Another area of especially strong interaction between machine learning and computational linguistics is unsupervised learning. In unsupervised learning, the training material is not labeled with correct answers. An example is clustering, which is used to induce classes of words, for example in language modeling or in the induction of selectional restrictions.

Intermediate between supervised and unsupervised learning is *bootstrapping*. In bootstrapping, the algorithm is given a very small amount of labeled data, and a large amount of unlabeled data. It can be viewed as supervised learning with supplementary unlabeled data, or as unsupervised learning in which the labeled set is used to give names to the clusters. It has been successfully applied to word-sense disambiguation and named entity classification, among other things.

SUMMARY

Statistical methods have become the standard paradigm in computational linguistics. They can be placed in the historical perspective of American structuralism, though they derive more immediately from statistical speech recognition and Shannon's noisy channel model. They can be grouped roughly into descriptive statistics, generative models (stochastic finite-state, context-free, and attribute–value grammars), and machine-learning methods (classification, clustering, bootstrapping).

References

Baum LE, Petrie T, Sopules G and Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**: 164–171.

Bloomfield L (1933) *Language*. New York, NY: Holt.

Chen SF (1996) *Building Probabilistic Models for Natural Language*. Doctoral dissertation, Harvard University.

Chomsky N (1956) Three models for the description of language. *IRE Transactions on Information Theory* **IT-2**(3): 113–124.

Dempster AP, Liard NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **B.39**: 1–38.

Finch SP (1993) *Finding Structure in Language*. Doctoral dissertation, University of Edinburgh.

Harris Z (1951) *Structural Linguistics*. Chicago, IL: University of Chicago Press.
Lauritzen SL (1996) *Graphical Models*. Oxford, UK: Clarendon Press.
Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* **27**(3–4): 379–423, 623–656.
Shannon CE (1951) Prediction and entropy of printed English. *The Bell System Technical Journal* **30**: 50–64.
Stolz W (1965) A probabilistic procedure for grouping words into phrases. *Language and Speech* **8**: 219–325.

Further Reading

Charniak E (1993) *Statistical Language Learning*. Cambridge, MA: MIT Press.
Jelinek F (1997) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.
Jurafksy D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech*. Upper Saddle River, NJ: Prentice-Hall.
Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Stress

Advanced article

Harry van der Hulst, University of Connecticut, Storrs, Connecticut, USA

CONTENTS

Stress is a phonetic manifestation of accent, which marks the phonological head of the word.

INTRODUCTION

In this article, I will discuss the phenomenon of linguistic stress as it applies to words. Units that are larger than words (such as phrases and sentences) can be said to have stress too, but I will not touch on these larger units here. Right away, in the very next section, I propose to shift our attention to the notion of accent, which I define as more fundamental than stress. Stress, as we shall see, can be seen as a phonetic manifestation of accent. I will provide a typology of the various ways in which accent is manifested besides through stress.

MANIFESTATIONS OF ACCENT

It seems prudent first to make an attempt to define what stress is, or, at least, how I will use the term in this article. Starting with what most people who are able to read this article know (i.e. people who know English), let us consider the following pairs of English words:

convíct cónvict
protést prótest
pervért pérvert (1)

If one pronounces these words, pairwise, one will notice a difference that seems to involve the (relative) prominence of the syllables that the words are composed of. Let us capitalize the prominent syllables:

conVÍCT CÓNvict
proTÉST PRÓtest
perVÉRT PÉRvert (2)

Stress, as I will define it, is (relative) syllable prominence. It is now fair to ask what is meant by 'prominence'. This brings us into the realm of phonetics, i.e. the study of the way speech is produced and perceived. Relative prominence corresponds, on the one hand, with greater articulatory effort in production and, on the other, with greater salience, or audibility, on the perceptual side. Stressed syllables, then, stand out and are easier to perceive than the unstressed, or lesser stressed syllables. Greater articulatory force can be the cause of several effects that can be measured by investigating the details of production, or the physical properties of the produced acoustic signal, e.g.:

Phonetic properties of stressed syllables
- The stressed syllable has greater duration
- The stressed syllable is louder (greater amplitude)
- The stressed syllable is pronounced at a higher pitch (higher fundamental frequency)
- The segments are pronounced with greater precision (3)

This list is not meant to be finite, nor is it couched in the latest language of the trained phonetician. Also, some or all of the phonetic properties may be exclusively or primarily manifested in only a part of the syllable such as its vowel or its rhyme. Whatever the details, a stressed syllable will differ from unstressed syllables in having 'more' of whatever 'stretchable' property any syllable may have (such as duration, pitch, loudness, manner of articulation).

Following researchers such as Hyman (1977), I propose to reserve the term 'stress' for prominence as signaled by the above collection of cues. Then, I will also follow these researchers in saying that stress, in the sense just defined, is a phonetic manifestation or exponent of an abstract property, accent.

Before we address the question of how accent is to be formally understood, let us include another

language in the discussion, namely Safwa (Bantu). Consider the following words or word combinations:

a'mi-ino	'teeth'
ga'mi-ino	'the very teeth'
mi-ino'	'it is teeth'
inko'ombe i'im-bisi	'uncooked beans'
inko'ombe m-bisi'	'the beans are uncooked'

(4)

Again, I have provided certain vowels with what is often (and appropriately) called an 'accent mark'. As in the case of English, speakers of Safwa perceive the syllables that contain these accented vowels as more prominent than the surrounding syllables. When we now look at the articulatory and acoustic properties of the vowels in question, it turns out that what distinguishes them from other vowels in the word is just (or mainly) their relative higher pitch. Thus, the relevant vowels are singled out by only one of the properties that cue the presence of accent in English. But if 'stress' is the collection of all the properties in (3), we cannot say that Safwa has stress. So what *do* we say? The obvious answer may be that Safwa has pitch. We can now capture the difference between English and Safwa terminologically by referring to the former as a stress-accent language and the latter as a pitch-accent language, as proposed in Hyman (1977).

Before we discuss the matter of accent locations, it will also be important to see that the accents can be cued by phonological properties instead of, or in addition to, nondistinctive phonetic cues, although the line between what is called 'phonetic' and what is called 'phonological' lies in different places for different researchers. One important way in which word accents can reveal themselves is by function as anchoring points for some of the tones that make up the intonation melody. (Because these tones, being pitch events, link to word accents, researchers often refer to them as 'pitch-accents', not to be confused with Hyman's notion of pitch-accent introduced earlier.) An intonation melody in a language like English consists of one or more 'pitch-accents' (which can be high tone, low tone, or a contour) and additional boundary tones coming at the beginning or the end of whatever word stretch of the sentence the melody is associated with (cf. Gussenhoven, 1984). This word stretch is usually called the 'intonational phrase', a unit that need not coincide with a syntactic phrase. In the following example the 'pitch accent' is taken to be a high tone (H), and no boundary tones are assumed:

```
H                    H
|                    |
In California they count votes manually
```

(5)

Both *California* and *manually* have several syllables, yet the H tone links to the one that we would call stressed or accented. I am not making a universal claim here on how intonation melodies are anchored to the 'text'. In languages other than English the tune-to-text rules may be different. In any event, in languages that work like English in this respect, 'pitch-accents' function as cues for word accent.

A second nonphonetic cue for accent lies in the notion of phonological contrast. Regularities in the phonological (or phonotactic) patterns of words can be broken down into statements about the inventory of phonological segments (or phonemes) and the possible combinations of these segments. It is not unusual to make statements about the segment combinations in terms of syllables, assuming that a well-formed word is a combination of well-formed syllables (plus, possibly, extra consonants at the beginning or end of the word). It is well known, however, that some syllables allow more segment types and more combinations than others, and at this point it will not come as a surprise to learn the syllables that allow more segment types are the ones that are accented.

Finally, we need to consider yet another type of cue. In English, the sound [p^h] (aspirated p) is restricted to initial position in accented syllables (if not preceded by an /s/). In unaccented syllables, instead, the sound [p] is found. In addition, in syllable final position we always only find [p]. Traditionally, [p^h] and [p] are called allophones (realizations) of one lexical phoneme /p/. The lexical representation of words only has a segment /p/. The aspirated segment is derived by an allophonic rule that forms part of the mapping from the lexical into the postlexical representation. Now, since [p^h] only occurs in accented syllables, its presence is a cue of accent. Thus, one might say that English has a postlexical constraint that bars the segment [p] from a stressed syllable onset. A process 'add aspiration' (called a 'repair rule' in some frameworks) ensures that the lexical form /pin/ is rendered as [p^hin] at the postlexical level. To account for the fact that English has no contrast between /p/ and /p^h/, we assume that the lexical phonology has a constraint that bars a phoneme /p^h/ altogether.

Both levels, then, are characterized by a set of 'wellformedness' constraints, and both levels are served by repair rules that will change forms

that violate these constraints before they can be accepted at that level. Lexical constraints can be violated by new words that are produced by the morphology or that enter the language through deliberate new formations or by loan words. Postlexical constraints can be violated by the forms that are provided by the lexical phonology. Postlexical constraints, unlike lexical constraints, are subject to variables that include style and rate of speech as well as sociolinguistic variables.

All the differences between the contrastive options that can occur in accented syllables and in unaccented syllables, as well as the differences in syllable types that are allowed in these two circumstances, are clear examples of phonological (or phonotactic) cues for accent.

Summing up, we have seen that the location of accent in English can be signaled by at least the following cues:

Cues for accent in English
- Inherent stretchable properties (duration, pitch, loundness, manner)
- Anchoring of intonational tones
- Lexical-phonotactic constraints
- Postlexical 'phonetic' constraints (and the processes that serve them) (6)

All of the above undermines the term 'stress-accent' because they show that accent is manifested in much more than just stress, which only covers the first item in (6). However, it strengthens the more important point that we must separate the notion of accent from the cues that signal its location. Large portions of the lexical and postlexical phonology are determined by the difference between the presence or absence of accent.

ACCENTUAL REPRESENTATIONS: A TYPOLOGY

In the preceding discussion, we have been assuming that accent is a local property of one particular syllable in the word. If this where so, a proper and simple representation of accent would be to assign some sort of mark to the syllable in question (or its vowel), much as in done in dictionaries or transcriptions where accent is marked by an 'accent mark' (as in our example in 1). This practice, although tolerable for some purposes, is inadequate for two reasons, which will almost sound contradictory. First, the use of a local accent mark fails to explain that accent is a 'once-per-word' property, i.e. only one syllable in the word can be accented. Accent is, as is often said, a culminative property. Thus, in order to represent accent formally in the phonological representation of words, it must that there can be only one accent. The second reason for thinking that the accent mark is inadequate is that words apparently can have more than one accent. Consider the examples:

an'tielope,
cro'codi,le (7)

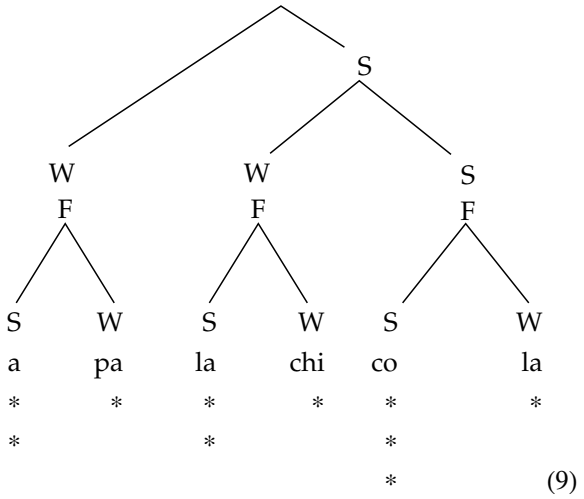
In English, if more than one syllable precedes the accent, a second accent is possible, often indicated by a second type of accent mark. The second mark is said to be a nonprimary or secondary accent (or secondary stress), as opposed to the primary accent (stress). The reader will realize that a contradiction between the two inadequacies of the local accent mark is apparent: the potential presence of a secondary accent does not invalidate the claim that there can be only one primary accent. However, there can be more than one secondary accent:

a, pa la, chi co' la (8)

Secondary accents clearly are a linguistic manifestation of rhythm. In some languages words can have an even greater length than six syllables, and in those cases words can have three rhythmic secondary accents or more. Hayes (1995) describes or refers to many cases of this type.

There are two views on the relationship between primary accent and secondary accent(s). In this section I will focus on one of these. The other view is discussed in the next section.

One view is that a primary accent is basically a promoted secondary accent. In this view, in other words, primary accent is determined on a foundation of secondary accents. A point in favor of this idea is that, in the examples in (7) and (8), the distribution of primary accents seems to follow the same rhythmic pattern that characterizes the secondary accents, in that an accented syllable is typically followed by an unaccented syllable. Hence each accent seems to create a strong–weak domain. Metrical theory formalizes this idea by assuming that the string of syllables of a word is organized in a sequence of binary trochaic (i.e. left-strong) feet. Essentially, a word is compared to a line of verse. The sequence of feet is then organized into a right-strong structure that designates the rightmost foot as the strongest foot in the word:



(9)

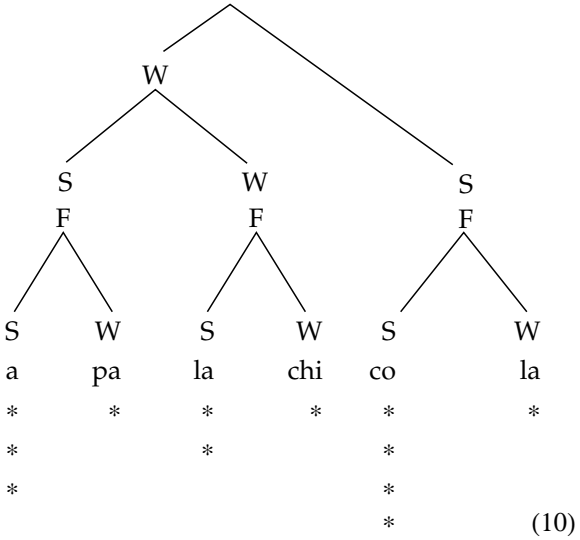
(Here, and in (10), ignore the 'grid' with asterisks, for a brief while.)

The terms 'strong' and 'weak' refer to the idea that accent is a relative notion. A syllable is said to be 'stressed' by virtue of being more prominent than a neighboring syllable. In later work, however, the idea has been expressed that a syllable by itself (e.g. in a monosyllabic word) can also be said to be stressed. In this way, one can make a difference between stressed monosyllabic words and unstressed monosyllabic words, the latter often called (phonological) 'clitics'. In line with the second, nonrelative view on stress/accent, it was also suggested to replace the terms 'strong' and 'weak' by the terms 'head' and 'dependent'.

The feet are binary branching, headed constituent, restricted to two syllables (we call that 'bounded'). The tree structure that organizes the feet is also thought of as binary branching, but not bounded because it can contain more than two feet. The idea that the phonological string of phonemes (just like the string of words that makes up sentences, and the morphemes that make up words) is organized in a headed (most likely binary branching) structure has become widely acknowledged. A systematic account of this view on phonology can be found in the framework of dependency phonology (Anderson and Ewen, 1987), that, in its earliest work, predates metrical phonology. (Below the level of the foot, we find that segments are organized into headed syllables consisting of binary headed syllabic constituents like onset and rhyme.)

Why is the structure in (9) adequate as a representation of the accentual structure of *apalachicola*? Notice, first, that only one syllable can be exclusively dominated by nodes labeled S, or head, (and the root node). This syllable, then, is the head of all heads, the ultimate head (UH), of the word and that seems an adequate formal representation of the notion of primary accent. Second, secondary accents are uniformly represented as heads of feet. The structure in (9) suggests that the two feet preceding the primary accent are not equal because they are structurally different within the word tree structure.

Indeed, according to speakers of English, the nonprimary accent on the first syllable is stronger than the nonprimary accent on the third syllable (which is sometimes called a 'tertiary accent'). Is (9) an adequate representation of that difference? It would be, if we posit the axiom that the more deeply embedded a weak foot is, the weaker it is. However, one might also wish to consider the structure in (10) that more directly shows that the second foot is subordinate to the first:



(10)

To date, it has been difficult to decide on such matters, which has seduced some researchers into using nonbinary organization of feet resulting in a 'flat' structure.

The issue can be resolved quite easily, however. After the introduction of metrical theory, there have been a number of 'internal debates' on certain notational issues (cf. Halle and Vergnaud, 1987, and van der Hulst, 1999). Confusing to the relative outsider may be the use of so called metrical grids. Originally, Liberman and Prince (1977) introduced two simultaneous structures to account for the accentual structure of words, the tree, and the grid. The grid consists of a series of columns, one for each syllable, the height of which indicates the degree of accent. The principle in (11) derives the grid from the tree:

Tree–grid correlation
In any constituent the head has one more asterisk than its dependent. (11)

Notice, how in accordance with this principle, the grid in (10) nicely makes a three-way distinction between three degrees of accentual strength, whereas the grid in (9) does not have the same distinction.

Given that the grid is merely an interpretative device, one might argue that it is strictly speaking redundant. Realizing this, for a brief while, reseachers considered abandoning the tree rather than the grid, thus giving up on constituency. Others argued that grids had to be abandoned. Halle and Vergnaud (1987) argue that constituency should not be eliminated, but to please all parties they adopt a notation that uses the grid, enriched with brackets to indicate constituency, which is not different, of course, from using the headed tree structures (although it does have the typographical advantage of not having to draw tree structures).

English, apparently has left-headed feet, with the rightmost foot being the head foot (cf. Kager, 1989, for many details and discussion):

English accent
- foot: left-headed (iterating through the word from right to left)
- word: right-headed (12)

The framework of metrical theory has shown to be very productive in accounting for cross-linguistic variation in accentual patterns by assuming that we can find variation along the two parameters (see (13)):

word \ foot	left-headed	right-headed
left-headed	initial syllable	second syllable
right-headed	penultimate syllable	final syllable

(13)

Feet must be assigned iteratively if the word has more than two syllables, and it must therefore also be specified whether this iteration works from right-to-left or from left-to-right, since that will make a difference in case the number of syllables in the word is uneven. For English, since the head foot is on the right, it can easily be shown that the iteration is from right-to-left. With the head foot being on the right, the location of primary accent would be dependent on the number of syllables in the word if the iteration was left-to-right, as shown in (14):

```
head right          *                       *   2
left-to-right  (*     *)     (*     *     *) 1
               (* *)(* *)  (* *)(* *)(*) 0

               1  2 3 4    1   2 3 4 5      (14)
```

I use here the bracketed grid notation mentioned above. Level 0 represents all the stressable units (the rhymes). On level 1 we represent the foot heads, and level 2 is for the word head. If a language were to have a system like that in (14), the location of the primary accent would be penultimate in words that have an even number of syllables and final where the number of syllables is odd. This description certainly fails to apply to English, which therefore must have right-to-left footing.

(Systems of the type in (14) have been claimed to exist, but I will argue below that metrical algorithms should not produce them directly. Thus, I will assume that the direction of footing is not independent from the edge choice for the head foot. Since in English the head foot is on the right, footing must be from right-to-left. I shall return to this issue later.)

We now turn to a factor that may influence and interrupt the regular way in which syllables are gathered into feet. So far, we have assumed that syllables are gathered in groups of two, monosyllabic feet arising only where we run out of syllables. In some languages, however, certain types of syllables (called 'heavy') may not appear as the dependent in a foot. This property is called 'weight-sensitivity'. The weight of a syllable is determined by its intrinsic, phonological properties. There are various types of intrinsic weight, and weight is typically (perhaps exclusively, depending on the analysis) binary: i.e. languages will split the set of syllables in to two, one called 'heavy', the other called 'light':

heavy	*light*
long vowel	short vowel
closed syllable	open syllable
checked vowel	unchecked vowel
low vowel	nonlow vowel
high-toned vowel	low-toned vowel
full vowel	reduced vowel

(15)

The first three types in (15) have also been called 'moraic weight', where the heavy syllables contain two units in their syllable rhyme – these units are called 'moras' (or weight units). In the remaining three cases (for which we might adopt the term

'sonority weight'), the heavy syllable is more salient by virtue of its greater aperture, its higher pitch, or its more complex articulation.

Intuitively, it may be seem clear that the properties in the left-hand column give more prominence to a syllable (or its rhyme) in terms of duration (long vowel, checked vowels, closed syllables, full vowels), high pitch, loudness (open vowel), manner of articulation (full vowel) – precisely those factors that can be found as phonetic cues of accent. It seems obvious that syllables that have more of those properties intrinsically (i.e. as distinctive properties) are reluctant to appear in positions that typically have fewer of them, i.e. unaccented positions. Conversely, syllables with such intrinsic properties will 'attract' accent.

The assignment of feet can also be influenced by lexical irregularity. Thus for example in Polish, which has weight-insensitive penultimate primary accent, some words have irregular final or antepenultimate accent. How can we account for that? The answer is that irregular final accent is achieved by assigning a lexical mark to the final syllable, and adding the convention that syllables with such marks may not appear in the dependent position of the foot. Elsewhere, I have referred to such marks as 'diacritic weight'. Such marks usually are historical residues of an earlier situation in which the relevant syllables had intrinsic weight. After a language has lost, e.g. a vowel length contrast, the accents can stay in the same position and thus, in a sense, become unpredictable. Thus, there are two types of weight:

Sensitivity of foot assignment: *the dependent cannot dominate*
- A syllable having certain phonological properties (intrinsic weight)
- A lexically marked syllable (diactritic weight) (16)

Lexical accent structure can be sensitive to both diacritic and intrinsic weight (English).

The antepenultimate exceptions require another type of lexical encoding: for example encoding the final syllable as being disregarded by the metrical algorithm. This is called 'extrametricality'.

LEXICAL AND POSTLEXICAL STRUCTURE

In the preceding section, I have proposed that the direction of footing and the edge choice of the head foot are correlated:

Direction (left-to-right) = head foot left
Direction (right-to-left) = head foot right (17)

Thus, with left-headed feet, we have assumed only two possible systems:

a. *Initial accent*

left-headed	*	*
left-to-right	(* *)	(* * *) 1
	(* *)(* *)	(* *)(* *)(*) 0
	1 2 3 4	1 2 3 4 5

b. *Penultimate accent*

right-headed	*	*
right-to-left	(* *)	(* * *) 1
	(* *)(* *)	(*)(* *)(* *) 0
	1 2 3 4	1 2 3 4 5

(18)

The English system presents a variety of (18b). If, however, the direction of foot assignment does not have to correlate with the choice of the head foot, two further systems can be produced:

a.

right-headed	*	* 2
left-to-right	(* *)	(* * *) 1
	(* *)(* *)	(* *)(* *)(*) 0
	1 2 3 4	1 2 3 4 5

b.

left-headed	*	*
right-to-left	(* *)	(* * *) 1
	(* *)(* *)	(*)(* *)(* *) 0
	1 2 3 4	1 2 3 4 5

(19)

In (19a), which is identical to (14), where the direction is from the left, and the head is on the right (the parameters have opposite values, so to speak), we derive a system in which the location of primary accent is actually dependent on the number of syllables; in even-numbered syllables primary accent is penultimate, while in odd-numbered words, it is final. (19b) would have initial accent in both cases, but the rhythmic structure would be odd. I am not aware of any such systems being reported in the literature. As mentioned in the previous section, systems as in (19b) do seem to occur, but they are rare. In only a few cases do we find that primary accent is truly dependent on rhythm. However metrical theory, with its bottom-up procedure of first building feet and then the word structure, predicts that the cases in (19) should be just as common as these in (18). Given their rarity, I would like to argue that we might want to exclude the possibilities in (19) from our basic apparatus.

We can do this by somehow assigning primary accent first. With primary accent in place, we can then account for rhythmic structure in terms of the assignment of secondary accents that typically 'ripple or echo away' from the primary accent. Rather than stipulating the order in which primary accent and secondary accents are assigned in terms of rule ordering, we can attribute the two aspects of the overall accentual pattern to the lexical and postlexical phonology, respectively.

Lexical and postlexical structure, which I will call here 'phonotactic' and 'prosodic', respectively, may differ in a number of ways. This supports the idea that there are, in fact, two algorithms. In (20) I give a number of examples of such differences:

	Lexical	*Postlexical*	
Foot	weight-sensitive	weight-insensitive	(English)
	weight-insensitive	weight-sensitive	(Finnish)
	left-headed	right-headed	(BigNambas, Marind)
	right-headed	left-headed	(Taga, Dari, Uzbek)
Word	right-headed	left-headed	(English)
	left-headed	right-headed	(Turkish)

(20)

In standard approaches to metrical structure, such mismatches are not interpreted as evidence for two structures but as evidence for rules that transform an initial lexical structure into a later structure (not necessarily referred to as postlexical). This is a typical derivational approach, stemming from the tradition of generative phonology. For example, Halle and Vergnaud (1987) propose that in English feet are assigned from right-to-left giving a right-headed tree and thus primary accent at the right edge. Then, to account for the fact that the secondary accents come from the left, they 'erase' all feet except the head foot and assign feet for a second time, now from left-to-right. In my approach, the apparent conflict between right-to-left and left-to-right footing is taken as evidence for a two-level analysis.

One might now ask how we can account for the cases in (19a) in which, contrary to the majority situation, primary accent does seem to be dependent on the prior existence of rhythmic structure. Space limitations prevent me from discussing this issue in detail. In this case, we need to say that the assignment of postlexical structure is such that the dependent in the postlexical feet cannot be rhythmically strong, while the distribution of rhythm is accounted for in terms of lexical footing. What remains to be explained is why in such cases the head of rightmost foot in the postlexical structure prevails over the ultimate head of the lexical structure (which is on the first syllable).

SOME FURTHER ISSUES

This article has discussed issues of representation and typology, but certain important issues that involve accent/stress have not been dealt with. I will mention two such issues briefly here.

This article has focused on word level accent. In this domain, it is relevant to consider the relationship between the accentual pattern and the morphological structure. One expects that only lexical metrification can be sensitive to morphological structure. Indeed, it has been argued that, for example, the English 'stress' rule is applied within domains that can be smaller than the word if the word is morphologically complex and either compounded or affixes with so-called level 2 affixes. I have not discussed the phenomenon of accent at higher levels than the word, but it should be clear that phonological structure is also relevant for syntactic organization. Here, going beyond the domain of the lexicon, the distinction between phonotactic and prosodic organization no longer applies. A discussion of higher level prosodic structure and its relation to syntactic structure requires a separate article (cf. Nespor and Vogel, 1986). I have also not discussed phenomena involving rules that 'shift' stress as in the famous pair (…) *thirteen'* versus *thir'teen* (men), where the location of accent in *thirteen* differs depending on the syntactic or prosodic context. In line with the suggested analysis of English stress, the different locations correspond to the lexical primary accent (right edge) and the postlexical primary accent (left edge). In the form *thir'teen* the postlexical primary accent has taken over primacy from the lexical accent in order to avoid a stress clash between the accent on *teen* and *men*. Rhythm, then, is an important determinant of the distribution of postlexical accents, not only at the foot level but also at higher prosodic levels. A full discussion of such shifts is also beyond the scope of this article. A thorough discussion of many of the relevant facts and analyses can be found in Visch (1999).

Finally, one might ask whether all languages are accentual at the word level. I suspect that the answer is affirmative. We have seen that accent determines much more than pitch or stress. It seems almost inconceivable to me that we would come across languages that would lack all of the possible cues for accent. We have also seen that accents correspond to the notion (ultimate) head, where this head is just a part of the overall structure that organizes the phonological structure of the word. Expecting to find languages that lack such hierarchical structure at the word level is like

expecting to find languages in which sentences are linear strings of words without any syntactic organization.

References

Anderson JM and Ewen CJ (1987) *Principles of Dependency Phonology*. Cambridge, UK: Cambridge University Press.

Gussenhoven C (1984) Intonation: a whole autosegment language. In: van der Hulst H and Smith N (eds) *Advances in Nonlinear Phonology*, pp. 117–133. Dordrecht: Foris Publications.

Halle M and Vergnaud J-R (1987) *An Essay on Stress*. Cambridge, MA: MIT Press.

Hayes B (1995) *A Metrical Theory of Stress: Principles and Case Studies*. Chicago, IL: University of Chicago Press.

van der Hulst H (1999) Word accent. In: van der Hulst H (ed.) *Word Prosodic Systems in the Languages of Europe*, pp. 3–116. Berlin and New York: Mouton de Gruyter.

Hyman L (1977) On the nature of linguistic stress. In: Hyman L (ed.) *Studies in stress and Accent*. Scopil 4: 37–82.

Kager R (1989) *A Metrical Theory of Stress and Destressing in English and Dutch*. Dordrecht: Foris Publications.

Liberman M and Prince A (1977) On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249–336.

Nespor M and Vogel I (1986) *Prosodic Phonology*. Dordrecht: Foris Publications.

Visch E (1999) The rhythmic organization of compounds and phrases. In: van der Hulst H (ed.) *Word Prosodic Systems in the Languages of Europe*, pp. 161–232. Berlin and New York: Mouton de Gruyter.

Further Reading

Beckman M (1986) *Stress and Non-stress Accent*. Dordrecht: Foris Publications.

Beckman M and Pierrehumbert J (1986) The intonational structure in English and Japanese. *Phonology* **3**: 255–309.

Dogil G (1999) The phonetic manifestation of word stress. In: van der Hulst HG (ed.) *Word Prosodic Systems in the Languages of Europe*, pp. 273–310. Berlin and New York: Mouton de Gruyter.

Garde P (1968) *L'Accent*. Paris: Presses Universitaires de France.

Goldsmith J (1988) Prosodic trends in the Bantu languages. In: van der Hulst HG and Smith N (eds) *Autosegment Studies on Pitch Accent*, pp. 81–94. Dordrecht: Foris Publications.

Gussenhoven CHM (1991) The English rhythm rule as an accent deletion rule. *Phonology* **8**: 1–35.

Gussenhoven CHM and Bruce G (1999) Word prosody and intonation. In: van der Hulst HG (ed.) *Word Prosodic Systems in the Languages of Europe*, pp. 233–272. Berlin and New York: Mouton de Gruyter.

Halle M and Idsardi W (1994) General properties of stress and metrical structure. In: Goldsmith J (ed.) *A Handbook of Phonological Theory*, pp. 403–443. Oxford: Blackwell.

Haraguchi S (1977) *The Tone Pattern of Japanese: An Autosegmental Analysis*. Tokyo: Kaitakusha.

Haraguchi S (1988) Pitch accent and intonation in Japanese. In: van der Hulst HG and Smith N (eds) *Autosegment Studies on Pitch Accent*, pp. 123–150. Dordrecht: Foris Publications.

Haraguchi S (1991) *A Theory of Stress and Accent*. Dordrecht: Foris Publications.

Harms RT (1981) A backwards metrical approach to Cairo Arabic stress. *Linguistic Analysis* **7**: 429–451.

Hayes B (1984) The phonology of rhythm in English. *Linguistic Inquiry* **13**: 227–276.

Hayes B (1995) *A Metrical Theory of Stress: Principles and Case Studies*. Chicago: University of Chicago Press.

van der Hulst HG (1984) *Syllable Structure and Stress in Dutch*. Dordrecht: Foris Publications.

van der Hulst HG (1996) Separating primary accent and secondary accent. In: Goedemans R, van der Hulst HG and Visch E (eds) *Stress Patterns of the World*, pp. 1–26. The Hague: HAG.

van der Hulst HG (1999) Issues in foot typology. In: Davenport M and Hannahs SJ (eds) *Issues in Phonological Structure*, pp. 95–127. Amsterdam: John Benjamins.

van der Hulst HG, Hendriks B and van de Weijer J (1999) A survey of European word prosodic systems. In: van der Hulst HG (ed.) *Word Prosodic Systems in the Languages of Europe*, pp. 425–476. Berlin and New York: Mouton de Gruyter.

Hurch B (1992) Accentuations. In: Hurch B and Rhodes R (eds) *Natural Phonology: The State of the Art on Natural Phonology*, pp. 73–96. Berlin and New York: Mouton de Gruyter.

Hyman L (1984) *A Theory of Phonological Weight*. Dordrecht: Foris Publications.

Inkelas S and Zec D (eds) (1990) *The Phonology–Syntax Connection*. Chicago: University of Chicago Press.

Kager R (1993) Alternatives to the iambic-trochaic law. *Natural Language and Linguistics Theory* **11**: 381–432.

Kager R (1995) The metrical theory of word stress. In: Goldsmith J (ed.) *A Handbook of Phonological Theory*, pp. 367–402. Oxford: Blackwell.

Lehiste I (1970) *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman M and Prince A (1977) On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249–336.

Prince A (1983) Relating to the grid. *Linguistic Inquiry* **14**: 19–100.

Prince A and Smolensky P (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report No. 2 of the Rutgers Center for Cognitive Science. Piscataway, NJ: Rutgers University.

Revithiadou A (1998) *The Prosody–Morphology Interface*. The Hague: HAG.
Roca I (1986) Secondary stress and metrical rhythm. *Phonology Yearbook* **3**: 341–370.
Salmons J (1992) *Accentual Change and Language Contact*. London: Routledge.
Selkirk E (1980) The role of prosodic categories in English word stress. *Linguistic Inquiry* **11**: 561–605.
Selkirk E (1984) *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA: MIT Press.

Syntactic Encoding and Syntactic Choice

Advanced article

Padraig G O'Seaghdha, Lehigh University, Bethlehem, Pennsylvania, USA

CONTENTS

Syntactic encoding creates a representation intermediate between a nonlinguistic conceptual level and a subsequent phonological level containing word forms. Lexical selection contributes to construction of a syntactic frame. Syntactic persistence, a form of implicit memory, suggests that syntactic structure is partly isolable from content. The relation between conceptual and syntactic formulation is a crucial question that is ripe for further exploration.

INTRODUCTION

Speakers continuously inject nonlinguistic thoughts into linguistic molds as they talk. Standard language production theory (e.g. Bock and Levelt, 1994; Garrett, 1975; Levelt, 1989) postulates a message level in which ideas and other mental contents are represented in a variety of prelinguistic codes. The theory is largely agnostic on the nature of these codes. There are two primary stages in the conversion of a message to the linguistic forms that eventually enables articulation to proceed, syntactic encoding and phonological encoding. This entry addresses the first of these, syntactic encoding. (*See* **Speech Production**)

The idea or message necessarily precedes its encoding. This, combined with the requirement of fluency in speech, means that syntactic encoding is incremental (de Smedt, 1996). That is, syntactic encoding need not proceed in elaborate full-sentence packages but rather progresses as soon as some viable component is ready for transmission to the next level of encoding. But what is viable? The answer to this is partly determined by the two-part nature of syntactic encoding itself. (*See* **Language Production, Incremental**)

The standard account distinguishes two primary subprocesses of syntactic encoding: functional encoding and positional encoding. At the risk of oversimplification, functional encoding begins with the activation and selection of key content words, and positional encoding involves the sequential arrangement, or linearization, of these words with their grammatical affixes, and of additional function words. Activation of potential keywords triggers certain syntactic options and restrictions. Properties such as number and gender of nouns as well as syntactic requirements of verbs are made available at the functional level when words are first activated (Garrett, 1975). Thus, for example, it has been shown that number agreement errors are determined at the functional level rather than by proximity in positionally specified speech (e.g., Bock and Miller, 1991). Likewise, classic studies of natural speech errors (e.g., Garrett, 1975) show that whole word errors mainly involve mis-selection of words performing similar grammatical roles in separate phrases. This highlights two crucial points: that units are defined by their syntactic functions, and that syntactic encoding at the functional level certainly reaches beyond the current phrase. (*See* **Speech Error Models of Language Production**)

This article first examines how lexical elements are selected and how they influence the sequencing process called linearization. It then briefly considers some recently discovered properties of syntactic structures as distinct mental entities that are at least partially dissociable from specific lexical or message content. An overarching constraint on language production is that more or less simultaneous and stable representations at one level are translated into smaller sequenced linguistic pieces at the next. The most crucial of these translations, that from thought into language, is discussed.

LEXICAL ACTIVATION AND SYNTACTIC CHOICE

In encoding a message, a speaker must select lexical items and sequence them according to the rules of the language. Most theorists assume that these

processes are intimately connected. At the functional level, key content words are activated, leading to the construction of syntactic fragments. The initial activation of content words is necessarily strongly influenced by the organization of the prelinguistic message (Bock, 1986; de Smedt, 1996). Thus, other things being equal, factors such as animacy influence the grammatical roles assigned to nouns (McDonald *et al*., 1993). Most importantly, verbs play a key role in functional assignment, influencing the selection and subsequent ordering of other words in the sentence.

To illustrate, consider a transaction in which Sue puts money in the hands of the curator of the poverty-stricken Museum of Dead Languages. In English, we may describe this event using the verb *give* or the verb *donate*. If *give* is chosen, the speaker may say either *Sue gave the curator money* OR *Sue gave money to the curator/ museum*. However, if *donate* is used, the speaker must say *Sue donated money to the museum*. Thus, in the process of incremental sentence generation, one or both of the verbs may be activated. Further, if *give* is activated, either alone or together with *donate*, two configurations of the remainder of the sentence are available. Finally, *give* prefers an animate but *donate* an impersonal recipient. The options for each verb are summarized in Figure 1.

The flexibility entailed by the existence of choice may pre-empt conflict and abet fluency if speakers choose the most available option without hesitation (Ferreira, 1996). However, in other circumstances, competition may arise at the choice points of verb selection and sentence continuation (see Dell and O'Seaghdha, 1994; Stallings *et al*., 1998). The configuration in Figure 1 presents a number of processing options. (*See* **Choice Selection**)

- *Give* and *donate* may compete for selection either by direct contest between their representations or indirectly via a selection mechanism that assesses their absolute or differential fitness (see Dell and O'Seaghdha, 1994).

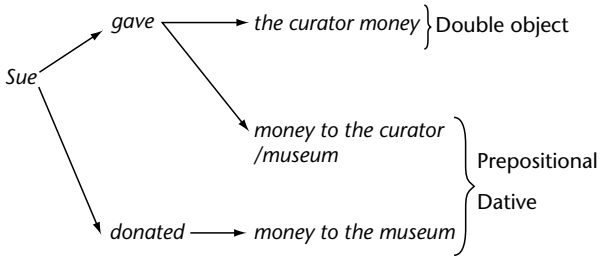


Figure 1. Some syntactic options for describing an act of pecuniary generosity.

- If *give* is ascendant, the alternative direct object and prepositional dative continuations may likewise compete directly or indirectly for selection. This contest may be between the structures themselves, between elements within them (in this case the *curator/museum* and the *money)*, or both.
- Most interestingly perhaps, even in the case that *donate* is not selected, its allegiance with the prepositional dative option may dispose the preferred *give* to that option. Evidence from the *fast priming* paradigm in sentence comprehension (Trueswell and Kim, 1998) suggests that verbs activate their syntactic preferences automatically and so may change the landscape for syntactic choices even when they are not overtly present.

This kind of sensitivity to verb preferences can be accounted for by constraint-based approaches in which statistics of use are coded with verbs (MacDonald *et al*., 1994). Verb preferences may influence not only the choice of local syntactic structure for immediately following words but also the ordering of entire phrases in postverbal position. Thus, Stallings *et al*. (1998) proposed that verbs that tend to have their objects nonadjacent in a variety of syntactic structures tend to show that disposition generally. More broadly, the constrained and sometimes rather arbitrary choices afforded and imposed by lexical options require that thinking, at least if it is to be expressed in words, be cognizant of syntactic obligations (Slobin, 1996). (*See* **Constraint-based Processing**; **Syntactic Form Frequency: Assessing**)

SYNTACTIC STRUCTURE PRIMING

The suggestion above that *donate*'s structural requirement for the prepositional dative might attract *give* toward that option is an instance of the intriguing phenomenon known as *syntactic structure priming*. Bock (1986) first brought this into the limelight using a paradigm in which participants were exposed to sentences in particular configurations and then described pictures bearing no meaningful relation to the preceding sentence. She found that speakers tended to echo the structure of preceding sentences in their descriptions of the pictures. Anecdotal evidence of such echoing existed before, but it lacked empirical substantiation and theoretical pertinence. Bock, her colleagues, and other researchers have now expanded the theoretical and empirical horizons considerably.

- The effect appears to be structural and not linked to semantics (Bock and Loebell, 1990).
- It is persistent over many intervening structurally unrelated sentences, and so is not a function of short-term

activation (Bock and Griffin, 2000). The exercise of a syntactic choice appears to leave a relatively enduring trace.

- It has been mimicked by a sequential learning model in which the structural effects are captured by long-lasting weight changes (Dell *et al*., 1999). Bock and Griffin identify these adjustments as 'dynamic vestiges' of processes involved in mastering the skill of actual speech production. (*See* **Implicit Learning**)
- However, the processes involved may be more abstract than this because syntactic structure priming occurs whether the persisting influence originates in a previously produced sentence or in a previously heard or read sentence. (*See* **Production–Comprehension Interface**)

Syntactic priming is akin to structural persistence at the lexical level (e.g., Church and Fisher, 1998). The latter kind of structure priming is observed in cases of lexical repetition among speakers called cryptomnesia, and both processes may be implicated in the phenomenon of conversational coordination (e.g., Branigan *et al*., 2000). Finally, both lexical and syntactic persistence, not mere retention of a literal record, appear to be necessary to explain the accuracy of immediate complex sentence recall (Potter and Lombardi, 1998).

HIGH-LEVEL PLANNING PROCESSES AND SYNTACTIC DECISIONS

Until recently, the preoccupation of psycholinguists with lexical and sentence-level processes has resulted in neglect of higher discourse-level issues. Much of the work on discourse level planning for production has been conducted by computational linguists in relative isolation from psycholinguistics (see Andriessen *et al*., 1996). However, the ongoing ferment in language production research has recently begun to work its way into the upper echelons of the language processing system.

The key question in this recent work is centered around an old question which we may call the Wundt–Lashley problem. The general problem arises between any two levels of planning for production when the functional requirements at the two levels lead to noncoincident organizations (see Lashley, 1951). The particular form of the problem that originates with Wundt (1900) concerns the coordination of nonlinguistic conceptual representations with linguistic ones. This problem may be more difficult than translation between levels of linguistic representation, especially if, as Wundt believed, conceptual representations are not subject to the same temporal sequencing imperatives as linguistic ones. In Wundt's view, conceptual representations are in important respects simultaneous and atemporal, whereas linguistic representations are strongly time-bound. The problem then is how to transduce the atemporal into the temporal. Wundt proposed a solution – that speech is planned conceptually in simultaneous components of at least the size of a clause, and then encoded syntactically in smaller phrase-sized chunks – that has received good support from two recent studies.

Griffin and Bock (2000) took a very direct approach by examining the timing of eye fixations during inspection of simple scenes under varying instructions. Patterns of eye fixations showed that speakers first apprehended an entire event before proceeding to elaborate the linguistic structure via a more gradual process in which eye fixations and lexical-phrasal choices were strongly correlated. This method shows great promise of providing further insight into processes at the conceptual–lexical interface during syntactic encoding.

Using a less direct measure, sentence initiation time, Smith and Wheeldon (1999) examined the scope of conceptual and syntactic planning in more complex sentences. They found that conceptual planning included at least an entire clause, often with partial look-ahead beyond it, and that syntactic encoding was primarily regulated by phrase structure. The broad consistency of the findings of Smith and Wheeldon with those of Griffin and Bock (as well as with earlier studies by Dell and O'Seaghdha, 1992; Ferreira, 1991; Meyer, 1996, among others) despite use of very different methods inspires confidence. Although the experimental study of conceptual–syntactic coordination has barely begun, further progress in this important underresearched area of cognitive science is imminent.

References

Andriessen J, de Smedt K and Zock M (1996) Discourse planning: empirical research and computer models. In Dijkstra T and de Smedt K (eds) *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing*. London, UK: Taylor & Francis.

Bock JK (1986) Syntactic persistence in language production. *Cognitive Psychology* **18**: 355–387.

Bock K and Griffin ZM (2000) The persistence of structural priming: transient activation or implicit learning. *Journal of Experimental Psychology: General* **129**: 177–192.

Bock K and Levelt WJM (1994) Language production: grammatical encoding. In Gernsbacher MA (ed.) *Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Bock K and Loebell H (1990) Framing sentences. *Cognition* **35**: 1–39.
Bock K and Miller CA (1991) Broken agreement. *Cognitive Psychology* **23**: 45–93.
Branigan HP, Pickering MJ and Cleland AA (2000) Syntactic coordination in dialogue. *Cognition* **75**: 13–25.
Church BA and Fisher C (1998) Long-term auditory word priming in preschoolers: implicit memory support for language acquisition. *Journal of Memory and Language* **39**: 523–542.
Dell GS and O'Seaghdha PG (1992) Stages of lexical access in language production. *Cognition* **42**: 287–314.
Dell GS and O'Seaghdha PG (1994) Inhibition in interactive activation models of linguistic selection and sequencing. In Dagenbach D and Carr TH (eds) *Inhibitory Processes in Attention, Memory, and Language*, pp. 409–453. San Diego, CA: Academic Press.
Dell GS, Chang F and Griffin ZM (1999) Connectionist models of language production: lexical access and grammatical encoding. *Cognitive Science* **23**: 517–542.
de Smedt K (1996) Computational models of incremental grammatical encoding. In Dijkstra T and de Smedt K (eds) *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing*. London, UK: Taylor & Francis.
Ferreira F (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* **30**: 210–233.
Ferreira VS (1996) Is it better to give than to donate? The consequences of syntactic flexibility in language production. *Journal of Memory and Language* **35**: 724–755.
Garrett MF (1975) The analysis of sentence production. In Bower GH (ed.) *The Psychology of Learning and Motivation*. New York, NY: Academic Press.
Griffin ZM and Bock JK (2000) What the eyes say about speaking. *Psychological Science* **11**: 274–279.
Lashley KS (1951) The problem of serial order in behavior. In Jeffress LA (ed.) *Cerebral Mechanisms in Behavior*. New York, NY: John Wiley.
Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
MacDonald MC, Pearlmutter NJ and Seidenberg MS (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 672–703.
McDonald J, Bock K and Kelly MH (1993) Word and world order: semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* **25**: 188–230.
Meyer AS (1996) Lexical access in phrase and sentence production: results from picture-word interference experiments. *Journal of Memory and Language* **35**: 477–496.
Potter MC and Lombardi L (1998) Syntactic priming in immediate recall of sentences. *Journal of Memory and Language* **38**: 265–282.
Slobin DI (1996) From 'thought and language' to 'thinking for speaking'. In Gumperz JJ and Levinson SC (eds) *Rethinking Linguistic Relativity*, pp. 70–96. Cambridge, UK: Cambridge University Press.
Smith M and Wheeldon L (1999) High level processing scope in spoken sentence production. *Cognition* **73**: 205–246.
Stallings L, MacDonald MC and O'Seaghdha PG (1998) Phrasal ordering constraints in sentence production: phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language* **39**: 392–417.
Trueswell JC and Kim AE (1998) How to prune a garden path by nipping it in the bud: fast priming of verb argument structure. *Journal of Memory and Language* **39**: 102–123.
Wundt W (1900) *Die Sprache*. Leipzig, Germany: Kroner.

Further Reading

Bock K (1990) Structure in language: creating form in talk. *American Psychologist* **45**: 1221–1236.
Bock K (1996) Language production: methods and methodologies. *Psychonomic Bulletin & Review* **3**: 395–421.
Dell GS (1995) Speaking and misspeaking. In Gleitman LR, Liberman M *et al.* (eds) *Language: An Invitation to Cognitive Science*, 2nd edn, vol. 1, pp. 183–208. Cambridge, MA: MIT Press.
Ferreira F (2000) Syntax in language production an approach using tree-adjoining grammars. In Wheeldon L (ed.) *Aspects of Language Production*, pp. 291–330. Hove, UK: Psychology Press.
Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Syntactic Form Frequency: Assessing

Intermediate article

Marc Brysbaert, Royal Holloway College, Egham, UK
Don C Mitchell, University of Exeter, Exeter, UK

CONTENTS

Syntactic form frequency research aims to find out how often words and sequences of words in a given context fulfil particular syntactic roles in a sentence. This is achieved by analyzing representative samples of text or speech materials. The research is motivated by the findings that artificial grammars perform better when they take syntactic form frequencies into account, and that humans also seem to be sensitive to this kind of information.

THE NEED FOR ASSESSING SYNTACTIC FORM FREQUENCIES

Syntactic form frequency refers to the number of times words and sequences of words in a given context fulfil particular syntactic roles in a sentence. When a human or a machine tries to convey a message, it is important not only to use the correct words, but also to assign the correct roles to the different parts of the sentence (i.e. to explain 'who did what to whom'). This is equally vital when listeners or readers try to recover the meaning intended by the speaker or the writer.

The assignment of the correct syntactic structure is less straightforward than it might seem. First, the structure of a sentence is not uniquely defined by the position of the words in the sentence. Although the order subject–verb–direct object–indirect object is the basic order in English, this is by no means the only possible sequence (as shown by the examples 'Lyn gave Charles the pencil' and 'The pencil was given by Lyn to Charles'). Conversely, small differences in word order can introduce large differences in meaning (compare 'He showed her baby the pictures' with 'He showed her the baby pictures'; Frazier and Clifton, 1996). Second, many sentences include regions that allow more than one interpretation, even when the parser is able to make a distinction between the subject, the verb, and the object. For instance, in the sentence 'The cop informed the motorist that he had followed…' the final clause could either be a complement structure ('… that he had followed the instructions') or a relative clause ('…that he had followed the whole day, that …'). In principle, the number of syntactic ambiguities grows exponentially with sentence length. This became apparent when one of the first artificial grammars was applied to some example input sentences (Martin *et al*., 1983):

List the sales of products in 1973.
(3 analyses possible: the products in 1973, the sales in 1973, or the listing in 1973) (1a)

List the sales of products produced in 1973.
(10 analyses possible) (1b)

List the sales of products produced in 1973 with the products produced in 1972. (455 analyses) (1c)

Not all possible syntactic forms occur equally often, however. Some are more frequent than others. Because researchers believe those differences in frequency are important both for practical purposes (e.g. to build an artificial language device) and to understand the ways in which humans interpret the syntactic structure of sentences (see below), they have tried to get more precise estimates of the relative frequencies of different structures.

FINDING CORPORA

To assess syntactic form frequencies, one needs corpora. Usually, these are machine-readable text files that consist of written texts or transcriptions of human speech. Most of the corpora that have been collected are available on the internet and can be found relatively easily with the existing search engines. Ideally, a corpus must be very large and

contain a representative sample of general language. The larger the corpus, the more reliable the frequency estimates become and the more representative the corpus is for the kind of texts covered. In the early days of corpus research, corpora with one million words were considered huge; nowadays, owing to the massive availability of digital text sources, corpora can include up to a billion words. One much used source of corpora, for example, is the CD-ROMs made available on a yearly basis by newspaper publishers.

The problem of the representativeness of the corpus depends to some extent on the research question. With respect to syntactic structures there is, for instance, the fact that newspaper publishers have their articles text-edited before publication. This may raise problems for a researcher who is interested in actual usage within a certain language community (as opposed to use prescribed in grammar textbooks). Another limitation of newspapers and magazines is that they cover only a limited range of language registers (i.e. language use in a particular context). Therefore, a few corpora with a wider variety of texts have been collected for research purposes. Unfortunately, most of these corpora are rather limited in extent (to a few million words). Another source of corpora that is currently attracting a great deal of attention is the internet, where in discussion groups and in chat channels millions of sentences are produced weekly on a great variety of topics and without any stylistic supervision.

A final concern with the existing corpora is that most are based on written texts. This raises a number of issues. One is the extent to which the calculated syntactic form frequencies on the basis of written materials can be extended to spoken materials. Another is that written materials may tell us little about the language children are exposed to during their preschool years.

ESTIMATING FREQUENCIES

After the corpus has been chosen (or assembled), the sentences must be parsed in order to gain access to the different syntactic forms and frequencies. Depending on the research question, two different strategies are used. For some topics (for example, the development of an automatic sentence parser) the breadth of coverage is important. This means that the program must be able to handle a great variety of texts, but that it does not usually have to provide a full analysis of the sentences and may make occasional mistakes (in general, a two to four percent error rate is acceptable). For other topics (such as the resolution of specific syntactic ambiguities) the analysis must be complete and accurate. In this case, however, the amount of material that has to be handled is much more limited.

Thus far, there is no easy technique that produces flawless results for the parsing of large text corpora, not even parsing by humans, unless very stringent criteria are adhered to. Reporting on their first experiences with the annotation of a corpus (i.e. the addition of markers that make it easy to retrieve and analyze information about the language), Marcus *et al*. (1993) reported that trained human annotators needed 44 minutes on average to tag 1000 words (hence requiring nearly 100 working days to go through a corpus of one million words) and showed an inter-annotator disagreement of 7.2 percent. Part of the problem is that the syntactic structure of a sentence can become quite complicated when the sentence is long and contains nested structures. Another reason is that a lot of sentences are ambiguous at a purely syntactic level and can have the ambiguity removed only by looking at the meaning of the sentence or the discourse context.

The performance of the annotators in the study of Marcus and coworkers was twice as good (in terms of both speed and accuracy) when the materials were preprocessed by an automatic parser and needed to be corrected only for the mistakes made by the algorithm. As a result of this finding, the annotation of large corpora nowadays is nearly always carried out semi-automatically (that is, the sentences are first parsed by a computer program and the output is then post-edited by humans). In this case, however, care has to be taken to avoid the possibility of annotators being biased by the suggestions of the algorithm. Luckily, owing to the efforts of previous researchers, in many cases it is not necessary to annotate a new corpus. For many research issues, one can make use of an existing corpus that has already been tagged. This is done, for instance, to test linguistic and psycholinguistic hypotheses, or to measure the performance of newly developed software and to provide the training input for these computer programs.

For other research topics, researchers do not need a fully parsed corpus. Often, their question is confined to one particular syntactic structure or to a small set of words, about which they want an in-depth analysis. For such purposes, it is usually feasible (and desirable) to do the analysis by hand. Sometimes this can be achieved simply by scanning the corpus for particular words or combinations of words. An example of this type of research concerns the issue of to which syntactic form

frequencies humans are sensitive when they are parsing sentences.

TYPE FREQUENCY VERSUS TOKEN FREQUENCY

Another issue researchers have to face when they are assessing syntactic form frequencies is how to define the different categories. In frequency counts, there is always the issue of which instances to group and which to separate, because there is rarely a full mapping of verbal forms with theoretical categories. Consider the word 'that'. It can have at least three syntactic functions, as shown below:

He showed the girl that painting. (demonstrative pronoun) (2a)

He showed the girl that he was strong. (complementizer) (2b)

He showed the girl that he had just met that he was strong. (relative pronoun) (2c)

In addition, the syntactic function of 'that' in (2a) overlaps with the function of the words 'this', 'these', and 'those', raising the question whether they have to be grouped or not. This problem is known as the issue of type versus token frequency. Roughly, types refer to theory-based distinctions, whereas tokens refer to the number of occurrences of these types in a corpus (e.g. the number of times 'that' is used as a demonstrative pronoun, as a complementizer, or as a relative pronoun). Mitchell *et al*. (1995) listed some examples of syntactic form frequency distributions that differed significantly depending on how the types had been defined (a phenomenon these authors called the grain-size problem). For instance, in the examples above, it may be that the word 'that' is used much less often as a demonstrative pronoun than as a complementizer after the sequence 'he showed the girl', but the same need not be true for the more general syntactic categories introduced by the word 'that' (i.e. a noun phrase versus a complement clause).

THE USE OF SYNTACTIC FORM FREQUENCIES

Syntactic form frequencies are useful for two purposes. First, it has been shown that automatic sentence parsing algorithms (such as those needed for artificial speech perception) perform better when they take into account not only the syntactic features of the individual words and phrases but also the frequencies of the different syntactic forms given the preceding context. So, when confronted with the sentence 'She put the dress *on the rack*', the algorithm will do a better job in interpreting the ambiguous final phrase 'on the rack' when it takes into account the probability of such a prepositional phrase following the verb 'put' versus the probability of such a phrase following the noun 'dress' (as in 'She saw the dress on the rack'). By taking this probabilistic information into account, the computer may be more likely to come to the correct attachment.

Second, information about form frequencies is important to find out whether humans also make use of this kind of probabilistic information when they are parsing a sentence, and if they do, whether this information is used immediately or in a second, reanalysis stage after the initial analysis has failed. The first theories of human sentence parsing assumed such information did not play a role in the initial syntactic analysis because of the limitations in working memory capacity. More recently, researchers have argued that the build-up of the syntactic structure by the human parser cannot be understood without taking into account this type of information. Still others (e.g. Mitchell *et al*., 1995) accept an influence of syntactic form frequencies, but only at the level of the syntactic structure, not at the level of the individual words that make up the sentence. As this debate is largely based on comparisons of corpus findings with experimental reading data, assessment of syntactic form frequencies has become an important research tool in psycholinguistics as well as in (computational) linguistics.

References

Frazier L and Clifton C Jr (1996) *Construal*. Cambridge, MA: MIT Press.

Marcus MP, Santorini B and Marcinkiewicz MA (1993) Building a large annotated corpus of English: The Penn Treebank. In: Armstrong S (ed.) *Using Large Corpora*. Cambridge, MA: MIT Press.

Martin W, Church K and Patil R (1983) Preliminary analyses of a breadth-first parsing algorithm: theoretical and experimental results. In: Bolc L (ed.) *Natural Language Parsing Systems*. Berlin, Germany: Springer-Verlag.

Mitchell DC, Cuetos F, Corley MMB and Brysbaert M (1995) Exposure-based models of human parsing: evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, **24**: 469–488.

Further Reading

Biber D, Conrad S and Reppen R (1998) *Corpus Linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Mitchell DC, Cuetos F, Corley MMB and Brysbaert M (1995) Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research* **24**: 469–488.

Pickering MJ, Traxler MJ and Crocker MW (2000) Ambiguity resolution in sentence processing: evidence against frequency-based accounts. *Journal of Memory and Language* **43**: 447–475.

Tabor W, Juliano C and Tanenhaus MK (1997) Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* **12**: 211–271.

Thomas J and Short M (eds) (1996) *Using Corpora for language research*. London: Longman.

Syntax, Acquisition of

Introductory article

Stephen Crain, University of Maryland, College Park, Maryland, USA
Rosalind Thornton, University of Maryland, College Park, Maryland, USA

CONTENTS

Research shows that young children adhere to core grammatical principles early in the course of language development. Experimental studies of child language confirm the theory of universal grammar, and support the continuity assumption for language development.

INTRODUCTION

From the vantage point of linguistic theory, all normal children are expected to have full command of a rich and intricate system of linguistic principles as soon as these principles can be assessed, roughly around age three. Experimental investigations of child language, however, have often led to a different picture of language development. Several studies are interpreted as showing that language learning takes several years, and that children make numerous missteps along the way. The article begins by reviewing the reasons, based on current linguistic theory, for anticipating the rapid growth of linguistic knowledge. Then it turns to the laboratory, to consider a sample of findings that do not sit well with the expectations of linguistic theory, as well as some findings that do comport with theory.

PRINCIPLES AND PARAMETERS

Despite the complexity of human languages, children rapidly converge on a grammatical system that is equivalent to other members of their linguistic community. This remarkable acquisition scenario unfolds every day, all across the globe, yielding the truism that any human child can learn any human language. To explain the universal mastery of language in the species, linguists have sought more and more restrictive theories of the knowledge that is acquired. These theories restrict the space of possible human languages while still providing explanatory accounts of a range of linguistic phenomena. The result of these linguistic investigations is a road-map for grammar formation, a universal grammar, which establishes the boundaries on the space of possible human languages. The crossroads on the road-map must be clearly marked, so different learners can take different roads to acquire any one (or more) of the roughly 6000 existing languages in just a few years.

The road-map children use to plot a course in grammar formation enables them to project beyond their experience, rather than being securely tied to it. Children are therefore expected to form grammars that deviate in certain respects from those of adult speakers of the target language. But, like Rome, all roads lead to the same destination; at some point, children achieve a stable state that is equivalent to that of adults in the linguistic community. From this perspective, the 'errors' that arise in the course of language acquisition are not the result of defective grammars; rather, language-learners sometimes speak a foreign language (metaphorically speaking). This is the continuity assumption: the proposal that child language can differ from the language of the linguistic community only in ways that adult languages can differ from each other. As a general research strategy, advocates of the continuity assumption advance explanations of behavioral differences between children and adults which postulate minimal differences in cognitive mechanisms.

As children navigate the terrain of universal grammar, they will face forks in the road, leading to different classes of grammars. The different routes are the parameters of natural language. Presumably, there are only a finite number of

parameters, each one with a limited number (perhaps two) of innately specified values, or settings. The values of a parameter may be intrinsically ordered: if the language that is generated on one setting of a parameter properly contains the language that is generated on another setting, then the parameter must be set initially to the value that generates the smaller language. This is the subset condition on parameters. Children's adherence to the subset condition obviates the need for negative evidence that informs children that certain expressions are not permitted in the local language. It is widely acknowledged that negative evidence (e.g., corrective feedback) does not occur in sufficient quantity to be a major contributor to language development. Adherence to the subset condition ensures that parameters can be changed on the basis of positive evidence alone; that is, on the basis of input from other speakers. In cases when parameters are not constrained by the subset condition, children are free to try out various options before they settle on a grammar that is equivalent to that of an adult speaker in the linguistic community. However, the linguistic options are severely constrained in advance of any experience, so, at some point, every normal child is able to successfully stabilize on a grammar for the local language. (*See* **Learnability Theory**)

Experience matters with the principles and parameters model, just as it does with other models of language development. After all, children exposed to English learn English, and children exposed to Basque learn Basque. Once children achieve a stable state, however, there appear to be no lingering reminders of any wrong turns that children might have taken in the course of grammar formation. It is as if each child had access to all the primary linguistic data at once and had reached the stable state instantaneously. The idea of instantaneous acquisition is an idealization, of course. But the fact that the actual course of language development does not leave any indelible marks on the adult grammar is important for linguistic research. This fact permits linguists to investigate the grammars of human languages without concern for any childhood 'errors' that are made in language development. But any actual missteps in language development must somehow be set straight, to redirect children to a grammar that is equivalent to that of adults. Setting a new course requires positive evidence from speakers of the local language. Depending on the properties of the input available to children, it is conceivable that children spend some time using grammars that differ in certain respects from those of other speakers and which are more like grammars used by speakers of other human languages.

This leads us to ask how long such developmental 'stages' last. If parameter values can be revised on the basis of simple and readily available features of every child's linguistic experience, then the stages of language development should not last long at all. The logic of the situation dictates that the available evidence for parameter setting must be both simple and abundant (Lightfoot argues that all parameters can be set with simple input). If the evidence needed for parameter setting were exotic or required excessive computational resources, then some (perhaps many) children would not encounter the requisite data, or would be unable to perform the necessary computation. These children would not successfully converge on an adult grammar. Of course this does not happen: all normal children learn to speak the language of the local community. Therefore it is safe to conclude that the evidence for parameter resetting is simple and readily available. So, *ceteris paribus*, children should rapidly converge on the target grammar.

THEORY VERSUS DATA

In contrast to the expectations of linguistic theory, experimental assessments of child language paint a different picture of the course of language development. The experimental findings often suggest that language learning takes many years, and that children make numerous missteps along the way. We will review some of the recalcitrant data (from the standpoint of linguistic theory), and we will review several ways researchers have attempted to deal with them.

It has been amply demonstrated that the experimental findings do not accurately reflect children's underlying linguistic knowledge in some cases; several of children's apparent difficulties have turned out to have been experimentally induced. The data from experiments in child language have withstood scrutiny, however, in other cases. Still, there are other avenues to explore short of abandoning linguistic theory. One useful observation is that the emergence of linguistic knowledge may be impeded by aspects of child development that are known to take time, such as the accrual of real world knowledge. Or if a linguistic structure requires computations to be performed within more than a single component of the language apparatus, this could delay language development. For example, some syntactic constructions coincide with specific speech acts, and others are associated with particular stress

patterns. It could take time for children to master the so-called interface conditions that relate these particular linguistic structures to the relevant speech acts or stress patterns.

Here is the structure of the remainder of this article. First, we use the acquisition of relative clauses to illustrate how the experimental assessment of children's linguistic competence is fraught with methodological perils, sometimes leading to the erroneous conclusion that children's grammars are deficient. Then we report the findings of an experimental study that successfully probed children's knowledge of a linguistic constraint on both form and meaning, called the 'head movement constraint'. To illustrate children's difficulties at the interface of linguistic systems, we consider two examples. The first is a linguistic constraint on the interpretation of pronouns like 'him' and 'her'. In adult grammars, the interpretation of pronouns receives assistance from both phonological and pragmatic principles, but the confluence of syntactic knowledge and these nonsyntactic properties apparently takes time for some children to grasp. The second example of late acquisition is a parameter that involves children's production of Why-questions. We review the findings of a longitudinal study of one English-speaking child's Why-questions, in which both adultlike and nonadult constructions were recorded. We conclude that the child's nonadult Why-questions represent a genuine parametric option, but one for which parameter resetting requires the child to adjust the range of speech acts that are associated with auxiliary raising in English. The final discussion briefly turns to the syntax of two-year-olds to provide evidence that some parameters are set quite early in the course of language acquisition.

THE EMERGENCE OF RELATIVE CLAUSES

An example of the apparently late acquisition of syntactic knowledge was uncovered in research conducted in the 1970s. In several studies children were found to commit systematic errors in responding to sentences with a restrictive relative clause. Children's errors appeared in experiments using an act-out methodology. These studies found that four- and five-year-old children consistently acted out sentences like (1) in a nonadult fashion.

The dog pushed the sheep that jumped over the pig. (1)

When asked to act out the meaning of (1) the majority of the child subjects had the dog push the sheep and then jump over the pig. For adults, (1) is understood to mean that the sheep jumped over the pig. Children's nonadult responses led researchers to claim that children assigned a structure that was appropriate for conjoined clauses, as if sentence (1) had the structure appropriate for (2), according to which the dog, not the sheep, jumped over the pig. Accordingly, this proposal was called the *conjoined clause analysis*.

The dog pushed the sheep and jumped over the pig. (2)

Using other experimental procedures, it was demonstrated that English-speaking children and Italian-speaking children have mastery of sentences with a relative clause at a much younger age, even before their third birthday The innovation in procedure was motivated by the observation that a sentence with a restrictive relative clause, such as (1), bears two kinds of presuppositions. First, felicitous use of the noun phrase, *the sheep that jumped over the pig*, presupposes that there are at least two sheep in the conversational context. If there is only a single sheep, there is no need to add a modifier; the speaker could just as well have said 'The dog pushed the sheep.' In short, a restrictive relative clause is appropriate when some restricting needs to be done. The second presupposition of sentence (1) is that the event mentioned in the relative clause (the jumping event) should have taken place prior to the event mentioned in the main clause (the pushing event).

In research that evoked high error-rates from four- and five-year-old children, neither of these presuppositions was satisfied. Only one sheep was present in the experimental workspace for (1), and no event occurred prior to the presentation of a test sentence; the child's task was simply to 'act out' the test sentences from scratch, using objects that were placed in front of him or her in the experimental workspace. Based on these observations, two experiments were designed to satisfy the presuppositions associated with relative clauses.

In one experiment, additional sheep were added to the workspace for a sentence like (1). This simple change resulted in a much higher percentage of correct responses by children, including children as young as three. A second experiment provided even more compelling evidence that young children command the principles underlying relative clause formation. Using an elicited production technique, pragmatic contexts were constructed to satisfy the presuppositions of restrictive relatives. The task required children to distinguish (verbally)

one salient figure from a set of identical figures, by describing some action that it alone was performing. The child subjects described the figure (and what it was doing) to a blindfolded person; this person then removed the blindfold and attempted to pick out the referent of the child's description. Using the elicited production task, even children as young as 2.8 consistently produce well-formed relative clauses (e.g. *Point to the kangaroo that's eating the strawberry ice cream*). Thus, the apparent late acquisition of restrictive relative clauses can be explained away as an experimental artifact.

The moral is that it is not possible to test the predictions of linguistic theory without the assistance of appropriate experimental methodologies.

THE HEAD MOVEMENT CONSTRAINT

The next example illustrates the claim that children do not venture beyond the boundaries established by universal grammar. The discussion focuses on a linguistic constraint known as the *head movement constraint*. Interpreted more broadly, the constraint restricts children to hypotheses based on structure-dependent linguistic operations, but not those solely based on linear order. Consequently, the space of possible human languages becomes so restricted as to exclude even apparently simple rules that are compatible with much of the primary linguistic data available to children. The example is Chomsky's parade case of 'the logical problem of language acquisition', the formation of Yes/No questions.

The formation of Yes/No questions is interesting because, as Chomsky observed, the principles for forming these questions are more complex than meets the eye, at least when the focus is simple examples. A structure-independent hypothesis such as 'move the first {*is, can, will*, …} to the front of the string of words' yields the right results in forming Yes/No questions from many simple declarative sentences, as (3) illustrates.

(a) Bill can play the sax. ⇒ Can Bill play the sax?
(b) The sky is blue. ⇒ Is the sky blue? (3)

However, the simple structure-independent hypothesis 'move the first {*is, can, will*, …}' produces the wrong Yes/No questions for sentences that contain a relative clause, such as (4). In the Yes/No question corresponding to (4), the auxiliary verb, *is*, following the entire subject noun phrase (NP), *the man who is feeding a donkey*, has to move, as in (5). Moving the first occurrence of *is*, from inside the relative clause (*who is feeding a donkey*), results in a deviant Yes/No question, (6).

The man who is feeding a donkey is happy. (4)

Is the man who is feeding a donkey _ happy? (5)

*Is the man who _ feeding a donkey is happy? (6)

On one linguistic analysis, Yes/No questions are formed by 'local' movement. The auxiliary verb that moves is the inflectional head, I, of the inflection phrase (IP) projection; movement of the auxiliary verb in I takes it to the adjacent head position, C, within the CP projection (see Figure 1). Movement of phrasal heads is governed by the head movement constraint, a putatively universal linguistic principle. According to the head movement constraint, the heads of phrases can *only* move locally, as in I-to-C movement. I-to-C movement of the first occurrence of *is*, from inside the relative clause in (4), would violate the head movement constraint, because this would move the auxiliary verb past the heads of *two* phrasal projections. Metaphorically, a relative clause is an island, where linguistic elements are stranded.

We noted that the structure-independent hypothesis is consistent with much of the input children encounter, so it is reasonable to suppose that some children would adopt it unless this kind of hypothesis is ruled out in advance of any experience. Children who succumbed, however, would face three tasks in converging on the adult grammar. First, they would need to figure out that Yes/No questions are formed by I-to-C movement. This requires complex input, such as (5). But even if such complex input were available in the primary

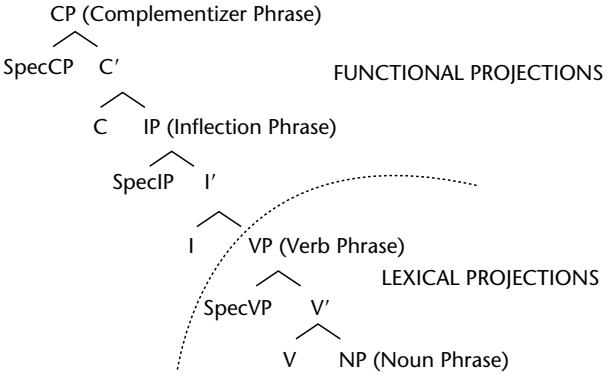


Figure 1. The hierarchy of syntactic projections (head-complement order as in English).

linguistic data, these children would also require access to negative evidence, such as corrective feedback, to accomplish the second task, which is to expunge the nonadult sentences, such as (6), from being produced by their grammars. Without negative evidence, both the local and the nonlocal movement options could coexist in children's grammars. Specifically, positive evidence (for adding 5) and negative evidence (for removing 6) would have to be available in sufficient quantities during the relevant stage of language development.

The third task children would face is that of learning the prohibition against nonlocal movement of heads: the head movement constraint. Even if children were able to rid their grammar of the unacceptable sentences, this would not explain why all languages enforce a ban on the nonlocal movement of the heads of phrases. The explanation, according to the theory of universal grammar, is that the head movement constraint is not learned from experience but is innately specified. One hallmark of innate specification is early emergence. If young children demonstrate knowledge of the constraints that characterize adult grammars, such as the head movement constraint, that would compress the acquisition problem considerably, and would correspondingly strengthen the argument for innate specification. (*See* **Innateness and Universal Grammar**)

To test young children's knowledge of the head movement constraint, an experiment was developed to elicit Yes/No questions from children. The experiment used a toy figure, Jabba the Hutt (from *Star Wars*), and some pictures. Each child was asked to pose questions to Jabba the Hutt. To elicit Yes/No questions, instructions to the child were couched in the carrier phrase, 'Ask Jabba if …' Following each question, Jabba the Hutt was shown a picture and responded 'Yes' or 'No'. The experimenter inserted a variety of declarative sentences into the carrier phrase, including (4). The result is (7).

Ask Jabba if *the man who is feeding a donkey is happy*. (7)

The outcome was as predicted by the theory of universal grammar. It was found that as soon as children could be tested, shortly after their third birthday, they never produced incorrect Yes/No questions like (7). Therefore, there is no evidence of a stage at which children produce structure-independent Yes/No questions, in violation of the head movement constraint.

Advocates of the continuity assumption contend that the head movement constraint is not learned but is innately specified as part of the theory of universal grammar. No one produces ill-formed Yes/No questions, such as (6), but no one tells children that such questions are ill-formed; yet even young children know this. By incorporating constraints on form and meaning, children's grammars go beyond the primary linguistic data along several dimensions. These phenomena are the stock in trade of nativists, because these phenomena form the basis for poverty of stimulus arguments for innate linguistic knowledge.

THE BINDING THEORY

This section continues the discussion of linguistic constraints: the principles of the *binding theory*. Like the head movement constraint, the principles of the binding theory explain why language users judge certain sentences to be unacceptable, and why they judge certain acceptable sentences to lack particular meanings. There are three principles of the binding theory, A, B, and C. Each one governs a different type of noun phrase. We will discuss two of the three principles, principles A and B. Principle A governs the placement of reflexive pronouns, as in (8). The indices i, j, k,… are used to indicate the interpretive options that are available in (8); if two expressions have the same index, then one of the expressions is said to be referentially dependent on the other. Roughly, two expressions that are related in this way are interpreted as referring to the same individual(s) in the conversational context. In (8), the reflexive pronoun *himself* is referentially dependent on the noun phrase *Papa Bear*, so the sentence means that Bill said that Papa Bear covered himself. Illicit referential dependencies are marked by an * . In (8), the reflexive pronoun *himself* cannot refer back to *Bill*, so the sentence cannot mean that Bill said that Papa Bear covered Bill. (*See* **Binding Theory**)

Bill_i said that Papa Bear_k covered $\text{himself}_{*i/*j/k}$

Principle A: a reflexive pronoun *must* be referentially dependent on a local noun phrase. (8)

The second principle of the binding theory, Principle B, governs the use and interpretation of ordinary pronouns, like *him*, *her*, and *them*. Consider the example in (9) in which the reflexive pronoun in (8) has been replaced by the accusative pronoun 'him'. Notice that the pronoun 'him' can

be referentially dependent on 'Bill' but not on 'Papa Bear'. So, (9) can mean that Bill said that Papa Bear covered Bill. It can also mean that Bill said that Papa Bear covered some other male individual, other than Papa Bear or Bill. What it cannot mean is what (8) must mean, that Bill said that Papa Bear covered himself.

Bill_i said Papa Bear_k covered $\text{him}_{i/j/*k}$

Principle B: a pronoun *cannot* be referentially dependent on a local noun phrase. (9)

Young children's knowledge of the principles of the binding theory has been the subject of much research, including cross-linguistic investigations. Regardless of methodology, the findings of experimental research reveal a strikingly similar pattern of linguistic behavior for both children and adults in response to sentences governed by Principle A, disallowing those meanings ruled out by the principle. By contrast, children and adults sometimes give different responses to sentences like (10). Some English-speaking children, for example, accept sentence (10) as a description of a story in which Papa Bear covered himself.

Papa Bear covered him (10)

On one account, children's acceptance of referential dependence between 'him' and 'Papa Bear' in (10) appeals to children's lack of real world knowledge. This proposal is based on the observation that there are special circumstances in which even adults accept an interpretation of sentences like (10) which takes the pronoun to be referentially dependent on the name. One such circumstance is where Papa Bear is behaving in a manner that is uncharacteristic of bears. In such circumstances, the interpretation in question is possible because the two NPs represent the same referent, Papa Bear, in two different guises. It is also worth noting that adults allow this interpretation only if the pronoun receives stress, as indicated by the upper case letters in (11). On this account, children must learn from experience what is characteristic or expected (e.g., of bears), and they must learn how stress is associated with interpretation.

Papa Bear covered HIM (11)

In summary, the late acquisition of syntactic knowledge has been explained by invoking real world knowledge as well as the lack of knowledge of an interface condition on the application of a linguistic principle: its association with stress.

WHY-QUESTIONS: A CASE STUDY IN CONTINUITY

As mentioned in the introduction, the continuity assumption supposes that children and adults share a common core of linguistic knowledge. Essentially, child language can differ from adult language only in ways that adult languages can differ from each other. So far, the literature that we have reviewed on child language is consistent with the continuity assumption: children do not appear to entertain grammatical hypotheses that extend beyond the boundary conditions imposed by universal grammar (UG). Advocates of a principles and parameters theory should not be surprised, however, if some English-speaking children exhibit some features of grammar that appear in other Germanic languages, Romance or East Asian languages, in the absence of evidence for these properties in the primary linguistic data. Evidence of children's nonadult (but UG-compatible) productions may be the strongest argument for the theory of universal grammar. An argument of this kind can be made using evidence from the acquisition of Why-questions.

Some languages make finer distinctions than others do. In some languages, Why-questions are found to show a different pattern of word order from questions formed with other Wh-phrases (e.g., Who-questions, What-questions). For example, in French, *pourquoi* (why) cannot remain *in situ* in questions, and cannot undergo stylistic inversion. In Spanish, and Basque also, Why-questions show a different pattern; in contrast to other Wh-phrases, adjacency is not maintained between Why-phrases and the verb. Whatever underlies the (semantic) distinction among Wh-phrases in these other languages, the distinction is clearly blurred in the grammar of English. Why-questions are formed in the same way as other questions; they exhibit I-to-C movement and, therefore, manifest the same pattern of word order.

The kind of 1-to-1 mapping of form and meaning that is characteristic of French, Basque, and Spanish is also found in the language of English-speaking children. That is, English-speaking children often differentiate Why-questions from other questions, by failing to carry out I-to-C movement in their one-clause Why-questions. The following examples are from a transcription of the speech of a child, A.L., at 3.5 years. All of the examples were from a 3-week period.

Why you have your vest on?
Why she doesn't have any hanger?
Why he's woofing?
Why that guy has tookened Walker?
Why he's following the guy?
Why that kind of thing could break? (12)

In contrast to these one-clause Why-questions, which lacked I-to-C movement, during the same period A.L.'s two-clause Why-questions consistently showed I-to-C movement. One explanation of this fact is that in two-clause questions, the syntax forces the auxiliary verb to raise from I to C. In two-clause questions, such as (13), the moved Why-phrase moves over a sentence barrier, in contrast to one-clause Why-questions, as in (12), where *Why* may be generated in place, and not moved to sentence-initial position.

Why do you think he goed in the paddling pool?
Why do you think that Gonzo went in that truck? (13)

In short, A.L. was clearly able to produce adult-like structures, with I-to-C movement of the auxiliary verb; this was characteristic of A.L.'s more complex Why-questions, but not of the simpler Why-questions. Before the experimental tools were developed to elicit complex questions from children, accounts of question formation in early child language were often misdirected by placing undue emphasis on their inability to perform I-to-C movement. Experience presumably informs children like A.L. that one-clause Why-questions should not be distinguished from two-clause Why-questions. But this takes time, because A.L.'s grammar distinguishes between them. The existence of abundant counter-examples in the input from English-speaking adults must seem inexplicable to children like A.L.

To recap, the case of Why-questions offers further evidence of delayed acquisition but is also in keeping with the continuity assumption. The account we have presented appeals, once again, to the natural seams (or parameters) of natural language. According to the continuity assumption, child language should develop along these seams, though child language can diverge from that of adults until the pressure of experience ultimately propels them towards a grammar that is equivalent to that of adults, but children are never expected to try out a language that violates core principles of universal grammar.

FIRST STEPS IN LANGUAGE ACQUISITION

The younger children are, the more their grammars appear to differ from those of adults. When children first begin to speak, their utterances are composed of just one or two words and, in the absence of context, it is often difficult to determine the meanings they are attempting to convey. One might suppose that this early stage of language development poses a serious threat to the continuity assumption. Examination of cross-linguistic data, however, especially of languages with rich morphological systems, leads to a picture that is consistent with the continuity assumption.

In English, syntactic categories like nouns and verbs carry little morphological information, making it difficult to assess children's knowledge of the morphological properties of the language. The observation that young children often leave off what little morphology English has (for example, the third-person singular *-s* in *he walks*, the past tense *-ed* in *he walked*, the genitive *-s* of *mommy's sock*, and so on) contributed to an initial pessimism about the continuity assumption. Based on children's omissions, some researchers proposed that the grammars of young children contain only lexical (content-bearing) projections such as verb phrase (VP) and noun phrase (NP) but do not contain functional projections that carry grammatical information, such as inflection phrase (IP, which carries tense and agreement) and complementizer phrase (CP, which hosts question words). More specifically, one claim was that functional projections mature, being biologically timed to emerge later in language development.

In contrast to English, Dutch, German, and Italian are languages with rich morphological paradigms. The results of research on the acquisition of these languages weigh in against the idea that children's grammars initially generate only partial syntactic phrase markers. Although children learning these languages do make errors of omission, as do their English-speaking counterparts, when children learning these other languages use inflection, overwhelmingly they use it correctly. Several studies have documented these facts for children under 2.6. For example, a study of person inflection by three Italian-speaking children found the percentage of correct usage at between 97 to 99 percent. Similarly, a study by Poeppel and Wexler of a German-speaking child (age 2.1) revealed

agreement errors of only 3 percent. Children's correct use of agreement is compelling evidence that they have access to the functional projection, IP, where information about agreement is maintained. The same study revealed mastery of linguistic principles that involve both the IP projection and the CP projection. For example, the child Andreas proved to be aware of the fact that, in German, (1) finite verbs must be raised to C and must be preceded by some other constituent (the verb second, or V2 effect), and that (2) infinitival forms remain in final position. To demonstrate that Andreas moved finite verbs to C, they analyzed utterances of at least three words. This ensured that a verb in C position could be distinguished from a verb in final position. Aged 2.1, Andreas produced finite verbs in C position in 95 percent of the utterances that were analyzed. In short, the cross-linguistic data support the view that the full inventory of functional projections is available in early grammars.

These German data also provide compelling evidence that the verb movement parameter is set early. This parameter encodes variation in the position in the phrase structure in which a finite verb appears. In some languages (e.g., German), the verb is in the Complementizer position; in others (e.g., French) it is positioned in the Inflection position, or it can remain inside the verb phrase (as in English). Data from languages other than English provide important evidence that indicates children set the relevant parameters early. For example, young children acquiring French can be shown to know that finite verbs appear to the left of negation (= *pas*) in Inflection (that is, in I), and infinitival forms must appear to the right of the negative form. In a study of young French-speaking children, finite verbs appeared to the left of *pas*, as in (14a), 97 percent of the time. If a verb appeared to the right of *pas*, it was an infinitive, as in (14b).

a. elle roule pas (Grégoire 1.11)
she roll not
'It doesn't roll'
b. pas rouler en vélo (Philippe 2.2)
not roll-INF on bike
'Not go by bike' (14)

It is more difficult to establish what young English-speaking children know about verb movement. It is not easy to evaluate this by investigating negation, because there is a complicating factor; in negative sentences, the dummy verb *do* appears to the left of negation (e.g., *He does not like cheese*). However, even before children acquire do-support, their production of sentences with negation suggests that they know that main verbs remain inside the verb phrase. Negative imperatives, as in (15), also illustrate this fact.

Not sing in the car! (A.L 1.11)
Not go here! (15)

In early child grammars, there are one or two apparent counterexamples to the conclusion that children's nonadult productions are mainly errors of omission, and not errors of substitution. One of these is the phenomenon known as 'optional infinitives' or 'root infinitives'. Although children learning languages like Dutch, German, and French appear to know that finite verbs must raise (in accordance with the verb movement parameter), they do not always raise the verb to the appropriate position. Instead, they sometimes produce utterances with a (nonfinite) infinitival form of the verb. The error is shown for French in (14b), and for German in (16).

du das haben (Andreas 2.1)
you that have-INF
'You have that' (16)

English has no special infinitival form; the dictionary form of the verb (e.g., *to eat*) is the same as the verb stem (i.e., the form of the verb when it is stripped of inflection). Given this fact, combined with the word order facts of English, it is difficult to see whether English-speaking children also produce infinitival forms. It has been argued, however, that their frequent failure to use the third-person *-s* ending and the past tense *-ed* inflection is a manifestation of the same phenomenon.

Children's optional infinitive errors are a potential problem for the continuity assumption, because root infinitives are not a core phenomenon of adult languages. On one current account, optional infinitives arise because the young child's grammar has not matured, and children can optionally omit tense or agreement features from a phrase structure representation. Omission of agreement results in an utterance with default (accusative) case; omission of tense results in children's failures to produce present or past tense marking. Another account attempts to further minimize the differences between child and adult grammars. The only difference on this account is that children do not require clauses to project up to the CP level; until this knowledge matures, children allow truncated structures. A third account is more in keeping with the continuity assumption: children's syntax is intact but optional infinitives result from a processing bottleneck

that sometimes prevents children from merging the inflection and the main verb. Only time will tell whether the optional infinitive phenomenon represents a genuine violation of the continuity assumption or if it can be squared with it.

CONCLUSION

Empirical investigations of the acquisition of syntactic knowledge have reached a new level of maturity in recent years. If developments in methodology and theory keep pace, we will continue to achieve a deeper understanding of children's universal mastery of syntactic competence, and we will find out if the continuity assumption can be maintained.

Further Reading

Chomsky N (1971) *Problems of Knowledge and Freedom*. New York, NY: Pantheon Books.

Chomsky N (1975) *Reflections on Language*. New York, NY: Pantheon Books.

Crain S (1991) Language acquisition in the absence of experience. *Behavioral and Brain Sciences* **14**: 597–650.

Crain S and Nakayama M (1987) Structure dependence in grammar formation. *Language* **63**: 522–543.

Crain S and Pietroski P (2000) Nature, nurture and Universal Grammar. *Linguistics and Philosphy*.

Crain S and Thornton R (1998) *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.

Crain S, McKee C and Emiliani M (1990) Visiting relatives in Italy. In: Frazier L and de Villiers J (eds) *Language Processing and Language Acquisition*, pp. 335–356.

Grodzinsky Y and Reinhart T (1993) The innateness of binding and the development of coreference: a reply to Grimshaw and Rosen. *Linguistic Inquiry* **24**: 69–103.

Guasti MT (1993) Verb syntax in Italian child grammar: finite and non-finite verbs. *Language Acquistion* **3**: 1–40.

Hamburger H and Crain S (1982) Relative acquisition. In: Kuczaj S (ed.) *Language Development: Syntax and Semantics*, pp. 245–274. Hillsdale, NJ: Lawrence Erlbaum.

Heim I (1998) Anaphora and semantic interpretation: a reinterpretation of Reinhart's approach. In: Sauerland U and Percus O (eds) *MIT Working Papers in Linguistics*, vol. 25, pp.205–246.

Lasnik H and Crain S (1985) On the acquisition of pronominal reference. *Lingua* **65**: 135–154.

Lebeaux D (1988) *Language Acquisition and the Form of the Grammar*. Doctoral dissertation, University of Massachusetts Amherst.

Lightfoot D (1991) *The Development of Language: Acquisition, Change and Evolution*. Oxford, UK: Blackwell.

McDaniel D, Cairns H, and Hsu J (1990) Binding principles in the grammars of young children. *Language Acquisition* **1**: 121–139.

McKee C (1992) A comparison of pronouns and anaphors in Italian and English acquisition. *Language Acquisition* **2**: 21–54.

Pierce A (1992) *Language Acquisition and Linguistic Theory: A Comparative Analysis of French and English Child Grammars*. Dordrecht, Netherlands: Kluwer.

Pinker S (1984) *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Poeppel D and Wexler K (1993) The full competence hypothesis of clause structure in early German. *Language* **69**: 1–33.

Radford A (1990) *Syntactic Theory and the Acquisition of English Syntax: The Nature of Early Child Grammars of English*. Oxford, UK: Basil Blackwell.

Rizzi L (1990) *Relativized Minimality*. Cambridge, MA: MIT Press.

Rizzi L (1993) Some notes on linguistic theory and language development: the case of root infinitives. *Language Acquisition* **3**: 371–395.

Sheldon A (1974) The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior* **13**: 272–281.

Schütze C and Wexler K (1996) Subject case licensing and English root infinitives. In: Stringfellow A, Cahana-Amitay D, Hughes E and Zukowski A (eds) *Proceedings of the 20th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Tavakolian S (1981) The conjoined-clause analysis of relative clauses. In: Tavakolian S (ed.) *Language Acquisition and Linguistic Theory*, pp. 167–187. Cambridge, MA: MIT Press.

Thornton R and Wexler K (1999) *Principle B, VP Ellipsis and Interpretation in Child Grammar*. Cambridge, MA: MIT Press.

Thornton R (2000) Let's change the subject: focus movement in early grammar. Ms., University of Maryland.

Travis L (1984) *Parameters and Effects of Word-Order Variation*. Doctoral dissertation. Cambridge, MA: MIT Press.

Uriagereka J (1999) Minimal restrictions on Basque movements. *Natural Language and Linguistic Theory* **17**: 403–444.

Vainikka A (1993) Case in the development of English syntax. *Language Acquisition* **3**: 257–325.

Wexler K (1994) Finiteness and head movement in early child grammars. In: Lightfoot D and Hornstein N (eds) *Verb Movement*, pp. 305–351. Cambridge, UK: Cambridge University Press.

Wexler K (1998) Very early parameter setting and the Unique Checking Constraint: a new explanation of the optional infinitive stage. *Lingua* **106**: 23–79.

Syntax

Introductory article

Colin Phillips, University of Maryland, College Park, Maryland, USA

CONTENTS

Syntactic theory aims to explain how people combine words to form sentences, and how children attain knowledge of sentence structure.

GOALS OF SYNTACTIC THEORY

Syntactic theory aims to provide an account of how people combine words to form sentences. A common feature of all human languages is that speakers draw upon a finite set of memorized words and morphemes (i.e. minimal meaning-bearing elements) to create a potentially infinite set of sentences. This property of *discrete infinity* allows speakers to express and understand countless novel sentences that have never been uttered before, and hence forms the basis of the creativity of human language. Syntactic theory is concerned with what speakers know about how to form sentences, and how speakers acquire that knowledge.

For example, speakers of English know that 'dogs chase cats' and 'cats chase dogs' are possible sentences of English, but have different meanings. Speakers know that 'chase dogs cats' is not a possible sentence of the language, and that 'cats dogs chase' is possible in specific discourse contexts, as in 'cats, dogs chase, but mice, they flee'. Speakers' knowledge of possible word combinations is often referred to as the *(mental) grammar*.

An accurate model of a speaker's knowledge of his or her language should minimally be able to generate all and only the possible sentences of the language. For this reason, syntactic theory is often known as *generative grammar*. In the 1950s, early attempts by Noam Chomsky and others to create explicit generative grammars quickly revealed that speakers' knowledge of syntax is a good deal more complex than had been anticipated. Research on syntactic theory has relied primarily upon speakers' intuitive judgments about the well-formedness ('grammaticality') of sentences of their language. Since grammaticality judgments can be gathered relatively easily, syntactic theory has amassed a large database of findings about an ever more diverse set of languages.

The complexity of syntactic knowledge sharpens the problem of how language is learned. Research on language acquisition has demonstrated that children know much of the grammar of their language before they are old enough to understand explicit instruction about grammar. Therefore, a primary challenge for syntactic theory has been to understand how a child can learn any language, relatively effortlessly, and without explicit instruction. Research on comparative syntax has met this challenge by seeking to characterize human languages in terms of universal syntactic properties, which may reflect the child's innate knowledge, and non-universal clusters of syntactic properties that pattern together across languages, and hence may be learned as a group. Thus, the study of comparative syntax and the study of language learning are closely related. (*See* **Innateness and Universal Grammar**)

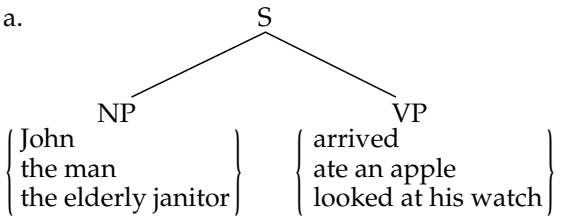
FUNDAMENTALS OF SYNTACTIC THEORY

Discrete Infinity

Almost all accounts of the discrete infinity property of natural language syntax start from the notion that sentences consist of more than just sequences of words. In the minds of speakers and listeners, sentences are hierarchically structured representations, in which words are grouped together to form phrases, which in turn combine to form larger phrases. For example, a minimal sentence of English, such as 'John arrived', contains a subject and a predicate, but the roles of subject and predicate may be replaced by phrases of arbitrary complexity. By representing possible subjects and predicates as *noun phrases* (NPs) and *verb phrases* (VPs)

respectively, the structure of many possible sentences (S) can be captured. This basic 'template' for sentences of English can be expressed as a tree structure, as in (1a), or as a phrase structure rule, as in (1b).

a.



b. S → NP VP (1)

Just as rules like S → NP VP provide templates for sentences, templates can also be specified for the internal structure of noun phrases, verb phrases, and many other phrase-types. Even a small number of phrase structure rules and a small lexicon can generate large numbers of sentences. With only the five phrase structure rules in (2) and a 30-word lexicon (consisting of 10 nouns, 10 determiners, and 10 verbs) 122,100 different sentences can be generated.

S → NP VP
VP → V NP
VP → V
NP → Det NP
NP → N (2)

Rules that allow a phrase to be embedded inside another phrase of the same type are known as *recursive* rules. Coordination (3), modification (4), and sentential complementation (5) all involve recursion. They can thus be invoked arbitrarily many times in a single sentence. Such rules increase the expressive power of the grammar from merely vast to clearly infinite. There are obvious practical limitations on the length and complexity of naturally occurring sentences, but such limitations are typically attributed to independent limitations on attention and memory.

NP → NP Conj NP
VP → VP Conj VP
Conj → *and* (3)

VP → VP PP
NP → NP PP (4)

VP → V S′
S′ → Comp S
Comp → *that* (5)

Although the rules listed in (1)–(5) fall far short of the expressive power of English, even this small fragment shows how natural language syntax uses finite means to generate infinitely many sentences. (*See* **Phrase Structure and X-bar Theory**)

Motivating Structures: Constituency

The syntactician's toolbox includes a number of structural tests that can be used as aids in diagnosing sentence structures; for example, *constituents* of sentences can generally be conjuncts in coordinate structures, as is shown for NPs and VPs in (6a, b). Other tests that show the constituency of VPs include substitution of the expression 'do so' for a VP (7a), and fronting of the VP to a clause-initial position (7b).

a. Wallace fetched[$_{NP}$ the cheese] and [$_{NP}$ the crackers]
b. Wallace[$_{VP}$ sliced the cheese] and [$_{VP}$ opened the crackers] (6)

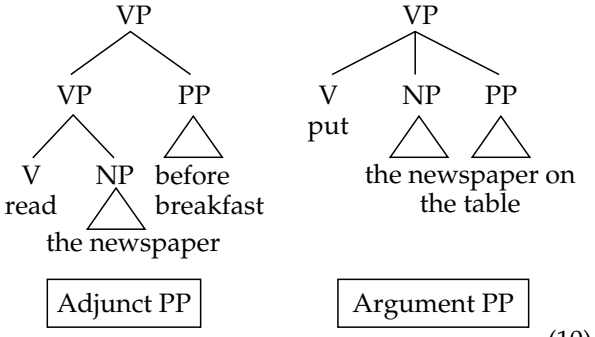
a. Wallace [$_{VP}$ read the newspaper] and Gromit [$_{VP}$ did so] too.
b. Wallace wanted to [$_{VP}$ impress Wendolene], and [$_{VP}$ impress Wendolene] he did. (7)

Constituency tests like those shown in (7) can be used to demonstrate that prepositional phrases (PPs) that are adjuncts (i.e. optional phrases) of VP recursively expand the VP, whereas PPs that are arguments of the verb (roughly, required phrases) do not. (8) shows that when *do so* substitution applies to a VP containing an adjunct-PP, the PP may be targeted or ignored by *do so* substitution. This indicates that there is a smaller VP-constituent that excludes the adjunct-PP. In contrast, (9) shows that an argument-PP cannot be ignored by *do so* substitution. If the argument-PP is 'stranded' by substitution (9b), the result is ungrammatical (indicated by the asterisk). This indicates that argument-PPs are contained within the smallest VP constituent. These contrasts motivate the VP-structures shown in (10).

a. Wallace [$_{VP}$[$_{VP}$ read the newspaper] [$_{PP}$ before breakfast]], and Gromit [$_{VP}$[$_{VP}$ did so]] too.
b. Wallace [$_{VP}$[$_{VP}$ read the newspaper] [$_{PP}$ before breakfast]], and Gromit [$_{VP}$[$_{VP}$ did so]] [$_{PP}$ at lunchtime] (8)

a. Wallace [$_{VP}$ put the newspaper [$_{PP}$ on the table]], and Gromit [$_{VP}$ did so] too.
b. *Wallace [$_{VP}$ put the newspaper [$_{PP}$ on the table]], and Gromit did so [$_{PP}$ on the floor]. (9)

Adjunct PP : VP → VP PP
Argument PP : VP → V (NP) PP

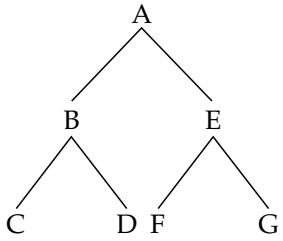


(10)

Motivating Structures: Hierarchy

In addition to tools that show which groups of words form constituents, other tests diagnose the *hierarchical* relations among positions in a structure. A striking finding of syntactic research is that many grammatical phenomena are sensitive to a structural relationship known as *c-command*, which is similar to the logical notion of *scope*. A node c-commands its sister and any nodes contained inside its sister. Thus, in the structure in (12), node B c-commands its sister, node E, and nodes F and G contained inside its sister. On the other hand, node C c-commands its sister, node D, and no others.

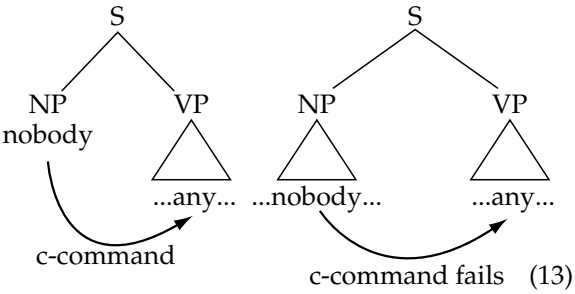
C-Command
A node c-commands its sister and all nodes dominated by its sister. (11)



(12)

One syntactic phenomenon that is sensitive to the c-command relation involves *Negative Polarity Items* (NPIs), such as 'anybody' or 'ever', which are only possible when they are c-commanded by an appropriate licenser, typically a negative expression such as 'not' or 'nobody'. NPIs are possible when the negative expression c-commands the NPI (13a–d) but are impossible when the negative expression fails to c-command the NPI (13e–h), because the negative expression is embedded inside a subject NP.

a. Wallace didn't find any cheese.
b. Nobody found any cheese.
c. Wallace didn't think that he would ever refuse cheese.
d. Nobody thought that Wallace would ever refuse cheese.
e. *[$_{NP}$ The fact that Wallace didn't like the cheese] amazed anybody.
f. *[$_{NP}$The fact that nobody liked the cheese] amazed anybody.
g. *[$_{NP}$ The person that Wallace didn't notice] thought that Gromit would ever return.
h. *[$_{NP}$The person that nobody noticed] thought that Gromit would ever return.



(13)

Multiple Roles: Transformations

In any sentence, each word or phrase occurs in a unique position in the linear sequence of words. However, many words and phrases appear to participate in *multiple* structural relations in the sentence. A central concern of syntactic research since the 1950s has been to understand how individual phrases can assume multiple roles in a sentence.

The multifunctionality of phrases has been most fully explored in the case of NPs. First, speakers represent the *thematic role* of each NP in a sentence. *Agent* thematic roles are canonically realized on subject NPs, *theme* thematic roles are canonically realized on direct object NPs, and thematic roles such as *goal*, *beneficiary*, or *location* are canonically realized inside PPs (14).

Wallace sent the flowers to Wendolene
agent *theme* *goal* (14)

However, it is important to distinguish thematic roles from *grammatical relations* such as *subject* and *object*, since thematic roles can be realized in different grammatical relations. The theme argument of the underlined verb 'steal' is realized as a direct object in the active sentence in (15a), as a subject in the passive sentence in (15b), and as the subject of a higher clause in the raising construction in (15c).

a. The penguin stole the diamond.
theme *active*

b. The diamond was stolen by the penguin.
theme
c. The diamond seemed to have been stolen.
theme

It is also important to distinguish thematic roles and grammatical relations from the structural positions in which NPs receive their *scopal* interpretation. In English, the scope of *wh*-phrases such as *who*, *what*, or *which boy* is marked by their surface position: they are positioned sentence-initially in direct *wh*-questions (16a), and in clause-initial position of an embedded clause in indirect *wh*-questions (16b).

a. Which story did the teacher know that the children like?
theme
direct object
matrix scope
b. The teacher knew which story the children always like.
theme
direct object
embedded scope (16)

In English the scope of quantified NPs such as 'every boy' is not encoded in surface word order, with the consequence that sentences like (17) are scopally ambiguous. However, in other languages, such as Hungarian and Japanese, the scope of quantificational NPs is often fixed by surface word order.

Some student answered every question correctly.
some > every: *a single student got all the questions correct*
every > some: *for each question, there was at least one student who got it right* (17)

Since speakers must distinguish for any NP a representation of its thematic role, its grammatical relation, and its scope, it is important to ask how these different structural positions are related to one another. One well-known answer to this question is that speakers represent multiple structures for each sentence, one structure each for thematic role assignments, grammatical relations, and scope relations. The thematic representation is typically known as the *deep structure* (or *d-structure*) of the sentence, and it is related to the other levels of representation – *surface structure* (or *s-structure*) and *logical form* (LF) – by transformational processes which move NPs from thematic positions to positions associated with grammatical relations, and scope positions. This *Transformational Grammar* approach was proposed by Chomsky in the 1950s and has been elaborated and modified by many linguists since that time.

(18) shows a sample transformational derivation of a *wh*-question. The NP 'who' receives the theme thematic role as the sister of the verb in the d-structure representation in (18a). In (18b) it moves to subject position of the embedded passive clause. Finally, in (18c) it moves from subject position to its clause-initial scope position. Following a now standard elaboration of the transformational approach, (18) represents the residue of movement as phonologically null elements known as *traces* (*t*).

a. the announcer believed had been elected who
b. the announcer believed who_i had been elected t_i
c. who_i did the announcer believe t_i had been elected t_i (18)

Some alternative approaches share with Transformational Grammar the assumption that sentence structure representations consist of multiple levels, but deny that there is a derivational relationship among the levels of representation (e.g. Lexical Functional Grammar, Categorial Grammar). Other approaches assume that the thematic, grammatical relation, and scope properties of an NP are all represented in a single enriched phrase structure, but that only one of these positions is normally encoded phonologically (e.g. Head-Driven Phrase Structure Grammar: see further below).

Types of Dependencies

Speakers' syntactic representations encode a variety of structural relations among words and phrases. The most basic relations are the groupings of words and phrases that form hierarchical constituent structures. These relations can be expressed entirely in terms of *sisterhood* and *dominance* among phrase structure nodes. The discussion above has introduced a number of other syntactic relations, or *dependencies*.

The syntactic relation between a pronoun or reflexive and the NP from which it takes its reference is known as a *binding* dependency. Such dependencies have been studied extensively under the rubric of *Binding Theory*.

a. Wallace_i likes himself_i *local*
b. Wallace_i thinks that Wendolene likes him_i *nonlocal* (19)

(*See* **Binding Theory**)

In *VP-ellipsis* constructions the VP in the second conjunct is dependent on the VP in the first conjunct for its interpretation (20). Transformational analyses of *wh*-questions (21a) and relative clauses (21b) treat the relationship between the *wh*-phrase and the trace as a binding relation between a *wh*-operator and a variable.

Wallace $[_{VP}$ likes cheese$]_i$ and Gromit does $[_{VP}\ ]_i$ too. (20)

a. Who$_i$ did the voters elect t_i
b. The man who$_i$ the voters elected t_i (21)

(*See* **Ellipsis**)

A leading question in research on referential dependencies involves how closely related the different types of referential dependencies are: does each type of dependency follow independent principles, or do they follow the same principles?

CONSTRAINTS ON DEPENDENCIES

Wh-movement and related phenomena have been among the most extensively investigated topics in syntactic research, giving rise to a wealth of findings. By virtue of the length of *wh*-dependencies, syntacticians can manipulate which structural positions participate in the *wh*-dependency, and which structural positions the dependency crosses. *Wh*-dependencies have thus served as a kind of 'magnifying glass' for the investigation of syntactic dependencies.

Wh-dependencies can span many clauses, in fact arbitrarily many clauses, and thus they are often referred to as *unbounded dependencies*. In (22), the *wh*-phrase has been *extracted* from a number of embedded clauses, each of which is the *complement* (direct object) of the next higher verb.

Which candidate$_i$ did the court fear [that the public might conclude [that the voters had elected t_i]] (22)

However, in a tradition of research beginning with influential work in the late 1960s by John Robert Ross, it has been found that there are many syntactic environments which *wh*-extraction cannot cross. Following Ross's terminology, the environments that block extraction are known as *islands*, and restrictions on extraction are known as *island constraints*.

Relative clauses create islands for extraction (23a), as do indirect questions (23b), complements of NPs (23c), subjects (23d), and adjunct clauses (23e). Extraction from definite NPs or NPs with a possessor is highly marked, although indefinite NPs create no such difficulties (24). If a phrase is extracted from one conjunct of a coordinate structure, it must also be extracted from the other conjuncts (25a, b).

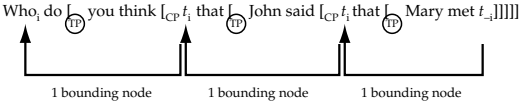
a. *Who$_i$ did the court upset the voters [who favored t_i]
b. *Who$_i$ did Bill wonder [whether his new outfit would shock t_i]
c. *What$_i$ did Sarah believe [the rumor that Ed was willing to spend t_i]
d. *Who$_i$ did [the fact that the president nominated t_i] upset the opposition party?
e. *What$_i$ did Wallace eat the cheese [while he was reading t_i] (23)

Who$_i$ did Sally hear a? the *Helen's story about t_i? (24)

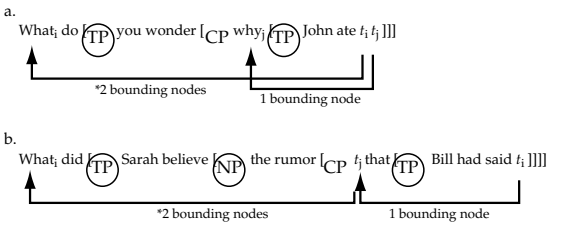
a. *What$_i$ did [Gromit read the newspaper] and [Wallace eat t_i]
b. What$_i$ did [Gromit read t_i] and [Wallace eat t_i] (25)

The examples in (22)–(25) raise the question of why *wh*-dependencies can be arbitrarily long as in (22), but not as in (23)–(25). Work on this question led to the proposal that, contrary to appearances, all *wh*-dependencies are *local*. If all *wh*-movement is local, proceeding one clause at a time, then apparent long-distance *wh*-movement turns out to be a series of local movements, each one targeting a landing site in the next higher CP (*Complementizer Phrase*), as in (26). If all *wh*-movement must be local, then it follows that relative clauses and embedded *wh*-questions create islands, because these are cases in which an intermediate landing site of movement is already filled (27a).

The best-known implementation of this proposal is the *Subjacency Constraint*, proposed by Chomsky in the early 1970s. In its original formulation the constraint blocks any *wh*-dependency that spans more than two *bounding nodes*, where the bounding nodes are defined as NP and TP (*Tense Phrase*: a more recent term for the 'S' node, recognizing *Tense* as the head of a clause). This formulation also explains the complex NP constraint violation in (23c) and (27b).



(26)



(27)

The proposal that a long-distance *wh*-dependency involves a sequence of local dependencies receives interesting support from a number of languages that show a syntactic residue of local movement. In certain varieties of Spanish, for example, subject–auxiliary inversion occurs in every clause in the path of *wh*-movement ((28): compare this to the English translation, in which inversion occurs only in the highest clause).

a. Juan pensaba que Pedro le había dicho que la revista había publicado ya el articolo.
Juan thought that Pedro him had told that the journal had published already the article
'Juan thought that Pedro had told him that the journal had published the article already.'

b. Qué pensaba Juan que le había dicho Pedro que había publicado la revista?
What thought Juan that him had told Pedro that had published the journal
'What did Juan think that Pedro had told him that the journal had published?'

c. *Qué pensaba Juan que Pedro le había dicho que la revista había publicado?
What thought Juan that Pedro him had told that the journal had published (28)

The island constraints restrict the nodes that a *wh*-dependency may cross. All of the examples presented so far involve extraction of a direct object *wh*-phrase. In addition, subject and adjunct *wh*-phrases in English are subject to tighter restrictions than object *wh*-phrases. For example, extraction of an embedded direct object *wh*-phrase is possible, irrespective of whether the embedded clause contains an overt complementizer 'that' (29a, b). However, extraction of an embedded subject *wh*-phrase is impossible if the complementizer is overt (29c). This constraint is known as the *that-trace constraint*, and it has been observed in many languages, as discussed further below.

a. Who$_i$ do you think t_i that John met t_i?
b. Who$_i$ do you think t_i John met t_i?
c. *Who$_i$ do you think t_i that t_i met John?
d. Who$_i$ do you think t_i t_i met John? (29)

There are also differences in extraction possibilities between argument *wh*-phrases such as 'what' and 'which books', and adjunct *wh*-phrases such as 'why' and 'how' (see Further Reading).

A long-standing goal of syntactic research on unbounded dependencies has been to uncover a set of general principles that can explain the full variety of constraints on *wh*-dependencies. Although there have been many different attempts to unify the constraints on movement, two observations have been pervasive, and have featured in many different theories. First, if movement is required to be *local*, then it is subject to *intervention effects*, when a required landing site of movement is occupied by another element. Second, movement paths that include noncomplement nodes (subjects or adjuncts) are consistently more restricted than paths that include only complement nodes (sisters of heads). (*See* **Constraints on Movement**)

CROSS-LANGUAGE SIMILARITIES AND DIFFERENCES

A fully general theory of the mental representation of syntax clearly must handle the facts of all human languages. In addition, cross-linguistic investigations are important to accounts of how natural language syntax is learnable. *Universals* of syntax, or *principles*, may be part of the child's innate endowment, and thus not need to be learned. *Non-universal* syntactic properties must also be learnable within the constraints imposed by the time and evidence available to the child. When a set of syntactic properties *covaries* across languages, it is possible that the learner only needs to learn one member of the set of properties in order to draw appropriate conclusions about the entire set. Thus, an important goal of cross-linguistic syntax research is to find clusters of covarying syntactic properties, or *parameters*. This *Principles and Parameters* (P&P) approach to syntax has been most intensively investigated in transformational approaches to syntax but it can be applied equally well to other syntactic approaches. (*See* **Government–Binding Theory**)

Research on comparative syntax has discovered a number of striking cross-linguistic parallels between languages that appear very different on the surface. An example from Mohawk serves as an illustration.

One constraint on pronouns in English is that a pronoun cannot co-refer with an NP that it c-commands. This constraint ('Binding Condition C') accounts for the contrast between (30a) and (30b). The pronoun inside the subject NP fails to c-command the direct object in (30a), thereby allowing co-reference. On the other hand, the subject pronoun c-commands the name inside the object NP in (30b), thereby preventing co-reference.

a. [$_{NP}$ The book that he$_i$ bought] offended John$_i$
b. *He$_i$ bought [$_{NP}$ the book that offended John$_i$] (30)

Unlike English, which exhibits strict subject–verb–object (SVO) word order, Mohawk, an Iroquoian language spoken in Quebec and upstate New York, exhibits free word order, allowing all six possible permutations of subject, verb, and object, and also allows 'discontinuous constituents' in the form of split noun phrases (31). Based on such properties, languages like Mohawk have sometimes been described as 'nonconfigurational'. However, Mark Baker has demonstrated that Mohawk exhibits similar configurational asymmetries to English, as the contrast in (32) shows. This contrast can be explained by Binding Condition C, just as in English, provided that we attribute some degree of underlying configurational structure to Mohawk sentences.

Ne kíke wa-hi-yéna-' ne kwéskwes
NE this FACT-1sS/MsO-catch-PUNC NE pig
'I caught this pig.' (31)

a. Wa-ho-nakuni-'
tsi Sak wa-hi-hrewaht-e'
FACT-NsS/MsO-anger-PUNC
that Sak FACT-1sS/MsO-punish-PUNC
'That I punished Sak$_i$ made him$_i$ mad.'
(co-reference possible)
b. Wa-shako-hrori-'
tsi Sak wa-hi-hrewaht-e'
FACT-MsS/FsO-tell-PUNC
that Sak FACT-1sS/MsO-punish-PUNC
'He$_i$ told her that I punished Sak$_i$.'
(co-reference impossible) (32)

The similarity between English and Mohawk is striking, given how different the languages appear on the surface. Furthermore, evidence from many other languages suggests that Binding Condition C is a universal of natural language syntax, which may be part of innate linguistic knowledge. Consistent with this suggestion, studies by Stephen Crain and his colleagues have shown that children exhibit knowledge of Binding Condition C by their third birthday, which is as early as it has been possible to test this knowledge. This finding has been replicated in children learning other languages (e.g. Italian, Dutch, Russian). It is particularly encouraging news for a child learner of Mohawk that Binding Condition C need not be learned, since it is unlikely that the presence of the constraint could be guessed from the input to the Mohawk child.

Although syntactic research has uncovered many universals of language, there are clearly many properties that vary across languages and hence must be learned. The search for parametric clusters of syntactic properties has turned up a number of cross-language correlations. The most useful correlations are those that link abstract (and hence difficult-to-observe) syntactic properties with more easily observable syntactic properties. For example, the *that-trace* constraint introduced above does not apply in all languages: it applies in English (33a), but not in Italian (33b). This is not easily inferred from the language input to children, since it is not easy to observe the *absence* of a particular construction. However, the availability of *that-trace* sequences correlates cross-linguistically with the availability of postverbal subjects, which are readily observable in the input to the learner. Italian allows postverbal subjects (34b), but English does not (34a). This connection has been reinforced based on the study of many other languages. Therefore, the child learner may be able to learn whether the *that-trace* constraint applies in his language, by observing whether postverbal subjects are available.

a. *Who$_i$ did you say that t_i has written this book?
b. Chi$_i$ hai detto che t_i ha scritto questo libro?
who have-you said that has written this book
'Who did you say has written this book?' (33)

a. *Have arrived many students
b. Hanno arrivato molti studenti.
have-3pl arrived many students
'Many students have arrived.' (34)

VARIANTS OF SYNTACTIC THEORY

Since the 1960s syntactic theory has undergone a number of changes, and has spawned a variety of different grammatical theories, each with a different title, such as Relational Grammar (RG), Head-Driven Phrase Structure Grammar (HPSG),

Lexical-Functional Grammar (LFG), Categorial Grammar (CG), Government-Binding Theory (GB), Tree Adjoining Grammar (TAG), etc. While it is tempting to view these as monolithic alternatives, to do so would be misleading.

First, all approaches provide only fragments of a full theory of grammatical knowledge; sometimes these fragments only partially overlap between approaches. Second, there are many fundamental points of agreement between the different approaches. Third, the differences among practitioners of the same general framework can be as large as or even larger than the differences between frameworks. The differences of opinion that engender different 'named' grammatical theories draw greater attention, but they have no special status. Therefore, rather than reviewing different named grammatical theories, this section focuses on a selection of fundamental issues on which syntactic theories diverge.

Syntactic Atoms and How They Combine

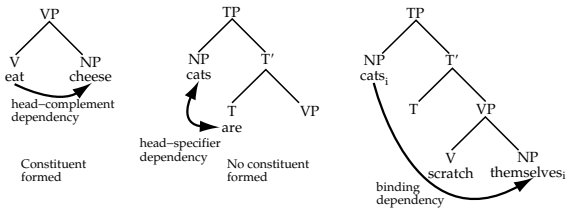
First, syntactic theories differ on the issue of what are the 'atoms' of syntax, that is, the pieces of sentences that are stored in a speaker's long-term memory. At one extreme are certain versions of Transformational Grammar (including the recent *Minimalist Program*), which claim that the atoms of syntax are smaller than words – either morphemes or individual syntactic features. Under this approach, underlying syntactic structures are formed by selecting a set of these atomic units, and combining them based on highly general principles of structure-building. Under this approach, syntax is responsible even for the formation of word-sized units. For example, an inflected verb such as 'runs' may be formed by independently selecting the verbal head 'run' and the inflectional head *[3rd person singular, present]*, and applying a transformation which combines them to form a complex syntactic head, which is spelled out as the word 'runs'.

At the other end are approaches that assume much larger atoms, in the form of templates for phrases or even clauses. Construction Grammar and some versions of Tree Adjoining Grammar are examples of such approaches. Under these approaches, the representation of idiomatic expressions is little different from the representation of other types of phrases. Construction Grammar has provided some insightful analyses of constructions that have been largely overlooked in mainstream transformational syntax. (*See* **Construction Grammar**)

Despite disagreements about the size of the atoms of syntax, there has been a quiet convergence of opinion on the role of the atoms of syntax. In early generative theories it was standard to distinguish the terminal elements of syntax (i.e. lexical items) from the phrase structure rules that determine how the terminals combine. In most current theories this distinction has been eliminated, and the work once done by phrase structure rules is replaced by a set of highly general conditions on how syntactic atoms combine. In these *lexicalized* grammars, information about the combinatorial possibilities of syntactic atoms is built into the lexical entries of the atoms themselves. Lexicalism is a common feature both of theories that assume very small syntactic atoms and of theories that assume much larger syntactic atoms.

Types of Structural Dependencies

A second issue involves the question of how syntactic elements enter into structural dependencies. As a starting point, in a typical phrase structure grammar syntactic elements enter into two basic types of dependencies, illustrated in (35). First, when syntactic elements enter into a *sisterhood* relation, this both forms a dependency between the two elements and creates a new syntactic *constituent*. Constituents may participate in a variety of different syntactic processes, such as coordination, movement, and ellipsis. On the other hand, many syntactic dependencies do not involve the formation of new constituents, for example subject–verb agreement and reflexive binding.



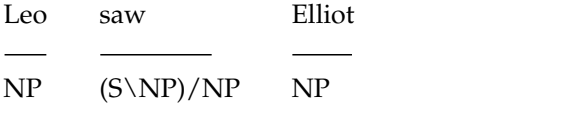
(35)

The phrase structure notions that create this division among structural dependencies continue to dominate thinking about syntax, but there are a number of interesting alternative proposals that reduce or eliminate this distinction. First, *Dependency Grammars* treat all syntactic dependencies in a parallel fashion, and do not single out constituent-forming dependencies as special (36).

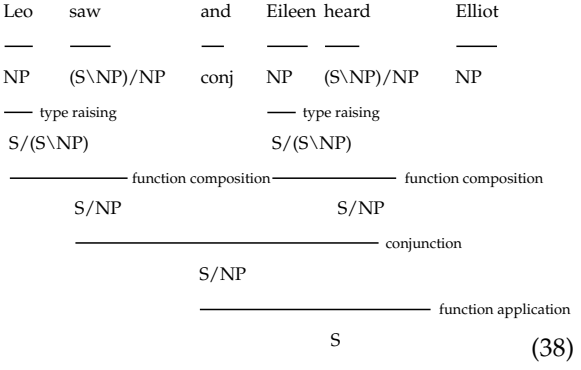
Dependency Grammar representation of argument and agreement relations (36)

Combinatory Categorial Grammars (CCG) also reduce the division of dependency types, but in the opposite manner from Dependency Grammars. In CCG grammars, information about the elements that a syntactic atom may combine with is encoded in enriched category labels. For example, an intransitive verb such as 'run' might have the category label S\NP, read as 'a category that combines with an NP to its left to form an S'. Dependencies between syntactic sisters are formed by the rule of *function application*. Dependencies between non-sisters are also formed by function application, thanks to the mediating effects of *function composition* rules, which allow the combinatorial requirements of a syntactic atom to be passed up through a series of larger units.

By virtue of their more uniform treatment of structural dependencies, both Dependency Grammar and Combinatorial Categorial Grammars have led to innovative proposals about the treatment of constituency phenomena. Whereas most phrase structure grammars impose a clear boundary on which syntactic relations form constituents, DG and CCG do not. Therefore, these approaches have been used to analyze syntactic phenomena that do not fall straightforwardly under standard notions of constituency, such as 'nonconstituent coordination' involving subject–verb sequences. In CCG, flexible constituency relations are made possible by the introduction of *type raising* rules, which allow the combinatorial requirements of a category to be satisfied in a different order (38).



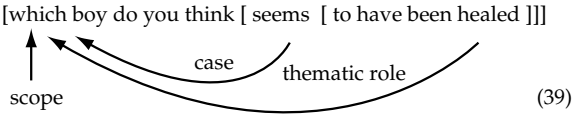
(37)



(38)

Multiple Roles: Alternatives to Transformations

All syntactic theories must address the fact that individual syntactic elements can enter into a variety of different structural relations. In a multiclause sentence, a single phrase may take scope in one clause, be case-marked by the predicate of a second clause, and receive a thematic role in a third clause (39).



(39)

Since the 1950s and 1960s, transformational approaches to syntax have famously argued that phrases can bear multiple syntactic roles because there are multiple syntactic levels of representation, which are related to one another via movement through a series of different structural positions. As a result, a leading area of research on transformational syntax has been the investigation of constraints on possible movement operations, as outlined above.

The syntactic frameworks *Lexical-Functional Grammar* (LFG) and certain versions of *Combinatory Categorial Grammar* (CCG) share with Transformational Grammar the assumption that words and phrases bear multiple roles because they appear in multiple different levels of representation. These approaches diverge from Transformational Grammar in the respect that they do not assume that each different level is a hierarchical constituent structure, or that the levels are related by movement transformations. LFG assumes independent levels of *a-structure* (argument structure: representation of argument/thematic roles), *f-structure* (function structure: representation of subject, object, etc. roles), and *c-structure* (constituent structure: surface syntax). There are rules for mapping between these levels, but the mappings are not assumed to be transformational. In some versions of CCG separate representations of *argument structure* and *surface structure* are posited. (*See* **Lexical-Functional Grammar**)

In *Head-Driven Phrase Structure Grammar* (HPSG) only one syntactic level of representation is assumed. This single level of representation combines words and phrases into the surface constituent structure that is familiar from many other syntactic theories. However, the terminal elements of these structures contain highly articulated feature structures, which encode a great deal of information about argument structure,

phonology, 'moved' arguments, and so on. Whereas transformational grammars use movement operations to handle the multiple roles problem, the same work is done largely internal to individual syntactic heads in HPSG. For example, a verbal head may encode the information that a *wh*-phrase, which is represented as the *focus* argument of the clause, is to be treated as the filler of one of the slots in the verb's argument structure list. Constraints on movement operations in transformational approaches must be replaced in nontransformational theories by related constraints on the relations between scope and argument slots. (*See* **Phrase Structure Grammar, Head-driven**)

Causes of Ungrammaticality

Standard approaches to syntactic theory assume a set of syntactic atoms and a relatively small number of formal principles or constraints that determine how these atoms may be combined to form sentences. A sentence is assumed to be grammatical if it violates no constraints. An ungrammatical sentence is a sentence that violates one or more constraints. Some variants of syntactic theory have explored broader notions of what causes a sentence to be (un)grammatical.

Functional grammars typically emphasize the role of meaning or of communicative efficiency in determining the well-formedness of a sentence. Such approaches typically do not appeal directly to semantics or processing efficiency to explain ungrammaticality. Rather, semantics or processing efficiency are used to provide a functional motivation for a set of formal grammatical constraints.

Both *Optimality Theory* and certain versions of the *Minimalist Program* question the standard assumption that a grammatical sentence is a sentence that violates no constraints. This characterization is replaced in these approaches with the requirement that a well-formed sentence is the *optimal* candidate from a set of possible structures/derivations for that sentence. In other words, a sentence may be deemed ungrammatical for the simple reason that there exists a better way of expressing the same thing. (*See* **Optimality Theory**)

CHALLENGES AND FUTURE PROSPECTS

Universal Grammar

The Principles and Parameters approach to syntax, which is compatible with any of the syntactic frameworks discussed above, aims to explain how a child can attain rich knowledge of any language. It does so by seeking universal syntactic principles and clusters of syntactic properties that covary across languages. Syntactic research since the 1970s has uncovered a wealth of cross-linguistic findings, and a number of good candidates for universals of syntax have been found. However, the search for parameters has met with mixed success. The prospect that each parametric cluster may be linked to an easily observable surface property of the language appears to be viable, but parametric clusters of properties appear to be both narrower and more numerous than originally expected. It remains to be seen whether all of natural language syntax, including the idiosyncrasies of each language, can be handled in terms of a Principles and Parameters approach.

The Unification Problem

Syntactic theories are theories that aim to characterize the mental representations underlying knowledge of language, and of how people acquire that knowledge. However, most syntactic theories characterize knowledge of which sentences are grammatical and which sentences are ungrammatical, with few suggestions about how speakers successfully access this knowledge in real-time speaking or understanding, or about how children acquire this knowledge. Even less is known about how to encode this knowledge in brain structures. The overall goals of syntactic theory may be significantly affected by findings in these areas. In addition, a complete syntactic theory will have to provide answers to questions about how sentence structures are learned, used in real time, and encoded in the brain.

Further Reading

Baker M (2001) *The Atoms of Language*. New York: Basic Books.

Baltin M and Collins C (2000) *Handbook of Contemporary Syntactic Theory*. Malden, MA: Blackwell.

Bresnan J (2000) *Lexical Functional Grammar*. Malden, MA: Blackwell.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Crain S and Lillo-Martin D (2000) *An Introducion to Linguistic Theory and Language Acquisition*. Malden, MA: Blackwell.

Culicover P (1997) *Principles and Parameters*. New York: Oxford University Press.

Radford A (1988) *Transformational Grammar*. Cambridge, UK: Cambridge University Press.

Roberts I (1997) *Comparative Syntax*. London: Edward Arnold.

Sag I and Wasow T (2000) *Syntactic Theory: A Formal Introduction*. Stanford, CA: CSLI Publications.

Steedman M (2000) *The Syntactic Process*. Cambridge, MA: MIT Press.

Tense and Aspect

Advanced article

Carlota S Smith, University of Texas, Austin, Texas, USA

CONTENTS

Tense and aspect convey temporality in language. Tense relates situations to a reference time and the moment of speech. Aspectual information classifies events and states into situation types, and presents them from a perfective or imperfective viewpoint.

INTRODUCTION

The cognitive science perspective analyzes the formal systems of language in terms of the information they provide. Temporality is pervasive in human experience, and in language. Natural language expresses temporal information with the linguistic forms of adverbials, tense, and aspectual categories. The aspectual notion of temporality has to do with how situations unfold in time: it recognizes beginnings and endings, intervals, the dynamic and the static.

Tense and adverbials temporally locate the situation talked about in a sentence; aspect categorizes that situation. For instance, consider the temporal information in the sentences (a) 'Mary was opening the door' and (b) 'Mary opened the door'. The situations are located before the time of speech or writing; this information is conveyed by the verb tense. They are durative events with a final endpoint; this is aspectual information, conveyed by the verb and its arguments. In (a), the event is in progress: this information is conveyed by the auxiliary verb, which expresses the progressive aspectual viewpoint; in (b) the event is complete, conveyed by the simple verb form which expresses perfective aspectual viewpoint.

In what follows these statements will be explicated so that the reader will have a grasp of temporal location, aspectual situation type, and aspectual viewpoint. The next section is concerned with temporal location, tensed and tenseless languages, tense systems, the conceptual and semantic interpretation of tense, and the contribution of context. The following section discusses aspect: aspectual situation types and viewpoints, coercion, and global categories. The final section concludes with comments about the interaction between tense and aspect.

THE INTERPRETATION OF TEMPORAL LOCATION

Adverbials, Tense, and Time

Time is a single unbounded dimension without landmarks; it is analogous to space, though simpler. As space requires a landmark to indicate a position, time requires an orientation point to locate or anchor a situation. The speaker is the canonical center of linguistic communication: the basic orientation points in language are the speaker's place ('here') and the speaker's time ('now'). Linguistic forms that anchor to these coordinates are known as *deictic*. In sentences out of context, situations are located with respect to Speech Time, which is always the Present. Context can change the interpretation significantly. (Capital letters refer to times, lower-case letters to tenses.)

All known languages express temporal notions and divide time into the general notional categories of Past, Present, and Future. Strikingly, this holds for languages embedded in cultures that are very different from the Indo-European. The Hopi language, for instance, was thought by Benjamin Whorf to encode temporal distinctions entirely different from those of the West. But recent studies show that Hopi has a full, productive set of temporal adverbials which make the familiar distinctions.

It is important to distinguish between tense and time. Tense is not necessary for conveying temporal information: time adverbials suffice. There are languages without tense, but all languages have time adverbials ('yesterday', 'when John arrived', 'earlier', etc.). Situations are located at the moment of speech, Speech Time; or before or after Speech Time. Past and Future time are symmetrical in their relation to Speech Time, but their interpretations

are not entirely symmetrical. The Future is unknown, and so reference to it has the modal quality of prediction. Indeed, the English future *will*-future patterns with other forms that have modal force ('may, might', etc.).

In tenseless languages (Thai, Mandarin Chinese, classical Hebrew, and many others) temporal adverbials function with aspectual information to locate situations temporally. Temporal adverbials may give specific or relational times ('Mary called at noon/earlier'), or locate one situation in terms of another ('John left when Mary arrived'). Adverbials may be anchored to Speech Time ('now, yesterday, tomorrow') or a time specified in the linguistic context ('earlier, later').

This article will focus on English as a typical tensed language. The basic meaning of tense is deictic, anchored to Speech Time. Tense locates the situation in a clause by tacitly invoking two times: the time of the situation expressed, which I will call Situation Time; and Speech Time. A third time, Reference Time, is introduced below. The tense of a sentence locates the situation talked about from Speech Time. The common labels for tense reflect this: present, past, and future tense indicate times at, before, or after Speech Time.

Tense Systems in Language

Tense is a grammatical form associated with the verb, with temporal meaning. The category includes verb affixes ('Mary walk<u>s</u>', 'Mary walk<u>ed</u>'), infixes ('Mary r<u>a</u>n'), and perhaps verb auxiliaries such as the *will*-future in 'We will walk' ('will' also has modal meanings, beyond the scope of this discussion). Other forms that convey temporal meaning are not technically tenses.

Tense systems vary among languages. In a study of 64 languages, Dahl (1985) found that almost all have a past tense of some kind, and that over three-quarters of the languages have future tense. Present tense is often conveyed by the absence of an overt morpheme – it is unmarked, as in the English 'We <u>like</u> snow'. Tense is usually a suffix, though Bantu languages have tense prefixes. In less than half of the languages studied the future tense is conveyed by an auxiliary verb ('periphrastic' tense). The most common periphrastic tense across languages is the 'perfect', a complex construction with temporal and aspectual meaning. The perfect in English is conveyed by auxiliary 'have' and a participle, as in 'They <u>have locked</u> the door', 'They <u>had locked</u> the door'. Some languages have a partial tense system. Navajo, for instance, has verb affixes that convey Future time; sentences without them are located in the Past or Present.

Tenses may be 'absolute' or 'relative'. Absolute tenses relate to Speech Time, while relative tenses require a linguistically specified anchor time. The past perfect is a relative tense in English: for instance, in 'At noon they had locked the door' the adverbial 'at noon' provides an anchor time. Again, the modal 'would' allows a future-in-past interpretation requiring an explicit anchor: in 'Jane said last week that Bill would leave soon' the main clause provides the anchor for 'would' in the complement clause.

Degrees of remoteness from Speech Time are coded into some tense systems, including Aboriginal languages of Australia, some American Indian languages, and Bantu languages. Degrees of remoteness are more frequent and more complex for the past than the future. For instance Haya, a Bantu language, has a three-way distinction in the Past – today, before-today, remote – and a two-way distinction in the Future.

In English and in tensed languages generally, every independent clause has a tensed verb, known as a 'finite' verb. Most languages also have tenseless verb forms temporally dependent on the tensed verb in a sentence. The tenseless forms in English are bare stems ('I saw them <u>go</u>'), infinitives ('We want <u>to go</u>'), and participles ('<u>Singing</u>, John left'; '<u>Exhausted</u>, John left').

Tenses vary in interpretation: past and present tense do not always indicate Past time and Present time. For instance, the present tense with a past adverbial conveys Past time – the 'historical present', as in 'Last week, this guy comes up and offers me a lottery ticket'. The past tense may be anchored to a time in the Future. The variation can be explained by treating the semantic meaning of tense as relational: the present tense conveys simultaneity, the past tense anteriority, the future posteriority.

Conceptual Interpretation of Tense: Reference Time

There is an important conceptual dimension to temporal information. The speaker's centrality implies an organizing consciousness, a temporal perspective. The temporal perspective may differ from the time of a situation. Consider (a) and (b); they have the same truth-conditions yet differ subtly in meaning: (a) 'Henry arrived.' (b) 'Henry has arrived.' (a) is set squarely in the Past, while (b), a present perfect, has a Present perspective implying that the arrival is relevant to the Present. Hans

Reichenbach (e.g. 1947) noted this contrast and attributed it to a difference in perspective, or Reference Time. (a) has a Past Reference Time, while in (b) the Reference Time is Present. In both, Situation Time is in the past.

Reference Time explains the difference between the past and the present perfect tense. The past tense conveys that Reference Time and Situation Time are the same, prior to Speech Time. With the present perfect, Reference Time and Speech Time are the same, and Situation Time is prior. Generalizing, all tenses involve three times.

Shifted deixis is also clarified by the notion of Reference Time. Deictic adverbials such as 'now', 'in three days', etc., normally anchor to the moment of speech. But they can also anchor to another time, as in 'Mary sat down at the desk. Now she was ready to start work'. In the second sentence 'now' is anchored to the Past time, and suggests the temporal perspective of Mary. Certain effects of context on interpretation can also be explained with Reference Time. When there are simultaneous or overlapping situations, Reference Time gives an indispensable locus for the relation between them. (*See* **Reasoning about Time**)

Semantic Interpretations of Tense

In formal semantics the classical approach of tense logic was important as a research tool. It has now been supplanted by richer, more natural systems. Tense logic analyzes tense as a sentence operator that determines the time of evaluation for the proposition expressed. For a past tense sentence expressing the proposition P (past (P)), the truth of P is evaluated at a time prior to the moment of speech. Although reasonable for simple sentences, the approach cannot be extended to complex cases. Complex tenses do not function as nested operators, one within the scope of the other; and many predicted combinations do not occur. Further, the treatment does not extend to other temporal expressions. Adverbials are not operators on tense: in the sentence 'Mary left yesterday', 'yesterday' does not indicate a time prior to the time indicated by the past tense. Rather, the adverbial further specifies the time. The tense logic approach cannot deal with explicit or implicit contextual dependencies.

There is also an objection of a different kind. Tense cannot be an operator on a proposition, because nominal referents that figure in the proposition are temporally independent. To see this, consider 'Every fugitive is now in jail'. The referents of the subject noun phrase (NP) do not have fugitive status at the time of the sentence ('now'): they were fugitives at some earlier time. Another key example: 'The president was a fool'. There are two readings: either the person who is now president was a fool earlier during the presidency, or the person was a fool at an earlier time when not the president. These examples show that the NP referents are outside the scope of tense (Enç, 1986). The operator approach fails on both scopal and compositional grounds.

The referential analysis of tense, in which tenses refer to times, is now widely accepted. Tense information involves three times, following Reichenbach: Speech Time, Reference Time, and Situation Time ('event time' for Reichenbach). Each tense conveys information about the relation between Speech Time and Reference Time (SpT and RT), and the relation between Reference Time and Situation Time (RT and SitT). In the present tense, SpT equals RT; in the past tense, RT precedes SpT; in the future, RT follows SpT. For the simple tenses SitT and RT are the same; for complex tenses they differ. The notion of Reference Time is now well established; other parts of Reichenbach's original system have been revised and extended.

In current generative grammar, Tense heads its own functional projection and encodes semantic information. In dynamic semantic theories such as Discourse Representation Theory, tenses are associated with the appropriate relational information. The appearance of a given tense in surface structure licenses the introduction of times and conditions stating their relations in semantic representation (Kamp and Reyle, 1993). (*See* **Semantics, Dynamic**)

Context and the Interpretation of Tense

Context may affect the interpretation of tense. Information in the context may determine a specific time, as in 'anaphoric' interpretations of tense (Partee, 1984). For instance, in 'I didn't turn off the stove', the past tense implies a particular earlier time. In the sequence 'I went to John's party last night. Mary was there' the second past tense refers to the time specified by the first tense and time adverbial together.

Tense anchoring may also be determined by syntactic context. Complement clauses, for instance, may be anchored to a main clause time. Consider the complement clause in 'Tomorrow at midnight the Prime Minister will announce that he burned the documents an hour earlier'. The anchor time for the past tense in the complement is the Future time of the main clause. Unlike complement

clauses, relative clauses may be temporally dependent or independent. With relative tenses, an independent sentence may provide the anchor time: 'We arrived at noon. John had already left.' The past perfect and the adverb 'already' anchor to the time in the first sentence.

In discourse, tense is interpreted in three different patterns, according to the context. In narrative and procedural contexts tense conveys continuity; in descriptive contexts tense is anaphoric to an established time; in other contexts tense is deictic. (*See* **Discourse Processing**)

Finally, tenses have atemporal meanings in certain contexts. The past tense conveys irrealis in conditionals, as in 'If you left early I would be pleased', and politeness as in 'I wanted to ask you a question'. These meanings for the past tense are not unique to English, but are found in many languages of the world. (*See* **Story Understanding**)

ASPECT: INTERNAL TEMPORAL STRUCTURE

Aspect is a semantic domain that is reflected in the syntax and morphology of a language. Two kinds of aspectual information are conveyed about the eventualities, or situations, expressed in sentences. Situations are presented from a particular perspective, or viewpoint; and they are indirectly classified as a state or an event of a certain type.

Situation Type

Situation type information classifies situations according to their internal temporal properties. For instance, 'John is tall' expresses a state, 'John walked by the river' an event. Events may be telic, with natural final endpoints, or goals: 'John walked to school' is telic. The traditional term for such distinctions is *Aktionsarten* (kinds of action). The terms *situation* and *eventuality* are neutral between event and state.

Since Aristotle, situations have been classified by temporal features. The features are based in perceptual and cognitive abilities, and realized indirectly in language. Situation types are concepts, idealized classes of situations. Vendler (1967) distinguished four types, using the features static-dynamic, telic-atelic, and durative-instantaneous. The list below adds a fifth, semelfactives (Smith, 1997):

States: static, durative (know the answer, love Mary)
Activities: dynamic, durative, atelic events (laugh, push a cart, stroll in the park)
Accomplishments: dynamic, durative, telic events consisting of a nondetachable process with an outcome (build a house, walk to school, learn Greek)
Semelfactives: dynamic, atelic, instantaneous events (tap, knock, flap a wing)
Achievements: dynamic, telic, instantaneous events (win the race, reach the top)

Temporal features are concepts associated with classes of situations. They are realized when a situation unfolds in time. Situations with the feature [Dynamic] are events, and occur at successive stages in time. [Atelic] events can continue indefinitely, with arbitrary final endpoints. [Telic] events have an intrinsic final endpoint with a change of state. [Durative] events take place over an interval, while [Instantaneous] events occur at a moment, in principle. [Static] situations have an undifferentiated interval with no structure. States do not have endpoints: the coming about and ending of a state are events in themselves. Events take time, whereas states do not.

Abstracting away from duration, situations are classified as States, Processes, and Events, using the features [stative/dynamic] and [homogeneous/heterogeneous]. Situations have patterns of entailment due to their part structure. States and Processes are homogeneous, with no difference between part and whole. There is an entailment from part to whole that follows from this homogeneity: if one has done any walking, for instance, one has walked. When a State holds for an interval, it holds equally for any smaller interval of that interval; this is known as the sub-interval property. The classes differ in dynamism and in uniformity: States are Static and uniform, Processes are dynamic and less so.

Events are dynamic and quantized, with a heterogeneous part structure. For Events there is no entailment from part to whole: no proper part has the status of the event. Telic events have intrinsic bounds; they are instantaneous or involve an incremental process. In the incremental process, the part structure of the event maps onto that of the referent. The event proceeds incrementally and as it does so the referent is traversed, built, used up, etc. Take the event 'Mary ate an apple'. As the event progresses, more and more of the apple is eaten: at completion no more apple remains and no subpart is an event of eating an apple. The process realizes a homomorphism from the argument denotation to the event (Krifka, 1989). This analysis captures the close association between count and mass nouns ('an apple', 'wine') and quantized and nonquantized situations ('Lee drank a glass of wine/wine').

Situation types are realized by the verb and its argument, the verb constellation. The verb is key, but the nature of the arguments is also a factor. For instance, 'Lee ate apples' expresses an Activity, while 'Lee ate an apple' expresses an Accomplishment. In the latter the object is quantized, discrete. Some languages convey this difference by case marking: in Finnish, for instance, accusative case NPs are quantized, partive case NPs are not. Languages may have marked situation types, such as the Tentative in Mandarin, a limited event indicated by verb reduplication.

The temporal features have grammatical correlates with co-occurrence restrictions. For instance, the semantic property of dynamism is expressed by the progressive, verbs and adverbs of volition, and pseudo-cleft 'do'. Stative sentences do not appear with these forms: '*Mary was knowing the answer' or '*What Mary did was know the answer' (*indicates ill-formedness). Adverbs of duration occur with states and atelic events ('She walked for an hour'); adverbs of completion occur with telic verb constellations but are odd with atelics, e.g. 'She wrote the letter in an hour', '??She strolled by the river in an hour'. Due to such facts, situation types are 'covert' grammatical categories: recognized in the grammar though not morphologically marked.

Aspectual Viewpoint

Aspectual viewpoints focus situations like the lens of a camera. As the camera lens makes a scene available, so viewpoint makes visible all or part of the situation expressed in a sentence. The information made visible is available for semantic interpretation and pragmatic inference. The aspectual viewpoints of a language are usually signaled morphologically. The term *grammatical aspect* is also used.

The main aspectual viewpoints are perfective and imperfective. In English the choice between them is obligatory, conveyed by the form of the verb: the simple verb form is perfective, as in 'Mary swam in the pond', the *be+ing* auxiliary is imperfective, as in 'Mary was swimming in the pond'. Perfective viewpoints present situations as bounded, often with endpoints; imperfectives present ongoing situations with no information about beginning or end. The perfect is a special case, discussed below. The progressive, a type of imperfective viewpoint, appears neutrally only with non-statives: 'Mary is singing', '*John is knowing the answer'. There is a third type, the neutral viewpoint; it allows bounded and unbounded interpretations, depending on contextual information. The neutral viewpoint arises in sentences with no viewpoint morpheme; it is not applicable to English.

Aspectual systems differ most in the viewpoint component. English has one perfective and one imperfective viewpoint. Some have richer systems: Mandarin, for instance, has two imperfective viewpoints and three perfectives, none obligatory. In Russian, viewpoint information is salient and is conveyed by verb prefixes and suffixes. Linguistic expression of tense and aspectual viewpoint is distinct in some languages, intertwined in others. The two vary independently in English. In French, past tenses code aspectual viewpoint: the *imparfait* is a past imperfective tense, the *passé composé* and *passé simple* are past perfectives. Some languages, for example Finnish and Icelandic, have minimal viewpoint systems.

Perfective viewpoints generally present situations as discrete, bounded. If the situation is telic, the perfective expresses a completed event. The interpretation of completion can be demonstrated with conjunction. It is semantically contradictory to conjoin a perfective event sentence with an assertion that the event continues: '#Mary opened the door, but she didn't get it open'; '#Donald fixed the clock and he is still fixing it'. The English perfective presents statives as unbounded, as shown by the felicity of 'Mary knew the answer and she still knows it'. This is a language-particular feature; French perfective viewpoints present statives as bounded.

Imperfective viewpoints focus part of a situation, without endpoints. Since it is unbounded, the focused interval has the sub-interval property. Imperfectives of internal or preliminary intervals have an intensional component. They focus a stage of a process, which, if continued to its final endpoint, results in an event. Thus one cannot infer from a progressive 'They were building a house' that the building was completed. This gap between knowledge of the type of situation and its outcome is known as the Imperfective Paradox (Dowty, 1979). Imperfect viewpoints may also focus resultant intervals. For instance, the Japanese *-te iru* is an imperfective verb form that focuses resultant or internal intervals.

Coercion

People talk about situations in more than one way: states can be presented as events, events can be presented as states, and the endpoints are events. Language has mechanisms for 'coercion'

or 'situation type shifts' that change the class of verb constellations.

There is a basic-level set of associations between situations in the world and situation types. However, verb constellations typically of one type may be coerced into another. For instance, 'Teresa understood the problem' is a typical state; but with a point adverbial it changes, expressing an event: 'At that moment Teresa understood the problem.' Verbs such as 'start, begin' also convey this meaning. Coercion is triggered by material in the context, often adverbial.

Events and specific states can be coerced in generalizing-habitual sentences, for instance: 'Teresa always played tennis on Tuesday.' Such sentences are semantically stative, expressing a pattern of recurrence. Instantaneous events are coerced into processes with multiple stages with durative adverbials: 'John knocked at the door for five minutes.' There is a general rule for coercion, the rule of External Override. If an adverbial and verb constellation are incompatible, the feature of the adverbial coerces the verb constellation. Again, durative adverbials coerce telic verb constellations into atelics: 'Mary read a book for an hour.' This sentence conveys that Mary did some book-reading. Other examples of coercion include stative verb constellations with the progressive viewpoint, presenting a state as an event, as in 'I'm liking this play'.

The *perfect* coerces a situation into a consequent state, as in 'Mary has eaten dinner'. This is the aspectual meaning of the perfect. There is also a temporal component. Temporally, the perfect conveys that Situation Time precedes Reference Time. English has a constraint on adverbials and the present perfect: for instance, the sentence '*Mary has left yesterday' is impossible. Such adverbs are good with the pluperfect and future perfect, however, and in some other languages.

Supercategories include information from both aspectual situation type and viewpoint. The classes are Statives, homogeneous and unbounded; and Events, heterogeneous and bounded. They contribute differently to narrative time. Events advance narrative time; Statives do not. Whether a given sentence moves narrative time constitutes a test for its global aspectual category.

The class of Statives consists of basic-level states, perfects, habitual-generalizing sentences, and sentences with the progressive viewpoint. States are intrinsically unbounded; perfect sentences and habitual-generalizing sentences are stative by coercion. The progressive, a viewpoint, focuses dynamic situations without endpoints. Thus homogeneity arises in more than one way.

The class of linguistically realized Events consists of bounded, dynamic situations expressed with the perfective viewpoint.

INTERACTION BETWEEN TENSE AND ASPECT

Tense and aspect information interact at Situation Time. Depending on aspectual value, the situation of a sentence overlaps Situation Time, or is included within it.

The boundedness of a situation determines its relation to Situation Time. Bounded events are included in the Situation Time interval: for instance, 'Lee built a sandcastle yesterday' takes place within the interval of 'yesterday'. Ongoing events and states overlap or surround Situation Time. In the sentences 'Lee was building a sandcastle' and 'Henry was at school', the situations hold during Situation Time and are understood to hold before and after it.

There is a general constraint on linguistically realized situations in the Present which has far-reaching implications. The constraint allows only unbounded situations at Speech Time. Thus Present events must be ongoing, in English with the progressive: 'John is talking', 'Mary is drawing a circle'. The constraint involves a general principle of communication that is pragmatic and semantic. In the temporal perspective of the Present, there is a tacit convention that communication is instantaneous. The perspective of the present time is incompatible with a bounded event, because the bounds would go beyond the moment of the communication.

Languages realize the Bounded Event Constraint differently. In English, simple present event sentences undergo coercion and are interpreted as habitual statives, as in 'Tom feeds the cat', 'Sue speaks French'. Apparent exceptions include performatives ('I hereby christen this ship') and sports-announcer reports ('Now Jones throws the ball'). In these uses, time is telescoped into a notional moment. Russian has another strategy: present tense perfective sentences express Future time.

The constraint against bounded events in the Present explains an important pattern of temporal interpretation in tenseless languages. Commonly, states and progressives – both unbounded – are located in the Present, while bounded events are located in the Past. The Bounded Event Constraint prevents uncertainty about bounded events.

References

Dahl O (1985) *Tense and Aspect Systems*. Oxford, UK: Blackwell.
Dowty D (1979) *Word Meaning and Montague Grammar*. Dordrecht, Netherlands: Reidel.
Enç M (1986) Towards a referential analysis of temporal expressions. *Linguistics and Philosophy* **9**: 405–426.
Kamp H and Reyle U (1993) *From Discourse to Logic*. Dordrecht, Netherlands: Kluwer.
Krifka M (1989) Nominal reference, temporal constitution and quantification in event semantics. In: Bartsch R *et al*. (eds) *Semantics and Contextual Expressions*, pp. 75–115. Dordrecht, Netherlands: Foris.
Partee B (1984) Nominal and temporal anaphora. *Linguistics and Philosophy* **7**: 243–286.
Reichenbach H (1947) *Elements of Symbolic Logic*. London, UK: Macmillan.
Smith CS (1997) *The Parameter of Aspect*, 2nd edn. Dordrecht, Netherlands: Kluwer.
Vendler Z (1967) *Verbs and Times*. Ithaca, NY: Cornell University Press.

Further Reading

Bach E (1986) The algebra of events. *Linguistics and Philosophy* **9**: 5–16.
Comrie B (1976) *Aspect*. Cambridge, UK: Cambridge University Press.
Comrie B (1985) *Tense*. Cambridge, UK: Cambridge University Press.
Mourelatos A (1978) Events, processes and states. *Linguistics and Philosophy* **2**: 415–434.
Parsons T (1990) *Events in the Semantics of English*. Cambridge, MA: MIT Press.
Tenny C (1994) *Aspectual Roles and the Syntax–Semantics Interface*. Dordrecht, Netherlands: Kluwer.
Verkuyl H (1993) *A Theory of Aspectuality*. Cambridge, UK: Cambridge University Press.

Typology

Intermediate article

William Croft, University of Manchester, Manchester, UK

CONTENTS

Typology is an empirical, cross-linguistic approach to the study of grammatical structures in the world's languages.

Typology is an empirical, cross-linguistic approach to the study of grammatical structures in the world's languages. The typological method was first formulated by Joseph H. Greenberg in the 1960s (Greenberg, 1966). The typological method consists of three steps: classification, generalization and explanation. In the first step, languages are classified into types. For example, English places its adjectives before nouns, as in *red book*, while Spanish places its adjectives after nouns, as in *libro rojo*. English and Spanish belong to two different types, adjective–noun and noun–adjective respectively. In the second step, the range of attested language types is compared to the range of possible language types, and generalizations (universals) are formulated that constrain languages to types attested or assumed to exist. Finally, explanations are offered for the universals. Explanations range from processing explanations to semantic-pragmatic explanations to diachronic (historical) explanations. As in any empirical science, typological research involves an interplay among classification of data, generalization, and explanation.

TYPOLOGICAL CLASSIFICATION: BASIC METHODOLOGICAL ISSUES

The classification of languages into grammatical types presupposes that there is a cross-linguistic basis for the comparison of grammatical structures in different languages. Languages in fact vary considerably in their grammatical structure. For example, many languages do not have a separate word class of adjectives. Typologists address this problem by comparing how meanings are encoded in grammatical form. Thus, to identify adjective–noun order in a language that lacks adjectives as a separate word class, one examines the order of words denoting properties such as 'red'. One thus examines the full range of grammatical variation without excluding certain grammatical types *a priori*.

A major challenge in typological classification is to construct a proper sample of the world's languages from which one may make valid generalizations or statistical inferences. In order to make valid generalizations, one must construct a variety sample which maximizes the likelihood of capturing all relevant gramamtical variation. In order to make valid statistical inferences, one must construct a probability sample which reflects the proportion of different types among the world's languages.

A variety sample is constructed by seeking the most diverse sample of languages for the size of the sample population. Typological diversity can be obtained by minimizing the likelihood that the languages in the sample are similar due to historical accident, that is by sharing a common ancestor or by contact. For example, English and German are both adjective–noun because they both belong to the Germanic language family. Finnish and Russian have subject–verb–object (SVO) order because of centuries of contact. Hence, typological diversity is maximized by constructing a genetically and geographically stratified language sample.

A probability sample is more difficult to construct. Statistically, a probability sample should contain only historically independent cases. However, some linguistic phenomena are so stable that reasonably large sample sizes will inevitably include genetically or geographically related language types. Also, the proportion of languages can be distorted by recent expansions of language families, such as Bantu and Austronesian. Various statistical techniques are used to address these problems.

Typology depends heavily on descriptive linguistic research, and various souces are drawn upon. Native language consultants provide the greatest detail in description but cannot be relied on exclusively in large studies. Reference grammars are the most practical source for large studies, and are adequate for basic grammatical phenomena, but cannot always be relied on for more specialized studies. Recorded texts are used for studies of grammatical phenomena connected to discourse.

IMPLICATIONAL UNIVERSALS AND COMPETING MOTIVATIONS

Typological generalizations are universals of language. The vast majority of typological generalizations are implicational universals: e.g. if a language has adjective–noun (AN) word order, then it also has numeral–noun (NumN) order. Implicational universals accommodate a certain degree of cross-linguistic variation. The implicational universal given above allows languages that are AN/NumN, NA/NumN, or NA/NNum; it excludes only languages that are AN/NNum.

Typological generalizations are formed by comparing attested language types to possible language types. Further research and larger samples may bring up counterexamples to proposed implicational universals. Sometimes the universal is refined to account for the putative counterexamples. In other cases, even when all types are attested, the frequency is often highly skewed, and this skewing calls for an explanation.

A competing motivations model is used to explain many typological generalizations. Explanatory principles (motivations) in this model cannot be simultaneously satisfied. The motivations compete; existing language types represent the range of most optimal satisfaction of the competing motivations. (*See* **Constraint Satisfaction**; **Optimality Theory**)

In the classic paper on typology, Greenberg (in Denning and Kemmer, 1990) demonstrated the existence of three common clause word orders, SOV, SVO, and VSO; he noted the existence of VOS and even OVS; the existence of OSV languages is still uncertain (see for example Comrie, 1989). Greenberg discovered a number of implicational universals linking clause order to a number of other clausal and phrasal word orders (see also Dryer, 1992).

Greenberg proposed a competing motivations model for word order patterns, using dominance (an inherent preference for one order, e.g. NA and NumN are dominant) and harmony (a parallel alignment of orders, e.g. NA/NNum and AN/NumN). The three language types in the example given above satisfy either harmony (NA/NNum, AN/NumN) or dominance (NA/NumN). The dispreferred type, AN/NNum, is disharmonic and neither order is dominant. In affix order, there is a harmony pattern which competes with a general suffixing preference.

Deeper explanations for word and affix order fall into two types, processing and diachronic. The most recent processing explanation of word order argues that human beings attempt to parse the constituents of a sentence as early as possible and as quickly as possible (Hawkins, 1994). Hawkins successfully predicts a frequency distribution of word order types reflecting degrees of optimality of the attested types for parsing. The processing account implies that harmonic word orders are most optimal, but also longer constituents are more optimally placed after their sister elements. Hawkins also argues for a processing account of affix order, based on psychological research on processing the beginnings and ends of words. The beginning of a word is most salient for word recognition, hence prefixes would interfere with recognition, and so most inflections are suffixes.

Diachronic explanations of word order rest on the fact that the same grammatical constructions are used for different word order patterns. In some cases, there is historical evidence of the extension of the construction from one word order pattern to another. A particularly common pattern is the extension of the noun–genitive construction for prepositional phrases: compare English *inside the house* (preposition–noun) to its historical antecedent *in the side of the house* (noun–genitive). Likewise, there is evidence that the position of inflections as prefixes or suffixes is a consequence of their position as independent words before they became affixed.

GRAMMATICAL CATEGORIES: MARKEDNESS, ECONOMY, AND ICONICITY

Languages vary in the grammatical categories they possess. Nevertheless, universals of grammatical categories exist. One class of such universals is markedness. The concept of markedness in typology differs considerably from the concepts given the same name in other linguistic theories. In typology, markedness is a property of conceptual categories or, more precisely, values of a category (e.g. the values 'singular' and 'plural' of the category

'number'). Typological markedness is a universal property of values of a category. Typological markedness is embodied in universals of the formal expression of the values of a category. Typological markedness accounts for many asymmetries in grammatical expression.

Typological markedness is manifested in two ways, structural coding and behavioral potential. Structural coding is the number of morphemes used to express a value. For example, in English plural is expressed with a suffix (*book-s*) but singular is expressed without any suffix (*book-Ø*). This asymmetry is evidence for the markedness of plural compared to singular. Not all languages are like English, in that they have symmetrical expression of singular and plural. For example, Lithuanian has overt coding of both singular and plural, and Mandarin has zero coding of both (i.e. there is no number distinction). Hence the typological markedness pattern must be formulated as an implicational universal: if a language overtly codes the singular value, then it overtly encodes the plural value.

Behavioral potential is the expression of other inflectional categories, or the range of distribution of the values of the category in question. For example, the English third person singular pronoun has three different forms that distinguish gender (he/she/it), while the third person plural does not distinguish gender (they). This asymmetry is further evidence for the markedness of the plural. Again, not all languages are like English: Spanish has gender distinctions in both singular and plural while Turkish has no gender distinctions in either singular or plural.

A competing motivations model can account for the variation between asymmetric and symmetric expression of category values. The asymmetric pattern is motivated by economy: express no more than one has to. Thus, one value can be expressed by zero. The value that is chosen to be zero is the most frequent value (there is a long established correlation between frequency of use and shortness of form). The symmetric pattern is motivated by iconicity: use a one-to-one mapping between meaning and form in an utterance.

The correlation of frequency with economy accounts for two other typological generalizations: the unmarked form is most likely to be irregular, and the unmarked form is most likely to form the base in the analogical leveling of morphological paradigms. Economic motivation has been explained in terms of an activation network model of morphological organization in which more frequent (i.e. less marked) word forms are more independently entrenched. Economy and iconicity also compete in the storage of morphemes and words and their associated meanings.

Economy and iconicity are both ultimately explainable as processing constraints. Economy is clearly a processing efficiency factor: minimize overt expression where possible, particularly in high-frequency values. Iconicity also represents processing efficiency. A simple mapping of meaning onto form is easier to store and to use in production and comprehension than a noniconic mapping, which will inevitably be more complex.

HIERARCHIES, PROTOTYPES, AND THE SEMANTIC MAP MODEL

Many grammatical categories have more than two values, and often those values are ranked on a hierarchy. For example, many languages have a dual number value as well as singular and plural, which leads to a hierarchy singular < plural < dual, such that singular is the least marked value and dual the most marked value.

Two hierarchies are manifested in many areas of grammar, the extended animacy hierarchy and the grammatical relations hierarchy. The extended animacy hierarchy is a composite hierarchy: first or second person < third person pronoun < proper name < human common noun < nonhuman animate common noun < inanimate common noun. The extended animacy hierarchy can be broken down into three constituent hierarchies, to which may be added the definiteness hierarchy:

- Person (speech act participant): first (speaker), second (addressee) < third (other)
- Referentiality: pronoun < proper name < common noun
- Animacy: human < animate < inanimate
- Definiteness: definite < specific indefinite < nonspecific indefinite

The extended animacy hierarchy allows constrained variation in the conditions on grammatical constructions. For example, in some languages, a verb agrees with only first or second persons, while in other languages, a verb will agree with first or second persons and nouns referring to humans. The hierarchy excludes a language type in which agreement occurs with first or second persons and inanimate nouns but not human or animate nouns.

The extended animacy and definiteness hierarchies operate in any construction that includes reference to a person, animal, or thing, including expression of number and gender; deixis; agreement of the verb, adjective, or noun; and case

marking in grammatical relations between a verb and its dependents.

The grammatical relations hierarchy was first discovered in constraints on the formation of relative clauses. A relative clause expresses the modification of a referent (the head noun) by a proposition (the relative clause itself). For example, in *the book that I wrote last year*, the head noun is book and the relative clause itself is that I wrote last year. There is a great variety of grammatical structures used to encode relative clauses, ranging from nonfinite participial clauses such as English *the child* SLEEPING *on the couch* to relative clauses that contain their heads, as in Imbabura Quechua. However, all relative clauses have in common the fact that the head noun has a grammatical relation to the verb in the relative clause. For example, in *the book that I wrote last year*, *book* is the object of *wrote*.

Languages impose different constraints on the grammatical relations allowed by specific relative clause constructions: for example, the English participial construction allows only subjects. The constraints conform to a single grammatical relations hierarchy (Keenan and Comrie, 1977): subject < direct object < indirect object < oblique < possessor. The grammatical relations hierarchy also controls the encoding of grammatical relations by agreement and case marking.

However, not all languages categorize grammatical relations in the same fashion. English categorizes the subject of intransitive verbs and the subject of transitive verbs in the same way but categorizes the object of transitive verbs differently. This pair of categories is implicit in the traditional terms 'subject' and 'object'; the terms 'nominative' and 'accusative' are also used. But many languages categorize the subject of intransitive verbs and the object of transitive verbs in the same way and the subject of transitive verbs differently. These categories are called 'absolutive' and 'ergative' respectively. The categories in these languages nevertheless conform to a universal hierarchy absolutive < ergative < …

The semantic map model is employed to represent variation in grammatical categories across languages. The semantic map model posits an underlying universal conceptual space. For grammatical relations, this conceptual space includes the semantic roles for transitive subjects (abbreviated A), transitive objects (P), and intransitive subjects (S). Grammatical categories of particular languages are represented as semantic maps on the conceptual space. The grammatical categories for case marking found in the world's languages are given in Figure 1.

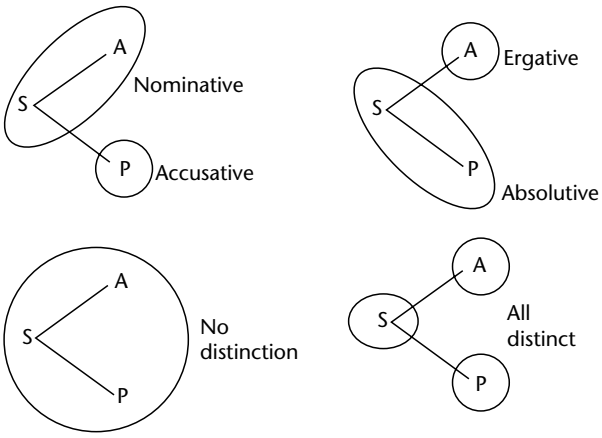


Figure 1. Attested semantic maps of grammatical relations defined by case marking.

Typological universals are represented as constraints on the structure of conceptual space. Grammatical categories may only be mapped onto connected regions of the conceptual space, represented by links between points. This constraint captures the fact that no language encodes A and P in the same way and S distinctly. The conceptual space itself is asymmetric from left to right in Figure 1, representing the hierarchical patterns nominative < accusative and absolutive < ergative. The highest grammatical relation in both hierarchies includes the leftmost role in the conceptual space. There is similar variation in the categories of objects of transitive verbs and ditransitive verbs (verbs with two objects), subject to the same constraints.

The semantic map model has been used to describe grammatical relations, number, aspect, modality, voice, indefinite pronouns, adverbials, and parts of speech. The semantic map model allows one to represent cross-linguistic variation in the grammatical categories of a language while simultaneously capturing universals across all languages regardless of the categories they employ. The conceptual space represents those universals, which are presumed to reflect the structure of the human mind.

Grammatical relations, animacy, and definiteness all interact in a complex way, leading to a wide range of linguistic structures for encoding grammatical relations in a variety of constructions. All of these interactions are subject to universal constraints which can be represented in the semantic map model. A variety of explanations has been offered for the typological generalizations found in the realm of grammatical relations.

In the encoding of grammatical relations, there is an association of the A role with higher animacy

and/or definiteness, and conversely an association of the P role with lower animacy and/or definiteness. This complementary pattern of association describes two prototypes. The prototypical (transitive) subject is a cluster of features <A role, high animacy, high definiteness>. The prototypical object is <P role, low animacy, low definiteness>.

Three explanations have been offered for the grammatical patterns that manifest the subject and object prototypes. A semantic explanation is natural agency: high animacy, particularly human status, is indicative of natural agency, and thus will be zero coded as A but overtly coded as P. A processing explanation is efficiency: a zero coded high animacy referent will be assumed to be A unless overtly coded as a P (or indirect object). Conversely, a zero coded low animacy referent will be assumed to be P unless overtly coded as A. The third explanation involves the pragmatic (discourse) notion of topicality: the subject represents the most topical participant in the verbal event, and highly topical referents are normally human and/or definite. Conversely, nontopical participants are those that are referred to by full noun phrases rather than pronouns (which represent continuing topics). A highly robust cross-linguistic universal is that the transitive subject (A role) is almost always pronominal in spoken language. Full noun phrases are found almost exclusively in the S and P roles, and this may account for the common occurrence of ergative categorization among nouns in contrast to nominative categorization of pronouns.

The typological prototype analysis has also been applied to parts of speech. As noted above, languages vary as to their parts of speech. However, there exist universals in the encoding of semantic classes in the relevant constructions, namely those for the propositional acts of reference, predication, and modification. These constructions are mapped in a conceptual space defined by semantic class and propositional act. Asymmetric typological markedness patterns among semantic maps for word classes across languages reveal the existence of typological prototypes for parts of speech:

- Nouns = <reference, objects>
- Adjectives = <modification, properties>
- Verbs = <predication, actions>

Nonprototypical points in the conceptual space for parts of speech are usually encoded by special grammatical constructions, listed in Table 1.

The prototypes for parts of speech can be accounted for by the characteristics of the semantic classes and their suitability for the discourse functions of the propositional acts. Object identity is unchanging and lasting over time. Objects are therefore most suitable for the function of reference, which opens a discourse file for an entity that will be used for an extended piece of discourse. Actions are dynamic, transitory processes. Actions are therefore most suitable for predications, which provide a continuous sequence of assertions in a discourse. Properties are values associated with objects on a single semantic dimension, and are most suitable as modifiers, which qualify the description of a referent. Other combinations of semantic class and discourse function are possible (see Table 1), but they are less suitable and hence are typologically marked. This is essentially a processing explanation for the typological generalizations underlying parts of speech constructions.

ICONIC MOTIVATION AND SYNTACTIC STRUCTURE

A number of aspects of syntactic structure conform to typological generalizations. These aspects can be accounted for by iconic motivation, this time in terms of syntactic structure and not merely categorization. Of these, the most robust universal is linguistic (grammatical) distance. Linguistic distance can be measured in terms of the degree of morphological or syntactic integration of two elements in a construction. Languages often have two or more constructions that differ structurally in terms of linguistic distance. These two constructions often have semantic differences in their use as well. The semantic differences sometimes define a difference in conceptual distance.

Table 1. Constructions in nonprototypical points of the conceptual space for parts of speech

	Reference	*Modification*	*Predication*
Objects	unmarked nouns	genitive, adjectivalizations, PPs on nouns	predicate nominals, copulas
Properties	deadjectival nouns	unmarked adjectives	predicate adjectives, copulas
Actions	action nominals, complements, infinitives, gerunds	participles, relative clauses	unmarked verbs

The iconic distance generalization is that if two similar constructions differ in linguistic distance and also in conceptual distance, the linguistically more tightly linked construction will be used for the conceptually tighter relation. This generalization has been found to hold for constructions indicating directness of causation, semantic integration of complement clauses to their main verbs, alienability of possession, and distinctness of semantic roles in reflexive events. The generalization also accounts for the ordering of inflectional and derivational affixes in words. The ordering of verbal inflectional affixes follows a hierarchy from nearest to farthest from the root: aspect < tense < mood < person/number agreement, which largely conforms to the degree of semantic effect of the inflection on the meaning of the root. More generally, derivational affixes, which substantially alter the meaning of the root, are closer to the root than inflectional affixes.

DIACHRONIC TYPOLOGY AND GRAMMATICALIZATION

Language types represent current states of language structure in a speech community. However, languages change, and language types therefore change as well. In diachronic typology, language types are treated as evolving entities. Diachronic typology involves two theoretical shifts. The first is the dynamicization of typology. It chiefly involves the reinterpretation of competing motivations as dynamic processes resulting in cross-linguistic variation, and the reinterpretation of semantic maps as the reflection of paths of change through conceptual space. The second shift is the discovery of universals of language change, in particular universals of grammaticalization.

Grammaticalization is the evolution of grammatical morphemes and constructions from full words and nonconventionalized grammatical structures. The most common grammaticalization processes are:

- full verb > auxiliary > tense–aspect–mood affix
- noun > reflexive pronoun > middle/passive voice affix
- verb > adposition
- noun > adposition
- adposition > case affix
- adposition > subordinator
- verb > subordinator
- emphatic personal pronoun > clitic pronoun > agreement affix
- cleft sentence marker > highlighter
- noun > classifier
- verb > classifier
- demonstrative > article > gender/class marker
- demonstrative or article > complementizer or relativizer
- collective noun > plural affix
- numeral 'one' > indefinite article
- numerals 'two' or 'three' > dual/paucal/plural affix

Grammaticalization involves three more or less simultaneous, unidirectional changes: phonological, morphosyntactic, and functional (semantic/pragmatic). Recent research has refined and narrowed the exact set of processes that are part of grammaticalization, in the light of putative counterexamples that have been posed in the literature.

The phonological processes involved in grammaticalization are reduction of phonological form and evolution from an independent word to a bound morpheme (compound or affix). The morphosyntactic processes in grammaticalization are the fixing of the word order of the construction; making the relevant morphemes obligatory; a reduction in the range of forms that can occur in the construction; and the eventual loss of independent syntactic status of the relevant morphemes. The functional processes are the idiomatic specialization of the meaning of the construction and the shift from a 'lexical' meaning to a 'grammatical' one. Grammaticalization can be illustrated by the evolution of the English motion construction *She is going to the store* to the future *She's gonna become an architect*. The phrase *is going to* is phonologically reduced to *'s gonna*, and *to* has lost its independent syntactic status. Only the contracted auxiliary is found with *gonna*. The meaning of the grammaticalized construction is specialized, and indicates future tense rather than a motion event.

The grammaticalization process can be explained as a cycle of change driven by the nature of communication. The first step in the cycle is the use of a novel, unconventionalized periphrastic expression for a particular meaning to avoid misunderstanding. The second step in the cycle is the conventionalization of that expression as the normal way to convey that meaning. The third and final step is the erosion or reduction of the that expression, phonologically and morphosyntactically.

CONCLUSION

Typological research has uncovered a wide range of empirical generalizations about grammatical structure by studying a representative sample of linguistic diversity. The universals discovered cover word and affix order, grammatical categories and relations, syntactic structures and processes of grammatical change. Explanations offered for the

universals uncovered by typology are formulated in terms of language processing, historical processes, and the coding of semantic and discourse- pragmatic concepts by grammatical categories and constructions. Although a significant body of empirical results has been produced in the past four decades, important language universals continue to be discovered.

References

Comrie B (1989) *Language Universals and Linguistic Typology*, 2nd edn. Chicago, IL: University of Chicago Press.

Denning K and Kemmer S (1990) *On Language: Selected Writings of Joseph H. Greenberg*. Stanford, CA: Stanford University Press.

Dryer M (1992) The Greenbergian word order correlations. *Language* **68**: 81–138.

Greenberg JH (1966) *Language Universals, with Special Reference to Feature Hierarchies*. The Hague: Mouton.

Hawkins JA (1994) *A Performance Theory of Order and Constituency*. Cambridge, UK: Cambridge University Press.

Keenan EL and Comrie B (1977) Noun phrase accessibility and universal grammar. *Linguistic Inquiry* **8**: 63–99.

Further Reading

Bybee JL, Perkins RD and Pagliuca W (1994) *The Evolution of Grammar*. Chicago, IL: University of Chicago Press.

Croft W (2001) *Radical Construction Grammar: Syntactic Theory in Typological Perspective.* Oxford, UK: Oxford University Press.

Croft W (2001) *Typology and Universals*, 2nd edn. Cambridge, UK: Cambridge University Press.

Givón T (1979) *On Understanding Grammar*. New York, NY: Academic Press.

Haiman J (1985) *Natural Syntax*. Cambridge, UK: Cambridge University Press.

Haspelmath M (2002) The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In: Tomasello M (ed.) *The New Psychology of Language*, vol. 2. Mahwah, NJ: Lawrence Erlbaum Associates.

Heine B, Claudi U and Hünnemeyer F (1991) *Grammaticalization: A Conceptual Framework*. Chicago, IL: University of Chicago Press.

Hopper P and Thompson SA (1980) Transitivity in grammar and discourse. *Language* **56**: 251–299.

Hopper P and Traugott EC (1993) *Grammaticalization*. Cambridge, UK: Cambridge University Press.

Acalculia

Intermediate article

John Whalen, University of Delaware, Newark, Delaware, USA

CONTENTS

Impairment of the ability to calculate; research into numerical impairments in patients with brain damage has revealed that human calculation abilities are composed of several distinct cognitive processes.

INTRODUCTION

What are the brain processes that underlie our fundamental numerical abilities such as calculation, estimation, and reading and writing numbers? A particularly successful line of research in answering this question has come from the study of impairments in numerical abilities as a result of brain injury.

While it was once thought that there was one calculation center in the brain that could be impaired (hence the term 'Acalculia' – the inability to calculate), it is now believed that there are many distinct numerical abilities. This change in our understanding is based on evidence that only some arithmetic abilities may be impaired as a result of brain injury, while others are spared. Individual brain-injured patients show some remarkably specific impairments. Note that these are not isolated cases: there are several documented reports of people with similar patterns of performance.

NUMERICAL IMPAIRMENTS AFTER BRAIN DAMAGE

Numeral Comprehension and Production Processes

A distinction has been drawn between the ability to perform calculations and the ability to comprehend and produce numbers. Patients such as the person known as D. R. C. reveal striking impairments after brain damage, for example the inability to remember simple arithmetic facts such as $2 \times 4 = 8$, even though other abilities such as comprehending and producing written and spoken numerals are unimpaired. This pattern implies the brain systems for comprehending and producing numerals are separate from those for calculation. Several cases with the same as well as the opposite pattern of impairment have been reported.

The ability to comprehend and produce numbers is also composed of several functional subcomponents. For example, some patients reveal highly specific impairments strictly limited to writing arabic numerals, such as writing down '100206' in response to hearing 'one hundred and twenty-six'. This example reveals that numerical processing has several distinct components, including numeral comprehension and numeral production (only production was impaired), verbal and arabic numeral processes (only arabic numeral production was impaired), and the ability to retrieve a single digit and the ability to put together several digits to make a larger number within one component (only number composition was impaired).

Dissociations between number and language processing

Skills such as reading 'number words' (e.g. 'five', 'sixty') were originally thought to be subsumed by language processes. However, at virtually all levels of processing, dissociations have been observed between numerical and linguistic abilities, suggesting a surprising degree of specificity in the human brain. There are several case reports of near-total impairment in one domain, with remarkable sparing of similar function in another. One brain-damaged patient reveals the ability to write numbers with remarkable ease, even performing complex arithmetic, while failing to be able to write simple words and even their own name. In contrast, other patients reveal word reading to be largely spared, while 'number word' reading was almost totally lost as a result of brain damage.

Localization of numeral comprehension and production processes

The brain regions involved in comprehending and producing written and spoken numerals generally

mirror those involved in written and spoken language, despite the apparent functional uniqueness of number processing. Impairment in producing spoken and written numerals (both word and arabic forms) nearly always originates from damage to left hemisphere language production areas, while impairments in comprehending spoken numerals typically results from damage localized to left hemisphere language comprehension regions.

In contrast, arabic numeral comprehension is distributed across both brain halves (or hemispheres) more than other numerical and linguistic abilities. Evidence from patients with disconnected brain hemispheres (which permits the study of each brain hemisphere acting independently) reveal excellent comprehension abilities in both hemispheres.

Impairment of Calculation

Evidence from people with brain damage, as well as those with normal and impaired development, has revealed that several independent cognitive processes are required in order to perform calculations such as 274×59. First, there is a distinction between the ability to remember simple arithmetic facts (such as $6 \times 4 = 24$), and the ability to perform calculation procedures (such as those required for carrying numbers in multidigit multiplication). Several patients with brain damage have impairments that are specific to either fact retrieval or multidigit calculation procedures, indicating that the ability to retrieve arithmetic facts and the ability to perform multidigit calculations are represented by different neural substrates.

Within the simple process of remembering arithmetic facts like $6 \times 4 = 24$, there are multiple processes involved. Brain-damaged patients have also revealed selective impairment of the ability to recognize the arithmetic operator (e.g. $+$ or $-$) and the ability to recognize the digits.

The independence of fact retrieval and numerical competence

Evidence from people with brain damage has revealed strong divisions between the ability to remember the answer to simple arithmetic problems (e.g. remembering $2 + 2 = 4$) and the ability to calculate an answer based on arithmetic principles. For example, one patient revealed a marked inability to retrieve previously known facts from memory (e.g. 8×7). However, given enough time the patient was able to answer a problem such as 8×7 by producing an elaborate strategy, such as using the answer to 8×10 and adjusting in accordance with the appropriate algebraic principle: $8 \times 7 = 8 \times 10 - 8 \times (10 - 7)$. Others patients are clearly able recall some facts, but are simply unable to use mathematical principles to derive other answers.

Representing arithmetic facts in memory

There is debate as to the form in which arithmetic facts are stored. One view is that arithmetic facts are stored in an abstract, meaning-based form that is not tied to a specific modality (e.g. spoken or written). In several cases of brain damage the patients have revealed calculation impairments that are consistent regardless of the form in which the problems are presented or answered. Perhaps more surprising is the fact that patients can successfully calculate in spite of severe impairments in representing the spoken form of numerals. For example, a patient can be presented with a problem such as '7×3', read the problem aloud as 'five times eight', say the answer as 'twenty-six' but nevertheless write the correct answer to the original problem: '21'. This pattern of performance appears incompatible with the notion that we store arithmetic facts purely in a sound-based form.

However, an alternative possibility is that arithmetic facts are stored and retrieved in a sound-based representation (like a nursery rhyme). According to this position, abstract magnitude representations are not related to arithmetic fact retrieval, but instead are involved in estimation and mathematical reasoning. Some people with brain damage are unable to determine exact arithmetic responses (e.g. $2 + 2 = 3$) but nevertheless can reject a highly implausible answer such as $2 + 2 = 9$, indicating that there is an approximate number representation which provides the meaning of numbers and may be separate from the process of retrieving exact arithmetic facts (Dehaene, 1997).

Localization of arithmetic processes

The ability to retrieve arithmetic facts from memory seems to be strongly localized to the left hemisphere. Patients with disconnected hemispheres perform at nearly normal levels in their left hemisphere, while having essentially no calculation abilities in the right hemisphere. A majority of fact-retrieval impairments are suffered as a result of left parietal lobe damage; more rarely they are the result of damaged subcortical structures and left language centers. In contrast, the ability to perform multidigit arithmetic appears to draw on the planning and working memory capacity of the frontal lobes (dorsolateral prefrontal cortex).

Exact and Approximate Calculation

Response latencies for determining the larger of two numerals (derived from healthy adults) suggests that we represent numerical quantities in terms of a magnitude representation similar to that used for light brightness, sound intensity and time duration. This magnitude representation is used to compare two or more numbers. Moyer and Landauer (1967) found that humans are faster at comparing numerals with a large difference (e.g. 1 and 9) than comparing two numerals with a small difference (e.g. 4 and 5). Further, when the differences between the numbers are equated (e.g. 2 and 3, versus 8 and 9), the smaller number pair (2 and 3) is compared more quickly than the larger number pair (8 and 9). These findings led to the conclusion that numerals are being translated into a magnitude representation along a mental number line with increasing imprecision the larger the quantity being represented (a psychophysical representation which conforms to Weber's law).

These magnitude representations are found in human adults, infants and children, and also in animals. This magnitude system allows us to meaningfully represent numerical quantities, estimate, and compare numerical quantities. It appears to be related to our representation of other magnitudes, such as time duration and distance (Whalen *et al*., 1999). A crucial challenge for the development of numerical literacy is the formation of mappings between the exact number symbol representations that are learned in school, and the approximate numerical quantity representations that are present very early in life.

There is strong evidence that parietal regions of both brain hemispheres represent numerical magnitude. Patients with disconnected hemispheres have revealed the ability to perform number comparison in either hemisphere (Dehaene, 1997). Electrical brain signatures during number comparison reveal that approximately 0.1 s, after the presentation of two numbers there is bilateral activation of the parietal lobes, and that the signature varies according to the difficulty of the comparison (Dehaene, 1996).

INTEGRATION: THE TRIPLE CODE MODEL OF NUMERICAL PROCESSING

Stanislas Dehaene and colleagues were the first researchers to provide a detailed theory of number processing that includes both the different functional components and their localization in the brain (Dehaene, 1997). According to the 'triple code' model there are three separate number systems in the brain: verbal, visual, and magnitude systems (Figure 1).

The verbal system is located in the left hemisphere adjacent to language areas, and is responsible for comprehending and producing spoken numerals, as well as storing arithmetic table facts in memory. The visual number system takes input from primary visual centers, and is used to recognize and produce numerals in written word and arabic forms. The magnitude system is located in the parietal lobe of both hemispheres, and provides the ability to estimate and compare numbers.

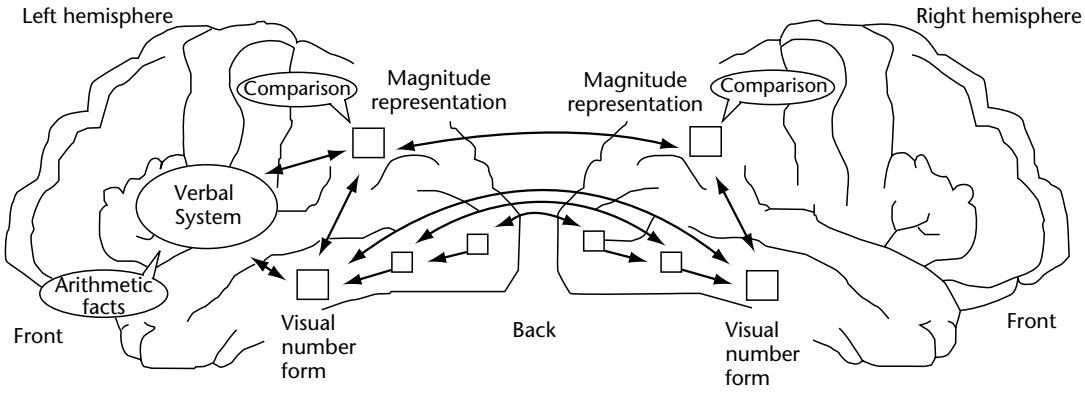


Figure 1. The triple code model of numerical processing is composed of three separate number systems. The verbal system is responsible for comprehending and producing spoken numbers, as well as storing arithmetic table facts such as $2 \times 2 = 4$ in memory. The visual system recognizes numbers written in arabic digit or written-word form. The magnitude system is responsible for deciding which of two numbers is the larger, estimating, and giving an approximate sense of how much a number represents.

According to the triple code model, even simple calculation (e.g. $8 + 9 = ?$) may require several processes and brain regions, including rote verbal memory (verbal system), elaboration (e.g. $8 + 9 = 7 + 10$, magnitude system), and strategy use (frontal lobes). This theory illustrates that even the simplest arithmetic (e.g. $2 + 2 = ?$) may require several processes and brain regions.

In summary, the cases of brain damage reported here reveal that acalculia is not a single impairment. Rather, we now know that many cognitive processes distributed across the brain together provide our arithmetic and numerical competencies. The challenge for the future will be the precise identification of the brain regions involved in arithmetic and the elucidation of how these regions work together to produce our considerable arithmetic abilities.

References

Dehaene S (1996) The organization of brain activations in number comparison: event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience* **8**: 47–68.

Dehaene S (1997) *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Moyer RS and Landauer TK (1967) Time required for judgments of numerical inequality. *Nature* **215**: 1519–1520.

Warrington EK (1982) The fractionation of arithmetical skills: a single case study. *Quarterly Journal of Experimental Psychology* **34A**: 31–51.

Whalen J, Gallistel CR and Gelman R (1999) Non-verbal counting in humans: the psychophysics of number representation. *Psychological Science* **10**: 130–137.

Further Reading

Butterworth B (1999) *What Counts: How Every Brain is Hardwired for Math*. New York: Free Press.

Dehaene S (1997) *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Addiction, Neural Basis of

Intermediate article

Roy A Wise, National Institute on Drug Abuse, Baltimore, Maryland, USA

CONTENTS

Addictive drugs affect brain function by acting at specialized receptors for the endogenous chemical messengers that affect communication between nerve cells. The understanding of addiction depends on our understanding of these receptors, the brain circuits they are embedded in, and the functions that these circuits normally serve.

INTRODUCTION

Addiction is a term that is widely used but poorly defined, even by specialists (Maddux and Desmond, 2000). It is most commonly used to refer to compulsive drug-seeking and drug-taking behaviors, particularly when those behaviors are continued, despite repeated attempts to change, in the face of clearly harmful consequences. The term is also used increasingly to refer to compulsive eating, compulsive gambling, compulsive sexual behavior, and other compulsions that are maintained despite harmful consequences.

The 'official' definitions of the World Health Organization and the *Diagnostic and Statistical Manual of Mental Disorders* of the American Psychiatric Association continue to change as theoretical, legal, and political factors influence the committees in charge. The major problem is that the distinction between 'ordinary' habits and the habits termed 'addiction' is subjective and quantitative rather than based on principle. There was considerable debate in the 1970s as to whether cocaine and nicotine are addictive; by the 1990s it was widely accepted that they are. Part of the problem was theoretical; in the 1950s and 1960s physiological dependence was taken as the defining property of addiction. However, it has become clear that cocaine and nicotine can establish habits as compulsive and harmful as those associated with heroin and alcohol, despite their lack of dramatic dependence syndromes. As the conceptual strength of dependence theory eroded (see below), addiction became defined less by physiological criteria than by subjective ones. The criteria for the terms 'compulsive' and 'harmful', central to the contemporary definition of addiction, are largely matters of personal judgment.

VIEWS OF ADDICTION

Negative Reinforcement View

Lower animals can be trained to work compulsively for intravenous injections of such drugs as cocaine, nicotine, amphetamine, and heroin; thus these drugs can serve as reinforcers, stamping in response habits much as food does for a hungry animal. Despite the fact that lower animals are not subject to the peer pressures or the stresses of poverty or city life that are often blamed for human addiction, it appears that all mammals are at risk of addiction to such drugs as cocaine and heroin. Laboratory rats and monkeys learn to self-administer intravenous heroin and will do so to the point of physical dependence; they learn to self-administer intravenous cocaine, and will do so to the point of death. Thus the observation that addictive drugs are reinforcers has become the common denominator of contemporary addiction theory.

A drug can be viewed as reinforcing because it is an extra 'treat' in one's life – like an after-dinner mint – or because it is a 'treatment' for a need state – like the aspirin that alleviates headache or the meat and potatoes that alleviate hunger and restore energy balance. The once-dominant negative reinforcement view of addiction held that drug-taking becomes compulsive when the nervous or metabolic system has adapted to continued use of a drug and the drug has become necessary for normal bodily homeostasis. The obvious and objective physiological distress of an opiate addict in the early stages of drug withdrawal – the defining evidence of physical dependence – gave rise to this view (dependence theory), which dominated

addiction theory until quite recently (Wise, 1987). In this view, initial drug-taking was attributed to peer pressure, thrill-seeking, or simple boredom, but subsequent drug-taking was seen as reflecting the need to self-medicate the withdrawal syndrome that soon developed. Because the dependence syndrome becomes stronger with continued drug use, apparent tolerance develops to dependence-producing drugs and progressively stronger doses of the drug are required to alleviate withdrawal distress. A variation of dependence theory, the self-medication hypothesis, raised the possibility that some individuals have preexisting conditions of stress or anxiety that, like withdrawal symptoms, are medicated by addictive drugs. This view was offered to explain why not all individuals who try addictive drugs come to take them compulsively. The view that all addiction reflects self-medication of preexisting distress syndromes has largely been discounted on evidence that happy, healthy animals, healthy suburban human adolescents – and, indeed, physicians – appear to be as readily addicted to cocaine or opiates as are the inner-city adolescents who were once the primary concern of addiction specialists.

Dependence theory, at least in its classic form, has largely been discredited as a sufficient explanation of addiction. Injections of heroin, the prototypical addictive drug, can establish compulsive drug-seeking and drug-taking even when it is available for too short a portion of each day to establish the classic opiate dependence syndrome (Deneau *et al*., 1969). Indeed, it has been found that the classic opiate dependence syndrome is alleviated by drug injections into the periaqueductal gray matter, whereas the reinforcing effects of the drug are caused when the drug is injected into the nearby ventral tegmental area (Bozarth and Wise, 1984). People with alcoholism often forgo alcohol during periods of maximum withdrawal distress, only to begin working for alcohol when withdrawal stress has largely subsided. Thus the notion that drug-taking is compulsive only when needed to alleviate withdrawal distress does not fit with the basic facts of even the most classic addictions: those of opiates and alcohol.

Nor does classic dependence theory fit with the facts of compulsive use of nicotine, cannabis, or cocaine. Extensive use of these drugs does not produce the classic somatic dependence syndromes seen with the opiates, barbiturates, benzodiazepines, or alcohol. The withdrawal symptoms associated with barbiturates, benzodiazepines, and alcohol are similar for all three classes of drugs, and are alleviated by drugs from any of the other classes, suggesting a common mechanism for dependence. The withdrawal symptoms associated with opiates are similar, and are partially alleviated by these agents and thus at least partially mediated by the same mechanisms. However, withdrawal from cocaine, amphetamine, nicotine or cannabis produces mild symptoms by comparison, and in the case of cocaine and amphetamine, the somatic withdrawal symptoms are the converse of those of opiates. The general mood associated with opiate withdrawal is hyperexcitability and irritability, whereas cocaine or amphetamine withdrawal is associated with hypoexcitability and depression.

Thus, although there is a withdrawal syndrome associated with termination of regular use of many drugs, there is no common somatic withdrawal syndrome that can explain the compulsive use of the full range of addictive drugs. For these reasons, classic dependence theory is no longer accepted as an explanation or a defining property of addiction (Wise, 1987). As discussed below, however, a variant of negative reinforcement theory remains influential.

Positive Reinforcement View

An alternative view of addiction holds that the primary motivation for drug-taking is the seeking of euphoria or a drug 'high' (McAuliffe and Gordon, 1974). Rather than simply returning the addicted individual to a normal feeling state, this view holds that the drug is taken because it produces a better-than-normal feeling state. This view is strengthened by the fact that humans and lower animals will work for electrical stimulation of certain brain regions, stimulation that brings a state of pleasure which satisfies no biological need and was never experienced in mammalian evolutionary history. The view that addictive drugs produce elevated states of pleasure also explains, as dependence theory does not, why addictive drugs are strongly habit-forming prior to development of dependence, and why there is such a strong probability of relapse after detoxification.

The negative and the positive reinforcement views are not mutually exclusive. The alleviation of withdrawal distress certainly brings pleasure, and it is difficult to imagine that a larger dose than is needed to alleviate pain would not be desirable to someone who is self-medicating. This would explain the fact that people given methadone to medicate withdrawal distress still desire and often use street heroin when it is available. In the case of heroin addiction, it is reported that

initial drug-taking results in drug euphoria, but that as chronic use progresses drug euphoria and positive reinforcement become progressively weaker while withdrawal-associated dysphoria and negative reinforcement become progressively more dominant factors in the maintenance of compulsive drug-seeking.

ADDICTION AND THE MESOLIMBIC DOPAMINE SYSTEM

The 1970s, when injected amphetamine and intranasal cocaine were becoming increasingly popular, brought the positive reinforcement view to the forefront of addiction theory. Cocaine users reported that cocaine caused euphoria and that no great distress was felt when the available supply of the drug had been used up. Nonetheless, the drug was taken compulsively for as long as it was available. It was widely accepted that this drug, when used by normal and successful individuals, was taken to 'get high' and not to self-medicate a withdrawal state or any preexisting abnormal distress state. As cocaine came to be taken by injection or by smoking freebase or 'crack' cocaine, it became much more obvious that this drug was strongly addictive. While many people became addicted to intranasal cocaine, intravenous cocaine or smoked freebase caused addiction much more rapidly, and 'graduation' from nasal use to smoking or injecting became a frequent consequence. Rats and monkeys given unlimited access to intravenous cocaine will take the drug to the point of death, losing a third or more of their body weight with a week or two of drug self-administration, punctuated by minimal sleep and food intake. This weight loss and sleep disturbance are not a withdrawal syndrome, however; they are exacerbated rather than alleviated by continued intoxication.

Cocaine is an inhibitor of monoamine neurotransmitter reuptake. The monoamine neurotransmitters are noradrenaline (norepinephrine), dopamine and serotonin; cocaine shares with amphetamine the ability to elevate the concentration of each of these substances in the junctions between communicating neurons. Whereas amphetamine causes the direct release of these transmitters, cocaine blocks their synaptic inactivation by blocking their reuptake by the cells that released them (thus prolonging and elevating their actions). The drug-induced elevation of extracellular dopamine levels accounts for the habit-forming effects of the cocaine and amphetamine; drugs that block the effects of dopamine block the ability of cocaine or amphetamine to establish drug-seeking response habits or preferences for the places in the environment where the drug has been experienced. Lesions of the mesolimbic branch of the forebrain dopamine projections also block the reinforcing effects of these drugs. Lesions or pharmacological blockade of the noradrenergic or serotonergic systems have no such effects. Dopamine-blocking drugs also eliminate the ability of normally rewarding brain stimulation to establish or maintain intracranial self-stimulation. Thus cocaine and amphetamine are habit-forming because they activate the reward system that was once termed a 'pleasure center in the brain' (see Wise and Bozarth, 1987).

Food, sex, and rewarding brain stimulation, as well as the addictive drugs amphetamine, cocaine, morphine, heroin, nicotine, cannabis, and alcohol, each elevate brain dopamine levels in the nucleus accumbens, where the mesolimbic dopamine system has its main termination. When cocaine, amphetamine, or heroin is self-administered, dopamine levels are elevated severalfold; animals allowed to self-administer these drugs do so whenever their dopamine levels fall to about 200% of normal. Food for hungry animals and sex for experienced males cause dopamine levels to increase, but the increases tend to peak at 150–200% of normal. Thus animals self-administering cocaine, amphetamine or heroin, at least, maintain their dopamine levels at higher values than are usually produced by the normal pleasures of life (Wise, 1998).

SITES OF THE REINFORCING ACTIONS OF ADDICTIVE DRUGS

The major classes of addictive drug act at receptors that are the normal targets of endogenous neurotransmitters or neuromodulators. Nicotine acts at the nicotinic class of receptors for the neurotransmitter acetylcholine. Cannabis acts at the receptors for the neuromodulator anandamide. Phencyclidine acts at the *N*-methyl-D-aspartate (NMDA) class of glutamate receptor. Opiates act at mu, delta, and kappa opioid receptors, receptors for the endogenous opioids enkephalin, β-endorphin, dynorphin, and endomorphin. Cocaine and amphetamine act at transporters for dopamine, noradrenaline, and serotonin, elevating the extracellular levels of these transmitters which then act at their own endogenous receptors.

Each of these drugs acts in multiple anatomical pathways, and thus is involved in multiple physiological functions. The sites of reinforcing actions of some addicting drugs have been localized, and each thus far identified is associated with the mesolimbic dopamine system or with the cells it targets

in the nucleus accumbens. Nicotine's reinforcing action involves nicotinic cholinergic receptors localized to the mesolimbic dopamine neurons themselves, stimulating these cells to fire and to release dopamine. Morphine and heroin have reinforcing actions on cells containing γ-aminobutyric acid (GABA) that are found near to the dopamine cells and that normally inhibit dopaminergic cell firing. The opiates inhibit the GABA-containing cells, thereby disinhibiting the dopaminergic cells and increasing their firing rates. The opiates also inhibit, as does dopamine, the medium-sized spiny output neurons of nucleus accumbens.

The psychomotor stimulants amphetamine and cocaine have their reinforcing actions at the dopamine transporters in the nucleus accumbens. Amphetamine reverses the transporter causing it to expel rather than take up dopamine, and cocaine blocks the transporter, blocking the reuptake of dopamine released from cell firing. Dopamine, in turn, acts at its own receptors on nucleus accumbens neurons. Phencyclidine blocks NMDA-type receptors for the excitatory amino acid transmitter glutamate, thus reducing the excitatory input to the medium-sized spiny output neurons of nucleus accumbens. Thus the direct or indirect inhibition of medium spiny neuron output from nucleus accumbens appears to be the critical common consequence for the reinforcing actions of each of these addictive drugs. Phencyclidine is also habit-forming when injected into the medial prefrontal cortex. Again, the drug's ability to block NMDA receptors accounts for its rewarding action, but it is not yet known which population of NMDA receptors or which type of cortical neuron is involved (Wise, 1998).

The sites of reinforcing actions of cannabis, alcohol, barbiturates, benzodiazepines, and caffeine remain to be identified. Some but not all of these are expected to prove to have rewarding actions involving the mesolimbic dopamine system or its associated circuitry.

CONTEMPORARY THEORIES OF ADDICTION

Classic dependence theory, attributing the compulsive dimension of addiction to the need to alleviate the somatic, largely autonomic, symptoms of withdrawal distress, has been described above. While the conscious desire to alleviate somatic withdrawal distress doubtless contributes to drug-taking in people who are addicted, it is no longer seen as the defining property of addiction. Contemporary addiction theories stress the actions of addictive drugs in the reward circuitry of the brain: the circuitry that is activated by the natural pleasures of life as well as by the laboratory rewards of brain stimulation and addictive drugs.

The psychomotor stimulant theory of addiction (Wise and Bozarth, 1987) suggests that the dominant sedative effect of a number of addictive drugs is unrelated to the addictive liability of these drugs. Rather, it postulates that even the addictive depressants and sedatives (e.g. opiates, alcohol, cannabis) activate the brain circuitry of arousal and forward locomotion. Forward locomotion – the central behavioral component of approach behavior – has been postulated to be the unconditioned response to all positive reinforcers. In the cases of amphetamine and cocaine, the prototypic psychomotor stimulants, drug-induced locomotion and stereotyped orofacial movements associated with activation of the mesolimbic and adjacent nigrostriatal dopamine systems dominate the behavior of intoxicated animals. With opiates and alcohol this action is not obvious because it is masked by the dominant depressive actions of strong doses of these drugs. These depressants activate the mesolimbic dopamine system, however; they cause locomotion and act as stimulants at low doses or in the early stages of intoxication, when only low levels of the drug have reached the brain. The psychomotor stimulant theory attributes the habit-forming effects of addictive drugs to their ability to activate the brain mechanisms of reward and approach behaviors, and attributes the high-dose depressant effects of the drugs to actions in other parts of the brain. The psychomotor stimulant theory fits well with current data on cocaine, amphetamine, opiates, and nicotine. While alcohol and cannabis each activates the postulated reward circuitry, the sites and circuitry through which they do so are not yet known. Whether barbiturates or benzodiazepines activate this circuitry is controversial, and evidence suggests that the habit-forming effects of caffeine are likely to involve an independent mechanism. Thus the psychomotor stimulant theory offers a unified theory for some, but probably not all, addictive substances.

The psychomotor stimulant theory also introduced the notion of 'incentive motivation' to addiction theory. Incentive motivation is a construct from learning theory, designed to deal with the circularity of reinforcement theory. A reinforcer is an object or event which, when made contingent upon a given act by a given animal, increases the probability of recurrence of that act. To suggest that reinforcement explains the act it is defined by is circular; it implies the teleological view that the

cause of an event can follow the event. It is the reinforcement history, not the coming reinforcer, that explains the probability of the act in question. Incentive motivational theory was an attempt to identify the precipitating precursor of the habitual act. It stressed the role of incentives that are present prior to the act – incentives that are attractive because of their past association with the drug – in eliciting and guiding the act. Incentive motivation is the principle by which addictive drugs are seen to establish conditioned place preferences: learned preferences for the portions of the environment where the drug has previously been experienced.

Advocates of the psychomotor stimulant theory pointed out – largely on the basis of observations from brain stimulation reward studies – that reinforcers not only have the ability to increase the probability of recurrence of acts that precede them, but that they (and their associated environmental predictors) have the ability to 'prime' a response habit, energizing the animal and focusing attention on the previously established habit. This proactive feature of reinforcers is the cornerstone of animal models of relapse to addiction. Because it encompasses the ability of a reward to precipitate anticipatory excitement and behavioral arousal as well as to consolidate the memory trace for recent acts, the psychomotor stimulant theory addresses drug reward (embracing both the proactive and the retroactive processes) rather than simply drug reinforcement.

Opponent process theory (Solomon and Corbit, 1973) is a negative reinforcement theory, a general view that subsumes classic dependence theory. However, whereas classic dependence theory attributed compulsion to the need to medicate somatic withdrawal distress involving the autonomic nervous system, contemporary opponent process theories (Dackis and Gold, 1985; Koob and Bloom, 1988) attribute the compulsive nature of addiction to neuroadaptations in the reward circuitry discussed above. The core postulate of classic opponent process theory is that homeostatic controls adjust in compensation for chronic intoxication, opposing the direct effects of the drug and desensitizing the nervous system to that drug. Such opponent processes are found in all the targets of a given drug, including the thermoregulatory system and neurotransmitter receptors in the gut. It is the unopposed effect of the postulated opponent-process neuroadaptations that explains the fact that the withdrawal symptoms associated with a given drug are the opposite of the acute effects of the drug. Thus, while one of the direct effects of opiates is to constrict the intestine and cause constipation, the compensatory relaxation, which is masked so long as the drug continues to oppose it, results in diarrhea when the drug wears off.

The contemporary versions of opponent process theory hold that the reward pathway itself becomes desensitized by chronic drug intoxication, and becomes progressively more difficult to activate by both normal rewards and by the addictive drug itself. This results in loss of responsiveness to (and, as a consequence, loss of interest in) the normal pleasures of life. It is also seen to result in the need to escalate the dosage in order to achieve the expected drug effect. Opponent process theory offers an explanation of tolerance and dependence, not only in the reward pathway but also in the autonomic nervous system and all the systems activated by a given drug. An early version of this theory that focused on the reward system was the dopamine depletion hypothesis (Dackis and Gold, 1985), which held that one consequence of such opponent processes is a decrease in extracellular levels of mesolimbic dopamine during psychomotor stimulant withdrawal. While there remains controversy as to its magnitude and significance, depletion of extracellular dopamine has been reported in animals withdrawn from cocaine, amphetamine, and opiates. Moreover, there is now evidence of development of multiple drug-induced opponent processes in the reward pathway.

First, animals undergoing withdrawal from amphetamine or cocaine have elevated brain stimulation reward thresholds; that is, it takes stronger stimulation of the reward system to motivate an animal undergoing withdrawal from psychomotor stimulants. While there is controversy as to whether there is tolerance or its opposite (sensitization) to the reward-specific effects of addictive drugs, under some circumstances animals have been shown to escalate drug intake as drug exposure is extended. Also, there is now considerable evidence for intracellular neuroadaptations within the dopamine system and within its target neurons in the nucleus accumbens following repeated cocaine and morphine treatments (Nestler and Aghajanian, 1997). Some of these neuroadaptations can be mimicked experimentally, and they reduce the effectiveness of cocaine in producing conditioned place preferences. Thus it is clear that opponent processes are called into play by chronic use of addictive drugs. It remains to be determined how important a role these neuroadaptations have in compulsive drug-taking. The neuroadaptations are clearly consequences of chronic drug intake, but it is not clear to what extent they become causes of subsequent intake.

Incentive salience theory (Robinson and Berridge, 1993), like contemporary opponent process theory, builds on the psychomotor stimulant theory, attributing the habit-forming actions of drugs of abuse to their ability to activate the reward circuitry associated with the mesolimbic dopamine system. Incentive salience theory, however, differs from opponent process theory in that it invokes proponent processes to explain the compulsive nature of addiction. It is well known that the psychomotor stimulant effects of amphetamine, cocaine, and opiates become progressively stronger with repeated intermittent intoxication. The incentive salience theory postulates that progressively increasing sensitivity to the drug gives drug-associated incentives progressively more control over behavior, and that it is this sensitization that makes drug-seeking compulsive. This sensitization has several known correlates, all involving the mesolimbic dopamine system and its associated afferents and efferents. Again, however, it is not yet clear what causal role psychomotor sensitization plays in the increasingly compulsive drug-seeking habits of addiction.

A major issue highlighted by incentive salience theory is the relative importance of the reinforcing effects of the drug (effects of the drug after it has been taken) and incentive motivational effects of the drug (effects related to the drug history and environmental associations with that history). Incentive salience theory calls attention to this underappreciated incentive motivational postulate of traditional learning theory by relating the specialist terms 'incentive motivation' and 'reinforcement' to their subjective correlates 'wanting' and 'liking'. One wants the drug prior to having it: this is the cognitive correlate of incentive motivation and is presumably elicited by drug-associated stimuli that are present before the drug is. One likes the drug after having it: this is the presumed cognitive correlate of the reinforcement that stamps in the stimulus–stimulus associations between the drug actions and the environmental stimuli in the situation in which the drug is encountered. Opponents of this view point out that reinforcing events need not involve conscious pleasure: monkeys have been trained to lever-press for painful footshock, and we can learn to drink initially noxious solutions if they contain alcohol.

CONCLUSION

Much has been learned – though much remains to be learned – about the neural basis of addiction. Many addictions result from the ability of the addictive drug to activate primitive brain circuitry involved in the formation of simple and normal response habits such as feeding and sexual activity. Most addictive drugs activate this system and do so more strongly than do the normal pleasures of life. The activation of this system by addictive drugs seems to be associated, at least initially, with pleasure, though the nature of the pleasure is vaguely and variously described by human subjects.

Repeated drug use leads to adaptations within the central and peripheral nervous systems, and these adaptations are usually opposite in direction to the acute effects of the drugs that produce them. When a drug is taken often enough to build up significant neuroadaptations, withdrawal symptoms opposite to the effect of the drug are seen when drug use is terminated. Such withdrawal symptoms can be aversive, and the self-medication of these withdrawal symptoms is reported to be a significant factor in the inability of people with addiction to simply discontinue the use of opiates and barbiturates. Against the claim that self-medication is a sufficient explanation of continued heroin use, however, is the fact that people given methadone to medicate withdrawal distress often continue to use street heroin when it is readily available. Alleviation of withdrawal symptoms seems to play a minimal part in other addictions, and compulsive drug-seeking can be established with some drug regimens that cause minimal signs of addiction.

If repeated use is sufficiently intermittent, the stimulant effects of addictive drugs can become progressively stronger. It has been suggested that this might explain the progressively stronger control of behavior by addictive drugs. Drug tolerance and drug sensitization have each been demonstrated, but the optimal conditions for the two are different. Tolerance most clearly results from chronic intoxication, whereas sensitization results from intermittent intoxication. Since addiction involves periods of prolonged intoxication and also periods of intermittent drug withdrawal, each factor seems likely to play some part in addiction. Pavlovian conditioning – association of the drug with the environment in which the drug is experienced – is known to contribute both to tolerance and to sensitization, and the degree to which each results from simple pharmacological exposure or rather to the interaction of the drug with the environment remains to be clarified.

It is widely assumed that stressful life experiences predispose some individuals to addiction, and the role of stress in addiction is a topic of

great contemporary interest. Some forms of stress can clearly reinstate drug-taking behavior in animal models, and a good deal of attention is turning from the factors that establish initial drug habits to the factors that can reinstate these habits once they have been broken. A priming 'taste' of the drug itself (such as just one drink or just one cigarette) is one of the strongest reinstating stimuli, and some stress stimuli are comparably effective. The mechanisms of interaction between stress systems and reward systems have not yet been identified.

The cognitive correlates of the various stages of addiction are complex and not well defined or identified. Pleasure, euphoria, craving, stress, wanting, liking, and satiety are terms frequently attributed to phases of addiction, but it remains to be determined which of these relate to causal factors and which to after-the-fact correlates.

References

Bozarth MA and Wise RA (1984) Anatomically distinct opiate receptor fields mediate reward and physical dependence. *Science* **224**: 516–518.

Dackis CA and Gold MS (1985) New concepts in cocaine addiction: the dopamine depletion hypothesis. *Neuroscience and Biobehavioral Reviews* **9**: 469–477.

Deneau G, Yanagita T and Seevers MH (1969) Self-administration of psychoactive substances by the monkey: a measure of psychological dependence. *Psychopharmacologia* **16**: 30–48.

Koob GF and Bloom FE (1988) Cellular and molecular mechanisms of drug dependence. *Science* **242**: 715–723.

Maddux JF and Desmond DP (2000) Addiction or dependence? *Addiction* **95**: 661–665.

McAuliffe WE and Gordon RA (1974) A test of Lindesmith's theory of addiction: the frequency of euphoria among long-term addicts. *American Journal of Sociology* **79**: 795–840.

Nestler EJ and Aghajanian GK (1997) Molecular and cellular basis of addiction. *Science* **278**: 58–63.

Robinson TE and Berridge KC (1993) The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Research Reviews* **18**: 247–292.

Solomon RL and Corbit JD (1973) An opponent-process theory of motivation: II. Cigarette addiction. *Journal of Abnormal Psychology* **81**: 158–171.

Wise RA (1987) The role of reward pathways in the development of drug dependence. *Pharmacology and Therapeutics* **35**: 227–263.

Wise RA (1998) Drug-activation of brain reward pathways. *Drug and Alcohol Dependence* **51**: 13–22.

Wise RA and Bozarth MA (1987) A psychomotor stimulant theory of addiction. *Psychological Review* **94**: 469–492.

Further Reading

Berke JD and Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**: 515–532.

Carlezon WAJ, Thome J, Olson VG *et al*. (1998) Regulation of cocaine reward by CREB. *Science* **282**: 2272–2275.

Goldstein A (1994) *Addiction: From Biology to Drug Policy*. New York, NY: Freeman.

Nestler EJ (1997) Molecular mechanisms of opiate and cocaine addiction. *Current Opinion in Neurobiology* **7**: 713–719.

Robbins TW and Everitt BJ (1999) Drug addiction: bad habits add up. *Nature* **398**: 567–570.

Shaham Y and Stewart J (1995) Stress reinstates heroin-seeking in drug-free animals: an effect mimicking heroin, not withdrawal. *Psychopharmacology* **119**: 334–341.

Stewart J and Eikelboom R (1987) Conditioned drug effects. In: Iversen LL, Iversen SD and Snyder SH (eds) *Handbook of Psychopharmacology*, pp. 1–57. New York, NY: Plenum.

Wise RA (1996) Neurobiology of addiction. *Current Opinion in Neurobiology* **6**: 243–251.

Wise RA (2000) Addiction becomes a brain disease. *Neuron* **26**: 27–33.

Wise RA, Newton P, Leeb K *et al*. (1995) Fluctuations in nucleus accumbens dopamine concentration during intravenous cocaine self-administration in rats. *Psychopharmacology* **120**: 10–20.

Aggression and Defense, Neurohormonal Mechanisms of

Introductory article

R J Blanchard, University of Hawaii, Honolulu, Hawaii, USA
C Markham, University of Hawaii, Honolulu, Hawaii, USA
D C Blanchard, University of Hawaii, Honolulu, Hawaii, USA

CONTENTS

Aggression and defense are important behavioral strategies throughout the animal kingdom, and are linked to activation of brain areas in the prefrontal cortex, limbic system, hypothalamus, and periaqueductal gray matter. While substantial progress has been made toward understanding the neural and neurotransmitter systems underlying defense, and the neurotransmitter and hormonal systems involved in aggression, the relationship between aggression and defense systems is poorly understood, as are the applicability of these findings to human aggression and violence.

INTRODUCTION

The medical, social, economic, and societal problems associated with human aggressive and defensive behaviors often create an incorrect view of these as essentially maladaptive or pathological. A comparative approach to analysis of aggression and defense is necessary to understand that these represent crucial life-crisis management strategies for all mammals and most inframammalian species. Research on situations eliciting aggression and defense and the hormonal mechanisms modulating them has provided information about the evolutionary functions of these biobehavioral systems, while work on their neural systems and pharmacological control may provide links to some specific psychopathological conditions.

In humans and in other mammalian species, agonistic interactions between members of the same species (conspecific) almost always consist of some mixture of aggressive and defensive behaviors. Both combatants are likely to be aggressive, as a nonaggressive participant could usually preclude or conclude the encounter by leaving. However, fights also typically involve pain and sometimes injury: that nonhuman animals always settle conspecific disputes by nonpainful or noninjurious displays is a myth. Thus self-defense is also typical of one or both participants in real-world encounters. This situation creates difficulties in making clear differentiations between aggression and defense, and has substantially hindered analysis of these two different patterns, and of their evolutionary functions.

STUDIES OF AGGRESSION AND DEFENSE IN ANIMALS

Laboratory studies can polarize aggressive and defensive tendencies of two combatants, making one extremely aggressive and the other exclusively defensive, in order to make the two behavior patterns easier to analyze. While such studies are not adequate for determining the range of conditions that will elicit aggression or defense, they do provide information on some focal antecedents for each of them, and enable relatively uncontaminated descriptions of the actual behaviors involved in the two biobehavioral systems. When findings from such laboratory research (typically involving rodents) are combined with ethological studies on a much broader range of mammalian and inframammalian species in their own natural habitats, a consistent view of the adaptive value of both aggression and defense begins to emerge.

CAUSES OF AGGRESSION AND DEFENSE

Offensive aggression is coming to be conceptualized as a behavior pattern elicited in the context of resource disputes, particularly in response to challenge (generally but not always from a member of the same species) to the individual's control over these resources or rights. While aggression may be adaptive in reducing conspecific challenge over crucial elements such as food, water, access to nesting sites or materials, and the like, in an evolutionary context a particularly valuable resource is access to a breeding partner. Thus male on male aggression is especially common during the breeding season, or in the presence of reproductive females. Similarly, for females offspring are a particularly important evolutionary resource, and many instances of female aggression occur in a maternal context. The common tendency to refer to such situations as involving 'defense' of mates, offspring or other resources produces an additional, but purely semantic, problem in conceptualizing the difference between the biobehavioral systems for aggression and defense. The defense system is a response to threat to bodily integrity. Aggression is a response to challenge to the individual's control of a resource. It does not aim at 'defending' the integrity of the resource, but in promoting the aggressor's claim to that resource. Successful aggression is thus an important means of social control.

Most mammalian species have evolved mechanisms that reduce or limit intraspecies aggression by limiting the circumstances under which one animal is likely to challenge another for crucial resources. Territoriality and dominance relationships are two such mechanisms. The first operates through avoidance by one animal or group of the territory of another animal or group, while the second involves recognition by animals within a group of the relative ranks of other animals, such that subordinates typically do not challenge those higher in rank to themselves. However, territorial encroachment and rank challenges do occur, and for most species acquisition of territory and rise in rank depend on successful aggression by the protagonist and sometimes its allies.

In contrast to these rather complex causes of aggression, defensive behaviors occur in response to bodily threat, from predators or from conspecific individuals. Antipredator situations provide an opportunity to describe and analyze these behaviors in situations that do not contain elements of aggression: predation is not aggression. Defensive behaviors represent attempts to avoid, evade, escape, conceal, bluff, threaten or otherwise neutralize threat from dangerous stimuli, thereby enhancing the animal's likelihood of coping successfully with the threat.

FACTORS CONTROLLING THE SUCCESS OF AGGRESSION AND DEFENSE

Successful aggression is adaptive, but aggression itself can be dangerous. This puts a premium on the ability of animals including humans to evaluate the probability of success prior to entering into an aggressive encounter. The individual's own history of victory or defeat is a particularly important determinant of future success in aggression, and a strong predictor of aggressive behavior. Members of most mammalian (and many other) species also pay particular attention to features of a potential opponent that indicate size, strength, health, and fighting ability, before entering a fight. Thus hyperexpression of such features becomes adaptive in discouraging challengers, a situation that has resulted in the evolution of ever-larger bodies (particularly for males of polygynous species, a factor in sexual dimorphism); more impressive weapon systems such as antlers; and enhancement of other body features that may be the focus of comparison, such as the mouth of the hippopotamus. Particular postures and movements that optimally present these features, sometimes eliciting retreat from a challenger, constitute many of the aggressive displays that have wrongly been believed to replace actual agonistic behavior in nonhuman animals.

Age and fighting ability are also factors in the success of aggression, and these are often evaluated in fights among pubertal and young adult males. However, play fights in prepubertal mammals appear to bear little direct relationship to adult fighting between the same animals: prepubertal males who later become dominant are less likely to initiate play fights than are those who later become subordinates, and the contact sites and behavior patterns seen in the fights of prepubertal mammals appear to be more similar to adult sexual behavior than to adult agonistic behaviors.

Successful defensive behaviors must counter the attacks of predators in addition to conspecific threats, and must be effective in a variety of habitats and situations. Thus few higher animals rely on a single type of defensive behavior. Most mammals appear to have a relatively consistent group of about half a dozen focal forms of defensive

behavior, with some of these emphasized or rarely used in a particular species, but all represented to some degree. Given this range of defensive behavior, the major single factor in the success of defense is how well the particular defensive behavior fits the situation, i.e. the specific threat stimulus and features of the environment relevant to the success of a particular defense. Thus, although the success of both aggression and defense relies on utilization of information about both the opponent and the situation, the topography or form of defensive behavior is geared more to features of the environment (such as presence of an escape route, location of places of safety, and distance between predator and self), while the form of aggression largely reflects features and behaviors of the opponent.

AGGRESSION AND DEFENSE IN THE RAT

Conspecific aggressive and defensive behaviors have been described in greatest detail in laboratory rodents, particularly the rat. Conspecific attack in rats involves elements of approach and investigation, typically focused on olfactory assessment of the sex, age, and breeding condition of the opponent. The attack itself involves attempts to bite the back and flanks of the opponent, as well as a number of maneuvers that enable the attacker to reach this location. Conspecific defense in rats also relies heavily on the targeting of the attack towards the back of the opponent, in that if the defender can remove this site from its attacker, it is relatively safe from being bitten. One general method of removing a target site for attack is flight, for which the countering attack tactic is chasing. The defensive animal may also adopt an upright posture facing the attacker; this keeps its back – the attacker's target – out of reach. An experienced defender can also pivot smoothly to maintain frontal orientation if the attacker attempts to circle around to the defender's back, utilizing a lateral attack sequence. This consists of the attacker arching its back and moving laterally toward the upright defender, sometimes pushing it off balance or lunging in a forward circular motion toward the defender's back and flanks. In an even higher-level defense tactic the rat lies on its back, again concealing the target of conspecific attack. However, the attacker typically stands over the supine defender, sometimes pushing at it to induce movement and reveal the back target.

These defense tactics are adaptive because laboratory rats, as well as wild *Rattus norvegicus* and *R. rattus*, are reluctant to bite targets other than the back of a conspecific opponent. In particular they fail to bite an opponent's ventrum, even if this is the only body area exposed. While the strength of this prohibition on biting particular targets may vary somewhat from one species to another, some targeting of bites and blows appears to be common across mammalian species. The specific target of attack is typically one in which there is relatively little chance of lethal damage. Thus such bites or blows may produce pain, discouraging further resource or status challenge from the other, without killing a conspecific that may be a relative, or part of the attacker's social group. Some dedicated structures, such as antlers, provide weapons that are used in offensive attack, and also serve as the target of this attack. Typically such structures are not used in defense against predators.

Antipredator defensive behaviors are much less dependent on defense of particular body structures, since the attacks of predators are typically aimed at vulnerable body sites rather than away from them. Thus while defensive tactics like flight remain adaptive, some defenses that are effective against conspecifics (such as lying on the back) would not be useful against a predator. For most mammals, flight is a dominant antipredator defense when an escape route is available or a place of safety within reach. Immobility, often involving specific freezing postures, is also a common defensive behavior. This may be adaptive by helping the prey animal to avoid detection or because some predators selectively attack fleeing prey. In close encounters, defensive threat followed by defensive attack may be highly effective in discouraging the attacker: typically, an injured predator is at a severe disadvantage in future hunts. When there is a potential rather than a clear and obvious threat (e.g. novel stimuli or predator odors) defenses such as flight and defensive threat or attack may be useless or even counterproductive. To such stimuli the most prominent defensive behavior is risk assessment, a highly motivated information-gathering pattern involving orientation to the potential threat source, sensory (visual, auditory, olfactory) scanning, and approach with a low-back posture that minimizes detection of the defensive animal while allowing it to investigate the potential threat source.

NEURAL MECHANISMS OF DEFENSIVE BEHAVIOR

Neural systems underlying defensive behaviors are the focus of intense research interest. They have been investigated using a variety of research

paradigms, including evaluation of behavioral effects of lesions or stimulation of particular brain sites, and analysis of regional intermediate early gene expression during confrontation with a threat source. These studies, in conjunction with tract tracing from cells in sites thus implicated, has yielded a detailed, and relatively consistent, view of the nervous control of particular defensive behaviors.

Studies of neuronal activation in association with aggression or defense have often used regional expression of c-*fos* messenger ribonucleic acid (mRNA) as a marker of activity in specific brain sites. A number of such studies, using a variety of different species (rat, mouse, hamster) and both conspecific defeat and predator exposure paradigms, have provided relatively consistent findings. Defensiveness is associated with activity in several limbic areas (e.g. amygdala, lateral septum), and in a ventral zone from the preoptic area through the hypothalamus to the midbrain periaqueductal gray (PAG) matter. Fos expression in both the septum and the central nucleus of the amygdala (CeA) following confrontation with a threat source (conspecific defeat) appears to vary from the initial to later exposures. Fos expression provides a good, though not necessarily comprehensive, overview of brain areas that respond to particular experiences or activities. However, Fos studies are somewhat difficult to interpret as they do not differentiate between areas that are directly involved in a particular function and those that respond to that function or to nonspecific events such as arousal, motor activity, or autonomic changes that may be associated with it. In addition, Fos does not indicate the organization of the systems that may be involved. Stimulation, lesion, and tract-tracing studies add significantly to systems research on defense, and such studies have shown that several of the areas in which Fos expression is seen during defense are indeed important in the elicitation and maintenance of defensiveness.

Stimulation of the PAG can elicit flight, freezing, and defensive threat responses in both rats and cats. These different responses are organized in terms of longitudinally coursing columns within the PAG. At more rostral levels, dorsal stimulation tends to elicit defensive vocalizations, while at more caudal levels such stimulation elicits flight and sometimes freezing. Stimulation of the ventrolateral column, in its caudal aspects, elicits an immobility response that has been variously interpreted as pain-related quiescence or as freezing. However, as lesions of this area sharply reduce kyphosis – an upright, crouched, posture shown by rat mothers during nursing – it is also possible that this area of the PAG is broadly involved with the maintenance of all active immobile reactions rather than with defense-related immobility only, or with specific types of defense-related immobility.

The dorsal premamillary nucleus of the hypothalamus has also received particular attention in a defense context. Lesions in this area appear to virtually abolish flight and freezing, and retrograde tracing studies indicate that it receives direct afferent connections from many of the telencephalic and diencephalic structures that show Fos expression after exposure to a predator. Other well-developed research efforts have been directed toward analysis of neural systems involved in specific defense-related behaviors. Siegel and his colleagues used combinations of lesions, stimulation, and local and systemic drug administration to characterize the control of defensive threat vocalizations (hissing) in the cat. They found that a number of sites in both limbic areas and in the ventral zone from preoptic area to PAG act to initiate and modulate this component of defense. Several such neurotransmitter-specific modulatory systems have been described, some involving direct connections from several areas of the amygdala to the PAG, while others consist of direct and indirect hypothalamic–PAG connections.

Similarly, Davis and his colleagues have concentrated on systems involved in potentiation of the startle reflex. Although the basic startle reflex is a simple three-synapse brain stem–spinal cord circuit, the amplitude of startle can be increased by conditioned cues (stimuli previously associated with shock) or unconditioned defense-enhancing manipulations such as testing in a brightly lit chamber. A particularly interesting finding is that the systems involved in conditioned and unconditioned potentiation of startle may be quite different: the CeA is strongly involved in conditioned potentiation, but is not crucial for unconditioned potentiation of startle, while damage to the bed nucleus of the stria terminalis abolishes unconditioned but not conditioned potentiation. These findings are consonant with the central role of the CeA in conditioning of freezing to shock cues, and with findings that CeA Fos expression changes with repeated threat exposure – all of which suggest that this nucleus may be more involved in plastic processes (learning, habituation) than in defense *per se*.

NEURAL MECHANISMS OF AGGRESSIVE BEHAVIOR

Research into the physiological mechanisms of aggression has tended to look at the pharmacological control of aggression rather than the specific neural systems involved in this behavior. Analysis of aggressive behaviors is additionally complicated by the fact that attack behaviors may involve offensive aggression, defensive attack, or predation; these may be difficult to differentiate under standard laboratory conditions. However, research programs have focused on sites in the hypothalamic area, with some differences in specific sites for different species. For example, in the hamster, vasopressin manipulations in the nucleus circularis, an area rich in vasopressin-expressing neurons, may profoundly affect aggressive behavior. However, in the rat, the nucleus circularis does not contain vasopressin-expressing neurons, suggesting that aggression changes associated with manipulations in this area must involve a different mechanism of action. In a number of hypothalamic sites, including the nucleus circularis (hamster), the intermediate hypothalamic area and part of the ventromedial nucleus (rat), electrical stimulation elicits an attack response, while lesions may reduce offensive attack. Neuronal tracing studies indicate connections to a number of areas similar to those implicated in defense, such as the prefrontal cortex, amygdala, septum and the PAG. Because stimulation and lesioning of some of these structures may impact both aggressive and defensive behaviors, it is not clear that the changes in one set of behaviors are independent of effects on the other. Similarly, because of the reliance of both aggressive behavior and sexual and maternal behaviors on pheromones, areas such as the medial nucleus of the amygdala that are strongly involved in the accessory olfactory system may influence both types of behavior. However, aggressive, defensive and reproductive behaviors appear to be well differentiated at the level of the hypothalamus. In addition, there are a number of hypothalamic sites in which stimulation elicits grooming, suggesting the existence of a number of roughly parallel systems in this area that are strongly involved in species-typical behaviors of great evolutionary importance.

Serotonin (5-hydroxytryptamine, 5-HT) is perhaps the most frequently investigated neurotransmitter in terms of effects on aggression. Deficiencies in serotonin have been implicated as potentially important in impulsivity and aggression in species ranging from lower mammals to humans. In the latter species, low levels of a serotonin metabolite in cerebrospinal fluid may signal reduced serotonin activity in the brain, which, interacting with other genetic and environmental factors, may be associated with instances of impulsive violence, or other impulsive actions. Two specific serotonin receptor subtypes, 5-HT_{1A} and 5-HT_{1B}, are of particular interest as the principal targets of a class of 'serenic' drugs that have proved to be extremely effective in reducing offensive aggression in a variety of animal models.

Vasopressin activity in the hypothalamus and related structures is associated with aggression enhancement, and serotonin may act at the level of the vasopressin receptor to modulate this behavior. In hamsters, pretreatment with fluoxetine, a serotonin reuptake inhibitor, increases serotonin levels in the anterior hypothalamus (an area known to influence aggressive behavior) and inhibits aggression. Systemic fluoxetine also reduces the increase in aggression seen when vasopressin is injected into the ventrolateral hypothalamus. This area contains both the 5-HT_{1A} and 5-HT_{1B} receptor subtypes, which have been particularly implicated in the control of aggression in animal models.

HORMONAL MECHANISMS OF AGGRESSION AND DEFENSE

The evidence for a relationship between hormones and aggressive behavior predates science: castration of male animals to reduce their aggressiveness has been common throughout human history. Testosterone generally enhances aggression across a range of mammalian species, including humans, but the relationship is highly variable, and may be more specific to times or situations involving reproductive behaviors, or dominance and territoriality related to reproductive advantage. One mechanism by which testosterone, like serotonin, may influence aggression is through hypothalamic vasopressin receptors. In the ventrolateral hypothalamus of male hamsters, vasopressin receptor binding disappears after castration, but can be maintained by treatment with testosterone. Findings that microinjections of vasopressin in this area enhance aggression in untreated male hamsters but fail to do so in castrates suggest that testosterone maintenance is important for the functioning of vasopressin receptors in this area.

Attempts to understand the mechanisms of the relationship between testosterone and aggression have emphasized that both testosterone and its metabolites may influence relevant neural systems and that the direction of effect may be different for some of these. In the brain, testosterone may

be aromatized to estradiol, or metabolized by 5α-reductase to dihydrotestosterone. The functional pathways underlying male aggression may be modulated by either of these metabolites, being either estrogen-sensitive or androgen-sensitive, at various points. A finding of particular interest, in view of the centrality of 5-HT_{1A} and 5-HT_{1B} receptors in reducing aggression, is that in male mice treated with androgens, stimulation of either 5-HT_{1A} or 5-HT_{1B} receptors or combinations thereof reduces aggression. However, in estrogen-treated males, only combined, high-dose, 5-HT_{1A} plus 5-HT_{1B} receptor activation was effective. In animals given testosterone, which metabolizes into both androgens and estrogens, aggression fails to decrease with low doses of 5-HT_{1A} or 5-HT_{1B} alone or in combination, suggesting that estrogens may protect the neural systems involved in aggression from suppression by serotonin receptor activation. These relationships may also reflect action at particular sites along the brain pathways serving aggression, in that estrogen or androgen treatments differentially modulate effects of microinjections of 5-HT_{1A} or 5-HT_{1B} agonists into the lateral septum, or the medial preoptic nucleus of the hypothalamus. Such findings suggest the possibility of complex modulation of aggressive behavior, or of aggressive behavior as elicited in different sorts of situations, by reproduction-relevant steroids. These interactions are also interesting in the context of widespread environmental or food contamination by estrogenic chemicals.

Anabolic steroids are not uncommonly used illegally by athletes to enhance the development of muscle mass. Reports of enhanced aggressiveness or irritability accompanying such use have spurred experimental studies in animals. High-dose anabolic steroid treatment during early adolescence increases the aggressive response of male hamsters towards intruders, and these increases may be partly ameliorated by treatment with vasopressin receptor antagonists in the anterior hypothalamus. Treated hamsters showed enhanced vasopressin fiber density and peptide content in this area.

Gonadal and adrenal steroid hormones may interact in complex ways in association with aggression and defense. Both corticosterone and testosterone levels respond to acute and chronic defeat, with the former increasing and the latter decreasing. In fact, high circulating levels of free corticosterone may contribute to reductions in testosterone. However, corticosterone has recently been reported to enhance offensive aggression in male rats, while adrenalectomy produces a more rapid but 'deviant' attack aimed at the head of the opponent. The specificity of this enhancement is not yet entirely clear, as other behavioral consequences of corticosterone are poorly understood, but reports that blocking the mineralocorticoid receptor by spironolactone dramatically reduced territorial aggression are consistent with a view that adrenal hormones may have an important permissive role in aggression.

AGGRESSION AND DEFENSE IN HUMANS

Studies of aggression in humans have largely consisted of criminological data and articles on laboratory elicitation of aggression-like behaviors in response to some type of insult, irritation or other provocation. In general there has been little attempt to differentiate between offensive and defensive aggression, although a case has been made for anger or moral outrage as the human equivalent of offense, elicited by challenge to important rights or resources of the individual. Conversely, it has been suggested that human aggression represents defensive aggression. Because of the complexity of human verbal representations (to self and others) of the rationale or motivation for behavior (particularly behavior that is considered socially undesirable), the task of determining whether a particular instance of aggression is primarily offensive or defensive may be extremely difficult. However, this distinction is important in the treatment of some individual acts of violence by virtually every legal system yet devised, and enhanced attention to the differences in behavioral, as well as circumstantial, aspects of violent acts may yet shed light on these differences. As noted above, human aggression findings such as those involving low levels of serotonin metabolites, or effects of anabolic steroids, serve as important spurs to experimental research in animals. However, research attempting to link cognitive factors in human aggression to the circumstances that elicit aggression in lower animals has been sadly neglected. Similarly, normal human defensive behavior has received little attention. However, the potential involvement of particular defensive behaviors in specific psychopathological disorders, notably various types of anxiety disorders, is receiving increasing consideration. This is due in part to agreement between the effects of drugs on particular defensive behaviors and their efficacy against relevant anxiety disorders, and also reflects specific similarities of behavior between major symptoms of the anxiety disorder and those of the parallel defensive behavior.

Further Reading

Blanchard DC and Blanchard RJ (1984) Affect and aggression: an animal model applied to human behavior. In: Blanchard RJ and Blanchard DC (eds) *Advances in the Study of Aggression*, vol. 1, pp. 2–62. New York, NY: Academic Press.

Blanchard DC, Griebel G, Rodgers RJ and Blanchard RJ (1998) Benzodiazepine and serotonergic modulation of antipredator and conspecific defense. *Neuroscience and Biobehavioral Reviews* **22**(5): 597–612.

Canteras NS, Chiavegatto S, Valle LE and Swanson LW (1997) Severe reduction of rat defensive behavior to a predator by discrete hypothalamic chemical lesions. *Brain Research Bulletin* **44**(3): 297–305.

Delville Y, De Vries GJ and Ferris CF (2000) Neural connections of the anterior hypothalamus and agonistic behavior in golden hamsters. *Brain, Behavior and Evolution* **55**(2): 53–76.

Depaulis A, Keay KA and Bandler R (1992) Longitudinal neuronal organization of defensive reactions in the midbrain periaqueductal gray region of the rat. *Experimental Brain Research* **90**(2): 307–318.

Haller J, Halasz J, Mikics E, Kruk MR and Makara GB (2000) Ultradian corticosterone rhythm and the propensity to behave aggressively in male rats. *Journal of Neuroendocrinology* **12**(10): 937–940.

Kollack-Walker S, Don C, Watson SJ and Akil H (1999) Differential expression of c-fos mRNA within neurocircuits of male hamsters exposed to acute or chronic defeat. *Journal of Neuroendocrinology* **11**(7): 547–559.

Pope HG, Kouri EM and Hudson JI (2000) Effects of supraphysiologic doses of testosterone on mood and aggression in normal men: a randomized controlled trial. *Archives of General Psychiatry* **57**(2): 133–140.

Roeling TAP, Veening JG, Kruk MR *et al*. (1994) Efferent connections of the hypothalamic 'aggression area' in the rat. *Neuroscience* **59**(4): 1001–1024.

Siegel A, Roeling TA, Gregg TR and Kruk MR (1999) Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience and Biobehavioral Reviews* **23**(3): 359–389.

Simon NG, Cologer-Clifford A, Lu SF, McKenna SE and Hu S (1998) Testosterone and its metabolites modulate 5HT1A and 5HT1B agonist effects on intermale aggression. *Neuroscience and Biobehavioral Reviews* **23**(2): 325–336.

Aging, Neural Changes in

Introductory article

Elizabeth A Kensinger, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
Suzanne Corkin, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

Dementia is not an obligatory consequence of aging. Normal aging does, however, result in changes in cognition, caused by a combination of neurotransmitter abnormalities and alterations in the structure and function of the brain.

INTRODUCTION

The increasing proportion of older adults in the population of many countries has heightened interest in the cognitive and neural changes that accompany normal aging. The following sections elucidate the effect of aging on a range of cognitive capacities, and the neural changes that may explain the pattern of spared and impaired function. Particular attention is devoted to short-term and working memory (allowing the temporary storage and manipulation of information); declarative memory (requiring conscious awareness); and nondeclarative memory (formed without conscious awareness). The evidence presented here comes from behavioral studies in healthy volunteers (cognitive psychology), patients with brain lesions (neuropsychology), analyses of brain structure (volumetric magnetic resonance imaging) and observations of focal task-related changes in normal brain activation during performance of specific cognitive operations (functional neuroimaging using positron emission tomography or magnetic resonance imaging).

DECLARATIVE MEMORY

Long-term memory is broadly divided into two components: declarative and nondeclarative memory. Declarative (explicit) memory is formed with conscious awareness, and requires the participation of medial temporal lobe structures, including the hippocampus. Declarative memory is what we use to help us remember the items we need to pick up at the grocery store, or the name of a friend whom we haven't seen in years. In general, declarative memory is more affected by normal aging than is nondeclarative memory. As an example of declarative memory, read the following words, and try to remember them. After reading through the list, write down as many words as you can remember, without looking back at the list: table, orange, calendar, computer, paper, needle, napkin, chair, sneeze, movie, sleep, castle, build, lunch, flower, dragon, plant, cushion, dolphin, muscle.

SENSORY MEMORY (PERCEPTION)

The sensory systems are those through which we receive direct input from the world via receptors: touch, taste, smell, vision, and hearing. Our perception of that input, however, can be influenced by a number of factors, such as our internal state, or the context in which the sensation occurred. Imagine you hear a loud banging sound while you are walking by a construction site. Now imagine that you hear the identical sound while walking alone on a deserted street late at night. Although the sound (sensory information) is identical in the two instances, the interpretation of the sound (perception) may differ greatly because it occurs in different contexts.

Some sensory systems are degraded as part of the normal aging process. Most commonly, older adults experience hearing loss. By the age of 80 years the majority of older adults have significant hearing loss. They also have visual deficits, including poorer color and luminance contrast, and many

have a loss of central vision due to macular degeneration. Sensory and perceptual deficits can hinder adults' performance on many tasks. For example, Murphy and colleagues found that older adults are more affected by background noise when trying to remember word pairs than are young adults. These researchers proposed that part of this increased effect is due to degraded sensory representations, though attentional reductions probably also contribute. Older adults also may have more difficulty discriminating isoluminant colors such as blue and green, and are slower and less accurate on tasks that require color discrimination, such as color-naming tests, the Stroop Test, and the Wisconsin Card Sorting Test.

Nevertheless, with modifications to testing procedures (such as louder stimuli) most older adults are successfully able to perceive information. Perceptual priming, requiring visual processing, is spared with normal aging. Older adults are also able to repeat word lists or digit strings (presented either aurally or visually), suggesting that their sensory deficits are not profound enough to affect immediate memory. It is, nonetheless, important to control for perceptual confounds when interpreting the performance of older adults, and to match (equate) the perceptual capabilities of young and older adults (either by matching individuals or, more feasibly, matching the stimuli so that young and older adults perceive them equally). Without taking these measures, it is unclear whether impaired performance in older adults stems from a purely cognitive deficit or from impaired perception. Particularly in memory studies, reduced perception may result in older adults having a degraded memory representation, subject to faster disruption over time.

SHORT-TERM MEMORY AND WORKING MEMORY

Short-term memory is a limited-capacity storage buffer for information to be remembered over a very short duration (a few seconds). The term 'short-term memory' is commonly used to mean recent memory, but that definition is not used by cognitive psychologists. Short-term memory consists of two components: a passive information store and an active rehearsal system. Working memory, in contrast, not only stores information but also updates and manipulates that information.

Read the following words, and try to keep them in mind for 10 s: hill, milk, goat, tool, foot, pie. This type of storage requires short-term memory. To succeed in repeating the words 10 s later, you might also have felt that you were 'rehearsing' those words (e.g. internally vocalizing them) to allow yourself to remember them. This phenomenon highlights the active rehearsal component of short-term memory. Now, read the words again, look away, and this time try to say them in alphabetical order. Simply rehearsing the words is insufficient to complete this task; rather, you also need to manipulate the words to place them in the proper sequence. This task, therefore, requires working memory.

The most widely accepted model of working memory, proposed by Baddeley and Hitch, defines working memory as consisting of three components: the central executive, the phonological or articulatory loop, and the visuospatial sketchpad. The central executive controls the allocation of attention, as well as the coordination and monitoring of activities, while the phonological loop and visuospatial sketchpad are slave systems of the central executive that temporarily maintain and manipulate verbal and nonverbal material, respectively.

Short-term memory is usually spared with aging, whereas working memory shows age-related decrements. This decline probably does not occur equally across all components of working memory, but rather targets only a subset of processes. Three components probably account for the majority of age-related working memory decline: processing speed, storage capacity, and inhibitory ability.

Processing Speed

Older adults are known to have a slowed speed of processing. Salthouse and colleagues proposed that decreased processing speed could account for some of the age-related declines in cognition. They suggested that cognitive performance suffers because (a) the slowed mental operations cannot be carried out within the necessary time frame, and (b) the increased time between mental operations makes it more difficult to access previously processed information. Processing speed can affect encoding because the quality and availability of perceptual information degrades over time, so information that is processed quickly will be encoded more effectively and, therefore, will have a more durable representation or memory trace.

The hypothesis of a relation between processing speed changes and cognitive decline has been confirmed in a number of studies. Longitudinal studies have shown that changes in speed of processing

may predict longitudinal cognitive decline, and a number of researchers have found that controlling for speed eliminates age effects on various memory tasks.

Storage Capacity

Storage capacity is one component of short-term and working memory: the passive storage buffer that dictates how much information can be stored without rehearsal being used to 'refresh' that information. A reduction in storage capacity is likely to contribute to age-related working memory decline: older adults may be able to hold less information in mind. Reduced storage capacity could provide an alternate explanation to reduced processing speed. Thus, remembering what information was processed, or carrying out mental operations, would be restricted not by time pressure but by reduced storage capacity. Although storage capacity declines with age, it is not clear that this deficit is sufficient to explain the cognitive decrements in working memory that occur with aging.

Inhibitory Ability

Hasher and Zacks proposed the inhibitory deficit theory to account for changes in cognitive performance with age. 'Inhibition', in this theory, is the ability to ignore irrelevant information while focusing attention on pertinent information. The inability to filter out irrelevancies causes older participants' working memory to be filled with unneeded information, leaving less space for task-relevant memories. This explanation, therefore, is not completely dissociable from a storage capacity explanation for cognitive aging.

Researchers have found evidence for inhibitory deficits in older adults on a variety of tasks. Commonly used paradigms for assessing inhibition are task-switching or set-shifting. On these tasks, participants must first remember one set of rules or pay attention to one salient characteristic, and then must switch rules or attend to a different characteristic. Most investigators have found that these tasks are sensitive to aging effects, with older adults being less able to ignore the previously relevant information.

LONG-TERM MEMORY

Long-term memory can be divided into two categories: episodic and semantic. Episodic memory entails retrieving information from a particular episode, localized in space and time (e.g. remembering seeing the Eiffel Tower on your first trip to Paris), while semantic memory requires retrieving factual information independent of any specific episode (e.g. knowing that the Eiffel Tower is in Paris). Recall of the word list given at the beginning of this article required episodic memory. You had to bring the word to mind, and also correctly remember that the word was on the list you had just read. Accessing the meaning of the words, however, required semantic memory.

Episodic Memory

Episodic memory appears to be more affected by normal aging than other memory processes. All aspects of episodic memory are not affected uniformly, however.

Factual and source memory

Episodic memory can be subdivided into two components: factual memory and contextual or source memory. Normal aging results in a disproportionate impairment in source memory as compared with fact memory. Even when older adults remember a fact or event, they have more difficulty than younger adults pinpointing the specific contextual details, such as where and when they learned a fact. For example, Spencer and Raz tested young and older adults on a test requiring them to remember facts, some true and some fictitious (e.g. 'Angela Lansbury regularly consults with astrologists'). After a delay, participants were asked to complete the fact ('Angela Lansbury regularly consults with …'), and to say where they had learned the fact (experiment or elsewhere) and whether the fact had been presented on a blue or pink card. Older adults were disproportionately impaired on the source recall than on the fact recall.

Source memory is believed to rely on the brain's frontal lobes. Measures of frontal lobe function correlate with measures of source memory, and reductions in source memory have been shown to occur in amnesic patients with frontal lobe lesions. The frontal lobes are also critical for linking events together in time. Aging results in frontal lobe dysfunction, probably connected to the source memory deficits seen in older adults.

Recall and recognition

Older adults show poorer performance on recall tests ('What words were on the word list?') where no cue is provided, than on recognition tests ('Was "cloud" or "table" on the word list?') where retrieval cues are provided. In general, older adults show improved performance on episodic memory

tests when cues are provided during encoding or retrieval phases.

The source memory decrement, and the benefit provided to older adults with cues, are probably related to the robustness of the memory trace encoded by the older adults. Aging seems to affect the quality of the representation, such that general gist-based information is more easily encoded and retrieved than richer, item-specific information that includes not only the to-be-remembered information, but also the context in which it was learned. This hypothesis is supported by the finding that on recognition tasks, older adults are more likely than young adults to say that an item is 'familiar' (they feel they have encountered the item before), but less likely to say that they 'recollect' the item (remember something specific about the item's presentation).

Semantic Memory

One of the most readily reported complaints by older adults is their declining ability to recall the names of people and objects. Word finding difficulties are among the most severe deficits in normal aging. Naming deficits result in slower speed of picture naming, a greater number of speech disfluencies, and an increased number of tip-of-the-tongue effects.

Picture naming

Older adults' naming deficit is particularly pronounced for proper names, though studies have also reported longer naming times for nonproper objects. The difficulty may be related in part to deficits in associative memory: the ability to form associations between a name and an object may be reduced in normal aging.

Tip-of-the-tongue effect

The tip-of-the-tongue effect occurs when a person has access to a word's meaning, but is unable to produce the phonological code. Older people report more tip-of-the-tongue experiences with everyday objects and with proper names than younger people. In addition, the accuracy of available information during a tip-of- the-tongue state is higher for young than old adults. For example, younger participants are more likely to state correctly the first letter of the word they are trying to remember than older participants. As with naming deficits, tip-of-the-tongue effects are more pronounced for proper names than for everyday objects. Better performance with everyday objects may be related to what Burke and colleagues refer to as 'summation of priming'. With everyday objects, connections from a variety of semantic associates converge on the correct name; but with proper names, older adults are handicapped without this type of summation.

NONDECLARATIVE MEMORY

Nondeclarative (implicit) memory is encoded and strengthened, across trials, without conscious awareness. It encompasses a heterogeneous group of processes and kinds of performance, including skill (motor) learning, repetition priming, and classical conditioning. These domains rely on distinct and separable neural substrates. Because of the task diversity, and range of necessary neural substrates, it is perhaps logical that nondeclarative memory is not uniformly impaired with aging.

Skill (Motor) Learning

In the 1960s, Milner demonstrated that the amnesic patient HM, while unable to form new declarative memories, could successfully learn a new motor skill. She asked HM to perform a mirror tracing task, in which he had to trace the outline of a star seen only in mirror-reversed view. Over 3 days of practice, his error scores decreased dramatically, and he maintained the learning from one day to the next, but he had no conscious recollection that he had done the task before. Corkin and colleagues administered additional skill-learning tasks to HM, confirming that he generally showed preserved learning. Other investigators have also reported that amnesic patients can learn and retain motor skill learning without awareness of prior exposure to the task.

These results indicate that the brain structures that support conscious, declarative memory and which are damaged in amnesia (the hippocampus and other medial temporal lobe structures) are not critical for skill learning. Skill learning is thought to rely on the motor cortex, supplementary motor area, cerebellum, basal ganglia, and posterior parietal cortex. Older adults have reductions in the amount of dopamine and acetylcholine in the basal ganglia; they also show cerebellar dysfunction. These changes may result in slower acquisition of some motor learning tasks.

No consensus exists as to how skill learning is affected by aging. Researchers have found every possible outcome: equal performance in young

and older adults, better performance in older adults, and poorer performance in older adults.

Repetition Priming

Priming is broadly defined as a faster or biased response to a stimulus based on prior exposure to that stimulus, or a related stimulus. As an example of priming, try to complete these word stems with the first word that comes to mind: nap—, dol—, cas—, cus—, tab—, dra—. You may have responded with words from the list given at the beginning of this article, without being consciously aware that you had done so. This effect, based on prior exposure to a stimulus, is an example of repetition priming.

Priming is not a unitary construct; rather, multiple processes contribute to priming effects. For discussion purposes, we will divide priming into two categories: perceptual priming and conceptual priming. These types of priming are dissociable and rely on separate neural substrates.

Perceptual Priming

Perceptual priming is based on the sensory characteristics of a stimulus. For example, if participants are shown the pseudoword 'pabhan', they will later be more likely to recognize that pseudoword when it is flashed briefly, than another pseudoword flashed at the same rate. Keane and colleagues proposed that perceptual priming effects are mediated by a structural–perceptual memory system localized to the occipital lobe; this hypothesis has been supported by neuropsychological and functional imaging studies.

Older participants frequently perform as well as younger adults on perceptual priming tasks. For example, Schacter and colleagues presented young and older adults with black-and-white drawings of three-dimensional objects in either structurally possible or impossible configurations. When participants had to judge whether the briefly presented stimuli were possible or impossible objects, young and older adults showed the same magnitude and pattern of priming, with robust priming for possible objects and no priming for impossible objects.

The finding of spared perceptual priming with aging is consistent with its reliance on the occipital and temporoparietal cortex because aging is thought to spare primary cortices and modality-specific association areas, including the occipital lobe.

Conceptual Priming

In contrast to perceptual priming, conceptual priming relies primarily on the semantic representation of the stimulus. For example, if participants are first presented with the category word 'fruit', they will be faster at determining that the word 'apple' is a real word than if they were first presented with the category word 'furniture'. Keane and colleagues proposed that conceptual priming is mediated by a lexical–semantic memory system recruiting temporoparietal regions. This hypothesis has been supported also by neuropsychological and neuroimaging studies.

Some studies have reported age-related deficits in priming experiments that are conceptual in nature, including lexical priming and priming for new word associations. Other researchers, however, have reported spared performance in older adults. The discrepancy may have stemmed from different task designs, or individual variation within the older populations.

CLASSICAL CONDITIONING

One of the most commonly used forms of classical conditioning is the eyeblink response. In delay conditioning, a neutral stimulus (a tone) is followed repeatedly by a biologically relevant stimulus (an air puff to the eye), and the two stimuli coterminate. The measure of learning is the subsequent ability of the tone, by itself, to elicit a biologically relevant response (an eyeblink), the conditioned response. The strength of the conditioned response increases gradually with repetition, making it possible to document the number of trials needed to learn to a particular criterion. Older rabbits and older humans require significantly more trials than younger ones to acquire the association between the tone and the air puff, but considerable variability exists among older individuals. Results from neuroimaging and neuropsychology converge on the conclusion that the cerebellum is the critical neural substrate for delay conditioning. Because the cerebellum is affected by normal aging, the reduction in classical conditioning with normal aging is believed to result from less efficient cerebellar communication and output.

Trace conditioning differs from delay conditioning in that there is an unfilled interval between the offset of the neutral stimulus (the tone) and the onset of the biologically relevant stimulus (the air puff). The participant must therefore build up a representation, across trials, as to the relation between the tone and air puff. In addition to

cerebellar recruitment, the hippocampus is critical for trace conditioning. The hippocampal contribution is likely to stem from the fact that delay conditioning is not purely a nondeclarative memory task; conscious awareness of the relation is mandatory for successful conditioning.

On the trace conditioning paradigm, young and middle-aged adults condition at a similar rate, but older animals and humans are impaired. These deficits may occur at an earlier age than deficits in delay conditioning, and may be more pronounced.

AGE OF COGNITIVE DECLINE

Methods of Assessment

The age of cognitive decline can be assessed using one of two designs: cross-sectional or longitudinal.

Cross-sectional studies use data collected from individuals considered to be representative of an entire population, and interpret differences among those individuals as indicative of differences across two or more populations. For example, a cross-sectional study of aging might examine the performance of adults in their twenties, fifties and eighties. If the 80-year-olds performed more poorly than the other groups, this difference would be attributed to age. This design requires that groups be equated (matched) on as many variables as possible (e.g. overall intelligence and perceptual ability, as well as lifestyle, psychological and medical factors) to assure that group differences are due to age and not to other differences.

A longitudinal study avoids many of these confounds by tracking the same group of individuals across time, and comparing their performance at different time points. For example, a group of adults might be tested every 5 years for 20 years. Because each individual serves as his or her own baseline, the investigator does not have to worry about confounds such as intelligence or education level. Changes in overall health or perceptual ability over time must still be considered, and longitudinal studies can also be confounded by nonrandom drop-out rates (e.g. in a memory study, individuals who believe their memory is failing might be more likely to drop out of the study than those who believe their memory is good).

Cognitive Performance

The worsening performance across an extensive age range has led many researchers to divide the older adult population into 'young-old' and 'old-old' subgroups. This dichotomy was first proposed by Neugarten, who noted that these groups were dissociable not only by chronological age, but also by lifestyle changes. The young-old have fewer health limitations than the old-old, and the old-old are more likely to be widows or widowers than the young-old.

Researchers have used this dissociation to examine the progression of cognitive changes into the later decades of life. Most studies have confirmed that memory loss does not reach a plateau in the sixth or seventh decades; rather, memory decline continues throughout the later decades. Adults over the age of 70 years perform significantly worse on a range of recognition tasks compared with individuals in their seventh decade of life. Deficits in semantic memory and conditioning can also become more pronounced in the old-old.

The age at onset of decline differs depending on the type of function assessed. Semantic memory, as measured by tip-of-the tongue effects, has been found to be altered between the fifth and sixth decades. Woodruff-Pak and colleagues, however, found that eyeblink classical conditioning decrements began almost a decade earlier, with 40-year-olds showing significant impairments. Episodic memory, in contrast, remains relatively stable until around the seventh decade.

ANATOMICAL CHANGES

Longitudinal studies have found decreases in overall grey and white matter volumes with age, as well as increases in volumes of ventricular cerebrospinal fluid. The changes are not uniform across all brain regions, however. For example, the prefrontal cortex and medial temporal areas are more affected than primary association cortices. The pattern of neural changes helps to clarify why some types of cognition are particularly affected by normal aging, while other cognitive processes are relatively spared.

Hippocampus and Other Medial Temporal Lobe Structures

Hippocampal function is impaired by normal aging. Functional neuroimaging studies have shown that the hippocampus and other medial temporal lobe structures are less activated by memory tests with aging, and these functional changes often correlate with memory performance. A quantitative imaging study assessing the volume of different brain regions also found that hippocampal volume is significantly correlated with

performance on delayed recall tests. In fact, out of a variety of brain regions measured (including overall brain volume), hippocampal volume was the best predictor of delayed memory performance.

It is unclear whether there is substantial cell loss in this region, or whether the hippocampal dysfunction is related to neuropathological changes and cellular dysfunction affecting neuronal communication. On postmortem examination, adults over the age of 55 years typically show at least some entorhinal neurons that contain tangles, or where tangles are beginning to form. Brain neurochemistry also appears to be altered, with reductions in synaptic signaling. For example, long-term potentiation, thought to be a critical neural mechanism for learning and memory, is reduced with normal aging. Reductions in the number of NMDA (*N*-methyl-D-aspartate) receptors in the hippocampus, or reductions in the efficiency of the receptor, may mediate some of the age-related hippocampal dysfunction. Glucocorticoids, too, mediate hippocampal function, and increases in glucocorticoid levels may contribute to dysfunction.

Even studies that have found cell loss do not agree on which medial temporal lobe regions are most affected. While a number of studies found evidence for cell loss in the CA1 region of the hippocampus, not all studies have replicated this finding, using unbiased stereologic counting methods.

Cerebellum

Studies of humans, rats and rabbits suggest that the cerebellum, in particular the Purkinje (output) cells, is affected by aging. Older animals have fewer Purkinje cells, and those that remain have a reduced efficiency. Because Purkinje cells are the major output system of the cerebellum, damage to these cells results in dramatically reduced cerebellar output. Evidence for structural changes in humans comes from a magnetic resonance imaging (MRI) study showing significant negative correlations between age and grey matter volume in the cerebellar vermis and hemispheres.

Prefrontal Cortex

The function of the prefrontal cortex is affected by aging. Older adults perform more poorly than younger adults on tasks that measure frontal lobe capacities, including the Wisconsin Card Sorting Test and the Stroop Test. Neuroimaging studies have indicated changes in prefrontal activation, particularly in dorsolateral prefrontal cortex. Even on tasks where young and older adults perform at similar levels, prefrontal regions in older individuals show different patterns of activation, including recruitment of additional areas, and reduced activation in other regions relative to young adults.

As with the hippocampal region, it is unclear what neuropathological changes account for deficits in frontal lobe capacities. Researchers have proposed that neuronal shrinkage, or reductions in the number of presynaptic terminals, may be responsible for some of the age-related impairment. Axonal abnormalities may also underlie age-related deficits. In a volumetric MRI study, Double and colleagues found frontal lobe white-matter atrophy, suggestive of reductions in axonal processes. They suggested that slowed cognitive processing may occur because of a decrease in the speed of nerve conduction due to such axonal changes. These alterations may account for the working memory deficits with normal aging.

Neurotransmitter and Neuromodulator Abnormalities

A neurotransmitter is a chemical messenger that is released by one neuron, travels across a space between two neurons (a synapse), and binds to the second neuron. In this way, information is passed between neurons. A neuromodulator is a chemical that is not itself a transmitter, but affects the release of neurotransmitters.

Dopamine

Age-related changes in the dopaminergic system are well documented in humans, monkeys, and rodents. Levels of dopamine and tyrosine hydroxylase (an enzyme important for the production of dopamine) decrease with normal aging, and these reductions are particularly pronounced in the frontal lobes and basal ganglia. Postsynaptic alterations are also reported to occur with aging, including reductions in D2 dopamine receptors and some increases in D1 receptors.

Age-related reductions in dopamine levels may contribute to age-related working memory impairments. Dopamine depletion in the frontal lobes impairs performance on working memory tasks and dopamine may be particularly important for inhibitory ability. Prefrontal cortex must sort out task-relevant information, and maintain that information in the face of other distractors. Dopaminergic systems may provide the basis

for that allocation of attention. Dopamine may potentiate synapses associated with a reward (e.g. correct recall), thereby intensifying links between task-relevant computations, and weakening others.

Acetylcholine

Considerable evidence links acetylcholine to learning and memory. Acetylcholine is released when animals perform spatial memory tasks, and injection of cholinergic antagonists such as hyoscine (scopolamine) impairs memory acquisition in humans and nonhuman primates.

In aged animals, memory loss is correlated with hypoactive cholinergic neurons. For example, older rats show reduced excitatory postsynaptic potential amplitudes resulting from stimulation of the CA1 region of the hippocampus, suggesting that cholinergic neurons are less responsive in older animals. Older animals also show reductions in cholinergic receptor density that are particularly pronounced in the medial and caudal parts of the striatum, and in the frontal lobes.

Adrenal glucocorticoids

Stress hormones, too, are linked to the neural loss and dysfunction associated with normal aging. The adrenal cortex (in the adrenal glands, located near the kidneys) secretes glucocorticoids, which underlie our physical responses to threatening stimuli. In the short term, glucocorticoids are essential to our survival because under stressful conditions they increase the availability of energy substrates (blood glucose). Prolonged exposure to elevated glucocorticoid levels, however, can be detrimental, suppressing anabolic processes and depleting existing energy stores. With age, the stress response is not terminated as efficiently, causing glucocorticoid levels to remain elevated for significantly longer following stress.

The hippocampus, one of the main target sites for glucocorticoids, seems to be hardest hit by prolonged glucocorticoid exposure. When rats underwent experimental removal of the adrenal glands, disrupting glucocorticoid production, aged rats showed little or no evidence of hippocampal neuron loss as compared with control rats. These results link the production of glucocorticoids to the hippocampal atrophy that occurs with aging. Further evidence for this hypothesis comes from studies in Sapolsky's laboratory, showing that young rats treated with corticosterone show patterns of hippocampal cell loss similar to that in aged rats. Sapolsky and colleagues suggest that the effect of glucocorticoids on hippocampal neurons is probably related to metabolic changes stemming from the fact that glucocorticoids inhibit glucose uptake.

Rate of Decline

As discussed above, different cognitive processes decline at differing phases of the aging process. These differences are probably related to the times that neuropathological abnormalities appear in different brain regions. Cerebellar atrophy, thought to cause changes in the acquisition of a conditioned response, may start at an earlier age than most other brain changes, with significant atrophy present by the fifth decade of life. Other regions such as the medial temporal lobe or frontal lobe may not be altered until the seventh or eighth decades. Similarly, neurotransmitter changes, such as dopaminergic reductions, are thought to start around the seventh decade of life and to continue throughout the remaining adult years.

CONCLUSION

Aging does not affect all aspects of cognition uniformly. It does, however, affect a range of cognitive capacities. These changes are not static, but rather continue to intensify. Cognitive alterations are intimately linked to age-related changes in the neurotransmitter systems and in the structure and function of the brain.

Further Reading

Baddeley AD and Hitch GJ (1974) Working memory. In: Bower GH (ed.) *The Psychology of Learning and Motivation*. New York, NY: Academic Press.

Burke DM, MacKay DG, Worthley JS and Wade E (1991) On the tip of the tongue: what causes word finding failures in young and older adults? *Journal of Memory and Language* **30**: 542–579.

Craik FIM and Salthouse TA (1999) *The Handbook of Aging and Cognition*. Mahwah, NJ: Lawrence Erlbaum.

Double KL, Halliday GM, Kril JJ *et al.* (1996) Topography of brain atrophy during normal aging and Alzheimer's disease. *Neurobiology of Aging* **17**: 513–521.

Golman-Rakic PS and Brown RM (1981) Regional changes of monoamines in cerebral cortex and subcortical structures of aging rhesus monkeys. *Neuroscience* **6**: 177–187.

Hasher L and Zacks RT (1988) Working memory, comprehension, and aging: a review and a new view. In: Bower GH (ed.) *The Psychology of Learning and Motivation*, vol. 22, pp. 193–225. New York, NY: Academic Press.

Light LL and Burke DM (1993) *Language, Memory, and Aging*. New York, NY: Cambridge University Press.

Makman MH and Stefano GB (eds) (1993) *Neuroregulatory Mechanisms in Aging*. New York, NY: Pergamon Press.

Murphy DR, Craik FIM, Li KZ and Schneider BA (2000) Comparing the effects of aging and background noise on short-term memory performance. *Psychology and Aging* **15**: 323–334.

Park DC and Schwarz N (1999) *Cognitive Aging: A Primer*. Philadelphia, PA: Psychology Press.

Perfect TJ and Maylor EA (eds) (2000) *Models of Cognitive Aging*. New York, NY: Oxford University Press.

Salthouse TA (1996) The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–428.

Schacter DL, Cooper LA and Valdisseri M (1992) Implicit and explicit memory for novel objects in older and younger adults. *Psychology and Aging* **7**: 299–308.

Schultz W, Dayan P and Montague PR (1997) A neural substrate of prediction and reward. *Science* **275**: 1593–1599.

Spencer WD and Raz N (1994) Memory for facts, source, and context: can frontal lobe dysfunction explain age-related differences? *Psychology and Aging* **9**: 149–159.

Sullivan EV, Desmond JE, Deshmukh A, Lim KO and Pfefferbaum A (2000) Cerebellar volume decline in normal aging, alcoholism, and Korsakoff's syndrome: relation to ataxia. *Neuropsychology* **14**: 341–352.

Woodruff-Pak DS (1997) *The Neuropsychology of Aging*. Malden, MA: Blackwell.

Alzheimer Disease

Intermediate article

Elizabeth A Kensinger, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
Suzanne Corkin, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

Alzheimer disease is the leading cause of dementia among older adults. It results in neurochemical and neuroanatomical brain changes that cause increasing cognitive dysfunction.

INTRODUCTION

Alzheimer disease (AD) was first described by Alois Alzheimer in 1907. A German neuropathologist working in Emil Kraepelin's laboratory in Heidelberg, Alzheimer reported a case study of a 51-year-old woman who had psychiatric symptoms and memory problems. Upon her death, Alzheimer noted an abundance of neuritic plaques and neurofibrillary tangles in her brain. Although neuritic plaques had been described previously, Alzheimer was the first to recognize that their presence in large numbers was abnormal, and that the coexistence of the plaques and tangles signaled a previously unidentified disease of the cerebral cortex. This combination of neuritic ('senile') plaques and neurofibrillary tangles is now recognized as the neuropathologic signature of AD.

Alzheimer disease is the most common cause of dementia, accounting for approximately two-thirds of all cases. It is estimated that over 4 million people in the USA have AD, and by the eighth decade of life as many as one in two adults will develop the disease, which is characterized by neuronal changes, neurotransmitter abnormalities and decreased brain volume. These changes underlie cognitive deficits including memory loss, language dysfunction, and visuospatial and temporal disorientation.

DIAGNOSIS

Dementia is defined as an overall loss of intellectual function severe enough to impede daily activities. It consists of a group of behavioral symptoms that must occur together; these symptoms can have various etiologies. Alzheimer disease is defined as the presence of memory impairment plus one other area of cognitive dysfunction: language, motor, attention, executive function, personality, or object recognition, according to the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV). The deficits must have a gradual onset, and a continuous (and irreversible) progression. A definitive diagnosis of AD can be made only at postmortem examination by observing the hallmark neuritic plaques and neurofibrillary tangles. Antemortem, the diagnosis of 'probable' AD is given when an individual meets the criteria of the DSM-IV, the National Institute of Neurological Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINDS-ADRDA), and when all other causes of dementia have been eliminated. When made by a trained clinician, this exclusionary diagnosis is accurate in 80–90% of cases.

INCIDENCE AND PREVALENCE

The incidence of AD increases exponentially with advancing age. At age 70–75 years the incidence of AD is approximately 1% per year, but at age 80–85 years the annual incidence is over 6% per year (e.g. Hebert *et al*., 1995). Overall, women remain at greater risk of developing AD than men. The prevalence of AD also increases with age (e.g. Fratiglioni *et al*., 1991), with 30–50% of adults aged 70 years and above having a diagnosis of probable AD.

NEURONAL CHANGES

Alzheimer disease results in a range of neuronal changes, including cellular dysfunction and death.

Eventually, AD affects nearly all brain structures. Early in the disease, however, some brain regions (e.g. limbic structures) are affected to a much greater extent than others (e.g. the primary sensory cortices).

Neuropathologic Changes

The hallmarks of AD are neurofibrillary tangles (intracellular) and neuritic plaques (extracellular). These changes reduce the efficiency of neural communication. While normal aging is associated with these neuropathological changes as well, the number of plaques and tangles seen in the AD brain is far greater than in nondemented individuals.

Neurofibrillary tangles

Neurofibrillary tangles consist of small, paired, helical filaments (i.e. two fiber strands that are twisted around one another). They are found in the cell body and dendrites of neurons, and often appear flame-shaped, with a rounded cell body and a threadlike apical dendrite. They can also have a more spherical shape. Neurofibrillary tangles are composed of hyperphosphorylated tau protein. Typically, tau protein is a soluble component of the cell. When tau is overphosphorylated, however, it becomes insoluble and forms tangles. Because neurofibrils are frequently used for the transport of chemicals that will be made into neurotransmitters, the tangling of these fibrils renders them useless and can prevent neurotransmitter synthesis. Neurofibrillary tangles are not specific to AD, but also appear in other neurological disorders including Parkinson disease, Down syndrome, progressive supranuclear palsy and other forms of dementia.

Neuritic plaques

Neuritic plaques are dense deposits found outside the brain's nerve cells (extracellularly). They are spherical structures with a dense core of amyloid-β protein surrounded by a halo and a ring of abnormal (dystrophic) neurites. The halo component consists of other types of brain cells (astrocytes) and inflammatory cells (microglia). The dystrophic neurites represent dying nerve terminals and are small, threadlike structures consisting of abnormal neuronal dendrites. In addition to these 'typical' plaques, AD brains may also show 'diffuse' plaques, which have a loose accumulation of amyloid-β rather than a dense core, and no surrounding dystrophic neurites.

Patterns of deposition

Neurofibrillary tangles and neuritic plaques show different patterns of accumulation. Early in the course of AD, neurofibrillary tangles are confined primarily within the limbic structures. Neuritic plaques, however, appear throughout the cortex, even early in AD (Arriagada *et al*., 1992).

Relation between neuropathology and disease

It is unknown whether these neuropathological changes cause AD, or whether they are epiphenomenal. Nonetheless, there does appear to be a link between the amount of tangles in the brain and the severity of AD. Researchers are now working on therapeutic approaches to reduce the formation of tangles and plaques in the brain, hoping that this reduction will halt, or reverse, disease progression.

Brain Atrophy

One of the most prominent features of AD is atrophy (shrinkage) of the medial temporal lobe. The entorhinal cortex (a gateway for information into the hippocampus) and the hippocampus are among the first regions affected. These regions lose about 50% of their neurons, a finding that accounts for the shrunken brain tissue. The volume of these regions, measured by neuroimaging techniques, can be used to identify people with early AD, and may even identify individuals with memory impairments who will later develop AD (e.g. Jack *et al*., 1999). Another region of the brain that shows substantial cell loss early in AD is the nucleus basalis. This region of the ventral forebrain contains many of the brain's cholinergic neurons. Damage to this region reduces neurotransmission in pathways using the neurotransmitter acetylcholine.

As AD progresses the brain changes become more widespread. The ventricles of the brain expand as the surrounding tissue deteriorates. Sulci (the 'valleys' between the brain's folds) widen. Neocortical areas, including temporal and parietal neocortex, show increased atrophy. Eventually nearly all of the brain, including secondary and even primary sensory areas, is affected.

Neurotransmitter Abnormalities

The damage to the nucleus basalis results in reduced levels of choline acetyltransferase, the enzyme needed for acetylcholine formation. These reductions occur relatively early in the disease, but not all cholinergic pathways are affected equally:

the long-axon cholinergic neurons (e.g. those connecting the nucleus basalis and the cerebral cortex) are particularly vulnerable. Because of the dramatic reduction in cholinergic transmission, the first approved therapies for AD were aimed at increasing the amount of acetylcholine in the brain. Current treatment for AD is administration of acetylcholine esterase inhibitors. Acetylcholine esterase is the enzyme that breaks down acetylcholine into its constituent parts. Acetylcholine esterase inhibitors, therefore, enhance cholinergic neurotransmission by blocking the breakdown of the neurotransmitter. Acetylcholine levels remain higher and can have a longer-lasting effect before being broken down. This therapy, however, provides only a transient increase in memory performance, if any, and has no effect on disease progression.

The minimal effectiveness of acetylcholine esterase inhibitors suggests that acetylcholine deficiency is not the only cause of the cognitive dysfunction in AD. In fact, as the disease progresses, nearly all neurotransmitter systems become depleted. There appears to be much individual variation in the neurotransmitters most affected by AD and the absolute reductions. Estimates, however, suggest that levels of neurotransmitters including noradrenaline (norepinephrine), dopamine and serotonin can show reductions of up to 50% in the late stages of AD.

GENETICS

Alzheimer disease can be divided into two types: familial (inherited) and sporadic. Familial AD is relatively rare, representing less than 5% of total cases, and typically affects individuals at a younger age than sporadic AD (often before age 50 years, with cases reported of people developing the disease in their mid-20s). Sporadic AD usually has a later age at onset (after age 65 years). Some research suggests that familial and sporadic AD differ not only in terms of age at onset, but also in their cognitive profile. Familial AD may be associated with a more rapid cognitive decline and shorter time to death. It also may be linked to more verbal deficits and fewer visuospatial deficits than sporadic AD (Filley *et al.*, 1986).

Familial or Early-onset AD

Familial AD is linked to mutations in three genes: those coding for presenilin 1 (PS-1) on chromosome 14, presenilin 2 (PS-2) on chromosome 1, and amyloid precursor protein (APP) on chromosome 21 (Table 1). These mutations are causative: a person who has one of these genetic mutations will develop AD. Familial AD is inherited following an autosomal dominant pattern. This pattern means that if one parent has this form of AD, each offspring has a 50% chance of developing AD. Interestingly, all of these mutations appear to have a common effect: increased production of amyloid-β peptide (Aβ) 42, the main constituent of the amyloid plaques in the AD brain. The peptide is part of a larger precursor protein (APP), which can be cleaved (cut apart) in two different places, leading to the formation of two types of amyloid-β, one with 42 amino acids (Aβ42) and one with 40 (Aβ40). The first type appears to be more likely to form plaques in the brain than Aβ40; it may also be the more toxic form, and its presence may lead to neuronal death. Researchers are now working on ways to reduce the amount of Aβ42 formed from APP, in the hopes of stopping the formation of amyloid plaques, and perhaps also preventing further clinical decline.

Table 1. Molecular genetics of Alzheimer disease

AD Group	*Chromosome*	*Gene*	*Protein*
Familial AD, onset 50s	21	*APP*	Amyloid
Familial AD, onset 40s	14	*PS1*	Presenilin 1
Familial AD	1	*PS2*	Presenilin 2
Late-onset AD	19	*APOE*	ApoE ε2, ε3, ε4

AD, Alzheimer disease.

Sporadic or Late-onset AD

Sporadic AD probably has a multifactorial basis, including possible genetic and environmental influences. Although there are no causative mutations, there is a major genetic susceptibility factor. A gene that encodes apolipoprotein E (ApoE) is found on chromosome 19 (see Table 1). Everyone has this gene: it is essential for carrying cholesterol in our bloodstream. The gene has many alleles (or forms); the most common ones produce the ApoE variants ε2, ε3 and ε4. One allele is inherited from each parent. Being homozygous for the ε4 allele (i.e. having two ε4 alleles) is associated with an increased risk of developing AD (Saunders *et al.*, 1993). Being homozygous for the ε2 allele, in contrast, is associated with a reduced likelihood of developing AD. The allele is not predictive of AD, however; individuals without an ε4 allele can develop AD, and those who are homozygous for the ε4 allele can remain unaffected by AD. The ApoE

alleles are thought to influence the development of late-onset AD in about one-third of the population. Dozens of other genetic risk factors have been suggested, but their roles have been researched less thoroughly than the role of ApoE.

Sporadic AD is also associated with other, nongenetic risk factors. The greatest risk factor is advanced age. One well-researched correlation is with decreased estrogen levels (e.g. Paganini-Hill and Henderson, 1994). It is believed that the reason women are at greater risk of developing AD than men is because of the postmenopausal drop in estrogen levels. Taking estrogen after menopause appears to decrease the likelihood of developing AD, though it does not seem to alter its progression in those who already have the disease. Head injury is another risk factor for AD. Particularly when unconsciousness has occurred, it increases the likelihood of developing AD, and more severe injury is associated with greater risk (e.g. Guo *et al.*, 2000). Head injury as far back as early childhood appears to influence the rate of AD development later in life.

ANIMAL MODELS

Animal models of AD can provide insight into the genetics, pathological progression and treatment of AD. Several transgenic mouse models have been created with the objective of clarifying the role of genetic factors. Researchers have engineered mice that express genes implicated in AD, such as those coding for APP, the presenilins and tau; for reviews see Janus and Westaway (2001), van Leuven (2000) and Duff and Rao (2001). The characteristic pathologic findings in these mice differ depending on the genetic alterations. Transgenic mice created by introduction of APP develop neuritic plaques, and show deficits in learning and memory; however, they do not develop neurofibrillary tangles, which are the other neuropathological hallmark of AD. Transgenic mice created by inserting human tau genes develop abnormal clumping of tau filaments (neurofibrillary tangles), as well as neuronal degeneration, but do not develop neuritic plaques. Mice engineered to express PS-1, in contrast, do not display abnormal pathology or cognitive impairment, but do show elevated levels of Aβ42 (the peptide associated with plaque formation). These models, therefore, provide insights into the contributions of genes implicated in the development of familial AD. Transgenic mice can also be used to examine whether and how genetic risk factors (e.g. expression of ApoE ε4) influence disease progression.

In addition, animal models can provide clues about the pathological progression of AD. Because AD appears to occur naturally only in humans, it has not been possible to examine the neuropathological changes in the brain at various stages of the disease. Rather, the samples available have by necessity come at the time of death. While much information can be garnered from analysis of endpoint data, animal models provide a means for systematic tracking of pathological changes. Hybridizing mice genetically altered to develop neuritic plaques with those engineered to manifest neurofibrillary tangles, has resulted in a strain of mice showing both of the neuropathological hallmarks of AD. This animal model may be particularly important in contributing to our understanding of how these two neuropathological features relate. Researchers are optimistic that use of such animal models will provide information about the relation between amyloid deposits and tau-containing tangles, and about the role they play in the development and progression of AD.

Animal models will also be important in testing potential treatments for AD. Once researchers have established an animal model that closely approximates the pathological and cognitive characteristics of AD, it will be possible to test the efficacy of treatments on these animals.

COGNITIVE FUNCTION

The signs of AD develop slowly, and it is often difficult to pinpoint the time of disease onset. Initial symptoms are mild, and can include forgetfulness, passivity, decreased work productivity, word-finding difficulties, and disorientation. As the disease progresses, nearly all aspects of function are affected, including memory, language, attention, vision, audition, and motor control.

Impaired Capacities

Episodic memory

Impairment of episodic memory – the recollection of events that occupy a specific spatial and temporal context – is typically the earliest and most prominent deficit in AD. Patients have difficulty forming new episodic memories (anterograde amnesia), and this impairment worsens with disease progression. Deficits in episodic memory, including delayed recall of verbal and nonverbal material, are the best way of distinguishing people with AD from healthy older adults (e.g. Locascio *et al.*, 1995). In contrast, however, people with AD remain capable of retrieving some long-term episodic memories. While remote memory is impaired, it

does not show the pronounced decrements seen in the formation of new episodic memories. The loss of retrograde memory (retrograde amnesia) also appears to be temporally graded, with recent memories showing more degradation than remote memories. In fact, the capacity to recollect events from the remote past is often quite resilient in AD. Patients can even become preoccupied with the past, and can confuse their current environment with that of their youth. The degrees of anterograde and retrograde memory deficits are not significantly correlated (e.g. Greene and Hodges, 1996).

The episodic memory deficit is consistent with the neural changes early in AD: brain structures that support long-term memory (medial temporal lobe regions, including the hippocampus) are compromised in early AD, while other regions of the brain are less affected.

Emotional memory

Individuals are often better at remembering emotional compared with neutral information. This emotional memory enhancement effect appears to result from interactions between the amygdala and other regions of the medial temporal lobe, including the hippocampus. Alzheimer disease results in a substantial volumetric reduction in the amygdala, and this amygdaloid atrophy appears to reduce the emotional enhancement effect: people with AD do not show better memory for emotional information than for neutral information (Kensinger *et al*., 2002) and their ability to remember emotional stimuli appears to correlate with amygdaloid volume.

Semantic memory

Semantic memory – general knowledge about the world – is relatively spared early in AD, but as the disease progresses significant deficits arise. Deficits occur on tasks requiring general knowledge retrieval, word definitions, word–picture matching, or picture naming. It is unclear whether the semantic deficit is related to a breakdown in the structure of semantic memory, to impaired access of semantic information, or to a combined deficit in structure and access.

The extent of the language deficits is useful for assessing the severity of AD (Locascio *et al*., 1995). Initial deficits include increased 'tip of the tongue' effects, reduced fluency scores, and a difficulty with tests requiring confrontation naming. In later stages of the disease, deficits can include forgetting the names of spouse or children, and the inability to recall names of common objects. The progression of semantic memory deficits is associated with the expanding pathological changes of advancing AD: as perisylvian areas become affected, semantic memory deficits increase.

Visuospatial function

While early AD is associated with some disorientation, visuospatial dysfunction increases with disease progression. In the middle stages of the disease, it is common for patients to become lost while driving their car or on a walk, even when following a route that they have taken on many occasions.

Executive functions

People with AD show deficits in short-term memory and in on-line processing of information (Corkin, 1982). At least some of these deficits may be due to deficits in the 'central executive' in Baddeley's model of working memory. Becker (1988) suggested that AD might have two main deficits: one paralleling that of amnesia, and the other in the central executive. In support of the central executive hypothesis, people with AD have frequently been found to have poor dual-task performance while being capable of performing each component task at a normal level. Baddeley and colleagues also found that dual-task performance declined with disease progression, while single-task performance remained stable.

Classical conditioning

Most forms of nondeclarative memory are spared in AD. One notable exception, however, is seen with delay conditioning, in which the unconditioned stimulus occurs just before the offset of the conditioned stimulus. People with AD are impaired at acquiring a conditioned response (such as an eyeblink in response to a tone). They require more trials to learn this type of relation (e.g. Woodruff-Pak *et al*., 1996). This deficit is probably not related to damage to the medial temporal lobe because amnesic patients with damage to these structures are capable of acquiring a conditioned response. The deficit is more likely to be related to cholinergic or cerebellar dysfunction.

Vision

Alzheimer disease results in changes in basic sensory and perceptual capabilities. People with AD often have more difficulty perceiving visual stimuli than do nondemented older adults. A significant correlation exists between the severity of perceptual deficits and the extent of cognitive dysfunction (e.g. Cronin-Golomb *et al*., 1995). It is unclear whether this correlation is causal (e.g. visual deficits could cause poorer performance on a task of

visual memory) or associative (e.g. patients with greater brain atrophy are likely to have both sensory deficits and memory dysfunction). The visual deficits are related to reductions in contrast sensitivity, color perception and discrimination, and visual acuity. The visual dysfunction is probably related to neuropathological changes in primary visual cortices and visual association areas because AD does not appear to affect the retina or optic nerve.

Preserved Capacities

Despite the range of capacities affected in AD, some domains remain relatively preserved until late in the course of the disease. Most prominently, many types of nondeclarative (implicit) memory are unaffected by early to moderate AD.

Priming

Priming can be broadly broken down into two subsets: conceptual and perceptual priming. People with AD show a dissociation in performance on these priming tasks: their performance on conceptual priming tasks is impaired, whereas their perceptual priming is normal. This dissociation probably reflects a disproportionate reliance on temporoparietal regions in conceptual priming. Conversely, perceptual priming appears to rely predominantly on occipital lobe regions that are less affected by AD.

Skill learning

Until the late stages of AD, patients are capable of learning new skills, ranging from motor learning to visual adaptation. The preservation of such learning is likely to be related to the relative preservation of brain regions important for nondeclarative learning, including the basal ganglia and frontal lobe.

TREATMENT

The treatment of AD consists predominantly of three types of approaches: (a) mitigating noncognitive disorders (psychiatric symptoms such as anxiety or paranoia, sleep disturbance), (b) restoring neurotransmitter function, and (c) protecting neurons from further death. The majority of drugs prescribed to restore neurotransmitter function have been cholinesterase inhibitors. These drugs, however, have had only minimal effectiveness in slowing the disease progression and have not been able to halt or reverse the disease's effects. There is some evidence that antioxidants (vitamin E), antiinflammatory drugs (ibuprofen), estrogen, and lipid-lowering agents (statins) may be neuroprotective, in as much as they slow the progression of AD. To date, however, there is no evidence that these treatments can slow disease progression in individuals already affected.

SIGNIFICANCE OF RESEARCH FOR UNDERSTANDING BRAIN FUNCTION

Research into AD has improved our understanding not only of the neurologic disorder but also of brain function. By observing the pattern of spared and impaired functions, researchers have learned that some types of memory (e.g. declarative) are affected by diffuse damage to the medial temporal lobe, while other types of memory (i.e. nondeclarative) remain relatively unaffected. Similar dissociations, such as between preserved perceptual priming and impaired conceptual priming, or between impaired anterograde memory and only moderately affected retrograde memory, have helped to uncover the dissociable neural mechanisms responsible for these cognitive functions. Similarly, the finding of reduced cholinergic function in AD sparked interest in the role of acetylcholine in long-term memory formation. Further research has demonstrated the necessity of the cholinergic system for successful episodic encoding.

By comparing performance in AD and amnesia, researchers have also been able to determine what memory dysfunction in AD is caused specifically by damage to the medial temporal lobe system, and what deficits may be related to neocortical damage (Corkin, 1982). This complementary interaction between neurology, neuropsychology, and neuroscience has allowed simultaneous advancements in the diagnosis and treatment of AD, and in our understanding of brain–behavior relations.

References

Arriagada PV, Growdon JH, Hedley-Whyte ET and Hyman BT (1992) Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology* **42**: 631–639.

Becker JT (1988) Working memory and secondary memory deficits in Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology* **10**: 739–753.

Corkin S (1982) Some relationships between global amnesias and the memory impairments in Alzheimer's disease. In: Corkin S, Davis KL, Growdon JH, Usdin E and Wurtman RJ (eds) *Alzheimer's Disease*, vol. 19, A Report of Progress in Research, pp. 149–164. New York, NY: Raven Press.

Cronin-Golomb A, Corkin S and Growdon JH (1995) Visual dysfunction predicts cognitive deficits in

Alzheimer's disease. *Optomology and Visual Science* **72**: 168–176.
Duff K and Rao MV (2001) Progress in the modeling of neurodegenerative diseases in transgenic mice. *Current Opinion in Neurology* **14**: 441–447.
Filley CM, Kelly J and Heaton RK (1986) Neuropsychologic features of early- and late-onset Alzheimer's disease. *Archives of Neurology* **43**: 574–576.
Fratiglioni L, Grut M, Forsell Y *et al*. (1991) Prevalence of Alzheimer's disease and other dementias in an elderly urban population: relationship with age, sex, and education. *Neurology* **41**: 1886–1892.
Greene JDW and Hodges JR (1996) The fractionation of remote memory: evidence from the longitudinal study of dementia of Alzheimer's type. *Brain* **119**: 129–142.
Guo Z, Cupples LA, Kurz A *et al*. (2000) Head injury and the risk of AD in the MIRAGE study. *Neurology* **54**: 1316–1323.
Herbert LE, Scherr PA, Beckett LA *et al*. (1995) Age-specific incidence of Alzheimer's disease in a community population. *Journal of the American Medical Association* **273**: 1354–1359.
Jack CR, Peterson RC, Xy YC *et al*. (1999) Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* **52**: 1397–1403.
Janus C and Westaway D (2001) Transgenic mouse models of Alzheimer's disease. *Physiology and Behavior* **73**: 873–886.
Kensinger EA, Brierley B, Medford N, Growdon JH and Corkin S (2002) Effects of normal aging and Alzheimer's disease on emotional memory. *Emotion* **2**.
van Leuven F (2000) Single and multiple transgenic mice as models for Alzheimer's disease. *Progress in Neurobiology* **61**(3): 305–312.
Locascio JJ, Growdon JH and Corkin S (1995) Cognitive test performance in detecting, staging, and tracking Alzheimer's disease. *Archives of Neurology* **52**: 1087–1099.
Paganini-Hill A and Henderson VW (1994) Estrogen deficiency and risk of Alzheimer's disease in women. *American Journal of Epidemiology* **140**: 256–261.
Saunders A, Strittmater W, Schmechel D *et al*. (1993) Association of apolipoprotein E allele e4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**(8): 1467–1472.
Woodruff-Pak DA, Papka M, Romano S and Lo YT (1996) Eyeblink classical conditioning in Alzheimer's disease and cerebrovascular dementia. *Neurobiology of Aging* **17**: 505–512.

Further Reading

Baddeley AD, Bressi S, Della Sala S, Logie R and Spinnler H (1991) The decline of working memory in Alzheimer's disease: a longitudinal study. *Brain* **114**: 2521–2542.
Corkin S (1998) Functional MRI for studying episodic memory in aging and Alzheimer's disease. *Geriatrics* **53**: S13–S15.
Growdon JH, Wurtman RJ, Corkin S and Nitsch RM (eds) (2000) The molecular basis of dementia. *Annals of the New York Academy of Sciences*, vol. 920. New York, NY: New York Academy of Sciences.
Hodges JR (2000) Memory in the dementias. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*. New York, NY: Oxford University Press.
Katzman R, Terry RP and Bick KL (eds) (1994) *Alzheimer's Disease*. New York: Raven Press.
Keane MM, Gabrieli JD, Fennema AC, Growdon JH and Corkin S (1991) Evidence for a dissociation between perceptual and conceptual priming in Alzheimer's disease. *Behavioral Neuroscience* **105**: 326–342.
Morris RG and Kopelman MD (1986) The memory deficits in Alzheimer-type dementia: a review. *Journal of Experimental Psychology* **A38**: 575–602.
Nebes RD (1989) Semantic memory in Alzheimer's disease. *Psychological Bulletin* **106**: 377–394.
Selkoe DJ (1999) Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature* **399**: A23–A31.
Villareal DT and Morris JC (1998) The diagnosis of Alzheimer's disease. *Alzheimer's Disease Review* **3**: 142–152.

Amnesia

Intermediate article

AR Mayes, University of Liverpool, Liverpool, UK
NM Hunkin, University of Sheffield, Sheffield, UK

CONTENTS

Amnesia means loss of memory; more specifically, the term refers to the amnesic syndrome. In this syndrome, usually caused by regional brain damage, there is impaired recall and recognition of facts and episodes encountered both before and after the onset of the disorder, although intelligence and short-term or immediate memory are preserved.

INTRODUCTION

The term 'amnesia' is used in a general sense to mean any kind of loss of memory. The term also has a technical sense in which it refers to the amnesic syndrome. This memory disorder syndrome has four main features. The first is an impairment in the ability to recall or recognize facts or personally experienced episodes encountered following the brain injury; this is known as anterograde amnesia. The second feature is an impairment in the ability to recall or recognize facts and personally experienced episodes that were encountered and put into memory premorbidly; this is known as retrograde amnesia. In patients, the severity of anterograde amnesia is only weakly correlated with the severity of retrograde amnesia, and these disorders are often accompanied by two further features: preserved intelligence, and preserved short-term memory. Research has confirmed earlier, more informal observations, the first of which was made by Claparède in 1911, that people with amnesia can show good memory although they have no recall or recognition of how they acquired such memory. These 'implicit' memories are only shown by changed behavior, because patients are not aware that they are remembering.

Amnesia in the technical sense is, therefore, far from being a general disorder of all kinds of memory. It is referred to as a syndrome because some of its component symptoms probably have different causes, although it remains unclear to what extent the cause or causes of anterograde and retrograde amnesia differ. Most often these memory disorders are caused by damage to any one of several different brain regions, each of which plays a key role in certain kinds of memory (organic amnesia); but it is also known that amnesic symptoms can result from psychiatric causes that are not necessarily triggered by brain damage, and which typically only affect premorbid autobiographical memories (psychogenic amnesia).

Although Lawson was one of the first to describe a relatively selective memory disorder associated with chronic alcoholism in 1878, it was the Russian physician Korsakoff who, between 1887 and 1891, gave the classic description of the amnesic syndrome. Most of the cases he described were of chronic alcoholism, but he also described patients with neoplasm, carbon monoxide poisoning, diabetes and persistent vomiting. Korsakoff syndrome, the variant of amnesic syndrome that bears his name, is believed to be caused by thiamin deficiency. Most often thiamin deficiency is caused by chronic alcoholism, but in the twentieth century it became clear that other forms of nutritional disorder such as anorexia nervosa or persistent vomiting may cause thiamin deficiency and amnesia. Korsakoff also noted that patients with his syndrome usually showed lack of insight into their condition and often falsely recollected information (confabulated). These symptoms have subsequently been found not to be universal, or even particularly common, features of other kinds of amnesia.

Later work clearly showed that the amnesic syndrome could arise from a wide range of other causes including several kinds of vascular incident (such as infarctions of the posterior cerebral artery, and rupture and repair of anterior communicating artery aneurysms), herpes simplex encephalitis and some kinds of meningitis, as well as surgical

treatment of temporal lobe epilepsy involving bilateral removal of the medial temporal lobes. The most famous amnesic patient, known by his initials as HM, had such an operation in 1953, and nearly fifty years later still showed a very severe anterograde amnesia and a less severe retrograde amnesia mainly affecting memories acquired in the years immediately preceding his surgery. It was first suggested by Ribot in 1882 that memories acquired further in the past before brain damage occurred are more resistant to impairment. In other words, he predicted that retrograde amnesia should be temporally graded such that older memories are less impaired.

The history of research on amnesia may be broken into three overlapping stages. The first stage, which began with Lawson, Korsakoff and Ribot, involved systematic clinical observation of patients. The second stage, which began after the Second World War, involved the application to amnesic patients of more formal experimental procedures drawn from mainstream cognitive psychology. In the third stage, modern *in vivo* brain imaging technologies made it possible to determine exactly what brain lesions cause which kinds of memory deficits. This work is complemented by the exploration of which brain regions are activated when normal people use the memory processes that are impaired in amnesic patients. For decades, work on human amnesic patients has been complemented by research on animal models of amnesia in which it is possible to localize damage more focally.

FORMS OF AMNESIA

Amnesia is divisible into organic and psychogenic forms. Psychogenic or functional amnesia, which includes fugue states and multiple personality disorders, typically has no apparent physical cause and is believed to have a psychiatric origin. People with psychogenic amnesia are believed to be neither merely pretending to be unable to remember, nor consciously trying not to remember or create new memories, although it is often possible to identify emotional benefits arising from such patients being unable to remember. This contrasts them with malingerers, who are pretending or consciously trying not to remember.

Fugue states consist of a sudden loss of all autobiographical memory together with the loss of personal identity. Typically, these states are short-lived, lasting only a few hours or days, although there are reports of fugues being maintained over long periods. Fugues tend to be precipitated by a traumatic event such as bereavement or marital strife, and the memory loss may be interpreted as a form of adaptive response in which patients dissociate themselves from their own identity as a means of escaping from the traumatic experience. In multiple personality disorder, a patient appears to have at least two different personalities, and sometimes as many as 20. The different personalities are used to compartmentalize the patient's experiences, and a particular personality needs to be active for particular experiences to be recalled. In addition to fugues and multiple personality states, there are cases when loss of autobiographical memory for the premorbid past is less complete. In these cases the person usually has a retrograde amnesia that is selective for autobiographical memory, but with no impairment of the ability to acquire new autobiographical memories. The ability to acquire new memories is only rarely disrupted in cases of psychogenic amnesia.

Organic and psychogenic amnesias can usually be distinguished from each other by examining the patient's medical history. If the patient has suffered a brain injury, and there is damage to brain areas known to be involved in memory functioning, an organic basis to the memory disorder is likely. In contrast, if there has been no brain trauma or evidence of brain disease, it is more likely that the memory disorder has a psychogenic basis. Some cases are more difficult to interpret; in these, there is evidence of brain damage that is either inconsistent with the pattern of memory impairment observed or does not seem extensive enough to explain the severity of the observed memory impairment. In such cases psychogenic factors may overlie an organic disorder or, conversely, organic factors such as epilepsy may exacerbate an underlying psychogenic disorder.

Like psychogenic amnesia, organic amnesia may be transient, although more often it is permanent. Transient organic amnesia may occur following electroconvulsive therapy and epileptic seizures. It also occurs in post-traumatic amnesia, during the early period following head injury, or emergence from coma in severely injured patients when patients have difficulty in storing and/or retrieving new information. Finally, best known is transient global amnesia, in which there is an abrupt onset of severe anterograde amnesia and a more variable, temporally graded retrograde amnesia. The attack typically lasts for a number of hours and then gradually resolves, leaving the patient with a dense loss of memory covering the period of the attack and a few hours preceding it, but with a more or less restored ability to lay down

new memories. It has been shown that blood flow to the medial temporal lobes and sometimes the thalamic region is temporarily reduced during attacks. The cause of this reversible dysfunction of memory-related regions (see next section) is unknown, but is sometimes associated with migraine.

The severity of the memory deficits in permanent organic amnesia usually remains stable, although there may be some recovery following the initial brain injury. However, amnesia is also typically the major characteristic of the early stages of Alzheimer disease: in this neurodegenerative disease not only does the amnesia became worse as the disease progresses, but other cognitive and memory deficits also appear.

Psychogenic amnesia typically involves a retrograde amnesia for personal episodes and no anterograde amnesia. It is interesting, therefore, that it has recently been claimed that brain damage can also cause a retrograde amnesia in the context of minimal anterograde amnesia. Cases of such focal retrograde amnesia have been used to identify which brain structures are involved in remote memory function. However, such research is controversial because (a) it is difficult to explain how a patient can fail to recall a premorbidly formed memory while being able to store and retrieve new memories fairly normally, and (b) it is difficult to exclude the possibility that the amnesia has a psychogenic explanation.

NEUROANATOMY OF AMNESIA

Damage to any one of several distinct brain regions, which include the medial temporal lobes, the midline diencephalon, the basal forebrain and possibly parts of the prefrontal neocortex, is sufficient to cause the amnesic syndrome. Although these regions comprise distinct gray matter structures, they are all interconnected and there is evidence that damage to the fibre tracts, such as the fornix, which link the different regions can also cause the syndrome, or at least some of its characteristic memory impairments (Tranel and Damasio, 1995).

The medial temporal lobes include the perirhinal and parahippocampal cortices, which receive processed information from the posterior and anterior association neocortices. They provide nearly two-thirds of the cortical input to the hippocampus via an intermediate projection to the entorhinal cortex (Figure 1). The hippocampus, therefore, receives inputs of processed modality-specific and polysensory information, semantic information and possibly other kinds of information related to planning and movement activities from the association

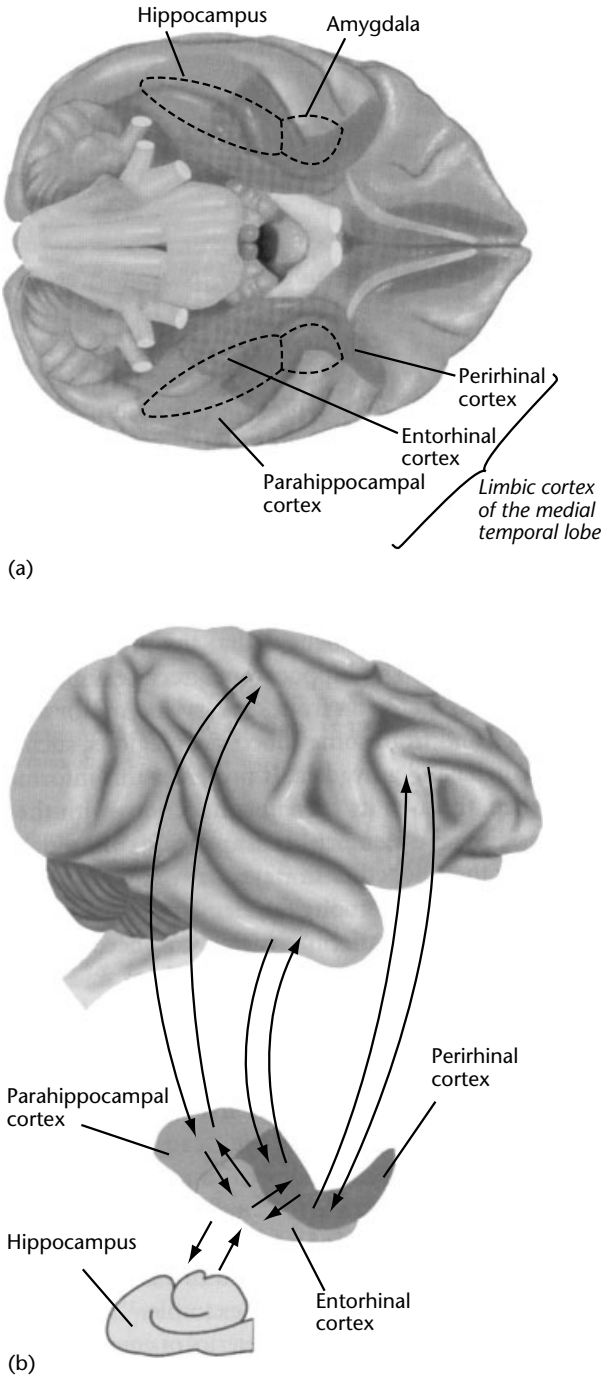


Figure 1. (a) Ventral view of a typical nonhuman primate (monkey) brain which illustrates the position of the medial temporal lobe cortices and the hippocampus and amygdala (which lie just above them). (b) Sagittal view of the same structure, that illustrates the two way connections which exist between the hippocampus and entorhinal cortex; the entorhinal, parahippocampal, and perirhinal cortices; and the parahippocampal and perirhinal cortices, and overlying neocortical structures.

cortices. Combined damage to all medial temporal lobe regions produces severe anterograde and retrograde amnesia in humans and monkeys (Zola-Morgan and Squire, 1993). Research with monkeys has shown that perirhinal cortex lesions alone cause severe anterograde amnesia with badly impaired recognition, whereas parahippocampal cortex lesions may primarily disrupt some kinds of spatial memory. Hippocampal lesions are known to produce mild anterograde and retrograde amnesia, although there is controversy about the nature of this impairment. In humans, this is because selective hippocampal damage is rare, and some of the patients reported have shown equivalent recognition and recall deficits, whereas others have shown moderately severe recall deficits but mild or undetectable recognition deficits. These selectively impaired patients show milder retrograde amnesias than are found in patients with large medial temporal lobe lesions. There is also evidence that severe (and perhaps focal) retrograde amnesia may depend on damage that includes parts of the anterior temporal neocortex and possibly parts of the prefrontal neocortex. The medial temporal lobes contain the amygdala, which is interconnected with the hippocampus and lies above the perirhinal cortex. Although studies with monkeys indicate that amygdala lesions do not usually impair recognition, there is growing evidence that this structure is essential to some aspects of emotional memory and that damage to it reduces the memory advantage of 'emotional' stimuli over neutral ones (Phelps *et al*, 1998).

Evidence from both animals and humans indicates that damage to either the mamillary bodies or the nuclei and fibre tracts in the midline thalamus also causes amnesia. It is not agreed, however, precisely which thalamic regions are implicated in amnesia, although there is good evidence that damage to either the anterior thalamic or the dorsomedial thalamic nuclei causes memory impairments. The memory disorder in Korsakoff syndrome is particularly associated with lesions of the mamillary bodies, and of the anterior and dorsomedial thalamic nuclei.

Structures in the basal forebrain modulate neocortical and medial temporal cortex activity, so it is not surprising that damage to these structures can produce amnesia. The septum and diagonal band of Broca project to the hippocampus via the fornix, and lesions to these basal forebrain structures have been found to cause amnesia. The same has also been claimed for lesions of the nucleus accumbens, a structure that also has hippocampal and other limbic system connections.

The prefrontal neocortex is a large region, and although lesions to some parts of it (e.g. the dorsolateral prefrontal cortex) have been found to disrupt particular kinds of memory in humans and animals, it is often argued that, in contrast to organic amnesia, these deficits are usually mild and are secondary to failures of planning that impair encoding and retrieval of complex information, and hence disrupt memory. It remains possible, however, that damage to those frontal regions, such as the orbitofrontal cortex, which receive the heaviest projections from relevant thalamic nuclei and the medial temporal lobes (Tranel and Damasio, 1995), causes a memory disturbance that is more similar to the amnesic syndrome.

Damage to the fornix, which connects the hippocampus directly (and indirectly via the mamillary bodies) to the anterior thalamus, has been shown to cause a mild amnesia. This includes a mild retrograde amnesia, although – as with hippocampal lesions – it is disputed to what extent the anterograde amnesia caused by the lesion involves recognition as well as recall. A similar disagreement extends to the mamillary bodies and the anterior thalamus.

The extent to which all the structures discussed above function as a unitary and integrated system subserving memory for facts and episodes, and to what degree damage to the different structures impairs these forms of memory in subtly different ways, is not yet clear. Figure 2 illustrates some of the connections between the structures implicated in amnesia. These connections may provide clues about the kinds of processing that may underlie fact and episode memory.

IMPLICIT AND EXPLICIT MEMORY

The terms 'explicit memory' and 'implicit memory' were introduced by Graf and Schacter (1985). Explicit memory is the form of memory required when people know that they are remembering (often a particular episode). It is believed to be essential for performing tasks involving free recall, cued recall or recognition of previously presented information. In contrast, implicit memory is the form of memory required to mediate performance on tasks in which memory is indicated by some behavioral change without the person being aware of remembering something. Implicit memory is functionally heterogeneous and includes skill learning, various forms of simple conditioning, and perceptual learning. It also includes an information-specific kind of implicit memory called priming, which mediates performance on tasks

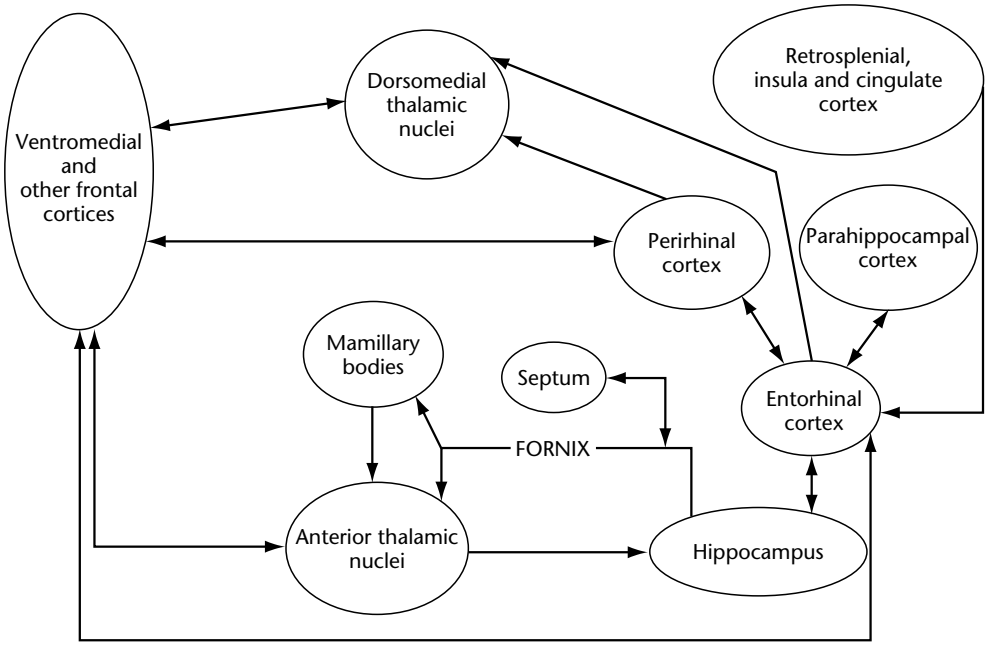


Figure 2. Flow chart illustrating the connections which exist in the primate brain between the components of the medial temporal lobes, the midline diencephalon, and the prefrontal neocortex. Damage to any of these structures is believed to cause at least some of the memory symptoms of the amnesic syndrome. The connections between the structures may provide a useful constraint on theories about what functional deficits underlie the amnesic syndrome.

such as fragment completion, perceptual identification and speeded reading. In experimental investigation into these tasks, participants first study specific information (e.g. words or pictures) and then, after a delay, are required to perform some operation related to that information (such as identifying a briefly displayed studied word). Techniques are used to minimize the degree to which participants use explicit memory to influence their performance. Priming is indicated to the extent that study influences subsequent behavior (e.g. identification) towards the studied information when the behavior is not affected by explicit memory. For example, the speed of reading studied words, compared with similar unstudied words, may be greater; or briefly displayed studied words may be better identified, even when the words are not recognized as having been studied.

Although people with organic amnesia are typically impaired at all tasks involving explicit memory, it has been suspected since the time of Claparède that implicit memory may be preserved. There is good evidence that such patients are relatively unimpaired at acquisition and retention of motor, perceptual and even cognitive skills. Impairments have only become apparent when normal participants' performance, unlike that of the amnesic patients, has been enhanced by their ability to use explicitly remembered strategies. Lacking explicit memory of the learning experiences, amnesic patients are often surprised by their own skills. Similar preservation of simple kinds of classical conditioning has been found. There is some evidence that patients show impaired conditioning when there is a short delay between the end of the conditioned stimulus and the start of the unconditioned stimulus, but it has been argued that this occurs because trace conditioning depends on explicit memory for the relationship between the stimuli. Patients also show preservation of several kinds of perceptual learning and memory.

Evidence about whether people with organic amnesia show preserved priming for all kinds of information is more conflicting. Most but not all researchers believe that patients show preserved priming when the studied information – whether perceptual or semantic in nature – is already familiar and in memory prior to study (e.g. words or famous faces). However, a meta-analysis by Gooding *et al* (2000) indicates that although this is true, amnesic patients' priming of novel items and novel associations is impaired. It might be argued that people with amnesia only show impairments on priming tasks when performance in normal people is facilitated by their superior explicit memory; but this explanation is somewhat implausible because the meta-analysis deliberately compared priming of similar kinds of familiar and

novel information on which participants performed similar priming operations at test (for example, made word or non-word judgments as quickly as possible). Also, it cannot explain why priming of novel associations is impaired even when normal participants have no recognition of studied stimuli. There has been no study of whether people with amnesia show preservation of priming for premorbidly acquired memories, so it is unknown whether such priming might differentiate between organic and psychogenic amnesia. However, when both priming and explicit memory for the same information are impaired it is likely that the information never has been or is no longer in long-term storage.

WHAT CAUSES AMNESIA?

Theories about the causes of amnesia must specify what processes break down to cause the syndrome, what lesions are responsible for this, and why. Initially, theorists treated amnesia as a unitary deficit in which only one functional process was disrupted, and all the lesions that produce amnesia disrupted this process although perhaps to differing degrees. Most researchers now believe that the syndrome comprises several distinct functional deficits, each caused by a different lesion (or, more probably, set of lesions), because it is also believed that processes are mediated by systems of structures rather than individual structures.

Theories may differ with respect to: (a) whether encoding, storage or retrieval processes are disrupted in amnesia; (b) what kinds of fact and episode information the impaired processes directly affect; and (c) what specific structures are thought to be critically damaged and, thus, necessary for the execution of the process in question. It has been proposed that defective encoding (processing and representing) of semantic information causes amnesia, or at least the amnesia of Korsakoff syndrome (Butters and Cermak, 1980). The patients' relatively normal intelligence and ability to show normal memory benefits from directed semantic encoding argue against this view. Also, people with amnesia can answer questions normally about semantic (and some other kinds of) information when the questions are asked immediately following presentation. Amnesia is, therefore, unlikely to be caused by a semantic encoding deficit. However, some evidence suggests that perirhinal cortex lesions may impair integrative encoding of high-level sensory information. This possible visual object encoding deficit remains to be systematically tested in humans.

There have been few retrieval deficit theories of amnesia. The most popular such account suggests that people with amnesia may suffer excessively from interference during retrieval and so should show abnormal numbers of intrusion errors, and benefit abnormally from cues restricting the number of possible responses (Warrington and Weiskrantz, 1974). The evidence favoring these predictions is not good, and the theory also leaves unexplained why people with amnesia have this retrieval problem unless it is a secondary effect of a storage defect. Such a defect could cause a secondary problem with interference during retrieval if contextual associations to studied information were not stored. In such a case, contextual markers would not be available to identify which of several competing alternatives was the appropriate memory in a given situation, with the result that intrusion errors would be made unless appropriately restrictive cues were provided. In general, however, retrieval theories imply that people with amnesia store information normally (in anterograde amnesia) or retain it normally (in retrograde amnesia). This possibility is hard to refute decisively so it still needs to be proved convincingly that amnesic patients do not store fact and episode information normally. If such patients do store this information normally, then they should show preserved priming for the same fact or episode information for which their explicit memory is impaired.

Most theories have suggested that amnesia is caused by a failure of the consolidation processes that put all fact and episode information into long-term storage. More specifically, a failure to put the associations linking together the components of facts and episodes normally into memory storage in the minutes or hours following exposure is postulated to cause anterograde amnesia. Retrograde amnesia is explained in terms of a retention failure for the same kinds of association. However, many researchers believe that anterograde and retrograde amnesia are closely linked deficits because there is a slow consolidation process, perhaps involving rehearsal and repetition, through which long-term storage of memories is eventually achieved in the neocortex (Squire and Alvarez, 1995). If they are correct, then memories acquired closer in time to the brain injury should be more disrupted than memories acquired earlier. The alternative view proposes that such temporally graded retrograde amnesia does not occur because memories about episodic incidents (although perhaps not facts) continue to depend on the medial temporal lobes and perhaps other structures

implicated in amnesia for as long as they last (Nadel and Moscovitch, 1997). The evidence is conflicting.

Aggleton and Brown (1999) have proposed one of the currently most influential theories. Their theory states that amnesia is functionally heterogeneous. They argue that lesions of the hippocampus, fornix, mamillary bodies or anterior thalamus disrupt rapid associative memory consolidation such that recall memory is greatly impaired, but item recognition – except when it relies heavily on recollection – is relatively intact. Another version of this view is that these lesions selectively impair recall because they impair the consolidation of contextual associations such as those involving spatial and temporal relations. In contrast, the theory postulates that lesions to the perirhinal cortex or some of its projections disrupt 'true recognition' or familiarity. Perirhinal cortex lesions also disrupt recollection because they prevent critical inputs reaching the hippocampus. This theory is controversial, as is the view that anterior temporal neocortex lesions disrupt long-term storage of facts and episodes so perhaps causing focal retrograde amnesia (see Kapur, 1992). However, both theories illustrate the increasing belief that the amnesic syndrome is functionally heterogeneous.

TREATMENT OR REMEDIATION

Organic amnesia is permanent and stable when it is caused by destruction of memory-relevant brain regions. In the future, the condition might possibly be treated effectively by transplantation of stem cells to replace some of the lost neurons and form appropriate connections; at present, however, only remediation using strategies to compensate for the permanently impaired memory processes is feasible. Psychogenic amnesia is typically transient, but when it is long-lasting it may be treated by hypnosis or by drugs such as amobarbital, both of which therapies can reduce patients' resistance to retrieving memories.

Remediation of permanent organic amnesia is important because severe memory impairment typically means that patients will be easily disoriented, have only a hazy sense of their past, cannot work, and may require careful supervision in the home. They may also be bored because social interactions and interesting pastimes that depend on memory, such as reading and watching television, can no longer be pursued.

Recent research has focused on alleviating the consequences of having a deficient memory rather than attempting to improve the impaired underlying memory processes. Four kinds of technique have been moderately successful. First, the use of cognitive strategies (such as imagery in teaching the name of a new acquaintance) has been effective, although patients rarely use these strategies unless specifically instructed to do so.

Second, Baddeley and Wilson (1994) have developed an 'errorless learning' technique, in which the possibility of making errors during training is eliminated by repeatedly providing the information to be learned and thereby preventing guessing. For example, patients learn a series of face–name associations by being repeatedly shown a picture of a face immediately followed by its name, which has to be read or copied. This method has been used successfully to teach a range of information. It is uncertain whether it relies on the patients' use of implicit memory or priming, or their residual explicit memory.

Third, Glisky and Schacter (1987) developed the method of 'vanishing cues' for teaching specific factual information to facilitate activities in a particular domain. They taught an amnesic patient a set of computer terms and procedures by gradually reducing the strength of cues given. The patient was able to return to employment where this knowledge could be used. Like the errorless learning technique, this method leads to very few errors being made, and evidence that the memory it produces is dependent on the cues matching those used during learning suggests that it may depend on priming.

Fourth, the ability of people with amnesia to cope may be improved by modifying their immediate environment – for example, by keeping to a regular routine or encouraging the use of a diary. The repetition produced by regularizing the local environment in this way may help patients by allowing them to form important long-term memories using their relatively intact, but slow to consolidate, neocortical storage sites. Like the methods of errorless learning and vanishing cues, the approach could also be effective because the patients' relatively good memory for local (and other) information may depend to a greater extent than does the memory of normal people on implicit memory or priming. There is evidence that this form of memory is much more sensitive than explicit memory to the effects of interference (see Mayes *et al*, 1987), so that the generation of errors would be much more likely to impede appropriate learning in people with amnesia than it would in normal people. Thus, the prevention of errors by making the local environment more regular and repetitive

might facilitate learning in amnesic individuals as well as providing them with a less memory-demanding set of surroundings.

References

Aggleton JP and Brown MW (1999) Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences* **22**: 425–443.

Baddeley A and Wilson BA (1994) When implicit learning fails: amnesia and the problem of error elimination. *Neuropsychologia* **32**: 53–68.

Butters N and Cermak LS (1980) *Alcoholic Korsakoff's Syndrome: An Information Processing Approach to Amnesia*. New York, NY: Academic Press.

Glisky EL and Schacter DL (1987) Acquisition of domain-specific knowledge in organic amnesia: training for computer-related work. *Neuropsychologia* **25**: 893–906.

Gooding PA, Mayes AR and van Eijk R (2000) A meta-analysis of indirect memory tests for novel material in organic amnesics. *Neuropsychologia* **38**: 666–676.

Graf P and Schacter DL (1985) Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory and Cognition* **11**: 501–518.

Kapur N (1992) Focal retrograde amnesia in neurological disease: a critical review. *Cortex* **29**: 217–234.

Mayes AR, Pickering A and Fairbairn A (1987) Amnesic sensitivity to proactive interference: its relationship to priming and the causes of amnesia. *Neuropsychologia* **25**: 211–220.

Nadel L and Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology* **7**: 217–227.

Phelps EA, LaBar KS, Anderson AK *et al*. (1998) Specifying the contributions of the human amygdala to emotional memory: a case study. *Neurocase* **4**: 527–540.

Squire LR and Alvarez P (1995) Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology* **5**: 169–177.

Tranel D and Damasio AR (1995) Neurobiological foundations of human memory. In: Baddeley AD, Wilson BA and Watts FN (eds) *Handbook of Memory Disorders*, pp. 27–50. New York, NY: Wiley.

Warrington EK and Weiskrantz L (1974) The effect of prior learning on subsequent retention in amnesic patients. *Neuropsychologia* **12**: 419–428.

Zola-Morgan S and Squire LR (1993) The neuroanatomy of amnesia. *Annual Review of Neuroscience* **16**: 547–563.

Further Reading

Cohen NJ and Eichenbaum H (1993) *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT Press.

Mayes AR (1988) *Human Organic Memory Disorders*. Cambridge, UK: Cambridge University Press.

Mayes AR and Downes JJ (1997) *Theories of Organic Amnesia*. Hove, UK: Psychology Press.

Schacter DL and Glisky EL (1986) Memory remediation: restoration, alleviation and the acquisition of domain-specific knowledge. In: Uzzell BP and Gross Y (eds) *Clinical Neuropsychology of Intervention*. Dordrecht: Martinus Nijhoff.

Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review* **99**: 195–231.

Squire LR and Knowlton BJ (1995) Memory, hippocampus and brain systems. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Parkin AJ and Leng NRC (1993) *Neuropsychology of the Amnesic Syndrome*. Hove, UK: LEA.

Amygdala

Intermediate article

Ralph Adolphs, University of Iowa, Iowa City, Iowa, USA

CONTENTS

The amygdala is a collection of nuclei in the telencephalon that connect with the sensory neocortex, frontal cortex and a variety of structures that regulate physiological responses. The amygdala participates in the regulation of emotion, attention, memory, and decision-making.

INTRODUCTION

The amygdala is an almond-shaped collection of nuclei which connect with disparate brain structures. In mammals, it is located in the anterior medial temporal lobe, just rostral to the hippocampal formation. The amygdala has a role in motivational, emotional, and cognitive processes that involve information of biological value to the organism. This function has been dissected at the finest grain in animal studies, whereas studies in humans have typically focused on the amygdala's contribution to social cognition.

NEUROANATOMY AND CONNECTIVITY

The amygdala's connections with other brain structures are enormously diverse: it connects with sensory and association neocortex, with hypothalamus, brainstem nuclei, basal forebrain, ventral striatum, thalamus, and hippocampus, and with other structures (Amaral *et al*., 1992). This broad architecture permits the amygdala to link sensory representations of stimuli (in sensory and association neocortex) with modulation of both body state (via hypothalamus and brainstem) and cognition (via frontal cortex, basal forebrain, hippocampus, and other structures). The amygdala's complex connectivity with other brain regions is complemented by an equally complex internal connectivity among its various component nuclei.

Sensory Input to the Amygdala

The amygdala receives sensory information from all sensory modalities. Signals directly from the olfactory bulb are a prominent form of input in mammals other than primates. Projections from brainstem nuclei and insular cortex provide somatic visceral information, and projections from thalamus and temporal cortex provide auditory and visual information.

In primates, the visual inputs to the amygdala have been best studied. Visually responsive high-level neocortices in the anterior temporal lobe project to the lateral amygdala, and provide highly processed visual information. The amygdala in turn projects back to all those regions from which it receives inputs (via its basal nucleus), and also projects back to some earlier sensory cortices from which it does not receive direct inputs. In fact, the amygdala provides nonreciprocal feedback to all earlier stages of visual processing in the ventral visual stream (which subserves object recognition), including the striate cortex (Figure 1). The feedback from the amygdala terminates in layers 1 and 2 of the cortex and provides an interesting architecture by which the amygdala could, in principle, modulate perception even at the earliest stages of processing.

Connectivity with Brainstem and Hypothalamus

The amygdala connects with nuclei in the brainstem and hypothalamus via its central nucleus. These targets permit the amygdala to modulate the organism's somatic state, a function that constitutes an important component of emotion. For instance, emotionally arousing stimuli are known to trigger increases in sympathetic autonomic tone. The amygdala participates in such triggering via its projections to the paraventricular hypothalamus and periaqueductal gray matter, structures that in turn contain neurons that project to the intermediolateral cell column in the spinal cord, constituting the preganglionic sympathetic neurons. It is thus possible in many cases to trace a causal chain from

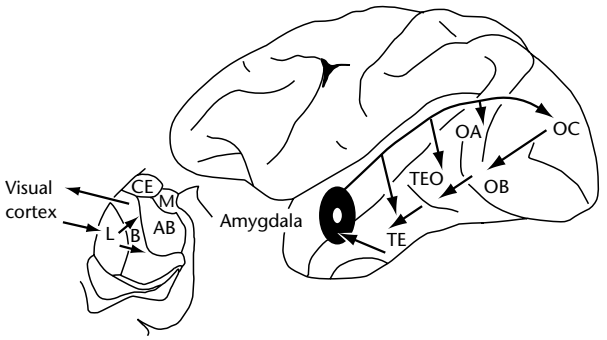


Figure 1. Connections of the amygdala with sensory neocortex. In the monkey, visual inputs to the amygdala originate in high-level visual cortices in the anterior temporal lobe. Feedback from the amygdala is both reciprocal and nonreciprocal, providing projections to all levels of visual processing, including V1. This architecture in principle permits the amygdala to participate in perceptual processing at even the earliest cortical stages. Inputs to, and outputs from, the amygdala rely on the lateral and the basal nucleus, respectively (inset at left). Reproduced with permission from Amaral *et al*. (1992), copyright John Wiley. **Key**: AB, accessory basal; B, basal; CE, central; L, lateral; M, medial; OA, OB, OC, sectors of occipital cortex; TE, TEO, sectors of temporal cortex.

the central nucleus of the amygdala to various effector structures in the brainstem and hypothalamus, all of which regulate autonomic, visceral, and endocrine states (Figure 2).

As with the sensory information, the amygdala's effector communications are not one-way. The amygdala receives information about the somatic changes it effects, via multiple channels. In addition to receiving visceral somatic information from the insular cortex, it also receives more direct information from brainstem nuclei such as the parabrachial nucleus and the nucleus of the solitary tract, as well as direct input from the spinal cord. The vagus nerve is an important route by which body-state information is conveyed to the brainstem and then to the amygdala.

Connectivity with Structures Mediating Cognitive Processes

The above two sets of connections – with sensory cortices on the one hand, and with autonomic and visceral effector structures on the other – would suffice to permit the amygdala to trigger emotional body-state changes in response to sensory stimuli. While this is certainly one function of the amygdala, it is only a small part of the whole story. Perception of emotionally salient stimuli triggers changes not only in the functioning of an animal's body, but also in the animal's cognition. Attention, memory, decision-making and other processes are all prominently modulated by emotion and motivation, and all involve the amygdala.

The amygdala's participation is by virtue of its connections with many other brain regions. For instance, the amygdala projects to cholinergic nuclei in the basal forebrain, which modulate attention; to the hippocampal formation, which is involved in memory; and to the prefrontal cortex, which is involved in planning, reasoning, and decision-making. All these structures, and all the cognitive processes they mediate, can be influenced by the amygdala. As a consequence, emotionally salient information from a sensory stimulus will lead to changes not only in an animal's body, but also in the way that its brain processes that information.

To put the amygdala's role in full perspective, we need to add yet another consideration. The amygdala links sensory percepts with both bodily and cognitive changes, but it does not do so on its own. Rather, it does so in conjunction with other structures that have similar roles. In particular, the amygdala functions in tandem with the orbitofrontal cortex and the ventral striatum, both of which are intimately involved in motivation and emotion.

Intrinsic Processing of the Amygdala

Inputs to the amygdala from the sensory neocortex terminate in the lateral nucleus. Outputs from the amygdala to brainstem and hypothalamus originate in the central nucleus, and outputs to many other structures, such as ventral striatum, cingulate cortex and orbitofrontal cortex, issue from the lateral and basal nuclei, among others. Connections between the lateral nucleus and the basal, central and other nuclei mediate considerable internal information processing (Pitkanen *et al*., 1997). Put simply, there is a flow of information from the lateral to the central parts of the amygdala, either directly or via intermediate nuclei. Lesions of the lateral nucleus can therefore impair all aspects of an animal's conditioned response to a stimulus, whereas damage to other nuclei impairs only specific components of such a conditioned response (Amorapanth *et al*., 2000).

AMYGDALA FUNCTION IN ANIMALS

The amygdala has been studied extensively in rats, cats and primates, with some convergent findings. Ever since the classical studies of Klüver and Bucy in the 1930s, the amygdala has been implicated in emotional and social behavior. This rather vague function has now been dissected by studies

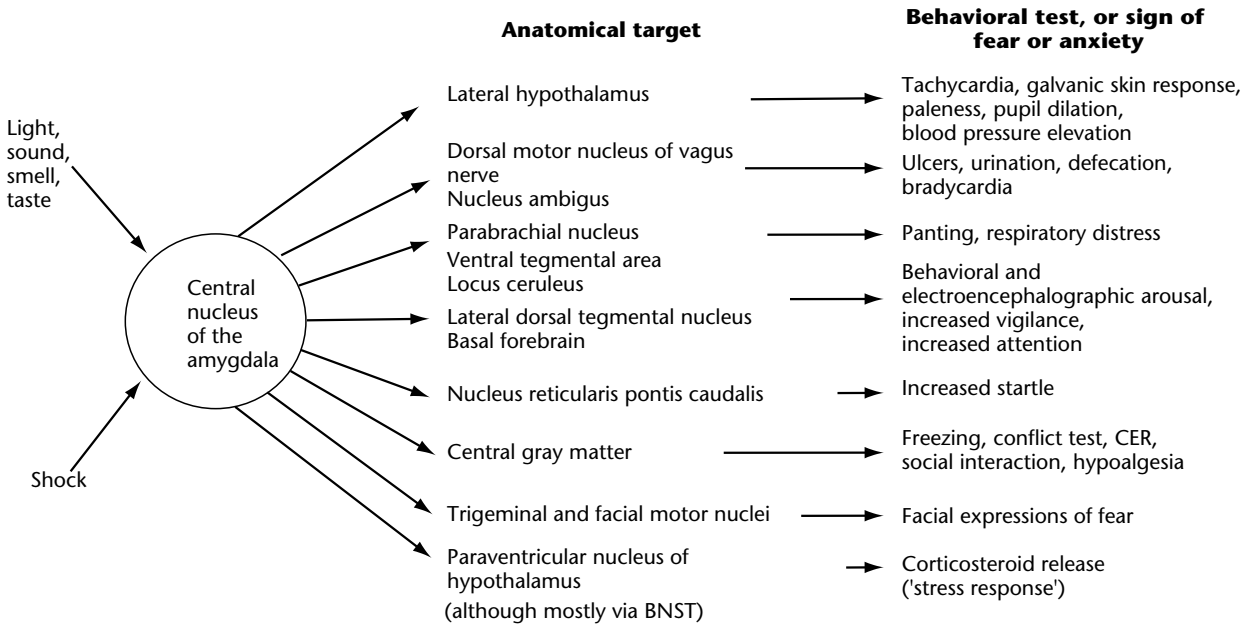


Figure 2. Outputs of the amygdala that mediate fear conditioning. Conditioned and unconditioned stimuli are associated in the lateral amygdala, which in turn projects to the central nucleus of the amygdala. The central nucleus can effect autonomic, endocrine, and visceromotor fear responses by virtue of its connections with structures in brainstem and hypothalamus. Reproduced with permission from Davis (1992), copyright John Wiley.

demonstrating a more specific role in associating sensory stimuli with their rewarding or punishing contingencies. Different nuclei within the amygdala contribute to different aspects of this function. It remains an open question whether the amygdala's basic function is related to reward and punishment in general, or whether it is more specifically related to processing the significance of social stimuli, although findings in animals are more consistent with the former possibility than with the latter.

Regulation of Physiological Responses

Consistent with its connections to hypothalamus and brainstem nuclei, the amygdala participates in regulating a wide array of physiological responses (Figure 2). Direct evidence for this role comes from experiments in which electrical stimulation of the amygdala by implanted electrodes produced changes in physiological measures such as blood pressure and heart rate, corroborating the amygdala's role as an autonomic control structure, and consistent with its known connectivity.

Fear Conditioning

The topic that has seen the most intense research in animal studies is fear conditioning. In fear conditioning an animal is presented with two different kinds of stimuli: an unconditioned stimulus (US) that is highly aversive, such as an electric shock, and a conditioned stimulus (CS) that is neutral to begin with, such as a light. At the beginning of the experiment the animal shows a physiological reaction to the electric shock, but not to the light. During the conditioning experiment, the electric shock is paired with the light stimulus. After several such CS–US pairings the animal learns that the light predicts the shock, a form of associative emotional memory. The final stage of the experiment is to present the CS without the US. The fear-conditioned animal will behave as though it is afraid of the light, a conditioned response that can be measured in a number of ways as indicated in the right-hand column in Figure 2. (*See* **Conditioning**)

Conditioned fear responses critically depend on the amygdala, and cells within the lateral amygdala increase their firing rates and the synchrony of their discharges following the CS. Importantly, the dependency of the acquisition and expression of fear-conditioned responses on the basolateral amygdala is not attributable solely to the amygdala's role in the behaviors used to assess fear conditioning, such as 'freezing' (Maren, 1999).

Current debates

Despite the popularity of the fear-conditioning paradigm, there are many issues still to resolve. Is

the amygdala essential for all components of fear conditioning? Is the memory of the US–CS association actually stored in the amygdala? Affirmative answers to these questions have been given by some, but contested by others. One general issue is whether the amygdala is essential to the acquisition of fear-conditioned responses, or whether its function is better thought of as modulatory and important rather than essential (e.g. Wilensky *et al*., 2000). Another issue concerns the role of fear conditioning during development, and its relation to emotional and social development. Fear conditioning is not seen at birth in all species; possibly its absence early in life permits an animal to form an attachment to its parent regardless of the circumstances (that is, even when the presence of the parent occurs together with aversive events).

Reconsolidation

If the amygdala is indeed the repository for the associations between conditioned and unconditioned stimuli in fear-conditioning experiments, it remains an open question precisely how such storage is implemented. One possibility is that a relatively permanent association is stored in a way that is insensitive to future perturbations. This appears unlikely at first, because it is known that fear-conditioned responses can be extinguished if the CS is presented for many times without the US. However, this extinction appears to rely not on an erasure of the memory trace within the amygdala, but rather on active inhibition of the amygdala's fear memory by the prefrontal cortex. It thus appears that some components of the fear memory may indeed be stored in the amygdala permanently, an aspect that may have implications for indelible fear responses in humans, for instance in posttraumatic stress disorder and phobias.

Although the memory of the fear conditioning in the amygdala may be relatively permanent, it is not static. Experimenters have found that reactivating the fear memory by presenting the CS rendered the memory trace plastic and susceptible to alteration. After reactivation of the memory, there was a period of 'reconsolidation' during which the memory could indeed be erased. Injections of drugs that inhibited protein synthesis, such as anisomycin, were found to erase the fear-conditioned memory, but only when given during such a period of reconsolidation (Nader *et al*., 2000). This finding may have broad applicability to memory in general, including memory in humans, and emphasizes the dynamic nature of memory representations in the brain. (*See* **Memory Consolidation**)

Molecular mechanisms

What are the molecular mechanisms whereby the amygdala stores the associations learned during fear conditioning? There is evidence that such associative memory may rely on long-term potentiation (LTP) within the amygdala. The acquisition of fear-conditioned responses requires within the amygdala some of the molecular machinery known to be involved in the induction and maintenance of LTP, such as protein kinase A, protein synthesis, and perhaps even NMDA (*N*-methyl-D-aspastate) receptors (although this has not been conclusively shown). It thus appears likely that at least some components of the amygdala's role in associative memory are similar at the molecular level to those required for memory consolidation in the hippocampus.

Other Forms of Conditioning

The amygdala has also been implicated in other forms of conditioning besides the classical (Pavlovian) fear response. Conditioning to stimuli that are rewarding rather than aversive does not appear to rely on the amygdala to the same extent that conditioning to aversive stimuli does. However, the amygdala has been found to be important in reinforcer devaluation effects, in which the value of a rewarding US is changed: for instance, by using food reward as the US and then later feeding the animal to satiation on that food so that the US no longer has the same motivational value that it did at the time of conditioning. Normal animals are sensitive to such manipulations, but animals with amygdala damage are not. These studies suggest that a comprehensive function of the amygdala in associative emotional memory is the provision of access to current values of reinforcing stimuli, a function in which it participates intimately with the orbitofrontal cortex (Baxter *et al*., 2000).

Modulation of Motivated Learning

Aversively motivated learning can be modulated by various neurotransmitter systems acting within the amygdala (Cahill and McGaugh, 1996; McGaugh, 2000). For instance, injection into the amygdala of one of a variety of drugs that modulate neurotransmission mediated by γ-aminobutyric acid (GABA), noradrenaline, or opiates immediately after training influences long-term memory for stimulus avoidance. Research suggests that the amygdala can influence such memory by modulating consolidation that actually takes place within other brain structures, such as the hippocampus and basal ganglia.

The amygdala thus participates in at least two distinct types of memory function: associative emotional memory (such as fear conditioning), which may be stored within the amygdala, and modulation of memory (including declarative memory in humans) that is itself dependent on other brain structures.

Influences on Attention

The amygdala's role in attentional processes has been investigated in a number of experiments (Holland and Gallagher, 1999). One component of attention, orienting behavior towards cues that have become associated with rewarding contingencies, has been found to rely on a circuit involving the central nucleus of the amygdala and its connections with the substantia nigra and the dorsal striatum. Another important component of attention, increased allocation of processing resources toward novel or surprising situations, appears to depend on the integrity of the central nucleus of the amygdala and its connections with cholinergic neurons in the substantia innominata and nucleus basalis, structures in the basal forebrain. Thus, the amygdala could modulate cholinergic neuromodulatory functions of the basal forebrain nuclei, and consequently modulate attention, vigilance, signal-to-noise and other aspects of information processing that depend on cholinergic modulation of cognition. Through circuits including components of amygdala, striatum, and basal forebrain, emotion may thus help to select particular aspects of the stimulus environment for disproportionate allocation of cognitive processing resources; in other words, the organism preferentially processes information about its environment that is most salient to its immediate survival and wellbeing. Recent studies in humans have demonstrated such a role (Anderson and Phelps, 2001).

Social Behavior

Although early experiments concerning the effects of amygdala lesions highlighted impairments in social behavior, this issue has received scant attention. In general, the findings of the effects of focal amygdala lesions have corroborated impairments in social behavior – typically an increase in tameness and placidity towards others (Weiskrantz, 1956; Emery *et al*., 2001). However, how such an impairment plays out in a group of animals depends on the species and on the environmental circumstances in which the social behavior is assessed: monkeys whose amygdala has been removed can be the subjects of either increased affiliative or increased aggressive behavior from other monkeys in the troop. It remains an open question to what extent the deficits in social behavior that are seen following amygdala damage could be understood in terms of (or could be reducible to) impairments in more basic processing of reward and punishment.

AMYGDALA FUNCTION IN HUMANS

In contrast to animal studies, which implicate the amygdala in processing reward and punishment in a general sense, studies with humans point to a more restricted function in processing socially relevant stimuli. Furthermore, findings in humans indicate a role of disproportionate importance in processing stimuli related to negatively valenced stimuli that signal threat, ambiguity or distress. However, the apparent differences between animals and humans may be due more to the kinds of tasks and stimuli employed than to actual differences in amygdala function.

Conditioning

Damage to the human amygdala impairs the ability to acquire conditioned fear responses. One study showed that a patient with bilateral damage to the amygdala, but with an intact hippocampal memory system, was unable to acquire conditioned autonomic responses, but was nonetheless able to acquire declarative memory about the CS–US pairings: in essence, the patient could tell the experimenter that the light had been paired with the shock – so the patient knew, as a declarative fact, that light predicted shock – even though there was no fear conditioning (Bechara *et al*., 1995). Functional imaging studies have likewise found evidence that the amygdala is activated in normal people during fear conditioning.

Recognition of Emotion

Both lesion and functional imaging studies of the brain have confirmed a role for the amygdala in perception and recognition of emotional stimuli. The evidence is strongest in the case of recognition of emotions from facial expressions, although there is some evidence for emotion perception from auditory, olfactory, and gustatory stimuli as well. A rare patient with bilateral amygdala lesions was found to have an impaired ability to recognize fear from faces, a finding subsequently replicated in several other studies, and corroborated by the finding that

viewing faces showing expressions of fear activates the amygdala in normal people.

Despite these studies, it remains unclear precisely which emotions the amygdala is most involved in processing. While fear recognition is notably impaired in some people with amygdala damage, this is not always the case, and others with similar brain lesions are more impaired on recognition of anger, sadness or disgust (Adolphs *et al*., 1999). Functional brain imaging of psychiatric patients has shown that some people with phobias or similar disorders have abnormal activation of the amygdala (for example, an abnormally high activation of the amygdala occurs in response to 'neutral' faces, in patients who are socially phobic). All these findings have led to proposals that the amygdala processes predominantly information about threat, about ambiguity or about distress. At this stage, all we can say for sure is that the amygdala participates in processing knowledge of emotion from a variety of social stimuli, such as facial expressions – but whether there is specificity regarding particular emotions, and what that specificity consists of, are issues requiring further research. Some functional imaging studies have found activation of the amygdala by pleasant stimuli also.

Another important point is that the amygdala does not appear essential for all knowledge regarding emotions. In particular, knowledge of emotions from explicit lexical stimuli is unaffected by amygdala damage. People with bilateral amygdala damage are quite able to use words such as 'fear' appropriately in conversation, and possess considerable knowledge about the kinds of situation that would normally make people feel afraid, or the kinds of behavior elicited by fear. However, the concept of fear that people with bilateral amygdala damage have is not entirely normal: it has gaps. It seems likely that those gaps are most prominent for emotional knowledge that cannot be explicitly encoded into language – such as knowing what a fearful face looks like.

Functional imaging studies have confirmed that the amygdala is activated by emotional facial expressions, especially ones showing fear. Such activation has been observed even when the face stimuli were presented so briefly that they could not be consciously recognized (Whalen *et al*., 1998) There is evidence to suggest a differential role for the right amygdala in processing emotional visual stimuli below the level of conscious awareness, and for the left amygdala in processing emotional stimuli that are consciously perceived (Morris *et al*., 1999).

It is less clear to what extent the amygdala is involved in recognition of emotional auditory stimuli; some lesion and functional imaging studies have suggested such a role, whereas others have not. The most robust activation of the amygdala to sensory input with emotional significance has been observed with olfactory stimuli, where unpleasant smells reliably activate the amygdala (Royet *et al*., 2000).

Emotional Memory

As in animals, the amygdala in humans modulates memory for emotionally arousing stimuli. In humans, this has been studied in regard to declarative memory. When normal people are shown emotionally arousing pictures or film, they remember best those stimuli that they found the most emotionally arousing. Functional imaging studies have shown that activation of the amygdala at the time of encoding (when the stimuli were shown, while the subject was in the scanner) correlated with how well the person later remembered those same stimuli when questioned several weeks later (Canli *et al*., 2000). Consonant with these findings, people with bilateral amygdala damage were less able to remember emotionally arousing pictures – they remembered them only as well as neutral pictures. Electrical stimulation of the vagus nerve – which provides visceral body-state information to brainstem nuclei and the amygdala – has been shown to enhance memory, consistent with the idea that emotional body states can modulate cognitive processes such as memory, at least in part via the amygdala.

Social Cognition

The human amygdala plays a more general role in social cognition as well. In one study, participants were asked to judge how trustworthy or approachable they found other people; those with bilateral amygdala damage gave abnormally positive ratings (Adolphs *et al*., 1998). Other studies have found activation of the amygdala in response to viewing faces of people of another race: the amygdala habituated more rapidly when viewing faces of one's own race than of another race, and amygdala activation to unfamiliar faces of a different race correlated with implicit measures of race evaluation (Phelps *et al*., 2000). Such racial outgroup responses may fit into the general scheme of threat detection and vigilance by the amygdala. Finally, there is evidence linking amygdala pathology to autism (Baron-Cohen *et al*., 2000): some

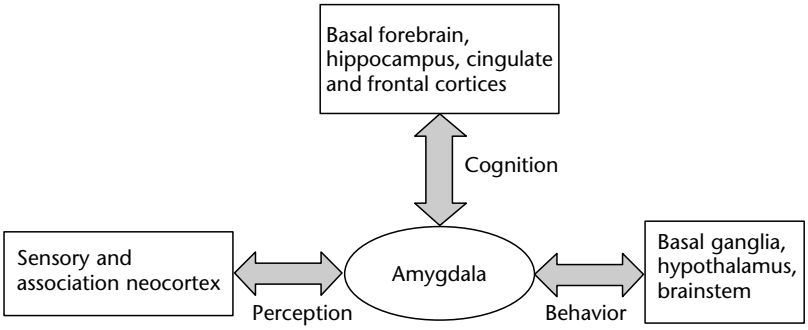


Figure 3. Connectivity of the amygdala at the system level. The amygdala's participation in multiple effector mechanisms permits mediation of a concerted, integrated emotional response in both body and brain.

structural and functional imaging studies suggest such a link, and people with autism are impaired on some of the same tasks as people with bilateral amygdala damage. Clearly, the amygdala is important in social cognition.

A SYSTEMS-LEVEL VIEW OF THE AMYGDALA

While it is clear that the amygdala broadly participates in emotion and social behavior, the mechanisms that underlie this function remain poorly understood. In general terms, the amygdala can participate in emotion in three different ways (Figure 3). First, it can link perception of stimuli to an emotional response, by virtue of its inputs from sensory cortices and its outputs to control structures such as the hypothalamus, brainstem nuclei, and periaqueductal gray matter; second, it can link perception of stimuli to modulation of cognition, by virtue of its connections with structures involved in decision-making, memory, and attention; and third, it can link early perceptual processing of stimuli with modulation of such perception via direct feedback. The amygdala's participation in all three of these mechanisms enables it to contribute globally to affective processing in a concerted manner, by modulating numerous processes simultaneously.

To which aspects of emotion could such mechanisms contribute? Certainly, the amygdala contributes to emotional responses to stimuli. It also contributes to the recognition of emotions, but this function is likely to be indirect: possibly the amygdala's modulation of perception by feedback, or its elicitation of components of an emotional response, can contribute to the brain's ability to reconstruct knowledge about the emotion signaled by a stimulus. Finally, the amygdala's role in feeling an emotion (in the conscious experience of emotion) is unclear, and may well be inessential.

References

Adolphs R, Tranel D and Damasio AR (1998) The human amygdala in social judgment. *Nature* **393**: 470–474.

Adolphs R, Tranel D, Hamann S *et al*. (1999) Recognition of facial emotion in nine subjects with bilateral amygdala damage. *Neuropsychologia* **37**: 1111–1117.

Amaral DG, Price JL, Pitkanen A and Carmichael ST (1992) Anatomical organization of the primate amygdaloid complex. In: Aggleton JP (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pp. 1–66. New York, NY: Wiley-Liss.

Amorapanth P, LeDoux JE and Nader K (2000) Different lateral amygdala outputs mediate reactions and actions elicited by a fear-arousing stimulus. *Nature Neuroscience* **3**: 74–79.

Anderson AK and Phelps EA (2001) Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* **411**: 305–309.

Baron-Cohen S, Ring HA, Bullmore ET *et al*. (2000) The amygdala theory of autism. *Neuroscience and Biobehavioral Reviews* **24**: 355–364.

Baxter MG, Parker A, Lindner CCC, Izquierdo AD and Murray EA (2000) Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *Journal of Neuroscience* **20**: 4311–4319.

Bechara A, Tranel D, Damasio H *et al*. (1995) Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* **269**: 1115–1118.

Cahill L and McGaugh JL (1996) Modulation of memory storage. *Current Opinion in Neurobiology* **6**: 237–242.

Canli T, Zhao Z, Brewer J, Gabrieli JDE and Cahill L (2000) Event-related activation in the human amygdala associates with later memory for individual emotional experience. *Journal of Neuroscience* **20**: RC99 (91–95).

Emery NJ, Capitanio JP, Mason WA *et al*. (2001) The effects of bilateral lesions of the amygdala on dyadic

social interactions in rhesus monkeys. *Behavioral Neuroscience* **115**: 515–544.
Holland PC and Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* **3**: 65–73.
Maren S (1999) Neurotoxic basolateral amygdala lesions impair learning and memory but not the performance of conditional fear in rats. *Journal of Neuroscience* **19**: 8696–8703.
McGaugh JL (2000) Memory – a century of consolidation. *Science* **287**: 248–251.
Morris JS, Ohman A and Dolan RJ (1999) A subcortical pathway to the right amygdala mediating 'unseen' fear. *Proceedings of the National Academy of Sciences of the USA* **96**: 1680–1685.
Nader K, Schafe GE and LeDoux JE (2000) Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**: 722–726.
Phelps EA, O'Connor KJ, Cunningham WA *et al*. (2000) Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience* **12**: 729–738.
Pitkanen A, Savander V and LeDoux JE (1997) Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. *Trends in Neurosciences* **20**: 517–523.
Royet JP, Zald D, Versace R *et al*. (2000) Emotional responses to pleasant and unpleasant olfactory, visual, and auditory stimuli: a positron emission tomography study. *Journal of Neuroscience* **20**: 7752–7759.
Weiskrantz L (1956) Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative Physiology and Psychology* **49**: 381–391.
Whalen PJ, Rauch SL, Etcoff NL *et al*. (1998) Masked presentations of emotional facial expressions modulate amygdala, activity without explicit knowledge. *Journal of Neuroscience* **18**: 411–418.
Wilensky AE, Schafe GE and LeDoux J (2000) The amygdala modulates memory consolidation of fear-motivated inhibitory avoidance learning but not classical fear conditioning. *Journal of Neuroscience* **20**: 7059–7066.

Further Reading

Adolphs R (1999) The human amygdala and emotion. *Neuroscientist* **5**: 125–137.
Aggleton J (ed.) (2000) *The Amygdala: A Functional Analysis*. New York, NY: Oxford University Press.
Davis M (1997) Neurobiology of fear responses: the role of the amygdala. *Journal of Neuropsychiatry and Clinical Neurosciences* **9**: 382–402.
Emery NJ and Amaral DG (1999) The role of the amygdala in primate social cognition. In: Lane RD and Nadel L (eds) *Cognitive Neuroscience of Emotion*, pp. 156–191. Oxford, UK: Oxford University Press.
LeDoux J (1996) *The Emotional Brain*. New York, NY: Simon & Schuster.
Rolls ET (1999) *The Brain and Emotion*. New York, NY: Oxford University Press.

Anosognosia

Introductory article

Alfred W Kaszniak, University of Arizona, Tucson, Arizona, USA

CONTENTS

Anosognosia is a clinical syndrome in which a patient with brain dysfunction appears unaware of a neurological or neuropsychological impairment, such as paralysis or amnesia.

INTRODUCTION

In 1914, Joseph Babinksi described several patients who were paralyzed on the left side of their body due to damage in the right hemisphere of the brain. Most startling was his observation that these patients appeared to have no awareness of their paralysis and never complained about their impairment. Although others had made similar observations in the late 1800s and early 1900s, it was Babinski who coined the term 'anosognosia' to describe this syndrome. Subsequently, the term has been extended to encompass similar unawareness of deficit phenomena in persons suffering from a variety of neurological impairments. Anosognosia has been documented for deficits as diverse as blindness, impaired recognition of familiar faces, hemiplegia (paralysis of one side of the body), and aphasia (acquired disorder of language comprehension and expression).

Historically, there has been debate about how best to conceptualize anosognosia. In their influential book entitled *Denial of Illness*, published in 1955, neurologist Edwin Weinstein and psychologist Robert Kahn argued that anosognosia is a motivated unawareness, reflecting the operation of psychodynamic defense mechanisms that block symptoms or deficits from awareness. Although this conceptualization is intuitively appealing, neurologist and neuroscientist V. S. Ramachandran has recently noted that the psychodynamic interpretation does not account for two important aspects of anosognosia. First, anosognosia typically occurs only when there is damage to particular brain structures. These structures include lower portions of the parietal cortex, frontal lobes, and the subcortical structures connecting to these structures, and anosognosia is seen predominantly following right, rather than in left hemisphere brain damage. Anosognosia is typically not observed when peripheral, spinal, lower brain stem, or cortical primary sensory or motor areas alone are affected. Second, the unawareness is often quite specific. For example, a patient may deny his/her hemiplegia but readily admit to other disabling or distressing symptoms. Occasionally, persons with hemiplegia of both their upper and lower extremities may admit that their leg is paralyzed but insist that their arm is not. In contrast to earlier psychodynamic conceptualizations, most current theories of anosognosia adopt a neuropsychological explanation, attributing the impaired awareness of deficit to dysfunction in brain circuits important for self-monitoring. In a frequently cited review paper published in 1989, neuropsychologists Susan McGlynn and Daniel Schacter emphasized the specificity and dissociations (i.e. intact awareness for some impairments and impaired awareness for other impairments) found in different forms of anosognosia in individuals with focal neurological damage. Based on their review of clinical-pathological correlation studies, they concluded that damage to right parietal and/or frontal brain regions is of particular importance in causing anosognosia.

CLINICAL STUDIES OF ANOSOGNOSIA FOR AMNESIA AND DEMENTIA

Some of the most informative recent research for understanding the neuropsychology of anosognosia comes from studies of amnesia and dementia. Amnesia refers to an acquired loss of memory. Although amnesia may be due to psychological factors, such as intensely frightening personal experiences, it is most often due to neurological causes. In 1889, the physician S. S. Korsakoff described the amnesic disorder that is now called Korsakoff's syndrome and observed that most of his patients had little apparent awareness of their

memory deficits. Korsakoff's syndrome most typically occurs in persons with a history of chronic alcohol abuse and thiamine deficiency, with postmortem examinations of the brain revealing damage in the dorsomedial nuclei of the thalamus, the mammillary bodies, and other nearby brain structures. Computerized tomographic (CT) and magnetic resonance imaging (MRI) studies have confirmed the presence of thalamus and mamillary body damage in persons with Korsakoff's syndrome and have also revealed damage to the orbitofrontal and mediotemporal areas of the cerebral cortex. In addition to severely impaired memory (thought to be associated with the thalamic, mamillary body, and mediotemporal cortex damage), persons with Korsakoff's syndrome also show other cognitive deficits, such as impairments in shifting and dividing attention, in performing tasks requiring hypothesis generation, testing, and problem solving, and in preserving the temporal order of information. These deficits appear to be associated with the frontal cortical damage.

In contrast to persons with Korsakoff's syndrome, individuals with equally severe memory impairment, due to restricted medial temporal (particularly hippocampus) damage, do not show frontally related cognitive dysfunction and are generally well aware of their memory deficits. Persons with amnesia due to medial temporal damage may spontaneously comment about their memory difficulty and rely upon mnemonic aids such as reminder notes and schedule books. Thus, amnesic patients without frontal lobe damage appear to have intact awareness of their memory deficits, while those with frontal damage (i.e. those with Korsakoff's syndrome) are anosognosic for their amnesia.

The study of persons with missile wound brain injuries has supported the association of impaired awareness of memory deficit with frontal lobe damage. Those who lack awareness of their memory impairment have been found to have damage involving both frontal lobes (among other areas), whereas those with intact awareness of their memory deficits show no evidence of frontal lobe damage. Other persons with amnesic syndromes due to brain damage that includes the frontal lobes (e.g. anterior communicating artery aneurysm rupture, closed head injury) have also been found to be anosognosic for their memory deficits.

Dementia refers to progressive impairment of multiple cognitive functions (including memory) due to acquired brain disease. Clinical accounts of anosognosia in dementia syndromes began appearing in the early 1980s. The most prevalent cause of dementia is Alzheimer disease (AD), a neurodegenerative disorder with neuron loss and other microscopic brain changes that are most prominent in the medial temporal, posterior temporal, parietal, and frontal brain regions. Persons with AD show relatively severe memory impairment early in the course of their dementia, along with milder deficits in other cognitive functions (e.g. perception, language, judgment). These other cognitive functions become increasingly impaired with disease progression. The memory and other cognitive impairments correlate with the pattern of observed neuropathological changes, particularly involving medial temporal and frontal brain regions. Many persons with AD, particularly as their illness progresses, show impaired awareness of their memory and other cognitive deficits.

Impaired awareness of deficits has also been noted in clinical descriptions of other types of dementia besides AD. These include vascular dementia (typically associated with multiple and bilaterally distributed small cerebral strokes), frontotemporal dementia (presenting with impaired impulse control, poor social judgment, and cognitive deficit, and typically involving severe atrophy of frontal and temporal cortices), and Huntington's disease (an autosomal dominant genetic neurodegenerative disorder characterized by motor, cognitive, and emotional dysfunctions associated with caudate nucleus atrophy and frontal lobe changes).

AD and other dementia types involve impairments in a wide range of cognitive, behavioral, emotional, and functional areas. Thus, the study of deficit awareness in dementia provides an opportunity for examining the degree to which impaired awareness is general or limited to particular deficits. Clinical observations have suggested that awareness of deficit can be selective in dementia. For example among persons with AD some deny all cognitive deficits, while others claim that their memory is good but admit to difficulty in finding words, and still others deny memory impairment but admit to reading difficulties.

Some clinical observers have suggested that loss of awareness of deficits occurs earlier in the course of illness for frontotemporal dementia than for AD. Both of these progressive dementia types are typically associated with signs of frontal lobe pathology, but frontal damage is generally more severe in the early stages of frontotemporal dementia than AD. This is consistent with the association between frontal lobe pathology and impaired awareness of deficit that has been observed in amnesic syndromes.

Studies of the neuropsychological correlates of anosognosia for dementia in AD have also supported a relationship to frontal lobe damage. For example, an association has been demonstrated between decreased fluency of speech (as measured in tests requiring the generation of words within particular categories or beginning with particular letters) that is typically seen following left or bilateral frontal cortex damage, and clinical ratings of anosognosia for memory deficit in persons with AD. Similarly, anosognosia for dementia in AD has been shown to be associated with both lower general mental status test performance and specific impairment on measures of 'executive functions' (involving the capacity to plan and carry out complex, goal-oriented behavior) that are sensitive to frontal lobe dysfunction. More direct evidence of a relationship to frontal brain dysfunction is provided by research showing that clinical ratings of anosognosia for memory loss in persons with AD are associated with decreased blood flow (as measured by single photon emission computed tomography) in the right dorsolateral frontal brain region. (*See* **Alzheimer Disease**; **Amnesia**; **Huntington Disease**)

LABORATORY STUDIES OF ANOSOGNOSIA FOR AMNESIA AND DEMENTIA

Although important, clinical observations alone allow for only limited and tentative conclusions concerning both the neurobiological correlates and nature of anosognosia. There are several reasons for the inferential limitations of clinical observations. First, the methods employed in making clinical observations of impaired awareness (e.g. clinical rating scales, patients' responses to interview questions) are generally unsystematic and have varied across investigators. This creates problems for any attempt to compare findings across studies. Further, clinical observations alone do not allow for the development of any articulated theory of anosognisia. One goal of research on anosognosia is to make theoretical inferences concerning human metamemory functions. The term 'metamemory' refers to those processes involved in the conscious monitoring and control of, as well as beliefs about, one's own memory functioning. Theoretically important questions such as whether anosognosia represents inaccurate self-efficacy beliefs (e.g. a person's beliefs about how well their memory generally functions), poor self-monitoring, or some combination of these or other factors, cannot be adequately addressed by purely clinical observations. Recently, systematic laboratory studies have become available, allowing quantitative measurement of different aspects of awareness in amnesia and dementia. Most of these laboratory studies have focused on awareness of deficits in Korsakoff's and other amnesic syndromes or AD.

Laboratory approaches have typically used one of three different methods to study awareness of memory deficit: comparisons of patient self-report and others' ratings of patient disability; evaluation of the accuracy of patients' predictions for their performance on specific cognitive tasks; or feeling-of-knowing paradigms (e.g. confidence ratings or rankings regarding the likelihood that recently learned information or long-term knowledge which the individual failed to recall would later correctly be recognized from among multiple alternatives). These different methods provide information relevant to theoretical formulations concerning the nature of metamemory impairment. Psychologists Christopher Hertzog and Roger Dixon, reviewing metamemory research involving healthy adult participants, have distinguished three aspects of metamemory: (1) knowledge about how memory functions and the utility of different strategies in memory tasks; (2) memory self-monitoring, defined as awareness of the current state of one's own memory system; and (3) self-referent beliefs about memory (memory self-efficacy beliefs). In general, the first two methods described above (patient self-report versus others' ratings; performance prediction accuracy) have provided information relevant primarily to those aspects of metamemory concerning knowledge of how memory functions and memory self-efficacy beliefs. The third method (feeling-of-knowing paradigms) provides information that can be interpreted as more specifically relevant to questions about memory self-monitoring.

Results of studies using these different methods are consistent with the conclusion that anosognosia for amnesia occurs in memory-impaired persons who also have frontal lobe dysfunction. Further, evidence suggests that it is the memory self-monitoring aspect, rather than knowledge of how memory operates, that breaks down in disorders such as Korsakoff's syndrome and AD. However, the question of whether at least some of the relevant research findings may reflect a failure to update self-efficacy beliefs (due to the memory impairment itself), rather than impaired self-monitoring, awaits an answer from future research.

Prospective studies are needed to determine whether premorbid patient characteristics may be related to anosognosia. Further research is also needed to simultaneously compare different

methods of assessing anosognosia, particularly contrasting those approaches relevant to theoretically distinct aspects of metamemory (e.g. memory self-monitoring versus memory self-efficacy beliefs). Finally, there is a need for additional studies concerning the practical implications of anosognosia for amnesia and dementia: for example relationships may exist between anosognosia for dementia and the tendency to engage in potentially risky behavior such as driving. It has been hypothesized that impairment in the ability to recognize cognitive and behavioral limitations may play a role in both driving and other risky behavior among persons with AD. Given the practical implications, research designed to systematically test this hypothesis is of high priority.

Further Reading

Duke LM and Kaszniak AW (2000) Executive control functions in degenerative dementias: a comparative review. *Neuropsychology Review* **10**: 75–99.

Hertzog C and Dixon RA (1994) Metacognitive development in adulthood and old age. In: Metcalfe J and Shimamura AP (eds) *Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press.

Kaszniak AW and Zak MG (1996) On the neuropsychology of metamemory: contributions from the study of amnesia and dementia. *Learning and Individual Differences* **8**: 355–381.

McGlynn SM and Schacter DL (1989) Unawareness of deficits in neuropsychological syndromes. *Journal of Clinical and Experimental Neuropsychology* **11**: 143–205.

Prigatano GP and Schacter DL (eds) (1991) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. New York, NY: Oxford University Press.

Ramachandran VS (1995) Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* **4**: 22–51.

Shimamura AP (1994) The neuropsychology of metacognition. In: Metcalfe J and Shimamura AP (eds) *Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press.

Weinstein EA and Kahn RL (1955) *Denial of Illness: Symbolic and Physiologic Aspects*. Springfield: Charles C. Thomas.

Aphasia

Intermediate article

Argye Elizabeth Hillis, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
Alfonso Caramazza, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

Aphasia is impairment of language caused by brain damage. The term may be restricted to acquired disorders of language caused by focal brain lesions such as stroke, or more broadly applied to developmental language disorders or impairments due to diffuse brain damage such as head trauma or dementia.

INTRODUCTION

The study of language disturbance resulting from brain lesions has fascinated researchers over the years. In 1836, Dax reported one of the first cases of aphasia and proposed that the disorder reflected damage to a language center in the left frontal lobe. Nearly thirty years later, Paul Broca made a similar proposal on the basis of several patients with minimal speech output who were found to have lesions in the left, posterior frontal lobe at autopsy. That report, published in 1861, set the stage in the following decades for the theories of Wernicke and Lichtheim, which postulated the existence of various 'brain centers' that were critical for separate aspects of language. However, over subsequent decades the Wernicke–Lichtheim proposal of aphasia classification based on damage to specific brain centers was severely criticized by Freud, Marie, Goldstein, Luria and others. At the same time, competing classification schemes were being proposed (Table 1). The resulting confusion concerning the relationship between brain damage and aphasia led to a virtual abandonment of aphasia research until 1965, when Norman Geschwind revived and expanded the Wernicke–Lichtheim theory. In this scheme, three distinct domains of language were said to be mediated by specific regions of the left cortex: fluency of speech production mediated by Broca's area (posterior-inferior frontal lobe); comprehension of language mediated by Wernicke's area (posterior superior temporal gyrus); and repetition mediated by the arcuate fasciculus – a white-matter tract between Broca's area and Wernicke's area. The relative sparing or impairment of these domains formed the basis of the 'classical aphasia' syndromes: Wernicke, Broca, conduction, transcortical sensory, transcortical motor, and mixed transcortical aphasias. (Note that a similar effort to revive this same classification scheme had been made by Nielsen in 1936, but with far less impact.) These aphasia syndromes, identified in chronic stroke patients, roughly coincide with distinct cortical lesions on radionuclide brain scans or computed tomographic head scans. Perhaps because of these early claims of a correspondence between the site of brain damage and the resulting clinical aphasia syndrome, Geschwind's proposal gained credibility. It has since formed the basis for much of the current conceptualizations of aphasia in neurology, speech–language pathology and neuropsychology. Although it is not evident that Geschwind's classification scheme is any more valid than other proposed classification schemes, it merits description because of its historical influence.

CLASSICAL APHASIA SYNDROMES

Wernicke Aphasia

Wernicke aphasia is characterized by prominent impairment in the understanding of spoken words and sentences, and effortless production of utterances with the basic structure and melody of sentences but mostly devoid of clear meaning. Speech

Table 1. Proposed aphasia classifications: syndrome clusters comparable to Geschwind's 1965 classification

Other classification systems (year)	*Geschwind classification*							
	Broca aphasia	*Wernicke aphasia*	*Conduction aphasia*	*Global aphasia*	*TCM*	*TCS*	*MTC*	*Anomic aphasia*
Broca (1865)	Aphemia	Verbal						Amnesia
Wernicke (1881)	Cortical	Cortical	Conduction	Total	TCM	TCS		
Lichtheim (1885)	Motor	Sensory						
Head (1926)	Verbal	Syntactic				Nominal		Semantic
Pick (1931)	Expressive	Impressive		Total				Amnestic
Weisenberg and McBride (1935)	Expressive	Receptive		Expressive/ receptive				Amnestic
Kleist (1934)	Word muteness	Word deafness	Repetition					Amnestic
Neilsen (1936)	Broca	Wernicke	Conduction	Global	TCM	TCS		Amnestic
Wepman (1964)	Syntactic	Jargon				Pragmatic		Semantic
Luria (1966)	Efferent motor	Sensory	Afferent motor		Dynamic	Acoustic		Semantic

MTC, mixed transcortical; TCM, transcortical motor; TCS, transcortical sensory.

in Wernicke aphasia is described as 'empty', consisting of strings of real words in apparently meaningless combinations ('yeah, that was the pumpkin furthest from my thanks') or phrases replete with word-like neologisms ('the scroolish prastimer ate my spanstakes'). Repetition resembles spoken output – mostly fluent jargon. Naming is generally poor. The syndrome has been ascribed to lesions of the posterior-superior temporal cortex (Wernicke's area), often along with the nearby parietal cortex.

Broca Aphasia

The hallmark of Broca aphasia is nonfluent, 'agrammatic' speech output with relatively preserved comprehension ability. Spoken output is characterized by the omission of grammatical morphemes, such as prepositions, conjunctions and verb inflections. To illustrate, a patient with Broca aphasia described her recent onset of neurological impairments in this way: 'Stroke…Sunday…arm, talking – bad.' Often there is concomitant effortful production of words with distorted articulation. Repetition of sentences mirrors spontaneous speech – it is effortful and telegraphic, but the gist of the sentence is mostly retained. Naming is generally impaired, with verbs named more poorly than nouns. Comprehension of single words is mostly spared, but comprehension of syntactically complex sentences is often impaired (Caramazza and Zurif, 1976). The full syndrome is typically ascribed to damage to Broca's area, along with adjacent frontal fields, and the underlying white matter and basal ganglia. Damage to Broca's area, or the underlying white matter, alone causes selective impairment in planning and executing the complex movements to articulate words, a disorder known as 'aphemia'. Often, cases of Broca aphasia evolve from an initial global aphasia, with severely impaired comprehension as well as speech production. In at least some cases, this 'evolution' may be due to initial poor blood flow to Wernicke's area posterior to the stroke, causing impaired comprehension. If blood flow is restored to that area, comprehension improves (Figure 1).

Global Aphasia

The combination of severely impaired comprehension of language and extremely limited speech output is known as global aphasia. Spontaneous speech and repetition are limited to a few perseverative words or nonword utterances (e.g. 'dee dee dee'). Cursing or the production of frequently used phrases, such as 'I don't know' or 'How do you do?', may be spared in this and other aphasia types. Global aphasia is usually ascribed to large lesions involving both Broca's and Wernicke's areas, and the basal ganglia, but can also result from two lesions – one anterior and one posterior.

Transcortical Aphasia

In transcortical aphasias the fluency and content of sentence repetition far exceed those of spontaneous speech. In transcortical sensory aphasia, comprehension of language and content of speech are markedly impaired (as in Wernicke aphasia) but sentence repetition is relatively accurate. A lesion posterior to Wernicke's area, near the temporooccipital junction, is frequently implicated. In transcortical motor aphasia, comprehension and content of

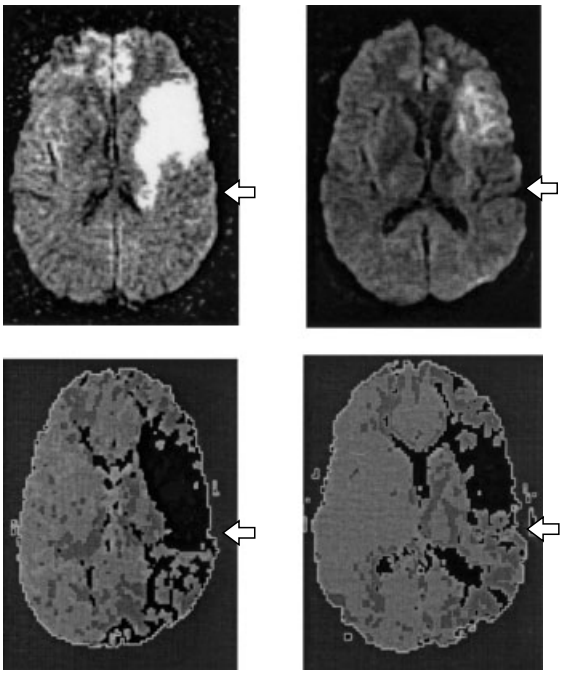


Figure 1. (a) Magnetic resonance diffusion-weighted (top) and perfusion-weighted imaging showing acute stroke involving Broca's area (and other frontal regions, and caudate), with hypoperfusion of Wernicke's area (arrow) in a patient with early global aphasia. Darker areas are regions of poor perfusion. (b) The same patient, after partial reperfusion of Wernicke's area and concomitant resolution of the comprehension deficit.

speech are relatively intact, while fluency and articulation are impaired in spontaneous speech but not in sentence repetition. Stroke or poor blood flow to the left dorsolateral frontal lobe, anterior and superior to Broca's area, has been associated with this form of aphasia. Finally, in mixed transcortical aphasia, also known as 'isolation syndrome', all language functions are impaired except repetition. Such patients can repeat sentences verbatim, with no evidence of comprehending them. A combination of lesions causing transcortical motor and transcortical sensory aphasia has been postulated as the cause of mixed transcortical aphasia.

Conduction Aphasia

In contrast to the transcortical aphasias, conduction aphasia is associated with disproportionately impaired sentence repetition. Speech is fluent and grammatical, although phonemic paraphasias are often present. The frequently observed phenomenon of progressive self-correction of speech errors, such as 'tormano, tornano, tornado', has been called '*conduit d'approche*'. In at least some cases the impaired repetition has been shown to be due to severely limited auditory short-term memory, such that the patient is unable to retain the entire sentence in short-term memory while articulating it (Warrington and Shallice, 1969). Conduction aphasia was initially claimed to result from lesions of the arcuate fasciculus (resulting in a putative disconnection between Wernicke's area and Broca's area), but there is little evidence for this particular lesion–deficit association. More recently, conduction aphasia has been ascribed to lesions of the left supramarginal gyrus and/or the insula or Wernicke's area.

Reading and Writing Disorders

In all of the above aphasia syndromes, reading comprehension is generally impaired at least to the degree of auditory comprehension, and written output is typically impaired at least as much as spoken output, often mirroring the content of speech. However, there are cases of pure (auditory) word deafness, in which comprehension of spoken language is severely impaired, but comprehension of written language is intact (Denes and Semenza, 1975). There have also been reported cases in which written naming accuracy far exceeds spoken naming accuracy. There are also a variety of patterns of pure reading impairment (alexia), writing impairment (agraphia) or both (alexia with agraphia), associated with different lesion sites.

Anomic Aphasia

In anomic aphasia the predominant problem is in naming, or word retrieval. Occasionally, naming is selectively impaired for certain categories, such as proper names, verbs or nouns (Goodglass *et al.*, 1966). More often in category-specific aphasias, both naming and comprehension are impaired or spared in selective semantic categories, such as that of living things (Warrington and Shallice, 1984). In pure anomic aphasia, words are understood but poorly retrieved. Anomic aphasia was initially considered to be nonlocalizing and to be the residual state of other aphasic syndromes, or broadly overlapping with Wernicke aphasia. However, one type of naming disorder that has a well-documented association with lesion site is that of optic aphasia, in which naming of pictures and other visual stimuli is severely disrupted, but naming in response to definitions and naming of stimuli presented in auditory or tactile form are spared. The patient appears to recognize the visual stimulus, and will

often mime how to use the object (Lhermitte and Beauvois, 1973). This clinical syndrome was initially described in 1881 by Dejerine, who proposed that optic aphasia occurred when there were two lesions: one in the left occipital lobe causing right homonymous hemianopia, such that all visual processing takes place in the right occipital lobe; and a second lesion in the splenium of the corpus callosum causing a disconnection of the right occipital lobe from the left hemisphere language areas (Figure 2). Thus, objects can be seen and recognized (in the right hemisphere) but not named (because of the disconnection to the left perisylvian region).

This classification scheme is summarized in Figure 3. It is important to note, however, that this 'syndrome' approach advocated by Geschwind and his colleagues is a clinical nosology, based on frequently co-occurring deficits in individuals who have sustained strokes. It is not based on any defensible theory of language: that is, it has not been proposed that there is a single underlying deficit that gives rise to all of the symptoms observed in a single aphasia syndrome. For example, Broca aphasia refers to the co-occurrence of nonfluent, agrammatic speech, difficulty in comprehending syntactically complex sentences, and effortful articulation. Although it has been postulated that damage to a 'central syntactic processor' might give rise to both the agrammatic speech and the impaired comprehension of syntactically complex sentences, such a proposal does not account for impaired speech articulation. Rather, it is likely that the frequent co-occurrence of these various symptoms reflects the fact that large, consistent regions of the brain are typically supplied by distinct cerebral arteries, the occlusion of which results in stroke. Suppose the large area supplied by a vessel such as the superior division of the left middle cerebral artery (MCA) consists of a number of smaller regions each responsible for a specific language function (e.g. grammatical sentence formulation, computation of syntactic relations, and articulation); occlusion of this vessel would typically result in impairment of all three functions. According to this hypothesis, occlusion of a different cerebral artery would spare these functions and

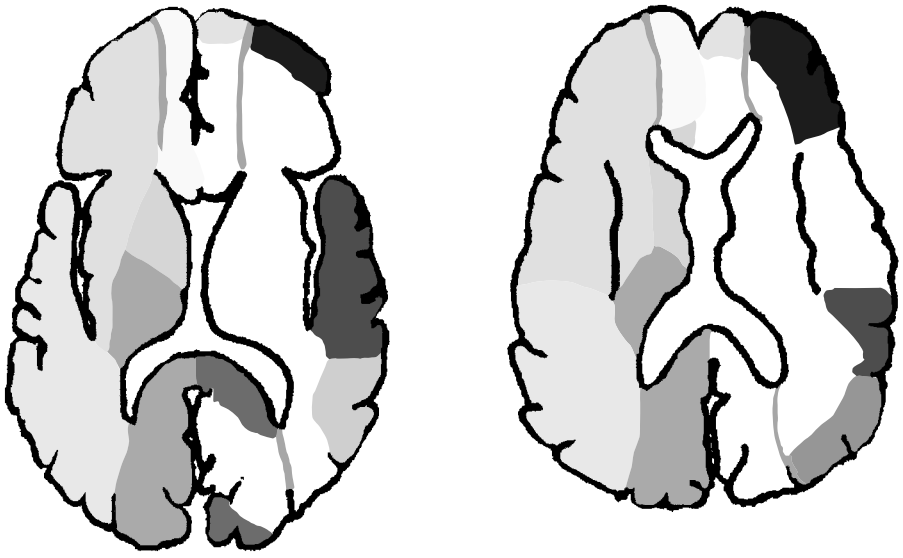


Figure 2. [*Figure is also reproduced in color section*.] Approximate vascular distributions of major arteries supplying language cortex are shown on the left side of each brain cut (the right hemisphere on computed tomographic and magnetic resonance imaging). Pale yellow, anterior cerebral artery (ACA); beige, anterior choroidal branch of internal carotid artery; light blue, superior division of the middle cerebral artery (MCA); pink, inferior division of the MCA; lavender, posterior cerebral artery (PCA); gray, border of ACA/MCA and MCA/PCA (potential 'watershed' areas). Regions of the cortex implicated in classical aphasia syndromes are shown on the right side of each brain cut (the left hemisphere). Bright yellow, Brodmann's area (BA) 10, associated with some cases of transcortical motor aphasia; dark blue, BA 44 and 45, Broca's area; red, BA 22, Wernicke's area; dark green, BA 39, the angular gyrus; light green, BA 37, the posterior middle temporal gyrus; purple, BA 31, the splenium of the corpus callosum and BA 18, visual cortex (areas associated with optic aphasia); gray, 'watershed' areas associated with the transcortical aphasias (the width of this territory depends on the degree of diminished flow in the ACA/MCA or MCA/PCA). Adapted from Damasio and Damasio (1989).

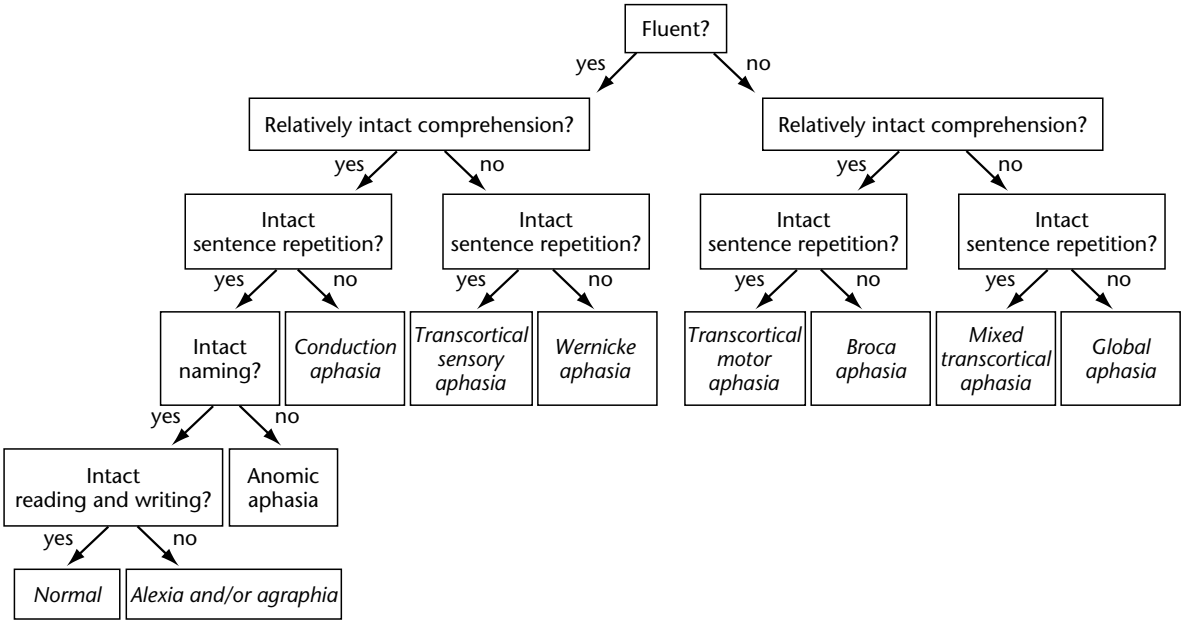


Figure 3. A decision tree for classifying aphasia according to Geschwind's 1965 scheme.

damage other functions. Thus, the clinical syndromes might serve to localize the region of the brain affected by the lesion, and identify the artery involved.

ANATOMICAL CORRELATES OF APHASIA SYNDROMES: VASCULAR DISTRIBUTIONS

There is a wealth of data consistent with the hypothesis that the syndromes described above are manifestations of occlusion of specific arteries. There is evidence that Broca aphasia reflects a blockage of the superior division of the left MCA which supplies the left posterior, inferior frontal lobe and much of the basal ganglia (Figure 2). Because this branch also serves the cortex responsible for motor function of the face and arm on the contralateral side, most patients with Broca aphasia have right face and arm weakness. In contrast, Wernicke aphasia is thought to reflect blockage of the inferior division of the left MCA, resulting in stroke in the left posterior temporoparietal area. This branch also supplies the optic tract (visual pathway), so that patients with Wernicke aphasia often have a contralateral visual field cut. Global aphasia typically reflects occlusion of the left proximal MCA before it divides, resulting in a large stroke involving the left frontotemporoparietal cortex and subcortical structures. Transcortical motor aphasia has most often been attributed to left anterior cerebral artery strokes, involving the left medial frontal lobe, but in other cases has been attributed to strokes in the watershed distribution on the border between the left anterior cerebral artery (ACA) and left MCA (see Figure 2). In these cases, blood flow in both the ACA and MCA is so diminished that it does not reach the borders of the normal territories. Similarly, the collection of symptoms that characterize transcortical sensory aphasia are frequently caused by watershed strokes between the left MCA and left posterior cerebral artery (PCA). Not surprisingly, mixed transcortical aphasia results from strokes involving both watershed regions (ACA–MCA and MCA–PCA), which can occur in the presence of low blood pressure resulting in poor blood flow through all of these vessels. Of course, there can be strokes involving only a portion of the territory of any of these vessels, which would plausibly result in only a subset of the symptoms that form a clinical syndrome. Thus, cases of anomia might be due to damage to a portion of the inferior division of the MCA territory, causing only part of the clinical syndrome of Wernicke aphasia. Anomia can also be caused by lesions involving only portions of the brain regions responsible for other aphasic syndromes, since it can be present in every aphasia type.

WEAKNESSES OF THE LESION–SYNDROME APPROACH

If the above clinical syndromes reflect the relative consistency of vascular beds, such that strokes

generally affect a typical collection of brain regions that may each be crucial for different language functions, it follows that brain damage caused by nonvascular lesions, such as tumor or trauma, would be likely to affect different brain areas and, consequently, different sets of language functions. In other words, profiles of language disturbance would not be expected to correspond with the classical aphasia syndromes. Indeed, tumor (primary or metastases), trauma, abscesses and other infections can all cause aphasia, but the patterns of impairment frequently do not fit the classification scheme outlined above. For example, herpes encephalitis tends to affect the mesial temporal and frontal lobes. When the encephalitis predominantly occurs in the left temporal lobe, a selective impairment in naming and comprehending animals and other living things is frequent. This type of category-specific language deficit has also been reported after closed head trauma, which preferentially affects the temporal lobes, and in the temporal variant of frontotemporal dementia (also called semantic dementia). Patients with frontotemporal dementia may have a primary progressive aphasia with deterioration in language functions before other cognitive abilities. In the temporal variant speech is often fluent, but progressively devoid of content or meaning. In contrast, the frontal variant of frontotemporal dementia is often heralded by a nonfluent, primary progressive aphasia, characterized by increasingly halting, telegraphic speech, with relatively spared comprehension. Tumors and abscesses cause widely disparate language disturbances, depending on both the location of the lesion and the rate of growth. Slow-growing lesions may cause no neurologic dysfunction until late in the disease, when associated brain swelling may acutely compress a ventricle or crucial structure. Even in stroke, there are notable variations in the human vasculature, such that the territory of a given vessel is somewhat different between individuals and may be markedly so in a few. This fact may account for the observation that in the best of circumstances, only about 50% of patients with aphasia due to stroke can be easily classified into one of the classical syndromes.

Furthermore, despite early reports documenting a close relationship between aphasia classification and site of lesion in chronic stroke, the correlation of aphasia type with location of lesion has not withstood recent attempts at replication. In the early studies, exclusion of patients with lesions but without aphasia, or patients with short-lived aphasia, may have resulted in overestimation of the value of lesion location for positively predicting aphasia. Another potential reason for the conflicting results is that early studies may have included only the 'best' or 'cleanest' cases of each syndrome type, whereas other studies may have included cases that were not as easily classifiable. Nevertheless, large-scale studies that have used similar measures to classify patients, and have included a category of 'others' for patients who do not fit well into any of the syndromes, have not provided evidence for a strong localization of these aphasia types (Kreisler *et al*., 2000). Kreisler and colleagues found that the most common site of lesion in patients with Wernicke aphasia and those with Broca aphasia was the insula-external capsule region in both groups. Only 60% of patients with Broca aphasia had strokes involving the inferior frontal gyrus (Broca's area), and only 70% of patients with Wernicke aphasia had strokes involving the left posterior temporal gyrus (Wernicke's area). However, stronger associations were found between impairment of specific language tasks (naming, repetition) or speech characteristics (fluency) and the site of lesion. For instance, 92% of patients with poor auditory comprehension had lesions involving the posterior superior and middle temporal gyri. This study indicates that there may be stronger localization of impairment for specific language tasks than for aphasia syndromes.

However, even the separation of impairments to different tasks may be too gross a characterization of language disorders to identify the neuroanatomical substrates. Kreisler *et al*. (2000) found, for example, that damage to any one of several regions of the brain results in impaired naming. The presence of damage to any one of five areas – the insula–external capsule plus the white matter underlying temporoparietal cortex, the medial temporal gyrus, superior frontal gyrus, middle frontal gyrus, or the thalamus – was associated with impaired picture naming. These observations probably reflect the fact that naming is a complex process, consisting of a number of cognitive processes that may take place in different brain regions. Naming a picture involves, at the very least, early visual processing resulting in recognition of the picture as something familiar; accessing the semantic representation of the object (the collection of features that define how that object is distinguished from other objects with different names); accessing the phonological form of the word (the lexical representation, or stored pronunciation of the

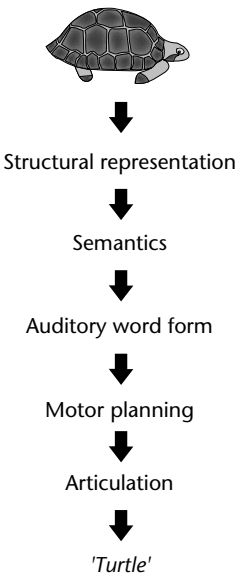


Figure 4. The cognitive processes underlying naming.

word); programming the movements of the lips, tongue, jaw and palate to produce the name; and finally, articulating the word by implementing the planned movements (Figure 4). It is plausible that these components of the naming task are carried out by separate regions of the brain. In this case, damage to any one of the regions would disrupt the person's ability to name an object, although the mechanism would be different across cases.

LOCALIZATION OF SPECIFIC LEXICAL PROCESSES AND REPRESENTATIONS

Studies using advanced neuroimaging of function or dysfunction have indicated that there may be more reliable localizations of each of these component processes (e.g. access to semantic information) than of broader language tasks (e.g. naming).

Structural Representations for Object Recognition

Evidence from surgical ablation studies in primates and functional activation imaging in humans indicate that visual object recognition entails a neural network consisting of retinotopic visual representations in primary visual cortex; visual feature perception, such as perception of color and movement in visual association cortex; unimodal visual/structural descriptions of the shape, color and motion of familiar objects in the medial superior temporal, lateral intraparietal, and the posterior and anterior inferior temporal cortices; and a polymodal representation of familiar objects in the midtemporal cortex.

Semantic Representations

Some lesion studies indicate that impairments of semantic representations (word meaning) are associated with lesions in Wernicke's area and the left middle temporal gyrus. Likewise, electrical stimulation of Wernicke's area interferes with lexical-semantic processing. Positron emission tomography (PET) studies, which show areas of 'activation' or increased regional blood flow (rCBF) associated with increased metabolism, also show activation of Wernicke's area and the left midtemporal cortex in semantic processing. Similar results are found with another method of functional imaging, perfusion-weighted imaging (PWI), which shows regions of poor perfusion, or blood flow, resulting in tissue dysfunction in the setting of cerebrovascular disease. Studies confirm a strong correlation between impaired access to semantics and poor perfusion of Wernicke's area (Hillis *et al*., 2001). Furthermore, when blood flow was restored to Wernicke's area in a subset of these patients, access to semantics was restored, indicating that this region is involved in semantic processing (see Figure 1).

Such findings do not allow the conclusion that Wernicke's area (or any other brain region) is alone responsible for a given cognitive process. For example, although Wernicke's area appears to be important for the understanding of words, it is unclear what – if any – semantic information is represented in this area. It is more likely that semantic representation is distributed across a variety of temporal and parietal brain regions and that the role of Wernicke's area is in linking phonological (spoken) word forms to meanings. Although it is often difficult to establish that patients with damage to Wernicke's area are impaired in mapping words to their meanings, rather than impaired at the level of meanings themselves, rare cases of dissociation between these two functions have been reported in patients with extraordinarily small lesions. For example, Hillis *et al*. (1999) described a patient, J. B. N., who spoke in fluent jargon and failed to comprehend spoken words or sentences, despite normal hearing and 'early' auditory processing. Nevertheless, J. B. N. had intact writing and comprehension of written language. Thus, this patient did not have impaired semantics, or word meanings, but was impaired in linking spoken

words to their meanings and vice versa, owing to poor blood flow in the sylvian branch of the anterior temporal artery, supplying Wernicke's area. Interestingly, this proposal about the role of Wernicke's area in linking words to their meanings was first put forward by Carl Wernicke himself in the 1870s.

Auditory Word Forms

Most functional imaging studies fail to distinguish between access to auditory word forms and mapping auditory word forms to semantics, since recognizing or saying a word is likely to 'automatically' activate its meaning. Hence, not surprisingly, PET studies show increased blood flow in Wernicke's area during tasks of auditory word-form processing. As discussed above, this region may be where auditory word forms are linked to semantics, not where auditory word forms are represented. Evidence from PWI in acute stroke patients indicates that impaired naming without impaired comprehension occurs with hypoperfusion (poor blood flow) just posterior and inferior to Wernicke's area. To illustrate, the patient whose scans are shown in Figure 5 had selective impairment in accessing auditory word forms for oral naming and oral reading when the posterior middle temporal gyrus was hypoperfused (receiving poor blood flow), but this impairment resolved when this region was reperfused the following day. Such studies indicate that the posterior middle temporal gyrus was crucial for access to auditory word forms for output, but not for semantics, in this patient.

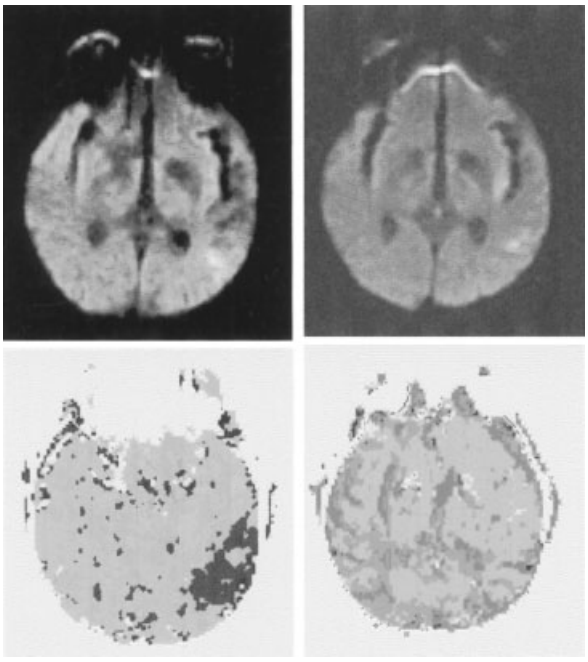


Figure 5. [*Figure is also reproduced in color section.*] (a) Magnetic resonance diffusion-weighted imaging (top) and perfusion-weighted imaging (bottom) showing hypoperfusion of the posterior middle temporal gyrus on day 1 in a patient with impaired access to auditory word forms for output, causing poor oral naming and oral reading, but intact semantics. (b) Day 2, showing reperfusion of the posterior middle temporal gyrus when naming and oral reading abilities had recovered.

Motor Speech

Studies of stroke patients indicate that Broca's area and the medial third of periventricular white matter are critical for motor speech. Similarly, intraoperative cortical stimulation, causing temporary, focal dysfunction of regions within Broca's area, disrupts motor speech.

WEAKNESSES OF THE LESION–DEFICIT APPROACH

We have discussed the role of various cortical regions that seem to be involved in specific components of the naming process. However, subcortical structures may also play a role in carrying out these processes, since subcortical stroke can cause aphasia. Studies of dysfunctional tissue, using PET, single photon emission computed tomography or PWI, have shown that at least some cases of aphasia associated with lesions restricted to subcortical structures show perfusion abnormalities in the temporoparietal cortex (Figure 6). These studies indicate that language deficits in such cases might be due to low blood flow to the cortex, rather than to the subcortical lesion itself.

Most lesion–deficit studies are complicated by the fact that the neural substrates of a given function may change after stroke. Reorganization of motor and sensory (and perhaps language) 'maps' shift over the course of days to months after brain injury or peripheral lesions. Therefore, the neural regions responsible for any given language process are probably identified best by experimental temporary lesions (such as those produced by transcranial magnetic stimulation) or by imaging stroke and dysfunctional tissue in the first day of stroke symptoms before significant reorganization occurs. Investigations of the correlation between language deficits in the first day of stroke and concurrent

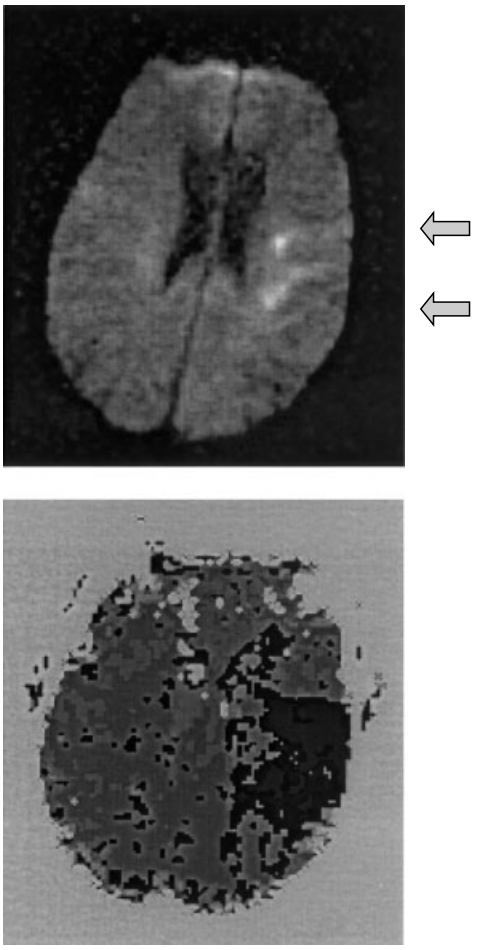


Figure 6. Diffusion-weighted imaging (top) showing tiny subcortical strokes (arrows), and perfusion-weighted imaging done at the same time, showing cortical hypoperfusion of the left temporal lobe, in a patient with global aphasia. Darker areas show regions of poor perfusion.

regions of abnormality defined by diffusion-weighted imaging (DWI), which shows areas of acute stroke within minutes of onset, and PWI are likely to be less contaminated by varying degrees of reorganization or recovery than traditional studies of chronic deficits and lesions. Preliminary studies using these advanced imaging techniques have revealed some strong associations between areas of dysfunction (due to hypoperfusion) and impairments of specific lexical functions in acute stroke before reorganization. These data converge with results from activation studies (using functional magnetic resonance imaging and PET) in support of the hypothesis that language tasks such as naming involve a network of brain regions that are each responsible for relatively discrete types of language representations or processes.

CONCLUSION

A profusion of reports spanning the last century and a half of brain lesions associated with patterns of language disturbance has shown that occlusion of a given cerebral artery or branch often results in a particular aphasia syndrome – a set of frequently co-occurring language deficits. More recent studies, using advanced functional imaging to show areas of activation in normal people engaged in a particular task or to show regions of dysfunctional brain tissue at the time of onset of specific language deficits, have demonstrated that distinct processing components of any language task, such as naming, are subserved by separate brain areas. These findings are consistent with research in cognitive neuropsychology and other branches of cognitive science, showing that each language task entails numerous levels of representation and processing which can be differentially impaired by brain damage. However, we are still far from defining a theory of the neural substrates for language processing that goes much beyond a coarse taxonomy of the major components of language: semantics, auditory (phonological) word forms and the like. This is in contrast to the rich data that aphasia provides for inferring the structure of normal language processing. For example, there are many reports of exquisitely fine-grained dissociations involving selective damage or sparing of spoken but not written word retrieval, nouns relative to verbs, content words (nouns and verbs) relative to function words, inflectional (e.g. number and tense) relative to derivational morphology (such as the '-er' rule of noun formation from verbs, e.g. 'hunter'), consonants relative to vowels, and so on. Dissociations such as these impose biologically motivated constraints on the theories of lexical processing. However, we still do not have an equally fine-grained analysis of the neural mechanisms corresponding to the functional units of language processing revealed through the detailed analysis of language deficits in aphasia.

References

Caramazza A and Zurif E (1976) Dissociation of algorithmic and heuristic processes in language comprehension. *Brain and Language* **3**: 572–582.

Damasio H and Damasio A (1989) *Lesion Analysis in Neuropsychology*. New York, NY: Oxford University Press.

Denes F and Semenza C (1975) Auditory modality-specific anomia: evidence from a case of pure word deafness. *Cortex* **11**: 401–411.
Goodglass H, Klein B, Cary P and James KJ (1966) Specific semantic word categories in aphasia. *Cortex* **12**: 145–153.
Hillis AE, Boatman D, Hart J and Gordon B (1999) Making sense out of jargon: a neurolinguistic and computational account of jargon aphasia. *Neurology* **53**: 1813–1824.
Hillis AE, Kane A, Barker P *et al*. (2001) Neural substrates of the cognitive processes underlying reading: evidence from magnetic resonance perfusion imaging in hyperacute stroke. *Aphasiology* **15**: 919–932.
Kreisler A, Godefroy O, Delmaire C *et al*. (2000) *Neurology* **54**: 1117–1123.
Lhermitte E and Beauvois MF (1973) A visual-speech disconnexion syndrome: report of a case with optic aphasia, agnosic alexia and colour agnosia. *Brain* **96**: 695–714.
Warrington E and Shallice T (1969) The selective impairment of auditory verbal short-term memory. *Brain* **92**: 885–896.
Warrington E and Shallice T (1984) Category specific semantic impairments. *Brain* **107**: 829–853.

Further Reading

Alexander MP (1997) Aphasia: clinical and anatomical aspects. In: Feinberg TE and Farah MJ (eds) *Behavioral Neurology and Neuropsychology*, pp. 133–150. New York, NY: McGraw-Hill.
Caplan D (1992) *Language: Structure, Processing and Disorders*. Cambridge, MA: MIT Press.
Caramazza A (2000) Aspects of lexical access: evidence from aphasia. In: Grodzinsky Y, Shapiro L and Swinney D (eds) *Language and The Brain: Representation and Processing*, pp. 203–228. San Diego, CA: Academic Press.
Goodglass H (1993) *Understanding Aphasia*. San Diego, CA: Academic Press.
Goodglass H and Kaplan E (1972) *The Assessment of Aphasia and Related Disorders*. Philadelphia, PA: Lea & Febiger.
Ojemann GA (1994) *Cortical Stimulation and Recording in Language*. London, UK: Academic Press.
Sarno MT (1998) *Acquired Aphasia*. San Diego, CA: Academic Press.

Apraxia

Intermediate article

Michael Peters, University of Guelph, Guelph, Ontario, Canada

CONTENTS

Apraxias are movement disorders that cannot be attributed to primary motor problems such as paralysis or weakness, or to mental incompetence. Apraxias manifest in the failure to perform, accurately and smoothly, simple or complex movements. In studying apraxias we learn much about purposive motor behavior in general.

INTRODUCTION

In the absence of a generally agreed definition, apraxia is identified here as a higher-order movement disturbance that involves many facets of motor behavior (Freund, 1992). However, more restrictive definitions are often used in clinical applications. A widely used definition states that apraxia is 'a disorder of skilled movement that is not caused by weakness, akinesia, deafferentation, abnormal tone or posture, movement disorders – such as tremors or chorea – intellectual deterioration, poor comprehension, or uncooperativeness' (Heilman and Rothi, 1993). It manifests itself in the failure to perform, accurately and smoothly, simple or complex movements by imitation, or in response to verbal command.

In approaching the problem of apraxia, we can subdivide the production of skilled movement into several stages. First, there is the general idea of an action plan. For instance, we may decide to peel an apple. Second, there is the translation of the general 'I want to …' idea into a specific plan of action. In this case, we need to find a suitable knife, and then the apple. Once we have the knife and the apple, we can execute the action plan 'peel the apple'. There are two different aspects to execution. First, we need to have a clear idea as to how we peel an apple. One hand will have to hold the apple and this hand will have to position the apple so that the knife can act on it. This requires us to monitor the orientation of the apple in space, with constant adjustments of the position as the act of peeling progresses. Simple as the act of peeling is, it requires us to orient the apple so that the peeling strokes are not wasted. We do not wish to peel the same area twice, or rotate the apple so that bits of peel are missed. The second aspect of execution lies in the actual motor aspects of the peeling action. That is, we know what we want to do, but now we must do it as well as we can. The angle of the knife must be appropriate and the strokes must be such that they remove as little of the apple flesh and as much of the skin as possible. In addition, the actions of the two hands have to be coordinated so that the knife stroke does not begin before the apple is oriented properly. This example illustrates the very simple case where an action is implemented out of the person's own volition. When a person is asked to demonstrate an action, for example how to use a hammer, there is the important element of memory which is implicated both in the association of the object with hammering as a concept as well as the association of the object with a specific type of action. Here, access to motor planning areas through visual or auditory or even tactile avenues becomes an important component of performance.

Considering the number of variables involved, it is easy to see that problems in carrying out this task can be affected by disturbances at different levels, from the initial development of an action plan to the technical aspects of carrying out the movement. By consensus, the definition of apraxia excludes problems that arise in the actual implementation of movement, or what might be called the very last stage of producing skilled movement. In addition, the definition also excludes problems at the 'top end' of the process, so that difficulties in carrying out skilled movement due to general intellectual deterioration, an inability to understand, or a lack of willingness to cooperate, are also excluded from the definition of apraxia.

The definition places the problem of apraxia somewhere between actual motor implementation at one extreme and general mental competence at the other. The variety of factors that can produce

apraxia of skilled movement, once these two end points are excluded, is still dauntingly large. Indeed, because the causes of the disturbance can be so varied, investigators have attempted to provide a classification of apraxia in terms of both the body parts affected, and the particular level at which the disturbance occurs. There is no ideal or compelling classification of apraxia, but in the following we shall discuss the major categories of apraxia that have been recognized. In particular, we shall see that it is not always possible to separate general mental state and the state of the motor system from intervening variables that are thought to cause apraxia.

VARIETIES OF APRAXIA

Ideomotor Apraxia

By far the best-researched and understood type of apraxia is ideomotor apraxia, especially that involving the upper limbs. For this reason, we will consider some of the general aspects of how to test for apraxia, and the principal categories of disturbances, within the framework of this form of apraxia.

Testing for apraxia

Recognition of the different kinds of apraxia depends on clinical examination and experimental measurement of motor behavior. Although there is no standardized methodology, certain approaches are common to many investigations. In examining motor performance, individuals can be asked verbally to perform a task, or they can be asked to imitate a movement that is demonstrated for them. Investigators distinguish between transitive movements, where the individual acts on an object or operates a tool, and intransitive movements where no object is acted on. Examples of intransitive movements would be saluting, taking up the stance of a boxer, or waving goodbye. A further distinction is occasionally made between movements that are representational and those that are not.

Part of testing for ideomotor apraxia involves an assessment of actual motor performance, to evaluate the possible role of 'lower level' causes such as tremor or weakness. In the past, data were mostly collected by recording observations of the performance of the patients, and subtle deficits in motor execution were often missed or ignored. More recently, precise recording of movement in space and time has become possible, and studies in which precise measurement of movement supplements clinical observation are becoming more widespread.

General disturbances in apraxia

Because of the complex neurological machinery that underlies skilled motor behavior, one can expect that lesions in different areas will produce different kinds of problems. Such disturbances can manifest themselves in whatever motor systems are involved (such as movements of the speech musculature, the body axis, or limbs). For the limbs, the following problems have been noted.

Problems in spatial function

Current consensus holds that much of motor planning – which by definition includes spatial factors because movement implies movement in space – relies on regions in the parietal lobe of the cortex. We can expect that lesions in this region will lead to problems both in the planning and execution of movement, and in spatial problems. A distinction is made between personal and extrapersonal space, the former referring to the spatial relations of parts of the body to each other, while the latter refers to external objects and their spatial relations to each other and the body of the observer. Common problems in personal space would involve the incorrect positioning of a tool: thus, if asked to pantomime cutting a loaf of bread, an individual might position the hand and imaginary knife in the horizontal rather than the vertical plane. An incorrect coordination of joints will also lead to spatial errors during movement. Spatial errors also result when there is a faulty integration of proximal and distal parts of the limb: for instance, an individual might position the arm correctly for a salute while assuming an unrelated posture of the hand.

Apraxic patients may show problems not only in the execution of movement in space but also in imagining movement. Interestingly, such difficulties can be quite specific to imagined movement, leaving spatial imagery involving objects intact (Ochipa *et al.*, 1997).

Problems in timing, sequencing and initiation of movement

Timing is often confounded with sequencing because in sequencing of movement the timing of the initiation and cessation of component movements determines successful 'running off' of a sequence. It may be helpful to distinguish between

two kinds of sequencing. First, there is the rapid sequencing of progressive movements, such as in those involving several joints. Much of the timing is automatic, and may be considered 'process' timing. There is also the sequencing of separate functional components in the gain of a goal-oriented action. For example, in preparing an envelope for mailing, the separate components involve inserting the letter, sealing the envelope, attaching the stamp and writing the address. Problems in timing and sequencing in ideomotor apraxia can be observed following parietal or frontal damage, with different causation of timing errors for each location. Modern techniques allow a clear visual presentation of apraxic sequencing problems in three-dimensional space (Poizner *et al*., 1990). Timing anomalies are also of great importance in the documentation of apraxia of speech. Clinical observations suggest that not only is the sequencing of movements across joints in the same limb affected by apraxia, but that sequencing of movement between hands (bimanual movements) is often the most conspicuously impaired activity in apraxic patients, to the point that bimanually coordinated movements may not just be poor but impossible to perform.

Simple movements

The study of simple movements is important for methodological reasons, because such movements can be measured with some precision. The many regions in the brain that are responsible for the production of skilled movement add to the controversy about the execution of 'simple' movements. For instance, rapid finger-tapping has been said to be affected in apraxia by some, while others fail to find an effect (Heilman, 1975; Haaland *et al*., 1980). The initiation of movement as such may also be affected in apraxia, but here too the evidence is not clear. It is in the study of simple movements where the dividing line between apraxia and problems in motor execution becomes difficult to define. We may draw a parallel to agnosia, where individuals have difficulties with the recognition of objects. While it is felt that such difficulties cannot be attributed to problems in sensory processing as such, it is extremely rare to encounter an agnosia without some degree of sensory impairment. Similarly, it is rare indeed to encounter a person with apraxia who does not have some degree of motor impairment, however subtle. Nevertheless, because impairments in motor execution do not lead to apraxia in themselves, it is felt that impairment in movement execution may coexist with, but does not cause apraxia.

Ideational Apraxia

Ideational apraxia is rare but striking, and involves problems in sequencing the separate acts that lead to goal-directed behavior. For instance, when asked to prepare a cup of tea, the difficulties would not lie so much in the execution of the individual acts – such as reaching for and filling a kettle, placing tea in the teapot, letting it steep, and then pouring the tea – as in the correct sequencing of the component acts. We note here that the component acts are in themselves complex, and they may be performed quite well in isolation. For instance, the individual might place the teabag reasonably well – but in the kettle, not in the teapot. It is true that ideational apraxia is rarely seen without ideomotor apraxia, but they vary in severity independently from each other. Ideational apraxia affects normally occurring activities while ideomotor apraxia is more often observed in precisely the opposite setting, when movements are to be performed out of context. Some see ideational apraxia as part of a disconnection syndrome, where the motor planning substrate has impaired or no access to the motor execution machinery. The specificity of ideational apraxia to sequencing of motor acts has been illustrated by showing that patients will not be able to arrange in sequence a number of pictures that show progressive phases of actions with given objects, but are nevertheless able to correctly sequence pictures that show a chain of events that did not involve a sequence of movements (Lehmkuhl and Poeck, 1981).

Conceptual Apraxia

Conceptual apraxia and ideational apraxia are often considered to be synonymous. However, it is reasonable to distinguish between a sequencing problem for a succession of movements, and a problem that involves the proper selection of a specific movement. Heilman and Rothi (1993) emphasize that conceptual apraxia often involves the conceptual aspects of tool use, such as impairment in performing the appropriate motion with a given tool, not being able to pair a tool with the object that the tool would normally act on, choosing substitute tools that do not possess the essential features needed for a given application, or the ability to fashion a tool. As in the case of ideational apraxia,

conceptional apraxia is often seen together with ideomotor apraxia, but not necessarily so. The converse, ideomotor apraxia without conceptual apraxia, is quite common.

Apraxia of Speech

Apraxia of speech, also referred to as 'anarthria', has been controversial because it is difficult to disentangle anarthria from Broca aphasia, where patients have difficulties in producing fluent speech, or dysarthria, where the final elements of speech articulation are affected. In speech apraxia, problems arise regardless of the context in which speech is produced. This contrasts with Broca aphasia, where it does matter under what conditions vocalizations are produced. For instance, in Broca aphasia singing may be relatively much less affected than spontaneous speech, while in speech apraxia the patient will show no such difference. If the person with limb apraxia has difficulties in accessing skilled movements and postures of the limbs, the person with speech apraxia has difficulties in accessing and organizing the 'postures' and movements necessary in speech production. The problem lies at a level one step removed from the actual production of speech sounds. This can be demonstrated in patients who have buccofacial apraxia (problems in assuming tongue and mouth postures) without speech apraxia. However, more often than not speech apraxia is accompanied by such problems (Martin, 1974). As might be expected, speech apraxia is sensitive to the length of utterances.

Apraxia in Dementia and Other Brain Disorders

Apraxias are part of the defining clinical symptoms of dementias and of Alzheimer disease in particular. Strictly speaking, because such apraxias are part of a disease that involves general intellectual deterioration, they do not fall under the general exclusionary definition. However, because apraxias in Alzheimer disease are an important part of the clinical picture, they are considered under this label. It is not surprising, considering the widespread neurological changes in Alzheimer disease, that numerous varieties of apraxia have been described. Conceptual apraxia, ideomotor apraxia, dressing apraxia, constructional apraxia and ideational apraxia have all been described. Systematic work in this area is only now beginning, and some claims have been made that patients with Alzheimer disease are more impaired in performing intransitive than transitive movements, compared with patients with specific left hemisphere brain damage. Contrary claims are also made. While apraxia in Alzheimer disease is common, it also occurs in other dementias and degenerative brain diseases (Leiguarda *et al*., 1997).

Other Varieties of Apraxia

Limb-kinetic apraxia

In limb-kinetic apraxia, we find that the idea of the movement plan is spared but that the execution of even simple and practiced movements is coarse. Limb-kinetic apraxia is found contralateral to the lesion (in contrast to ideomotor limb apraxia, where damage may be ipsilateral or contralateral to the apraxic limb). Because limb-kinetic apraxia is close to the implementation aspect of movement, some have suggested the label 'apraxia of execution' (Freund, 1992). In prototypical patients, it has been claimed that there is no evidence of direct primary defects in motor function and the defect is one of 'the breakdown of fine skillfulness of fingers' in the absence of other types of apraxia (Denes *et al*., 1998).

Apraxia of lid opening

Apraxia of lid opening shares with limb-kinetic apraxia the debate of whether it should be considered a proper apraxia. This apraxia presents itself as inability to close the eyelids voluntarily even though they close or open during reflex blinking (Chapanis and Gropper, 1968; Boghen, 1997; Defazio *et al*., 1998).

Axial apraxia

In axial (or truncal) apraxia neck and trunk movements are affected. Most interesting are cases where aphasia (in this case language comprehension) affects limb movements, while truncal and gait movements can be performed without the patient having any clear comprehension of the verbal command asking for trunk and gait movements. This is probably due to some capacity of the right hemisphere to perform bilateral trunk and gait movements.

Gait apraxia

Whether or not gait apraxia should be considered an apraxia is not clear. Because gait movements have been observed in spite of intact sensation or limb weakness some clinicians feel that a true gait apraxia may exist. However, a predominant aspect of gait apraxia is a lack of initiative to move the legs

and a lack of spontaneity in movement. To the extent that difficulties with gait can be caused by many factors it is likely that cases of true gait apraxia are very rare. The point has been made that the problem with gait apraxia is not so much one of incorrect or inappropriate movements but one of problems with initiation (Brown, 1972).

Dressing apraxia

Dressing apraxia may denote a true apraxia, such that the organization of the learned and skilled movements required for dressing are affected, but the failure to dress properly may also be due to more general problems, such as lateral neglect.

Constructional apraxia

In the clinical demonstration of constructional apraxia, individuals fail to arrange objects such as building blocks according to a visually presented scheme. Such cases, often associated with right hemisphere lesions, are more appropriately considered a secondary outcome of visuospatial processes.

NEUROANATOMY

The Role of the Left Hemisphere

There is little doubt that the left hemisphere has a predominant role in apraxia. Beginning with the earliest descriptions of apraxia, in the vast majority of cases the patients have left hemisphere damage. This supports the general understanding that in right-handed people especially, the conception and generation of movements relies heavily on left hemisphere function. Beginning with Liepmann in 1905, attempts have been made to separate different types of apraxia in terms of cortical lesion location. Thus, Liepmann suggested that lesions associated with limb-kinetic apraxia would be straddling the primary motor and sensory cortex, lesions causing ideomotor apraxia would be posterior to this in the higher-order sensory cortex, and lesions causing ideational apraxia even further posterior, close to the visual cortex.

Subsequent work is less confident of a precise allocation of cortical region for specific apraxias. Three general regions of damage are associated with apraxia. First, there is the parietal region in general and the angular and supermarginal gyri in particular, which are associated with the representation of movement schemes in time and space. Large lesions in this region would lead not only to apraxia, but also to an inability to judge whether a movement demonstrated to the patient is flawed or not. Lesions that leave this region intact but are slightly more anterior (without infringing on premotor regions in the frontal cortex) would lead to apraxia, but affected patients would have the ability to judge whether a movement demonstrated to them is flawed or not. Such lesions would be responsible for 'disconnection' apraxias, where the problem arises from a lack of access from the posterior areas involved in planning and movement image generation to the anterior areas that are more directly involved in implementation. Small lesions that sever the arcuate fasciculus, a tract that connects posterior to anterior cortex, can produce apraxia. This type of disconnection apraxia is probably the only kind where a circumscribed small lesion can cause apraxia (Tanabe *et al*., 1987).

In contrast to posterior lesions, anterior lesions (anterior to the primary motor cortex) are implicated in disturbances of implementation and specific sequencing of consecutive motor acts. The supplementary motor area (SMA) has long been suspected of involvement in apraxia and a number of cases are known where the SMA has been specifically implicated. It is of interest that these cases involve bilateral apraxia; the SMA is known to be involved in bimanual coordination. In addition, the premotor cortex has been identified with limb-kinetic apraxia, also involving both arms.

The Role of the Right Hemisphere

The role of the right hemisphere in apraxia is contested. Soon after the emphasis on left hemisphere lesions in apraxia was pointed out, cases of apraxia after right hemisphere lesions were described (Brun, 1922). Isolated cases are known of right-handed people who are fully apraxic in the right hand after right brain lesions. However, there is a debate about whether in larger series of brain-damaged individuals, some degree of apraxic impairment is seen in those with right hemisphere lesions. It appears that in such series, apraxia after right brain damage is rare. However, a number of researchers show that subtle problems exist. For instance, patients with right-side lesions performed consistently much better than patients with left-side lesions on most tests for apraxia, but they performed worse than control subjects on verbal command when asked to make intransitive movements. In addition, of the 11 patients with right hemisphere lesions, five performed at a level below that of the worst control subject. Other researchers similarly found that while individuals with right-brain damage showed fewer apraxic problems, they performed less well than normal

participants on specific tests (Goldenberg and Hagmann, 1998). It is clear that much of the negative evidence on right hemisphere contributions cannot be considered conclusive because it matters what movements were tested, and to what extent the right hemisphere would be involved in a given movement (Roy *et al*., 1991). An important distinction is between the frequency of apraxic disturbances and the severity; the frequency of apraxia tends to be less after right hemisphere lesions while the severity of apraxia when it does occur is comparable after right and left hemisphere lesions.

The Corpus Callosum

Implicit in the idea that the left hemisphere is predominant in the formulation of movement plans is the assumption that such plans reach the right hemisphere regions responsible for the movements of the left arm via the corpus callosum, from the left hemisphere. 'Pure' lesions in the anterior regions of the corpus callosum that carry motor commands to be implemented by the right hemisphere might therefore lead to apraxia in the left arm – even though the right arm will not be affected. Some cases corresponding to this scheme have been described.

Subcortical Involvement

In general, lesions of subcortical structures that lead to apraxia also involve cortical damage, and it is difficult to disentangle the effects of such multiple lesions. Perhaps the strongest evidence for a distinct subcortical lesion associated with apraxia (usually ideomotor apraxia) comes from apraxic patients with restricted thalamic lesions (Pramstaller and Marsden, 1996). Speculative interpretations implicate part of the thalamus that is intimately connected with a cortical region known to be involved in apraxia. One possible involvement of the basal ganglia in particular may relate to reports of slight difficulties in the initiation of movement in some apraxic patients. Here, it may be suggested that in addition to the cortically derived signs of apraxia, subcortical contributions may be involved. The failure to see slowness in initiation of movement in some apraxic patients may therefore stem from a noncortical source.

APRAXIA AND APHASIA

The association between aphasia and apraxia is well known and documented. In the demonstration of ideomotor apraxia a language problem was implicit in the very first descriptions, because it was the inability to perform in response to verbal commands that defined the disturbance. In addition, even if patients are selected not on the basis of apraxia but on the basis of brain lesions in the left or right hemispheres, patients with left-brain damage will show a strong association of aphasia with ideomotor apraxia. Similarly, in the much rarer cases of ideational apraxia, lesions that tend to produce this apraxia will almost always also produce aphasia. Thus, there are strong links between apraxia and aphasia for different apraxias. However, it is here where individuals with anomalous cerebral lateralization of praxis and language functions provide important information. Despite their rarity, over time a considerable number of patients have been described who have dissociations between aphasia and apraxia. Such patients usually have what is called 'crossed apraxia', where apraxia is caused by right hemisphere lesions. In some of these aphasia is also produced by right hemisphere lesions, but in others there is no aphasia. Of greatest importance here are left-handers who will tend to have language presentation in the left hemisphere. While there is some debate as to whether left-handers show apraxia after right hemisphere lesions, specific and careful testing suggests that this is the case. Careful testing is indicated because left-handers may show less severe apraxia than right-handers and faster recovery after lesions in either hemisphere.

While it is tempting to speculate that the strong general association of aphasia with apraxia is due to higher-order mechanisms that are involved in generation action in either language or body movement, it is more likely that the coincidence of disturbances in both domains is due to lesions that are large enough to affect both substrates involved with language and body movement. Even allowing for the fact that some apraxias often are unrecognized, apraxias are less common after general brain damage than aphasias.

Aphasias therefore appear to be related to apraxias in two ways. First, directly, as in apraxias that emerge when movement to verbal command is tested and when the access from language areas to motor implementation is impaired. Second, there is a general correlation between presence and severity of aphasia and presence and severity of apraxia. Two possible explanations can be considered. First, the regions involved with language comprehension and motor planning in the posterior cortex are coextensive and lesions that affect one function are likely to affect the other. This is slightly

less of a problem in the frontal lobe, but the same principle obtains. For instance, lesions producing Broca aphasia are likely to impinge on regions associated with buccofacial apraxia. Second, it is possible that especially at the top level of motor planning and language preparation, there are higher-order mechanisms that operate at a level above the specific modality and when these are damaged, both apraxia and aphasia ensue because mechanisms common to both are affected.

CONCLUSION

The systematic study of apraxia is in its infancy. If we adopt a broad definition of apraxia, we can see that damage in many parts of the central nervous system can lead to apraxia, and that the causes for the observed 'higher order motor disturbance' are manifold, ranging from relatively low-level problems in the timing and sequencing of unfolding movements to problems in the conceptual plan for movements. Future work will probably focus on an aspect of apraxia that has been relatively neglected by research: the representation of and access to motor memory.

In the course of trying to understand apraxia, we are also trying to understand purposeful motor behavior and its generation by analyzing the stages from conception of a movement plan to the final realization of that plan with a specific set of muscles. This will benefit neural network applications of movement control. Neural network approaches have been extremely successful in simulation of language recognition and the recognition of stimuli in the various sensory modalities. There is little doubt that some of the neural network techniques applied to recognition of incoming information can also be of use in the guidance of robotic movement. For instance, there is no reason why specified motor concepts such as 'reaching' and 'grasping' cannot be used in algorithms that control simple movements. Nevertheless, there is a large gap between neural networks that can implement a concept such as 'reaching' and neural networks that can generate such concepts.

References

Boghen D (1997) Apraxia of lid opening: a review. *Neurology* **48**(6): 1491–1494.

Brown JW (1972) *Aphasia, Apraxia and Agnosia*. Springfield, IL: Charles C Thomas.

Brun R (1922) Klinische und anatomische Studien über Apraxie. II Zur Lokalisation der Apraxie. *Schweizer Archiv für Psychiatrie und Neurologie* **10**: 186–209.

Chapanis A and Gropper BA (1968) The effects of operator's handedness on some directional stereotypes on control-display relationships. *Human Factors* **10**: 303–320.

Defazio G, Livrea P, Lamberti P *et al*. (1998) Isolated so-called apraxia of eyelid opening: report of 10 cases and a review of the literature. *European Neurology* **39**: 204–210.

Denes G, Mantovan MC, Gallana A and Cappelletti JY (1998) Limb-kinetic apraxia. *Movement Disorders* **13**: 468–476.

Freund HJ (1992) The apraxias. In: Ashbury AK, McKahann GM and McDonald WJ (eds) *Diseases of the Nervous System*, pp. 751–767. Philadelphia, PA: WB Saunders.

Goldenberg G and Hagmann S (1998) Tool use and mechanical problem solving in apraxia. *Neuropsychologia* **36**: 581–589.

Haaland KY, Porch BE and Delaney HD (1980) Limb apraxia and motor performance. *Brain and Language* **9**: 315–323.

Heilman KM (1975) A tapping test in apraxia. *Cortex* **11**: 259–263.

Heilman KM and Rothi LJG (1993) *Apraxia*. In: Heilman KM and Valenstein E (eds) *Clinical Neuropsychology*, pp. 141–163. New York, NY: Oxford University Press.

Lehmkuhl G and Poeck K (1981) A disturbance in the conceptual organization of actions in patients with ideational apraxia. *Cortex* **17**: 153–158.

Leiguarda RC, Pramstaller PP, Merello M *et al*. (1997) Apraxia in Parkinson's disease, progressive supranuclear palsy, multiple system atrophy and neuroleptic-induced parkinsonism. *Brain* **120**: 75–90.

Martin AD (1974) Some objections to the term apraxia of speech. *Journal of Speech and Hearing Research* **39**: 53–64.

Ochipa C, Rapcsak SZ, Maher LM *et al*. (1997) Selective deficit of praxis imagery in ideomotor apraxia. *Neurology* **49**: 474–480.

Poizner H, Mack L, Verfaellie M, Rothi LJG and Heilman KM (1990) Three-dimensional computergraphic analysis of apraxia. Neural representations of learned movement. *Brain* **113**: 85–101.

Pramstaller PP and Marsden CD (1996) The basal ganglia and apraxia. *Brain* **119**: 319–340.

Roy EA, Square-Storer P, Hogg S and Adams S (1991) Analysis of task demands in apraxia. *International Journal of Neuroscience* **56**: 177–186.

Tanabe H, Sawada T, Inoue N *et al*. (1987) Conduction aphasia and arcuate fasciculus. *Acta Neurologica Scandinavica* **76**: 422–427.

Further Reading

Brown JW (1972) *Aphasia, Apraxia and Agnosia*. Springfield, IL: Charles C Thomas.

Goldenberg G (1995) Imitating gestures and manipulating a mannikin – the representation of the human body in ideomotor apraxia. *Neuropsychologia* **33**: 63–72.

Goldenberg G, Hermsdorfer J and Spatt J (1996) Ideomotor apraxia and cerebral dominance for motor control. *Cognitive Brain Research* **3**: 95–100.

Heilman KM and Rothi LJG (1997) *Apraxia: The Neuropsychology of Action*. Hove, UK: Psychology Press.

Leiguarda RC and Marsden CD (2000) Limb apraxias: higher-order disorders of sensorimotor integration. *Brain* **123**: 860–879.

Liepmann H (1905) Die linke Hemisphäre und das Handeln. *Münchener Medizinische Wochenschrift* **52**: 2322–2326, 2375–2378.

Poeck K (1986) The clinical examination for motor apraxia. *Neuropsychologia* **24**: 129–134.

Rothi LJG, Ochipa C and Heilman KM (1997) A cognitive neuropsychological model of limb praxis and apraxia. In: Heilman KM and Rothi LJG (eds) *Apraxia: The Neuropsychology of Action*, pp. 29–49. Hove, UK: Psychology Press.

Roy EA, Black SE, Blair N and Dimeck PT (1998) Analysis of deficits in gestural pantomime. *Journal of Clinical and Experimental Neuropsychology* **20**: 628–643.

Roy EA, Heath M, Westwood D *et al*. (2000) Task demands and limb apraxia in stroke. *Brain and Cognition* **44**: 253–279.

Schnider A, Hanlon RE, Alexander DN and Benson DF (1997) Ideomotor apraxia: behavioral dimensions and neuroanatomical basis. *Brain and Language* **58**: 125–136.

Seddoh SA, Robin DA, Sim HS *et al*. (1996) Speech timing in apraxia of speech versus conduction aphasia. *Journal of Speech and Hearing Research* **39**: 590–603.

Attention, Neural Basis of

Introductory article

Peter De Weerd, University of Arizona, Tucson, Arizona, USA

CONTENTS

Attention, the ability to selectively process sensory information, is generated by the modulation of sensory processes by frontoparietal neural networks.

INTRODUCTION

The neural basis of attention resides in the ability of structures in the brain representing the behavioral relevance of stimuli to alter sensory processing, so that relevant stimuli are processed effectively and irrelevant ones are ignored. Attention, however, is not a unitary cognitive ability. Attention can alter the sensory processing of stimuli in all sensory systems. In addition, there are different kinds of attention which are related to specific behavioral requirements in different cognitive tasks. Thus, depending on the cognitive task one faces, and the sensory systems involved, different neural networks in the brain will become active during attentional operations.

Definitions of Attention

At the end of the nineteenth century the psychologist William James described attention as a 'concentration of consciousness', with the purpose of 'withdrawing from some things in order to deal effectively with others'. An experimental demonstration of that idea was given by Hermann von Helmholtz in 1894. He filled a large screen with letters, which were spread too widely to be read at once without moving the eyes. During a brief illumination of the screen with a flash of light, which made it impossible to use eye movements to explore the screen, he measured the number of letters that could be read. Only a few letters (up to about five) could be read during each flash. Interestingly, the letters that could be read were not necessarily located at the center of gaze. Letters in any region of the screen away from the center of gaze could be read if the region of interest was selected in advance. This experiment demonstrates several aspects of attention which are still being investigated today: the dissociation of directed attention from eye gaze ('covert' attention), the selective nature of attention, and the limited capacity of attention.

The experiment of Helmholtz indicates that the ability to selectively attend to information is an answer to limitations in sensory information processing. The ability to select information also avoids our behavior being determined by the strength of random stimuli in our environment. Behavior otherwise would inevitably become uncoordinated and chaotic. Hence, the ability to select information is crucial, because it allows us to willfully give precedence to behaviorally relevant information, at the cost of irrelevant information. The 'grabbing' of attention by powerful stimuli is referred to as 'bottom up' (exogenous) attention, and the intentional selection of behaviorally relevant stimuli is referred to as 'top down' (endogenous) attention.

There are different kinds of top-down attention. Selective attention, the main topic discussed here, is the ability to detect or discriminate relevant stimuli (targets) in the presence of competing irrelevant stimuli (distracters). Selective attention is required to find a familiar face in a crowd at a party. A second attentive process is vigilance: the ability to orient attention and respond to randomly occurring, relevant events in the environment over an extended period. A guard who is keeping an eye on the entrance of a building, monitoring for suspicious activity throughout the night, is carrying out a vigilance task. In addition, it has been proposed that there are attentional functions that maintain 'executive control' over goal-directed behavior and thought processes. Executive control functions are especially important when stimuli contain conflicting information. The onset of reading of the word 'green' will be faster when the word is printed in green than when it is printed in red, a phenomenon known as the Stroop effect. Applying

a level of analysis to a stimulus that contains conflicting information, appropriate for the required behavioral response, is a function of executive processes.

Circuits of Selective Attention in the Visual System

The ability to attend to a stimulus is best understood as an emergent property of the functional architecture of sensory systems. This concept will be explored within the context of the visual system (Figure 1).

The primary visual area (V1) receives retinal information through the lateral geniculate nucleus, a relay station in the thalamus. Following lesion and anatomical studies in the monkey, Ungerleider and Mishkin proposed in 1982 that forward projections from V1 give rise to two visual processing streams. One is directed ventrally to the temporal lobe, and is involved in the analysis of objects (ventral pathway, or 'what?' pathway). The other one is directed dorsally to the parietal lobe, and is involved in the analysis of locations of objects relative to each other and relative to the observer (dorsal pathway, or 'where?' pathway). A relative segregation between ventral and dorsal streams is maintained in their projections to the prefrontal cortex in the frontal lobe (Figure 1). Prefrontal cortex is involved in the maintenance and manipulation of information in working memory, response selection and a variety of executive processes.

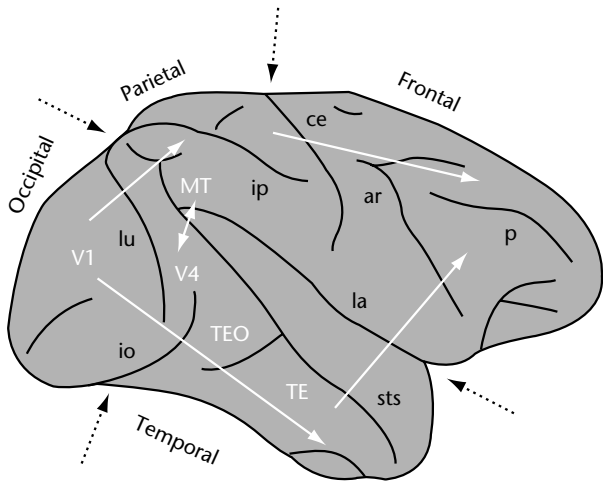


Figure 1. A neural network for attention in the primate brain. Shown is a lateral view of the brain of a rhesus macaque, with the back of the brain towards the left. Dotted arrows correspond to the boundaries between occipital, temporal, parietal, and frontal lobes. White labels indicate visual areas, white arrows indicate anatomical pathways. Visual area V1 gives rise to a ventrally directed pathway which includes V2 (buried in the lunate sulcus), V4, and areas in the temporal lobe (TEO and TE). It also gives rise to a dorsally directed pathway which includes V3 (buried in the lunate sulcus), MT (buried inside the superior temporal sulcus), and areas in the parietal lobe. There is cross-talk between areas belonging to these two visual processing pathways (e.g. double arrow between V4 and MT). Roughly 40 visual areas have been described, and only a few are indicated here. Projections from parietal and temporal lobes towards the prefrontal region of the frontal lobe retain a significant degree of segregation. Forward projections (white arrows) from one to the next area are always returned by backward projections (not shown). ar, arcuate sulcus; ce, central sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus; la, lateral sulcus; lu, lunate sulcus; p, principal sulcus; sts, superior temporal sulcus.

Each of the two processing streams is composed of a number of areas, which are hierarchically organized. Lower-order areas (closer to V1) each contain an orderly retinotopic map of the environment, and are composed of neurons with small receptive fields (RFs), which process simple aspects of the stimulus, such as the orientation of edges. Higher-order areas do not show retinotopy, their neurons have large RFs, and they process complex aspects of the stimulus, such as its shape. Owing to this hierarchical organization, lesions at low levels result in severe sensory deficits, while lesions at intermediate and higher levels leave elementary sensory processes intact and interfere in more subtle ways with visual processing. Lesions at higher levels of the ventral stream induce a disorder in visual object recognition that is referred to as 'agnosia'. Parietal damage causes severe spatial deficits. Because these lesion effects reflect at least in part attention deficits, they are discussed in the following section.

When we look at an object it is automatically perceived as a whole, with a particular location, color, shape, and orientation. Because different aspects of a stimulus are processed in a distributed manner in different cortical regions, this raises the question of how the different features of a stimulus are combined in order to generate a holistic perception. This question is referred to as the 'binding' problem, and the quest for a solution to that problem is intimately related to the quest to understand mechanisms of selective attention. A fundamental insight that will help to resolve both issues is the fact that the separation of the different processing streams is only relative: at various levels in the system, there are anatomical connections between pathways that permit cross-talk. Furthermore, feedforward projections are returned by feedback

projections. Hence, neural activity in a given area can be modulated by activity in other areas or structures to which it is connected. These modulatory influences are a functional property that reflects the structural layout of the system, which is crucial to understanding attention and other cognitive abilities.

PARADIGMS OF ATTENTION IN EXPERIMENTAL PSYCHOLOGY

Before explaining how modulation of activity creates the cognitive ability of attention, it is important to review the different paradigms psychologists have used to investigate attention. These paradigms have led to a precise characterization of attentional behavior, and have become the cornerstone of most experiments on attention in the broad field of cognitive neuroscience (Figure 2).

Spatial Cueing

The spatial cueing paradigm was developed by Posner and colleagues to measure the effects of directed attention (Figure 2(a)). Participants are instructed to keep their eyes on a central fixation cross presented on a computer monitor, and to covertly attend to a location in the left or right half of the screen, indicated by an arrow close to the fixation cross (the cue). The cue is followed in time by a target (e.g., a small square) briefly presented away from fixation. The target can be presented in the cued location (valid cue), or in a minority of cases in a location in the other half of the screen (invalid cue). Participants have to make a button response as soon as they detect the target. On trials in which the target is preceded by a valid cue, response times are shorter (about 250 ms) than on trials in which the target is preceded by an invalid cue (about 300 ms). These results support the idea of a 'focus of attention'. After a valid cue, participants move their focus of attention to the expected location in anticipation of target presentation, and target detection benefits from the presence of attention at the cued location. After an invalid cue, attention moves to the invalid location, and the target is presented outside the focus of attention. The associated cost, an increased reaction time, may be due to the reorienting of the attentional focus to the target.

The reorienting of attention to a location in response to a cue is controlled by the participant, and

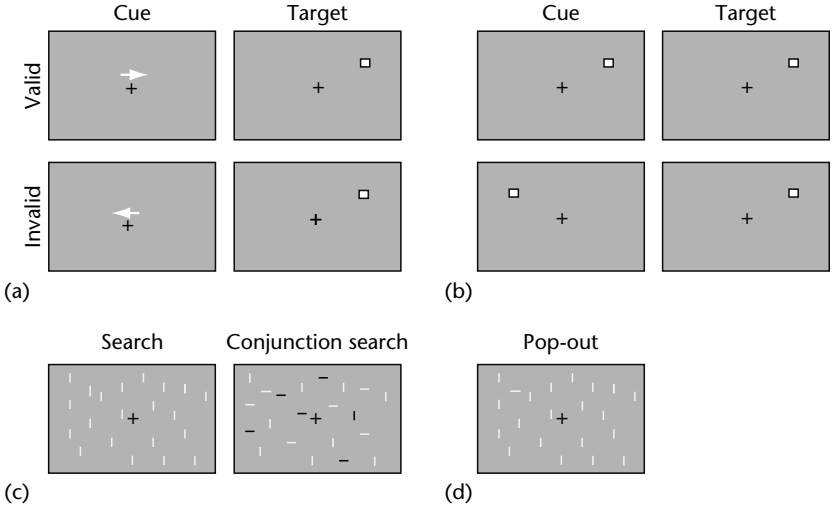


Figure 2. Paradigms in attention research. (a, b) Spatial cueing paradigms. (a) The top two panels show a symbolic cue (arrow) close to a fixation spot (cross) on a computer monitor (gray rectangle) which validly predicts the location of the target (square), presented briefly after offset of the cue. The bottom two panels show target presentation after an invalid cue. When cues are valid on most trials, attention is shifted to the cued location in anticipation of target presentation, and this speeds up responses to the target. (b) Valid (top row) and invalid (bottom row) cueing of attention by a physical stimulus. Even if half of the cues are invalid (rendering the cue unpredictive of target location), the cue has specific attention effects on targets presented in the cued location. (c, d) Search paradigms. (c) Searching for a target among distracters is slow and time-consuming when the target is similar to the distracters (left panel), or if the target is defined by a conjunction of features on more than a single stimulus dimension (right panel). (d) Search is fast and seemingly effortless (pop-out) when the target differs from distracters on a single dimension, and when the difference is large.

this type of cueing is referred to as endogenous cueing ('top down' attention). On the other hand, attention is often automatically oriented to new stimuli. When a student enters class late, it is almost inevitable that everybody turns attention to the latecomer. This exogenous cueing effect ('bottom up' attention) has been investigated in a variation of the spatial cueing paradigm (Figure 2(b)). In this variation, small cues are presented in random locations on the computer screen while the participant fixates a small cross in the middle of the screen. The cue location is unpredictive for the location of the subsequently presented target stimulus. Nevertheless, the cue affects processing of the target if the target happens to be presented in the location of the preceding cue. This effect depends upon the time interval between the cue and target presentation. A target that follows a cue within 250 ms will be responded to with a decreased reaction time, while a target lagging behind the cue by more than about 300 ms will be responded to with an increased reaction time, compared with target presentations not preceded by a cue.

Thus, the mere presentation of a stimulus in a given location can attract attention to that location for a limited time. After that time, attention is disengaged from the stimulus location, and a reorientation of attention to the same location is suppressed. This effect is referred to as 'inhibition of return'. It is important that attention can be attracted easily to unexpected but significant events in the environment, but it is equally important that attention then can be freed from that event to make it available for new, potentially important stimuli. If there were no easy disengagement of attention, attention could get stuck to a stimulus, and new incoming stimuli might go unnoticed. The brief period during which attention is engaged in the processing of a particular stimulus, and during which limited or no attention is available for the processing of new stimuli, is often referred to as the 'attentional blink'.

Search Paradigms

The metaphor of a 'focus of attention' has also been used to explain performance in visual search tasks, a second major paradigm to study attention. In a search task, participants are presented with displays filled with distracter stimuli, which may or may not contain a specific target. Participants are instructed to respond with a button press as soon as they find the target, or to press another button to indicate its absence. Performance is assessed by measuring the time required to find the target as a function of the number of distracters in the display. The resulting plot is referred to as a 'search curve'. A typical experiment suggesting serial search would show an increase in search time by about 30 ms per item.

Search performance becomes dependent on the number of distracters in the display when the target is difficult to distinguish from the distracters (Figure 2(c)). An example of such a search display is one in which the target is a line element slightly tilted away from vertical, surrounded by vertical distracters. Another example is one in which the target is defined by a conjunction of features: the target could be a vertical red line, surrounded by horizontal red, vertical green, and horizontal green lines. The latter task (conjunction search) is difficult because different properties of the objects have to be analyzed and combined before it can be determined whether the object matches the target being searched. The time-consuming nature of these types of search suggests that a focus of attention is continually relocated during search to scan items in the display serially, and in the case of conjunction search also suggests that focal attention plays a role in solving the binding problem.

While search displays have been used to demonstrate serial search, 'pop-out' displays have been used to demonstrate the existence of 'parallel search'. In pop-out displays, the target (e.g. a horizontal line) differs strongly from the surrounding distracters (e.g. vertical lines) and the time required to detect the target appears independent from the number of distracters (Figure 2(d)). The identification of a target in pop-out displays is often considered to occur 'preattentively', or to result from a parallel attentional operation.

A COGNITIVE NEUROSCIENCE APPROACH TO ATTENTION

While the experimental paradigms introduced above characterize important aspects of attentional behavior, they do not address the important question of how modulations of sensory processes produce attentive behavior. Recent neurophysiological findings have brought us closer to an answer.

Findings in the Temporal Lobe

In a series of ground-breaking recording studies, Desimone and colleagues flashed pairs of stimuli inside the RF of single neurons of temporal lobe areas V2, V4, and TE in monkeys trained to identify one of the stimuli and ignore the other.

The neurons' responses were determined by the attended stimulus (target), while the influence from unattended stimuli inside the RF upon the neurons' responses was greatly attenuated (Figure 3). In the period during which the monkeys were waiting for the presentation of each stimulus pair, increased baseline activity was observed compared with the spontaneous firing rate. Depending on the task, increased baseline activity might represent the instruction to the monkey to attend in a particular location, or to expect a stimulus of a given identity, and it might facilitate processing of the target stimulus.

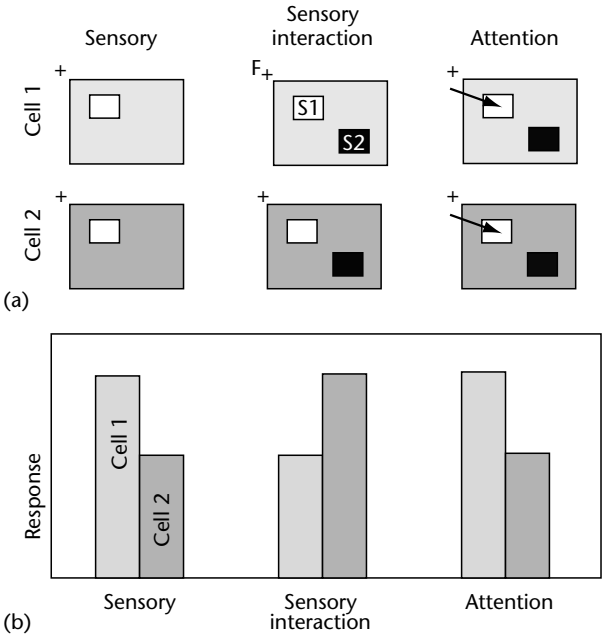


Figure 3. Elimination of sensory interaction (competition) by attention. (a) Different configurations of stimuli and covert attention inside the receptive fields (RF) of two cells (cell 1 and 2). The monkey fixates a cross (F) while covertly attending (arrow) or ignoring stimuli away from fixation placed in the RF of cells 1 or 2, whose activity is recorded. The RFs of cell 1 and 2 are shown as the light and dark gray squares, respectively. Stimuli are symbolized by small white (S1) or black (S2) squares. (b) Pattern of responses illustrating competition and effects of attention. For cell 1, stimulus S1 is more effective than S2 (response to S2 alone not shown). Because stimuli compete for control over the firing rate of the cell, the addition of the less effective stimulus (S2) to the more effective stimulus (S1) drives the response down (sensory interaction). That suppressive effect is eliminated when attention is covertly directed to S1 (arrow in A). Cell 1 now responds as if only S1 were present in its RF, despite the presence of S2 in the RF. Thus, attention eliminates competitive effects caused by behaviorally irrelevant stimuli. In cell 2, stimulus S2 is more effective than S1, and adding S2 to S1 in the RF increases the response compared with the response obtained with S1 alone. Attention to the less effective stimulus S1 will screen out the effect of the more effective stimulus S2, resulting in a decreased response. Thus, whether attention will increase or decrease the activity of a given neuron depends upon the selectivity of the neuron for the stimuli in its RF.

These data provide a neural foundation for the behavioral finding that target stimuli are processed more efficiently than nonattended distracters. In 1995, Duncan and Desimone proposed a 'biased competition' model of attention to account for these findings. In this model, objects presented simultaneously in the visual field compete with each other for control over the firing rate of cells within whose RF they are presented. The competition is decided in favor of stimuli that receive a biasing signal which can be generated 'bottom up' (by a salient stimulus) or 'top down' (by an instruction to identify a particular target among distracters, or to monitor objects in a particular location). The increased spontaneous activity reported during attention tasks may be a correlate of such a biasing signal. Thus, the process of attention is viewed as a biased competition between populations of cells representing different objects or locations (Figure 4).

In agreement with the idea that temporal lobe areas are important for attention, cortical lesions of monkey areas V4 and TEO in the temporal lobe cause an inability to discriminate properties of target objects surrounded by distracters, especially when the target is weak (e.g. smaller or dimmer than the distracters). Without distracters, deficits in discriminating the target are absent or limited. These lesion deficits in the monkey resemble agnosia, a neurological syndrome found after occipitotemporal damage in humans. Humans with agnosia have trouble recognizing visually presented objects embedded in a complex scene, or visually presented complex objects consisting of multiple parts; yet simple stimuli presented by themselves can be discriminated with great accuracy. These general characteristics of agnosia suggest that attention deficits contribute to the syndrome. In the terminology of Duncan and Desimone, agnosia may reflect, at least in part, a failure of biased competition, because of damage to the substrate in which the competition takes place.

Findings in the Parietal Lobe

The description of attention as a process of biased competition leads to the question: where does the

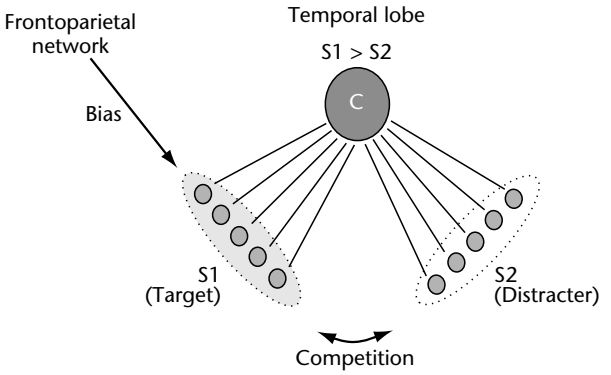


Figure 4. Competition in the temporal lobe is biased by signals generated in a frontoparietal network. Cell C receives input from a population of cells coding S1, and another population of cells coding S2. Each population provides a mixture of excitatory and inhibitory inputs to cell C. The balance of excitatory and inhibitory inputs to C provided by each population determines the size of C's response to S1 and S2 presented alone in C's RF. The response to S1 and S2 presented together in C's RF is determined by the total balance of all excitatory and inhibitory inputs from both populations taken together (competition). When one of these two stimuli (e.g. S1) is made behaviorally relevant (a target), and is therefore attended to, a biasing signal renders inputs from cells coding the target more efficient compared with the distracter. As a result, C's activity will reflect the balance of excitatory and inhibitory inputs provided by S1 alone, rather than the balance of excitatory and inhibitory inputs provided together by the two stimuli in the RF, and S1 wins a competition with S2 for control over the firing rate of cell C (S1 > S2). 'Bottom up' stimulus-driven bias can have effects similar to 'top down' bias generated in a frontoparietal network. This formalization of biased competition theory was first proposed by Reynolds and Desimone in 1999.

bias come from? To answer this question, we turn briefly to a parallel line of work on attention in the parietal cortex. Mountcastle in 1976 and Wurtz in 1982 showed that neurons in parietal cortex of the monkey were more active when the stimulus in their RFs was behaviorally relevant (attended) than when it was not. Several imaging studies in humans have now confirmed that covertly orienting one's attention to a stimulus or location leads to enhanced activity in parietal cortex. Further neurophysiological work in monkeys suggests that the effects of behavioral relevance on parietal activity can depend upon the type of action planned by the monkey towards the target (e.g. an arm versus an eye movement).

Lesions in the parietal cortex induce a complex set of deficits in the representation of space and action. Unilateral parietal damage leads to neglect of the contralateral side of the body, and of stimuli in contralateral extrapersonal space. Neurologists often use the line bisection test to assess neglect: patients with neglect will bisect a horizontal line towards one end of the line, as the other part of the line will become neglected when they look at it. When a strong stimulus is presented on the neglected side, patients can recognize the stimulus perfectly, but when two such stimuli are presented, one in the neglected hemifield and one in the other hemifield, then the stimulus in the neglected hemifield will remain unnoticed, a phenomenon referred to as *extinction*. Extinction and neglect do not always co-occur after parietal damage, indicating that both phenomena may constitute dissociable syndromes after different types of parietal damage.

Using a spatial cueing test, Posner and colleagues found that patients could shift attention to the neglected hemifield when cued correctly, and engage in attentional operations in that hemifield, but had trouble redirecting attention from the ipsilateral to the contralateral (neglected) hemifield. It was therefore postulated that parietal lesions induce a failure of a disengagement operation. Other experiments suggest that the pulvinar nucleus, a thalamic nucleus that provides weak but direct projections to the parietal lobe, plays a part in the shifting and engaging of attention.

Bilateral parietal damage leads to a constellation of symptoms, including optic ataxia (deficit in visually guided reaching), paralysis of gaze, inattention to peripheral stimuli, and sustained hyperattention to single objects or locations (Balint syndrome). Patients with this syndrome do not know where they are, and have no idea of spatial relations between objects. However, once attention is directed to an object, they can identify it perfectly. In pop-out displays, their report of the presence of the target is not influenced by the number of distracters, as in non-afflicted individuals, but they cannot tell where the target is once it has been reported. In conjunction search tasks, people with Balint syndrome require extraordinary amounts of time to find the target (about 1 s per item). In addition, they often make 'illusory conjunctions' (e.g. joining the orientation of a given item with the color of a neighboring item). These illusory conjunctions can also be demonstrated in non-afflicted individuals when the search display is presented very briefly (and followed by a mask). In both situations, there may be not enough attention available to focus accurately on one of the items.

A compelling account of the function of focused attention was given in 1980 by Treisman and Gelade in a model referred to as the 'feature integration' theory. According to this theory, different elementary features in the image (e.g. color, orientation, motion) are analyzed in separate feature maps – an idea at least in part supported by physiological evidence – and the read-out of activity within single maps occurs fast and in parallel. This would explain simple pop-out phenomena. However, when targets have to be identified that are defined by a combination of two or more features, the activity in different feature maps must be combined, which is a lengthy operation requiring focused attention. Hence, the time-consuming nature of conjunction search may reflect a mixture of factors, including the disengagement, orienting, and engagement of attention, as well as the time required to conjoin features. Imaging data support the idea that parietal cortex is involved in the directing of attention to targets, in the shifting of attention in response to both endogenous and exogenous cues, and in conjunction search, but not in feature search or pop-out. Since conjunction search implies shifts of attention and conjoining of features, parietal activity could indicate either or both.

Findings in the Frontal Lobe

An additional line of research that is related to the question where the biasing signal comes from was initiated by Fuster and Alexander in 1971. They found that neurons recorded around the principal sulcus in prefrontal cortex showed enhanced activity during delayed-response tasks, in which the monkey was required to remember a previously cued location that would become the target of a directed action after a delay of a few seconds. Later research has demonstrated that neurons in prefrontal cortex can be activated during tasks that require working memory for both location and object identity.

Many attentional operations require working memory. For example, while a participant looks for a target in a search array, incoming sensory information is matched against a template of the target that is kept on-line in working memory. Many physiological and imaging studies have confirmed that working memory and attention both rely heavily on prefrontal cortex. Limitations on attentional capacity, witnessed by the impossibility of identifying more than a few targets at once, are reminiscent of limitations to hold more than a few stimuli in working memory.

Modulation of Competition in the Occipitotemporal Lobe by a Frontoparietal Biasing Signal

Competition in the occipitotemporal lobe

The term 'competition' refers to the sensory interactions that take place automatically when two stimuli S1 and S2 are placed inside a cell's RF. If this cell (C) receives inputs from a pool of neurons that codes S1, and another pool that codes S2, then the response of the cell will be determined by a mixture of the influences of S1 and S2. Attention biases the efficacy of one set of inputs relative to the other set, such that the activity of the cell will reflect preferentially the attended stimulus (Figure 4). Cells encoding a target could enhance their influence on the activity of cell C by slightly depolarizing their cell membranes. This enhances the probability that these cells will generate action potentials whenever the target is presented in their RFs. Alternatively, the pool of neurons that represents the target may synchronize its spontaneous and stimulus-driven activity. This enhances the probability of coinciding spikes in the pool of neurons coding the target. Because of the summation properties of neurons, rate enhancements and synchronization could both increase the control of the pool of neurons coding the target over the activity of the postsynaptic cell C. The pool of neurons coding the target is thought to include neurons coding various aspects of the stimulus, distributed in temporal lobe areas and other parts of the brain. The synchronization of the activity of neurons coding different properties of an object could be a mechanism for the binding of those object properties, a proposal championed by Singer and colleagues.

A frontoparietal source for the biasing signal

Where does the biasing signal come from? Alternatively, what makes a stimulus behaviorally relevant? A stimulus is behaviorally relevant when there is a requirement to make a saccade to it, to reach and grab it, and to keep it in working memory. It is precisely under those conditions that specific regions in parietal and frontal cortex become active. Thus, to the extent that requirements in an attention task engage frontoparietal regions, activity in those regions could bias activity in ventral stream areas through connections between retinotopically matched regions in ventral and frontoparietal areas.

Evidence suggests that parietal activation during endogenous directing, shifting, and maintenance of

spatial attention may be controlled by the prefrontal cortex, while parietal activity related to exogenous cueing may be controlled by subcortical thalamic input to the parietal cortex. From the perspective of a theory of binding, the bias would be considered part of the representation of the attended object. Indeed, a complete object representation would consist of a synchronized ensemble of neurons distributed in the frontal, parietal, and temporal lobes, representing the object's identity and location, and possible actions towards it. Similarly, the appropriate binding of features (e.g. color and orientation) in conjunction stimuli may be related to the synchronization of activity in neurons coding the color and orientation of those stimuli by a common biasing signal. Thus, selective attention, synchronization, and binding may be intimately related.

Implications for the Serial versus Parallel Search Debate

Attentional bias signals affect sensory processing at the earliest levels of sensory systems. This settles a long debate between 'early selection' proposals (pioneered by Broadbent in 1958) and other proposals advocating 'late selection'. Not unlike feature integration theory, early selection theory predicts that nonattended stimuli are incompletely processed, and that full processing of all aspects of a target stimulus requires focusing of attention, which implies that search must have a serial component. However, strong arguments have been made against the existence of serial search. It has been argued that target search can be mediated by biasing sensory processing towards the target everywhere in the visual field in parallel. Increases in search time as a function of increases in the number of distracters would merely reflect a thinner spread of a limited amount of attention over all items in the display.

Evidence is mounting, however, that performance in search tasks depends on a mixture of serial and parallel operations. For instance, it has been reported that detection of a 'pop-out' target is impaired when an attention-demanding task is carried out at fixation during stimulus presentation. This does not exclude a parallel component in the detection of pop-out, but it does argue against the idea that pop-out is entirely preattentive. Thus, some degree of directed attention may have a role, even in pop-out. Furthermore, performance in spatial cueing tasks, designed to reveal serial relocations of a focus of attention, is influenced by a segmentation of visual space into surfaces and objects, which takes place in parallel across the field. Specifically, experiments by Driver and colleagues show that reorienting attention to a target location, after an invalid positional cue, occurs faster if the invalid cue and target are both presented within the same shape, compared with when the target and cue are presented in different shapes. Although those experiments are set up to equate the distance over which attention has to be reoriented, reorientation takes significantly longer between than within objects. Hence, spatial attention does not operate within unsegmented space.

Findings in patients with neglect due to unilateral parietal damage support this hypothesis. Neglect is often described as a purely spatial imbalance in the distribution of attentional operations between hemifields. However, experiments in a patient with right parietal damage showed that details on the right side of an object placed in the left (neglected) hemifield are perceived more accurately than details on the left side of an object placed in the right (normal) hemifield. The idea that parietal cortex manipulates the distribution of attention in an object-driven way, rather than a purely spatial way, is compatible with the role of parietal cortex in representing potential actions towards target objects.

In sum, processing of a visual scene may begin with a fast quasiautomatic segmentation based on parallel processes that require little attention. Subsequently, attention may be distributed across this roughly segmented scene, in ways that reflect behavioral demands, and which may not be exclusively parallel or serial. When relevant objects are expected in a single location, attentional resources will be focused on that location. When switches of attention between locations are required, this serial redistribution of attention will be influenced by the preceding segmentation of space. Furthermore, when a specific target object is searched for without knowledge of its location, a 'top down' parallel bias may contribute to the identification of potential targets, but does not exclude the possibility that potential targets are serially inspected. Thus, while items in a search display may not be scanned individually, parallel processes (including automatic 'bottom up' segmentation and parallel 'top down' biases) may lead to the determination of likely targets, which may then be inspected serially. The exact distribution of attention in response to task demands is controlled by executive processes. These and similar ideas have led to various mixed parallel–serial models of attention.

CONCLUSION

The combination of behavioral paradigms developed by psychologists and new methodological approaches developed in the field of cognitive neuroscience has greatly enhanced our understanding of attention and of the neural processes that underlie it. Attention can be best understood as a modulation of sensory processes driven by regions in the brain that represent the behavioral relevance of stimuli. These modulations ensure that behaviorally relevant objects (or locations) in the environment are preferentially processed at the cost of irrelevant ones.

Further Reading

Andersen RA, Snyder LH, Bradley DC and Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–330.

Broadbent DE (1958) *Perception and Communication*. London, UK: Pergamon Press.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–349.

Desimone R and Duncan J (1995) Neuronal mechanisms of selective attention. *Annual Review of Neuroscience* **18**: 193–222.

Desimone R and Ungerleider LG (1989) Neural mechanisms of visual processing in monkeys. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 2, pp. 267–299. New York, NY: Elsevier.

Desimone R, Wessinger M, Thomas L and Schneider W (1990) Attentional control of visual perception: cortical and subcortical mechanisms. In: *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 55, pp. 963–971. Cold Spring Harbor Press.

Farah MJ (1990) *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. Cambridge, MA: MIT Press.

Gazzaniga MS (1995) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Gazzaniga MS, Ivry RB and Mangun RM (1998) *Cognitive Neuroscience: The Biology of the Mind*. New York, NY: WW Norton.

Hillyard SA and Picton TW (1987) Electrophysiology of cognition. In: Plum F (ed.) *Handbook of Physiology*, Section 1: The nervous system, vol. 5, Higher functions of the brain, part 2, pp. 519–584. Bethesda, MD: American Physiological Society.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* **24**: 167–202.

Parasuraman R (1998) *The Attentive Brain*. Cambridge, MA: MIT Press.

Pashler HE (1999) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Singer W and Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–586.

Treisman A (1999) Solutions to the binding problem: progress through controversy and convergence. *Neuron* **24**: 105–110.

Wolfe JM (1994) Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review* **1**: 202–238.

Audition, Neural Basis of

Introductory article

Shihab Shamma, Institute for Systems Research, University of Maryland College Park, College Park, Maryland, USA

CONTENTS

Sound is translated into neural signals by the organs of the auditory system. Key to this process is the cochlea, which converts sound pressure waves into spatially ordered patterns of membrane vibrations and then transforms these vibrations into neural patterns in the auditory nerve.

INTRODUCTION

Sounds in the environment are produced when a force excites a structure and causes it to vibrate. Sounds may be sudden and non-repetitive (e.g. a clap or the snapping of a twig), sustained and irregular (e.g. the burbling of water or the whisper of a friend) or sustained and regular (e.g. a singing voice or the rhythm of a drumbeat). In all cases the emitted sound carries information about the exciting force, the vibrating structure, and other physical features, such as the resonating chambers that modify the sound on its way to our ears. Our auditory systems, and those of most other animals, have evolved ingenious ways of extracting all of this information and they use it to detect, locate and identify sound sources of danger, food, courtship and companionship.

THE NATURE OF SOUND CUES

When the driving force causes a structure to vibrate (e.g. a plucked string), or is itself repetitive in nature, the sound produced consists of a succession of almost identical copies of the same emission. The fact that the emissions are so similar is an excellent clue to their common origin, so we might expect organisms to have evolved special mechanisms for recognizing repetition. Furthermore, the time intervals between repetitions are meaningful, as they characterize both the driving force and the vibrating structure of the sound source. When repetitions are slow (less than 50 repetitions per second – or Hz), they are heard as distinct sounds that reflect the dynamics of the driving force. For example, the sequence of notes in a melody or the rhythmic tapping of a drum result from finger movements, and the succession of speech sounds in an utterance reflect movements of the mouth. When the dynamics or frequency of the driving force become faster (50–4000 Hz), we begin to perceive a sustained sound with a distinct pitch that is proportional to the frequency of vibrations. This percept is critical to our appreciation of melody in music, and to the recognition of human voice. Finally, excited structures often exhibit complex patterns of vibrations consisting of numerous simultaneous frequencies that may extend to very high rates (even exceeding the audio range of humans at 20 000 Hz). These complex vibrations evoke distinct sound qualities (or timbres) that reflect the shape of the structure (e.g. a violin versus a cello), its material composition (e.g. wood versus brass) and dynamics (e.g. a muted versus a free string).

COCHLEA FUNCTION

When sound reaches the ears as pressure waves in air or water, it causes the eardrum to vibrate. It in turn transmits the vibrations to the cochlea of the inner ear via an attached chain of three tiny bones located in the middle ear. The cochlea is the key hearing organ. It consists of an elongated fluid-filled cavity with elaborate sensory cells that are embedded within exquisitely sensitive membranes that extend along its entire length. The cochlea has two main functions as illustrated in Figure 1.

First, it converts the sound pressure wave into a spatially ordered pattern of membrane vibrations. Specifically, the mechanics of the cochlear membranes gradually change their electromechanical properties along its length in a manner that causes different places to vibrate best (or be tuned) to

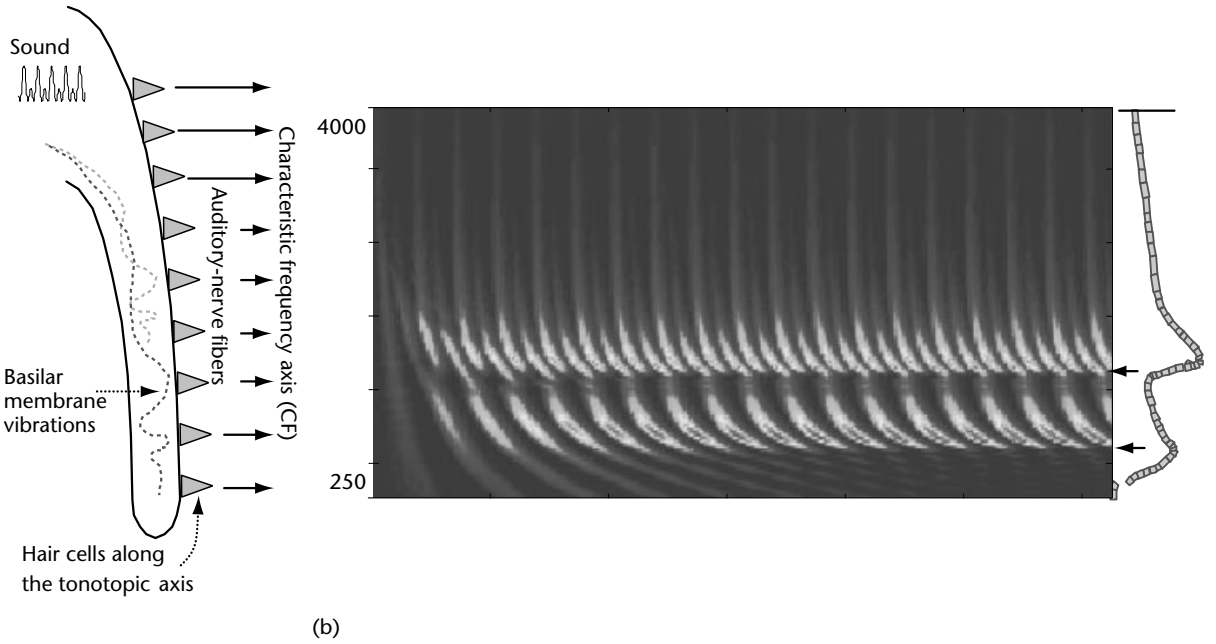


Figure 1. [*Figure is also reproduced in color section*.] Schematic model of the early stages in auditory processing. (a) Cochlear analysis. Sound enters the cochlea via the eardrum and the middle ear, initiating travelling wave displacement patterns on the basilar membrane. Vibrations due to low frequencies propagate and achieve their maximum amplitude further down the cochlea (broken red line) compared to high frequencies (broken green line). This mapping of sound frequency components onto different places along the cochlea creates the tonotopically ordered (spatial) axis of the auditory system. Basilar membrane vibrations are transduced into spatiotemporal responses on the auditory nerve by an array of hair cells distributed along the length of the cochlea. (b) Auditory nerve responses. The spatiotemporal response patterns in the auditory nerve due to a two-tone stimulus (300 and 600 Hz). The ordinate represents the tonotopic axis (labeled by the characteristic frequency axis (CF) at each location). The response at each CF represents the instantaneous probability of firing in the nerve fiber at that CF. Note that each component in the stimulus initiates a localized travelling wave pattern that abruptly ends creating a prominent discontinuity near the appropriate CF (marked by the arrow heads to the right of the panel). The stimulus responses depicted here are at a low sound level, such that it does not saturate the nerve responses. The response amplitudes are therefore strongest near the CFs of the two tones, resulting in clear peaks in the average response curve (red plot to the right).

different frequencies. Consequently, sounds with very high frequencies cause large membrane vibrations near the entrance to the cochlea, whereas those with low frequencies cause maximal vibrations near the end of the cochlea. In this way, the cochlea effectively separates the different components of the sound according to their frequency, sending them off to different places and creating a frequency-organized axis known as the tonotopic axis of the cochlea. Each complex sound therefore creates a unique spatial pattern of strong and weak vibrations along the tonotopic axis that reflects the amplitudes of its different frequency components – or its frequency spectrum.

The second main function of the cochlea is to transform these vibrations into neural patterns on the auditory nerve, to be interpreted by the brain subsequently. This is accomplished by more than 3000 specialized sensory cells (known as the hair cells) that are distributed along the cochlea. Hair cells possess channels that open and close rapidly, modulating the flow of electric current into them. The currents initiate a cascade of electrochemical events, culminating in neural signals on the auditory nerve that faithfully encode the phase (or are phase locked to the time course) of the vibrations at each point up to fairly high frequencies (4000 Hz in some mammals, and 9000 Hz in some birds). Since stronger vibrations also lead to a more vigorous neural response, the auditory nerve in effect encodes the spectrum of the sound both by the level and by the phase-locked structure of the responses along the tonotopic axis.

NEURAL PATHWAYS

The first neural structure beyond the auditory nerve is the cochlear nucleus, which consists of

several anatomically elaborate subdivisions that receive parallel direct projections from the nerve. Multiple pathways emerge from the cochlear nucleus up through the midbrain and thalamus to the auditory cortex, each passing through different neural structures, repeatedly converging on to and diverging from other pathways along the way. This complexity reflects the rich and varied auditory percepts extracted from the sound, the integration of these percepts into a total auditory sensation, and its final fusion with vision and other sensory modalities, and with motor actions.

Although there is still much to be learned about the exact mechanism whereby all of these neural pathways and structures process sound, it is nevertheless clear which signal cues the nervous system must extract in order to give rise to a few important auditory percepts that include loudness, pitch, timbre and sound location.

LOUDNESS, PITCH, AND TIMBRE

Perhaps the most intuitive of these percepts is that of loudness. It is normally associated with increasing the volume (amplitude or intensity) of the sound. However, there is another physical dimension that correlates strongly with loudness, namely the range of frequencies that make up the sound, or its bandwidth. Loudness can therefore be generally viewed as being mediated by the total volume of neural activity in the auditory nerve, and thus increasing the activity (by raising either the sound intensity or the bandwidth) leads to a louder sound.

The sensation of pitch is also a readily understood attribute of sound that is normally associated with musical scales and melodies, or with the low and high voices of men and women, respectively. Pitch is strongly correlated with the repetition rates or frequencies in a sound. However, unlike loudness, the neural basis of pitch is a much more contentious topic, largely because 'pitch' is an imprecise term that is ascribed to multiple sensations which have distinct origins, and most probably different neural mechanisms. For example, the pitch of a pure tone is directly related to its frequency, and is felt over a very broad range of approximately 50 to 20 000 Hz. From a neural perspective, this percept is readily encoded by the location of best tone-evoked response along the tonotopic axis. A second example is so-called rattle pitch, namely a relatively weak sensation that is correlated with the modulation rate of the amplitude of a noise or a tone. This pitch is typically heard only up to a few hundred hertz (< 400 Hz), and is likely mediated by neural responses (in the auditory nerve and beyond) that are explicitly entrained to the modulation rate. The final and most salient sensation of pitch is that of musical instruments and voices. This percept exists over a moderate range of frequencies (< 4000 Hz), and is exclusively associated with harmonic sounds composed of frequencies that are integer multiples of a common fundamental frequency. An interesting fact about this pitch is that its value remains that of the fundamental frequency, even if the fundamental component in the sound is missing (hence the common description of this percept as the 'pitch of the missing fundamental'). That is, the pitch value is derived from the harmonic relationship between the components, and not simply from the fundamental frequency *per se* as illustrated in Figure 2a. The neural basis of this percept remains uncertain. One plausible theory proposes that the brain stores (or learns) harmonic templates of all pitch values, and the percept is derived according to which templates best match the spectrum of the incoming sound (Figure 2b). According to another theory, the pitch is computed directly from the incoming sound without resort to any templates, and as such it does not explicitly distinguish between this percept and the 'rattle' pitch described earlier (Figure 2b). Both of these theories (and many other variations) can account for most relevant psychoacoustical findings. The major missing piece in the pitch puzzle is the lack of firm biological understanding of the mechanisms underlying pitch processing in general.

Timbre is best regarded as the quality of a sound. It is the percept that allows us to distinguish between a violin, an oboe and a piano playing at the same pitch, or to perceive the difference between vowels (or other phonemes) in spoken language. Timbre is a multidimensional percept that is difficult to reduce to a simple scale (e.g. the low-to-high scale of pitch and loudness). Instead, it has been customary to propose several 'descriptive' scales to quantify it, using intuitive notions such as 'sharp to dull' and 'continuant to transient'. However, experimental evidence makes it abundantly clear that the shape and dynamics of a sound spectrum directly influence its timbre. Spectral shape is extracted early in the auditory system, perhaps as early as the cochlear nucleus, where the neural activity of a specific type of neuron has been found to represent faithfully and robustly the acoustic spectrum along its tonotopic axis. Further elaboration of this representation occurs in the midbrain and cortex, where more complex spectral features are selectively detected. The neural correlates of

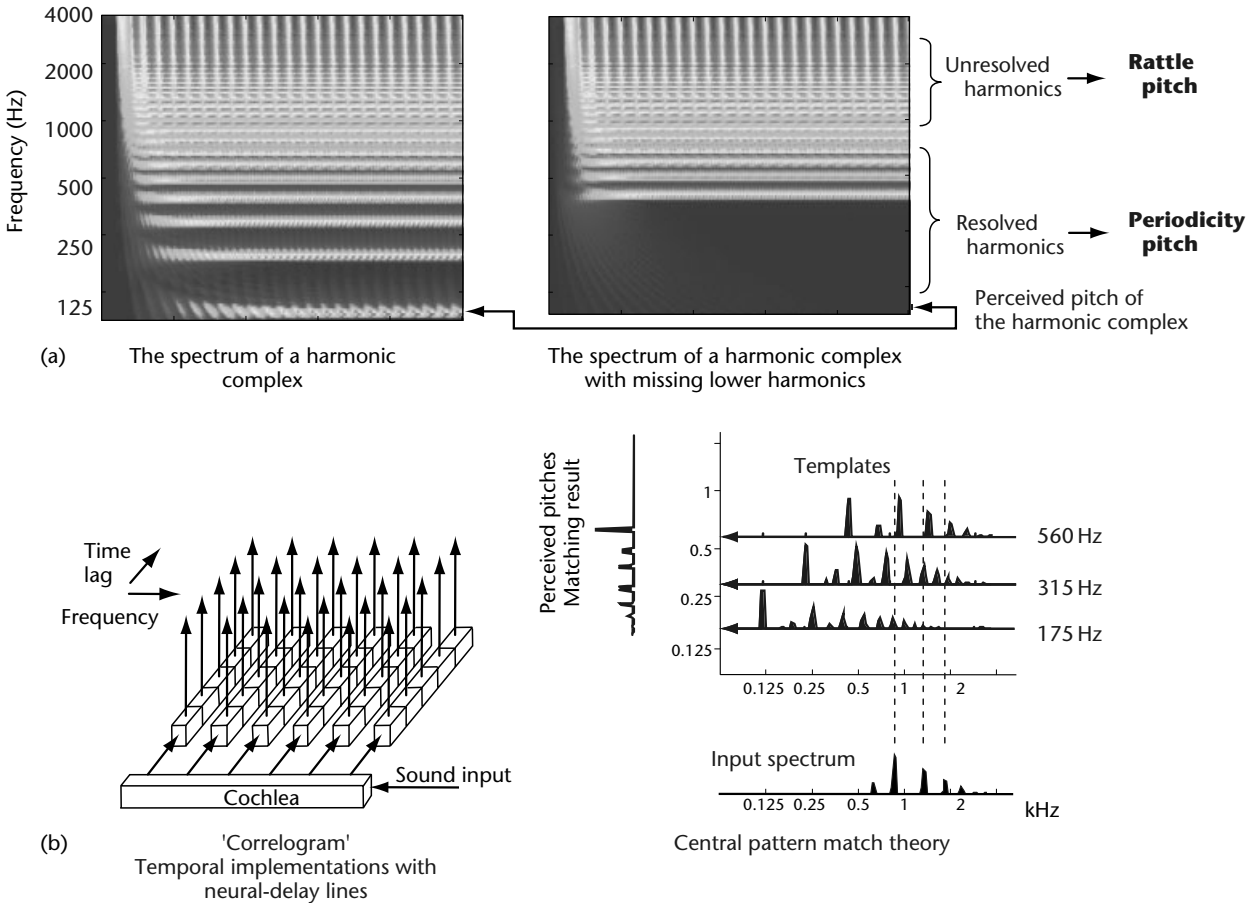


Figure 2. [*Figure is also reproduced in color section*.] Representation and extraction of stimulus periodicity of a harmonic complex. (a) The auditory spectra of two harmonic series of a 125 Hz fundamental. The low-order harmonics are well *resolved* along the tonotopic axis. These harmonics evoke a pitch sensation at the fundamental frequency of the series (i.e. at 125 Hz). The high-order harmonics (> 8th harmonic or 1 kHz) are largely *unresolved* and they instead evoke a pattern that 'beats' at the difference frequency of 125 Hz. These harmonics evoke a weaker pitch (called the 'rattle pitch') at the beating frequency (i.e. 125 Hz in this case). (Left) The harmonic complex here contains all 40 harmonics and evokes a strong pitch at 125 Hz. (Right) The harmonic complex here lacks the lowest three harmonics (125, 250, 375 Hz); nevertheless, it still evokes a strong pitch at the 'missing fundamental' frequency of 125 Hz. (b) Two algorithms for extracting this missing fundamental pitch. (Left) The schematic illustrates an autocorrelogram implementation. It presumes the existence of organized delay lines to compute the autocorrelation of the responses from each auditory nerve channel (or fiber) prior to computing the pitch. (Right) Schematic illustrates a *template-matching idea*. It presumes the existence of harmonic templates in the brain that are matched to incoming spectra so as to measure the pitch.

spectral dynamics also undergo significant transformations in different auditory stages, with faster temporal features (> 50 Hz) mainly being evident in the pre-cortical stages. In the auditory cortex (as in other sensory cortices) the dynamics of the responses explicitly represent the temporal evolution of the sound spectrum over relatively slow rates (< 50 Hz). These timescales are commensurate with the dynamics of the vocal tract in speech, the rate of change in pitch in musical melody, the transient dynamics that differentiate a string that is struck from one that is bowed (e.g. a piano versus a violin), and the rhythms of percussion instruments. Finally, there are other more complex representations of sound combining spectral and dynamic features which have been found to endow cortical cells with elaborate response selectivity. Examples of these include selectivity to downward or upward frequency-swept sounds in bats, to phrases of species-specific calls in primates, to phoneme or phonemic clusters in humans, or even to entire songs in birds. However, it is still unclear what neural mechanisms and architectures give rise to these representations and how, and the

way in which these representations ultimately relate to our perception of timbre.

SPATIAL HEARING

Finally, we consider the perception of sound location in space. This auditory function is critical for survival – for example, in escaping predators, following prey and finding mating partners. Despite the enormous variability in the nature of the cues and mechanisms involved in this task, some principles are common to most species. The first of these is the use of differences in the sound impinging on the two ears, especially differences in time of arrival and in sound intensity. Specifically, when a sound source is centered in front of the head, it reaches the two ears simultaneously. If it moves to the right, the path to the right (relative to the left) ear shortens, and thus the sound reaches the right ear sooner – by a fraction of a thousandth of a second, a detectable difference for many animals (especially those with larger heads). An analogous difference in sound level occurs when a source is closer to one ear, especially for high-frequency sounds where the head shadow is more effective. Most animals have evolved neural mechanisms for detecting, extracting and utilizing these inter-aural differences in order to locate the sound source. For example, there are specialized coincidence cells in all mammals and birds that receive synchronized inputs from the two ears, and which are tuned to detect a particular time-delay between them. In barn owls, such cells are highly organized topographically so as to create an optimum time-delay axis. Another commonly found cell type is tuned to detect specific level differences between the two ears.

The second important localization principle concerns the use of special spectral cues (from one or both ears) to locate the elevation of a sound source or to characterize the acoustic environment. These spectral cues are usually introduced by auxiliary structures such as the pinna (the external ear), the shoulders, and nearby walls and floors. In the case of the pinna, its highly convoluted cavities function as mini-resonators that absorb or amplify certain sound frequencies depending on the direction of arrival of sound. This is extremely useful because, when a source is located on the midline, the two paths to the ears are equal regardless of the source elevation, and thus the only way to localize it is based on pinna-originated spectral cues. In mammals, there is some evidence that specialized neural pathways have evolved as early as the cochlear nucleus to detect these unique cues and process them in conjunction with the binaural cues. Another useful function of certain spectral cues is to convey information about the reverberant qualities of a room, and hence (indirectly) about its size and material structure. This issue is extremely important with regard to both the architectural design of music halls and auditoriums and the assessment of the quality of communication channels and equipment.

CONCLUSION

There are many other auditory attributes and tasks in the animal world that we have not touched upon, and that are as important to these animals as the perception of sound timbre, pitch and location is to us. Examples include the ultrasonic echolocation in bats and dolphins that enables them to locate prey and avoid obstacles in cluttered environments, the infrasonic (very-low-frequency) communication signals among many terrestrial animals (e.g. elephants), and the unique auditory adaptations of many animals that help them to deal with the limitations of their small size (e.g. insects) or their aquatic environments (e.g. fish). Finally, we have not considered a number of key questions. For example, how does the auditory system assemble all of these disparate percepts into an integrated whole that identifies the sound source as an entity amidst the clutter of other simultaneous sound sources in the environment? And how do auditory percepts become integrated with visual, motor and other neural processes of attention and memory so as to give rise to the typical active auditory behaviors that we normally associate with this amazing modality? The answers to these and countless other auditory mysteries lie in the findings of future exciting research.

Further Reading

Blauert J (1996) *Spatial Hearing: The Psychophysics of Human Sound Localization*, translated by J Allen. Cambridge, MA: MIT Press.

Moore BCJ (1997) *An Introduction to the Psychology of Hearing*. London, UK: Academic Press.

Pickles JO (1998) *An Introduction to the Physiology of Hearing*. London, UK: Academic Press.

Autonomic Nervous System

Intermediate article

Gary G Berntson, Ohio State University, Columbus, Ohio, USA
Martin Sarter, Ohio State University, Columbus, Ohio, USA
John T Cacioppo, University of Chicago, Illinois, USA

CONTENTS

The autonomic nervous system has been viewed as a reflexive system for maintaining internal homeostasis. It is now clear, however, that the autonomic nervous system has reciprocal interactions with higher neurobehavioral substrates, influencing both autonomic control and cognitive/behavioral processes.

INTRODUCTION

The autonomic nervous system (ANS) is the designation applied by John Langley to a complex network of peripheral nerves and ganglia, together with associated regulatory systems of the brain and spinal cord, which serve to control smooth muscles and glands of the viscera (Langley, 1921). Implicit in the term 'autonomic' is the view that the ANS is rigidly regulated, and not subject to the vagaries of 'volitional' control like the somatic nervous system. This was even more strongly implied by an earlier name: the involuntary nervous system.

The influential physiologist Walter Cannon argued that the ANS was specialized for what he termed 'homeostasis', or the maintenance of stability of the internal fluid matrix, necessary to sustain life (Cannon, 1939). This function was suggested to be implemented by an array of feedback-regulated autonomic reflexes responding to perturbations in visceral states with compensatory adjustments to restore homeostatic balance. An example is the baroreceptor reflexes, whereby an increase in blood pressure, signaled by baroreceptor afferent activity, triggers reflex responses including relaxation of vascular smooth muscle, as well as decreases in heart rate and myocardial contractility that reduce cardiac output. Together, these responses serve to decrease blood pressure and compensate for the initiating perturbation (Figure 1).

Early research focused on the plethora of autonomic reflexes organized at lower brainstem and spinal cord levels. An important historical development, however, was an emerging recognition that central regulatory systems extended well above these lower levels of the neuraxis. Although many aspects of ANS regulation may be based on simple reflexes, autonomic adjustments are also closely linked to cognitive and behavioral processes that arise from higher neurobehavioral substrates, including the limbic system and cerebral cortex. Another important recognition is the fact that the ANS is as much a sensory system as a motor system. In addition to contributing to lower-level reflexive regulation, visceral afferent information now appears to bias processing in higher neural systems.

ANATOMY AND PHYSIOLOGY

Peripheral Components

In addition to the intrinsic enteric system that is sometimes considered to be part of the ANS, the ANS consists of two major peripheral divisions, the sympathetic and parasympathetic branches. These two branches have distinct central origins, and differ in their peripheral anatomy and physiology. The sympathetic nervous system has its central origins in the intermediolateral cell column of the thoracic and lumbar divisions of the spinal cord (Figure 1), and so has also been termed the thoracolumbar division. Spinal sympathetic motor neurons give rise to preganglionic efferents which exit the spinal cord in the ventral roots and enter an interconnected set of sympathetic chain ganglia which lie along each side of the cord. On entering the chain ganglia, preganglionic fibers may ascend or descend before terminating on sympathetic

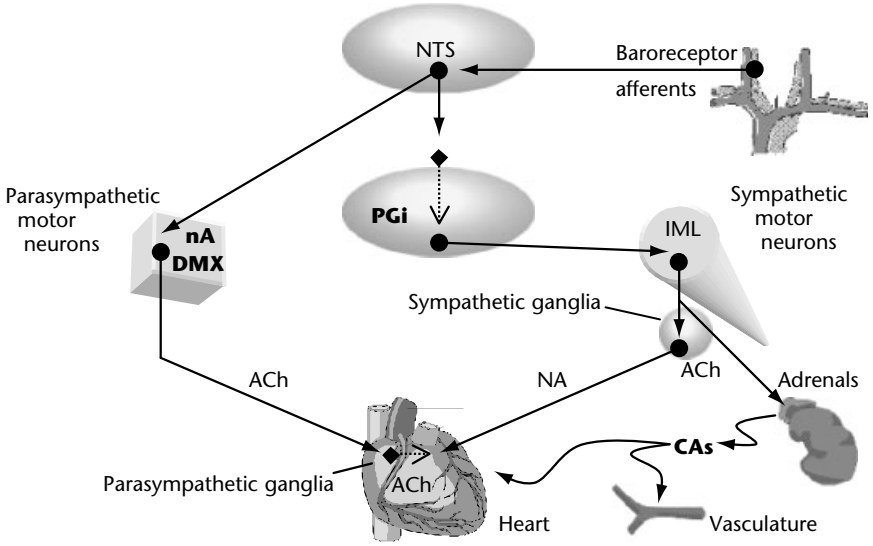


Figure 1. Some features of the baroreflex circuits and peripheral organization of the autonomic nervous system. Baroreceptor afferents project to the nucleus of the tractus solitarius (NTS), by which baroreceptor activity leads to activation of parasympathetic motor neurons in the nucleus ambiguus (nA) and dorsal motor nucleus of the vagus (DMX). The NTS also indirectly inhibits the nucleus paragigantocellularis (PGi) within the rostral ventrolateral medulla, leading to a withdrawal of excitatory drive on the sympathetic motor neurons in the intermediolateral cell column of the cord (IML). ACh, acetylcholine; CAs, catecholamines; NA, noradrenaline (norepinephrine). (Adapted from Cacioppo JT, Tassinary LG and Berntson GG (2000) *Handbook of Psychophysiology*, p. 465, with permission from Cambridge University Press.)

ganglion cells, which give rise to postganglionic axons that in turn project to visceral organs. Because of the extensive ganglionic interconnections within the sympathetic nervous system, this division was often considered to discharge as a whole (i.e., in sympathy). We now know, however, that the sympathetic system can exert much more precise and organ-specific actions. The primary neurotransmitter at the ganglionic synapse is acetylcholine (ACh), whereas the postganglionic synapse employs the catecholamine neurotransmitter noradrenaline (norepinephrine). There are some deviations from this general anatomical plan, as preganglionic fibers innervate the adrenal medulla directly, without synaptic interruption in the chain ganglia. Hence, sympathetic synapses onto the adrenal medulla release ACh, and the adrenal secretory cells release the catecholamines noradrenal and adrenaline (epinephrine). In contrast to postganglionic sympathetic innervation, however, these adrenomedullary catecholamines are released into the general circulation where they can act humorally on many organ systems, including some that do not receive direct innervation. Because many of the peripheral actions of the sympathetic system are activational and promote energetic metabolism, this division has been considered to be a mobilization system involved in responding to adaptive challenges.

The other branch of the ANS, the parasympathetic division, differs from the sympathetic in its central origin, peripheral anatomy, neuropharmacology, and functions. The lower central motor neurons of the parasympathetic division lie in the intermediolateral column of the sacral spinal cord, and in numerous nuclei within the brainstem (e.g. the nucleus ambiguus, dorsal motor nucleus of the vagus, and salivatory nuclei; see Figure 1). Because of this anatomy, the parasympathetic division has been termed the craniosacral branch, and is also sometimes referred to as the 'vagal' branch, after the vagus nerve (10th cranial nerve) that carries parasympathetic efferents. Strictly, however, the latter term applies only to the vagal component of the parasympathetic branch. Like the sympathetic division, the parasympathetic system includes peripheral ganglia, but these are not collected into coherent ganglionic chains, but rather are generally located in or near the visceral organs innervated. Because of this anatomical difference the parasympathetic system has been thought to be capable of more localized action, although considerable regional specificity can also be demonstrated for the sympathetic branch. As in the sympathetic system,

the preganglionic axons of the parasympathetic branch employ ACh as a neurotransmitter (both divisions acting primarily via nicotinic cholinergic synapses on the ganglia); but in contrast to the sympathetic division, postganglionic axons of the parasympathetic system also employ ACh (generally acting via muscarinic cholinergic receptors). This simple distinction in neurochemical coding among the peripheral autonomic branches belies the tremendous complexity of neurotransmitter, neuromodulatory, and neurohormonal interactions within the peripheral ANS.

Opposing versus Synergistic Actions and Modes of Autonomic Control

Many visceral organs are dually innervated by both autonomic branches, and the two divisions are often opposing in their actions. For example, the sympathetic cardiac innervation increases heart rate via β adrenergic receptors that speed the depolarization of the sinoatrial pacemaker potential. In contrast, the parasympathetic innervation slows the beat of the heart via ACh, acting at muscarinic receptors on the sinoatrial node, which increases potassium conductance and slows the rate of pacemaker depolarization. More generally, in contrast to the mobilizing functions of the sympathetic branch, the parasympathetic system has been viewed as a conservation system that functions to promote energy intake, reduce energy expenditure, and preserve energy reserves. Historically, the autonomic branches sometimes have been considered to be subject to reciprocal central control, with increased activity of one division associated with decreased activity of the other (Berntson *et al*., 1991). Indeed, to the extent to which branches are functionally opposed, the reciprocal mode of regulation would yield the widest dynamic range of autonomic control over target organs (Berntson *et al.*, 1991).

All organs are not dually innervated, however, and it is often the case that actions of the two branches on a given organ are not precisely opposite. Major arterioles, for example, receive only sympathetic innervation, and both sympathetic and parasympathetic activity can stimulate salivary secretion. Even when generally opposing in their actions, as in the control of heart rate, the two branches may operate by different cellular mechanisms with distinct features and temporal dynamics. These differences can lead to distinct functional states, with differing levels of activity of the two branches, which cannot be duplicated by simple variations along a reciprocal bipolar continuum extending from maximal sympathetic to maximal parasympathetic activity. Penile erection and ejaculation, for example, require the coactivation of both autonomic branches.

Higher Central Controls and Neurobehavioral Systems

Many basic autonomic homeostatic reflexes, such as baroreceptor reflexes, do display a reciprocal pattern of control over the autonomic branches. However, the ANS is far from being a simple homoeostatic mechanism controlled by reflex systems of the lower brainstem. Indeed, it is now apparent that autonomic outflow is regulated by neurobehavioral systems at the highest levels of the neuraxis, including the cerebral cortex. Rostral brain systems not only modulate lower reflex mechanisms, but issue descending projections that terminate directly on autonomic source nuclei in the brainstem and spinal cord. In accord with the general increase in flexibility and integrative capacity of higher neural systems, rostral neurobehavioral mechanisms appear to exert more variable and flexible control over autonomic outflows. Consequently, in addition to the reciprocal mode of control often seen in reflex regulation, a wider range of control modes can be observed in behavioral contexts, including independent changes in the autonomic branches as well as coactivation or coinhibition of both divisions.

Autonomic Afferents and Ascending Central Pathways

An important aspect of the autonomic nervous system is its sensory function (Dworkin, 2000). In fact, over 75% of the fibers in the largest autonomic nerve, the vagus, are afferents. Visceral afferents carry a range of information concerning the internal state of the body, from baroreceptors, chemoreceptors, and other interoceptors. Some visceral afferents enter the spinal cord via the dorsal root (along with somatic afferents) and terminate in the dorsal horn, where second- and higher-order neurons may participate in local autonomic reflexes, or relay visceral information to higher central structures. One such structure is the nucleus of the tractus solitarius (NTS), a major visceral relay station in the brainstem (Figure 1). Additional visceral afferents, such as those carried by the vagus and other cranial nerves, terminate directly in the NTS. The NTS is a key structure in brainstem autonomic reflexes and serves as an important relay in

ascending pathways to higher levels of the neuraxis where they can modulate the processing of rostral neural systems. Although the functional contributions of this ascending visceral information have not been fully elucidated, it has been shown, for example, that baroreceptor activation can reduce cortical arousal, suppress spinal reflexes, and attenuate pain transmission (Dworkin, 2000). The impact of this ascending information on rostral neurobehavioral mechanisms, and its role in cognitive and behavioral processes, has become an active area of research.

FUNCTIONS OF THE ANS: HOMEOSTASIS AND ALLOSTASIS

A historically recognized role of the ANS is in the maintenance of internal homeostasis. Central and ganglionic autonomic reflex circuits react to perturbations in internal states and generate responses that compensate for these perturbations and restore internal conditions. On standing up from a sitting position, for example, gravitational forces result in a pooling of blood in the legs, which could lead to a dangerous drop in blood pressure and circulatory compromise. This orthostatic challenge becomes a serious problem in autonomic failure and other conditions of impaired autonomic function, where it may lead to syncope (fainting). In healthy individuals, however, this postural maneuver results in the unloading of the carotid baroreceptors and an associated decrease in baroreceptor afferent activity. Baroreceptor reflexes then trigger a compensatory increase in sympathetic outflow and a reciprocal decrease in parasympathetic activity. The resulting increase in heart rate and cardiac output, together with sympathetically mediated vasoconstriction, serves to restore normal blood pressure. This illustrates a classical feedback-regulated (servocontrolled) homeostatic reflex. However, the homeostatic model of autonomic function is overly restrictive, and does not adequately reflect the adaptability and flexibility of autonomic regulation.Feedback-regulated systems, for example, can respond only after a disturbance has taken place, and hence do not effectively prevent these disturbances. Fortunately, central autonomic control systems provide for a broader range of regulatory adjustments, including anticipatory responses.

The early work of Pavlov, demonstrating the conditioning of autonomic responses, foreshadowed current models of autonomic regulation that extend well beyond simple feedback-regulated control mechanisms. Through central associative processes, both exteroceptive and interoceptive stimuli can come to control ANS activity in an anticipatory fashion, and can effectively prevent or minimize perturbations prior to their occurrence (Dworkin, 2000). Conditioned anticipatory autonomic responses, for example, contribute to cardiovascular adjustments to orthostatic stress prior to severe blood pressure perturbations, and trigger anticipatory insulin release prior to the onset of a meal. Specific associations between an innocuous stimulus (conditioned stimulus, CS) that is paired with an aversive event (unconditioned stimulus, US) can result in conditioned fear reactions to the CS in which preparatory somatic and autonomic responses can be emitted in anticipation of the aversive stimulus or event (Ohman *et al.*, 2000). Similarly, the exaggerated affective and autonomic reactions to a wider (and sometimes poorly defined) range of stimuli and contexts in anxiety states probably reflects a pattern of hyperattentional processing of threat-related stimuli based on a broader and less specific associative structure (Berntson *et al.*, 1998).

Moreover, although central autonomic regulatory systems do contribute to homeostasis, steady state conditions are not always optimal, and central regulatory systems may also promote explicit deviations from homeostasis. Psychological stress, for example, can lead to an inhibition of the baroreceptor heart rate reflex, allowing an increase in both heart rate and blood pressure that might contribute to an adaptive response (see Berntson and Cacioppo, 2000). This is indicative of what has been termed 'allostasis' (McEwen, 1999), rather than homeostasis, and represents a class of autonomic control by which central mechanisms can actively adjust internal states in accord with adaptive demands (see McEwen, 1999; Berntson and Cacioppo, 2000). This is important, as it is not always optimal to maintain homeostatic, steady state conditions. During physical exercise or in the face of a survival threat, for example, there would be considerable adaptive advantage to increasing cardiac output and blood pressure, to enhance blood perfusion of muscles. In such instances, integrative neural systems shift the regulatory set point to a new level (allostasis), rather than maintaining a single homeostatic state. In fact, regulation of autonomic activity is often more dynamic than static, especially in behavioral contexts, so central neurobehavioral systems could more appropriately be considered to exert a pattern of allodynamic regulation over autonomic states (Berntson and Cacioppo, 2000). The construct of allodynamic regulation more appropriately encompasses the

complexity and flexibility of central autonomic control and its integration with neurobehavioral function.

EMOTION, ANXIETY, AND ANS ACTIVITY

Autonomic reactions have long been recognized as integral aspects of emotional states such as fear and anxiety, and it is now apparent that there is no clear distinction between higher neural systems underlying behavioral, autonomic, and neuroendocrine regulation. Increasingly, a close integration of these diverse response domains is apparent, so that behavioral responses, for example, are accompanied by the requisite autonomic, and neuroendocrine support. An illustration of this comes from the broad actions of corticotrophin releasing hormone systems at many levels of the neuraxis, which appear to coordinate and integrate the affective, behavioral, autonomic, and neuroendocrine states underlying stress reactions.

In addition, cerebral cortical areas, including the medial prefrontal cortex, have been implicated in the cognitive aspects of affective states, and have been shown to have relatively direct projections to lower autonomic mechanisms (Berntson *et al.*, 1998). These descending pathways (Figure 2) represent important routes by which rostral neural systems can modulate and regulate ANS activity. Indeed, autonomic adjustments are ubiquitous in cognitive and behavioral contexts, and these adjustments frequently do not adhere to expectations based on simple metabolic demands. Rather, these autonomic adjustments are likely to reflect an integrated pattern of central control over cognitive, behavioral, and autonomic functions. Consequently, it is not surprising that psychological dysfunctions are often associated with altered autonomic control.

The fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) published by the American Psychiatric Association recognizes altered autonomic activity as a frequent feature of anxiety disorders. However, there is no single autonomic feature that uniformly characterizes anxiety states; indeed, there appear to be considerable differences between the anxiety disorders, and even between individuals within a given diagnostic category (Berntson *et al.*, 1998). This variability attests to the flexibility of autonomic control by higher neural substrates. It remains the case that altered autonomic function is a common accompaniment of anxiety disorders, and may reflect both cause and consequence of these conditions.

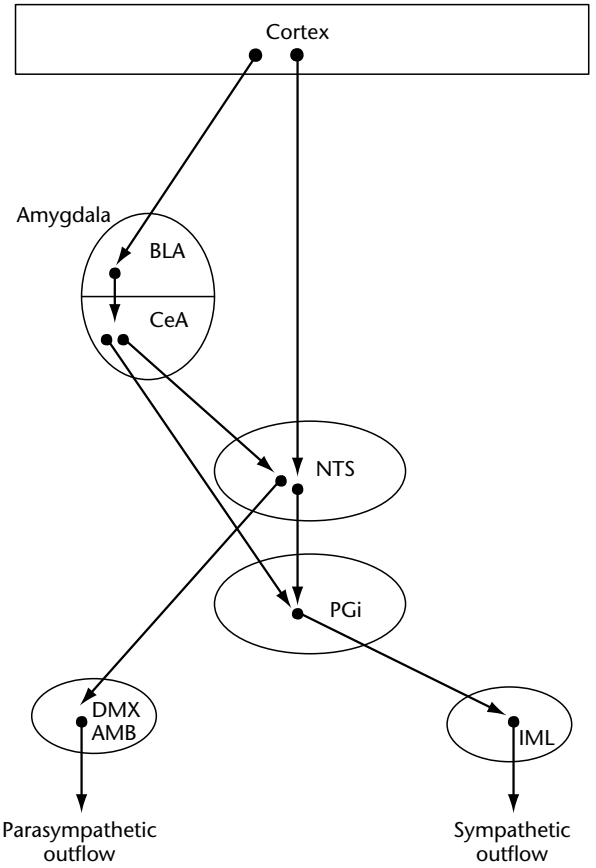


Figure 2. Descending neural systems that affect fear, anxiety, and autonomic control. AMB, nucleus ambiguus; BLA, basolateral amygdala; CeA, central nucleus of the amygdala; DMX, dorsal motor nucleus of the vagus; IML, intermediolateral cell column of the cord; NTS, nucleus of the tractus solitarius; PGi, nucleus paragigantocellularis.

The pathways considered above (Figure 2) provide ample routes by which cognitive and behavioral processes can influence autonomic states. Autonomic and neuroendocrine activation in response to stressors serves to mobilize visceral resources in support of the requirements of 'fight or flight', and of the attentional and cognitive demands of adaptive challenges. The stressors of contemporary society, however, do not require – or even allow – a behavioral response of fight or flight, and hence this visceral mobilization may not be readily resolved. Indeed, the metabolic demand for somatic activation imposed by chronic stressors has diminished at the very time that the requirement for visceral support of attentional, cognitive, and behavioral adaptation strategies has increased. Thus, the physiological response to stress that worked well in human evolution may have

maladaptive aspects, which become more evident as urban societies develop and life expectancy increases well beyond the reproductive years. Because the somatovisceral mobilizing functions of the autonomic nervous system appear to have evolved to deal with transient challenges (Dworkin, 2000), the prolonged somatovisceral activation associated with chronic psychological stressors in today's society may have long-term health costs (McEwen, 1999). Consequently, the cortical and cognitive processes that underlie this chronic activation are especially important to understand.

COGNITIVE INTERACTIONS WITH ANS ACTIVITY

The descending pathways of Figure 2 represent important substrates for 'top down' regulation of the ANS, whereby cognitive and behavioral states can affect autonomic function. Indeed, cognitive imagery alone is sufficient to trigger autonomic responses. Because cognitive and attentional biases lie at the core of anxiety disorders such as posttraumatic stress disorder and generalized anxiety disorders, these descending pathways probably contribute to the altered autonomic function in anxiety states. The interactions between cognitive and autonomic functions are not one-way, however, and the sensory functions of the ANS may allow autonomic states to bias cognitive processing.

The nineteenth-century psychologist William James proposed that sensory feedback from physiological responses might constitute emotional experience (see Cacioppo *et al.*, 1992). Although it now appears that somatovisceral feedback is not essential for emotional reactivity, this feedback may importantly bias cognitive and emotional processing (see Berntson *et al.*, 1998). Thus, visceral afference, carried by vagal afferents with a relay in the NTS, has been shown to enhance emotional memory in animals (McGaugh *et al.*, 2000), and feedback from autonomic responses may be sufficient to trigger panic attacks in people with panic disorder. More generally, somatovisceral and environmental feedback may have a crucial role in organizing and regulating broad aspects of behavior, affect, and cognition (see, for example, Bechara *et al.*, 2000).

One anatomical route by which this type of 'bottom up' priming may be effected is suggested by work on the nucleus paragigantocellularis (PGi) and the locus ceruleus (LC) in the brainstem (Aston-Jones *et al.*, 1996). These structures could be considered as parts of an ascending visceral afferent system. The PGi receives direct input from the NTS, and modulates sympathetic outflow via descending projections and LC activity via ascending projections (Figure 3). Thus, activity of LC neurons would be expected to be highly responsive to the state of activity of the sympathetic branch. Noradrenergic neurons of the LC in turn issue excitatory projections to the basal forebrain cortical cholinergic system, as well as to the cortex directly.

The ascending pathways of Figure 3 represent a potentially important route by which autonomic activity and associated visceral afference might modulate cognitive and emotional processing. In addition to the known descending projections from rostral neurobehavioral systems, this ascending pathway suggests an additional class of interactions between autonomic activity and cortical/cognitive processes. Moreover, the potential for both 'top down' (Figure 2) and 'bottom up' (Figure 3) influences may represent the substrate for a vicious circle of reciprocal

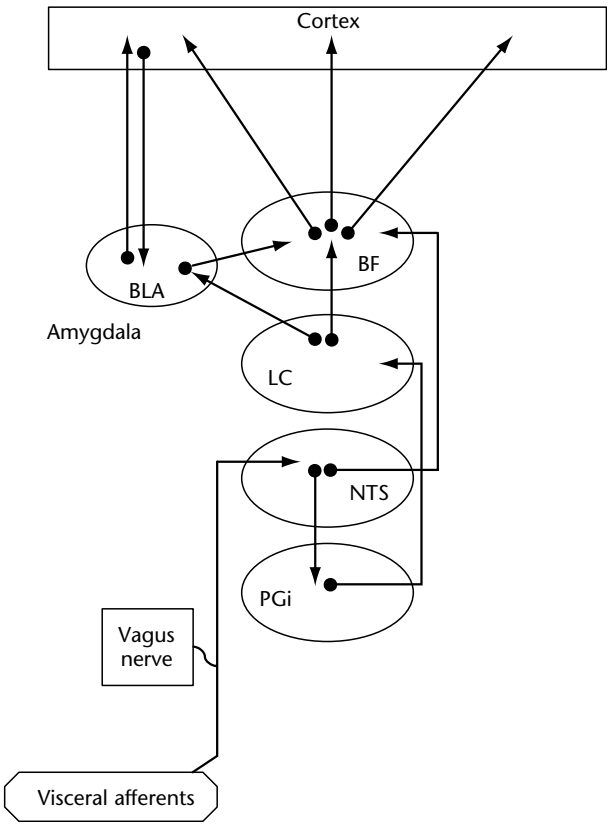


Figure 3. Ascending neural systems that may affect cognitive and emotional processing. BF, basal forebrain; BLA, basolateral amygdala; LC, locus ceruleus; NTS, nucleus of the tractus solitarius; PGi, nucleus paragigantocellularis.

cognitive/autonomic priming that might account for the apparent irrationality and exaggerated hyperattentional processing in anxiety states.

Anxiety, Anxiolytics, and the Basal Forebrain Cortical Cholinergic System

Because cognitive processes appear to contribute substantially to the clinical features of anxiety, including the hyperattentional processing of threat-related stimuli, it is not surprising that cortical systems have been implicated in anxiety disorders. An illustration comes from studies on anxiety and the basal forebrain cholinergic system (Berntson *et al.*, 1998). The widespread corticopetal projections of the basal forebrain provide the primary source of cortical ACh, and ACh appears to enhance cortical processing generally (Sarter and Bruno, 2000). This is illustrated by the global dementia of Alzheimer disease, which is associated with degeneration of the basal forebrain cholinergic system. In view of these considerations, cortical cholinergic activity would be expected to exaggerate the cognitive processing underlying anxiety. This is consistent with the finding that selective immunotoxic lesions of the cholinergic neurons of the basal forebrain in the rat attenuate the exaggerated behavioral and autonomic reactions in anxiogenic contexts (Berntson *et al.*, 1998).

Additional evidence supports this view and explains the antianxiety actions of benzodiazepine receptor agonists such as chlordiazepoxide and diazepam (for a review, see Berntson *et al.*, 1998). The activity of basal forebrain cholinergic neurons is regulated in part by an inhibitory input mediated by γ-aminobutyric acid (GABA), which in turn is bidirectionally modulated through a benzodiazepine binding site on the GABA receptor complex. The anxiolytic benzodiazepine receptor agonists enhance GABA-mediated inhibition and reduce basal forebrain cholinergic activity. Conversely, benzodiazepine receptor inverse or partial inverse agonists, which decrease GABA-mediated inhibition and enhance basal forebrain cholinergic activity, have notable anxiogenic properties (see Berntson *et al.*, 1998). Especially relevant in the link between anxiety and autonomic function may be the medial prefrontal cortex, an area that has been repeatedly implicated in both affect and autonomic control. Selective immunotoxic lesions of the basal forebrain cholinergic projections to the medial prefrontal cortex, like lesions of the basal forebrain itself, eliminate the exaggerated autonomic reactions in anxiety-related contexts (see Berntson *et al.*, 1998).

The ascending pathways in Figure 3 are intended to be representative rather than exhaustive, and there are many important questions that remain in this area. It is increasingly apparent, however, that comprehensive models of cognitive processing or autonomic functions will require a deeper understanding of the reciprocal interactions between these functional domains. In contrast to the Jamesian view that somatovisceral feedback constitutes emotion, the ascending pathway of Figure 3 provides a route by which visceral afference may more subtly trigger or modulate cognitive/emotional processing, even in the absence of conscious perception of the visceral stimulus.

CONCLUSION

Classical views of the autonomic nervous system often focused on brainstem reflexes and the autonomic regulation of internal homeostasis. With the elucidation of higher neurobehavioral influences over autonomic outflow, and the important sensory functions of the ANS, these limited perspectives are no longer tenable. Beyond lower sensory and motor neurons, there is no clear distinction between central systems that regulate behavioral, autonomic, neuroendocrine, and immune processes. Moreover, it appears that the higher the level of neural organization, the greater is the integration among these response domains, culminating in the supreme integrative functions of cerebral cortical systems. Consequently, an understanding of cortical/cognitive processes and the systems and mechanisms that regulate this processing is a prerequisite for a comprehensive model of the autonomic nervous system and its central regulation.

References

Aston-Jones G, Rajkowski J, Kubiak P, Valentino RJ and Shipley MT (1996) Role of locus coeruleus in emotional activation. *Progress in Brain Research* **107**: 379–402.

Bechara A, Damasio H and Damasio AR (2000) Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex* **10**: 295–307.

Berntson GG and Cacioppo JT (2000) From homeostasis to allodynamic regulation. In: Cacioppo JT, Tassinary LG and Berntson GG (eds) *Handbook of Psychophysiology*, pp. 459–481. Cambridge, UK: Cambridge University Press.

Berntson GG, Cacioppo JT and Quigley KS (1991) Autonomic determinism: the modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychological Review* **98**: 459–487.

Berntson GG, Sarter M and Cacioppo JT (1998) Anxiety and cardiovascular reactivity: the basal forebrain

cholinergic link. *Behavioural Brain Research* **94**: 225–248.
Cacioppo JT, Berntson GG and Klein DJ (1992) What is an emotion? The role of somatovisceral afference, with special emphasis on somatovisceral 'illusions'. *Review of Personality and Social Psychology* **14**: 63–98.
Cannon WB (1939) *The Wisdom of the Body*. New York, NY: WW Norton.
Dworkin BR (2000) Interoception. In: Cacioppo JT, Tassinary LG and Berntson GG (eds) *Handbook of Psychophysiology*, pp. 482–405. Cambridge, UK: Cambridge University Press.
Langley JN (1921) *The Autonomic Nervous System*. Cambridge, UK: Heffer.
McEwen BS (1999) Protective and damaging effects of mediators of stress: elaborating and testing the concepts of allostasis and allostatic load. *Annals of the New York Academy of Sciences* **896**: 30–47.
McGaugh JL, Roozendall B and Cahill L (2000) Modulation of memory storage by stress hormones and the amygdala complex. In: Gazanniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 1081–1098. Cambridge, MA: MIT Press.
Ohman A, Hamm A and Hugdahl K (2000) Cognition and the autonomic nervous system: orienting, anticipation, and conditioning. In: Cacioppo JT, Tassinary LG and Berntson GG (eds) *Handbook of Psychophysiology*, pp. 533–579. Cambridge, UK: Cambridge University Press.
Sarter M and Bruno JP (2000) Cortical cholinergic input mediating arousal, attentional processing and dreaming: differential afferent regulation of the basal forebrain and brainstem afferents. *Neuroscience* **95**: 933–952.

Further Reading

Appenzeller O (1999) *The Autonomic Nervous System. Part I, Normal Functions*. New York, NY: Elsevier.
Appenzeller O (2000) *The Autonomic Nervous System. Part II, Dysfunctions*. New York, NY: Elsevier.
Cacioppo JT, Tassinary LG and Berntson GG (2000) *Handbook of Psychophysiology*. Cambridge, UK: Cambridge University Press.
Goehler LE, Gaykema RP, Hansen MK *et al.* (2000) Vagal immune-to-brain communication: a visceral chemosensory pathway. *Autonomic Neuroscience* **85**: 49–59.
Schulkin J, Gold PW and McEwen BS (1998) Induction of corticotropin-releasing hormone gene expression by glucocorticoids: implication for understanding the states of fear and anxiety and allostatic load. *Psychoneuroendocrinology* **23**: 219–243.
Williams CL and Clayton EC (2001) Contribution of brainstem structures in modulating memory storage processes. In: Gold PE and Greenough WT (eds) *Memory Consolidation: Essays in Honor of James L. McGaugh*, pp. 141–162. Washington, DC: American Psychological Association.

Basal Forebrain

Intermediate article

Eve De Rosa, Stanford University School of Medicine, Stanford, California, USA
Mark G Baxter, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

The basal forebrain is a complex of subcortical nuclei that project widely to cortical and limbic areas involved in cognitive function. Damage to the basal forebrain is associated with cognitive deficits. The contributions of particular neuroanatomical and neurochemical components of the basal forebrain to different aspects of cognitive function can be dissociated to some extent.

INTRODUCTION

The term 'basal forebrain' commonly refers to an extended continuum of subcortical neurons that projects to diverse limbic and neocortical areas implicated in various aspects of cognitive function. Damage to the basal forebrain region can result in global cognitive impairments. For instance, aneurysms of the anterior communicating artery that injure the basal forebrain are associated with amnesia and impairments in executive function. Cognitive deficits in both normal aging and age-related pathological conditions have also been associated with the basal forebrain. The severity of cognitive impairment observed in Alzheimer disease is correlated with the extent of deterioration of cholinergic neurons in the basal forebrain. A similar relationship between cognitive impairment and alterations in basal forebrain cholinergic neurons is seen in normal aging. For these reasons, cholinergic neurons have been central to most explanations of the cognitive effects of basal forebrain damage. Hypotheses regarding the involvement of the basal forebrain cholinergic system in global aspects of cognitive function have been gradually revised as more and more selective lesion methods have become available for experimental studies of this region (Wenk, 1997).

Attempts to identify the specific cognitive functions of basal forebrain cholinergic neurons have suggested that these neurons do not play a specific part in memory processing *per se*, or a general role in sustaining the function of their cortical targets. Instead, damage limited to basal forebrain cholinergic neurons produces highly restricted impairments in aspects of sensory information processing and attention.

ANATOMY AND CONNECTIONS OF THE BASAL FOREBRAIN

Cholinergic basal forebrain neurons are intermingled with a substantial population of noncholinergic neurons that share similar projection patterns (Gritti *et al.*, 1997). Estimates of the proportion of cortically projecting basal forebrain neurons that are cholinergic vary from study to study but are generally in the range 30–50 percent. Many noncholinergic neurons in the basal forebrain may be local circuit neurons, receiving cortical input and modulating activity of cortically projecting cholinergic and noncholinergic neurons (Zaborszky *et al.*, 1997).

The basal forebrain can be divided into four groups of cells: the medial septum, projecting primarily to the hippocampus; the vertical limb of the diagonal band of Broca, projecting primarily to the hippocampus and cingulate cortex; the horizontal limb of the diagonal band of Broca, projecting primarily to the olfactory bulb, piriform cortex and entorhinal cortex; and the nucleus basalis magnocellularis or substantia innominata, projecting primarily to the neocortex and amygdala (Figure 1). These cholinergic nuclei have also been designated Ch1 to Ch4 (Mesulam *et al.*, 1983); this nomenclature approximately corresponds to the above divisions. The organization of the basal forebrain is similar in the primate and in the rat, although subdivisions of the nucleus basalis can be identified reliably in the primate but not in the rat. Some areas of the basal forebrain are reciprocally connected with

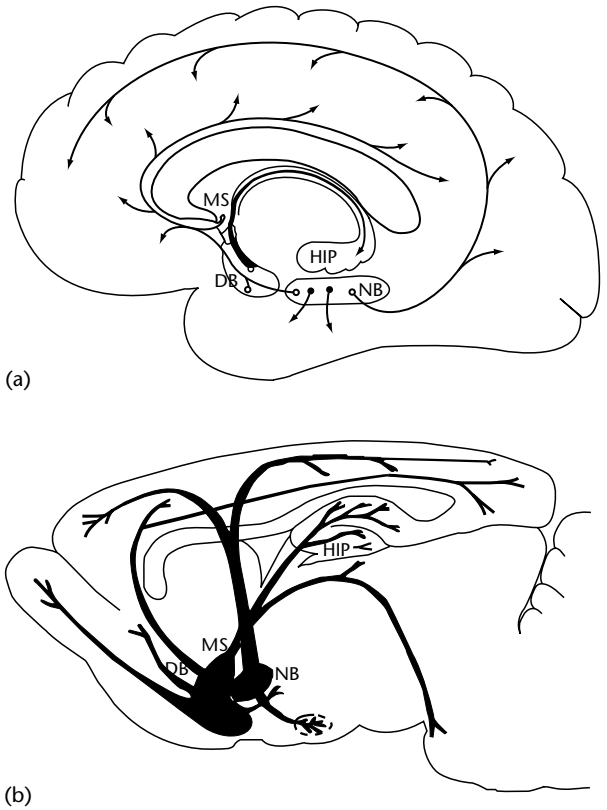


Figure 1. The basal forebrain cholinergic system, schematically represented in sagittal views of (a) human and (b) rat brain. The basal forebrain can be roughly divided into three major divisions (rostral to caudal): the medial septum (MS), projecting primarily to hippocampus (HIP); the diagonal band nuclei (DB), consisting of the vertical limb of the diagonal band of Broca, projecting to the hippocampus and cingulate cortex, and the horizontal limb of the diagonal band of Broca, projecting to the olfactory bulb and entorhinal cortex; and the nucleus basalis (NB), projecting to neocortex and amygdala. These cell groups share similar projection patterns in both species. The substantia innominata, which forms a less discrete nucleus in the rat than in the primate, sends cholinergic projections to the neocortex.

their targets; anatomical experiments in rats have shown that the medial septum and diagonal band project to the hippocampus and medial prefrontal cortex and receive projections back from these structures. Similarly, the nucleus basalis is reciprocally connected with the amygdala. The basal forebrain also receives inputs from the hypothalamus, as well as from the midbrain and upper pons (Wainer and Mesulam, 1990). Inhibitory projections from the nucleus accumbens to the basal forebrain may have a particular role in regulating cortical acetylcholine release (Sarter and Bruno, 2000).

INVOLVEMENT OF THE BASAL FOREBRAIN IN MEMORY

Considerable attention has been devoted to the magnocellular cholinergic neurons of the basal forebrain and their function in memory, because of the hypothesis that cholinergic dysfunction specifically is partially or wholly responsible for the memory deficits seen after damage to the basal forebrain (the 'cholinergic hypothesis'). Alzheimer disease involves loss of these neurons. Postmortem examinations of brains from people with Alzheimer disease have found a marked cell loss in the basal forebrain relative to the brains of healthy people. In Alzheimer disease there is a loss of choline acetyltransferase and acetylcholinesterase in the cortical targets of the basal forebrain. The enzyme choline acetyltransferase synthesizes acetylcholine; once released from the presynaptic terminal, acetylcholine is degraded by the enzyme acetylcholinesterase. Cognitive impairment in people with Alzheimer disease is correlated with the extent of cholinergic loss. These findings suggest that the loss of cholinergic transmission in the basal forebrain contributes to the mnemonic dysfunction observed in this disease (Collerton, 1986).

Additional support for the presumed role of the basal forebrain in memory dysfunction is derived from patients with amnesia resulting from an anterior communicating artery aneurysm. Infarcts associated with these aneurysms do not typically damage the traditional cerebral areas implicated in amnesia (i.e. medial temporal and diencephalic structures). Amnesic patients who sustained infarcts resulting in lesions confined to the basal forebrain with no evidence of diencephalic or medial temporal involvement have relatively intact immediate recall and intact implicit memory, but impaired delayed recall and an increased susceptibility to proactive interference. The increased susceptibility to proactive interference may be due to deficient attentional inhibitory mechanisms. These patients have also shown a reduced information processing speed, poor divided attention ability and increased distractibility (DeLuca and Diamond, 1995).

The findings in the above disorders suggest that the basal forebrain contributes to mnemonic dysfunction, motivating the development of animal models of basal forebrain damage to elucidate the exact role of the cholinergic basal forebrain system in mediating such dysfunction. In these models, the basal forebrain cholinergic system or its cholinergic targets are damaged and behavioral tests are performed to determine the consequent pattern of

mnemonic function. The validity of these models depends on the similarity of the cognitive processes being tested in animals and in humans, as well as on the anatomical homology of the basal forebrain in animals and humans.

To confirm the cholinergic hypothesis in an animal model, two challenges must be met: any memory deficits observed must be due to a specific impairment in memory processes, as opposed to disruption of other cognitive or noncognitive processes; and the deficits must not be attributable to destruction of other populations of noncholinergic neurons in the basal forebrain (Sarter and Bruno, 1997).

Most initial studies used stereotaxic placement of electrolytic or excitotoxic lesions in different regions of the basal forebrain in rats. Interpretation of these studies has been made difficult by the fact that electrolytic lesions damage mixed populations of neurons as well as passing fibers, and that excitotoxic lesions – produced, for example, by ibotenic acid, kainic acid, quisqualic acid or α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) – preserve the passing fibers but still damage mixed populations of neurons. Ibotenate and other excitotoxin-induced lesions of the nucleus basalis magnocellularis disrupt learning in working memory tasks (such as delayed alternation) and reference memory tasks (measuring spatial or conditional visual discrimination learning).

The cholinergic hypothesis was called into question, however, because lesions of the basal forebrain made with other excitotoxins (quisqualic acid and AMPA) resulted in greater destruction of cortically projecting cholinergic neurons of the nucleus basalis magnocellularis than that seen following ibotenate lesions, but induced only a few of the learning and memory deficits caused by ibotenate. Consequently, many of the deficits induced by ibotenate lesions of the basal forebrain cannot be attributed to a primary destruction of cholinergic neurons in the nucleus basalis magno cellularis (Everitt and Robbins, 1997).

Because these lesion techniques are not selective for cholinergic neurons and therefore the resulting behavioral deficits may be due in part to damage in some neighboring neuronal system, explicit tests of the cholinergic hypothesis require a neurotoxin specific for basal forebrain cholinergic neurons. The immunotoxin, 192 IgG-saporin, provides a route for directly targeting the toxin saporin to cholinergic neurons *in vivo* in the rat without damaging noncholinergic neurons at the lesion site. Immunotoxic lesions restricted to specific cholinergic nuclei of the basal forebrain have generally failed to produce mnemonic deficits. It is noteworthy that the more specific the cholinergic lesion, the less dramatic the effects on learning and memory (Everitt and Robbins, 1997; Baxter and Chiba, 1999). The studies with less specific excitotoxic lesions underscore the importance of careful validation that basal forebrain lesions intended to be specific to cholinergic neurons do actually spare noncholinergic neurons at the lesion site (for related discussion see Chappell *et al.*, 1998).

The foregoing experiments in rats suggesting a role for the basal forebrain in memory function but no specific role for cholinergic basal forebrain neurons are supported by experiments in nonhuman primates. These animals have a more spatially distinct nucleus basalis of Meynert, and comparisons of cognitive deficits between humans and monkeys are more plausible than between humans and rats. Voytko (1996) and her colleagues extensively and systematically investigated the effect of administration of ibotenic acid lesions placed stereotaxically into the basal forebrain cholinergic nuclei of Old World (macaque) monkeys. Their battery of tests (delayed nonmatching to sample, delayed response, simultaneous discrimination, and a visuospatial attention task) revealed no mnemonic deficits. There is some contradiction between the findings of ibotenic lesions of the basal forebrain nuclei in Old and New World monkeys. Ibotenate lesions to the nucleus basalis of Meynert led to impairments in retention of preoperatively learned visual discriminations and reversal learning of visual discriminations postoperatively in New World monkeys. This was accounted for by possible nonspecific damage in the lesions made in New World monkeys. Administration of the putative cholinergic toxin ME20.4 IgG-saporin to the basal forebrain nuclei of New World monkeys led to perceptual impairments on retention and acquisition of visual discriminations (Fine *et al.*, 1997), although sparing of noncholinergic neurons at the lesion site in these monkeys was not verified.

Aims to increase acetylcholine levels in the clefts of the cholinergic synapse with acetylcholinesterase inhibitors, or to directly replace acetylcholine with cholinomimetic drugs, have largely been unsuccessful in reversing cognitive deficits in people with Alzheimer disease. This may be because direct enhancement of acetylcholine levels may disrupt any remaining intact cholinergic transmission. An alternative approach, trans-synaptic modulation, describes the effects of changes in the activity of an afferent neuronal system on the excitability of its target neurons, e.g. the effects of afferents mediated by γ-aminobutyric acid (GABA), originating locally

or distally, on the cholinergic neurons of the basal forebrain. Another complementary, pharmacological strategy to treatment of cognitive deficits consequent to cholinergic loss would be to modulate the cholinergic system by trans-synaptic mechanisms, for example by affecting cholinergic transmission with GABA. This approach may be more effective in alleviating the cognitive deficits since this acts to amplify the endogenous physiological cholinergic signal (Sarter and Bruno, 1997). However, given that loss of cholinergic neurons does not seem to be a primary cause of memory impairment in these conditions, even pharmacological treatments aimed at enhancing function of remaining cholinergic neurons by trans-synaptic modulation might be expected to have little beneficial effect on memory.

In summary, the most recent and stringent tests of the cholinergic hypothesis in animal models have not supported an essential role for cholinergic involvement in learning and memory functions *per se*. In contrast, basal forebrain cholinergic neurons do appear to have an essential role in regulating attentional processing capacity.

INVOLVEMENT OF THE BASAL FOREBRAIN IN ATTENTION

Mechanisms of attention are thought to enable the efficient allocation of processing resources in order to respond to important events in the environment. There are different types of attentional processes: selective or focused attention targets attentional resources on a restricted number of sensory stimuli and excludes other sensory stimuli; sustained attention or vigilance reflects a state of readiness to detect and respond to unpredictable or rare events; and divided attention refers to simultaneous monitoring of several different channels of sensory information. Attention filters what sensory information is introduced into memory. The basal forebrain appears to regulate many of these attentional functions; recent studies of lesions of the cholinergic forebrain in rats and monkeys have suggested deficits in several different paradigms for measuring attention, including selective spatial, divided and sustained attention.

The five-choice serial reaction time task, based on a human test of continuous performance, measures sustained spatial attention in rats. In this task, rats are required to monitor the occurrence of a light stimulus in one of five locations. With this paradigm, excitotoxic lesions (e.g. ibotenate acid, quisqualic acid and AMPA) of the nucleus basalis magnocellularis (NBM) produce impairments in performance, which are ameliorated by increasing the duration of the stimuli. In support of the involvement of the cholinergic basal forebrain, this deficit is alleviated by an acetylcholinesterase inhibitor, physostigmine (Everitt and Robbins, 1997). A role for the NBM in attention has also been demonstrated in a temporal divided attention task, where at times rats are required to divide their attention between two simultaneous stimuli: rats with excitotoxic lesions of the NBM are specifically impaired in their ability to time two stimuli simultaneously (Olton *et al.*, 1988). In a cross-modal divided attention task, infusion of 192 IgG-saporin into the basal forebrain selectively increases the response latencies in a bimodal condition, without affecting the accuracy or the retrieval and execution of the stimulus–response action in either the visual or auditory modality (Turchi and Sarter, 1997).

Basal forebrain neurons, including the cholinergic neurons, receive direct contacts from GABA-mediated afferents and are inhibited by GABA. Like cholinergic lesions, infusions of a benzodiazepine receptor agonist into the basal forebrain impair the performance of rats tested in a sustained attention without altering the efficacy of perceptual processes. Infusion of the $GABA_A$ receptor agonist, muscimol, directly into the basal forebrain both decreases the activity of the cortically projecting NBM neurons and impairs accuracy on the five-choice task. Two-lever vigilance tasks present the rat with successive trials of changes in the intensity of continuously delivered auditory stimuli (e.g. white noise) or visual stimuli (e.g. illumination). One lever measures the 'hits' and the other lever measures the correct rejections. In two-lever vigilance tasks for rats, both 192 IgG-saporin lesions of the NBM and intra-NBM administration of $GABA_A$ benzodiazepine receptor agonists have impaired vigilance performance (Sarter and Bruno, 1997).

In a visuospatial attention task, Old World monkeys with basal forebrain lesions showed impairments in shifting attention (Voytko, 1996). A brief cue indicates the expected location of a target stimulus. The difference in response time to a target that appears at expected and unexpected locations is used as the measure of shifting of attention. A slower response time on invalid trials, in which the target appears at an unexpected location, compared with valid trials in which the target appears at the expected location, indicates that the subject has paid attention to the cue that is signaling the likely location of a target. Monkeys with basal forebrain lesions showed disproportionately longer response times on invalid trials, suggesting an

impairment in disengaging attention from the invalidly cued location. This was not due to a general impairment in motor skills, since the response to the target following the initiation of a response was comparable in monkeys with and without basal forebrain lesions. A similar result has been described in rats with selective lesions of cholinergic neurons in the NBM (Chiba *et al.*, 1999). It is important to note that this attentional impairment was selective; the monkeys with basal forebrain lesions in the study by Voytko were not impaired on tests of visual learning and memory, and rats with selective cholinergic lesions of the NBM were not impaired on a variety of cognitive tasks including tests of spatial learning and memory (reviewed by Baxter and Chiba, 1999).

Experiments that examine the regulation of attentional processing in associative learning paradigms have identified dissociable roles for rostral (septal) and caudal (NBM) regions of basal forebrain cholinergic neurons in these aspects of attention (reviewed by Baxter and Chiba, 1999). Increases in conditioned stimulus processing, brought about by violations of learned contingencies, require the integrity of cholinergic neurons in the NBM, but not cholinergic neurons in the medial septum or vertical limb of the diagonal band of Broca. Decreases in conditioned stimulus processing, which occur when conditioned stimuli are preexposed in the absence of reinforcement, or are consistent predictors of another event, require the integrity of cholinergic neurons in the latter structures, but not in the NBM.

The basal forebrain has reciprocal connections with the central nucleus of the amygdala, which has potential for widespread influences on cortical processing directly through projections to both lower-order and higher-order sensory areas and indirectly through projections to cholinergic neurons in the basal forebrain. Stimulation of the amygdala central nucleus has been found to influence basal forebrain electroencephalographic patterns consistent with acetylcholine release in rats and humans (Kapp *et al.*, 1994). Cortical acetylcholine release stimulated by the amygdala is thought to induce a state of cortical readiness for processing sensory information. This suggests that the basal forebrain cholinergic system might provide an output pathway for the regulation of attention and cortical information processing by the amygdala. Indeed, evidence from crossed-lesion studies suggests that projections from the central nucleus to the NBM are critical for producing increases in conditioned stimulus processing: disconnection of the NBM from the central nucleus of the amygdala by crossed unilateral lesions also disrupts the ability to increase attentional processing in response to a violation of expectancy (Baxter and Chiba, 1999).

Thus far, selective cholinergic lesions in the basal forebrain system appear to have a limited effect on measures of learning and memory, but rather impair attentional processing. It is assumed that attentional processes are involved in the primary stages in information processing, so it is remarkable that no dramatic effects in various learning and memory tasks have been observed following damage to basal forebrain cholinergic neurons. If attentional processes are impaired, then a performance deficit would be expected in most learning and memory tasks, but it appears that this is not the case after selective cholinergic lesions. However, the statement that basal forebrain damage impairs attention is an oversimplification. Instead, the effects of these lesions might be better characterized as producing an impairment in the ability to regulate attentional processing appropriately in response to task or environmental demands (Baxter and Chiba, 1999), rather than a reduction in the ability to attend to stimuli *per se*.

CONCLUSION

The basal forebrain is anatomically situated to regulate information processing in cortical and limbic structures involved in a wide array of cognitive functions. Indeed, extensive damage to the basal forebrain (to both cholinergic and noncholinergic components) results in a broad array of cognitive impairments. In contrast, damage limited to basal forebrain cholinergic neurons produces more restricted impairments in regulation of attentional processing. Hence, basal forebrain cholinergic neurons appear to have a selective role in regulating cortical information processing, rather than a generalized role in supporting the functions of their target areas. This apparently restricted function of basal forebrain cholinergic neurons suggests that a loss of these cells is probably not the core factor in producing cognitive deficits in conditions such as Alzheimer disease, or in the amnesia consequent to aneurysms in the basal forebrain.

The characterization of the function of basal forebrain cholinergic neurons as 'attentional' may fail to capture the array of functions performed by these neurons. These neurons are probably involved in other aspects of sensory processing and representational plasticity in primary sensory cortical areas (Baskerville *et al.*, 1997; Kilgard and Merzenich, 1998). The involvement of basal

forebrain cholinergic projections to other cortical areas in similar types of functions has not yet been investigated. Hence, a general operating principle for the role of these cholinergic neurons in cognition remains elusive, as does a role for the basal forebrain generally in cognitive function. The basal forebrain, particularly the cholinergic component, may be more broadly involved in gating cortical information processing and regulating aspects of conscious experience (Sarter and Bruno, 2000); hence abnormalities in basal forebrain function could contribute to a wide variety of cognitive impairments in both neurodegenerative disease and in psychopathology. These hypotheses represent a useful guiding principle for future studies of basal forebrain function, both in animal models and in human clinical populations.

References

Baskerville KA, Schweitzer JB and Herron P (1997) Effects of cholinergic depletion on experience-dependent plasticity in the cortex of the rat. *Neuroscience* **80**: 1159–1169.

Baxter MG and Chiba AA (1999) Cognitive functions of the basal forebrain. *Current Opinion in Neurobiology* **9**: 178–183.

Chappell J, McMahan R, Chiba A and Gallagher M (1998) A re-examination of the role of basal forebrain cholinergic neurons in spatial working memory. *Neuropharmacology* **37**: 481–487.

Chiba AA, Bushnell PJ, Oshiro WM and Gallagher M (1999) Selective removal of cholinergic neurons in the basal forebrain alters cued target detection in rats. *Neuroreport* **10**: 3119–3123.

Collerton D (1986) Cholinergic function and intellectual decline in Alzheimer's disease. *Neuroscience* **19**: 1–28.

DeLuca J and Diamond BJ (1995) Aneurysm of the anterior communicating artery: a review of neuroanatomical and neuropsychological sequelae. *Journal of Clinical and Experimental Neuropsychology* **17**: 100–121.

Everitt BJ and Robbins TW (1997) Central cholinergic systems and cognition. *Annual Review of Psychology* **48**: 649–684.

Fine A, Hoyle C, Maclean CJ *et al*. (1997) Learning impairments following injection of a selective cholinergic immunotoxin, ME20.4 IgG-saporin, into the basal nucleus of Meynert in monkeys. *Neuroscience* **81**: 331–343.

Gritti I, Mainville L, Mancia M and Jones BE (1997) GABAergic and other noncholinergic basal forebrain neurons, together with cholinergic neurons, project to the mesocortex and isocortex in the rat. *Journal of Comparative Neurology* **383**: 163–177.

Kapp BS, Supple WF and Whalen PJ (1994) Effects of electrical stimulation of the amygdaloid central nucleus on neocortical arousal in the rabbit. *Behavioral Neuroscience* **108**: 81–93.

Kilgard MP and Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* **279**: 1714–1718.

Mesulam MM, Mufson EJ, Wainer BH and Levey AI (1983) Central cholinergic pathways in the rat: an overview based on an alternative nomenclature (Ch1–Ch6). *Neuroscience* **10**: 1185–1201.

Olton DS, Wenk GL, Church RM and Meck WH (1988) Attention and the frontal cortex as examined by simultaneous temporal processing. *Neuropsychologia* **26**: 307–318.

Sarter M and Bruno JP (1997) Trans-synaptic stimulation of cortical acetylcholine and enhancement of attentional functions: a rational approach for the development of cognition enhancers. *Behavioural Brain Research* **83**: 7–14.

Sarter M and Bruno JP (2000) Cortical cholinergic inputs mediating arousal, attentional processing and dreaming: differential afferent regulation of the basal forebrain by telencephalic and brainstem afferents. *Neuroscience* **95**: 933–952.

Turchi J and Sarter M (1997) Cortical acetylcholine and processing capacity: effects of cortical cholinergic deafferentation on crossmodal divided attention in rats. *Cognitive Brain Research* **6**: 147–158.

Voytko ML (1996) Cognitive functions of the basal forebrain cholinergic system in monkeys: memory or attention? *Behavioural Brain Research* **75**: 13–25.

Wainer B and Mesulam MM (1990) Ascending cholinergic pathways in the rat brain. In: Steriade M and Biesold D (eds) *Brain Cholinergic Systems*, pp. 65–119. Oxford, UK: Oxford University Press.

Wenk GL (1997) The nucleus basalis magnocellularis cholinergic system: one hundred years of progress. *Neurobiology of Learning and Memory* **67**: 85–95.

Zaborszky L, Gaykema RP, Swanson DJ and Cullinan WE (1997) Cortical input to the basal forebrain. *Neuroscience* **79**: 1051–1078.

Further Reading

Baxter MG and Murg SL (2001) The basal forebrain cholinergic system and memory: beware of dogma. In: Squire LR and Schacter DL (eds) *Neuropsychology of Memory*, 3rd edn, pp. 425–436 (2000). New York, NY: Guilford Press.

Berger-Sweeney J, Stearns NA, Murg SL *et al*. (2001) Selective immunolesions of cholinergic neurons in mice: effects on neuroanatomy, neurochemistry, and behavior. *Journal of Neuroscience* **21**: 8164–8173.

Burk JA and Sarter M (2001) Dissociation between the attentional functions mediated via basal forebrain cholinergic and GABAergic neurons. *Neuroscience* **105**: 899–909.

De Rosa E, Hasselmo ME and Baxter MG (2001) Contribution of the cholinergic basal forebrain to proactive interference from stored odor memories

during associative learning in rats. *Behavioral Neuroscience* **115**: 314–327.

Himmelheber AM, Sarter M and Bruno JP (2001) The effects of manipulations of attentional demand on cortical acetylcholine release. *Cognitive Brain Research* **12**: 353–370.

McGaughy J, Everitt BJ, Robbins TW and Sarter M (2000) The role of cortical cholinergic afferent projections in cognition: impact of new selective immunotoxins. *Behavioural Brain Research* **115**: 251–263.

Sarter M and Bruno JP (1999) Abnormal regulation of corticopetal cholinergic neurons and impaired information processing in neuropsychiatric disorders. *Trends in Neurosciences* **22**: 67–74.

Turchi J and Sarter M (2000) Cortical cholinergic inputs mediate processing capacity: effects of 192 IgG-saporin-induced lesions on olfactory span performance. *European Journal of Neuroscience* **12**: 4505–4514.

Waite JJ, Wardlow ML and Power AE (1999) Deficit in selective and divided attention associated with cholinergic basal forebrain immunotoxic lesion produced by 192-saporin; motoric/sensory deficit associated with Purkinje cell immunotoxic lesion produced by OX7-saporin. *Neurobiology of Learning and Memory* **71**: 325–352.

Basal Ganglia

Introductory article

Lucy L Brown, Albert Einstein College of Medicine, Bronx, New York, USA
Samuel M Feldman, New York University, New York, USA

CONTENTS

Nuclei near the base of the brain integrate information from the entire cerebral cortex and serve a regulatory function for movement and cognition.

INTRODUCTION

The basal ganglia are a group of brain structures that regulate movement in mammals. Damage to these subcortical structures is the primary cause of involuntary, abnormally sequenced movements such as tremor, rigidity, athetosis, tics, chorea and ballism. However, the role of the basal ganglia goes well beyond the regulation of the sequencing of movements. They also have a major role in the learning of motor and cognitive skills. They control the expression of voluntary behaviors that are essential for normal mammalian interaction with an ever-changing and challenging environment. Unlike the spinal cord, which is necessary for withdrawal reflexes and the final stages of the movements we execute, the basal ganglia influence the selection and sequence of the final movement, and participate in adaptive motor control.

A neuroscientist's gag is: 'Why are the basal ganglia like the Dean's office?' Answer: 'Because they take up a lot of space and nobody knows what they do.' This old joke turns out to be surprisingly on target. A serious answer might be that they facilitate the global operations of a large 'institution', planning and making decisions based on many varied inputs. Such an executive role in brain function is difficult to analyze, and our knowledge about the function of these nuclei is still evolving. Accordingly, the descriptions of sensorimotor and cognitive functions of the basal ganglia that follow must be seen as descriptions of work in progress.

Anatomically, the basal ganglia are a collection of nuclei in the mammalian forebrain and midbrain. The forebrain nuclei include the caudate nucleus and putamen, the nucleus accumbens, the globus pallidus (internal and external segments) and the subthalamic nucleus. Two subunits of the substantia nigra, the pars reticulata and pars compacta, make up the midbrain component. Similar structures exist in other vertebrates, including reptiles, amphibians, birds and fish. Although the amygdala and other basal forebrain nuclei were once included in discussions of the basal ganglia, current concepts of anatomy and function restrict the definition now used.

Alternative designations are used for some of the basal ganglia nuclei. For example, in primates the caudate and putamen are anatomically distinct, but lower mammals such as the rat lack this dichotomy and have a single undifferentiated nucleus called the striatum. In the primate brain the putamen and globus pallidus are physically close and often hard to distinguish, and are referred to collectively as the lenticular nucleus.

The striatum receives its major synaptic input from the entire extent of the cerebral cortex. This diverse cortical input reflects the many functions in which the basal ganglia participate, ranging from sensorimotor feedback and motor control to spatial working memory. Both the anatomy and physiology suggest that the basal ganglia act collectively as a 'suppress and release' mechanism for a wide range of behaviors. Dysfunction of these mechanisms is seen in basal ganglia disorders; for example, people with Parkinson disease have difficulty initiating movements, while people with Huntington chorea cannot stop. These disorders are involuntary: people with Huntington chorea have little or no control over their flailing movements, nor can people with Parkinson disease voluntarily overcome their 'frozen' state. Thus, the suppress and release mechanism operates at the unconscious level.

ANATOMY AND CHEMOARCHITECTURE

Location and Gross Structure

The caudate–putamen and globus pallidus are situated on either side of the thalamus, beneath the mantle of the cerebral cortex. The corpus callosum and external capsule form a border around the caudate–putamen. In primates, the caudate and putamen are separated by the internal capsule, which contains axon projections between the cortex, brainstem and spinal cord. In the primate, the caudate has a head, body and tail. The head borders the anterior lateral ventricle while the body and tail follow the contour of the lateral ventricle into the temporal lobe (Figure 1). The nucleus accumbens is continuous with the caudate and putamen, extending anteriorly and downward to the inferior surface of the brain. The subthalamic nucleus, at the junction of the midbrain and forebrain, and substantia nigra, in the midbrain, are smaller than the caudate–putamen. The right basal ganglia control the left side of the body and vice versa. Accordingly, parkinsonian tremor of the left hand is evidence of pathological changes in the right basal ganglia.

Neuroanatomists have described connections of the basal ganglia in the mammal as loops that originate in different regions of the cortex, and project back to subregions of the originating cortex (Figure 2). For example, a sensorimotor loop has its origins in sensorimotor cortex. Excitatory sensorimotor cortex projections to striatum activate cells whose output to the globus pallidus is inhibitory. The pallidal neurons also have an inhibitory effect on their target neurons in the ventral thalamus,

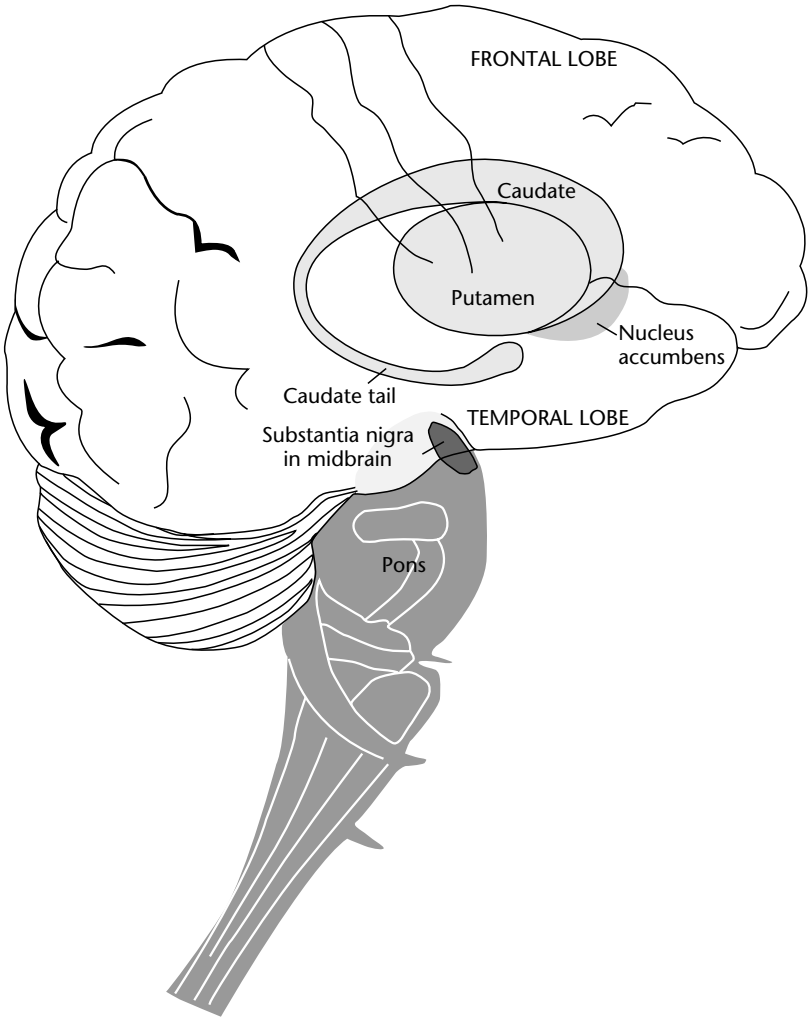


Figure 1. Sagittal (side) view of the human brain shows the central locations of the largest basal ganglia nuclei. The caudate, putamen, globus pallidus and nucleus accumbens are in the forebrain. The substantia nigra is in the midbrain while the subthalamic nucleus (not shown) is at the junction of the midbrain and forebrain.

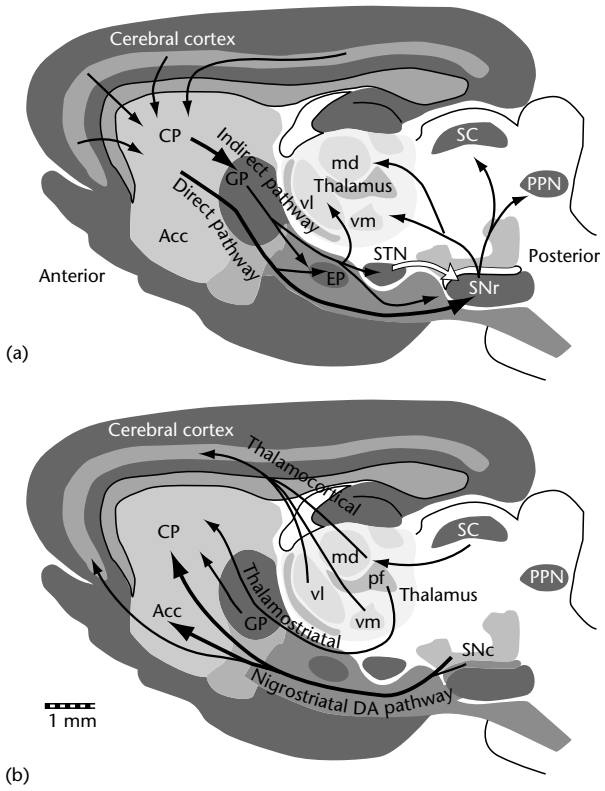


Figure 2. Nuclei and major pathways of the basal ganglia in the rat. The pathways exist also in primates. The view is sagittal, thus showing the extent of the basal ganglia from the forebrain to the midbrain. (a) Descending pathways. All of the cerebral cortex projects to the caudate–putamen (CP). The direct pathway projects to the internal segment of the globus pallidus, called the entopeduncular nucleus (EP) in rats, and to the substantia nigra reticulata before projecting out of the basal ganglia nuclei to the thalamus, superior colliculus and pedunculopontine nucleus. The indirect pathway projects to the subthalamic nucleus and globus pallidus externa (GP in rats) before going to the entopeduncular nucleus and substantia nigra and leaving the basal ganglia. (b) Feedback pathways. The substantia nigra compacta, which contains dopaminergic cells, sends an important projection called the nigrostriatal dopamine pathway to the caudate–putamen and nucleus accumbens. The thalamocortical projection completes the loop from cortex to basal ganglia to thalamus and back to cortex. Acc, nucleus accumbens; CP, caudate–putamen; DA, dopamine; EP, entopeduncular nucleus; GP, globus pallidus; md, mediodorsal nucleus; pf, parafascicular nucleus; PPN, pedunculopontine nucleus; SC, superior colliculus; SNc, substantia nigra compacta; SNr, substantia nigra reticulata; vl, ventrolateral nucleus; vm, ventromedial nucleus. Adapted from Gerfen and Wilson, 1999.

which in turn has an excitatory projection to the supplementary motor cortex, a region known to be involved in the planning of complex movements. The loop circuits are examples of brain regulatory systems that use inhibition of a tonic inhibitory drive as a gate to release normally suppressed output. Other important basal ganglia circuits include an oculomotor loop, which has a major output pathway to the superior colliculus, via the reticulata region of the substantia nigra, in addition to the cortical frontal eye fields, via the thalamus; a dorsolateral prefrontal cortex loop; and a so-called limbic loop.

In addition to the fundamental circuit described above, a critically important input to the striatum comes from the compacta region of the substantia nigra via the nigrostriatal tract (Figure 2). One of the major dopamine pathways in the brain, its integrity is essential for normal functioning of the basal ganglia. Parkinsonism is the direct result of interference with nigrostriatal transmission. Finally, complex and specialized interactions among basal ganglia structures involve the subthalamic nucleus through the 'indirect pathway' (Figure 2).

Chemoarchitecture

The basal ganglia are a collection of heterogeneous nuclei. The striatum contains small zones with a dense concentration of opioid receptors, called 'striosomes' or 'patches' (Figure 3), within a 'matrix' rich in cholinesterase. Many other neurally active substances (such as somatostatin, calbindin and substance P) and receptor proteins (such as the $GABA_A$ receptor) are unevenly distributed between the striosome and matrix compartments. Although the functional significance of this chemoarchitectonic heterogeneity is largely unknown, cells in the patch compartment are active following administration of some dopaminergic drugs, and may play a major part in regulation of dopamine in substantia nigra.

Other components of the basal ganglia have a distinctive complex of cell types, neurochemicals, receptor proteins and synaptic connections that interact in a highly structured fashion. For example, regions thought to have more cognitive than motor functions are rich in calbindin, a calcium binding protein, while those thought to regulate motor functions are calbindin-poor. In addition, the major target of the neostriatum, the globus pallidus, has two segments, internal and external, divided by different inputs; the external segment also has a much higher concentration of enkephalin. Finally, the substantia nigra has two components: reticulata

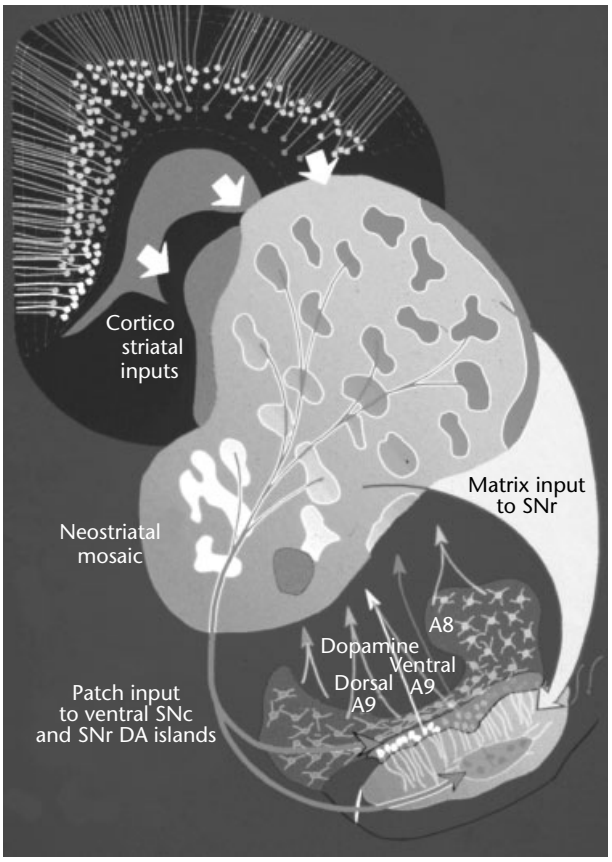


Figure 3. [*Figure is also reproduced in color section.*] The patch–matrix design of the striatum in the rat, seen in the right side of the rat cortex, striatum, nucleus accumbens and substantia nigra. The patches (red, orange and yellow zones), also called striosomes, are rich in an opioid receptor. The surrounding matrix is rich in acetylcholinesterase and calbindin. The anatomical connections of the patches and matrix are color-coded to show, for example, that the cells in the deep layers of lateral cortex project to the lateral patches in the striatum, which in turn project to the ventral substantia nigra compacta dopaminergic cells. SNr, substantia nigra reticulata; SNc, substantia nigra compacta; A9, nomenclature that refers to substantia nigra compacta cells; A10, refers to dopaminergic cells of the midbrain tegmentum that project to the nucleus accumbens. Diagram courtesy of C. Gerfen. >)

and compacta (Figure 3). The reticulata component contains GABA-mediated neurons, while compacta neurons are dopaminergic, their axons being the nigrostriatal pathway.

FUNCTIONS OF THE BASAL GANGLIA

Sensorimotor Functions

Tennis professionals and pianists practice constantly. Eventually their skills become unconscious habits, and during performance they do not think about them in any detail. Indeed, we all have large repertoires of highly practised motor skills (buttoning a shirt, riding a bicycle, driving a car) that require sensory feedback during learning, but are eventually executed without thought or verbalization. Consider the task of driving a car with a manual gear shift, in which each of the four limbs is simultaneously engaged in a different task! Evidence is currently accumulating that the basal ganglia have a critical role in regulation, learning and execution of such motor habits.

We know from animal studies that electrical stimulation of small areas of the caudate–putamen produces movement in localized body regions such as the tongue or contralateral wrist. Stimulation of larger areas produces contralateral head movements or turning, or flexion of the limbs. This reflects the fact that the output of the basal ganglia affects the elements of which the skilled behaviors are constructed. They play a critical part in innate motor sequences (e.g. grooming in rodents), voluntary movements (e.g. limb trajectory) and the sensory feedback that guides directional movement, all of which are essential for developing complex trained sequences or motor habits.

Injection of dopamine directly into the rat neostriatum produces contralateral turning, while small neostriatal lesions prevent the normal sequencing of movements seen during grooming. In rodents and subhuman primates, an abnormal increase in dopamine at synapses in the basal ganglia causes repetitive, stereotyped behaviors: uncontrolled, purposeless, repetitive movements. Rodents will continually chew or sniff, or move back and forth repetitively. In humans, stereotypies may be observed in following administration of dopaminergic drugs, including cocaine.

Note, however, that all of these affected elements of motor behavior are organized at the spinal cord and brainstem level. It is their release, suppression and timing that seem to be regulated by the basal ganglia. Rather than programming these behaviors directly, the basal ganglia apparently implement sequences of behavioral elements that achieve complex behavior.

Additional insight comes from electrophysiological studies. Striatal cells fire in relation to the serial order of innate grooming sequences in rodents, or to learned movement sequences in primates. Inactivating the globus pallidus neurons by cooling disrupts smooth performance of movements by causing continuous flexion of the limbs. Also, cells in the caudate–putamen and globus pallidus fire before or during a learned movement,

when a limb is passively moved, or when the animal produces a spontaneous, unlearned movement. Interestingly, cells do not fire in relation to the amplitude or speed of a movement. In addition, neurons in substantia nigra reticulata change their activity before and during an eye movement. The substantia nigra and superior colliculus are more involved in head and eye movements, while the globus pallidus appears to be involved in limb and trunk movements.

Note, again, that none of these functions is exclusive to the basal ganglia. Basal ganglionic damage may prevent a movement by compromising the release mechanism, even while other parts of the brain (e.g. cerebellum, cortex) retain both essential and redundant circuits for execution of the task. The result of such damage is serious disruption and even prevention of behavior. The basal ganglia can also influence muscle tension and extensor/flexor balance, which are largely controlled by the brainstem and spinal cord. Rather than acting as a motor control system, the basal ganglia and its related structures apparently function as a gating mechanism for innate and complex movement patterns that are organized at lower levels.

Cognitive Functions

Just as the basal ganglia have a role in the unconscious aspects of motor control, they also affect the unconscious aspects of cognition. Such processes are difficult to study and require subtle investigative approaches. Spatial working memory, one of the first cognitive functions of the basal ganglia to be recognized, is linked closely to similar functions of the cortex. Recall that the targets of basal ganglia projections are regions of the cortex involved in motor planning and higher-order cognitive functions, such as spatial working memory. Caudate–putamen lesions consistently produce deficits in spatial learning tasks, and imaging studies in humans show that learning a complex spatial task such as the Tower of London puzzle involves the caudate. In addition, cells of the basal ganglia are involved in detecting and evaluating the learned context of an environmental stimulus. They fire in relation to conditioned stimuli and also when a stimulus becomes relevant, which suggests that the basal ganglia may affect cognitive behavior on a global scale. For example, detecting a primary reward such as a sweet taste, or learning that money is rewarding, or expecting a reward, all profoundly affect our behavior and and its planning. In human brain-mapping studies, several categories of reward such as cash, the 'rush' and 'high' of cocaine, and verbal feedback, activate the caudate, putamen and nucleus accumbens. Activation in the nucleus accumbens tracks the value of a monetary reward during a gambling task, even when there is only a prospect of winning money. Thus, the anticipation of reward, a highly influential – and global – role in behavior is also represented in the basal ganglia.

Studies of people with basal ganglia disease add further insight. People with Parkinson disease experience attention deficits and difficulty in shifting mental set. They are poor at learning to predict probabilistic classifications, which are similar to unconscious, learned, motor habits. This involves unconscious memory and selection of a group of objects that are correct more often than not during a training series. People with globus pallidus lesions describe their mental life as empty; they have no spontaneous thoughts. In Tourette syndrome, people experience racing thoughts. People who suffer Huntington disease exhibit symptoms of obsessive–compulsive disorder prior to the onset of severe motor symptoms.

CONCLUSION

The basal ganglia are subcortical forebrain and midbrain nuclei that process information from the cortex to affect movement and cognition. Physiological and pathophysiological studies indicate that the basal ganglia influence stopping, starting and switching behaviors at the unconscious level; that they play a role in complex visuospatial tasks; and that they have access to learned motor and cognitive habits. An understanding of the normal functions of the basal ganglia nuclei remains unclear, perhaps because global, executive, decision-making processes are based on so many factors. 'Adaptive motor control' may be a good description of their functions, not only for motion, but also for strategic planning of movements and tasks.

Further Reading

Breiter HC, Aharon I, Kahneman D, Dale A and Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**: 619–639.

Brown LL, Schneider JS and Lidsky TI (1997) Sensory and cognitive function of the basal ganglia. *Current Opinion in Neurobiology* **7**: 157–163.

Cromwell HC and Berridge KC (1996) Implementation of action sequences by a neostriatal site: a lesion mapping study of grooming syntax. *Journal of Neuroscience* **16**: 3444–3458.

Gerfen CR and Wilson CJ (1996) The basal ganglia. In: Swanson LW, Bjorklund A and Hokfelt T (eds) *The Handbook of Chemical Neuroanatomy* vol. 12, Integrated Systems of the CNS, Part III, pp. 371–468. New York, NY: Elsevier.

Jog MS, Yaso K, Connolly CI, Hillegaart V and Graybiel AM (1999) Building neural representations of habits. *Science* **286**: 1745–1749.

Kawagoe R, Takikawa Y and Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience* **1**: 411–416.

Kermadi I and Joseph JP (1995) Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology* **74**: 911–933.

Knowlton BJ, Mangels JA and Squire LR (1996) A neostriatal learning system in humans. *Science* **273**: 1399–1402.

Knutson B, Adams CM, Fong GW and Hommer D (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience* **21**: RC159.

McDonald RJ and White NM (1994) Parallel information processing in the water maze: evidence for independent memory systems involving dorsal striatum and hippocampus. *Behavioral Neural Biology* **61**: 260–270.

Behavioral Neuropharmacological Methods

Introductory article

Roy A Wise, National Institute on Drug Abuse, Baltimore, Maryland, USA

CONTENTS

The neural systems responsible for cognitive function are segregated by anatomical and neurochemical specificity. Characterization of the effects of drugs on thought and action helps us to identify the basic functional units of neuronal and cognitive organization.

INTRODUCTION

Neuropharmacology deals with the 'chemical coding' of brain circuitry and the ability to influence and study this circuitry selectively, either through the use of drugs or in an effort to understand their action. Neurons are normally activated or inhibited by endogenous substances released from nerve cells and variously termed neurotransmitters, neuromodulators or neurohormones. These chemical messengers act at specialized receptor proteins embedded in the surface membranes of nerve cells. Because of peculiarities in their geometry, the receptors bind their appropriate messenger very selectively; transmitters and hormones fit their receptors in much the same way as a key fits a lock. Neuropharmacology deals with the activation or inhibition of neurons by chemicals, including exogenous substances – drugs – that can mimic, augment or block the functions of endogenous transmitters. Traditionally neuropharmacologists used electrophysiological recording techniques to measure the responses of single neurons to chemicals delivered systemically or locally. Behavioral neuropharmacology is the study of neurochemical control of brain function as it is reflected in behavior.

Behavioral neuropharmacological methods have key roles in the analysis of behavior. Much of behavior is dominated by neuropharmacological variables. The roles of hormones in stress and sexual behavior, the roles of nutrients and hormones in feeding and drinking behavior, the roles of endogenous opioids in control of pain, and the roles of drugs in addiction illustrate the importance of neuropharmacology in behavioral analysis. Behavioral neuropharmacology not only teaches us ways to control behavior; it contributes to our understanding of neuronal organization and function. The chemical selectivity of brain circuitry gives us major clues to the structure and function of the brain. As pointed out several decades ago by Donald Hebb, our best theories about the nature of cognition are always constrained by our current understanding of the structure and activity of the functional units of the brain.

MODIFYING NEUROTRANSMITTER SYSTEMS

The neurotransmitter systems of the brain can be selectively manipulated in several ways. The most direct ways involve drugs that bind to the receptors for endogenous neurotransmitters. Such drugs share the geometrical features that allow the neurotransmitter to bind to receptors on a target neuron. If the drug binds to the receptor for a given transmitter but does not trigger a transmitter-like action there, it is an antagonist of the transmitter, often termed a 'receptor blocker' because it physically blocks the access of the endogenous transmitter to its normal binding site. The effects of drugs on a neurotransmitter system are usually temporary.

If the drug's geometry is sufficiently similar to that of the transmitter it will not only bind to the receptor but also trigger the biological action of the transmitter; in such cases the drug is termed an 'agonist', because it mimics the transmitter action. Nicotine is an agonist at a subset of acetylcholine receptors; morphine is an agonist at receptors for endogenous 'opioid' peptide neurotransmitters. Phencyclidine is an antagonist at a subset of glutamate receptors.

Amphetamine and cocaine are termed 'indirect' monoamine agonists because although they do not

bind to monoamine receptors, they act at the transporter molecules that dispatch or take back up the monoamine transmitters. By reversing or blocking the reuptake mechanism, they elevate the synaptic concentrations of the three monoamine transmitters – dopamine, noradrenaline (norepinephrine), and serotonin. These drugs have the same behavioral effects as the transmitters because they increase the transmitter concentration in the local extracellular fluid. Injections of drugs often have more dramatic effects than injections of the transmitters themselves because the drugs are generally much more resistant than the transmitters to the normal deactivation mechanisms.

Neurotransmitter agonists and antagonists have varying degrees of selectivity for different receptors and different receptor subtypes. Cholinergic receptors (defined by their common sensitivity to the neurotransmitter acetylcholine) are of two major subtypes: nicotinic (binding and responding to nicotine but not to muscarine) and muscarinic (binding and responding to muscarine but not to nicotine). There are five or more variations of muscarinic receptors: five slightly different receptor molecules that each bind and respond to muscarine and acetylcholine. There are five known subtypes of dopamine receptor and 13 known subtypes of serotonin receptors. The selectivity of a given agonist or antagonist depends on sometimes subtle geometric peculiarities of the neurotransmitter, the drug and the receptor. One goal of neuropharmacology is to identify drugs that are highly selective for a given receptor molecule.

It is possible to cause permanent damage to single neurotransmitter systems or portions of single neurotransmitter systems by the use of neurotoxins. The substance 6-hydroxydopamine is a general neurotoxin if it is given in sufficient concentration. However, if given in low concentration, this molecule (which resembles dopamine) is taken up selectively by neurons that express and take up the closely related monoamines dopamine and noradrenaline. These neurons will concentrate the toxin intracellularly. Thus the toxin can be used to cause selective degeneration of dopaminergic and noradrenergic neurons. If a selective blocker for the noradrenergic uptake mechanism is given, the drug will be taken up and damage only dopamine systems. If the toxin is microinjected into a local brain region, it will damage only those dopaminergic or noradrenergic neurons found in that region. There are analogous neurotoxins for serotonergic neurons. Other toxins selectively damage noradrenergic systems; one substance, MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine), is metabolized to a selective toxin for dopamine neurons by an enzyme found in monkeys but not in rats.

Another selective degeneration method involves coupling the general ribosome-inactivating protein saporin to an antibody that gives it selectivity for neurons of a given neurotransmitter type. Saporin-linked antibodies have been used successfully to destroy forebrain cholinergic neurons and, more recently, noradrenergic pathways. The linkage of saporin to other antibodies will allow it to be used to target other transmitter systems. Other neurotoxins, particularly excitatory amino acids, have little selectivity for particular neurotransmitter systems but, when injected locally, can be used to damage the cell bodies of a given region while sparing passing fiber systems.

In addition to neuropharmacological methods for selective modulation of neurotransmitter systems, new molecular biological approaches are rapidly being developed. The expression of receptors, enzymes or neurotransmitters can be blocked at the level of gene transcription by antisense oligonucleotides, and viral vectors can be used to insert genetic material that transiently increases expression of such gene products. There is thus an expanding range of methods for selective activation, blockade or destruction of particular neurotransmitter systems. Each of these methods can be used to study the behavioral role of a given neurotransmitter or the behavioral sensitivity to a given drug.

MEASURING NEUROTRANSMITTER ACTIVITY

The spontaneous activity of various neurotransmitter systems can be measured in several ways. Extracellular recordings of the electrophysiological activity of single neurons are useful when the cell type can be identified by anatomical location, firing patterns or sensitivity to different drugs. Unique firing patterns have been characterized from intracellular recordings in simplified preparations *in vitro* where a dye can be injected into the cell for subsequent identification of neurotransmitter type. Unfortunately, not all cell types have unique electrophysiological signatures, and the electrophysiological characteristics of some cell types differ considerably between conditions *in vitro* where identification is positive and conditions *in vivo* where it is not.

There are two approaches to measuring or estimating the concentrations of various neurochemicals in the extracellular fluid of a given brain

region. One involves sampling extracellular fluid from some local region and subjecting that fluid to bench-top assay. The extraction of extracellular fluid was originally accomplished with 'push-pull' cannulas where artificial cerebrospinal fluid was injected through one line and withdrawn through the other. Neurochemicals from the region mixed with the injected fluid and were withdrawn with it. This method is now infrequently used because it collects not only the transmitters but also the enzymes that continue to degrade them. In a more recent method the tips of push-pull cannulas are encapsulated in microdialysis tubing which allows the small neurotransmitters to diffuse into the perfusate, but blocks entry of the larger enzymes. This technique for collecting brain chemicals is termed 'microdialysis' (Figure 1).

When microdialysis samples are assayed, unambiguous identification of neurochemicals is possible. Microdialysis is currently used widely in conjunction with high-performance liquid chromatography to assay acetylcholine, glutamate, γ-aminobutyric acid (GABA) and the monoamine transmitters dopamine, noradrenaline (norepinephrine) and serotonin. The many neuropeptide transmitters tend to be released in very low concentrations, and are more difficult to assay with current methods. Mass spectrometric methods with much greater sensitivity are currently under development for this purpose.

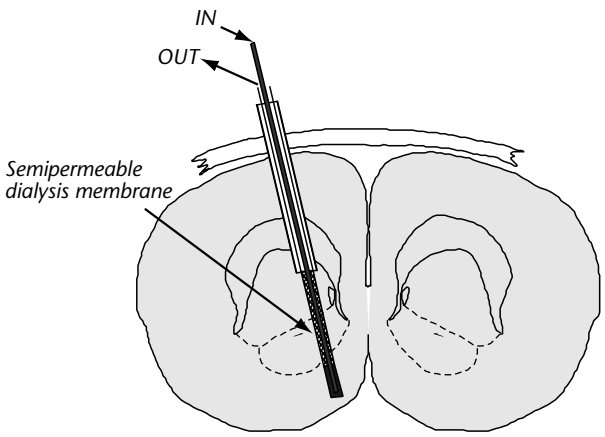


Figure 1. A microdialysis probe is inserted into a rat brain. The membrane portion of the probe is semipermeable, and (like a blood vessel) allows the exchange of brain chemicals between the artificial extracellular fluid which is perfused through it and the brain's extracellular fluid which surrounds it. The perfusing fluid is pumped slowly into the internal cannula, passes back up between the inner cannula and the membrane itself – where it absorbs neurotransmitters and other substances from the endogenous brain fluids – and is collected in a vial connected by flexible tubing to the outer cannula. The neutral perfusion fluid and the collected brain chemicals that it carries out of the brain are then analyzed to determine what neurotransmitters are being released under the conditions of testing.

The microdialysis approach allows confident identification of a given chemical species, but the temporal resolution is poor for behavioral studies. It often takes several minutes to collect detectable levels of a given transmitter, and it is not possible to identify peaks of neurotransmitter release within a given sampling period. Thus, even though it is clear which chemicals were released, it is difficult to determine precisely when they were released. A technique with greater temporal resolution is *in-vivo* voltammetry. Here a neurochemical assay is performed within the brain. Different chemicals oxidize at different voltages, and a voltage appropriate for the chemical of interest is applied locally. The voltage-specific oxidation of the chemical in question induces a measurable increment in the current flow between the electrodes. Fluctuations of these oxidation currents are used to estimate fluctuations in the concentration of the chemical of interest. Unfortunately, more than one chemical species may oxidize at a given voltage, and thus neurochemical resolution is not as good as in microdialysis.

Each of these methods is undergoing rapid refinement. The temporal resolution of microdialysis is being improved by development of more sensitive laboratory assays. The neurochemical resolution of voltammetry is being improved by development of different electrodes and by the use of pharmacological tools in conjunction with locally stimulated release of transmitters. The development of more precise methods for monitoring neurochemicals in freely moving animals is receiving active attention in several laboratories.

Another use for microdialysis is to infuse substances into local brain regions. Just as local neurochemicals diffuse from the brain across the dialysis membrane and into dialysis samples, neurochemicals in the dialysis fluid can diffuse into the brain. Drug infusion by dialysis is more localized and less stressful for local tissue than are bolus injections by the more traditional hydraulic pressure. If two dialysis probes are implanted in appropriate regions, it is possible to infuse a drug at one site and observe the consequences on behavior and on neurotransmitter release at the second site. Such studies begin to identify the connections between sites of drug action and help characterize the distal transmitter release caused by a drug.

THE STUDY OF COGNITION

The coding of information flow in the brain is both spatial and neurochemical. Our eventual understanding of the brain processes underlying various cognitive processes will require us not only to identify the portion of the brain that is sending and receiving relevant messages, but also which chemical messenger – among the dozens that can be found in most brain regions – is carrying the signals.

CONCLUSION

The realization that the circuits of the brain contain a wide range of chemical messengers, and that dozens of chemically coded messages can be transmitted simultaneously in the same brain region, has motivated the development of methods for distinguishing the neurochemical constituents of brain fluid, measuring their fluctuations during thought and action, and determining their consequences by microinfusing them into the brains of freely moving animals. Anatomists have come far in identifying the cells of different brain regions, the source of their inputs, and the targets of their outputs. They have also made great progress in identifying – through fluorescence and immunohistochemical techniques – the chemical messengers in various cell groups.

Neuropharmacologists build on such information to probe the functions of chemically distinct neuronal populations. Neuropharmacological methods help us to understand the effectiveness of various drugs in alleviating the symptoms of such conditions as schizophrenia, depression and Parkinson disease. Such methods are used increasingly to develop new drugs – with greater potency and fewer side effects – for the treatment of mood and thought disorders. Neuropharmacological methods also help us to understand basic brain function; they help us to parse and discover the syntax of the activity of the brain. Relative to the complexity of brain function, current neuropharmacological methods are still crude; none the less, the rate of technological advance is rapid and neuropharmacological methods have already given us major insights into the functional units of mood and movement.

Further Reading

Benveniste H (1989) Brain microdialysis. *Journal of Neurochemistry* **52**: 1667–1679.

Boulton AA, Baker GB and Adams RN, eds (1995) *Neuromethods. Voltammetric Measurements in Brain Systems*. Totowa, NJ: Humana Press.

Emmett MR and Caprioli RM (1994) Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. *Journal of the American Society for Mass Spectrometry* **5**: 605–613.

Hebb DO (1955) Drives and the CNS (conceptual nervous system). *Psychological Review* **62**: 243–254.

Kiyatkin EA and Rebec GV (1998) Heterogeneity of ventral tegmental area neurons: single-unit recording and iontophoresis in awake, unrestrained rats. *Neuroscience* **85**: 1285–1309.

Miller NE (1965) Chemical coding of behavior in the brain. *Science* **148**: 328–338.

Morari M, O'Connor WT, Darvelid M *et al*. (1994) Functional neuroanatomy of the nigrostriatal and striatonigral pathways as studied with dual probe microdialysis in the awake rat – I. Effects of perfusion with tetrodotoxin and low-calcium medium. *Neuroscience* **72**: 79–87.

Quan N and Blatteis CM (1989) Microdialysis: a system for localized drug delivery into the brain. *Brain Research Bulletin* **22**: 621–625.

Routtenberg A (1972) Intracranial chemical injection and behavior: a critical review. *Behavioral Biology* **7**: 601–641.

Wise RA and Hoffman DC (1992) Localization of drug reward mechanisms by intracranial injections. *Synapse* **10**: 247–263.

You ZB, Herrera-Marschitz M, Nylander I *et al* (1996) Effect of morphine on dynorphin B and GABA release in the basal ganglia of rats. *Brain Research* **710**: 241–248.

Blindsight, Neural Basis of

Intermediate article

Carlo A Marzi, University of Verona, Verona, Italy

CONTENTS

Blindsight is the presence of unconscious visually guided behavior elicited by stimuli presented within the visual field loss of patients with damage to the primary visual cortex.

INTRODUCTION

The term 'blindsight' to describe the presence of unconscious visually guided behavior in people with a lesion of the primary visual cortex was coined by Weiskrantz *et al*. (1974) and should not be confused with the term 'residual vision', which defines conscious visually guided behavior following a visual cortical lesion. The essence of blindsight lies primarily in the lack of conscious awareness in the presence of an above-chance visual performance. The first demonstration of blindsight was the target localization by eye movements reported by Poeppel *et al*. (1973), followed by target localization by manual pointing and target detection (Weiskrantz *et al*., 1974). Many other functions have been since then tested with different success in humans and in nonhuman primates (Stoerig and Cowey, 1997). The importance of blindsight in neuroscience research is threefold. First, it has opened the way to the scientific investigation of conscious awareness, a topic before relegated to the domain of philosophy. Second, it has reconciled the discrepancy between human and nonhuman primates as far as the effect of the lesion of the primary visual cortex is concerned; lesions of the primary visual cortex in monkeys (or in cats), although severely impairing visual acuity, leave some visually guided behavior intact, while this is typically not the case in humans. Research has shown, however, that visually guided behavior in humans may persist following a primary visual cortex lesion, but it remains unconscious. Third, blindsight might be an initial stage in the return of vision; unfortunately, the correlation between presence of blindsight and successful rehabilitation of conscious visual function is not good. However, blindsight can improve with training and there is evidence of an increase in unconscious visual sensitivity over the years (Stoerig and Cowey, 1997). Whether this may lead to recovery of conscious vision is still an open question.

An important distinction is between direct and indirect methods of testing blindsight. The former include a forced-choice procedure similar to that used in animal testing, in which the person is asked to guess despite lack of stimulus awareness. In contrast, in the indirect procedure, the person does not have to guess but the presence of blindsight is inferred from the influence of unseen stimuli on the response to stimuli presented to the normally sighted hemifield.

LESION SITE

Unilateral complete lesions of the primary visual cortex (also known as striate cortex, area 17 in the Brodmann nomenclature, or area V1 in the jargon of electrophysiology) result in contralateral homonymous hemianopia: that is, the entire hemifield on the side opposite to the lesion is blind as assessed by clinical perimetric examination. Partial lesions, result in a scotoma which may affect various portions of the contralesional hemifield. Finally, bilateral complete lesions of the visual cortex result in cortical blindness that affects the whole visual field.

Which Regions Must Be Affected to Produce the Deficits?

Damage to visual centres other than V1 may result in a hemianopia but such lesions may not be compatible with blindsight when crucial centres are visually deafferented. This is the case with lesions of the optic tract, which funnels visual information not only to the geniculostriate system but also to the superior colliculus and other visual midbrain areas, or to thalamic areas such as the pulvinar

which relay visual information to cortical areas bypassing the primary visual cortex. When these areas are deafferented in addition to primary visual cortex, no blindsight is possible.

WHICH REGIONS MEDIATE SPARED FUNCTION?

There are two main candidates as centres mediating the spared unconscious functions characterizing blindsight: the superior colliculus and the extrastriate visual cortex. The superior colliculus (SC) projects indirectly through the thalamic pulvinar to extrastriate cortical areas such as V2, V3, V4 and V5 (also known as MT, see below) and to visually responsive temporal areas. Therefore, in the absence of V1, visual input can still activate these cortical areas. In addition, a small number of cells – mainly located in interlaminar zones of the dorsal lateral geniculate nucleus (dLGN) – project directly to extrastriate cortex, rather than to V1 like the majority of dLGN neurons. The difference between the relative contribution of these structures has been investigated by comparing the effects of lesions restricted to the primary visual cortex with the effects of lesions including the extrastriate cortex and those of loss of an entire hemisphere (Azzopardi *et al*., 2001). An intriguing conclusion emerging from those studies is that following hemispherectomy – that is, in the absence of both V1 and extrastriate cortex – the remaining SC and pulvinar cannot subserve voluntary responses such as those employed in the direct methods to test blindsight described above. In contrast, there are indications that some hemispherectomy patients can still show blindsight when tested with indirect methods, as reported by Tomaiuolo *et al*. (1997).

More evidence on the areas mediating blindsight comes from brain imaging studies in which visual stimuli are presented within the perimetrically blind area of the visual field. It has been shown with functional magnetic resonance imaging (fMRI) that, despite absence of activation of their lesioned V1, hemianopic patients show preserved responsiveness in area V5 with moving stimuli, and in areas V4 and V8 within the fusiform gyrus with colored images of objects (Goebel *et al*., 2001). The former is an area, or rather a complex of areas, which belong to the so-called dorsal system including a series of cortical regions mediating spatial perceptual and visuomotor functions. The area MT is named for its anatomical location in the monkey's middle temporal sulcus in the proximity of the junction of temporal, parietal and occipital lobes. It contains neurons selectively sensitive to the direction and velocity of motion stimuli. The human homolog of area MT is also known as area V5: its bilateral lesion results in a severe impairment in the detection of moving stimuli: see Zeki (1991) for a review.

In contrast to the dorsal stream, that is, a series of cortical regions, the ventral stream includes a host of cortical visually responsive areas whose neurons respond preferentially to forms, colors and natural objects. Areas within the fusiform gyrus (V4 and V8) belonging to this system have been found to be selectively activated by appropriate stimuli presented within the hemianopic field. It is important to consider that both the dorsal and the ventral system activations were not accompanied by conscious awareness of the stimuli. This shows that activation of extrastriate cortex *per se* is not sufficient to yield stimulus awareness.

WHICH FUNCTIONS ARE SPARED?

Blindsight functions can be divided into two categories: direct or voluntary responses usually elicited with a forced-choice procedure, and indirect responses usually tested by assessing the influence of unseen stimuli (presented to the hemianopic field) on stimuli presented to the normal hemifield.

Direct Response

Direct responses include basic functions such as simple detection of stationary or moving stimuli, spatial localization and discrimination of direction of motion, stimulus displacement, wavelength and orientation. The latter ability seems to be crucial for the apparent form discrimination exhibited by some people with blindsight. In fact, it has been shown that shape discrimination in the hemianopic field is impossible when orientation cues are eliminated. One interesting dissociation has been repeatedly described, namely that between action and perception. Some people who are unable to discriminate forms or objects because of cortical damage can none the less show reaching or grasping hand movements that are appropriate for the size and orientation of the stimuli presented. This means that information that is not available for conscious perception can be used for motor action on the same object.

Indirect Response

An example of this approach is the use of the redundant target effect (RTE), with bilateral stimuli

presented across the vertical meridian of the visual field. With this paradigm, simple manual reaction time in response to bilateral brief visual stimuli is typically faster than for unilateral single stimuli. It has been shown that this is the case even when one stimulus in the pair has been presented to the hemianopic hemifield of patients with a V1 lesion. Despite their claim of having seen only one stimulus they show an RTE with bilateral stimuli. The speeding up of reaction time by an unseen stimulus can thus be taken as indirect evidence of blindsight (Marzi *et al*., 1986). A similar implicit RTE has been found in patients with an hemianopia resulting from hemispherectomy performed as an extreme therapy for intractable epilepsy (Tomaiuolo *et al*., 1997). Another example of indirect approach is the demonstration that distractor signals in the blind half of the visual field could inhibit saccades toward targets in the intact visual field (Rafal *et al*., 1990).

RELEVANCE TO THEORIES OF CONSCIOUSNESS

One of the merits of research on blindsight is to have given impulse to a neuroscientific study of consciousness and to have aroused the interest of philosophers in the neural basis of conscious experience. A few years ago an experiment was attempted to answer this fundamental question (Sahraie *et al*., 1997). In an fMRI experiment on a blindsight patient extensively investigated in other studies (GY), stimulation of the blind hemifield yielded conscious or unconscious above-chance visual performance depending upon the velocity of motion stimuli. The patient was required to discriminate the direction of a moving stimulus; with slower stimuli, discrimination was as good as with faster stimuli but stimulus awareness was lost. The main thrust of the study was to provide evidence for a shift in the pattern of neural activation from cortical to subcortical neural structures associated with the change from conscious to unconscious vision. Notably, the superior colliculus was activated in the unconscious mode alone. These results confirm earlier views that blindsight may be mediated by subcortical structures. However, as pointed out by Searle (2000), they cannot be extended to consciousness in general because the patients only exhibit blindsight if they are already conscious. For a clue to the neural mechanisms of consciousness in general it is necessary to demonstrate that there are neural structures that are crucial for shifting from unconsciousness to a conscious state. So far, this evidence has not been provided.

OTHER POSSIBLE INTERPRETATIONS

In principle, blindsight effects could be the result of various spurious factors (Campion *et al*., 1983). One possibility is light-scattering: this possible source of artefact has been taken care of by minimizing the light intensity of the stimuli and, ingeniously, by introducing control trials in which stimuli are presented to the blind spot of the hemianopic hemifield, that is, to an area corresponding to a retinal zone without photoreceptors. Real blindsight cannot survive a blind spot presentation, and this is what has been found in the majority of cases; see for example Tomaiuolo *et al*. (1997).

Another possible source of spurious blindsight comes from a shift in the decision criteria adopted by the hemianopic patient in moving from clinical visual field assessment to experimental blindsight testing. In the former situation the patient may adopt a more conservative criterion and deny having seen something, despite some near-threshold residual vision. In contrast, under laboratory conditions, especially when using forced-choice procedures, the patient may adopt a more liberal criterion and performance might improve. A specific answer to this type of criticism has been provided by Azzopardi and Cowey (1997) who measured the sensitivity of a hemianopic patient independently of his response criterion. They found that, in contrast to normal control subjects, sensitivity was higher during the forced-choice task than during a procedure similar to that used in clinical visual field testing in which a conscious report is required ('Do you see the stimulus?'). This means that blindsight cannot be simply assimilated to normal near-threshold vision and that a mere shift of response criterion cannot explain it.

Finally, it has been proposed (Campion *et al*., 1983; Fendrich *et al*., 1992) that blindsight may be related to islands of spared visual cortex yielding correspondingly small areas of visual field preservation that can be documented only with special techniques (Fendrich *et al*., 1992). In contrast to this possibility, however, it has been found that two blindsight patients studied with fMRI did not show any activation of V1 (Goebel *et al*., 2001) and therefore it is unlikely that their blindsight might have been related to spared V1, although this might explain other cases. All in all, one can safely conclude that blindsight is a genuine phenomenon, but its investigation requires careful control of all possible sources of artefact.

References

Azzopardi P and Cowey A (1997) Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Science USA* **94**: 14190–14194.

Azzopardi P, King SM and Cowey A (2001) Pattern electroretinograms after cerebral hemispherectomy. *Brain* **124**: 1228–1240.

Campion J, Latto R and Smith Y M (1983) Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behavioural Brain Sciences* **6**: 423–486.

Fendrich R, Wessinger CM and Gazzaniga MS (1992) Residual vision in a scotoma. Implications for blindsight. *Science* **258**: 1489–1491.

Goebel R, Muckli L, Zanella FE, Singer W and Stoerig P (2001) Sustained extrastriate cortical activation without visual awareness revealed by fMRI studies of hemianopic patients. *Vision Research* **41**: 1459–1474.

Marzi CA, Tassinari G, Aglioti S and Lutzemberger L (1986) Spatial summation across the vertical meridian in hemianopics: a test of blindsight. *Neuropsychologia* **24**: 749–758.

Poeppel E, Frost D and Held (1973) Residual visual function after brain wounds involving the central visual pathways in man. *Nature* **243**: 295–296.

Rafal R, Smith J, Krantz J, Cohen A and Brennan C (1990) Extrageniculate vision in hemianopic humans: saccade inhibition by signals in the blind field. *Science* **250**: 118–121.

Sahraie A, Weiskrantz L, Barbur JL *et al*. (1997) Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Science USA* **94**: 9406–9411.

Searle JR (2000) Consciousness. *Annual Review of Neuroscience* **23**: 557–578.

Stoerig P and Cowey A (1997) Blindsight in man and monkey. *Brain* **120**: 535–559.

Tomaiuolo F, Ptito M, Marzi CA, Paus T and Ptito A (1997) Blindsight in hemispherectomized patients as revealed by spatial summation across the vertical meridian. *Brain* **120**: 795–803.

Weiskrantz L, Warrington EK, Sanders MD and Marshall J (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* **97**: 709–728.

Zeki P (1991) Cerebral akinetopsia (visual motion blindness). A review. *Brain* **114**: 811–824.

Further Reading

Holmes G (1945) Ferrier lecture. The organisation of visual cortex in man. *Proceedings of the Royal Society (London) Series B* **132**: 348–361.

Marzi CA (1999) Why is blindsight blind? *Journal of Consciousness Studies* **6**: 12–18.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. New York: Oxford University Press.

Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA and Mansfield RJW (eds) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Weiskrantz L (1986) *Blindsight: A Case Study and Implications*. Oxford: Clarendon Press.

Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford: Oxford University Press.

Zeki S (1993) *A Vision of The Brain*. Oxford: Blackwell.

Zihl J (2000) *Rehabilitation of Visual Disorders after Brain Injury*. Hove, UK: Psychology Press.

Brain Asymmetry

Introductory article

Albert M Galaburda, Harvard Medical School, Boston, Massachusetts, USA
Glenn D Rosen, Harvard Medical School, Boston, Massachusetts, USA

CONTENTS

The two cerebral hemispheres are specialized for different functions. The discovery of anatomic asymmetries in the brain has given new light to our understanding of the cognitive differences between the left and right hemispheres.

INTRODUCTION

It is generally accepted that in humans the left hemisphere is specialized for the processing of some aspects of language while the right hemisphere dominates over many spatial, emotional and musical functions. The exact relationship between brain asymmetry and side differences in function, however, is not known. In fact, a century and a half after the initial discoveries of lateralization of language the neural basis for language is still not clearly understood, and as part of this incomplete knowledge the relationship between language lateralization and cerebral asymmetry is at best tentative. Moreover, the biological substrates underlying non-language-based functional lateralization is perhaps even less clear. That being said, there is a wealth of information concerning asymmetries in the brain, giving intriguing insights into cognitive function.

GROSS BRAIN ASYMMETRIES

Sylvian Fissure and Planum Temporale

Left–right asymmetries in the sylvian fissure (Figure 1) have been noted since the end of the nineteenth century. In general, the left sylvian fissure tends to be longer and have a flatter trajectory than its right hemisphere counterpart, which deviates dorsally at the posterior end. This obvious asymmetry led the neurologists Norman Geschwind and Walter Levitsky to examine the portion of the sylvian fissure subsumed by a structure known as the planum temporale in 100 autopsied human brains. The planum temporale, which contains several auditory association cortices, is thought on the left side to be an important portion of the language network of the left hemisphere (Figure 1(b)). Lesions affecting significant portions of the planum temporale, usually on the left side in right-handed people, often lead to Wernicke's aphasia. Geschwind and Levitsky found that 65% of the sample showed a longer left planum temporale, whereas the right planum was longer in only 11% and both plana were equal in the remaining 24%. The results were highly significant and were essentially confirmed by subsequent studies using both postmortem and *in vivo* imaging techniques. (*See* **Aphasia**; **Geschwind, Norman**)

Inferior Frontal Area

The inferior frontal area, composed of the pars opercularis, triangularis and orbitalis of the inferior frontal gyrus, the frontal operculum and the subcentral cortex, is complex, highly folded, and variable among individuals (Figure 1(a)). Asymmetries have been demonstrated in this part of the brain. The ascending limb of the sylvian fissure, which separates the pars opercularis posteriorly from the pars triangularis anteriorly, has been reported to be more often branched on the left side, giving off the diagonal sulcus. Others have found the pars opercularis to be more developed on the left, and still others have found greater surface area on the left side. Wada and colleagues, on the other hand, measured the surface portion of the pars opercularis together with part of the pars triangularis and found it to be larger on the right. (*See* **Frontal Cortex**)

Asymmetries in Fetal and Infant Brains

Gross anatomical asymmetries are present in the cerebral cortex before birth. Even though clearly cerebral dominance can be modified after birth

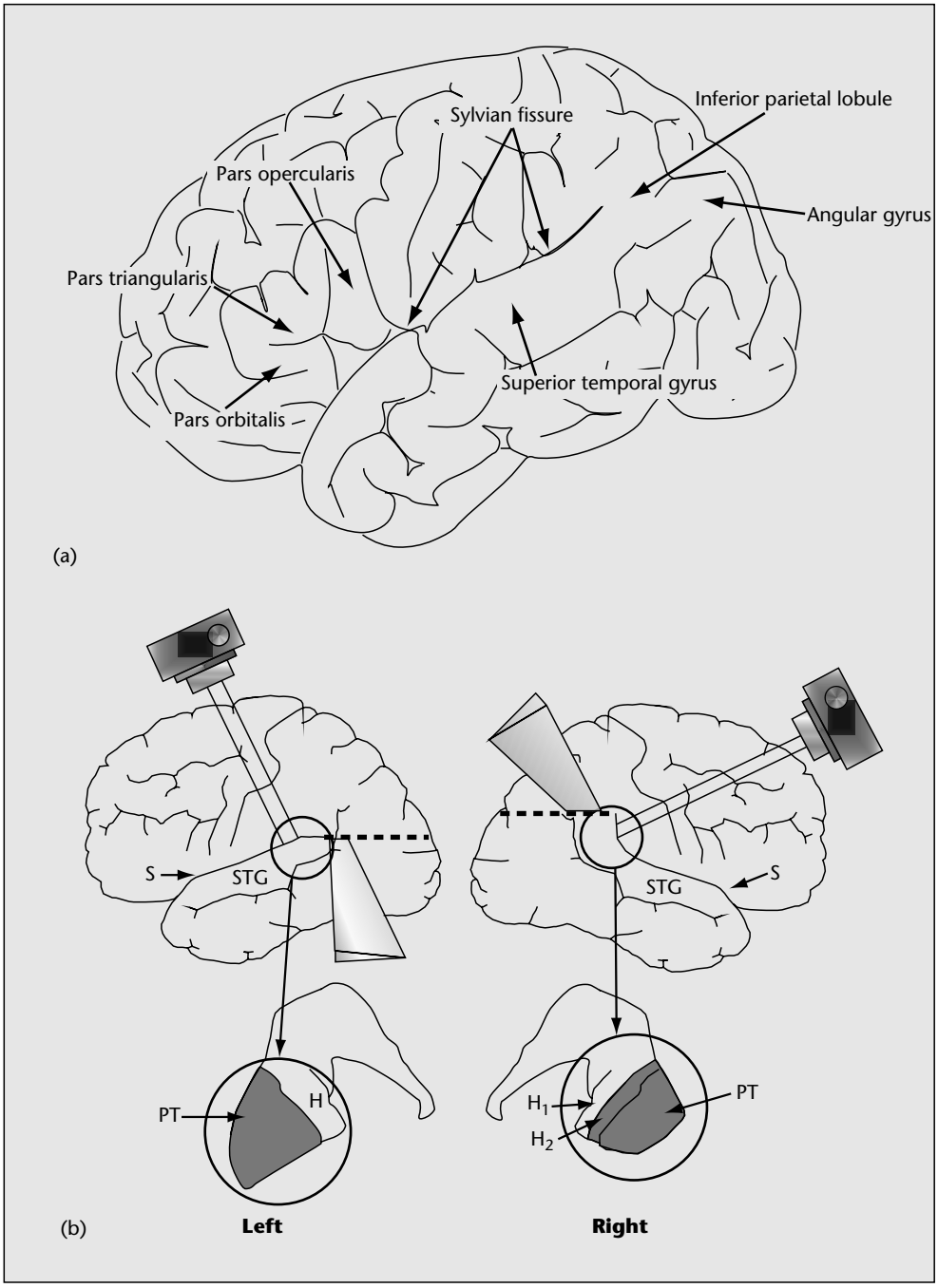


Figure 1. Locations of various structures in the human brain. (a) Lateral view of the left hemisphere identifying areas known to be asymmetric. (b) Lateral views of the right and left hemispheres: note the asymmetric patterns of the sylvian fissure (S), with the left fissure being longer and the right more angled posteriorly. To view the planum temporale (PT), a cut (dotted line) is made at the end of the sylvian fissure (encircled) exposing the superior surface of the superior temporal gyrus (STG). The resulting images are shown below. Here, the planum temporale is larger on the left than the right, and there are two Heschl's gyri (H) on the right side.

(e.g. recovery from early hemispheric injury and ability to switch handedness), the anatomical asymmetries that may underlie functional lateralization are fixed, at least in their gross design, before birth. This is not to say that recovery from early lesions is perfect at any age or that switching handedness leads to equivalent degrees of dexterity with the new hand. We do not know about this because controlled experiments cannot be performed in individual patients and variability in

the population gets in the way of interpreting the results. Thus, we cannot tell in advance how well language would have developed in the ordinary hemisphere before the lesion, or how good the right hand would have become without having had to switch to the left.

The sylvian fissures are visibly asymmetric from about the middle of the gestational period, demonstrating the pattern usually seen in the adult human. The planum temporale is also grossly asymmetric before the end of pregnancy. The study by Witelson and Pallie, which included 14 brains from newborns, found the planum temporale to be larger on the left side in 79% of the cases. Wada and colleagues made planimetric measurements of 100 adult and 100 newborn and fetal brains and found that the average ratio of the surface area of the right to left planum temporale was 67% in the younger brains and 55% in the adults. These authors also noted that the planum temporale asymmetry was visible beginning in the 29th week of gestation. Others have found that the right Heschl's gyrus (Figure 1(b)) was doubled on the right side in 54% of the cases and on the left in 18%.

ARCHITECTONIC ASYMMETRIES

Gross anatomical asymmetry reflects to a great extent asymmetries in underlying architectonic areas, which are subdivisions of cortex with specific cellular architecture, connectivity, and physiology. Cortical folding, however, may also be related to physical forces imposed by the skull during growth, and gross anatomical asymmetries may, therefore, reflect in part bony asymmetries unrelated to brain function. It is important, therefore, to ascertain whether asymmetry in the cerebral cortex can be demonstrated at a level that heralds functional differentiation more strictly than the gross anatomical level of lobes, gyri and sulci – the cytoarchitectonic level. It is possible with experience to draw architectonic borders around many cytoarchitectonic areas and to measure their volume in each hemisphere. This has shown asymmetries in several cortical areas implicated in language function. For instance, area Tpt, which is located on the posterior third of the superior temporal gyrus extending onto the planum temporale, is larger on the left side in the majority of brains. Asymmetry of area Tpt correlates positively with asymmetry of the planum temporale, and in some cases the left can be several times the size of the right. Damage to Tpt may result in problems with language comprehension.

The frontal operculum contains predominantly cytoarchitectonic areas 44 and 45, the former covering most of the pars opercularis and the latter mainly the pars triangularis (Figure 1(a)). Both of these cortices belong to the category of motor association cortex, with 44 being an inferior premotor cortex and 45 an inferior prefrontal cortex, and are concerned with speech production. Area 44, which was enhanced by special stains, was found to be of greater volume in six out of ten brains, nearly symmetrical in three, and larger on the right side in one brain from a collection of neurologically normal brains. There is no information regarding asymmetry of area 45.

A study of the inferior parietal lobule (Figure 1(a)) in 10 normal human brains, a region concerned with word meaning containing areas PF and PG, revealed a predominance of the left area PG in eight of those brains. Area PG (Brodmann's area 39) lies on the angular gyrus, is highly functionally multimodal, and is anatomically interposed between cortices dealing respectively with somesthetic, auditory and visual functions. Lesions of the angular gyrus, which often lead to anomic aphasia and acquired reading and writing deficits probably destroy the bulk of area PG. (*See* **Parietal Cortex**)

The same brains with leftward asymmetry of area PG also exhibited predominance of the left planum temporale and of the left area 44. However, a more dorsal parietal area, area PG, which is less clearly related to language and more likely to be related to hemispatial attention, tended to be larger on the right side, and asymmetry in this area did not correlate with asymmetry of the planum temporale or area PG. This finding could suggest that asymmetries in one region are correlated with asymmetries in another region as long as the two regions are functionally linked. Animal studies showed similar correlations of directional asymmetry between adjacent, visually related cortices.

ASYMMETRIES IN OTHER SPECIES

The cerebral cortex emerges for the first time in its six-layered organization in mammals. Asymmetries in structures other than cortex are present, however, in birds, fish and amphibians. In the case of birds, the asymmetries are mainly in the functional domain and consist of differential effects of neural lesions on the ability of birds to sing, with the left predominating. Slight left anatomical superiority of one of the song-relevant nuclei has also been demonstrated. Fish and amphibians often show asymmetries in the habenular nuclei, the functional significance of which is not clear.

In nonhuman primates, the types if not the degree of brain asymmetry are similar to those found in the human brain, which raises questions about what lateralized functions, if any, these structural asymmetries might serve. For instance, sylvian fissure asymmetries have been found in other primate brains. If one is permitted to attach significance to the sylvian asymmetry in the human *vis à vis* linguistic capacity, what then is the meaning of asymmetry in the same structure in the baboon, the orangutan and the chimpanzee? Adding fuel to the fire, recent reports suggest that the pattern of human planum temporale asymmetry is present in chimpanzees. This again has potentially important implications for the role of cortical asymmetry and language function.

Asymmetries have also been noted in fossil skulls. The best-known of these is the asymmetry in the Neanderthal fossil from La Chapelle-aux-Saints, about 60 000 years old, which showed a sylvian fissure asymmetry similar to that seen in modern humans. There was a suggestion of a comparable asymmetry in the endocast of Peking man, which dates from 500 000–600 000 years ago. Asymmetries in perisylvian cortex also appear to exist in australopithecines, *Homo habilis* and *Homo erectus*, which date to up to 3.5 million years. Again, as is the case in nonhuman primates, asymmetries in the sylvian fissure are found in individuals whose language capacity is in question. Whether these asymmetries represent linguistic capacities in early humans or some preadaptive behavior is likely to remain unknown. It will help to find out what it is about asymmetry of the modern human brain, if anything, that can explain linguistic capacity.

MECHANISMS OF CEREBRAL ASYMMETRY

The phrenological approach to neuropsychology suggests that the basis for cognitive capacity is size. For example, our brains are uniquely able to carry out complex cognitive tasks because they are bigger than the brains of other less cognitively gifted primates. Similarly, the left hemisphere is able to support language because some of the perisylvian cortices involved in language are larger than corresponding areas on the right. An alternative to the phrenological concept suggests that highly specialized functions may depend more on specificity of organization. Thus, building a phonologically competent brain area may require increasing the specificity of the brain substrate by customized reduction rather than just enlargement of its components. In this sense, an asymmetric language area would result from the pruning of one side rather than from the enlargement of the other.

We have gained some insight as to the underlying substrates of brain asymmetry through the use of an animal model. Rats, as it turns out, also exhibit brain asymmetry. While individual rats may exhibit a large degree of asymmetry (magnitude), there is no leftward or rightward directional bias of the population as a whole. Exploiting this distinction, researchers have begun to ask questions concerning what distinguishes a symmetric from an asymmetric brain, rather than what factors bias a population to one side or the other. Interestingly, Collins has convincingly demonstrated in mice that while magnitude of lateralization, in this case paw preference, is under genetic control, direction is not. Using rats, in combination with examination of human asymmetries, we can begin to gain a hold on the phrenologic debate outlined above.

Symmetry of cortical areas is associated with changes in the volume of architectonic areas – symmetric brain regions are larger than asymmetric ones. Further, side differences in overall volume are the result of different numbers and densities of some neurons. Also, with increasing asymmetry, the proportion of the targeted area receiving callosal connections diminishes. In other words, the more asymmetrical areas have relatively fewer callosal connections than similar areas that are more symmetric, suggesting that the one key component of anatomic asymmetry may be greater intrahemispheric – as opposed to interhemispheric – connectivity. Indirect evidence suggests, furthermore, that asymmetry is determined during the earliest stages of brain development.

CONCLUSION

The histological characteristics of symmetric and asymmetric cortical areas do not support the notion that cerebral dominance represents simply a case of storing functional areas in one hemisphere or the other, but rather that there are storage factors as well as factors of network size and detailed connectivity. Thus, cerebral dominance reflects variation in functional properties of symmetric and asymmetric cortical areas, which provides the species with desirable and sometimes problematic individual variation.

For the first time, it is possible to assess the normally functioning human brain with sophisticated methods that can address issues about localization and lateralization of function. With

improving anatomical imaging it will be possible to extract individual information rather than sample averages, which will be useful for studies on individual variation of localization and lateralization. This knowledge can then be applied to studies of variability in response to brain injury and of developmental variation in cognitive style, response to brain injury, and learning disorders. (*See* **Neuroimaging**)

Further Reading

Denenberg VH (1981) Hemispheric laterality in animals and the effects of early experience. *Behavioral and Brain Sciences* **4**: 1–49.

Gannon PJ, Holloway RL, Broadfield DC and Braun AR (1998) Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog. *Science* **279**: 220–222.

Geschwind N and Galaburda AM (1987) *Cerebral Lateralization. Biological Mechanisms, Associations, and Pathology*. Cambridge, MA: MIT Press/Bradford Books.

Geschwind N and Levitsky W (1968) Human brain: left-right asymmetries in temporal speech region. *Science* **161**: 186–187.

Rosen GD (1996) Cellular, morphometric, ontogenetic and connectional substrates of anatomical asymmetry. *Neuroscience and Biobehavioral Reviews* **20**: 607–615.

Shapleske J, Rossell SL, Woodruff PW and David AS (1999) The planum temporale: a systematic, quantitative review of its structural, functional and clinical significance. *Brain Research Reviews* **29**: 26–49.

Brain Damage, Treatment and Recovery from

Intermediate article

Barbara A Wilson, MRC Cognition and Brain Sciences Unit, Cambridge, and the Oliver Zangwill Centre for Neuropsychological Rehabilitation, Ely, UK

CONTENTS

Natural recovery can and does occur in children and adults following brain injury. Rehabilitation can also result in improvements of functioning.

IS THERE RECOVERY FROM NONPROGRESSIVE DAMAGE?

The term 'recovery', when applied to nonprogressive brain damage, is interpreted in different ways by people according to their professional background, knowledge, expectations, and experience. Some will focus on biological repair of brain structures, others may regard survival rates as a measure of recovery, while others will be looking for signs of recovery of cognitive functions or motor skills. Some might argue that recovery can be thought of only in terms of the complete restoration of functions lost or impaired after brain injury. However, recovery in this sense is rarely achievable for the majority of brain-injured people. Others regard a good measure of recovery as resumption of normal life, even though there might be minor neurological and psychological deficits that do not disappear over time; this is certainly achievable for some people with nonprogressive brain damage. Yet another interpretation of recovery sees it as a diminution of impairments in behavioral or physiological functions over time; and experience suggests this is certainly true for the majority of people with brain injury. Observations by those involved in the treatment of brain injury suggest that recovery from nonprogressive brain damage typically involves partial recovery of function together with substitution of function, and can be operationally defined as complete or partial resolution of deficits incurred as a result of an insult to the brain.

People who sustain brain damage from a moderate or severe traumatic head injury (the most common cause of brain damage in people under the age of 25 years) usually undergo some – and often considerable – recovery. This is likely to be fairly rapid in the early weeks and months after the injury, followed by a slower recovery which can continue for many years. A typical pattern is an initial period of coma (the patient makes no verbal response, does not obey commands or open the eyes spontaneously or after stimulation), followed by posttraumatic amnesia (PTA), a period in which the patient is confused and disoriented, suffers from retrograde amnesia, and seems to lack the capacity to store and retrieve new information. The next stage is when the patient emerges from PTA, possibly with a number of motor, cognitive, and behavioral problems; these may resolve or partially resolve over time. Variations on this pattern may occur in other kinds of nonprogressive brain injury such as encephalitis (a viral infection of the brain), hypoxia (brain damage caused through shortage of oxygen), or cerebral vascular event.

A number of factors influence the extent of recovery, some of which we can do nothing about once the damage has occurred. These include the age of the person at the time of injury, the severity of damage, the location of damage, the status of undamaged areas of the brain, and the premorbid cognitive status of the brain. Other factors such as motivation, emotions, and the quality of rehabilitation programs available can be manipulated.

Age, often thought to be an important factor, has a less clear-cut influence than many believe. There appears to be a general belief that younger people recover better from injury to the brain than older people; this is known as the 'Kennard principle' after Kennard (1940), who showed that young primates with lesions in the motor and premotor cortex exhibited sparing and partial recovery

of motor function. Even Kennard, however, recognized that such sparing did not always occur and some problems became worse over time for certain younger individuals. Once severity, etiology and other demographic factors are taken into account, age is not always predictive of good outcome and younger people sustaining severe head injury often do worse than older people in terms of behavior problems and social deficits. Age, then, must be regarded as just one factor in the recovery process, to be considered alongside other factors such as whether the lesion is focal or diffuse, the severity of the injury, and the time since acquisition of the function under consideration: for example, someone who has only just learned to read is more likely to show reading deficits than someone who learned to read many years earlier.

One factor thought to interact with brain injury is level of 'cognitive reserve', suggesting that people with more education and higher intelligence may show less impairment than those with poor education and low intelligence. Most clinicians are aware of the fact that any injury of the same severity can produce profound damage in one patient and minimal damage in another. The concept of cognitive reserve has been used in studies of HIV (human immunodeficiency virus) infection and Alzheimer disease, and may also prove useful in understanding recovery from nonprogressive brain injury. A neurologist once said, 'It is not only the kind of head injury that matters but the kind of head.' Support for the idea of cognitive reserve comes from the field of language therapy, where severity of aphasia and site of lesion are not unfailing predictors of improvement.

Individuals with high intelligence and possibly superior education may process tasks in a more efficient way. Consequently, in cases of Alzheimer disease, task impairment may occur later in people with superior cognitive reserve than in those who are less intelligent and less well educated when both groups are matched for severity.

TREATMENT, REHABILITATION, AND OTHER FACTORS THAT INFLUENCE RECOVERY

Before considering intervention, one needs to have some understanding of the natural course of recovery from nonprogressive brain injury. This has been reported in a number of studies (Wilson, 1998). For example, a young girl developed meningococcal meningitis at the age of 14 months and, as a result, became prosopagnosic. When last seen at the age of 11 years and 7 months, there was no change in her prosopagnosia. Another study reported the case of a child who developed viral encephalitis at the age of 9 years, and was left with a visual object agnosia. When the girl was last seen at the age of 16 years, there had been a limited degree of recovery that the authors put down to her intact spatial abilities, which enabled her to compensate for her recognition difficulties. Scans using imaging techniques such as computerized tomography (CT) and magnetic resonance imaging (MRI) showed little change over time, leading the authors to believe that neural plasticity for visual processes is limited.

In contrast, a boy with Sturge–Weber syndrome showed dramatic recovery of language. The boy was mute until he underwent a left hemispherectomy at the age of 9 years. Language functioning then developed, and he was reported to have achieved clearly articulated speech with well-structured and appropriate language. At the age of 15 years he had the language skills appropriate for a child of 8–10 years. Possibly language skills are more likely to recover than other cognitive functions: children who develop hippocampal damage and consequent memory impairments early in life appear to show little, if any, improvement in memory.

These findings are confirmed in adults. Some people with language deficits go on to make a reasonable recovery, unlike those who sustain memory impairments. The patient HM, first reported by Scoville and Milner (1957), is perhaps the most famous amnesic patient in the world and has shown no recovery since his operation to relieve epilepsy in the 1950s. An amnesic musician who survived encephalitis, CW, also showed no recovery of memory functioning over a 10-year period. One study found that about two-thirds of memory-impaired people showed no change in memory functioning 5–10 years following discharge from a rehabilitation center, although most were compensating better.

As well as the nature of the cognitive deficit, the cause of the brain damage probably influences recovery. Thus people with head injuries often do better than people with other diagnoses: for example, a cohort of Second World War servicemen who sustained missile wounds to the brain showed considerable preservation of ability as a group despite some selective impairments related to the specific loci of the lesions. Most people with head injuries do not sustain penetrating wounds. Age and duration of coma are often found to be significant predictors of recovery from closed head injury (although recall the earlier caveats about

age). Different patterns of recovery are to be expected following severe head trauma, with some people doing well and others remaining in a minimally conscious state even after several years. In contrast to the results of head injury, people with bilateral surgical lesions, such as HM, and people with encephalitis such as CW, often show less recovery and little change over time. People with hypoxic brain damage also tend to show little improvement. On the other hand, recovery from a cerebrovascular event can be considerable for some people: Taub *et al*. (1993) showed it was possible to improve function in the hemiparetic upper limbs of stroke patients by preventing them using the corresponding unaffected limbs. After two weeks of training with the affected limbs (and immobilization of the unaffected ones), a significant improvement in motor functioning occurred.

According to some research, the recovery of stroke patients can be predicted by attentional skills. It is possible that restitution of functioning following stroke may be possible after small lesions while compensatory procedures are more likely to underlie recovery from larger lesions. We will return to this issue later. Meanwhile, note that in addition to natural recovery, treatment or rehabilitation of people with nonprogressive brain damage can also result in improvement of functioning.

Rehabilitation is a process whereby people disabled by injury or disease work together with professional staff, relatives and members of the wider community to achieve their optimum physical, psychological, social, and vocational wellbeing. Rehabilitation programs may work by teaching people to compensate for their problems, by helping them to learn more efficiently, or by achieving restoration (or partial restoration) of functioning through plasticity or regeneration. In clinical practice there appears to be a tendency to aim for restitution of function in the early days and weeks, with compensatory strategies coming into play when natural recovery stops or slows. Robertson (1999), however, believes that the choice of restitution or compensation as an aim depends on the extent (and possibly the location) of the lesion. He believes that restitution of function is achievable for some people with unilateral neglect, and states that 'there is growing evidence that restitution of basic function may be influenced by appropriate behavioural and cognitive inputs' (p. 691).

It is not clear whether such restitution is possible for other skills. The lack of change in memory function over time has already been mentioned. This may also be true for other cognitive functions, such as object recognition skills and spatial localization abilities. For such patients, compensation techniques taught during rehabilitation may offer the best hope of reducing everyday problems, and plenty of evidence exists to demonstrate that this is possible, as we shall see later.

EVIDENCE OF THE RECOVERY OF SENSORIMOTOR FUNCTIONS

Like other deficits, sensorimotor problems may show spontaneous recovery after nonprogressive brain damage. Zihl (2000) discusses spontaneous recovery of visual field disorders as well as recovery following training. He reports a finding that nearly a third of 41 patients with unilateral homonymous field loss showed recovery over an eight-month period. In several studies of people with cerebral blindness, only about 6% made a complete recovery but a further 67% showed some recovery of vision of varying degrees. Recovery tended to follow a particular pattern, with light perception recovering first, then perception of 'vague' contours with 'foggy vision', and finally perception of objects, faces, and surroundings. An interesting self-report of recovery of a visual field deficit following an occipital stroke is that of Bryan Kolb, a neuropsychologist (Kapur, 1997). Kolb kept a diary of his visual field deficit, his recovery, and his emotional reactions to this over a period of four years.

With regard to the recovery of motor functions, it has been found that the average patient receiving a program of focused stroke rehabilitation performed better than the majority of patients in control groups. The work of Taub *et al*. (1993) has already been mentioned, in which the unaffected upper limbs of hemiplegic stroke patients were immobilized for a period of two weeks while the patients were encouraged to use the affected limbs; considerable recovery was achieved. Furthermore, the recovery was maintained over the subsequent two-year period. Others have found that improvements in function could be obtained long after the natural recovery process was over and the hemiparetic limb had ceased to improve. Taub's team felt that temporary loss of use of a limb resulted in 'learned nonuse': the patients lost the habit of using the limb even though the underlying mechanisms enabling limb use had been repaired.

Some studies suggest that the incidence of visual, tactile, proprioceptive, and motor deficits is higher in patients whose stroke occurred in the right hemisphere rather than the left. This is thought to be due to a greater incidence of attentional deficits in patients with a right-sided injury, attention having an important role in influencing the function of

primary sensory and motor circuits, and consequently in recovering function within these circuits after a brain lesion. The functional recovery of 47 brain-damaged patients after a right-hemisphere injury was monitored over a two-year period, and it was found that sustained attention (assessed two months after the stroke using tasks from the Test of Everyday Attention) predicted motor recovery two years after the stroke. In addition, in comparison with a group of patients with left brain damage, those with right brain damage showed less functional ability and poorer attention skills.

Positron emission tomography (PET) has been used to look at the recovery of stroke patients. There appears to be considerable plasticity in the recovery of motor functions and this recovery is mediated by (a) recruitment of cortical areas in the undamaged hemisphere, and (b) extension of specialized areas adjacent to the site of the lesion. Another kind of imaging study looked at a 17-year-old girl who had sustained a unilateral brain injury at birth. Functional MRI revealed that the healthy hemisphere was controlling motor movements via direct ipsilateral corticospinal projections together with the contralateral cerebellum.

Relatively little has been written on the recovery of motor functioning following traumatic brain injury (TBI), although physical impairments are common resulting from both the brain damage and concomitant orthopedic trauma. It is believed that nearly half of people with TBI have physical problems, with about one quarter of these having orthopadic and/or muscular skeletal injuries as well as problems caused by the brain damage. A year after injury, however, most people with TBI are ambulant, and a study of Vietnam war veterans with penetrating head injuries and associated movement disorders found that at follow-up all were independent despite continuing mobility problems. It would appear that motor recovery after TBI follows, to a large extent, a developmental sequence.

EVIDENCE OF THE RECOVERY OF COGNITIVE FUNCTIONING

There appear to be four major approaches to cognitive rehabilitation: cognitive retraining through stimulation or exercises; strategies derived from cognitive–neuropsychological theoretical models; techniques combining methodologies and theories from a number of different fields, particularly behavioral psychology, neuropsychology, and cognitive psychology; and holistic approaches that address social and emotional problems alongside the cognitive ones. The nature of these approaches and an examination of their strengths and weaknesses are addressed in further detail by Wilson (1997).

There is increasing evidence that rehabilitation can improve cognitive functioning. In the field of memory disorders, the method of choice for reducing everyday problems is probably to teach compensatory approaches. Several publications show that people with memory impairments can function independently if they are able to use strategies to bypass their difficulties. Over a 10-year period, JC, a young man who became densely amnesic following a ruptured aneurysm on the left posterior cerebral artery, developed a sophisticated system of memory aids enabling him to live alone, hold down a job, and be totally independent despite remaining severely amnesic. The natural history of the development of his compensatory system shows that he began by writing on scraps of paper a few weeks after his stroke before gradually developing his highly successful system.

Another encouraging study describes the rehabilitation of a young woman who became amnesic following status epilepticus. She was taught to use a personal organizer (datebook), to refer to it regularly, and to monitor a number of daily events. Over a period of several weeks the young woman was able to learn the different sections of the datebook and use this to manage her life. After leaving the rehabilitation center, she worked in a voluntary capacity (still using her system) and eventually was taken on as a paid employee.

Even for people who are unable to return to work, memory compensations can assist independent living. One device that has been helpful to a number of people with memory impairment is a specifically designed alphanumeric pager worn on a belt, which sends out daily reminders. People with memory and/or planning problems all showed statistically significant improvements in achieving everyday target behaviors, such as taking medication, feeding the cat, and preparing meals, when using this pager, in comparison with a baseline period.

In addition to compensations in the form of external memory aids, techniques for improving learning in people with impaired memory are a major part of rehabilitation. It has been demonstrated that people with amnesia learn better if they are prevented from making mistakes during the learning process. Most of us can benefit from trial-and-error learning if we can remember our previous mistakes. People with amnesia, of course, cannot do this. Once a mistake has been made, it may be strengthened or reinforced or be

indistinguishable from the correct response. Consequently it is better to prevent the incorrect response being made in the first place. Several studies have applied the 'errorless learning' principle to teaching real-life tasks. Although a number of questions, both clinical and theoretical, remain to be answered about errorless learning, it appears to be more effective than trial-and-error learning when used to impart useful information to people with memory difficulties.

One group of people often considered difficult to treat are those with frontal lobe or executive deficits such as problems with planning, organization, and problem-solving. Even this group, however, have been helped by appropriate rehabilitation. Specific problem-solving training can help such people. Furthermore, the learning of patients with a dysexecutive syndrome and severe behavior problems can be enhanced through the provision of exaggerated feedback on their performance. A stroke patient with problems of attention, planning, distractibility, and perseverative behavior was able to overcome a number of her everyday problems through a combination of use of a pager and a checklist. More recently, a problem-solving strategy called Goal Management Training has been designed to help people with executive deficits and disorganized behavior manage their daily lives. This involves five stages, each corresponding to an important aspect of goal-directed behavior (e.g. 'define the main task', 'list the steps').

Several studies have demonstrated improvement of language functioning in both children and adults using training exercises presented through computer games. In one study children with language impairments had 8–16 hours of training over 20 days. This resulted in marked improvements of the children's ability to recognize sequences of nonspeech and speech stimuli. In another study, also using computer-based exercises, language comprehension improved with acoustically modified speech. Computer training was also carried out with 55 adults each of whom had sustained a left-hemisphere stroke resulting in aphasia. A reading treatment (visual matching and reading comprehension tasks) led to better results than computer stimulation (involving nonverbal tasks and games) or absence of treatment.

Numerous studies discuss the treatment of unilateral visual neglect, and a number of these discuss the rehabilitation of patients with visual disorders after brain injury. As with the studies quoted above, substantial evidence accrues to suggest that improvement of cognitive functioning can be achieved through rehabilitation.

MECHANISMS OF RECOVERY

The process of recovery from brain injury is not well understood and probably involves different biological processes. Changes seen in the first few minutes (for example after a mild head injury) presumably reflect the resolution of temporary dysfunction without accompanying structural damage. Recovery after several days is more likely to be due to resolution of temporary structural abnormalities such as edema or vascular disruption, or to the depression of metabolic enzyme activity.

Recovery after months or years is even less well understood. There are several ways this might be achieved, including regeneration, diaschisis, and plasticity. Regeneration in the central nervous system can occur and is more likely to do so early in life, although it does occur in adults. Thus the view held for many years that cerebral plasticity is severely restricted in the adult human brain is no longer credible. Although the limits of neurorehabilitation have been significantly influenced by the basic premise that brain cells can never regenerate, this is now known to be false and our horizons may well be extended. What is less clear is the extent to which regeneration can lead to functional gains in coping with real-life problems.

Diaschisis assumes that damage to a specific area of the brain can result in neural shock or disruption elsewhere in the brain. The secondary neural shock can be adjacent to the site of the primary insult or much further away. In either case, the shock follows a particular neural route. Similar to this, but not identical, is the idea of inhibition. In inhibition, however, the shock is more diffuse and affects the brain as a whole. Robertson and Murre (1999) interpret diaschisis as 'a weakening of synaptic connections between the damaged and undamaged sites, contingent on the reduced level of activity in the lesioned area' (p. 547). Because cells in the two areas are no longer firing together, synaptic connectivity between them is weakened and this results in the depression of functioning in the undamaged but partly disconnected remote site.

Plasticity implies anatomical reorganization based on the idea that undamaged areas of the brain can take on the functions subserved by a damaged area. Until recently this idea was discredited as an explanation for recovery in adults, although views are now changing. Robertson and Murre (1999), in a thought-provoking paper, suggest that plastic reorganization may occur initially because of a rapidly occurring alteration in synaptic activity taking place over seconds or minutes, followed by structural changes taking place over

days and weeks. The authors focus in particular on people who are likely to show recovery provided they have assistance and rehabilitation. According to Robertson and Murre, some individuals show autonomous recovery, others show little recovery even over a period of years, while others show reasonably good recovery provided they receive rehabilitation. They refer to this as a triage of spontaneous recovery, assisted recovery, and no recovery. Robertson and Murre believe that the strategy of choice for people in the 'no recovery' group is to teach compensatory approaches, and that the 'spontaneous recovery' group do not need rehabilitation as they will get better anyway. Consequently, they focus on the 'assisted recovery' group to address issues about brain plasticity. They also believe that the severity of the lesion relates to this triage, with mild lesions resulting in spontaneous recovery, moderate lesions benefiting from assisted recovery, and severe lesions necessitating the compensatory approach.

Although heuristically useful, this idea may be too simplistic. For example, people with mild lesions in the frontal lobes could be more disadvantaged in terms of recovery than people with severe lesions in the left anterior temporal lobe. The former group might have attention, planning, and organization problems precluding them from gaining the maximum benefit from the rehabilitation on offer, whereas the latter group with language problems could show considerable plasticity by transferring some of the language functions to the right hemisphere. Nevertheless, Robertson and Murre make some interesting arguments and present a model of self-repair in neural networks based on a connectionist model of recovery of function.

One technique that has potential to help us understand the nature of recovery is the employment of imaging techniques. Grady and Kapur (1999) suggest that imaging studies could enable us to measure specific changes occurring in the brain during recovery and allow us to determine whether recovery is the result of reorganization of functional interactions within an existing framework, recruitment of new areas into the network, or plasticity in regions surrounding the damaged area.

A few studies using imaging techniques to look at recovery following brain injury have been reported. Perhaps the first paper in this area reported on changes in regional cerebral blood flow (rCBF) following cognitive rehabilitation for people who had sustained encephalopathy after exposure to toxins. Later in 1997, PET was used to identify the neural correlates of stimulation procedures employed in language rehabilitation in people with dysphasia. Single photon emission computed tomography (SPECT) has been used to evaluate rCBF during recovery from brain injury. Specific changes in rCBF appeared to be related to (a) the location of the injury, and (b) strategies used in cognitive rehabilitation. Continued improvements in three patients were documented in rCBF, functional abilities, and cognitive skills up to 45 months following the injury.

The following year functional imaging was employed to monitor the effects of rehabilitation for unilateral neglect. The brain regions most active after recovery were almost identical to the areas active in control subjects engaged in the same tasks. This would appear to support the view that some rehabilitation methods repair the damaged network and do not simply benefit patients through compensation or behavioral change.

The use of imaging studies is certain to grow in research into recovery from brain injury. To what extent findings from such imaging studies will help us plan or improve rehabilitation remains an open question.

CONCLUSION

The human brain is capable of more plasticity than previously thought, since natural recovery can and does occur over time not only in children but also in adults. In addition to natural recovery there is increasing evidence that rehabilitation can result in improvements of both sensorimotor and cognitive functioning. Despite this, there is considerable variability in the nature and extent of such recovery and there are clearly limits to the amount of recovery and improvement possible in people with nonprogressive brain injury. We are now faced with a number of questions. What factors limit the recovery process? Can neurogenesis, for example, lead to cells that can survive in sufficient numbers and integrate in ways to improve everyday functioning? When should rehabilitation programs begin, during the spontaneous recovery process or later? How should we determine whether to aim for plasticity, regeneration, or compensation? Should we be influenced solely by the severity of the lesion, the cognitive function affected, the time after injury, or by all of these?

References

Grady CL and Kapur S (1999) The use of imaging in neurorehabilitative research. In: Stuss DT, Winocur G and Robertson I (eds) *Cognitive Neurorehabilitation:*

A Comprehensive Approach, pp. 47–58. New York, NY: Cambridge University Press.
Kapur N (1997) *Injured Brains of Medical Minds: Views from Within*. Oxford, UK: Oxford University Press.
Kennard MA (1940) Relation of age to motor impairment in man and in subhuman primates. *Archives of Neurology and Psychiatry (Chicago)* **44**: 377–397.
Robertson IH (1999) Theory-driven neuropsychological rehabilitation: the role of attention and competition in recovery of function after brain damage. In: Gopher D and Koriat A (eds) *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*, pp. 677–696. Cambridge, MA: MIT Press.
Robertson IH and Murre JMJ (1999) Rehabilitation after brain damage: brain plasticity and principles of guided recovery. *Psychological Bulletin* **125**: 544–575.
Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21.
Taub E, Miller NE, Novack TA *et al*. (1993) Technique to improve chronic motor deficit after stroke. *Archives of Physical Medicine and Rehabilitation* **74**: 347–354.
Wilson BA (1997) Cognitive rehabilitation: how it is and how it might be. *Journal of the International Neuropsychological Society* **3**: 487–496.
Wilson BA (1998) Recovery of cognitive functions following non-progressive brain injury. *Current Opinion in Neurobiology* **8**: 281–287.
Zihl J (2000) *Rehabilitation of Visual Disorders After Brain Injury*. Hove, UK: Psychology Press.

Further Reading

Broman M, Rose AL, Hotson G and Casey CM (1997) Severe anterograde amnesia with onset in childhood as a result of anoxic encephalopathy. *Brain* **120**: 417–433.
Fawcett JW, Rosser AE and Dunnett SB (2001) *Brain Damage, Brain Repair*. New York, NY: Oxford University Press.
Katz RC and Wetz F (1997) The efficacy of computer-improved reading treatment for chronic aphasic adults. *Journal of Speech and Language Hearing Research* **40**: 493–507.
Newcombe F (1996) Very late outcome after focal wartime brain wounds. *Journal of Clinical and Experimental Neuropsychology* **18**: 1–23.
Schiavetto A, Decarie J-C, Flessas J, Geoffroy G and Lassonde M (1997) Childhood visual agnosia: a seven year follow-up. *Neurocase* **3**: 1–17.
Tallal P, Miller SL, Bedi G *et al*. (1996) Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science* **271**: 81–84.
Vargha-Khadem F, Carr LJ, Isaacs E *et al*. (1997) Onset of speech after left hemispherectomy in a nine-year-old boy. *Brain* **120**: 159–182.
Vargha-Khadem F, Gadian DG, Watkins KE *et al*. (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **227**: 376–380.
Wilson BA (1999) *Case studies in Neuropsychological Rehabilitation*. New York, NY: Oxford University Press.
Young AW and Ellis HD (1989) Childhood prosopagnosia. *Brain and Cognition* **9**: 16–47.

Brain Self-stimulation

Introductory article

Peter M Milner, McGill University, Montreal, Quebec, Canada

CONTENTS

Most vertebrates and some invertebrates learn to make responses that deliver electrical stimulation to parts of the nervous system. This provides a tool for studying the anatomy and neuropharmacology of reward mechanisms. Knowledge so gained has proved useful for understanding motivation and addiction.

INTRODUCTION

The observation that rats seek electrical stimulation of certain brain areas was made in 1953 by James Olds and Peter Milner, working in the laboratory of Donald Hebb at McGill University. The publication in 1949 of Hebb's influential book, *The Organization of Behavior*, brought a group of scientists to Hebb's laboratory, keen to explore his theories. James Olds was a postdoctoral fellow with a degree in social psychology from Harvard; I was a graduate student with a degree in communication engineering from Leeds University. A year before the discovery Seth Sharpless, a philosophy student from the University of Chicago, and I, seeking to improve on Hebb's account of motivation, performed a pilot study to test our notion that reward involved the recently discovered reticular activating system (RAS). Our experiments indicated, however, that RAS stimulation was punishing. As pain pathways pass through the RAS this outcome was not deemed sufficiently exciting to justify a change of thesis research.

When Olds arrived on the scene it fell to me to show him how to implant electrodes in the rat brain; Olds intended to implant his first practice electrode in the RAS. When the rat recovered from the operation he placed it on a large table and tested the effect of short bursts of stimulation. In the initial experiments the stimulation was a few volts of 60 Hz alternating current from a transformer connected to a power outlet. Most later experiments used brief (>1 s) trains of short (0.1–1.0 ms) rectangular current pulses of negative polarity. It was soon obvious that the rat found any place where it had been stimulated attractive. If removed from the place it would quickly return, sniffing and searching. Later tests, in which the rat learned to press a lever to deliver electric current to its brain, gave further reason to believe that the stimulation served as a reward.

This was very exciting. I immediately operated on another rat, using the same stereotaxic settings but, as before, without success. Suspecting that the rewarding electrode of the successful rat was not at the intended site we X-rayed its head, confirming our suspicion. Probably the dental cement holding the electrode to the skull had not set before the rat was released from the stereotaxic instrument. Unfortunately, when the rat finally came to autopsy, its brain was too badly damaged for precise localization of the electrode. Based largely on the x-ray, we estimated that the electrode was in the vicinity of the septal area, but after the relation of self-stimulation behavior to electrode site had been established by subsequent mapping studies, it became evident that the behavior exhibited by the first rat had been more similar to that of rats with hypothalamic electrodes than to rats with septal area electrodes. Before he embarked on a systematic search for reward areas in the brain, Olds designed an electrode that was held to the skull by screws rather than by dental cement.

MAJOR CHARACTERISTICS OF SELF-STIMULATION

The characteristics of self-stimulation depend greatly on the site of the electrode, owing in part to concomitant stimulation of paths not related to reward. Rewarding sites have been reported in many areas of the brains of most, if not all, vertebrates that have been tested, and even in some invertebrates. In mammals, stimulation is very rewarding in the vicinity of the medial forebrain bundle (MFB) where it passes through the lateral hypothalamus (LH), a nucleus at the base of the

brain involved in appetitive behaviors (Figure 1). Most rats with electrodes in that region quickly learn to press a lever for stimulation. Typically, after a few successful responses, they become excited and assail the lever vigorously, pausing momentarily at intervals to run in a tight circle. Some gnaw on the lever or other parts of the apparatus. If given the opportunity they may continue working at a high rate for as long as 24 h, without sleeping or eating.

Stimulation at some sites in the ventral midbrain (near the MFB) and the ventral pons elicits similar enthusiastic responding, but at most other reward sites, including parts of the frontal cortex, septal nuclei, basal ganglia, thalamus, and amygdala (too lateral to be shown in Figure 1), the results are less dramatic. At these sites, learning to bar-press may take several sessions, and responding is usually intermittent with long pauses. Sometimes the pressing is slow and deliberate and in some cases the rat recoils from the bar after pressing it, as if frightened or hurt, but nevertheless returns for more.

Stimulation at almost all reward sites appears to become aversive if the train lasts for more than a second or two, depending on its intensity. In most experiments the duration of each train of stimulation is determined by a timer, but if it is not, rats almost always release the lever within half a second. If stimulation is turned on automatically, rats will learn to press a lever to turn it off, so the effect is not due to stimulation-induced hyperactivity.

Rewarding brain stimulation can replace food or water reward in most animal learning experiments, but there are some differences. Rats with LH electrodes, for example, quickly learn to press a lever or run through a maze to receive stimulation, but they appear to forget equally quickly. They also show severe 'overnight decrement', initially ignoring, possibly even avoiding, the lever or goal-box for which they showed great enthusiasm the previous day. 'Priming' by a few bursts of stimulation delivered by the experimenter restores the motivation.

MECHANISMS OF SELF-STIMULATION

Quite soon after mapping the reward system, Olds discovered that self-stimulation was profoundly depressed by antipsychotic drugs. Subsequent discovery of dopamine fibers in the MFB led to the theory that stimulation fires these fibers, releasing dopamine in forebrain structures to reinforce synaptic connections. This theory was too simple to be the whole story. Dopamine fibers, which are all unmyelinated, are not easily fired by the short pulses of current commonly used for self-stimulation.

It was next postulated that the current fires myelinated fibers, which then synapse with dopamine neurons in the ventral tegmental area (VTA), thus indirectly increasing the delivery of dopamine to forebrain structures. However, such an increase has proved difficult to measure. Dopaminergic activity seems to increase mainly during unexpected

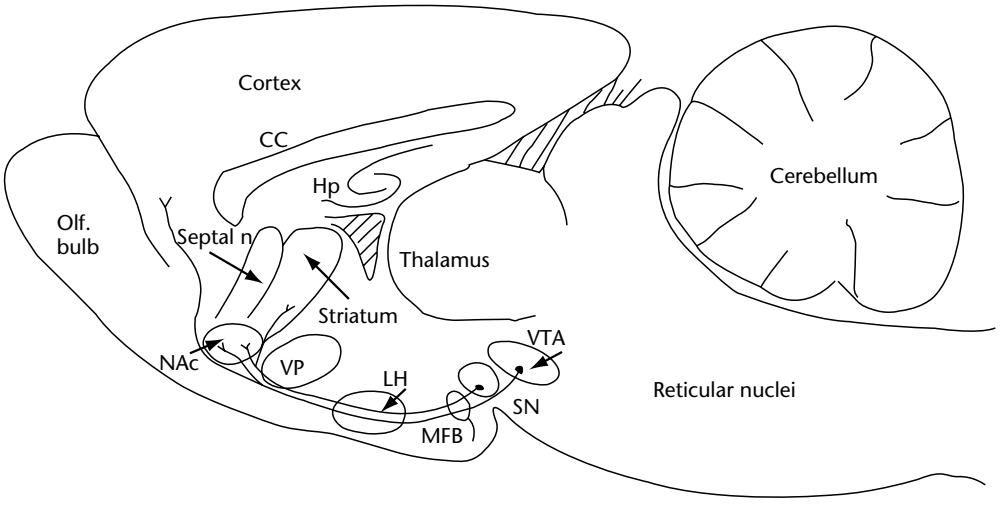


Figure 1. Longitudinal section of rat brain showing the locations of some brain-stimulation reward sites. CC, corpus callosum; Hp, hippocampus; LH, lateral hypothalamus; MFB, medial forebrain bundle; NAc, nucleus accumbens; Olf., olfactory; SN, substantia nigra; VP, ventral pallidum; VTA, ventral tegmental area.

events, which is of course when most learning takes place.

It certainly appears that dopamine is essential for brain stimulation reward from MFB sites; blocking its action in the accumbens nucleus by injecting a small quantity of a dopamine-blocking agent quickly diminishes self-stimulation. The accumbens is a nucleus that until the 1970s was considered part of the septal area, but its cell type and connections indicate that it is functionally part of the ventral striatum. It receives dopaminergic input via the MFB.

Using the fact that there is a rapid increase in the expression of the protein Fos in neurons when they are active, it has been found that self-stimulation of the LH or VTA activates neurons in many subcortical nuclei. Lesions indicate, however, that only a few of them play a vital role in self-stimulation. Some neurons are activated directly by the electrical current, others via synaptic connections from directly stimulated neurons. Some of the neurons that express Fos may be those involved in bar-pressing. An activated area that seems important for reward is the ventral pallidum, which lies just anterior to the LH. Some of its connections are to midbrain nuclei, including those that deliver dopamine to forebrain areas.

Signals from hunger, thirst, and other motivational states are delivered to the ventral part of the striatum, where they elicit appropriate response plans. It is there that incipient response activity can be either suppressed, or amplified to become an overt movement. The amplification circuit, which is modulated by dopamine, normally receives input from innately rewarding stimuli (such as water, or the smell of food) via the hypothalamus or the amygdala. The response suppressing system, which may also be modulated by dopamine, receives innately aversive input (pain or predator smell) mainly via the thalamus or the amygdala.

Innately neutral sensory input reaches the striatum via the cerebral cortex. In the striatum this input may acquire associations (by classical conditioning) with either the response amplifier or the response inhibitor, depending on which type of innate reinforcer influences the ventral striatum at the time. Thus an animal may at first be attracted only by the smell of food in a dish, but in time it is also attracted by the sight of the dish.

Rewarding brain stimulation is assumed to stimulate pathways that would normally be activated by innate rewards such as food, thereby triggering the response release mechanism in the striatum. As a consequence, sensory input and plans for motor activity that reach the striatum during rewarding stimulation acquire the ability to amplify and release a planned response.

It is generally assumed that brain self-stimulation has an effect similar to a conventional reward except that the modulatory effect of bodily need is bypassed. If an animal is not hungry, for example, the smell and taste of food do not amplify response plans, but self-stimulation, by acting directly on the reward system, is not greatly influenced by the intensity of bodily need. Self-stimulation has proved a useful tool for studying the basic mechanism of reward, providing direct access to the anatomical and physiological features of the system.

SELF-STIMULATION AND THE STUDY OF ADDICTION

Harmful addictions are the penalty we pay for an effective motivation system. Understanding the reward system in the brain is highly relevant to the way we approach the problems of drug abuse and other obsessive indulgences. At one time the most widely held view of drug dependency was that aversive after-effects of abused substances are alleviated by taking more of the substance, leading to a vicious spiral. Neither the sites of action nor the pharmacological processes involved were known. After the discovery of brain-stimulation reward, parallels were drawn between the behavior of the self-stimulating rat and that of addicts.

All creatures are in a sense addicted to certain essentials like air, water, heat, and food, all of which can be harmful in excess. Fortunately, protective regulators evolved along with the appetites. Automatic regulation of stimulants that act directly on the reward system is more difficult. Electrical stimulation has a similar direct effect on reward mechanisms, but with the advantage that it is easier to identify the brain structures involved, so that their connections and pharmacological properties may be studied.

When the importance of dopamine for self-stimulation was discovered, dopamine soon became a prime focus of investigations into drug abuse. It was pointed out that enhancers of dopaminergic action, such as amphetamine and cocaine, are addictive. Many dopaminergic neurons are sensitive to other abused substances such as opiates and nicotine.

The knowledge concerning the anatomy of reward systems was helpful in designing experiments to explore addictive processes; the techniques used for electrical stimulation were also

adapted for use in these experiments. Rats learn to press a lever to deliver many of the substances that are addictive to humans directly to a point in the brain via implanted cannulae. This makes it possible to determine or confirm the sites of action of a drug.

The dopamine theory of reward changed the direction of addiction research, but like most new theories it proved to be an oversimplification. Opiate receptors, for example, are widespread in the brain and almost certainly affect learning and other systems directly, as well as via dopamine. The mechanism by which dopamine itself influences reward is still not fully understood. Today the incentive for a great deal of the research on self-stimulation is the desire to resolve the problem of drug dependency. This quest, having been reoriented by the discovery of electrical self-stimulation, is now the tail that wags the dog.

SELF-STIMULATION AND THE STUDY OF COGNITION

Although philosophers used to draw a sharp distinction between knowledge and the will, behavior cannot be so clearly compartmentalized. Attention provides a strong link between motivation and perception, for example. Research on brain stimulation reward contributes to cognition by providing a better understanding of motivational mechanisms.

It is probably true to say that most perceptual research and theory during the twentieth century ignored attention. Sensory paths were treated as one-way streets from receptors to some internal representation of a stimulus in memory, or to the response mechanism. The basic function of sensory systems to extend the capacity of the motor system went unnoticed, or was forgotten.

Decisions to make a response are made on the basis of the kind of information we call motivational. Once the decision to act has been established, motor adjustments required to implement it are guided by the sort of stimuli that are most usually studied in perceptual experiments: the shape, color, distance and so on of objects. Of the very large number of stimuli usually present in the environment, the perceptual systems select for processing only those that pertain to the task in hand. In order to study the mechanism by which this is achieved it is necessary to understand how motivating stimuli influence response plans, and how these plans then interact with the reception and transmission of sensory signals.

Incentive learning is another example of the interplay between perceptual input and motivation. Important motivational paths meet sensory input in the striatum, suggesting that it is a place where stimuli that are not innately motivating may acquire value. Patterns of sensory input acquire the ability to release behavior or inhibit it. By the reverse process, motivational signals become associated with sensory patterns that are correlated with innately reinforcing events. Hunger becomes associated with the image of a food dish, for example. It is this image that determines the pattern of stimulation to which the sensory system is sensitized. In other words, reverse pathways in the sensory systems carry attentional facilitation. They can lower the perceptual threshold for goal-related stimuli and the motor threshold for intention-related actions. Thresholds for distracting stimuli are raised.

Further Reading

Bozarth MA (1987) Ventral tegmental reward system. In: Engel J and Oreland L (eds) *Brain Reward Systems and Abuse*, pp. 1–17. New York, NY: Raven Press.

Gallistel CR, Shizgal P and Yeomans JS (1981) A portrait of the substrate for self-stimulation. *Psychological Review* **88**: 228–273.

Milner PM (1989) The discovery of self-stimulation and other stories. *Neuroscience and Biobehavioral Review* **13**: 61–67.

Milner PM (1991) Brain-stimulation reward: a review. *Canadian Journal of Psychology* **45**: 1–36.

Milner PM (1999) *The Autonomous Brain*. Mahwah, NJ: Lawrence Erlbaum.

Olds J (1973) Commentary: the discovery of reward systems in the brain. In: Valenstein ES (ed.) *Brain Stimulation and Motivation*, pp. 81–99. Glenview, IL: Scott, Foresman.

Olds J and Milner PM (1954) Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–427.

Rolls ET (1975) *The Brain and Reward*. Oxford, UK: Pergamon Press.

Shizgal P and Murray B (1989) Neuronal basis of intracranial self-stimulation. In: Liebman JM and Cooper SJ (eds) *The Neuropharmacological Basis of Reward*, pp. 106–163. Oxford, UK: Clarendon Press.

Vaccarino FJ, Schiff BB and Glickman SE (1989) Biological view of reinforcement. In: Klein SB and Mowrer RR (eds) *Contemporary Learning Theories: Instrumental Conditioning Theory and the Impact of Biological Constraints on Learning*, pp. 111–142. Hillsdale, NJ: Lawrence Erlbaum.

White NM and Milner PM (1992) The psychobiology of reinforcers. *Annual Review of Psychology* **43**: 443–471.

Wise RA, Bauco P, Carlezon WA and Trojniar W (1992) Self-stimulation and drug reward mechanisms. *Annals of the New York Academy of Sciences* **654**: 192–197.

Yeomans J (1988) Mechanisms of brain-stimulation reward. *Progress in Psychobiology and Physiological Psychology* **13**: 227–266.

Cerebellum

Intermediate article

Frank A Middleton, State University of New York Upstate Medical University, Syracuse, New York, USA
Stephen I Helms Tillery, Arizona State University, Tempe, Arizona, USA

CONTENTS

The cerebellum ('little brain') contains more neurons than the rest of the entire nervous system, and its complex organization has been the subject of much research. Although it is not essential for cognitive thought processes, the cerebellum is clearly involved in the performance of some cognitive tasks and in the improvement of motor control.

INTRODUCTION

The human cerebellum is a fist-sized structure that sits in the back of the skull, located just behind the midbrain and below the cerebral hemispheres (Figure 1). It has been the subject of considerable scientific attention for much of the past two hundred years from researchers in such diverse fields as phrenology, psychiatry, evolutionary and comparative biology, neurology, genetics, and developmental biology. Cerebellum is translated literally as 'little brain' and indeed, despite its relatively small size, this single structure contains more neurons than the rest of the nervous system combined. While this feature alone might seem daunting at first glance, there are several aspects of cerebellar organization that have greatly simplified its study and led to its widespread use as a model system for addressing a large number of structural and functional issues in brain research.

ANATOMY

The cerebellum is often depicted as a composition of three basic elements: the cerebellar cortex (which consists of a granular layer, Purkinje layer and molecular layer), the deep cerebellar nuclei, and the large white-matter tracts that run between these structures and connect the cerebellum with other brain structures. In cerebellar cortex, five main types of neurons have been identified – the Purkinje, Golgi, granule, basket and stellate cells (Table 1, Figure 2) – as well as three minor classes of neurons – Lugaro, unipolar brush and pale cells (not shown). In the deep cerebellar nuclei, there are only two main types of neurons – projection neurons and local interneurons – although several minor classes have also been identified. In both the cortex and deep nuclei, the connections and neurochemical markers of these neurons have been well characterized (Table 1, Figure 2). In addition to these neuronal classifications, there are also three different types of nonneuronal cells in the cerebellum: radial (or Bergmann) glial cells in the Purkinje and molecular layers of cerebellar cortex; bushy astroglia in the granular layer; and oligodendrocytes in the white-matter layer between the cerebellar cortex and deep nuclei. Perhaps the most salient feature of cerebellar organization is the orderly arrangement of connections between all of these different cell types in the cerebellum (Figure 2). This circuitry was worked out by careful and time-consuming anatomical and electrophysiological mapping studies (see Ito, 1984). Because this basic circuitry exists throughout the entire cerebellum, researchers have been able to understand how the whole cerebellum works by studying only small parts of it.

Superimposed on the repetitive microcircuitry of the cerebellum are larger structural and functional subdivisions which have proved extremely useful for understanding cerebellar organization. At the gross anatomical level, the cerebellum is divided into multiple lobes or lobules, often smooth in appearance in smaller species but composed of numerous folds and fissures in more complex animals (Table 2, Figure 1). Each of these subdivisions contains essentially the same microcircuitry shown in Figure 2. The somewhat complicated nomenclature listed in Table 2 (and defined and described in detail by Larsell and Jansen, 1972) was actually developed as a simplifying scheme of the

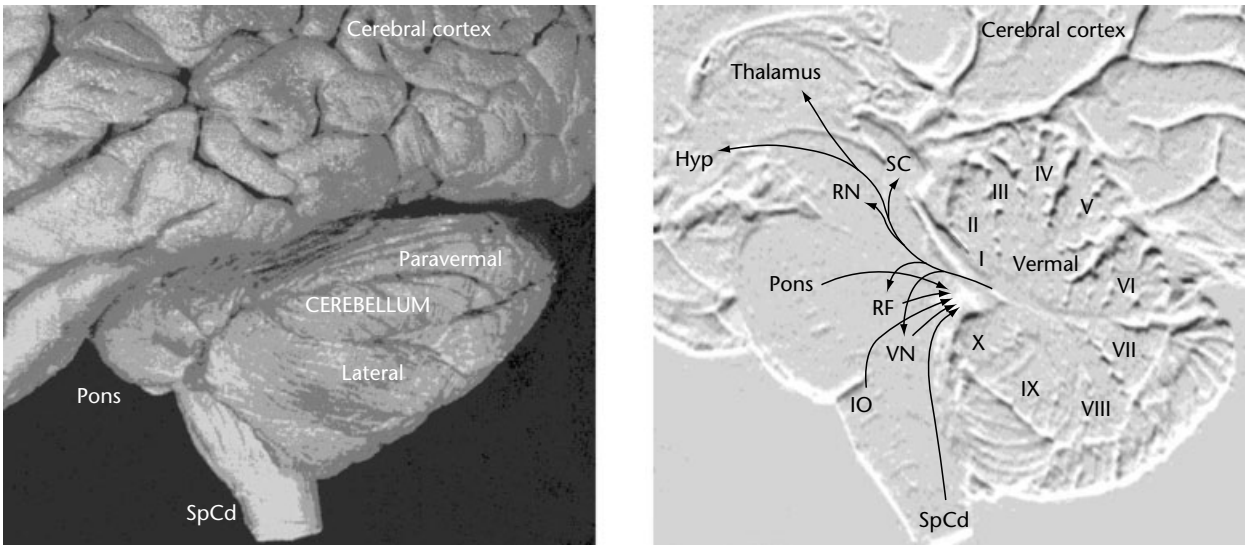


Figure 1. The cerebellum. (a) External view. (b) Midline view showing major inputs and outputs. Roman numerals refer to Larsell's nomenclature of the vermal lobules as listed in Table 2. Hyp, hypothalamus; IO, inferior olive; RF, reticular formation; RN, red nucleus; SC, superior colliculus; SpCd, spinal cord; VN, vestibular nuclei.

Table 1. Main cell types in the cerebellum

Location	*Name*	*Type*	*Markers*	*Afferents*	*Efferents*
Cortex, Purkinje layer	Purkinje	Inhibitory neuron	GABA, GAD, GABA-T, zebrin, motilin	Inferior olive, via climbing fibers Granule cells, via parallel fibers Other cortical neurons (basket, stellate, Lugaro)	Deep nuclei
Cortex, Purkinje layer	Basket	Inhibitory neuron	GABA, GAD	Granule cells, via parallel fibers	Purkinje cell somas, initial axon segments
Cortex, Purkinje layer	Stellate	Inhibitory neuron	GABA?	Granule cells, via parallel fibers	Purkinje cell dendrites
Cortex, granular layer	Golgi	Inhibitory neuron	GABA	Granule cells, via parallel fibers	Granular layer cells
Cortex, granular layer	Granule	Excitatory neuron	GABA, glutamate	Pontine and brainstem nuclei via mossy and mossy-like fibers	Dendrites of other cerebellar cortical cells
Deep nuclei	Projection	Excitatory neuron	Glutamate	Purkinje cells, collaterals of other afferent systems	Thalamus, magnocellular red nucleus, brainstem, reticular nuclei, superior colliculus, cerebellar cortex
Deep nuclei	Projection	Inhibitory neuron	GABA	Purkinje cells, other afferent systems	Inferior olive
Deep nuclei	Interneuron	Inhibitory neuron	GABA, glutamate	Other deep nuclei cells	Other deep nuclei cells

GABA, gamma-aminobutyric acid; GABA-T, gamma-aminobutyric acid transaminase; GAD, glutamate acid decarboxylase.

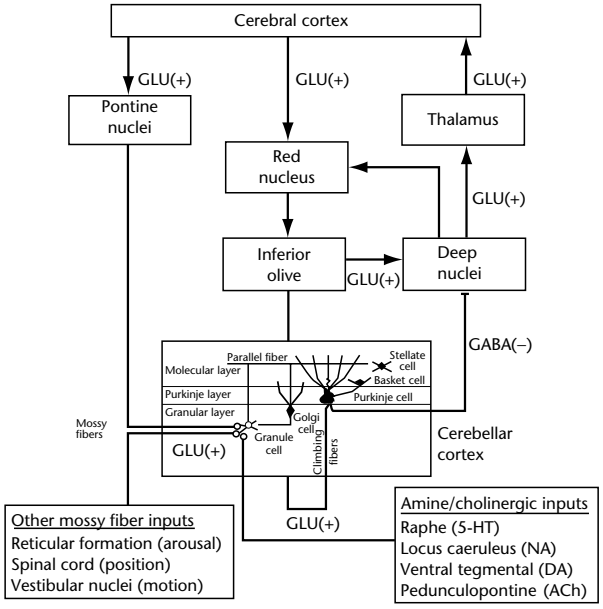


Figure 2. Connections and microcircuitry of cerebellar structures. 5-HT, 5-hydroxytryptamine (serotonin); ACh, acetylcholine; DA, dopamine; GABA, gamma-aminobutyric acid; GLU, glutamate; NA, noradrenaline (norepinephrine). Plus and minus signs refer to excitatory and inhibitory projections, respectively.

transverse structural organization of the cerebellum that holds across all species of animals. However, the functional organization of the cerebellum as it is currently understood is simpler, and consists of different transverse 'zones' (see Jansen, 1972; Voogd, 1975). In the sagittal plane, the cerebellum can be divided into midline (vermal), intermediate (paravermal) and lateral (hemispheric) zones (Table 3, Figure 1). Each of these zones receives a certain type of functional input at the level of the cerebellar cortex and sends its output largely through a single deep nucleus. For example, most of the vermal zone is concerned with maintaining posture and balance and coordinating reflexes (including reflexive eye movements and some autonomic functions) and directs its output through the medial deep cerebellar nucleus. In a similar manner, most of the intermediate cerebellum is concerned with sensory and motor processing that facilitates and coordinates voluntary movement, and directs its output through the intermediate deep nuclei. Lastly, most of the lateral cerebellum (also called 'neocerebellum' by some investigators) appears to be concerned with the planning, generation and control of complex behavior and directs its output through the lateral deep nucleus (Table 3).

The cerebellum is known to be connected with a large variety of different brain regions, including the spinal cord, brainstem, vestibular and reticular nuclei, as well as several major nuclei that are known to be the site of synthesis of specific neurotransmitters, such as the raphe, ventral tegmental area, locus caeruleus, and pedunculopontine nuclei (Figures 1 and 2). The sole output structures of the cerebellum are the four deep cerebellar nuclei and the vestibular nuclei (Table 3). These nuclei send signals back to many of the same brain regions that provide input to the cerebellum. Not surprisingly, the area of the cerebellum that is reciprocally connected with these extracerebellar regions has often been shown to be involved in functions that are related to that structure. For example, vestibular signals from the inner ear regarding movement of the head are sent primarily to the vermis. The activity of neurons in certain parts of the vermis has been shown to be related to both these sensory signals and to the commands for movement that are sent from the cerebellum to the vestibuloocular system to help maintain orientation of the head and eyes while the body is in motion. A similar type of arrangement has been shown to exist for the connections between the cerebral cortex and cerebellum. Most of the cerebral cortex sends fibers to the pontine nuclei that are a source of input to the cerebellar cortex. In turn, the output nuclei of the cerebellum project back, via the thalamus, to many of the cortical areas from which they receive input. Defining the full scope of areas that participate in these types of reciprocal cerebellar loops has important implications for understanding the functions with which the cerebellum is most likely to be involved.

The precise nature of the input projection to the cerebellum depends on the structure from which it originated. Signals from the pontine nuclei, vestibular nuclei, trigeminal nuclei and spinal cord are conveyed to the cerebellum by mossy fiber projections to granule cells. Signals from the red nucleus are conveyed by a projection to the inferior olive, which sends climbing fibers to Purkinje cells (Figure 2). Within the cerebellum, a great deal of processing of these signals takes place using cerebellar microcircuitry. For many functions, this processing has been shown to improve the accuracy of the signals, so that the behavior or function that modulates the activity in the circuit is seen to improve. For this reason, the cerebellum has often been thought of as a type of external teaching device that improves the performance of any brain region participating in cerebellar feedback loops. In addition, the cerebellum has also been depicted as a

Table 2. Main structural divisions of the cerebellum

Lobule	*Vermal name*	*Human hemisphere name*	*Mammalian hemisphere name*
I	Lingula	Vinculum lingulae	Anterior lobe
II	Central	Ala lobuli centralis	
III			
IV	Culmen	Anterior quandrangular lobule	
V			
VI	Declive	Posterior quandrangular lobule	Lobulus simplex
VIIA	Folium	Superior semilunar lobule	Ansiform lobule, crus I Ansiform lobule, crus II
VIIB	Tuber	Inferior semilunar lobule Gracile lobule	Paramedian lobule
VIII	Pyramis	Biventral lobule	
IX	Uvula	Biventral lobule Tonsilla Accessory paraflocculus	Dorsal paraflocculus Ventral paraflocculus
X	Nodulus	Flocculus	Flocculus

Lobule numbers according to Larsell and Jansen (1972).

Table 3. Functional zones of the cerebellum

Cerebellar zone	*Cortex region*	*Deep nuclei nomenclature*		*Putative motor function*	*Putative cognitive function*
		Human	*Mammalian*		
Midline	Vermis	Fastigial Vestibular	Fastigius or medialis Vestibularis	Balance, eye movement, reflexes	Autonomic arousal, limbic regulation
Intermediate	Paravermal hemisphere	Globose Emboliform	Interpositus anterior Interpositus posterior	Sensorimotor integration, movement execution	Simple verbal responses to commands
Lateral	Lateral hemisphere	Dentate	Dentatus or lateralis	Preparation and planning of movements, fine motor dexterity, eye movements, imagined movements	Verbal association, rule-based learning, working memory, problem-solving, monitoring performance, temporal perception

type of feedforward processor that integrates much of the low-level sensory information it receives into signals that help direct the activity of higher-level brain areas.

Three types of evidence are used to determine the functions that the cerebellum is involved in. First, anatomical data help define what areas of the brain participate in cerebellar loops, and thus what types of signals the cerebellum is likely to process. Second, physiological data from imaging studies and recordings of cerebellar cells have indicated which cerebellar regions and neurons are active during specific tasks. Finally, data from clinical, pathological and behavioral studies have revealed what the functional consequences of cerebellar damage are.

ROLE OF THE CEREBELLUM IN MOVEMENT

Theories about the role of the cerebellum as a component of the motor control system have a venerable history, dating back at least to the eighteenth century. In the nineteenth and early twentieth centuries, Flourens, Luciani and Holmes separately described the results of cerebellar damage (for a thorough review of these and later studies see Dow and Moruzzi, 1958). During the First World War, Holmes described several discrete movement disorders that resulted from gunshot wounds of the cerebellum, and attempted to characterize these disorders based on the site of damage. In so doing, Holmes essentially defined the classic

'cerebellar syndrome' that has guided much of the research on the cerebellum ever since.

Cerebellar Lesions and Motor Function

Following unilateral lesions limited to the lateral lobes of the cerebellum, Holmes and others noted the loss of muscle strength and tone, most often involving the arms (but occasionally the legs), along with ataxia, rebound disorder, oculomotor disorders, postural disorders and scanning speech. The most conspicuous of these symptoms is ataxia, a generalized disorder of coordinating and executing voluntary movements. In their attempts to coordinate movement, cerebellar patients often exhibit asynergia, or difficulty coordinating muscular actions, and decomposition of movement, in which normal complex movements become broken down into single movements involving single joints. In executing movements, patients with cerebellar damage suffer from dysmetria, poorly directed movements which often miss their targets, and deviations from the line of movement, where a movement path does not follow the shortest line between two points. Possibly related to dysmetria is intention tremor, or involuntary shaking of the reaching hand or limb as it performs an action. Examples of these types of problems are easily seen when people with cerebellar damage are asked to touch their nose with one of their fingers. At first, the movement becomes decomposed. To reach the finger to the nose requires coordinated movements of the shoulder, elbow and hand, but is nonetheless readily performed in one smooth motion by normal individuals. People with cerebellar damage, however, might perform this movement by first lifting the shoulder until the arm is horizontal, and then flexing the elbow separately to bring the hand close to the nose. Finally, when the finger approaches the nose, involuntary alternating contractions of the muscles in the hand and wrist cause the finger to successively undershoot and overshoot the nose. Interestingly, high levels of ethanol consumption produce a similar effect, probably due to its action on the cerebellum, forming the basis of the modern sobriety test.

Another specific problem of coordination often seen in people with cerebellar damage is dysdiadochokinesis, or difficulty in performing rapidly alternating movements, such as slapping the surface of a desk repeatedly first with the palm of the hand and then the back of the hand. In dysdiadochokinesis the movements of the affected limb slow down and decrease in amplitude.

Midline cerebellar lesions often also lead to oculomotor and postural disorders. Some of these may be analogous to the weakness and dysmetria of the limbs seen with lateral cerebellar lesions. For example, people with cerebellar damage often exhibit difficulty maintaining their gaze on moving objects, and may display nystagmus – continuous oscillations of the eyes as if they were watching a train pass. Speech is often also affected in cerebellar lesions, causing scanning speech in which articulation is difficult and often staccato. Importantly, however, none of the effects of cerebellar lesions on motor performance that have been described implies that the cerebellum is essential for the direct control of movement. Indeed, lesions of the cerebellum do not prevent the initiation of movement, but rather cause problems in the optimal regulation and performance of ongoing movement.

Another aspect of cerebellar organization revealed by Holmes's early work, and reinforced by many anatomical and physiological mapping studies, has been the realization that the functions of the cerebellum are somatotopically organized: that is, the regions of the cerebellum concerned with the sensation and movement of the arms, legs, eyes and face are largely separate from each other. This is true even though these body parts are each represented multiple times throughout the cerebellar cortex and deep nuclei.

Cerebellar Physiology

Knowledge of the effects of cerebellar damage formed a backdrop to the earliest physiological studies of the cerebellum. Research by Eccles and his colleagues (Eccles *et al*., 1967) and more recently by Rodolfo Llinas (Llinas and Sotelo, 1992) has dissected the basic physiology of the cerebellum and particularly the cerebellar cortex. Indeed, the basic firing properties of the major cell types in the cortex have been worked out in detail. The main output cell of the cerebellar cortex is the Purkinje cell; these distinctive cells are notable in that they fire two distinct types of action potentials, simple spikes and complex spikes (Thach, 1968). (*See* **Single Neuron Recording**)

Simple spikes are typical neuronal action potentials, in which the neuron's membrane voltage changes rapidly from about −50 mV to +50 mV, and then returns again to its resting membrane potential of about −50 mV. This entire cycle occurs in the space of less than 3 ms when the sum of a Purkinje cell's inputs from other sources, notably parallel fibers, exceeds a threshold voltage.

These simple spikes can occur at rates of up to $100\,s^{-1}$. Simple spikes are known to occur in a predictable fashion with a variety of volitional movements. The relations between behavior and simple spike discharge have been especially well characterized during arm movements, head and neck movements, smooth pursuit and vestibulooocular reflex eye movements, and during walking. Generally, the firing rates of Purkinje cells during movement can be related to specific aspects of the movement, such as velocity of joint rotation or eye movement, or direction of arm movement.

Complex spikes, on the other hand, are not typical at all. During a complex spike, the neuron's membrane voltage increases rapidly, but then stays at a high value for an extended period, lasting perhaps 20 ms or longer. While the membrane potential is depolarized, the neuron typically fires many action potentials. The complex spikes occur in a one-to-one relationship with the arrival of action potentials on the climbing fibers. Thus, a single action potential from the inferior olive produces a profound and long-lasting depolarization in the membrane potential of a Purkinje cell. Typically, complex spikes occur only once or twice per second. Complex spiking in Purkinje cells can often be reliably produced by application of stimuli to distinct patches of skin. Interestingly, during behavior, complex spiking can often be most readily related to the occurrence of unexpected stimuli. During walking, for example, if an unexpected obstacle prevents the swing of the leg, complex spikes are often seen in the lateral lobes.

Explaining the relations between simple spikes, complex spikes and behavior in terms of the operation that the cerebellum performs, however, has proved remarkably difficult. Two general types of models, not necessarily exclusive, have been proposed to describe the operations of the cerebellum in motor control: cerebellar learning hypotheses and cerebellar timing hypotheses. According to learning hypotheses, the cerebellum learns maps between input patterns and output patterns. The core idea, proposed independently by David Marr and James Albus around 1970, is that the strength of synapses between parallel fibers and Purkinje cells can be modified, and that the changes in the strength of those synapses are regulated by the action of climbing fibers. While evidence for this idea has proved both elusive and controversial (Bell *et al*., 1996), a great deal of cerebellar research has been inspired by this theory. Work by Masao Ito and colleagues (Ito, 1984) suggested that complex spikes induce a long-term change in the responsiveness of Purkinje cells to parallel fiber inputs, termed 'long-term depression' (LTD). While LTD is reliably produced in experimental situations, it is not clear whether it also occurs in more physiological conditions – see work by Bloedel and Kelley, in Llinas and Sotelo (1992). It also seems likely that the cerebellum plays a part in classical conditioning such as the nictitating membrane reflex, but the original idea that the engram for this conditioning lies in the cerebellum itself has not been borne out by all researchers (see Bell *et al*., 1996, and special issues of *Trends in Neurosciences* and *Trends in Cognitive Science* listed under 'Further Reading'). In contrast, there is accumulating evidence that the cerebellum is a critical storage site for the storage of the engram established during the learning of some operantly conditioned behaviors (see Milak *et al*., 1997). (*See* **Long-term Potentiation and Long-term Depression**)

Another idea consistent with the observations of asynergia and decomposition of movement is that the cerebellum participates in coordination by handling problems of timing. The core idea was initially proposed by Valentino Braitenberg, who noted that the parallel fibers stretching across a row of Purkinje cells looked something like a delay line, and proposed that this feature might enable the cerebellum to select the relative timing of output events. Research by Ivry and Keele has indeed shown that cerebellar patients often have difficulty judging and reproducing temporal intervals. More recently, evidence from several laboratories, notably the experimental work of Fred Miles and Steve Lisberger in eye movements and Jim Bloedel in limb movements, as well as the theoretical work of Chris Miall, has suggested that complex spikes might serve as a generalized error signal, reporting that in some way behavioral output is not meeting expected outcomes. Thus, the inferior olive would communicate an error signal to the cerebellar cortex that it needed to modify its output. The specific mechanism by which this signal leads to a change in the processing of information by the cerebellum is not yet clear, but it does seem to be consistent with a number of lines of evidence, including the error signal idea. (*See* **Time Perception and Timing, Neural Basis of**)

DOES THE CEREBELLUM PLAY A PART IN COGNITION?

The cerebellum has received a great deal of attention regarding its potential involvement in cognitive functions. In discussing this possibility, it is helpful to define precisely what is meant by the term 'cognition'. For many researchers, the term is

often operationally defined as an 'awareness' of behavior and the ability to voluntarily modify it. Using this definition, the evidence for the cerebellum's involvement in cognitive behavior is compelling. However, this involvement does not imply that the cerebellum is essential for that behavior. Animals and humans lacking a functioning cerebellum can interact meaningfully and demonstrate awareness of their behavior. Nonetheless, studies have shown that when the cerebellum is badly damaged, there are alterations in certain types of cognitive behavior that can be assessed with well-designed formal tests. When evaluating such studies, however, it is important to keep in mind the nature of the damage and the nature of the behavior being measured before reaching any conclusions. For example, it has been reported that individuals with localized cerebellar damage have deficits on many standardized intelligence and psychological tests. However, a close inspection of the results in some of these studies reveals that most of the tests that the cerebellar patients performed badly placed considerable demands on hand–eye coordination (visuomotor abilities) and the speed of responding, parameters that have little to do with formal cognitive processing. Thus, from these types of studies it is often not clear what type of cognitive deficits are present that are independent of motor deficits. In order for the cerebellum to be accepted as a true nodal point for normal cognitive processing, it is necessary to show that cerebellar damage can produce cognitive deficits without affecting motor performance. To date, very few studies have reported this type of finding for cerebellar patients.

Fiez and colleagues performed psychological testing on a patient with a localized lesion resulting from a small stroke in the lateral portion of one cerebellar hemisphere (Fiez *et al*., 1992). This patient had excellent scores on all standard intelligence and language measures. However, when he was given tasks to perform that involved rule-based learning, verbal association and planning, he showed deficits in the ability to improve his performance with practice, and mild deficits in the actual tasks themselves. Similar types of practice-related and performance deficits were also reported in studies of cerebellar patients by Jordan Grafman and colleagues (1992) at the National Institutes of Health. In addition, Akshoomoff and Courchesne (1992) studied a group of people with widespread cerebellar damage and concluded that one of the main consequences of cerebellar damage in humans was reduced verbal fluency and verbal association capacity, and impaired visuospatial skills. Finally, work by Ivry and Keele (1989) has shown that patients with lesions of the lateral cerebellum, but not the medial cerebellum, display impaired perception of temporal intervals. Thus, it appears that certain cognitive tasks do involve the cerebellum, particularly the lateral portions of it, and that improving the performance of these tasks, or detecting performance errors, requires the activity of an intact cerebellum. Moreover, in some cases the cognitive deficits appear to be present without considerable motor deficits. Therefore, it is possible that the deficits in verbal association, visuospatial skills, rule-based learning, planning and error detection form the core cognitive symptoms of cerebellar damage.

In addition to studies of cerebellar damage, perhaps the best evidence in support of the cerebellar involvement in cognitive function is derived from studies of cerebellar activation during cognitive tasks using brain imaging techniques such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI); for a review of this topic see Desmond and Fiez (1998). Importantly, with these approaches, it is possible to compare the activation seen during cognitive tasks with that seen during motor tasks, and thus reach more accurate conclusions regarding the patterns of activity one sees. In one of the early PET studies of cerebellar activation, Fiez and colleagues (1992) examined the potential cerebellar involvement in cognitive and language tasks. Participants in the study were asked to vocalize a verb after being presented with a noun; for example, given the noun 'dog', a participant might generate the verb 'bark'. These nouns were presented by auditory or visual means, and the levels of cerebellar activity during this task were compared with the activity seen during two types of control task. In the sensory control task, participants only looked at or listened to the noun, and in the motor control task for speech output, participants immediately repeated the noun they were presented with. In the motor control task, activation was confined to the midline cerebellar hemisphere and to motor and premotor areas of the cerebral cortex normally seen activated during speech. In the verb generation task, however, in addition to the areas activated during the motor control task, there was activation of the lateral cerebellar cortex and prefrontal regions of the cerebral cortex. Whether this additional activation was related to purely language functions or a working knowledge of the strategy involved in performing the task is not clear, and both are likely possibilities. This same task, however, was the one that the cerebellar

patient studied by Fiez and colleagues was found to be profoundly deficient on. In another imaging study, Decety *et al*. (1990) reported activation of the cerebellum in participants who were asked to imagine swinging a tennis racket, and to silently count the number of times they hit the ball. In comparison with people at rest, the people performing the mental imagery had a 10% increase in lateral cerebellar activity during silent counting and a 20% increase during the mental imagery condition. Finally, using fMRI, Kim *et al*. (1994) examined the activation of the lateral output nucleus of the cerebellum (the dentate) while participants sat in a scanner and either thought about ways to solve a difficult pegboard puzzle, or made visually guided movements of pegs in a pegboard. The participants had only mild increases in activity during the motor task, but displayed extensive activation of the dentate while mentally working out possible solutions to the puzzle. Taken together, these imaging studies show that whether or not observable movement takes place, there are regions of the cerebellum that display prominent activation during cognitive tasks, and that these areas appear to be different from those activated during simple motor tasks. This finding supports the idea that the cerebellum can process both cognitive and motor information using parallel circuits. (*See* **Planning: Neural and Psychological**; **Neuroimaging**)

ROLE OF THE CEREBELLUM IN COGNITIVE FUNCTION

The cerebellum is not essential for cognitive thought processes. However, the cerebellum is clearly involved in the performance of some cognitive tasks. The demonstration that cerebellar damage can alter affect behaviors that are known to rely on specific cognitive areas of the cerebral cortex suggests that the cerebellum must be communicating with these areas to exert its influence. In monkeys and cats, reciprocal cerebellar connections have been described with regions of the parietal cortex concerned with spatial orientation and attention and also areas of prefrontal cortex concerned with planning, rule-based learning, short-term working memory and verbal association. There are also reciprocal cerebellar connections with neuroendocrine areas of the brain, including parts of the hypothalamus. Together, these connections allow the possibility of direct cerebellar involvement in modulating limbic and autonomic functions as well as cognitive and motor functions. Such connections with primitive and highly advanced brain areas indicate that the cerebellar involvement in modulation of nonmotor function has a long-standing precedent and is not unique to humans. (*See* **Spatial Attention, Neural Basis of**; **Frontal Cortex**; **Hypothalamus**)

Although we know a great deal about the functions of the cerebellar microcircuitry and how it operates within the domain of vestibular, motor and sensory processing, we do not yet have any real insight into what the cerebellum might be doing during cognitive processing, or any idea at all how it utilizes the signals it receives from limbic, neuroendocrine and neurotransmitter synthesizing nuclei. We assume that the cerebellum does the same thing for cognitive tasks that it does for noncognitive tasks – namely, that it improves overall task performance by using its circuitry to fine-tune the speed, efficiency and accuracy of the response or planned response. However, one of the challenges of future research will be to show how this is achieved at the cellular, molecular and physiological levels for all types of behavior that fall under the influence of the 'little brain'.

References

Akshoomoff NA and Courchesne E (1992) A new role for the cerebellum in cognitive function. *Behavioral Neuroscience* **106**: 731–738.

Bell C, Cordo P and Harnad S (eds) (1996) *Controversies in Neuroscience IV: Motor Learning and Synaptic Plasticity in the Cerebellum*. Special issue of *Behavioral and Brain Sciences* **19**(3).

Decety J, Sjoholm H, Ryding E, Stenberg G and Ingvar DH (1990) The cerebellum participates in mental activity: tomographic measurements of regional cerebral blood flow. *Brain Research* **535**: 313–317.

Desmond JE and Fiez JA (1998) Neuroimaging studies of the cerebellum: language, learning and memory. *Trends in Cognitive Sciences* **2**: 355–362.

Dow RS and Moruzzi G (1958) *The Physiology and Pathology of the Cerebellum*. University of Minnesota Press, Minneapolis.

Eccles JC, Ito M and Szentagothai J (1967) *The Cerebellum as a Neuronal Machine*. New York, NY: Springer.

Fiez JA, Petersen SE, Cheney MK and Raichle ME (1992) Impaired non-motor learning and error detection associated with cerebellar damage. *Brain* **115**: 155–178.

Grafman J, Litvan A, Manaquoi S, Stewart M, Sirigu A and Hallett M (1992) Cognitive planning deficit in patients with cerebellar atrophy. *Neurology* **42**: 1493–1496.

Ito M (1984) *The Cerebellum and Neural Control*. New York: Raven Press.

Ivry RB and Keele SW (1989) Timing functions of the cerebellum. *Journal of Cognitive Neuroscience* **1**: 136–152.

Jansen J (1972) Features of cerebellar morphology and organization. *Acta Neurologica Scandinavica* **51**: (supplement) 197–217.

Kim SG, Ugurbil K and Strick PL (1994) Activation of a cerebellar output nucleus during cognitive processing. *Science* **265**: 949–951.

Larsell O and Jansen J (1972) *The Human Cerebellum, Cerebellar Connections, and Cerebellar Cortex*. University of Minnesota Press, Minneapolis.

Llinas RL and Sotelo C (eds) (1992) *The Cerebellum Revisited*. New York, NY: Springer.

Milak MS, Shimansky Y, Bracha V and Bloedel JR (1997) Effects of inactivating individual cerebellar nuclei on the performance and retention of an operantly conditioned forelimb movement. *Journal of Neurophysiology* **78**: 939–959.

Thach WT (1968) Discharge of Purkinje and cerebellar nuclear neurons during rapidly alternating arm movements in the monkey. *Journal of Neurophysiology* **31**: 785–797.

Voogd J (1975) Bolk's subdivision of the mammalian cerebellum. Growth centres and functional zones. *Acta Morphologica Neerlando-Scandinavica* **13**: 35–54.

Further Reading

Altman J and Bayer SA (1997) *Development of the Cerebellar System: In Relation to Its Evolution, Structure, and Functions*. CRC Press.

De Zeeuw CI, Strata P and Voogd J (eds) (1997) *The Cerebellum: From Structure to Control*. Special issue of *Progress in Brain Research* **114**.

Middleton FA and Strick PL (eds) (1998) *The Cerebellum*. Special issues of *Trends in Neurosciences* (**21**) and *Trends in Cognitive Science* (**2**).

Schmahmann JD (ed.) (1997) *The Cerebellum and Cognition*. Special issue of *International Review of Neurobiology* (**41**).

Cerebral Commissures

Intermediate article

Sandra F Witelson, McMaster University, Hamilton, Ontario, Canada
Debra L Kigar, McMaster University, Hamilton, Ontario, Canada
Alison Walter, McMaster University, Hamilton, Ontario, Canada

CONTENTS

The cerebral commissures house the fibers that interconnect the two hemispheres of the brain. The main fiber tract, namely the corpus callosum, varies greatly in size in the human brain. Factors such as sex of the individual and chronological age affect its size. Relationships exist between callosal size, degree of functional asymmetry and cognitive ability. The callosum may play a crucial role in the experience of conscious unity.

INTRODUCTION

With the increasing prominence of the neocortex in the mammalian brain, a system of neocortical commissures developed which provide a direct route for interhemispheric exchange of sensory and stored information. The importance of these commissures for cognitive function is now well established, but this issue has had a chequered history. As early as the eighteenth century, when neuroanatomists still did not know of the existence of neurons and their axons, the corpus callosum was believed by some to be the seat of the soul or the center of rationality or of higher thought. However, during the first half of the twentieth century, most authors considered the corpus callosum to have no function at all, except perhaps as a mechanical support for the cerebral cortex. There were exceptions, like Ramón y Cajal, who postulated that since memory centers for language are unilateral, but perceptual centers are present on both sides, the corpus callosum must be necessary to allow information to flow between the two hemispheres. By the 1960s, experimental research in animals and then studies of patients in whom the corpus callosum was surgically divided demonstrated the functional necessity of the callosum. Most recently of all, the role that the corpus callosum may play in the experience of human consciousness and personal identity is being reconsidered.

STRUCTURE

Gross Anatomy and Functional Topography

The corpus callosum (CC) is the main fiber tract connecting the cortical neurons between the two cerebral hemispheres. It is a major feature of the mammalian brain. The human CC is a large structure, about 7 cm in length from front to back, and it contains about 300 million fibers or axons. Axons from callosal neurons pass through the CC in a systematic pattern. Figure 1 is a photograph of a postmortem brain showing the CC and some anatomical regions. In general, the most anterior curved region of the CC, namely the genu, interconnects the prefrontal regions. The body of the CC interconnects the precentral and postcentral regions. The posteriormost part of the CC body, referred to as the isthmus due to its narrowing in height relative to the other regions, connects the posterior temporal and anterior parietal cortical regions between the hemispheres. The size of the isthmus has been particularly linked to the asymmetry in function between the two hemispheres, and to sex differences in CC anatomy (issues which will be discussed in more detail later). The posterior bulbous region is the splenium, which connects mainly occipital lobe cortex. The splenium has also been the subject of numerous studies concerning differences in CC size between the sexes.

Experimental research using methods of retrograde degeneration following cortical lesions and autoradiographic tracing techniques in monkeys have shown the specificity of the systematic pattern of the location where axons course through the CC to the opposite hemisphere (Figure 2). This topography is consistent with results obtained from neuropsychological studies of commissurotomized (split-brain) patients who underwent sectioning of

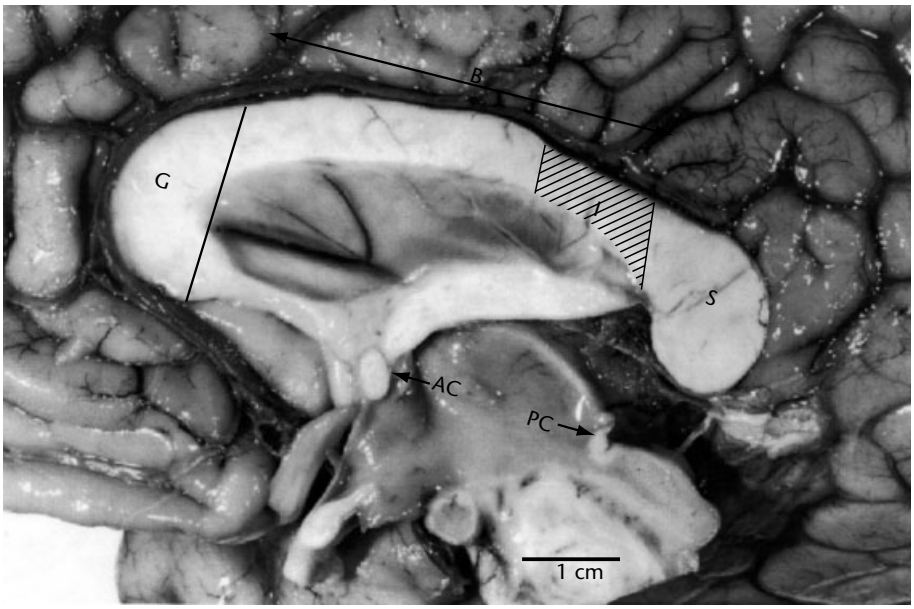


Figure 1. Photograph of the midsagittal view of an adult brain highlighting the three commissures, namely the corpus callosum (CC) and its main subregions [G, genu (most anterior region); B, body; I, isthmus (shaded); S, splenium] the anterior commissure (AC) and the posterior commissure (PC). Scale bar represents 1 cm.

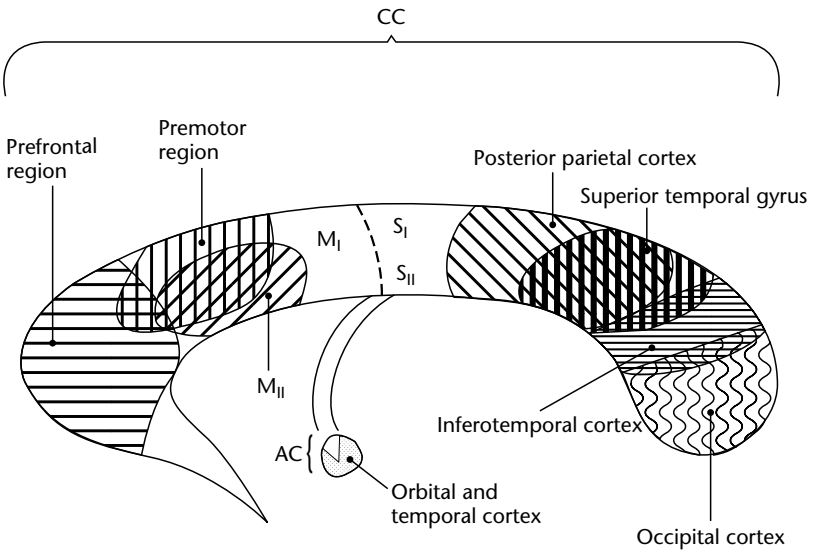


Figure 2. Diagrammatic representation of the midsagittal section of the corpus callosum (CC) and anterior commissure (AC) of the rhesus monkey, showing the topography of commissural fibers from different parts of the cerebral cortex. M_I, precentral motor region; M_{II}, supplementary motor region; S_I and S_{II}, postcentral somesthetic gyri. (Adapted from Pandya DN and Seltzer B (1986) The topography of commissural fibers. In: Leporé F, Ptito M and Jasper HH (eds) *Two Hemispheres – One Brain: Functions of the Corpus Callosum*, pp. 47–73. New York: Alan R. Liss.)

the CC for the relief of epilepsy that was uncontrollable by drug treatment. In some cases, the complete CC is not cut. Sectioning the posterior CC regions results in loss of transfer of visual or tactual information between the hemispheres. Sectioning only the anterior regions does not interfere with modality-specific cognition, but does interfere with more general cognitive functions such as semantic processing and aspects of memory (Sperry, 1974). The precise topography of the CC, as well as

the fractionation of cognition that study of the CC makes possible, are revealed by other clinical cases, such as those with callosal lesions. The latter indicated that the anterior to middle splenium is involved in the transfer of pictorial information from the language-nondominant hemisphere to the language-dominant side, and that the ventroposterior part is involved in the transfer of letter information.

Sex Differences

There is considerable variation in the size and shape of the CC between individuals. Similar to the larger volume and weight of the male brain compared with that of the female, anatomical studies at the turn of the last century revealed that the CC was larger in area in the midsagittal plane in men than in women. However, during the last few decades, several studies reported that some subregions, mainly the splenium (de Lacoste-Utamsing and Holloway, 1982) and the isthmus (Witelson, 1989), were larger in women than in men relative to total callosal size. These sex differences have been interpreted as a possible anatomical correlate of the sex differences in the pattern and degree of hemispheric functional asymmetry. The relatively larger CC subregions in women would be consistent with the greater bihemispheric representation of some aspects of language in women than in men. A larger CC could allow for more transfer of information back and forth between hemispheres. These first studies were based on direct measurement of postmortem brains. However, magnetic resonance imaging (MRI) scans provide relatively clear pictures of the CC, although not without some ambiguity of exact boundaries (Figure 3). This new technology has been used in numerous studies of CC size. The results of these studies revealed considerable inconsistency, which led to an extensive review and meta-analysis of 49 postmortem and MRI studies conducted since 1980. It was concluded that the CC is larger in absolute size in men than in women, and that subregions (when CC size is taken into account) are not significantly different in size between the sexes, with the exception of the isthmus, for which there is some consistency of evidence for a larger isthmus in women, albeit with a difference of small magnitude (Bishop and Wahlsten, 1997).

Most studies have used one of two methods to measure the subregions of the CC. The earlier method employs an arithmetic parcellation of the CC into seven subregions that are determined mainly by dividing it into halves and thirds along the maximal longitudinal CC length (Figure 4a).

Figure 3. MR image showing the midsagittal view of a normal adult brain, illustrating the clear depiction of the corpus callosum (CC) that is obtained on scans. However, some regions of the CC are difficult to distinguish from the surrounding structures in scans, such as the boundaries of the rostrum (R) of the CC, and the junction of the CC body and fornix (F). G, genu; S, splenium.

The cortical regions related to each region may be roughly estimated. As further study of CC size ensued, a more sophisticated computerized method was developed which obtained the widths (height) of 99 percentile divisions of the CC along its curved longitudinal axis (Figure 4b). Studies then measured the area of several groups of widths that approximated the subregions of the arithmetic method, or that were obtained by factor analysis. Studies of CC size raise the important issue for anatomical studies in general of how to correct or control for variation in brain size as a baseline. Different normalization techniques using ratio scores, analyses of covariance, or stereotaxic methodology which scales MRI scans into standardized space were used on the same data sets and revealed different results with regard to the differences in CC size (Bermudez and Zatorre, 2001). (*See* **Sex Differences in the Brain**)

Microscopic Anatomy

Variation in size of the CC could reflect various histological differences, such as a different total number of axons, a different proportion of thick

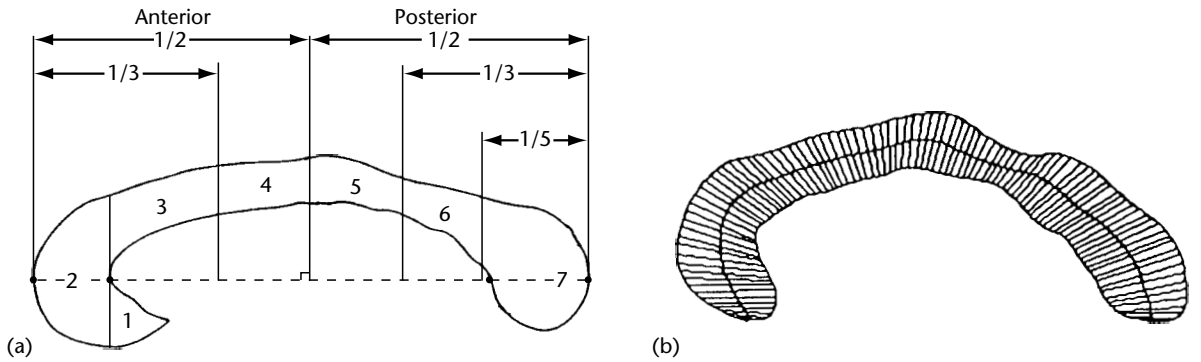


Figure 4. Sketches showing the two different methods for dividing the CC into subdivisions for measurement. (a) Arithmetic method: diagram of the midsagittal view of the CC of the human adult, showing seven subregions. The broken line is the linear axis used to subdivide the CC arithmetically into anterior and posterior halves, as well as anterior, middle and posterior thirds, and the posterior one-fifth region (region 7) which is roughly congruent with the splenium. The boundary line for the genu (region 2) is drawn perpendicular to the linear axis. Regions 3, 4, 5 and 6 constitute the body of the CC. (Adapted from Witelson SF (1989) Hand and sex differences in the isthmus and genu of the human corpus callosum: a postmortem morphological study. *Brain* **112**: 799–835.) (b) Computerized method: a tracing of a human CC showing the division into 99 minimum widths along the curved longitudinal axis. The widths are determined by dividing the dorsal and ventral CC perimeters into 99 percentiles and connecting the corresponding points. (Adapted from Denenberg *et al.* (1991) A factor analysis of the human's corpus callosum. *Brain Research* **548**: 126–132, with permission from Elsevier Science.)

(large-diameter) and thin fibers, or variation in the amount of myelination. A larger CC could have the same number of fibers as a smaller CC, but with a smaller packing density, or it could have a greater number of fibers with the same packing density. Different microscopic anatomy could have functional consequences. Using light microscopy, Aboitiz and colleagues found that the overall density of callosal fibers was not significantly correlated with callosal area, indicating that an increased CC area housed an increased total number of fibers (Aboitiz *et al.*, 1992). However, there was regional differentiation in the density of fibers of different size. Thin fibers were most dense in the anterior third of the CC, and their density decreased to a minimum in the posterior body. Thicker fibers had a complementary pattern, having a peak density in the posterior body. Thicker and more myelinated fibers have faster rates of conductivity. Relatively more thin fibers interconnect higher-order association cortex. It may be that the longer interhemispheric transfer time over thin fibers may be irrelevant for higher-order cognition compared with more direct sensory information.

Small Commissures

Fibers also cross the midline in two other commissures, namely the anterior and posterior commissures (Figure 1). Each is small, on average less than $10\,mm^2$ in cross-section. The anterior commissure houses fibers from the orbitofrontal regions and from the anterior and midregions of the temporal lobe. The posterior commissure connects the diencephalic and midbrain structures. These two commissures are important features in brain imaging, as they are readily viewed and serve as stereotactic landmarks to define the horizontal axis of the brain.

The anatomy of the anterior commissure has undergone an interesting evolutionary change which has functional relevance. Phylogenetically it is older than the neocortex. In lower mammals (e.g. the lemur), it primarily connects regions related to olfaction. Marsupials (e.g. kangaroos) are the one group of mammals that do not have a CC, and in this case the anterior commissure plays a key role in transmitting information between the hemispheres. In humans and other higher primates, the anterior commissure has anterior and posterior limbs, the anterior limb being part of the olfactory system. The posterior limb has neocortical fibers from auditory and visual association regions of the temporal lobe. In cases of agenesis or lack of development of the CC, the anterior commissure appears to play a role in functional plasticity and to mediate some information that is usually transmitted via the CC. The anterior commissure is highly variable in size, which appears to be related to gender and the pattern of hemispheric functional asymmetry. (*See* **Reorganization of the Brain**)

CORPUS CALLOSUM DEVELOPMENT OVER THE LIFESPAN

Fetal and Early Development

Early in brain development, a temporary bridge of glial cells (known as the glial sling) forms across the hemispheric midline, and commissural fibers cross supported by the sling. In some cases of abnormal development this process does not occur, resulting in callosal agenesis. The anterior commissure may then play a greater role in interhemispheric connectivity, but the pattern of hemispheric functional asymmetry is atypical. Initially in development there is an exuberance of callosal fibers, and subsequently some axons are eliminated. For example, this was demonstrated dramatically by electron microscopic analysis in the monkey (LaMantia and Rakic, 1990). In addition, axons are eliminated in a systematic pattern (e.g. in the monkey, up until 1 month before birth there are callosal axons connecting the primary and secondary somesthetic cortex, which are then eliminated, resulting in the adult pattern). Research on fetal and infant brains has been conducted via postmortem study. A review of such studies has indicated that the size of the CC doubles in fetal life and then doubles again within the next 2 years. CC size was found to correlate with gestational age, a feature that is useful for identifying normal versus abnormal brain development *in utero*. The CC is thin during the first few months of postnatal life, but the genu and splenium grow rapidly over the next few months. During this period myelination begins, and the adult shape and orientation of the CC are achieved by about 8 months of age.

There has been relatively little research on CC anatomy during childhood and adolescence. In a recent large longitudinal MRI study, it was found that the area of the CC increases substantially from 5 to about 18 years of age (Giedd *et al*., 1999), when it reaches adult size. This increase is probably due to myelination. Now that high-resolution structural MRI scans are possible, growth of the CC and its subregions can be studied in relation to cognitive development over periods of marked cognitive changes.

Adulthood

Even after the mature CC has developed, there is great variation among individuals, some of which is related to the sex of the person (see above), while some may be associated with variation in cognitive ability (see below). Age is another major factor. Both postmortem and MRI studies have demonstrated negative correlations between CC size and age, with men showing an earlier and more rapid change than women. Frontal CC regions show the greatest change with age, paralleling the more marked change in frontal lobe regions with advancing age. Table 1 summarizes the results of the studies of CC size as a function of age. As can be seen from the table, the results of the MRI studies are somewhat inconsistent with regard to the relationship between CC area and age, as well as with regard to the larger overall CC in men than in women. Although the CC is one of the most clearly visualized structures in MRI scans, the exact boundaries may be ambiguous in some regions, depending on the contrast of voxels in the surrounding structures (Figure 3). (*See* **Neuroimaging**)

Changes in the overall size of the CC must reflect some microscopic change. Only electron microscopy, whose demanding requirements for tissue preparation are problematic in non-experimental situations, can reveal all unmyelinated axons reliably. Consequently, little research has been done in this area. However, one such study reported no relationship between age and the total number of CC fibers, suggesting that the decreasing size of the CC with age may be due to loss of myelin or to a reduction in axon diameter, rather than being due to loss of axons.

Hormonal and Environmental Effects on Corpus Callosum Anatomy

Experimental research has demonstrated that sex hormones contribute to the larger overall CC in the male rat brain compared with that of the female. The rat brain is sensitive to the effects of hormones after birth. During this period of sexual differentiation of the brain, testosterone results in an increase in CC size and, in contrast, estrogen actively inhibits CC growth (Mack *et al*., 1996).

Several recent studies have documented the strong genetic component in CC size. Studies of monozygotic versus dizygotic twin pairs have shown that heritability may account for as much as 70–90% of the variation. A related issue is whether environmental/experiential factors affect callosal anatomy. Later in this article it will be noted that although left-handers or non consistent-right-handers (people who use their left hand for even a few tasks) have a larger CC than consistent-right-handers, the experience of an early forced shift from left- to right-hand writing does not appear to affect CC size (Witelson, 1989). However, there is some evidence that rats living in enriched

Table 1. The corpus callosum and age

Study	*Subjects*	*Method*[a]	*Findings*[b]
Witelson, *Brain*, 1989	50 brains from cognitively normal subjects 25–70 years 15M; 35W	Post mortem Total CC	M: $r = -0.6$ W: ns CC area: M > W
Witelson, *New England Journal of Medicine*, 1991	62 brains from cognitively normal subjects 25–70 years 23M; 39W	Post mortem Total CC Three CC subregions	M: total CC, $r = -0.7$ Genu: $r = -0.7$ Body: $r = -0.5$ Splenium: $r = -0.5$ W: ns
Doraiswamy *et al*., *Journal of Neuropsychiatry*, 1991	36 normal subjects 26–79 years 16M; 20W	MRI Total CC	M: $r = -0.6$ W: ns No sex difference in CC area
Cowell *et al*., *Developmental Brain Research*, 1992	73 pairs of age-matched M and W 2–79 years	MRI Total CC 99 widths (see Figure 4b)	M: CC area peaks at about 20 years W: CC area peaks at about 50 years Peaks followed by decline, especially in genu and posterior splenium Little decline in body, isthmus and anterior splenium
Parashos *et al*., *Journal of Neuropsychiatry*, 1995	80 normal subjects 30–91 years 28M; 52W	MRI Total CC Five CC subregions	CC areas decrease with age No sex difference in CC areas
Salat *et al*., *Neurobiology of Aging*, 1997	76 subjects 65–95 years 31M; 45W	MRI Three CC subregions	M: ns W: anterior region, $r = -0.4$ mid region, $r = -0.4$ Posterior CC region: W > M
Sullivan *et al*., *Neurobiology of Aging*, 2001	92 subjects 22–71 years 51M; 41W	MRI Total CC Three CC subregions	No decrease with age CC areas: M > W

M, men; W, women.
[a]Method: midsagittal areas.
[b]Findings: *r*-values are CC area as a function of age.

as opposed to deprived environmental conditions have enlarged callosa. In humans, literacy (as opposed to being illiterate) and early musical training have been associated with larger CC size. Thus environment as well as heredity may affect CC size.

RELATIONSHIP OF THE CORPUS CALLOSUM TO FUNCTIONAL ASYMMETRY AND COGNITION

The Corpus Callosum and Hemispheric Functional Asymmetry

The pattern and degree of hemispheric specialization varies between individuals. It is possible that variation in CC anatomy is related to brain lateralization. Several studies have examined the relationship between callosal anatomy and neuroanatomical asymmetries, with varying results. Studies of asymmetries of the Sylvian fissure or the posterior surface of the superior temporal gyrus (planum temporale) in relation to CC size have found that the anatomical asymmetries were correlated with posterior CC regions. Since the posterior CC houses fibers which connect right and left posterior temporo-parietal regions which are asymmetrical in function between the two hemispheres, these results support the hypothesis that the extent of callosal connectivity is related to the direction and degree of anatomical asymmetry

Table 2. The corpus callosum and hemispheric functional asymmetry

Study	*Subjects*	*Method*[a]	*Lateralization measure*	*Findings*
Witelson, *Science*, 1985	42 brains from cognitively normal subjects 25–65 years 12M; 30W 27 CRH; 15 nonCRH	Post mortem Total CC Seven CC subregions (see Figure 4a)	Hand preference (tested, 12 items)	Total CC, anterior half, posterior half and posterior fifth areas Total group: NonCRH > CRH
O'Kusky *et al.*, *Annals of Neurology*, 1988	50 normal controls (for study of 50 epileptic subjects) 17–60 years M and W	MRI Total CC Five CC subregions	Hand, foot and eye preference Verbal dichotic listening 50 epileptic subjects, 44 had Wada testing for speech lateralization	No association of CC area with hand preference No sex difference in CC In patients, those with right-hemisphere speech lateralization had larger CC
Witelson and Goldsmith, *Brain Research*, 1991	22 brains from cognitively normal subjects Mean age 54 years All M 13 CRH; 9 nonCRH	Post mortem Total CC Seven CC subregions (see Figure 4a)	Hand preference (tested, 12 items)	Total CC: nonCRH > CRH Isthmus and handedness score: $r = -0.5$
Habib *et al.*, *Brain and Cognition*, 1991	53 normal subjects 18–51 years 35M; 18W 26 CRH; 27 nonCRH	MRI Total CC Six CC subregions	Hand preference (10-item questionnaire)	Total CC, subregions: M: NonCRH > CRH W: ns No sex difference in CC
Hines *et al.*, *Behavioral Neuroscience*, 1992	28 subjects All W 20–45 years	MRI Total CC Three CC subregions	Hand preference (tested, 18 items) Language lateralization	Posterior CC area and language lateralization: $r = -0.3$
Cowell *et al.*, *Neurology*, 1993	104 normal subjects 18–49 years 51M; 53W 52 RH; 52 LH	MRI Total CC CC divided into 99 widths, and widths grouped into 7 regions (see Figure 4b)	Hand preference (defined by writing hand only)	Rostral body (widths 22–39) and posterior midbody (widths 49–62) W: RH > LH M: LH > RH
Yazgan *et al.*, *Neuropsychologia*, 1995	11 normal subjects Mean age 34 years 11M; 2W RH	MRI Total CC Five CC subregions	Verbal dichotic listening Verbal–manual interference (VMI) test	CC and REA, $r = -0.7$ CC and VMI, $r = -0.6$
Jancke *et al.*, *Cerebral Cortex*, 1997	120 normal subjects 18–45 years 71M; 49W 54 CRH; 28 CLH; 38 MH	MRI Total CC Four CC subregions	Hand preference (tested, 12 items)	No sex difference for absolute CC areas Total CC and subregions: W > M for relative size Middle third CC area: CRH > CLH, MH

Moffat *et al*., *Brain*, 1998	16 subjects Mean age 23 years All M, LH 10 REA; 6 LEA RH male archival data	MRI Total CC Six CC subregions	Verbal dichotic listening Hand preference (tested, 8 items)	Total CC and posterior regions: LH > RH LH-REA subjects > LH-LEA or RH subjects Smaller isthmus associated with LEA

M, men; W, women; CRH, consistent-right-handed; nonCRH, non consistent-right-handed; RH, right-handed; LH, left-handed; MH, mixed-handed; REA, right-ear advantage; LEA, left-ear advantage.
[a]Method: midsagittal areas.

and possibly functional asymmetry. (*See* **Brain Asymmetry**)

One of the first behavioral measures that was found to be related to CC size was hand preference (Witelson, 1989). This result indicated that structure and function may be related in very direct ways (i.e. that the size of parts of the human brain may be correlated with behavioral measures). Since left-handers have a greater degree of bihemispheric representation of cognitive functions, it was suggested that the larger CC may provide a substrate for greater interhemispheric communication in left-handers. Subsequent studies using MRI measures of CC generally replicated and supported the finding of a larger CC and particularly some subregions in left-handers or in individuals who did not have the typical pattern of left-sided speech representation. As research progressed to address further issues, it appeared that the CC varies with both hand preference and speech lateralization independently to some degree. The CC was smaller in individuals who had what may be called 'ipsilateral' (same-sided) representation of hand preference and speech than in people who had 'contralateral' representation (e.g. left-handers with right-sided hand representation and left-sided speech representation), but it was not smaller than in left-handers with right-sided speech (also 'ipsilateral' cases) (Table 2).

The Corpus Callosum and Cognitive Correlates

Corpus callosum anatomy may provide a window on the anatomy and function of the cerebral cortex, and consequently it may reflect variation in cognitive processing. Several MRI studies have focused on CC size and its relationship to performance on verbal and spatial tasks in predominantly right-handed individuals (Table 3). In general, small but statistically significant positive correlations were observed. If larger callosa have a greater number of fibers and provide greater interhemispheric communication, these results suggest that, at least in strongly lateralized individuals, such anatomy may confer some functional advantage.

Although it seems remarkable to find relationships between ability and brain size – in this case CC area and, by inference, cortical anatomy – much further research is needed to elucidate the structure–function relationships and possible mechanisms. Such research also raises major ethical issues. MRI scans have the potential to be used as tests or indicators of ability by educational and professional bodies. Clearly caution and proactive consideration are needed with regard to such applications of basic research findings, which will likely only increase with time.

Split-brain Studies and Theoretical Implications

Commissurotomy is an operative procedure which permanently severs the CC or all tracts between the right and left hemispheres. Each hemisphere then only has access to the stimulation it receives and the neural processes that are established on that side. The neuropsychological study of split-brain patients carried out by neuroscientist Roger Sperry and colleagues yielded unambiguous information about functional lateralization of the brain. In such patients, the two sides of the brain are isolated from each other, so that neuropsychological testing can determine not only which side is dominant for specific cognitive functions, but also the limitations of each hemisphere, by determining which functions it cannot support. This has been done in a series of ingenious studies (Sperry, 1974). (*See* **Split-brain Research**)

The research with split-brain patients also raised the issue of the nature of conscious unity. For the development of this issue, in addition to

Table 3. The corpus callosum and cognition

Study	*Subjects*	*Method*[a]	*Cognitive measure*	*Findings*
Hines *et al.*, *Behavioral Neuroscience*, 1992	28 W 20–45 years	MRI Total CC Three CC subregions	Verbal fluency tests Visuospatial ability	Posterior CC and verbal fluency: $r = 0.6$ Anterior CC and visuospatial ability: $r = 0.4$
Strauss *et al.*, *Journal of Clinical and Experimental Neuropsychology*, 1994	47 epileptic subjects 12–57 years	MRI Total CC Five CC subregions	Wechsler Adult Intelligence Test (WAIS) (Full-scale IQ – FSIQ)	Splenial area and FSIQ: $r = 0.4$
Yazgan *et al.*, *Neuropsychologia*, 1995	11 normal subjects Mean age 34 years 11 M; 2W RH	MRI Total CC Five CC subregions	Line bisection WAIS-R	CC and line bisection: $r = 0.6$ CC and WAIS: ns
Atkinson *et al.*, *Journal of Neuroimaging*, 1996	20 age- and sex-matched controls (for epilepsy study) 12–47 years 8M; 12W	MRI Total CC area	WAIS-R Wechsler Memory-R	CC and FSIQ: $r = 0.3$ CC and Performance IQ: $r = 0.4$
Salat *et al.*, *Neurobiology of Aging*, 1997	76 subjects 65–95 years 31M; 45W RH	MRI Total CC Three CC subregions	Weschler Memory-R (logical memory and visual reproduction subscales) WAIS-R (block design subscale)	W: CC and visual memory: $r = 0.3$ M: ns in all tests
Davatzikos and Resnick, *Cerebral Cortex*, 1998	114 highly educated subjects 56–85 years 68M; 46W RH	MRI Total CC and subregions	Card rotations Boston naming test Letter fluency Verbal recognition memory test Figural recognition memory test	W: CC areas and all five neuropsychological tests: positive *r*-values M: ns
Peterson *et al.*, *Human Brain Mapping*, 2001	138 control subjects (for a study of multiple syndromes) 6–88 years M and W CRH = 86%	MRI Total CC area, widths, bending angle, splenial bulbosity, etc.	WAIS-R Performance IQ and Verbal IQ	A factor representing a thinner, more concave anterior body of the CC predicted higher IQ scores

M, men; W, women; CRH, consistent-right-handed; nonCRH, non consistent-right-handed; RH, right-handed; LH, left-handed; MH, mixed-handed; REA, right-ear advantage; LEA, left-ear advantage.
[a]Method: midsagittal areas.

experimental neurobiological work, Dr Sperry was awarded the Nobel Prize for Medicine in 1981. More recently, it has been suggested that this tract of white matter enables the human condition. The CC may have eliminated the necessity for having two redundant systems, by linking the two hemispheres to such a degree that information can be easily shared and transmitted to appropriate sensory or motor systems. It allows for specialization of the two hemispheres, and it provides the anatomical substrate for the development of completely new functions. Gazzaniga (2000) suggested that it is processing by the left hemisphere which interprets events and allows one to attempt to find

reason or logic in life events. It may be left-hemisphere processing which is crucial for generating hypotheses to explain inexplicable events. This capacity of the left hemisphere may be what makes a person feel as if they control their own actions. Although the right hemisphere has access to what is going on in the outside world, its role may be to keep track of this information so that the left hemisphere can process and interpret it. Thus the two hemispheres may work synergistically to provide an adaptive system of experience and to interpret a person's environment and internal thoughts. (*See* **Neural Correlates of Consciousness as State and Trait**)

References

Aboitiz F, Scheibel A, Fisher RS and Zaidel E (1992) Fiber composition of the human corpus callosum. *Brain Research* **598**: 143–153.

Bermudez P and Zatorre RJ (2001) Sexual dimorphism in the corpus callosum: methodological considerations in MRI morphometry. *NeuroImage* **13**: 1121–1130.

Bishop KM and Wahlsten D (1997) Sex differences in the human corpus callosum: myth or reality? *Neuroscience and Biobehavioral Reviews* **21**: 581–601.

de Lacoste-Utamsing C and Holloway RL (1982) Sexual dimorphism in the human corpus callosum. *Science* **4553**: 1431–1432.

Gazzaniga MS (2000) Cerebral specialization and interhemispheric communication. Does the corpus callosum enable the human condition? *Brain* **123**: 1293–1326.

Giedd JN, Blumenthal J, Jeffries NO *et al*. (1999) Development of the human corpus callosum during childhood and adolescence: a longitudinal MRI study. *Progress in Neuropsychopharmacology and Biological Psychiatry* **23**: 571–588.

LaMantia A-S and Rakic P (1990) Axon overproduction and elimination in the corpus callosum of the developing rhesus monkey. *Journal of Neuroscience* **10**: 2156–2175.

Mack CM, McGivern RF, Hyde LA and Denenberg VH (1996) Absence of postnatal testosterone fails to demasculinize the male rat's corpus callosum. *Developmental Brain Research* **95**: 252–254.

Sperry RW (1974) Lateral specialization in the surgically separated hemispheres. In: Schmitt FO and Worden FG (eds) *The Neurosciences: Third Study Program*, pp. 5–19. Cambridge, MA: MIT Press.

Witelson SF (1989) Hand and sex differences in the isthmus and genu of the human corpus callosum: a postmortem morphological study. *Brain* **112**: 799–835.

Further Reading

Gazzaniga MS and LeDoux JE (1978) *The Integrated Mind*. New York, NY: Plenum Press.

Leporé F, Ptito M and Jasper HH (eds) (1986) *Two Hemispheres – One Brain: Functions of the Corpus Callosum*. New York, NY: Alan R. Liss.

Rakic P and Yakovlev PI (1968) Development of the corpus callosum and cavum septi in man. *Journal of Comparative Neurology* **132**: 45–72.

Witelson SF and Kigar DL (1988) Anatomical development of the human corpus callosum: implications for individual differences and cognition. In: Molfese DL and Segalowitz SJ (eds) *Developmental Implications of Brain Lateralization*, pp. 35–57. New York, NY: Guilford Press.

Zaidel E and Iacoboni M (eds) *The Parallel Brain: Cognitive Neuroscience of the Corpus Callosum*. Cambridge, MA: MIT Press.

Circadian Rhythms

Introductory article

Ralph E Mistlberger, Simon Fraser University, Burnaby, British Columbia, Canada

CONTENTS

Circadian rhythms are daily (about 24 h) rhythms of behavior, physiology and biochemistry that are controlled by internal clocks. These rhythms are entrained by environmental cues, and modulate cognitive performance.

INTRODUCTION

The rotation of the earth about its axis creates daily cycles of light, temperature, humidity and other geophysical variables that have had a profound impact on the evolution of life. Most living organisms, from single-celled bacteria to fungi, plants and animals, exhibit daily rhythms in their biochemistry, physiology and behavior that mirror the dramatic environmental changes that define the solar day. Some daily rhythms may represent a direct response to environmental stimuli, but most are controlled by one or more internal, 'circadian' clocks (from the Latin *circa*, 'about' and *dies*, 'day'). The primary function of these circadian clocks is to organize and synchronize the organism's cellular and behavioral activities with the outside world. By internalizing the mechanism for rhythmicity, the organism can anticipate and prepare in advance for predictable changes in its environment. Circadian clocks have been further exploited in some species to regulate seasonal reproductive cycles by measuring day length, and to enable certain cognitive operations that require internal representations of time of day.

While daily rhythms in the activity of plants and animals have undoubtedly always been recognized, the existence of endogenous circadian clocks was a matter of contention until the latter half of the twentieth century. The most compelling evidence for these clocks is the observation that daily rhythms persist (free run) when organisms are maintained in environments lacking 24 h time cues, such as in a controlled laboratory, at the poles, or in orbit aboard the space shuttle. The average duration ('period') of one complete cycle under these conditions usually deviates slightly from 24 h, hence the designation 'circadian'. The period varies both within and between species; in temporal isolation, the circadian clock of the typical mouse completes a cycle in less than a day (about 23.5 h), whereas in most humans the clock cycles at a rate slower than one solar day (the most current estimate is about 24.1 h).

Research on circadian rhythms attempts to answer the following fundamental questions. How are circadian rhythms synchronized with the environment? What are the neural and genetic mechanisms of the circadian clock? What are the consequences of circadian rhythmicity for human performance, health and welfare?

ZEITGEBERS AND ENTRAINMENT

According to the terminology used in the study of biological clocks, a rhythm is any process that repeats itself at regular intervals. A device that produces a rhythm can be said to 'oscillate'. An oscillator can be used like an hourglass to measure the passage of time. If the oscillator is synchronized to the solar day, it can also be used as a 24 h 'clock' to recognize local time, analogous to a sundial or digital wristwatch. A circadian clock is therefore an oscillator that has a mechanism by which it is synchronized to the solar day. Without such a mechanism, the circadian clock would generate daily rhythms that would drift in and out of optimal alignment with the environment. The mechanism for synchronization is therefore vital to the adaptive function of circadian clocks.

Circadian rhythms are strongly synchronized to 24 h light–dark (LD) cycles by a process of 'entrainment', defined as phase and period control of one oscillation by another. 'Phase' refers to any discrete point in the cycle, such as the wake-up time within the daily sleep–wake cycle. Thus, when the circadian sleep–wake rhythm is entrained to a

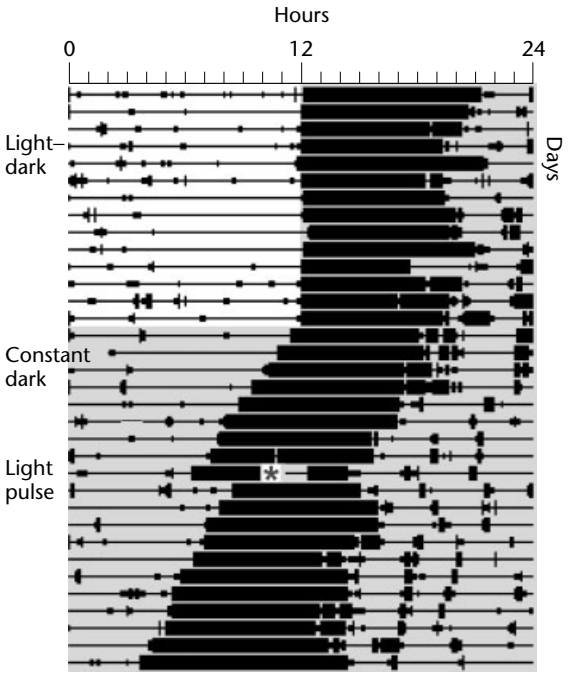


Figure 1. Wheel-running activity of a mouse in a light–dark cycle (12 h dark period indicated by shading), and in constant darkness. Each line represents one day, plotted in 10 min time bins from left to right. Vertical deflections (heavy bars) indicate time bins when wheel running occurred. In the light–dark cycle mouse activity is nocturnal and has a periodicity of exactly 24 h. In constant dark, the activity rhythm 'free runs' with a period of less than 24 h. A 30 min light pulse (asterisk) on day 10 of constant darkness induced a 'delay' phase shift of approximately 2 h.

24 h LD cycle, its period is exactly 24 h, and wake-up time occurs at about the same time within each LD cycle (e.g. when the lights come on in diurnal species, and when the lights go off in nocturnal species). Any stimulus that can entrain another periodic process is a 'zeitgeber' (from the German, 'time-giver').

The mechanism of LD entrainment has been investigated by exposing animals to brief pulses of light (from seconds to hours in duration) during prolonged recordings in constant dark. Light exposure early in the 'subjective night' (that portion of the circadian cycle when the animal acts as if it were night) resets the circadian cycle back to an earlier phase (a 'delay' phase shift; Figure 1). Light exposure late in the subjective night resets the clock forward to a later phase (an 'advance' phase shift). The brighter or longer the light pulse, the larger the shift. Light exposure during most of the 'subjective day' (when the sun would normally be up) typically has little or no effect. Entrainment is accomplished by small, light-induced phase shifts at the dawn and dusk transitions that precisely compensate for the difference between the periodicity of the LD cycle (normally exactly 24 h) and that of the circadian cycle (normally about 24 h). In humans, light during the first half of the night (before the body temperature reaches its minimum, about 2–4 h before the habitual wake-up time) induces delay shifts, while light later in the night induces advance shifts.

Although the LD cycle is the dominant zeitgeber for most species, several nonphotic zeitgebers have also been identified. Daily cycles of temperature can entrain rhythms in some species, although even in plants and poikilothermic animals this does not involve a direct effect of temperature on the rate at which the clock cycles. The biochemical machinery of the circadian clock includes a mechanism for temperature compensation, such that the circadian cycle is relatively constant across a range of tissue temperatures. If this were not the case, the 'clock' of poikilotherms would serve better as a thermometer than as a timepiece.

A number of vertebrate species can be entrained by daily schedules of food intake, and some can also be entrained or phase shifted by scheduled bouts of physical exercise or nonspecific arousal. These behavioral zeitgebers appear to be most effective at times of day when the animal would normally be asleep. The circadian timekeeping system thus appears to have the flexibility to adjust daily rhythms to important nonphotic stimuli that might be of more immediate consequence to survival than is the LD cycle.

NEURAL MECHANISMS

The biological system generating circadian rhythmicity has at its core three integrated components: a self-sustaining circadian clock, an entrainment pathway by which zeitgebers can influence the clock, and one or more output pathways by which the clock confers rhythmicity to other systems (Figure 2). In mammals, the 'master' circadian clock is located in the suprachiasmatic nucleus (SCN) in the hypothalamus. The SCN receives LD information directly from the retina, via a pathway that does not participate in form vision. Damage to the SCN disrupts or eliminates circadian rhythms in all mammals so far studied. Electrical or chemical stimulation of SCN neurons can phase shift circadian rhythms. Individual SCN neurons, grown from embryonic cells in culture, oscillate with a circadian periodicity. Remarkably, transplants of embryonic SCN tissue to adult animals

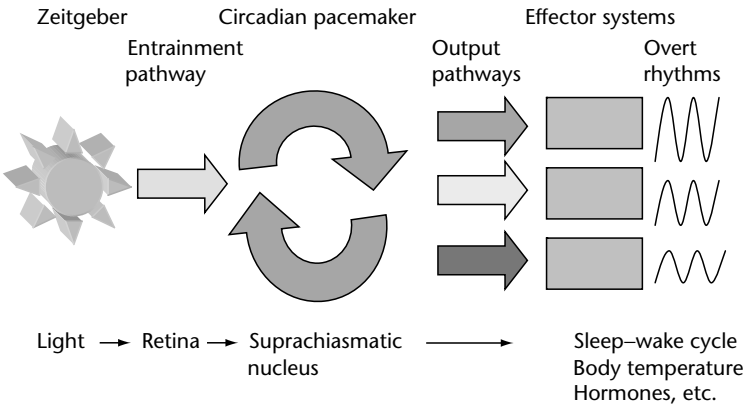


Figure 2. Conceptual model of the core elements of the system generating circadian rhythms in mammals. Overt rhythms may vary in amplitude and phase, but when entrained to light, all exhibit the same 24 h periodicity and a relative stable phase relation with other rhythms. This defines internal temporal order.

can rapidly restore circadian rhythmicity lost by SCN damage. These observations converge in support of a hypothesis that the SCN is the master circadian clock in mammals. This is an exceptional example of the use of complementary techniques to establish the function of a discrete brain structure.

Because of its importance for normal circadian organization of behavior and physiology, the SCN is accorded the status of 'master' oscillator, or 'pacemaker'. However, other circadian oscillators may also exist. One example is the retina, which contains a circadian oscillator that drives daily rhythms of its own local processes, such as photoreceptor disc renewal. In birds and reptiles, the pineal gland, retina and hypothalamus may all contain circadian oscillators, although the role of these as pacemakers or secondary oscillators varies across species.

The SCN receives major inputs from several other brain structures, some of which are important for nonphotic entrainment (e.g. inputs from the thalamic intergeniculate leaflet and pontine raphe nuclei). Neurons of the SCN in turn send axons to a limited number of structures, primarily within the hypothalamus, which presumably disperse circadian timing information more widely. The SCN may also send rhythmic signals by a diffusible factor, conveyed in the extracellular fluid, cerebrospinal fluid or blood.

GENETIC MECHANISMS

Circadian rhythms are genetically programmed and, although sensitive to environmental cycles, do not require these cycles to develop normally. Circadian phenotypes can be selected by breeding, and can be altered by gene mutations. Genetic differences are thought to underlie variability in human rhythms, such as the 'night owl' versus 'early bird' phenotypes.

In all species studied so far, circadian rhythms appear to be generated by an intracellular feedback loop involving a set of 'clock' genes and their protein products. Activation of these clock genes induces the synthesis of proteins, which then feed back upon and temporarily inhibit further clock gene expression. As clock proteins gradually degrade, clock genes are released from inhibition and protein production is renewed. Positive and negative regulators of clock gene expression are articulated in such a way as to produce an approximately 24 h cycle of cellular activity. Zeitgebers shift and entrain the clock by altering clock protein levels. Putative clock genes have been identified and cloned in fungi, fruit flies and mammals, and the feedback principle and some of the specific genes appear highly conserved across a range of phylogenetic groups. A complete molecular description of the circadian 'clockworks' can be expected soon for several species.

CONSEQUENCES OF CIRCADIAN RHYTHMS FOR COGNITION

Consistent with its ubiquitous role in physiological regulation, the circadian timing system also influences performance on cognitive tasks. This is in part secondary to circadian regulation of arousal or alertness, which in LD-entrained humans is low early in the morning and peaks in the early evening, in parallel with the circadian rhythm of body temperature. Performance speed on cognitive tasks that stress simple, repetitive throughput of information, as in tests of vigilance (e.g. detection

of an infrequent signal), serial search, card dealing, additions, and reaction times, follows the same circadian function, with the best scores achieved in the evening. Performance accuracy on some of these tasks varies inversely with body temperature. Tasks stressing short-term memory also generate better scores earlier in the day, before the body temperature reaches its maximum. In addition, performance on some tests of long-term memory is best when participants are tested at the same time of day at which they were trained, even when retest intervals are weeks apart. This implies that circadian time may be incorporated within the neural representations that mediate learning and recall.

The circadian clock has been adapted by many species to provide the sense of time necessary for foraging, navigating and migrating. These species can use timing information provided by the circadian clock to compensate for the movement of the sun and stars, and thereby use the position of these celestial landmarks to guide travel over great distances. Some species, including bees, fish, birds and at least some mammals, can learn and remember the time of day when food is available at specific places in the environment, without the aid of external time cues. According to one theory, circadian time may be encoded within all memories, for use in cognitive computations that guide the temporal aspects of many behaviors.

In humans, the ability to estimate the passage of time is also regulated in part by the circadian system. In temporal isolation, subjective estimates of hourly intervals are proportional to the duration of the circadian sleep–wake cycle. Estimation of intervals in the seconds to minutes range, however, are independent of circadian time.

Disruption of circadian rhythms, induced by transmeridian jet travel, shift-work rotation or a change in laboratory LD cycles, is associated with impaired performance on many cognitive tasks in humans and animals. Some of these effects can be attributed to the partial sleep deprivation that typically accompanies travel and shift rotation in humans, but impairments are also evident in laboratory studies that minimize sleep disruptions. Cognitive deficits that emerge across the life span may also be related to the disruptions of sleep and circadian rhythms that are a physiological hallmark of old age.

CONCLUSION

Circadian rhythmicity is a pervasive feature of life on earth. At all levels of analysis, from molecules to behavior, the daily cycles of the environment are reflected in the functions of the organism. The circadian clocks that have evolved to meet the challenges of the solar day modulate cognitive processes and mediate forms of animal behavior that require precise recognition of time of day or day length.

Further Reading

Aschoff J (ed.) (1981) *Handbook of Behavioral Neurobiology*, vol. 4, Biological Rhythms. New York: Plenum Press.

Aschoff J (1989) Temporal orientation: circadian clocks in animals and humans. *Animal Behaviour* **37**: 881–896.

Brown FM and Graeber RC (1982) *Rhythmic Aspects of Behavior*. London: Lawrence Erlbaum.

Dunlap J (1999) Molecular basis for circadian clocks. *Cell* **96**: 271–290.

Harrington ME and Mistlberger RE (2000) Anatomy and physiology of the mammalian circadian system. In: Kryger MH, Roth T and Dement WC (eds) *Principles and Practise of Sleep Medicine*, 3rd edn. pp. 334–345. Philadelphia: WB Saunders.

Hinton SC and Meck WH (1997) The 'internal clocks' of circadian and interval timing. *Endeavour* **21**: 82–87.

Klein DC, Moore RY and Reppert SM (1991) *Suprachiasmatic Nucleus: The Mind's Clock*. New York: Oxford University Press.

Mistlberger RE (1994) Circadian food anticipatory activity: formal models and physiological mechanisms. *Neuroscience and Biobehavioral Reviews* **18**: 171–195.

Mistlberger RE and Rusak B (2000) Circadian rhythms in mammals: formal properties and environmental influences. In: Kryger MH, Roth T and Dement WC (eds) *Principles and Practise of Sleep Medicine*, 3rd edn. pp. 321–333. Philadelphia: WB Saunders.

Monk T (1994) Circadian rhythms in subjective activation, mood and performance efficiency. In: Kryger MH, Roth T and Dement WC (eds) *Principles and Practise of Sleep Medicine*, 2nd edn. pp. 321–333. Philadelphia: WB Saunders.

Refinetti R (2000) *Circadian Physiology*. Boca Raton: CRC Press.

Takahashi JS, Turek FW and Moore RY (eds) (2001) *Handbook of Behavioral Neurobiology*, vol.12, *Circadian clocks*. New York: Kluwer Academic.

Winfree AT (1987) *The Timing of Biological Clocks*. New York: Scientific American Books.

Color Vision, Neural Basis of

Intermediate article

Bevil R Conway, Harvard Medical School, Boston, Massachusetts, USA
Margaret S Livingstone, Harvard Medical School, Boston, Massachusetts, USA

CONTENTS

The brain determines color from different wavelengths. Specialized cells in the retina, thalamus, primary visual cortex and higher brain areas achieve color perceptions by taking into account chromatic context, which explains various color illusions and color constancy.

INTRODUCTION

A world without color is bleak. Despite this, the benefits of color vision are difficult to quantify. Picasso said, 'When I run out of blue I use red', by which he meant that it is the brightness of a pigment and not its color that describes the three-dimensional shape of objects. Matisse demonstrated this point beautifully in his painting *Femme au Chapeau* (Figure 1). A gray-scale reproduction shows that the values of the pigments do not interfere with an accurate representation of the play of light across his subject's face. That the painting reads well as a face despite the radical color transitions shows that color is not an important cue to shape. In fact, object shapes are easily recognizable even in dim light when color vision is absent. Moreover, many people function perfectly well with impaired color perception: about 1 in 12 men are red–green color-blind and many of them are unaware of it. However, color cues are useful. In monkeys, they assist the discrimination of ripe fruit and of suitable procreative partners, and in humans, color is more than a cue for discriminating objects, for unlike shape and texture, color has emotional significance. One is 'green' with envy, 'red' with anger, 'blue' with sadness. Indeed, Matisse used this to push his portrait past mere representation. Moreover, his picture is much more appealing in color than in black and white. It is probably the emotional quality of color that has fueled color vision research, and it also helps explain the passionate controversies that fill this field's history.

COLOR VISION: WHAT IS IT?

Color vision is the ability to discriminate surfaces based on the spectral content of the light reflected from them, taking into account the light reflected from surrounding objects. Color vision, which has evolved in many animal groups such as insects and birds, is crude in most mammals except some primates, such as humans. Research has focused on color vision in Old World monkeys because their color vision is virtually identical to that in humans (De Valois *et al*., 1974).

THE ROLE OF CONES

In 1802 the English physician Thomas Young proposed that color was subserved by three classes of sensors, each maximally sensitive to a different part of the visible spectrum (Figure 2a). This trichromatic theory, extended by Helmholz in 1866, resolved a profound problem. Though we can see millions of different colors, our retinas simply do not have enough space to accommodate a separate detector for every color at every retinal location, as proposed by Newton. Given that almost all hues can be matched by the combination of three primary colors, and that the number of receptors for color at every retinal location must be small, Young's proposal of three sensors was reasonable. Moreover, it changed the way we think about color: the bottleneck imposed by the small number of sensors implied that color was a neural construction reflecting both physical properties of light and biological properties of photopigments, neurons and networks of neurons. Color is a perception, and not a property of the world.

(a) (b)

Figure 1. [*Figure is also reproduced in color section.*] (a) Henri Matisse, *Femme au Chapeau* (Paris, autumn 1905). Oil on canvas, 80.5 cm × 60 cm; San Francisco Museum of Modern Art (bequest of Elise S. Haas). The gray-scale reproduction (b) shows that the surprising color transitions do not interfere with an accurate representation of the woman's face. Yet the color reproduction is obviously more appealing – why? The dissociation of color and form, clear in this picture, shows that color is processed by the visual system separately from other stimulus attributes, like form.

We now know that color perception begins in the retina with photoreceptors called cones (Figure 2b, c). Cones are more densely packed in the portion of the retina corresponding to the center of gaze (the fovea), and become less dense in the periphery. A second type of photoreceptor, rods, are absent from the fovea. Rods function best in dim light, when the cones do not function well. Because we only have one class of rods, and a comparison between at least two classes of photoreceptor is required for color vision, rods for the most part are not involved in color perception, and we do not see color in very dim light.

Cones are divided into three classes according to their peak absorptions: the S cones absorb shorter wavelengths optimally (peak 440 nm); the M cones absorb middle wavelengths (peak 535 nm) and the L cones absorb long wavelengths (peak 565 nm). All cone classes are somewhat sensitive to wavelengths throughout most of the spectrum (Figure 2b). Thus a single class of cones is color-blind because it cannot distinguish between a dim light of optimal wavelength and an intense light of less optimal wavelength. Moreover, at any given point in the retina there is only one cone, so the retina is color-blind on a spatial scale of single cones. Despite these facts, the cones are often loosely called 'blue', 'green' and 'red', because these names are somehow more intuitive, but we must be cautious because these are not even the color names that we assign to the region of the spectrum to which each class is maximally sensitive. That single cone classes do not code the perception of single colors is proof that the simple trichromatic theory cannot fully explain color perception.

It was once assumed that cones in the primate retina would be regularly distributed (to facilitate uniform sampling of wavelength), as is the case in the goldfish. Primate S cones are distributed fairly regularly (Curcio *et al*., 1991), but L and M cones are surprisingly patchy (Roorda and Williams, 1999) (Figure 2c). The resulting clumpiness may help us detect fine-grained luminance variations, but only at the cost of color resolution. Indeed, variations in color are harder to resolve than variations in luminance (for a review see Livingstone and Hubel, 1987). Two objects that differ only in color are described as equiluminant, and you can find examples of them in the Matisse painting (equiluminant colors will come out roughly the same gray in a gray-scale copy).

The very center of the fovea (0.1° of visual angle) is devoid of S cones. Our eyes are focused for about 550 nm light, where the L and M cones have

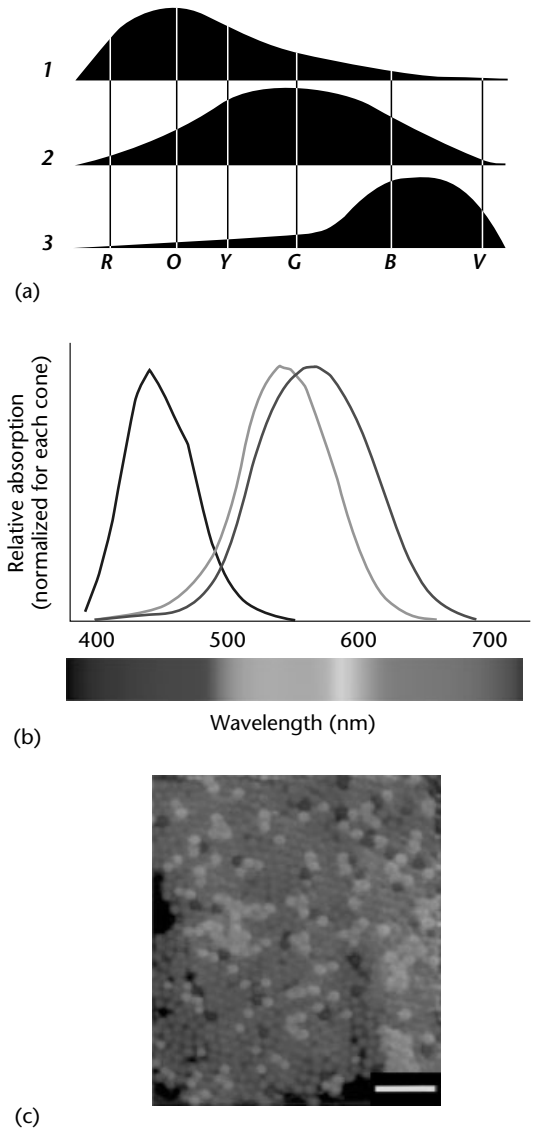


Figure 2. [*Figure is also reproduced in color section.*] Color perception begins in the retina of the eye with three classes of photoreceptors called cones. (a) The absorption spectra of the three detectors proposed by Helmholtz in 1866: shorter wavelengths (V, or violet) were represented on the right. (b) The actual cone absorption spectra of the three cone classes, L, M and S, based on the cone fundamentals of Smith and Pokorny (1972). Convention today puts shorter wavelengths on the left. Below the plot is the visible spectrum. (c) The cone mosaic of a patch of living human retina made visible with adaptive optics, from Roorda and Williams (1999). The S cones are represented by blue, M by green and L by red. Scale bar, 5 arc minutes of visual angle.

their peak sensitivities. Consequently shorter-wavelength light will be blurred. Evolution may have selected against having S cones in the center of the fovea where high spatial acuity is the goal because the short-wavelength light to which the S cones are most responsive will be an unreliable source of spatial information. The absence of S cones in the center of the fovea has surprisingly little impact on color vision, probably because the spatial extent of the S cone hole is finer than the coarse resolution of color vision.

The wavelength sensitivity of a given cone cell is attributed to the specific photopigment protein that it expresses (Nathans, 1999). The M and L photopigment genes, which are encoded on the X chromosome, are fairly similar in sequence, suggesting that the M and L photopigments arose from a common ancestral gene that duplicated not so long ago – around 30–40 million years ago, shortly after the continents of Africa and South America separated. The similarity of the M and L gene sequences predisposes them to recombination during meiosis. This has led to a polymorphism of the L and M photopigments, which is more commonly manifest in males because they only have one X chromosome on which to rely for their M and L photopigments. The polymorphism can be a complete loss of L cones (protanopia), loss of M cones (deuteranopia) or, more frequently, the expression of a mutant M/L hybrid. It is these polymorphisms that underlie the range of so-called red–green color blindness, the most famous case of which is that of Sir John Dalton (a deuteranope), who in 1794 was the first to describe the condition. The deletion of the S cone gene, on the seventh chromosome, is possible (tritanopia), but rare.

COLOR IS AN OPPONENT PROCESS

In 1880 the German psychologist Ewald Hering proposed that color was mediated not by trichromacy but by opponency. Hering observed that we cannot perceive a continuous mixture of colors as predicted by the trichromatic theory – we cannot perceive (or even conceive of) reddish-greens or bluish-yellows. Some colors are mutually exclusive of others. So, Hering argued, color must be determined by the activity of opponent mechanisms, and he proposed three: a red–green mechanism, a blue–yellow mechanism and a black–white mechanism. Today we can appreciate another reason why color must be an opponent process: the L and M cone fundamentals are so similar that to perceive long-wavelength light as distinct from middle-wavelength light (i.e. red and not yellow) the responses of the M cone must be subtracted from those of the L cone.

Each opponent mechanism can be thought of as one axis in a three-dimensional space that

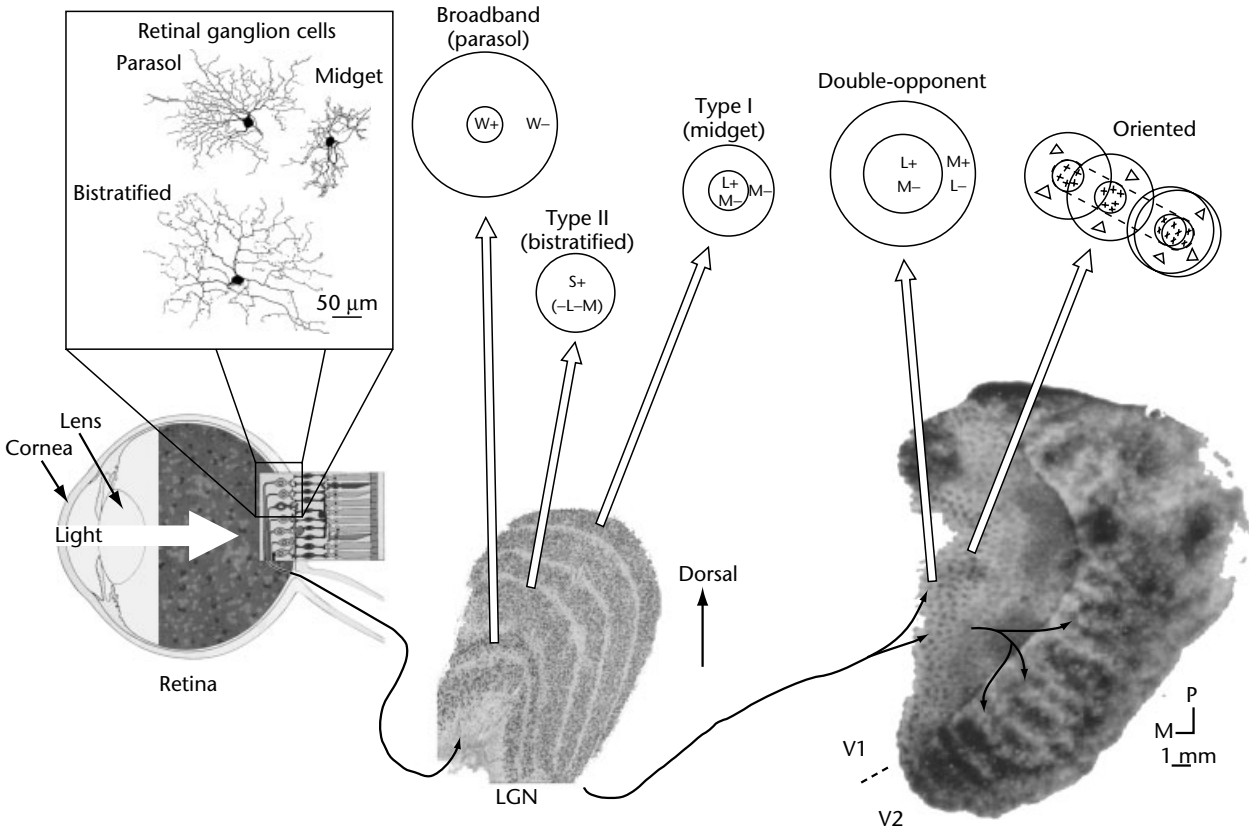


Figure 3. [*Figure is also reproduced in color section.*] A summary of color processing in the visual system. Light enters the eye and is focused on the retina by the cornea and lens. The three classes of cones respond to the light. Different retinal ganglion cells (inset; adapted from Dacey and Lee, 1994) sample the cone mosaic and provide input to the lateral geniculate nucleus (LGN). The retinal ganglion cell names 'midget' and 'parasol' reflect the relative sizes of their dendritic fields, which in turn reflect the relative sizes of their receptive fields. The cells of the LGN, here stained with Nissl substance, comprise six well-defined layers: four dorsal (or parvocellular) layers and two more darkly staining ventral (or magnocellular) layers. Each purple dot is a single cell, about 10 µm in diameter. The parvocellular layers contain type I cells; the receptive field of an L-ON center/M-OFF surround type I cell is given. The magnocellular layers contain the broadband cells. Between the darkly staining parvocellular and magnocellular layers are the koniocellular layers. Type II cells reside in these layers. Broadband cells, type II cells and type I cells are the LGN targets of parasol, bistratified and midget ganglion cells, respectively.

Neurons in the LGN send their axons to the primary visual cortex (V1). In this figure, V1 is represented by a tangential section of one hemisphere of unfolded and flattened squirrel monkey cortex that has been stained with the metabolic enzyme cytochrome oxidase (M, midline; P, posterior). Cytochrome oxidase (CO) staining clearly demarcates the border between V1 and the second visual area, V2, and reveals CO blobs in V1 and the CO stripes in V2. Color information is processed by the double-opponent cells, which reside in the V1 blobs and send their axons to the thin CO stripes of V2 (arrows). Between the blobs are cells that are sensitive to the orientation of a visual stimulus.

encompasses all colors. So any given color could be uniquely defined by three variables: the activity along the red–green axis, the activity along the blue–yellow axis and the activity along the black–white axis. The scientific dispute between Hering and Helmholtz and their supporters is the source of much animosity in the field of color, even today, although studies have now reconciled them.

The retinal ganglion cells, which receive input from the cones, project to neurons in the lateral geniculate nucleus (LGN). These in turn project to neurons in the primary visual cortex (Figure 3). The most common retinal ganglion cells are the midget cells and the parasol cells. Midget cells project to type I cells in the four parvocellular layers of the LGN. Parasol cells project to broadband cells in the two magnocellular layers. Most type I cells receive antagonistic inputs from M and L cones and therefore respond in opposite ways to different colors (DeValois *et al*., 1958; Wiesel and Hubel, 1966; Reid

and Shapley, 1992). A type I cell may be excited by long-wavelength light and suppressed by middle-wavelength light. These cells represent the sort of building block for Hering's red–green opponent process; the three cone types represent Young and Helmholtz's trichromacy.

A third type of retinal ganglion cell is the bistratified cell (Dacey and Lee, 1994). Bistratified cells are excited by S cones and suppressed by a mixture of L and M cones, making them likely candidates for Hering's blue–yellow mechanism. Bistratified cells project to the (S versus M + L) type II cells in the koniocellular layers of the LGN (Figure 3).

Black–white cells must contribute to our perception of color. For example, the addition of black changes the color of orange to brown. Many parvocellular cells do not show cone opponency, and could therefore represent black–white, but it is unclear whether these or the broadband cells of the magnocellular layers (or an unidentified class of cells) underlie the black–white color axis (Wiesel and Hubel, 1966).

COLOR CONSTANCY

Perhaps the greatest misperception about color is that it is equated to wavelength. This misperception is cultivated early in our education when we are taught (incorrectly!) that long-wavelength light is 'red' and short-wavelength light is 'blue'. Though wavelength is the critical determinant of color, it is not the only determinant. For example, our perception of white depends on responses from all three cone classes, which is normally achieved when we see broadband light. However, after viewing a colored surface, one class of cones may become fatigued, perhaps by bleaching of the photopigment, and a previously 'white' surface will appear as a colored afterimage (Figure 4a).

Chromatic context also affects our perception of color. This is well known by artists, who place red against a green background to make it redder. Another example where wavelengths do not correlate directly with color is the phenomenon of induced colors: a gray spot can be made to appear colored if it is surrounded by a large colored annulus (the larger the annulus and the less sharp the boundaries, the stronger the effect). In fact spatial context even colors afterimages (Figure 4b). The discrepancies between physical cues and our perceptions can leave us confused. However, as Edwin Land pointed out, one should avoid asking the question, 'What color is it really?' as if to imply that our visual systems are deceiving us. Color is a product of our physiology and its interaction with the physical world; visual illusions simply point out what our visual systems are constantly (and usually effectively) doing.

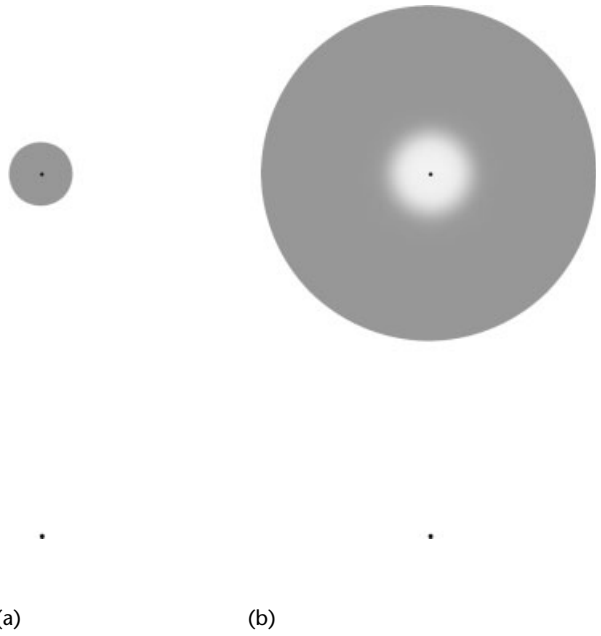


Figure 4. [*Figure is also reproduced in color section.*] There is more to color than meets the eye! Stare at the fixation dot in the middle of the green disk (a), being careful to hold your gaze steady. After 20 s or so, transfer your gaze to the fixation dot below; you should see a reddish afterimage. Now consider the small, fuzzy gray disk in the center of the colored annulus (b). After prolonged viewing the gray seems to adopt a weak reddish tinge. Such induced colors are much more striking when the colored annulus occupies the entire visual field surrounding the central gray spot. Try generating an afterimage to the gray spot. The afterimage to the gray spot is surprisingly green! This shows that the spatial configuration of a scene affects both the color of perceived images and the color of afterimages.

An object will appear colored if it selectively absorbs some wavelengths and reflects others. The spectral distribution of reflected light is a product of the absorptive properties of the object's surface and the spectral properties of the light source (the illuminant). So if the illuminant changes, the reflected light will change too. Illuminants are constantly changing. A bright sunny day, under a blue sky, will contain a large proportion of shorter wavelengths, while a tungsten light bulb will produce longer wavelengths. The paradox of color vision is that despite these different illumination conditions, and the resulting difference in reflected light, the color of objects is fairly constant. A red apple is red, for example. It is not that a red apple is red only when viewed under a certain illumination

condition such as a blue sky. This color constancy is mostly a property of our visual systems and not a function of memory.

It is easy to see why our visual systems have evolved in this way. If we were to assign a color to an object based solely on the light reflected from it then we would assign different colors to the same object depending on the conditions under which the object was viewed. Color constancy means that colors are properties of objects (which are constant) and not viewing conditions (which are continually changing).

Edwin Land, the inventor of instant photography, went to great lengths to reiterate that the color we assign an object is largely independent of the spectral content of the illuminant but is correlated with the absorption properties of the object or surface (Land, 1977). He was prompted by the familiar problem faced by color photographers: a scene photographed under tungsten light comes out with a reddish cast, and one photographed outside on a sunny day with a bluish cast. This is in contrast to our perceptions of the scene under the two illumination conditions – we see neither a reddish cast nor a bluish cast. Land concluded that our visual systems do not simply equate color and reflected wavelengths. In his experiments, Land used different light sources to illuminate different patches of a colored 'Mondrian' display, and varied the spectral content of his illuminants. He was able to show that two differently pigmented patches could be made to reflect the same spectral distribution and yet, remarkably, the patches still appeared as different colors. He also showed that an identical surface could be made to reflect a different spectral distribution and yet appear the same color.

The puzzle remained. How could the visual system achieve different color judgments for two areas if the light from the two areas were the same? Land devised the retinex algorithm, which is capable of determining illuminant-independent colors (Land, 1977). According to this algorithm, the critical determinant of the color of a surface is the chromatic context in which the surface appears. This might sound like a reiteration of what artists already knew empirically, but it went further. It provided a testable hypothesis about the visual system. It claimed that color is determined by abrupt changes in the relative cone activities across a scene. For example, retinex would identify a region as 'red' only when the long wavelength light reflected from it is surrounded by regions reflecting shorter wavelengths. Thus we would not expect to 'see' the reddish cast of a tungsten light because the cast is diffuse.

DOUBLE-OPPONENT COLOR CELLS

The cone-opponent retinal ganglion and LGN cells could subserve wavelength discrimination. However, color is not simply wavelength discrimination (see above). Rather, color is achieved through a spatial comparison of wavelengths across an image. A cell having a receptive field fed by a single cone class in a spatially opponent fashion might be the building block for such a comparison: for example, a cell excited by L cones in one part of visual space but suppressed by L cones in an adjacent part of visual space (an L-ON center/L-OFF surround cell); but despite intensive searches, no retinal ganglion cells or LGN cells like this have been found. The cone-opponent type I cells (see Figure 3) have spatially opponent receptive fields, but the centers and surrounds are fed by different cone classes, and the opponency is in the wrong direction to subserve color constancy. So where in the primate visual system is such a comparison made?

In 1968, Nigel Daw showed that some cells in goldfish retina have receptive fields that are both chromatically and spatially opponent, and therefore capable of computing simultaneous color contrast (Daw, 1968). Computational studies have shown that such 'double-opponent' cells could subserve color constancy in primates: a single double-opponent cell exceeds the requirement of a spatial comparison for one cone class: it is a spatial comparison for two cone classes. A common type of double-opponent cell, for example, is L-ON center/L-OFF surround and M-OFF center/M-ON surround.

The existence of double-opponent cells in the primary visual cortex (V1) of primates has been controversial, but there is now a consensus that they do exist (Conway, 2001; Johnson *et al*., 2001) (Figure 5). Cortical cells that show simple chromatic opponency (such as LGN type II cells) also exist, although they probably do not represent a distinct cell class but rather the end of a continuum of cone-opponent cells that show very weak surrounds.

In addition to mediating spatial color contrast, double-opponent cells may also play an important part in temporal color contrast (a red spot is redder if preceded by a green spot) because they respond to both the onset and cessation of a stimulus (Cottaris and DeValois, 1998) and show stronger responses to sequences of oppositely colored stimuli, e.g. green and then red (Conway *et al*., 2002).

Cortical cone-opponent cells represent about 10% of the total population of cells in V1. Both red–green double-opponent cells – i.e. L versus M

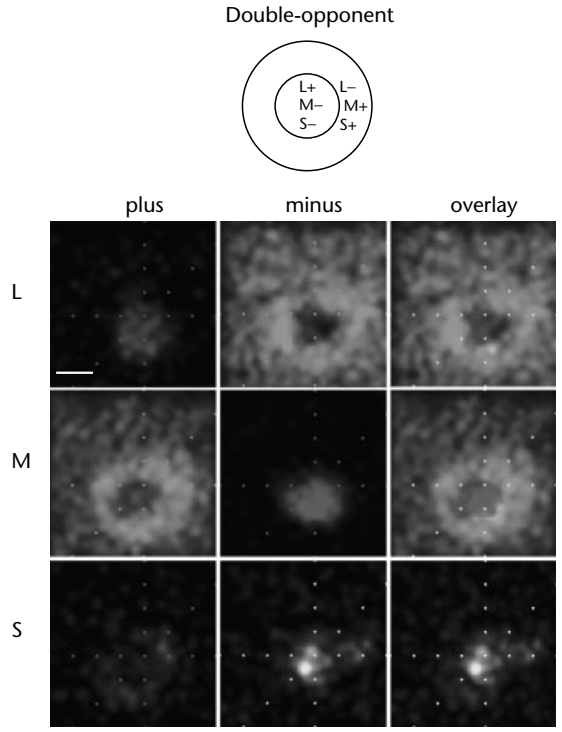


Figure 5. [*Figure is also reproduced in color section.*] The receptive field of a double-opponent cell in monkey V1. The left-hand column shows the spatial extent of the cell's response to increasing activity of the three cone classes (L, top; M, middle; S, bottom); the middle column shows the same cell's response to decreasing the activity of the three cone classes. Comparing the maps (right-hand column) shows that this cell's receptive field is both spatially and chromatically opponent. This double-opponent structure is critical to color constancy – our ability to determine an object's color despite changing illumination conditions. From Conway (2001).

– and blue–yellow double-opponent cells – i.e. S versus (L + M) – are found, and these, with a class of opponent achromatic cells, could be the sole basis for color perception despite their relative scarcity, because color perception is coarse and would require many fewer cells than (say) form perception. Perhaps not by coincidence, color in color televisions requires only about 10% of the bandwidth.

Curiously, some red–green cells appear to receive S cone input. This needs to be studied further. It also remains to be shown how double-opponent cells are constructed from the LGN inputs and why there are more red–green double-opponent cells than blue–yellow ones.

Double-opponent cells reside in the cortex in clusters; moreover, these clusters are coarsely localized in metabolically distinct regions of cortex called 'blobs' (Figure 3, bottom right) (Livingstone and Hubel, 1984). Blobs are easily identified in primate visual cortex by staining with the metabolic enzyme cytochrome oxidase. Why double-opponent cells are localized to the cytochrome oxidase blobs remains a mystery – perhaps these cells require more energy and therefore express higher levels of this metabolic enzyme. Surprisingly, other animals with poor color vision have cytochrome oxidase blobs (although the blobs in some of these mammals, such as cats, are not as prominent). Thus, it may be that the blobs represent regions of cortex dedicated to a more generic function, like parsing surfaces, of which color is only one component.

Nevertheless, the segregation of color cells in the LGN (in the parvocellular and koniocellular layers) and in the primary visual cortex (in the cytochrome oxidase blobs) shows that color is largely processed separately from other visual attributes. This segregation of color processing is evident perceptually (see below) and was used to advantage by Matisse (see Figure 1).

COLOR IS PROCESSED SEPARATELY FROM FORM AND MOTION

Color signals carried by the blobs of V1 are relayed to V2 and then to higher visual areas. Like V1, V2 displays an interesting pattern of staining for the enzyme cytochrome oxidase; unlike V1, the staining consists of alternating thick and thin stripes separated by interstripes (see Figure 3). Cells residing in the blobs of V1 send their axons to the thin stripes of V2 (Livingstone and Hubel, 1984). Not surprisingly, cells in the thin stripes are more likely to be color selective than cells in the thick stripes. The color cells in the V2 thin stripes respond best to colored spots, but they do so over a larger area of visual space. They are not responsive to a large field of color that encompasses the entire region over which small spots are effective. Such 'complex' color cells may be useful in identifying color boundaries present anywhere within a large area.

Color signals carried by cells in V1 and V2 are relayed to subsequent areas where, presumably, color percepts are elaborated. The V2 thin stripes project to V4; the V2 thick stripes, on the other hand, project to the middle temporal area, an area specialized for analyzing motion. Many V4 neurons respond better to some wavelengths than to others (Zeki, 1983), suggesting they are involved in color vision. Their receptive fields are much larger than the receptive fields of V1 cells, suggesting they are involved in elaborating color constancy.

Extrastriate visual areas are better described in the macaque monkey than in the human, but it becomes difficult to compare the areas of humans and monkeys the further the areas are from V1. In humans, for example, an area (V8) situated in the inferotemporal cortex is specialized for computing color (Hadjikhani *et al.*, 1998); it is debated if monkeys have a homolog of this area. Perhaps V4 is the monkey equivalent of human V8. Neurons in V4 are color biased (Zeki, 1980) but they are also selective for other attributes of a stimulus, such as the stimulus orientation (Schein and Desimone, 1990), suggesting that V4 is not simply a color area. To complicate the matter further, unlike lesions of V8 in humans, partial lesions of V4 in monkeys have little effect on tasks requiring color vision. Perhaps V4 is involved in piecing together form and color information; conversely, V4 may actually be a complex of areas, one devoted to form processing and another to color processing. The increasing resolution of functional magnetic resolution imaging may soon make it possible to address these issues.

Certain people who have damage to V8 following a stroke show a profound loss of color perception. Remarkably, this acquired achromatopsia does not interfere with their perception of form and motion – further suggesting that color is processed separately from other visual attributes. Oliver Sacks described one such stroke patient who was an artist (Sacks, 1995). After the patient had lost his color vision, he made peculiar color choices in his paintings; but he was still able to represent luminance and shape, because the areas of his brain dedicated to processing those aspects of the visual world were unaffected by the stroke. Moreover, he had no loss of motion perception.

CONCLUSION

An observer would say that the color and the shape of an object are inextricably linked. If color and form are processed separately by the cortex, how do they then become bound? How would this binding be manifest in the brain? The binding might simply be found in the correlated activity of the two pathways: an orange ball would produce separate sensations of 'round' in the form pathway and 'orange' in the color pathway, but the sensations would be elicited simultaneously. The reliability of the simultaneous activation of the two pathways, possibly reinforced by neural connections joining separate areas, could be enough to bind the 'round' ball with its 'orange' color. But how to test this hypothesis, or any hypothesis for binding for that matter, is challenging.

Parallel processing is an efficient means of computing information because it is fast: multiple aspects of a scene can be processed simultaneously. Although many lines of evidence support a parallel processing model for visual perception (Livingstone and Hubel, 1988), some anatomical and physiological studies show that the visual system does not operate according to such a simple model. There are cells in the superficial layers of V1 (not confined to blobs) that are both selective for stimulus orientation and also more responsive to some wavelengths than others. What contribution do such 'oriented color' cells make to color perception? What are the LGN inputs to red–green double-opponent cells? Do the same LGN cells provide input to sharply orientation-tuned cells? How is the activity of the three chromatic axes in V1 integrated to bring about the perception of specific hues, and how is hue then bound with form? Perhaps most compelling, how do colors bring about emotional responses? These and many more questions will keep the field of color vision research alive for many years to come.

References

Conway BR (2001) Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1). *Journal of Neuroscience* **21**: 2768–2783.

Conway BR, Hubel DH and Livingstone MS (2002) Color contrast in macaque V-1. *Cerebral Cortex* **12**: 915–925.

Cottaris NP and De Valois RL (1998) Temporal dynamics of chromatic tuning in macaque primary visual cortex. *Nature* **395**: 896–900.

Curcio CA, Allen KA, Sloan KR *et al*. (1991) Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *Journal of Comparative Neurology* **312**: 610–624.

Dacey DM and Lee BB (1994) The 'blue-on' opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. *Nature* **367**: 731–735.

Daw N (1968) Goldfish retina: organization for simultaneous color contrast. *Science* **158**: 942–944.

De Valois RL, Smith CJ, Kitai ST and Karoly AJ (1958) Response of single cells in monkey lateral geniculate nucleus to monochromatic light. *Science* **127**: 238–239.

De Valois RL, Morgan HC, Polson MC, Mead WR and Hull EM (1974) Psychophysical studies of monkey vision. I. Macaque luminosity and color vision tests. *Vision Research* **14**: 53–67.

Hadjikhani N, Liu AK, Dale AM, Cavanagh P and Tootell RB (1998) Retinotopy and color sensitivity in human visual cortical area V8. *Nature Neuroscience* **1**: 235–241.

Johnson EN, Hawken MJ and Shapley RM (2001) The spatial transformation of color in the primary visual

cortex of the macaque monkey. *Nature Neuroscience* **4**: 409–416.
Land EH (1977) The retinex theory of color vision. *Scientific American* **237**: 108–128.
Livingstone MS and Hubel DH (1984) Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience* **4**: 309–356.
Livingstone MS and Hubel DH (1987) Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience* **7**: 3416–3468.
Livingstone M and Hubel D (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240**: 740–749.
Nathans J (1999) The evolution and physiology of human color vision: insights from molecular genetic studies of visual pigments. *Neuron* **24**: 299–312.
Reid RC and Shapley RM (1992) Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature* **356**: 716–718.
Roorda A and Williams DR (1999) The arrangement of the three cone classes in the living human eye. *Nature* **397**: 520–522.
Sacks O (1995) *An Anthropologist on Mars*. New York, NY: Knopf.
Schein SJ and Desimone R (1990) Spectral properties of V4 neurons in the macaque. *Journal of Neuroscience* **10**: 3369–3389.
Smith and Pokorny (1972) Spectral sensitivity of color-blind observers and the cone photopigments. *Vision Research* **12**: 2059–2071.
Wiesel TN and Hubel DH (1966) Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *Journal of Neurophysiology* **29**: 1115–1156.
Zeki S (1980) The representation of colours in the cerebral cortex. *Nature* **284**: 412–418.
Zeki S (1983) The relationship between wavelength and color studied in single cells of monkey striate cortex. *Progress in Brain Research* **58**: 219–227.

Further Reading

Conway BR (2002) *Neural Mechanisms of Color Vision*. Boston: Kluwer.
Gegenfurtner KR and Sharpe LT (1999) *Color Vision: From Genes to Perception*. Cambridge, UK: Cambridge University Press.
Hubel DH (1995) *Eye, Brain and Vision*. New York, NY: Scientific American Library.
Hurvich LM (1981) *Color Vision*. Sunderland, MA: Sinauer.
Livingstone MS (2002) *Vision and Art: The Biology of Seeing*. New York, NY: Abrams.
Zeki S (1993) *A Vision of the Brain*. Cambridge, MA: Blackwell.

Computational Neuroscience: From Biology to Cognition

Intermediate article

Randall C O'Reilly, University of Colorado, Boulder, Colorado, USA
Yuko Munakata, University of Denver, Colorado, USA

CONTENTS

Computational neuroscience involves the construction of explicit computational models that implement neural mechanisms to simulate cognitive functions such as perception, learning and memory, motor function, and language.

INTRODUCTION

This article describes computer models that simulate the neural networks of the brain, with the goal of understanding how cognitive functions (perception, memory, thinking, language, etc.) arise from their neural basis. Many neural network models have been developed over the years, focused at many different levels of analysis, from engineering, to low-level biology, to cognition. Here, we consider models that try to bridge the gap between biology and cognition. Such models deal with real cognitive data, using mechanisms that are related to the underlying biology.

THE RELATIONSHIP BETWEEN COGNITIVE AND NEURAL THEORIES

Computational models provide an important tool for linking data across multiple levels of analysis. The cognitive implications of cellular and network properties of neurons are often not immediately apparent: there are too many factors at many different levels interacting in complex ways. Trying to develop behavioral predictions that capture the complexity of the neural level can be like trying to predict the weather from a number of satellite measurements. A computational model, of the weather or of the brain, can help by formalizing information and relating it through complex, emergent dynamics. Cognitive properties can thus be understood as the product of a number of lower-level interactions, and neural properties can be understood in terms of their functional role in cognitive processes. Further, the effects of manipulations of lower-level interactions (e.g. through genetic knockouts or lesions) can be simulated and reconciled with the observed behavioral effects. Importantly, these simulations can make sense of much more subtle behavioral effects than the generic impairment of behavior on a cognitive task.

Although models thus have the potential to clarify brain–behavior relations, they do not always do so. Models can be underconstrained by neural and behavioral data, and thus be of questionable value in showing how the brain actually subserves behavior. Moreover, models may be devised merely as demonstrations that a behavior can be simulated, but this is insufficient for understanding why the models behave as they do. Thus, models must be evaluated in a balanced way for whether they advance understanding of specific phenomena, provide general principles, and make useful links between brain and behavior.

This article reviews a number of neuroscience-based computational models of various cognitive phenomena, with an emphasis on the general principles embodied by these models and their implications for understanding the general nature of cognition. Specifically, we examine models of: vision, including topography and receptive fields

in primary visual cortex and spatial attention emerging from interactions between parietal and temporal streams of processing; episodic memory subserved by the hippocampus; conditioning and skill learning subserved by the basal ganglia and cerebellum; working memory and cognitive control subserved by the prefrontal cortex; and language processing guided by neuropsychological cases. For a more comprehensive treatment of many of these models and the ideas behind them, see O'Reilly and Munakata (2000).

COMPUTATIONAL MODELS OF VISION GUIDED BY NEUROSCIENCE

Vision is one of the best-studied domains in cognitive neuroscience, having a long tradition of integrating biological and psychophysical levels of analysis. Computational models of vision have been influential in both the vision and computational research communities. We review two areas of visual modeling here: topography and receptive fields in primary visual cortex; and spatial attention and the effects of parietal lobe damage. Other major areas of visual processing that have been modeled include object recognition, motion processing, and figure–ground segmentation.

Topography and Receptive Fields in Primary Visual Cortex

The primary visual cortex (V1) provides an interesting target for computational models, because it has a complex but relatively well-understood organization of visual feature detectors (a 'representational structure') subject to considerable experience-based developmental plasticity (Hubel and Wiesel, 1962; Gilbert, 1996). Thus, the major question behind many of the V1 models has been: can we reproduce the complex representational structure of V1 through principled learning mechanisms exposed to realistic visual inputs?

First, we summarize the complex representational structure of V1. V1 neurons are generally described as edge detectors, where an edge is simply a roughly linear separation between regions of relative light and dark. These detectors differ in their orientation, size, position, and *polarity* (i.e. whether they detect transitions from light to dark or dark to light, or dark–light–dark or light–dark–light). The different types of edge detectors (together with other neurons that appear to encode visual surface properties) are packed into the two-dimensional sheet of the visual cortex according to a particular topographic organization. The large-scale organization is a 'retinotopic map', which preserves the topography of the retinal image in the cortical sheet. At the smaller scale are 'hypercolumns' (see Figure 1), containing smoothly varying progressions of oriented edge detectors, among other things (Livingstone and Hubel, 1988). The hypercolumn also contains 'ocular dominance columns', in which V1 neurons respond preferentially to input from one eye or the other.

Many computational models have emphasized only one or a few aspects of the many detailed properties of V1 representations; for reviews, see Swindale (1996) and Erwin *et al*. (1995). For example, models have demonstrated how ocular dominance columns can develop based on a Hebbian learning mechanism, with greater local correlations in the neural firing coming from within one eye than from across eyes (Miller *et al*., 1989). Hebbian learning encodes correlational structure by strengthening the weights between neurons that fire together, and decreasing the weights between those that do not. (See Oja (1982) and Linsker (1988) for mathematical analyses of Hebbian correlational learning.)

Several models have demonstrated how a realistic set of oriented edge-detector representations can develop in networks presented with natural visual scenes, preprocessed in a manner consistent with the contrast-enhancement properties of the retina (e.g. Olshausen and Field, 1996; Bell and Sejnowski, 1997; van Hateren and van der Schaaff, 1997; O'Reilly and Munakata, 2000). The Olshausen and Field (1996) model demonstrated that sparse

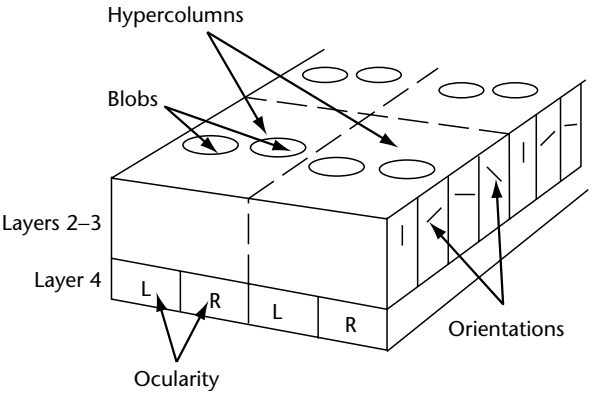


Figure 1. Structure of a cortical hypercolumn, representing a full range of orientations (in layers 2–3), ocular dominance columns (in layer 4, one for each eye), and surface features (in the blobs). Each such hypercolumn is focused within one region of retinal space, and neighboring hypercolumns represent neighboring regions.

representations (with relatively few active neurons) provide a useful basis for encoding real-world (visual) environments, but this model was not based on known biological principles. Subsequent work has shown how biologically-based models can develop oriented receptive fields, through a Hebbian learning mechanism with sparseness constraints in the form of inhibitory competition between neurons (a known property of cortex) (O'Reilly and Munakata, 2000). Furthermore, lateral excitatory connections within this network (another known property of cortex) produced a topographic organization consistent with several aspects of the hypercolumn structure (e.g. gradients of orientation, size, polarity, and phase tuning and pinwheel discontinuities: see Figure 2).

To summarize, these V1 models demonstrate how Hebbian learning mechanisms exposed to naturalistic stimuli, with certain kinds of biological prestructuring (e.g. connectivity patterns and inhibition), can produce aspects of the observed representational structure of V1. However, many complex aspects of early visual processing remain to be addressed, including motion, texture, and color sensitivity of different populations of V1 neurons.

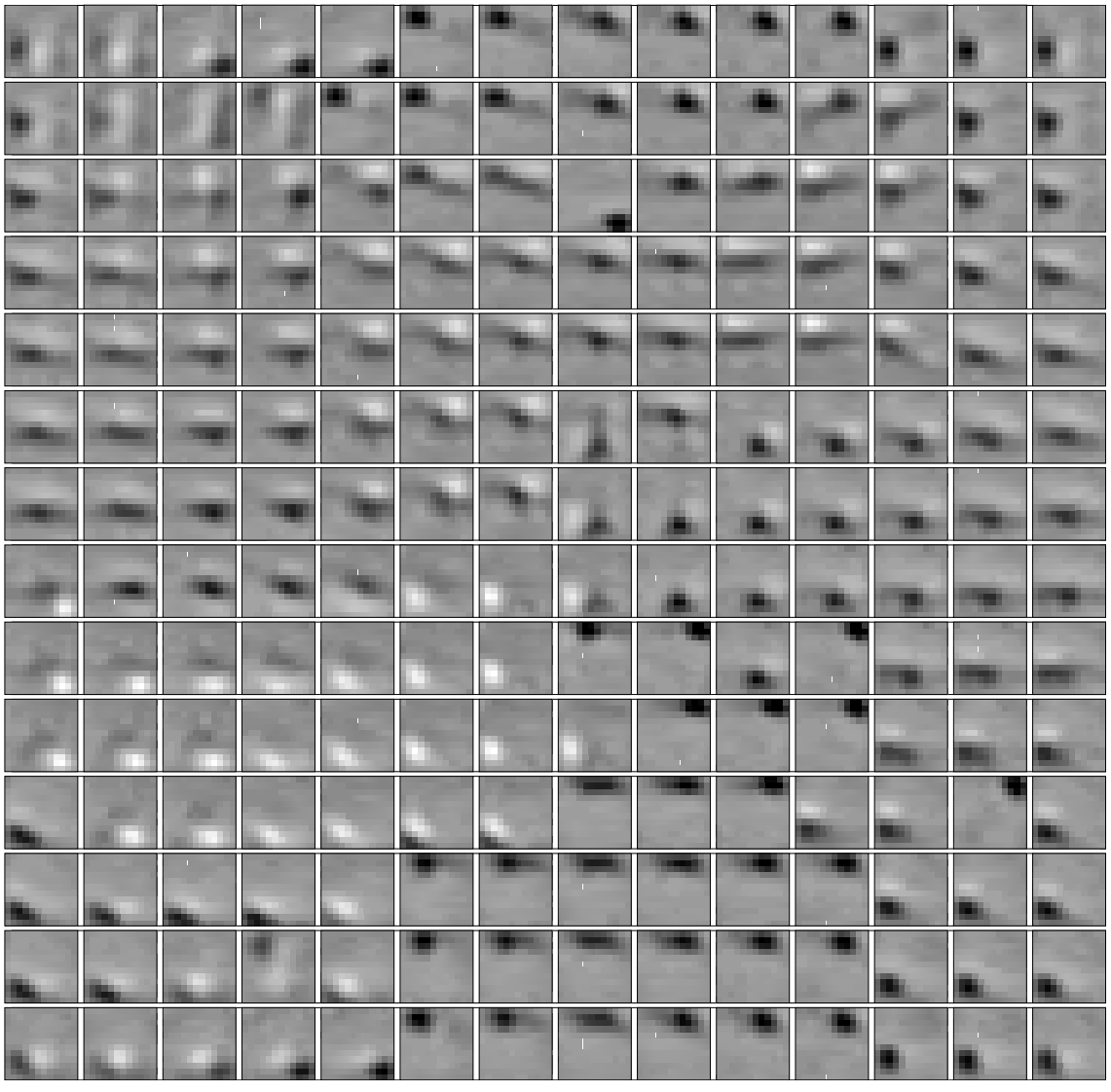


Figure 2. The receptive fields of model V1 neurons (O'Reilly and Munakata, 2000). Lighter shades indicate areas of on-center response, and darker shades indicate areas of off-center response. Individual units are shown by smaller grids (showing weights into those units from different locations in the retinally-organized input). These are organized into a larger grid representing the location of each unit within the simulated V1 hypercolumn.

Spatial Attention and the Effects of Parietal Lobe Damage

Many computational models of higher-level vision have explored object recognition (e.g. Mozer, 1991; Fukushima, 1988; LeCun *et al*., 1989) and spatial processing (e.g. Pouget and Sejnowski, 1997; Mozer and Sitton, 1998; Vecera and O'Reilly, 1998). Here we describe a model of spatial attention (Cohen *et al*., 1994) that demonstrates how biologically-based computational models can provide alternative interpretations of cognitive phenomena. Spatial attention has traditionally been operationalized according to the Posner spatial cuing task (Posner *et al*., 1984: see Figure 3). When attention is drawn, or cued, to one region of space, participants are faster to detect a target in that region (a validly cued trial) than a target elsewhere (an invalidly cued trial). Patients with damage to the parietal lobe have particular difficulty with invalidly cued trials.

According to the standard account of these data, spatial attention involves a 'disengage' module associated with the parietal lobe (Posner *et al*., 1984). This module typically allows one to disengage from an attended location to attend elsewhere. This process of disengaging takes time; hence the slower detection of targets in unattended locations. Further, the disengage module is impaired with parietal damage, so that patients have difficulty disengaging from attention drawn to one side of the space.

Biologically-based computational models, based on recurrent excitatory connections and competitive inhibitory connections, provide an alternative explanation for these phenomena (Cohen *et al*., 1994; O'Reilly and Munakata, 2000). In this framework, the facilitative effects of drawing attention to one region of space result from excitatory connections between spatial and other representations of that region: this excitatory support makes it easier to process information in that region. The slowing observed in the invalid trials results from inhibitory competition between different spatial regions. Under this model, damage to the parietal lobe simply impairs the ability of the corresponding region in space to have sufficient excitatory support to compete effectively with other regions.

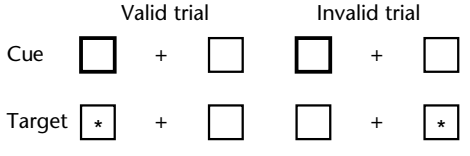


Figure 3. The Posner spatial attention task. The cue is a brightening or highlighting of one of the boxes, which focuses attention on that region of space. Reaction times to detect the target are faster when this cue is valid (the target appears in that region) than when it is invalid (the target appears elsewhere).

The two models make distinct predictions (Cohen *et al*., 1994; O'Reilly and Munakata, 2000). For example, following bilateral parietal damage, the disengage model predicts disengage deficits on both sides of space (Posner *et al*., 1984), but the competitive inhibition model predicts reduced attentional effects (smaller valid and invalid trial effects). Data support the latter model (Coslett and Saffran, 1991; Verfaellie *et al*., 1990), demonstrating the utility of biologically-based computational models for alternative theories of cognitive phenomena.

COMPUTATIONAL MODELS OF EPISODIC MEMORY AND THE HIPPOCAMPUS

Damage to the hippocampus, in the medial temporal lobe, can produce severe memory deficits, while leaving unimpaired certain kinds of learning and memory (Scoville and Milner, 1957; Squire, 1992). Many computational models have been developed to explore the exact contribution of the hippocampus, and these models have had a major influence (e.g., Marr, 1971; Treves and Rolls, 1994; Hasselmo and Wyble, 1997; Moll and Miikkulainen, 1997; Alvarez and Squire, 1994; Levy, 1989; Burgess *et al*., 1994; Samsonovich and McNaughton, 1997).

One framework has combined known biological features of the hippocampal formation with computationally motivated principles about learning and memory to further clarify the unique contributions of the hippocampus in memory (McClelland *et al*., 1995; O'Reilly and Rudy, 2000, 2001; O'Reilly and McClelland, 1994; O'Reilly *et al*., 1998). The central idea is that there are two basic types of learning that an organism must engage in – learning about specifics and learning about generalities – and that because the computational mechanisms for achieving these types of learning are in direct conflict, the brain has evolved two separate brain structures to achieve them. The hippocampus appears to be specialized for learning about specifics, while the neocortex is good at extracting generalities.

Learning about specifics requires keeping representations separated (to avoid interference), whereas learning about generalities requires overlapping representations that encode shared

structure across many different experiences. Furthermore, learning about generalities requires a slow learning rate to gradually integrate new information with existing knowledge, while learning about specifics can occur rapidly. This rapid learning is particularly important for episodic memory, where the goal is to encode the details of specific events as they unfold.

These computational principles provide a satisfying and precise characterization of the division of labor between the hippocampus and the neocortex. The models that implement these principles have been shown to account for a wide range of specific learning and memory findings, including nonlinear discrimination, incidental conjunctive encoding, fear conditioning, and transitive inference in rats (O'Reilly and Rudy, 2001) and human recognition memory (O'Reilly *et al*., 1998). However, these models fail to incorporate important aspects of the hippocampal formation (e.g. the subiculum and the mossy cells in the hilus), and many more complex behaviors that depend on the hippocampus (and its interactions with other brain areas) remain to be addressed.

COMPUTATIONAL MODELS OF CONDITIONING AND SKILL LEARNING IN THE BASAL GANGLIA AND CEREBELLUM

A convergence between biological, behavioral and computational approaches has been achieved in the domain of conditioning (learning to associate stimuli and actions with rewards). In the computational domain, reinforcement learning mechanisms can change the behavior of a simulated animal according to reward contingencies in the environment (Sutton and Barto, 1998). Such learning mechanisms, including the 'temporal differences' algorithm (Sutton, 1988), not only work well mathematically (e.g. Dayan, 1992), but also correspond with aspects of neural recordings made in the reward-processing area of the brain (Montague *et al*., 1996; Schultz *et al*., 1997).

Specifically, a straightforward neural implementation of the temporal differences algorithm involves a systematic transition of reward-related neural firing similar to that observed in dopamine neurons in the midbrain. During a simple conditioning task where a sensory stimulus (e.g. a tone) reliably predicts a subsequent reward (e.g. orange juice), these neurons initially fire in response to the reward, but then after some trials of learning they respond to the sensory stimulus that predicts the reward and no longer to the reward itself (Schultz *et al*., 1993; Schultz *et al*., 1995). This transfer of reward-related firing from the actual reward to predictors of the reward is an essential property of the temporal differences mechanism as implemented by Montague *et al*. (1996), which thus provides a principled, provably effective explanation for why the brain appears to learn in this manner.

Models of motor performance and skill learning have been developed based on the biological properties of the relevant underlying brain areas, including the basal ganglia (which includes the striatum, globus pallidus, substantia nigra, subthalamic nucleus, and nucleus accumbens) and the cerebellum (e.g. Beiser *et al*., 1997; Wickens, 1997; Houk *et al*., 1995; Berns and Sejnowski, 1996; Schweighofer *et al*., 1998a, b; Contreras-Vidal *et al*., 1997). These models accord well with detailed neural properties of these areas, but tend to focus on simpler aspects of motor performance: complex motor skills remain to be addressed.

COMPUTATIONAL MODELS OF WORKING MEMORY, COGNITIVE CONTROL, AND PREFRONTAL CORTEX

The prefrontal cortex is important for a range of cognitive functions, which can be described generally as higher level cognition, in that they go beyond basic perceptual, motor, and memory functions. For example, frontal cortex has been implicated in problem-solving tasks like the Tower of Hanoi (e.g. Shallice, 1982; Baker *et al*., 1996; Goel and Grafman, 1995), which requires executing a sequence of moves to achieve a subsequent goal. Many theoretical perspectives summarize the function of frontal cortex in terms of 'executive control', 'controlled processing', or a 'central executive' (e.g. Baddeley, 1986; Shallice, 1982; Gathercole, 1994; Shiffrin and Schneider, 1977), without explaining at a mechanistic level how such functionality could be achieved. Computational models provide an important tool for exploring specific mechanisms that might achieve 'executive-like' functionality.

Working Memory and Active Maintenance

One proposal is that the fundamental mechanism underlying frontal function is 'active maintenance', which then enables all the other 'executive' functionality ascribed to the frontal cortex (Cohen *et al*., 1996; Goldman-Rakic, 1987; Munakata, 1998;

O'Reilly *et al*., 1999; O'Reilly and Munakata, 2000; Roberts and Pennington, 1996). For example, a flexible, adaptive, active maintenance system can meet information processing challenges by using the strategic activation and deactivation of representations (activation-based processing) instead of weight changes (weight-based processing) (O'Reilly and Munakata, 2000). There are tradeoffs between these types of processing (e.g. activations can be more rapidly switched than weights, but they are also transient); so using both kinds of processing is better than using either alone.

There is considerable direct biological evidence that the frontal cortex subserves the active maintenance of information over time, as encoded in the persistent firing of frontal neurons (e.g. Fuster, 1989; Goldman-Rakic, 1987; Miller *et al*., 1996). Many computational models of this basic active maintenance function have been developed (Braver *et al*., 1995; Dehaene and Changeux, 1989; Zipser *et al*., 1993; Seung, 1998; Durstewitz *et al*., 2000; Camperi and Wang, 1997). Some models have further demonstrated that active maintenance can account for frontal involvement in a range of different tasks that might otherwise appear to have nothing to do with maintaining information over time.

Inhibition, Flexibility, and Perseveration

For example, several models have demonstrated that frontal contributions to 'inhibitory' tasks can be explained in terms of active maintenance instead of an explicit inhibitory function. Actively maintained representations can support (via bidirectional excitatory connectivity) correct choices, which will therefore indirectly inhibit incorrect ones via standard lateral inhibition mechanisms within the cortex. A model of the Stroop task provided an early demonstration of this point (Cohen *et al*., 1990). In this task, color words (e.g. 'red') are presented in different colors, and people are instructed to either read the word or name the color in which the word is written. In the conflict condition, the ink color and the word are different. Because we have so much experience of reading, we naturally tend to read the word, even if instructed to name the color, so that responses are slower and more error-prone in the color-naming conflict condition than in the word-reading one. These color-naming problems are selectively magnified with frontal damage. This frontal deficit has usually been interpreted in terms of the frontal cortex helping to inhibit the dominant word-reading pathway. However, Cohen *et al*. (1990) showed that they could account for both normal and frontal-damage data by assuming that the frontal cortex instead supports the color-naming pathway, which then collaterally inhibits the word-reading pathway. Similar models have demonstrated that, in infants, the ability to inhibit perseverative reaching (searching for a hidden toy at a previous hiding location rather than at its current location) can develop simply through increasing ability to actively maintain a representation of the correct hiding location (Dehaene and Changeux, 1989; Munakata, 1998). Again, such findings challenge the standard interpretation that inhibitory abilities *per se* must develop for improved performance on this task (Diamond, 1991).

The activation-based processing model of frontal function can also explain why frontal cortex facilitates rapid switching between different categorization rules in the Wisconsin card sorting task and related tasks. In these tasks, subjects learn to categorize stimuli according to one rule via feedback from the experimenter, and then the rule is changed. With frontal damage, patients tend to perseverate in using the previous rule. A computational model of a related intradimensional/extradimensional (ID/ED) categorization task demonstrated that the ability to rapidly update active memories in frontal cortex can account for detailed patterns of data in monkeys with frontal damage (O'Reilly *et al*., 2002; O'Reilly and Munakata, 2000).

Computational models of frontal function can provide mechanistic explanations that unify the various roles of the frontal cortex, from working memory to cognitive control and planning and problem-solving. However, it remains to be shown whether complex 'intelligent' behavior can be captured using these basic mechanisms.

COMPUTATIONAL MODELS OF LANGUAGE USE GUIDED BY NEUROPSYCHOLOGICAL CASES

Damage to language-related brain areas causes a wide variety of impairments. One class of such impairments, the dyslexias (also known as alexias), have been the subject of a series of influential computational models of the normal and impaired reading process (Seidenberg and McClelland, 1989; Plaut and Shallice, 1993; Plaut *et al*., 1996). These models simulate the pathways between visual word inputs (orthography), word semantics, and verbal word outputs (phonology), and can account for different kinds of dyslexias in terms of differential patterns of damage to these pathways.

These models have been influential in part because they suggest an alternative, somewhat

counterintuitive, interpretation of how words are represented and how language processing works. Traditional models have assumed that the brain contains a 'lexicon', with distinct representations for different words. Furthermore, these models assume that reading a word aloud (i.e., mapping from orthography to phonology) can occur via two different routes: pronunciation rules (for 'regular' words like 'make'), or a mechanism like a lookup table (for 'exception' words like 'yacht') (Pinker, 1991; Coltheart *et al*., 1993; Coltheart and Rastle, 1994). In contrast to these dual-route models, the neural network models posit a single pathway to process both regular and exception words, and they employ a distributed lexicon without centralized, discrete lexical representations. Lexical processing occurs in pathways that map between different aspects of word representations (see Figure 4).

In general, neural networks can learn all kinds of different mappings: fully regular ones, like the spelling-to-sound mapping of the 'a' in words like 'make' and 'bake', as well as irregular mappings that occur in words like 'yacht'. Nevertheless, networks are sensitive to both the degree of regularity and the frequencies of different mappings. Specifically, neural network models predict frequency-by-regularity interactions that would not be expected in dual-route models, and which are observed in behavioral tests (Plaut *et al*., 1996). Furthermore, these network models can account for patterns of deficit with brain damage that would seem improbable under dual-route models. For example, people with surface dyslexia can read non-words (e.g. 'nust'), but they are impaired at retrieving semantic information from written words, and have difficulty in reading exception words. It would be natural in neural network models to interpret this as damage in the pathway between orthography and semantics. Interestingly, surface dyslexics' difficulty with exception words is generally limited to low-frequency exceptions (e.g. 'yacht'); they do not have difficulty reading high-frequency exceptions (e.g. 'are'). This pattern suggests that the remaining 'direct' pathway between orthography and phonology can handle both regular words and high-frequency exceptions, as in the network models. This pattern of data is not easily explained by the dual-route models: with two pathways, either regular words or exceptions should be affected, but not both, and independently of frequency.

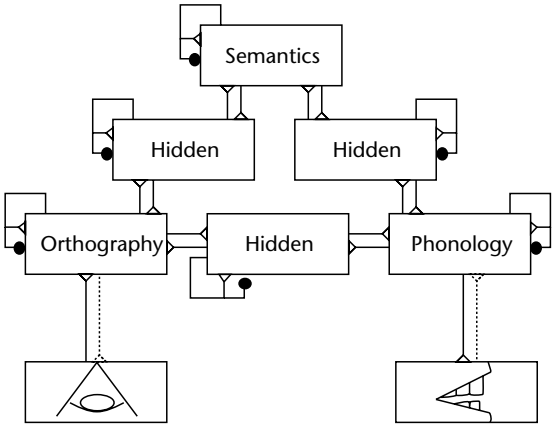


Figure 4. A neural network model of reading aloud. Words are represented in a distributed fashion across orthographic (visual word recognition), phonological (speech output), and semantic areas.

Neural network models of language can provide alternative, counterintuitive ways of explaining some of the complex patterns of deficits that occur with brain damage. Nevertheless, such models remain controversial: neural network accounts are challenged by revised versions of dual-route models, and by the complexity of different neuropsychological profiles associated with damage to different language areas.

CONCLUSION

Computational models based on the neural networks of the brain can provide important insights. Many models have applied a set of basic principles to a range of phenomena, and arrived at explanations completely different from those based on purely verbal cognitive theories. Hence, these models have played an important role in guiding empirical research and theorizing in a number of domains.

Despite these successes, many researchers remain skeptical of models. A common concern is that different models may employ different sets of mechanisms to explain the same data, so that it may not be very significant that a given model can simulate a set of data. Several points have been made in response to this concern.

Firstly, it applies not only to computational models, but to scientific theorizing in general (several theories can account for the same data). Competing theories and models can be evaluated by many other criteria than simply accounting for a set of data, such as the accuracy of their predictions, the coherence of their theoretical framework, and the ease of accounting for new data (Munakata and Stedron, 2002).

Secondly, mechanisms developed independently can turn out to be equivalent (e.g. O'Reilly, 1996), providing converging evidence for their utility, and indicating more coherence to principles than might otherwise be evident.

Thirdly, a common set of mechanisms appears to be emerging as the field continues to mature. For example, over 40 different phenomena (including most of what has been described above) have been modeled using a common set of mechanisms (O'Reilly and Munakata, 2000). This set of mechanisms was developed over many years by many different researchers, and has now been consolidated and integrated into one coherent framework (O'Reilly, 1998).

Therefore, there is a largely consistent set of ideas underlying many neural network models, and this framework provides an important way of understanding the connections between cognition and underlying neural systems.

References

Alvarez P and Squire LR (1994) Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences* **91**: 7041–7045.

Baddeley AD (1986) *Working Memory*. New York, NY: Oxford University Press.

Baker SC, Rogers RD, Owen AM *et al*. (1996) Neural systems engaged by planning: a PET study of the Tower of London task. *Neuropsychologia* **34**: 515–526.

Beiser DG, Hua SE and Houk JC (1997) Network models of the basal ganglia. *Current Opinion in Neurobiology* **7**: 185–190.

Bell AJ and Sejnowski TJ (1997) The independent components of natural images are edge filters. *Vision Research* **37**: 3327–3338.

Berns GS and Sejnowski TJ (1996) How the basal ganglia make decisions. In: Damasio A, Damasio H and Christen Y (eds) *Neurobiology of Decision-Making*. Berlin, Germany: Springer-Verlag.

Braver TS, Cohen JD and Servan-Schreiber D (1995) A computational model of prefrontal cortex function. In: Touretzky DS, Tesauro G and Leen TK (eds) *Advances in Neural Information Processing Systems*, pp. 141–148. Cambridge, MA: MIT Press.

Burgess N, Recce M and O'Keefe J (1994) A model of hippocampal function. *Neural Networks* **7**: 1065–1083.

Camperi M and Wang XJ (1997) Modeling delay-period activity in the prefrontal cortex during working memory tasks. In: Bower J (ed.) *Computational Neuroscience*, chap. XLIV, pp. 273–279. New York, NY: Plenum Press.

Cohen JD, Dunbar K and McClelland JL (1990) On the control of automatic processes: a parallel distributed processing model of the Stroop effect. *Psychological Review* **97**: 332–361.

Cohen JD, Romero RD, Farah MJ and Servan-Schreiber D (1994) Mechanisms of spatial attention: the relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience* **6**: 377–387.

Cohen JD, Braver TS and O'Reilly RC (1996) A computational approach to prefrontal cortex, cognitive control, and schizophrenia: recent developments and current challenges. *Philosophical Transactions of the Royal Society, Series B* **351**: 1515–1527.

Coltheart M and Rastle K (1994) Serial processing in reading aloud: evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 1197–1211.

Coltheart M, Curtis B, Atkins P and Haller M (1993) Models of reading aloud: dual route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608.

Contreras-Vidal JL, Grossberg S and Bullock D (1997) A neural model of cerebellar learning for arm movement control: cortico-spino-cerebellar dynamics. *Learning and Memory* **3**: 475–502.

Coslett HB and Saffran E (1991) Simultanagnosia. To see but not two see. *Brain* **114**: 1523–1545.

Dayan P (1992) The convergenece of TD(λ) for general λ. *Machine Learning* **8**: 341–362.

Dehaene S and Changeux JP (1989) A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience* **1**: 244–261.

Diamond A (1991) Neuropsychological insights into the meaning of object concept development. In: Carey S Gelman R (eds) *The Epigenesis of Mind*, chap. III, pp. 67–110. Mahwah, NJ: Lawrence Erlbaum.

Durstewitz D, Seamans JK and Sejnowski TJ (2000) Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology* **83**: 1733–1750.

Erwin E, Obermayer K and Schulten K (1995) Models of orientation and ocular dominance columns in the visual cortex: a critical comparison. *Neural Computation* **7**: 425–468.

Fukushima K (1988) Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*: **1**: 119–130.

Fuster JM (1989) *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe.* New York, NY: Raven Press.

Gathercole SE (1994) Neuropsychology and working memory: a review. *Neuropsychology* **8**: 494–505.

Gilbert CD (1996) Plasticity in visual perception and physiology. *Current Opinion in Neurobiology* **6**: 269–274.

Goel V and Grafman J (1995) Are the frontal lobes implicated in 'planning' functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia* **33**: 623–642.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Brookhart JM and Mountcastle VB (eds) *Handbook of Physiology. The Nervous System*, vol. V, pp. 373–417. Baltimore, MD: American Physiological Society.

Hasselmo ME and Wyble B (1997) Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural Brain Research* **67**: 1–27.

van Hateren JH and van der Schaaff A (1997) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society, Series B* **265**: 359–366.

Houk JC, Davis JL and Beiser DG (eds) (1995) *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press.

Hubel D and Wiesel TN (1962) Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–154.

LeCun Y, Boser B, Denker JS *et al*. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1**: 541–551.

Levy WB (1989) A computational approach to hippocampal function. In: Hawkins RD and Bower GH (eds) *Computational Models of Learning in Simple Neural Systems*, pp. 243–304. San Diego, CA: Academic Press.

Linsker R (1988) Self-organization in a perceptual network. *Computer* **21**(3): 105–117.

Livingstone M and Hubel D (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240**: 740–749.

Marr D (1971) Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society, Series B* **262**: 23–81.

McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionst models of learning and memory. *Psychological Review* **102**: 419–457.

Miller EK, Erickson CA and Desimone R (1996) Neural mechanisms of visual working memory in prefontal cortex of the macaque. *Journal of Neuroscience* **16**: 5154–5167.

Miller KD, Keller JB and Stryker MP (1989) Ocular dominance column development: analysis and simulation. *Science* **245**: 605–615.

Moll M and Miikkulainen R (1997) Convergence-zone episodic memory: analysis and simulations. *Neural Networks* **10**: 1017–1036.

Montague PR, Dayan P and Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**: 1936–1947.

Mozer MC (1991) *The Perception of Multiple Objects: A Connectionist Approach*. Cambridge, MA: MIT Press.

Mozer MC and Sitton M (1998) Computational modeling of spatial attention. In: Pashler H (ed.) *Attention*, pp. 341–393. London, UK: UCL Press.

Munakata Y (1998) Infant perseveration and implications for object permanence theories: a PDP model of the A-not-B task. *Developmental Science* **1**: 161–184.

Munakata Y and Stedron JM (forthcoming). Memory for hidden objects in early infancy. In: Fagen J and Hayne H (eds) *Advances in Infancy Research*, vol. XIV. Norwood, NJ: Ablex.

Oja E (1982) A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology* **15**: 267–273.

Olshausen BA and Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**: 607–609.

O'Reilly RC (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation* **8**: 895–938.

O'Reilly RC (1998) Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences* **2**: 455–462.

O'Reilly RC and McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a tradeoff. *Hippocampus* **4**: 661–682.

O'Reilly RC and Munakata Y (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

O'Reilly RC and Rudy JW (2000) Computational principles of learning in the neocortex and hippocampus. *Hippocampus* **10**: 389–397.

O'Reilly RC and Rudy JW (2001) Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review* **108**: 311–345.

O'Reilly RC, Norman KA and McClelland JL (1998) A hippocampal model of recognition memory. In: Jordan MI, Kearns MJ and Solla SA (eds) *Advances in Neural Information Processing Systems*, vol. X, pp. 73–79. Cambridge, MA: MIT Press.

O'Reilly RC, Braver TS and Cohen JD (1999) A biologically based computational model of working memory. In: Miyake A and Shah P (eds) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, pp. 375–411. New York, NY: Cambridge University Press.

O'Reilly RC, Noelle D, Braver TS and Cohen JD (2002) Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cerebral Cortex* **12**: 246–257.

Pinker S (1991) Rules of language. *Science* **253**: 530–535.

Plaut DC and Shallice T (1993) Deep dyslexia: a case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**: 377–500.

Plaut DC, McClelland JL, Seidenberg MS and Patterson KE (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115.

Posner MI, Walker JA, Friedrich FJ and Rafal RD (1984) Effects of parietal lobe injury on covert orienting of visual attention. *Journal of Neuroscience* **4**: 1863–1874.

Pouget A and Sejnowski TJ (1997) Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience* **9**: 222–237.

Roberts RJ and Pennington BF (1996) An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology* **12**(1): 105–126.

Samsonovich A and McNaughton BL (1997) Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* **17**: 5900–5920.

Schultz W, Apicella P and Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience* **13**: 900–913.

Schultz W, Apicella P, Romo R and Scarnati E (1995) Context-dependent activity in primate striatum reflecting past and future behavioral events. In: Houk JC, Davis JL and Beiser DG (eds) *Models of Information Processing in the Basal Ganglia*, pp. 11–28. Cambridge, MA: MIT Press.

Schultz W, Dayan P and Montague PR (1997) A neural substrate of prediction and reward. *Science* **275**: 1593–1599.

Schweighofer N, Arbib M and Kawato M (1998a) Role of the cerebellum in reaching quickly and accurately. I: A functional anatomical model of dynamics control. *European Journal of Neuroscience* **10**: 86–94.

Schweighofer N, Arbib M and Kawato M (1998b) Role of the cerebellum in reaching quickly and accurately. II: A detailed model of the intermediate cerebellum. *European Journal of Neuroscience* **10**: 95–105.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **20**: 11–21.

Seidenberg MS and McClelland JL (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* **96**: 523–568.

Seung HS (1998) Continuous attractors and oculomotor control. *Neural Networks* **11**: 1253–1258.

Shallice T (1982) Specific impairments of planning. *Philosophical Transactions of the Royal Society, Series B* **298**: 199–209.

Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing. II: Perceptual learning, automatic attending, and a general theory. *Psychological Review* **84**: 127–190.

Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**: 195–231.

Sutton RS (1988) Learning to predict by the method of temporal diferences. *Machine Learning* **3**: 9–44.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Swindale NV (1996) The development of topography in the visual cortex: a review of models. *Network: Computation in Neural Systems* **7**: 161–247.

Treves A and Rolls ET (1994) A computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**: 374–392.

Vecera SP and O'Reilly RC (1998) Figure–ground organization and object recognition processes: an interactive account. *Journal of Experimental Psychology: Human Perception and Performance* **24**: 441–462.

Verfaellie M, Rapcsak SZ and Heilman KM (1990) Impaired shifting of attention in Balint's syndrome. *Brain and Cognition* **12**: 195–204.

Wickens J (1997) Basal ganglia: structure and computations. *Network: Computation in Neural Systems* **8**: 77–109.

Zipser D, Kehoe B, Littlewort G and Fuster J (1993) A spiking network model of short-term active memory. *Journal of Neuroscience* **13**: 3406–3420.

Consciousness, Disorders of

Introductory article

Fred Plum, Weill Medical College of Cornell University, New York, USA
Nicholas D Schiff, Weill Medical College of Cornell University, New York, USA

CONTENTS

Disorders of consciousness include coma, stupor, confusion and other abnormal states of acute brief or moderately sustained unconsciousness.

INTRODUCTION

The brain generates the mind, and the healthy, wakeful mind formulates consciousness. The American psychologist William James in 1890 stated, 'Consciousness is the [indispensable] fundamental awareness of the self's internal ego.' He then expanded that self-centered focus to identify the self's greater qualities of memory, attention, intention, chronological time, emotion, learned behavior, and several other less general psychological qualities. At that early time, only philosophical thinking interpreted gross anatomic knowledge in trying to understand how the awake brain might lose conscious functions.

Modern neurological medicine has defined several distinct behavioral pathological states that arise from inherited and acquired brain injuries and lead to disorders of consciousness. Brain injuries that reflect global disorders of consciousness include stupor and coma, the vegetative state, akinetic mutism, absence and partial complex seizures, delirium, and severe dementia. These global disorders, described below, totally disable the capacity of the individual for intentional behaviors. Though different in pattern, 'focal' disorders of consciousness can exist in several serious illnesses. A patient suffering a focal disorder of consciousness can be awake and interact with the environment, and yet exhibit severe alterations in awareness. These disorders uniquely illustrate the constructed nature of conscious experience.

THE FORMULATION OF CONSCIOUSNESS

All people with a healthy brain and body can recognize themselves, their thoughts and their intentional conscious activity. Descriptions may vary in detail, but ask people what they think about the quality called their consciousness and the first reply is likely to be, 'I'm awake and I'm here'. Proof often follows with (for example) 'I'm John Smith!'

An educated person recognizes conscious awareness as a continuously unfolding, automatic sense of being awake, alive, and logically thoughtful. Actually, one's mind is being continuously filled with flowing thoughts, normal language, recent memories, learned motor behavior, or novel discoveries. Even the most educated person, however, sometimes wonders about how the brain automatically experiences normal emotions, how it generates logical thinking, and how it induces the smooth flow of relevant or original thoughts and coordinated deeds.

'Now, how did I come to think about that?' is an often-expressed question, but usually not one that is part of everyday conversation. Nor do we wonder what preceding activities our brain generates when we automatically take our daily walk down the same lane. Even when we 'instinctively' jump out of the way of an unseen, oncoming vehicle, we often fail to realize that our awake, preconscious frontal lobe thought and acted first. Only after we have jumped away from the danger do we become aware of our act and experience an emotional feeling of fright. This example illustrates how we often act or even speak before we consciously think.

The normal brain's cognitive processing systems organize intentional behaviors drawing on a rapidly accessed, vast store of relevant memories. It preconsciously formulates either incoming or spontaneous information in less than a quarter of a second. It is astonishing to realize that the functions of memory, intention, and perception may largely occur before any act or expressed thoughts enter immediate conscious awareness. We may

think it strange that when we hold a conversation, our mind has preconsciously formulated what we are going to say a half-second or more before we actually say it. The Nobel laureate in medicine, Gerald Edelman, recognized this normal, instinctive preconscious formulation of thoughts, words and athletic acts in the ingenious title of his book, *The Remembered Present, A Biological Theory of Consciousness*.

Neuropsychological Dimensions of Consciousness

Consciousness is a time-ordered, egocentric process that interweaves inner and outer perceptions, stored memories, and immediately innovative thoughts. Emotional feelings imbue conscious awareness and sharpen intentional actions. Memory provides not only the ultimate storehouse of explicit conscious knowledge; it also develops the preconscious, implicit brain learning of motor skills and physical practice. Memory qualities and quantities depend on the combination of our innate cognitive talents, our subsequent schooling, our continuously thoughtful appraisal of new objects, and our interpersonal learning from and about people. The goal can be athletic, intellectual, or both. All evidence indicates that the earlier the young begin to learn and continue lifelong studies, the greater will be their future mental and behavioral capacities. Indeed, the longer a person's education and thought-requiring occupation last, the greater the brain's and body's functional longevity.

How this serially time-ordered, organized process of consciousness incorporates outer information with inner attention and immediate evaluation is the subject of intense neuroscientific investigation. Several distinct neuropsychological qualities can be ascribed to distributed networks of brain regions that selectively contribute to organized, wakeful human consciousness. These networks include the brainstem and allied arousal systems which control the sleep–wake cycles of the entire forebrain; prefrontal cortical regions (e.g. anterior cingulate cortex, frontal eye fields) which support continuous attention to self and environment and immediate intention; and posterior cortical regions of the temporal lobes (superior temporal gyrus) and parietal lobes (inferior parietal cortex) which support self-sensory perceptions and instinctive, and automatic awareness of inner and outer spatial relationships. Memory systems of the brain are widely distributed and depend on the integrity of the medial temporal lobe (hippocampus, entorhinal cortex) for initial storage, and on multiple cortical association areas (frontal, parietal, temporal, and occipital) and parts of the thalamus for functions of both storage and retrieval.

Additional neuropsychological qualities include the mind's chronological ordering of events (of unknown localization but disordered by injuries to the thalamus), moods and emotions (contributed by distributed regions of the 'limbic' brain). Learnt symbolic abstractions of verbal (left hemisphere), musical (right hemisphere), and geometric languages (left posterior parietal regions) contribute to humans' singular qualities of normal awareness.

SPECIFIC DISORDERS OF CONSCIOUSNESS

Coma

Coma is a totally unconscious and unarousable brain state resembling sleep, in which the eyes are closed and which lasts 24 h or more, due to any of several major causes. One is the use of sustained therapeutic anesthesia. More frequent causes are direct brain injury or diseases affecting the brain's cerebral hemispheres and arousal systems. Table 1 compares the loss of neuropsychologic

Table 1. Global disorders of consciousness

	Coma	*PVS*	*ASZ*	*AKM*	*HKM*	*CPS*	*DEL*
Arousal	−	+	+	+	+	+	+
Attention	−	−	−	+	+/−	+/−	+/−
Intention	−	−	−	−	+	+/−	+/−
Memory	−	−	−	−	−	−	+/−
Awareness	−	−	−	−/?	+/−	+/−	+/−

AKM, akinetic mutism; ASZ, absence seizures; CPS, complex partial seizures; DEL, delirium; HKM, hyperkinetic mutism; PVS, persistent vegetative state; −, absent; + present (in crude form for attention, AKM, and intention, HKM), +/− incompletely expressed; −/? apparently absent.

components incurred in coma with those of other disorders of consciousness.

Stupor

Stupor is a condition of deep sleep or behaviorally similar unresponsiveness from which the person cannot be aroused except by vigorous and repeated exogenous stimulation. As soon as such stimulation ceases, the person relapses into the unresponsive state. Light stupor is typical in cases of overdoses of soporific drugs or alcohol. Deep stupor more frequently reflects severe pharmacological, metabolic, or traumatic injury to the brain. The term 'semi-coma' is occasionally used in nonmedical writing to describe patients in stupor or persistent vegetative state but is not considered a diagnostic category.

Persistent Vegetative State

The vegetative state is a condition in which physiologically active, systemic organs continue to sustain the life of a body that has become at least temporarily devoid of a conscious brain. In most cases of coma, wakefulness will return spontaneously in a matter of days or weeks; but in some people, despite a wakeful appearance, the mind may be absent for many weeks, months or forever. This tragedy has been named the *persistent vegetative state* (PVS). Such patients express irregularly timed sleep–wake patterns, but all feeding and bodily care must be provided by external sources. The term 'arbitrarily' identifies patients who remain psychologically unconscious for at least a month. They are alive, but totally unaware of self or their environment. The vegetative state presents the fundamental clinical dissociation of arousal from all other components of consciousness (Table 1).

The clinical judgment of unconsciousness in PVS has been supported by the results of positron emission tomography (PET) scan studies that reveal overall cerebral metabolism to be reduced by 50% or more below the normal rate. The observed metabolic levels are equivalent to those found in persons undergoing deep surgical anesthesia. In a study of behavioral and physiological variations in a few patients in the vegetative state, one woman randomly expressed occasional single, understandable words. Her PET studies revealed isolated islands of left frontotemporal cerebral structures that operated at an abnormally low metabolic rate but at nearly twice the rates of the remaining brain. Similar isolated expressions have been encountered in several other vegetative patients. Typically, the patients express easily identifiable, stereotypical, emotional-limbic responses. These emotional expressions probably reflect distinct and isolated limbic mechanisms; their preservation is likely to depend on integrative brainstem structures that lie outside the corticothalamic systems that typically undergo overwhelming injury in PVS patients.

Syncope

Syncope (fainting) consists of brief unconsciousness caused by reduction of systolic arterial blood flow through the brain. Most syncope is benign and occurs in persons younger than 50 years. Termed 'vasogenic', it reflects sudden dilation of the body's cholinergic and sympathetic neurovascular systems, reduces systemic blood pressure and deprives the erect brain's critical oxygen supply. A second type is less frequent and affects older people suffering from postural orthostatic hypotension. A third type affects elderly people with severe cardiac, cardiopulmonary, or systemic atheromatous illness. Such patients rarely regain normal brain function if they fail to gain accurate awareness in more than 4–5 minutes.

Concussion

Concussion is an unconscious state that immediately follows a severe traumatic head injury. Since its ultimate duration cannot be predicted accurately, some surgeons call post-traumatic lack of consciousness 'concussion' for 24 h; after that, the term is changed to 'coma'.

At its least, concussion interrupts the brain's organized thoughts and impairs or blocks its recent memory. Acute severe concussions may suddenly and briefly suppress vegetative brainstem functions, thereby invoking transient breathlessness, slowed heart rate, low blood pressure, and widening of the pupils. Boxing knockouts for 10 s or more vividly exemplify moderate to severe concussion, as the bewildered athlete staggers from the ring and sometimes falls. A few knocked-out boxers will remain unconscious after the count, and a very few may die from acute brain hemorrhage. A measurable group may gradually develop dementia during their early sixth decade. Many drivers or passengers in serious road traffic accidents can suffer brief knockouts of a few seconds, followed by several hours of confused memory and, frequently, light coma or unsteady behavior. Lack of arousal during this time is sometimes regarded as short-term concussion, but brief coma

is a more accurate label to apply until the person awakens.

Confusion

Confusion can be either temporary or permanent. Temporary confusion refers to disturbed memory and an inexact orientation of time, place, or person. Awakening from deep sleep after moderate sedation, suffering the effects of using excessive alcohol or street drugs, or awakening in a strange room, are typical examples. Chronic, waking confusion relates to sustained difficulties in identifying time, date, the environment and the failure to recognize long-known persons. It is also a gentle term for dementia.

Absence and Complex Partial Seizures

Seizures reflect severe impairments of self-aware consciousness, accompanied by unique forms of behavior. Absence seizures typically occur in children and are often noted as 'staring spells'. During the event the eyes typically fix in a forward stare, motion ceases, and movements of the lips or eyelids may be noted. People who undergo frequent absence attacks (once called *petit mal*) may lose extended self-awareness for a matter of hours and sometimes longer. During these states they remain awake and usually continue vaguely purposeful behavior.

Absence seizures originate from the cerebral cortex but involve brainstem and thalamic neuronal networks. People suffering severe complex partial seizures, a different neurological disorder often emanating from the temporal lobes (see below), lose their cognitive memories, but may also express a variety of learned behaviors. Both types of event exhibit attentional and intentional failure, loss of working memory, and perceptual dissociation. In their classic form, absence seizures may be interpreted to represent momentary vegetative states (see Table 1).

Akinetic Mutism

The term 'akinetic mutism' covers different behavioral expressions that relate to damage of several cerebral and subcortical structures. While sometimes confused with the vegetative state, akinetic mutism may resemble a state of constant hypervigilance. Such patients appear attentive and vigilant but remain motionless. The preservation of visual tracking in the form of following persons or moving objects with smooth, roving eye movements can differentiate this condition from the vegetative state. Classic akinetic mutism as listed in Table 1 reflects the recovery of a crude wakeful attentiveness without the apparent recovery of any other neuropsychologic function.

A similar picture, but excluding absence of eye movements, can rarely be a feature of untreated, rigid Parkinson disease. A strong clinical resemblance to this syndrome has been identified in some forms of variant Creutzfeldt–Jakob disease. The hyperattentive form of this disorder is typically seen in patients with large bilateral injuries to the medial and ventral frontal lobes (see below).

Hyperkinetic Mutism

Hyperkinetic mutism is a wakeful, continuous movement disorder accompanied by at least partial loss of global self-awareness. Patients with hyperkinetic mutism exhibit totally unrestrained but coordinated motor activity in the absence of external evidence of awareness of the environment. The patients also demonstrate an inability to develop conditioned responses, and produce no apparent memory of self.

Hyperkinetic mutism is the converse of akinetic mutism, with preserved unconscious expression of frontal intentional mechanisms, loss of sustained directed attention presumably requiring posterior attentional components of the inferior parietal lobe or posterior temporal lobe (see below), and a state of behavioral unawareness despite a whirlwind of activity. In contrast to the akinetic mute state, these people demonstrate minimally expressed intention and attention. The fragment of intention expressed in the meaningless motor activity of the hyperkinetic mute person is a reciprocal of the crude form of attention seen in akinetic mutism. Both examples reveal the fundamentally unconscious nature of such fragmentary neuronal activity.

Similar examples of such unconscious motor activity include the repetitive, uncontrollable production of words in the neurological disorder known as Tourette syndrome.

Delirium

Delirium is generally perceived as an acute or semiacute temporary deficit of attention and working memory. A salient component is temporal disorientation. Delirium may follow acute, moderately severe head injuries, encephalitis, bacterial meningitis, exceptionally high fever, heat stroke, or withdrawal from chronic alcoholism or drug misuse. Delirium in patients less than 45 years old usually

subsides without serious reduction in intelligence, but in alcoholism the person must abandon alcohol completely after the first or second delirious bout or begin to lose intellectual capacities permanently. Elderly people suffering mild dementia often become delirious during acute systemic illness or frequently changed surroundings. Visual hallucinations or impaired perceptions often occur in systemic delirium, whereas auditory hallucinations appear more often, but not solely, in people with schizophrenia.

Dementia

Dementia is characterized by two different conditions. One is a permanent, sometimes fluctuating loss of short-term or long-term memory. It can follow severe brain trauma, a sudden, sustained loss of oxygen to the brain, or surgical removal of the anterior areas of both temporal lobes. The other consists of an insidious, gradual loss of (first) short-term and (later) long-term memory. This process results from degeneration and death of nerve cells in the cerebral cortex.

Focal Unconsciousness: Agnosia, Anosognosia, and Neglect

Agnosia is a term specifically applied to different types of focal losses of awareness. Examples include an inability to see or feel objects as a whole greater than the sum of several parts, and a loss of specific capacities to hear aspects of sounds. A rare bilateral injury to the ventral temporal occipital lobe may produce the loss of perception of motion, leading to an experience of life as if seen constantly through a stroboscope, never in continuous motion.

Anosognosia is a term specifically applied to a loss of awareness and an inability to consciously perceive. Examples of anosognosia include denying that one's hand is one's own and unable to move intentionally. This form of focal unconsciousness is also labeled 'neglect' and is typically applied to a syndrome arising from damage to the right parietal lobe. This normally provides automatic knowledge of the contralateral body as well as the immediate outer space that surrounds it. Neglect increasingly appears to be a disorder of entry of primary sensory information into the appropriate internal context to be integrated into the construction of consciousness. Neglect can be seen following damage to either frontal or posterior (inferior parietal or superior temporal) cortices (see below).

RELEVANCE TO UNDERSTANDING HUMAN CONSCIOUSNESS

Anatomic Relationships

Disorders of consciousness are often generated by selective brain injuries. Specific neuropsychological deficits accompanying these disorders reflect the relatively segregated cerebral neuronal networks that generate human consciousness and complex behavior. Autopsies over almost two centuries and the increasing knowledge of functional anatomy provided by modern brain imaging have greatly added to neuropsychological understanding of conscious or unconscious behavior. Several brain regions are implicated in these disorders, including the two cerebral hemispheres, each of which possesses approximately half of the cerebral cortex, the thalamus, and the basal ganglia. Near the mid-brainstem, they connect with the large cerebellum and the arousal systems that lie within the brainstem. To discover just how this network generates consciousness has become a major scientific effort. (*See* **Cerebellum**; **Basal Ganglia**)

Nonspecific arousal is generated largely in the brainstem and is indispensable to supporting sleep–wake cycling and the wakeful states of consciousness. Cholinergic (pedunculopontine, later dorsal tegmental nuclei), noradrenergic (locus ceruleus), and other neuronal populations located within the upper brainstem, hypothalamus, and basal forebrain have a key role in organizing this large-scale human behavior. By itself, however, arousal is independent of expressed neuropsychological qualities, as is evident in the vegetative state or in 'absence seizures' (see Table 1). Brain mechanisms that govern sleep and its various dreams and perambulations only partially overlap the circuitry of normal wakeful consciousness. The integrity of both distributed cortical and other subcortical structures as is necessary for the expression of integrated cerebral activity to generate consciousness.

Cognitive capacities expressed in the conscious state depend on the moment-to-moment continuity of short-term memory (disordered in delirium) with other neuropsychological components. Short-term or working memory appears to depend strongly on the integrity of the prefrontal and parietal cortices along with subcortical structures. The richness of mental life contained in the storage of long-term memories is a distributed capacity of the association regions of the cerebral cortex and is severely degraded in dementia.

The cortical regions indispensable for conscious behavior are the frontal lobes: these largely govern and express behavior, both immediate and learnt. Their functions provide the executive generator and dictator of consciousness, organizing mood, behavior, and mind. Within the frontal lobes the basal forebrain area has evolved from ancient mammalian brains and occupies most of the undersurface. It participates in generating emotional feelings and social behavior as well as stimulating the person's intentional purposes. The lateral and medial prefrontal areas (including the dorsolateral prefrontal regions, supplementary motor zones, and anterior cingulate cortices) largely influence physical coordination and participate in volitional and cognitive aspects of attention and working memory.

The most posterior regions of the lateral and medial frontal lobe generate and regulate coordinated expressions of logical manipulations, language, intended eye movements and, ultimately, all coordinated, intentional behavior. Examples include skilled athletics, the expression of well-learnt and practiced instrumental music, and other rapidly expressed activity.

Functional generation of self-directed attention and intention are mapped strongly in the ventral-medial frontal lobes and less frequently the posterior thalamus and rostral mesencephalon.

Akinetic mutism reflects the disabling of the ventral-medial and medial frontal and prefrontal networks (including a large contribution from the deep gray-matter structures of the basal ganglia, which interact with the cortex via long-loop connections with the thalamus and underpin much routine learnt behavior), providing volitional drive and self-directed (executive) attention. The crude aspect of attention remaining in this state of impaired consciousness may originate from automatic orienting systems driven by posterior parietal and subcortical structures (thalamus).

The posterior parts of the cerebrum, including the parietal, occipital and temporal lobes, in conjunction with the thalamus, generate the perceived contents of thoughtful consciousness. They receive their direct signals of attention and intention from the frontal lobes and express their immediate demands. The occipital lobes receive inputs of retinal vision, which are processed further within the adjacent temporal lobe. Auditory stimuli are also processed in the temporal lobe. Abstract cognitive icons represent the verbal, musical, mathematical, geometric, and pictorial languages that make up our intellectually conscious knowledge. Most of these particular cognitive qualities and contents are dominantly expressed by the left cerebral hemisphere:

The right inferior parietal lobe and adjacent superiorlateral temporal lobes, however, normally provide a person's dominant preconscious attentive perception and awareness of both the left side of the body and its surrounding environment. Severe acute damage to the right parietal-temporal areas as described in the paragraph on focal unconsciousness may cause total unawareness of the entire left side of the individual's personal universe. Lost is the memory of being able to see, or to remember normal vision; lost is the accurate perception of any existing left-spatial noises; lost is total awareness or memory of the absent hemiworld to the left and, remarkably, the person's ability to recognize his or her own left arm, leg or ear. This remarkable clinical syndrome demonstrates that our conscious experience is instinctual and can be lost in parts.

Evidence from neurological disorders of consciousness demonstrates that subcortical structures are also essential for normal integrative brain function associated with consciousness. Most causes of the global disorders of consciousness reviewed above appear to arise from either large bilateral injury to frontal (e.g. bilateral medial-basal frontal injuries and akinetic mutism) or posterior association cortices (bilateral temporal-parietal association areas and hyperkinetic mutism). In addition, it is known that selective subcortical injuries (generally damage to medial aspects of the thalamus or upper brainstem) may produce identical or very similar disorders. The subcortical injuries that may produce transient coma, vegetative state, akinetic mutism, or conditions resembling hyperkinetic mutism also implicate brainstem and thalamic structures. These include the brainstem arousal systems important for sleep and wake cycling and related brainstem and thalamic substructures that play a part in the complex, large-scale integration of many cerebral networks. The contribution of these deep brain structures may lie in the selective facilitation of activity patterns that allow widely separated brain regions to briefly communicate

Further Reading

Crick F (1964) *The Astonishing Hyphothesis. The Scientific Search for the Soul*. New York, NY: Charles Scribner's Sons.

Edelman GM (1987) *Neural Darwinism, the Theory of Neuronal Group Selection*. New York, NY: Basic Books.

Edelman GM (1989) *The Remembered Present A Biological Theory of Consciousness*. New York, NY: Basic Books.

Edelman GM and Tononi G (2000) *A Universe of Consciousness. How Matter Becomes Imagination*. New York, NY: Basic Books.

Plum F (1991) Coma and related global disturbances of the human conscious state. *Cerebral Cortex* **9**: 359–425. New York, NY: Plenum Press.

Plum F and Posner JB (1982) *Stupor and Coma*, 3rd edn. New York, NY: Oxford University Press.

Wilkinson IMS (1999) *Essential Neurology*, Chap. 11 Unconsciousness, pp. 171–186. London, UK: Blackwell Science.

Cortical Columns

Intermediate article

Geoffrey J Goodhill, Georgetown University Medical Center, Washington, DC, USA
Miguel A Carreira-Perpiñán, Georgetown University Medical Center, Washington, DC, USA

CONTENTS

In many regions of the cortex, neuronal response properties remain relatively constant in a direction perpendicular to the surface of the cortex, while they vary in a direction parallel to the cortex. Such columnar organization is particularly evident in the visual system, in the form of ocular dominance and orientation columns.

INTRODUCTION

The most prominent feature of the architecture of the cortex is its horizontal organization into layers. Each layer contains different cell types, and forms different types of connections with other neurons. However, a strong vertical organization is often also apparent: neurons stacked on top of each other through the depth of the cortex tend to be connected and have similar response properties despite residing in different layers. This type of vertical structure is called a *cortical column*, and has been hypothesized to represent a basic functional unit for sensory processing or motor output. Columnar organization has been most extensively studied in the somatosensory and visual systems.

DISCOVERY OF COLUMNAR ORGANIZATION

Cortical columns were first discovered electrophysiologically by Mountcastle (1957). When he moved an electrode obliquely to the surface of somatosensory cortex, he encountered neurons that responded to different sensory submodalities (e.g. deep versus light touch). However, when the electrode was moved perpendicularly to the cortical surface, all neurons had similar response properties. He summarized his findings as follows:

> These data … support an hypothesis of the functional organization of this cortical area. This is that the neurons which lie in narrow vertical columns, or cylinders, extending from layer II through layer VI make up an elementary unit of organization, for they are activated by stimulation of the same single class of peripheral receptors, from almost identical peripheral receptive fields, at latencies which are not significantly different for the cells of the various layers.

Shortly following this, vertical uniformity was also found in the visual system by Hubel and Wiesel (1977). Here, response properties that vary across the cortical surface but not through the depth of the cortex include the location of the neuron's receptive field in visual space, and the degree to which neurons are dominated by one eye. Columnar organization has also since been found in the auditory cortices of cat and monkey, where alternating bands related to monaural or binaural responses occur. A number of techniques have been employed for the experimental determination of cortical columns since the original use of electrode penetrations. These include methods based on axonal transport of substances such as horseradish peroxidase; on the differential consumption of radioactive 2-deoxyglucose by neurons; on optical imaging techniques, where cortical activity is converted to a visual signal by changes in reflectance or by voltage-sensitive dyes; and most recently on functional magnetic resonance imaging (fMRI).

There are some difficulties with defining exactly what is meant by a column. In some cases it is relatively clear: for instance, 'barrels' in somatosensory cortex and ocular dominance columns in visual cortex have fairly discrete boundaries with neighboring columns. In other cases, for instance orientation columns, there is a smooth variation in response properties moving parallel to the cortical

surface, rather than a series of discrete jumps. Another problem is that the term 'column' has been used to refer to structures at several different scales. At one extreme, from an anatomical point of view, are narrow vertical chains of neurons seen in Nissl-stained sections, barely more than one cell diameter wide, sometimes called *minicolumns*. At the other extreme, largely from a theoretical point of view, are complete functional units up to 1 mm in size, sometimes called *hypercolumns*. In between, Szentágothai (1978) specifies a generic column to be 200–300 μm wide. This article avoids such definitional issues by focusing on well-characterized examples of columnar organization. (*See* **Neuroimaging**)

COLUMNS IN THE VISUAL SYSTEM

Ocular Dominance Columns

Moving parallel to the surface of the primary visual cortex (V1) of several mammalian species, notably ferrets, cats, monkeys, and humans, there is a regular alternation between groups of neurons that respond best to input in the left eye and neurons that respond best to input in the right eye. The anatomical basis of this physiological pattern is the segregation of the thalamic input fibers – lateral geniculate nucleus (LGN) afferents – representing the left and right eyes to the visual cortex (Figure 1(a)). Although these fibers terminate primarily in layer 4 of the cortex, and this is where ocular preference is most sharply defined, a similar bias is also seen in higher and lower layers. This vertical structure of monocular preference is called an 'ocular dominance column' (reviewed by Hubel and Wiesel, 1977). When the entire pattern of eye preference is visualized in V1, for instance by injection of a radioactive tracer into one retina and its subsequent transport to the cortex, an alternating pattern of stripes is observed (Figure 1(b)). The periodicity of this pattern varies depending on the species and location in the cortex, and also varies substantially between individuals (Horton and Hocking, 1996):

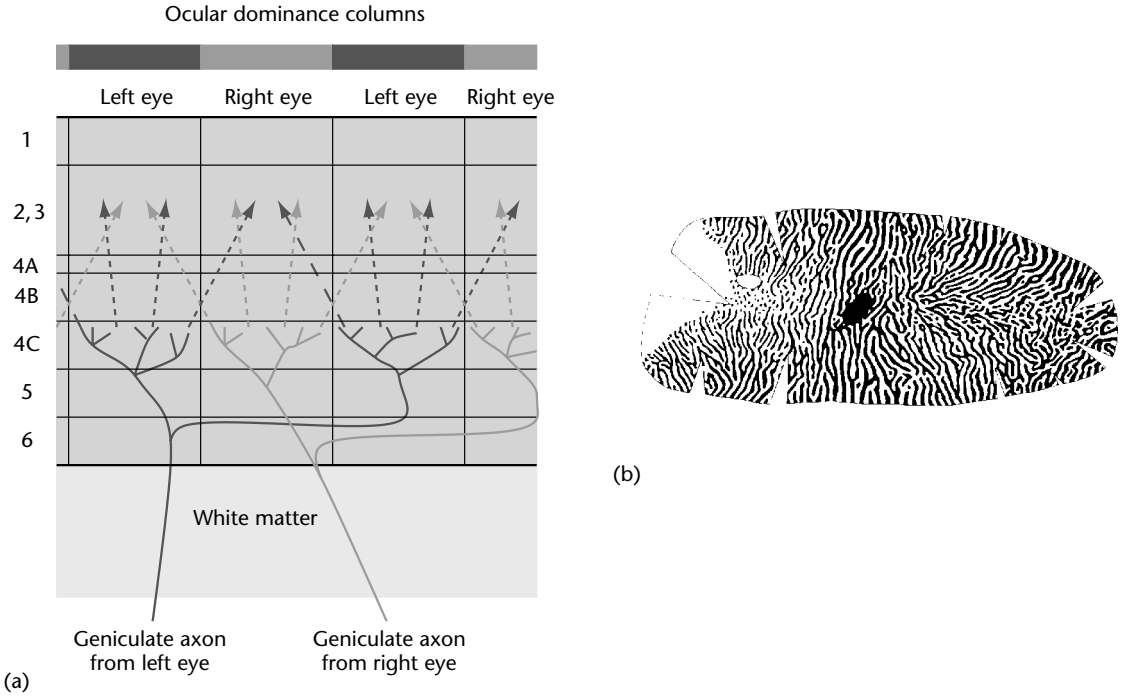


Figure 1. Ocular dominance columns in a monkey. (a) Anatomical basis. Each afferent axon from the lateral geniculate nucleus ascends through the deep layers of V1 (layers 5, 6) subdividing repeatedly and terminating in layer 4C in a couple of 0.5 mm-wide clusters separated by 0.5 mm gaps (approximately). Axons from the two eyes alternate, giving ocular dominance columns in 4C. The presence of horizontal connections and the arborization between different layers brings about overlapping and blurring of ocular dominance columns beyond layer 4: the ocular dominance of a given cell varies then between pure monocularity and pure binocularity. Adapted from Hubel (1995). (b) The pattern of ocular dominance columns from the primary visual cortex of a macaque monkey. White represents regions of cortex dominated by input from one eye, black the other eye. The width of individual columns is 0.5–1 mm. Source: LeVay *et al*. (1985) *Journal of Neuroscience* **5**: 486–501, © 1985 by the Society for Neuroscience.

in fact, each ocular dominance pattern is apparently as unique as a fingerprint. It can be seen from Figure 1 that these columns are in fact more like slabs, being long and relatively narrow rather than short and round.

Orientation Columns

Another type of columnar organization observed in the visual cortex is the orientation column. Many neurons in V1 respond best to an edge or bar of light at a specific orientation. This preferred orientation remains roughly constant through the depth of the cortex but varies mostly smoothly across the surface of the cortex. The overall pattern of orientation columns can be visualized by optical imaging methods. Cortical tissue changes its reflectance properties very slightly when neurons are active, and so by examining changes in reflected light from the cortical surface as visual stimuli of varying orientations are presented one can build up a picture of the complete map. An example is shown in Figure 2. A notable feature is the presence of pinwheels, point singularities around which all orientations are represented in a radial pattern. Superimposing the ocular dominance and orientation maps from the same animal, one observes regular geometric relationships between the two columnar systems. For instance, ocular dominance and orientation columns tend to meet at right angles, and orientation pinwheels tend to lie at the center rather than at the borders of ocular dominance columns.

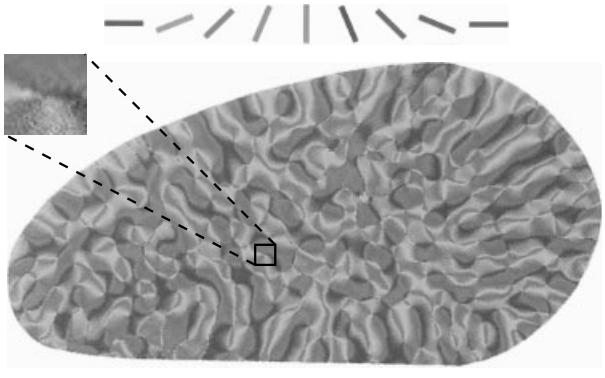


Figure 2. The orientation map in primary visual cortex of a tree shrew. The different degrees of shading represent patches that have different orientation preferences. The detail shows a pinwheel, where the orientation preference changes by 180° along a closed path around the center. Adapted from Bosking *et al*. (1997) *Journal of Neuroscience* **17**: 2112–2127, © 1997 by the Society for Neuroscience.

Other Types of Columns

Besides ocular dominance and orientation columns, several other types of columns are also present in the visual cortex. The most fundamental of these are what might be called position columns. Neurons in V1 have small receptive fields localized at specific positions in visual space. Moving vertically through the cortex, neurons have receptive fields at similar positions, while moving horizontally there is a smooth progression of visual field position versus cortical position, forming a topographic map of visual space in the cortex. This locality of information processing in visual cortex can also be seen from the fact that a small injury (e.g. a tumor or stroke) in part of V1 can cause blindness in a localized area in the visual field (a scotoma) with normal vision elsewhere, rather than an overall worsening of vision. Other receptive field properties that are organized into columns include preference for the spatial frequency of a stimulus across the receptive field, preference for the direction of movement of a stimulus, and disparity of inputs from the two eyes. All these columnar systems occupy the same cortical territory as the ocular dominance and orientation columns, and show complex geometric relationships that have yet to be fully characterized. Color-sensitive cells in layers 2–3 of monkey visual cortex (although not in other layers) are grouped in blobs, in which neurons respond to the color of a stimulus, but are mostly insensitive to orientation – unlike cells outside the blobs (the interblobs) which show marked orientation selectivity. (*See* **Color Vision, Neural Basis of**; **Depth Perception**)

Factors Driving the Formation of Columns

A number of different experiments on visual deprivation, where the visual experience that an animal receives is distorted, have shown that it is possible to produce physiological and structural changes in the columnar organization of visual cortex. For example, if one eye is sutured closed or strabismus is induced then most cells become monocular; if animals are presented with only bands at a specific orientation angle, then the proportion of cells that respond to that angle increases; if movement in a particular direction is excluded, the cells that would have responded to that movement direction no longer do so. Recovery to normal structure is also possible. However, both deprivation and recovery are effective only in an early period of the life when the connections are

developing. These experiments suggest that the development of ocular dominance columns is the result of two competing processes: segregation is promoted when neural activity is equal in each eye but not correlated between both eyes; and binocular innervation of neurons and merging of the two sets of columns is promoted by the correlation in activity between corresponding retinal areas of the two eyes that results from normal binocular vision.

However, the relative importance of intrinsic, or genetically programmed, factors versus extrinsic, or activity-driven, factors is still not clear. On the one hand, Crowley and Katz (1999) found that total removal of retinal influence early in visual development did not prevent segregation of geniculocortical axons into ocular dominance columns of normal periodicity. They thus propose that ocular dominance column formation relies on molecular cues present on thalamic axons, cortical cells, or both. On the other hand, Sur and colleagues (e.g. Sharma *et al*., 2000) have surgically rewired the optic nerve of newborn ferrets to feed into auditory thalamus (itself deprived of auditory inputs), which in turn projects to primary auditory cortex (A1) – rather than the normal pathway, optic nerve to LGN to V1. Such rewired ferrets develop in A1 a pattern of orientation columns with some similarities to that normally present in V1, though with a less regular periodicity. Such new cortical structure perceptually acts as visual; that is, the animals use the rewired A1 to see, rather than hear – although the resulting visual acuity is lower than normal. This suggests that retinal inputs can drive the formation of columns.

COLUMNS IN THE SOMATOSENSORY SYSTEM

The first physiological indication of cortical columns came from experiments by Mountcastle (1957) in the somatosensory cortex of cats. He found three types of neurons: those activated by light pressure on the skin, those activated by movement of hairs, and those activated by deformation of deep tissues (as occurs during for instance joint movement). As summarized by Mountcastle:

> Cells belonging to each subgroup were found in all the cellular layers. In 84 per cent of penetrations across the cellular layers which were directed perpendicularly, all the neurons encountered belonged to either cutaneous or deep subgroups. These modality-specific vertical columns of cells are intermingled for any given topographical region.

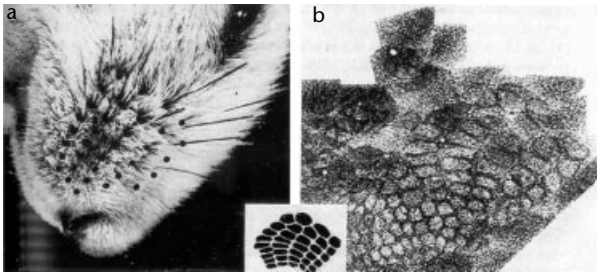


Figure 3. Posteromedial barrel subfield from a mouse's muzzle. Reprinted from Woolsey TA and van der Loos H (1970) The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. *Brain Research* **17**: 205–242, © 1970, with permission from Elsevier Science.

More recent results have amplified this. For instance, Favorov and Diamond (1990) found discrete jumps in receptive field location between neighboring columns in cat primary somatosensory cortex, with no significant receptive field shifts within a column. However, the most striking example of columns in somatosensory cortex are the 'barrels' discovered by Woolsey and van der Loos (1970). In animals such as mice and rats, the long whiskers of the face are present in a stereotyped spatial pattern of rows. This is reflected in the posterior-medial barrel subfield of primary somatosensory cortex by a similar spatial pattern of columns, one column for each whisker (Figure 3). These are best defined in layer 4 where the thalamic afferents terminate. However, specialization to a single whisker is also apparent in higher and lower layers, and owing to their three-dimensional shape these columns were dubbed 'barrels'. The number and layout of these barrels can be altered by manipulations of the sensory periphery, such as removing a whisker.

COLUMNS IN OTHER SYSTEMS

The primary auditory cortex (A1) of animals such as cats and bats shows a systematic, spatially distributed representation of several independent auditory stimuli (reviewed in Schreiner, 1995). However, these auditory maps appear somewhat disordered because the local scatter of receptive field properties can vary over a wide range. The most regular map is that of preferred frequency, organized along a tonotopic axis without gradient reversals that mimics the tonotopic organization of the cochlea. Orthogonal to this axis, no systematic change of the preferred frequency is observed, with neurons being arranged along isofrequency contours. Other response parameters vary along the

isofrequency contours in a systematic way, such as the bandwidth and shape of tuning curves. Further maps also appear to be represented in a columnar fashion, such as the coding of intensity and sound localization, but the details of their organization are still unclear. (*See* **Audition, Neural Basis of**)

Inferotemporal (IT) cortex is a visual area essential for object perception and recognition. Using intrinsic signal imaging and extracellular recording in macaque monkeys, Tsunoda *et al*. (2001) showed that the neural activity evoked in IT by complex objects is laid out spatially as distributed patches. This result suggests that an object is represented by a combination of cortical columns, each of which represents a visual feature. However, not all the columns related to a particular feature were necessarily activated by the original objects. Thus, objects would be represented by using a variety of combinations of active and inactive columns for individual features, rather than simply by the addition of feature columns. It is unclear, though, whether an object is represented by a combination of modules, each specific to a visual feature or a part of the object (feature-based or part-based representation), or whether modules are specific to the object (object-based representation). (*See* **Temporal Cortex**; **Object Perception, Neural Basis of**; **Vision: Object Recognition**)

Columnar maps also exist outside the cortical areas, such as in the brainstem (maps of interaural delay, of interaural intensity difference, and of auditory space) and the superior colliculus (map of motor space, or gaze direction). (*See* **Motor Areas of the Cerebral Cortex**)

INTRACOLUMNAR AND INTERCOLUMNAR CIRCUITRY

Cortical columns are also distinguished from each other by their patterns of circuitry. The majority of intracortical circuits are local, connecting neurons within the same columns, with only a minority of connections being between columns. Again, this organization has been most extensively studied in the visual system.

Callaway (1998) proposed a generic model of vertical connectivity linking layers within a column in primary visual cortex of cats and monkeys (Figure 4). The model is based on three simplifying assumptions: only excitatory synapses are considered; each cortical layer provides its primary output to only one layer; and only two types of connections are considered (feedforward and feedback). A direct path from inputs (coming from the LGN) to outputs (going mainly to other areas in the cortex) passes through cortical layers 4C and 2–4B, with layers 6 and 5 providing feedback (modulatory) connections, respectively. Since these dense connections are mostly confined within a column, this provides a great deal of purely intracolumnar – and therefore local – information processing. Long-range connections (generally up to a few millimeters long) between columns mostly project within layer 2/3. They are generally sparse and patchy, and tend to connect spatially separated columns with the same feature preference, such as the same orientation or ocular dominance preference (although some recent experiments do not fully agree with this cluster-like connection pattern). It is easy to imagine how such specific patterns could arise as a result of Hebbian learning, since columns with similar feature preferences would be expected to have highly correlated activity. Likewise, it has been suggested that color blobs are preferentially linked to color blobs of the same ocular dominance, and interblobs to interblobs.

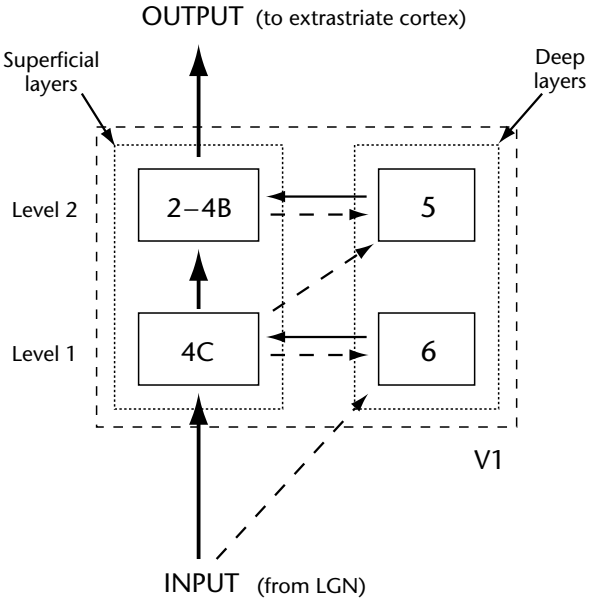


Figure 4. Two-level model of local cortical circuitry in V1. A direct path from input (from the lateral geniculate nucleus, LGN) to output (to extrastriate cortex) is provided by the two feedforward superficial layers 2–4B and 4C; feedback deep layers 5 and 6 modulate the activity of each level. Adapted from Callaway (1998).

COLUMNAR DEVELOPMENT AND COMPUTATIONAL MODELS

The high degree of order displayed by columnar structures and the large amount of data acquired,

especially regarding the development of ocular dominance columns in primary visual cortex, has inspired several computational models of columnar development. Such models are useful to explain the processes at work as well as to produce predictions that can guide future experiments. They should also be able to account for interspecies variations and be generalizable to models for other areas of the cortex, assuming that the underlying mechanisms of cortical development are reasonably universal. An excellent review of such developmental models can be found in Swindale (1996). (*See* **Neural Development**; **Neural Development, Models of**)

Most models of visual cortical development are based on the following assumptions (which are partially supported by experimental data): patterned retinal activity in the afferents to cortical neurons; Hebb synapses; radially symmetric, short-range excitatory and long-range inhibitory lateral cortical connections; and normalization of synapse strength. Thus, most of these models largely disregard genetic factors and assume that the columns in the primary visual cortex appear during development from an apparently uniform cortical sheet by a process of activity-dependent self-organization that modifies synaptic strengths in response to patterns of visual stimulation. These patterns can be produced both externally by the world, and generated internally by spontaneous activity in the retina (Meister *et al*., 1991). The rule by which synaptic strengths appear to change is roughly the one proposed by Hebb (1949): 'neurons that fire together wire together'. The models often represent the cortex as a two-dimensional array of neural units (each representing a collection of real neurons) and thus directly embody the definition of column. The visual stimulus is represented either in an abstract, low-dimensional way, as a vector of independent components representing ocular dominance, orientation preference, retinotopic position or direction preference; or in a concrete, high-dimensional way, as a vector containing the connection strengths between a cortical cell and a set of receptor cells in the retina. (*See* **Hebb Synapses: Modeling of Neuronal Selectivity**; **Hebb, Donald Olding**)

A common characteristic of these models is that they try to maximize coverage as well as continuity, as originally suggested by Hubel and Wiesel. Coverage refers to the fact that all combinations of eye and orientation preference occur at least once within any region (of a certain, small size) in stimulus space – otherwise, the animal might be blind to the unrepresented stimulus (although it has been suggested that higher cortical areas could interpolate between incomplete representations in lower cortical areas). Continuity refers to the fact that the preferences of neighboring neurons in cortex tend to be similar. Representing a high-dimensional stimulus space in a two-dimensional cortex results in coverage and continuity competing at the expense of each other, with the striped organization observed being perhaps a locally optimal solution to their trade-off. (*See* **Vision: Occlusion, Illusory Contours and 'Filling-in'**)

Two particularly important types of models are correlational (e.g. Miller *et al*., 1989) and competitive (e.g. Goodhill, 1993). In correlational models the input–output function of neurons is linear, and receptive field development is driven by the eigenvectors of an operator dependent on the correlation of the input patterns, the intracortical connections, and the LGN arborization. In competitive models the input–output function of neurons is highly nonlinear, and such models implement something more akin to cluster analysis. Generally speaking, these models account for much of the observed phenomenology of cortical maps, including the striped structure of ocular dominance and orientation columns with the appropriate periodicity and interrelations, and the location of pinwheels and fractures. However, no model so far can account for all observed features for both ocular dominance and orientation maps at the same time, or for some of the more elusive data. (*See* **Pattern Recognition, Statistical**; **Receptive Fields**)

WHY A COLUMNAR ORGANIZATION?

The presence of a columnar organization in various regions of the cortex of many mammalian species has suggested that columns form the basic information processing elements of the cortex, with each column being responsible for analyzing a small range of stimuli, and the same modular unit being repeated multiple times to span the entire range of stimuli (e.g. Szentágothai, 1978). As such, columns have been considered to be a fundamental functional feature important for perception, cognition, memory, and even consciousness (Szentágothai, 1978; Eccles, 1981). However, there is no general agreement about the reason for the existence of such groupings. Such columnar structure has not been found in some mammalian species closely related to other species that do have columns (Purves *et al*., 1992). Thus, it has been argued that the columnar organization of the cortex may not imply a functionally modular organization (Swindale, 1990; Purves *et al*., 1992). In particular,

Purves *et al*. suggested that the production of iterated patterns of circuitry might be an incidental consequence of the activity-dependent elaboration of synaptic connections and be of little significance to cortical function. In other words, a given cortical system might work just as well if columns did not form. Purves *et al*. suggest several factors that could drive such origin. (*See* **Modularity in Neural Systems and Localization of Function**; **Synaptic Plasticity, Mechanisms of**)

CONCLUSION

Many areas of the cortex, particularly in the visual and somatosensory system, can be divided into repeating modules characterized by discrete patterns in both function and anatomy. The best-studied examples are 'barrels' and touch-modality columns in primary somatosensory cortex, and orientation and ocular dominance columns in primary visual cortex. There are many vertical connections linking neurons within a column, and a few horizontal connections linking different columns. Columnar development may be driven by activity-dependent self-organization, and can often be modeled using Hebbian learning rules – although the relative importance of genetic factors and patterned activity is not clear. As yet no compelling justification has emerged for why columnar structure exists. (*See* **Vision, Early**; **Pattern Vision, Neural Basis of**; **Cortical Map Formation**)

References

Callaway EM (1998) Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience* **21**: 47–74.

Crowley JC and Katz LC (1999) Development of ocular dominance columns in the absence of retinal input. *Nature Neuroscience* **2**: 1125–1130.

Eccles JC (1981) The modular operation of the cerebral neocortex considered as the material basis of mental events. *Neuroscience* **6**: 1839–1856.

Favorov OV and Diamond ME (1990) Demonstration of discrete place-defined columns – segregates – in the cat SI. *Journal of Comparative Neurology* **298**: 97–112.

Goodhill GJ (1993) Topography and ocular dominance: a model exploring positive correlations. *Biological Cybernetics* **69**(2): 109–118.

Hebb DO (1949) *The Organization of Behaviour*. New York, NY: John Wiley.

Horton JC and Hocking DR (1996) Intrinsic variability of ocular dominance column periodicity in normal macaque monkeys. *Journal of Neuroscience* **16**: 7228–7339.

Hubel DH and Wiesel TN (1977) Functional architecture of the macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B* **198**: 1–59.

Meister M, Wong ROL, Baylor DA and Shatz CJ (1991) Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science* **252**: 939–943.

Miller KD, Keller JB and Stryker MP (1989) Ocular dominance column development: analysis and simulation. *Science* **245**: 605–615.

Mountcastle V (1957) Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology* **20**: 408–434.

Purves D, Riddle DR and LaMantia AS (1992) Iterated patterns of brain circuitry (or how the brain gets its spots). *Trends in Neurosciences* **15**(10): 362–368.

Schreiner CE (1995) Order and disorder in auditory cortical maps. *Current Opinion in Neurobiology* **5**: 489–496.

Sharma J, Angelucci A and Sur M (2000) Induction of visual orientation modules in auditory cortex. *Nature* **404**: 841–847.

Swindale NV (1990) Is the cerebral cortex modular? *Trends in Neurosciences* **13**(12): 487–492.

Swindale NV (1996) The development of topography in the visual cortex: a review of models. *Network: Computation in Neural Systems* **7**(2): 161–247.

Szentágothai J (1978) The neuron network of the cerebral cortex: a functional interpretation. *Proceedings of the Royal Society of London, Series B* **201**: 219–248.

Tsunoda K, Yamane Y, Nishizaki M and Tanifuji M (2001) Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience* **4**: 832–838.

Woolsey TA and van der Loos H (1970) The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. *Brain Research* **17**(2): 205–242.

Further Reading

Erwin E, Obermayer K and Schulten K (1995) Models of orientation and ocular dominance columns in the visual cortex: a critical comparison. *Neural Computation* **7**: 425–468.

Hubel DH (1995) *Eye, Brain, and Vision*. New York, NY: WH Freeman.

Jones EG and Diamond IT (eds) (1995) *The Barrel Cortex of Rodents*, vol. 11 of *Cerebral Cortex*. London, UK: Plenum Press.

Nicholls JG, Martin AR and Wallace BG (2000) *From Neuron to Brain: A Cellular and Molecular Approach to the Function of the Nervous System*, 4th edn. Sunderland, MA: Sinauer.

Peters A and Rockland KS (eds) (1994) *Primary Visual Cortex in Primates*, vol. 10 of *Cerebral Cortex*. London, UK: Plenum Press.

Cortical Map Formation

Intermediate article

Jon H Kaas, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

Cortical maps are orderly representations of sensory receptors or body movements that form during the development of the brain. Proper development depends on genetic information, molecular signals based on relative position in the brain, and information in spontaneous and evoked neural activity patterns.

INTRODUCTION

The functional machinery of the brain includes cortical areas and subcortical nuclei that are interconnected to form systems. Many of these nuclei and cortical areas systematically represent a sensory surface or the motor control of muscles. In the visual system, the inputs from the retina terminate in topographic patterns in thalamus and midbrain nuclei to form maps of the retina or visual space. The map in the laminated lateral geniculate nucleus in the thalamus projects to primary visual cortex, V1, to form a map of the hemiretinas of the two eyes that see the contralateral half of the visual world. Area V1 projects in turn to several additional visual areas to activate further maps of the contralateral visual hemifield. Monkeys and humans have over 25 visual areas, most of which can be described as containing maps of the contralateral visual hemifield (Kaas, 1989). The interconnected maps form a processing hierarchy for visual information in which the early maps, especially the one in V1, most precisely map the visual hemifield, while higher-order representations progressively become less retinotopic. However, these maps may represent other aspects of visual information in systematic ways, as they start to reflect more of the visual outcomes of cortical computations. The collections of interconnected cortical and subcortical maps constitute the visual system, and the auditory and somatosensory systems are constructed similarly.

The motor system is defined somewhat differently. Motor cortex includes a primary area and other areas where neural activity at any specific location in the map elicits a specific movement or set of movements. The movements can be elicited by natural neural activity or by neural activity evoked by focal electrical stimulation within these areas. By electrically stimulating a large number of sites across a motor area, the representation of movements can be experimentally revealed. For example, primary motor cortex, a mediolateral strip of cortex in the caudal portion of the frontal lobe, represents foot movements medially near the brain midline, trunk movements more laterally, forearm and hand movements next, and face movements most laterally (Penfield and Boldry, 1937). As motor areas also receive rather indirect sensory inputs, they can be said to represent somatosensory or other sensory inputs, but these sensory maps are seldom discussed. Likewise, movements can be evoked from somatosensory areas of cortex, at higher levels of stimulating current, and these somatosensory areas can also be said to contain motor maps. Monkeys and humans have as many as ten areas in each cerebral hemisphere that map movements of the muscles of the contralateral half of the body, as well as two main areas for moving the eyes and directing gaze into the contralateral visual hemifield. Because the internal representational organizations of the sensory and motor areas can be revealed by successively recording from neurons or stimulating neurons in many locations in each representation with microelectrodes, the recording or stimulating procedures are sometimes referred to as 'mapping' the brain.

In addition to recording or stimulating with microelectrodes, there are other ways of recording or imaging brain activity in response to sensory stimuli or repeated movements that also allow the organizations of brain maps to be revealed. The use of 'noninvasive' imaging procedures has revealed the locations and internal organizations of visual, auditory, somatosensory, and motor maps in humans. The existence of orderly maps in the

cortex has been implied for some time by the nature of perceptual and movement deficits that follow focal lesions of regions of cortex in human patients. Motor maps were directly revealed in the 1870s when investigators exposed the surface of the brains of dogs and monkeys and described regions where body and eye movements could be evoked by electrical stimulation. Comparable evidence for sensory maps came later in the 1930s and 1940s when the invention of the oscilloscope and amplifiers made it possible to record the weak electrical activities of cortical neurons as they responded to sensory stimuli. These recording and stimulation methods have long been applied to humans when their brains were exposed during surgery. With the existence of such maps now well established, the interesting and challenging question of how they emerge in the developing brain can be addressed.

THE DEVELOPMENT OF CORTICAL MAPS

The primary sensory and motor maps occupy the same relative positions in the neocortex of most mammalian species. Thus, the visual cortex is at the back, auditory cortex is lateral, somatosensory cortex is toward the front, and motor cortex is just ahead of somatosensory cortex. There are several theories as to how this happens.

One theory is that the subdivision of the cortex into areas begins early in brain development, well before the neocortex has even been generated by the migration of cells from a generative zone along the cerebral ventricles. According to this theory, cellular interactions occur in the pool of dividing cells along the ventricle so that the cells become committed to certain fates and collectively form an early plan of the overall organization of neocortex. When the cells migrate in a point-to-point fashion to cortex, they carry with them the basic organization of where emerging cortical maps are relative to each other and even the internal organizations of maps. Thus, the basic organization of cortex is determined by interactions between progenitor cells before they even generate the cells that form cortex. This theory has been called the *protomap hypothesis* (Rakic, 1988), and it suggests that there is something inherent within each emerging cortical region that directs the region to become a specific cortical area.

A second view is that neocortex is generated as a *uniform sheet* and that patterns of inputs from the thalamus and later the emerging cortical connections subdivide the cortex into areas. Thus, each region of cortex starts out with no particular identity, but it acquires an identity when axons from another part of the brain arrive and tell it how to differentiate. The problem with the view that regions of cortex have no particular identity is that this does not explain by itself why visual cortex is always located toward the back and motor cortex toward the front of the neocortex. One possibility would be to utilize well-known differences in the front to back and lateral to medial neurogenic gradients in the emerging neocortex and thalamus to form patterns or connections. Axons could grow out in maturational order to arrive at cortical targets in maturation order, providing an overall organization in cortex. If axons maintained neighborhood relations with each other as they grew to cortex, as they largely seem to do, considerable order in cortex could result. Thalamic axons, according to this possibility, only need to be attracted to cortex, and the sorting mechanism would be a consequence of positional differences in developmental order, which in turn would be dependent on position effects on gene expression and the consequent molecules. Refinement of the initial order and the addition of local circuits in cortex could then follow.

Alternatively, and most likely, some sort of *positional signaling* is used. An early hypothesis of how growing axons reach and recognize an appropriate target was the chemoaffinity theory proposed by Roger Sperry in the 1960s (Sperry, 1963). According to this proposal, both the guidance of axons to the target region and the recognition of the appropriate target neurons are achieved by the operation of highly specific chemical affinities between substances in the growing axons and in the neurons of the target structure. Such a proposal nicely accounts for the formation of highly specific patterns of connections, but it fails to explain all observations. Most notably, lesions of neocortex very early in development (before the thalamic axons arrive) do not completely abolish some divisions of cortex and leave others intact; instead, smaller than normal areas form on the smaller sheet of neocortex. This type of observation is more compatible with a modified form of Sperry's chemoaffinity theory.

One proposal is that molecular gradients across developing neocortex indicate a general front to back and medial to lateral direction for patterns of thalamic connections, but not specific locations for connections (Fukuchi-Shimogori and Grove, 2001). Thus, axons from the visual thalamus would seek a high (or low) region of expression in the gradient and always would grow to the back of the neocortex (Huffman *et al*., 1999). If the back part of

neocortex had been removed, these axons would still grow to the back of the portion of neocortex that remained. Local competition between axons guided to positions along a molecular gradient would then lead to winners and losers in the reduced cortical sheet, but not all of the visual axons would be losers. In contrast, if appropriate target neurons were completely prespecified either in a map in the generative cells along the ventricle or in a similar protomap formed in the early cortical sheet, then removal of the target cells would completely abolish a cortical area. In principle, chemical gradients in cortex could provide the signals, given adjustable thresholding mechanisms and competition for synaptic space, for both cortical areas emerging in the correct places with the correct thalamic inputs, and at least approximately correct topographic patterns within cortical areas. Thus, molecular patterns intrinsic to cortex guide the development of cortex while allowing several outcomes. A pluripotential cortical sheet is differentiated by molecular gradients.

All cortical areas may not be specified in the same way. Possibly a few cortical areas, especially the primary areas, are specified very early in development and can only become those areas, although this does not seem to be the case for the visual cortex of opossums or somatosensory cortex of rats, where the outcomes have been experimentally altered. However, an argument has been made for primary visual cortex of monkeys and humans, as this cortex has more neurons across the thickness of cortex than other cortical areas, and the posterior ventricular zone that generates neurons for primary visual cortex generates more neurons. These observations suggest that the protomap of primary visual cortex already exists in the ventricular generative zone. Yet a major role in the specification of visual cortex by thalamic inputs is clearly indicated. In humans in whom the eyes fail to develop and in monkeys with eye removal early in fetal life, many of the neurons in the developing lateral geniculate nucleus of the thalamus degenerate without visual input, and the number of thalamic neurons projecting to visual cortex is greatly reduced. As a result the primary visual cortex is only a fraction of its usual size (Rakic, 1988). Some of the cortex that would normally become primary visual cortex develops features of nonprimary visual cortex. Thus, it appears that the region of primary visual cortex becomes primary visual cortex only if it receives an adequate input of visual axons from the thalamus.

The internal organization of cortical maps probably depends on a balance of intrinsic factors leading to regional differences in gene expression and gene products, and extrinsic factors including regional and local differences in neural activity patterns evoked by sensory stimuli, self-movement, or by correlated spontaneous activity. A combination of such molecular and activity-dependent mechanisms would best account for the basic features of cortical map development listed below.

- At least the crude topographic features of cortical maps appear to develop independently of any information from sensory receptors and afferents. In other words, the basic features of cortical maps develop without instruction from the receptor sheet. Activity patterns relayed from the periphery are not necessary. The most compelling evidence for this comes from studies of thalamocortical connections in genetically eyeless (anophthalmic) mice where the connections between the lateral geniculate nucleus of the visual thalamus and primary visual cortex are at least approximately normal in topography. Similar results have been reported in other mammals reared after the removal of retinal afferents early in development. Also, the basic pattern of connections between the ventroposterior nucleus of the somatosensory thalamus and primary somatosensory cortex of rats appears to form before the thalamus is activated by somatosensory inputs. The topographic order of the thalamocortical pattern is said to develop independently of a 'template' of the periphery.
- Detailed cortical maps of the receptor sheets often develop prenatally, and thus they develop without postnatal experience. Newborn mammals of many species have orderly normal maps of the body surface in primary somatosensory cortex at birth, and in newborn lambs and monkeys the retinotopic organization of primary visual cortex appears to be normal at or soon after birth.
- The internal organizations of sensory maps closely reflect the features of the sensory sheet. The maps reflect the densities of receptors across the skin or retina so that the maps are proportional to receptor densities rather than skin or retinal territories, although a behaviorally important receptor surface may achieve extra space in cortical maps. Even errors in the arrangement of receptors or the projections of afferents are reflected in the order of cortical maps. For example, the number and arrangement of the vibrissae on the face of rats and mice is highly consistent across individuals, and the maps in primary somatosensory cortex (S1) separately and precisely represent each whisker. When rare individuals with an extra whisker or two are examined, an extra anatomical module in S1 is found for each of the extra whiskers. Likewise, S1 of star-nosed moles with an array of 11 ray-like sensory appendages protruding from each side of the nose has 11 modules or bands in the face portion of S1, one for each ray. However, rare individuals who develop with 10 or 12 rays have 10 or 12 bands in S1 (Catania and

Kaas, 1997). Similarly, Siamese cats have a mutant gene that alters skin and eye pigment and also changes the projection pattern from the retina somehow so that the segment of the retina that projects to the contralateral lateral geniculate nucleus in the thalamus is extended by some 20° of visual angle. This extra input is often incorporated into the order of cortex maps so that primary visual cortex (V1) represents an extra 20° of visual space in a single retinotopic pattern (Kaas and Guillery, 1973).

- There is evidence from the study of rats that the segregation of afferents from the two eyes, or from different whiskers on the face, in sensory cortex depends on having intact inputs from the eyes or face in early development. In rats reared after section of sensory nerves at birth, the cortical arbors of thalamocortical neurons are unusually large and overlapping, with less focused distributions of synapses. In visual cortex, inputs relayed from the two eyes normally divide the space in visual cortex into alternating ocular dominance bands. If one eye is removed early in development, thalamocortical axons related to the remaining eye occupy nearly all of cortex. Similarly, if the activity of retinal neurons in one eye is simply reduced by rearing a cat or monkey with one eye sutured shut, the thalamocortical axons related to the normal eye acquire most of the cortical space (Hubel *et al*., 1977).
- The development of cortical maps is altered if they are abnormally innervated. Several investigators have induced the sensory projections from one sensory system to innervate a thalamic nucleus for another sensory system early in development, thereby causing the sensory map in cortex to receive the wrong sensory input. When visual inputs from the retina were induced to innervate the medial geniculate nucleus of the auditory thalamus, the medial geniculate neurons thereby relayed visual rather than auditory information to primary auditory cortex (A1). This cortex then developed a map of the retina that had some, but not all, of the features of visual cortex. The researchers concluded that the nature of the sensory input, visual or auditory, influenced, but did not totally determine, cortical map development (Roe *et al*., 1990).

MOLECULAR MECHANISMS OF CORTICAL MAP FORMATION

Regional differences in the expression of molecules in the developing cortex appear to have major consequences for the course of cortical development. Many molecular correlates of developmental stages and features have been described, including the expression of nerve growth factors and receptors for those factors, growth-associated proteins, structural proteins associated with synapses, membrane-associated glycoproteins, and axon guidance molecules. The underlying question concerns the factors that relate to differences in gene expression allowing various signaling and neuron construction molecules to be expressed at the proper time and place in the developing neocortex.

According to the early views of chemospecificity postulated by Roger Sperry, molecular signals – either regionally expressed or expressed as gradients across developing cortex – would at least set up the basic organization of cortex by guiding incoming connections to appropriate locations and maintaining synaptic contacts at those locations. Other local features of cortical organization might then emerge as a result of gene expression related to the neural activity patterns induced by the initial connections. The key assumption here is that the early regional differences in the expression of molecular signals depend on factors intrinsic to developing cortex.

Important evidence on intrinsic differences in the expression of molecular signals comes from studies of mutant mice in which thalamocortical afferents failed to develop. Despite a total lack of thalamic inputs, the emerging neocortex demonstrated both graded and sharp patterns of gene expression, including genes thought to be important in axon adhesion and in neuronal differentiation. Thus, developing neocortex has intrinsic signals that seem to be capable of directing regional differentiation. Further evidence comes from studies of abnormal thalamic connections in mutant mice lacking certain regulating genes. In these mice, the early expression of specific molecular markers are shifted toward the back in the developing cortex. As a result, thalamic connections are also shifted toward the back. For example, the somatosensory inputs that normally go to the middle of neocortex terminate instead in the back where visual inputs normally terminate. The visual inputs in turn are displaced to the very margin of the posterior neocortex. Thus, these experiments provide compelling evidence for the existence of intrinsic molecular positional cues for incoming thalamic axons.

In summary, it appears that differences in gene expression in cortical neurons based on their position in the cortex or their previous history as precursor cells in the ventricular generative zone provide the molecular guidance necessary for at least the gross pattern of thalamocortical connections to develop. This pattern includes a basic set of cortical sensory and motor areas, and at least the crude topographic order of the sensory and motor maps within those areas. However, refinements of crude patterns within maps appear to depend on information carried from the receptor sheet to the thalamus and then to cortex.

ROLE OF EXPERIENCE AND NEURAL ACTIVITY

The prevailing view is that cortical maps emerge in development as a result of a combination of mechanisms that require neural activity and those that do not. The activity-independent mechanisms are thought to be important in forming cortical areas and the crude topography of cortical areas, while activity-dependent mechanisms are thought to be essential in achieving precision in the connection patterns that make up cortical maps.

The early evidence of a role for activity in the formation of cortical maps came from the landmark experiments of David Hubel and Torsten Wiesel. In the early 1960s these investigators experimentally sutured together the lids of one eye in kittens whose eyes had not yet opened, thus depriving that eye of pattern vision. When the visual system of these cats was examined after they matured, few neurons in primary visual cortex were responsive to the previously closed eye and the anatomical relay of connections from the deprived eye to visual cortex were sparse. The interpretation of these results was that axons from neurons in the lateral geniculate nucleus of the thalamus that were activated by the deprived eye were less active than those related to the normal eye, and the two types of axons were in competition with each other for synaptic space on cortical neurons. The more active axons for the normal eye grew and took over most of the cortical neurons, while the axons for the deprived eye lost contacts with cortical neurons and failed to grow with the expanding brain and innervated little of visual cortex.

Subsequently there have been many deprivation studies that demonstrate that it is possible to alter the development of the visual system. Such results provide further evidence for activity-dependent mechanisms. Nevertheless, the roles of activity and especially experience continue to be questioned. For example, the separate alternating territories in primary visual cortex with inputs from one or the other eye, the ocular dominance columns, develop before birth in many mammals. Thus, they do not depend on pattern vision. However, the development of these columns has been attributed to spontaneous or nonvisual activity in the neurons of the two eyes before they are exposed to light. Neurons close together in any structure, including the eye, tend to be interconnected and spontaneously active at the same time. Thus the patterns of activity relayed to the thalamus from each developing eye would differ, and this difference would be relayed to cortex as a basis for segregating axons related to each eye in ocular dominance columns. Evidence suggests that such segregations of thalamocortical terminations develop even without eyes. However, this observation, if valid, does not eliminate the possibility of spontaneous activity being essential, as differences in spontaneous activity patterns could occur in the layers of the lateral geniculate nucleus normally related to each eye.

CONCLUSION

Over years of active investigation, a general theory of cortical map development has emerged. In brief, a few major factors appear to be highly important for the formation of map topology, modular organization within maps, and neuron response properties.

First, molecular mechanisms that are independent of external influences seem to be responsible for guiding axons to their approximate locations in cortex. Which molecules govern this guidance and establishment of a crude topographic order for connections remains largely uncertain, but a number of promising candidate substances are under investigation. Another uncertainty is how these molecules come to be expressed in regionally specific patterns. When and how are position cues recognized? Spatiotemporal gradients in neuron generation and maturation, and substrate cues for gene expression, are likely to be important. Several guidance and chemoaffinity factors have been demonstrated.

Neural activity patterns provide another important source of information, and this information is used to select some synaptic contacts over others. Neurons are induced by activity patterns to locally grow and form more connections, or to retract and lose connections. Much research supports the premise that, in both the mature and developing nervous system, synapses on a neuron that are active while the neuron itself is active are strengthened, while those that are inactive during such discharges are weakened. The physiological consequences of such changes in synaptic effectiveness, which are rapidly induced, have been called *long-term potentiation* (LTP) and *long-term depression* (LTD). The changes are often referred to as 'Hebbian' plasticity after the early proposal of cellular mechanisms for learning by Donald Hebb.

The key variable is whether overlapping inputs on cortical neurons are active at the same time. Activities that are correlated in time across receptors in the receptor sheet are based on proximity and receptor transducing factors. The same stimuli

are likely to activate nearby receptors, but classes of receptors in the same location are differentially responsive to the same stimulus. In the central nervous system, proximity is additionally important because of local interconnections, and local computations add to response diversity by producing neurons that have new selectivities for features of sensory stimuli. Selections of synapses based on correlated and discorrelated activities fine-tune the internal representational order of sensory maps so they more closely reflect the proximities and disjunctions on the receptor sheets. Such selection also distributes and segregates inputs differing in correlated spontaneous activity based on proximity levels. Neurons projecting from one structure to the next would be more densely interconnected if adjacent than widely separate, and the local interconnections would induce correlations in activity. Thus, activity patterns could promote the formation of topographically matched connections between cortical areas. Neurons driven by different receptor classes would be separately activated and their overlapping projections would segregate. Similarly, neurons that became selective for different stimulus attributes, as a result of the computations of central neurons, would develop segregated projections. Thus, activity patterns would induce sharply defined modules such as ocular dominance columns, stimulus orientation modules, and light-on or light-off layers in cortical maps.

Early-maturing maps, largely the primary cortical areas, would be most influenced by prenatal activity patterns, especially correlated spontaneous activity but also sensory responses evoked by self-generated movements, while later-maturing maps would have the opportunity to use information from a more complex external environment and the processing outcomes of early maturing cortical areas and thalamic nuclei.

Finally, the susceptibility of neurons to both local substrate cues in the molecular environment and to the effects of neural activity patterns changes over the course of maturation, leading to the concept of critical or susceptible periods for developmental change.

References

Catania KC and Kaas JH (1997) The mole nose instructs the brain. *Somatosensory and Motor Research* **14**: 56–58.

Fukuchi-Shimogori T and Grove EA (2001) Neocortex patterning by the secreted signaling molecule FGF8. *Science* **294**: 1071–1074.

Hubel DH, Wiesel TN and LeVay S (1977) Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London Series B* **278**: 377–409.

Huffman KJ, Molnai Z, VanDellen A *et al*. (1999) Formation of cortical fields on a reduced cortical sheet. *Journal of Neuroscience* **19**: 9939–9952.

Kaas JH (1989) Why does the brain have so many visual areas? *Journal of Cognitive Neuroscience* **1**: 121–135.

Kaas JH and Guillery RW (1973) The transfer of abnormal visual field representations from the dorsal lateral geniculate nucleus to the visual cortex in Siamese cats. *Brain Research* **59**: 61–95.

Penfield W and Boldry E (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* **60**: 389–443.

Rakic P (1988) Specification of cerebral cortical areas. *Science* **241**: 170–176.

Roe AW, Pallas SL, Hahm JO and Sur M (1990) A map of visual space induced in primary auditory cortex. *Science* **250**: 818–820.

Sperry R (1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Science of the USA* **50**: 703–710.

Further Reading

Kaas JH (1988) Development of cortical sensory maps. In: Rakic P and Singer W (eds) *Neurobiology of Neocortex*, pp. 115–136. New York, NY: John Wiley.

Kaas JH (1997) Topographic maps are fundamental to sensory processing. *Brain Research Bulletin* **44**: 107–112.

Kaas JH (2000) Organizing principles of sensory representations. In: Bock G and Cardew G (eds) *Evolutionary Developmental Biology of the Cerebral Cortex*, Novartis Foundation Symposium 228, pp. 188–205. New York, NY: John Wiley.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* **274**: 1133–1138.

Kennedy H and Dehay C (1993) Cortical specification of mice and men. *Cerebral Cortex* **3**: 171–186.

Krubitzer L and Huffman KJ (2000) Arealization of the neocortex in mammals: genetic and epigenetic contributions to the phenotype. *Brain, Behavior and Evolution* **55**: 322–335.

Levitt P (2000) Molecular determinants of regionalization of the forebrain and cerebral cortex. In: Gazzaniga MS (ed.) *The New Cognitive Neuroscience*, pp. 23–32. Cambridge, MA: MIT Press.

O'Leary DDM (1989) Do cortical areas emerge from a protocortex? *Trends in Neurosciences* **12**: 401–406.

O'Leary DDM, Schlaggar BL and Tuttle R (1994) Specification of neocortical areas and thalamocortical connections. *Annual Review of Neuroscience* **17**: 419–439.

Shatz CJ (1990) Impulse activity and the patterning of connections during CNS development. *Neuron* **5**: 745–756.

Wiesel TN (1982) Postnatal development of the visual cortex and the influence of environment. *Nature* **299**: 583–591.

Wong ROL (1999) Retinal waves and visual system development. *Annual Review of Neuroscience* **22**: 29–47.

Decoding Neural Population Activity

Introductory article

Peter Foldiak, University of St Andrews, St Andrews, UK

CONTENTS

The brain uses the activity patterns of a large number of neurons to represent information about the world. Decoding methods allow us to read out this neural code and interpret the neural activity pattern.

INTRODUCTION

Brain activity can be observed using a variety of technical methods. Functional imaging techniques such as positron emission tomography and functional magnetic resonance imaging allow us to observe which brain areas show increased activity during the performance of various tasks. However, these methods have relatively poor resolution in space and time, which means that we can only observe average brain activity over many thousands of neurons and over many consecutive brain states. This prevents us from seeing functionally important details in brain activity. Yet we need to understand more than just which areas are involved in certain tasks; we are interested in the computations within these areas. As much of the interesting difference between brain states is visible on the level of much smaller groups of neurons and on a faster time scale, we need more detailed indicators of brain activity. In fact, the detailed response properties of individual neurons even within a small area of the brain seem to be substantially different from each other, which means that if we want to understand information processing in the brain, we need to observe the activity of single cells. Single cell recording is the only technique that gives us the precise times of each action potential (a 'spike train') from an isolated neuron.

In the sensory system, we try to make sense of these spike trains in terms of the sensory input stimuli, while in the motor system we try to understand the relationship between neural spikes and the resulting movements. One view of this relationship is captured by the concept of neural 'tuning curves'. It assumes that most of the information in the spike train can be summarized by a single number, the neural response. This number is often thought to be the firing rate, which is the number of action potentials generated by a neuron during a time interval, divided by the length of that interval. Other aspects of the spike train, such as the more precise timing of the spikes in relation to each other or to the onset of a stimulus, may also be carriers of information. A tuning curve is the (average) response of the neuron as a function of certain properties or parameters of the stimulus input or motor output. An example would be the orientation tuning curve of neurons in primary visual cortex, where the stimulus is a grating of equally spaced parallel dark and bright bars (or two-dimensional sine wave grating), and the neuron responds differently to gratings of different orientations. The orientation tuning curve plots the response as a function of the grating orientation, showing a peak at the 'preferred' orientation of the neuron and lower values elsewhere. A more complete description of the neural code also includes a measure of the response variability, as the neuron's response can be different on each stimulus repetition depending on unpredictable external and internal factors (e.g. stimulus and transduction noise, physiological state, attention, neural interactions).

The other, complementary way of asking questions about the relationship between the spike train and the stimulus (or movement) is by trying to identify the stimulus or movement based on the observed response from a neuron or responses from several neurons. This is the problem of decoding.

POPULATION VECTOR METHOD

A simple method of decoding the response of a neural population is based on the peaks of tuning

curves. It was originally developed for predicting arm movements based on the responses of a population of neurons in motor cortex but it can also be applied to populations of sensory neurons to decode a sensory stimulus. Essentially it is a response-weighted sum of the preferred values of the encoding neurons. Each neuron has a vector associated with it that points in the direction that corresponds to the peak of its tuning curve. In the case of arm movements, the direction of this vector is the direction of the arm movement in physical space that maximized the neuron's response. The direction of this vector stays fixed but the length of this vector for each neuron is scaled at each moment to the size of the neuron's response. The population vector at any moment is the sum of the vector contributions of the individual neurons in the population (Figure 1). Under some assumptions about the shapes of the tuning curves and the distributions of their peaks, the population vector will be a reasonable prediction of the stimulus or movement parameters (e.g. the direction of arm movement). To understand this procedure, imagine that each neuron represents a single 'preferred' direction and the strength of each neuron's response indicates how close the actual direction is to the neuron's preferred direction. As the most active neurons will contribute the longest vectors to the vectorial sum, the neurons with the preferred directions closest to the actual direction will have the greatest weight in determining the resulting population vector. This weighted sum is then an estimate of the actual direction, and the estimate should get better as the number of neurons in the population increases. While the population vector method is relatively simple to apply, there are several problems with it:

- Neurons in a population do not generally fulfill the condition that the peaks of the tuning curves of the neurons in the population have to be uniformly distributed in all directions (i.e. roughly the same number of neurons have to have peaks in each direction). If there is an excess of peaks in a certain direction, the population vector will be biased in that direction.
- It assumes a specific form of the tuning curve (falling off as a cosine function from the peak). In many cases, these assumptions are not good approximations of the actual tuning curves and their distribution in the population. The only information it uses about the actual tuning curve is the location of its peak, and we have no way of incorporating any additional information (e.g. breadth of tuning, or multiple peaks) we might have about the true tuning curves.
- Even if all conditions necessary for the population vector method are met, there are other methods that give better estimates.
- It has no way to take neural variability into account.
- It generates a single estimate, with no indication of confidence in the estimate or distribution of alternatives.
- In some cases the stimulus parameters do not naturally fall in a space where vectors make sense (e.g. discrete or nominal variables), and the population vectors cannot be calculated.
- Implementing the population vector method in neural machinery is not particularly easy or plausible. Alternative methods that take weighted sums of the neural responses are not equivalent to it, and are biologically much simpler and fit better with what is known about neural response properties.

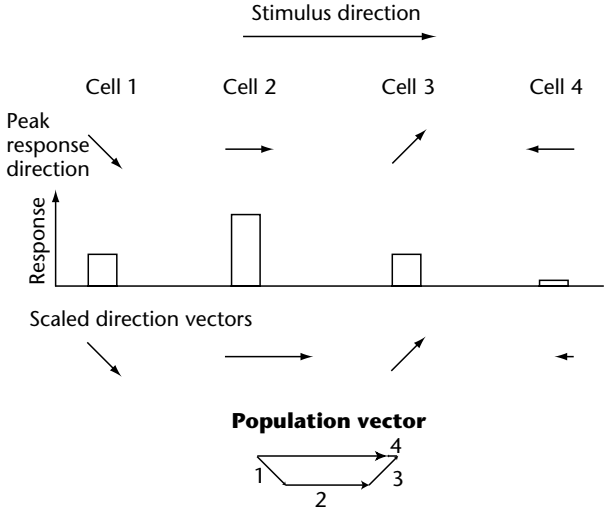


Figure 1. Calculation of the population vector for a horizontal stimulus (or movement) direction from the responses of four neurons.

BAYESIAN METHODS

Most of the above problems (except for the last) can be solved by thinking of both tuning curves and estimates as probability distributions. A description of a neuron's response properties is then given by a conditional probability distribution: $P(r|s)$. This distribution describes the probabilities of observing any response value (r) given that a particular stimulus (s) is presented. The conditional probability is also related to the joint probabilities $P(r,s)$ of the response and the stimulus:

$$P(r|s) = P(r,s)/P(s) \quad (1)$$

where $P(s)$ is the prior probability of seeing stimulus s, without any knowledge of the response. If these conditional distributions can be estimated for all stimuli, we have a full description of the neuron's tuning and variability. Decoding then

becomes easy by applying Bayes' theorem, which allows us to calculate the probabilities of each of the possible stimuli given the observed responses: $P(s|r)$. Bayes' theorem follows from the definition of conditional probabilities

$$P(s,r) = P(r|s)\ P(s) = P(s|r)\ P(r) \tag{2}$$

and it states that

$$P(s|r) = P(r|s)\ P(s)/P(r) \tag{3}$$

where $P(r)$ is the probability of a given response, which is summed across all stimuli as

$$P(r) = \Sigma P(r|s)\ P(s) \tag{4}$$

and it is used to normalize the $P(s|r)$ distribution to sum to 1. An example of two discrete stimuli and two discrete responses from a single neuron is given in Figure 2. The formulas given above are applicable to discrete sets of stimuli and responses, but can also be expressed for continuous variables in terms of continuous probability distributions. When the responses of several neurons are available, the response can be considered to be a vector variable, and the distributions need to be estimated over such a vector space. The problem of estimating the density of a multidimensional probability distribution is usually a difficult one for several reasons. One is that the volume of response space increases exponentially with the number of neurons, and therefore the number of data points needed for estimation increases quickly. The other problem is experimental and is related to whether the responses can be recorded simultaneously from the neurons (e.g. using multiple electrodes). If simultaneous recording is not available then the trial-by-trial relationship between the responses is not observable. Estimates of this dependence and of the prior distribution can influence the results substantially. These apparent disadvantages of the Bayesian method are not avoided, only ignored by the alternative methods. The Bayesian method has a far greater flexibility and power than the population vector method. Not only is it free of assumptions about the shape and distribution of the tuning curves but it is also applicable in cases where vectors could not be defined at all (e.g. in some unknown shape space).

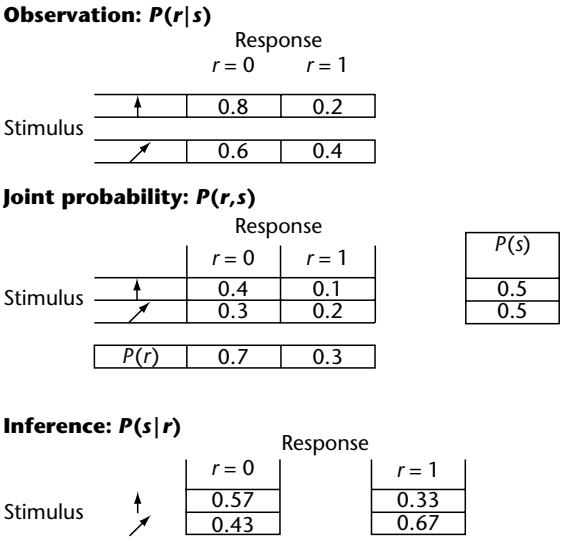


Figure 2. Bayesian decoding of a binary neural response (0 or 1) of a single neuron to a set of two possible stimuli (vertical or diagonal). Such binary responses could be observed if measured in a short time window. The conditional probability $P(r|s)$ is measured separately for the two stimuli, and these are the relative frequencies of the responses. The two stimuli are assumed to occur with equal (0.5) probability. Bayes' theorem (Eqn 3) gives the decoding separately for the two possible responses.

STIMULUS DISCRIMINATION USING POPULATION RESPONSES

The result of applying Bayes' theorem is $P(s|r)$. You can think of this distribution as the answer to the question of what the population response means. The Bayesian method is optimal in the sense that if you have the correct $P(r|s)$ values it gives you the true $P(s|r)$ values, and optimal decisions or discriminations can be made based on this. If the goal, for instance, is to make the smallest number of errors in guessing the stimulus, the most probable stimulus – the one corresponding to the peak of the $P(s|r)$ – should be chosen. Other values of this distribution provide possible alternatives with lower probabilities. A highly peaked probability distribution across the stimuli indicates a high degree of certainty, while a broad, flat distribution is a sign of uncertainty in the decoding due to the variability in the response and the overlap between the conditional distributions. Other optimality criteria can be expressed as a loss function $L(s,s')$, where s is the actual and s' is the chosen stimulus. It can be expected that confusing some pairs of stimuli (e.g. similar stimuli, or where both are members of the same category) will be much less costly than confusing other pairs (e.g. members of different categories, or dissimilar stimuli). The loss function can capture the relevant structure, and Bayesian inference gives the decision that minimizes expected cost.

APPLICATIONS

Decoding methods are not usually considered to be models of physiological processes, as it is not likely that explicit stimulus (or movement) decoding takes place in the brain. Information in inputs, outputs, as well as internal representations are distributed to some extent. There is no clear evidence that any component of such a representation could be considered a decoding of a stimulus. Decoding is still interesting from a theoretical perspective, as it tells us the amount and nature of information available for processing in a certain population, and sets upper limits on performance for a mechanism operating on inputs from this population, just as an 'ideal observer' is a useful theoretical concept in studying human psychophysical performance. It allows the calculation of the efficiency of an actual neural mechanism by providing the best possible level of performance for comparison and the calculation of the amount of information available. Such an ideal observer of the neural activity allows an investigation and interpretation of neural representations in real experimental situations. It does not, however, in itself tell us how the brain actually uses the information in a neural population.

Further Reading

Dayan P and Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, chap. 3, Neural Decoding, pp. 87–122. Cambridge, MA: MIT Press.

Oram MW, Foldiak P, Perrett DI and Sengpiel F (1998) The 'Ideal Homunculus': decoding neural population signals. *Trends in Neurosciences* **21**: 259–265.

Rieke F (1997) *Spikes*. Cambridge, MA: MIT Press.

Decoding Single Neuron Activity

Introductory article

James J Knierim, University of Texas–Houston Medical School, Houston, Texas, USA

CONTENTS

Neurons communicate with each other by firing trains of action potentials (spikes). Understanding what information is represented by this activity, and how to decipher the code in which the neural messages are encrypted, are the central issues of single neuron physiology.

INTRODUCTION

One of the most powerful techniques for deciphering the brain mechanisms that underlie cognition is the recording *in vivo* of single neuron activity with the use of microelectrodes. From low-level brain areas (such as primary sensory and motor cortex) to high-level areas (such as prefrontal cortex and hippocampus), single neuron recordings have provided a wealth of data about how information is processed and stored in the brain. One of the great challenges of neuroscience, however, is how to interpret the firing patterns of single neurons. Is information encoded in the average firing rate of neurons, in the precise temporal dynamics of a train of action potentials ('spikes'), or in some other parameter (or set of parameters)? Does the code differ depending on the brain area under study, the experimental task or stimulus at hand, or the amount of experience that the organism has? These questions are at the forefront of neurophysiological research.

REPRESENTING SPIKE TRAINS

Spike trains are usually represented in a few standard formats. Figure 1a shows an oscilloscope trace from an extracellular neuronal recording, with time on the x axis and spike amplitude (voltage) on the y axis. The line underneath the trace denotes the presentation time of the stimulus used to evoke a response from the cell. There are a number of discrete spikes of different amplitudes that rise above the baseline level of activity; each amplitude level presumably corresponds to the firing of a different neuron. Let us concentrate on the cell that fires the largest spikes in the example (denoted by asterisks). The neuron fired few spikes in the periods before and after the stimulus presentation; the rate of firing in the absence of stimulation is referred to as the background level or spontaneous activity level. When the stimulus was presented, the neuron fired a rapid series of spikes. There was a burst of activity shortly after the onset of the stimulus, and the activity decreased over time (although in some neurons the response might stay strong throughout the duration of the stimulus). These raw recordings are usually converted to a number of formats to represent the spike trains. Figure 1b is a spike raster plot from five repetitions of the same stimulus. The first line of the raster is a representation of the oscilloscope trace in Figure 1a. Each mark corresponds to the firing of a spike at that particular time. The neural responses to further repetitions of the stimulus are shown in the subsequent lines. As all responses are aligned to the onset of the stimulus, notice that the first spike of each response can occur after a variable delay period (see below). (*See* **Single Neuron Recording**)

Such spike rasters can be converted into a peristimulus time histogram (PSTH), representing the cumulative firing of the neuron over many trials (Figure 1c). The trial is divided into a series of time bins aligned to the onset of the stimulus, and all of the spikes that occurred within a time bin are summed to create the PSTH. Figure 1c illustrates two superimposed histograms (one black and one shaded), each corresponding to a different stimulus, making it clear that the neural response to one stimulus was greater than the response to the other. Finally, the PSTHs of many neurons can be averaged together, as in Figure 1d. Although the single

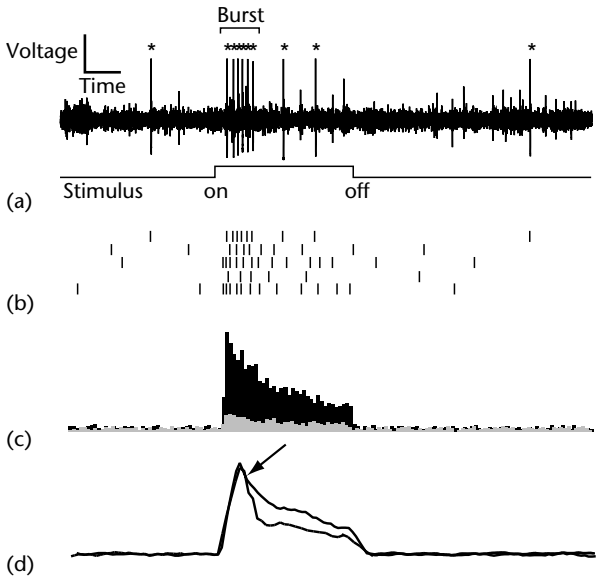


Figure 1. Representations of spike trains. (a) Oscilloscope trace of spikes. (b) Spike raster diagram. (c) Peristimulus time histogram. (d) Peristimulus time histogram averaged over a population of neurons.

neuron in Figure 1c had very different responses to the two stimuli, Figure 1d demonstrates that the average response of a population of many neurons did not distinguish the two stimuli as strongly; moreover, the average response was equal during the initial part of the response, and the difference in response became apparent only after a delay period (arrow). Such population averages provide a 'snapshot' of the overall neural population response to a stimulus or to a behavioral event, which sometimes may prove more informative than the response of any single neuron. (*See* **Decoding Neural Population Activity**)

As understanding of neural firing properties increases and theoretical advances occur, more sophisticated mathematical techniques are being employed to represent spike trains and to decipher the information they encode. Such approaches include the use of information theory to quantify rigorously the amount of information encoded in a spike train, and computational modeling to simulate how neurons or networks of neurons represent and store information. A growing legion of physicists, mathematicians, computer scientists and engineers are joining traditional neuroscientists to develop the analytical tools that will be necessary to understand fully the firing patterns of neurons. (*See* **Information Theory**)

SOURCES OF NEURONAL VARIABILITY

It is obvious from Figure 1b that there is a large degree of variability in the spike trains elicited on each of the five trials, both in the number of spikes fired on each trial and in the temporal pattern of spikes. Such variability is ubiquitous in the brain and derives from two broad sources. The first source is the set of uncontrolled variables that are present in a particular experiment. These uncontrolled variables may be behavioral variables in an alert animal preparation (e.g. changes in the animal's attentional state, changes in gaze direction, and changes in arousal or motivation); stimulus parameters that may change slightly from trial to trial; or other factors, such as changes in anesthesia level in an anesthetized preparation. The second source derives from properties that are intrinsic to the neuron or neural network. For example, the stochastic nature of the opening and closing of individual ion channels may contribute to whether a neuron that is near firing threshold actually generates a spike at a particular time. As a second example, repeated presentations of a stimulus can cause an adaptation of the response, in which the response strength decreases over time as the result of changes in biochemical cascades within a neuron or in the strengths of synaptic connections in the network. Of these intrinsic sources, it is useful to distinguish random sources (e.g. the stochastic properties of single ion channels) from deterministic sources (e.g. response habituation).

MOST EFFECTIVE STIMULI

The spike trains illustrated in Figure 1 are often reduced to a single number, the firing rate, by dividing the number of spikes in a certain period by the length of that period (e.g. the amount of time that the stimulus was presented). Neurons are often characterized by the type of stimulation that produces their greatest rate of firing. For example, the responses of a neuron in the visual cortex to different colors of light can be plotted as a tuning curve that illustrates the overall response selectivity of the cell. Figure 2 illustrates a hypothetical cell that fired strongly to a green light, with little or no response to light of other wavelengths. Such a cell is said to be tuned to, or selective for, a particular stimulus. In contrast, a second cell responded very similarly to wavelengths of all colors. This cell is said to be nonselective for color, although it might

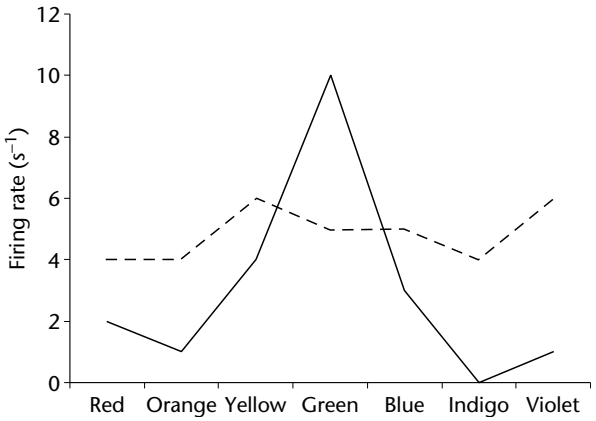


Figure 2. Hypothetical tuning curves for visual neurons. The solid line represents a neuron that is selective for the color green; the dashed line represents a neuron that is not selective for color.

be selective for some other parameter (for example, the size of a stimulus). The characterization of the most effective stimulus for a particular neuron is one of the most prevalent and important analyses of single-unit data. Neurons have been identified in the visual cortex that are selective for color, size, depth, speed of motion, direction of motion, spatial frequency, and other, more complex, properties of visual stimuli. Neurons in the auditory system can be selective for sound frequency, loudness, or the spatial location of a sound source. Neurons in the thalamus and limbic cortex can be selective for the direction in which the animal is facing. These are a few examples of the many different types of selective responses displayed by neurons in different parts of the brain.

Although this type of research is important, a few caveats are necessary. First, finding a cell that is selective for color, for example, does not necessarily mean that the function of the cell is color perception *per se*. The neuron may be part of a network that is actually involved in extracting information about boundaries between objects based on differences in color between the two objects. Second, finding the most effective stimulus that drives a neuron is limited to the set of stimuli that are tested in a given experiment. Thus, a cell may respond best to a red spot of light when tested against other spots of light, but its true function may be related more closely to the perception of complex three-dimensional shapes. If the neuron's response to such shapes is never tested, however, the investigator may incorrectly attribute color selectivity as the primary correlate of that neuron. This point leads to the third caveat, which is that neurons often are tuned along multiple stimulus dimensions. Thus, a cell may be sharply tuned to red on the color dimension, horizontal on the orientation dimension, and far away on the depth dimension. Is such a cell specifically tuned for a horizontally oriented, red bar that is located far away from the observer? Or does this cell act as a filter for many different stimulus parameters in a multiplexed, population code? Most current thought favors the latter interpretation, but how the cell performs this task and how the brain interprets the code is not well understood.

REVERSE CORRELATION METHODS

The analyses described above are based on the methods of presenting a set of discrete stimuli to the animal and measuring the response to each member of the set. Another powerful method of finding the most effective stimulus for a cell is to record the activity of a neuron continuously, during behavior or during the presentation of a continuous stream of stimuli, and to use the method of reverse correlation to determine what aspect of the behavior or stimulus drove the cell to fire. For example, a neuron in visual cortex might be recorded while an animal views a succession of scenes on film. Whenever a spike occurs, the frame displayed at that moment is tagged. After the presentation of the whole film sequence, each tagged frame is analyzed to find the common element (or elements). One way to do this would be to create a prototypical tagged frame by averaging each pixel from all of the tagged frames. If the averaged frame showed a vertically oriented edge, one might infer that the cell was acting as a filter for vertical orientation. As another example, cells in the hippocampus of rats can be recorded as the animal freely explores an environment. For each spike, the spatial location of the animal is recorded, and by reverse correlation of the spikes of the cell with the position of the rat, it can be shown that the cell fires whenever the rat enters a particular part of that environment. Each pyramidal cell in the rat hippocampus fires in different selected regions of different environments, and this finding was one of the principal lines of evidence for the cognitive map hypothesis of hippocampal function. (*See* **Place Cells**)

ENCODING VERSUS DECODING

It is important to distinguish the encoding of neural activity from its decoding. This discussion has so far been concerned mainly with the former operation; that is, how does the brain take information

(either from the external world via sensory receptors or from other brain areas) and transform it into a representation made of spiking neurons. The other side of the problem is understanding how the brain later on decodes this representation. To understand encoding, we want to know how a neuron will respond to a particular stimulus:

$$\text{stimulus} \rightarrow \text{neural response}$$

To understand decoding, we want to know how a downstream neuron interprets this neural response in order to determine what stimulus produced the response:

$$\text{neural response} \rightarrow \text{stimulus}$$

These two processes can be thought of in terms of probabilities, and they can be related by an equation known as Bayes' rule:

$$P(\text{stimulus}|\text{response}) = P(\text{response}|\text{stimulus}) \times P(\text{stimulus})/P(\text{response}) \quad (1)$$

where $P(\text{stimulus}|\text{response})$ is the probability that a particular stimulus was presented given that the neuron produced a particular response (decoding); $P(\text{response}|\text{stimulus})$ is the probability that the neuron produced that response given that the particular stimulus was presented (encoding); $P(\text{stimulus})$ is the general probability of the animal being presented with that particular stimulus out of all possible stimuli; and $P(\text{response})$ is the general probability of the neuron producing that response regardless of what stimulus is presented. (*See* **Neurons, Representation in**)

Understanding encoding is a prerequisite for understanding decoding, and much research focuses on exactly what coding schemes are used by neurons. For example, one of the great debates of neurophysiology is the degree to which the firing of neurons constitutes a rate code or a temporal code: is the information encoded by neurons contained in the firing rate of the neuron, averaged over a certain time interval, or is the information contained in the precise temporal pattern of spikes emitted by the neuron? How neurons perform the decoding operation is also a subject of much research. Although the complexity of the brain and the complexity of cognitive behavior combine to make these tasks daunting, they are necessary for a complete understanding of the brain mechanisms that underlie cognition. (*See* **Rate versus Temporal Coding Models**)

Further Reading

Abbott LF (1994) Decoding neuronal firing and modelling neural networks. *Quarterly Reviews of Biophysics* **27**: 291–331.

DeCharms RC and Zador A (2000) Neural representation and the cortical code. *Annual Review of Neuroscience* **23**: 613–647.

DeYoe EA and Van Essen DC (1988) Concurrent processing streams in monkey visual cortex. *Trends in Neuroscience* **11**: 219–226.

Laurent G (1999) A systems perspective on early olfactory coding. *Science* **286**: 723–728.

Optican LM and Richmond BJ (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology* **57**: 162–178.

Rieke F, Warland D, de Ruyter van Steveninck R and Bialek W (1997) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Shadlen MN and Newsome WT (1994) Noise, neural codes and cortical organization. *Current Opinion in Neurobiology* **4**: 569–579.

Descending Motor Tracts

Introductory article

Michael Peters, University of Guelph, Guelph, Ontario, Canada

CONTENTS

The human ability to 'compose' new kinds of movement allows us to speak and to generate technology. Evolutionary changes in the descending motor tracts and especially the corticospinal tracts lie at the core of this ability.

INTRODUCTION

Action implies movement, and movement is implemented by nerve tracts that descend from the brain to the spinal cord, where they activate motor neurons. The activation can take place by direct contact of the terminals of the tracts on motor neurons, or indirectly, by termination on 'interneurons' in the spinal cord which then activate the motor neurons. Motor neurons are cells that directly activate muscle contraction. Movement ranges from simple reflexive responses to highly complex voluntary movement, and to serve this range of movement, a number of descending tracts are needed.

THE CORTICOSPINAL TRACTS

Of all descending motor pathways, the corticospinal pathways are by far the most massive. In humans there are about a million corticospinal fibers on each side, although there is considerable variability, with values as low as 750 000 and as high as 1 400 000. The vast majority of these fibers cross over from one side to the other, in the pyramidal decussation. The fibers in the corticospinal tract that cross over to the contralateral side form the dorsolateral corticospinal tracts (or funiculi) in the spinal cord (Figure 1). These tracts are thought to be primarily involved in the control of muscles in the distal parts of the limbs (especially the muscles that operate the hand). A minority of fibers originating from the left hemispheres of the brain descend in the left half of the spinal cord. These fibers are largely found in the ventral and medial portion of the spinal cord (Figure 1) and form the ventromedial or anterior corticospinal tracts. The proportion of fibers that remain uncrossed is variable from individual to individual, ranging from individuals in whom the normally crossed portion descends ipsilaterally, to individuals who lack an ipsilateral portion of that tract. The ventromedial corticospinal tracts are largely concerned with the central parts of the body, such as the torso.

Source of Corticospinal Tract Fibers

Where do the corticospinal tract fibers come from and where do they go to? The majority of the fibers come from the cortex that lies anterior to the central fissure. Immediately anterior to the central fissure is the primary motor cortex (usually referred to as M1). Primary motor cortex means the area of the anterior cortex that was first recognized as being specialized for motor function because stimulation in this area would produce movement in the limbs. This is the area from which the densest fiber projections are sent towards the spinal cord. Anterior to the M1 lie several other motor areas. Current work with primates suggests that there are at least seven of these areas. One of these, the supplementary motor area (SMA) also contains a rough topographical map of the body musculature, like the M1 area, and the SMA also sends fibers directly to the spinal cord. Other motor areas are collectively called premotor cortex. Some of these send fibers directly to the spinal cord and others connect to each other and the M1, without any direct output to the spinal cord. In humans, the corticospinal tract fibers that come from anterior of the central fissure amount to some 80% of the tract, with some 20% originating posterior to the central fissure. In monkeys the ratio is closer to 60:40, indicating a relatively greater weight on motor cortex origin in humans.

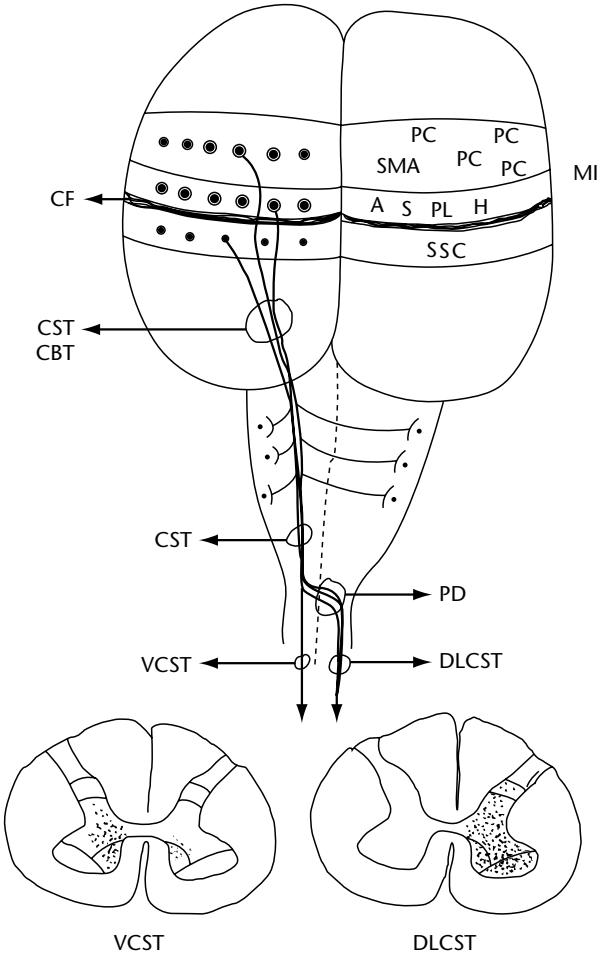


Figure 1. Origin and path of corticospinal and corticobulbar fibers. The bulk of fibers (80%) run from the cortex anterior to the central fissure and 20% from the somatosensory cortex. The bulk of corticobulbar fibers contact the motor nuclei of the nerves that control movements of the head and neck, eyes, facial musculature, jaws and tongue. Some corticospinal fibers do not cross the midline and descend on the ipsilateral (same) side, cross over to the contralateral side to descend the spinal cord as the ventral (anterior) corticospinal tract (VCST). The bulk of corticospinal fibers cross to the opposite side in the pyramidal decussation and descend as the dorsolateral corticospinal tract (DLCST). In the primary motor cortex (M1) there is a topographical relation to body parts, where 'A' denotes axial – the midline parts of the body, 'S' denotes the shoulders, 'PL' denotes the proximal limbs and 'H' denotes the hand. The two spinal cord sections (see Figure 2) show the terminations of the two corticospinal tracts. VCST ends mostly in the ipsilateral intermediate region, on premotor neurons concerned with the axial and proximal parts of the body. DLCST terminates throughout the grey matter, on sensory interneurons, premotor neurons in the intermediate region and on the motor neurons for proximal and distal body parts. Dots in the spinal cord section show predominant terminations of corticospinal fibers. The size of dots in the originating cortical areas indicates the relative contribution of those areas to the descending tracts. Abbreviations: CBT, corticobulbar tract: CF, central fissure; CST, corticospinal tract; MI, primary motor cortex; PC, premotor cortex; PD, pyramidal decussation; SMA, supplementary motor area; SSC, somatosensory cortex.

Destination of Corticospinal Tract Fibers

The termination of corticospinal fibers in the spinal cord reflects their region of origin (Figure 2). Fibers that terminate in the dorsal horn, where the sensory relay cells are located, come mostly from the sensory cortex, posterior to the central fissure. Fibers that terminate on the premotor neurons in the intermediate zone come mostly from the premotor cortex anterior to the primary motor cortex. Finally, fibers that terminate on the motor neurons in the spinal cord come mostly from the primary motor cortex. That fewer components of the corticospinal tract in humans originate from the sensory cortex is supported by the observation that fewer corticospinal tract fibers terminate in the dorsal horn of the spinal cord than is the case for monkeys. While humans share with other primates direct terminations of the corticospinal tract fibers on motor neurons that innervate finger muscles, humans are unique in also having direct terminations on the motor neurons that operate the thoracic musculature. This is needed because of the precise use of the thorax as a 'bellows' in speech.

Finally, descending corticospinal tract fibers end in the gray matter of the spinal cord in an orderly fashion, such that the termination on motor neurons that activate distal muscles, such as hand muscles, is located laterally (that is, near the left or right outer edges of the ventral part) in the gray matter. In contrast, fibers that activate muscles that operate the center of the body, such as the torso, terminate medially (that is, in the gray matter close to the center of the spinal cord).

Loss of the Corticospinal Tracts

Thorough observations about what happens after the corticospinal tracts are severed are only available for monkeys, because in humans accidental damage to the corticospinal tracts is rarely restricted to these tracts. In monkeys with destruction of the pyramids – and therefore of the cortical spinal tracts – there are extremely severe deficits immediately after the damage. However, recovery does take place. Researchers observed a striking

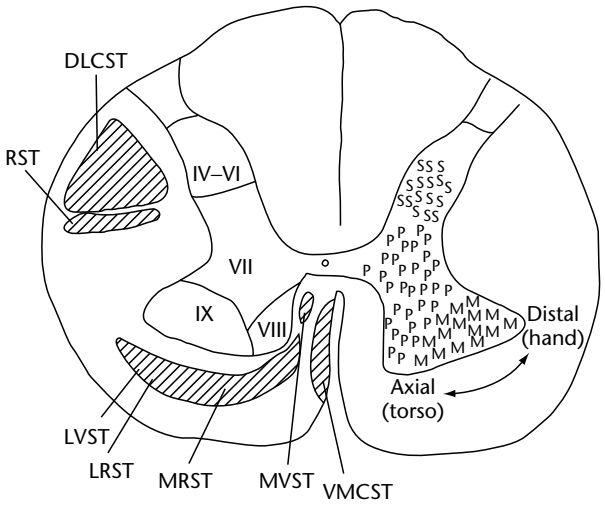


Figure 2. Cross-section of the spinal cord. The central butterfly-shaped region contains mostly cell bodies, and is know as the 'gray matter'. Around it lie the regions through which fiber tracts descend and ascend in the spinal cord, known as 'white matter'. The left half of the section shows the regions or Lamina of Rexed, and the right half shows the major contents of these regions. Cross-sections vary in appearance depending on the level at which the cord is cut. This level is roughly in the region of the shoulders. Motor neurons are arranged in a rough topographical order. The terms 'distal' and 'axial' denote the location of motor neurons that innervate distal portions of the body, such as the hand, and axial portions, such as the torso. Abbreviations: S, sensory neurons in the dorsal horn; P, intermediate region of the gray matter that contains mostly premotor neurons; M, motor neurons that connect directly to muscles. DLCST, dorsolateral corticospinal tract; LRST, lateral reticulospinal tract; LVST, lateral vestibular tract; MRST, medial reticulospinal tract; MVST, ventral vestibulospinal tract; RST, rubrospinal tract (very small in humans); VMCST, ventral or anterior corticospinal tract.

difference in recovery between movements that involved walking and climbing, and movements of the individual fingers. Ultimately the animals were able to walk and climb relatively well, even though the movements remained slow and there were signs of fatigue. In addition, independent movements of the arms recovered so that the animals could reach for food quite quickly and accurately. What did not recover, however, was the ability to use manipulative individual finger movements in, for instance, winkling a morsel of food out of a hole with a single finger. Interestingly, there were also difficulties with releasing an item of food once it was grasped with the whole hand even though the animals had no trouble grasping and releasing a bar of the cage during climbing. All of this suggests a prominent role of the corticospinal tracts in the voluntary execution of precise and independent finger movements.

A final observation concerns the capacity for recovery in animals with sparing of some corticospinal tract fibers. Individual finger movement seems to recover regardless of the precise location of the spared fibers in the tract. This suggests a remarkable plasticity in the role of descending corticospinal fibers, allowing considerable room for reorganization in the relation between corticospinal tract fibers and their target neurons.

THE CORTICORUBROSPINAL TRACT

While it is true that the corticorubrospinal connections show much less prominence in nonhuman primates and humans than in subprimate species, it is also true that the red nucleus itself remains undiminished in relative size. The answer to this puzzle lies in the nature of the red nucleus. It has a magnocellular (large-celled) part that is associated with the contralaterally descending corticorubrospinal tract, and appears to be involved in the direct control of limbs. It also has a parvocellular (small-celled) part that does not seem to be involved in projections to the spinal cord and also seems to play no part in the direct control of the muscles. Instead, this part is richly interconnected with the cerebellum. It is this part that has grown at the expense of the magnocellular part in primates, especially in humans (Figure 3).

Loss of the Corticorubrospinal Tract

In monkeys, selective lesions of the rubrospinal tract lead to slowing of movement as well as limpness and weakness in the hands. However, these problems show recovery to the point where no deficits are noted. If the rubrospinal tract is damaged on one side after loss of the corticospinal tracts, there are permanent problems with the ability to perform those hand and arm movements that showed recovery after corticospinal tract loss. Thus, the animals were no longer able to close the affected hand around a morsel of food, and reaching movements were poor. For instance, if animals were held close to the bars of the cage, they would try to move the affected arm into position by moving the shoulder rather than making use of the arm itself. In contrast, the animals could use the hand for climbing and clinging onto a bar even though there was weakness in the affected hand. In addition, animals showed the ability to right themselves and body posture was relatively normal.

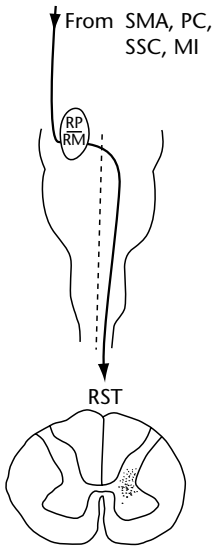


Figure 3. The cortical input from the supplementary motor area. The fibers from the RM cross the midline and descend as the rubrospinal tract (RST) in the spinal cord. They are shown to terminate mostly on premotor neurons in the intermediate region of the gray matter. PC, premotor cortex; SSC, somato-sensory cortex; MI, primary motor cortex to the large-celled portion of the red nucleus; RM, magnocellular portion of red nucleus; RP, parvocellular portion of red nucleus.

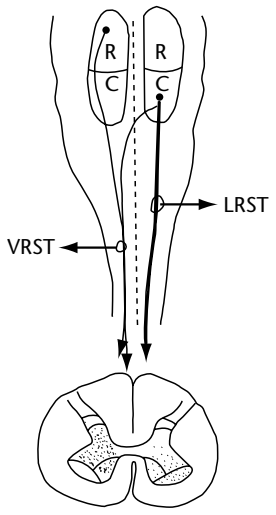


Figure 4. Two of the reticulospinal tracts are shown, in a much simplified form. VRST, ventral reticulospinal tract; LRST, lateral reticulospinal tract; R, rostral (toward the top of the brain) portion of the reticular formation; C, caudal (toward the spinal cord portion of the reticular formation).

However, the arm on the side affected by the lesion was not held in a normal position but hung loosely from the shoulder.

It is clear that postural control of the body is managed by a system other than the corticorubral or the corticospinal system, because such control survives loss of both of these systems.

THE VENTROMEDIAL BRAINSTEM–SPINAL SYSTEM

Several tracts originate in the brainstem, notably from the reticular formation in the region of the medulla oblongata and the region of the pons (Figure 4). Another important component of this system comes from the vestibular nuclei. There are two separate paths within the reticulospinal system; one originates more dorsally and rostrally in the reticular formation and descends ipsilaterally in the medial reticulospinal tract, while a more caudal and ventral portion descends bilaterally in the lateral reticulospinal paths. One hint as to the possible roles of these tracts comes from the observation that axons from neurons in the reticulospinal tract may make contact with target neurons in the region of the spinal cord where information flows out to the forelimbs (cervical level) and then continue to regions concerned with the hind limbs (lumbar levels). In addition, axons may also send collateral branches across the midline. It is likely that these pathways not only manage basic locomotion and postural adjustments but are also involved in the integration of these more basic motor behaviors with voluntary movement.

Loss of Function in the Reticulospinal System

Damage to the descending paths from the reticulospinal tracts and parts of the vestibulospinal tracts in monkeys leads to lasting impairment in body posture and movements. The animals walk unsteadily, and limbs and the trunk are flexed. The shoulders are raised; the head is not held upright in the normal position and falls forward. When the animals jump toward the bars of their cage, they often miss and when they reach for a morsel of food, they cannot move the hand smoothly toward the target. This is due to an ataxia (problem with muscular coordination) in the portions of the arm close to the body. In contrast, once the hands are near a target, the animals can perform delicate individual movements of the fingers; this would be expected because the corticospinal tracts are intact.

The description of the behavior of these animals provides the clearest illustration of the principle that in order to function properly, the corticospinal tracts need the support of the reticulospinal system.

Thus, the increased specialization of the corticospinal system in the guidance of voluntary and skilled movements in human as opposed to nonhuman primates does not render this 'older' system obsolete in humans. If anything, because the requirements of 'composing' and learning new movements demand strong support by the postural and integrative contributions of the brainstem–spinal systems, these systems are as important as ever.

Further Reading

Armand J (1982) The origin, course and terminations of corticospinal fibers in various mammals. In: Kuypers HGJM and Martin GF (eds) *Anatomy of Descending Pathways to the Spinal Cord*, vol. 57, pp. 229–360. Amsterdam: Elsevier Biomedical Press.

Canedo A (1997) Primary motor cortex influences on the descending and ascending systems. *Progress in Neurobiology* **51**: 287–335.

Donoghue JP and Sanes JN (1994) Motor areas of the cerebral cortex. *Journal of Clinical Neurophysiology* **11**: 382–396.

Galea MP and Darian-Smith I (1995) Postnatal maturation of the direct corticospinal projections in the macaque monkey. *Cerebral Cortex* **5**: 518–540.

Ghez C (1991) The control of movement. In: Kandel ER, Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, 3rd edn, pp. 533–547. New York: Elsevier.

Kennedy PR (1990) Corticospinal, rubrospinal and rubro-olivary projections: a unifying hypothesis. *Trends in Neurosciences* **13**: 474–479.

Kuypers HGJM (1981) The descending pathways to the spinal cord, their anatomy and function. In: Brooks VB (ed.) *Handbook of Physiology*, sect. 1: The nervous system, vol. II, Motor control, part 1, pp. 597–666. Bethesda, MD: American Physiological Society.

Kuypers HGJM (1982) A new look at the organization of the motor system. In: Kuypers HGJM and Martin GF (eds) *Anatomy of Descending Pathways to the Spinal Cord*, vol. 57, pp. 381–403. Amsterdam: Elsevier Biomedical Press.

Lawrence DG and Kuypers HGJM (1968) The functional organization of the motor system in the monkey. I. The effects of bilateral pyramidal lesions. *Brain* **91**: 1–14.

Lawrence DG and Kuypers HGJM (1968) The functional organization of the motor system in the monkey. II. The effects of lesions of the descending brain-stem pathways. *Brain* **91**: 15–36.

Rizzolatti G, Luppino G and Matelli M (1998) The organization of the cortical motor system: new concepts. *Electroencephalography and Clinical Neurophysiology* **106**: 283–296.

Schwartzman RJ (1978) A behavioral analysis of complete unilateral section of the pyramidal tract at the medullary level in Macaca mulatta. *Annals of Neurology* **4**: 234–244.

Diffusion Models and Neural Activity

Intermediate article

Luigi M Ricciardi, Federico II University, Naples, Italy
Petr Lánský, Academy of Sciences of the Czech Republic, Prague, Czech Republic

CONTENTS

Neuronal interspike intervals can be characterized in terms of the first-passage time probability density of stochastic diffusion processes under steady state and periodic stimulation. The Wiener and Ornstein–Uhlenbeck models, and models with multiplicative noise, can be used to elucidate neuronal activity.

INTRODUCTION

One of the basic modes of signaling in the nervous system is by the frequency of action potentials. This is true even if the input signal is time-varying, in which case the firing rate is expected to be modulated in time to reflect the time course of the input. The rate coding that is introduced via the stochastic description of single neurons is the focus of this article. A common way to introduce stochasticity is by the assumption that the incoming signal includes a random component, generally denoted as 'noise'. Another source of stochasticity can originate in the neuron itself, where a random component is added to the signal. Taking these circumstances into account, one is led to the conclusion that the deterministic differential equations usually describing the membrane response should be completed by insertion of a noise term, thus becoming stochastic differential equations. The solutions of such equations, under certain regularity conditions on their coefficients, can be viewed as 'sample paths' or 'realizations' of stochastic processes belonging to the class of 'diffusion' processes. This is the rationale for the rise of diffusion models for the description of neuronal activity. Here, the center of interest is single-point models of interspike intervals generation. Such a one-point representation implies severe restrictions, discussed below. This type of simplification neglects the spike's duration and its detailed shape, so that the entire neuronal activity is schematized in the form of identical point-size signals occurring in time according to a stochastic point process. On the other hand, it permits us to quantify not only the mean of the spiking activity, but also its variability – even its probability distribution.

DETERMINISTIC LEAKY INTEGRATE-AND-FIRE NEURONAL MODEL

The electrical circuit model of a neuronal membrane is composed of a resistor and a capacitor in parallel charged by a battery (RC circuit). It can be described by the first-order differential equation

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\frac{v(t)}{\tau} + i(t), \quad v(t_0) = v_0 \qquad (1)$$

where $v(t)$ denotes the difference between the membrane potential at time t and the membrane potential in resting conditions (i.e. the membrane depolarization), τ is the membrane time constant, $i(t)$ is the input to the neuron and v_0 is the initial voltage after spike generation that for simplicity – and without loss of generality – is sometimes set to zero ($v_0 = 0$). This is also known as the Lapicque model, or the RC circuit model. In this model, under a constant input, $i(t) = i$, as time grows to infinity the depolarization tends to the asymptotic level $i\tau$, and if the input is removed at some instant, say $i(t) = 0$, the depolarization tends exponentially to zero (i.e. to the resting level), the speed of approach to zero depending on time constant τ. The extreme simplicity of eqn (1) witnesses that the action potential generation mechanism is not an inherent part of the model, so that a firing threshold S, with $S > v_0$, has to be postulated. Within this model the neuron is assumed to fire whenever the threshold voltage is reached, and the membrane depolarization $v(t)$ is assumed to be instantly reset to v_0 after each firing. The interspike intervals (ISIs)

are identified with the first-passage time of $v(t)$ through S, namely with the variable

$$T = \inf\{t \geq t_0 : v(t) > S\}, \quad v(t_0) = v_0 < S \qquad (2)$$

In the case of a constant input, the condition for evoking a spike is $i\tau > S$, which defines suprathreshold stimulation. Otherwise the input is unable to produce a spike and the neuron remains silent. For a standard treatment of model described by eqn (1) see Keener *et al*. (1981) and Knight (1972).

STOCHASTIC LEAKY INTEGRATE-AND-FIRE NEURONAL MODEL

One of the stochastic versions of eqn (1), of interest in the present context, is formally obtained by adding to the right-hand side a term to account for the random components that are present as a part of the global signal acting on the neuron. This random component is usually identified, as a useful approximation, with the 'white noise' $\xi(t)$:

$$\frac{dv(t)}{dt} = -\frac{v(t)}{\tau} + i(t) + \sigma\xi(t), \quad v(t_0) = v_0 \qquad (3)$$

By definition, the white noise is a stationary Gaussian process characterized by zero mean and delta-type correlation function. In eqn (3) σ is a positive parameter representing the amplitude of the random fluctuations of the noise, a sort of measure of the degree of unpredictability of the deviations of the noise sample paths from its mean trajectory. Although the mathematical definition of the process $\xi(t)$ is a rather complicated and subtle issue, by analogy with white light as the superposition of all colors, the white noise may be intuitively envisaged as a random process in which spectral components of all frequencies are present and equally represented. For constant $i(t)$ the process defined in eqn (3) is called the Ornstein–Uhlenbeck model of the membrane potential. Its properties in the absence of the firing threshold are as follows: (1) at each time the probability density function of the membrane depolarization is Gaussian; (2) an equilibrium regime exists since in the limit, as $t \to +\infty$, the probability density function that describes the membrane depolarization attains mean $i\tau$, and variance $\tau\sigma^2/2$. Note that the asymptotic variance is independent of the input signal.

Generation of an action potential in the model defined in eqn (3) is again described by the first-passage time of the trajectories of the process across the firing threshold S, i.e. by the random variable (eqn (2)). However, since now the trajectories are different realizations of the stochastic diffusion process modeled by eqn (3), the ISIs are of different length even for constant input. Thus T is a nondegenerate random variable. For this model under constant input $i(t) = i$, as well as for others, the neuronal output is a renewal process, namely intervals between successive firings are independent and identically distributed random variables. Note that now, in contrast with model (1), spikes can be generated due to the presence of noise even for subthreshold stimulations.

Figure 1 depicts the spike generation process. It indicates the effects of a constant stimulation of magnitude i. In the absence of noise (Lapicque model), the depolarization grows exponentially towards the asymptotic value $i\tau$. However, as soon as the neuron's threshold S ($S < i\tau$) is reached, a spike is generated and the depolarization is instantly reset at its initial value. The process then starts afresh. Within such a model, a rigorously periodic sequence of spikes (arrows, Figure 1) at time T, $2T$,… is generated. In the case of the leaky integrate-and-fire model (cf. eqn (3)) the presence of noise alters dramatically the picture, in that the trajectories of the neuron's depolarization now exhibit an erratic time course. In Figure 1 three such trajectories are plotted and the instants t_1, t_2, t_3 of attainment of threshold S are indicated together with the corresponding generated spikes. The first-passage time of the process modeling the neuron's depolarization is now a random variable whose probability density function mimics the probability density of the ISIs. Figure 2 shows an ISI histogram for a finite number of trajectories, and the ISI probability density obtained when all possible trajectories are taken into account.

The properties of this model, including the moments of the random variable T, are complicated (Ricciardi, 1977; Ricciardi and Sacerdote, 1979). Inoue *et al*. (1995) have attempted to compare systematically this model with experimental data, and Lánský (1983) describes a method to estimate parameters based on intracellular recording of the trajectories of $v(t)$.

Until recently the models encountered in applications of stochastic diffusion processes to neuroscience have been predominantly time-homogeneous, reflected by the circumstance that i and σ in eqn (3) do not depend on time. An interest in stochastic resonance (a cooperative effect that arises out of the coupling between deterministic and random dynamics in nonlinear systems) evoked studies on diffusion neuronal models with time-dependent parameters. Analogous to the model in eqn (1), the Ornstein–Uhlenbeck model operates in two distinct regimes – the deterministic

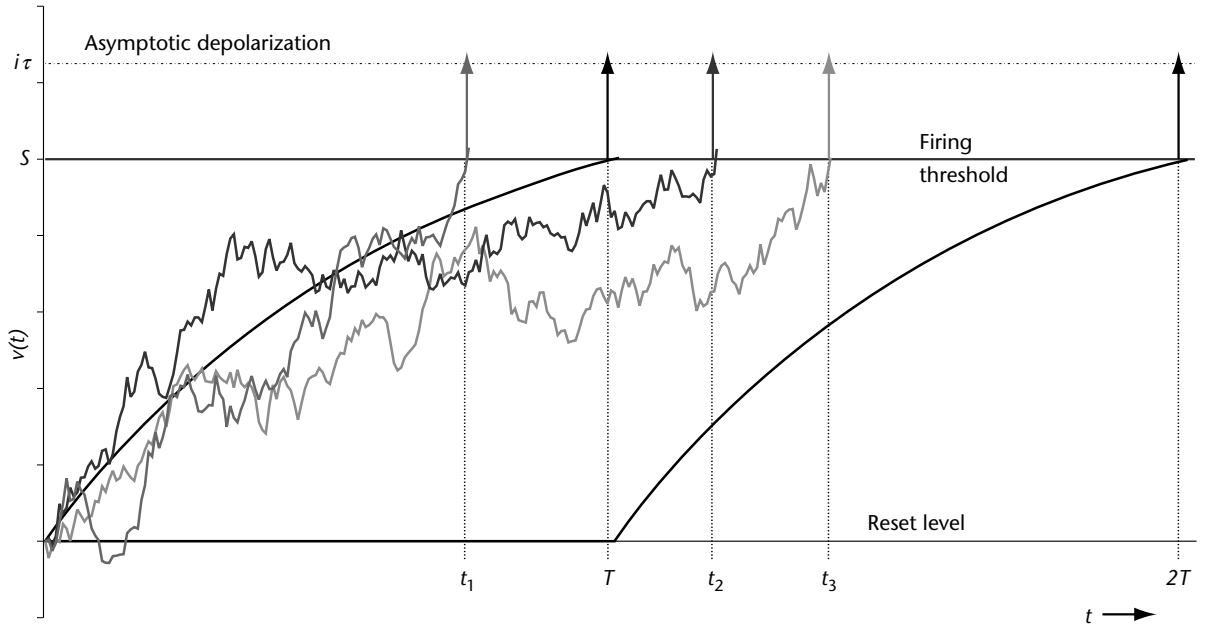


Figure 1. Spike generation mechanism for Lapicque and stochastic leaky integrate-and-fire models.

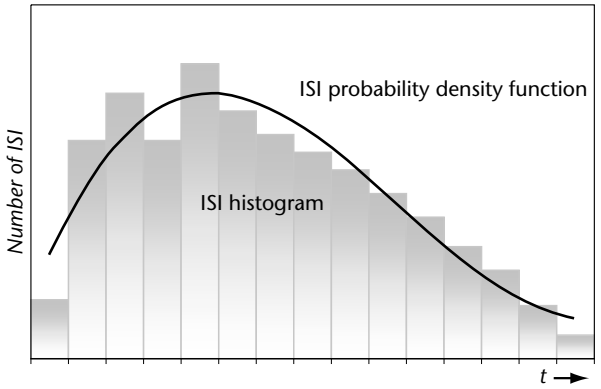


Figure 2. An interspike interval (ISI) histogram and the corresponding first-passage time probability density.

one (suprathreshold stimulation) and the stochastic one (subthreshold stimulation). In the first regime the signal $i(t)$ is large enough, and thus the firing events occur even in the absence of noise, the effect of which is merely to distort the output. In the second (noise-activated) regime the signal alone is insufficient to cause a firing, but the noise itself contributes to activating the neuron. The methods of stochastic resonance extend this view mainly for subthreshold periodic signals. Two sources of periodicity can be expected in the signal: either an endogenous periodicity, or a periodicity of external input, the exogenous periodicity (Lánský, 1997). In both cases, an optimum level σ of noise exists, for which the input frequency is best reflected in the output signal (Bulsara *et al*., 1996; Shimokawa *et al*., 1999). Another type of nonhomogeneity in model (3) is achieved if the natural assumption is made that the amplitude of the noise depends on the signal (Lánský and Sacerdote, 2001). All generalizations mentioned in this section can be applied to any stochastic neuronal diffusion model.

WIENER PROCESS AS A NEURONAL MODEL

If the time constant τ is very large ($\tau \to +\infty$), then eqn (3) takes the form

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = i(t) + \sigma\xi(t), \quad v(t_0) = v_0 \tag{4}$$

This is called the Wiener neuronal model. Although eqn (4) can be taken as a definition of this model, it can also be obtained from first principles using the formalism of diffusion equations. To this purpose, let us initially assume that the neuron is subjected to a sequence of excitatory and inhibitory postsynaptic potentials of constant magnitudes $a > 0$ and $b < 0$ occurring in time in accordance with two independent Poisson processes. The membrane potential is thus viewed as a stochastic process $X(t)$ in continuous time with a discrete space consisting of the lattice $x_0 + ka + hb (h, k = \ldots, -1, 0, 1, \ldots)$ with the points of discontinuity randomly occurring in time. Ricciardi (1977) shows that if the input rates are taken larger and larger while simultaneously taking smaller and smaller postsynaptic potentials with

suitable constraints, the membrane potential 'converges' to the diffusion process generated by eqn (4). This was the method initially used in the well-known paper by Gerstein and Mandelbrot (1964), in which the diffusion approach to neuronal modeling was first considered. By including the spontaneous decay of the membrane potential, which is reflected by a finite value of the time constant τ, in this model with discrete jumps, model (3) can be derived as well.

The basic properties of the Wiener model with constant input, $i(t) = i$, are the following: the distribution of the membrane potential at any instant is normal, with mean $it + v_0$ and variance $\sigma^2 t$. Hence, there is no steady state distribution of the membrane potential, in contrast to the Ornstein–Uhlenbeck model. By means of the methods outlined for instance in Ricciardi (1977), one can prove that the first-passage time density function of the Wiener model is given by

$$g(t|v_0, S) \equiv \frac{\partial}{\partial t} Prob\{T < t\} = \frac{S - v_0}{\sigma\sqrt{2\pi t^3}} \exp\left\{ -\frac{(S - v_0 - it)^2}{2\sigma^2 t} \right\} \qquad (5)$$

which is known in statistical literature as the inverse Gaussian distribution. In this model, for $i \geq 0$, neuronal firing is a sure event. If one takes $i < 0$, density (5) can be interpreted as the firing density conditional upon the event 'firing occurs'. The form of the density (5) permits evaluation of the moments of the ISI and the firing rate, as well as estimation of the parameters of the model. The assumptions of model (4) are oversimplified and many important electrophysiological properties of neuronal membrane are not taken into account. Therefore, the Wiener process is more suitable as a statistical descriptor of the data than as a realistic biological model.

GENERAL DIFFUSION MODELS

Consider a general deterministic model characterized by the dynamic equation

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = h(v, t) + e(v, t)i(t), \quad v(t_0) = v_0 \qquad (6)$$

where $i(t)$ is an input, $e(v, t)$ describes the effect of this input on the depolarization and $h(v, t)$ is the function describing the rate of change of the depolarization in the absence of input. Both e and h are assumed to be sufficiently smooth functions. Then if $i(t)$ can be written as the sum of signal and white noise, more general models than those previously considered arise (Hanson and Tuckwell, 1983). It is indeed a well-known fact, also reflected in the Hodgkin–Huxley model, that the change of the membrane depolarization by a synaptic input depends on its actual value. Basically, the depolarization of the membrane caused by an excitatory postsynaptic potential decreases linearly with decreasing distance of the membrane potential from the excitatory 'reversal potential', say V_E. In the same manner, the hyperpolarization caused by inhibitory postsynaptic potential is smaller if the membrane potential is closer to the inhibitory reversal potential, V_I. In this way, unlike previous models, depolarizations are confined to finite interval (V_I, V_E). Natural conditions relating the reversal potentials, the initial depolarization and the firing threshold is $V_I < 0 < S < V_E$. The diffusion models, which take into account the existence of the reversal potentials result always in the multiplicative noise effect as in eqn (6). This is in agreement with the general notion that an additive noise is generated by external events that transmit messages, whereas the multiplicative noise is generated inside the processing unit, namely inside the neuron.

FEEDBACK, SPATIAL PROPERTIES AND REFRACTORINESS

The analogy of the process describing the time course of the neuron's membrane potential with the laws describing the diffusion of a substance in a liquid provides an intuitive justification for the use of the term 'diffusion model'. It must be stressed that the neuronal behavior described by diffusion models ultimately assumes that for time-constant input, the output is a renewal process. However, one can conceive models aimed, for instance, at simulating clustering effects in spike generation. Serial dependence among ISIs can be modeled in various ways, for instance by adjusting the reset value after each spike (afterhyperpolarization). Another possibility consists of inclusion into the model of some kind of feedback, usually self-inhibition, often experimentally observed. A further generalization is achieved by taking into account the spatial properties of a neuron. In the simplest way, it can be done by assuming that the model neuron is composed of two compartments: the dendritic compartment, described by a standard diffusion model, and the trigger zone compartment, including the spontaneous decay of depolarization and the firing mechanism. Also, the phenomenon of refractoriness can be included in stochastic diffusion models, usually by postulating the existence of time-varying thresholds,

instead of constant thresholds as assumed in the foregoing.

References

Bulsara AR, Elston TC, Doering CR, Lowen SB and Lindberg K (1996) Cooperative behavior in periodically driven noisy integrate-and-fire models of neuronal dynamics. *Physical Review E* **53**: 3958–3969.

Gerstein GL and Mandelbrot B (1964) Random walk models for the spike activity of a single neuron. *Biophysical Journal* **4**: 41–68.

Hanson FB and Tuckwell HC (1983) Diffusion approximations for neuronal activity including synaptic reversal potentials. *Journal of Theoretical Neurobiology* **2**: 127–153.

Inoue J, Sato S and Ricciardi LM (1995) On the parameter estimation for diffusion models of single neurons' activities. *Biological Cybernetics* **73**: 209–221.

Keener JP, Hoppensteadt FC and Rinzel J (1981) Integrate-and-fire models of nerve membrane response to oscillatory input. *SIAM Journal of Applied Mathematics* **41**: 503–517.

Knight BW (1972) Dynamics of encoding in a population of neurons. *Journal of General Physiology* **59**: 734–766.

Lánský P (1983) Inference for diffusion models of neuronal activity. *Mathematical Biosciences* **67**: 247–260.

Lánský P (1997) Sources of periodical force in noisy integrate-and-fire models of neuronal dynamics. *Physical Review E* **55**: 2040–2044.

Lánský P and Sacerdote L (2001) The Ornstein-Uhlenbeck neuronal model with the signal-dependent noise. *Physics Letters A* **285**: 132–140.

Ricciardi LM (1977) *Diffusion Processes and Related Topics in Biology*. Berlin, Germany: Springer.

Ricciardi LM and Sacerdote L (1979) The Ornstein-Uhlenbeck process as a model for neuronal activity. *Biological Cybernetics* **35**: 1–9.

Shimokawa T, Pakdaman K and Sato S (1999) Time-scale matching in the response of a leaky integrate-and-fire neuron model to periodic stimulus with additive noise. *Physical Review E* **59**: 3427–3443.

Further Reading

Chhikara RS and Folks JL (1989) *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. New York, NY: Marcel Dekker.

Gardiner CW (1983) *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Berlin, Germany: Springer.

Karlin S and Taylor HM (1981) *A Second Course in Stochastic Processes*. New York: Academic Press.

Lánský P and Sato S (1999) The stochastic diffusion models of nerve membrane depolarization and interspike interval generation. *Journal of Peripheral Nervous Systems* **4**: 27–42.

Ricciardi LM and Sato S (1994) Diffusion processes and first-passage-time problems. In: Ricciardi LM (ed.) *Lectures in Applied Mathematics and Informatics*, pp. 206–285. Manchester, UK: Manchester University Press.

Tuckwell HC (1988) *Introduction to Theoretical Neurobiology*. Cambridge, UK: Cambridge University Press.

Disorders of Body Image

Introductory article

Giovanni Berlucchi, University of Verona, Verona, Italy
Salvatore M Aglioti, University of Rome, Rome, Italy

CONTENTS

Body image (or schema) is the complex of perceptions, beliefs, and representations about one's own body that is included in the notion of self. Amputations, cerebral lesions, and psychiatric disorders may induce dramatic alterations of this image.

INTRODUCTION

The ability to distinguish self from nonself is a hallmark of all living organisms. In its most elementary expression, in unicellular animals, the distinction is based on simple physicochemical reactions which serve nutritional and self-defence purposes. Multicellular organisms retain such simple reactions for the same purposes, but if they possess a complex brain (particularly a human brain), the self–nonself distinction takes up an entirely different meaning, insofar as it provides the basis for self-perception and personal identity.

Our brain is housed in a body that is at the same time the instrument of all brain-generated patterns of behavior, and the container of all sensory receptors that inform the brain about the external world and about the body. How can our brain distinguish sensory reports about the external world from those about our own body?

A seemingly logical hypothesis is that we know the external world through visual, auditory, and olfactory receptors, all of which respond to stimuli from objects remote from the body, while we know our body through the somatosensory receptors for touch, thermoception, nociception, and proprioception, all of which are acted upon by stimulants directly applied to the body itself. Accordingly, the physiologist Sherrington distinguished between exteroception, which informs the brain about the external world, particularly through the distance receptors, and interoception and proprioception, which inform the brain about states and changes of states in the body.

However, awareness of one's own body can be gained not only through interoception and proprioception, but also through distance receptors that monitor body postures and movements. Vision, for example, is a major source of information not only about the external world, but also about one's body parts that can be seen directly or through mirrors or other reflecting surfaces. Similarly, somatosensory receptors can supply the brain with direct information about the body, but can also be used to explore one's surroundings and to recognize objects by contact.

In short, all kinds of receptors and all sense organs can make their own special contributions to the separate representations of the body and the external world that the brain concurrently entertains.

SELF-CONSCIOUSNESS

Self-consciousness as awareness of one's own being refers to knowledge of one's own mind as well as to knowledge of one's own body. In contrast to purely mental views of self-consciousness, many current psychological and physiological approaches to the concept of self-consciousness emphasize the importance of the awareness of one's own body.

In the nineteenth century the psychologist William James stated that the nucleus of the self is always the bodily existence felt to be present at the time, and that the entire feeling of one's own mental activities is really a feeling of bodily activities, mainly in the head (motor adjustments of eyeballs, eyelids and eyebrows) and the throat (changes of breathing due to movements of the soft palate, posterior nares, glottis, and so on).

Modern analyses suggest that from earliest infancy, simultaneously perceived flows of multisensory information enable us to gain a realistic and accurate perception of the relations between our

own body and our physical environment. This perception, which is perhaps the earliest form of self-knowledge, is complemented – again at an early age – by the perception of reciprocated relations between our own behavior and that of other people: that is, by the sense of the self as an agent and target of social interactions. Babies a few weeks old engage in elaborated social exchanges, loaded with affective meanings, with their mothers and other individuals, and vocalize in response to heard language. Between one and two years of age, interactions with adults lead children to start to build up organized beliefs and memories, and to think that they have traits, attributes, and values. It is around this age that children not only develop language, but also begin to recognize themselves in a mirror – an ability that may be regarded as an objective index of self-consciousness (Figure 1). Among mammals, only humans, chimpanzees, and orangutans are thought to be endowed with the ability of mirror self-recognition.

BODY SCHEMA, BODY IMAGE, AND CORPOREAL AWARENESS

In classical neurology, notions about the functional representation of the body in the brain were first inspired by observations about postural regulation and perception. Even a minor movement of a single body part entails a widespread adjustment of the postural tone affecting most other body parts, suggesting that the brain keeps a moment-to-moment record of the postural state through the entire musculature. This usually unconscious record surfaces to consciousness when a person is asked to report a postural change imposed on the whole body or a part of it by an examiner.

The term 'body schema' was originally coined to denote this current internal model of one's own postural state, serving as the basis for the appreciation of subsequent postural changes. The term has later been extended to refer to a variety of normal performances which are indicative of a general awareness of the body, whether conscious or unconscious. In addition to the ability to appreciate active and passive postural changes and movements, such performances include the abilities to localize tactile stimuli, to move, name and point to specified body parts, and in general to map sensory inputs and motor outputs onto an orderly topographical model of the external anatomy of the human body.

All these abilities suggest the existence of a mental construct, termed 'corporeal or body awareness', which comprises the sense-impressions, perceptions, memories, and ideas about the dynamic organization of one's own body and its relation to other bodies. In current terminology sensory-perceptual components of corporeal awareness are preferentially named 'body schema', while conceptual and imaginative components are preferentially named 'body image'. This distinction, however, is not easy to make in practice, and the two terms are often used as synonyms. The term 'body image' is also employed to refer to the evaluative and emotional judgment, either self-appreciative or

Figure 1. Children begin to recognize themselves in a mirror at between one and two years of age. Before that, as shown by the behavior of this 7-month-old girl in front of the mirror, children tend to interact with their image as if it were another child.

self-critical, that one has of one's own body. Severe distortions of the body image are now thought to underlie psychogenic eating disorders, particularly the syndrome of anorexia nervosa in which people feel grossly fat even if they are extremely underweight.

The Extended Body Schema: Enlargement of the Body's Boundaries

Noncorporeal objects bearing some functional relation of contiguity to the body, such as clothes, ornaments and tools, may come to be felt as parts of the body itself. When one drives a nail into a wall with a hammer, the perceived end of the arm wielding the hammer is not the hand but the head of the hammer itself. This enlargement of the body's boundaries can be accounted for by the same mechanisms that are involved in the representation of the body *per se*. The execution of a voluntary movement of a limb under direct view activates a coherent multimodal representation of that limb in the body schema, based on the congruence between the internal knowledge of the moving command and the visual and proprioceptive feedbacks from the moving limb. An object held by the limb and moving jointly with it takes part in generating such feedbacks and is thus incorporated into the body schema.

A putative neuronal mechanism for including a noncorporeal object into awareness of the body has been found in the anterior parietal cortex of the monkey brain, where there exist neurons that respond to somatosensory and visual stimuli arising from the monkey's hands. If the monkey retrieves food with its hand, the visual receptive fields of these neurons are limited to that hand; but if the retrieval is helped by a rake, the visual receptive fields expand to include both hand and tool, as if they were a unified body part.

Vision and the Body Schema

Inanimate objects can be incorporated into the body schema even without visual feedback, as exemplified by the stick used by blind people for assistance in walking; but the importance of vision is evident from the fact that blindness, especially if congenital, can conspicuously distort the mental representation of the body.

Vision is also crucial for detecting the correspondence between a part of one's own body and the matching part of another person's body, as well as for the imitation of gestures and other movements. A strong innate tendency to imitate sounds and motor acts sets apart humans from other primates, and is probably a prerequisite both for social communication and for the self–nonself distinction. Hours or even minutes after delivery, babies can imitate orofacial and head movements performed by adult models in front of them. This deceivingly simple performance indicates that babies are able to identify a movement of a specific bodily part of the adult model, and to produce a similar movement in the corresponding part of their own anatomy.

This early capacity for visual imitation of elementary actions has suggested that humans are born equipped with a rudimentary body schema which during maturation undergoes a gradual refinement as a consequence of orderly interactions between visual, tactile, proprioceptive, and vestibular inputs. The latter inputs provide the sensory data for the appreciation of head position in the gravitational field, and for the detection of head acceleration and deceleration during linear and circular movements. In adults, the persistence of a systematic reciprocal relation between the visual perception of another person's body part and the representation of one's own corresponding body part is suggested by the finding that visual discrimination of postural changes in the arms of another person is facilitated during movements of the observer's arms but not legs, and vice versa.

Brain Regions Related to the Body Schema

The premotor cortex of the monkey contains neurons which become active when the monkey either performs a goal-directed movement or views a corresponding movement made by another monkey or human. These neurons, if present in the human brain, may provide a simple mechanism for the imitation of actions as well as for the detection of a correspondence between one's own body parts and the matching parts of another individual's body.

Analysis of cortical activation in normal observers during imitation of finger movements has suggested that a region of the left frontal lobe, corresponding to Broca's area, encodes the general goal shared by observed and imitated movements, whereas the precise kinesthetic representation of the same movements is encoded by specific regions of the right posterior parietal lobe, which may also distinguish between observed and self-produced movements.

Anatomically and physiologically the body is represented in a topographic fashion in somatosensory maps in the anterior parietal cortex,

corresponding to the sensory homunculus, and in motor maps in the posterior frontal lobe, corresponding to the motor homunculus, but the relations between these cortical homunculi and corporeal awareness is far from simple. According to Melzack, corporeal awareness relies upon a large neural network where the somatosensory cortex, the parietal lobe, and the insular cortex play crucial and different roles, as suggested by studies of functional brain imaging, brain stimulations, and brain lesions.

Functional brain imaging has shown that a posterior parietal system (superior parietal cortex, intraparietal sulcus, and the adjacent, most rostral part of the inferior parietal lobule) is involved in mental transformations of body in space. The insular cortex is also involved in body awareness, particularly in relation to emotional aspects of this. Insular lesions can cause somatic hallucinations, and electrical stimulation near the insula induces illusions of body position changes and feelings of being outside one's body. Lesions of the cortical sensory homunculus, or of the cortical motor homunculus, induce tactile, proprioceptive, and motor deficits, but there is no evidence that the body parts numbed or paralyzed by such lesions are eliminated from the mental representation of the whole body.

Body awareness may be altered in the context of a diffuse cognitive impairment involving general mental functions such as attention, memory or language; but can some kind of brain damage selectively impair body awareness in the absence of major deficits in other cognitive domains? A positive answer to this question is provided by the striking alterations or mutilations of the mental representation of the body that can be observed after lesions of the cerebral hemispheres, especially the right cerebral hemisphere and the posterior parietal lobe.

Cerebral Lesions and Body Awareness

Patients with lesions of the right posterior parietal lobe may show a neglect (failure to attend to stimuli) of the left half of their body or parts of it. In many cases such symptoms occur in the setting of a general neglect of the left hemispace, both personal and extrapersonal, insofar as patients do not respond to (or mislocalize) all kinds of stimuli coming from their left side. In these cases, neglect of the left half of the body is best attributed to an overall impairment of spatial attention or space representation rather than to a selective disruption of the body schema. In other cases, however, at least some disturbances may be so pronounced or selective as to hint at a specific alteration of body awareness in the form of a *hemisomatagnosia* – that is, a defective or absent knowledge of half of the body.

Patients with right posterior parietal lesions may fail to detect tactile stimuli to the left side of the body even in the absence of basic sensory impairments, or may refer such stimuli to a corresponding position on the right side. Such mislocalization is called *allesthesia* or, in the case of stimuli to the hand, *allochiria*. Patients may even fail to groom or dress the left half of the body, or leave the left half of the face unshaved.

A pathologic attitude towards conspicuous motor and sensory deficits caused by a right hemisphere lesion on the left half of the body may manifest itself in various ways. Patients may appear unaware of their severe deficits to the point of vehemently denying any impairment at all, or they may admit the existence of deficits but show a lack of concern regarding them, or, on the contrary, exhibit hatred for the affected limbs.

The most severe manifestation of neglect for the left body is *somatoparaphrenia*, a condition in which otherwise rational and objective patients exhibit feelings that a body part does not belong to them, and express a resolute denial of ownership of such part. The neglected or disowned body parts are expunged from the mental representation of the body, so that in order to account for the material existence of these parts the patients resort to improbable rationalizations and confabulations: for example, they may claim that their disowned arm belongs to the examiner or to the previous occupant of the hospital bed. Noncorporeal objects previously associated with a disowned hand and included in the body schema, such as a ring, are similarly disowned, although ownership of such objects is immediately acknowledged when they are removed from the disowned hand (Figure 2). Somatoparaphrenia following posterior parietal lesions may also manifest as a distinct and irrepressible feeling of the existence of supernumerary limbs on the affected body half, perhaps suggesting that different aspects of body awareness are localized in different components of the parietal cortex.

Neglect of one side of the body is most commonly observed following lesions of the right posterior parietal lobe, but may also occur following frontal, cingulate or insular lesions, as well as following subcortical lesions in thalamus and basal ganglia, again on the right. Reports of similar symptoms on the right side of the body following left hemisphere lesions are extremely rare. There is,

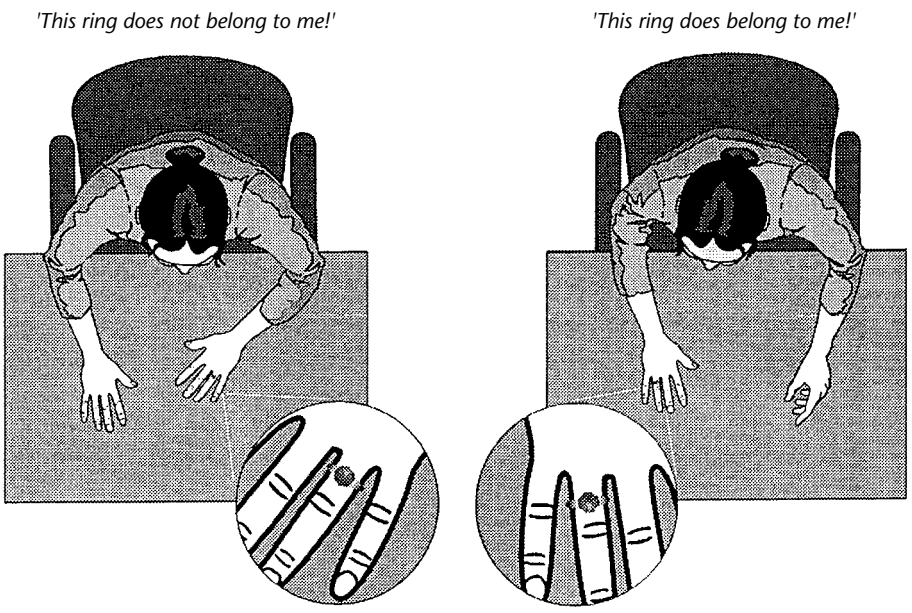


Figure 2. A patient with a right hemisphere lesion exhibited a strong denial of ownership of the left side of the body. She also denied ownership of a ring that she had worn on the left hand as long as the ring was on that hand, but she promptly recognized the ring as her own when it was moved to the right hand. (See Aglioti, Smania, Manfredi and Berlucchi (1996) Disownership of left hand and objects related to it in a right-brain-damaged patient. *NeuroReport* **8**: 293–296).

however, a rare alteration of body awareness, *autotopagnosia*, which seems to depend on a left parietal lesion. Patients with this condition are unable to point to parts of their own or other people's body on verbal command, a disability that cannot be imputed to mental deterioration or language incomprehension because patients can carry out successfully verbal commands unrelated to the body, such as 'touch the pedal of a bicycle'. These patients also have difficulties in describing the spatial relations between body parts: for example, they may say that the mouth is between the nose and the eyes.

Often the difference in efficiency between responses to body-related commands and responses to body-unrelated commands is relative rather than absolute, and it can be argued that at least some cases of autotopagnosia suffer from a general disability to analyze a whole into parts, though such disability is maximally evident in relation to the body.

Autotopagnosia may occur in the setting of the Gerstmann syndrome, also associated with left parietal lesions, which is characterized by dysgraphia (writing disturbances), dyscalculia (calculation difficulties), right–left disorientation, and finger agnosia, the latter being a specific inability to identify and differentiate the fingers of both hands. The relation between autotopagnosia and finger agnosia is a complex one, because patients classified as autotopagnosic may show no finger agnosia, and patients with finger agnosia may perform normally in response to verbal commands targeted to other parts of the body.

The contrast between the symptoms of deranged body awareness associated with left parietal lesions and those associated with right parietal lesions has prompted the following simplistic but effective hypothesis about the relation between the parietal lobes and body awareness: the left parietal lobe houses a conceptual representation of the body, strongly based on a linguistic mediation, whereas the right parietal lobe houses a spatial, nonverbal representation of the body.

Phantom Limb Phenomena

Many people who have undergone amputation report rich and vivid perceptions originating from the amputated body part, which often occur in the form of excruciating pain precisely referred to that part. These 'phantom' perceptions are most common following limb amputation, though they also occur after amputation of other body parts such as the breast in women or the penis in men. Moreover, people with a congenital absence of one or more limbs have been reported to experience phantom perceptions from the limbs that they have never possessed. Even though phantom perceptions are in some sense illusory, they are usually

so distinct and lifelike that patients can, for example, fall down in an attempt to walk with an amputated foot.

The analysis of phantom sensations provides important insights into the mechanisms of corporeal awareness. Peripheral activation of sensory nerves at the amputation scar can contribute to such experience, but there is clear evidence that phantom phenomena have major cerebral causes.

It has been proposed that phantom phenomena are primarily caused by the persisting activity of brain centers that have been deprived of their normal inputs, and by the brain's interpretation of this activity as originating from the lost part. The relevance of cerebral components in determining phantom perceptions is also suggested by their disappearance after lesions to the right posterior parietal lobe. The fact that phantom perceptions are most frequent and vivid following limb amputation is probably due to the functional relevance and the extensive cerebral representation of these body parts. A phantom limb may be perceived as identical in shape to the former real limb, thus suggesting that structures in the central nervous system are committed to the representation of that body part. Furthermore, the persistence of phantom limbs over decades after an amputation attests to a certain degree of stability of somatic representations. This strong tendency of the brain to maintain an intact representation of a lost body part, even a massive one, is perhaps an index of the preservation of the integrity of the self.

Changes in phantom sensations may reflect adjustments or reorganizational changes in the neural substrates representing the lost body part. Phantom limbs, for example, particularly when painful, may shrink in such a way that the hand is perceived as attached directly to the shoulder (the telescoping phenomenon). Complex dynamic aspects of the body schema are also revealed by the recent evidence in limb or breast amputees that vivid phantom sensations can arise as a result of tactile stimulations applied to body regions distant from the amputation line. Sensations in the phantom hand, for example, can be elicited by tactile stimuli delivered to the lower face on the side of the amputation. Like the concurrent veridical facial sensations, the evoked phantom sensations may convey precise information about different features of the facial stimuli. Given the representational contiguity of face and hand, phantom hand sensations from facial stimulation are probably caused by an appropriation of the original cerebral representation of the lost hand by sensory afferents to the adjacent face representation.

NEUROLOGIC OR PSYCHIATRIC DERANGEMENTS OF CORPOREAL AWARENESS

Body dysmorphic disorder, or *dysmorphophobia*, is an enduring excessive concern with a selected bodily flaw which is totally imaginary or grossly exaggerated. Targets for dysmorphophobic concerns may be the shape of the nose, the thickness of the hair, the size of the penis or breast, the appearance of the facial skin, and so forth. Some individuals with dysmorphophobia are aware of the absurdity of their concerns, while many are not; nevertheless, all of them suffer from an intense emotional distress which can disrupt their social and occupational functioning, and indeed their entire life. Patients tend to avoid social contacts in order to conceal their 'ugly' feature from others' view, may obsessively check their appearance in mirrors or, on the contrary, be morbidly afraid of seeing themselves in mirrors or photos. They may seek unnecessary surgical corrections, which invariably fail to eliminate their emotional problem, and may even be driven to such extreme decisions as physically injuring the offensive body part or committing suicide.

Dysmorphophobia differs from anorexia nervosa insofar as people affected by the former condition are concerned with a single feature of their bodily appearance, while in the latter condition it is the overall shape or size of the body that is at the center of the pathologic preoccupation. Moreover, the physical appearance and eating habits of people with dysmorphophobia are normal, whereas those of people with anorexia nervosa are not. Dysmorphophobia usually also differs from obsessive–compulsive disorder, because all patients with the latter typically recognize the absurdity of their disturbing thoughts, which they can at least temporarily hold in check with ritualistic behaviors that are of no help to patients with dysmorphophobia. The modest therapeutic success with serotonin reuptake inhibitors obtained in a few cases of obsessive–compulsive disorder or dysmorphophobia suggests that malfunctions of brain activity regulation by serotonin, a major central neurotransmitter, may underlie both disorders, but no definite notion about these putative pathogenetic mechanisms is yet available.

Body-centered delusions such as underestimation of the size of bodily parts are often observed in major psychiatric illnesses like schizophrenia, where such symptoms are more frequently related to the left body side, and depression, where they are more frequently related to the right body side.

Also, hypochondriac patients tend to refer more frequently to the right side of the body when expressing their complaints (e.g., a pain in an arm). Other deformations of the body image may be experienced during epileptic or migraine attacks, as reported by the British neurologist Macdonald Critchley. For example, a patient with a left parietal meningioma suffered from recurrent attacks of migraine during which the right side would feel bigger and stronger, as if there were a sharp line down the middle. The left side, however, would remain 'calm, cool and collected, while the right side would be tense, anxious, agitated, and highly strung'. These side differences may perhaps be pathologic expressions of the asymmetric functioning of the cerebral hemispheres, but no decisive evidence to support this possibility has been offered.

In heautoscopy (autoscopy) individuals experience a veritable visual hallucination of themselves. A celebrated case is that of the Swedish naturalist Linnaeus who, while examining plants and flowers in his garden, would sometimes see at a little distance his alter ego performing the same actions. A persistent feeling of living outside one's own body characterizes the depersonalization syndrome, while patients with the Cotard syndrome are affected by delusions that their body does not exist, suggesting a specific disorder of corporeal and/or egocentric space awareness.

All the disturbances of the body image described above must have an immediate neural basis, but so far knowledge about the brain activities involved in these conditions is discouragingly small. Even more obscure are the mechanisms by which cultural and social factors act on the brain to produce derangements of the body image. Deep-seated psychological problems related to family and occupational conditions are suspected to be involved in anorexia nervosa, which typically occurs in middle-class young women and shows a high incidence in professional models and ballet dancers. The pathologic attitude toward the body and the related drastic cutdown in eating may be triggered by an exacerbation of a culturally imposed view of thinness as a supreme hallmark of beauty. The endocrine manifestations of anorexia nervosa leave no doubt that the hypothalamus and the adenohypophysis are malfunctioning in this condition, but it is far from clear whether the hypothalamic dysfunction constitutes the primary neural problem or is secondary to starvation and weight loss. Perhaps, in a not too distant future, approaches based on functional brain imaging and other modern neurotechnologies will allow an understanding of the internally generated or externally triggered patterns of cerebral activity which result in this and other distortions of the body image.

CONCLUSION

Studies in nonhuman and human primates indicate that the brain generates separate representations of the body and external world. The term 'body image' (here used as a synonym of body schema) indicates the complex of beliefs, memories, and knowledge about one's own and others' anatomy. This complex mental construct, on which the concept of the self is based, is the product of a continuous interaction between central neural systems dedicated to the representation of the body and peripheral inputs including somatic, vestibular, and visual signals. The ways in which the human body is perceived and represented may change dramatically as a consequence of amputations and cerebral lesions, the effects of which can help localize the neural systems specialized in representing the body. The neural bases of body-related disorders in neurologic and psychiatric diseases are probably due to dysfunctions of these systems as well, but knowledge in this area is still preliminary, though potentially open to improvement thanks to modern approaches to the study of the brain.

Further Reading

Berlucchi G and Aglioti S (1997) The body in the brain: neural bases of corporeal awareness. *Trends in Neurosciences* **20**(12): 560–564.

Bermudez JL, Marcel A and Eilan N (eds) (1995) *The Body and the Self*. Cambridge, MA: MIT Press.

Critchley M (1979) *The Divine Banquet of the Brain*. New York, NY: Raven Press.

Head H and Holmes G (1911) Sensory disturbances from cerebral lesions. *Brain* **34**: 102–254.

Iriki A, Tanaka M and Iwamura Y (1996) Coding of modified body schema during tool use by macaque postcentral neurones. *NeuroReport* **7**: 2325–2330.

James W (1890) *The Principles of Psychology* [reprinted 1950]. New York, NY: Dover.

Meltzoff AN (1990) Towards a developmental cognitive science: the implications of cross-modal matching and imitation for the development of representation and memory in infancy. *Annals of the NY Academy of Science* **608**: 1–31.

Melzack R (1992) Phantom limbs. *Scientific American* **266**(4): 90–96.

Rizzolatti G, Fadiga L, Fogassi L and Gallese V (1999) Resonance behaviors and mirror neurons. *Archives Italiennes de Biologie* **137**: 85–100.

Snodgrass JG and Thompson RL (eds) (1997) *The Self Across Psychology*. New York, NY: New York Academy of Sciences Press.

Dyslexia

Introductory article

Max Coltheart, Macquarie University, Sydney, New South Wales, Australia

CONTENTS

Dyslexia is a specific impairment in the ability to read; it may be acquired (impaired reading caused by brain damage in a previously literate person) or developmental (failure ever to have learnt to read adequately).

INTRODUCTION

When we learn to read, we build up a system in our mind that is capable of turning the printed word into a pronunciation (that is how we read aloud) or a meaning (that is how we understand text). This is an information processing system: it takes in information in one form (the visual appearance of printed words) and gives out information in a different form (the pronunciation or the meaning of the printed words). In some people, however, this system no longer works, even though it used to; and other people never manage to learn the system properly in the first place. Both kinds of difficulty with reading are referred to as 'dyslexia': the first is acquired dyslexia and the second is developmental dyslexia.

There are many people who learnt to read perfectly but who then, because of brain damage (due to a stroke, for example, or an injury to the head), lost some of their ability to read. Such people may still be able to perform normally other mental activities such as remembering, or recognizing people and objects; sometimes they are also normal at speaking, writing, and spelling – but they can no longer read adequately. The brain damage has specifically impaired the mental information processing system which had been used for the job of reading. This is acquired dyslexia.

Other people never manage to learn to read properly – even people who are normal in intelligence and who had no difficulty as children in learning how to carry out other mental activities such as remembering, recognizing people and objects, and arithmetic calculations. This is developmental dyslexia.

If we are to understand these abnormalities of reading, we first need to understand normal reading. We need to know what the mental information processing system used for reading is like, how it works, and how reading is learnt. Experimental psychologists discovered a great deal about this in the last quarter of the twentieth century.

Suppose normal readers are asked to read aloud some nonsense word such as 'bloof'. All will be able to do so. This fact is simple yet instructive. It tells us that reading aloud does not depend solely on being able to look up, in a kind of mental dictionary, what pronunciation had been learnt for the string of letters. If that were so, no one would be able to read 'bloof' aloud, since no pronunciation could have been previously learned for that particular letter string, because the reader would never have seen it before. How do normal readers read aloud such unfamiliar material? This can be achieved because part of what normal readers have learnt is knowledge about the rules relating letters to sounds; they have learnt how 'b' is pronounced, how 'l' is pronounced, how 'oo' is pronounced, and how 'f' is pronounced. These rules apply even to letter strings never before seen; so they can be used for reading aloud such unfamiliar material.

Could this system of rules be all that is needed for adequate reading aloud? No, because there are many words in English that disobey these rules. The rule for how to pronounce 'oo', used to read 'bloof' aloud, is disobeyed by the real words 'blood' and 'good'. Despite this, normal readers will read 'blood' and 'good' aloud correctly if asked to. How do they do this? It can be achieved because these are words these readers have seen before, and this has allowed them to learn the correct pronunciations (and meanings) of the words and store these in a kind of mental dictionary.

ACQUIRED DYSLEXIA

The mental information processing system used for reading thus contains two different procedures for reading aloud – one a system of rules specifying the relationships of letters to sounds; the other a 'dictionary' procedure that allows the reader to retrieve information previously learnt about familiar words. To read aloud unfamiliar material requires the first procedure. To read aloud words that disobey the rules requires the second procedure.

This mental information processing system must be located somewhere in the brain. If different parts of this system are located in different parts of the brain, theoretically brain damage could harm the rule system without affecting the dictionary look-up system. What would the reading of such a person be like? Any familiar word could still be read aloud correctly, since the dictionary look-up system allows this, but unfamiliar letter strings such as 'bloof' could no longer be read aloud. This kind of acquired dyslexia does occur: it is known as 'phonological' dyslexia, and the first case was reported in 1979; numerous other cases have since been reported.

Suppose instead that brain damage harmed the dictionary look-up system without affecting the rule system. The affected person would still be able to read unfamiliar letter strings such as 'bloof', and real words that obey the rules (such as 'bloom') could still be read aloud. However, words that disobey the rules would be affected; if 'blood' could not be looked up in the dictionary, the rules would have to be used to read it aloud, and that will give an error with the 'oo' part of the word. This kind of acquired dyslexia is known as 'surface' dyslexia; the first case was reported in 1973, and numerous other cases have since been reported.

Several other types of acquired dyslexia have been discovered. For example, if it is true that words disobeying the rules are read aloud by looking them up in a mental dictionary, perhaps this involves reading them via information about the meanings of the words: the route is from print to meaning and then from meaning to speech. If so, anyone in whom brain damage has affected knowledge of word meaning would be impaired at reading words that disobey the rules. This turns out not to be so. In Alzheimer disease and other forms of dementia, knowledge about what words mean will sooner or later be lost. In some unfortunate people to whom this has happened, reading aloud of words, even words that disobey the rules, can still be normal. The person might no longer have any idea what the word 'blood' means, yet would still be able to read it aloud correctly. So here we have a form of acquired dyslexia in which what is impaired is reading *comprehension*, with reading aloud being unaffected.

In yet another well-documented type of acquired dyslexia, known as 'deep' dyslexia, the brain-damaged reader will read a word as some other word similar in meaning; the word 'canary' might be read as 'parrot', or the word 'wrist' as 'watch'. Reading by people with deep dyslexia shows a number of other symptoms too. Words that have a concrete meaning ('leopard', 'cigar') are much more likely to be read aloud correctly than words that are abstract in meaning ('idea', 'character'). The small grammatical words of the language, even though they are generally very short and very common ('the', 'and', 'if', 'but'), are rarely read aloud correctly. Unfamiliar letter strings such as 'bloof' cannot be read aloud at all.

Finally, there is 'letter-by-letter reading', so called because sufferers from this kind of acquired dyslexia typically spell out aloud words they are trying to read, letter by letter, rather than being able to recognize the word immediately as a whole. If they manage to name each letter in the word correctly, then they can generally say the whole word correctly. Despite this severe difficulty in reading, such people can be perfectly normal at writing and spelling.

The discovery of these distinct patterns of acquired dyslexia tells us not only that the mental reading system consists of numerous different components, but also that these different components are located in different parts of the brain; if that were not so, brain damage could not affect one part of the reading system while leaving other parts unaffected.

Only a limited amount is so far known about exactly which parts of the brain are damaged in these different kinds of acquired dyslexia. There is persuasive evidence that the limited reading achievable by people with deep dyslexia depends upon the use of reading mechanisms in the right hemisphere of the brain which may play little or no part in normal reading. Letter-by-letter reading also involves the right hemisphere, since these patients typically have damage to the visual areas of the left hemisphere (so that visual identification of letters cannot occur in the left hemisphere), and also damage to one part of the corpus callosum (which is responsible for transmitting information between the two hemispheres of the brain). The effect of this damage is that after letters are identified in the right hemisphere their transmission to

the reading system in the left hemisphere is abnormally slow and inefficient.

When knowledge about word meanings is affected in Alzheimer disease and other forms of dementia (in patients with impaired text comprehension but intact ability to read aloud) this is due to progressive deterioration of the temporal lobes of the brain, especially the left temporal lobe. Virtually nothing is currently known about which brain regions are specifically impaired in surface dyslexia or phonological dyslexia.

One might assume that, since in acquired dyslexia some part of the brain that is needed for reading is permanently damaged, there could be no effective treatment to improve the reading of such people; but this turns out not to be so. Previously normal readers who no longer recognize the word 'blood', and so can only try to read it using the rules, can learn to recognize it again; that is, surface dyslexia can respond to treatment if this is designed appropriately. Phonological dyslexia, deep dyslexia, and letter-by-letter reading have also been shown to be remediable by appropriate treatment.

DEVELOPMENTAL DYSLEXIA

Turning now to a consideration of difficulties in learning to read – developmental dyslexia – a natural question to ask is the following. If the different components of the reading system can be separately impaired by various forms of brain damage to produce different kinds of acquired dyslexia, is the particular component of the reading system causing the difficulty in learning to read different in different children? If that were so, there would be different kinds of developmental dyslexia, just as there are different kinds of acquired dyslexia. This turns out to be the case.

Some children with developmental dyslexia have a particular problem in learning or using the system of rules specifying the relationships of letters to sounds. This is known as 'developmental phonological dyslexia'. Other children with developmental dyslexia have a particular problem learning or using the dictionary look-up system needed for fluently recognizing words as wholes – this is developmental surface dyslexia.

It is clear that developmental dyslexia is partly a genetic condition involving abnormalities on at least two specific chromosomes, namely chromosome 6 and chromosome 15. It may be the case that the abnormality of chromosome 6 is characteristic of developmental phonological dyslexia and is associated with some general difficulty on processing speech sounds; and it may also be the case that the abnormality on chromosome 15 is characteristic of developmental surface dyslexia.

Since there are genetic influences at work here, one might wonder whether it is possible to remedy these developmental dyslexias. However, it does not follow from the fact that a disorder has a genetic basis that it cannot be treated; and in fact research has shown that both of these kinds of developmental dyslexia respond well to appropriate treatment.

Many children with developmental dyslexia exhibit the symptoms of both of these kinds of developmental dyslexia: they are poor at both learning or using the system of rules specifying the relationships of letters to sounds and also at learning or using the dictionary look-up system needed for fluently recognizing words as wholes. The most plausible explanation for the prevalence of this commonly occurring 'mixed' developmental dyslexia is that each of these reading procedures assists the child in learning the other. A child who is, say, 8 years old will have an auditory vocabulary of perhaps 10 000 words, but a small vocabulary of words that can be recognized in print. Such a child will therefore constantly come across printed words which cannot be recognized in print but which would be recognized if heard. If the child could apply rules relating letters and sounds to such words, sounding out the words would allow them to be recognized. This is a method by which children could teach themselves to recognize whole words; but it is unavailable to the child who is poor at using such rules. So being poor at the rule-based reading route will make it difficult to build up the dictionary look-up reading route.

The reverse is probably also true. How do children learn what these rules are? One way might be by reflecting on the spellings and pronunciations of words they already know, and working out from these what the rules must be. If so, children who have few words in their mental dictionary will have an impoverished database upon which to reflect, and that will limit what these children can learn about letter–sound rules.

CONCLUSION

Great advances have been made in our understanding of developmental and acquired dyslexia. The crucial step was the recognition that the system we use to read is a mental information processing system that contains a number of different processing components, including a letter recognition system, a mental dictionary storing the spellings,

meanings and pronunciations of the words the reader knows, and a system of rules specifying the relationship between letters and sounds. Studies of people with acquired dyslexia have shown that brain damage can impair particular components of the reading system while not affecting others. Studies of people with developmental dyslexia have shown that such people can have difficulty learning particular components of the reading system while being able to learn others well; and such specific difficulties in learning to read are associated with specific genetic abnormalities.

Further Reading

Castles A and Coltheart M (1993) Varieties of developmental dyslexia. *Cognition* **47**: 149–180.

Castles A, Datta H, Gayan J and Olson RK (1999) Varieties of developmental reading disorder: genetic and environmental influences. *Journal of Experimental Child Psychology* **72**: 73–94.

Coltheart M and Jackson N (1998) Defining dyslexia. *Child Psychology and Psychiatry Reviews* **3**: 12–16.

Habib M (2000) The neurological basis of developmental dyslexia: an overview and working hypothesis. *Brain* **123**: 2373–2399.

Jackson N and Coltheart M (2001) *Routes to Reading Success and Failure*. Philadelphia, PA: Psychology Press/Taylor & Francis.

Marshall JC (1989) The description and interpretation of acquired and developmental reading disorders. In: Galaburda AM (ed.) *From Reading to Neurons: Issues in The Biology of Language and Cognition*, pp. 69–86. Cambridge, MA: MIT Press.

Seymour PHK (1990) Developmental dyslexia. In: Eysenck MW (ed.) *Cognitive Psychology: An International Review*, pp. 135–196. New York, NY: John Wiley.

Electroencephalography (EEG)

Introductory article

Terence W Picton, Rotman Research Institute, Toronto, Ontario, Canada
Ali Mazaheri, Rotman Research Institute, Toronto, Ontario, Canada

CONTENTS

Electroencephalography is a measurement of the brain's electrical activity. It provides information about the timing of intracerebral processes which can be used in conjunction with anatomical information derived from hemodynamic studies to learn about events in the human brain.

INTRODUCTION

Electroencephalography (EEG) is the recording of the electrical activity of the brain. When the neurons of the brain process information, they do so by changing the flow of electrical currents across their membranes. These changing currents, particularly those caused by the synaptic excitation and inhibition of cortical neurons, generate electric fields that can be recorded using small electrodes attached to the surface of the scalp. The potentials between different electrodes are amplified and displayed as they fluctuate over time.

The human EEG was first recorded in the 1920s by a German psychiatrist named Hans Berger. Since then, it has been used widely to investigate normal and abnormal brain function.

The EEG is characteristically recorded simultaneously from multiple locations on the scalp. The amplitude of the EEG is very small – usually several tens of microvolts. The recordings use special amplifiers that record the differences in potential between two electrodes. These differential amplifiers cancel out other large electrical activities, such as line noise, and allow us to see the small EEG signals. Two types of recordings are possible: the electrical activity can be recorded from each location relative to a common reference, or the activity can be recorded from each electrode relative to an adjacent electrode. The latter technique effectively records the slope or change of the scalp potentials over space rather than their absolute value relative to a single reference. Figure 1 shows the EEG signals recorded from ten scalp electrodes. The recordings are all relative to a linked-mastoid reference. The signals in Figure 1 are plotted according to the usual EEG convention that negativity at the first electrode relative to the second is shown as an upward deflection.

Electroencephalographic recordings fluctuate in time and are often 'rhythmic' in the sense that they alternate regularly. The most prominent rhythm in the EEG is the alpha (α) rhythm, which has a frequency of 8–13 cycles per second (Hz) and is recorded mainly over the posterior regions of the scalp close to the regions of the brain that process visual information. When the eyes are open the α rhythm is very small and when the eyes are closed it becomes large.

During a rhythm the neurons tend to fire synchronously, and their fields overlap and add to each other to cause the large scalp potentials. Interactions between cortical and thalamic neurons can cause the cortical neurons to fire periodically and thus generate rhythmic scalp potentials. When the visual areas are activated by real or imagined information, the neurons fire independently and their fields tend to cancel each other out. The EEG is then said to be 'desynchronized'. The transition between the α rhythm and the desynchronized EEG occurring with eye opening is shown in Figure 1.

As well as activity generated in the brain, the electrodes also pick up electrical potentials from other sources in the head. The eyes are the most prominent of these sources. Large potentials occur in the anterior regions of the scalp as the eyes or eyelids move. In Figure 1, the anterior regions become more negative as the eyes open, and there is a brief positive deflection during a blink. The scalp muscles also generate activity that is picked up in the EEG. This activity is characteristically faster than the activity arising from within the brain.

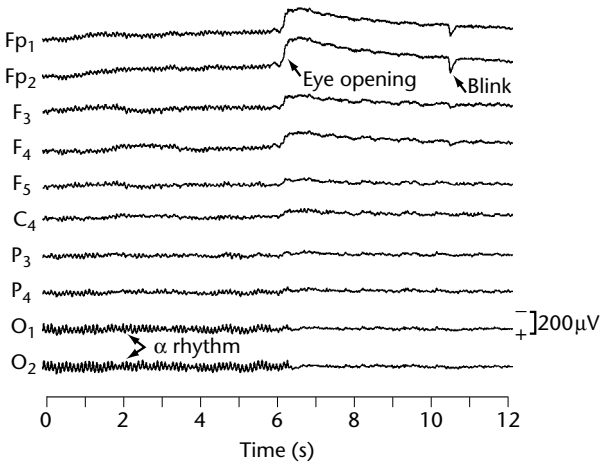


Figure 1. Human electroencephalogram (EEG). This 12 s recording was taken from a normal young woman. The EEG was recorded from ten scalp electrodes, with the activity at each electrode measured relative to a linked-mastoid reference. Negativity at the scalp electrode relative to the reference is plotted upwards. The electrodes are named according to their location on the scalp: Fp, frontopolar (forehead); F, frontal; C, central; P, parietal; O, occipital. Odd-numbered electrodes are on the left and even-numbered electrodes are on the right, and the number varies with the distance from the midline. At the beginning of the recording, the eyes were closed; halfway through the recording the woman opened her eyes. Movements of the eyes and eyelids when the eyes opened were recorded as a large negative wave in the Fp electrodes. While her eyes were open, the woman blinked; this was recorded as a brief positive wave in the Fp electrodes. While the eyes were closed there was a sustained rhythmic oscillation at 10 Hz in the posterior electrodes (O_1 and O_2) – the alpha rhythm.

FREQUENCY ANALYSIS

Electroencephalographic signals are usually plotted as changes in voltage over time, as in Figure 1. Because of the rhythmic nature of the signals, it is often informative to plot the signals according to the frequencies they contain. This change from the time domain to the frequency domain is accomplished using a Fourier transform. The frequency spectrum of the EEG signal then shows various peaks that denote the particular rhythms in the EEG. The frequency representation of the EEG recorded from the right occipital electrode at the back of the head (labeled O_2 in Figure 1) is shown in Figure 2. The α rhythm stands out as a peak in the spectrum recorded when the eyes are closed.

The spectrum of frequencies present in the EEG is broad. As well as the (α) activity recorded at frequencies of 8–13 Hz, the EEG also contains slower activity – theta (θ) activity at 4–7 Hz and delta (δ) activity at 0–3 Hz – and faster activity – beta (β) activity at 14–25 Hz and gamma (γ) activity at 25–50 Hz.

TEMPORAL ANALYSIS OF THE EEG SPECTRUM

The way in which the frequencies of the EEG change over time can show the changing state of the brain. Figure 3 shows the desynchronization of the α rhythm as the eyes open. The upper tracing shows the time-domain EEG recorded from the O_2 region of the scalp. The middle tracing shows the way the spectrum changes over time (EEG spectrograph). Spectra similar to those shown in Figure 2 are plotted with the frequency on the *y* axis and the amplitude of the activity demonstrated using the color scale (*z* axis). Whereas the two spectra in Figure 2 combine activity over several seconds, the multiple spectra plotted in Figure 3 are computed about 40 times a second. The α activity, which is present with the eyes closed and then goes away with eye-opening, shows up as the prominent dark red line at a frequency of 10 Hz. The lower tracing shows the amplitude of activity in the α frequency band over time. This follows the α rhythm over the time course portrayed in Figure 1.

The α rhythm can attenuate even when the eyes are closed. This can occur when the person is using the visual areas of the brain in processes such as problem-solving or imagining visual information. Sometimes the changes in the rhythms following an event (such as a stimulus) are small and can be measured only if the spectral changes are averaged over multiple trials.

SPATIAL ANALYSIS

The EEG signals are characterized by their distribution over the scalp as well as by the way they change in time. The EEG is spatiotemporal in nature. The visually reactive α rhythm is usually distinguished from other EEG activity by its posterior scalp location. Figure 4 shows the scalp topography of the α activity recorded in Figure 1. The α rhythm is typically slightly larger over the right posterior scalp than over the left.

Other rhythmic activities are recorded from different regions of the scalp. The mu rhythm has a similar frequency to the α rhythm but is located over the sensorimotor regions of the cortex and reacts to somatosensory input, motor activity or the thought of motor activity in the contralateral

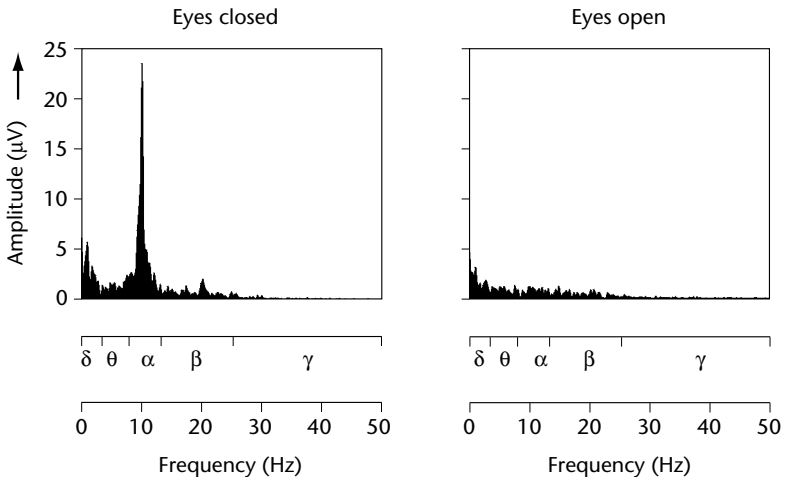


Figure 2. Frequency analysis of the electroencephalogram. The frequency content of the EEG signal at the O_2 electrode (lowest tracing in Figure 1) is shown. The spectrum on the left shows the amplitude of activity at the different frequencies when the eyes are closed (from 1 s to 5 s in Figure 1) and the spectrum on the right shows the pattern when the eyes are open (from 7 s to 11 s). The eyes-closed spectrum shows a prominent peak at 10 Hz.

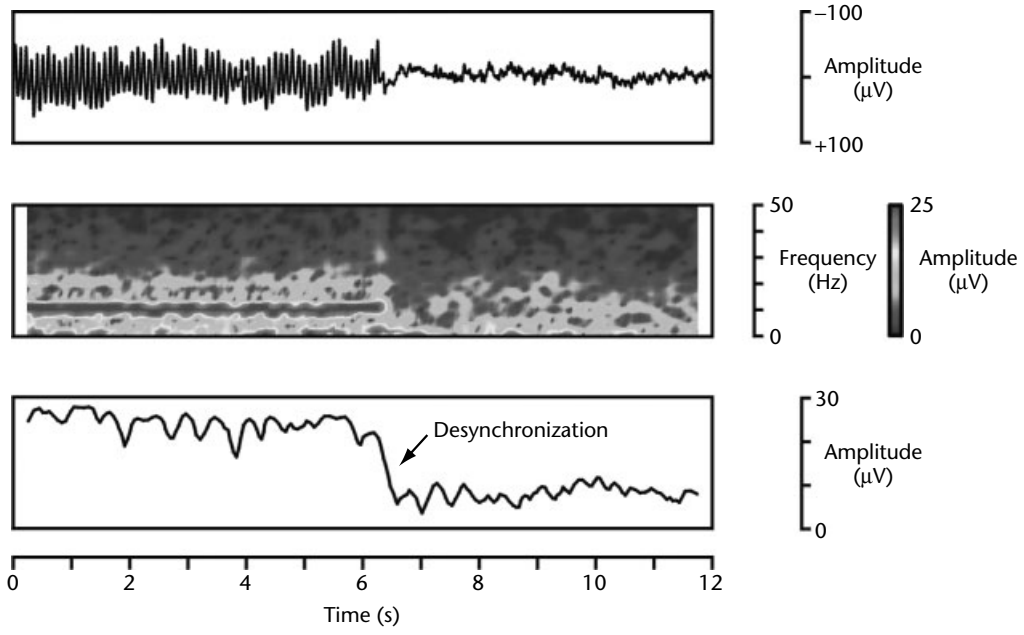


Figure 3. [*Figure is also reproduced in color section*.] Electroencephalographic changes over time. The upper tracing shows the EEG signal recorded from the O_2 electrode in Figure 1. The middle panel shows the spectrogram. This plots the changes in the spectrum over time: the frequency (on the *x* axis in Figure 2) is plotted vertically (*y* axis) and the amplitude of the activity is plotted on the *z* axis as color. A dark red line at 10 Hz represents the α rhythm. This line is clearly recognizable when the eyes are closed and disappears when the eyes are open. The lower tracing represents the amount of activity present in the 8–13 Hz frequency band as it changes over time. When the eyes open, this activity decreases (event-related desynchronization).

hand. Its name comes from the 'm' shape of the waves.

The γ rhythms of the brain have been studied in relation to memory and perception. These rapid rhythms may serve to bind separate aspects of a perceived object by synchronizing relevant neurons from different regions of the brain. The rhythms may also act as a signature for that

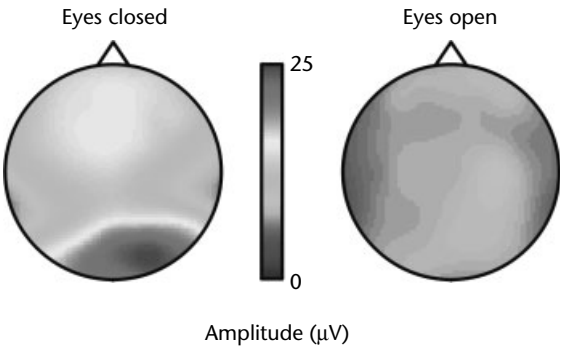


Figure 4. [*Figure is also reproduced in color section.*] Topography of the alpha rhythm. The scalp is viewed from above using an azimuthal equidistant projection centered at the vertex. The outside of the circle reaches the level of the ear canal. The α activity when the eyes are closed is maximally recorded from the posterior regions of the scalp, and is slightly greater on the right than on the left. These maps were based on activity from 65 scalp electrodes (10 of which are shown in Figure 1).

bound information. Bursts of EEG γ activity can occur as a stimulus is perceived or maintained in short-term memory. (*See* **Neural Oscillations**)

CHANGES IN THE STATE OF THE BRAIN

Changes in the EEG as the brain changes its state are most clearly seen during sleep. Many specific patterns of activity define various sleep stages. During a normal night of sleep, the brain will cycle through these various stages once every 90 min or so. In one stage, often associated with vivid visual dreams, there are many rapid eye movements (REM). Another stage of sleep is associated with widespread δ activity – slow-wave sleep. Bursts of activity with frequencies around 14 Hz – sleep spindles – are prominent during the transition between REM sleep and δ sleep. (*See* **Sleep and Dreaming**)

The EEG is used extensively to assess the abnormal brain states that occur with neurological disorders. An abnormal decrease of brain activity is usually associated with slow EEG waves. These can occur in localized regions of the scalp over areas of focal brain damage, or can be more widespread in cases of generalized brain dysfunction. After extensive brain damage there may be no electrical activity recorded from the brain – a state known as 'brain death'. An abnormal excitability of the neurons in the brain, which can occur in epilepsy, is usually associated with high-frequency EEG activity. The most common abnormality is a brief, sharp change in voltage called a spike. Spikes may be generated focally in one region of the brain or may be more widespread in association with a generalized disturbance of consciousness. A common generalized pattern is a combination of spike and wave that recurs at a rate of $3\,s^{-1}$. (*See* **Epilepsy**)

OTHER ANALYSES OF THE EEG

In order to look at the response of the brain to a stimulus or the activity in the brain occurring before a motor act, the process of averaging can be used to distinguish the cerebral activity specifically related to the event from the other activity of the EEG: the stimulus or the act is repeated and the EEG signals related to each occurrence of the event are averaged together. The specific activity – 'event-related potential' – remains the same during averaging, whereas the unrelated EEG activity tends to cancel itself out. The event-related potentials can show what is happening in the brain when a person processes the information in a stimulus and then prepares and executes a behavioral response. Averaging can also be used to assess spectral information (instead of time waveforms) in order to show subtle changes in the frequencies of the EEG in association with stimuli or responses – a process called 'event-related synchronization and desynchronization.' (*See* **Event-related Potentials and Mental Chronometry**; **Auditory Event-related Potentials**; **Visual Evoked Potentials**)

The flow of information from one area of the brain to another can be evaluated by measuring the correlation or coherency between the EEG signals recorded in each area. The amount of correlation can indicate the amount of information transferred, and the time lag can indicate the direction of the transfer. Unfortunately, since the EEG signals recorded from the scalp spread quite widely, much of the correlation between these signals may be caused by current spread in the scalp rather than information spread in the brain. Techniques that calculate the radial currents or extrapolate back from the scalp to the cortical surface can make these studies more precise. It would be even better to determine the actual intracerebral generators for the scalp-recorded activity and to correlate the waveforms generated in these sources. Although there is no unique solution to the problem of calculating the intracerebral sources for electrical signals recorded at the scalp surface, constraints imposed on the solutions from other imaging techniques, which provide anatomy and cortical activation patterns, can lead to sensible solutions.

RELATIONS TO OTHER BRAIN IMAGING METHODS

As well as generating electric fields in the extracellular fluid, the passage of currents through neuronal membranes generates magnetic fields. A magnetoencephalogram (MEG) can therefore be recorded from the surface of the scalp using specialized techniques. These signals are similar to EEG signals since they derive from the same currents in the brain; they differ, however, in terms of which neurons contribute to the scalp-recorded signals and how these signals spread to the recording sensors.

Both the EEG and the MEG are 'functional' tests. They differ from simple structural measurements of brain anatomy by measuring the activity of the brain as it is working. The EEG is therefore related to other tests of brain function that measure cerebral blood flow, such as positron emission tomography or functional magnetic resonance imaging. One difference between these hemodynamic measurements and the electromagnetic measurements is in the timing. The electromagnetic activity changes simultaneously with the neuronal activity, whereas the blood flow changes after a delay. However, blood flow measurements are much more accurate in terms of their anatomical localization. Activation patterns from hemodynamic studies may provide localization for the temporal changes in the EEG – either event-related potentials or event-related desynchronizations. (*See* **Neuroimaging**)

Further Reading

Aminoff MJ (1999) *Electrodiagnosis in Clinical Neurology*, 4th edn. New York, NY: Churchill Livingstone.

Dale A, Liu AK, Fischl BR *et al*. (2000) Dynamic statistical parametric mapping: combining fMRI and MEG for high resolution imaging of cortical activity. *Neuron* **26**: 55–67.

Gevins AS, Le J, Brickett P, Reutter B and Desmond J (1991) Seeing through the skull: advanced EEGs use MRIs to accurately measure cortical activity from the scalp. *Brain Topography* **4**: 125–131.

Gloor P (1969) *Hans Berger on the Electroencephalogram of Man* (*Electroencephalography and Clinical Neurophysiology*: supplement 28). Amsterdam, Netherlands: Elsevier.

McFarland DJ, Miner LA, Vaughan TM and Wolpaw JR (2000) Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topography* **12**: 177–186.

Niedermeyer E (1997) Alpha rhythms as physiological and abnormal phenomena. *International Journal of Psychophysiology* **26**: 31–49.

Niedermeyer E and Lopes da Silva F (1998) *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, 4th edn. Philadelphia, PA: Lippincott, Williams & Wilkins.

Nunez PL (1995) *Neocortical Dynamics and Human EEG Rhythms*. Oxford, UK: Oxford University Press.

Nunez PL, Silberstein RB, Shi Z *et al*. (1999) EEG coherency II: experimental comparisons of multiple measures. *Clinical Neurophysiology* **110**: 469–486.

Pfurtscheller G and Lopesda Silva F (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology* **110**: 1842–1857.

Singer W (2000) Response synchronization: a universal coding strategy for the definition of relations. In: Gazzaniga MS (ed.) *The New Cognitive Neuroscience*, pp. 325–338. Cambridge, MA: MIT Press.

Steriade M (1998) Cellular substrates of brain rhythms. In: Niedermeyer E and Lopes da Silva F (eds) *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, 4th edn, pp. 28–75. Philadelphia, PA: Lippincott, Williams & Wilkins.

Tallon-Baudry C and Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences* **3**: 151–162.

Wong PKH (1991) *Introduction to Brain Topography*. New York, NY: Plenum.

Emotion, Neural Basis of

Introductory article

Jeannine V Morrone-Strupinsky, University of Arizona, Arizona, USA
Richard D Lane, University of Arizona, Arizona, USA

CONTENTS

Emotion is information about the extent to which goals are being met in interaction with the environment.

INTRODUCTION

Emotion is information about the extent to which goals are being met in interaction with the environment. Emotion is a mechanism that emerged in evolution to solve the problem of attributing value to classes of stimuli based on experience, and serves as a means of adapting rapidly to critical stimuli in the environment. It can be viewed as an extension of more rudimentary or reflex-based behavioral systems. For example, mating behavior that is reflexive in reptiles is elaborated and extended in mammals and contributes to the formation of social bonds.

Emotion provides a mechanism for rapidly resetting the organism in response to environmental contingencies. This resetting occurs cognitively (e.g. in attention and memory systems), physiologically (e.g. preparing for exertion, diminishing functions not needed during a crisis, such as digestion), and behaviorally (e.g. shifting the propensity for approach or avoidance behavior). Emotional information can be conveyed internally (experiential) or externally (expressive, for instance gestures or facial expressions), the latter reflecting the inherent social element in emotion, that of signaling to or communicating with other animals.

Thus, emotion involves evaluating the motivational significance of a stimulus and subsequently implementing motivated behavior. The function of emotion is manifold. In response to some change in the environment, emotions (1) interrupt behavior and focus attention on particular elements in one's surroundings, or on one's internal sensations or appraisals; (2) physiologically prepare the organism for alternative action; (3) serve as a means of communication among conspecifics; and (4) help label or 'tag' memories for significance, which can guide approach/avoidance tendencies in related situations in the future, as well as trigger memories of previous experiences that may affect progression of the current behavioral agenda.

WHAT INFLUENCES THE DEVELOPMENT OF EMOTIONS?

There has been spirited discussion about whether emotions are rooted in nature (i.e. genetically hardwired neural circuits) or nurture (learning via experience). Indeed, it has been demonstrated that we possess certain rudimentary emotional functions at birth. However, these basic functions are then subjected to learning and environmental experiences of many kinds. For instance, Rene Spitz's observations of infants raised in hospitals illustrate that social development is disrupted if infants are deprived of appropriate emotional responses from the environment. More than a third of the infants died, and of those who survived, most experienced developmental delays. The effect of the environment is not well understood, but it is likely to be an important source of the heterogeneity observed between people in their emotional behavior. For instance, research shows that offspring of depressed caregivers are at increased risk of maladaptive development and emotional difficulties.

THE STRUCTURE OF EMOTION

Precisely how emotion is organized in the brain remains an unresolved issue. There is debate over whether different emotions should be considered as discrete entities, or viewed from a dimensional

perspective. Proponents of the discrete or basic emotions model contend that there are unique circuits for particular emotions, such as fear, anger, joy, and disgust. Darwin suggested that distinct facial expressions reflected distinct neural circuits for each emotion. This is supported by evidence of the generation and recognition of prototypical facial expressions of emotion in all known cultures. Furthermore, specific neural circuits have been identified in animals by Panksepp for basic emotional states such as rage, fear, joy, and sorrow.

The dimensional approach postulates that emotion is founded on separable motivational systems, such as approach and avoidance (appetitive versus aversive), and can be subdivided into components of valence (unpleasantness–pleasantness) and arousal (unaroused–aroused). Dimensional models of valence are further bifurcated into bivariate and bipolar models; in the former positive and negative emotions are conceptualized as separate dimensions, in the latter they are viewed as opposite ends of the same continuum. Dimensional models can incorporate discrete emotions models, but discrete emotions models cannot encompass dimensional models. For example, discrete emotions can be mapped onto the plane defined by the arousal dimension on the vertical axis and the valence dimension on the horizontal axis.

Substantial evidence from psychometric and psychophysiologic studies in healthy people supports a dimensional perspective. For instance, Watson, Tellegen and colleagues have found that self-report of mood is organized along positive and negative affective dimensions. In addition, Lang and colleagues have demonstrated the validity of the dimensional perspective using emotional stimuli in different modalities (pictures and sounds) to evaluate both subjective emotional and psychophysiological response. Patterns of psychophysiological responses are consistent across positive or negative valence of stimuli, but there are no unique physiological signatures for discrete emotions. Furthermore, a cognitive neuroscientific perspective suggests that the brain mediates various cognitive functions by combining component processes from different brain regions (such as the visual system) rather than dedicating specific neurons or specific regions to one particular function, such as recognition of a particular face. Indeed, attempts to define unique circuitry for discrete emotions in human imaging studies have failed to yield nonoverlapping activation patterns. For instance, Lane and colleagues found that the discrete emotions of happiness, sadness, and disgust were each associated with increases in activity in the thalamus and medial prefrontal cortex. These three emotions were also associated with activation of anterior and posterior temporal structures, particularly when induced by film. Recalled sadness was associated with increased activation in the anterior insula. Happiness was distinguished from sadness by greater activity in the vicinity of ventral mesial frontal cortex. From these findings Lane and colleagues concluded that there are both common and unique components of the neural networks mediating discrete emotions.

A function as complex as emotion is likely to be mediated by the coordinated effort of a variety of subsystems. Much is known about which structures contribute to the cognitive elaboration of emotion. This process requires the conscious awareness of emotions in order to understand the origin and meaning of one's emotions, which can subsequently be incorporated into thought and behavior. However, it is not yet known which brain structures should definitely be included or excluded in this network, and our understanding of how the system works as a whole to mediate the variety of emotions that are observed behaviorally is limited.

HISTORICAL THEORIES OF THE NEURAL BASES OF EMOTION

Study of the neural bases of emotion was neglected throughout most of the nineteenth and twentieth centuries because the experience of emotion was considered too subjective and vague to study scientifically. Furthermore, behaviorist and cognitive approaches predominated in psychology during the second half of the twentieth century.

Visceral feedback theories of emotion are useful from a historical perspective and remain influential in contemporary work on emotion. Their main contribution is the realization that the nature of emotion is a combination of brain and bodily states. Although the brain probably has the more prominent role in the generation of emotions, the experiential aspect of emotions is influenced to a large degree by bodily state and the transmission of this information to the brain.

There are three main visceral feedback theories of emotion. The James–Lange theory of emotion states that the experience of emotion follows bodily feedback: that is, perceiving an emotional stimulus generates visceral sensations or autonomic arousal, which then mediates the experience of particular emotions. Put another way, emotion is simply the perception of visceral sensations, and each discrete emotion has a unique pattern of physiological

arousal. The Cannon–Bard theory focuses on the central nervous system (which includes the brain and the spinal cord), particularly on the thalamus, a structure responsible for gating sensory input. Perceiving an emotional stimulus simultaneously generates an emotional state and induces bodily arousal via the sympathetic nervous system. The pattern of physiological arousal does not correspond with specific emotions. According to the Schachter–Singer theory, the experience of emotion is a combination of generalized physiological arousal and cognitive appraisal. Arousal occurs first in the chain of events, and one's appraisal of this inner state 'labels' the specific emotion being experienced. Thus, the Schachter–Singer theory agrees with the James–Lange theory that physiological arousal is necessary for emotion, but it differs from James–Lange theory in that it argues that each discrete emotion does not have a unique pattern of physiological arousal.

The three models therefore address 'bottom up' and 'top down' mechanisms that determine what the emotion is and how it is experienced. It is difficult to identify clearly a unique autonomic signature for basic emotions. To the extent that autonomic differences are observed between emotions, they are subtle and difficult to detect. (*See* **Autonomic Nervous System**)

Several neuroanatomical theories were proposed in the first half of the twentieth century as the basis for emotional function. They are not generally accepted today in their original forms, but elements of these theories hold some validity and served as the basis for later elaborations. The Papez circuit was the first specific neuroanatomical theory of the subjective experience of emotion. It identified a region of the brain for which the function was not known and attributed to it a function (emotion) that did not have a neural basis at the time. The Papez neural circuit of emotion detailed the flow of information from the sensory nuclei of the thalamus to the sensory cortex and on to the cingulate cortex, and from the thalamus to the mamillary bodies of the hypothalamus. It did not include the amygdala and orbitofrontal cortex, regions that are well supported today as integral to emotional responses, but did include the hippocampus and cingulate cortex; the latter probably participates in emotional experience. MacLean's theory of the triune brain proposed that the generation of emotions occurs in an evolutionarily old region of the brain of mammals ('paleomammalian'), which he deemed the limbic system or the 'visceral' brain. This region is located between the 'reptilian' brain, which controls basic and reflex motor functions, and the 'neomammalian' brain, which mediates higher cognitions. Like visceral feedback theories, the triune brain theory suggested that the subjective experience of emotion stems from the confluence of sensory input from the external world and internal visceral sensations in the limbic system. The concept of the limbic system is a subject of intense debate, as it has not been proved whether all of the structures included in the concept behave as a system in relation to emotional function, and the number of structures to be included is still undecided. The term persists, however, because a better alternative has not yet been found.

EMPIRICAL STUDIES OF EMOTION

The neural substrates of emotion have received considerable attention from researchers such as Damasio, LeDoux, Panksepp, and Rolls. Their work has largely been based on experimental findings in animals and observations in people with brain lesions. More recently functional brain imaging techniques such as positron emission tomography and functional magnetic resonance imaging have been used to study the neural substrates of emotion in healthy volunteers and, to a lesser extent, in clinical populations. Figure 1 illustrates the main brain structures that contribute to emotional function.

Studies of Emotion in Animals

Understanding of the neural bases of emotion has been informed substantially by experimental studies in animals. Lesion studies by Cannon and Bard led to the conclusion that the hypothalamus is essential to organized emotional responses. Since the hypothalamus regulates the autonomic nervous system, the Cannon–Bard theory proposed that bodily responses during emotional states are controlled by higher centers via the function of the hypothalamus. Concurrently, the subjective experience of emotion relies on input from the hypothalamus to the cortex, and thus emotional responses in decorticate animals were labeled as 'sham rage'.

In the 1930s Klüver and Bucy demonstrated that extensive lesions of the anterior temporal lobe of rhesus monkeys produced profound behavioral changes, including psychic blindness (the tendency to approach animate or inanimate objects without hesitation or fear), excessive oral tendencies, hypermetamorphosis (the tendency to react and attend to all visual stimuli), decreased fear and aggression, and hypersexuality. Weiskrantz demonstrated in 1956 that the behavioral abnormalities of the

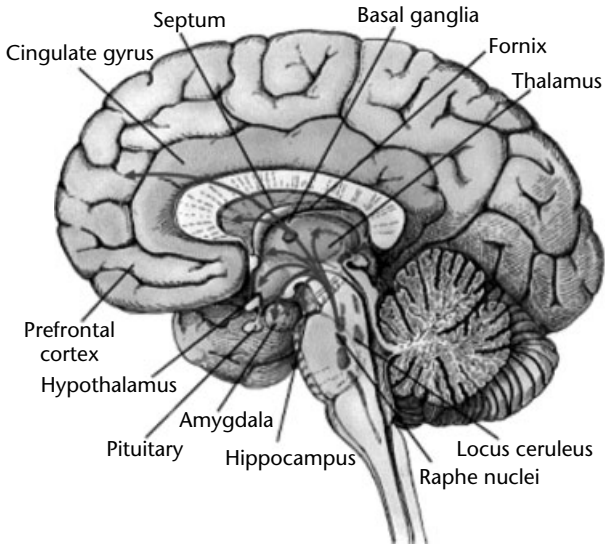


Figure 1. Brain structures that mediate emotion. In the brainstem, the locus ceruleus and raphe nucleus are the source of noradrenergic and serotonergic efferents to widespread regions of the cortex. These diffuse transmitter systems modulate neural activity related to emotion and other functions. The thalamus participates in evaluating the emotional significance of stimuli and organizing emotional behavior. The hypothalamus controls visceral functions. The amygdala and hippocampus mediate emotional conditioning and memory for emotional stimuli and their contextual surrounds, respectively. The basal ganglia facilitate the transition from motivation to action. A paralimbic structure, the cingulate gyrus, regulates attention and may contribute to the conscious experience of emotion. The prefrontal cortex coordinates functions throughout the emotion network, integrates information from the external world, and plays a major part in emotion regulation.

Klüver–Bucy syndrome could be produced by lesions of the amygdala alone.

More recent animal work has delineated neural circuits for specific emotions. From this work, a consensus has developed that the amygdala is essential to emotion. In a programmatic line of research, LeDoux has mapped out the circuitry underlying fear. His work on auditory fear conditioning has demonstrated that there are separate subcortical and cortical pathways that contribute to the fear response. The subcortical pathway, or 'low road', enables quick responses to potentially dangerous stimuli, by the transmission of information from the sensory thalamus directly to the amygdala. The cortical pathway, or 'high road', responds more slowly, since information is passed from the sensory thalamus to the cortex and then to the amygdala. However, it more accurately represents the nature of the stimulus. Rolls has argued that LeDoux's 'low road' probably does not apply to most complex stimuli, but instead applies only to simple stimuli, such as tones – or perhaps evolutionarily prepared stimuli, such as sight of a snake. More complex stimuli require cortical processing before information about them is transferred to the orbitofrontal cortex and amygdala for the assignment of reinforcement value. The orbitofrontal cortex modulates amygdala activity and can suppress or override stimulus–reinforcement associations established in the amygdala. (*See* **Amygdala**)

Lesions of the amygdala prevent fear conditioning, which suggests that the amygdala is a primary structure involved in emotional function. Current concepts of the amygdala view it as a threat detector. Whalen theorizes that the amygdala monitors the environment for stimuli that signal an increased probability of threat. Although they indicate the presence of threat in the environment, fearful cues may be ambiguous in terms of the source of the threat. As a result, ambiguous stimuli that convey fear encourage processing of the contexts in which they are present, because they derive their meaning from their immediate surroundings. Lesions to the amygdala interfere with the ability to learn about both specific cues and contextual stimuli. In this vein, Whalen suggests that the amygdala facilitates activity in brain regions that encode contextual stimuli, such as the hippocampus. Indeed, hippocampal damage prevents the conditioning or association between contextual cues present with discrete stimuli and emotional states.

The ventral striatum, particularly the nucleus accumbens, has a primary role in motivational circuitry and reward-related behavior. The nucleus accumbens integrates inputs from the prefrontal cortex, amygdala, and the ventral tegmental area (which provides dopamine input). Dopaminergic input to the nucleus accumbens may play an important part in the reinforcing effects of substances of abuse.

Studies of Emotion in Humans

Studies in humans, particularly of people with brain damage, point towards a basis of emotional function similar to that in nonhuman animals. The most famous patient in the study of emotion, Phineas Gage, sustained damage to the frontal lobe, including the ventromedial prefrontal cortex, when a tamping iron pierced his skull in 1848. He changed from being a moral, upstanding man and a hard worker to spouting profanity, showing little regard for social conventions, and behaving

irresponsibly. In general, damage to the lower middle portion of the frontal lobe (as in Phineas Gage) is associated with impulsive, disinhibited behavior, consistent with its modulatory influence on subcortical structures, including the amygdala. In contrast, damage to the upper portion of the frontal lobe is associated with amotivational behavior (i.e. failure to initiate behavior) consistent with the likely role of the dorsomedial frontal cortex in planning and executing goal-directed behavior.

Damasio has theorized that feedback from the body contributes to emotional experience. People with ventromedial frontal lobe damage lack physiological signs of emotional response, such as the skin conductance response (a measure of arousal), but they do have a conscious body state representation. As demonstrated by Bechara and colleagues, these patients also exhibit poor decision-making in gambling tasks in which risk-taking and reward assessment is involved. In particular, people without brain damage show electrodermal responses to risky choices when participating in the gambling task, prior to conscious awareness of how the game works, whereas people with ventromedial frontal lobe damage do not show such physiological responses and never learn the strategy of the gambling task. There is evidence that other brain regions, including the amygdala, and somatosensory cortices (S1, S2, and insula) are part of this neural system underlying the generation of somatic feedback. The insula has a particularly important role in detecting visceral sensations. Consequently, Damasio suggests that somatic feedback or 'markers' provide important information about the current state of the body to the brain, and influence subsequent decision-making.

Patients with bilateral amygdala damage but intact hippocampi show deficits in recognizing certain facial expressions of emotion, particularly fear and anger, although facial identification is unaffected. Patients with hippocampal damage but intact amygdalas display unconscious emotional conditioning, but have no memory for the conditioning events.

Studies of people with epilepsy have provided insight into the neural basis of emotion, particularly negative emotions, which possibly are related to the adaptive significance and greater elaboration of varieties of negative emotions. Psychomotor epilepsy is often rooted in abnormal electrical activity in the temporal lobe. Prior to seizure activity patients report an 'aura', which is characterized by negative emotions such as fear, anger, and dejection, as well as by positive emotions such as affection and ecstasy. During psychomotor epileptic seizures, which often begin in the amygdala, patients report experiencing fearful feelings.

Functional imaging studies are beginning to reveal the neural substrates of emotional disorders. Activity in paralimbic structures, which include the anterior cingulate, orbitofrontal cortex, insula, and temporal pole, appears to differ in people with depression and certain anxiety disorders compared with controls. These brain structures are intermediate between limbic structures and neocortex in their phylogenetic origin, and are involved in coordination of memory and higher-level emotional functions. Mayberg, for example, proposes that major depression is associated with dysfunction of specific subregions of the anterior cingulate cortex, and that recovery from depression is associated with a reversal of some of these effects. A similar pattern of dysfunction has been observed in individuals with post-traumatic stress disorder (PTSD). In response to trauma-related stimuli, the amygdala and certain paralimbic structures such as the orbitofrontal cortex are more activated in the brains of individuals with PTSD. Owing to learning related to the trauma, individuals with PTSD may have a lower threshold in the detection of fearful stimuli by the amygdala. Importantly, the anterior cingulate was found to not be activated in individuals with PTSD in response to trauma-related stimuli. This may be related to the numbing of emotional experience that many PTSD patients report. (*See* **Anxiety Disorders**; **Affective Disorders: Depression and Mania**)

Thus, these regions subserve a variety of functions, including the integration of cognitive and emotional functions via infusion of cognitive functions with emotional significance. This could help to explain, for example, how cognitive functions may be altered in the context of depression or anxiety. Notably, there is overlap in the neural circuitry implicated in emotional disorders and in substance abuse. Breiter and colleagues demonstrated in people addicted to cocaine that this drug activates a number of limbic and paralimbic structures, including the nucleus accumbens, caudate, putamen, insula and cingulate, and deactivates the amygdala and temporal pole.

Emotion Measurement in Humans

A variety of methods are used to measure emotional function in humans. Subjective methods include self-report (by questionnaire, structured or open-ended interview, response to emotion induction or other experimental tasks) and behavioral observation. Psychophysiological data indirectly

tap into neural function through peripheral indices such as muscle activity (electromyography) and autonomic nervous system activity, measured by electrodermal or skin conductance activity, heart rate, pupillary dilation, and the startle reflex (eyeblink). Psychophysiological measures are aligned more closely with dimensional models of emotion than with the discrete emotion theory because there is insufficient evidence for a specific profile of psychophysiological measures for particular emotions. For instance, activity of the corrugator muscle, which lies under the eyebrow and creates a furrowed brow, is associated with negative emotion, whereas activity of the zygomatic muscle, which lies under the cheek and raises the corners of the mouth during smiling, is associated with positive emotion. Arousal is positively correlated with both electrodermal activity and pupillary dilation, pleasantness is related to both heart rate increases and decreased startle response, and unpleasantness is related to pupillary dilation and an enhanced startle reflex.

Neuroimaging methods used in the study of emotional function include electroencephalography (EEG), single photon emission computed tomography (SPECT), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). In several PET and fMRI studies of induced emotion, both pleasant and unpleasant pictures activated the thalamus and medial prefrontal cortex more than neutral pictures did. Most studies point towards brainstem structures, the amygdala, hypothalamus, basal forebrain, ventromedial prefrontal cortex, cingulate cortex, and orbitofrontal cortex as active in neural circuits of emotion in a variety of emotional states. (*See* **Neuroimaging**)

On the basis of neuroimaging evidence, particularly that obtained with EEG, Davidson has theorized that there is hemispheric asymmetry or lateralization of emotions. People with greater left prefrontal cortical activation report greater positive and less negative affect and respond more strongly to positive compared with negative film clips than people who exhibit greater right frontal activation, who show the opposite pattern. Thus, left prefrontal activity is proposed to be related to approach or appetitive tendencies and right prefrontal activity is related to avoidance or aversive tendencies. Some studies, however, do not support this model. Issues to be resolved include the state versus trait nature of these hemispheric associations, and the time course and contexts in which this pattern is observed.

There are dissociations or loose coupling among the different components of emotion. Consequently, there is some question as to whether emotion represents a unified concept. Thus, one view holds that although data acquired using different methods is correlated, it should not be assumed that the findings all relate to a single causative influence or neural basis of emotion. From the same data, others conclude that there is a physiologic system mediating emotion, and this system was designed to have loosely coupled components that allow for optimal flexibility. For instance, it is adaptive in certain high-arousal situations to dampen one's subjective emotional responses. As more becomes known about the neural bases of emotion, these debates will move towards resolution.

CONSCIOUS AND UNCONSCIOUS EXPERIENCE OF EMOTION

It is now well established that people can display emotional behavior in the absence of concomitant emotional experience. Emotional responding can be elicited unconsciously, which allows for rapid responding in dangerous situations (fearful or angry stimuli both indicate danger), even before the stimulus is consciously perceived, conferring an evolutionary advantage. In his work on priming, Ohman used a backward masking procedure in which the fearful stimulus is presented for a very brief period, followed by a neutral stimulus of longer duration. Ohman demonstrated that fear-relevant stimuli are more resistant to extinction than fear-irrelevant stimuli, and people show autonomic responses to conditioned angry faces without conscious recognition of the stimuli. Concordantly, Whalen found bilateral amygdala activation and deactivation during unconscious processing of fearful and happy faces, respectively. In addition, the sublenticular substantia innominata, which reciprocally connects the amygdala and the hypothalamus, was activated during unconscious processing of both happy and fearful faces. Morris, Ohman, and Dolan extended this finding in a PET study examining neural activity during the conscious and unconscious processing of aversively conditioned angry faces. The main findings were that the right amygdala was activated during unconscious processing and the left amygdala was activated during conscious processing of the conditioned faces. These findings are consistent with the thesis that unconscious processing of emotional stimuli occurs primarily at the subcortical level, where more basic, evolutionarily old functions are executed.

These findings are to be contrasted with the activation of paralimbic structures observed in studies

of the conscious experience of emotion. LeDoux argues from animal research that emotional experience is represented in working memory. Human neuroimaging studies have demonstrated that the rostral anterior cingulate cortex and the medial prefrontal cortex participate in the representation of emotional experience, and lesions to these areas produce blunting of emotional experience. The conscious experience of emotion permits planning and flexibility of response, which would benefit long-term survival. It is likely that the structures needed for the conscious experience of emotion are not emotion-specific, which could explain the variability observed across individuals in the extent to which emotion is consciously experienced.

EMOTION AND MEMORY

Cahill and McGaugh have demonstrated in animals that the amygdala is involved in the formation of enhanced declarative memory for emotionally arousing events. The amygdala is not a site of long-term memory storage, but serves to influence memory storage processes in other brain regions, such as the hippocampus, striatum, and neocortex. Human studies show that administration of a drug that blocks the effects of noradrenaline (norepinephrine) decreases the enhancement of memory by emotional arousal. Imaging studies in humans demonstrate that activity in the amygdala while viewing emotional films correlates highly with memory of the films. Studies of emotion-dependent memory in healthy volunteers suggest that episodic memory, which reflects conscious awareness of one's personal experiences, appears to be organized along emotional lines. In other words, it is easier for one to recall emotionally significant memories if one is in a similar mood to the original experience.

INDIVIDUAL DIFFERENCES IN EMOTION

There is abundant research on variability across individuals in how they evaluate emotional stimuli, respond to emotional stimuli, and experience subjective feelings of emotion. From the perspective of the neural bases of emotion, pharmacological and neuroimaging work is quite informative.

For example, Depue and colleagues have demonstrated in humans that hormonal and affective response to a drug that stimulates the neurotransmitter dopamine through blockade of the reuptake transporter is associated with the personality trait of extraversion. Extraversion reflects the tendency to experience positive emotions based on sensitivity to rewarding stimuli, such as enjoying and valuing close interpersonal bonds, pursuing leadership roles, behaving assertively, and experiencing a subjective sense of potency in accomplishing goals. (*See* **Neurotransmitters)**

In an MRI study of personality influences on brain reactivity to emotional stimuli, Canli and colleagues found that extraversion was correlated with brain reactivity to positive picture stimuli in both cortical (frontal, temporal, including the cingulate gyrus) and subcortical (amygdala, caudate, putamen) regions, while neuroticism, or the tendency to experience anxiety, was correlated with brain reactivity to negative pictures in left frontal and temporal cortical regions. Extraversion was uncorrelated with response to negative pictures, relative to positive pictures, and neuroticism was uncorrelated with response to positive pictures, relative to negative pictures. Importantly, Canli and colleagues propose that studies of individual differences in the patterns of neural activity during emotional states provide a potential explanation for inconsistent findings across studies of emotion.

It is typically assumed that patterns of neural activity during emotional states are activated similarly across individuals, but evidence suggests that there is significant variability in the neural bases of emotion.

CONCLUSION

Much has been learned about the neural substrates of emotion from animal studies, observations in patients, and functional brain imaging studies in healthy volunteers and clinical populations. The fundamental constituents of the neural circuitry mediating emotion are being defined with increasing precision, and the neural basis of the interactions between emotion and cognitive functions is increasingly being understood. However, the precise mechanisms by which the neural networks mediating emotions work to orchestrate the range of normal and abnormal emotional responses in humans remain to be determined.

Further Reading

Damasio A (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Grosset/Putnam.

Damasio A (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt.

Davidson RJ and Irwin W (1999) The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences* **3**: 11–21.

Johnston VS (1999) *Why We Feel: The Science of Human Emotions*. Reading, MA: Perseus Books.

Lane RD, Nadel L, Ahern GL *et al*. (2000) *Cognitive Neuroscience of Emotion*. New York, NY: Oxford University Press.

Le Doux J (1996) *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon & Schuster.

Le Doux J (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* **23**: 155–184.

Lewis M and Haviland-Jones JM (2000) *Handbook of Emotions*. New York, NY: Guilford Press.

Panksepp J (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York, NY: Oxford University Press.

Rolls E (1999) *The Brain and Emotion*. New York, NY: Oxford University Press.

Encoding and Retrieval, Neural Basis of

Intermediate article

Lars Nyberg, Umeå University, Umeå, Sweden

CONTENTS

Cognitive theories describe encoding and retrieval as two distinct but interacting memory processes. This way of characterizing encoding and retrieval processes converges with analyses of their neural basis in that encoding and retrieval seem to engage specific as well as common brain areas.

INTRODUCTION

Encoding and retrieval are two fundamental memory processes. Encoding refers to the acquisition of information into memory, and retrieval to the use of previously encoded information. Retrieval can be *implicit* in the sense that previously encoded information affects subsequent behavior even though no conscious attempt is made to retrieve that information. Here, focus is on *episodic* memory retrieval as measured by tasks such as recall and recognition. Such tasks require *explicit* retrieval of previously encoded information in the sense that one has to think back to a previous study episode and retrieve information that was acquired at that particular time and place. Episodic information can be verbal or nonverbal; it can involve sensations of smell, taste, and touch; and it can represent emotional states.

THEORIES OF ENCODING AND RETRIEVAL AND THEIR INTERACTION

The division of the learning/memory process into encoding and retrieval (and an intermediate storage stage) became a major issue during the 1960s. Studies by Tulving and colleagues set the stage for subsequent empirical and theoretical work on retrieval processes. In a very influential paper (Tulving and Pearlstone, 1966), it was demonstrated that if encoding (and storage) conditions are held constant, the amount of information retrieved depends on the retrieval conditions.

During the 1970s, a considerable amount of research was focused on encoding. Much of this work was inspired by the *levels-of-processing* framework (Craik and Lockhart, 1972). In this framework, encoding was described in terms of various processing levels, and deeper processing was suggested to lead to better retention than shallow processing. It became clear that intention *per se* is not crucial for effective encoding but rather the way information is processed.

Another important topic of research during the 1970s concerned the interplay between encoding and retrieval processes. One set of findings gave rise to the *encoding specificity hypothesis* (Thomson and Tulving, 1970), by showing that retrieval cues are effective to the extent that they overlap with the encoded information. Another set of findings qualified predictions from the levels-of-processing framework by showing that deeper processing at encoding does not always lead to superior retention. Rather, what is optimal processing at encoding is dependent on the retrieval conditions. These findings formed the basis for the *transfer appropriate processing* principle (Morris *et al*., 1977). The importance of interplay between encoding and retrieval processes is also salient in other theoretical accounts, such as Kolers' procedural viewpoint (Kolers, 1973).

Encoding–retrieval interplay has a central role in contemporary accounts of dissociations between measures of retrieval, most notably dissociations between explicit and implicit retrieval. Based on a classification of retrieval measures as 'conceptually-driven' (dependent on the encoded meaning of information) or 'data-driven' (dependent on the perceptual match between encoding and retrieval) (Jacoby, 1983), Roediger and colleagues have put forward a transfer-appropriate account of such dissociations (Roediger *et al*., 1989).

CAN ENCODING AND RETRIEVAL BE DISTURBED SEPARATELY? LESION AND PHARMACOLOGICAL EVIDENCE

A fundamental question in the study of the neural basis of encoding and retrieval is whether encoding and retrieval can be selectively affected by brain damage. Here it should be noted that it is difficult to isolate the effects of brain damage to encoding or retrieval processes. If a lesion is made/has occurred prior to encoding and an effect is noted on subsequent retrieval performance, it is not clear whether the lesion affected encoding or retrieval processes. If a lesion occurs after encoding and prior to retrieval, it is still difficult to conclude that it affected retrieval processes. This is because there is evidence that consolidation of information in memory goes on for some time after the initial encoding. Moreover, it is possible that the lesion affected sites for memory storage rather than sites involved in the retrieval of stored information. With these caveats in mind, some suggestions regarding process-specific effects of brain damage will be considered.

Organic amnesia is characterized by dysfunctional episodic memory. This syndrome can result from lesions in the medial temporal lobes, diencephalon, or the basal forebrain. Collectively, these sites have been referred to as the 'expanded' limbic system (Markowitsch, 2000). This system, or various combinations of its constituent parts, has been associated with the transfer of incoming registered information for long-term memory storage. It has been proposed that the limbic structures can be regarded as 'bottleneck structures' in the sense that bilateral damage to any of them will lead to anterograde amnesia (Markowitsch, 2000).

Limbic regions may also play a role in retrieval. Several studies indicate that damage to limbic structures impairs retrieval of recent episodes, but it is a matter of debate whether retrieval of remote events is also affected. Moreover, specific regions in the infero-lateral prefrontal cortex and in the temporo-polar cortex (especially in the right hemisphere) seem to be critical for retrieval. Lesions to this regional combination result in retrograde amnesia (Markowitsch, 2000).

In a recent study (Rossi *et al*., 2001), repetitive transcranial magnetic stimulation was used to transiently interfere with prefrontal brain activity during encoding and retrieval of pictures. It was found that the left dorsolateral prefrontal cortex was crucial for encoding, whereas the right dorsolateral prefrontal cortex was crucial for retrieval. This study provides strong evidence that encoding and retrieval can be disturbed separately, and highlights the role of prefrontal regions in this regard (cf. the section below on differences in functional brain activity).

Psychopharmacological studies also provide some hints that encoding and retrieval processes can be disturbed separately. A much-studied class of drugs is benzodiazepines. These drugs facilitate the transmission of gamma-aminobutyric acid (GABA), which is the major inhibitory neurotransmitter in the brain. Administration of benzodiazepine drugs before the encoding of new information leads to poor retention. By contrast, if the drug is administered at the test phase, performance is not affected (Curran, 2000). A similar effect has been observed for scopolamine, which acts as a cholinergic blocker. The observed pattern of effects suggests that these drugs impair encoding processes rather than retrieval processes, possibly by affecting the ability to form item–item or item–context associations (Curran, 2000).

ARE THERE SEPARATE NEURAL SYSTEMS FOR ENCODING AND RETRIEVAL?

The results from lesion and pharmacological studies indicate that brain damage or drug administration can have selective effects on encoding and retrieval processes. These patterns of results provide tentative support for the existence of separate neural systems for encoding and retrieval. Additional evidence comes from functional neuroimaging studies. In these studies the neural correlates of encoding and retrieval can be studied separately, and hence it can be explored whether there are distinct encoding and retrieval systems in the brain. Two main classes of functional neuroimaging techniques are hemodynamic methods and electromagnetic methods. This discussion is limited to hemodynamic registrations by positron emission tomography (PET) and functional magnetic resonance imaging (fMRI).

FUNCTIONAL NEUROIMAGING STUDIES

PET and fMRI rely on the changes in cerebral blood flow that accompany neural activity. To identify changes in brain activity that are associated with a process of interest, it is common to measure activity in at least two conditions (experimental and control condition). These conditions are carefully matched in all respects, except for the process of interest. By subtracting out brain activity associated with the

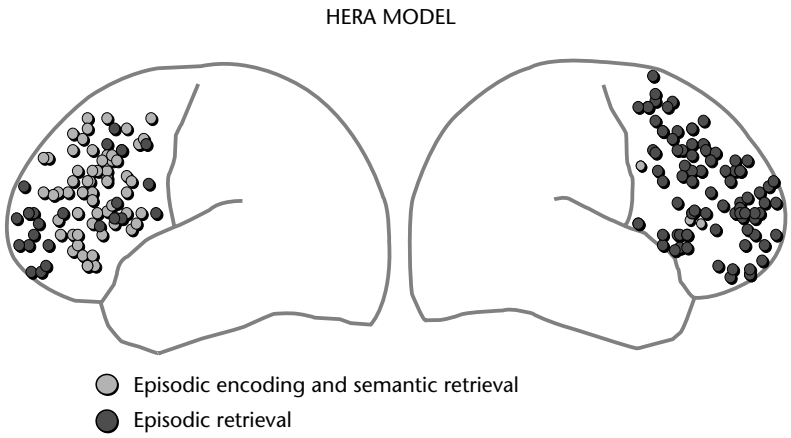


Figure 1. Differential activation of prefrontal cortex during encoding and retrieval of episodic information. Prefrontal activations from published PET and fMRI studies up to December 1999 are plotted on lateral brain outlines. Courtesy of Roberto Cabeza.

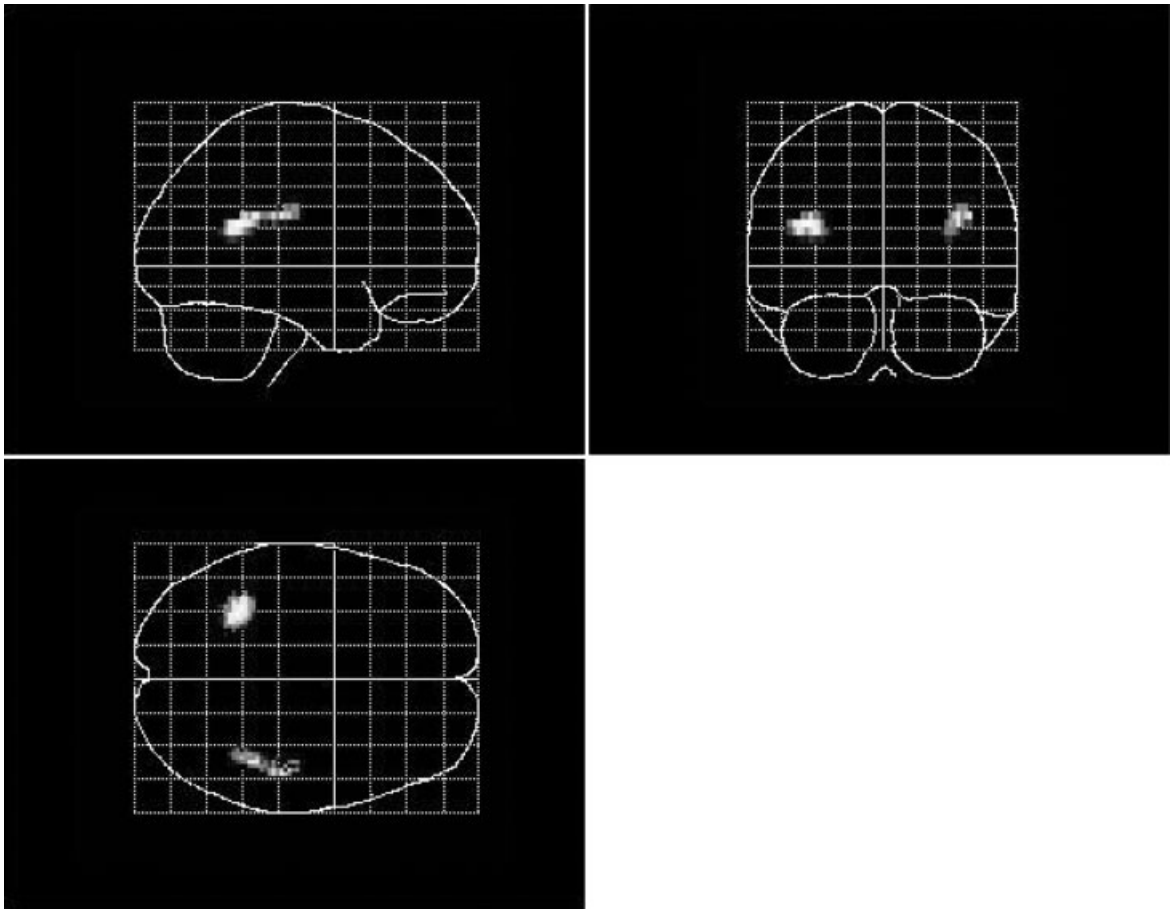


Figure 2. [*Figure is also reproduced in color section.*] Overlap in parietal activation during encoding and retrieval of spatial information. Activations are outlined in sagittal (top left), coronal (top right), and horizontal (bottom left) transparent brain maps. Reprinted with permission from 'Conjunction analysis of cortical activations common to encoding and retrieval' (Persson and Nyberg, 2000).

control condition from that associated with the experimental condition, it is possible to isolate brain regions associated with the process of interest. In what follows, a summary of results from PET and fMRI studies of encoding and retrieval is given (Cabeza and Nyberg, 2000).

Encoding

Episodic encoding is associated with the left prefrontal cortex and medial-temporal lobe regions. Left prefrontal activation has been observed for intentional encoding and also for incidental encoding (e.g. comparisons of deep and shallow semantic tasks). Medial-temporal activation during encoding has been related to novelty detection. There is some evidence that medial-temporal regions interact with material-specific regions during encoding (e.g. occipito-temporal regions during picture encoding), and that the laterality of encoding-related activity is modulated by material.

Retrieval

Retrieval is strongly associated with right prefrontal and medial parietal activation. Increased activity has also frequently been observed in the medial-temporal cortex and in the cerebellum. Activity in some regions is higher during successful than unsuccessful retrieval, whereas other regions' activity tends to be unaffected by level of retrieval or increase during more demanding (and less successful) retrieval conditions.

Differences and Similarities

As indicated above, encoding and retrieval have been associated with different regions of the frontal cortex, with encoding being associated with left prefrontal regions and retrieval with right prefrontal regions. This asymmetric involvement of prefrontal regions during encoding and retrieval is captured by the HERA (Hemispheric Encoding/Retrieval Asymmetry) model shown in Figure 1 (Nyberg *et al*., 1996).

In addition to encoding–retrieval differences, several studies have reported overlap in activation patterns for encoding and retrieval (Nyberg, 2002). One example is that encoding and retrieval of spatial locations activate overlapping regions in the 'where-pathway' as shown in Figure 2 (Persson and Nyberg, 2000). Such overlap suggests that formation and recovery of memory representations engage the same brain regions. This possibility is in line with the view that memories are represented in a distributed fashion in/near regions that are involved during initial perception/encoding.

CONCLUSION

Encoding and retrieval are two distinct, but highly interactive, memory processes. The neural bases of these processes are only beginning to be understood. At this stage, there is converging evidence that encoding and retrieval engage specialized neural systems. There is also evidence that specific association areas are engaged during encoding of specific information as well as during subsequent retrieval of the same information. Areas where encoding and retrieval processes meet in the brain may represent storage sites.

References

Cabeza R and Nyberg L (2000) Imaging cognition II: an empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience* **12**: 1–47.

Craik FIM and Lockhart RS (1972) Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–684.

Curran HV (2000) Psychopharmacological perspectives on memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 539–554. New York, NY: Oxford University Press.

Jacoby LL (1983) Remembering the data: analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior* **22**: 485–508.

Kolers PA (1973) Remembering operations. *Memory & Cognition* **1**: 347–355.

Markowitsch HJ (2000) Neuroanatomy of memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 465–484. New York, NY: Oxford University Press.

Morris CD, Bransford JD and Franks JJ (1977) Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* **16**: 519–533.

Nyberg L (2002) Where encoding and retrieval meet in the brain. In: Squire LR and Schacter DL (eds) *Neuropsychology of Memory*, pp. 193–203. New York, NY: Guilford Press.

Nyberg L, Cabeza R and Tulving E (1996) PET studies of encoding and retrieval: the HERA model. *Psychonomic Bulletin & Review* **3**: 135–148.

Persson J and Nyberg L (2000) Conjunction analysis of cortical activations common to encoding and retrieval. *Microscopy Research Techniques* **51**: 39–44.

Roediger HL III, Weldon MS and Challis BH (1989) Explaining dissociations between implicit and explicit measures of retention: a processing account. In: Roediger HL III and Craik FIM (eds) *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*, pp. 3–41. Hillsdale, NJ: Erlbaum.

Rossi S, Cappa SF, Babiloni C *et al*. (2001) Prefrontal cortex in long-term memory: an 'interference' approach using magnetic stimulation. *Nature Neuroscience* **4**: 948–952.

Thomson DM and Tulving E (1970) Associative encoding and retrieval: weak and strong cues. *Journal of Experimental Psychology* **86**: 255–262.

Tulving E and Pearlstone Z (1966) Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior* **5**: 381–391.

Further Reading

Dolan RJ and Fletcher PC (1997) Dissociating prefrontal and hippocampal function in episodic memory encoding. *Nature* **388**: 582–585.

Kelley WM, Miezin FM, McDermott KB *et al*. (1998) Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. *Neuron* **20**: 927–936.

Lepage M, Ghaffar O, Nyberg L and Tulving E (2000) Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Sciences USA* **97**: 506–511.

Mayes AR (1995) Memory and amnesia. *Behavioural Brain Research* **66**: 29–36.

Nadel L and Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology* **7**: 217–227.

Nyberg L, McIntosh AR, Houle S, Nilsson L-G and Tulving E (1996) Activation of medial temporal structures during episodic memory retrieval. *Nature* **380**: 715–717.

Rugg MD and Allan K (2000) Event-related potential studies of memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 521–537. New York, NY: Oxford University Press.

Schacter DL and Wagner AD (1999) Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus* **9**: 7–24.

Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**: 195–231.

Wagner AD, Schacter DL, Rotte M *et al*. (1998) Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281**: 1188–1191.

Epilepsy

Introductory article

W McIntyre Burnham, University of Toronto, Toronto, Ontario, Canada

CONTENTS

The term epilepsy refers to a group of neurological disorders characterized by spontaneous, recurrent seizures. In most cases drug therapy is effective, but intractable epilepsy can have a profound effect on quality of life.

INTRODUCTION

The epilepsies are a group of neurological disorders characterized by spontaneous, recurrent seizures. About 1% of the population has epilepsy at any given time, and about 4% of the population will have epilepsy during their lifetime. Another word for epilepsy is 'seizure disorder'.

Seizures themselves are periods of self-sustained neural hyperexcitation. During a seizure, the neurons of the brain cease their normal activities and begin to fire in massive, synchronized bursts. Blood flow to the brain is increased, an6d there is a greater use of glucose and oxygen. After seconds or minutes, when the inhibitory mechanisms of the brain regain control, the seizure ends. During a seizure the hyperactivity in the brain can be seen as a series of spikes, or spikes and waves, in electroencephalographic (EEG) recordings: these are called the 'electrographic' seizure. The behavior of the patient during the epileptic attack – which may or may not be convulsive – is called the 'clinical' seizure. If the clinical seizure involves muscle spasms, it is called a 'convulsion'. It should be noted that many clinical seizures are nonconvulsive. Other terms for seizure are 'attack' and 'ictus'. The terms 'preictal', 'ictal' and 'postictal' refer to events that occur before, during and after a seizure; 'interictal' refers to events that happen between seizures.

To be classified as epilepsy, seizures must be spontaneous, meaning that there is no outside cause for the attack. Seizures caused by fevers in children (febrile seizures), by convulsant drugs or by metabolic imbalances are not considered to be epilepsy. Seizures must also be recurrent. A single seizure – estimated to occur in about 10% of the population – is not considered to be epilepsy: two or more seizures must occur before epilepsy is diagnosed.

It is important to note that seizures do not equal epilepsy. Every brain has a 'seizure threshold', and every brain will generate a seizure if it is subjected to a high level of excitatory stimulation. In most people the seizure threshold is high, and seizures do not occur spontaneously. People with epilepsy are distinguished by the fact that their brain – or a part of their brain – has a seizure threshold so low that from time to time the brain's own activity triggers an attack. This low seizure threshold is always present, and stimuli that would not affect an ordinary person, such as hyperventilation or flashing lights, will sometimes trigger a seizure in a person with epilepsy.

ONSET OF EPILEPSY

The onset of epilepsy can occur at any time during life. In many patients seizures begin in childhood, often before the age of 15 years. If seizures begin later, during adulthood, they may indicate the presence of an expanding tumor in the brain. In recent years – perhaps because people are living longer – there has been an increased onset of seizures in elderly people. Many of these are thought to be the result of small strokes.

FREQUENCY OF SEIZURES

The frequency of seizures varies enormously. Some patients may experience only a few seizures during the whole course of their life. At the other end of the spectrum are patients who experience many seizures every day.

EFFECT ON LIFE

In general, seizures do not constitute a major medical emergency. They are not painful, and except for the condition known as 'status epilepticus' (see below), they do not harm the individual. Admittedly, there are risks. People with seizures sometimes injure themselves as they fall, and if seizures are frequent they are advised not to swim alone, to shower rather than to bathe, and to avoid high places and the edges of subway platforms. In general, however, the risks of seizures are manageable.

Seizures often have social and economic effects on life, however, that go far beyond their medical importance. Seizures may be frightening to watch, and people who have public seizures face both social and economic discrimination. The impact that seizures have on life thus relates to whether public seizures can be controlled. If seizures are responsive to medication, the patient may never have another attack. Such patients have to take medication regularly, but if they do so, epilepsy will have little effect on their life. If seizures are only partly responsive to medications, however, the patient will live a fairly normal life, but will continue to have some attacks. If they occur in public, the patient will probably have some social and economic problems. In most countries, drivers' licences are canceled with the first seizure. The licence is usually returned if the patient is seizure-free for a certain period (often a year). A patient who has even one seizure a year, however, may never be able to legally drive again, which may severely limit the possibilities of employment. If seizures are drug-resistant, they will continue to occur despite the best drug therapy. Children with drug-resistant seizures often have serious problems at home and at school, while adults are frequently unemployed or underemployed. Serious emotional problems may develop, and there is a risk of suicide. These patients must turn to nonpharmacological therapies, such as surgery, the ketogenic diet and vagal stimulation (see below).

CLASSES OF EPILEPTIC SEIZURE

There are a number of different types of epileptic seizure, some of which occur primarily in childhood. Table 1 sets out the international classification of epileptic seizures, formulated in 1981 by the International League Against Epilepsy. The following discussion is limited to four types of seizure commonly seen in adults.

Table 1. International classification of epileptic seizures, formulated in 1981 by the International League Against Epilepsy

- I *Partial (focal, local) seizures*
  - A Simple partial seizures (consciousness unimpaired)
    - With motor signs
    - With somatosensory or special sensory symptoms
    - With autonomic symptoms or signs
    - With psychic symptoms
  - B Complex partial seizures (consciousness impaired)
    - Simple partial onset followed by impairment of consciousness
    - With impairment of consciousness at onset
  - C Partial seizures evolving to secondarily generalized seizures
    - Simple partial seizures evolving to generalized seizures
    - Complex partial seizures evolving to generalized seizures
    - Simple partial seizures evolving to complex partial seizures, evolving to generalized seizures
- II *Generalized seizures (convulsive or nonconvulsive)*
  - A Absence seizures
    - 1 Typical absences
    - 2 Atypical absences
  - B Myoclonic seizures
  - C Clonic seizures
  - D Tonic seizures
  - E Tonic–clonic seizures
  - F Atonic seizures (astatic seizures)
- III *Unclassified epileptic seizures*

Modified from: Proposal for revised clinical and electroencephalographic classification of epileptic seizures. *Epilepsia* 1981; **22**: 489–501.

Generalized Seizures

In generalized seizures the whole brain is involved in epileptic activity (EEG 'spiking') during the attack, and consciousness is usually lost.

Absence seizure

Absence seizures (formerly called 'petit mal' seizures) are nonconvulsive. The clinical seizure consists of a few seconds of blank staring and immobility, while the electrographic seizure consists of 'three per second spike and wave' activity in the EEG. Consciousness is lost, and the person has no memory of the attack. After the attack the person resumes whatever he or she was doing, and may be unaware that an attack has taken place. These mild seizures usually start in childhood, and may be outgrown by the teenage years. While each seizure is mild, absence seizures can occur many times each day, and children may have trouble following what is going on around them.

Tonic–clonic seizure

The second common type of generalized seizure is the tonic–clonic seizure (formerly called 'grand mal' seizures). These are the dramatic convulsive seizures that people think of when they hear the term 'epilepsy'. The clinical seizure involves a convulsion which consists of stiffening of the whole body (tonus) and then jerking of the whole body (clonus). The electrographic seizure consists of constant spiking in the EEG. Consciousness is lost, and the patient will have no memory for the period of the attack. After the attack – which lasts a few minutes – the person is comatose for a short time (postictal coma) and then may be conscious but confused. After that, the person returns to normal, perhaps being able to go back to work, or wanting to lie down or sleep.

Partial Seizures

In addition to the generalized epilepsies, there are two types of 'partial' epilepsy (formerly called 'focal' epilepsies). The term 'partial' indicates that, at least at the onset, only part of the brain is involved in epileptic activity (EEG spiking). The person is usually conscious during a partial seizure, although consciousness may be impaired. Partial seizures are often nonconvulsive.

Simple partial seizure

During simple partial seizures (formerly called 'cortical focal' seizures), epileptic spiking is limited to one part of the brain. The symptoms of the seizure depend on the part of the brain activated by the epileptic discharge. They are very variable: for instance, they may involve an experience of flashing lights, a buzzing sound, a bad smell or taste, or a feeling of fear. If a motor area of the cortex is involved, convulsive jerking (clonus) will be seen on the opposite side of the body. The person is conscious throughout the seizure, and will remember it afterwards.

Complex partial seizure

In complex partial seizures (formerly called 'temporal lobe' or 'psychomotor' seizures), epileptic spiking is often seen on both sides of the person's brain in the temporal lobe areas. During such seizures the person does not convulse, and does not lose consciousness – however, consciousness is said to be impaired. The person is nonresponsive and seems to be out of contact with the environment. Repetitive meaningless movements may occur, such as lip smacking or fumbling with the clothes (automatisms). Although not unconscious, the person will have no memory of the attack, possibly because the memory mechanisms in the both temporal lobes are involved. Complex partial seizures are the most common type of seizure in adults.

Auras and Prodromes

A simple partial seizure may spread through the brain (generalize) to become a complex partial attack or a tonic–clonic attack. The simple partial seizure – which is remembered – is then termed an 'aura' or warning. It is important to distinguish auras – which are simple partial seizures – from prodromes. Prodromes are signs, recognized by the person or the person's family, that a seizure is going to occur later in the day. Like auras, prodromes precede seizures, but, unlike auras, they do not involve seizure activity.

Status Epilepticus

Very long seizures, or seizures that repeat without the patient regaining consciousness, are termed 'status epilepticus'. Status can consist of a prolonged period of absence seizures, partial seizures or tonic–clonic seizures. Nonconvulsive status is not considered a major medical emergency. Convulsive status, however (i.e. tonic–clonic status), may lead to brain damage or death if it is not terminated. Epilepsy associations often suggest that families should call for an ambulance if a tonic–clonic seizure has not stopped within 5 min.

EPILEPTIC SYNDROMES

An attempt has been made to fit epileptic seizures into larger classifications or epileptic syndromes. An epileptic syndrome includes not only a seizure type (or types), but also a prediction about the time of seizure onset, a possible cause, a prognosis for control, and, in some cases, a likely response to a particular medication. One example is juvenile myoclonic epilepsy: in this syndrome the seizures consist of isolated muscle jerks (myoclonus) and tonic–clonic attacks. Onset is between ages 8 years and 18 years. The young person is perfectly normal except for the seizures, which often occur in the morning, just after waking. There is often a history of juvenile myoclonic epilepsy in the family, and the cause is believed to be genetic. There is a good response to anticonvulsant drugs, especially valproate. Other well-known epileptic syndromes include West syndrome, which has its onset in the first year of life, and Lennox–Gastaut syndrome,

which usually begins between the ages of 1 year and 8 years. These serious epileptic syndromes involve intractable seizures, an abnormal interictal EEG and in most cases mental retardation. However, many seizures do not fit into any recognized syndrome.

CAUSES OF EPILEPSY

In about 30–40% of people with epilepsy, there is a clear-cut structural abnormality in the brain. These abnormalities include scars, tumors, infections, areas of improperly developed cortex (cortical dysplasia) and blood vessel (arteriovenous) malformations. Cases of epilepsy where there is a clear abnormality in the brain are described as 'symptomatic'. They are often hard to control with anticonvulsant medications. Table 2 indicates the frequency of different causes of epilepsy found in a large American study.

In about 60–70% of people with epilepsy, the brain appears to be entirely normal. In these people the abnormality causing the seizure is assumed to be an ionic or biochemical imbalance. Cases of epilepsy in which the brain appears to be entirely normal are described as 'idiopathic', and their cause is generally thought to be genetic. The term 'cryptogenic' is sometimes applied to cases of epilepsy where no clear brain abnormality can be found, but where there are symptoms that resemble those of the symptomatic epilepsies. It is assumed in these cases that an abnormality exists, but that it cannot be found.

Genetic Factors

There is a genetic factor in most cases of epilepsy. Both the partial and the generalized epilepsies show significant hereditability. In most cases of epilepsy, however, the genetic contribution is not a simple one.

The simplest mode of inheritance is inheritance involving a single gene. This is often called Mendelian inheritance, and the rules that govern dominant and recessive Mendelian inheritance are well understood. In other cases, inheritance is dependent on a number of genes, and is described as 'multifactorial'. The rules governing multifactorial inheritance are much less clear. In most cases of epilepsy inheritance is multifactorial. The disorder tends to 'run in families', but does not follow simple Mendelian rules. The contribution of multifactorial genetics is clearest in the idiopathic epilepsies, where there is no structural abnormality in the brain. These are generally considered to be genetic disorders. What is inherited may be a collection of mild abnormalities, related to several different genes. None is sufficient to cause epilepsy by itself, but in combination, they contribute to a chronic low threshold.

There is a genetic predisposition even in cases of symptomatic epilepsy, where the seizures are related to a structural abnormality, such as a scar on the brain. Some people with such scars develop epilepsy, whereas others with similar scars do not. Again, the genetic predisposition seems to be multifactorial.

People with seizures often wonder whether their seizures will be passed along to their children. With multifactorial epilepsies, the chance of transmission is about 6% overall. It is about 9–12% if the epilepsy is idiopathic.

Table 2. The frequency of causes of epilepsy in a large American study

Cause	*Frequency (%)*
No clear cause (idiopathic epilepsy)	68.7
Known cause (symptomatic epilepsy)	31.3
Cerebrovascular disease	13.2
Developmental	5.5
Brain trauma	4.1
Brain tumor	3.6
Cerebral infection	2.6
Degenerative disease	1.8
Other	0.5

Based on Annegers JF (1994) Epidemiology and genetics of epilepsy. *Neurologic Clinics* **12**: 15–29.

The Genetic Epilepsies

In some epilepsies genetic factors have a much greater role. These are the epileptic syndromes that relate to single gene mutations, the inheritance of which follows Mendelian rules. More than 140 genetic epileptic syndromes are now recognized; most of them are rare.

In some cases a single gene abnormality gives rise to a subtle change in the brain, which is otherwise normal. Some of these subtle changes have been found to be abnormalities in ion channels, which has given rise to the concept of the genetic epilepsies as 'channelopathies'. These epileptic syndromes are idiopathic: an example is juvenile myoclonic epilepsy. Within this class there are a number of benign syndromes. In a benign syndrome, seizures appear in childhood, but spontaneously disappear as the child grows older. Examples are benign familial neonatal convulsions

and benign rolandic epilepsy. In other cases of inherited epilepsy, the genetic abnormality gives rise to a major malformation of the brain. These epilepsies are described as 'symptomatic'. An example is tuberous sclerosis: in this condition the brain is malformed, and the child frequently suffers from mental retardation and intractable seizures. Tuberous sclerosis is often a cause of West syndrome. In still other cases, the inherited defect is a metabolic abnormality. This makes the brain unable to metabolize some compound produced in neurons. The compound gradually builds up over the years, until it interferes with neural function. A child with this type of genetic abnormality will be normal at first, but later, as the compound begins to compromise brain function, will experience a progressive loss of mental function, intractable myoclonic seizures, and eventually death. These serious symptomatic syndromes are known as progressive myoclonus epilepsies and sometimes as 'storage diseases'. Lafora disease and Baltic myoclonus epilepsy are examples of progressive myoclonus epilepsies.

With epilepsies related to a single gene, the possibility of transmission to children is high. Families afflicted with genetic epilepsies should seek genetic counseling.

Finally, in some cases epilepsy is caused by genetic abnormalities that are not inherited but arise during fetal development. Examples are Rett syndrome, which involves retardation and seizures, and Down syndrome, in which seizures occur in about 15% of cases.

TREATMENT

Drug Therapy

The most common therapy for epilepsy is treatment with anticonvulsant drugs. It is almost always the first therapy to be tried. The term 'anticonvulsant' is something of a misnomer, since many seizures do not involve convulsions. Some writers, therefore, prefer the terms 'antiepileptic' or 'antiseizure' drug. 'Anticonvulsant', however, is still the term most widely used. A wide variety of anticonvulsant drugs are available. Among the most commonly prescribed are ethosuximide, used for absence seizures, and phenytoin and carbamazepine, used for tonic–clonic and partial seizures. Phenobarbital (phenobarbitone) is an older drug, sometimes used for tonic–clonic and partial seizures. Valproic acid and clonazepam are wide-spectrum drugs, effective against many different seizure types. Table 3 lists the most commonly prescribed anticonvulsant drugs, grouped according to use. In addition to the standard anticonvulsants, a number of new drugs have been introduced, including lamotrigine, gabapentin, topiramate, vigabatrin, levetiracetam, zonisamide and oxcarbazepine. These drugs are available in some countries, but not in others. While the new drugs generally have fewer side effects than the older drugs, it is not yet clear that they are more effective at stopping seizures. They are considerably more expensive than the older drugs.

Table 3. Commonly prescribed anticonvulsant drugs

Type of Seizure	*Drug*
Absence seizures	Ethosuximide
Tonic–clonic and partial seizures	Carbamazepine
	Gabapentin
	Phenobarbital (phenobarbitone)
	Phenytoin
	Primidone
Broad-spectrum drugs (absence, tonic–clonic and partial seizures)	Clonazepam
	Lamotrigine
	Sodium valproate, valproic acid
	Topiramate
Status epilepticus (administered intravenously in hospital)	Diazepam
	Lorazepam

Modified from Burnham WM (1998) Antiseizure drugs. In: Kalant H and Roschlau WHE (eds) *Principles of Medical Pharmacology*. New York: Oxford University Press.

Anticonvulsant therapy is begun cautiously. At the initial consultation of a patient experiencing seizures, the physician first searches for the cause of the attacks. If the seizures are the result of an active disease process such as an expanding tumor or a brain infection, the physician treats the disease rather than prescribing anticonvulsant drugs. If no active pathological condition is found, the physician prescribes anticonvulsant drug therapy. The first step is to establish which type of seizure is occurring. This is important, because different types of seizure require different drugs. Drugs used to treat tonic–clonic seizures may make absence seizures worse, and vice versa. It is particularly hard to distinguish between some of the nonconvulsive seizure types, such as absence and complex partial seizures. The distinction, however, is important, since they require different medications.

In current practice the medical management of seizures will begin with a single drug (monotherapy). If the first drug fails, the physician may try another single drug before going on to drug

combinations (polypharmacy). With most anticonvulsants, it is wise to start treatment with a low dose, and gradually build up to the full therapeutic dosage. Drug doses are commonly adjusted on the basis of 'blood levels'. After the patient has started taking the medication, a blood sample is taken to determine whether the dose administered has produced blood concentrations within the therapeutic range. If concentrations are not in the therapeutic range, the dose may be adjusted. It is important to note, however, that some patients achieve seizure control with blood levels below the therapeutic range, while other patients fail to show toxicity with blood levels above it. In these patients there is no need to adjust dosage to therapeutic levels: in the long term, patient response – not blood levels – should determine dosage.

The anticonvulsant medications do not cure epilepsy. They simply suppress seizures on a temporary basis. Patients must continue to take their medication, once, twice or three times daily, sometimes for the rest of their lives.

The success of drug therapy varies from patient to patient. When appropriate drugs and dosages are administered, therapy is completely successful in about 60% of patients. These patients may never have another seizure. In these cases of complete seizure control, after 2 years and if the EEG is normal, the patient may be offered the choice of gradually discontinuing medication. This is a serious decision, since the patient should not drive for 3–6 months after the drugs are discontinued. About half of these patients find that they have 'outgrown' their epilepsy, and that the drugs are no longer necessary.

Another 20% of patients are only partially responsive to drugs. These patients have fewer attacks while on medication. Unfortunately, about 20% of patients have seizures that are completely resistant to drugs – 'refractory' or 'intractable' seizures.

While any type of seizure can be intractable, intractability is most often seen in association with structural brain abnormalities, or when certain types of seizures are involved. The seizures associated with two rare childhood syndromes, West syndrome and Lennox–Gastaut syndrome, are frequently intractable, as are those associated with complex partial epilepsy. The intractability of complex partial seizures presents a major therapeutic problem, since these are the most common seizures in adults. If two appropriate single drugs have been tried and have failed to control the seizures, there is a good chance that the seizures will be intractable. The patient may then be given a combination of two or more drugs, but even polypharmacy may not stop the attacks. Other, nondrug therapies should then be considered.

Nondrug Therapies

If the patient suffers from intractable partial epilepsy, and attacks always start in the same spot in the brain (the 'focus'), the physician may suggest surgery. The most common surgical treatment for seizures involves removal of the site in the brain that initiates the attacks. If the patient is suffering from complex partial seizures of temporal lobe origin, the surgeon might remove the anterior part of one of the temporal lobes (temporal lobectomy). Surgery may stop the seizures altogether, or may make them more controllable with medication. Other surgical procedures are corpus callosotomy, where the surgeon cuts the connections between the left and right cortex, and multiple subpial resection, where the surgeon makes many small cuts in an area of epileptic cortex that is too important to be removed (e.g. the speech center).

Another therapy for intractable seizures is the ketogenic diet. The ketogenic diet is a high-fat diet containing adequate protein and limited carbohydrate. In its classic form, the diet consists of three or four parts of fat to one part of everything else that is eaten. Patients may find the diet unpalatable, and it is difficult to maintain, since all of the patient's food must be weighed and measured. The ketogenic diet, however, stops seizures in about a third of patients who have found drug therapy ineffective, and decreases seizures in another third. Owing to worries about the diet's nutritional effects, patients usually stay on the diet for only 2–3 years. Traditionally the diet has been used in children, although reports suggest that it is effective in adults as well. The diet's mechanism of action is unknown.

A third treatment for intractable seizures is vagal stimulation. A device similar to a cardiac pacemaker is permanently implanted in the patient's chest muscle, and used to stimulate the vagus nerve, which originates in the brain. For unknown reasons, stimulation of the vagus nerve suppresses seizures in some patients. Vagal nerve stimulation is not as effective as surgery or the ketogenic diet, but it is an alternative to consider when the other therapies fail.

Therapy for the Emotional Consequences of Epilepsy

Intractable epilepsy has major emotional consequences. These must be dealt with if the seizures cannot be stopped.

Seizures involve a frightening loss of control, which occurs at random, unpredictable intervals. People with intractable epilepsy, therefore, often suffer a major loss of self-confidence. This is compounded by social rejection, economic decline, and the side effects of the anticonvulsant drugs. It is estimated that about 50% of children with intractable epilepsy suffer from emotional disturbances – often anxiety or depression – severe enough to require therapy. The numbers are probably similar for adults with intractable epilepsy. The emotional problems related to seizures are responsive to therapeutic drugs, and in particular to antidepressants. Unfortunately, these problems are seldom recognized or treated. Therapy centers on seizure control, and emotional problems are often overlooked. Patients with emotional problems should talk to their doctors and ask for help.

CONTRIBUTIONS OF EPILEPSY RESEARCH TO NEUROSCIENCE

Much of our knowledge about the organization of the human neocortex originally came from the study of simple partial seizures. When epileptic discharge activates a part of the brain, the seizure that occurs reflects the function of that brain area. When the visual cortex is involved, for instance, the patient sees flashing lights; when the auditory cortex is involved, the patient hears a buzzing sound. In the late nineteenth century, the British neurologist Hughlings Jackson worked out the organization of the human somatosensory and motor areas on the basis of simple partial seizures. Much of what he deduced was later confirmed by Penfield and Jasper, working at the Montreal Neurological Institute, and using electrical impulses to stimulate the exposed brain of waking patients during seizure surgery.

CONCLUSION

Epilepsy is a common neurological disorder, usually treated with anticonvulsant drugs. With the current drugs more than half of patients are seizure-free, and others have fewer seizures. Roughly 20% of patients, however, obtain no benefit from drugs, and these patients must seek other forms of therapy. It is the job of clinicians and researchers to find better therapies for people whose epilepsy is intractable. It is the job of the general public to develop tolerance and positive attitudes toward people whose seizures cannot yet be controlled.

Further Reading

American Journal of Medical Genetics (2001) **106**. [This whole volume is devoted to the genetics of epilepsy.]

Burnham WM (1998) Antiseizure drugs. In: Kalant H and Roschlau WHE (eds) *Principles of Medical Pharmacology*. New York: Oxford University Press.

Burnham WM, Carlen PL and Hwang PA (eds) (2002) *Intractable Seizures: Diagnosis, Treatment and Prevention*. New York: Plenum Press.

Dodson WE and Pellock JM (eds) (1993) *Pediatric Epilepsy Diagnosis and Therapy*. New York: Demos.

Engel J and Pedley TA (eds) (1997) *Epilepsy: A Comprehensive Textbook*. New York: Lippincott–Raven.

Freeman JM, Kelly MT and Freeman JB (1994) *The Epilepsy Diet Treatment*. New York: Demos.

Guberman A and Bruni J (1999) *Essentials of Clinical Epilepsy*. Woburn, MA: Butterworth-Heinemann.

Luders HO (ed.) (1992) *Epilepsy Surgery*. New York: Raven Press.

McLachlan R (1997) Vagus nerve stimulation for intractable epilepsy: a review. *Journal of Clinical Neurophysiology* **14**: 358–368.

McNamara JO (1996) Drugs effective in the therapy of the epilepsies. In: Hardman JG, Limbird LE, Molinoff PB, Ruddon RW and Gilman AG (eds) *The Pharmacological Basis of Therapeutics*. New York: McGraw-Hill.

Rowan AJ and Ramsay EE (1997) *Seizures and Epilepsy in the Elderly*. Boston: Butterworth-Heineman.

Wallace S (ed.) (1996) *Epilepsy in Children*. London: Chapman & Hall Medical.

Face Cells

Intermediate article

Martin Tovée, Newcastle University, Newcastle upon Tyne, UK

CONTENTS

Face cells are neurons in the primate visual system that seem to be specialized for face recognition.

INTRODUCTION

It has been known since the 1970s that there are neurons in the monkey visual system that are sensitive to faces. These face cells have been studied in most detail in the anterior inferior temporal (IT) cortex and the upper bank of the superior temporal sulcus (STS), but they also occur in other areas such as the amygdala and the inferior convexity of the prefrontal cortex (Figure 1).

Characteristically, the optimal stimuli of face cells cannot be deconstructed into simpler component shapes (Wang *et al*., 1996). In general, these cells show virtually no response to any other stimulus tested (such as textures, gratings, bars, and the edges of various colors), but respond strongly to a variety of faces, including real ones, plastic models, and video display unit images of human and monkey faces. The responses of many face cells are invariant to size and position; the cell's response is maintained when there is a change in the size of the face, or if the position of the face within the cell's receptive field is altered. Face cells do not respond well to images of faces that have had the components rearranged, even though all the components are still present and the outline is unchanged (e.g., Perrett *et al*., 1992). Face cells are even sensitive to the relative position of features within the face; particularly important is inter-eye distance, distance from eyes to mouth, and the amount and style of hair on the forehead (e.g., Young and Yamane, 1992).

Moreover, presentation of a single facial component elicits only a fraction of the response generated by the whole face, and removal of a single component of a face reduces, but does not eliminate, the response of a cell to a face. This suggests that the face cells encode holistic information about the face, because the entire configuration of a face appears to be critical to a cell's response.

ORGANIZATION OF FACE CELLS

Studies that have combined optical imaging with single cell recording have revealed a patchy distribution of cellular activity on the cortical surface in response to faces, consistent with face cells being organized into functional columns (Wang *et al*., 1996). However, the imaging also showed that rather than discrete columns with little overlap, there was significant overlap in activity to different face orientations (Figure 2). This may mean that stimuli are mapped as a continuum of changing features (Tanaka, 1997). Such a

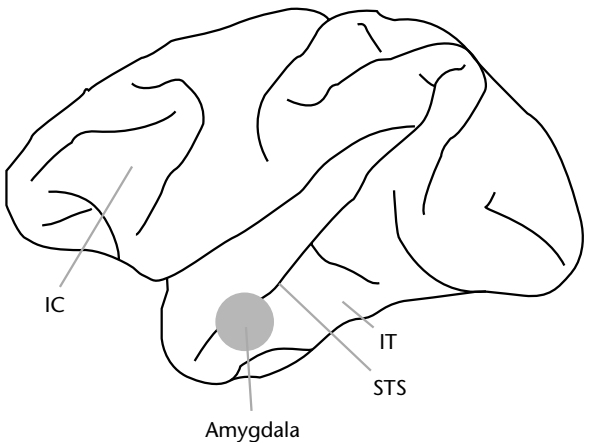


Figure 1. The cortex of a rhesus monkey showing the principal sites in which face cells have been reported: the inferior temporal cortex (IT), the superior temporal sulcus (STS), the amygdala (which is located within the temporal lobe), and the inferior convexity (IC) of the prefrontal cortex.

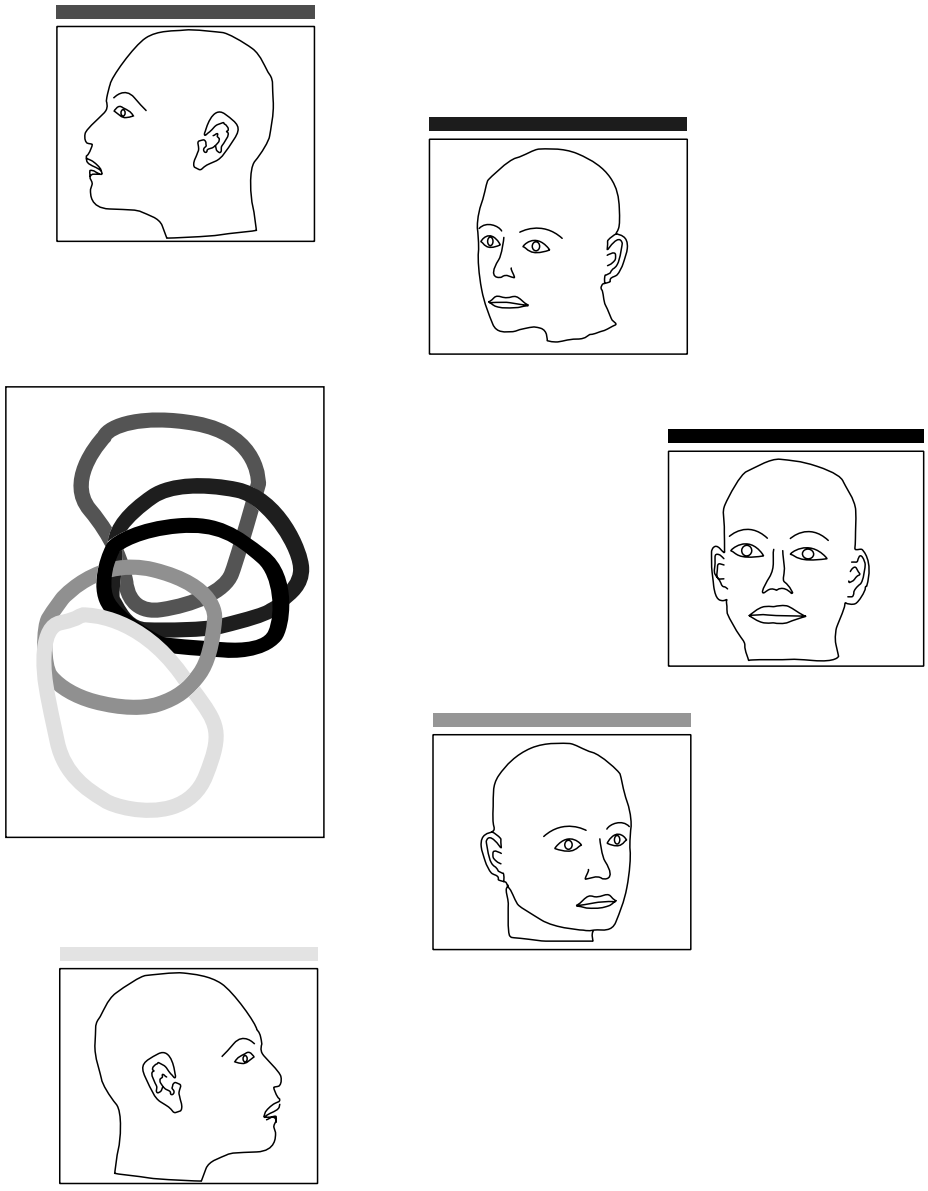


Figure 2. [*Figure is also reproduced in color section*.] Spatial pattern of activation on the surface of the inferior temporal cortex to successive presentations of a head viewed at different angles. The color of the strip above the image of the head indicates which activation pattern corresponds to which head (from Wang *et al*., 1996).

continuous map could produce a broad tuning of cortical cells for certain directions of feature space, which would allow the association of different, but related images, such as the same object from different viewpoints or under different illumination. This would obviously be an important mechanism in the development of a stimulus-invariant response.

However, feature space is a vast multidimensional area in which even the simplest 'real world' stimulus will possess a wide variety of elementary features, such as depth, color, shape, orientation, curvature, and texture, as well as specular reflections and shading (Young, 1995). Thus, a continuous representation would have to be reduced in some way to fit the limited dimensions possible in the cortex. Ultimately, a columnar organization is more likely, with cells in several columns responsive to stimuli that have features in common, and becoming jointly active as appropriate, a scheme that can also give rise to stimulus invariance.

FUNCTIONAL DIVISIONS: IDENTITY, EXPRESSION, AND DIRECTION OF GAZE

Faces can vary in a number of 'dimensions', such as identity, expression, direction of gaze, and viewing angle. Different populations of face cells seem to be sensitive to specific facial dimension, and insensitive to others. For example, Hasselmo *et al*. (1989) studied face cells in the STS and anterior IT cortex with a set of nine stimuli consisting of three different monkey faces, each displaying three different expressions. Neurons were found to respond to either dimension independently of the other. Cells that responded to expressions tended to cluster in the STS, whereas cells that responded to identity clustered in the anterior IT cortex. Further investigation has shown that there are also face cells in the STS that are responsive to gaze direction and orientation of the head (both of which are cues to the direction of attention) rather than expression (Hasselmo *et al*., 1989; Perrett *et al*., 1992). There seem to be five 'classes' of face cell in the STS, each class tuned to a separate view of the head – full face, profile, back of the head, head up, and head down (Perrett *et al*., 1992). There are an additional two subclasses, one responding to the left profile and one to the right profile.

Consistent with this finding of an anatomically segregated, functional specialization in processing different dimensions of facial information, removal of the cortex in the banks and floor of the STS of monkeys results in deficits in the perception of gaze direction and facial expression, but not in face identification (Heywood and Cowey, 1992). Perrett *et al*. (1992) suggested that the STS face cells may signal social attention, or the direction of another individual's gaze, information clearly crucial in the social interactions of primates.

Other face-cell populations also seem to be responsive to a specific dimension. The face cells in the amygdala seem to be sensitive to facial expression and may have a role in influencing and activating the emotional state of the monkey (Rolls, 1992). The face cells in the prefrontal cortex are sensitive to facial identity and seem to play a role in working memory (O'Scalaidhe *et al*., 1997). The functional organization of the different face-cell populations suggests the existence of a neural network containing processing units that are highly selective to the complex configuration of features that make up a face, and which respond to different facial dimensions (Gauthier and Logothetis, 2000).

There seem to be some homologies between the human and monkey face-processing systems. An area of the fusiform gyrus in humans has been implicated in face identification and may be the homologue of the face area in anterior IT cortex. There is also a region in the STS of both humans and monkeys that appears to be important for the processing of eye gaze and other facial expressions. Additionally, the human amygdala is active during the viewing of emotional facial expressions, particularly fear, which would be consistent with the finding of face cells responsive to expression in the monkey amygdala.

FACE CELLS AND THE POPULATION CODE

Face cells appear superficially to resemble the infamous 'grandmother cells' of the visual recognition system. These cells were described as being at the top of a processing pyramid that began with line and edge detectors in the striate cortex and continued with detectors of increasing complexity until a unit was reached that represented one specific object or person, such as your grandmother, leading to the name by which this theory became derisively known. This idea had two serious problems. First, the number of objects you meet in the course of your lifetime is immense, much larger than the number of neurons available to encode them on a one-to-one basis. Second, such a method of encoding is extremely inefficient as it would mean that there would need to be a vast number of uncommitted cells kept in reserve to code for the new objects one would be likely to meet in the future.

Although individual cells respond differently to different faces, there is no evidence for a face cell that responds exclusively to one individual face (e.g., Young and Yamane, 1992; Rolls and Tovée, 1995). Face cells seem to be part of a distributed network for the encoding of faces, just as other cells in the IT cortex probably make up a distributed network for the coding of general object features. Faces are thus encoded by the combined activity of populations, or ensembles, of cells. The representation of a face would depend on the emergent spatial and temporal distribution of activity within the ensemble (e.g., Rolls and Tovée, 1995; Rolls *et al*., 1997). Representation of specific faces or objects in a population code overcomes the two disadvantages of the grandmother cell concept. The number of faces encoded by a population of cells can be much larger than the number of cells that make up that population; so it is unnecessary to have a one-to-one relationship between stimulus and cell.

Second, no large pool of uncommitted cells is necessary. Single cell experiments have shown that the responses of individual neurons within a population alter to incorporate the representation of novel stimuli within the responses of existing populations (e.g., Rolls *et al*., 1989).

The size of the cell population encoding a face is dependent on the 'tuning' of individual cells: that is to say, how many or how few faces do they respond to? If they respond to a large number of faces, then the cell population of which they are a part must be large to signal accurately the presence of a particular face. A large cell population containing cells responsive to a large number of faces is termed 'distributed' encoding. If a cell responds to only a small number of specific faces, then only a small number of cells in the population is necessary to distinguish a specific face; this is termed 'sparse' encoding. Single cell recording experiments in monkey IT cortex have found that the face cells are quite tightly tuned and show characteristics consistent with sparse encoding (Young and Yamane, 1992; Abbott *et al*., 1996).

ARE FACE CELLS SPECIAL?

It has been suggested that face cells are a 'special' dedicated system that has arisen owing to the importance in social primates of facially conveyed information (Tanaka, 1997). All other nonface objects are represented by the distributed activity of a large number of cells selective to the different components that make up these objects. Each of these active cells (sometimes called elaborate cells) responds to only a comparatively simple shape; but across a population of cells responsive to different simple shapes, the representation of a complex object can be developed. In this framework, the face cell is qualitatively different from the neural representation of any other object class. Alternatively, there may be a 'plasticity' in the representation of visual stimuli. If you only have to make comparatively coarse discriminations, such as between different categories of objects (for example, cat versus dog), then this may be mediated by a distributed code across populations of elaborate cells. However, if you have to make fine, within-category discriminations, such as between faces, then a population of cells may become specialized for this purpose – in which case face cells are only special to the extent that they are an example of the neural machinery necessary to mediate within-category 'expert' discriminations (Tovée, 1998).

To resolve this question, one would have to train a monkey to become expert in recognizing and discriminating within a category of objects sharing a number of common features. Logothetis and Pauls did just that. They trained monkeys to discriminate within two categories of computer-generated three-dimensional shapes: wire frames or spheroidal 'amoeba-like' objects (Figure 3). The animals were trained to recognize these objects presented from one viewpoint and were then tested on their ability to generalize this recognition. Single cell recording from the anterior IT cortex during this recognition task revealed a number of cells that were highly selective to familiar views of these recently learned objects (Logothetis and Pauls, 1995; Logothetis *et al*., 1995). These cells exhibited a selectivity for objects and viewpoints that was similar to that found in face cells. They were largely size- and translation-invariant, and some cells were very sensitive to the configuration of the stimuli. In short, these cells showed the same response properties as face cells, but to computer-generated object categories.

These results suggest that the properties displayed by face cells can be duplicated for other

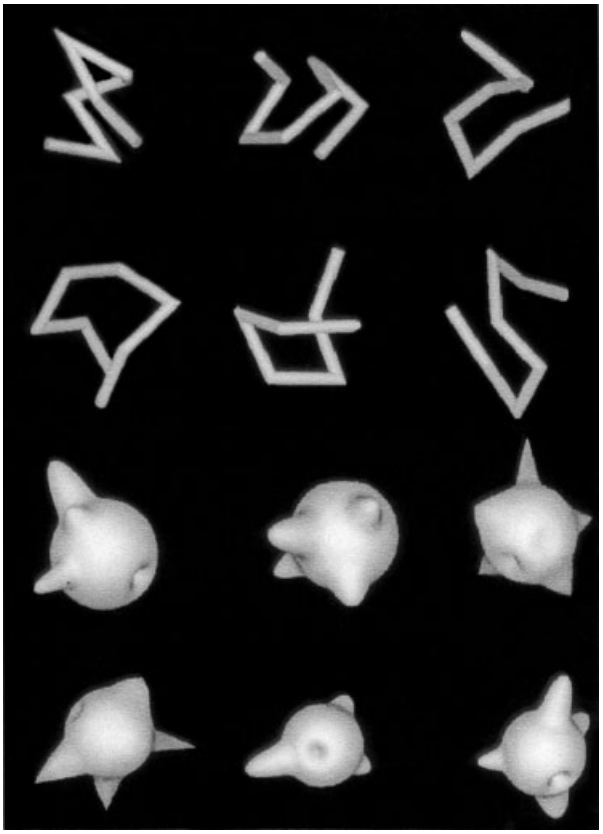


Figure 3. Examples of the wire-frame objects and 'amoeba-like' blobs used by Logothetis and Pauls (from Gauthier and Logothetis, 2000).

object categories which require fine within-category discrimination over a sustained period. Face cells are only 'special' because of the difficulty of the task in discriminating and interpreting facially conveyed information that requires a dedicated neural network. Equally difficult tasks can produce also a similar neural substrate to mediate this discrimination.

References

Abbott LF, Rolls ET and Tovée MJ (1996) Representational capacity of face coding in monkeys. *Cerebral Cortex* **6**: 498–505.

Gauthier I and Logothetis NK (2000) Is face recognition not so unique after all? *Cognitive Neuropsychology* **17**: 125–142.

Hasselmo ME, Rolls ET and Baylis GC (1989) The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Experimental Brain Research* **32**: 203–218.

Heywood CA and Cowey A (1992) The role of the 'face-cell' area in the discrimination and recognition of faces by monkeys. *Philosophical Transactions of the Royal Society of London B* **335**: 31–37.

Logothetis NK and Pauls J (1995) Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex* **5**: 270–288.

Logothetis NK, Pauls J and Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. *Current Biology* **5**: 552–563.

O'Scalaidhe SP, Wilson FAW and Goldman-Rakic PS (1997) Areal segregation of face-processing neurons in prefrontal cortex. *Science* **278**: 1135–1138.

Perrett DI, Hietnan JK, Oram MW and Benson PJ (1992) Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London B* **335**: 23–30.

Rolls ET (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society of London B* **335**: 11–21.

Rolls ET and Tovée MJ (1995) The sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology* **73**: 713–726.

Rolls ET, Baylis GE, Hasselmo ME and Nalwa V (1989) The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research* **76**: 153–164.

Rolls ET, Treves A and Tovée MJ (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research* **114**: 149–162.

Tanaka K (1997) Mechanisms of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology* **7**: 523–529.

Tovée MJ (1998) Is face processing special? *Neuron* **21**: 1239–1242.

Wang G, Tanaka K and Tanifuji M (1996) Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* **272**: 1665–1668.

Young MP (1995) Open questions about the neural mechanisms of visual pattern recognition. In: Gazzanida MS (ed.) *The Cognitive Neurosciences*, pp. 463–474. London, UK: MIT Press.

Young MP and Yamane S (1992) Sparse population coding of faces in the inferotemporal cortex. *Science* **256**: 1327–1331.

Further Reading

Fujita I, Tanaka K, Ito M and Cheng K (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* **360**: 343–346.

Kanwisher N (2000) Domain specificity in face perception. *Nature Neuroscience* **8**: 759–763.

Tarr MJ and Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* **3**: 764–769.

Tovée MJ, Rolls ET and Ramachandran VS (1996) Rapid visual learning in the neurons of the primate temporal visual cortex. *NeuroReport* **7**: 2757–2760.

Tsunoda K, Yamane Y, Nishizaki N and Tanifuji M (2001) Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience* **4**: 832–838.

Face Perception, Neural Basis of

Intermediate article

Shlomo Bentin, Hebrew University of Jerusalem, Jerusalem, Israel

CONTENTS

There is neurophysiological evidence to support the existence of a specific neural mechanism in the extrastriate cortex, which is tuned to detect physiognomic information in the visual field and form a mental image, capturing those aspects of the face that enable us to distinguish it from other faces, a stage that is considered to precede personal identification.

INTRODUCTION

The outstanding expertise of humans in recognizing faces has led scientists to suspect that this perceptual process is based on dedicated neural structures. Indeed, face recognition in humans has often been paralleled to phonetic perception, as opposite poles on a hemispherical asymmetry continuum. Whereas the latter has been attributed to dedicated structures buried in the supra-temporal plane of the left hemisphere, the normal processing of faces has been assumed to engage primarily (but not exclusively) posterior-temporal structures in the right hemisphere. The view that the neural structures that are devoted to face perception are also *exclusively* dedicated to this process is controversial. In this article I shall review evidence suggesting that neural structures are implicated in face processing, and discuss their domain specificity. In particular, I shall focus on perceptual processes aimed at constructing a visual representation that is sufficiently detailed and complete to allow unequivocal categorization and identification of the face.

Current models of visual perception distinguish between levels of integration and representation in the visual system (Marr, 1982; Biederman, 1987, 1995; Ullman, 1995, 1996). Despite differences in emphasis and general approach, all of these models assume that the categorization of visually presented stimuli is based on the formation of a mental representation, which is reconstructed from changes in light energy impinging on the retina. Whereas the initial detection of these changes, in terms of edges between different levels of illumination, contrast, orientation and color, is performed in the striate and peristriate visual cortices, it is well accepted that the process of integrating the basic visual primitives into higher-level visual representations is a product of extrastriate neural mechanisms, most of which are distributed along the parvocellular ventral pathway of the visual system, the so-called 'what' pathway. To remind the reader, sensory information from the primary visual cortex reaches the temporal and parietal lobes to form two relatively (but not completely) separate pathways (Ungerleider and Mishkin, 1982; Desimone and Ungerleider, 1989). One of these is the 'where' system, which passes dorsally in the extrastriate cortex to end in the posterior parietal lobule (and additional ramifications to the frontal lobe). This is the dorsal, magnocellular pathway, which contains cells that are particularly sensitive to the stimulus location in space, and to movement. The other pathway is the 'what' system, which passes ventrally in the extrastriate cortex to reach the inferior temporal cortex, which includes (among other structures) the inferotempral gyrus (IT), the occipito-temporal (fusiform) gyrus, the lingual gyrus and the superior temporal sulcus (STS). This is the parvocellular ventral pathway, which contains cells that are involved in the formation of object-oriented, category-specific visual representations (Van Essen and Deyoe, 1995).

NEUROPSYCHOLOGICAL EVIDENCE

Evidence that faces are processed by dedicated neural mechanisms located in the ventral and inferior-temporal regions comes from a number of different sources. The most notable of these are neuropsychological observations of patients with impaired face recognition (prosopagnosia), electrical activity of single cells recorded in monkeys and of cell assemblies recorded in humans (either directly from the cortical surface or on the scalp

surface), and imaging of brain activity using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI).

Although in most cases of acquired prosopagnosia (impaired face recognition due to brain damage) face-processing deficits are often the most conspicuous aspect of a more general visual agnosia (Gauthier *et al*., 1999a), there are a few reports of prosopagnosic patients whose ability to recognize objects remained intact (McNeil and Warrington, 1993; Farah *et al*., 1995, 2000; Bentin *et al*., 1999). Conversely, there are reports of patients suffering from associative object agnosia whose face recognition ability was spared (Moscovitch *et al*., 1997). The double dissociation between face and object recognition suggests that the two abilities may be neurologically as well as functionally distinct.

EVIDENCE FROM SINGLE-CELL STUDIES

Single-cell recordings in the monkey demonstrated the existence of cells in the temporal lobes that are tuned to respond selectively to monkey (as well as human) faces, but not to other complex and emotion-arousing stimuli such as snakes, spiders or food (Perrett *et al*., 1987, 1990; Desimone, 1991; Gross, 1992; Logothetis and Scheinberg, 1996). Face-selective cells were predominantly found in the superior temporal sulci (STS) and inferior temporal gyri (IT), as well as in the amygdala and the inferior convexity of the prefrontal cortex. Whereas some of these cells responded more vigorously to isolated eyes than to whole faces, other cells responded only to the entire face-view configuration – that is, they were sensitive to the holistic face shape rather than to the existence of individual features. The selectivity of these neurons for faces was maintained despite changes in stimulus size and position. Furthermore, many of these cells responded only to particular orientations of faces or particular directions of gaze. A detailed investigation revealed five types of face-specific cells in the STS, each being maximally responsive to one view of the head, namely full face, profile, back of the head, head up and head down. In addition, two subtypes have been discovered that respond only to left profile or only to right profile, which confirms that these cells are involved in the structural analysis of the face, rather than in representing specific social or emotional responses that faces might elicit (Perrett *et al*., 1985). In summary, the activity of single cells in the monkey suggests that with regard to visual analysis of faces, there is some specificity in the IT and STS. The sensitivity of different cells to different stimulus characteristics suggests a high degree of specialization, providing a complex mechanism of face encoding in the extrastriate cortex. However, the fact that the same cells respond to human faces as well as to within-species (monkey) faces also suggests a high degree of neural plasticity and susceptibility to visual experience.

NEUROIMAGING EVIDENCE

Suggestive as they are, data characterizing the visual system of the monkey cannot be immediately generalized to humans. Indeed, from a phylogenic perspective we see a continuous trend of *reduced* specialization of individual cells, at least in the primary visual cortex. Fortunately, modern technology has enabled the recording of stimulus-linked brain activity in humans. In particular, functional brain imaging using PET and fMRI, and the recording of event-related potentials (ERPs), have yielded pertinent data that suggest face-processing specificity.

The distribution of hemodynamic changes in brain tissue is correlated with neural activity, because elevated metabolism requires more oxygen. Since oxygen is transported to the tissue by hemoglobin cells, the amount of blood flowing through a particular region correlates with its relative activity. PET and fMRI are two techniques which may reveal task-related changes in regional cerebral blood-flow. A series of PET studies suggested that there is posterior-ventral localization of face-specific visual processing (Huxby *et al*., 1996). For example, in one of the first attempts to isolate components of face processing from the processing of objects and nonsense shapes it was found that, compared with nonsense gratings and sinusoid shapes, both faces and objects activate extensive areas in the occipito-temporal cortex. However, whereas this activation was more pronounced in the right hemisphere for faces, it was greater in the left hemisphere for objects (Sergent *et al*., 1992). More recently, fMRI studies have provided a more detailed description. Faces activated the fusiform and middle occipital gyri, the lateral occipital sulcus and more anteriorly the superior temporal sulcus (STS), and this activation was greater in the right than in the left hemisphere (Kanwisher *et al*., 1997; McCarthy *et al*., 1997) (see also Figure 2). In addition, fMRI studies showed distinct activation for perceptual categories other than faces in regions adjacent to the face areas. Although the neuroimaging studies were consistent in showing neuroanatomical specificity for face processing (but see

Gauthier *et al*., 1999b; Gauthier *et al*., 2000), they did not establish when such processing occurs. A tentative answer to the time-course question is provided by the ERP studies.

ELECTROPHYSIOLOGICAL EVIDENCE

The flow of ionic currents across the cell membranes of active neurons and glia cells produces electrical potentials that can be recorded from the cortical surface or from the scalp by disk electrodes (several millimeters in diameter) and represented in the electroencephalogram (EEG) as voltage as a function of time. ERPs are processing-induced changes in these potentials – changes that are time-locked to the brain event that elicited them. These changes become conspicuous on the background of the ongoing EEG following the averaging of designated time epochs that encompass the onset of the stimulus under investigation and are time locked to it. ERPs are usually defined in terms of polarity relative to a neutral reference, onset and peak latency, peak or mean amplitude, and scalp distribution. Recorded non-invasively on the scalp, this technique can be used to investigate normal functioning of the human brain. However, scalp recording does not allow firm conclusions to be drawn about the location of the relevant active structures in the brain. One way to (partly) overcome this limitation is to record the ERPs directly from the cortical surface (and/or from deeper structures) when such invasive recordings are clinically indicated. In a series of studies, investigators at the Yale New Haven Medical Center examined the neural specificity for face processing by recording the electrical activity directly from the ventral and lateral cortex of the temporal lobes. This procedure was possible in patients with medically refractory epilepsy who were implanted with electrodes for localizing the focus of the seizures.

Extensive and systematic electrophysiological investigation of about 100 patients revealed discrete regions in the human extrastriate cortex which were activated by faces but not by other categories of visual stimuli such as cars, flowers, human hands, butterflies or printed words (Allison *et al*., 1999; McCarthy *et al*., 1999; Puce *et al*., 1999). A region was considered to be face specific if the amplitude of any of the ERP components (or a combination of them) recorded at this site was at least twice as large in response to a face than in response to other stimulus categories (note that the same criterion was used to determine the face selectivity of single cells). From a total of more than 7500 recording sites (across patients), about 100 sites were face specific. Most importantly, whenever a component was significantly larger in response to faces, its amplitude was similar across all other stimulus categories (Figure 1). Consistent with the fMRI studies, the vast majority of these sites were clustered on the ventral occipito-temporal cortex, and some were found on the posterior lateral surface of the temporal lobes. A further differentiation of the face-specific ventral sites was between the posterior clusters and an anterior cluster, with no overlap between them. Whereas the posterior ventral and lateral face areas were bilaterally distributed, the anterior face area was only found in the right hemisphere.

Face specificity was evident during three latency windows (from stimulus onset). The earliest face-specific component was a robust negative (compared with the reference) potential that peaked at around 200 ms (N200). This component was elicited in the posterior ventral and lateral face areas. Its amplitude was similar in men and in women, and its latency was slightly more delayed in men than in women, probably reflecting the relatively larger brains in men. The second face-specific component was a positive potential with an average peak of about 350 ms (P350). It was recorded bilaterally from the posterior ventral face area, bilaterally from the lateral face area, and in the right hemisphere from the anterior ventral face area. The third face-specific component was a sustained negativity (N700) observed in the waveforms elicited at about half of the sites at which the N200 was face specific, but also at some sites where the N200 was not specific to faces.

The response properties of this neuronal activity were examined in a series of experiments. In general, the results of these investigations were similar to those obtained by recording face-specific activity of single cells. Neither the N200 nor the N700 were sensitive to the color of the face or to its size. Moreover, all face-specific components were similar in response to unfamiliar and famous faces. Face inversion (i.e. presenting the face upside down) had a complex effect, depending on the location of the face in the visual field. Compared with the face specificity of single cells in the monkey, the human N200 was more species-selective. Although faces of cats and dogs elicited an N200 response which was significantly larger than that elicited by non-face stimuli, the amplitude of this N200 was half the size and its latency was significantly delayed compared with the response to human faces. This pattern was replicated in scalp recordings, in which human and primate but not other animal faces elicited a distinctive component, the N170 (Bentin

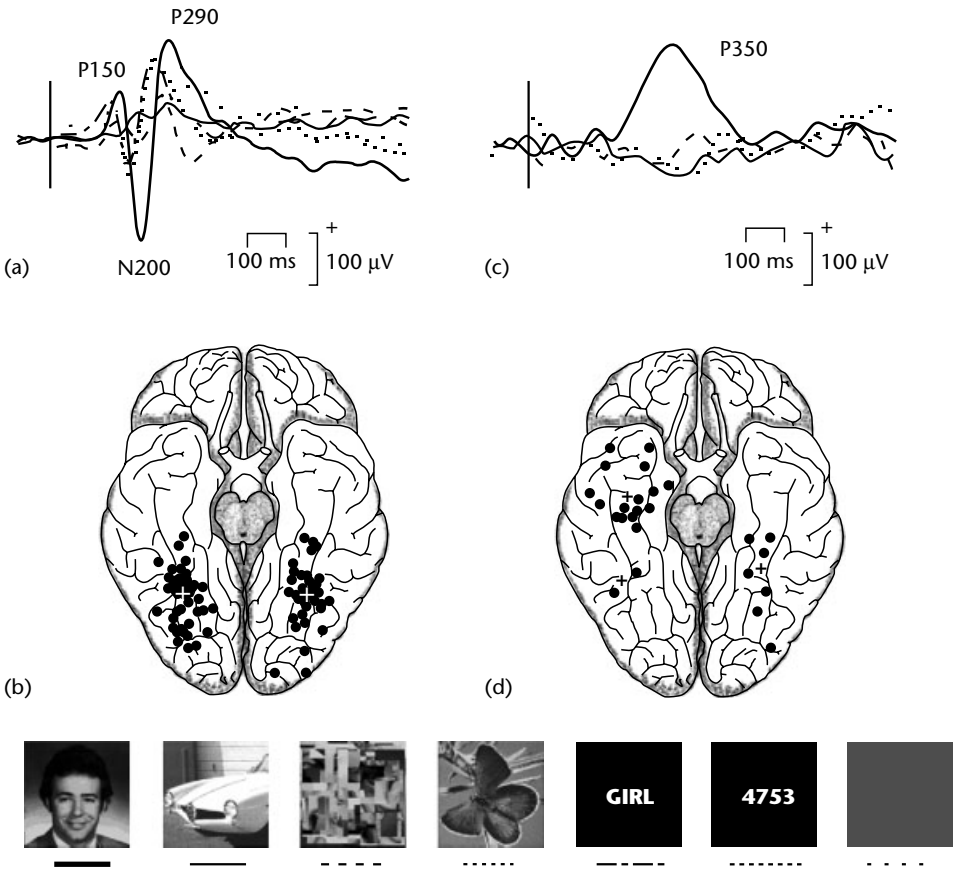


Figure 1. Face-specific ERP components recorded on the cortical surface. (a) N200 and its adjacent positive peaks. (b) Face-specific sites (black dots) in the posterior-ventral temporal lobes. (c) Face-specific P350. (d) Distribution of the P350. Note the unilateral right hemisphere distribution of the anterior sites. (Data provided courtesy of Dr Aina Puce.)

et al., 1996; Carmel and Bentin, 2002). Taken together, the ERP results suggest that the face-specific components recorded on the surface of the fusiform gyrus as well as those recorded from the posterior inferior temporal gyrus (IT) and those recorded on the scalp are associated with neuronal mechanisms that analyze the structural representation of the face, forming its internal representation, and that they are not involved in within-category face identification.

Additional results from intracranial as well as scalp recordings have revealed some characteristics of the face-encoding process. Behavioral and neuropsychological evidence strongly indicates that face identification is based on a holistic process – that is, the gestalt configuration of the face is analyzed first, whereas the analysis of individual face components (if necessary at all) is a late process, independent of face identification (Tanaka and Farah, 1993). Nonetheless, the possibility that the holistic process on which identification is based is only part of a more complex face-encoding mechanism that may also include the analysis of face parts has not been ruled out. Indeed, the finding that despite its significant detrimental effect on face identification, face inversion does not have dramatic effects on either the N200 or the N170 suggests that impeding configurative analysis of the face has only a minor influence on the activity of the face-encoding neuronal process with which these ERP components were associated. To investigate this hypothesis, the sensitivity of the N200 to face parts was directly examined. The results demonstrated that the N200s elicited by isolated eyes, lips, noses or face contours were significantly distinct from those elicited by objects such as cars, flowers and items of furniture, which suggests that the associated neuronal mechanism processed face parts (whether internal or external) as face-like rather than as object-like stimuli. Yet at all face-specific sites the N200 amplitude was reduced relative to that elicited by full-face gestalt. Moreover, as for inverted or degraded faces, the N200 latencies to face parts were significantly longer than those to full faces, suggesting

that analysis of unusual or partial views of a face requires additional processing time. However, the fact that neither the latency nor the amplitude of the N200 to faces is the linear sum of the response to face parts does not support a hierarchy of processing in which face-part cells send their output to a next stage of face processing. It is possible that individual face components are analyzed in parallel by other neuronal mechanisms, and that the accumulated information is integrated by the face-specific mechanism manifested in the N200. The latter hypothesis is supported by the finding of certain regions, lateral to the ventral face-specific N200 sites, which respond preferentially to face components, particularly eyes. These part-sensitive sites are primarily located in the inferior temporal gyrus, at the border between the ventral and lateral cortex. Given their location and the orientation of the cellular columns in that part of the cortex, their activity probably accounts for the modulation of the face-specific N170 recorded on the scalp. Indeed, one of the most consistent findings is that the amplitude of the N170 elicited by isolated eyes is significantly larger and its latency is longer than the N170 elicited by full faces.

Is the face-specific encoding modulated by top-down processes engaging previous knowledge, contextual information and expectations? A series of intracranial as well as scalp recordings designed to answer this question revealed a complex pattern of results. On the one hand, the intracranial recorded N200 was not significantly affected by

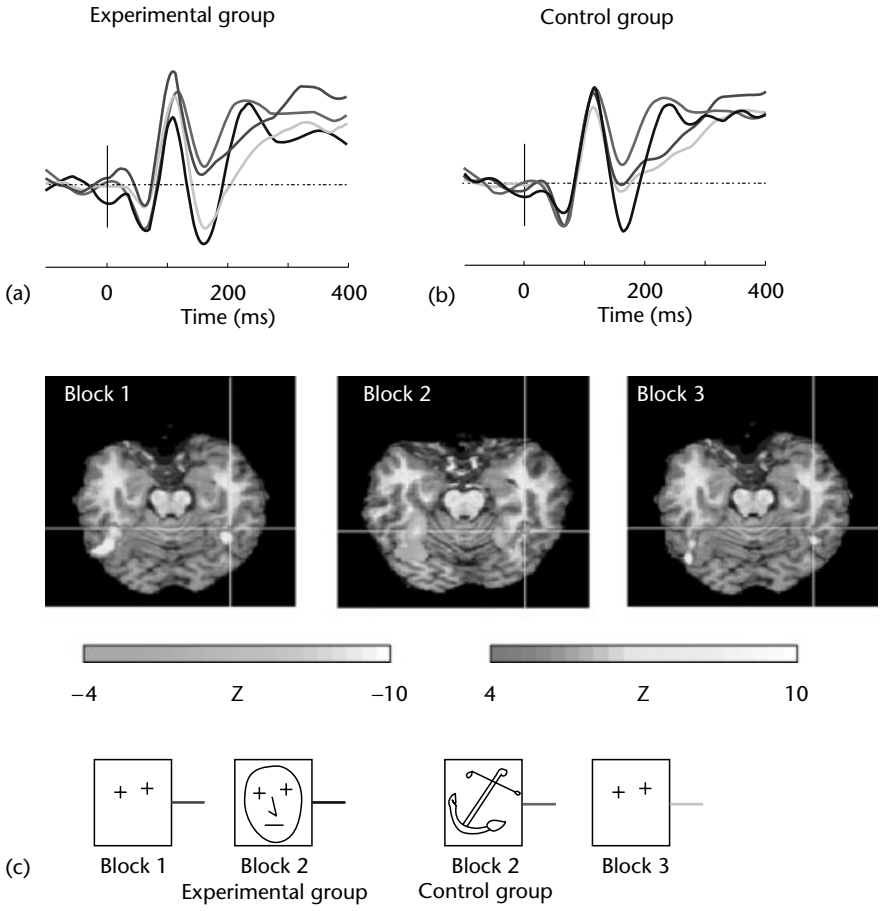


Figure 2. [*Figure is also reproduced in color section*.] Priming effects on face-specific neural activity. (a) The ERPs elicited by two crosses seen in block 1 resemble those elicited by objects. The same stimuli seen in block 3, after a face context was suggested, elicit an N170 similar to that elicited by faces. (b) If objects rather than faces are presented in block 2, the face-specific mechanism is not switched on. (c) Priming effects seen by fMRI. Note in the middle panel the right-hemisphere-lateralized activity elicited by faces (warm colors), and the bihemispheric activity elicited by objects. The two crosses shown in block 1 activate areas similar to those activated by objects. When provided with a face context, the same stimuli activate more restricted areas, albeit bilateral ones. (The fMRI data were collected at the Max-Planck Institute of Cognitive Neuroscience in collaboration with Axel Mecklinger, Yves von Cramon and Angela Friederici.)

stimulus repetition, showing no habituation. It was identical for familiar and unfamiliar faces, and it was not affected by semantic priming. On the other hand, the scalp N170 seems to be sensitive to conceptual influences. First, it is elicited as efficiently by schematic drawings of faces as by natural faces, provided that the schematic representation is simple and clear (Sagiv and Bentin, 2001). Secondly, whereas simple pairs of lines do not normally elicit an N170, they would do so if the subject was primed to interpret the lines as eyes in a schematic face (Figure 2) (*see* Bentin *et al*., 2002).

SUMMARY

The neurophysiological data suggest that faces are encoded in the human visual system by a specialized neural system located in the lateral posterior fusiform gyrus and inferotemporal gyrus. Some of these areas (particularly those located in the fusiform gyrus) probably process the face as a gestalt. The more lateral sites appear to be sensitive to face components, particularly the eyes. The shorter latency of the scalp-recorded N170 relative to the intracranial N200 suggests that the processing of the face parts starts slightly earlier than processing of the gestalt. Yet the exact relationships between these two components of the face-encoding system are unclear. Whether the face-specific activity is a manifestation of an innate module, as the visual preference of infants for faces suggests, whether it is imperative and domain specific, as is suggested by the insensitivity of the N170 to task-relevance of the face (Carmel and Bentin, 2002), or whether this is the manifestation of a change in visual processing induced by expertise (e.g. from an analytic, part-based strategy to a holistic strategy), as is suggested by some behavioral studies (Gauthier and Tarr, 1997) and ERP studies (Tanaka and Curran, 2001), has yet to be established.

Acknowledgements

This work was supported by the US-Israel Bi-National Science Foundation. The author wishes to thank Dr Leon Deouell for his critical comments on this article.

References

Allison T, Puce A, Spencer DD and McCarthy G (1999) Electrophysiological studies of human face perception. I. Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex* **9**: 415–430.

Bentin S, Allison T, Puce A, Perez E and McCarthy G (1996) Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience* **8**: 551–565.

Bentin S, Deouell LY and Soroker N (1999) Selective streaming of visual information in face recognition: evidence from congenital prosopagnosia. *Neuroreport* **10**: 823–827.

Bentin S, Sagiv N, Mecklinger A, Friederici A and von Cramon DY (2002) Conceptual priming in visual face-processing: electrophysiological evidence: *Psychological Science* **13**: 190–193.

Biederman I (1987) Recognition-by-components: a theory of human image interpretation. *Psychological Review* **94**: 115–147.

Biederman I (1995) Visual object recognition. In: Kosslyn M, Osherson DN (eds) *Visual Cognition*, pp. 121–166. Cambridge, MA: MIT Press.

Carmel D and Bentin S (2002) Domain specificity versus expertise: factors influencing distinct processing of faces. *Cognition* **83**: 1–29.

Desimone R (1991) Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience* **3**: 1–8.

Desimone R and Ungerleider LG (1989) Neural mechanisms for visual processing in monkeys. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, pp. 267–299. Amsterdam: Elsevier.

Farah MJ, Levinson KL and Klein KL (1995) Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia* **33**: 661–674.

Farah MJ, Rabinowitz C, Quinn GE and Liu GT (2000) Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology* **17**: 117–123.

Gauthier I and Tarr MJ (1997) Becoming a 'Greeble' expert: exploring mechanisms for face recognition. *Vision Research* **37**: 1673–1682.

Gauthier I, Behrmann M and Tarr MJ (1999a) Can face recognition be dissociated from object recognition? *Journal of Cognitive Neuroscience* **11**: 349–370.

Gauthier I, Tarr MJ, Anderson AW, Skudlarski P and Gore JC (1999b) Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience* **6**: 568–573.

Gauthier I, Skudlarski P, Gore JC and Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience* **3**: 191–197.

Gross CG (1992) Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London* **B-335**: 3–10.

Huxby JV, Ungerleider LG, Horowitz B *et al*. (1996) Face encoding and recognition in the human brain. *Proceedings of the National Academy of Sciences of the USA* **93**: 922–927.

Kanwisher N, McDermott J and Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**: 4302–4311.

Logothetis NK and Scheinberg DL (1996) Visual object recognition. *Annual Review of Neuroscience* **19**: 577–621.

McCarthy G, Puce A, Gore JC and Allison T (1997) Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience* **9**: 605–610.

McCarthy G, Puce A, Belger A and Allison T (1999) Electrophysiological studies of human face perception. II. Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral Cortex* **9**: 431–444.

McNeil JE and Warrington EK (1993) Prosopagnosia: a face-specific disorder. *Quarterly Journal of Experimental Psychology* **46A**: 1–10.

Marr D (1982) *Vision*. San Francisco, CA: WH Freeman.

Moscovitch M, Winocur G and Behrmann M (1997) What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience* **9**: 555–604.

Perrett DI, Smith PAJ, Potter DD *et al*. (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London* **B-223**: 293–317.

Perrett DI, Mistlin AJ and Chitty AJ (1987) Visual neurons responsive to faces. *Trends in Neuroscience* **10**: 358–364.

Perrett DI, Harries MH, Mistlin AJ *et al*. (1990) Social signals analyzed at the cell level: somone is looking at me, something touched me, something moved! *International Journal of Comparative Psychology* **4**: 25–54.

Puce A, Allison T and McCarthy G (1999) Studies of human face perception. III. Effects of top-down processing on face-specific potentials. *Cerebral Cortex* **9**: 445–458.

Sagiv N and Bentin S (2001) Structrual encoding of human and schematic faces: holistic and part-based processes. *Journal of Cognitive Neuroscience* **13**: 937–951.

Sergent J, Ohta S and MacDonald B (1992) Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain* **115**: 15–36.

Tanaka JW and Farah MJ (1993) Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology* **46A**: 225–245.

Tanaka JW and Curran T (2001) A neural basis for expert object recognition. *Psychological Science* **12**: 43–47.

Ullman S (1995) The visual analysis of shape and form. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 339–350. Cambridge, MA: MIT Press.

Ullman S (1996) *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press.

Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ (ed.) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Van Essen DC and Deyoe EA (1995) Concurrent processing in the primate visual cortex. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 383–400. Cambridge, MA: MIT Press.

Fragile X Syndrome

Introductory article

Louise W Gane, MIND Institute, University of California, Davis, California, USA

CONTENTS

Fragile X syndrome (FXS) is a leading cause of mental retardation, associated with a broad spectrum of involvement, ranging from mild emotional problems or learning disabilities through all levels of mental retardation. Since 1991 it has been known that the disorder is caused by a single gene on the X chromosome.

INTRODUCTION

Fragile X syndrome (FXS) is the leading cause of *inherited* mental retardation. It is associated with a broad spectrum of involvement, ranging from mild emotional problems or learning disabilities through all levels of mental retardation. One in approximately 4000 males in the general population is affected with FXS. A similar number of females will also experience problems, but females are generally less affected than males. Since 1991 it has been known that the disorder is caused by a single gene on the X chromosome. In some people, this gene undergoes a change that causes an expansion in the gene which leads to FXS. The growth of knowledge with regard to FXS has been incremental, covering a span of 60 years, and is still continuing.

CLINICAL FEATURES OF FRAGILE X SYNDROME

The classic physical features of FXS include a long face, prominent ears, prominent chin, large head circumference, high arched palate, and loose flexible joints. Many of the physical features are associated with the connective tissue problems linked with FXS. Low muscle tone, hyperextensibility, hernias, scoliosis, and flat feet are common. Young children affected with FXS are often floppy and have recurrent ear infections, sinusitis, and/or gastric reflux. Sleep problems in childhood are common. Children with the disorder often show rapid growth. However, as adults, stature is usually short. Girls with FXS are at risk of precocious puberty. At puberty, large testicles are often seen in males. Mitral valve prolapse (a heart condition) is found in 50% of males who have FXS. The commonest neurological problem associated with the disorder is seizures, which occur in 20% of affected males and approximately 5% of affected females.

Girls with FXS often have the same physical findings as boys. However, since the physical findings in girls are more subtle, they can easily be overlooked.

The classic behavioral features associated with FXS are present in early childhood. Language delay, hyperactivity, and short attention span are often noted by the second or third year. Rapid and repetitive speech patterns are associated with the syndrome. Repetitive patterns of behavior, such as watching the same video or television program over and over again, are also characteristic of the disorder. Often both boys and girls with FXS will have difficulty in making direct eye contact. In addition, shyness, anxiety and over-sensitivity to environmental stimuli, often expressed in tantrums, are commonly seen in children with FXS. These tantrums may also be associated with changes in routine or environment, crowded situations, overstimulation and/or too much activity. In addition, autistic-like features are associated with FXS, including poor eye contact, shyness, social anxiety, hand biting, hand flapping, hand posturing, and repetition of speech and/or behaviors.

Girls are very likely to have behavioral difficulties related to shyness and anxiety. Sometimes the latter can become incapacitating and lead to lack of verbal communication with selected people. In addition, short attention span (with or without hyperactivity) and impulsivity may be seen in girls with FXS.

Approximately 20–30% of children with FXS will be diagnosed with true autism. Those who are more severely cognitively impaired and those

who are non-verbal or very late in developing speech are more likely to be diagnosed with autism and FXS. However, children who have FXS and autism or FXS with autistic-like features respond well to therapeutic interventions. The latter should include speech and language therapy and occupational therapy that emphasizes sensory integration therapy. In addition, applied behavioral analysis (ABA) and discrete trial training with positive reinforcement may help some of these children. With such interventions, children who have autism and FXS will often outgrow the dual diagnosis and no longer meet the criteria for autism.

The classic cognitive profile of FXS includes impairment of understanding of information provided from other sources, remembering sequential information that is given verbally, and the ability to think abstractly or complete tasks that require complex reasoning. Object memory is a strength, as is simultaneous processing. This means that learning from their experiences at home and school as well as those with whom they interact are very helpful to individuals with FXS. Mathematical skills are weak in both males and females with the disorder. However, early strengths in language and long-term memory are seen, and these are often reflected in the social abilities and sense of humor of affected individuals. In addition, many of those with FXS have strong copying and mimicking skills which strengthen their ability to learn, particularly in the early school years.

Approximately 50–70% of females with FXS will have an IQ in the low or borderline range (70 to 85). However, 30–50% of females with the disorder will have a normal IQ (above 85). However, of these females who have an IQ in the normal range, 60% will have emotional or behavioral problems. These problems often take the form of shyness, anxiety, mood swings, and/or problems with mathematics. However, tasks involving attention, organizational skills, and the ability to inhibit impulses will also often reflect the difficulties experienced by females.

GENETIC BASIS OF FRAGILE X SYNDROME

The gene responsible for FXS is located near the end of the X chromosome, (Xq27.3), and is known as the *F*ragile X *M*ental *R*etardation 1 (FMR1) gene. Although everyone has the FMR1 gene, in some people it undergoes a mutation. The mechanism of the change is such that the gene will expand in one place where there are normally a number of repeats involving two (cytosine (C) and guanine (G)) of the four chemical bases (nucleotides) of which a gene is composed. These repetitions are referred to as *CGG repeats*. This repeat characteristic means that FXS is also referred to as a *trinucleotide repeat disorder*. When the change occurs in the CGG repeats, the number of such repeats becomes larger. The increase in the number of CGG repeats will continue to be transmitted from one generation to the next as the X chromosome containing the gene change is passed on to offspring. The FMR1 gene change will increase dramatically in repeat size when it is passed to a child through a female. Eventually the mutation will jump to a large expansion of over 200 CGG repeats (full mutation), and the child will have FXS. One in 259 women carry the FMR1 gene change, with the CGG repeats ranging from 55 to 200. When the repeats are in this range, the gene change is called a *premutation*. One in 755 men carry the premutation. This means that those men and women who are carriers of the premutation are not affected cognitively by the FMR1 gene. However, women who carry the premutation are at risk of having a child with the 'full mutation' (more than 200 repeats) and affected with FXS.

With expansion of the CGG repeats to more than 200, the FMR1 gene within the cell is no longer able to produce normal levels of the *F*ragile X *M*ental *R*etardation *P*rotein (FMRP) which is required for normal brain development. It is this lack of FMRP that leads to FXS.

GENETIC COUNSELING AND TESTING FOR FRAGILE X SYNDROME

When an individual is diagnosed with FXS, it is important that all family members receive genetic counseling. Meeting with a genetic counselor enables the relatives to gain an understanding of the condition itself and how the FMR1 gene expands through the generations, and to identify family members at risk of carrying the gene change and having offspring with FXS. The family pedigree obtained by the genetic counselor is a tool that is used to gain a detailed family history, to identify previously undiagnosed family members, and to identify the inheritance pattern within the family and determine which family members should have testing to assess their FMR1 gene status. In addition, the genetic counselor will coordinate the family testing and interpret the test results in a manner that can be understood by all family members.

The genetic counselor can also be instrumental in interfacing with other professionals, advocating in support of the needs of those diagnosed with FXS,

and identifying and helping to meet the needs of other family members. Obtaining the necessary services from the medical, educational, and community services that will be of help to the family is often part of the role of the genetic counselor. He or she is often the person who will first bring the support services to the attention of the family.

It is also the genetic counselor who may be helpful to the family in working through the emotional impact of the diagnosis. Having a child with a genetic diagnosis can have a serious although often temporary effect on self-esteem. Grief over the loss of a healthy child is part of the complicated process of coping with such a diagnosis. Family members may also experience feelings of guilt, blame, embarrassment, stigmatization, denial, disappointment, and loss of expectations that they had for their child. Although it may be a relief to obtain the diagnosis and to know what is wrong, at the same time this may mean that hopes and dreams are shattered and will need to be rebuilt.

When a family member receives the diagnosis of FXS, the reproductive plans of the childbearing generation are affected. Genetic counseling is recommended for any family member who is pregnant or who is considering a future pregnancy. It is at this time that prenatal testing options can be reviewed and family members can make informed choices.

CONCLUSION

FXS is a complex genetic disorder that has already led the way to understanding other genetic disorders. Although the clinical picture of FXS is well defined, the focus of research is on understanding the consequences of the gene change in relation to the clinical findings, and on characterizing the role of FMRP. In addition, better and more specific medical and therapeutic interventions are being sought in order to enhance the potential of all those diagnosed with FXS.

Further Reading

Amaria RN, Billeisen LL and Hagerman RJ (2001) Medication use in fragile X syndrome. *Mental Health Aspects of Developmental Disabilities* (in press).

Braden ML (ed.) (1997) *Fragile, Handle with Care: Understanding Fragile X Syndrome*, 2nd edn. Chapel Hill, NC: Avanta Publishing.

Hagerman RJ (ed.) (1999) *Neurodevelopmental Disorders: Diagnosis and Treatment*. New York, NY: Oxford University Press.

Hagerman RJ and Hagerman PJ (eds) (2002) *Fragile X Syndrome: Diagnosis, Treatment and Research*. Baltimore, MD: Johns Hopkins University Press.

Mazzocco MM, Pennington BF and Hagerman RJ (1993) The neurocognitive phenotype of female carriers of fragile X: additional evidence for specificity. *Journal of Behavioral and Developmental Pediatrics* **14**: 328–335.

National Fragile X Foundation. http://www.fragilex.org.

Rousseau F, Rouillard P, Morel ML, Khandjian EW and Morgan K (1995) Prevalence of carriers of premutation-size alleles of the FMR1 gene – and implications for the population genetics of the fragile X syndrome. *American Journal of Human Genetics* **57**: 1006–1018.

Sobesky WE, Taylor AK, Pennington BF *et al*. (1996) Molecular/clinical correlations in females with fragile X. *American Journal of Medical Genetics* **64**: 340–345.

Staley-Gane LW, Flynn L, Neitzel K, Conister A and Hagerman RJ (1996) Expanding the role of the genetic counselor. *American Journal of Medical Genetics* **64**: 382–387.

Tassone F, Hagerman RJ, Chamberlain WD and Hagerman PJ (2000) Transcription of the FMR1 gene in individuals with fragile X syndrome. *American Journal of Medical Genetics* **97**: 195–203.

Verkerk AJ, Pieretti M, Sutcliffe JS *et al*. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**: 905–914.

Frontal Cortex

Intermediate article

Donald T Stuss, The Rotman Research Institute of Baycrest Centre for Geriatric Care; University of Toronto, Toronto, Ontario, Canada
Darlene Floden, The Rotman Research Institute of Baycrest Centre for Geriatric Care; University of Toronto, Toronto, Ontario, Canada

CONTENTS

The frontal lobes (or cortex) serve as the primary bases of social interactions, self-awareness, and the regulation of more modular functions such as perception, memory, and language.

INTRODUCTION

The frontal lobes (or cortex) are considered to be the anatomical area related to the highest of human functions, including the temporal integration of sensory information for the organization of behavior such as memory and language, and the cohesion of thinking and feeling to structure social behavior (Fuster, 1985; Stuss and Benson, 1984).

A HISTORY OF FRONTAL LOBE FUNCTIONS

Conflicting historical opinions about the role of the frontal lobes stemmed from limitations at both the methodological and the conceptual levels. Methodologically, there has been difficulty in identifying patients with selective frontal lobe damage. There is no natural clinical condition (with the possible exception of frontal lobe dementia in its early stages) that produces damage selective to the frontal lobes, making the study of patients with focal frontal lobe damage a formidable logistical problem. Furthermore, some of the cognitive functions dependent on frontal activity, such as language, are intrinsically human, and for these animal models are of only limited use.

Prior to the 1960s, and even later, frontal lobe patients were often not identified until clinical signs were well advanced; overt neurological signs were often absent, and imaging techniques were either not readily available or inadequately sensitive. In such cases many brain regions beside the frontal lobes were often damaged. It was observations in such patients that formed the basis of the 'frontal lobe syndrome'. Research in patients who had undergone a frontal lobotomy provided some localizing information. However, the disorder leading to the lobotomy, such as schizophrenia, confounded these results to some degree.

The conceptual limitations for the study of frontal lobe function were equally formidable. Behaviorism favored an approach without appeal to mental mechanisms. By the early 1950s, interest in behaviorism had waned. Cognitive scientists, with their emphasis on information processing, turned to the deconstruction of higher cognitive processes into smaller and more tractable units. Even then, psychological constructs related to less modular and more integrative functions remained difficult to define, and even more difficult to relate to the frontal lobes. Most early cognitive scientists were 'functionalists' who thought that the underlying brain mechanisms had no direct bearing on the operations of mental processes.

Significant advances towards delineating the roles of the frontal lobes have been made since the 1980s. Technical advances in structural and functional imaging have shifted the focus from the pure study of mental processes to the study of the elemental mental operations of these processes as they relate to specific neural substrates. Such research has revealed a certain degree of anatomical and functional independence among frontal areas.

NEUROANATOMY

The frontal lobes constitute a quarter to a third of the human brain, occupying the region anterior to

the fissure of Rolando and superior to the Sylvian fissure (Petrides and Pandya, 1994). It includes primary motor cortex (Brodmann area 4: generation of movement), premotor cortex (Brodmann area 6: programming of sequential movements) and prefrontal cortex. This review is limited to the prefrontal cortex. The prefrontal cortex is roughly divisible into distinct regions (Figure 1): dorsolateral (Brodmann areas 9–12, 45–47), orbitofrontal (10–15, 47), superior medial (8, 9, 24, 25) and inferior or ventral medial (9–13, 24, 25, 32). A major anatomical factor underlying the functional role of the prefrontal cortex is its reciprocal connections with virtually all other brain regions. The frontal lobes therefore are capable of integrating and influencing all types of information, cognitive and affective.

As a general observation, hemispheric asymmetry within the frontal lobes is relative, and appears to be related primarily to the dorsolateral regions. The medial frontal lobes are functionally divisible in the superior/inferior plane, with minimal evidence of hemispheric asymmetry.

A THEORETICAL FOUNDATION

A general concept accepted by many is that the frontal lobes are related to supervisory or control processes, regulating more posterior/basal systems that are more automatic in nature. In this view the frontal lobes are necessary when new information is being processed, when processing is complex, or when old information has to be analyzed in new ways. Although this distinction between control and automatic processes may be an adequate overall framework, it is too global. Most researchers today agree that the prefrontal cortex cannot be reduced to the function of a single, unitary cognitive processor. New imaging techniques are helping to separate complex cognitive processes into simpler components, as described below. This idea is compatible with the anatomy of the frontal lobes; they are not homogeneous monolithic structures, but are composed of morphologically distinct areas interconnected with each other and with posterior and basal brain regions to constitute complex anatomical networks.

Processes required for traditional neuropsychological tests of frontal lobe functioning are most often related to the dorsolateral frontal lobes. They are not often represented in the inferior medial frontal lobe. The dorsolateral region is most relevant for staying 'on task' and performing normally when the environment is unstructured. Damage to this inferior medial region does affect the understanding of the emotional consequences of behavior, or more generally, the ability to respond appropriately to reinforcers. Patients may fail to alter their behavior when contingencies change, leading to inappropriate behavior and misinterpretation of others' emotions, and under some circumstances may appear to have an 'acquired sociopathy'.

The superior medial frontal cortex, in contrast, is necessary for performance on many commonly used 'frontal' tasks. A major impairment after superior medial damage, left or right, is a reduction in the ability to initiate and maintain actions or processes. The functional deficits after superior medial damage are often similar but less severe than damage to either dorsolateral frontal cortex, since the superior medial regions provide the activation and drive for the functions of these dorsolateral areas.

COGNITIVE FUNCTIONS OF THE FRONTAL LOBES

Anterior Attentional System

Many – if not all – of the functions described in this section on the role of the frontal lobes in attention have also been considered as 'executive' or 'supervisory' processes. We use 'attention' as the term of choice for several reasons. Most importantly, some type of impaired attentional control is clearly evident in most patients with damaged frontal lobes. There is also a conceptual advantage. An 'anterior attentional system' working in conjunction with a posterior attentional system immediately provides a framework for understanding brain functions as integrated networks. Attention is more easily defined operationally than executive functions, and the fractionation of component processes of attentional tasks has provided a new focus for understanding the complexity of the frontal lobes. This assists in moving theoretical positions of the functions of the frontal lobes away from a general unitary process.

The anterior attentional system can be dissociated into different processes related to distinct brain regions within the frontal lobes (Stuss *et al*., 1994; Knight *et al*., 1995; Stuss and Alexander, 2000). Visual directed attention deficits, as evident in neglect after damage to right parietal regions, are also present after frontal lobe damage. The right frontal lobe in particular provides the exploratory-motor functions of a complex directed attention network. The frontal lobes are important for the stability of gaze fixation. Frontal lesions lessen the benefit of preparatory directional cues on location-based

target detection. Malfunction in the regions of the dorsolateral frontal eye field, supplementary motor area, and perhaps the anterior cingulate gyrus, causes a deficit in the inhibition of inappropriate visuomotor responses to distracting external stimuli, and also results in slowness in moving the eyes voluntarily away from a cue. Patients with frontal lobe damage have difficulty detecting relevant

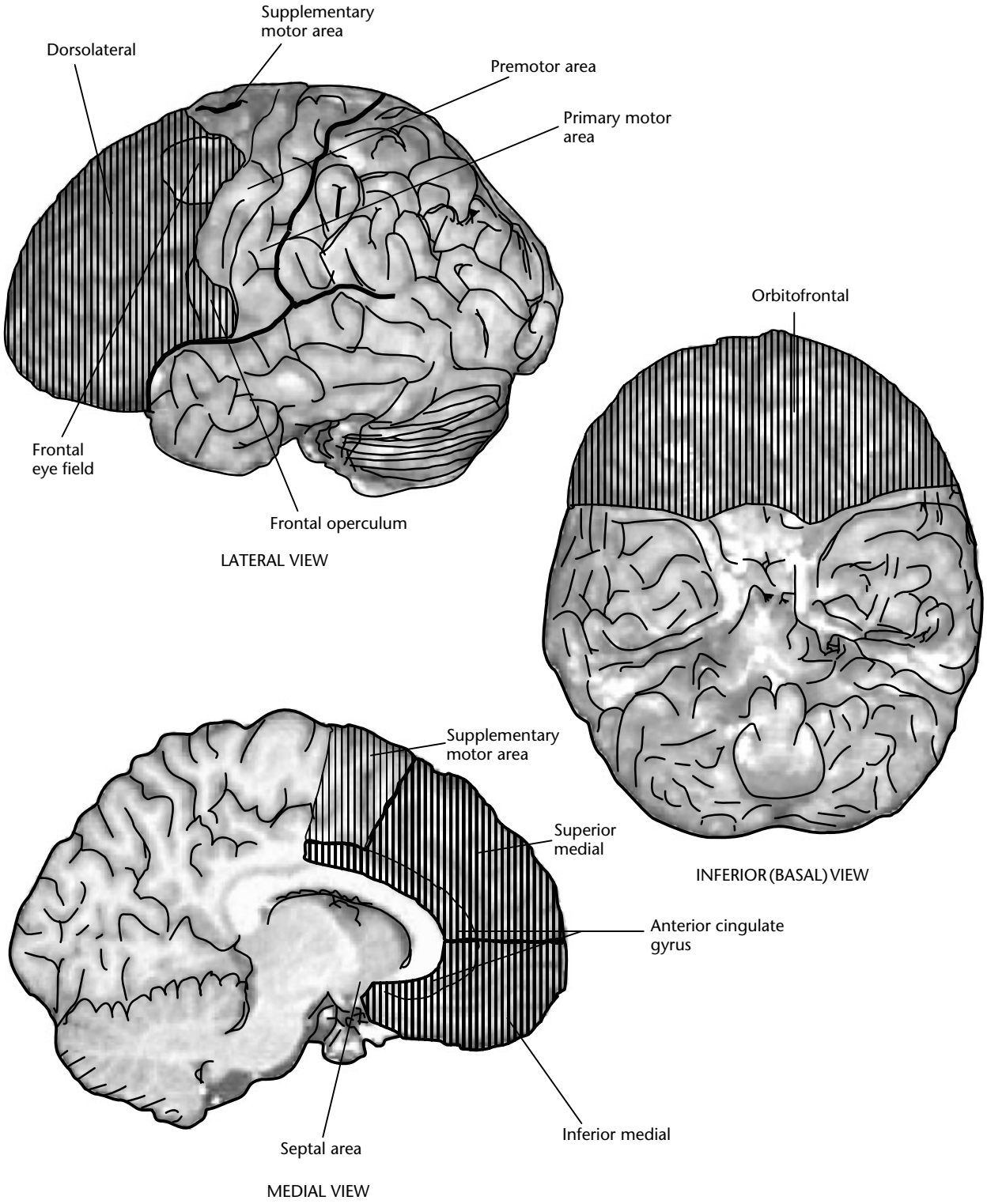


Figure 1. The prefrontal cortex.

stimuli presented on the side contralateral to the lesion. Almost incongruously, electrophysiological responses to unattended stimuli are larger when the stimuli are presented to the side contralateral to the lesion. There is some hemispheric asymmetry in this attentional impairment, with distractibility more noticeable after right dorsolateral frontal lobe lesions.

Patients with frontal lobe damage (particularly on the right) have impairment in sustained attention, defined as the ability to perform a task correctly over a period of time. This appears to be more severe when the task is simple, or stimuli are presented slowly. The frontal lobes are also involved in selective attention, which can be observed in problems of differentiation between novel and familiar stimuli, and in comparisons of attended and unattended information. For example, pathological damage in the anterior cortex diminishes the initial response to novel stimuli. There is also diminished habituation to repeated stimuli.

The frontal lobes also play a role in attentional switching. The Wisconsin Card Sorting Test requires the trial-and-error learning of a criterion by which to sort a deck of cards. After the participant has achieved a defined number of correct responses, the examiner then shifts to a new criterion without warning. Dysfunction in switching categories of response on this type of test is evident after right or left dorsolateral frontal and superior medial, but not inferior medial, lesions. Switching deficits can be observed after inferior medial pathology, however, if the task requires switching initiated in response to affective feedback.

Memory Systems

Damage to the frontal lobes does not normally result in a classic amnesia. The possible exception to this rule involves the septal region, the posterior part of the inferior medial frontal lobe. The septal region is part of the hippocampal and mesial temporal memory system and is often concomitantly damaged with pathological changes to the inferior frontal lobe. Most patients with frontal lobe damage outside the septal region function within normal limits on standard tests of memory such as story recall. However, meta-analysis of focal lesion studies indicates that frontal lobe lesions do affect performance on tasks of free recall, cued recall, and even recognition memory. Functional imaging has identified a role for the frontal lobes in the strategic aspects of semantic and episodic retrieval, and episodic encoding. While the neural networks for these functions have yet to be detailed, there is general agreement that the left frontal lobe is involved in episodic encoding and semantic retrieval, and the right frontal lobe in episodic retrieval.

While the frontal lobes have a role in explicit memory, their function in implicit memory, considered to be more automatic, is less clear. Frontal lobe processes may be necessary in implicit memory, but only under specific conditions which require more strategic search and retrieval, such as word stem (sold—) as opposed to word fragment (s—d-er) completion. This role of the frontal lobes in implicit memory tasks calls into question a simple interpretation of the theoretical dichotomy of control and automatic processes.

Strategic processes in learning and memory

As might be expected from the hypothesized general role of the frontal lobes, the effect of damage to the frontal lobes on memory processes occurs primarily when monitoring and supervision of basic processes is required. This has been called 'working with memory' (Moscovitch and Winocur, 1995) to indicate the influence of the frontal lobe regulatory processes on the more posterior and more automatic memory functions. Following frontal lobe damage, deficits may frequently arise on tasks such as self-ordered pointing, conditional associative learning, metamemory, and memory for source and temporal order. The self-ordered pointing task requires participants to recall which item, out of a defined set of items, they selected on the previous trial in order to avoid selecting the same item consecutively. Conditional associative learning ('if this, then that') requires subjects to form associations between stimuli and responses depending on the context in which they are encountered. Metamemory involves evaluating one's own knowledge to estimate the likelihood that some information is available in memory. In tests of memory for source or temporal order, content or item information is recalled but the details about where this information was encountered (source or context) or the order in which it was presented (temporal order) are lost. To explain the deficits frontal patients exhibit in these complex memory tasks, the processes have to be examined in terms of finer component strategic functions such as subjective organization, self-initiation, monitoring, filtering, and response selection.

Working memory

Working memory tasks involve temporarily holding a small amount of information 'online'. In

research with nonhuman primates, working memory tasks were originally considered to involve only the maintenance of the information over the period of delay to complete a subsequent task when a cue was given. In human research, on the other hand, the concept of working memory incorporated some type of manipulation of information. For example, a task may involve rearranging items during the period of delay in order to list them alphabetically. In both cases, information must be maintained for a brief period in which the information is not available in the environment.

Single cell recording in animals during different kinds of working memory tasks reveal different roles for prefrontal cells. Some cells show an increase in the number of spikes per second firing when a stimulus is presented, others show sustained firing during the delay alone, and still other cells increase firing rate as the end of the delay approaches in anticipation of a response. These cells are organized into local networks to coordinate the three types of information in real time. As in the occipital lobe, there appears to be a cortical columnar organization containing cells with similar roles.

The major role in maintaining information 'online' in the absence of an external stimulus was originally assigned to the lateral prefrontal regions, with regional specificity for the type of information acted upon: location of an object ('where') cells, dorsolateral frontal; identity ('what') cells, ventrolateral frontal. Integration of location and identity information occurred both through organization in overlapping regions of 'what' cells and 'where' cells as well as through 'what and where' cells that respond selectively to the conjunction of location and identity. There is now increasing recognition that this rehearsal or memory maintenance function cannot be completely contained within the prefrontal cortex, but must involve interactions with other regions. For example, systems of frontal and parietal neurons have been reported, with similar cellular recordings and neuroimaging activations in both regions in response to stimuli.

The research in humans has provided a context for a more elaborated position than simple memory maintenance to explain the functional roles related to the firing of specific cells during delays. That is, working memory tests may assess not (or perhaps not just) memory maintenance but executive or higher-level attentional functions. This has also been demonstrated in animal research (Rainer *et al*., 1998). In a task requiring a motor response after a delay if a target stimulus occurred in the same location as a probe stimulus, neurons first fired in the prefrontal cortex, followed by firing in inferior temporal regions responsible for the visual processing of the target. This sequential pattern suggests a 'top down' influence of the frontal lobes on attentional selection in visual processing. There is also context specificity for different cells. Some cells fire during the formation of associations between arbitrary stimuli. Other cells reflect firing in response to the context of the association (i.e., whether a stimulus is currently associated with a reward). Still others show firing that is selective for the type of reward expected within that context. In this more complex formulation there are overlaps between working memory and attentional processes. Such overlaps remain to be fully investigated.

Language

Since the famous case described by Paul Broca, a correlation between frontal lobe damage and language disorders has been accepted. There has been a gradual delineation of the roles of distinct areas within the left and right frontal lobes in communication competence (Alexander *et al*., 1989). The interactive nature of different systems (in the case of language, motor, activation, cognitive, and paralinguistic) for final output are clearly demonstrated in the following examples. For both the left and right frontal lobes, the following descriptions of language disorders will start with the consequences of more posterior frontal region damage and move anteriorly.

Left frontal

Damage to the left inferior motor cortex results in speech (motor) impairment without significant language disorder. If the lesion is considerably more extensive (involving the entire left frontal operculum, posterior sections of the inferior and middle frontal gyri, anterior insula, periventricular and subcortical white matter, as well as putamen, head of the caudate and anterior parietal region), the classic Broca aphasia results with motor (articulation), language (e.g., word finding, agrammatism) and activation (language fluency) deficits. A smaller lesion of inferior frontal operculum, lower motor cortex and adjacent white matter yields a 'small Broca' aphasia, highlighted by dysarthria, initial muteness and subsequent slowness, and language impairment that is similar to but less severe than in Broca aphasia. Damage anterior to the lower motor cortex, superior and anterior to the operculum, or in white matter underlying the operculum, results in normally articulated speech but a verbal output that

is primarily sparse and delayed. Superior medial damage, probably including the supplementary motor area and anterior cingulate gyrus, affects speech activation primarily. When damage involves the left anterior lateral area and polar medial areas, the language deficit is predominantly 'paralinguistic'. Impairments in reasoning, spontaneous utilization of complex syntax, repetition of sentence form, and omissions of elements are often observed. Inferior medial damage does not result in any language or speech activation disturbance.

Right frontal

Damage to the right frontal lobe does not cause aphasia in normal right-handed individuals. Although the localization is not as exact as in the left frontal lobe, damage to the right frontal areas can affect communication in a manner analogous to the left frontal lobe, with motor, cognitive, activation, and formulation impairments. Posterior lateral frontal and subcortical damage alters the affective prosody of language, resulting in flat, monotonous speech. If the pathological change is similar to the small Broca aphasia, deficits appear to occur in the pragmatics of communication such as sustaining attention to communication. Right superior medial damage affects initiation of speech in a manner similar to the left superior medial area. Finally, in more anterior regions, deficits in humor appreciation, sarcasm, critical self-assessment, and establishing a coherent narrative context for communication may result. Confabulation has been described as a consequence of medial damage, particularly on the right.

SOCIAL BEHAVIOR, PERSONALITY, AND SELF-AWARENESS

Defective comportment, a profound apathy, deficient social interactions, and impaired interpersonal skills often constitute the most striking observations in patients with frontal lobe damage, particularly bilateral damage to the orbital frontal (ventral medial) areas. Certain evidence points to a possible preeminent role of the right frontal lobe in many of these functions. The deficits may be such that significant others may consider the individual not to be the same person, as in Harlow's classic description of the case of Phineas Gage – 'he was no longer Gage'. Performance on measures of intelligence and standard neuropsychological tests, including traditional frontal lobe cognitive tests, is often normal. The impairment is most often elicited in real life situations, where the environment is less constrained.

The behavior of some patients with ventral medial frontal lobe damage has been described as an 'acquired sociopathy'. There is indeed a superficial similarity to sociopathic behavior. The patient may have impaired ability to understand and take into account the feelings of others (empathy), and may lack or fail to demonstrate appropriate feelings (sympathy). Behaviors may appear totally self-interested. Patients with frontal lobe damage may be humorless, or conversely show inappropriate jocularity. At other times, there may be impulsive outbursts of anger, or inappropriate, irresponsible and sometimes risky behavior. In frontal lobe patients, however, there is usually no considered decision underlying such behavior, and no deliberate deception. Moreover, the behavior may be demonstrated only under certain conditions such as unstructured environments.

These behaviors can be experimentally demonstrated. Patients with right frontal lobe damage in particular do not appreciate the subtleties of humor as in jokes that depend on a 'twist' at the end, although they can grasp slapstick humor. These same types of patients, particularly if the damage is to inferior medial, find it difficult to take the perspective of others to understand or guide their own behaviors. They may not grasp the implications of any *faux pas* they make. While such functions also require cognitive capacity of different kinds, these deficits do not seem to be reducible to cognitive impairment.

The frontal lobes (with a particular emphasis on the right frontal lobe) are most important for higher levels of awareness. They provide the ability of the individual to use past personal knowledge to understand current behaviors, and to select and guide future responses to integrate the personal self into a social context. This self-reflective ability has been called autonoetic (self-knowing) consciousness, and is considered to be important for episodic memory – that memory related to personal, warm, and emotionally relevant past episodes (Wheeler *et al*., 1997). Frontal lobe patients may have dulled or absent affective responsiveness to past memories. New personal memories may be acquired only via the semantic memory system.

CONCLUSION

Understanding the functions of the frontal lobes remains a formidable task. Future progress will depend on the successful advances in, and integration of, many areas of knowledge and technology: comparative psychology, cognitive neuropsychology, neuropsychiatry, structural and functional

imaging, and neuropsychopharmacology. Discussions will center not on the functions of the frontal lobes but on the role of the frontal lobes in any variety of human behavior.

The ultimate challenge is to use this knowledge for the rehabilitation of patients with damage to the frontal lobes. Efforts in this arena are very much in their infancy. Effective rehabilitation must target the specific deficits resulting from damage to different frontal brain regions. For example, sustained attention deficits after right dorsolateral frontal damage may be assisted by the use of intact left frontal self-verbalization abilities. A combination of 'top down' and 'bottom up' strategies should be employed. In many cases, the use of external environmental support may be partially or completely necessary, taking over the role of the frontal lobes of these patients to compensate for their deficits.

References

Alexander MP, Benson DF and Stuss DT (1989) Frontal lobes and language. *Brain and Language* **37**: 656–691.

Fuster JM (1985) The prefrontal cortex, mediator of cross-temporal contingencies. *Human Neurobiology* **4**: 169–179.

Knight RT, Grabowecky MF and Scabini D (1995) Role of human prefrontal cortex in attention control. In: Jasper HH, Riggio S and Goldman-Rakic PS (eds) *Epilepsy and the Functional Anatomy of the Frontal Lobe*, pp. 21–36. New York, NY: Raven.

Moscovitch M and Winocur G (1995) Frontal lobes, memory, and aging. *Annals of the New York Academy of Sciences* **769**: 119–150.

Petrides M and Pandya DM (1994) Comparative architectonic analysis of the human and macaque frontal cortex. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology* vol. 9, pp. 17–57. Amsterdam, Netherlands: Elsevier.

Rainer G, Asaad WF and Miller EK (1998) Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**: 577–579.

Stuss DT and Alexander MP (2000) Executive functions and the frontal lobes: a conceptual view. *Psychological Research* **63**: 289–298.

Stuss DT and Benson DF (1984) Neuropsychological studies of the frontal lobes. *Psychological Bulletin* **95**: 3–28.

Stuss DT, Eskes G and Foster J (1994) Experimental neuropsychological studies of frontal lobe functions. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology* vol. 9, pp. 149–185. Amsterdam, Netherlands: Elsevier.

Wheeler M, Stuss DT and Tulving E (1997) Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin* **121**: 331–354.

Further Reading

Damasio AR (1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London, Biology* **351**: 1413–1420.

Fuster JM (1997) *The Prefrontal Cortex*, 3rd edn. Philadelphia, PA: Lippincott-Raven.

Goldman-Rakic PS (1996) Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences of the USA* **93**: 13473–13480.

Grafman J, Holyoak KJ and Boller F (eds) (1995) *Structure and Functions of the Human Prefrontal Cortex.* New York: New York Academy of Sciences.

Levin HS, Eisenberg HM and Benton AL (eds) (1991) *Frontal Lobe Function and Dysfunction*. New York, NY: Oxford University Press.

Miller EK (1999) The prefrontal cortex: complex neural properties for complex behavior. *Neuron* **22**: 15–17.

Passingham R (1993) *The Frontal Lobes and Voluntary Action*. New York, NY: Clarendon Press.

Perecman E (ed.) (1987) *The Frontal Lobes Revisited*. New York: IRBN Press.

Stuss DT and Benson DF (1986) *The Frontal Lobes*. New York, NY: Raven Press.

Golgi Staining

Introductory article

Arnold B Scheibel, UCLA Medical Center, Los Angeles, California, USA

CONTENTS

Golgi staining describes a group of methods that depend on the reaction of chromate salts with heavy metals, usually silver or mercury, to reveal the nerve cell body and all of its processes for microscopic study.

INTRODUCTION

The earliest attempts to understand the organization of the nervous system during the first half of the nineteenth century were based on teasing apart small fragments of brain tissue with dissecting needles, either in dilute saline (salt) solutions, or with the addition of small amounts of chromic acid or potassium dichromate solution which tended to weaken the adhesive forces between the cells (today, small amounts of detergent solution would be used). Microscopic studies of these preparations often revealed the cell body and the beginnings of its extensions (dendrites and axons). However, the visibility of these structures was poor since they were virtually as transparent as the aqueous (water-based) medium in which they were suspended.

At the beginning of the 1870s, brain tissue began to be stained (colored) with dyes such as carmine and salt solutions such as gold chloride which made certain parts of the neural components more easily visible. Shortly thereafter, it was found that tissue fixation (inactivation of tissue enzymes and hardening) in dichromate solution followed by sectioning (cutting a number of thin sections of the tissue), staining in hematoxylin and selective removal of excess stain produced a reliable way to visualize all myelinated fibers. However, the fibers became invisible when the myelin sheath was lost as the fiber approached its synaptic (terminal) field. The introduction of aniline stains – chemically basic stains which selectively bind to acidic portions of the cell such as deoxyribose nucleic acid (DNA) and ribonucleic acid (RNA) – also helped reveal a number of structures internal to the cell body and its dendrites. However, there was still no single technique that could reveal the cell body and its entire retinue of dendritic and axonal processes.

DISCOVERY OF THE GOLGI STAIN

The term 'Golgi' has become generic for a group of histological staining methods that rely on exposing neural tissue to solutions of chromate or dichromate salts, followed by heavy metal ions, either silver (e.g. in the form of $AgNO_3$ solution) or mercury (e.g. as $HgCl_2$). The method was first reported by a young Italian physician, Camillo Golgi, in a single paragraph of a short paper he published in 1873 on the cerebral cortex. Golgi was working at that time as resident physician in a hospital for 'incurables' at Abbiategrasso. He is certain to have realized the extraordinary power of the method he described, a method that was to invigorate and restructure our knowledge of brain organization. And yet, we have no clear report of how the revolutionary technique was discovered.

A frequently told, probably apocryphal, version suggests that an overzealous cleaning woman threw some pieces of the young histologist's dichromate-fixed tissue into a refuse jar containing discarded silver nitrate solution. The next morning an angry Golgi rescued the tissue, and while trying to assess the damage, discovered a few nerve cells stained in their entirety. Another, more likely, story suggests that he stumbled on the method while using silver in an attempt to impregnate (stain) the pial (protective) membrane on dichromate-hardened brain tissue.

In either event, he proceeded to describe, in the most complete detail yet achieved, the structure of many kinds of neurons in the central nervous system, their dendrite ensembles and axonal ramifications. Unfortunately, the technique remained cranky and unpredictable as Golgi used it, and

he went on to draw several incorrect conclusions from his material – conclusions that were to start one of the early epic intellectual struggles in neurohistology.

Major improvements in the Golgi method were made by a young Spanish professor of histology, Santiago Ramón y Cajal. He was the first to demonstrate the benefit of using brain tissue from embryonic or young animals, before the myelin sheaths (fatty insulating coverings) of axons were laid down. In addition he was one of the first to add osmic acid to the dichromate fixative, thereby changing Golgi's familiar 'slow method' involving several weeks or months of fixation in dichromate to the 'rapid method' wherein fixation in the osmic–dichromate solution occurred in a few days. The increasingly powerful results of this modified method revolutionized our knowledge of the fine structure of the nervous system, and in its new form attracted histologists from all over Europe and America. So widely was the method used and on such a variety of tissue that one investigator suggested, 'Next we will Golgify potatoes'.

It is not difficult to understand why 'Golgimania' swept through the community of neural investigators towards the end of the nineteenth century. There in all their beauty and delicate complexity were a small number of nerve cells and neuroglia (physiological support cells), each appearing as a black silhouette with all its branches and connections depicted on a pale translucent background. Ramón y Cajal described the appearance as that of 'Chinese ink drawings on transparent Japanese paper'. For the first time one could trace out the course of individual elements (axons or dendrites) from their source, often to their termination. Basic questions about the pathway and nature of connections in the brain could be answered simply by looking, drawing and photographing (Figure 1).

EFFECT OF THE DISCOVERY OF THE GOLGI STAIN ON NEUROSCIENCE

The main strength of the Golgi stain lies in its ability to reveal all the components of the nervous system, including neurons, neuroglia and blood vessels. Moreover it does so in a selective manner, so that only a few of each may be visualized in any one microscopic field. This is advantageous since if all elements were stained, the investigator could see only an inextricable tangle of cells and their processes.

In Golgi's hands the method clearly revealed the differences between axons and dendrites, the multiplicity of nerve cell types and the variety of patterns generated by cell processes. He described the difference between cells with long (projective) axons (Golgi type I) and those with short (local circuit) axons (Golgi type II), a differentiation that continues to be basic to our understanding of how nerve cells are connected into circuits. He erred in claiming that his material proved that all dendritic and axonal processes flowed together or fused, forming a continuous network of processes, or syncytium.

Ramón y Cajal's research was summarized in a great two-volume work *Histology of the Nervous System* now translated into English by the Swansons and published by Oxford University Press. Among his discoveries we may enumerate the mode of termination of axons, the axonal growth cone, development of the concept of polarity and one-way conduction along nerve cell circuits, dendritic spines, the first diagrams of reflex pathways, and the suggestion that learning might depend on the selective strengthening of synapses, an idea that would be reformulated by Donald Hebb in 1949 and become the basis for modern learning theory.

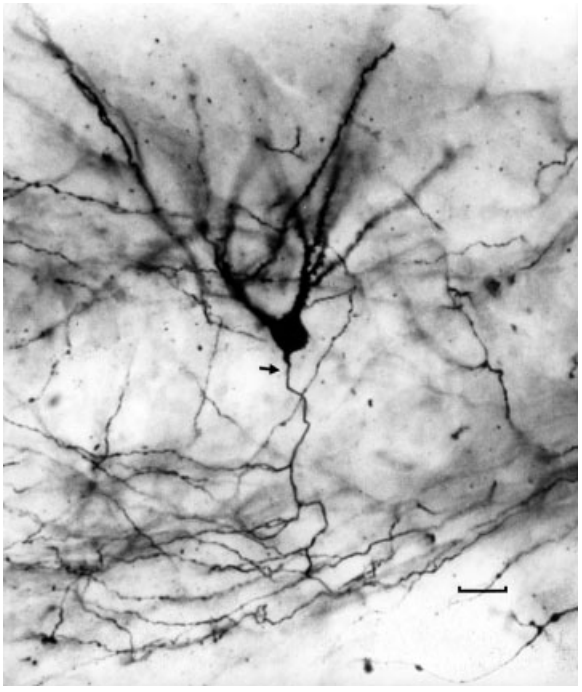


Figure 1. Microphotograph of a small neuron (granule cell) in the dentate gyrus of the hippocampus of a young cat stained by the rapid Golgi method. This is a thick section (100 μm) and several of the dendrites become blurred as they leave the focal plane of the photograph. The axon (arrow) develops a complex series of branchlets (axonal plexus) typical of many local circuit neurons. Bar, 10 μm.

However, the most important single achievement by Ramón y Cajal was undoubtedly his defense and elaboration of the neuron theory. In contradistinction to those (including Golgi) who claimed that dendritic and axonal processes flowed together to form a continuous syncytium or reticulum, Ramón y Cajal became a spokesman for the group of investigators who believed that each neuron was an individual element and that communication from neuron to neuron occurred across small gaps between axons and dendrites or cell bodies, forming junctions called synapses (a functional term originally suggested by the English physiologist Sherrington). Ramón y Cajal made brilliant use of his Golgi material in defending this idea. Since the actual dimensions of the gap or synaptic cleft were of the order of 20–40 nm and therefore below the resolution of the best light microscope, he had to reason from a few special cases, including a special gear-like synapse in the cerebellum. In 1906, Golgi and Ramón y Cajal shared the Nobel prize for physiology or medicine, awarded for their contributions to the fine structure of the nervous system. In his Nobel address Golgi used data from his method to defend the syncytial or reticularist theory; Ramón y Cajal, using material similarly stained by the method of Golgi, defended the neuron theory. (The two men never spoke to each other again.) It took another half-century before the electron microscope with its vastly increased resolution revealed that the neuronists were generally correct and that most interactions between nerve cells occurred across a synaptic cleft. The complex chemistry of neurotransmitters operating across these clefts, and the equally complex study of membrane physiology and receptor-channel structure which provide substrate to these activities, now constitute one of the most active fields of investigation in neuroscience.

It might be added that as in all hotly contested arguments, the truth was finally found to lie somewhere between the two opposing positions. Careful electrophysiological analysis of synaptic interactions supported by electron microscopy has revealed another type of synaptic junction (the gap junction) in which the two apposing neuronal membranes (the presynaptic and postsynaptic elements) are only 2–4 nm apart. Protoplasmic bridges cross this tiny space. Ions flow directly through these minute bridges (connexons) and no chemical neurotransmitters are necessary. The increased rapidity of conduction across these gap junctions fits them for certain types of interactions where flexibility of response seems less important than speed of response.

ROLE OF THE GOLGI STAIN IN MODERN NEUROSCIENCE

Today, the field of neuroscience is rich in histological techniques which seem capable of directly, or indirectly, answering almost any question put to them. There are methods that can trace an axon from the cell body of origin to its terminations or from its termination back to its source; immunohistochemical techniques that can selectively identify by antigen–antibody reaction any one (or several) of literally thousands of chemical substances of which the cell is made up; injection techniques that can outline the individual cell body and all its processes with remarkable clarity; and molecular techniques (*in situ* hybridization) that can identify the very RNA which the cell uses in constituting and maintaining itself – to name only a few. Yet the Golgi technique remains uniquely viable and useful because of the range of its applications.

Since the 1950s Golgi methods have been increasingly used in evaluation of pathological changes in brain tissue, including epilepsy, the senile dementias, degenerative processes such as Parkinson and Huntington diseases, schizophrenia, autism and various retardation syndromes. Refinements in techniques have made it possible to study Golgi stained material at the light microscope level, and then through resectioning, to isolate specific cell groups or even specific synaptic zones for study under the electron microscope.

Of at least equal significance has been the application of Golgi methods to quantitative neurohistology. Through proper selection methods, and increasingly aided by computer technology, it is possible to make quantitative studies of dendrite length, branching pattern and number, and to compare these over several brain areas, or between brains. Similarly, it is possible to follow, quantitatively, the changes in neurons during development and maturation and to describe the variation in dendrite length and complexity as the organism is exposed to enriched (or deprived) sensory input. Literally any perturbation of brain function that involves a physical change in axonal and dendritic size and pattern is available for quantitative description through the use of Golgi-stained tissue.

The enormous power of the Golgi method is essentially a latent one, depending on the interpretative gifts of each particular investigator. The method is almost endlessly adaptable and its constantly broadening usefulness to qualitative and quantitative neurohistology and neuropathology suggests that it will remain relevant in an

increasingly molecular age, as long as investigators have the ingenuity to put new questions to it.

Further Reading

Golgi C (1967) The neuron doctrine – theory and facts. In: *Nobel Lectures: Physiology and Medicine 1901–1902*, pp. 189–217. Amsterdam: Elsevier.

Huttenlocher FR (1974) Dendritic development in neocortex of children with mental defect and infantile spasm. *Neurology* **24**: 203–210.

Ramón y Cajal S (1937) *Recollections of My Life*, translated by E. Hornie Craigie. Cambridge, MA: MIT Press.

Ramón y Cajal S (1967) *The structure and connections of nerve cells*. In: *Nobel Lectures: Physiology and Medicine 1901–1902*, pp. 185–256. Amsterdam: Elsevier.

Ramón y Cajal (1995) *Histology of the Nervous System*, translated by N Swanson and L Swanson, 2 vols. New York: Oxford University Press.

Scheibel ME and Scheibel AB (1978) The methods of Golgi. In: Robertson RT (ed.) *Neuroanatomical Research Techniques*, pp. 89–114. New York, NY: Academic Press.

Hebb Synapses: Modeling of Neuronal Selectivity

Intermediate article

Harel Z Shouval, Brown University, Providence, Rhode Island, USA

CONTENTS

Hebb synapses are modifiable synaptic connections that increase their efficacy when the presynaptic and postsynaptic neurons are coactive. Such a learning rule has appealing computational properties but is not stable. Several modifications have been proposed to stabilize this rule. There is mounting experimental evidence that modifiable synapses exist in brain and form the basis for learning, memory and some aspects of development.

INTRODUCTION

Information processing in the brain is carried out by networks of neurons. Neurons are linked to other neurons by synaptic connections, which convey information from one neuron to another. These synaptic connections are plastic, that is their strength can be altered. The prevailing hypothesis is that the major mechanism underlying learning, memory and many aspects of development is synaptic plasticity. Therefore it is interesting to know what rules govern these changes.

The first influential proposal was made by Donald Hebb in 1949, long before any direct evidence existed about such changes. Hebb's original abstract hypothesis was formed to account for results of experiments such as classical conditioning. This idea was generalized and developed over the years in order to overcome flaws in the original formulation and to account for an increasing number of experimental results. Today, modifiable synapses with modification rules that follow generalizations of Hebb's postulate are often termed Hebb synapses. A combination of experimental and theoretical work in neuroscience is providing mounting evidence that Hebb synapses are indeed the basis for learning memory and many forms of neural development.

HEBB'S ORIGINAL POSTULATE

Hebb originally postulated that:

> When an axon in cell A is near enough to excite cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency in firing B, is increased. (Hebb, 1949)

How can such a rule produce learning? A simple scenario (Figure 1) is the case of a classical conditioning experiment: a dog hears a tone and is then given a food reward; when the animal is given a food reward it salivates; if this is repeated many times the tone alone becomes sufficient to cause salivation. Suppose that neuron A is an auditory neuron that responds to the tone, and neuron B is a neuron that is activated by the food reward and contributes to triggering salivation. According to Hebb's postulate, pairing of the tone, which causes neuron A to fire, and the food reward, which causes neuron B to fire, will cause a strengthening of the synaptic connection from neuron A to B. Eventually activation of neuron A will be sufficient for activating neuron B, even without a food reward. Thus this rule can account for how the tone alone eventually causes salivation, owing to the conditioning procedure.

Mathematically the Hebb rule can be formulated as:

$$\Delta w_i = \eta x_i y \quad (1)$$

where w_i is the strength of the synaptic weight connecting the presynaptic neuron i to the postsynaptic neuron, Δw_i is the change of this synaptic weight in each learning step, x_i is the activity state of the presynaptic neuron i, y is the activity of the postsynaptic neuron and η is a learning-rate constant (Figure 2a). It is simplest to think of the variables x_i and y as binary variables that are equal to

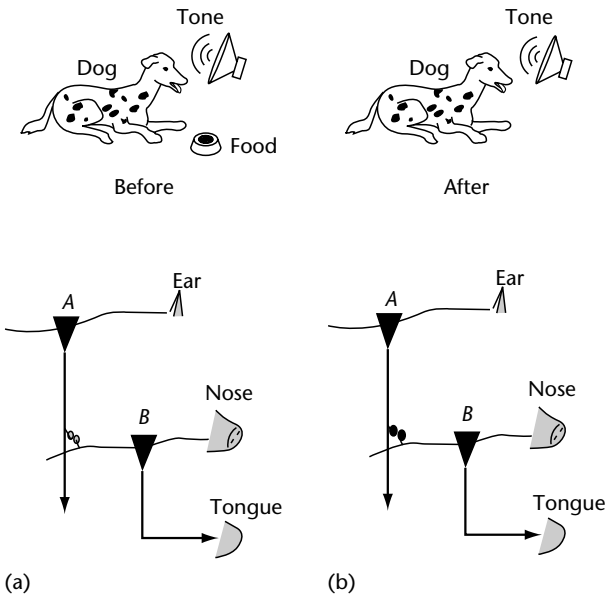


Figure 1. A dog who is presented with food will normally salivate. If the food reward is paired with a tone (a), the dog learns to associate the tone with the food. Eventually (b) the tone will be sufficient to cause the dog to salivate, even in the absence of the food reward. Hebb's postulate can account for this learned association. Neuron A, which responds to the tone, is initially weakly connected to neuron B which responds to the food and drives salivation (a). Pairing the food with the tone will result in an enhancement of the Hebbian synapse connecting neuron A to neuron B (b). Thus activity of auditory neuron A becomes sufficient for activation of neuron B, producing salivation.

1 if the neuron fires, 0 if it does not. It is also possible to think of these variables as continuous variables that correspond to the firing rate of these neurons.

As we have illustrated, the Hebb rule may account for classical conditioning. There are many types of learning tasks, but we will concentrate here on the ability of learning rules to produce selective neurons. Selective neurons are neurons that respond strongly only to a subset of the inputs they receive. Neuronal selectivity is the rule for most cortical neurons; moreover, if neurons responded equally to all inputs they could not perform computations. For example, neurons A in the conditioning example above would be active all the time, hence the dog would salivate all the time.

INSTABILITY OF HEBB'S RULE AND POSSIBLE SOLUTIONS

The Hebb rule, as formulated above, has some fundamental flaws. In particular, under many conditions it would not produce selective neurons. The variables we have used (x_i, y, η) are positive numbers, as a result weights can only grow. Even if the activity of a presynaptic neuron is very weakly correlated with that of a postsynaptic neuron, any small correlation due to spontaneous activity would cause synapses to grow indefinitely. As this is not physiologically plausible, additional constraints must exist. If a simple constraint is imposed, that all synapses have a maximal saturation value, eventually all synapses would attain that maximal value. If all synapses attain their maximal value a neuron will not be selective. One solution is to modify the Hebb rule so it will allow for both decreases and increases in synaptic weights. A simple generalization that produces increases as well as decreases in synaptic weights is the correlation rule of the form:

$$\Delta w_i = \eta(x_i - x_0)(y - y_0) \tag{2}$$

where x_0 and y_0 are constants (Figure 2b). In the absence of saturation limits synaptic weights will either increase or decrease indefinitely. With saturation limits, synaptic weights will converge to either the upper or lower saturation limit. Neurons employing this rule can become selective if the appropriate choice of an environment, constants (x_0, y_0) and saturation limits are chosen (Linsker, 1986).

Synaptic Competition

Another way of preventing the growing weights problem is to assume that the sum of all synaptic weights cannot exceed a constant, possibly because of a restricted amount of some resource (Stent, 1973). Such was the approach used by von der Malsburg (1973) when he used a modified Hebb rule to model the development of orientation selective cells in visual cortex. As a result, when the Hebb rule causes some synapses to increase their synaptic efficacy, the synaptic efficacy of other synapses would decrease. This rule (Figure 2c) implicitly required that the sum of synaptic weights be known to every synapse in the neuron and that after every Hebbian update the value of each synapse would be divided (normalized) by the sum of synaptic weights. A more plausible, local rule has been postulated by Oja (1982). The Oja rule uses a local decay term that causes the sum of the square of the weights $(\Sigma_i\, w_i^2)$ to converge to a constant value. The Oja rule has the form

$$\Delta w_i = \eta(x_i y - w_i y^2) \tag{3}$$

The decay term, which stabilizes this learning rule,

can cause synaptic depression in synapses that exhibit no presynaptic activity: this is referred to as heterosynaptic depression. This approach for preventing the growing weights problem is often called synaptic competition – that is, synapses compete for limited resources (Stent, 1973).

The synaptic competition approach seems to be supported by a set of ocular dominance plasticity experiments. Cells in the visual cortex of cats are mostly binocular; that is, they respond to inputs from both eyes. If one of these eyes is sutured shut during the animal's first months of life (monocular deprivation) the deprived eye rapidly loses its effect on cortical cells (Wiesel and Hubel, 1965). This probably stems from a depression in the synaptic efficacies of synapses connecting cortical cells to afferents from the deprived eye. In contrast, if both eyes are sutured (binocular deprivation) a much slower disconnection occurs (Wiesel and Hubel, 1965). These results are often taken as evidence that synaptic competition does indeed occur.

The Sliding Modification Threshold

An alternative mechanism to prevent the growing weights problem has been proposed by Bienenstock, Cooper and Monro (the BCM rule; Bienenstock *et al*., 1982). This rule (Figure 2) assumes that for moderate levels of postsynaptic activity ($y < \theta_m$) synaptic weights are depressed, whereas for larger levels ($y > \theta_m$) they are potentiated. In both cases presynaptic activity is required for synaptic plasticity, that is both potentiation and depression are homosyaptic. Mathematically the rule takes the form

$$\Delta w_i = \eta x_i \Phi(y, \theta_m) \tag{4}$$

where Φ (Figure 2c), which is a function of the postsynaptic activity (y) and had a zero crossing at θ_m, has both potentiation and depression regions, and θ_m is the modification threshold separating the depression and potentiation regions. This rule alone is not stable; in order to stabilize learning the BCM rule assumes a sliding modification threshold. When synaptic weights and activity levels are increased θ_m will increase as well, faster than the growth of activity. This enhances the depression region and reduces the potentiation region thus stabilizing weight growth. If on the other hand significant synaptic depression occurs resulting in a decrease in synaptic activity, the modification threshold would be lowered. Mathematically the modification threshold could be taken to be the temporal average of some superlinear function of postsynaptic activity, for example $\theta_m = < y^2 >_\tau$,

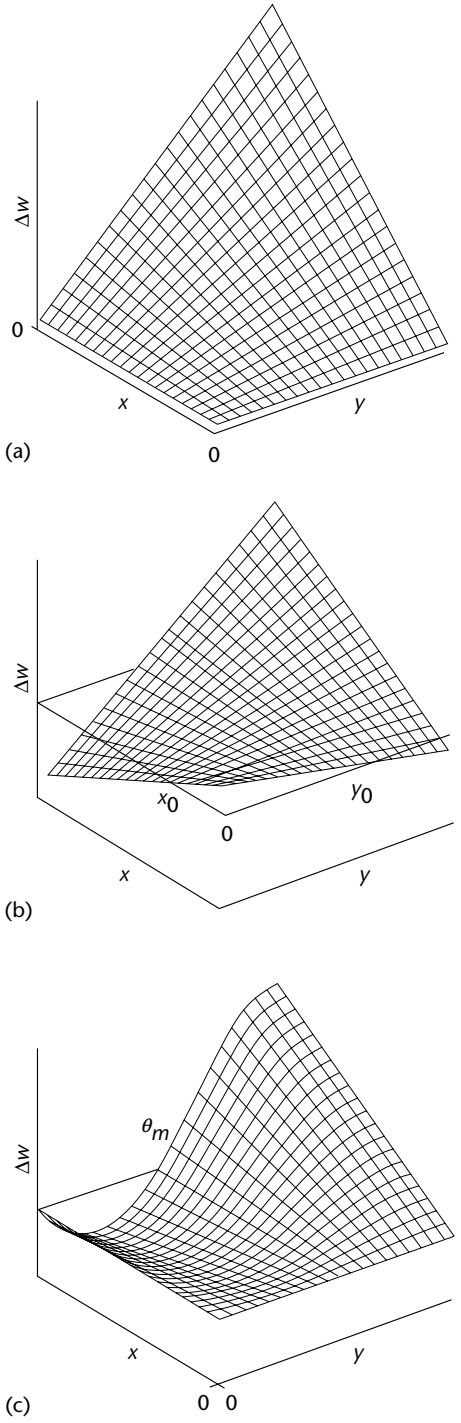


Figure 2. Change in synaptic weights (Δw) due to coincidence of presynaptic activity (x) and postsynaptic activity (y): for the simple Hebb rule (a), described in eqn (1), for the correlational rule (b) described in eqn (2), and for the Bienenstock, Cooper and Monro (BCM) rule (c) described by eqn (4). Note that for the BCM rule, the zero crossing-point (θ_m) is itself a dynamic variable that changes with time.

where < > denotes a temporal average and τ is the time scale of the temporal average.

How can such a modification rule account for paradigms such as monocular and binocular deprivation? A combination of experimental and theoretical work (Rittenhouse *et al*., 1999; Blais *et al*., 1999) has shown that this rule is actually more likely to explain experimental results than rules that depend on synaptic competition.

STATISTICAL PROPERTIES OF HEBBIAN RULES

Different variants of the Hebb rule have different statistical properties. The dynamics of the simple Hebb (eqn (1)) and correlational (eqn (2)) rules are dominated by the second order statistics of their inputs. In the absence of saturation limits the weight vectors become asymptotically parallel to the eigenvector with the largest eigenvalue of a two-point correlation matrix. For the simple Hebb rule (eqn 1), this matrix is simply the cross-correlation matrix of the inputs. The correlational rule (eqn 2) is a somewhat modified version of the input correlation matrix (Linsker, 1986). The Oja rule (eqn (3)) converges to the normalized eigenvector of the correlation matrix. With saturation limits, the stable state is usually dominated by this eigenvector. However, this is a nonlinear system and exact predications are hard to make. If the inputs have zero mean, the first eigenvector is also called the principal component, therefore the Oja rule is often called principal component analysis (PCA). The principal component is a projection in the input space that has the largest variance. For a Gaussian channel, the principal component is the projection that maximizes information transfer.

The BCM rule in contrast depends both on second and third order statistics. For an input space composed of a linearly separable set of vectors the weight vector becomes orthogonal to all vectors but one, thus being maximally selective. For natural input environments it becomes highly selective as well, forming a sparse representation (Blais *et al*., 1998). It can also be understood as a method for finding projections that are highly non-Gaussian (Intrator and Cooper, 1992).

Single cell learning rules can be derived from cost functions that maximize statistical measures such as kurtosis and skew (Blais *et al*., 1998); these can have a roughly Hebbian form. These types of learning rules are closely associated with independent component analysis rules, which try to decompose their inputs into a superposition of independent sources (Hyvrinen, 1999).

EXPERIMENTAL EVIDENCE

There is strong experimental evidence that Hebbian mechanisms exist in cortex. Most research has concentrated on the processes of long-term potentiation (LTP) which is a long-term enhancement in synaptic efficacy due to high-frequency stimulation (Bliss and Lomo, 1973). Homosynaptic long-term depression (LTD) due to low-frequency stimulation has also been observed (Dudek and Bear, 1992). The frequency dependence of these two forms of synaptic plasticity follows qualitatively the form proposed by the BCM theory. Evidence has appeared for existence of the sliding threshold as well (Wang and Wagner, 1999). There is also some evidence of heterosynaptic LTD, as required for synaptic competition (Christie and Abraham, 1992). Recently evidence has appeared that synaptic plasticity depends on the precise timing of presynaptic and postsynaptic spikes (Markram *et al*., 1997). If the presynaptic spike preceds the postsynaptic spike then LTP is induced; if on the other hand the postsynaptic spike comes before the presynaptic spike then LTD is induced.

Pharmacologically blocking LTP and LTD blocks ocular dominance plasticity in visual cortex (Kleinschmidt *et al*., 1987), showing that LTP and LTD are indeed involved in development. Intervention in these processes has also been shown to affect certain types of learning and memory (Morris, 1989).

CONCLUSION

Synapses in the brain exhibit activity-dependent synaptic plasticity. This plasticity is believed to be the basis of learning, memory and many forms of development.

The first influential proposal for a learning rule that can govern these synaptic changes was made by Donald Hebb in 1949; this rule has been the cornerstone for many different theoretical models of synaptic plasticity which have been proposed since then. Hebb's original postulate had several significant problems. Contemporary models of synaptic plasticity overcome some of the problems exhibited by Hebb's original proposal such as the growing weights problem. Different models also have different statistical properties and become selective to different features in the environment.

Plastic synapses have been experimentally characterized in the brain. Their properties are in general agreement with the Hebb postulates as well as with some of the modifications of Hebb's original rule. Further, there is strong evidence that Hebbian plasticity is indeed involved in learning, memory and development.

References

Bienenstock EL, Cooper LN and Munro PW (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* **2**: 32–48.

Blais BS, Intrator N, Shouval H and Cooper LN (1998) Receptive field formation in natural scene environments: comparison of single cell learning rules. *Neural Computation* **10**(7): 1797–1813.

Blais BS, Shouval HZ and Cooper LN (1999) The role of presynaptic activity in monocular deprivation: comparison of homosynaptic and heterosynaptic mechanisms. *Proceedings of the National Academy of Sciences of the USA* **96**: 1083–1087.

Bliss TVP and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. *Journal of Physiology, London*, **232**: 331–356.

Christie BR and Abraham WC (1992) NMDA-dependent heterosynaptic long-term depression in deniate gyrus of anesthetized rats. *Synapse* **10**: 1–6.

Dudek SM and Bear MF (1992) Homosynaptic long-term depression in area CA1 of hippocampus and the effects on NMDA receptor blockade. *Proceedings of the National Academy of Sciences USA* **89**: 4363–4367.

Hebb DO (1949) *The Organization of Behavior; A Neuropsychological Theory*. New York: John Wiley.

Hyvrinen AA (1999) Survey on independent component analysis. *Neural Computing Surveys* **2**: 94–128.

Intrator N and Cooper LN (1992) Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability connections. *Neural Networks* **5**: 3–17.

Kleinschmidt A, Bear MF and Singer W (1987) Blockade of NMDA receptors disrupts experience-dependent plasticity of kitten striate cortex. *Science* **238**: 355–358.

Linsker R (1986) From basic network principles to neural architecture: emergence of orientation selective cells. *Proceedings of the National Academy of Sciences USA* **83**: 7508–7512, 8390–8394, 8779–8783.

Markram H, Lübke J, Frotscher M and Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science* **275**: 213–215.

Morris RG (1989) Synaptic plasticity and learning: selective impairment of learning rats and blockade of long-term potentiation in vivo by the *N*-methyl-D-aspartate receptor antagonist AP5. *Journal of Neuroscience* **9**: 3040–3057.

Oja E (1982) A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology* **15**: 267–273.

Rittenhouse CD, Shouval HZ, Paradiso MA and Bear MF (1999) Evidence that monocular deprivation induces homosynaptic long-term depression in visual cortex. *Nature* **397**: 347–350.

Stent G (1973) A physiological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences USA* **70**: 997.

Von der Malsburg C (1973) Self-organization of orientation sensitive cells in striate cortex. *Kybernetik* **14**: 85–100.

Wang H and Wagner JJ (1999) Priming-induced shift in synaptic plasticity in the rat hippocampus. *Journal of Neurophysiology* **82**(4): 2024–2028.

Wiesel TN and Hubel DH (1965) Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *Journal of Neurophysiology* **28**: 1029–1040.

Further Reading

Abraham WC and Bear MF (1996) Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neuroscience* **19**: 126–130.

Bear MF and Malenka RC (1994) Synaptic plasticity: LTP and LTD. *Current Opinions in Neurobiology* **4**: 389–399.

Bi G and Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* **24**: 139–166.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience*. MIT Press.

Johnston D and Wu S-M (1995) *Foundations of Cellular Neurophysiology*. MIT Press.

Shouval HZ, Bear MF and Cooper LN (2002) A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of The National Academy of Sciences of the USA* **99**: 10831–10836.

Hebbian Cell Assemblies

Intermediate article

Yves Frégnac, Unité de Neurosciences Intégratives et Computationnelles, Gif-sur-Yvette, France

CONTENTS

A Hebbian cell assembly is defined as a collective dynamic activity process reverberating in a closed set of recurrent connections, which are part of a larger association network.

INTRODUCTION

This article summarizes the historical foundations and subsequent theoretical and experimental elaboration of a simple activity-dependent algorithm for synaptic plasticity proposed by Hebb in 1949. This algorithm is based on the assumption that the efficacy of synaptic transmission is regulated in a predictive manner by the temporal correlation between pre- and postsynaptic activities. Validation of Hebb's postulate has been sought at different levels of neural integration. Numerous experimental reports show that the same phenomenological rule can be used to predict activity-dependent changes in synaptic efficacy in a variety of neural structures and species, ranging from invertebrate sensorimotor ganglia to vertebrate neocortex (Brown *et al*., 1990; Frégnac *et al*., 1988; Frégnac, 2002).

Hebb's postulate appears to be valid both for individual synapses and for the global dynamics of functional coupling between neurons embedded within densely connected networks. Irrespective of the presence of direct anatomical contacts, its predictive power holds, under certain additional constraints, for functional links measured solely on the basis of spike-based correlations (Ahissar *et al*., 1992). It has also been suggested that similar Hebbian forms of plasticity, acting on different time scales, participate in the transient build-up of mental representations and illusory percepts, as well as in long-lasting associative adaptation in behavior (Von der Malsburg, 1981; Frégnac and Bienenstock, 1998).

PRE-HEBBIAN THEORIES

The most obvious feature of synaptic transmission is the causal relationship between the presynaptic action potential and the amplitude or probability of response in the postsynaptic cell. Philosophers of antiquity theorized how causal relationships between external events could be established in the brain, and they pointed out the necessity of repeating sequences of activation in order for mental links to emerge:

> Acts of recollection, as they occur in experience, are due to the fact that one movement has by nature another that succeeds it in regular order. If this order be necessary, whenever a subject experiences the former of two movements thus connected, he will (invariably) experience the latter. (Aristotle, *De Memoria et Reminiscentia*)

This view, which was also apparent in the writings of empiricist philosophers such as John Locke in the seventeenth century, was not translated into more specific brain-activity-related processes until the end of the nineteenth century. The foundations of association theories in the brain can be traced back to 1890. According to William James, the 'spiritual father' of the American school of experimental psychology, the brain is not constructed to think abstractly, but to ensure a permanent fit between the perception and actions of the organism in relation to its environment. James suggested that the adaptive capacities of the human brain depend on mechanistic laws of association that operate under the guidance of central neural structures such as the cerebral cortex in higher vertebrates:

> When two elementary brain processes have been active together or in immediate succession, one of them, on re-occurring, tends to propagate its excitement into the other. (James, 1890, p. 256)

James imagined a local integration process where one point in the brain had properties of summation and adaptation very similar to those that we now know regulate neuronal integration of synaptic potentials. The memory trace of association between events was thought to depend on competitive processes which could interfere with the result of previous binding operations:

> The amount of activity at any given point in the brain-cortex is the sum of the tendencies of all other points to discharge into it, such tendencies being proportionate (1) to the number of times the excitement of each other point may have accompanied that of the point in question, (2) to the intensity of such excitements and (3) to the absence of any rival point functionally disconnected with the first point, into which the discharges might be diverted.…When two neurons are simultaneously active, their exchanges of dynamic influxes are more intense than the influxes they send to resting regions. (James, 1890, p. 257)

These concepts of association are immediately applicable to an understanding of behavioral learning. Pavlovian classical conditioning is an example of association between events that are sequentially phased. The presentation of a conditioned stimulus (CS) is followed, at a fixed inter-stimulus interval (ISI), by the application of an unconditioned stimulus (US). The repeated presentation of these paired events results in the development of a conditioned response (CR) that is revealed *de novo* after conditioning when the CS is reapplied in the absence of the US. In general, the time course of the conditioned response predicts the temporal occurrence of the absent US or reward, as if it had been applied. To account for the new behavioral repertoire acquired through classical conditioning, Pavlov introduced the analogy of a circuit closure ('Zamykatielnost') – it was as if the repeated CS–US associations had switched on a broken or previously dormant connection.

The concept of associative conditioning was developed in cellular terms by Jerzy Konorski (1948), a contemporary of Donald Hebb, who assumed that 'before a conditioned reflex is established, there exist potential interneuronic connexions, directed from the emitting neuron, or neural center', representing the conditioned stimulus, 'to the receiving neuron or center' whose activity is forced by the reinforcing stimulus (or US). Extending the suggestion originally made by Cornelius Ariëns Kappers in 1921, that new connections could be established by the simultaneous activation of the neural centers fed by the CS and the US, Konorski formulated a detailed law of association transforming these potential interconnections into 'actual' ones:

> When the excitation of a given center is synchronous with the rise of excitation in another center, conditioned excitatory connexions are formed from the first of these centers to the latter. …Conversely, the excitation of a given cortical center synchronized with the fall of excitation in another center results in the establishment of inhibitory connexions between these two centers. (Konorski, 1948, p. 106)

A second field of application of association theories is the transient formation of mental representations during perception or dreams. Here the effect produced by repeated associations is not restricted to sequences of external events, but extends to autonomous activity of the brain. Rather than being transformed into a long-lasting 'mnesic' form, the trace of the association is seen as a reversible facilitation, promoting neural links over the time required for the establishment of the percept (a few hundred milliseconds).

In his seminal essay, 'Le rêve: étude psychologique, philosophique et littéraire', the visionary French neuroanatomist Yves Delage sought to establish the physiological basis of the psychological processes involved in the genesis of mental images or thoughts (Delage, 1919). Impressed by the work of Louis Lapicque on chronaxy and the finding of a vibratory mode shared by both the muscle and the innervating motoneuron, Delage postulated that each cortical neuron exhibits an intrinsic and characteristic periodicity. The relative diversity of intrinsic frequencies adopted by possible future functional partners ('heterochrony') would be dynamically restructured during perception and generate transient and highly synchronized states ('parachrony') among the activated members of the functional assembly.

> Every modification engraved in the neuron's vibratory mode as a result of its co-action with others leaves a trace that is more or less permanent in the vibratory mode resulting from its hereditary structure and from the effects of its previous co-actions. Thus its current vibratory mode reflects the entire history of its previous participations in diverse representations. (translated from Delage, 1919, pp. 118–119)

Unfortunately, we have no way of establishing how influential these different proposals were in the formulation of Hebb's neurophysiological postulate (Hebb, 1949). Nevertheless, it seems important to recognize that this general plasticity rule, which was to become a winner in the growing field of neurobiology, is the composite echo of a diversity of concepts in the field of learning and perception, already shared by many preceding and contemporary scientists.

A NEUROPHYSIOLOGICAL POSTULATE TO BUILD ASSEMBLIES

Interestingly, the major contribution of Donald Hebb as a scientist of the brain is probably not what modern neurobiology now remembers. His physiological association principle was in fact just one of several keystones incorporated in a multilevel model of cerebral functioning during perceptual and learning processes. The primary aim of Hebb's model was to explain the waxing and waning of psychological constructs during a stream of thoughts. The concept of a 'cellular assembly', pivotal to his theory, designates an activity process which reverberates in 'a set of closed pathways'. The postulate of Hebbian synapses was introduced as one of the possible ways to reinforce functional coupling between co-active cells and thus of growing assemblies. Similar hypotheses were developed which allowed cognitive events to be linked to their recall, in the form of temporally organized series of activations of assemblies. Hebb referred to this binding process as a 'phase sequence'. Thus Hebb's neurophysiological postulate, articulated at a late stage in his famous book, *The Organization of Behavior*, appears simply as a putative low-level biophysical mechanism for establishing the persistence of activity among assemblies.

The wording of Hebb's postulate is impressive in its globality and polymorphism, which may explain its predictive power when applied to highly diverse but specific biological systems:

> When an axon of cell A is near enough to excite cell B, and repeatedly or consistently takes part in firing it, some growth process or metabolic change takes part in one or both cells such that A's efficiency, as one of the cells firing B, is increased. (Hebb, 1949, p. 62)

The formulation of Hebb's postulate requires both temporal coincidence and spatial convergence of one neuron on to another, supporting a causal relationship between the afferent activity and the postsynaptic spike. It provides a specific prediction, namely that a period of maintained positive temporal correlation between pre- and postsynaptic activity will lead to an increase in the efficacy of synaptic transmission. Hebb did not indicate whether the modifications responsible for this decrease in 'synaptic resistance' were presynaptic, postsynaptic or both, nor did he describe the biophysical substrate responsible for the modification, leaving an open choice between 'metabolic change' and 'oriented growth'.

Historically, Hebb's postulate referred exclusively to excitatory synapses, probably of the axosomatic or axodendritic type, since at that time inhibitory connections were almost unknown (but see the behavioral generalization made by Konorski, 1948). A symmetrical if not synergetic version of Hebb's postulate was proposed much later for the case of inhibitory synapses (Stent, 1973), where functional coupling can be increased by reducing the strength of inhibitory synapses activated at the same time as action potentials are fired in the postsynaptic cell:

> When the presynaptic axon of cell A repeatedly and persistently fails to inhibit the postsynaptic cell B while cell B is firing under the influence of other presynaptic axons, metabolic changes take place in one or both cells such that A's efficiency, as one of the cells inhibiting B, is decreased. (Stent, 1973, p. 1000)

The use of Hebbian plasticity rules in formal networks does not always follow this distinction between adaptive excitatory synapses and fixed inhibitory weights. Some models introduce inhibitory plasticity as well, and it has been demonstrated analytically that negative weights are required for optimal mapping and memory capacity in associative memories. Thus the 'effective' gain between input and output, defined at an ideal Hebbian synapse, should be allowed to vary between positive and negative boundaries, depending on whether the net effect induced by the input is excitatory or inhibitory. This could be done by assuming that a biological synapse may change in sign, or that the gain of the ideal Hebbian junction reflects the balanced influence of an elementary circuit, composed of dual parallel monosynaptic excitatory and polysynaptic inhibitory connexions (Figure 1).

The possibility of a change in sign may seem unlikely for any single particular synapse, since it violates Dale's principle, according to which a neuron releases the same transmitter at all of its processes. However, unexpected support was found in *Hermissenda crassicornis*, a nudibranch mollusk, during Pavlovian conditioning. Here associative pairing between visual input and iontophoresis of a $GABA_B$ agonist at the axon terminal of a photoreceptor cell transforms the GABA-induced hyperpolarization into a depolarizing postsynaptic potential. Some theoretical studies also proposed that synaptic gain could change sign during development (Bienenstock *et al.*, 1982), and this suggestion was later found to apply to inhibitory neocortical or hippocampal circuits. It has been shown that, early in development, GABA-receptor activation promotes postsynaptic depolarization at a stage when the chloride equilibrium potential (E_{Cl^-}) is less negative than the resting potential.

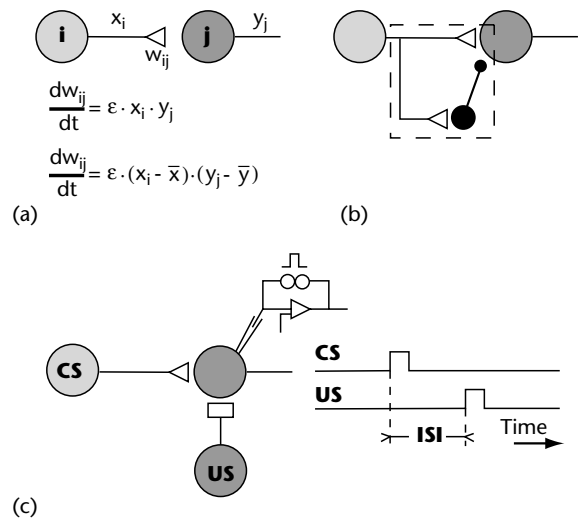


Figure 1. The multiple identities of a Hebbian synapse. (a) Identified monosynaptic connection between a presynaptic axon (open triangle) and a postsynaptic target cell (dark gray). The Hebbian algorithm postulates that the change in synaptic efficacy is given by the product (or logical AND) of pre- and postsynaptic activities at any point in time. The covariance algorithm postulates that the change in synaptic efficacy is given by the product of the differences between the instantaneous pre- and postsynaptic activities from their mean. (b) A dual presynaptic excitatory/inhibitory circuit (dotted box) equivalent to an ideal Hebbian synapse, the gain of which varies between negative and positive boundary values. Open circles and triangles denote excitatory input cell and synapse, respectively. Solid symbols denote inhibitory cell and synapse. (c) A cellular analog of classical conditioning. The Hebbian synapse transmits the neural information fed by the conditioned stimulus (CS). The unconditioned stimulus (US) activates the postsynaptic cell in an all-or-nothing fashion through a non-modifiable synapse or through depolarizing current injection directly applied by the experimenter in the postsynaptic cell (dark gray). The conditioned response stems from the repeated association of the CS input (light gray) followed by the US input activation.

The more classical hyperpolarizing or shunting effect of inhibition is observed at a later stage, when E_{Cl^-} becomes more negative than or equal to the resting membrane potential.

Hebb did not envisage specific forms of use-dependent weakening of synaptic gain, nor did he propose rules to explain plasticity occurring at synapses impinging on the same postsynaptic cell but remaining silent at the time of the association (heterosynaptic plasticity). However, the first cyberneticians who tried to build computational machines incorporating Hebb's principle were faced with the inherent instability of positive feedback and the divergence of synaptic weights. The consequence of this problem of synaptic weight saturation is that any network theory of memory storage based on Hebb's postulate, where associations will be overlaid on common sets of synapses, requires bidirectional, reversible changes in synaptic gain. However, this symmetry in the phenomenological necessity of 'what comes up must eventually come down' should not be taken literally at the substrate level, and there is continuing debate as to whether the molecular processes underlying homosynaptic forms of potentiation and depression are inversely related.

POST-HEBBIAN THEORIES: DYNAMIC BINDING OF ASSEMBLIES AND GRAPH MATCHING

The two main fields of application of Hebb's postulate have been the formation of memory traces during learning, and the self-organization of neural assemblies during development. However, Hebb's theory also addressed perception. Although the dynamic nature of the binding process which is required to form a transient percept had already been predicted by Delage (1919), Peter Milner was probably the first theoretician to propose explicit rules for the composition of assemblies. The repeated activation of a given cell assembly would reinforce synaptic links within this assembly, and in addition would 'prime' a restricted number of cells, allowing future binding and thus composition with other associative processes. The latent labeled (or 'tagged') synapses would remain transiently eligible for further potentiation by the contiguous firing of other assemblies. Repeated sequential activation would reinforce the primed connections, which would then become an integral part of the next active assembly, thereby resulting in second-order associations. Thus their firing would allow the recall of a complete 'phase sequence'. Twenty years later, similar ideas were reworded with the addition of a 'glue' mechanism, namely temporal synchrony, to bind elementary representations into a cognitive 'whole':

> If adjacent, or nearly adjacent, cells interact when excited, in such a way as to synchronize and perhaps intensify each other's activity, this could provide the unifying characteristics that tie the elements of a figure together. At subsequent levels of the pathway, impulses from the cells fired by one whole would arrive as synchronous volleys, whereas impulses from different figures would have a random temporal relationship to each other. (Milner, 1974, p. 526)

A related formalism was revived 7 years later, which postulated fast binding processes during visual shape recognition that depend on ongoing dynamic changes in the temporal correlation of firing between co-stimulated cells. In 1981, Von der Malsburg proposed a 'correlation theory of brain function' which offers a new field of validation for Hebbian associative theories, on a millisecond timescale rather than on the classic developmental scale. The internal representation of mental objects, used in particular in sensory coding and form recognition, here relies on the topological properties of the connection patterns or 'graphs', whose nodes would be formed by the neurons which are co-active at a given point in time. The architecture of the graph, which defines the relational coding within the assembly, would be independent of the physical location of the activated neurons in the brain. According to this view, the representation of a generic object, such as a 'chair', is neither the activation of a grandmother cell nor the activation of a static population code. Rather, it consists of the activation of dynamic links between representation units at distinct levels of a hierarchy, in accordance with composition rules defined at each level. Invariance of the percept with, for example, position or orientation of the 'chair' in the visual field, will be signaled by a homeomorphic match between the graphs representing the various instances of the 'chair' (Frégnac and Bienenstock, 1998).

The neural counterparts of cognitive entities are implemented in the form of high-order statistics of multidimensional spiking processes established over the full network, providing further substrate for compositionality. The theoretical work fostered by Von der Malsburg and colleagues showed that the combined use of 'fast' and reversible Hebbian-like synaptic changes and 'slow' Hebbian plasticity provides interesting properties when applied to a specific type of assembly called 'synfire chains'. The hypothesis of chains of synchrony, first introduced by Moshe Abeles (1982, 1991) and further developed by others (Bienenstock, 1994), implies that at fixed points in time, ensembles of neurons that may be physically far apart in the neural tissue fire together. Re-ordering active units as a function of time, rather than as a function of space or node location in the full network, reveals the sequential recruitment of assemblies firing in bursts.

In their initial version, synfire chains consisted of discrete sequences of pools of neurons in which feedforward conduction times were identical for all connections, from one pool to the next (Figure 2). This is the simplest architecture that guarantees the preservation of synchrony (Abeles, 1991). Simultaneous recordings of cortical single-unit activity in behaving monkeys indeed show that a significant proportion of activity correlated with a specific cognitive task can be described as a wave of synchrony relayed between sets of co-active neurons, with delays ranging from a few to several hundreds of milliseconds (Vaadia *et al.*, 1995). However, debate continues concerning the validity of the statistical analysis (Oram *et al.*, 1999). Bienenstock and colleagues extended this concept to 'synfire braids' (Bienenstock, 1994). They assumed that transient binding will be promoted between polysynaptic circuits whose individual connections have non-uniform conduction times but which, by their composition, satisfy some type of transitivity rule in transmission (Figures 2b and 3a). The reinforced connections linking one set of neurons that are co-active at time t_0 to another set which will become co-active at a later point in time $(t_0 + \Delta t)$ will be those for which the global transmission delay satisfies the same timing constraint Δt. This global delay is defined along any possible polysynaptic path joining the two synfire ensembles by adding each individual delay encountered, and Hebbian processes select out only links for which the global delay is kept constant and equal to Δt. This mechanism, and the possibility that the same neuron could participate in the same synfire chain at different times, allow the graph topology to extend beyond a pure feedforward structure and generate a complex hierarchy of representations (Bienenstock, 1994; see Bi and Poo, 2001 for experimental validation).

This 'Lego-like' interlacing of dynamic assemblies provides neuronal expression of abstract compositional models of cognition based on dynamic binding operations between symbols that are located at various levels of a representational hierarchy (Von der Malsburg, 1981; Hummel and Biederman, 1992; Bienenstock, 1994; Frégnac and Bienenstock, 1998). Relational binding can be specified in a syntactic way (e.g. in terms of perceptual grammar between geometric features for visual object representation and in terms of linguistic features and composition for speech representation) (Shastri and Ajjanagadde, 1993) (Figure 3b).

BREAKING THE SYMMETRY: CAUSALITY RATHER THAN SYNCHRONY

Despite the indiscriminate use of Hebb's rule by experimenters and modelers, it is surprising to note that the temporal specificity of the association

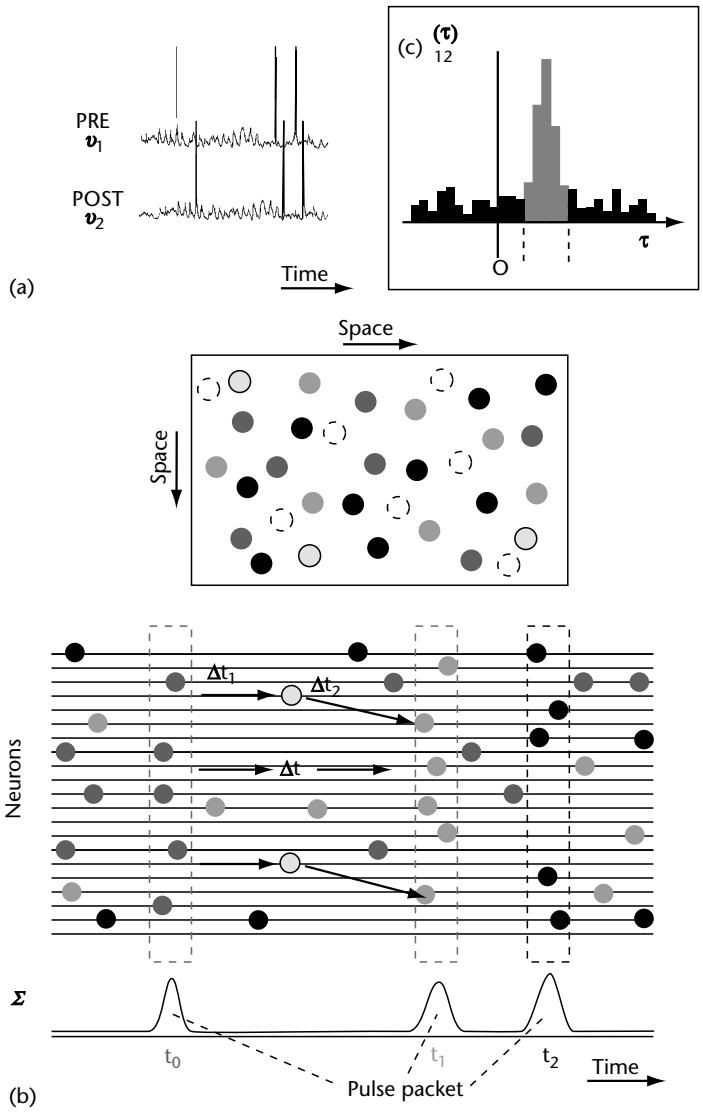
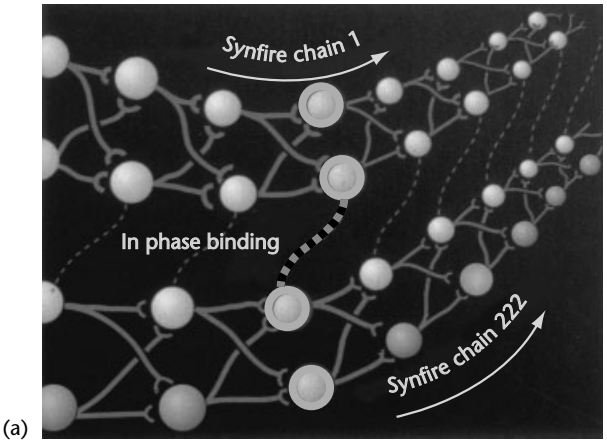
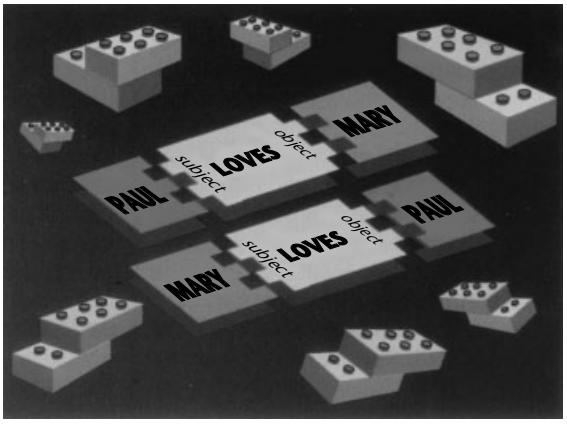


Figure 2. [*Figure is also reproduced in color section*.] Hebbian assemblies and synfire chains. (a) Functional link established between the spiking activity of two cells recorded simultaneously in the same network. In the right inset, the light-gray shaded area in the cross-correlogram indicates the probability that cell 2 (pre) fires at a given temporal delay $\Delta\tau$ after cell 1, and allows the quantification of the 'effective' gain of this functional connection. The effective gain is determined both by the strength of the direct connections between the two units and by the statistics of the firing of the other units influencing the activity of each cell. (b and c) Network dynamics and synfire chains. (b) The physical location of units that are co-active at time t_0, t_1 and t_2 is labeled in red, blue and black, respectively. Open dotted circles denote cells that are randomly active and out of synchrony. (c) Chronogram of the activity of the same units irrespective of their physical location in the network. Transitivity rules constrain the selection of transmission delays corresponding to the polysynaptic connections (via relay cells, shown in yellow) between each set of co-active units. The lower line shows at each point in time the dispersion of activity within each synchrony packet, which remains constant along the chain.

principle itself may have been initially misread. Most applications of Hebbian theories based on pairing protocols highlighted the importance of co-activity, ignoring the interval of a few milliseconds that separates a presynaptic spike from the triggering of postsynaptic activation. For those applications, no temporal ordering was required between pre- and postsynaptic activation. This temporal symmetry may not apply strictly to tetanization protocols where a high-frequency stimulation train is the primary causal event inducing the long-term synaptic change, namely long-term potentiation (LTP) or long-term depression (LTD). However, it is important to observe that associative



(a)



(b)

Figure 3. [*Figure is also reproduced in color section.*] Abstract compositionality in Hebbian assemblies. The same formalism as in Figure 2, applied to neural networks underlying logogenesis (adapted from Bienenstock (1991) with permission). (a) Composition of two synfire chains by specific sets of synapses or neurons (highlighted in pink) interfacing the two activity chains. (b) The 'Lego-brick' analogy is applied to language, where words encoded by individual synfire chains are composed according to syntactic binding rules.

conditioning of a weaker test pathway was usually obtained by applying the test input in phase or in synchrony with the postsynaptic depolarization induced by the tetanized pathway. The most obvious cases of associative learning that depart from co-activity rules are those which require associations between neural events separated by long delays, such as the optimal inter-stimulus intervals (of the order of seconds) of association in classical conditioning (Figure 1c), or between the sample and choice periods in delayed-matching-to-sample comparison tasks (Amit, 1995).

However, the wording of Hebb's original synaptic plasticity principle implies a temporal constraint in the millisecond range – that is, it dictates that presynaptic activity should precede the spike initiation in the postsynaptic element to which it contributes in a causal way ('A's efficiency, as one of the cells firing B'). In this respect, temporally asymmetrical Hebbian and anti-Hebbian rules (Figure 4), which were only introduced in recent years, are in closer agreement with the original concept. Recent research, in some cases using dual patch recordings *in vitro*, in preparations as diverse as cultured hippocampal networks, the developing retinotectal system of frog, the adult electrosensory lobe of electric fish and the sensory neocortex of rat, suggests an even tighter temporal contingency rule (of the order of 10 ms). The temporal order between the onset of the postsynaptic subthreshold potential reflecting the arrival of the presynaptic spike and the postsynaptic spike retropropagating in the dendrite determines whether potentiation or depression occurs (Bi and Poo, 2001; Frégnac, 2002).

The obvious consequence is that models which incorporate spike-timing-dependent plasticity (STDP) rules account most accurately for the emergence of causal chains within neuronal assemblies, and best support phase-sequence learning (Frégnac, 2002). Nevertheless, it is possible that both synchrony and causality are required to promote reverberating assemblies rather than the simple build-up of open-ended chains of feedforward activity, as exemplified by 'synfire chains'. It is plausible that STDP reinforces the progressive establishment of a transmission mode through 'pulse packets', since the asymmetrical nature of this plasticity rule will indirectly control the jitter that could be observed in the timing spread within each pulse packet. Indeed, the STDP rule for selecting connections may be applied to any feedforward pathway – direct or polysynaptic – jumping from one set of co-active units to any other belonging to the same synfire chain, independently of their respective rank with regard to temporal activation. Consequently, the same level of temporal precision in the selection process of individual transmission delays may be preserved all along the chain (within the millisecond range), without the dilution of within-packet synchrony that one might expect from a purely sequential structure.

CONCLUSION

Although synfire chains and graph matching appear to take us far beyond the simple application of local activity-dependent rules of synaptic regulation, it is likely that Donald Hebb would still recognize the heritage of his theory of assemblies

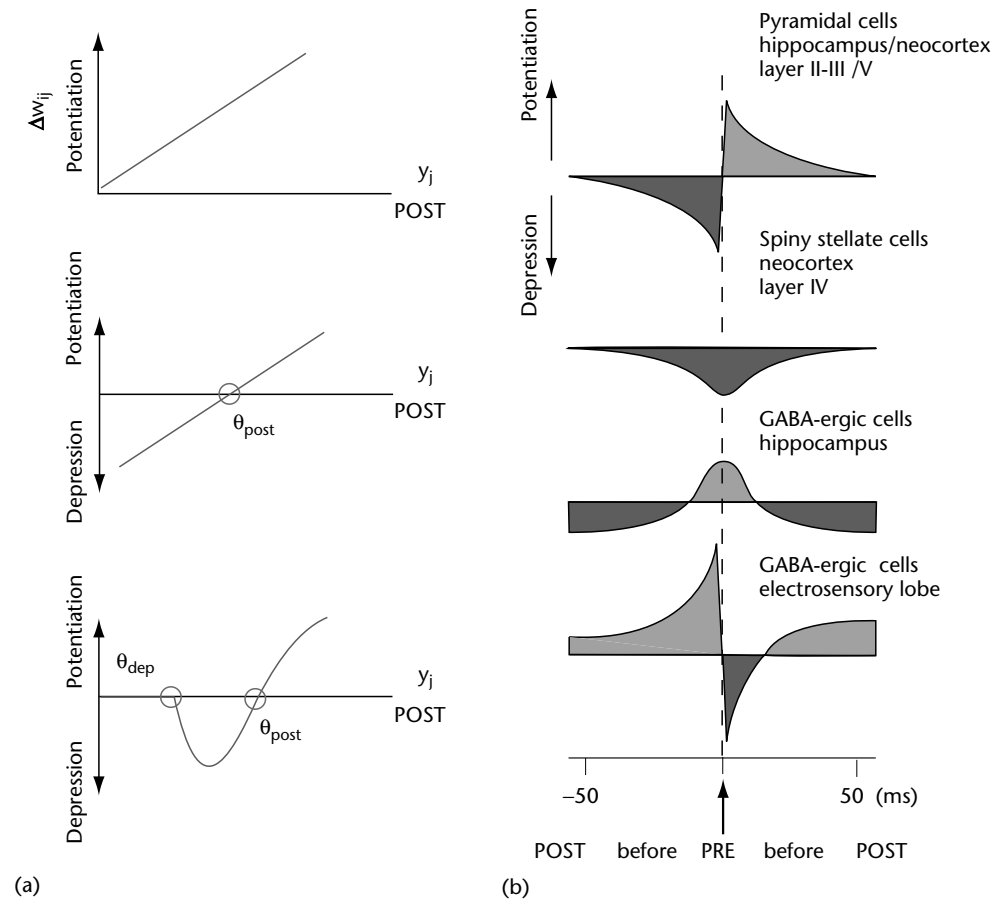


Figure 4. Hebbian synapses and spike-timing-dependent plasticity. (a) Hebb's rule (top) and the most often observed rules of homosynaptic plasticity established in biological networks, *in vivo* and *in vitro* (Frégnac, 2002). Each graph expresses the relationship between the induced synaptic change (positive ordinates for potentiation, negative ordinates for depression) and postsynaptic activity (abscissa) at the time of the association. The slope is proportional to presynaptic activity. The simple Hebbian rule (top) predicts potentiation only. The covariance rule (middle) (Bienenstock *et al*., 1982; Frégnac *et al*., 1988) and A–B–S rule (bottom) (Artola *et al*., 1990) predict both depression and potentiation, respectively, with one (θ_{post}) or two postsynaptic plasticity thresholds (θ_{dep}, θ_{pot}) (Artola *et al*., 1990). (b) Different forms of spike-timing-dependent plasticity rules established *in vitro* in co-cultures and acute slices. The induced synaptic change is expressed as a function of the temporal delay separating postsynaptic firing from presynaptic firing (taken here as the zero-time reference, and indicated by an arrow) that has been imposed during the pairing protocol. From top to bottom: pyramidal cells in the hippocampus or in non-granular layers in the neocortex; granular spiny stellate cells in the neocortex; GABA-ergic neurons in hippocampal culture; GABA-ergic medium ganglionic layer cells in the electrosensory lobe of the electric fish (Abbott and Nelson, 2000; Frégnac 2002).

as a key to perception. Some 50 years later, Hebb's postulate still impacts on fields of research as diverse as empiricist philosophy, associationism, experimental psychology, cybernetics, connectionism, neurocomputing and cellular neurobiology, and it raises the question of the generality of application of the underlying model. Taken in its most general terms, Hebb's principle becomes almost impossible to falsify, and an easy and often made criticism is that the information content that one can derive from its prediction remains poor. However, when Hebb's principle is further constrained by defining pre- and postsynaptic variables adapted to the specific study of a biological network, its predictive power remains surprisingly good, as numerous independent validations have been obtained in extremely diverse experimental preparations. The fact that the same principle still holds whatever application field and biological system is being considered, regardless of the level of analysis chosen and the timescale studied, is perhaps the best argument that Hebbian assemblies constitute a major abstract computational principle for explaining the inner dynamics of the mind.

References

Abbott LF and Nelson SB (2000) Synaptic plasticity: taming the beast. *Nature Neuroscience* **3**(Supplement): 1178–1183.

Abeles M (1982) *Local Cortical Circuits. An Electrophysiological Study*. New York: Springer-Verlag.

Abeles M (1991) *Corticonics: Neuronal Circuits of the Cerebral Cortex*. Cambridge, UK: Cambridge University Press.

Ahissar E, Vaadia E, Ahissar M *et al*. (1992) Dependence of cortical plasticity on correlated activity of single neurons and on behavioral context. *Science* **257**: 1412–1415.

Amit DJ (1995) The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences* **18**: 617–657.

Ariëns Kappers C (1932) On structural laws in the nervous system: the principles of neurobiotaxis. *Brain* **44**: 125–149.

Artola A, Bröcher S and Singer W (1990) Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of the rat visual cortex. *Nature* **347**: 69–72.

Bi G and Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* **24**: 139–166.

Bienenstock E (1994) A model of neocortex. *Network* **6**: 179–224.

Bienenstock E, Cooper LN and Munro P (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* **2**: 32–48.

Bienenstock E (1991) Un jeu de construction. *Science et Vie* **177**: 132–141.

Brown TH, Ganong AH, Kairiss EW and Keenan CL (1990) Hebbian synapses: biophysical mechanisms and algorithms. *Annual Review of Neuroscience* **13**: 475–511.

Delage Y (1919) *Le Rêve: Etude Psychologique, Philosophique et Littéraire*. Paris: Presses Universitaires de France.

Frégnac Y (2002) Hebbian synaptic plasticity. In: Arbib M (ed.) *Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press.

Frégnac Y and Bienenstock E (1998) Correlational models of synaptic plasticity: development, learning and cortical dynamics of mental representation. In: Carew T, Menzel R and Shatz CJ (eds) *Mechanistic Relationships Between Development and Learning: Beyond Metaphor*, pp. 113–148. Chichester: John Wiley and Sons.

Frégnac Y, Shulz D, Thorpe S and Bienenstock E (1988) A cellular analogue of visual cortical plasticity. *Nature* **333**: 367–370.

Hebb DO (1949) *The Organization of Behavior*. New York: John Wiley and Sons.

Hummel JE and Biederman I (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* **99**: 480–517.

James W (1890) *Psychology: Briefer Course*. Cambridge, MA: Harvard University Press.

Konorski J (1948) *Conditioned Reflexes and Neuron Organization*. London: Cambridge University Press.

Milner PM (1974) A model for visual shape recognition. *Psychological Review* **81**: 521–535.

Oram MW, Wiener MC, Lestienne R and Richmond BJ (1999) Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *Journal of Neurophysiology* **81**: 3021–3033.

Shastri L and Ajjanagadde V (1993) From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences* **16**: 417–494.

Stent G (1973) A physiological mechanism for Hebb's postulate of learning. *Proceedings of the National Academy of Sciences of the USA* **70**: 997–1001.

Vaadia E, Haalman I, Abeles M *et al*. (1995) Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature* **373**: 515–518.

Von der Malsburg C (1981) *The Correlation Theory of Brain Function*. Göttingen: Max-Planck Institute for Biophysical Chemistry.

Further Reading

Abbott LF and Song S (1999) Temporally asymmetric Hebbian learning, spike timing and neuronal response variability. In: Cohn DA (ed.) *Advances in Neural Information-Processing Systems*, pp. 69–75. Cambridge, MA: MIT Press.

Aertsen A, Diesman M and Gewaltig MO (1996) Propagation of synchronous spiking activity in feedforward neural networks. *Journal of Physiology (Paris)* **90**: 243–247.

Ahissar E, Abeles M, Ahissar M, Haidarliu S and Vaadia E (1998) Hebbian-like functional plasticity in the auditory cortex of the behaving monkey. *Neuropharmacology* **37**: 633–655.

Alkon DL, Sanchez-Andrés JV, Ito E *et al*. (1992) Long-term transformation of an inhibitory into an excitatory GABAergic synaptic response. *Proceedings of the National Academy of Sciences of the USA* **89**: 11862–11866.

Bi G and Poo M (1999) Distributed synaptic modification in neural networks induced by patterned stimulation. *Nature* **401**: 792–796.

Bienenstock E and Von der Malsburg C (1987) A neural network for invariant pattern recognition. *Europhysics. Letters* **4**: 121–126.

Cherubini E, Gaiarsa JL and Ben-Ari Y (1991) GABA: an excitatory transmitter in early postnatal life. *Trends in Neurosciences* **14**: 515–519.

Cho K, Aggleton JP, Brown MW and Bashir ZI (2001) An experimental test of the role of postsynaptic calcium levels in determining synaptic strength using perirhinal cortex of rat. *Journal of Physiology* **532**: 459–466.

Dale H (1935) Pharmacology and nerve-endings. *Proceedings of the Royal Society of Medicine* **28**: 319–332.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Frégnac Y (1986) Aplysia: Hebbian or not? *Trends in Neurosciences* **9**: 410.

Frégnac Y (1995) Comparative and developmental aspects of Hebbian synaptic plasticity. In: Arbib M (ed.) *Handbook of Brain Theory and Neural Networks*, pp. 459–464. Cambridge, MA: MIT Press.

Frégnac Y and Shulz D (1994) Models of synaptic plasticity and cellular analogs of learning in the developing and adult vertebrate visual cortex. In: Shinkman P (ed.) *Advances in Neural and Behavioral Development*, pp. 149–235. Norwood, NJ: New Jersey Neural Ablex Publishers.

Frey U and Morris RGM (1997) Synaptic tagging and long-term potentiation. *Nature* **385**: 533–536.

Lisman JE (1989) A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences of the USA* **86**: 9574–9578.

Milner PM (1957) The cell assembly: mark II. *Physiological Reviews* **64**: 242–252.

Nass MM and Cooper LN (1975) A theory for the development of feature-detecting cells in visual cortex. *Biological Cybernetics* **19**: 1–18.

Pavlov IP (1927) *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. London: Oxford University Press.

Rochester N, Holland JH, Haibt LH and Duda WL (1956) Tests on a cell-assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory* **IT2**: 80–93.

Sahley CL (1985) Co-activation, cell assemblies and learning. *Trends in Neuroscience* **8**: 423–424.

Sejnowski TJ (1977) Storing covariance with non-linearly interacting neurons. *Journal of Mathematical Biology* **4**: 303–321.

Willshaw D and Dayan P (1990) Optimal plasticity from matrix memories: what goes up must come down. *Neural Computation* **2**: 85–93.

Hippocampus

Intermediate article

John O'Keefe, University College London, UK

CONTENTS

The hippocampus is a cortical structure located in the temporal lobes of the brain. Its role in memory has been extensively investigated in humans and other animals.

INTRODUCTION

There is universal agreement that the hippocampus, a cortical structure located in the temporal lobes, is involved in memory. Theories of hippocampal function are based on studies of amnesic patients and experimental animals. The cognitive map theory is the most clearly specified and is supported by the most evidence, but the declarative memory and flexible relational theories also have their proponents.

ANATOMY

Historically, the hippocampal formation has been considered to be a part of the limbic system, a set of structures traditionally believed to be constituents of the emotional brain. It was also included in the Papez circuit, a network of structures envisaged by Papez to subserve emotion. As we shall see, there is good evidence that the hippocampal formation and the Papez circuit are not involved in emotion but instead are part of the memory system of the brain. In humans, the hippocampus is an elongated structure lying on the medial surface of the temporal lobe (Figure 1a). In rats, it is more banana-shaped and extends dorsally upwards from the temporal lobes to lie beneath the parietal cortex (Figure 1b). There are two major pathways by which the hippocampus communicates with the rest of the brain: the fornix–fimbria fiber bundle which connects it with the septal nuclei and the hypothalamus, and the perforant path which connects it with the cerebral cortex. The fornix–fimbria conveys information to the hippocampal formation about the animal's movements, attentional and bodily states and takes information from the hippocampus to control movements, in particular those which take the animal towards goals or away from punishments. An important function of the fornical input is to provide both the cholinergic and γ-aminobutyric acid (GABA) inputs which are a necessary condition for the theta electroencephalographic (EEG) state of the hippocampus (see below). Sensory information about the external world enters the hippocampus through the perforant path, which arises in the entorhinal cortex. The entorhinal cortex in turn receives information from other parts of the temporal lobe, in particular the perirhinal and parahippocampal cortices and ultimately from many of the sensory analysing regions of the cerebral cortex. These areas in the parahippocampal gyrus appear to be responsible for at least some aspects of the global amnesia syndrome which follows extensive damage to the medial temporal lobes. The hippocampal return output to the neocortex may provide the context for selection of cortical processing, or the results of hippocampal computations may be stored in a more permanent form there. One suggestion is that the hippocampus provides allocentric spatial information whereas the parietal cortex deals in egocentric spatial information, and this pathway may be the way in which these two different representations interact.

Interestingly, with minor exceptions, a slice taken at any level through the hippocampus of either the rat or the human shows the same internal structure. Figure 1b shows this profile in the rat. As seen in Figure 2, information originating in the entorhinal cortex (EC) flows in one direction through each of the subsections of the hippocampal formation, taking either a slow stopping route or an express route. Information via the slow route passes first to the dentate gyrus (DG), from there to the CA3 field, next to the CA1 field, finally to end in the

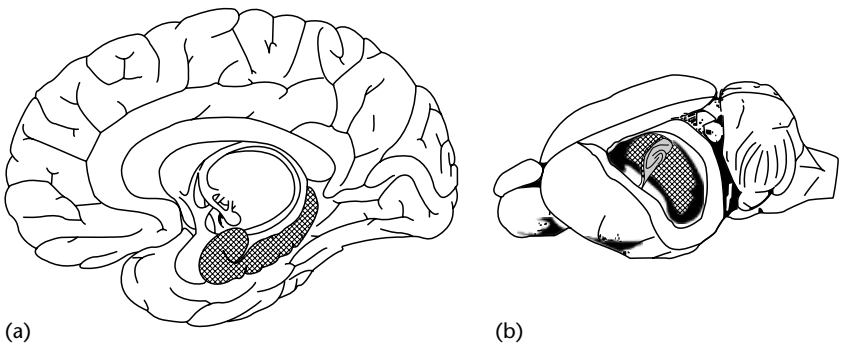


Figure 1. Location of the human (a) and rat (b) hippocampus within the brain (indicated by cross hatching).

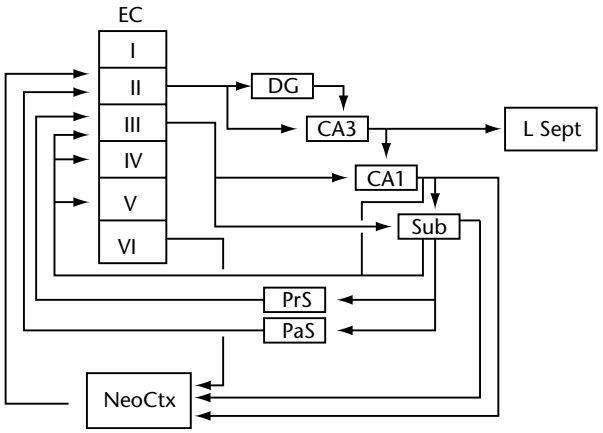


Figure 2. Internal pathways of the hippocampus (after Amaral and Witter, 1995). Abbreviations: EC, entorhinal cortex; DG, dentate gyrus; CA, cornu ammonis; L Sept, subiculum. Alternatively, a rapid transit direct connection exists from the entorhinal cortex to each of the subsections. The function of these separate pathways is not yet clear although there is speculation that different frequencies of entorhinal inputs may take different routes, or that each subfield can compare the information coming via the two routes. It is also not clear yet whether different functions can be ascribed to the different hippocampal fields. For example, the properties of the spatial cells in the different subfields do not seem to vary much, and – with the exception of long-term potentiation (LTP) in the mossy fiber pathway from the dentate to CA3 – the physiological properties of the cells are broadly similar.

THE ROLE OF THE HIPPOCAMPUS IN MEMORY

The Case of H.M.

Interest in the role of the hippocampus in memory stems from the original description of the amnesic syndrome which followed bilateral damage to the medial part of the temporal lobes in the classic case of the patient H.M. (Scoville and Milner, 1957). This patient suffered from severe epileptic seizures which were not well controlled by medication, and underwent surgical removal of the hippocampus and other temporal lobe structures on both sides in an attempt to cure this epilepsy. The patient's seizures were reduced by the operation but he was also left with a severe memory problem. He could no longer acquire new information, and he had lost information which he had acquired prior to the operation. The loss of previous information (retrograde amnesia) was not complete but appeared to extend for a period of 11 years prior to his operation. This patient, who has been tested exhaustively since his operation, fails to remember events or episodes he has experienced, and seems incapable of learning new vocabulary words or facts. Squire (1992) has termed the type of memory lost in H.M. 'declarative memory' to distinguish it from 'procedural memory' which appears to be intact. For example, H.M. can learn new perceptual and motor skills such as those needed to identify line drawings from fragments of the original or to draw images seen in a mirror. One question which arises from the study of patients such as H.M. concerns which structure or structures in the medial temporal lobes are responsible for the amnesia. Are

different brain areas responsible for different parts of the syndrome, or are we dealing with a single system in which damage to any part produces an impairment of all aspects of declarative memory, with greater amounts of damage leading to a denser, but qualitatively similar, amnesia? In order to answer this question it is necessary to look at the memory capacities of patients with damage restricted to parts of the mesial temporal lobe. The reader should be warned, however, that there is always the possibility that small, subtle amounts of damage outside the hippocampal formation might go undetected by even the most sophisticated modern histological or imaging techniques.

Patients with More Restricted Damage to the Hippocampus: R.B., V.C. and Jon

Patients with apparently similar damage restricted to the hippocampus appear to vary in the extent and nature of their memory deficits. For example, some have a recognition memory deficit, while others do not; some have a graded retrograde memory loss, others do not. Nevertheless, one area of agreement is that they all suffer from a deficit in episodic memory.

The patient R.B. suffered anoxic damage to the hippocampus as an adult following a series of heart attacks (Zola-Morgan *et al*., 1986). He died soon after and histological examination of his brain revealed that damage was mainly restricted to the CA1 field of the hippocampus. He showed a much milder amnesia than H.M. and little or no forgetting of information acquired prior to his brain damage. In support of the declarative memory theory, R.B. had a memory deficit when tested by either recognition or recall. The patient V.C. (Cipolotti *et al*., 2001) had a slightly larger lesion which encompassed the entire hippocampal formation but spared the entorhinal cortex and other temporal lobe structures. Like R.B., his memory deficits included public as well as personal events whether tested by recognition or recall. In contrast to R.B., he had an extensive retrograde amnesia as well as a profound anterograde amnesia. Furthermore, there was no indication of a gradation in his retrograde amnesia, his memory for 30-year-old events being as poor as for more recent ones. Both these patients acquired their hippocampal damage as adults. In contrast, the third patient, Jon, is one of three developmental amnesic patients who acquired their hippocampal damage early in life (Figure 3) and whose profile of memory deficit is different from the adult amnesic patients (Vargha-Khadem *et al*., 1997). All three early-onset patients were brought to the clinic by their parents who complained that they were forgetful about events, places and times. However, while they share the difficulties in recalling events of their past with adult amnesics, these patients seem to be much less impaired when required to recognize items they have recently seen. They also seem to be much more impaired in memory for episodes than in their memory for facts, although this dissociation has been less well studied. These findings suggest that some of the deficits seen in adult amnesic patients following hippocampal damage might be due to covert extra hippocampal involvement, or that brain injury to the hippocampal system can be compensated for by remaining intact tissue if the damage occurs early in life.

One interesting facet of the relatively spared recognition memory of these patients is that there are two domains in which they are still impaired: object in place recognition memory, and voice/face associative recognition memory. Extensive testing of the recognition memory capacity of the patient Jon shows that he is normal when asked to remember even difficult associations such as which two doors were seen together, but is very poor at allocentric spatial tasks such as finding his way around a virtual reality environment or identifying the correct location of an object when asked to do so from a viewpoint different from that from which it was originally seen (King *et al*., 2002). In this last task, healthy volunteers appear to solve the problem by imagining their viewpoint rotated to the appropriate location or by rotating the array of objects themselves. An inability to do either might explain the inflexibility attributed to these patients by both the cognitive map and flexible relational theories (see below).

Another interesting aspect of the memory deficit exhibited by many (but not all) people with amnesia is a gradient of retrograde amnesia, such that older memories are relatively spared in comparison to more recent ones. Declarative memory theory proposes that this is due to the memories being shifted from the hippocampus into the neocortex over time by a process of consolidation. In the early years following acquisition of a memory, interactions between the elementary component traces in the disparate regions of the neocortex require hippocampal support, but over time these bind directly to each other and become independent of the hippocampus. An alternative view, put forward by Nadel and Moscovitch (1997), is that some memories, such as those of a personal autobiographical nature, always rely on the integrity of the hippocampus, but that others, such as those

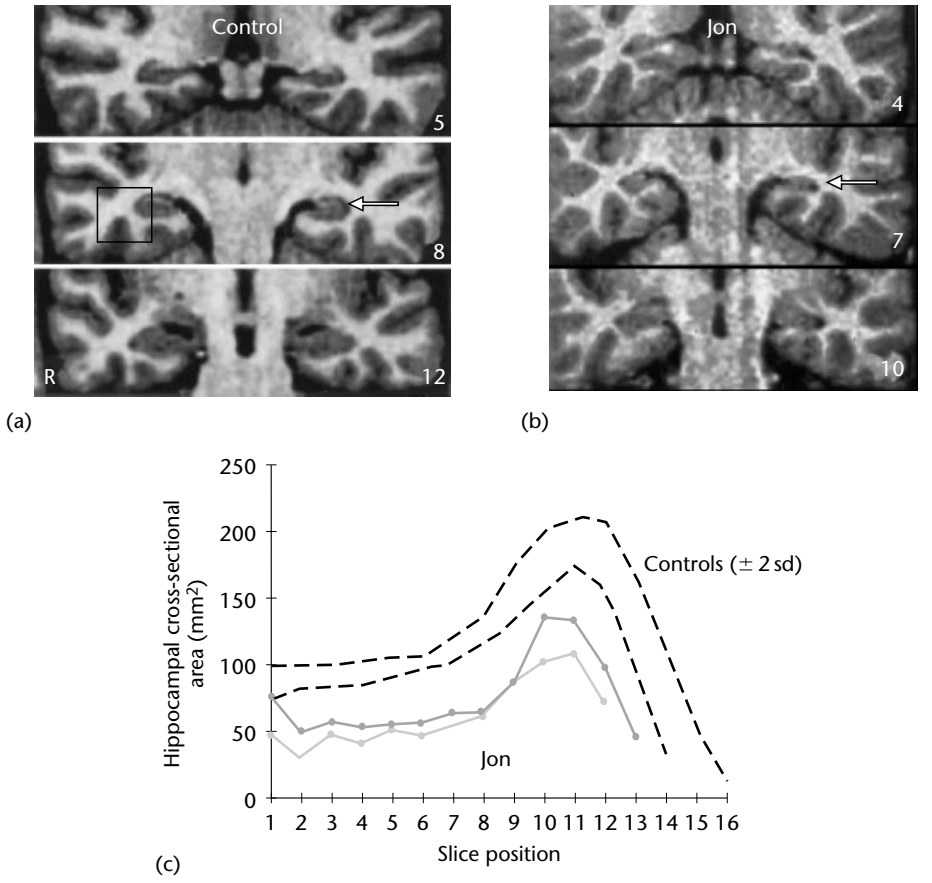


Figure 3. Hippocampal damage in the amnesic patient Jon. Arrows point to the right hippocampus in the control brain (a) and in Jon's brain (b). (c) Cross-sectional area measurements showing a 50% reduction in Jon's hippocampus. Dotted lines show 2 standard deviations of control hippocampal cross-section (after Spiers *et al.*, 2001a).

about public events or personalities, can show a gradient depending on how many times they have been reactivated since acquisition. According to this multiple trace theory, each reactivation lays down a new trace of the memory, rendering the multiply-represented memory less vulnerable to selective damage. The greater the amount of damage outside the hippocampus proper in the medial temporal lobe, the more likely it is that there will be multiple memories and that these nonpersonal memories will show temporal gradients of forgetting. At this stage it is difficult to choose between these two notions but the existence of patients with flat gradients of retrograde amnesia for personal autobiographical memories and results showing equal activation of the hippocampus in imaging studies for remote and recent memories support the multiple trace theory.

ANIMAL MODELS OF AMNESIA

Models of Amnesia in Monkeys

The declarative memory theory suggests that the hippocampus stores information about consciously recallable facts, faces and meanings of words as well as personal episodes. Although it is not possible to interrogate an animal's conscious recollection in the same way as that of the human, Squire and Zola-Morgan have proposed several analogs of declarative memory tasks which can be performed by monkeys. These include the ability to learn several visual discrimination tasks concurrently, delayed retention of object discriminations, delayed response, and delayed nonmatching-to-sample. Performance on all of these tasks is impaired in monkeys with large temporal lobe lesions of the sort that H.M. has. Unfortunately,

more selective lesions of the hippocampus itself produced by using neurotoxins or ischemia do not cause such extensive impairments (Zola-Morgan *et al*., 1992) suggesting that the deficits are due to other areas in the temporal lobe, in particular the perirhinal and parahippocampal cortices which surround the hippocampus. The delayed nonmatching-to-sample task is particularly instructive since it is also a good example of a flexible relational task. On this task, the animal is shown a novel object and must remember it for long enough to avoid it when it is subsequently paired with a second novel object in a forced-choice paradigm. Monkeys with selective hippocampal lesions can perform this task normally with delays over tens of minutes between the first and second presentation of the object (Murray and Mishkin, 1998). Of equal importance, single unit recordings taken from the hippocampus of animals performing this task have reported virtually no neurons selectively responding to the novelty of the object, in marked contrast to neurons in the perirhinal cortex which do show such responses and which are damaged in the large lesions produced in the study by Squire and Zola Morgan. One possible explanation for the absence of a deficit on this task is that it could in principle be solved on the basis of the relative familiarity of the two objects as well as on the basis of the encoding of an episodic memory that a particular object was seen at a particular time in the testing location. This explanation, however, would not account for the absence of single unit responses related to the task.

Models of Amnesia in Rats

A second approach to understanding the role of the hippocampus in memory involves work on the hippocampus of the rat. Here there have been two different approaches. In the first, the functions of the rat hippocampus are considered with respect to the animal's lifestyle and allowances are made for the possibility that additions and/or modifications to the basic system as found in the rat have been made during the course of evolution to produce a more general episodic memory system of the sort which would account for the global memory problems of humans with amnesia. The main theory here is the cognitive map theory. The second approach assumes that the functions of the rat and human hippocampi are the same and seeks to come up with a general function for both: the dominant hypothesis here is the flexible relational theory.

COGNITIVE MAP THEORY OF HIPPOCAMPAL FUNCTION

The original impetus for the development of the cognitive map theory was the discovery by O'Keefe and Dostrovsky (1971) of spatially coded neurons in the hippocampus of freely moving rats. These neurons were named 'place cells' by O'Keefe (1976) who studied their properties in more detail (*See* **Place Cells**). The primary correlate of each place cell is a location in a familiar environment. Different cells code for different locations such that a small group of them provide coverage of an entire environment. Some cells respond solely to location, whereas others incorporate information about objects in that location. Most importantly, the cells signal that the animal has entered the preferred location irrespective of its heading direction: they provide what is termed 'allocentric' spatial information, which is independent of the orientation of the body axis and can be contrasted with 'egocentric' spatial information within which objects are located relative to body-centered frameworks such as the eye, the head or the body. On the basis of these findings, O'Keefe and Nadel (1978) proposed that the hippocampus formed a cognitive map, consisting of a set of place representations which completely cover an environment and which are linked together by vectors that specify the direction and distance between places. The discovery of the head direction cells in the neighboring subicular area a few years later (Ranck, 1984; Taube *et al*., 1990) and the demonstration that the place cell firing rate varies as a function of the animal's running speed (Czurko *et al*., 1999) have provided strong support for the cognitive map theory since they confirm that the information required by a mapping system is available to the hippocampus. Further support for the theory comes from experiments on the effects of damage to the hippocampus on the spatial abilities of rats. As part of their general theory, O'Keefe and Nadel (1978) suggested that there were several ways in which animals could find their way around an environment. In addition to the use of a cognitive map-like representation of the environment, they could also move from one place to another using 'routes'. Routes were viewed as stimulus–response–stimulus (S–R–S) chains similar to those proposed by behaviorists such as Hull to explain all of behavior. They direct attention to specific stimuli in the environment and specify the appropriate behaviors in response to those stimuli.

Let us see how these two types of strategies differ in the way in which they direct an animal to solve a simple T-maze problem, in which an animal in the stem of a T-shaped maze must go to the left-hand arm of the T in order to receive reward. An example of a successful route would instruct the animal to leave the start arm and go to the T-shaped intersection of the maze (S) , turn left and run straight ahead (R) until it came to the food well (S). This can be contrasted with behavior based on the cognitive map, which would identify both the animal's position in the start arm and the location of the goal in its cognitive representation of the maze, and then instruct the motor system to locomote in the appropriate direction for the appropriate distance to reach the goal. When rats are first trained in such a maze they tend to use place hypotheses in the early stages of learning, but then with continued practice the route hypothesis begins to dominate. This can be shown in a simple probe test in which the stem of the T maze is rotated by 180° to lie on the opposite side of the T. This forces the animal to choose between going to the goal location (making the opposite body turn to that usually made), or sticking with a route strategy and making the same body turn (taking it to the opposite arm of the T from usual). The dependence on the hippocampus of place, but not route, strategies has been nicely shown in an experiment by Packard and McGaugh (1996) in which the injection of local anesthetic solution into either the hippocampus or the caudate nucleus (a good candidate for the neural basis of route strategies) was done during probe trials given either just after rats had learned the T maze or following substantial overtraining. Just after learning the rats tended to use place hypotheses and performance was blocked by hippocampal injections; later on, the rats used route hypotheses which were blocked by caudate injections. The blocking of route hypotheses in overtrained animals uncovered place hypotheses which were still intact but overshadowed by the stronger route hypotheses.

The cognitive map hypothesis makes three strong predictions about the behavioral deficit following hippocampal damage: it predicts that animals will lose their allocentric spatial memories, will be unable to navigate to a hidden goal when started from different locations in a familiar environment, and will not explore novel environments or be sensitive to changes in the location of objects in a familiar environment. All three predictions have been consistently confirmed. The best available tests of one-trial spatial memory in the rat are the T-maze alternation test and the Olton radial arm maze. Each trial on the T-maze alternation task consist of two runs. On the first run, one of the arms is blocked, forcing the animal into the other where it is rewarded. On the second run, both arms are available and the correct rewarded choice is the one not visited on the first run. The animals have to remember which arm they were forced to enter on the first run in order to choose correctly on the second. Animals with hippocampal damage are very poor at this task when the delay between the two runs is longer than a few seconds (e.g. Aggleton *et al*., 1986). The Olton maze can be viewed as a more testing elaboration of this. In this task, eight arms radiate from a central platform each with a food well at its peripheral end. At the beginning of each trial there is food in each of the wells but this is not replenished when the animal eats it. The animal's optimal strategy on each trial is to retrieve all of the eight food rewards without reentering any of the depleted arms. There are two versions of the task, one in which the goal arms are identified by their spatial location in the room and another in which each arm contains a distinctive set of cues which the animal can use as an *aide-mémoire* instead of the spatial cues. While normal animals learn either version, animals with hippocampal lesions are very poor at the spatial version but succeed on the cued version (Jarrard, 1993). This implies that they are motivated to solve the task and have preserved motor skills but lack the ability to form the spatial memories needed to solve it. A similar dissociation between their ability to use spatial and nonspatial information can be seen in their navigational performance. The most widely used test of navigation is the water maze task devised by Richard Morris. At the start of each trial, the animal is placed in a pool filled with cloudy water and must escape by swimming to a hidden platform in a fixed location. The start location varies from trial to trial so that the animal cannot simply learn to head in a specific direction or towards a specific stimulus. Raising the platform above the level of the water transforms it into a visible cue which can substitute for information about its location. Rats with hippocampal lesions are deficient on the spatial version of the task but learn the visible cued version as quickly as normal rats (Morris *et al*., 1982).

The last prediction of the cognitive map theory is that rats with damage to the hippocampus should not be sensitive to changes in the spatial layout of objects and should not explore them as do normal rats. This follows from a core postulate of the theory that the motivation for originally building and subsequently altering maps is the cognitive

one of curiosity. Animals build maps either because they have no hippocampal representation of an environment or because there is a mismatch between the stored representation and the current experience of that environment. Recent experiments have shown that hippocampal-damaged rats do not respond to the movement of objects or stimuli in a familiar environment, but that they do respond to the introduction of a new stimulus (Xavier *et al*., 1990; Save *et al*., 1992).

It is clear that the cognitive map theory accounts for the function of the rat hippocampus very well (although, as we shall see in a moment, there have been suggestions that it is not broad enough). How does it deal with the global amnesia seen in humans following damage to the hippocampal formation? The theory begins by acknowledging that the human hippocampus has a wider and more elaborate function than that of the rat. This partly reflects the development of language and consequent lateralization of cortical functions in the human, and the development of a sense of linear time. This latter, assumed to be dependent on the dramatic expansion of the prefrontal cortex in higher mammals, provides a time stamp for each individual visit to a location in the map, forming the basis for memories embedded in a spatiotemporal context or what Tulving has called 'episodic' memories. Most language functions are localized to the left side of the brain and this lateralization is honored by the hippocampal memory system. Historically it has been recognized that lesions of the right human hippocampus result in spatial and visuospatial memory deficits, lesions to the left side result in verbal memory deficits (Milner, 1971; Frisk and Milner, 1990). The cognitive map theory views spatial language as the prototype for all language. Instead of both hippocampi storing the locations of physical objects, as in the rat, the left human hippocampus uses the same system to store the symbols originating in the auditory modality which stand for these representations. In support of this hypothesis, analysis of the meanings of English prepositions shows how they code for the shapes of place fields and the locations of objects relative to other landmarks (O'Keefe, 1996). In this view, the function of the left human hippocampus is to store episodes and narratives. Evidence in support of the lateralization of spatial memory to the right and episodic memory to the left comes from experiments on patients who have had unilateral temporal lobectomies for the relief of epilepsy (Spiers *et al*., 2001b) and from the bilateral hippocampal patient Jon who is deficient in both (Spiers *et al*., 2001a). Further evidence comes from functional imaging studies which show that the right hippocampus is selectively activated during imagined navigation in taxi drivers or in virtual reality environments (Figure 4), while the left hippocampus is more active during episodic memory tasks.

ASSOCIATIVE MEMORY THEORIES OF HIPPOCAMPAL FUNCTION

Alternative explanations of hippocampal function do not dispute its role in spatial memory but seek to broaden its function to include nonspatial associative memory as well. Noting that the cognitive map theory needs to broaden the functions it ascribes to the hippocampus in order to account for the nonspatial functions of the left human hippocampus, they ask whether a broader theory might apply to the rat hippocampus as well. Currently there are two versions of this broader associative theory with some support: configural associative theory (Sutherland *et al*., 1989) and flexible relational theory (Cohen and Eichenbaum, 1993). Configural association theory suggests that the hippocampus binds together two separate representations to make a third 'configured' representation. This would enable an animal to respond differently to the configuration than to either of the elements presented independently, for example enabling it to respond positively to either a light or a tone but not to the combination of the two. While there is some evidence that rats with hippocampal lesions might have difficulties in solving some problems of this nature, two carefully performed experiments (Gallagher and Holland, 1992; Davidson *et al*., 1993) have found them to be normal on these tasks. Both demonstrated that the animals with lesions could learn that a stimulus (e.g. a light) on its own had one valence but in combination with a second stimulus (e.g. a tone) had the opposite valence. These results show that animals can use configurations, and argue for a more restricted role for the hippocampus than that envisioned by the configural association theory. This has been acknowledged by the authors in a more recent paper (Rudy and Sutherland, 1995). The flexible relational theory of Cohen and Eichenbaum views the hippocampus as storing many of the different types of relationships which stimuli or objects can have to each other. Some examples are, 'A is associated with a particular smell', 'A is more desirable than B', 'A came before B in time'. Like the cognitive map theory, the relational theory emphasizes the flexibility of the information stored in the hippocampus, either enabling it to be

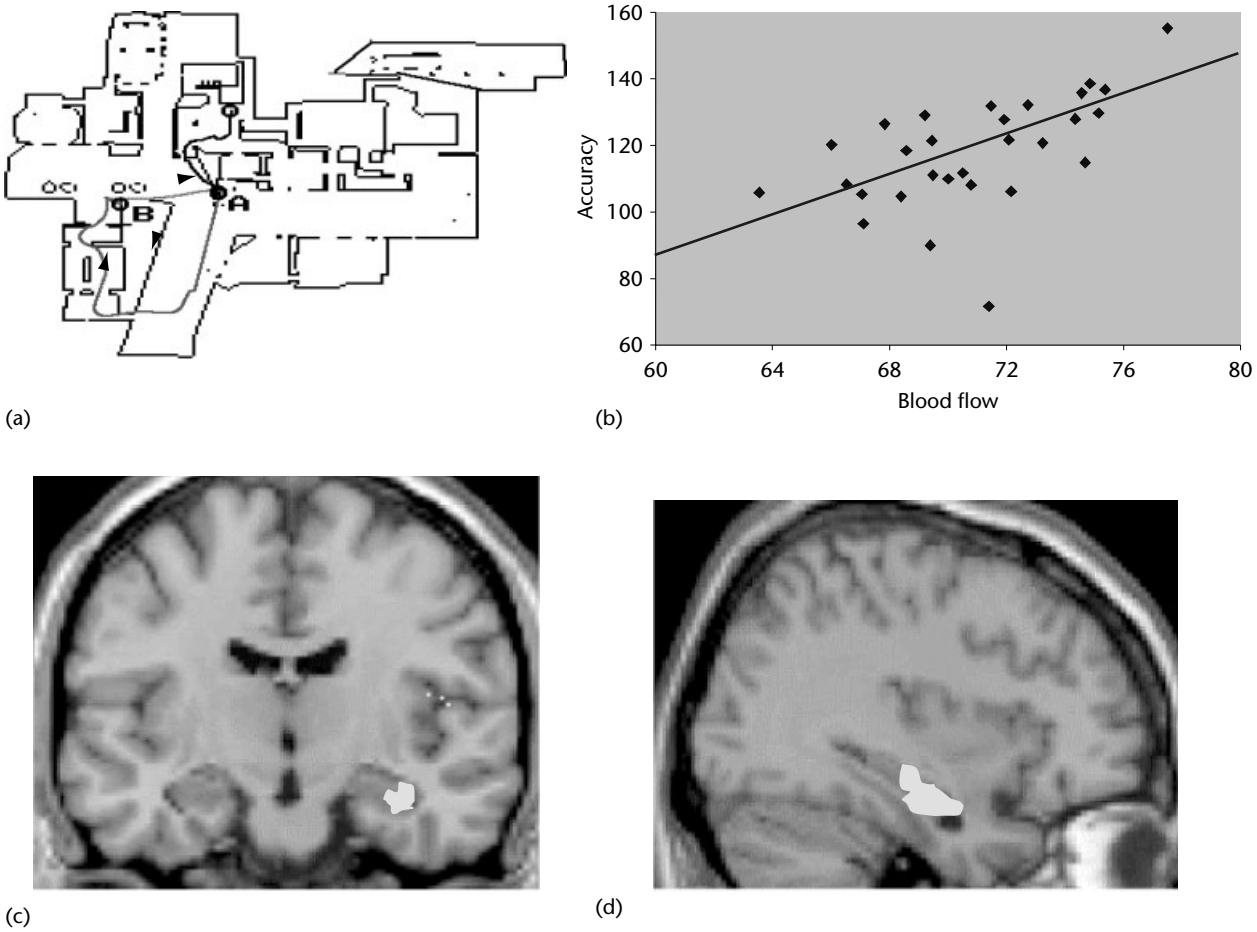


Figure 4. Hippocampal activation during successful navigation in a virtual reality environment. (a) Plan view of the virtual environment with three paths shown. (b) Correlation of navigational accuracy in degrees (max = 180°), with hippocampal blood flow in rCBF units. Correlation coefficient, $r = 0.56$. (c) Coronal and (d) parasagittal sections through human brain showing positron emission tomography activation of hippocampus correlated with accurate navigation (after Maguire *et al*., 1998).

used in contexts other than those in which it was acquired or serving as the basis for inferences not explicit in the originally learned relations but deducible from them. The main evidence for this theory comes from lesion and single unit recording experiments in rats. Rats with lesions of the hippocampus or of structures that send fibers to the hippocampus such as the entorhinal cortex or fornix have been reported to fail nonspatial tasks requiring these relational abilities. To take two examples: first, rats are social animals and use their interactions to communicate information about edible foods; after smelling the odor of food on the breath of a familiar rat, a normal rat will choose it in preference to an unknown food. Rats with hippocampal lesions were reported to forget this association faster than normals (Winocur, 1990; Bunsey and Eichenbaum, 1995). In another experiment, the role of the hippocampus in the ability of the rats to carry out transitive inferences was tested. Rats were trained that following the smell of odor A, odor B rather than C was correct, and independently that following odor B, odor D and not E was correct. Subsequently, presented with A followed by a choice between D and E they were able to infer that odor D was the correct choice. Rats with hippocampal damage could learn the original relations but could not deduce the correct inference, suggesting that they lacked the relational machinery (Bunsey and Eichenbaum, 1996).

Unfortunately, the reliability of these findings has not been established by replication in other laboratories. Burton and coworkers (2001) could not reproduce the social transmission of olfactory information finding, and Li *et al*. (1999) found no

deficit in the transitive inference task. Therefore, for the time being, one must suspend judgment on the claim made by these theories that there is a set of non-spatial memory tasks that are reliably failed by animals with hippocampal damage. A similar caveat applies to the results of single unit recording studies which have been adduced to support the relational theory. The simplest type of relationship between two entities is the identity relationship $A = A$. The relational theory should predict hippocampal involvement in tasks which rely on this capacity for reidentification, particularly when the animal is required to act flexibly and make a different response to the two presentations of the object or stimulus. The ideal task for measuring this capacity is the delayed nonmatching-to-sample (DNMS) task in which the animal is rewarded for responding to the first presentation of an item but then must reject it and choose the alternative item on the second presentation. As we have seen above, it was originally believed that animals with hippocampal lesions were deficient at these simple memory tasks (Zola-Morgan and Squire, 1985) but it is now clear that this is not the case for either monkeys (Murray and Mishkin, 1998) or rats (Mumby *et al*., 1996). Furthermore, single unit experiments in primates have failed to find cells reflecting this relationship. The adherents of the relational theory originally claimed this putative deficit as evidence in its favor but no longer view DNMS as a good test of this type of memory.

THE ROLE OF EEG THETA

There are reasons to believe that the hippocampus acts as an integrated structure. One mechanism by which this integration takes place is reflected in the hippocampal EEG. During some behaviors, the electroencephalogram recorded from the hippocampus of rats and other animals exhibits a striking sinusoidal rhythm called the theta rhythm. At other times, the EEG is desynchronized. The behavioral correlates of the theta waves vary slightly among species but in general can be found in response to arousing stimuli and during behaviors that change the animal's location in the environment. Recordings of the theta waves from different parts of the hippocampus suggest they are locked in synchrony and thus may act to bind the hippocampus into one coordinated system. Both pyramidal cells and interneurons fire in bursting patterns with phase relationships to the ongoing EEG theta, indicating that the mechanism that generates theta has functional significance. The interneurons always maintain a fixed phase relationship to the theta waves whereas the pyramidal cells have a more interesting correlate: as the animal enters the place field, the earliest firing of the cell occurs at a particular phase of the theta wave (Figure 5), but as the animal continues through the field each subsequent firing burst moves to an earlier phase of the wave (O'Keefe and Recce, 1993). The best correlate of the phase of firing is the animal's location on the maze and this improves the animal's ability to identify its location by 43% (Jensen and Lisman, 2000). Another important role for theta is in gating the efficacy of afferent inputs in producing long-term potentiation (LTP) and long-term depression (LTD), the phenomena in which a brief train of electrical pulses delivered to the afferent fibers to a cell result in an increased (or decreased) response for periods of hours, or even days (Bliss and Lomo, 1973; Bliss and Collingridge, 1993). When the afferent synaptic input arrives at the theta peak, LTP results; at the trough, LTD (Huerta and Lisman, 1995; Orr *et al*., 2001). We can conclude that the theta system is important in integrating activity across the hippocampus, in acting as a timing device against which place cell activity can be clocked, and in setting up the conditions for synaptic modification.

CONCLUSION

The hippocampus is involved in the storage of memories in both animals and humans. In animals such as the rat the memories are primarily or exclusively spatial, while in humans they include memories for episodes and narratives as well. Evidence for the role of the rat hippocampus in spatial memory comes from place-coded cells recorded there and from the deficits in spatial memory and navigation following damage to the hippocampus. A broader role in other kinds of relations has been suggested for the rat hippocampus but the single unit recording and lesion evidence in support of this position is much less robust than for the cognitive map theory. Both the cognitive map and relational theories agree that the human hippocampus has a broader function than that of the rat and that one of its functions is to store episodic memories. The declarative memory theory takes an even broader view, suggesting that factual and semantic memories in humans are also dependent on the hippocampus. This latter theory clearly captures the functions of the medial temporal lobe as a whole, but recent results suggest that different structures within this lobe subserve different components of the total declarative memory function.

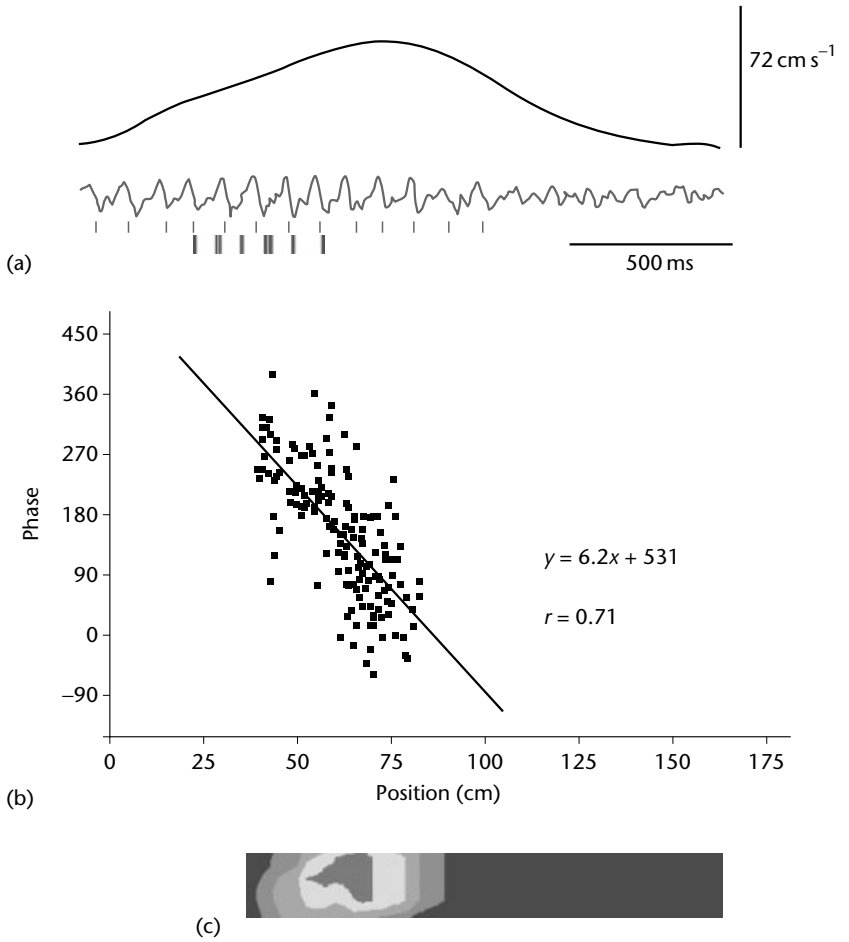


Figure 5. Hippocampal place cell activity changes phase with the electroencephalographic (EEG) theta activity. (a) Velocity profile, EEG theta, and place cells firing during a single run on a linear track. (b) Phase of theta at which unit fires versus position on the track. (c) Location of place field on the track (after O'Keefe and Burgess, 1999).

References

Aggleton JP, Hunt PR and Rawlins JN (1986) The effects of hippocampal lesions upon spatial and non-spatial tests of working memory. *Behavioural Brain Research* **19**: 133–146.

Amaral DG and Witter MP (1995) The hippocampus. In: Paxinos G (ed.) *The Rat Nervous System*, pp. 443–493. New York: Academic Press.

Bliss TV and Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**: 31–39.

Bliss TV and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–356.

Bunsey M and Eichenbaum H (1995) Selective damage to the hippocampal region blocks long-term retention of a natural and nonspatial stimulus-stimulus association. *Hippocampus* **5**: 546–556.

Bunsey M and Eichenbaum H (1996) Conservation of hippocampal memory function in rats and humans. *Nature* **379**: 255–257.

Burton S, Murphy D, Qureshi U, Sutton P and O'Keefe J (2000) Combined lesions of hippocampus and subiculum do not produce deficits in a nonspatial social olfactory memory task. *Journal of Neuroscience* **20**: 5468–5475.

Cipolotti L, Shallice T, Chan D *et al*. (2001) Long term retrograde amnesia. The crucial role of the hippocampus. *Neuropsychologia* **39**: 151–172.

Cohen NJ and Eichenbaum H (1993) *Memory, Amnesia and the Hippocampal System*. Cambridge, MA: MIT Press.

Czurko A, Hirase H, Csicsvari J and Buzsaki G (1999) Sustained activation of hippocampal pyramidal cells by 'space clamping' in a running wheel. *European Journal of Neuroscience* **11**: 344–352.

Davidson TL, McKernan MG and Jarrard LE (1993) Hippocampal lesions do not impair negative patterning: a challenge to configural association theory. *Behavioral Neuroscience* **107**: 227–234.

Frisk V and Milner B (1990) The role of the left hippocampal region in the acquisition and retention of story content. *Neuropsychologia* **28**: 349–359.

Gallagher M and Holland PC (1992) Preserved configural learning and spatial learning impairment in rats with hippocampal damage. *Hippocampus* **2**: 81–88.

Huerta PT and Lisman JE (1995) Bidirectional synaptic plasticity induced by a single burst during cholinergic theta oscillation in CA1 in vitro. *Neuron* **15**: 1053–1063.

Jarrard LE (1993) On the role of the hippocampus in learning and memory in the rat. *Behavioral Neural Biology* **60**: 9–26.

Jensen O and Lisman JE (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *Journal of Neurophysiology* **83**: 2602–2609.

King JA, Burgess N, Hartley T, Vargha-Khadem F and O'Keefe J (2002) The human hippocampus and viewpoint dependence in spatial memory. *Hippocampus* (in press).

Li H, Matsumoto K and Watanabe H (1999) Different effects of unilateral and bilateral hippocampal lesions in rats on the performance of radial maze and odor-paired associate tasks. *Brain Research Bulletin* **48**: 113–119.

Maguire EA, Burgess N, Donnett JG *et al*. (1998) Knowing where and getting there: a human navigation network. *Science* **280**: 921–924.

Milner B (1971) Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin* **27**: 272–277.

Morris RGM, Garrud P, Rawlins JN and O'Keefe J (1982) Place navigation impaired in rats with hippocampal lesions. *Nature* **297**: 681–683.

Mumby DG, Wood ER, Duva CA *et al*. (1996) Ischemia-induced object-recognition deficits in rats are attenuated by hippocampal ablation before or soon after ischemia. *Behavioral Neuroscience* **110**: 266–281.

Murray EA and Mishkin M (1998) Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. *Journal of Neuroscience* **18**: 6568–6582.

Nadel L and Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology* **7**: 217–227.

O'Keefe J (1976) Place units in the hippocampus of the freely moving rat. *Experimental Neurology* **51**: 78–109.

O'Keefe J (1996) The spatial prepositions in English, vector grammar and the cognitive map theory. In: Bloom P, Peterson M, Nadel L and Garrett M (eds) *Language and Space*, pp. 277–316. Cambridge, MA: MIT Press.

O'Keefe J and Burgess N (1999) Theta activity, virtual navigation and the human hippocampus. *Trends in Cognitive Sciences* **3**: 403–406.

O'Keefe J and Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* **34**: 171–175.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

O'Keefe J and Recce ML (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**: 317–330.

Orr G, Rao G, Houston FP, McNaughton BL and Barnes CA (2001) Hippocampal synaptic plasticity is modulated by theta rhythm in the fascia dentata of adult and aged freely behaving rats. *Hippocampus* **11**: 647–654.

Packard MG and McGaugh JL (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory* **65**: 65–72.

Rudy JW and Sutherland RJ (1995) Configural association theory and the hippocampal formation: an appraisal and reconfiguration. *Hippocampus* **5**: 375–389.

Save E, Poucet B, Foreman N and Buhot MC (1992) Object exploration and reactions to spatial and nonspatial changes in hooded rats following damage to parietal cortex or hippocampal formation. *Behavioral Neuroscience* **106**: 447–456.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21.

Spiers HJ, Burgess N, Hartley T, Vargha-Khadem F and O'Keefe J (2001a) Bilateral hippocampal pathology impairs topographical and episodic but not recognition memory. *Hippocampus* **11**: 715–725.

Spiers HJ, Burgess N, Maguire EA *et al*. (2001b) Unilateral temporal lobectomy patients show lateralised topographical and episodic memory deficits in a virtual town. *Brain* **124**: 2476–2489.

Squire LR and Zola-Morgan S (1991) The medial temporal lobe memory system. *Science* **253**: 1380–1386.

Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review* **91**: 195–231.

Sutherland RJ, McDonald RJ, Hill CR and Rudy JW (1989) Damage to the hippocampal formation in rats selectively impairs the ability to learn cue relationships. *Behav Neural Biology* **52**: 331–356.

Taube JS, Muller RU and Ranck JB (1990) Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience* **10**: 420–435.

Vargha-Khadem F, Gadian DG, Watkins KE *et al*. (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–380.

Winocur G (1990) Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioral Brain Research* **38**: 145–154.

Xavier GF, Stein C and Bueno OF (1990) Rats with dorsal hippocampal lesions do react to new stimuli but not to spatial changes of known stimuli. *Behav Neural Biology* **54**: 172–183.

Zola-Morgan S and Squire LR (1985) Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia. *Behavioral Neuroscience* **99**: 22–34.

Zola-Morgan S, Squire LR and Amaral DG (1986) Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. *Journal of Neuroscience* **6**: 2950–2967.

Zola-Morgan S, Squire LR, Rempel NL, Clower RP and Amaral DG (1992) Enduring memory impairment in monkeys after ischemic damage to the hippocampus. *Journal of Neuroscience* **12**: 2582–2596.

Hodgkin–Huxley

Intermediate article

Alwyn Scott, University of Arizona, Tucson, Arizona, USA

CONTENTS

Quantitative representation of transmembrane ionic currents by the Hodgkin–Huxley equations allows the calculation of nerve-impulse dynamics, including all-or-nothing behavior, threshold, temperature dependence, decremental conduction and the nature of the refractory zone.

INTRODUCTION

In 1952, a detailed study of ionic current flow through the membrane of the giant axon of the squid (*Loligo*) was presented by Hodgkin and Huxley, culminating in a mathematical formulation for the dynamics underlying a nerve impulse. In addition to allowing computations of the ionic permeability and transmembrane voltage of a propagating impulse, their formulation predicts several new features, including an unstable impulse of lower amplitude and speed, threshold conditions for launching an impulse, temperature dependence of propagation, the effects of 'narcotizing' an axon by reducing the maximum ionic conductivities, and the nature of the refractory zone following the passage of an impulse.

SPACE AND VOLTAGE CLAMPING

When conducting their research, Hodgkin and Huxley used the techniques of *space clamping* and *voltage clamping* to characterize a localized patch of squid membrane. Although these terms sound alike, they describe quite different experimental techniques (*See* **Single Neuron Recording**)

Geometrically, a squid nerve is a cylindrical tube of conductive axoplasm encased in a lipid-bilayer membrane, as shown in Figure 1(a). Variations in transmembrane voltage with distance along the structure can be eliminated by inserting a conducting wire longitudinally through the nerve as a terminal for the internal voltage. With such space clamping, the total current per unit area flowing through the membrane is given by an expression of the form

$$\frac{\text{total membrane current}}{\text{membrane area}} = J_{\text{ion}} + C\frac{dV}{dt} \qquad (1)$$

where the first term on the right-hand side represents the transmembrane ionic current per unit area through the membrane, and the second term is the displacement current per unit area through the capacitance of the lipid bilayer.

The term 'voltage clamping', on the other hand, implies the use of a *negative feedback amplifier* to hold the potential difference across a nerve membrane at a fixed value, unaffected by flows of ionic currents.

IONIC CURRENTS THROUGH A PATCH OF SQUID MEMBRANE

Using the techniques of space clamping and voltage clamping, Hodgkin and Huxley were able to measure individually the dynamics of sodium and potassium ion currents. For example, the sodium current consists of two components, namely a *conduction current* that flows in response to the transmembrane voltage difference, and a *diffusion current* that flows in response to the transmembrane difference in sodium ion concentrations on the two sides of the membrane. These two components are proportional to each other because the transmembrane mobility μ_{Na} (which determines the sodium conduction current) is related to the transmembrane diffusion constant D_{Na} (which determines the sodium diffusion current) by the *Einstein relation*, $D_{\text{Na}} = kT\mu_{\text{Na}}/q$, where k is the Boltzmann constant, T is the absolute temperature and q is the electronic charge. In other words:

$$\text{sodium current} = \widetilde{G}_{\text{Na}}(V,t)(V - V_{\text{Na}}) \quad \text{and} \qquad (2)$$

$$\text{potassium current} = \widetilde{G}_{\text{K}}(V,t)(V - V_{\text{K}}) \qquad (3)$$

where $\tilde{G}_{\text{Na}}(V,t)$ and $\tilde{G}_{\text{K}}(V,t)$ are sodium and potassium conductances, and the first and second

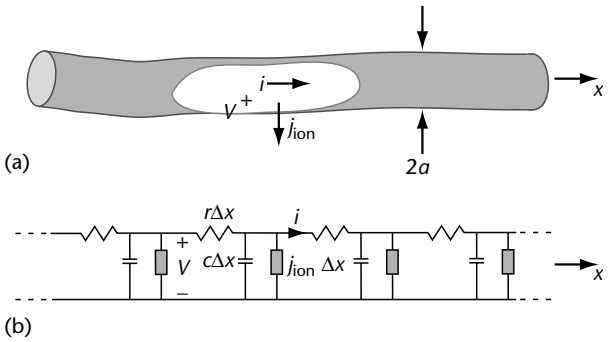


Figure 1. (a) Sketch of a squid axon. (b) A corresponding differential circuit diagram that can be used to derive the cable equation for impulse propagation.

terms in parentheses represent conduction and diffusion current, respectively.

If the membrane is at rest (steady state), the sum of the total sodium and potassium currents must be zero, leading to a *resting potential* V_R which is about 65 mV negative inside the axon with respect to the outside. Measuring the transmembrane voltage V with respect to the resting potential, and counting an increase in the inside voltage as positive, it follows from the Einstein relationship that

$$V_{Na} = 25 \log\left(\frac{[Na^+]_o}{[Na^+]_i}\right) - V_R \tag{4}$$

where $kT/q = 25$ mV at room temperature, $[Na^+]_o$ and $[Na^+]_i$ are the outside and inside concentrations of sodium ions, respectively, and similarly for the potassium ions. Normally $[Na^+]_o > [Na^+]_i$, whereas $[K^+]_o < [K^+]_i$; thus sodium ions diffuse into the axon, whereas potassium ions diffuse out of it.

To determine the individual dynamics of sodium and potassium ions at several fixed voltages, Hodgkin and Huxley first measured the total membrane current as a function of time at those voltages. They then reduced the external sodium concentration ($[Na^+]_o$) to approximately zero, effectively eliminating the inward sodium current and allowing measurements of the potassium current at the same voltages. Finally, subtraction of the second measurements from the first gave the corresponding sodium currents.

When this procedure was completed, it was discovered that the sodium conductance (or permeability) initially rises to a maximum value in less than a millisecond and then decays back to approximately zero over several milliseconds. Thus the dynamics of the sodium ion current at voltage V are represented by the formula

$$J_{Na}(V) = G_{Na} m^3(V,t) h(V,t)(V - V_{Na}) \tag{5}$$

where G_{Na} is a maximum sodium conductance per unit area, m is a 'sodium turn-on' variable and h is a 'sodium turn-off' variable.

Since the potassium conductance was observed to rise without falling back to zero, the dynamics of the potassium ion current at voltage V are represented by the formula

$$J_K(V) = G_K n^4(V,t)(V - V_K) \tag{6}$$

where n is a 'potassium turn-on' variable.

The total ionic current per unit area across the membrane, as expressed in Equation (1), then becomes

$$J_{ion} = G_{Na} m^3 h(V - V_{Na}) + G_K n^4 (V - V_K) + G_L(V - V_L) \tag{7}$$

where the last term is a small 'leakage current' accounting for ionic current missed by the measurements of sodium and potassium components. The empirical relationships governing the dynamics of m, h and n are as follows.

The membrane turn-on and turn-off variables are solutions of first-order rate equations with voltage-dependent parameters. Thus:

$$\begin{aligned} \frac{dm}{dt} &= \alpha_m(1-m) - \beta_m m \\ \frac{dh}{dt} &= \alpha_h(1-h) - \beta_h h \\ \frac{dn}{dt} &= \alpha_n(1-n) - \beta_n n \end{aligned} \tag{8}$$

where at 6.3°C the voltage dependencies of the coefficients are given by the following:

$$\alpha_m = \frac{0.1(25 - V)}{\exp[(25 - V)/10] - 1}$$
$$\beta_m = 4\ \exp(-V/18)$$
$$\alpha_h = 0.07\ \exp(-V/20)$$
$$\beta_h = \frac{1}{\exp[(30 - V)/10] + 1}$$
$$\alpha_n = \frac{0.01(10 - V)}{[\exp(10 - V)/10] - 1}$$
$$\beta_n = 0.125\ \exp(-V/80) \quad (9)$$

in units of milliseconds^{-1} with V measured in millivolts. At other temperatures, changes in these rates are accounted for by multiplying by the following factor

$$\kappa = 3^{(\text{temp}-6.3)/10} \quad (10)$$

where 'temp' is the temperature in centigrade.

The parameters in Equations (1) and (7) are shown in Table 1, where the standard value of V_L has been selected to make the steady-state value of $J_{ion} = 0$ at $V = 0$.

SPACE-CLAMPED ACTION POTENTIALS

Suppose that a length of squid axon is space clamped, and to simplify the arithmetic take this area to be 1 cm^2. Then from Equation (1) the basic equation governing the dynamics of membrane voltage is

$$\frac{dV}{dt} = \frac{I_0(t) - J_{ion}}{C} \quad (11)$$

where $I_0(t)$ is a current injected into the axon by the experimenter and J_{ion} is given in Equation (7).

A short pulse of injected current will begin to charge the membrane capacitance, raising the transmembrane voltage from its resting value ($V = 0$) to a threshold level. Near threshold, a positively charged sodium ion current flows into the axon, further increasing V, which rises rapidly to approximately V_{Na}. At this point, the sodium current falls to zero, and the potassium current begins to carry positive charge out of the axon, causing V to relax slowly back to the resting value. This cycle of activity is termed a space-clamped *action potential* or spike.

To compute these dynamics, Hodgkin and Huxley integrated Equation (11), together with Equations (7) and (8), for a variety of experimental conditions, demonstrating quantitative agreement with measurements of $V(t)$ on space-clamped squid membranes.

Table 1. Parameter values measured by Hodgkin and Huxley (1952) for the giant axon of the squid

Parameter	*Mean*	*Range*	*Standard*	*Units*
C	0.91	(0.8–1.5)	1.0	μF/cm^2
G_{Na}	120	(65–260)	120	mmhos/cm^2
G_K	34	(26–49)	36	mmhos/cm^2
G_L	0.26	(0.13–0.5)	0.3	mmhos/cm^2
V_{Na}	+109	(95–119)	+115	mV
V_K	−11	(9–14)	−12	mV
V_L	+11	(4–22)	+10.5995	mV

NERVE-IMPULSE PROPAGATION

Consider next how this local switching activity propagates along a nerve that is not space clamped, with the following parameters (Figure 1(b)):

- r is the longitudinal resistance per unit length of the fiber, measured in units of ohms per centimeter. For a cylindrical axon of radius a, $r = \rho/\pi a^2$, where ρ is the resistivity of the cytoplasm in ohm-centimeters.
- c is the membrane capacitance per unit length of the fiber, measured in units of farads per centimeter. For a cylindrical fiber of radius a, $c = 2\pi aC$, where C is the capacitance per unit area of the membrane in farads per square centimeter.
- j_{ion} is the ionic current flowing across the membrane (from inside to outside) per unit length of the axon, measured in units of amperes per centimeter. For a cylindrical fiber of radius a, $j_{ion} = 2\pi a J_{ion}$, where J_{ion} is the transmembrane current per unit area given in Equation (7).

Applying Kirchhoff's circuit laws to the differential circuit diagram shown in Figure 1(b) in the limit $\Delta x \to 0$ yields the following *nonlinear diffusion equation*:

$$\frac{\partial^2 V}{\partial x^2} - rc\frac{\partial V}{\partial t} = rj_{ion} \quad (12)$$

from which emerges the nerve impulse. A solution of Equation (12) that represents nerve impulse is a *traveling wave* for which

$$V(x,t) = \widetilde{V}(x - vt) = \widetilde{V}(\xi) \quad (13)$$

with the dependencies on x and t constrained by the variable

$$\xi \equiv x - vt \quad (14)$$

where v is the propagation velocity of the traveling wave.

This assumption implies that the partial derivatives with respect to x and t in Equation (12) are related as $\partial/\partial x = (-1/v)\partial/\partial t = \partial/\partial\xi$. Thus

Equation (12) – a partial differential equation (PDE) – is reduced under the traveling wave assumption to the ordinary differential equation (ODE):

$$\frac{d^2\widetilde{V}}{d\xi^2} + rcv\frac{d\widetilde{V}}{d\xi} = rj_{\text{ion}} \tag{15}$$

Although Equation (15) is easier to solve than Equation (12), it contains less information because the impulse speed v is an undetermined parameter. (*See* **Diffusion Models and Neural Activity**)

In attempting to solve Equation (15), Hodgkin and Huxley were faced with two problems. (1) In order to determine the traveling wave speed v, it was necessary to carry through many tedious integrations without the benefit of an electronic computer. (2) The parameter values for real squid nerves vary over a rather wide range, as indicated in Table 1.

To limit the number of integrations, they selected one axon, with an internal (axoplasmic) resistivity (ρ) of 35.4 ohm-cm, a radius (a) of 0.0238 cm, and the 'standard' membrane values indicated in Table 1, for detailed numerical analysis. For this particular axon, the parameters of Equation (15) are given in Table 2. Since the action potential was measured at 18.5°C, the numerical calculations of the impulse dynamics were also performed for this temperature.

As shown in Figure 2, the shape of the traveling-wave solution calculated for this nerve was found to be in good qualitative agreement with experimental observations, and the calculated impulse velocity was $18.8\,\text{ms}^{-1}$, whereas the measured speed was $21.2\,\text{ms}^{-1}$, again in substantial agreement. Therefore, like a celebrity who is 'famous for being well known', this nerve has become widely recognized and studied as the *standard Hodgkin–Huxley axon*, defined by the parameters listed in Tables 1 and 2.

Although it was impractical for Hodgkin and Huxley to perform the necessary calculations in the early 1950s, there are in fact *two* impulse solutions at two different values for the traveling-wave speed, which are shown in Figure 2. In this figure, the higher-amplitude solution (at $v = 18.8\,\text{ms}^{-1}$) corresponds to the experimentally observed impulse.

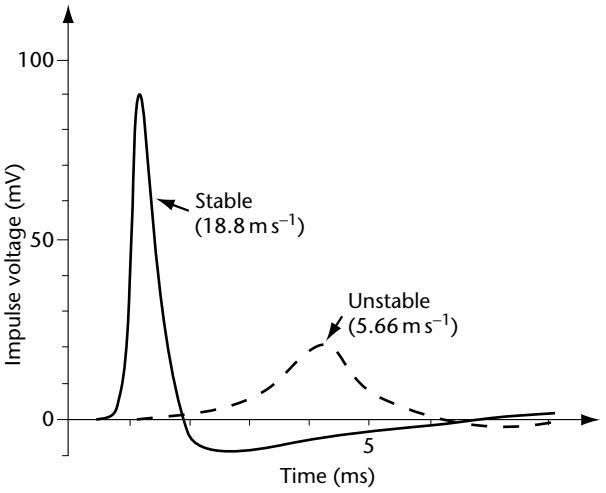


Figure 2. A full-sized impulse (at $v = 18.8\,\text{ms}^{-1}$) and an unstable threshold impulse (at $5.66\,\text{ms}^{-1}$) for the standard Hodgkin–Huxley axon at 18.5°C. (From data in Hodgkin and Huxley (1952) and Huxley (1959)).

The smaller-amplitude traveling-wave solution, with a speed of $5.66\,\text{ms}^{-1}$, was found by Huxley in 1959 using an electronic computer. This solution is unstable in the sense that deviations from it diverge with increasing time. Slightly smaller solutions decay to zero and slightly larger solutions grow to become the fully developed nerve impulse. Thus this unstable solution defines *threshold conditions* for igniting an impulse (Scott, 1973).

Table 2. Standard Hodgkin–Huxley parameters for the giant axon of the squid

Parameter	*Value*	*Units*
ρ	35.4	ohm-cm
a	0.0238	cm
r	2.0×10^4	ohms/cm
c	1.5×10^{-7}	F/cm

DEGRADATION OF A NERVE IMPULSE

In 1966, Cooley and Dodge described the effects of 'narcotizing' a Hodgkin–Huxley fiber by reducing the maximum sodium and potassium conductances by a *narcotization factor*, $\eta < 1$. Thus

$$G_{\text{Na}} \to \eta G_{\text{Na}} \text{ and } G_{\text{K}} \to \eta G_{\text{K}}$$

are the values used in numerical computations.

One qualitative effect of this change is to reduce the speed of the larger-amplitude stable traveling-wave solution as shown in Figure 3. Conversely, the speed of the smaller-amplitude unstable traveling-wave solution is increased. Eventually, a critical point is reached (at $\eta_c = 0.261$) where the two solutions merge together. For yet smaller values of η, no traveling-wave solutions exist. Along the locus shown in Figure 3, a condition of power balance obtains, at which the rate of energy release by the impulse is equal to its dissipation rate.

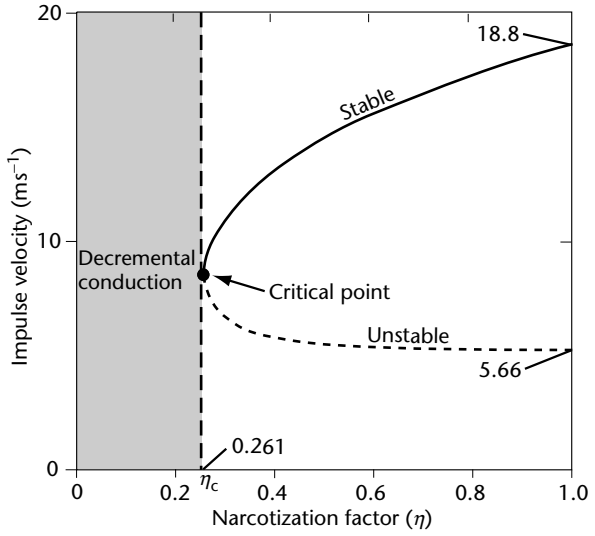


Figure 3. Power balance loci, showing impulse speeds for the Hodgkin–Huxley equations as a function of a 'narcotization factor' (η). (From data in Cooley and Dodge (1966)).

Although Adrian's concept of all-or-nothing propagation holds for $\eta > \eta_c$, its logical basis evaporates for $\eta < \eta_c$. However, in this regime one can find *decremental* propagation of a nerve impulse, where the rate at which energy is generated is only slightly less than the rate of dissipation, so the impulse amplitude relaxes rather slowly to zero. As was emphasized by Lorente de Nó and Condouris in 1959, this phenomenon was long overlooked by electrophysiologists, who had concentrated their attentions on the properties of standard nerves.

The qualitative conclusions stemming from the computations of Cooley and Dodge apply to several other situations in which the ability of a nerve to conduct impulses is degraded, including the following: (1) increasing external concentrations of potassium ions (Adelman and FitzHugh, 1975); (2) higher temperatures, at which the temporal rates of membrane conductance increase according to Equation (10), thereby shortening the potassium turn-on time (Cole, 1968); (3) increasing the leakage conductance (Cole, 1968); (4) abruptly increasing the cross-sectional area, which has implications for information processing at the axonal or dendritic branchings of a neuron (Berkinblit *et al.*, 1970); (5) finally, the propagation of a periodic train of impulses is degraded when the interval between individual pulses becomes less than a certain value, because each impulse then propagates in the *refractory zone* of the preceding impulse.

REFRACTORY AND ENHANCEMENT ZONES

The problems encountered by a nerve impulse that follows too closely on the heels of another have been studied empirically using *double-impulse* measurements (Donati and Kunov, 1976; Scott and Vota-Pinardi, 1982). In such experiments, pairs of impulses are launched on single fibers, and the ratio of impulse speeds is measured as a function of the impulse interval (T).

The lower part of Figure 4 shows measurements of the ratio

$$R(T) \equiv \frac{\text{speed of second impulse}}{\text{speed of first impulse}}$$

for impulses propagating on an axon of the squid (*Lologo vulgaris*). A significant feature of this ratio is the impulse interval $T = T_1$ at which $R(T) = 1$.

In terms of T_1, four regions can be distinguished. (1) *Absolute refractory zone*: for impulse intervals less than $T < 0.4T_1$, it is impossible for the second impulse to propagate; thus $R(T) = 0$. (2) *Relative refractory zone*: for impulse intervals in the range $0.4T_1 < T < T_1$, the second impulse is able to propagate, albeit with diminished speed. (3) *Enhancement zone*: in the range $T_1 < T < 1.8T_1$, the second impulse is observed to travel *faster* than the first impulse. (4) *Uncoupled zone:* if the impulse interval is greater than about $1.8T_1$, both impulses travel at the same speed, with the second impulse uninfluenced by the first. Data from 13 squid axons at temperatures between 14.4°C and 20.4°C give the relationship $T_1 = 58.2 \times 3^{-\text{Temp}/10}$ ms ($\pm 13\%$), in agreement with Equation (10) (Scott and Vota-Pinardi, 1982).

The upper part of Figure 4 is an oscilloscope photograph of the trailing edge of a single squid impulse, into which a second impulse would attempt to propagate. T_0 is the time between the point of maximum slope on the leading edge of $V(t)$ (occurring at about 50 mV, this is off the scale of the figure) and the point on the trailing edge where $V(t)$ crosses over from being hyperpolarizing (more negative than the resting value) to being depolarizing. Both T_0 and T_1 have been measured on the above cohort of 13 axons, showing that $T_0 \approx T_1$ to within experimental errors of about $\pm 15\%$.

The qualitative similarity of these two measurements in the neighborhood of the crossover points indicates that the enhancement zone stems from the depolarizing phase on the trailing edge of the first impulse. From a physical perspective, a depolarizing phase occurs because near its resting level a nerve membrane is oscillatory.

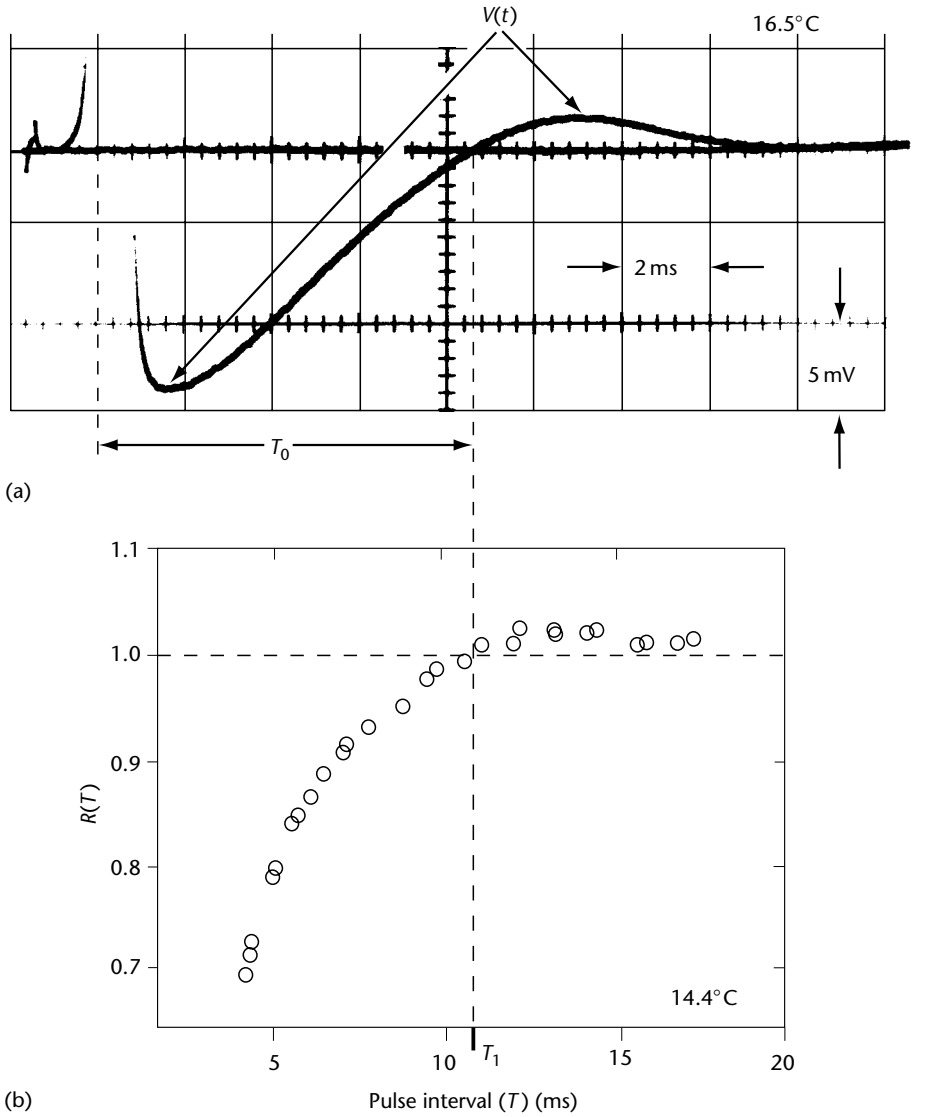


Figure 4. Two data sets related to the refractory zones in the wake of a propagating impulse on a squid axon. (a) An oscilloscope photograph showing the oscillatory tail of a single action potential $V(t)$. (b) Double impulse measurements, where T is the time interval between impulses and $R(T)$ is the ratio of the speeds of the two impulses. (Note that the crossover points in the two data sets are aligned, and the timescales have been adjusted for different temperatures.)

Discovered by Cole and Baker (1941) and noted by Hodgkin and Huxley (1952), subthreshold membrane oscillations of the Hodgkin–Huxley equations were investigated numerically by Sabah and Liebovic (1969), and both experimentally and numerically by Mauro *et al*. (1970). These studies show that a patch of squid membrane near its resting voltage has a damped resonance (Q–3) at around 50–100 Hz, depending on the temperature. (*See* **Neural Oscillations**)

References

Adelman WJ Jr and FitzHugh R (1975) Solutions of the Hodgkin–Huxley equations modified for potassium accumulation in a periaxonal space. *Federation Proceedings* **34**: 1322–1329.

Berkinblit MB, Vvedenskaya ND, Gnedenko LS *et al*. (1970) Computer investigation of the features of conduction of a nerve impulse along fibers with different degrees of widening. *Biophysics* **15**: 1121–1130.

Cole KS (1968) *Membranes, Ions and Impulses: A Chapter of Classical Biophysics*. Berkeley, CA: University of California Press.

Cole KS and Baker RF (1941) Longitudinal impedance of the squid giant axon. *Journal of General Physiology* **24**: 771–788.

Cooley JW and Dodge FA (1966) Digital computer solutions for excitation and propagation of the nerve impulse. *Biophysical Journal* **6**: 583–599.

Donati F and Kunov H (1976) A model for studying velocity variations in unmyelinated axons. *Institute of Electrical and Electronic Engineers Transactions on Biomedical Engineering* **BME-23**: 23–28.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* **117**: 500–544.

Huxley AF (1959) Can a nerve propagate a subthreshold disturbance? *Journal of Physiology* **148**: 80P–81P.

Lorento de Nó R and Condouris GA (1959) Decremental conduction in peripheral nerve: integration of stimuli in the neuron. *Proceedings of the National Academy of Sciences of the USA* **45**: 593–617.

Mauro A, Conti F, Dodge FA and Schor R (1970) Threshold behavior and phenomenological impedance of the squid giant axon. *Journal of General Physiology* **55**: 497–523.

Sabah NH and Liebovic KN (1969) Subthreshold responses of the Hodgkin–Huxley cable model for the squid giant axon. *Biophysical Journal* **9**: 1206–1222.

Scott AC (1973) Strength duration curves for threshold excitation of nerves. *Mathematical Biosciences* **18**: 137–152.

Scott AC and Vota-Pinardi U (1982) Velocity variations on unmyelinated axons. *Journal of Theoretical Neurobiology* **1**: 150–172.

Further Reading

Cole KS (1968) *Membranes, Ions and Impulses: A Chapter of Classical Biophysics*. Berkeley, CA: University of California Press.

Jack JJB, Noble D and Tsien RW (1975) *Electric Current Flow in Excitable Cells*. Oxford: Clarendon Press.

Keener J and Sneyd J (1998) *Mathematical Physiology*. New York: Springer-Verlag.

Khodorov BI (1974) *The Problem of Excitability: Electrical Excitability and Ionic Permeability of the Nerve Membrane*. New York: Plenum Press.

Koch C (1999) *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press.

Leibovic KN (1972) *Nervous System Theory: An Introductory Study*. New York: Academic Press.

Scott AC (1975) The electrophysics of a nerve fiber. *Reviews of Modern Physics* **47**: 487–533.

Scott AC (2002) *Neuroscience: A Mathematical Primer*. New York: Springer-Verlag.

Hormones, Learning and Memory

Introductory article

Joe L Martinez Jr, University of Texas, San Antonio, Texas, USA
William J Meilandt, University of Texas, San Antonio, Texas, USA
Haixiang Peng, University of Texas, San Antonio, Texas, USA
Edwin J Barea-Rodriguez, University of Texas, San Antonio, Texas, USA

CONTENTS

Hormones modulate learning and memory processes by increasing or decreasing the strength with which information is stored.

INTRODUCTION

Memory is often defined as the ability to acquire, store, and retrieve new information. One's memory of one's first teacher at school is represented in the brain by an interconnected network of neurons. If a memory is with you for life, then some permanent change took place in your brain that represents the memory. This usually involves the strengthening of synapses within a specific neural network for a given memory. Hormones do not affect the memory trace directly; instead, they influence modulatory systems that in turn regulate associative strength, thereby making memories stronger or weaker. Hormones are considered to be memory modulators because different doses of a hormone can have opposing effects on memory.

Different forms of memory (declarative, procedural, spatial, etc.) are often divided into two groups: those that depend on the hippocampus, and those that do not. Although the mechanisms for memory storage may differ for different kinds of memory, as a memory modulator, hormones can influence multiple forms of memory by enhancing or impairing learning. Hormones can modulate:

- acquisition – the ability to learn a task,
- consolidation – the transition of information from short-term to long-term memory,
- retention – the ability to remember information,
- extinction – the ability to decrease a behavioral response to conditioning because of a lack of reinforcement,
- retrieval – the ability to recall information.

THE NEUROENDOCRINE SYSTEM

Why are extremely emotional or stressful events remembered so vividly? For instance, who in America will ever forget the events that took place on 11 September 2001? Moments such as these activate our autonomic and neuroendocrine system, resulting in the release of hormones that elicit an increase in heart rate, sweating, nausea, and anxiety. The hormonal release and the ensuing cellular and physiological responses in the brain can have a dramatic influence on the strength with which an event is to be remembered.

Methods of Communication

Endocrine communication occurs when endocrine glands secrete their chemical messengers, hormones such as the glucocorticoids, into the bloodstream. These hormones, in turn, activate target cells throughout the body. Only cells containing the specific receptor will be activated and remain activated until hormone levels decrease. Autocrine and paracrine forms of communication occur when endocrine cells secrete hormones that act on the releasing cell or neighboring cells, respectively. Some chemical messengers, such as noradrenaline (norepinephrine), can act both as a hormone (when released from the adrenal gland) and as a neurotransmitter (when released from neurons in the brain).

Endocrine glands were previously thought to be the only source of hormone secretion into the bloodstream. However, in 1928 Ernst Scharrer discovered that certain neurons in the hypothalamus secrete chemical messengers, termed neurohormones, into the blood, thus establishing the field of neuroendocrinology.

The mechanisms by which the nervous and endocrine systems communicate are displayed in Figure 1. Neurons secrete chemical messengers, known as neurotransmitters, into a specialized region between neighboring neurons called the synapse. Neurotransmitters are modified amino acids that cross the synapse and rapidly activate specific receptors (proteins located on the cell surface), which increase or decrease the activity of the postsynaptic target cell. In addition to neurotransmitters, neurons can also contain and release neuropeptides, such as adrenocorticotrophic hormone (ACTH). Neuropeptides are peptide hormones derived from larger precursor proteins that act as neuromodulators to regulate the activity of neurons. A second method of neuronal communication is neuroendocrine secretion. Specialized neurons, termed neurosecretory cells, release neurohormones into the bloodstream to activate distant target cells; for example, neurosecretory

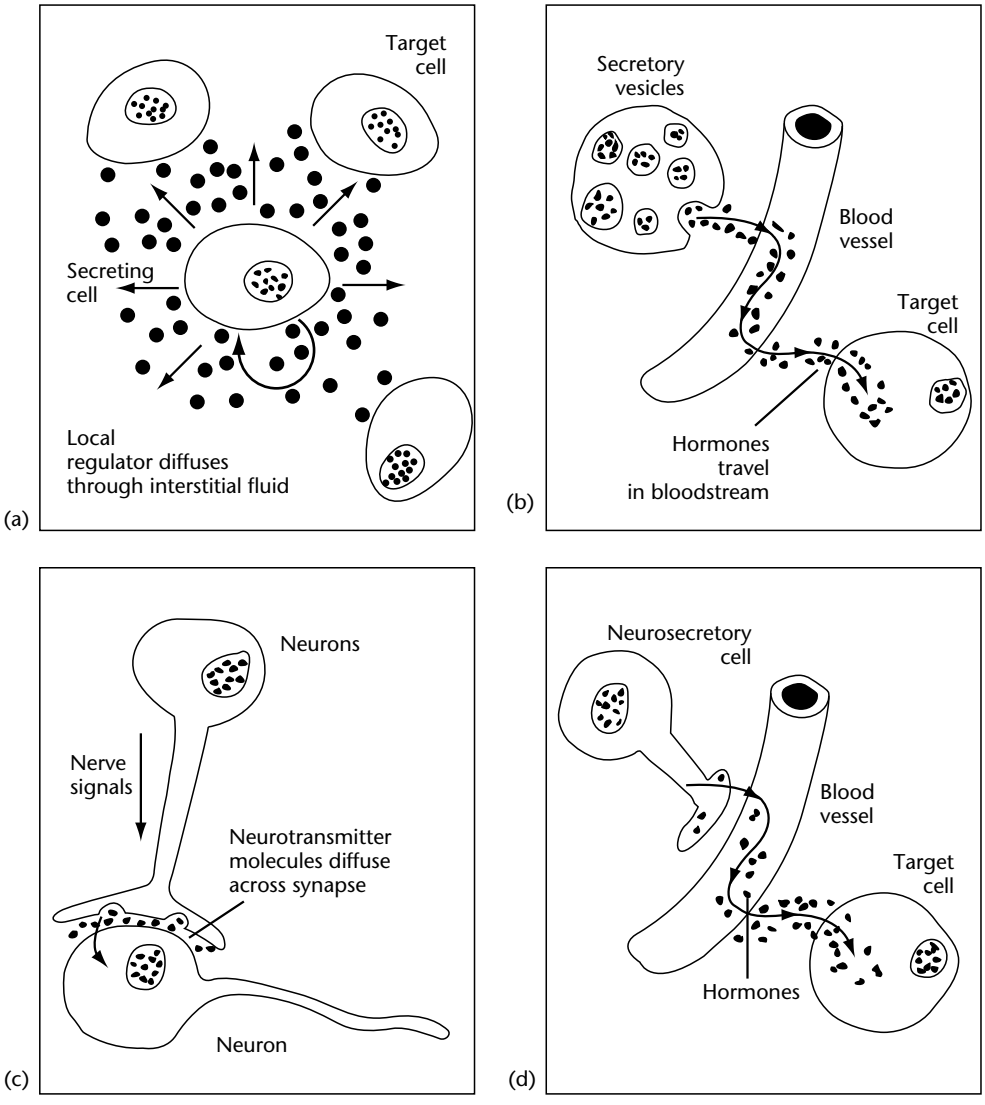


Figure 1. Endocrine and neuronal systems of communication. The endocrine systems of communication include (a) autocrine and paracrine forms in which secretory cells release hormones that act on the releasing cell or adjacent cells, and (b) endocrine forms in which cells release hormones into the bloodstream, activating target cells throughout the body. Neurons communicate by (c), releasing neurotransmitters across the synapse to activate neighboring neurons. Neuroendocrine communication (d) occurs when neurosecretory cells release their hormones into the bloodstream to activate distant target cells. Adapted from Waltz L (2001) *Environmental Estrogens and Other Hormones*. Website of the Center for Bioenvironmental Research at Tulane and Xavier Universities (http://www.som.tulane.edu/ecme/eehome/basics /endosys/hormones.html).

cells of the posterior pituitary secrete vasopressin to activate kidney cells.

The Endocrine Glands

An endocrine gland is a ductless organ that secretes hormones into the bloodstream. The release of hormones from these endocrine glands is monitored and controlled by the pituitary gland, which in turn is controlled by the hypothalamus (Figure 2). The posterior pituitary is not a gland but is actually an extension of the brain that contains nerve terminals which release hormones into the bloodstream.

Let us return to the question of why extremely emotional or stressful events are remembered so vividly. It has been found that during particularly stressful events, the hypothalamic–pituitary–adrenal (HPA) axis is activated, creating behavioral and physiological responses to the stressor that can alter the strength with which memories are stored (Figure 3). Incoming sensory input (sight, sound, and pain) is relayed to higher-order processing centers in the brain, which in turn release neurotransmitters activating hypothalamic neurosecretory cells to release a hormone called corticotrophin releasing hormone (CRH). This hormone travels through the portal blood supply (connecting the hypothalamus with the anterior pituitary) to stimulate cells of the anterior pituitary to release ACTH into the bloodstream. This ACTH activates the adrenal cortex to secrete glucocorticoids such as cortisol, which bind to target cells throughout the periphery and brain. The released glucocorticoids can readily reenter the brain and provide negative feedback to the hypothalamus and pituitary to prevent further release of CRH and ACTH when glucocorticoid levels become too high. Activation of the autonomic nervous system during stressful events additionally stimulates the adrenal medulla to release adrenaline (epinephrine) and noradrenaline (norepinephrine) into the bloodstream. The released glucocorticoids and noradrenaline then interact with memory mechanisms to mark the importance of the event, as described below.

HORMONAL INFLUENCE ON LEARNING AND MEMORY

The influence of hormones on memory often follows a U-shaped curve. For example, when an

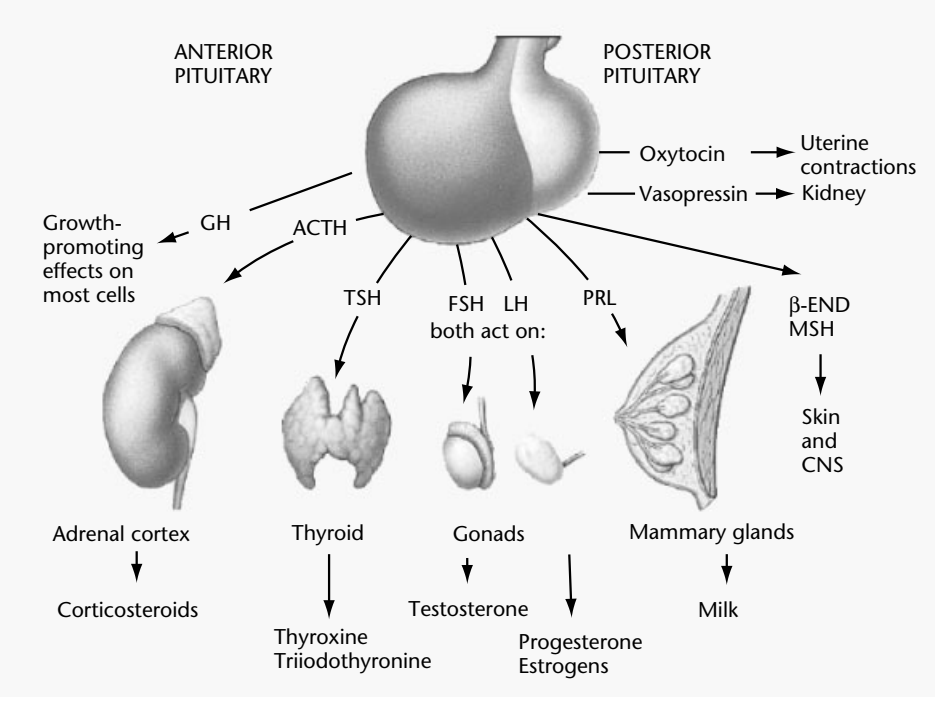


Figure 2. Anterior and posterior pituitary hormones travel through the bloodstream activating target cells throughout the body. ACTH, adrenocorticotrophic hormone; CNS, central nervous system; β-END, β-endorphin; FSH, follicle stimulating hormone; GH, growth hormone; LH, luteinizing hormone; MSH, melanocyte stimulating hormone; PRL, prolactin; TSH, thyroid stimulating hormone. Adapted from Starr and Taggart (2001) *Biology: The Unity and Diversity of Life*, 9th edn.

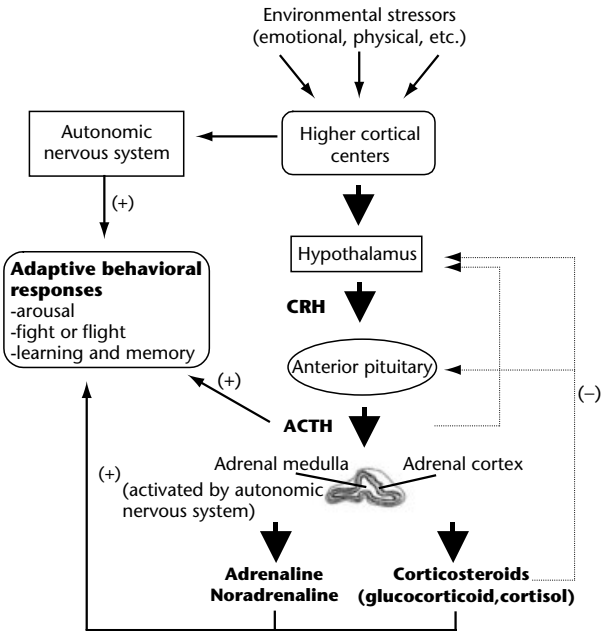


Figure 3. Activation of the hypothalamic–pituitary–adrenal axis during stressful events produces adaptive behavioral responses following the release of corticosteroids and the catecholamines, adrenaline and noradrenaline. Glucocorticoids can provide negative feedback (–) to the hypothalamus and pituitary to prevent further release of hormones. ACTH, adrenocorticotrophic hormone; CRH, corticotrophin releasing hormone.

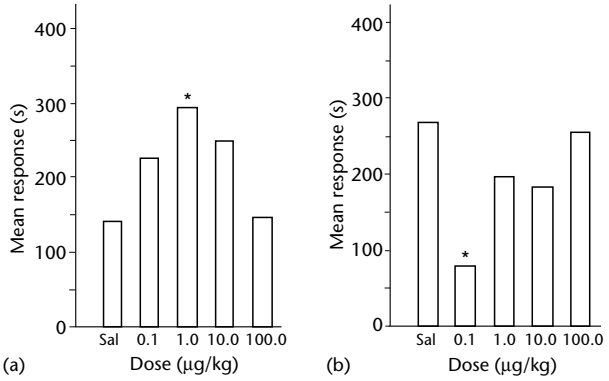


Figure 4. Hormones characteristically produce U-shaped dose–response curves. Some hormones produce either an (a) inverted-U or (b) U-shaped dose–response curve for inhibitory avoidance behavior. Injections of intermediate doses of a hormone are usually the most effective (asterisk) in either (a) enhancing or (b) impairing memory when compared with saline (Sal) control-treated animals.

animal receives a low, intermediate, or high dose of a hormone, only the intermediate dose affects learning and memory (Figure 4). There can also be a biphasic response, where the low and high doses impair memory and the intermediate dose enhances it, or vice versa. These behavioral responses are the hallmark of neuromodulators.

This section will look at three types of behavioral training that require certain limbic structures, such as the hippocampus and amygdala. The first uses food as a motivation to learn a maze task, to study appetitive learning. The second task requires the animal to move actively away from a particular location, or passively 'freeze' when placed in a chamber, to avoid receiving a foot shock, to study active and passive avoidance learning respectively. The final task requires animals to use spatial cues to navigate through a maze either on land or in water, to study spatial learning. The effects of a given hormone will depend on the strength of training (e.g. the intensity of foot shock), the hormone dose, the time of administration (before or after training), and the type of memory being tested.

ACTH

The release of ACTH from the anterior pituitary and subsequent release of glucocorticoids from the adrenal cortex are a critical component of the endocrine and metabolic response to stress. The release of ACTH is also important for behavioral adaptations (learning and memory) to stressful or arousing experience. As early as 1955 it was demonstrated that surgical removal of the pituitary gland – specifically the anterior pituitary – impairs active avoidance learning. However, if these animals were given ACTH, there was no learning impairment. In rodents, systemic injections of ACTH facilitate the acquisition of active and passive avoidance learning and enhance retention of passive avoidance response. This hormone also prolongs extinction of active avoidance behavior and appetitively motivated behavior in a T maze. The amount (or dose) of ACTH injected into the animals will determine whether enhancement, impairment, or no effect is seen. In addition, the same dose of ACTH may either enhance or impair memory, depending on the intensity of the foot shock used to train the animals. Together, these results highlight the modulatory nature of ACTH on learning and memory processes.

It is generally agreed that ACTH cannot cross the blood–brain barrier. So, does ACTH modulate memory directly by activating the central nervous system, or indirectly by activating the adrenal cortex? The answer is likely to be both. The hormone must work through the central nervous system, because ACTH restores performance in

animals that have had their pituitary and adrenal glands removed. Analogues of ACTH that can cross the blood–brain barrier, but do not stimulate adrenal cortex to release glucocorticoids, can mimic the effects of ACTH on conditioning. Furthermore, ACTH receptors are abundant in regions of the limbic system (hippocampus and amygdala) classically associated with learning and emotion. Lesions of these structures block the effects of ACTH analogues on extinction of conditioned avoidance responses.

Evidence suggests that systemic administration of ACTH exerts its effects, at least in part, through peripheral mechanisms. First, it is unknown whether ACTH and its analogues influence learning and memory processes through the same mechanisms. Second, while lesion studies suggest that an intact limbic system is necessary for the effects of ACTH analogues on extinction, this does not prove that these effects are due to the binding of ACTH to receptors within the limbic system. The effects of ACTH on conditioning may be initiated outside the blood–brain barrier, but the limbic system probably is required for the expression of these effects. Finally, glucocorticoids also have modulatory effects on learning and memory processes (discussed below). Together, these findings suggest that the effects of systemically administered ACTH on learning and memory could be initiated at sites in the brain or periphery.

Glucocorticoids

Glucocorticoids (corticosterone in rodents, and cortisol in humans) are steroid hormones, derived from cholesterol, that are secreted by the adrenal cortex into the general circulation. Glucocorticoids often take several minutes to be released in response to a stressor, and several hours for their effects to emerge. This is generally viewed as the second wave of the autonomic response that follows the rapid release of catecholamines, which trigger second-messenger cascades within seconds.

The most common effect of glucocorticoids on memory formation is an impairment following high, sustained doses of glucocorticoids, produced by multiple injections or chronic stress. However, the effects of glucocorticoids on memory consolidation and retention follow an inverted U-shaped dose-dependent relationship. Following training, peripheral or central administration of low doses of glucocorticoids enhances memory, whereas high doses are less effective or may even impair memory. Single injections of moderate doses of corticosterone or the synthetic glucocorticoid dexamethasone enhance memory for passive avoidance response. Similar modulatory effects of corticosterone on learning and memory processes have been observed in contextual-cue fear conditioning and a spatial water-maze task in rats. In addition to the dose, the biphasic effects of glucocorticoids also depend on the time and the relative aversiveness of the task. It has been shown that injections of moderate doses of dexamethasone following training can impair spatial memory in the water-maze task. However, if the training conditions are made less stressful by increasing the water temperature of the maze, such injections can enhance memory. Drug effects are greatest when the hormone is administered immediately after training and are generally ineffective when administered several hours later. This supports the belief that memories of an experience are not fixed or consolidated at the time of the experience, but involve a time-dependent process.

Glucocorticoids readily penetrate the blood–brain barrier and act on receptors in the central nervous system, especially in the hippocampus and amygdala, to influence learning and memory processes. The hippocampus is rich in glucocorticoid receptors, and is the primary neural structure where glucocorticoids exert their actions. In rats, both enhanced secretion and complete removal of corticosterone levels induce loss of hippocampal neurons and atrophy of dendritic structures. This is accompanied by an impairment of spatial memory performance, suggesting that the morphological changes induced by glucocorticoids have an impact on hippocampal function. Glucocorticoids also affect memory storage through influences involving the basolateral nucleus of the amygdala and requiring β-adrenergic activity (i.e. noradrenaline: see below). The amygdala integrates hormonal and neural influences to modulate associative memories in other brain regions, including the hippocampus. For example, lesions of the amygdala or inactivation of β-adrenergic receptors within the basolateral amygdala also block the memory-modulatory effects of intrahippocampal administration of glucocorticoids.

Catecholamines

The catecholamine hormones adrenaline (epinephrine) and noradrenaline (norepinephrine) are secreted by the adrenal medulla in response to the activation of the sympathetic nervous system (see Figure 3). In addition, noradrenaline is the major neurotransmitter released by postganglionic neurons of the sympathetic nervous system.

Together with glucocorticoids, the blood-borne catecholamines help coordinate different parts of the body for adaptation to stressful situations.

Marked elevations in plasma catecholamine levels are detected after a single shock to the foot of an intensity commonly used in avoidance training experiments, suggesting that peripheral catecholamines are responsive to training. Destruction of peripheral sympathetic nerve terminals or removal of the adrenal medulla produces performance deficits in avoidance learning trials, whereas pretraining injections of adrenaline reverse the performance deficit. Administration of adrenaline or noradrenaline following training enhances both active and passive avoidance retentions in rodents. Finally, adrenaline enhances the retention of an appetitive discrimination task in rats and mice, indicating that the modulatory effects of catecholamines are not restricted to aversive training situations. Taken together, these results suggest that catecholamines play a part in modulation of learning and memory processes. Interestingly, the animal may not even have to be conscious for this process to occur, because it has been reported that adrenaline promotes conditioning in anesthetized animals.

Depending on the strength of training, catecholamines may produce either enhancement or impairment of memory formation in a dose- and time-dependent fashion. Moderate shock produces faster learning than mild shock. Adrenaline injections plus mild shock produce learning curves that are nearly identical to those produced by moderate shock. Thus, the levels of endogenous catecholamines achieved in plasma, in response to different footshock intensities, correspond well to the magnitude of retention performance. It has been suggested that the role of catecholamines is to potentiate noxious stimuli, making them especially salient and more likely to be remembered.

Unlike glucocorticoids, catecholamines do not readily enter the brain and the effects of peripheral catecholamines on memory storage processes apparently are mediated at sites outside the blood–brain barrier. Two hypotheses have been developed for how catecholamines affect memory formation. First, activation of peripheral β-adrenergic receptors by catecholamines influences memory through changes in concentration of blood glucose, which readily penetrates the blood–brain barrier and affects neural metabolism, neural activity, and neurotransmitter synthesis. Second, during and after an emotionally stressful event release of catecholamines activates peripheral adrenergic receptors, which indirectly stimulates the amygdala via the vagus nerve and nucleus of the solitary tract. The amygdala, in turn, influences hippocampal memory storage processes. However, because catecholamine receptors are also present in the central nervous system, the effects of systemically administered catecholamines may involve direct actions on central adrenergic receptors. Synthesized primarily in the brainstem, central noradrenaline acts as a general regulator of neurotransmission through extensive projections in the brain. Noradrenaline is directly involved in the general degree of alertness and arousal, as well as in exerting central control over the endocrine and autonomic nervous systems. Evidence shows that the effects of adrenaline on memory consolidation depend on the integrity of the noradrenergic system in the amygdala, and systemic injections of adrenaline induce the release of noradrenaline within the amygdala.

Vasopressin

Vasopressin is a peptide hormone generally known for its antidiuretic and pressor functions. It is predominantly synthesized by two hypothalamic cell types, the magnocellular and parvocellular neurons. Magnocellular neurons send axons into the posterior pituitary, where vasopressin is secreted into the bloodstream to exert its hormonal actions. In contrast, parvocellular neurons send vasopressin-containing projections to both the median eminence (where the hormones are released into the portal vessel system to regulate anterior pituitary function) and the brainstem and spinal cord. Furthermore, extrahypothalamic sources of vasopressin have been identified in the stria terminalis, septum, amygdala, and locus ceruleus. These structures as well as the hippocampus are the major targets of vasopressin-containing projections.

The first report on the involvement of vasopressin in learning and memory appeared in 1965 when de Wied demonstrated that removal of the posterior pituitary prior to training in an active avoidance experiment did not alter acquisition of learning, but markedly accelerated its extinction. This deficit was reversed by the administration of an extract of a portion of the pituitary, later identified as vasopressin. Exogenously administered vasopressin also influences the acquisition, retention, and extinction of learned behavior. In active avoidance experiments in which the conditions were such that the initial tendency to avoid was low, pretraining systemic injections of vasopressin enhanced acquisition of the avoidance response. Pretraining

vasopressin injections are ineffective when the aversive stimulus is stronger, suggesting an interaction between training strength and hormonal treatment, such as we have seen for other hormones. Vasopressin and a variety of fragments and analogues all enhance retention of a shock-motivated passive avoidance response following systemic administration. Avoidance extinction is also delayed by central injections of vasopressin. Moreover, the mutant Brattleboro rat, which lacks the ability to synthesize vasopressin, displayed an impairment in passive avoidance retention and showed more rapid extinction of an active avoidance response compared with control animals. However, injections of vasopressin normalized the learning deficit.

Vasopressin does not readily pass the blood–brain barrier, although there is evidence for active transport of this molecule across the barrier. The memory-modulatory effects of vasopressin in the central nervous system are probably due to vasopressin metabolites, which are endogenous peptides that affect the central nervous system without resulting in endocrine effects on renal function and blood pressure. They are generally more potent than the parent molecule in enhancing passive avoidance retention. Lesion studies and local applications of receptor agonists and antagonists show that vasopressin may modulate retrieval processes through actions in the amygdala and hippocampus, and consolidation processes through actions in the dorsal septum, dorsal raphe and hippocampus.

However, peripheral vasopressin also enhances retention in a variety of tasks through action on receptors outside the blood–brain barrier. Systemic administration of an antagonist that blocks the effects of vasopressin on blood pressure reverses the enhancement of retention and the delay of extinction produced by vasopressin. This effect is most likely due to the peripheral autonomic effects of vasopressin, which alters the state of arousal. Taken together, parallel central and peripheral systems may exist upon which vasopressin acts to influence the strength of memory.

Opioid Peptides

Learning and memory are important aspects of adaptive behavior, which are regulated by opioid peptides in the neocortex, hippocampus, and amygdala. Opioid peptides are derived from the proteolytic cleavage of the precursor proteins preproenkephalin, preprodynorphin and proopiomelanocortin (POMC), producing enkephalins, dynorphins (A and B), and β-endorphins respectively. Two additional opioid-related peptides have been identified, named endomorphin and nociceptin (or orphanin FQ). Binding of opioid peptides to μ, κ, and δ opioid receptors or the nociceptin/orphanin FQ receptor produces a variety of behavioral and physiological responses.

Enkephalins are produced and released from neurons in the brain, and are jointly released with adrenaline from the adrenal medulla in response to stress. Beta-endorphin and ACTH are also released from the anterior pituitary in response to stress. The hormonal and neuromodulatory actions of enkephalins and β-endorphin on learning and memory have been identified in a variety of species and training conditions. The most frequently reported effects of enkephalins and β-endorphins are impairments in the acquisition and retention of learning tasks, although there are some reports of enhancement. For instance, peripheral administration of enkephalins impairs the acquisition and retention of active avoidance performance, and passive avoidance retention in rats and mice, with a U-shaped dose–response curve. Systemic injection of β-endorphin following training generally impairs retention of active and passive avoidance behavior. Thus, enkephalins and endorphins have a modulatory role in learning and memory, because the interaction is dependent on the strength of training and concentration of the drug.

There are several reports on the effects of dynorphins on memory formation. Injection of dynorphin A following training in rats does not affect inhibitory or shuttle avoidance responses, but enhances retention, while it impairs aversive and appetitive learning in chicks. Injection of dynorphins into the hippocampus impairs spatial learning in the water maze, and working memory in the radial arm maze. It is also believed that spatial memory deficits in aged rodents are due, in part, to elevated levels of dynorphin A. The newly identified opioid peptides endomorphin and nociceptin also display a modulatory role in leaning and memory. Injection of endomorphins into the hippocampus impairs spatial memory, working memory, and passive avoidance memory; μ-opioid receptor antagonists block this effect. Nociceptin impairs retention but not acquisition of passive avoidance performance, while intrahippocampal injections induce spatial learning impairments in the water-maze task. Mice containing deletions of the nociceptin gene have enhanced spatial learning ability, greater working memory performance, and enhanced synaptic plasticity such as long-term potentiation.

Evidence thus suggests that opioid peptides can enhance or inhibit memory formation. Memory modulation is dependent on the strength of training and whether the drug is administered centrally or peripherally. Following systemic injections, enkephalins, dynorphins, and β-endorphin apparently act on opioid receptors that lie outside the blood–brain barrier to exert their modulatory actions on learning and memory. It has been suggested that opioids interact with peripheral β-adrenergic systems to regulate the release of noradrenaline in the amygdala. Following central administration opioid peptides appear to act on modulatory systems in the central nervous system to influence learning and memory, which argues for the existence of parallel central and peripheral mechanisms whereby opioid peptides might influence the strength of the memory.

Sex Hormones

The function of gonadal hormones in the nervous system was once considered to be solely the regulation of sexual behavior. Studies implicating estrogen-related effects on verbal memory tests, performance in spatial learning tasks, and reduction in the incidence of Alzheimer disease in postmenopausal women have changed this view. The hypothalamus and anterior pituitary control the cyclic release of estrogen from the ovary throughout the menstrual cycle. One of the first identified nonreproductive actions of estrogen was on the cyclic regulation of synaptogenesis (or synapse formation) in neurons of the hypothalamus and hippocampus. As estrogen levels increase and decrease throughout the cycle, the potential number of modifiable synapses (or spines) used to store information changes. Furthermore, ovariectomized rats having no detectable levels of circulating estrogens have fewer hippocampal spines than rats given two injections of estrogen. Similar increases in spine density have been observed in cultured hippocampal neurons when treated with estrogen.

The structural changes induced by estrogen treatment may alter the ability of the hippocampus to process and store information. When estrogen levels are highest during proestrus in rats, synaptic plasticity such as long-term potentiation (LTP) is enhanced. Estrogen enhances LTP in the CA1 region of the hippocampus but impairs LTP in the dentate gyrus of estrogen-treated ovariectomized rats. In humans, when estrogen levels decline during menopause, there is a decrease in cognitive ability and a greater risk of neuronal cell damage. There have been some studies, although controversial, demonstrating that estrogen replacement therapy may provide some protection against the development of Alzheimer disease, and may improve cognitive function.

The effects of estrogen on learning and memory are varied and somewhat contradictory, with findings indicating enhancement, impairment or no change, depending on the dose of estrogen used. Ovariectomized rats treated with low doses of estrogen have been reported to improve acquisition and choice accuracy in the radial arm maze, while high doses of estrogen impair working memory performance. Prolonged estrogen treatment is reported to improve working memory performance in a radial arm maze, but this does not occur following brief estrogen treatment. Aging female rats with low levels of estrogen during 'estropause' display impaired reference memory in the spatial water-maze task, but improved working memory in the radial arm maze. In women, estrogen treatment impairs performance of spatial tasks, while enhancing verbal memory performance.

Although the mechanism of action for estrogen's effects on learning and memory has yet to be determined, there are some promising models. Estrogens may bind cell surface receptors to rapidly alter neuronal excitability through ion channels and second-messenger systems, and may directly alter gene expression. It is also believed that estrogen interacts with and maintains the basal forebrain cholinergic system by regulating the activity of key enzymes including choline acetyltransferase. Estrogens are therefore considered to have multiple effects on brain structures and behaviors, including learning and memory.

CONCLUSION

It is widely believed that acquisition, storage, and retrieval of information involve the functional modification of synaptic connections between neurons. Hormones have the ability to modulate learning processes by either increasing or decreasing the strength of the memory trace. This can be achieved directly, by altering the structure of neurons, generating greater or fewer numbers of synapses and thereby increasing or decreasing the amount of information that can be stored and affecting the cellular process that results in stronger synapses; or it can be achieved indirectly, by activating the autonomic nervous system and increasing arousal and energy metabolism and hence the efficiency of brain function. The modulatory role of hormones in memory is important because

hormones inform the organism of events that ultimately support survival.

Further Reading

Brown RE (1994) *An Introduction to Neuroendocrinology*. Cambridge, UK: Cambridge University Press.

Cahill L (2000) Neurobiological mechanisms of emotionally influenced, long-term memory. *Progress in Brain Research* **126**: 29–37.

Cahill L and McGaugh JL (1996) Modulation of memory storage. *Current Opinion in Neurobiology* **6**(2): 237–242.

Martinez JL, Schulteis G and Weinberger SB (1991) How to increase and decrease the strength of memory traces: the effects of drugs and hormones. In: Martinez JL and Kesner RP (eds) *Learning and Memory, A Biological View*, pp. 149–198. San Diego, CA: Academic Press.

McEwen BS and Alves SE (1999) Estrogen actions in the central nervous system. *Endocrine Reviews* **20**: 279–307.

McEwen BS and Sapolsky RM (1995) Stress and cognitive function. *Current Opinion in Neurobiology* **5**(2): 205–216.

McGaugh JL and Cahill L (1997) Interaction of neuromodulatory systems in modulating memory storage. *Behavioural Brain Research* **83**: 31–38.

Schulteis G and Martinez JL (1992) Peripheral modulation of learning and memory: enkephalin as a model system. *Psychopharmacology* **109**: 347–364.

Thompson RF (2000) Peptides, hormones, and the brain. In: *The Brain: A Neuroscience Primer*, 3rd edn, pp. 159–195. New York, NY: Worth.

Hunger, Meals, and Obesity

Introductory article

Jan H Strubbe, Department of Animal Physiology, University of Groningen, Haren, Netherlands

CONTENTS

Hunger, meals, and obesity are strongly related topics discussed in this article. Obesity is a severe state of being overweight, which is increasing rapidly in the Western World and is a major health problem. Obesity is often associated with an increased desire for food (hunger) which will be taken as larger bouts of feeding activity (meals). Therfore, to get more insight in the causes of obesity, knowledge of the physiological background of feeding behavior is extremely important. This will be discussed in this article.

INTRODUCTION

In humans and animals, food intake depends on several internal and external variables. They can adapt to changes in the caloric content of the diet so that, over a wide range of content, intake matches energy expenditure and body weight remains essentially unchanged. On a diet with a constant caloric content, an increase in energy expenditure results in increased food intake. One example of this phenomenon is the hyperphagia (overeating) that results from a lower ambient temperature, where the animal has to compensate for its heat loss. In this case, the energy content of the body remains at a constant level, since the subject eats the same amount of calories as it expends. This is one of the arguments for the energy content of the body and therefore the adipose (fat-storing) tissue mass being regulated within narrow limits. Another argument derives from the phenomenon whereby after temporary starvation or overfeeding the disturbed body weight returns towards its normal level through compensatory adjustment of food intake. This indicates that the energy content of the body is subject to homeostatic control (maintenance of internal equilibrium) by the regulation of food intake.

In humans, feeding occurs during bouts of energy intake called meals. Meals are regarded as bouts of behavior that, although necessary for supplying nutrients to the body, result in perturbations of homeostatic controlled parameters. The main strategy is to change average meal size in order to prevent homeostatic disturbances in the body. If conditions dictate that more food be consumed than can be accommodated by means of a small number of large meals, more meals will be eaten. The size of individual meals is the focus of intense research as scientists investigate compounds known as satiety factors that regulate the termination of meals and therefore require animals to consume smaller or larger meals.

This article will discuss the effects of meals on homeostasis, and will propose that the initiation and termination of meals is a well-anticipated and orchestrated behavior intended to maximize the passage of energy into the body, while minimizing homeostatic disruption to the body. Disturbances of this regulation lead to disorders in food intake, such as hyperphagia, and consequently to the development of obesity.

EFFECTS OF MEALS ON HOMEOSTASIS

The meal patterns of humans and animals are strongly related to the light–dark cycle. For example, rats eat 80% of their daily food intake during the dark phase. This means that during each day the energy content of the rat varies considerably, as is also indicated by the alteration in net lipogenesis (fat formation) during the dark phase and net lipolysis (fat reduction) during the light phase. Body weight is constant only over a longer timescale. This rhythmic control of food

intake is governed by oscillators located in the suprachiasmatic nucleus (SCN) in the hypothalamus (a region of the forebrain that exerts profound regulatory influences over physiological and behavioral processes which are essential for survival). These so-called *circadian pacemakers* are synchronized by the light–dark change, and they largely determine the temporal organization of food intake and interact with the short-term regulation of energy intake. A lesion in the SCN will induce arrhythmicity of feeding behavior (i.e. meals are spread equally over the daily cycle).

These rhythmic factors are modulating and not regulatory factors in the sense that they are involved in the regulation of body energy content. Rather, they constitute one of the constraints within which the regulation of food energy intake must operate. There are many constraints of this type, ranging from learned habits in feeding behavior (e.g. adjustment of meal size to expectations of caloric content of the diet, based on gustatory experience), to inhibition of feeding by the demands of other urgent behaviors such as sleeping or social confrontations.

Thus although many non-regulatory factors determine the switching of meals in the 'on' or 'off' direction, there is also a wide variety of regulatory negative feedbacks, or satiety signals, which act in the short term as well as in the long term, reporting to the central nervous system about the current energy content of the body. These signals are released by different compartments where (1) food is ingested (oropharyngeal regions), (2) nutrients are digested and absorbed (stomach and intestines), and (3) fuels are stored (liver and adipose tissues).

The afferent (inward) pathways to the central nervous system may be either nervous or hormonal. In this respect the gut hormones cholecystokinin (CCK), bombesin, and glucagon-like peptide 1 (GLP-1) may play a major role as short-term satiety factors. There is evidence that CCK can act on the afferent pathways of the vagus nerve (which runs from the brain through the thorax to the abdomen), since subdiaphragmatic cutting of the nerve appears to abolish the suppressive effects of CCK on food intake (Figure 1). When the area of entrance of the vagus nerve in the CNS, namely the nucleus tractus solitarius (NTS), is damaged, similar effects are obtained. In conclusion, since CCK-containing neurons and receptors are found in the whole chain from the intestine via the vagus to the NTS, and from there via the parabrachial nucleus to the ventromedial hypothalamus (VMH), it is possible that CCK influences feeding motivation at all of these levels. Since lesioning of the VMH results in strong

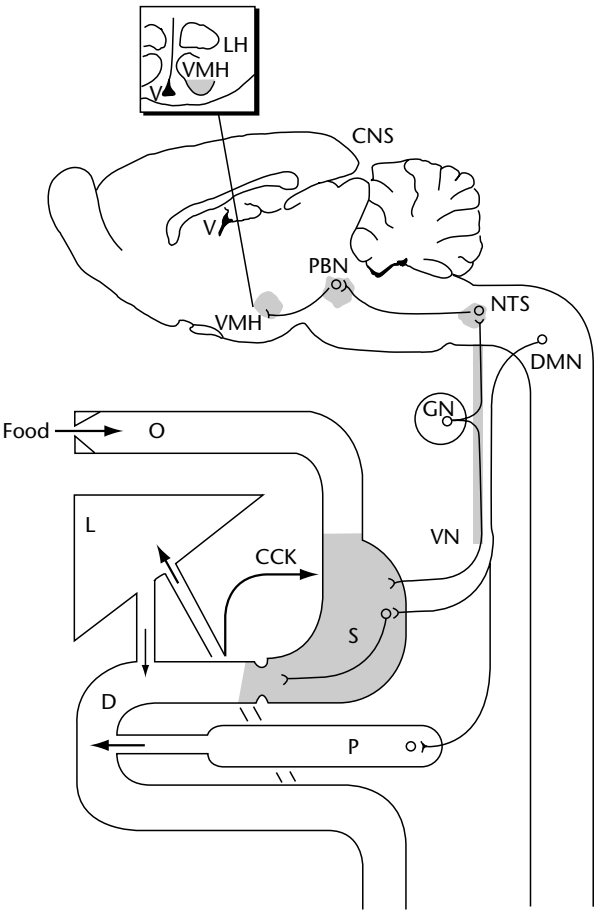


Figure 1. Schematic presentation of anatomical structures showing where cholecystokinin (CCK) is transmitted and where the CCK receptors are located (shaded areas). S, stomach; D, duodenum; P, pancreas; L, liver; O, oropharyngeal cavity; CNS, central nervous system; VN, vagus nerve; PBN, parabrachial nucleus; NTS, nucleus tractus solitarius; GN, ganglion nodosum; DMN, dorsomotor nucleus of the vagus. From Strubbe JH (1994) Regulation of food intake. In: Westerterp-Plantenga MS, Fredrix EWHM and Steffens AB (eds) *Food Intake and Energy Expenditure*, pp. 141–154. London: CRC Press.

hyperphagia and finally in obesity, it has been suggested that this area is the main target for feedback control of satiety signals. However, lesioning of the lateral hypothalamus (LH) caused aphagia (cessation of eating) and a decrease in body weight. On the basis of these findings, the LH is regarded as a feeding or hunger center, whereas the VMH is more likely to be a satiety center.

Among the absorbed fuels, glucose might play a prominent role together with insulin, according to the 'glucostatic theory'. The metabolism of these fuels induces heat production, and the resulting increased body temperature could inhibit feeding activities. Meals are terminated at very similar

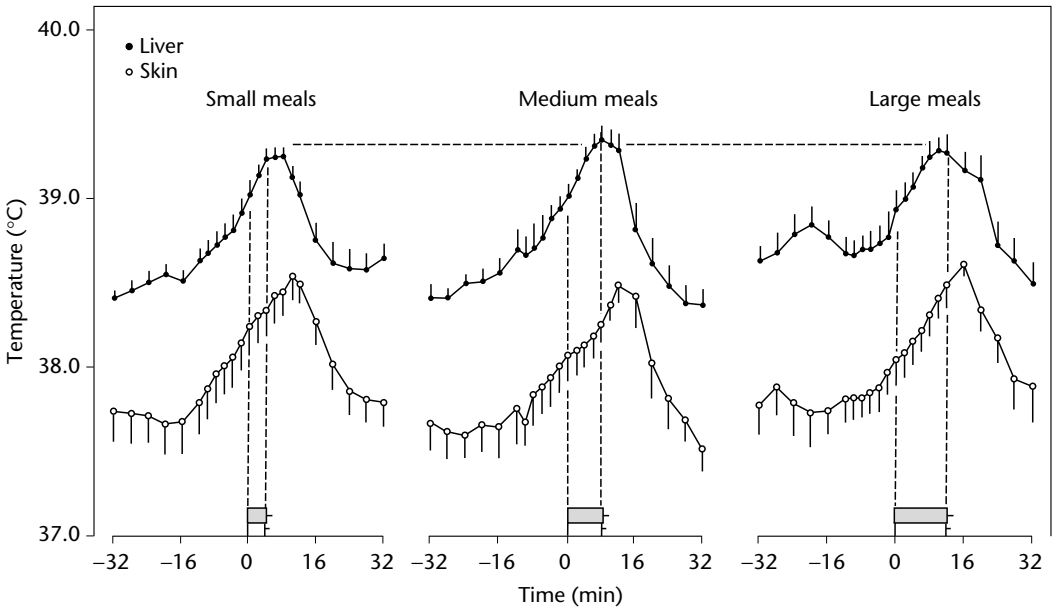


Figure 2. Hepatic and skin temperatures for separate groups of rats consuming small, medium, and large meals. From De Vries J, Strubbe JH, Wildering WC, Gorter J and Prins AJA (1993) Patterns of body temperature during feeding in rats under different ambient temperatures. *Physiology and Behavior* **53**: 229–235.

relatively high liver temperatures, which might provide a signal to stop all feeding activities (Figure 2). This concept has been formulated as the 'thermostatic theory'. The reserve tissues may also inform the central nervous system about their content. Signals may arise in association with the size of adipose tissue ('lipostatic theory'). Among these signals, the hormones insulin and leptin may be important determinants of the regulation of feeding behavior by controlling the size of the adipose tisssue mass. This control holds for the long-term regulation of food intake, setting the motivational background level for feeding behavior in the brain. Other signals may be involved in reporting the size of the glycogen reserves, and would therefore contribute more to the short-term regulation of food intake.

In addition to the negative satiety signals, positive feedbacks from the oropharyngeal regions can also be distinguished. Once a meal has been started, these feedback signals keep the feeding motivation at a relatively high level. Thus they stabilize feeding in the sense that they help to postpone meal termination. These 'positive and negative signals' consist of fluctuating physiological factors (e.g. hormones, blood glucose level, body temperature) which are kept within physiological limits by the animal's decisions to start, continue or stop feeding activities so as to prevent adverse perturbations of homeostasis. Moreover, anticipatory and conditioned processes will help to minimize the perturbations of homeostasis.

EFFECTS OF CONDITIONING ON PREMEAL PHYSIOLOGICAL CHANGES AND HUNGER

The body responds to meals so as to minimize both the magnitude and the duration of homeostatic perturbation. One way in which it does this is by anticipating meals and initiating homeostasis-preserving responses as early in the food-intake process as possible. This meal-anticipating process can also be used whenever the environment dictates that an animal has to eat relatively large meals in order to obtain adequate nutrition. However, the animal does not change its meal pattern simultaneously with the change in environment. Rather, the changes in meal size occur over a period of several days as the animal learns what to expect and what the physiological consequences of a particular meal may be. When rats must eat all of their daily food within a short window of time, there is a maximum meal size possible, and if more food per meal must be eaten, the rat sacrifices weight instead of consuming that much food all at one time. Thus the disruption of other related systems that is caused by increasing the size of a meal may elicit corrective responses (in this case, the

premature cessation of meals) in order to protect other perturbed systems.

A certain flexibility is built into the system so that when large meals must be consumed for survival, and when the environment is completely predictable, anticipatory responses enable a further increase in meal size. Adaptations occur as increases in the autonomic responses to gastrointestinal processing of food and of insulin secretion. Conditioned increases in insulin levels were seen only when rats expected to eat a very large meal (Figure 3). Since this conditioned insulin response was abolished by pharmacological treatment with the muscarinic receptor-antagonist atropin, it can be concluded that the response is caused by the parasympathetic nervous system. These conditioned insulin responses may therefore prevent wider fluctuations in blood glucose levels when the meal is eaten.

When animals expect to eat a large meal, it is also advantageous to keep body temperature low in order to allow a longer period of temperature increase. This indeed occurs in the voluntarily taken meals shown in Figure 2, where the preprandial increases in body temperature are lower when animals expect to eat a larger meal. The thermostatic theory predicts that rats can eat for longer before they reach a 'setpoint' temperature which may terminate the meal. Therefore it is possible that lower body temperatures, as often observed in obese rodents, enable them to eat larger meals.

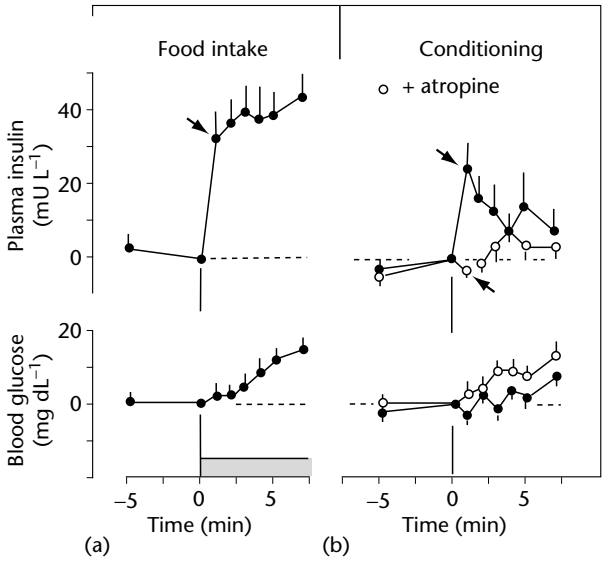


Figure 3. Mean plasma insulin and blood glucose responses (a) during feeding and (b) after opening of doors in front of an empty food hopper, in rats habituated to a feeding schedule of 2 meals per day. In the latter condition, premedication with atropine was also given (open circles). From Strubbe JH (1992) Parasympathetic involvement in rapid meal-associated conditioned insulin secretion in the rat. *American Journal of Physiology* **263**: R615–R618.

INCREASING PREVALENCE OF AND FACTORS CONTRIBUTING TO OBESITY

The prevalence of obesity and overweight type 2 diabetes mellitus is increasing rapidly in the Western world. However, during the last few years major increases in the incidence of overweight have also been seen in other parts of the world. Although obesity in humans can be regarded as a risk factor for many serious diseases, including coronary heart disease and maturity-onset (type 2) diabetes mellitus, the occurrence of gallstones, arthritic problems and varicoses, it is not a disease in its own right except in very severe cases. Even the presence of filled adipose tissues can be of survival value to overcome periods of famine. It is therefore possible that this adaptation has developed more in some regions of the world than in others (e.g. obesity observed in Pima Indians). In a similar way to the situation in hibernation where, for instance, marmots become very fat prior to wintertime, humans in ancient times might have filled their adipose tissues in times of plenty and depleted their stores in times of famine. Under the force of selection pressures this may have led to an adapted 'thrifty' genotype which is able to survive unfavorable external conditions such as fluctuations in the supply of food due to dryness or low ambient temperatures.

Although the range over which body fat stores are filled under normal conditions may vary widely between individuals, each individual or group of individuals has its own genetically determined setpoint of adipose tissue size which is dependent on age and external conditions. There is evidence that the body mass index (BMI) (i.e. body weight in kilograms divided by the square of the height in meters) of children correlates strongly with those of their biological parents and not with those of their adoptive parents. Moreover, if both parents are obese, there is a greater likelihood that their children will also become obese.

These phenomena provide evidence for a genetic background of human obesity. However, we should also take into account the behavioral component, which again is very similar between

parents and children, as the behavior and genetic background may act in the same direction.

NEW PERSPECTIVES ON HUNGER, MEALS, AND OBESITY

The mechanism of the setpoint control (i.e. the regulation of food intake) has been the subject of recent research. Genetically determined disturbances in this regulation may lead to an imbalance between intake and subsequent storage and expenditure. It is the goal of obesity research to determine the nature of the genetic background of obesity and the point of action on behavior and physiology. The many genetic obese animal models may help to establish the origin of such defects, although it is very difficult to distinguish between primary and secondary effects.

It is known that in many genetically obese rodents, such as the ob/ob mouse, the rate of thermogenesis (heat production) is decreased. The thermogenic effect of food in obese humans is also often reduced. It has been suggested that they eat more in order to increase heat levels in the body and to prevent heat loss by means of insulation with extra adipose tissue. Thus a lower preprandial temperature or decreased prandial slope may be one of the major causes of eating larger meals, which may result in obesity. Indeed, obese rodents such as the obese Zucker rat eat larger and less frequent meals. Therefore the deviating factor in food-intake regulation of obesity may occur not in the initiation of meals but in their termination. In fact the strategy is quite similar to those in lactating and diabetic rats, and in rats kept at low ambient temperatures, although there may be different mechanisms underlying the control.

Although decreased thermogenesis may be one of the final physiological processes that regulates the behavior of obese rodents, there may be several physiological processes involved in this defect, including decreased feedback control in the short term (CCK and GLP-1) or in the long term (insulin and leptin). In recent years the discovery of leptin, a hormone that is secreted by adipocytes and which is thought to act as a signal for fat stores, has provided new insights into the interrelationships between obesity and energy regulation. Studies have shown that mice strains which lack either the gene that is responsible for the production of leptin (e.g. ob/ob mice) or its receptor (e.g. db/db/mice) become obese, while administration of recombinant leptin results in direct weight loss in the ob/ob mice. However, the ineffectiveness of leptin in humans suggests that there is marked resistance, in contrast to the situation in non-obese rodents. The targets of leptin include pro-opiomelanocortin (POMC) neurons in the arcuate nucleus which is located close to the VMH. Leptin receptors are located on the POMC neurons that synthesize melanocortins. Studies in humans have shown that polymorphisms of the melanocortin (MC-4) receptor may be associated with obesity. Future research on agonists of this MC-4 receptor will reveal the many physiological processes that are governed by this receptor, including the stimulation of thermogenesis, which may be a final causal factor allowing humans and animals to reduce their food intake. Thus the development of selective agonists will be a useful approach in the treatment of obesity.

CONCLUSION

In humans and animals, body weight is often maintained within narrow limits on a variety of diets and under different external conditions by homeostatic control involving the regulation of food intake. The tendency to display feeding behavior (hunger) or to stop meals (satiety) at any particular point in time will depend on (1) constraints such as day–night rhythms, and all kinds of learned habits and behavioral interactions, (2) the current level of energy expenditure, (3) possible shifts in the desired level of body energy content, and (4) the combined result of all of the above-mentioned satiety signals pooled in the central nervous system. These 'positive and negative signals' consist of fluctuating physiological factors (e.g. hormones, blood glucose levels, body temperature) which are kept within physiological limits by the animal's decisions to start, continue, or stop feeding activities in order to prevent adverse perturbations of homeostasis of body weight and physiological variables. Moreover, anticipatory and conditioned processes will help to minimize such perturbations of homeostasis.

Further Reading

Barsh GS, Farooqi IS and O'Rahilly S (2000) Genetics of body weight regulation. *Nature* **404**: 644–651.

Bray GA and Tartaglia L (2000) Medicinal strategies in the treatment of obesity. *Nature* **404**: 672–677.

Friedman JM (2000) Obesity in the new millenium. *Nature* **404**: 632–634.

Kopelman PG (2000) Obesity as a medical problem. *Nature* **404**: 635–643.

Lowell BB and Spiegelman BM (2000) Towards a molecular understanding of adaptive thermogenesis. *Nature* **404**: 652–660.

Schwartz MW, Woods SC, Porte D, Seeley RJ and Baskin DG (2000) Central nervous system control of food intake. *Nature* **404**: 661–671.

Strubbe JH (1994) Obesity. In: Westerterp-Plantenga MS, Fredrix EWHM and Steffens AB (eds) *Food Intake and Energy Expenditure*, pp. 183–194. London, UK: CRC Press.

Strubbe JH (1999) Circadian organization of feeding behavior. In: Westerterp-Plantenga MS, Steffens AB and Tremblay A (eds) *Regulation of Food Intake and Energy Expenditure*, pp. 135–157. Milan, Italy: EDRA.

Westerterp-Plantenga M, Steffens AB and Tremblay A (eds) (1999) *Regulation of Food Intake and Energy Expenditure*. Milan, Italy: EDRA.

Woods SC and Strubbe JH (1994) The psychobiology of meals. *Psychonomic Bulletin and Review* **1**: 141–155.

Huntington Disease

Introductory article

James F Gusella, Massachusetts General Hospital/Harvard Medical School, Charlestown, Massachusetts, USA

CONTENTS

Huntington disease (HD) is a neurodegenerative disorder that involves the loss of neurons in the basal ganglia (particularly the caudate nucleus and the putamen) and in the cerebral cortex. It is characterized by a peculiar writhing movement disturbance, but also has intellectual and psychiatric manifestations. All cases of HD are genetically caused, resulting from an expanded CAG trinucleotide tract that encodes an elongated stretch of consecutive glutamine residues in huntingtin, a large protein of unknown function.

SYMPTOMS OF HUNTINGTON DISEASE

Huntington disease (HD) is named after George Huntington, a nineteenth-century family physician on Long Island, New York, who first described the symptoms of the disorder and recognized its inherited nature. The characteristic feature of HD is a peculiar motor disturbance that begins subtly with manifestations such as clumsiness, slight adventitious movements, awkward gait or lack of smooth eye pursuit. However, the movement disorder inexorably progresses to blatant chorea that consumes all parts of the body and is only absent during sleep. Indeed, HD was originally termed Huntington chorea because of the distinctiveness and exaggerated nature of its hallmark symptom. However, a century later the name was changed to HD in recognition of the presence of other disease manifestations, including intellectual decline and psychiatric disturbances. Psychiatric symptoms, which include chronic depression, impulsive behavior, personality changes and occasionally frank psychosis, are variably present and may in some cases precede the motor disturbance by many years. However, age at onset of HD is usually judged on the basis of the first detected neurological symptoms, which are more reliably detected.

The age at neurological onset of HD varies from the juvenile years to late in life, but most cases manifest in middle age, usually after more than half a lifetime without any detectable abnormality. As the disorder progresses, the HD patient suffers increasing lack of motor control in combination with decreasing intellectual capacity, until the capacity for independent functioning and eventually communication is lost. The HD patient requires constant care and protection from self-injury. Late in the course of the disorder, which extends for 10–20 years after onset, the patient will typically have exhibited dramatic weight loss and the chorea may have given way to rigidity, particularly in juvenile-onset cases. Death then ensues, often due to heart disease resulting from the constant motions and dramatic weight loss, or to aspiration pneumonia, resulting from difficulties in taking food.

BRAIN DAMAGE ASSOCIATED WITH HUNTINGTON DISEASE

Post-mortem examination of brains from HD patients has revealed a characteristic pattern of neuronal loss that produces the disorder's clinical symptoms. The region that is affected earliest and most severely is the caudate nucleus, whose architecture is eventually completely destroyed by the effects of the mutation. Neuronal cell loss shows a gradient of progression across the caudate that eventually encompasses the adjacent putamen. Medium-sized spiny neurons that use the neurotransmitter γ-aminobutyric acid (GABA) and send axons out of the caudate are most vulnerable to the disease, while other neuronal types are relatively spared. As the GABA neurons are lost, the inhibition that they would normally cause in downstream target neurons is released, presumably leading to the characteristic uncontrolled movements. In late-stage HD and in juvenile-onset

cases, the patient may display rigidity rather than uncontrolled movements, as a result of the spread of neuropathology beyond the caudate to other regions. In late-stage HD brains, the caudate and putamen are completely destroyed, and it is evident that neuronal cell loss also affects other regions, particularly layers III, V and VI of the cerebral cortex. Overall brain weight may be reduced by 35% or more, but the specific neuronal populations that contribute to all of this loss have not been clearly delineated.

GENETIC BASIS OF HUNTINGTON DISEASE

HD has been detected in all races, but is most frequent in Caucasians, with a prevalence of *c*. 1 in 10 000. The disorder is due to a dominant genetic defect in a gene on chromosome 4 that has a segment of consecutive CAG codons in the 5′ coding sequence. On normal chromosomes, the CAG trinucleotide repeat is polymorphic with alleles consisting of 10–34 CAGs that are inherited in a stable mendelian fashion. The trinucleotide stretch is expanded to more than 35 CAGs on HD chromosomes, and shows a high mutation rate through meiotic transmission. Most HD chromosomes contain 40–50 CAG units, but they may have as many as 250 units. When transmitted to the succeeding generation, most of them undergo modest size changes of a few CAGs, with a bias towards length increases. Occasionally, transmissions from a male parent involve much larger length increases, up to a doubling or more in the number of CAG units, due to particular instability during spermatogenesis.

The CAG expansion mutation is the only aberration that has been found in patients with the disorder, and its length is the major factor determining age at onset of neurological symptoms. Most HD patients display midlife onset due to 40–50 CAGs. Individuals with 35–39 CAGs typically show HD onset late in life or not at all. By contrast, those with the longest CAG arrays show juvenile onset. After considering the effect of CAG tract length, the remaining variation in onset age may be due to other genetic, environmental or stochastic factors. Only one genetic modifier of relatively modest effect, namely a subunit of the glutamate receptor, has been identified to date. Interestingly, individuals with two doses of the HD defect and no normal HD gene copy do not show significantly accelerated onset. It is also noteworthy that progression to death, which typically occurs around 15 years after onset, does not show any correlation with CAG tract length. This suggests that once the mechanism of pathogenesis has been fully triggered, factors other than the genetic defect become paramount.

TEST FOR THE HUNTINGTON GENE

Prior to the discovery of a linked genetic marker for HD in 1983, there was no reliable method for determining whether asymptomatic 'at-risk' individuals had inherited the disease gene. As a result of mapping HD to chromosome 4, a linkage test became available which was applicable to prenatal and predictive testing in families with sufficient living members to identify the disease chromosome reliably via the alleles at linked polymorphic DNA markers. With the cloning of the HD gene and the discovery of the nature of the mutation, a direct DNA test based on polymerase chain reaction (PCR) amplification across the CAG repeat became possible. The direct test does not require DNA from other family members, and it is 100% diagnostic of HD. It can be used for prenatal determinations, predictive testing in 'at-risk' individuals, and for HD confirmation or differential diagnosis in those with extant movement disorders. The decision to undergo predictive testing is an intensely personal one that cannot be predicted on the basis of such factors as education, socioeconomic status or religious beliefs. For some individuals, the 'need to know' derives from a strong desire to eliminate the uncertainty of being 'at risk', for peace of mind, life planning or reproductive decisions. However, as there is currently no treatment for preventing onset or slowing the progression of HD, many 'at-risk' individuals choose not to undergo predictive testing, as they have learned to cope with the uncertainty. For some, this decision is made after the potential consequences of predictive testing have been explained through intensive genetic counseling. The provision of such counseling before proceeding with actual testing has, since the advent of the linkage test, been considered essential for a disorder such as HD, which has inherent psychiatric effects and a suicide rate far above normal.

EFFORTS TO PREVENT HUNTINGTON DISEASE

Prenatal testing clearly provides one means of reducing the impact of the HD defect in future generations. However, for individuals who are born with the HD defect, any hope of developing a treatment is focused on understanding the mechanism of pathogenesis and thereby developing a rationale

for therapy. Current evidence favors the HD mutation acting through the polyglutamine tract of huntingtin, via a 'gain of function' rather than through a simple loss of the protein's inherent activity. Notably, at least seven other neurodegenerative disorders, spinal and bulbar muscular atrophy, dentatorubropallidoluysian atrophy and spinocerebellar ataxias 1, 2, 3, 6 and 7, are all due to similar CAG expansion mutations extending the length of polyglutamine tracts in various regions of other proteins. This suggests that a similar 'gain of function' in each of these proteins results in neuronal toxicity, but the different pattern of neuronal loss in each disorder indicates that the specificity of neuronal toxicity is conferred by the function, structure or localization of the particular protein that is hosting the polyglutamine segment.

Huntingtin's primary structure and pattern of expression do not provide any direct clues about the specificity of neuronal loss in HD. Huntingtin shows no significant similarity to other proteins, and its normal function is still unknown, although roles in a wide range of cellular activities, including transcription, RNA splicing, cytoskeletal function, intracellular signaling and protein trafficking have been suggested. The protein is present in many different cell types (neuronal and non-neuronal), and is essential for normal development. However, the normal polyglutamine tract varies widely in length in other species, and is entirely absent from fruitfly huntingtin, which suggests that it is not necessary for huntingtin's normal function. The expanded glutamine tract probably alters the conformation or folding of the mutant huntingtin protein, conferring a new physical property that acts in a dominant manner to trigger pathogenesis.

The relationship between the genotype (number of CAGs) and phenotype (expression of HD) of HD patients has provided a series of criteria for studying the mechanism that triggers pathogenesis. These suggest that the polyglutamine tract must exceed a critical threshold length in order to trigger the disorder within a normal lifespan. Once the threshold (> 35 CAGs in humans) is exceeded, the mechanism is progressive with increasing numbers of consecutive CAGs, is dominant over the normal HD gene copy, and is more dependent on polyglutamine length than on the presence of two copies of the mutant gene. One property that matches these criteria is the capacity of a small amino-terminal fragment of huntingtin to self-aggregate in a test tube, converting from a soluble to an insoluble form. Production of similar fragments inside cells promotes the formation of various soluble and insoluble cellular inclusions, which in some cases are similar to inclusions observed in surviving neurons in post-mortem HD brain. It is not clear whether the intracellular inclusions in HD brain are crucial mediators of pathogenesis, or whether they are secondary consequences of huntingtin being degraded in cells that have been damaged by the effects of the mutation.

Although the simplicity of the former possibility is attractive, the lack of specific and consistent cell death as a result of inclusion formation in cell-culture and mouse models casts doubt on it. Instead, the novel conformational property of mutant huntingtin may actually trigger the disorder in patients by acting within the full-length protein, where its effect may be to alter protein–protein interactions rather than to produce insoluble inclusions. This alternative possibility has been supported by studies in 'knock-in' mice, in which the HD mutation has been introduced into the mouse equivalent gene, resulting in faithful expression of the full-length mutant huntingtin protein. These mice display altered behavior of the full-length protein many months before the formation of intraneuronal inclusions. The alterations in mutant huntingtin are seen specifically in striatal neurons and fulfill the genetic criteria outlined above, perhaps indicating an early event in the pathogenetic pathway.

Whether the trigger for pathogenesis acts at the level of full-length huntingtin or via a truncated fragment, assays using amino-terminal fragments still offer an effective method for monitoring the novel conformational property conferred by the HD mutation. Therefore they may be useful for identifying small molecules that interfere with or promote this process. Screening of compound libraries has already identified a few such compounds. It is expected that the continued pursuit of this strategy, followed by testing of the compounds in cell-culture, lower-organism and mouse models, offers the best current hope for identifying a drug to prevent the onset of HD. Such a therapy might also be expected to be effective in each of the other polyglutamine disorders, since it would be aimed at the fundamental pathogenetic property produced by the CAG expansion mutation.

An alternative route to a specific therapy for HD would be to delineate further the pathway of pathogenesis, which presumably consists of an initial impact of the mutant protein in susceptible neurons, followed by a cascade of changes that eventually lead to cell death. It has been postulated that the first effect of the expanded glutamine tract is to produce an altered protein–protein interaction that is particularly deleterious to vulnerable neurons.

Huntingtin has been shown to interact with more than 30 other proteins, but none of these has yet been established as a partner that participates in a mechanism that initiates pathogenesis. Identification of such a partner would provide an additional target for therapeutic drug screens. Once the pathogenetic pathway is triggered, each of the changes that eventually lead to cell death might also provide a later target for development of neuroprotective therapeutics. Various potential targets, such as the processes of energy metabolism, caspase proteases involved in cleaving huntingtin, and other proteins that participate in apoptotic cell death pathways have been suggested and are currently being tested in model systems. However, no intervention has yet been shown to be effective in HD patients. Indeed, at neurological onset an HD patient may already have lost 30–40% of vulnerable caudate neurons, and many of the remainder may be dysfunctional. Efforts to rescue the latter might be expected at best to halt the progression of disease symptoms. Finally, the decimation of the striatum in severely affected individuals suggests that neuronal implants may be the only available possibility for treatment in these cases. This approach is being attempted, but has not yet shown unequivocal success.

CONCLUSION

HD differs from more common neurodegenerative disorders in having an exclusively genetic cause. Delineation of the precise genetic alteration that causes specific neurodegeneration in HD has provided the unequivocal starting point and initial clues to the subsequent steps in the process of pathogenesis. Continued analysis of this pathway using genetic and biochemical approaches to define potential targets for therapeutic intervention offers the best hope for complementing effective prenatal and presymptomatic diagnosis with effective treatments to prevent the onset and progression of this devastating disorder.

Further Reading

Gusella JF and MacDonald ME (2000) Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. *Nature Neuroscience* **1**: 109–115.

Gusella JF, Wexler NS, Conneally PM *et al*. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234–238.

Huntington G (1872) On chorea. *Medical and Surgical Reporter of Philadelphia* **26**: 317–321.

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971–983.

McGeer PL (ed.) (2000) *Focus on Huntington's Disease. Neuroscience News 3*.

Martin JB and Gusella JF (1986) Huntington's disease – pathogenesis and management. *New England Journal of Medicine* **315**: 1267–1276.

Hypothalamus

Introductory article

Charles W Malsbury, Memorial University of Newfoundland, St John's, Newfoundland, Canada

CONTENTS

The hypothalamus is a collection of cell groups (nuclei) at the base of the brain that influences nearly every aspect of physiology and behavior. Hypothalamic neurons participate in neural networks that integrate endocrine, autonomic and behavioral responses that must occur together for an animal to interact successfully with its environment.

INTRODUCTION

The hypothalamus is an aggregation of discrete cell groups, or nuclei, that form the ventral part of the forebrain. This small part of the brain is critically important for the homeostatic regulation of most aspects of physiology, including the regulation of the endocrine system. Hypothalamic neurons also participate in neural networks that regulate eating and drinking, and defensive and reproductive behaviors. Thus the functions in which the hypothalamus participates are vital for both individual and species survival.

A central idea is that the hypothalamus has an integrative role in regulating physiology and behavior. It participates in the coordination of three types of responses – endocrine, autonomic and behavioral – that must occur together for an animal to interact successfully with its environment. For example, when an animal confronts a threat, the hypothalamus is required for the coordinated release of stress hormones, activation of the sympathetic nervous system, and activation of the appropriate behavior (fight or flight).

ANATOMY

The hypothalamus extends from the region of the optic chiasm to the posterior border of the mamillary bodies at the base of the brain (Figure 1). The thalamus forms its dorsal boundary. The hypothalamus has a medial division that contains discrete cell groups (nuclei) and a lateral division, the lateral hypothalamus (LH), in which the neurons are distributed diffusely.

Because of the variety of responses in which it participates, the hypothalamus must communicate with many regions of the brain and with the pituitary gland. Some of these connections carry sensory information to the hypothalamus. Although visual and auditory stimuli can influence hypothalamic activity, the hypothalamus is not involved in visual or auditory perception. For example, the visual input to the hypothalamus conveys information about light and dark, not visual patterns. A direct projection from the retina allows the light/dark cycle to 'set' the 24 h (circadian) clock mechanism within the suprachiasmatic nucleus. Taste, and sensory information from visceral organs such as the stomach, liver and heart, are also relayed to the medial and lateral hypothalamus directly from neurons within the posterior brainstem (medulla). Olfactory stimuli are extremely important in guiding the social behavior of rodents. Olfactory information reaches the hypothalamus after it is processed in limbic system structures such as the amygdala. The hypothalamus has extensive reciprocal connections with the limbic system. The limbic system can be loosely defined as the network of interconnected forebrain structures that is particularly important for emotion. It includes the hippocampus (important for memory) and the amygdala, itself a collection of cell groups with different functions. The hypothalamus is often included within definitions of the limbic system.

Some of the pathways that carry connections to and from the hypothalamus are organized into discrete and conspicuous bundles like the fornix and the stria terminalis, while others are diffuse and difficult to trace. Important fibre systems of the latter type include the medial forebrain bundle and the ventral amygdalofugal pathway. Despite the word 'medial', the medial forebrain

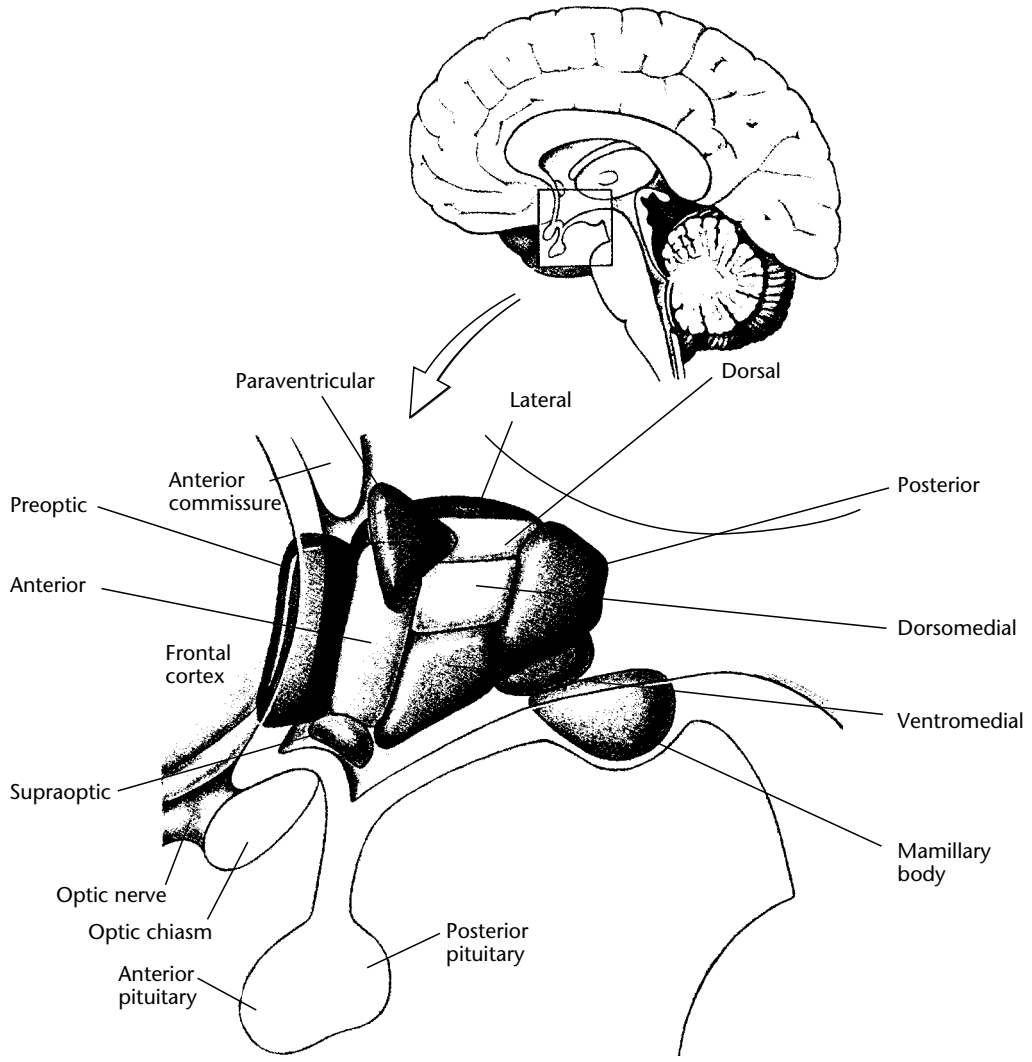


Figure 1. The nuclei of the human hypothalamus are shown in midline views of the right hemisphere.

bundle (MFB) is located within the LH, and despite the word 'bundle', these axons are not arranged into a discrete bundle, but are spread out over the whole lateral hypothalamic area. The MFB contains numerous ascending and descending components. Many of these form synapses with neuronal cell bodies within the LH and also send branches into the medial hypothalamus. However, some MFB axons only pass through the hypothalamus. An important example is the nigrostriatal pathway: this ascending pathway originates in the substantia nigra of the ventral midbrain and supplies the monoamine, dopamine, to the basal ganglia of the forebrain. Degeneration of the nigrostriatal pathway is the cause of the movement disorder known as Parkinson disease. Axons from other monoamine cell groups, i.e. those producing serotonin and noradrenaline (norepinephrine), ascend from the posterior brainstem, enter the MFB and terminate within the hypothalamus. The MFB also carries certain outputs of the medial hypothalamus. For example, neurons within the medial preoptic area are critically important for male sex behavior. The axonal projections of the preoptic area that facilitate sexual arousal descend to the midbrain within the MFB.

The amygdala is a major source of input to the medial hypothalamus. Input from this part of the temporal lobe arrives via two pathways: the stria terminalis, and the ventral amygdalofugal pathway. These provide a route for cortical and limbic influences on hypothalamic activity. This is one way in which a variety of sensory information, for example olfactory stimuli, reaches the hypothalamus. It is

also through these pathways from the amygdala that psychological stress and emotions such as fear and anger influence hypothalamic activity.

RELATION OF THE HYPOTHALAMUS TO THE ENDOCRINE SYSTEM

The hypothalamus is the final common pathway in the neural control of the endocrine system. The glands that secrete hormones into the bloodstream (e.g. the gonads, thyroid and adrenal glands) are each ultimately controlled by a particular set of hypothalamic neurons. These neurons are different from others in one respect. When excited, instead of releasing neurotransmitters that act at synapses, they release hormones near small blood vessels at the ventral surface of the hypothalamus (the median eminence). These capillary loops within the median eminence carry hypothalamic hormones down the pituitary stalk and into the anterior pituitary gland. The neurons that produce hypothalamic hormones are distributed throughout the medial hypothalamus, more or less concentrated within particular nuclei (Figure 2).

The hypothalamus participates in homeostatic regulation of hormone secretion through negative feedback. For example, the hypothalamus controls the blood level of the adrenal steroid hormone, cortisol, by secreting adrenocorticotrophic hormone (corticotrophin) releasing hormone (CRH). When one experiences stress or fear, input from the limbic system activates the hypothalamic neurons that release CRH, which in turn stimulates cells in the anterior pituitary to release adrenocorticotrophic hormone (ACTH). The release of ACTH stimulates the adrenal cortex to secrete cortisol into the bloodstream. Cortisol feeds back onto both the hypothalamus (neurons of the paraventricular nucleus which produce CRH) and the anterior pituitary (ACTH-producing cells) to limit

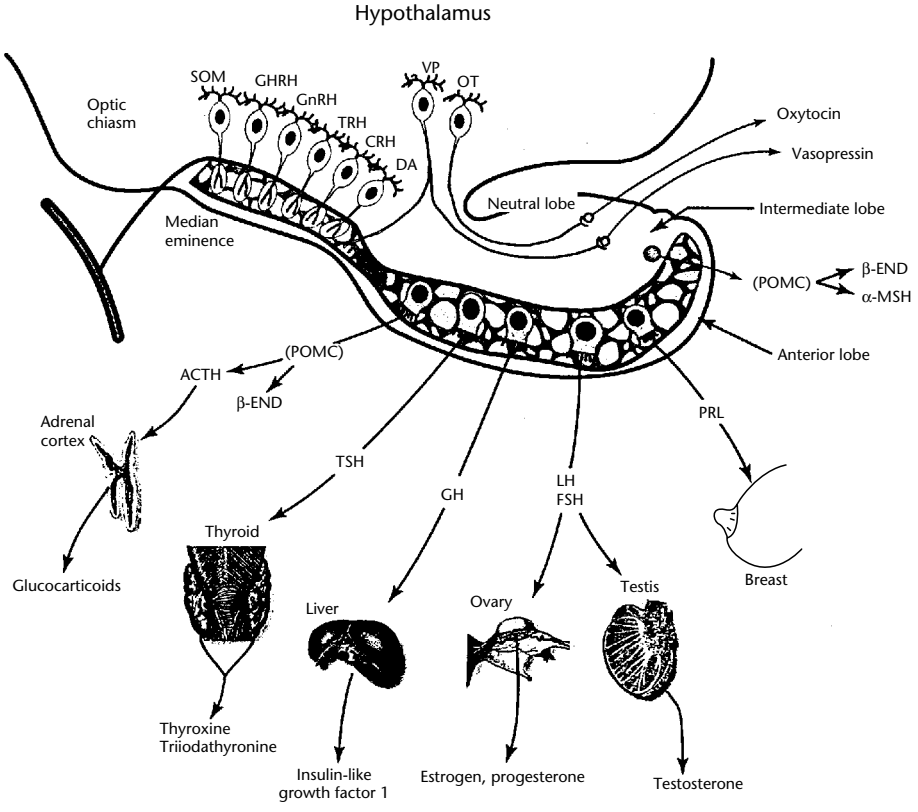


Figure 2. Organization of the hypothalamic control of the endocrine system. Hypothalamic release and release-inhibiting hormones: CRH, corticotropin (ACTH) releasing hormone; DA, dopamine, the prolactin release-inhibiting hormone; GnRH, gonadotropin releasing hormone; GHRH, growth hormone releasing hormone; TRH, thyrotropin releasing hormone; SOM, somatostatin (growth hormone release-inhibiting hormone). Anterior pituitary hormones: ACTH, adrenocorticotropic hormone; β-END, beta-endorphin; FSH, follicle stimulating hormone; GH, growth hormone; LH, luteinizing hormone; PRL, prolactin; TSH, thyroid stimulating hormone. Posterior pituitary hormones: VP, vasopressin; OT, oxytocin. POMC, proopiomelanocortin; MSH, melanocyte stimulating hormone.

further increases in cortisol secretion. Negative feedback is mediated by receptors for the various endocrine hormones within the hypothalamus, for example cortisol feeds back directly to CRH neurons via cortisol (glucocorticoid) receptors within those neurons. Hypothalamic hormone receptors do more than mediate negative feedback. For example, receptors for the gonadal steroid hormones, androgen and oestrogen, not only participate in negative feedback regulation of androgen and oestrogen secretion, but also mediate the stimulatory effects of these hormones on social and reproductive behaviors. During early development, gonadal steroid hormones act on these receptors to guide the sexual differentiation of the hypothalamus.

The posterior pituitary hormones, vasopressin and oxytocin, are produced by other, larger (magnocellular) neurons within the paraventricular and supraoptic nuclei. These hormones are transported down the pituitary stalk within the axons of the hypothalamic neurons that make them and stored in axon terminals within the posterior pituitary. When the hypothalamic neurons are excited by the appropriate neural inputs, oxytocin and vasopressin are released from the posterior pituitary into the blood. For example, nursing, through stimulation of the nipples, excites hypothalamic neurons that release oxytocin into the bloodstream. Oxytocin acts on the breasts to produce milk ejection. This is an example of how some hypothalamic neurons act as neuroendocrine transducers. The brain responds to nipple stimulation and relays this somatosensory information to the hypothalamus via neural (synaptic) signals. Oxytocin neurons respond to these neural signals and produce a hormonal output.

FUNCTIONS OF HYPOTHALAMIC NUCLEI

Examples of the many aspects of physiology and behavior in which hypothalamic nuclei participate are shown in Table 1. It is important to note that more than one nucleus may be involved in a particular function. This is illustrated in the following discussion of the hypothalamic control of eating behavior.

The neural mechanisms for chewing, swallowing, and even reacting to the taste of food, are located posterior to the hypothalamus. Early animal research suggested that the hypothalamus is important for the motivation to eat (appetite), but this idea has been disputed. The idea that neurons within the medial and lateral hypothalamus interact to regulate appetite was first proposed in the 1950s, based on the dramatic effects of hypothalamic lesions on eating and body weight in animals. Lesions of the medial hypothalamus produce overeating and obesity, while lesions of the LH reduce appetite and body weight. This led to the 'dual center' model: the idea that the LH was the location of a 'hunger center', while the medial hypothalamus was the location of a 'satiety center'. Satiety refers to the natural state that follows eating when appetite is inhibited. Increasing activity of the medial hypothalamic satiety center during and following eating was thought to inhibit activity in the lateral hypothalamic hunger center and thus terminate a meal. This model was eventually rejected because of the following findings. First, damage to the LH does not selectively reduce appetite. Animals with extensive bilateral damage to the LH initiate little or no eating or drinking; in fact, they initiate very little behavior of any kind. Second, destruction of axons that merely passed through the LH could duplicate the effects of LH lesions. As mentioned above, the nigrostriatal dopamine pathway passes through the LH on its way to innervate the basal ganglia. The finding that selective destruction of dopamine neurons at any point along the nigrostriatal pathway produced the same deficits as nonselective damage to the LH cast doubt on the idea that cell bodies intrinsic to the LH play a role in eating behavior. Finally, the idea of a medial hypothalamic satiety center was rejected when it was found that the overeating and obesity that follow bilateral damage to the medial hypothalamus could be explained by a metabolic change produced by the lesions. Bilateral lesions of the medial hypothalamus that include the ventromedial or paraventricular nuclei produce chronic hyperinsulinemia, the continuous oversecretion of insulin. Damage to the medial hypothalamus disrupts the neural regulation of insulin secretion by removing an inhibitory influence on the parasympathetic innervation of the pancreas. The resulting abnormally high blood insulin levels cause much of the glucose that is absorbed after a meal to be immediately put into storage in the form of fat. Thus it is possible that animals with medial hypothalamic lesions overeat not because of a lack of satiety, but because they must keep eating to ensure that they have enough glucose and other nutrients in their blood to meet their immediate energy requirements. However, recent discoveries have renewed interest in the idea that the hypothalamus regulates appetite, the motivation to eat, not merely the autonomic regulation of insulin secretion.

Table 1. Hypothalamic nuclei and some of the functions in which they participate. Nuclei are listed according to their longitudinal position, i.e. anterior nuclei are at the top of this list, posterior nuclei at the bottom

Preoptic	Fertility: neurons scattered through the preoptic area produce GnRH, essential for fertility in male and female. In humans and other primates GnRH neurons are found throughout the longitudinal extent of the hypothalamus
Medial	Gonadal hormones act here to promote male sexual behavior and parental behavior. The medial preoptic area contains the sexually dimorphic nucleus, which is larger in male rats. The human hypothalamus also contains a sexually dimorphic nucleus, the INAH-3, in a similar location Temperature-sensitive neurons regulate body temperature by autonomic responses (dilation of blood vessels) and behavioral responses (seeking or avoiding warmth)
Ventrolateral	Neurons here induce sleep through inhibition of histamine neurons in the posterior hypothalamus
Suprachiasmatic	Neurons generate 24 h (circadian) rhythms of endocrine and behavioral responses
Supraoptic	Source of blood-borne oxytocin and vasopressin
Anterior	Neurons facilitate defensive and aggressive behavior
Paraventricular	
Magnocellular	Source of blood-borne oxytocin and vasopressin
Parvocellular	Regulates thyroid via production of thyrotropin releasing hormone (TRH) Integrates responses to stress: • produces CRH • projections to medulla and spinal cord increase behavioral arousal and activate sympathetic nervous system to increase heart rate and blood pressure Regulation of eating and body weight: • a site where peptides act to stimulate or inhibit appetite • projections to medulla regulate parasympathetic innervation of gastrointestinal tract and pancreas
Ventromedial	In rats, estrogen and progesterone act here to induce female sexual receptivity
Arcuate	Produces growth hormone releasing hormone (GHRH) Neurons here release dopamine into the median eminence to inhibit pituitary prolactin secretion Leptin acts here to inhibit appetite
Lateral hypothalamus	Activity here increases arousal, appetite and thirst

CRH, corticotropin releasing hormone; GnRH, gonadotropin releasing hormone; INAH, interstitial nucleus of the anterior hypothalamus.

The discovery of leptin in 1994 provided new insight into the regulation of appetite and body weight. Leptin is a hormone made by fat cells. The blood level of leptin informs the brain of the amount of fat the body is carrying. The more fat, the more leptin in the blood. Leptin inhibits appetite and thus serves as a negative feedback signal to help regulate eating and body weight. Leptin receptors are present in the hypothalamus, including the medial hypothalamus, and injections of leptin into the cerebral ventricles inhibit eating in animals. Together these findings suggest that leptin acts on neurons within the medial hypothalamus to inhibit appetite.

The discovery of leptin, along with the discovery of several hypothalamic peptides that stimulate appetite, has led to a new model of how neurons within the medial and lateral hypothalamus interact to regulate appetite (Figure 3). Neuropeptide Y, orexin A and melanin-concentrating hormone all stimulate eating when injected into the rat brain. Neuropeptide Y is made by neurons of the arcuate nucleus, part of the medial hypothalamus, while the other two peptides are made by separate populations of LH neurons. The demonstration that LH peptides stimulate eating has revived the idea that neuronal cell bodies within the LH have a role in regulating appetite. So although the LH is no longer thought of as the 'hunger center', LH neurons are part of the neural network that gives rise to hunger.

CONCLUSION

The hypothalamus is a critical interface between the brain and the rest of the body. It influences nearly every aspect of physiology and behavior. It integrates endocrine, autonomic and behavioral responses that must occur together for an animal to interact successfully with its environment. Although its role in initiating goal-directed behavior (motivation) has been controversial, evidence

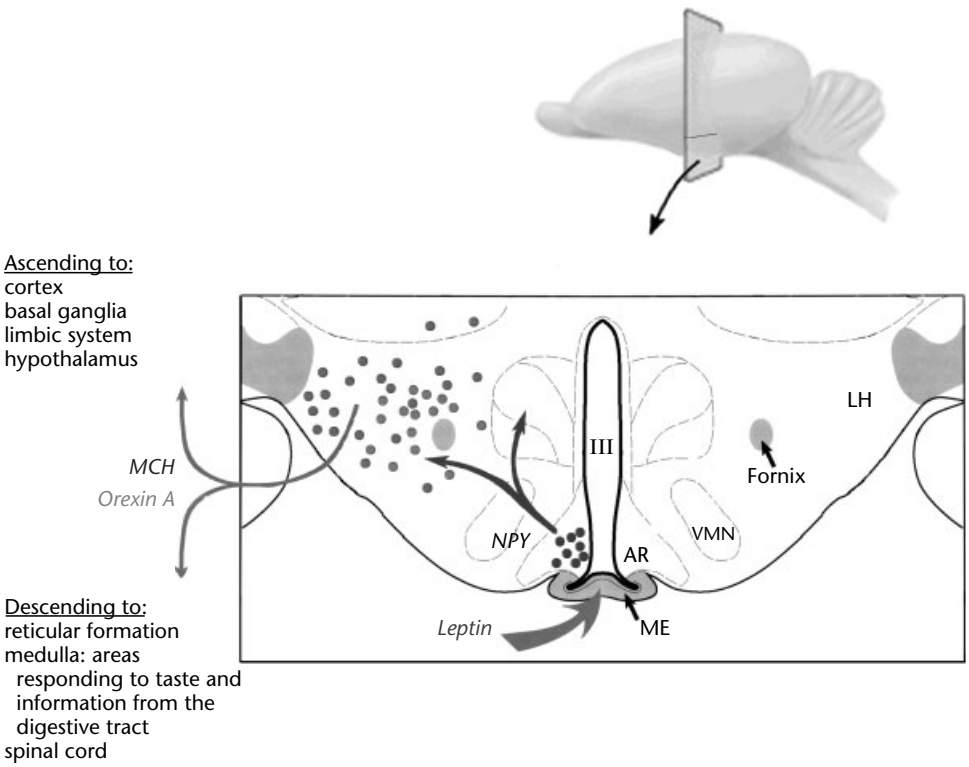


Figure 3. A section through the rat hypothalamus depicting how interactions between the hormone leptin and neurons of the medial and lateral hypothalamus (LH) regulate appetite. Anatomical structures are labeled on the right half of the section, and peptide-containing neurons are shown on the left half. Neuropeptide Y (NPY), orexin A and melanin-concentrating hormone (MCH) are appetite-stimulating peptides. Neurons in the arcuate nucleus (AR) release NPY that excites orexin A and MCH neurons in the LH. Orexin A and MCH neurons have similar widespread projections to other brain regions including those concerned with behavioral activation and arousal which accompany food deprivation and food-seeking. In this model leptin inhibits appetite by directly inhibiting NPY neurons in the medial hypothalamus. This reduces the excitatory influence of NPY neurons on the MCH and orexin neurons in the LH. It should be noted that the MCH and orexin neurons influence other behaviors as well as eating. For example, it has recently been discovered that the sleep disorder, narcolepsy, is caused by the loss of hypothalamic orexin neurons. ME, median eminence; VMN, ventromedial nucleus; III, third ventricle.

suggests that it does have this role in regulating eating behavior. That is, hypothalamic neurons are part of the network of neurons that mediates the motivation to eat (hunger or appetite).

Further Reading

Akil H, Campeau S, Cullinan WE *et al*. (1999) Neuroendocrine systems I: Overview – thyroid and adrenal axes. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 1127–1150. San Diego, CA: Academic Press.

Baum MJ (1999) Psychosexual development. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 1229–1244. San Diego, CA: Academic Press.

Card JP, Swanson LW and Moore RY (1999) The hypothalamus: an overview of regulatory systems. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 1013–1026. San Diego, CA: Academic Press.

Kilduff TS and Peyron C (2000) The hypocretin/orexin ligand-receptor system: implications for sleep and sleep disorders. *Trends in Neuroscience* **23**(8): 359–365.

Petrovich GD, Canteras NS and Swanson LW (2001) Combinatorial amygdalar inputs to hippocampal domains and hypothalamic behavior systems. *Brain Research Reviews* **38**(1–2): 247–289.

Risold PY, Thompson RH and Swanson LW (1997) The structural organization of connections between hypothalamus and cerebral cortex. *Brain Research Reviews* **24** (2–3): 197–254.

Sawchenko PE (1998) Toward a new neurobiology of energy balance, appetite, and obesity: the anatomists weigh in. *Journal of Comparative Neurology* **402**: 435–441.

Siegel A, Roeling TAP, Gregg TR and Kruk MR (1999) Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience and Biobehavioral Reviews* **23**: 359–389.

Kindling

Introductory article

Deborah M Saucier, University of Saskatchewan, Saskatoon, Canada
Michael E Corcoran, University of Saskatchewan, Saskatoon, Canada

CONTENTS

Kindling is a model of complex partial epilepsy and neuroplastic change within the nervous system.

INTRODUCTION

The first description of a kindling-like effect was the observation made by Watanabe in 1936, who electrically stimulated the cerebral cortex of freely moving dogs and found that convulsive seizures developed progressively, eventually resulting in the occurrence of spontaneous convulsions. Studies from Jose Delgado's laboratory at Yale in the late 1950s and early 1960s found that behavioral seizures developed during ictal electroencephalic (EEG) discharge and that ictal EEG grew as the number of stimulations progressed. However, these observations were largely left unnoticed by the general scientific community until the seminal paper by Graham Goddard in 1967.

Intriguingly, research on intracranial self-stimulation (ICSS) had previously documented the development of seizures (both electrographic and behavioral manifestations) as a side effect of the ICSS procedure, which is a means of investigating the neural basis of reinforcement. However, the seizures that sometimes accompanied ICSS were viewed primarily as an artifact and a potential confound for those studying the neural basis of reinforcement, rather than as a phenomenon interesting in itself.

The discovery of kindling arose primarily from a serendipitous error that occurred during Graham Goddard's doctoral research, which was to study the effects of electrical stimulation of the amygdala on conditioning in rats. Owing to an equipment malfunction he accidentally delivered repeated trains of fairly high-intensity stimulation to the amygdala of some rats, one of which developed seizures after a few repetitions. The rat continued to display seizures even after the equipment had been repaired, and Goddard observed that similar stimulation given to naive rats did not cause seizures. In other words, it was the repeated application of stimulation that led to seizure, rather than the stimulation itself. Goddard's additional insight was that the brain itself might have been changing in response to the repeated application of stimulation, and that this phenomenon might model neural mechanisms of learning.

Goddard's subsequent research not only established the parameters of kindling, but also demonstrated a number of the fundamental characteristics associated with this phenomenon. He and his colleagues found that repeated application of brief trains of electrical stimulation leads to the development of generalized convulsions in rats and that this state of induced susceptibility to convulsion is seemingly permanent. They demonstrated that kindling depends on the distributed application of high-frequency stimulation and that it is unlikely to be due to pathological effects of electrical stimulation such as tissue damage, poison, edema, or gliosis. Finally, they confirmed that kindling occurs when similar treatments are applied to the cat and the monkey; it is not unique to the rat's central nervous system.

Although Ronald Racine may not have been the first to recognize kindling as an important phenomenon, the field of kindling has been (and continues to be) profoundly influenced by his research, beginning in 1969, and especially by two papers published in 1972. Racine was the first to recognize that the EEG events accompanying kindling must be recorded, allowing the subtle neural changes (which are not evident in the convulsions) to be studied. Racine's research was the first to characterize the EEG changes occurring in response to the repeated application of electrical stimulation. Racine developed a rating scale for classifying behavioral seizure development that is now used

almost universally in kindling research. He also clarified the relations between repeated stimulation, ictal discharge, and kindling. Finally, he imagined that the neural changes underlying the kindled state might be widespread throughout the brain and he developed procedures to test this hypothesis. Using these techniques, Racine demonstrated that kindling involves widespread transsynaptic changes. Together with Goddard's initial experiments, Racine's research defined the kindling phenomenon.

Although other people made significant contributions during the early days of kindling research, the final person that we shall focus on here is Juhn Wada, a clinical neurologist. Wada performed seminal experiments in cats in which he discovered that establishment of a cortical focus with ethyl chloride and aluminum hydroxide led to the development of independent foci in other cortical and subcortical sites and to the evolution of behavioral seizures. In the early 1970s Wada began to investigate the susceptibility of Senegalese baboons (*Papio papio*) to kindling. Further research resulted in the study of the characteristics and mechanisms of kindling in rats, cats, monkeys, and baboons. This work produced, among other things, the first systematic and extensive descriptions of kindling in cats, baboons, and monkeys. Importantly, Wada was among the first to describe the occurrence of spontaneous seizures during kindling. Less tangibly but no less importantly, the endorsement of kindling by a prominent neurologist such as Wada, in his research papers, in presentations at meetings, and in the series of books he edited, eased its acceptance by the neurology community as a legitimate and important preparation in experimental epilepsy.

MAIN FEATURES OF KINDLING

Kindling involves the progressive development of seizures and an enduring increase in seizure susceptibility in response to repeated electrical stimulation in the brain. Convulsions can be triggered when low-intensity current of brief duration is repeatedly passed through the uninsulated tip of chronically implanted electrodes within certain areas of the brain. Initial simulations result in epileptiform afterdischarges of short duration. Stimulation intensity is usually at the afterdischarge threshold, which is arbitrarily defined as the lowest intensity of stimulation that evokes an afterdischarge at the stimulation site. The initial seizure activity consists of low-amplitude afterdischarge spikes, and there is a limited propagation to other sites in the brain. Repetition of the stimulation at a fixed interval (usually 24 h) and intensity will lead to an increase in seizure activity manifested by typical behavioral responses and EEG patterns. Typically kindling stimulations are a 1 s train of constant current biphasic square wave pulses, each pulse with a duration of 1.0 ms. Biphasic pulse pairs are delivered at a rate of $60\,s^{-1}$.

Kindling is persistent. In limbic sites there is typically no remission of the phenomenon observed. Kindled animals can display spontaneous seizures, typically after many repeated stimulations. Kindling does not appear to result from focal tissue damage. Lesioning of a kindled site does not affect subsequent kindling from other sites, which indicates that lesion-induced damage does not itself kindle seizures or facilitate electrical kindling. Further, lesions at the tip of the implanted electrode interfere with kindling from the stimulated site. This suggests that intentional gross damage to the stimulated site does not facilitate kindling; rather, it interferes with it. Additionally, homotopic contralateral unstimulated sites will show a saving in the number of stimulations required to produce a generalized convulsion, indicating that kindling is a trans-synaptic plastic change that occurs throughout much of the brain of the kindled animal.

Kindling occurs in response to chemical as well as electrical stimulation. Carbachol, a cholinergic agonist, effectively kindles seizures. When carbachol is infused into the brain at doses that are initially subconvulsive, there is a progressive increase in seizure activity and severity of convulsions. Picrotoxin, a γ-aminobutyric acid (GABA) antagonist, kindles seizures when injected either into the amygdala or intraperitoneally at subconvulsive doses. Morphine or opioid peptides also kindle seizures when repeatedly injected intracerebrally at initially subconvulsive doses.

Chemically induced kindling is fully transferable to electrically induced kindling. The strength of transhemispheric transfer between chemical and electrical kindling is as strong as transhemispheric transfer between electrical and electrical kindling. The transfer data provide strong evidence that the mechanisms underlying chemically induced kindling are the same as those underlying electrically induced kindling. In general, any procedure that evokes sustained and rapid burst firing of neurons has the potential to kindle convulsions. Thus, the most important requirement of kindling is not the form that the stimulus takes, but whether it evokes burst firing of neurons, as do carbachol and specific forms of electrical stimulation.

The developing brain appears to be more seizure-prone than the adult brain. Rats given a

single subconvulsive exposure neonatally to a proconvulsant treatment (e.g. kindling simulation, exposure to heat) and examined as adults displayed decreased thresholds for kindling and an increased rate of kindling. However, the rate of kindling in immature rats is more rapid than that observed in adults, suggesting that the immature brain is more hyperreactive than the mature brain. Given the results from early exposure to epileptogenic experiences and their ability to permanently affect future plasticity, we have good evidence that the immature brain has a predisposition toward hyperexcitability.

The immature brain is not completely homologous with the adult brain. For instance, the immature brain is hypersensitive to the glutamate agonist, *N*-methyl-D-aspartate (NMDA), injections of which result in intense seizures and hippocampal damage, whereas injections of kainic acid produce intense seizures with relatively little damage. In the adult brain, this pattern is reversed: NMDA injections produce much less severe seizures, whereas kainic acid injections are much more toxic and produce widespread damage. Further discrepancies between the immature and the mature nervous systems can be observed in the response to infusions of muscimol (a $GABA_A$ receptor agonist). In the developing nervous system, muscimol is proconvulsive, whereas in adult rats it is an anticonvulsant. These results suggest that the immature brain is differently disposed to hyperexcitability, although regardless of the treatment, epileptiform activity in the infant enhances future seizure susceptibility.

KINDLING AS A MODEL: EPILEPSY AND NEUROPLASTICITY

Kindling as a Model of Epilepsy

Epilepsy is a disorder characterized by recurrent and unpredictable episodes of seizure activity. The hallmark for diagnosis is two or more unprovoked seizures combined with epileptic spikes occurring while an EEG is recorded. The prevalence of a second seizure is approximately 40% in those who have had a single unprovoked seizure. Treatment of epilepsy typically aims to prevent or suppress seizures through the use of antiepileptic drugs.

Many types of epileptic seizures occur and the frequency and form of seizures is variable. However, two major types of seizures are typical and are usually described as either partial or generalized. Partial seizures are ones in which seizure activity is limited to one area of the brain. Complex partial seizures (also referred to as temporal lobe epilepsy) may involve complicated motor acts and impairments of consciousness during the motor component of the convulsion. Generalized seizures refer to the involvement of the whole brain and typically involve two-phase generalized convulsions: the tonic phase, which involves the loss of consciousness and falling with body rigidity; and the clonic phase, during which the body extremities jerk and twitch.

Many facets of kindling make it a good model of complex partial epilepsy in humans. First and foremost, kindling is the progressive development of seizures, which makes it very similar to post-traumatic epilepsy, and it can result in the development of repeated spontaneous seizures, a hallmark of epilepsy. However, kindling also exhibits many other features that make it a good model of complex partial epilepsy.

Although complex partial seizures can occur throughout the brain, in humans they typically occur in the temporal lobe. In nonhuman animals, parallel limbic areas (including the amygdala, hippocampus, and septum) kindle readily. In humans, partial seizures are initially limited to one site in the brain, which is what is observed during the initial stages of kindling in limbic sites. The afterdischarge and subsequent EEG that are observed following a kindling stimulation in animals are very similar to those of a human undergoing an epileptic seizure. The form of the convulsions resulting from early kindling in the amygdala is very similar to that observed in humans with complex partial epilepsy. Kindling, like epilepsy, may progress from complex partial seizures to become more severe and generalized. Like generalized convulsions in humans, kindled generalized seizures in nonhuman animals are marked by tonic–clonic convulsions, including falling with body rigidity, and body extremity jerking. Thus, kindling shares anatomical, EEG, and convulsive features with epilepsy.

Importantly, drugs that are effective in treating epilepsy have similar effects on kindling. Treatments with barbiturates, GABA agonists or benzodiazepines are all effective in retarding kindling. Kindling therefore enables researchers to test new treatments for epilepsy on brains that are truly epileptic. This is not the case with seizures that are triggered once only, as is the case with electroconvulsive shock or PTZ seizures. However, for those studying the causes or treatment of epilepsy the most valuable feature of kindling as a model of complex partial epilepsy is its predictable development. In the limbic system the reliable rate of

kindling provides researchers with excellent parametric control over the seizures.

There is one major difficulty in firmly asserting that kindling thoroughly models complex partial or generalized seizures. The definition of both complex partial and generalized seizures requires the impairment or loss of consciousness during the seizure. Although nonhuman animals may exhibit a number of features that are suggestive of the loss of consciousness, nonhuman animals are essentially nonverbal. As our window onto consciousness is primarily verbal, we cannot definitively state that kindling results in an impairment or loss of consciousness in nonhuman animals.

Kindling as a Model of Neuroplasticity

In 1948 Jerzy Konorski coined the phrase 'synaptic plasticity', conceiving it as a long-lasting change in neural activity. Although Konorski may have coined the term, it was Donald Hebb in 1949, with the publication of *The Organization of Behavior*, who has had the greatest impact on our thinking about plasticity phenomena. Hebb suggested that it was the plastic nature of neurons and their ability to change their responses that was the basis of learning in the nervous system. In 1949 Hebb stated, 'any two cells or systems of cells that are repeatedly active at the same time will tend to become associated, so that activity in one facilitates activity in the other.' This definition still describes what is understood as neuroplasticity today.

Kindling involves an increase in the responsiveness of neurons to previously weak or ineffective stimuli. This change is persistent, and may even be permanent – certainly it is a long-lasting change in neural activity. We now know that neuroplasticity can result from the growth of new synapses, the activation of previously inactive synapses, and changes in efficacy of synapses, among other things. Importantly, the changes associated with kindling are widespread and trans-synaptic, suggesting that the entire nervous system is responding to the kindling stimulations. Kindling may modify the efficacy of synapses, and it has been suggested that kindling may result from the activation of previously dormant synapses. Further, research suggests that in adult rats, markers for synaptic growth appear following the application of initial kindling stimulation, leading to speculation that kindling results in the production of new synapses. Thus, at a basic level, kindling results in plastic change of the nervous system that is specific to kindling itself.

Recall, however, that when Goddard first observed the kindling phenomenon he thought that perhaps kindling was a model for learning. Today kindling is not thought of as a model for learning; rather it is a model of pathological neuroplastic change. In fact, kindling in the limbic system has been shown to interfere with a number of forms of learning. Moreover, the learning impairments are relatively specific to the site that was kindled: it has been demonstrated that hippocampal kindling, but not kindling of the amygdala, interferes with tasks that are thought to be mediated by the hippocampus.

Kindling does share superficial similarity with another model of neuroplasticity, long-term potentiation (LTP). The latter is marked by a long-lasting increase in the responsiveness of neurons that have been exposed to a brief electrical stimulation. Similar potentiation can occur following kindling stimulation, and is known as kindling-induced potentiation (KIP). The fact that potentiation occurs with kindling, and the similarities between the techniques for the induction of LTP and kindling, have resulted in many discussions of whether common mechanisms underlie these two phenomena. Although some similarities are observed at the molecular layer (for example, the participation of various protein kinases in the production of both KIP and LTP), there are some significant differences between KIP and LTP. Most importantly, unlike kindling, LTP does not require the production of afterdischarge, and repeating the stimuli required to produce LTP does not result in afterdischarge or kindling. Unlike kindling and KIP which are extremely persistent, LTP typically decays to baseline within several days or weeks. Another difference is that various brain structures often show very different propensities to kindle and undergo LTP.

There are large distinctions in the neural activity that produce LTP and KIP. Kindling-induced potentiation results from the summation of activity in the many pathways (excitatory and inhibitory) that are activated by the kindling stimulation. The widespread trans-synaptic activation observed in kindling and KIP does not typically occur in LTP paradigms. Typically, LTP results from local activation of relatively few neural circuits. As such, it may be that the difficulties in comparisons between LTP and kindling or KIP result from the fundamental difference between these phenomena in evoking widespread trans-synaptic change, rather than in any meaningful difference between the mechanisms underlying the production of these phenomena.

PUTATIVE MECHANISMS OF KINDLING

The form of the stimulation required to elicit the kindled state can vary; the only requirement is that it must trigger an afterdischarge. Stimuli that do not produce an afterdischarge do not result in kindling. Unlike the parameters and features of kindling, the mechanisms underlying kindling are little known and frequently contentious. We will focus on the anatomical explanations of kindling as well as the intriguing results provided by fast- and slow-kindling strains of inbred rats.

Neuroanatomical Correlates of Kindling

Brain sites vary in the number of stimulations required to produce behavioral seizures. In order of decreasing sensitivity to kindling, the following pattern was found: amygdala, globus pallidus, piriform cortex, olfactory bulb, septal area, preoptic area, caudate putamen, and hippocampus.

Although the piriform and perirhinal cortices produce generalized convulsions rapidly, they may have a critical role in the development of kindling. Limbic sites such as the amygdala are unable to directly support convulsive motor responses, although kindling of the amygdala results in eventual development of convulsive motor responses. It appears that the recruitment of perirhinal cortex may be essential for triggering convulsions, and participation by the piriform cortex may be required for additional amplification of convulsions. The perirhinal cortex has numerous close and reciprocal connections with the frontal motor cortices. Depressing activity within the frontal motor cortex of the rat retarded the development of kindling – reducing the duration of both afterdischarges and convulsions.

Examinations of kindling in the claustrum have demonstrated that this structure is extremely sensitive to kindling, rapidly producing intense and prolonged seizures. Furthermore, unlike many other forebrain structures, convulsive motor responses are produced during the stimulation itself. Conversely, lesions of the claustrum significantly retard the kindling of seizures with stimulation occurring elsewhere in the brain. These results suggest that the claustrum also plays a part in the production of the convulsive motor responses that are characteristic of kindled seizures. Taken together, these results suggest that the claustrum along with the piriform, perirhinal, and frontal motor cortices has a major role in the production of convulsions and thus in the progression of kindling from complex partial seizures to generalized seizures. Moreover, destruction of any one of these structures does not result in the complete cessation of kindling, although kindling is significantly retarded. It therefore appears that there are multiple pathways by which kindling can proceed and develop.

Hippocampal degeneration is often observed in people who have temporal lobe epilepsy. It has been suggested that it is the loss of cells critical for the regulation of normal inhibition within the hippocampus that underlies kindling. Although it does appear that cell loss occurs following status epilepticus (typically induced by kainic acid), the evidence for cell loss following kindling is much less compelling. Some researchers have observed that there is cell loss in the hilus of the hippocampus following kindling; however, others have observed that the volume of the hilus is changed following kindling, confounding calculations of cell density. Furthermore, kindling does not result in increased numbers of microglia, or increases in the numbers of apoptotic cells. Although kindling does appear to increase the number of reactive astrocytes, these increases have also been observed in rats that were not fully kindled, suggesting that astrocytic activation may have been related to neuronal activity rather than to kindling itself. Thus, it does not appear that kindling is strongly associated with cell loss within the hippocampus. The evidence from kindling studies suggests that the hippocampal degeneration observed in patients with chronic epilepsy may be the result of episodes of status epilepticus, not the cause of the epilepsy.

Another area of intense focus within the hippocampus has been the observation that patients with temporal lobe epilepsy also exhibit abnormal sprouting in the excitatory pathways of the hippocampus (the mossy fibers). The mossy fibers are axons of the granule cells of the dentate gyrus and project to pyramidal cells of the area CA3 of the hippocampus. Many people who have epilepsy may have experienced some type of traumatic incident (fever, head injury) prior to the development of epilepsy. It has been suggested that the time course for the development of these abnormal mossy fibers parallels observed time courses between initial trauma and subsequent seizure syndromes. Thus, mossy fiber sprouting (MFS) has been suggested as a putative mechanism of enhanced seizure susceptibility.

Both kindling and status epilepticus can result in the induction of MFS. However, a number of researchers using a variety of species and techniques have demonstrated that MFS can be dissociated from the progressive development of seizures.

These demonstrations include observations that the amount of MFS is not related to the intensity of kindled seizures, and the protein synthesis inhibitor, cyclohexamide, blocks status epilepticus-induced MFS without affecting the expression of status epilepticus itself. Researchers have reported that MFS targeted inhibitory cells and may actually restore inhibitory responses within the hippocampus. Furthermore, MFS has been shown to occur with stimulations that produce LTP and that do not result in kindling. Thus, MFS may actually reflect neural activation, something that undoubtedly occurs during kindling, rather than synaptic modifications that result in enhanced ability of the hippocampus to produce seizures.

Fast- and Slow-kindling Rats

Ronald Racine developed two strains of rats that were highly inbred based on their rate of kindling. The fast strain kindled 40–50% more quickly than normal rats, whereas the slow strain kindled 200–300% more slowly than normal rats. These differences are observed primarily within the amygdala, as both fast and slow strains tend to kindle at normal rates when stimulation occurs in the hippocampus. The fast strain is more susceptible to picrotoxin or bicuculine, which are GABA antagonists, whereas the slow strain require smaller doses of barbiturates (GABA agonists) to produce anesthesia. Thus, GABA-mediated inhibition may have a role in the observed differences between the fast and slow strains.

The fast and slow strains exhibit profound differences in the expression profile of the α subunit of the $GABA_A$ receptors. The fast strain express approximately 50% fewer of the adult inhibitory form of the subunit, $\alpha 1$, whereas they overexpress the neonatal excitatory forms of the GABA α subunits. The slow strain underexpresses neonatal excitatory subunits and expresses approximately 200% more of the inhibitory, $\alpha 1$ subunits. These results are consistent with the fast strains demonstrating an arrested development of the $GABA_A$ receptor system, suggesting that the molecular correlate of the fast kindling phenotype is related to differences in GABA-mediated inhibition. Other researchers have found that hippocampal commissure kindling results in the conversion of $GABA_A$ receptor subunits to include a greater prevalence of neonatal, excitatory forms in otherwise normal adult rats. Additionally, following status epilepticus in adult rats, researchers have observed an upregulation of the messenger RNA for the excitatory, neonatal forms of $GABA_A$ simultaneous with downregulation of the inhibitory, adult-form mRNA in otherwise normal adult rats. There have been numerous demonstrations that changes in $GABA_A$ inhibition occur following seizures and kindling. It seems likely that the mechanisms that underlie kindling are related to the results obtained by examining these fast and slow strains.

CONCLUSION

Kindling is the predominant model of clinical epilepsy. Kindling can occur following electrical stimulation or the injection of proconvulsant chemicals as long as the stimulation produces afterdischarges. Studies of the mechanisms of kindling will continue to provide important insights into the causes and treatment of epilepsy.

Further Reading

Corcoran ME and Moshe S (1998) *Kindling 5*. New York, NY: Plenum.

Goddard GV (1967) Development of epileptic seizures through brain stimulation at low intensity. *Nature* **214**: 1020–1021.

Goddard GV, McIntyre DC and Leech CK (1969) A permanent change in brain function resulting from daily electrical stimulation. *Experimental Neurology* **25**: 295–330.

Racine RJ (1972) Modification of seizure activity by electrical stimulation. I. After-discharge threshold. *Electroencephalography and Clinical Neurophysiology* **32**: 269–279.

Racine RJ (1972) Modification of seizure activity by electrical stimulation. II. Motor seizure. *Electroencephalography and Clinical Neurophysiology* **32**: 281–294.

Levels of Analysis in Neural Modeling

Introductory article

Peter Dayan, University College London, London, UK

CONTENTS

Neural systems are analyzed by building multiple levels of descriptive and explanatory mathematical models. Parallel to these are interpretive computational models, showing how information is represented and manipulated.

LEVELS OF ORGANIZATION IN THE BRAIN

The construction and refinement of qualitative and quantitative models for observed phenomena is the standard practice in neuroscience, just as it is in all other scientific disciplines. What particularly distinguishes neural modeling, apart from the sheer complexity of the phenomena, is that brains are computational devices, transducing, representing, manipulating and storing information, and using this information to control actions. The most comprehensive neural models must therefore play the dual role of accounting for experimental data and interpreting it in terms of underlying computations. Understanding what neural models are, what their connections to experimental data are, and how to build and use them in practice, is tricky because of the complexity and the multiple goals for modeling.

One key idea for structuring the enterprise of modeling that has many, and confusingly different, manifestations is that of different levels of organization. The confusion comes because there are at least four different, though partially overlapping, threads to this idea. Understanding these threads of levels of organization is fundamental to understanding the construction and critique of neural models.

The first thread is the standard one of scientific reduction, describing observable phenomena in qualitative and quantitative detail, and explaining them in terms of descriptions of their underlying substrates at lower and less abstract levels.

The second thread is that of the construction or synthesis of systems to execute some particular task. Here, the use of different levels is a conventional divide-and-conquer strategy in the face of design complexity. For instance, programmers building large computer systems decompose complicated functions (at higher levels) using simpler subroutines and procedures (at lower levels). Modern object-oriented programming methods take this one stage farther.

The third thread, originally suggested by David Marr, is associated specifically with computational modeling. It involves a computational level, which is an abstract description of the goal of the task, and the logic of the strategy by which it is to be satisfied; an algorithmic level, which is a description of the way that information should be represented and algorithmically manipulated in the service of carrying out the task; together with an implementational level, which is a description of the way that these representations and manipulations are instantiated in neural hardware. To avoid confusion, we will reserve the word 'planes' for these three computational levels.

The fourth and final thread is that of levels of processing as a strategy for manipulating and extracting information from input. This is well exemplified by the visual system of primates. In some primates, tens of different structurally and functionally defined areas of the brain are devoted wholly or partially to analyzing visual information. On the basis of various sources of evidence, including characteristic layers of termination of connections, these areas appear to be organized in a somewhat loose hierarchy. This hierarchy starts from the lowest level in primary visual cortex, where cells tend to have comparatively small receptive fields (i.e. are directly responsive only to a small area of visual space) and simple response properties (for instance, responding to small bars of light at particular orientations). The hierarchy extends all the way up to much higher levels involving areas such as the inferotemporal cortex,

where cells have large receptive fields and complex response properties (for instance, responding to pictures of particular classes of faces, but not to other apparently similar objects). Similarly, when building machines to process images for such tasks as object recognition, one can build a processing hierarchy, in which the visual input is subjected to sequences of manipulations leading to the answer.

The idea in the fourth thread that visual and other inputs might be processed hierarchically is quite distinct from the idea in the other three threads about the hierarchical organization of the analysis of a complex system such as a brain. The latter, which is the main focus of this article, applies whatever form neural computations happen to take – it is about describing and explaining the behavior of systems of any sort.

TYPES OF NEURAL MODEL

Conventional Reductive Models

The first thread to the idea of levels concerns standard reductive modeling. For the practical (as opposed to the philosophical) aspects of this, models are useful in at least two ways: describing neural phenomena, and providing a means for reductionist explanations of the phenomena, by appealing to the mechanisms that might actually be responsible for generating the phenomena in the first place. These mechanisms themselves are usually understood in terms of models. Thus, the modeling process is recursive in the sense that explanatory models at one level are built from descriptive or explanatory models at a lower level; explanatory models at the lower level are built from descriptive or explanatory models at a yet lower level. Although models need not be expressed as a set of mathematical equations, quantifying them in some way is essential to be able to check whether the mechanisms postulated (or rather the models of the mechanisms postulated) are really capable of capturing the phenomena to the desired degree of accuracy.

One example of this reduction is the phenomenon of the behavior of an interconnected network of neurons. This can be described in terms of a (relatively complicated) dynamic state description, but it might also be explained in terms of descriptions of the properties of the individual neurons and their connections. In turn, the descriptions of the individual neurons can be explained in terms of their detailed geometry and the membrane-bound channels they contain.

Another example is modeling the shape over time of the voltage inside an axon (relative to extracellular space) during the passage of an action potential or spike. The data can be characterized at some level of detail, such as the time course of the voltage averaged over a large number of spikes. The data can be summarized, at least within the approximation of experimental error, by any number of different descriptive quantitative models. For instance, one could tune the parameters of various high-order polynomials or piecewise linear functions to replicate the shape of the spike accurately. In contrast to these nonmechanistic descriptions, Alan Hodgkin and Andrew Huxley, in their classic study, were interested in explaining how action potentials arise, and therefore constructed a model on the basis of what they expected about the way that cells might manipulate the potential difference between inside and outside. They had evidence that the basic mechanism was some way for the axon to change its permeability to particular ions, so they postulated as a mechanism a form of channel or gate in the cell's membrane. They built a quantitative explanatory model of the action potential from a quantitative descriptive model of the gate, fitting the parameters of the model to make correct the form of the overall action potential it produced.

Following the advent of more powerful experimental techniques, we can now understand Hodgkin and Huxley's deterministic phenomenological channel model in terms of a large number of tiny individual channels in the membrane which open and close in a stochastic manner. The summed effect of all these individual channels matches Hodgkin and Huxley's model very closely. So we can now build an explanatory model of the action potential using descriptive models of these gates. Based on even more recent experimental evidence, we could go further, and try to build explanatory models of these gates from our knowledge of their molecular structure.

The general conclusions from such examples are that there are different levels of model, corresponding to different levels of reduction of a phenomenon, and, often in neuroscience, different levels of anatomical detail. At a level, there are descriptive models which capture the behavior without much regard to the substrate, and explanatory models, which capture the behavior by reducing it to models at lower levels. Quantitative models tie together the different levels because they allow proof, or at least numerical demonstration, that the behavior assumed at one level can truly account for the behavior that it is intended to explain.

Almost always, the models at a given level are only approximate, with models at lower levels being faithful to more intricate details of the experimental observations.

One can think of the models at different levels as constraining each other – the model of the stochastic channel has to behave correctly to produce the overall spike, if such gates are really to be responsible for generating action potentials. Similarly, explanatory models of the action potential are constrained by what is possible at lower levels. One can also think of the different levels as liberating each other – there are many different gating mechanisms that have the same qualitative behavior – so if one is only interested in the effect of gating and not its cause, then one need not be concerned with the explanatory details of the model. In practice, there is not really such a strong separation between explanatory and descriptive models. For instance, Hodgkin and Huxley did not just build any descriptive model of their phenomenological gate. Indeed, so close was their descriptive model to the true underlying mechanism, that we can now interpret in terms of the latter some of the parameter values that they found by their fitting process.

Computational Interpretive Models

Computational models start from the premise that many of the tasks for brains are best characterized as involving computations. That is, brains solve computational problems, and it is a matter only of taste whether they are referred to as electrochemical computers. For example, for a bee to forage optimally in a field containing two types of flowers with different characteristics of provision of nectar, it must continually compute the choice between sorts of flower to land on. Equally, to catch a flying ball in the hand, the brain has to transform the visual input about where the ball is and how it is moving into a sequence of motor commands to move the hand to an appropriate place at an appropriate time, with the fingers arranged correctly. This transformation happens in various parts of the visual and motor systems of the brain.

Computational modeling is about imputing a computational task and interpreting the collective behavior of the neural components of the system in terms of this task. In these cases, the tasks involve generating the appropriate output, such as the choice of the better flower or the sequence of actions to catch the ball, from the input, such as the past qualities of the flowers or the visual impression of the ball over time.

More concretely, computational modeling involves understanding (i.e. interpreting) how neural machinery can represent and store information about aspects of the outside world, and how it can transform information from one form to another in the performance of tasks. Typically, representation involves the simultaneous spiking activities of large, distributed populations of neurons; storage involves both the persistent activity of neuronal populations and changes in the efficacies or strengths of synaptic connections between neurons; and transformations are occasioned by the propagation of activity within and between populations of neurons. As in a standard computer, the semantics of the computation, that is, its computational meaning, are implemented by the syntax of the physical substrate, for example the biophysical rules according to which activity propagates and synapses change their strengths.

Computational models can themselves be decomposed into the three planes of Marr described above, from an abstract description of the underlying task, through a more concrete description of the representations and algorithms adopted to satisfy the task, to a description of how these representations and algorithms are actually implemented. There is not necessarily a single description in the computational and algorithmic planes – there can be many different algorithmic ways of expressing the same transformation. For instance, to work out the product $a \times b$ of two positive integers, one can add a to itself b times, or b to itself a times, or use the standard long-multiplication method most of us learned at school, or add together the natural logarithms of a and b, and exponentiate the result. These different algorithms share the same description on the computational plane.

Computational models satisfy many of the same properties as conventional models. First, in the same way that there are different levels of conventional models, there are different levels of computational models, paralleling a decomposition of the underlying computation. Second, there are both descriptive and explanatory computational models. For example, take the case of a single, spiking, neuron. One can build a descriptive model of what the output of a neuron represents by presenting all possible classes of input to the animal and recording what spikes it produces. However, one can also build an explanatory model by working out what inputs the neuron has, what these inputs represent, and what transformations the neuron performs on these inputs.

The third similarity between computational and conventional models is functionalism, that

the same abstractly defined computation can, in principle, be carried out by many different representations and algorithms, and the same algorithm instantiated in many different forms of hardware. However, there is a caveat to this which has important implications for research programs. When we try to write down computational descriptions for complex tasks, such as identifying the structure of a visual scene, we tend to do so rather succinctly and abstractly. The neural substrate, as an evolved rather than a designed system, and facing implementational constraints such as limited connectivity and speed, is likely to be capable of instantiating only approximations to such abstract requirements. Phenomena such as visual illusions can result: cases in which the visual system fails to satisfy an apparently natural abstract description of a visual task. These then force us to enrich our computational descriptions, and also provide insight into the neural substrate. Therefore, the system does not really possess functional equivalence between the planes – the implementational planes implement functions that might be slightly different from the idealized ones described in the computational planes.

From the perspective of analyzing neural systems, computational and conventional modeling should work hand in hand. Consider a single level. Experimentally observable phenomena lie at the implementational plane of the computational model. Conventional modeling provides an explanatory account of these phenomena by reducing them to a lower level, where, in turn, they comprise the implementational plane of the lower level of the computational model. At each level in the computational model, algorithmic and computational planes are associated with the implementational plane. Thus, parallel to the conventional reduction from one level to the level below, there is a computational reduction from the planes at one level to the planes at the level below.

The ultimate goal is to have conventional and computational models for neural function that are mutually consistent, extending from the lowest levels of molecular dynamics to the highest levels of ethology. We would then have a complete, quantitative, multi-level understanding of *how* the brain executes *which* tasks. Of course, such exhaustiveness is far off at the moment. The coverage of conventional and computational models of any sort is very poor, and appropriate reductions and interpretations are few and far between.

In practice, computational models are often used synthetically rather than analytically. That is, networks of neurons, described down to some level of abstraction, are constructed to perform a computational task, and the behavior of the model neurons is compared with that observed in physiological (and other) experiments, justifying the model. In such synthetic treatments, the multiple levels of computational modeling can be made explicit, and the reduction between the levels can be shown to be exact. One popular synthetic technique is to start from an idea about optimal or normatively correct ways to perform a task, and to seek implementations involving biologically reasonable components.

DEGREES OF MODELING DETAIL

To a very coarse approximation, one can separate out three different classes of quantitative models in common use: conductance-based models, integrate-and-fire models, and firing-rate models. The different models are naturally couched at different levels of abstraction and, as described in general terms above, are used to account for data that are similarly collected at different levels of abstraction. Not all the models that are built and used fit neatly into these categories, but they nevertheless give a flavor of the differing degrees of neural and analytical detail that are regularly employed.

Conductance-based Models

Conductance-based models place their emphasis on describing single or just a few neurons with a high degree of detail. They typically approximate the structure of an individual neuron by multiple, interconnected compartments, each of which is treated as being electrically compact. The whole set of compartments is designed to be more or less faithful to the geometry of the neuron, including such facets as branching points of dendrites and the diameters and lengths of different parts. There is an elaborate art to representing cells that have complex geometrical structures in terms of just a few compartments, or even just a single compartment, a reduction that is often necessary to make computations involving the models adequately quick. In standard conductance-based models, each compartment is given an assortment of active channels, such as voltage-sensitive or synaptic channels.

Conductance-based models of single cells are ideal for explaining phenomena to do with spikes and the thresholds for initiating spikes, the precise effects of synaptic input, bursting, spike adaptation, spikes that propagate backwards up the dendritic tree, and the like. One problem with these models is that there is rarely good experimental

data on the actual locations on the dendritic tree or the strengths of the active channels. Such data are essential to making the models faithful to the neural substrate. More generally, conductance-based models involve a large number of parameters, and the values of only a few of these can be determined from experiments. A second problem with these models is that they are so complex that they cannot readily be analyzed, and yet can exhibit a huge range of behaviors depending on the exact values of the parameters.

There are various extensions of this single-cell use of conductance-based models. One is that it is common to model networks of neurons by connecting together simple (single-compartment) conductance-based models through model synapses. This allows simple network properties to be explored; more complicated networks are typically modeled using integrate-and-fire models. A second extension is to model important internal states, such as intracellular calcium levels, using the same compartmental structure. This is particularly important for things like calcium-sensitive potassium channels, which can lead to phenomena such as spike-rate adaptation. Although compartmental models capture the electrical geometry of single cells, they rarely capture the three-dimensional milieu in which the cells live.

Integrate-and-fire Models

Integrate-and-fire models lie at a level of abstraction above conductance-based models. Rather than using active channels to implement action potentials, they make the approximation of using a symbolic model of spike generation coupled with a leaky integrator model of a cell when its voltage is below the threshold for spike initiation. They also radically simplify the geometry of cells, eliminating the compartmentalization and including, at best, a stereotyped time course for synaptic input and for other time-dependent factors such as those allowing spike-rate adaptation.

Integrate-and-fire models (and the related spike-response model) are good for simulating large, recurrently connected, networks of neurons. Many mathematical issues about networks, such as the synchronization and desynchronization of spiking across the whole population, and the effects of different sorts and sources of noise, have been explored through using them. Furthermore, evidence suggests that precise time differences between presynaptic and postsynaptic activity control such things as synaptic plasticity. The integrate-and-fire model is the simplest form that still outputs spikes, and so can be used to address these issues. However, integrate-and-fire models pose substantial analytical difficulties themselves, and sometimes end up being an unhappy medium between more realistic, but analytically intractable, compartmental models, and highly abstract, but tractable, firing- rate models.

Firing-rate Models

The most abstract level of characterization of neurons abandons spiking altogether, and instead treats the output of cells as being continuous-valued, time-varying firing rates. This can be derived as an abstract approximation to integrate-and-fire neurons, under some assumptions about the time constants of processes inside cells. Networks of firing-rate models can be constructed, in which the influence of one cell on another is given by the product of the presynaptic cell's firing rate and the synaptic strength for the connection.

The main advantages of the firing-rate models are their empirical and analytical tractability. Firing-rate models usually involve a mild nonlinearity, turning an internal continuous variable like somatic voltage or current into a firing rate (which must be positive). Therefore, networks of neurons described using firing-rate models can be treated as coupled, nonlinear differential equations that can be shown to exhibit dynamic behaviors such as attraction to one of a set of fixed points, oscillations or chaos. In contrast to conductance-based models, these often evolve to relatively simple fixed-point or limit-cycle attractors. The regularities implied by attractor and oscillatory dynamic behaviors make them ideal as substrates on which to hang analyses of network computation.

Although it is possible to study the effects of synaptic plasticity in the context of conductance-based or integrate-and-fire models, by far the bulk of the work on computational analyses has been performed using firing-rate models. Here, with the exception of studies on recurrent attractor networks, a majority of the work has focused on non-recurrent, feed-forward network models, which are analytically much simpler to handle. These models take information represented in one way at one level of a processing hierarchy by the firing rates of a population of neurons, and transform and manipulate it to represent it in a different way by the firing rates of a population of neurons at a higher level of the hierarchy. Rules for plasticity have been seductive for computational modeling, since they offer an obvious way for large

networks of fairly simple processing units to come to perform computationally sophisticated tasks, apparently without requiring sophisticated programming.

Although the boundaries between them are somewhat blurred, there are three main classes of learning model. Unsupervised learning models act in a self-organizing manner, extracting statistical structure from input data. They are often used to capture activity-dependent adaptive processes that are assumed to be operational during development. Reinforcement learning models use evaluative information (rewards and punishments) in temporally complex and controllable environments (such as mazes) and specify how to predict future returns and choose actions in order to maximize these returns. Supervised learning models are rather special in that adaptation is based on information about both the inputs to the network and the desired outputs.

Supervised learning is extensively used outside the context of neural modeling. However, in two incarnations, it is important for neural modeling too. One case is when the output of one network is used to train another network. This requires an intricate dance by which neurons in the trained network must have two sets of input, one of which controls synaptic plasticity in the other. It has been suggested that this might take place under the control of neuromodulators. The other case for supervised learning is when the key question is whether a particular design of network of neurons is capable of executing a particular computational task or exhibiting some particular behavior, or whether the activity of some neurons in a network is consistent with their playing a role in the solution of a task. In this case, one can attempt to train the network (for instance by minimizing the discrepancy between the target and the actual outputs), using procedures that need have no relation to the rules governing neural adaptation.

CONCLUSION

Partnering conventional, multi-level, reductive models of experimental observations on neural mechanisms are models offering computational interpretations, couched at exactly the same multiple levels. These models indicate how the neural mechanisms implement computations, in the sense of representing information and performing algorithmic manipulations which are appropriate for solving a computationally well-specified task. The computational and conventional models should ideally mesh completely at all the levels. Few existing conventional and computational models actually achieve this degree of mutual coherence. Most current modeling is either conventional, at the compartmental and integrate-and-fire level, or computational, using sophisticated models of synaptic adaptation in the face of extremely simplified models of individual neurons, but rarely both conventional and computational.

There are suggestions that modeling should really proceed top-down, from the definition and computational decomposition of abstract tasks to the reductive neural implementation. However, such a strict policy is not productive as a strategy for analyzing neural systems, because it means throwing away constraints from the experimental data, and because it relies on adequate, multi-level accounts of the underlying computational tasks, which we currently lack. Instead, both conventional and computational modeling at multiple levels should progress together.

Further Reading

Churchland PS (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland PS and Sejnowski TJ (1992) *The Computational Brain*. Cambridge, MA: MIT Press.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

Foster CL (1992) *Algorithms, Abstraction and Implementation: Levels of Detail in Cognitive Science*. London, UK: Academic Press.

Haugeland J (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* **1**: 215–226.

Haugeland J (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Marr D (1982) *Vision*. New York, NY: Freeman.

Montague PR and Dayan P (1998) Neurobiological modeling: squeezing top down to meet bottom up. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 526–542. Oxford, UK: Basil Blackwell.

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Long-term Potentiation and Long-term Depression

Introductory article

Joe L Martinez Jr, University of Texas, San Antonio, Texas, USA
David Haixiang Peng, University of Texas, San Antonio, Texas, USA
William J Meilandt, University of Texas, San Antonio, Texas, USA
Edwin J Barea-Rodríguez, University of Texas, San Antonio, Texas, USA

CONTENTS

Long-term potentiation and depression are enduring changes in synaptic efficacy following brief high-frequency or low-frequency electrical stimulation of nerve fibers. Both are considered to be candidate mechanisms of learning and memory.

INTRODUCTION

It is widely believed that memories are stored in the brain as modifications of the synaptic strength between neurons. As new events are experienced, the synaptic connections between neurons increase and decrease in order to encode the memory within the specific network of neurons that are active during the event. When memories are recalled, it is believed that the same patterns of synaptic connections are reactivated in order to recreate the image or event being remembered. The leading experimental models used to study the cellular mechanisms underlying learning and memory include an activity-dependent increase in synaptic strength, known as long-term potentiation (LTP), and an activity-dependent decrease in synaptic strength, called long-term depression (LTD). Most studies continue to examine these two forms of plasticity in the hippocampus, a region of the brain critical for learning and memory processes. In addition, these forms of plasticity have been found in other structures of the brain such as neocortex, cerebellum, and peripheral nervous system. (*See* **Long-term Potentiation, Discovery of**)

The Hippocampus

The hippocampus derives its name from the Latin word for 'seahorse', because of its characteristic shape like the letter C (Figure 1). The hippocampus is composed of an evolutionarily older cortex that contains only three layers of cells, unlike the newer cortex (neocortex), which contains seven layers of cells. The hippocampus consists of the hippocampus proper (fields CA3, CA2, and CA1) and the dentate gyrus. The interconnections of neurons within the hippocampus make up what is known as the 'trisynaptic circuit'. Information enters the hippocampus through axons of the perforant path, which synapse onto dentate granule cells. The axons of dentate granule cells form a bundle of

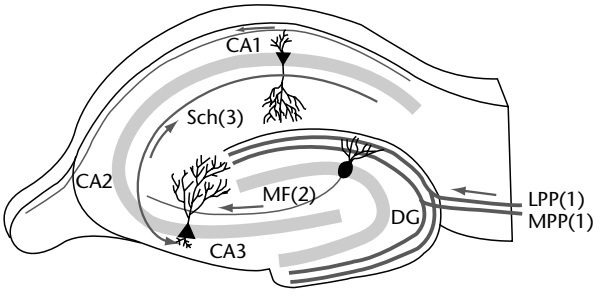


Figure 1. The trisynaptic circuit in a transverse section through the hippocampus of the rat. Shaded areas show the principal cell layers: granule cells of the dentate gyrus and the pyramidal cells of areas CA3, CA2, and CA1. There are three major excitatory pathways: 1, the perforant pathway (lateral/medial) from the entorhinal cortex to the dentate gyrus granule cells; 2, the mossy-fiber pathway that extends from the granule cells to the CA3 pyramidal cells; 3, the Schaffer collaterals which project from the CA3 pyramidal cells to the CA1 pyramidal cells. The arrows show the direction of information flow. DG, dentate gyrus; MF, mossy fibers; LPP, lateral perforant path; MPP, medial perforant path; Sch, Schaffer collaterals. Adapted from Bliss and Collingridge (1993).

fibers (the mossy fibers), and these synapse onto CA3 pyramidal cells. The CA3 pyramidal cells then send their axons to the pyramidal cells in CA1 via the Schaffer collateral pathway. Several forms of LTP and LTD have been identified in each of these pathways, some of which are discussed below.

LTP and LTD as Information Storage Devices

One of the major questions in neuroscience is whether LTP and LTD act as information storage devices. To answer this question, we must first explain how LTP works.

To date, most studies have focused on the properties of LTP in the Schaffer collaterals to CA1 (Figure 2). Electrodes can be placed within the cell body or dendritic layer of CA1 in order to record the basal synaptic activity of pyramidal neurons evoked by single electrical stimulations. After a stimulation, the axons of the CA3 neurons will release neurotransmitters (glutamate) into the synapse between the CA3 axon (presynaptic element) and the dendrites of the CA1 cells (postsynaptic element), causing the CA1 cell to respond by producing an excitatory postsynaptic potential (EPSP). By measuring the slope of the EPSP or the amplitude of the response, one can determine how the synaptic connections between these neurons change following electrical stimulation. Applying high-frequency stimulation (100 Hz) to the axons of the Schaffer collaterals causes a large, long-lasting increase in the slope and amplitude of the EPSP recorded in the CA1 neurons; this long-lasting increase in EPSP response is referred to as LTP. Alternatively, low-frequency stimulation (1–5 Hz) causes a long-lasting decrease in the EPSP response known as LTD (Figure 3).

How might this form of plasticity be used to store information? It is believed that information can be stored within the synaptic connections between neurons. If a group of neurons are activated following high-frequency stimulation causing the postsynaptic cells to persistently respond with greater intensity, then a 'memory' for that event has been stored at these synapses, allowing an enhanced response to the same synaptic input. Similarly, LTD could be induced in the surrounding fibers to reduce the synaptic activity, thereby increasing the signal-to-noise ratio, or LTD could be induced in the same fibers to return formerly potentiated responses back to baseline to perhaps erase a memory. Together, these two forms of plasticity provide an attractive model to explain how memories are stored in the brain.

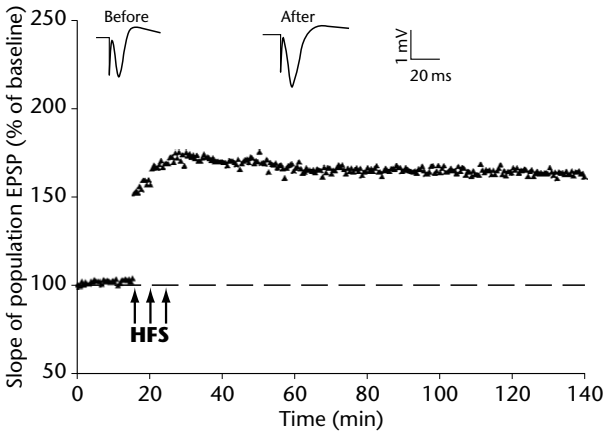


Figure 2. Long-term potentiation (LTP) in the Schaffer collateral CA1 pathway of rat hippocampus *in vivo*. The slope of a population excitatory postsynaptic potential (EPSP), i.e. the initial slope of the larger and slower deflection (see the traces in the insets), is plotted versus time. The population EPSPs are evoked by single shocks in presynaptic Schaffer collaterals (axons) and recorded extracellularly in the postsynaptic CA1 dendritic area at 15 s intervals. Delivery of three trains of high-frequency stimulation (HFS, 1 s of 100 Hz) separated by 5 min induces LTP lasting for hours. Traces in the insets (before/after tetani) show the change of population EPSP (i.e. the postsynaptic response).

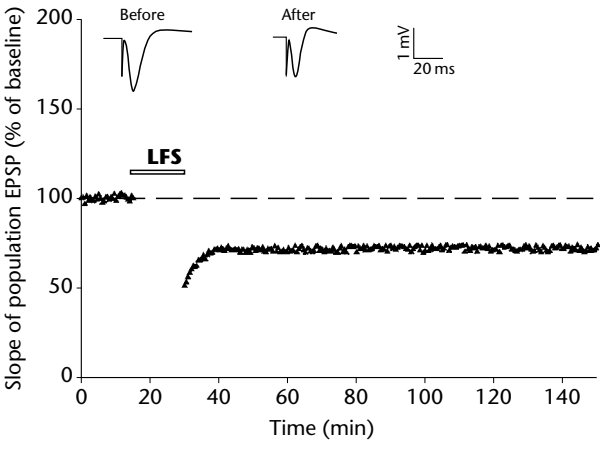


Figure 3. Homosynaptic long-term depression (LTD) in the Schaffer–CA1 pathway in rat hippocampal slices. The extracellular population excitatory postsynaptic potential (EPSP) recorded in CA1 dendrites is plotted against time. The LTD is induced by low-frequency stimulation (LFS, 900 pulses at 1 Hz). Traces in the insets (before/after tetani) show the change of population EPSP. Adapted from Lee *et al*. (2000) *Nature* **405**: 955–959.

DISCOVERY OF LTP

Although it is generally agreed that LTP was discovered by Bliss and Lomo in 1973, the story dates back to the 1950s, when the precisely organized afferent systems in hippocampal and associated structures were revealed by scientists in Oxford and Oslo. At that time, just a few groups of neurophysiologists, such as Per Andersen in Oslo, John Green and Ross Adey in Los Angeles, and Eric Kandel and Alden Spencer in Karl Frank's laboratory at the National Institutes of Health, were interested in taking advantage of this unique hippocampal arrangement to study central synaptic transmission and epileptic phenomena.

In the early 1960s, Kandel and Andersen intracellularly recorded the synaptic transmission to the pyramidal cells in the hippocampus. At that time, these researchers generally knew that repetitive electrical stimulation with moderate strength and duration could increase signals of neural cell discharges, and this proved useful in their experiments. When the signals were fading, they could be easily recovered by turning the stimulus rate to a higher frequency for a few seconds. When Terje Lomo joined Andersen's laboratory, this phenomenon of frequency potentiation elicited his interest. In 1966, he reported his pioneering observations that repetitive stimulation (6–50 Hz) of the perforant path fibers induced a large potentiated response. This response was seen as an increase in the amplitude and a decrease in the latency of the population spike (population action potentials, i.e. synchronous firing of action potentials by a group of neural cells) lasting for hours, depending upon the stimulation frequency.

Later, Timothy Bliss joined Lomo in Oslo, and started to systematically examine this phenomenon of frequency potentiation in the hippocampus. Before long, their collaborative effort led to the well-known paper published in the *Journal of Physiology* in 1973. This was the first time that long-term potentiation was quantitatively described. In anesthetized rabbits, Bliss and Lomo measured the amplitude of the population EPSP (i.e. summation of postsynaptic potentials) and the amplitude and latency of the population action potential extracellularly, following stimulation of the perforant path fibers to the dentate area with single shocks. In general, the levels of the population EPSP and action potential remained stable over time; but when a short train of shocks was applied at a high frequency (e.g. 15 Hz for 15 s) to the synapses, both measurements rose sharply and remained elevated for hours afterwards. In a subsequent paper, Bliss and Gardner-Medwin repeated these experiments on unanesthetized rabbits, and the effect of potentiation lasted for months. However, in most of the original publications, the term 'long-lasting potentiation' was used until the term 'long-term potentiation' was introduced by Douglas and Goddard in 1975. Since then, the acronym LTP has been generally adopted because it is easier to pronounce than alternatives such as LLP (long-lasting potentiation), LTE (long-term enhancement) or LTSE (long-term synaptic enhancement). In 1977, the first intracellular measurements of LTP in transverse hippocampal slices were made by Andersen and co-workers. They observed both an enhanced EPSP and an increased probability of firing action potential, verifying Bliss and Lomo's original findings through extracellular measurements.

Not unexpectedly, LTP has attracted great attention because it has been and still may be the best candidate mechanism for the information storage in the brain. Much of the subsequent work on LTP has focused on exploring its properties, the conditions required for its induction, mechanisms of its induction and maintenance, and its correlations with learning and memory. In addition to the high-frequency electrical stimulation used by Bliss and Lomo, many other stimulus patterns have been found to efficiently induce LTP. These include theta burst stimulation (TBS) – several short, high-frequency (e.g. 100 Hz) bursts of a few pulses delivered at an interburst interval of 200 ms, which mimics the endogenous hippocampal theta rhythm in rats, and primed-burst stimulation – a single priming stimulus followed 200 ms later by a burst of four shocks at 100 Hz. Subsequently, different forms of LTP have been observed in a variety of brain structures and species as described below.

LONG-TERM POTENTIATION

Properties of LTP: The Hebbian Synapse

Since its discovery, LTP has attracted the attention of numerous scientists because it has properties that are highly suggestive of an information storage device. In addition to its long-lasting nature (several hours in an anesthetized animal, and days or even weeks in an awake, freely moving animal), LTP has three distinguishing properties: cooperativity, associativity, and input specificity. Cooperativity occurs following stimulation of a sufficient number of afferent fibers to collectively surpass a

stimulus threshold necessary to induce LTP. Subthreshold stimulation or stimulation of too few fibers will not, on its own, induce LTP. One of the more intriguing properties of LTP, known as associativity, led some researchers to suggest that it is a memory mechanism. For instance, if a weak (subthreshold) stimulation in one afferent is simultaneously paired with a strong (suprathreshold) stimulation in a separate afferent to the same cell population, then the weakly stimulated afferent also exhibits LTP. This may be relevant to associative or relational features of learning and memory, because associative induction implies the capacity to relate two arbitrary patterns of pre- and postsynaptic neural excitation. Finally, LTP is input-specific, because only the tetanized afferents show potentiation, whereas other inputs to the same neurons that are not active at the time of the tetanus do not exhibit LTP. A synapse-specific mechanism might endow greater storage capacity than would changes in cell excitability.

As a physiological model, LTP more closely resembles the plastic synapse that Hebb postulated for associative learning than anything else presently available. In his now-classic 1949 book, *The Organization of Behavior*, Donald Hebb proposed:

> When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process of metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

In other words, the process of learning involves intense or repeated activity at particular synapses, which become stronger as a result of the activity of the pre- and postsynaptic cells. This associative strengthening of synapses that are active during the learning process in turn forms the 'memory' of what has been learned. The synapses that are modified in this manner are referred to as 'Hebbian synapses'. Hebb's postulate is very close to the definition of LTP.

The Hebbian concept that memory depends on the modification of neuronal synapses, or synaptic plasticity, has led scientists interested in the cellular basis of memory to focus their research on elucidating the processes of LTP and the molecular mechanisms underlying these processes.

Mechanisms of LTP

Two major forms of LTP have been identified in the hippocampus and a variety of other tissues (see below). One form requires the activation of *N*-methyl-D-aspartate (NMDA) receptors, known as NMDA receptor-dependent LTP, while the other, NMDA receptor-independent LTP, does not require activation of such receptors. (*See* **Neurotransmitters**)

NMDA receptor-dependent LTP

In most regions of the hippocampus, the induction properties of LTP can largely be attributed to a specialized receptor called the NMDA receptor. These receptors are a unique subtype of voltage-dependent glutamate receptors. For the induction of LTP, the NMDA receptor must be activated by the neurotransmitter glutamate and simultaneously there must be sufficient depolarization of the postsynaptic membrane to relieve a magnesium ion block in the NMDA receptor-associated ion channel. Release of this Mg^{2+} block subsequently allows the entry of calcium ions into the postsynaptic terminal. The calcium ion influx activates any number of Ca^{2+}-sensitive second-messenger cascades, including various protein kinases and phospholipases. Because optimal activation of NMDA receptor channels requires both presynaptic transmitter release and postsynaptic depolarization, the NMDA receptors act as Hebbian coincidence detectors. Thus, activated NMDA receptors at synapses that are close to active sites of depolarization may be depolarized sufficiently to relieve the magnesium ion block and initiate the cascade of events that leads to LTP induction. This cascade may occur even though the activity of that particular synapse alone was not sufficient to induce LTP. Thus, NMDA receptors can account for the association of two separate afferent projections to the same cell, one strongly and the other weakly active, and for the cooperative requirement that a threshold number of fibers be active.

The maintenance of NMDA receptor-dependent LTP is less well understood. It is generally agreed that LTP has distinct temporal phases. The short-term early phase of LTP lasting about 1 h requires only covalent modification of preexisting proteins, such as protein kinase activation and protein phosphorylation. Long-term potentiation can be further maintained and converted to the more stabilized form, late-phase LTP, when accompanied by new protein synthesis from existing messenger ribonucleic acid (mRNA) molecules and gene expression. There are at least three possible explanations for the cascades of events that lead to the expression and maintenance of NMDA receptor-dependent LTP: an increase of presynaptic glutamate release; an increase in the number of postsynaptic glutamate receptors; and morphological changes that favor efficacy of synaptic transmission.

NMDA receptor-independent LTP

In 1986, Harris and Cotman discovered that unlike LTP in CA1, the induction of mossy-fiber LTP was not dependent on the activation of NMDA receptors, and could be reliably induced even in the presence of NMDA receptor antagonists, suggesting that this pathway uses different mechanisms for the induction and expression of LTP. This form of LTP is also be found in the lateral perforant path to the dentate gyrus and CA3 (see Figure 1).

In accordance with the finding that LTP in the mossy-fiber and lateral perforant pathways does not require the activation of NMDA receptors, immunohistochemical studies demonstrate that the synaptic terminals of these pathways contain sparse numbers of NMDA receptors and large numbers of opioid receptors. In an attempt to identify the mechanisms for mossy-fiber LTP induction, it was found that the mossy-fiber and lateral perforant pathways contain and release opioid peptides. Studies have shown that the induction of mossy-fiber LTP is blocked by the opioid receptor antagonist naloxone. At least two mechanisms of opioid receptor-dependent LTP induction exist in the hippocampus. Induction in the mossy fiber to CA3 and lateral perforant to CA3 pathways depends on the activation of μ, but not δ, opioid receptors, whereas induction in the lateral perforant to dentate pathway depends on both μ and δ opioid receptors.

The cellular and molecular mechanisms required for the induction and expression of mossy-fiber LTP and the site of induction, either presynaptically or postsynaptically, remain a heavily debated issue. There is evidence to suggest that mossy-fiber LTP induction is dependent on brief high-frequency stimulation (with a minimum of 30 pulses at 100 Hz) sufficient to (a) depolarize the postsynaptic cell and activate voltage-gated Ca^{2+} channels, and (b) release opioid peptides and activate μ opioid receptors. This leads to an increase in intracellular Ca^{2+} levels in addition to the activation of group I metabotropic glutamate receptors, which results in the release of calcium ions from internal stores. This, in turn, acts through a signal transduction pathway mediated by cyclic adenosine monophosphate (cAMP) to activate protein kinase A and induce new protein synthesis. The requirement for cAMP-mediated signal transduction has also been found to be involved in the presynaptic mechanisms associated with mossy-fiber LTP induction and expression. Additionally, endogenously released opioids may induce mossy-fiber and lateral perforant path LTP indirectly by reducing inhibition, or disinhibiting the synapses, allowing LTP to occur. Taken together, these data suggest that activation of μ opioid receptors plays an important part in the induction of mossy-fiber and lateral perforant path LTP in area CA3.

Like the NMDA receptor-dependent LTP, the induction of mossy-fiber LTP in hippocampal slices (*in vitro*) also has a protein-independent early phase and a protein-dependent late phase. However, Barea-Rodríguez and co-workers have provided evidence suggesting that protein synthesis is necessary for mossy-fiber LTP induction in anesthetized rats (*in vivo*). This may be due to the significant differences in time course between these two forms of LTP: in contrast to its rapid onset *in vitro*, mossy-fiber LTP *in vivo* has a very slow onset and takes about 1 h to reach its maximum augmentation.

LTP at Various Brain Sites

Most of our knowledge about the properties and potential mechanisms of LTP comes from studies in the rodent and rabbit hippocampal formation. This is due not only to the discovery of LTP in the hippocampus, but also to the distribution of LTP which, in various forms, is evident at the three principal synaptic connections in hippocampus (see Figure 1): in the dentate gyrus granule cells by stimulation of the perforant pathway as originally described by Bliss and Lomo, in the CA3 pyramidal cells by stimulation of the mossy fibers, and in the CA1 pyramidal cells by stimulation of the Schaffer collateral branches of the CA3 neurons.

Long-term potentiation has also been well examined in the intrinsic connections of the neocortex such as the somatosensory, visual, and motor neocortex, which function as a structure for long-term storage. Associative Hebbian LTP can be reliably induced in the superficial layers of the neocortex, using high-frequency stimulation at the middle layers of cortex in anesthetized animals or in slice preparations. It is input-specific and dependent on activation of postsynaptic NMDA receptors. In contrast, for the induction of neocortical LTP in adult, freely moving rats, the stimulation trains must be spaced and repeated over a period of days. Tetanic stimulation of the white matter, in sharp contrast, consistently failed to elicit LTP in the superficial layer of cortex unless a γ-aminobutyric acid (GABA) type A receptor antagonist was applied, or the white matter and the middle layer simultaneously received tetanic stimulation.

It is well known that the entorhinal cortex sends much of its output both to the hippocampus and back to the neocortex. Although the entorhinal

cortex receives inputs from various cortical sites, its major inputs are the projections from the perirhinal cortex and the parahippocampal cortex, which receive their inputs from sensory and association cortices. Long-term potentiation is found in all pathways between the hippocampus and neocortex, but the induction and the decay are slower in the neocortex than in the hippocampus and associated structures, which may reflect some aspects of LTP in memory processing.

With few exceptions, such as the lateral olfactory tract, LTP can be produced throughout the limbic system (amygdala, entorhinal cortex, dentate gyrus, subiculum, septum, and hippocampus), although the LTP is relatively larger in the hippocampal pathways. Because of its good correlation with emotional learning and memory, LTP in the amygdala is also extensively studied as a laboratory model. The first examination of LTP in the amygdala was performed in awake rats by stimulation in the piriform cortex. More recently, LTP has been reported in both thalamoamygdala and hippocampoamygdala projections *in vivo*. There are two forms of LTP in the amygdala: the NMDA receptor-dependent LTP in the basolateral nucleus, and a form of LTP dependent on α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) receptors in the interneurons.

In addition to the brain structures described above, LTP can also be induced in the deep nuclei of cerebellum by stimulation of the inferior peduncle fibers, the ventral horn of spinal cord by stimulation of the dorsal horn–intermediate nucleus region, and the superior cervical ganglion by stimulation of the preganglionic nerves. Furthermore, LTP is not only limited to the mammalian brain, such as rabbits, rats, monkeys, and humans, but has been described in other vertebrates as well, including goldfish, bullfrogs, birds, and lizards. It has also been found in some invertebrates, suggesting that phenomena fitting the general description of LTP occur ubiquitously throughout the nervous system.

LONG-TERM DEPRESSION

Long-term depression is characterized by a depression of evoked synaptic responses. It was first described in the hippocampus and was induced by persistent low-frequency stimulation (typically 1–5 Hz for 5–15 min). Studies have now shown that 900 pulses of 1 Hz stimulation can reliably induce LTD in the Schaffer collaterals to CA1 (Figure 3). This form of LTD is induced only in the specific fibers that receive stimulation, and hence is termed 'homosynaptic' LTD. This process is interesting because it suggests that it could serve as a mechanism for forgetting, following studies demonstrating that LTD induced in previously potentiated synapses would return synaptic responses back to baseline levels. It could also serve to enhance recent representations by depressing synapses that are less active, and therefore sharpen sparsely encoded memories. The finding that LTD requires postsynaptic factors, including NMDA receptor activation, suggests a more direct role for LTD in information storage.

Long-term depression is also found in the cerebellum. Evidence indicates that the cerebellum is important in eyeblink classical conditioning. Indeed, the memory trace of this form of learning is localized in the cerebellum. Although LTP may be induced in the cerebellum, the predominant form of plasticity observed in this structure is LTD. Studies have indicated that LTD may be a neural substrate for the conditioned eyeblink response in the cerebellum. One research group found that mice lacking one type of glutaminergic receptors did not exhibit LTD and were impaired in learning a conditioned eyeblink response.

Mechanisms of LTD

Studies investigating the mechanisms of LTD induction in the hippocampus focus on homosynaptic LTD in the CA1 region. Interestingly, the mechanisms of LTD induction are similar to those underlying homosynaptic LTP in that both postsynaptic calcium ion influx and NMDA receptor activation are necessary. Although both calcium ions and NMDA receptors are necessary for LTD induction in CA1, it seems that the intracellular Ca^{2+} concentration determines whether LTP or LTD is induced. According to some studies, a small influx of calcium ions, which produces a low intracellular Ca^{2+} concentration, may activate selective protein phosphatases (calcineurin) whose actions lead to the induction of LTD. Selective inhibitors of protein phosphatases block the induction of LTD in CA1. Thus, LTD induction, at least in CA1, may require the activation of protein phosphatases.

Researchers suggest that LTD occurs when conditions approach, but do not meet, the requirements for LTP induction. This may serve as a mechanism of contrast enhancement, and ensure that a minimum number of synapses contribute to a given representation.

CONCLUSION

Although their relationship with learning and memory remains controversial, we believe that LTP and LTD are the best candidates for a cellular process of synaptic change that underlies learning and memory in the vertebrate brain. The absence of proof that LTP and LTD are involved in memory results from the current uncertainties about what a memory is and how we should observe it. Are the different forms of LTP and LTD associated with the different forms of memory known to exist? The fact that there are multiple forms of LTP and LTD, together with the distributed nature of information storage, makes it difficult to identify the processes necessary for memory known to implicate specific cellular processes in behavioral measures of memory. The same may be true for cerebellum-related learning.

Further Reading

Bear MF and Malenka RC (1994) Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology* **4**: 389–399.

Bennett MR (2000) The concept of long term potentiation of transmission at synapses. *Progress in Neurobiology* **60**: 109–137.

Bliss TVP and Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**: 31–39.

Christie BR, Kerr DS and Abraham WC (1994) Flip side of synaptic plasticity: long-term depression mechanisms in the hippocampus. *Hippocampus* **4**: 127–135.

Johnston D, Williams S, Jaffe D and Gray R (1992) NMDA-receptor-independent long-term potentiation. *Annual Review of Physiology* **54**: 489–505.

Landfield PW and Deadwyler SA (1988) *Long-term Potentiation From Biophysics to Behaviour*. New York, NY: Alan R Liss.

Linden DJ (1994) Long-term synaptic depression in the mammalian brain. *Neuron* **12**: 457–472.

Martinez JL and Derrick BE (1996) Long-term potentiation and learning. *Annual Review of Psychology* **47**: 173–203.

Long-term Potentiation, Discovery of

Introductory article

Preston E Garraghty, Indiana University, Bloomington, Indiana, USA

CONTENTS

Long-term potentiation is an increase in the strength of connections between neurons due to changes in the inputs to a neural circuit. It is widely regarded as the best model for studying the ways the brain changes as we learn and remember.

INTRODUCTION

Long-term potentiation (LTP) is a persistent change in the strength of connections between neurons. It has been studied intensively and extensively, as it represents the best current model of how memories might be stored in the brain. More specifically, it serves as a potential means by which associations might be formed and maintained.

The first formal explication of learning through association can be attributed to the seventeenth-century British philosopher, Thomas Hobbes. Hobbes was the first proponent of a philosophy of mind based on the assumption that nothing exists aside from matter and energy, and consequently that all human behavior can be understood materialistically – that is, in terms of physical processes, particularly in the brain. Hobbes' principle of association stated that ideas come to be thought together if their corresponding sensations occurred together initially. These ideas included ones of reward and punishment, and thus, how the anticipated consequences of actions could influence their probabilities. Hobbes' ideas proved to be profoundly influential, helping to stimulate the development of a philosophy of mind that came to be known as British empiricism. According to the British empiricists, the mind of a newborn is a *tabula rasa* (blank slate) on which is written the associations between elementary ideas that arise from sensory experience. These associations permit thought, and the construction of consciousness.

This notion of learning by association proved to be a powerful idea in the emergence of experimental psychology, but it was not until the serendipitous discovery of classical conditioning by the Russian physiologist, Ivan Pavlov, that an experimental paradigm became available to rigorously evaluate learning by association in a scientific context.

Pavlov demonstrated that after a formerly neutral stimulus (a tone) has been systematically paired with a stimulus (food powder) that naturally elicits a reflexive response (in this case, salivation), the tone comes to 'stand for' the food powder, and elicit a salivation response. In this particular instance, Pavlov termed the food powder the 'unconditional stimulus' (US) and the salivation it elicited the 'unconditional response' (UR), as the former elicited the latter unconditionally. The tone and the salivation response it came to elicit were termed the 'conditional stimulus' (CS) and 'conditional response' (CR) respectively, as the former came to elicit the latter only because of the experimentally imposed conditional relationship between the CS and the US. Pavlov believed that the process of conditioning reflected a strengthening of the connection between a brain region representing the CS and a brain region representing the US. Thus, in his view, as conditioning proceeds, increasing levels of CS-evoked neural excitation spread to the US brain area, and evoke the response. In the language of the British empiricists, the 'idea' of the CS (evoked by its presentation) becomes associated with the 'idea' of the US because the conditioning procedures ensured that the two ideas would be temporally contiguous. Learning by association became an assumption, yet the neural mechanisms were completely unknown.

Memories reflect changes in an organism arising from experience. The minimal requirement for a physiological demonstration of learning is simply to demonstrate that existing synaptic connections can be strengthened as a function of experience.

The first demonstration of learning-related neural plasticity was made quite by chance. In 1935, Durup and Fessard were recording electroencephalographic (EEG) activity in a human subject, as they investigated how a particular EEG rhythm (alpha) ceases when the subject is in an illuminated versus a darkened environment. Thus, they simply switched the lights on and off, and monitored the resulting changes in the recordings. This proceeded for a number of transitions, but on one trial, the switch failed to illuminate the room as expected because the bulb failed. Even though the room remained darkened, the alpha rhythm ceased as if the room had been illuminated. Additional experiments revealed that the sound of the light switch had become a CS because it had been reliably paired with illumination (the US), and thus, came to elicit a CR (alpha reduction). This clearly represented a learning-related change in brain activity, but it reflected changes at literally millions upon million of synapses. Certainly, the acquisition and storage of some bit of information, say the abbreviation of 'long-term potentiation', has affected far fewer of your synapses.

Scientists continued to search for evidence of changes in synaptic strength at individual synapses, and, in 1947, Larrabee and Bronk provided such evidence at a far more 'local' level. They recorded the summed activity of neurons in the stellate ganglion of cats while stimulating preganglionic axons. Any change in the amplitude of the postganglionic response would reflect the effects of the stimulation. They found that a brief period of repetitive (tetanic) stimulation was followed by an increase in the magnitude (potentiation) of the postganglionic response that persisted for about three minutes (post-tetanic potentiation). Clearly, three minutes of strengthened synaptic transmission cannot account for the persistence of memory, but these results clearly demonstrated changes in the strength of connections as a function of experience.

DISCOVERY OF LTP

In 1973, Bliss and Lomo reported that repeated, intense stimulation of projections to and within the hippocampus (the perforant pathway, the mossy fibers or the Schaffer collaterals; Figure 1) resulted in enhanced neuronal responsiveness of postsynaptic hippocampal neurons. This increased synaptic strength persisted for weeks, making this the first demonstration of sustainable changes in synaptic connectivity in the mature brain. Donald Hebb had predicted in his 1949 book, *The Organization of Behavior*, that such sustainable changes in the strength of neural connections must be possible if one is to account for such behavioral phenomena as learning and memory. Many behavioral neuroscientists and computational modelers now refer to changes in synaptic strength (or weight) as being due to 'Hebbian-like' processes.

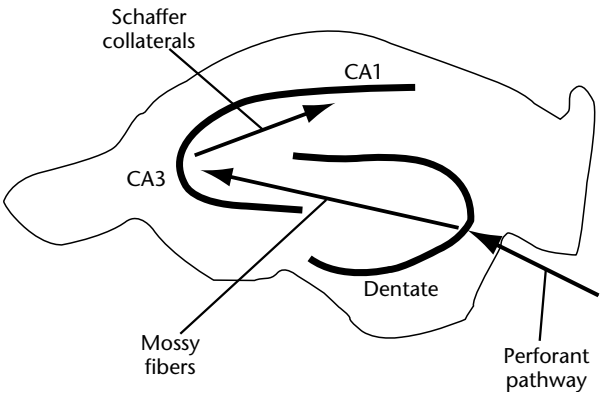


Figure 1. A cross-section through the hippocampal formation. The perforant pathway represents the primary extrinsic input into the hippocampus. Fibers of this pathway synapse with neurons in the dentate. Dentate neurons project as mossy fibers onto neurons in the CA3 region of the hippocampus; CA3 neurons synapse with CA1 neurons via Schaffer collaterals. Tetanic stimulation of any of these fiber tracts can produce long-term potentiation in the target neurons.

Since the classic Bliss and Lomo study, LTP has been observed in many other brain areas, such as the cerebellum and cortex, and has been fairly well accepted as an exemplar of a biological model of learning and memory.

PUTATIVE MECHANISMS OF LTP

It must be noted at this juncture that there are multiple forms of what can be called LTP. Here we will look at a form of hippocampal LTP that is dependent upon a particular subset of glutamatergic receptors (glutamate is the principal excitatory neurotransmitter in the brain), but even in mammalian hippocampus at least one other form of LTP exists, and it is dependent on opioid peptides and not glutamate. (*See* **Neurotransmitters**)

To address issues related to the mechanisms of LTP, at least a rudimentary knowledge of hippocampal anatomy is necessary. Information coming into the hippocampus arrives in the dentate via the perforant pathway. From there, mossy fibers project to area CA3 (CA is derived from *cornu ammonis*, another name for the hippocampus), which in turn projects to area CA1 via Schaffer collateral fibers

(Figure 1). Feedforward inhibitory modulation is present in the dentate and CA1 layer, utilizing the inhibitory neurotransmitter GABA (γ-aminobutyric acid). Importantly, these inhibitory loops are activated only during repeated bursts of activity (i.e. tetanus: a level that is significantly higher than 'normal' synaptic transmission).

The development of LTP requires not only a sufficient level of presynaptic activity, but also temporally correlated activation of postsynaptic neurons. Stimulation of the presynaptic circuits (inputs) can clearly account for the first requisite, and a reduction in postsynaptic inhibition has been proposed to 'prime' the circuit for synaptic potentiation. That is, hippocampal feedforward inhibitory circuits are active only when the stimulation intensity is sufficient to evoke bursting of projection neurons (as opposed to tonic firing levels). Support for the notion that changes in inhibitory influences play a part in the development of LTP has come from studies showing that the application of compounds that reduce inhibition facilitates the induction of LTP, while increasing inhibition attenuates LTP.

Glutamate is the primary neurotransmitter used for propagation of information through the hippocampus. As with all known neurotransmitters, glutamate is a ligand for several classes of postsynaptic receptors, including NMDA and AMPA receptors – so called because exogenous compounds with those acronyms (NMDA, *N*-methyl-D-aspartate; AMPA, α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid) bind selectively to those subsets of glutamate receptors. These receptor subtypes differ in two important functional respects. First, unlike the classic excitatory postsynaptic receptor that opens sodium channels when bound with a neurotransmitter molecule, the NMDA receptor gates calcium channels. Second, this gating of calcium ions occurs only when the ligand is present and the postsynaptic membrane is relatively depolarized with respect to its resting potential. Thus, AMPA receptors are ligand-gated while NMDA receptors are ligand- and voltage-gated. Therefore, activation of AMPA receptors simply requires presynaptic activity, whereas activation of NMDA receptors requires presynaptic activity and residual depolarization due to recent excitatory and/or disinhibitory inputs. In sum, AMPA receptors are used for routine cellular communication, whereas the NMDA receptor system appears to be recruited when inputs are out of the ordinary (e.g. tetanic stimulation).

At least one form of LTP is dependent upon the NMDA subset of glutamate receptors for its initiation. That is, application of compounds that block NMDA receptors attenuate the induction of LTP in the CA1 region of the hippocampus, while compounds that activate NMDA receptors accentuate the response. After LTP is induced, however, blockade of NMDA receptors fails to prevent its continued expression. This finding suggests that other mechanisms are responsible for the maintenance of the increased synaptic strength once it has been established, and it has been reported that the number of postsynaptic AMPA receptors was elevated in the hippocampus of rats after LTP had been established. This increase in AMPA receptor number would clearly result in larger depolarizations arising from excitatory synaptic inputs. Thus, for this form of long-term potentiation, NMDA receptors are necessary for its induction while AMPA receptors are necessary for its maintenance.

ATTEMPTS TO RELATE LTP TO LEARNING AND MEMORY

There are a number of reasons why LTP has been so fervently embraced as a model of learning and memory.

1. LTP can be induced by conditioning stimuli within the physiological range; that is, the frequencies of stimulation used are not unlikely to occur in the intact nervous system.
2. Unlike post-tetanic potentiation, LTP can last for weeks.
3. It is most prominent in brain regions that are strongly implicated in learning and memory (for example, the hippocampus and neocortex).
4. It does not occur in the absence of tetanic stimulation, and, presumably, such increased firing rates in the intact system would reflect the processing of inputs having increased relevance to the organism (Figure 2).
5. Associativity can occur whereby a weak input can be strengthened if its stimulation is paired with tetanic stimulation of strong inputs (Figure 2).
6. Long-term potentiation requires the Hebbian property of correlated presynaphic and postsynaptic activity.
7. Pharmacological interventions that block LTP induction have also been shown to impair some forms of learning. For example, NMDA receptor antagonists that block the induction of LTP also interfere with spatial learning in rats. Blocking NMDA receptors impairs the learning of a Pavlovian discriminated eyeblink reversal in rabbits and the acquisition of a tone-signaled avoidance response after appetitive instrumental training in rats.
8. Mutant mice in whom hippocampal LTP is severely attenuated show learning deficits.
9. The blockade of NMDA receptors interrupts neural plasticity in animal models in studies of injury-induced brain plasticity, and it seems reasonable to

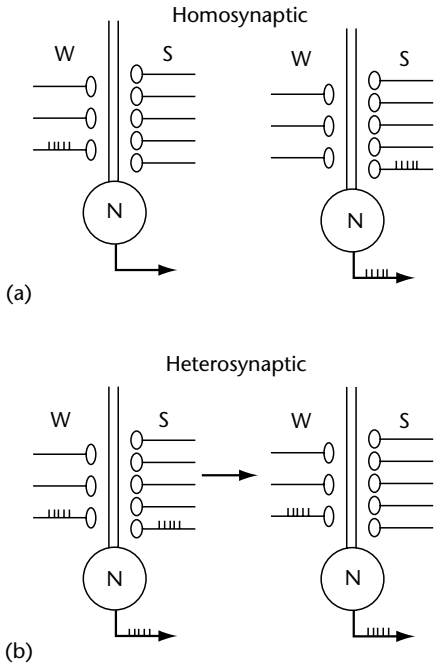


Figure 2. (a) Homosynaptic facilitation reflects the strengthening of a set of connections due to appropriate patterns of stimulation of the presynaptic fibers. On the left, the 'W' reflects either weak stimulation or the stimulation of a set of inputs with weak connections (the series of vertical lines represents stimulation of this pathway). There is no change in the output of the neuron receiving the synapses (N). On the right, the 'S' input represents a strong tetanic stimulation, and the increase in the responsiveness of the postsynaptic neuron is represented by the series of vertical lines on its axon. The association here relates to the increased correlation between pre- and postsynaptic activity. (b) Heterosynaptic facilitation reflects the strengthening of a weak set of inputs because of a correlation with a strong set of inputs. This is reflected on the right where the weak inputs have gained strength because of their correlation with the strong inputs (left).

assume tentatively that altering inputs using nerve injuries is akin to altering inputs with tetanic stimulation or the demands of some learning tasks.

CONCLUSION

The reasons that NMDA-dependent LTP has been suggested as a model of learning and memory are compelling. It must be recognized, however, that they represent a series of inferences. Using the language of jurisprudence, the above-listed set of observations represent circumstantial evidence, not an eyewitness account. Thus, it is prudent to view NMDA-dependent LTP as an interesting model of synaptic plasticity that may represent the neural substrate for some manifestations of learning and memory. It is also important not to lose sight of the fact that several forms of synaptic facilitation have been demonstrated that are not dependent on NMDA glutamatergic receptors. Thus, the brain has at its disposal a number of mechanisms by which synapses can be strengthened; no one of them is likely to emerge as the sole substrate of learning and memory. Finally, the emphasis throughout this brief review has been on how synaptic couplings can be strengthened. It is of crucial importance to recognize that reductions in synaptic efficacy are likely to be as important for learning and memory in particular, and neural flexibility in general, as LTP.

Further Reading

Bashir ZI, Berretta N, Bortolotto ZA *et al*. (1994) NMDA receptors and long-term potentiation in the hippocampus. In: Collingridge GL and Watkins JC (eds) *The NMDA Receptor*, 2nd edn, pp. 294–312. New York, NY: Oxford University Press.

Bliss TVP and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–356.

Brown TH, Kairiss EW and Keenan CL (1990) Hebbian synapses: biophysical mechanisms and algorithms. *Annual Review of Neuroscience* **13**: 475–511.

Buonomano DV and Merzenich MM (1998) Cortical plasticity: from synapses to maps. *Annual Review of Neuroscience* **21**: 149–186.

Collingridge GL and Bliss TVP (1987) NMDA receptors – their role in long-term potentiation. *Trends in Neuroscience* **10**: 288–293.

Martinez JL and Derrick BE (1996) Long-term potentiation and learning. *Annual Review of Psychology* **47**: 173–203.

Morris RGM and Davis M (1994) The role of NMDA receptors in learning and memory. In: Collingridge GL and Watkins JC (eds) *The NMDA Receptor*, 2nd edn, pp. 340–375. New York, NY: Oxford University Press.

Myers WA, Churchill JD, Muja N and Garraghty PE (2000) Role of NMDA receptors in adult primate cortical somatosensory plasticity. *Journal of Comparative Neurology* **418**: 373–382.

Rosenzweig MR, Leiman AL and Breedlove SM (1999) *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience*, 2nd edn, pp. 504–510. Sunderland, MA: Sinauer.

McCulloch–Pitts Neurons

Intermediate article

James A Anderson, Brown University, Providence, Rhode Island, USA

CONTENTS

The 'McCulloch–Pitts neuron' is an abstraction of the computational functions of a biological neuron. This model 'neuron' has two states, on or off, it sums activation from other neurons, and it enters the on-state when the sum of its inputs exceeds a threshold. Networks composed of such abstract "neurons" can compute any finite logical expression.

ASSUMPTIONS

In 1943, Warren S. McCulloch and Walter H. Pitts wrote a paper for the *Bulletin of Mathematical Biophysics* entitled 'A logical calculus of the ideas immanent in nervous activity'. In this article, generally considered to be the first and most influential paper on neural network modeling or computational neuroscience, McCulloch and Pitts boldly proposed a model both of the neuron and of what groups of such neurons could compute.

The paper proposed several assumptions that formally defined what has since been known as the *McCulloch–Pitts neuron*. First, 'the activity of the neuron is an "all-or-none" process' – that is, it is a binary unit that can exist in only two states. These states can be variously interpreted as on or off, zero or one, active or inactive or, most portentously of all, true or false. Secondly, the neuron can be connected to other neurons and to input lines. For a neuron to become active, a certain number of input connections ('synapses') had to be excited within a specific time period. This assumption as usually implemented contained three parts. (1) All synapses have the same strength. The connections, as well as the neurons, are 'all or none'. (2) Time in the network is quantized in units of time based on the integration time needed to sum the active synapses. (3) Each neuron has a threshold – that is, the number of simultaneously active synapses required to put the neuron in the active state.

There are several less important assumptions. First, the only significant time delay in the nervous system is the delay that even then was known to exist at synapses. Secondly, there is no learning – that is, connections and neuron thresholds do not change with use. Thirdly, inhibition is absolute – that is, any active inhibitory synapse forces the neuron into the inactive state. The paper referred to above itself shows that this form of inhibition is not necessary, and that subtractive inhibition gives the same results.

Figure 1 shows a simple McCulloch–Pitts neuron. The original notation used in the 1943 paper is opaque and is now rarely used. The notation used here is that employed by Minsky in his classic book, *Computation: Finite and Infinite Machines* (Minsky, 1967), which discusses the properties and uses of McCulloch–Pitts neurons in considerable detail. The neuron in Figure 1 has two excitatory inputs, a and b, denoted by arrows. It has a threshold of 1. Suppose that during the nth time quantum neither synapse a nor synapse b is active. Therefore there are no active synapses and the neuron threshold of 1 has not been exceeded, so during the $(n+1)$st time quantum this neuron is in the inactive state. Suppose, however, that during the nth quantum input a is active and input b is inactive. The number of active inputs is one, the threshold has been reached, and the neuron becomes active during the next time step. Similarly, if input b is active and input a is inactive, or if both input a and input b are active, the unit will become active.

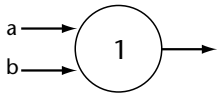


Figure 1. A single McCulloch–Pitts neuron with threshold 1 and two inputs, 'a' and 'b'. This neuron realizes the logic function 'Inclusive OR' because, in words, the unit becomes active when 'input a is active' OR 'input b is active' OR 'both inputs a and b are active'. Arrowheads correspond to excitatory (positive) connections.

There are four possible activity states of the two inputs, a and b, each of which leads to the unit becoming active or inactive during the next time quantum. If these states are enumerated, it can be seen that the unit is realizing the truth-table of the logic function '*Inclusive or*'. In other words, this neuron becomes active if input a is active *or* input b is active *or* both input a and input b are active.

Suppose that we interpret activity or inactivity of the unit as corresponding to particular patterns of truth (activity) or falsity (inactivity) at the neuron's inputs. Then we see that, singly or in groups, McCulloch–Pitts neurons can be considered to be computing logic functions based on the truth-values of the inputs together with the neuron's threshold. The first line of the abstract in the 1943 paper states this precisely: 'Because of the "all-or-none" character of nervous activity, neural events and the relations between them can be treated by means of propositional logic' (McCulloch, 1965, p. 19).

IMPLICATIONS

This paper does a great deal more than merely suggest a biological abstraction that can be used to compute logic functions. The main finding of the paper is reported in the next sentence: 'It is found that the behavior of every net can be described in these terms … and that for any logical expression satisfying certain conditions one can find a net behaving in the fashion it describes' (McCulloch, 1965, p. 19). This result is of breathtaking generality and power, as it means that *any finite logical expression can be realized by a properly constructed network of McCulloch–Pitts neurons*. With a memory added to the system, a network can become a fully general computing device – that is, a Turing machine.

The implication of the paper is that we now know exactly what the brain is doing – it is computing logic functions. As the paper puts it, 'in psychology, introspective, behavioristic or physiological, the fundamental relations are those of two-valued logic' (McCulloch, 1965, p. 38). The strong suggestion which is made in the paper is that psychology and neuroscience have become branches of logic.

GENESIS

Every scientific paper, even one of such great originality and elegance, has roots. It is striking that the paper itself has only three references (Carnap, *The Logical Syntax of Language*; Hilbert and Ackerman, *Foundations of Theoretical Logic*; and Whitehead and Russell, *Principia Mathematica*). Clearly, the sources for the paper were in the fields of logic and philosophy. There were no biological references.

This conclusion is confirmed by Professor Jerry Lettvin in his description of the paper's origin:

> And so early in 1942 Warren invites Walter and me to live with him and his family. Warren and his wife, Rook, were always enormously generous. We settled in, and it was in the evenings then that Walter and Warren got together on 'A Logical Calculus of Ideas Immanent in Nervous Activity.' Walter at that time had read Leibniz, who had shown that any task which can be described completely and unambiguously in a finite number of words can be done by a logical machine. Leibniz had developed the concept of computers almost three centuries back, and had even developed a concept of how to program them.
>
> I didn't realize that at the time. All I knew was that Walter had dredged this idea out of Leibniz, and then he and Warren sat down and asked whether or not you could consider the nervous system [as] such a device. So they hammered out the essay at the end of '42.
>
> (Anderson and Rosenfeld, 1998, p. 3)

LIMITATIONS

Groups of McCulloch–Pitts neurons can build a general-purpose computing device that will have the same limitations and computational power as any digital computer. In fact, one can regard any current digital computer as being built from a limited subclass of McCulloch–Pitts neurons. However, this does not mean that the neurons as described in the paper do everything equally efficiently. Technically, McCulloch–Pitts neurons realize what is sometimes called 'threshold logic', a logic family at one time commercially available in very large-scale integration (VLSI). Certain types of computations can be performed more easily with threshold logic than with more standard logic. Figure 2 shows one example, namely a 'majority' neuron that becomes active if any three or more of its inputs are active. Such a function requires a

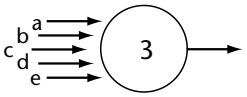


Figure 2. A McCulloch–Pitts neuron can easily compute some logic functions that can require more complex networks of standard logic functions. This unit with threshold 3 and five inputs computes the function 'majority', that is, it becomes active when any three, four, or five of its inputs become active.

combination of several simpler logic functions to realize with traditional semiconductor logic families. However, the slight advantages that threshold logic shows for functions such as 'majority' are lost when computing other simple logic functions (e.g. exclusive OR). A threshold logic unit by itself is basically monotonic in the way in which it responds to excitatory inputs. Once the threshold has been exceeded, the neuron stays on as more and more inputs are excited. However, Exclusive OR and related logic functions such as parity are non-monotonic – that is, increased input excitation can make the unit become inactive.

Figure 3 shows what is required to build a network of McCulloch–Pitts neurons that computes Exclusive OR. The open circle represents an inhibitory synapse. Clearly, this logic function is nowhere near as simple to compute for McCulloch–Pitts neurons as Inclusive OR. (Many simple neural networks also have considerable difficulty in computing highly non-monotonic functions such as Exclusive OR, and for similar reasons, a computational limitation was pointed out in the late 1960s for simple neural networks by Minsky and Papert (1969) in their book, *Perceptrons*.)

More fundamental problems arose with McCulloch–Pitts neurons when the latter were considered for their proposed role as models of the brain. These abstract neurons were simply an inadequate model of the biological neuron. Real neurons frequently seem to act more like analog computers, often using a continuous value as an output parameter (e.g. frequency of action potentials).

McCulloch was very aware of the results that were being obtained from neuroscience. In 1947, he and Pitts wrote another paper, 'How we know universals' (Pitts and McCulloch, 1947), which took a much more modern and also much less general approach to nervous system computation in the cortex and superior colliculus. The models that these researchers proposed used distributed representations and were specialized to perform highly specific computations. In their model for the colliculus, a structure which controls eye movements in primates, they suggested that the eyes were directed to the center of brightness of the visual image – a distributed computation using continuous values, and one which is very similar in spirit to modern models of the colliculus. The great logical generality of the 1943 paper was replaced by a more realistic and specialized approach to neural computation that made extensive use of continuous mathematics.

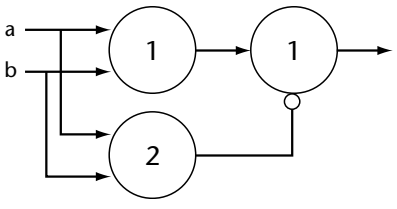


Figure 3. Some elementary logic functions are more difficult to compute with McCulloch–Pitts neurons than others. Although 'Inclusive OR' is easy (see Figure 1), 'Exclusive OR' is more difficult and requires a small network to compute. In words, the output unit in this network becomes active when 'input a is active' OR 'input b is active' but NOT 'both inputs a and b are active'. The circle corresponds to an inhibitory (negative) connection.

There are a number of areas where the human nervous system presents a dual computational aspect. Higher-level cognition, especially language, presents considerable evidence for rule-based operations on discrete symbols, as seen for example in syntax. These are clearly the types of operations for which McCulloch–Pitts neurons seem to be well suited. Yet ample evidence also indicates that there are powerful 'analog' aspects to many nervous system operations, even at the highest levels of cognition.

McCulloch was aware of this problem. He and Pitts had put the discrete, logic-based operations at the lowest level of nervous system hardware – at the neuron level. However, in an essay from 1951 entitled 'Why the mind is in the head', McCulloch suggested that the picture was more complex:

> The nervous system is par excellence a logical machine. Sense organs and effectors are analogical. For example, the eye and ear report the continuous variable of intensity by discrete impulses, the logarithm of whose frequency approximates the intensity. But this process is carried to extremes in our appreciation of the world in pairs of opposites. As Alcmaeon, the first of the experimental neurophysiologists, so well observed, 'the majority of things human are two – white–black, sweet–bitter, good–bad, great–small.' (McCulloch, 1965, p. 73)

McCulloch suggests here that there is an important continuous-to-discrete transition taking place in mental computation. The McCulloch–Pitts neuron and its resulting networks locate the transition at the level of the single neuron. However, it is possible that they had it the wrong way round. Low-level neural processing may be largely continuous, and the discreteness that is so valuable in cognition may arise at the higher levels through other mechanisms. One modern model for such a transition would be discrete attractor states arising from an underlying continuous nonlinear dynamic system.

INFLUENCE

The 1943 paper by McCulloch and Pitts was influential in a large number of places, some of them unexpected. In the realm of mathematics itself this paper is often given credit for founding the important field known as finite state automata theory.

However, its influence went even further. The paper was published at the height of the Second World War. At that time there were a number of projects in progress to build practical computing machines for various military purposes. The teams involved became aware of the McCulloch–Pitts paper very early on. As Norbert Wiener comments:

> In the summer of 1943, I met Dr J. Lettvin … who was … a close friend of Mr Pitts and made me acquainted with his work. He induced Mr Pitts to come out to Boston. … From that time it became clear to us that the ultra-rapid computing machine, depending as it does on consecutive switching devices, must represent an almost ideal model of the problems arising in the nervous system … the construction of computing machines had proved to be … essential for the war effort … and was progressing at several centers. … There was a continual going and coming of those interested. … Everywhere we met with a sympathetic hearing, and the vocabulary of the engineers soon became contaminated with the terms of the neurophysiologist and the psychologist. (Wiener, 1948, pp. 21–23)

One of those influenced was John von Neumann, who sketched what has become known as the 'von Neumann computer architecture' in a famous 1945 technical report. In the report he made the following comment:

> In existing digital computing devices, various mechanical or electrical devices have been used as elements. … It is worth mentioning that the neurons of the higher animals are definitely elements in the above sense. … Following W. PITTS and W. S. MACCULLOCH [sic] … we ignore the more complicated aspects of neuron functioning. … It is easily seen that these simplified neuron functions can be imitated by telegraph relays or by vacuum tubes. (von Neumann, 1945, republished 1982, pp. 388–389)

The proposed similarity between the computer and the architecture of the brain was taken very seriously by computer scientists at the time. When early computer scientists referred to computers as 'giant brains' they were not just using a metaphor, but were referring to what they thought were two computing systems based on the same principles but using different hardware.

From the early 1940s the McCulloch–Pitts neuron was considered by many non-neuroscientists to be the most appropriate way to approach neural computation, largely because the work of McCulloch and Pitts was so well known. The idea that biological neurons were acting primarily as logic devices was an article of faith for decades. Paul Werbos described a discussion he had with Marvin Minsky in the early 1970s. Werbos reported the gist of Minsky's comments as follows:

> Everyone knows a neuron is a 1–0 spike generator. That is the official model from the biologists. … If you and I come out and say the biologists are wrong, and this thing is not producing 1s and 0s, no one is going to believe us. (Anderson and Rosenfeld, 1998, p. 344)

The McCulloch and Pitts paper suggests first that a paper does not have to be correct in its initial domain of application – in this case brain theory – for it to be immensely valuable in many other places and at many different levels, and secondly, that a tight coupling between brain science and computer science has existed from the earliest beginnings of both fields, and has enriched both.

References

Anderson JA and Rosenfeld E (1998) *Talking Nets: An Oral History of Neural Networks*. Cambridge, MA: MIT Press.

McCulloch WS (1951) Why the mind is in the head. In: Jeffress LA (ed.) *Cerebral Mechanisms in Behavior*, pp. 42–111. New York: John Wiley & Sons.

McCulloch WS (1965) *Embodiments of Mind*. Cambridge, MA: MIT Press.

McCulloch WS and Pitts WH (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–133.

Minsky ML (1967) *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice Hall.

Minsky ML and Papert S (1969) *Perceptrons*. Cambridge, MA: MIT Press.

Pitts WH and McCulloch WS (1947) How we know universals: the perception of auditory and visual form. *Bulletin of Mathematical Biophysics* **9**: 127–147.

von Neumann J (1945, republished 1982) First draft of a Report on the EDVAC. In: Randell B (ed.) *The Origins of Digital Computers*, 3rd edn, pp. 383–397. Berlin: Springer.

Wiener N (1948) *Cybernetics*. New York: John Wiley & Sons.

Further Reading

Anderson JA (1996) *An Introduction to Neural Networks*. Cambridge, MA: MIT Press.

Anderson JA and Rosenfeld E (1988) *Neurocomputing. Foundations of Research*. Cambridge, MA: MIT Press.

Anderson JA, Rosenfeld E and Pellionisz A (1990) *Neurocomputing 2*. Cambridge, MA: MIT Press.

Haykin S (1999) *Neural Networks*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Hebb DO (1949) *The Organization of Behavior*. New York: John Wiley & Sons.

McClelland JL and Rumelhart DE (1986) *Parallel Distributed Processing*, vol. 2. *Psychological and Biological Models*. Cambridge, MA: MIT Press.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and retrieval in the brain. *Psychological Review* **65**: 386–408.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing*, vol. 1. *Foundations*. Cambridge, MA: MIT Press.

Memory Consolidation

Introductory article

Morris Moscovitch, University of Toronto and Rotman Research Institute – Baycrest Centre, Toronto, Ontario, Canada

CONTENTS

In memory consolidation, biochemical and psychological processes form new memories and assimilate them into a person's pre-existing body of knowledge.

INTRODUCTION

The term 'consolidation' was introduced in 1900 by G. E. Müller and A. Pilzecker to describe a time-dependent process that was needed to assimilate an experience and store it permanently as a memory that was relatively immune to disruption. Integrating what was known at that time, W. H. Burnham identified two processes that were implicated in consolidation: (1) a physiological or biochemical process required for the formation and storage of a memory trace or engram and (2) a psychological process needed to assimilate the newly acquired memory into an already existing body of knowledge and allow it, in turn, to influence what will be learned subsequently. Elucidating these processes remains at the heart of research on memory and consolidation.

In 1957, the modern era of research into the neural substrates of memory was ushered in by W. B. Scoville and B. Milner's publication on the effects of excision of the anterior and medial temporal lobes bilaterally to control intractable epilepsy in a single patient, H.M. (Figure 1). Although the surgery was effective in controlling the epilepsy of this patient, one of its unanticipated consequences was that H.M. became profoundly amnesic. The memory loss which characterized his amnesia was typical of that observed in other patients with medial temporal lobe damage who were studied subsequently, and indeed in many individuals with organic amnesia of different etiologies which affected midline thalamic nuclei. Although they lacked substantial portions of the medial temporal lobes or critical diencephalic nuclei, these individuals had a normal short-term memory as measured by a number of tests, including digit-span performance, which involves repeating back a series of numbers. Similarly, it was reported that deficits in remote memory were limited to retrieval of events that had occurred within the past few years, suggesting that older memories were stored, and could be retrieved readily, without the medial temporal lobes. These observations were interpreted as showing that the medial temporal lobes and related diencephalic structures were involved neither in processing short-term memories nor in storing remote memories. Instead, their function was to help to consolidate memories and to encode, store and retrieve them until that process was complete. Indeed, the standard model of consolidation was based on these initial observations, and has been modified little since then although, as will be discussed later in this article, there is some evidence to suggest that the standard model needs to be modified.

The Standard Model

According to the standard model, memory consolidation begins when information, which is registered initially in the neocortex, is bound into a memory trace by the medial temporal lobes and related structures in the diencephalon. This initial binding into a memory trace involves short-term processes, the first of which may be completed within seconds and the last of which may be completed within minutes or at most days. This is referred to as *rapid consolidation* or *cohesion*. A process of *prolonged consolidation* is also believed to occur. During this process, the medial temporal lobes and related structures are needed for storage and recovery of the memory trace, but their contribution diminishes as prolonged consolidation proceeds, until the neocortex alone is capable of sustaining the permanent memory trace and mediating its retrieval. Thus according to the standard model the medial temporal lobes and related structures are considered to be temporary memory systems,

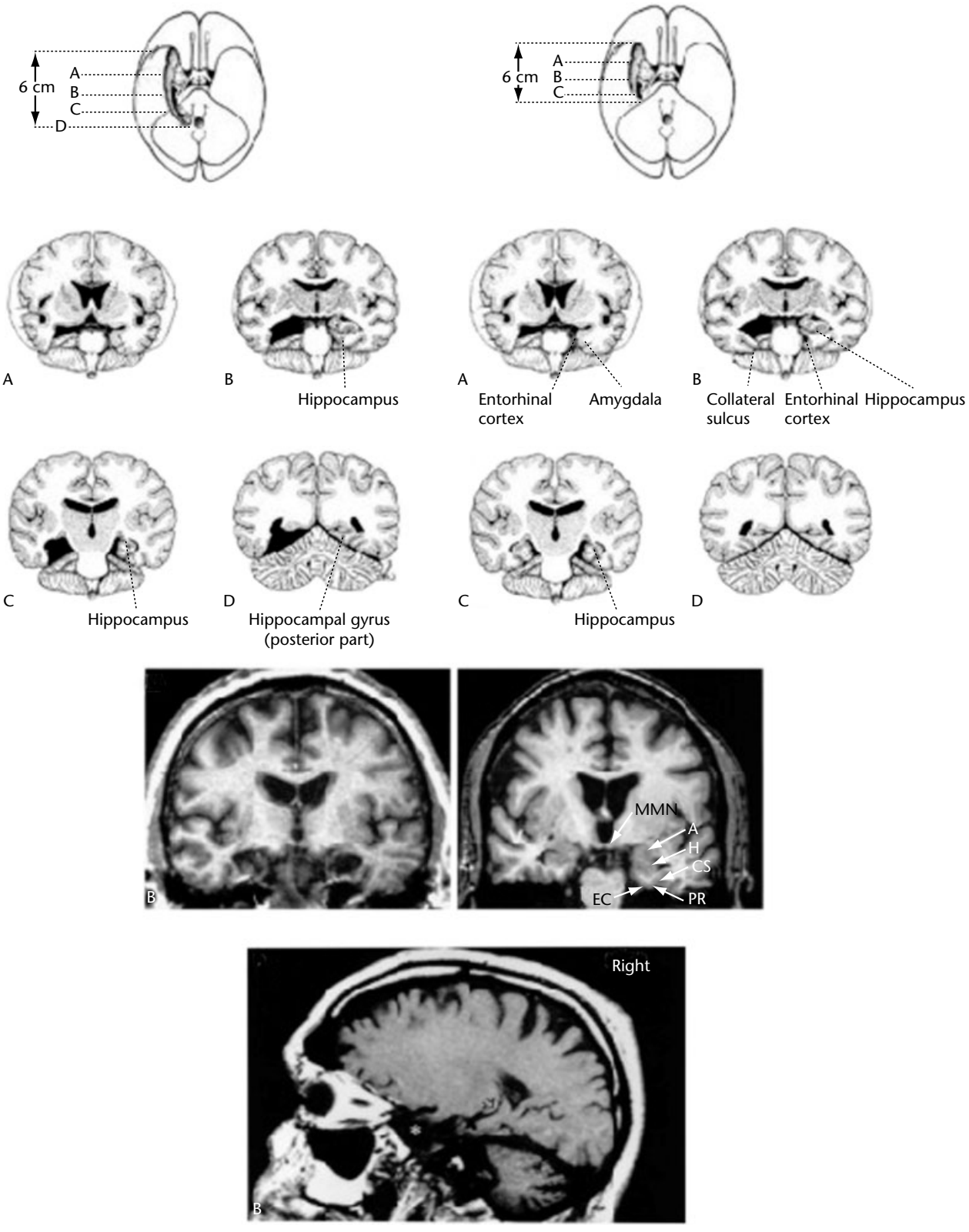


Figure 1. A recreation of H.M.'s lesion from the surgeon's report (Scoville and Milner, 1957) and a recent MRI scan of the lesion (Corkin *et al*., 1996).

which are needed to store and retrieve memories until prolonged consolidation is complete. The time it takes for consolidation to be completed is determined by the temporal extent of retrograde amnesia following lesions of the medial temporal lobes and diencephalon, other insults (concussions, closed head injuries or electrical currents), or the administration of pharmacological agents which permanently disrupt memory.

Rapid Consolidation

The existence of rapid consolidation is not in dispute. Indeed, much has been learned about its cellular and neurochemical (molecular) basis which seems to be similar not only across species but also across different memory systems in the same species. Stimulus presentation initiates a cascade of neurochemical events at the synaptic membrane and within the cell which increase the synaptic strength or efficiency with which neurons that form the memory trace can communicate with (activate) one another. The first of these processes involves local, transient molecular modifications that lead to an increase in neurotransmitter release at the affected synapse, and which may mediate short-term or intermediate memory. If the stimulus is intense enough and/or repeated, additional processes are activated. These involve genetic transcription and protein formation, which lead to long-lasting cellular changes, including the creation of new synapses, that support the formation and maintenance of long-term memory. These processes may last from hours to days. Although we are well on our way to understanding the basic cellular and molecular mechanisms of memory, we are far from understanding memory at a systems level which includes the psychological processes that Burnham emphasized. It is this problem that is inextricably tied to the debate with regard to prolonged consolidation.

Prolonged Consolidation and Memory Systems

By the 1960s, the outlines of the central debate concerning the validity of the standard model were already clearly crystallized in research with amnesic patients, reflecting the assumption that it was damage to the medial temporal lobes and diencephalon that was primarily responsible for the amnesia. The debate continues to center on three questions. (1) What types of memory are implicated? (2) Which neuroanatomical structures in the medial temporal lobes and diencephalon are involved? (3) What is the extent and duration of retrograde amnesia and, by implication, of consolidation, and how is it affected by lesion location and memory type?

TYPES OF MEMORY

Explicit and Implicit Memory

One of the most important recent discoveries of memory researchers is that memory is not unitary but consists of various types, each influenced by different variables, governed by different principles, possibly concerned with different materials, and each mediated by different neural structures and mechanisms that form distinguishable and dissociable systems. Two broad classes of memory were identified, namely *explicit memory*, which refers to conscious recollection of experiences and facts (also called declarative memory), and *implicit memory*, which is memory without awareness. The latter is revealed by the effects that prior experience has on behavior without the individual consciously retrieving the memory or even being aware of having it (non-declarative memory). One example of implicit memory is perceiving a picture of a face or a word more quickly after it has already been seen, although the person may deny that the face or word was familiar (perceptual priming). Other examples include learning a repeated, complex motor sequence, even though the individual may be unaware of the sequence or that it was repeated (procedural memory), or learning to form conditioned responses, although the individual may be unaware of the stimuli controlling the response (conditioning). What is important is that many if not all types of implicit memory can be acquired, retained and retrieved normally even by individuals who are profoundly amnesic as a result of medial temporal or diencephalic damage. It is believed that implicit memory is mediated by the neural structures involved in acquiring information (e.g. the posterior neocortex for perceptual priming) and in executing action, (e.g. the basal ganglia for motor learning). We know little about prolonged consolidation effects in implicit memory, or even whether it occurs. Our discussion of prolonged consolidation will therefore be restricted to explicit memory.

Explicit memory: episodic versus semantic

Explicit memory is itself divisible into two types, namely episodic and semantic memory. Episodic memory is memory for particular, autobiographical episodes that have a distinct spatio-temporal

context, and it involves a detailed re-experiencing of the initial event – what E. Tulving called 'mental time travel'. Semantic memory, on the other hand, is memory for knowledge that lacks a spatio-temporal context, such as knowledge of vocabulary and facts about the world (e.g. history, geography, people) and knowledge about ourselves (e.g. where we were born, where we lived, who our friends were, what schools we attended, what jobs we held). Some authors have called the latter type of memory *personal semantics*, to distinguish it from memory for autobiographical episodes. Episodic memory is assessed by descriptions of autobiographical episodes from different periods in an individual's life, vocabulary (by recognition and definition of words that came into use at different times) and knowledge of people (by recognition of faces and names in the news from different time periods).

Although different types of tests are used to assess episodic and semantic memory, they are considered to be similar with regard to consolidation by proponents of the standard model. According to supporters of the latter, damage to the medial temporal lobes and diencephalon leads to a graded, temporally limited retrograde amnesia for both types of memory. Memories acquired most recently are most severely affected, with more remote memories being retained normally, having been fully consolidated before the insult.

The standard model of consolidation had been challenged by Warrington and her colleagues, who showed that retrograde amnesia can be severe and of long duration following medial temporal lesions. This led them to favor the view that amnesia results from a deficit in retrieval rather than in consolidation. On the basis of the evidence that they collected, M. Kinsbourne and F. Wood argued that amnesia is a deficit only of episodic (autobiographical) memory, and does not distinguish between recent and remote memory. Although few authors endorsed their ideas at the time, L. Nadel and M. Moscovitch have noted a number of problems with the standard model with regard to both the types of memory that are affected and the duration and extent of retrograde amnesia. The latter varies with memory type, decreasing in severity and extent from the autobiographical to the semantic. Retrograde amnesia for details of autobiographical events after large medial temporal (or diencephalic) lesions can extend for decades, or even a lifetime – far longer than would be biologically plausible for even prolonged consolidation to be completed. However, retrograde amnesia for public events and personalities is less extensive and is often temporally graded. This is even more true of semantic memory that pertains to vocabulary, to facts about the world and to personal semantics.

The distinction between temporally extensive and temporally limited retrograde amnesia also applies to spatial memory. Schematic cognitive maps of old neighborhoods that are adequate for navigation are retained, but they lack topographical details and local environmental features, such as the appearance and location of particular homes, that would allow the person to have detailed cognitive maps of their locale.

NEUROANATOMY OF MEMORY

It has also been noted that lesion size and location play a role in determining the nature, severity and extent of retrograde amnesia. The initial studies on retrograde amnesia implicated the medial temporal lobes and diencephalon. These areas themselves consist of a number of separate but related structures (Figures 2 and 3). Initially, the focus of attention shifted quickly from the medial temporal lobes to the hippocampal formation, and then to the hippocampus itself. More recently, however, investigators have begun to appreciate the importance of the other structures, and the different functions each serves, as well as their relationship to each other and to corresponding regions in the diencephalon. Although the precise function of each of the areas is still being debated, a consensus is emerging. It has been proposed that there are two integrated medial–temporal–diencephalic memory systems. One system, consisting of the hippocampus and its connections to the mamillary bodies and anterior thalamic nuclei, mediates recollection

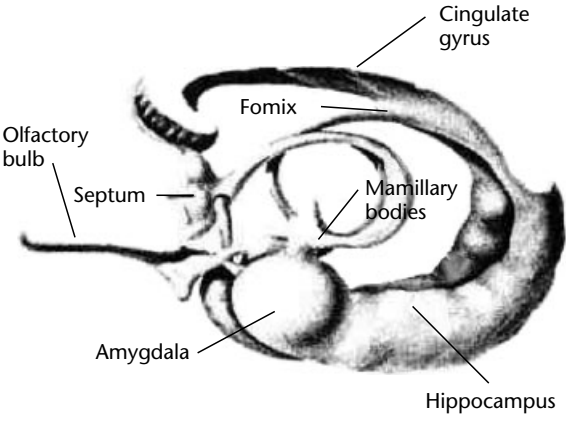


Figure 2. The limbic system and memory circuit (Kolb and Whishaw, 1996).

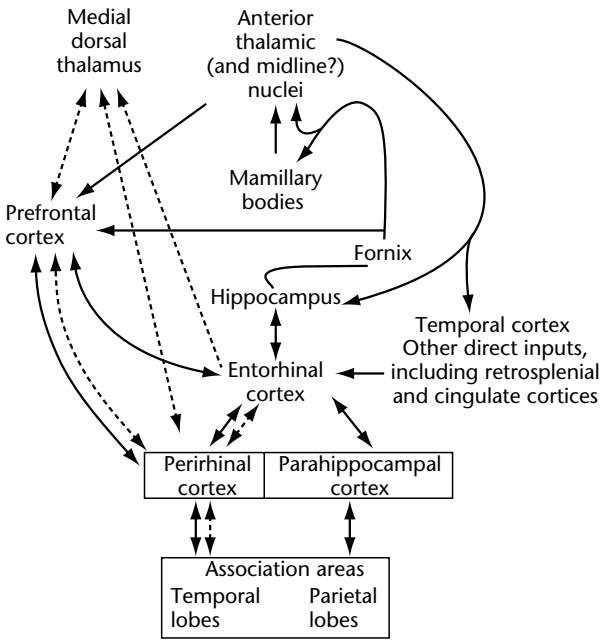


Figure 3. The hippocampal–diencephalic systems (modified from Aggleton and Brown, 1999).

which relies on relational information, including the temporal–spatial context of the memory. Damage to this system causes deficits in spatial memory and in memory for complex relational information that typifies memory for autobiographical episodes. The other system, which consists of the perirhinal cortex and its connections to the dorsomedial nucleus of the thalamus, is necessary for item recognition based on familiarity judgments without a spatial–temporal context. Damage to this system will impair recognition of even single items. The function of the parahippocampal cortex is less well understood, although it seems to be necessary for forming memories of places or for associating objects with particular locations.

Defenders of the standard theory argue that the temporally limited memory applies only to the hippocampus, and that permanent memories are consolidated in the adjacent regions. Such an argument would represent a significant departure of the standard theory from its origins in response to evidence brought against it. Until recently, standard theory assumed the locus of consolidated memory to be in lateral and posterior neocortex, not in the parahippocampus or other structures in the medial temporal lobes adjacent to the hippocampus. Those structures were regarded as part of the extended hippocampal system or complex in the medial temporal lobes, whose function was to help to consolidate memories, not to be involved permanently in storing and retrieving them. Indeed, as we learn more about the separate functions of these regions, it may be reasonable to consider the possibility that each of them is involved in retention and retrieval of those aspects of information which they specifically process. Moreover, more recent studies have shown that damage which is confined largely or exclusively to the hippocampus can often lead to extended retrograde memory loss for autobiographical events, the severity of this loss being dependent on the extent of hippocampal damage (Figure 4).

Damage to extrahippocampal structures in the medial temporal lobes can lead to loss of remote memories for facts, events and people, the latter being particularly associated with damage to the anterior temporal pole. Loss of semantic memory, including vocabulary and personal semantics, is associated with damage to posterior neocortical structures, particularly the lateral aspects of the temporal lobe. Such loss is evident in many patients with dementia and neocortical degeneration, but is most revealing in individuals with semantic dementia whose medial temporal lobes are relatively spared. They show a reverse temporal gradient, with recent memories being preserved and remote memories being impaired. Although the reverse gradient has been observed in some patients for both episodic and semantic memory, in other patients it has been noted only for semantic memory, with episodic, autobiographical memories being spared, presumably because they are dependent on the medial temporal lobes. It has yet to be determined what accounts for the individual differences in the patterns of preservation and loss among semantic dementia patients.

Multiple-Trace Theory: An Alternative to the Standard Model

Based on the functional and neuroanatomical evidence that they reviewed, Nadel and Moscovitch concluded, contrary to the traditional consolidation model, that the function of the medial temporal system is not temporally limited, but that it is needed to represent even old memories in rich detail (be it autobiographical or spatial) for as long as the memory exists. Neocortical structures, on the other hand, are sufficient to form domain specific and semantic representations based on regularities extracted from repeated experiences with words, objects, people, and environments. This applies even to autobiographical episodes that one recollects repeatedly, creating a gist of each episode which lacks the details that make re-experiencing it

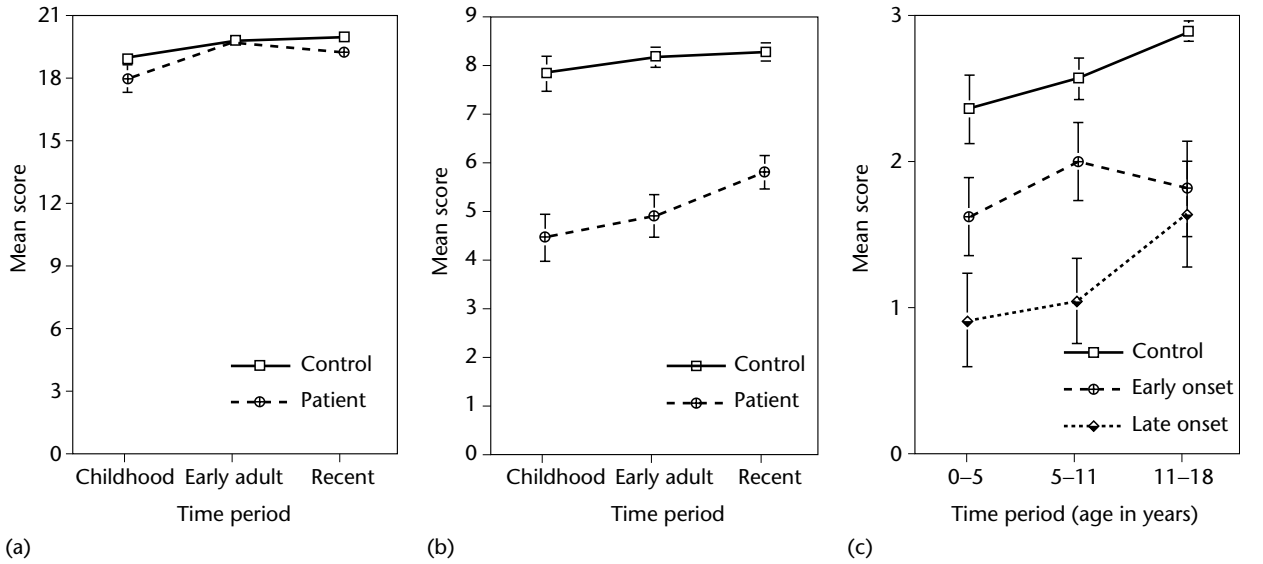


Figure 4. Mean scores on the Autobiographical Memory Inventory (AMI) for control ($n = 22$) and patient ($n = 25$) groups. Vertical lines depict standard errors of the means. (a) Personal semantic component. The maximum score is 21 per time period. (b) Autobiographical episodic component. The maximum score is 9 per time period. (c) Mean scores on episodic components of AMI for control ($n = 22$), late seizure onset ($n = 11$) and early seizure onset ($n = 8$). Late seizure onset describes patients who reported first seizures after the age of 18 years, and early seizure onset describes patients who reported first seizures before the age of 5 years. The maximum score is 3 per time period.

possible. The medial temporal lobe system may have a role in the initial formation of these neocortical representations, but once they have been formed they can exist on their own.

Recent evidence from studies of children whose hippocampus was damaged at birth or shortly thereafter supports this view. F. Vargha-Khadem and colleagues found that these children acquired sufficient general knowledge (semantic memories) to complete high school even though their memory for autobiographical episodes was impaired, although L. R. Squire and S. M. Zola have presented some counterarguments.

Corroborating evidence is also provided by neuroimaging studies of recent and remote autobiographical and semantic memory. These studies found that the hippocampus is activated equally during retrieval of recent and remote autobiographical memories, but not during retrieval of memories for public events or personal semantics (Figure 5). However, remote memory for famous faces shows the expected temporal gradient, particularly in the entorhinal cortex, with recent memories being much more strongly activated than remote ones.

To account for this evidence, Nadel and Moscovitch proposed a *multiple-trace theory (MTT)*. The hippocampal complex (and possibly diencephalon) rapidly and obligatorily encodes all information that is attended (consciously apprehended), and binds the neocortical (and other) neurons which represent that experience into a memory trace. This information is sparsely encoded in a distributed network of hippocampal complex neurons which act as a pointer, or index, to the neurons that represent the attended information. Thus a memory trace of an episode consists of a bound ensemble of neocortical and hippocampal/medial temporal lobe (and possibly diencephalic) neurons which represent a memory of the consciously experienced event. Formation and consolidation of these traces, or cohesion, is relatively rapid, lasting for a period of the order of seconds or at most days.

In this model there is no prolonged consolidation process that slowly strengthens the neocortical component of the memory trace so that, with time, it becomes independent of the hippocampal complex. Instead, each time an old memory is retrieved, a new hippocampally mediated trace is created so that old memories are represented by more or stronger traces than are new ones, and are therefore less susceptible to disruption due to brain damage than are more recent ones. Because the memory trace for autobiographical episodes is distributed in the hippocampal complex, the extent and severity of retrograde amnesia and perhaps the slope of the gradient are related to the amount and

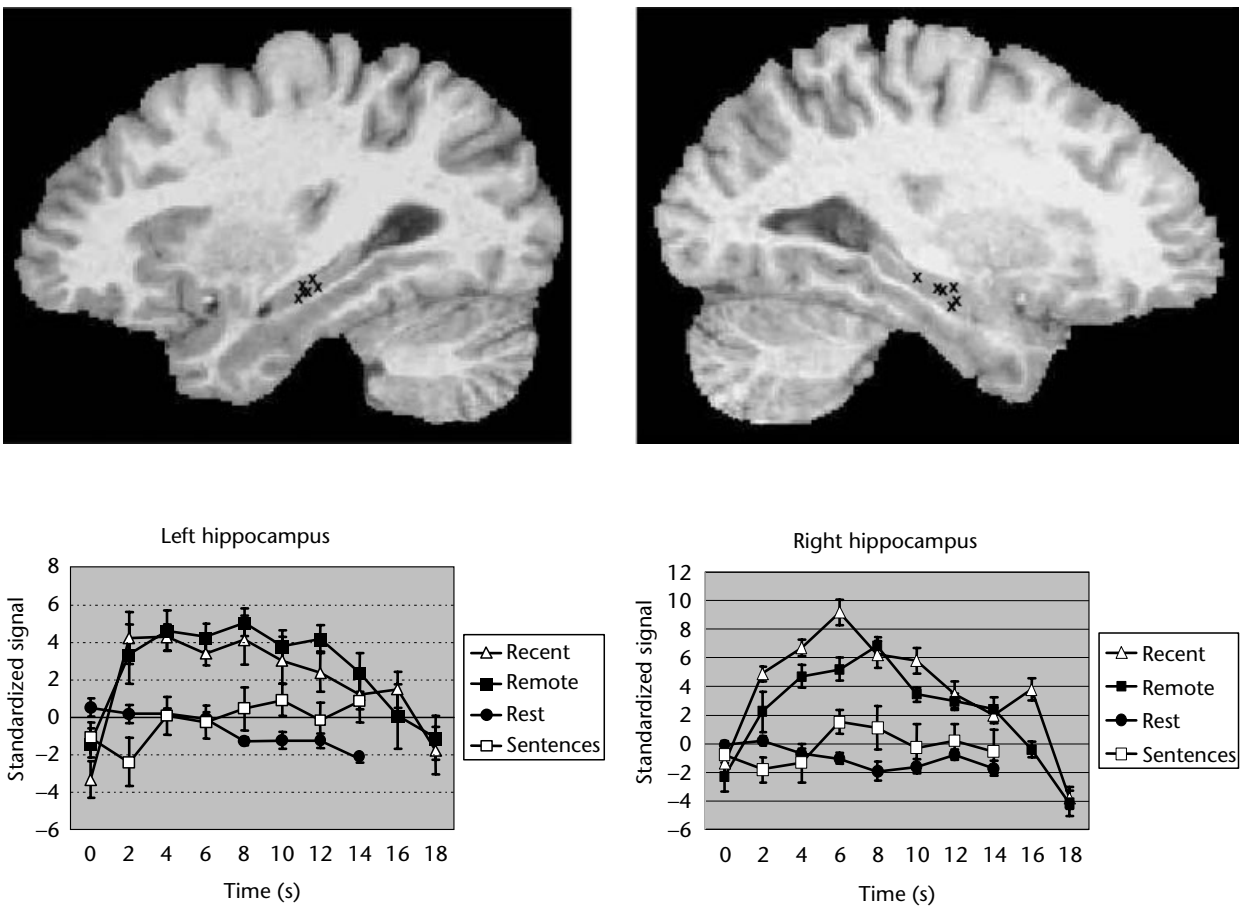


Figure 5. Hemodynamic response of the hippocampus during recall of recent and remote memories, and two baseline conditions (rest and sentence completion) (Ryan *et al*., 2001).

location of damage to the extended hippocampal complex.

Whereas each autobiographical memory trace is unique, the creation of multiple, related traces facilitates the extraction of the neocortically mediated information which is common to all of them, and which is shared with other episodes. This information is then integrated with pre-existing knowledge to form semantic memories that can exist independently of the hippocampal complex. Thus facts about the world, people and events that are acquired in the context of a specific episode are separated from it and ultimately stored independently of it. This process of increased semanticization may give the impression of prolonged consolidation (see below).

Prolonged consolidation for semantic memory: two alternatives

Remote memories for the gist of events, and for personal and public semantics, are not similarly dependent on the continuing function of the hippocampal complex. Instead, the hippocampal complex is needed only temporarily, until the memory is represented permanently in neocortical structures that are specialized for processing the information and capable of being modified while doing so. Semantic memory therefore behaves in a manner consistent with both the standard model and the MTT. However, the two models' explanations of the effects differ. According to the standard model, the memory that is held temporarily in the medial temporal lobes is identical to the memory that is stored permanently in the neocortex. Indeed, many believe that prolonged consolidation effects a transfer of the same memory from one location to another by strengthening neocortical connections. On the other hand, the MTT assumes that the temporary medial temporal memory is fundamentally different to the permanent neocortical one. The former retains its episodic flavor, such that the semantic content is tied to the spatio-temporal (autobiographical) context in which it was acquired. However, the latter is

stripped of its episodic context and retains only the semantic core. By this view, prolonged consolidation refers to the establishment of a semantic trace that can survive on its own, but it does not involve the loss of the episodic trace, nor is it identical to it.

Indeed, according to the MTT, the two types of memories can coexist, so that an individual can have both an episodic and a semantic representation of the same event, object or fact, and they can lose one type of representation without losing the other. The evidence from individuals with semantic dementia supports this interpretation. Having lost the neocortical, semantic representation, they rely on medial temporal representations to identify objects, people, places and facts. Thus they will recognize an object that has autobiographical significance for them (e.g. their own vase or kettle) but not the same type of object which is not their own (e.g. another person's vase or kettle). Conversely, amnesic people with medial temporal damage will recognize objects and individuals regardless of their autobiographical significance, but will be unable to evoke an autobiographical event related to them. However, it should be noted that the detailed episodic trace may also fade in normal people, as most memories do unless they are rehearsed, leaving only the general, semantic memory behind.

Animal Models of Consolidation and Retrograde Memory Loss

A brief examination of the animal literature is instructive because in animal research greater control can be exercised over lesion size and location and over pharmacological and other interventions, many of which cannot be attempted in humans.

For the most part, developments in the animal literature parallel those in human research. Despite the widely held view that retrograde amnesia is relatively brief and temporally graded, even in animals, not all of the evidence is consistent with this pattern. For example, a temporally graded, time-limited memory loss has been observed for a socially acquired food preference, for object discrimination learning and for contextual fear conditioning. On the other hand, extensive retrograde memory loss with a flat gradient has been observed for tests of allocentric, spatial memory, whether measured in a water maze or in a radial arm maze. Such evidence has forced investigators to reconsider the state of premorbid memories following medial temporal and in particular hippocampal damage. As in humans, the extent and severity of retrograde memory loss depend on the type of task and on lesion location. To account for these discrepancies, Rosenbaum and colleagues distinguished between context-dependent and context-free memories, which in humans correspond to episodic and semantic memories, respectively. Only memories that are dependent on relational context, such as spatial ones, show extensive retrograde loss following hippocampal lesions. However, context-free memories, such as memory for objects, may show only a temporally limited, graded memory loss after hippocampal lesions, but an extensive loss with a flat gradient following lesions to extrahippocampal structures, such as the perirhinal cortex. Extrapolating from these results, one can predict that severe and extensive retrograde memory loss in animals, as in humans, would require damage to the entire medial temporal lobe.

Studies on reconsolidation testify to the continued participation of different regions of the medial temporal lobe in memory long after the memory has been learned and presumably consolidated in the neocortex or other brain structures. The point of departure for these studies is the finding that electroconvulsive shock (ECS), lesions and protein-synthesis inhibitors are effective amnestic agents only if administered shortly after learning occurs. Delaying their administration beyond a critical period renders them ineffective, presumably because consolidation is already complete. However, if the memory is re-instated by having the animal perform the learned task long afterwards (even days or weeks later), when consolidation is certainly complete, then the amnestic agent again becomes effective. In other words, the effectiveness of the amnestic agent depends on when it is administered relative to the time when the memory was most recently activated, not relative to the time when it was acquired initially. With regard to consolidation, these studies indicate that memories are vulnerable when they are reinstated, and that they require a rapid reconsolidation process in order to remain permanent. The fact that damage to the hippocampus, or inhibition of protein synthesis in the hippocampus or amygdala, can lead to amnesia of well-learned memories, suggests that these structures – contrary to the standard model – continue to be involved in representing remote memories long after consolidation is presumably complete. These findings, and the idea that each time a memory is reinstated it is again re-encoded (rapid consolidation or cohesion), are consistent with the predictions of the MTT.

CONCLUSION AND FUTURE DIRECTIONS

We are beginning to understand consolidation at the cellular level, particularly those processes that mediate rapid consolidation. Undoubtedly advances will be made in discovering more about the molecular mechanisms that are implicated, and in developing drugs and gene therapies that can be used to treat memory disorders and even improve normal memory.

However, there is still much to be learned about consolidation at the systems level. Are different types of memories consolidated in different structures? Is the time course of consolidation similar for all of them? Are they each susceptible to different amnestic agents? What are the interactions between various structures and memory systems in consolidating information? These questions are particularly relevant when considering prolonged consolidation. As was noted above, recent evidence has challenged the standard model of consolidation, and there is even some dispute as to whether prolonged consolidation exists, at least for some types of memory. The alternative model, namely the MTT, seems to be more consistent with evidence that there are different memory systems, each of which is governed by different principles and mediated by different structures, and has different consolidation processes. It is hoped that future research will help to elucidate the exact nature of these memory systems, the role of various structures in memory representation and consolidation, and the nature of the interaction between the hippocampus and the neocortex, and between these and other brain structures.

Further Reading

Aggleton JP and Brown MW (1999) Episodic memory, amnesia and the hippocampal anterior thalamic axis. *Brain and Behavioral Sciences* **22**: 425–489.

Burnham WH (1904) Retroactive amnesia: illustrative cases and a tentative explanation. *American Journal of Psychology* **14**: 392–396.

Cermak LS (ed.) (1982) *Human Memory and Amnesia*. Hillsdale, NJ: Erlbaum.

Cermak LS and O'Connor M (1983) The anterograde and retrograde retrieval ability of a patient with amnesia due to encephalitis. *Neuropsychologia* **21**: 213–234.

Cipilotti L, Shallice T, Chan D *et al*. (2001) Long-term retrograde amnesia…the crucial role of the hippocampus. *Neuropsychologia* **39**: 151–172.

Conway MA, Turk DJ, Miller SL *et al*. (1999) A positron emission tomography (PET) study of autobiographical memory retrieval. *Memory* **5–6**: 679–702.

Corkin S (1984) Lasting consequences of bilateral medial temporal lobectomy: clinical course and experimental findings in HM. *Seminars in Neurology* **4**: 252–262.

Corkin S, Amaral DG, Gonzalez G, Johnson KA and Hyman BT (1997) H.M.'s medial temporal lobe lesion: findings from magnetic resonance imaging. *Journal of Neuroscience* **17**: 3964–3979.

Crovitz HF and Schiffman H (1974) Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society* **4**: 519–521.

Eichenbaum H (1999) The hippocampus and mechanisms of declarative memory. *Behavioural Brain Research* **103**: 123–133.

Fujii T, Moscovitch M and Nadel L (2000) Consolidation, retrograde amnesia and the temporal lobe. In: Boller F and Grafman J (eds) *The Handbook of Neuropsychology*, vol. 4, pp. 223–250. Amsterdam: Elsevier.

Graham KS and Hodges JR (1997) Differentiating the roles of the hippocampal system and the neocortex in long-term memory storage. *Neuropsychology* **11**: 77–89.

Graham KS, Patterson K and Hodges JR (1999) Episodic memory: new insights from the study of semantic dementia. *Current Opinion in Neurobiology* **9**: 245–250.

Kandel ER (2001) The molecular biology of memory storage: a dialogue between genes and synapses. *Science* **294**: 1030–1038.

Kapur N (1999) Syndromes of retrograde amnesia: a conceptual and empirical analysis. *Psychological Bulletin* **125**: 800–825.

Kim JJ and Fanselow MS (1992) Modality-specific retrograde amnesia of fear. *Science* **256**: 675–677.

Kinsbourne M and Wood F (1975) Short-term memory processes and the amnesic syndrome. In: Deutsch D and Deutsch AJ (eds) *Short-Term Memory*. New York: Academic Press.

Kolb B and Whishaw IQ (1996) *Fundamentals of Human Neuropsychology*, 4th edn. New York: WH Freeman.

Kopelman MD, Wilson BA and Baddeley AD (1989) The autobiographical memory interview: a new assessment of autobiographical and personal semantic memory in amnesic patients. *Journal of Clinical and Experimental Neuropsychology* **5**: 724–744.

Kopelman MD, Stanhope N and Kingsley D (1999) Retrograde amnesia in patients with diencephalic, temporal lobe or frontal lesions. *Neuropsychologia* **37**: 939–958.

Korsakoff SS (1889) Etudes medico psychologique sur une forme du maladi de la memoire. *Revve Philosophique* **28**: 501–530 (translated and republished by Victor M and Yakovlev PI (1955) *Neurology* **5**: 394–406).

McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–457.

McGaugh JL (2000) Memory – a century of consolidation. *Science* **287**: 248–251.

Maguire EA (in press) Neuroimaging studies of autobiographical event memory. *Philosophical*

Transactions of the Royal Society of London B **356**: 1441–1451.
Milner B (1966) Amnesia following operation on the temporal lobe. In: Whitty CWM and Zangwill OL (eds) *Amnesia*, pp. 109–133. London: Butterworth.
Milner B, Squire LR and Kandel ER (1998) Cognitive neuroscience and the study of memory. *Neuron* **20**: 445–468.
Morris RG, Nunn JA, Abrahams S, Feigenbaum JD and Recce M (1998) The hippocampus and spatial memory in humans. In: Burgess N and Jeffery KJ (eds) *The Hippocampal and Parietal Foundations of Spatial Cognition*, pp. 259–289. London: Oxford University Press.
Moscovitch M (1992) Memory and working with memory: a component process model based on modules and central systems. *Journal of Cognitive Neuroscience* **4**: 257–267.
Moscovitch M (1995) Recovered consciousness: a hypothesis concerning modularity and episodic memory. *Journal of Clinical and Experimental Neuropsychology* **17**: 276–291.
Moscovitch M (2001) *Amnesia*. In: Smesler NJ and Baltes P (eds) *International Encyclopedia of Social and Behavioral Sciences*. Amsterdam: Elsevier.
Moscovitch M and Nadel L (1998) Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Current Opinion in Neurobiology* **8**: 297–300.
Moscovitch M, Vriezen E and Goshen-Gottstein Y (1993) Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In: Boller F and Spinnler H (eds) *The Handbook of Neuropsychology*, vol. 8, pp. 133–173. Amsterdam: Elsevier.
Moscovitch M, Yaschyshyn T, Ziegler M and Nadel L (1999) Remote episodic memory and amnesia: was Endel Tulving right all along? In: Tulving E (ed.) *Memory, Consciousness and the Brain: the Tallinn Conference*. New York: Psychology Press.
Müller GE and Pilzecker A (1900) Experimentelle beiträge zur lehre vom gedächtnis. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* **S1**: 1–288.
Mumby DG, Astur RS, Weisend MP and Sutherland RJ (1999) Retrograde amnesia and selective damage to the hippocampal formation: memory for places and object discriminations. *Behavioural Brain Research* **106**: 97–107.
Murray EA and Bussey TJ (2001) Consolidation and the medial temporal lobe revisited: methodological considerations. *Hippocampus* **11**: 1–7.
Nadel L and Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*. **7**: 217–227.
Nadel L and Land C (2000) Memory traces revisited. *Nature Reviews: Neuroscience* **1**: 209–212.
Nadel L and Moscovitch M (2001) The hippocampal complex and long-term memory revisited. *Trends in Cognitive Science* **5**: 228–230.
Nadel L, Samsonovich A, Ryan L and Moscovitch M (2000) Multiple trace theory of human memory: computational, neuroimaging and neuropsychological results. *Hippocampus* **10**: 352–368.
Reed JM and Squire LR (1998) Retrograde amnesia for facts and events: findings from four new cases. *Journal of Neuroscience* **18**: 3943–3954.
Rosenbaum RS, Priselac S, Kohler S *et al*. (2000) Remote spatial memory in an amnesic person with extensive bilateral hippocampal lesions. *Nature Reviews: Neuroscience* **3**: 1044–1048.
Rosenbaum RS, Winocur G and Moscovitch M (2001) New views on old memories: re-evaluating the role of the hippocampal complex. *Behavioral Brain Research* **127**: 183–197.
Rozin P (1976) The psychobiological approach to human memory. In: Rozenzweig R and Bennett EL (eds) *Neural Mechanisms of Learning and Memory*. Cambridge, MA: MIT Press.
Ryan L, Nadel L, Keil K *et al*. (2001) The hippocampal complex and retrieval of recent and very remote autobiographical memories: evidence from functional magnetic resonance imaging in neurologically intact people. *Hippocampus* **11**: 707–714.
Schacter DL and Badgaiyan RD (2001) Neuroimaging of priming: new perspectives on implicit and explicit memory. *Current Directions in Psychological Science* **10**: 1–4.
Schacter DL and Tulving E (eds) (1994) *Memory Systems*. Cambridge, MA: MIT/Bradford Press.
Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21.
Snowden JS, Griffiths HL and Neary D (1994) Semantic dementia: autobiographical contribution to preservation of meaning. *Cognitive Neuropsychology* **11**: 265–288.
Snowden JS, Griffiths HL and Neary D (1996) Semantic-episodic memory interactions in semantic dementia: implications for retrograde memory function. *Cognitive Neuropsychology* **13**: 1101–1137.
Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review* **99**: 195–231.
Squire LR and Alvarez P (1995) Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology* **5**: 169–177.
Squire LR and Zola SM (1998) Episodic memory, semantic memory and amnesia. *Hippocampus* **8**: 205–211.
Squire LR, Cohen NJ and Nadel L (1984) The medial temporal region and memory consolidation: a new hypothesis. In: Weingartner H and Parker E (eds) *Memory Consolidation*, pp. 185–210. Hillsdale, NJ: Erlbaum.
Teng E and Squire LR (1999) Memory for places learned long ago is intact after hippocampal damage. *Science* **400**: 675–677.
Teyler TJ and DiScenna P (1986) The hippocampal memory indexing theory. *Behavioral Neuroscience* **100**: 147–154.
Tulving E (1972) Episodic and semantic memory. In: Tulving E and Donalson W (eds) *Organisation of Memory*, pp. 381–403. New York: Academic Press.

Tulving E (1983) *Elements of Episodic Memory*. Oxford: Clarendon Press.

Tulving E (1985) Memory and consciousness. *Canadian Psychologist* **25**: 1–12.

Tulving E and Schacter DL (1990) Priming and human memory systems. *Science* **247**: 301–306.

Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*. Oxford: Oxford University Press.

Vargha-Khadem F, Gadian DG, Watkins KE *et al*. (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–380.

Viskontas IV, McAndrews MP and Moscovitch M (2000) Remote episodic memory deficits in patients with unilateral temporal lobe epilepsy and excisions. *Journal of Neuroscience* **20**: 5853–5857.

Warrington EK and Weiskrantz L (1970) Amnesic syndrome: consolidation or retrieval? *Nature* **228**: 628–630.

Warrington EK and Sanders HI (1971) The fate of old memories. *Quarterly Journal of Experimental Psychology* **23**: 432–442.

Warrington EK and McCarthy RA (1988) The fractionation of retrograde amnesia. *Brain and Cognition* **7**: 184–200.

Westmacott R, Leach L, Freedman M and Moscovitch M (2001) Different patterns of autobiographical memory loss in semantic dementia and medial temporal lobe amnesia: a challenge to consolidation theory. *Neurocase* **7**: 37–55.

Wiggs CL and Martin A (1998) Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology* **8**: 227–233.

Winocur G (1990) Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioural Brain Research* **38**: 145–154.

Zola-Morgan S and Squire LR (1990) The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science* **250**: 288–290.

Memory, Neural Basis of: Cellular and Molecular Mechanisms

Introductory article

Mark R Rosenzweig, University of California, Berkeley, California, USA

CONTENTS

Through the ages, there has been much speculation about how memory works. The current consensus is that memories are stored in circuits of neurons by plastic neurochemical and neuroanatomical changes at junctions between neurons (synapses), but it remains to be determined which of these changes are necessary and sufficient for memory formation.

INTRODUCTION

Recorded speculations about mechanisms of memory go back over two thousand years, but only in the last fifty years has substantial empirical progress been made in solving this mystery. In Greek mythology, memory was the province of the goddess Mnemosyne. She was the mother of the nine Muses, goddesses who presided over learning and the arts and sciences. This relationship demonstrated the necessity of memory for creativity. We invoke the name of Mnemosyne whenever we call methods to aid memory 'mnemonics' or 'mnemonic devices'.

Speculations about the bodily mechanisms of memory have been related to the technology of each period of history. Thus, classical Greek and Latin authors used as their models or metaphors of memory processes the then-current technology of wax slates and of signet rings impressing wax seals. Socrates assumed that there is a block of wax in our souls, the gift of Mnemosyne. He suggested that the wax varied in quality in different individuals, with finer wax allowing sharper, more detailed impressions. More recent models of memory mechanisms have ranged from telephone exchanges, to computers, to storage of genetic information.

In the Renaissance, when water was used to activate mechanical devices, nerves were thought to be tubes that conducted a fine fluid called 'animal (or animate) spirits'. By the mid-nineteenth century, nerve cells were visualized with the aid of microscopes and dyes, and in the latter part of the century, psychologists had begun to speculate that training could cause the proliferation of contacts between neurons. Such speculations appeared even before the formal announcement of the 'neuron doctrine' – that is, that neurons are separate cells that can affect other neurons but do not interpenetrate them. Some likened the nervous system to a telegraph system, but after telephones entered into commercial use in the 1880s, others likened the nervous system to a telephone system where connections can be made or broken. Synaptic junctions between neurons were named only near the end of the nineteenth century. Not until the middle of the twentieth century was it accepted that transmission at most synaptic junctions is accomplished by neurochemical processes, and then began attempts to understand synaptic plasticity in terms of neurochemistry. Success of this research was soon followed by findings of plastic neurophysiological and neuroanatomical changes at synapses.

CELLULAR MECHANISMS

Hypotheses about neural bases of memory became more specific and detailed as knowledge about the nervous system increased. To understand some of these hypotheses, it will be helpful to review some essential information presented in Figure 1. A typical neuron contains a cell body, or soma, that subserves maintenance of the cell and integration of signals. There are also many dendrites, which are involved in receiving signals, and an axon along which signals are transmitted. A neuron receives multiple inputs at its soma and dendrites, and sends its output through its axon to multiple target sites. The axon terminals form connections with dendrites or cell bodies of recipient neurons through specialized structures called synapses.

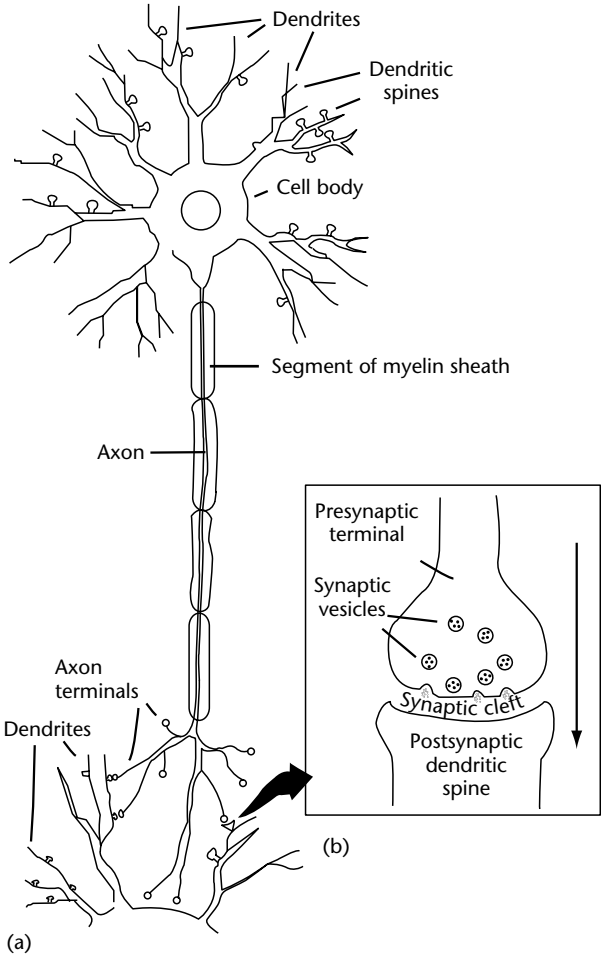


Figure 1. (a) A typical neuron, consisting of a cell body, dendrites involved in the reception of information, and an axon along which signals are transmitted. (b) A synaptic junction. When a neural impulse arrives at an axon terminal, it may lead synaptic vesicles to discharge transmitter molecules which move across the synaptic cleft to receptor sites on the postsynaptic neuron. Adapted from Rosenzweig *et al*. (1999).

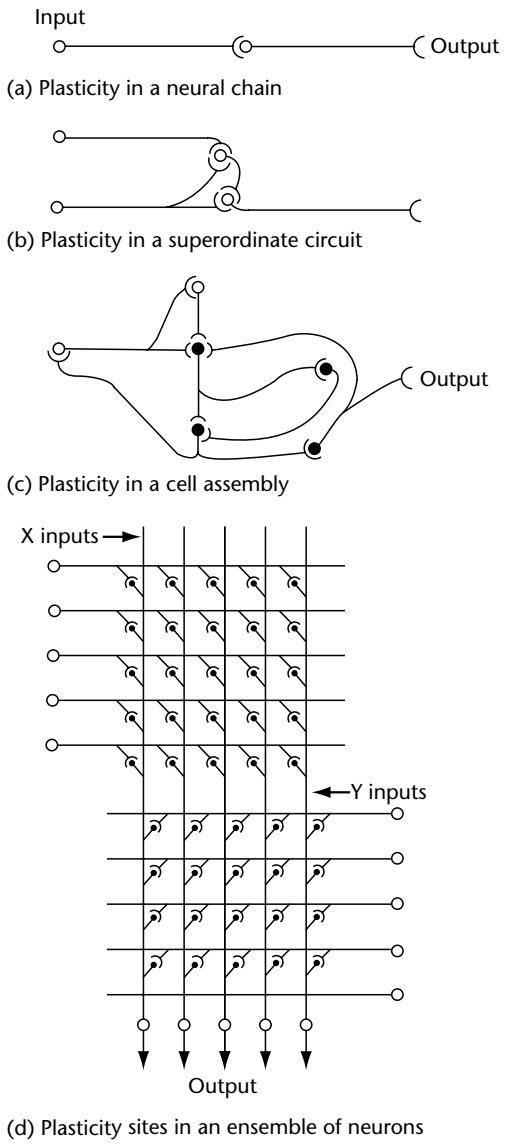


Figure 2. Some basic neural circuits; solid circles indicate sites of plasticity. (a) A neural chain with a plastic synapse. (b) Plasticity in a higher-order segment of a circuit. (c) A cell assembly with several plastic sites. (d) Many or all the synapses may be plastic in a parallel distributed circuit of neurons. Changes at plastic sites could underlie memory storage. Adapted from Rosenzweig *et al*. (1996).

The small gap (about 20 nm) between presynaptic terminals and postsynaptic sites requires signal transmission to be mediated by chemicals called neurotransmitters. There are many different chemicals that serve this purpose at different neurons, and some neurons employ more than a single kind of transmitter. Direct electrical transmission across specialized types of synapses also occurs at some sites in the nervous system.

Changes that underlie learning occur in circuits of neurons, and several kinds of circuits have been studied. Some basic circuits and sites of plasticity are illustrated in Figure 2, with plasticity in (a) a neural chain, (b) a superordinate circuit, (c) a cell assembly, and (d) an ensemble of neurons. Usually, for simplicity, plasticity is studied at a single synapse or at parallel synapses, but it should be recognized that most behavior is governed by elaborate circuits, and changes with learning may occur at several places within the circuit.

Before taking up theories of memory mechanisms, let us consider briefly some different types of memory.

TYPES OF MEMORY

Memories can be classified in a variety of ways, and studied in a variety of animals. A basic classification contrasts nonassociative with associative learning and memories. In nonassociative learning the presentation of a particular stimulus alters the strength or probability of a response. Thus, in habituation, the repeated presentation of a (usually weak) stimulus decreases the strength of the response to it. In contrast, in the kind of nonassociative learning called sensitization, presentation of a strong or painful stimulus makes the organism more responsive to most stimuli. In the sea slug *Aplysia californica*, habituation is found early in development, whereas greater maturity is required before sensitization appears. (*See* **Learning, Psychology of**; **Animal Learning**)

In associative memories an association is formed between two stimuli or between a stimulus and a response. Associative memories are frequently subdivided into declarative and nondeclarative memories. A declarative memory is one that the learner can state or describe, whereas a nondeclarative memory is shown by performance rather than by conscious recollection. One type of declarative memory is called episodic; this refers to a person's ability to recall the events of the person's own life. A different type of declarative memory is called semantic; this refers to general knowledge about facts, people, and events. Nondeclarative memories are sometime called procedural because they are responsible for motor actions such as walking, riding a bicycle or serving a tennis ball. Other kinds of nondeclarative memory include perceptual representation systems and conditioned responses. Studies involving brain lesions and/or brain imaging indicate that learning of these different classes tends to involve different brain regions. Even kinds of learning and memory that seem quite similar may involve different brain structures.

Consider, for example, two kinds of eyelid conditioning that occur when a tone (the conditioned stimulus, CS) precedes a puff of air to the eye (the unconditioned stimulus, US). After repeated pairing of the two stimuli, the CS comes to evoke the eyeblink. If the CS continues until the US begins (in what is called delay conditioning), the circuit responsible for conditioning includes structures in the cerebellum. But if the CS ceases shortly before the US starts (in trace conditioning), then the hippocampus is required.

Another classification is in terms of the duration of the memory. Investigators distinguish between short-term memory (STM), intermediate-term memory (ITM), and long-term memory (LTM). Often new information, unless it is rehearsed, fades away in less than a minute; this is called STM. Other information (such as 'Where did I park my car this morning?') may disappear in a few hours; this is called ITM. Other memories last for days or even a lifetime; this is LTM. Some investigators have found evidence that memories of these three different durations have different underlying neurochemical mechanisms. The fact that a single brief learning trial may, in some cases, be recalled minutes or hours later is often explained by the hypothesis that, for a limited time, changes in synaptic strengths or weights alter how information may flow in neural circuits. The fact that a single trial may, in some cases, lead to LTM suggests that labile STM can be transformed to ITM and then to LTM, and much research supports this hypothesis. On the other hand, some studies indicate that STM, ITM, and LTM may all be initiated at the time of learning and that they develop independently of each other, rather than sequentially.

CURRENT THEORIES OF MEMORY

It is generally accepted that the plasticity that mediates memory occurs at synaptic junctions; this could occur in several different ways, including the following.

1. After training, each nerve impulse in the relevant neural circuit causes increased release of transmitter molecules, thus increasing the size of the postsynaptic potential (PSP).
2. An increase in the number of postsynaptic receptor molecules causes a larger response to the same amount of transmitter release.
3. An increase in receptor affinity for the neurotransmitter or an increase in the mean time the receptor channel remains open when activated also causes a larger response.
4. An interneuron modulates the polarization of the axon terminal and causes release of more transmitter molecules per nerve impulse.
5. A neural circuit used more often increases the number of synaptic contacts.
6. A more frequently used neural pathway can take over synaptic sites formerly occupied by a less active competitor. (*See* **Encoding and Retrieval, Neural Basis of**; **Synaptic Plasticity, Mechanisms of**)

In the 1950s, critics noted that experimental evidence had not been obtained for any proposed neurochemical or neuroanatomical mechanisms of memory. Then, beginning in 1960, a series of papers by an interdisciplinary group at the University of California at Berkeley (Edward L. Bennett, Marian C. Diamond, David Krech, and Mark R. Rosenzweig)

showed that formal training or exposure to an enriched environment caused neurochemical and neuroanatomical changes in the cerebral cortex of laboratory rats. These findings also encouraged other researchers to investigate neural mechanisms of memory. Now evidence has been found for each of the changes listed above, varying with brain region and with kinds of learning; the different mechanisms need not be mutually exclusive. Research is now being done to find which of these changes is (or are) necessary and sufficient for learning and memory storage to occur.

As mentioned above, evidence has been found that STM, ITM, and LTM have different underlying neurochemical mechanisms. In fact, they require different stages of a neurochemical cascade. The cascade can be initiated when sensory stimulation activates sensory receptor organs, which stimulate afferent neurons by using various synaptic transmitter agents such as acetylcholine (ACh) or glutamate. Agents that inhibit ACh activity can prevent STM, and so can inhibitors of glutamate receptors, including both the *N*-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) receptors. Alteration of ion channels in neural membranes can also inhibit STM formation. Preventing formation of STM also prevents formation of ITM and LTM. If STM is formed, inhibition of some further stages in the cascade can inhibit formation of ITM or LTM. Agents that inhibit calcium–calmodulin protein kinases (CaM kinases) prevent formation of ITM and consequently of LTM. Agents that do not inhibit CaM kinases but inhibit protein kinase A or protein kinase C do not inhibit formation of ITM but prevent formation of LTM. Agents that inhibit protein synthesis, such as anisomycin, also prevent formation of LTM, while permitting formation of STM and ITM. Research with the fruit fly *Drosophila melanogaster* shows that alteration of different genes disrupts separately the formation of STM, ITM, or LTM. (*See* **Memory, Long-term**)

DISCUSSION OF ISSUES AND EVIDENCE

Many investigators have proposed what they considered to be universal cellular/molecular mechanisms for memory, in spite of the fact that they had studied only a single kind of learning in a single species. It is true that some similar mechanisms have been found by investigators studying animals as diverse as the fruit fly *Drosophila*, the sea slug *Aplysia californica*, and mammals. Nevertheless, some investigators, such as Seymour Kety, have cautioned that the proposed unity of memory processes may be an oversimplification. Kety noted that so profound and powerful an adaptation as learning and memory is not apt to rest on a single set of mechanisms but rather is likely to employ every opportunity provided by evolution. There were forms of memory before organisms developed nervous systems, and after that remarkable leap forward, it is likely that every new neural complexity, every new neurotransmitter or hormone, was incorporated into processes of learning and memory. So we must be prepared to trace out the cellular and molecular mechanisms for a variety of kinds of memory in a variety of species, noting both the similarities and the differences as they appear. (*See* **Learning in Simple Organisms**)

In the 1980s and 1990s there were disputes among those who held that synaptic plasticity occurred solely or mainly presynaptically in the axon terminals and those who held it occurred solely or mainly on the postsynaptic side of the junction. Gradually evidence accumulated for important changes on both sides of the synaptic junctions as a consequence of training, and some investigators spoke of a trans-synaptic dialogue. Evidence has also shown that, at least at some synapses, a retroactive messenger from the postsynaptic to the presynaptic junction plays an important part in establishing memory. Here again, the picture has grown more complex. Much of this research has involved the phenomenon of long-term potentiation (LTP), which is a candidate mechanism for synaptic changes in memory formation. Although some kinds of LTP show parallels with some kinds of memory formation, it has not yet been established that any of the kinds of LTP is the mechanism for any of the kinds of memory formation.

The discussion above has been limited, for simplicity, to increases in activity and excitatory processes at synaptic junctions, but memory may equally well involve decreases in activity and inhibitory processes.

CONCLUSION

The establishment of memory appears to employ a cascade of neurochemical events operating on both sides of the synaptic junction. In some cases this is complemented by structural changes in axon terminals, dendritic spines, and location of receptor molecules. It remains to be determined which of these changes may be necessary and sufficient for memory formation.

Further Reading

Draaisma D (2000) *Metaphors of Memory: A History of Ideas about the Mind*. Cambridge, UK: Cambridge University Press.

Finger S (1994) *Origins of Neuroscience: A History of Explorations into Brain Function*. New York, NY: Oxford University Press.

Hayashi Y, Shi SH, Esteban JA *et al*. (2000) Driving AMPA receptors into synapses by LTP and CaMKII: requirement for GluR1 and PDZ domain interaction. *Science* **287**: 2262–2267.

Kandel ER (2000) Learning and memory. In: Kandel ER, Schwartz JM and Jessell TM (eds) *Principles of Neural Science*, 4th edn. New York, NY: McGraw-Hill.

Kety SS (1976) Biological concommitants of affective states and their possible role in memory processes. In: Rosenzweig MR and Bennett EL (eds) *Neural Mechanisms of Learning and Memory*, pp. 321–326. Cambridge, MA: MIT Press.

Martin SJ, Grimwood PD and Morris RGM (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience* **23**: 649–711.

Rosenzweig MR (1996) Aspects of the search for neural mechanisms of memory. In: *Annual Review of Psychology* **47**: 1–32. Palo Alto, CA: Annual Reviews.

Rosenzweig MR (1998) Historical perspectives on the development of the biology of learning and memory. In: Martinez JL and Kesner RP (eds) *Neurobiology of Learning and Memory*, 3rd edn, pp. 1–33. San Diego, CA: Academic Press.

Rosenzweig MR, Breedlove SM and Leiman AL (2001) *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience*, 3rd edn. Sunderland, MA: Sinauer.

Mirror Neurons

Intermediate article

Giacomo Rizzolatti, University of Parma, Parma, Italy
Vittorio Gallese, University of Parma, Parma, Italy

CONTENTS

Mirror neurons are a class of neurons, discovered in monkey premotor cortex, that become active both when the monkey makes an action and when it observes another individual making a similar action. Similar systems are found in humans.

INTRODUCTION

A fundamental problem in cognition is how we understand actions made by other individuals. How do we know that a person sitting in front of us is eating an apple? A possibility is that the visual occipital and temporal areas provide a pictorial description of the observed scene. On the basis of it, we infer that the person is eating an apple. Another possibility is that we do not infer cognitively 'person eating apple', but understand the observed action by matching its pictorial description on an internal motor representation. Although the two alternatives are not mutually exclusive, recent neurophysiological data render the second hypothesis highly plausible.

The strongest evidence in favor of a direct observation/execution matching comes from the discovery in the monkey of a class of neurons that become active both when the monkey makes an action and when it observes another individual making a similar action. These cells are called 'mirror neurons' (Gallese *et al*., 1996; Rizzolatti *et al*., 1996). Such observation/execution matching systems exist also in humans, and preliminary evidence suggests that they too are involved in imitation.

BASIC PROPERTIES

Motor Activity

Mirror neurons are located in frontal area 5 (F5) of the premotor cortex (Figure 1). Like other neurons of this area, mirror neurons discharge in association with actions of the hand and/or mouth.

Using the effective action as the classification criterion, mirror neurons can be subdivided into various classes, such as 'grasping', 'holding', and 'tearing'. Grasping neurons are the most common; many of these are selective for a particular type of prehension, such as precision grip, finger prehension, or whole-hand prehension. Some mirror neurons discharge during the whole action they code; others are active during a phase of the action such as the opening or closure of the fingers.

Mirror neurons (and F5 neurons more generally) form a 'vocabulary' of potential motor actions. The 'words' composing this vocabulary are constituted by populations of neurons. Some of them indicate the general category of an action (hold, grasp, tear, manipulate). Others specify how the action should be made (e.g., precision grip). Finally, others are concerned with the action's temporal segmentation (Rizzolatti *et al*., 1988).

Visual Properties

Mirror neurons become active when the monkey sees an interaction between an agent and an object. Object presentation alone, even of an interesting object such as food, is ineffective. Similarly ineffective are emotional stimuli such as the sight of faces or threatening actions. The agents most effective in driving mirror neurons are natural effectors (hands, mouth). Actions made using tools typically do not activate these neurons (Figure 2). Similarly, mimicking an action is usually ineffective.

The observed hand actions effective in triggering the neurons are the same as those that excite the neurons when actively executed. More than half of mirror neurons are active during the observation of one action only (e.g., grasping). The remainder responded to two (e.g., grasping and holding) or, rarely, to three actions.

There is a large amount of generalization as far as the precise physical aspects of the effective

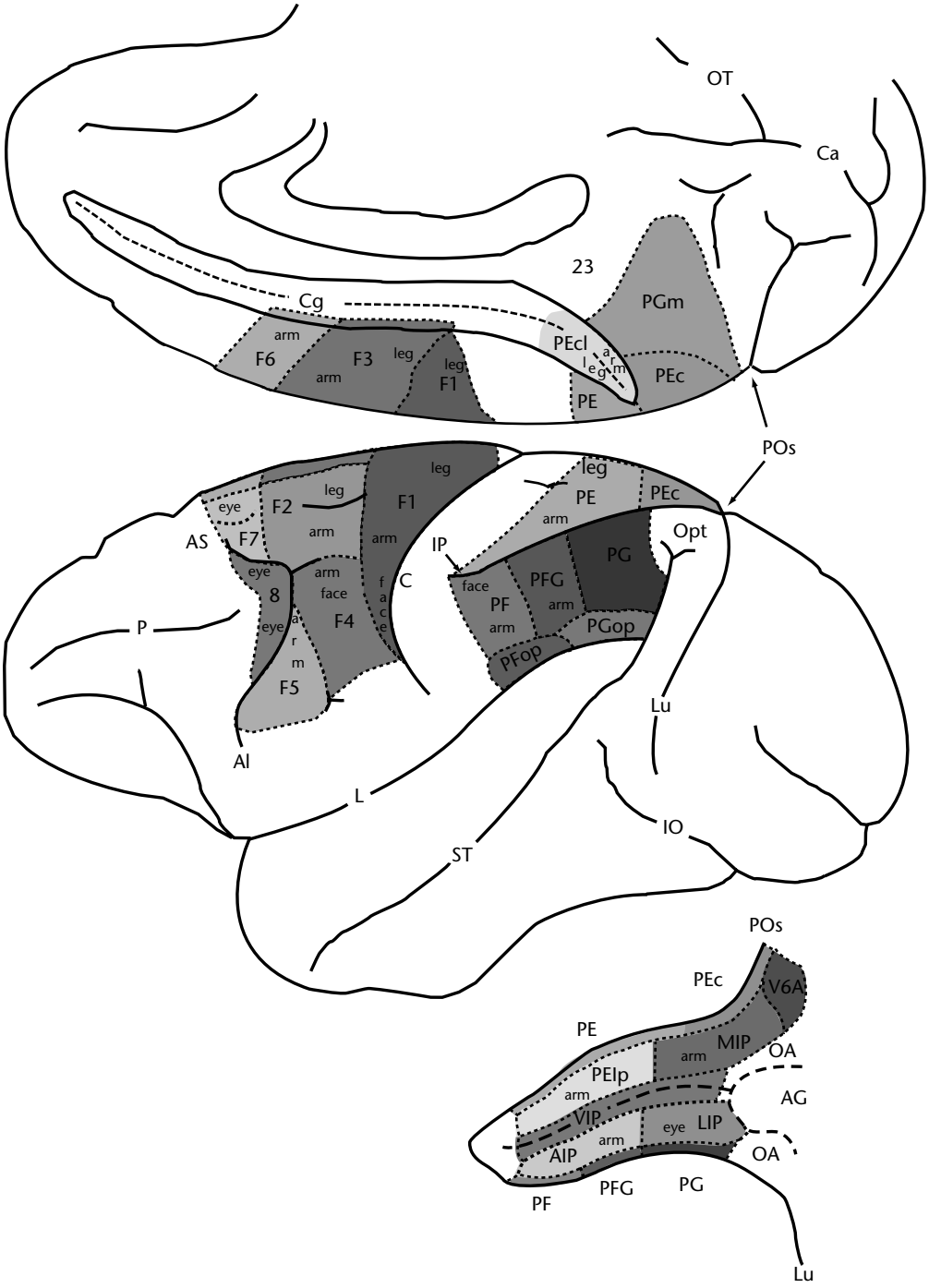


Figure 1. Mesial and lateral views of the macaque brain showing the cytoarchitectonic parcellation of the agranular frontal cortex and of the posterior parietal cortex. The areas located within the intraparietal sulcus (IP) are shown in an unfolded view of the sulcus in the lowest part of the figure. On the basis of the available data, the different body-parts representations are indicated. In the prefrontal cortex the frontal eye field FEF is also defined according to physiological criteria. AG, annectant gyrus; C, central sulcus; Ca, calcarine fissure; Cg, cingulate gyrus; IO, inferior occipital sulcus; IP, intraparietal sulcus; L, lateral fissure; Lu, lunate sulcus; OT, occipitotemporal sulcus; P, principal sulcus; POs, parietooccipital sulcus; ST, superior temporal sulcus; AI, inferior arcuate sulcus; AIP, anterior intraparietal area; AS, superior arcuate sulcus; LIP, lateral intraparietal area; MIP, medial intraparietal area; Opt, occipitoparieto-temporal area; VIP, ventral intraparietal areas.

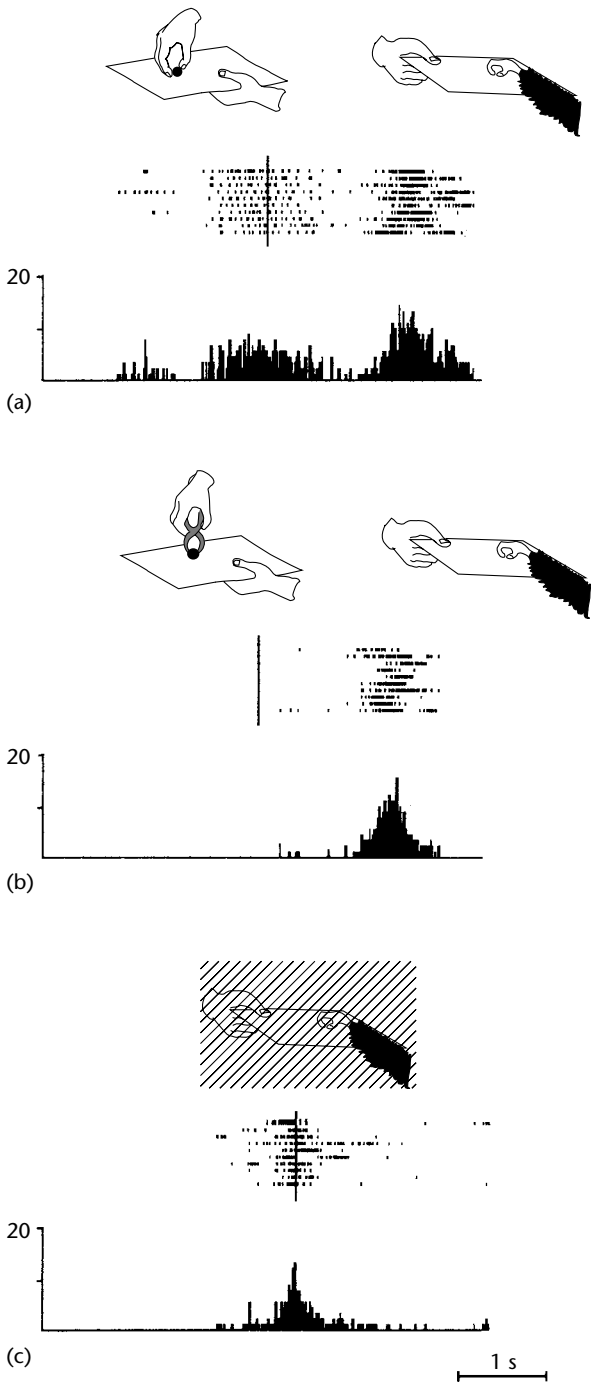


Figure 2. Visual and motor responses of a mirror neuron. Testing conditions are schematically represented above the rasters. Response histograms represent the sum of ten consecutive trials (raster display). (a) a tray with a piece of food is presented to the monkey, the experimenter grasps the food, puts the food again on the tray and then moves the tray toward the monkey who grasps the food. The phases when the food is presented and when it is moved toward the monkey are characterized by the absence of neuronal discharge. In contrast, a strong activation is present during grasping movements of both the experimenter and the monkey. (b) The same procedure is followed, except that the experimenter grasps the food with pliers. (c) The monkey grasps the food with no visual feedback. Rasters and histograms are aligned (vertical line) with the moment at which the food is touched either with the experimenter's hand (a) or with the pliers (b), and when the monkey touches the food with its hand (c). Ordinates, spikes/bin; abscissae, time; bin width, 20 ms (modified from Rizzolatti *et al*., 1996).

agent are concerned. For many neurons the precise hand orientation is not crucial for their activation. Similarly, the distance at which the action is executed does not influence the response in most cases.

Almost all mirror neurons show congruence between the observed and executed action. This congruence can be extremely strict; that is, the effective motor action (for example, precision grip) coincides with the action that, when seen, triggers the neurons (precision grip). Sometimes the congruence is broader: in these cases the motor requirements (for example, precision grip) are usually stricter than the visual ones (any type of hand grasping).

As far as the object that is the target of the observed action is concerned, its meaning does not influence the neuronal discharge. The responses to meaningful objects such as food are the same as those to any three-dimensional solid. The size of the object is relevant in about one-third of the neurons. The selectivity is related to the real size of the object, and not to its size on the retina.

ANATOMICAL ASPECTS

Figure 1 shows a cytoarchitectonic map of the macaque monkey cortex. Mirror neurons are mostly located in the F5 convexity. Visually responsive neurons are present also in the part of F5 hidden in the arcuate sulcus. These 'canonical' neurons, unlike mirror neurons, do not require an agent/object interaction, but respond to the presentation of 3D objects (Rizzolatti *et al*., 1998).

The convexity of F5 is strongly connected with area PF. Many PF neurons respond to the sight of biological actions, and some of them discharge also in association with motor actions, showing properties similar to those of F5 mirror neurons.

An important input to PF comes from the region of the superior temporal sulcus (STS). The studies of Perrett *et al*. (1989) showed that in the anterior part of STS there are a variety of neurons responsive to biological actions. Some of them encode body postures, others code specific body movements, and others discharge during goal-directed

actions. The spectrum of body parts and body movements that are coded in the temporal lobe is wide and, at variance with F5, includes arbitrary postures and movements. Although the issue was not specifically addressed, it does not appear that STS neurons discharge during active movements of the monkey (see Carey *et al*., 1997).

There is thus a rich description of biological actions in STS. Such descriptions, via PF (where some mirror neurons are present), are sent to premotor cortex where they are matched with the motor representation of the same actions. This matching is the origin of the F5 mirror neurons.

FUNCTIONAL ROLE

Although it is likely that the human matching systems use sets of neurons similar to mirror neurons, their functional properties may not necessarily be identical to those of F5 mirror neurons. Two explanations for the activity of F5 mirror neurons come immediately to mind. The first is that their firing depends on abortive movements passed unnoticed to the experimenter; the second is that their activity is preparatory to an impending movement. Controlled experiments have ruled out both these hypotheses. Mirror neurons discharge in response to the appropriate stimulus both when the monkey is still and when it performs movements that have no relation to the observed action. Furthermore, mirror neurons do not fire when an object is moved toward the monkey and made more available for action on it – the opposite of what would be predicted by the motor preparation hypothesis. Discarding these possibilities, the most likely interpretation of mirror neurons is that their activity generates an internal representation of the observed action. This interpretation fits the general notion that motor representations are elicited in motor cortex, in a potential form, by a variety of stimuli. These representations become actions only if specific contingencies render their transformation into action useful.

Potential actions are frequently generated by the sight of an object. In the case of mirror neurons they are generated by the sight of an action. The problem now is to elucidate the functional role of a motor representation generated by the sight of an action. The most likely possibility is that this representation is used for making sense of the observed action. The assumption here (difficult to deny) is that when making (or preparing) an action, the individual predicts ('knows') its consequences. Mirror neurons extend this knowledge to actions performed by others. When a person observes an action performed by another individual, mirror neurons that represent that action become active. The action is hence recognized because it corresponds to that generated during action programming. Might this mechanism lead to confusion between self-generated actions and actions made by others? Hardly so. Signals preceding action initiation as well as proprioceptive signals following movement onset can easily discriminate (at least in normal subjects) self-generated action from action made by others.

Another possibility is imitation. Monkeys, however, do not imitate – or at least do not imitate manual gestures made by another individual. Thus, although endowed with a mechanism that generates internal copies of actions made by others, they are unable to use them for replicating those actions. The intentional use of internal copies of actions appears to have developed only later in evolution.

OBSERVATION/EXECUTION MATCHING SYSTEMS IN HUMANS

Seeing another person making movements activates motor areas in our brain. This phenomenon has been demonstrated in humans by electrophysiological studies (e.g., Hari *et al*., 1998) as well as by transcranial magnetic stimulation. This latter technique showed that the increase of motor evoked potentials during the observation of hand movements is highly specific. It involves those muscles of the observer that the actor is using. Motor evoked potentials increase also during the observation of intransitive (nongoal-directed), meaningless movements (Fadiga *et al*., 1995).

The localization of areas active during action observation was made using brain imaging techniques. Initial studies showed that during the observation of different types of grips performed on a variety of objects there is an activation of the left STS region, the inferior parietal lobule, and Broca's area. This circuit corresponds to that of mirror neurons in the monkey (see Grafton *et al*., 1996).

In a more recent study using functional magnetic resonance imaging, participants were shown actions made with mouth, hand and foot, some of which were object-directed while others were mimed, no object being present. Both object-related and nonobject-related actions determined a somatotopically organized activation of premotor cortex. During mouth actions there was a bilateral activation of the ventral premotor cortex plus an activation of Broca's area. During hand actions a more dorsal part of ventral area 6 plus Broca's area

were recruited. Finally, the observation of foot actions elicited an activation of dorsal area 6 (Binkofski *et al*., 2001). In addition, a difference in activation was found between observations of object-related and nonobject-related actions. When an object is the target of an action the parietal lobe is strongly activated. Parietal activations are also somatotopically organized and depend on the effector used.

These results indicate that when individuals observe an action, a replica of that action is automatically generated in their cortex, recruiting the same circuits that are active when the observed action is internally generated by the observer. The presence of a strong parietal activation in the case of object-related actions is probably due to the fact that the observer makes a 'pragmatic' analysis of the objects similar to that necessary in order to act on the object.

While in the experiments described above the only instruction to participants was to carefully observe the action, in subsequent experiments participants were instructed to observe the actions in order to imitate them either after or during scanning. The most important finding was the observation (in addition to premotor activation) of an activation of the right superior parietal lobe (Grèzes *et al*., 1998; Iacoboni *et al*., 1999), probably due to the need to make an internal description of the precise movements made by the actor, not just to understand what the actor is doing. This requires activation of the right superior parietal lobe, where there is a proprioceptive description of movements.

Studies in humans cannot specify the precise mechanisms that match observed and executed actions. They showed, however, that in humans a mechanism similar to that of the monkey is present during action observation. In addition, in humans observation of intransitive meaningless movements activates the motor areas. Finally, and most importantly, direct matching is used in humans not only for action understanding, but also for imitation. It appears, therefore, that other cognitive functions have developed out of an evolutionarily older mechanism.

References

Buccino G, Binkofski F, Fink GR *et al*. (2001) Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* **13**: 400–404.

Carey DP, Perrett DI and Oram MW (1997) Recognizing, understanding and reproducing action. In: Boller F and Grafman J (eds), *Handbook of Neuropsychology*, vol. XI, pp. 111–129. Amsterdam, Netherlands: Elsevier.

Fadiga L, Fogassi L, Pavesi G and Rizzolatti G (1995) Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* **73**: 2608–2611.

Gallese V, Fadiga L, Fogassi L and Rizzolatti G (1996) Action recognition in the premotor cortex. *Brain* **119**: 593–609.

Grafton ST, Arbib MA, Fadiga L and Rizzolatti G (1996) Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Experimental Brain Research* **112**: 103–111.

Grèzes J, Costes N and Decety J (1998) Top-down effect of strategy on the perception of human biological motion: a PET investigation. *Cognitive Neuropsychology* **15**: 553–582.

Hari R, Forss N, Avikainen S *et al*. (1998) Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences of the USA* **95**: 15061–15065.

Iacoboni M, Woods RP, Brass M *et al*. (1999) Cortical mechanisms of human imitation. *Science* **286**: 2526–2528.

Perrett DI, Harries MH, Bevan R *et al*. (1989) Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology* **146**: 87–113.

Rizzolatti G, Camarda R, Fogassi L *et al*. (1988) Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Experimental Brain Research* **71**: 491–507.

Rizzolatti G, Fadiga L, Gallese V and Fogassi L (1996) Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* **3**: 131–141.

Rizzolatti G, Luppino G and Matelli M (1998) The organization of the cortical motor system: new concepts. *Electroencephalography and Clinical Neurophysiology* **106**: 283–296.

Further Reading

Gallese V and Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* **12**: 493–501.

Jeannerod M (1994) The representing brain: neural correlates of motor intention and imagery. *Behavioral Brain Science* **17**: 187–245.

Jeannerod M, Arbib MA, Rizzolatti G and Sakata H (1995) Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neuroscience* **18**: 314–320.

Jellema T and Perrett DI (2002) Coding of visible and hidden actions. In: Prinz W and Meltzoff A (eds) *The Imitative Mind: Development, Evolution, and Brain Bases*, pp. 356–380. Cambridge, UK: Cambridge University Press.

Rizzolatti G and Arbib MA (1998) Language within our grasp. *Trends in Neuroscience* **21**: 188–194.

Rizzolatti G, Fogassi L and Gallese V (2000) Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, 2nd edn, pp. 539–552. Cambridge, MA: MIT Press.

Rizzolatti G, Fogassi L and Gallese V (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* **2**: 661–670.

Rizzolatti G, Fadiga L, Fogassi L and Gallese V (2002) From mirror neurons to imitation: facts and speculations. In: Prinz W and Meltzoff A (eds) *The Imitative Mind: Development, Evolution, and Brain Bases*, pp. 247–266. Cambridge, UK: Cambridge University Press.

Modularity in Neural Systems and Localization of Function

Introductory article

Carlo Umiltà, University of Padua, Padua, Italy

CONTENTS

Modularity assigns particular cognitive functions to particular structures (modules) in the brain. These functional units are viewed as operating relatively independently from one another. This view contrasts with the hypothesis that such functions correspond to the highly interactive operations of many brain components.

INTRODUCTION

The dispute between 'modularists' (or 'localizationists') and their 'distributionist' opponents goes back to the dawn of neuroscience. Modularists assign particular cognitive functions to separate 'organs' or 'modules' in the brain, while distributionists see such functions as corresponding to the interactive operations of many brain components. Modularists see the brain's functional units as operating relatively independently from one another; distributionists regard the brain's components as highly interactive.

In the Middle Ages, mental faculties were thought to be localized in the brain's ventricles. The faculties of the mind, derived from the theories of Aristotle, were distributed among the ventricles, as described by Galen. One ventricle was thought to receive input from all the sense organs and to be the site of the *sensus communis*, or common sense, which integrated across modalities. Fantasy and imagination were located there too. The second ventricle was considered to be the site of cognitive processes: reasoning, judgment and thought. The third ventricle was considered to be the site of memory. These ideas remained prevalent until the sixteenth century, when Vesalius attacked the ventricular doctrine of brain functions. In the eighteenth century, von Haller maintained that all parts of the brain were functionally equivalent.

The localization of different psychological functions in different regions of the brain (of the cerebral cortex, to be more precise) began with Gall, in the early nineteenth century. The central ideas of his phrenological system were that the brain is a machine for producing cognition and emotion, and that it is composed of a set of organs with different functions. Gall and his collaborator Spurzheim postulated nearly forty affective and intellectual faculties and assumed that these were localized in specific organs of the cerebral cortex. Gall argued that 'the brain is composed of as many individual and independent organs as there are forces of the soul.'

The most influential critique of Gall came from Flourens in the mid-nineteenth century. Flourens reported that lesions of the cerebral cortex had devastating effects on mental functions. However, the site of the lesion was irrelevant, demonstrating that all cortical areas contributed equally to these functions. In Flourens' view, 'there are not … different seats for different faculties, nor for different sensations'.

Despite their temporary eclipse under the influence of Flourens' experiments, Gall's notions of punctuate localization stimulated the search for relations between the site of the brain injury and specific psychological deficits in patients as well as in experimental animals. The second half of the nineteenth century thus saw an intense search for the localization of sensory centers in the brain. Gall's general ideas were considered to be confirmed by Broca's demonstration of an association between damage to the left frontal lobe and aphasia, and, later, by Fritsch and Hitzig's experiments on the stimulation of the motor cortex. In 1861 Broca presented the case of a patient with severe difficulties in language production (aphasia) and showed that his brain had a lesion in the left frontal lobe. This case corroborated the idea that the human mind could be subdivided into specific functions and that these specific functions were mediated by discrete brain structures. It should be kept in mind, though, that the notion of brain

equipotentiality (or holism) survived, through Charcot and Freud, until Lashley in the 1950s.

THE CONCEPT OF MODULARITY

In the 1970s the concept of modularity became central to cognitive science. Marr thought that this concept was so important as to be elevated 'to a principle, the principle of modular design', according to which 'any large computation should be split up into a collection of small, nearly independent specialized sub-processes'. There is certainly abundant evidence for modularity in the sense of specialized functions. In particular, it is a well-established fact that different brain areas are dedicated to representing information from specific sensory and motor channels. Undoubtedly, the visual system has a modular organization, in which representations of color, shape, motion and location are separable physiologically, anatomically and psychologically at very early stages of information processing. However, one must distinguish between two different meanings of the term 'module'. This word is often used in a general sense to denote any mechanism with a specialized function; in this sense, it actually means 'component or component process that performs a specific computation'. Posner and Shallice among others endorsed this weaker and nonspecific version of modularity, according to which components of the functional architecture, that is 'isolable subsystems', can be distinguished in terms of their specialized functions. For example, according to the modular model of McCloskey, basic number processing consists of three functionally distinct systems specialized for number comprehension, calculation and production. In turn, the calculation system includes memory subsystems for basic numerical facts, rules and procedures. (*See* **Binding Problem**; **Acalculia**; **What and Where/How Systems**)

There is a stronger, more specific, and perhaps more interesting meaning for the term 'module'. It was Fodor who first explicitly discussed what should be meant by 'cognitive module'. In his view, a cognitive module accepts representations as input, transforms them through a set of specialized computations, and transmits the transformed representations onward. Thus, the output of one module is the input to the next module down the line. According to Fodor, cognitive modules are distinguished from other computational mechanisms by strict criteria. Modules are domain-specific, innately specified, informationally encapsulated, hard-wired, autonomous and not assembled. They process information automatically (i.e. their operations are mandatory) and produce a shallow output. They are associated with a fixed neural architecture, when lesioned they manifest a specific breakdown pattern, and, during development they manifest a specific pace and sequencing.

According to Moscovitch and Umiltà, the most important criteria of modularity are domain specificity, informational encapsulation (or cognitive impenetrability) and shallow output. Domain specificity indicates that each module processes information from only a restricted domain, that is, it only responds to stimuli of a particular class (e.g. human faces). Informational encapsulation indicates that the computations that take place within a module are not affected by computations that take place within other modules or in higher-order central systems. Thus, modules are cognitively impenetrable and deliver only a shallow output to other modules and to central systems. An output from a module is considered to be shallow when it is not semantically interpreted. The interpretation of the output is the responsibility of central systems that relate it to general knowledge (i.e. to semantic memory).

Different modules do not interact with one another, except when one has completed its processing, at which point it makes the end product available to other components downstream. That is, a module carries out its computations without being affected by other information available in other modules or in central systems. These characteristics of modules ensure that the computations they perform are not distorted by beliefs, motivations and expectancies. Modules, therefore, act as special-purpose devices, especially suited to picking up information from their restricted domains, processing it efficiently and automatically, and delivering precise – but narrow and presemantic – information to other modules or to central systems for semantic interpretation. As Fodor put it, they are 'stupid' but efficient mechanisms that are necessary for 'representing the world veridically, and making it accessible to thought'.

Moscovitch and Umiltà have argued that modules may be decomposed into a set of smaller modules and this decomposition stops when a level is reached at which all the components are primitive processors. In this view, there can be higher-order systems that have an internal modular structure. These assembled modules, made up of more basic components, do not violate the most important criteria of modularity: they perform specialized functions, do not interact with other modules, are not influenced by semantic systems,

and their outputs are not semantically interpreted – i.e. they are domain-specific, informationally encapsulated and deliver shallow outputs.

AN ALTERNATIVE APPROACH

The parallel distributed processing (PDP) framework is the most recent example of the alternative, distributionist view. The distributed, interactive, graded processing in PDP systems contrasts sharply with the encapsulated, staged, all-or-none processing in modular systems.

According to Farah, the three essential features of PDP systems are distributed representations, gradedness and interactivity. In PDP systems representations consist of patterns of activations distributed over a population of processing units. Different inputs can therefore be represented in the same set of units, because the patterns of activations over the units is distinctive. Knowledge (i.e. semantic memory) is distributed too, because knowledge is encoded in the pattern of connection strengths (weights, as opposed to activations) distributed over a population of units. In PDP systems processing is graded rather than all-or-none. Representations can be partially active. Similarly, partial knowledge can be embodied in connection strengths. The units in PDP systems are highly interconnected and thus mutual influence among different components of the system occurs. (*See* **Backpropagation**; **Hebb Synapses: Modeling of Neuronal Selectivity**; **Synapse**)

The PDP approach therefore assumes that human information processing is graded, distributed and interactive. The representation that is being processed is distributed over a number of units of the network, rather than being localized in a few units. The knowledge involved in a process is distributed throughout a number of connectionist weights of the network, rather than being localized in particular components. Processing is graded and continuous rather than being staged and all-or-none, making partial results in one part of the network continually available to other parts. Groups of units representing different types of information are highly interactive, instead of receiving inputs from only a few other components.

RECONCILING THE TWO APPROACHES

It is unnecessary to represent modular models of cognitive processes as being in conflict with connectionist models. Behind this putative dispute there is a conflation of levels of analysis. If modular models and PDP models are viewed as descriptive of the same processes but at different levels of analysis, the dispute dissolves. This becomes clear if one considers Marr's division of levels of analysis. Functions that need to be computed in a complex cognitive process are specified at the computational level. How those functions are to be computed is specified at the algorithmic level. How the algorithms are physically realized is specified at the implementational level. Cognitive processes can be modular at the computational level, whereas at the implementational level they are distributed and interactive. (*See* **Computational Models: Why Build Them?**; **Computer Modeling of Cognition: Levels of Analysis**; **Connectionism**)

A similar way of reconciling the modular and the PDP approaches is to consider both the macrostructure and the microstructure of a cognitive process. Modular models concern the macrostructure: that is, the level of description that identifies the components of the functional architecture. In contrast, PDP models concern the microstructure: that is, the nature of the operations that go on within the architectural components. Thus, in explaining cognitive process, one should first derive the macrostructural model, in which the relevant components of the functional architecture are identified, and then investigate the microstructure of each component's internal operations. Modular models say very little about the underlying computational mechanisms, which instead are addressed in details by PDP models.

CONCLUSION

The modularity hypothesis assumes that the human brain possesses a functional architecture made up of cognitive modules, that is, of a set of relatively independent information-processing computational mechanisms. A module accepts representations as input, transforms them through specialized computations, and transmits the transformed representations onward. Each module carries out its computations without being affected by information available in other modules. Modules interact only when one has completed its computations, at which point it makes the end product available to a relatively small number of other processing systems.

In contrast, the PDP hypothesis assumes that human information processing is graded in nature, highly interactive, and characterized by distributed representation of knowledge. The two hypotheses may be reconciled by proposing that modular

models identify the components of the functional architecture of the brain, whereas PDP models concerns the operations that are performed within each component.

Further Reading

Block N (1995) The mind as the software of the brain. In: Smith EE and Osherson DN (eds) *Thinking: An Invitation to Cognitive Science*, pp. 21–63. Cambridge, MA: Bradford Books/MIT Press.

Coltheart M (1999) Modularity and cognition. *Trends in Cognitive Sciences* **3**: 115–120.

Farah MJ (1994) Neuropsychological inference with an interactive brain: a critique of the 'locality' assumption. *Behavioral and Brain Sciences* **17**: 43–104.

Fodor JA (1983) *The Modularity of Mind*. Cambridge, MA: Bradford Books/MIT Press.

Gross CG (1998) *Brain, Vision, Memory: Tales in the History of Neuroscience*. Cambridge, MA: Bradford Books/MIT Press.

Marr D (1982) *Vision*. San Francisco: Freeman.

Moscovitch M and Umiltà C (1990) Modularity and neuropsychology: modules and central processes in attention and memory. In: Schwartz MF (ed.) *Modular Deficits in Alzheimer-type Dementia*, pp. 1–59. Cambridge, MA: Bradford Books/MIT Press.

Posner MI (1978) *Chronometric Explorations of Mind*. Hillsdale, NJ: Erlbaum.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, Foundations. Cambridge, MA: Bradford Books/MIT Press.

Shallice T (1988) *From Neuropsychology to Mental Structure*. Cambridge, UK: Cambridge University Press.

Motion Perception, Neural Basis of

Introductory article

Norberto M Grzywacz, University of Southern California, Los Angeles, California, USA
David K Merwine, University of Southern California, Los Angeles, California, USA

CONTENTS

Animals perceive motion by extracting velocity information from their visual inputs. Differing aspects of the information are computed in a hierarchical series of sequential stages from the retina through the temporal cortex.

INTRODUCTION

For most animals, the ability to perceive motion is vitally important. For example, a predator must have the ability to track moving prey in order to hunt and survive, whereas prey must detect the smallest movement of a potential predator. Another example is that of locating and communicating with a potential mate, which often involves motion detection. Self-navigation is also critically dependent on one's perception of motion relative to surrounding objects. It is therefore not surprising that the ability to perceive motion appears throughout evolution. Animals as simple as the fly contain motion detectors within their visual systems. Most vertebrates are able not only to detect motion, but also to extract its parameters. This article will focus on vertebrate vision. (*See* **Motion Perception, Psychology of**)

According to Newton, in order to know the motion of a point (or particle), one needs to measure three things, namely direction, speed and acceleration. The first computation related to motion direction occurs within the eye, and this information is sent to both cortical and subcortical areas of the brain. Within the visual cortex, motion direction is independently determined again from non-directional inputs, and is subsequently refined through a succession of cortical stages. However, retinal and most cortical neurons are not speed selective. Some individual neurons in higher cortical areas, such as the middle temporal cortical area (MT), appear to be speed selective. However, it is generally believed that speed is determined by the combined responses of several neurons. Biological systems are extremely good at determining the direction of a motion, and are quite good at determining its speed. However, they are poor at determining acceleration, an issue that we shall not explore further here.

One of the most fundamental sources of visual motion is an animal's own movement, namely egomotion. The overall motion of the visual field that results from egomotion is called *optic flow*. For example, if one moves forward in a straight line, the visual flow will spread out from the center-of-heading (Figure 1a), a type of optic flow that is termed *expansion*. In contrast, self-movement backwards results in *contraction* of the optic flow. Sideways movements yield *translation* (Figure 1b), and tilting the head sideways yields *rotation* (Figure 1c). Figures 1a and c illustrate how despite being globally coherent, local-motion directions can be very different and even opposite. Such visual inputs are termed *complex motions*. Many neurons at the higher stages of the motion-processing pathway selectively respond to these types of complex motions.

Motion perception is computed hierarchically – that is, in successive stages within the brain. This article will follow the hierarchy. The computation begins in the retina with a local computation related to the direction of object movement. In the first visual cortical center, namely the primary visual cortex (V1), motion direction is computed again, independently and somewhat differently. From there, a subset of cells sends information to the MT, and from the latter to the middle superior temporal cortex (MST), and so on. The response properties of the neurons along this path are progressively refined. Thus cells in the higher cortical areas respond specifically to complex aspects of

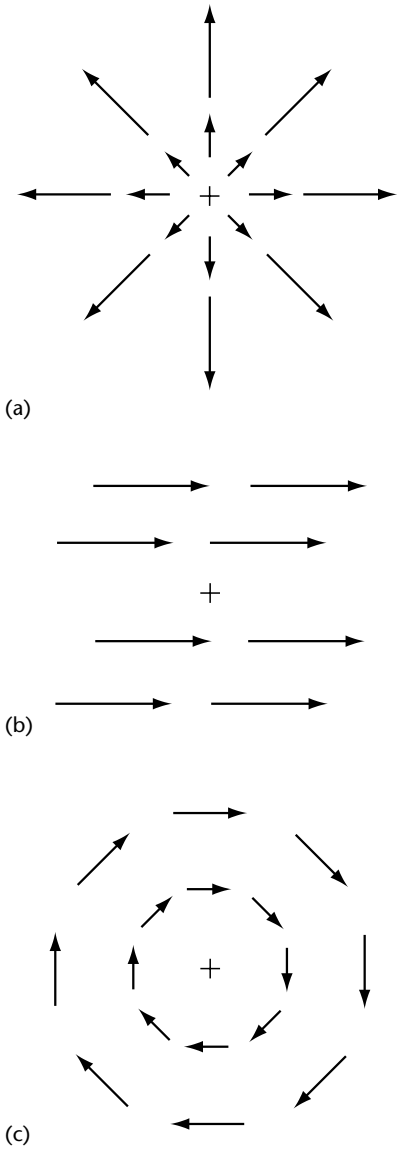


Figure 1. Examples of optic flow. The arrows indicate the direction and speed of motion at each point in space. (a) Expansion. (b) Translation. (c) Rotation.

motion, such as expansion of the optic flow or object rotation in three-dimensional space. (*See* **Modularity in Neural Systems and Localization of Function**)

DIRECTIONAL SELECTIVITY IN THE RETINA

Basic Description

The electrophysiological study of retinal information processing began with H. Keffer Hartline, who later won the Nobel Prize for his research. He discovered that ganglion cells (the retinal output) would alter their firing rates in response to local changes in brightness. The area in which these luminance changes could influence the firing of a cell was termed the receptive field (RF). Horace B. Barlow subsequently made the first reports of motion-selective ganglion cell responses in frog retinas in 1953. He found that there were many cells in the frog retina that would fire continuously so long as a visual object moved within the cell's RF. (Interestingly, when the target neurons of a motion-detecting neuron were electrically stimulated, the frog would snap at the corresponding point in space. The motion-detecting neurons were therefore termed 'bug detectors'. This was one of the first demonstrations of a link between neural activity and behavior.) Later, together with Richard M. Hill and William R. Levick, Barlow recorded the first directionally selective (DS) motion responses in a mammalian retina. (*See* **Single Neuron Recording**; **Receptive Fields**)

Barlow and his colleagues found two types of DS cells in the rabbit retina. The commonest one is the On–Off DS cell. An On–Off DS cell will respond weakly to a small spot that is flashed on or off anywhere within its RF. However, it will fire strongly if a spot (light or dark) is moved through its RF in an appropriate, 'preferred' direction (Figure 2). Motions in the opposite or 'null' direction yield essentially no response, while orthogonal motions yield intermediate responses. The On–Off DS cells are divided according to their preferred directions into up, down, left and right subgroups, each of which independently tiles the retinal surface. The other DS cell type is the on DS cell. These cells will only respond to bright objects, and they prefer larger objects and slower speeds than do the

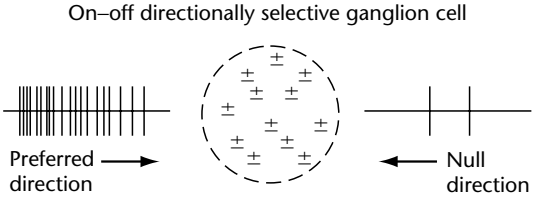


Figure 2. On–Off DS ganglion cell responses. The rabbit retinal on–off DS cell will respond to both the onset and offset of a small spot flashed within its RF (the area enclosed by the broken circle), as indicated by the ± symbol. This cell will respond vigorously to a spot, slit or edge moved in its preferred direction (to the right in this example), but will respond poorly to motions in the opposite or null direction. Upward or downward motions elicit intermediate responses. In this and subsequent figures, the vertical lines denote spikes.

On–Off DS cells. There are three subgroups of on DS cells, whose preferred directions of motion align with the semicircular canals of the vestibular (balance) system.

Species Comparison

The retinas of mammals, birds, reptiles and amphibians have all been shown to contain DS cells (fish are the only major vertebrate class in which DS cells have rarely been reported). However, as with the two rabbit DS cell types described briefly above, there are many differences in DS cell properties between species. Preferences for light versus dark objects, slow versus fast speeds, sensitivity to object shape and level of dendritic ramification all vary according to the species. Even the percentage of DS cells in the retinal population varies widely. Directionally selective cells are reported to represent between 2% (in cats) and 40% (in turtles) of the retinal ganglion cell population. These percentage differences largely reflect the differences in neocortex available to these species for processing visual information. Thus species with little or no neocortex devote many retinal neurons to computing directional selectivity, while those with large neocortices do not. Nevertheless, despite differences in detail, there is a highly conserved plan among vertebrate species with regard to the use of DS cell output. For all species this information is sent to the accessory optic and pretectal brainstem nuclei. These structures participate in the vestibulo-ocular and optokinetic systems which stabilize the eyes during rotatory head movements, or during rapid global image movements on the retina. Thus these systems are crucial for determining whether image motions on the retina are due to eye, head or body movements or reflect the movement of external objects.

Detailed Response Properties of On–Off DS Cells

Many studies have been undertaken to elucidate further the properties of the rabbit retinal On–Off DS cell, the aim being to understand the mechanisms responsible for its behavior. For example, it has long been known that an object does not have to pass through the entire RF of an On–Off DS cell in order to generate a DS response (the opposite applies to simple cortical DS cells). Anatomically, the dendritic trees of the On–Off DS cell contain many 'loops', which are consistently around 40–50 μm in diameter. Therefore it was proposed that these cells contained multiple 40–50 μm subunits, each of which performed the DS computation independently. This proposal was consistent with earlier measurements by Barlow and Levick, who attempted to determine the shortest motions that produce DS responses. However, it was recently demonstrated that these cells can discriminate direction of motion for movements as small as 26″ of visual angle (*c.* 1 μm on the retinal surface!). This result severely limits the possible mechanisms for computing directional selectivity, implying that they act very locally, they probably involve only a few synapses, and they exist essentially everywhere within the cell's dendritic tree (or those of its inputs). At the time of writing, the function of the dendritic loops remains unclear. (*See* **Neurons, Computation in**)

Another interesting feature of these cells' responses is that their directional selectivity is invariant to other aspects of a stimulus. For example, consider speed. It is true that an On–Off DS cell will respond best to motions within a narrow speed range. Nonetheless, this cell will respond significantly better to a preferred-direction motion than to a null-direction motion, regardless of the speed. The same is true for temporal frequency. Responses to moving sine- and square-wave gratings may be better for certain temporal frequencies than for others, but the cell always responds better to a preferred-direction stimulus, regardless of the temporal frequency. Finally, imagine that one presents an On–Off DS cell with two simultaneous, nonparallel, drifting sine- or square-wave gratings (so that their overlap forms a 'plaid', as shown in Figure 3). In this case, the cell will respond best when the 'plaid' motion is in the preferred direction. In other words, retinal DS responses are independent of the orientation of the gratings. (This is similar to pattern cells higher in the cortical motion pathway; see below.) In essence, if the cell can 'see' the motion stimulus, it will respond to it in a directionally selective manner. Thus it seems that this cell sacrifices information regarding the exact location, speed, size and shape of stimuli, so that

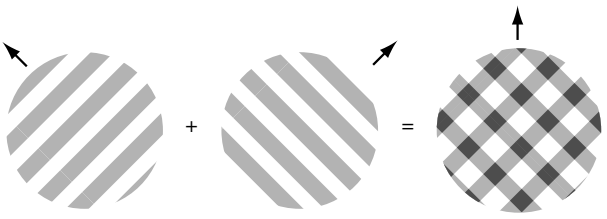


Figure 3. Plaid motion. Two moving gratings are superimposed to form a moving plaid.

it may robustly indicate their motion direction. The 'lost' information is encoded along the cortical motion pathway.

DIRECTIONAL SELECTIVITY IN THE CORTEX

The Motion Hierarchy

Inputs from the retina pass through the thalamus on their way to the primary visual cortex, also known in primates as V1 (Visual Area 1). From V1, visual information is split along two primary pathways. One path extends into the temporal cortex and contributes to the recognition of objects. The other path, which we shall discuss here, is extended into the parietal cortex and is used to analyze visual motion. This pathway is hierarchical – that is, the information is processed in a series of stages. At each stage, the information is progressively refined. Thus cells at the beginning of the hierarchy respond only when presented with an appropriately oriented, moving edge in a well-defined spatial location. In contrast, cells at higher stages in the hierarchy respond to complex motions, regardless of spatial location. However, it should be noted that the computation of motion is not simply sequential. There is crosstalk with other cortical as well as subcortical areas, and there are back-projections to earlier stages along the hierarchy. Recent experiments suggest that these connections may be crucial for attention, visual awareness and image segmentation (the parsing of the image into regions of relatively homogeneous properties). (*See* **Parietal Cortex**)

The parietal-motion pathway begins with signals from directionally selective V1 cells. These cells project to the middle temporal cortex (MT or V5) and to V2, whose neurons also project to the MT. From the MT, the motion pathway proceeds to the middle superior temporal cortex (MST). In turn, MST cells project to the lateral and ventral intraparietal areas (LIP and VIP), as well as to V7a. These latter stages in the pathway also carry information for motor planning and control, which are among the main goals of the motion-processing system.

V1 Cells

The first DS cells along the cortical path appear in the input layers of V1. Nearly all of the so-called 'simple' cells there show some directional selectivity. Indeed, nearly all neurons in V1 are directionally selective to some degree. However, this selectivity is weaker than that in the retina. Moreover, the simple-cell selectivity is dependent on the size, location, spatial frequency and orientation of the visual stimulus. A simple cell contains separate on and off subregions and responds weakly to a spot flashed in these locations. The best response of this cell occurs when an oriented edge passes entirely through the cell's RF in its preferred direction. As with retinal DS cells, the preferred direction of motion cannot be explained by any spatial organization of the cell's dendritic tree. However, it is often possible to predict the preferred direction of motion for a simple cell by examining its response to flashed spots. Thus directional selectivity in these cells has been explained using a simple additive scheme, which we shall discuss in the section on models and mechanisms of directional selectivity below.

The simple cells project to the upper and lower layers of V1 where 'complex' DS cells are found. These cells do not respond to flashed spots, nor do they have separate on and off subregions. Like simple cells, complex cells prefer oriented edges moving in a particular direction. However, it is no longer necessary for an object to pass through the complex cells' entire RF in order to generate a DS response. It is still unclear to what extent (if at all) the directional selectivity of simple cells contributes to that of complex cells.

MT Cells

A subset of DS V1 cells projects to the MT. Cells in the MT have large RFs, often 10 times larger than those of cells in V1. MT cells will respond to an object moving in their preferred direction anywhere within their RF. In addition, J. Anthony Movshon and colleagues found that approximately one-third of the cells in the MT (called pattern cells) could detect the motion of a plaid in a DS manner. That is, suppose that the cells are presented with two overlapping sine waves moving in different directions (Figure 3). In this case, these cells will respond best when the composite motion of the sine waves, not their individual motions, is in the preferred direction. This behavior is illustrated in Figure 4. In contrast, the component MT cells (the other two-thirds) behave similarly to cells earlier in the cortical hierarchy – that is, they respond whenever either of the two gratings is moving in the preferred direction of the cell. Not surprisingly, human subjects see the coherent plaid motion under many stimulus conditions (i.e. subjects often see the motion 'reported' by the pattern cells). However, under some conditions (e.g. when the sine waves have

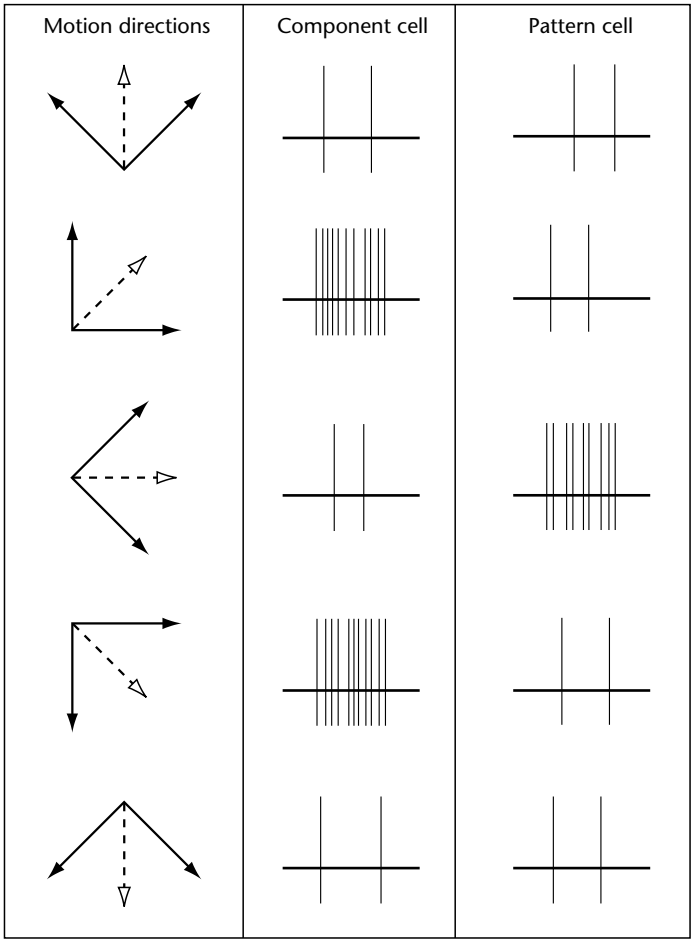


Figure 4. Responses of component and pattern MT cells. The left-hand column shows motion stimuli, while the center and right-hand columns show the responses of component and pattern MT cells, respectively. For the motion stimuli, the solid arrows indicate the motion directions of two identical sine-wave gratings. The broken, open arrows show the motion of the plaid formed by the combination of the gratings (Figure 3). A component cell will respond whenever either grating is moving in the preferred direction (to the right for both cells in this figure), as shown in the second and fourth rows of the second column. This cell does not respond when the plaid motion is in the preferred direction (third row of second column). A pattern cell, on the other hand, will not respond to the individual gratings, but will respond to the plaid motion.

very different contrasts and spatial frequencies), subjects see the sine waves sliding past each other. This percept is termed transparency.

MST Cells

Cells in the MT primarily send their outputs to the middle superior temporal cortex (MST). Here DS information from several cells converges to create neurons with very large RFs. Neurons in the MST are the first motion neurons in which RFs are bilateral – that is, they extend across the visual midline. Some neurons in the dorsal region of the MST have RFs that cover most of the visual field. These cells combine DS inputs in such a way that they are sensitive to complex visual motions, such as contraction, expansion or rotation of the optic flow. A subregion of the MST is also devoted to encoding translational motions. Interestingly, the cells that respond to translational motions are active so long as the stimulus is present. Cells that are attuned to rotational or expansive motions, although they show a strong sustained component to their response, also display a strong transient response to the onset of a complex motion. This signal may prove crucial for navigating through the environment while simultaneously moving one's head and eyes. We shall discuss these responses in more detail in a later section.

Before we leave this section, we would like to mention an area located lateral and anterior to human MT/MST. (This newly discovered area

and the MT/MST complex are both in the superior-temporal sulcus.) This area is interesting because it extends the capabilities of the MST to the detection of biological motions, such as 'point-light walkers'. Point-light walkers are images produced by placing lights on the joints of a moving person. Subjects readily recognize these displays, and can even report the gender of the 'walker'. The newly discovered area is specifically activated by such displays.

Correlation between Cell Responses and Perception

William T. Newsome and colleagues have performed a fascinating series of experiments with MT neurons. The aim was to investigate whether motion perception is genuinely determined by their activity. These scientists trained monkeys to report their perception of the direction of a group of moving dots. Some of these dots were correlated (i.e. moving in the same direction), while others were not. By varying the percentage of correlated dots, Newsome and colleagues discovered that the monkeys' discrimination mirrored directly the activity of the cells in the MT. Furthermore, as the responses varied because of noise, the discrimination also varied in a predictable manner. These researchers further showed that they could control the monkeys' reported perception by injecting small amounts of current into the MT during the experimental trials. Thus they provided strong evidence that the firing rate of these neurons directly encoded the animals' perception.

MODELS AND MECHANISMS OF DIRECTIONAL SELECTIVITY

Models of Directional Selectivity

During their studies of retinal directional selectivity in insects, Werner T. Reichardt, Tomaso Poggio and their colleagues described the theoretical requirements for any model of directional selectivity. The first requirement is a spatial asymmetry. That is, if a neuron responds better to motions coming from the left than to motions coming from the right, then there must be some difference in the neural inputs from the left and right sides of the cell's RF. Essentially, all DS models propose that this asymmetry is temporal. In other words, some difference in time course exists between the left- and right-side inputs. However, this need not be the case. For example, the left-side input could 'gate' the right-side input. In this case, the cell will only fire if the motion comes from the left and opens the gate before the right-side input is activated.

The second requirement for producing directional selectivity is a nonlinear mechanism (i.e. a mechanism that does not simply additively sum its inputs to determine its output). A spatial (plus temporal) asymmetry as described above can generate a directional difference in the response (i.e. differing response waveforms depending on the direction of stimulus motion). However, this alone is not sufficient to produce two different single-number responses for preferred- and null-direction motions. The work by Poggio and Reichardt proves that, without a nonlinear mechanism, the numbers obtained by integrating (summing over time) the waveforms of the responses must be equal for all directions of motion. Some nonlinearity, perhaps as simple as a threshold, must be present. (A threshold nonlinearity operates by only allowing responses when the input exceeds some minimum value.)

Figure 5a illustrates the simplest model proposed for insect retinal directional selectivity, known as the Reichardt model. For clarity, the model uses inputs from only two locations, although more complex versions are possible. (The mechanism is presumably replicated many times within the RF of the DS cell.) The proposed spatial asymmetry in this model is temporal. Inputs from the side first encountered by an object moving in the preferred direction (left in the figure) propagate to the interaction site slowly compared with those from the null side. The proposed nonlinearity for their interaction is multiplication. Therefore if an object moves in the preferred direction at the right speed, the slowness of the left-side path is compensated for by the earlier arrival time of the stimulus. This allows both inputs to arrive at the interaction site at the same time, yielding a positive multiplication. For null-direction motions the inputs will arrive separately, and the result of the multiplication will be zero. Interestingly, many psychophysical models of human motion perception follow variants of this multiplicative Reichardt model. (*See* **Computational Neuroscience: From Biology to Cognition**)

Retinal Mechanisms for Directional Selectivity

As part of their study of rabbit DS cells, Barlow and Levick performed a series of two-slit apparent-motion experiments. These scientists drew attention to the responses to two adjacent slits that were flashed as if an object was moving in the

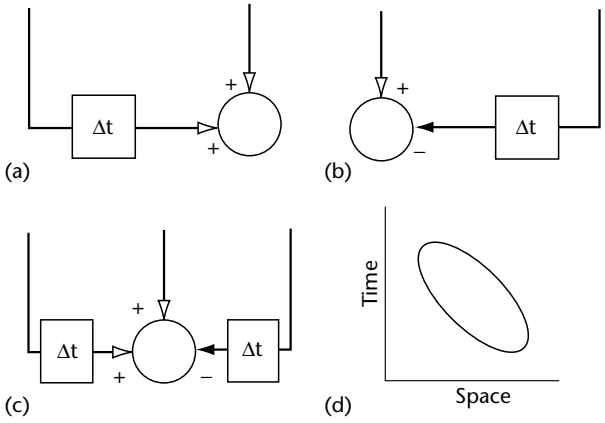


Figure 5. Models of theoretical DS mechanisms. In the retinal models (a, b and c), the lines with arrows represent inputs to nonlinear interaction sites (denoted by circles). Boxed Δt symbols denote slow or delayed input lines. Movement proceeding from the slow line to the fast line can create signals that arrive at the interaction site simultaneously. In this figure, these signals are such that all of the illustrated models prefer rightward motions. (a) In the Reichardt model, the proposed interaction is facilitatory (multiplication). (b) In the Barlow and Levick model, the proposed interaction is inhibitory (now believed to be division-like). (c) In the two asymmetrical pathways model, both facilitation and inhibition operate, allowing for robust DS to a variety of motion stimuli. (d) Spatio-temporal inseparability. The horizontal axis indicates the spatial location of a visual stimulus, while the vertical axis shows the time of peak response. For a cell with spatio-temporally inseparable responses, the tilted oval bounds the response region. Thus there is an orderly progression between space and time, such that only for motion in a particular direction will the responses from several points in space coincide in time.

null direction. The response generated by the DS cell under study was far less than the sum of the responses for each slit flashed alone. Barlow and Levick concluded from this that directional selectivity in these cells was produced by a nonlinear, inhibitory mechanism that 'vetoes' responses to null sequences. As shown in Figure 5b, their proposed spatial asymmetry has two components. The central component is excitatory and propagates quickly. The second, inhibitory component is offset to the null side and propagates slowly. Thus an asymmetry exists in both the sign and the time course of the two spatially separated components. Therefore motions in the preferred direction will yield responses, while null-direction motion responses are vetoed. Studies performed more recently by Franklin R. Amthor and colleagues have shown that these cells cannot perform a perfect veto, and that a better description of the inhibitory interaction is a division-like nonlinearity.

In addition, the response of a DS cell to a preferred-direction apparent motion is greater than the sum of the responses to each slit alone. This is called preferred-direction facilitation. Under some circumstances, preferred-direction facilitation can be as strong as null-direction inhibition. The excitation responsible for facilitation appears to originate from starburst amacrine cells which release the neurotransmitter acetylcholine. It has been shown that there is a spatial asymmetry in the input–output relationship of the dendritic trees of these cells. Their dendrites receive excitatory inputs along their entire length, but they release excitatory neurotransmitter and may receive inhibitory inputs only at their tips. (These neurons do not have axons!) Several lines of evidence indicate that the tip inhibition acts in a division-like manner. Therefore each dendrite contains a spatial asymmetry and a nonlinearity, and can act as an autonomous DS unit. For this reason it has been proposed that DS signals appear presynaptically (before the DS ganglion cell itself), and flow from the dendrites of the starburst amacrine cell to the DS cell. A DS cell could therefore generate its responses by preferentially sampling from starburst-cell dendrites with a common preferred direction.

However, there is accumulating evidence that both pre- and postsynaptic asymmetries may be involved in producing retinal directional selectivity. Complete blockade of starburst-cell outputs with acetylcholine antagonists does not entirely eliminate this selectivity to moving bars. The residual direction selectivity appears to be postsynaptic. In turn, blocking the inhibitory input to a DS cell (with antagonists of the neurotransmitter GABA) does not always eliminate retinal directional selectivity. Moreover, such blockade occasionally even reverses the cell's preferred and null directions. Computer simulations of the starburst dendritic directional selectivity can account for these reversals because of synaptic saturation. Thus asymmetrical postsynaptic inhibition and asymmetrical presynaptic facilitation may act cooperatively (Figure 5c) to produce robust directional selectivity in On–Off DS ganglion cells.

Cortical Mechanisms for Directional Selectivity

There are two classes of models that have been proposed to account for cortical directional selectivity. One class, advanced by George Sperling and colleagues, is based on human psychophysics,

and is similar to the Reichardt model shown in Figure 5a. Another class consists of the motion-energy models advanced by Edward H. Adelson and James R. Bergen. As with all DS models, motion-energy models require a spatial asymmetry. The proposed spatial asymmetry is that successive adjacent locations in the cell's RF will respond with gradually decreasing sluggishness. The cell's RF profile is therefore tilted in space and time, as shown in Figure 5d by the slanted oval. This property is known as space–time inseparability. As can be seen in the figure, for this type of space–time arrangement only one direction of object motion (to the right) can result in the DS cell's inputs all arriving simultaneously. As before, simple linear summation will result in differential response waveforms for preferred- and null-direction motions, and some nonlinearity must exist to convert this directional difference into directional selectivity. The most commonly proposed nonlinearity is squaring, which is described as extracting the motion energy from the directional difference. (*See* **Psychophysics**)

Robert M. Shapley and colleagues obtained physiological evidence for space–time inseparability in both the on and off subregions of the inputs to the simple DS cells found in V1. However, the correlation between the simple cell space–time profile and direction selectivity varies widely in V1. Cells in some layers of V1 show a very high correlation, while those in other layers show a very low correlation, despite equivalent directional tuning. Thus space–time structure alone cannot completely account for simple DS cell responses. In addition, these models of DS simple cells generally overestimate non-preferred responses and sometimes underestimate preferred responses. Moreover, these models do not predict onset transients, which are commonly observed. Therefore both inhibitory and excitatory feedback interactions between cortical cells have been proposed to account for these discrepancies. The exact mechanisms for producing directional selectivity in these cells are still the subject of debate.

Because the complex DS cells of V1 do not respond to flashed spots, they cannot show space–time-oriented RFs. However, it has been demonstrated that the interactions between two sequentially stimulated locations in a complex cell's receptive field are space–time inseparable. This behavior is termed second-order space–time orientation. Dynamic nonlinearities have been proposed to account for the directional selectivity of complex DS cells. These nonlinearities would facilitate or inhibit, respectively, the responses to preferred- or null-direction motions. Similarly, space–time-separable simple cells have also been shown to display some second-order space–time structure. Because there is ample evidence for interactions between complex and simple cells, it is possible that second-order space–time inseparability in simple cells arises from complex-cell inputs.

The directional selectivity in MT neurons is more sophisticated than that in V1, especially in pattern cells (Figure 4). Many models have been advanced to account for the orientation independence of pattern cells. These models often begin by pointing out the aperture problem of V1 cells – that is, the difficulty in determining the true direction of motion through small apertures (e.g. small RFs). This is because they only reveal small, straight portions of contours, and it is impossible to tell the direction of motions parallel to straight edges. Models of pattern cells combine the responses of V1 (or component) cells with many preferred directions to disambiguate the direction of motion of multi-orientation patterns.

SELECTIVITY TO SPEED

The visual system measures not only the directional component of the velocity vector but also its magnitude – the speed. Measuring speed is important for vision. For example, humans appear to use both direction and speed signals to obtain a precision of 1–2° in the estimation of heading direction. Furthermore, Richard Andersen and colleagues have shown that both humans and primates benefit from speed information when estimating the three-dimensional structure of objects from motion cues. Psychophysical findings also indicate that one can segment images based solely on gradients of speed signals. Moreover, smooth pursuit and saccades are made precise because speed signals are used in their computations. Finally, the visual system must somehow use speed signals to achieve effective deblurring. (Motion blurring occurs when one leaves the shutter of a camera open for too long when there is motion in the image.)

However, measuring speed is more difficult than measuring the direction of motion. It is difficult to measure the former because it is the magnitude of the derivative of position over time, and thus requires precise spatio-temporal information. In contrast, to measure direction one only requires two relatively imprecise position measurements. Psychophysically determined measurements of local and instantaneous speed have been found to be very noisy. The relative errors in the measurement of local speed are in the range 30–100%.

Nevertheless, humans can measure speed precisely if they are provided with a relatively long trajectory of motion. Under these conditions, the errors made when discriminating speed can be as low as 5%. This high level of precision seems to be achieved by integrating relatively imprecise local-speed signals over time. The 5% precision of velocity determination occurs in many experimental conditions, including motions of dots, edges, sine-wave gratings (of varying spatial and temporal frequencies), plaids and frequency-modulated stimuli. (In the latter, the local-intensity profile is not moving, but the contrast of local portions of the profile increases transiently, and this perturbation propagates with fixed velocity.)

Perhaps the simplest method of measuring local visual speed is by its derivative definition. This involves finding the positions of an image feature in two discrete instances in time and then computing the ratio between the positional distance and the temporal delay. It is essentially the approach proposed by Shimon Ullman in his minimal mapping theory. He proposed that the main problem facing the visual system when measuring motion is to solve the correspondence problem – that is, to find correspondences between image features at one instant in time and (hopefully) the same features at the next instant in time. Ullman suggested that the features correspond to minimize the total distance traveled. After correspondence has been established, distance, delay and thus the velocity of a feature can then be calculated. Serious challenges to Ullman's emphasis on the correspondence problem, and thus his method for measuring visual speed, have been raised by motion psychophysicists, who pointed out that his theory was not immediately consistent with known neural processes underlying motion perception. Moreover, they argued that the correspondence problem is essentially non-existent when one is dealing with real neural RFs. (*See* **Computational Neuroscience: From Biology to Cognition**)

When one looks directly at neural responses for clues as to how the brain measures speeds, a puzzling finding arises. Neurons have a sharp optimal speed if they are stimulated with moving edges, especially in the MT. However, cells in V1 and 40% of the cells in the MT do not seem to detect speeds. If these cells were speed tuned, then if one were to raise the temporal frequency of the stimulus, the spatial frequencies that yield the optimal responses should increase in proportion (Figure 6a). Rather, these motion sensitive cells are tuned to spatio-temporal frequencies, regardless of speed (Figure 6b). Nevertheless, the work of

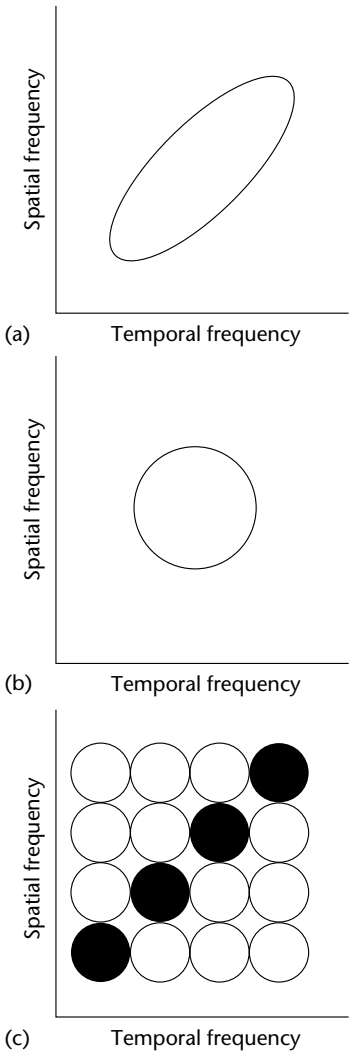


Figure 6. Theories for speed encoding. For (a), (b) and (c), the vertical axis indicates the spatial frequency of a moving sine-wave grating and the horizontal axis indicates the temporal frequency at which the intensity of every point oscillates. (a) Idealized relationship between optimal spatial and temporal frequencies for a cell that can encode speed. For any given speed, as the spatial frequency of the grating increases, the optimal temporal frequency increases in proportion. Therefore speed is indicated by the slant (slope) of the oval response region. (b) Spatio-temporal profile of a non-speed-selective cell on the motion pathway. Most motion-selective cells are band-pass in both the spatial and temporal domains. Non-speed-selective cells do not show the organized relationship (shown in a) that is required for speed selectivity. (c) A population code for speed. Each cell is represented by a circle whose center denotes the optimal spatial and temporal frequencies of the cell. Solid and open circles illustrate responding and non-responding cells during a translation. By looking at a population of cells with band-pass spatio-temporal response properties, a higher-level unit can determine speed regardless of the spatial and temporal frequencies of the stimulus.

John A. Perrone and Alexander Thiele revealed that 60% of MT cells show true speed selectivity. So how do 60% of MT neurons compute speed from inputs that are not speed selective?

A Population Code for Speed

Several investigators have proposed that speed selectivity arises from a population code. In such a code, a speed-selective cell high in the motion-pathway hierarchy could read the speed of the stimulus from the simultaneous firing of many earlier cells. Eero Simoncelli, David Ascher and colleagues obtained psychophysical data that support the concept of a population code. We shall now explain this idea in more detail.

Population-code models for the measurement of local speed begin by looking at the responses of a population of DS neurons with different receptive-field sizes. It is known that V1 neurons with increasingly larger receptive-field sizes are tuned to increasingly lower spatial frequencies when they are stimulated with sinusoidal gratings. Moreover, different neurons tend to be tuned to different temporal frequencies of the gratings. The population-code models for local speed measurement consider a three-dimensional space defined by these tunings. In this space, two axes are the optimal horizontal and vertical spatial frequencies that drive different DS cells, while the third axis represents their optimal temporal frequencies. A point in this space corresponds to a DS cell. This space is of interest when a visual translation of given velocity covers the receptive fields of these cells. The optimal responses tend to fall on a plane in this space. Figure 6c illustrates the projection of this plane on to the plane formed by the temporal-frequency axis and one of the spatial-frequency axes. To measure local speed in population-code models, one must detect the slant of the projecting plane relative to the temporal-frequency axis. Several schemes for detecting this plane have been proposed in the literature. These schemes depend on the exact spatio-temporal properties of the cells, although Ascher and colleagues described how to design good schemes under broad conditions.

Finally, although we first introduced the concept of population codes in the context of determining speed, such codes must occur generally in the brain. Even if a cell is tuned to a particular property, the response of that cell may be modulated by other properties. This would mean that a cell's firing rate was not unique. Consequently, the brain must look at the firing of several cells in order to disambiguate any particular property. In the motion domain, in addition to speed, such a disambiguation also occurs for direction of motion. For example, one can obtain local direction from the tilt of the plane described above relative to the spatial frequency axes. (*See* **Decoding Neural Population Activity**)

SELECTIVITY TO COMPLEX MOTIONS

Interesting Types of Complex Motion and MST

The two preceding sections focused on the measurement of two local variables of motion, namely direction and speed. However, as was shown in Figure 1, the motion of optic flow is in general globally complex. This complexity arises because, for example, as one moves through the environment, neighbor velocity vectors are statistically dependent. A similar dependence occurs when a rigid or quasi-rigid object moves in front of one's eyes. Jan J. Koenderink showed that if one considers small planar patches in the surface of moving, rigid objects, their general motion can be decomposed in terms of translational, radial (expansion and contraction), rotational and shear motions. These types of motion can be modeled with a few parameters describing the dependence of the local velocity vectors. For instance, one can describe a rotation by specifying its center and its angular velocity. In turn, one can specify an expansion by specifying the focus and rate of expansion. Thus if the brain could measure the few parameters of translational, radial, rotational and shear motions, then it could estimate such things as direction of egomotion heading. As was explained earlier, the MST is the first cortical area to be selective for these motions. Neurons in the dorsal portion of this area (MSTd) respond selectively to these motions, either alone or in combination (e.g. spiral motions). Figure 7 illustrates how one sees this selectivity in the electrophysiological recordings. An expansion-selective cell (second column) responds strongly to expansion (a) but not to contraction (b) or rotations (c and d). In contrast, a cell that is selective for clockwise motion (second column) responds strongly to this motion (c) but not to counter-clockwise rotation (d) or radial motions (a and b). These cells also do not respond to translation (not shown in the figure).

Roles of MST Neurons

What are the behavioral functions of the MST cells that are selective for complex global motions? Kenneth J. Britten and Richard J. van Wezel showed that microstimulation of neurons in this

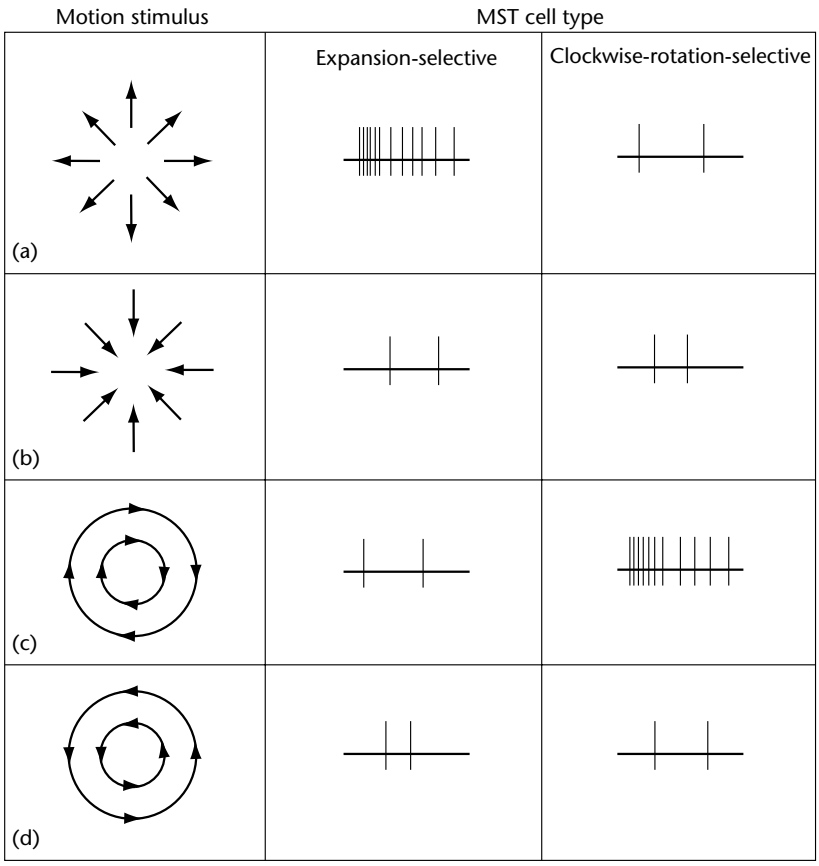


Figure 7. Responses to complex motions by MST cells. The left-hand column shows various complex-motion stimuli, while the center and right-hand columns indicate the responses of expansion-selective and clockwise-rotation-selective cells, respectively. Each cell type responds only to a specific global coherent motion presented within its RF. (a) Expansion. (b) Contraction. (c) Clockwise rotation. (d) Counter-clockwise rotation.

area influences the heading behavior of primates. Related to this, Andersen and colleagues showed that many MSTd neurons shift their focus-of-expansion tuning curves, compensating for retinal motions during eye movements. (This tuning curve is the cellular activity as a function of the focus location of the stimulus.) Because the focus of expansion indicates the direction of heading, this and the results of Britten and Wezel demonstrate the contribution of these neurons to this function. Another of their roles was demonstrated by lesions of the MST, which impaired the animal's ability to execute a smooth-pursuit eye movement when the target moved towards the lesioned hemisphere. (A similar deficit was seen for optokinetic nystagmus movements.) Furthermore, microstimulation within the MST influenced the velocity of smooth-pursuit eye movements. Thus MST neurons must also contribute to the control of eye movements. A final role that we shall mention here is that MST neurons can code the abstract concept of a motion as well as the actual motion of a visual stimulus. In the experiment demonstrating this role, animals saw some stimuli appear and then move in some trials. In other trials, the stimuli appeared, disappeared, and then reappeared at the same final location as the moving stimuli. This simulated a motion that might have occurred behind an occluder. Many MST neurons coded the direction of the occluded motion, which strongly suggested that the MST has a central role in the perception of motion.

Mechanisms of MST Receptive Fields

It is thought that the receptive field properties of MST neurons emerge from the combination of properties of earlier neurons. This is very similar to pattern cells in MT being built out of component cells. For example, one could build a clockwise-rotation-selective cell by using DS cells with leftward, upward, rightward and downward preferred directions. All one has to do is to place these

cells in the bottom, left, top and right visual fields, respectively. However, one problem with this idea is that these cells show position and scale invariance. In other words, although the magnitude of the response may vary with location, the selectivity for a particular pattern of motion remains the same throughout the receptive field, regardless of size. Thus building MSTd RFs is not simple, and may involve dendritic subunits such as those described earlier for the retina. A further complication is that the complex selectivities in MST are also independent of the cues that convey the motion. For instance, effective expansion stimuli could be generated by illusory contours. Consequently, one must explain how any motion that appears perceptually as an expansion will activate an expansion neuron, even if the luminance pattern is not expanding.

It also remains unclear whether MST computations include estimations of all of the parameters of the optic-flow components described by Koenderink. Koenderink himself, and some of his colleagues, performed psychophysical experiments which showed that humans do not metrically discriminate all of the parameters of complex motions. For example, according to these experiments, humans may not discriminate between angular velocities. However, many of the experiments used impoverished visual displays. Experiments that have been conducted more recently with richer displays indicated that humans could measure quantitatively certain complex-motion parameters, including angular velocity.

Further Reading

Andersen RA (1997) Neural mechanisms of visual motion perception in primates. *Neuron* **18**: 865–872.

Borg-Graham LJ and Grzywacz NM (1991) A model of the direction selectivity circuit in retina: transformations by neurons singly and in concert. In: McKenna T, Davis J and Zornetzer SF (eds) *Single Neuron Computation*, pp. 347–375. Orlando, FL: Academic Press.

Croner LJ and Albright TD (1999) Seeing the big picture: integration of image cues in the primate visual system. *Neuron* **24**: 777–789.

Grzywacz NM and Yuille AL (1991) Theories for the visual perception of local velocity and coherent motion. In: Landy MS and Movshon JA (eds) *Computational Models of Visual Processing*, pp. 231–252. Cambridge, MA: MIT Press.

Grzywacz NM and Merwine DK (2002) Directional selectivity. In: Arbib M (ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

Miles FA and Wallman J (1993) *Visual Motion and its Role in the Stabilization of Gaze*. Amsterdam: Elsevier Science.

Smith AT and Snowden RJ (1994) *Visual Detection of Motion*. London, UK: Academic Press.

Watanabe T (1998) *High-Level Motion Processing: Computational, Neurobiological and Psychophysical Perspectives*. Cambridge, MA: MIT Press.

Wurtz RH and Kandel ER (2000) Perception of motion, depth and form. In: Kandel ER, Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, pp. 548–571. New York: McGraw-Hill.

Zanker JM and Zeil J (2001) *Motion Vision: Computational, Neural and Ecological Constraints*. Berlin, Germany: Springer Verlag.

Zeki S (1990) The motion pathways of the visual cortex. In: Blakemore C (ed.) *Vision: Coding and Efficiency*, pp. 321–345. Cambridge, UK: Cambridge University Press.

Motor Areas of the Cerebral Cortex

Intermediate article

Marc H Schieber, University of Rochester School of Medicine and Dentistry, Rochester, New York, USA
Andrew J Fuglevand, College of Medicine, University of Arizona, Tucson, Arizona, USA

CONTENTS

The motor areas of the cerebral cortex are those four regions most directly involved in deciding which movements to make and in executing the selected movements – posterior parietal, dorsolateral prefrontal, secondary motor, and primary motor cortex.

INTRODUCTION

In a sense, the entire nervous system functions to animate the body. Based on sensory inputs about external and internal conditions, on stored information about past experience, and on considerations of current and future needs, the nervous system organizes and executes movements intended to optimize the life of the animal. Some of these movements, such as those associated with subtle adjustments of the legs needed to maintain upright posture, are reflexive and are mediated primarily through spinal or brainstem circuits. Others serve more complex purposes, and involve goal-directed voluntary activities such as acquisition of food, creation of shelter, production of tools, and communication through gesture, vocalization, speech, writing, dance, or music. The mammalian cerebral cortex performs much of the complex information processing used in generating these voluntary, goal-directed movements.

With this broad view, the entire cerebral cortex could be considered 'motor'. Here, however, we will focus on four regions most directly involved in deciding which movements to make and in executing the selected movements. Information traditionally is considered to flow from sensory input and perception, to associative decision-making, to execution of motor output. We will therefore consider how the posterior parietal cortex processes visual and somatosensory inputs to guide movements in space; how the dorsolateral prefrontal cortex decides which responses to make; how secondary motor areas participate in selection of movements to perform; and how the primary motor cortex executes movements. During any given movement, neurons in many interconnected regions of the cerebral cortex (and other parts of the brain as well) are active concurrently, and their discharge may share many features.

Figure 1 illustrates where these regions lie in the cortex of a macaque monkey, and how they are interconnected.

ROLES OF THE FOUR MOTOR AREAS

The Posterior Parietal Cortex: Guiding Voluntary Movements in Space

In order to reach out, grasp, and handle an object, two fundamental operations need to be performed. Firstly, the body part that is to interact with the object needs to move into the vicinity of the target object. In primates, this typically is the hand, but could also be the mouth or the foot. Inevitably, this transport function is the responsibility of the appendage (the arm, head, or leg) upon which the interacting body part is fixed. Secondly, the manipulator, such as the hand, needs to be configured appropriately so as to conform to the shape of object to be grasped.

Different types of spatial information are required to perform these operations successfully. For the transport function, information is needed about both the location of the object and the location of the limb. In order to manipulate objects proficiently, additional information is needed about the shape and orientation of the target object. Two major corticocortical inputs converge in the posterior parietal cortex to provide this information: inputs from the dorsal visual stream of the occipital

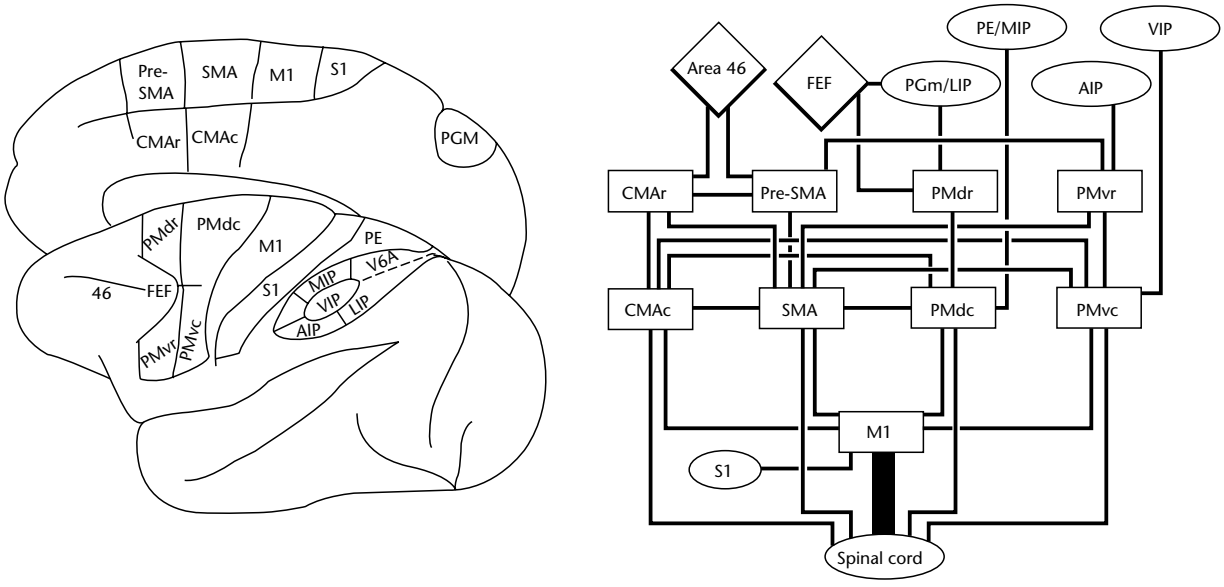


Figure 1. Cortical motor areas of the macaque monkey. (a) Anatomic location. The primary motor cortex (M1) is equivalent to Brodmann's area 4. Brodmann's area 6 on the hemispheric convexity has been subdivided into ventral and dorsal premotor cortex (PMv and PMd, respectively) on the convexity, each with further caudal (c) and rostral (r) subdivisions. Brodmann's area 6 on the medial surface of the hemisphere (shown above), formerly considered the supplementary motor area (SMA), recently has been subdivided into SMA proper and pre-SMA. Additional motor areas have been identified in the cingulate cortex (Brodmann's areas 23 and 24); this cingulate motor area (CMA) again has caudal and rostral subdivision. (b) Major corticocortical and corticospinal connections of cortical motor areas are illustrated in a wiring diagram. Most of the corticocortical interconnections are reciprocal. Thinner lines to the spinal cord indicate that the corticospinal projections from these areas are not as strong as that from M1. Abbreviations: M1, primary motor cortex; S1, primary somatosensory cortex; CMAc, cingulate motor area, caudal; CMAr, cingulate motor area, rostral; PMdc, premotor cortex, dorsal, caudal; PMdr, premotor cortex, dorsal, rostral; PMvc, premotor cortex, ventral, caudal; PMvr, premotor cortex, ventral, rostral; AIP, anterior intraparietal area; MIP, medial intraparietal area; VIP, ventral intraparietal area; LIP, lateral intraparietal area; PE, parietal region E; PGm, parietal region G, subregion m; FEF, frontal eye field; V6A, superior parietal lobe, caudal. Modified from Schieber (1999).

lobe, which processes information on the spatial features of visual objects, and inputs from the somatosensory cortex of the anterior parietal lobe, which processes tactile and proprioceptive information on the position and motion of body parts. (*See* **What and Where/How Systems**; **Parietal Cortex**)

Combining visual and somatosensory information, the posterior parietal cortex plays a major role in guiding movements of the body to interact with external targets. Lesions here impair the ability to guide and orient movements accurately towards visually perceived objects in external space, without impairing the ability to identify and describe an object, or to discriminate it from other objects. In humans, lesions in the posterior parietal cortex may produce the syndrome of apraxia, in which the patient makes inappropriate (though dextrous) movements when handling objects, such as an inability to pour water accurately from one glass into another.

Though much of the posterior parietal cortex is active during any natural movement to reach out and manipulate objects, various regions make different contributions to such movements. These regions have been defined most precisely in macaque monkeys and include area 5 in the superior parietal lobe, area 7 in the inferior parietal lobe, and several regions that lie in the intraparietal sulcus between areas 5 and 7 (Figure 1).

Input to area 5 arises mainly from the primary somatosensory cortex (S1) through which it receives information about the status of joints, muscles, and skin. However, unlike neurons in S1, neurons in area 5 show little response to passive somatosensory stimuli. Instead, neurons in area 5 are active predominately when the monkey reaches toward a desired object in the space immediately surrounding the animal (Mountcastle *et al.*, 1975). Limb movements that are not associated with this type of reaching, such as might occur during aggressive

or aversive behaviors, fail to elicit much activity in these neurons. Moreover, individual neurons in area 5 are broadly tuned to a preferred direction of movement in that they fire most briskly when the arm moves in a particular direction and fire little when the arm moves in the opposite direction.

Interestingly, this directional tuning is unaffected by the application of static loads to the arm which changes the pattern of muscular activity needed to move the arm in a particular direction (Kalaska *et al*., 1990). Therefore, the collective activity of neurons in area 5 seems to provide a robust representation of the trajectory of a limb associated with transporting the hand toward a desired object. Because neurons in area 5 often respond prior to the onset of movement, their activity may not simply reflect the sensory consequence of movements but is probably intimately related to the planning of goal-directed movements. Indeed, area 5 is richly interconnected with the premotor and primary motor areas of the frontal lobe, two regions of the primate brain most directly associated with planning and executing voluntary movements.

While neurons in area 7 are mostly visual in nature, in several respects their behavior resembles that of neurons in area 5. Many area 7 neurons respond when the animal gazes upon or reaches toward an object of interest (Hyvärinen and Poranen, 1974). For example, these neurons become active when a monkey looks at a morsel of food, but will not discharge when the monkey looks at a book, a laboratory instrument, or other object of little interest to a monkey. In addition, the object of interest must be within arm's reach in order to evoke consistent activity in these neurons. Furthermore, activity ceases when the animal becomes satiated if the object is food or a target associated with delivery of food or liquid reward.

Many neurons in area 7 are directionally sensitive and will discharge when looking at or reaching toward an object placed in one location within the peripersonal space but not in other locations. Similarly, some neurons in area 7 are active only during smooth pursuit of an object of interest while it is moved in one direction but not in other directions within the monkey's peripersonal space. Therefore, it seems that one role of area 7 is to encode the spatial attributes of objects in the region immediately surrounding the animal. However, this does not occur in an obligatory fashion; instead, spatial representation by neurons in area 7 is allocated only to the object at any given moment that the animal desires to grasp or touch. Such spatial information is then conveyed to premotor areas in the frontal cortex where it is essential in formulating the motor plan associated with reaching.

The intraparietal sulcus forms the boundary between areas 5 and 7. Several regions within the intraparietal sulcus have been identified and neurons in these regions typically respond to a blend of visuospatial and somatosensory stimuli. Different regions within the intraparietal sulcus, however, seem to make particular contributions to movements of the hand, arm, head, and eyes. For example, the anterior intraparietal (AIP) area lies in the anterior portion of the lateral bank of the intraparietal sulcus and plays a specific part in hand movements (Figure 2).

This area contains neurons that respond differentially depending on the visual shape, size, and orientation of small objects that the monkey might grasp using different hand postures (Taira *et al*., 1990; Murata *et al*., 2000). Some of these neurons respond not only when the monkey sees the object, but also when the monkey grasps the object, even in the dark. When the homologous region in humans is affected by a lesion, the shaping of the hand is abnormal as the patient reaches out to grasp a visually perceived object. The AIP thus makes a crucial contribution to adjusting the posture of the hand to grasp an object based on visual or somatosensory inputs. This area is closely connected to a region in the ventral premotor cortex (see below) that is active during the production and planning of various grasping movements of the hand. Thus, AIP primarily subserves the manipulation component of reaching and grasping, whereas areas 5 and 7 contribute primarily to the spatial representation of the arm and target, respectively, needed to successfully transport the hand to an object of interest.

Visual and somatosensory input associated with arm movement influences the activity of neurons in the caudal aspect of the superior parietal lobe (V6A) and in the medial bank of the intraparietal sulcus (medial intraparietal area, MIP). In V6A, individual neurons respond prior to and during arm reaches or eye movements toward targets located in a particular region of peripersonal space. The response is strongest, however, when both the eyes and arm are allowed to move together toward the target (Battaglia-Mayer *et al*., 2000). This suggests that V6A combines information about target location and about arm position. It thus represents an important node in the parietofrontal network involved in planning and executing goal-directed arm movements.

In area MIP, somatosensory and visual information is combined in a slightly different way.

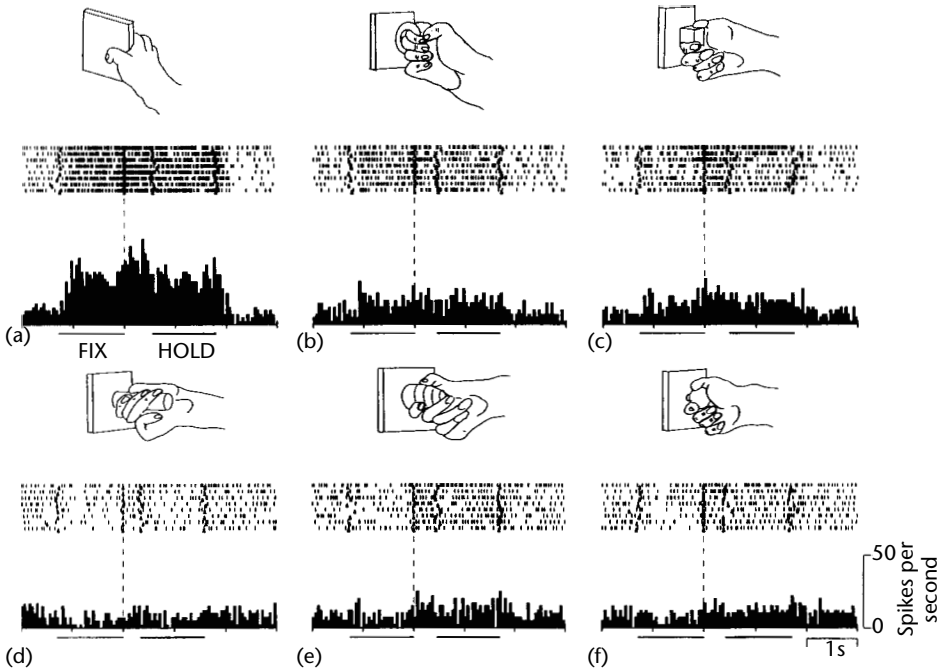


Figure 2. Activity of a neuron in the anterior intraparietal (AIP) area recorded as a monkey grasped six objects of different shape. This neuron became active as the monkey visually fixated a small vertical plate (a, FIX) and continued to discharge as the monkey grasped, pulled and held (a, HOLD) the plate. The same neuron showed less activity if the object was a vertical ring (b) or a small cube (c), and little if any modulation when the object was a cylinder (d), cone (e) or sphere (f). The neuron showed similar, though somewhat lower, discharge if the monkey looked at the plate without reaching out to grasp it, or if the monkey reached out to grasp the plate in the dark (not illustrated). Reproduced from Murata *et al.* (2000).

Neurons in MIP seem to respond to visual targets that are in the vicinity of the arm with activity increasing the closer the target is to the arm. Many of these neurons are bimodal with tactile receptive fields on the arm and visual receptive fields surrounding the arm. Area MIP therefore seems to represent the spatial location of objects with respect to the arm or hand. Intriguingly, if a small rake is used by a monkey in order to obtain food morsels located outside the region that could be normally reached by the arm, the visual receptive field of neurons in MIP expand to encompass the territory that can be reached by the rake (Iriki *et al*., 1996).

The ventral intraparietal (VIP) area, lying in the fundus of the intraparietal sulcus, plays a role in head movements. This area contains neurons that respond to moving visual stimuli directed toward tactile receptive fields of the same neurons located on the head and face. The visual receptive fields remain spatially anchored to the tactile receptive fields on the head even if the eyes move. The VIP area thus is likely to play a role in guiding movements of the head in relation to nearby objects.

Likewise, the lateral intraparietal (LIP) area, lying posterior to AIP in the lateral bank of the intraparietal sulcus, plays a role in eye movements. Neurons of the LIP have retinotopic (visual) receptive fields, though their responses are modulated by proprioceptive information regarding the position of the eyes in the orbit. The responses of LIP neurons are enhanced when the subject must pay attention to a stimulus in its receptive fields, and will continue to discharge after the stimulus disappears if the subject must remember the location. These neurons also discharge just before saccadic eye movements that will shift the line of sight onto a stimulus in the subject's receptive fields. The LIP area, which is interconnected with the superior colliculus and the frontal eye fields, thus seems to participate in selecting visual targets for saccadic eye movements.

The Dorsolateral Prefrontal Cortex: Deciding Which Voluntary Movements to Make

The cortex on the surface convexity of the frontal lobe, anterior to the primary motor and premotor

cortex in the precentral gyrus (see below), can be viewed as one of the most complex association areas of the brain. Along the phylogenetic scale of mammals – from rodents, to carnivores, to primates, to humans – the dorsolateral prefrontal cortex has enlarged more extensively than other cortical regions. Dorsolateral prefrontal lesions in humans impair foresight and long-range planning ability, leading to the general notion that this cortical region serves executive functions that supervise behavior, selecting which behaviors to execute, and which to withhold.

At the neuronal level, some prefrontal neurons have brief responses to particular sensory stimuli, and some discharge as movements are executed, but the striking property of prefrontal neurons is their discharge between an instructional cue and execution of the instructed movement (Funahashi, 2001). This property is brought out experimentally by presenting an instructional cue only briefly, and then requiring a monkey to wait several seconds before finally performing the instructed movement when a separate, nonspecific triggering cue appears (Figure 3).

Many prefrontal neurons will start to discharge shortly after an appropriate instructional cue, and then maintain their discharge after the instructional cue disappears until finally the monkey receives the trigger cue and executes the movement. (Similar activity appears during instructed delays, albeit progressively less frequent, in secondary and primary motor areas.) This activity of prefrontal

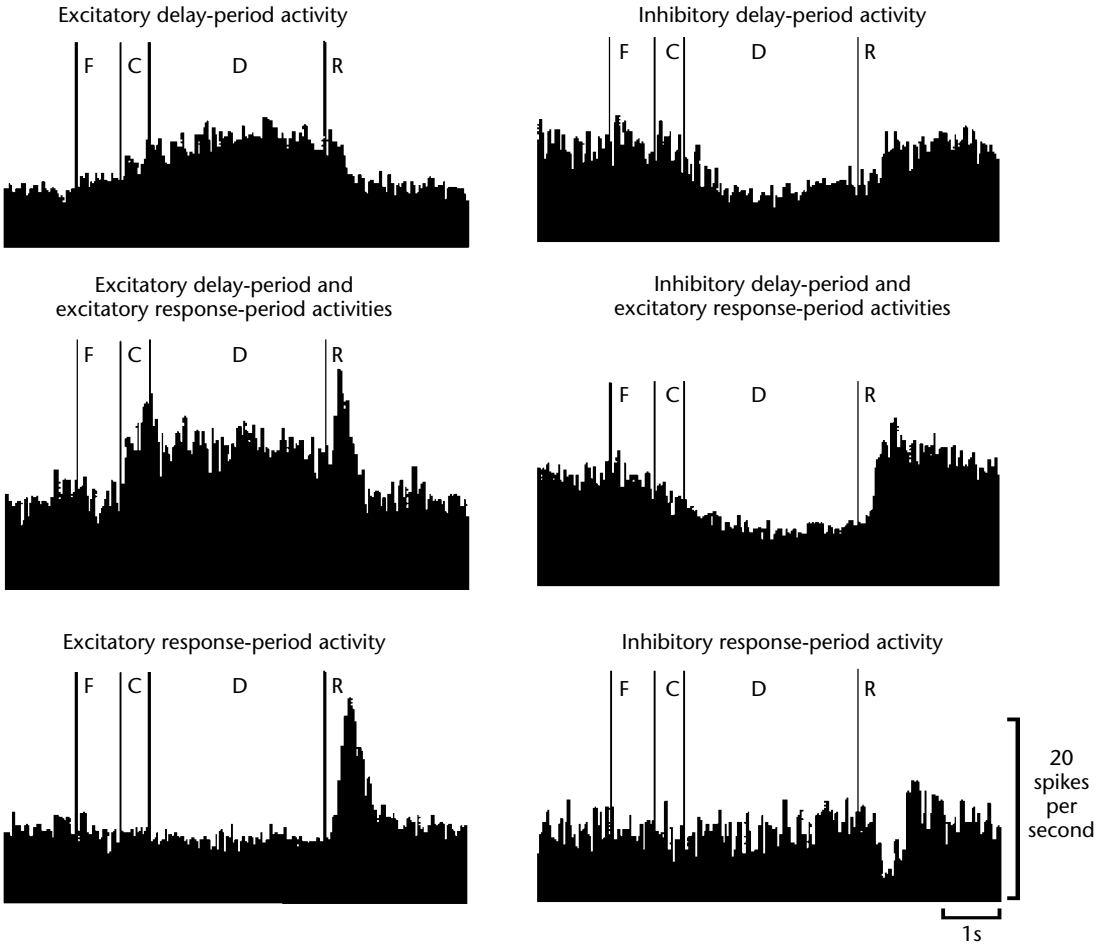


Figure 3. Activity of six different prefrontal neurons during an oculomotor delayed-response task in which the monkey first fixes (F) its eyes on a central target, then receives a transient cue (C) to respond to one of several peripheral targets, and then must wait through a delay (D) period before receiving a second nonspecific triggering instruction, after which the monkey responds (R) by turning its eyes to fix on the cued target. Individual neurons show different combinations of excitatory and inhibitory changes in their discharge immediately after the cue appears, during the delay period, and when the eye movement response is made. Delay-period activity is a prominent feature of neurons in the dorsolateral prefrontal cortex. Reproduced from Funahashi (2001).

neurons during such an instructed delay period appears to serve as a 'working memory', holding information either on which cue was presented, or on which movement to perform, while the monkey waits for the trigger cue. Indeed, when the delay period discharge of such neurons is inappropriate given the instructional cue, the monkey often executes an erroneous movement after the trigger cue. Moreover, lesions of the dorsolateral prefrontal cortex impair the ability to perform delayed-response tasks. Although the subject can make the correct response if the response can be made immediately, if the response must be delayed the subject seems to gradually forget which response was correct.

Different regions of the dorsolateral prefrontal cortex receive different combinations of processed sensory inputs. Inputs from the posterior parietal cortex carrying spatial information are directed chiefly to area 46 around the principal sulcus. Inputs from the inferotemporal cortex of the ventral visual pathway that processes visual information for recognition and discrimination of complex visual stimuli are directed chiefly to areas 45 and 12 inferior to the principal sulcus. Inputs from different regions of auditory cortex also are directed differentially to areas 46, 45, and 12. Consequently some prefrontal neurons discharge during delayed-response tasks when the spatial location of the stimulus must be remembered, while others discharge when nonspatial attributes of the stimulus must be remembered. The discharge of dorsolateral prefrontal neurons also reflects the rule used in selecting the next response; for example, whether the response must be selected based on the spatial location of the instructional stimulus (e.g. is the stimulus on the right or on the left) or the nonspatial attributes of the stimulus (e.g. is the stimulus a picture of a banana or a picture of a whale) (White and Wise, 1999). Dorsolateral prefrontal neurons thus appear to 'keep in mind' various types of information as the subject applies rules to decide what to do next. (*See* **What and Where/How Systems**)

The Wisconsin Card Sorting Test reveals this classic aspect of dorsolateral prefrontal function in humans. Cards with different numbers of colored geometric shapes are presented to the participant, who is asked to sort them into stacks according to a rule that the participant must discern simply from being told whether each card in turn has been sorted correctly or incorrectly. The rule may require sorting according to color, number, or shape. Unbeknownst to the subject, the rule is periodically changed by the experimenter (e.g. from color to shape). Normal participants quickly realize that the rule has been changed and proceed to discern the new rule, but people with dorsolateral prefrontal lesions may fail to look for a new rule even after being told that they have sorted card after card incorrectly. They are unable to adapt flexibly to changing conditions and remain 'stuck in set', following the old rule long after the responses determined by that rule have become inappropriate. Such repeated production of the same voluntary response, initially but no longer appropriate, is termed 'perseveration'. Perseveration may affect behaviors from simple movements to complex acts.

The dorsolateral prefrontal cortex thus appears to have an important role in selecting the next behavioral response based on the combination of recent stimuli and consideration of complex rules and conditions.

The Secondary Motor Areas: Selecting Voluntary Movements

Movements can be evoked by electrical stimulation in only a limited region of the mammalian neocortex, located in the posterior frontal lobe. Corresponding to Brodmann's areas 6 and 4, this region of electrically 'excitable cortex' contains neurons in layer V with corticospinal axons that descend to the ventral horn of the spinal cord. Such corticospinal neurons are most concentrated in area 4. (*See* **Descending Motor Tracts**)

Though once considered altogether as the motor cortex, this region now is considered to contain multiple cortical motor areas distinguished by their interconnections with other cortical areas and with the thalamus, their cytoarchitectonics and neurotransmitter distributions, and their contributions to control of voluntary movement. While multiple cortical motor areas are present in the human premotor cortex, these areas have been distinguished most thoroughly in experimental studies on nonhuman primates.

As shown in Figure 1(a), area 4 is now considered to be the primary motor cortex (M1), whereas area 6 contains multiple secondary motor areas. The premotor cortex on the dorsolateral surface of the hemisphere is subdivided into ventral and dorsal portions, with each of these portions further subdivided into rostral and caudal sections (PMvc, PMvr, PMdc, and PMdr). On the medial aspect of the hemisphere, area 6 contains the supplementary motor area (SMA), and a more rostral pre-SMA. Area 23 in the cingulate sulcus contains rostral and caudal cingulate motor areas (CMAr

and CMAc) as well. As indicated by the interconnection diagram shown in Figure 1(b), the more caudal of these secondary areas have direct, reciprocal interconnections with M1, and have their own corticospinal projections as well. Though varying degrees of activity may be present in many of these areas during any given voluntary movement, the secondary motor areas participate differentially in movement performed in various contexts.

The ventral premotor cortex appears most involved in selecting movements in which the body will be brought into contact with visually perceived objects. Considerable visual and somatosensory input arrives in PMv from the posterior parietal lobe, particularly from areas VIP and AIP in the intraparietal sulcus (see above). Neurons in PMvc have large, often bilaterally symmetric, somatosensory receptive fields on the arms, hands, or face. These neurons also respond to visually perceived objects approaching their somatosensory receptive field. When the body part carrying a neuron's somatosensory receptive field is moved in space, the visually responsive region of space for that neuron moves with the body part. These neurons also discharge more when the monkey reaches for targets in response to visual cues than when the monkey reaches for the same targets based on a remembered sequence (Figure 4).

In PMvr, neurons tend to discharge when the monkey uses its hand in particular ways. One neuron may discharge when the monkey uses a precision pinch to pick up a raisin, while another neuron discharges when the monkey uses a power grip to hold a large bar. Moreover, the same PMvr neuron that discharges when the monkey performs a precision pinch may also discharge when the monkey watches another monkey, or even a human, pick up a raisin with a precision pinch. Consistent with a role for PMv in selecting and guiding movements to interact with visual objects, lesions in the ventral premotor cortex decrease the likelihood that the subject will use the contralateral extremity or move toward targets in contralateral space.

The dorsal premotor cortex appears most involved in selecting movements based on abstract instructions or conditional cues. For example, if a monkey is taught to touch a button on the right when a green light flashes in the midline but a button on the left when a red light flashes in the midline, many PMdc neurons will discharge after the appearance of one light or the other, signaling the impending movement. Indeed, if a neuron seems to discharge in error – for example, if a neuron that normally discharges after a green flash starts to discharge after a red flash – the monkey typically will make the wrong movement.

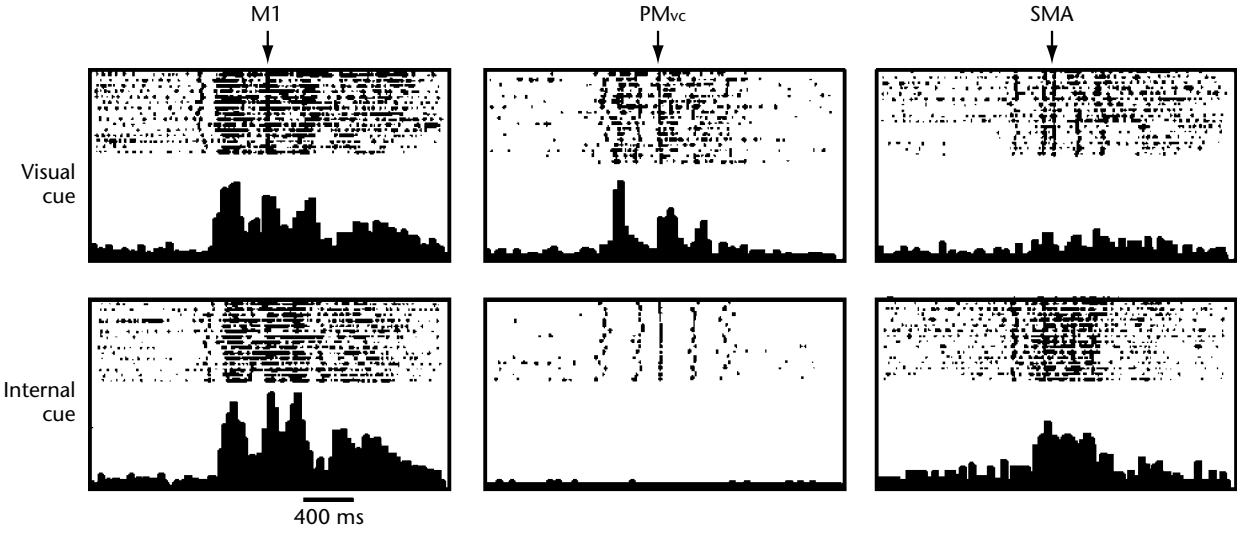


Figure 4. Activity of three neurons – one from M1, one from PMvc, and one from SMA – recorded as a monkey pressed three buttons in a sequence specified at first by lighting the buttons in sequence (visual cues), and then later from memory (internal cues) without the buttons lighting on each trial. Whereas the M1 neuron showed similar activity whether the monkey performed from visual or internal cues, the PMvc neuron was much more active when movements were selected on the basis of visual cues, and conversely the SMA neuron was much more active when movements were selected on the basis of internal cues. Modified from Mushiake *et al.* (1991).

Like neurons in the prefrontal cortex (see above), if the monkey is also taught to wait for many seconds after seeing the red or green light flash, many PMdc neurons will discharge continuously through the delay period after the red (or green) flash until the movement finally is made.

Neurons in the PMdc also participate in learning the arbitrary associations between particular abstract cues and specific movement responses. When the monkey is first presented with a novel cue, a given PMdc neuron may not discharge during the delay period. As the monkey learns by trial and error to make a particular movement in response to the new cue, however, the same PMdc neuron may gradually develop more and more delay-period discharge over several trials. Ultimately, when the monkey has learned to associate the once novel cue with a particular movement response, the PMdc neuron discharges consistently during the delay period.

The discharge properties of PMdr neurons in many ways resemble those of PMdc neurons. The PMdr receives input from the dorsolateral prefrontal cortex, however, and PMdr neurons show more learning and rule-dependent modulation than PMdc neurons. The sector of PMdr closest to the midline participates in selection of eye movements, and is commonly referred to as the supplementary eye field (SEF). Neurons in the SEF discharge when the response to be made is a saccadic eye movement, and they show typical directional selectivity during instructed delay periods. Moreover, as the subject learns to associate a particular instructional cue with a particular saccade direction, individual SEF neurons will develop delay-period discharge selective for that direction. When saccades are made to particular locations on visual objects, SEF neurons discharge selectively for saccades to certain locations on the object, regardless of where the object is in space, indicating an object-centered reference frame.

On the medial aspect of the hemisphere, the SMA and pre-SMA appear most involved in selecting movements based on internally remembered sequences (Figure 4). The SMA neurons are more active when a monkey touches buttons in a remembered sequence than when the monkey touches the same buttons following their illumination (Mushiake *et al.*, 1991). These neurons also may be active preceding a particular movement element of a sequence. The pre-SMA neurons, in contrast, are more likely to be active as the monkey shifts from performing one sequence to performing another.

The Primary Motor Cortex: Executing Voluntary Movements

The primary motor cortex (M1) receives corticocortical inputs from secondary motor areas described above, additional corticocortical input from the primary somatosensory cortex (S1, or areas 3, 1, and 2), area 5, and strong inputs from the basal ganglia and cerebellum arriving via the ventrolateral thalamic nuclei. The primary motor cortex represents the final stage of cortical processing before information controlling movement is delivered via the corticospinal pathway to the spinal interneurons and motoneurons most directly in control of muscle contractions. In addition to the direct corticospinal pathway, two indirect pathways – rubrospinal and reticulospinal – convey M1 output to the spinal cord. The red nucleus (origin of the rubrospinal pathway) and the pontomedullary reticular formation (origin of the reticulospinal pathway) each receive inputs from descending M1 axons.

The primary motor cortex shows an overall organization according to major body parts; i.e. a somatotopic organization. In the M1 of primates, including humans, the face is represented laterally, the lower extremity (or hindlimb and tail) medially, and the upper extremity (or forelimb) in between (Figure 5). Lesions on the lateral convexity of human M1 cause weakness or paralysis of the contralateral face, more medial lesions on the convexity affect the contralateral hand and arm, and lesions on the medial wall of the hemisphere affect the leg and foot. Distal parts of the extremities and acral parts of the face (lips and tongue) are most heavily represented caudally in M1, whereas proximal parts of the extremities and axial movements are most heavily represented rostrally. Those parts of the body that are used for fine manipulative movements (such as lips, tongue, and fingers in primates) are generally represented over more cortical territory than body parts used in gross movements such as ambulation.

At a finer level, however, M1's somatotopic organization is not a one-to-one mapping of body parts, muscles or movements (Schieber, 2001). Within the arm representation, used here as an example because it has been studied most intensively, the cortical territory representing any particular part of the arm overlaps considerably with the territory representing nearby parts. This overlap results from three features of M1's organization. First, M1 neurons whose axons converge on any given spinal motoneuron pool are distributed over

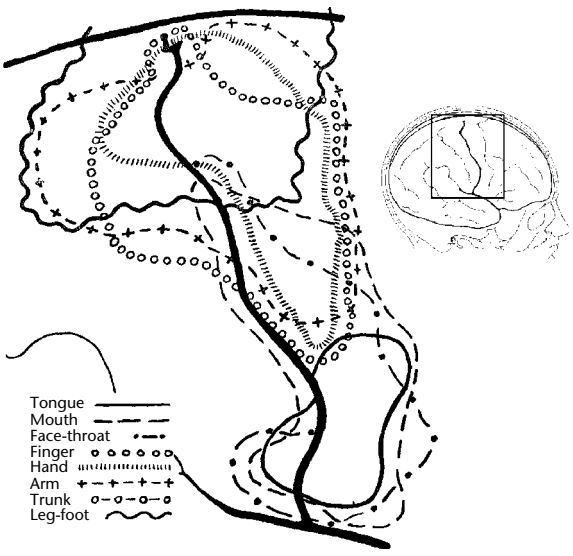


Figure 5. Organization of primary motor cortex. Stimulation of the cortical surface in patients undergoing neurosurgical procedures demonstrated that stimuli delivered laterally along the precentral gyrus evoked movements of the face; more medially stimuli evoked movements of the upper extremity; and most medially, movements of the leg. The region of the brain shown enlarged is indicated by the rectangle drawn on the inset at upper right. Modified from Penfield and Boldrey (1937).

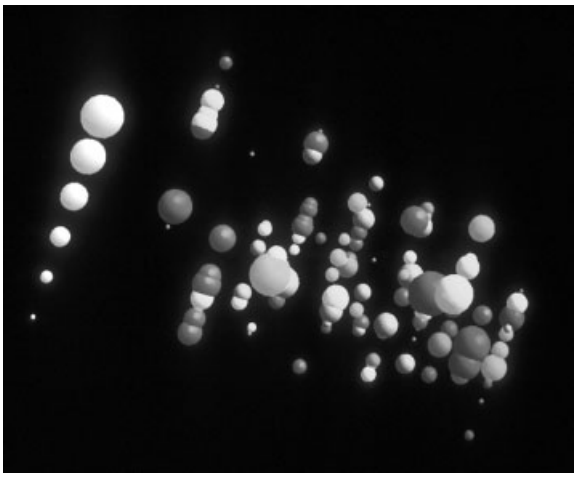


Figure 6. [*Figure is also reproduced in color section.*] Colored spheres each represent a single neuron recorded in the left hemisphere M1 as a monkey performed individuated movements (flexion or extension) of each right-hand digit and of the right wrist. The sphere representing each neuron is centered at the location of the recorded neuron in the anterior bank of the central sulcus, with the hemispheric surface above, white matter below, lateral to the viewer's right and medial to the left. Each sphere is sized according to its greatest change in discharge frequency during any of the movements; the white spheres at left constitute a scale from 0 to 200 spikes per second, with centers 1 mm apart. Each sphere representing a neuron is colored according to the movement for which that neuron's greatest discharge occurred: thumb, red; index finger, orange; middle, yellow; ring, green; little, blue; wrist, violet. Neurons best-related to movements of each digit or the wrist were intermingled throughout the same cortical territory, reflecting the extensive overlap of the representations of movements of different fingers. Modified from Schieber and Hibbard (1993).

a relatively large cortical territory that overlaps extensively with the cortical territory, giving rise to outputs that influence other nearby muscles. Second, single M1 neurons often have axons that diverge to innervate the motoneuron pools of multiple nearby muscles. Third, horizontal axon collaterals interlink neurons through considerable distances within the M1 arm representation. As a consequence of this convergence, divergence and horizontal interconnectivity, neuronal activity is widely distributed in the M1 arm representation even when only a single finger or one more proximal joint is moved (Figure 6) (Schieber and Hibbard, 1993). Such distributed representation provides a substrate for considerable plasticity (Nudo *et al*., 2001). The map of body representation in M1 has been shown to reorganize as normal people acquire different motor skills, as people adapt to peripheral injuries, and as patients recover from central nervous system injuries. (*See* **Reorganization of the Brain**)

The discharge of M1 neurons carries information more closely related to the kinematics and dynamics of incipient movement than the discharge of neurons in other cortical areas. Within the region representing a particular body part, M1 neuron firing rates typically begin to change approximately 50–100 ms before the corresponding movement. The firing rate of individual M1 neurons correlates with the movement direction, force exerted, joint position, and/or velocity of the relevant body part. Most M1 neurons fire at different rates for different values of any given parameter over a considerable range. The actual movement direction, for example, therefore can be specified more precisely by considering the discharge of a population of M1 neurons.

Studies have revealed, however, that M1 also participates with other cortical areas in the sensorimotor transformation from stimulus to movement. In behavioral paradigms that dissociate stimulus direction from movement direction – for example if rightward target movement cues leftward arm movement – the discharge of some M1 neurons is

more closely related to stimulus direction than to movement direction, even as the movement is occurring. Conversely, discharge related to kinematic and dynamic parameters of the movements being executed are found in other cortical areas as well, such as PMdc and parietal area 5.

CONCLUSION

We have considered how different regions of the cerebral cortex participate in various aspects of selection, guidance, and generation of voluntary movements, including deciding which responses to make, selecting which particular movements to perform, guiding these movements in space, and generating the movement *per se*.

As indicated above, many cortical regions are active during any particular movement. Techniques that examine these processes at the neuronal level in real time have revealed, however, that none of the functions described above is performed exclusively by a single region. The time required for transmission of information from one cortical region to the next is short compared with the time the nervous system uses to process input, select action, and execute movement; i.e. reaction time and movement time. Hence during any given response, neurons in many of the cortical regions considered above are active concurrently. When a monkey suddenly sees a morsel of food and reaches out to pick it up, for example, the discharge of dorsolateral prefrontal neurons may participate in the decision to get the food, while almost simultaneously, certain posterior parietal, premotor and primary motor cortex neurons participate in transporting the arm toward the morsel, and other parietal, secondary and primary motor cortex neurons participate in adjusting the configuration of the hand to grasp the morsel. Complex cooperation of multiple motor areas in the cerebral cortex thus underlies the apparently simple act of picking up a raisin.

References

Battaglia-Mayer A, Ferraina S, Mitsuda T *et al*. (2000) Early coding of reaching in the parietooccipital cortex. *Journal of Neurophysiology* **83**: 2347–2391.

Funahashi S (2001) Neuronal mechanisms of executive control by the prefrontal cortex. *Neuroscience Research* **39**: 147–165.

Hyvärinen J and Poranen A (1974) Function of the parietal association area 7 as revealed from cellular discharge in alert monkeys. *Brain* **97**: 673–692.

Iriki A, Tanaka M and Iwamura Y (1996) Coding of modified body schema during tool use by macaque postcentral neurones. *NeuroReport* **7**: 2325–2330.

Kalaska JF, Cohen DAD, Prud'homme M and Hyde ML (1990) Parietal area 5 neuronal activity encodes movement kinematics, not movement dynamics. *Experimental Brain Research* **80**: 351–364.

Mountcastle VB, Lynch JC, Georgopoulos A, Sakata H and Acuna C (1975) Posterior parietal association cortex in monkey: command functions for operations within extrapersonal space. *Journal of Neurophysiology* **38**: 871–908.

Murata A, Gallese V, Luppino G, Kaseda M and Sakata H (2000) Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *Journal of Neurophysiology* **83**: 2580–2601.

Mushiake H, Inase M and Tanji J (1991) Neuronal activity in the primate premotor, supplementary, and precentral motor cortex during visually guided and internally determined sequential movements. *Journal of Neurophysiology* **66**: 705–718.

Nudo RJ, Plautz EJ and Frost SB (2001) Role of adaptive plasticity in recovery of function after damage to motor cortex. *Muscle and Nerve* **24**: 1000–1019.

Penfield W and Boldrey E (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* **37**: 389–443.

Schieber MH (1999) Voluntary descending control. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 931–949. San Diego, CA: Academic Press.

Schieber MH (2001) Constraints on somatotopic organization in the primary motor cortex. *Journal of Neurophysiology* **86**: 2125–2143.

Schieber MH and Hibbard LS (1993) How somatotopic is the motor cortex hand area? *Science* **261**: 489–492.

Taira M, Mine S, Georgopoulos AP, Murata A and Sakata H (1990) Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Experimental Brain Research* **83**: 29–36.

White IM and Wise SP (1999) Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research* **126**: 315–335.

Further Reading

Ashe J (1997) Force and the motor cortex. *Behavioural Brain Research* **86**: 1–15.

Bock G, O'Connor M and Marsh J (eds) (1987) *Motor Areas of the Cerebral Cortex.* Ciba Foundation Symposium 132. Chichester, UK: John Wiley.

Boussaoud D, di Pellegrino G and Wise SP (1995) Frontal lobe mechanisms subserving vision-for-action versus vision-for-perception. *Behavioural Brain Research* **72**: 1–15.

Hepp-Reymond MC (1988) Functional organization of motor cortex and its participation in voluntary movements. In: Seklis HD and Erwin J (eds)

Comparative Primate Biology, pp. 501–624. New York, NY: Alan R Liss.

Jeannerod M, Arbib MA, Rizzolatti G and Sakata H (1995) Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* **18**: 314–320.

Kalaska JF and Crammond DJ (1992) Cerebral cortical mechanisms of reaching movements. *Science* **255**: 1517–1523.

Lacquaniti F and Caminiti R (1998) Visuo-motor transformations for arm reaching. *European Journal of Neuroscience* **10**: 195–203.

Passingham RE (1996) Attention to action. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **351**: 1473–1479.

Rizzolatti G, Luppino G and Matelli M (1998) The organization of the cortical motor system: new concepts. *Electroencephalography and Clinical Neurophysiology* **106**: 283–296.

Wise SP, Boussaoud D, Johnson PB and Caminiti R (1997) Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. *Annual Review of Neuroscience* **20**: 25–42.

Multiple Sclerosis

Introductory article

KJ Smith, King's College, London, UK
MK Sharief, King's College, London, UK

CONTENTS

Multiple sclerosis is a neurological disorder characterized by symptoms that either wax and wane or slowly progress, in which immune cells invade the central nervous system and damage nerve fibers, thereby disrupting impulse conduction.

INTRODUCTION

Multiple sclerosis (MS) is the most common disabling neurological disease of young adults, affecting 1 individual per 500–1000 of population in northern Europe, an area with high risk. The cause of the disease is not known, but it is more common in women (the male to female ratio is approximately 2:3), and although MS is neither infectious nor inherited, people can inherit genes that predispose them to acquiring the disease. The nature and pattern of the disease in different patients is highly unpredictable and variable, and there is no known cure. Multiple sclerosis is rare in children, but it is believed to be acquired before puberty, becoming expressed later in life, most often during the third and fourth decades.

PATHOLOGY

The hallmark of MS is the presence of inflammatory demyelinating lesions within the central nervous system (CNS). The distribution of lesions is unpredictable and widespread, and many new lesions (up to a hundred in severe cases) can occur in a year, due to what is believed to be an autoimmune process. An early event in the formation of a new lesion is breakdown of the blood–brain barrier, with the passage into the brain tissue of inflammatory cells, including T cells and macrophages, many of which congregate in a cuff around the blood vessels: local microglial cells become activated. The myelin sheaths of neighboring axons are attacked and removed by macrophages, leaving the axons locally demyelinated, i.e. denuded of their insulating layer of myelin. The fate of the affected oligodendrocytes (the myelin-forming cells of the CNS) is unclear, but many appear to undergo degeneration. Astrocytes (glial cells) become activated, forming a glial (sclerotic) scar. The multiple sclerotic lesions underlie the common name of the disease, also known as disseminated sclerosis. Some lesions, particularly in the earlier stages of the disease, are repaired by remyelination, but remyelinated axons are not protected from future attack, so that demyelinated lesions accumulate over time and are common at autopsy. Apart from demyelination, affected axons can also undergo degeneration, becoming transected within the lesions. Most degenerating axons die during the inflammatory phase of lesion development, but degeneration proceeds at a 'slow burn' once the inflammation has subsided, perhaps owing to the lack of trophic support. Although attention has focused on lesions occurring within the white matter, gray-matter lesions are also common. The extent to which neuronal cell bodies undergo degeneration is unclear, but such degeneration probably occurs, and might be substantial.

ONSET AND COMMON SYMPTOMS

The initial symptoms of MS are often visual (in approximately 50% of patients) typically due to a lesion in the optic nerves, resulting in blurred or double vision, or perhaps partial or complete blindness. Another 40% of MS sufferers present with limb weakness or sensory symptoms (e.g. numbness), and the remaining patients may present with impairments of bladder or bowel function, or

blunting of the intellect. The symptoms can appear quite suddenly, usually upon waking or over a few days. The expression of such symptoms is typically only temporary (e.g. persisting for a few weeks) in relapsing–remitting MS, but it may be more permanent in progressive disease. No specific event – such as viral or bacterial infection, or trauma – can be reliably identified as preceding the onset of the disease. The semirandom distribution of new lesions means that various parts of the brain and spinal cord are affected in different patients, and therefore the range of neurological complaints expressed by patients during the course of their disease is very wide. Most lesions fortunately occur in clinically 'silent' areas of the brain (i.e. they do not cause symptoms), but lesions also occur in more 'eloquent' tracts (such as the optic nerves) where they result in functional deficits appropriate for the pathway affected. Fatigue is common, and can be the most important complaint of patients. Lesions within the cerebellum, or that interrupt the motor and sensory pathways, can cause tremor and interfere with balance and other coordinated movements, making walking difficult (ataxia) and preventing fine hand movements; speech can become slurred. Tingling sensations are common in MS, often appearing intermittently, although they may be persistent. Pain occurs in about 30% of patients at some stage of the disease process. The expression of most symptoms can be surprisingly sensitive to body temperature (warming is deleterious and cooling beneficial), underpinning the earlier 'hot bath' diagnostic test for MS. As the disease advances, spasticity of affected limbs and sexual dysfunction are not uncommon, and there may be incontinence of urine or feces.

COURSE

The course of MS is variable and unpredictable. However, the majority of people with the disease (about 70%) start with an episodic pattern of attacks, constituting a relapsing and remitting course. Clinical relapses (exacerbations) cause seemingly random impairments of neurological functions. The symptoms during these relapses may remit completely, leaving no residual deficit, or partially resulting in variable degrees of permanent impairment. The frequency of clinical relapses varies widely from 0.1 to 10 per year, but in people who are considered to have active MS, the average frequency of relapses is 1.5 per year. In general, relapses in the early stages of MS are followed by complete remission, but in the later stages relapses become less frequent and may be followed by incomplete recovery. The relapsing and remitting course eventually transforms in many patients into a 'secondary progressive' form where clinical fluctuations are replaced by a gradual accumulation of permanent neurological deficit. The remaining patients with relapsing and remitting MS may have only minimal disability even after many years. A minority of MS patients have primary progressive MS where the clinical course is progressive from the onset. In contrast, there is a mild form of MS in which people may have only minor neurological impairments that do not interfere with activities of daily living. This condition, known as benign MS, is seen in approximately 10–15% of cases. The activity of MS, and the distribution of demyelinated plaques, can be monitored by brain magnetic resonance imaging, which is also widely used as a noninvasive diagnostic tool. Although MS is not a fatal disease, the average life expectancy for MS sufferers is 7 years less than normal, or 75–85% of expected survival. The majority of deaths are due to secondary complications such as pneumonia and pulmonary embolism, but the suicide rate in MS is also increased, being more than six times that of the general population.

PATHOPHYSIOLOGY

The symptoms of MS are determined not only by the location of the individual lesions, but also by their pathology: inflammation, demyelination and axonal loss each appear to have major roles. Conduction block is perhaps the most important deficit, contributing to symptoms such as blindness, paralysis and numbness, depending upon the pathway affected. Conduction block arises from demyelination, but it appears that inflammation can also play a part, and inflammatory mediators such as cytokines have been implicated, with attention focusing on nitric oxide in particular. The conduction block and the consequent expression of symptoms may only be temporary, because inflammation can resolve and, even where demyelination persists, conduction can be restored along the demyelinated axons accompanied by changes in the distribution of sodium channels. However, in demyelinated axons, restored conduction is slower than normal (resulting in a diagnostically valuable conduction delay upon electrophysiological examination), and is also less secure, so recovery from symptoms can be incomplete. Demyelinated axons, particularly sensory ones, may not only have conduction restored, but can become spontaneously active, wrongly generating trains of impulses that can be interpreted by the brain as a tingling

sensation referred to the body part normally innervated by the axons. Demyelinated axons can also become sensitive to physical distortion, resulting in movement-induced sensations: for example, people with lesions in the spinal cord in the neck can experience tingling sensations radiating down their body when they bend their neck. Massed, synchronous spontaneous firing of axons by ephaptic (nonsynaptic) transmission might contribute to spasms. Where remyelination occurs, it is accompanied by the restoration of fast, secure conduction, contributing to the restoration of normal vision, walking, and other activities. Axonal degeneration contributes to the permanent loss of function, although the persisting deficit may be ameliorated by compensatory (plastic) changes in the brain.

ETIOLOGY

The cause of MS is unknown. The disease is more common in temperate climates, and people of northern European descent are more likely to develop MS than people from most other ethnic backgrounds. Epidemiological studies have identified two principal etiological factors: a genetically determined susceptibility, and (probably) exposure to environmental agents. Analysis of migration patterns has shown that MS is acquired before an age around puberty, perhaps due to an environmental agent such as climate, diet, sunlight or toxins, although many believe that the agent is likely to be infective, most probably viral. However, direct evidence for a viral cause of MS is still lacking, and if viruses have a role it may be by initiating autoimmunity against CNS antigens. On the other hand, genetic studies have suggested multiple genetic influences on the development of MS. To date, the best-characterized susceptibility-associated gene has been mapped to the human leukocyte antigen (HLA) complex, located on chromosome 6. Different HLA associations have been identified in MS. The HLA-DR2, DR(1*1501), DQ(1*602), DQA102 and DW2 haplotypes are frequently associated with MS. The role of genetic factors has been confirmed by family studies, which have shown that the risk of developing MS increases significantly if other family members are affected. While the risk of developing MS in the general population in the UK is about 1 in 800, first-degree relatives of people with MS have a more than 1 in 100 risk of developing the disease. However, it is important to note that MS is not directly inherited, and that more than 80% of people with the disease do not have affected first-degree relatives.

AUTOIMMUNE DISEASE

Multiple sclerosis is thought to be mediated by autoimmune attacks due to a dysregulated immune system that allows T cells to become activated against brain antigens, enter the CNS and mediate tissue damage through inflammatory demyelination and axonal damage. Abnormalities of the immune system in MS have been detected locally within the brain and cerebrospinal fluid, and also systemically in the peripheral circulation. The exact target of the autoimmune response in MS remains unclear, but may involve components of the myelin layer that surrounds nerve fibers. It is also possible that such antibodies are a consequence of the disease, rather than part of its cause. However, if they are part of its cause one explanation for this autoimmune reactivity against myelin antigens could be molecular mimicry, where a shared molecular structure that exists between a viral protein and a normal human CNS protein might cause an antiviral immune response to mediate CNS damage. Another possibility is that autoimmunity may result from stimulation of T cells by molecules known as superantigens, which interact with a large number of immune cells. Such superantigens, often of viral or bacterial origin, may bind to some T-cell receptors and produce nonspecific stimulation of a large number of T cells. This stimulation may cause expansion of T cells reactive to myelin components, and subsequent demyelination.

ANIMAL MODELS

Multiple sclerosis does not occur in animals, but research has been greatly advanced by the study of animal models that mimic different aspects of the disease. The immunology underlying MS has been studied using experimental autoimmune (allergic) encephalomyelitis (EAE) induced in animals, most commonly rats and mice, by immunization with myelin or myelin components (active EAE), or by administration of T cells activated against myelin components (passive EAE). In either case T cells enter the CNS and attack the myelin sheath, forming scattered lesions that closely resemble those of MS. Affected animals can develop a weak or paralyzed tail and hindlimbs for a few days, but these deficits typically resolve within a week or two. New drugs for the therapy of MS are typically developed and tested using EAE models. Animal models have also been essential for the development of transplantation techniques to introduce cells capable of repairing lesions by remyelination,

and trials of such techniques in humans with MS have now commenced.

TREATMENT

There is no cure for MS, and no treatment completely halts the disease process. However, many of the symptoms can be treated. Antidepressants and antiepileptic drugs may reduce painful sensory symptoms, numbness or fatigue. Fatigue and spasticity may also be relieved by drugs such as 3,4-diaminopyridine, amantadine or methylphenidate for fatigue, and baclofen, tizanidine, diazepam or dantrolene for spasticity. Urinary symptoms may be relieved by drugs that relax the bladder muscles. In addition to symptomatic therapies, the course of MS can be modified by several strategies. Generalized immunosuppression may provide modest reduction of MS activity. Steroids may ameliorate many of the symptoms during clinical relapses by reducing swelling and inflammation, but do not alter the course of the disease. A more specific immunosuppression (or immunomodulation) in MS can be achieved with interferon beta, which reduces relapses by about one-third, probably by reversing some of the T-cell defects, or reducing the ability of T cells to invade the central nervous system. Glatiramer acetate also reduces relapses by about one-third, probably by suppressing immune responses to brain antigens. Experimental therapies now being tested in people with MS include specific targeting of immune cells or inflammatory mediators, stem cell transplantation, and strategies to protect nerve fibers from degeneration.

Further Reading

Blakemore WF and Franklin RJM (2000) Transplantation options for therapeutic central nervous system remyelination. *Cell Transplantation* **9**: 289–294.

Compston A, Ebers G, Lassmann H *et al*. (eds) (1998) *McAlpine's Multiple Sclerosis*, 3rd edn, pp. 145–190. Edinburgh, UK: Churchill Livingstone.

Kalman B and Lublin FD (1999) The genetics of multiple sclerosis. A review. *Biomedical Pharmacotherapy* **53**: 358–370.

Lassmann H (1998) Neuropathology in multiple sclerosis: new concepts. *Multiple Sclerosis* **4**: 93–98.

McDonald WI, Compston A, Edan G *et al*. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Annals of Neurology* **50**: 121–127.

Noseworthy JH (1999) Progress in determining the causes and treatment of multiple sclerosis. *Nature* **399**: A40–A47 (supplement).

O'Connor KC, Bar-Or A and Hafler DA (2001) The neuroimmunology of multiple sclerosis: possible roles of T and B lymphocytes in immunopathogenesis. *Journal of Clinical Immunology* **21**: 81–92.

Rudick RA (1999) Disease-modifying drugs for relapsing-remitting multiple sclerosis and future directions for multiple sclerosis therapeutics. *Archives of Neurology* **56**: 1079–1084.

Smith KJ and McDonald WI (1999) The pathophysiology of multiple sclerosis: the mechanisms underlying the production of symptoms and the natural history of the disease. *Philosophical Transactions of the Royal Society of London Series B* **354**: 1649–1673.

Navigation and Homing, Neural Basis of

Introductory article

Verner P Bingman, Bowling Green State University, Bowling Green, Ohio, USA

CONTENTS

Navigational ability in animals often relies on memory-based maps of environmental space, typically assisted in birds by external 'compass' clues and in rodents by internal positioning information represented by specialized neurons such as place cells. The hippocampus and related brain structures appear to be integral to the ability to navigate.

INTRODUCTION

During the course of animal evolution, the parallel emergence of the ability to move in space and a structured nervous system to coordinate that movement has contributed to the complexity of brain and behavior observed in modern animals. Virtually all animals who move can do so in a goal-directed manner. Such behavior may be simple and reflexive, for example the positive phototaxis displayed by a hermit crab as it moves toward a light source. Alternatively, such behavior can be complex and learned, for example the map-like spatial memory of a homing pigeon used to navigate home from unfamiliar locations tens of kilometers away.

Broadly defined, navigation is the ability to move to a goal location. This article, however, is specifically concerned with two types of navigation and their neural basis. One is the use of a memorized representation of environmental stimuli, something like a map of the spatial distribution of landmarks, to locate a goal. The other is the use of memories of stimuli generated by the animal's own movement, 'path integration', to locate a goal. Although the following discussion focuses on birds and mammals, memory-based navigational ability is characteristic of all vertebrates and many invertebrate species as well.

HOMING IN BIRDS

The term 'homing' refers to the ability of an animal to return to a familiar location (home) after a journey. Although in birds homing is usually associated with pigeons returning to their loft after being displaced to some distant, unfamiliar location, the term can easily apply to a migratory bird returning to a previous breeding site or overwintering area. In the case of homing pigeons, successful homing is based primarily on three distinct spatial mechanisms that resemble the map and compass of human navigation.

To identify its location relative to home and to return home from distant, unfamiliar locations, homing pigeons use a 'navigational' map. This map is often a memory representation of the distribution of atmospheric odors in the environment, with differences in the distribution of odors serving to uniquely define different locations in space. Similarly, from familiar areas, homing pigeons can additionally navigate using memory of the spatial distribution of previously experienced landmarks, identified not just visually but using other senses as well, that provide a stable spatial reference. This 'landmark' map resembles most what is often referred to in the scientific literature as a 'cognitive' map.

The navigational map and landmark map together enable a pigeon to determine its location in space relative to the home loft. This map information must then be used to compute a compass direction, which the pigeon would rely on to orient itself in a homeward direction. In birds, two sources of compass or directional information are known to be used: the sun, and the earth's magnetic field. These compasses can tell a bird nothing about

where it is, but are crucial in enabling a bird to orient in an appropriate direction – for example north, once the bird has determined it is south of home based on map information. Finally, compass information is probably also used to provide a directional framework for learning about the spatial relationship among atmospheric odors and landmarks that leads to map learning. In other words, pigeons, and probably many other vertebrate species, are thought to exploit the directional reference provided by compasses to learn the memory-based, map-like representations of space necessary for navigation.

DEAD RECKONING, PLACE CELLS, AND HEAD DIRECTION CELLS IN NAVIGATION

Dead Reckoning

Birds are generally diurnal creatures with good eyesight who travel large distances over open countryside, explaining in part why they rely heavily (albeit not exclusively) on the sun and visual landmarks for navigation. Most rodents, in contrast, have poor vision, are nocturnal, and tend to move in natural labyrinths or take paths close to physical boundaries such as walls. It is not surprising, therefore, that although animals such as laboratory rats can certainly use visual landmarks to locate a goal, they also rely heavily on 'dead reckoning' for navigation.

Dead reckoning, or path integration, is a form of navigation in which an animal integrates information about its own motion to locate its present position or return to a starting location. The source of the self-motion information can come from the vestibular system, efference copy from movement commands (information from ongoing motor neuron activity sent to other regions of the nervous system) and other sensory information created by an animal's movement.

Dead reckoning is a reliable source of navigational information over short distances (a few meters), but is generally believed to be prone to the accumulation of errors, and therefore less effective over longer distances. However, this generalization does not preclude the possibility that some animal species have evolved more accurate path integration systems that could support navigation for longer distances. Just like the sun compass of birds, dead reckoning in rodents is thought to serve also as a spatial framework or reference for learning about the spatial relationship among landmarks, which would ultimately lead to the acquisition of a landmark map. Once learned, such a landmark map would provide rodents with a navigational mechanism to locate a goal from distances beyond those for which dead reckoning would be reliable.

Place Cells

The presence of map-like representations of environmental space used by birds and mammals for navigation raises the fundamental question of how these representations are structured in the brain. Although numerous brain regions participate in the representation of space, one structure, the hippocampus, seems to have a particularly important role.

In both birds and mammals, experimental interference with the normal operation of the hippocampus seriously impairs their ability to navigate using map-like representations of space. It should be emphasized that interfering with the hippocampus does not disrupt all spatial ability, but specifically navigation based on map-like representations, and possibly also some spatial behavior mechanisms that support the learning of such map-like representations.

For example, while interference with the hippocampus does not impair sun compass orientation in birds, it has been reported to impair dead reckoning or path integration in rodents. This suggests that in birds the hippocampus may exploit sun compass information for learning a map of local landmarks, but is not essential to the bird's orientation by the sun. In contrast, in rodents the hippocampus exploits the use of path integration information in learning a map of local landmarks, and it may also participate in path integration itself. However, it remains controversial whether the hippocampus is necessary for path integration in rodents.

At the cellular level, how is spatial information in the environment coded in the activity of single neurons? Several different types of neurons in the rodent hippocampus display what are called 'place fields'. Specifically, in a freely moving rat, a typical hippocampal neuron with a place field would display low levels of spontaneous activity (occurrence of action potentials) in much of a familiar local environment. However, that neuron would show a large increase in activity when the rat passes through a particular small part of the local environment – the neuron's place field. Such hippocampal neurons are called 'place cells'. As a rat moves through a familiar environment, the change in activity across a large number (network) of place

cells, varying depending on where the rat is and where it is moving, is believed to keep the rat continually updated on where it is in space. Different place cells would show higher levels of activity in different portions of the environment, and collectively the thousands of place cells in the hippocampus would structure the map-like representations of environmental space used for navigation.

However, although the properties of place cells have been exhaustively studied in rats, it is uncertain whether there are similar hippocampal place cells in birds and even primates. Preliminary research in birds and primates suggests that the activity of some hippocampal neurons changes as a consequence of space, but in ways somewhat different from those found in rats. For example, in primates, and maybe pigeons, it seems that where an animal is looking, rather than its specific location in space, is more important in determining whether a place cell is active or not. The suspected differences in the properties of place cells in primates and birds compared with rats may be related to the highly visual, diurnal lifestyle of the former groups.

Head Direction Cells

In addition to place cells, some brain areas closely related to the hippocampus contain neurons that are preferentially active when a rat is headed or pointed in a particular direction, relative to some directional reference, independent of where it is located in space. Such neurons are referred to as 'head direction cells'. These cells are thought to work together with place cells in structuring the neural representation of space in the hippocampus.

For example, one possibility in rats is that head direction cells provide essential velocity information about an animal's movement in space (dead reckoning) and that this information could then be used to determine the place fields of place cells and the map-like representation of local landmarks that emerges from them.

Head direction cells have yet to be discovered in the hippocampus of birds, but if they were found, it would not be surprising to discover that as directional references they were sensitive to compass information gathered from the sun or earth's magnetic field.

SPATIAL BEHAVIOR IN HUMANS

Although in developed countries much of human navigation is now structured by the grid-like pattern of streets and the routes people take through them, nomadic people who roam freely in open environments navigate by map-like representations seemingly similar to the landmark maps of birds and nonhuman mammals.

Consistent with the idea that humans too learn maps of their environment is the finding that the human hippocampus is an essential brain structure for spatial memory. Individuals who have experienced damage to the hippocampus, perhaps because of stroke or surgical interventions to control epilepsy, are strikingly impaired in navigating in their neighborhoods and remembering where they left objects, much like patients with Alzheimer disease who also experience hippocampal dysfunction. Some of this work further suggests that the hippocampus of the right hemisphere may be more important for spatial memory than the left hippocampus.

The development of powerful brain imaging techniques, such as positron emission tomographic scans and functional magnetic resonance imaging, has revolutionized the ability to investigate regional differences in brain activity that correlate with ongoing cognitive activity in healthy people. With respect to spatial memory tasks, whether people are asked to navigate a virtual maze or taxi drivers are asked to describe a route to an address, the parahippocampal region invariably shows an increased level of activity. Although the hippocampus itself may or may not show higher levels of activity during a spatial memory task, the heightened activation found in the neighboring parahippocampal area, a brain region with strong neural connections to the hippocampus, indicates that a circuit of brain structures that includes the hippocampus participates in memory-based navigation. The imaging work also is consistent with the idea that the right hippocampal area may be more involved in spatial memory than the left.

Human navigation is complex, and it is simplistic to attribute such a complex behavior to a restricted cluster of brain areas like the hippocampus and parahippocampus. The imaging work has identified other brain structures that show heightened activity when people engage in tasks of spatial memory. One interesting brain area is the posterior parietal cortex: this region is important for egocentric orientation, in other words the ability to relate the location of objects in space to one's own body axis. In contrast, the cognitive map-like hippocampus is important for understanding the spatial relationship among environmental stimuli independent of where one is located in space. Imaging data suggest that successful memory-based

navigation requires the integration of the posterior parietal lobe representation of egocentric space (where objects are relative to oneself) with the hippocampal representation of allocentric space (the map of environmental landmarks), which does not change as one moves through space.

COMPUTATIONAL MODELS

The research described above allows attribution of a memory-based, map-like representation of the environment to the hippocampus and related structures. Recordings of the electrical activity of many hippocampal cells have revealed that these cells have place fields, which may represent specific locations in the environment.

However, the map of the environment found in the hippocampus cannot be reduced to a collection of independent single neurons, each responsible for the representation of a restricted portion of space. Rather, at any given time, it is logically assumed that space and navigating through space are represented by the activity in networks of thousands of neurons in the hippocampus and related structures. Unfortunately, investigating the activity in networks of thousands of neurons is beyond the techniques available to neuroscience today, yet ultimately the neural representation of space will have to be understood at this level.

Computational models, most often neural network or connectionist models, are designed to compute a behaviorally interpretable output function based on input variables and system parameters. In the case of understanding how the hippocampus structures memory representations of space, the behaviorally interpretable output could be the place field of a place cell, the input variables could be the network of neurons that modulate the place field of a place cell, and the system parameters could be the connection strengths among the elements in the network.

The properties of neural network and connectionist computational models resemble the properties of networks of neurons in the brain. This includes experience-dependent changes in the strength of connections between elements in a neural network or connectionist model that resemble changes in synaptic strength occurring among neurons in the brain during learning. Consequently, computational models have been successfully exploited to explain how the hippocampus can represent space at the level of networks of neurons.

Computational models are powerful research tools because they can capture the complexity of neural networks in the brain, and make explicit predictions about how the hippocampus and related structures may represent space. In the absence of techniques to measure the simultaneous activity of thousands of neurons, artificial computational models are important in elucidating the complexity of interactions among networks of neurons and their relationship to behavior and cognition.

Further Reading

Alerstam T (1990) *Bird Migration*. New York, NY: Cambridge University Press.

Balda RP, Pepperberg IM and Kamil AC (eds) (1998) *Animal Cognition in Nature*. San Diego, CA: Academic Press.

Berthold P (ed.) (1991) *Orientation in Birds*. Basel, Switzerland: Birkhäuser.

Best PG, White AM and Minai A (2001) Spatial processing in the brain: the activity of hippocampal place cells. *Annual Review of Neuroscience* **24**: 459–486.

Burgess N, Jeffery, KJ and O'Keefe J (eds) (1999) *Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford, UK: Oxford University Press.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Healy S (ed.) (1998) *Spatial Representation in Animals*. Oxford, UK: Oxford University Press.

Hippocampus Special Issue on Place Cells (1999) *Hippocampus* **9**(4).

McNaughton B, Barnes CA, Gerrard JL *et al*. (1996) Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *Journal of Experimental Biology* **199**: 173–186.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford, UK: Oxford University Press.

Redish AD (1999) *Beyond the Cognitive Map. From Place Cells to Episodic Memory*. Cambridge, MA: MIT Press.

Rumelhart DE and McClelland JL (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press.

Taube JA (1998) Head direction cells and the neurophysiological basis for a sense of direction. *Progress in Neurobiology* **55**: 225–256.

Neglect

Introductory article

Jon Driver, University College London, London, UK
Patrik Vuilleumier, University College London, London, UK

CONTENTS

Unilateral spatial neglect is the failure to perceive, orient, and act on objects or events towards the side of space opposite a unilateral brain lesion (typically to the right hemisphere). These characteristic deficits are not attributable to primary sensory or motor loss, but involve deficits in higher-level spatial representation and attention.

INTRODUCTION

Unilateral spatial neglect is a common and disabling syndrome following unilateral stroke, especially after infarction of the right middle cerebral artery. The patients typically exhibit a constellation of deficits, but their tendency to ignore information towards the contralesional side of space (i.e. usually the left side) is particularly striking, especially given that this can arise even in patients who have no primary sensory or motor deficit on the affected side. Recent progress in understanding such neglect has come from relating the disorder to the underlying neural systems that are damaged or preserved, and also from relating certain aspects of neglect to 'attentional' phenomena in normal people, who can analogously fail to consciously perceive or respond to stimuli if they do not attend to them.

CLINICAL SIGNS

Although the exact pattern of deficit varies from one patient to another, the following clinical signs may be present in prototypical cases. The patients behave as if half of their spatial world were lost (usually the left side, following right hemisphere damage). They may ignore people or objects on the affected side of space, eat food from only the other side of their plate, shave or make up only one side of their face, dress only one side of their body, miss letters, words or even a whole page on the affected side when reading, and overlook or forget contralesional turns in routes. Notably, many patients with neglect often do not realize that they are missing anything on the affected side (anosognosia). If the brain damage causes paralysis of contralesional limbs, the patients may even remain unaware of this if they also suffer from spatial neglect for that side of space.

In the clinic, neglect is not only observed in daily life, but also demonstrated with a variety of simple paper-and-pencil tests. For example, when patients are required to search for and mark all target shapes on a page (cancelation task), they may mark only a few of those on the ipsilesional side and ignore the presence of the others even when given unlimited time (Figure 1a). When asked to mark the midpoint of a horizontal line (bisection task), some patients may deviate towards its ipsilesional end, as if ignoring or compressing its contralesional extent (Figure 1b). When drawing from memory or copying a picture, they may omit most details from the contralesional side (Figure 1c), even for familiar objects with important unique features on either side that have a prescribed layout (e.g. the numbers on a clock face). Such spatial biases do not arise only in visually guided behavior. Even when blindfolded and exploring objects haptically, the patients may search only for those on the ipsilesional side. When asked to point straight ahead in darkness, or to move a laser-pointer to the subjective straight ahead position, neglect patients may deviate ipsilesionally. When searching for invisible targets in darkness, their eye movement path may show an analogous bias towards the ipsilesional side. Sound localization may also be affected, with inaccurate spatial coding on the contralesional side, and/or mislocations towards the ipsilesional side.

Importantly, such failures to detect, perceive, localize, explore or act towards the affected side

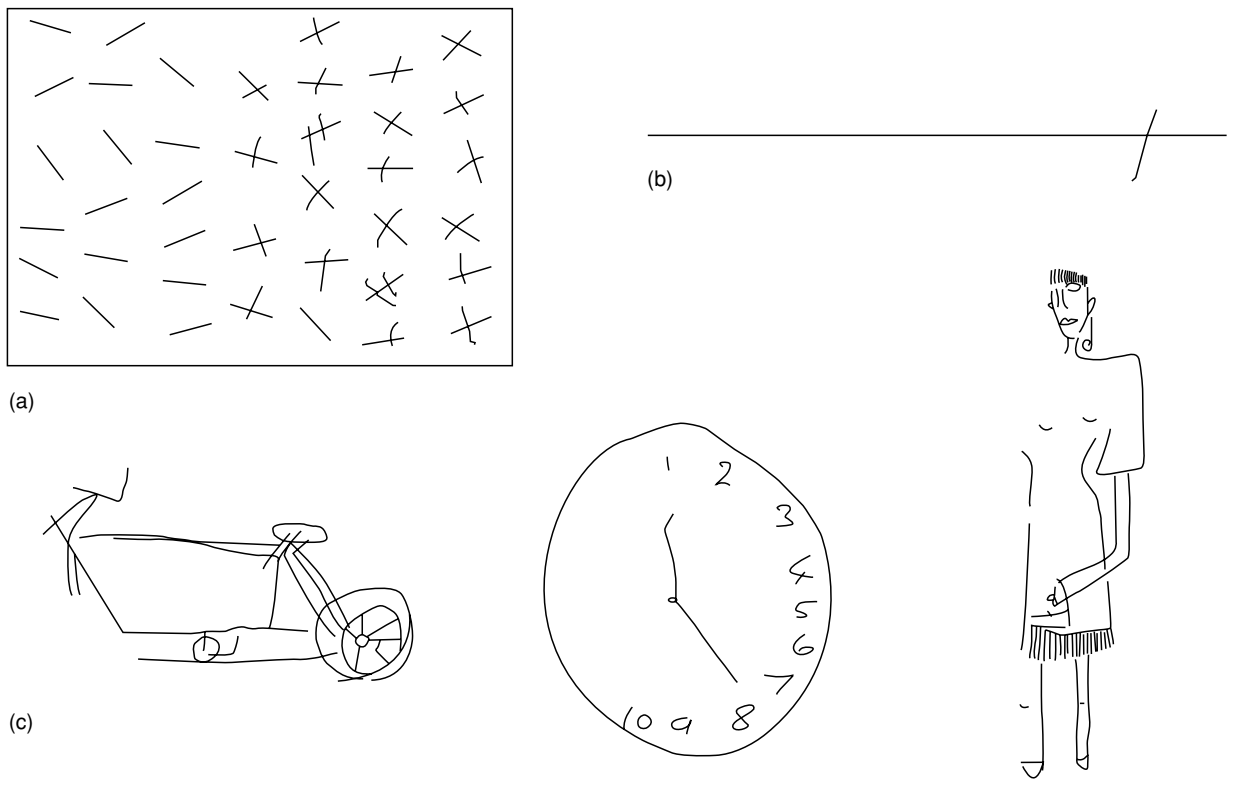


Figure 1. Examples of left-sided neglect after damage to the right hemisphere of the brain, in clinical paper-and-pencil tests. (a) Cancelation; (b) line bisection; (c) drawings of bicycle, clock and woman.

of space in neglect patients are typically not attributable to primary sensory or motor loss. Many patients who do suffer from such primary losses do not exhibit neglect; conversely, other patients can exhibit neglect despite having no paralysis, no visual field cut, and no other primary sensory loss. This is further illustrated by another common component of the syndrome. Perceptual 'extinction' arises in patients who can detect and report a unilateral stimulus presented alone on the affected side of space, but who miss the same stimulus if presented concurrently with another detected stimulus on the ipsilesional side. The ipsilesional stimulus 'extinguishes' awareness of a contralesional stimulus that would otherwise be detected. Such extinction can arise within vision, touch, audition, or even smell. It can also be elicited crossmodally, as when a visual stimulus on the ipsilesional side extinguishes awareness of a touch on the contralesional hand that would otherwise have been felt. It has been suggested that extinction may reflect a pathological bias in attention, which becomes most apparent when multiple concurrent stimuli compete to attract attention. Note that multiple stimulation tends to be the rule rather than the exception in daily life; and that competition between multiple stimuli may also be involved in many other tests for neglect (e.g. cancelation tasks).

Unlike sensory deficits, neglect is often reduced if the patient's attention is directed towards the affected side, even in the presence of competing stimuli. Moreover, neglect is often graded (becoming increasingly severe for increasingly contralesional locations), without the step-function at anatomical midlines that characterizes lower-level sensory disorders (e.g. hemianopia). Finally, unlike sensory disorders, neglect can be influenced by changes in posture. The same retinal stimulation in the contralesional visual field may be detected or neglected, depending on the current direction of the eyes in their orbits, and/or of the head on the trunk. Remarkably, neglect patients may see objects in the left visual field better when just their trunk is twisted leftwards (so that the same retinal stimulation now falls further to the right of their body). Analogously, tactile detection may improve for the contralesional hand if this is placed further towards the ipsilesional side of the body.

In some patients neglect can arise in tasks of mental imagery or of memory, again demonstrating some independence from sensory deficits. The

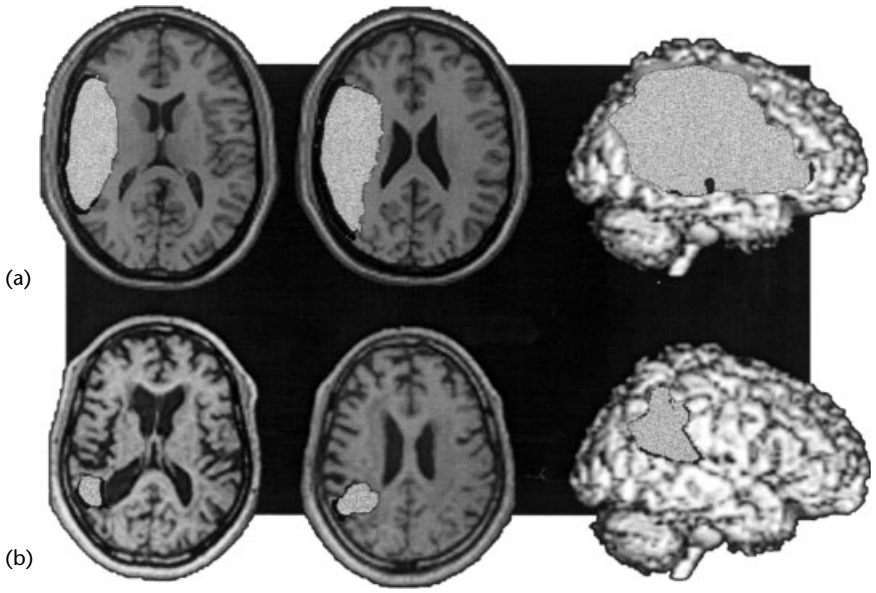


Figure 2. Reconstructions of typical lesions in neglect. (a) Damage after a massive stroke involving middle and anterior cerebral arteries in the right hemisphere, in a patient with enduring left neglect and extinction. Many combined deficits would be expected after such diffuse damage; these different deficits may interact and exacerbate each other to produce the clinically observed syndrome. (b) More focal damage centered on the right inferior parietal lobule, in another patient with enduring left neglect and extinction. Although the lesion is smaller, it will still affect many distinct parietal areas, each with subtly different functions, to produce a combined deficit. Note that underlying white matter may also be damaged, potentially affecting other areas via remote connections.

patients may fail to describe elements that would be situated on the contralesional side for well-known settings or imagined scenes. Moreover, when asked to adopt a different perspective in their mind's eye (e.g. the view from the other side of a familiar city square), the previously omitted details may now be reported, while previously reported details now shifted to the neglected side become missed instead. In some cases, neglect in mental imagery can apparently arise without ostensible neglect for externally present stimuli, and vice versa.

ANATOMICAL CONSIDERATIONS

Neglect can arise after focal damage in a number of different brain areas, all reciprocally interconnected in a distributed network that is considered critical for the elaboration of high-level multimodal spatial representations, for the control of spatial attention, and for the spatial preparation of intentional motor responses. Although some aspects of neglect can be seen in the immediate aftermath of left hemisphere damage, neglect is more common and much more enduring after right hemisphere damage. Prototypically, severe neglect is seen after major strokes in the territory of the right middle cerebral artery, causing damage to numerous brain areas and also to underlying white matter that affects remote connections. Such large lesions (Figure 2a) may cause numerous deficits, each of which can potentially exacerbate the others to produce florid neglect symptoms. Severe neglect can also follow more focal damage centered around the sylvian fissure, involving right inferior parietal cortex (supramarginal and angular gyri, Brodmann areas 39 and 40; Figure 2b). However, even such relatively focal lesions may affect many distinct parietal areas, and also disrupt fiber connections in the paraventricular white matter. It has been suggested that the superior temporal cortex may have a critical role in spatial neglect, although this remains controversial.

Other areas implicated include the prefrontal cortex (either close to the frontal eye field or in the more ventral inferior frontal gyrus), medial frontal cortex (supplementary motor and cingulate areas), thalamus (anteromedial nuclei and posterior pulvinar) and basal ganglia. It has been suggested that subcortical lesions might produce neglect by disconnection or functional disruption of distant parietal areas. Parietal lesions themselves can cause remote dysfunction in frontal and cingulate areas, and this has been found to correlate with neglect

severity. Even though neglect is characterized by a dramatic loss of conscious perception and of goal-directed action towards the affected side of space, the typical brain lesions are quite remote from primary sensory and motor cortices. This further underlines the point that patients can exhibit severe neglect despite not being blind, deaf, insensitive or paralyzed. For instance, patients can show severe visual neglect despite primary (striate) and secondary (extrastriate) visual cortical areas in the occipital and temporal lobe remaining structurally intact.

Some aspects of neglect have been related to neurophysiological data on parietal, premotor or prefrontal cortex, as gleaned from single cell recordings in monkeys. Particularly relevant properties of certain intraparietal neurons include the following: convergence of spatial information from different modalities; modulation of sensory responses (e.g. to stimuli at a particular point on the retina) by current posture; some ipsilateral receptive fields in addition to a predominance of contralateral receptive fields; highly selective responses to currently attended or salient stimuli; and involvement in the initial translation of sensory information into particular spatial motor responses. Computational models have shown that all of these aspects of parietal neurons may help to explain seemingly disparate aspects of neglect (specifically, its often multimodal nature; its modulation by posture; its graded nature; its attentional aspects; and the combination of perceptual and motoric biases; see below). Nevertheless, it remains controversial whether brain lesions in monkeys can produce deficits closely mimicking human neglect. Extinction-like deficits, as well as biases in search and exploration, may follow unilateral destruction of structures associated with human neglect, or reversible chemical lesions within specific intraparietal subareas. However, these deficits seem relatively short-lived and less severe than the human syndrome, perhaps reflecting the much larger size of damage in stroke patients (see Figure 2) or some laterality of function that is unique to humans.

VARIETIES OF NEGLECT: DISSOCIATIONS AND SUBCOMPONENTS

The exact pattern of deficit can differ between patients with neglect, presumably in accord with differences in the lesion. Dissociations have been reported between different tests for neglect (e.g. cancelation versus bisection or extinction); between perceptual versus motor aspects of neglect; between perceptual versus imaginal neglect; and between neglect in different spatial domains (e.g. in near within-reach space, versus far in space; or within objects versus between objects). Bisection and cancelation tasks can indeed provide contrasting measures of neglect. Difficult cancelation tests may constitute the paper-and-pencil measure correlating best (to date) with clinical severity of neglect and with function in daily life, whereas bisection tests may show substantial influences from other deficits (e.g. coexisting visual field cuts). Extinction can dissociate from neglect in some cases, although as noted earlier, there is an analogous competitive aspect to many of the standard tests for neglect.

The issue of perceptual versus motor aspects arises because many tasks in which patients show neglect not only assess perception or attention for the affected side, but also require a spatial motor response to stimuli there (as in the cancelation task). In principle, pathological neglect behavior (e.g. failures to cancel contralesional targets) could either have a perceptual/attentional basis, or instead reflect some deficit in executing movements towards the affected side (perhaps even with the ipsilesional limb), or a combination of such deficits. Some studies have sought to disentangle such components with ingenious methods for opposing the spatial direction of sensory information versus the required motor response (e.g. by means of video systems, reversing mirrors, pulley systems, or spatially reversed mouse–cursor relations in computerized studies). Such work initially suggested a relatively clear-cut distinction between 'perceptual' neglect after posterior lesions and 'motor' neglect following more anterior prefrontal lesions. However, the apparent frontal involvement may primarily be due to the unusual requirement to move away from the direction indicated by target stimuli in many of these experimental situations. Frontal lesions are known to produce general impairments in such incompatible tasks. Further work has separated perceptual and motor requirements by changing the start position and hence the direction of movement, while still reaching directly towards the same visual target. In this situation, patients with inferior-parietal (but not frontal) lesions were not only impaired at perceiving contralesional visual stimuli, but over and above this were slower to initiate movement in the contralesional direction. Such a combined perceptual-motor deficit would accord with recent views of parietal areas as sensorimotor interfaces.

Nevertheless, with appropriate tests, purely sensory and purely motor aspects of the neglect

syndrome can still be identified. For instance, tests of perceptual extinction do not require spatially directed motor responses, yet can reveal clear deficits. On the other hand, purely motor neglect may be apparent in patients who fail spontaneously to use the contralesional hand, even when this is not paretic and can be moved skillfully when prompted. 'Motor extinction' can be observed for tasks requiring unseen bimanual action (e.g. raising both hands, or making repeated movements with both, either in or out of phase), with deficient contralesional movements during bimanual but not unimanual actions. Note that this may reflect competitive aspects analogous to those for perceptual extinction, but possibly arising within different neural structures. Motor extinction or underuse of a nonparetic contralesional hand have been observed after unilateral lesions to frontal cortex, supplementary motor area, parietal cortex, basal ganglia or thalamus.

Pioneering lesion work in monkeys showed that unilateral damage to different areas of premotor cortex could produce phenomena resembling neglect or extinction for different spatial domains: within space close to the head following damage to area 6, sparing behavior for more distant stimuli; and vice versa following damage to area 8 instead. Human studies have reported that neglect is sometimes more severe for stimuli in near space than in far space (in bisection tasks), or conversely worse in far space than near space for other patients across a variety of tasks. This dissociation between near versus far neglect suggest that it may affect different types of spatial representation independently, rather than arising only within some single 'master map' of space. A similar conclusion follows from demonstrations that neglect can arise either 'within' visual objects, or 'between' visual objects. Such a dissociation may actually arise within the same patient, but affect opposite sides of space. For instance, one patient with bilateral brain damage showed neglect of the left side of individual objects (e.g. missing the initial letters in words), but neglect of whole objects on the right side of space (e.g. missing whole words on that side of a page). Importantly, this pattern was not just specific to reading, but was found across a variety of different tasks and stimuli.

The various dissociations described above indicate that neglect is a multicomponent disorder, consistent with the anatomy reviewed earlier. In addition to the different types and different domains of neglect, some further components exist that might not produce neglect in isolation, but exacerbate symptoms when combined with the other deficits. One example of this is the bias towards 'local' rather than 'global' aspects of a visual scene, associated with damage to the right temporoparietal junction. Some patients with focal damage here may exhibit local biases without neglect, whereas patients with equivalent damage in the left hemisphere contrastingly show a global bias. Because the large right hemisphere lesions in neglect will often include the temporoparietal junction, such patients are likely to suffer from local biases in addition to their perceptual, attentional and motor biases towards the ipsilesional side of space. This may greatly exacerbate their deficit: because their attention tends to lock onto local rather than global aspects of complex displays, as in cancelation tasks, they may consequently neglect a much wider area of contralesional space.

Another deficit that might apply to all locations following right hemisphere damage, yet still exacerbate spatial neglect when combined with biases towards the ipsilesional side, concerns an impairment in tonic arousal. The noradrenergic alerting system, originating in the brainstem and projecting to frontal cortex, shows greater right lateralization cortically in humans. Extensive right hemisphere damage may cause many neglect patients to be chronically underaroused, with particular difficulties in maintaining self-arousal endogenously over lengthy periods. This might in turn aggravate their neglect. In apparent support of this, phasically alerting neglect patients with warning stimuli can temporarily reduce their abnormal ipsilesional bias.

A further deficit that might apply to both sides of space, yet exacerbate contralesional neglect, concerns a general reduction in perceptual capacity following right hemisphere damage. As described earlier, perceptual extinction is a failure to detect contralesional stimuli specifically in the presence of multiple competing stimuli. This may reflect reduced processing capacity in addition to the ipsilesional bias. Neglect patients may thus exhibit a degree of simultanagnosia (a difficulty in perceiving multiple stimuli simultaneously) in addition to their lateral bias. In support of this, patients with extinction following right parietal damage may show an impaired capacity for reporting more than one target even in vertical arrays briefly presented on the ipsilesional (i.e. supposedly intact) side. Moreover, if instructed to report any contralesional item first in brief bilateral displays, they may show paradoxical extinction of items on the ipsilesional side. Abnormally prolonged attentional dwell-time has also been observed in people with right hemisphere neglect, even when

all stimuli are presented at central fixation and in rapid succession. Finally, it has recently been proposed that, in addition to the lateral biases in attention, neglect may involve deficits in spatial working memory (or more specifically, in maintaining representations of locations already examined during search, across saccades). Spatial working memory tasks in the normal brain activate a predominantly right-lateralized network that is strikingly similar to the areas of typical lesions in neglect patients. If neglect patients fail to form a stable representation of previously searched locations across multiple saccades, then their search might return recursively to those locations most favored by their lateral attentional bias, without the patient realizing this, exaggerating the bias still further.

The possible involvement of local biases, deficits in alertness, general reductions in perceptual capacity, and impairments in spatial working memory, all as exacerbating factors, may provide some explanation for the strong association of neglect with right hemisphere damage. Further accounts for this asymmetry include proposals that in humans the right hemisphere is specialized for spatial cognition (perhaps as the flipside of left hemisphere specialization for language); or that some right hemisphere areas normally represent locations within both sides of space, while left hemisphere homologues primarily represent contralateral locations (so that a right hemisphere lesion would be more devastating in terms of the spatial locations whose representations are lost). None of these various explanations need be regarded as mutually exclusive. Analogous hemispheric specialization may not exist in other animals.

RESIDUAL UNCONSCIOUS PROCESSING

Despite escaping the patient's awareness, neglected or extinguished stimuli can nevertheless undergo residual unconscious processing. Such findings again underscore the fact that neglect or extinction can arise despite sparing of some basic sensory and motor processes. Much of the evidence for residual processing has stemmed from analogies between the fate of neglected and extinguished stimuli in the patients, and the fate of unattended information in neurologically healthy people. While normal people can show little or no explicit awareness for stimuli that are not selectively attended, some implicit processing of these can nevertheless be revealed by indirect measures such as priming. As a first approximation, the residual unconscious processing found in patients with neglect or extinction corresponds well with that thought to take place 'preattentively' in the normal brain. In people with or without brain damage, this processing can determine which information will attract attention and reach awareness, and which will escape awareness instead.

Residual unconscious processing for contralesional stimuli can manifest in different ways. Extinguished or neglected stimuli can influence the patients' responses to consciously detected ipsilesional stimuli (e.g. affecting the speed of such responses), revealing that the presence, color, shape, or even the semantics of a contralesional stimulus can sometimes be extracted despite unawareness of it. In some cases, particular relationships between concurrent contralesional and ipsilesional stimuli can influence the degree of extinction or neglect. For instance, extinction is reduced when bilateral stimuli can be linked into a single object through image-segmentation principles, connecting the stimulated spatial locations to yield a common perceptual event. Such effects imply that image-segmentation processes may operate normally on contralesional visual inputs, prior to the level at which extinction arises. Similar conclusions follow from the various manifestations of so-called 'object-based' neglect (i.e. neglect affecting segmented visual objects, rather than unparsed spatial regions of the retinal image). Extinction or neglect can also be influenced by the particular identity of the stimuli used, again suggesting that some residual processing occurs prior to stages where contralesional inputs become extinguished or instead detected. For instance, stimuli of particular emotional significance (e.g. angry faces) may undergo less extinction, in keeping with the fact that such stimuli can readily capture attention even in healthy people.

The anatomical basis of such residual processing has been studied using functional imaging and electrophysiological measures of neural activity in the patients' brains, in response to extinguished or neglected stimuli. To date, functional magnetic resonance imaging (fMRI) studies have confirmed that, although escaping awareness, such stimuli can still activate anatomically intact areas in striate and extrastriate visual cortex, including category-specific areas in temporal cortex (e.g. the fusiform face area for extinguished faces), plus more remote limbic regions (e.g. amygdala and orbitofrontal cortex for extinguished fearful faces). This supports prior proposals that residual unconscious processing in neglect and extinction may reflect preserved afferents into primary visual cortex (V1), and

thence along the ventral visual system into temporal lobe (typically spared by prototypical parietal lesions). On the other hand, even early visual responses to extinguished stimuli may not be entirely normal. Visual evoked potentials have shown that components around 100 ms after stimulus onset can be attenuated for extinguished compared with perceived contralesional stimuli. Functional MRI results have also shown greater visual activation associated with conscious perception versus extinction, together with a greater covariation of striate visual cortex with parietal and frontal areas in the intact hemisphere during conscious perception. These results may accord with proposals that such parietal and frontal areas control spatial attention in the normal brain, by modulating activity in early sensory areas.

REHABILITATION

Neglect is disabling and carries a poor prognosis for independent living after a stroke. Spontaneous recovery occurs in some cases, but is often partial and poorly understood, though it may correlate with improved cerebral blood flow in spared areas of both the damaged and intact hemispheres. There have been some encouraging demonstrations of plasticity in the adult brain from other research areas, but rehabilitation of neglect remains a major challenge despite considerable effort and some recent progress. The evidence described above of considerable residual processing, and of factors influencing the degree of deficit, suggests a potential platform to build on.

One approach has been to train patients to direct search or attention towards the contralesional side. Neglect is improved by this in the short term, but there is little generalization to daily life. Another approach has been to modulate arousal, but again the challenge is to generalize this successfully to self-arousal during daily life. Use of the contralesional limb (in nonparetic patients) on the contralesional side of space can reduce neglect. However, many patients may not use this limb spontaneously in daily life, favoring the ipsilesional limb instead. Ipsilesional arm restraint may be considered, since this can ameliorate mild contralesional paresis, and may also reduce neglect. Patching of the whole ipsilesional eye has produced inconsistent results. More recently, patching of the ipsilesional visual hemifield in both eyes has produced some encouraging results. The logic behind this intervention is to correct the asymmetrical spatial bias by inducing a reverse bias at the input stage.

Caloric stimulation of the vestibular system (by iced water in the left ear, or warm water in the right ear, to induce vestibular signals similar to those from actually turning left) has been shown to ameliorate perceptual and exploratory neglect, plus associated deficits such as imaginal neglect, or anosognosia for hemiplegia. Again, the asymmetric stimulation may oppose the pathological bias to the ipsilesional side caused by the lesion. Vestibular signals also contribute to coding of external space relative to the body, being integrated with other sensory signals within parietal cortex. Similar principles may explain the analogous beneficial effects of vibratory neck-muscle stimulation (inducing proprioceptive signals similar to a leftwards turn of the trunk relative to the head). Both vestibular and proprioceptive manipulations produce consistent reductions in neglect, but their effects dissipate when the stimulation ends. The same applies for optokinetic stimulation.

Longer-term benefits have been reported following prism adaptation. The patients wear prisms shifting visual information towards the ipsilesional side, so that their reaches to visual targets initially err in the ipsilesional direction. Following adaptation, the hand is directed further to the contralesional side for a given retinal position. Just 10 min of this procedure has been claimed to improve neglect over several days, although the exact basis for this remains unknown. Prism adaptation in the reverse direction has no analogous impact.

Finally, drugs such as dopamine agonists have been used to increase arousal and enhance contralesional exploration, but with only mixed success to date.

Research into the processes that are impaired or spared in spatial neglect, and into the underlying neural mechanisms, has uncovered many important new findings in recent years. It is hoped that these findings will contribute to improved rehabilitation procedures in the future.

Further Reading

Bisiach E and Vallar G (2000) Unilateral neglect in humans. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, pp. 459–502. Amsterdam: Elsevier.

Driver J and Vuilleumier P (2001) Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* **79**: 39–88.

Heilman KM, Watson RT and Valenstein E (1993) Neglect and related disorders. In: Heilman KM and Valenstein E (eds) *Clinical Neuropsychology*, pp. 279–336. New York, NY: Oxford University Press.

Mesulam MM (1999) Spatial attention and neglect: parietal, frontal and cingulate contributions to the

mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society of London Series B Biological Science* **354**: 1325–1346.

Rafal RD (1994) Neglect. *Current Opinion in Neurobiology* **4**: 231–236.

Robertson IH and Marshall JC (1993) *Unilateral Neglect: Clinical and Experimental Studies*. Hillsdale, NJ: Lawrence Erlbaum.

Neural Basis of Memory: Systems Level

Introductory article

Arthur P Shimamura, University of California, Berkeley, California, USA

CONTENTS

A systems level approach to the neuroscientific investigation of memory looks at the ways in which different areas of the brain interact and communicate to achieve this function.

INTRODUCTION

What did you eat for dinner two days ago? What is the name of the mouse in the story *Dumbo*? How many turns do you need to make to get from your front door to your bedroom? How are you able to understand the visual symbols in this sentence? Memory – our remarkable capacity to learn and retain information – offers the answers to these questions. Across a lifetime, we encounter, store and retrieve an enormous amount of information. We have knowledge about ourselves, about world facts, and even about things that we do not usually associate with our memory, such as habits and skills. How does the brain accomplish such extraordinary feats of memory?

The study of memory has been approached from many scientific disciplines. Neurobiological approaches address the microstructure of memory, such as the biochemical mechanisms within neurons or the formation of synaptic connections between neurons. Such processes allow the brain to adapt to changes in the environment. At a more global level of analysis, or what is often described as a systems level, neuroscientists describe how brain regions interact and communicate to serve human memory. Take, by analogy, the levels at which one could understand how an automobile works: at a microscopic level, one could study how the fuel molecules, through combustion, react to make an automobile move, whereas at the systems level, one could study the function of specific components, such as the engine or the transmission, and then determine how these components work together. This article approaches human memory from a systems level, focusing on brain regions that contribute to various components of human memory.

Our understanding of human memory has been significantly advanced by integrating neuroscience with cognitive science – a field of research called cognitive neuroscience. Over the past century medical scientists have studied patients with brain injury as a way to understand the biology of human memory. Around 1970, cognitive scientists became interested in neuropsychological studies, because such studies often demonstrated modularity in cognitive function. For example, brain injury could severely impair an aspect of memory, but leave intact other cognition functions such as attention, language, and problem-solving skills. More recently, studies using neuroimaging techniques – such as functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) – have led to finer analyses concerning the relationship between brain regions and memory function. In such studies, it is possible to show that specific brain regions become active when individuals engage in memory processes, such as learning a list of words or trying to remember a past event. The following sections will identify components of memory from a cognitive neuroscience perspective.

ORGANIC AMNESIA AND THE MEDIAL TEMPORAL CORTEX

A significant advance in understanding the biology of human memory occurred by chance through the study of a now-famous neurological patient known by his initials, HM. This patient underwent an experimental surgical procedure in 1953 to relieve severe epilepsy. The surgery involved removal of brain tissue in the medial temporal cortex, a region

of the cerebral cortex that encompasses the inner (medial) surface of the temporal lobe. The operation included removal of the hippocampus, a brain structure adjacent to the cerebral cortex. Following surgery, HM's epileptic seizures were reduced, but he was left with a profound organic amnesia – that is, he was unable to remember events and information encountered since his operation. For example, when HM was asked what he had had for lunch half an hour previously, he could not remember what he had eaten, or even if he had eaten at all. Despite this severe memory impairment, there was little impairment in intellectual ability or language skills. Indeed, HM had a normal IQ (intelligence quotient) and could communicate fluently with others. There was some memory impairment for information that occurred before his operation, a disorder called retrograde amnesia. For example, HM could not recall the layout of the hospital in which he was treated, or recall the death of a favorite uncle who had died three years previously. Yet HM's retrograde amnesia was not gross, as indicated by the fact that he performed as well as others on a face recognition test of celebrities who became famous prior to his operation. Also, he was capable of recollecting events and episodes from his childhood.

Clinical observations indicate that HM's memory for ongoing events is still severely impaired. He is somewhat aware of his condition, as indicated by the following quote:

> Right now, I'm wondering. Have I done or said something amiss? You see, at this moment everything looks clear to me, but what happened just before? That's what worries me. It's like waking from a dream; I just don't know. It's like waking from a dream.

Indeed, HM lacks the ability to acquire and retain events and facts encountered since his operation. The impairment affects information received from all types of sensory receptor and includes impairment of both verbal and nonverbal (e.g. spatial) memory. For example, HM has failed to add to his vocabulary new words such as 'jacuzzi', because such words have been added to the language since his surgery. He also exhibits severe impairment on laboratory tests in which he is asked to recall or recognize recently presented words and pictures.

Studies of neurological disorders that affect the medial temporal cortex have confirmed the importance of this brain region for memory. For example, tumors or strokes occurring in the medial temporal cortex can cause organic amnesia. Also, other neurological disorders – such as viral infections, ischemia (loss of blood flow to the brain) or hypoxia (loss of oxygen to the brain) – particularly damage the medial temporal cortex. In these cases, the inability to retain newly learned information is the outstanding cognitive disorder. Finally, in Alzheimer disease, organic amnesia is usually the most significant impairment in the early stages. However, this is a degenerative disease, and as more brain areas become affected more cognitive functions become disrupted.

Events in your life consist of a variety of psychological experiences, such as sensory inputs, ideas, feelings, and actions. Indeed, every moment is completely unique in that these experiences – perceptions, thoughts, emotions, actions – will happen together only once. If something significant occurs at any particular moment, such as a life-threatening or life-enriching event, it would be useful to be able to bind or associate the varied aspects of the event as one encapsulated 'episodic' memory. Likewise, if you are asked to take an examination, say a history test, it is important to be able to relate or associate a diverse set of facts into an encapsulated 'semantic' or conceptual memory. To form such encapsulated memories, it important to relate quickly many aspects of an event or concept.

A prominent view of medial temporal function is that it allows for the rapid linking of these experiences, sometimes called 'relational binding'. The medial temporal cortex is ideally situated for such relational binding to occur. Many parts of the cerebral cortex extend neural projections that converge onto regions within the medial temporal cortex, from which neural projections extend back to these cortical areas. It is believed that the medial temporal cortex enables neural activations in any given moment to be bound as an encapsulated set of associations. Without this ability, memory would dissolve into fragments unattached or unrelated to any particular time, place, or concept.

WORKING MEMORY AND THE PREFRONTAL CORTEX

Despite HM's amnesic disorder, he is able to think, communicate, and even retrieve childhood memories. In addition, he has the capacity to keep information in mind for brief periods: for example, he can hold in mind a short series of digits, such as a telephone number. Yet, as soon he is distracted, the information is lost. These findings suggest that organic amnesia does not affect the online processing of information, or what psychologists call 'working memory'. Working memory refers to the

processes that allow us to keep information in mind. These processes include the ability to select, retrieve, and maintain information in a short-term or transient manner. With respect to the example of holding a telephone number in mind, working memory is essential when it is necessary to maintain or use information after it has been presented. Animal studies have been invaluable in defining a brain region – the prefrontal cortex – that is essential for working memory. In such studies, neurons in the prefrontal cortex are activated when an animal must maintain information, such as the location of a food item after it has been hidden. Moreover, when this brain region is damaged, the ability to maintain information is disrupted.

The prefrontal cortex is part of the most anterior section of the frontal lobes. It is a large area, comprising 28 percent of the human cortex. Disruption of working memory is apparent in people with damage to the prefrontal cortex. Such individuals have difficulty paying attention and keeping things in mind. For example, people with frontal lobe lesions exhibit severe deficits in the ability to keep in mind a series of items, such as digits (telephone numbers), sounds or spatial locations. In neuroimaging studies using PET and fMRI, increased activation in prefrontal cortex occurs when individuals are asked to keep information in mind. The left prefrontal cortex is involved in holding verbal information whereas the right prefrontal cortex is involved in holding spatial information.

Working memory is important for efficient learning and retrieval. For example, try to learn the following series of words: sister, soda, milk, shoe, father, pants, juice, brother, hat. Maybe you used a strategy to help you organize the words. Perhaps you noticed that the words could be encoded or grouped into three meaningful categories: family relatives, drinks, and clothing. Learning is significantly benefited by the meaningful organization of information. Making new information meaningful to you enhances the integration of new information into your existing database. This strategy involves working memory, because it is necessary to manipulate and update the information in mind. That is, the more you think about the information and why it is meaningful to you, the more you actively integrate the new information with what you already know. Individuals with prefrontal damage exhibit problems in organizing their thoughts and memories. These people do not develop learning strategies or think deeply about what they need to learn. For example, they do not group words into meaningful categories. In neuroimaging studies, the prefrontal cortex is particularly active when individuals are asked to consider the meaning of words. Also, prefrontal activation during the learning of words is greater for words that are later remembered than for words that are forgotten. This finding shows that activity in the prefrontal cortex during learning increases the chances of retrieving the information at a later time.

Working memory is also important in retrieving information. Consider a fairly simple memory retrieval task – try to retrieve as many animals that you can think of in 60 s. This task requires you to search your memories and retrieve specific items. Perhaps you developed a strategy, such as retrieving animals within certain categories, such as pets, farm animals and animals one sees at the zoo, or maybe you went down the alphabet and used each letter as a cue to retrieve animal names. Such strategies are efficient because they facilitate the retrieval of different items and prevent repeating the same ones. Keeping track of which animals were already elicited requires you to keep in mind the animals and update this list each time you retrieve an item. Individuals with prefrontal damage have difficulty keeping track of previously retrieved items. It is as if these items get in the way, making it harder to retrieve others. Thus, these people report only four or five different animals in a minute and often repeat the same ones over and over. Neuroimaging studies have corroborated the importance of the prefrontal cortex during memory retrieval: when brain activity is measured while individuals are asked to retrieve information, the prefrontal cortex is particularly active.

Findings from studies of patients with prefrontal damage and findings from neuroimaging studies suggest that the prefrontal cortex enables efficient organization of information in mind. That is, when one is working with memories, it is necessary to select relevant information, update that information, and use it to respond. Cognitive scientists use the term 'executive control' to characterize the processes associated with the selection, control, and manipulation of information in working memory. Related issues include selective attention and response supervision. Selecting and using information in working memory is essential for efficient processing in many cognitive domains. Indeed, the prefrontal cortex, which is most closely tied to executive control, appears to be involved not only in the control of memory, but also in the selection of perceptual features, thinking, problem-solving, and response selection. Different areas of the prefrontal cortex appear to be responsible for the control of difference aspects of information processing.

REPRESENTING KNOWLEDGE AND PAST EXPERIENCES

So far, we have looked at how the medial temporal cortex binds experiences together to form new memories and how the prefrontal cortex monitors and controls memory activations. But what actually is stored in memory? This intriguing question has bemused scientists, philosophers and poets. Karl Lashley, the noted neuroscientist, argued that there was not a single location where a particular memory was stored. Instead, memories were distributed widely in the brain. To demonstrate this notion, Lashley trained rats to follow a particular pathway in a maze in order to obtain food. Lashley found that lesions of many brain regions, not just one region, impaired performance on this task; moreover, the severity of impairment was determined by the size of the lesion rather than its location. This idea – that memories were distributed in many parts of the brain – countered 'filing cabinet' views of memory in which specific memories were localized in specific areas in the brain.

Today, most cognitive neuroscientists take a middle-of-the-road position between distributed and localized views of memory storage. It is not likely that any memory, such as memory of a past birthday party or your knowledge about a particular historical event, is stored in a small, localized area in your brain. Indeed, there is no evidence that a brain injury would produce an absence of a specific piece of information in one's memory, which would be akin to throwing out a specific folder in your filing cabinet. Thus, memory storage appears to be distributed or represented in many parts of the brain. However, certain features of a memory, such as memory for faces, can be disrupted by specific brain damage. In many cases, these aspects are tied to the way the brain uses or processes such information. Thus, although memories may be distributed widely in the brain, each area may contribute to the memory in a different way. For instance, memories of the sights, sounds, feelings, and actions experienced during a birthday party are likely to be stored in different parts of your brain. When you recollect such past events, you draw on many of these different aspects of your memory, not just any localized one.

Episodic and Semantic Memory

Episodic memory refers to memories that are tied to one's autobiography. These memories are embedded within a context of time and place. Semantic memory refers to the vast database of factual knowledge stored in memory. Such knowledge is usually learned on many occasions and thus not tied to any specific time in one's life. For example, your knowledge of how memory works comes from many experiences (hopefully one of which is from reading this text). The distinction between episodic and semantic memory was first developed by Endel Tulving in 1973. It has been useful in denoting two major divisions of memory representation.

There are both similarities and differences in the manner in which episodic and semantic memory are represented in the brain. One similarity is that the acquisition of new episodic and semantic memory is dependent upon the integrity of the medial temporal cortex. That is, people with organic amnesia have difficulty remembering autobiographical events experienced since the onset of amnesia as well as learning new factual information. Indeed, the outstanding feature of organic amnesia is a deficit in forming new episodic memory. However, these people also fail to learn new semantic knowledge. As mentioned earlier, HM has no knowledge of words that entered his culture after the onset of his amnesia. In another example, a patient with amnesia had been vice-president of an optical firm and was very knowledgeable about the field of optics. However, when he became amnesic, this knowledge did not grow as he was unable to add to this knowledge base. Thus, the acquisition of both episodic and semantic memories depends upon the medial temporal cortex.

One difference between episodic and semantic memory is the extent to which they are represented in the two cerebral hemispheres. It has been known for over a century that the left hemisphere in most individuals is specialized for language processes. Much of our verbal processes and verbal knowledge, such as knowledge about phonology and word definitions, are largely represented in the left cerebral hemisphere. The right hemisphere in most individuals is specialized for spatial processes. For example, damage to the right hemisphere often produces spatial disorders, such as a 'neglect' syndrome in which individuals fail to attend to stimuli on one side. As a large proportion of episodic memory is spatial in nature, damage to the right hemisphere tends to produce greater disruption of episodic memory than semantic memory. As semantic memory is often more tied to verbal processing, such as factual knowledge read in books, it is often the case that semantic memory is more disrupted by damage to the left cerebral hemisphere. As mentioned earlier, such

memories are widely distributed. Thus, extensive damage to one hemisphere is needed to observe disruption in episodic or semantic memory. Moreover, such damage does not appear to reflect a loss of a specific memory, but instead impairment is indicated by diffuse or fuzzy memories.

Category-specific Knowledge

The notion of a widely distributed memory system suggests that memory is multiply represented in many brain regions. This kind of memory representation is efficient because damage to one region may degrade memory but it is unlikely that it will totally abolish it. Yet, as mentioned above, there appear to be features or categories of knowledge that are grouped together. Interestingly, in rare cases of brain injury, there appear to be category-specific deficits of semantic knowledge. That is, some affected individuals exhibit greater problems in recollecting certain features of knowledge than others. One outstanding impairment involves a disorder in remembering faces. This impairment (prosopagnosia) is often caused by damage to ventral–posterior regions of the cerebral cortex; it affects visual recognition of all faces, including the patient's own face. Recent fMRI studies have identified an area within this region, called the fusiform gyrus, that is highly active during the viewing of faces. Although this area is involved in the visual memory of faces, it does not disrupt all memories for people. Patients with prosopagnosia can recognize familiar individuals, such as friends and family, by their voices.

In rare cases, specific loss of a semantic category can occur. For example, a person can exhibit a severe loss in the ability to recognize drawings of common animals, whereas the same individual can recognize drawings of manufactured objects, such as common tools. Interestingly, other affected individuals may exhibit exactly the opposite pattern – an inability to recognize common tools while retaining the ability to name common animals. Such cases suggest that the representations of certain aspects of knowledge are localized in different brain regions. In these people it is difficult to ascertain the brain regions responsible for such category-specific deficits, because it is often the case that the brain injury is large or diffuse.

Do category-specific deficits in semantic knowledge suggest that our memories are stored in a localized rather than a distributed manner? It is unlikely that a specific memory, such as memory of a pet you once owned, is represented in a specific location in the brain. How, then, can one interpret findings of a separation of the concept of animals from that of tools? Neuroimaging studies have shed some light on this issue. In one PET study, participants viewed drawings of animals, tools, or nonsense forms. The nonsense forms were used as control stimuli, and PET activation in response to drawings of animals and tools was compared with activation to these control stimuli. The study showed that both animals and tools activated bilateral areas in the ventral temporal cortex, suggesting that semantic knowledge of animals and tools has broad and overlapping (i.e. distributed) representation.

There were, however, some category-specific activations. Drawings of animals but not tools activated areas in the occipital cortex, an area of the brain associated with visual processing. Drawings of tools but not animals activated premotor areas in the frontal cortex, an area that is also activated during imagined hand movements. Such findings suggest that category-specific knowledge may be tied to the manner in which such information is encoded or used. Animals tend to be encoded and encountered visually, whereas tools are manipulated. It appears that our memories are not stored as purely abstract, semantic representations. Instead, they are associated with the manner in which they have been encoded and used. Apparently, the distribution of our semantic knowledge is related to the distribution of other cognitive processes in the brain, such as visual, auditory, verbal, spatial, emotional, and motor processes.

Procedural Memory: Habits and Skills

The ability to read, drive a car or swim certainly requires memory. This kind of memory, however, is not represented in the same manner as other forms of memory, such as semantic or episodic memory. Skills involve the tuning or coordinating of sensory and motor associations. Procedural memory refers to the processes and representations associated with habits and skilled behavior. Thus, procedural memory is linked specifically to neural circuits associated with the modification of perceptual and motor function. For example, a skilled behavior such as throwing a ball or playing a musical instrument involves the modification and tuning of various perceptual-motor circuits. It is inherently 'procedural' in that skilled behavior is expressed as a sequence of finely tuned movements. Procedural memory can occur without explicit or conscious knowledge of the training session. Indeed, it is often the case that conscious intervention or feedback can disrupt the expression of skills.

Procedural memory is unique in that it appears to be independent of the medial temporal cortex. One of the most striking findings is that people with amnesia can learn simple skills in an entirely normal fashion. For example, HM showed learning and retention in a mirror-drawing task in which he was required to trace the outline of a star while viewing the star through a mirror. The task is difficult at first but then becomes easier with practice. HM exhibited normal learning and retained this skill over days of practice. Another skill learning task, the pursuit-rotor task, has been used to test people with brain injury. In this task, the patient learns to keep a stylus on a rotating target. In these tests, HM and other amnesic patients perform well, even when tested the next day. However, such patients have no memory of having performed the task before.

Habits or dispositions are forms of procedural memory that can be acquired without conscious awareness. For example, during an interview with an amnesic patient, a neurologist hid a pin between his fingers and surreptitiously pricked the patient on the hand. At a later time during the interview, he once again reached for the patient's hand, but the patient quickly withdrew her hand. The patient did not acknowledge the previous incident, and, when asked why she withdrew her hand, she simply stated, 'Sometimes pins are hidden in people's hands.' This anecdote is an example of stimulus–response habit learning without awareness, or what is called 'fear conditioning'.

Other habitual forms of learning can be demonstrated in studies of Pavlovian classical conditioning, in which a tone is paired with a puff of air to the eye, causing the subject to blink. After several trials, the person will blink on hearing tone by itself. This kind of habit conditioning is normal in individuals with amnesia. These people retained the eyeblink response for as long as 24 h, even though they did not recognize the test apparatus.

Such examples of habits and skill learning have been used to study subcomponents of procedural memory, such as timing, sequencing, and storing of perceptual-motor associations. Procedural memory is presumed to involve both cortical and subcortical circuits. At the cortical level, procedural memory is assumed to depend significantly upon unimodal sensory systems in posterior cortex, and motor and premotor systems in frontal cortex. At the subcortical level, the cerebellum and basal ganglia appear to be significantly involved in skilled behavior.

Functional imaging studies have provided some additional support for the participation of the basal ganglia in motor skill learning. In one study, participants underwent PET scanning while learning a finger-tapping skill. Comparisons between activations seen during initial stages of learning and skilled performance revealed reductions in basal ganglia activation over the course of learning, perhaps reflecting increased processing efficiency. Changes in activation over the course of motor skill learning have also been observed in motor cortex, somatosensory processing areas of thalamus and cortex, and the cerebellum – indeed, in all areas thought to be involved in sensorimotor behavior. These findings suggest that the products of skill learning are represented in widely distributed networks involving many, if not all, of the brain areas that contribute to performance of the skill.

CONCLUSION

Findings from both neurological patients and neuroimaging studies have suggested that certain brain regions contribute to human memory. The medial temporal cortex is essential for the learning of new information. People with organic amnesia following damage to this brain region have difficulty learning new episodic and semantic information. Conceptually, the medial temporal cortex appears to be involved in relational binding, which enables new information to be linked or associated with existing knowledge. The prefrontal cortex is critical for the selection and control of activated memories. The notion of executive control is used to describe the role of the prefrontal cortex in activating memories.

We know less about the manner in which memories are actually stored. Cognitive scientists suggest that memories are stored as a widely distributed network of knowledge. Neuroimaging studies suggest that our memories are not simply abstract concepts but are instead stored with respect to the way they were encoded and used. Thus, some memories are tied to sensory processes whereas others are tied to motor functions. Moreover, procedural memory appears to be represented in a different manner from episodic or semantic memory. That is, our memory for skills and habits is less available to conscious awareness and appears to be represented as the tuning of associations between sensory and motor functions.

Based on this analysis, cognitive neuroscience provides a useful framework in which to conceptualize the intricacies of human memory. It has been possible to identify components of memory function, such as relational binding, working

memory, episodic memory, semantic memory, and procedural memory. Just as one component of an automobile, such as the engine or transmission system, cannot completely define its working, it is impossible to identify one component or region of the brain that completely defines the workings of human memory. It is the interplay of these components that provides us with the rich and seemingly endless experience of memories.

Further Reading

Baddeley A (1986) *Working Memory*. Oxford, UK: Oxford University Press.

Farah MJ and Aguirre GK (1999) Imaging visual recognition: PET and fMRI studies of the functional anatomy of human visual recognition. *Trends in Cognitive Sciences* **3**: 179–186.

Gazzaniga MS (ed.) (2000) *The New Cognitive Neurosciences*, 2nd edn. Cambridge, MA: MIT Press.

Nolde SF, Johnson MK and Raye CL (1998) The role of prefrontal cortex during tests of episodic memory. *Trends in Cognitive Sciences* **2**: 1399–1406.

Roberts AC, Robbins TW and Weiskrantz L (eds) (1999) *The Prefrontal Cortex*. Oxford, UK: Oxford University Press.

Shimamura AP (2000) The role of the prefrontal cortex in dynamic filtering. *Psychobiology* **28**: 207–218.

Squire LR (1987) *Memory and Brain*. New York, NY: Oxford University Press.

Squire LR (ed.) (1992) *Encyclopedia of Learning and Memory*. New York, NY: Macmillan.

Neural Correlates of Consciousness as State and Trait

Intermediate article

Petra Stoerig, Heinrich-Heine-University, Düsseldorf, Germany

CONTENTS

INTRODUCTION

Consciousness, filed as 'unmentionable' for many decades, 'both the most obvious and the most mysterious feature of our minds' (Gregory, 1987, p. 160), came back into the scientific arena with a vengeance in the 1990s. As there is still no accepted scientific definition in sight, the general definition has been circumscribed as what you lose when you become comatose or anesthetized, and what you regain upon recovery (Searle, 1992). In this sense it is a state of an organism, and by studying the differences between conscious and unconscious states, insights into the neuronal basis of consciousness, as well as its function(s), can be gained.

If all losses of consciousness were to result from damage to the same structures of the central nervous system (CNS), one could regard these structures as hypothetical correlates of conscious states. The alternative view, which regards consciousness as an emergent property of the overall activity of the brain, could then be relegated to the back burner. However, 'consciousness' is not only a state of an organism, it is also a trait of a perception, a thought, a wish, a memory or an intention. How does the brain effect this 'consciousness of something'? How does it effect the conscious representation of one piece of information but not another at any one point in time? Do particular macroscopic structures of the CNS (nuclei, pathways, cortical areas) need to be activated? Or is consciousness an emergent property of particular CNS functions? Does it entail coherent patterns of activation, possibly involving recurrent loops between cortical areas? Since, in order to have consciousness of anything, the organism needs to be in a conscious state, should the neural correlate of conscious representations involve the same substrate as the conscious state, or a different one, or an interaction between the two? Or is any neuronal pattern in principle accessible to conscious representation so long as the organism is in a conscious state?

This article will deal with empirical findings that bear on these questions, focusing on consciousness in these two senses of the term, namely conscious as opposed to unconscious (as an attribute of an organism), and conscious as opposed to implicit (as an attribute of a representation).

CONSCIOUSNESS AS A STATE OF ORGANISMS

Consciousness is a graded state, with levels ranging from somnolent and drowsy to alert and hyperactive. These correlate with different levels of perceptiveness and attention, speed and complexity of behavior, and electroencephalographic (EEG) activity (which is of lower amplitude and higher frequency the more awake the subject) (Moruzzi and Magoun, 1949). A characteristic feature that may help to target the system that regulates the state of consciousness is its lack of specificity: an organism is conscious regardless of the specific content of that consciousness.

Coma and Vegetative State

Of the unconscious states, the rarest and least reversible is that of coma, which can result from toxic, metabolic, degenerative, vascular, or traumatic causes. It probably comes closest to what most people intuit unconsciousness to be – something akin to a state of deepest sleep. Reflexes such as the pupil light reflex may be present, flexor or

tensor muscle activity is absent, verbal behavior (if present) is restricted to grunts, and brain metabolism is reduced by approximately 50%. If the coma does not end in recovery or death, it develops into a more or less persistent vegetative state (PVS) (Jennett and Plum, 1972). This state differs markedly from coma in that the sleep–wake cycle is resumed, and during the waking phases the eyes open spontaneously and in response to stimulation. In addition, the patients may move limbs, grimace, smile, cry, utter words, and briefly follow a moving object with gaze or head. Although overall brain metabolism is still as reduced as in coma, islands of relatively preserved functional cortical tissue have recently been found using brain imaging techniques (Menon *et al*., 1998; Schiff *et al*., 1999, 2002). If in contrast to the 'locked-in syndrome' in which the patient is conscious but, due to paralysis, almost incapable of expressing him- or herself, the patient in the vegetative state is indeed fully unconscious, then it follows that wakefulness is not sufficient for consciousness.

It is still uncertain whether, despite the wide range of pathology that causes coma or PVS, a common anatomical substrate exists that needs to be impaired for consciousness to be abolished. On the one hand, very widespread cortical degeneration as well as diencephalic, mesencephalic, and brainstem lesions may all cause loss of consciousness (Plum and Posner, 1980). On the other hand, massive bilateral lesions of the frontal, temporal, parietal, or occipital lobes do not cause unconsciousness (Bogen, 1997), nor does even a very extensive hemispherectomy (Austin and Grant, 1955). Therefore the substrate necessary for consciousness cannot reside in any one pair of the cortical lobes, and must be present in both hemispheres.

The major contenders at present differ primarily in their delineation. The wider concept (Penfield, 1938) postulates a centrencephalic network consisting of the ascending reticular activating system (ARAS) of the brainstem, the diencephalic nuclei that it targets, and the diffuse thalamocortico-thalamic projections that it receives. The narrower hypothesis focuses on the nonspecific thalamic nuclei, specifically the intralaminar nuclei of the thalamus (ILN) (Bogen, 1995, 1997). The latter theory has gained some support from the neuropathological findings in the famous PVS case of Karen Ann Quinlan which, despite the common widespread lesions that develop in the course of longlasting vegetative states, showed disproportionately dense bilateral destruction of the thalamus (Kinney *et al*., 1994).

Anesthesia

It was only very recently that general anesthesia was first reported to target these same structures preferentially. Previously, a homogenous overall reduction in brain metabolism was regarded as characteristic of the anesthetized state (Alkire *et al*., 1995). However, analysis of functional brain imaging data recorded under widely used anesthetics (isoflurane, halothane, and propofol) has revealed a disproportionate deactivation in both narcotic-specific and (at least for the small number of substances tested so far) narcotic-independent foci of selective depression (Fiset *et al*., 1999; Alkire *et al*., 2000). In addition to frontoparietal structures, the latter include the thalamus, forging the first link between anesthetically and pathologically induced unconsciousness.

Sleep

The neuroanatomy of dreamless sleep which is often counted among the unconscious states (e.g. Rees *et al*., 2002), involves the ARAS of the brainstem, the pontine nuclei, the locus coeruleus, the raphe, and the dorsal tegmental nuclei (Moruzzi and Magoun, 1949; Hobson and Steriade, 1986). Neurons within these structures defacilitate thalamic neurons (Steriade and McCarley, 1990) and, via the reticular thalamic nucleus (RTN) that mediates between the thalamus and the cortex, entrain the thalamocorticothalamic network into the spindle and delta rhythmicity that is expressed in the spontaneous EEG characteristics of sleep stages 2 and 3–4, respectively (Steriade *et al*., 1994). In contrast, rapid eye movements occur predominantly in stages characterized by a low-amplitude, high-frequency EEG resembling the waking pattern, the 'REM' phases. As dreams are reported more often when subjects are woken during REM rather than non-REM sleep, neuroimaging studies revealing a regional decrease in the thalamus during slow-wave sleep (SWS) that covaried with the delta activity simultaneously recorded for sleep-stage assessment (Maquet *et al*., 1990, 1992) again point to a central role for the thalamus in regulating the level of consciousness. However, dreams and other more thought-like forms of mentation have also been sufficiently often reported by subjects woken from non-REM sleep (Foulkes, 1962; Bosinelli, 1995) to forbid a simple equation of SWS with an unconscious state; indeed, any differences between dream and non-dream sleep may thus be obscured. After all, getting a dream report necessitates waking the subject, which abolishes the

sleeping state one would like to study, and relegates whatever mentation may have occurred before to memory, which in turn may be differentially affected by different sleep stages. In view of these unsolved methodological issues, it is in fact uncertain whether any truly unconscious phases are part of normal sleep at all. Even if sleep entailed a switch more in what one can be conscious of than in whether one is conscious, it is still noteworthy that, like the vegetative states and anesthesia, SWS appears to be associated with disproportionate deactivation of thalamic and dorsal brainstem structures. (*See* **Sleep and Dreaming**)

CONSCIOUSNESS AS AN ATTRIBUTE OF REPRESENTATIONS

Loss of a conscious representation can be induced in conscious organisms either by non-invasive experimental manipulation, or following lesions in the central nervous system. The advantage of the former approach is that pathology does not interfere with the conclusions that can be drawn from the results, but this is balanced by two disadvantages. First, the effects are transitory, and secondly, the differences in the neuronal activation patterns are difficult to assess. The reverse is true for the neuropsychological approach. Here the lesion prevents the conscious representation permanently, allowing extensive prompting of the implicit faculties remaining in its absence. However, conclusions about the neuronal basis are restricted by the fact that a CNS lesion that causes loss of a conscious representation both anatomically and functionally also affects neurons that are distant from the lesion itself and/or unrelated to the loss.

To date, the majority of both types of study have concentrated on the visual system, but the findings agree with those obtained from other modalities.

There are several types of experimental manipulation that transiently block a particular conscious visual representation. In visual masking, a second stimulus is presented in close temporal conjunction with the target (Alpern, 1953; Breitmeyer, 1984), or a transcranial magnetic impulse is applied (Amassian *et al*., 1989). In change blindness, there is an impaired capacity to detect the position or identity of a changing item in a complex visual pattern when the original and altered image are successively presented and separated by brief blanks (Simons and Levin, 1997). In binocular rivalry, two incompatible patterns (e.g. an upward and a downward moving grating) are presented simultaneously, one to each eye, and the observer's perception, instead of fusing the patterns, alternates between them (Blake, 1989). In all of these paradigms, a transient blindness is induced for a stimulus or change that would be easily detected in isolation. How does the manipulation interfere with the processes that are necessary for conscious vision? And how is the neuronal response to unseen targets different from that to seen targets?

Both backward and forward visual masking and binocular rivalry have been amply studied psychophysically and physiologically in order to determine the most effective conditions and learn how the neuronal coding of seen and unseen information may differ. The results of the latter approach show that backward and forward masking stimuli interfere with different aspects of the neuronal response to the masked target in the primary visual cortex (V1). The mask that precedes the target obliterates the early part of the discharge, whereas the mask that follows the target obliterates its late components (Bridgeman, 1980; Macknick and Livingstone, 1998). (*See* **Neural Correlates of Visual Consciousness**)

Neuronal activity both in V1 and in higher extrastriate visual cortical areas has been studied during rivalrous dichoptic stimulation while a monkey behaviorally indicated which pattern it perceived at any one point in time (Logothetis, 1998). The results showed that some neurons responded to their preferred stimulus only when the monkey reported seeing it (although it was always present), while others responded independently of the perceptual status. Interestingly, the proportion of neurons that followed the percept increased markedly from V1/V2 (*c*. 20%), to high extrastriate visual cortical areas in the temporal lobe (90%). This increase may indicate that the higher the visual cortical area, the more its activity reflects what the animal is aware of. But why do some neurons in earlier areas already follow the percept? Do they decide what will happen at the later stages? Or do they receive the output of the higher areas via re-entrant connections?

The latter possibility is more consistent with recent functional magnetic resonance imaging (fMRI) data which demonstrate that activity in areas V1, V2, and V3 covaries with the percept when stimuli that differ in contrast and are thus tuned to the properties of the early areas are used (Polonsky *et al*., 2000). The cause of the characteristic time course of the perceptual switch, which is approximately of the order of 1–3 s, was addressed in an fMRI study that compared activity during rivalry with that seen during successive presentation of the same stimuli, mimicking the rivalry percept. The results revealed that only

frontoparietal areas in the right hemisphere responded preferentially in the rivalrous condition, which suggests that the perceptual switch may be initiated outside the visual system proper (Lumer *et al*., 1998). The view that awareness may be mediated by a special module outside the visual system, and not by particular types of neuronal response patterns within it, is also championed by Crick and Koch (1998). Like J. H. Jackson (Taylor *et al*., 1931), they focus on prefrontal areas, surmising that sensory awareness needs to be directly accessible to executive functions.

However, even if that is indeed the major purpose of conscious representations, neuropsychological evidence demonstrates that while frontal lesions can impair the appropriate use of consciously represented information, it is only destruction of the primary retino–geniculo–striate cortical visual system that causes blindness (e.g., Wilbrand and Sänger, 1900). Neither circumscribed prefrontal lesions (complete destruction of both prefrontal lobes has not been reported, and would impair a multitude of cognitive and executive operations) nor lesions confined to higher visual cortical areas cause a loss of conscious vision. The latter type of lesion instead affects selective aspects of conscious vision. Destruction of the cortical color complex causes cerebral color blindness (Meadows, 1974), lesions in the human motion complex cause motion perception deficits (Zeki, 1991), and extensive damage that predominantly affects the higher occipitotemporal areas produces the visual agnosias (inability to recognize objects, faces or scenes as what they are or mean) (Benson and Greenberg, 1969; Grüsser and Landis, 1991). Conscious phenomenal vision (the ability to see color and brightness), although impaired in these conditions, remains so long as the lesion does not completely disconnect the higher cortical areas from areas V1 and V2. The intrinsic architecture of conscious vision that is expressed in the way in which the deficits map on to the lesioned areas (Stoerig, 1996), together with the loss of conscious vision caused by V1/V2 damage (although higher extrastriate areas can still be activated via non-geniculo-striate cortical pathways), points to a pivotal role for V1 and V2. Whether this is due to their providing a necessary input, to their receiving the results of the processing in higher areas via reentrant projections, and/or to their allowing the establishment of reverberating inter-areal circuits is still unknown.

These possibilities relate to computational aspects of the neuronal processing that mediates awareness (Cariani, 2000; Atkinson *et al*., 2000). What structures are involved in the process of making neural information conscious is one question, and whether neurons in these structures need to fire or interact with others in a particular fashion is another. Synchronization of neuronal firing in neurons that code for similar features or objects (Singer, 2000), stability (O'Brien and Opie, 1999) or strength (Greenfield, 1995) of responses, and the duration of the spike train may have important roles.

The latter was demonstrated in a pioneering series of experiments on human subjects performed by Libet *et al*. (1964). These researchers studied the time period required for peripherally and (via electrodes applied to the subject's brain) centrally applied electrical stimulation to evoke a reportable sensation. They found this duration to be at least 100 ms, a value that could increase to several hundred milliseconds with low-intensity stimuli. Although these findings could be seen to indicate that the duration of the spike train is more important than the type or position of the stimulated neuron (so that if the timing was right, any neuron could contribute to conscious experience), the fact that a very large proportion of neuronal processing is in principle inaccessible to direct conscious representation renders this possibility unlikely. The regulation of our hormonal status, blood sugar level, digestion, and sleep, and also the concerted action of the muscles that underlie our every action, are not consciously representable. This indicates that a combination of anatomical and computational features is ultimately more likely to explain the way in which information is consciously represented.

CONSCIOUSNESS AS STATE AND TRAIT

Implicit Functions in the Absence of Consciousness

To a varying degree, information can be processed in unconscious states. Comatose patients may show reflexive responses. PVS patients may, in response to sensory stimulation, show regional increases in cerebral bloodflow which reflect the particular rudimentary behavior patterns that these patients exhibit (Menon *et al*., 1998; Schiff *et al*., 1999, 2000). Anesthetized subjects who are presented with lists of spoken words via headphones during surgery may score well above the levels expected by chance in postoperative tests of implicit recognition (Andrade, 1995; Bonebakker *et al*., 1996). It is common knowledge that information is also processed during sleep. Examples of this include the

phenomenon of sudden awakening in response to small sounds of distress from one's child, while sleep continues undisturbed by much louder but irrelevant noises on the street outside. External information is also processed during sleepwalking and other forms of sleep activity, which include sensory-guided navigation and fixed-action patterns such as walking and eating. As pointed out earlier, proof that normal sleep is unconscious is still outstanding, and it is safer to assume that this kind of complex behavior occurs in altered states of consciousness rather than in its absence.

More evidence for information processing in the absence of a conscious representation comes from studies of conscious subjects who were rendered incapable of seeing a stimulus either by an experimental manipulation or because of a circumscribed visuocortical lesion. Using forced-choice procedures that make the subject guess whether or not they are or have previously been confronted with a stimulus, and indirect procedures, such as the measurement of reaction times as a function of presenting unseen stimuli in addition to seen ones, evidence has been found for the processing of information that is neuronally but not consciously represented. Examples from neuropsychology show that prosopagnosic patients respond differently to familiar and unfamiliar faces (Bruyer *et al*., 1983), colorblind patients can see borders or movement, although these features are defined solely by 'unseen' chromatic information (Heywood *et al*., 1991), and patients with fields of cortical blindness can guess the presence, position, and identity of stimuli in their blind field.

The latter phenomenon, termed 'blindsight' (Weiskrantz *et al*., 1974), is one of the best-known examples of implicit functions. It demonstrates that a visual input processed in the extra-geniculostriate cortical system can be put to behavioral use even when it is not consciously represented (Stoerig and Cowey, 1997). Whether the dissociation between visually guided responses and consciously acknowledged vision that characterizes 'blindsight' also occurs in monkeys with hemianopia due to unilateral V1/V2 lesions has been tested using a combination of two forced-choice paradigms (Cowey and Stoerig, 1995). In a first step, the monkeys learned to almost perfectly localize targets in their blind field, as was expected from extensive previous research (Weiskrantz and Cowey, 1970; Pasik and Pasik, 1982). In order to then 'ask' them whether they had seen the targets to which they had responded (something that is usually done verbally with human subjects), in addition to localizing a target wherever it had appeared by pressing its position, as before, they had to respond to newly introduced blank trials by pressing a no-stimulus response area on the monitor whenever no stimulus was given under otherwise identical conditions. Once they had mastered this new signal detection task in their normal visual field, stimuli that they could localize well in the blind field were presented there, interspersed with blanks and stimuli in the normal hemifield, in order to determine whether they would touch them or whether they would touch the no-stimulus field instead. As they consistently chose the latter option, it seems that monkeys also have blindsight, and not some type of reduced but conscious vision in their visual field defects (Stoerig *et al*., 2002).

Consciousness in Animals

The results obtained for blindsight in monkeys suggest that they, like humans, lose conscious vision when their primary visual cortex is lesioned. They do not show that monkeys normally have conscious vision, but instead they posit that this is the case. The same is true for the experiments on binocular rivalry in monkeys, which do not demonstrate that the stimulus the monkey indicates as seen is indeed consciously represented, but which assume that this is so and then attempt to elucidate the difference in the neuronal processing of the dominant and suppressed stimuli. Old World monkeys in particular have a visual system whose function and architecture are very similar to ours. Although this does not prove that they have conscious vision, it makes it likely for those who assume that conscious vision is but one of the functions performed by the system.

Studies that explicitly address the question of consciousness in animals have largely focused on demanding cognitive tasks, and then discussed whether the ample evidence for communication, scheming, planning, and abstracting thus revealed is sufficient to allow the conclusion that their subjects need to be conscious of the task at hand (Griffin, 1976; Ristau, 1991). It is obvious that the answer will depend on what one takes the function of consciousness to be; if it had none, obviously every cognitive task could be solved without consciousness ('Zombie-hypothesis'); if it was necessary for creative problem solving, then all animals capable of complex behavior need it (Stamp Dawkins, 1998). As all consciousness *of* requires the organism to be in a conscious state, and mammals, birds, fish, and even insects lose consciousness just like humans as a result of appropriate brain lesions or under general anesthesia, it seems

straightforward to assume that they are conscious when brain-normal and not anesthetized; after all, this reasonable assumption is the rationale for using general anesthesia during surgery on animals. What they are or can be conscious of when in a conscious state should depend on their sensorium as well as their capacity to deal with the information in terms of planning, memorizing, problem-solving, abstraction, attribution, enjoyment, and so on. Animals whose sensorium includes echolocation and sideline organs will have sensations that we cannot imagine, but this in no way precludes the sensation's conscious representation. In view of the limited behavioral repertoire that we see in our unconscious conspecifics, animals that are able to communicate and respond to challenges in a creative manner are very likely to be conscious in both senses of the term.

It seems obvious that an organism needs to be in a conscious state in order to be conscious of anything, but the reverse dependence – that is, the state depending on a conscious content of whatever denomination – appears to be less certain. However, the effects first on content (in the form of hallucinations), and later on the conscious state, caused by both sensory deprivation and the anesthetic ketamine may indeed be indicative of an interdependence. If the state and content of consciousness were indeed mutually interdependent, the function of the specific systems that furnish the content would depend on the nonspecific system mediating the state, and equally the nonspecific system's function would depend on some specific system's ability to produce a conscious sensation. Evidence for such interaction (e.g., between the ILN and the reticular formation on the one hand, and specific thalamic and cortical responses on the other hand) shows that stimulation of nonspecific nuclei has a substantial influence on the processing of sensory information (Jasper, 1949; Munk *et al*., 1996). Furthermore, recent research has revealed that nonspecific and specific neurons are quite intermingled, and do not respect the traditional boundaries of the nuclei that are attributed to one or the other system (Jones, 1998), suggesting that interaction between the specific and nonspecific systems is much more common than was previously suspected.

CONCLUSION

Consciousness is a state that is lost in coma and general anesthesia, and altered in sleep. It enables the organism to have conscious experiences, communicate purposefully, and respond in a novel, adaptive fashion to unexpected events. In mammals, it appears to be mediated by a nonspecific system of neurons distributed in the brainstem, thalamus, and cortex, which interacts with the specific systems that mediate the content of consciousness. Thalamocortical structures are currently the most likely site both for making the conscious representation (visual or otherwise) and for the interaction between the systems. To date, the evidence indicates that the correlate of an individual conscious representation consists of a network of activated neurons which may require a minimum number of participants, a minimum duration, and a cortical component with a minimum level of activation and complexity.

References

Alkire MT, Haier RJ, Barker SJ *et al*. (1995) Cerebral metabolism during propofol anesthesia in humans studied with positron emission tomography. *Anesthesiology* **82**: 393–403.

Alkire MT, Haier RJ and Fallon JH (2000) Towards a unified theory of narcosis: brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness and Cognition* **9**: 370–386.

Alpern M (1953) Metacontrast. *Journal of the Optical Society of America* **43**: 648–657.

Amassian VE, Cracco RQ, Maccabe PJ *et al*. (1989) Suppression of visual perception by magnetic coil stimulation of human occipital cortex. *Electroencephalography and Clinical Neurophysiology* **74**: 458–462.

Andrade J (1995) Learning during anesthesia: a review. *British Journal of Psychology* **86**: 479–506.

Atkinson AP, Thomas MSC and Cleeremans A (2000) Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences* **4**: 372–382.

Austin GM and Grant FC (1955) Physiologic observations following total hemispherectomy in man. *Surgery* **38**: 239–258.

Benson DF and Greenberg JP (1969) Visual form agnosia. *Archives of Neurology* **20**: 82–89.

Blake RR (1989) A neural theory of binocular rivalry. *Psychological Review* **96**: 145–167.

Bogen JE (1995) On the neurophysiology of consciousness. I. An overview. *Consciousness and Cognition* **4**: 52–62.

Bogen JE (1997) Some neurophysiologic aspects of consciousness. *Seminars in Neurology* **17**: 95–103.

Bonebakker AE, Bonke B, Klein J *et al*. (1996) Information processing during general anesthesia: evidence for unconscious memory. *Memory and Cognition* **24**: 766–776.

Bosinelli M (1995) Mind and consciousness during sleep. *Behavioural Brain Research* **69**: 195–201.

Breitmeyer B (1984) *Visual Masking: an Integrative Approach*. Oxford, UK: Clarendon Press.

Bridgeman N (1980) Temporal response characteristics of cells in monkey striate cortex measured with metacontrast masking and brightness discrimination. *Brain Research* **196**: 347–364.

Bruyer R, Laterre C, Seron X *et al*. (1983) A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition* **2**: 257–284.

Cariani P (2000) Anesthesia, neural information processing and conscious awareness. *Consciousness and Cognition* **9**: 387–395.

Cowey A and Stoerig P (1995) Blindsight in monkeys. *Nature* **373**: 247–249.

Crick F and Koch C (1998) Consciousness and neuroscience. *Cerebral Cortex* **8**: 97–107.

Fiset P, Paus T, Daloze T *et al*. (1999) Brain mechanisms of propofol-induced loss of consciousness in humans: a positron emission tomography study. *Journal of Neuroscience* **19**: 5506–5513.

Foulkes WD (1962) Dream reports from different stages of sleep. *Journal of Abnormal and Social Psychology* **65**: 14–25.

Greenfield S (1995) *Journey to the Centres of the Mind*. WH Freeman: San Francisco.

Gregory RL (1987) Consciousness. In: Gregory RL (ed.) *The Oxford Companion to the Mind*, pp. 160–164. Oxford, UK: Oxford University Press.

Griffin DR (1976) *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. New York: Rockefeller University Press.

Grüsser OJ and Landis T (1991) *Visual Agnosias and Other Disturbances of Visual Perception and Cognition*. London, UK: Macmillan.

Heywood CA, Cowey A and Newcombe F (1991) Chromatic discrimination in a cortically colour-blind observer. *European Journal of Neuroscience* **3**: 802–812.

Hobson JA and Steriade M (1986) The neuronal basis of behavioral state control: internal regulatory systems of the brain. In: Bloom F and Mountcastle V (eds) *Handbook of Physiology* vol. 4, pp. 701–823. American Physiological Society.

Jasper HH (1949) Diffuse projection systems: the integrative action of the thalamic reticular system. *Electroencephalography and Clinical Neurophysiology* **1**: 405–420.

Jennett B and Plum F (1972) Persistent vegetative state after brain damage. *Lancet* **1**: 734–737.

Jones EG (1998) A new view of specific and nonspecific thalamocortical connections. In: Jasper HH, Descarries L, Castellucci VF and Rossignol S (eds) *Consciousness at the Frontiers of Neuroscience. Advances in Neurology* vol. 77, pp. 49–71. Philadelphia: Lippincott-Ravell.

Kinney HC, Korein J, Panigrahy A, Dikkes P and Goode R (1994) Neuropathological findings in the brain of Karen Ann Quinlan. *New England Journal of Medicine* **330**: 1469–1475.

Libet B, Alberts W, Wright E *et al*. (1964) Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex. *Journal of Neurophysiology* **27**: 546–578.

Logothetis NK (1998) Single units and conscious vision. *Philosophical Transactions of the Royal Society of London* **353**: 1801–1818.

Lumer E, Friston K and Rees G (1998) Neural correlates of perceptual rivalry. *Science* **280**: 1930–1934.

Macknick SL and Livingstone MS (1998) Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience* **1**: 144–149.

Maquet P, Dive D, Salmon E *et al*. (1990) Cerebral glucose utilization during sleep–wake cycle in man determined by positron emission tomography and [^{18}F]2-fluoro-2-deoxy-D-glucose method. *Brain Research* **513**: 136–143.

Maquet P, Dive D, Salmon E *et al*. (1992) Cerebral glucose utilization during stage 2 sleep in man. *Brain Research* **571**: 149–153.

Meadows JC (1974) Disturbed perception of colours associated with localized cerebral lesions. *Brain* **97**: 615–632.

Menon DK, Owen AM, Williams EJ *et al*. (1998) Cortical processing in persistent vegetative state. *Lancet* **352**: 200.

Moruzzi G and Magoun W (1994) Brainstem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* **1**: 455–473.

Munk MHJ, Roelfsema PR, König P, Engel AK and Singer W (1996) Role of reticular activation in the modulation of intracortical synchronization. *Science* **272**: 271–274.

O'Brien G and Opie J (1999) A connectionist theory of phenomenal experience. *Behavioural Brain Science* **22**: 127–196.

Pasik P and Pasik T (1982) Visual functions in monkeys after total removal of visual cerebral cortex. *Contributions to Sensory Physiology* **7**: 147–200.

Penfield W (1938) The cerebral cortex in man. I. The cerebral cortex and consciousness. *Archives of Neurology and Psychiatry* **40**: 417–442.

Plum F and Posner JB (1980) *The Diagnosis of Stupor and Coma*. Philadelphia, PA: FA Davies Company.

Polonsky A, Blake R, Braun J and Heeger DJ (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience* **3**: 1153–1159.

Rees G, Kreimann G and Koch C (2002) Neural correlates of consciousness in humans. *Nature Reviews Neuroscience* **3**: 261–270.

Ristau CA (1991) *Cognitive Ethology – The Minds of Other Animals*. Hillsdale, NJ: Lawrence Erlbaum.

Schiff ND, Ribary U, Moreno DR *et al*. (2002) Residual cerebral activity and behavioural fragments can remain in the persistently vegetative brain. *Brain* **125**: 1210–1234.

Schiff N, Ribary U, Plum F and Llìnas R (1999) Words without mind. *Journal of Cognitive Neuroscience* **11**: 650–656.

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Simons DJ and Levin DT (1997) Change blindness. *Trends in Cognitive Sciences* **1**: 261–267.

Singer W (2000) Phenomenal awareness and consciousness from a neurobiological perspective. In: Metzinger T (ed.) *Neural Correlates of Consciousness. Empirical and Conceptual Questions*, pp. 121–137. Cambridge, MA: MIT Press.

Stamp Dawkins M (1998) *Through Our Eyes Only? The Search for Animal Consciousness*. Oxford, UK: Oxford University Press.

Steriade M and McCarley R (1990) *Brainstem Control of Wakefulness and Sleep*. New York, NY: Plenum Press.

Steriade M, Contreras D and Amzica F (1994) Synchronized sleep oscillations and their paroxysmal developments. *Trends in Neuroscience* **17**: 199–208.

Stoerig P (1996) Varieties of vision: from blind processing to conscious recognition. *Trends in Neuroscience* **19**: 401–406.

Stoerig P and Cowey A (1997) Blindsight in man and monkey. *Brain* **129**: 535–559.

Stoerig P, Zontanou A and Cowey A (2002) Aware or unaware: assessment of cortical blindness in four men and a monkey. *Cerebral Cortex* **12**: 565–574.

Taylor J, Holmes G and Walshe FMR (1931) *Selected Writings of John Hughlings Jackson*. London, UK: Hodder & Stoughton.

Weiskrantz L and Cowey A (1970) Filling in the scotoma: a study of residual vision after striate cortex lesions in monkeys. *Progress in Physiological Psychology* **3**: 237–260.

Weiskrantz L, Warrington EK, Sanders MD and Marshall J (1974) Visual capacity in the hemianopic field following a restricted cortical ablation. *Brain* **97**: 709–728.

Wilbrand H and Sänger A (1900) *Die Neurologie des Auges*, vol. III.

Zeki S (1991) Cerebral akinetopsia (visual motion blindness). *Brain* **114**: 811–824.

Further Reading

Edelman GE (1989) *The Remembered Present: a Biological Theory of Consciousness*. New York, NY: Basic Books.

Jasper HH, Descarries L, Castellucci VF and Rossignol S (eds) (1998) *Consciousness at the Frontiers of Neuroscience. Advances in Neurology* vol. 77. Philadelphia, Lippincott-Ravell.

Libet B (1993) *Neurophysiology of Consciousness. Selected Papers and New Essays by Benjamin Libet*. Boston, MA: Birkhäuser.

Metzinger T (ed.) (2000) *Neural Correlates of Consciousness. Empirical and Conceptual Questions*. Cambridge, MA: MIT Press.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford, UK: Oxford University Press.

Savage-Rumbaugh S and Lewin R (1994) *Kanzi: The Ape at the Brink of the Human Mind*. Chichester, UK: John Wiley.

Tononi G and Edelman GE (1998) Consciousness and complexity. *Science* **282**: 1846–1851.

Zeki S (1993) *A Vision of the Brain*. Oxford, UK: Blackwell Science.

Neural Degeneration

Introductory article

Paul John Lucassen, University of Amsterdam, Amsterdam, The Netherlands
Vivi M Heine, University of Amsterdam, Amsterdam, The Netherlands
Karin Boekhoorn, University of Amsterdam, Amsterdam, The Netherlands
Harm Krugers, University of Amsterdam, Amsterdam, The Netherlands

CONTENTS

As adult neurons generally do not divide, dysregulation or unwanted induction of apoptosis at an adult stage can have dramatic consequences for brain function.

INTRODUCTION

During the early development of the central nervous system (CNS), new cell generation and cell division are widespread phenomena. Notably, they often occur in close association with massive cell death. Many of the initially generated cells extend protrusions into their surroundings which contain growth cones; these structures enable subsequent nerve fiber migration by which new cells search for contacts with protrusions from other cells close by. Competition for growth promoting factors (neurotrophins), which are secreted by the target cells, eventually determine the establishment and maintenance of functional neuronal connections. In this way, a myriad of interconnected neurons is formed which is the basis for our brain.

It is during this phase of extensive migration that many of the immature cells either fail to make contact, or establish nonfunctional, unwanted connections. These cells are rapidly eliminated by activation of an intrinsic death program present in every cell, which kills the cell in a well-controlled and orderly manner. This cellular 'suicide' consists of various intermediate steps and is known as 'programmed cell death'. Its morphological appearance is referred to as 'apoptosis', which is different from the more passive, necrotic type of cell death, both of which are discussed below.

A neuron's chance of survival during development depends on the extent of its connections, the neurotrophic support it receives from other neurons and glia cells, and in adulthood, also on the presence of steroid hormones.

ROUTES OF NEURODEGENERATION: APOPTOSIS AND NECROSIS

Traditionally, a cell anywhere in the body had only two ways to die: through apoptosis or necrosis. For over 125 years 'necrosis' was the main term used to describe pathological degeneration such as that following physical injury (i.e. cell 'murder'). This usually involved the death of large groups of cells within a specific region. Necrosis is characterized by initial swelling of the cell, modest condensation of the deoxyribonucleic acid (DNA), occurrence of holes, or vacuoles, in the cytoplasm, breakdown of cell organelles, and rupture of the outer cell membrane. This causes the release of cellular contents in the space surrounding the neurons and often elicits a local inflammatory response.

Although apoptosis (the name is derived from an ancient Greek term for the seasonal dropping of leaves from trees) was recognized in 1885, when specific morphological alterations were observed during egg development, the process went relatively unnoticed for decades. Only in the early 1970s, when apoptosis was described in extensive detail and in relation to pathology, did the term became widely adopted. Subsequently, in the early 1990s, specific histological techniques (*in situ* end labeling, ISEL) were developed that allowed identification of dying cells in tissue sections. Together with the clarification of the molecular pathways of the apoptotic cascade, this has greatly facilitated the study of apoptosis in various disciplines, such as development, immunology, oncology and neurodegeneration.

Apoptosis denotes a 'suicidal' type of cell death that involves the sporadic and self-chosen loss of primarily individual cells, without any tissue inflammation. Although preparation for apoptosis may take days, the actual execution is rapid and

dying cells are detectable only for minutes to hours. Consequently, if massive loss occurs over a few days, only a small minority (< 1%) of apoptotic cells will be visible at any one moment. The chance of 'trapping' ongoing apoptosis in thin tissue sections, taken for example from a person with a chronic brain disorder, is hence very low. This may explain why the magnitude and importance of apoptosis in many conditions has gone unnoticed for so long. Necrosis received much more attention because it was present in larger quantities and for longer periods of time, and hence was more easily detectable in tissue.

An additional feature that distinguishes the two types of cell death is their pattern of DNA fragmentation. The highly condensed and supercoiled DNA that is wrapped around specific nuclear proteins, the nucleosomes, is fragmented during apoptosis. This DNA fragmentation involves the activation of specific enzymes that cut DNA particularly between these nucleosomes. This gives rise to characteristic DNA fragment sizes that are multiples of 200 base pairs, i.e fragments of 400, 600, 800 bp and so on. When separated on agarose gels these produce characteristic DNA 'ladders', whereas in necrosis a 'smear' is found, due to the presence of randomly cut DNA which yields all kinds of different fragment sizes.

Furthermore, a neuron undergoes extensive structural reorganization during apoptosis as it dismantles its nucleus, membranes and dendrites. Under the microscope, the hallmarks of apoptosis are: shrinkage in cellular and nuclear size, membrane blebbing, DNA fragmentation, pycnotic chromatin condensing against the nuclear membrane, and cytoplasmic organelles that remain intact as the cytoplasm and nucleus break up into characteristic apoptotic bodies, which are subsequently cleared by macrophages or neighboring cells (Figure 1). Although initially strictly separate, the morphological distinction between apoptosis and necrosis has gained in nuances over the past years, and intermediate forms also have been observed, with (for example) nuclear enlargement rather than condensation in the very early phase of apoptosis, accompanied by DNA fragmentation.

Initially considered to be 'wasteful', the massive cell death during development is now seen as a fundamental mechanism for controlling cell number. The ability to ablate cells at will is as essential for an organism's survival as the ability to replicate and differentiate them. For example, during brain development, apoptosis eliminates transient brain structures that are only there to support additional cortical layers built on top of them, after which they are no longer needed. As such, apoptosis shapes and sculpts the first structural blueprint of the brain. Moreover, it serves as an important quality control and error correction mechanism, removing cells with inappropriate phenotypes and thus preventing the transmission of genetic errors.

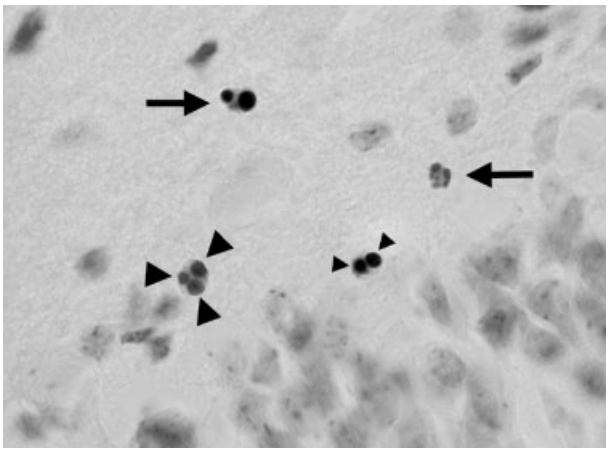


Figure 1. Apoptotic cells (arrowed) in the rat hippocampus. Note the different morphological stages of apoptotic cell death. All include shrinkage in size, dark staining due to *in situ* end labeling, indicating extensive DNA fragmentation, as well as the clear presence of two, three or more apoptotic bodies (arrowheads).

The incidence of apoptosis in the brain rapidly decreases postnatally, making cell death virtually absent in the adult central nervous system, where neuron number, and the network derived from it, remain fairly constant. As adult neurons generally do not divide anymore (with some exceptions), dysregulation or unwanted induction of apoptosis at an adult stage has dramatic consequences for brain function. The underlying molecular mechanisms that control apoptosis appear quite similar in different models and situations. Therefore, knowledge of the factors that regulate developmental apoptosis might improve our understanding of possible ways to attenuate pathological apoptosis in adult neurodegenerative disorders.

MOLECULAR MECHANISMS OF NEURONAL APOPTOSIS

Molecular Mediators

Several gene products have been identified in the pathways that (a) prepare for and (b) eventually execute the apoptotic death of a cell. Initial genetic studies in the worm *Caenorhabditis elegans* have

revealed three genes, *ced-3*, *ced-4* and *ced-9*, whose products were implicated in the developmental cell death in these animals. Later, several mammalian homologs such as interleukin-1 converting enzyme (ICE), with analogous functions, were identified. In contrast, *bcl-2*, a human oncogene overactive in a tumor named follicular lymphoma, was found to block apoptosis. The product of this oncogene, Bcl-2, allows the survival of cells for which death would have been more appropriate, and is therefore generally considered to be a neuronal survival factor. It belongs to a larger protein family with many other members, such as those encoded by *bcl-Xl*, *bcl-Xs*, *bid* and *bax*, each exerting either a specific proapoptotic or antiapoptotic action. The balance between these factors determines whether a cell lives or dies (Figure 2).

This is demonstrated by inducing alterations in specific *bcl-2* family members in a model of neuronal damage. These determine the extent of neuronal death occurring after nerve fiber transsection (axotomy) in a motor system. Increasing the endogenous levels of Bcl-2 reduced motor neuron death after axotomy, as the ratio of Bcl-2 to Bax and other proapoptotic family members is tilted, favoring cell survival. An opposite effect is accomplished in mice lacking the proapoptotic death factor gene *bax*, where approximately 90% of the motor neurons survive the axotomy. This confirms that the balance between specific mediators of the apoptotic cascade strongly influences cell death.

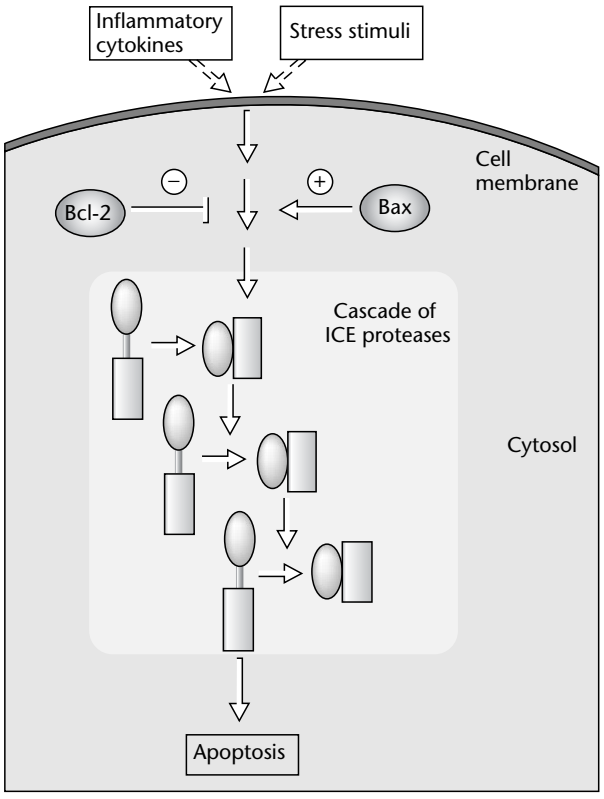


Figure 2. The apoptotic cascade. ICE, interleukin converting enzyme. Redrawn from: Cosulich S and Clarke P (1996) Apoptosis: does stress kill? *Current Biology* **6**: 1586–1588. Copyright 2002, with permission from Elsevier Science.

Apoptosis Signaling Routes

Elucidation of the molecular biology of apoptosis has, so far, revealed a complex suicide program. However, although apoptosis can be triggered by a wide variety of cellular stresses, including DNA damage, ultraviolet radiation, ionizing radiation, heat shock and oxidative stress, as well as by extracellular stimuli acting through activation of specific receptors at the cell surface, most cells undergoing apoptosis go through a similar series of changes that are well conserved across different cell types and models. Interestingly, various apoptotic triggers all converge to activate a single group of enzymes, which cleave their substrates after specific aspartic residues and are named 'caspases'.

All the biochemical and morphological alterations occurring during apoptosis, such as DNA fragmentation, membrane blebbing and the formation of apoptotic bodies, are the direct consequence of the activation of one or more of these caspases. In living cells caspases are generally inactive, but they can be activated by specific death receptors, such as the Fas receptor or tumor necrosis factor (TNF) receptor at the cell surface. The Fas ligand and TNF together form a complex that eventually recruits procaspase 8 in the cell. The subsequent activation of this enzyme activates other downstream caspases, such as caspase 3, caspase 6 and caspase 7, which are involved in the actual execution of cell death.

Another apoptotic route through which caspases can be activated is by the release of cytochrome C (CytC) from the mitochondria. This molecule is involved in electron transfer important for energy metabolism, and is exclusively located in the intermembrane space of mitochondria. During apoptosis, the outer membranes of the mitochondria become permeable to CytC, which can then bind a cytosolic factor called Apaf-1. This CytC–Apaf-1 complex recruits and activates procaspase 9, which in turn activates the downstream caspases 3, 6 and 7. Interestingly, CytC release not only initiates caspase activation, but also disrupts mitochondrial function, diminishing the energy essential for neuronal function and generating more

damaging oxygen species, further contributing to the neurons' demise. The CytC pathway in particular is regulated by various members of the Bcl-2 protein family.

Calcium Signaling

Another signaling mechanism that is particularly relevant to neuronal degeneration involves the redistribution of intracellular calcium ions (Ca^{2+}). Various physiological stimuli increase intracellular Ca^{2+} levels in a transient way. Under pathological stimulation conditions, however, Ca^{2+} levels can become elevated for prolonged periods. Although neurons are equipped with various Ca^{2+} buffering systems, high Ca^{2+} levels increase neuronal vulnerability, reducing neuronal function as well as viability. Extrusion of calcium ions is necessary, yet costly in terms of energy demand. Chronic calcium overload therefore leads to exaggerated energy expenditure. Furthermore, excess calcium ions can activate damaging enzymes that attack the cytoskeleton and DNA. Interestingly, steroid hormones also have an important role in maintaining neuronal Ca^{2+} levels (see below).

APOPTOSIS IN NEURODEGENERATIVE DISEASE

In several human neurodegenerative disorders, neuronal loss is obvious. Not surprisingly, in many of them a causal role for apoptosis has been suspected, either in selected neuronal populations, or in relation to genetic defects affecting the entire brain or spinal cord. An example of the latter is the most severe form of human spinal muscular atrophy, Werdnig–Hoffmann disease. This disorder is caused by an autosomal recessive mutation in a neuronal apoptosis inhibitory protein (NAIP). Under normal conditions this gene product inhibits apoptosis in vertebrate cells. In Werdnig–Hoffmann disease, however, the mutation in this protein induces failure to inhibit apoptosis in early development, and the consequent massive death of motor neurons during the perinatal period results in respiratory failure and death at a young age.

In many other neurodegenerative disorders, aberrant apoptosis is, however, not linked to such a gene mutation. Although end-stage diseased human brain tissue often shows increased expression of apoptosis-associated genes, it is difficult to prove a causal role for apoptosis, owing to its rapid time kinetics, and its probable early occurrence in the disorder. Nevertheless, apoptotic neurodegeneration is also suspected in other systems and disorders as diverse as the spinal cord in amyotrophic lateral sclerosis, the nigrostriatal system in Parkinson disease and the basal ganglia in Huntington disease. Disturbances in memory and higher cognitive functions have been linked to apoptosis, involving brain areas such as the hippocampus and cortex in Parkinson and Huntington disease, acquired immune deficiency syndrome (AIDS) dementia complex, stroke, prion diseases and Alzheimer disease.

Stroke

Stroke or cerebrovascular accident (CVA) is a major cause of death and physical impairment. It can be caused by occlusion, rupture or bleeding of one of the cerebral vessels, resulting in cerebral ischemia. Stroke may be of ischemic or hemorrhagic origin: ischemic CVA is caused by occlusion of one of the major arteries or small perforating vessels, which occurs in 80% of all CVA cases; bleeding from one of the intracerebral or the subarachnoidal arteries results in hemorrhagic CVA. (*See* **Stroke**)

The incidence of stroke is between 150 and 250 per 100 000 persons. This incidence rises sharply with increasing age (up to 2000 per 100 000 in those aged 80 years and over). About 60–80% of people who have suffered a stroke die within 30 days, while of those who survive, many remain disabled. Extensive research is therefore carried out to explore possible therapeutic interventions.

In primates, the normal cerebral blood flow is 50–60 mL per 100 g per minute. Profound reduction of cerebral blood flow (below 10–15 mL min^{-1}) produces rapid neuronal cell death. Importantly, between this ischemic core and the normally perfused brain there is tissue in which blood flow is moderately reduced, which forms a target for acute therapeutic intervention. The amount of blood flow in this zone (the penumbra) depends largely on the collateral blood supply from surrounding arteries. Ultimately, oxygen delivery becomes insufficient in the penumbra to allow normal energy metabolism. This results in lactic acidosis and further reduces the production of the main energy source of ionic pumps, adenosine triphosphate, in turn disturbing the homeostasis of cellular ionic processes, and the ensuing rapid loss of neuronal potassium ions triggers massive depolarization. This opens voltage-sensitive Ca^{2+} channels in the cell wall and leads to a sustained influx of calcium ions (Figure 3).

Prolonged periods of elevated intracellular Ca^{2+} levels can lead to the generation of reactive oxygen species, free radicals. They can cause severe

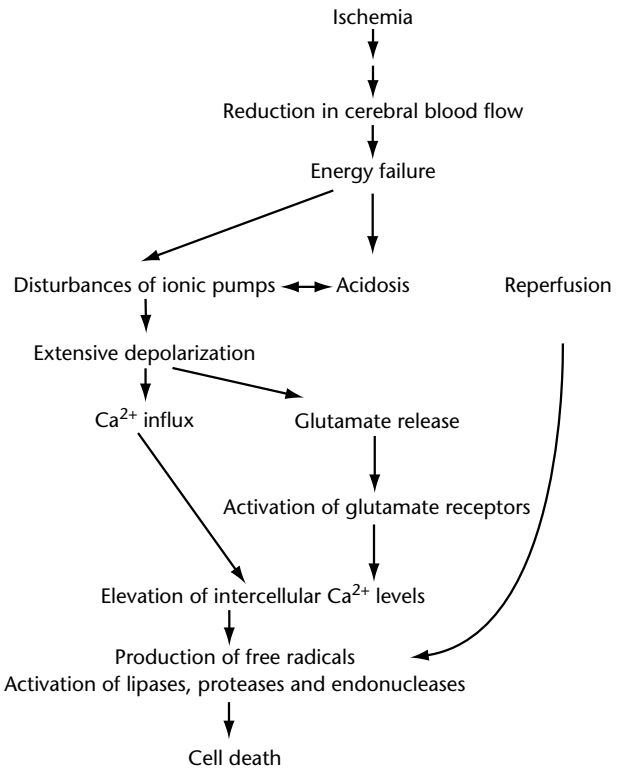


Figure 3. Sequence of events that might ultimately result in neuronal degeneration.

damage to membrane proteins, impairing membrane integrity which results in leakiness to substances that do normally not cross the cellular membrane. A second cytotoxic pathway activated by excessive intracellular Ca^{2+} levels involves Ca^{2+}-dependent phospholipases, proteases and endonucleases, all of which have deleterious consequences for cellular function. After ischemia necrotic cell death is accompanied by apoptosis in the penumbral area. When blood flow is restored (reperfusion), enhanced free radical production may further increase the oxidative damage.

Alzheimer Disease

Alzheimer disease (AD) is a well-known type of progressive dementia with a high incidence in the elderly population. Patients suffer from severe memory and cognitive disturbances, which involve selective neuronal circuits in the neocortex, hippocampus and cholinergic system. Conclusive diagnosis depends on postmortem histopathological findings, on the basis of quantitative rather than qualitative scoring of the main neuropathological hallmarks: these are extracellular (senile) plaques, associated with an amyloid protein, and dystrophic neurites and neurofibrillary tangles composed of an altered cytoskeletal protein, tau. (*See* **Alzheimer Disease**)

The early-onset forms of AD are mostly familial and show an autosomal dominant inheritance pattern. This small group (2–3% of all cases) involves mutations in three identified genes: those encoding the amyloid precursor protein (APP), presenilin 1 (PS-1) and presenilin 2 (PS-2). These mutations all induce an altered processing of APP, which gives rise to aberrant levels of amyloid beta (Aβ) protein. The normal biological function of APP, the presenilins and how they cause AD remains so far elusive, but most PS-1 mutations alter APP processing and increase the relative Aβ levels, which may stimulate apoptosis through oxidative damage or through modulation of apoptosis mediators.

A fourth gene important in AD (*APOE*) encodes apolipoprotein E, which has three isoforms: ε2, ε3 and ε4. Apolipoprotein E is involved in lipid transport in the periphery and has a separate pool in the brain. Epidemiological studies have shown that people with one or two copies of the ε4 allele have increased risks of developing AD earlier and in a more severe form. Another new and interesting aspect to AD is the possible contribution of frame shift mutations at the ribonucleic acid level which can give rise to toxic, or aberrant, proteins.

Early studies *in vitro* pointed to a neurotoxic and primary role for Aβ. This would induce massive cell death, which could explain the considerable weight loss in the brain observed in AD. Indeed, altered expression of apoptosis-related proteins, such as Bax, Bcl-2 and Par-4, has been found in AD brain. The observation in AD of many cells containing fragmented DNA has also been interpreted as evidence for apoptosis. However, necrosis and DNA damage were also detected with these techniques, and so far, no morphologically convincing evidence for apoptosis in AD has been observed.

Detailed morphometric studies have now revealed that cell loss in AD is certainly not massive, but takes place only in a few limited yet important brain areas: the locus coeruleus, the CA1 area of the hippocampus, and layer II of the entorhinal cortex. Specific unbiased counting techniques have failed to demonstrate any difference in total neuron number between the neocortex of control and AD cases. Massive cell loss due to apoptosis is thus unlikely in AD. More and more research contradicts the notion that Aβ is an early and toxic factor in AD; not only is there a poor relation between amyloid deposits and cognition, many studies *in vitro* have even established

growth-promoting capacities of amyloid fragments. Also, in various mouse models in which Aβ levels are increased, learning deficits do indeed occur, but do so months prior to the formation of Aβ deposits, and in absence of massive hippocampal cell loss in most models. This suggests that Aβ is not acutely neurotoxic, and at least not sufficient to cause pathological changes or apoptosis in AD on its own. The considerable weight loss is now attributed to atrophy and shrinkage of neurons rather than to neuronal loss.

HORMONAL CONTROL OF NEURONAL APOPTOSIS

Hormones, in particular steroid hormones, affect various processes in the brain, such as learning and memory, and notably also neuronal viability. In the adult vertebrate nervous system there are several examples of steroid-related neural degeneration. One example is the seasonal turnover of specific neuronal populations in the adult canary brain. These populations are considered to be a neural substrate for learning specific song patterns. During spring, when circulating levels of testosterone are highest, canaries sing frequently and the turnover rate of projection neurons in the song control region – the high vocal center (HVC) – is relatively low. In contrast, in autumn when testosterone levels drop and song patterns are modified, considerable numbers of older neurons in the HVC die, in parallel with incorporation of new neurons into the HVC.

In another example, steroids have been shown to be crucial for neuronal viability and function in the adult rat. This is apparent in the hippocampus, a brain structure involved in learning and memory. In intact animals, the steroid hormone corticosterone (Cort) is released from the adrenal gland in response to stress. Cort readily occupies two types of nuclear receptors which are present in large densities in the hippocampus and brain: the high-affinity mineralocorticoid receptor (MR) and the lower-affinity glucocorticoid receptor (GR). Owing to the difference in affinity the MR is almost always occupied, whereas the GR is only occupied during stress. Upon binding of Cort, the ligand–receptor complex translocates to the nucleus, where it binds to DNA and interferes with gene transcription for prolonged periods. After a rise in Cort, hormone binding in the brain causes adrenal Cort production to shut off again. This feedback inhibition of the stress system helps maintain Cort levels within physiological boundaries: this is important as both too high and too low corticosterone levels affect neuronal viability. Depletion of Cort by removing the adrenal glands induces massive apoptosis in the hippocampus three days later (Figure 1). This apoptotic cell loss directly affects hippocampal function and memory and, interestingly, can be completely prevented by replacing adrenalectomized animals with low levels of Cort just sufficient to occupy MR. Cort and MR are thus essential for neuronal survival in the rat hippocampus. (*See* **Hippocampus**; **Apraxia**)

Over-exposure to steroids too can affect neuronal function and viability. As the hippocampus exerts an (indirect) inhibitory control on Cort secretion, hippocampal damage or cell loss can disrupt the feedback inhibition and induce a feedforward cascade of cumulative hormone exposure. In rat, prolonged stress or overexposure to Cort impairs memory and learning as well as electrical activity of the hippocampus. Also, transient (but still reversible) atrophy of neurons is induced. Following prolonged Cort exposure, even massive neuronal loss has been reported, in rat and monkey hippocampus in some but certainly not all studies. This suggested that chronic stress or steroid overexposure kills neurons through apoptosis. This seemed relevant to human aging and AD, where the extent of hippocampal atrophy and disease correlate well with the moderately elevated Cort plasma levels.

A human condition even more relevant for steroid overexposure is major depression, in which plasma levels of cortisol and other indications of enhanced stress axis activity are much more pronounced – indeed, decreased hippocampal volume and adrenal enlargement are commonly observed. Hence, hippocampal damage would be expected. However, little (if any) neuropathological change, or evidence of massive cell death could be found in the hippocampus of people with major depressive disorder or steroid-treated patients. This agrees with the general clinical experience of depression, that treatment can relieve both the depressive symptoms and hippocampal atrophy, which would be unlikely if many cells had indeed been lost. In conclusion, although hippocampal function can certainly be affected by steroid overexposure, hippocampal structure shows reversible and adaptive rather than neurotoxic phenomena, particularly in the case of primates, including humans. (*See* **Affective Disorders: Depression and Mania**)

Modulation of Ischemic Neurodegeneration by Corticosteroids

Neurological insults such as ischemia activate the hypothalamic–pituitary–adrenal axis, with a

subsequent rise in Cort levels that bind MR and GR in brain. Considerable evidence indicates that Cort increases cellular vulnerability in the hippocampus when combined with additional challenges such as ischemia. This glucocorticoid endangerment potentiates the rise in extracellular glutamate levels after an insult and enhances intracellular Ca^{2+} levels. In contrast, low levels of Cort have been shown to reduce neuronal damage after ischemic insults, promoting neuronal viability.

In animal studies this has stimulated research to assess whether therapeutically relevant interventions might exert neuroprotective actions. Using the steroid synthesis inhibitor metyrapone it was shown that preventing the rise in Cort concentration during or after ischemia reduced neuronal damage and preserved hippocampal function. This was accompanied by attenuation of ischemia-related changes. However, metyrapone is likely to have more effects than reducing Cort levels alone. For example, other steroids with potential neuroprotective actions are found to have elevated levels after treatment with metyrapone. In addition, since activation of GR is considered to enhance neuronal vulnerability, one might speculate that GR antagonists such as RU 38486 would have neuroprotective properties. However, RU 38486 appears to be ineffective *in vivo* in preventing neuronal damage after ischemia. Thus, while elevated corticosteroid levels are generally believed to increase neuronal vulnerability, it is far from clear whether interventions that reduce Cort levels or prevent Cort from binding GR are of therapeutic importance.

CONCLUSION

In at least some neurodegenerative disorders, neuronal apoptosis is likely to be involved. More important than the question of whether or not apoptosis is involved (or detectable) is the question of which factors induce cell death and degeneration. For many conditions this is not known, but may be related to an enhanced vulnerability to triggers of apoptosis, or to a tissue-specific environment that favors apoptosis. However, neither of these have been proved. Also, cell death in postmitotic systems in neurodegenerative progressive disorders may in this respect be different from classic developmental apoptosis.

Steroids can interfere with and modulate the extent of neuronal damage in specific models. However, as yet there is no effective way to prevent neuron loss in neurodegenerative disorders such as stroke and AD, even though several candidate therapeutic agents have been developed based on interference with the apoptotic cascade. However, it is important to realize that these treatments can only be effective when administration is possible at the right time (i.e. early in the disorder) and in the right place (i.e. in a selected brain area), and preferably only when the factors that initially induced cell death (whether intrinsic, like local tissue environment conditions, or extrinsic, like free radicals), have been identified and are no longer operative. Furthermore, systemic inhibition of apoptosis throughout the body, affecting for example the immune system, should for obvious reasons be prevented. As apoptosis outside the adult brain is often highly functional, inhibition might have serious consequences for normal physiology.

Despite considerable progress in the clarification of the mechanisms of programmed cell death, the entire intracellular cascade leading to neuronal death in chronic progressive neurodegenerative disorders remains to be elucidated. Understanding these mechanisms in detail is crucial for the rational development of protective strategies and novel therapies for these devastating disorders.

Further Reading

De Keyser J, Sulter G and Luiten PG (1999) Clinical trials with neuroprotective drugs in acute ischaemic stroke: are we doing the right thing? *Trends in Neurosciences* **22**: 535–540.

De Vrij FM, Sluijs JA, Gregori L *et al*. (2001) Mutant ubiquitin expressed in Alzheimer's disease causes neuronal death. *FASEB* **15**(14): 2680–2688.

Krugers HJ, Maslam S, Korf J and Joels M (2000) The corticosterone synthesis inhibitor metyrapone prevents hypoxia/ischemia-induced loss of synaptic function in the rat hippocampus. *Stroke* **31**: 1162–1172.

Lucassen PJ, Muller M, Holsboer F *et al*. (2001) Hippocampal apoptosis in major depression is a minor event and absent from subareas predicted to be at risk for glucocorticoid overexposure. *American Journal of Pathology* **158**: 453–468.

Majno G and Joris I (1995) Apoptosis, oncosis, and necrosis. An overview of cell death. *American Journal of Pathology* **146**: 3–15.

Raff M (1998) Cell suicide for beginners. *Nature* **396**: 119–122.

Swaab DF, Lucassen PJ, Hofman MA *et al*. (1998) Reduced neuronal activity and reactivation in Alzheimer's disease. In: Van Leeuwen FW, Salehi A, Giger RJ *et al*. (eds) *Progress in Brain Research*, vol. 117, Neuronal Degeneration and Regeneration: from Basic Mechanisms to Prospects for Therapy, pp. 343–377. Amsterdam: Elsevier Science.

Neural Development, Models of

Introductory article

Geoffrey J Goodhill, Georgetown University Medical Center, Washington, DC, USA

CONTENTS

Models of neural development attempt to produce precise mathematical and computational descriptions of the biological mechanisms that regulate the formation of the nervous system.

INTRODUCTION

Neural development is the process whereby the nervous system of a developing animal is formed. It involves several stages, each of which is regulated by different types of mechanisms. At each stage a complex interaction of genetic and environmental influences determines the final outcome. The effect of the environment becomes especially important in later stages of development, so that the mature nervous system is specifically adapted to the environment in which it must survive. Mathematical/computational models (hereafter referred to as 'models') have been formulated for several stages of development. As with all modeling, the goal is to take qualitative hypotheses about possible mechanisms and make them quantitatively precise. This rigorously defines the conditions under which such qualitative hypotheses are indeed plausible candidate mechanisms, and makes explicit the implicit assumptions that often underlie them. A desirable result of such modeling is that it makes novel and testable predictions about the outcome of experiments.

Neural development can be conceptually divided into four stages, although they overlap temporally to some degree. These stages are neural induction and neural-tube formation, pattern formation and regionalization, establishment of connectivity, and activity-dependent refinement of connections. The majority of modeling research has concentrated on the last of these stages, attempting to understand the principles by which connection strengths between neurons are adjusted in response to specific patterns of neural activity from the sense organs. This article will consider these stages sequentially.

NEURAL INDUCTION AND NEURAL-TUBE FORMATION

The first stage of neural development is the specification of neural ectoderm by the notochord, which occurs within the first few weeks of gestation in humans. This region, which is called the neural plate, then folds in on itself to form the neural tube. The neural tube gives rise to the peripheral and central nervous systems. As the neural folds come together to form the neural tube, neural-crest cells migrate from the neural folds away from the neural tube. The neural crest forms the peripheral nervous system, while the rest of the neural tube forms the central nervous system. The molecular interactions that lead to the specification of one region of the gastrula as the neural plate have been extensively investigated experimentally. Diffusion of certain molecules from the organizer region is crucial, and mathematical modeling of diffusion processes in general is very well understood. However, application in this area is complicated by the fact that these molecules may be transported by mechanisms other than free diffusion. Furthermore, as yet there are virtually no models of the mechanical forces that drive the formation of the neural tube.

PATTERN FORMATION

In the second stage of neural development, the neural tube expands and changes shape as new cells are generated and migrate to new locations. In conjunction with this, certain regions start to become specialized for particular fates. Moving from anterior to posterior, boundaries begin to form between the prosencephalon (embryonic forebrain), mesencephalon (embryonic midbrain) and rhombencephalon (embryonic hindbrain). Further subdivisions then occur as the prosencephalon divides into the telencephalon and diencephalon, and the rhombencephalon divides into the

metencephalon and myelencephalon. Within these general regions discrete structures are constructed, such as the hippocampus and cerebrum from the telencephalon, the retina and thalamus from the diencephalon, and the cerebellum and pons from the metencephalon. The main question that modeling can attempt to address at this stage of development is how such a spatially non-uniform structure can emerge from an apparently spatially uniform tube.

Reaction–Diffusion Systems

One way to approach this question is to treat it as a very general one. Pattern formation occurs in many other biological systems, such as the spatially regular pattern in which leaves are added to a plant stem, and the division of a fly embryo into individual body segments. In fact, evolutionary conservation of genetic programs has led to some of the same genes controlling regionalization of the nervous system as control fly body segment formation. Pattern formation has also been studied in many physical systems, such as convection patterns in heated liquids and certain types of chemical reactions. The most influential class of mathematical models in this area has been that of reaction–diffusion systems, first proposed by Turing in 1952 and subsequently developed in specific biological contexts by Meinhardt. The principal assumptions are that cells in an initially spatially uniform array produce several chemicals or 'morphogens' that diffuse at different rates from cell to cell through the array. In its simplest form there is one 'activator' chemical and one 'inhibitor', often called an activator–inhibitor system. The activator diffuses slowly and stimulates its own production, while the inhibitor diffuses more rapidly and suppresses production of the activator. In any real system some noise will be present, which means that each cell in the array will not initially have exactly the same concentrations of morphogen. The activator–inhibitor system amplifies such small fluctuations dramatically, leading to a stable state of regularly spaced regions dominated by either the activator or the inhibitor. This elegant mathematical idea has been developed in numerous ways to model biological pattern formation. One aspect of neural development where such algorithms may operate is the selection of just one cell in a proneural cluster to become a neuroblast. The ligand Delta, operating through its receptor Notch, acts as a lateral inhibitor, so that if one cell is specified as neural, then neighboring cells are inhibited from assuming that fate.

Morphogen Gradients

However, there is no evidence that reaction–diffusion mechanisms underlie the formation of overall regionalization in the embryonic nervous system. Rather than asking how symmetry is broken in a spatially uniform array, the key question seems to be how patterns arise from an initial gradient of some morphogen over the array. This initial polarization is created by the positioning of the egg inside the mother. Wolpert first framed this question in 1969 as the 'French flag' problem, and subsequently discussed how various types of chemical reaction kinetics could transform an initially smoothly varying morphogen concentration into sharply defined regions of different concentrations. The fundamental process by which the nervous system becomes initially regionalized seems to be a biochemical cascade, whereby initial gradients of maternal gene products turn on genes in a concentration-dependent manner, creating gradients of new proteins and so on, to form structure at ever finer levels of detail. One of the first signs of patterning along the anterior–posterior axis of the embryo is the expression of homeobox genes of the Pax family in different regions. These gene products then control the regional expression of other genes. As yet there has been no theoretical research addressing the structure and outcome of these complex networks of interacting proteins in the context of nervous system regionalization. However, some models have recently been proposed for fly segment formation, which may turn out to be relevant to the nervous system of vertebrates as well. In such models, certain gene products are assumed to regulate the expression of other genes in a network, somewhat akin to an artificial neural network with weighted connections. Gene products also diffuse within the embryo. Starting initially with a network of random weights (strength of regulatory influences), the outcome of the developmental process for this network is calculated. Using methods such as gradient descent and simulated annealing, weights are then incrementally updated until the developmental outcome matches that which is seen biologically. These optimal weights represent quantitative predictions of the strengths of the regulatory influences that exist in the real biological system, and predictions can be made about the effect of specific gene deletions.

ESTABLISHMENT OF CONNECTIVITY

The task that is achieved in the third stage of neural development is the establishment of initial patterns

of connectivity between neurons. Structures such as the retina send out axons which navigate through a complex environment to find appropriate targets (e.g. the superior colliculus and lateral geniculate nucleus). On their journey, they often have to make many guidance decisions, where for instance some axons in a group are routed in one direction and other axons are routed in another direction (e.g. the split between axons originating from the nasal and temporal halves of the retina that occurs at the optic chiasm). On reaching the correct target region, both the tip of the axon and the target have to change their structure in order to form a specialized synapse. The principal way in which the growing axon senses its environment is via the growth cone (a specialized structure at its tip). The growth cone continuously extends and retracts thin filopodia which sample the local region. The types of molecular guidance cues that can be detected include pathways of growth-permissive factors (e.g. laminin), barriers of growth-inhibitory factors (e.g. myelin-associated glycoprotein) and concentration gradients of molecules (e.g. netrins and semaphorins) which can be attractive or repulsive. These cues can be bound to or expressed by the substrate, or they can diffuse more or less freely. This type of diffusion process (e.g. from a target region) is one way in which a concentration gradient can be established.

Models of Growth-Cone Guidance

Models have been proposed that address several different types of questions at this stage of development. One class considers the types of gradient shapes that axons might encounter *in vivo* or *in vitro*, and how this shape constrains the maximum distance over which it is possible for an axon to be guided. Such models assume that gradient detection can only occur when two principal constraints are satisfied. First, the absolute concentration of the guidance factor at the growth cone must exceed a certain minimum, and secondly, the fractional change in concentration across the growth cone must exceed a certain steepness. Such models predict that the maximum distance over which a growth cone could be guided *in vivo* is about 1 mm for a target-derived diffusible factor, and about 1 cm for a substrate-bound gradient with shape optimized for maximum guidance length. Both of these are reasonably consistent with experimentally measured values. Another class of models attempts to understand how growth-cone gradient detection is limited by noise. The external ligand gradient causes different numbers of receptors to be bound on one side of the growth cone than the other. However, since receptor binding is a stochastic process, at each instant a small difference from the mean level of binding will be measured. One influential hypothesis has been that the threshold for reliable gradient detection occurs when the average size of these small differences becomes comparable with the true difference in concentration between the two sides of the growth cone. Such models can produce numerical values for growth cones of about 1%, which is similar to experimentally measured values. Another class of models focuses on the role of filopodia in growth-cone guidance. Such models propose rules (usually stochastic) with regard to the way in which filopodia are generated in response to environmental cues. Complete axonal trajectories can then be simulated, and both short-term and filopodial dynamics and overall trajectories can then be compared with reality. The mechanisms whereby a difference in receptor binding is internally amplified to give a turning response toward or away from the direction of higher concentration have not yet been considered for growth cones. This is partly because the signal transduction pathways in the growth cone that cause this response are only how beginning to be revealed experimentally. However, several models which may be applicable have been proposed for analogous signaling pathways in bacteria, white blood cells and slime molds. One hypothesis is that the polarization of a cell in response to the gradient may occur by a similar type of reaction–diffusion mechanism previously described in the context of spatial regionalization.

Retinotectal Map Formation

Perhaps the most popular area for theoretical modeling at this stage of development is the formation of topographic maps. There are numerous examples in the brain, such as the retina and superior colliculus (optic tectum), where the connections between two structures form a continuous map, so that neighboring neurons in one structure connect to neighboring neurons in the target structure. Although, as will be discussed later, neural activity plays an important role in fine-tuning the structure of such maps, their initial formation is probably governed by gradient-directed axon-guidance mechanisms. The basic idea, first proposed by Sperry several decades ago, is that gradients of some molecules in the input structure give each neuron in that structure a unique regional identity. Analogous gradients are postulated to exist in the target structure, and the map forms because each

axon follows the gradients in the target structure to a topographically appropriate location. During the past few years, direct evidence for this type of process has been discovered in the form of gradients of Eph receptors in structures such as the retina, and their ligands (the ephrins) in target structures such as the optic tectum. Starting in the 1970s, theoreticians proposed models that made the qualitative ideas of Sperry and others more precise. An early model suggested that, for retinal and tectal gradients of 'labels' (receptors and ligands) that run in the same direction, all retinal axons seek to connect to tectal regions with the highest level of tectal label. Competition for tectal space ensures that only those retinal axons with the highest available level of retinal label can connect to any particular region. A map forms because the 'best' retinal axons occupy the 'best' regions of tectum, and so on down the line. In more abstract multiple constraint models, systems matching arises as the optimal balance between the opposing forces of competition for tectal space and matching by stable gradients. Other models address retinal axon navigation over a continuous substrate in response to gradients running in opposite directions. In the future it will be important for such models to take into account the rapidly expanding experimental literature on Eph/ephrin gradients and their interactions.

ACTIVITY-DEPENDENT REFINEMENT OF CONNECTIVITY

In the fourth stage of neural development, the initially crude patterns of connections that were established by molecularly based axon-guidance mechanisms are refined by neural activity. Patterns of firing among groups of neurons, generated either internally or by sensory transduction of signals from the environment, help to mold the architecture of the nervous system for optimum matching with the structure of the external world. This process of continued refinement continues throughout life, and it is generally believed that the mechanisms of synaptic plasticity that operate during development are similar to those that govern adult learning. Recently it has also become clear that as well as refinement there is also growth (i.e. the generation of new neurons and new connections), and that this continues throughout life.

Hebbian Learning

The most basic principle of activity-dependent synaptic plasticity is the rule proposed by Hebb in 1949, which can be paraphrased as 'cells that fire together wire together'. That is, the connection between a presynaptic neuron and a postsynaptic neuron is strengthened when their activity is correlated. This is directly illustrated experimentally by the phenomenon of long-term potentiation (LTP). This was first discovered in the hippocampus, but has now also been observed elsewhere (e.g. in the visual cortex). Here synaptic strengths can be increased when an input pathway is strongly stimulated, causing a correlated depolarization of the postsynaptic cell, or when weaker stimulation of the input pathway is paired with direct depolarization of the postsynaptic neuron. A related phenomenon is long-term depression (LTD), where under certain circumstances synaptic strengths can decrease due to a lack of correlation between pre- and postsynaptic activity. Further experimental evidence for Hebbian learning mechanisms comes from the development of the neuromuscular junction and the visual system, which will be discussed below.

Hebb's qualitative statement of his rule can be made mathematically precise in a number of different ways. Each of these versions leads to different outcomes when the same series of input patterns is presented. Since postsynaptic activity is crucial, one important source of variability in the mathematical statements of Hebb is how this activity is calculated as a function of the activity of presynaptic neurons and the strengths of connections ('weights'). One common assumption is that the output is just a linear weighted sum of activities multiplied by weights. The change in the weight vector at each time can often then be expressed as the covariance matrix of the presynaptic input patterns multiplied by the current weight vector. The final weight pattern that then emerges is the principal eigenvector of the covariance matrix. This is computationally useful, since this weight vector is the principal component of the input patterns. This is the direction of maximum variance in the input data, which in one sense is the most 'interesting' projection of the input data.

An alternative common assumption with regard to how the output of postsynpatic neurons is calculated is that some form of nonlinear competition exists between these neurons. This is considered (either explicitly or implicitly) to arise via lateral connections between postsynaptic neurons. In the simplest case this is implemented computationally by a 'winner-takes-all' mechanism, whereby the postsynaptic neuron in the group that has the largest initial activity in response to an input pattern is the only one that has its weights updated. What

tends to emerge from this type of rule is weight vectors that point to the centers of clusters of input points. This is computationally useful since such cluster analysis can help to divide the input patterns into useful categories based on the degree of overlap of different patterns.

In both of these linear and nonlinear cases an important component is often competition in the form of normalization. Since Hebb's rule specifies only increases in synaptic strengths, without some such mechanism for decreasing weights they would all increase without limit. Therefore the total weight impinging on a postsynaptic neuron is often constrained so that when some weights are increased, others are decreased.

The Neuromuscular Junction

One particular process in neural development that has been modeled using Hebbian learning is synaptic elimination at the neuromuscular junction. When growth cones first encounter their targets, molecular signals are passed back and forth which induce the development of the presynaptic and postsynaptic specializations that form a synapse. This has been most extensively studied in the development of synapses of motor neurons on muscle fibers. Initially, several motor neurons contact each muscle fiber (polyneuronal innervation). However, by a process of activity-dependent competition all but one of these connections is eliminated. Several Hebbian-based models of this process have been proposed, and again an important component is normalization. In fixed-resource models, terminals (e.g. measured by synaptic strength or arbor size) compete for a fixed amount of a resource (e.g. total synaptic strength or space). In fixed-rate models, terminals compete for a resource that is generated at a fixed rate (e.g. neurotrophins). In these models a critical variable is the function which describes the dependence of receptor production rate on the amount of neurotrophin that is bound. Such models make testable predictions with regard to experimental perturbations such as a reduction in the supply of neurotrophin or a change in the level of activity.

Visual System Development

Processes of activity-dependent development have been most extensively modeled in the visual system, due to the large amount of experimental data that is available to constrain and inspire theories. As mentioned previously, the map from the retina to the optic tectum is topographic. If neural activity in the retina is blocked, a topographic map forms that is cruder than normal, indicating that activity is important for map refinement. This refinement can be modeled using Hebb-type rules, with two important assumptions. The first assumption is that 'neighborliness' between neurons in the retina is encoded by spatially correlated activity, so that nearby neurons are more highly correlated than more distant neurons. This is true of natural images – the correlation between pixels tends to drop off smoothly with distance. Similar correlations can also be generated earlier in development, before eye opening, by spontaneous activity of retinal neurons in the form of waves or blobs of activity that sweep periodically across the retina. The second assumption is that neighborliness between neurons in the tectum is encoded by lateral connections, so that nearby neurons have stronger connections and are thus more highly correlated than more distant neurons. These connections are often assumed to be in the form of short-range excitation and longer-range inhibition. Although there is some evidence that such connections exist in reality, there is also evidence for longer-range excitatory connections which are patchy rather than spatially uniform.

Another important area of theoretical modeling of activity-dependent processes in the visual system is the development of ocular dominance and orientation columns. In the adult visual cortex of mammals such as cats and monkeys, a beautiful mosaic pattern of varying response properties is seen. Each neuron responds better to an input in one eye than to one in the other eye, and to an edge or bar of light at a specific orientation angle within its receptive field. The response properties of nearby neurons mostly vary smoothly as one moves across the surface of the cortex, so that a patch of neurons that prefers one orientation blends into a patch that prefers a similar but slightly different orientation, and there is a smooth variation in preference for one eye compared with the other eye (Figure 1). However, these patterns of stimulus preference are not present in very young animals, and they may emerge by a process of activity-dependent Hebbian learning during development. The evidence for this is particularly strong in the case of ocular dominance columns, since their overall structure can be altered by changing the correlations in the visual stimuli that are seen during the critical period for ocular dominance development. For instance, if one eye is sutured shut early in development, the size of the regions representing that eye in V1 shrink relative to the regions representing the other eye.

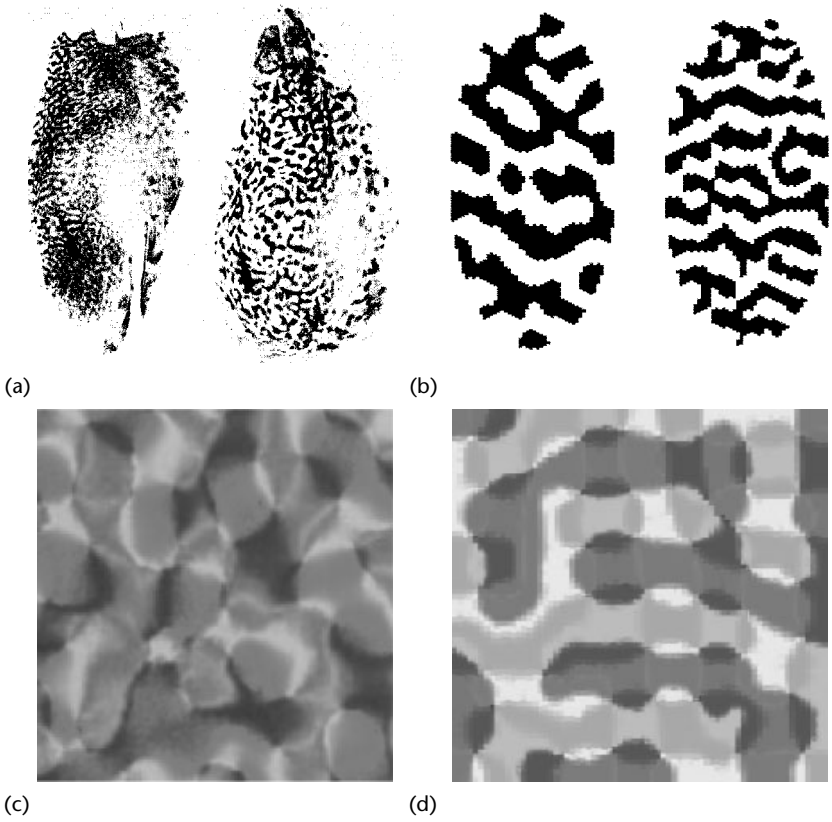


Figure 1. [*Figure is also reproduced in color section*.] (a) The pattern of ocular dominance columns in cat primary visual cortex. Dark patches are those regions of cortex dominated by one eye, and light patches are regions dominated by the other eye. Left: normal cat; right: strabismic cat. Note that columns are wider in the strabismic cat than in the normal cat. (From Löwel (1994) *Journal of Neuroscience* **14**: 7451–7468. Copyright Society for Neuroscience.) (b) Results of the author's analogous simulations using the elastic net algorithm. Left: 'normal' case; right: 'strabismic' case. (c) The orientation map in primary visual cortex of the tree shrew. The different colors represent patches that have different orientation preferences. (From Bosking *et al*. (1997) *Journal of Neuroscience* **17**: 2112–2127. Copyright Society for Neuroscience.) (d) Results of the author's analogous simulation using the elastic net algorithm.

The development of ocular dominance and orientation columns has been modeled by a variety of different versions of Hebbian learning. Three of the most important classes of models are high-dimensional, low-dimensional and filtering models. In high-dimensional models there is an explicit representation of each neuron (or each small group of neurons) in both the cortex and input layers (retina and/or lateral geniculate nucleus), and of all of the connections between layers. Patterns of stimulation are presented to the retina representing images either from the environment or generated by spontaneous activity. Activity is propagated to the cortex, and the lateral spread of activity through intracortical connections is taken into account. Weights are then updated by a Hebbian rule. Under appropriate conditions, orientation and ocular dominance columns can emerge in the cortical layer. Some high-dimensional models include hypotheses about the role of neurotrophins in mediating competition between cortical neurons. Low-dimensional models are more abstract, and instead of presenting explicit patterns of pixel intensity they consider that input images can be characterized by a small number of features, such as position in the visual field, orientation and ocular dominance. Each of these features is represented by an orthogonal dimension, and the weights of a cortical neuron now represent the selectivity of that neuron in the low-dimensional space of features. The learning rules can be seen as trading off 'matching' and 'stretching', so that all feature combinations in the input space are represented, while at the same time maximizing the degree to which neighboring cells in the cortex represent similar features. Although such models are more difficult to interpret biologically than high-dimensional models, they can often be more

robust, and they actually tend to produce a better match with the fine structure of real orientation and ocular dominance maps (Figure 1). More abstract still are the filtering models, which essentially repeatedly convolve the cortical layer with a filter, such as a difference of gaussians, as an abstract representation of the effect of intracortical processing on the cortical input.

SUMMARY

The way in which the immense complexity of the nervous system arises during development is still somewhat nuclear. Theoretical modeling has been applied to several different stages of neural development in the hope of shedding light on some of these mysteries. The best developed areas are axon growth, branching and gradient detection, and activity-dependent development in the visual system and at the neuromuscular junction. In each case, specific quantitative predictions can be derived. The comparison of these predictions with experimental data allows the assumptions underlying the models to be developed and refined.

Further Reading

Brown MC, Keynes RJ and Lumsden A (2001) *The Developing Brain*. Oxford: Oxford University Press.

Churchland PS and Sejnowski TJ (1992) *The Computational Brain*. Cambridge, MA: MIT Press.

Goodhill GJ and Richards LJ (1999) Retinotectal maps: molecules, models, and misplaced data. *Trends in Neurosciences* **22**: 529–534.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* **274**: 1133–1138.

Mueller BK (1999) Growth-cone guidance: first steps towards a deeper understanding. *Annual Review of Neuroscience* **22**: 351–388.

Murray JD (1993) *Mathematical Biology*, 2nd edn. New York: Springer.

Price DJ and Willshaw DJ (2000) *Mechanisms of Cortical Development*. Oxford: Oxford University Press.

Swindale NV (1996) The development of topography in the visual cortex: a review of models. *Network* **7**: 161–247.

Wolpert L *et al*. (1998) *Principles of Development*. Oxford: Oxford University Press.

Neural Development

Introductory article

Dale R Sengelaub, Indiana University, Bloomington, Indiana, USA

CONTENTS

The structure of the nervous system is determined developmentally through the interaction of genetic and nongenetic processes. These processes are progressive, wherein cells are added, connections are established, and functions mature, as well as regressive, involving substantial death of cells and loss of connections.

INTRODUCTION

The human brain contains over 100 billion neurons, with hundreds of differing forms and multiple functions. These neurons are organized into populations and circuits that communicate through 100 trillion synapses, intricately and precisely connected in a complex array. The amount of information that must be specified in assembling the nervous system is consequently vast. While some of this information is genetic, the surprising smallness of the human genome (perhaps at most containing 40 000 genes) clearly indicates that sources of information that are not genetic (epigenetic) are heavily involved in neural development.

Indeed, the development of the nervous system has traditionally been viewed as an interaction between a variety of factors including genes and changes in their expression, and the timing and spatial distribution of cues in the cellular environment which organize a host of developmental programs. These cues can be intrinsic, involving proteins inherited with each cell division, or extrinsic, utilizing such things as diffusible molecules, cell-membrane proteins, or even the early experience of the organism. Importantly, though it may seem counterintuitive, the developmental processes that sculpt the nervous system are regressive as well as progressive in nature, and the expected addition of newly generated cells, their growth and differentiation to mature form, and the establishment of their complex connections, is accompanied by massive remodeling of connections and even the death of substantial numbers of cells.

NEUROEMBRYOLOGY

All the cells of the nervous system – and of the entire body for that matter – arise from a single cell, the fertilized egg. Because this one cell will through repeated divisions produce all of the different cell types of the body, it is said to be totipotent, or capable of giving rise to cells of all kinds. This one totipotent cell contains all of the genetic information involved in developing all of these cell types (and as you will see, this is only a small part of the information required). As development proceeds, and this cell and the cells it generates divide and divide again, decisions about which subset of these genes will be expressed by the newly generated cells are programmed through cellular interactions, initially limiting the potential fates of those cells to a narrow range of outcomes (pluripotency). This 'progressive determination' continues as cells divide and differentiate, ultimately resulting in the restriction to one outcome.

The Germ Layers

As early as one week after fertilization, three distinct layers of cells can be recognized in the developing human embryo. These three layers are known as the 'germ layers', and each will give rise to a specific set of tissues in the embryo. The outermost layer, the ectoderm, is the origin of the cells that will become the nervous system and skin. The underlying mesoderm will give rise to many different things, including the skeletal muscles, most bones, and the cardiovascular and urogenital systems. The last layer, the endoderm, will give rise to other organs of the gut, the lungs, and the liver.

Neural Tube Formation

The formation of the nervous system begins in earnest two weeks after fertilization. The

ectodermal cells of the embryo will differentiate into skin unless directed not to by a signal that comes from the underlying mesoderm. This kind of interaction between tissues occurs throughout embryonic development and is known as 'induction', a phenomenon in which signals from one tissue modify the development of another tissue. In the case of the nervous system, chemical signals from the mesoderm block differentiation of ectoderm into skin and cause neural tissue to form instead. Several candidate molecules have been identified that act to direct the differentiation of the ectoderm into nervous system tissue.

As a consequence of these chemical cues, ectodermal cells differentiate into the neural plate, where changes in the shape of cells, addition of cells, and rearrangements of cells result in the formation of a groove along the midline of the embryo. Over the next few days this groove deepens and its edges fuse, forming a long, hollow tube along the back of the embryo (Figure 1). This neural tube is the embryonic origin of the central nervous system (the brain and spinal cord). By six weeks after conception, several swellings develop at the anterior end of this tube, corresponding to the major subdivisions of the brain. This tube will continue to distort its shape and size, eventually obscuring the fact that the structures of the brain and spinal cord are elaborations along a long tube, whose interior constitutes the fluid-filled ventricles of the brain and the central canal of the spinal cord.

The origin of the peripheral nervous system – the various ganglia and their connecting nerves that lie outside the brain and spinal cord – derives from a further differentiation of the neural plate called the 'neural crest'. Neural crest cells form at the edges of the neural plate, and come to rest on the top of the newly formed neural tube. These neural crest cells then migrate away from this position, and depending on the route they take, differentiate into a variety of tissue types including the facial bones, pigmented cells of the skin, and the cells of the peripheral nervous system. Again, the instruction of what type of tissue to differentiate into is an inductive event. The inductive cues appear to come from the interaction of these migrating neural crest cells with the cells that surround them. Molecules from these surrounding cells direct neural crest migration by permitting it (through providing adhesion) or inhibiting it (through repulsive factors). Once the crest cells have been guided to their final destinations, other adhesion molecules cause the cells to stop migrating and aggregate at the new location.

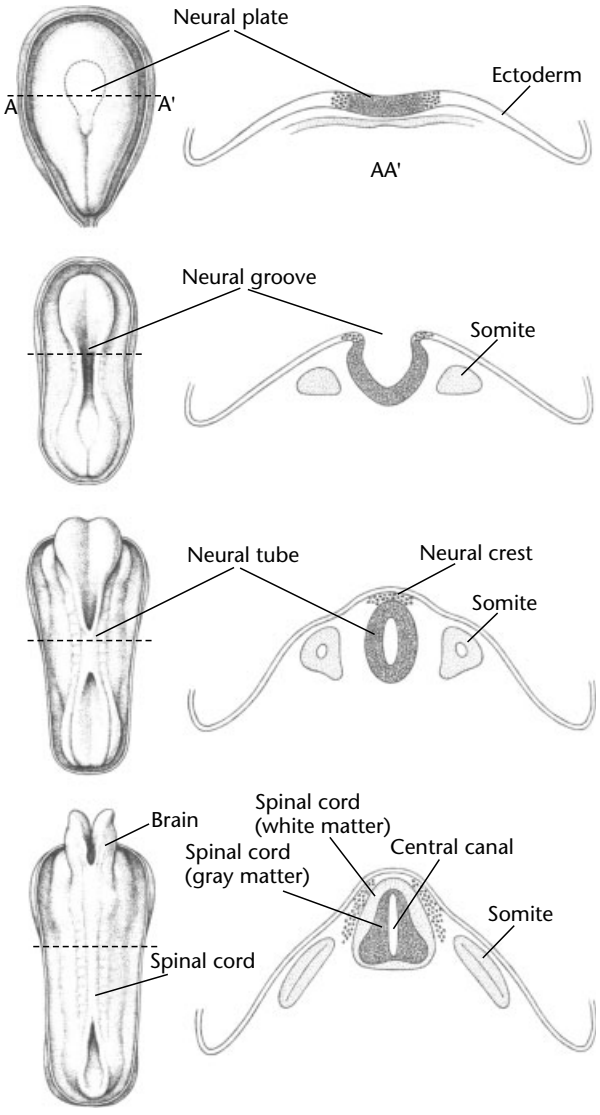


Figure 1. The nervous system has its origins during the first weeks after conception. External views of the embryo are shown on the left as well as cross-sections corresponding to the dashed lines on the right. At 18 days after conception, the neural plate develops from the ectoderm. Over the next few days the plate invaginates into a groove, and eventually closes itself into the neural tube, the embryonic origin of the brain and spinal cord. Structures of the peripheral nervous system arise from a population of cells called the neural crest, which migrate down the sides of the neural tube and aggregate into the peripheral ganglia. From Cowan WM (1979) The development of the brain. *Scientific American* **241**: 112–133.

DEVELOPMENTAL PROCESSES

Cell Proliferation

While it has recently been shown that the generation of neurons continues throughout life, the

majority of neurons are generated prenatally or early postnatally. Neural proliferation occurs primarily in ventricular zones that develop in the inner lining of the neural tube. Here, as cells divide, they undergo a series of well-defined stages known as the 'cell cycle'. During the cell cycle, germinal cells in the ventricular zone shuttle back and forth as they go through mitotic divisions. Early in this process, divisions are symmetrical, with each resulting in a doubling of the germinal cells; as development proceeds, these divisions become asymmetrical, and with each mitotic cycle, one germinal cell and either a neuron or glial cell will be produced. (Glial cells serve a variety of functions in the nervous system both during development as well as throughout life, and actually outnumber neurons.) The production of neurons (neurogenesis) during development is impressive, generating over 250 000 neurons per minute. Neurogenesis proceeds in a variety of spatial and temporal patterns, as well as a progressive generation of large neurons first followed by intermediate and then small neurons.

Migration

At some point, cells in the ventricular zones stop dividing and withdraw from the cell cycle. These cells migrate away from their site of generation and begin to differentiate into their final form, soon expressing proteins unique to their neural or glial type. The routes these migrating cells take are often tortuous, and can include secondary migrations as well. Neurons appear to use a variety of cues and mechanisms to guide their migration. Some cells actively move, while others are passive and are simply pulled and pushed to their destination by their active neighbors.

In the peripheral nervous system, sheets of neural crest cells are pulled through the surrounding cells by active pioneer cells. The mechanism that guides these migrations appears to be simply permissive, relying on adhesion or repulsion by surrounding cells. In the brainstem, displacement of cells from the ventricular zone is followed by mass migrations regulated by unknown mechanisms. The migrating cells then collect into masses called 'nuclei'. Again, expression of neural adhesion molecules is thought to be involved in establishing these cellular masses.

Perhaps the best-known mechanism for migration is provided by radial glia, specialized glial cells whose long processes appear to provide a guide along which migration occurs. These radial glia differentiate quite early and extend fibers from the ventricular zones out to the expanding edges of the growing nervous system (Figure 2). As neurons are generated and exit the ventricular zone, they attach themselves to the radial glia, and literally pull themselves along the glial fiber. The radial glia provide not only a means by which migrating neurons can travel, but also provide a direction for this primary migration.

The migration of neurons can contribute to the development of highly organized structures. Consider, for example, the formation of the neocortex. The neocortex is massive, covering most of the surface of the forebrain, and is organized into six layers, with large neurons projecting to subcortical areas residing in the deep layers, neurons receiving major inputs from the thalamus in middle layers,

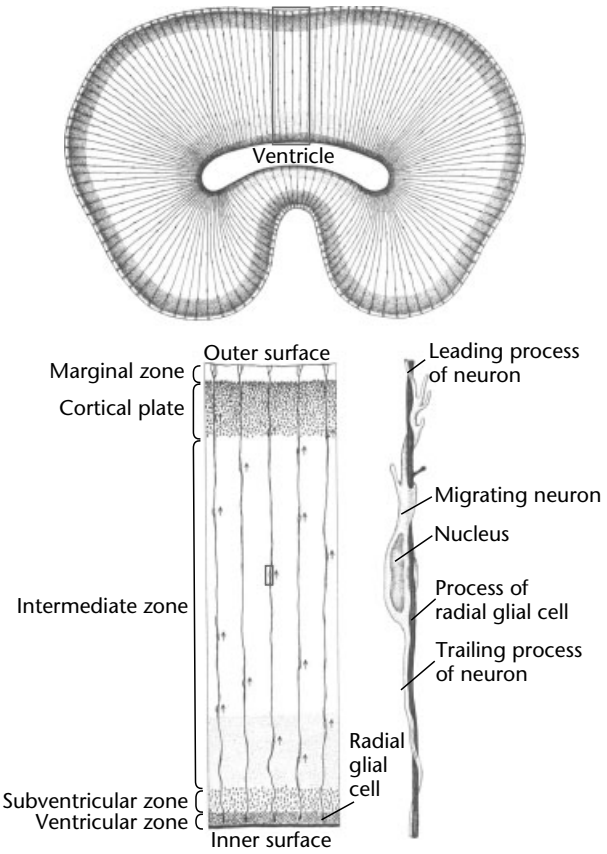


Figure 2. Processes from specialized cells called radial glia span the width of the developing brain and spinal cord, providing a path for migrating neurons. The illustration at the top shows the arrangement of radial glia in a transverse section through the developing cerebral hemisphere of a monkey. The successive enlargements below show migrating neurons attached to the radial glia, actively pulling themselves along the fibers. From Cowan WM (1979) The development of the brain. *Scientific American* **241**: 112–133.

and smaller neurons involved in sharing information with cortical areas on the opposite side of the brain occupying the superficial layers. The layers of the neocortex arise sequentially through successive waves of migration. The deepest layers of the neocortex form first, and as new cells are generated and migrate into the developing neocortex they migrate past the previously deposited neurons, establishing more superficial layers. Thus, the layers of the neocortex contain neurons that differ in age, with the oldest (earliest generated) neurons lying in the deep layers, and progressively younger (generated later) neurons in the outer, more superficial layers. Migration in the neocortex involves more than the simple radial migration along glia into successive layers, as tangential migrations across the cortex also occur.

Complex migrations of neurons using mixed strategies are seen elsewhere in the developing nervous system. For example, the cortex of the cerebellum also contains layers, with small granule cells lying below a layer of large Purkinje cells, which in turn lie below a surface layer of fibers. In the development of the cerebellum, two migrations occur. In the first, the developing Purkinje cells migrate from the ventricular zone to the cerebellar plate. Presumptive granule cells then stream in and spread over the Purkinje cell layer. These granule cells then extend axons horizontally across the cerebellar cortex, forming the superficial fiber layer. Lastly, the cell bodies of the granule cells then migrate down through the Purkinje cell layer, guided by specialized glial cells (the Bergmann glia).

Other patterns of migration exist as well. As in the developing brain, spinal cord neurons exit the ventricular zone surrounding the central canal and migrate radially along glial cell fibers. Curiously, these migrations can be bidirectional, with neurons migrating away from the ventricular zone out to the edges of the developing spinal cord, and then reversing course. Tangential migrations along bundles of axons within the spinal cord further complicate this process. Finally, not all of the neurons in the central nervous system are generated in ventricular zones. Certain neurons found in the hypothalamus are actually generated outside the neural tube in a structure known as the nasal placode. Cells from this placode give rise to structures of the nose, as well as producing cells that migrate into the developing hypothalamus.

Cell Differentiation

Migrating neurons have already begun to express proteins that identify them as neurons, but their differentiation into their adult form is far from complete. Migrating neurons might have leading and trailing processes, but their size and certainly the elaboration of their dendrites and axons that uniquely characterize them as particular types of neurons, as well as most of their functional properties including neurochemistry and receptor expression, have yet to be realized. The assignment of cellular identity, and the processes through which neurons develop their individual forms and functions, are controlled by factors that are both intrinsic as well as extrinsic to the neuron. The bulk of the evidence suggests that even intrinsic factors, such as differential gene expression, are in fact themselves regulated by extrinsic control.

As an example, neurons communicate primarily through chemical synapses, and different types of neurons use different types of neurotransmitters. For at least some neurons, the decision of which neurotransmitter to express appears to come from an interaction with the neuron's extracellular environment or target tissues, another example of an inductive event. In the peripheral nervous system, neurons aggregate into groups called 'ganglia', and these ganglia differ with respect to which neurotransmitter they will use in communicating with their target tissues. Neurons in ganglia that serve the sympathetic branches of the autonomic nervous system typically use the neurotransmitter noradrenaline (norepinephrine), while neurons in ganglia serving the parasympathetic branches use the neurotransmitter acetylcholine. The assignment of which neurotransmitter to express appears to depend at least in part on the influence of the local cellular environment. Different ganglia have different migratory routes, and cues in these routes influence the different neurotransmitter assignments for these neurons. In addition, the target tissues for these neurons also appear to have a role, and the neurotransmitter expressed can be switched by altering the nature of the target tissues they contact.

The type of neuron that differentiates also seems to be controlled by induction. For example, motor neurons are commonly organized into nuclei that are found in the ventral portion of the hindbrain and spinal cord. Neurons in these ventral areas develop into motor neurons early, and this differentiation is regulated through a protein cue that originates from cells in the neighboring notochord. The notochord is part of the mesoderm, and lies just below the growing spinal cord. Cells in the notochord express a protein called 'sonic hedgehog', which diffuses away from the notochord, causing the ventral spinal cord neurons to differentiate into motor neurons.

Dendritic and Axonal Growth

As neurons differentiate, they establish characteristic morphologies which correspond to their function and physiological properties. Neuronal morphologies vary widely, and these anatomical differences, including the branching patterns and distribution of processes, and overall shape of dendritic arbors, reflect important functional differences. To date, a variety of factors have been identified which may be involved in regulating the growth, distribution, and orientation of dendritic arbors. These factors include interactions with the axons that contact them (their afferents), their own targets, and a variety of trophic substances. Dendritic development begins after migration of the cell body is complete. The characteristic number and orientation of dendrites that are initially extended appears to depend on information contained in the neuron, perhaps specified early in differentiation, as neurons grown artificially in cell cultures nonetheless develop appropriate dendritic morphologies. However, further elaboration of the dendritic arbor, including the pattern of branching and the length of dendritic segments, is defined by agents and actions occurring outside of the cell. Electrical and chemical activity in a neuron's afferent inputs, factors from its targets, and other variables all interact to shape the developing dendritic array.

While the development of dendrites typically occurs after migration of the cell body, the growth of a neuron's axon often begins before migration is complete, and in some cases the axon has already grown into its target before the cell body reaches its final migratory destination. The growth of the axon occurs at the tip, in a structure called a growth cone, a small, spiky blob that sends out filaments in many directions which can adhere to particular surfaces, dragging the axon along behind it. The routes that growing axons take are often long and complicated, and as a consequence there appear to be multiple mechanisms involved in guiding the growing axon. For example, some growing axons rely on guidepost cells spaced along the route. These guidepost cells direct axon segments to the next set of cells containing guidance information for the subsequent leg of the journey. In many cases the guidance of the growing axons is quite precise, and chemical cues such as molecules from the surrounding cells, adhesion or cell recognition molecules on the surfaces of cells, growth factors, and signaling molecules organized in gradients, have all been identified as contributing to the establishment of correct axonal projections and complex mappings. Simple mechanisms also exist; for example, growing axons will travel along previously grown axons, or use the presence of blood vessels to guide them. Finally, in some systems axons will grow diffusely, sending branches into several potential target areas. Because many of these early diffuse branches are in fact in error, their later elimination (see below) is an essential process in the development of precise axonal projections.

With the arrival of the growing axon at its target, the formation of synaptic contacts (synaptogenesis) begins. Synaptogenesis is essential for the continued differentiation, growth, and even survival of developing neurons. As in so many earlier cases, this process involves interactions between cells. For example, axons of motor neurons grow into the muscles they will control and begin to form specialized synapses called neuromuscular junctions. The motor axons appear to signal the formation of the synapse by secreting chemicals that cause the muscle to make and cluster neurotransmitter receptors at the site of the developing synapse. The muscle in turn secretes chemicals that cause the motor axon to branch. The process of synaptogenesis can extend well into postnatal development. For example, in the human neocortex, synapse numbers increase dramatically through the first year of postnatal life.

Normally Occurring Neuron Death

The formation of synaptic contacts between neurons and their targets ushers in the next major process in neural development, the death of enormous numbers of neurons. The death of neurons during normal development is a surprising but well-documented phenomenon that occurs throughout the nervous system. Initially neurons are overproduced and their numbers are subsequently reduced during a period of normally occurring cell death. This death appears to be controlled by target availability, synaptic activity, and to some extent afferent input as well.

Normally occurring neuron death, or apoptosis, is sometimes called 'programmed' cell death because it is a predictable process that involves the switching on or off of a set of genes coding for enzymes that are used in the destruction of the cell. Which neurons survive and which die during development in the human nervous system is not genetically determined, but instead results from an apparently competitive process in which otherwise healthy, normal neurons are eliminated. This death of developing cells is not at all unique to neurons,

and in fact occurs in a variety of developing tissues, allowing the sculpting of specialized structures. For example, the development of the fingers and toes of the hands and feet includes the death of cells that lie between the forming digits. In the case of the developing nervous system, the death rate can range up to 80 percent of the initial neuron population in a given structure, with the deaths occuring over a brief period. Normally occurring neuron death contributes to the formation of a variety of structural specializations throughout the nervous system, creating local differences in neuron numbers that underlie functional specializations ranging from high-acuity vision to sex differences (Figure 3). In addition, neuron death appears to set the numbers of cells appropriately in interconnecting populations, as well as eliminating neurons that have established erroneous connections through diffuse growth of axons.

The onset of normally occurring neuron death coincides with the establishment of connections between neurons and their targets. Moreover, the activity in those connections plays an important part in regulating the amount and timing of neuron death. For example, altering the size of the target has a profound effect on the number of developing neurons that will survive the cell death period. Removal of target tissue, for example the target muscles of developing motor neurons, results in a proportionately larger death of motor neurons innervating those muscles. Conversely, adding additional target tissue, for example by grafting on an extra limb, results in a sparing of motor neurons that would otherwise have died. However, the simple presence of neuronal targets is not sufficient, as the process of neuronal death appears to also require communication between neurons and their targets. For example, blockade of the neuromuscular junction with drugs that prevent synaptic transmission (e.g. curare) reduces the amount of motor neuron death. The amount of input a neuron receives may also influence neuronal survival, as large removals of projecting neurons increases the amount of death seen in the target population.

The regulation of the amount of neuron death by targets and afferents suggests that developing neurons depend on factors that are in short supply: gaining access to sufficient amounts or types allows a neuron to survive, while failing to do so results in cell death. Several competition models have been proposed in which developing neurons must make adequate synaptic connections or obtain special chemicals that promote survival. Such chemicals are known as 'trophic factors', and numerous candidates have been identified. These trophic factors are thought to regulate neuron survival by either preventing the expression of genes that cause the death of the cell or triggering the expression of protective molecules that block products of lethal gene products. For example, neurons in the

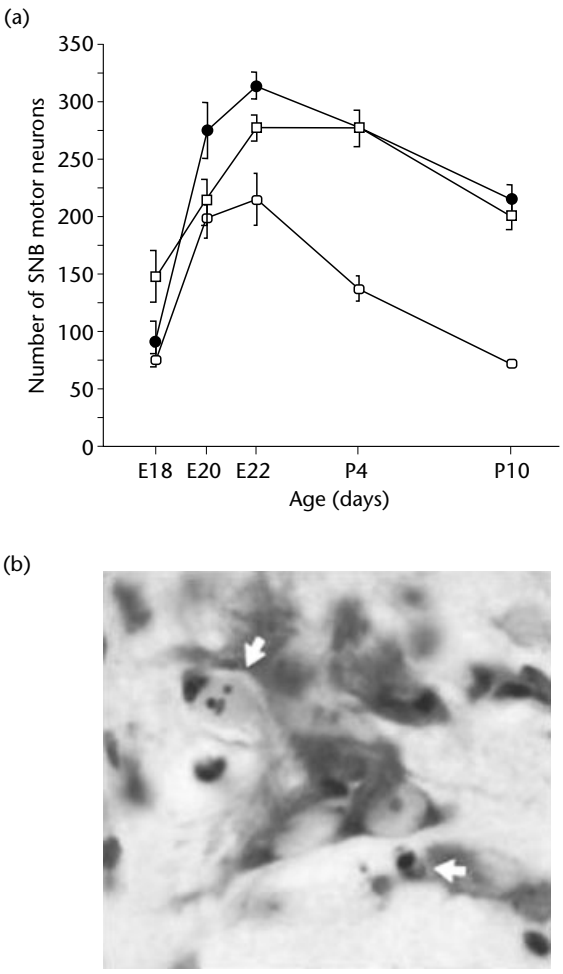


Figure 3. Normally occurring neuron death eliminates substantial portions of the developing nervous system. This death can be regulated by a variety of factors, including sex hormones, producing sex differences in neuron number. (a) Hormonally dependent death of motor neurons in the spinal nucleus of the bulbocavernosus (SNB) in the rat spinal cord over the late embryonic (E) and early postnatal (P) days. Female rats (white circles) lose many more SNB motor neurons than do male rats (black circles), and treatment of females with testosterone propionate (white squares) decreases that loss to male levels. (b) The photomicrograph shows degenerating cells (arrowed) in the SNB. Adapted from Nordeen EJ, Nordeen KW, Sengelaub DR and Arnold AP (1985) Androgens prevent normally occurring cell death in a sexually dimorphic spinal nucleus. *Science* **229**: 671–673.

sympathetic ganglia require nerve growth factor supplied by their target tissues, motor neurons require proteins that are produced in their target muscles, and even glia in the peripheral nervous system need trophic factors from axons to survive. Given that the survival of developing neurons also depends on activity, it is possible that activity might regulate access to trophic factors by influencing axonal branching or synapse formation.

Synapse Elimination

Just as neurons in the developing nervous system are overproduced and then reduced to mature amounts and distributions through the regressive process of neuron death, the synaptic connections made during development are also subject to an initial overproduction and then subsequent retraction. In some cases, it is the pattern of synaptic connections that is remodeled (and the actual number of synapses may actually increase), while in others, the overall amount of synapses declines dramatically. In the human nervous system, this process of synapse elimination typically occurs after the period of neuron death, and can take several years to reach completion (Figure 4).

A simple case of synaptic reorganization occurs in neuromuscular systems. At maturity, a typical muscle fiber is contacted by a single motor neuron, but developmentally, several motor neurons synapse onto a single muscle fiber. This polyneuronal innervation in neuromuscular systems is reduced through a retraction of axon branches from some neuromuscular junctions, eliminating the multiple neuron inputs to single fibers. Simultaneously, the remaining axons grow and branch, making numerous new contacts resulting in a net increase in synaptic numbers. This synaptic remodeling is driven by neuronal activity in the system, wherein synaptic activation maintains the active synapses and removes inactive ones. As in the case of activity-dependent neuron death, preventing activity at the neuromuscular junction arrests synapse elimination.

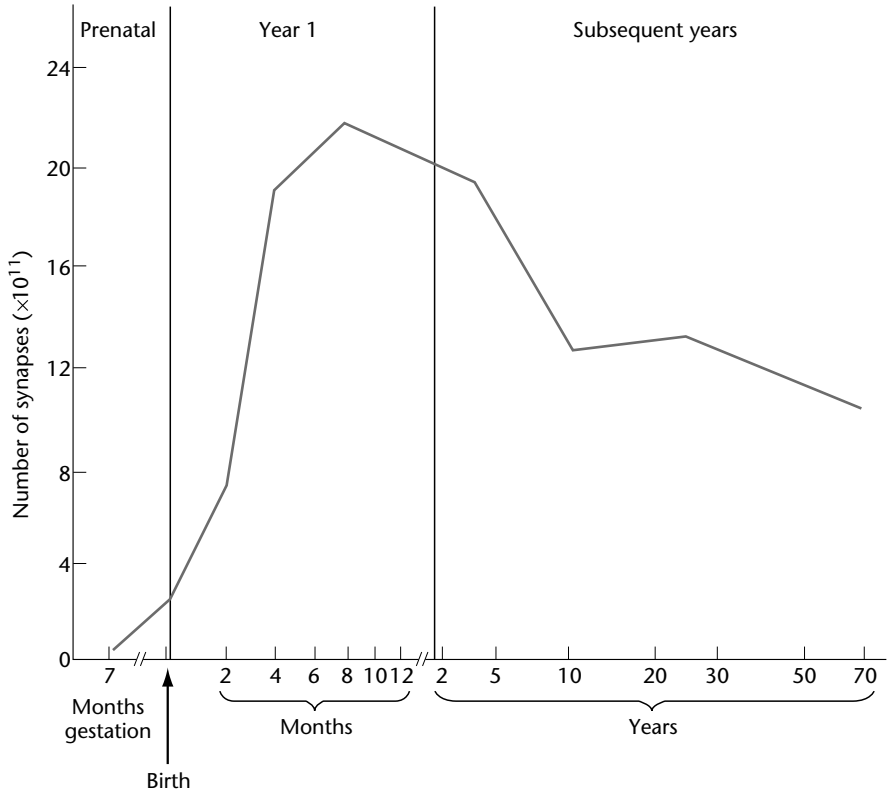


Figure 4. Synapse elimination in the human cortex occurs over a period of years. In the visual cortex, the number of synapses increases rapidly after birth, peaking at about one year of age. The number of synapses then declines dramatically over the subsequent 10 years. The number of synapses is stable through young adulthood, and then slowly declines thereafter. From Kolb B and Wishaw IQ (2001) *An Introduction to Brain and Behavior*. Worth Publishing.

Synapse elimination is involved in the establishment of more complex patterns of connectivity as well. In the mature visual system, projections from the retina to central visual targets such as the lateral geniculate nucleus of the thalamus and the superior colliculus of the midbrain show highly organized topographic maps such that a neuron's location on the retina (and hence the part of the visual world it responds to) is faithfully preserved in the representation across these structures. The initial retinal projections to these structures are far less organized and must be refined through activity-driven synapse elimination.

The classic example of synapse elimination in development comes from work examining the representation of visual input in the cortex. At maturity, neurons in the visual cortex receive information via the thalamus from both the right and left eyes. These neurons are organized into discrete, alternating columns which are more easily activated by information from one eye than from the other – hence they are called 'ocular dominance columns'. This organization is not as readily apparent at birth, and the distinction between columns is blurred because the thalamic axons carrying information from the two eyes are initially intermixed. The activity supplied by normal vision during postnatal development drives the process of synapse elimination in the visual cortex, and the overlapping axons segregate, retracting from inappropriate areas and elaborating in appropriate ones, refining the boundaries between ocular dominance columns.

This same phenomenon also occurs in the lateral geniculate nucleus, the thalamic nucleus that supplies the axons carrying the visual information to the cortex. In the lateral geniculate nucleus, information from both eyes is initially intermixed early in development. Through synapse elimination this mixing is removed, and at maturity, information from the left and right eyes are held separately in distinct layers.

Thus, in both the thalamus and the visual cortex, an activity-dependent sorting of synaptic connections refines the structure and function of visual neurons.

Early Experience

The development of normal vision depends on the presence of normal visual experience during the period in which synapse elimination, axonal segregation, and synaptic remodeling are occurring. If normal vision is disturbed during this period, permanent perceptual deficits will result.

This illustrates the concept of the critical period, a specific time in development when a neuron, system or organism is particularly sensitive to – and permanently altered by – environmental or other outside influences. Critical periods exist for all of the developmental processes, and the timing of these periods differs across neural systems, corresponding to their particular developmental programs. In the case of the developing visual system, depriving the eyes of normal vision during this postnatal critical period either through cataracts, vision problems such as uncorrected amblyopia or astigmatism, monocular eye patches, or rearing in complete darkness will result in permanent perceptual deficits. All of these conditions will alter the pattern of activity produced by stimulation of the two eyes, and as a consequence will affect the interaction between synapses that drives the normal segregation and elaboration or appropriate connections in the developing visual cortex. Importantly, if these same manipulations are applied after the critical period (that is, after synapse elimination has been completed), they have no effect.

The fact that the early experience of an organism can have profound effects on development has been amply demonstrated. Altering the early experience of a developing organism will alter both neural structure and function, and indeed the organism's behavior. Early experience is an essential element of development, providing a variety of stimuli and a wealth of information. Importantly, this early experience is not limited to laboratory experiment manipulations; for example, children who do not play or are rarely touched have significantly smaller brains than those of children who experience normal everyday childhood interactions.

The influence of the early environment on neural development has received a great deal of attention. Animal studies have repeatedly demonstrated that depriving developing organisms of normal social, cognitive, or environmental experiences has a profound detrimental effect on both the brain and behavior. Importantly, despite popular claims to the contrary, these animal studies do not support the hypothesis that enhancing the environment or experience above what is normally available has any positive effect whatsoever. Rather, these animal studies demonstrate that deficiencies in the early environment can be partly compensated for by adding stimulation, complexity, and novelty. In other words, the 'enriched' environments used in these studies are only poor approximations of the animal's natural normal environment.

CONCLUSION

The development of the human nervous system is a complex and protracted phenomenon, involving many processes spanning prenatal as well as postnatal development.

At birth, while the generation of neurons is by no means finished, the number and locations of neurons have essentially been established, and their differentiation to adult form is complete or well under way. However, the development of important structures still has a long way to go. Establishment of the basic structure of major brain areas such as the neocortex, hippocampus, and cerebellum will continue for many months postnatally. Major axon bundles are still forming at birth, including the corpus callosum which will bridge the cerebral hemispheres, and the corticospinal tract with its motor projections to the spinal cord. The development of the myelin sheaths that speed the conduction of action potentials along axons will proceed throughout adolescence, and the major changes in connectivity allowed by synaptic remodeling through synapse elimination will take years to complete.

This development involves both progressive and regressive processes that are influenced by multiple interacting factors. The dynamic assembly and destruction of the nervous system during normal development ultimately requires both genetic as well as epigenetic sources of information, including the behavioral experience of the organism.

Further Reading

Brown MC, Hopkins WG and Keynes RJ (1991) *Essentials of Neural Development*. Cambridge, UK: Cambridge University Press.

Cowan WM (1979) The development of the brain. *Scientific American* **241** (September): 112–133.

Jacobsen M (1991) *Developmental Neurobiology*. New York, NY: Plenum.

Purves D and Lichtman JW (1985) *Principles of Neural Development*. Sunderland, MA: Sinauer.

Sanes DH, Reh TA and Harris WA (2000) *Development of the Nervous System*. San Diego, CA: Academic Press.

Neural Inhibition

Introductory article

Kevan AC Martin, Institute of Neuroinformatics, Zurich, Switzerland

CONTENTS

Neural inhibition is an active process that reduces or suppresses the excitatory activity of synapses, neurons or circuits.

INTRODUCTION

Neural inhibition achieves its effect by reducing the excitability of neurons. Inhibition is not simply 'all or nothing' in its action, but provides a graded, analogue control of excitation. The most important actions of inhibition are effected through a large array of specialist inhibitory neurons integrated within brain circuits. Thus inhibition always acts in tandem with excitation. Although inhibition is usually regarded primarily as a synaptic event, any process that acts to reduce the excitability of synapses, neurons, circuits or systems is inhibitory. Paradoxically, without inhibition, no perception, decision or movement is possible. This is because inhibition not only acts to check excitation, but may also release excitation by inhibiting other inhibitory neurons that keep the excitatory neurons in check. The push–pull interplay between excitation and inhibition generates the rhythms in the brainstem and spinal cord that allow us to breathe, walk and perform most other basic functions. Inhibition thus expresses itself in different forms at all levels of organization of the nervous system.

MECHANISM OF INHIBITION

The basic inhibitory mechanisms apply to all neurons. Inhibition is based on channels in the membrane, which are selectively permeable to specific ions (e.g. sodium, potassium, chloride, calcium) (Figure 1). The channel permeability can be altered by neurotransmitters released at synapses or by changes in voltage, or indirectly by the concentration of ions such as calcium. The 'resting' membrane potential (about −70 mV between the inside and outside of the neuron) is determined by the ratio of the concentrations of the different ions inside and outside the neuron and the relative permeabilities of the ion-selective channels.

When synaptic current flows into the neuron, the membrane depolarizes and this activates voltage-sensitive channels, which open to allow sodium ions to move down their concentration gradient. This inward current depolarizes the cell further, opening even more sodium channels in a positive feedback loop. If the depolarization is sufficiently large (about 20 mV), the membrane potential crosses a threshold for generating an all-or-nothing action potential. Individual sodium channels only open briefly, and then inactivate when they initiate the nerve impulse. Simultaneously, the depolarization activates voltage-sensitive potassium channels, which open to allow an opposing outward current to flow across the membrane. Together with the inactivation of the sodium channels, the outward potassium current returns the membrane to the resting potential.

Other types of membrane ion channels are located at synapses (Figure 1). In their simplest form, these ion channels have receptors that bind the neurotransmitters which are released at synapses. These are termed directly gated channels, and the associated receptors are called ionotropic receptors. When a neurotransmitter binds, the ion channel opens and ions flow down their concentration gradients. The direction of the net current (i.e. whether it flows into or out of the neuron) determines whether that particular synapse is excitatory or inhibitory. The activation of an individual excitatory synapse evokes an inward current, which depolarizes the neuron slightly (0.1–1.0 mV) from its resting potential of about −70 mV. If many excitatory synapses are activated simultaneously, this can trigger an action potential. Inhibitory synapses also have ionotropic receptors. Unlike the excitatory synapses, the binding of the inhibitory neurotransmitter results in an outward current, which

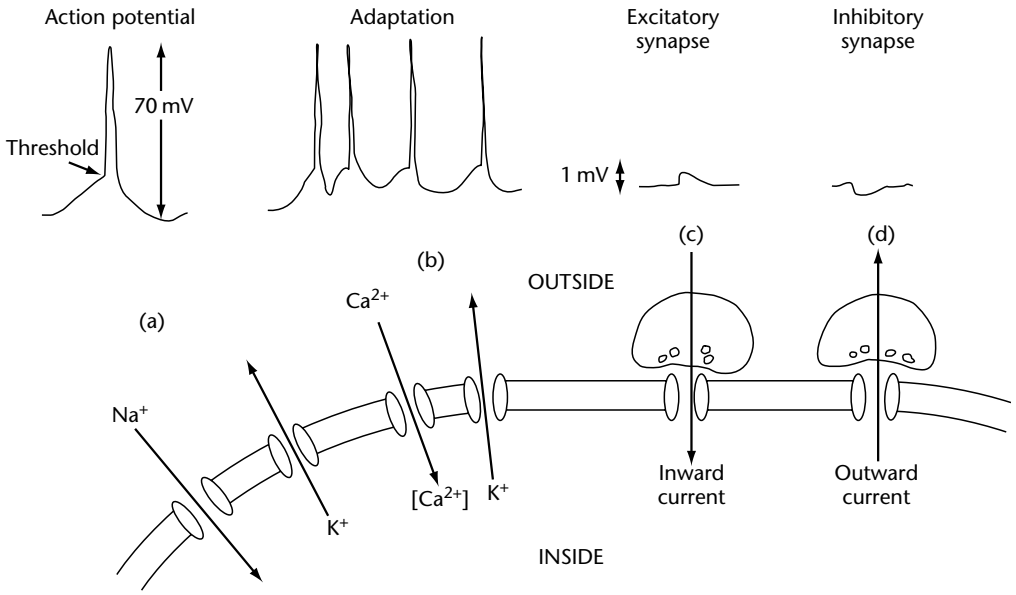


Figure 1. A segment of a neuronal membrane. Schematics indicate the direction of the currents associated with (a) the sodium and potassium channels responsible for the action potential (nerve impulse), (b) voltage-dependent calcium channels and the calcium-dependent potassium channels, which slow the rate at which the nerve impulses are produced (firing frequency adaptation), and the channels associated with (c) an excitatory synapse and (d) an inhibitory synapse. More than one ion species generally flows through the synaptic channels. The effects of activation of these channels on the membrane is shown above the schematic membrane. The magnitude of the action potential is about 70 times greater than the potentials generated by individual synapses.

tends to drive the membrane to more negative values, thus opposing the inward depolarizing current generated by excitatory synapses.

Although neurotransmitters are often described as 'excitatory' or 'inhibitory', it is not the neurotransmitter that makes the synapse inhibitory or excitatory, but rather the channel associated with the particular neurotransmitter receptor. Inhibitory synapses in the central nervous system are typically associated with potassium- or chloride-selective channels, both of which generate outward currents when activated at membrane voltages below -70 mV. In addition, the opening of the chloride channels significantly lowers the electrical resistance of the neuron, which allows the excitatory current to leak away. This is called 'shunting' inhibition.

FORMS OF INHIBITION

Synaptic Inhibition

There are two forms of inhibition that are mediated by synapses. In the more common form, inhibitory synapses on the target neuron directly reduce the neuron's excitability. This is called *postsynaptic inhibition*. In the less common form, namely *presynaptic inhibition*, the inhibitory synapse is on the terminal of the excitatory synapse itself. The presynaptic inhibitory mechanisms act by reducing the probability that excitatory neurotransmitter will be released from the presynaptic terminal. The role of both forms of inhibition is the same, namely to negate or reduce the effects of excitation. Most inhibitory synapses are formed by axons, but some neurons form their inhibitory synapses with their dendrites. Under the electron microscope the inhibitory synapses appear morphologically distinct from the excitatory synapses.

Intrinsic Inhibition

Not all inhibition is due to synaptic action. Many neurons have intrinsic mechanisms that reduce their excitability. These mechanisms usually involve potassium channels that are either directly activated by changes in the membrane voltage or indirectly activated via changes in the intracellular concentration of certain ions, especially calcium. In 'voltage-gated' potassium channels, the voltage of the membrane determines the activation or deactivation of the channels. When the membrane depolarizes, potassium channels open and the outward current opposes the depolarization.

Voltage-dependent channels also have a very important role in allowing neurons to fire repetitively at low frequencies, because they slow the rate at which the neuron can depolarize between impulses. Calcium-activated potassium channels (Figure 1) are sensitive to the free calcium concentration in the neuron, and respond to elevated calcium levels by repolarizing or hyperpolarizing the neuron. Other potassium channels are activated by internal 'second-messenger' systems, usually involving G-protein pathways, which are themselves often triggered by ions such as calcium either released from internal stores or entering via voltage-gated channels in the membrane. The effect is that more excitatory current is required to move the membrane to threshold, and the firing rate slows or even stops. This effect is known as firing frequency adaptation. These intrinsic inhibitory mechanisms are particularly significant with regard to producing pacemaker cells that have a spontaneous rhythm. Both voltage-dependent channels and channels gated by second messengers are influenced by neurotransmitters or neuromodulators.

NEUROTRANSMITTERS AND RECEPTOR TYPES

The principal neurotransmitters involved in inhibitory transmission in the central nervous system (CNS) are gamma-aminobutyrate (GABA), glycine and acetylcholine. Each of these binds to its own specific receptors, but there may be more than one receptor for the same neurotransmitter. Acetylcholine is an example of a neurotransmitter that can be either excitatory or inhibitory, depending on the synaptic receptor. When acetylcholine binds to nicotinic-type acetylcholine receptors, its action is excitatory, but when it binds to muscarinic receptors, its action is inhibitory. The muscarinic acetylcholine receptors are one of the family of 'metabotropic receptors', so called because the receptors are not directly linked to the ion channels, but are linked to them indirectly via a G-protein pathway. The binding of acetylcholine leads to a biochemical cascade that opens potassium channels, which then drive the resting membrane voltage to more negative values and so oppose the excitatory depolarization. Compared with ionotropic receptors, the onset of response is slower and the duration of response tends to be longer. Both GABA and glycine bind to ionotropic receptors on a chloride channel. Like potassium, the equilibrium potential of chloride is very negative relative to the action potential threshold. Therefore activation of the GABA and glycine receptors coupled to the chloride channels opposes the excitatory depolarization of the membrane. GABA also binds to a metabotropic receptor that is coupled indirectly to a potassium channel.

SHORT-RANGE AND LONG-RANGE EFFECTS

It is quite common to find that when a projection pathway is stimulated, its target neurons respond with monosynaptic excitation and disynaptic inhibition. This indicates that while most long-distance projections are excitatory, they connect to both excitatory and inhibitory neurons within their target area. Inhibitory neurons tend to connect only locally, and one of the most fundamental uses of inhibition in the CNS is to sharpen the spatial pattern of excitatory responses. The action of such *lateral inhibition* is to confine excitation to a core of the most active cells, and to prevent activity in the surrounding less excited cells. Lateral inhibition is a strong feature of retinal processing, where the retinal circuitry produces a spatial receptive field whose excitatory center is surrounded by a moat of inhibition. These lateral inhibitory influences are the cause of a number of visual illusions (e.g. Mach bands, the Craik–Cornsweet illusion, and the Hermann grid illusion). However, lateral inhibition is not confined to the sensory periphery – it is ubiquitous in the CNS. For example, in the spinal cord, the Renshaw cells (see below) mediate lateral inhibition of the motor neuron pool.

ROLE IN MICROCIRCUITS

A particularly well-studied microcircuit in the spinal cord coordinates muscle action around a joint (Figure 2). When a muscle is stretched, stretch receptors in the muscle excite the motor neurons which innervate that muscle and cause the muscle to contract. This is the *stretch reflex*, which acts to resist increases in muscle length. However, the same stretch receptors excite so-called '1a' inhibitory neurons that provide a *feedforward inhibition* to the motor neurons that excite the antagonistic muscles. If the 1a inhibitory neurons are themselves inhibited, co-contraction of opposing muscles can occur. Thus this circuit enables the stiffness of the joint to be set by controlling the relative amount of excitation and inhibition within the microcircuit.

The 1a inhibitory neuron itself makes only local connections in the spinal cord, but it is excited by long-distance connections from the muscle stretch

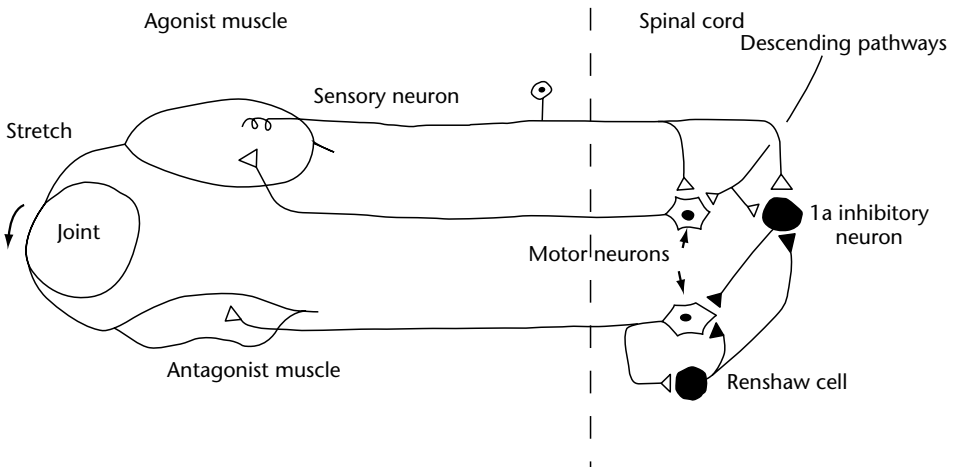


Figure 2. The circuit of the stretch reflex. When a muscle is stretched, spindles in the muscles sense the change in length and excite the motor neurons that cause the muscle to contract. The sensory neuron also excites the 1a inhibitory neuron, which inhibits the excitatory motor neurons of the antagonistic muscle. This is feedforward inhibition. The motor neurons also excite a class of inhibitory neurons called Renshaw cells, which provide feedback or recurrent inhibition to the same motor neurons. The Renshaw cells also inhibit the 1a inhibitory neurons, thus producing disinhibition. The excitatory neurons and synapses are indicated in outline, and the inhibitory neurons and synapses are indicated in black. The different inhibitory pathways are only illustrated for the antagonistic muscle. Identical circuits exist for the agonist muscle. Descending pathways from the cortex, cerebellum and spinal nuclei excite both motor neurons and inhibitory neurons.

receptors and descending pathways from the brain. This means that the same stretch reflex circuit can be used to coordinate voluntary movements and produce both the reciprocal activity that is required for walking and the co-contraction of opposing muscles that is required for precision movements.

The motor neurons themselves can be inhibited by another type of neuron, called the Renshaw cell, which also makes only local connections. A Renshaw cell lies in the vicinity of the motor neuron and is excited by it. In turn, it inhibits the motor neuron and its synergists. This is called *recurrent inhibition*, and the negative feedback that it provides regulates the excitability of the motor neurons and confines activity to the most active motor neurons through the mechanism of lateral inhibition. The Renshaw cells also inhibit the 1a inhibitory neurons of the antagonistic motor neurons, and so produce *disinhibition*. The motor neurons, the Renshaw cells and the 1a inhibitory neurons also receive long-range inhibition and excitation from descending spinal pathways. These descending pathways to inhibitory neurons can thus alter the excitability of all of the motor neurons around a joint. Motor neurons do not show strong spike frequency adaptation, so the major mechanism of inhibition in these spinal circuits is synaptic rather than intrinsic.

Local circuits involving feedforward and feedback inhibition and excitation are found throughout the central nervous system. Even simple circuits of excitatory and inhibitory neurons can generate great functional variety. Small variations in the spatial patterns and strength of connections, small differences in the intrinsic excitability of neurons and the timing of the interactions can produce large differences in the overall behavior of the system. For example, in four-legged animals the same basic spinal circuits can produce standing, walking, trotting and galloping.

NEUROMODULATORS AND INHIBITION

Neuromodulators act on inhibitory cells to vary their action. The thalamus is a nucleus that relays sensory information from the eyes, ears, touch receptors, etc. It also receives a variety of relatively diffuse inputs from the brainstem. When the parabrachial nucleus of the brainstem releases the transmitter acetylcholine from its terminals in the thalamus, it dramatically switches the firing mode of the thalamic neurons from burst to tonic firing. Nitric oxide, which is contained in the same parabrachial terminals, may also promote the transition from burst to tonic firing. The ability to change the mode of firing underlies the changing pattern that is seen in the electroencephalogram (EEG) when there is a transition from rapid eye movement (REM) to non-REM sleep. The former is

desynchronized and high frequency, whereas the latter is characterized by slow wave rhythms and is associated with a lack of activity in the acetylcholine-containing parabrachial terminals. In non-REM (slow-wave) sleep, inhibition is strongly active and a cycle of excitation and inhibition creates large rhythmic potentials in the thalamus and cortex. The effect of acetylcholine inhibits the inhibitory cells in the thalamus and excites the excitatory cells. This paradoxical action occurs because the two cells have different receptors for acetylcholine. As a result, the inhibitory cells are relatively inhibited and thus cannot contribute the inhibition that is necessary to maintain the rhythm.

In other neurons (e.g. in the olfactory bulb), noradrenaline (norepinephrine) reduces the release of inhibitory neurotransmitter. Noradrenaline and acetylcholine are also modulators of the intrinsic inhibition involved in firing frequency adaptation. Noradrenaline acts on channels that make the membrane slightly more negative, so potentiating the strength of adaptation. However, it also blocks firing frequency adaptation by blocking the activation of the membrane channels that normally open and act to slow the firing frequency. Acetylcholine has a similar effect. Both of these neurotransmitters are produced by neurons that form the diffusely projecting systems of the brain. Since these systems produce their relatively long-duration effects over a wide area of brain, they are probably involved in processes that alter the state of arousal or vigilance.

Further Reading

Evarts EV, Wise SP and Bousfield D (eds) (1985) *The Motor System in Neurobiology*. Amsterdam: Elsevier Medical Press.

Koch C (1999) *Biophysics of Computation*. New York: Oxford University Press.

Mize RR, Marc RE and Sillito AM (eds) (1992) *GABA in the Retina and Central Visual System. Progress in Brain Research*, vol. 90. Amsterdam: Elsevier Science Publishers.

Nicholls JG, Martin AR and Wallace BG (eds) (1992) *From Neuron to Brain*, 3rd edn. Sunderland, MA: Sinauer Assoc.

Shepherd GM (ed.) (1998) *The Synaptic Organization of the Brain*, 4th edn. New York: Oxford University Press.

Neural Oscillations

Intermediate article

Xiao-Jing Wang, Brandeis University, Waltham, Massachusetts, USA

CONTENTS

Oscillations represent a general feature of neural firing patterns, produced by dynamical interplay between cellular and synaptic mechanisms. Large-scale synchronous neural population rhythms reflect different behavioral states, play a role in neural synchronization that contributes to the neuronal encoding of sensory stimuli, or may be correlated with cognitive processes such as attention and working memory.

INTRODUCTION

Integrative operations in the brain involve coordinated neural firing patterns in large-scale neural networks. Ever since Hans Berger discovered distinct electroencephalogram (EEG) wave patterns in sleep and waking states in 1929, synchronous brain rhythms have been recognized as one of the most conspicuous types of neural population dynamics. Describing a 'mathematical outlook' of the cortex, Charles C. Sherrington wrote a half-century ago:

> A scheme of lines and nodal points, gathered together at one end into a great ravelled knot, the brain, and at the other trailing off to a sort of stalk, the spinal cord. Imagine activity in this shown by little points of light. Of these some stationary flash rhythmically, faster or slower. Others are travelling points, streaming in serial trains at various speeds. The rhythmic stationary lights lie at the nodes. The nodes are both goals whither converge, and junctions whence diverge, the lines of travelling lights. (Sherrington, 1951, p. 176)

Today, there is a resurgence of interest in brain rhythms for two reasons. First, it is now known that neural rhythms occur in alert and behaving states (as well as in sleep), and that they may play important roles in cortical functions. Secondly, technical advances have led to significant progress in the elucidation of the cellular and synaptic mechanisms of these rhythms.

Broadly speaking, there are three categories of brain rhythms.

1. *Spontaneous rhythms* occur in brain states characterized by the absence of sensory inputs. These spontaneously occurring oscillations usually have low frequencies (< 15 Hz), and are remarkable for their large-scale synchronization across brain structures, detectable as large-amplitude brain waves by the EEG. Examples include the spindle rhythm (4–12 Hz) and delta rhythm (3–4 Hz) in the thalamocortical system, which are the hallmarks of non-dreaming sleep and disappear during dreaming sleep or wakefulness (Figure 1(a)). Spindle and delta rhythms are often temporally nested with another very slow rhythm (< 1 Hz).
2. *Induced rhythms* have been observed in waking states. They are typically evoked by external sensory stimulation and correlated with certain behaviors. In the olfactory bulb, fast synchronous oscillations (gamma at 30–60 Hz) can be induced by olfactory stimulation. For a given odor stimulus, the 'induced rhythm' (so named by E. D. Adrian, who discovered it) is synchronized transiently, in time (for a few hundreds of milliseconds), and only among a neural subpopulation which is selectively responsive to that particular odorant. Similarly, in the visual cortex of the cat and monkey, gamma oscillations were observed in evoked neural responses to optimal visual stimuli (Figure (1b)), and synchronization is realized within subgroups of neurons with similar stimulus specificity. Gamma oscillations are also present in the hippocampus and nearby limbic structures, where they are temporally nested with a slower, theta (4–10 Hz) rhythm. Theta rhythm occurs during exploratory movements and spatial navigation. Physiological studies of behaving rodents show that at any given time, only a subset of hippocampal neurons that encode the animal's current (or immediate future) spatial location is synchronized at a particular phase of the theta cycle (Figure 1(c)).

Therefore, in contrast to the 'spontaneous rhythms', neural rhythms of the waking brain are usually characterized by the presence of a prominent fast oscillatory component (gamma at 30–60 Hz). Synchronization is subtle, typically confined to a restricted (possibly small) subpopulation within a brain area, and occurs intermittently by short episodes in time. The synchronous events propagate from one neural

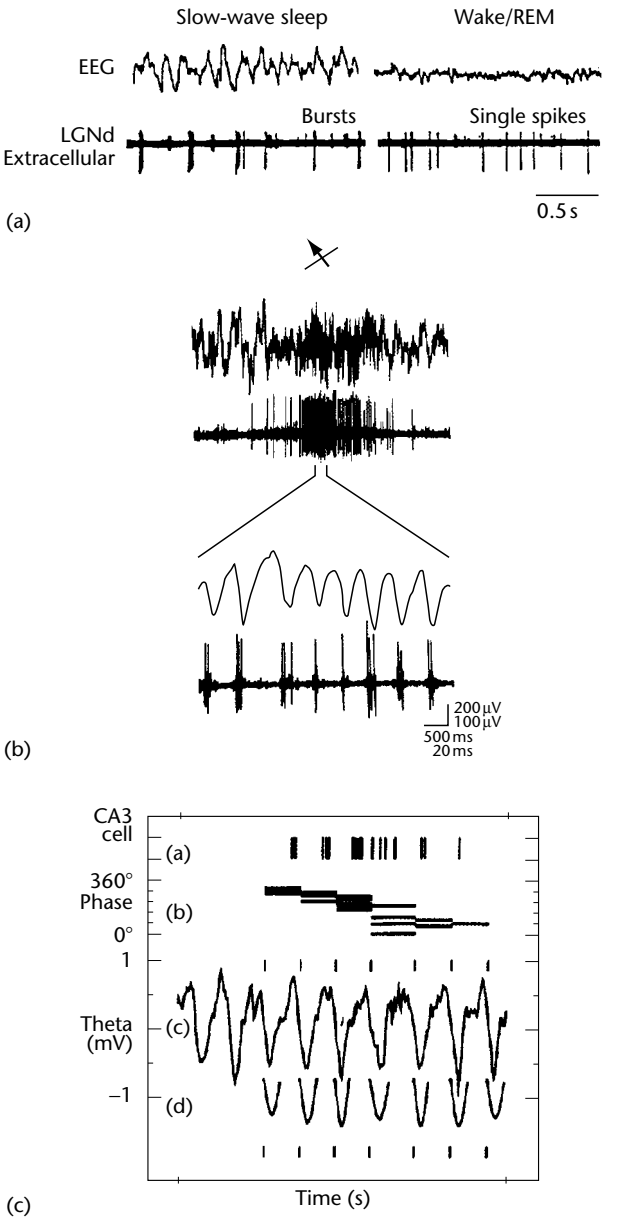


Figure 1. Neural rhythms correlated with behavior. (a) During slow-wave sleep, EEG shows slow synchronous oscillations and thalamic neurons in the dorsal lateral geniculate nucleus (LGNd) fire bursts of action potentials (left panels). In contrast, during waking or dreaming sleep (waking/REM), large-amplitude EEG oscillations are absent, and thalamic neurons fire single spikes tonically (right panels). Adapted from McCarley RW, Benoit O and Barrionuevo G (1983) *Journal of Neurophysiology* **50**: 798–818. (b) Multi-unit and local field potential from the cat primary visual cortex, in response to an optimally oriented light bar stimulus. In the upper two traces, the onset of the response is associated with an increase in *c.* 40 Hz oscillations, which are shown at an expanded timescale in the lower two panels. From Gray CM and Singer W (1989) *Proceedings of the National Academy of Sciences of the USA* **86**: 1698–1702. (c) Burst firing pattern of a hippocampal place cell and its relationship to the EEG theta rhythm as the rat runs through the place field on a narrow linear track. Upper panel: each spike is represented as a single vertical line. Middle panel: the phase of each spike relative to the theta cycle within which it falls is represented by a horizontal line. Lower panel: hippocampal theta rhythm is recorded at the same time as the neural spikes. Note that each successive burst moves to an earlier phase of the theta cycle than the previous burst, as shown by the descending staircase of the phase correlates in the middle panel. Adapted from O'Keefe J and Recces ML (1993) *Hippocampus* **3**: 317–330.

assembly to another (e.g., in response to varying stimuli).

3. *Pathological rhythms* are associated with certain neurological conditions, such as spike-and-wave patterns (3–4 Hz) of epileptic seizures, or tremors (3–8 Hz) characteristic of Parkinson disease. They may be viewed as pathological cases of low-frequency and extremely synchronous oscillatory brain states.

Are there general principles for the rhythmogenesis in the brain? Typically, a coherent neural network rhythm is generated within a relatively localized brain circuit, which is nevertheless composed of many hundreds or thousands of neurons. Therefore studies of the neural mechanisms underlying brain rhythms provide an excellent means of investigating how circuit dynamics emerge from the interplay between synaptic and intrinsic cellular properties. There are two general questions with regard to the mechanisms of a coherent oscillation.

1. What determines its oscillation frequency? Are there pacemaker neurons or is the rhythmicity largely a network phenomenon?
2. What are the synaptic mechanisms for network synchronization?

These two issues will be examined in turn.

CELLULAR PACEMAKER MECHANISMS

Single neurons in the CNS are endowed with a large repertoire of voltage- and calcium-gated ion channels, distributed across the dendritic and somatic membrane, which can give rise to complex neuronal dynamics. In general, oscillation occurs in a single cell, when a strong fast positive feedback (generating the rising phase of membrane voltage) interacts with a slower negative feedback (producing the decay phase of the cycle). Positive feedback within a cell can be provided by activation of

voltage-gated inward Na^+ and/or Ca^{2+} currents, whereas negative feedback is mediated by either inactivation of inward currents or activation of outward K^+ currents. A group of neurons is described as pacemakers for a brain rhythm if (1) they are endowed with intrinsic membrane properties to display robust oscillations in a well-defined frequency range, and (2) that brain rhythm is critically dependent on the integrity of these cells.

Spindle sleep rhythm and rebound bursts in thalamic neurons

Spindle oscillations during quiet sleep originate in the thalamus, and have been reproduced *in vitro* in thalamic brain slices. It was discovered by Jahnsen and Llinás that thalamocortical projection cells and inhibitory neurons in the nucleus reticularis show two modes of firing patterns. On depolarization they discharge single spikes tonically, whereas on hyperpolarization they fire bursts of spikes, possibly in a rhythmic fashion (Figure 2(a), upper panel). During quiet sleep, thalamic cells are in the bursting mode and entrain the spindle oscillations in the entire thalamocortical system. Waking is associated with a switch of thalamic cells from the bursting to tonic firing mode, due to an increase in the neuromodulatory (cholinergic, noradrenergic and other) inputs.

The bursts of spikes are produced by a low-threshold voltage-gated Ca^{2+} ion channel I_T (of the T-type), which de-inactivates during hyperpolarization, and a hyperpolarization-activated cation channel I_h. Such a bursting mechanism is demonstrated by the single thalamic neuron model in Figure 2(a) (lower panel). Intuitively, rhythmic bursting can be generated as follows (Figure 2(a), upper panel). A hyperpolarizing input slowly activates I_h and de-inactivates I_T. The build-up of the I_T eventually leads to a depolarization wave that triggers a *rebound burst* of rapid (250–500 Hz) spikes. The burst is terminated by the inactivation of the same I_T at depolarized voltage, and the oscillatory cycle can start over again. The period of oscillation (*c.* 100 ms) is determined by the inactivation time constant of I_T and the activation time constant of I_h during hyperpolarization.

Gamma (c. 40 Hz) rhythm and chattering neurons in the neocortex

The mechanisms for the fast gamma (30–70 Hz) oscillations, commonly observed in waking and behaving states, have not yet been fully elucidated. In the neocortex, intrinsic gamma oscillations have been reported in a subclass of neurons termed 'chattering cells' (Figure 2(b), upper panel). These cells display repetitive burst firing in the gamma-frequency range, with intraburst spike rates of 300–500 Hz. A compartmental model suggests that the fast rhythmic bursting in chattering neurons is generated by a Ca^{2+}-independent ionic mechanism (Figure 2(b), lower panel). Instead, it relies on voltage-gated Na^+ currents in the dendrite. In this model, perisomatic action potentials propagate back to the dendritic sites, where an Na^+-dependent slow depolarization is produced, which in turn triggers more spikes in the soma. This somatodendritic 'ping-pong' interplay underlies a burst of spikes, which is terminated by the activation of a K^+ current. The de-activation of the K^+ current during hyperpolarization leads to recovery and eventually to the start of a new burst. Experimental evidence in support of such an Na^+-dependent, Ca^{2+}-independent mechanism has recently been reported from cortical slice studies. It has yet to be demonstrated that chattering neurons indeed play the role of pacemakers for gamma oscillations in the neocortex *in vivo*.

Hippocampal theta (7–10 Hz) rhythm and pacemaker neurons in the medial septum

The theta rhythm in the hippocampus and surrounding limbic structures is believed to be critically dependent on the input pathway from the medial septum, where pacemaker-like neural discharges have been observed. There are two major cell types in the septum, which are thought to play distinct roles in the generation of theta rhythm. Cholinergic cells slowly modulate the excitability of hippocampal neurons, whereas gamma-aminobutyrate (GABA)-ergic cells play the role of pacemakers. Recent physiological studies have revealed that non-cholinergic (putative GABA-ergic) neurons in the medial septum display robust intrinsic oscillations in the theta-frequency range, where clusters of spikes are inter-nested in time with episodes of *subthreshold membrane oscillations* (Figure 2(c), upper panel). Interestingly, similar membrane oscillations have been observed in single principal neurons of the rat olfactory bulb, which displays prominent gamma and theta rhythms. A conductance-based model (Figure 2(c), lower panel) suggests that such intrinsic rhythmicity can be generated by a low-threshold, slowly inactivating K^+ current I_{KS}. When the cell fires, the I_{KS} slowly de-inactivates during spike after-hyperpolarization, and a sufficiently large I_{KS} eventually terminates the spiking episode. When the cell does not fire spikes, I_{KS} slowly decreases due to inactivation, until the cell is sufficiently

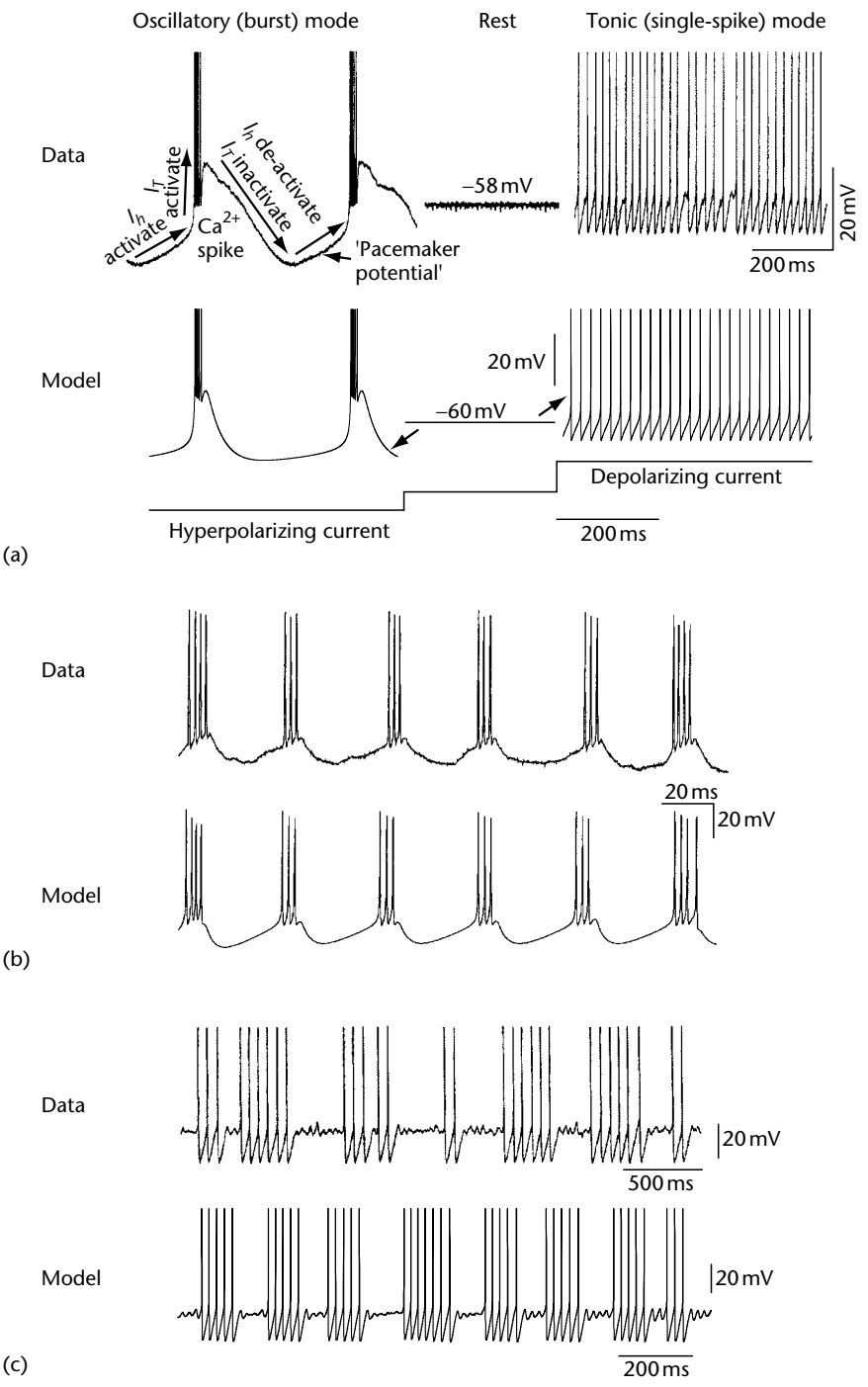


Figure 2. Intrinsic membrane oscillations of single neurons. Each panel shows experimental data and model simulations. (a) A thalamic relay cell displays two distinct spiking modes, namely tonic firing on depolarization, and burst discharges on hyperpolarization. Upper trace adapted from McCormick DA and Pape HC (1990) *Journal of Physiology* **431**: 319–342; lower trace adapted from Wang X-J (1994) *Neuroscience* **59**: 21–31. (b) A 'chattering' neuron from the cat visual cortex shows rhythmic bursting in the gamma-frequency range. Upper trace from Gray CM and McCormick DA (1996) *Science* **274**: 109–113; lower trace from Wang X-J (1999) *Neuroscience* **89**: 347–362. (c) A non-cholinergic (putative GABA-ergic) cell in the rat medial septum displays rhythmic alternations at theta frequency between 'clusters' of spikes and epochs of subthreshold membrane potential oscillations. Upper trace from Serafin M *et al*. (1996) *Neuroscience* **75**: 671–675; lower trace from Wang X-J (2002) *Journal of Neurophysiology* **87**: 889–900. The simulated oscillation is faster than the experimental data (see the different timescales), because the model simulation was performed at body temperature (37°C), whereas the *in-vitro* trace was recorded at 32°C.

recovered and can start to fire again. The subthreshold oscillations are produced by the interplay between an Na^+ current and the low-threshold activation of I_{KS}. In this model, the periodicity of the theta rhythm is largely controlled by a single current (the I_{KS}) in septal GABA-ergic cells. This hypothetical mechanism has not yet been tested experimentally.

Summary

To summarize, some general remarks can be made. First, a single neuron can display different dynamic (e.g., single-spiking and bursting) modes, which depend on the membrane potential level and are under neuromodulatory control. Secondly, there are at least two general classes of ionic mechanisms for rhythmogenesis, one of which depends on an interplay between Na^+ and K^+ currents, while the other depends on Ca^{2+} currents. Putative pacemaker neurons for the gamma and theta rhythms of the waking brain seem to rely on Na^+ and K^+ currents, whereas pacemaker neurons for the spindle and delta sleep rhythms are critically dependent on Ca^{2+} currents.

Thirdly, a given set of ion channels can generate qualitatively different membrane dynamics, depending on their relative strengths or their distributions across the dendro-somatic membrane. For example, I_T gives rise to subthreshold oscillations (at about 10 Hz) rather than bursts in inferior olive cells that send climbing fibers to the cerebellum.

Finally, subthreshold oscillations and repetitive bursting may have different implications for synchronization of coupled neurons. Subthreshold oscillations could subserve a signal carrier for phase-locking and resonance, by virtue of the cell's sensitivity to small but precisely timed inputs (at the peak of the membrane oscillation cycle). On the other hand, bursts may provide a reliable signal for the rhythmicity to be transmitted across probabilistic and unreliable synapses between neurons.

SYNAPTIC NETWORK MECHANISMS

A neural circuit, whether it is thalamic, neocortical or hippocampal, consists of two major cell types, namely excitatory principal neurons and inhibitory interneurons. It follows that three types of synchronization mechanisms by chemical synapses can be envisaged: recurrent excitation between principal neurons, mutual inhibition between interneurons, and feedback inhibition through the excitatory–inhibitory loop.

Recurrent Excitation Model

Recurrent excitatory connections have historically been the first synchronization mechanism to undergo detailed experimental and computational analysis. This was motivated by the observation that blockade of synaptic inhibition in a cortical network led to extremely synchronous neural firing patterns which resembled epileptic discharges. Intuitively, mutual excitation is expected to synchronize coupled neurons, if cells that fire earlier in time excite the others and advance their firing times, so that the network is brought to fire in phase. However, model simulations of biophysically realistic coupled neurons have shown that synaptic excitation often delays rather than advances the firing time in the postsynaptic cell (e.g., if its predominant effect is to increase a voltage-gated K^+ current). Therefore the ability of mutual excitation to synchronize depends on the intrinsic membrane properties of the constituent neurons (Hansel *et al.*, 1995; van Vreeswijk *et al.*, 1995). In general, synchronization of normal brain rhythms is not realized by excitation alone, but depends critically on synaptic inhibition.

Interneuronal Network Model

Computational studies have revealed that reciprocal synaptic inhibition is capable of synchronizing certain rhythmic activities in an interneuronal network (Wang and Rinzel, 1992; van Vreeswijk *et al.*, 1995). One general requirement for this mechanism is that the decay time of the inhibitory synaptic current should be long relative to the intrinsic membrane recovery time, and comparable with the oscillation period. For example, $GABA_B$-receptor-mediated inhibition with a time constant of 100–200 ms could in principle synchronize slow oscillations at a few hertz. $GABA_A$-receptor-mediated inhibition with a time constant of about 10 ms is too fast to syncronize an oscillation at a few hertz, but is sufficiently slow for a 40-Hz oscillation (with a period of about 25 ms) (Figure 3(a)). Indeed, a physiological study using *in-vitro* slices provided evidence that $GABA_A$-receptor-mediated inhibition in a hippocampal interneuronal network, without the involvement of pyramidal neurons, could give rise to coherent 40-Hz oscillations (Whittington *et al.*, 1995).

Thus the interneuronal network model suggests a candidate scenario for the gamma rhythmogenesis in the hippocampus. Because this mechanism requires an optimal match between the synaptic time constant and the oscillation period, coherent

network oscillations are possible only within a frequency range. In other words, synchronous oscillations with a well-defined frequency can be realized even without pacemaker neurons.

Feedback Inhibition Model

A competing network mechanism for coherent gamma oscillations is based on feedback between excitatory and inhibitory neural populations. W. J. Freeman first proposed this scenario to explain 40-Hz oscillations observed in the olfactory bulb. Similar models have been applied to the olfactory cortex and to the hippocampus.

A recent study (Fisahn *et al*., 1998) demonstrated such a scenario in hippocampal slices, where spontaneously occurring 40-Hz oscillations have been shown to depend on both excitatory and inhibitory synaptic transmissions. This experiment can be reproduced robustly in a network of pyramidal cells and interneurons, even in a randomly connected network model (Figure 3(b)). Thus the interneural network mechanism and the feedback inhibition mechanism do not have to be mutually exclusive. Rather, the two mechanisms are likely to operate cooperatively in a cortical network.

Note in Figure 3(b) that while the neural population as a whole oscillates in the 40-Hz frequency range, individual neurons fire more randomly and intermittently in time (at about 10 Hz for interneurons and only 2 Hz for principal cells), as is the case in the experiment of Fisahn *et al*. (1998). These intermittently firing dynamics are similar to neural firing activities during gamma rhythms of the intact brain. Such network dynamics cannot be adequately described in terms of coupled oscillators, with individual neurons firing regularly like a clock. A new conceptual framework is needed (Brunel, 2000) to describe coherent oscillations that emerge from complex and irregular firing patterns involving many thousands of neurons in a random network, in a very similar manner to the vivid picture envisioned by Sherrington 50 years ago.

Summary

A major conceptual advance resulting from experimental and computational work was the recognition that inhibitory neurons play a critical role in the synchronization of cortical networks. With the exception of pathological brain rhythms, synchronization of cortical oscillations appears to rely primarily on synaptic inhibition within interneuronal networks and/or through feedback between excitatory and inhibitory cells. Recent studies have revealed that electrical coupling as well as chemical synapses between interneurons could contribute to network synchronicity, and that diverse subclasses of interneurons could be involved in different types of cortical rhythms.

POSSIBLE FUNCTIONS

Neural oscillations are obviously the modus operandi of central pattern generators – neural circuits that produce rhythmic motor movements such as respiration and walking. At the other end of the sensorimotor continuum, in the early stages of sensory information processing, the functional role of network rhythms is less obvious. Several proposals have been advanced, all centered on the idea that neural representation of sensory inputs is dynamic (as opposed to static), distributed in both space and time. This *dynamic stimulus-encoding hypothesis* is supported by data from the olfactory system, where an odor stimulus induces fast oscillations in a number of odor-specific neural assemblies, each of which is briefly synchronized at a different time epoch of the population response. Such activity patterns of oscillating neural assemblies in temporal sequence may be important for encoding, perhaps for temporally dissecting, and ultimately for identifying an odor stimulus. In the hippocampus, the spike times of a pyramidal cell progressively shift to earlier phases of the population theta cycle when the animal passes through the cell's spatial field (Figure 1(c)). Therefore spike firing time relative to a network oscillation cycle, in addition to the firing rate, may contribute to the hippocampal coding of spatial trajectory for animal navigation.

It has been proposed that neural synchronization provides a substrate for feature integration. For example, groups of neurons that detect different local features of a visual object can encode ('bind') the integrated whole image by synchronization of their spike discharges. Note that synchronization between neurons (which is a much more general phenomenon) does not necessarily imply oscillations. Indeed, unit-recording data from the visual cortex indicate that synchronization between nearby neurons can occur without oscillations, but long-range (> 2 mm) synchronization often occurs concomitantly with fast neural oscillations. Similarly, synchronization of a neural assembly in the motor cortex could reflect an integrated representation and association of movement features in a skilled action. This *temporal correlation hypothesis* for the 'binding problem' is still the subject of active debate. The contentious issues include the

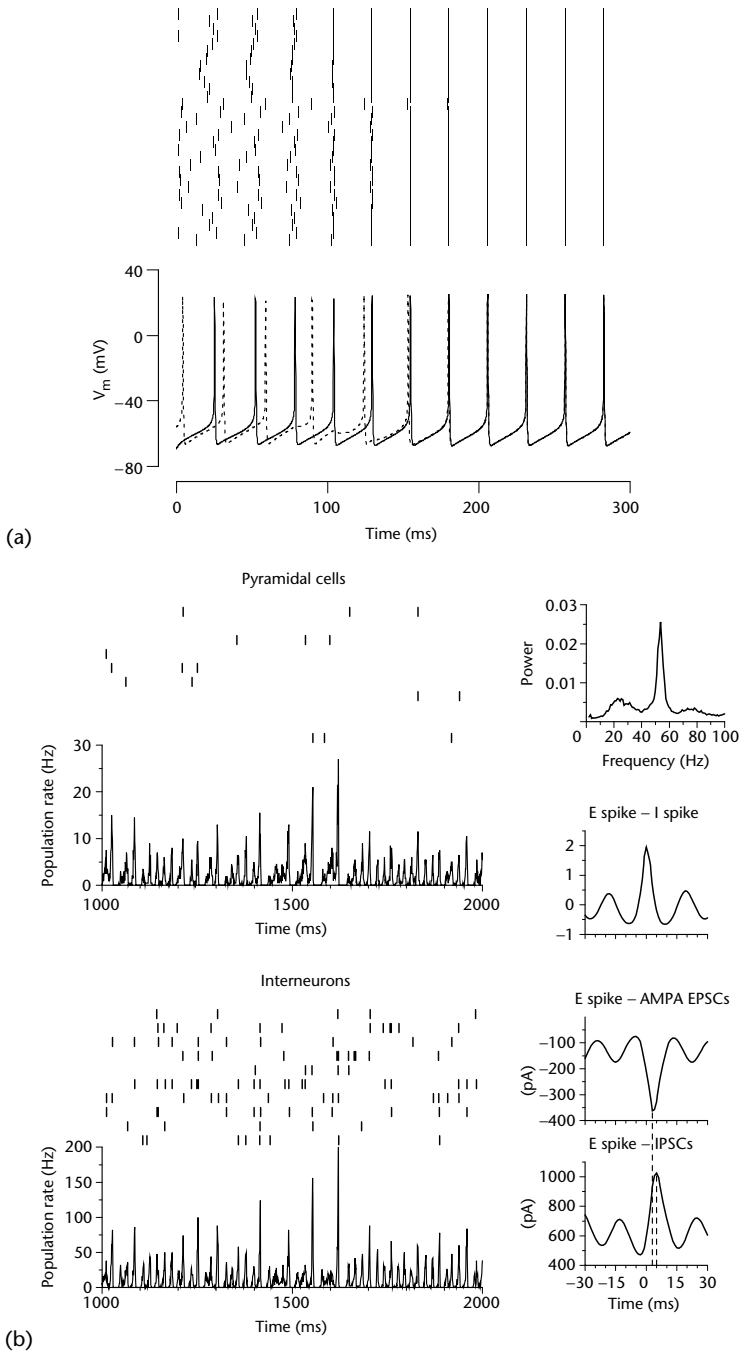


Figure 3. Synaptic mechanisms for network synchronization. (a) The interneuronal network model. A population of interneurons is synchronized at gamma frequencies, by mutually inhibitory connections mediated by $GABA_A$ receptors. Upper panel: the rastergram where each row of vertical bars represents spikes discharged by one of the neurons in the network. Neurons are intially out of phase, but quickly become perfectly synchronous after a few oscillation cycles. Lower panel: the membrane potentials of two neurons. From Wang X-J and Buzsáki G (1996) *Journal of Neuroscience* **16**: 6402–6413. (b) The feedback inhibition model. Computer simulation of a model with two neural populations (pyramidal cells and interneurons) in a sparsely connected random network. The network shows a collective oscillation at 55 Hz (see population rates, and the power spectrum), whereas single neurons fire spikes intermittently in time at low rates (2 Hz for pyramidal cells and 10 Hz for interneurons; see rastergrams). The spike discharges are synchronized to zero-phase between the two populations (second trace on the left), whereas the inhibitory synaptic current shows a phase lag of about 2 ms compared with the excitatory synaptic current (third and bottom traces on the left). From Brunel N and Wang X-J, unpublished data.

prevalence of the oscillatory activities in evoked neural responses, and the extent to which neural oscillations are correlated with sensory perception or motor behavior. Several physiological studies show that fast gamma-frequency-band oscillations in the primate motor cortex usually occur during the preparatory phase of a motor task, before the movement onset, rather than during the motor action itself. Therefore synchronous neural discharges may be associated with anticipation or planning.

Anticipation, planning, and other cognitive processes depend on association cortical areas. Neural circuits in these areas are believed to be highly recurrent, since reverberatory excitations are thought to generate the mnemonic persistent neural activities that are observed in association cortices. Since strongly recurrent network dynamics readily give rise to oscillations, it is possible that in the alert brain of behaving animals synchronous fast oscillations occur as a result of the activation of a highly recurrent cortical circuit (Wang, 1999). If this is so, one would expect an increased occurrence of cortical gamma oscillations with working memory or attention which requires top-down signals from the parietal and frontal cortices. Indeed, working memory (Pesaran *et al*., 2002) and selective attention (Fries *et al*., 2001) have been shown to enhance gamma oscillations in monkeys' cortices. These recent studies, if confirmed, suggest that fast (gamma) rhythm may be a characteristic sign of the engagement of strongly reverberatory cortical circuits in cognition and memory.

References

Brunel N (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of Computational Neuroscience* **8**: 183–208.

Fisahn A, Pike FG, Buhl EH and Paulsen O (1998) Cholinergic induction of network oscillations at 40 Hz in the hippocampus *in vitro*. *Nature* **394**: 186–189.

Fries P, Reynolds JH, Rorie AE and Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**: 1506–1507.

Hansel D, Mato G and Meunier C (1995) Synchrony in excitatory neural networks. *Neural Computation* **7**: 307–337.

Pesaran B, Pezarls JS, Sahani M, Mitra PP and Andersen RA (2002) Temporal structure in neuronal activity during working memory in Macaque parietal cortex. *Nature Neuroscience* **5**: 805–811.

Sherrington CC (1951) *Man on His Nature*. Cambridge, UK: Cambridge University Press.

van Vreeswijk C, Abbott LF and Ermentrout GB (1995) When inhibition, not excitation, synchronizes neural firing. *Journal of Computational Neuroscience* **1**: 313–322.

Wang X-J (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *Journal of Neuroscience* **19**: 9587–9603.

Wang X-J and Rinzel J (1992) Alternating and synchronous rhythms in reciprocally inhibitory model neurons. *Neural Computation* **4**: 84–97.

Whittington MA, Traub RD and Jefferys JG (1995) Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. *Nature* **373**: 612–615.

Further Reading

Buzsáki G and Chrobak JJ (1995) Temporal structure in spatially organized neuronal assemblies: a role for interneuronal networks. *Current Opinion in Neurobiology* **5**: 504–510.

Gray CM (1994) Synchronous oscillations in neuronal systems: mechanisms and functions. *Journal of Computational Neuroscience* **1**: 11–38.

Laurent G (1996) Dynamical representation of odors by oscillating and evolving neural assemblies. *Trends in Neurosciences* **19**: 489–496.

Llinás RR (1988) The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science* **242**: 1654–1664.

Marder E (1998) From biophysics to models of network function. *Annual Review of Neuroscience* **21**: 25–45.

Rinzel J and Ermentrout GB (1998) Analysis of neural excitability and oscillations. In: Koch C and Segev I (eds) *Methods in Neuronal Modeling*, 2nd edn, pp. 251–291. Cambridge, MA: MIT Press.

Steriade M, McCormick DA and Sejnowski T (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* **29**: 679–685.

Traub R, Jefferys JGR and Whittington MA (1999) *Fast Oscillations in Cortical Circuits*. Cambridge, MA: MIT Press.

Neural Prostheses

Introductory article

John K Chapin, SUNY Health Science Center at Brooklyn, New York, USA

CONTENTS

A neural prosthesis is an electronic device that uses electrode stimulation to artificially activate muscle, nerve or brain tissue that has lost its normal functional capacity.

INTRODUCTION

A broad range of neural prostheses (or 'neuroprostheses') are now being developed to palliate neurological dysfunctions throughout the body. Nearly all of these devices transform electrical signals into neural signals by stimulating tissues through electrodes placed on the skin or implanted in peripheral or central nervous system (CNS) structures. The full potential of neural prostheses for restoring human nervous system function has not yet been achieved, though a few such devices are already widely used.

AUDITORY PROSTHESES

The cochlear implant is widely recognized as the most effective means for restoring auditory sensation in patients with sensorineural hearing loss. This disorder involves partial degeneration of the peripheral processes of spiral ganglion cells that normally carry auditory signals from the auditory receptor (hair) cells in the cochlea. The central processes of these ganglion cells, which normally transmit these auditory signals through the auditory nerve to the brain, are less affected. Cochlear implants bypass the degenerated peripheral processes by using microphones to record incoming sounds and then processing them into electrical signals that directly stimulate the neural fibers in the cochlea through implanted electrodes. Since the cochlea contains a 'tonotopic' map, multiple electrodes are needed to transfer differentially high and low tones to the auditory system. Recent research suggests, however, that this stimulation does not by itself produce a sufficiently high fidelity of sound transmission to allow patients to understand normal human speech. Instead, auditory perception in some patients with cochlear implants appears to be sufficiently enhanced by a general increase in auditory nerve cell activity produced by the electrical stimulation (the activity of these nerve cells is severely degraded in sensorineural hearing loss). One hypothesis is that merely increasing the noise input to the auditory nerve cells allows them to more easily transmit natural auditory signals from the receptor cells.

In cases where the hearing loss is caused by damage to the auditory nerve itself, auditory brainstem implants can be placed over the surface of the cochlear nucleus, which is the primary CNS target of the auditory nerve. Current research is focusing on increasing the accuracy and efficiency of such stimulation by implanting stimulating electrodes directly into the cochlear nucleus. Advanced thin-film technologies allow such electrodes to be designed and fabricated to include relatively large numbers of stimulus contacts on arrays of thin 'daggers' that penetrate through the auditory representation in the nucleus. By increasing the packing density of these contacts it will be possible to activate selectively large numbers of functionally specific neural targets using minimal stimulus currents. Some issues remain to be resolved, particularly that of decreasing the nonspecific and artifactual effects of electrical brain stimulation to ensure that the sensory perceptions evoked in the patient are as natural as possible. Although these technologies are currently far from adequate, the use of high-density penetrating electrode arrays might one day allow the extension of the neural prosthetic approach to the restoration of function in the brain itself.

VISUAL PROSTHESES

Despite the possibility of implanting neural prostheses in higher brain areas, most attention still focuses on peripheral sites. Thus, while early

investigations found some success in using electrode arrays to stimulate the surface of the visual cortex, current efforts to develop visual prostheses are focusing more on the development of microfabricated multi-electrode arrays for implantation on the retina. Although this technology is still in its infancy, it could theoretically allow one to input relatively detailed stimulation patterns into the visual system. It remains to be determined, however, whether such large numbers of individual stimuli can be perceived as an integrated visual percept.

SOMATOSENSORY PROSTHESES

As with the other sensory systems, most investigations of neural prostheses in the somatosensory system are focused on the periphery. Nerve cuff electrodes are being developed not only to stimulate nerve fibers that have been disconnected by injury or amputation, but also to record impulse activity of intact peripheral sensory nerves. The latter may ultimately be important for the restoration of somatosensory function after spinal cord injury. In this scenario, neural activity recorded in peripheral sensory nerves would be used to actuate stimulating neural prostheses implanted in somatosensory pathways on the proximal side of the spinal cord lesion.

FUNCTIONAL ELECTRICAL STIMULATION

Many neural prosthesis applications involve the use of implanted or surface electrodes for functional electrical stimulation (FES) of muscles or muscle nerves (functional neuromuscular stimulation, FNS). The cardiac pacemaker might be considered as the prototypical FES neural prosthesis in that it uses electrical stimulation to restore function normally mediated by endogenous neuromuscular control mechanisms.

Implanted FES devices are also used to remedy autonomic or motor dysfunctions caused by spinal cord injury (SCI). For example, SCI-related bowel and bladder dysfunction can be remedied by using implanted neural prostheses to stimulate the nerves that normally control these organs. Similarly, neural prosthetic stimulators implanted on the phrenic nerves or the diaphragm muscles are used to pace breathing movements in patients with SCI-induced respiratory muscle paralysis. Finally, FES stimulators have been widely studied for their potential to activate the limb musculature of paralysis patients. This approach is useful not only because of its ability to produce movements, but also because it can allow paralyzed patients to exercise their muscles in order to maintain their trophic integrity. Traditionally, stimulating electrodes mounted on the skin surface have been used to activate specific groups of skeletal muscles in patients with various paralysis syndromes. An early example was the alleviation of 'foot-drop' by stimulating the nerve that lifts the foot up during stepping. More recently, FES stimulation electrodes have been implanted in or near target muscle groups. Although these implanted systems present somewhat greater risks of infection, they can produce more specific movements with lower stimulus currents and hence less tissue pathology than is caused by surface electrodes. Much effort is currently being devoted to the development of FES stimulating electrodes that do not require implanting wires through the skin, but instead are implanted with data receivers and inductive power systems that allow them to be controlled and sustained through wireless connections.

There are, however, unresolved problems with the use of FES to activate muscles directly. For example, such stimulation activates both sensory and motor fibers within these nerves. Thus the same nerve stimulation that produces bladder contraction also stimulates a reflex loop that closes the bladder's sphincter, preventing urine from flowing out. For this reason, such patients often require surgical sectioning of the sensory nerves that mediate this reflex (posterior rhizotomy) to ensure that the sphincter is relaxed during bladder voiding. Another problem is that muscle nerve stimulation tends to recruit muscle fibers in an incorrect order, producing movements that are jerky and difficult to control. This is just one of many problems that prevent the clean and selective stimulation of all the muscles that are normally used to produce smoothly coordinated movements.

One possible solution to these problems may be to activate movements by stimulating through electrodes implanted in interneuron pools in the spinal cord. In theory, spinal cord neural prostheses could be more effective and less cumbersome than traditional FES devices. Moreover, spinal circuits are thought to help integrate and finely coordinate the timing and recruitment of different muscles, and also to regulate sensory feedback from the periphery. Current investigations are determining whether spinal cord implants can selectively activate these circuits and thereby elicit smoothly coordinated patterns of muscle contraction. For example, through stimulation of the appropriate subzone of the sacral spinal cord one might

produce simultaneous bladder contraction and urethral sphincter relaxation, thus obviating the need to perform posterior rhizotomy. Recent research suggests that spinal cord stimulation may also improve neuroprosthetic control of limb movements. Localized microstimulation of spinal interneuron pools can produce integrated multi-muscle directional movements and also a relatively normal order of muscle fiber recruitment.

PROSTHETIC CONTROL

Along with technological improvements in FES systems and other neural prostheses to activate motor systems, one must also improve the methods for controlling them. Current motor prostheses tend to employ either mechanical or electromyographic (EMG) recordings from unparalyzed body parts (e.g. the shoulder) to allow the patient to control the prosthesis. Electroencephalographic (EEG) recordings from scalp electrodes have even been used to allow paraplegics to move computer cursors slowly across a screen to spell words. To obtain natural and efficient control of a body limb, however, the command signals would best be recorded directly from the areas of the nervous system that are normally used to process that motor function. It has long been known that voluntary movements are planned and executed through neural processing in the primary and premotor cortices, and that various mechanical parameters of such movements are encoded in the spiking activity of neurons in these brain areas. Although single neuron recordings cannot provide enough information to drive a motor prosthesis, technology is now available for simultaneously recording more than 100 neurons from the brains of rats and monkeys performing trained limb movements. Moreover, it has recently been shown that it is possible to use online computers to 'decode', in real time, the motor control information from such simultaneously recorded neuronal populations and to use this motor information to move a robot arm. These results have demonstrated the essential feasibility of using cortical recording arrays to allow paralyzed patients to control an external robot arm, or even an FES device.

One potential problem is that it may be difficult to extract appropriate motor signals from the motor cortices of long-term paralysis patients. Indeed, it is generally believed that functionally deprived cortical areas tend to be 'remapped' to assume the functions of adjacent areas. This potential problem may, however, be solved by the patient's own ability to adapt and learn. In fact, recent studies have shown that animals using their cortical neurons to control a robot arm can rapidly adapt these neurons to optimize this control. Thus human quadriplegics might, with practice, learn to use their motor cortical 'arm' areas to control a robot arm, an orthotic device, or even their real arm via neural prosthetic stimulators implanted in the muscles or the spinal cord.

Further Reading

Chapin JK (2000) Neural prosthetic devices for quadriplegia. *Current Opinion in Neurology* **13**: 671–675.

Chapin JK and Moxon KA (eds) (2001) *Neural Prostheses for Restoration of Sensory and Motor Function*. Boca Raton: CRC Press.

Davis R, Patrick J and Barriskill A (2001) Development of functional electrical stimulators utilizing cochlear implant technology. *Medical Engineering and Physics* **23**: 61–68.

Grill WM and Kirsch RF (2000) Neuroprosthetic applications of electrical stimulation. *Assistive Technology* **12**: 6–20.

Grill WM, McDonald JW, Peckham PH *et al*. (2001) At the interface: convergence of neural regeneration and neural prostheses for restoration of function. *Journal of Rehabilitation Research and Development* **38**: 633–639.

Loeb GE, Peck RA, Moore WH and Hood K (2001) BION system for distributed neural prosthetic interfaces. *Medical Engineering and Physics* **23**: 9–18.

Maynard EM (2001) Visual prostheses. *Annual Review of Biomedical Engineering* **3**: 145–168.

Prochazka A, Mushahwar VK and McCreery DB (2001) Neural prostheses. *Journal of Physiology* **533**: 99–109.

Neural Regeneration

Introductory article

Christine E Bandtlow, University of Innsbruck, Innsbruck, Austria
Thomas Oertle, University of Zurich, Zurich, Switzerland

CONTENTS

In the mammalian nervous system nerve fibers of the peripheral nervous system can regrow upon injury; in contrast, fibers of the central nervous system lack this regenerative capacity. Research into the molecular and cellular processes of nerve repair is improving the prospects of successful treatment of spinal cord injury.

INTRODUCTION

Restoration of axonal function after nerve damage is not only a topic of great research interest, it is also an area of vital concern to people with spinal cord injuries. A prerequisite for functional recovery is the successful regeneration of the injured axons; this depends on many factors, such as survival of the corresponding cell body, the elongation and growth of newly sprouted fibers, and establishment of contacts with the appropriate targets. Cold-blooded vertebrates such as fish and amphibia have been used extensively to study the biochemistry of successful regeneration of both the central nervous system (CNS) and the peripheral nervous system (PNS). In adult mammals, however, an intriguing difference is observed when comparing the capacity to regenerate axons in the CNS versus PNS. The mammalian PNS is able to support axonal regeneration, often resulting in functional restoration of nerve circuits; in the CNS, however, although nerve fibers initially start to sprout, they fail to regenerate over long distances to contact their target cells, resulting in the permanent loss of function.

REACTION OF NEURONS TO INJURY

Injury or trauma of the nervous system can vary from minor contusions or crushes to open injuries associated with partial or even complete disruption of the nerves. Any type of axonal injury induces an array of dramatic molecular and cellular responses at the level of the cell bodies and at the site of the injury. Following an injury, a sequence of reactions takes place that has a direct impact on the potential for successful regeneration. One of the most immediate changes is the destruction of tissue at the site of the injury, or lesion, which can vary greatly depending on how the injury occurs. Damage to the axons of nerve cells may lead in its most severe form to the complete transection of the nerve fiber and thus to the subsequent death of the neuronal cell body. The most important requirement for regeneration of injured axons is the successful survival of the nerve cell bodies and their capacity to meet the new commands to initiate axonal regrowth and reconnection with the original targets. The closer the axonal insult occurs to the cell body, the less likely it is that the neuron can be maintained in a healthy state. Injuries that sever axons some distance from the cell body produce a severe reduction in neuronal size (atrophy) several weeks after axotomy, but the neurons will normally survive. Surviving neurons switch their metabolism in such a way that restoration of the axon will be facilitated. They undergo morphological and molecular changes involving major rearrangements of the endoplasmic reticulum and Golgi apparatus, a process referred to as chromatolysis. Chromatolysis is often accompanied by an increase in expression of a set of genes associated with axon growth and protein synthesis. These metabolic changes probably meet the affected cell's requirements to initiate regeneration. Even if the cell ultimately survives the injury, the nerve terminals and the entire segment distal to the lesion site (i.e. the one that has lost contact with the cell body) will degenerate. This process, called 'Wallerian degeneration', is associated with myelin breakdown and the removal of axon and myelin debris. Early experimentation in the field of regeneration in

mammals indicated that regrowth of axons could occur in the peripheral nervous system, such as the nerves of the arms and legs. However, similar experiments performed in the CNS suggested that damaged neurons of the CNS could not regrow. For this reason, it was long believed that there is a primary difference between the regenerative capacity of central and of peripheral neurons which may explain the lack of successful regeneration of fibers after spinal cord injuries.

NEURONAL REGENERATION IN THE PERIPHERAL NERVOUS SYSTEM

Damage to the PNS is frequently reversible. Under certain circumstances neurons are able to survive and regenerate their axonal projections to reestablish contact with their former target cells. Provided that such connections are regained, a considerable degree of function can be restored. The reason peripheral neurons are more likely to survive axonal damage may be due to many factors, both intrinsic and extrinsic to the neurons. Studies of the development of peripheral nerves demonstrated that neurons require an array of neurotrophic factors for survival and maintenance. These factors are not only important in the development of the PNS but also in regeneration. Neurotrophic factors are polypeptides required for survival of discrete neuronal populations (sensory, sympathetic, motor neurons). They are divided into three major families: the neurotrophins, the cytokines, and the transforming growth factor family. Such trophic factors are produced by many different sources, such as the innervated target cells and the axon-surrounding glial cells, the Schwann cells. Production of neurotrophic factors by Schwann cells may explain why the closer the axonal injury is to the cell body, the less likely the neuron is to survive.

Upon damage to the PNS a remarkable sequence of cellular and molecular changes, associated with Wallerian degeneration, occur in the distal nerve segment. Morphological changes appear in the distal nerve segments during the first 3 days after injury, the segments becoming fragmented and starting to shrink, and within a few weeks the entire distal segment is destroyed. One of the first cellular responses is the proliferation and infiltration of macrophages. Macrophages are predominantly recruited from the circulating pool of hematogenous monocytes. This response begins within 2 days of axonal injury and reaches its peak by 4–7 days. Macrophages have an enormous phagocytic ability and can remove myelin and axonal debris within a few days. This is in sharp contrast to the CNS, where myelin clearance may take months. After the initial extrusion of myelin debris by macrophages, Schwann cells that are normally in close contact with the axon detach, dedifferentiate and proliferate. Dividing Schwann cells line up along the preserved basal lamina tube that surrounded the original fiber to form endoneurial Schwann cell tubes (bands of Büngner) which function as channels for regenerating sprouts from the proximal segment. Most importantly, however, Schwann cells prepare the local environment that supports and guides regenerating nerve fibers to their denervated targets. They not only produce a variety of cell surface molecules and basement membrane components that support neurite growth, they also express neurotrophic factors and their corresponding receptors to ensure the survival of regenerating axons and their corresponding cell bodies.

Full recovery from nerve injury requires reestablishment of connections between neurons and their appropriate targets. Reinnervation is more accurate after crush lesions than after transections. In a crush injury axons are severed, but the integrity of the nerve sheaths and the basement membrane of myelinated axons is maintained. Regenerating axons seem to prefer to grow inside the internal structures of endoneural Schwann cell tubes, suggesting that the basement membrane is essential as a guidance structure for growing axons to find their appropriate targets. In contrast, a completely transected nerve with separation of the proximal and distal nerve stumps will not successfully regenerate without surgical intervention, despite the proliferation of Schwann cells and fibroblasts at the site of the lesion. The ability of axons to sprout from the proximal stump towards the distal stump is limited to a few millimeters. To facilitate axonal crossing of the gap, microsurgical repair is often required to bring the ends of the nerve together as closely as possible. Large gaps, however, need to be bridged in some way. A common method is to use an autografted nerve segment which already contains Schwann cells, essential for axonal regeneration. However, if several pieces of nerve are required (e.g. to repair complex, highly branched nerves like the brachial plexus) the material available for grafting is often limited. In view of this problem, alternative methods are being developed which encourage axonal regeneration between the cut ends of the nerve stumps. Using tubes of silicone or other materials to bridge the gaps, axons can regenerate up to 1 cm between the stumps of the cut nerves. To encourage axons to grow over even

longer distances, a variety of materials including Schwann cells or trophic factors and extracellular matrix (ECM) molecules which are known to stimulate axonal growth have been inserted into the tubes.

The rate of regeneration depends not only on the type of injury but also on the age and type of fibers that are affected. In experimental animals such as rodents, the rate of axonal elongation is 3–4 mm per day in a crushed nerve, but only 2.5 mm per day after a complete transection. In humans the distances over which nerves must regenerate are correspondingly longer and the daily growth rate may be only 1–1.5 mm. Since delay in reinnervation may result in atrophy of the endoneurial tubes and (more importantly) degeneration of the denervated muscles, there has been considerable interest in speeding up regeneration. Various methods have been used including electromagnetic stimulation and treatment with hormones or trophic factors that have been shown to stimulate neural growth in a culture dish. Reinnervation of denervated target organs such as muscle and skin is a prerequisite for functional recovery. Most often it occurs with a relatively high specificity, even from mixed sensory and motor nerves. However, experimental observations also indicate that regenerating fibers explore several pathways by sending out sprouts (collateral branches) into inappropriate endoneurial tubes, which results in disorganization and imprecision of their innervation patterns.

NEURONAL REGENERATION IN THE CENTRAL NERVOUS SYSTEM

In contrast to peripheral nerves, damaged axons in the adult mammalian brain and the spinal cord do not spontaneously regenerate and consequently there is little functional recovery. When a spinal cord injury occurs, many axons are damaged or destroyed, which disrupts communication of signals between the brain and neurons below the level of the injury. As a result, normal bodily functions are impaired or lost. The extent of paralysis depends on the severity and location of the damage. If the site of injury occurs at the waist, only movement and sensation to the legs are lost. The closer the level of injury is to the brain, the more extensive the loss.

The CNS Reaction to Injury

An injured CNS axon does have an encouraging initial response to axotomy. As in the PNS, the surviving proximal segment of the axon is seen to exhibit a regenerative response within 6 h of injury as 'sprouts' develop at the severed nerve end which are directed towards the injury site. Disappointingly, these sprouts are able to extend for only up to 1 mm before growth is aborted and the new sprouts are gradually resorbed and retracted. The reason why the regenerative process cannot be furthered is likely to be multifactorial, given the complexity of the CNS.

Intrinsic Neuronal Properties

One major hurdle is that injured CNS neurons often fail to reinitiate and maintain a general growth program required for axonal elongation. In the PNS, particular growth-associated proteins (GAPs) and cytoskeletal proteins typically present in growing axons during development are reexpressed upon injury, and remain increased throughout the regenerative period. In contrast, the initial upregulation of such proteins is often only transitory or completely absent in the majority of injured CNS neurons. The sustained reactivation of the genes expressing these proteins appears to be crucial in determining the success or failure of axon regeneration.

Extrinsic Factors

The CNS as a nonconducive growth environment

Besides the differential regulation of intrinsic properties that determine a neuron's commitment to regrow, the local environment of the CNS tissue is nonconducive to neuronal growth. Even peripheral neurons which are able to regrow in PNS tissue show only limited growth when transplanted into the brain or spinal cord of adult vertebrates. In contrast, injured CNS neurons were seen to grow through peripheral nerve grafts which were transplanted as bridges into the brain or spinal cord of adult rats. Fibers that entered the grafts ceased growth, however, when they reentered CNS tissue. These experiments demonstrated that (a) adult CNS neurons process mechanisms for axonal growth over long distances if they are provided with an appropriate environment; and (b) adult CNS tissue is unfavorable for neuronal growth. What are the molecular signals that make the CNS tissue such a poor environment for nerve repair? Several different cell types are involved in the CNS injury response, and these are recruited at different times. Thus, the environment of a cut axon that is

trying to regenerate will vary over time, particularly in the first 2 weeks after injury.

Secondary cell death

An important event induced within the first few hours after CNS injury is the development of secondary injury changes; these take place at the damaged area for a prolonged period after the injury and can include tissue swelling, chemical imbalances within the tissue, and delayed death of cells within the lesion. The secondary injury was long thought to be due to the continuation of cellular destruction through necrotic (or passive) cell death. However, evidence from brain injury and ischemia suggested that cellular apoptosis, an active form of cell death, could have a role in CNS injury in adulthood. Apoptosis occurs mainly in populations of neurons and oligodendrocytes. The death of oligodendrocytes in white-matter tracts continues for many weeks after injury and may contribute to postinjury demyelination. The mediators of injury-induced apoptosis are not well understood, but there is a close relationship between recruitment of CNS microglia, peripheral monocytes and cell death, suggesting that proinflammatory responses may play an important part. Microglial cells are activated by CNS injury, then divide and migrate to injury sites. After 24 h CNS injuries therefore contain large numbers of activated microglial cells, and they may also contain large numbers of blood-derived macrophages if the vasculature has been damaged. These cells are certainly capable of producing toxic molecules. Activated microglia have most of the properties of macrophages, and when stimulated can undergo a respiratory burst and release free radicals, nitric oxide, arachidonic acid derivatives and other toxic molecules. However, there is also evidence that microglia and macrophages (as in the PNS) can be neuroprotective and help to stimulate the repair response. Whether these cells are beneficial or harmful for nerve repair in the CNS is therefore somewhat controversial.

The glial scar at lesion sites

Traumatic injury to the adult CNS results furthermore in a rapid response from resident astrocytes around the injury site, a process often referred to as reactive astrogliosis or glial scarring. Although the functional role of glial scarring is not completely understood, it has been suggested to be an attempt made by the CNS to restore homeostasis through isolation of the damaged region. There are many different glial cells and cellular signals involved in the formation of the glial scar. Besides astrocytes, leptomeningeal cells, which are normally found on the surface of the CNS, as well as fibroblasts migrate towards the injured site. Within 3–5 days after injury, these cells create a dense plexus by interdigitating their processes and thereby walling off the lesioned region. In order for axons to regenerate, they must grow across the damaged tissue left by the injury, as well as this scar tissue. Observations of the course of fibers at lesion sites show that in adult mammals these scars are not penetrated by regrowing fibers. The high density of the astrocytic process network may provide a physical barrier to the axons. More recent experimental studies question whether this limited growth through scar tissues is simply the result of a physical barrier created by the density or the three-dimensional geometry of the scar. It seems likely that changes in the molecular properties of cells forming the scar create a nonpermissive or inhibitory environment that prevents fiber penetration. Associated with the glial scar is an increased deposition of a number of ECM molecules including tenascin-R, collagens III and IV and chondroitin sulfate proteoglycans. Many of these molecules have been shown to repulse or inhibit neurite outgrowth of a variety of neurons *in vitro* and are therefore strong candidates that prevent regenerative responses of CNS neurons *in vivo*. The molecular signals that induce the upregulation of these proteins are not well established, however.

Myelin-associated neurite growth inhibitors

Central nervous system injuries – in particular spinal cord injuries – are associated with direct damage to myelin sheets leading to the immediate release of myelin debris at the lesion site. It has long been recognized that CNS myelin and the cells that produce it (the oligodendrocytes) inhibit neurite growth *in vitro*. Consistent with this view is the observation that CNS injuries in very young mammals, i.e. before the onset of fiber myelination, are not detrimental for axonal regeneration and often lead to full functional restoration. Furthermore, a delay of myelination in experimental models prolonged the period during which regeneration of injured fibers was possible. Following the realization that myelin in the CNS impedes axonal regrowth, a number of myelin-specific inhibitors have been identified, such as the protein NI250 or Nogo-A, myelin-associated glycoprotein (MAG), tenascin-R and chondroitin sulfate proteoglycans. These molecules are capable of preventing the growth of nerve fibers *in vitro*. The main evidence for the inhibitory effect of myelin on CNS regeneration *in vivo* comes from experiments using

antibodies that neutralize one of the myelin inhibitory molecules, NI250/Nogo-A. In the damaged spinal cord, blockade of this molecule with antibodies can improve regrowth of axons and motor function in experimental animals. It seems that intact myelin may not be inhibitory *per se* but is most inhibitory to axon regeneration where there has been substantial trauma to the CNS. Inhibitory components or inhibitory domains of proteins will become exposed immediately after injury where the damage has actually disrupted myelinated axons. Myelin debris containing inhibitory components will be present not only at the lesion site, but also distal to the lesion site in degenerating myelinated tracts when axons and their surrounding myelin fragment over a period of days following axotomy. In contrast to the PNS, the debris is removed only slowly owing to the low phagocytic activity of the CNS microglial/macrophage cells, and may remain for several months.

REGENERATION IN FISH AND AMPHIBIA

In contrast to mammals, lower vertebrates show spontaneous recovery after injury. This observation is reflected in the difference of epimorphic regeneration in general. When a planarian worm is transected, the head fragment is able to regenerate the tail, and the tail fragment regenerates a new head structure. An adult urodele is capable of regenerating its limbs, tail, jaws, retinas and eye lenses, neural crest, spinal cord and heart ventricle. A teleost fish can regenerate not only its fins but also its spinal cord. Crucial for the absence of these epimorphic regenerative processes is the lack of apparent cellular dedifferentiation in higher vertebrates, i.e. they lack the capacity to successfully reenter the cell cycle from the differentiated state.

Behavioral observations in fish and amphibians also provide evidence for a successful axonal regeneration process in the CNS giving rise to functional recovery. Fish are paralyzed immediately after a spinal transection. They tend to lie on their sides, unable to move their tail and caudal fins. After several weeks, however, their swimming ability is regained. Although there is evidence that not all classes of axons regenerate equally well when severed, the question remains why anamniotes show an excellent capacity for axonal regeneration. One likely explanation is the observation that the CNS tissue in anamniotes responds to an insult in ways that promote regeneration rather than impeding it as seen in amniotes. Spinal cord axotomy in an adult fish appears not to induce cell death among neurons projecting from the brain to the spinal cord. At present the reasons that enable fish CNS neurons to survive after axotomy remains unknown. The identification of genes involved in the axotomy-induced death pathways in mammals may help to elucidate whether similar mechanisms are present in fish and how they are regulated after a lesion.

At the morphological level two observations stand out that distinguish the reaction of the fish CNS from the mammalian. There is no conspicuous glial scar formation, and there is a robust infiltration of macrophages. After a spinal cord lesion in the goldfish, astrocytes seem to dedifferentiate and together with fibroblasts and Schwann cells fill in the gap resulting from the injury. Repopulation of the lesion gap by immature astrocytes has been proposed to support axonal regeneration in the fish as they may produce factors that enhance axonal regrowth. Furthermore, following injury, the number of macrophages and microglia increases dramatically at the lesion site. As in the mammalian PNS, these cells help to remove axonal and myelin debris and may conceivably speed up the removal of putative inhibitory proteins from the local environment. In addition, macrophages may also secrete factors that support axonal regeneration. Another striking difference between the fish and the mammalian CNS is the long-lasting upregulation of growth-associated molecules in neurons of the fish brain that show relatively high rates of regenerative success. Whether the presence of inhibitory molecules associated with oligodendrocytes and myelin, such as MAG and Nogo-A, is a fundamental property distinguishing mammals from fish is not entirely clear, and is a question that will only be solved when the corresponding fish orthologs are cloned. The possibility that such inhibitors are present, although in lower amounts than in the mammalian CNS, cannot be excluded. While in adult mammals regeneration of somatic axons is often inaccurate, resulting in severe motor dysfunction, regeneration of sensory and motor axons in amphibia is quite accurate. In urodeles, regenerating axons find their correct destinations although they may not use their original trajection. It appears that the selective innervation of muscles is based on positional information and the involvement of target recognition molecules. In newts and frogs, motor nerves can be induced to innervate incorrect muscles, but the synapses formed are less stable and release of neurotransmitter per nerve impulse is reduced. In mammals, in contrast, the synapses formed when a nerve innervates an incorrect muscle are as effective as those in correct muscles.

TREATMENT OF SPINAL INJURY

For many years it was believed that the brain and the spinal cord were a hard-wired system where any type of fiber regrowth was truly impossible. However, research into the molecular and cellular signals that impair nerve repair in the CNS has raised hopes that regeneration can be induced under certain conditions.

Current Treatment

A therapy that is already in use in humans with spinal cord injuries is treatment with methylprednisolone, a steroid drug which must be administered to a patient within a short period after the injury. It is believed that this drug acts by preventing the secondary changes after injury, allowing more spinal cord tissue to remain healthy and functional. Patients who have received methylprednisolone treatment show a greater recovery of motor and sensory function than do patients with similar injuries who have not received treatment. The success of methylprednisolone in treating spinal cord injuries has stimulated research to characterize all the damaging events that take place in the spinal cord in the days and weeks following an injury and to develop drugs that can block these events. Efforts to control the glial scar, to block inhibitory proteoglycans expressed in these regions, and to prevent the secondary cell damage resulting in cavitation are required to improve regeneration.

Transplantation

One important area of continuing research is the transplantation of different types of tissue and materials into the site of the spinal cord injury to connect the ends of the damaged spinal cord. These materials can produce an environment that is highly conducive to the growth of axons. Success has already been achieved in animals using transplants of nerves from the PNS and tissue from the spinal cords of embryonic animals. Tissue from the PNS is believed to be helpful because regeneration of nerves does occur in this part of the nervous system. Embryonic tissue may promote growth because it is still developing and may provide a supportive environment for growing nerve fibers. Transplantation technology has been advancing at a rapid rate, and many different techniques are currently being optimized for use in spinal cord injuries. To overcome the cavitation resulting from tissue necrosis, Schwann cell tissue, olfactory ensheathing (glial) cells or tanycytes have been successfully grafted. These transplanted bridging tissues not only provide a permissive pathway to nerve regeneration, but could also represent sources of trophic support. Schwann cells use cell adhesion molecules and tight junctions to provide morphological stabilization of the contact with the elongating axon, as well as small-scale gap junctions to facilitate traffic of substances between them.

Further Reading

Bandtlow CE and Schwab ME (2000) NI-35/250/nogo-a: a neurite growth inhibitor restricting structural plasticity and regeneration of nerve fibers in the adult vertebrate CNS. *Glia* **29**(2): 175–181.

Brittis PA and Flanagan JG (2001) Nogo domains and a Nogo receptor: implications for axon regeneration. *Neuron* **30**(1): 11–14.

Fournier AE and Strittmatter SM (2001) Repulsive factors and axon regeneration in the CNS. *Current Opinion in Neurobiology* **11**(1): 89–94.

Fu SY and Gordon T (1997) The cellular and molecular basis of peripheral nerve regeneration. *Molecular Neurobiology* **14**(1–2): 67–116.

Jones LL, Oudega M, Bunge MB and Tuszynski MH (2001) Neurotrophic factors, cellular bridges and gene therapy for spinal cord injury. *Journal of Physiology* **533**: 83–89.

Kury P, Stoll G and Muller HW (2001) Molecular mechanisms of cellular interactions in peripheral nerve regeneration. *Current Opinion in Neurology* **5**: 635–639.

McGraw J, Hiebert GW and Steeves JD (2001) Modulating astrogliosis after neurotrauma. *Journal of Neuroscience Research* **63**(2): 109–115.

Qiu J, Cai D and Filbin MT (2000) Glial inhibition of nerve regeneration in the mature mammalian CNS. *Glia* **29**(2): 166–174.

Raisman G (2001) Olfactory ensheathing cells – another miracle cure for spinal cord injury? *Nature Reviews Neuroscience* **2**(5): 369–375 [review].

Ramon-Cueto A (2000) Olfactory ensheathing glia transplantation into the injured spinal cord. *Progress in Brain Research* **128**: 265–272.

Neural Transplantation

Introductory article

Stephen B Dunnett, Cardiff University, Cardiff, Wales, UK

CONTENTS

Neural transplantation is the transplantation of nerve cells and tissues into the brain and spinal cord. First developed as an experimental strategy to study development and plasticity in the brain, it has led to new strategies to treat injury or disease throughout the nervous system.

INTRODUCTION

For fifty years after the publication of *Degeneration and Regeneration of the Nervous System* by Ramóny Cajal, neurology was dominated by the widespread belief that all regeneration was 'abortive' in the mature central nervous system (CNS), with the consequence that early attempts to transplant nerve cells in the brain were widely ignored – even when successful. The turning point came in 1969 and 1970 with two seminal studies: the demonstration by Geoff Raisman in London that adult axons can sprout to reinnervate vacated synapses, and a second by Lars Olson and colleagues in Stockholm that neuroendocrine cells from the adrenal gland and developing neurons of the embryonic brain can survive transplantation into the adult eye. These studies confirmed that growth and plasticity are indeed possible in the mature CNS, and quickly led to the discovery of new methods for transplantation of neurons in the adult mammalian brain, leading to novel strategies for cellular repair of damage in the brain and spinal cord associated with a variety of forms of trauma, injury, and neurodegenerative diseases. (*See* **Neural Development**; **Neural Degeneration**; **Neural Regeneration**; **Brain Damage, Treatment and Recovery from**)

NEURAL TRANSPLANTATION

Methodology for Fetal Cells

For transplantation of CNS neurons it is necessary to employ donor tissues harvested from embryos at the time of initial formation of those cells in early development. A variety of methods are available (Figure 1), including implantation as solid pieces of tissue into natural cavities of the brain (such as the ventricles), or injection of the tissue in fragments or as suspensions of dissociated cells directly into the host brain tissue. Provided the critical developmental time window is selected and that the grafts are positioned in a site where they can acquire the necessary nutrients by access to the host blood circulation, transplantation of embryonic neuron into the brain and spinal cord can be achieved reliably and reproducibly, whether the recipient is immature, adult, or aging.

Peripheral Tissues

In contrast to CNS neurons, glial cells and peripheral neurons (which continue to divide throughout life) can survive transplantation from neonatal and adult donors as well as from embryos.

Immunology

The brain has been found to be an immunologically privileged site. Neural tissues exhibit relatively low immunogenicity. Moreover, in contrast to other organs and tissues of the body, allografts (i.e. from the same species) implanted into the brain are protected behind the blood–brain barrier and are not recognized as foreign by the host immune system. However, once species boundaries are crossed, xenografts (i.e. from different species) are much more susceptible to rejection but can still be protected using immunosuppressant drugs.

FUNCTIONAL NEURAL TRANSPLANTATION

Neuroendocrine Grafts

The first systematic study of functional recovery following transplantation into the brain was

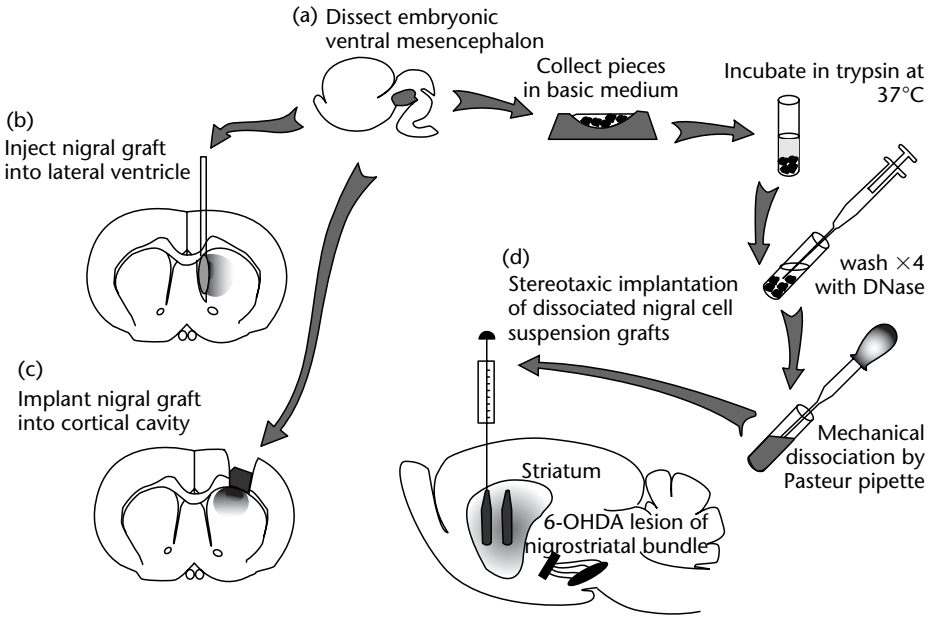


Figure 1. Methods of cell transplantation. (a) Dissection of the ventral mesencephalon, containing dopamine cells, from the embryonic brain. (b) Implantation of solid graft tissue into the lateral ventricle. (c) Implantation of solid graft tissue into an artificial cortical cavity. (d) Preparation and injection of dissociated tissue suspension into brain parenchyma. Abbreviations: 6-OHDA, the dopamine-specific neurotoxin, 6-hydroxydopamine; DNase, deoxyribonuclease, to inhibit reaggregation.

achieved in the 1960s, using neuroendocrine tissues to alleviate pancreatic dysfunction in rats. Subsequently, similar studies used hypothalamic tissues to alleviate a variety of neuroendocrine deficits associated with genetic and lesion-induced dysfunction in hypothalamic hormonal control, including loss of gonadotrophins controlling sexual maturation and behavior, diabetic symptoms associated with loss of antidiuretic hormone neurons, and disrupted circadian rhythms associated with mutation or degeneration of the suprachiasmatic nucleus. Although these grafts could in principle alleviate deficits by diffuse secretion of the missing neurohormones, in fact it has generally been found that the transplanted neurons provide a more complete repair by integrating into the host brain and extending axons to make specialized contacts with the appropriate host capillary circulation to restore normal neuroendocrine control.

Dopaminergic Grafts

A second main area of investigation has been the recovery of motor deficits following transplantation into the basal ganglia in animal models of Parkinson disease. This disease is characterized by degeneration of dopamine neurons of the substantia nigra and an associated loss of dopaminergic innervation in the neostriatum and other nuclei of the basal ganglia, resulting in impairments of voluntary motor control. A comparable syndrome, involving akinesia and impairments in the initiation of voluntary movements, can be reproduced in experimental animals by selective toxins. (*See* **Parkinson Disease**)

Many of these symptoms are alleviated by transplantation of embryonic dopamine neurons into the striatum, the primary area of terminal loss. In contrast, implants in the substantia nigra itself survive but have no functional benefit, indicating that recovery is dependent upon restitution of an appropriate dopaminergic innervation and reactivation in the striatum itself. Recovery of parkinsonian symptoms with embryonic nigral grafts, first demonstrated in rats, has been replicated in other species including mice, rats, and monkeys, laying the foundation for clinical trials in human patients.

CLINICAL APPLICATIONS IN PARKINSON DISEASE

Adrenal Grafts

The first neural transplants in humans were undertaken in the early 1980s, by grafting adrenal medullary tissues in patients with Parkinson disease (PD).

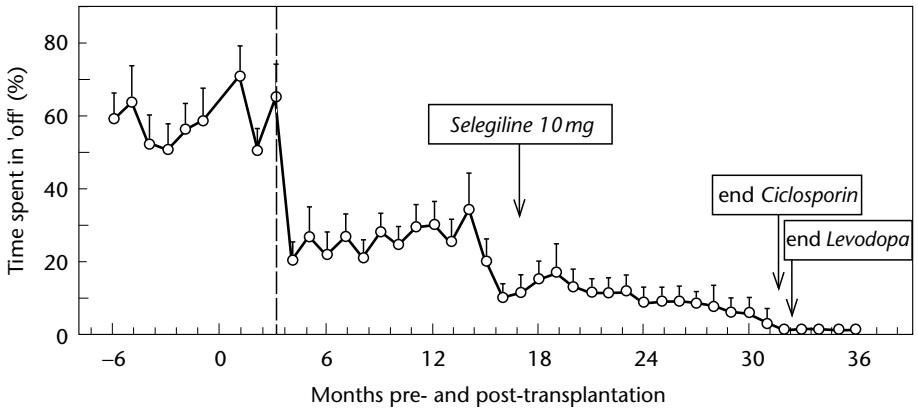


Figure 2. Progress of neurological symptoms in a patient with Parkinson disease, from the University of Lund series in Sweden. The 'off' period is the percentage of time during each day when the patient is suffering full parkinsonian symptoms without alleviation by the administered levodopa. Data kindly provided by Professor Olle Lindvall.

This tissue was selected as a peripheral source of catecholamine neurons which could be harvested from the patients themselves, circumventing both the poorly understood issues surrounding immunological compatibility and the ethical issues associated with using fetal tissues. However, the grafts did not survive well and resulted in significant adverse morbidity and mortality for limited clinical benefit. Therefore, after a flurry of interest and several hundred operations worldwide, a consensus was reached by the early 1990s that use of this source of tissue for transplantation in PD was not warranted.

Fetal Tissues

Since biological factors favored the efficacy of primary fetal tissues for transplantation, the ethical use of aborted human fetal tissues for research and therapy has been and is still much debated. This debate has resulted in acceptance in many (but not all) Western countries that it is ethical to use fetal tissues for developing new therapies, subject to a variety of regulations and guidelines to ensure that the decision and conduct of the abortion is not influenced by the subsequent use of the tissue.

Nigral Grafts in PD

Since the first operations in 1988, several hundred further patients with PD have now received transplants of fetal dopamine cells. The procedure is both safe and effective provided that the tissue collection and surgical implantation are firmly based on strict biological principles. Thus, systematic longitudinal follow-up is now available from several centers, and shows that patients can exhibit a significant alleviation of neurological signs, improved manual dexterity and control, reduced dependence on drugs, and an improved efficacy and reduced side-effects from the drugs that are taken (Figure 2). These effects take 1–3 years to develop and stabilize, which corresponds to the time required for normalization of the images in positron emission tomography scans of a functional dopamine system. The benefits can last for at least 11 years, the longest period studied. In contrast, other centers that have diverged from experimentally validated methods – whether for ease of tissue preparation, availability, or surgical convenience – have generally achieved poorer results, and in a few cases unacceptable side-effects. Neural transplantation is therefore now demonstrated to be effective in some patients, but efficacy is highly dependent on the procedures adopted.

Improving Graft Viability

The biggest technical problem facing clinical neurotransplantation is the limited availability of suitable fetal tissues. This problem is exacerbated by the fact that, at best, only 5–10% of the dopamine cells survive transplantation, requiring three to seven donor fetuses to be used for each hemisphere of the brain, for maximum efficacy. Consequently, there is currently a major research effort into discovering ways of improving graft survival, including treating the cells with growth factors, antioxidants, anti cell-death proteins and other neuroprotective agents. A combination of treatments can now more than double cell survival, but methods are still far from optimal.

ALTERNATIVE SOURCES OF CELLS

Even if 100% cell survival can be achieved, there will always be ethical, practical, and safety concerns limiting the availability of primary fetal cells for transplantation. Consequently a major topic for research is to identify alternative sources of cells that can be available 'off the shelf' for transplantation, in unlimited supply, properly standardized and fully characterized for safety. A number of alternatives have been suggested.

Stem and Other Precursor Cells

Stem cells are early precursor cells that have the capacity to divide exponentially to generate multiple duplicate copies of themselves as well as to differentiate down diverse lineages to produce a multiplicity of different cell types. Stem cells have been isolated from adult as well as developing brains, massively expanded, and differentiated to form large samples of both neurons and glia. The difficulty with applying this technology for cell therapy is that it is not yet possible to control precisely the fate of the progeny, and the majority of expanded cells when transplanted differentiate into immature neuroglia and astrocytes. The problem of understanding the signals that control differentiation of individual types of cells of therapeutic interest, such as developing and mature dopamine neurons, is potentially soluble but is not yet resolved.

Xenotransplantation

Grafting from other species would allow breeding of animals to provide well-defined cells and tissues for human transplantation. Pigs are widely considered the species of choice; they are of similar physical size and rate of development, there is considerable experience of pig husbandry, they can be bred under sterile conditions, and they produce regular large litters which can provide large quantities of fetal donor tissues. The two main problems relating to nigral xenotransplantation are that the methods for effective long-term immunoprotection across the species divide are still not well understood, even within an immunologically privileged site such as the brain, and there is a widespread safety concern about the theoretical risk of introducing new pig pathogens and viruses into the human population.

Cellular Gene Therapy

If we can understand the developmental signals that determine cell phenotype, then it should theoretically be possible to harvest cells from the patients themselves (thereby obviating any immunological problems) and engineer them to express the particular phenotypes required therapeutically. Although a variety of cells, such as fibroblasts, and cell lines have now been engineered to expressed particular components of a relevant phenotype (e.g. to synthesize dopamine), the requirements of developing cells to exhibit all the features of a developing neuron – to grow, extend neurites, seek appropriate targets, establish synaptic connections, transduce inputs, transmit electrochemical signals, and regulate synaptic release of transmitter-mediated outputs – remain poorly understood, and clinical applications are certainly not imminent.

Neuroprotection

Most transplantation therapies are based on the principle of cellular repair, as distinct from strategies based on delivery of neuroprotective molecules that might slow the course and progression of neurodegenerative disease. However, many of the neuroprotective molecules discovered in the 1990s – including a variety of growth factors, transcription factors, antioxidants, and antiapoptotic molecules – do not cross the blood–brain barrier, so targeted delivery to specific sites in the depth of the brain is proving to be an obstacle. Neural transplantation of cells engineered to express particular neuroprotective molecules, in particular when combined with molecular triggers for controlled release, provides one of the most likely prospects for providing targeted delivery for therapeutic application, although again a variety of technical and safety issues remain to be solved.

PROSPECTIVE APPLICATIONS IN OTHER NEURODEGENERATIVE DISEASES

Huntington Disease

Huntington disease is characterized by a selective degeneration of intrinsic striatal neurons. Following a similar strategy to that used in PD, striatal neurons can be transplanted into the patient's brain, and there is now preliminary evidence

from the first clinical trials that this is not only safe but can also alleviate cognitive as well as motor symptoms of the disease. (*See* **Huntington Disease**)

Stroke

Loss of blood supply in a circumscribed area of the brain leads to infarct and extensive cell loss at the focus. There have been claims that neuronal precursor cells can migrate to, and repopulate, the site of injury with functional benefit in rats, and this has already led to an initial clinical trial. The data remain controversial. (*See* **Stroke**)

Multiple Sclerosis

In multiple sclerosis the oligodendrocytes that myelinate long-distance axons of the brain and spinal cord are lost in 'plaques', the foci of sporadic acute inflammatory reactions. Experimental models of the focal plaque are available, and these can be effectively remyelinated by implants either of immature oligodendrocytes or their precursors. The problems yet to be solved for this to become a clinical therapy are how to stimulate migration of implanted oligodendrocyte precursors to widely scattered plaque sites; and second, how to inhibit the formation of new plaques which can rapidly lead to further deterioration in this disease. (*See* **Multiple Sclerosis**)

Spinal Cord Injury

It is now possible to reduce the formation and spread of spinal cord cysts using embryonic implants and to use a variety of cellular and matrix materials to bridge a site of transection injury. However, the main problem in achieving functional repair in spinal cord injury is how to stimulate extensive regrowth of cut long-distance sensory and motor axons through the host spinal cord and back to remote targets distal to the site of injury. Although a major area of investigation, the technical problems still to be overcome remain massive.

Further Reading

Barker RA and Dunnett SB (1999) *Neural Repair, Transplantation and Rehabilitation*. Hove, UK: Psychology Press.

Dunnett SB and Björklund A (1999) Parkinson's disease: prospects for novel restorative and neuroprotective treatments. *Nature* **399**: S32–S39.

Dunnett SB and Björklund A (2000) *Functional Neural Transplantation*, vol. II. *Novel Cell Therapies for CNS Disorders*. Amsterdam, Netherlands: Elsevier Science.

Dunnett SB, Björklund A and Lindvall O (2001) Cell therapy in Parkinson's disease – stop or go? *Nature Reviews, Neuroscience* **2**: 365–369.

Fawcett JW, Rosser AE and Dunnett SB (2001) *Brain Damage, Brain Repair*. Oxford, UK: Oxford University Press.

Freeman TB and Widner H (1998) *Cell Transplantation for Neurological Disorders*. Totowa, NJ: Humana.

Gage FH and Christen Y (1997) *Isolation, Characterization and Utilization of CNS Stem Cells*. Berlin, Germany: Springer-Verlag.

Lindvall O (1997) Neural transplantation: a hope for patients with Parkinson's disease? *NeuroReport* **8**(14): iii–x.

Lindvall O, Björklund A and Widner H (1991) *Intracerebral Transplantation in Movement Disorders*. Amsterdam, Netherlands: Elsevier.

Olanow CW, Kordower JH and Freeman TB (1996) Fetal nigral transplantation as a therapy for Parkinson's disease. *Trends in Neurosciences* **19**: 102–109.

Neurogenesis

Introductory article

Brandi K Ormerod, University of British Columbia, Vancouver, Canada
Erin M Falconer, University of British Columbia, Vancouver, Canada
Liisa A M Galea, University of British Columbia, Vancouver, Canada

CONTENTS

Adult neurogenesis has been found to occur in the olfactory bulb and the hippocampus of all vertebrate species studied but has been most extensively characterized in birds and mammals. Although its function is unknown, neurogenesis is influenced by physical activity, learning, environmental complexity, season and photoperiod, exposure to stress, and experimental brain pathology.

INTRODUCTION

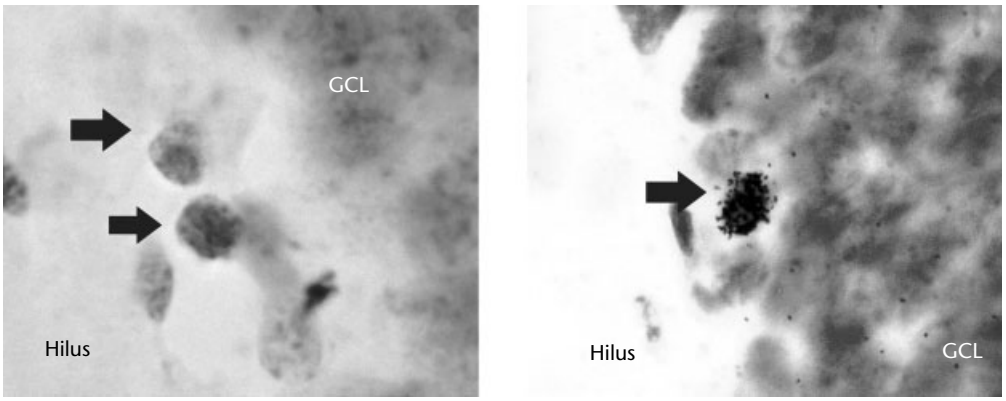
For many years, researchers thought that plasticity within the adult vertebrate brain consisted only of changes in synapse efficacy or number. However, research has now conclusively demonstrated that regions of the adult vertebrate brain retain the capacity to produce and/or incorporate new neurons. In avian species, new neurons are incorporated into the entire telencephalon, preferentially into the song-learning circuit and hippocampal complex. In mammals, new neurons are incorporated more discretely into the dentate gyrus of the hippocampus, olfactory bulbs, and perhaps the association areas of the nonhuman primate neocortex, albeit to a lesser extent.

Neurogenesis occurs when precursor cells divide, producing daughter cells that migrate to target locations where they differentiate into neurons. To evaluate neurogenesis in adult vertebrates, investigators capitalize on the fact that cells in the synthesis phase of mitosis will incorporate the traceable markers [^{3}H]thymidine or bromodeoxyuridine (BrdU; a thymidine analogue) into their deoxyribonucleic acid (DNA). Figure 1 shows examples of cells labeled in this way. The synthesizing precursor cells and their progeny (precursor cells, neurons, and glia) can be counted and the phenotype verified immunohistochemically by applying antibodies raised against the proteins specific for neurons versus glia. Using a combination of techniques, investigators have identified factors that influence adult neurogenesis by altering cell proliferation, migration, differentiation, and/or survival.

ADULT NEUROGENESIS IN BIRDS

Early work investigating song acquisition in canaries described the phenomenon that, although adult females typically do not sing, adult female canaries produce male-like song following androgen treatment. Along with the male-like song acquisition, the hyperstriatum ventrale, pars caudalis (or higher vocal center, HVc), the primary telencephalic song-control nucleus, doubles in size in adult females treated with androgens. Fernando Nottebohm and Steven Goldman proposed the hypothesis that the expansion in HVc volume following androgen treatment in adult female canaries reflected the addition of new neurons. After treating adult female canaries with exogenous testosterone and injecting the birds with [^{3}H]thymidine, they found that many HVc cells had recently undergone mitosis. Later studies indicated that these new cells responded to auditory stimuli, received synaptic input, and had neuronal morphology, suggesting that the new cells were incorporated into functional neural circuitry. As a cautionary note, the initial studies did not use immunohistochemical markers for neuron identification (because they were not yet available), but characterized the new cells as neurons using [^{3}H] thymidine, Nissl staining, and observations of ultrastructural characteristics such as dendrites and axonal hillock.

The discovery of neurogenesis in the HVc of songbirds, along with evidence that these neurons were responsive to auditory stimuli, prompted Barnea and Nottebohm to suggest that these new cells contributed to avian song learning. This suggestion



(a) (b)

Figure 1. (a) Photomicrograph of a cell labeled with bromodeoxyuridine in the subgranule zone of the dentate gyrus, having the morphology of an immature neuron of an adult rat. (b) Photomicrograph of a [^{3}H]thymidine-labeled cell in the subgranule zone of the dentate gyrus in an adult female meadow vole. GCL, granule cell layer.

was supported by studies identifying that juvenile male zebra finches, during a sensitive period for song acquisition, add more neurons to the HVc than do juvenile females, which do not learn to sing. Further, in adult canaries there is a seasonal change in the expression of new song patterns that corresponds with a seasonal change in the number of neurons in the HVc.

Neurogenesis in the adult bird is not limited to the HVc. Indeed, neurogenesis exists in the ventricular zone in adult birds that do not learn song in adulthood (such as the chicken, quail, parakeet, and dove), suggesting that this phenomenon may be widespread in the brain across different bird species. Barnea and Nottebohm found that new cells are produced in the hippocampus of wild blackcapped chickadees in a seasonally dependent manner, with more cells produced in the autumn than in spring. Interestingly, during the autumn these birds store food in unique cache sites, never reuse a cache site, and can take up to 4 weeks to retrieve food from a particular cache site, suggesting that they need a large capacity for spatial memory in the autumn and winter.

The study of adult avian neurogenesis, coupled with earlier reports of new cells in the hippocampus of adult rats, prompted further study of hippocampal and olfactory bulb neurogenesis in mammals.

ADULT NEUROGENESIS IN THE MAMMALIAN HIPPOCAMPUS

Joseph Altman described neurogenesis within the dentate gyrus of adult rats and cats in the 1960s, but few studies investigated adult neurogenesis following Altman's discovery. Early researchers were skeptical that the newborn hippocampal cells described by Altman could be neurons; perhaps the most obvious reason for this was that immunohistochemical markers that could positively identify new cells as neurons had not yet been developed.

Renewed interest in adult neurogenesis followed Fernando Nottebohm and colleagues' report in the early 1980s of season-dependent fluctuations of neuron production within the adult canary song circuit, described above. However, hope that the adult primate brain could produce and incorporate new neurons was diminished when Pasko Rakic reported that although new cells are produced within the adult nonhuman primate hippocampus, they could not be definitively identified as neurons.

In the early 1990s, Heather Cameron and colleagues used immunohistochemical markers to establish that many new neurons are produced within the dentate gyrus of the adult rat. At about the same time, Mark West and colleagues developed an unbiased stereological technique to estimate the total number of new neurons in the granule cell layer of the dentate gyrus. In fact, work combining immunohistochemical and stereological methods suggests that as many as 4000 new neurons could be produced in the dentate gyrus of untreated adult laboratory rats following 2 h of proliferation.

A team headed by Elizabeth Gould, Bruce McEwen and Eberhard Fuchs in 1997 provided the first hint that neurogenesis might occur within the primate hippocampus. They discovered new neurons in the dentate gyrus of the adult tree shrew, an insectivore phylogenetically related to primates. Then Gould and colleagues reported neurogenesis in the hippocampus of adult

marmoset monkeys and, concurrently with David Kornack and Pasko Rakic, the dentate gyrus of adult macaque monkeys.

Finally, a study headed by Peter Eriksson and Fred Gage confirmed the presence of neurogenesis within the adult human dentate gyrus. They found evidence of new neurons in the brain tissue of human patients who had received the DNA marker BrdU up to 2 years prior to succumbing to cancer. This study confirmed that investigating the factors that influence the proliferation, migration, differentiation, and survival of new neurons could have profound relevance for adult human brain trauma or disease.

Neurogenesis in the Adult Mammalian Brain

Although neurogenesis in adult mammals is characterized most extensively in the hippocampus, studies have shown that the olfactory bulbs of mammals and perhaps the neocortical association areas of nonhuman primates incorporate new neurons throughout adulthood. However, new neurons are probably produced in only two areas of the mammalian brain: the subventricular zone that lines the anterior lateral ventricles, and the subgranular zone that borders the granule cell layer and hilus of the dentate gyrus. Because precursor cells derived from many areas of the adult vertebrate brain will divide *in vitro*, the subventricular zone and subgranular zone are probably the only areas of the adult brain to retain an environment that permits cell proliferation postnatally.

The daughter neuroblasts (immature neurons) of precursor cells located in the mammalian subventricular zone migrate tangentially into the olfactory bulbs and possibly into association areas of the neocortex. Upon reaching their target location in the olfactory bulb, the neuroblasts differentiate into glutamatergic, GABAergic or tyrosine hydroxylase-secreting neurons. The daughter neuroblasts of precursor cells located in the subgranular zone migrate, probably along radial glia, into the granular cell layer where they differentiate into glutamatergic granule neurons. Some evidence suggests that as granule neurons age they migrate deeper into the granule cell layer where they may eventually die and be replaced by new neurons. Many new cells are born daily but studies have shown that the total number of labeled cells diminishes between 1 week and 3 weeks, because many new cells die before becoming incorporated into the granule cell layer.

Cell Development in the Dentate Gyrus

The majority of labeled cells in the dentate gyrus of adult mammals become neurons. For example, labeled neuroblasts migrate into the granule cell layer where they acquire a granule neuron morphology. Like other neurons, they receive presynaptic input and extend an axon into the mossy fiber pathway to target CA3 pyramidal neurons. Antibodies raised against proteins expressed specifically by neurons during various stages of maturation can be used to further verify the phenotype of newborn cells. For example, in adult mammals the protein TUC-4 is expressed by precursor cells undergoing division and by immature neurons in the process of axon extension. Because axon extension is probably completed by 4–10 days after birth, anti-TUC-4 can be used to verify the phenotype of new immature neurons. In contrast, neuron-specific enolase is expressed by adult-generated neurons 2–3 weeks after division, presumably when the new neuron establishes a synapse with a CA3 pyramidal cell during development. New granule neurons produced within the dentate gyrus are likely to be functional, as they demonstrate paired-pulse facilitation similar to mature granule neurons but – interestingly – are more plastic electrophysiologically than mature granule neurons.

PUTATIVE MECHANISMS AND FUNCTIONS

Many factors regulate the different facets of neurogenesis (cell proliferation, migration, differentiation, and survival) within the adult vertebrate brain, suggesting that their mechanisms are complex (Table 1). To understand the mechanisms and function of adult neurogenesis, researchers must delineate the factors that affect adult neurogenesis at each stage of maturation (i.e., cell proliferation versus cell survival). For example, a treatment administered prior to or during the uptake of a DNA marker will affect cell proliferation (brains are examined within 24 h), but a treatment administered during migration, differentiation, or maturation will affect cell survival (brains are examined between 24 h and 3 weeks). Cell survival can be influenced independently of cell proliferation: for example, learning a hippocampus-dependent task enhances the number of new neurons surviving without influencing cell proliferation. Factors that increase the number of new neurons in existing circuitry would probably more readily influence behavior relative to factors that increase cell

Table 1. Factors affecting cell proliferation and granule cell survival

	Cell proliferation	*Cell survival*
Adrenal steroids	↓	No change
Adrenalectomy	↑	↑
NMDA	↓	
NMDA-R antagonist	↑	↑
Insulin growth factor	↑	↑
Testosterone (canaries)		↑
High levels of estradiol (rats, voles, canaries)	↑	↑ (canaries)
Serotonin (rats)	↑	
Aging (rats)	↓	Possibly ↓
Hippocampal-dependent learning (rats)	No change	↑
Short photoperiod (voles, hamsters, gray squirrels, chickadees, canaries)	↑ (no change in gray squirrels)	↑ (no change in gray squirrels)
Rearing in enriched environment (mice)	No change	↑
Lesions to dentate gyrus	↑	↑
Kindling (rats)	↑	No change
Running (mice)	↑	↑
Exposure to stress (rats, tree shrews, marmoset monkeys)	↓	Possibly ↓

NMDA, *N*-methyl-D-aspartate; NMDA-R, NMDA receptor.

proliferation, owing to their immediate potential to influence circuitry.

Regulation of Neurogenesis in Birds

Although neurogenesis occurs in other areas of the avian brain, the HVc has been the site most studied in songbirds. In male canaries, most neurons outside the telencephalon are born before hatching, whereas new neurons continue to be added to the telencephalon during adulthood. New neurons found in the adult bird HVc are derived from precursor cells in the walls of the lateral ventricles.

The newly proliferated cells remain in the subventricular/ventricular zone for at least 4 days before migrating into the forebrain parenchyma. Studies *in vitro* have identified factors that might regulate the time taken before the new cell initiates migration (one should recognize that mechanisms *in vivo* might differ). The cell adhesion molecule N-cadherin has been implicated as a factor that might restrict migration, and factors such as insulin growth factor (IGF) increase the number of neurons that migrate. The new cells migrate over distances of up to several millimeters, aided initially by radial glia. About two-thirds of these new neurons die on their migratory route, close to the radial glia. Evidence suggests that estrogen aids the survival of these cells by enabling neural cell adhesion molecules (NCAM) to couple with calcium signaling pathways, and that this process may be necessary for cell survival during migration. After 30 days, these neurons are differentiated, form synaptic contacts, and respond to auditory stimuli with action potentials. Many of the new neurons project outside the HVc and connect to distant targets.

Factors other than estrogen have also been shown to influence the survival of new neurons in the avian brain. In the adult canary, testosterone induces an increase in HVc neuron number and HVc volume, and it is thought that testosterone exerts its effects through brain-derived neurotrophic factor, a neurotrophin that affects the survival of neurons in the developing and adult brain. Other neurotrophins may also play a part in neuron survival.

Potential Function of Adult Neurogenesis in Birds

The existence of neurogenesis in the song nuclei of adult songbirds suggests that the new cells may regulate song learning. As mentioned above, neurogenesis in the songbird HVc varies seasonally along with testosterone level. During the autumn there is an increase in cell numbers, commensurate with the time when songbirds modify their song patterns. It has been suggested that changes in neurogenesis may have a role in modifying song control circuitry to allow plasticity in song behavior. In adult male white-crowned sparrows, an increase in HVc neurogenesis is accompanied by an increase in the development of song stereotypy. However, in songbirds that retain the same song for life (e.g., the song sparrow), there is still a seasonal fluctuation in neuron density and number in the HVc. Although the seasonal difference in neuron number and density in the song sparrow is not as large as that found in the canary, this evidence suggests that neurogenesis in the HVc may serve a function other than song learning and this function has yet to be determined.

An area of the avian brain not involved in song production, the hippocampus, also incorporates new neurons in adulthood. In birds, differences in hippocampal density and cell number are seen between species that store food and those that do not, with food-storers showing increased hippocampal size and neuron number. David Sherry

and colleagues have found that food-storers (but not non-storers) have a seasonal change in the rate of new cell birth. Hippocampal cell survival peaks at the time when the blackcapped chickadee engages in the most food-storing behavior. When given experience in foodstoring, juvenile laboratory marsh tits increase the number of neurons and size of the hippocampus compared with birds deprived of food-storing experience. Such experience-dependent neurogenesis suggests that neurogenesis in the hippocampus may have a role in spatial behavior, as behaviors such as food-storing are dependent on the integrity of the hippocampus.

Regulation of Neurogenesis in the Adult Mammalian Hippocampus

Corticosterone and glutamate

Many factors have been shown to suppress cell proliferation within the dentate gyrus of adult rodents. The first inhibitory factors discovered were adrenal steroids, via corticosterone, and glutamate, via *N*-methyl-D-aspartate receptor (NMDA-R) activation. While high levels of either corticosterone or NMDA-R activation suppress cell proliferation, low levels increase cell proliferation. In fact, corticosterone and glutamate coregulate cell proliferation within the dentate gyrus of adult rodents. Specifically, NMDA-R activation prevents adrenalectomy-induced increases in cell proliferation, and NMDA-R inactivation blocks corticosterone-induced suppression of cell proliferation. However, corticosterone and glutamate indirectly coregulate cell proliferation as precursor cells do not express the NMDA-R NR1 subunit and express type I and type II glucocorticoid receptors 24 h and 2 weeks after cell division, respectively.

Corticosterone may also inhibit cell migration within the dentate gyrus of adult rats. Polysialylated neural cell adhesion molecule (PSA-NCAM) is only expressed in the dentate gyrus and olfactory regions of adult mammals by immature neurons in the process of migration and differentiation. Administering corticosterone can reverse adrenalectomy-induced increases in PSA-NCAM expression in the young adult rat.

Estradiol

Exposure to high levels of estradiol for 4 h enhances cell proliferation, but 48 h after this exposure there is a suppression in cell proliferation within the dentate gyrus of adult female meadow voles. The effect of estradiol on cell proliferation is likely to be time-dependent within the dentate gyrus of adult female rats. Female rats exposed to high endogenous levels of estradiol (proestrus) 7 h prior to labeling have higher rates of cell proliferation than rats exposed to a high level of estradiol 30 h prior to labeling (estrous or diestrous). The effect of estradiol on cell proliferation may be indirect, as traditional estrogen receptors are not expressed in the precursor cells within the dentate gyrus.

Dopamine, serotonin, and growth factors

While dopamine has been shown to suppress cell proliferation, serotonin and growth factors have been shown to enhance cell proliferation within the dentate gyrus of adult rodents. When D2 dopamine receptors are inactivated the rate of cell proliferation increases, but when dopamine synapses are activated the rate of cell proliferation decreases in the dentate gyrus of adult gerbils, relative to animals that receive an injection of saline.

When the dorsal raphe nucleus, the main serotonin-mediated input into the dentate gyrus, is removed, the rate of cell proliferation in the dentate gyrus of adult rats decreases. Similarly, when serotonin (5-hydroxytryptamine, 5-HT) is inactivated by an antagonist, cell proliferation is diminished. Alternatively, cell proliferation is increased when 5-HT release is enhanced by 5-HT agonists. The effect of 5-HT may be mediated by the 5-HT_{1A} receptor, because 8-hydroxy-2(di-n-propylamino) tetralin (8-OH-DPAT, a 5-HT_{1A} receptor agonist) increases cell proliferation in the dentate gyrus of adult rats. Serotonin may also affect the migration of new cells into the granule cell layer. When the dorsal raphe nucleus is removed, PSA-NCAM expression decreases in the dentate gyrus of adult rats.

Growth factor peptides are expressed largely during development. When adult rats are given multiple injections of insulin-like growth factor I, cell proliferation is increased and, furthermore, a greater percentage of new cells differentiate into neurons than in rats treated with vehicle only. Other growth factors, therefore, may affect both cell proliferation and cell differentiation in the dentate gyrus of adult rats.

Signals from dying cells

Some researchers have postulated that dying cells trigger cell birth within the adult mammalian hippocampus. For example, high levels of corticosterone reduce both cell birth and cell death within the granule cell layer. Furthermore, the degree of unilateral lesion-induced damage to the granule cell layer is related to the degree of proliferating cells on the same side. The number of new cells

produced in the dentate gyrus can be increased by experimentally induced ischemia (which prevents oxygen flow to the brain) in rats. Longer bouts of ischemia, which increase the level of cell death, produce higher rates of cell proliferation.

Possible Function of Neurogenesis in the Adult Mammalian Hippocampus

Stress

Separate studies have shown that cell proliferation in the dentate gyrus of adult rats, tree shrews, and marmoset monkeys is suppressed by exposure to an ecologically relevant stressor. Rats exposed to predator odor, subordinate tree shrews, and 'intruder' marmoset monkeys, all show a rapid suppression of cell proliferation within the dentate gyrus. It is not clear what long-term effect, if any, an acute exposure to stress has on the animal. Exposure to a stressful event elevates a number of factors that have been shown to suppress cell proliferation within the dentate gyrus of adult rats (such as adrenal steroids and glutamate). Preliminary evidence has shown that scopolamine (an acetylcholine muscarinic receptor antagonist) blocks the suppression of cell proliferation induced by predator odor in the dentate gyrus of adult rats.

Aging

Both the rate of cell proliferation and the percentage of new cells that differentiate into neurons are reduced within the dentate gyrus of senescent (21 months old) adult rats relative to young (6 months old) adult rats. Levels of PSA-NCAM are also reduced in the dentate gyrus of senescent rats, suggesting that the reduced percentage of new neurons observed might reflect a diminished ability of new immature neurons to migrate into the granule cell layer.

Kindling

Repeated low-intensity electrical stimulation of various brain areas can produce progressively intense seizure activity in a process called 'kindling'. Kindling to a moderate seizure intensity (class 5) is accompanied by a drastic increase in cell proliferation within the dentate gyrus of adult rats. The mechanism mediating amygdala kindling-induced increases in cell proliferation may become desensitized. For example, if BrdU is administered before each kindling session, increases in cell proliferation can be observed following four class 5 seizures, but if BrdU is administered after kindling sessions have ended, increases in cell proliferation can only be observed following nine (and not four) class 5 seizures. Alternatively, the process of kindling could enhance the survival of new neurons.

Season

Within the dentate gyrus of adult female meadow voles, cell proliferation varies seasonally. During the non-breeding season cell proliferation is elevated, concurrent with an increase in territory size and an enhancement in spatial performance in the laboratory. However, the same relationship between increased territory size and increased cell proliferation is not seen in male meadow voles, indicating that female brains may be more plastic than male brains. Seasonal changes in cell proliferation have also been reported in the adult male golden hamster, with an enhancement in cell proliferation during the short photoperiod relative to the long photoperiod. However, no seasonal change in cell proliferation in wild adult gray squirrels of various ages has been demonstrated, despite the fact that these animals are known to cache food during the autumn. This suggests that the mechanisms and factors controlling cell proliferation may not be conserved across all species.

Hippocampus-dependent learning

Successful performance in the Morris water-maze task relies upon the integrity of the hippocampus. In this task, rats use information in the environment to locate a platform hidden beneath water made opaque by nontoxic paint or milk; removal of the hippocampus impairs the rats' ability to find the platform. When training on this task commences a week after BrdU has been administered (at a time when these new neurons are in the process of extending dendrites and an axon), more labeled neurons are found within the dentate gyrus compared with rats trained on tasks not involving the hippocampus. However, training does not enhance the survival of 1-day neurons or neurons that are more than 12 days old. In fact, new neurons may only be susceptible to the survival-enhancing effects of hippocampus-dependent learning when they are immature (about 7 days old) and in the process of dendrite and axon extension.

Experience

Mice living in an enriched environment are exposed to social (cage partners) and inanimate (running wheel, tunnels and toys) stimulation. Such mice have more new neurons (but no change in cell proliferation) in the dentate gyrus than mice living in standard laboratory conditions. This

finding is interesting because the hippocampus appears to have a role in rats' abilities to learn about different contexts or environments. In addition, mice given access to a running wheel have enhanced cell proliferation, better performance in the Morris water maze, and increased long-term potentiation (a putative cellular model of learning) relative to mice without access to a running wheel.

CONCLUSION

New cells are produced within the subventricular/ventricular zone of birds and mammals and the subgranular zone of mammals. Many of the new cells are incorporated into the song-learning circuit and hippocampus of birds, and into the hippocampus and olfactory bulbs of mammals. Many factors have been shown to influence cell proliferation and/or cell survival in both birds and mammals. Although the function of new neurons is unknown, adult neurogenesis in the dentate gyrus is modified by hippocampus-dependent behavior. However, manipulations that enhance cell proliferation in the hippocampus of adult rodents (e.g., adrenalectomy, NMDA-R antagonist administration, hippocampal lesions, and kindling) also impair hippocampus-dependent learning. In contrast, factors that enhance cell survival, such as housing in a complex environment, generally enhance hippocampus-dependent learning. Thus, the challenge for future research is to determine not only which factors affect adult neurogenesis but also how they do so, as well as the consequences of alterations in adult neurogenesis on behavior.

Further Reading

Alvarez-Buylla A and Kirn JR (1997) Birth, migration, incorporation and death of vocal control neurons in adult songbirds. *Journal of Neurobiology* **33**: 585–601.

Cameron HA and McKay RDG (1998) Stem cells and neurogenesis in the adult brain. *Current Opinion in Neurobiology* **8**: 677–680.

Goldman SA (1998) Adult neurogenesis: from canaries to clinic. *Journal of Neurobiology* **36**: 267–286.

Gould E and Cameron HA (1996) Regulation of neuronal birth, migration and death in the rat dentate gyrus. *Developmental Neuroscience* **18**: 22–35.

Kempermann G and Gage FH (1999) New nerve cells for the adult brain. *Scientific American* **May**.

Neurogenetic Approaches

Introductory article

Hans-Peter Lipp, University of Zürich, Zürich, Switzerland
David P Wolfer, University of Zürich, Zürich, Switzerland

CONTENTS

The relation between genes and cognition is analyzed using two approaches: identification of genetic traits and statistical mapping to chromosomes, and by studying the effects of targeted deletion of genes on brain and behavior.

INTRODUCTION

Does heredity control cognition and intelligence, and if so, to what extent? This question has been a conceptual battleground for two hundred years. The debate is not only a scientific one but carries implications for social sciences, education, and even politics. It can be traced to two concepts of individuality.

In the older and mechanistic view of the late sixteenth and the seventeenth centuries, humans and animals were considered as invariant prototypes created by god or nature. Thus, differences between individuals were considered as random deviations from a predefined norm. This interpretation of individual variation is still prevalent in engineering, where individuality is considered as error.

A biological view of mental processes was introduced in the early nineteenth century as phrenology (initially the science of mind) by Gall, and advanced further by Darwin. Both of them focused on individuality as starting point. Gall assumed that behavioral individuality of animals and humans was rooted in different volumes and connections of brain modules, and Darwin offered a functional role for such biological variation, namely as a substrate for natural selection. With the discovery of genes as units of hereditary transmission and the acceptance of evolutionary theory, the study of heredity of mental traits became the focus of a subspecialty of genetic research known as behavior genetics. This individualistic approach is merging today with the expansion of mechanistic molecular techniques into the fields of cognition and learning. (*See* **Concepts, Philosophical Issues about**; **Epistemology**; **Scientific Thinking**; **Evolutionary Psychology: Applications and Criticisms**)

Correspondingly, there are two conceptual approaches to study the role of genes in cognition. One, the top-to-bottom approach, searches for individual differences in talents and temperaments, and tries to identify whether they have a genetic base, and, eventually, to identify the genes involved. The other approach starts with defined mutations of genes that either occur naturally, or were induced experimentally, and studies changes in higher brain functions that can be attributed to the inactivated gene.

FROM GENE TO COGNITION: THE ROLE OF BRAIN DEVELOPMENT

In order to understand the relations between genes and cognition, one must consider how genes interact with the developing brain, and how they might specifically affect cognition. This is most easily achieved if one assumes that the brain forms a hierarchy of systems (Figure 1): the lowest level is formed by the spinal cord and brain systems which represent several input channels for sensory information, and the only output channel, the motor system. These channels are controlled and modulated by a superimposed hierarchy of neural systems that can be compared with a bureaucracy. The topmost levels with respect to cognition are the frontal, parietal, and temporal parts of the cerebral cortex, the ensemble known as associational cortex. This cerebral system hierarchy has one goal, keeping sensory and motor systems finely tuned under all possible conditions, even in the presence of disturbing factors such as mutated genes or brain diseases. Such adaptability is known as 'brain plasticity' and is the main mechanism in masking gene effects.

The second point to keep in mind is brain development. It starts early in embryogenesis, but takes

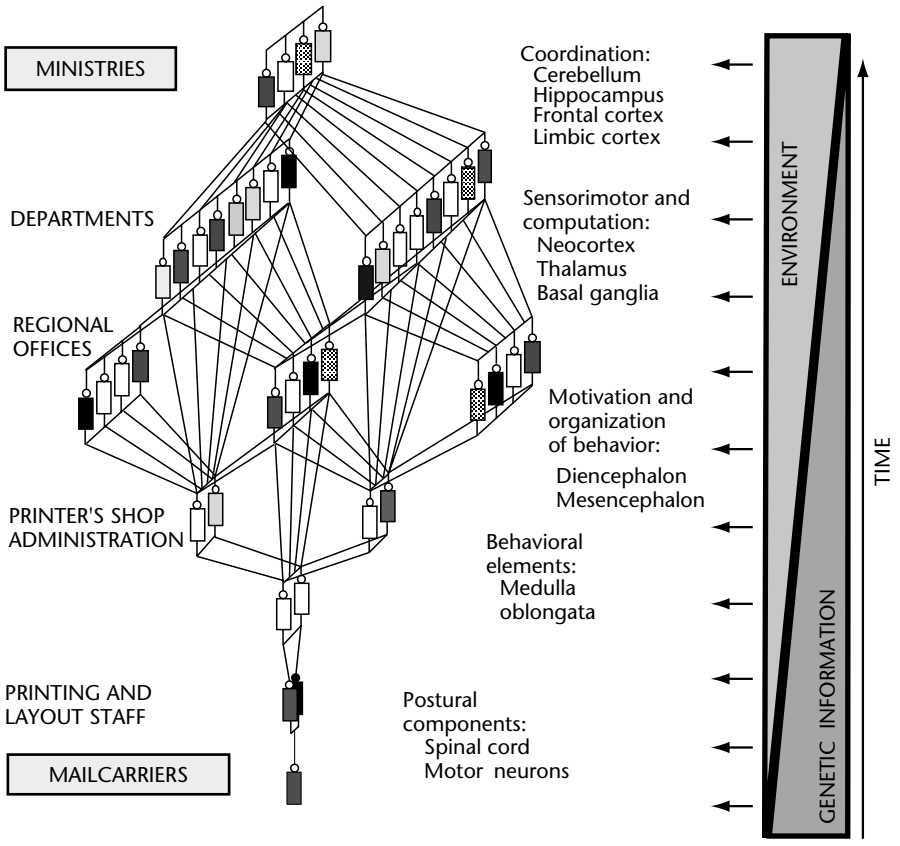


Figure 1. The brain as a cerebral bureaucracy which can be targeted by single gene effects. The relations between system levels and anatomical systems are arbitrary. The penetrance (visibility and predictability) of a mutation is low when it affects the middle levels of the bureaucracy because of homeostasis and other buffer mechanisms. It is high if the final common pathway is affected, but will appear as a mixture of neurological and behavioral deficits that can barely be disentangled. Mutations affecting the top layers are likely to be specific for cognitive processes, but will tend to have discrete effects that will become visible only in a test challenging the affected system. System specificity of gene effects can be obtained by proper timing (see time arrow at right). Early-acting genes encode primarily the layout of body and brain, and depend little on environmental factors. Genes with late activation have no other targets than the late differentiating coordination systems of the brain, and their effects are coincident with the action of environmental factors. Thus, developmental syndromes affecting behavior may have both a genetic and an environmental origin, but the two are not necessarily intertwined.

considerable time. There is always a postnatal phase of brain development and maturation which extends until puberty or even beyond, depending on the species. Mammalian brain development is characterized by dynamic plasticity. Developmental inaccuracies or even insults are often compensated for by rearrangement of connections and compensatory growth. This represents a second factor masking genetic influences. (*See* **Neural Development**; **Neural Development, Models of**; **Reorganization of the Brain**)

How could a particular mutation have a predictable impact on cognition, despite such powerful masking factors? The easiest way would be by timing. The neural system hierarchy is built up progressively: simpler systems form first, complex ones later. The transition from a crying baby to a talking child reflects such developmental stages accurately. Yet, development is under genetic control. Genes are not activated all at once. Some are switched on early and remain so, others are switched on late, and others are switched on and off. It is obvious that lately activated developmental genes can influence specifically the differentiation of the top level of the system hierarchy, without being masked by developmental brain plasticity. Such late-acting genes are mostly of no particular importance for the proper functioning of the rest of the brain. Because spontaneously occurring mutations among them are not rapidly eliminated by natural selection, one can expect to find many such mutated genes in the gene pool of a population.

This implies that a considerable proportion of individual differences in cognition is caused by such subtle mutations. On the other hand, it is also obvious that environmental factors are modifying the impact of these mutations: the later-acting the gene, the stronger the influence of the environment (Figure 1). (*See* **Language Development, Critical Periods in**; **Memory, Development of**; **Piaget, Jean**; **Neuropsychological Development**; **Individual Differences**; **Early Experience and Cognitive Organization**)

LINKAGE STUDIES

What is a Trait?

How can the effects of such subtle mutations be identified, and how is it possible to associate them with a particular gene in both humans and animals?

The classic 'top down' approach tries first to identify genetic variation of mental traits. A trait means a difference between individuals. It is defined as qualitative when it describes a noncontinuous difference, such as blond or black hair. A quantitative trait is a measurable difference between groups; however, individual members of the groups show considerable overlap, such as body height differences between men and women. Because mental traits depend heavily on the external environment and brain plasticity, they are invariably quantitative.

Behavior genetics assumes that a mental trait can be caused by actions of single genes (monogenic), or by joint action or interaction of many genes (polygenic). In either case, effects from the environment are expected. In the broadest sense, they include anything not coded directly by genes and may encompass the internal environment of the body, including internal factors such as brain plasticity, and external factors such as education. Note, however, that the frequently used term 'gene' denotes a defined string of two deoxyribonucleic acid (DNA) sequences located in a defined part of the genome (locus). Thus, a gene itself cannot cause observable differences between individuals unless there are corresponding differences in their nucleotide sequences. Alternative variants at the same locus are known as 'alleles', and their occurrence as 'genetic polymorphism'. For the sake of simplicity, we will keep the expression 'gene' as a shorthand for alleles.

Testing Twins and Inbred Mice

The classic way to recognize whether a mental trait has a genetic component is to compare individuals with an identical set of genetic information. This is done most easily by subjecting inbred mouse strains to various behavioral tests. The resulting scores will often reveal a difference between the mean values, which roughly reflects the effect of one or several genes, while the variation around the mean reflects the dissimilarity caused by environmental factors.

In humans, the approach is modified because genetically identical individuals are found only as twins or triplets, which is insufficient for an analysis of replicates. Here, two strategies prevail. One is to compare many pairs of concordant (genetically identical) with discordant (dissimilar) twins, both combinations growing up in the same environment which is assumed to be constant. If mental traits are more similar between concordant twins than between discordant ones, it is possible to calculate the impact of genetic factors in the form of an index dubbed 'heritability'. If, on the other hand, concordant and discordant twins show on average equal differences, then the mental trait must have a purely environmental origin. The other approach is to compare the cognitive abilities of concordant twins raised apart in different environments. Here, small quantitative differences indicate strong heritability, large differences weak heritability.

The sophisticated quantitative partitioning of genetic and environmental components is the hallmark of classical quantitative genetics and has been useful in evidencing genetic variation of mental traits. The technique has been broadly accepted for animals but has met strong opposition when applied to humans – partly for ideological reasons, but also because the complex statistical procedures and underlying assumptions can be criticized on technical grounds.

From Population Genetics to Linkage Analysis

For a long period, the population genetic approach focused on demonstrating heritability and was less interested in identifying the genes and brain processes causing mental traits. This scenario has changed with the rapid progress of molecular genetics which has permitted both the deciphering of the DNA sequences of the human and mouse genome, and the mapping of the location of genes on the chromosomes. Moreover, these techniques have also identified many short nucleotide sequences, some of which are part of functional (coding) genes, while others belong to noncoding DNA sequences. This has enabled new forms of genetic linkage analysis combining sophisticated statistical techniques with molecular biology.

What is a marker?

Genetic transmission involves the random breaking up and recombination of chunks from maternal and paternal chromosomes in the germ-line cells, a process called 'segregation'. These chunks still contain hundreds and thousands of genes which cosegregate together; such genes are said to be 'linked'. In principle, linkage analysis requires easily identifiable markers with genetic polymorphism and known chromosomal location. For example, in the mouse the *PrnP* gene encoding for the prion protein is located not too distant from the *albino* locus. In creating knockout mice (see below), the chromosomal segment from the donor strain carrying the inactivated *PrnP* gene encoded a brown coat color. Using coat color as a marker made it possible to discriminate without molecular testing between mutant and wild-type mice.

Molecular markers

Today, many thousands of polymorphic DNA markers of different types are mapped to chromosomal locations, and computerized databases exist for both mice and humans. This progress permitted new forms of quantitative genetics such as quantitative trait loci (QTL) analysis. In mice, QTL techniques are used to analyze the cosegregation of molecular markers and both behavioral and cerebral traits. This is done by cross-breeding inbred strains known to carry unique DNA markers and to show differences in cognitive abilities. This strategy can identify statistical associations between markers and mental traits in the offspring. By using available databases and software, putative chromosomal locations can be computed which may harbor known or unknown candidate genes contributing to the genetic variation of the trait. It is then a matter of molecular biology to isolate DNA strings and to test whether the parental strains show discordant DNA sequences.

In humans, the QTL mapping method is employed in twin research, and for screening people with outstanding cognitive performance. Since it is not known *a priori* whether these groups carry a specific marker, it is necessary to test associations with a large battery of DNA markers with the hope of finding a significant association between certain markers and mental traits. Moreover, such findings need replication in other samples. Nonetheless, this approach has resulted in the identification of several candidate loci responsible for individual differences in cognitive abilities and disabilities of humans. (*See* **Behavior, Genetic Influences on**)

Another line of linkage research focuses on mental disabilities which are linked to an identifiable chromosome. For example, forms of nonspecific learning difficulties are known to be associated with suspected mutations on the X chromosome. In contrast to traits, these more severe manifestations of mutations are termed 'phenotypes'. By mapping the X chromosome for markers found in afflicted families, candidate regions on the chromosome were identified and subsequently analyzed by DNA cloning. This led to the identification of mutations in a gene (*GDI1*) encoding a protein essential for vesicular transport within neurons. (*See* **Fragile X Syndrome**)

GENE KNOCKOUT TECHNIQUES

The Rationale of a 'Bottom Up' Approach

The second conceptual approach to the analysis of relations between genes and mental abilities is a 'bottom up' strategy derived from the classical studies of known single gene mutations (such as albinism) and cognitive correlates. The advances in molecular biology have now added a new perspective: targeted inactivation of genes suspected of having a fundamental role in memory and learning, followed by an analysis of brain structures, neurophysiology, and behavior in order to recognize the chain of action between genes and mental processes. The target may be chosen for theoretical reasons, using gene deletion as a refined lesion technique to inactivate brain processes inaccessible by other methods. Alternatively, the target is selected because mutations of a human gene have been found in association with mental retardation. In fact, mouse mutants with a deletion of the above-mentioned *GDI1* gene have been produced; these show impaired memory for aversive tastes and inappropriate social behavior. Up to now, these techniques have been confined to mice for technical reasons. This restriction poses some problems in assessing mutation-induced cognitive changes in a species with fairly limited mental abilities. On the other hand, this handicap is compensated for by the vast knowledge amassed through behavioral genetic studies of mice, a factor that facilitates comparative behavioral analysis.

Conventional Knockout Mice

Conventional gene targeting is done in mice in dividing embryonic stem (ES) cells maintained in cell cultures, mostly by inserting an artificially constructed DNA strand which binds to the complementary targeted gene during replication

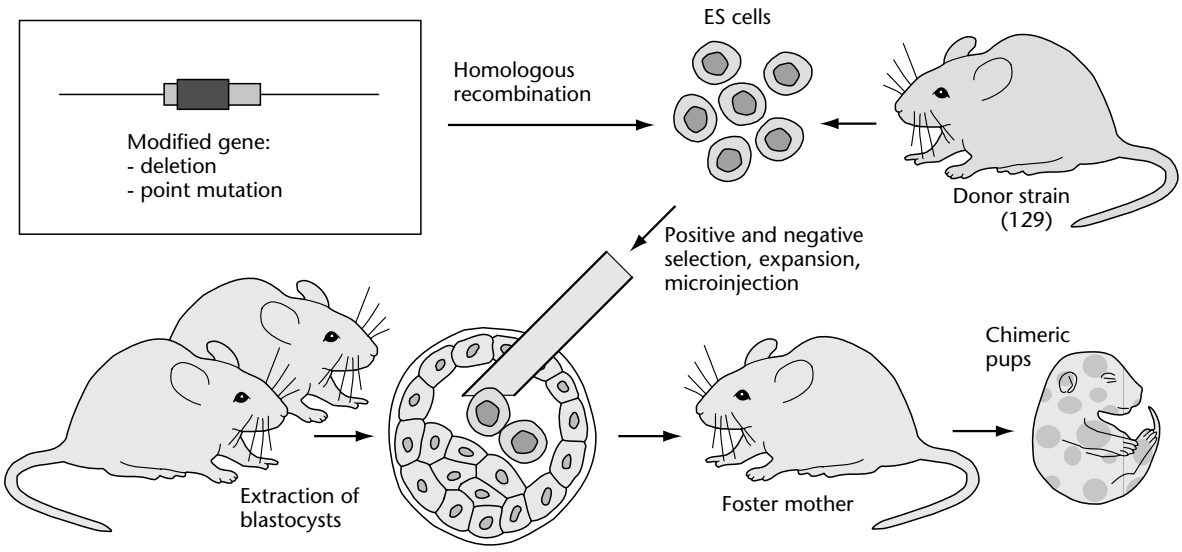


Figure 2. Production of transgenic mice with general knockout of target genes (conventional knockout). The genetic material is injected into embryonic stem (ES) cells, and binds specifically to the target gene which it may replace during recombination of chromosomes (homologous recombination). The inserted nonfunctional gene is often linked with molecular markers for easy detection. Inserting a normal or slightly modified gene is called a 'knockin'.

(Figure 2). Stable topological incorporation of the foreign strand into the chromosome can prevent the reading of the entire gene (null mutation). Alternatively, the allele may still function but yield much less protein, in which case it is termed 'hypomorphic'. The ES cells with stable targeted mutations are then identified in cell cultures, merged with blastocytes from another mouse strain and implanted into foster mothers. This procedure results in 'chimeric' mice, among whom a few individuals might be found in which the targeted mutation has been incorporated in germ-line cells differentiating into sperm and eggs. The entire process thus takes considerable time and is costly.

Like a spontaneous mutation in germ-line cells, gene knockout in ES cells results in inactivation of a gene in all tissues of the body. Also, inactivation starts with conception and lasts throughout life. For reasons explained above, alterations of brain and behavior in such mutants are often difficult to interpret and it may sometimes be impossible to establish a direct causal link to the gene product. This is because compensatory mechanisms may not only mask phenotypic changes, resulting in variable expression (penetrance of the mutation) but also because they interact unpredictably, entailing different patterns of expression (pleiotropy). Thus, it can never be excluded that the phenotype reflects compensation mechanisms rather than the functional inactivation of the gene. On the other hand, mutations may have such devastating consequences that affected individuals die during development and cannot even be examined. All these difficulties have motivated molecular biologists to make remarkable efforts to develop new techniques that permit 'conditional' gene targeting; that is, to introduce genetic changes in a tissue- and time-dependent manner.

Tissue-specific Knockouts

The frequently used tissue-specific technique exploits the activity of the enzyme Cre recombinase which removes pieces of DNA in a chromosome that are flanked by the enzyme's recognition sequences, the so-called loxP sites. In nature, the enzyme is used by viruses to recombine their own DNA with that of the infected bacterial cell. Thus, neither Cre recombinase nor loxP sites occur in normal mammalian cells. This makes it possible to exploit the system to generate tissue specific knockouts in a process that involves three main steps:

- conventional insertion of an intact gene flanked with loxP sites in a mouse line
- production of transgenic mice carrying the gene for Cre recombinase in specific parts of the body
- crossing the two lines.

During recombination, the 'floxed' gene will be excised in those brain cells which contain the recombinase. For details see Figures 2 and 3.

This technique has proved particularly useful in the examination of genes whose loss is fatal if it

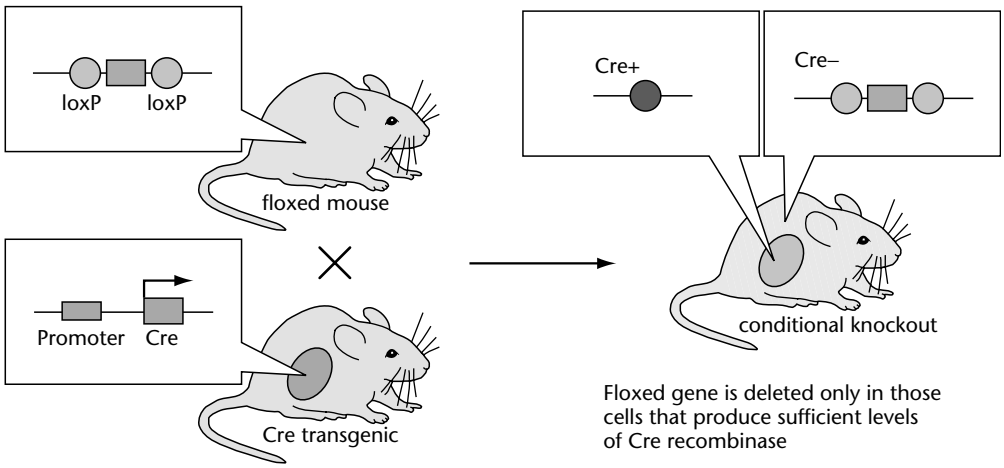


Figure 3. Production of mice with tissue-specific knockout of genes, the so-called 'conditional knockout' by Cre/loxP-mediated recombination. First, conventional gene targeting in embryonic stem (ES) cells is used to replace the gene of interest with a copy that is flanked by loxP sites but otherwise remains fully functional. The resulting 'floxed' mouse is completely normal and cannot be distinguished from wild-type mice without specialized assays that detect the presence of loxP sites in mammalian tissues. In a second step, using conventional pronuclear injection, a transgenic mouse line is created that expresses Cre recombinase in the tissue where one wants to inactivate the gene of interest. Depending on the questions that one is asking, the transgene will be constructed in such a way that Cre recombinase will become active early in development or only later when the animal approaches adulthood. The last step consists in crossing the 'floxed' mouse with the transgenic line expressing the Cre recombinase. In double mutant offspring, the 'floxed' gene of interest will be destroyed in every cell that expresses sufficient amounts of Cre recombinase.

occurs early in development and in the whole body. One example is the cell membrane protein trkB which serves as receptor for brain-derived growth factor (BDNF). It has been suggested that trkB plays an important role in the modifications of neural circuitry that underlie learning and memory in mammals, including humans. However, conventional gene targeting could not be used to test this hypothesis, because inactivation of the trkB gene is lethal if it occurs in the whole body and early in development. Using the technique described above, a tissue-specific knockout has been generated in which the trkB gene is removed only in the forebrain, and not before completion of development. These mutant animals survive and are healthy, making it possible for the first time to study brain function and behavior in adult animals in the absence of trkB receptors. Several experiments with these mice confirmed the hypothesis that trkB receptors are essential for both synaptic function and learning. Conceptually, the Cre–lox technique is an example of late-acting mutation effects.

Inducible and Reversible Transgenes

An even more flexible time-dependent control is achieved through the introduction of genetic changes that put the activity of a transgene under the external control of a drug. Feeding the drug or withholding it from a mutant mouse switches the mutation on or off. Several such systems have been developed or are under development.

The technique which thus far has made the most substantial contribution to the investigation of cognition is the control of the expression of a transgene by the antibiotic drug group known as tetracyclines. This approach exploits the fact that gene expression in living cells is controlled by the binding of endogenous regulatory proteins to specialized chromosome regions called 'promoters'. By combining bacterial and viral DNA sequences, molecular biologists have engineered an artificial system consisting of the tetracycline-controlled transactivator (tTA) and a corresponding promoter (tetO). The tTA turns on any transgene that is connected to tetO, but this activation is blocked by tetracycline. Normal mammalian tissues contain neither tetO promoters nor a gene that codes for tTA. Therefore, as with the Cre/loxP system described above, generation of an inducible mouse model requires three main steps:

- production of mice with a transgene for tTA in selected tissues
- production of mouse lines linked to the tetO promoter
- crossing the two lines and feeding or withholding the antibiotic to double mutant offspring (Figure 4).

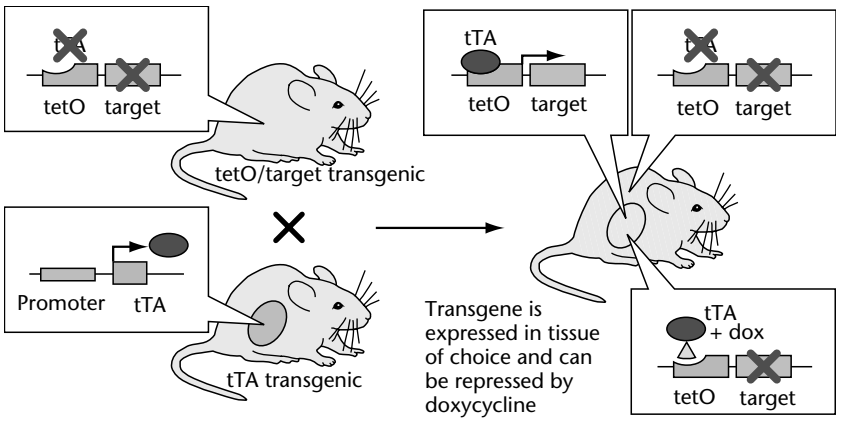


Figure 4. Producing switchable knockout mice using a tetracycline-controlled transactivator (tTA) system which permits reversible transgene repression by doxycycline. First, pronuclear injection (see Figure 5) is used to create a transgenic mouse line that bears an artificial gene expressing tTA in selected tissues, such as the forebrain. Next, a second mouse line is engineered that carries the transgene of interest linked to the artificial promoter (tetO). Finally, the two engineered lines are crossed. In double mutant offspring, expression of the transgene occurs only in selected tissues and can be reversibly suppressed at any time by feeding the animals a tetracycline antibiotic such as doxycycline. More recently, the system has been made even more flexible by engineering a reversed transactivator (rtTA). If rtTA is used instead of tTA, tetracycline activates rather than suppresses the expression of the transgene. The rtTA/tTA and Cre/loxP systems can be combined by putting the expression of Cre recombinase under the control of tTA or rtTA. The generation of such triple mutant mice is complicated, but allows flexible control of the time point at which a 'floxed' gene is effectively knocked out. In other studies, the tTA/rtTA system has been used to control transgenes that interfere directly with cellular function.

It can be shown that such a manipulation does indeed affect memory and learning. The approach has been used to create mice that overexpress the phosphatase calcineurin whenever they are fed tetracycline. It has been known for a while that transient protein phosphorylation in neurons has an important role in learning and that this phosphorylation is controlled by a fine balance between kinases and phosphatases, enzymes that add phosphate residues to proteins or remove them, respectively. However, conventional genetic models have not addressed the question whether protein phosphorylation in neurons is needed only to form new memories, or also to recall previously established memories. (*See* **Memory, Neural Basis of: Cellular and Molecular Mechanisms**; **Synaptic Plasticity, Mechanisms of**; **Memory Consolidation**)

Because overexpression of the phosphatase calcineurin potently disturbs protein phosphorylation in neurons, the rtTA/calcineurin mice were most useful in addressing this question. When calcineurin overexpression was switched on while the mice had to learn to find a hidden platform in a water pool, they failed to learn and were later unable to find the platform. This confirmed the hypothesis that protein phosphorylation was needed to form new memories. In a second experiment, the mice were first taught to find the hidden platform and were fed tetracycline only after learning. This treatment clearly blocked their ability to recall what they had previously learned without problems.

Together, these two elegant experiments provide strong evidence that fine control of protein phosphorylation in neurons is needed not only to form new memories, but also to recall those that are already formed. They also come closest to the premise of gene knockout techniques as an ultrafine brain lesioning tool for the study of memory and learning.

GENE REPLACEMENT TECHNIQUES

Molecular biologists have also succeeded in replacing nonfunctional or artificially deleted genes.

Transgenic Animals

The first and most simple way of inserting genes is a kind of shotgun approach. Isolated and cloned genetic material can be injected into an oocyte (pronuclear injection; Figure 5). The inserted genes integrate randomly into the chromosomes of the recipient egg cell. The oocyte is then implanted into a foster mother, yielding an animal that carries one or many copies of the inserted gene. Further

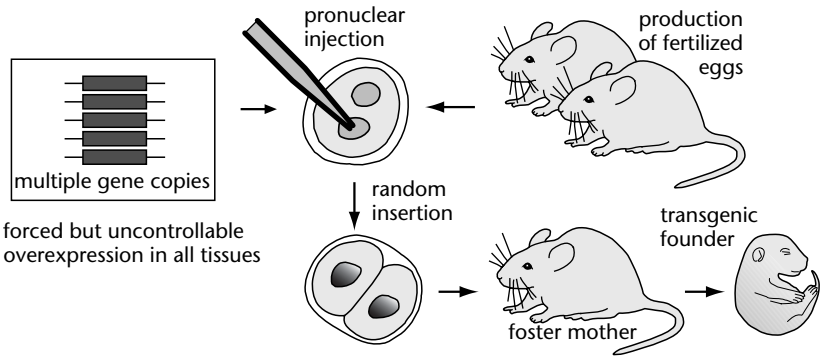


Figure 5. Production of transgenic mice by injection of genetic material into fertilized eggs (pronuclear injection). Historically this was the first rather primitive genetic manipulation.

breeding of such animals leads to a transgenic line. The technique has been applied to many animal and plant species.

The analysis of cognition in such transgenic mice has become somewhat obsolete, because the number of gene copies and their location cannot be determined. Moreover, uncontrolled integration can entail additional gene deletions obscuring the behavioral phenotype. At present, two applications of transgenic animals remain. One is the creation of mouse models of degenerative diseases. For example, there are transgenic mouse lines carrying multiple copies of a human mutant gene for the β-amyloid precursor protein (β-APP) which show increased amyloid deposits in their brain. Since they show also deficits in learning tasks, they are used to test drugs designed for the treatment of Alzheimer disease. The other application is to rescue learning performance in knockout mice by reinserting copies of the deleted gene. This technique serves mainly to validate knockout effects, although it does not really allow discrimination of direct gene effects from compensatory processes. Finally, transgenic mice are used for the production of advanced gene targeting techniques.

Gene 'Knockins'

The gene 'knockin' technique is a much more sophisticated variant of the knockout methods described above. Instead of targeted insertion of a disabling DNA sequence, an allele can be inserted which encodes a modified protein. This has been used, for example, to study memory formation in mice with a knocked-in gene carrying a point mutation that blocks the autophosphorylation of α-calcium–calmodulin kinase II. These mice showed a spatial memory deficit combined with blockade of long-term potentiation (a change in the signal transduction property of neurons thought to underly memory formation).

Viral Gene Transfer

The techniques above have a severe limitation: mutations must be incorporated into the germ line, they must become heritable in order to study their effects, and most of them are confined to mice. Somatic mutations, on the other hand, die with the cells of their carrier, and would seem desirable for medical reasons as well as for studies of gene–brain interactions. The problem is how to insert such manipulated genes into the brain and into the right neurons. A potential solution is to use viruses with low infectivity that can enter brain cells without damaging them. Transferring specific DNA into cells maintained in culture by the intermediary of a virus is relatively easy. However, applying the technique to living organisms faces numerous obstacles, and is still in an early experimental stage. Nonetheless, the procedure bears great promise.

POTENTIAL OF GENETIC RESEARCH IN THE STUDY OF COGNITION

Scientific Development

It would seem realistic to expect rapid developments in the molecular genetics of cognition. The lead is likely to be taken by the 'bottom up' approach of molecular biology. The number of scientists working in this field is large, and the development of new methods is more dynamic than in psychology and classic behavioral genetics. Moreover, the techniques that are evolving for analyzing other parts of the body can be transferred readily to the brain.

One can expect many new conventional knockout models, owing to the rapid spread of this more simple procedure. Unless natural genes can be targeted which are known to act selectively on given brain systems, they may be less useful for studying relations between genes and cognition. On the other hand, they may serve well to analyze hereditary mental disabilities because they share the same developmental pathways as the natural mutations.

The new generation of knockout and knockin models is likely to generate insights into the general cellular mechanisms underpinning memory and learning. Success in microanalyzing brain systems will depend, however, on better knowledge of brain systems and their mode of interaction.

The experimental stage in the uses of somatic mutation induced by viruses or other carriers will probably continue for some time, certainly in the field of genetics and cognition. However, the enormous potential for medical application is a strong driving force that will eventually result in substantial progress, whose ramifications cannot be estimated yet.

The 'top down' approach will profit from a merger with molecular biology. Its main asset is the ability to identify unknown genes by behavioral screening, and the traditional knowledge about relations between complex behavior and brain systems. Thus, its chief role will be to provide empirically relevant target genes for molecular biology, and to analyze the pathways from gene to behavior.

Impediments

The technical impediment of the main molecular approach is the limitation to germ-line mutations in mice, which is likely to persist for several reasons. Another related obstacle is the simplistic approach to analysis of using a few not always appropriate tests of cognition in mice and extrapolating the results to humans.

A nontechnical impediment is the confrontation between mechanistic and evolutionary biological thinking, as the two have different priorities in setting research goals. However, the dividing line cannot be drawn clearly between molecular biologists and psychologists, because the latter include both categories. In order to merge the two fields fruitfully, psychologists have to abandon, or to soften at least, some deeply rooted prototypical thinking, while molecular biologists have to understand the meaning of biological variability and plasticity of the brain.

The last and perhaps most serious impediment is the public perception of this research. Members of the public tend to think in mechanistic terms, assuming that genes are the first link in a machinery in which every step predicts the next one – hence genes control the fate of the individual. It means also that experiments with genes are perceived as attempts to control cognition and thinking; moreover, they seem to challenge divine or natural wisdom, thus bearing the risk of punishment. Similarly, such mechanistic oversimplification carries political implications, provoking unnecessary opposition by concerned people. Thus, meaningful progress will depend on the public and politicians as well as scientists understanding the probabilistic nature of behavior. The only way in which a mutated gene can manifest itself, be it of natural or technical origin, is to bias the probability of innate or acquired motor actions such as movements and speech. Likewise, natural genetic variation of mental traits is a property of populations useful for evolution but of minor significance for the carrier, because individual genetic biases are automatically counterbalanced by brain plasticity, learning, and appropriate environment – the latter two of which are in the hands of people and politicians.

Neuroimaging

Intermediate article

James B Rowe, Wellcome Department of Imaging Neuroscience, London, UK
Richard SJ Frackowiak, Wellcome Department of Imaging Neuroscience, London, UK

CONTENTS

Neuroimaging techniques include magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT). These techniques provide information about the structure and function of the brain in health and disease, and have increased our understanding both of regional specialization and functional integration of different brain regions.

INTRODUCTION

Since 1980 there has been an explosion in brain imaging technologies. Neuroimaging has a central role in cognitive neuroscience, including the organization of language, memory, sensory perception, control of action, and the mechanisms of emotional experiences and learning. It is advancing our understanding of the neuropsychological consequences of brain injury, and the mechanisms of recovery.

This article attempts to provide the reader with a rational structure for understanding these emerging techniques, and introduce key concepts in the acquisition, analysis and interpretation of data. A broad distinction is made between functional and structural neuroimaging. However, this distinction is not absolute. Increasingly, functional studies incorporate prior anatomical knowledge. Conversely, the analysis of brain structure is guided by our understanding of local functions. In both functional and structural neuroimaging lie two complementary approaches to the organization of the brain. Firstly, that there is some regional specialization by which different regions serve different functions. Secondly, that connections within and between specialized regions support functional integration, such that tasks utilize a network of regions distributed across the brain.

PRINCIPLES OF FUNCTIONAL NEUROIMAGING TECHNIQUES

Table 1 provides an overview of the common neuroimaging techniques. Magnetic resonance imaging (MRI) is perhaps the most versatile method, used to characterize structural (MRI), functional (BOLD-fMRI, blood oxygenation level dependent-functional MRI) and chemical (MRS, magnetic resonance spectroscopy) changes in the brain. Positron emission tomography (PET) and single photon emission computed tomography (SPECT) can be used to study neuropharmacology as well as neuronal metabolic function. These methods differ in their spatial and temporal resolution (illustrated in Figure 1), as well as cost and safety considerations.

Blood Flow and Aerobic Metabolism

Functional neuroimaging techniques exploit the close relationship between neuronal activity and metabolism, and the secondary effects on local vasculature (see Figure 2). Although some details of neurovascular coupling remain controversial, particularly for fMRI, there is a linear relationship between neuronal activity and the measured signal (Logothetis *et al.*, 2001). Increased local neuronal firing causes greater post-synaptic activity in a population of neurons.

The increased neuronal activity leads to greater aerobic metabolism of glucose. This may be measured directly by PET using [^{18}F]-2-fluoro-2-deoxyglucose (FDG). Alternatively, $H_2{}^{15}O$ can be used to measure the distribution of blood flow to active brain tissue. The isotopes decay to produce a positron, which in turn produces two photons, detected by an array of sensors around the subject's

Table 1. A summary of common neuroimaging techniques by principal subject and mechanisms together with abbreviations

Neuroimaging subject	*Principal mechanism*	*Object of measurement*	*Method*	*Abbr.*
Neuronal activity	Endogenous metabolites	E.g. actate, choline, n-acetylaspartate	Magnetic resonance spectroscopy	MRS
	Exogenous metabolites	Radiolabeled metabolites	Positron emission tomography	PET
		E.g. fluoro-deoxy-glucose	Single photon emission computed tomography	SPECT
	Neurovascular responses	Perfusion	Positron emission tomography	PET
			Single photon emission computed tomography	SPECT
			Arterial spin labeling	ASL
		Blood oxygen level dependent (BOLD) MRI signal	Functional magnetic resonance imaging	fMRI
Neuropharmacology	Radioligand binding	Radiolabeled agonists/ antagonists	Positron emission tomography	PET
		E.g. fluoro-dopa	Single photon emission computed tomography	SPECT
Structural	Nuclear magnetic resonance	Tissue type and distribution	Magnetic resonance imaging	MRI (VBM, TBM, DBM)
	X-ray absorption		X-ray computed tomography	CT
Neuronal electrical activity	Focal coherent neuronal electrical activity	Induced electric field	Event related potentials	EEG/ ERP
		Induced magnetic field perturbations	Magneto-encephalography	MEG

head. With adjustments for the attenuation of photons by brain and skull tissues, the computed tomograph of the photons' sources may be read as a 'map' of regional cerebral metabolism or blood flow. Multiple scans are made, with subjects performing different cognitive or motor tasks. The maps may be contrasted in order to determine regions in which a particular task was associated with greater neuronal activity.

fMRI exploits the properties of oxygenated blood flowing in a magnetic field. Increasing aerobic metabolism requires increased local oxygen extraction from blood. There is increased blood flow locally. The increase in blood flow is greater than that required to supply the increase in oxygen consumption, by a factor of about two. The result is a paradoxical fall in deoxyhemoglobin concentration near increased neuronal activity. Unlike oxygenated hemoglobin, deoxyhemoglobin is paramagnetic. This is the basis of the BOLD signal used for fMRI. A 'map' of the changes in brain activity may be inferred from the changes in BOLD signal. For normal subjects in typical scanning environments, there is a very gradual drift in the baseline BOLD signal over minutes or hours. In contrast, the BOLD signal change due to neuronal activation occurs over 5–15 seconds.

Complex Metabolic Indices

The BOLD signal used in fMRI depends on the nuclear magnetic resonance of water protons. However, protons in complex molecules (e.g. creatine, choline, glutamate, n-acetyl-cysteine or lactate), and nuclei of other elements (e.g. ^{31}P) also exhibit nuclear magnetic resonance, at slightly different frequencies. This forms the basis of MRS. It is very informative about metabolic states, and therefore subtle differences between pathological processes. It has been used to try to distinguish different brain tumor types, or demyelinating diseases. However, MRS has poor spatial and temporal resolution, because the signal is weaker than that of BOLD-fMRI.

Neuroimaging can also be used to identify the distribution and function of specific neurotransmitter receptors, using radiolabeled agonists or antagonists such as 6-[^{18}F]-dihydroxy-phenylalanine

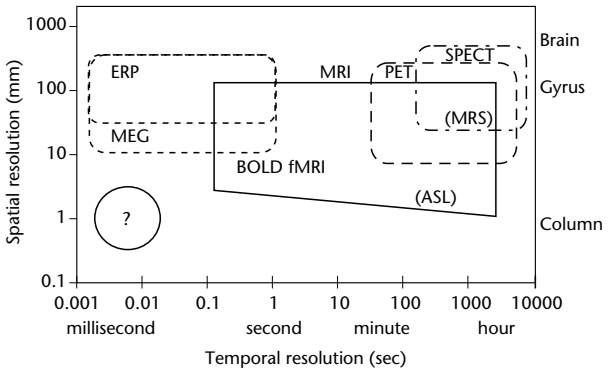


Figure 1. Diagrammatic representation of spatial and temporal resolution of minimally invasive human functional imaging techniques (adapted from Belliveau *et al*. (1992) *Investigative Radiology* **27**: S59–65). More invasive methods such as implanted electrodes' recording may offer higher combined spatial and temporal resolution in limited brain regions, but are rarely possible in humans. The ideal method (suggested by '?') would provide high spatial and temporal resolution across the whole brain at low cost and with high safety. The highly versatile platform of MRI may eventually be extended, alone or in combination with electrophysiological techniques, to provide this optimal imaging technique.

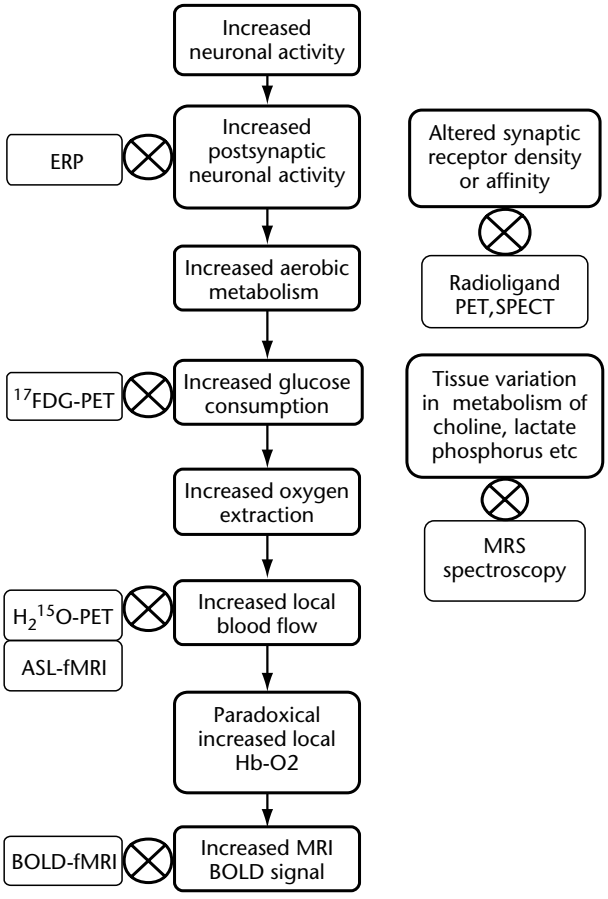


Figure 2. Increased neuronal activity has a series of electrical and metabolic consequences (dark outlines), which are detectable by different neuroimaging techniques (light outlines).

(fluoro-DOPA) and [^{11}C]-raclopride. Fluoro-DOPA acts as a false neurotransmitter, localizing in proportion to dopamine receptors in the striatum, and detected by PET. The monoamines are most commonly studied, because of their central role in the pathophysiology of major neurological diseases like Parkinson's disease, and psychiatric illnesses including depression and schizophrenia.

Application of PET and fMRI

The majority of functional neuroimaging at present uses $H_2{}^{15}O$ PET and BOLD-fMRI. Their applications are diverse, published throughout the neurological, psychological, physiological and general scientific literature, including many other parts of this encyclopedia. The early methods for neuroimaging, like PET and SPECT, required many seconds or minutes to acquire each brain image. During acquisition, subjects could perform many repetitions of the same task, or a complex task with many components. The overall activity might be due to any or all of these components.

The different tasks have some features in common, such as sensory stimulation (visual or auditory), motor actions (to indicate the chosen response, or learned motor sequences) or cognitive processes (reading, speech, working memory, long term memory encoding or recall, attention). The topic of research is reflected in critical differences *between* tasks. Areas of differential activation can be associated with the particular sensory, motor or cognitive elements of interest.

The different elements of a task might interact. For example the activation related to observation of object movement (versus static objects) might depend on whether or not attention was directed towards movement. To understand these interactions, different tasks must be designed that vary each factor independently of the others. This is known as a factorial design.

More elaborate designs include parametric variation of a factor, e.g. the speed of movement of visual stimuli, the number of items to be remembered, or the rate at which a task must be performed. The passage of time in a learning study may also be considered as a parametric modulator of activation.

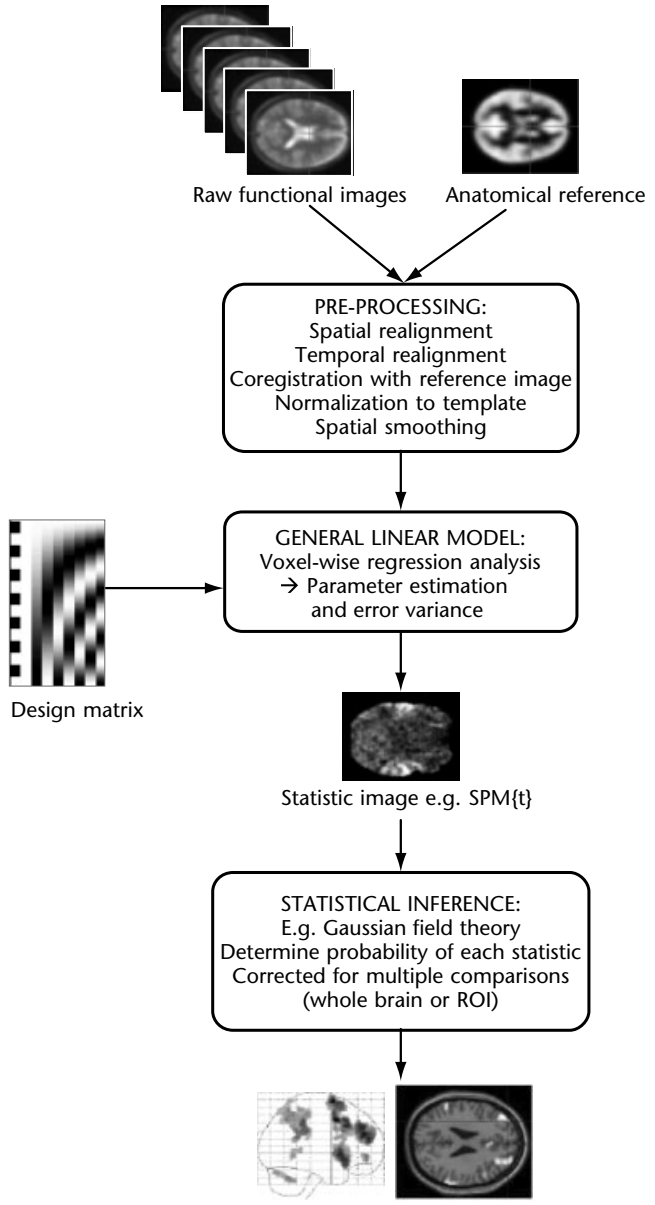


Figure 3. General scheme for analysis of functional and structural images. The three main stages are pre-processing, application of a general linear model, and statistical inference. Not all pre-processing steps are essential, but the interpretation and description of final results will vary for different pre-processing steps. The data are then subject to a general linear model, the basis for parametric tests including t-test, F-test, ANOVA, correlation, multiple regression, etc. In hierarchical models, such as random effects analyses, the application of a general linear model is repeated at each level of the hierarchy. Lastly, the probability of regional activations is determined, indicating where there are significant activations associated with the experimental conditions or tasks. Images courtesy of SPM, Wellcome Department of Imaging Neuroscience, London.

PRINCIPLES OF IMAGE ANALYSIS

A general scheme for analysis of functional or structural brain images is given in Figure 3. It can be divided into three main stages: pre-processing; application of a general linear model; and statistical analysis and inference.

Data Pre-processing

All forms of neuroimaging data undergo some form of processing before formal analysis. The minimum pre-processing is to reconstruct the raw data into a 3-D matrix, each data point representing a finite volume of brain tissue, called a voxel, for

each time point during an experiment. The subsequent processing stages will be illustrated using BOLD-fMRI data, but the principles may be generalized.

Pre-processing is aimed at allowing fair comparison of images, with optimal signal to noise ratio. To correct for movement between scans, the images are realigned to the same frame of reference. The realigned functional images may be coregistered with a structural scan of the subject, so that any regional task-related activity may subsequently be viewed in relation to the anatomy of the subject. When studying a group of subjects, it is preferable to accommodate the variation in size and shape of brains by 'warping' the brains to an international standard shape (an average shape of many normal brains). This process is called normalization. A third pre-processing step is often undertaken called smoothing, by which the data from each voxel are averaged with the data from the voxel's neighbors. Smoothing of data improves signal to noise ratio, at only a modest cost to spatial resolution. Smoothing data is necessary to use the theory of Gaussian fields to draw statistical inferences about regionally specific effects (see below). The fMRI data are often band-pass filtered during analysis. The filter may have two components: a high pass filter and temporal smoothing. This removes very low frequency effects such as the slow drift in BOLD signal over minutes.

The General Linear Model

A general linear model describes the activity of a voxel in terms of a weighted linear combination of explanatory variables (e.g. sensory, motor or cognitive events), plus some error. Given a series of images, it is possible to calculate the values of these weights, for each explanatory variable for each voxel, with minimal error. These best linear unbiased estimates of the weighting factors for each variable are known as the parameter estimates.

For accurate and meaningful parameter estimation, and subsequent statistics, it is essential to design the experiment carefully. Special consideration is given to how complex experimental tasks can be described in terms of the presence or absence of simpler explanatory variables such as stimulus presentation, memory encoding, or motor responses. Common statistical tests including t-tests, ANOVA (analysis of variance), multiple regression analysis, and correlation analysis are all based on general linear models.

Statistical Inference

The commonest statistics derived from the general linear model are the parametric t- and F-statistics. For t-tests, the magnitude of the parameter estimates must be considered against variability in activity due to chance alone, i.e. the error variance. To contrast the effects of two experimental conditions (e.g. activation in a task versus resting state) one can weight the parameter estimates, say $+1$ and -1, for the two conditions of interest. The statistic is calculable for every voxel in the brain producing a statistical parametric map, known as the SPM{t}. Statistical Parametric Mapping, SPM, refers to this general process of image analysis and is also the name of one of several software packages available for this process.

The F-statistic is the ratio of variance explained by inclusion of a variable in the model (or weighted combination of variables), divided by the error variance of the whole model. It may be used to compare models in a hierarchy of complexity, or to test a set of multiple linear hypotheses in a model. A map of the F-statistic can be drawn for the whole brain, called the SPM{F}.

To draw statistical inference, it is necessary to assign a probability value to the t- and F-statistic at each voxel. However, hundreds or thousands of voxels may have been scanned. Bonferroni correction for multiple comparisons is not appropriate because a voxel's data are not independent of the data in neighboring voxels. One solution is to consider the voxels to be a lattice representation of a continuously varying field across the brain. The theory of Gaussian random fields may then be used to make clear predictions about the expected features of the SPM{t} or SPM{F} (Worsley *et al.*, 1996), and assign a probability value to each voxel for a given contrast of experimental conditions.

There are different types of inference that can be drawn from the statistical parametric maps. There is a trade-off between regional specificity of the inference and the sensitivity. The most regionally specific test is to ask whether the t- or F-value at a voxel exceeds a threshold. For example, were voxels in the temporal cortex more active during reading than rest. In contrast, there is the omnibus test of whether the overall number of suprathreshold clusters of voxels is significant. For example, is the brain map different when reading compared with rest, without specifying where those differences lie. Of intermediate specificity and sensitivity are cluster level and set-level tests. The cluster level test determines the probability of spatial extent for each suprathreshold cluster. For example, when

comparing reading with rest, is the group of activated voxels in the temporal lobe significantly large. The set-level test determines the probability of the observed number of suprathreshold clusters that each exceed a specified size. For example, are there significantly more regions of activation above a certain size than would be expected by chance.

INTERPRETATION OF VARIANCE

General Principles

When studying a group of subjects under several conditions, there are many possible reasons for differential activation of a brain region. For example, it may be that the experimental condition is associated with different activity at that part of the brain. This is called 'condition variance'.

However, activation may differ because of structural brain differences; or because subjects perform the task differently; or that they have inherently different neurovascular coupling. These factors contribute to 'subject variance'. Even if these factors are controlled, the sensitivity and baseline of the BOLD-fMRI signal changes between scanning sessions. This leads to 'session variance'.

Session variance is generally large in comparison with condition variance in fMRI. Session by condition interactions may significantly explain changes in regional brain activation. If so, task-related activations for a given subject cannot be generalized from a single session to infer 'typical' activation patterns from this subject. In addition, differences of task-related activation between sessions might be due merely to session by condition interactions rather than the factor of interest (e.g. scanning before and after drug therapy).

Fixed and Random Effects Analyses

One often wants to extend inferences from particular subjects to the general population. A group of subjects may be considered to have been sampled randomly from the general population to which they belong, e.g. all men of that age, or all patients with a particular disease. This contrasts with the experimental conditions which are 'fixed' by the experimenter. Random and fixed variables must be treated differently in the analysis of variance of repeated measures, including neuroimaging.

For SPM{t}s, one calculates the ratio of the contrasted parameter estimates to a denominator variance term. If all variables are fixed (a fixed effects analysis), then the denominator term is the error variance within subjects (mainly noise from scan to scan). However, with a random variable like the selection of subjects, the appropriate denominator term must include the variance between subjects (random) as well as within subjects (fixed).

Models with both fixed and random effects are generally difficult to analyze. However, the balanced design of most imaging experiments enables a two-staged approach to mimic a mixed effects model. The first stage is a fixed effects model within each subject, to generate a summary image from each subject, such as the map of activation differences between reading and rest. These summary images can then be assessed across subjects in a second level (random) effects model. For balanced designs, the variance due to fixed and random effects is in the right ratio to assess the overall population effect.

In a fixed effects model combining data from several subjects, a significant activation may be due to a large effect in just one or two subjects, rather than a general effect across the group. However, it is still possible to make qualitative inferences about the population effects from a fixed effects model, if one stipulates that an effect is significant in all subjects. The refutation of the null hypothesis in all of the subjects is known as conjunction analysis. If a result is found in all subjects, even from a small group, it may be regarded as typical of the general population.

Group Data and Clinical Studies

The problem of extending inference to the subjects' general population is highlighted in clinical studies. The expression of disease will vary at a subject level, both between patients and over time in the same patient. If patients are rare or heterogeneous, how should they best be compared with control subjects, or studied over time? A random effects analysis may not always be possible. It requires a large group of patients (typically >12) who differ from the comparison group only in terms of the disease of interest. Many studies have used fixed effects models, and had to accept that the inference only extends to the particular patients studied. It is possible to use a conjunction analysis approach to show that each and every patient in a group differs from the control subjects with regard to a particular effect.

Ambiguity can arise when interpreting differences in regional brain activations in patient-control studies. These problems arise whether the patients have focal lesions, are subject to different drug therapies or are being studied during recovery

from disease. Interpretation depends on whether the performance, behavior and cognitive strategy in the patient group are the same as the control group. If a patient does not perform a task, or manipulations of the task suggest a different strategy is employed, then the normal pattern of brain activity cannot be expected. Neuroimaging cannot resolve whether the performance difference is attributable to the abnormal brain activity or vice versa.

ADVANCED MODELING TECHNIQUES

Event-Related fMRI

In early fMRI studies, the experimental paradigms were similar to those used in PET studies. Tasks were repeated during a specific 'block' or 'epoch' lasting approximately 30 seconds. fMRI can also characterize the BOLD response to transient events, such as a brief visual stimulus or single movement (Buckner *et al*., 1996). Although these events are transient, the BOLD response can be averaged across many repetitions. This has many advantages.

First, it exploits the excellent temporal resolution of fMRI, to identify separate patterns of brain activation caused by events even less than a second apart (Menon and Kim, 1999). Second, some cognitive and neurological processes can only be studied as occasional events, e.g. responses to unexpected or unusual events can only be studied if the events are rare, and not if they are recurrent in a block design. Third, some events cannot be controlled or predicted by the experimenter, e.g. correct versus incorrect responses to questions. The times of these events must be determined *post hoc*.

If stimuli can be presented individually, what is the best rate and order of presentation of different stimulus types? The rate of events may be defined by the stimulus onset asynchrony, SOA. With shorter SOAs, the BOLD response to one event becomes attenuated by previous events, but more events can be studied. If multiple event types are studied, the optimal SOA depends on the type of comparison to be made. In general, SOAs in the range 2–8 seconds are more efficient. Shorter SOAs are more efficient to determine differential responses between event types, whereas longer SOAs are more efficient to estimate the responses to each event type.

Effective Connectivity

Functional connectivity describes temporal correlations between spatially remote neurophysiological events. Effective connectivity describes the influence one area or neuronal system has over another. Functional connectivity between two areas may arise from effective connectivity, but it can also arise if there are common inputs from a third area.

Changes in effective connectivity are better understood when associated with particular experimental conditions. For example, the causal relationship between two regions' physiological activity may change under different psychological conditions. This is a psychophysiological interaction. There are many approaches to characterizing such interactions.

The first approach uses a general linear model. The model estimates the condition-specific covariance between the activation of a 'voxel of interest' and all other voxels. A significant difference in condition-specific covariance implies that the remote voxel covaried with the voxel of interest more under one experimental condition than another. This difference in covariance is not affected by the main effect of experimental condition, and is not dependent on an overall correlation between the voxel and the region of interest.

An alternative approach is to assess effective connectivity within a more restricted anatomical model. The anatomical model may be based on known primate anatomy; information obtained from structural neuroimaging studies (see below); or proposed on the basis of distributed activation patterns in functional imaging studies. Effective connectivity may be identified using structural equation modeling (Buechel and Friston, 1997). The model is used to calculate path coefficients (connection strengths) among the regions that best explain the variance-covariance structure of the data. The connection strengths take into account potential common inputs from other areas in the model. The significance of specific connection strengths may be tested by comparing two models, with and without exclusion or constraint of that connection. Connections are significant if their exclusion or constraint significantly changes the overall goodness of fit (measured by the chi-squared statistic, rather than t- and F-tests).

Structural equation modeling is a powerful and flexible tool for image analysis. Models can be used to compare connectivity under different experimental conditions. For example, administration of drugs or changing cognitive context may change the connectivity among regions, without changing overall activity within each region. Models can also be used to describe the contribution of one cortical region to changing connectivity between two others. For example, the covariance between

primary sensory cortex and association cortex may itself covary with parietal or prefrontal cortical activity, consistent with top-down modulation of sensory processing. There are, of course, limitations to structural equation modeling and other analyses of effective connectivity. Although demonstrating the influence of one area on another in the model, they do not prove that that model is the best of all possible models, nor that the connection is direct.

Causal influences are also suggested by the temporal order of activation in different regions, and experimental manipulation of activity by stimulation or lesion studies. Even event-related fMRI currently lacks the temporal resolution to show temporal precedence in distributed neural networks. This may be possible in combined studies, with fMRI and electrophysiological measures such as implanted electrodes, electroencephalograms (EEG) or magnetoencephalograms (MEG). In humans, direct experimental manipulation of cortical activity may be achieved by transcranial magnetic stimulation, TMS.

A caveat to the interpretation of connectivity from neuroimaging is that it refers to an excitatory or inhibitory effect of one region over the other at a systems level. Inferences cannot be made about the excitatory or inhibitory nature of the synaptic links, nor whether the connections are direct or polysynaptic.

STRUCTURAL NEUROIMAGING

Computational Morphometrics

The BOLD signal used for fMRI is based on the T2* MRI signal. In contrast, the T1 signal may be used to achieve very detailed anatomical information about gray or white matter. This may be used to examine subtle but systematic anatomic differences between populations, a process known as computational morphometry (Ashburner and Friston, 2000). It is possible, for example, to compare men and women, young and old, or patients and controls. Any of the information given by the structural brain images may in principle be used, but there are three common approaches.

Firstly, voxel-based morphometry, VBM, which compares the main tissue compartments of the brain. If brain images are placed in a standard anatomic framework, then at any given coordinate there is a certain probability of that voxel being gray matter, white matter or cerebrospinal fluid (CSF). The probability distributions for each tissue type for each subject may be entered into a general linear model, in the same way as functional imaging data above. If two groups are included, then for each voxel one can derive a statistic of the group difference, and thereby an SPM{t} of structural differences between the groups.

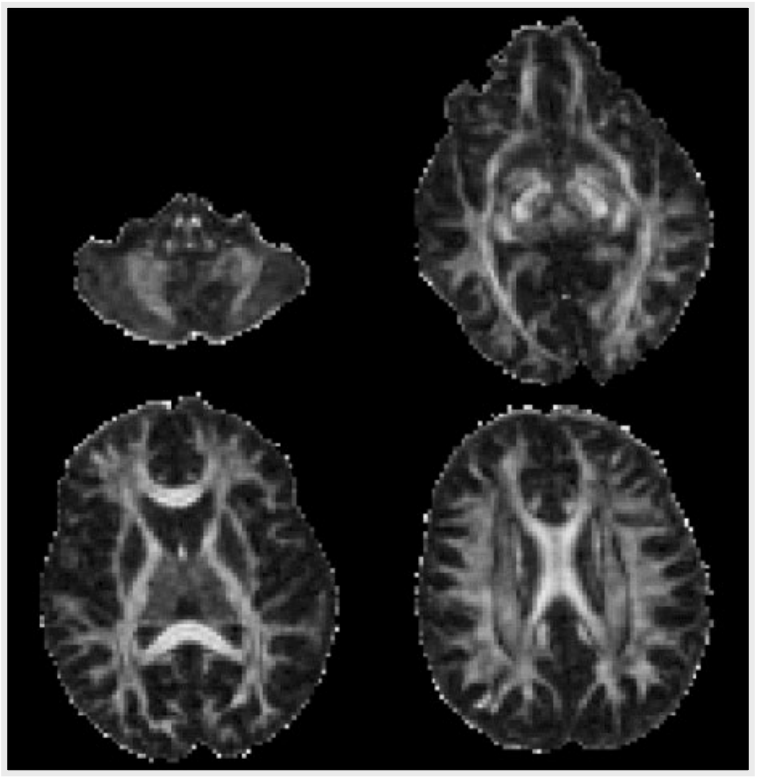
The second approach uses the information derived from the normalization (warping) of each brain image to a standard template image. Multivariate statistics may be applied to the deformation fields, to reveal global differences in brain shape between groups (deformation-based morphometry, DBM). In the third method, spatial derivatives of the deformation fields may be used to reveal significant differences in local shape or structure (tensor-based morphometry, TBM).

VBM, DBM and TBM can each be used to quantitatively characterize and compare structural brain MRI scans. The first to examine local differences in anatomy, the second to examine large-scale differences in shape and the third to characterize local shape differences. For some specific hypotheses, e.g. hippocampal gyral changes in Alzheimer's disease, one technique may be preferred. For other hypotheses, e.g. sex differences in cortical anatomy, these three methods are complementary.

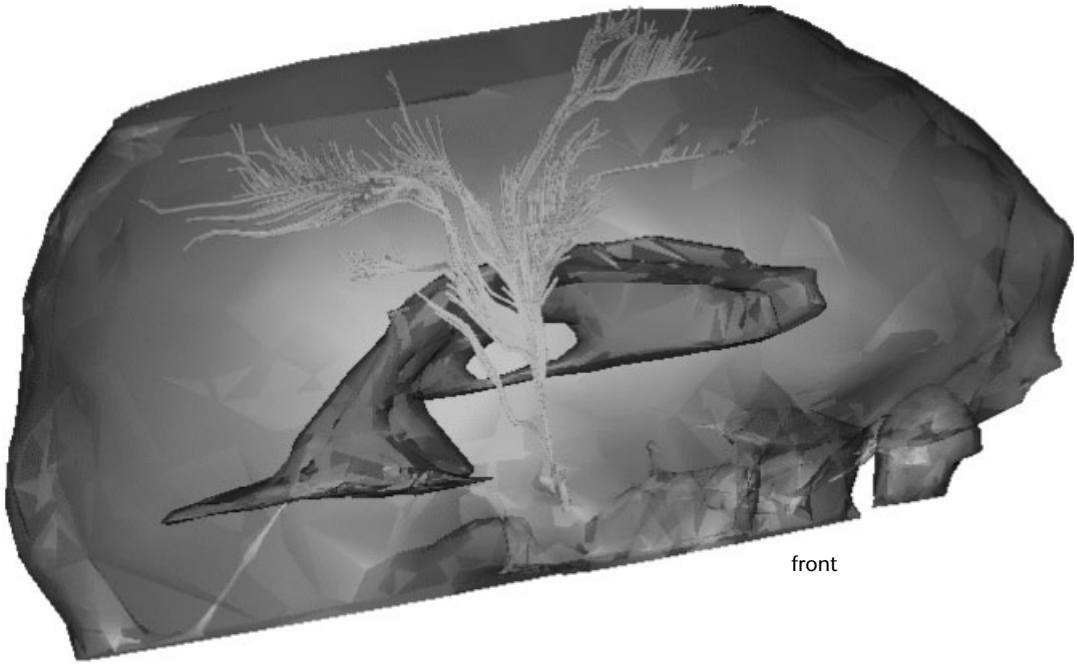
Diffusion Weighted Imaging

In parallel with characterization of functional connectivity by fMRI and PET imaging, methods have also been developed to determine anatomical connectivity. New techniques allow investigation of human brain connections *in vivo*. Several of these techniques depend on the property of anisotropy of water diffusion. Water diffusion in a magnetic field gradient leads to MRI signal dropout. The diffusion of water molecules along axons is greater than diffusion across the axon membranes. In an organized bundle of axons there will be asymmetric diffusion, called anisotropy. Anisotropy is high in normal white matter tracts, and low in gray matter, or disordered white matter.

The anisotropy may be measured for the whole brain, or specific regions, by diffusion tensor imaging, DTI. The diffusion tensor is a kind of matrix that contains information about the magnitude and direction of anisotropy for each voxel of the brain. The anisotropy may be displayed in terms of its magnitude or direction, or both. This provides an immediate impression of interconnected white matter tracts as in Figure 4a. Formal analysis of these images is the basis of tractography, which is the mapping of white matter tracts and nerves. If a tract runs between two voxels, then they will have similar anisotropy. Conversely, if two voxels have similar anisotropy, it is more likely that they



(a)



(b)

Figure 4. [*Figure is also reproduced in color section.*] (a) Anisotropy map through four axial slices of the human brain, with higher anisotropy indicated by lighter gray. The map clearly highlights the organization of the white matter tracts. (b) The anisotropy information may be used for tractography. In this example the likely connections from a 'seed' voxel in the pyramidal tract are shown in green from a right-lateral viewpoint, within a transparent brown cranium, extending upwards and fanning out as part of the corona radiata. The positions of the ventricles are shown in red. Images courtesy of Dr Geoff Parker, Institute of Neurology, London.

lie on the same tract than if the anisotropy is different. It is therefore possible to define how two brain regions are most likely to be connected, by finding the connecting route that minimizes the cumulative difference in tensors. This is known as a 'streamline' approach (Conturo *et al*., 1999). Alternatively, one can start with a 'seed' voxel, and link together neighboring voxels of similar tensors, building up a branching tree of likely connections. The results of this approach are shown in Figure 4b.

FUTURE DEVELOPMENTS

Progress in our understanding of integrated functional brain systems is fueled by improvements in the temporal and spatial resolution of neuroimaging, and the theoretical framework for analysis and interpretation of brain images. This is partly due to advances in the technology of image data acquisition, but also new analytical techniques and experimental designs, such as event-related fMRI. In the future, the analysis of functional imaging may directly incorporate prior information of cortical anatomy and human anatomic connectivity.

The integration of functional imaging with other techniques also promises a fuller characterization of integrated neural systems, with defined effective connectivity, dependent on specific neuropharmacological processes. For example, it will become easier to combine EEG with fMRI data, both by simultaneous acquisition and by mutually informative analysis of activation foci. This combination may offer the temporal resolution of EEG with the spatial resolution of fMRI (Dale *et al*., 2000). The addition of transcranial magnetic stimulation or new pharmacological agents in combination with fMRI or PET, will allow one to study the effects of experimentally controlled manipulation of brain function, to understand the normal human brain and the mechanisms of neurological and psychiatric disease.

Acknowledgements

Both authors are supported by the Wellcome Trust.

References

Ashburner J and Friston KJ (2000) Voxel-based morphometry – the methods. *Neuroimage* **11**: 805–821.

Buckner RL, Bandettini P, O'Craven K *et al*. (1996) Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the USA* **93**: 14878–14883.

Buechel C and Friston KJ (1997) Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex* **7**: 768–778.

Conturo TE, Lori NF, Cull TS *et al*. (1999) Tracking neuronal fibres in the living human brain. *Proceedings of the National Academy of Sciences of the USA* **96**: 10422–10427.

Dale AM, Liu AK, Fischl BR *et al*. (2000) Dynamic statistical parametric mapping: combining fMRI and MEG for high resolution imaging of cortical activity. *Neuron* **26**: 55–67.

Logothetis NK, Pauls J, Augath M, Trinath T and Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**: 150–154.

Menon RS and Kim SG (1999) Spatial and temporal limits in cognitive neuroimaging with fMRI. *Trends in Cognitive Sciences* **3**: 207–216.

Worsley KJ, Marrett S, Neelin P *et al*. (1996) A unified statistical approach to determining significant signals in images of cerebral activation. *Human Brain Mapping* **4**: 58–73.

Statistical Parametric Mapping [http://www.fil.ion.ucl.ac.uk/spm/].

Further Reading

Cabeza R and Kingstone A (2001) *Handbook of Functional Neuroimaging*. Cambridge, MA: MIT Press.

Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ and Mazziotta JC (1997) *Human Brain Function*. London, UK: Academic Press.

Mazziotta JC, Toga AW and Frackowiak RSJ (2000) *Brain Mapping The Disorders*. London, UK: Academic Press.

Toga AW and Mazziotta JC (1996) *Brain Mapping The Methods*. London, UK: Academic Press.

Toga AW and Mazziotta JC (2000) *Brain Mapping The Systems*. London, UK: Academic Press.

Organisation of Human Brain Mapping [http://www.fmri-world.net].

Neuron Doctrine

Introductory article

Leonard K Kaczmarek, Yale University School of Medicine, New Haven, Connecticut, USA

CONTENTS

A simple statement of the neuron doctrine is that neurons are the fundamental units of the nervous system, and that the entire activity of the brain can be understood in terms of the actions and interactions of these cells. While this is generally accepted by current-day neurobiologists, theories of exactly how neurons achieve control over the workings of the brain have undergone multiple modifications since the concept was first introduced.

THE NEURON AS A CELLULAR ENTITY

The neuron doctrine, which was first formulated explicitly in a review article by Wilhelm Waldeyer in 1891, emerged through a debate at the end of the nineteenth century between the neuroanatomists Camillo Golgi and Santiago Ramón y Cajal. Although Golgi had developed a staining procedure that allowed an entire neuron to be visualized, together with its complex axonal and dendritic branches within the brain, his position in the debate was that the nervous system constituted a complex continuous syncytium of tissue. In contrast, Ramón y Cajal championed the concept that each of these 'Golgi-stained' structures with their complex arborizations represented a single functional unit (i.e. a single neuronal cell) (Figure 1). Although Golgi's accomplishments were far-reaching in many spheres of cell biology, and both men received the Nobel Prize for their work, it is Ramón y Cajal's name that is spoken in hallowed tones by today's neurobiologists.

A philosophical implication of the neuron doctrine is that it will become possible to explain fully the workings of the mind through an understanding of the way in which neurons work. Whether this is true or not, there are many more experimentally tractable questions that are related to the role of neurons in the brain. For example, what exactly constitutes 'a functional unit'? What aspects of the anatomical, physiological or chemical make-up of a neuron are central to its function? Over the more than 100 years since the neuron doctrine was formulated, views of how the combined activity of neurons gives rise to simple and complex patterns of behavior, as well as learning and memory, have become modified considerably, and there is no reason to believe that we are yet close to an answer.

If neurons as individual cells are the fundamental units of the nervous system, there must exist

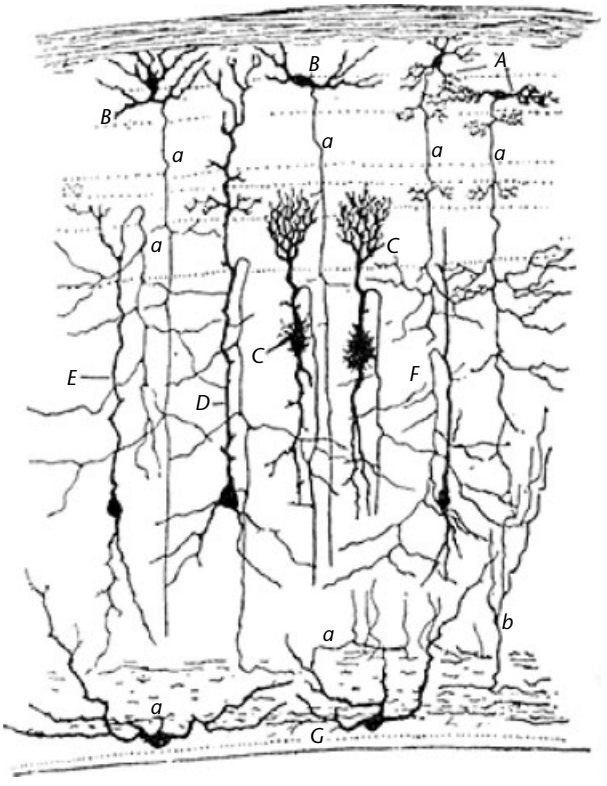


Figure 1. A drawing made by Santiago Ramón y Cajal, using the Golgi staining technique, showing different types of neurons in the optic tectum of a bird. (Reproduced by permission of the Cajal Institute, CSIC, Madrid).

mechanisms whereby they communicate. As is true of all cell types in the body, there is a host of chemical and mechanical ways in which neurons and other cells of the nervous system can influence their partners. These include soluble molecules known as growth factors, which in general instruct cells to grow, divide or even die, as well as molecules on the surfaces of neurons, which influence the properties of neighboring cells by direct physical contact. Nevertheless, the interaction between neurons that is most closely identified with the neuron doctrine is *synaptic transmission*. The term *synapse* was coined by Sir Charles Sherrington at the end of the nineteenth century, close to the time when the neuron doctrine itself was formulated. It refers to the connection between two neurons, usually at the end of the axon of one neuron where it is in close contact with the dendrite of the second neuron. At this site, unidirectional transfer of information occurs from the first neuron to the second one. Sherrington used the concept of synapses to explain how information is passed from a sensory organ to muscles in simple spinal cord reflexes, such as those that produce the withdrawal of a limb in response to strong mechanical stimulation of touch- or pain-sensitive neurons in the skin. It is now known that the synapse is a specialized site at the end of an axon where chemical neurotransmitters are released upon the arrival of a nerve impulse. These neurotransmitters diffuse across a narrow gap between the two neurons, and bind to neurotransmitter receptors on the second neuron. This in turn may excite the postsynaptic neuron, either triggering a nerve impulse in the second cell or increasing the probability that an impulse will occur. Alternatively, a neurotransmitter may inhibit the postsynaptic cell, thereby reducing the probability that a nerve impulse will occur.

NEURONS AS 'ON–OFF' SWITCHES

Some of the earliest studies of synaptic transmission were carried out at two very specialized synapses, namely the neuromuscular junction of vertebrates and the giant synapse of the squid. Indeed, the mechanism of the nerve impulse itself was first discovered as a result of studies of the giant axon of the squid. Transmission across the synapse on to the giant axon eventually results in the contraction of muscles that force water through the mantle of the animal. This allows squid to propel themselves through seawater at speeds that match those of fast automobiles. The occurrence of nerve impulses (action potentials) in both the presynaptic and postsynaptic neuron is 'all or none' – that is, if the neuron is excited sufficiently it generates an impulse whose form is independent of the initial stimulus. The impulse then propagates along the axon to its destination at the axon terminal. The release of neurotransmitter at the synaptic junction produces excitation of the postsynaptic neuron that is sufficient to drive another impulse in its axon. The properties of this synapse are generally similar to those of the neuromuscular junction, where the postsynaptic cell is a muscle cell rather than another neuron, and where the impulse results in contraction of the muscle. Both types of synapse are considered to be very 'secure'.

The mode of action of these two systems proved very influential in theories about how neurons may function within the brain. In particular, a neuron could be considered to be either active (i.e. conducting an impulse) or inactive. According to this line of thinking, the wiring diagram of the nervous system (i.e. the exact way in which the neurons are connected) is the single most important factor determining the way in which the nervous system functions. A variety of analogies could be drawn to provide an intuitive understanding of neuronal function. The wiring diagram of the nervous system has been said to resemble a telephone network. A key aspect of this view is that, because an individual neuron is either on (firing an action potential) or off (silent), knowledge of the biochemical and/or physiological properties of the component neurons is not particularly relevant to an understanding of the way in which the nervous system controls behavior. However, the strength of the synaptic connections between the neurons, and whether they are excitatory or inhibitory, is clearly of paramount importance, and studies of learning and memory have focused primarily on mechanisms that can alter the strength of the synaptic connections between two neurons. Indeed, the current most popular models for the investigation of learning and memory are phenomena termed long-term potentiation and long-term depression, in which specific patterns of stimulation lead to either an increase or a decrease in the strength of synaptic transmission.

According to the view of neuronal activity described above, in which a neuron can be considered to be either 'on' or 'off', binary logic can be applied to the elements of the neuronal network. As a result, many types of computer models have been implemented over the years to mimic the behavior of networks of neurons. While some of these models constituted real progress in engineering

and artificial intelligence, such simple models have so far proved to be spectacularly uninformative with regard to understanding real systems of neurons. This is partly because real neurons do much more than simply act as an 'on–off' switch. While no neurobiologist would dispute the importance of either connectivity or mechanisms that change the strength of synaptic connections, it is now evident that other mechanisms, related to the biological properties of individual neurons, are also key elements of neuronal function (Figure 2).

CONTROL OF BEHAVIOR BY NEUROPEPTIDES

The simple picture of the brain as a carefully wired circuit-board was dealt a very significant blow by the discovery of *neuropeptides*. These are basically a class of neurotransmitters whose function is similar to that of the 'classical' neurotransmitters that are released at synaptic endings, and which excite or inhibit the postsynaptic cells. Chemically they are larger and more complex than the classical neurotransmitters. More significantly, however, the action of these neuropeptide neurotransmitters is not delimited by the synaptic connections of the neuron from which they are released. Even more strikingly, the release of these substances is not usually dependent on the occurrence of single nerve impulses. Typically, a prolonged burst of neuronal firing is required to evoke the release of neuropeptides.

Neuropeptides are thought to be associated with relatively complex animal behaviors, rather than with simple reflexes. For example, they regulate feeding and reproductive behaviors. A very large number of neuropeptides have been identified within the brain, and some of these also function as hormones in other organs of the body. One of the first neuropeptides to have its role in animal behaviors thoroughly investigated was egg-laying hormone (ELH), which is released from certain neurons in *Aplysia*, a marine organism that, like the squid, has been widely used to investigate neuronal function. ELH is released by a train of impulses in the neurons, and brings about a long and complex sequence of reproductive behaviors that culminate in egg-laying. However, no specific synaptic connections are required for ELH to act on its target neurons. Instead, the neuropeptide simply diffuses away from the cells from which it was released until it encounters other neurons that have specific receptors for this transmitter. Another early example of neuropeptide action is that of luteinizing hormone-releasing hormone (LHRH) which, when it is released from certain neurons in rodent brain, brings about lordosis, a stereotyped pattern of sexual behavior. As with other neuropeptides, the neurons that respond to LHRH do not need to be connected synaptically to the neurons that release it.

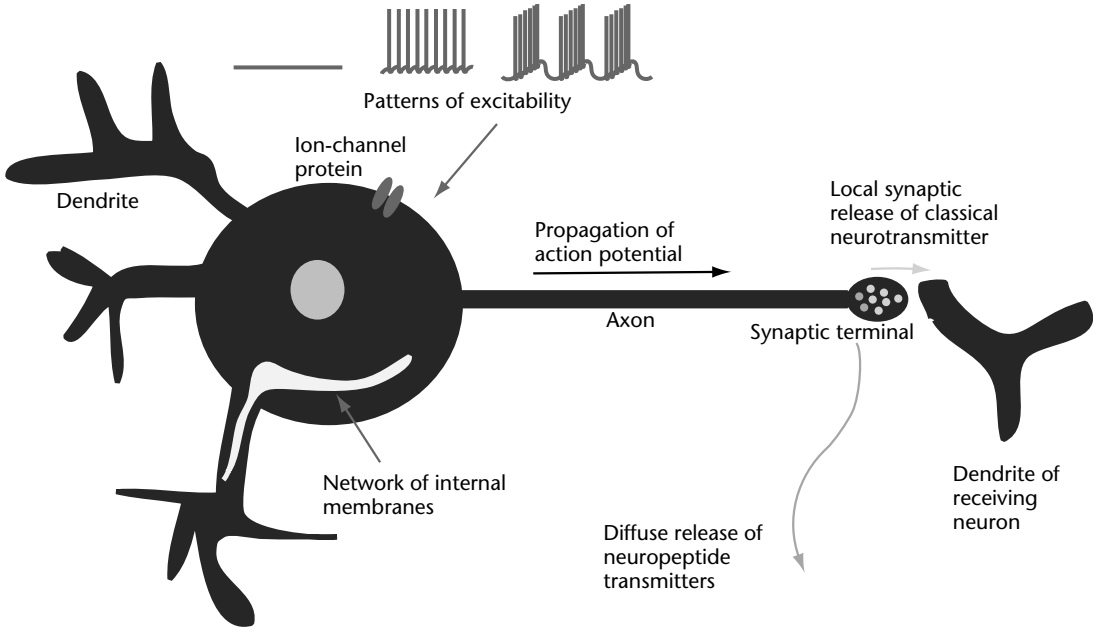


Figure 2. Schematic drawing of a neuron, showing some of the modes of neuronal information transfer described in the text.

Neuropeptides are found within neurons at all stages of sensory and motor processing within the brain, where they generally coexist with classical neurotransmitters. However, communication between neurons via their neuropeptide network does not depend on anatomical connectivity between the neurons. In many of these 'networks' the occurrence or timing of single neuronal impulses has no particular significance; these impulses exist simply as a mechanism to promote the secretion of the neuropeptides. The significant information now resides in which particular neurons bear receptors for the peptides and the exact way in which exposure to the neuropeptide alters the properties of these neurons (see below).

EACH NEURON HAS ITS OWN ELECTRICAL PERSONALITY

Another major modification to the simple 'on–off' view of neuronal function has arisen from a closer examination of the electrical properties of neurons within the central nervous system. These are much more varied than those of the squid giant axon or the neuromuscular junction. Indeed, even the concept of the nerve impulse as 'all or none' has changed. Although the propagation of an action potential down a long axon is indeed an 'all-or-none' event, the action potentials that occur in dendrites, cell bodies or synaptic terminals of a neuron can vary in their shape, height and width. The particular way in which a neuron is stimulated can produce changes in the shape of these action potentials. When these occur at synaptic endings, they alter the amount of neurotransmitter that is released. Changes in action potential shape at the cell body produce biochemical changes that determine how the neuron will respond to future simulation.

The shape of its action potential is not the only electrical feature of a neuron that is subject to change over time. Some neurons are entirely silent unless they receive an excitatory synaptic input from another neuron. In contrast, other neurons are capable of generating long spontaneous trains of action potentials in the absence of any synaptic input. Yet other neurons produce rhythmic bursts of impulses separated by brief periods of silence. Neurons with the latter feature can frequently be found in neuronal networks that generate rhythmic motor outputs (e.g. for locomotion or breathing). The way in which a neuron responds to sustained stimulation of its synaptic inputs is also varied. Some neurons only respond when they are first stimulated, and cease to generate impulses if the stimulation is maintained. This neuronal property underlies the adaptation that occurs in response to many types of sensory stimulation. Other types of neuron fire action potentials as long as they are being stimulated, while yet others continue to generate impulses spontaneously long after the initiating stimulus has ceased.

The most important feature of these electrical properties is that neurotransmitters and neuropeptides produce both short-term and long-term changes in these intrinsic firing patterns. In this case, the complex computational decisions that shape the output of the network are made within the neurons themselves, rather than at specific synaptic connections. The intrinsic electrical excitability of a neuron and the way in which this changes as a result of stimulation are both determined by the complement of *ion-channel proteins* that are found in the membrane of the cell. The relatively recent discovery of a huge number of genes that encode ion channels is testament to the importance of these intrinsic electrical properties. It has been estimated that a neuron selects its ion channels from literally billions of possible combinations.

ARE GLIAL CELLS ALSO NEURONS?

The neuron doctrine arrogantly ascribes the business of the brain to neurons alone. Yet neurons are not the only cell type in the brain. The *glial cells* that surround neurons are in fact far more abundant than the neurons themselves. Even before the neuron doctrine was formulated, glial cells had been named by the anatomist Rudolf Virchow as *neuroglia* or 'nerve glue', with the assumption that they exist primarily to hold the neurons together. Several different types of glial cells exist, and a variety of supporting roles have been convincingly attributed to them. Nevertheless, recent research has also suggested that one class of glial cells, termed *astrocytes*, may participate in extensive cell-to-cell communication. For example, the calcium ion concentration within a cell is known to be a key regulator of many cell functions, and in neurons it directly controls the release of neurotransmitters. Traveling waves involving rapid fluctuations of calcium ions have now been measured in networks of astrocytes that span wide areas of brain tissue. The significance of these waves of glial signaling is not yet known, but speculation abounds that they play a functional role in shaping the output of the nervous system, and that they may even play a role in learning and memory.

NEURONS USE MULTIPLE PATHWAYS FOR SIGNALING

To communicate across the relatively large distances (in cellular terms) between different regions of the brain, and between the brain and peripheral organs, the nervous system developed the mechanism of rapid axonal conduction of nerve impulses coupled to synaptic transmission. However, it is now evident that other mechanisms exist for the flow of information from one part of a neuron to another. These include an extensive array of membranes that lie within the cell itself. Composed of organelles termed endoplasmic reticulum and mitochondria, these membranes form an internal network within a single neuron. Information that is largely mediated by fluctuations in calcium ion concentration may be propagated from one site in a neuron to another across this network without invoking the traditional generation of nerve impulses. Indeed, this concept of extended internal membranes has been dubbed a 'neuron within a neuron'. In addition to this relatively rapid means of communication, slower biological processes that transfer molecules from the cell body of a neuron to its axon or dendrites, and vice versa, also exist. The way in which the nervous system adapts to its environment in the long term almost certainly involves these slower mechanisms of information transfer.

SUMMARY

The neuron doctrine, expressed as the belief that neurons are fundamental units of the nervous system, and that their activity controls both the behavior of humans and animals and the way in which they perceive their environment, is in good experimental shape. The anatomical, biochemical and physiological processes and the organizational principles that allow networks of neurons to do this are being determined by neuroscientists at many levels, ranging from the molecular to the behavioral.

Further Reading

Berridge MJ (1998) Neuronal calcium signaling. *Neuron* **21**: 13–26.

Gold I and Stoljar D (1999) A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Science* **22**: 809–869.

Levitan IB and Kaczmarek LK (2002) *The Neuron: Cell and Molecular Biology*, 3rd edn. Oxford: Oxford University Press.

Shepherd GM (1991) *Foundations of the Neuron Doctrine*. Oxford: Oxford University Press.

Neurons, Computation in

Introductory article

James M Bower, University of Texas Health Science Center, San Antonio, Texas, USA

CONTENTS

Detailed knowledge of the complexities of single-neuron computation – the way in which single neurons process synaptic input – is essential to understanding the function of the brain itself.

INTRODUCTION

Arguably the most famous, and certainly one of the most influential, neuroscientists of all time was the Spanish neuroanatomist Santiago Ramón y Cajal. Working in the later nineteenth and early twentieth centuries, Cajal used an anatomical staining technique developed by the Italian biologist Camillo Golgi that – for reasons that remain mysterious – results in dark staining of only a small number of the many trillions of neurons found in the brain. Cajal spent his entire professional career drawing these labeled cells and carefully describing their sizes, shapes and distributions. An example taken from his studies of the cerebellum is shown in Figure 1. For this work, Cajal was awarded a Nobel Prize in 1906, a prize he shared with Golgi.

Cajal's work established some of the most fundamental assumptions underlying brain science. He demonstrated that the somewhat amorphous-looking brain tissue was actually made up of individual and separate neurons: this is now known as the 'neuron doctrine'. In the nineteenth century and even well into the twentieth, there were many philosophers and scientists, including Golgi, who believed that mental function could not depend on individual cells, but instead required what they called a 'reticulum' of directly interconnected processes. Cajal's claim that each neuron was a distinct entity with processes separated from other neurons by an actual physical space, now known as the synaptic cleft, was hotly debated until the mid-1950s when the invention of the electron microscope finally proved that Cajal had been right. The neuroscience literature is full of papers using modern tools to re-establish conclusions first drawn by Cajal using a simple microscope.

While Cajal's observation turned out to be correct, the neuron doctrine in its full form makes a much more fundamental and important assertion. As formulated in the famous 1891 review of Cajal's work by the German physician Wilhelm Waldeyer, Cajal's efforts led to the conclusion that 'the nerve cell is the anatomical, physiological, metabolic, and

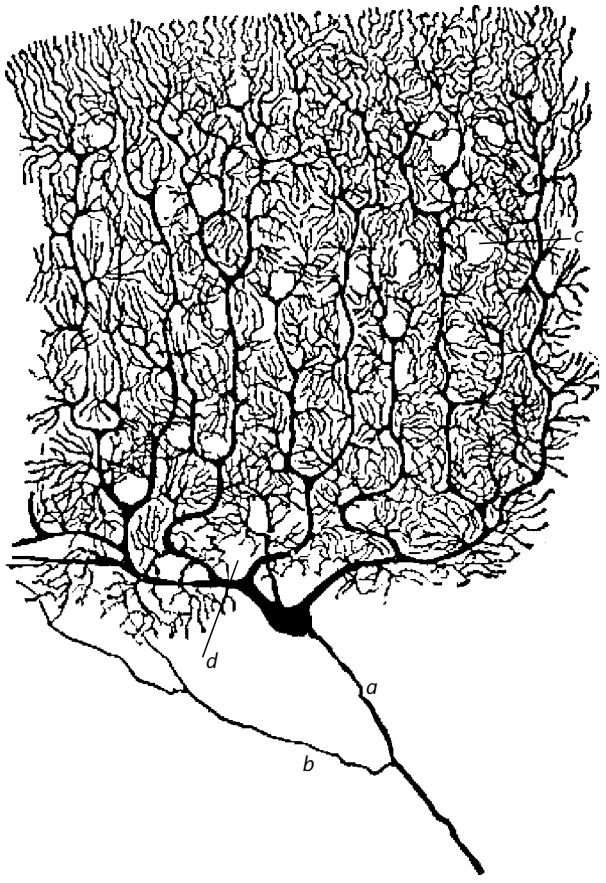


Figure 1. Drawing of a single Purkinje cell made by Ramón y Cajal in the late 1800s.

genetic unit of the nervous system.' Today we would add to that list that the neuron is also the brain's most fundamental computational unit. It is the electrical–chemical interactions within and between neurons that are believed to underlie all behavior, from the simplest to the most complex. Accordingly, the ultimate understanding of brain function will almost certainly depend on an understanding of how neurons act and interact.

INFORMATION PROCESSING IN NEURONS: THE TRADITIONAL VIEW

It can be argued that the development of neuroscience has been characterized by a slow and continuous growth in our understanding of the complexity and sophistication of the brain as a computational device. While this statement applies to all levels of brain studies, from behavior to molecules, it is perhaps nowhere as clear as in the study of the physiological and computational properties of single neurons. The traditional description of the neuron divides each cell into three distinct functional parts: the dendrite (from the Greek *dendron*, or tree), the soma or cell body, and the axon. The classical view of neuronal computation then follows from these divisions, with the dendrite receiving synaptic input from other neurons, the soma determining if the neuron should generate an output, and the axon conveying that signal in the form of an action potential to other neurons. Remarkably enough, Cajal inferred these relationships by looking at the shapes of neurons. At the time, the idea that dendrites received synaptic input, the soma synthesized the input and the axon distributed the results was known as the 'law of dynamic polarization', and remained highly controversial for many years. Once again, however, Cajal turned out to be basically right.

In the years between Cajal's publications and the 1960s, a great deal of attention was directed at efforts to understand the biophysical mechanisms underlying the generation and propagation of electrical signals along the axon. Primarily through the convenient use of a very large axon found in the squid (the 'giant axon'), a detailed understanding of the mechanism underlying this signal, also called the action potential, was obtained by the late 1950s. This effort was capped by a careful series of experiments performed by Alan Hodgkin and Andrew Huxley that led to a mathematical model for the generation of the action potential and resulted in a Nobel Prize in 1963. The mathematics used to describe this biophysical mechanism, known now as the Hodgkin–Huxley equations, remains at the base of most efforts to model neurons today.

To understand the action potential, or any neuronal process for that matter, one needs to realize that neurons are electrochemical devices. Critical to the function of neurons is the fact that their cellular membranes not only isolate the conditions of the inside of the cell from the outside, but also maintain those conditions through the selective and controlled passage of ions and other substances through the membrane. Charged ions such as sodium, potassium, calcium and chloride, which are particularly important to the electrical behavior of neurons, can move rapidly in and out of the membrane through physical channels that open or close under different electrical or chemical conditions. In addition, neuronal membranes contain ion pumps that control the intracellular concentrations of these important ions over the longer term.

In neurons, these membrane properties result in a maintained or 'resting' electrical difference at the soma between the inside of the cell and the outside of, on average, 65–75 millivolts, the inside of the cell being more negative than the outside. During the generation of an action potential, this resting membrane potential rapidly reverses polarity, with the inside of the cell briefly becoming more positive than the outside. Hodgkin and Huxley showed that this membrane potential 'depolarization' is driven by the rapid movement of positively charged sodium ions into the neuron through membrane channels specific for sodium. Under normal steady-state conditions, these channels are closed. When the channels are opened, positively charged sodium ions rush into the cell, driven by the difference in electrical potential (they are positive, the inside is more negative), as well as the fact that an ion pump maintains ten times less sodium inside the cell than outside. Observationally, once the action potential reaches its peak voltage (usually within a millisecond or less), membrane potential returns fairly rapidly to the resting level, with the inside once again more negative than the outside. Analysis by Hodgkin and Huxley demonstrated that this 'repolarization' of the membrane is driven by the movement of potassium ions out of the axon through their own set of membrane channels. Potassium moves out of the cell, following its own concentration gradient, because the same ion pump that maintains low levels of intracellular sodium maintains high levels of intracellular potassium. The flow of positively charged potassium to the outside of the cell leaves the inside more negative.

What Hodgkin and Huxley demonstrated was that the key to understanding this sequence of events lies in the factors controlling the opening and closing of the sodium and potassium channels. It turns out that sodium channels are sensitive to membrane potential, with a higher probability of opening as neuronal membranes are depolarized. As the membrane depolarizes, more sodium channels open, leading to an influx of positively charged sodium ions, producing more depolarization, causing more channels to open, and so on. It should therefore be easy to see why an action potential is a very fast, even explosive event. The return of membrane potential to the resting state is similarly dependent on the control of channel opening and closing. In this case, after opening in response to membrane depolarization, sodium channels automatically close after a short period of time. The return of the membrane to resting levels, however, is due to the voltage-dependent opening of potassium channels, which also open with membrane depolarization after a delay. Taken together, these channels result in a very rapid change in membrane potential, followed by a relatively rapid return to baseline conditions.

The feature of the action potential that is critical to its function in neurons is its seemingly automatic propagation down the axon once initiated. This propagation is also a consequence of the voltage dependence of the sodium channel. Specifically, depolarization at the beginning of the axon, driven by the sodium channel, results in a passive spread of electric current to adjacent axonal regions and the depolarization of that membrane. Sodium channels in these adjacent regions then start to open, further depolarizing the membrane, leading to depolarization of the next part of the axon, the opening of sodium channels there, and so on. In this way the action potential propagates down the axon, signaling the output of one neuron to many others.

NEURONS AS SYNAPTIC SUMMING DEVICES

The discovery that the sodium channel is sensitive to depolarization of the membrane also provided a ready and simple explanation for the influence of synapses on neuronal output. In general, synapses are found at the junction between the axons of one neuron and the dendrites or soma of another. When an action potential in one neuron reaches a synapse, neurotransmitters are released into the gap between the neurons, diffuse across that gap and influence another set of membrane ion channels in the postsynaptic neuron. These channels, which are different from the sodium and potassium channels already mentioned, open to allow positively charged ions (sodium and calcium, for example) into the postsynaptic cell. These positively charged ions result in a local depolarization of the postsynaptic cell membrane.

While brain function would be much easier to understand if a single synaptic depolarization of the postsynaptic membrane led directly to the initiation of an action potential in the next neuron, even early neurobiologists realized that a single input producing a single output did not allow much room for computing. Instead, action potential generation usually requires input from more than one synapse, and it is here that most neurobiologists believe the real core of neuronal computation is to be found. Not surprisingly, it is also here that one finds the greatest number of mysteries about how neurons work.

In the classical view of neuronal computation, shared by a remarkable number of neurobiologists and theorists even today, the fundamental operation performed by the dendrite of a neuron is a spatial and temporal summing of synaptic input. The idea is that sufficient synaptically driven depolarization of the postsynaptic membrane will lead to the initiation of an action potential. In its most common theoretical form, this type of cell is referred to as an 'integrate and fire' neuron. Such a neuron combines its synaptic input (usually by simple summation), and then 'fires' an output when a certain threshold is reached. This type of cell has formed the basis for many models of brain function, as well as most neural network models of information processing and cognition. It is also the most common description of neuronal computation found in textbooks even to this day.

While it is remarkable how many interesting models can be constructed from such simple neurons, growing evidence suggests that, in fact, many if not most neurons in the brain are not performing anything like a simple summation of synaptic inputs. Further, it is now clear that different types of neuron treat their synaptic inputs very differently. How neurons actually 'integrate' synaptic input is, at the moment, a very active and contentious area of research. The truth is that we are probably still very far from any real understanding of the true complexity of neuronal computation. The following sections will consider several features of neurons, currently largely ignored in higher-order theories and models to date, that are likely to be important in an eventual understanding of real neuronal computation.

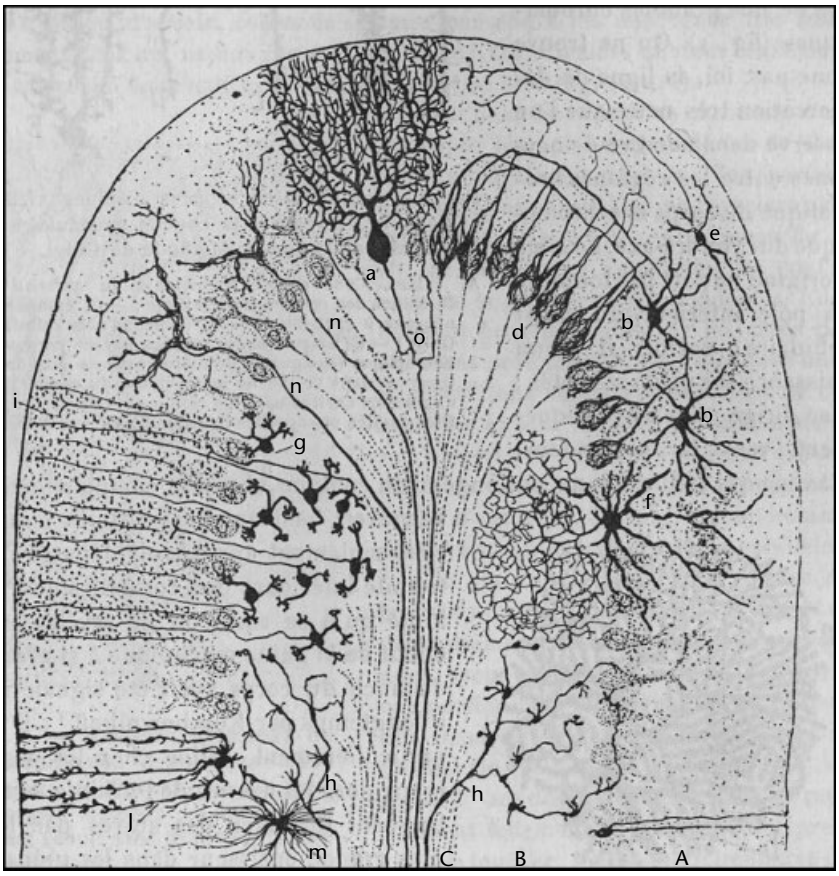


Figure 2. Drawing by Cajal of the different types of neurons found in the cerebellum.

NEURONAL FUNCTION IS RELATED TO NEURONAL SHAPE

When Cajal looked at his first stained tissue, he must have been amazed at the diversity of neuronal shapes, especially those of the dendrites. Considering either a single brain structure, or the brain as a whole, neurons come in a bewildering number of shapes and sizes. Figure 2 is an example of the cell types Cajal described in the cerebellum. As we have already noted, he believed that the shape of a neuron and especially of its dendrite directly reflected its function. In fact, the dendritic shapes and sizes he observed served as the basis for many of his and his contemporaries' speculations on how these neurons worked.

What is remarkable to note in the history of neuroscience is that from Cajal's day until very recently, many if not most theorists have simply ignored dendritic morphology. Instead, most discussions of neuronal computation have focused on the mechanisms of action potential generation. In fact, well into the 1960s many prominent neurobiologists believed that synapses on the distal dendrites of neurons had very little influence on neuronal output. Sir John Eccles (who shared the 1963 Nobel Prize with Hodgkin and Huxley for his work on the spinal cord) and others argued that the small depolarizations associated with such far-away synapses would dissipate as a result of the internal electrical resistance of the cells long before they could influence action potential generation. This assumption, of course, made it much easier to speculate about brain function.

In the 1950s, a physicist named Wilfrid Rall left the Manhattan project intent on applying physical principles and quantitative modeling techniques to neurobiological systems. Rall, who actually trained with Eccles, suspected that the complex geometries of neuronal dendrites held the key to neuronal computation. Taking advantage of his mathematical and analytical skills, Rall developed modeling tools that allowed, for the first time, quantitative studies of the consequences of dendritic morphology for synaptic integration. In a 15-year battle with Eccles and others, Rall applied techniques intended to predict how electrical signals would propagate through long-distance telephone cables

to show that dendritic synapses could, in fact, exercise considerable influence on action potential generation. He demonstrated not only that dendritic synapses have an effect on somatic output, but also that the complex geometry of neuronal dendrites provides a rich opportunity for complex computation. Using this approach, he began to study the interaction of excitatory and inhibitory synapses, a subject that is only today beginning to be taken seriously by many neural modelers. Through this pioneering work, Rall and his students confirmed Cajal's suspicion that the shape of a neuron had a profound effect on its function. While we are only at the very beginning of understanding this relationship, it is very likely that the geometrical complexity of a neuron's dendrite is a direct reflection of the complexity of its computation.

NEURONS ARE NOT PASSIVE CABLES

The discussion in the 1950s and 1960s about the role of dendrites in neuronal computation, as well as Rall's early modeling efforts, assumed that neuronal dendrites were passive electrical cables and did not themselves contain ion channels of the sort Hodgkin and Huxley had shown produced somatic action potentials. Over the last 20 years, it has become increasingly clear that many if not most dendrites are not passive but 'active'. Dendrites are now known often to have a bewildering set of ion channels responding not only to changes in membrane voltage but also to changes in the chemical state of the neuron or its surroundings. It has become equally clear that these active processes result in much more complex synaptic interactions than synaptic potential summation. Even studies of the ion channels underlying the action potential have made it clear that the idea of a simple threshold for neuronal activation is a considerable oversimplification of the real situation. In reality, there is no clear threshold, but instead the generation of an action potential results from the complex interplay of the probabilities of opening and closing of sodium, potassium and other ion channels.

While Rall himself suspected early on that the active properties of neurons would substantially change their computational properties, even his conclusions based on passive dendrites were controversial, and too far out of the mainstream to be taken seriously by many of the neurobiologists of the time. Nevertheless, in the early 1960s Rall introduced a new modeling technique that allowed single neurons to be represented in their full anatomical and physiological complexity. By breaking the cell and its dendrite up into many small compartments, it was possible to study the effects of arbitrary dendritic geometry and non-uniform ion channel distributions on neuronal computation. An example of this modeling approach is shown in Figure 3, which represents part of a compartmental model of a Purkinje cell, a type of nerve cell that occupies the middle layer of the cerebellar cortex. Rall's approach now provides the basis for modeling systems such as GENESIS and NEURON in which computational neuroscientists are discovering all kinds of new computational properties by constructing anatomically and physiologically realistic single-neuron models. Using these techniques, we are slowly realizing, once again, that our simplified notions of computation in neurons are going to have to change dramatically.

THE NEURON IN THE CONTEXT OF ITS NETWORK

The detailed analysis of neuronal anatomy and physiology allowed by compartmental modeling techniques has also made it clear that to study the computational complexity of any particular neuron, one must place that neuron in the context of its network. Thus, as Cajal suspected, and took great pains to illustrate (Figure 4), it is both the detailed structure of the neurons and the detailed geometry of their networks that determines how they compute. It is clear that neurons are not devices designed to receive arbitrary patterns of synaptic input. Instead, their structure anticipates the kinds of input that they naturally receive from their network. Slight changes in the spatial and temporal pattern of synaptic inputs can produce widely different neuronal output. This is especially true when the interaction of excitatory and inhibitory synaptic inputs is considered. For this reason, to understand the computational properties of an individual neuron, one will almost certainly need to construct an equally realistic model of the circuit to which it belongs. In this way, modeling at the neuronal level inevitably leads to modeling at the network level. Whether it will be possible to extract general principles from such models, or whether the computation will always be in the details, is yet to be seen. It seems certain, however, that oversimplifications of neuronal or network structure at the outset are likely to result in essential computational features of both being overlooked.

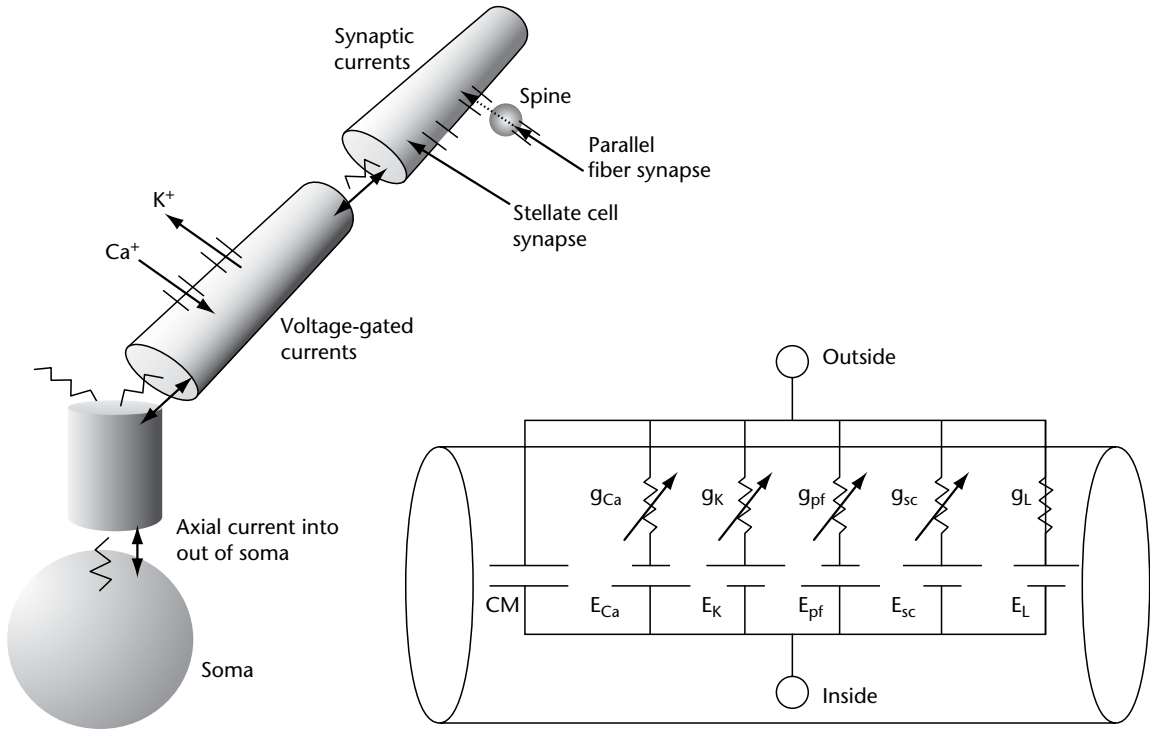


Figure 3. Representation of a neuron within a compartmental model. As shown on the left, the neuron is divided into a series of small compartments. On the lower right, the electric circuit diagram illustrates the way in which the electrical properties of membranes and ion channels are represented in each compartment.

EXAMPLE: A NEW PURKINJE CELL MODEL

In order to illustrate the potential complexity of single neuron computation, we will conclude with a brief description of an anatomically and physiologically realistic model of a Purkinje cell developed and studied by the author and his students over the last twelve years. As shown in Figures 1, 2 and 4, Cajal provided a remarkably detailed description of the morphological features of the cerebellar Purkinje cell, as well as its relationship to other neurons in the cerebellum. From anatomical analysis such as Cajal's, we know that the surface area of this cell's dendrite is one of the largest in the mammalian brain. It also has a remarkable geometrical shape, being flattened in one dimension like a fan. The cerebellar circuitry takes full advantage of the large Purkinje cell dendrite, providing it with between 100 000 and 200 000 synaptic inputs, believed to be the largest number for any neuron in the brain.

As befits a neuron of this size and complexity, we have modeled the neuron using almost 2000 individual compartments, several of which are shown in Figure 3. In these compartments we have included the electrical properties of the dendrite, such as its capacitance and resistance, as well as what information is available on the active properties of the membrane. Through the pioneering electrophysiological studies of the Purkinje cell by Mutsuyuki Sugimori and Rodolfo Llinas, we know that this cell's dendrite includes many different types of voltage-dependent ion channels and, in particular, has a very high concentration of voltage-sensitive calcium channels. From modeling and experimental work, we estimate that the current produced through these channels can be five times as large as that produced by the cell's hundreds of thousands of synaptic inputs. We also know from the work of Sugimori and Llinas that the ion channels in the soma are more complex than is assumed in most simple models of neurons. In addition to the standard sodium and potassium channels associated with the action potential, there are several other channels that also influence action potential generation. In fact, our model predicts that the soma of the Purkinje cell can generate action potentials on its own, without any synaptic input.

Given the complexity of the anatomy and physiology of this neuron, the construction of a model is the only way in which its function can be understood. The process used to construct this type of

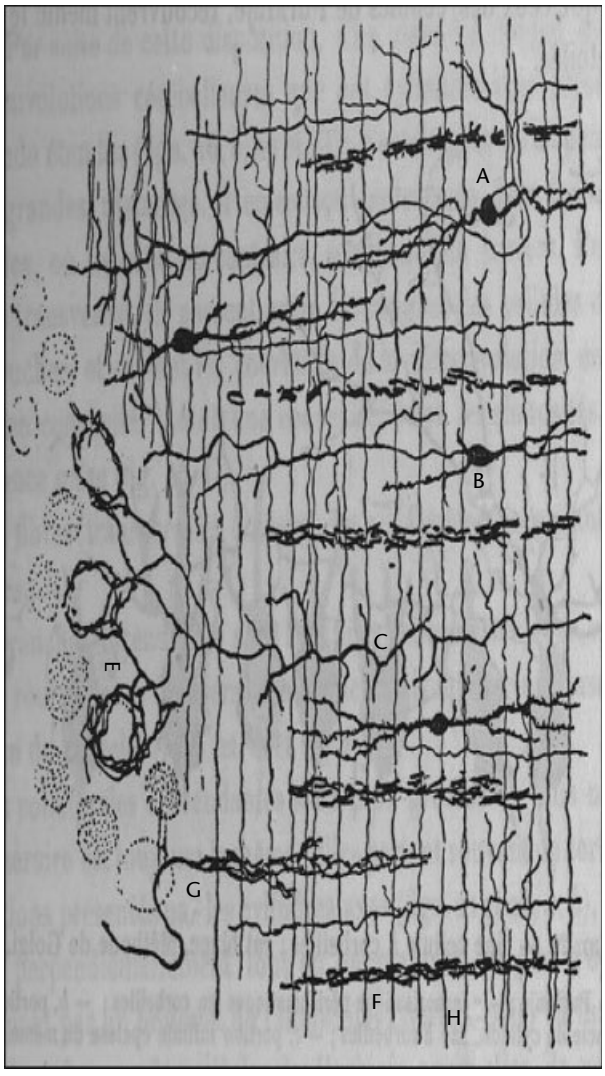


Figure 4. Drawing by Cajal of one aspect of the circuitry of the cerebellum. The cell labeled F is the dendrite of a Purkinje cell through which the long axons called parallel fibers are running.

model is as follows. First, one obtains a complete and anatomically correct description of the Purkinje cell. The cell is then divided into compartments following rules first established by Rall. The passive electrical properties of the model (membrane resistances and capacitances, for example) are then adjusted until they match the experimental data. At this point, mathematical equations that represent ion channels are added and adjusted until the model replicates real physiological responses. Typically, the data used to tune ion channel properties are obtained experimentally by injecting electric current directly into the cell's soma and looking at its electrical response. This technique, called voltage and current clamping, was used by Hodgkin and Huxley to understand the action potential and is very useful in assessing the basic electrical properties of the neuron. Once the model replicates these patterns of activity, model parameters are fixed, and synapses attached to the cell to start investigating model responses to synaptic input. The built-in realism of the model allows modeling results to be compared directly against experimental data, often contributing to the design of new experiments. While easy to describe, this iterative process is still ongoing after almost 12 years for the model described here.

Consequences of Dendritic Calcium Ion Channels for Purkinje Cell Function

It is asserted above that the presence of voltage-sensitive ionic membrane channels in dendrites has a profound influence on the processing of synaptic information. This point will be illustrated here by briefly considering the response of the Purkinje cell model to two of the different types of synaptic input.

One of the interesting and historically puzzling features of electrical activity in the Purkinje cell is that the soma of this neuron generates both classical action potentials and an unusual burst of activity known as a complex spike (bottom right graph in Figure 5). It was Llinas, working as a graduate student in Eccles' laboratory in Australia, who first demonstrated that the complex spike was associated with a special type of synaptic input arising from a single axon that wound its way around the central stalk of the Purkinje cell dendrite. We now know that each Purkinje cell is contacted by only one of these axons, known as a climbing fiber. That one axon, however, makes several hundred synaptic contacts with the central region of the Purkinje cell dendrite.

Figure 6 shows a computer simulation of the response of the Purkinje cell to activation of this one axon and its several hundred synapses. The top set of images of the modeled Purkinje cell shows changes in membrane voltage, while the bottom set shows changes in intracellular calcium concentration. Instead of a simple summing of postsynaptic potentials, the synaptically induced activation of calcium channels in the Purkinje cell dendrite results in an explosive calcium-dependent depolarization of the dendrite beginning in the central location of the climbing fiber synapses and extending rapidly out to the farthest tips of the dendrites. Climbing fiber synapses do not even extend to these regions of the dendrite. This enormous depolarization of the dendrite results in the generation

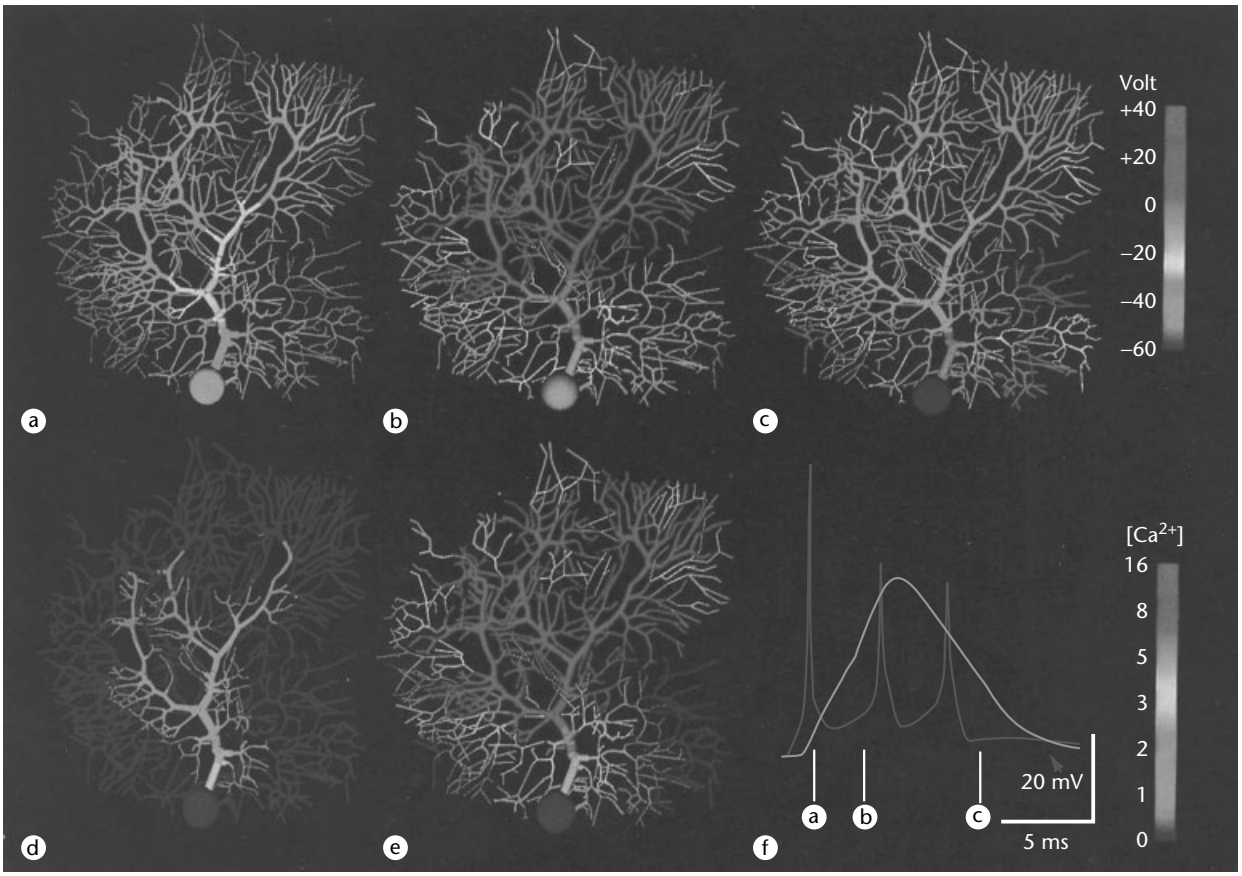


Figure 5. [*Figure is also reproduced in color section.*] Simulation results from climbing fiber activation of the Purkinje cell model. The upper images (a–c) use color to represent membrane potential, while the lower images (d, e) represent intracellular calcium concentration. Parts (b) and (e) were obtained at the same time and 2.5 milliseconds after parts (a) and (d). The voltage record shown in part (c) was obtained 7 milliseconds later than (b) and (e). The graph at the lower right (f) shows changes in membrane voltage in the dendrite (green trace) and at the soma (red trace) in response to climbing fiber activation and resulting from the active properties of the Purkinje cell dendrite. Figure reproduced with permission from De Schutter and Bower (1994) *Journal of Neurophysiology* **71**: 401–419.

of a burst of action potentials in the soma of the cell, as shown by the blue trace in the graph at the bottom right. In contrast to an integrate-and-fire neuron, the presence of active voltage-dependent channels in the dendrite of the Purkinje cell results in a single climbing fiber axon producing a massive depolarization of the dendrite, followed by a burst of generated action potentials rather than only one.

In addition to demonstrating the complex properties of the Purkinje cell dendrite, the modeling results in Figure 5 also indicate the importance of considering neurons in the context of the circuitry in which they are found. The climbing fiber input is a unique feature of the cerebellum, and it is clear that the anatomical and physiological structures of the Purkinje cell are organized accordingly. However, the results shown in Figure 6 illustrate that this cell's response to even a more traditional type of synaptic input is also profoundly affected by the active properties of its dendrite.

The results in Figure 6 were obtained by synchronously activating a small number of synapses on the most remote dendrite of the model. As discussed previously, before the work of Rall there was considerable controversy as to whether synapses this remote from the soma would have any significant influence on neuronal output. In the results shown here, we can see that activation of this dendrite is very effective in producing a somatic membrane depolarization. Analysis of the model predicts that the same dendritic calcium channels responsible for the explosive response of the Purkinje cell to climbing fiber activation are also responsible for this more subtle and constrained effect. In response to the climbing fiber input, these channels drive an explosive entry of calcium into the cell; whereas in this second case they serve

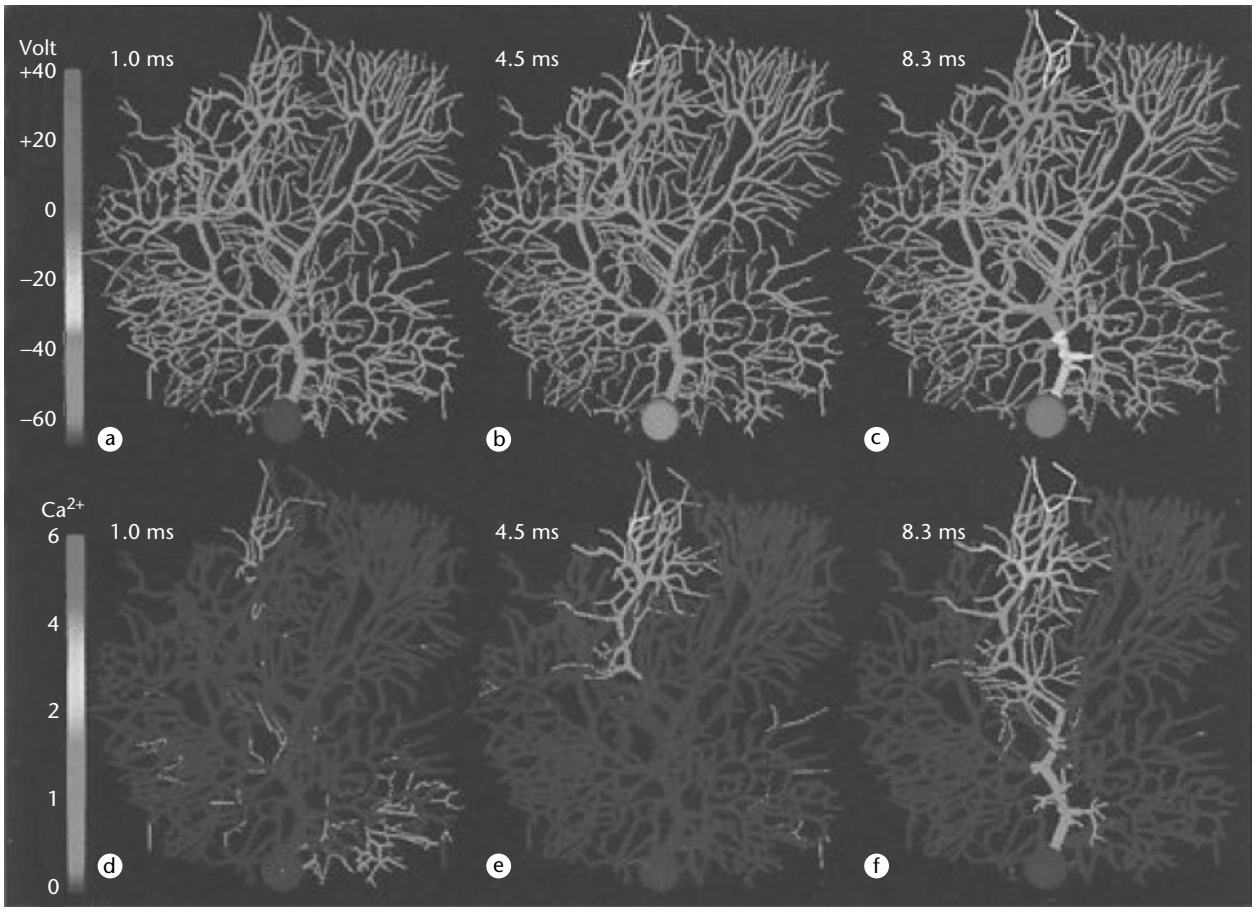


Figure 6. [*Figure is also reproduced in color section.*] The response of the Purkinje cell model to synaptic activation distant from the cell's soma. As in Figure 5, the upper images (a–c) use color to represent membrane potential, while the lower images (d–f) represent intracellular calcium concentration. Each set of images is labeled to indicate the time after the onset of synaptic activation at which they were obtained. This illustration shows the development of a calcium-dependent amplification mechanism that appears to assure that synapses some distance from the soma can still influence action potential generation. Figure modified from De Schutter and Bower (1994) *Proceedings of the National Academy of Sciences of the USA* **91**: 4736–4740.

to amplify the synaptic depolarization and route its influence to the soma. The difference has to do with the detailed distribution of the channels, the nature of the synaptic activation, and the detailed anatomy of the Purkinje cell dendrite. These relationships would not be apparent without a model.

While our study of this model to date has revealed many interesting new features of this neuron, the computational significance of these behaviors is more elusive. It is already clear that to elucidate this question, we must construct an equally detailed model of cerebellar circuitry. Whatever the Purkinje cell is computing, the detailed anatomical and physiological characteristics of its dendrite will be key, and hence realistic models are necessary. Given the dependence of the function of the brain on the computations performed by single cells, our understanding even of cognitive function may ultimately depend on analysis at this level of resolution.

Further Reading

Bower JM (2002) The organization of cerebellar cortical circuitry revisited: implications for function. In: Highstein S and Thach T (eds) *Recent Developments in Cerebellar Research*. New York: New York Academy Press.

Bower JM and Beeman DB (1998) *The Book of GENESIS*, 2nd edn. New York: Springer-Verlag.

Segev I, Rinzel J and Shepherd GM (eds) (1995) *The Theoretical Foundation of Dendritic Function*. Cambridge, MA: Bradford Books.

Shepherd GM (1991) *Foundations of the Neuron Doctrine*. New York: Oxford University Press.

Neurons, Representation in

Introductory article

Peter Földiák, University of St Andrews, St Andrews, Scotland, UK

CONTENTS

The brain uses the activity patterns of a large number of neurons to represent information about the world. The way in which information items are encoded in these patterns has fundamental consequences for information storage, recall, generalization and learning.

INTRODUCTION

A fundamental scientific problem concerns how to account for mental activity in terms of information processing in the brain, and how such information processing is in turn implemented by the functioning of nerve cells. The problem can be broken down into two distinct but interrelated questions. (1) How is information represented in the brain? (2) How are these representations transformed and manipulated in a way that eventually results in evolutionarily appropriate behavior? This article will focus on the first of these questions, with an emphasis on some general principles.

There are infinitely many ways of choosing a physical representation for a piece of information. For instance, numbers can be represented by notches on a piece of wood, by the positions of beads on an abacus, by ink marks on a piece of paper or by electrical voltages in a computer's memory. All of these representations are equivalent in the sense that the same meaning can be read out from them, and each item can be distinguished from other items. However, the representations differ significantly with regard to which operations are easy and which are difficult to perform using them. Addition on an abacus is easy, whereas multiplication using roman numerals on paper is difficult. Representations may vary with regard to what aspects of the items are made explicit in the representation (i.e. can be read out directly) and what other aspects are implicit (i.e. can only be calculated from it). These considerations will also be relevant when considering the encoding of information in the brain. Representation in the brain is determined both by the nature of the available biological 'hardware' and by the characteristics of the tasks that need to be performed.

Most computers today have one highly complex processing unit that is capable of performing a long series of complicated operations at high speed, receiving and sending out complex messages as inputs and outputs. In contrast, the brain's biological hardware consists of an enormously high number (approximately 10^{11} in humans) of relatively simple processing units, namely the neurons. Each neuron has a relatively simple range of internal states (such as 'inactive' or 'active') in comparison with the complex internal states of the central processing unit of a computer. An astronomical number (of the order of 10^{14}) of highly specific, adaptable connections between the units play a fundamental role in the computation, while the message transmitted on each individual connection is a sequence of electrical impulses that can also be regarded as simple by comparison with the messages that are sent within a conventional electronic serial computer. The millions of neurons, on the other hand, generate their outputs and send them to specific targets simultaneously, in parallel, thereby giving the system its computational power. Some of these fundamental functional properties can be abstracted in mathematical models and simulated on conventional electronic computers. These models, which are known as artificial neural networks, can give us some basic understanding of the nature of computational style and the capabilities of such highly, specifically and adaptively interconnected networks of simple units. Finally, the brain also differs from current general-purpose electronic computers in that although the system as a whole can exhibit highly flexible behavior, its processors are dedicated to specialized tasks. The central nervous system is organized into special-purpose areas and modules,

which are specialized and optimized by natural selection for representing certain aspects of the outer world and solving specific computational problems. Within these dedicated areas, each neuron is specialized even further. For example, the brain area specialized for processing touch information may have a neuron dedicated to signaling only when a light pressure is applied to the inner surface of the middle section of the left hand's ring-finger. The precise tuning of each neuron may change gradually to some extent over longer periods of time, but the neuron's response properties remain largely constant during a computation. In some early sensory areas at least, we may think of a neuron as a 'labeled line' with a fixed description on the label, while the activity of the neuron would indicate whether the label is applicable at the time. However, the appropriate description on the labels can often be difficult to formalize, and the labeled line analogy for many types of neurons may even break down completely.

The sensory system, and the visual system in particular, is perhaps the part of the brain that is most accessible to experimental study, as at least the stimuli that enter it are under the experimenter's control. Most of our knowledge of neural representation comes from sensory brain areas, and the rest of this article will focus on these areas.

In order to make sense of sensory representations in the brain, we need to consider the tasks that the sensory system has to solve. One of the main tasks is to reconstruct the causes of the physical signals that are provided by sensory receptors in terms of the significant objects and events in the environment. Light receptors can measure the pattern of light reflected from the surfaces of objects, and mechanical receptors in the auditory system can provide signals about the vibrations propagating through the air (i.e. the sounds that an object makes), but the relationship between these physical signals and the real objects and their properties is usually extremely complex. Why is the apparently simple and effortless act of recognizing a friend on the street computationally difficult? The reason is that the relationship between the identity of a friend and the pattern of light reflected from a face and falling on the retina can be extremely complicated, in view of the unlimited number of ways in which this pattern can change due to variations in the illuminating light, the position of the person in the visual field, distance, various occlusions, shadows, rotations and shape changes of the face itself due, for example, to altering facial expressions. Almost any simple measure of similarity can be smaller between different images of the same person than between images of this person and those of a very large number of other individuals whose facial surfaces differ only in tiny details. Such problems of recognition exist not just for vision but also for hearing and other modalities as well. In addition to simple pattern recognition of familiar objects or people, a more interesting problem is the classification of and appropriate response to novel, unfamiliar objects or situations. In these cases some type of description of the scene is necessary. A novel situation may show some resemblance to a previously encountered situation in certain respects, and it is not just the degree of similarity to an earlier situation, but also the nature or aspect of the similarity, that should determine the appropriate behavior in the new situation. It is these aspects of similarity that should form the basis of a high-level representation, making the 'meaning' or real-world causes of the stimuli explicit. Our goal is to understand how such a representation can be constructed, using the known neural machinery (mainly in the cerebral cortex of the brain), from the physical signals that come from the sensory receptors.

TUNING CURVES AND RECEPTIVE FIELDS IN DIFFERENT CORTICAL AREAS

Our most direct knowledge of the way in which neurons process information comes from neurophysiological experiments. Despite the huge number of neurons in the brain, it is possible to measure and record the electrical impulses from single nerve cells within the functioning neural system using microelectrodes. The single-cell recording method was developed in the 1950s, and has since been applied systematically to explore the relationship between stimuli and neural responses (or 'response properties') of neurons in different areas of the cerebral cortex.

The neural response in all cortical areas consists of electrical impulses (also called action potentials) of similar amplitude and shape. However, the timing and number of these impulses vary, and a large part of this variation is caused by changes in the stimulation that the neuron receives. The most commonly considered statistic of such a sequence of impulses is the total number of impulses or the rate of the impulses. Other measures of response, such as the time of appearance of the first impulse after a new stimulus, can also convey additional information in neurons.

Each neuron responds above its resting rate only when the stimulus has certain critical or 'trigger'

features. For instance, the earliest stages of visual processing take place in the retina, and in frogs this processing involves quite specific 'trigger features'. As frogs are interested in catching small insects, their retina has neurons that respond only when a small dark spot moves across a certain part of the visual field on a light visual background, just as would happen if a small bug was to fly across the frog's visual field against the background of the sky. Consequently, these specialized cells were called 'bug detectors'. Higher animals have retinas with less specialized retinal trigger features to allow more flexible and complex processing in the higher stages of the visual system.

After finding the critical stimulus features of a sensory neuron, the most interesting question from the viewpoint of the neural representation is how the neuron's response varies with various changes in the stimulus. In most experiments, the experimenter chooses to vary some properties or parameters of the stimulus, and then measures the changes in the neural response as a function of the stimulus parameters. The resulting function is called a tuning curve. The tuning curve often has a single peak at the point corresponding to the 'preferred' stimulus (i.e. the stimulus parameter that evokes the strongest response). The value of the tuning curve falls off as one moves away from the peak. If the tuning curve falls off sharply around this point, the neuron is considered to be narrowly tuned to this parameter; otherwise it is considered to be broadly tuned. The width of the peak can provide a measure of the breadth of tuning.

Another concept that is frequently used to describe neural representation is the receptive field. The receptive field of a sensory neuron is the space within the receptive sheet in which it transduces stimuli. For example, the receptive field of a somatic sensory mechanoreceptor (for touch) is the portion of the skin that is directly innervated by the receptor terminals. For a visual sensory neuron, the receptive field is the region of the visual field in which stimuli affect the activity of a neuron. The concept of the receptive field originates in studies of retinal ganglion cells. The trigger features of these cells are small spots of light (or dark spots), and the main parameter to which these neurons are tuned is the spatial position of the spot. Each ganglion cell receives excitation from a small group of photoreceptors, and receives inhibition from receptors in the surrounding small region (or vice versa). These are the excitatory and inhibitory zones of the receptive field. Such a center–surround organization means that these cells will not be optimally activated by a uniform field of light, but only by some change in the light intensity within their receptive field. The concept of a receptive field is clearly useful in the retina, as the responses of these ganglion cells to any stimulus can be predicted by the position of the light spot, or by the distribution of light across the receptive field. If it is inside the excitatory zone, a positive response can be expected, so here the receptive field refers mainly to a spatial region. However, even in the retina this concept starts to refer to more than just a spatial region, as it also begins to refer to the pattern of optimal stimulation within this region.

Tuning curves in the early stages of processing can be determined by the physical properties of the receptor cells. For instance, the touch-sensing (somatosensory) mechanical sensor has a particular location in the skin, and has a physical tuning determined by its anatomical construction and by the mechanical properties of the surrounding tissues. The tuning curves of the neural signals to position and pressure coming from these receptors are determined simply by these physical factors. The tuning curves of higher-order neurons are constructed neurally from these signals. Although the number of cortical neurons is higher than the number of receptors, many peripheral receptors converge on to a single sensory neuron in the central nervous system. Consequently, the receptive field of a central neuron is larger, and its response is more specific. These sensory neurons in turn converge on higher-order neurons. This process is then repeated, which leads to both a progressive increase in the receptive field size and a change in the trigger features of the neurons. The response properties of each neuron are shaped by the specific pattern of connections (and the pattern of connection strengths) from lower-level neurons, as well as by the pattern of inputs from other neurons at the same level of processing.

A major feature of cortical organization is that the cortex is divided into distinct cortical areas defined by anatomical and functional criteria. Cells within a particular area have a similar composition, neurochemistry, and input and output connectivity patterns, and their response properties may be similar. Neurons within an area can form continuous maps of visual space, and specific lesions or damage to cells within an area will lead to similar types of functional deficits.

PRIMARY VISUAL CORTEX

The primary visual cortex (or V1) is the first cortical stage where neural signals arrive from the eye, via

a thalamic stage (the lateral geniculate nucleus or LGN). The specific convergent connections from the LGN to V1, as well as connections within V1, give rise to novel trigger features. V1 neurons are no longer activated effectively by spots of light, and they require elongated features such as lines, edges or gratings of a certain orientation. These receptive fields are thought to be constructed by convergence from a row of LGN cells with receptive fields aligned along a certain orientation. Connections within V1 also have a role in shaping the response properties and the orientation tuning curves. V1 cells have a wide range of receptive field sizes, and the population of V1 neurons covers the whole visual field, more densely with smaller receptive fields near the center of the visual field, and less densely with larger receptive fields around the periphery. V1 neurons are tuned to several stimulus parameters simultaneously. For example, for an oriented line segment one could find optimal parameters such as position (receptive field size is of the order of 1 degree), orientation, length and width of the line segment (or density or spatial frequency of a periodic grating), spatial phase, flicker rate (response can be better to non-static images), direction of motion, and velocity. Some neurons in the input layers are driven preferentially by the left or right eye, while most other cells are driven by both eyes in a similar way. Some other neurons also show selectivity to particular color contrasts (color changes from a background), and some show selectivity to stereoscopic disparity (i.e. a difference in the position of an image of a point between the two eyes due to three-dimensional depth). Any given V1 neuron could be selective to a subset of but not necessarily all of these features. The neurons that are particularly sensitive to the spatial phase of a grating (or the position of a line segment) are called simple cells. Their response properties could be constructed mainly by convergent input from the LGN. Their input can be modeled as an appropriately weighted linear sum of LGN cell activities, with spatially separate excitatory and inhibitory receptive field regions. Here excitatory synaptic inputs would count as positive weights, and inhibitory connections would count as negative weights. V1 cells can of course only give positive responses, so any weighted sums that would be negative result in a zero response (effectively rectifying the response). Phase-insensitive neurons are referred to as complex cells, and they have larger receptive fields. Their responses could be constructed by convergent pooling of the responses of simple cells with similar tuning within the complex-cell receptive field, resulting in a nonlinear relationship between image components and the cell response.

The responses of V1 cells can also be modified by larger stimuli, probably reflecting neural inputs from other neurons within the area.

One of the most remarkable characteristics of the organization of V1 is the systematic and mostly smooth change in response properties as a function of position on the cortex. This may be necessary as neurons representing similar tuning values may need to interact more during processing. Thus placing neurons representing similar parameter values closer to each other on the cortex minimizes the length of the necessary connecting axons. There is in fact a continuous mapping of position of the receptive field in the visual field to cortical position, so that the left-hand side of the visual field is mapped at one side of the cortical map and the right-hand side is mapped at the other ('retinotopic map'). Other tuning parameters, such as orientation, are embedded in this map so that a small region (a 'hypercolumn') containing neurons that have similar receptive field positions is mapped in such a way that tuning parameters such as orientation preference change mostly smoothly across the hypercolumn.

EXTRASTRIATE VISUAL AREAS

Visual areas are organized approximately according to a hierarchical scheme. Pattern recognition is carried out mainly by the 'ventral stream', consisting of areas $\mathrm{V1} \rightarrow \mathrm{V2} \rightarrow \mathrm{V4} \rightarrow \mathrm{PIT}$ (posterior infero-temporal) $\rightarrow$ AIT (anterior infero-temporal). The major input to each of these areas is provided by the area that precedes it in this stream (with some shortcuts and inputs from other areas as well). The main characteristic of subsequent areas is an increase in receptive field size and the development of new trigger features. Increases in the receptive field size do not mean that the optimal stimulus has to get larger for these areas. Instead, the same feature size can be optimal for a higher area as in an earlier area, but the tolerance of the precise position within the receptive field (and of some other properties) is increased at each stage.

For instance, area V2 has cells that have similar tuning to those in V1, but some V2 cells are unoriented and respond to small spots of colored light ('spot cells'), while others respond to contours defined by non-luminance cues such as color borders, texture borders, borders defined by stereoscopic depth or illusory contours (which are like contours behind occluding surfaces). V2 is also likely to be a

significant stage in the processing of local curvature or combinations of orientations.

After V2, the two main visual streams diverge, with form and color being processed primarily by the ventral stream, whereas movement is processed primarily by the dorsal visual stream. The latter continues in areas V3 → MT → MST. MST cells are tuned to complex motions such as expansion/contraction or rotation. Many of these cells are influenced by stimuli beyond the classically defined receptive field. Their responses are suppressed when motion in the surround is in the same direction and has the same speed as in the center. Effectively, this means that MST cells are not driven by global motion, but only by motion relative to a background.

The next stage in the ventral pathway after V2 is V4, where receptive field size increases further. A large class of V4 cells shows tuning to color in a way that displays color constancy. This means that these cells are tuned to the color that a surface reflects (spectral reflectance – the fraction of light that the surface reflects at each wavelength), disregarding to a large extent the incidental changes in the color of the illuminating light. Other V4 cells encode well-localized shape primitives, particular combinations of colored shape primitives and contour features, such as angles and curves, independently of global shape configuration.

The last purely visual, high-level visual areas are in or near the inferior temporal (IT) cortex in monkeys. The cells in this region are sensitive to a variety of stimulus attributes, such as shape, color and texture. Posterior IT cells make complex discriminations of visual form, and have more complex shape tuning, yet they still respond quite well to shapes that are simpler than those that typically occur in the natural world. IT cells are often selective for configurations of several simple shapes with a specific relative spatial relationship. This is especially interesting, as the same cells are insensitive to the absolute position of these configurations within their large receptive fields. Some IT neurons respond to a given shape independently of the cues that define the contours or shape outlines in an image. They respond similarly to forms defined by light-intensity edges, or edges defined by color change, texture change or changes in stereoscopic disparity (cue invariance).

A large proportion of cells in the anterior IT cortex and in the neighboring area of the superior temporal sulcus (STS) are selective for complex, ethologically important patterns such as hands and particularly faces, and views of the whole body. In some cases, cells are tuned even more specifically to faces with a particular orientation, faces with certain characteristics, or the direction of gaze of the face within the stimulus image. IT cells have such large receptive fields that they include most of the visual field. Optical imaging of the surface of the IT cortex suggests that cortical maps in IT are organized in such a way that columns of neurons with related shape-tuning properties are located close to each other in the IT cortex.

These high-level visual cortical cells could be considered to have successfully solved the basic pattern recognition problem, as they respond to specific classes of objects while generalizing across many of the stimulus changes that are unrelated to the identity or 'meaning' of the object. Such representations provide useful inputs to semantic and polysensory areas involved in higher cognitive processes.

The response properties of IT cells can also change with experience. Extensive exposure to images of a set of objects will eventually increase the number of IT cells that respond to those objects. This indicates that the IT cortex plays a critical role not only in the representation and discrimination of objects but also in learning effects, object recognition and visual memory.

SPECIALIZATION AND GENERALIZATION

The example of the monkey ventral visual stream reflects some general computational principles. As one moves from lower to higher sensory areas, one can see that neurons in higher areas are more specialized (i.e. they are tuned to increasingly abstract combinations or configurations of lower-level features). These are not random combinations, but instead they are non-random co-occurrences, or 'suspicious coincidences' in the natural sensory environment of animals. The statistics of the sensory signals from the natural environment provide many different types of regularities or structure, and different stages of sensory processing are well adapted to this structure. For example, a 'face cell' in IT cortex is well suited to an environment in which facial features (e.g. eye shapes and mouth shapes) tend to appear and disappear together. The mechanism for specialization seems to be selective convergence of the signals from detectors of the coincident stimulus features. This convergence allows a higher-order neuron to respond only when a significant number of the appropriate lower-order features are present. The connections necessary for the appropriate convergence can

be learned by experience-dependent synaptic modification.

The other general principle that can be observed is that the tuning to those parameters of the stimuli which are related only to incidental conditions of image formation, or which are not related to real objects or events in the world, is gradually eliminated or generalized over. An example of such a non-object-related feature is the location of a feature within the visual field. The gradually increasing size of the receptive field in higher sensory areas illustrates the corresponding generalization. The most likely mechanism for such generalizations is also by convergence or 'pooling' of the signals about the appropriate set of specific lower-level features. The appropriate pattern of pooling connections can also be acquired through sensory experience, exploiting statistical structure in the normal sensory environment.

These principles do not seem to be unique to the visual system. Despite the quite different nature of the input from the somatosensory (touch) and auditory systems, the center–surround organization, and at later stages orientation and directional motion selectivity, seems to be present in these other modalities as well.

The concept of a receptive field was clearly useful at the earliest stages of sensory processing, where the major neural tuning properties were spatial. The concept has to be stretched to its limits in higher areas to include more pattern-specific information. The appearance of strong nonlinearity in responses, even during the first cortical processing stage, and the significant effect of specific long-range dynamic interactions from beyond the classic receptive field makes that concept of a well-defined and descriptive receptive field less justifiable.

STIMULUS–RESPONSE CORRELATIONS

There is a remarkably close relationship between sensory stimuli, neural responses and perceptual experience. In the previous sections we considered examples of the highly specific nature of this relationship between the stimuli and responses of cortical neurons. Even though these results strongly suggest that these cells take part in the perceptual decisions and experience, we cannot be certain of this without studying perceptual performance itself.

Humans, and in some cases other animals as well, can be asked to report their sensory experiences by simple behavioral decisions (e.g. the pressing of buttons depending on the detection or discrimination of stimuli). Performance in these psychophysical tasks can be highly informative about the nature of the actual mechanisms involved in perception, especially in difficult perceptual judgments, that are near the limits of the capabilities of the perceptual system. When stimulus levels are close to threshold, the observer makes errors with repeated presentations of the sensory stimuli. A sensory threshold can be defined as the stimulus level that supports a certain probability of making a correct decision. Thus, at their thresholds, perceptual judgments are subject to statistical variation.

Neural responses also show statistical variability, as the neural response is not identical even when the same stimulus is presented to the animal many times. This variability can have several sources. It may be due to factors internal to the neuron, or it may be due to fluctuations in the stimulus itself, noise in the receptor cells, or noise at any point in the neural processing pathway between the stimulus and the neuron.

Neural detection or discrimination thresholds can be determined for a population of cells using stimuli that are matched to the tuning characteristics of the neurons. The same stimuli can be used in a psychophysical task to determine perceptual thresholds. If the neurons as a group take part in the perceptual decision-making, their joint performance should be able to support the observed perceptual performance.

There are several models of the way in which information from multiple neurons can be combined to give perceptual levels of performance. The 'lower envelope principle' was proposed as an explanation for the detection of low-contrast gratings. Individual neurons in V1 have tuning curves to spatial frequency (related to the spacing of bars in a grating pattern). The peaks of the neural tuning curves also correspond to the spatial frequencies where the detection threshold curve has a minimum (i.e. where the cells' response can be used to detect the presence of a low-contrast grating most efficiently). This minimum lies at different spatial frequencies for different V1 neurons. However, subjects performing detection tasks have lower thresholds across a much wider range of frequencies than their individual neurons. In fact, the psychophysical threshold curve approximately follows the minima (or 'lower envelope') of the individual threshold curves of a population of neurons. The lower envelope principle therefore states that human perceptual performance in the detection task is always based on a single neuron (or a small subset of neurons), which has a

minimum threshold at the tested frequency. In other words, the hypothesis states that humans can select on each trial the neurons that are most sensitive to the particular stimulus, and can base their decision on these neurons. Although the lower envelope principle seems to be consistent with much of the psychophysical and physiological data, it assumes that the neural decision mechanism has completely accurate prior information about the parameters of the stimulus to be detected, and can also switch the decision mechanism to receive its signal from the right set of neurons without interference from other neurons. These seem to be unrealistic assumptions.

Alternative decision models combine signals from appropriately selected pools of neurons either by addition or by taking optimally weighted sums. The pooled response is normally less noisy and more reliable. Other statistical methods depend on the probability distributions of responses to determine the best decision of a statistically ideal observer of the neural responses. (*See* **Decoding Neural Population Activity**)

The strongest evidence for the involvement of a set of neurons in a perceptual task comes from experiments in which neural and psychophysical data are acquired simultaneously. In one experiment, neural impulses were recorded in humans (using microneurographic methods) from the fibers carrying information from the mechanoreceptive fibers of the hand. Psychophysical responses to pressure on the skin were recorded simultaneously. The psychophysical thresholds for the finger were similar to those based on the response of a single neuron. Furthermore, there was almost perfect trial-to-trial variation in that on trials when the stimulus failed to elicit a neural spike, detection was also missed, but when the spike was detected on a trial, the subject also detected the stimulus. This implies that in this case a single action potential in the fiber is sufficient to cause a conscious sensation. However, another neural fiber carrying mechanoreceptive information from the palm of the hand, was significantly more sensitive than the response of the human subject. This seems to imply that access of conscious decision-making to all incoming activity is not guaranteed. Moreover, electrical stimulation of a particular mechanoreceptive fiber caused the sensation of a light indentation at the location on the skin surface that corresponded to the previously determined receptive field of that particular fiber.

In another interesting set of experiments, responses from area MT of the visual cortex were recorded simultaneously with behavioral responses from a monkey in a direction discrimination task matched to the preferred stimulus of a directionally tuned motion-detector cell. For a population of cells, the distribution of ratios of neural and behavioral thresholds was close to 1, indicating that the level of sensitivity of single neurons was very close to the sensitivity of the monkey's response. The response of the monkey could, at least in theory, be based on the response of a single cell or a very small pool of cells. The trial-to-trial correlation between the neural and behavioral responses was small but significant. For repeated presentation of a given stimulus, the monkey tended to choose the preferred direction of the neuron more often when that neuron yielded a larger than average number of spikes. This implies either that the recorded neuron was one of a set of independently noisy neurons that contributed to the perceptual decision, or that there was a correlation between the variability of this cell and other cells that did contribute to the decision. Further evidence for the involvement of cells in perceptual decision-making comes from experiments in which the recording electrode was used to alter the activity of a relatively small group of cells by injecting a small current during the decision task. Stimulation of a column encoding upward motion induced a bias in the monkey's choices towards the upward direction. Such experiments establish that activity in directionally specific MT circuits of the monkey visual cortex underlies behavioral judgments of motion direction in the psychophysical task.

In conclusion, although there is a clear link between the activation of a relatively small subset of neurons and perceptual experience, it is difficult to infer precisely the number of neurons participating in even such simple perceptual tasks as were discussed above. While the sensitivity and reliability of individual neurons can be similar to the sensitivity and reliability of the whole behaving organism, it is not possible to conclude that we have identified the single neuron from the pool of many millions of neurons that actually performs the discrimination or detection task. Selecting the signal of the appropriate neuron to control the response for a particular task, as the lower envelope principle would assume, without interference from the irrelevant signals of millions of other neurons, must be an enormously difficult task. The imperfection and unknown noise that are introduced by this selection process must be compensated for by the improvement of the signal quality by the combination of responses from multiple neurons. Both the degree of improvement and the amount of signal

combination necessary for a given level of improvement are as yet unknown.

LOCAL, DISTRIBUTED AND SPARSE CODING

Despite having a relatively detailed knowledge of the response properties of most neurons in some parts of the visual system, due to nonlinearities and dynamic interactions we are still unable to predict reliably their responses to arbitrary novel stimuli. Most neurophysiological recording experiments from the visual cortex still use a relatively small set of simple abstract stimuli, so our knowledge of the way in which these cells respond to the type of complex and dynamic natural sequences of scenes that animals experience in their natural environment is to a large extent unknown. Furthermore, we would need simultaneous access to the activation patterns of whole populations of cells in order to describe the neural code. Single-cell recording techniques currently allow us to record only a relatively small number of neurons (often just one) at a time. Various functional imaging techniques may give us a vague idea of the involvement of brain regions in certain tasks, but their spatial and temporal resolution misses most of the computationally relevant details. A further problem concerns the way in which an experimenter chooses the range of stimuli to be tested on a neuron. This is a highly subjective and assumption-laden process, and it is almost certain to bias the results.

Despite these difficulties, there are some aspects of the neural code about which we can make some inferences. For instance, it seems to be a general observation in sensory areas that neurons have fairly specific tuning preferences, often to several parameters of the stimulus simultaneously. A random selection of natural stimuli will therefore relatively rarely cause high activation in a given sensory neuron, which means that neurons are narrowly tuned across the set of all possible stimuli. A separate but related issue concerns the distribution of activity across the whole population of neurons for a given stimulus, which determines the amount of sparseness. A code is sparse when the number of active units divided by the total number of units is low. Due to the limitations of the recording techniques, we have no way of observing sparseness directly. However, on average, the more narrowly tuned the neurons are across the stimuli, the smaller the number of neurons that is expected to be active for a given stimulus. To encode the same information with fewer active neurons, the visual system must eliminate the statistical redundancy (i.e. the uninformative replication) in its coding. For instance, the center–surround organization of retinal ganglion cells eliminates responses except at the boundaries of visual surfaces. This type of organization has also been found in higher visual areas. For instance, the responses of motion-sensitive cells in area MT are suppressed by the same motion in the surround, so they only respond when the motion involves solely the receptive field of the cell. These types of specific suppressive effects have been found in all cortical areas where they were tested, presumably resulting in a significant decrease in the number of active cells.

A sparse code is useful because it combines the advantages of local and fully distributed codes. It allows the discrimination of a much higher number of states than local codes, and it allows generalization based on the overlap of representations, while still keeping the mapping from the representation to an output pattern relatively simple and easy to learn by means of simple learning algorithms. A neural network model that aimed specifically to maximize a measure of sparseness trained on natural images resulted in localized wavelet-like receptive-field profiles that resemble V1 simple-cell profiles. Finally, a sparse code may be a desirable way of encoding information, as the number of real causes of any given sensory stimulus is probably much lower than the total number of potential causes of all possible natural images. Each image could then be encoded as a relatively small subset of active units corresponding to the current causes.

Further Reading

Abbott LF and Sejnowski TJ (eds) (1999) *Neural Codes and Distributed Representations*. Cambridge, MA: MIT Press.

Arbib MA (2002) *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press.

Baddeley R, Hancock P and Foldiak P (2000) *Information Theory and the Brain*. Cambridge, UK: Cambridge University Press.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

deCharms RC and Zador A (2000) Neural representation and the cortical code. *Annual Review of Neuroscience* **23**: 613–647.

De Valois KK (ed.) (2000) *Seeing*. Academic Press.

Hinton G and Sejnowski TJ (eds) (1999) *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: MIT Press.

Kandel ER, Schwartz JH and Jessell TM (2000) *The Principles of Neural Science*, 4th edn. McGraw-Hill.

Parker AJ and Newsome WT (1998) Sense and the single neuron: probing the physiology of perception. *Annual Review of Neuroscience* **21**: 227–277.

Penfield W and Perot P (1963) The brain's record of auditory and visual experience. *Brain* **86**: 596–696.

Rockland KS, Kaas JH and Peters A (eds) (1997) *Cerebral Cortex*, vol. 12. *Extrastriate Cortex in Primates*. New York: Plenum Press.

Simoncelli EP and Olshausen BA (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* **24**: 1193–1216.

Tanaka K (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* **19**: 109–139.

Walsh V and Kulikowski JJ (eds) (1998) *Perceptual Constancy: Why Things Look As They Do*. Cambridge, UK: Cambridge University Press.

Neuropsychological Disorders, Animal Models of

Introductory article

Gordon Winocur, Rotman Research Institute, Toronto, Canada

CONTENTS

Animal models are important to the study of brain damage and neurological disease. When brain function is affected, there are direct psychological and cognitive symptoms that must be taken into account to achieve a full understanding of the disease.

PROS AND CONS OF ANIMAL MODELS

Animal models are an essential part of medical science and have contributed significantly to our understanding of various diseases. Indeed, the treatment of serious illnesses, such as cancer and diabetes, would still be very primitive without the knowledge derived from animal research. This approach is possible because humans and other animals (especially mammals) share important anatomical and physiological characteristics, so that meaningful generalizations across species are possible. Apart from practical considerations, the use of animals is necessitated on moral and ethical grounds; it is obviously unacceptable to subject humans to unproven experimental procedures and treatments, no matter how promising they might be.

In the study of brain damage and neurological disease, we must consider psychological and cognitive symptoms as well as physical ones. Thus, for example, in studying Alzheimer disease (AD), in addition to describing the neuropathological changes that characterize the disease, it is necessary to identify parallel behavioral changes (e.g., memory loss), because of their obvious importance for diagnosis and treatment. With the emergence of the discipline of neuropsychology – the study of relationships between brain function and behavior – sensitive tests have become available to study changes in cognitive abilities suffered by humans with brain damage. Of equal importance is the progress that has been made in devising tests that measure similar processes in animals, and allow for the development of models that represent these changes.

When properly used, animal models of neuropsychological disorders can be instructive on several levels. If the model actually duplicates the disorder it is meant to represent, it can provide valuable information as to cause, symptoms, and treatment. Unfortunately, because our knowledge of complex neurological diseases is limited, it is virtually impossible to model all aspects of a disease. Consequently, we must often resort to creating artificial, approximate models. This can be accomplished by physiological or biochemical manipulations that can involve, for example, surgical destruction of brain tissue or genetic alterations. In such cases, it is usually necessary to focus on some subset of symptoms or causative factors; but even a partial characterization, if accurate, can lead to significant progress. An important measure of the success of a model is the extent to which it expresses cardinal behavioral features of the disease.

Animal models can provide valuable insights into the mechanisms of diseases. For example, we are becoming increasingly aware that heredity and environmental factors interact in complex ways in the expression of certain diseases. Only by carefully controlling genetic make-up, environmental influences, and related experiential factors (e.g., early trauma, stress) can their effects be assessed. Similarly, as neurological diseases progress, new and distinct neuropathological markers appear, along with new cognitive symptoms. Typically, the type of control needed to study these relationships can be exercised much more readily in animal populations than in humans. Although the advent of sophisticated neuroimaging techniques has created new opportunities in human research, properly-constructed animal models represent the best hope for this line of investigation.

Another advantage of animal models is that they are well suited to preclinical experimentation with drugs and other treatments. Apart from the study of short- and long-term benefits of a treatment, animals allow for the study of side effects that could compromise its effectiveness. Unfortunately, it is often easier to reduce cognitive impairment in brain-damaged animals than in comparable human populations, and therein lies a limitation of animal models. Nevertheless, demonstrating that a particular drug can effectively relieve a major symptom in any form of the disease is usually an important step in the right direction.

There are other problems with trying to model neuropsychological disorders in animals. For example, some have argued on compassionate and even scientific grounds that alternative methods (e.g., computer modeling) are preferable. It goes without saying that animal researchers must follow procedures that are as safe and humane as possible and that safeguards must be in place to protect animals from maltreatment. While a certain amount of controlled pain is sometimes inevitable, it is kept to a minimum, and practice is guided by the principle of the 'greater good'. As long as we are prepared to attach special significance to the quality of human life, there is little doubt that the benefits of animal research outweigh the costs.

It may be argued that, because neuropsychological problems are not necessarily life-threatening, animal sacrifice is not warranted. In reality, just as animal research has contributed significantly to the development of drugs for treating serious mental illnesses (e.g., schizophrenia) and alleviated much suffering in the lives of patients, animal-based findings can lead to the discovery of cognitively enhancing medication. One must not underestimate the misery of those suffering a dementing disease, and the accompanying psychosocial problems that take an enormous toll on the quality of an individual's life. The argument that cognitive disorders do not represent a serious enough health problem to justify animal experimentation must be considered against the plight of the many elderly individuals, stroke victims, and AD patients, who are mentally imprisoned within a shrinking cognitive world.

The neuropsychological investigation of cognitive disorders in brain-impaired humans is still in its infancy, but, thanks to a growing recognition of the problem and the development of new technologies (e.g., sensitive tests, structural and functional brain imaging), great strides are being made. This article will review advances in the use of animal models. It first reviews developments in assessing cognitive deficits associated with brain damage. The emphasis here is on the circumscribed damage that results, for example, from brain surgery, cardiovascular disease (e.g., stroke), or head trauma. We then look at the unique features and cognitive symptoms of neurodegenerative diseases, with AD, Parkinson disease (PD), and Huntington chorea (HC) as specific examples.

ASSESSING COGNITIVE IMPAIRMENT IN ANIMAL MODELS OF BRAIN DAMAGE

We would ideally want to use the same tests to measure cognitive function in animals and humans. While this is possible to some degree, several factors – most prominently, cross-species differences in cognitive capacity – render this approach impractical. In addition, animals and humans have different ways of doing things and are biased towards different types of information. As an obvious example, humans are predisposed to make extensive use of language, monkeys respond primarily to visual information, and rodents rely heavily on olfactory and spatial cues. A more realistic strategy is to administer tasks that are appropriate to each species' stimulus–response characteristics, but that require the same cognitive processes for successful performance. Thus, for example, we assume that short-term memory (STM), the ability to retain small amounts of information for short durations, is distinct from long-term memory (LTM), where information is stored for long periods of time, but that both are part of the cognitive make-up of all relatively sophisticated animals.

In most cases, even circumscribed brain damage produces variable patterns of cognitive deficits. However, memory loss is a common feature of most neuropsychological disturbances. For this reason, and to exemplify the modeling approach in animal research, this review will focus on memory as a primary domain of cognitive impairment associated with brain damage. The challenge, as with all forms of cognitive assessment, is to devise tests that measure memory impairment in ways that are appropriate to each species.

The development of suitable tests for studying the effects of brain damage on memory function was stimulated by attempts to reproduce features of human amnesia in animals. In a pioneering investigation of neurological patients in the 1950s, Brenda Milner and her colleagues at the Montreal Neurological Institute discovered that damage to the hippocampus, a major structure deep within the medial temporal lobe (MTL) of the brain,

produced a profound memory loss for experiences subsequent to the injury (anterograde amnesia). In general, the impairment was restricted to conscious recollection of specific events (e.g., a news event that occurred several days ago), whereas memory for general knowledge (e.g., that the sun rises in the east and sets in the west) or learned skills (e.g., chess) was relatively intact. There was also evidence of time-dependent retrograde amnesia, in which experiences that occurred a relatively short time before the hippocampal damage were forgotten, while much older memories were retained.

Milner's studies were conducted on epileptic patients who underwent surgical removal of hippocampal tissue to reduce the frequency and severity of debilitating seizures. However, similar patterns of memory function are observed in patients who suffer hippocampal damage as a result of neurological diseases, such as herpes encephalitis, an inflammation of the brain that selectively attacks the medial temporal lobe. Also, vascular infarcts that block the supply of blood, glucose, and oxygen to the hippocampus will produce ischemic damage and similar memory deficits.

Research with various clinical populations has revealed that other structures also contribute to memory function. In addition to implicating other brain regions, such as the frontal lobes (FL), the basal ganglia (BG), and various thalamic nuclei, this work has demonstrated important differences between the contributions of the different regions. As we will see later, the BG are involved in a form of (procedural) memory that supports learning and remembering non-episodic information. By comparison, the FL are essential when recalling information that is dependent on strategic search and retrieval operations.

The importance of the thalamus – a subcortical structure within the diencephalon – for memory emerged from studies of Korsakoff syndrome, a condition that is associated with chronic alcoholism and vitamin B deficiency. Korsakoff patients often suffer extensive neuropathology, but the profound memory loss that they experience seems to relate to damage to specific (anterior and dorsomedial) nuclei in the thalamus. Studies of patients with restricted damage to the thalamus, resulting from stroke or tumor, confirmed the link, although their memory profile is different from that associated with MTL damage. Thalamic patients suffer a more generalized impairment, which can include STM and LTM function, and even extend to certain types of learning. On the other hand, when damage is restricted to the thalamic region, premorbid memory appears to be intact.

Thus, rather than comprising parts of a single, integrated memory system, as was once thought, the hippocampus and thalamus probably control different memory-related cognitive processes. Animal studies have contributed significantly to the resolution of this issue. To model selective brain damage in animals, lesions are produced in various ways, such as passing electrical current through the structure, aspirating tissue, or infusing neurotoxins that destroy neural cells. Ischemic damage, similar to that seen in stroke patients, is typically produced by temporary occlusion of major arteries that supply blood to targeted brain areas. Extensive behavioral testing, involving a wide range of tasks, has been conducted in different species of animals (e.g., monkeys, rats) with selective hippocampal or thalamic lesions produced by these procedures.

Animal research has confirmed that lesions to either structure produce comparable degrees of memory loss for specific information, such as the location of food rewards. However, when more sensitive tests are administered, it becomes apparent that the hippocampus and thalamus mediate different cognitive processes. On tests that provide separate measures of general learning ability, STM, and LTM, the memory problems of thalamus-damaged animals were related to difficulties in learning the task. This pattern, which is also observed in thalamic amnesic patients, suggests that the thalamus is involved in the early stages of learning and possibly in attending to relevant task-related information. By comparison, when tested on similar tasks, animals with hippocampal damage generally exhibit good rule-learning ability but are reliably impaired when required to recall specific responses after extended delays, or when distracted by interfering influences. There are several views of hippocampal function, but this pattern suggests that the hippocampus participates in transferring information from STM to LTM, a process that is sometimes called 'consolidation'.

Memory function is also affected when the FL are damaged, but, again, in different ways. Patients with FL lesions resulting from injury or disease do not suffer generalized memory loss, but they have difficulty recalling information without the benefit of external aids, or when a conscious effort is required to locate the information in their memory. Thus, they may recognize a familiar picture, but the picture is not likely to conjure up memories of experiences associated with the scene. It is difficult to characterize FL function in precise terms, but there is broad agreement that the structure is involved in strategic processes that direct thought

and goal-oriented behavior. This function is important not only for free recall, but also for other cognitive activities that require the organization of information and behavioral planning. Therefore, FL patients are also impaired in complex learning tasks, in solving problems, and in organizing daily routines.

Our understanding of the functional significance of the FL, as well as their interactions with other brain regions, has been advanced significantly by animal research involving a variety of tasks. One such task involves conditional associative learning (CAL), in which different stimuli must be associated with different responses. Thus, if stimulus A is presented, the individual must perform response A; if stimulus B is presented, response B, and so on. Both humans and monkeys with FL damage are severely impaired on this rule-learning task. Rats with FL lesions are also impaired in a Skinner-box version of CAL, but, depending on the task requirements, other structures are also important. Thus, while rats with FL lesions are impaired in learning the basic rule, rats with hippocampal lesions learned it as well as normal controls. Increasing the delay between stimulus presentation and the opportunity to respond does not affect FL-damaged rats, but the hippocampal group's performance deteriorates sharply. This dissociation is significant because it shows that the primary effect of the FL lesion was on the strategic aspect of the task – that is, in using relevant information to learn the rule and select correct responses. In contrast, the behaviour of the hippocampal rats confirmed that the hippocampus is involved in relatively straightforward memory processes.

The above examples illustrate the feasibility of devising animal tests that measure memory processes analogous to those affected by focal lesions to the MTL, diencephalic, and FL systems. Investigators have used these tests to confirm the precise function of specific brain regions, and to show how localized functions are integrated within a multiple-systems model. This type of research has created a strategy for studying the complex pattern of lost and spared memory function following brain damage. It is a strategy that has already led to important insights into the nature of memory disturbances, and it has practical implications for the development of treatment programs.

ANIMAL MODELS OF ALZHEIMER DISEASE

Alzheimer disease, the most common cause of dementia, affects about 10% of people over the age of 65, and as much as 40% of the population over 80. The early manifestations are seemingly innocent bouts of forgetfulness, but as the disease progresses, memory lapses become more frequent and other forms of cognitive impairment begin to appear. Over a period of years, attention, learning ability, language and speech skills decline as part of a deterioration of general intellectual function.

In line with the early symptoms of memory loss, AD begins with abnormalities in the hippocampus, which spread to other regions of the brain. Apart from considerable neuronal loss, certain pathological features are unique to AD, although their precise relation to the disease is uncertain. Cardinal features include intracellular accumulation of neurofibrillary tangles that result from changes in the protein *tau*. Normally, *tau* molecules provide internal support to nerve cells, but in AD, they become twisted and clog the cells, causing them to die. There are large extracellular deposits of neuritic plaques formed by an overproduction of β-amyloid protein, which is normally involved in cell membrane function. In large numbers, the plaques trigger an inflammatory response, which may initiate the sequence of events that ultimately leads to cell death.

AD is accompanied by depletion of several neurotransmitters, although most attention has focused on acetylcholine (Ach) and cholinergic neurons of the basal forebrain and septal and hippocampal regions. Cholinergic activity normally declines in old age, but the loss is substantially greater in AD patients. It has been estimated that the brains of AD patients lose as much as 70% of the Ach-synthesizing enzyme, choline acetyltransferase, and up to 90% of the normal supply of Ach. There is little doubt that these changes contribute significantly to the severe memory loss associated with AD.

The nature of AD presents several obstacles to developing a comprehensive animal model of the neuropsychological component. AD does not occur naturally in animals, and certain essential features that affect brain function (e.g., neurofibrillary tangles) have not been reproduced reliably in non-human species. Also, because AD is progressive, it is difficult to capture all the cognitive and related psychological impairments that are expressed as the disease passes through its various stages. In light of these considerations, most attempts to model AD have focused on the early stages, where the pathology is less complex and where memory loss is the primary cognitive symptom.

Investigators have used surgical techniques and pharmacological manipulations to demonstrate that disruption of cholinergic systems in animals produces cognitive deficits similar to those seen in AD patients. Lesions to the medial septum, hippocampus, and nucleus basalis of Meynert (NBM) in the forebrain result in significant reductions in cholinergic activity and cognitive deficits in nonhuman primates and rodents. These experiments often involve conditional learning tasks, such as CAL (described above). On such tasks, animals with lesions to the area of the medial septum and hippocampus typically learn the basic responses within normal limits, but their performance deteriorates when challenged to remember specific information that is critical to response selection. In CAL, this occurs when a delay is introduced between the stimulus and the response. In contrast, lesions to the NBM, the primary source of cholinergic input to frontal cortex, interfere with basic rule learning, but increasing stimulus–response intervals does not add significantly to the deficit. These are important findings because they show that the two branches of the cholinergic system are involved in different functions that are also implicated in AD. The evidence directly links the septal–hippocampal region to the type of memory function that is affected in AD, whereas the NBM–frontal system appears to control strategic or attentional processes that are essential for various kinds of new learning.

In the 1970s, researchers found that drugs, such as scopolamine and atropine, that block cholinergic receptor function at the synaptic level severely disrupt memory function in young adult humans. An equally important finding, and one with clinical implications, was that performance deficits can be reversed by cholinergic agonists (e.g., physostigmine) that promote Ach synthesis. The same effects have been demonstrated in young adult rats and monkeys.

These results provided direction for treating the cognitive problems associated with normal aging and diseases such as AD. In fact, studies showed that cholinergic drugs (e.g., physostigmine, donepezil) were often very effective in overcoming learning and memory deficits in aged animals or in animals with damage to cholinergic brain regions. Currently, such drugs are the only pharmacological treatments approved for AD, and while their effectiveness is limited, they provide relief and hope to many patients, where only a few years ago there was none.

Animal models have also been useful in studying the connection between β-amyloid and cognitive impairment in AD. Given the large deposits of β-amyloid in AD brains, it is not surprising that its presence correlates with neuronal damage and cognitive impairment. Similar relationships have been reported in rats infused with β-amyloid protein in various brain regions, especially the hippocampus. Importantly, the pattern of cognitive impairment associated with β-amyloid build-up is similar to that seen in early AD. In one study, the performance of aged dogs on a battery of learning and memory tests was related directly to β-amyloid levels in the brain. A significant correlation was detected on tests where successful performance depended on remembering specific information. On the other hand, tasks that emphasized stimulus–response rule learning were less affected by age and β-amyloid levels.

Similar results have been reported in investigations of strains of 'senescence-accelerated mice' that are specifically bred to produce β-amyloid deposits that correlate with learning and memory deficits. It follows that it should be possible to limit such deficits by reducing β-amyloid levels. Although not yet attempted in humans, there are indications that this is a promising line of investigation. Several studies have shown that injecting antibodies to β-amyloid into senescence-accelerated mice significantly reduces their memory deficits.

New and exciting possibilities for studying AD emerged from the discovery that some forms of the disease can be traced to genetic mutations and that these mutations can be introduced into the brains of mice. For example, an abnormality on chromosome 19 that results in underproduction of APOE4, a protein involved in cholesterol production, has been linked to over 60% of AD cases. When similar APOE4 deficiency was created in transgenic mice, scientists discovered AD-like pathology that included abnormalities in β-amyloid production and impaired cholinergic function. These mice also exhibited learning and memory impairment that was directly linked to reduced Ach levels in the hippocampus and FL.

APOE4 is a susceptibility gene for late-onset AD. This means that not everyone who carries it will necessarily acquire the disease. Other recently-discovered genes have proven to be stronger predictors of early-onset familial AD, although these genes account for only about 5% of the total number of AD cases. An especially important discovery was a mutation of the amyloid precursor protein (APP) on chromosome 21. APP is normally involved in producing the β-amyloid protein, but in its mutant form it leads to the development of amyloid plaques. Transgenic mice carrying this

mutation develop AD-like neuropathology that includes deposits of plaque and abnormalities in the *tau* protein (which, as indicated above, is implicated in the formation of neurofibrillary tangles). APP mice also show cognitive deficits that correlate with biochemical and pathological abnormalities symptomatic of AD. The model was enhanced when the investigators showed that the constellation of symptoms increased with the age of the animals, as would be expected in a progressive disease like AD. Mutations of other genes (e.g., presenilin 1 on chromosome 14, presenilin 2 on chromosome 1) have been similarly linked to forms of AD. Behavioral work in this area is just beginning, but the encouraging early results raise hopes that this approach will be useful in relating the developmental pattern of cognitive deficits to the progression of neuropathological symptoms of AD.

ANIMAL MODELS OF BASAL GANGLIA DISEASE

The basal ganglia, located centrally in the forebrain, are a group of structures that include the caudate nucleus (CN) and putamen (known together as the neostriatum) and the globus pallidus. These structures form part of the extrapyramidal motor system that regulates movement in all parts of the body. The BG and, most notably, CN are also involved in cognitive functions. The CN has direct connections with the FL, so that damage to the CN often produces cognitive symptoms similar to those seen in patients with FL lesions. We focus here on two degenerative brain diseases – PD and HC – that attack the BG and produce motor and neuropsychological disturbances.

PD, which affects about 0.1% of the population, originates in the substantia nigra of the upper brain stem, and spreads along dopaminergic pathways to the BG and ultimately to the FL. The earliest symptom is loss of fine motor control, but as the disease progress, patients suffer generalized motor slowing, tremor, rigidity, and speech and swallowing difficulties. A variety of cognitive symptoms are also part of the disease. PD patients sometimes develop a form of dementia that includes memory and intellectual loss similar to that of AD. Nondemented PD patients typically exhibit a neuropsychological profile that reflects damage to CN and its frontal connections. The defining cognitive deficit in PD is impairment in procedural learning, as expressed in various problem-solving and skill-learning tasks that require strategic planning. While learning ability is affected in PD, straightforward memory function is relatively spared. When PD patients do experience memory difficulties, it is usually for inaccessible memories and can be traced to interruption of CN–frontal pathways.

HC is a relatively rare genetic disorder that results from an abnormality on chromosome 4. It is characterized initially by involuntary, jerky (choreiform) movement of fingers and toes, which later spreads to most regions of the body. Cognitive and psychiatric problems are common in HC, including dementia and changes in mood and personality. Initially, the disease attacks the CN and putamen; it then spreads to other brain regions including the FL.

In the early stages of HC the cognitive profile resembles that of nondemented PD patients. Thus, HC patients are impaired in learning and remembering new skills and procedures but not necessarily in consciously recalling specific experiences. This point is illustrated in experiments in which HC patients are compared with MTL or diencephalic amnesics on tests of skill learning and specific memory function. Such experiments reveal a double dissociation, in which HC patients display procedural learning deficits but normal memory, while the opposite pattern is observed in amnesic patients.

It is important to note that procedural-learning tasks on which PD and HC patients are typically impaired often have a substantial motor component. Consequently, one has to be careful not to confound effects of motor disabilities in assessing neuropsychological test performance. Careful study has shown that the cognitive and motor symptoms of these diseases, as well as those with focal lesions to CN, are largely unrelated, and linked to different brain systems.

Scientists have attempted to model the cognitive symptoms of PD and HC by studying the effects of damage to CN in animals. Early research focused on the disruptive effects of CN lesions on spatial-learning tasks in which animals had to perform a series of turning responses to locate an object in their environment. Deficits on such tasks mirror the disorientation problems of PD patients, and have been related to the requirement that critical associations be formed between specific stimuli and discrete responses. Other studies have shown that CN lesions disrupt performance on a variety of discrimination-learning tasks where positive and negative stimuli are associated with different responses. These findings parallel those of studies with humans, and provide strong evidence that the CN is a critical part of the neural system which forms stimulus–response associations: a

process that is essential for many rule- or skill-based procedural-learning tasks.

Attempts to develop more specific animal models of PD and HC have utilized neurotoxin lesion techniques that selectively destroy neurotransmitters directly linked to the diseases. Thus, to mirror PD, investigators have infused dopamine antagonists, such as 6-hydroxydopamine, into the CN, causing selective damage to dopamine cells. With respect to HC, similar experiments have been performed with various pharmacological agents (e.g., kainic acid, quinolinic acid, 3-nitrapropinic acid) that produce changes in CN similar to the pathology of HC. Such research has confirmed that even selective damage to CN produces generalized stimulus–response learning deficits. It is not yet clear whether traditional lesion techniques will be sufficient to capture the range of symptoms of PD or HC, and other strategies need to be explored.

The development of models using transgenic mice has been a major advance in studying HC. These models have the important advantage of taking into account the genetic basis of the disease. Mice carrying the human HC gene display motor disturbances and many of the neuropathological features of the disease. Behavioral studies with these mice are just beginning, but early results indicate a deficit pattern consistent with that observed in other animal models and in HC patients. Of particular importance is the observation that these mice develop progressive impairments that reflect the progression of the disease to other regions of the brain, and specifically to the FL.

A promising model of PD has emerged from tragic and unusual circumstances. In the 1980s, William Langston, a neurologist working in California, reported that over 200 drug addicts developed parkinsonian symptoms after mistakenly injecting themselves with a substance, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyradine (MPTP). MPTP is a powerful neurotoxin that selectively attacks the substantia nigra, destroying dopaminergic pathways that connect this region to BG. The entire group developed typical signs of PD, including the motor and cognitive disturbances. MPTP does not produce PD-like symptoms in rodents, but monkeys show virtually the full range of pathological signs and motor deficits within three days of injection.

Investigations of MPTP effects have consistently revealed learning and memory deficits on tasks that are sensitive to impairment in the FL and CN brain regions: a pattern that is consistent with human findings. An important feature of the MPTP model is that the cognitive deficits correlate with dopamine depletion and respond to treatment with dopamine agonists. Thus the tragedy of the MPTP-exposed drug addicts has led to a primate model that may lead to major breakthroughs in understanding and treating a menacing brain disease.

With respect to treating BG diseases, there has been some progress in recent years. Again, animal models have played a central role. In PD, the most common treatment is to prescribe dopamine agonists like levadopa and deprenyl, which temporarily reduce the symptoms in most patients. More invasive treatments, such as surgical destruction of anatomically-related subcortical structures (e.g., thalamic nuclei), can be effective in reducing motor symptoms. HC is sometimes treated with dopamine antagonists to restore BG function, but the benefits tend to be modest. Because of its genetic origin, advances in genetic engineering and the development of transgenic-mice models of HC offer greater prospects for success.

Another radical treatment, which takes a very different approach to restoring lost brain function, involves grafting or transplanting live neural tissue. In the most common application, brain cells are extracted from aborted fetal tissue and transferred to damaged sites in the host brain. Properly implanted, the cells migrate to appropriate locations and begin to function both physiologically and neurochemically. Behavioral experiments, in which cholinergic or dopaminergic brain cells were transplanted into brain-damaged or aged animals, have shown that learning and memory function can also be restored. This technique has proved to be especially effective in the MPTP monkey model of PD, where CNS transplantation of dopaminergic cells reduced the motor and cognitive symptoms of the disease. Indeed, the results have been so encouraging that the procedure has been attempted on PD patients, but with mixed results.

CONCLUSION

Animal models are becoming as important to the study of neuropsychological deficits following brain injury as they are to any other medical condition. Substantial progress followed the emergence of surgical, pharmacological, and genetic techniques that produce, in animals, the neuropathological and cognitive changes reliably seen in patient populations. Equally important was the need for sensitive behavioral tests that assessed cognitive impairment associated with localized damage, and dissociated specific deficits from those

resulting from coexisting damage to other regions. Important advances in these areas led quickly to improved understanding of underlying mechanisms, and to experimental treatments that promise to alleviate functional deficits associated with brain injury of various etiologies. This article has focused on the use of animal models of cognitive deficits associated with focal brain lesions and such neurodegenerative diseases as AD, PD, and HC. These examples serve to make the point with respect to major neurological problems, but the list could be extended to include Down syndrome, multiple sclerosis, and neuropsychiatric diseases (e.g., schizophrenia, obsessive–compulsive disorder) that have cognitive symptoms.

Animal models have their limitations, and the practice, with respect to neuropsychological disorders, has its critics. Nevertheless, the benefits derived are incontrovertible, and as they continue to accrue, the central role of animal models in this research enterprise will be assured.

Acknowledgment

The preparation of this article and some of the research reported were supported by grants from the Canadian Institutes for Health Research and the Natural Sciences and Engineering Research Council of Canada.

Further Reading

Bakay RAE and Herring CJ (1993) CNS transplantation in nonhuman primates to relieve parkinsonism. In: Schneider JS and Gupta M (eds) *Current Concepts in Parkinson's Disease Research*, pp. 299–334. New York, NY: Hogrefe and Hubert.

Bartus RT (2000) On neurodegenerative diseases, models, and treatment strategies: lessons learned and lessons forgotten a generation following the cholinergic hypothesis. *Experimental Neurology* **163**: 495–529.

Cummings BJ, Head E, Ruehl W, Milgram NW and Cotman CW (1996) The canine as an animal model of human aging and dementia. *Neurobiology of Aging* **17**: 259–268.

Di Monte D and Langston JW (1993) MPTP-induced Parkinsonisn in nonhuman primates. In: Schneider JS and Gupta M (eds) *Current Concepts in Parkinson's Disease Research*, pp. 159–179. New York, NY: Hogrefe and Hubert.

Guénette SY and Tanzi RE (1999) Progress toward valid transgenic mouse models for Alzheimer's disease. *Neurobiology of Aging* **20**: 201–211.

Hsiao K, Chapman P, Nilsen S *et al*. (1996) Correlative memory deficits, Aβ elevation, and amyloid plaques in transgenic mice. *Science* **274**: 99–102.

Ingram DK, Spangler EL, Iijima S *et al*. (1994) Rodent models of memory dysfunction in Alzheimer's disease and normal aging: moving beyond the cholinergic hypothesis. *Life Science* **55**: 2037–2049.

McDonald MP and Overmier JB (1998) Present imperfect: a critical review of animal models of the mnemonic impairments in Alzheimer's disease. *Neuroscience and Biobehavioral Reviews* **22**: 99–120.

McDonald RM, Ergis A-M and Winocur G (1999) Functional dissociation of brain regions in learning and memory: evidence for multiple systems. In: Foster JK and Jelicic M (eds) *Memory: Systems, Process, or Function?*, pp. 66–103. Oxford, UK: Oxford University Press.

Ridley RM and Baker HF (1991) A critical evaluation of monkey models of amnesia and dementia. *Brain Research Reviews* **16**: 15–37.

Squire LR (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**: 195–231.

Thomas RK (1996) Investigating cognitive abilities in animals: unrealized potential. *Cognitive Brain Research* **3**: 157–166.

Winocur G (1991) Functional dissociation of the hippocampus and prefrontal cortex in learning and memory. *Psychobiology* **19**: 11–20.

Winocur G and Moscovitch M (1999) Anterograde and retrograde amnesia after lesions to frontal cortex in rats. *Journal of Neuroscience* **19**: 9611–9617.

Neurotransmitters

Introductory article

Gary L Wenk, University of Arizona, Tucson, Arizona, USA

CONTENTS

Neurotransmitters are chemicals that are released in tiny amounts by neurons in response to an action potential.

INTRODUCTION

Neurotransmitters are chemicals that are released in very small amounts from neurons in response to the arrival of an action potential. Once released, the neurotransmitter travels across the synaptic cleft that separates one neuron from the next in order to transfer information. Almost everything interesting that relates to information processing is thought to rely upon the action of the neurotransmitters.

The first neurotransmitters to be discussed are those that form diffuse regulatory pathways. These pathways originate in the brainstem region; they include those releasing the neurotransmitters acetylcholine, dopamine, noradrenaline (norepinephrine), serotonin, and the various neuropeptides. These projection pathways consist of small clusters of cell bodies that mostly reside within the brainstem and send branching axons to terminate in multiple and widespread areas of the brain. Thus, a single neuron may modulate the activity of a large number of target cells. The diffuse pathways contrast with other neurotransmitter pathways which provide a rapid, point-to-point innervation to specific neurons.

PRINCIPLES OF CHEMICAL NEUROTRANSMISSION

It is important to realize the fact that behavior influences and is influenced by the neurotransmitter function within the brain. Specificity of neurotransmitter action is fundamental to normal brain function. This specificity is achieved by either localized release within a discrete area (anatomic specificity) or by limited receptor distribution (chemical specificity). Anatomic specificity is imparted through spatial sequestration of transmitter–receptor interaction; i.e. the actions of the neurotransmitter are restricted to the space of the synaptic cleft. Chemical specificity is dependent upon the moment-to-moment composition of interstitial fluid that surrounds the neurons at a given receptor population, and hence neurotransmitters may act at sites distant from their release. The two main things that you need to know about what a neurotransmitter is doing are (a) where it is released, and (b) what the receptor does in response to the presence of the neurotransmitter. Simply stated, neurotransmitter systems that innervate specific brain regions influence the function of that brain region. For example, because noradrenaline (norepinephrine) innervates the hippocampus it has a role in hippocampal function, and because the hippocampus has a role in memory, so then does noradrenaline.

CLASSES OF NEUROTRANSMITTERS

With the introduction of sophisticated histological techniques in the early part of the twentieth century, neuroanatomists were able to map neuronal pathways on the basis of transmitter content. Only the 'classical' neurotransmitters – acetylcholine and the monoamines – are discussed in detail here; owing to the multiplicity of neuropeptide neurotransmitters (over fifty have been identified to date), these substances are considered only in general terms. Finally, the two major amino acid neurotransmitters, glutamate and γ-aminobutyric acid (GABA), are introduced.

An important concept to keep in mind when considering the role of the various neurotransmitters in the brain is that they occur at very different levels of concentration. The neuropeptides exist at nanomolar levels, the classical neurotransmitters exist at 100 to 1000 times the concentration of

neuropeptides, and the amino acid neurotransmitters exist at concentrations 100 to 1000 times greater than that of the classical neurotransmitters.

In addition, the manner in which the brain produces and discards the various neurotransmitter molecules in each class is quite different. The neuropeptides are processed from much larger polypeptides and are not reused after release from the synaptic terminal. In contrast, the classical neurotransmitters, such as acetylcholine, dopamine, and noradrenaline, begin as simple amino acid nutrients from the diet and are modified by a series of enzymes prior to being released. Once released, these transmitters are largely used again by a process called reuptake; essentially, the neurotransmitter is removed from the synaptic cleft and repackaged into a new synaptic vesicle. Finally, the amino acid neurotransmitters glutamate and GABA are obtained from the diet and are not modified at all prior to being stored in synaptic vesicles and then released into the synaptic cleft. These differences also highlight another difference in the nature of these neurotransmitter systems: the neuropeptides are metabolically expensive to produce and use compared with the amino acid transmitters.

RECEPTOR SUBTYPES

Most of the neurotransmitter receptors are proteins that float within the neural membrane. They are the targets of action of the neurotransmitter molecules as well as of drugs that influence brain function. They respond with specificity to the neurotransmitters, have a finite rate of operation (the speed with which they transmit information) and are saturable (they have a finite maximal effect). The binding of neurotransmitter molecules to the receptors is reversible.

There is usually more than one type of receptor within each neurotransmitter system. For example, there are at least five different dopamine receptors and two different receptors that respond to acetylcholine. The dopamine receptors all respond to dopamine, but the nature of the response varies considerably: for example, some dopamine receptors open ion channels directly, while others activate a series of second messenger systems. The different subtypes of neurotransmitter receptors are distinguished also by the drugs that interact with them: for example, the two subtypes of acetylcholine receptor are distinguished by the fact that one responds to nicotine while the other responds to muscarine.

Receptors are further distinguished by whether they lead to an increase or decrease in neural function. For example, serotonin might bind to one of its receptor subtypes and produce stimulation, while binding within the same brain region to a different subtype of receptor might lead to inhibition of the activity of the postsynaptic neuron. Therefore, it is impossible to state that a particular neurotransmitter is either excitatory or inhibitory; this response depends entirely upon the nature of the receptor.

NEUROTRANSMITTERS IMPORTANT IN COGNITION

Acetylcholine: Learning, Memory, and Attention

Acetylcholine is an important neurotransmitter throughout the animal kingdom. In mammals, four major classes of acetylcholine-containing (cholinergic) neurons have been identified.

The first category includes neurons that are contained entirely within the peripheral autonomic nervous system; all postganglionic parasympathetic neurons and some specialized postganglionic sympathetic neurons (innervating sweat glands and some blood vessels) are cholinergic.

The second category of cholinergic neurons are those that originate within the central nervous system (brain and spinal cord) and project to the periphery; these include the skeletomotor neurons and the preganglionic neurons of the parasympathetic and sympathetic nervous systems.

The final two categories of cholinergic neurons are contained entirely within the central nervous system: local circuit or interneuronal cholinergic neurons, and cholinergic projection pathways. Cholinergic interneurons are found in the striatum and retina. The other cholinergic cells of the central nervous system are within the basal forebrain. The basal forebrain region includes the medial septal area and the nucleus basalis. Cholinergic cell bodies in the medial septal area project axons to the hippocampal formation; those in the nucleus basalis innervate the amygdala and send a topographic projection to the entire cortical mantle; and those in the reticular formation and periventricular area send a major projection pathway to the thalamus.

Acetylcholine is synthesized in neurons that express the enzyme choline acetyltransferase using two precursors from the diet. The first, choline, is found in many foods, particularly those that contain lecithin (a common food additive). The acetyl

group used to produce acetylcholine is obtained from glucose. Acetylcholine is synthesized in the axonal terminal within the cytoplasm and is then actively transported into synaptic vesicles for storage and release. Each synaptic vesicle contains approximately 5000 to 10 000 molecules of acetylcholine. The regulation of its synthesis is controlled by end-product inhibition: higher levels of acetylcholine within the terminal inhibit production. The production of acetylcholine is also controlled by the availability of both of its precursors. Giving additional choline alone will not produce or release more acetylcholine; therefore, it is not possible to enhance cholinergic function by eating a diet high in choline. Once released, acetylcholine is metabolized by the enzyme acetylcholinesterase into choline and acetic acid. Most of the choline released into the synaptic cleft is taken back up into the terminal and used to produce new acetylcholine molecules.

Cholinergic neurons are implicated in several forms of higher-order behavior. Most notably, the inputs to hippocampus, amygdala, and cortex appear to participate in sleep, learning, memory, and attention. The general function of any neurotransmitter can sometimes be inferred from the consequences of its loss from the brain or from the consequences of pharmacological manipulation. For example, drugs that impair cholinergic function, such as hyoscine (scopolamine) and atropine, also impair memory; drugs that enhance cholinergic function, such as physostigmine, tend to enhance learning and memory abilities. In addition, a characteristic feature of Alzheimer disease is the marked loss of cholinergic cells within the basal forebrain that roughly parallels the decline in cognitive function.

Noradrenaline (Norepinephrine): The Arousing Transmitter

Noradrenaline (norepinephrine) belongs to the class of neurotransmitters known as catecholamines (this group also includes dopamine and adrenaline). Catecholamines have been identified as neurotransmitters in a variety of species ranging from insects to humans. This transmitter is used for intercellular communication in both the central nervous system and peripheral nervous system.

Almost all of the postganglionic sympathetic neurons are noradrenergic; i.e. they use noradrenaline as a major transmitter. Within the central nervous system, there are several discrete cell clusters (nuclei) of noradrenergic neurons and all of these are located in the lower brainstem and pons. The locus ceruleus is the largest cell group with at least half of all central noradrenergic cells being located at this site in the caudal pontine central gray matter (in humans, about 12 000 cells per side). The noradrenergic cells of the locus ceruleus send axons that terminate throughout the brain and spinal cord (mainly ipsilaterally) to innervate all areas of the brain. Within the locus ceruleus, the more caudally located cell groups send descending projections to the brainstem and spinal cord, whereas the more rostrally located cell groups innervate the forebrain and telencephalon, with the exception of the basal ganglia.

The production of noradrenaline and the other catecholamine neurotransmitters is from dietary precursors and requires five different enzymes. These enzymes are very similar to each other and are produced by genes that lie close together, somewhat like beads on a string. These enzymes exist within the brains of all five classes of vertebrates and several invertebrates. The high degree of nucleotide homology between these enzymes suggests that they may have evolved from duplication of a common ancestral gene.

It is important to appreciate that not all of these synthetic enzymes are expressed in every catecholamine neuron. When more than one enzyme is expressed in a neuron, their regulation is usually linked in a coordinated fashion leading to the production of a specific neurotransmitter. Occasionally during embryogenesis, selected enzymes are transiently expressed and then disappear. This is why some neurons begin life containing noradrenaline but later on contain dopamine.

The main rate-limiting enzyme in the production of the catecholamine neurotransmitters is tyrosine hydroxylase. Therefore, the production of noradrenaline requires the amino acid tyrosine from the diet. For this reason, the content and balance of dietary amino acids is important for synthesis of the brain's major neurotransmitters. Shifting the concentration of different precursor amino acids in the diet can alter the uptake of each nutrient into the brain and may limit or alter neurotransmitter production. With a balanced diet, uptake of all amino acids is fairly constant and in correct proportion. However, too much of one amino acid will offset the metabolism of the others.

Once released from the terminal, noradrenaline acts upon a distinct class of receptors whose classification, as with many other transmitter systems, is based on anatomical and pharmacological criteria. Noradrenergic neurons are thought to govern the activity of preganglionic autonomic neurons and to participate in neuroendocrine regulation. In

addition to the brain, the chromaffin cells of the adrenal medulla also release noradrenaline into the bloodstream upon exposure to stressful stimuli.

Noradrenergic pathways within the central nervous system have been implicated in virtually all aspects of brain function from homeostatic processes to advanced behaviors such as learning and memory. Locus ceruleus projections may also mediate selected feeding behaviors and a global-orienting and arousal function to environmental stimuli. Drugs that enhance the function of noradrenaline, such as amphetamines, also enhance arousal.

Dopamine: The Rewarding Transmitter

Although dopamine is the predominant neurotransmitter in species that evolved prior to molluscs, this neurotransmitter has a rather limited distribution within the central nervous system compared with the other classical neurotransmitters.

Dopamine is used by several pathways within the central nervous system. The tuberohypophysial system (also known as the tuberoinfundibular system) consists of cell bodies in the arcuate nucleus of the hypothalamus and periventricular nucleus that project axons to the median eminence and the intermediate and neural lobes of the pituitary and control the release of prolactin. Another dopamine projection system originates in the ventral mesencephalon. The mesocorticolimbic dopamine neuronal system is composed of cell bodies located in the ventral tegmental area which send projections primarily to frontal lobe structures, including the nucleus accumbens, olfactory tubercles, amygdala, septal area, and the prefrontal, cingulate, piriform, and entorhinal areas of the cortex. The nigrostriatal pathway consists of cell bodies in the substantia nigra that send axons to the striatum (caudate and putamen).

These dopaminergic systems serve different functions. For example, the nigostriatal system regulates postural reflexes and inhibition of motor output. The mesocorticolimbic dopamine system is a critical element of brain reward pathways, whereas the tuberoinfundibular dopamine system governs pituitary hormone release.

Decreased levels of dopamine in the nigrostriatal pathway result in Parkinson disease, the symptoms of which include tremors, spasticity, and akinesia (difficulty in initiating or stopping movements). In contrast, increased levels of dopamine in the forebrain are associated with stereotypy in rats and schizophrenia in humans. One of the best understood actions of forebrain dopamine is its role in reward and euphoria. Virtually every known drug or rewarding experience stimulates, directly or indirectly, the forebrain dopamine neurons. This system is stimulated by the consumption of all known addicting substances, such as cocaine, heroin, and nicotine. Interestingly, its activity is not influenced by ingestion of caffeine.

Serotonin: The Most Complicated Neurotransmitter System

Serotonin or 5-hydroxytryptamine is found in many non-neuronal sites such as platelets, mast cells, and enterochromaffin cells. This amine is also used by central nervous system projection pathways that provide diffuse innervation throughout the brain and spinal cord.

There are nine major serotonergic nuclei and these are located in the midline areas of the pons and upper brainstem. The more caudal nuclei project to terminal fields within the brainstem and spinal cord. The rostral nuclei send diffuse ascending projections throughout the telencephalon and cerebellum; many of the terminal fields are overlapping. All areas of the forebrain receive input from these nuclei.

Serotonin was first found in serum and determined to have tonic effects on the cardiovascular system, which explains its name. It is found in the gastrointestinal tract of virtually every animal species examined and has also been found in the venom of amphibians, wasps, and scorpions – its effects upon the vasculature (vasodilation) assist in the action and absorption of the venom. Even planaria have serotonin receptors.

In humans, approximately 90% of total body serotonin is found in the gut, about 8% is found in platelets and mast cells in the blood, and only about 1–2% is located in the central nervous system, most of it within the pineal gland, which is not really part of the nervous system. The remainder is found in the raphe nuclei within the reticular region of the brainstem. Ascending pathways from these nuclei pass via the medial forebrain bundle to innervate the limbic structures, hypothalamus, septum, cingulate gyrus, cerebellum, superior colliculi (to influence vision), hippocampus, and neocortex. Serotonin terminals also make contact with glial cells and blood vessels; this might explain the role of serotonin in the onset of migraine headaches.

Although serotonin neurons represent less than 0.01% of neurons in the brain, their axonal terminals ramify widely to innervate almost every region of the brain. Serotonin neurons have a regular, slow spontaneous firing rate (about 0.5 to 2.5 spikes per

second) which suggests that they are constantly dribbling serotonin everywhere in the brain. This pattern of activity does not suggest that real information transfer is occurring. It is thought that these neurons provide only modulation of other neural activity within the region they innervate. This may explain why it has been so difficult to define a specific function for this neurotransmitter system. The spontaneous activity of these neurons is slowed significantly during slow wave sleep and eliminated completely during dream sleep. Undoubtedly, serotonin neurons play a role in the control of sleep onset and maintenance. Serotonin is also involved in adjusting the sleep–wake cycle to the photoperiod and regulation of the circadian rhythms.

The production of serotonin requires the amino acid tryptophan as a precursor from the diet. Uptake of tryptophan depends upon the level of other amino acids in the blood. A high-carbohydrate diet enhances tryptophan uptake into the brain. Approximately 1% of the tryptophan taken up into the brain is converted by the enzyme tryptophan hydroxylase into the transmitter serotonin. The production of serotonin is tightly controlled. It is known that lack of tryptophan in the diet will impair the formation of serotonin, but the presence of additional tryptophan does not lead to increased function of these neurons in spite of the fact that the action of the synthetic enzyme tryptophan hydroxylase is not rate-limiting. Any excess serotonin molecules that are produced appear to be metabolized within the cytoplasm, not released. Synthesis occurs within the neuronal terminals, so tryptophan hydroxylase is made in the cell body and shipped to the terminals.

Once released, serotonin can act upon a large variety of receptors: at least fourteen have been identified thus far. Pharmacologists have developed many different drugs that target these receptors; these drugs have been used to treat anxiety and depression as well as migraine headaches. Many naturally occurring drugs such as mescaline, dimethyltryptamine, and psilocybin can also influence serotonergic receptor function and produce hallucinations. The termination of action of serotonin, as for dopamine and noradrenaline, is by reuptake. Catabolism by various enzymes is a minor component in the termination of the action of serotonin.

Elevation of brain hippocampal levels of serotonin has been associated most often with impairment of learning and memory abilities. Lesion of the serotonin neurons can induce hyperactivity and hyperresponsivity to novel stimuli and increased motor activity. Serotonin neurons may also have a role in feeding behavior, thermoregulation, and sexual behavior. Serotonergic pathways are also implicated in mediating arousal behaviors. Finally, abnormal serotonergic transmission is thought to underlie psychiatric disorders such as major depressive illness and obsessive–compulsive disorder.

Neuropeptides

Most peptides that serve as intercellular messengers within the central nervous system range in size from 3 to 50 amino acid residues. More than 50 neuropeptides have been identified to date and there is no doubt that countless more will be discovered. Neuroanatomists have described multiple peptide pathways throughout the central and autonomic nervous systems and several peptides are known to exist within the dorsal root ganglion sensory neurons. Many peptides found within the nervous system are also produced in non-neural tissues such as the gastrointestinal tract and endocrine system.

Neuropeptides exert potent effects upon a variety of processes when administered into the brain, and specific peptides are thought to be involved in mediating several diverse functions. It is well established, for example, that certain peptides known as hypothalamic releasing hormones are the primary regulators of pituitary hormone secretion. Interestingly, some of the hypothalamic releasing hormones are also distributed outside the hypothalamus in brain areas controlling behavioral and autonomic outputs. On the basis of brain distribution and actions, some neuropeptides are believed to play a unique part in homeostatic processes by integrating appropriate endocrine, autonomic, and behavioral responses to environmental stimuli.

Amino Acid Neurotransmitters

The neurotransmitters discussed above are released throughout the brain and are, in general, involved with modulation of neural activity within the brain regions they innervate. For example, they may increase the likelihood of a memory, but are not responsible for the actual molecular mechanisms underlying the neural plasticity. In addition, these transmitter molecules are produced primarily for the purpose of neuronal communication. In contrast, some amino acids have multiple duties within the central nervous system. These amino acids are used for intermediary metabolism related

to protein synthesis or neuronal repair as well as for neurotransmission.

The two most important amino acid neurotransmitters that have been identified are glutamate and GABA. Neurons that contain and release glutamate and GABA are found throughout the central nervous system. Chemically these two neurotransmitters are very similar: indeed, glia actively convert glutamate into GABA, and vice versa, by the addition or subtraction of a molecule of carbon dioxide. This chemical similarity related to the presence or absence of carbon dioxide is a common phenomenon in nature; for example, some mushrooms contain both ibotenic acid (which acts like glutamate in the brain) and muscimol (which acts like GABA in the brain). As expected, ibotenic acid and muscimol differ only by a molecule of carbon dioxide.

Glutamate

Glutamate and GABA are the brain's major excitatory neurotransmitters, providing point-to-point innervation of brief and fast excitatory input. Almost everywhere that glutamate is released it produces a strong postsynaptic excitation. This excitation can lead to either a change in function of the synapse, often called neuroplasticity, or to the destruction of the postsynaptic neuron. These two actions of glutamate have different roles in the brain at different times of life, and are influenced by the response of specific glutamate receptors to the presence of elevated levels of glutamate.

For example, during development and maturation of the brain, elevated extracellular levels of glutamate may lead to the pruning of excess neuronal synapses or the destruction of selected neurons. The loss of excess synapses and neurons is thought to improve information transfer and processing within the maturing brain. After these processes are completed, the brain alters the way in which it uses glutamate and exerts greater control over the concentration of glutamate within the synaptic cleft. During adulthood, extracellular levels of glutamate are carefully controlled in order to avoid overstimulation of its receptors and the inappropriate destruction of postsynaptic neurons. With advanced aging, extracellular levels of glutamate once again may become involved with the degeneration of selected neural systems.

Recent research has focused upon pharmacological methods to attenuate the degenerative actions of glutamate in order to slow the progress of neurological disorders such as Alzheimer and Huntington diseases.

GABA

Gamma-aminobutyric acid neurons can be found throughout the brain. They exist mostly as small interneurons that provide local circuit control of regional neural networks. However, there are also a few prominent long pathways that release GABA. The widespread nature of these underlies an interesting feature of this neurotransmitter system, which is that a large percentage of synapses are influenced by GABA and are therefore inhibitory. This feature underlies an important proclivity of neural function, that most neural processing involves inhibition rather than excitation.

Receptors for GABA come in two types that are distinguished by the type of drugs that bind to them. These receptors serve as binding sites for drugs that depress central nervous system function. Alcohol, barbiturates, and the benzodiazepine drugs (e.g. diazepam) all bind to GABA receptors, enhancing their function and thereby increasing the level of inhibition in neural networks throughout the brain. The consequence is a slowing a neural activity and the subsequent loss of consciousness.

COTRANSMISSION OF NEUROTRANSMITTERS

It is now known that a single neuron may contain several transmitters in different combinations. Most, if not all, neurons contain two or more different neurotransmitters, such as a neuropeptide in combination with a classical neurotransmitter. Depending on the neuronal firing rate, coexisting transmitters may be released separately or together. Cotransmitters may either potentiate or attenuate each other's effects. For example, the neuropeptide galanin is present in acetylcholine-containing neurons that innervate the hippocampus; the release of galanin appears to attenuate the action of acetylcholine upon the postsynaptic neurons within this brain region.

Although the functional significance of coexisting transmitters is as yet not completely understood, a clear implication is the enormous potential for highly specific and differentiated intercellular communication.

CONCLUSION

Neurons communicate with target cells and other neurons through chemical means. A large number of chemically diverse compounds serve as neurotransmitters in the central and peripheral nervous

systems. Many neurons contain more than one transmitter substance. Projection pathways typically contain the classical transmitters (acetylcholine, noradrenaline, dopamine, serotonin), while the peptide transmitters coexist within these neurons. Many neuropathological disorders are associated with reductions or excesses of particular neurotransmitters. The rich diversity of neurotransmitter substances underscores the complexity of neural communication and also provides pharmacological opportunities to modify information processing within the central nervous system.

Further Reading

Charney DS, Nestler EJ and Bunney BS (eds) (1999) *Neurobiology of Mental Illness*. New York, NY: Oxford University Press.

Cooper JR, Bloom FE and Roth RH (eds) (1996) *The Biochemical Basis of Neuropharmacology*. New York, NY: Oxford University Press.

Julien RM (1999) *A Primer of Drug Action*. New York, NY: WH Freeman.

Kandel ER, Schwartz JH and Jessell TM (1991) Elementary interactions between neurons: synaptic transmission. In: *Principles of Neural Science*, pp. 23–269. Norwalk, CT: Appleton & Lange.

Purves D, Augustine GJ, Fitzpatrick D *et al*. (1997) Neurotransmitters. In: Purves D, Augustine GJ, Fitzpatrick D *et al*. (eds) *Neuroscience*, pp. 99–120. Sunderland, MA: Sinauer.

Siegel GJ, Agranoff BW, Albers RW, Fisher SK and Uhler MD (eds) (1999) *Basic Neurochemistry*. Philadelphia, PA: Lippincott-Raven.

Non-Alzheimer Dementias

Introductory article

Esma Dilli, University of British Columbia, Vancouver, British Columbia, Canada
Bruce L Miller, University of California School of Medicine, San Francisco, California, USA

CONTENTS

Significant degenerative disorders other than Alzheimer disease that cause dementia include the frontal temporal lobe dementias, Parkinson-like dementias, and prion-related diseases.

INTRODUCTION

The most common type of dementia is Alzheimer disease; however, there are other significant degenerative disorders that cause dementia. These non-Alzheimer dementias vary in their symptomatology and treatment. Therefore, understanding the pathophysiology and anatomy of these dementias will help to provide the most effective treatment for these patients. Non-Alzheimer dementias can be categorized into frontal temporal lobe dementias, Parkinson-like dementias, and prion-related dementias. Each of these categories can be further divided according to the location of the disease process in the brain and/or the signs and symptoms surrounding the disease.

FRONTAL TEMPORAL LOBE DEMENTIA

Frontal temporal lobe dementia (FTLD) is the third most common type of dementia after Alzheimer disease and Lewy body dementia. It accounts for approximately 16% of all patients with primary degenerating dementia. There are three subtypes of FTLD, which are characterized by their presenting symptoms as well as by the involvement of the frontal temporal lobe: frontal temporal dementia, progressive nonfluent aphasia, and semantic dementia.

Frontal Temporal Dementia

Frontal temporal dementia (FTD) causes a gradual deterioration in behavior and language. The behavioral signs include social inhibition and disordered social conduct such as a decrease in manners, social graces, and decorum. Also, patients demonstrate an early impairment in the regulation of personal conduct where they act passively, or talk too much and constantly interrupt others. Other behavioral features of FTD include loss of insight and distractibility. Difficulty with planning and organizing at home or work reflects the decrease in executive function in the frontal lobe. Hyperorality, weight gain, altered preference for food (such as craving sweet foods only) and/or excessive smoking and alcohol consumption may be due to a combination of frontotemporal dysfunction and a decrease in the amount of brain serotonin. Stereotyped perseverative behaviors such as hand rubbing, collecting items, and impulse shopping are also common symptoms. People with FTD demonstrate lack of concern for others and emotional blunting. Criminal behavior such as theft is higher in FTD patients (10%) than in Alzheimer patients (1–2%), reflecting the socially inappropriate behaviors that are present in FTD. Some people with FTD also develop amyotrophic lateral sclerosis (ALS). Around 60% of FTD cases occur sporadically and the other 40% are familial. Approximately 80% of these familial cases are autosomal dominant. Mutations in the tau gene on chromosome 17 account for a small percentage of familial cases.

Although there is no cure for FTD, treatment is available to alleviate its symptoms. Antidepressant drugs such as selective serotonin reuptake inhibitors (SSRIs) can be used to treat the depression as well as reducing compulsions, disinhibition, and hyperorality.

Progressive Nonfluent Aphasia

Progressive nonfluent aphasia is a disorder of expressive language. Patients present with word-finding deficits, trouble in organizing their thoughts, and slowing of word output, which can lead to stuttering. Speech is sparse (decreased word output and phrase length) and effortful. Patients

have problems both expressing and understanding the grammatical elements of speech. 'Anomia' is the term used for the difficulty in naming manifested by inability or slowness to find the correct word. Other signs that suggest progressive nonfluent aphasia are difficulty with reading and writing, as well as impaired repetitions. People with progressive nonfluent aphasia also display verbal apraxia (speech disarticulation). The behavioral changes emerge late in the disease, and usually present as apathy, depression, and social withdrawal. Insight and personal awareness remain intact in these patients. Memory is relatively spared in the early stages of the disease. The brain atrophy is usually asymmetrical and involves the left frontal temporal lobe. The pathology of progressive nonfluent aphasia shows a classical FTLD pattern.

There is no cure for progressive nonfluent aphasia, but speech therapy can be used to improve verbal apraxia.

Semantic Dementia

Semantic dementia is a temporal variant of FTLD causing progressive deterioration of semantic knowledge about people, objects, facts, and words. People with this disorder lose the meaning of words even though they might be able to read, repeat or spell the words phonetically. They perform poorly on tests that require conceptual knowledge such as picture naming, category fluency (e.g. the patient is asked to name all the words in a particular category such as animals) and word–picture matching (matching written or spoken words to pictures). Difficulties in naming (anomia) and impaired verbal comprehension are common complaints. The speech is fluent and grammatically correct with normal flow, but empty in content. Phonology, syntax, and other nonverbal aspects of speech are preserved. Memory for day-to-day events (episodic memory) such as remembering recent personal events and appointments is well preserved, unlike the loss of recent memory in Alzheimer disease. Visuoperceptual and spatial abilities, and nonverbal problem-solving, are commonly preserved until late in the disease. However, as the syndrome progresses a number of patients develop visual deficits such as prosopagnosia, the inability to recognize a familiar face such as that of a relative.

Neuroimaging typically reveals bilateral focal atrophy of the inferolateral temporal neocortex, especially the left anterior temporal cortex, with sparing of the hippocampus. Because the hippocampus is involved in the acquisition and storage of recent episodic and semantic memory, patients with semantic dementia have better recall of recent autobiographical memory while Alzheimer patients have better recall of past events. Most postmortem studies show non-Alzheimer pathological changes with either nonspecific neuronal loss, gloss or spongiform changes.

PARKINSON-LIKE DEMENTIA

Lewy Body Dementia

Lewy body dementia (LBD) is the second most common type of dementia after Alzheimer disease, being responsible for approximately 15–25% of all cases of dementia. It usually begins in the fifth to seventh decade of life, and is twice as common in men than in women. Patients show a progressive cognitive decline with deficits in attention and visuospatial ability. The symptoms include fluctuating alertness, recurrent visual hallucinations (seeing things that are not real) and motor symptoms of Parkinson disease. Patients demonstrate a significant degree of cognitive impairment such as progressive loss in memory, language, visuospatial construction, and reasoning.

Visual hallucinations occur in about 60–80% of people with LBD and are usually recurrent, well formed, and detailed, with themes of people and animals. The hallucinations can occur at any time but are often worse during times of acute confusion. Auditory hallucinations are less common than visual ones in LBD. The neurochemical basis of the hallucinations involves an imbalance between acetylcholine and dopamine. Besides the visual and auditory hallucinations, people with LBD also present with other neuropsychiatric symptoms, such as delusions (including paranoid delusions) and depression. Rigidity and slowing of movement are common in LBD. Slow, shuffling gait, stooped posture, mask-like face, and hypophonic speech are also signs of the parkinsonism typical of LBD. Other clinical features include a history of recurrent unexplained falls or loss of consciousness.

Lewy bodies are the main histopathological feature. Classic LBD shows tau-negative, spherical, red-staining neuronal inclusion bodies. These Lewy bodies stain for ubiquitin and α-synuclein. They may be single or multiple within the neuron and may vary in size. Other associative pathological features of LBD include Lewy-related neuritis, plaques, neurofibrillary tangles, regional neuronal loss of substantia nigra, locus ceruleus and nucleus basalis of Meynert, and microvacuolation. There is no known etiological or risk factor for

LBD. There are a few rare familial cases of LBD related to an autosomal dominant pattern of inheritance. Lewy body disease usually ends in death; mean survival is 3–9 years after diagnosis.

The visual hallucinations may be alleviated by cholinesterase inhibitor agents such as donepezil. If extrapyramidal motor signs are prominent in an individual with LBD, moderate use of antiparkinsonian drugs to enhance dopamine function can be helpful, but may exacerbate psychotic symptoms. It is important that people with LBD avoid neuroleptic tranquilizer medications, which might worsen the parkinsonian symptoms. Therefore atypical antipsychotic neuroleptic medications are recommended to alleviate the psychotic symptoms of hallucinations and delusions.

Progressive Supranuclear Palsy

Progressive supranuclear palsy (PSP) causes progressive nonfluent aphasia due to frontal lobe degeneration. The incidence is approximately 1 in 100 000. Around 8% of people diagnosed with clinical Parkinson disease eventually present with PSP. The onset of the disease is often in the sixth decade (age range 45–75 years). Patients often present with supranuclear ophthalmoplegia (inability to move the eyes). At first the patient has difficulty in looking up and down; later, as the disease progresses, there is deterioration in horizontal voluntary eye movements. This impairment can affect the patient's daily activities such as reading, cooking, or walking. Patients with PSP also have an increased incidence of repetitive falls. Other symptoms of PSP include hypophonia (low volume in speech) and dysarthria due to weakness and loss of control of throat and mouth muscles, related to a condition called 'pseudobulbar palsy'. The patient may also complain of stuttering, and have trouble swallowing and chewing (dysphagia). The parkinsonian symptoms of PSP include axial rigidity (slow and stiff movement) of neck and trunk with relative sparing of the limbs. The axial rigidity is particularly demonstrated with abduction and lateral rotation of the neck. Additional motor symptoms of PSP consist of hyperextended gait and little or no tremor.

The behavioral symptoms of PSP include profound apathy, loss of motivation, problems with planning and organization, and irritability. However, patients appear less frustrated and disinhibited than people with FTD. Dementia is present in most patients. They tend to become forgetful and slow in their thinking. On physical examination, the patient with PSP shows square wave jerks (small, inappropriate saccades that intrude on steady fixation) and decreased saccade velocity. In contrast, people with corticobasal ganglionic degeneration show preserved saccade velocity but an increase in saccade latency on the ipsilateral side of the apraxia. This distinguishing feature can be assessed by an electrooculogram (EOG), which tests eye movement velocity and latency.

The cause of PSP is unknown, but rare familial clusters have been described in which the pattern of inheritance appears to be autosomal dominant. Postmortem examinations show bilateral loss of neurons and gloss in periaqueductal gray matter, superior colliculus, subthalamus, red nucleus, pallidum, dentate nucleus, and oculomotor nuclei. The cerebral cortex is relatively spared. Positron emission tomographic scans show dorsal frontal hypoperfusion, while magnetic resonance images show midbrain atrophy.

There is no cure for PSP; however, levodopa can be effective in reducing the motor symptoms. In addition, speech therapy and encouraging the patient to eat soft food and drink thick liquids may be helpful in coping with the swallowing difficulties. Death usually occurs around 5–7 years after diagnosis.

Corticobasal Degeneration

Corticobasal degeneration is a rare, sporadic, progressive neurological disease characterized by a combination of parkinsonian and cortical symptoms. It often begins in the sixth decade. The symptoms begin initially on one side of the body then spread to affect both sides. This asymmetric pattern is a distinguishing feature of this disease. These asymmetrical motor deficits include focal dystonia (abnormal posture) and 'alien hand' – a combination of abnormal hand postures, involuntary movements, and changes in muscle tone. The early motor involvement in corticobasal degeneration includes bradykinesia and rigidity, poor coordination, and dysequilibrium (impaired balance). Additional motor symptoms are myoclonus and apraxia (loss of ability to make familiar purposeful movements). Additionally, patients show left or right parietal lobe deficits.

Corticobasal degeneration usually starts in the parietal lobe and spreads to the basal ganglia. The pathological changes include neuronal loss, gliosis, neuronal inclusions, and achromasia (ballooned neurons) in the superior frontal gyrus and parietal cortex. The absence of amyloid plaques and of neurofibrillary tangles differentiate this disorder from Alzheimer disease. Astrocytic plaques and

tau proteins are also present. The substantia nigra shows neuronal loss with gliosis.

There is no treatment to slow the course of corticobasal degeneration, and the symptoms of the disease are generally resistant to therapy. Antiparkinsonian drugs do not produce any significant or sustained improvement in the motor symptoms. Clonazepam may help the myoclonus. Occupational, physical, and speech therapy may help in managing disability. The course of corticobasal degeneration results in death, usually 6–8 years after diagnosis.

HUMAN PRION DISEASES

Human prion diseases are transmissible spongiform encephalopathies involving proteinaceous infectious particles that lack nucleic acids (DNA or RNA). The disease is due to biochemical modification of the prion protein (PrP) that is normally present in human cells. The normal PrP is converted to an abnormal isoform, which results in aberrant folding of the protein structure. The normal prion protein contains more of the tertiary α-helix conformation, while the abnormal prion protein contains more of the β-sheet tertiary conformation. There are five types of human spongiform encephalopathies: kuru, Creutzfeldt–Jakob disease (CJD), fatal familial insomnia, Gerstmann–Sträussler–Scheinker disease, and variant CJD.

Kuru

Kuru is a transmissible prion disease that occurs only in certain inhabitants of Papua New Guinea, and is believed to be due to cannibalism. The brains of patients with this disease show unicentric plaques.

Creutzfeldt–Jakob Disease

Creutzfeldt–Jakob disease is a rapidly progressive, fatal prion disease that can be familial or sporadic in its development. Eighty percent of the sporadic cases occur in the age range 50–70 years, and 10–15% of the familial cases are autosomal dominant. The mean survival is 5 months, with death occurring within 1 year in 90% of the sporadic cases. There are three separate groups of symptoms that appear early in the disease. One group consists of generalized systemic symptoms such as fatigue, poor sleep, and loss of appetite. The second is nonfocal neurological symptoms which include memory loss and confusion. The final group is focal neurological symptoms consisting of ataxia and aphasia. All patients later present with myoclonus, behavioral changes, and dementia. The diagnosis of CJD requires an EEG showing periodic complexes in the wave pattern, a distinct protein band in cerebrospinal fluid, and the presence of dementia and myoclonus in the patient. Neuropathological examination shows spongiform degeneration, neuronal loss, astrocytosis, and amyloid plaques in 5–10% of sporadic CJD cases.

Fatal Familial Insomnia

Fatal familial insomnia is a rare fatal prion disease usually due to a mutation in the prion protein gene. The onset is usually in people aged 35–61 years. Death occurs 7–36 months after the initial symptoms, which consist of untreatable insomnia, progressive dysautonomia (hypertension, irregular breathing, tachycardia) and motor signs such as ataxia and myoclonus. The family history, the symptoms, and the EEG recording (showing a slow background activity without the periodic complexes that occur with CJD) make up the diagnosis.

Gerstmann–Sträussler–Scheinker Disease

Gerstmann–Sträussler–Scheinker disease is a fatal autosomal dominant spongiform encephalopathy in which patients present with ataxia associated with other cerebellar signs. The incidence is approximately 5 per 100 million, making this disorder very rare. Pathological examination of the brain shows multicentric amyloid plaques that form in the cerebellum and cerebrum.

Variant CJD

Variant CJD (vCJD) is another fatal progressive prion disease which has characteristic differences from sporadic CJD. This prion disease is due to ingestion of meat that contains the bovine spongiform encephalopathy agent. The mean onset of the disease is at age 29 years, which is earlier than that of sporadic CJD. The duration of symptoms is longer in vCJD and patients survive approximately 14 months. In addition, people with vCJD present with psychological symptoms first, while people with sporadic CJD often develop the dementia first. Cerebellar signs are almost always present in vCJD, whereas only 40% of people with sporadic CJD show cerebellar signs such as ataxia. People with vCJD do not show periodic complexes in the EEG.

Further Reading

Brun A (1993) Frontal lobe dementia of the non-Alzheimer type revisited. *Dementia* **4**: 126–131.

Demaerel P, Heiner L, Robberecht W *et al*. (1999) Diffusion-weighted MRI in sporadic Creutzfeldt–Jakob disease. *Neurology* **52**(1): 205–208.

Feany MB and Dickson DW (1996) Neurodegenerative disorders with extensive tau pathology: a comparative study and review. *Annals of Neurology* **40**: 139–148.

Hodges JR and Patterson K (1996) Nonfluent progressive aphasia and semantic dementia: a comparative neuropsychological study. *Journal of the International Neuropsychological Society* **6**: 511–524.

Hong M, Zhukareva V, Vogelsberg-Ragaglia V *et al*. (1998) Mutation-specific functional impairments in distinct tau isoforms of hereditary FTDP-17. *Science* **282** (5395): 1914–1917.

Litvan I (1998) Progressive supranuclear palsy: staring into the past, moving into the future. *Neurologist* **4**: 13–20.

Miller BL, Darby AL, Swartz JR, Yener GG and Mena I (1995) Dietary changes, compulsions and sexual behavior in fronto-temporal degeneration. *Dementia* **6**: 195–199.

Miller BL, Boone K, Geschwind D and Wilhelmsen K (1999) Pick's disease and frontotemporal dementias: emerging clinical and molecular concepts. *Neurologist* 5205–5212.

Mori H, Nishimura M, Namba Y and Oda M (1994) Corticobasal degeneration: a disease with widespread appearance of abnormal tau and neurofibrillary tangles, and its relation to progressive supranuclear palsy. *Acta Neuropathologica (Berlin)* **88**(2): 113–121.

Neary D, Snowden JS, Gustafson L *et al*. (1998) Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology* **51**: 1546–1552.

Steele JC, Richardson JC and Olszewski R (1964) Progressive supranuclear palsy. A heterogeneous degeneration involving the brain stem, basal ganglia, and cerebellum with vertical gaze and pseudobulbar palsy, nuchal dystonia and dementia. *Archives of Neurology* **10**: 333–354.

Stevens M, van Duijn CM, Kamphorst W *et al*. (1988) Familial aggregation in frontotemporal dementia. *Neurology* **50**: 1541–1545.

Will RG, Zeidler M, Stewart GE *et al*. (2000) Diagnosis of new variant Creutzfeldt–Jakob disease. *Annals of Neurology* **47**: 575–582.

Nootropic Drugs

Introductory article

Warren H Meck, Duke University, Durham, North Carolina, USA
Christina L Williams, Duke University, Durham, North Carolina, USA

CONTENTS

Nutrients and other chemicals that amplify cognition are sometimes called 'nootropics'. These 'smart drugs' are thought to work by one of two basic processes, either by influencing glucose metabolism and blood flow in the brain, or by elevating the levels of one of the many neurotransmitters or neuromodulators that are known to play a role in attention, memory and decision processes.

CRITERIA THAT MUST BE MET BEFORE A DRUG CAN BE ESTABLISHED AS NOOTROPIC

The term 'nootropic' was coined by the pharmacologist Cornelius Giurgea in the 1970s (its origin is Greek, meaning 'acting on the mind'). In terms of chemistry, nootropic drugs are a highly heterogeneous group. Giurgea argued that to qualify as a nootropic, a drug had to exhibit the following characteristics. There must be reliable evidence of enhanced cognition, especially under conditions of disturbed neural metabolism resulting from hypoxia, electroconvulsive shock or age-related changes. Nootropic substances must also increase the efficacy of tonic cortical/subcortical control mechanisms and facilitate information flow between the cerebral hemispheres. The prophylactic effects of the drug were also taken into consideration (i.e. the importance of enhancing the general resistance of the brain to physical and chemical injuries was emphasized). Finally, the drug had to be devoid of any other psychological or physiological effects, including addictiveness or neurotoxicity. In order to streamline these criteria, most neuroscientists now define nootropics as a class of drugs that act as cognitive enhancers and which have no undue side-effects or toxicity.

One group of nootropics works by improving cerebral blood flow, and may be used to improve the deficits in attention, language or memory that are commonly seen following the neurological damage associated with stroke. The damage is often caused by constriction of the arteries that supply blood, and therefore oxygen and glucose, to the brain. Consequently, drugs that increase cerebral blood flow and reduce hypertension might be expected to 'improve the performance' of otherwise energy-starved neurons. This is why propranolol, phenytoin and hydergine are used as smart drugs. Because drugs such as hydergine also act as free-radical scavengers, they may be effective in preventing strokes and combating age-related neuronal insults.

The calcium-channel blockers, such as verapamil, diltiazem, nifedipine, nitrendipine and nimodipine, are another group of potential smart drugs. They are widely used to treat hypertension, and thus might affect cognition by increasing cerebral blood flow, but they also block the entry of calcium ions into neurons. A key stage in the molecular cascade of memory formation appears to be the opening of membrane channels through which calcium ions flow. In aged rodents, the mechanisms that open and close these calcium channels seem to become defective, and consequently calcium accumulates to dangerously high levels inside the cells. The calcium-channel blockers might improve memory by countering this effect. Ginseng, that all-purpose elixir of life, contains an agent that blocks calcium channels.

There is some evidence that calcium-channel blockers, glutamate agonists, serotonin antagonists and certain piracetam derivatives can help laboratory animals to learn highly specific tasks if the drug is injected around the time of training. Yet in many cases it is not always clear how the drugs act. Some may work by interfering with the molecular events that lead to memory formation, while others may affect the animal's general level of arousal. Clearly, even the simplest psychological tests involve many

more aspects of performance than merely learning and memory. Drugs that alter sensations of hunger or thirst, or which reduce sensitivity to pain or simply make a participant more active, could all improve performance. For these reasons it is important to separate the effects of a drug on different memory processes (e.g. encoding, consolidation and retrieval) as well as its effects on different types of memory (e.g. episodic and semantic).

A third group of nootropics has been developed as a result of discoveries about the neurochemical deficits associated with Alzheimer disease. One of the characteristic features of that disease is progressive loss of memory function. Post-mortem studies of the brains of Alzheimer patients show dramatic destruction of neurons, particularly the basal forebrain neurons which secrete or utilize the neurotransmitter acetylcholine. Research conducted during the 1970s in humans and other animals also provided evidence that acetylcholine might be a key neurotransmitter in memory formation. For example, acetylcholine-receptor-blocking drugs (e.g. atropine and scopolamine) that act by reducing the effective level of brain acetylcholine also impair memory in young healthy subjects. In the early 1980s, these types of observations gave rise to the 'cholinergic hypothesis' of memory loss. If memory loss was due to low levels of acetylcholine or to loss of functioning within the cholinergic system, then researchers reasoned that what was needed were smart drugs to increase the brain's supply of acetylcholine and to amplify its actions. As a result of this reasoning, several smart drugs were developed, including piracetam, together with food additives that are potential metabolic precursors of acetylcholine, such as choline and acetylcarnosine.

Clinical trials have provided convincing evidence that some drugs (e.g. the acetylcholine-enhancing drug tacrine) may be of help in the early stages of Alzheimer disease. In addition, the results of several naturalistic studies suggest that perimenopausal estrogen replacement may reduce the risk of developing Alzheimer disease by up to 50%, although other studies did not find this benefit. Preliminary results from one large epidemiological study suggest that estrogen replacement may delay the memory deficits that are associated with normal aging.

However, additional controlled clinical trials are needed to determine how and under what circumstances estrogen replacement has nootropic effects. Estrogen influences language skills, mood, attention and a number of other functions in addition to memory. Estrogen receptors are present in several regions of the brain, including those involved in memory (such as the hippocampus), and increases in synapse formation (i.e. increased brain plasticity) with estrogen treatment have been observed. In addition, estrogen may in effect raise the levels of certain neurotransmitters within the brain. These include acetylcholine (implicated in memory), serotonin (implicated in mood), noradrenaline (implicated in mood and other autonomic functions) and dopamine (implicated in attention, interval timing, motor coordination and reward). However, it is important to note that the evidence for these beneficial effects of estrogen on cognition is still preliminary.

EXAMPLES OF SUBSTANCES THAT HAVE BEEN CLAIMED TO IMPROVE MEMORY IN NORMAL SUBJECTS

Although some drugs do improve memory in different tasks in normal human subjects, many neuroscientists are careful to point out that the brain is a very complex system, and that cognition results from the coordination of many processes. It seems unlikely that a single chemical could improve memory. However, some experimental results suggest that there are certain drugs which improve memory. The nootropics that have been studied can be divided into several different categories as follows.

- Cholinergic agonists (e.g. strychnine and picrotoxin as well as nicotine and related substances) act by influencing or emulating the release of acetylcholine in the hippocampus. It is known that this neurotransmitter plays an important role in synaptic plasticity.
- Stimulants (e.g. caffeine and amphetamine) probably act by increasing the effective level of dopamine and/or norepinephrine, which leads to increases in alertness and attention, which can in turn improve memory consolidation.
- The piracetam family of compounds (e.g. piracetam, aniracetam, oxiracetam and others) serve as representative AMPAkine agents for which the hippocampal AMPA-glutamate receptor is often the target for effects on memory. The AMPAkine agents cover a broad class of drugs and vary widely in terms of brain penetration and efficacy. A number of the compounds in this class increase synthesis of neuronal membrane lipids such as phosphatidylinositol, phosphatidylcholine and phosphatidylethanolamine. AMPAkines appear to act by influencing synaptic strengthening or formation.
- Substances such as *Ginkgo biloba* extract and acetyl-L-carnitine may support increased attention or neural health in adults. These chemicals are thought to improve blood flow in cerebral structures and to facilitate memory.

- Hormones such as estrogen, vasopressin, epinephrine, norepinephrine and ACTH probably act by modulating the cholinergic input to the hippocampus from the basal forebrain although, as mentioned above, estrogen may directly modulate synaptic plasticity. Each of these substances may have a separate role in the brain and in the periphery, depending on their site of origin and their ability to cross the blood–brain barrier.
- The last category includes essential nutrients such as choline, folic acid, vitamin B_{12} and vitamin E. Choline in particular, which is a major component of cellular membranes and the precursor of the neurotransmitter acetylcholine, appears to affect the early development of brain structures involved in attention and memory through metabolic imprinting. This process results in organizational changes in the way in which acetylcholine is synthesized, released and recycled – leading to lifelong enhancement of cognitive processes and resistance to age-related memory decline.

EVALUATION OF EVIDENCE FOR THE ABOVE-MENTIONED EXAMPLES

For the most part, nootropics act at the level of cellular or intracellular mechanisms such as gene expression, transmitter release and associated changes in neural plasticity. It is at this level that adjustments can affect higher-level processes such as memory encoding. To date, most studies have been conducted with laboratory animals rather than with humans, and most of the human studies have focused on treating dementia or aging because of the lack of effect in healthy individuals. However, positive results have been obtained for healthy adults, and given that the memory systems of a wide range of animals appear to be quite similar, there is good reason to believe that memory-enhancing substances could work in healthy humans. There is also reason to believe that some aspects of brain aging could be slowed or potentially ameliorated.

There are fewer substances that are known to improve other aspects of cognition. Although we are well aware that drugs such as nicotine and the stimulants (e.g. caffeine, amphetamine and acetyl-L-carnitine) improve reaction time and attention, it is less clear whether higher-level processes such as intelligence or creativity are influenced by drugs. This may be because intelligence and creativity are difficult to measure. There might be whole classes of chemicals we have not yet found that influence higher-level cognition, although it is best to remain cautious about this.

The effects of nootropics can be highly individual, possibly linked to the amount of glucose, oxygen and steroids in the bloodstream and also to brain neurochemistry. For example, caffeine appears to improve recognition and recall differently depending on the level of internal arousal and the time of day. Nootropics need to be tuned to the brain chemistry of their users in order to be effective.

Finally, it is well known among researchers that the timing of drug administration can be an important factor. In many cases the drug has no effect if it is given before or during the learning experience, and is only effective if given after a certain time. Other drugs only work if they are administered before the learning experience. In fact, some drugs can even be given before birth. It has been found that if the diet of pregnant animals is supplemented with choline, their offspring will exhibit enhanced memory capacity and precision throughout their entire life. This appears to be due to permanent alterations in the septal–hippocampal pathway. Whether or not it would be a good idea to try this treatment in humans is an open question (it might be interesting to compare the children of mothers who eat choline-rich and choline-poor diets).

There are also the problems of dose response and pharmacokinetics. High doses of the drugs do not necessarily improve performance, and in fact they usually decrease it. The memory improvement typically follows an 'inverted U-curve', where there is an optimal dose that is dependent on the ways in which other environmental and genetic factors are affecting the participant's neurochemistry, as well as drug absorption and metabolism. Primarily for this reason the phrase 'the dose makes the drug' was coined – that is, the same drug can have very different behavioral effects depending on the dose and how it interacts with the contextual factors that determine pharmacokinetics.

It is important to note that in addition to the benefits provided by nootropics, it is likely that some of these improvements in cognition come at a price. For example, caffeine has been shown to improve sustained attention and vigilance, but it impairs the ability to deal with contradictory or uncertain stimuli. If we learn too well, we might remember details at the expense of general knowledge or abstraction ability. Enhanced memory capacity would be expected to lead to increased interference unless information was also stored with greater precision. All of these problems suggest that while nootropics can enhance memory, using them optimally is much less straightforward than merely swallowing a smart pill with one's coffee or orange juice in the morning.

PROSPECTS FOR THE FUTURE DEVELOPMENT OF NOOTROPICS

It is known that nootropic drugs and cognitive training are complementary and probably act on different systems. In general, it seems that improving the hardware (e.g. neural networks) can give rise to a quantitative improvement, while improving the software (e.g. hierarchical structures) can bring about a qualitative improvement. A well-integrated combination of nootropic drugs and cognitive strategies could become very powerful. It is likely that the future of cognitive amplification will be found in the gradual combination of different behavioral and pharmacological tools into unified systems, where cognitive science and neurobiology provide mutual support for each other.

Further Reading

Bartus RT (2000) On neurodegenerative diseases, models and treatment strategies: lessons learned and lessons forgotten a generation following the cholinergic hypothesis. *Experimental Neurology* **163**: 495–529.

Blusztajn JK (1998) Choline, a vital amine. *Science* **281**: 794–795.

Gouliaev AH and Senning A (1994) Piracetam and other structurally related nootropics. *Brain Research Reviews* **19**: 180–222.

Joseph JA, Shukitt-Hale B, Denisova NA *et al*. (1998) Long-term dietary strawberry, spinach or vitamin E supplementation retards the onset of age-related neuronal signal transduction and cognitive behavioral deficits. *Journal of Neuroscience* **18**: 8047–8055.

Kleijnen J and Knipschild P (1992) *Ginkgo biloba*. *Lancet* **340**: 1136–1139.

Levin ED, Conners CK, Silva D *et al*. (1998) Transdermal nicotine effects on attention. *Psychopharmacology* **140**: 135–141.

Meck WH and Williams CL (2002) Metabolic imprinting of choline by its availability during gestation: implications for memory and attentional processing across the lifespan. *Neuroscience and Biobehavioral Reviews* (in press).

Mondadori C (1996) Nootropics: preclinical results in the light of clinical effects. Comparison with tacrine. *Critical Reviews in Neurobiology* **10**: 357–370.

Rusted JM, Graupner L, Tennant A and Warburton DM (1998) Effortful processing is a requirement for nicotine-induced improvements in memory. *Psychopharmacology* **138**: 362–368.

Sandstrom NJ and Williams CL (2001) Memory retention is modulated by acute estradiol and progesterone replacement. *Behavioral Neuroscience* **115**: 384–393.

Sarter M, Dudchenko P, Moore H, Holley LA and Bruno JP (1992) Cognition enhancement based on GABA–cholinergic interactions. In: Levin ED, Decker MW and Butcher LL (eds) *Neurotransmitter Interactions and Cognitive Function*, pp. 329–354. Boston, MA: Birhauser.

Williams CL (2002) Hormones and cognition in nonhuman animals. In: Becker JB, Breedlove SM, Crews D and McCarthy MM (eds) *Behavioral Endocrinology*, pp. 527–577. Cambridge, MA: MIT Press.

Windisch M (1996) Cognition-enhancing (nootropic) drugs. In: Baskys A and Remington G (eds) *Brain Mechanisms and Psychotropic Drugs*, pp. 239–257. Boca Raton, FL: CRC Press.

Object Perception, Neural Basis of

Introductory article

Minoru Koyama, University of Tokyo School of Medicine, Tokyo, Japan
Isao Hasegawa, University of Tokyo School of Medicine, Tokyo, Japan
Yasushi Miyashita, University of Tokyo School of Medicine, Tokyo, Japan

CONTENTS

Object recognition is a process through which sensory inputs reactivate distributed neural networks subserving memory of objects in the temporal and other association cortices.

INTRODUCTION

Imagine yourself looking at an apple on a dish (Figure 1). Without conscious effort, you will perceive a red round object on a white background and recognize it as an 'apple', a kind of fruit with sweet and sour taste. It may remind you of the story of Adam and Eve, or of the discovery of the law of gravity by Isaac Newton. You might even think of a particular type of personal computer by means of association with the word 'apple'. Thus, visual information is automatically processed to lead to the percept of an object, which should be recognized with reference to the memory in our mind. Furthermore, knowledge or personal experience related to the object can be retrieved by association.

What, exactly, is happening in your brain while you are looking at an apple? The red light entering your eye is focused on the retina, where optical sensation is converted into neural signals. Retinal output is relayed to the lateral geniculate nucleus of the thalamus and transmitted to the primary visual cortex (V1) which is located at the posterior pole of the brain (Figure 2). In V1, individual neurons are activated only when a restricted part of the visual field (receptive field) is stimulated. Within their receptive fields, V1 neurons detect elementary visual features such as orientation of the outline of the apple, or the wavelengths of light reflected from the apple. Various visual features detected in V1 undergo hierarchical and parallel-distributed processing along two major cortical visual pathways. A dorsal visual pathway from the occipital to the parietal lobes is specialized for spatial and motion perception and for the visual control of actions, e.g. rotating and cutting the apple. A ventral visual pathway from the occipital to the inferior temporal (IT) cortex is specialized for identification of objects. Different features such as shape, color or texture of an apple are analyzed in discrete neural modules, and are reintegrated to give a unified percept of a whole apple. This integration occurs in the IT cortex, the final stage of the ventral visual pathway. The IT cortex plays a central role in visual identification as well as recognition of objects. Electrical recording of neuronal activity provides compelling evidence that visual long-term memory is coded as coordinated activities of neuronal populations in the IT cortex in primates.

However, many questions remain. How are different classes or depths of categories organized? Are neural activities better correlated with the physical properties of an object, or with subjective percept of the object? How is representation of a newly learned stimulus generated in the adult IT cortex? What is the mechanism underlying reactivation of neural representations in the IT cortex by

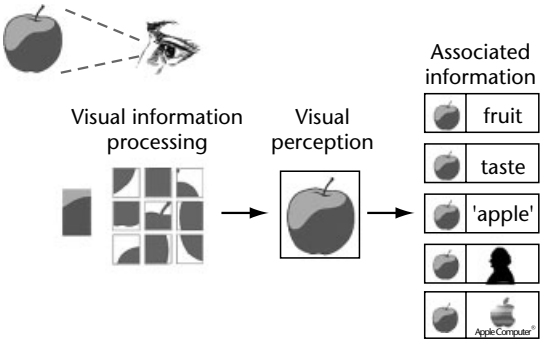


Figure 1. Object recognition.

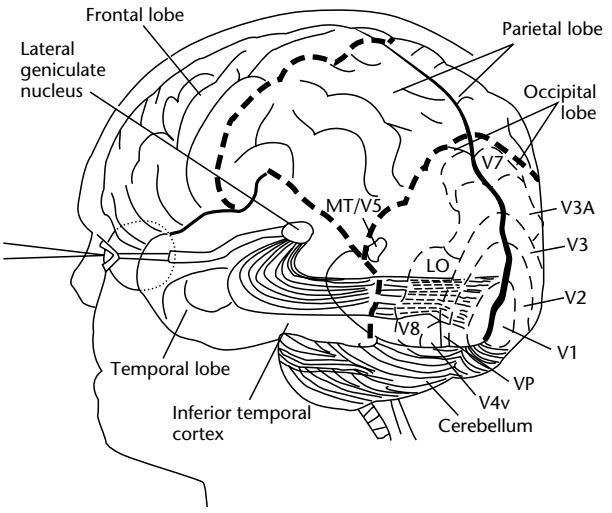


Figure 2. Human visual pathway. Starting with the eyes, the pathway extends to V1 through the lateral geniculate nucleus. The visual cortex is divided into visual areas V1, V2, etc. The inferior temporal cortex has a central role in object recognition. Modified from Logothetis NK (1999) Vision: a window on consciousness. *Scientific American* **281**(5): 69–75.

association? The following sections include recent evidence that partly answers these questions. It should be emphasized here that recognition is a process through which retinal input ultimately reactivates distributed neural networks subserving memory of objects in the temporal and other association cortices.

FAILURE OF RECOGNITION FOLLOWING BRAIN DAMAGE

Behavioral studies of humans with brain damage provide important insights into the organization of neural modules that participate in visual object recognition. Klüver–Bucy syndrome is a mosaic of behavioral abnormalities following extensive brain lesions to bilateral temporal lobes including medial temporal lobe structures. In addition to emotional, sexual, and personality changes, people with this syndrome exhibit a characteristic failure to recognize objects in front of them. Although they have normal visual acuity without any visual field defect, they behave as if they are 'blind' to the objects. Instead, they tend to put everything into their mouth in order to identify it by oral contact. Thus, visual recognition of objects can be selectively impaired without loss of elementary visuosensory functions, and without inability to identify the same object by other sensory modalities. This symptom is called visual agnosia. Agnosic patients can 'see' a visual image and sometimes even draw a copy of it, but cannot tell what it is or how to use it. A typical cause of visual object agnosia is ischemic infarcts in the IT cortex or adjacent occipital visual association cortex (Figure 2). This contrasts with the finding that more posterior-occipital lesions often affect visual sensation itself (cortical blindness). These clinical findings suggest that object recognition and basic visuosensory functions may be dissociable processes implemented in different anatomical structures.

Agnosia sometimes disproportionately involves certain object categories more than other categories. Various types of category-specific agnosia have been reported. Some patients who can recognize 'living things' are relatively impaired in the recognition of 'nonliving things' which are most typically exemplified by small, easily handled, manufactured objects. Other patients have the reverse problem: a patient who can quickly recognize a saw or a screwdriver will become distressed by attempts to recognize animals and foods, and may hesitatingly answer, 'Perhaps they are some kind of animal or plant.' Visual agnosia specific to letter strings is also reported. This syndrome is called alexia. Patients with alexia cannot read normally, despite normal visual capabilities, and despite their ability to understand spoken language and to write. Typically, they read words letter by letter, spelling each word before identifying it. Interestingly, some patients with alexia can easily read single and multiple digits without error. These findings indicate that distinct neural machinery exists for different categories.

A special type of agnosia is prosopagnosia, which renders people incapable of recognizing personal identity on the basis of face but spares their ability to identify someone from their voice or the visual characteristics of their gait. People with prosopagnosia have the ability to recognize common objects. Furthermore, they can assign faces to the 'face' category and correctly recognize various facial expressions on the faces they fail to identify individually.

Identification inability within a category such as seen in prosopagnosia does not seem to be restricted to faces. When people with prosopagnosia try to recognize nonface objects that are unique to themselves, for example their own pets, houses or cars, their failure is generally as marked as it is for faces. Moreover, there are case reports of a bird-watcher who had lost the ability to differentiate visually between birds, and of a farm worker who could no longer recognize his cows. These reports suggest that focal brain damage interferes with the

patient's ability to perform within-category identifications without affecting the recognition of the generic class to which the stimulus belongs. This suggests that a distinct system exists for the purpose of identification.

The variety of deficits in the visual processing of objects reviewed here suggests the existence of multiple systems for recognition. However, naturally occurring lesions tend to be several cubic centimeters in size and located according to the vagaries of vascularization rather than functional distinctions. Moreover, purely focal cognitive impairment is rare in clinical cases; patients often have other neurological deficits. For these reasons, the precise localization of the area responsible for specific perceptual processing is difficult to determine solely from currently available lesion data.

BRAIN MODULES FOR OBJECT RECOGNITION IN HUMANS

Imaging Studies in Humans

Functional activation studies in normal volunteers supplement the lesion data in identifying the neural correlates for the object recognition system at a finer spatial resolution. Functional magnetic resonance imaging (fMRI) and other neuroimaging methods enable comparison of regional brain activity while the person tested views various types of objects. Comparison of the activity evoked by faces with that evoked by nonface objects has identified a region in the fusiform gyri (part of the IT cortex) that responds preferentially to faces rather than objects, especially in the right hemisphere (Figure 3). This area is called the fusiform face area (FFA). Furthermore, regional brain activity in FFA does not increase when the person studied views photographs of 'scrambled' faces, houses and hands, suggesting that this is a face-specific brain area. In the field of orthography, one study comparing passive viewing of letters with passive viewing of digits found segregated regions responding to letters that were localized in the left fusiform gyrus. The parts of the visual association cortex implicated in these studies are consistent with the locations of lesions observed in prosopagnosia and alexia.

Some researchers claim the existence of a brain area that specifically processes geographic entities and locations, which is not suggested by the commonly observed dissociation in agnosic patients. A region in the parahippocampal cortex responds specifically to buildings or places in a local environment (Figure 3). This area is called the parahippocampal place area (PPA). There are thus anatomically separate modules selectively processing different categories, such as faces, letter strings, and geographic locations.

An enduring question concerning face perception is whether faces are the only objects processed by the FFA. The fact that some people with prosopagnosia cannot perform within-category identification in nonface categories indicates that the FFA is not restricted to processing faces. Gauthier and colleagues propose that this putative 'face area' might be the result of our extensive experience with faces. They found that expertise with homogeneous sets of stimuli newly created for the experiment (Figure 4) recruits face-selective areas in

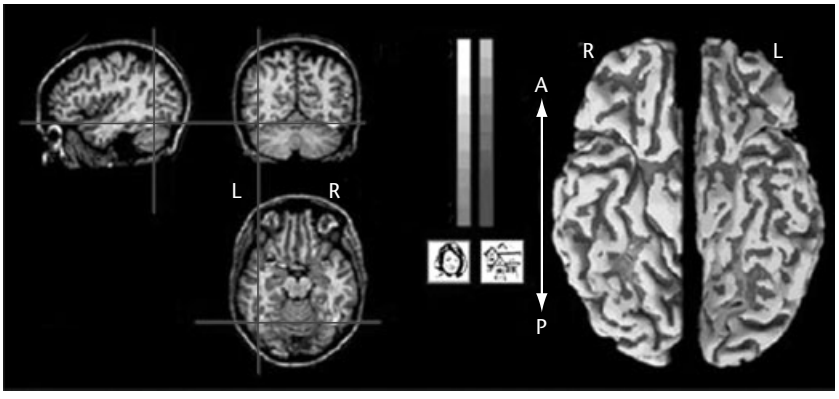


Figure 3. Example of face- and building-related regions in one observer. Preferential activation to faces versus buildings (orange) and to buildings versus faces (blue), shown on sagittal, coronal and axial slices (left), and on a three-dimensional reconstructed brain (right). The three-dimensional brain is shown in a ventral view. R, right; L, left; A, anterior; P, posterior. Modified from Levy I, Hasson U, Avidan G, Hendler T and Malach R (2001) Center-periphery organization of human objects areas. *Nature Neuroscience* **4**(5): 533–539.

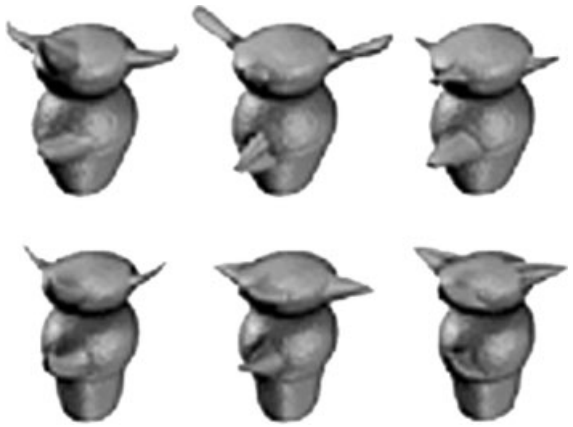


Figure 4. 'Greebles': a homogeneous set of nonface unfamiliar stimuli specially created for research into the 'face area' of the brain. Modified from Gauthier I, Tarr MJ, Anderson AW, Skudlarski P and Gore JC (1999) Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience* **3**(8): 764–769.

the fusiform area (FFA) and occipital lobe (OFA). They extended this finding to other homogeneous categories. Study participants who were experts on birds or cars were subjected to fMRI in two task conditions: (1) identification of faces, cars and birds in homogeneous categories, and (2) coarse categorization of familiar objects. Homogeneous categories activated the FFA to a greater extent than did familiar objects. Moreover, in the face-selective areas in the right hemisphere, significantly greater activation was observed in experts than in novices. This suggests that the level of categorization and expertise, rather than geometrical properties of objects, determines the specialization of the FFA.

In summary, neuroimaging methods have clarified that anatomically distinct regions may serve for the recognition of different categories, such as face, letter strings, and places in the local environment. The specificity for face in FFA may be the result of our extensive experience with faces.

OBJECT RECOGNITION IN MONKEY IT CORTEX

Evidence from Lesion Studies

Macaque monkey preparations provide efficient animal models to investigate detailed anatomical and functional architectures responsible for visual recognition of objects. The Klüver–Bucy syndrome was originally described in the monkey. As in humans, monkeys with bilateral temporal lobe resection appear indifferent to objects that are presented visually, even if there is no sign of visual acuity impairment, visual field defects or deficits in modalities other than vision. The monkeys have the compulsion to examine all objects by sniffing and by oral contact. The visual aspects of this syndrome proved to be caused by small lesions limited to the temporal neocortex, but not by resection of the amygdala, hippocampus or other medial temporal lobe structures. These findings indicate that in both humans and macaques, the IT cortex serves as the center for visual recognition of objects.

By forming focal lesions systematically in monkeys, it is possible to further characterize the functional organization of the IT cortex. Lesions in the posterior part of the IT cortex cause severe impairments in visual pattern discrimination. The deficits are more marked when the object to be discriminated is presented in a different size, position or angle. On the other hand, lesions in the anterior part of the IT cortex produce milder visual discrimination deficits, and the performance may be relatively invariant to physical transformations of the visual pattern. Anterior IT lesions are also reported to impair the ability to retain an object's identity for up to several minutes, or the ability to associate objects with other objects. Thus, it is likely that the posterior part of the IT cortex is involved in more sensory aspects of visual perception, while the anterior part is involved in more mnemonic and associative aspects of object recognition.

Neuronal Selectivity for Stimulus Features

To determine how objects are represented in the neural network of the IT cortex, we can examine the stimulus-coding properties of single neurons. The activity of single neurons can be monitored by recording their electrical activity using an extracellular microelectrode.

The majority of IT cells respond to visual stimuli. Many IT neurons are selective for various stimulus attributes, such as color, texture or shape. Of particular interest is the sensitivity of IT neurons to stimulus shape. Although shape selectivity in earlier visual areas such as V4 has also been reported, only in IT is this selectivity notably observed. Neurons in this area respond selectively to various natural or synthetic objects, or to mathematically constructed two-dimensional patterns, such as Fourier descriptors. Groups of cells in IT also respond to the sight of biologically important objects such as faces or hands. A population of

neurons that selectively respond to disparity-defined three-dimensional shapes has also been reported.

The stimulus features necessary for maximal activation of the cell can be determined by reducing the complexity of an effective visual stimulus. Maximal activation of the neurons in IT cortex is usually achieved by moderately complex stimuli, such as star or T shapes. The properties of IT cells change significantly along the anteroposterior axis of the IT cortex. In the most posterior part of the IT cortex, neurons prefer simple stimuli and have their receptive fields constrained to the contralateral side, like those in area V4. In the most anterior part of IT, neurons rarely respond to such simple stimuli and have larger receptive fields extending to the ipsilateral side. In addition, a group of IT neurons preserve their selectivity for particular shapes even if the size, position and orientation of the shapes are changed.

In summary, IT neurons possess suitable properties for coding visual objects in a efficient way, such as a large receptive field, complex shape selectivity, and invariance to size, position or orientation. This tendency becomes more pronounced in neurons towards the anterior part of the IT cortex.

Functional Architecture of IT Cortex

Columnar organization is a well-established property of many different cortical areas. In the early visual system, the clustering of neurons responding selectively to simple stimulus attributes, such as position in the visual field, ocular preference, orientation of line segments or direction of movement, is observed. In the IT cortex, modular organization is less related to retinotopic organization and reflects instead similar preferences for combinations of shapes and other stimulus attributes. A pair of neurons (within 100 μm) that were recorded simultaneously using a single electrode had more similar feature selectivity than neurons recorded using different electrodes (more than 1 mm apart). Fujita and colleagues found neurons with similar stimulus selectivity when electrodes were advanced nearly at a right angle to the cortical surface over a distance of 0.6–1.4 mm, whereas when electrodes were advanced obliquely they did so only over a distance of 0.2–0.5 mm (Figure 5). In the latter type of penetrations, there were two or three separate clusters of neurons that showed similar selectivities. All of these results support the conclusion that the anterior IT cortex consists of modules in which neurons with similar selectivities cluster across cortical layers. Optical recording studies have further confirmed the clustering of cells that respond to a moderately complex feature, and demonstrated that similar but different stimuli activated partially overlapped spots. The averaged diameter of individual spots roughly coincides with the width of columns inferred from the single unit recording experiments.

In summary, neurons with similar selectivities cluster in a columnar pattern in the IT cortex. Various calculations, such as those of similarity for categorization and difference for identification between objects, may be performed in these columnar modules.

Generating the Mnemonic Representation

The stimulus selectivity of IT neurons can be acquired through learning in adulthood. Sakai and Miyashita have demonstrated the effects of associative learning in adults on the stimulus selectivity of IT cells. In a visual stimulus–stimulus association task, monkeys were required to retrieve a visual stimulus specified by another visual stimulus from long-term memory. Recordings made during this task identified a class of neurons exhibiting significantly correlated visual responses to arbitrarily assigned picture pairs (Figure 6). Thus, the stimulus selectivity of IT neurons can be acquired through learning in adulthood. Moreover, the activation of IT neurons can link the representations of temporally associated but geometrically dissimilar stimuli.

The plasticity of stimulus selectivity was recently supported by other studies. In adult monkeys trained to discriminate between complex visual stimuli, the proportion of IT cells maximally responsive to some members of the stimuli used for the training was significantly greater than that observed in the control untrained monkeys. Some researchers found a compact cluster of cells that responded selectively to the visual stimuli used in the task that the monkeys had learned: this suggests that specific visual experience induces development of clusters of anterior IT neurons with similar stimulus preferences. The implication of dynamic reorganization of columns after learning are consistent with Gauthier's proposal that the putative 'face area' in the human brain may be the result of our extensive experience with faces.

Supporting evidence for the associative effect on IT neurons has been accumulating. Guided by a theoretical conjecture that the modification of response properties for temporally associated stimuli might also be a mechanism for learning to associate different three-dimensional views of the same

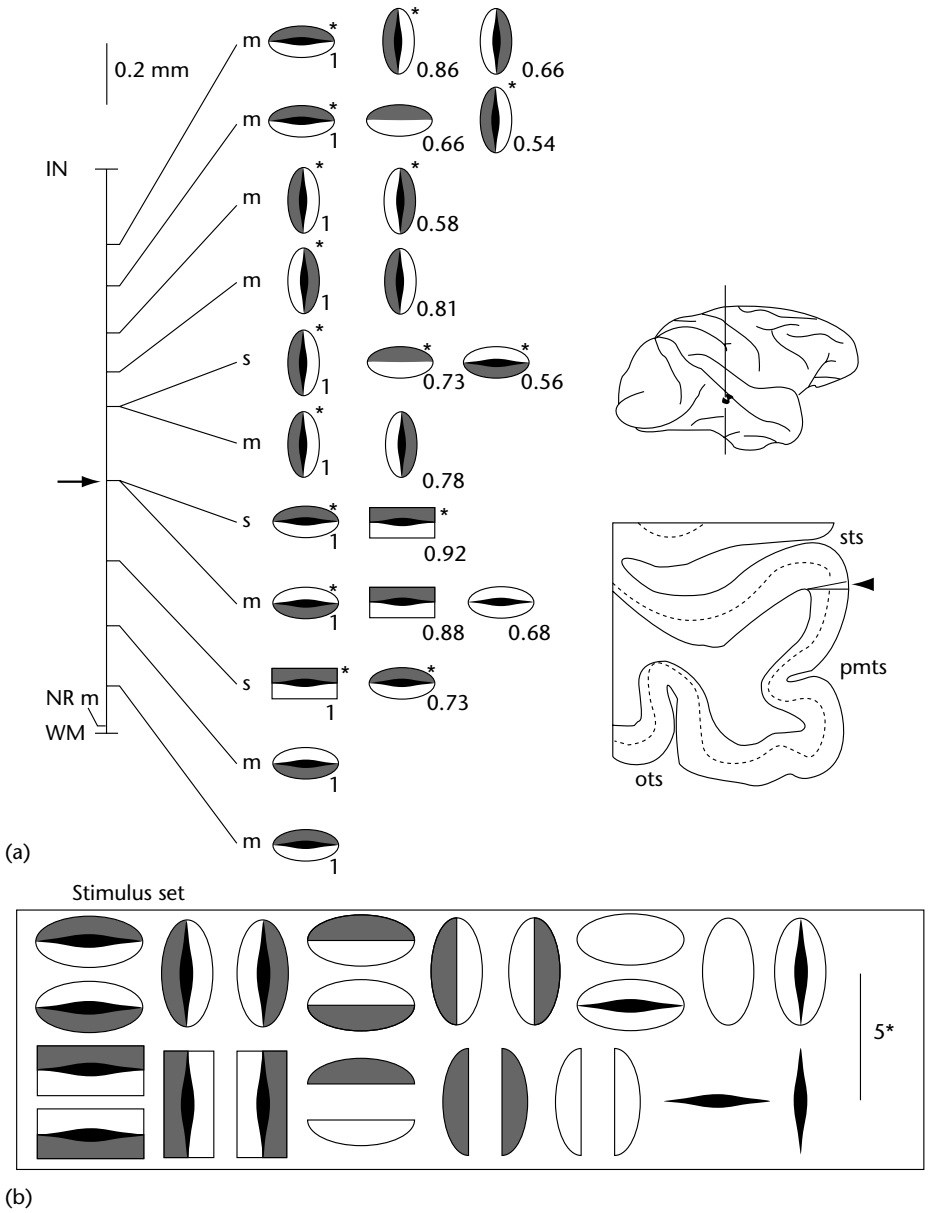


Figure 5. Response strength profiles of inferior temporal (IT) cells or multiple units recorded in a penetration directed vertically to the cortical surface. (a) The penetration is reconstructed on a coronal section through the IT (arrowhead in right panel); pmts, posterior middle temporal sulcus; sts, superior temporal sulcus; ots occipitotemporal sulcus. A schematic drawing of the recording track (left) shows recording depths; IN, site where the first neuron was recorded in this penetration; WM, entry into white matter; NR, no response. The effective stimuli for the cell, seventh from the top (arrow), was determined first; then other cells distributed throughout the gray matter along this penetration were tested; s, single cell recording; m, multiple cell recording. Numbers show response magnitude relative to the maximal response. (b) Set of 24 stimuli used for the experiment, which included effective stimuli for the cell and their modifications. Modified from Fujita I, Tanaka K, Ito M and Cheng K (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* **360**: 343–346.

object, some researchers found that IT neurons indeed responded more consistently to different three-dimensional views of the same object than would be expected by chance. Furthermore, when visual stimuli are shown in a fixed temporal order repeatedly, the few images that evoke activity in a given neuron are often neighboring stimuli in the sequence, suggesting this activity could work as a vehicle for generating long-term visual associations.

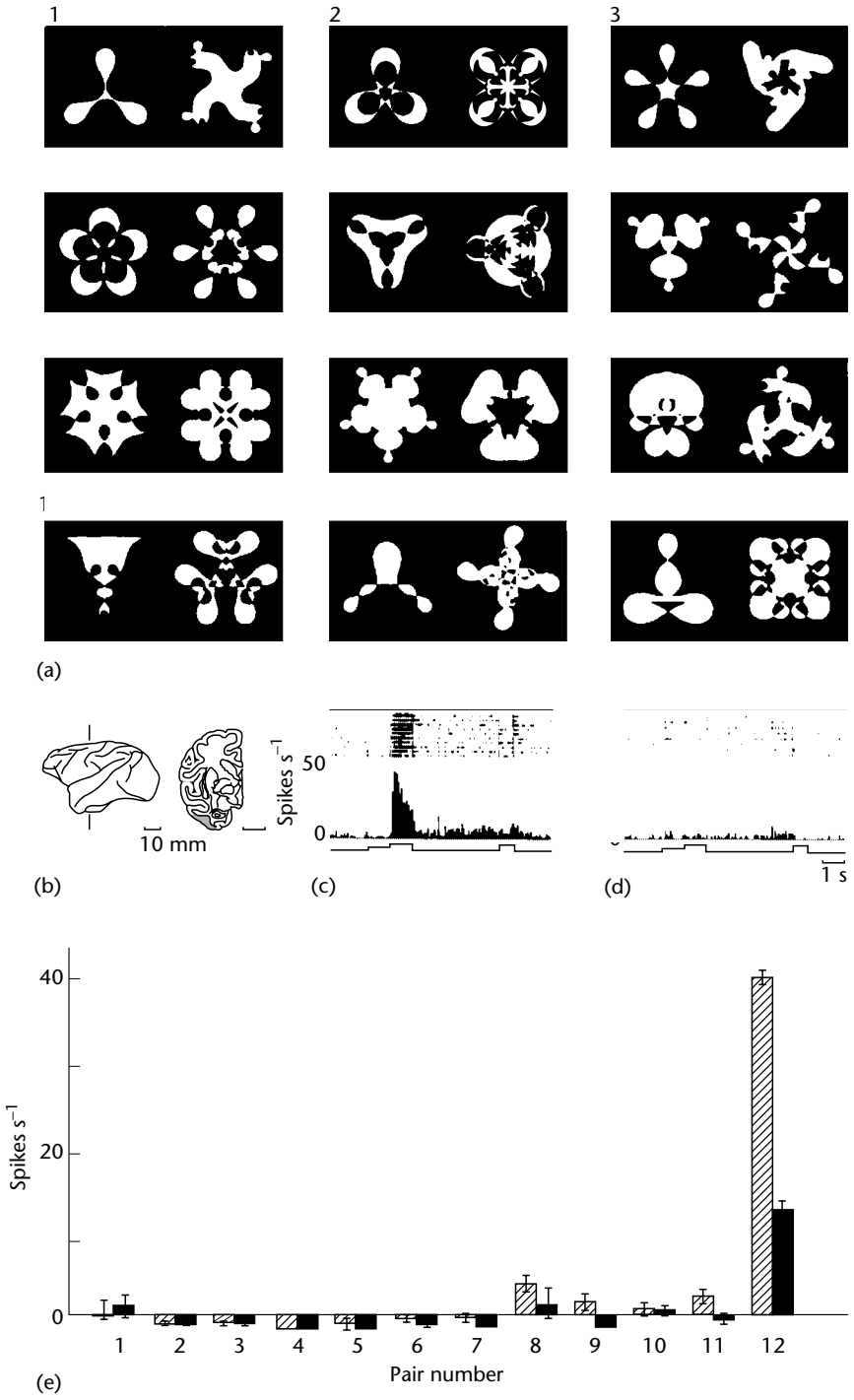


Figure 6. Stimulus sets used for the pair association task and responses of a pair-coding neuron. (a) Twelve pairs of Fourier descriptors used as stimuli in the task. (b) Location of recorded neurons. Left, lateral view of a monkey brain. Right, cross-section indicated by a vertical line on the lateral view. The stippled area represents the range of recording sites. (c) Rastergrams of neural discharges in each trial (upper panel) and spike density histograms (lower panel) obtained from a single neuron. Bin width 80 ms. Trials were collected for cue 12. (d) Trials for cue 7, which elicited no response at all in the same cell as in (c). (e) Mean discharge rates for each cue presentation relative to the spontaneous discharge rate (denoted by arrowhead) in the same cell as in (c) and (d). Modified from Sakai K and Miyashita Y (1991) Neural organization for the long-term memory of paired associates. *Nature* **354**: 152–155.

It has thus been clarified that IT neurons have the ability to establish new linkages between different stimuli that have arbitrary but cognitively meaningful connections through dynamic reorganization of the columnar architecture.

Active Representation and Subjective Perception

A striking feature of neural representation in the monkey IT cortex, compared with other earlier visual areas, is that single-unit activities are more contingent upon subjective perception. When different images are presented to the left and right eyes, the individual does not perceive a combination of the two images, but rather perceives each image alternately. By presenting a stimulus that evoked strong responses from a recorded cell to one eye, and a control stimulus that did not activate the recorded neuron to the other eye, Logothetis and colleagues compared the responses of IT cells with the monkey's reported perception. Only a small fraction of the cells in V1 responded consistently with the monkey's report. This proportion increased moderately in V4, and nearly all the IT cells responded consistently with the monkey's perception. These results indicate that the conscious perception of objects is better correlated with cell activities in IT than in earlier visual cortices.

In human studies using functional brain imaging, cortical regions whose activity is related to rivalrous perceptions have been systematically investigated. The extrastriate areas in the fusiform gyri, but not in the early visual areas, have been found to exhibit significant enhancement of neural activity during perceptual reversals. When images of a house and a face are presented to different eyes during binocular rivalry, changes in stimulus-selective magnetic resonance signals in the FFA and PPA are differentially accompanied by changes in the person's perceptions. Thus, activity in these object-representation areas seems to reflect the perceived stimulus rather than the actual retinal stimulus.

CONCLUSION

Object recognition is a process wherein retinal input ultimately reactivates distributed neural networks subserving memory of objects in IT cortex. Neural representations in the IT cortex are organized in a columnar fashion and dynamically change their organization even after birth. This dynamic reorganization may be the neural mechanism of memory encoding of objects. Furthermore, the activity of neurons in IT cortex is closely associated with subjective perception. This indicates that conscious percept is mediated by activation of neural representations in the IT cortex.

Although the IT cortex serves as the center for visual recognition of objects in both humans and macaques, the detailed correspondence of anatomical structures between human and monkey remains to be established. One promising approach to clarify this correspondence is the use of fMRI of the monkey brain, enabling us to bridge the gap between data from human fMRI studies and those from monkey single-unit analysis. These comparisons should provide a powerful means of discovering how neural codes are linked to conscious experiences.

Further Reading

Damasio AR (1990) Category-related recognition defects as a clue to the neural substrate of knowledge. *Trends in Neurosciences* **13**(3): 95–98.

Farah MJ (1995) *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. Cambridge, MA: MIT Press.

Gross CG (1972) Visual functions of inferotemporal cortex. In: Jung R (ed.) *Handbook of Sensory Physiology*, vol. 8/3B, pp. 451–482. Berlin: Springer.

Logothetis NK (1998) Single units and conscious vision. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **353**: 1801–1818.

Mishkin M and Appenzeller T (1987) The anatomy of memory. *Scientific American* **256**(6): 80–89.

Miyashita Y and Hayashi T (2000) Neural representation of visual objects: encoding and top-down activation. *Current Opinion in Neurobiology* **10**(2): 187–194.

Tanaka K (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* **19**: 109–139.

Treisman AM and Kanwisher NG (1998) Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology* **8**(2): 218–226.

Ungerleider LG and Haxby JV (1994) 'What' and 'where' in the human brain. *Current Opinion in Neurobiology* **4**(2): 157–165.

Zeki S (1993) *A Vision of the Brain*. Oxford: Blackwell.

Occipital Cortex

Introductory article

Peter De Weerd, University of Arizona, Tucson, Arizona, USA

CONTENTS

The occipital cortex is the thin, 6-layered sheet of gray matter that envelopes the white matter of the most posterior lobe of the brain, and that contains several visual areas involved in the analysis of the retinal image.

INTRODUCTION

The occipital cortex is the gray matter (consisting of neuronal cell bodies) of the occipital lobe, which is the most posterior lobe of the brain. Occipital cortex is strongly involved in visual processing. It contains the primary visual area (V1), which receives retinal information via the lateral geniculate nucleus (LGN) in the thalamus. Depending on the species, it may contain additional visual cortical areas, including the secondary visual area (V2). This article will focus on the anatomy and visual functions of V1 and V2.

In the human brain, V1 is mostly buried in the calcarine sulcus, on the medial side of the brain (Figure 1a and b), whereas in many other species, V1 is largely exposed on the surface of the occipital lobe. The primary visual cortex is often referred to as 'area 17'. This terminology can be traced back to the German anatomist Korbinian Brodmann, who at the beginning of the twentieth century performed detailed anatomical studies to subdivide the human cerebral cortex. Brodmann numbered cortical areas in the order that he studied them, and the 17th area that he delineated is what we now call V1. This region is distinguishable from the immediately surrounding cortex by a prominent stripe of white matter (consisting of myelinated axons) in layer IV, which is called the stria of Gennari. This explains the origin of the name 'striate cortex' – yet another name referring to V1. The area surrounding V1 is V2, or Brodmann's area 18, which performs further cortical analysis of the visual input. In humans, the occipital lobe contains visual territory in addition to areas V1 and V2, which is beyond the scope of this article.

FUNCTIONAL ORGANIZATION OF AREAS V1 AND V2

Properties of Single Neurons

Until the 1950s, light spots had been the most conventional tool for exploring the function of visual neurons. The use of light spots during recordings from ganglion cells in the retina and from LGN cells had revealed a center-surround concentric organization of receptive fields (RFs). However, light spots proved to be an ineffective stimulus for most cortical neurons. The function of the visual cortex remained a mystery until the early 1960s, when Hubel and Wiesel discovered that a straight edge or line was a much more effective stimulus for activating V1 cells. They recorded from single V1 neurons in the anesthetized cat, and stimulated the RFs by back-projecting stimuli on a translucent screen placed in front of the cat's eyes. They made their discovery when they observed that the edge which was projected on the screen while inserting a new stimulus slide in the slide projector was more effective in activating V1 cells than the various dots and disks they had used so far – one more example of the fact that great discoveries are often accidental!

This discovery marked the beginning of an explosive growth in our understanding of the functioning of the visual cortex. Hubel and Wiesel discovered that V1 neurons (except for input-layer IV-C, where RFs have a concentric center-surround organization) were 'tuned' to orientation. The neurons responded optimally to an edge or line of a particular orientation placed in their RF (the 'preferred' orientation), whereas their responses diminished when the stimulus was oriented away from the preferred orientation, until an activity rate was reached that was similar to the rate obtained without a stimulus (Figure 2a). Two main classes of neurons were described, namely simple cells and complex cells (Figure 2b). These

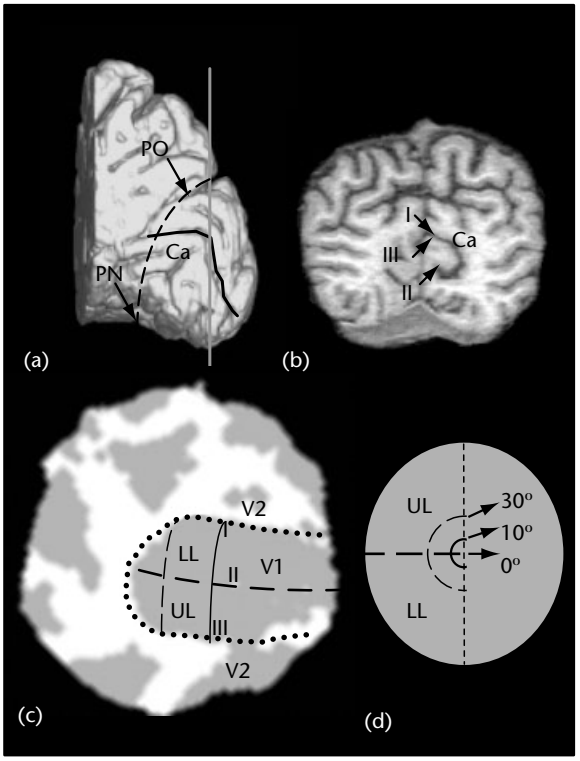


Figure 1. Visual areas in the occipital lobe. (a) Extent of the occipital lobe on a medial view of the right hemisphere. The occipital lobe is located posterior to a line (broken line) connecting the parieto-occipital sulcus (PO) and the pre-occipital notch (PN). The calcarine sulcus (Ca) is indicated by the bold black line. (b) Coronal section through the occipital lobe, at the level of the gray bar in Figure 1a. The right half of the section corresponds to the right hemisphere, and the left half corresponds to the left hemisphere. The view of the calcarine sulcus in Figure 1a corresponds to a view from the left in Figure 1b, after removing the left hemisphere. (c) Flat map of V1 and surrounding cortex (oriented such that posterior aspect is on the right, and dorsal aspect is on top). A flat map is based on an unfolding of cortex, such that cortex inside the sulci becomes exposed. During this process, the upper lip (arrow I in figure 1b) and lower lip (arrow III in Figure 1b) of the calcarine sulcus are pulled apart, such that all of the cortex inside the sulcus becomes exposed, including the fundus of the sulcus (II). In the resulting flat map, cortex that comes from inside a sulcus is shown in gray, and cortex on the surface is shown in white. (d) Representation of the left hemifield and its projection onto flattened cortex in V1 (Figure 1c). The vertical meridian and its representation on the flat map are shown by dotted lines. The horizontal meridian is denoted by a coarse dashed line. The upper left quadrant (UL) of the visual field is projected onto the lower bank of the calcarine sulcus, and the lower left quadrant (LL) is projected onto the upper bank. The upper lip of the calcarine sulcus (I) corresponds to the lower part of the vertical meridian, and the lower lip of the calcarine sulcus (III) corresponds to the upper part of the vertical meridian. The fundus of the calcarine sulcus (II) corresponds to the horizontal meridian. Iso-eccentricity lines in the visual field (fine solid and broken lines in Figure 1d) are projected onto straight lines going across the calcarine sulcus (positions shown in Figure 1c are approximate). A disproportionate amount of cortex is devoted to the central 10° of the visual field (cortical magnification). The vertical meridian forms the border between areas V1 and V2. The projection of the left hemifield onto V2 is mirror-reversed compared with the projection of the left hemifield onto V1 (not shown). The projection of the right hemifield on to visual areas in the left hemisphere is analogous to the projections described here for the left hemifield on to the right hemisphere. The posterior part of the right hemisphere from a human volunteer shown in this figure is a reconstruction based on high-resolution structural MRI scans (3-dimensional Spoiled Gradient Echo Sequence, $512 \times 384 \times 128$ matrix, Echo Time = 5 ms, Reception Time = 24 ms, flip angle 45°) obtained on a 1.5-tesla GE Signa Horizon EchoSpeed system at the Laboratory for Brain and Cognition at the National Institutes of Health. Flattening was done with software from Brian Wandell, enhanced with rendering software by Peter Jezzard.

neurons differ in the specificity of the stimulus that is required to trigger action potentials. To activate simple cells, the edge or line must have a specific contrast polarity in a given RF position. This is due to the fact that simple RFs consist of multiple elongated and juxtaposed subregions with opposite preferences for lightness and darkness (similar to the antagonistic subregions in the RF of LGN neurons). For complex cells, the position and contrast polarity of the stimulus in the RF is irrelevant. Simple cells derive their properties at least in part from excitatory input from LGN RFs aligned along the preferred orientation of the simple cell. Complex cells in turn derive their properties in part from converging input from simple cells with overlapping RFs (Figure 2b). In addition, local inhibitory circuits contribute to the properties of simple and complex cells. Orientation-selective neurons are also present in many extrastriate areas.

Functional Organization of V1 and V2

Columns, retinotopy and cortical magnification

By comparing perpendicular and tangential electrode penetrations in V1, it became clear that neurons with the same orientation preference are grouped together in *orientation columns*, extending

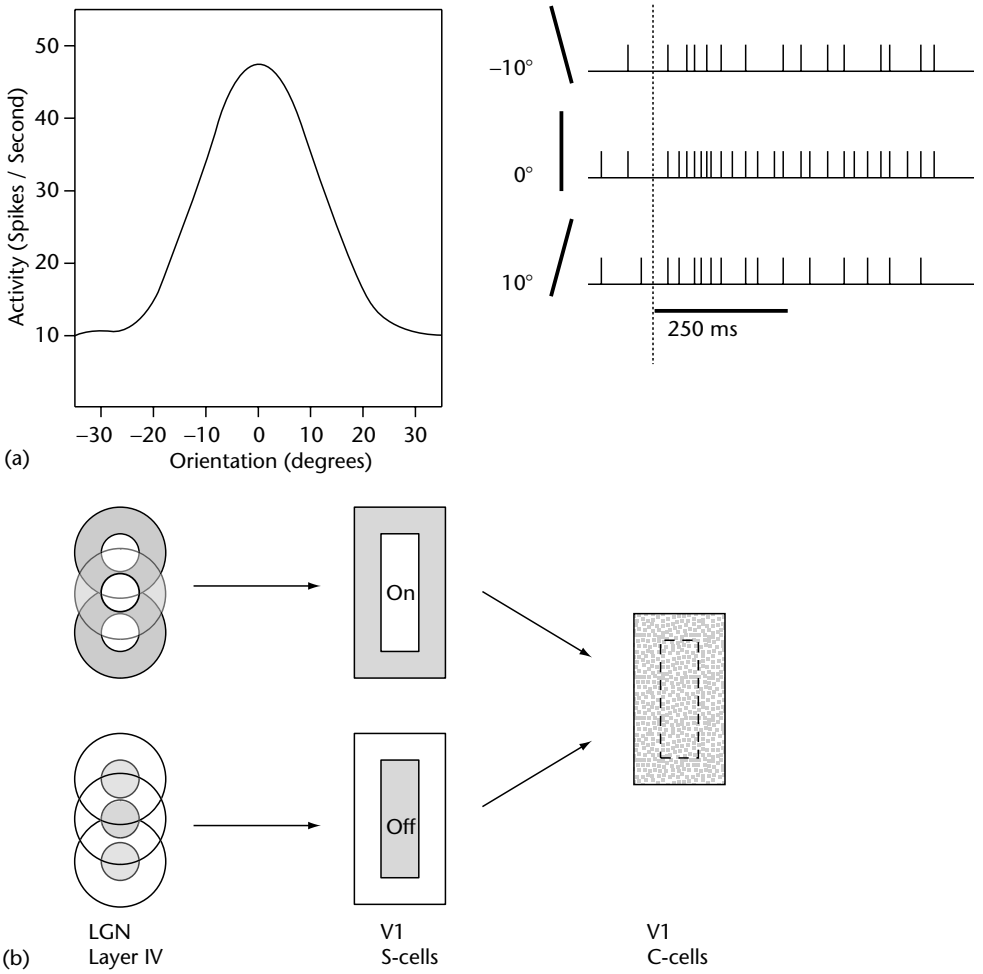


Figure 2. Orientation tuning of visual neurons in the occipital lobe. (a) Idealized graph in which activity of a neuron is plotted against the orientation of a bar stimulus placed in its receptive field. The activity is expressed as spikes (action potentials) per second. On the right-hand side, a single-trial histogram of activity elicited by three different stimuli is shown. Spikes (vertical marks) are shown as a function of time. The vertical stippled line indicates when the stimulus is presented in the receptive field, and the duration of the stimulus (250 ms) is denoted by the solid horizontal bar at the bottom. For each stimulus, the frequency of firing increases after stimulus onset, and then returns to a level similar to that when no stimulus was present (left of the dotted line). While the stimulus is present, a higher firing rate can be observed for the vertical bar (denoted as 0°) compared with bars tilted away from vertical. This indicates that in this example the neuron is 'tuned' to vertical. Orientation tuning curves, such as that shown on the left-hand side, are based on a large number of trials in which the response of the neuron to many different orientations is measured, and expressed in terms of an average number of spikes per second. For the present example, vertical would be referred to as the 'preferred orientation'. (b) Hypothetical scheme proposed by Hubel and Wiesel to explain orientation tuning in V1. White regions in the receptive fields of LGN cells and of V1 simple cells (S-cells) correspond to parts of the receptive field that are activated by bright stimuli on a dark background, and gray regions correspond to regions of the receptive field that are activated by dark stimuli on a bright background. The hatching of the receptive field of the V1 complex cell (C-cell) indicates that the contrast polarity of the stimulus is irrelevant. Orientation tuning properties of S-cells are explained by hierarchical inputs from LGN cells, while the properties of C-cells are explained by hierarchical inputs from S-cells. (For further explanation, see text.)

from the superficial layers (II and III) to the deep layers of cortex (V and VI), interrupted by layer IV. In addition, neurons receive inputs predominantly from one eye or the other, forming ocular dominance columns. Using optical imaging, Blasdel showed that orientation groupings within ocular dominance columns are organized as sectors around a central point (Figure 3). Two ocular dominance columns taken together form a *hyper-column*, which processes all orientation inputs from both

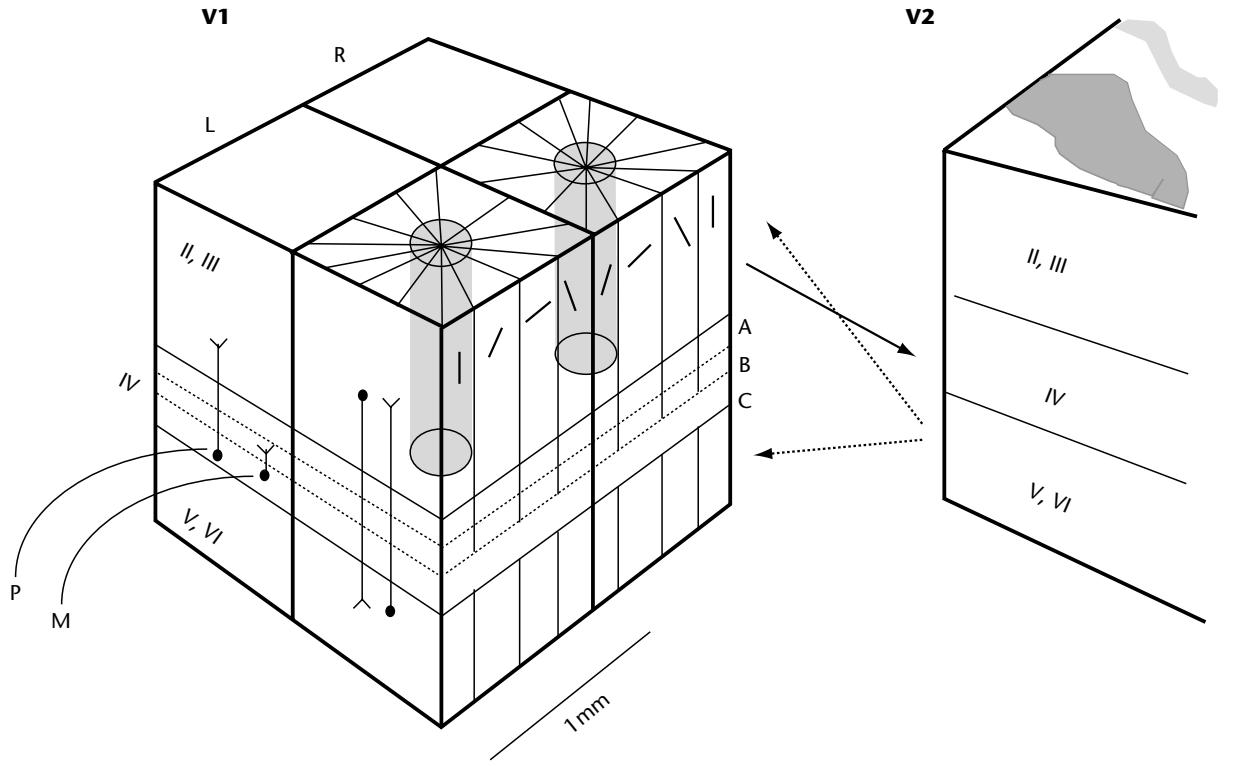


Figure 3. Schematic illustration of elementary processing units in areas V1 and V2, and of their inputs and outputs. The cubes of cortex shown illustrate the anatomical divisions and functional specialization of cortex in V1 and V2. Cortex in V1 and V2 consists of six layers, which are labeled from I to VI. Layers I to III are referred to as superficial layers, and layers V and VI are referred to as deep layers. Layer I is not shown because it contains few or no neurons. In the diagram for V1, further subdivisions of layer IV are shown. Layer IV is subdivided into sub-layers (A–C), and layer IV-C is subdivided into layer IV-Ca (more superficial) and layer IV-Cb (deeper). The latter subdivision is not shown. Layer IV-Cb receives parvocellular (P) input from the LGN, and layer IV-Ca receives magnocellular (M) input. Four symbolic neurons (the dot corresponds to the cell body, and the linear extension corresponds to the axon) show important connections between different cortical layers in area V1. From layer IV, information tends to be sent to superficial layers first, after which interactions between superficial and deep layers occur through vertical connections. Similar communication between layers exists in V2 (not shown). Functional subdivisions shown in V1 include orientation columns (labeled by a bar representing preferred orientation) and ocular dominance columns (labeled by the preferred eye; left 'L' or right 'R'). Other functional subdivisions shown are blobs (gray cylinders) and inter-blobs (the rest of the cortex) in V1, and thick stripes (dark gray), thin stripes (pale gray) and inter-stripes (white) in V2. Arrows show feedforward (solid arrow) and feedback (stippled arrows) connections between areas V1 and V2. (For further explanation, see text.)

eyes. Thus a hyper-column is an elementary processing module with all of the equipment necessary to process elementary information in a small part of the image. V1 uses a juxtaposition of such hyper-columns to process local information from the entire image.

Neighboring hyper-columns in V1 process neighboring points from the retinal image, and this principle is referred to as *retinotopy*. Studies of retinotopy in V1 show that many more neurons are devoted to central vision than to peripheral vision. This reflects the fact that in the fovea of the retina there is a one-to-one relationship between receptors and ganglion cells, and this relationship is preserved in the projections from the fovea to the LGN and V1. The more peripheral vision is, the more convergence there is in the retina from receptors onto ganglion cells, and the sparser are the projections from retina to cortex. The resulting disproportionate representation of central vision in V1 is referred to as *cortical magnification* (Figure 1c and d). The extra processing power allocated to central vision maximizes acuity, and is important for the analysis and recognition of objects.

Area V2 receives projections from V1 that are retinotopically organized, and this area also shows significant, albeit less pronounced cortical magnification. In contrast to V1, area V2 is not organized

into a clear system of hyper-columns, which suggests that it may code image features which are less elementary than the orientation of boundary segments (see below).

Two visual streams in areas V1 and V2

Areas V1 and V2 maintain a relative separation of signals generated by the two main classes of retinal ganglion cells (β and α cells). β cells (with small RFs) are specialized for analyzing object features at high spatial resolution, while α cells (with larger RFs) are specialized for detecting motion. Ultimately, those two different retinal signals form the basis for a subdivision into two cortical systems. One of these systems extends into the temporal lobe and is specialized for object processing, while the other extends into the parietal lobe and is specialized for the processing of motion and the spatial relationships between objects. In non-human primates the occipital lobe is entirely occupied by areas V1 and V2, whereas in humans several additional areas belonging to the two visual processing streams are located within the occipital lobe. In the following paragraphs, the origins of the two visual processing streams in areas V1 and V2 will be discussed (Figure 3).

β cells project to parvocellular (P) cells in the LGN, which in turn project predominantly to layer IV-Cb of area V1, from where inter-neurons send axons to neurons outside layer IV. V1 cortical neurons in layers II and III can be subdivided into two main compartments, namely 'blobs' and 'inter-blobs'. Blobs tend to be more specialized for the coding of wavelength, while the inter-blobs tend to code orientation and disparity (depth) in addition to wavelength. Blobs occupy the center of hyper-columns, whereas inter-blobs occupy the remaining space. Neurons in blobs and inter-blobs strongly project to separate compartments in V2 (thin stripes and inter-stripes, respectively), from where a cascade of anatomical projections starts, directed towards the temporal lobe. This anatomical pathway is specialized for the recognition of objects.

α cells in the retina project to magnocellular (M) cells in the LGN, which in turn project to inter-neurons in layer IV-Ca, which make synapses onto neurons in layer IV-B. Neurons in layer IV-B project to a separate sub-compartment of V2 (thick stripes), and to area MT, from where a cascade of anatomical projections starts, directed towards the parietal lobe. This pathway is specialized for the processing of object- and self-motion, and for spatial aspects of perception.

COMPLEX PERCEPTUAL PROCESSING IN V1 AND V2

The classic view that V1 neurons perform a local analysis of edges and boundaries inside their RF was quickly challenged by unexpected and somewhat paradoxical findings. Indeed, studies conducted during the 1970s showed that responses to stimuli presented inside the RF of V1 cells could be modulated by stimuli *far outside* the classical RF (context sensitivity), while the presentation of those modulatory stimuli by themselves did not result in any effect. Anatomical findings showing lateral (horizontal) connections in V1 running over distances of up to 5 mm provided an anatomical basis for these effects. The context sensitivity of V1 neurons may explain a variety of perceptual phenomena.

For example, surrounding a line segment inside an RF by identical line segments just outside the RF (in the surround) causes a suppression of the response which is smaller than the suppression caused by orthogonal line segments in the surround. This finding provides a neural basis for the perceptual effect of 'pop-out' in texture displays, in which a line segment that is different in orientation from the surrounding elements captures attention, and is detected effortlessly (Figure 4a). Furthermore, the response of V1 neurons to a visual texture covering their RFs is modulated by whether that texture is perceived as a figure or a background, suggesting that V1 neurons contribute to figure–ground segregation. These and many other findings have caused a dramatic shift in views on the function of V1, from an area that contributes to local analysis of edges in the image to an area that contributes to complicated contextual perception, figure–ground segmentation, and even cognitive factors such as attention.

If V1 is doing so much, then what is the function of the other areas? This question is probably put too sharply. Many of the more complex properties of V1 may depend at least in part on feedback from extrastriate cortex, and not just on horizontal connections within V1. Indeed, forward projections from the superficial layers of V1 (mainly layer III) to layer IV of V2 and other extrastriate areas are returned by feedback projections which arrive predominantly in layers II, III and VI (see Figure 3). Thus a fast feedback signal (top-down influence) may influence the information that is sent forward from V1 to extrastriate cortex. Perception may thus depend on several feedforward–feedback loops, and the complex properties of V1 neurons may partly reflect such iterative processing.

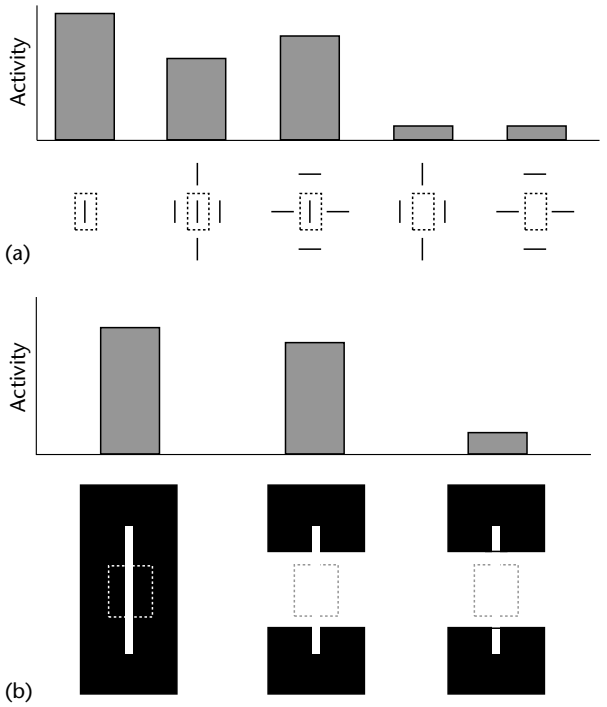


Figure 4. Perceptual effects that have been attributed to functional properties of visual neurons in the occipital lobe. (a) Response of V1 neurons to a single bar in the preferred orientation is suppressed more by lines of the same orientation outside the classical receptive field (dotted outline) than by lines orthogonal to the preferred orientation. This constitutes a possible basis of the perceptual effect of 'pop-out' (see text). Stimuli that are presented outside the receptive field, without any stimulus inside the latter, only have minimal effects on the firing rate of the cell. The drawing summarizes work by Knierim and Van Essen. (b) The activity of many V2 neurons is influenced not only by optimally oriented 'real' bars (left) placed inside the receptive field (dotted outline), but also by 'illusory' bars, whose presence is suggested by the alignment of white cut-outs in the black inducing stimuli. The white bar which seems to be occluding the two black rectangles generates a response in V2, even though no physical stimulation corresponding to the bar is present in the receptive field (middle). When the cut-outs are closed, which is incompatible with the perception of a white bar in the foreground overlying black surfaces in the background, then the response of the V2 neuron to the illusory bar is eliminated. The drawing summarizes work by von der Heydt and Peterhans. (For further explanation, see text.)

Consistent with this idea, neurons in V2 and extrastriate cortex show complex properties that are not present, or that are less pronounced, in V1. The increased complexity of visual processing as one moves from V1 to V2, and to higher-order extrastriate cortex, is referred to as the hierarchical processing of visual information. For example, V2 neurons show responses to illusory contours, which are boundaries perceived in the absence of physical discontinuities in the image. In V1, those responses are far less prevalent. Illusory contours are perceived in two-dimensional displays in which the arrangement of local stimuli suggests a three-dimensional arrangement in which one surface is occluding other surfaces (Figure 4b, middle). The responses of V2 neurons to illusory contours (Figure 4b) therefore suggest that they are involved in completion processes which take place during the processing of objects in the foreground which occlude other objects in the background. Area V2 contributes to a variety of other functions that help to segregate objects from their background. In line with the hierarchical processing of visual information, lesions in V2 cause deficits in grouping and texture segregation, but leave the perception of isolated luminance-defined lines and edges intact. Findings from imaging studies (e.g. functional magnetic resonance imaging) have confirmed that cortical activity related to the perception of illusory and other complex contours is stronger in extrastriate than in striate cortex.

EARLY AND LATE PLASTICITY AND THE EFFECTS OF RETINAL DAMAGE

The modular organization of occipital cortex takes shape during a 'critical period' that extends up to approximately 6 months after birth. During this period, diffuse thalamic inputs to V1 neurons compete with each other such that stronger inputs become strengthened and weaker inputs are suppressed. This competition is molded by the probability of exposure to natural stimuli in the environment. According to this view, orientation selectivity in V1 reflects the abundance of edges in natural visual images. Sensory deprivation experiments during which animals in their critical period are exposed to an impoverished visual environment support this idea. For example, exposure to edges of a single orientation during the critical period leads to an over-representation of neurons tuned to that orientation in V1 and extrastriate cortex.

Many aspects of occipital cortex can also be altered during adult life. Retinal damage, which leads to a scotoma, induces V1 neurons that had RFs within the scotoma to become sensitive to stimuli presented outside the scotoma. Eventually this leads to a permanent remapping of the visual field onto the cortex. Some studies of skill learning suggest that the orientation tuning of V1 neurons, and the extent to which activity of V1 neurons is influenced by stimuli outside the classic RF, can be

altered by experience. These changes in V1, resulting from experience or learning, form the basis of the procedural memory for visual skills. Effects of sensory damage and skill learning on the properties of V1 neurons may reflect changes in the strength of horizontal connections between visual neurons, following the principles of Hebbian learning.

CONCLUSIONS

Areas V1 and V2 in the occipital lobe form the beginning of a cortical network which processes visual input. Although V1 neurons analyze elementary aspects of the visual image, they also contribute to a more complex analysis of surfaces, depth and figure–ground segregation. The more complex properties of V1 depend on lateral connections within V1 and on feedback from extrastriate cortex. The functional organization of V1 and V2 is shaped by the statistics of visual stimulation during a time-limited critical period shortly after birth, but it remains malleable throughout life.

Further Reading

Barlow HB and Mollon JD (eds) (1982) *The Senses*. Cambridge, UK: Cambridge University Press.

Bear MF, Connors BW and Paradiso MA (1996) *Neuroscience: Exploring the Brain*. Baltimore, MD: Williams & Wilkins.

Gazzaniga MS, Ivry RB and Mangun RM (1998) *Cognitive Neuroscience: The Biology of the Mind*. New York: WW Norton & Company.

Lamme VA and Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience* **11**: 571–579.

Ungerleider LG and Desimone R (1989) Neural mechanisms of visual processing in monkeys. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol 2, pp. 267–299. Amsterdam, Netherlands: Elsevier Science.

Wurtz RH and Kandel ER (2000) Central visual pathways. In: Kandel ER, Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, pp. 492–506. New York: McGraw Hill.

Zeki S (1993) *A Vision of the Brain*. Oxford, UK: Blackwell Scientific Publications.

Olfaction and Gustation, Neural Basis of

Introductory article

Peter A Brennan, University of Cambridge, Cambridge, UK

CONTENTS

The senses of taste and smell convey information that is vital for the survival and reproduction of most animals. This information is conveyed by the pattern of activation of receptor cells with different chemical specificities.

INTRODUCTION

Olfaction, by which we mean the sense of smell, and gustation, which is the sense of taste, both provide information about the chemical environment. Taste allows substances entering the mouth to be screened to assess their food qualities and elicit ingestive or rejection behaviors. The olfactory system has the more complex role of detecting and analyzing the vast range of airborne chemicals to provide information about the nature of their source.

ADAPTIVE SIGNIFICANCE OF TASTE AND OLFACTION

As the dietary requirements of different species vary so do their taste systems, but most omnivores such as humans have the ability to sense four classical taste qualities: sweet, salt, bitter, and sour. It has also been acknowledged that some amino acids, especially L-glutamate, give rise to a fifth taste called umami. The purpose of these taste sensations is to provide the body with information about food quality needed to maintain the correct ion balance and energy supply and to avoid potentially harmful foods. (*See* **Hunger, Meals, and Obesity**)

All animals need a supply of energy. Land-dwelling omnivores obtain much of this from carbohydrates, which are digested to yield simple sugars such as sucrose, glucose and fructose. Sensitivity to these sugars in terms of the sweetness of food therefore provides valuable information about its energy content. Maintenance of the body's normal ionic balance depends upon being able to regulate the intake and excretion of sodium ions. Taste plays a vital part in this by identifying the concentration of salt present in food and water. Salt deprivation leads to a craving for salty foods and promotes their ingestion, whereas foods containing high levels of sodium ions are usually rejected, as their consumption would lead to salt overload and dehydration. Sour tastes result from sensitivity to the hydrogen ions found in acids. Although acidity is not necessarily desirable or harmful, it does provide information about the state of food. Unripe fruit has high levels of acidity, as does food that has spoiled owing to bacterial activity. Bitter tastes are frequently caused by toxins that are produced by plants to deter animals from eating them. Animals have evolved the ability to sense these compounds and bitterness is a warning signal that the food may be unsafe to eat. Both sour and bitter tastes elicit rejection behaviors such as spitting and mouth wiping. However, many bitter-tasting substances can in fact be tolerated in the diet in small amounts. Familiarity with these substances can result in an animal acquiring a taste for them, to the extent that they can even become pleasurable.

It should be noted that the terms 'flavor' and 'taste' are often confused. The flavor of a food is a complex sensory perception, most of which is contributed by the smell of the food detected through the olfactory system. More basic information is provided by taste sensation from the taste buds on the tongue, with other receptors providing information about the temperature and texture of the food, and its irritant qualities (such as the burning sensation from chilli peppers). Taste and smell both contribute to the pleasurable qualities of food and the acquisition of food preferences. Animals can also learn to avoid the flavor of foods that cause sickness, especially if they are novel, and

this powerful rejection response may last a lifetime. A further role of the senses of smell and taste is to prepare the body for the digestion of food by stimulating salivation, gut secretions and insulin production from the pancreas.

For many animals, olfaction is the major sensory system through which they are able to sense their environment. Olfaction allows the recognition of food and the fine assessment of its palatability before it encounters the taste receptors. Rats can even learn the smell of a novel food on the breath of other rats, and use it to guide their food preference. Similarly, many young animals develop a preference for the smell of food eaten by their mother, even during their development in the uterus. In addition to food preferences, the olfactory system is involved in prey detection and tracking, as well as predator detection and avoidance. Some animals are able to use odors for navigating their way around their environment. For example, salmon learn the distinctive smell of the stream in which they grew up and use olfactory cues to return there for breeding after spending two years at sea. However, probably the most important use for olfaction is in social communication, especially territorial and sexual behaviors. For instance, male mice use olfactory cues to determine whether a female is sexually receptive and for choosing among potential mates. The sense of smell is much less important for humans than it is for other animals; this is largely because the complexities of human social interactions demand a greater flexibility of behavior than can be governed by chemical cues. Nevertheless, the vast expenditure on perfumes, fragrances and flavorings testifies to the subtle ways in which olfaction can influence human behavior. (*See* **Social Learning in Animals**)

GENETICS OF TASTE AND OLFACTION

Taste and olfaction rely on the interaction of chemicals with receptor proteins or ion channel proteins on the surface of the receptor cells. These sensory proteins are coded by specific genes, and variation in the structure of these genes or their absence can result in differences in the perception of taste and smell among individuals. However, olfaction does not depend solely on genetics, and experience has been shown to have dramatic effects on olfactory sensitivity and discrimination in animals.

Taste Receptor Genes

Chemicals that elicit the sensation of taste are known as 'tastants'. There are a number of different sweet and bitter tastants. However, although they bind to a number of different receptor proteins, they only give rise to the simple sensations of sweetness or bitterness. Several variants of receptors for sweet and bitter tastes have been identified, although they have not yet been associated with the binding of specific tastants. Comparisons of individual sensitivities to a broad array of bitter-tasting substances suggest that at least four types of receptor mediate bitter taste sensation in humans. Individual differences in the genes that are expressed may lead to individual differences in the sensitivity to specific compounds. For example, there are substantial individual differences in sensitivity to the bitter substance *n*-propylthiouracil. The availability of the human genome sequence now makes it possible to identify the genes at specific chromosomal locations, which have been linked with different taste sensitivities, allowing more genes for taste receptors to be identified.

Olfactory Receptor Genes

Airborne chemicals that elicit the sensation of smell are known as odorants. The perception of an odor may occur in response to a single odorant or a complex mixture of odorants. For instance, isobutyric acid on its own has a sweaty odor, whereas a complex mixture of hundreds of individual odorants make up the smell of fresh coffee. The ability of the olfactory system to distinguish between a vast number of odors depends on the binding of odorants to a large array of different receptor proteins. Variability of the amino acid sequence within the odorant binding pocket of a receptor protein causes odorants to be accepted or excluded, based on molecular features such as size, shape and the presence of functional groups. There are about a thousand types of these receptor proteins in rodents such as mice, each binding and responding to a different chemical feature of the odorants. Gene duplication and subsequent mutation have led to the evolution of receptors with different odorant binding specificities, representing the largest superfamily of genes in the genome. However, in contrast to the thousand or so functional receptor types possessed by rodents, humans are only thought to have around 350, which reflects our relatively poor olfactory abilities.

Individual differences in the olfactory receptor repertoire are evident from the existence of specific anosmias, in which individuals may have a greatly reduced sensitivity for certain odorants, such as the sweaty smell of isobutyric acid. There are also individual differences in the perception and naming

of odors. For example, the odorant bangolol is perceived by different individuals as having a pleasant sandalwood odor, an unpleasant urinous odor, or as being odorless.

NEURAL MECHANISMS OF TASTE AND OLFACTION

Taste and olfaction are dependent on the delivery of stimuli to the receptor cells. In the case of taste, saliva and the action of chewing distribute the tastants to taste buds, which contain the taste receptor cells (TRCs). For olfaction, the exploratory behavior of sniffing creates turbulence in the airstream, facilitating the delivery of odorants to the olfactory receptor neurons (ORNs) located in the olfactory epithelium at the top of the nasal cavity. Another important route for odorant access is the retronasal route by which odorants from the mouth and throat reach the olfactory epithelium in exhaled air. This route is particularly important in the perception of the flavor of foods. An additional layer of complexity in odorant delivery is imposed by the layer of mucus that covers the ORNs. Special transport proteins may play an important part in transporting odorants across the mucus layer to the receptors.

Taste Transduction Mechanisms

Taste receptor cells are specialized neuroepithelial cells which are arranged in concentric rings to form a taste bud, similar to the layers of an onion (Figure 1). They are found not only on the tongue but also in the wall of the mouth and throat. Tastants gain access to the receptor proteins on the tip of the TRCs through an opening in the top of the taste bud, called the taste pore. Taste transduction is the process by which the chemical stimulus of a tastant is converted into a change in the electrical potential of the TRC. This activates nerves that contact the TRC and transmit the message to the brain. The sense of taste employs a diversity of transduction mechanisms, reflecting the diversity of taste stimuli. Salt and sour tastants are ions and can affect the electrical potential of the TRC relatively directly. Sodium ions can flow into the TRC through selective ion channels, carrying positive charge into the cell. Hydrogen ions may be able to flow into the TRC through the same channels. In addition, other transduction mechanisms for sour taste appear to exist. One such mechanism may involve an effect of hydrogen ions to block the flow of positively charged potassium ions out of the cell through a selective potassium channel.

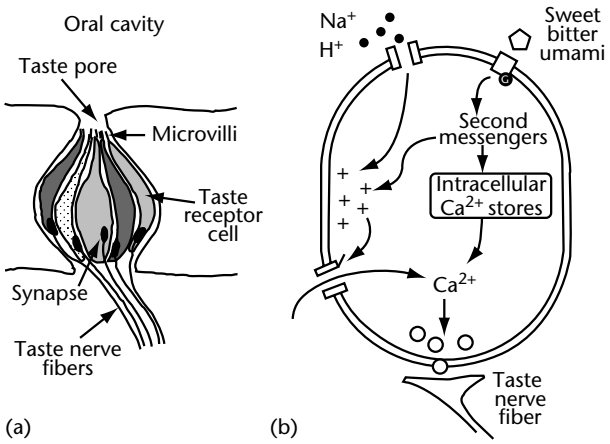


Figure 1. Taste transduction mechanisms. (a) Tastants in the oral cavity gain access to the taste bud via the taste pore. Transduction occurs in microvilli at the apex of the taste receptor cells leading to activation of taste nerve fibers, which contact the basal part of the taste receptor cells. (b) The two main pathways for taste transduction. Ionic stimuli (Na^+, H^+) can enter the receptor cell directly via ion channel proteins to cause a depolarization of the membrane potential. In contrast, complex molecules bind to receptor proteins on the microvilli and activate second messenger systems. These can release calcium ions (Ca^{2+}) from intracellular stores or affect membrane potential by opening or closing ion channels. Flow of positive charge in the taste receptor cell leads to influx of calcium ions via voltage-gated channels and neurotransmitter release, which activates the taste nerve fibers. G, guanine nucleotide binding protein.

Bitter, sweet and amino acid tastants interact with specific receptors in the cell membrane, in a similar manner to odorants in the olfactory system. Binding of a tastant to these receptors activates a cascade of biochemical reactions in the TRC changing the levels of certain chemical messengers, such as cyclic adenosine monophosphate (cAMP) or calcium ions. The identity of these messengers differs among sweet, bitter and umami tastes. However, in each case, changes in the level of the intracellular messenger leads to a change in electrical membrane potential of the TRC and stimulation of the taste nerve fibers. The extent to which combinations of taste transduction pathways are present in individual TRCs is still unclear.

Neural Analysis of Taste Information

Recordings from single nerve fibers show the greatest response to one particular taste sensation but also smaller responses to other tastes. Thus, a fiber that responds best to salt tastants typically

also responds, at a lower level, to sour and sweet tastants. A certain rate of firing in a salt-preferring fiber could therefore be due to a low concentration of salt or to a high concentration of acid. This ambiguity of response means that information about both taste identity and intensity must be represented in the relative rates of activity across a population of taste nerve fibers with different stimulus preferences.

Taste buds are found within papillae, which vary across the tongue in their appearance and the stimulus preferences of their TRCs. The fungiform papillae of the anterior part of the tongue contain taste buds innervated by the chorda tympani branch of the facial nerve. These fibers respond best to pleasant sweet and salt tastants and have little response to bitter tastants. In contrast, taste buds in circumvallate and foliate papillae found on the posterior part of the tongue are innervated by a branch of the glossopharyngeal nerve; they typically respond well to sour and bitter tastants, but have little response to sweet stimuli. Taste buds in the larynx and upper esophagus respond mainly to salt, sour and distilled water stimuli. These taste buds have the role of monitoring the ionic environment of the throat.

The pathway by which taste information travels to the brain is similar in most vertebrates. Taste nerve fibers project primarily to the nucleus of the solitary tract in the brainstem. Many reflexes concerned with the ingestion, digestion or rejection of food are integrated at this level. The response specificity of taste neurons in more central brain regions broadens to include thermal and tactile stimuli. From the nucleus of the solitary tract, there is a projection via the thalamus to the cerebral cortex. It is here that higher-order taste information is linked to olfactory information for the perception of flavors.

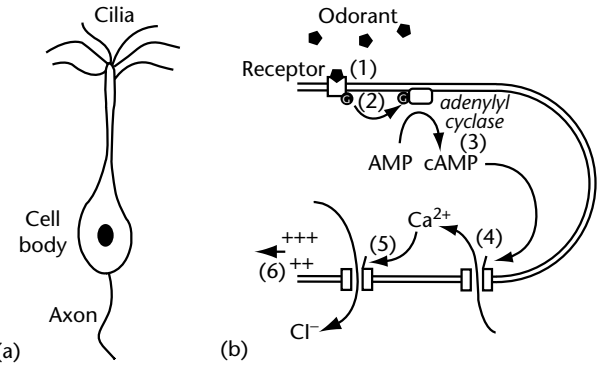


Figure 2. The mechanism of olfactory transduction. (a) Olfactory receptor neurons are bipolar cells. Transduction occurs in the apical cilia leading to the influx of positive charge that spreads passively to the cell body and triggers activity in the olfactory nerve fiber axon. (b) The transduction mechanism in an olfactory cilium. The binding of an odorant to a receptor protein (1) activates a guanine nucleotide binding protein (G) as step (2), which in turn increases cyclic adenosine monophosphate (cAMP) production by adenylyl cyclase (3). Cyclic AMP opens channels that allow positively charged sodium ions (Na^+) and calcium ions (Ca^{2+}) into the cell (4). The increased level of calcium ions in the cilium opens chloride ion (Cl^-) channels allowing negatively charged chloride ions to flow out (5). The influx of Na^+ and Ca^{2+} and the efflux of Cl^- result in net flow of positive charge into the receptor cell, which triggers nerve activity (6).

Olfactory Transduction

Olfactory transduction (Figure 2) is the mechanism by which the binding of an odorant to receptor proteins produces a change in the electrical potential of the ORN. In vertebrates, binding of an odorant to a receptor protein activates a cascade of biochemical reactions resulting in an increase in the level of the intracellular messenger cAMP. This cAMP directly opens an ion channel that allows both sodium and calcium ions to flow into the ORN. In addition to carrying positive charge into the cell themselves, they open a chloride ion channel that causes negatively charged chloride ions to flow out of the ORN. This change in electrical potential triggers activity in the olfactory nerve fibers, which transmit the information to the brain. Mice that have been genetically manipulated to lack these cAMP-activated ion channels have no sense of smell. An increase in the level of calcium ions in the ORN also reduces its activity in response to sustained stimulation. This important feature of olfaction reduces the sensitivity to the background odors to which an animal is constantly exposed, enhancing the response to important odors.

Coding of Olfactory Information

Each ORN expresses only a single type of receptor protein. However, the receptors are not thought to be highly tuned to respond to only a single molecule. Instead, they appear to respond to a particular chemical feature of a range of related odorants, such as the number of carbon atoms in a chain, or the presence of a functional group such as —OH. It may seem surprising that olfactory receptors are not specific in their response to odorants, but this feature allows the olfactory system to respond to an enormous range of chemical stimuli. In the case of

humans, if each of the 350 or so olfactory receptors responded specifically to a single odorant, the olfactory system would only be sensitive to 350 odorants. Instead, the olfactory system analyzes the pattern of responses generated by an odorant across the whole receptor array. It is this overall pattern that is specific for a particular odorant, rather than the ambiguous response of an individual ORN, and the olfactory system can thus respond to many thousands of different odorants. In general, similar molecular structures will tend to fall within a broad odor class, such as 'fruity' or 'minty'. However, there are exceptions, and similar structures can have very different odors. For example, (*R*)- and (*S*)-carvone have the same molecular formula but are mirror images of each other, and they have different smells. Moreover, molecules with quite different structures can smell remarkably similar. Not only will a particular olfactory receptor respond to a range of similar odorants (known as the molecular receptive range), but a single odorant will also activate a number of different olfactory receptor proteins, which have different affinities. Therefore, different concentrations of an odorant may activate different populations of ORNs. This means that the olfactory system must be able to generalize these slight differences in the patterns of receptor activity to give consistent odor perception across a range of odor intensities.

Neural Analysis of Olfactory Information

The ORNs send their odor information to a region of the brain called the olfactory bulb. Here they provide input to mitral cell neurons in specialized spherical structures called glomeruli. Although ORNs that express a particular type of receptor protein are scattered across large areas of the olfactory epithelium, they send their axons to only one or two glomeruli in the olfactory bulb. As each glomerulus receives information from only one receptor type, it may be thought of as a fundamental unit of odor processing. The task of identifying the pattern of receptors that are activated by an odor is therefore reduced to that of identifying the pattern of activated glomeruli in the olfactory bulb.

The initial stage of odor analysis is performed in the olfactory bulb (Figure 3). Mitral cells collect information from a single glomerulus and have inhibitory interactions with neighboring mitral cells via inhibitory interneurons. This lateral inhibition integrates the information about glomerular activity into a pattern of mitral cell activity that represents that odor. Most natural odors are complex mixtures of different odorants and it is often the subtle differences in the minor components that are biologically important. Learning plays an important role in associating the components of a complex odor to form a distinct odor representation that can be distinguished from similar odors with different meanings. This learning involves substantial neural changes at the level of the olfactory bulb. Odor information also has to be linked with other sensory information and appropriate behavioral responses. This aspect of learning occurs in more central brain areas. (*See* **Neural Inhibition**)

The main output of the olfactory bulb is to the piriform cortex. This simple cortical structure has extensive interconnections with the rest of the brain and may have a role similar to higher-order association areas in other sensory systems. From here, odor information reaches the orbitofrontal cortex. This region is probably important for attributing meaning to odors and flavors. Medial temporal lobe structures such as the hippocampus are thought to be important for flexibly learning relationships among odor stimuli and among stimuli from other senses. The large olfactory projection to the amygdala is important for learning the emotional response elicited by meaningful odors. There is also a large olfactory input to a region of the brain called the hypothalamus that is involved with feeding and reproductive behaviors.

DISORDERS OF TASTE AND OLFACTION

Many people who complain of a lack of taste in their food actually have an olfactory deficit. True taste disorders are uncommon. There may be a loss of the ability to detect salt, sweet, sour and bitter tastes, or abnormal tastes may be present in the mouth. The most common cause of a loss of taste sensation is damage to the nerve supply resulting from surgery, dental treatment or the presence of a tumor.

The complete absence of a sense of smell is known as 'anosmia'. This condition can be congenital and have a genetic basis. However, this is rare. Anosmia usually arises from injury or illness that results in damage to the ORNs or their nerves, or blockage of the nasal passages. Physical trauma such as a blow to the head can shear the olfactory nerves, disrupting transmission of olfactory information. Being exposed to the environment, ORNs can also be damaged by noxious chemical vapors,

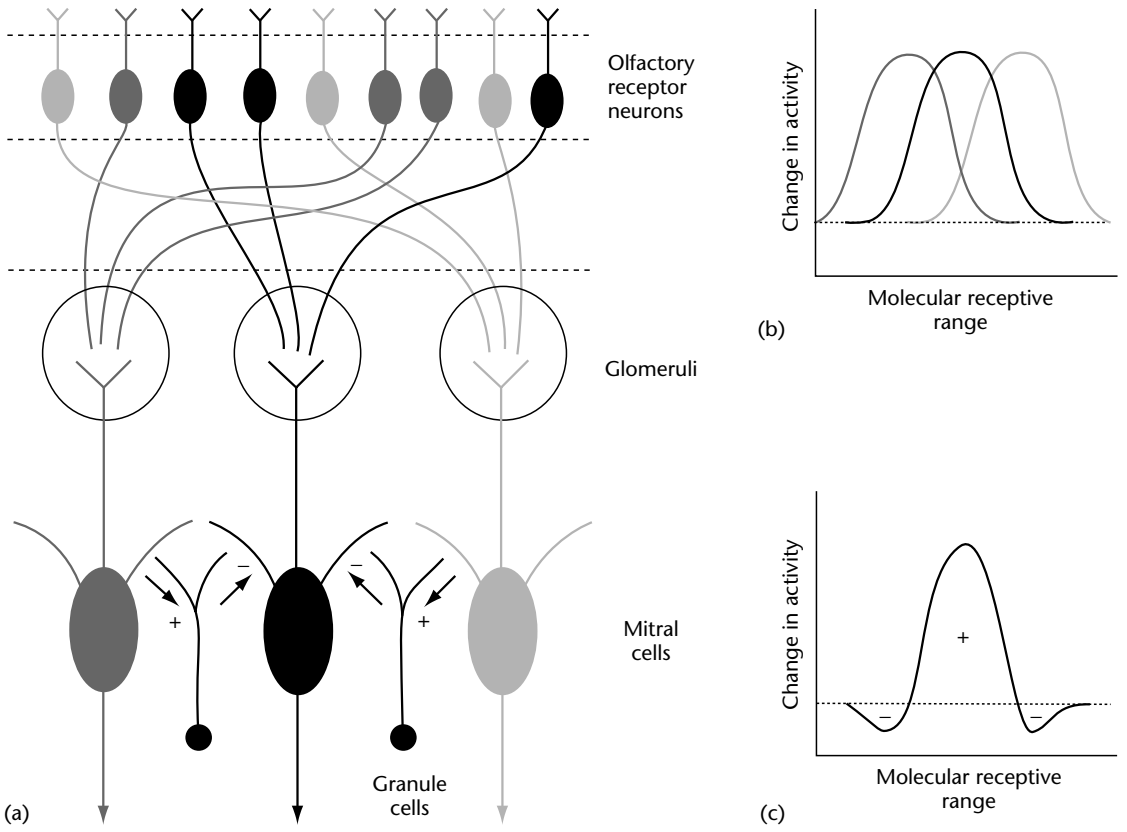


Figure 3. The initial stages of odor processing in the olfactory bulb. (a) Olfactory receptor neurons expressing a single receptor type are randomly distributed in the olfactory epithelium but their axons converge onto one glomerulus in the olfactory bulb. Mitral cells receive information from a single glomerulus and therefore from a single receptor type. The first stage of comparison of the activity across receptor types occurs in the olfactory bulb, by lateral inhibition from neighboring mitral cells, via granule cell inhibitory interneurons. (b) Olfactory receptors respond to a variety of similar odorants and the molecular receptive ranges for individual olfactory receptor neurons overlap. (c) Lateral inhibition acting on the central mitral cell from neighboring mitral cells that respond less well to the same stimuli, sharpens its molecular receptive range. The mitral cell is excited by a narrower range of similar odorants and is inhibited by odorants just outside that range, which stimulate neighboring mitral cells more effectively. This increases the contrast in the pattern of mitral cell activity compared with that of the olfactory glomeruli.

especially caustic ones such as ammonia. The receptor neurons are also vulnerable to infection, particularly by viruses, which can lead to a total loss of function. However, probably the most frequent cause of anosmia is a physical obstruction of the nasal passages that prevents odorant access to the olfactory epithelium. This can be caused by a displaced nasal septum, nasal polyps, or swelling due to infection or irritation of the nasal passages. Many people with anosmia do not complain about the condition and live quite normally, although they are unaware of their body odor, which can be a cause of anxiety. However, such people are subject to serious risks due to the lack of the warning function provided by the olfactory system. They need to install smoke protectors to warn of fire and take precautions when using gas appliances. Additionally, without a sense of smell to elicit an aversive reaction, they need to be careful not to eat food that has gone bad.

Hyposmia is a less severe olfactory deficit, in which the sensitivity to odorants is reduced. It is commonly associated with smoking and with allergies such as hay fever. The incidence of hyposmia increases after the age of 60 years, and olfactory sensitivity drops severely after the age of 70 years. This can have adverse consequences for the diet of elderly people, as most of the flavor of food and the pleasure derived from eating it relies on olfactory sensation. The loss of odor sensitivity in elderly patients can lead to a loss of appetite, which contributes to slow recovery from illness. However, to some extent this can be overcome by adding extra flavoring to food. A more distressing olfactory

disorder is that of parosmia in which the identification of odors is distorted; for example, coffee may be perceived as having an unpleasant chemical or rotting odor. Rarely, people may suffer from phantosmias, in which a persistent perception of a (usually unpleasant) odour is continually present in the absence of odor stimuli, and pervades their life.

Olfactory function may be disrupted in various diseases and can even be used in their diagnosis: for example, 42% of people with Parkinson disease are reported to be anosmic. Moreover, olfactory deficits show up at an early stage of Parkinson disease and are so distinctive that they can be used to differentiate the condition from related motor disorders. Olfactory function also appears to be abnormal in depression and Alzheimer disease, although in these cases the disorder may reflect abnormal functioning of central brain systems rather than solely an olfactory deficit. They may have problems naming odors, although this is generally difficult to test satisfactorily owing to their limited attention span. People with depressive disorders show a lower arousal to pleasant odors and a greater arousal to unpleasant odors than normal. Liver or kidney disease and some metabolic disorders result in the accumulation of toxins in the bloodstream, which can affect olfaction. Endocrine disorders can also be associated with reductions in olfactory sensitivity: 75% of people with diabetes are reported to be hyposmic or anosmic. As ORNs are exposed to damage from the environment, they are continually replaced by new receptor cells. This can lead to the restoration of olfactory sensitivity in cases of damage to the mature receptors. However, conditions that affect this replacement can result in olfactory deficits. Hypothyroidism impairs the maturation of olfactory receptors, resulting in reductions in ORN density in the olfactory epithelium. This is associated with hyposmia. Chemotherapy or radiotherapy in the head region may also disrupt the replacement of the receptor neurons and is likely to lead to hyposmia. However, in most cases the deleterious effect of the primary disorder on quality of life is far greater than that caused by olfactory dysfunction. (*See* **Alzheimer Disease**; **Affective Disorders: Depression and Mania**)

Further Reading

Griff IC and Reed RR (1995) The genetics of olfaction. *Current Opinion in Neurobiology* **5**: 456–460.

Kinnamon SC and Margolskee RF (1996) Mechanisms of taste transduction. *Current Opinion in Neurobiology* **6**: 506–513.

Lindemann B (2000) A taste for umami. *Nature Neuroscience* **3**: 99–100.

MacLeish PR, Shepherd GM, Kinnamon SC and Santos-Sacchi J (1999) Sensory transduction. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 671–717. San Diego, CA: Academic Press.

Mombaerts P (1996) Targeting olfaction. *Current Opinion in Neurobiology* **6**: 481–486.

Mori K (1995) Relation of chemical structure to specificity of response in olfactory glomeruli. *Current Opinion in Neurobiology* **5**: 467–474.

Smith DV and Shepherd GM (1999) Chemical senses: taste and olfaction. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL and Squire LR (eds) *Fundamental Neuroscience*, pp. 719–759. San Diego, CA: Academic Press.

Smith DV and St John SJ (1999) Neural coding of gustatory information. *Current Opinion in Neurobiology* **9**: 427–435.

Pain and Analgesia, Neural Basis of

Introductory article

AD Craig, Barrow Neurological Institute, Phoenix, Arizona, USA

CONTENTS

Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage. Analgesia is the alleviation or the absence of pain.

INTRODUCTION

Pain is a multidimensional sensory experience that is essential for survival. It is a somatic distress signal that indicates unhealthy physiologic conditions in the body, warns of the danger of physical harm, or alerts to injury. Although subjective, pain is a specific, discriminative sensation that can be cognitively evaluated, localized, scaled, timed, and qualitatively described. Pain has a defining, affective unpleasantness that motivates behavior, which can lead to the long-term psychological experience of 'suffering'. The unpleasantness can be graded and qualitatively differentiated, ranging from the discomfort of fatigued muscles or a small cut to the agony of a toothache or gallstone.

Pain was once regarded as an exteroceptive somatosensory modality (like touch), but pain signals the condition of the body itself, not the features of an impinging external object or the movements of the body, and skin is but one organ from which pain can originate. Like hunger, thirst, itch, and other 'feelings' from the body, the affective drive (or motivation) that is an essential characteristic of pain reveals an evolutionary origin in the process of homeostasis (the regulatory mechanisms that maintain the body and generate basic drives). This is evidenced by the reflex autonomic responses and interactions that pain generates and by the integration with emotion that is inherent in pain. The neuroanatomy of pain in humans substantiates the concept that pain is a specific aspect of interoception, the sense of the physiologic condition of the body itself. A direct interoceptive representation in the cerebral cortex incorporates pain, temperature, itch, and other feelings from the body; thus, pain is part of a system that provides an important basis for awareness of the material self. This cortical representation is essentially unique to primates, and it is well developed only in humans. Conceptual recognition that pain is an emergent sensation based on the homeostatic and motivational systems of the central nervous system provides a fundamental framework for understanding the psychological interactions between pain and general health, as well as a sound reason for treating pain clinically with an integrated, multidisciplinary approach.

QUALITIES OF PAIN

Psychophysical studies of cutaneous pain indicate that mechanical or thermal pain can be localized nearly as well as touch (within 4–10 mm on the hand and arm). Perceptual thresholds for pain vary with body region, time of day, and perhaps gender. The mean force threshold for mechanical cutaneous pain is approximately 10 g per mm circumference of a flat probe; the mean cutaneous pain thresholds to heat and cold are about 45°C and 15°C, respectively. Different levels of painful heat can be discriminated at steps of about 1°C, and paired forced-choice trials indicate that differences of about 0.4°C can be detected on the face and the hand. Whereas differences much smaller than 1°C can be detected at innocuous cool temperatures, steps of over 4°C are required for discrimination in the noxious cold range.

Qualitatively different feelings are associated with pain caused by different cutaneous stimulus modalities (mechanical, hot, cold, or chemical stimuli) or with pain from different tissues (muscle or viscera), indicating different neural mechanisms.

Sudden noxious stimuli (whether by electrical or natural means) elicit two distinct pain sensations: an immediate sharp, pricking sensation, and about 1 s later an unpleasant burning sensation. These 'first' and 'second' pain sensations are associated with fast-conducting A fibers and slow-conducting C fibers respectively in peripheral nerves, and different sets of spinal neurons. Activity of C fibers from deep (noncutaneous) tissues produces a dull aching or cramping sensation.

Brief (less than 1 s) noxious heat stimuli repeated at short intervals (3 s) cause an augmenting second pain sensation but only a weak first pain sensation. This temporal summation occurs centrally (in the spinal cord and brain). In normal people it shows a striking reset phenomenon; that is, the summation caused by one series of repeated stimuli vanishes after 4–8 s and begins again from baseline in a subsequent series. Dysfunction of the mechanism underlying the rapid inhibition of this central augmentation may be critical in people with chronic pain. Noxious mechanical stimuli (e.g. pressure from a thin point) generate a first pain sensation that also augments if maintained for minutes, which similarly results from central summation.

A surprising sensation of burning pain can be elicited by the thermal grille illusion displayed in many science museums. Innocuous (nonpainful) warm and cool bars that are spatially interlaced produce a painful, ice-like feeling that mimics the burn of noxious cold. This unmasking (or disinhibitory) phenomenon demonstrates that there is a central mechanism for the inhibition of pain by cooling that is reduced by simultaneous warming; this mechanism reflects the importance of the integration of pain and temperature sensibilities for homeostasis (i.e. thermoregulation). The burning pain felt when lukewarm water is applied to feet that are numb from cold is a similar phenomenon. Painful cold evoked by the cold pressor test, in which the hand is held in circulating water at 4°C, is often used to test the effect of drugs on pain tolerance.

Pathologic activity from visceral tissues can produce referred pain localized to a cutaneous zone represented in the same spinal segments. Many pain conditions produce hyperesthesia, or increased sensitivity to any stimuli in a cutaneous zone. For example, pain from cardiac tissue is referred to the left shoulder in angina pectoris, and cutaneous facial hypersensitivity often accompanies a toothache. A zone of hyperalgesia, or increased pain from noxious stimuli, usually surrounds an injury site. Pain evoked by normally nonpainful mechanical or thermal stimuli, called allodynia, may also occur.

In general, pain is influenced by many physiological processes of homeostasis, not only by coolness and warmth, but also by exercise, food ingestion and cardiovascular hypertension, as well as by attention and cultural factors. Pain is also modulated by hormones, particularly those released by stressors, including corticosteroids, catecholamines, androgens, and immune-related cytokines, for which pain-processing neurons have specific receptors.

EFFECTS OF INJURY

Damage or infection elicits a local tissue injury response (inflammation). This includes the production of eicosanoids (metabolites of arachidonic acid from cell membranes) including prostaglandins, bradykinin (from plasma), lactic acid, and superoxide free radicals (from leukocytes and mast cells), as well as serotonin, histamine, and proinflammatory cytokines. These antimicrobial mediators activate or sensitize A and C fiber endings. Activated C fibers contribute a neurogenic component to inflammation by peripheral release of glutamate, peptides, and adenosine triphosphate (ATP), which further enhance vasodilatation and plasma extravasation (increased permeability between capillary endothelial cells), activate mast cells, and influence release from sympathetic nerve fibers. This increases tissue infiltration with blood cells, plasma, and cytokines, which in small quantities results in erythema (redness or flare), and in larger quantities in edema (swelling). The spread of this inflammatory cascade is restricted by systemic corticosteroids (released by activation of the hypothalamo–pituitary–adrenal axis). Trophic agents, such as nerve growth factor, are also induced and may play a part in neuropathic pain, in addition to their role in healing.

Sensitization of nociceptors produces the condition of primary hyperalgesia; for example, following sunburn, even light touch or warmth hurts (lowered threshold) and a hot shower is more painful than normal (increased activation). Inflammatory hyperalgesia is important in many diseases, such as arthritis. A zone of secondary hyperalgesia also surrounds an injury, due to increased spinal neural excitability (central sensitization). Preventing C-fiber activity during an injury from reaching the spinal cord can avoid this; this is the rationale for prophylactic use of analgesic agents prior to surgery (preemptive analgesia). Central sensitization depends on a chemical

cascade elicited by immune-competent glial cells that release cytokines in the spinal cord.

Unhealthy physiologic tissue conditions such as ischemia, hypoglycemia, or excessive muscular activity also cause pain. Loud noises that threaten eardrum rupture cause pain (due to excessive middle ear muscle contraction). These situations cannot be rectified by homeostatic mechanisms alone and require behavioral compensations for survival. Thus, hyperalgesia produces protective behavior that allows tissue healing without disturbance. Children or animals born lacking peripheral C fibers have congenital pain insensitivity and generally do not survive into adulthood.

NEURAL MECHANISMS OF PAIN

Peripheral sensory fibers that innervate all tissues of the body are sensitive to chemical or metabolic stimuli (low pH, hypoxia, hypoglycemia, ATP, bradykinin, and lactic acid), and to strong mechanical or thermal stimuli. These include nociceptors that respond selectively to noxious (or potentially damaging) stimuli. These have thinly myelinated (A) or unmyelinated (C) fibers and small cell bodies in the dorsal root ganglia. The receptor terminals are free nerve endings. Some A-fiber nociceptors respond specifically to noxious mechanical stimuli, noxious heat stimuli, or both. Most C-fiber nociceptors are polymodal (sensitive to noxious mechanical, thermal, and chemical stimulation). Some C fibers are unresponsive unless sensitized by inflammatory agents. The discharge rate of nociceptors is low (C fibers rarely fire faster than 20 Hz), and central summation is required for the perception of pain. Above the perceptual threshold, the intensity of pain directly correlates with the number of action potentials fired.

The small-diameter primary afferents develop in a second wave of neurogenesis after the large-diameter fibers (the mechanoreceptors). They enter the spinal cord through the medial division of the dorsal roots, arborize over one to three spinal segments, and terminate in laminae I and II of the superficial dorsal horn. Collaterals terminate also in lamina V (the neck of the dorsal horn) and lamina X (near the central canal).

Spinal nociceptive neurons in lamina I and lamina V send ascending axons to the brain; these provide specific and integrative activity, respectively. In lamina I, there are distinct types of neurons that respond to specific stimuli within small receptive fields (in skin, muscle, joint, or viscera). Nociceptive-specific (fusiform) lamina I cells dominated by A-fiber input correlate with first pain; polymodal nociceptive (multipolar) lamina I cells dominated by C-fiber input correlate with second pain. Lamina I cells that respond specifically to innocuous cooling, to warming, or to histamine correspond directly with thermal or itch sensations, respectively. The nociceptive neurons in lamina V have large receptive fields and considerable background activity, and they respond both to nonnoxious and noxious mechanical stimuli (over a wide dynamic range). They integrate afferent activity from all somatic tissues; as a population, their activity is related to the cumulative intensity of stimulation. Nociceptive cells that respond to visceral stimulation can have an additional cutaneous receptive field, which is probably the basis for referred pain. The substantia gelatinosa (lamina II) was once considered to be the pain relay, but it contains only small inhibitory interneurons that arborize locally; the functional organization of this modulatory region is a mystery.

The modality-specific lamina I neurons project to hierarchically organized homeostatic regions, including sympathetic spinal neurons and homeostatic brainstem areas. These projections provide the basis for somatoautonomic reflex adjustments of cardiorespiratory function in response to noxious, thermal, or metabolic stimuli (including exercise). Accordingly, lamina I receives descending controls from brainstem homeostatic areas and from the master autonomic control center, the hypothalamus. The integrative lamina V neurons project to spinal motoric regions and to the brainstem reticular core, where they affect somatomotor integration and behavioral state. (*See* **Autonomic Nervous System**)

Thus, the parabrachial nucleus, the major homeostatic integration site in the brainstem, is involved in pain processing; it receives both lamina I (sympathetic) and vagal (parasympathetic) afferent input. It is interconnected with most other homeostatic sites in the brainstem and forebrain, including the periaqueductal gray matter, and it projects pain-related activity to the hypothalamus and amygdala. The amygdala is involved in fear conditioning and in opiate-induced analgesia. The periaqueductal gray matter is the major brainstem site controlling homeostasis, vocalization, facial expression, and emotional, sexual and defensive behavior, and it drives descending antinociceptive control of spinal neurons. Finally, indirect nociceptive inputs to the hypothalamus from the parabrachial nucleus and catecholamine neurons in the medulla affect basic emotional drives, neuroendocrine responses, and thermoregulatory, osmoregulatory, and immune system control. (*See* **Amygdala**)

ROLE OF THE CEREBRAL CORTEX

The main pathway for pain sensation is the lateral spinothalamic tract (STT), which crosses the midline to ascend contralaterally in the middle of the lateral funiculus and consists primarily of lamina I axons. Human cordotomy cases indicate that this pathway is clinically and behaviorally critical for pain, temperature, itch, and other feelings from the body. Lamina I STT axons terminate in a specific relay nucleus in the posterior thalamus (VMpo), in a portion of medial dorsal thalamus (MDvc) that projects to the anterior cingulate cortex (ACC), and in a part of the main somatosensory nucleus of thalamus (VP), which relays mechanoreceptive lemniscal activity from the dorsal column nuclei to somatosensory cortex. (Lamina V axons ascend in the anterior STT in the ventrolateral funiculus and terminate in patches within the VP.) Ancillary multisynaptic pathways (e.g. the parabrachial nucleus) provide pain information to the forebrain. Nociceptive projections in subprimates are fundamentally different. (*See* **Somesthesis, Neural Basis of**)

The specific pain and temperature relay nucleus VMpo in turn projects topographically to a distinct portion of the mid/posterior insular cortex (and to Brodmann's area 3a near the central sulcus). This insular cortical field, which is well developed only in humans, constitutes a limbic sensory area that is interconnected with the amygdala and hypothalamus, and it represents the physiologic condition of the entire body. In turn, this area projects to more anterior insular and orbitofrontal cortical areas on the right (nondominant) side that are neurologically associated with emotional evaluation of the physical self, which is consistent with the 'somatic marker' hypothesis of consciousness and the James–Lange theory of emotion. (*See* **Neural Correlates of Consciousness as State and Trait**)

Pain sensation clearly involves the cerebral cortex. The long-held (incorrect) view that pain is sensed in the thalamus was based on the failure of lesions of the somatosensory cortex (which did not damage the insular cortex) to produce analgesia. Analysis of cortical activity during pain is a vigorous field of study. Functional neuroimaging and evoked potential studies indicate that mid/posterior insular cortex and the ACC are coactivated by painful stimuli (Figure 1). Somatosensory cortical activation is seen only in some studies, perhaps because noxious stimuli inhibit the primary somatosensory area 3b as they activate the adjacent area 3a. Other regions activated include dorsolateral prefrontal cortex, striatum, cerebellum, hypothalamus,

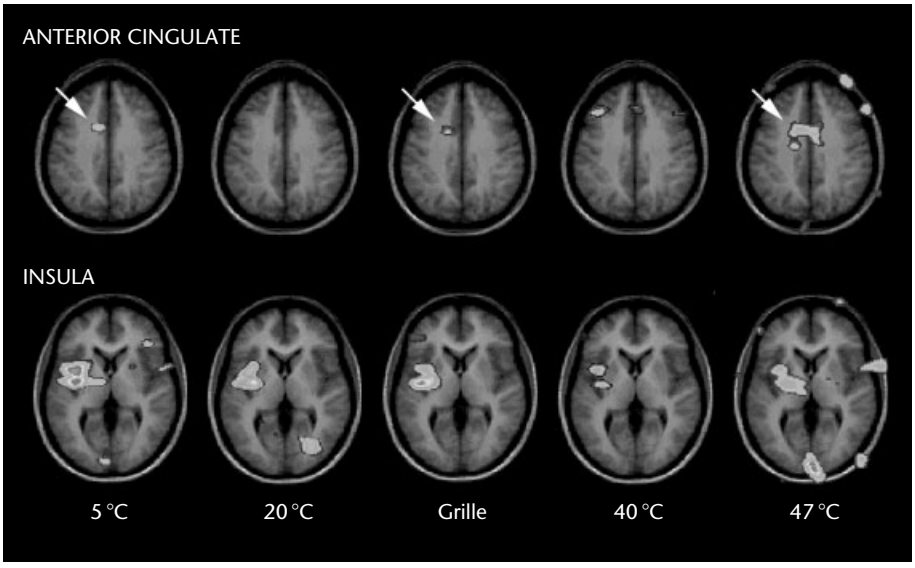


Figure 1. Functional imaging by positron emission tomography of activation in the human brain due to application of noxious heat, noxious cold, innocuous warmth, innocuous cool, and the thermal grille (an illusion of pain) to the right hand. Activation occurred in the left insular cortex with all stimuli, but activation was observed in the anterior cingulate cortex only with the noxious stimuli and the grille stimulus (in which interlaced innocuous warm and cool bars elicit a feeling of ice-like burning pain). This experiment showed that the thermal grille elicits a pattern of brain activation indistinguishable from that produced by noxious cold, validating the psychophysical inference that the illusory feeling it unmasks is like the burn of noxious cold. These images demonstrate the two main regions of the brain activated by painful stimuli in all imaging studies. From Craig *et al*. (1996) *Nature* **384**: 258–260.

amygdala, and periaqueductal gray matter. Many of these areas are interconnected, and the experience of pain clearly involves multiple forebrain regions that interact within a complex network; yet, associations of particular regions with different roles are likely. The VMpo projection field in insular cortex seems to serve as a primary sensory field for pain, temperature, itch, muscle and visceral sensations, and be associated with sensory identification, memory and homeostasis; abnormal activation of this region has been observed in people with chronic pain. The ACC appears particularly important for motivation (urgency), attention and response guidance; its activation is selectively correlated with the perception of thermal pain and with hypnotic modulation of pain unpleasantness. Expectation of pain activates more rostral areas in the ACC and the insula, reflecting further processing stages. Lesions involving VMpo or insular cortex can severely impair pain and temperature sensation, but they can also result in a paradoxical central pain syndrome. Lesions of ACC may blunt the emotional (but not the sensory) aspect of pain, but the great anatomical variability between human brains in this region presents a serious confound. (*See* **Neuroimaging**)

NEURAL MECHANISMS OF ANALGESIA

The classic examples of endogenous neural pain control are soldiers with massive wounds who do not complain of pain, and the placebo analgesia that can be elicited in patients with organic causes of pain. Antinociceptive (analgesic) circuits probably evolved because they enhance survival (by enabling defense behaviors), and because they facilitate recuperative illness responses. Central control of pain makes sense because healing requires considerable time. Several mechanisms are recognized.

Fast inhibition of second (C-fiber) pain is initiated by low-threshold A-fiber activity (e.g. rubbing, vibration) and by A-nociceptor activity (counterirritation). These effects can be demonstrated during a progressive pressure block of peripheral nerve conduction and occur at spinal and cortical levels, but they are still poorly understood. They suggested the heuristic gate control theory of pain integration. Dysfunction of these mechanisms could underlie touch-evoked neuropathic pain (mechanical allodynia). These mechanisms may be engaged by transcutaneous electrical nerve stimulation (TENS) and by dorsal column stimulation, procedures used by some clinicians for pain control.

Innocuous thermal stimuli inhibit C-fiber evoked pain. Cooling has a peripheral palliative effect on inflammation and on sensitized nociceptors, and it inhibits pain processing in the forebrain. This interaction is demonstrated by the thermal grille illusion of pain, in which simultaneous cooling and warming of the skin minimizes ascending thermosensory activity; this unmasks (disinhibits) the cold-evoked burning sensation associated with noxious cold that is normally inhibited by ongoing thermosensory activity. Accordingly, the thermal grille produces the same pattern of forebrain activation as noxious cold, which differs from the insular cortical activation produced by innocuous thermal stimuli because it also produces activation of the ACC (Figure 1). Thermosensory inhibition of pain processing may occur by direct and indirect interactions between these two cortical areas.

Electrical or chemical stimulation of descending and ascending pathways from the periaqueductal gray matter can cause stimulation-produced analgesia. These pathways activate serotonergic, enkephalinergic and catecholaminergic cells in homeostatic control regions that modulate nociceptive processing in the spinal dorsal horn. Descending bulbospinal projections also include pronociceptive (facilitatory) components, which may be imbalanced in neuropathic pain. These projections have direct actions on sympathetic neurons, but in certain conditions deep brain stimulation can alleviate chronic pain without autonomic changes or emotional experiences.

The endogenous pain control systems are activated by environmental stressors (stress-induced analgesia) or danger signals, which can be classically conditioned or learned. Meditation or biofeedback training may access these systems. Conversely, endogenous antianalgesia (pronociceptive) circuits are activated by learned safety signals that indicate when danger will not occur. Such circuits are distinguishable from the analgesia circuits and oppose or reverse their effects at the spinal level by the spinal release of peptides (particularly cholecystokinin) which can block analgesia produced by stress or even morphine. Such antianalgesia mechanisms may play a key part in chronic pain and in the development of morphine tolerance.

ROLE OF OPIATES AND ASPIRIN IN PAIN

The main analgesic drugs used today are nonsteroidal antiinflammatory drugs (NSAIDs) and the opiates. These agents are useful for acute, organic

pain, yet development of better methods for the alleviation of both acute and chronic pain is urgently needed, because pain has enormous costs for both individuals and society.

The prototypical NSAID is acetylsalicylic acid (aspirin), originally derived from willow tree bark. Aspirin permanently blocks cyclooxygenases (COX), which catalyze the production of prostaglandins from arachidonic acid (a cell membrane lipid). Prostaglandins produced during tissue inflammation sensitize nociceptors to other metabolites (ATP, lactic acid, bradykinin) and decrease their thresholds to mechanical and thermal stimuli. Thus, blocking the peripheral production of prostaglandins reduces ongoing pain and hyperalgesia. Corticosteroids have a similar effect by preventing the induction of COX and proinflammatory cytokines (which also sensitize nociceptors).

The NSAIDs inhibit both the COX-1 (constitutive) and COX-2 (inducible) isozymes, which produce different prostaglandins and are localized in different tissues. The constitutive form COX-1 is essential for the health of the stomach and kidney, where chronic high doses of aspirin cause damage. The prostaglandins made by COX-2 are involved in inflammation, sensitization and fever, and thus selective COX-2 inhibitors ('superaspirins') have been developed to minimize the gastrointestinal effects of aspirin.

These analgesics also have central effects. Prolonged activity in C-nociceptors causes production of prostaglandins and proinflammatory cytokines by astrocytes and microglia in the spinal cord, which sensitizes spinal nociceptive neurons and increases neurotransmitter release from primary afferents. Spinal (intrathecal) administration of small amounts of NSAIDs (or cytokine antagonists) reduces pain behavior and neurogenic inflammation (caused by antidromic activity in nociceptive fibers and the peripheral release of peptides). In addition, systemic NSAIDs act synergistically with endogenous opiate-mediated pain control mechanisms in the spinal cord and in the brainstem.

The opioids (which include classic opiate drugs such as morphine and also peptides with similar actions) are still the most potent analgesics available. They have both peripheral and central actions. Peripheral opioid injections affect immune cell activity and reduce inflammatory sensitization of nociceptive primary afferents. In the spinal cord, opioids inhibit nociceptive afferents and the postsynaptic nociceptive spinal neurons. In contrast to NSAIDs, both sensitized and normal pain transmission are reduced by opiates. All three types of opiate receptors (mu, delta, and kappa) are involved, but lumbar epidural injections of fast-acting μ-selective opiates (e.g. fentanyl) are clinically most effective. Such injections usually include the adjuvant bupivacaine, a long-lasting local anesthetic agent that blocks small-diameter fibers, reducing the dosage of opiate needed and the danger of respiratory depression due to rostral spread. Kappa opioids seem to be effective for deep pain selectively in women.

Systemically administered opiates engage endogenous pain control mechanisms in the spinal cord and the forebrain, particularly in the brainstem periaqueductal gray matter and the amygdala. Mu-selective agonists activate the descending pathways involving serotonergic and noradrenergic fibers that are activated by stimulation-produced analgesia. Another antinociceptive pathway normally activated by endogenous β-endorphin in the periaqueductal gray matter is not engaged by morphine.

Nociceptive primary afferent fibers activate dorsal horn cells with the same neurotransmitters (glutamate or aspartate) used by most neurons, and agents that act at these receptors (e.g., dextromethorphan) can affect pain. They also release other substances (e.g. ATP), including the peptides substance P and calcitonin gene-related peptide (CGRP), which have long-lasting postsynaptic actions. Spinal substance-P receptive neurons are located almost exclusively in lamina I, and their abolition reduces chronic pain behavior in animal models. The inhibitory transmitters in the spinal cord include γ-aminobutyric acid (GABA), glycine, enkephalin, dynorphin, and adenosine, which are used by lamina II interneurons that modulate nociceptive processing.

MANAGEMENT OF CLINICAL PAIN

Effective relief from clinical pain of organic origin should be managed according to the 'three-step ladder' of the World Health Organization. Medications with the fewest side effects and the least addictive potential – NSAIDs – are offered first, in addition to treating the organic cause. Their efficacy is limited by a 'ceiling' dose beyond which they are no longer therapeutic, and their potential for gastrointestinal and renal injury. The second step is the combined use of NSAIDs and weak opioids such as codeine. This achieves a synergistic effect which decreases the doses needed; compounded preparations are used that limit the doses, based on NSAID toxicity. The third step is the administration of strong opioids such as morphine orally, rectally, intravenously, or transdermally. Significant side

effects include respiratory depression, pruritus, sedation, and constipation. Opiates have traditionally been underused for acute pain relief, because of dependency fears. However, opioids are now strongly recommended in order to restore patients' function and minimize the deleterious homeostatic effects of pain, based on studies showing improved recovery and an addiction risk of less than 1%. Coadministration of adjuvants that reduce the development of tolerance (the pharmacological need for larger doses) has enabled opiate usage in bedside pumps for patient-controlled analgesia. Nonetheless, regular outpatient administration must still be carefully monitored because of dependency, tolerance, and potential abuse.

Pain interacts with many aspects of homeostasis, including immune system function. Illness and immune challenges induce hyperalgesia. Proinflammatory cytokines have direct effects on nociceptors and central neurons. Increased headache and migraine incidence occurs during systemic immune responses. An exaggerated pain state can be regarded as a natural concomitant of neurally organized, immune-activated illness behavior (which includes fever, decreased activity, decreased food and water intake, and increased sleep), consistent with the role of C fibers and pain processing in metabo-reception and homeostasis.

However, pain itself can have strong negative effects on health. Incessant pain, like stress or surgery, inhibits immune function, enhances tumor growth, and increases morbidity and mortality by sleep deprivation, loss of appetite, immobilization, cardiovascular stress, severe impairment of general health, and even suicide. Pain *can* kill. Unrelieved pain can produce a neural pain 'memory' that persists even when the organic cause is removed. Management of patients with intractable chronic pain presents a great challenge. Appropriate multidisciplinary and behavioral therapy for pain is an important aspect of clinical team management, including techniques of biofeedback, visualization, and relaxation, and if pain is complicated by anxiety or depression, psychiatric treatment. Pain reports can be influenced by psychological factors such as culture, context, emotional status, experience, social support, and beliefs. Attention and expectation have strong effects on pain; in the laboratory, anticipation of a noxious stimulus can cause a warm stimulus to feel painfully hot. Neural blocks, deep brain stimulation, and neurosurgical lesions may be used in particular patients if pharmacologic and behavioral therapies prove insufficient.

Chronic, pathological pain can also occur through damage to the nervous system itself. The brain itself is insensate, but injury to peripheral nerves can result in neuropathic pain. Peripheral nerve damage (e.g. carpal tunnel syndrome) or pathogenic processes (e.g. diabetic neuropathy, postherpetic neuralgia) can produce abnormal neural activity, molecular phenotype changes, or anatomical sprouting. For example, novel sensitivity of nociceptors to circulating adrenaline (epinephrine) and sympathetic efferent activity may underlie reflex sympathetic dystrophy; i.e. complex regional pain syndrome or causalgia. Sympatholysis, systemic lignocaine (lidocaine) or orally administered gabapentin may relieve neuropathic pain by interfering with these mechanisms. Deafferentation pain occurring with a brachial plexus root evulsion can be effectively treated by a lesion of the dorsal root entry zone that eliminates hyperactivity in the spinal cord. Phantom limb pain following amputation, particularly if there was significant preoperative pain, may involve changes in both peripheral nerve and brain function.

Intractable pain can also result from damage to the brain. In Wallenberg syndrome (anesthesia dolorosa), an infarct in the caudal medulla produces loss of pain and temperature sensations in the ipsilateral face (due to damage to the trigeminal dorsal horn) and the contralateral body (due to interruption of the spinothalamic tract), and paradoxical burning pain arises in these regions. Similarly, a spinal lesion (e.g. in multiple sclerosis) or stroke-induced damage in the posterolateral thalamus or the insular cortex can result in a central pain syndrome, if normal temperature and pain sensation is eliminated in the contralateral body, in which paradoxical ongoing burning pain arises which is exacerbated by cooling or touch. Both syndromes may result from imbalanced homeostatic integration in the brain, due to the loss of sensation and disruption of endogenous analgesia mechanisms. These syndromes are unresponsive to opiates, but tricyclic antidepressant (amitriptyline) or antiepileptic (carbamazepine) agents can be efficacious.

Further Reading

Besson JM, Guilbaud G and Ollat H (eds) (1995) *Forebrain Areas Involved in Pain Processing*. Paris, France: John Libbey Eurotext.

Belmonte C and Cervero F (1996) *Neurobiology of Nociceptors*. Oxford, UK: Oxford University Press.

Bonica JJ (ed.) (1990) *The Management of Pain*. Philadelphia, PA: Lea & Febiger.

Fields HL (1987) *Pain*. New York, NY: McGraw-Hill.

Perl ER (1984) Pain and nociception. In: Darian-Smith I (ed.) *Handbook of Physiology. 1: The Nervous System*.

Vol. 3: *Sensory Processes*, pp. 915–975. Bethesda, MD: American Physiological Society.
Price DD (1988) *Psychological and Neural Mechanisms of Pain*. New York, NY: Raven Press.
Wall PD and Melzack R (eds) (1999) *Textbook of Pain*, 4th edn. Edinburgh, UK: Churchill Livingstone.
Willis WD (ed.) (1992) *Hyperalgesia and Allodynia*. New York, NY: Raven Press.
Yaksh TL, Lynch C, Zapol WM, *et al*. (1998) *Anesthesia: Biologic Foundations*. Philadelphia, PA: Lippincott-Raven.

Parietal Cortex

Introductory article

Jody C Culham, University of Western Ontario, London, Ontario, Canada

CONTENTS

Parietal cortex consists of the cortical surface of the parietal lobes. It contains numerous specialized subregions involved in encoding bodily sensations and integrating sensory information to perform motor actions.

INTRODUCTION

Parietal cortex forms approximately 20% of the cerebral cortex, the crumpled outer layer of the mammalian brain that performs higher intellectual functions. The parietal lobes form the top, back region of each of the two cerebral hemispheres (Figure 1). Parietal cortex can be divided into two main regions: the somatosensory cortex and posterior parietal cortex (PPC). The somatosensory cortex, in the frontward section of parietal cortex, receives input about touch from the opposite side of the body. Posterior parietal cortex, behind the somatosensory area, is 'association cortex' which integrates information from many sensory inputs and performs higher-level processing. The posterior parietal cortex is well situated to take input from sensory regions (including visual areas in the occipital lobe, auditory areas in the temporal lobe and tactile areas in parietal somatosensory cortex) and to provide output to premotor and motor regions within the frontal lobe. (*See* **Somesthesis, Neural Basis of**; **Motor Areas of the Cerebral Cortex**; **Frontal Cortex**)

To review the functions of parietal cortex, consider the stages necessary in using sensory input to guide motor output in everyday tasks such as taking a sip of coffee from a cup. The cluttered spatial environment, such as an office desk, must be encoded and whittled down to focus on the target object, the cup in this example. Once the cup has been located, the eyes are directed to look at it and the hand reaches towards it with the wrist oriented vertically and the fingers opened to an appropriate size. The initial visual processing codes the cup with respect to its location on the retina, but the motor output requires that the hand move relative to its location on the body. When the hand contacts the cup, touch receptors on the fingers send sensory information that is used to determine the cup's shape, size, orientation, weight and material. Sensory feedback is also used to optimize the grip on the cup and guide it to the mouth using the appropriate force, more for a full cup than an empty one. This seemingly simple everyday task suggests the range of functions the parietal cortex performs in sensorimotor processing – spatial coding, attention, visuomotor control, coordinate transformation and tactile exploration.

ANATOMY

Figure 1 shows the boundaries of the parietal lobes in the human brain. The parietal lobes are divided from the frontal lobe by the central sulcus, from the occipital lobes by the parietooccipital fissure (and its extrapolation to the lateral side), and from the temporal lobes by an arbitrary boundary. Between the central sulcus and the postcentral sulcus, within the postcentral gyrus, lies primary somatosensory cortex (S I). The somatosensory strip in each hemisphere receives information about touch sensations from the opposite side of the body. Below S I lies the secondary somatosensory cortex (S II). The posterior parietal cortex lies behind somatosensory cortex and is divided by the intraparietal sulcus (IPS) into the superior parietal lobule (SPL) and inferior parietal lobule (IPL). The IPL consists of two main subregions, the supramarginal gyrus and the angular gyrus. The continuation of the SPL onto the medial side is called the precuneus.

During the course of evolution parietal cortex first emerged in mammals, in conjunction with their greater reliance on visual and auditory information than on the sense of smell that predominates in lower animals. Parietal regions are limited in animals that use their limbs for locomotion (e.g.

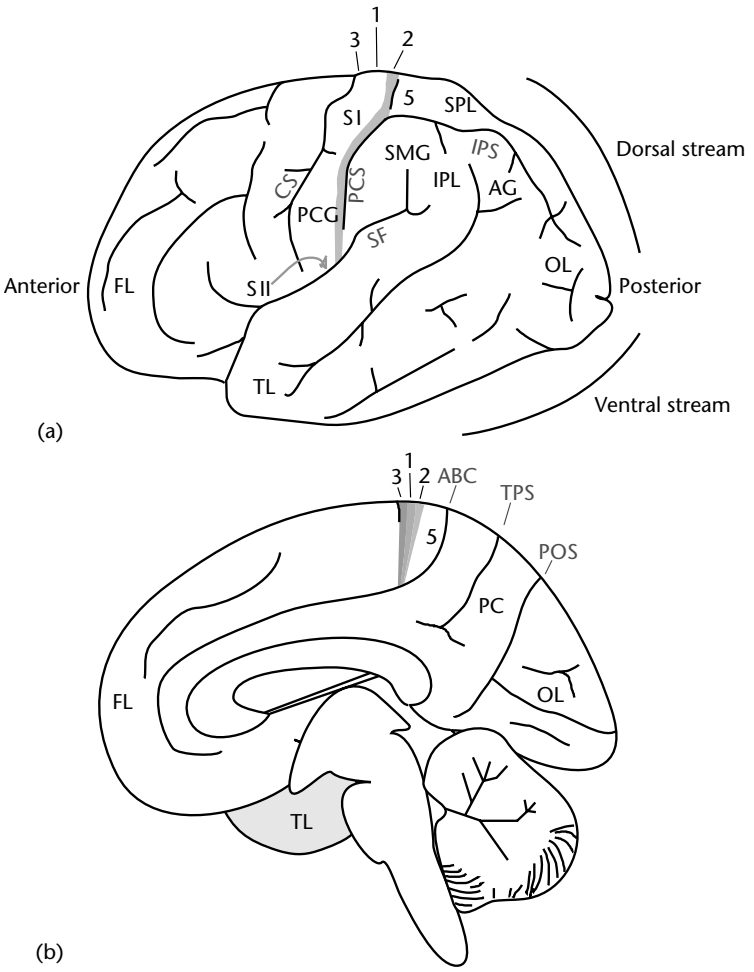


Figure 1. Regions and boundaries of human parietal cortex seen from (a) lateral and (b) medial views. Primary somatosensory cortex (S I, Brodmann's areas 3, 1 and 2) is shown between the central sulcus (CS) and postcentral sulcus (PCS) in the postcentral gyrus (PCG). Secondary somatosensory cortex (S II) is buried within the parietal operculum above the sylvian fissure (SF), as indicated by the arrow. The posterior parietal cortex is divided by the intraparietal sulcus (IPS) into the superior parietal lobule (SPL) and inferior parietal lobule (IPL). The IPL is composed of the supramarginal gyrus (SMG) and angular gyrus (AG). The continuation of the SPL on the medial side is called the precuneus (PC), which lies between the parietooccipital sulcus (POS) and the ascending band of the cingulate sulcus (ABC), and contains the transverse parietal sulcus (TPS). FL, frontal lobe; OL, occipital lobe; TL, temporal lobe.

bears), but are more conspicuous in arboreal mammals (e.g. tree shrews and lemurs), probably related to the additional sensorimotor processing required for locomotion through the trees. Parietal cortex is larger in primates (especially monkeys and chimpanzees) than in lower mammals and larger still in early hominins and modern humans. The surface area of parietal cortex is 20 times larger in the human than in the macaque monkey (in contrast to visual cortex which is twice as large, or temporal cortex which is nine times as large). One suggestion is that the evolutionary growth of parietal cortex (particularly PPC) occurred in conjunction with increased use of the forelimbs (or other limbs such as the monkey's tail or elephant's trunk) for action and interaction with objects, including in humans the increased use of tools. (*See* **Human Evolution**; **Human Brain, Evolution of the**)

Although human parietal cortex has not been extensively mapped, much has been learned from the study of closely related species, particularly monkeys. The best-studied nonhuman species to date has been the macaque monkey (Figure 2). However, it can sometimes be difficult to determine which regions of the monkey brain, if any, are comparable to those in the human. Interspecies comparisons are made difficult by the differences in techniques available for studying monkeys and

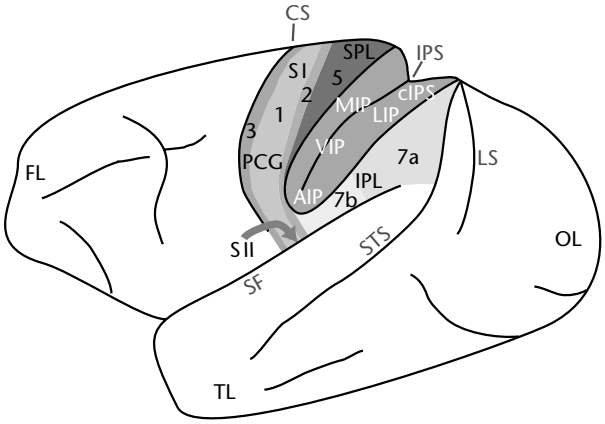


Figure 2. Regions and boundaries of parietal cortex in the macaque monkey, the most intensively studied non-human primate species, seen from a lateral view. As in the human, somatosensory areas 3, 1 and 2, forming SI, lie behind the central sulcus (CS). Although macaques also have superior parietal lobule (SPL) and inferior parietal lobule (IPL) regions, their comparability to human SPL and IPL remains questionable. Monkey SPL contains only area 5, whereas the IPL comprises areas 7a and 7b. Numerous highly specialized areas exist within the macaque intraparietal sulcus (IPS), which is indicated here by gray shading to indicate that the sulcus has been opened up to expose the areas lining its banks. These areas include the caudal IPS (cIPS), the lateral intraparietal area (LIP), the medial intraparietal area (MIP), the ventral intraparietal area (VIP) and the anterior intraparietal area (AIP). The sylvian fissure (SF), superior temporal sulcus (STS) and lunate sulcus (LS) are also shown. FL, frontal lobe; OL, occipital lobe; TL, temporal lobe.

humans. In the monkey, neurophysiologists have used electrode recordings of single neurons to study the function of numerous highly specialized parietal regions, particularly within the IPS (Figure 2). Data from humans has historically come from studies of neuropsychological patients with lesions to regions of parietal cortex. The introduction of brain imaging techniques such as functional magnetic resonance imaging (fMRI) has made possible the detailed study of parietal function in normal humans. More recently scientists have begun to perform fMRI experiments on monkeys, facilitating the possibility of comparing brain activation patterns between the two species. To date, many of the comparisons are based on cytoarchitectonic mapping, using microscopic examinations of the layers of cortex to distinguish subregions. One of the more popular cytoarchitecture schemes was proposed by Brodmann in 1905 for both the human and the monkey. Primary somatosensory cortex includes Brodmann's areas 3, 1 and 2 which appear comparable between the two species. However, in PPC the correspondence between Brodmann's areas (e.g. area 7) in the two species is less clear and is somewhat controversial. (*See* **Neuroimaging**; **Single Neuron Recording**)

FUNCTIONS OF POSTERIOR PARIETAL CORTEX

Theorists have suggested that the processing of visual information proceeds through two distinct pathways, a dorsal stream from early stages of visual processing in visual cortex to PPC and a ventral stream from visual to temporal cortex. Although there is general agreement that the ventral stream is involved in the processing of visual information for the purpose of object recognition, the function of the dorsal stream has been debated. In 1982, Ungerleider and Mishkin suggested that the overarching function of the dorsal stream, including PPC, was the processing of spatial locations. They suggested that the PPC analyzes 'where' objects are in the world, in contrast to the ventral stream, including regions of temporal cortex, which analyzes 'what' the objects are in order to recognize and remember them. However, in 1992 Goodale and Milner argued that the PPC might instead analyze 'how' actions can be accomplished. They suggested that one reason for the division of labor between the two visual systems arose from the necessity of coding space in different frameworks. The ventral or temporal stream represents the world in object-centered coordinates regardless of its location or orientation. For example, a coffee cup must be recognizable regardless of whether the handle is on the left or right or whether it is 20 cm or 200 cm away. In contrast, the dorsal or parietal stream represents the world in egocentric, or person-centered, coordinates. To grasp the coffee cup requires a computation of how to guide the hand from where it is to the location of the handle and how to orient and shape the hand appropriately to hold the handle. Although the majority of theory and research on dorsal versus ventral streams has focused on the processing of visual information, there are suggestions that a similar dichotomy may also exist for the other spatial senses – hearing and touch. (*See* **What and Where/How Systems**)

The profound deficits of patients with parietal damage suggest the fundamental role these functions have in everyday behavior. For example, people with extensive damage to both parietal lobes often demonstrate the phenomena characteristic of Balint syndrome, including an inability to

notice more than one object at a time (simultanagnosia); 'sticky fixation', where the eyes do not normally explore space (gaze apraxia); highly inaccurate movements towards objects (optic ataxia), such as pouring water from a jug but missing the glass; and spatial disorientation, such as being unable to find one's bedroom or to find the bed within the room.

Spatial Processing

One of the fundamental functions of parietal cortex is the processing of information about where things are in the world. Following damage to parietal cortex, humans or monkeys demonstrate gross deficits in the perception of space. For example, monkeys with PPC lesions fail at a task that requires them to discriminate locations, although they succeed at tasks that require them to discriminate objects. Humans with parietal lesions also show profound difficulties in conceptualizing space: for example, patients with right hemisphere parietal damage demonstrate great difficulty in copying two-dimensional drawings or three-dimensional block arrangements (constructional apraxia). Neuroimaging studies have also found enhanced brain activity in the parietal lobes when normal participants perform a task that requires spatial judgment (versus an object recognition task).

The 'space' represented in PPC may not be limited to the real world. Parietal cortex also plays a part in spatial mental imagery (e.g. imagining the locations of shops on a street) and mental rotation (e.g. imagining whether, when rotated to another orientation, a piece will fit into a slot on a piece of furniture you are trying to assemble). People with parietal damage may show deficits in these functions, and neuroimaging studies have reported increased parietal activation for such tasks. Parietal 'space' may be even more abstract, as with numerical representations in mathematics that appear to be encoded in the angular gyrus. (*See* **Mental Rotation**; **Acalculia**)

Spatial Attention

Parietal processing of information at a particular spatial location depends on the observer's attention. In any visual scene, certain regions of space or objects within them may be particularly important. For example, one might be searching for a particular object, be it the coffee cup or the children's book character Waldo. The posterior parietal cortex is essential for directing attention to relevant items so that they can be processed more effectively. Damage to this area can lead to severe problems in the way people process space, particularly when they must attend to a particular region or objects within it. (*See* **Attention**; **Attention, Neural Basis of**; **Spatial Attention, Neural Basis of**)

People with unilateral PPC lesions can develop a bias to attend to one side. In one syndrome, extinction, the patient fails to notice a stimulus on the side opposite to the lesion when it is presented at the same time as another stimulus. In the classic clinical test for this condition the doctor faces the patient with both arms extended horizontally and asks the patient to report whether the doctor is moving the fingers of one hand or of both. A patient with extinction arising from right hemisphere damage would correctly report when the doctor moved the fingers of either the left hand or the right hand alone. However, when both fingers were moved simultaneously, the patient would report seeing only the movement of the doctor's left fingers (on the patient's right side), oblivious that the fingers of the doctor's right hand (on the patient's left side) had also moved. Unilateral neglect is a related but dissociable syndrome that is even more striking. It occurs most often in patients with damage to the right hemisphere. These patients show an inability to attend to the left side of space: they may ignore the left side of the world, the left side of objects, and even the left side of their own bodies or the left side of an imagined object or scene. Neglect has traditionally been attributed to damage to the right inferior parietal lobule (supramarginal and angular gyri); however, more recent evidence suggests that the focus may instead be in the right superior temporal cortex.

People with PPC damage may show problems in switching attention from one location to another. In one task, participants are given a cue that a target will appear to one side of where they are looking. If the cue is valid and the target appears on the cued side, people with parietal damage are nearly as fast as normal participants. However, if the cue is invalid and the target appears on the opposite side, they are much slower to respond. Thus it is thought that damage to the parietal cortex leads to problems in disengaging and shifting attention from an expected location. People with such damage may also show problems in attending to more than one item at a time (simultanagnosia). For example, when looking at an array of objects on a desk, a patient might report seeing only the pencil, then only the stapler. In such cases, attention becomes locked on one item to the exclusion of others.

Visual attention can lead to enhanced firing of neurons in PPC. Higher firing rates in monkey area

7 neurons are found when the monkey keeps its eyes fixed straight ahead but must pay attention to a target in another (peripheral) location regardless of whether or not an eye movement is then made to the target. One unresolved debate concerns whether attention and eye movements are separable mechanisms. Certainly, one can pay attention to something without making an eye movement towards it, but does that simply involve planning an eye movement that is never executed? Neurophysiological studies of parietal cortex suggest that the processing of a target may depend not so much on attention towards it as the 'intention' to interact with the target via the eyes (by a saccadic eye movement) or the hand (by a reaching movement).

Human neuroimaging studies have reported robust PPC activation (in the IPS, SPL and temporoparietal junction, as well as frontal and subcortical structures) during the performance of attentional tasks. A similar network of areas is activated during eye movements (saccades). The PPC is thought to be the source of where attention is directed. Once attention has been directed to a particular location, the relevant features of that item (e.g. its shape, motion or color) can be processed by sites in the temporal lobe that are specialized for that feature. Although the majority of attention studies have focused on vision, other varieties of attention also rely on parietal lobe function, including auditory attention, tactile attention, and motor attention.

Sensorimotor Control

The PPC as a whole has a key role in using sensory information to guide motor actions. Furthermore, sensorimotor control is specialized within PPC in that numerous subregions are each tuned to a particular action such as moving the eyes towards a target, reaching towards it, or grasping it. Given that the inputs to PPC association cortex originate in different coordinates (retina-centered for vision, body-centered for touch, head-centered for hearing), the guidance of action requires the computation and transformation of a variety of coordinate frames. For example, by combining what the retina sees with information about where the eyes are looking allows the system to calculate where an object is with respect to the head. (*See* **Action**)

Eye movements

The eyes are reflexively drawn towards a target that suddenly appears – such as the flashing lights of an ambulance appearing in a rear-view mirror. Similarly, when the attentional system flags a potential item that one is seeking, such as the red and white stripes on the shirt worn by Waldo, the eyes are then directed to it so that it can be processed fully to determine whether it is indeed Waldo or simply a distractor. The PPC plays a crucial role in guiding not only attention, as described above, but also eye movements. (*See* **Eye Movements**)

Bilateral PPC lesions can cause problems with movements of the eyes. Although people with these lesions can move their eyes to a target that suddenly appears, if they are asked to look voluntarily at a target or to move their eyes to search for a target, they show pronounced problems including slow and abnormal eye movements (gaze apraxia or 'sticky fixation'). Other eye movement problems are also reported with parietal damage, including deficits in keeping the eyes fixated on a target, tracking a smoothly moving target with the eyes (smooth pursuit), or focusing the eyes at the appropriate depth (accommodation and convergence). Even unilateral parietal lesions can lead to a slowness in making eye movements towards the opposite visual field.

One PPC area in particular, within the lateral intraparietal sulcus (area LIP), represents space as it is explored by the eyes. Neurons in this area respond when a monkey makes an eye movement to a visual target. These cells also respond without an eye movement if the monkey pays attention to a visual target or remembers its location after it disappears. Although LIP neurons code space in an eye-centered (retinotopic) framework, their responses are more complex than those in simple visual areas. Neuroimaging studies in humans have also found activation for eye movements in an IPS area that probably corresponds to LIP.

Arm and hand movements

When one goes to pick up an object, such as a coffee cup, the hand must be transported to the location of the cup and then the wrist and fingers must be oriented appropriately to grasp it. The process of reaching to a target location, and then grasping it by preshaping the hand, depends critically on regions within the PPC. Lesions in these regions, particularly in the anterior IPS and area 5, cause problems with movements of the hands towards visible targets in actions such as reaching and grasping (optic ataxia). People with parietal lesions may show problems in reaching towards visible objects even though movements without visual guidance are made normally. For example, people with ataxia can quickly and accurately touch parts of their own bodies. This suggests the problem is

not purely a motor one, but a visuomotor one. Although misreaching (optic ataxia) and eye movement disorders (gaze apraxia) both occur in Balint syndrome, there are cases of patients who can guide their eyes but not their hands to a target, indicating the misreaching is not merely a symptom of more general spatial disorientation.

An area in the medial IPS (area MIP) responds when monkeys make reaching movements, particularly those guided by vision. Cells in this area respond best to nearby stimuli, close enough to be reached with the arms. Interestingly, these cells respond not only to the space which the arm can reach, but when a tool is used, also to the space around the tool.

In some ataxic patients, the deficit can be specific to grasping with intact reaching movements. People with grasping deficits can guide their hand to the location of a target, but are unable to process object shape and orientation to preshape the hand appropriately. They will fail to pick up an object correctly because they do not open the hand to the right size or place the fingers on the most stable points for grasping, as is normally done. The site of damage in patients with grasping disorders is the anterior IPS, a region that is activated in brain imaging experiments during grasping. The anterior IPS region in humans appears to correspond to an anterior intraparietal (AIP) region in monkeys which has a key role in using vision to guide grasping. Monkey AIP contains neurons that respond when the monkey sees an object and grasps it. Some cells respond to sight of the object alone, others respond to the act of grasping alone (even in the dark), and many respond to the combination of seeing the object and grasping it. If this area is inactivated by injection of an inhibitory drug, the monkey can still reach towards an object, but loses the ability to shape the hand to grasp the object correctly, behaving much like the human patients with lesions to the anterior IPS. The AIP receives input from the caudal intraparietal sulcus (cIPS), which contains neurons that encode object properties such as surface orientation.

Parietal arm movement disorders are not limited to direct actions on real objects. Apraxia is the inability to make skilled motor movements in the absence of pure sensory, motor or cognitive deficits. It typically results from IPL damage opposite to the dominant hand, and can take several forms. The deficits can be relatively stronger for pantomiming actions (ideomotor apraxia), imitating actions (conduction apraxia), using tools (conceptual apraxia), or performing a sequence of movements (ideational apraxia). (*See* **Apraxia**)

Head-centered motion

The ventral part of the monkey IPS contains a region (area VIP) that responds well to motion in the visual, tactile or auditory domains. This area codes motion towards the head, regardless of where the eyes are looking, and has neurons that responds to touch to a particular part of the face or to motion towards that part. This region represents space centered around the head, especially the mouth, and is particularly responsive to motion in ultra-near space. Many VIP neurons respond to optic flow (mainly expansion but also contraction) like the motion one would experience when moving through the world. Brain imaging experiments have identified a human IPS area with similar properties: it responds to motion in different modalities (visual, auditory, tactile) and it responds both to motion towards the head and touching of the face. The VIP area may play an important part in head movement and perhaps in guiding the hand towards the head in actions such as feeding or grooming.

FUNCTIONS OF SOMATOSENSORY REGIONS

Primary somatosensory cortex (SI, Brodmann's areas 3, 1 and 2) lies along the postcentral gyrus and contains neurons tuned to somatosensory (touch) stimulation on the opposite side of the body. SI contains a tactile representation of the human body, first discovered by Wilder Penfield and colleagues during surgery on conscious epilepsy patients. When SI was electrically stimulated in one hemisphere, patients reported tactile sensations such as numbness, tingling or pressure on a specific body part on the other side of the body. This enabled Penfield to determine maps of the somatosensory homunculus – the 'little man' represented on the somatosensory strip along the postcentral gyrus. The somatosensory representation is distorted. First, it is largely upside down: the foot is represented at the top of the strip, with the legs, torso, arm, hand and head subsequently below it. Second, the amount of cortex devoted to each body part differs based on the resolution of that part – the hands and face have relatively large representations compared with the torso, for example. The representation in SI is quite plastic. For example, following amputation, the representation of the amputated limb is taken over by adjacent body part representations. Output from SI goes to the secondary somatosensory cortex (SII, in the upper bank of the sylvian fissure) and PPC association areas (5 and 7 in the monkey), which elaborate

further on complex properties such as direction selectivity and are more susceptible to cognitive factors such as attention. Vestibular input from the inner ear provides information about head movement to several regions of parietal cortex. (*See* **Penfield, Wilder**)

People with damage to somatosensory regions (postcentral gyrus) lose the sense of touch, temperature, pain and muscle sense (proprioception) from the opposite side of the body (hemianesthesia). Their finger movements may be clumsy because of the loss of feedback regarding finger position. Even when the sense of touch itself is intact, these people may be unable to recognize objects felt by touch, even though they can recognize them visually (astereognosis). Damage to secondary somatosensory cortex can lead to somatosensory extinction, an inability to register two simultaneous touches even though either one alone would be perceived.

Some of the more unusual parietal lesion disorders are disturbances in the way people represent the knowledge of their bodies and the relationship of their bodies to the environment. These asomatognosias can include denial of or indifference to bodily illnesses (anosognosia and anosodiaphoria, respectively, typically on the left side of the body following right PPC lesions), a lack of aversion to pain (asymbolia to pain, possibly resulting from a disconnection between parietal sensation regions and emotional centers of the brain), or an inability to localize and name body parts (autotopagnosia, usually resulting from left parietal damage). In one form of autotopagnosia known as finger agnosia, patients are unable to identify or distinguish their own fingers or those of another person. Patients with right–left disorientation cannot differentiate between left and right on their own bodies or others' bodies. (*See* **Anosognosia**)

CONCLUSION

The parietal cortex is a somewhat arbitrary territory with few clear-cut boundaries, making it difficult to attribute a common function to the region. However, the continuity of parietal cortex with adjacent cortex in the other lobes is consistent with its associative sensorimotor role. That is, parietal cortex receives sensory input from somatosensory parietal cortex, occipital visual cortex and temporal auditory cortex; it encodes spatial locations in a variety of coordinates, particularly for attended targets that are behaviorally relevant; it generates specific instructions to interact with the world using the eye, arm, hand, mouth and body that are then sent to the frontal premotor cortex for motor programming; and it uses somatosensory information about the resulting tactile responses and proprioceptive feedback about the positions of the joints, along with visual feedback, to fine-tune motor actions during visually guided actions and exploratory movements.

Further Reading

Andersen RA (1989) Visual and eye movement functions of the posterior parietal cortex. *Annual Review of Neuroscience* **12**: 377–403.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–349.

Critchley M (1953) *The Parietal Lobes*. London: Arnold.

Culham JC and Kanwisher NG (2001) Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion in Neurobiology* **11**: 157–163.

De Renzi E (1982) *Disorders of Space Exploration and Cognition*. Chichester, UK: John Wiley.

Hyvärinen J (1982) *The Parietal Cortex of Monkey and Man*. Berlin: Springer.

Kolb B and Whishaw IQ (1996) *Fundamentals of Human Neuropsychology*, chap. 12, pp. 265–286. New York, NY: Freeman.

Mountcastle VB (1995) The parietal system and some higher brain functions. *Cerebral Cortex* **5**: 377–390.

Rizzolatti G, Fogassi L and Gallese V (1997) Parietal cortex: from sight to action. *Current Opinion in Neurobiology* **7**: 562–567.

Sakata H, Taira M, Kusunoki M, Murata A and Tanaka Y (1997) The TINS Lecture. The parietal association cortex in depth perception and visual control of hand action. *Trends in Neurosciences* **20**: 350–357.

Thier P and Karnath HO (1997) *Parietal Lobe Contributions to Orientation in 3D Space*. Berlin: Springer.

Parkinson Disease

Intermediate article

Brian C Rakitin, Columbia University, New York, New York, USA
Yaakov Stern, Columbia University, New York, New York, USA

CONTENTS

Parkinson disease is a neurodegenerative disorder resulting in the loss of dopamine-producing neurons and disruption of basal ganglia-dependent motor and cognitive function.

INTRODUCTION

The study of Parkinson disease (PD) provides a unique opportunity to investigate functional aspects of dopamine regulation and cortical–basal ganglia–thalamic anatomy. The disease is associated with a variety of cognitive changes; in a large proportion of people, these are quite subtle. In addition, a subset of sufferers develop a dementia that differs from that seen in Alzheimer disease.

To date, cognitive research in PD has focused on three issues. First, what types of dementia are associated with PD? Second, what specific cognitive changes are found in nondemented patients? Third, how can the cognitive deficits be related to the affected neurotransmitter and anatomical systems? These questions are considered following a description of the disease and its pathology.

DEFINITION, DIAGNOSIS, AND FREQUENCY

A cluster of motor signs known as parkinsonism characterizes Parkinson disease. The three primary signs of parkinsonism are tremor, rigidity, and akinesia (Agid, 1991). The tremor is seen at rest, has an approximate rate of 9 cycles per second, and is most evident in the distal portion of the limbs as a 'pill-rolling' motion of the thumb and forefinger. Rigidity refers to stiffening of the limbs evident during external manipulation. Akinesia is difficulty in initiating movements, sustaining repetitive movements, and performing simultaneous movements, as well as a general paucity of movement ranging from slight to complete immobility. This difficulty is most evident when patients attempt to avoid obstacles or change directions, and is often described as 'freezing'. Akinesia can be attenuated in the presence of external stimulation. In some classification schemes the third primary symptom is bradykinesia, or slowness in the execution of movements, and freezing is considered a secondary symptom (Fahn, 1995).

One common set of criteria for the diagnosis of PD requires the presence of either resting tremor or bradykinesia in addition to one other feature such as rigidity, flexed posture, loss of postural reflexes, or freezing (Fahn, 1995). A major difficulty in diagnosis is distinguishing between PD and other parkinsonian syndromes. These syndromes are divided into four classes according to their etiology (Fahn, 1995). Parkinson disease is classified as idiopathic parkinsonism, an allusion to the absence of a firmly established cause, either genetic or environmental. The other classes are symptomatic parkinsonism, Parkinson-plus syndromes, and heterodegenerative disorders. Nonresponsiveness to levodopa usually excludes a diagnosis of PD, but responsiveness is not sufficient to rule out certain non-PD syndromes: symptomatic syndromes related to methylphenyl tetrahydropyridine (MPTP) toxicity and encephalitis, and early-stage Parkinson-plus syndromes like striatonigral degeneration and olivopontocerebellar atrophy, will also respond to levodopa. Differential diagnosis therefore relies on additional features, such as a syndrome presenting with a known cause (i.e. symptomatic syndromes) or with unique symptoms (e.g. the eye-movement difficulty associated with progressive supranuclear palsy).

The prevalence of PD is approximately 160 cases per 100 000 individuals. The incidence rate is about 20 per 100 000. Because PD is an age-related neurodegenerative disorder, both prevalence and incidence increase with age. Thus, at 70 years of age the prevalence is 550 per 100 000 and the incidence is 120 per 100 000. The mean age of onset is 55 years

with a range of 20–80 years. The disease is somewhat more common in men, with a male to female ratio of 3 : 2.

NEUROPATHOLOGY

The principal neuropathological change in PD is degeneration of neurons in the substantia nigra pars compacta (SNc), a pigmented midbrain nucleus that is the principal source of dopamine in the striatum. This results in dopamine depletion of the striatum (Agid, 1991; Fahn, 1995) and large alterations to the normal patterns of connectivity within the rest of the basal ganglia (Delong, 1990). Reduced striatal output downregulates the 'direct' pathway between the striatum and the internal segment of the globus pallidus, reducing inhibition of the latter structure. Reduced striatal output also upregulates the 'indirect' pathway that includes the external segment of the globus pallidus and the subthalamic nucleus, resulting in increased excitation of the globus pallidus. The net result is an increase in activity of inhibitory projections from the globus pallidus to the central medial, ventral lateral and ventral anterior thalamic nuclei, and a corresponding decrease in activity of excitatory thalamocortical projections. Since the globus pallidus projects in large part to cortical areas that have direct input to the striatum, the ultimate effect of reduced nigral dopamine output is reduced activity of the thalamocortical–basal ganglionic circuitry.

By the time the first signs of PD appear, as many as 70% of the dopamine-producing neurons are already dead (Agid *et al*., 1989; Agid, 1991; Fahn, 1995). This suggests that normal motor function of the basal ganglia can persist until a threshold is surpassed. This extensive cell death is associated with uneven dopamine depletion through the striatum. The putamen evidences extensive depletion early in disease progression. This portion of the striatum forms a circuit via the thalamus with the motor cortices, explaining the relatively early expression of motor symptoms in the disease course. Dopamine depletion then extends to the dorsal regions of the caudate nucleus, which has principal connections with the dorsal–lateral prefrontal cortex. This presumably leads to the expression of cognitive changes (see below), although certain cognitive symptoms are apparent early in the disease course as well.

Standard treatment with dopamine replacement therapy consisting of levodopa has dramatic effects on the motor symptoms of PD and leads to vast improvements in the quality of life of patients (Fahn, 1999). Long-term treatment of PD with levodopa often results in the appearance of new symptoms, however, including dyskinesias, involuntary spastic movements of the limbs and neck, and the on-off effect – increasingly rapid fluctuations in the efficacy of dopamine replacement therapy. Both dyskinesia and the on-off effect are thought to result from hypersensitization of striatal neurons.

While dopamine depletion is the most pervasive and evident cause of PD symptoms, other neurotransmitter systems are also affected (Agid *et al*., 1989). The locus ceruleus, the pigmented midbrain source of noradrenaline (norepinephrine), and the nucleus basalis of Meynert, a principal basal forebrain source of acetylcholine, both show extensive loss of neurotransmitter-producing neurons early in the disease course. This pattern of multiple system involvement may underlie the varied efficacy of dopamine replacement therapy in treatment of the numerous symptoms of PD. It also belies the notion that PD is a perfect model of SNc lesions, and thus limits attribution of cognitive deficits to the dopaminergic systems.

There are three major hypotheses concerning PD pathogenesis (Agid, 1991). The first is that cell death results from increased oxidative stress. Oxidative stress stems from three sources: first, lipid peroxidation is increased, resulting in an increase in free radicals apparently from an increase in detoxifying enzymes in the SNc; second, neuromelanin production in pigmented neurons results in free radical production; and third, dopamine degradation by monoamine oxidase B results in free radical production that may accelerate in remaining neurons as cell death proceeds. This is a particularly important aspect of the free radical hypothesis since dopamine replacement therapy drives dopamine production in remaining cells, potentially exacerbating free radical production.

The second hypothesis is that cell death results from impairment of cellular metabolism. Some reports indicate that complex I mitochondrial metabolism is impaired in the platelets of PD patients. This is an important but unconfirmed observation since methylphenyl pyridinium (MPP+), a metabolite of the MPTP toxin that produces a parkinsonian syndrome, impairs complex I activity.

The third hypothesis involves the role of Lewy bodies. These structures are found with greatest frequency in the dopamine-producing neurons of the SNc. While clearly pathological, the degree to which Lewy bodies are associated with cell death and dopamine depletion is in question owing to the fact that the Lewy bodies consist of normal cellular components. In PD, Lewy bodies are

characteristically found in subcortical areas. However, cortical Lewy bodies are found in PD patients with dementia with greater frequency than in nondemented patients, suggesting that they may play a part in the pathogenesis of that aspect of PD.

DEMENTIA

While PD has been traditionally considered the archetypal extrapyramidal motor disorder, it has long been evident that PD patients exhibit pervasive cognitive deficits. Some PD patients exhibit dementia, but most do not. Those who are not demented have patterns of cognitive impairment that suggest specific problems with implicit learning and memory, executive function, and interval timing.

Dementia is more prevalent in patients with PD than in control populations. Estimates of the frequency of dementia in PD range from 10% to 50%. Variation in these estimates stems at least in part from differences in assessment methods, criteria for labeling an individual as demented, control populations and other methodological issues. One major predictor of the incidence of PD dementia is the age of the patient, and age at onset of disease. Beyond this, a prospective incidence study found that the raw frequency of incident dementia in PD patients (about 19%) over a 3.5-year period was greater than that found in control subjects (about 15%) after accounting for differences in age, education and gender (Marder *et al*., 1995). Major predictors of dementia in PD patients included a high score on the Hamilton Rating Scale for Depression and more severe extrapyramidal motor signs.

The question naturally arises as to whether dementia in PD is due to concomitant Alzheimer disease (AD). Postmortem studies suggest that AD pathological change is indeed present in a proportion of PD dementia cases. However, the cognitive profiles of the dementia in these two diseases differ. One informative study compared demented and nondemented patients with PD to individuals with AD who were matched to the PD patients' overall level of intellectual function (Stern *et al*., 1993). Demented patients with PD performed worse than their matched AD controls on measures of verbal fluency and visuospatial tasks, whereas the AD patients performed worse on a memory task after delay. A similar study also found differences between probable AD and PD patients matched for moderate dementia. Patients with PD showed large number of intrusion errors, but not as many as the AD patients, and derived a much larger benefit from cues given during recall. It was therefore concluded that the explicit memory deficits in PD dementia are associated primarily with retrieval of previously stored items. These studies indicating that PD dementia differs from AD dementia suggest that, while some PD patients with dementia have comorbid AD, most are probably demented for other reasons. Likely candidates include damage to the cholinergic basal forebrain nuclei and spread of Lewy bodies into the cortex.

COGNITIVE DEFICITS IN PATIENTS WITHOUT DEMENTIA

Nondemented patients with PD exhibit a variety of cognitive changes that are evident on standard neuropsychological tests, as well as on more focused tests of specific cognitive functions. Neuropsychological tests consistently show PD-related changes such as reduced verbal fluency, increased perseverative errors on the Wisconsin Card Sorting Test and the Odd Man Out Set Shifting Test, slowed responses on the Trail-making Test (particularly, slow responses on Trails B, which requires set shifting), increased interference on the Stroop Test, impaired block construction and picture assembly (such as those used in the Weschler Adult Intelligence Scale version III-Revised: WAIS III-R), among many others. Overall, the neuropsychological profile of PD patients is one of impairment in executive function, or in the executive component of working memory. Memory changes have also been noted in nondemented patients with PD; such patients recall fewer items in list-learning tasks than do matched control subjects, but recognition memory remains intact. Studies of MPTP-induced parkinsonism, which primarily affects the dopamine system, suggest that at least some of these observed cognitive changes are related to loss of dopamine (Stern *et al*., 1990).

Many studies have investigated the degree to which procedural learning is impaired in PD. Patients show abnormally slow learning of sensorimotor tasks including serial reaction time, pursuit rotor, and mirror reading. The serial reaction time task has been particularly informative. Patients fail to show the typical speeding of reaction time when several repetitions of a sequence of targets are surreptitiously introduced into a spatial choice reaction time task (Ferraro *et al*., 1993). This effect persists even after those who could explicitly report part of the sequence are removed from the analysis. In addition, ability to explicitly report part of the sequence does not differ between PD patients and controls. It thus appears that implicit procedural

learning is especially impaired in PD and may be dependent on intact striatal function.

Another line of research has sought to elaborate the nature of PD executive dysfunction, exemplified by findings of increased perseverative errors on the Wisconsin Card Sorting Test. One possibility is that executive dysfunction in PD is primarily a deficit of set-switching, or the remapping of responses to stimuli within an experimental session. This hypothesis is supported by the observation that reaction times increase when set changes occur. For example, in one experiment (Hayes *et al*., 1998) the signal 'AA' indicated the key sequence 1-2-3-1-2-3 and 'BA' indicated 3-1-2-1-2-3. The reaction time to press the '1' key in the fourth position was longer following 'BA' than 'AA', indicating slowed processing when moving between two different sequences of three elements. Reaction times are similarly slow when PD patients switch from mapping color to keys to mapping shape to keys, compared with two sequential mappings of the same type. Importantly, the appearance and magnitude of both set-switching effects depended on the degree of motor impairment evident in the patient immediately before testing. More severely akinetic patients had the slowest set-switching reaction times. Since the motor symptoms of PD stem from dopamine depletion, these results suggest that set-switching is a dopamine-mediated cognitive function, and is dependent on intact basal ganglia.

The timing abilities of PD patients are abnormal in several ways. In one study, a single PD patient performed a repetitive tapping task with both hands. As with many PD patients, one hand was more impaired than the other. The inter-tap intervals produced by the more impaired hand were more variable. This was attributed to increased variability in a central pattern generator rather than to peripheral motor sources (Wing *et al*., 1984). Increased timing variability was attenuated by dopamine replacement therapy, suggesting that the deficit is related to dopamine depletion (O'Boyle *et al*., 1996). Similar deficits were evident when patients compared the duration of two tones, but not when loudness or pitch were compared, suggesting that the problem is specific to timing rather than with some cognitive process related only to motor control.

Two additional dopamine-dependent deficits are evident when patients with PD are asked to reproduce an interval by delaying a key-press for a period of time equal to a previously demonstrated standard (Malapani *et al*., 2001). When PD patients are tested on these tasks while on dopamine replacement therapy, performance is equal to or better than age-matched controls. Off medication, the same patients overestimate the shorter of two intervals (6 or 8 s in duration) and underestimate the longer of the two (17 or 21 s), a pattern referred to as the 'migration effect'. Patients show a second deficit if they learn either one or two intervals while off medication, but are tested on medication.

CONCLUSION

Parkinson disease is a complex and variable condition. Its signs and symptoms can vary widely from person to person, and its cause is still unknown. The variability in the expression of symptoms stems at least in part from the fact that many brain systems are affected by the disease, giving rise to a wide variety of motor and cognitive manifestations. Nonetheless, PD is one of the best naturally occurring opportunities for understanding the cognitive neuroscience of the basal ganglia and other dopamine-dependent brain structures.

Not all cognitive functions that have been identified as impaired in PD patients can be directly attributed to dopaminergic and basal ganglia dysfunction. Among those that can are timing and set switching. Uncertainty in attributing the neural origins of cognitive deficits in PD can be reduced by considering three major issues. First, many diseases other than idiopathic PD present the parkinsonian syndrome, so patients must be carefully screened to insure that a homogeneous sample is tested. Second, dementia is prevalent in PD, and probably represents the effects of additional neuropathology beyond that found in nondemented patients. Experiments aimed at investigation of specific cognitive function should carefully screen out demented PD patients. Third, many neurotransmitter systems are involved in PD. To insure that a cognitive deficit stems from dopamine depletion, patients should be compared on and off dopamine replacement therapy, or in different states of dopamine replacement efficacy.

References

Agid Y (1991) Parkinson's disease: pathophysiology. *Lancet* **337**: 1321–1324.

Agid Y, Cervera P, Hirsch E *et al*. (1989) Biochemistry of Parkinson's disease 28 years later: a critical review. *Movement Disorders* **4**(suppl. 1): S126–S144.

Delong MR (1990) Primate models of movement disorders of the basal ganglia. *Trends in Neuroscience* **13**(7): 281–285.

Fahn S (1995) Parkinson's disease. In: Rowland LP (ed.) *Merrit's Textbook of Neurology*, 9th edn. Baltimore, MD: Williams & Wilkins.

Fahn S (1999) Parkinson disease, the effect of levodopa, and the ELLDOPA trial. Earlier vs later L-. *Archives of Neurology* **56**: 529–535.

Ferraro FR, Balota DA and Connor LT (1993) Implicit memory and the formation of new associations in nondemented Parkinson's disease individuals and individuals with senile dementia of the Alzheimer type: a serial reaction time (SRT) investigation. *Brain and Cognition* **21**(2): 163–180.

Hayes AE, Davidson MC, Keele SW and Rafal RD (1998) Toward a functional analysis of the basal ganglia. *Journal of Cognitive Neuroscience* **10**(2): 178–198.

Malapani C, Deweer B and Gibbon J (2002) Separating storage from retrievel dysfunction of temporal memory in Parkinson's disease. *Journal of Cognitive Neuroscience* **14**(2): 311–322.

Marder K, Tang MX, Cote L, Stern Y and Mayeux R (1995) The frequency and associated risk factors for dementia in patients with Parkinson's disease. *Archives of Neurology* **52**: 695–701.

O'Boyle DJ, Freeman JS and Cody FJW (1996) The accuracy and precision of timing of self-paced, repetitive movements in subjects with Parkinson's disease. *Brain* **119**: 51–70.

Stern Y, Tetrud J, Martin WR, Kutner SJ and Langston JW (1990) Cognitive change following MPTP-exposure. *Neurology* **40**: 261–264.

Stern Y, Richards M, Sano M and Mayeux R (1993) Comparison of cognitive changes in patients with Alzheimer's and Parkinson's disease. *Archives of Neurology* **50**: 1040–1045.

Wing AM, Keele SW and Margolin DI (1984) Motor disorder and the timing of repetitive movements. In: Gibbon J and Allan L (eds) *Timing and Time Perception*, pp. 183–192. New York, NY: New York Academy of Sciences.

Further Reading

Brown RG and Marsden CD (1988) Internal versus external cues and the control of attention in Parkinson's disease. *Brain* **111** (Pt. 2): 323–345.

Brown RG and Marsden CD (1990) Cognitive function in Parkinson's disease: from description to theory. *Trends in Neurosciences* **13**(1): 21–29.

Davis CG, Williams AC, Markey SP *et al*. (1979) Chronic parkinsonism secondary to intravenous injection of mepedrine analogues. *Psychiatric Research* **1**: 249–254.

Flowers K and Robertson C (1985) The effect of Parkinson's disease on the ability to maintain a mental set. *Journal of Neurology, Neurosurgery, and Psychiatry* **48**: 517–529.

Pillon B, Dubois B, Lhermitte F and Agid Y (1986) Heterogeneity of cognitive impairment in progressive supranuclear palsy, Parkinson's disease and Alzheimer's disease. *Neurology* **36**(9): 1179–1185.

Pattern Vision, Neural Basis of

Introductory article

Daniel C Kiper, Institute of Neuroinformatics, University of Zurich, Switzerland
Matteo Carandini, Institute of Neuroinformatics, University of Zurich, Switzerland

CONTENTS

From the retina to the cerebral cortex, the early stages of the visual system decompose the visual scene into a number of relevant features, while discarding redundant information. Successive stages of the visual system perform an increasingly complex analysis of the visual scene.

INTRODUCTION

The visual system analyzes the two-dimensional images projected on the retinas, and from these images extracts informative features. These features allow the perception of form and texture, and enable the observer to distinguish and recognize objects and ultimately construct an internal three-dimensional representation of the world. The computations that extract the informative features constitute 'pattern vision'.

Pattern vision is achieved by a succession of neural operations performed by the early stages of the visual pathway. These operations have been studied extensively by recording the activity of individual neurons in cats and in primates such as macaque monkeys. Studies in these animals have revealed principles that probably apply to humans as well. Indeed, recent studies employing functional magnetic resonance imaging (fMRI) in humans have confirmed a tight relationship between neural activity and the perceptual aspects of pattern vision. (*See* **Single Neuron Recording**; **Vision: Early Psychological Processes**)

Anatomy of the Visual System

The early stages of visual processing are illustrated in Figure 1. The optics of the eye focus images onto the retina. Through a network of retinal cells, electrical signals generated by photoreceptors are transmitted to retinal ganglion cells. These cells

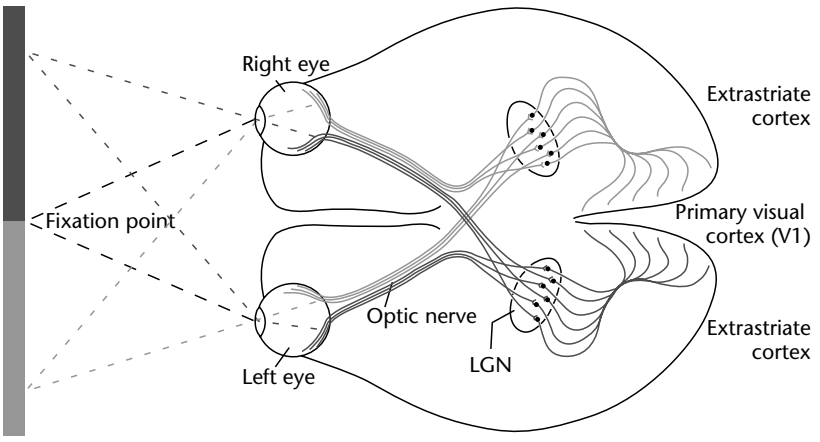


Figure 1. [*Figure is also reproduced in color section*.] Basic organization of the visual pathways. Images originating in the left and right visual hemifields are projected onto opposite parts of each retina. Retinal ganglion cells connect through the optic nerve to the lateral geniculate nucleus (LGN). Neurons in the LGN then project to the primary visual cortex. The visual information is further analyzed in an array of subsequent visual areas. Note that the left visual hemifield is analyzed in the right cerebral hemisphere, and vice versa.

produce spikes of electrical activity, which are carried by the optic nerve to the lateral geniculate nucleus (LGN). The neurons of the LGN in turn send their own spikes to the cerebral cortex, in a large area called the primary visual cortex, or area V1. Area V1 is a key center of visual processing, whose outputs are sent to numerous other visual cortical areas and to the rest of the cortex. (*See* **Vision: Early Psychological Processes**; **Occipital Cortex**; **What and Where/How Systems**)

PATTERN VISION AS IMAGE FILTERING

Neurons in the visual system respond to the distribution of light presented in a particular region of the visual field – the neuron's 'receptive field'. The study of the receptive fields of visual neurons is fundamental to our understanding of visual function, and provides the basis for the current knowledge about pattern vision. In particular, receptive fields can be seen as filters through which cells view the visual scenes; filtering is a well-known concept in engineering and computer science. Image filtering by the receptive fields in the early stages of the visual system is illustrated in Figure 2. Starting with a visual stimulus (a), this follows the operation of successive stages in the retina (b, c) and in the primary visual cortex (d, e). The operation of each stage is illustrated through a 'neural image', a map of the responses that would be obtained if an array of identical cells were looking at the physical image in all possible positions. In a neural image, gray indicates no response, whereas white and black indicate higher and lower responses. (*See* **Vision: Early Psychological Processes**)

Image Filtering in the Retina

Photoreceptors in the retina have small receptive fields (Figure 2(b), lower panel), and transform light into electricity. The gain of this transformation depends on the prevalent light conditions, a phenomenon known as 'light adaptation'. Thanks to light adaptation, rather than signaling absolute light intensity, photoreceptors encode the strength of light relative to the average in the recent past in that local region of the retina. This relative measure of light is called 'contrast', and it entails discarding information about absolute light levels. The advantage of computing contrast is illustrated in the

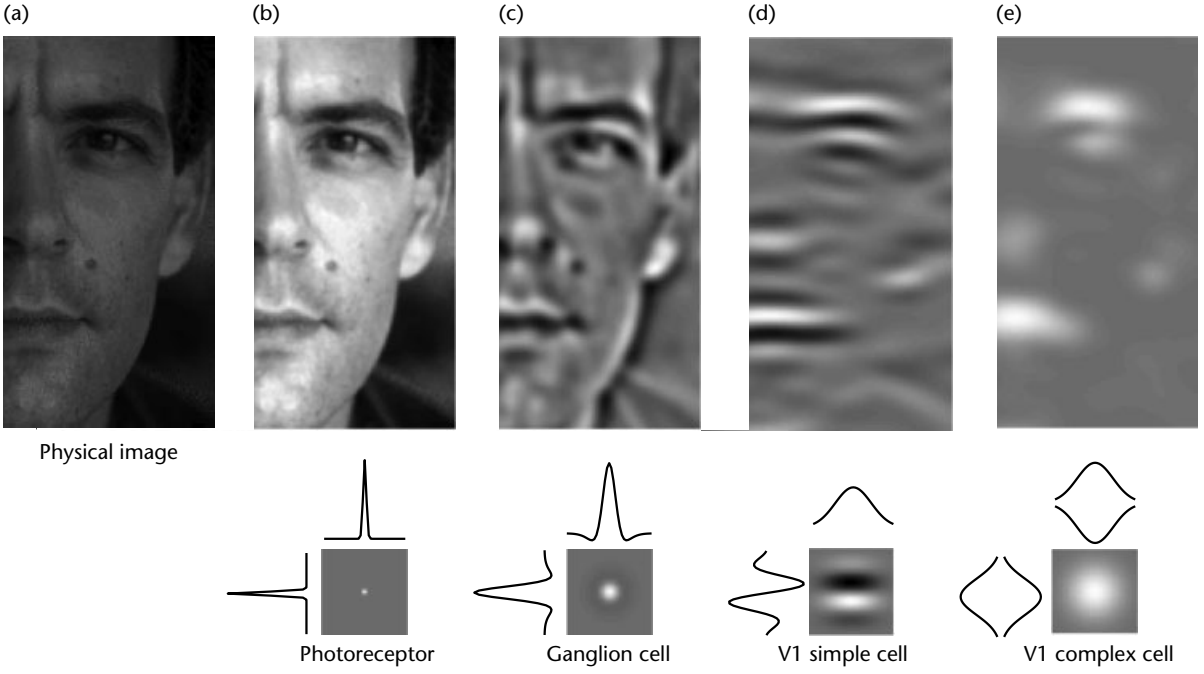


Figure 2. Receptive fields and neural images in the retina and primary visual cortex (V1). (a) A physical image obtained in conditions of dim illumination. (b–e, top row) Receptive fields of a photoreceptor, an on-center ganglion cell, a horizontally tuned V1 simple cell, and a horizontally tuned complex cell. (b–e, bottom row) Neural images corresponding to the cell types above. A neural image is a map of the responses of a given cell as it scans the whole image. White indicates a positive response, black indicates a negative response, and gray means no response.

neural image given by the responses of the whole array of photoreceptors (Figure 2(b), top). This neural image uses the full available range of responses (represented as white to black), even though the visual stimulus was illuminated by dim light (Figure 2(a)). We are mostly oblivious to light adaptation, but can become briefly aware of it by stepping from a dark room into bright daylight or vice versa.

Retinal ganglion cells have receptive fields organized in a center–surround fashion, and respond to the difference between the intensity of light in the center and in the surround. Cells in our example are on-center (Figure 2(c), lower panel), so they are optimally stimulated by a light spot on a dark background: they thus give a negative response to the mole just to the right of the person's nose, which is darker than the surrounding skin (Figure 2(c), top). Center–surround receptive fields enhance contours such as those at the edges of the face, the mouth, and the eyes. By extracting these features, they discard information about light intensity in uniform regions. Uniform regions, such as the cheek or the dark area behind the face in our example, stimulate both the center and the surround, whose contributions average out (resulting in gray regions in the neural image, Figure 2(c), top).

Image Filtering in Primary Visual Cortex

While the receptive fields of LGN neurons are similar to those of retinal ganglion cells, the receptive fields of neurons of cortical area V1 are radically different. Most V1 cells show receptive field properties that are absent at earlier stages, including selectivity for stimulus orientation and direction of motion. In particular, because of their orientation selectivity, V1 cells respond strongly to bars and edges of the appropriate orientation. (*See* **Vision: Early Psychological Processes**; **Occipital Cortex**; **Motion Perception, Neural Basis of**)

The receptive field of 'simple' V1 neurons is composed of separate 'on' and 'off' elongated regions (Figure 2(d), lower panel). The optimal stimulus to obtain a positive response is one that stimulates the 'on' regions with positive intensity while stimulating the 'off' regions with negative intensity. In our example, this happens near the mouth, the eye, and the eyebrow (Figure 2(d), top). In contrast, the edge of the face is mostly vertical, so it stimulates the 'on' regions and the 'off' regions at the same time. The inputs from these regions cancel out, and the resulting response of the cell is zero (gray in the neural image). The eyebrow is mostly horizontal, and evokes a negative response flanked by two positive responses, evoked by the regions above it and below it. Indeed, just like cells in earlier stages of the visual pathway, simple cells are sensitive to the precise location of a stimulus within their receptive field, and to the 'polarity' of a stimulus; i.e., whether the stimulus is brighter or darker than the background. This dependence on stimulus location and polarity is overcome by V1 'complex' cells. Complex cells have receptive fields where 'on' and 'off' regions overlap. For example, the 'on' region of our model complex cell takes up the whole receptive field (Figure 2(e), lower panel). As a result, the cell gives roughly the same positive response to a bar presented anywhere within its receptive field, whether it is lighter or darker than the background.

The neural image produced by complex cells (Figure 2(e), top) shows responses corresponding to the location of horizontal features such as the mouth, eye, and eyebrow. Complex cells such as the one in Figure 2(e) can be thought of as summing the positive outputs of many simple cells such as that in Figure 2(d), whose receptive fields are elongated in the same direction, and whose spatial location is displaced. By summing positive outputs from displaced simple cells, complex cells would overcome the dependence on stimulus location and polarity shown by simple cells. Invariance for location and polarity might be useful to maintain continuous responses in spite of small eye movements.

The neural images in Figure 2(d) and (e) illustrate only two of the many concurrent representations that are formed in our primary visual cortex when we look at a stimulus such as that in Figure 2(a). Indeed, the primary visual cortex contains cells selective for the full range of orientations and for a variety of spatial scales. (*See* **Occipital Cortex**; **Cortical Columns**)

An example of this variety is illustrated in Figure 3 for four simple cells selective for different combinations of orientation and spatial scale. The two cells in the left quadrants are selective for horizontal orientations, whereas the two cells in the right quadrants are selective for vertical orientations. The two cells in the top quadrants are selective for lower spatial scales than the two cells at the bottom. These cells respond to very different features in the visual stimulus (Figure 3(c)). The cell in the top right quadrant gives a negative response to the edge of the face, where the 'on' region is strongly stimulated by the dark background, and the lighter skin stimulates the 'off' region. None of the other cells responds well to these features. In contrast, the cell in the lower right quadrant responds strongly

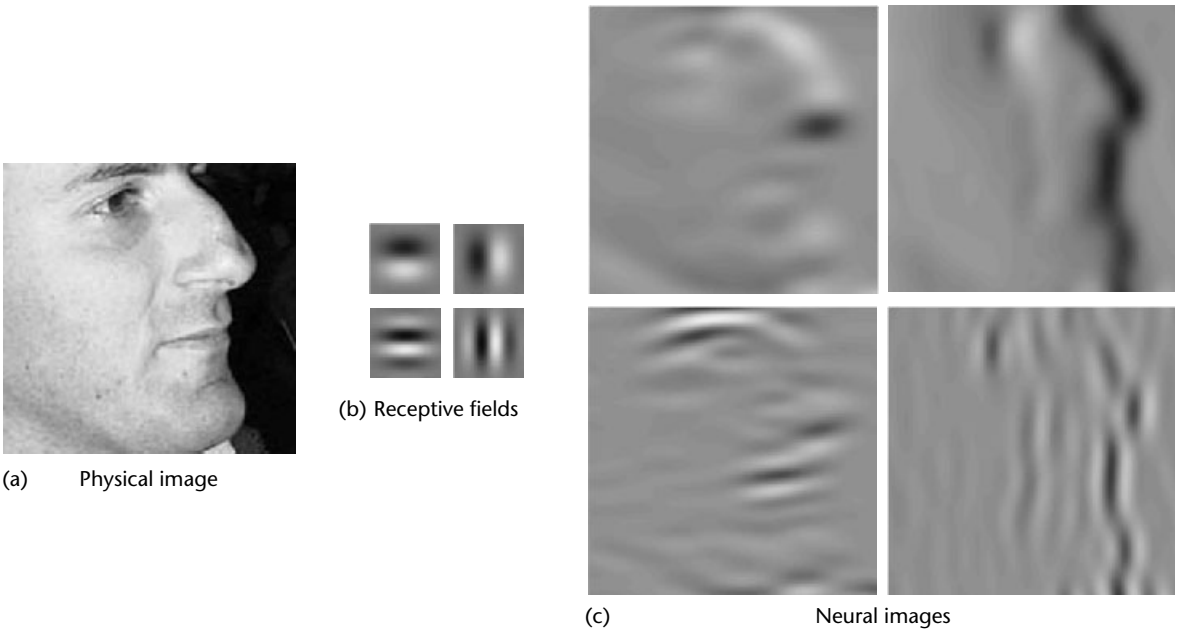


Figure 3. Neural images in V1. (a) Physical image. (b) The two-dimensional receptive fields of four simple cells, each sensitive to vertical, horizontal, or oblique orientations. The top set encodes orientations at a lower spatial scale than the bottom set. (c) The neural images resulting from the analysis of the image shown in (a) by the two sets of cells shown in (b). Each cell extracts one particular orientation at a given spatial scale. Adding together the four images in (c) would result in a coarse but faithful representation of the physical image (a).

to features such as the mouth and eyebrows, which are missed by the other cells.

In summary, cells in the primary visual cortex extract relevant elements of the visual image such as contours, lines and edges, at a variety of spatial scales. For the image in Figure 3(a), simple and complex cells stimulated by the edges of the face will give a strong response. Those centered on the background or on uniform parts of the face will remain mostly silent. Thus, the primary visual cortex extracts the face's edges and the contours of the face's various features. Simple cells provide an accurate description of the location of these features, as well as the indication of their polarity. For example, they indicate that the face is brighter than the background. The combined activity of simple and complex cells can yield a stable, faithful representation of the objects present in the environment.

CONTEXTUAL EFFECTS ON PATTERN VISION

We have described responses of visual neurons as entirely determined by the instantaneous distribution of light on their receptive field. In fact, these responses also depend on factors such as the previous history of stimulation, the overall contrast within the receptive field, and the distribution of contrast in the surrounding regions. These factors mostly affect the responses of neurons in the visual cortex.

Pattern Adaptation

Prolonged exposure to a visual pattern perturbs visual perception, reducing the perceived contrast and altering the appearance of subsequently viewed patterns. A classic example of this alteration can be experienced by looking at a pattern made of oblique bars for 30s or so and then looking at a pattern made of vertical bars. The vertical pattern will appear briefly as if it were tilted in the opposite direction to the preceding pattern. Effects of this kind are termed 'pattern adaptation'. A simple explanation for the perceptual effects of pattern adaptation states that, first, perception is the result of a weighted sum of the outputs of sensory neurons, and second, that pattern adaptation fatigues the neurons that respond most strongly. By causing these neurons to respond less strongly than they normally would, adaptation biases perception away from the adapting pattern.

The physiological substrate of pattern adaptation appears to lie in the primary visual cortex and in subsequent cortical areas. The responses of V1

neurons are sharply reduced after only a few seconds of stimulation, in a manner that could account for the perceptual effects. In contrast, the responses of neurons in the LGN and in the retina are essentially unaffected.

Pattern adaptation reflects a self-calibration mechanism in the visual cortex. This mechanism is continuously at work to adapt the responses according to the prevailing statistics in the stimulus. Its goal might be to increase the independence of subsequent neuronal responses, so that the response at a given time cannot be predicted by an earlier response. This makes maximal use of the resources, as it uses neuronal responses to encode the information that is not redundant over time. We tend to notice this calibration mechanism only when it misbehaves: when a prolonged stimulus is turned off, the calibration mechanism is caught off-guard, and visual perception is then briefly perturbed.

Interactions Between Superimposed Patterns

In addition to interacting in the time domain, different patterns interact when they are presented simultaneously. An example of this kind of interaction can be seen by superimposing two oriented patterns: for example, superimposing an orthogonal mask on an oriented test pattern impairs the perception of the test pattern. This phenomenon is known as 'masking', and reduces the apparent contrast of the test pattern.

The physiological correlate of masking appears to lie in mechanisms present in the retina and especially in the primary visual cortex. These mechanisms control the responsiveness of neurons and depend on overall contrast in a visual region centered on the receptive field. When two patterns are superimposed, overall contrast is increased, and neuronal responsiveness is reduced. This reduction in responsiveness has been observed in visual neurons of cats and monkeys, as well as in the electroencephalogram (EEG) obtained from visual cortex of humans. (*See* **Electroencephalography (EEG)**)

The current interpretation of masking effects involves inhibition between V1 neurons that have overlapping receptive fields but are selective for different orientations. Two superimposed patterns differing in orientation stimulate two sets of neurons, which inhibit each other and thus respond less than to a single pattern. This explanation, however, is being challenged and may not be entirely correct.

Interactions Between Spatially Displaced Patterns

Patterns do not have to be spatially superimposed to interact with each other. Strong interactions can also be observed between patterns that are spatially segregated. Perceptually, these interactions can be suppressive or enhancing, depending on stimulus configuration.

The physiological substrate of these interactions between spatially separate visual patterns seems to lie in area V1 and in the successive cortical areas. The responses of V1 neurons are influenced by stimuli presented in the regions surrounding their receptive field. This influence is commonly considered to originate in 'lateral inhibition', reciprocal inhibitory connections between pools of V1 neurons with displaced receptive fields.

Among the phenomena that would be explained by lateral inhibition is 'surround suppression'. Oriented patterns situated in the region surrounding the receptive field of a V1 neuron can substantially reduce the responses of the neuron, especially if their orientation is similar to that preferred by the neuron. For example, the responses of a V1 neuron to an optimally oriented bar can decrease if the bar is made to extend beyond the receptive field of the neuron. Cells showing this property are often called 'end-stopped'. One of the roles of surround suppression might be to enhance the independence of V1 neurons. The statistics of images commonly encountered by the visual system are such that contours and edges tend to be surrounded by other contours and edges. For example, in Figure 3 the horizontal contour given by the mouth is close to the vertical contour given by the edge of the face. Without surround suppression, knowing that a V1 neuron is strongly activated would also say something about the nearby neurons: they would be more likely to be active than other neurons. This lack of independence between neurons would be wasteful, as it would encode redundant information.

In addition to suppression, interactions between different regions of the receptive field can in some conditions involve enhancement. This enhancement may result from a sort of double inhibition: if one reduces the activity of a population of neurons that normally inhibits another population, the activity of the second population may be enhanced.

Anatomically, the interactions between different regions are thought to be mediated by local circuits as well as by long fibers running parallel to the cortical surface, which connect V1 cells over

distances of several millimeters. Long-range connections tend to link cells that share the same preferred orientation. Moreover, they preferentially connect cells along an axis that corresponds to the cells' preferred orientation.

SPATIAL REPRESENTATION OF VISUAL PATTERNS

Area V1 contains a full representation of the visual field, with nearby regions in the visual field corresponding to nearby regions in area V1. This kind of representation is called 'retinotopic'. Retinotopy is derived by a similar organization in the retina, optic nerve, and LGN; it coexists with other organizational principles that group cells according to characteristics such as orientation preference and ocular dominance. (*See* **Cortical Columns**)

Retinotopy can be observed in Figure 4(a), which illustrates an idealized visual stimulus resembling a bull's-eye, and the corresponding map of responses across area V1 in one hemisphere (Figure 4(b)). This representation of the visual stimulus is heavily distorted: just like the retina and the LGN, area V1 devotes most of its cells to the center of fixation (fovea). Indeed, while the portion of the bull's-eye that is inside the first solid ring takes up a tiny portion of the surface of the stimulus, the region that represents it occupies about half of area V1. Another major difference between the stimulus and the corresponding map of activity is the geometrical arrangement of the features. Concentric circles and radial lines in the stimulus become vertical and horizontal stripes of activity in the cortical map. This geometrical distortion could be important to help the rest of the brain recognize objects regardless of their distance and of their orientation. Indeed, as we move closer to an object, its size on the retina is scaled up. If we tilt our head (or the object), the object's image on the retina rotates around the fovea. Thanks to the transformation operated by the cortex, such scalings and rotations become simple translations, which may be easier to analyze for the subsequent stages of visual processing.

A striking example of the consequences of cortical magnification is illustrated in Figure 5. In this simulation, a mathematical model was fed a photograph of a room (a) and generated a prediction of the distribution of responses across the primary visual cortex in one hemisphere (b). The model simulates an observer who focuses her eyes on the letter 'O' on an eye chart (see detail) located at the back of the room. The representation in V1 is amazingly distorted. Most of the cortical surface devoted to the stimulus is taken up by the eye chart, and particularly by the letter 'O' and by the nearby letters. Barely any cortical space is given to the rest of the room, even though it occupies by far the largest portion of the visual stimulus.

PATTERN VISION BEYOND PRIMARY VISUAL CORTEX

Area V1 is by no means the only visual area in the cerebral cortex: in primates there are tens of visual areas, each containing a full representation of the visual field. These areas receive input directly or indirectly from V1. From what we understand, different areas emphasize the analysis of different aspects of the visual world. In particular, areas V2, V4, and IT appear to be closely related to the perception of visual patterns. (*See* **Occipital Cortex**; **What and Where/How Systems**)

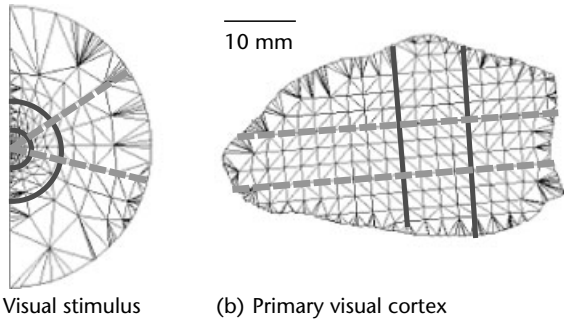


Figure 4. Retinotopy in primary visual cortex. Mathematical model of the transformation between a visual stimulus (a) and area V1 in one hemisphere (b). Solid and dashed lines indicate the transformation of concentric and radial features in the visual stimulus into vertical and horizontal lines of activity in V1. From Frederick C and Schwartz EL (1990) Conformal image warping. *IEEE Computer Graphics and Applications* (March): 54–61.

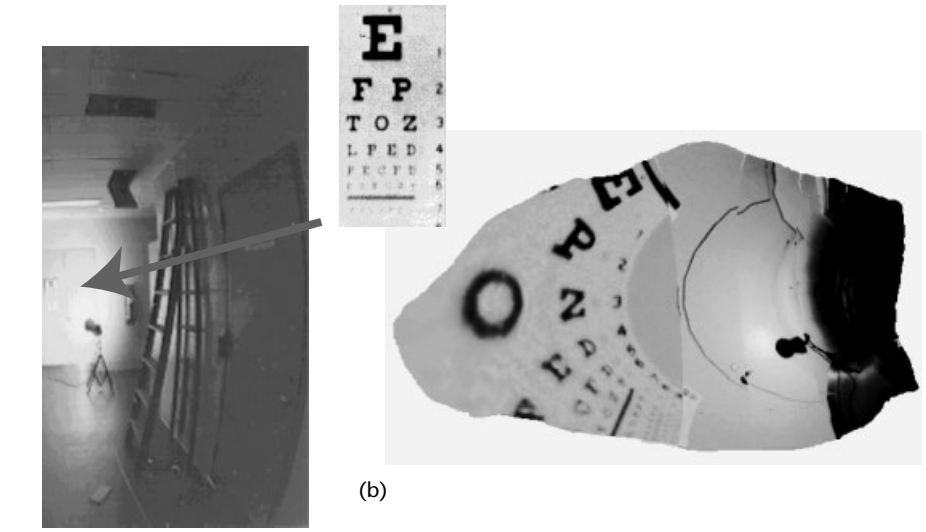


Figure 5. Application of the model in Figure 4 to simulate the representation in area V1 of the image of a room. In the back of the room is an eye chart (see detail), and the retina is centered on the letter O. The representation in cortex greatly magnifies the letters in the eye chart at the expense of the rest. From Schwartz EL, Merker B, *et al*. (1988) Applications of computer graphics and image processing to 2D and 3D modeling of the functional architecture of visual cortex. *IEEE Computer Graphics and Applications* (July): 13–23.

Cortical Area V2

The next stage of visual processing after V1 is performed in the secondary visual area, V2. The precise role of V2 is unknown; some differences between the responses of V2 and of V1 have been found, but most receptive field properties are similar in both visual areas. At equivalent positions in the visual field, the receptive fields of V2 cells are slightly larger than in V1. A functional property that has been demonstrated for V2 cells is their ability to respond to illusory contours, contours that are defined by figural clues rather than by lines or edges. Such contours could result from partially occluded objects, and the capacity to interpolate them is necessary to segment the visual scene into various objects. On the other hand, there are suggestions that also V1 cells can respond to such contours, so it is not clear that this property is specific to V2.

Cortical Area V4

The properties of V4 cells are even less well understood than those at earlier stages, but several results suggest that V4 cells have an important role in pattern vision. First, receptive fields in V4 are considerably larger than in V1 or V2, making V4 cells suitable for the analysis of large areas of the visual field. Large receptive fields could set the stage for the detection and identification of whole objects. (*See* **What and Where/How Systems**)

Second, V4 transmits information to the inferior temporal (IT) cortex, a cortical region that is known to be crucial for object recognition (see below). Third, a number of studies have shown that the activity of V4 cells is strongly modulated by attention. The response of a V4 cell to a given stimulus is high if the stimulus has a behavioral relevance (i.e., if the observer is paying attention to it), and low otherwise. Fourth, the foveal representation is even more enhanced in V4 than it is in either V1 or V2. This would seem reasonable if V4 were to be concerned with pattern vision rather than with attributes of the whole visual field.

Finally, recent physiological results, although controversial, add further support to a role of V4 in pattern vision. Although most V4 neurons, like those of V1 and V2, are selective for the orientation of a visual stimulus, a number show a preference for stimuli that differ from the bars and edges that seemed optimal to stimulate cells at earlier levels. For example, some V4 neurons respond better to stimuli made of concentric rings, or of radially oriented line segments, than to simpler bar and edge stimuli.

Thus, individual V4 cells seem able to encode visual patterns more complex than those encoded by individual cells at earlier levels. Taken together,

these facts strongly suggest that V4 is important for pattern vision.

Inferotemporal Cortex

In the cortical areas of the inferior temporal lobe, neurons prefer stimuli even more complex than those described in area V4. These cells often produce their highest activity for stimuli that have complex, irregular shapes, or for stimuli made of numerous features arranged in complex patterns. There are, for example, neurons in the anterior IT cortex that respond selectively to faces, or to other body parts. The responses of IT neurons are often invariant relative to stimulus position within their receptive field. For familiar objects, they can even be invariant relative to the angle of view. A given neuron could thus respond to an object regardless of the object's rotation, even though the object's two-dimensional projection on the retinas varies considerably in these conditions. These neurons are therefore important for the perception of complex patterns, a function necessary to achieve object identification. (*See* **Face Cells**)

The role of IT cortex in complex pattern perception and object perception is supported by lesion studies. Humans and monkeys with IT cortex lesions are severely impaired in object recognition tasks. They fail to recognize familiar objects, and are incapable of learning new ones. A particular example of such a loss is the incapacity to recognize faces, a condition called 'prosopagnosia'. This condition is thought to result from lesions to the parts of IT cortex specialized in the analysis of faces. Studies of the neurons' receptive fields and those of cortical lesions thus converge to indicate that IT cortex has an important role in pattern vision.

THE PERCEPTION OF PATTERNS

We have seen that visual stimulation activates neurons at numerous stages in the visual pathway. In the responses of these neurons we have also found counterparts for perceptual effects such as pattern adaptation, masking, and the interactions between spatially displaced patterns. These results suggest a link between the responses of cortical neurons and the perception of visual patterns.

Neural Activity and Pattern Vision

Other links between the responses of cortical neurons and the perception of visual patterns have been demonstrated by lesion studies and by stimulation experiments. Studies have been made of people with lesions of the visual cortex due to disease or to accident, for example patients with gunshot wounds after World War I. These studies established that the absence of a part of visual cortex yielded some deficit in visual perception. Stimulation experiments employ localized injections of current to stimulate small groups of cortical neurons, and have been mostly performed on surgical patients or on blind volunteers. Electrical stimulation of the visual cortex leads to the perception of illusory visual stimuli.

More recently, a strong link between neural activity and pattern vision has been demonstrated using fMRI. This imaging technique is not invasive, and allows the experimenter to measure brain activity in people who are awake, while they make perceptual decisions on the presence or absence of a visual stimulus. Using this method it has been shown that the detection of visual patterns relies on neurons in V1 and in subsequent areas of the visual cortex. The activity of these neurons predicts the performance of the participants and even correlates with perceptual decisions on a trial-by-trial basis. When making a difficult perceptual judgment, participants make a number of mistakes, such as indicating that a pattern is present when it is absent (false alarms), or that it is absent when it is present (misses). The responses of visual cortical neurons correlate with the perceptual decision of the participant rather than with the physical stimulus. Responses in visual cortex determine whether the person will perceive the stimulus.

The Binding Problem

Most of the theoretical and experimental work described above implies that the early stages of the visual system decompose an image into features, or components. Our visual perception, however, is one of unity and coherence, not a juxtaposition of independent elements. This suggests that at some stage of the visual pathways the image components must be bound together to form a coherent perception. To date, the solution to this problem is largely unknown. (*See* **Decoding Single Neuron Activity**)

In an attempt to provide an answer to this question, it was proposed that the neuronal substrate that glues different features together is found in the temporal pattern of neuronal activity. In that view, two cells coding for features belonging to the same object would synchronize their activity. The synchronicity of cell firing would thus serve as a tag for a given object, which is propagated through the various stages of visual processing. The so-called 'binding by synchrony' hypothesis received some

experimental support with the discovery of cells in the visual cortex of cats and monkeys that, for example, synchronize their activity when they are simultaneously activated by a single bar, and not when activated by two distinctly separate bars. Whether this synchronization reflects the process of 'binding' is the subject of intense debate and experimental scrutiny and there is, to date, no consensus on its validity.

Further Reading

Barlow HB and Mollon JD (eds) (1982) *The Senses*. Cambridge, UK: Cambridge University Press.

De Valois RL and De Valois K (1988) *Spatial Vision*. Oxford, UK: Oxford University Press.

Harris CS (1980) *Visual Coding and Adaptability*. Mahwah, NJ: Lawrence Erlbaum.

Hubel DH (1988) *Eye, Brain and Vision*. New York, NY: Scientific American Library.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer.

Wandell B (1995) *Foundations of Vision*. Sunderland, MA: Sinauer.

Wurtz RH and Kandel ER (2000) Central visual pathways. In: Kandel ER, Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, pp. 523–547. New York, NY: McGraw-Hill.

Wurtz RH and Kandel ER (2000) Perception of motion, depth and form. In: Kandel ER, Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, pp. 548–571. New York, NY: McGraw-Hill.

Perceptual Systems: The Visual Model

Introductory article

Brian J Stankiewicz, University of Texas, Austin, Texas, USA

CONTENTS

Perception is the interpretation of specific sensation from the visual, auditory, somatosensory and olfactory sensation systems. Computational models of perceptual systems attempt to understand the connection between the input (sensation) and the output (perception) of these systems.

INTRODUCTION

Humans are equipped with a series of sensing systems that give us information about the environment around us. Ranging from tactile sensors that allow us to grasp and feel an object, to our visual system that allows us to recognize family and friends across a crowded room, our senses provide us with vital information about our immediate and distant environment. One of the goals of cognitive scientists is to understand how the human brain converts the raw data from each of these sensing systems into a coherent and meaningful interpretation of the outside world. One of the approaches to understanding how this is done is to develop computational models of perception that use either real or simulated sensory data, which provide a meaningful interpretation of the data.

Because our cognitive system is designed to interpret sensory data with such remarkable speed and accuracy, it is often difficult to appreciate some of the computations that are involved in a common event. For example, imagine that you are walking down the sidewalk and you suddenly hear a loud sound behind you. This sound would produce a significant startle response, causing you to turn your head immediately to observe the event. Your visual system would then provide you with distal information about the scene (*See* **Visual Scene Perception**). Seeing a bicycle careening out of control in your general direction, you might decide to jump off the sidewalk on to the grass lawn. Wanting to ensure that you would not hurt yourself too badly during this move, you might hold your arms out to 'break' the impact of your fall. On feeling the grass touching your hands, you would gently collapse the muscles in your arms so that you landed safely on the grass.

The above example demonstrates how we use our senses every day with little thought. In this example your auditory system received a series of sound pressure waves indicating a loud sound. Your perceptual system processed the sound waves in order to localize the source as being behind you. Your visual system then received visual information about a moving object and quickly calculated the motion vector of the object to determine that it was heading in your general direction. Finally, your tactile system allowed you to determine the time of contact so as to break your fall safely after jumping out of the way of the careening cyclist.

Each and every day we perform thousands of processes on the sensations that our body and brain receive. These senses provide the fundamental input to our cognitive system that enables us to interpret the world around us. Whether they are determining whether Stilton cheese and walnuts taste good together or they are avoiding an out-of-control cyclist, our senses are providing our cognitive system with vitally important information about our environment.

PROCESSING SENSORY INFORMATION

Our senses provide us with a wealth of information. However, in its raw form this data is not very informative. Somehow the cognitive system must utilize the raw sensory data and interpret it informatively. In the above example, the visual system provided the pedestrian with essential information about the direction of travel of the cyclist. If the cyclist had been traveling in the

opposite direction, it would not have been necessary to jump out of the way. Using the raw images that are projected on to the retina is insufficient for estimating the direction of an object. In order to calculate the heading direction of the cyclist accurately, the visual system must make multiple computations across time to calculate the object's heading precisely.

Research scientists who are interested in understanding human perception often attempt to model the input–output relationship between sensation and perception. That is, using the information that is available to the human visual system (e.g. a series of raw images), these researchers will postulate how the raw information is processed and develop computational models that make explicit calculations about the external environment (e.g. the heading of an object, given a sequence of images). These models then make predictions about how other stimuli will be processed by the human observer, and if they model human perception accurately, these predictions will be supported by empirical (behavioral) data.

INFLUENCE OF NEURAL CODING STRATEGIES

Although perceptual models are mainly influenced by behavioral data, these models are also strongly influenced by knowledge from neuroscience. Neuroscientists are interested in understanding how specific types of information are processed and encoded in the brain. To do this, they use a technique called single-cell recording. Single-cell recordings are made by inserting an electrode into a neuron (or a group of neurons) and measuring the neural responses of these neurons to specific stimuli (e.g. an image or a sound). From an understanding of the properties of a group of neurons in a specific area of the brain, neuroscientists try to deduce the computations completed by this group of neurons.

Understanding how specific neurons in the brain respond to specific stimuli has greatly influenced our understanding of visual perception. The human eye serves as the initial sensing device for the visual perceptual system. The eye is equipped with an array of light-sensitive neurons that convert light energy into neural responses. Research scientists have presented carefully controlled stimuli to a visual system while making recordings from different areas of the eye and/or brain. These studies have revealed a number of important facts about the neural processing of visual stimuli. Although the details of these studies are beyond the scope of this article, the following sections provide an introduction to some of the properties elucidated by the use of this technique.

Receptive Fields

An important concept in the subject area of perception and neuroscience is that of a receptive field (*See* **Amygdala**). Most sensing systems have some type of receptive field coding. However, the concept of a receptive field is perhaps most intuitive with regard to the visual system. To understand the concept of a receptive field fully, one needs to understand the initial input into the visual system. The retina is located at the back of the human eye and is composed of approximately 130 million individual cells called photoreceptors that measure the light from a very small part of the visual field. These cells convert that light energy into neural responses. The retina is similar to the charge-coupled device (CCD) on a digital camera or digital video camera that converts light energy into electrical signals that can be recorded on to a digital medium. However, instead of recording these light intensities, the visual system subjects them to a series of online processing to convert them into meaningful pieces of information about the external world. In fact, there is major information processing of the image within the eye itself.

Each photoreceptor in the eye has a specific receptive field. That is, each cell is responding to light that is projected from a specific area of the visual field. If one was to shine a light on a specific photoreceptor, the neuron would become active, but shining the light anywhere else on the retina would produce no response. One can infer the receptive field of a photoreceptor by recording where the light is when the cell fires. The cells that leave the eye are called ganglion cells. The latter have dramatically different receptive fields to those of the photoreceptors. Ganglion cells take input from multiple photoreceptor cells that try to excite or inhibit the response of a given ganglion cell. Figure 1 illustrates the influence of photoreceptors on ganglion cell responses.

Using single-cell recording techniques, researchers have determined that there are two types of ganglion cells, both of which have a 'center–surround' receptive field (see Figure 1). One type of ganglion cell has an 'on-center–off-surround' receptive field. These cells respond maximally to an image that has a bright center with a dark ring surrounding it. The 'off-center–on-surround' cells have the opposite configuration.

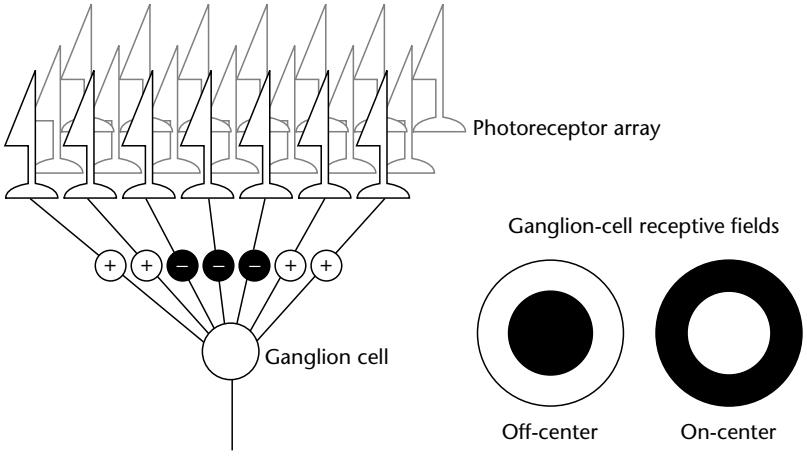


Figure 1. Two types of center–surround ganglion cell receptive fields ('off-center–on-surround' and 'on-center–off-surround'). The diagram on the left demonstrates the relationship between ganglion cells and photoreceptors. Activation of some photoreceptors will excite a ganglion cell ('+' connection), while other cells when activated (i.e. shown light) will inhibit the response of the ganglion cell ('−' connection). This type of relationship between photoreceptor and ganglion cell would produce an 'off-center' ganglion cell receptive field. The two diagrams on the right illustrate the optimal stimulus (i.e. the stimulus that gives the largest response) for the two types of ganglion cell. One responds best to a stimulus with a dark center and a bright surround, while the other responds best to a stimulus with a bright center and a dark surround.

Perceptual Influence of Ganglion Cells

The famous Hermann illusion is shown in Figure 2. This can be explained by the output of the ganglion cells from the eye. Almost immediately upon viewing the image in Figure 2 you will observe a series of gray areas between the white intersections. If you inspect the image closely (perhaps by covering up any four squares with your fingers while leaving the four-way intersection visible), you will see that these gray areas are not in fact part of the image. That is, the areas are solid white. This graying is an illusion produced by our visual system.

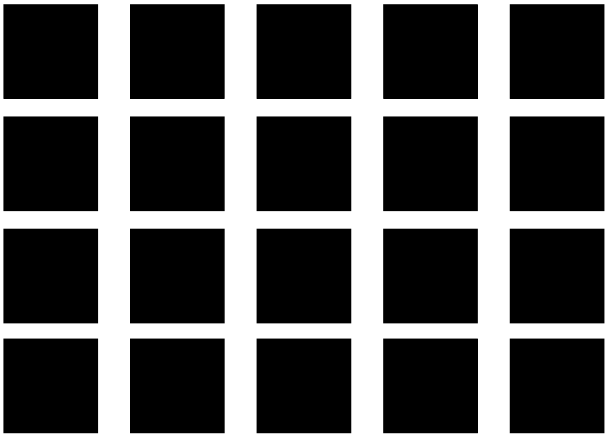


Figure 2. The Hermann illusion. Note the perception of gray dots at the intersection of each white '+' symbol. These gray dots are a perceptual illusion that can be explained by the output of visual ganglion cells.

In order to understand the illusion, we can refer back to the center–surround receptive fields of the ganglion cells. Figure 3 provides an explanation as to why we perceive the stripes as being brighter than the four-way intersections. The explanation relies on the output of the center-surround receptive fields found in the ganglion cells of the eye (see Figure 1). The on-center–off-surround cells have different levels of activation on the stripes to those that they have on the intersections. The ganglion cells that are spatially centered on the stripes will receive positive input from the white bar, and will receive a small amount of negative input from the white bar. By contrast, the ganglion cells that are spatially centered on the intersections will receive the same positive input as the cells on the stripes, but they will receive much more negative input from the two sets of stripes (the two white stripes; see right-hand side of Figure 3). This increase in negative input into a ganglion cell will cause the ganglion cells on the stripes to be more active (indicating a higher contrast) than the cells on the intersections. The brain then interprets this difference in ganglion cell output activation as the stripes being brighter than the intersections.

The Hermann grid illusion provides a good illustration of the basic principles underlying perceptual

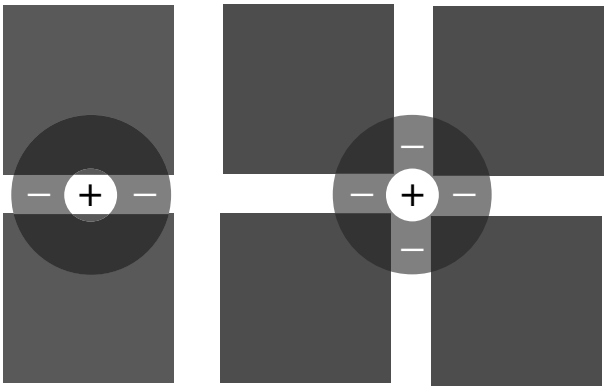


Figure 3. Application of the center–surround receptive fields to the Hermann grid illusion. The ganglion cells that are 'on-center' receive more negative input at the four-way white intersections than at the white stripes. This increase in negative input to the 'on-center' ganglion cells produces a smaller response for ganglion cells at the intersections than for those on the stripes. This reduced response translates into an interpretation by the visual system that the area is not as 'bright', thus producing the illusion observed in Figure 2.

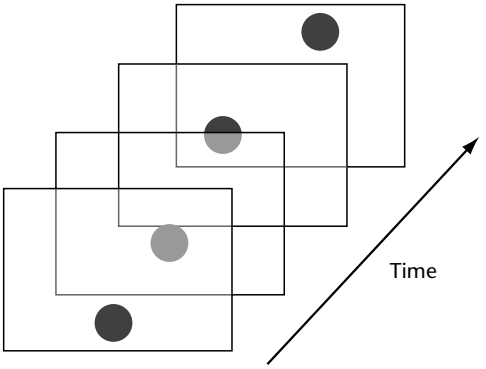


Figure 4. Illustration of apparent motion. A circle moves from the lower part of the image in frame 1 (the foremost frame) to the top of the image (the last frame). If these four images are viewed in succession with a fraction of a second presented between each frame, the human visual system perceives the circle to be moving from the bottom of the image to the top of the image.

modeling. The endeavor typically begins with a perceptual observation – in this case the dark spots in the Hermann grid. This is then followed by an observation in neuroscience or a hypothesis about how the neural substrate may be processing visual information to produce the behavior. In this example, the explanation started with a neuronal observation, namely the center–surround receptive fields of the ganglion cells. Following the observation or hypothesis, a model is developed. In this case, the model is a rather simple one in which the excitatory and inhibitory responses from the retina are simply added together to obtain a ganglion response. Although this is a simple example, it illustrates how all perceptual modeling is typically completed.

MOTION PERCEPTION

The Hermann grid illusion provides an important illustration of modeling perceptual processing for static images (*See* **Vision: Early Psychological Processes**). However, most of the time we do not process static images, but rather we need to deal with our highly dynamic environment in which objects move relative to one another, producing the perception of motion. As was illustrated by the example of the out-of-control cyclist, it is often important to judge the motion direction of other objects relative to one's own body. At other times, perhaps while driving, one needs to adjust one's heading or speed in relation to other objects. Both of these tasks rely on the visual system providing an accurate cue of how objects in the environment are actually moving. To make these computations, one needs to analyze more than one image (unlike analysis of the Hermann grid illusion). Figure 4 illustrates a motion effect. The figure contains a series of four images. If the sequence of images was to be presented rapidly, one's visual system would interpret the scene as a circle moving from the bottom to the top of the page. One can observe this motion phenomenon with 'flip-books', which are composed of several hundred images in which the objects are moved slightly from one page to the next in the book. By rapidly 'flipping' the pages (typically with one's thumb) one perceives apparent motion of the objects in the images. This technique of apparent motion is the basis of cartoons and motion picture production. By drawing a series of images, or taking a series of pictures across time and then presenting them in rapid succession, the objects in the movie are perceived to be actually moving.

Computing Motion

Unlike our earlier analysis of the ganglion system that could be understood by modeling a single image, motion computations rely on computations from multiple frames (*See* **Motion Perception, Psychology of)**. There are two important parameters that define the movement of an object in space, namely the direction in which the object is moving and the speed at which the object is moving. Computational models of motion attempt to calculate both of these properties of motion from a sequence of images presented to the retina.

Measuring motion direction

Figure 5 shows a simple model of motion estimation using a delayed-circuit approach. There is a series of receptors that respond to changes in luminance. When there is a change in luminance (either from light to dark or from dark to light), these cells try to activate the cells to which they are connected. If the cell is connected to a 'delay' unit, the latter waits a certain period of time before sending the activation to the direction-selective cells to which it is connected. In Figure 5 the direction-selective cells have to receive activation from two cells simultaneously in order to be active. If we assume that the delay is one time slice, then the second receptor from the top will receive a change in luminance on time slice 3, and will send that activation instantaneously to the delay unit. On time slice 4, the uppermost receptor will receive a change in luminance, and will send that activation to the vertical direction-selective unit. At the same time, the delay unit will also send its stored activation (from time slice 3) to the vertical motion-selective cell. In this case, the vertical direction-selective cell will receive activation from two units simultaneously (from the delay unit and from the uppermost receptor).

If we analyze the downward direction-selective cell, we find that it does not receive activation from two cells at the same time. On time slice 1, the unit receives activation from the lowest receptor. On time slice 2, the second lowest receptor receives a change in luminance and sends its activation to the delay circuit. After a delay of one time slice (on time slice 3), the delayed activation is sent to the downward direction-selective cell. The latter receives activation on time slice 1 and time slice 3, thus not satisfying the requirement for receiving activation from two sources simultaneously in order to become active.

Measuring motion speed

As was mentioned earlier, there are two important components to the perception of motion, namely direction and speed. Figure 5 illustrates how motion direction is computed, but it does not specify how the speed of the motion is computed. To measure the speed of the motion in addition to its direction, one simply needs to modify the amount of delay. In the above example the delay was set to one time slice. If we add another unit that receives direct input and input that is delayed by three time slices, then we have one unit that is sensitive to upward motion which is fast (a delay of one time slice) and a second unit that is sensitive to slow upward motion (a delay of three time slices). Using this model can probe the motion-direction cells to determine the direction of motion in a particular location in the visual field, as well as estimating the speed of the motion.

VISUAL OBJECT RECOGNITION

An impressive component of human perception is our ability to recognize three-dimensional objects from their two-dimensional retinal projections (*See* **Vision: Object Recognition**). We are so proficient at human object recognition that it can be difficult to understand completely some of the difficulties faced by our human object recognition

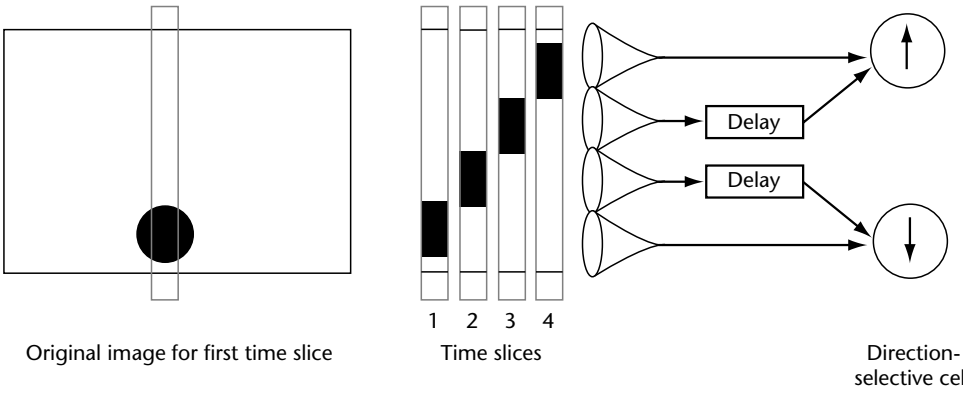


Figure 5. A simple motion-detection model. The image on the left-hand side is the original image for the first time slice shown in Figure 4. The black box is selecting a single column of pixels for the analysis. The right-hand side shows the same column of pixels for time slices 1 to 4 from Figure 4. The far right-hand part of the figure shows a method for detecting upward motion and downward motion using a delay circuit. When the circle moves from one receptor to the next at the appropriate delay, the responses from the upper two cells activate the vertical-motion detector simultaneously. This simultaneous activation indicates that there was vertical motion at a specific rate. The lower two receptors are wired to detect downward motion.

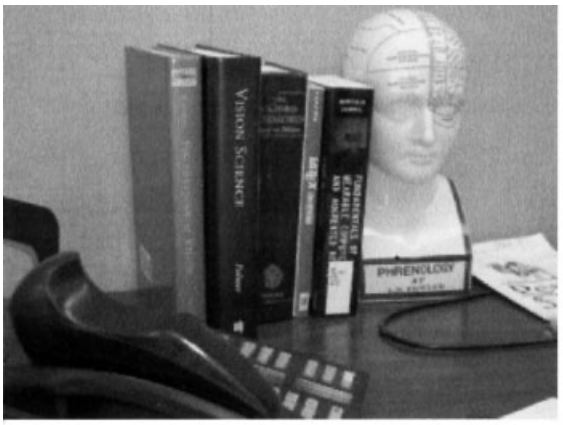


Image 1

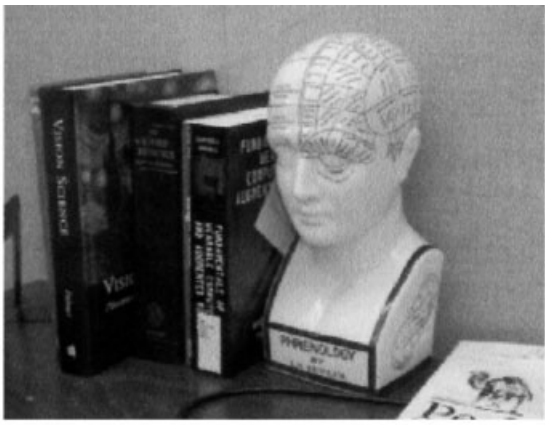


Image 2

Image 1 – Image 2

Figure 6. Example of object recognition performance. The upper two images are two views of the same scene. Immediately upon viewing these images one is able to determine that the two images are projections of the same set of objects. However, the sets of images are dramatically different, as is illustrated by the lower image that is the difference between the two upper images. If the images were identical, the lower image would be black.

system. Perhaps one of the most amazing aspects of human object recognition is that we are able to classify an image as a projection of a known object despite the fact that we will never see the same image projection twice. Figure 6 illustrates this. The two images are projections of the same three-dimensional object from two different viewpoints. Almost immediately upon viewing the two images you were probably able to determine that the two objects were the same. However, closer inspection shows that the two images are dramatically different. To illustrate just how different these two images are, we can simply take a difference in the image intensities at each image pixel. Figure 6 illustrates this difference. As you can see, these two images differ quite substantially.

Figure 6 illustrates one of the fundamental challenges faced by the human visual system, namely recognizing objects from different images when a certain imaging variable is changed. In this illustration the imaging variable that has been changed is viewpoint, but you can achieve a similar effect by changing the lighting direction, or by occluding part of an object by placing one object in front of another. However, to illustrate the modeling components of visual object recognition I shall mainly consider issues associated with viewpoint.

Feature Models of Human Object Recognition

One approach to modeling human object recognition is to propose that the visual system does not store in memory the precise pixel array to match new views of an object, but instead stores a collection of features in memory. The features might consist of extracted edges, color and/or vertices. For each object stored in memory, there would be a feature vector that would specify the set of features for a particular object. According to this

approach, the location of the features in the image is not important. Simply having the same set of features would be sufficient to enable the object's identity to be determined.

Limitations of feature models

Because feature models do not store the locations of the features, these models can recognize new views of an object if the same set of features is present in the image. However, this has a consequence. Simply specifying the set of features without specifying their spatial relationships in the image produces some strange behavior. For example, a feature model would treat the two images shown in Figure 7 as equivalent (the features are all present – they have just been rearranged from one image to the other). Although we might classify the image on the right as a mug, our perception of the objects in the two figures are not equivalent. Feature models would classify these two images as containing an identical object. A second limitation of feature models is that as an object is rotated in depth, the set of image features in one view becomes occluded (unobservable) in another view (for example, in Figure 6 the left ear of the phrenology head in Image 2 is not visible in Image 1). This makes the correspondence problem difficult for these models.

Alignment Models

More sophisticated versions of the feature models are alignment models (*See* **Mental Rotation**; **Vision: Object Recognition**). Unlike the feature models, alignment models utilize the spatial arrangements of the features in the image. However, in order to compensate for viewpoint changes, these models allow for specific types of image transformation on all of the collection of image features and their locations. For example, some simple image transformations that can be achieved include changing the size of the image (scale), changing the position of the image (translation) and rotating the object around the line of sight. By applying these image transformations from one view of an image to another, we can recognize new views of an object by applying these transformations to the set of features stored in memory and seeing whether they can be aligned with the locations of the features in the viewed image.

Limitations of alignment models

Alignment models are capable of aligning one view of an object with another if the change in viewpoint is due to a change in the object's size (scale) or location in the image (translation), or to rotation in the picture plane. However, these models encounter problems with rotations in depth when features become occluded (e.g. the left ear of the phrenology head in Figure 6). Because these models use features and align the positions of those features from one view to another, it is usually necessary to solve the correspondence problem (i.e. to determine which feature in one image corresponds to a feature in another image). When objects are rotated in depth, new features can appear while other features become occluded (*See* **Vision: Occlusion, Illusory Contours and 'Filling-in'**). Under these conditions, aligning the set of features from one view to another can be a very difficult task.

Part-Based Models

Another way to model human object recognition is by using a part-based approach. Part-based models begin by decomposing an object into a set of simple

Figure 7. A challenge faced by feature models of human object recognition. The features of the object on the left-hand side remain present when the image is scrambled. According to the feature approach, the two images are the same.

volumetric primitives. After the object has been decomposed into its constituent parts, the model then uses the image features to make shape estimations of the parts. For example, the cup part of the mug might be described as having a 'curved cross-section' and a primary axis that is 'straight'. The handle might also be described as having a 'curved cross-section', but with a primary axis that is 'curved'. In addition to describing the parts, these models also specify the relationships between the parts. For example, the simple volume describing the handle might be described as being 'beside' the cup portion of the mug. In these models the inter-part relationship can either be specified with respect to the viewer (e.g. 'beside' is specified as being to the left or the right of another part) or relative to the object as a whole (e.g. 'beside' is specified as being to the left or the right of the global primary axis of the object).

Limitations of part-based models

One of the primary limitations faced by these part-based models is segmenting the object from the background and then decomposing an object into its constituent parts. Although to most people it is obvious which parts of the image shown in Figure 6 are the object and which areas are the background, making these distinctions from the gray-level image can be difficult. A second limitation of these models is the ability to make the appropriate shape inferences from the gray-level image. That is, although there are image cues that may provide clues as to whether the cross-section is curved or straight, these cues can be very subtle in the image, and extracting them can prove challenging.

SUMMARY AND CONCLUSIONS

Perception plays an important role in our ability to interpret our environment. Whether we are watching a child ride a bicycle or listening to an old friend's voice on the telephone, perception is at work trying to convert the raw sensory data into a meaningful interpretation of the world around us. Because perception plays a vital role in our knowledge of the world in which we live, cognitive scientists are interested in understanding how the human mind converts these raw sensory patterns into meaningful interpretations. In order to understand better how a perceptual system might interpret raw sensory patterns, perceptual scientists will develop computational models of a specific perceptual process. These models will try to account for the current body of behavioral data as well as making novel predictions about human perception that can be tested.

Further Reading

Landy MS and Movshon JA (1991) *Computational Models of Visual Processing*. Cambridge, MA: MIT Press.
Marr D (1982) *Vision*. New York: WH Freeman.
Wandell BA (1995) *Foundations of Vision*. Sunderland, MA: Sinauer Associates.

Phonology, Neural Basis of

Intermediate article

KG Munhall, Queen's University, Kingston, Ontario, Canada

CONTENTS

The phonology of a language includes a representation of the set of sounds used in the language and a representation of the rules for producing and perceiving the sounds in sequences. Clinical evidence from aphasic patients indicates that phonological impairment on language perception and production results from both anterior and posterior cortical lesions. Functional imaging research has confirmed this pattern and using these techniques a more detailed neural mapping of the neural basis of phonology is beginning.

INTRODUCTION

At the heart of natural human language is an oral-, aural-based sound system called a phonological representation. To be a native speaker of a language means that you have learned to say and hear a set of sounds (the phonetic inventory) that is unique to that language and have learned a set of rules for ordering these sounds (phonotactic rules) and producing them in contexts (allophonic rules). The more than 5000 living languages exhibit a diverse set of phonetic inventories ranging in size from 11 sounds in languages such as Rotokas and Mura to more than 140 sounds in languages such as !Xu (see Maddieson, 1984). Languages also have quite different phonotactic and allophonic rule systems. So, for example, the vowel in the English word 'cat' does not occur in the Japanese phonetic inventory, nor do Japanese phonotactics permit a syllable final /t/. English voiceless stops such as /p/ have aspirated and unaspirated allophones whereas in other languages such as Bengali, Korean, and Sesotho aspiration is used contrastively. Words beginning with the aspirated versus unaspirated versions of /p/ will differ in meaning in these languages.

When you speak, listen to, and read your native language you access this sound representation or phonology and this processing is carried out by dedicated neural systems. Phonetics and phonology are two aspects of the same sound system. Traditionally phonetics refers to the physical realization of the more abstract phonological categories. While there is considerable overlap in the domains of these two components of language, this review will primarily focus on the the more abstract phonological representation and its use in spoken language processing. The localization of this phonological function has been at the heart of neurolinguistics since Broca and Wernicke first described patients with language production and perception disorders. In this article, the neural basis of phonological representation will be summarized and the contributions from recent imaging studies will be reviewed.

CLINICAL EVIDENCE

One of the earliest ideas about neurolinguistics was that separate components of the language system could be selectively impaired. While this selectivity is rare or perhaps nonexistent, it is clear that there are disproportionate deficits associated with particular lesions. There is abundant evidence that phonology is affected in this manner by neurological damage. Although it is difficult to distinguish the representation of phonology from processes that use this knowledge, lesions in a few areas have long been known to impair phonological aspects of language use. These include the left inferior frontal gyrus and the left superior temporal gyrus, extending into the supramarginal and angular gyri. Lesions in these areas are associated with well-known aphasic syndromes (Broca's aphasia, Wernicke's aphasia, conduction aphasia) that include phonological impairment as part of their spectrum of deficit.

Traditionally, anterior aphasias with frontal lobe lesions in Brodmann's areas (BA) 44 and 45 have been described as having a deficit in phonological output. Language output by these patients is slow and labored with many sound errors. Posterior lesions in the superior posterior temporal regions

have been characterized as producing problems with receptive phonology. Language output by these patients is generally fluent, but there are severe deficits in comprehension. This anterior–posterior typology, however, has been questioned on a number of grounds.

Both anterior and posterior aphasias show disturbances in phonological aspects of language input and output (Caplan, 1992). In experimental speech perception tasks, both anterior and posterior aphasias show deficits in judging phonological contrasts such as the first consonant in 'bear' vs. 'pear' (e.g. Blumstein *et al*., 1977). In experimental production tasks, all aphasias show difficulties in producing the correct phonological forms of words. These difficulties include problems selecting and ordering the sounds in words (e.g. Blumstein, 1973) with both anterior and posterior aphasias having a large number of sound substitution errors (Blumstein, 1998). While anterior and posterior aphasias are certainly distinguishable on the basis of patterns and types of phonetic and phonological errors, there is significant phonological involvement in both of these diagnostic categories of aphasia.

At first glance this broad cortical involvement suggests that phonological processing must involve a complex and distributed neural network. However, it has become clear in recent years that language processing, including phonological processing, is subtly sensitive to the requirements of experimental and clinical tasks. Tasks such as silent noun generation, rhyme judgments, visual object naming, and spontaneous speech all involve phonological processes but differ in what aspects of phonology are involved, how the phonological representation is accessed, and in the number of other language processes involved (see, for example, the task analysis in Indefrey and Levelt, 2000). Because of this complexity, the boundaries of the phonetic, phonological, lexical, and semantic components of language are difficult to distinguish in the tasks that are commonly used to test phonological processing. These difficulties have played a significant role in modern functional neuroimaging.

RECENT CONTRIBUTIONS FROM FUNCTIONAL IMAGING

Since the onset of positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) of language (Petersen *et al*., 1988, 1989), the comparison of neural activation patterns between two tasks has been at the heart of experimental designs. Regional cerebral blood flow measures during a baseline task are commonly subtracted from or compared to the regional blood flow estimates during a target activation task. Baseline conditions are selected with the assumption that they differ from the activation condition only in the processing component that is being tested. The validity of this assumption has been questioned for the study of phonology (e.g. Price and Friston, 1997) and other linguistic activities (e.g. Jennings *et al*., 1997). Further, the experimental tasks utilized in various imaging studies differ on many dimensions and thus potentially involve different aspects of phonology. These task factors have created controversy about the exact localization of the neurological substrates of phonological processing (Démonet *et al*., 1996; Poeppel, 1996).

In spite of these methodological problems, functional imaging has yielded some consistent patterns for phonological activity. The posterior region of the superior temporal gyrus is one of the most common cortical areas associated with phonological tasks. Studies that focus on auditory speech processing have consistently shown posterior temporal lobe activations (e.g. Petersen *et al*., 1988; Démonet *et al*., 1992; Zatorre *et al*., 1992; Fiez *et al*., 1995; Price and Friston, 1997; Binder *et al*., 2000; Scott *et al*., 2000; Wise *et al*., 2001).

Binder and colleagues (e.g. Binder *et al*., 1996; Binder 1999) have proposed that speech processing follows a dorsal to ventral pattern in the temporal lobe. The preliminary auditory processing takes place in the primary auditory area in the superior temporal lobe. Further processing takes place in more ventral areas within the superior temporal sulcus (STS) where phonemic pattern recognition takes place. Activations of the STS have also been observed for lipreading and audiovisual speech perception (e.g. Calvert *et al*., 1997, 2000). More ventrally in the medial and inferior temporal lobe lexical and semantic processing are thought to take place. Others have confirmed this dorsal to ventral pattern (e.g. Scott *et al*., 2000; Wise *et al*., 2001) but also have suggested that the anterior–posterior axis in the temporal lobe differentiates different phonological functions. In Scott *et al*.'s study the anterior STS only showed activation for intelligible speech while the posterior STS showed activation for auditory phonetic cues independent of the intelligibility of the utterance.

Studies that focus on speech production and reading also show activity near the posterior temporal region. Indefrey and Levelt (2000) concluded from their review of 58 word production studies that the posterior superior temporal gyrus is involved in retrieving phonological structure during

word production. Further, there are a number of studies showing evidence of temporal lobe processing of auditory feedback during speech (see Munhall, 2001 for a review). Studies of reading show posterior temporal activations during phonological tasks as well (Fiez and Petersen, 1998) and also in areas slightly posterior to this (e.g., BA 40; Price, 1998).

As suggested by the lesion data, the inferior frontal cortex plays a significant role in reading and spoken language. There is good evidence that the frontal cortex can be subdivided into areas involved in semantic processing (anterior inferior frontal/prefrontal areas) and phonological processing (posterior inferior frontal cortex; Fiez, 1997; Fiez and Petersen, 1998). In this latter region, Broca's area and the surrounding inferior frontal lobe show activation during a range of different speech and phonological tasks (e.g., Petersen *et al*., 1988, 1989; Sergent *et al*., 1992; Zatorre *et al*., 1992, 1996; Démonet *et al*., 1992, 1994; Paulesu *et al*., 1993; Price and Friston, 1997; Fiez and Petersen, 1998; Poldrack *et al*., 1999, 2001; Burton *et al*., 2000).

Meta-analyses of these studies have led to proposals for further subdivisions of the posterior inferior frontal gyrus for different types of phonological processing. For example, Burton (2001) suggests that auditory speech processing involving segmentation shows activation in superior and posterior parts of the inferior frontal gyrus while visual phonological judgments (mostly rhyme decisions) activate areas slightly inferior and anterior to those areas involved in auditory speech processing (Fiez and Petersen, 1998). Speech production is strongly associated with the frontal operculum and anterior insula (Wise *et al*., 1999), but similar areas have been implicated in the translation of semantic information to a phonological form (Price, 1998) and the translation of orthography to phonology (Fiez and Petersen, 1998).

CONCLUSION

Considerable progress has been made in the last 10 years in understanding the anterior and posterior cortical regions involved in phonological activity; however, a range of problems is still unresolved. First, the roles played by the right hemisphere and subcortical structures in phonological processing are not well understood. Clinical studies show that lesions in the left hemisphere produce significantly more phonological disorders than right hemisphere lesions, but the functional imaging work has frequently shown bilateral activation for phonological tasks (Hickok and Poeppel, 2000). Subcortical structures are a significant part of the language network including phonological processing (Nadeau and Crosson, 1997), but much remains to be learned about their role. Clearly, for both the right hemisphere and subcortical structures further work will be necessary to understand their full contribution to phonological processing. Second, functional imaging experimental designs use many different baseline and activation tasks. A complete understanding of the neural basis of phonology awaits a more comprehensive description of what actually is involved in phonological processing during these tasks. Finally, much of the work that has been done has focused on the English language. Recent work by Paulesu *et al*. (2000) demonstrates different patterns of activation for reading and naming tasks in Italian and English speakers. These results suggest that not only must we understand what is involved in phonological tasks but we must understand tasks within the context of a specific language system.

References

Binder JR (1999) Functional MRI of the language system. In: Moonen CTW and Bandettini PA (eds) *Functional MRI*. Berlin: Springer.

Binder JR, Frost JA, Hammeke TA, Rao SM and Cox RW (1996) Function of the left planum temporale in auditory and linguistic processing. *Brain* **119**: 1239–1247.

Binder JR, Frost JA, Hammeke TA *et al*. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex* **10**: 512–528.

Blumstein SE (1973) *A Phonological Investigation of Aphasic Speech*. The Hague: Mouton.

Blumstein SE (1998) Phonological aspects of aphasia. In: Sarno M (ed.) *Acquired Aphasia*, 3rd edn. New York: Academic Press.

Blumstein SE, Baker E and Goodglass H (1977) Phonological factors in auditory comprehension in aphasia. *Neuropsychologia* **15**: 19–30.

Burton MW (2001) The role of inferior frontal cortex in phonological processing. *Cognitive Science* **25**: 695–709.

Burton MW, Small SL and Blumstein SE (2000) The role of segmentation in phonological processing: an fMRI investigation. *Journal of Cognitive Neuroscience* **12**: 679–690.

Calvert G, Bullmore E, Brammer M *et al*. (1997) Activation of auditory cortex during silent lipreading. *Science* **276**: 593–596.

Calvert G, Campbell R and Brammer M (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology* **10**: 649–657.

Caplan D (1992) *Language: Structure, Processing, and Disorders*. Cambridge, MA: MIT Press.

Démonet JF, Chollet F, Ramsay S *et al*. (1992) The anatomy of phonological and semantic processing in normal subjects. *Brain* **115**: 1753–1768.

Démonet JF, Fiez JA, Paulesu E *et al*. (1996) PET studies of phonological processing: a critical reply to Poeppel. *Brain and Language* **55**(3): 352–379.

Démonet JF, Price C, Wise R and Frackowiak RJ (1994) A PET study of cognitive strategies in normal subjects during language tasks: influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain* **117**: 671–682.

Fiez JA (1997) Phonology, semantics, and the role of the left inferior prefrontal cortex. *Human Brain Mapping* **5**: 79–83.

Fiez JA, Raichle ME, Miezin FM *et al*. (1995) PET studies of auditory and phonological processing: effects of stimulus characteristics and task demands. *Journal of Cognitive Neuroscience* **7**: 357–375.

Fiez JA and Petersen SE (1998) Neuroimaging studies of word reading. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 914–921.

Hickok G and Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* **4**: 131–138.

Indefrey P and Levelt WJ (2000) The neural correlates of language production. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Jennings JM, McIntosh A, Kapur S, Tulving E and Houle S (1997) Cognitive subtractions may not add up: the interaction between semantic processing and response mode. *Neuroimage* **5**: 229–239.

Levelt WJ, Roelofs A and Meyer AS (1999) A theory of lexical access in speech production. *Behavior and Brain Sciences* **22**: 1–38.

Maddieson I (1984) *Patterns of Sounds*. Cambridge, UK: Cambridge University Press.

Mummery CJ, Ashburner J, Scott SK and Wise RJS (1999) Functional neuroimaging of speech perception in six normal and two aphasic subjects. *Journal of the Acoustic Society of America* **106**: 449–457.

Munhall KG (2001) Functional imaging during speech production. *Acta Psychologica* **107**: 95–117.

Nadeau S and Crosson B (1997) Subcortical aphasias. *Brain and Language* **58**: 355–402.

Paulesu E, Frith CD and Frackowiak RJ (1993) The neural correlates of the verbal component of working memory. *Nature* **362**: 342–345.

Paulesu E, McCrory E, Fazio F *et al*. (2000) A cultural effect on brain function. *Nature Neuroscience* **3**: 91–96.

Petersen SE, Fox PT, Posner MI, Mintum M and Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single word processing. *Nature* **311**: 585–589.

Petersen SE, Fox PT, Posner MI, Mintum M and Raichle ME (1989) Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience* **1**: 153–170.

Poeppel DA (1996) Critical review of PET studies of phonological processing. *Brain and Language* **55**: 317–351.

Poldrack RA, Wagner AD, Prull MW *et al*. (1999) Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage* **10**: 15–35.

Poldrack RA, Temple E, Protopapas A *et al*. (2001) Relations between the neural bases of dynamic auditory processing and phonological processing: evidence from fMRI. *Journal of Cognitive Neuroscience* **13**: 687–697.

Price CJ (1998) The functional anatomy of word comprehension and production. *Trends in Cognitive Sciences* **2**: 281–288.

Price CJ and Friston KJ (1997) Cognitive conjunctions: a new approach to brain activation experiments. *Neuroimage* **5**: 261–270.

Scott SK, Blank CC, Rosen S and Wise RJ (2000) Indentification of a pathway for intelligible speech in the left temporal lobe. *Brain* **123**: 2400–2406.

Sergent J, Zuck E, Lévesque M and MacDonald B (1992) Positron emission tomography study of letter andobject processing: empirical findings and methodological considerations. *Cerebral Cortex* **2**: 68–80.

Wise RJ, Greene J, Buchel C and Scott SK (1999) Brain regions involved in articulation. *The Lancet* **353**: 1057–1061.

Wise RJ, Scott SK, Blank SC *et al*. (2001) Separate neural subsystems within 'Wernicke's area'. *Brain* **124**: 83–95.

Zatorre RJ, Evans AC, Meyer E and Gjedde A (1992) Lateralization of phonetic and pitch discrimination in speech processing. *Science* **256**: 846–849.

Zatorre RJ, Meyer E, Gjedde A and Evans AC (1996) PET studies of phonetic processes in speech perception: review, replication, and re-analysis. *Cerebral Cortex* **6**: 21–30.

Place Cells

Intermediate article

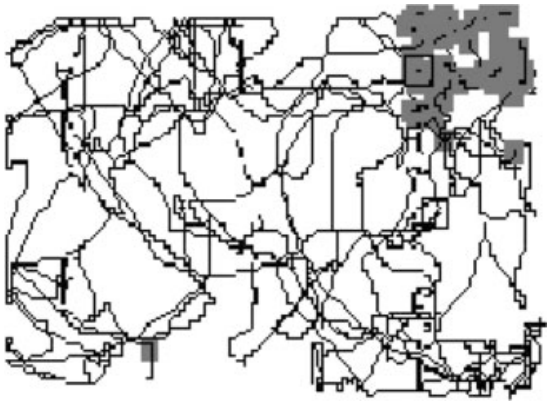
John O'Keefe, University College London, London, UK

CONTENTS

Research in the 1970s first suggested that neurons in the rat hippocampus signaled the animal's location in its environment. This prominent correlation with the animal's spatial location has led to these cells being termed place cells.

THE ROLE OF HIPPOCAMPAL PYRAMIDAL CELLS

In the early 1970s, O'Keefe and Dostrovsky reported that neurons in the rat hippocampus signaled the animal's location in its environment, and they suggested that these place-coded cells might provide the animal with a spatial map of its environment (O'Keefe and Dostrovsky, 1971). As the animal moves around a familiar environment, the typical hippocampal pyramidal cell will be silent for most of the time, only springing into activity each time the animal enters a delimited area. This prominent correlation with the animal's spatial location has led these cells to be called *place cells*, and the location where each cell fires is called its *place field*. An example is shown in Figure 1. The activity of this hippocampal unit was recorded as the animal searched for random pieces of food in a familiar rectangular-shaped environment. Place cells recorded in open-field environments fire regardless of the direction in which the animal is facing. Figure 2 shows the firing of a different place cell recorded while the rat foraged in a cylinder. The central panel shows the field regardless of the direction in which the animal is facing, and the four outer panels show the firing as the animal moved in different directions through the place field. One can rule out the possibility that the nondirectionality of this and similar cells is simply a response to olfactory or tactile cues on the floor of



(a)

(b)

Figure 1. [*Figure is also reproduced in color section*.] Place field of a hippocampal pyramidal cell. (a) Solid line shows the track of a rat foraging in a rectangular box for randomly located pieces of food. The red square shows the animal's location when the cell fires. (b) False color map of place field created by dividing the cell's firing rate in each location by the amount of time that the animal spent there.

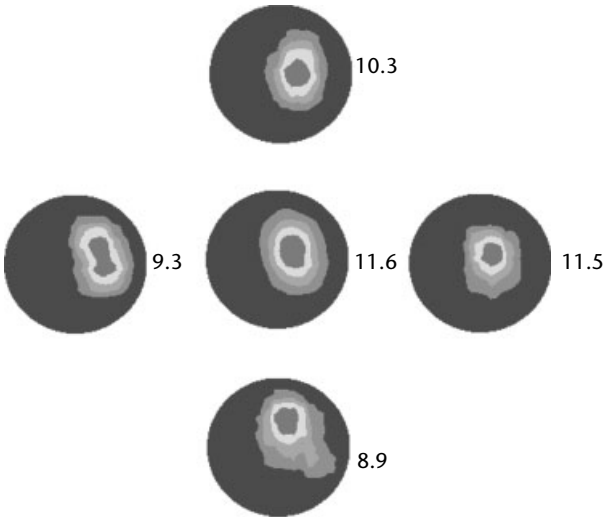


Figure 2. [*Figure is also reproduced in color section.*] Place cells are non-directional in open-field environments. The middle panel shows the firing field of a cell in a circular environment regardless of the direction in which the animal is facing. The four outer panels show the firing fields when the animal is facing north (upper), east (right), south (lower) and west (left). The number to the right of each panel is the firing rate in the hottest (red) region.

the enclosure, since the cells continue to fire in the same location when the floor is replaced or the box is rotated. We can conclude that, in open-field environments, the cells are signaling the animal's location rather than responding to some simple stimulus in that location.

Different cells identify different locations in an environment, and a collection of place cells taken together provides information about the animal's location at all points in the environment. Most place cells have a single compact place field in an environment, but a small proportion (*c*. 5%) have two or in rare cases even three such fields. Figure 3 shows the firing fields of 32 cells recorded simultaneously in a square-shaped environment. The firing fields have been organized so that cells with fields on the left-hand side of the box are shown on the left-hand side of the figure, and so on. It is clear that even in a small sample of cells there is one active field wherever the animal finds itself in the box. However, unlike the topographical representation of surfaces or receptors found in the neocortex, there appears to be no systematic relationship between the anatomical location of cells within the hippocampus and the places that they represent in an environment. Neighboring pyramidal cells are as likely to have fields distant from each other in an environment as to have fields close together (Redish *et al*., 2001). This simple basic fact points to a difference in reference frames between neocortical and hippocampal spatial representations.

In the neocortex, the reference frame is fixed to and centered on a sensory receptor surface such as the eye or the head, and there is systematic grouping of cells with similar response properties. These reference frames are collectively described as *egocentric*, since they are fixed to and move with the receptor surface. However, in the hippocampus no such fixed reference frame exists. This representation, which is not fixed to a body surface but to features of the environment, is described as *allocentric*. Cells are not systematically grouped within the hippocampus according to their place fields in an environment, or according to any other variable that has been identified. A given location is represented by a random subset of place cells distributed across the hippocampus. This may reflect the fact that the spatial relationships of the stimuli and features which make up a given environment cannot be known *a priori*, and the same cells can represent different locations in different environments. Therefore no predetermined topography can exist.

Place cells have been found in the CA3 and CA1 regions of the hippocampus proper, in the dentate gyrus, and in the subicular region and the entorhinal cortex. It is believed that the exteroceptive sensory information about environmental landmarks reaches these areas via the entorhinal cortex. The anatomical connections between these regions are such that the place representation in one area (e.g. CA1) may be built up through a series of stages (e.g. from EC through DG and CA3 to CA1), or it may be constructed independently on the basis of the information sent directly into each region separately from the entorhinal cortex. This anatomical organization suggests that under some circumstances, different parts of the hippocampal formation can act independently. Evidence for this view comes from experiments in which CA1 place cells can still be recorded following destruction of CA3 (Brun *et al*., 2002), and from experiments which show that CA1 place-cell activity can be dissociated from hippocampally dependent spatial learning (Jeffery *et al*., 2002). (*See* **Hippocampus**)

Not all cells have simple locational correlates. Although many complex-spike cells can be classified as simple place cells, others have more complex properties (O'Keefe, 1976). The firing rate of complex place cells is dependent on other factors in addition to location. For example, some cells increase their firing rates if the animal experiences a particular object or smell in the place field (Wood *et al*., 1999), or if it engages in a particular behavior

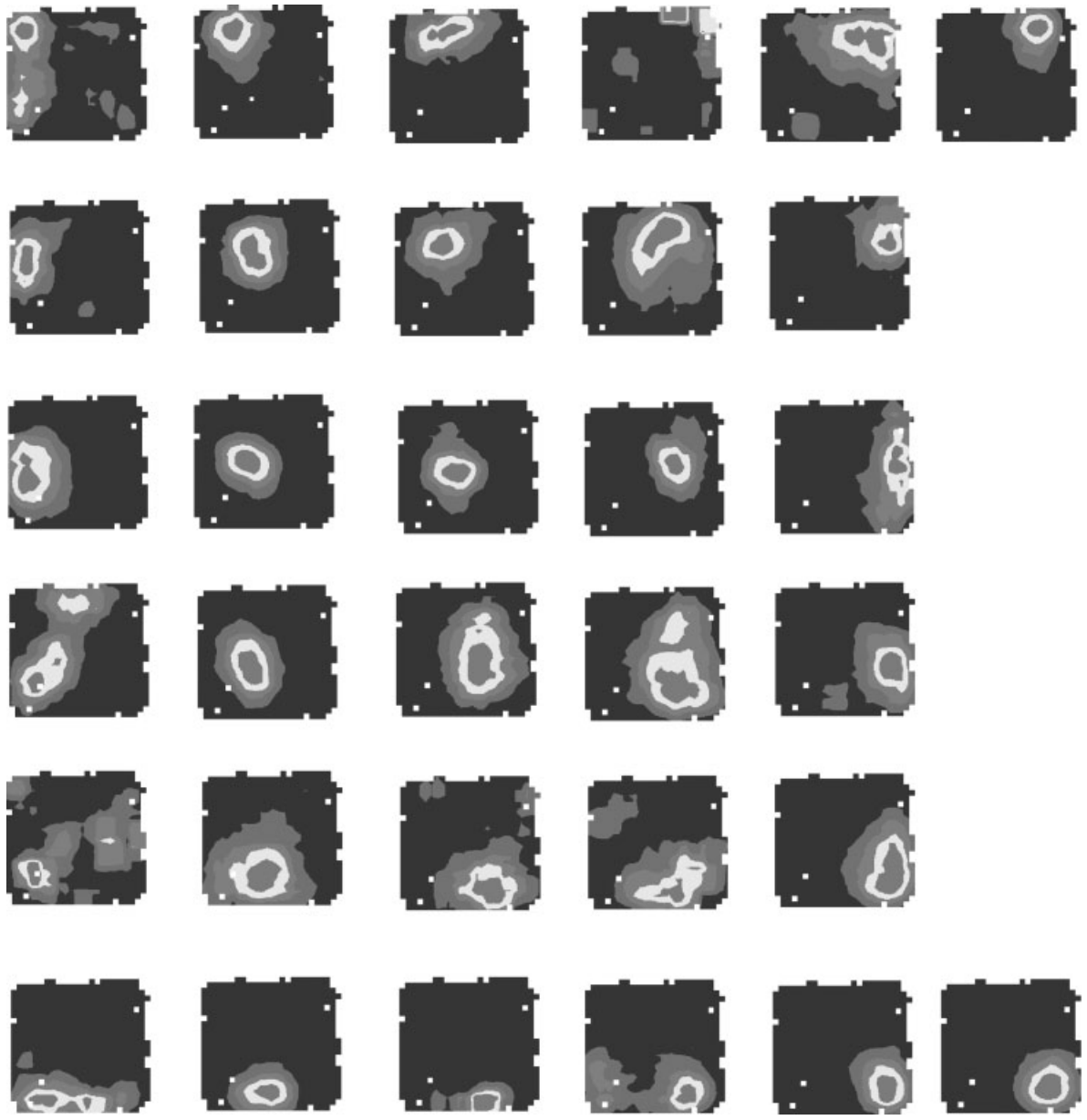


Figure 3. [*Figure is also reproduced in color section.*] Place fields of 32 simultaneously recorded hippocampal pyramidal cells. The firing fields have been organized so that fields located along the upper boundary of the testing box are located on the upper part of the figure, fields located along the left edge of the box are located on the left side of the figure, and so on. The 32 fields cover most of the environment. (After Lever *et al*., 2002.)

there. Others increase their firing rates when the animal either finds something new in the place field or fails to find something expected there. These cells are called *misplace* or *complex place* cells. As we have seen above, in open-field cylinders the place fields are non-directional (Figure 1). However, in more structured maze tasks they are highly directional. The same cells which have non-directional fields in a cylinder have directional fields in the radial arm maze (Muller *et al*., 1994). Although it is not known with certainty, the most important variable affecting directionality is likely to be whether the animal is free to face in different directions or to turn around in the place field.

Some place cells have fields in more than one environment. If the environments are different enough, there does not appear to be any relationship between the fields in terms of size, shape or location. The place cells are said to *re-map* when the animal moves from one environment to another, suggesting that the same cells participate in the representation of many environments, and that

any particular location is coded in the pattern of firing across many cells. One important exception to this rule is observed when similar environments are encountered initially, in which case there is a marked similarity between the fields in the two environments (see below).

THE COGNITIVE MAP THEORY

O'Keefe and Nadel (1978) suggested that a collection of place cells connected together in terms of the distance and direction of their fields from each other might form a spatial map. Such a map would identify the animal's location in an environment as well as the location of stimuli and objects in that environment, and would enable the animal to navigate from one location to another on the basis of any available route. As we shall see later, both directional and distance information is available to the place cells.

Are the place-cell fields hard-wired for each environment, or do they take time to develop, perhaps indicating that some learning process is involved? Wilson and McNaughton (1993) addressed this question and found that the place fields are established within the first 15 min. This suggests that any learning which takes place is very rapid, and that some aspects of the representation of an environment may perhaps be prefigured. For example, each cell may come with a small number of preselected environmental inputs to which it is sensitive, and exploratory behavior might involve the wiring together of these cells in accordance with the distance and direction between the parts of the environment that they represent. Evidence that the place fields of difference cells are linked together comes from experiments in which movements of spatial cues such as large visual stimuli at the edge of the environment cause the fields to move in concert (Muller and Kubie, 1987; Fenton and Muller, 2000). On the other hand, experiments in which changes in the shape of the environment cause place fields to move relative to each other show that the fields are not coupled to each other in any simple fashion which reflects a Euclidean metric (O'Keefe and Burgess, 1996). The form that the map takes and its exact location within the hippocampal formation are as yet unresolved.

SOURCES OF SPATIAL INFORMATION

Sources of spatial information can be external to an animal or internal, based on its own movements within the environment.

External or exteroceptive cues convey information about the large number of proximal and distal features and landmarks that are present in the environment. Internal or interoceptive stimuli are derived from idiothetic or path integration information generated by the animal's own movements. Once an animal has located itself in an environment on the basis of external sensory information, it could use this idiothetic system to calculate its subsequent angular rotations and linear translations on the basis of the predicted effect of its own movements or the optokinetic, proprioceptive and vestibular feedback from such movements.

IDENTIFICATION OF LANDMARKS

Place cells use information from different sensory modalities to identify landmarks or features of the environment. These cells are not totally dependent on any specific sensory modality for the identification of landmarks. They receive information from more than one sensory input, and they can use subsets of this total input to identify correctly their preferred place. Elimination of sensory modalities by occlusion or lesion indicates that there are still place cells after the elimination of olfactory, visual or combined visual and auditory inputs. However, vision appears to be the most important modality, even in such a relatively poorly endowed animal as the rat. The majority of place fields remain intact if the animal is placed in an environment with the lights on and the latter are then turned off (Quirk *et al*., 1990). However, if the animal is removed from the environment and then returned to it in the dark, the fields are altered in more than half of the cells tested. Furthermore, new fields can appear in the dark in the latter condition, and these are maintained even if the lights are then turned on with the rat in the environment. It appears that in the open-field cylinder or rectangle, once a pattern of place fields has been set up for a particular environment it is difficult to disrupt it without removing the animal from the environment. Another approach to the question of which cues control the place fields is to construct environments in which the distal spatial stimuli are explicitly identified and controlled (O'Keefe and Conway, 1978; Muller and Kubie, 1987). A strong visual cue or a set of four mixed modality cues (e.g. cards, lights, fans) hung around the periphery of the testing environment gain control over the angular orientation of the place fields. Rotation of the cues rotates the fields by the same amount, and removal of the cues leads to an unpredictable rotation of the field when the animal is returned to the

environment in most of the cells. In experiments where several distal spatial cues are available, removal of one or more of these shows that, while some fields are dependent on one of the four cues, the majority are maintained as long as any two of the four cues are present. Neither the shape of the field nor its distance from the walls of the box are affected in these experiments. Cues at the edge of a symmetrical environment or beyond can control the angular location of the place field, but they have much less control over its shape or its radial distance from the walls of the environment.

In contrast, the shape of the fields is primarily controlled by the walls or other barriers in the testing box or the room itself. This was demonstrated by experiments in which these walls or barriers were systematically moved relative to each other and the effect on place-field location and shape was studied (O'Keefe and Burgess, 1996; Gothard *et al*., 1996). In the O'Keefe and Burgess experiment, the same cell was studied while the rat foraged for food in four different-shaped boxes (two squares, one double the dimensions of the other, and two rectangles, each with a side equal to one of the sides of the squares). The shape and location of most fields were determined by the distance to two or more walls in orthogonal directions. For some cells, moving a wall moved the field without changing its shape, while for others, increasing the distance between two opposite walls caused the field center to shift so as to maintain the correct distance to both walls. This distorted the field shape, causing it to expand along that dimension, and it sometimes resulted in double peaks or the splitting of the field into two.

One important conclusion that can be drawn from the persistence of the dark-formed place fields when the lights are subsequently turned on in the cylinder (see above) is that visual cues cannot be the only cues that control place fields. Recent research suggests that, under these circumstances, the fields are maintained by olfactory and tactile cues on the floor and walls of the enclosure, and by internal idiothetic cues, in particular vestibular and proprioceptive ones, providing information about the animal's own movements.

Animals that were temporarily deprived of both visual and auditory sensory modalities (Hill and Best, 1981) and those that were blind from birth (Save *et al*., 1998) still had apparently normal place fields, and these were shown to be dependent on tactile and olfactory cues in combination with idiothetic information. In the experiment by Save *et al*. (1998), three large objects were placed inside the enclosure to allow the animal to orient itself, and the place fields only appeared after the animal had contacted the objects. In the experiment by Hill and Best (1981), many of the fields were disrupted by rotating the animal prior to placing it in the environment, which suggested that vestibular information was crucial. An alternative way to demonstrate this is by controlled rotation of the testing box with the animal in it, or by controlled rotation of the animal itself. In one experiment, the cylinder walls or floor were rotated separately and the manipulations were performed at either fast or slow speed, and either in the light or in the dark. It was assumed that the slow rotations were below the speed detectable by the vestibular system. The results showed that both visual and vestibular signals for rotation influence the angular location of place cells, and that the relative strength of these two influences is about equal (Sharp *et al*., 1995). Another experiment showed that slow rotation of the rat in a separate chamber outside the testing environment could also rotate the place fields by the same angle as that through which the animal was rotated (Jeffery *et al*., 1997).

PLACE CELLS ARE DEPENDENT ON HEAD DIRECTION INPUTS

Where does the directional information on which these place responses are based come from? How are the place fields being rotated around the center of the environment in the experiments described above? The most likely answer is that their angular position in an environment is dependent on inputs from the head direction system, originally discovered by Ranck and studied extensively by Taube and colleagues (Taube *et al*., 1990a, b). This second class of spatial cells is found in several parts of the brain, including the presubicular subdivision of the hippocampal formation. These cells have properties which complement those of the place cells – they ignore the animal's location, and instead signal its heading direction relative to the environmental frame. Each cell has its own particular prefered direction, and a number of these cells taken together provide information about the animal's heading across the entire 360° of the compass.

Evidence that the place cells are dependent on head-direction inputs comes from experiments by Muller and colleagues (1987), who showed that environmental manipulations which affect the place cells also affect the head-direction cells. For example, rotating prominent environmental landmark cues, such as large visual stimuli on the edge of the testing box, causes the fields of both the place

cells and the head-direction cells to rotate by an amount which maintains their spatial relationship to these stimuli. In an experiment in which place cells and head-direction cells were recorded simultaneously during manipulations which disoriented the animal's sense of direction, Knierim and colleagues (Knierim *et al*., 1995) showed that the two cell types maintained a fixed spatial relationship to each other even as they rotated relative to the environmental frame.

DETERMINANTS OF PLACE-CELL DISTANCE FROM LANDMARKS

What is the sensory source of distance, as opposed to directional, information to the hippocampal place cells? By analogy with the sources of *directional* control of these cells discussed above, information about *distances* from environmental landmarks might be given by optokinetic, proprioceptive and vestibular cues in addition to, or instead of, the distal visual cues. A good estimate of the linear distance traveled from an environmental landmark could be given by a visual cue such as the change in the angle of the cue above or below the horizon, or the size of the cue on the retina. Alternatively, the distance traveled could be derived by integrating the animal's speed of movement in a particular direction over time. At present there are three pieces of evidence which support the idea that distance could be represented in the hippocampus by information about speed of movement. First, O'Mara and his colleagues have described cells in the monkey hippocampus which respond to passive rotations or translations of the whole animal (O'Mara *et al*., 1994). Second, a small number of active speed cells have been recorded in the rat hippocampus (O'Keefe *et al*., 1998). The firing rate of these speed cells is a linear function of the speed at which the rat runs. Direction of movement appears to be irrelevant, and these cells are not activated by passive movement of the animal. Integration of the firing rate of such a cell over time would provide the distance information that is required by the place cells. Third, Czurko and colleagues have recently reported that the firing rates of the place cells themselves code for the speed of running in the place field (Czurko *et al*., 1999). They recorded from place cells with fields in a running wheel, and showed that the firing rates of the cells increased linearly with the speed of running in the wheel. Since there is no movement of the animal's head relative to the environment in this experiment, the speed signal cannot be coming from the vestibular system or from optic flow information, but must be derived from the motor system itself.

THE ROLE OF MOTIVATION, PROBLEM-SOLVING HYPOTHESES AND GOAL LOCATION

Place-field firing does not seem to depend on the reason why the animal runs through the location, as can be shown by the simple demonstration that interchanging the rewards at the ends of the different arms of a maze has no effect on place-cell firing. As is shown in Figure 3, in many environments the distribution of place fields is fairly even, with all regions being represented equally. However, there have been reports of fields shifting when rewards are moved (Breese *et al*., 1989), and of increased representation of rewarded locations (Hollup *et al*., 2001). Whether the animal is using its hippocampus to solve a spatial problem or not seems to have an influence on the number of cells with place fields in an environment (Zinyuk *et al*., 2000).

TEMPORAL PROPERTIES OF PLACE-CELL FIRING

Several studies have examined the temporal firing patterns of the pyramidal cells and sought to determine whether they convey any information. It has been known for some time that the firing of both the complex-spike cells and the theta cells bears a temporal relationship to the ongoing EEG theta wave. Theta activity occurs in the hippocampal EEG of the rat when the animal engages in behaviors which change its location in the environment. These include walking, exploring, swimming and jumping.

On the other hand, quite vigorous motor behaviors which do not change the animal's location in the environment, such as grooming, are accompanied by a different EEG pattern, with large-amplitude, non-rhythmical slow-wave activity (LIA). In addition, in some species low-frequency theta activity occurs in immobile animals in response to a stimulus such as a sound. However, in the rat it will only occur during immobility if the animal is highly aroused (e.g. in the presence of a predator) (Sainsbury *et al*., 1987).

During theta activity, all active hippocampal cells fire in bursts with the same frequency as or one close to that of the EEG theta rhythm. The theta cells do so whenever there is theta activity in the hippocampal EEG, and they always maintain a fixed correlation to a particular phase of the sinusoidal EEG wave. However, the phase relationship

of the place cells to the theta wave is different. O'Keefe and Recce (1993) studied the correlation between the two as the rat ran through the place field. In contrast to the theta cells, it was found that the phase correlation of the place cells did not remain constant, but changed in a systematic way. Each time the rat entered the field, the cell always began to fire on the same phase of theta. As the rat progressed through the field, the bursts of spikes occurred at an earlier phase of each successive cycle (Figure 4a). The phase of firing was strongly correlated with the animal's location within the place field (Figure 4b), and this correlation was stronger for position than it was for time after entry into the field. The location of an animal in an environment is coded by the temporal firing of the cells relative to the EEG theta wave, as well as by their absolute firing rate. The phase precession phenomenon is explained partly by the fact that the place-cell bursts occur at a higher frequency than the frequency of the concurrent EEG theta wave, and partly by the fact that there must be some adjustment in the relationship between the interburst frequency of the cells and the frequency of the EEG theta wave which compensates for the speed with which an animal runs through the field. Jensen and Lisman (2000) have shown that taking the phase correlate into account increases the spatial resolution of place cells by 43% above that provided by the firing rate alone.

PLACE-CELL PLASTICITY AND ITS ROLE IN SPATIAL MEMORY

Interest in the hippocampus largely stems from the belief that it is crucial for the formation and/or storage of long-term memory. There are three good examples of plastic changes in the place cells

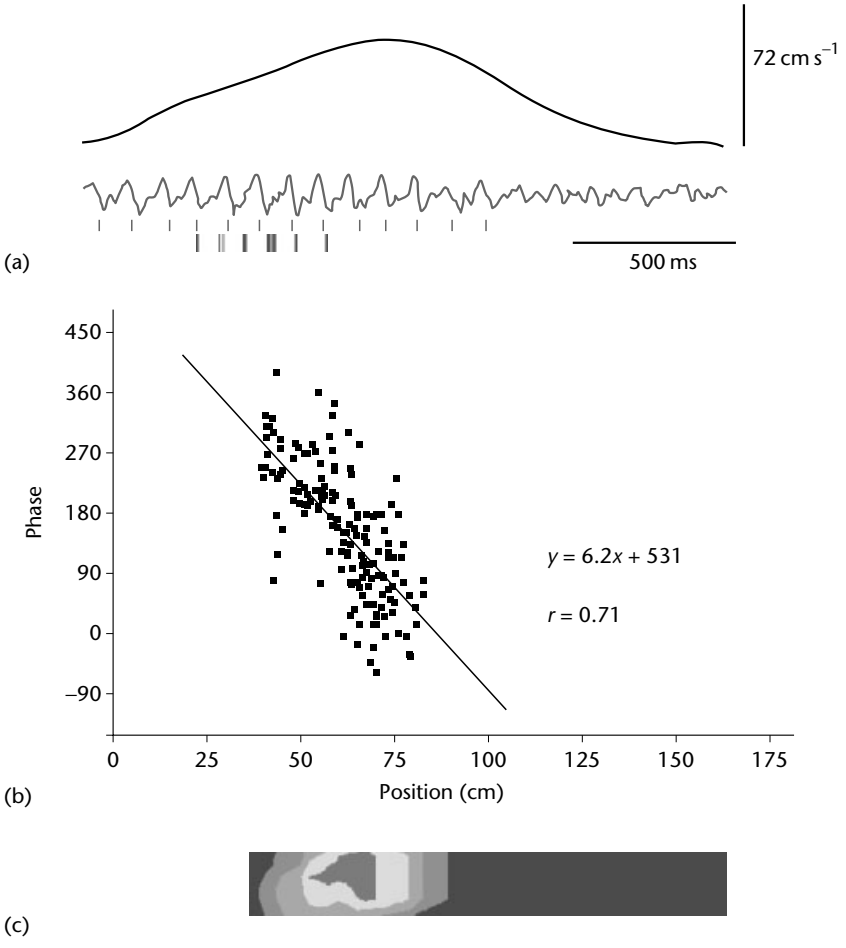


Figure 4. [*Figure is also reproduced in color section.*] Hippocampal place-cell activity changes phase with EEG theta activity. (a) Velocity profile, EEG theta rhythm and place-cell firing during a single run on a linear track. Ticks below EEG mark 0°/360°. (b) Phase of theta at which unit fires versus position on the track. (c) Location of place field on the track. (After O'Keefe and Burgess, 1999.)

which may form the bridge to the role of the hippocampus in human spatial and episodic memory, namely the short-term changes seen in the experiment of O'Keefe and Speakman (1987), the day-long changes seen in the experiment of Mehta *et al*. (1997, 2000) and the long-term longer-lasting changes observed in the experiment of Lever *et al* (2002). Finally, there is evidence that the *N*-methyl-D-aspartate (NMDA) receptor which has been implicated in long-term potentiation (LTP) also plays a role in the long-term stability of place-field firing.

O'Keefe and Speakman (1987) demonstrated that place cells exhibited a form of short-term memory in a delayed match-to-position task. Once place fields had been established on the basis of distal spatial cues, they were maintained after the removal of those cues. Tests on normal rats had previously established that well-trained animals could remember the location of the goal for periods of as long as 30 min after the cues had been removed, and lesions which disrupted hippocampal function resulted in a profound memory deficit. One suggested mechanism is a rotation of the head-direction system into alignment with the spatial cues, and a maintenance of that preferred direction after the removal of the cues. As we have seen, the orientation of the head-direction system controls the angular orientation of the place cells, and the maintenance of the new preferred head direction would be reflected in place-field location.

The second type of plasticity appears to last for at least a day. It has been demonstrated that place fields can change shape over a series of trials in which the animal always crosses the place field in the same direction. The change consists of an elongation of the field in the direction from which the animal is moving. The fact that the place fields have returned to their original shapes by the next day suggests that the change is temporary, lasting for less than a day (Mehta *et al*., 1997, 2000). Of particular interest is the fact that the change is NMDA-receptor-dependent (Ekstrom *et al*., 2001), which suggests that its synaptic basis may be the same as LTP.

The third type of plasticity takes the form of a long-term incremental change in the place fields, which may serve as the basis of long-term incidental spatial memory (Lever *et al*., 2002). Place fields in a square and circle were originally similar on initial exposure to the environments, but began to differentiate following prolonged experience. This differentiation (or re-mapping) persisted for at least a month without intervening exposure to the boxes.

Finally, NMDA-receptor antagonists do not affect the ability of place cells to differentiate between two different enclosures, but they do prevent the long-term stabilization of that differentiation (Kentros *et al*., 1998).

COMPLEX-SPIKE-CELL FIRING DURING SLEEP MAY BE MODULATED BY PRIOR SPATIAL LEARNING EXPERIENCES

It has been suggested that consolidation of memories might take place during sleep, and in particular during rapid eye movement. If this were the case, one might expect neuronal firing patterns to reflect this consolidation activity. It has also been suggested that consolidation involves the transfer of information from the hippocampus to the neocortex. In this case, one might expect an increased interaction between cells in these areas during and after sleep. McNaughton and his colleagues studied changes in connectivity between hippocampal neurons and found an increase between pairs of cells with overlapping place fields on a circular maze after the animal was allowed to run in the maze (Qin *et al*., 1997). In subsequent research there was also some indication that the temporal ordering of firing between cells with partially overlapping fields was replicated in a compressed form during sleep. Finally and most intriguingly, there was an increase in connectivity between pairs of hippocampal and neocortical cells. Wilson and his colleagues have gone one step further in this search for a reflection of the previous day's experience in the hippocampal unit activity during sleep (Louie and Wilson, 2001). They have found that when an animal is trained to run round a maze in a repetitive manner during the day, place cells become activated in the same sequence during REM sleep at night. This strongly suggests that the actual sequence of events in the maze is being replayed during dreaming sleep.

SUMMARY

The animal's location is the strongest correlate of hippocampal pyramidal-cell firing in animals such as the rat. A collection of these place cells may provide the animal with a spatial map of its environment, enabling it to navigate in a flexible, creative manner. Place cells identify locations both on the basis of environmental landmarks using visual, olfactory and tactile information, and on the basis of self-generated vestibular and proprioceptive cues which provide information about the distances and directions that the animal has moved since the last known location. One of the strongest

sources of place information is the distance of the animal in particular directions from large environmental features such as the walls of the enclosure. Directional information is provided by the head-direction cells of the presubiculum, which fire when the animal faces in a particular direction, regardless of its location in the environment. The time of firing of the place cells provides information about the animal's location in addition to the rate of firing. As might be expected from the role of the hippocampus in memory, place cells exhibit several types of plasticity on timescales lasting from a few minutes to several weeks. There is evidence that the events of the day influence the activity of hippocampal pyramidal cells during subsequent dreaming sleep.

References

Breese CR, Hampson RE and Deadwyler SA (1989) Hippocampal place cells: stereotypy and plasticity. *Journal of Neuroscience* **9**: 1097–1111.

Brun VH, Otnass MK, Molden S *et al*. (2002) Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* **296**: 2243–2246.

Czurko A, Hirase H, Csicsvari J and Buzsaki G (1999) Sustained activation of hippocampal pyramidal cells by 'space clamping' in a running wheel. *European Journal of Neuroscience* **11**: 344–352.

Ekstrom AD, Meltzer J, McNaughton BL and Barnes CA (2001) NMDA receptor antagonism blocks experience-dependent expansion of hippocampal 'place fields'. *Neuron* **31**: 631–638.

Fenton AA, Csizmadia G and Muller RU (2000) Conjoint control of hippocampal place cell firing by two visual stimuli. I. The effects of moving the stimuli on firing field positions. *Journal of General Physiology* **116**: 191–210.

Gothard KM, Skaggs WE and McNaughton BL (1996) Dynamics of mismatch correction in the hippocampal ensemble code for space: interaction between path integration and environmental cues. *Journal of Neuroscience* **16**: 8027–8040.

Hill AJ and Best PJ (1981) Effects of deafness and blindness on the spatial correlates of hippocampal unit activity in the rat. *Experimental Neurology* **74**: 204–217.

Hollup SA, Molden S, Donnett JG, Moser MB and Moser EI (2001) Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *Journal of Neuroscience* **21**: 1635–1644.

Jeffery KJ, Donnett JG, Burgess N and O'Keefe JM (1997) Directional control of hippocampal place fields. *Experimental Brain Research* **117**: 131–142.

Jeffery KJ, Gilbert A, Burton S and Strudwick A (2002) Preserved performance in a hippocampal dependent spatial task despite complete place cell remapping. *Hippocampus* (in press).

Jensen O and Lisman JE (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *Journal of Neurophysiology* **83**: 2602–2609.

Kentros C, Hargreaves E, Hawkins RD *et al*. (1998) Abolition of long-term stability of new hippocampal place-cell maps by NMDA-receptor blockade. *Science* **280**: 2121–2126.

Knierim JJ, Kudrimoti HS and McNaughton BL (1995) Place cells, head direction cells and the learning of landmark stability. *Journal of Neuroscience* **15**: 1648–1659.

Lever C, Wills T, Cacucci F, Burgess N and O'Keefe J (2002) Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature* **416**: 90–94.

Louie K and Wilson MA (2001) Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**: 145–156.

Mehta MR, Barnes CA and McNaughton BL (1997) Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the National Academy of Sciences of the USA* **94**: 8918–8921.

Mehta MR, Quirk MC and Wilson MA (2000) Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25**: 707–715.

Muller RU and Kubie JL (1987) The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience* **7**: 1951–1968.

Muller RU, Bostock E, Taube JS and Kubie JL (1994) On the directional firing properties of hippocampal place cells. *Journal of Neuroscience* **14**: 7235–7251.

O'Keefe J (1976) Place units in the hippocampus of the freely moving rat. *Experimental Neurology* **51**: 78–109.

O'Keefe J and Burgess N (1996) Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**: 425–428.

O'Keefe J and Burgess N (1999) Theta activity, virtual navigation and the human hippocampus. *Trends in Cognitive Science* **3**: 403–406.

O'Keefe J, Burgess N, Donnett JG, Jeffery KJ and Maguire EA (1998) Place cells, navigational accuracy and the human hippocampus. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **353**: 1333–1340.

O'Keefe J and Conway DH (1978) Hippocampal place units in the freely moving rat: why they fire where they fire. *Experimental Brain Research* **31**: 573–590.

O'Keefe J and Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Brain Research* **34**: 171–175.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

O'Keefe J and Speakman A (1987) Single-unit activity in the rat hippocampus during a spatial memory task. *Experimental Brain Research* **68**: 1–27.

O'Keefe J and Recce ML (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**: 317–330.

O'Mara SM, Rolls ET, Berthoz A and Kesner RP (1994) Neurons responding to whole-body motion in the

primate hippocampus. *Journal of Neuroscience* **14**: 6511–6523.
Qin YL, McNaughton BL, Skaggs WE and Barnes CA (1997) Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **352**: 1525–1533.
Quirk GJ, Muller RU and Kubie JL (1990) The firing of hippocampal place cells in the dark depends on the rat's recent experience. *Journal of Neuroscience* **10**: 2008–2017.
Redish AD, Battaglia FP, Chawla MK *et al*. (2001) Independence of firing correlates of anatomically proximate hippocampal pyramidal cells. *Journal of Neuroscience* **21**: RC134.
Sainsbury RS, Heynen A and Montoya CP (1987) Behavioral correlates of hippocampal type 2 theta in the rat. *Physiology and Behavior* **39**: 513–519.
Save E, Cressant A, Thinus-Blanc C and Poucet B (1998) Spatial firing of hippocampal place cells in blind rats. *Journal of Neuroscience* **18**: 1818–1826.
Sharp PE, Blair HT, Etkin D and Tzanetos DB (1995) Influences of vestibular and visual motion information on the spatial firing patterns of hippocampal place cells. *Journal of Neuroscience* **15**: 173–189.
Taube JS, Muller RU and Ranck JB Jr (1990a) Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience* **10**: 420–435.
Taube JS, Muller RU and Ranck JB Jr (1990b) Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *Journal of Neuroscience* **10**: 436–447.
Wilson MA and McNaughton BL (1993) Dynamics of the hippocampal ensemble code for space. *Science* **261**: 1055–1058.
Wood ER, Dudchenko PA and Eichenbaum H (1999) The global record of memory in hippocampal neuronal activity. *Nature* **397**: 613–616.
Zinyuk L, Kubik S, Kaminsky Y, Fenton AA and Bures J (2000) Understanding hippocampal activity by using purposeful behavior: place navigation induces place cell discharge in both task-relevant and task-irrelevant spatial reference frames. *Proceedings of the National Academy of Sciences of the USA* **97**: 3771–3776.

Further Reading

Best PJ, White AM and Minai A (2001) Spatial processing in the brain: the activity of hippocampal place cells. *Annual Review of Neuroscience* **24**: 459–486.
Taube JS (1998) Head directional cells and the neuropsychological basis for a sense of direction. *Progress in Neurobiology* **55**: 225–256.

Planning: Neural and Psychological

Intermediate article

Vinod Goel, York University, Toronto, Ontario, Canada

CONTENTS

To plan is to formulate a scheme or method for attaining some goal. Scientists are just beginning to understand the psychological processes and brain systems involved.

INTRODUCTION

Planning is a quintessential human cognitive activity that involves the mental formulation of future states of affairs. In this article, we shall differentiate between planning and general problem solving, further differentiate between ill-structured and well-structured planning, review the components of both types of planning, and summarize some of the data on the neural basis of planning.

DEFINITION AND CHARACTERIZATION OF PLANNING

Planning is a form of problem solving. Problem solving, according to most definitions, requires at least the following conditions: (1) there are two distinct states of affairs; (2) the agent is in one state and wants to be in the other state; (3) it is not apparent to the agent how the gap is to be bridged; (4) bridging the gap is a consciously guided (at least at the top executive level), multi-step process. Planning problems also require an agent to chart a path from A to B in a modeling space, without 'bumping' into the world. All of the 'bumping' must be done in the modeling space, and some satisfactory path extracted. That is, the whole idea of planning is that we want to know the consequences of an action before the action is executed. The only way to do this is to execute the action in a modeling space and 'observe' the consequences. If the results are satisfactory, the plan can be executed in the real world. If the results are not satisfactory, the plan can be revised. This definition tries to capture an interesting commonsense notion of planning, such that planning is neither a ubiquitous activity nor an overly narrow activity.

Planning in the world versus planning in the laboratory

Real-world planning problems are largely ill-structured problems – that is, problems where the information content of the start state, goal state, and transformation function is incompletely specified. For example, planning a meal for a guest is an example of an ill-structured task. The start state is incompletely specified (e.g., how hungry will they be? how much time and effort do I want to expend?). The goal state is also incompletely specified (e.g., how much do I care about impressing the guest? should there be three or four courses? would salmon be appropriate? would they prefer a barbecue or an indoor meal?). Finally, the transformation function is also incompletely specified (e.g., should I have the meal catered, prepare it myself, or ask everyone to bring a dish? if I prepare it, should I use fresh or frozen salmon?).

By contrast, most laboratory planning tasks are largely well-structured problems. They are characterized by the presence of information in each of the components of the problem vector. The Tower of Hanoi provides a relevant example. It is a disk transfer problem consisting of three pegs and a number of disks. The start state is completely specified (e.g., the disks are stacked in descending order on peg 1). There is a clearly defined test for the goal state (stack the disks in descending order on peg 3). The transformation function is restricted to moving disks within the following constraints: (1) only one disk may be moved at a time; (2) any disk that is not currently being moved must remain on a peg; (3) a larger disk may not be placed on a smaller disk.

Another very important difference between ill-structured and well-structured problems concerns the nature of the constraints in the two cases. In the Tower of Hanoi, as in all puzzles and games, the constraints of the task are logical or constitutive. That is, if one violates a constraint or rule, one is simply not playing that game. For example, if I place a larger disk on a smaller disk, I am simply not doing the Tower of Hanoi task. However, the constraints that we encounter in real-world planning situations are of a very different character. Some of these constraints are nomological, and many of them are social, political, economic, cultural, etc. Most of us quickly learn that these constraints are not definitional or constitutive of the task. On the contrary, they are negotiable/breakable, depending on the circumstances.

Ill-structured problems have no right or wrong answers, although there are certainly better and worse answers. In well-structured problems there are right and wrong answers, and clear ways of recognizing when they have been reached. Furthermore, most real-world problems have consequential costs associated with errors. Resources and lives are often at stake. In well-structured games, errors may cause some embarrassment to the subject, but that is about the extent of the 'damage.' Lastly, in many real-world situations there is no immediate feedback from the world. Therefore it must be simulated, or self-generated. This requires considerable resource allocation for modeling and performance prediction. In games like the Tower of Hanoi, there is genuine feedback after every operator application. However, it is local feedback, and the final solution needs to satisfy global constraints.

While both examples, planning a dinner and solving the Tower of Hanoi, can be viewed as planning tasks (though see Goel and Grafman (1995)), the above characterization makes it clear that the differences between them may be greater than the similarities, an issue we shall return to below.

COGNITIVE AND COMPUTATIONAL COMPONENTS OF PLANNING

Components of a well-structured planning system

The components of a planning system differ depending on whether we are talking about well-structured or ill-structured planning. Classic information-processing accounts of human problem solving are quite adequate to account for well-structured planning (Figure 1). Such systems typically involve an information-processing system (IPS), a task environment, and a problem space. An information-processing system is a computational system with a memory, a processor, and sensory receptors and motor effectors. There are actually three separate memories – a long-term memory, a short-term memory and an external memory. Each is characterized by its organization and its read/write times. The processor performs some basic elementary processes, such as read, write, test, compare, discriminate, and replace symbols, but there is no necessary or sufficient set.

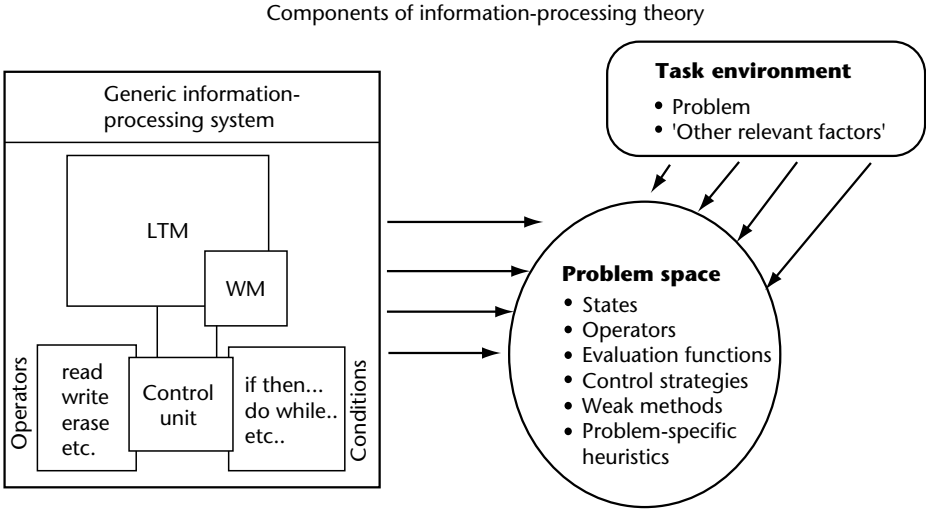


Figure 1. Characterization of a well-structured planning problem space. Abbreviations: LTM, long-term memory; WM, working memory.

These operations can be combined to perform any arbitrarily complex computation.

The task environment (TE) consists of a problem and 'any other relevant factors'. The problem space is a modeling space (constrained by the IPS and TE) where problem solving occurs as a computational process. It is defined by states, operators, evaluation functions and control strategies. States are symbolic representations of a problem at a given point in time. Operators are the procedures which transform one state into the next state. Evaluation functions measure the 'goodness' of the current state and guide the search. Weak method search strategies (e.g., means–ends analysis, breadth-first search, depth-first search) are hardwired into the information-processing system. Heuristic strategies use situation-specific knowledge to circumvent the search space. The strategy employed in any situation depends on the knowledge that the system has explicitly available.

Progress has been made in two directions with regard to understanding the cognitive and computational components of planning. Today there are many more sophisticated accounts of the structure of the information-processing system (e.g., connectionist and hybrid systems), and some progress has been made in specifying a set of mid-level computational procedures or 'modules' sufficient for planning tasks on top of the elementary operators of the information-processing system.

Components of an ill-structured planning system

The characterization of the problem space of ill-structured planning problems is more complex. Such problems have been characterized as having four component phases, namely *problem structuring*, *preliminary solutions*, *refinement* and *detailing* of solutions. Each phase differs with respect to the type of information that is dealt with, the degree of commitment to generated ideas, the level of detail that is attended to, the number and types of transformations that are engaged in, the mental representations needed to support the different types of information and transformations, and the corresponding computational mechanism. As one progresses from the preliminary phases to the detailing phases, the problem becomes more structured. This is depicted in Figure 2.

Problem structuring is an information-gathering phase in which explicit and implicit sources are used to fill in the information missing from the problem statement. Preliminary planning is a classic case of creative, ill-structured problem solving. It is a phase in which alternatives are generated and explored. This generation and exploration of alternatives is facilitated by the abstract nature of information being considered, a low degree of commitment to generated ideas, the coarseness of detail, and a large number of lateral transformations. A lateral transformation is one where movement is from one idea to a slightly different idea, rather than to a more detailed version of the same idea. Lateral transformations are necessary for the *widening of the problem space* and the exploration and development of kernel ideas.

The refinement and detailing phases are more constrained and structured. They are phases where commitments to a particular solution are made and then propagated through the problem space. They are characterized by the concrete nature of the information being considered, a high degree of commitment to generated ideas, attention to detail, and a large number of vertical transformations. A vertical transformation is one where

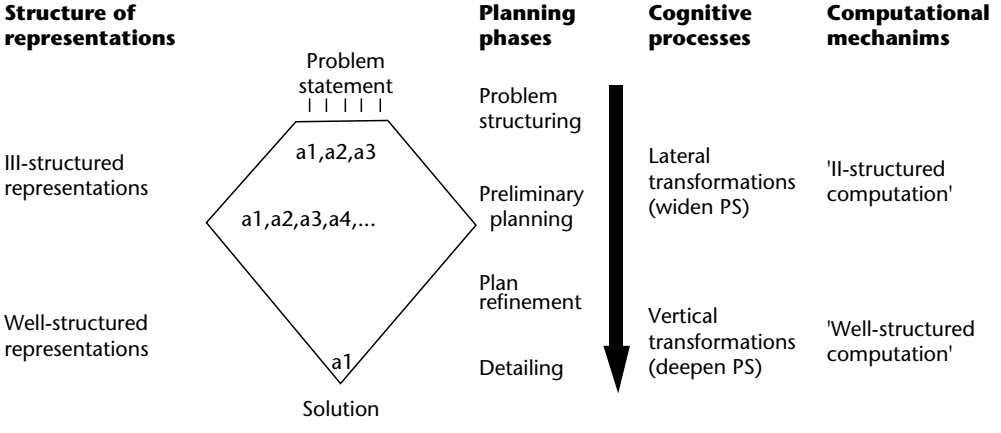


Figure 2. Characterization of an ill-structured planning problem space.

movement is from one idea to a more detailed version of the same idea. It results in a *deepening of the problem space*.

Lateral transformations are underwritten by mechanisms that support ill-structured mental representations, which are imprecise, ambiguous, fluid, indeterminate, vague, etc. In contrast, vertical transformations are underwritten by mechanisms that support well-structured mental representations, which are precise, distinct, determinate, and unambiguous. Furthermore, it has been argued that ill- and well-structured representations map onto different types of computational systems. This cognitive and computational dissociation between ill- and well-structured planning problems suggests a corresponding anatomical dissociation.

DISORDERS OF PLANNING FOLLOWING BRAIN DAMAGE

Disorders of planning have long been associated with frontal lobe lesions. The evidence comes both from informal anecdotal observations by physicians and clinicians and from laboratory experiments. Informal observations originate with Harlow's classic description of Phineas Gage, a railroad foreman who had a 4-foot iron tamping rod pass through his skull in a freak accident, and continue with Penfield's report on the behavior of his sister 15 months after the removal of her right frontal lobe. Harlow, a nineteenth-century physician who chronicled Gage's injury and subsequent behavior, noted that prior to the injury Gage was mild mannered and was 'considered a shrewd businessman, energetic and persistent in executing all his plans.' However, after the injury he would devise 'many plans of future operation but discard [ed] them before they were executed.' Penfield, upon observing his sister's inability to combine the individual components of a meal into a family dinner, noted that 'Although physical examination was negative and there was no change in personality or capacity for insight, nevertheless the loss of the right frontal lobe had resulted in an important defect. The defect produced was a lack of capacity for planned administration…'

Laboratory experiments on planning have been dominated by the two 'Tower tasks,' namely the Tower of Hanoi and the Tower of London. The Tower of Hanoi has been described above. The Tower of London is also a disk-transfer task, but without the constraint of not placing a larger disk on a smaller disk. The absence of the latter constraint, and the fact that the pegs accommodate different number of disks, makes the Tower of London a very different problem from the Tower of Hanoi (Goel and Grafman, 1995).

It is widely accepted that patients with lesions to the prefrontal cortex have difficulty completing the Tower of Hanoi (Goel and Grafman, 1995; Morris *et al*., 1997) and Tower of London (Owen *et al*., 1990) tasks. In fact, this connection seems so secure that the Tower tasks are considered to be 'frontal lobe' tasks, and they are used to determine frontal lobe involvement in a wide range of populations, including retarded children and adolescents, cerebellar patients, fragile X syndrome patients, Huntington disease patients, schizophrenics, and children with attention deficit hyperactivity disorder (ADHD).

During the past decade, a number of neuroimaging studies have confirmed the involvement of the prefrontal cortex in the Tower tasks. In positron emission tomography (PET) studies, the Tower of London task activates a large network, including prefrontal, cingulate, premotor, parietal and occipital cortices (Baker *et al*., 1996). In functional magnetic resonance imaging (fMRI) studies, the Tower of Hanoi task also activates the frontoparietal system, including the right dorsolateral and left inferior prefrontal cortex, bilateral parietal and premotor cortex (Fincham *et al*., 2002).

Recently, several experimental studies have used interesting real-world planning tasks to gain a better understanding of the role of the prefrontal cortex in ill-structured planning. Shallice and Burgess (1991) proposed and administered a Multiple Errands Test. Frontal lobe patients and normal controls were given a card with eight errands listed on it. These errands ranged from buying a loaf of bread to being at a certain place in 15 minutes' time, to finding yesterday's rate of exchange for the French franc. They were taken to an (unfamiliar) neighborhood near the hospital and asked to complete the errands within the following constraints: 'You are to spend as little money as possible (within reason) and take as little time as possible (without rushing excessively). No shop should be entered other than to buy something. Please tell one or other of us when you leave a shop what you have bought. You are not to use anything not bought on the street (other than a watch) to assist you. You may do the task in any order.' Frontal lobe patients engaged in more 'inefficient' actions (e.g., entering the same store twice), broke more rules (e.g., leaving a shop without paying), and failed to complete more tasks than normal controls.

Grafman and colleagues have used the 'scripts task' to explore planning deficits. Subjects are

requested to list as many actions as possible relevant to specified themes. The themes fall into three levels of familiarity, namely routine (e.g., preparing to go to work), nonroutine (e.g., taking a trip to Mexico) and novel (e.g., opening a beauty salon). After the list has been generated, subjects are asked to rank the actions in the correct order of execution. Finally, they are asked to rate each action with respect to importance on a 5-point scale ranging from 'least important' to 'most important'. Sirigu *et al*. (1995) reported no difference between frontal lobe patients and controls with regard to the size of the information domain, the speed of access, and the content retrieved. However, frontal lobe patients made more sequencing errors, boundary errors (i.e., failure to reach the end of a script or failure to stop at the end) and priority-setting errors (judgments of the importance of a particular action) than did controls. These results suggest that the content of scripts may remain intact after frontal lobe lesions, but that the organization, particularly temporal organization, is affected by the lesion.

Goel *et al*. (1997) administered a financial planning task to patients with focal lesions to the prefrontal cortex and to normal controls. The task involved a young couple needing help in planning their income, expenditures, investments and lifestyle to stabilize their finances, afford to purchase a home within the next two years, send their children to college in 15 to 20 years' time, and have sufficient funds to retire in 35 years. Subjects were required to help the couple achieve these four specific goals by various manipulations of income and expenditures and/or reallocation of certain assets. The findings indicate that patient performance is impoverished at a global level but not at a local level. That is, at a time-slice of the order of seconds, patient performance was indistinguishable from that of controls. They instantiated the same set of operators, in the same sequence and with the same frequency. However, when their performance was examined over a time scale of minutes to hours, differences began to emerge. Patients had difficulty in organizing and structuring their problem space. Once they began problem solving, they experienced difficulty in allocating adequate effort to each problem-solving phase. They also had difficulty in dealing with the fact that there were no right or wrong answers and there was no official termination point. In addition, they found it problematic to generate their own feedback. They invariably terminated the session before the details were fleshed out and all of the goals had been achieved. Finally, patients did not take full advantage of the fact that constraints on real-world problems are negotiable.

In a follow-up case study, Goel and Grafman (2000) tested an architect (patient PF), with a right prefrontal cortex lesion, in a real-world architectural design/planning task that required the patient to develop a new design for a laboratory space. His performance was compared with that of two age- and education-matched controls (an architect and a lawyer). The patient had superior memory and IQ scores and understood the task, and he even observed that 'this is a very simple problem'. His sophisticated architectural knowledge base was still intact, and he used it quite skillfully during the problem-structuring phase. However, the patient's problem-solving behavior differed from the controls' behavior in the following ways: (1) he had difficulty in making the transition from problem structuring to problem solving; (2) as a result, preliminary planning did not start until he was two-thirds of the way into the session; (3) the preliminary planning phase was minimal and erratic, consisting of three independently generated fragments; (4) there was no progression or lateral development of these fragments; (5) there was no carryover of abstract information into the preliminary planning or later phases; (6) the patient did not reach the detailing phase. This suggests that the key to understanding this patient's deficit is to understand the cognitive processes and mechanisms involved in the preliminary (ill-structured) planning phase.

FRACTIONATING THE NEURAL SUBSTRATES OF PLANNING

While there is mounting evidence for the involvement of the prefrontal cortex in planning, little is known about the differential involvement of the prefrontal cortex in different types of planning or different components of planning. However, it is possible to make a start along these lines. We have already discussed the cognitive and computational differences between ill- and well-structured planning. There is some evidence that the distinction may have a neural basis. Well-structured representations and computations may map on to left-hemisphere processes, while ill-structured representations and computations map on to right-hemisphere processes. That is, the inarticulate, ill-structured representational/computational system may be underwritten by the right prefrontal cortex, while the articulate, well-structured representational/computational system is underwritten by the left prefrontal cortex.

This claim is consistent with the pathology and performance of both Penfield's patient and patient PF reviewed above. Furthermore, at least five recent case studies have described frontal lobe patients with an anatomical and neuropsychological profile similar to that of PF. Like PF, these patients had difficulty in coping with real-world (ill-structured) situations despite having a high IQ. The majority, like PF, had lesions in the right dorsolateral prefrontal cortex (DLPFC). In addition, several imaging studies suggest that the right prefrontal cortex plays a special role in open-ended (i.e., ill-structured) inference tasks with no right or wrong answers (Goel and Dolan, 2000). By contrast, left hemisphere frontal lesions tend to affect more well-structured tasks, including the Tower tasks. For example, left-hemisphere patients are more impaired than right-hemisphere patients on Tower of Hanoi problems, with significant goal–subgoal conflict (Morris *et al*., 1997).

In terms of components of planning, there is an important distinction between *making* and *executing* a plan. The success of the former hinges on cognitive activity, while the success of the latter involves affective traits such as initiative, determination, and 'staying power'. There is growing evidence that the DLPFC is involved in cognitive components of tasks, while the medial ventral prefrontal cortex (MVPFC) plays a more affective role. If this is correct, one would expect patients with DLPFC lesions to have difficulty with plan formulation, and patients with MVPFC lesions to have difficulty with plan execution. Consistent with this, Phineas Gage, who devised 'many plans of future operation but discard[ed] them before they were executed', had a lesion to the MVPFC. Patient PF, who had difficulty in constructing plans, suffered a frontal dorsal lesion (Goel and Grafman, 2000). Neuroimaging studies of the Tower tasks are also attributing plan formulation (as opposed to execution) functions to the dorsolateral and rostral prefrontal cortex.

CONCLUSION

Planning is a human problem-solving activity that involves the need to model and predict the consequences of actions prior to their execution in the world. Planning problems can either be well- or ill-structured. While laboratory problems are typically well-structured, real-world planning has both ill- and well-structured components. There are cognitive and computational grounds for believing that the two forms of planning problems engage quite different mechanisms. Patient and neuroimaging data suggest that the prefrontal cortex plays a critical role in planning, but little is known about the specific contributions that different parts of the prefrontal cortex make to different types and components of planning. However, we can draw some tentative conclusions about the differential roles of the right and left prefrontal cortex in ill- and well-structured planning tasks, respectively, and of the DLPFC and MVPFC in plan formulation and execution, respectively.

References

Baker SC, Rogers RD, Owen AM *et al*. (1996) Neural systems engaged by planning: a PET study of the Tower of London task. *Neuropsychologia* **34**: 515–526.

Fincham JM, Carter CS, van Veen V, Stenger VA and Anderson JR (2002) Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences of the USA* **99**: 3346–3351.

Goel V and Dolan RJ (2000) Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience* **12**: 1–10.

Goel V and Grafman J (1995) Are frontal lobes implicated in 'Planning' functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia* **33**: 623–642.

Goel V and Grafman J (2000) The role of the right prefrontal cortex in ill-structured problem solving. *Cognitive Neuropsychology* **17**: 415–436.

Goel V, Grafman J, Tajik J, Gana S and Danto D (1997) A study of the performance of patients with frontal lobe lesions in a financial planning task. *Brain* **120**: 1805–1822.

Morris RG, Miotto EC, Feigenbaum JD, Bullock P and Polkey CE (1997) The effect of goal–subgoal conflict on planning ability after frontal- and temporal-lobe lesions in humans. *Neuropsychologia* **35**: 1147–1157.

Owen AM, Downes JJ, Sahakian BJ, Polkey CE and Robbins TW (1990) Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* **28**: 1021–1034.

Shallice T and Burgess PW (1991) Deficits in strategy application following frontal lobe damage in man. *Brain* **114**: 727–741.

Sirigu A, Zalla T, Pillon B, Grafman J, Dubois B and Agid Y (1995) Planning and script analysis following prefrontal lobe lesions. *Annals of the New York Academy of Sciences* **769**: 277–288.

Further Reading

Burgess PW (2000) Strategy application disorder: the role of the frontal lobes in human multitasking. *Psychological Research* **63**: 279–288.

Chandrasekaran B (1983) Towards a taxonomy of problem-solving types. *Artificial Intelligence Magazine* Winter/Spring: 9–17.

Damasio AR (1994) *Descartes' Error*. New York, NY: Avon Books.

Goel V (1995) *Sketches of Thought*. Cambridge, MA: MIT Press.

Goel V and Pirolli P (1992) The structure of design problem spaces. *Cognitive Science* **16**: 395–429.

Grafman J (1994) Alternative frameworks for the conceptualization of prefrontal lobe functions. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 9, pp. 187–202. Amsterdam, Netherlands: Elsevier.

Harlow JM (1868) Recovery after severe injury to the head. *Publications of the Massachusetts Medical Society* **2**: 327–346.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Penfield W and Evans J (1935) The frontal lobe in man: a clinical study of maximum removals. *Brain* **58**: 115–133.

Reitman WR (1964) Heuristic decision procedures, open constraints, and the structure of ill-defined problems. In: Shelly MW and Bryan GL (eds) *Human Judgments and Optimality*, pp. 282–315. New York, NY: John Wiley.

Shallice T (1988) *From Neuropsychology to Mental Structure*. Cambridge, UK: Cambridge University Press.

Prosopagnosia

Introductory article

Andrew W Young, University of York, York, UK

CONTENTS

Prosopagnosia is a neurologically based impairment of the ability to recognize faces.

THE DEFICIT AND BRAIN LESIONS THAT CAN CAUSE PROSOPAGNOSIA

Most of us recognize the faces of individuals whom we know with such ease that it comes as a surprise to discover that some people are unable to do this. Prosopagnosic individuals experience great difficulty with face recognition, and must rely on other cues (such as the person's voice) to identify those whom they know.

Prosopagnosia is by definition regarded as a form of visual agnosia. For a person to be considered prosopagnosic, it is therefore necessary to establish that the problem with regard to face recognition is not due to blindness (a prosopagnosic person can still see) or general intellectual impairment (people are still recognized from non-facial cues).

Prosopagnosia should not be confused with other types of impairment that can compromise face recognition. In prosopagnosia, recognition from cues other than the face is usually possible. In contrast, cases of inability to recognize people from face, voice and name have been described, and these seem to reflect loss of semantic information about the identities of individuals. There are also reports of problems with regard to name retrieval (anomia), in which familiar people are successfully recognized and appropriate semantic information is accessed, but their names cannot be recalled. Impairment of the ability to recognize faces therefore reflects damage to an underlying system which can break down in different ways.

Although prosopagnosia is generally considered to be a rare deficit, there are now several hundred case descriptions in the literature. These include cases where brain injury early or late in life has led to the inability to recognize previously familiar faces, and developmental cases where the disorder has been present from birth.

In cases that are due to brain injury, the underlying pathology involves lesions that affect ventro-medial regions of the occipito-temporal cortex. These include the lingual, fusiform, and parahippocampal gyri, and more anterior parts of the temporal lobes. Functional imaging studies of normal observers have confirmed the importance of these regions with regard to face perception, and especially part of the fusiform gyrus now known as the fusiform face area. Bilateral lesions are usually present in the relatively small number of cases of prosopagnosia that have come to autopsy, but cases involving unilateral lesions of the right cerebral hemisphere have been reported in several computed tomography (CT) and magnetic resonance imaging (MRI) studies, and the functional imaging findings for normal individuals also highlight the importance of the right hemisphere.

Deficits that are commonly associated with prosopagnosia include a visual field defect in the left upper quadrant, achromatopsia (loss of color vision), and topographical disorders (inability to find one's way about). These are useful clinical pointers, but they are thought to be due to the anatomical proximity of otherwise unrelated processes, rather than having any direct functional significance. Cases of prosopagnosia without each of these associated deficits have been described, and they have also each been reported in the absence of prosopagnosia.

People with prosopagnosia know when they are looking at a face, and can describe its features, but the loss of any sense of overt recognition is often complete, with no feeling of familiarity. In contrast, other aspects of face processing, such as the ability to interpret facial expressions or to match views of unfamiliar faces, can remain surprisingly well preserved in some (although not all) cases.

There is evidence that although these aspects of face perception may be well preserved in some cases of prosopagnosia, they may be supported by unusual mechanisms and strategies. For example, the matching of pictures of unfamiliar faces by prosopagnosic individuals is often achieved by a painstaking feature-by-feature inspection, as if the face's general appearance is somehow no longer perceived 'as a whole'. Many investigators have emphasized that recognizing a face involves sensitivity to the spatial positioning of the facial features (the face's configurational properties) as well as to the features themselves. For a normal perceiver, accurate perception of the facial configuration is strongly dependent on the face being seen in its usual upright orientation – we are not good at detecting configurational changes in inverted (upside-down) faces. This sensitivity to the face's orientation may be absent (or even paradoxically reversed) in prosopagnosia, which suggests that there is a disruption of configurational encoding.

PROSOPAGNOSIA AND OTHER DEFICITS OF RECOGNITION OF VISUAL OBJECTS

In prosopagnosia, even the most familiar faces may go unrecognized, such as those of famous people, friends and family, and the affected individual's own face when seen in a mirror. The deficit encompasses both premorbidly familiar faces and the faces of people who have been encountered only since the illness onset.

The ability to recognize everyday objects is usually not as severely affected as the ability to recognize faces, and many prosopagnosic patients are able to read without much difficulty. In a case of early acquired deficit, a child who had remained unable to recognize any faces since suffering meningitis and subsequent complications in infancy was nonetheless able to learn to read. This supports the view that reading and face recognition depend on different types of visual analysis.

A hypothesis that fits such facts neatly has been proposed by Farah, who suggested that the brain uses separate mechanisms to recognize visual stimuli consisting of decomposable parts (e.g., words, whose constituent letters are all easily recognized subunits) and visual stimuli that we recognize as wholes (e.g., faces, whose individual features are relatively difficult to recognize in isolation). In prosopagnosia, the whole-based mechanism is impaired but the part-based mechanism may be much less affected. This theory is consistent with the findings of altered inversion effects in cases of prosopagnosia.

IS FACE RECOGNITION SPECIAL?

The assumption underlying much of the interest in prosopagnosia has been that it might indicate whether the brain has an evolved neural substrate for the important social task of face recognition. Fodor has been one of the most influential of many authorities who have seriously entertained the idea of a 'module' for face recognition.

A strongly debated issue has therefore been whether prosopagnosia is specific to face recognition. There is now no doubt that most people with prosopagnosia have problems with certain other visual stimuli (as would be expected from Farah's hypothesis about whole and part-based recognition), but it is uncertain whether these problems with regard to recognition of objects other than faces are an essential feature of the condition (as Farah's hypothesis implies), or whether they reflect separate problems that are likely to coexist with impairment of face recognition. The latter view is supported by occasional reports of highly circumscribed impairments. One of De Renzi's cases could find his own belongings when they were mixed in with similar objects, identify his own handwriting, pick out a Siamese cat among photographs of other cats, recognize his own car without reading the number plate, and sort domestic coins from foreign coins. In all of these tasks he was therefore able to identify individual members of visual categories with high inter-item similarity, yet he could not recognize the faces of relatives and close friends. Just as strikingly, the deficit can affect mainly *human* faces. When McNeil and Warrington's prosopagnosic patient took up farming, he was able to learn to recognize his sheep, and he correctly identified several of them from photographs of their faces!

Although they are exceptionally rare, such cases imply that the possibility of face-specific deficits must be taken seriously. This has been thought to be consistent with evidence of an abundance of face-responsive cells in the brains of primate species. However, many of these cells are in regions of the brain that do not seem to be related to face recognition *per se*, but which are likely to be involved in other important social abilities.

The development of functional imaging methods has opened up the possibility of resolving the issue by studying the neurologically normal human brain. At present, however, functional imaging studies have simply intensified the debate.

Although the fusiform face area is strongly implicated in face recognition, it has been shown that parts of the fusiform gyrus can also acquire responsiveness to novel visual stimuli that share some face-like properties.

As with previous controversies in this field, some of the debate centers around differences in what is meant by the claim of a 'special' face recognition system. It is evident that fusiform cells show a degree of plasticity, but this does not in itself rule out the possibility that some of them form an evolved substrate for face recognition.

CONSCIOUS AND NON-CONSCIOUS FACE RECOGNITION

Prosopagnosic patients usually fail all tests of overt recognition of familiar faces. They cannot name the face, give the person's occupation or other biographical details, or even state whether the face is that of a familiar person (all faces seem unfamiliar to them). Surprisingly, however, there is substantial evidence of covert recognition from physiological and behavioral measures.

Bauer initiated this line of research by measuring skin conductance while prosopagnosic patient L.F. viewed a familiar face and listened to a list of five names. When the name belonged to the face at which L.F. was looking, there was a greater skin conductance change than when someone else's name was read out. Yet if L.F. was asked to choose which name in the list was correct for the face, his performance was found to be at chance level. Thus there was a marked discrepancy between L.F.'s inability to identify the face overtly and the relatively good recognition that was demonstrated using the indirect skin conductance measure.

A number of other indices of non-conscious, covert recognition in prosopagnosia have been developed. Matching of familiar faces is better than for unfamiliar faces, priming has been found from familiar faces on to the recognition of name targets, and learning of correct face–name pairings is better than learning of incorrect pairings. All of these effects can be demonstrated when overt recognition of the faces is at the level that would be expected by chance.

However, not all prosopagnosic patients show covert recognition. As might be expected, when there is evidence of substantial impairment of face perception, covert as well as overt performance can decline to chance level.

There is a parallel between preserved covert recognition abilities and those aspects of recognition that operate automatically for normal people. We cannot look at a familiar face and choose not to recognize it – the mechanisms responsible for visual recognition are not open to conscious control in this way. It seems that some of these automatic aspects of recognition continue to function in some cases of prosopagnosia. Young and Burton simulated this pattern by halving some of the connection strengths in a computer model of the architecture of the recognition system. The network was then no longer able to classify face inputs as familiar, yet it continued to display priming effects. This helps us to understand how covert responses can be preserved when there is no overt discrimination, but it does not solve the more difficult issue of what is involved in awareness of recognizing a face.

Despite the extensive range of covert effects demonstrated in the laboratory, prosopagnosic patients do not act as if they recognize faces in everyday life. Instead, most of them are acutely conscious of their problems in face recognition. However, overt recognition can sometimes be provoked in formal experimental settings. Sergent and Poncet found that their patient P.V. could achieve overt recognition of some faces if several members of the same occupational category were presented together. This only happened when P.V. could determine the category herself. Otherwise she continued to fail to recognize the faces overtly even when the occupational category was pointed out to her. The findings have been replicated in other cases, and it seems that the simultaneous presentation of several faces from the same category may temporarily raise their activation above the threshold needed for a sense of overt recognition. Computer simulation work by Morrison and his colleagues has demonstrated one possible mechanism whereby this might occur.

Findings of provoked overt recognition in prosopagnosia show that the boundary between awareness and lack of awareness is not as completely impassable as it seems to the patient's everyday experience. However, the circumstances under which this has been found to occur are currently very limited, and it has not yet been possible to find a way to provide significant remedial assistance in real life.

Further Reading

Bauer RM (1984) Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the guilty knowledge test. *Neuropsychologia* **22**: 457–469.

Bruyer R (1991) Covert face recognition in prosopagnosia: a review. *Brain and Cognition* **15**: 223–235.

De Renzi E (1986) Current issues in prosopagnosia. In: Ellis HD, Jeeves MA, Newcombe F and Young A (eds) *Aspects of Face Processing*, pp. 243–252. Dordrecht, Netherlands: Martinus Nijhoff.

Farah MJ (1991) Patterns of co-occurrence among the associative agnosias: implications for visual object representation. *Cognitive Neuropsychology* **8**: 1–19.

Farah MJ, Wilson KD, Drain HM and Tanaka JR (1995) The inverted face inversion effect in prosopagnosia: evidence for face-specific, mandatory perceptual mechanisms. *Vision Research* **35**: 2089–2093.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Haxby JV, Hoffman EA and Gobbini MI (2002) Human neural systems for face recognition and social communication. *Biological Psychiatry* **51**: 59–67.

Kanwisher N (2000) Domain specificity in face perception. *Nature Neuroscience* **3**: 759–763.

McNeil JE and Warrington EK (1993) Prosopagnosia: a face-specific disorder. *Quarterly Journal of Experimental Psychology* **46A**: 1–10.

Morrison DJ, Bruce V and Burton AM (2001) Understanding provoked overt recognition in prosopagnosia. *Visual Cognition* **8**: 47–65.

Sergent J and Poncet M (1990) From covert to overt recognition of faces in a prosopagnosic patient. *Brain* **113**: 989–1004.

Tarr MJ and Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* **3**: 764–769.

Young AW (1992) Face recognition impairments. *Philosophical Transactions of the Royal Society of London* **B335**: 47–54.

Young AW and Ellis HD (1989) Childhood prosopagnosia. *Brain and Cognition* **9**: 16–47.

Young AW and Burton AM (1999) Simulating face recognition: implications for modelling cognition. *Cognitive Neuropsychology* **16**: 1–48.

Reasoning and Thinking, Neural Basis of

Intermediate article

Jordan Grafman, National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA
Vinod Goel, York University, Toronto, Ontario, Canada

CONTENTS

Reasoning is the cognitive activity of drawing inferences from given information. All arguments in reasoning involve the claim that one or more propositions (the premises) provide some grounds for accepting another proposition (the conclusion).

INTRODUCTION

Reasoning is the cognitive activity of drawing inferences from given information. All arguments in reasoning involve the claim that one or more propositions (the premises) provide some grounds for accepting another proposition (the conclusion). Consider the following example. George returns from the grocery store accompanied by his young daughter and tells her that kiwi are perishable fruit, and that perishable fruit are placed in the refrigerator. Even though she has never seen kiwi before, she promptly places the kiwi in the refrigerator. How did she know that she should put them in the refrigerator? Furthermore, after seeing the purchased kiwi, she formed the belief that all kiwi have a brown, furry skin. Notice that she was not explicitly told that kiwi should be placed in the refrigerator, and she saw only the three purchased kiwi, yet we are not surprised by her actions. Her behavior is not a mystery. It is just an example of the reasoning brain at work. The former inference is an example of deductive reasoning, while the latter is an example of inductive reasoning.

A key feature of deductive arguments is that conclusions are contained within the premises and are independent of the content of the sentences. They can be evaluated for *validity*, a relationship between premises and conclusion involving the claim that the premises provide absolute grounds for accepting the conclusion. The reader is referred to Johnson-Laird's article on deduction for background information on the cognitive theories of deductive reasoning.

In inductive reasoning the conclusion reaches beyond the original set of premises, allowing for the possibility of creating new knowledge. Induction has long been a concern of philosophers, and it has been empirically studied by cognitive scientists for the past 25 years. Cognitive/computational models of induction typically view it as a form of hypothesis generation and selection, where one must search a large database and determine which items of information are relevant and how they are to be mapped on to the present situation. The determination of hypothesis relevance is the crucial component of induction. In the above example, the inference that 'all kiwi have a brown, furry skin', which was drawn after seeing three kiwi, seems to be plausible. However, each of the three kiwi also had a small sticker attached stating the name of the store at which they were purchased ('IGA'). Yet from seeing this sticker on three kiwi fruits we do not conclude that all kiwi have an IGA sticker on them. The former inference is plausible, but the latter is not. The logic of the two inferences is identical. However, we recognize that the property of having 'brown furry skin' is a relevant property for generalization across members of a species, but that the property of having an 'IGA' sticker is a matter of individual accident. The puzzle of induction is, to a large extent, the question of how we make these judgments of relevance. Given that almost nothing is known about the cognitive, computational, and neural basis of inductive inference, we shall restrict ourselves to deductive reasoning in this article.

REASONING AND THE BRAIN

There are two main ways in which cognitive neuroscientists investigate the neural mechanisms that

underlie reasoning processes. One approach is to study patients who have damage to specific brain regions and to observe whether they have difficulty in solving certain types of reasoning problems and, if so, which specific cognitive impairments can explain their reasoning deficits. The second approach is to study normal subjects and patients using functional neuroimaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), which allow the investigator to view which brain areas become selectively activated by specific reasoning processes and tasks.

Patient studies

Patient studies of logical reasoning have been few and infrequent. Gazzaniga and colleagues have administered simple reasoning tasks to split-brain patients and concluded that reasoning is a left-hemisphere phenomenon (Gazzaniga and Smylie, 1984). For example, they report that the left hemisphere will readily infer 'boiling water' when presented with 'water' and 'pan', while the right hemisphere seems to be incapable of such inferences. Gazzaniga goes on to postulate a 'left-brain interpreter' – that is, a mechanism which continuously elaborates and interprets information presented to it, and which readily draws inferences.

Caramazza *et al*. (1976) administered to brain-damaged patients two-term problems such as the following: 'Mike is taller than George. Who is taller?' They reported that left-hemisphere patients showed impairment for all forms of the problem, but right-hemisphere patients were impaired only when the form of the question was incongruent with the premise (e.g. 'Who is shorter?').

Read (1981) tested temporal-lobectomy patients by presenting them with three-term relational problems with semantic content (e.g. 'George is taller than Mary. Mary is taller than Carol. Who is tallest?'). The subjects were told that using a mental imagery strategy would help them to solve these problems. Read reported that the performance of left-temporal-lobectomy patients was more impaired than that of right-temporal-lobectomy patients. Left-temporal-lobectomy patients reported less use of mental imagery to solve the problems, and they also did not show deficits on standard neurological tests of verbal comprehension. Read concluded that 'The form of imagery utilized in solving deductive–recent problems is based upon verbal symbolic affirmation, and as such is mediated by the left hemisphere' (Read, 1981).

Golding (1981) evaluated responses to the Wason Four-Card Selection Task (Wason, 1966), which is perhaps the most widely used task for exploring the role of content in reasoning. In this task subjects are shown four cards (Figure 1). They can see what is on one side of each card, but not what is on the other side. They are given a rule of the form if *p* then *q* ('If a card has a vowel on one side, it has an even number on the other side') and asked which cards they would turn over in order to verify the rule. The visible values on the cards correspond to the *p*, *not-p*, *q*, and *not-q* cases of the rule. According to standard propositional logic, the correct choices are *p* (to verify that *q* is on the other side) and *not-q* (to verify that *p* is not on the other side). Although no control subjects and only one left-hemisphere-damaged subject selected the *p* card and *not-q* card (the logically correct answer), 50% of the right-hemisphere-damaged subjects chose both cards. Because control subjects tended to select those cards that matched the items described in specific test sentences, Golding proposed that the perceptual aspects of the task interfered with control subjects' verbal reasoning (e.g. when given the sentence 'Whenever there is a circle on one half of the card, there is yellow on the other half of the card,' control subjects selected the circle card or the circle and yellow cards). According to Golding, right-hemisphere-damaged patients with impaired visual processing showed superior verbal reasoning skills owing to a lack of visual perceptual interference.

Adolphs *et al*. (1996) administered a Wason selection task, using both familiar and unfamiliar stories, to patients with dorsolateral frontal lesions or ventromedial frontal lesions, and to normal controls. When given an arbitrary rule (as in Figure 1), typically fewer than 15% of normal subjects will turn over both the *p* card (A) and the *not-q* card (3). The introduction of meaningful content in a rule (e.g., 'If anyone is drinking beer, then that

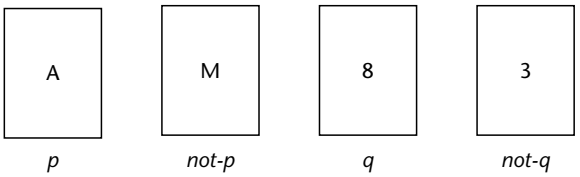


Figure 1. The Wason Four-Card Selection Task. Each card has a letter on one side and a number on the other side. The task is to determine which cards need to be turned over in order to verify the rule 'If a card has a vowel on one side, it has an even number on the other side.'

person must be over 19 years old') greatly facilitates performance. They reported that subjects in all three groups performed equally poorly when given unfamiliar stories (arbitrary rules). However, both dorsolateral patients and normal subjects chose the *p* and *not-q* cards (the logically correct selections) when the story involved familiar material. In contrast, the ventromedial patients, who had damage to the medial orbitofrontal cortex (five out of six cases had bilateral damage), did not show this facilitatory effect with familiar material. The authors concluded that in general people may reason by analogy, retrieving past experiences (including the emotion experienced) when confronted with a familiar situation. Ventromedial prefrontal patients may fail to retrieve or appropriately use past experiences when reasoning.

Waltz *et al*. (1999) found that compared with patients with damage to the temporal cortex, patients with frontal-lobe dementia were dramatically impaired in their ability to make inferences with regard to the integration of multiple relational representations. If problems varied on only one dimension, there were no between-group differences (including controls). If a problem only required judgments about canonical ordering, then patients with frontal-lobe dementia could solve the problem, but if they had to reorder the relationships to make the judgment, they failed to solve it.

In summary, studies of patients with focal lesions indicate that the left hemisphere helps to interpret stimuli as linguistic elements for further processing and inferential processes. Damage to the left hemisphere interferes with deductive reasoning. The right hemisphere may become more involved when distant relationships are being processed, and thus right-hemisphere lesions affect the processing of apparently incongruent stimuli. It is likely that left-hemisphere lesions lead to an inability to develop certain forms of mental imagery that are used when developing mental models of relationships. If a subject needs to use past experience to solve a problem, then the ventromedial prefrontal cortex may be important.

Neuroimaging studies

Goel *et al*. (1997) conducted an imaging study of deductive reasoning. Ten normal volunteers performed reasoning tasks while their regional cerebral blood flow pattern was recorded using the [^{15}O] H_2O positron emission tomography (PET) blood flow technique. Subjects were presented with arguments such as 'All apples are red; all red fruit are sweet; therefore all apples are sweet', and were then asked to make judgments about the validity of the conclusion. The deduction condition (versus a semantic baseline) resulted in activation of the left inferior frontal gyrus (Brodmann areas 45 and 47) and a region of the left superior occipital gyrus (Brodmann area 19). These brain areas are known to mediate aspects of language processing, and as such the results provide support for sentential theories of reasoning, at the expense of spatial models.

However, given that many people have the phenomenological experience of mapping syllogisms on to Venn-like diagrams, it is difficult to accept that the visuospatial system plays no role in deductive reasoning. Perhaps the absence of performance-associated brain activation in regions that are known to be involved in spatial processing (i.e. right hemisphere and parietal regions) was just a function of the fact that the deductive reasoning items which were used (a combination of categorical syllogisms, implications, disjunctions, and conjunctions) did not involve overt spatial relationships (e.g. John is standing behind Mary). Perhaps if the arguments explicitly required spatial encoding, then we might find right-hemisphere and parietal activation.

Goel *et al*. (1998) conducted another imaging experiment, this time using explicitly spatial arguments. Again they found unilateral (left-hemisphere) activation confined to the dorsolateral prefrontal cortex, the anterior cingulate, and the middle and superior temporal lobes. They did not find any significant activation in cortical regions known to be associated with spatial encoding and spatial working memory. These results were consistent with the first study (and a number of other patient studies), and demonstrated a lack of involvement of classic spatial encoding regions in reasoning.

However, the results of another study by Goel *et al*. (2000) may go some way towards resolving these counterintuitive results. In an fMRI study of deductive reasoning, using sentences with semantic content (e.g. 'All apples are red; all red fruit are sweet; therefore all apples are sweet') and without semantic content (e.g. 'All A are B; all B are C; therefore all A are C), they found evidence for the engagement of both linguistic and spatial systems, but under circumstances not predicted by cognitive theories. During content-based reasoning, a left-hemisphere temporal system was recruited. By contrast, a formally identical reasoning task that lacked semantic content activated a bilateral parietal system. This dissociation is remarkable because the logically relevant information is identical in

both conditions. Based on brain activation profiles, the two systems shared common neural components in bilateral basal ganglion nuclei, right cerebellum, bilateral fusiform gyri, and left prefrontal cortex. Goel *et al*. concluded that syllogistic reasoning is implemented in two distinct systems whose engagement is primarily a function of the presence or absence of semantic content. Furthermore, when a logical argument results in a belief–logic conflict, the nature of the reasoning process is changed by recruitment of the right prefrontal cortex.

The involvement of a parietal visual–spatial system in the abstract syllogism condition raises the question of whether argument forms involving three-term relational items (e.g. 'The apples are in the barrel; the barrel is in the barn; therefore the apples are in the barn' and 'Apples are more expensive than pears; pears are more expensive than oranges; therefore apples are more expensive than oranges') are sufficient to engage the parietal system. Goel and Dolan (2001) addressed this question in an event-related fMRI study of three-term relational reasoning, using sentences with concrete content (e.g. 'The apples are in the barrel; the barrel is in the barn; therefore the apples are in the barn') and abstract content (e.g. 'A are in B; B is in C; therefore A is in C'). They reported that both concrete and abstract three-term relational arguments activate a similar bilateral occipital–parietal–frontal network. However, the abstract reasoning condition engendered greater parietal activation than the concrete reasoning condition. It appears that arguments involving relationships that can be easily mapped on to explicit spatial relationships engage a visuospatial system, irrespective of concrete or abstract content.

The only other imaging study of deductive reasoning to date is that by Osherson *et al*. (1998), who performed a [^{15}O] H_2O PET study to compare deductive reasoning with probabilistic reasoning and language-comprehension tasks involving the same stimuli. They reported that deductive reasoning compared with probabilistic reasoning activated occipital and parietal regions, with a right-hemispheric prevalence. The comparison most relevant to us, namely deductive reasoning (using arguments such as 'None of the bakers play chess; some of the chess players listen to opera; some of the opera listeners are not bakers') versus language comprehension, resulted in activation in the left dorsal frontal gyrus (BA6), left cuneus (BA18), thalamus, caudate, and cerebellum.

In summary, the functional neuroimaging results demonstrate that the left hemisphere, and in particular the left prefrontal cortex, are very important for deductive reasoning. The more anterior the activation, the more likely it is that there is a semantic content requirement for reasoning to solve the problem.

ASSESSMENT OF THE STUDIES

The neuroimaging and patient studies to date support some form of dual-mechanism theory of deductive reasoning. The presence of semantic content engages the language system in the reasoning process. The absence of semantic content engages the visuospatial system in the identical reasoning task. At an anatomical level, these data predict that the left hemisphere is necessary and often sufficient for logical reasoning, whereas the right hemisphere is sometimes necessary but never sufficient. There is considerable consistency between the neuroimaging findings and the lesion data reported above.

As described above, both brain-lesion patients and functional neuroimaging studies suggest that each hemisphere and specific brain regions may contribute different cognitive processes to human reasoning. One interpretation of the results of these studies is that the right hemisphere may be involved in formal reasoning independent of the content of the task. Therefore the right hemisphere would be more adept at abstract reasoning. In contrast, the left hemisphere may use past experiences (factual, emotional), and would thus be more adept when reasoning involves familiar scenarios. The ventromedial prefrontal neural network would play a role when the behavioral relevance of possible responses could aid selection of the appropriate action.

The left hemisphere may also store in memory emotional states associated with personally experienced events. When an individual encounters a familiar scenario, representations of related past emotional experiences are retrieved by the left hemisphere and incorporated into the reasoning process. In the absence of or failure to activate such representations, the right hemisphere is engaged in the reasoning process.

Asymmetrical advantages of processing based on receptive-field size offer another explanation for hemispheric differences in reasoning skills. Beeman (1993) provided evidence that large semantic receptive fields account for the right hemisphere's role in understanding discourse and metaphor. A similar explanation may underlie hemispheric differences in reasoning. Large receptive fields in the right hemisphere would permit individuals to

activate all possible relationships, local and distant, between the items in the problem to be solved. Overlap between relational features from the activation of multiple relationships would give the right hemisphere an adaptive advantage over the left hemisphere for reasoning involving unfamiliar situations. By comparison, the left hemisphere's fine coding would allow individuals to focus on the main feature or event and the local relationships between the items in the problem. As a result, the left hemisphere would have an adaptive advantage over the right hemisphere in logical reasoning involving familiar content.

One clear conclusion to be drawn from the cognitive neuroscience investigation of reasoning is that there is no reasoning module, and that several different cognitive-processing components are required to reason, depending on the type of task with which the subject is faced and the particular strategies with which the subject prefers to reason.

FUTURE DIRECTIONS

The cognitive neuroscience investigation of reasoning is still in its infancy. It is clear from this article that we are only crudely able to map reasoning processes primarily to the left hemisphere and to the prefrontal cortex. More posterior cortices may contribute to reasoning when visuospatial representations are used by subjects to think about the elements of the reasoning task they are performing. The semantic content of the elements of the problem will require temporal-lobe processing, and more personal or familiar content will utilize the functions of the prefrontal cortex. The latter does not fully mature until the end of the teenage years in humans, and it is of interest to note that reasoning ability matures over the same time frame.

Reasoning does not stand alone as a cognitive process. Humans frequently reason during problem-solving tasks, when forming metaphors, and when developing a plan to achieve a goal. People also reason in social and non-social situations, and it may be that overlapping but different cognitive systems and neural regions are used for each type of reasoning. Future cognitive neuroscience studies that incorporate these factors into the design of reasoning tasks will enrich our understanding of the neural basis of reasoning processes.

References

Adolphs R, Tranel D, Bechara A, Damasio H and Damasio AR (1996) Neuropsychological approaches to reasoning and decision-making. In: Damasio AR (ed.) *Neurobiology of Decision-Making*, pp. 157–180. Berlin, Germany: Springer-Verlag.

Beeman M (1993) Semantic processing in the right hemisphere may contribute to drawing inferences from discourse. *Brain and Language* **44**: 80–120.

Caramazza A, Gordon J, Zurif EB and DeLuca D (1976) Right-hemispheric damage and verbal problem-solving behavior. *Brain and Language* **3**: 41–46.

Gazzaniga MS and Smylie CS (1984) Dissociation of language and cognition. *Brain* **107**: 145–153.

Goel V and Dolan RJ (2001) Functional neuroanatomy of three-term relational reasoning. *Neuropsychologia* **39**: 901–909.

Goel V, Gold B, Kapur S and Houle S (1997) The seats of reason: a localization study of deductive and inductive reasoning using PET (O^{15}) blood flow technique. *Neuroreport* **8**: 1305–1310.

Goel V, Gold B, Kapur S and Houle S (1998) Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience* **10**: 293–302.

Goel V, Buchel C, Frith C and Dolan RJ (2000) Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage* **12**: 504–514.

Golding E (1981) The effect of unilateral brain lesion on reasoning. *Cortex* **17**: 31–40.

Osherson D, Perani D, Cappa S *et al*. (1998) Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia* **36**: 369–376.

Read DE (1981) Solving deductive-reasoning problems after unilateral temporal lobectomy. *Brain and Language* **12**: 116–127.

Waltz JA, Knowlton BJ, Holyoak KJ *et al.* (1999) A system for relational reasoning in human prefrontal cortex. *Psychological Science* **10**: 199–125.

Wason PC (1966) Reasoning. In: Foss B (ed.) *New Horizons in Psychology*, pp. 135–151. Harmondsworth, UK: Penguin.

Further Reading

Evans JSBT, Newstead SE and Byrne RMJ (1993) *Human Reasoning: the Psychology of Deduction*. Hillsdale, NJ: Lawrence Erlbaum.

Garnham A and Oakhill J (1994) *Thinking and Reasoning*. Oxford, UK: Blackwell Science.

Goel V and Dolan RJ (2000) Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience* **12**: 1–10.

Goodman N (1955) *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press.

Henle M (1962) On the relation between logic and thinking. *Psychological Review* **69**: 366–378.

Johnson-Laird PN (1994) Mental models, deductive reasoning, and the brain. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 999–1008. Cambridge, MA: MIT Press.

McCarthy RA and Warrington EK (1990) *Cognitive Neuropsychology: a Clinical Introduction*. New York, NY: Academic Press.

Newell A (1980) Reasoning, problem solving and decision processes: the problem space as a fundamental category. In: Nickerson RS (ed.) *Attention and Performance VIII*, pp. 693–718. Hillsdale, NJ: Lawrence Erlbaum.

Rescher N (1980) *Induction: an Essay on the Justification of Inductive Reasoning*. Pittsburgh, CA: University of Pittsburgh Press.

Rips LJ (1994) *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.

Sloman SA (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* **119**: 3–22.

Receptive Fields

Introductory article

Aniruddha Das, Columbia University, New York, USA

CONTENTS

The receptive field of a neuron is the extended region of sensory space to which the neuron responds, including those stimuli that directly activate or suppress the neuron's activity, as well as those stimuli that can modulate the neuron's response to other stimuli, but have no direct effect by themselves.

INTRODUCTION

Neurons in the sensory systems of touch (somatosensory), hearing (auditory) and vision can be defined in terms of their receptive fields (RFs), namely the regions of sensory space that drive them most vigorously.

While neurons early in a sensory pathway have compact RFs that respond to simple stimuli, neurons at higher stages in the same pathway could have large RFs with complex properties, such as the 'face-selective' cells in the higher visual inferior temporal (IT) region. Determining the transformations in RF properties on passing from stage to stage in a sensory pathway and deducing the cortical circuitry that underlies such a transformation lies at the heart of understanding sensory processing.

The RFs of neurons, particularly at higher centers, can be highly plastic and modifiable. Their response properties may vary depending on the task in which the individual is engaged, may be modulated by attention, and can change over time with sensory training or changes in sensory stimulation. In particular, RFs can 'learn' to select features that are most important in the individual's environment and thus adapt constantly, although within limits, to changes in that environment.

This article will draw heavily on the visual system to illustrate the principles of RF structure, layout, response properties and plasticity, since vision is the best-studied sensory system. However, the principles described here apply to all sensory systems, and these parallels will be mentioned explicitly wherever possible.

HISTORICAL PERSPECTIVE

The concept of the receptive field (RF) was first suggested by the British physiologist Sir Charles Sherrington in the 1890s when he was studying the scratch reflex in dogs. He used this term in the book *The Integrative Action of the Nervous System* (published in 1906) to capture his observation that the reflex was spatially localized on the animal's body. Touching any spot on the dog's back elicited a scratch response directed to the same spot, which Sherrington named the receptive field for that reflex response.

In 1938, H. Keffer Hartline, then working at Johns Hopkins University, introduced the use of the term 'receptive field' to describe the response properties of single nerve fibers. Using a painstaking procedure he dissected out single fibers from the optic nerve of the horseshoe crab (*Limulus polyphemus*), laid them out on a cotton wick soaked in saline to make electrical contact, and then recorded the responses of these fibers while the *Limulus* eye was being stimulated with small spots of light. Each fiber (the axon of a retinal ganglion cell) only responded to light falling on a well-localized spot on the animal's retina, which Hartline defined as the RF for that particular nerve fiber. With similar recordings from frog and later mammalian (cat) optic nerve fibers, Hartline showed that some neurons responded only when a spot of light was turned on in their RF, some responded only when the spot was turned off, while others responded at both light on and light off. This was clear evidence that fairly elaborate visual processing was occurring even in the retina, with RFs reflecting the interactions between excitatory and inhibitory neurons.

In the 1940s, Stephen Kuffler, then also at Johns Hopkins University, exploited the invention of

metal microelectrodes (by Ragnar Granit) to refine Hartline's discoveries by showing that the RFs of retinal ganglion cells had a complex spatial structure. By stimulating different parts of a (cat) retinal ganglion cell RF using small spots of light, Kuffler showed that these RFs had two distinct regions – one excitatory and the other inhibitory – arranged in a concentric manner. He also introduced a more nuanced view of the RF by suggesting that the latter should encompass:

> not only the areas from which responses can actually be set up by the retinal illumination…but also areas which show a functional connection, by an inhibitory or excitatory effect on a ganglion cell. This may well involve areas which are somewhat remote from a ganglion cell and by themselves do not set up discharges.

This idea anticipates the current notion of RFs that includes regions exerting purely modulatory influences on a neuron (see later section on feedback and non-classical effects). Kuffler's colleagues David Hubel and Torsten Wiesel were the first to characterize the RFs of cortical neurons. They showed that neurons in primary visual cortex (V1) had RF properties (orientation preference, direction preference and binocularity) which were more complex than the properties of lateral geniculate nucleus (LGN) neurons providing input to the cortex. However, these RF properties could be explained by the input circuitry to V1. These findings led to the concept of each sensory system as a hierarchical sequence of processing stages, with the neural circuits at each stage generating progressively more elaborate RFs, leading up to the final sensory percept.

Early in the development of the concept of RFs, Humberto Maturana, Jerome Lettvin and their colleagues proposed the attractive hypothesis that RFs were specifically tuned to extract features of importance to the particular species. Working at Massachusetts Institute of Technology (MIT) in the late 1950s, they showed that the different types of RF in frog retina in effect acted as a 'bug detector' that was most sensitive to moving objects of the same size, visual contrast and speed as the worms that formed the animal's diet. Research conducted since then suggests that the visual systems in all species are tuned along broad 'ecological' principles. In species with more complex visual systems and ecological needs than those of frogs, such tuning does not emerge as a hard-wired response in the retina. Rather, at every level of visual processing, from the retina to the cortex, RFs appear to be attuned to those features of the visual surround that are important to the species. This tuning to ecologically relevant features becomes progressively more specialized in the higher levels of each visual system (see later sections on receptive-field organization and feedback and non-classical effects). Such principles of ecological tuning also operate in other sensory modalities. For example, songbirds have neurons that are precisely tuned for the particular song, of the species, with the ability to make fine discriminations between various 'dialects' in the song. Similarly, the auditory circuits that bats use for sonar ranging are tuned to pick up reflections from objects of the same size and speed as their prey, namely insects.

DIFFERENT SENSORY SYSTEMS

General Principles of Organization and Sensory Discrimination

The RF is a useful concept in the sensory systems of vision, touch and hearing because stimuli in all three systems form sensory spaces. Individual stimuli can be positioned in terms of a metric or distance relative to other stimuli, such that the more similar a pair of stimuli, the shorter the distance. This idea of a smooth and well-defined metric holds not only for simple stimulus attributes such as the position of a light touch to the body or the pitch of a pure tone, but also for complex stimuli such as faces that can be arranged along some continuum from 'more' to 'less' similar.

Individual neurons in all three systems respond over compact regions of sensory space. In other words, neurons are activated optimally by one particular stimulus, but stimuli with 'nearby' properties also activate the neuron in proportion to their similarity to the 'best' stimulus. As a corollary, neurons form smooth maps, with neurons that are closer to each other anatomically (e.g. on cortex) having RFs that are closer to each other in the stimulus space. Each stimulus activates a compact region of the neuronal map, its point image – that is, the patch of neurons whose RFs overlap with the given stimulus. The size of the minimal point image is uniform within a cortical region, independent of the stimulus. This means that the smaller the RF, the higher the cortical magnification (the cortical distance between two neurons whose RFs are separated by some unit distance in sensory space), and the better the level of discrimination between stimuli.

The senses of smell and taste are more difficult to describe in terms of uniform sensory spaces with

well-defined distances between distinct sensations, and the concept of the RF does not translate well to these two sensory systems.

Particular Sensory Systems

Here we shall consider the relationships between RFs, sensory spaces, cortical maps and stimulus discriminability.

Somatosensory system (touch)

Each neuron in the somatosensory system responds to stimuli in a circumscribed region of the body, which defines the RF (Figure 1). Some RFs do not lie on the body surface, but rather they carry proprioceptive information from muscle stretch receptors and limb joints. Neurons are selective for distinct features of the stimuli that fall within their RFs, such as discriminative touch, deep vibration, or sensations of heat or cold. In primates, the primary somatosensory cortex (S1) is subdivided into four distinct areas, namely 3a, 3b, 2 and 1, with neurons in each area being specialized for different stimulus features. Each cortical area contains a complete, independent map of the body surface that obeys the topological constraints of the body. Thus neurons that lie close together in the cortex have RFs that lie close to each other on the body, and RFs obey body boundaries, forming contiguous patches of the body and never consisting, for example, of unconnected islands on different fingers or different parts of the body. For touch stimuli, humans have the best two-point discrimination (i.e. the shortest separation between two light taps on nearby points that can still be told apart as distinct) on the lips and the fingertips, where the primary somatosensory cortical (S1) RFs are finest. On the arms, legs and body trunk, for which the

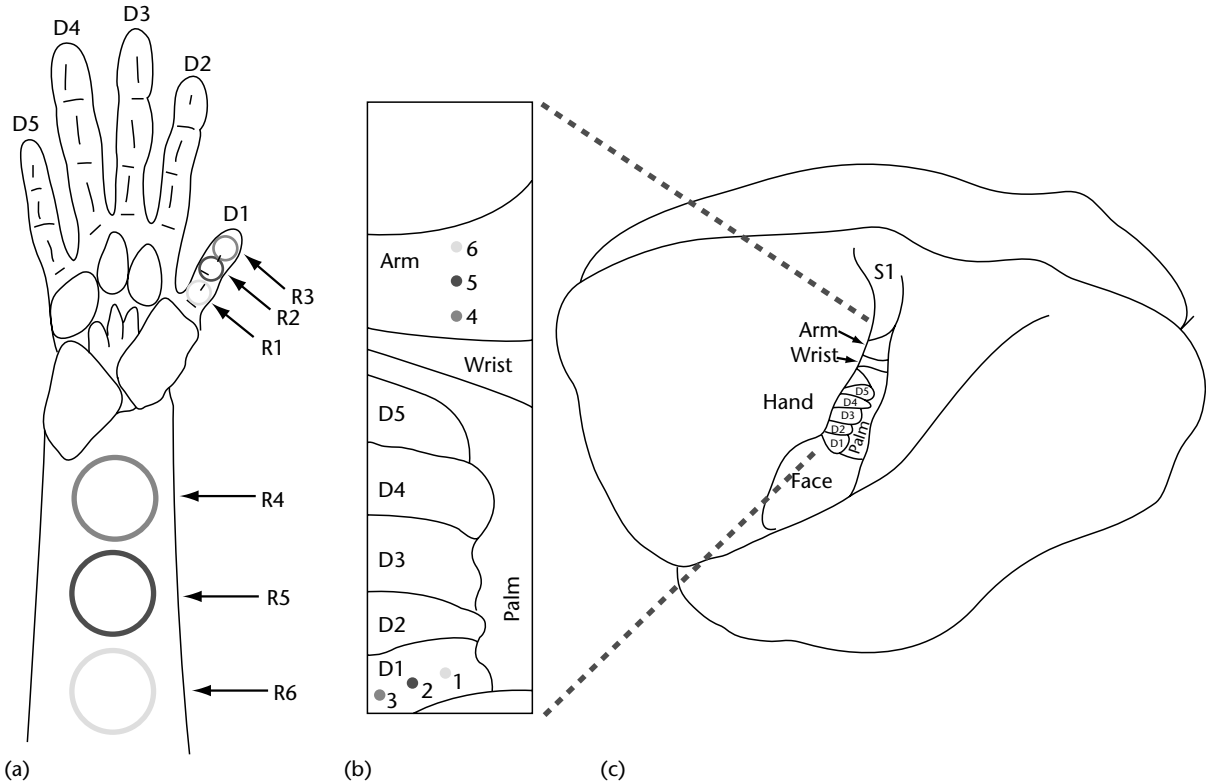


Figure 1. RFs and map layout of somatosensory cortex. (a) Schematic of an owl monkey arm. (b) Expanded schematic view of the region of primary somatosensory cortex (S1) that represents the arm, the wrist, and the palm and fingers of the hand. Cortical points 1, 2 and 3 represent three neurons with just non-overlapping RFs R1, R2 and R3 on the thumb (D1, i.e. digit 1). Points 4, 5 and 6 represent three neurons with just non-overlapping RFs R4, R5 and R6 on the arm. As cortical magnification is inversely proportional to RF size, points 4, 5 and 6 have the same separation between them, on the cortex, as points 1, 2 and 3. Thus the cortical point images due to stimuli at R1 and R2 will be as well separated as those due to stimuli at R4 and R5 – giving as good a two-point discrimination between light touches at points R1 and R2 as at R4 and R5. (c) Side view of the owl monkey brain showing the location of primary somatosensory cortex (S1), region 3a. (Modified from Kandel *et al*. (1991) *Principles of Neural Science*.)

RFs are much coarser, such spatial discrimination is correspondingly poorer (Figure 1).

Auditory system (hearing)

In the vertebrate auditory system, RFs for neurons early in the auditory pathway can be defined in terms of pure tone frequencies. Each neuron has a characteristic tuning profile with sharp tuning to a best frequency. At each stage in the auditory pathway, from the cochlea in the inner ear to the auditory cortex, neuronal best frequencies form smooth tonotopic maps that are laid out on an exponential scale, with frequencies increasing in geometric progression per unit distance on the map. Thus neurons that are equally spaced apart on a map have best frequencies that are multiples of each other. This probably explains why we perceive the frequency (pitch) scale in octaves (i.e. factors of 2), with tones an octave apart appearing to be equally spaced. Furthermore, a fixed small separation on the map translates in to a frequency difference that is not fixed, but rather which is proportional to the frequency at that point on the map. This explains why the frequency limen or just noticeable difference (JND) between tones (i.e. the minimum frequency difference required to tell two pure tones apart) is roughly proportional to the frequency of the tones.

Visual system

RFs in the visual system will be discussed in more detail in the section on receptive field organization. In the context of sensory discrimination, acuity is finest near the center of vision, the fovea, where the primary visual cortex (V1) RFs are smallest, and it drops off rapidly on moving away from the fovea, with increasing RF size. As in the somatosensory cortex, the cortical magnification factor (here the cortical territory devoted to unit angle of visual space) is inversely proportional to the RF size for that region of visual space.

Some regions of the brain contain RFs and maps in multiple sensory modalities. The superior colliculus (also known as the optic tectum), which has seven distinct neuronal layers, contains simultaneous visual, auditory, somatosensory and motor maps of space. These maps are spatially aligned over each other in the different layers. For example, neurons in the superficial layer that are visually driven with RFs at a particular location in space are located above neurons in the deeper auditory map that localize auditory stimuli to the same region of space. These sensory maps are aligned with the motor map in a yet deeper neuronal layer that guides the eyes to make saccadic (high-velocity) movements to the same point in space.

The relationship between individual neuronal RFs and sensory discrimination is not limited to such elementary sensory attributes as position or pitch, but extends to many complex properties of sensory stimuli. For example, our ability to discriminate between directions of motion of moving objects is linked to the direction-tuned responses of single neurons in the MT cortical region, part of the motion-detecting system in the visual cortex. Bill Newsome and his colleagues recorded from MT in monkeys that were shown patterns of moving dots and were trained to indicate which way they thought the dots were moving. Newsome showed that it was possible to predict the response of a monkey quite accurately from the responses of single MT neurons in whose RFs the pattern of moving dots fell.

RECEPTIVE-FIELD ORGANIZATION

In this section we shall trace the progressive transformation of RF sizes and response properties along the visual processing pathway – from stage to stage within the retina, and then via the lateral geniculate nucleus to the primary visual cortex and beyond. This detailed description will illustrate the following points.

- RF properties grow systematically more complex along a sensory processing pathway.
- The response properties at each stage can be explained on the basis of the underlying neural circuitry and the geometrical arrangement of inputs from previous stages.
- At each stage, RFs and response properties are well designed for extracting biologically important information and disregarding biologically irrelevant features.

The next section, on feedback and non-classical effects, will place caveats on some of the principles laid out in the current section. In particular, RFs do not show a strict hierarchy of complexity when moving along a sensory processing pathway. Rather, neurons at one cortical stage are strongly influenced by feedback within the same region as well as from 'higher' regions. However, it is very important to appreciate the simplified hierarchy of RFs and the neural circuitry that transforms RFs from one processing stage to the next, since much of the feedback and 'non-classical' effects appear to be built over the simpler hierarchical framework.

The emphasis here is on the visual system in primates and cats, since this is the system that has

been studied most thoroughly. Similar principles are expected to apply to other sensory systems.

The Retina: Photoreceptors to Bipolar Cells to Ganglion Cells

Photoreceptors (i.e. rods and cones) transduce light into electrical signals which are in turn transmitted to bipolar cells. Here the unimodal photoreceptor RFs are transformed into pairs of 'on' and 'off' RFs with (weak) concentric opponent surrounds (i.e. inhibitory input, mediated through horizontal cells, from neighboring photoreceptors on the retina). Bipolar cells feed into retinal ganglion cells (RGCs), with RFs that are 'on'–center 'off'–surround or 'off'–center 'on'–surround. 'On' bipolars feed 'on'–center RGCs and 'off' bipolars feed 'off'–center RGCs. The concentric RGC surrounds come from surrounding bipolar cells of the opposite type, mediated through amacrine cells. (Some RGCs, which carry blue–yellow signals, are not center–surround. Rather, they are blue 'on', from a blue-cone bipolar cell input with an overlapping yellow 'off', from a mix of red- and green-cone bipolar cell inputs. The blue-cone pathway is believed to be evolutionarily much older than the red- and green-cone pathway, as all mammals possess blue cones, while only the Old World primates, great apes and humans have red and green cones.)

RF sizes and distributions are matched to the optics of the eye. The aperture of the primate eye is such that the size and spacing of photoreceptors is well matched to the size and separation of just resolvable dots at the diffraction limit. Thus the photoreceptors, which form a close-packed mosaic over the retina, tile visual space fully and with optimal resolution. In the next stage up, each (foveal) cone photoreceptor connects to one 'on' and one 'off' bipolar cell. This generates parallel channels of opposite polarity looking at each point in a scene, each channel forming a complete mosaic that independently covers visual space. Similarly, 'on'–center and 'off'–center RGCs independently tile all of visual space.

The center–surround RF structure of RGCs is particularly suitable for two important and related tasks. First, it selectively identifies object boundaries in the visual input. Secondly, as a corollary, it compresses information for the bottleneck of the optic nerve. RGC centers and surrounds show a fine balance of excitation and inhibition. Thus RGCs respond vigorously to a contrast step (i.e. to brightness or color contrast within the RF) while ignoring uniform fields or smooth gradients. Visual input to the eye is thus tremendously compressed in the RGC response without losing much information since, in principle, it is only necessary to know edge positions and contrast in order to 'fill in' the rest and reconstruct an image. This high degree of visual processing and image compression in the retina (as opposed to further along the visual pathway) is likely to have evolved as a strategy for passing information efficiently along the optic nerve (i.e. the axonal output of RGCs). The optic nerve is constrained to be as lightweight and flexible as possible and with as small a cross-section as possible, to allow the eyes to move around rapidly and track moving objects. The image compression from photoreceptor to RGC allows the optic nerve to carry a large amount of biologically relevant information despite such constraints.

Having parallel 'on' ('on' bipolar to 'on'–center RGC, etc.) and 'off' channels looking at each point in space provides a high signal-to-noise ratio while also conserving metabolic energy. When scanning a visual scene, the number of visual edges that are encountered with a step-up in brightness equals, on average, the number of edges with a step-down in brightness. 'On'–center cells are quiet except at a step-up in brightness, while 'off'–center cells are quiet except at a step-down – each channel uses the full dynamic range of firing rate to signal the size of the step up or down. Over a region that is smooth and free from visual features, both channels are quiet, giving a stable baseline that signals 'no features'. If we had one common channel rather than separate 'on' and 'off' channels, then neurons would need to have a high, constant baseline level of firing, with increases and decreases signaling steps up and down, respectively. This would halve the dynamic range (i.e. from the midway point to maximum or zero), while being a constant drain on metabolic energy. Moreover, since neuronal firing is a noisy stochastic process, the baseline of constant, high neuronal activity would give a very noisy and variable measure of 'no features detected'. (As an aside, the existence of parallel and equivalent pathways for 'on' and 'off' is the likely reason why 'black' is perceptually as tangible a color as 'white', even though the two are physically so different – black being a reduction and white being an increase in light energy on the retina.)

Lateral Geniculate Nucleus (LGN)

RGC axons synapse with relay neurons in the lateral geniculate nucleus (LGN) with center–surround RFs very similar to the RGC RFs, with one addition, namely a systematic set of response delays. One class of LGN neurons relays RGC

inputs unchanged, while the class of lagged LGN neurons introduces a delay of 30 to 40 ms through local inhibitory interneurons. Lagged and direct LGN inputs are combined in the cortex in such a manner as to make cortical cells selective for the direction of motion of moving stimuli. By applying the principles of information theory, Joseph Atick and his colleagues showed that the range of LGN time lags was just right for an unbiased discrimination of visual motion over the range of speeds encountered in natural scenes.

Primary Visual Cortex (V1)

Neurons in the primary visual cortex (V1) – that is, the first cortical stage of visual processing – combine LGN inputs to generate RFs that are selective for oriented edges. In cat V1 this transformation occurs at the cortical input neurons whose RFs, described as 'simple cells' by their discoverers David Hubel and Torsten Wiesel, consist of alternating oblong excitatory and inhibitory subfields (unlike the concentric subfields of LGN RFs). Many models have been proposed, involving varying degrees of intracortical interactions, to explain the generation of this novel RF shape in V1. Current evidence favors Hubel and Wiesel's original suggestion, namely that each V1 neuron receives input from LGN cells whose RFs line up to form the oblong simple-cell subfields.

A number of RF response properties arise for the first time in V1. V1 neurons respond selectively to oriented lines. The orientation preference and sharpness of tuning is given accurately by a linear summation – or Fourier transform – of the RF subfields. In addition, these neurons respond selectively to a specific direction of motion of the oriented line, a property that can also be accounted for, at least qualitatively, by the linear combination of LGN inputs with their known phased time delays. Binocular RFs arise in V1 by combining, for the first time, inputs from the two eyes. The disparity or relative positional mismatch between the individual monocular inputs into a binocular RF is believed to inform depth perception. Simple-cell RFs are often modeled as Gabor functions, sinusoids defined over space and time (to incorporate directionality), with the appropriate orientation and period, within a Gaussian profile that determines the total extent of the RF. Disparity is modeled by introducing a relative phase shift between the Gabor sinusoid for each eye.

Simple-cell RFs are likely to be optimal for parsing natural scenes, which statistically consist of relatively redundant and featureless surfaces bounded by extended edges. A computer neural network that is instructed to form a sparse, optimal set of filters for picking out defining visual features and disregarding redundancies produces patterns remarkably reminiscent of simple-cell RFs. This suggests that simple-cell RFs are adapted to the statistics of the natural world. It also leads to the interesting question of whether these RF properties are the result of evolution and 'hard-wired' in the individual, or whether they are more plastic, adapting to the details of an individual's sensory history.

The architecture of V1 plays a critical role in the visual processing in this region. Neuronal RFs and response properties are organized systematically both in vertical depth and over its surface. Vertically, RF responses are organized in 'columns' that traverse all six cortical layers. All of the neurons in a column have the same RF location, orientation and color preference, etc., but neurons in the upper layers are more sharply orientation-tuned than those in the input layer, through interactions within the column. Across the surface of the cortex, RFs form smooth maps of space as well as of every other response property, and neurons that lie closer to each other on the cortex are more similar in RF position, preferred orientation, preferred color, binocularity, etc. (Figure 2). Each map is independent, though with a common periodic scale length such that each response property cycles through a full range of values over a cortical distance corresponding to a shift of one RF dimension in space. Thus V1 functions like a set of parallel computational modules over the cortex, simultaneously processing the input from each RF-sized tile of visual space, each module of cortical circuitry performing identical analyses over its segment of the image. Furthermore, the position of a neuron on the cortical map determines the response properties of other neurons within reach of its axonal and dendritic arbours, and thus the nature of the effective cortical circuits that determine the response of the neuron to complex visual stimuli (see the section below on feedback and non-classical effects).

Receptive Fields Beyond the Primary Visual Cortex

The primate visual cortex has about 40 distinct processing stages, which are organized roughly hierarchically, but with massive feedback from 'later' to 'earlier' stages as well as 'horizontal' interconnections between stages at the same level. These later stages form two parallel processing streams flowing out of V1, one of which is mainly for

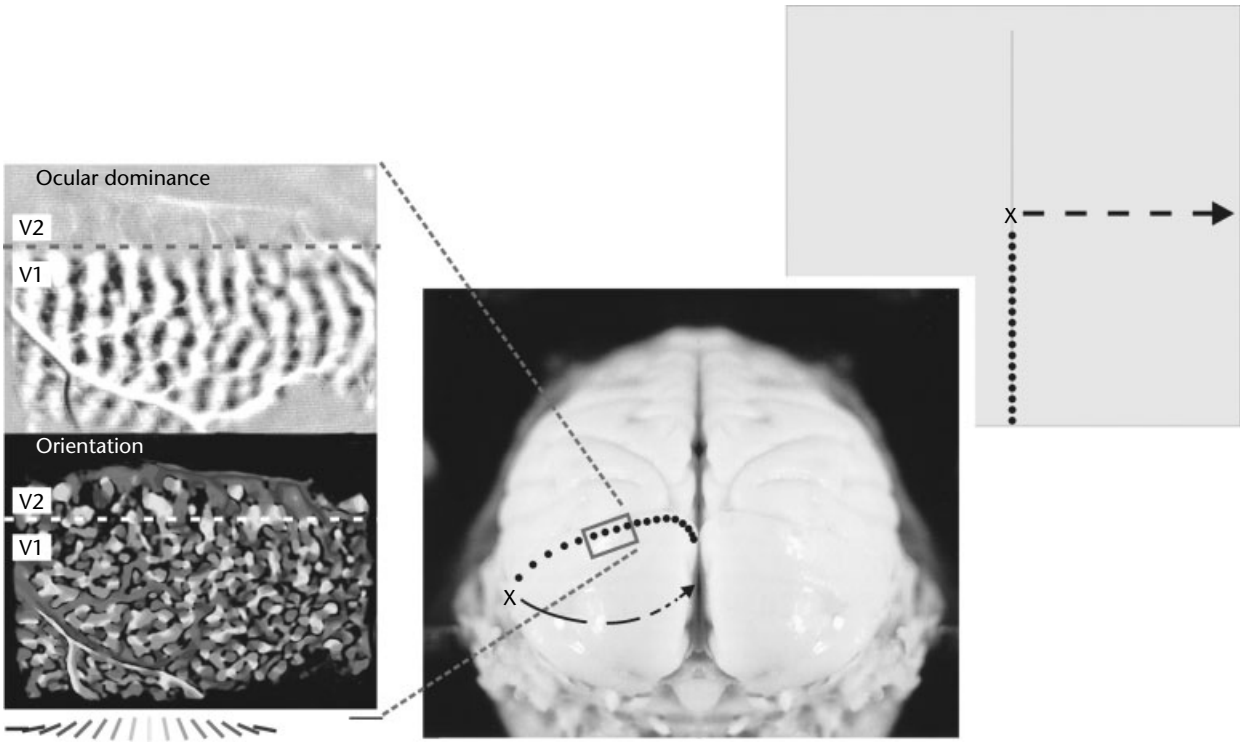


Figure 2. Layout of RFs on primary visual cortex. Center: the brain of a rhesus monkey, seen from the back. The large smooth region bounded by the dotted line on top is primary visual cortex, V1. If the monkey was looking at the screen shown on the right, with the letter X in its line of sight (fovea), then the X, the arrow and the lines on the screen would map on to the brain as shown. Note that even though the dots and dashes on the screen are uniform in size and separation, their projections on to the brain are magnified near the fovea and compressed near the periphery. (From Hubel DH (1988) *Eye, Brain and Vision*. Scientific American Library.) Left: optical images of ocular dominance (top) and orientation (bottom) obtained from the region of rhesus V1 similar to that enclosed in the red rectangle. In the ocular dominance map the dark green areas denote regions of V1 that are responsive to the right eye, while the white areas denote regions that are responsive to the left eye. In the orientation map, blue denotes those regions that respond best to horizontal lines, red denotes those responding to 45° lines, yellow denotes those responding to vertical lines and green denotes those that respond best to 135° lines. Note that the periodicity of the orientation map is similar to the spacing of the ocular dominance stripes.

extracting motion information, while the other is mainly for extracting form information (albeit with intercommunication between these two streams). Along both streams RFs become larger and progressively more complex. Thus in MT, the first stage in the 'motion' pathway, neurons not only respond to single moving edges as in V1, but they can also integrate motion cues over an extended region to signal the net motion of an object. In region MST, further along the motion pathway, neurons have RFs many tens of degrees in extent that are selective for complex motion patterns such as radial outflow or inflow and vortices. In the 'form' pathway, neurons in V4 respond better to specific feature combinations (e.g. a red triangle, or a green square) than to line edges. Neurons in the more advanced regions of the same pathway, such as in higher visual centers in the inferior temporal (IT) cortex, respond selectively even to such complex forms as faces. Specific excitatory and inhibitory circuits have been identified in some of the higher areas (e.g. TE) that can account for the complex feature selectivities of neurons in terms of interactions between neighboring groups of neurons with simpler response properties.

FEEDBACK AND NON-CLASSICAL EFFECTS

Current evidence suggests that the concept of strictly hierarchical sensory processing pathways discussed so far is an over-simplification. Rather, sensory perception appears to involve simultaneous processing of sensory features on a number of different scales, with interactions between different stages on the neural pathway.

At a perceptual level, our visual systems appear to be tuned to pick out not simple line elements, but rather features of intermediate complexity such as long contours, object boundaries or depth relationships between surfaces. These complex contextual effects often determine our perceptions of nominally simpler elements such as edges or surface colors. Moreover, our perceptions are strongly influenced by our state of attention or expectation, and they can change over time with specific perceptual training.

At a physiological level, the response of a neuron – even in V1 – can change by up to an order of magnitude in the presence of stimuli in the surround outside the neuron's RF. Furthermore, even V1 neurons are influenced by complex non-visual contexts such as the state of attention or the particular task in which an individual is engaged. The simple response properties of orientation, direction, color or disparity described in the last section are not adequate to describe neuronal response properties fully. Rather, these should be seen as forming a repertoire of components from which the more complex response of a neuron is put together. Again, most of the discussion will focus on the visual system, but only because this is the system that has been studied most thoroughly.

The Range and Variety of Modulatory Influences

Even in their earliest work on V1, Hubel and Wiesel noted that some V1 neurons had inhibitory zones outside the RF boundaries. For these neurons – which they labeled hypercomplex (now described as 'end stopped') – the response to an optimal bar stimulus inside the RF would start to diminish once the bar was lengthened beyond a critical value and started to encroach on the inhibitory end zones. Unlike RGC or simple-cell 'off' subfields, where a stimulus in the subfield alone suppressed the spike rate below spontaneous, these zones were 'silent' and nonlinear – a bar placed in the end zone alone had no effect.

Work in the 1970s by Colin Blakemore and Lamberto Maffei led to a distinction being made between the classical RF, as plotted using compact high-contrast stimuli, and the non-classical surround. A simple bar or grating stimulus inside the classical RF is adequate to activate a neuron. However, a stimulus in the non-classical surround has no effect by itself on a neuron's response, but facilitates or suppresses the neuron's response to a stimulus inside the classical RF.

A number of what were believed to be basic response properties of V1 neurons can be modified by the visual context. A neuron's orientation tuning can change scale, or can even be 'repelled' to a different preferred orientation by a surrounding field of oriented lines, in a manner reminiscent of tilt illusions. The RF size can be made effectively indefinite either by reducing the contrast of the stimulus bar, or by surrounding the RF with a noisy background. This appears to inactivate the inhibitory end zones of end-stopped neurons.

Modulation by visual context can be seen over a wide range of spatial scales and temporal latencies. Simple surround stimuli (small groups of lines or other figures in the immediate non-classical surround) lead to inhibition or facilitation with no perceptible delay. On the other hand, the modulatory effects of more complex attributes or background patterns show up after a delay of many tens of milliseconds. For example, Zipser and Lamme showed that the neuronal response to a given texture is on average higher if it is part of a closed 'figure' or 'foreground', rather than a 'background'. On another complex task, Roelfsema showed that the responses of V1 RFs to an optimally oriented line were consistently stronger if the animal was attending to the line rather than to a distractor. For both of these tasks, the difference only shows up late, around 80 to 100 ms after the start of the neuronal response.

Anatomical Substrates Underlying Non-classical Modulation

Each V1 neuron receives an estimated 20 000 to 40 000 synaptic inputs that could include LGN afferents, connections from other V1 neurons and feedback from higher visual areas. The LGN input – to neurons in the input cortical layers – is likely to set up the elementary response properties, while intra-V1 and feedback connections are likely to be responsible for the modulation by a larger visual context.

Intrinsic connections within V1

Pyramidal cells – that is, the excitatory neurons which constitute about 80% of the cortical population – form an extensive but precise network of connections with other excitatory and inhibitory neurons over V1. Locally they connect indiscriminately with neurons lying within range of the local arbour. At greater distances out to around 5 to 9 mm in the cortex their axonal collaterals connect selectively only with other neurons that are tuned to the same preferred orientation. However, the RFs of these distant target neurons are centered on very different parts of the visual field, covering

an area 10 times the size of a single RF. Thus the nearby connections sample a large range of orientations at the same point in space, while the distant connections sample a large visual context but at the same orientation. Through the orientation-selective distant interconnections, neurons with RFs along a smooth contour facilitate each other's activity, a process that is reflected in the way in which a flanking line can facilitate a neuron's response to (or our perception of) a dim-edge stimulus. This is likely to underlie our ability to pick out smooth curves from a jumbled background, and thus rapidly parse spatial scenes into contours and boundaries.

Feedback from higher visual areas

Neurons in V1 also receive extensive feedback connections (accounting for around 50% of the synaptic input, and arborizing at least as widely inside V1 as the intrinsic long-range connections) from areas V2, V3, V4 and MT. Feedback connections are likely to underlie the modulation of V1 responses by complex attributes of a figure, such as 'foreground' versus 'background', the level of attention being paid, or the task being performed. For example, Bullier and his colleagues showed that MT feedback is involved when V1 responses are modulated by large-scale motion. The response of a V1 neuron to a moving bar is enhanced when the bar is shown against a textured background moving in the opposite direction. This enhancement is reduced sharply and selectively on cooling and thus inactivating MT, the first stage in the 'motion detecting pathway'. However, cooling MT has no effect on the modulation of V1 responses by stationary backgrounds.

DYNAMIC PROPERTIES OF RECEPTIVE FIELDS

Recent research has shown that RF response properties do not form permanent, defining attributes of a neuron. Rather, even in the adult, RFs are dynamic and can change significantly, particularly following discrimination training or selective stimulation.

History: Phantom Limbs and RF Shifts following Surgical Manipulations

Plasticity of adult RF properties was first noted in the clinic through the phenomenon of phantom limbs. An amputee who has lost an arm retains a sensation of a 'phantom' arm that can feel pain and assorted bodily sensations. What is more interesting is that the face starts to develop a topographically accurate sensory map of the missing limb. For example, running a cotton wick down the patient's face generates the strong sensation of having it run down the missing arm, and the patient can even discriminate the phantom fingers that are being touched.

The systematic study of adult cortical plasticity, in monkey somatosensory cortex (S1), started in the 1980s. Mike Merzenich, John Kaas and their colleagues showed that surgically manipulating the sensory input coming to S1 from an animal's fingers (e.g. removing some input by amputating a finger, or shuffling the input by transposing an island of innervated skin from one finger to another, or synchronizing the input abnormally by suturing two fingers together) leads to dramatic shifts in the RFs of S1 neurons, and corresponding shifts in the cortical map of the hand and fingers. Similar surgically induced RF shifts were shown in the auditory system, where surgical ablation of a segment of the cochlea leads to shifts in the frequency map in the auditory cortex as well as subcortical stages (e.g. the inferior colliculus). Such RF shifts were also shown in the visual system, where a small scotoma or lesion of the retina leads to a remapping of visual space on V1. These RF shifts are very rapid, involving an immediate expansion by more than 10-fold in size, followed by a later consolidation to a new position (Figure 3).

Dynamic Changes in Receptive Fields and Perceptual Learning

The changes in RF position and size after lesions or injuries are just an extreme manifestation of the potential for dynamic RF plasticity which could last throughout life, and that presumably underlies our ability to learn new sensory discriminations.

RFs can be modified dynamically through passive sensory experience. For example, an 'artificial retinal scotoma' (a visual stimulus consisting of a blank patch within a bright and busy surround, simulating the effect of a lesion on the retina) causes V1 RFs inside the 'scotoma' to expand by 10-fold within seconds. This RF expansion, which can be reversed by stimulating inside the blank 'scotoma', is also believed to be related to the phenomenon of 'visual fill-in', where fixating steadily at a busy visual pattern with a small blank 'blind spot' causes the surrounding pattern to fill in the blind spot.

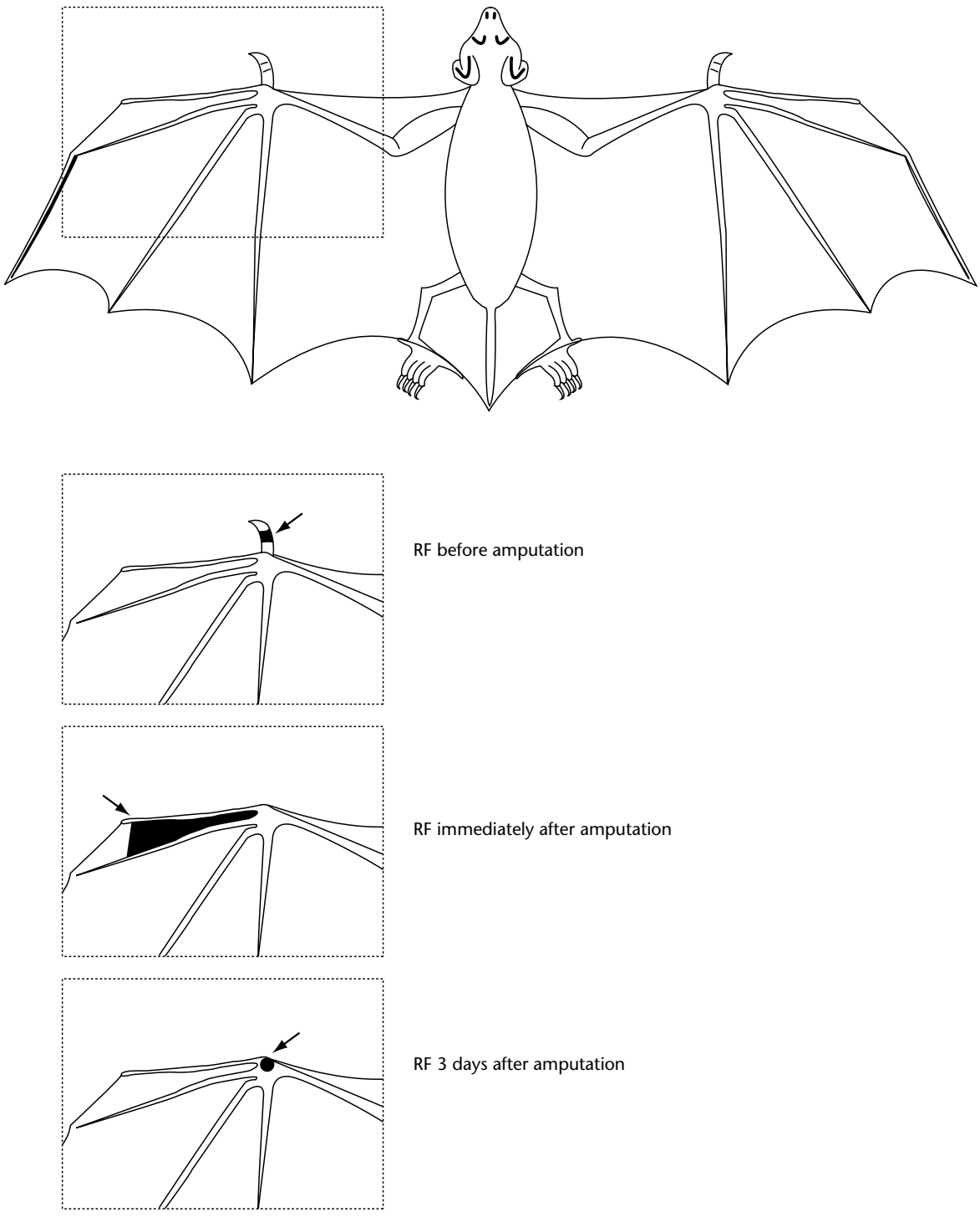


Figure 3. After a lesion in the periphery, RFs of neurons in the affected region of somatosensory cortex expand dramatically and rapidly. Then, over a period of days, they consolidate and shrink back to a new position bordering the lesion. The thumb of a flying fox (sketched at the top of the figure) was amputated after recording electrodes had been implanted in the somatosensory cortex to monitor RF size and position. The RF (see arrow pointing to RF, shown in red) of a typical neuron that received input from the thumb is shown in the sequence of sketches of the animal's forelimb. Immediately after amputation of the thumb, the RF of this neuron expanded to include a large area of the wing surface between digits 2 and 3. After 3 days, the neuron's RF had shrunk back to a more normal-sized region of the limb surface bordering the missing thumb.

RF shifts appear to be most marked and best consolidated when induced by specific sensory discrimination tasks. Furthermore, the magnitude of the shift is related to the degree of improvement of an individual's sensory discrimination thresholds. Recanzone and Merzenich trained monkeys to discriminate between frequencies of tactile vibration of a stylus pressed against a fixed spot on one finger and vibrating at a frequency (*c*. 25 Hz) higher than the normal flutter-frequency discrimination limit. At the end of the training, once the animal had improved its detection threshold manyfold – but *only* on the specific trained spot on that finger – not only did S1 show an abnormally large number of neurons with RFs on the trained spot, but also cortical neurons had developed an entirely new repertoire of responses, phase-locking to the unusually high frequency of the vibrating stylus. Training can induce changes in RF properties that are quite subtle. For example, training in a visual discrimination task leads to changes only in the modulation of a V1 neuron's response by the surrounding visual context. Furthermore, and even more interestingly, this changed RF response property was only manifested when the trained animal was engaged in the discrimination task, and not when it was passively shown the same stimulus.

Underlying Neural Networks and Mechanisms

Plastic changes are believed to be mediated by the same extensive neural networks that are responsible for feedback and non-classical effects. As was mentioned earlier, V1 pyramidal neurons connect with other neurons over the extensive range of their axonal collaterals, corresponding to a region of visual space more than 10 times larger than a single RF. Similar extensive networks of connections are seen in the somatosensory and auditory pathways. In each case the network for a given neuron involves tens of thousands of synapses with other neurons (both excitatory inhibitory). The tight constraints on the RF size, position and response properties measured at any given moment evidently involve a fine balance between excitation and inhibition over the extended input network to the neuron. The sudden imbalance caused by removing part of the input to the network (e.g. by surgically amputating a digit) (Figure 3) is presumed to underlie the rapid RF expansion that is seen. Furthermore, just as nonlinear interactions within the input network mediate the influence of a sensory surround on a neuron's RF, long-term plastic changes in synaptic strength in the network lead to long-term changes in RF properties.

The primary contributions to RF plasticity come from different stages of the pathway in the different sensory systems. In the visual system, the cortex (but not the LGN or retina) has extensive convergent and divergent connections where shifts in synaptic balance can alter RF properties significantly. In V1, the long-range axonal collaterals are involved in RF plasticity. However, little plastic change is seen, either in the retina or in the LGN. In the auditory and somatosensory systems, by contrast, many subcortical centers are capable of plastic change. For example, amputation causes extensive rewiring and RF shifts even at the dorsal horn of the spinal cord.

Modifying the Concept of Receptive Fields

With the complex sets of neural inputs that contribute to a neuron's response properties, many believe that the term 'receptive field' should actually incorporate the entire rich set of responses of which a neuron is capable – not just the 'classical' RF, but all of the contextual modulations and potential plastic modifications.

Further Reading

Gilbert CD (1996) Plasticity in visual perception and physiology. *Current Opinion in Neurobiology* **6**: 269–274.

Hubel DH (1988) *Eye, Brain and Vision*. New York: Scientific American Library.

Kandel ER, Schwartz JH and Jessel TM (eds) (2000) *Principals of Neural Science*, 3rd edn. New York: McGraw Hill.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: WH Freeman & Co.

Rakic P and Singer W (eds) (1988) *Neurobiology of Neocortex: Report on the Neurobiology of Neocortex*. New York: John Wiley.

Reinforcement Learning: A Biological Perspective

Intermediate article

P Read Montague, Baylor College of Medicine, Houston, Texas, USA
David M Eagleman, Salk Institute, La Jolla, California, USA
Samuel M McClure, Baylor College of Medicine, Houston, Texas, USA
Gregory S Berns, Emory University School of Medicine, Atlanta, Georgia, USA

CONTENTS

Reinforcement learning is an approach to learning problems that takes account of the interaction of the learner with their reactive environment. Despite its early connections to animal learning, reinforcement learning has been an active field of study in engineering and computer science. New research is now starting to connect reinforcement learning to identifiable biological systems. These connections may well provide direct insights into goal-directed learning and decision-making in biological organisms.

INTRODUCTION

Brain research during the last 50 years has focused on the metaphor of learning through interactions with an environment. The central idea here is interaction. All mobile creatures, including humans, are embedded in a rich, reactive environment. We move our eyes – the visual scene changes. We push against a tree branch – we feel its roughness, and it even pushes back. We shift our attention – suddenly a sight or sound becomes clearer and easier to interpret. Even our bodies act as part of the reactive environment, as mere consideration of an action can cause a change in an animal's physiological state whether or not it actually moves. The learning mechanisms that are embedded in our nervous systems have evolved to deal with just such reactive aspects of environments.

Reinforcement learning takes account of the idea that learners are situated in real-world settings which react to and may even adapt to the actions of the learner. In the words of modern reinforcement learning pioneers Sutton and Barto 'Reinforcement learning is learning what to do – how to map situations to actions – so as to maximize a numerical reward signal.'

In this framework, learning is specified as a complete problem of an agent interacting with an environment. Thus the properties of the agent and the environment and the dynamics of interaction must be specified in a reinforcement learning problem. As a class of problems, reinforcement learning problems are most closely related to mathematical problems of optimal control. One of the most important ideas that is built into all reinforcement learning systems is the concept of a *goal-seeking agent*. In reinforcement learning, the learner possesses goals that influence their selection of actions, the way they update their memory, and so on. From a biological perspective, the critical issue is to determine the physical mechanisms that define and control the goals and actions of the learner.

In this article we shall outline one computational approach to reinforcement learning and its biological realization in living systems. Reinforcement learning appeals to the computational community because it can be formally written down as equations or simulations. It appeals to biologists because of its plausibility as an architecture and its growing connection to real biological data. The discovery that reinforcement learning is a powerful approach to machine-learning should come as no surprise. Biological systems appear to have long exploited the approach to solve many problems of real-time adaptation in complex environments.

THE THREE BASIC COMPONENTS OF REINFORCEMENT LEARNING

Every reinforcement learning system has three basic components: (1) a reward function, (2) a value function and (3) a policy. These relatively

abstract terms capture the idea of immediate evaluation (reward function), long-term judgment (value function) and action selection (policy).

The *reward function* formalizes the idea of a goal for a reinforcement learning system. It assigns to each state of the agent a single numerical quantity – the reward. The reward function defines what is good 'right now', and can be viewed as a built-in assessment of each state that is available to the agent (learner). It is also used to define the agent's goal – to maximize the total reward.

The *value function* formalizes the notion of longer-term assessments (judgments) of each state of the agent. It provides a valuation of the current state of the agent, taking into account the succession of states that could follow. Formally, for each state, value is defined as the total amount of reward that the agent can expect from that state onward into the distant future. These values would have to be stored in some fashion within the agent. In practice, the learner uses the reward function to improve their internal estimate of the value function.

In shorthand, rewards are immediate and values are long-term. For example, a rat may take many steps across an electrified grid (low reward) to reach food (high reward). All of those intermediate states (steps on the grid) have a very low reward, but possess high value because they lead directly to future states with food (high reward).

A *policy* formalizes exactly what the word implies – 'given this, do that'. Formally, a policy maps states to actions. In both biological and machine-learning examples, a policy is usually probabilistic. For a given state, a policy defines the probability of taking one of many possible actions in order to end up in one of many succeeding states. For example, a rat seeking food in a maze encounters a three-way junction. The policy assigns probabilities separately to each of the three actions that are available to the rat.

TRIAL AND ERROR

The idea of trial-and-error learning is familiar to everyone – to solve a problem, try some solution, assess how well it performed, and try again if it did not perform up to expectations. Reinforcement is intimately linked to trial-and-error learning. The basic idea of reinforcement derives from psychology, took its clearest form initially as Thorndike's law of effect, and depends on the concept of reinforcers. There are two types of reinforcers – positive and negative. The law of effect is simple. Actions or internal states followed by positive reinforcers are later more likely to occur, and actions or internal states followed by negative reinforcers are later less likely to occur. In one form or another, this idea has become one basis (almost a principle) for thinking about trial-and-error learning. An animal tries an action, assesses its success in terms of positive and negative reinforcement, and adjusts its later likelihood of taking that action.

The idea of trial-and-error learning is also central to reinforcement learning. Trial-and-error learning systems have been investigated from the earliest days of artificial intelligence. In fact, in 1954 one of the pioneers of artificial intelligence, Marvin Minksy, wrote his PhD thesis (*Theory of Neural-Analog Reinforcement Systems and its Applications to the Brain-Model Problem*), at Princeton University on reinforcement learning. In addition to his many other contributions, Minsky was one of the first to specify clearly the central problem in trial-and-error learning where some type of reward signal is used to criticize the outcome of a series of actions. This problem is known as the *temporal credit assignment problem*.

Temporal credit assignment is intuitively familiar. For example, an animal seeking food (high reward) may make many decisions before actually acquiring the food and receiving a large reward signal. A natural question then arises. How does the animal's brain assign credit to each of the individual decisions? Some decisions are critical to eventually obtaining food, while others may have been completely irrelevant. This type of problem arises whenever the rewards received are delayed in time from the events that lead to reward. In any real-world setting, an animal must *learn associations through time* (e.g. 'I go left here' is associated with 'I get three units of reward' an hour later).

Depending on the situation, there are numerous ways to solve the temporal credit assignment problem. One way to achieve trial-and-error learning using a delayed reward signal is *temporal difference (TD) learning*.

TD LEARNING

Here we shall describe in detail the way in which temporal difference (TD) learning frames and solves the temporal credit assignment problem described above. We shall begin with a description in terms of animal learning – that is, at the level of behavioral learning. In the later sections of this article we shall show that this same formal description can account for the pattern of activity that is recorded in midbrain dopamine neurons while alert animals are actively learning. The latter connection is important because it suggests that

dopamine neurons may be one part of a mechanism that implements reinforcement learning in mammals.

Animals are taken as living in what is technically known as a Markov decision problem. This construct formalizes the characteristics of problems such as maze tasks, in which there are different states (e.g. different locations in the maze), actions (e.g. directions in which to move) and rewards. The rewards can either depend on the animal's actions or be provided without regard to what the animal does (as in classical conditioning). This framework is general enough to model many standard behavioral tasks. The most important assumptions underlying the TD approach involve the nature of the presumed computational task solved by an organism. There are two main assumptions in TD.

Assumption 1

First, the computational goal of learning is to use a set of sensory cues $\boldsymbol{x}(t) = \{x_1(t), _2(t), x_3(t), \ldots\}$ (e.g. characterizing the current state of an organism) to fit a 'value' function $V^*(\boldsymbol{x}(t))$ that 'values' the current state as the *average discounted sum of all future rewards from time t onward*:

$$V^*(\boldsymbol{x}(t)) = E\{\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots\}$$

E is the expected value operator (the average), $r(t)$ is the reward at time t, $r(t+1)$ is the reward at time $t+1$, and so on; γ is a discount factor that ranges between 0 and 1 and captures the idea that rewards in the near future are more valuable than rewards in the distant future. If the true (optimal) $V^*(\boldsymbol{x}(t))$ could be estimated by a system, then the system could use such an estimate to update its internal model of future rewards and future actions predicated on the expected receipt of those rewards. This would give the system a way to simulate possible future action sequences and value them according to their expected long-term returns. In this description, the indices on the x's denote different sensory cues.

Assumption 2

This is the Markovian assumption – that is, the appearance of future sensory cues and rewards depends only on the immediate (current) sensory cues and not on the past sensory cues. This is a relatively restrictive assumption about the structure of the environment. However, it has proved to be useful even in real-world settings.

Adjusting the Predictions (Weights)

The strategy of TD learning is to use a set of sensory cues $\boldsymbol{x}(t) = \{x_1(t), x_2(t), x_3(t), \ldots\}$ present in a learning trial along with a set of adaptable weights $\boldsymbol{w}(t) = \{w_1(t), w_2(t), w_3(t), \ldots\}$ to make an estimate $V(\boldsymbol{x}(t))$ of the true $V^*(\boldsymbol{x}(t))$. In this formulation, the weights act as predictions of future reward. For completeness, we shall include a remark about the weights. The weight associated with each sensory cue (e.g. $w_1(t)$ associated with sensory cue 1) is actually a collection of weights – one for each time point following the appearance of sensory cue 1.

Local Data Anticipate Long-Term Reward

The difficulty in actually adjusting weights to estimate $V(\boldsymbol{x}(t))$ is that the system (i.e. the animal) would have to wait to receive all of its future rewards in a trial $r(t+1), r(t+2), r(t+3)$ etc. in order to assess its predictions. The latter constraint would require the animal to remember over time which weights need changing and which weights do not. Fortunately, there is information available at each instant in time that can act as a surrogate prediction error. This possibility is implicit in the definition of $V^*(\boldsymbol{x}(t))$, since it satisfies a condition of consistency through time:

$$V^*(\boldsymbol{x}(t)) = E\ [r(t) + \gamma V^*(\boldsymbol{x}(t+1))]$$

Since the estimate V satisfies the same condition, an error δ in the estimated value function V (estimated predictions) can now be defined using information available at successive timesteps (i.e. taking the difference between both sides of the above equation and ignoring the expected value operator E for clarity).

$$\delta(t) = r(t) + \gamma V(\boldsymbol{x}(t+1)) - V(\boldsymbol{x}(t))$$

δ is called the *TD error* and it acts as a surrogate *prediction error* signal which is instantly available at time $t+1$. If the estimated predictions are correct, then $V(\boldsymbol{x}(t)) = V^*(\boldsymbol{x}(t))$, and the average prediction error is 0 (i.e. $E[\delta(t)] = 0$). In other words, if the system can adjust its weights (predictions) appropriately, then it can learn to expect future rewards predicted by the collection of sensory cues.

This section is somewhat technical, but it does emphasize the fact that local information can be used to make estimates of reward in the distant future. That is, it presents a method by which the temporal credit assignment problem is solved. It is not the only method for such a task, but its importance has recently been enhanced because this

theoretical description of TD learning appears to match the information that is encoded in the spike trains of midbrain dopamine neurons.

MIDBRAIN DOPAMINE SYSTEMS

In the mammalian brain, midbrain dopamine systems play a major role in reward processing and reward-dependent learning. Recent physiological and computational research strongly suggests that the information which is constructed and broadcast by midbrain dopamine systems represents a prediction error analogous to that described above.

Some of the basic electrophysiological findings are shown in Figure 1. After repeated pairings of visual and auditory cues followed by reward, dopamine neurons change the time of their phasic activation from just after the time of reward delivery to the time of cue onset. In one task, a naive monkey is required to touch a lever following the appearance of a small light. Before training, most dopamine neurons show a short burst of impulses following reward delivery. After training, the animal learns to reach for the lever as soon as the light is illuminated, and this behavioral change correlates with two remarkable changes in the dopamine neuron output. First, the primary reward no longer elicits a phasic response, and secondly, the onset of the (predictive) light now causes a phasic activation in dopamine cell output.

In trials where the reward is not delivered at the appropriate time following the illumination of the light, dopamine neurons are depressed dramatically below their basal firing rate at precisely the time when the reward should have been delivered. This well-timed decrease in spike output shows that the expected time of reward delivery based

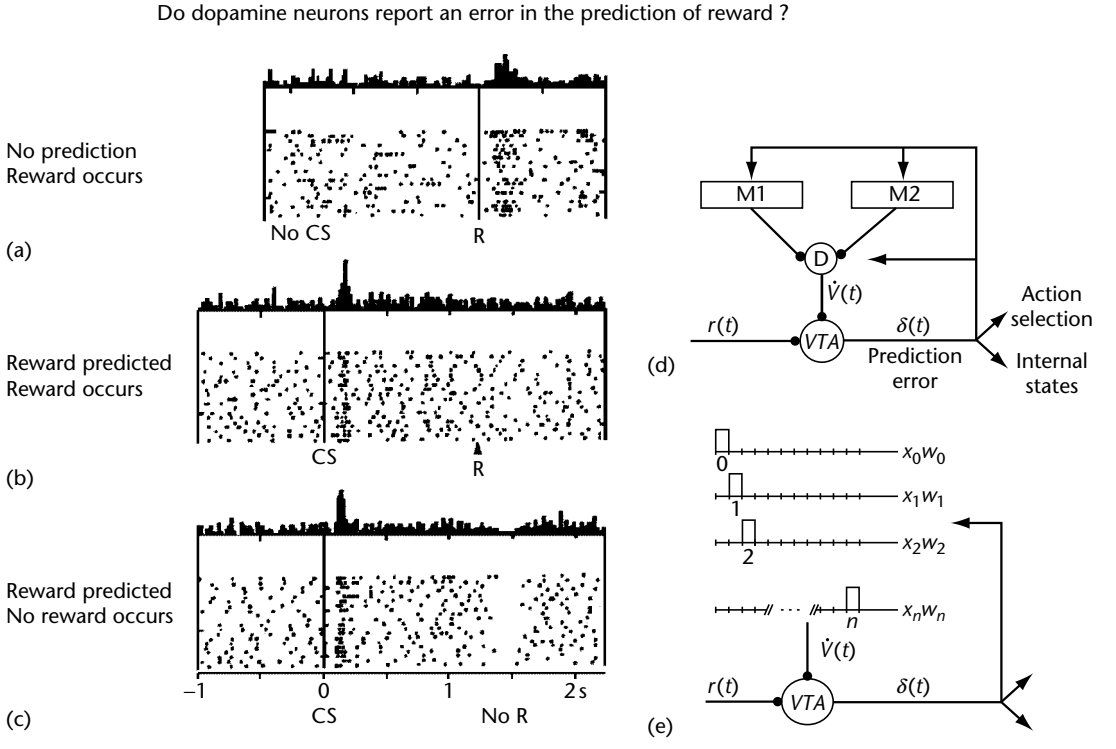


Figure 1. Left-hand panel. Dopaminergic predictor neurons during simple classical conditioning: plots of spike output of midbrain dopaminergic neurons recorded in alert primates during simple conditioning tasks. (a) Unpredicted delivery of reward (juice) causes a phasic increase in activity. This occurs for no conditioned stimulus (CS) (e.g. the light) or for a neutral CS (meaningless to a naive monkey). (b) After a number of trials involving a CS followed by juice delivery, the neurons stop responding to the presentation of the juice and instead give a phasic response just after the onset of the CS. (c) On error trials the cue is presented, but no reward is delivered. The neuron firing rate drops to 0 at the time when the reward would have been delivered if no mistake had been made. Right-hand panel. (d) Architecture of computational model mimics architecture of neuromodulatory systems in bees and humans. (e) Representation of a stimulus over time. There must be some representation of a stimulus over time in order to learn predictions. (Adapted from Schultz *et al*. (1997) *Science* **275**: 1593–1599.)

on the illumination of the light is also encoded in the fluctuations in dopaminergic activity.

From the results shown in Figure 1, it can be seen that dopamine neurons do not simply report the occurrence of rewarding events. Instead, their outputs appear to code for an error between the actual reward received and predictions of the time and magnitude of the reward. These neurons are only activated if the time of the reward is uncertain (i.e. unpredicted by any preceding cues). Dopamine neurons are therefore excellent feature detectors of the scalar 'goodness' of environmental events relative to learned predictions about those events. They emit a positive signal (increased spike production) if an appetitive event is better than predicted, no signal (no change in spike production) if an appetitive event occurs as predicted, and a negative signal (decreased spike production) if an appetitive event is worse than predicted. These observations have led to the hypothesis that these neurons are *predictor neurons* which distribute a prediction error signal (like $\delta(t)$ described above) to target neural structures in the form of changes in dopamine delivery.

Details of the way in which the dopamine signal is used neurally to implement full TD learning are beyond the scope of this article. The important point is that the electrical recordings from dopamine neurons show that they are part of a circuit in the brain that is capable of making predictions about future reward. This is one of the few instances where brain activity can be clearly connected to a well-understood computational learning procedure. Confirmation of the degree to which this description is incomplete or incorrect awaits future experimental and computational research. However, this work has provided a new way to interpret activity in dopamine neurons.

Dopamine neurons have long been known to be involved in reward processing. However, until recently it was thought that dopamine delivery was equivalent to reward. That is, the dopamine delivery itself was a positive reward signal. The experimental results described above, when combined with the computational interpretation in terms of TD learning, show that this cannot be the case. After training, there is no increased dopaminergic activity in response to the delivery of the rewarding fluid. The theory that dopamine is equivalent to reward is not consistent with this clear finding. However, an interpretation in terms of a prediction error is consistent, although this is not the end of the story. Dopaminergic activity is also seen in other contexts involving attention and orienting behavior. The connection to reinforcement learning has helped to frame the experimental issues in more sophisticated and formal terms.

One consequence of the TD model of dopaminergic function has recently been explored in human subjects using functional magnetic resonance imaging (fMRI).

PREDICTABILITY, REWARD AND HUMAN BRAIN RESPONSE

Reinforcement learning represents a general framework for describing how animals learn to interact with their environment. Therefore it is reasonable to extend this to the study of human reward pathways. Although these mechanisms are not thought to be fundamentally different in humans, the challenge lies in separating the core processes associated with reward prediction from secondary higher cognitive functions. For example, a temporal-difference model of reinforcement learning might predict that dopamine is released in proportion to prediction errors. While a monkey might learn a particular task by trial and error, humans are quite adept at using complex and abstract strategies that circumvent such laborious methods.

One relevant approach to studying humans is suggested directly by TD models. The core property that drives learning in TD models is predictability. Learning only occurs because there is some regular relationship between sensory cues and future reward. If there were no such relationship, then there would be nothing to learn. This can be used to advantage in humans. By simply altering the probability that one type of cue predicts a reward, one should be able to manipulate the response of dopaminergic systems in the brain. Furthermore, these manipulations can be subtle and therefore not consciously detectable by humans. These issues are important factors in experimental design if the aim is to prevent people from using conscious learning strategies.

It is worth noting that reward for a human may be different to reward for a monkey. We assume that human 'rewards' represent a lifetime of various types of conditioning that leads to the abstract assignment of reward to neutral objects (e.g. money). Understanding the way in which constructs such as money acquire reinforcing value are important but complex questions to address experimentally. However, it is worth first understanding how humans respond to the predictability of primary rewards, such as food and water.

In a simple experiment using fMRI, it has been shown that, just as in monkeys, the activity of dopaminergic projection sites in the human brain

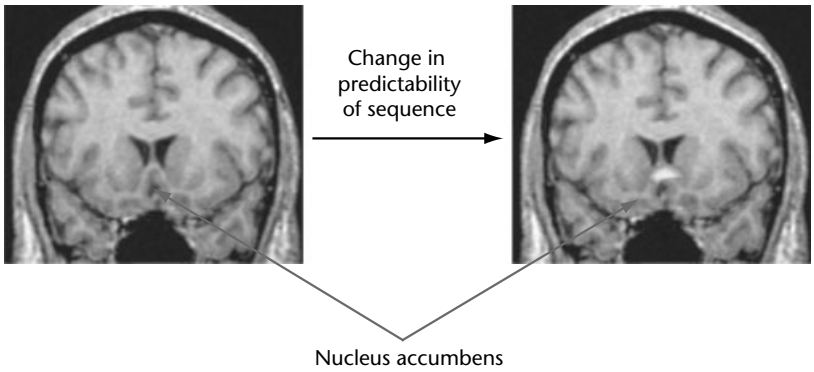


Figure 2. Change in predictability of a sequence of gustatory stimuli activates targets of dopamine projections. A sequence of squirts of fruit juice and water was delivered to the mouths of subjects. When the sequence of squirts changed from predictable to unpredictable, there was a larger activation in the nucleus accumbens, a region that is known to be involved in reward processing. (Adapted from Berns *et al*. (2001) *Journal of Neuroscience* **21**: 2793–2798.)

is modulated by the predictability of rewarding events. When subjects received squirts of fruit juice and water in their mouths, the activity in the nucleus accumbens was significantly amplified when the pattern of squirts was unpredictable (Figure 2). Since the participants were unaware of the differences in predictability, the fMRI activity changes cannot be ascribed to other top-down attentional processes.

SUMMARY

Reinforcement learning is one approach to the essential elements of learning problems that are encountered in real-world situations. There are numerous approaches to reinforcement learning, but here we have singled out temporal difference (TD) learning because it has growing connections to real biological data. We have reviewed one connection to the mammalian dopamine system and shown how a TD model accounts for changes in spike activity in dopaminergic neurons in alert monkeys during learning tasks. The same model provided an excellent starting point for the design of similar experiments in humans. One example of such an experiment using fMRI has been outlined, and the response of a target of dopamine projections, namely the nucleus accumbens, was shown to be activated by changes in the predictability of stimuli. This result was anticipated from the reinforcement learning model of the dopamine system, and it provides an excellent example of the way in which computational approaches to learning problems are informing the design and interpretation of experiments.

Further Reading

Bertsekas DP (1995) *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific.

Kaelbling LP, Littman ML and Moore AW (1996) Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* **4**: 237–285.

Montague PR, Dayan P and Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**: 1936–1947.

Schultz W, Dayan P and Montague PR (1997) A neural substrate of prediction and reward. *Science* **275**: 1593–1599.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Reorganization of the Brain

Intermediate article

Jon H Kaas, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

The complex normal organization of the brain in patterns of interconnected nuclei and areas emerges during brain development, but it can be altered or reorganized through life by sensory experience and damage to sensory and motor pathways. The reorganization process allows neurons to acquire new functional properties and roles in behavior.

INTRODUCTION

The mature mammalian brain is organized into systems of interacting nuclei and cortical areas. These nuclei and areas consist of neurons that are interconnected locally to form small computational groups. These local circuits then interconnect less strongly with other circuits within the same structure and to varying extents with circuits in other nuclei and areas to form processing systems. Systems then interconnect to produce a coherent, functioning brain. The complex normal organization of the brain emerges during development as a result of interactions between genetic instructions and information from the environment. Changes in experience or damage to part of a system will often alter the course of development to produce a different adult organization. The capacity for such organizational change is often referred to as 'developmental plasticity'.

Changes in the course of brain development typically occur during early sensitive or critical periods, after which the same damage or experiences fail to produce the same dramatic alterations. Because the developing brain is so much more responsive to injury or experience, many early investigators assumed that the organization of the mature brain is fixed and immutable. In addition, many felt that the stable performance of sensory, motor, and other systems must depend on a stable organization of the processing machinery. Of course we all realize that the brain must change when we learn, but often we assume that only certain learning centers in the brain are designed to be modified, and the rest of the brain is highly stable. Specialized learning systems do exist: in parts of the brain such as the hippocampus, neurons are capable of rapidly changing their functional connections during learning and experience. However, it is now known that other parts of the mature brain also are highly plastic. The internal organizations of nuclei and areas can be quite changeable. This has been best demonstrated in the sensory and motor systems, where normal organizations are well understood, and alterations are easily revealed by microelectrode recordings, electrical stimulation, and other procedures. These indicate that the basic sensory and motor processing structures in the brain are plastic throughout life, and that this plasticity is important for maintaining, changing, and improving brain functions.

Although evidence for brain plasticity largely stems from research on sensory and motor systems, conclusions based on this evidence may apply broadly, and all parts of the brain are likely to be susceptible to change.

REORGANIZATION IN SENSORY SYSTEMS AFTER DAMAGE

Sensory systems are processing hierarchies that start with a sensory surface of receptors, such as the rods and cones of the retina, the hair cells of the inner ear, and the mechanoreceptors of the skin. These sensory sheets send orderly projections to nuclei in the brainstem and thalamus, where they activate neurons in a pattern that reflects the spatial arrangement of the receptors in the receptor sheet. Thus, these target nuclei form topographic maps of the retina, cochlea, or skin. These nuclei in turn project in almost a point-to-point fashion to other

nuclei or cortical areas to create additional sensory maps, and this topographic relay of information continues through a number of brain structures. The early maps in the processing network most closely reflect the order of the receptor sheet, while higher-order maps become less topographic. The topographic orders of early maps in each system can be revealed by recording from neurons in many locations in the map, and determining what positions on the sensory surface stimuli activate neurons at each recording site. Because the normal organizations of these sensory maps can be revealed in detail by microelectrode recording methods, and normal maps prove to be highly constant in topographical organization from animal to animal, it is possible to determine whether these maps can be altered in internal organization by injury or experience.

Perhaps the most easily demonstrated changes in brain maps are those that follow partial sensory loss and deprivation. As one example, damage to a sensory nerve of inputs from the skin will remove a normal source of activation for neurons distributed across all maps of the skin in the somatosensory system. One might expect that the neurons in maps that normally respond to the removed inputs would no longer be activated by sensory stimuli. However, when this possibility was tested in somatosensory cortex of monkeys after section of the sensory nerve that contains the afferents from the thumb half of the glabrous hand, deprived neurons gradually became responsive to touch on other normally innervated parts of the hand (Florence *et al*., 1997).

In monkeys, primary somatosensory cortex consists of a long, narrow strip of tissue that extends mediolaterally across the central part of the cerebral hemisphere. The skin of the opposite side of the body is represented systematically from foot to face, in a medial to lateral sequence in this strip, with the hand representation lateral near that of the face. Digits one to five are represented in order in a lateromedial sequence. Section of the afferents from the glabrous surface of digits one to three deactivated about half of the map of the hand in primary cortex, but starting immediately and proceeding over a few weeks, the deprived cortical neurons became responsive to touch on the normally innervated dorsal surfaces of the digits and other parts of the hand. A similar reactivation of neurons occurs in the subcortical nuclei, the cuneate nucleus in the lower brainstem and the ventroposterior nucleus in the thalamus. Thus, the reorganization of the cortical map depends, at least in part, on the relay of reorganized patterns of responsiveness from subcortical structures. Primary somatosensory cortex relays a reorganized pattern to other areas of the somatosensory cortex, so that all areas in the system respond in a new manner. More extensive sensory losses can lead to even greater reorganization of sensory systems. Such changes are so large that they can be measured by noninvasive brain imaging techniques in humans. In individuals who had undergone amputation of an arm, the deprived portions of the map in primary cortex responded to touch on the stump of the limb or the face. Reorganization occurs in the thalamus as well, so that neurons formerly activated from the missing forelimb became responsive to the stump (Davis *et al*., 1998). Recordings from monkeys with major loss of sensory inputs indicate that reactivations of cortex can take as long as 6–8 months to complete (Florence *et al*., 1997).

Reorganization of sensory representations also occurs in auditory or visual nuclei and cortical areas after damage to the receptor sheet. After damage to the retina, the extent of reorganization appears to be much more extensive in the cortex than subcortically in the thalamus (Chino, 1997; Kaas *et al*., 2001). Thus, representations at higher levels in the visual system and perhaps in other systems may be more mutable than subcortical representations. Reorganizations also occur after damage to subcortical or cortical maps. For example, if part of the map of the hand in primary somatosensory cortex is removed by a lesion, adjacent parts of the map may change so that remaining neurons take on the role of those that were lost to different inputs from the hand, and the hand map reorganizes to recover some of the missing parts of the representation. Other cortical areas with inputs from primary cortex may reorganize even more extensively. If all of the hand representation is lesioned in primary somatosensory cortex, regions of higher-order maps that normally represent the hand may come to represent other parts of the body, such as the foot.

In summary, sensory maps respond to a loss of activating inputs, caused by damage to peripheral nerves or the central nervous system, by reorganizing so that deprived neurons become responsive to the remaining inputs. The recovery may be less extensive in subcortical structures and most extensive in higher-order cortical representations. Changes in maps can be rapid, but extensive reactivations may take months to emerge. These observations suggest the participation of a range of different mechanisms of recovery, and the possibility of several functional consequences.

REORGANIZATION IN MOTOR SYSTEMS AFTER DAMAGE

Reorganizations of the motor system follow damage to motor nerves, the loss of muscles for movement (such as after the loss of a limb), and damage to other parts of the motor system such as lesions of primary motor cortex (Nudo *et al*., 1997). The reorganization of primary motor cortex, M1, has been studied in rats after section of a motor nerve to the facial vibrissae, and in monkeys and humans who have suffered injuries requiring therapeutic limb amputations. In normal mammals, primary motor cortex systematically represents movements of the opposite side of the body. The internal organization of the representation can be demonstrated by stimulating locations throughout M1 with small electrical currents delivered with microelectrodes. In humans, the stimulations can be done harmlessly but somewhat less precisely through the skull with sudden, focused magnetic pulses that cause small, localized currents to flow in the brain (transcranial magnetic stimulation).

Rats normally devote a large part of M1 to moving the important sensory whiskers on the side of their face. As these whiskers move across objects, surface features can be detected by the rats. When the motor nerve to these whiskers is cut so the whiskers can no longer be moved, electrical stimulation of sites throughout the proportionally large whisker representation in M1 move the forepaw or eyelids instead. Similarly, in monkeys who have lost a forelimb, stimulating the large portion of M1 that normally moves digits and other parts of the forelimb moves instead the remaining stump of the limb (Wu and Kaas, 1999). Normally only a few sites are capable of evoking movements of the upper arm and shoulder. Comparable results have been obtained from deprived parts of M1 in humans with limb amputations. While some of the stimulation sites in reorganized cortex require higher than normal levels of current to be effective, many require only normal levels of current. Thus, primary motor cortex, when deprived of a normal muscle outlet, acquires control over the remaining adjacent muscles.

Motor cortex in primates includes not only the primary motor area, M1, but also the dorsal, ventral, supplementary, and cingulate premotor areas. Electrical stimulation in all of these premotor areas also evokes movement, but little is known about their capacity to reorganize after motor nerve damage or limb loss. However, limited results from monkeys suggest that the dorsal and ventral premotor areas do reorganize after forelimb loss, much as M1 reorganizes. The motor system also appears to reorganize after damage to motor cortex. There is considerable clinical evidence in humans for functional recovery after parts of motor cortex are damaged by stroke, but the nature of such reorganization has been difficult to determine. However, experiments in monkeys in which the normal organization of M1 is first determined with stimulating microelectrodes, followed by removal of a small portion of the region devoted to hand and digit movements, have found that the hand region of M1 reorganizes so that missing portions of the hand representation are recovered. This recovery seems to depend on training the monkey to use the impaired hand in a skillful manner. Thus, it appears that the reorganization of M1 after damage depends on use and practice, while the return of skilled movements depends on the cortical reorganization.

REORGANIZATION IN SENSORY SYSTEMS RESULTING FROM EXPERIENCE

Much of the research on the reorganization of sensory systems with experience has been on the somatosensory cortex of monkeys. When monkeys were given new sensory experiences, changes were observed in the organization of primary somatosensory cortex, S1 (Brodmann's area 3b). The observed changes were less extensive than those generally observed after damage to the somatosensory system. In one experiment, certain fingers received extensive amounts of sensory stimulation, and the representations of those fingers in S1 enlarged slightly. In other experiments, the simultaneous stimulation of several digits over long periods caused the detailed order of the representation of the fingers to become less precise, and neurons to acquire enlarged receptive fields.

The alterations in the map in somatosensory cortex appear to be predictable from the patterns of new stimulation. This is highly apparent with increased stimulation. Even electrical stimulation of receptors and afferents in a digit may increase the size of the cortical representation of that digit, and electrical stimulation of cortical sites increases the size of the representation of the body part that activates that site. Monkeys with long-standing abnormal use of a hand, such as after an induced motor disorder (dystonia) or hand injury, have abnormal maps of the hand in S1. In rats, the repeated stimulation of two adjacent whiskers on the face changed the responsiveness of S1 neurons to a second whisker. In humans, measures of patterns of brain activity

induced by stimulating fingers and other body parts have provided evidence for enlarged representations of the fingers in somatosensory cortex of Braille readers (Pascual-Leone and Torres, 1993) and skilled players of stringed instruments.

The effects of experience on somatosensory maps other than S1 are largely unknown, but presumably alterations take place in higher-order cortical maps as well, as changes in primary sensory areas would be relayed to higher-order representations. It is uncertain whether experience alters subcortical somatosensory representations.

There have been a number of studies of neurons in the primary auditory cortex of mammals (Weinberger, 1995). Neurons alter their responses to tones after a tone is paired with a foot shock so that they respond more to the paired tone and less to other tones. Shifts in the best frequency for groups of neurons alter the tonotopic organization of primary auditory cortex so that behaviorally significant tones have increased representations.

REORGANIZATION IN MOTOR SYSTEMS RESULTING FROM EXPERIENCE

It seems reasonable to suppose that as a skill is learned, motor cortex reorganizes to improve its mediation of that skill. This assumption has some experimental support, but individual differences in the detailed organization of motor cortex make it difficult to detect small changes. Nevertheless, there is evidence of use-dependent reorganization of the hand representation in M1 (Nudo *et al*., 1997). In humans, results from transcranial magnetic stimulation studies indicate that the representation in M1 of movements used in a highly repeated task are enlarged, and even short previous training on synchronous movements can produce a temporary change in the effects of stimulation on the motor map. However, functional magnetic imaging studies of cortical activation also indicate that complex motor tasks may actually cause less activation of motor areas of cortex when they are highly learned. This result suggests that fewer cortical neurons are needed for the same tasks after long-term practice. Thus, motor cortex sometimes reorganizes by reducing the number of circuits devoted to the task, possibly by increasing the effectiveness of those most critical for the task.

MECHANISMS OF REORGANIZATION

Reorganization of sensory and motor systems depends on two major mechanisms. First, the functional circuitry of these systems can be changed by increasing or decreasing the effectiveness of existing patterns of connections. There are many ways of altering the effectiveness of existing connections, and this type of change may be responsible for most of the plasticity that occurs in sensory and motor systems. Second, new connections may grow, even in the mature nervous system. Local growth and retraction of axons and dendrites may be common, but such limited growth would be difficult to detect by most anatomical procedures. Axons do extend considerable distances in the cortex and brainstem, but extensive new growth has been seen only under conditions of severe sensory deprivation with major reorganizations.

The responsiveness of neurons is normally under a number of excitatory and inhibitory influences. When these influences are in balance, the response properties of neurons are relatively stable. However, any change to the system would immediately alter the balance, and much of the immediate plasticity of brain systems is simply a reflection of rebalancing a dynamic system after inputs have been added or subtracted. For example, reduction in the excitatory drive of some inhibitory neurons causes their target neurons to become more excitable and respond to previously subthreshold inputs. To the extent that this creates new or larger receptive fields for these neurons, the details of sensory and motor representations are altered. Such changes are sometimes referred to as the 'unmasking' of 'silent synapses'. Altered responsiveness can also result from the release of neuromodulators during times of attention, distress, or motivation (Kilgard and Merzenich, 1998). Neurotrophic factors are also released to influence seasonal changes in neural circuits and behavior.

Perhaps most importantly, neurons are altered by activity patterns so that they become more responsive to frequent stimulus conditions and less responsive to other stimulus conditions. Experience and sensory events can lead to long-term potentiation (LTP) or long-term depression (LTD), which are long-lasting enhancements or reductions of synaptic effectiveness between neurons. The strengthening of synapses between neurons that are activated at the same time is known as Hebbian plasticity, after the speculations of the neuropsychologist Donald Hebb on the mechanisms of learning (Hebb, 1949). Hebbian plasticity appears to be based on LTP and LTD. The induction of LTP depends on the joint activation by excitatory inputs of both *N*-methyl-D-aspartate (NMDA) receptors and the non-NMDA glutamate receptors

(Rauschecker, 1991). The resulting depolarization of the activated neuron relieves a voltage-dependent block of the NMDA receptor, which allows calcium ions to flow through the receptor channel into the neurons to initiate cellular modifications that increase synaptic strengths. The cellular aspects of LTD are less well understood. Levels of neural activity also regulate the expression of neurotransmitters and transmitter receptors. Thus, a reduction of brain activity due to a nerve injury is followed by reduced expression by deprived neurons of the inhibitory neurotransmitter γ-aminobutyric acid (GABA) and the receptors for GABA (Jones, 1993). With less inhibition, the neurons become responsive to the remaining, previously ineffective connections.

Neurons also grow and form new connections. Increased activity may cause neurons to grow more dendrites and more dendritic spines, forming additional synapses to make circuits more effective. Axons may grow under a number of conditions. Groups of neurons that have become deactivated appear to cause growth in nearby neurons and attract new inputs. Such new growth, at least over longer distances, is normally inhibited, but the inhibition factors may be reduced or overcome by neurotrophic factors under conditions of deprivation. Finally, it appears that a small number of new neurons are generated, even in the mature brain. These new neurons may be important in some types of adult plasticity.

FUNCTIONAL CONSEQUENCES OF REORGANIZATION

While the reality of brain reorganization in mature mammals is well established, the functional consequences of these reorganizations are not so clear. The more limited reorganizations that occur with perceptual and skill learning are thought to be responsible for the improved performance (Pascual-Leone and Torres, 1993; Zahary *et al*., 1994). Changes in abilities have been shown to correlate with the reorganizations, but more direct evidence is lacking. Some progress might be made by evaluating the transfer of perceptual and motor skills from one part of the body to another: if these skills depend on reorganizations that are limited to sectors of a representation, they should not transfer when other body parts with unaltered representations are used. Some types of visual perception learning are highly specific for retinal location, suggesting that map reorganization with training does account for the learning. Individuals typically do recover to some extent from brain damage and sensory loss, and it is tempting to conclude that map reorganizations account for some or most of the recovery. Again, there is evidence that reorganization is at least sometimes correlated with recovery. For example, motor skills recover with motor map reorganization after partial lesions of motor cortex (Nudo *et al*., 1977). Further evidence on the role of reorganization in recovery is needed, however.

In some situations, reorganizations may mediate misperception and motor errors. Reorganizations probably account for elicited phantom sensations (Ramachandran and Blakeslee, 1998). Nearly all people with limb amputations have the sensation that the limb is still present, and this could be the result of spontaneous neural activity in the deprived portions of the somatosensory system. In addition, it is also possible to evoke sensations in the missing limb by touching other parts of the body. After forearm amputations in humans and monkeys, the deprived forearm portions of the somatosensory cortex and thalamus become responsive to touch on the limb stump and the side of the face. Touching the stump or face may be felt not only on the stump or face but also on the missing limb. It seems likely that these perceptions are the consequence of activating neurons in the stump or face portions of the central representations, as well as in the reactivated forearm portions. Although the forearm parts of the representation start to respond to new inputs, neural activity in these reactivated sectors continues to signal sensations to the missing limb. Such reorganizations may also be responsible for the perceptual filling in of scotomas after retinal lesions (Chino, 1997), tinnitus after damage to the auditory system (Muhlnickel *et al*., 1998), and focal dystonias after long, intense practice of motor skills (Byl *et al*., 1996).

In summary, map reorganizations may mediate desirable changes in abilities, and misperceptions and motor errors, depending to some extent on the magnitude of the reorganization. A goal of further research should be to understand how to foster reorganizations that mediate behavioral recoveries and new learning, while preventing and restricting those that mediate undesirable sensations and performances.

References

Byl N, Merzenich M and Jenkins W (1996) A primate genesis model of focal dystonia and repetitive strain injury: 1. Learning-induced de-differentiation of the representation of the hand in the primary somatosensory cortex in adult monkeys. *Annals of Neurology* **47**: 508–520.

Chino YM (1997) Receptive-field plasticity in the adult visual cortex: dynamic signal rerouting or experience-based plasticity. *Seminars in the Neurosciences* **9**: 34–46.

Davis KD, Kiss ZH, Luo L *et al*. (1998) Phantom sensations generated by thalamic microstimulation. *Nature* **391**: 385–387.

Florence SL, Jain N and Kaas JH (1997) Plasticity of somatosensory cortex in primates. *Seminars in the Neurosciences* **9**: 3–12.

Hebb DO (1949) *The Organization of Behavior*. New York, NY: John Wiley.

Jones EG (1993) GABAergic neurons and their role in cortical plasticity in primates. *Cerebral Cortex* **3**: 361–372.

Kaas JH, Collins CE and Chino Y (2001) Retinotopic maps in visual cortex reorganize after being locally deprived by damage to inputs. *Neuroscience News* **4**: 30–35.

Kilgard MP and Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* **279**: 1714–1718.

Muhlnickel W, Elbert T, Taub E and Flor H (1998) Reorganization of auditory cortex in tinnitus. *Proceedings of the National Academy of Sciences of the USA* **28**: 10340–10343.

Nudo RJ, Plantz EJ and Milliken GW (1997) Adaptive plasticity in primate motor cortex as a consequence of behavioral experiences and neuronal injury. *Seminars in the Neurosciences* **9**: 13–23.

Pascual-Leone A and Torres F (1993) Plasticity of the sensorimotor cortex representation of the reading finger in Braille readers. *Brain* **116**: 39–52.

Ramachandran VS and Blakeslee S (1998) *Phantoms in the Brain*. New York, NY: William Morrow.

Rauschecker JP (1991) Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. *Physiological Reviews* **71**: 587–615.

Weinberger NM (1995) Dynamic regulation of receptive fields and maps in the adult sensory cortex. *Annual Review of Neuroscience* **18**: 129–158.

Wu CWH and Kaas JH (1999) The organization of motor cortex of squirrel monkeys with longstanding therapeutic amputations. *Journal of Neuroscience* **19**: 7679–7697.

Zahary E, Celebrini S, Britten KH and Newsome WT (1994) Neuronal plasticity that underlies improvement in perceptual performance. *Science* **263**: 1289–1292.

Further Reading

Buonomano DN and Merzenich MM (1998) Cortical plasticity: from synapses to maps. *Annual Review of Neuroscience* **21**: 149–186.

Jones EA (2000) Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review of Neuroscience* **23**: 1–37.

Kaas JH (ed.) (1997) Functional plasticity in adult cortex. *Seminars in Neuroscience* **9**: 1–67.

Kaas JH (2001) *The Mutable Brain*. Amsterdam, Netherlands: Harwood.

Reticular Activating System

Introductory article

Martin Sarter, Ohio State University, Columbus, Ohio, USA
John P Bruno, Ohio State University, Columbus, Ohio, USA
Gary G Berntson, Ohio State University, Columbus, Ohio, USA

CONTENTS

The reticular activating system consists mostly of ascending noradrenergic and cholinergic projections originating in the brainstem. These projections enter the cortex, thalamus and basal forebrain, and mediate increases in wakefulness and arousal. Contemporary hypotheses describe these systems as actively regulated afferent components of forebrain circuits mediating defined cognitive processes.

INTRODUCTION

Beginning with the experiments carried out by Bremer and by Moruzzi and Magoun in the 1930s and 1940s, findings in support of the view that the brainstem contains a globally activating neuronal system gave birth to the concept of a homogeneously organized ascending reticular activating system. Interestingly, the term 'activation' originally referred to the increase in the readiness of forebrain neuronal networks to process other afferent information, and the generalization of this term to various forms of behavioral 'arousal' has suffered from insufficient conceptual definition and empirical support, to the extent that the scientific value of the construct 'arousal' has been questioned.

Traditional descriptions of the anatomy of the brainstem ascending systems focused on the 'reticular' character of these projections, a term that refers to the absence of clear anatomical segregation of the various cell groups in the brainstem giving rise to long ascending projections to the forebrain, and to the extensive branching or collateralization of these ascending projections. Modern anatomical techniques revealed a more precise anatomical organization of these projections and their terminal fields in the forebrain. Contemporary research designed to determine the functions of the different branches of the ascending activating system in interaction with their forebrain target circuits has yielded specific hypotheses about the dissociable functions of selected branches of this system in different behavioral and cognitive conditions. Two branches of the reticular activating system are most closely documented to mediate forebrain activating effects: the ascending noradrenergic system and the ascending cholinergic system. The ascending noradrenergic system is conceptualized as a neuronal system promoting cognitive processing in forebrain circuits particularly in response to novel, affective or stress-like stimuli and their associated changes in autonomic reactivity. Increases in activity in the brainstem ascending cholinergic system primarily promote dreaming cognition during rapid eye movement (REM) sleep via activation of mostly thalamic and basal forebrain corticopetal projections.

ANATOMY OF ASCENDING NORADRENERGIC AND CHOLINERGIC PROJECTIONS

Noradrenergic Projections

Although several noradrenergic cell groups in the dorsomedial and caudal ventrolateral medulla send their axons to the forebrain, the ascending projections of the locus caeruleus (LC) most extensively have been demonstrated to mediate 'arousal', awakening or increased alertness. The LC is situated close to the fourth ventricle in the pons, and in humans contains about 13 000 neurons per hemisphere; these project to all major regions in the mesencephalon, diencephalon and telencephalon. Projections from the LC to the forebrain are highly collateralized, with single neurons innervating widespread telencephalic areas. Although anatomical studies have qualified traditional statements about the 'diffuse' organization of noradrenaline (NA) inputs to the cortex by

demonstrating species-specific differences in the regional and laminar density of NA inputs, NA is released predominantly from nonjunctional varicosities, supporting the hypothesis that NA acts in a widespread, diffuse and modulating fashion, thereby globally gating the information processing in its target sites. Among the subcortical target sites that are important in cognitive functions (below) are the cholinergic neurons of the basal forebrain. The cholinergic basal forebrain, consisting primarily of the nucleus basalis of Meynert, the substantia innominata, the horizontal limb of the diagonal band of Broca and the preoptic area, has been conceptualized as the most rostral extension of the reticular activating system. Basal forebrain cholinergic neurons are predominantly depolarized via α_1-adrenergic receptors, driving cholinergic cells into a tonic mode of firing and increasing their rate of repetitive spike discharge. The widespread innervation of the cortex by cholinergic projections from the basal forebrain supports this notion, and NA inputs to the basal forebrain are organized to increase the responsivity of this structure's neurons to other (telencephalic) inputs.

The LC receives inputs primarily from two areas in the rostral medulla, the nucleus paragigantocellularis (nPG) and the nucleus prepositus hypoglossi (nPH). The nPG projects to sympathetic preganglionic neurons, mediating autonomic activation which, via the parallel projection to the LC, also activates the ascending NA system. The role of the inhibitory input from the nPH is less clear. Functionally, the afferent organization of the LC suggests that this system mediates the activational (specifically cognitive) consequences of stimuli that increase sympathetic reactivity. The finding that activation of the prefrontal cortex stimulates LC neurons furthers the hypothesis that the ascending NA system is wholly integrated in the neuronal systems mediating cognitive functions.

Cholinergic Projections

The cholinergic cell bodies in the brainstem which mediate forebrain activation, particularly during REM sleep, originate mainly from the pedunculopontine tegmental (PPT) and laterodorsal tegmental (LDT) region. The ability of these tegmental cholinergic neurons to induce cortical desynchronization has traditionally been assumed to depend on connections via the thalamus, mostly via dorsal, intralaminar, reticular thalamic nuclei. However, the PPT and LDT also project to the basal forebrain, and it appears that the increased cortical acetylcholine efflux during REM sleep is due mainly to stimulation of corticopetal cholinergic neurons in the basal forebrain by these ascending cholinergic projections. Thus, lesions of the PPT in animals resulted in a decreased cholinergic innervation of the basal forebrain, and electrical stimulation of the PPT appears to be sufficient to produce an increase in cortical acetylcholine release via projections terminating in the basal forebrain. Although the exact distribution of cholinergic terminals in the basal forebrain remains unsettled and may include direct contacts of cholinergic terminals with this structure's corticopetal cholinergic neurons, it seems more likely that the majority of cholinergic inputs do not synapse directly onto basal forebrain cholinergic neurons. It is not clear whether other PPT neurons, such as glutamatergic neurons, project to the basal forebrain.

EVOLUTION OF THE CONCEPT OF THE RETICULAR ACTIVATING SYSTEM

In addition to early anatomical work on the reticular formation and its relay to widespread cortical areas via 'nonspecific' thalamic nuclei, functional studies contributed to the construct of a generalized ascending reticular activating system. Stimulation of the reticular formation was reported to induce cortical activation and behavioral arousal, for example, whereas reticular lesions could yield cortical deactivation and coma. This was in keeping with the emerging concept of 'general arousal' in psychology, and these perspectives merged into the construct of a generalized arousal system.

The unitary anatomical-functional concept of the 'arousing' functions of the reticular activating system did not fare well empirically. More refined anatomical studies continued to reveal greater complexity and specificity in ascending activating systems, and the construct of 'generalized' arousal became increasingly untenable. For example, various psychophysiological (heart rate, electrodermal responses, cortical electroencephalography) and behavioral (e.g. reaction time) indices of arousal or activation are not well correlated, and the functional manifestations of manipulations of arousal do not correspond with descriptions that assume a linear continuum from low to high arousal. Rather, modern approaches have returned to the original meaning of 'activation', conceptualizing the functions of the ascending systems in terms of fostering or modulating the cognitive processes mediated through neuronal target circuits in the forebrain. Thus, the activating effects of the noradrenergic or cholinergic ascending systems may be limited to interactions with other afferent projections of their

respective target neurons in the forebrain. Such hypotheses include the view that the activating effects of ascending systems do not represent a mere passive, 'bottom up' contribution to forebrain functions. Rather, these ascending systems appear to be regulated by interactions among the stimulus-situation, task-demands, and 'top down' cognitive influences. In short, ascending systems represent integrated components of the forebrain neuronal systems mediating complex behavioral and cognitive functions.

ATTENTIONAL BIAS IN FEAR AND ANXIETY

The locus caeruleus, based on its anatomical connections described above, imports information about the sympathetic activation produced by novel, emotionally charged or stress-like stimuli to the forebrain. Functionally, LC-mediated increases in the responsivity of forebrain targets are thought to bias attentional information processing toward such stimuli and contexts. Thus, attentional processing in response to urgent (particularly aversive) stimuli is optimized by increased activity in noradrenergic ascending projections. A persistent attentional bias towards fear- and anxiety-related stimuli in people with anxiety disorders may be mediated in part by an overactive noradrenergic ascending system.

Although it is likely that a wide range of diencephalic and telencephalic (including cortical) target systems of the LC-NA system mediate such attentional biasing, the basal forebrain represents a prime target for this function. Corticopetal cholinergic projections from the basal forebrain are necessary for the mediation of attentional functions ranging from the detection, selection and discrimination of relevant stimuli to the allocation of processing resources to competing attentional tasks. Noradrenergic modulation of basal forebrain neurons may be sufficient to permit telencephalic (glutamatergic) afferents to stimulate basal forebrain. Collectively, noradrenergic inputs to the basal forebrain are hypothesized to initiate attentional processing via telencephalic inputs and the basal forebrain's corticopetal projection systems (Figure 1).

The crucial role of the corticopetal cholinergic system of the basal forebrain in the processing of fear- and anxiety-related information and in the associated increases in autonomic reactivity has been demonstrated in experiments involving cognitive processing (as opposed to the effects of unconditioned stimuli and classically conditioned stimuli for aversive events). Enhanced autonomic responses in such contexts were shown to depend on the integrity of basal forebrain cholinergic neurons, specifically on cholinergic projections to the medial prefrontal cortex.

The reasons for the affective nature of the attentional processing induced via noradrenergic activation may be at least twofold. First, activation of the LC itself reflects sympathetic activation and thus may signify the presence of affectively significant, novel or stressful stimuli. Second, the basal forebrain receives input from several sites involved in the descending processing of fear- and anxiety-related information – notably the amygdala. Interconnections between the basal forebrain, the amygdala and the prefrontal cortex represent a parsimonious neuronal system through which the processing in basal forebrain circuits is predominately enhanced in fear- and anxiety contexts (Figure 1). Additionally, descending activation of the LC by the prefrontal cortex may further foster the attentional processing of fear- and anxiety-related stimuli through a 'top down' influence.

The LC-mediated influence of sympathetic activity on forebrain cognitive processing requires additional comment. Although the assumption that peripheral variables can affect high-level processing may appear speculative, in fact such influences have been demonstrated for central regulation of autonomic, immune, and endocrine as well as behavioral processes. Indeed, sympathetic influences on the ascending LC system probably represent only one of a range of interactions among upper and lower levels of the neuroaxis in the generation of affective reactions and their cognitive components. In this regard, the LC represents a major ascending pathway contributing to the integration of lower-level information in higher-level cognitive processes, and experimental investigations into its role may eventually yield a scientifically useful version of William James's hypothesis about the role of visceral afferent feedback in affective reactions. (*See* **Autonomic Nervous System**)

CONCLUSION

The monolithic concept of a reticular activating system is rapidly being replaced by hypotheses concerning the precise functions of defined groups of neurons in the brainstem, their ascending (and descending) projections, and their interactions with other brainstem nuclei. Importantly, this development is not just fueled by the ever more detailed anatomical description of the efferent and afferent organization of brainstem neurons and their

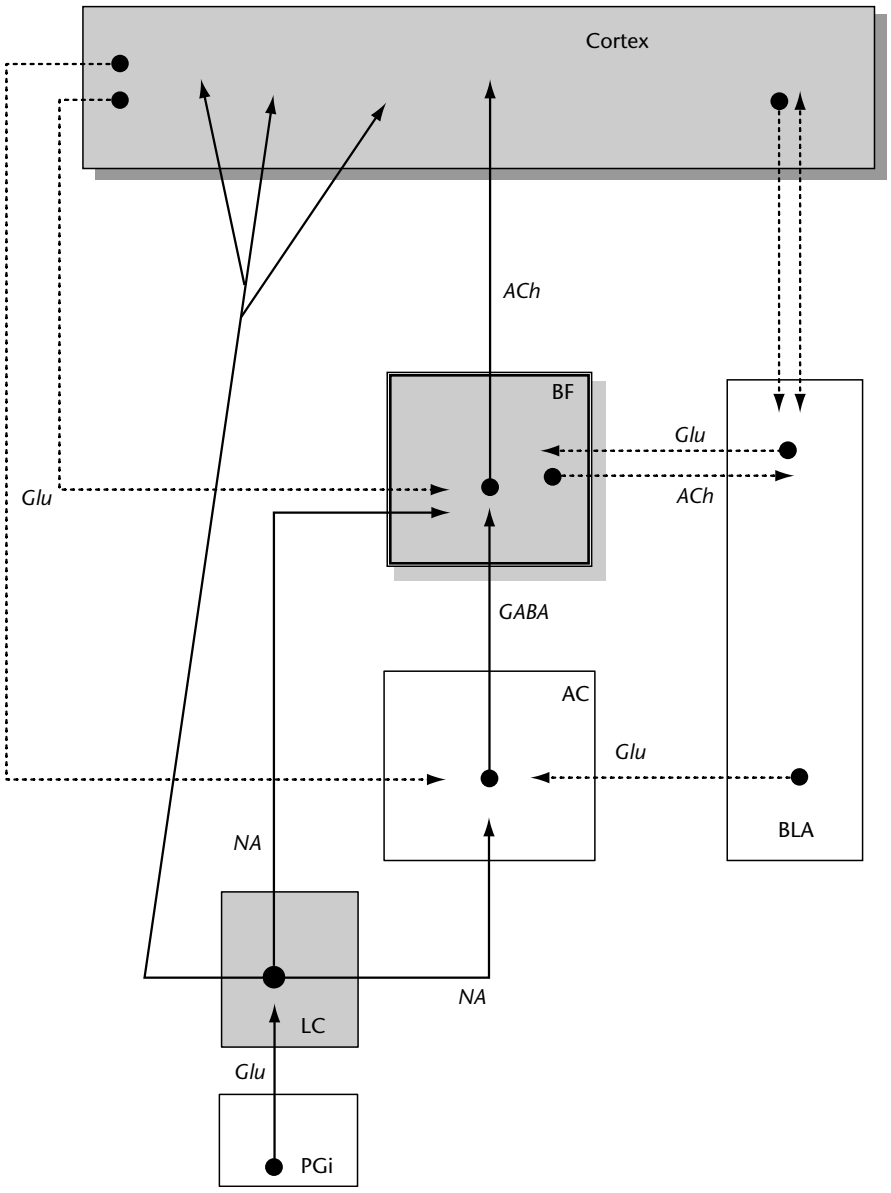


Figure 1. Role of the noradrenergic ascending projections to the cortex and basal forebrain. Novel, salient or emotionally charged stimuli activate ascending noradrenergic systems, possibly via afferents of the locus caeruleus originating in the nucleus paragigantocellularis, a sympathoexcitatory structure. Noradrenergic stimulation of neurons functionally recruits telencephalic circuits in the basal forebrain, consisting mostly of cortical, amygdaloid and accumbal inputs to corticopetal projections, for the processing of attentional information. AC, nucleus accumbens; ACh, acetylcholine; BF, basal forebrain; BLA, basolateral amygdala; GABA, γ-aminobutyric acid; Glu, glutamate; LC, locus caeruleus; NA, noradrenaline (norepinephrine); PGi, nucleus paragigantocellularis.

physiological analysis, but by hypotheses about the active regulation of the activity of ascending systems in different behavioral and cognitive contexts, and about the significance of activity in these systems for the cognitive processes mediated through their forebrain neuronal target circuits. The traditional, broad concepts of a 'reticular activating system' and 'arousal' cannot capture the significance of such hypotheses and thus should be replaced.

Further Reading

Aston-Jones G, Rajkowski J, Kubiak P, Valentino RJ and Shipley MT (1996) Role of the locus coeruleus in emotional activation. *Progress in Brain Research* **107**: 379–402.

Baghdoyan HA (1997) Cholinergic mechanisms regulating REM sleep. In: Schwartz WJ (ed.) *Sleep Science: Integrating Basic Research and Clinical Practice*, pp. 88–116. Basel: Karger.

Berntson GG, Sarter M and Cacioppo JT (1998) Anxiety and cardiovascular reactivity: the basal forebrain cholinergic link. *Behavioural Brain Research* **94**: 225–248.

Berntson GG, Cacioppo JT and Sarter M (2002) Bottom-up: implications for neurobehavioral models of anxiety and autonomic regulation. In: Davidson RJ, Goldsmith HH and Scherer KR (eds) *Handbook of Affective Science*. Oxford, UK: Oxford University Press.

Jones BE (1991) Paradoxical sleep and its chemical/structural substrates in the brain. *Neuroscience* **40**: 637–656.

Robbins TW and Everitt BJ (1995) Arousal systems and attention. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 703–720. Cambridge, MA: MIT Press.

Sarter M and Bruno JP (2000) Cortical cholinergic inputs mediating arousal, attentional processing and dreaming: differential afferent regulation of the basal forebrain by telencephalic and brainstem afferents. *Neuroscience* **95**: 933–952.

Steriade M and Buzsaki G (1990) Parallel activation of thalamic and cortical neurons by brainstem and basal forebrain cholinergic systems. In: Steriade M and Biesold D (eds) *Brain Cholinergic Systems*, pp. 3–60. Oxford, UK: Oxford University Press.

Steriade M and Llinas RR (1988) The functional states of the thalamus and the associated neuronal interplay. *Physiological Reviews* **68**: 649–742.

Szymusiak R (1995) Magnocellular nuclei of the basal forebrain: substrates of sleep and arousal regulation. *Sleep* **18**: 478–500.

Zaborszky L (1992) Synaptic organization of basal forebrain cholinergic projection neurons. In: Levin ED, Decker M and Butcher LL (eds) *Neurotransmitter Interactions and Cognitive Function*, pp. 329–354. Boston: Birkhauser.

Reward, Brain Mechanisms of

Introductory article

Roy A Wise, National Institute on Drug Abuse, Baltimore, Maryland, USA

CONTENTS

Direct electrical or chemical stimulation of the brain can be powerfully rewarding. The neurons activated by such rewards are partially known and are the focus of intense investigation. These neurons are thought to participate in normal motivation and in the control of behavior through the natural pleasures of life.

INTRODUCTION

Rewards are stimuli or events that increase the probability of the behavioral acts that they regularly follow. Specialists attempting to identify this class of events objectively have used the terms 'reinforcement' and 'incentive motivation' to characterize the effects of such stimuli; the approximations 'liking' and 'wanting' have captured recent attention in attempts to characterize these two aspects of reward. 'Reinforcement' refers to the effects of the stimulus on the brain circuitry that generated the act being reinforced; it is presumed that the neural elements of this circuitry remain in a plastic state for some time after the act, and that the reinforcing stimulus or event has the ability to stamp in or reinforce the synaptic connections between these elements. Two types of reinforcement have been suggested, one in which the associations between active stimulus elements are strengthened, and one in which the associations between stimulus and response elements are strengthened. Thus far, cellular correlates of only the first type – Pavlovian reinforcement – have been identified.

Pavlov made the important distinction between two types of stimuli: signal stimuli, which elicited reflex responses, and nonsignal stimuli, which did not. The distinction is not a clean one, because even so-called 'neutral' or nonsignal stimuli elicit a reflexive orienting response – a response consisting of eye, head, and whole body movements toward the stimulus, along with autonomic responses such as papillary dilation and changes in galvanic skin response. However, these responses to a nonsignal stimulus disappear (habituate) as the stimulus becomes familiar and predictable. A nonsignal stimulus can become a signal stimulus if it is associated in time with a signal stimulus. A well-known example is the nonsignal bell that becomes a conditioned (signal) stimulus when associated with the unconditioned (signal) stimulus of food. A conditioned stimulus can again become a nonsignal stimulus if it is not occasionally reassociated with its unconditioned stimulus. Pavlov gave the name 'reinforcement' to the renewal of the signal status of the conditioned stimulus by presentation of the conditioned and unconditioned stimulus together. The learning that Pavlov studied can be termed 'stimulus' learning, in that it deals with the learned association of one stimulus with another.

Thorndike used the phrase 'stamping-in' to suggest a similar strengthening of the association between a stimulus and a response. Thorndike came to use Pavlov's term – used even earlier by William James – to refer to the stamping-in associated with response learning. Skinner, in an attempt to better characterize response learning, also used the term 'reinforcement', defining a reinforcer as a stimulus that increases the probability of the acts that regularly precede it. In popular terminology, the term 'reward' is typically used in relation to this form of learning. However, it is clear that stimulus reinforcement has an important, perhaps critical, role in response learning.

Rewards not only stamp in stimulus and response associations; they also energize behavior and increase its probability proactively. The taste of a rewarding stimulus often focuses subsequent attention on stimuli previously associated with that reward and increases the probability of behaviors that led to that reward in the past. Familiar examples are the eating of one salted peanut or potato chip. The taste of one energizes searching

behavior that is satisfied only with more. The motivational arousal ('wanting') caused by initial exposure to a reward or to stimuli associated with a reward is termed 'incentive motivation'. Like reinforcement, it is one of the fundamental effects of rewards.

The familiar rewards are the everyday pleasures of life: food when we are hungry, water when we are thirsty, heat when we are chilled, and cold when we are overheated. The ability of these to establish strong response habits is well known; their ability to serve as incentive motivational stimuli is less widely articulated, but is also well known. Consider, for example, sexual arousal. Sexual arousal is not constant; it is strongly modulated by hormonal states in the female and by sensory stimuli in the male. In lower animals, it is strongly stimulated by the smell of a receptive female; in the human male it can be strongly stimulated by the sight of genital or perigenital flesh or, in some cultures, by the sight of a shapely breast. The sexual arousal activated by olfactory or visual stimuli illustrates powerful cases of incentive motivation.

BRAIN STIMULATION REWARD

Our understanding of the brain mechanisms of reinforcement and incentive motivation has been stimulated and advanced by studies in which deep structures in the brain were stimulated electrically in freely moving animals. Stimulation of the depths of the brain – the lateral hypothalamus and adjacent structures through which course common neuronal fibers (axons from motivationally important cells of unknown origin and termination) – can cause a sated animal to seek out and eat food, or to respond appropriately to a variety of other natural incentives such as prey or a sexual partner. Stimulation in the same brain sites – and, indeed, many more – is strongly reinforcing. The fibers passing near the sites where such stimulation has motivational effects are presumed to form a critical portion of the neural circuitry of reward.

The first experiments on the rewarding effects of direct brain stimulation were carried out by James Olds and Peter Milner, who noticed that rats would quickly return to a place in their environment where they had received the stimulation. That place had become, through association with the stimulation, an incentive stimulus, one that was approached in preference to others. Olds and Milner promptly noted the ability of this stimulation to establish lever-pressing habits. This behavior is termed 'intracranial self-stimulation'. Hungry rats will work for such stimulation in preference to food, and will, if access to food and stimulation is restricted to a brief period each day, starve as a consequence of their preference for the stimulation. Such studies have established the medial forebrain bundle as a locus of brain reward circuitry. Unfortunately, the medial forebrain bundle comprises fifty or sixty sets of nerve fibers, and it remains unclear which of these plays an important role in reward function. Pharmacological studies have, however, given some initial clues.

A ROLE FOR THE MESOLIMBIC DOPAMINE SYSTEM

The medial forebrain bundle fibers most clearly associated with reward function are the ascending fibers of the mesolimbic dopamine system which project from the ventral tegmental area of the midbrain to several forebrain structures including the nucleus accumbens. While lever-pressing habits can be disrupted by a variety of pharmacological agents (sedatives, noradrenergic antagonists, and, indeed, strong doses of drugs that are themselves rewarding), the drugs that affect self-stimulation in ways that parallel the effects of reducing the reward itself are antagonists of dopamine neurotransmission. Such drugs do not attenuate the lever-pressing of the first few seconds of a test session, where the animal has not yet learnt the consequences of the drug state, but they cause a gradual reduction of the behavior as if the animal finds the usual reward less than normally satisfying. Such drugs do not reduce the maximum levels of work that are sustained by the reward in question, but they increase the amount of reward that is required to motivate maximal response levels. Indeed, it is possible to determine the number of extra pulses of brain stimulation that is necessary to compensate for a given dose of a dopamine antagonist. Moreover, dopamine *agonists* have the opposite effect, reducing the number of brain pulses necessary to maintain a given rate of responding, and a dopamine agonist that reduces the number of required pulses by 10% will just offset the effects of a dopamine antagonist that increases the number of required pulses by 10%.

FIRST-STAGE FIBERS

There are perhaps fifty fiber pathways of the medial forebrain bundle that are directly activated by rewarding stimulation; it is unlikely that more than a few of these play a role in reward function. These directly activated neurons are termed

'first-stage' neurons, and they have not yet been identified. Pulse-pair studies have helped narrow the list of possibilities. Pulse-pair studies take advantage of the facts that (a) nerve fibers need time to recharge after each nerve impulse, and (b) different sizes and types of nerve fibers need different amounts of time for such recharging. Large, myelinated fibers recharge much more quickly than small, unmyelinated ones. By administering rewarding stimulation in trains of pairs of closely spaced pulses rather than in trains of equally spaced pulses, electrophysiologists have been able to determine the range of refractory periods of the populations of fibers at the electrode tip that participate in brain stimulation reward function. The procedure is to vary the interval between the pulses in each pair, identifying the amount of time needed for the second pulse in each pair to add reward value to that seen from the first pulse alone, and identifying the amount of time needed for full recovery such that the second pulse in each pair adds as much reward as does the first. This gives the range of refractory periods for the population of neurons contributing to the rewarding effects of the stimulation. The range of refractory periods is quite short (0.4–2.0 ms), suggesting that the reward-relevant fiber populations are large and myelinated. This rules out the dopamine-containing fibers as first-stage fibers, because they are small and unmyelinated. Which of several sets of large myelinated fibers contribute to the first-stage mechanism remains to be determined.

Further evidence that the dopaminergic fibers of the medial forebrain bundle do not serve as the first-stage transduction fibers of brain stimulation reward circuitry comes from pulse-pair studies involving two electrodes positioned along the length of the bundle. When the first pulse of a pulse pair is delivered at one electrode and the second is delivered some distance along the pathway, the time that must be given before the second pulse will be completely effective represents not only the time for the fiber to recover from the action potential triggered by the first pulse, but also the time it takes that action potential to reach the second electrode site. By subtraction one can estimate the conduction velocity of the neurons that contribute to the rewarding effect of the stimulation. The conduction velocities are fast – of the order of 2.5 meters per second – again implicating the large, myelinated fibers.

The conduction velocity experiment does not depend on which end of the first-stage neurons receives the first pulse of each pulse pair. However, if anodal current, which hyperpolarizes rather than depolarizes neurons, is given at one of the stimulation sites, and cathodal, depolarizing current is given at the other, the direction of conduction of the reward signal can be determined. If anodal stimulation at the caudal electrode blocks the rewarding effect of cathodal stimulation at the rostral electrode, and if anodal stimulation at the rostral electrode does not block the rewarding effect of cathodal stimulation at the caudal electrode, then the reward signal is conducted in the rostral–caudal direction. This test has shown that the bulk of the first-stage elements of medial forebrain bundle brain stimulation reward conduct in the rostral–caudal direction, opposite to the normal direction of conduction of the mesolimbic dopamine neurons.

Thus the critical role of the dopamine system in the rewarding effects of medial forebrain bundle stimulation must result from transsynaptic activation of the mesolimbic dopamine system by the large, myelinated, caudally projecting first-stage reward neurons or their efferents. The first-stage fibers project very close to the dopaminergic cell bodies of the ventral tegmental area, but the currently favored hypothesis is that they project beyond the ventral tegmental area to synapse on the cholinergic neurons of the pedunculopontine tegmental nucleus, which, in turn, projects back to the ventral tegmental dopamine cells.

CHEMICAL STIMULATION OF REWARD CIRCUITRY

The pharmacological studies implicating dopamine systems in brain stimulation reward suggest that pharmacological activation of the dopamine system should, like stimulation-induced activation, be rewarding in its own right. There is now considerable evidence to support this suggestion. Rats will learn to lever-press for direct injections of dopamine or direct dopamine agonists into the nucleus accumbens. Cocaine and amphetamine are indirect dopamine agonists – drugs that elevate synaptic levels of endogenous dopamine – and they too are self-administered into this nucleus. Morphine, heroin, nicotine, alcohol, and cannabis also elevate nucleus accumbens dopamine levels: nicotine by stimulating the dopamine neurons directly, morphine and heroin by disinhibiting the dopamine neurons, and alcohol, and cannabis by mechanisms that have not yet been elucidated. Morphine and other μ-opioids are self-administered directly into the region of dopamine cell bodies in the ventral tegmental area, and their potency for such self-administration is the same as their potency for elevating nucleus accumbens dopamine levels.

Not all drugs of abuse act at the level of the dopaminergic link in brain reward circuitry, however. It is not clear where barbiturates or benzodiazepines trigger their rewarding actions, but it seems likely that they do so in one or more synapse of the output neurons of nucleus accumbens. The output neurons of nucleus accumbens are medium-sized spiny neurons that express γ-aminobutyric acid (GABA) as their neurotransmitter, and barbiturates and benzodiazepines act at GABA receptors. The medium spiny neurons appear to be a site at which phencyclidine exerts its habit-forming actions; phencyclidine blocks excitatory amino acid receptors of the *N*-methyl-D-aspartate (NMDA) subtype. These receptors are localized on medium spiny neurons of the nucleus accumbens, and phencyclidine and other NMDA antagonists are self-administered into this region. While phencyclidine can alter local dopamine levels – like cocaine, it is a dopamine uptake inhibitor and can elevate synaptic dopamine levels – it blocks NMDA receptors at a concentration an order of magnitude lower than is needed to block dopamine uptake. Moreover, dopamine blockers do not block the rewarding effects of phencyclidine, and NMDA blockers that do not affect synaptic dopamine levels are as effective as phencyclidine in reward function.

Drugs of abuse thus implicate several neurotransmitters in brain reward circuitry. Amphetamine and cocaine implicate dopamine: their reward-relevant receptor in the brain is the dopamine transporter and their reward-relevant action is the blockade of that transporter and thus the blockade of clearance of neuronally released dopamine. Nicotine implicates cholinergic circuitry: acetylcholine is the endogenous transmitter for the receptors at which nicotine triggers its habit-forming action. Opiates implicate their own endogenous counterparts, endomorphin-1, β-endorphin, and enkephalin. They also implicate GABA, acting to disinhibit the dopamine system by inhibiting a GABA system. Phencyclidine implicates glutamate, which is the primary endogenous transmitter for the NMDA receptor that phencyclidine blocks for its habit-forming actions. Other transmitter substances are also likely to be involved, as the classic transmitters – acetylcholine, glutamate, GABA, and dopamine – are each colocalized with peptide cotransmitters such as cholecystokinin, substance P neurotensin, and enkephalin.

KNOWN ELEMENTS IN REWARD CIRCUITRY

Diagrams of brain reward circuitry (Figure 1) invariably center on the mesolimbic dopamine system. The cell bodies of the mesolimbic system receive reward-associated cholinergic inputs from the pedunculopontine and dorsomedial tegmental nuclei, glutamatergic inputs from the frontal cortex and amygdala, and GABA-mediated inputs from the nucleus accumbens, ventral pallidum, and substantia nigra. The terminals of the mesolimbic system synapse on medium spiny neurons of nucleus accumbens. A nondopaminergic mesolimbic projection involves GABA-mediated ventral tegmental neurons that project to cholinergic interneurons of the nucleus accumbens, which, in turn,

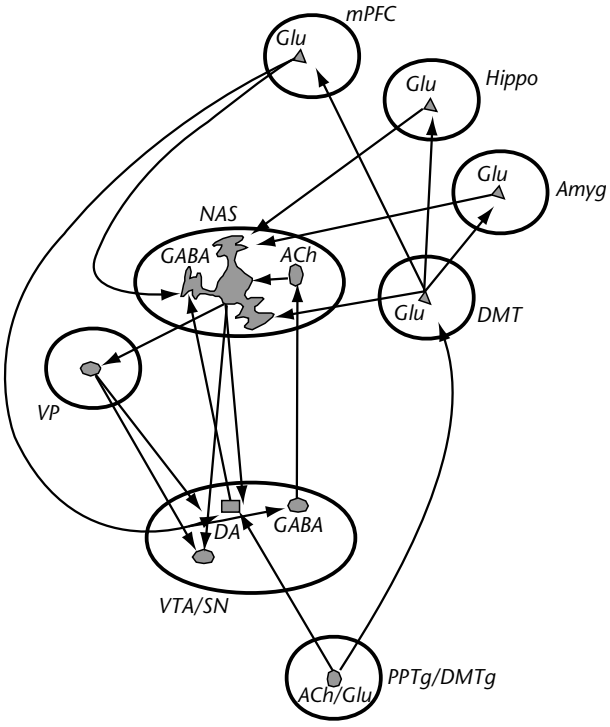


Figure 1. Some of the known complexity of brain reward circuitry. Subdivisions of the ventral tegmental area and nucleus accumbens are not shown because of uncertainties about the synaptic connections within them. Many related structures are not shown. Excitation, inhibition, and modulation are not differentiated. Colocalized peptide transmitters are not shown. Specific synaptic relations between afferents and efferents of each region are not shown in their known complexity. Receptor subtypes, sites of dendritic transmitter release, and sites of presynaptic (axon–axon) contacts are also not shown. Structures: Amyg, amygdala; DMT, dorsomedial thalamus; Hippo, hippocampus; mPFC, medial prefrontal cortex; NAS, nucleus accumbens; PPTg/DMTg, pedunculopontine and dorsomedial tegmental nuclei; VP, ventral pallidum; VTA/SN, ventral tegmental area and substantia nigra. Transmitters: ACh, acetylcholine; DA, dopamine; GABA, γ-aminobutyric acid; Glu, glutamate.

project to medium spiny neurons. The medium spiny neurons also receive glutamatergic input from the frontal cortex, amygdala, hippocampus, and dorsomedial thalamus. In large part, the system appears to parallel the extrapyramidal motor system with its dopaminergic link from the substantia nigra to the dorsal striatum.

Additional structures are clearly involved, as brain stimulation reward sites have been identified at all levels of the brain, from the sensory regions of the olfactory bulb and nucleus of the solitary tract, through the arousal systems of the brainstem and mesencephalon and the association areas of the diencephalon and telencephalon, to the region of the motor nucleus of the fifth cranial nerve. It is not clear to what degree these 25 or so reward sites connect to one another or converge on a common path, but while they have different sensitivities, none seems immune to modulation by dopaminergic drugs.

The role of dopamine in brain function is not well defined. The symptoms of schizophrenia, a mental disorder, are ameliorated by dopamine antagonists, and the symptoms of Parkinson disease, a movement disorder, are ameliorated by dopamine agonists. Extracellular dopamine levels are elevated not only by drug rewards, brain stimulation reward, sex, and food reward, but also by food deprivation. Stimulation at the lateral hypothalamic level of the medial forebrain bundle can serve not only as a reward but also as a drive. That is, rats will work for stimulation that makes them hungry; this is known as the drive–reward paradox. Dopaminergic neurons respond not only to reward-related stimuli but also to novel and painful stimuli and a variety of stressors. Thus brain dopamine and the circuitry in which it is embedded appear to play broad roles in some form or forms of arousal that are common to the range of stimuli that have behavioral significance for the animal. While the normal functioning of this circuitry is necessary for normally rewarding stimuli to exert their behavioral control, normal function of this circuitry is necessary for far more than simply reward function. Thus it should not be assumed that this circuitry is in any way 'specialized' for reward function.

When discussing the brain circuitry of reward it is frequently unconditioned rewards that are assumed: brain stimulation, addictive drugs, food, sex, and other biologically significant rewards. However, the cases of brain stimulation reward and injected drug reward make it clear that conditioned rewards must depend on this circuitry as well. In these two laboratory cases, there is no external incentive except the environmental stimuli that have been associated with unconditioned rewards in the experimental animal's reinforcement history. The stimuli that guide the animal's behaviors are the discriminative stimuli that signal reward availability and the incentive motivational stimuli and conditioned reinforcers that have been associated with the reward in question. In cases where the animal presses a lever for the drug or the stimulation, these include such stimuli as house lights, cue lights, the lever itself, and the noises that result from lever pressing. The drug itself (or the stimulation) does not present a sensory image, being delivered through an intravenous catheter. The drug is unexpected at first and subsequently signaled only by learned cues, and it is these cues – 'reinforced' by repeated association with the drug itself – that are present at the time the animal makes its habitual response. If we assume the rewarding impact of the drug to be experienced only when the drug elevates nucleus accumbens dopamine levels, the learned association may have to span minutes – a drug like cocaine, for example, must enter the bloodstream, move to the brain, cross the blood–brain barrier and bind to the dopamine transporter before a dopamine build-up will begin. Then dopamine cells must be stimulated to fire by other forces – cocaine itself, locally applied, actually decreases dopamine cell firing – before dopamine is released and allowed, by cocaine's action, to build up to significant levels. That learned associations are important is shown by the fact that an animal will press repeatedly for the drug-associated cue light alone. The fact that dopamine antagonists reduce responding for the cue light alone suggests that dopamine is important for conditioned as well as unconditioned reinforcement. Both nucleus accumbens and the central nucleus of the amygdala are important for the control of behavior by conditioned reinforcers.

Many of the elements of the brain circuitry of reward, and the way the various components of reward circuitry interact, remain to be characterized. Even the historically central element, the mesolimbic dopamine system, is not fully understood. Dopamine has important reward-relevant actions at both D1 and D2 type receptors in both nucleus accumbens and the ventral tegmental area. In the nucleus accumbens, D1 and D2 actions on medium spiny neurons are opposite (D1 action activates and D2 action inhibits the cyclic AMP second-messenger cascade in medium spiny neurons), while the behavioral actions are synergistic (neither selective D1 nor selective D2 agonists are rewarding, but the combination is). Dopamine in the nucleus accumbens seems to exert a

modulatory influence on glutamatergic activation of medium spiny neurons, but it also affects its own release from nerve terminals and acts on local interneurons and perhaps terminals. In the ventral tegmental area, dopamine acts at D2 receptors to limit the firing rate of its own neurons and acts at D1 receptors to inhibit glutamate and GABA release at incoming nerve terminals. Its D1-type actions in the ventral tegmental area are important for cocaine and probably other reward function.

The nucleus accumbens has complexities of its own. The shell subregion seems more important than the core subregion for phencyclidine and psychomotor stimulant reward, and shows greater dopamine elevations in response to drugs of abuse. The core region appears to have the more important role in opiate reward and conditioned reinforcement. Differences in the efferent connections of the two regions make this distinction a potentially important one.

Finally, the circuitry in which the ventral tegmental area and nucleus accumbens reside is complex. The dopamine-containing cells of the ventral tegmental area receive reward-related signals from the frontal cortex (as well as from drugs of abuse and from direct electrical stimulation) but so, apparently, do the GABA-containing cells of this region. While the dopamine release in nucleus accumbens has an important role in reward, it is not known what part the release of GABA plays. Dopamine is thought to inhibit the medium spiny output neurons of nucleus accumbens (both core and shell), while inputs from the frontal cortex, amygdala, hippocampus, and dorsomedial thalamus tend to excite these cells. One complication is thus that glutamate from frontal cortex not only directly excites the cells of nucleus accumbens but also excites the dopamine neurons that, in turn, inhibit the cells of nucleus accumbens. It also excites the ventral tegmental GABA cells which in turn inhibit cholinergic interneurons in nucleus accumbens, and may also contribute to GABA-mediated inhibition of the ventral tegmental dopamine neurons. Glutamate from frontal cortex projections also excites the pedunculopontine and dorsomedial tegmental pontine nuclei, which, in turn, excite the dopaminergic cells (and perhaps the GABA-containing cells) of the ventral tegmental area. As if this complexity on the input side of nucleus accumbens were not enough, some of the output cells of nucleus accumbens feed back to the ventral tegmental area, synapsing on dopaminergic or GABA-containing cells, or both. Other of the output cells of nucleus accumbens project to the GABA-containing cells of the ventral pallidum, which, in turn, project back to the ventral tegmental area, exciting the dopaminergic cells and perhaps others. Finally, the GABA-mediated output of the nearby substantia nigra and perhaps the ventral tegmental area projects to the pedunculopontine and dorsomedial tegmental nuclei and the dorsomedial thalamus.

CONCLUSION

The brain circuitry of reward is complex and as yet only partially understood. It involves many neurotransmitter substances, some of which are better known for their ability to alter the actions of other synaptic interactions than for exciting or inhibiting postsynaptic cells directly. The circuitry involves many anatomical links, with multiple opportunities for feedback and feedforward complications. It participates in multiple arousal functions, including arousal critical to response reinforcement as well as to reward anticipation and response initiation.

Further Reading

Gallistel CR (1987) Determining the quantitative characteristics of a reward pathway. In: Church RM, Commons ML, Stellar JR and Wagner AR (eds) *Biological Determinants of Reinforcement*, pp. 1–30. Hillsdale, NJ: Lawrence Erlbaum.

Gallistel CR, Shizgal P and Yeomans J (1981) A portrait of the substrate for self-stimulation. *Psychological Review* **88**: 228–273.

German DC and Bowden DM (1974) Catecholamine systems as the neural substrate for intracranial self-stimulation: a hypothesis. *Brain Research* **73**: 381–419.

Glickman SE and Schiff BB (1967) A biological theory of reinforcement. *Psychological Review* **74**: 81–109.

Olds J and Milner PM (1954) Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–427.

Wise RA (1978) Catecholamine theories of reward: a critical review. *Brain Research* **152**(2): 215–247.

Wise RA (1982) Neuroleptics and operant behavior: the anhedonia hypothesis. *Behavioral and Brain Sciences* **5**: 39–87.

Wise RA (1989) The brain and reward. In: Liebman JM and Cooper SJ (eds) *The Neuropharmacological Basis of Reward*, pp. 377–424. Oxford, UK: Oxford University Press.

Wise RA and Rompre PP (1989) Brain dopamine and reward. *Annual Review of Psychology* **40**: 191–225.

Wise RA, Newton P, Leeb K *et al*. (1995) Fluctuations in nucleus accumbens dopamine concentration during intravenous cocaine self-administration in rats. *Psychopharmacology* **120**: 10–20.

Yeomans JS, Mathur A and Tampakeras M (1993) Rewarding brain stimulation: role of tegmental cholinergic neurons that activate dopamine neurons. *Behavioral Neuroscience* **107**: 1077–1087.

Sensory Integration, Neural Basis of

Introductory article

Barry E Stein, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA
Terrence R Stanford, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA
Mark T Wallace, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA

CONTENTS

Multisensory integration is the ability of the brain to combine information from different sense receptors to enhance the organism's perception of its environment.

INTRODUCTION

Multisensory integration refers to the brain's ability to pool information from different senses. It is a process that allows the brain to construct an accurate picture of the world and to guide motor acts. Because all perceptions and behaviors take place in an environment rich with stimuli that activate very different sensory receptors, the processes underlying multisensory integration are continuously in operation. They add depth to our sensory experiences and accuracy to our judgments of environmental events.

CROSS-MODAL PERCEPTION

Basketball players don't seem so tall when we see them from the cheap seats or watch them play on television. We know that this misimpression is a result of judging the players against one another, and this is confirmed at the end of the game when we see one or more of them interviewed by someone of average height. The apparent change in a player's height may be amusing, but it is rarely startling, because we have learned very early in life that visual judgments are not absolute, and can be easily manipulated by the clever use of contrast and perspective.

Indeed, many elementary-school children are delighted to draw innumerable versions of railroad tracks that seem to disappear in the far distance. We also find it obvious that sounds appear louder or softer than usual if heard after sounds of contrasting volume, and are not surprised to find that normal speech has become inaudible after a rock concert. The relativity of sensory impressions characterizes each of our senses (i.e. sensory modalities).

Nevertheless, the idea that this relativity extends across different sensory modalities is less obvious. It is difficult at first to believe that the presence of a brief, low-intensity sound can markedly change one's estimate of the brightness of a light – but it does. It is also true that vibrating one side of the neck can make a spot of light in an otherwise darkened room appear to move. Similarly, activation of receptors in the vestibular and proprioceptive systems by rotating the body can make it seem as if that horizontal illuminated line we are viewing in the dark is now tilted, and stimulation of receptors in the body by the powerful gravitational forces produced by high-speed takeoffs from aircraft carriers can temporarily disrupt a pilot's visual estimates. These are just a few of the many intersensory, or 'cross-modal', influences that determine our daily perceptual experiences and on which we base our judgments about the events around us.

The difficulty of divesting oneself of the idea that each sensory modality functions in its own unique realm, from which it cannot alter judgments in other sensory modalities, is likely to be because each modality has its unique subjective impressions, or 'qualia'. Color is unique to vision, pitch

to audition, tickle and itch to somatosensation, salty to gustation, and so on. It is for this reason that one cannot really describe the color blue to someone who is born blind. Such people have had no way of constructing an equivalent experience through their intact senses and it is anyone's guess what their imagined sensations are like. Despite the unique nature of their qualia, the different sensory modalities have evolved to work in concert.

Because we know our world through our senses, and this knowledge is essential for survival, evolution has favored the creation of multiple senses that coexist in the same individual. It has also devoted massive amounts of brain tissue to ensure that the sensory information received is effectively processed so that it can modify perception and behavior. Each sensory modality carries somewhat different information because each is tuned to a different form of physical energy. Although this discussion is limited largely to sight, sound, and touch, many organisms depend on somewhat more exotic senses that are tuned to such stimuli as reflected sound (e.g. the echolocation and sonar capabilities of bats and dolphins), electrical currents (some fish), infrared heat (pit vipers), and magnetic fields and/or polarized light (migrating birds). Having multiple sensory modalities increases the range of possible stimuli that an organism can use to detect an event, allows the senses to substitute for one another when necessary (touch, smell, and hearing can substitute for vision in the dark) and can reduce ambiguities that often occur when evaluating an event through only one sensory dimension (two events may sound alike but look different).

For these reasons alone, developing multiple sensory modalities would have been extremely useful, and would have had a significant impact on evolution. However, there is a consequence of this evolutionary development that goes well beyond the obvious assets associated with the simple addition of new systems, and that is the ability to use the information carried by different sensory modalities in an integrative fashion. This 'multisensory integration' is possible, in large part, because many areas of the brain serve as common targets for multiple sensory channels and contain neurons that are able to share and compare the cross-modal information they receive. The final product of these processes is the enhancement of signals in the brain that are linked to common events, and hence are meaningful in an operational sense, as well as the suppression of nonmeaningful, or distracting, signals. This is obviously useful in determining which events are to capture an organism's immediate attention and overt responses. The efficiency of this system could not be achieved with independent channels of sensory information, and the neural bases for this multisensory integration are explored in more detail below.

Even without modern neurophysiological studies, the existence of neural mechanisms capable of synthesizing cross-modal cues – and even some of the principles by which these mechanisms must function – can be readily surmised from perceptual experience. As noted earlier, cues in one sensory modality can alter judgments in another. These cross-modal phenomena have been noted for some time, and it has become evident that one can often create cross-modal illusions by introducing slight discrepancies into the normal relationship between the different sensory stimuli that ordinarily are derived from the same event. This has been exploited by entertainers who have mastered the ventriloquist's art. We know that the dummy is not speaking, but the movement of its lips (and sometimes its head and eyes) in rough synchrony with what we hear produces a compelling illusion that the sound is coming from its mouth.

Actually, the ventriloquist's trick is less a consequence of the entertainer's skill (though few of us could drink water while reciting a limerick) than of the manner in which our brains integrate the visual and auditory cues associated with speech. The spatial resolution of the visual system is extraordinary, and it is no accident that it is this sensory modality on which we depend most heavily for locating events in space. Hence, the spatial discrepancy between what is received by the ear and the eye is resolved by the brain, attributing the source of the speech to the more compelling visual stimulus – the dummy. The far more common form of this illusion is experienced whenever we watch a film or a television show. The sounds are linked to events on the screen regardless of their location, when in reality all the sounds are coming from the same source: the speakers in the cinema or in the television cabinet.

There are many cross-modal illusions that involve different combinations of unimodal cues. Some involve translocation of the apparent source of the stimulus, others involve apparent changes in the meaning of the stimuli. Yet each serves as another illustration of the fact that the brain readily combines information from different senses in order to construct a representation of the outside world. This is a normal functional process, and in most circumstances the multiple sensory cues derived from a single event are coincident in space

and time. Such cross-modal cues produce an enhanced perceptual and/or behavioral product that can exceed statistical predictions based on responses to the individual unimodal cues.

THE SUPERIOR COLLICULUS

A Model for Understanding the Neural Bases of Multisensory Integration

Information from different senses is brought together in the brain by the convergence of their various projections onto the same neurons (Figure 1). Whenever and wherever such convergence takes place there is the opportunity to synthesize information from the sensory modalities involved and, it is important to note, this sort of sensory convergence takes place at all levels of the neuraxis. Thus, it is a common event, and the sites involved are generally found outside the primary projection pathways (those pathways that carry information in its most direct route from the peripheral sensory organ to a primary sensory cortex). One of the best known and most closely studied of these multisensory sites is the superior colliculus (SC).

The superior colliculi look like a pair of bumps on the surface of the midbrain (Figure 2). Each SC uses visual, auditory, and somatosensory information to initiate attentive and orientation movements, and it produces these movements through its connections with structures in the brainstem and spinal cord which are in more direct contact with muscles. The sensory cues received by the SC are used both individually and together in order to orient the sensory organs towards an interesting object or event that occurs on the opposite side of the body (i.e. in contralateral space). While the SC is best known for initiating and controlling eye and head movements such as gaze, it is involved in coordinated behaviors in which various peripheral sensory organs (eyes, head, ears, limbs, whiskers, and mouth) are moved contralaterally toward the source of stimulation. Most relevant to this discussion are the observations that have been made about how individual neurons in this structure integrate their various sensory inputs in order to produce these behaviors.

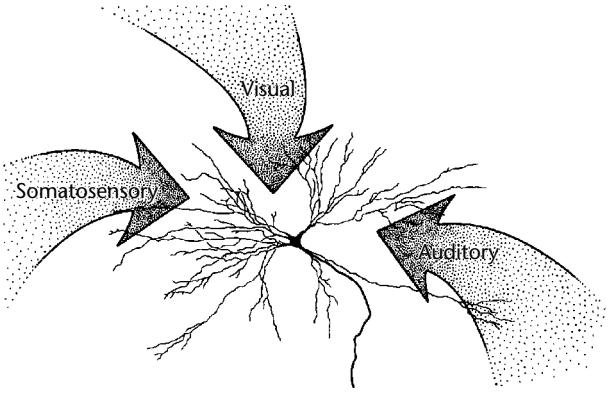


Figure 1. Visual, auditory, and somatosensory neurons often have converging projections onto other neurons, thereby rendering them multisensory. Adapted from Stein and Meredith (1993).

Spatial Register of Cross-modal Receptive Fields

The superior colliculus receives its visual, auditory, and somatosensory information via projections from many brain structures. However, all these projections terminate in the SC in a systematic manner, forming map-like or 'topographic' representations (Figure 3). These maps are composed of neurons, each of which responds to sensory cues only within a circumscribed region of space (its 'receptive field'). Although these inputs are considered sensory and confer sensory responsiveness onto SC neurons, it is important to emphasize that the maps they form are not based on the usual sensory referents that are found in structures more traditionally linked to the processing of unimodal visual, auditory, or somatosensory information. In these 'primary' sensory structures, the visual, auditory, and somatosensory topographies are linked to the retina, head, and body surface, respectively. In the SC, in contrast, each of the 'sensory' representations has been transformed from its own particular (and unimodal) reference frame to a common 'motor' frame of reference, a transformation that is critical for translating a sensory stimulus into a motor action.

The implications of this transformation for orienting can be illustrated by considering the problem of having to shift gaze to an interesting sensory stimulus, one of the primary functions of the SC, and one in which multisensory integration plays a critical role (see below).

Accordingly, there is a systematic arrangement of motor-related SC neurons. Each of these neurons is selective and is most active in association with the generation of a shift of gaze to a particular region of space. Neurons found in progressively more rearward positions of the (right) SC code progressively larger movements of the eyes and head (to the left), whereas neurons progressively further toward the midline code progressively more upward movements and those more lateral produce downward shifts of gaze. As a result, the

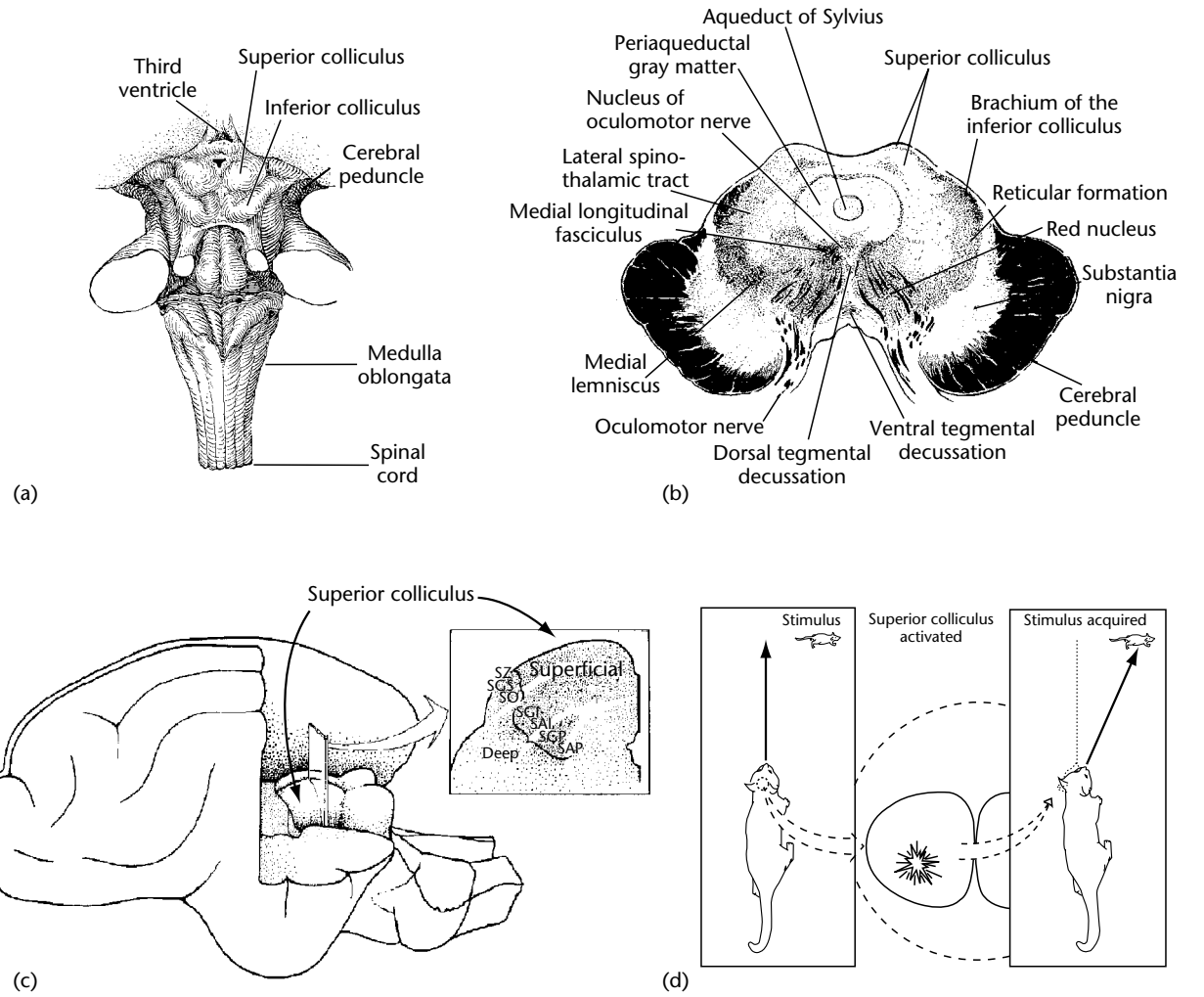


Figure 2. The structure and function of the superior colliculus (SC). (a) Dorsal view of the human brain, in which the overlying cortex and cerebellum have been removed. Note the location of the superior colliculi, which appear as a pair of bumps on the surface of the midbrain. (b) A transverse section through the human midbrain illustrates the location of the SC as well as subjacent tissue. (c) Cat brain, in which the caudal portion of the cerebral cortex has been removed to show the underlying midbrain and SC. A transverse section has been made through the middle of the structure. It is viewed from the front in the inset, and shows the highly laminated organization of the structure in which there are alternating darker (primarily cellular) and lighter (primarily fibrous) layers. The deeper layers are where multisensory neurons are located. SZ, stratum zonale; SGS, stratum griseum superficiale; SO, stratum opticum; SGI, stratum griseum intermediale; SAI, stratum album intermediale; SGP, stratum griseum profundum; SAP, stratum album profundum. (d) Role of the SC in orientation movements. In this example, the stimulus (a rodent) gives rise to a focus of neural activity within the SC's visual topographic map. This activity initiates promotor neural responses that are associated with the orientation of the animal's eyes, ears, and head toward the initiating stimulus. Parts (a) and (b) adapted from Chusid JG and McDonald JJ (1967) *Correlative Neuroanatomy and Functional Neurology*, 13th edn. Los Altos, CA: Lange Medical Publications; parts (c) and (d) adapted from Stein and Meredith (1993).

site of activity within the SC motor map dictates the size and direction of a gaze shift.

Returning to the problem of orienting gaze toward a sensory stimulus, it would seem an efficient coding strategy if the sensory and motor topographies were in register so that the site activated by a sensory stimulus would correspond exactly to the site at which motor-related activity would produce the shift of gaze necessary to look directly at that stimulus. So, for example, if one were looking directly ahead and an auditory stimulus were 10° to the right of midline, the activity it produces in the SC would be at the site at which the activity of motor-related neurons would produce a 10°

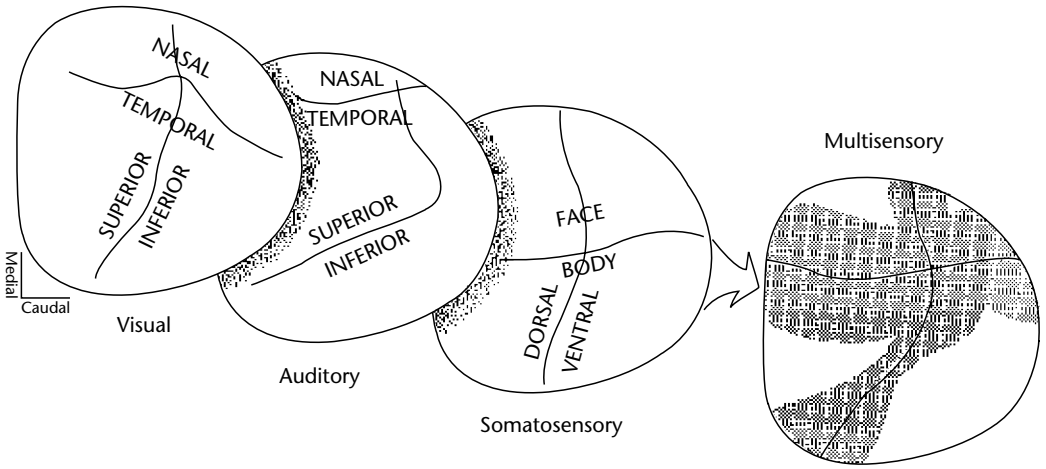


Figure 3. The superior colliculus is composed of maps of visual, auditory, and somatosensory space. These sensory maps have similar coordinate frames and involve many of the same multisensory neurons. Thus, these maps can be viewed as components of an integrated multisensory map. Adapted from Stein and Meredith (1993).

rightward shift of gaze. However, if the eyes happened to be pointed 10° to the left, a 20° rightward movement would be required to look at that same stimulus – a stimulus that has not changed in 'auditory' (head-centered) coordinates. The stimulus must now activate a different site in the SC motor map, one further rearwards in the SC. To ensure that any given auditory stimulus evokes activity at the SC site capable of producing the correct gaze shift, the auditory information has been transformed from a head-centered to an eye-centered frame of reference. The same logic can be applied for a tactile stimulus. Thus, insofar as orienting gaze is concerned, SC sensory topographies represent the position of stimuli with respect to current gaze position, not stimulus position in 'sensory space' as is more commonly done elsewhere in the central nervous system.

The presence of a common reference frame for representing sensory information has important consequences for multisensory integration. By definition, a multisensory neuron is responsive to stimuli from more than a single sensory modality. Accordingly, a multisensory neuron can be viewed as having multiple receptive fields, one for each modality to which it responds. Among SC multisensory neurons, 'bimodal' neurons are most common. These are, in order of incidence: visual–auditory, visual–somatosensory, and auditory–somatosensory. Trimodal (visual–auditory–somatosensory) neurons are also encountered, but they are comparatively rare. A common coordinate system for representing sensory information allows for topographic register among the modality-specific receptive fields, an important factor for the way in which these neurons integrate information.

The general alignment of the unimodal sensory representations has been demonstrated in studies carried out in animals with eyes, ears, head, and body facing forward so that the axes of visual (retina), auditory (head), and tactile (body) space are in approximate alignment with the direction of gaze. For each sensory modality, receptive fields shift systematically across space as one samples neurons in any given direction across the SC. For example, visual or auditory neurons in the front of the SC have their receptive fields in central space (consistent with sites in the motor map that produce small contralateral gaze shifts), whereas those located progressively further rearwards in the structure have their receptive fields shifted progressively more eccentrically into the peripheral aspects of contralateral space (in register with sites that produce larger contralateral gaze shifts). In short, when the animal is facing forwards, neurons in the front of the structure represent the space in front of the animal and those in the rear of the structure represent space in the periphery. The somatosensory representation corresponds to this organization in that neurons with receptive fields on the face are located rostrally (forwards) in the SC and those with receptive fields progressively further back on the body (towards the rump) are located progressively more caudally (rearwards). Neurons found towards the midline (medial in the structure) have receptive fields in upper space (or on the upper body) and laterally located neurons have receptive fields progressively lower in space (or lower on the body). The cross-modal

receptive field register among these neurons ensures that regardless of which sensory cues (visual, auditory, somatosensory) are derived from an event, they will activate the same neurons and thereby lead to activation of the same motor circuitry that is necessary to elicit an appropriate contralateral orientation response.

MULTISENSORY RESPONSE ENHANCEMENT AND DEPRESSION

Multisensory neurons do more than simply respond to cues in different sensory modalities: they are able to synthesize the unimodal inputs they receive into an integrated multisensory product. Operationally, multisensory integration has been defined as a statistically significant difference between the number of impulses evoked by a cross-modal combination of stimuli and the number evoked by the most effective of these stimuli alone. Multisensory integration can result in a response that is markedly enhanced, and that can exceed the arithmetic sum of the individual unimodal responses. It can also result in a response that is markedly depressed and sometimes eliminated. One of the most important determinants of this interaction is the spatial relationship among the cross-modal stimuli.

When stimuli are derived from the same event they also originate at approximately the same time and from the same location, thereby allowing them to stimulate their respective receptive fields of the same multisensory neurons. The two resultant excitatory signals interact in a synergistic fashion to enhance the neuron's response (Figure 4). The magnitude of this response enhancement is generally higher when the individual unimodal stimuli are weakly effective, suggesting that the benefits of multisensory enhancement are especially great when stimuli are difficult to detect individually. On the other hand, if these same stimuli are derived from different locations (e.g. they originate from different events), so that, for example, the visual stimulus falls inside the neuron's visual receptive field and the auditory stimulus falls outside the neuron's auditory receptive field, either response depression will occur, or there will be no interaction at all (Figure 4). These spatial relationships show how important it is to maintain the register among the receptive fields of multisensory neurons even if the peripheral sensory organs move relative to one another. If the register varied, the detection of important events might be compromised, and/or unrelated stimuli might gain preferential access to the circuitry of the SC.

Behavioral studies have shown that the principles that govern multisensory integration at the single neuron level in the SC also apply to SC-mediated overt behaviors. Specifically, cross-modal stimuli derived from the same location in space produce interactions that enhance behavioral responses (e.g. detection and orientation), whereas spatially disparate stimuli inhibit one another's effectiveness (Figure 5).

THE ROLE OF ASSOCIATION CORTEX

Although one might think that when cross-modal inputs converge on a particular neuron, the neuron would automatically be rendered capable of synthesizing these inputs, this is not correct. Some of the multisensory neurons in the SC are unable to synthesize the cross-modal inputs they receive, and nearly all of the others can do so only in the presence of influences from specific areas of association cortex. If these influences are eliminated by temporarily deactivating these areas of cortex, the SC neuron continues to respond to the individual unimodal inputs, but can no longer synthesize these inputs to produce the response enhancement that characterizes normal multisensory integration (Figure 6).

The importance of these cortical influences is also apparent behaviorally. Temporarily deactivating these association cortices has been shown to have no effect on an animal's ability to orient to a visual cue, but compromises its ability to use spatially coincident visual and auditory cues synergistically to enhance its performance (Figure 6). Similarly, the inhibitory effects of a spatially disparate auditory cue on its responses to the visual cue are significantly ameliorated. Thus, these 'association' areas of cortex produce at least some of their associative products via the SC and, in doing so, control some of the behaviors the SC mediates.

SUPERIOR COLLICULUS VERSUS MULTISENSORY CORTEX

Experiments like those described above in the SC have also been conducted in multisensory areas of cortex. Here, too, the different receptive fields of a multisensory neuron show the same sort of spatial register that typifies SC neurons. In addition, the same spatial principles of multisensory integration appear to apply (although the incidence of response depression is somewhat lower and it is less strong), and, as in the SC, the largest multisensory response enhancements are obtained with weakly effective unimodal cues. Lastly, the window

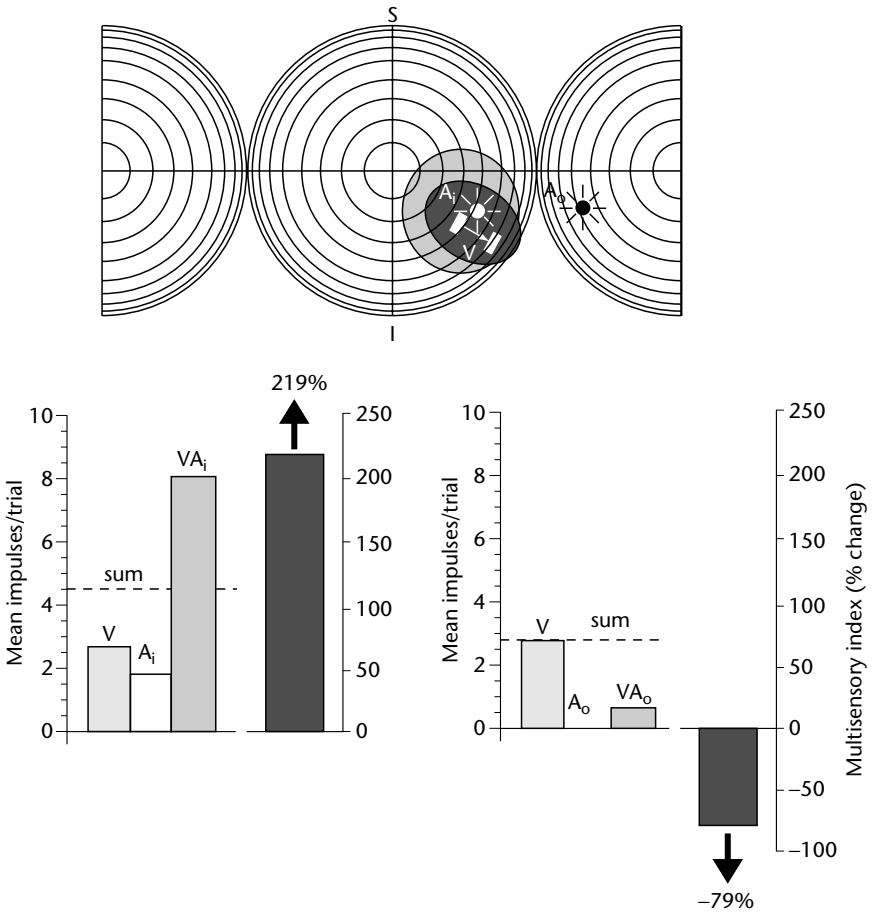


Figure 4. An example of the stimulus conditions that evoked multisensory response enhancement and multisensory response depression in a single SC neuron. At the top are shown the receptive fields (shaded) and locations of the stimuli presented (the visual stimulus is represented by the bar of light with arrow indicating direction of movement; the star represents the location of the auditory stimulus) for examining the multisensory responses of this visual-auditory neuron. The receptive fields are plotted onto a standardized representation of visual and auditory space in which the central 90° of space is depicted by the central sphere. Each concentric circle within the sphere represents 10°; the horizontal and vertical meridians are shown as straight lines. Auditory space extends around the animal, and its caudal aspect is depicted by the two hemispheres that have been split and folded forward. The bar graphs show the modality-specific and multisensory responses of this neuron to stimuli at two different configurations: when both the visual (V) and auditory (A_i) stimuli were within their respective receptive fields and, thus, at similar spatial locations (left), and when the auditory stimulus (A_o) was displaced from the visual and positioned outside its receptive field. The sum of the activity of the two modality-specific stimuli is indicated by the dashed line. Note the large multisensory response enhancement (VA_i) that resulted when both stimuli originated from a similar location, and the depression of the visual response (VA_o) when the auditory stimulus was moved outside its receptive field so that the two stimuli were spatially disparate. Dark shaded summary bar on the right illustrates the proportionate change induced by the stimulus combination (i.e. the multisensory index). This index is calculated by the following formula:

$$[(\mathrm{CM} - \mathrm{SM}_{\max})/(\mathrm{SM}_{\max})] \times 100 = \%\ \text{interaction}$$

where CM is the mean number of impulses evoked by the combined-modality stimulus and $\mathrm{SM}_{\max}$ is the mean number of impulses evoked by the most effective single-modality stimulus.

of time during which cross-modal stimuli can produce a multisensory interaction is similar to that seen in the SC. This parallel between the principles that govern multisensory integration in the midbrain and in the cortex is likely to reflect a more widespread condition that makes intuitive sense. A core of fundamental principles of multisensory integration operating across the brain would ensure that the salience of a given stimulus complex would be enhanced or degraded in the many

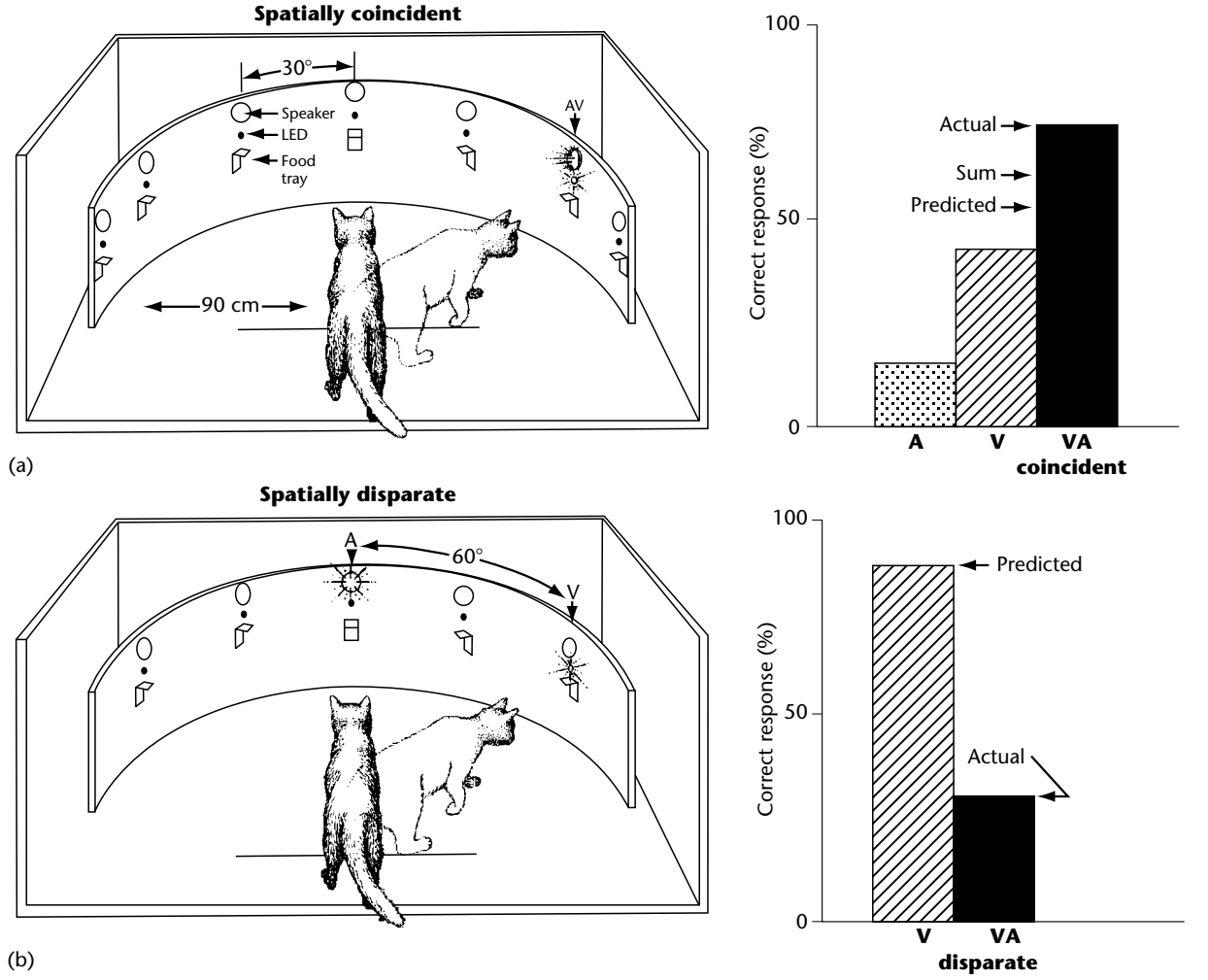


Figure 5. Conditions evoking multisensory response enhancement and multisensory response depression at the behavioral level. The animal is first trained, in the perimetry device shown, to orient to and approach a high-intensity visual stimulus provided by a light-emitting diode (LED). After learning, the intensity of the visual stimulus is decreased until the animal responds correctly on approximately 50% of the trials at each location within the perimetry (it is not trained to respond to the auditory stimulus). (a) Spatial coincidence: when the visual (V) and auditory (A) stimuli were presented at the same spatial location during testing, the cross-modal stimulus combination resulted in a significantly enhanced behavioral response. (b) Spatial disparity: in this case, the intensity of the visual stimulus was increased so that correct responses to the visual stimulus approximated 90%. The concurrent presentation of an auditory stimulus 60° disparate to the visual stimulus resulted in a significant behavioral depression. Adapted from Stein and Meredith (1993).

different brain regions that normally cooperate in producing a complex sensory experience and the behaviors that result from that experience.

Many of these same principles appear to be operative in very different species (e.g. carnivores, rodents, and nonhuman primates), and a number of studies have examined multisensory processes in the human cortex. Using techniques that can record electrical brain activity from the scalp, or that measure localized increases in blood flow by means of new imaging techniques, investigators have identified a number of multisensory areas of human cortex, including the right frontotemporal area, right insula-claustrum, left basal posterior temporal lobe, and the medial inferior parietal lobe. Furthermore, some of the same principles of multisensory integration that have been determined at the single neuron level in animals appear to apply to the activity of networks of neurons in human cortex.

It is not yet known whether the cross-species constancies in the fundamental principles governing multisensory integration that have been discussed here reflect a common origin or the convergent evolutionary pressures of ecological commonalities. Yet, however they came about,

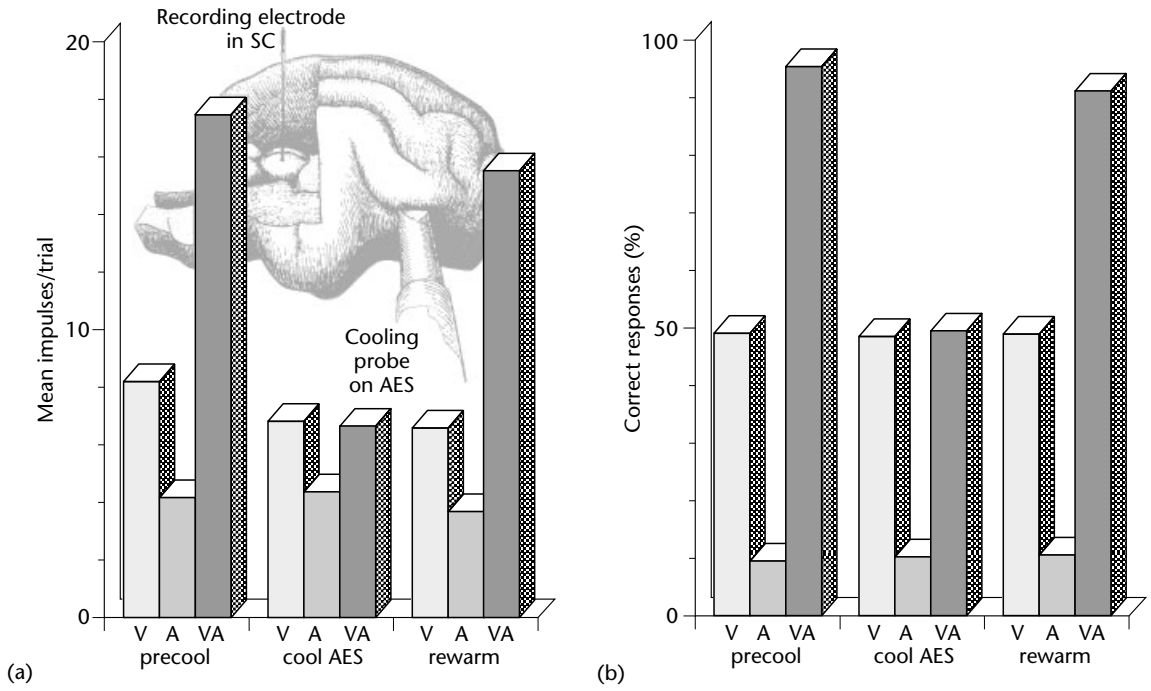


Figure 6. The cortex plays a critical role in mediated multisensory integration in the superior colliculus (SC). Neuronal (a) and behavioral (b) data illustrate the impact of deactivating the anterior ectosylvian sulcus (AES) on multisensory processes. Prior to cortical deactivation ('precool'), the response of the representative visual-auditory neuron to the combination (VA) of a visual stimulus (V) and an auditory stimulus (A) not only exceeded each individual modality-specific response (V or A), but it exceeded their arithmetic sum as well. This multisensory response enhancement was abolished during the deactivation of AES ('cool AES'). Now the response to the stimulus combination is no better than the response to the most effective modality-specific stimulus (V). Multisensory response enhancement returned when the AES was rewarmed ('rewarm'). Deactivation for the behavioral study was accomplished by means of small cooling coils implanted in the AES. Prior to cortical deactivation this animal showed a significant enhancement in correctly responding to the VA stimulus combination. However, deactivating AES eliminated this enhancement so that responses were no better than to the visual stimulus alone. Multisensory response enhancement was restored after cortical reactivation.

these principles appear to be capable of governing the integration of cross-modal information from several different senses. Although it may be reasonable to assume that these same principles are likely to govern the synthesis of information derived from even exotic sensory systems, a little caution may be in order. Because little is known about multisensory integration in systems other than the visual, auditory, and somatosensory systems, and comparatively few animal species have been examined thus far, it may be too soon to rule out the possibility that unique, species-specific principles of multisensory integration will be discovered.

Further Reading

Baron-Cohen S and Harrison JE (1997) *Synesthesia*. Cambridge, MA: Blackwell.

McGurk H and MacDonald J (1976) Hearing lips and seeing voices. *Nature* **264**: 746–748.

Meredith MA and Stein BE (1986) Visual, auditory and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology* **56**: 640–662.

Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.

Stein BE, Meredith MA and Wallace MT (1995) Neural mechanisms mediating attention and orientation to multisensory cues. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 683–702. Cambridge, MA: MIT Press.

Stein BE, Wallace MT and Stanford TR (2000) Neural mechanisms of cross-modal synthesis. In: Goldstein EB and Malden MA (eds) *Blackwell's Handbook of Perception*, pp. 709–736. Oxford, UK: Blackwell.

Walk RD and Pick LH (1981) *Intersensory Perception and Sensory Integration*. New York, NY: Plenum Press.

Welch RB and Warren DH (1986) Intersensory interactions. In: Boff KR, Kaufman L and Thomas JP (eds) *Handbook of Perception and Human Performance*, vol. I, *Sensory Processes and Perception*, pp. 25.1–25.36. New York, NY: John Wiley.

Sex Differences in the Brain

Introductory article

Melissa M Holmes, University of British Columbia, Vancouver, Canada
Liisa A M Galea, University of British Columbia, Vancouver, Canada

CONTENTS

Differences between male and female nervous systems exist in a multitude of species. In the human brain, sexual dimorphism occurs in areas related to sexual and maternal behavior, and also to spatial abilities. Gonadal hormones may affect these differences at the developmental stage.

INTRODUCTION

Despite obvious anatomical and behavioral differences between males and females across mammalian species, the discovery of structural and functional differences in the brain was, and often still is, fascinating and controversial.

It is important to clarify that there are sex differences in the nervous systems of a multitude of species, ranging from invertebrates to mammals. Furthermore, a number of nervous system regions have been found to be sexually dimorphic with respect to both gross and microstructural morphology (e.g. the number of neurons, synapses, dendritic branching). A list of common sexual dimorphisms in the brain is given in Table 1. This article discusses the best-characterized dimorphisms of structure that have putative functional significance in human cognition and behavior, particularly the sexually dimorphic nuclei of the preoptic area (SDN-POA), the corpus callosum, anterior commissure, and hippocampus.

DISCOVERY OF THE SEXUALLY DIMORPHIC NUCLEI

The SDN-POA of the hypothalamus is approximately five times larger in the male than in the female in many species, including the human. This dimorphism is intriguing as the SDN-POA is located in the medial preoptic area (mPOA), a brain region that is essential for the expression of sexual behavior in male rats and maternal and sexual behaviors in female rats.

Research focusing on this brain region in humans has been somewhat controversial. In the human brain, there are four interstitial nuclei of the anterior hypothalamus (INAH-1 to INAH-4) and INAH-1 is thought to be equivalent to the SDN-POA. Roger Gorski and colleagues did not find a dimorphism in INAH-1, but they did report that INAH-2 and INAH-3 were significantly larger in human males. Interestingly, Simon LeVay reported that the size or volume of INAH-3 in homosexual men was similar to that of women. Owing to the correlational nature of these data, it is impossible to determine the direction of causality – that is, does the size of INAH-3 determine sexual orientation, or does orientation alter INAH-3 size?

Other sexual dimorphisms in the human brain focus on the connectivity between the left and right hemispheres. The corpus callosum (CC) is the major connective pathway between the two hemispheres and is essential for the communication of information between the left and right sides of the brain. It is thought that females have a larger CC relative to total brain size than males, although not all studies find this particular sex difference. An additional connective structure, the anterior commissure, is also sexually dimorphic, with women having a larger midsagittal area of this structure then men. Further, homosexual men tend to have a larger midsagittal area than either women or heterosexual men. Taken with the aforementioned data relating sexual orientation to INAH-3 size, this further supports a link between sexual orientation and brain structure, although again direction of causality remains elusive. The massa intermedia is an area that can connect the two thalami and is larger in females than in males. Ultimately, these sex differences in connectivity between the

Table 1. A partial list of sexually dimorphic regions in the mammalian central nervous system (CNS)[a]

Region of the CNS	*Sex difference*	*Gonadal hormone effects*
Bed nucleus of the stria terminalis		
MP volume	M > F	Developmental estradiol
MA volume	F > M	Developmental estradiol
Corpus callosum	F > M (humans)	?
	M > F (rats)	Developmental and adult levels of testosterone
Cortex		
Cortical thickness (right hemisphere)	M > F	Development levels of testosterone on left hemisphere
Frontal cortex (dendritic arborization)	M > F	?
Occipital cortex	M > F	Developmental and adult levels of testosterone
Parietal cortex (dendritic arborization)	M > F	?
Hippocampus		
Total volume	M > F	Developmental levels of testosterone
Dentate gyrus	M > F	Developmental levels of testosterone
CA3 (apical dendritic spines)	M > F	?
CA3 (basal dendrites)	F > M	?
CA1 (transmitter activity)	M > F	Developmental levels of testosterone
CA1 (dendritic spine density)	?	Adult levels of estradiol
Medial amygdala	M > F	Developmental and adult levels of testosterone (some controversy)
Hypothalamus		
Periventricular nucleus	F > M	Developmental levels of estradiol
SDN-POA	M > F	Developmental levels of estradiol via aromatization
Suprachiasmatic nucleus	M > F	Developmental levels of testosterone?
Ventromedial hypothalamus		
(total volume)	M > F	Developmental levels of testosterone
(cell nuclei)	F > M	Developmental levels of testosterone
Spinal nucleus bulbocavernosus	M > F	Developmental levels of testosterone

[a]It is important to note that differences can be in area, volume, neuron size or number, spine density, etc., and the reader is directed to the primary literature for more specific details concerning these dimorphisms. Also, these sexual dimorphisms may not be consistent across mammalian species.

hemispheres have fascinating implications for lateralization of cognitive function.

The final sexual dimorphism in the human brain to be discussed here is the hippocampus. Males tend to have larger hippocampal volumes than females in a variety of species and, on average, perform better on spatial tasks than females. The hippocampus is thought to be an important structure for spatial learning and memory. Studies in behavioral ecology have shown that species with a sex difference favoring males in hippocampal volume are also the same species in which males traverse larger territories and have better spatial ability. However, in species in which there is no sex difference in hippocampal volume there is often no sex difference in territory size or spatial ability. Interestingly, female cowbirds have a larger hippocampus than the males and, as brood parasites, are required to navigate more than males.

In adult human males but not females, the right hippocampal formation is larger than the left. This is particularly interesting as the right hemisphere is historically considered to govern spatial ability. Although the sexual dimorphisms in the hippocampus are not as large as in other structures, work with other species suggests that the hippocampus is particularly sensitive to sex differences in experience as well as endocrinological environment (discussed further below).

EARLY HORMONES AND CEREBRAL DIMORPHISMS

Conventional theory regarding hormones and sexual differentiation states that androgens, notably testosterone, are responsible for masculinizing an animal and that estrogens, notably estradiol, are responsible for feminization. Further convention states that hormones can have two types of effects: organizational and activational. Organizational effects are those that occur during a critical or sensitive period during development and have

permanent effects on the nervous system. Activational effects occur later in life and serve to cause transient alterations in behavior. While many exceptions to these two rules have been demonstrated, they remain a suitable framework for considering how hormones alter the development of sexually dimorphic regions of the mammalian nervous system.

The role of testosterone and estradiol in masculinization of the SDN-POA has been well established in rats. Castration of males early in postnatal development results in the development of a feminine SDN-POA, while administration of testosterone to females during development will induce a masculine SDN-POA. Adult hormone manipulations do not appear to have any effects on SDN-POA size, implying that hormone level early in development, during a critical period, is essential for differentiating this sexually dimorphic structure. Interestingly, masculinization of the SDN-POA is not accomplished directly by testosterone in rodents. Testosterone can be metabolized to estradiol through a process called 'aromatization'. In rats, aromatized estradiol causes the masculinization of the SDN-POA. However, research with human and nonhuman primates has suggested that aromatization is not involved in the masculinization of primate brains. This, combined with the difficulty of investigating the human brain, results in little being known about the endocrinological mechanisms that govern masculinization of the SDN-POA, or the relevant INAH, in humans.

Testosterone is involved in the masculinization of the rat CC. Blocking androgen receptors prenatally followed by postnatal castration results in a feminine CC, while administration of testosterone to females results in a masculine CC. Unlike differentiation of the SDN-POA, this appears to be via androgen receptors and not by aromatization to estrogen. However, it is essential to consider that, unlike humans (where females are thought to have greater total CC volume), male rats have greater total CC volume than females, so the endocrine mechanisms underlying differentiation of this structure may not be identical across species. In adult human males, the size of the posterior portion of the CC (which is some studies is larger in males) is positively correlated with adult testosterone concentration. This further implicates a relationship between testosterone level and CC morphology, although these data do not conclusively demonstrate direction of causality, nor confirm the developmental time point at which the hormone may be having its effects.

Gross differentiation of the hippocampus also appears to occur early in development of rats, as sex differences can be seen in prepubertal animals. However, many of the more subtle sex differences in cellular connections or electrophysiological properties can be altered by hormone manipulations in adulthood. This is particularly true for dendritic synapse number, which can be altered in adult female or male rats by manipulating estradiol levels. The hippocampus is remarkably altered by gonadal hormone level (testosterone and estradiol) as well as by behavioral experiences such as stress and enriched environments. Indeed, a number of studies have shown that these interactions can be remarkably different across the sexes: for example, exposure to stress appears to have more dramatic detrimental effects on the male hippocampus compared with the female hippocampus.

It is clear that both androgens and estrogens are essential for sexual differentiation of the mammalian nervous system both in development and adulthood. The reader is directed to work done by Marc Breedlove and colleagues on the sexually dimorphic spinal nucleus of the bulbocavernosus for an excellent demonstration of the complex relationships between endocrinology, a sexually dimorphic structure, and behavior. Owing to the limitations of studying humans, present knowledge of the role of hormones in sexual differentiation of the human brain must come from individuals with endocrinological disorders and generalization from other species.

FUNCTIONAL SIGNIFICANCE OF HUMAN SEXUAL DIMORPHISMS

Many cognitive abilities are sexually dimorphic and these abilities may also be linked to sexually dimorphic brain regions. Females, on average, tend to perform better on tasks of verbal ability, verbal memory, fine motor control, and perceptual speed, while males, on average, tend to perform better on spatial tasks and, in some instances, mathematical reasoning. Interestingly, performance on these cognitive tasks varies with adult levels of hormones. For example, better performance on 'female advantage' tasks is associated with high levels of estradiol, while better performance on 'male advantage' tasks is associated with low levels of estradiol across the menstrual cycle in females. Spatial abilities also vary in human males, with low levels of testosterone being associated with better spatial performance. It is unclear where the neural substrates for each of these abilities lie, but

there is certainly evidence that the areas believed to control each of these cognitive abilities have some degree of sexual dimorphism. For example, the temporal superior gyrus and its subdivisions, areas linked to language abilities, are more often larger in females than males. In addition, the larger CC and anterior commissure size in human females has been purported to result in female brains being less lateralized than males in both structure and function. Studies using functional magnetic resonance imaging (fMRI) to visualize a working brain show that language is less lateralized in women than in men. Interestingly, Elizabeth Hampson has found that dichotic listening (a test of lateralization) is also related to adult levels of hormones in women, with low levels of estrogens being associated with less lateralization.

It is generally agreed that the hippocampus is involved at least partially in spatial learning and memory. Studies using MRI technology have demonstrated that London cab drivers have larger right posterior hippocampi than controls and that the right posterior hippocampal size is positively correlated with amount of cab driving experience. (London cab drivers undergo extensive training to navigate through the labyrinth of streets.) Taken together, this suggests that the sexually dimorphic spatial ability in humans is related to the known sex differences in hippocampus size.

Sex differences in the hypothalamus, specifically in the SDN or the INAH, are likely to be involved in differential expression of sexual behavior. It is well documented in many species that, along with elevated gonadal hormone levels in adulthood, masculinization of the hypothalamus results in the expression of stereotypical masculine sexual behaviors, whereas feminization of the hypothalamus results in the expression of stereotypical feminine sexual behaviors. Although the extent to which this occurs in humans remains unclear, it is certainly plausible that both endocrine level and sexually dimorphic neural structures serve to mediate the expression of sexual behavior in adult humans.

CONCLUSION

It is clear that there are numerous sexual dimorphisms in the human brain and that these dimorphisms probably play an important role in governing sex differences in behavior. The role of androgens and estrogens in the development and maintenance of these dimorphic structures continues to be a prolific area of research. Undoubtedly, androgens and estrogens are also essential for the differentiation of the human nervous system and behaviors, although future research must determine the precise nature of these effects.

Further Reading

Arnold AP and Breedlove SM (1985) Organizational and activational effects of sex steroid hormones on vertebrate behavior: a re-analysis. *Hormones and Behavior* **19**: 469–498.

Arnold AP and Schlinger BA (1993) Sexual differentiation of brain and behavior: the zebra finch is not just a flying rat. *Brain Behavior and Evolution* **42**(4–5): 231–241.

Cooke B, Hegstrom CD, Villeneuve LS and Breedlove SM (1998) Sexual differentiation of the vertebrate brain: principles and mechanisms. *Frontiers in Neuroendocrinology* **19**: 323–362.

Gorski RA (1991) Sexual differentiation of the endocrine brain and its control. In: Motta M (ed.) *Brain Endocrinology*, 2nd edn, pp. 71–104. New York, NY: Raven Press.

Kimura D (1987) Are men's and women's brains really different? *Canadian Psychology* **28**(2): 133–147.

Kimura D and Hampson E (1993) Neural and hormonal mechanisms mediating sex differences in cognition. In: Vernon PA (ed.) *Biological Approaches to the Study of Human Intelligence*, pp. 375–397. New Jersey: Ablex.

Single Neuron Recording

Intermediate article

Jeffrey S Taube, Dartmouth College, Hanover, New Hampshire, USA

CONTENTS

Single neuron recordings are the voltage or current responses monitored extracellularly or intracellularly from individual neurons over time.

INTRODUCTION

Brain functions are largely carried out by electrical activity arising in individual neurons and conveyed to other neurons. The pattern of electrical activity across many circuits within the brain gives rise to all our perceptions, feelings, and behaviors. Monitoring this electrical activity and understanding the neural information it encodes is a central theme of cognitive neuroscience.

In efforts aimed at exploring cortical localization of function, Richard Caton (1842–1926) was probably the first to record the spontaneous electrical activity of the brain (Finger, 1994). In 1875 he reported that an electrode placed on the surface of the skull recorded electrical activity that varied according to where on the skull he placed the electrode and the specific type of peripheral stimuli presented to the person being studied. (*See* **Electroencephalography (EEG)**)

Since these early experiments, many different techniques have been devised to record from populations of neurons in the brain, as well as from individual neurons. These techniques can be grouped into two general categories: intracellular and extracellular recording. Both techniques monitor neural activity by measuring voltage changes over time. This article focuses on recording from single neurons.

INTRACELLULAR RECORDING

Intracellular recording involves the placement of a small electrode (microelectrode) so that its tip is in contact with the inside of the cell. The electrode is then used to monitor changes in the cell's membrane potential over time, particularly the changes that occur during action potentials and postsynaptic events. Because neurons are small – a neuron's cell body (soma) is 10–50 μm in diameter – the tip of a microelectrode must be extremely small so that it can penetrate the cell without excessively damaging the plasma membrane and killing the cell. Although it is possible to record from inside a dendrite or axon, most intracellular recordings are made from the soma because it is the largest part of the neuron. The same microelectrode used to record from a neuron can also be used for passing electrical current or charged particles into the cell.

More recent techniques have used whole-cell patch clamping to record intracellularly. Patch clamping involves pressing a glass micropipette with a small opening against the cell's plasma membrane, as opposed to piercing it with a sharp electrode. If the pipette achieves sufficiently tight contact with the membrane, light suction can be applied so that the small patch of membrane circumscribed by the pipette is ruptured, providing access to the interior of the cell.

Because of the small size of neurons and the stability required for either of these manipulations, at present it is not possible to routinely make intracellular recordings from an awake, behaving animal. Thus, most intracellular recording is confined to neurons in anesthetized animal preparations, thin tissue slices, or tissue culture.

EXTRACELLULAR RECORDING

Extracellular recording monitors the electrical activity of a single neuron from an electrode positioned outside the cell. The electrical activity detected is composed of currents generated in the surrounding neuronal milieu by both action potentials and synaptic events. Extracellular recording from a single neuron (referred to as single unit recording) requires the electrode to be in close proximity to the cell; when the cell discharges, the magnitude of the extracellular current density generated decreases approximately as the square of the

distance from the cell. Single unit recording uses microelectrodes that are advanced into the tissue blindly with the hope that the electrode tip will terminate sufficiently close to a cell to enable detection of cell discharge. The chance of recording a single cell is increased by using multiple electrodes or microelectrodes that are moveable, usually along the dorsoventral axis. Thus, if there are no detectable cells at one location, the electrodes are advanced to a new location and each electrode is checked again to determine whether the waveform of an individual cell can be detected. This procedure is repeated until one of the electrodes can detect a neuron's waveform that is larger than the background electrical activity.

Before the recorded activity is displayed on an oscilloscope, the signal is usually amplified and filtered, so that frequencies below 300 Hz and above 10 000 Hz are significantly attenuated. The low-frequency filtering removes rhythmical brain activity that has higher voltage amplitudes than the single neurons the experimenter is trying to detect, while the high-frequency filtering removes electrical noise inherent in the recording equipment. The waveforms observed are usually biphasic, with the first phase being negative and the second phase positive with respect to ground. Occasionally the waveform is triphasic. The waveform's duration is about 1 ms, and although the peak-to-peak amplitude of the waveform is dependent on the distance between the recording electrode and the cell it is generally in the range 100–300 mV.

Frequently, more than one cell's waveform can be observed when recording with a single microelectrode. It therefore becomes important to distinguish the firing of one cell from another. Different methods have been developed to detect the occurrence of spike discharge. One of the more common techniques uses time–amplitude window discriminators, which detect the occurrence of signals passing through a range of voltages at specific times. Multiple windows in series are frequently used to detect biphasic and triphasic waveforms. More recent techniques use computers and employ template matching programs where the detected waveform must fall within a specified percentage of a 'master template'. Another method uses 'stereotrodes', which are two thin wires twisted together so that their tips are next to one another. Because they are close to one another, each wire usually detects the same nearby cells. However, the electrode that is closer to the cell will record a larger amplitude waveform than the second electrode. Graphs are constructed that plot the amplitude of the signal from one microelectrode against the amplitude of the similar signal recorded with the second microelectrode. Cluster analysis then distinguishes one recorded cell from another. To avoid the problem where multiple cells are simultaneously equidistant from both electrodes, bundles of four wires ('tetrodes') are used to distinguish the cells. Performing cluster analysis using stereotrodes or tetrodes is particularly helpful when recordings of single cells are desired from brain areas where cell density is high, such as the hippocampus.

Extracellular recordings are frequently used in conjunction with either electrical stimulation of afferent pathways or with natural occurring stimuli, such as a flash of light or brief sound. The responses observed are referred to as 'evoked responses'. Sometimes efferent pathways are stimulated and the experimenter is interested in monitoring the activity propagated back to the recording site. This type of activity is referred to as 'antidromic stimulation'.

SLICE RECORDING

Many of the methods devised to record from the brain are limited by the inability to visualize the placement of the electrode. Even when the electrode can be properly placed, the small size of neurons in a respiring animal with a heartbeat that continually produces small brain pulsations makes it difficult to maintain a recording of a single cell – even when the electrode is positioned extracellularly. To accomplish this task with an intracellular electrode requires an anesthetized preparation and considerable patience. Even then, the length of time a cell can be monitored is usually brief. The task of recording intracellularly in an awake, freely moving animal is impossible with current methodologies. However, many questions about neuronal functions and brain mechanisms can be answered only by using intracellular recordings. Thus, it is no wonder that our knowledge of how individual neurons function coincides with the development of the slice preparation.

The slice preparation involves the formation of thin brain slices (250–500 μm thick) from a larger piece of brain tissue using a vibrating tissue slicer or tissue chopper. The slices are placed on a fine mesh in a specialized chamber that warms, oxygenates, and bathes the tissue in a solution similar to the cerebrospinal fluid. Under such conditions, the slices remain reasonably healthy for more than 8 h. A microscope is positioned over the chamber and aids in the accurate placement of microelectrodes for recording and micropipettes for delivering

pharmacological agents into the slice. Thus, access to the tissue is excellent and the problems encountered *in vivo* are avoided. Depending on the plane in which the tissue is sliced, many brain regions retain a large portion of their afferent inputs, permitting the study of individual brain areas with most of their synaptic connections intact. The slice technique has therefore been of enormous benefit in exploring questions about ion channels, synaptic transmission, action potentials, microcircuitry and neuropharmacology.

SHARP AND PATCH ELECTRODES

Two types of electrodes have been used for recording: sharp microelectrodes made from glass micropipettes or fine metal wires, and patch electrodes made from glass micropipettes. Sharp electrodes were the first to be developed and are made from small glass (borosilicate) capillary tubes, known as micropipettes.

The electrodes are made by heating the capillary tube in the middle and slowly pulling the two ends apart. The resultant tip is sharp (1 μm or less for intracellular recording and 1–2 μm for extracellular recording) and in the case of intracellular recording can pierce the cell's plasma membrane with minimal damage. Despite the small tip size that results from the heat treatment, the capillary hole is preserved all the way to the tip. Thus, any solution contained in the capillary tube is in contact with the fluids outside the microelectrode. The shank of the glass micropipette is usually long (5–10 mm) in order to minimize tissue damage when the electrode is inserted into the brain, yet not so long that the shank becomes too brittle and breaks upon insertion into the tissue.

To facilitate filling of the micropipette, especially at the tip, capillary tubing is used that contains a small glass filament running the length of the pipette. This filament enhances the movement of fluid down to the tip through capillary action. The electrical resistance at the recording tip is usually 50–200 MΩ for an intracellular electrode and 5–10 MΩ for an extracellular electrode. A solution containing potassium chloride or potassium acetate is backfilled into the micropipette for intracellular recording, while a sodium chloride solution is used when recording extracellularly. With the electrode filled, a silver chloride coated wire connected to an amplifier is placed in contact with the salt solution. The silver chloride coating reduces potentials that form at the junction between the conductive solution and wire. Occasionally the solution contains a dye, which can be deposited into the cell or surrounding tissue by passing current through the electrode.

Another type of sharp electrode is the metal electrode, which is used almost solely for extracellular recording. These electrodes are constructed from fine metal wires made of tungsten, stainless steel, or nickel–chromium or platinum–iridium alloys. Except for the tip, the wires are insulated with epoxy. Irrespective of the metal used, it is important for the wire to be rigid enough to penetrate neural tissue without bending. Thus, some protocols use thick wires (about 1 mm) tapered at one end to a very fine point (0.5–3 μm). Other protocols, especially when multiple microelectrodes are used, pass several fine wires (each 15–25 μm in diameter) through a stainless-steel cannula that provides the rigidity. In these cases, the recording end protrudes 1–2 mm from the cannula. Sometimes a thin coat of an electrolyte is applied to the tips (e.g. gold plating) to reduce the electrode's impedance and thereby lower the noise characteristics of the electrode. Metal electrodes usually have resistances in the 2–10 MΩ range.

Patch clamp electrodes, like some types of sharp electrodes, are formed from small glass capillary tubes that are heated in the middle and the two halves slowly pulled apart. While the diameter of the opening at the end of the electrode is also small (about 1 μm), the shaft of the patch electrode tapers much more quickly than that of the sharp electrode, resulting in a much blunter tip. When the patch electrode is filled with a salt solution, its resistance is typically 1–5 MΩ. To record from a neuron, the patch electrode is placed next to the cell surface and mild suction is applied to the other end of the micropipette to form a tight, highly resistant seal (about 1 GΩ) between the fire-polished pipette tip and the cell membrane, so that no ions can pass through this junction. Additional suction is then applied and the membrane circumscribed by the electrode tip ruptures. The solution in the patch pipette is then in contact with the intracellular fluid and the cell's membrane potential can be measured over time.

An advantage of the whole-cell patch clamp technique over the sharp electrode technique is that the seal lowers the electronic noise and provides increased sensitivity, while the short, blunt electrode helps minimize the errors made in recording the magnitude and temporal changes in membrane potential.

REPRESENTING RESPONSES

Extracellular recording at the level of the single neuron is most often used when the goal is to

understand brain function in relation to the animal's behavior: experimenters record from single neurons and try to correlate the cell's activity with the animal's behavior. The ongoing behavior or sensory stimuli that best correlates with the unit's activity is referred to as the 'neuronal correlate' for that cell. Experiments using neuronal correlates have been applied to sensory, motor, regulatory, and cognitive systems. Although this is a useful and common technique for inferring the neural mechanisms underlying a particular behavior, the experimenter needs to be cautious because correlation does not imply causation. There are many examples where the activity of neurons in a particular brain area have correlated well with a particular behavior, only for researchers to discover later that lesioning the area has little effect on performance of a task that involves the purported function of the neurons. Thus, the neural mechanisms that underlie a particular behavior can be complex and the neuronal correlates thought to be involved can be redundant or not necessarily essential to the behavior in question. (*See* **Decoding Single Neuron Activity**; **Neurons, Representation in**)

The advantage of intracellular recording over extracellular recording is that it permits the recording of graded synaptic potentials that can lead to spike discharge. At the single cell level, extracellular recording can record the occurrence of spike discharge, but cannot detect the synaptic potentials or ionic currents from a single cell. The type of recording method selected depends upon the scientific issue under investigation and the kinds of measurements required to test a hypothesis. While an extracellular electrode can detect the discharge behavior of individual cells, it is not informative about what the population of cells in the circuit are doing or the nature of the underlying synaptic events that led to cell discharge. For example, the cell may have discharged because of excitatory events activating the neuron, or because of a decrease in inhibitory events impinging on the neuron. Intracellular recording can differentiate between these possibilities, but again, intracellular recording is not informative about population responses and it is nearly impossible to record intracellularly in the awake, behaving animal. Thus, all the different recording techniques have their own advantages and disadvantages, and are complementary. Therefore, in elucidating a particular neural process, several different techniques are often needed to bear on the problem. (*See* **Neural Inhibition**)

Reference

Finger S (1994) *Origins of Neuroscience*. New York, NY: Oxford University Press.

Further Reading

Conn PM (1991) *Electrophysiology and Microinjection. Methods in Neuroscience*, vol. 4. San Diego, CA: Academic Press.

Dingledine R (1984) *Brain Slices*. New York, NY: Plenum Press.

Hammond C (1996) The voltage-gated channels of action potentials. In: Hammond C (ed.) *Cellular and Molecular Neurobiology*, pp. 173–185. San Diego, CA: Academic Press.

Humphrey DR and Schmidt EM (1990) Extracellular single-unit recording methods. In: Boulton AA, Baker GB and Vanderwolf CH (eds) *Neurophysiological Techniques. Application to Neural Systems. Neuromethods*, vol. 15, pp. 1–64. Clifton, NJ: Humana Press.

McNaughton BL, O'Keefe J and Barnes CA (1983) The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. *Journal of Neuroscience Methods* **8**: 391–397.

Nistri A and Gutman A (1996) Advantages and disadvantages of sharp (intracellular) versus patch electrodes for measuring the resting membrane potential. In: Hammond C (ed.) *Cellular and Molecular Neurobiology*, pp. 114–115. San Diego, CA: Academic Press.

Sleep: Polyphasic

Introductory article

Claudio Stampi, Chronobiology Research Institute, Newton, Massachusetts, USA

CONTENTS

Polyphasic sleep describes the ubiquitous rest–activity pattern observed in nature, in which animals show bouts of activity and rest alternating several times per day.

INTRODUCTION

Humans are unique in that they mostly are monophasic sleepers, with a single consolidated sleep episode each night. Most other mammalian species are polyphasic sleepers, and some of them show ultra-short sleep–wake patterns, especially when living in dangerous natural environments. However, humans too have polyphasic patterns, such as in infancy, in sickness, or in around-the-clock operations. Especially in the latter conditions, when time available for sleep may be considerably reduced, and demands for human performance exceptionally high, polyphasic sleep appears to offer one of the most efficient strategies to minimize performance and cognitive impairment resulting from sleep reduction.

WHAT IS POLYPHASIC SLEEP?

The concept of *polyphasic sleep* was introduced for the first time in 1920 by Szymanski to describe the ubiquitous behavioral pattern observed in the animal kingdom, showing bouts of activity and rest alternating several times per day. Indeed, the majority (over 86 percent) of mammalian genera have typical polyphasic rest–activity patterns. The remaining species show a daily pattern of one consolidated period of sleep followed by a period of activity without any sleep, or with the intermission of a relatively brief rest–quiescence period. It appears that evolution led such minority of species to develop the ability to sustain wakefulness for relatively prolonged periods (about 16 hours), without apparent need for much intervening sleep. This is the monophasic activity pattern that is typical of, and well known for, the adult human.

Monophasic activity behavior appears therefore to be an exception to the ubiquitous and primordial behavior of the animal kingdom. However, are adult humans strictly monophasic, or is this just a variant of, or evolution from, an ancestral polyphasic behavior? In other words, do humans *need* to take sleep in one continuous, prolonged period, or can sleep be taken in several naps?

The question of whether we are irrevocably tied to a monophasic sleep pattern in order to function is becoming increasingly important in today's nonstop, around-the-clock society. Sleep occupies a substantial portion of our life span and its quality, duration, and pattern can dramatically affect wakeful functioning and performance. Insufficient or inappropriate rest has been the cause of, or a strong contributing factor to, several accidents that reached catastrophic proportions (e.g., Chernobyl, Three Mile Island, Space Challenger). Is polyphasic sleep then a suitable alternative to monophasic sleep when the latter is not an option?

NATURAL HISTORY OF POLYPHASIC SLEEP

Although the reasons why so many animal species are polyphasic (and – perhaps more importantly – why some of them *may not* be monophasic) are still open for speculation, the answers appear to be related to energy demands and exposure to risk. The negative correlation existing between body mass and metabolic rate is the reason why small mammals are forced to spend most of their waking time foraging. For example, species of shrew have to eat their own body weight of food each day to survive. Animals with such high

energy expenditure levels cannot afford to remain inactive or to sleep for prolonged periods: they must keep eating as often as possible. Indeed, owing to their high metabolic rates, some of these animals will not survive for more than a few hours if they are prevented from feeding.

When exposure to risk is examined, it appears that mammals living under dangerous or adverse conditions display highly polyphasic behavior. For instance, sleep is a particularly vulnerable state in the giraffe, an animal that takes 10 or more seconds to stand up. The giraffe lies down for periods between 3 and 75 minutes, three to eight times a night. Although most of this time is spent awake, it is assumed that the giraffe cumulates only 2 hours of (fragmented) sleep per 24 hours.

Polyphasic sleep naturally occurs in humans during certain situations or phases of the life span. Most typically, this is the natural sleep pattern during the first months of human life, where a 3- to 4-hour cycle in sleep–wake behavior is normally displayed. At the opposite end of the lifetime spectrum, nocturnal sleep is more fragmented in the elderly, and their daytime naps may become relatively frequent, even in healthy, alert individuals. Finally, it is well documented that adult humans do engage in more than one sleep episode per day when given the opportunity to do so.

If polyphasic sleep is then the primordial, ancestral pattern among animals, when in *Homo sapiens'* evolution did behavior become less polyphasic, and why did this happen? Anthropological studies suggest that one possible cause for change might have occurred as recently as circa 10 000 years ago, when the until then dominant hunter-gathering economy began to be gradually replaced by more settled and daytime-oriented agricultural societies. Support for this hypothesis comes from the partially polyphasic behaviors observed in a handful of hunter-gathering tribes that have survived isolated from modern civilization. Studies conducted on the Temiars and the Ibans of Malaysia indicate that their average nocturnal sleep is brief (4 to 6 hours) and that nighttime activities (fishing, cooking, watching over the fire, rituals) at any one time involve approximately 25 percent of the adult members. Daytime napping is widely practiced in both tribes: at almost any time of day, at least 10 percent of adults are asleep.

The next and most recent evolutionary milestone began with the Industrial Revolution and the invention of the light bulb. Having blurred the difference between day and night and dramatically extended the hours available to work, they eventually led to the so-called 24-hour society. Today, it is estimated that in industrialized societies, 20–25 percent of the workforce is involved in night- or shift-work, and this is indeed an important challenge to the practice of a monophasic, nocturnal sleep.

RHYTHMS OF SLEEP AND WAKEFULNESS

Are there any internal or external rhythms driving human sleep–wake and cognitive behaviors, in addition to the widely studied and prominent circadian (24-hour) light–dark cycle? Several have been described, and one of them is the circasemidian (12-hour) cycle. In humans and in most diurnal animals this is typically expressed by an increase in sleepiness in the early afternoon, and is also the explanation for the 'siesta' time observed in the many cultures that have grown to respect this biological drive.

The 90-minute ultradian REM/NREM sleep cycle represents the regular recurrence in sleep of REM (rapid eye movement) and non-REM sleep, two markedly different physiological states. The prominence of the REM/NREM cycle led to the speculation that this is the expression in sleep of a broader basic rest–activity cycle (BRAC), a never-ending beat which persists during both sleep *and* waking. A waking BRAC has been indeed observed for a variety of physiological and cognitive functions. In addition to the BRAC, a 4-hour ultradian periodicity has been observed in the multiple, spontaneous naps of subjects studied in isolation from time cues, in the rest–activity patterns of newborns, or in the pattern of involuntary daytime naps in narcoleptics.

The existence of the ultradian 4-hour cycle, and particularly of the BRAC, is one of the pillars supporting the theory that the ancestral sleep–wake pattern was polyphasic, which in turn helps to explain why humans appear to adapt to it. It is believed that one of the BRAC's functional values, or survival advantages, is to allow multiple transition opportunities, in the 24-hour continuum, from sleep to wake, or vice versa.

POLYPHASIC AND ULTRASHORT SLEEP RESEARCH: EFFECTS ON COGNITION AND PERFORMANCE

Is a polyphasic pattern a more effective solution when normal monophasic sleep is either disrupted, or especially when it is dramatically curtailed? Examples of possible applications range from the problems faced by parents attending the needs of

their newborns at night, to more extreme scenarios involving around-the-clock work, such as observed during emergency, survival, rescue or defense operations.

Two types of study have been conducted to answer these questions. One approach has been to *impose* a variety of polyphasic sleep scenarios in conditions of reduced sleep and then to compare the effects of such schedules to monophasic ones that allowed equivalent sleep time. A large number of field and laboratory sleep reduction studies conducted over the past three decades has shown that highly motivated individuals may be able to maintain acceptable levels of motor and cognitive performance if they sleep a minimum of 4.5 to 5.5 hours per day. Below that level, the effects of sleep deprivation invariably occur. For this reason, most studies comparing poly- versus monophasic sleep have chosen daily sleep quotas of 3 hours, thereby allowing the quantification of measurable decrements.

The most typical study design compared a polyphasic sleep schedule with 30 minutes of sleep every 4 hours to a monophasic sleep pattern allowing the same daily amount of sleep – 3 hours – in the middle of the night. Subjects had to follow each of these schedules under controlled laboratory environments for periods of about one month. These studies have shown that greater decrements in cognitive and psychomotor performance were observed when subjects adopted the monophasic pattern. That is, with only 3 hours' sleep per day, subjects under a polyphasic regimen performed significantly better than those with a monophasic pattern, albeit, as expected, both groups showed more errors than when performance was measured after 8 hours of normal nocturnal sleep.

SPONTANEOUS POLYPHASIC SLEEP DURING EXTREME PERFORMANCE

One of the first natural experiments with polyphasic sleep was apparently conducted as early as half a millennium ago. It has been claimed that Leonardo da Vinci would sleep 15 minutes out of every four hours, for a daily quota of only 1.5 hours' sleep per day. As one of the many unconfirmed yet intriguing enigmas surrounding Leonardo's life, if this were true it would certainly help explain the vastness of his productivity and legacy. More recently, the most compelling evidence of the potential effectiveness of polyphasic sleep came from studies of motivated individuals exposed to prolonged, extreme, and highly demanding around-the-clock work. When the fundamental need for sleep competes with an equally powerful call for staying awake 24 hours per day, will individuals spontaneously follow polyphasic patterns? If this is true, one may assume that they are adopting this behavior because it allows for a more effective waking performance.

Single-handed sailing races across the oceans offer the ideal field model to verify these hypotheses. Solo sailors are exposed for days and weeks to recurrent, and often extenuating demands to be awake, to adjust the yacht to varying conditions of sea and wind, to study complex weather data, and survey the tactics of competitors to obtain maximum speed. When asleep, autopilot yacht control may not be optimal, and risks of collision with other ships or obstacles (e.g., icebergs) increase. Motivation to win is extremely high and is incited by valuable prizes and sponsors' rewards.

Studies of over a hundred solo sailors conducted over the past 20 years, on races ranging from transatlantic (14 days minimum) to round-the-world non-stop (94 days), have confirmed that polyphasic sleep is the preferred strategy. Top competitors average 4.5 to 5.5 hours sleep per day, confirming the results of the sleep reduction studies mentioned earlier. Sailors' polyphasic behaviors showed some relationship with their circadian profile. Morning individuals tended to adopt shorter and more frequent naps, with up to 10 naps per day, averaging 25–30 minutes each. Evening-prone individuals, o the other hand, appeared to prefer fewer but longer sleep episodes. When the same sailors were studied in different races, as was the case with British Ellen MacArthur, the first and youngest woman to win the transatlantic race, it was found that their daily sleep amounts were reduced, and that sleep was more polyphasic, in the shorter races (transatlantic) compared to the longer ones (round-the-world).

CONCLUSION

Contrary to the common assumption that adult humans are dependent on a rigid, monophasic sleep–wake system, the issues briefly reviewed here point to the high level of plasticity of sleep regulatory mechanisms. Not only do adult humans adapt to polyphasic behaviors, but such strategies appear more advantageous during demanding conditions that involve dramatic sleep reduction. Such ability may represent the behavioral expression of the various underlying biological rhythms of sleep propensity. It also suggests that the recuperative value of sleep on performance and cognition is not linearly correlated with sleep duration.

Indeed, studies on polyphasic sleep have repeatedly indicated that short naps are responsible for remarkable recuperative effects, disproportionate to their duration. Rather, sleep recuperative power might be best represented by an exponential function, providing highest recuperative value at the beginning of sleep. As a result, under a restricted sleep regime, the strategy of choice is to recharge the sleep 'batteries' more often with short sleep episodes, and preferably prior to the occurrence of cognitive decrements.

Further Reading

Dinges DF and Broughton RJ (eds) (1989) *Sleep and Alertness*: New York, NY: Raven Press.

Stampi C (1989) Polyphasic sleep strategies improve prolonged sustained performance: a field study on 99 sailors. *Work & Stress* **3**: 41–45.

Stampi C (ed.) (1992) *Why We Nap: Evolution, Chronobiology, and Functions of Polyphasic and Ultrashort Sleep*. Boston, MA: Birkhauser.

Stampi C (1994) Sleep and circadian rhythms in space. *Journal of Clinical Pharmachology* **34**: 518–534.

Sleep Reduction, Cognitive Effects of

Introductory article

John PJ Pinel, University of British Columbia, Vancouver, British Columbia, Canada
Steven J Barnes, University of British Columbia, Vancouver, British Columbia, Canada

CONTENTS

Even minor sleep loss (e.g., 3 or 4 hours for one night) increases sleepiness, disturbs mood, and reduces vigilance. More complex cognitive and motor disturbances have been reported after sleep loss, but they have been proven to be less sensitive. Remarkably, the few studies of individuals who have adopted long programs of reduced sleep (e.g., 5.5 hours per night) suggest that their sleep gradually becomes more efficient, and their initial deficits diminish.

INTRODUCTION

Sleeping for fewer hours than one would ideally choose is commonly referred to as *sleep loss*. Identifying the adverse effects of sleep loss has been a major focus of research because it may provide insights into the function of sleep. The fact that all mammals and their evolutionary relatives spend a substantial proportion of their time sleeping suggests that sleep plays a critical role in maintaining normal function, but this role remains a mystery.

There are also practical reasons for determining the effects of sleep loss. Many individuals suffer from chronic sleep loss (e.g., medical interns, soldiers, mothers of young babies, or those suffering from sleep disorders), and it is important to understand their functional disabilities in order to counsel or treat them, or to design their environments in such a way as to limit the risk of mishaps.

THE NATURE OF SLEEP AND DIFFICULTIES IN ASSESSING SLEEP LOSS

Identifying the disruptive effects of sleep loss appears to be a simple matter. Virtually everyone has experienced the effects of sleep loss, and the debilitative effects seem to be compelling and obvious. However, there are several factors that complicate their identification.

One complicating factor is the fact that sleep is not a unitary phenomenon. Although it is commonly regarded as a single entity, it actually comprises distinct phases which may serve different functions. When one first falls asleep during a normal night's sleep, the cortical electroencephalogram (EEG) is characterized by a predominance of low-amplitude, short-duration waves. Then, as sleep ensues, the predominant waves gradually increase in amplitude and duration. For the purposes of analysis, this gradual progression from smallest to largest sleep EEG waves is artificially divided into four distinct phases, with the stage 1 sleep EEG being characterized by the highest density of the smallest, fastest sleep waves (theta waves), and the stage 4 sleep EEG being characterized by the highest density of the largest, slowest waves (delta waves). Although the average amplitude and duration of sleep EEG waves are the usual means of distinguishing between the various sleep stages, other differences exist. For example, sleep spindles (intermittent waxing and waning 1- to 2-second bursts of 12- to 14-Hz waves) tend to be restricted to stage 2 sleep.

Once a sleeping individual has progressed through sleep stages 1 to 4, the progression is reversed back through the stages. An entire cycle from stages 1 to 4 and back to 1 typically lasts about 90 minutes. Then the sleeper spends the rest of the night going back and forth through the sleep stages in a pendulum-like manner. However, an important change occurs as a typical night of sleep unfolds. The first half of a night's sleep contains much more slow-wave sleep, whereas the second half contains much more stage 1 and stage 2 sleep.

The initial stage 1 EEG episode is uneventful, but thereafter each period of stage 1 EEG tends to be accompanied by a variety of physiological changes, such as increased pulse and blood pressure, extreme relaxation of the muscles of the body core, irregularity of breathing, and a pattern of rapid eye movements (REMs). Accordingly, periods of stage 1 EEG sleep after the initial episode are typically referred to as *REM sleep*. The EEG of REM stage 1 differs slightly from the initial stage 1 EEG in several respects, the most obvious being the occasional appearance of bursts of slightly larger, slower, sharply pointed waves, which are termed *sawtooth waves* because of their appearance.

REM sleep is of particular interest because of its association with dreaming. When sleeping subjects are awakened during REM sleep, they tend to report that they have been dreaming, and they can recount the dream. However, although the relationship between REM and dreaming has been well documented, the association is not all or nothing in nature. On some occasions subjects who are awakened during REM sleep do not report dreams. On the other hand, on some occasions subjects who are awakened during slow-wave sleep (SWS) do report dreams. However, the psychological experiences associated with REM sleep tend to have more of a narrative quality than those that are associated with SWS. Because of the relationship between dreaming and REM, the fact that the proportion of SWS decreases and the proportion of REM sleep increases as the night progresses suggests that more dreaming occurs late in a night's sleep. In the present context, it is important to bear in mind that any detailed understanding of the effects of sleep loss must address the fact that SWS and REM sleep appear to serve different functions.

The following four factors also tend to complicate any interpretation of the effects of sleep loss. First, many situations in which there has been sleep loss also involve disturbances of circadian cycles, which themselves can produce physiological and behavioral disturbances. Secondly, when sleep loss occurs in natural situations, it has often been caused by stressful events (e.g., worry or an excessive workload), which themselves could have deleterious functional effects. Thirdly, there is often a surprising dissociation between the subjective feelings of sleep-deprived individuals and their actual abilities. Fourthly, the expectations of sleep-deprived individuals need to be taken into account.

Because of these problems with regard to interpreting the effects of sleep loss, efforts to identify those effects have relied on the results of controlled experiments. There are two fundamentally different types of sleep loss experiments (which will be discussed separately in the following two sections), namely short-term studies, which typically last for several days or less, and long-term studies, which typically last for several weeks or more.

EFFECTS OF SHORT-TERM SLEEP REDUCTION

Studies of short-term sleep reduction (those that involve sleep-reduction schedules of several days or less) can be divided into two distinct subcategories, namely those in which the deprivation is total (e.g., no sleep for 3 days) and those in which it is partial (e.g., 4 hours of sleep per night for 3 days). One of the early studies of short-term total sleep reduction was the classic 1922 study by Kleitman, who found that students were capable of staying awake continuously for several days with no major performance deficits that could not be attributed to their overwhelming tendency to fall asleep. Remarkably, the sleepiness of the participants did not develop monotonically. For the first 3 days they were much more sleepy during the night than during the day, with each successive day becoming more problematic than the one before. However, by the end of the third sleepless day, the circadian cycle of sleepiness became no more severe.

Unlike Kleitman's study, which was largely anecdotal, modern studies of sleep reduction have focused on the objective assessment of its effects. Effects of four different types have been widely investigated, namely physiological effects, deficits in motor performance, mood changes, and cognitive deficits. The implicit premise on which many of these studies have been based is that sleep serves a critical, albeit unknown survival function, and that it is maintained at a homeostatic level by unknown regulator mechanisms. The goal of these studies has been to document the adverse effects of disturbing this homeostasis. The fact that increases in negative mood and decreases in the time needed to fall asleep reliably occur after as little as 3 or 4 hours of sleep loss for one night is consistent with the idea that sleep is homeostatically regulated. However, the inconsistency of other effects of short-term sleep reduction has been less compatible with this view. For example, it has proven difficult to demonstrate adverse disturbances of physiological function or motor performance produced by sleep loss that are reliable from one study to another. For instance, various studies of up to 72 hours of sleep deprivation have not found reliable changes in muscle strength or

in cardiovascular or respiratory responses to heavy exercise. Although sleep-deprived individuals do feel that exercise requires more exertion, and they report exhaustion more quickly, these effects have not been confirmed by cardiovascular measures.

The effects of short-term sleep deprivation on cognitive abilities have been puzzling. Performance on complex demanding cognitive tests, which one would expect to be most sensitive to sleep loss, is sometimes unaffected. In contrast, simple tests of vigilance are consistently sensitive to sleep loss. For example, in one common test, sleep-deprived individuals listen to a series of tones and respond to those tones that are slightly shorter than the others. Loss of 3 or 4 hours of sleep on one night is sufficient to produce reliable deficits on this type of task, particularly if the task is lengthy.

It has been suggested that the cognitive deficits caused by sleep reduction are entirely attributable to microsleeps (3- or 4-second periods of sleep that commonly occur in sleep-deprived individuals, sometimes even when they are standing or sitting). According to this view, sleep-deprived people experience a powerful drive to sleep, which disrupts their performance, but while they are awake they perform normally. However, a careful analysis of the performance of sleep-deprived individuals does not support this interpretation. If it were correct, performance would be normal with occasional lapses in performance during the microsleeps, but this pattern has not been observed. For example, sleep-deprived people perform poorly on simple vigilance tasks even when they are not experiencing microsleeps; and on more complex cognitive tasks, sleep-deprived subjects often respond inappropriately, rather than just displaying occasional lapses of performance.

In 1997, Dinges and colleagues conducted a particularly thorough study of short-term sleep deprivation. The sleep of the participants was reduced to 5 hours per night for 7 days. As a result, they displayed deficits on several tests of sleepiness, mood, vigilance, and cognitive function. Of most interest was the finding that there was a gradual increase in the effects of sleep loss over the 7-day period, indicating that the effects of several nights of sleep reduction can accrue, at least in the short term.

In summary, studies of periods of partial or total sleep reduction lasting for several days or less have consistently shown that the loss of as little as 3 or 4 hours of sleep for one night increases sleepiness, disturbs mood, and decreases vigilance. In contrast, the performance of more challenging cognitive and sensorimotor tasks has proved to be less sensitive to sleep loss – deficits have tended to require more extreme sleep loss, and have been less consistently reported. In addition, various physiological changes have been reported following sleep loss (e.g., in the pattern of hormone release), but so far none of them has proved to be sufficiently large or reliable to suggest why sleep that is composed of both REM sleep and SWS has evolved in all mammals and birds.

EFFECTS OF LONG-TERM SLEEP REDUCTION

A second type of study of sleep reduction has focused on the effects of long-term programs of reduced sleep (e.g., 5.5 hours per night for 6 months). One method that has been used in these studies is to have the participants reduce their sleep gradually in small (e.g., 30-minute) increments until they reach the target level. These studies are particularly significant because they focus on an important point about sleep that is of major theoretical significance and is not addressed by studies of short-term sleep loss. This point is that once individuals begin to sleep less on a regular basis, their sleep seems to become progressively more efficient.

Several lines of evidence indicate that SWS is largely responsible for the restorative power of sleep in humans. For example, people who consistently sleep less tend to have sleep with a higher proportion of SWS, and they often get as much SWS as those who sleep much more. Moreover, following several nights of total sleep deprivation, there is a major increase in the proportion of SWS for the first few recovery nights, even though there is only a minor increase in total sleep.

What happens when people are placed on long-term restricted-sleep schedules? In one study, the eight participants gradually reduced their sleep time in 30-minute increments until they experienced prohibitive chronic fatigue. Two individuals reduced their sleep to 5.5 hours, four to 5 hours, and two to 4.5 hours. The participants were then required to sleep at their lowest level plus 30 minutes for the remainder of the experiment, which lasted 6 months. As the participants reduced their sleep, its efficiency increased. The proportion of SWS increased, the time needed to fall asleep decreased, and the number of night time awakenings decreased. A large battery of medical, personality, performance, mood, sleepiness, and cognitive tests (including a test of vigilance) failed to reveal any deficits. Indeed, when the participants were

revisited 12 months after the end of the study, all of them were sleeping 1 to 2.5 hours less per night than they had been doing before the study.

In a similar study, participants reduced their amount of sleep to 5.5 hours per night in one step, and then slept at that level for 2 months. They initially experienced daytime sleepiness and difficulty awakening from sleep in the morning, but as the efficiency of their sleep increased, these adverse effects disappeared. Again, a large battery of tests revealed no clear cognitive, performance, or mood deficits.

One currently active line of research on long-term sleep reduction involves polyphasic sleep schedules, in which the sleep of the participants is limited to several scheduled naps in each 24-hour period (e.g., 30 minutes every 4 hours). Naps have proved to be particularly effective in reducing the adverse effects of sleep loss. Consequently, individuals who learn to take all of their sleep in the form of naps can reduce their sleep to remarkably low levels in the absence of obvious dysfunction. (*See* **Sleep: Polyphasic**)

Despite the consistent results of studies of long-term sleep reduction and their provocative implications, these studies have one major weakness. They are rare, presumably because their duration is prohibitive for most researchers. Still, it seems likely that ultimate conclusions about the minimum requirements of sleep for effective cognitive functioning will of necessity rest on the results of these types of studies.

CONCLUSION

In the short term, even a few hours of sleep loss disturbs mood, increases sleepiness, and diminishes vigilance, and more extensive sleep loss can produce a variety of motor and cognitive deficits. It is important to appreciate the nature of these deficits because the demands of modern life impose short-term sleep loss on many people. Although there have been far fewer studies of long-term sleep reduction, their findings have been consistent and provocative. Most participants in long-term sleep reduction experiments have been successful in reducing their sleep to 5.5 hours per night or less. Initially they experience fatigue and other expected adverse effects, but if they persist, the efficiency of their sleep increases, and the obvious adverse effects of their new sleep regimen diminish.

Further Reading

Dinges DF, Pack F, Williams K *et al*. (1997) Cumulative sleepiness, mood disturbance and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep* **20**: 267–277.

Horne JA (1983) Mammalian sleep function with particular reference to man. In: Mayes AR (ed.) *Sleep Mechanisms and Functions in Human and Animals*, pp. 262–312. Wokingham, UK: Van Nostrand Reinhold.

Jewett ME, Dijk D-J, Kronauer RE and Dinges DF (1999) Dose–response relationship between sleep duration and human psychomotor vigilance and subjective alertness. *Sleep* **22**: 171–179.

Martin BJ (1986) Sleep deprivation and exercise. In: Pandolf KB (ed.) *Exercise and Sports Sciences Reviews*, pp. 213–229. New York, NY: Macmillan.

Monk TH (1991) *Sleep, Sleepiness and Performance*. New York, NY: John Wiley.

Pilcher JJ and Huffcut AJ (1996) Effects of sleep deprivation on performance: a meta-analysis. *Sleep* **19**: 318–326.

Pinel JPJ (2003) *Biopsychology*, 5th edn. Boston, MA: Allyn & Bacon.

Stampi C (1992) *Why we Nap: Evolution, Chronobiology and Functions of Polyphasic and Ultrashort Sleep*. Boston, MA: Birkhauser.

Van Cauter E and Copinschi G (2000) Interrelationships between growth hormone and sleep. *Growth Hormone IGF Research* **10**(Supplement B): S57–62.

VanHelder T and Radomski MW (1989) Sleep deprivation and the effect on exercise performance. *Sports Medicine* **7**: 235–247.

Wesensten NJ, Balkin TJ and Belenky G (2000) Does sleep fragmentation impact on recuperation? A review and re-analysis. *Journal of Sleep Research* **8**: 237–245.

Somesthesis, Neural Basis of

Intermediate article

Steven S Hsiao, Johns Hopkins University, Baltimore, Maryland, USA
Takashi Yoshioka, Johns Hopkins University, Baltimore, Maryland, USA
Kenneth O Johnson, Johns Hopkins University, Baltimore, Maryland, USA

CONTENTS

Somesthesis is the awareness of sensation from the skin, muscles, joints, and viscera, through specialized receptors in the skin and deep tissues.

INTRODUCTION

Somesthesis is the awareness of sensations from the skin, muscles, joints, and viscera. Its neural basis is the specialized receptors located in the skin and deep tissues that are selectively sensitive to mechanical stimuli of various kinds (e.g., pressure, vibration, body position, and movement), changes in skin temperature (warmth and cold), chemical stimuli of certain kinds (e.g., histamine, which activates itch receptors) and noxious stimuli. In order to convey the specificity of somesthesis, this discussion concentrates on the neural basis of discriminative tactile sensation from the human hand.

The human hand is a complex organ that is used to explore the world around us and to manipulate objects. We do so both directly and indirectly. In direct exploration and manipulation, information about surface features, object size, shape, and weight are communicated to the brain by specialized mechanoreceptive and proprioceptive receptors in the skin and muscles. The object's surface features and texture are signaled by one group of mechanoreceptors. Its shape is signaled by these mechanoreceptors together with proprioreceptors in the muscles and deep tissues. The proprioceptors also signal object weight, hand conformation, and finger location. Another group of mechanoreceptors in the skin provides feedback signals for grip control. Two groups of thermoreceptors signal object temperature: one group is selectively sensitive to warming, the other to cooling. In indirect exploration and manipulation we use tools or probes as extensions of our hands. Even though a probe or tool comes between our hands and the object of interest, we commonly perceive the object as though our fingers were present at the working surfaces of the tool or probe. This perception of distant events depends on (1) specialized mechanoreceptors in the hand which respond to minute vibrations created by interactions between the tool and the object being probed, and (2) other receptors that respond to reaction forces transmitted to the hand from the object.

Perception from the hand begins with the activation of 13 kinds of afferent fibers that receive information from receptors embedded in the skin and deep tissues. These afferents respond with a high degree of specificity to different kinds of internal and external stimuli. In everyday life all of the receptors (except the pain and itch receptors) are continuously active and send a constantly changing dynamic representation of the external world and the internal state of our tissues to the brain.

Four of the 13 afferent fiber types provide information about discriminative touch on the glabrous skin, and evidence discussed below suggests that these four kinds of mechanoreceptive afferents serve different sensory functions (for a detailed review see Johnson 2002). The slowly adapting type 1 (SA1) system, which terminates in Merkel receptors, provides the basis for the perception of local features (edges, corners, points) and the perception of surface form and texture. The rapidly adapting (RA) system, which terminates in Meissner corpuscles, provides the basis for the perception of low-frequency vibrations and provides feedback signals required for grip control. The Pacinian corpuscle (PC) system, which terminates in Pacinian corpuscles, provides the basis for the

perception of events distant from the hand signaled by minute vibrations. The slowly adapting type 2 (SA2) system provides information about hand conformation and about shear forces acting parallel to the skin surface (Edin, 1992).

SPATIAL DISCRIMINATION THRESHOLDS

There is much evidence that the threshold for spatial acuity is determined by the SA1 system. Three different psychophysical tasks – a gap detection task, a grating orientation task and a letter identification task – demonstrate that the threshold (element spacing that produces performance midway between chance and perfect discrimination) at the fingertip is about 1.0 mm (Figure 1; Johnson and Phillips, 1981). Acuity is not uniform across the body surface. It is highest at the lip and tongue (threshold about 0.5 mm; Van Boven and Johnson, 1994), lowest around the calf (about 30 mm; Stevens *et al*., 1996) and steadily decreases with age (threshold at the fingertip is about 3.0 mm for 80-year-old subjects; Stevens *et al*., 1996). Anatomical studies show that the performance at the fingertips is close to the theoretical limits imposed by the innervation densities of either the SA1 or RA and that it far exceeds the spatial capacity of the PC and SA2 afferent fibers. There are approximately 100 SA1 and 150 RA afferents per square centimeter (Johnson, 2002). These innervation densities correspond to mean spacings between individual fibers of about 1.0 mm and 0.82 mm respectively, which is similar to the limit of tactile spatial resolution at the fingertip. Evidence that acuity is determined by the SA1 system alone comes from neurophysiological studies in monkeys and humans demonstrating that only the SA1 afferent fibers are able to resolve spatial patterns with spatial feature separated by about 1.0 mm (Phillips and Johnson, 1981; Phillips *et al*., 1992).

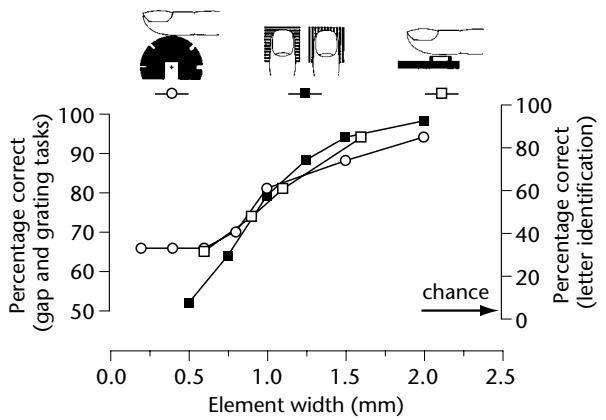


Figure 1. Human performance in gap detection (open circles), grating orientation discrimination (filled squares), and letter recognition (open squares). The abscissa represents the fundamental element width for each task: gap size for the gap detection task, bar width (half the grating period) for the grating orientation discrimination task, and the average bar and gap width within letters (approximately one-fifth of the letter height) for the letter recognition task. Threshold is defined as the element size producing performance midway between chance (50% correct for the gap and grating tasks, 1/26 for letter recognition) and perfect performance. Adapted from Johnson and Phillips (1981).

DISCRIMINATION OF COMPLEX SPATIAL PATTERNS

The perception of complex spatial patterns (e.g., Braille or embossed letters) is robust. It is relatively unaffected by contact force (over the range 16–128 grams; Loomis, 1985), by stimulus height (over the range 0.24–1.92 mm; Loomis, 1985), by scanning velocity (over the range 20–80 mm s^{-1}; Vega-Bermudez *et al*., 1991), by stimulus orientation (medial–lateral discrimination is the same as proximal–distal; Craig, 1999), whether patterns are scanned actively or scanned across the passively restrained hand (Vega-Bermudez *et al*., 1991), and whether the pattern is scanned or indented into the skin without horizontal movement (Phillips *et al*., 1983).

Humans discriminate surface curvature well. When a single finger pad is used, the threshold for discriminating a flat surface from a concave or convex surface is about 5 m^{-1} (radius 200 mm; Goodwin *et al*., 1991). People also discriminate very small changes in curvature accurately (e.g. 144 m^{-1} versus 158 m^{-1}). Goodwin and his colleagues have shown that only the SA1 mechanoreceptors can account for this perceptual capacity (Goodwin *et al*., 1995). When several fingers are used, the discrimination thresholds for flat surfaces decrease to around 0.5 m^{-1}; i.e. using three fingers we can distinguish a flat surface from one with a 2 m radius (Pont *et al*., 1997).

Slowly adapting type 1 afferents are sensitive to negative as well as positive surface curvature and provide the central nervous system with an isomorphic representation of the local two-dimensional and three-dimensional structure of stimuli on the skin (Figure 2). The neural responses of these afferents are only moderately affected by changes in contact force and scanning velocity (see Figure 2).

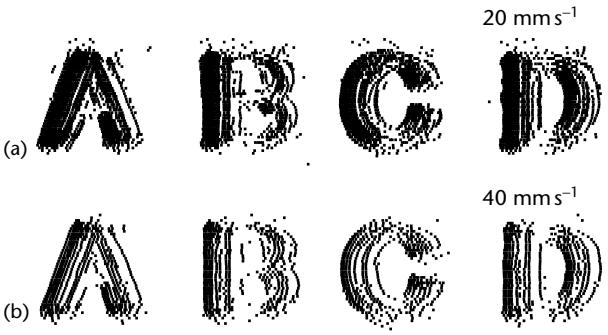


Figure 2. Spatial event plots from a representative slowly adapting type 1 (SA1) afferent fiber. Each tick represents the occurrence of an action potential evoked by the embossed letters A, B, and C (6.0 mm high) which were repeatedly scanned across the neuron's receptive field. After each scan, the embossed letter was shifted 200 μm in the vertical direction and the scan repeated. The plots show the responses after about 70 scans of the letter across the receptive field at two different scanning velocities – (a) 20 mm s^{-1} and (b) 40 mm s^{-1} – demonstrating that peripheral SA1 afferents provide an isomorphic representation of the letter (independent of scan velocity) to the central nervous system.

TEXTURE PERCEPTION

Multidimensional scaling studies show that texture perception is composed of two main dimensions (smooth–rough and soft–hard) and, possibly, a weaker third dimension that is not well understood (Hollins *et al*., 1993). Roughness, like form perception, is only mildly affected by contact force and scanning velocity. Studies using raised gratings as stimuli show that roughness increases as groove width increases and decreases as ridge width increases (Lederman, 1983). Studies using raised dot patterns show that roughness perception increases as the dot spacings increase to about 3 mm and decreases as spacings increase beyond 3 mm (Connor *et al*., 1990). Neurophysiological studies demonstrate that roughness, like form, is conveyed by the SA1 afferents, and perceived roughness is based on the variation in firing rates between groups of SA1 afferents (Yoshioka *et al*., 2001).

Experimentally, people are good at discriminating subtle differences in the softness of objects with deformable surfaces with and without active touch (Srinivasan and LaMotte, 1996). The essence of softness is progressive conformation to the contours of the fingers and hand in proportion to contact force. The degree of softness is signaled by the rate of growth of contact area with contact force and by the uniformity of pressure across the contact area. Conversely, the essence of hardness is invariance of object form with changes in contact force (Johnson, 2002). The ability of SA1 afferents to signal the spatial pattern of skin indentation suggests that they are responsible for softness perception (for a discussion see Johnson, 2002).

PERCEPTION OF MOTION AND VIBRATION

Rapidly adapting afferents are responsible for perceptual functions related to the detection of minute skin motion. This hypothesis is supported by two lines of evidence. The first is the demonstration that only RA afferents account for the detection of low-frequency vibrations (threshold of about 10 μm at 30 Hz; Mountcastle *et al*., 1972). The second is the demonstration that RA afferents are responsible for the detection of minute slips that occur between the object and skin when an object is lifted (Johansson and Westling, 1984). Rapid slip detection is needed for the minute adjustments in grip forces necessary to hold objects securely.

Numerous studies demonstrate that only PC afferents account for the detection of high-frequency vibrations (Mountcastle *et al*., 1972; Brisben *et al*., 1999). Humans can detect vibrations as small as 10 nm at 200 Hz and this is accounted for by the responses of the most sensitive PC afferents (Brisben *et al*., 1999). Studies of the detection of complex vibratory stimuli show that humans are sensitive to the temporal structure of high-frequency stimuli that activate only PC afferents (Formby *et al*., 1992), suggesting that PC afferents are responsible for the ability to perceive vibratory events at the working surfaces of tools held in the hand (Johnson *et al*., 2000).

ROLE OF ATTENTION IN HAPTIC PERCEPTION

Attention plays a critical role in haptic perception (Hsiao and Vega-Bermudez, 2001). Try consciously shifting your attention from one body location to another: notice how the other body locations disappear from your consciousness and how the location attended to dominates it. The degree to which attention can be focused is addressed in experiments in which the participants' ability to ignore distracting stimuli is tested. People asked to discriminate a vibratory stimulus presented to one finger (which depends on information conveyed by the RA system) can ignore distracting stimuli applied to the opposite hand, but not stimuli applied to other fingers on the same hand (Craig, 1985). However, a more recent study has shown that people

can make texture judgments (which depend on the SA1 system) with one finger and completely ignore distracting textures applied to adjacent fingers (Dorsch *et al*., 2001). This suggests that the ability to focus attention may differ between receptor systems – that we can focus attention narrowly when we rely on the system responsible for spatial information (the SA1 system) but not when we rely on the system responsible for the detection of motion on the skin (the RA system).

The psychophysical studies of attention are paralleled by neurophysiological studies in nonhuman primates and imaging studies in humans showing that the responses of neurons in both primary and secondary somatosensory cortex are modified by the attentional state (Hsiao *et al*., 1993; Burton and Sinclair, 2000). The modifications are of two kinds: one is a change in the firing rates of neurons processing information from the body part being attended to; the other is a change in the temporal firing patterns of the action potentials in these same neurons so they become more or less synchronous between neurons. The effect of enhanced firing rates and increased synchrony is enhancement of the saliency of the message conveyed by those neurons (Steinmetz *et al*., 2000).

CONCLUSION

Evidence from psychophysical and neurophysiological research indicates that the different aspects of tactile perception are based on different afferent fiber types. As in the visual system, one afferent system (SA1) appears to be specialized for form and texture perception and another (the RA system) for the detection of motion (Hsiao, 1998). The PC system, like neurons in the auditory system, is specialized to perceive vibratory events at a distance. Research is needed to determine whether modality specificity is preserved in the central processing pathways, and to identify the role and mechanisms of selective attention.

References

Brisben AJ, Hsiao SS and Johnson KO (1999) Detection of vibration transmitted through an object grasped in the hand. *Journal of Neurophysiology* **81**: 1548–1558.

Burton H and Sinclair RJ (2000) Attending to and remembering tactile stimuli: a review of brain imaging data and single-neuron responses. *Journal of Clinical Neurophysiology* **17**: 575–591.

Connor CE, Hsiao SS, Phillips JR and Johnson KO (1990) Tactile roughness: neural codes that account for psychophysical magnitude estimates. *Journal of Neuroscience* **10**: 3823–3836.

Craig JC (1985) Attending to two fingers: two hands are better than one. *Perception and Psychophysics* **38**: 496–511.

Craig JC (1999) Grating orientation as a measure of tactile spatial acuity. *Somatosensory and Motor Research* **16**: 197–206.

Dorsch AK, Hsiao SS, Johnson KO and Yoshioka T (2001) Tactile attention: subjective magnitude estimates of roughness using one or two fingers. *Society for Neuroscience Abstracts* **27**: 50–52.

Edin BB (1992) Quantitative analysis of static strain sensitivity in human mechanoreceptors from hairy skin. *Journal of Neurophysiology* **67**: 1105–1113.

Formby C, Morgan LN, Forrest TG and Raney JJ (1992) The role of frequency selectivity in measures of auditory and vibrotactile temporal resolution. *Journal of the Acoustic Society of America* **91**: 293–305.

Goodwin AW, John KT and Marceglia AH (1991) Tactile discrimination of curvature by humans using only cutaneous information from the fingerpads. *Experimental Brain Research* **86**: 663–672.

Goodwin AW, Browning AS and Wheat HE (1995) Representation of curved surfaces in responses of mechanoreceptive afferent fibers innervating the monkey's fingerpad. *Journal of Neuroscience* **15**: 798–810.

Hollins M, Faldowski R, Rao S and Young F (1993) Perceptual dimensions of tactile surface texture: a multidimensional-scaling analysis. *Perception and Psychophysics* **54**: 697–705.

Hsiao SS (1998) Similarities between touch and vision. In: Morley JW (ed.) *Neural Aspects of Tactile Sensation*, pp. 131–165. Amsterdam, Netherlands: Elsevier.

Hsiao SS and Vega-Bermudez F (2001) Attention in the somatosensory system. In: Nelson RJ (ed.) *The Somatosensory System: Deciphering the Brain's Own Body Image*, pp. 197–217. Boca Raton, FL: CRC Press.

Hsiao SS, O'Shaughnessy DM and Johnson KO (1993) Effects of selective attention of spatial form processing in monkey primary and secondary somatosensory cortex. *Journal of Neurophysiology* **70**: 444–447.

Johansson RS and Vallbo AB (1976) Skin mechanoreceptors in the human hand: an inference of some population properties. In: Zotterman Y (ed.) *Sensory Functions of the Skin in Primates*, pp. 171–184. Oxford, UK: Pergamon Press.

Johansson RS and Westling G (1984) Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research* **56**: 550–564.

Johnson KO (2002) Neural basis of haptic perception. In: Pashler H and Yantis S (eds) *Stevens Handbook of Experimental Psychology*, 3rd edn, vol. 1, *Sensation and Perception*, pp. 537–583. New York: Wiley.

Johnson KO and Phillips JR (1981) Tactile spatial resolution: I. Two-point discrimination, gap detection, grating resolution, and letter recognition. *Journal of Neurophysiology* **46**: 1177–1191.

Johnson KO, Yoshioka T and Vega-Bermudez F (2000) Tactile functions of mechanoreceptive afferents innervating the hand. *Journal of Clinical Neurophysiology* **17**: 539–558.

Lederman SJ (1983) Tactual roughness perception: spatial and temporal determinants. *Canadian Journal of Psychology* **37**: 498–511.

Loomis JM (1985) Tactile recognition of raised characters: a parametric study. *Bulletin of the Psychonomic Society* **23**: 18–20.

Mountcastle VB, LaMotte RH and Carli G (1972) Detection thresholds for stimuli in humans and monkeys: comparison with threshold events in mechanoreceptive afferent nerve fibers innervating the monkey hand. *Journal of Neurophysiology* **35**: 122–136.

Phillips JR and Johnson KO (1981) Tactile spatial resolution: II. Neural representation of bars, edges, and gratings in monkey primary afferents. *Journal of Neurophysiology* **46**: 1192–1203.

Phillips JR, Johnson KO and Browne HM (1983) A comparison of visual and two modes of tactual letter resolution. *Perception and Psychophysics* **34**: 243–249.

Phillips JR, Johansson RS and Johnson KO (1992) Responses of human mechanoreceptive afferents to embossed dot arrays scanned across fingerpad skin. *Journal of Neuroscience* **12**: 827–839.

Pont SC, Kappers AML and Koenderink JJ (1997) Haptic curvature discrimination at several regions of the hand. *Perception and Psychophysics* **59**: 1225–1240.

Srinivasan MA and LaMotte RH (1996) Tactual discrimination of softness: abilities and mechanisms. In: Franzen O, Johansson RS and Terenius L (eds) *Somesthesis and the Neurobiology of the Somatosensory Cortex*, pp. 123–135. Basel, Switzerland: Birkhauser.

Steinmetz PN, Roy A, Fitzgerald PJ *et al*. (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* **404**: 187–190.

Stevens JC, Foulke E and Patterson MQ (1996) Tactile acuity, aging, and braille reading in long-term blindness. *Journal of Experimental Psychology* **2**: 91–106.

Van Boven RW and Johnson KO (1994) The limit of tactile spatial resolution in humans: Grating orientation discrimination at the lip, tongue and finger. *Neurology* **44**: 2361–2366.

Vega-Bermudez F, Johnson KO and Hsiao SS (1991) Human tactile pattern recognition: active versus passive touch, velocity effects, and patterns of confusion. *Journal of Neurophysiology* **65**: 531–546.

Yoshioka T, Gibb B, Dorsch AK, Hsiao SS and Johnson KO (2001) Neural coding mechanisms underlying perceived roughness of finely textured surfaces. *Journal of Neuroscience* **21**: 6905–6916.

Spatial Attention, Neural Basis of

Introductory article

Michael S Worden, Weill Medical College of Cornell University, New York, New York, USA
Antigona Martinez, Weill Medical College of Cornell University, New York, New York, USA
Michael I Posner, Weill Medical College of Cornell University, New York, New York, USA

CONTENTS

The neural signals responsible for the control of spatial attention are presumed to stem from activity in specific cortical and subcortical networks and are transmitted via feedback projections to sensory cortex where they exert their modulatory effects.

INTRODUCTION

Suppose you are looking for a folder on the desktop of your computer. You might shift your eyes to different locations or, if the folders are close together, merely shift attention without moving your eyes while looking for the target. When you find a likely candidate you might examine the label carefully to make sure it is correct. When you move your eyes between locations or shift attention between words on the label, you employ a specific network of neural areas that operate to enhance the priority given to information from the selected location in comparison to other places in the visual field. While the acuity of the visual system is relatively fixed by the wiring between retina and primary visual cortex, the priority can be shifted at will to different places in the visual field. Once attention has moved to a new location, the location that had been previously attended appears to be slightly inhibited (inhibition of return), encouraging search of novel areas.

Experiments involving relatively empty visual fields often cue attention to a location by providing a marker at the location where the target might arrive, or by a central or symbolic cue informing the person of the target location. Performance in processing the target can then be compared with performance in the absence of cues providing information about target location. The improvement in reaction time (RT) and detection thresholds with a valid cue over a neutral condition is often slight; however, if the cue indicates a wrong location, the cost in performance is much larger. This illustrates the fact that, if not already engaged, attention can be rapidly summoned to the presentation of a target by luminance, motion or other cues that occur when a target is presented. For this reason a cue prior to the target produces only a small benefit. However, if attention is already engaged in processing a visual stimulus, costs due to disengaging and moving attention to the new location can be considerable.

When the visual field is cluttered with many objects, attention can be moved from location to location in a systematic fashion, or it may be guided by knowledge of the target properties (guided search). In addition, a new target can efficiently summon attention by luminance or motion cues. However, if these cues are artificially removed, the critical role of attention becomes clear. Powerful alterations of the meaning of objects in the field can be made without the viewer being aware of them as long as they are outside the current focus of attention. Thus the strong belief that we are attending to the entire visual field is not a correct one. In fact our attention is limited, but its ability to shift to a target that provides proper cues is impressive.

Visual spatial attention is closely related to the process of making eye movements. Indeed, one of the primary functions of spatial attention may be to identify locations in the visual field that contain pertinent information and to help the oculomotor system guide the eyes to those locations. Saccadic eye movements are preceded by an increase in spatial attention at the location to which the gaze will move. Although the exact correspondence between the network of brain areas responsible for visual spatial attention and that

responsible for planning and executing eye movements remains open to debate, it is clear from recent neuroimaging studies that there is considerable overlap between them.

Visual spatial attention is only one network of the attention system. However, because of its close relation to eye movements and the similarity of its operation between humans and other primates, the neural systems and pathways involved have been better analyzed than for other networks and it serves as a model for linking attention to neural systems.

PATHWAYS OF ORIENTING

Directing spatial attention to a location produces a modulation of the processing of sensory information. These are the sites on which attention operates and they can be distinguished from the sources that guide and control the spatial allocation of attention. The neural signals responsible for the control of spatial attention are presumed to stem from activity in higher cortical areas and are transmitted via feedback projections to sensory cortex where they exert their modulatory effects.

At a simple level, orienting can be viewed in terms of more elementary operations (such as disengage, move and engage) involved in the allocation of attention to locations or objects. Each operation is thought to be mediated by activity in different cortical or subcortical regions. The first two operations, disengaging and moving attention from the current location or object, involve the functioning of the parietal cortex together with the superior colliculus. Engaging attention at the relevant object or location seems to involve the pulvinar nucleus of the thalamus. The view that attentional control is mediated by a distributed network of brain areas is consistent with studies of people with brain damage who show impairments in their ability to direct attention spatially following localized lesions to a variety of brain regions. Different forms of deficit have been described in patients with damage to the temporoparietal junction, portions of frontal cortex, the anterior cingulate cortex, the basal ganglia, and the thalamus (in particular the pulvinar nucleus).

In recent years functional neuroimaging studies, using both position emission tomography (PET) and functional magnetic resonance imaging (fMRI), have played an important part in further investigating the sources of attentional control signals in the human brain. Using a variety of paradigms in which participants covertly direct attention to peripheral visual stimuli, a number of studies have reported activations within a common frontoparietal network of brain areas that seem to be recruited during directed spatial attention. These brain regions include the superior parietal lobe (SPL), temporal parietal junction, the frontal eye fields (FEF) and the supplementary eye fields (SEF), extending into the anterior cingulate cortex. An fMRI study by Kastner and colleagues compared activations in SPL, FEF and SEF in the presence and absence of visual stimulation. During the stimulus expectancy period (i.e. prior to delivery of the visual stimulus) there was an increase in activity within these frontoparietal areas that was greater than that observed in visual cortex. The sustained activity during the expectation period provides compelling evidence that the SPL, FEF and SEF areas were the sources of the signals producing attentional modulation of sensory activity seen in visual cortex.

One important characteristic of many directed attention tasks is the need for participants to hold in working memory information about the attended location in anticipation of stimuli to be presented at that location. Accordingly, it has been proposed that the signals controlling the deliberate allocation of spatial attention share neural mechanisms with those involved in spatial working memory. Consistent with this view is the finding that neurons in the lateral intraparietal sulcus (area LIP) of the monkey display elevated background firing rates when the animal performs a spatial working memory task as well as a spatial attention task. This functional link between spatial working memory signals and attentional control signals is further supported by human neuroimaging. Specifically, a number of neuroimaging experiments requiring the storage of spatial information in working memory have reported activations in posterior parietal and lateral frontal regions largely overlapping with those activated during spatial attention tasks.

The visual cortex in primates (including humans) and many other species is divided into a number of specialized regions that are organized in a hierarchical fashion. Visual information from the retina is relayed through the thalamus to the primary visual cortex (also called area V1), which is the first stage of cortical processing. The receptive fields (areas of space within which a stimulus will drive a cell) of V1 cells are small. From area V1, visual information propagates to successively higher areas including areas V2, V4, IT and MT. The structure of this hierarchical system is complex and it may contain over 35 separate areas specialized for visual processing. Among the characteristics that distinguish

one visual area from another are the type of stimuli to which each area responds and the size of receptive fields for individual cells in each area. Areas that are higher in the visual cortical hierarchy tend to respond to progressively more complex types of stimuli and have larger receptive fields.

Attention effects tend to become larger at higher levels of visual processing. These effects in primary visual cortex tend to be small: experiments in both monkeys and humans have shown that these responses become larger in area V2 and even larger in area V4.

One function of spatial attention is to bias the competition between visual objects when two or more objects are within the same receptive field. (*See* **Attention, Neural Basis of**)

One might imagine that attention could work by enhancing neural signals associated with the objects or spatial locations to which one's attention was directed, or by inhibiting the processing of nonattended items or locations. In fact, there is evidence that both types of mechanism – enhancement and suppression – are at work. Imaging studies in humans have shown that areas of visual cortex corresponding to the attended region of visual space show increased levels of activation not only when an object is present but also in the period between an attention-directing cue and the actual appearance of the object. This 'prestimulus' effect has been described in monkeys also in areas V2 and V4. In addition to areas of increased activity corresponding to attended regions, there is evidence for suppression of activity in nonattended regions.

Although orienting to visual stimuli has been most researched, there is increasing interest in orientation to information in other sensory modalities. Many of the mechanisms involved in vision are also used in orienting to both auditory and tactile modalities. There is evidence that orienting to one side of space can enhance information arising from different modalities at that location, even when the participant is not expecting a stimulus in that modality. The sensory sites at which attention operates can be very different, depending on the modality of the target. Whereas cueing to spatial locations dominates visual selection, auditory cueing of frequency of the target is more efficient than spatial cueing, perhaps because of the nature of cortical maps in the two modalities. Many of the sources of orienting effects are similar to those discussed for purely visual orienting. For example, the superior colliculus, superior parietal lobe, frontal eye fields and other eye movement structures are important to spatial orienting when the target information is not visual. Many areas of the parietal lobe are multimodal and may be involved both in integrating information across modalities and in orchestrating shifts of attention to novel stimuli. However, it is not yet understood how these areas are involved in the preattentive integration of information to form multimodal objects and in attentional orienting to the resulting objects.

FRONTAL CONTROL SYSTEMS

While spatial attention has a distinct anatomy, it can also act as part of a more general attention system that involves several other frontal areas and subcortical areas. When targets require difficult discriminations to separate them from nontarget events a network of frontal areas including the anterior cingulate gyrus, lateral frontal areas and sometimes the basal ganglia and cerebellum are active. These areas are involved in dealing with events in absence, in mediating conflict between targets and in generating novel responses. The exact function of each area is still unclear. A popular current view proposes that the medial frontal areas, including the anterior cingulate, may be involved in monitoring for cognitive and emotional conflict. Lateral frontal areas may serve to hold items on-line (working memory) and provide priority for selected information. The basal ganglia are often thought to be involved in switching between tasks. Orienting can have close ties to this more general frontal network, as when attention is drawn to places of interest during solving a problem or as indicated by a narrative. Because of this feature orienting is frequently used by cognitive studies to develop the time course of information processing during complex tasks.

Further Reading

Desimone R and Duncan J (1995) Neural mechanisms of selective attention. *Annual Review of Neuroscience* **18**: 193–222.

LaBerge D (1995) *Attentional Processing*. Cambridge, MA: Harvard University Press.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* **23**: 167–202.

Pashler HE (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Posner MI and Petersen SE (1990) The attention system of the human brain. *Annual Review of Neuroscience* **13**: 25–42.

Spatial Disorders

Introductory article

Giuseppe Vallar, University of Milan–Bicocca, Milan, Italy

CONTENTS

Cerebral lesions may impair a person's ability to perceive and remember objects, and to orient and navigate in the surrounding space. These neuropsychological disorders may be highly selective, suggesting that the internal anatomofunctional representation of space is multicomponential, although our phenomenal experience of it is largely unitary.

INTRODUCTION

Human beings, and other animals, live in a complex environment. They receive and process information concerning objects in space, and the spatial position of their body, through different sensory modalities (visual, auditory, somatosensory, vestibular); they continuously move, and are able to keep track of the position of their body, and of the location of objects in the space around them. These complex skills, essential to survival, comprise the perceptual processing of different sensory inputs from a continuously changing environment, and the programming and execution of motor acts. These include pointing to and reaching for objects through grasping, and locomotion. Sensory and motor information give rise to an internal representations of the body in space, and of the space around us. Our phenomenal experience of space is unitary. This is based, however, on the operation of multiple representations of different aspects of space. These may be selectively impaired by lesions of the brain, giving rise to specific disorders of spatial cognition.

MULTIPLE REPRESENTATIONS OF SPACE

The processing of sensory inputs first gives rise to representations of the stimulus specific to each sensory modality – retinotopic in vision, somatotopic in the tactile domain. Brain damage disrupting these levels of representation brings about primary sensory disorders, such as visual field defects (e.g. hemianopia) or impairments of tactile perception (e.g. hemianesthesia). The integration of visual, auditory, and somatosensory information with signals (eye position, vestibular, proprioceptive) concerned with the position of the body and of body parts in space results in two fundamental classes of representations of objects in space, including the subject's own body: egocentric and allocentric. In egocentric coordinate frames the position of the object is coded with reference to the subject's whole body or body parts, giving rise to representations, which may be head-centered (in the visual domain, resulting from the combination of the retinotopic map with information about eye position), trunk-centered (based also on information about the position of the head and about posture), arm-centered, and so forth (Figure 1). In allocentric coordinate frames, objects are primarily coded with reference to their spatial and configurational properties, such as the relationships between their component parts, and among different objects present in the environment. Egocentric representations may be used for the organization of goal-directed movements such as reaching a target or avoiding a harmful stimulus. Allocentric representations, encoding the configurational properties of objects and the relationships among them, may be useful for their identification and for navigation in space. In ecological conditions, objects are typically perceived from a variety of egocentric (observer-based) perspectives, suggesting a close interaction between these two frames of reference.

SPATIAL DISORDERS

Unilateral Spatial Neglect

Lesions to one cerebral hemisphere may severely impair the person's ability to detect and report sensory stimuli presented in portions of space contralateral to the side of the lesion (contralesional),

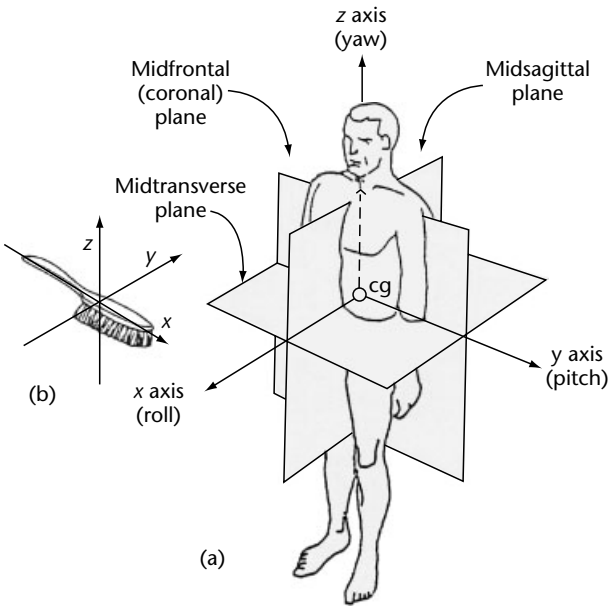


Figure 1. (a) The midsagittal plane of the body, which divides the extrapersonal space and the body into left and right sides, and other body coordinate systems and axes of rotation; cg: centre of gravity. (b) An object with its intrinsic axes. Adapted from Howard IP and Templeton WB (1966) *Human Spatial Orientation*. Chichester: John Wiley, with permission from John Wiley.

and to explore, through pointing or reaching movements or locomotion, that side of space. Neglect is a higher-order cognitive disorder, which can occur independently of primary sensory (e.g. hemianopia, hemianesthesia, deafness) or motor (e.g. hemiplegia) disorders. In humans, unilateral neglect is definitely more frequent and severe after damage to the right cerebral hemisphere, and concerns therefore the left side of space.

The spectrum of spatial neglect

Neglect of contralesional stimuli may occur in different sensory modalities – visual, auditory, tactile, or olfactory. Figures 2 and 3 show examples of left visuospatial unilateral neglect. The deficit may also concern the contralesional side of images evoked from long-term memory. Patients with this disorder, in an experiment requiring them to recall a familiar landscape (such as the Piazza del Duomo in Milan) from a given vantage point, failed to report landmarks on the left side. When, however, the vantage point was reversed, the neglected items became those that were left-sided with reference to the new perspective (Figure 4). This imaginal manifestation of neglect shows unambiguously that the disorder cannot be explained in terms of defective sensory input. In all these examples spatial neglect

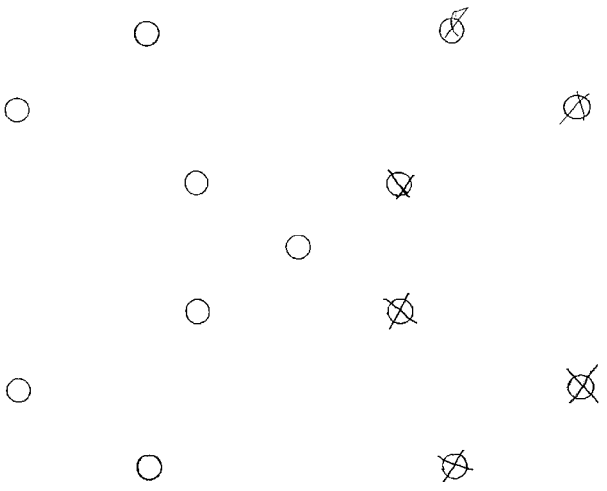


Figure 2. Spatial unilateral neglect. In a circle cancellation task a person with right-sided brain damage, who used his unaffected right hand, omitted to cross out circles on the left side.

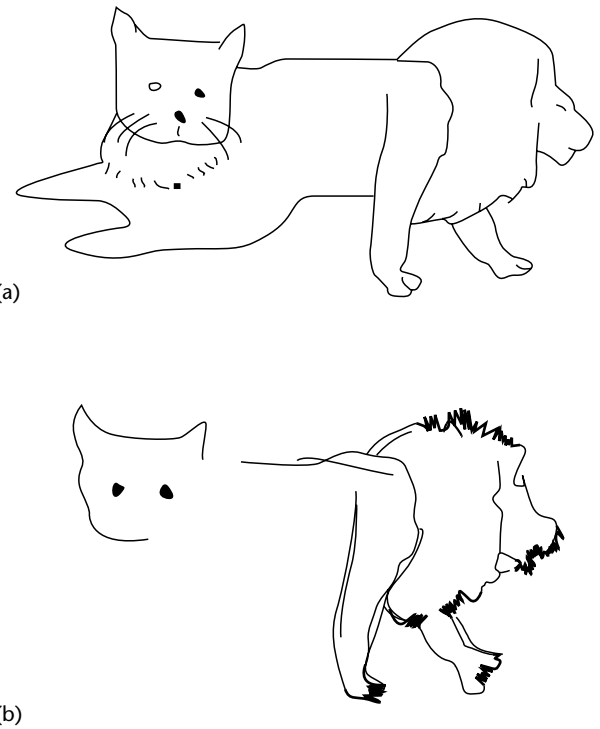


Figure 3. Patients with right brain damage and left unilateral neglect typically omit left-sided details in copying tasks. This cat–lion chimera (a) was identified by the patient as a lion on the basis of its right side (b). From Vallar G (1998) Spatial hemineglect in humans. *Trends in Cognitive Sciences* **2**: 87–97, with permission from Elsevier Science.

is defined with respect to an egocentric reference frame, such as the midsagittal plane of the body.

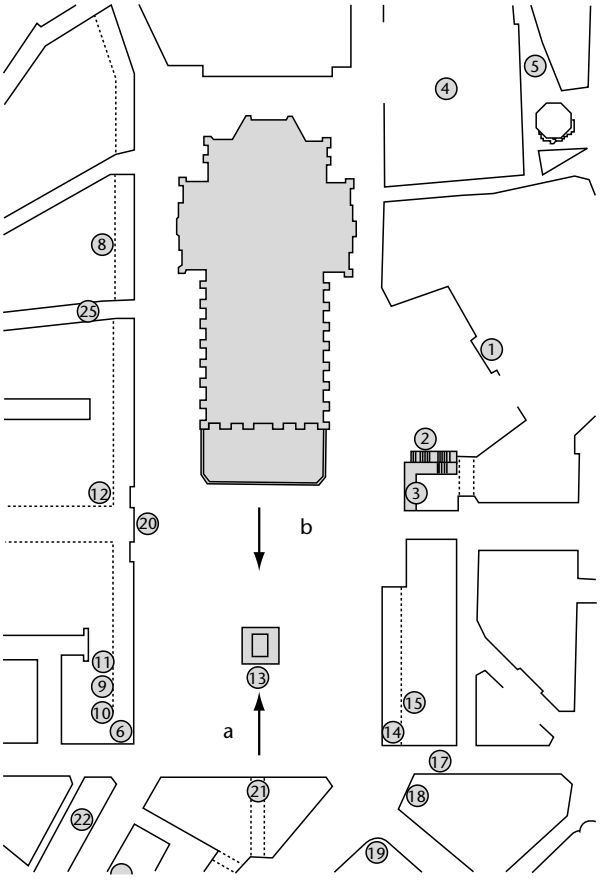


Figure 4. The Piazza del Duomo, Milan, Italy. Imagining the piazza from perspective (a) (looking at the cathedral with the steps in front of them), patients with left neglect fail to recollect left-sided details such as the front of the Galleria (20) or the Via San Raffaele (25), but not right-sided details such as the Royal Palace (1) or the Arengario (3). From the opposite perspective (b) (from the front doors of the cathedral) the deficit is reversed. From Bisiach E and Luzzatti C (1978) Unilateral neglect of representational space. *Cortex* **14**: 129–133, with permission from Masson.

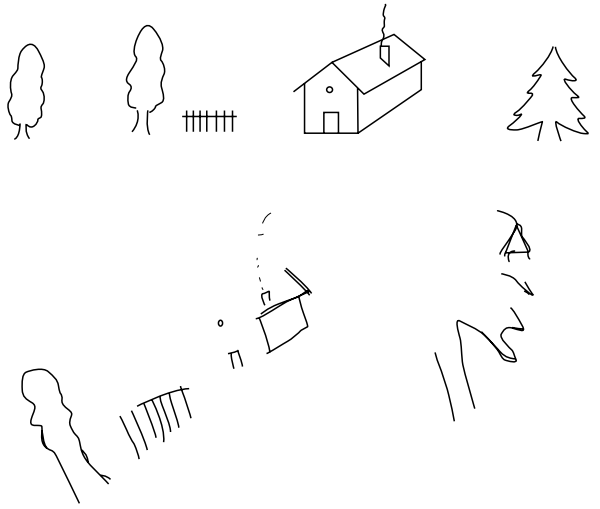


Figure 5. Object-based neglect. The patient copied many elements of the model, not only on the right but also on the left side of the sheet, but left unfinished the left half of various elements (the house and the fir tree). Reprinted from Gainotti G, Messerli P and Tissot R (1972) Qualitative analysis of unilateral spatial neglect in relation to laterality of cerebral lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **35**: 545–550, with permission from BMJ Publishing Group.

Within the space around us (extrapersonal) the disorder may be confined to specific sectors of it, such as 'near' (within hand-reach) and 'far' space. It may be also confined to specific types of objects (words, faces). Finally, although the lateral (left–right) dimension of neglect is the most frequently observed, the disorder may also manifest in other dimensions (e.g. altitudinal).

Spatial neglect may affect the contralesional left side of individual objects, rather than the left side of the whole field of perception (Figure 5). In a reading task, neglect may disrupt processing of the contralesional letters with reference to a canonical (e.g. horizontal, left to right, in a language such as English) representation of the word, independent of its physical presentation (vertical, mirror-reversed). These manifestations of neglect indicate that the disrupted level of representation concerns primarily an allocentric (object-centered or object-based) rather than an egocentric spatial reference frame.

These manifestations of neglect may reflect defective perception, impaired programming of movements towards the contralesional side of space, or both input and output factors. Patients with neglect may show evidence of entirely preserved processing of objects in the neglected side, up to the semantic level, provided such a knowledge is assessed by paradigms not entailing perceptual awareness (e.g. electrophysiological indexes, priming). Seen in this perspective, neglect is a case of defective access to conscious experience.

Neglect may affect the contralesional side of the body of the patient, who may be unaware of the presence of the left hand or arm, failing to reach them with the right hand (hemiasomatognosia). Patients may be selectively unaware of neurological disorders which affect their left side (anosognosia for hemiplegia, hemianopia, hemianesthesia) and actively deny the presence of a contralesional neurological deficit. They may develop delusional views about the contralesional side of their body (somatoparaphrenia). For instance,

patients might obdurately assert that their left arm is in fact their niece's arm, with this belief being resistant to persuasion. Anosognosia and somatoparaphrenia provide insights into the mechanisms of aware monitoring of function and of sense of ownership and personal identity.

In many patients these different manifestations of neglect coexist. However, empirical evidence has accumulated to show that each of these forms of neglect may occur in isolation, independent of the others. These dissociations of symptoms and signs provide a strong argument towards a view of the clinical syndrome of spatial unilateral neglect as a multicomponent deficit, and of spatial cognition as based on the integrated operation of multiple representational systems. These manifold manifestations of spatial neglect may be temporarily improved by a number of lateralized sensory inputs, including vestibular, proprioceptive–somatosensory, and visual stimulations. The observations that these modulatory effects concern virtually all manifestations of neglect, ranging from extrapersonal visuospatial neglect to somatoparaphrenia, suggest the existence of a common spatial representational medium, shared by different, more specific, representations.

Anatomic and functional basis

Unilateral spatial neglect is caused by a variety of cortical and subcortical lesions, which most frequently involve the right cerebral hemisphere, particularly in the parietal lobe. In the cortex, the inferior–posterior parietal region, at the temporoparietal junction, and the frontal premotor cortex (Brodmann's areas 6 and 44) represent main anatomical correlates (Figure 6). Damage to subcortical structures, such as the thalamus, the basal ganglia, and some white-matter fiber tracts, may also bring about neglect, suggesting that the disruption of a cortico-subcortical circuitry may be the pathological underpinning of the disorder. Damage confined to the primary sensory and motor cortices does not cause neglect, confirming the higher-order, cognitive character of the disorder.

The hemispheric asymmetry that characterizes neglect in humans may be explained by the hypothesis that both cerebral hemispheres include spatial representational and attentional systems devoted to perceptual processing and motor programming in the contralateral side of space. This organization replicates the general crossed anatomofunctional architecture of sensorimotor systems, whereby each hemisphere is mainly concerned with the contralateral side. In the case of spatial cognition, however, an hemispheric asymmetry is present.

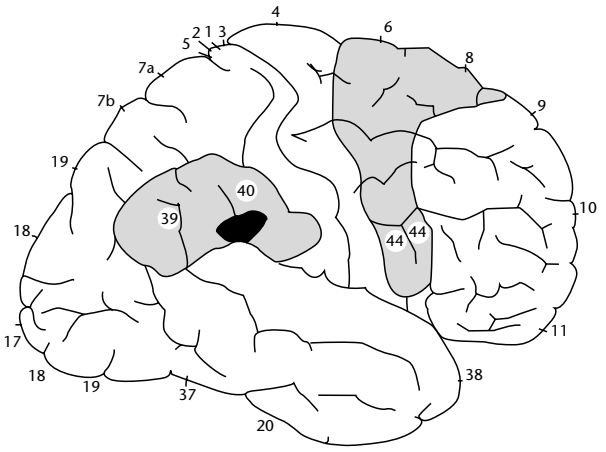


Figure 6. The anatomical basis of extrapersonal visual unilateral neglect (shaded area). Numbers are Brodmann areas. In the majority of patients the lesion involves the inferior parietal lobule, at the temporoparietal junction (black area), in the right hemisphere. Unilateral neglect after right frontal damage is less frequent and is usually associated with dorsolateral lesions of the premotor cortex. From Bisiach E and Vallar G *Unilateral Neglect in Humans*, Handbook of Neuropsychology, 2nd edn, vol. 1, pp. 459–502, with permission from Elsevier Science.

Only the right hemisphere is able to deploy spatial attention to both the left and the right sides of space, and has a complete lateral representation of it. The left hemisphere, by contrast, is mainly concerned with the contralateral right side of space. Accordingly, damage to the right hemisphere brings about a left-sided deficit, which cannot be compensated for by left-hemisphere neural networks. This hemispheric difference in the neuronal machinery subserving spatial representation and attention is likely to be a matter of degree. The difference between the 'neglected' and the 'preserved' sides of space, in a given reference frame, should be conceived as a contralateral–ipsilateral gradient, rather than a sharp divide.

Disorders of Spatial Perception and Reaching

Unilateral hemispheric damage may impair the patient's ability to localize visual and auditory stimuli, and to perceive the orientation of lines, in the visual and tactile modality. In general, the deficit is much more severe after damage to the right cerebral hemisphere. Patients with cerebral damage may be selectively impaired in reaching for visual objects with their arm (optic or visuomotor ataxia). The responsible deficit involves the spatial egocentric reference coordinates including

the representation of the position of the target with respect to the body and the reaching effector (e.g. the arm and hand). Allocentric reference frames are not involved, since these patients can correctly localize objects with respect to each other. Optic ataxia is associated with posterior parietal lesions, with a critical role of damage to the interparietal sulcus and the superior parietal lobule. The association of optic ataxia with 'gaze apraxia' (the inability to shift gaze, in order to bring peripheral visual stimuli into fixation), together with an extreme narrowing of the attentional field preventing the perception of more than one object at a time, and with defective estimation of distance and depth, constitutes the Bálint–Holmes syndrome. This symptom complex is usually caused by a bilateral parietooccipital lesion.

Disorders of Spatial Memory

A broad distinction within memory systems is between components concerned with the retention of a limited amount of material for short periods (seconds) (short-term memory) and components involved in the longer-lasting acquisition and retention of information (long-term memory). Unlike neglect, these spatial disorders do not show any lateral or altitudinal gradient.

Deficits of spatial short-term memory

Short-term memory for spatial locations is selectively disrupted by damage to the posterior, parietooccipital regions of both the left and right cerebral hemispheres, although the role of the latter appears to be more relevant. These patients show a selective deficit of immediate retention of short sequences of spatial locations, as assessed by a block-tapping task (a spatial analogue of digit span). The impaired retention of spatial locations cannot be accounted for in terms of visuoperceptual or response-related disorders (e.g. unilateral spatial neglect or misreaching). These patients may show no evidence of topographical disorientation or impairment in long-term learning of visuospatial information such as the path of a visual maze, suggesting a complete independence of short-term and long-term memory systems concerned with spatial locations.

Deficits of spatial long-term memory and topographical disorientation

Right hemispheric lesions involving the medial temporal regions in the limbic lobe and the posterior parietal areas impair learning and retention of the spatial position of objects and of pathways. Patients show a selective and disproportionate impairment in navigating in the environment, and are unable to find familiar routes, to learn new ones, to

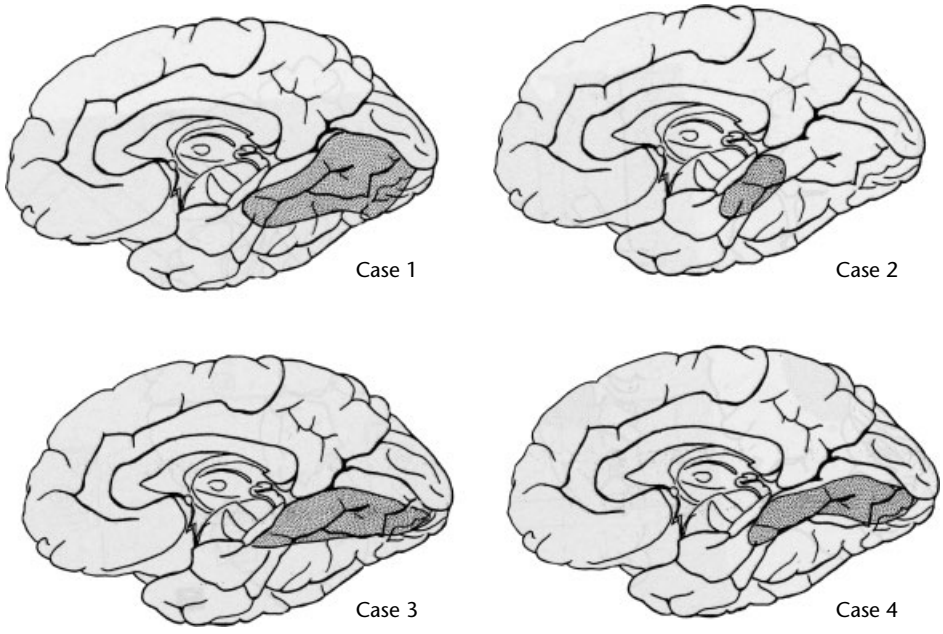


Figure 7. Lesions involving the right parahippocampal gyrus in four patients with pure topographical disorientation (medial view of the right hemisphere). From Habib M and Sirigu A (1987) Pure topographical disorientation: a definition and anatomical basis. *Cortex* **23**: 73–85, with permission from Masson.

draw a map, or to describe the path to reach one place from another (topographical amnesia). These patients are unable to draw or describe verbally the map of the Piazza del Duomo in Figure 4, or to walk from one place to another. Topographical disorientation, unlike unilateral neglect, does not exhibit a lateral gradient of impairment. These patients can recognize familiar landmarks but cannot give them any localizing value, being unable to recall their relationships. The disorder concerns mainly allocentric coordinate frames. The lesion responsible for purely topographical amnesia involves the medial temporal regions of the right hemisphere, with a main role of damage to the parahippocampal gyrus (Figure 7).

CONCLUSION

Neuropsychological investigations in people with brain damage suggest that spatial cognition, far from being monolithic, is articulated into a number of discrete systems, with specific neural correlates, which may be selectively impaired. These disorders include deficits characterized by a gradient of impairment, along a specific spatial axis (the syndrome of unilateral spatial neglect), disorders of reaching (optic ataxia), of spatial perception, and impairments of short-term and long-term spatial memory (topographical amnesia or disorientation). The neural correlates of spatial cognition, and of its neuropsychological disorders, are mainly based in the right hemisphere, including the premotor frontal, posterior parietal, and medial temporal regions. These spatial impairments may be qualified with respect to the deficit of a specific coordinate frame or representation, the main distinction being between egocentric and allocentric dimensions.

Further Reading

Andersen RA, Snyder LH, Bradley DC and Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–330.

Bermudez JL, Marcel A and Eilan N (1995) *The Body and the Self*. Cambridge, MA: MIT Press.

Bisiach E and Vallar G (2000) Unilateral neglect in humans. In: Boller F, Grafman J and Rizzolatti G (eds) *Handbook of Neuropsychology*, 2nd edn, vol. 1, pp. 459–502. Amsterdam, Netherlands: Elsevier.

Burgess N, Jeffery KJ and O'Keefe J (eds) (1999) *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford, UK: Oxford University Press.

De Renzi E (1982) *Disorders of Space Exploration and Cognition*. Chichester, UK: John Wiley.

Fink GR, Freund HJ, Marshall JC and Zilles K (eds) (2001) Action and visuo-spatial attention: neurobiological bases and disorders. *Neuroimage* **14**: S1–S146.

Heilman KM, Watson RT and Valenstein E (1993) Neglect and related disorders. In: Heilman KM and Valenstein E (eds) *Clinical Neuropsychology*, 3rd edn, pp. 279–336. New York, NY: Oxford University Press.

Howard IP (1982) *Human Visual Orientation*. Chichester, UK: John Wiley.

Lacquaniti F (1997) Frames of reference in sensorimotor coordination. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 11, pp. 27–64. Amsterdam, Netherlands: Elsevier.

Nichelli P (1999) Visuo-spatial and imagery disorders. In: Denes G and Pizzamiglio L (eds) *Handbook of Clinical and Experimental Neuropsychology*, pp. 453–477. Hove, UK: Psychology Press.

Prigatano GP and Schacter DL (eds) (1991) *Awareness of Deficit After Brain Injury. Clinical and Theoretical Issues*. Oxford, UK: Oxford University Press.

Rizzolatti G, Berti A and Gallese V (2000) Spatial neglect: neurophysiological bases, cortical circuits and theories. In: Boller F, Grafman J and Rizzolatti G (eds) *Handbook of Neuropsychology*, 2nd edn, vol. 1, pp. 503–537. Amsterdam, Netherlands: Elsevier.

Robertson IH and Marshall JC (eds) (1993) *Unilateral Neglect: Clinical and Experimental Studies*. Hove, UK: Lawrence Erlbaum.

Thier P and Karnath HO (eds) (1997) *Parietal Lobe Contributions to Orientation in 3D Space*. Heidelberg, Germany: Springer-Verlag.

Vallar G (1998) Spatial hemineglect in humans. *Trends in Cognitive Sciences* **2**: 87–97.

Split-brain Research

Introductory article

Maryse Lassonde, University of Montreal, Montreal, Quebec, Canada

CONTENTS

Split-brain research has led to an understanding of the role of the two hemispheres in perceptual and cognitive processes, and has raised important questions with regard to human consciousness and self-awareness.

INTRODUCTION

The corpus callosum, which has an estimated 600 million fibers, constitutes the largest communication pathway between the two cerebral hemispheres (Figure 1). The surgical division of the pathway which allows the two hemispheres of the brain to communicate with each other (the 'split brain') is performed for therapeutic purposes in cases of intractable epilepsy. For over 40 years this procedure has allowed testing of the specific cognitive and perceptual abilities of each cerebral hemisphere, thereby enhancing our knowledge of hemispheric specialization which, prior to the 1960s, had been derived mainly from brain lesion studies. However, whereas the effects of cerebral lesions could be obscured by many factors, including individual variability, the split-brain studies allowed researchers to investigate for the first time the abilities of each hemisphere *within the same individual*.

HISTORY OF SPLIT-BRAIN STUDIES AND THEIR SIGNIFICANCE

The interest in and mystery surrounding the role of this gigantic bundle of fibers have a long history in the field of neurosciences. In the eighteenth century, just as Descartes had proposed the pineal gland as the site where body and mind interact, Giovanni Lancisi, an Italian physician, proclaimed that the corpus callosum, because it was both unique and situated in the middle of the brain, was the seat of the soul.

In the nineteenth century, when questions were raised about the double nature of the brain (it is composed of two hemispheres that appear to be largely symmetrical), the corpus callosum was regarded as a structure that could reconcile this duality, thereby providing unity of consciousness and preventing the development of a 'double personality'. However, several cases were found in which no mental impairments could be observed in the absence of the corpus callosum. The notion of the callosal structure as an essential link to consciousness was therefore abandoned, and its exact role remained the subject of debate.

This situation was further complicated by reports published in the 1940s. Following the observation that, in animals, sectioning of the corpus callosum ('callosotomy') could prevent the spread of epileptic seizures from one hemisphere to the other, neurosurgeons decided to apply this operative technique to patients who had drug-resistant epilepsy. The psychologist Andrew Akelaitis, who conducted extensive behavioral studies in these first split-brain patients, failed to demonstrate any specific pattern of disconnection between the two hemispheres with regard to behavior, emotion or cognition. These negative findings prompted scientists such as Warren McCulloch to claim that the unique role of the corpus callosum was the propagation of epileptic seizures from one hemisphere to the other, or perhaps, as was facetiously proposed by Karl Lashley, its function was simply to hold the two hemispheres together!

Fortunately, a brilliant student of Karl Lashley, namely Nobel Laureate Roger Sperry, together with his colleague Ronald Myers, decided to investigate the role of this anatomically important

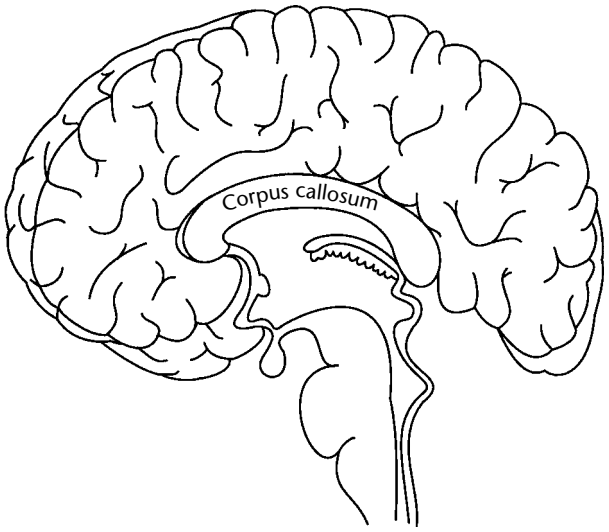


Figure 1. Midsagittal view of the brain, showing the internal portion of the disconnected right hemisphere. The corpus callosum is the large fiber tract that interconnects the two hemispheres.

structure further. They surgically divided the brain of cats by cutting the corpus callosum. They also sectioned another pathway – the optic chiasm – such that visual information from one eye would reach only the ipsilateral hemisphere. They then trained the animal to learn a simple visual discrimination task while one of the eyes was occluded. Once the animal had mastered the problem, the same task was presented to the untrained eye (and hemisphere) while the trained eye was occluded. Extraordinarily, the cat seemed to have no recollection of the task. It had to learn the task all over again with the other half of the brain. In fact, the animal behaved as if the section of the corpus callosum had produced two independent minds within its head, whereas outside the testing situation, nothing in its behavior had changed or distinguished it from its normal cage mates.

Soon afterwards, two Californian neurosurgeons, Philip Vogel and Joseph Bogen, decided to apply the surgical procedure used in the 1940s to patients with severe intractable epilepsy. Having heard about Sperry's work with cats and monkeys, Bogen invited Sperry and one of his colleagues, Michael Gazzaniga, to study these split-brain patients. Like the split-brain animals, these patients displayed normal behavior in their everyday lives, but under special testing conditions they behaved as if they had two distinct types of conscious awareness, one in each hemisphere. These results confirmed the animal research and contributed to a new understanding of the functions of the corpus callosum which eventually opened up a new era in the study of the relationships between brain and mind.

THE ROLE OF THE CORPUS CALLOSUM AND THE TWO HEMISPHERES IN COGNITION

Methodological Issues

In contrast to the situation in animals, the corpus callosum in humans joins two functionally different hemispheres. Sectioning of the corpus callosum thus allows the functions of each hemisphere in isolation to be studied. Callosotomy disrupts all of the direct connections between the left and right hemispheres, including the connections between the regions responsible for the motor and sensory control of each hand and those that interconnect the left and right visual fields. However, under normal circumstances the division of functions between the two hemispheres cannot be easily observed, due to the bilateral 'wiring' of the brain and the use of various strategies by the patients. This is one of the reasons why Akelaitis failed to observe any disconnection signs in the earlier series of split-brain patients that he studied. Sperry and his collaborators, on the other hand, devised specific testing procedures that restricted processing of sensory and motor information to a single hemisphere. An example of these procedures is shown in Figure 2.

Somatosensory and Motor Control

Following callosotomy, tactile information processed by one hand remains largely lateralized to the opposite hemisphere. Thus split-brain patients cannot compare two objects manipulated with one in each hand, although they are aware of holding something in each hand. However, noxious information from each hand, such as intense heat or pain, is conveyed to both hemispheres and is consciously perceived by each half of the brain as an unpleasant sensation.

The movements of each hand are also controlled by the opposite hemisphere. Early research indicates that split-brain patients experience difficulty in performing asynchronous, simultaneous movements with their two hands. In rare instances, split-brain patients perform conflicting actions with their hands (e.g. pouring water with one hand while taking away the glass with the other).

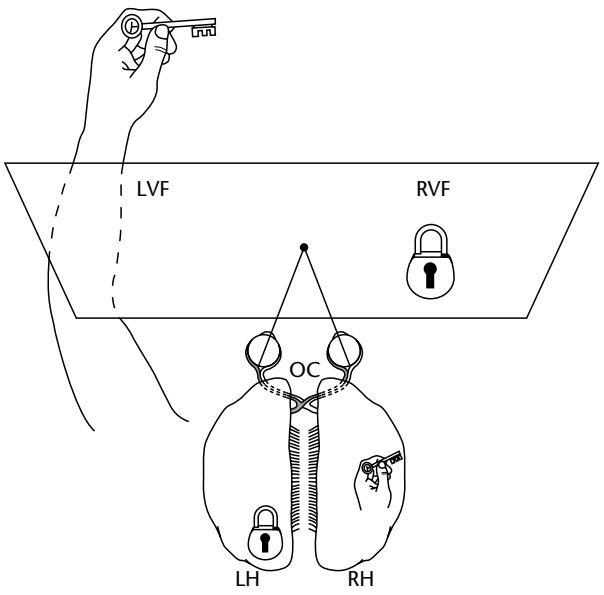


Figure 2. An example of the type of procedure used with split-brain patients. Visual information is lateralized to the right or left hemisphere by presenting it to the left or right side, respectively, of a central fixation point for a very short time in order to prevent eye-tracking and hence bilateral visual representation. Tactile information is kept within one hemisphere by asking the subject to feel an object out of view. In this set-up, the split-brain subject is incapable of comparing an object (key) held by the left hand (right hemisphere) with the visual stimulus flashed in the right visual field (left hemisphere). Moreover, only the object processed by the left hemisphere (key) can be named by the subject. RVF, right visual field; LVF, left visual field; OC, optic chiasm; LH, left hemisphere; RH, right hemisphere.

Patients who present with this so-called *diagonistic apraxia* act as if each hemisphere wanted to take control of the environment.

Language and Speech

Split-brain studies have consistently shown a marked left-hemisphere superiority for verbal processing, and this applies not only to right-handed subjects whose left hemisphere is dominant for motor control, but also to the majority of left-handed individuals. Although split-brain patients can readily name objects held in the right hand or presented in the right visual field (and hence projecting to the left hemisphere), they are unable to do so when the stimuli are processed by the right hemisphere, because the latter does not possess the mechanisms of speech. Patients may even deny having seen anything when the images are flashed in the left visual field, and some will report that their left hand has 'lost its sensitivity'. However, it appears that the patients have a full knowledge of the stimuli, even though they cannot name them, because when prompted they are able to retrieve with their left hand (connected to the right hemisphere) the corresponding object from an array of objects, or point to the picture that has been presented in their left visual field.

Although the disconnected right hemisphere has problems in producing speech, it does have some linguistic abilities. Thus many studies have shown that the right hemisphere possesses a lexicon, that it can make judgments about grammaticality, and that it can even sometimes develop expressive speech abilities, allowing the patient to report verbally sensations and perceptions experienced in their previously mute right hemisphere. In fact, research with split-brain patients has shown that there is enormous inter-individual variability in the linguistic abilities of the isolated right hemisphere.

Visuospatial Processing and Perception

There is a general consensus that the right hemisphere is superior with regard to processing spatial information. For example, split-brain patients – even right-handed ones – succeed better on the block design subtest of the Wechsler intelligence scale when using their left hand. This test, which involves reproducing a visual pattern using red and white blocks, is strongly correlated with the non-verbal intellectual quotient (performance IQ), suggesting that the right hemisphere has some specific higher-order cognitive abilities. Right hemisphere superiority has been found for other types of spatial abilities, such as perceiving the shape of a partially occluded object, deciding whether two visual images are identical or mirror-reversed, judging spatial relationships or recognizing faces. Drawings are also better accomplished by the left hand, which may sometimes even try to take over the right-hand execution (Figure 3).

Memory and Learning

Split-brain patients sometimes complain of memory problems following surgery, and indeed early reports indicated deficits in short-term memory in these patients. However, subsequent studies have reported improvements in memory following callosotomy. The most important difference between the earlier and later series of patients is the preservation of the anterior commissure in

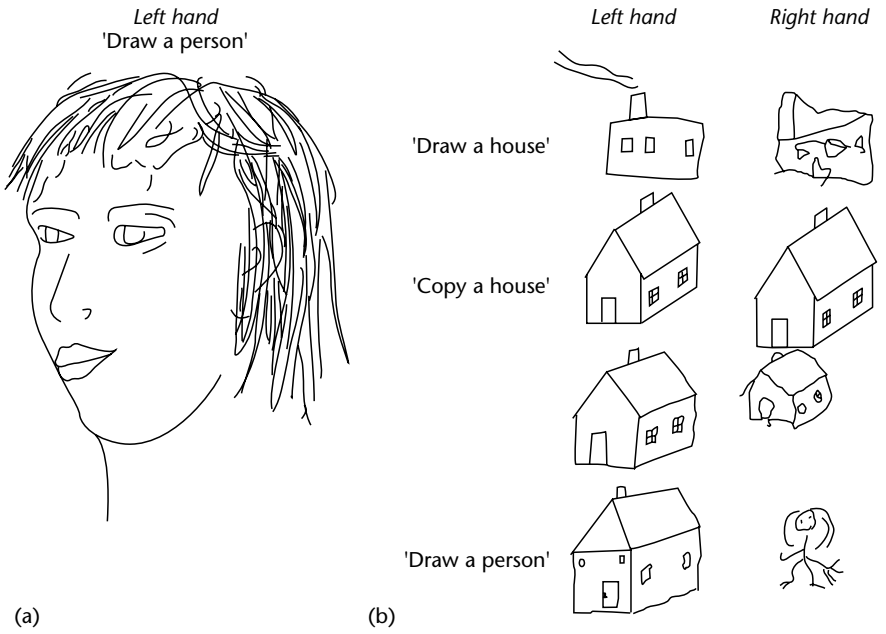


Figure 3. Drawings produced by one of our split-brain patients examined pre- and post-surgically. (a) Prior to surgery (pre-callosotomy), this left-handed patient used to draw as a hobby. (b) Following the operation (post-callosotomy), he can still produce drawings with his left hand (and thus his right hemisphere), although his spontaneous productions represent simplified versions of objects. Using his right hand, his spontaneous reproductions are rudimentary and often unrecognizable. In this particular instance, when the patient used his right hand to copy a house, the left hand tried to take over the copy but was stopped by the experimenter. When the patient was subsequently asked to draw a person with his left hand, the 'frustrated' right hemisphere produced a well-executed house. However, the patient complained verbally about the production, stating 'This is not a person'. The right hand (left hemisphere) then tried to draw a person, but with only limited success.

the latter, which is known to link areas in both hemispheres that mediate memory. More recent research suggests that in some cases callosotomy may interfere with the encoding of new information, but without affecting recognition of the previously learned material. Furthermore, consistent with previous results reported for unilateral brain lesions, the disconnected left hemisphere has been found to be more proficient in encoding verbal material (e.g. words), whereas the right hemisphere is superior with regard to remembering visual information (e.g. images and faces).

With regard to procedural memory, each hemisphere is capable of learning a routine or procedure, but the memory engram that results from this learning is not necessarily transferred to the other hemisphere. Furthermore, split-brain patients experience difficulties in implicitly (automatically) learning a routine that involves the simultaneous use of both hands, although they can explicitly (or consciously) describe all of the steps of the procedure. In fact, they show a dissociation between implicit (unconscious) and explicit (conscious) learning, which may explain the apparent unity in overt behavior in the presence of two cognitive systems that operate largely independently from one another.

Attention

Normal individuals can only attend to one visuospatial location at a time. However, studies in split-brain patients have shown that each of the two hemispheres is able to direct its attention to its own visual field. Thus the divided hemispheres can independently and concurrently scan their corresponding visual fields during visual search such that it takes a split-brain patient half the time to scan a given number of pictures when they are presented on either side of the midline compared with when they are viewed on only one side. In fact, because of the 'division of labor' between the two hemispheres, split-brain patients can outperform normal individuals in a test of visual retention when the information is distributed between the two visual half-fields.

IMPLICATIONS FOR PHILOSOPHY OF MIND AND SELF

Split-brain studies have thus shown that each disconnected hemisphere has its own set of percepts and cognitive abilities that have been variously described in terms of a number of dichotomies (e.g. verbal/visuospatial, analytic/holistic, propositional/appositional, global/local, rational/intuitive, or even intellectual/artistic). For several years now these distinctions have captured the attention of the general public, leading to popular (mis)interpretations. For example, it has been suggested that one could exercise one's right hemisphere in order to develop intuitive faculties or become more artistic (e.g. Betty Edward's 1980 book entitled *Drawing on the Right Side of the Brain*). Such popular beliefs fail to take into account the fact that the normal brain does possess a corpus callosum, which ensures that there is constant 'cross-talk' between the two hemispheres.

Apart from its impact on our knowledge of hemispheric specialization, split-brain research has also raised important questions about the nature and substrate of mind and conscious awareness. Transection of the corpus callosum apparently produces a division of the mind, since most conscious experience that is generated by one hemisphere appears to be inaccessible to the other. If this is the case, what does it signify about the unity of the conscious self in the normal intact brain?

One approach maintains that the mind and self remain unified because of the predominant control of the language hemisphere. The right hemisphere is viewed as subordinate, possessing few if any higher cognitive abilities and acting almost as an unconscious automaton. The intellectually superior left hemisphere, which alone is capable of verbal expression in most individuals, constantly formulates hypotheses and generates explanations about the environment and one's self. The interpretations given by the left hemisphere (labeled the 'Interpreter' by Gazzaniga) would thus provide the sensation of a unified mind even in the absence of available information from the right hemisphere.

Another view has emerged lately. Some recent research indicates that the right hemisphere may preferentially process self-face recognition, which has been regarded as an indicator of higher-order self-consciousness. Recognition of self (e.g. in a mirror) is indeed a predominantly human attribute that can be associated with other self-related tasks such as introspection and hence self-awareness. Interestingly, early reports indicated that split-brain patients displayed alexithymia (a difficulty in expressing feelings verbally). This would suggest that the verbal left hemisphere is unable to access emotions experienced by the (introspective?) right hemisphere. Such a theory would predict that division of the brain may disrupt the sense of self-awareness because an important part of the person's inner life cannot be conveyed explicitly by verbal means. However, as yet there have been no reports of a loss of personal identity or self-awareness in split-brain patients.

A more intermediate approach to the 'duality of mind', first expounded by Hughlings Jackson and endorsed by Sperry, views the mind as normally single and unified, with the two specialized hemispheres working in synchrony through the commissures, and functioning as an integrated whole rather than as a double or divided system. In fact, Sperry has proposed that this integrated activity leads to an emergent property whereby the combined activity of the two hemispheres is greater than the sum of each hemisphere functioning separately. He maintains that even in the split brain, where the cognitive processes of each hemisphere are largely separate, other aspects of consciousness are not necessarily divided. This is because, just like Siamese twins, the two surgically isolated hemispheres are exposed to the same experience. In the brain, this is achieved through bilateral distribution at several levels of sensory and motor pathways, as well as through behavioral means of compensation, such as explorative movements of the body and eyes which allow both hemispheres to be concurrently aware of a percept or sensation. However, whether the subsequent treatment of this information, either through the formulation of hypotheses or by introspection, can ultimately lead to the same high-level unified self-concept as exists in the undivided brain is still a matter of debate.

Acknowledgments

I wish to acknowledge the support of the National Science and Engineering Council of Canada and the Quebec Ministry of Science (Fonds FCAR). I also wish to thank Chris Davis for creating the illustrations.

Further Reading

Arguin M, Lassonde M, Quattrini A *et al*. (2000) Divided visuospatial attention systems with total and anterior callosotomy. *Neuropsychologia* **38**: 283–291.

Bogen JE (1993) The callosal syndrome. In: Heilman KM and Valenstein E (eds) *Clinical Neuropsychology*, pp. 308–359. New York, NY: Oxford University Press.

Bogen JE and Vogel PJ (1962) Cerebral commissurotomy in man. *Bulletin of the Los Angeles Neurological Society* **27**: 169–172.

De Guise E, del Pesce M, Foschi N *et al*. (1999) Callosal and cortical contribution to procedural learning. *Brain* **122**: 102–114.

Gazzaniga M (2000) Cerebral specialization and interhemispheric communication. Does the corpus callosum enable the human condition? *Brain* **123**: 1293–1326.

Keenan JP, Wheeler MA, Gallup GG Jr and Pascual-Leone A (2001) Self-recognition and the right prefrontal cortex. *Trends in Cognitive Sciences* **4**: 338–344.

Lassonde M and Jeeves M (1994) *Callosal Agenesis: A Natural Split-Brain?* New York, NY: Plenum Press.

Sperry R (1968) Hemisphere disconnection and unity in conscious awareness. *American Psychologist* **23**: 723–733.

Sperry R (1984) Consciousness, personal identity and the divided brain. *Neuropsychologia* **22**: 661–673.

Zaidel E (1998) Language in the right hemisphere following callosal disconnection. In: Stemmer B and Whitaker H (eds) *Handbook of Neurolinguistics*, pp. 369–383. San Diego, CA: Academic Press.

Stress and Cognitive Function, Neuroendocrinology of

Intermediate article

Michael J Meaney, McGill University, Montréal, Quebec, Canada
Sonia Lupien, McGill University, Montréal, Quebec, Canada

CONTENTS

Stressful events impose a very important set of demands on the body, and this is reflected by changes in cognitive function.

STRESS HORMONES AND COGNITIVE FUNCTION: OVERVIEW

Stress that occurs in the form of an actual or threatened insult produces states of increased emotional arousal which alter neuroendocrine function in a manner that is of obvious importance to an individual facing adversity. One way to appreciate the logic of these adaptive responses is simply to imagine the demands placed on the individual during virtually any form of stress. First, there are metabolic demands that arise from the increased level of cellular activity in tissues such as lung, liver, heart and brain. The organism meets these demands by mobilizing energy substrates (e.g. fats and sugars) from macromolecular storage sites and by increasing blood flow (and thus the distribution of energy substrates) through enhanced cardiovascular tone. These responses are produced as a result of increases in sympathoadrenal activity, resulting in increased levels of stress hormones, such as the highly catabolic glucocorticoids and the catecholamines, adrenaline and noradrenaline, which directly regulate lipolysis, gluconeogenesis and cardiovascular activity. These same hormones also influence learning and memory for information related to the stressful events, as well as the ability of the individual to process information that is not related to the stressor.

The release of these stress hormones occurs in response to signals from the hypothalamus as well as extrahypothalamic sites that assess the salience of the threat and regulate the release of stress hormone from sympathetic and adrenal sources. The most important of these regions are the basolateral regions of the amygdala, the hippocampus and the orbitofrontal cortex (or the medial prefrontal cortex in rodents). In addition, cognitive demands are placed on the individual that relate to coping with the immediate challenge and learning to avoid the same conditions in the future. During or immediately following a period of stress, cognitive function is focused on the events associated with the stress response. Not surprisingly, there is a price to pay for this. Thus the processing of information that is not related to the stressor is impaired, and this appears to be the case for both storage and retrieval of information.

Many of the adaptive, cognitive processes that occur during stress involve these same brain regions and are regulated by the same stress hormones. Thus glucocorticoids and catecholamines mediate stress-induced changes in attention, as well as learning and memory, through their effects on neuronal populations in the hippocampus and amygdala. In large measure, the effects of stress hormones on learning and memory can be understood in terms of the timing of the stressful event. The increase in emotional arousal that is associated with stressful events enhances learning and memory for those events. Thus the elevated levels of both glucocorticoids and catecholamines that are provoked by the stressor enhance memory storage. In contrast, these same hormones will impair storage of information that is related to events which follow the stressor, or the retrieval of information that is not related to the stressor. These effects may in part be explained by the effects of glucocorticoids and catecholamines on the hippocampus and amygdala.

THE PRINCIPAL STRESS HORMONES

The hypothalamic–pituitary–adrenal (HPA) axis is the primary endocrine branch of the stress response (de Kloet *et al*., 1998). Activation of the HPA axis results in increased adrenal release of glucocorticoids into the circulation. Under most conditions, the HPA axis is under the dominion of specific releasing factors that are secreted by neurons located in the paraventricular nucleus (PVNh) of the hypothalamus, notably corticotropin-releasing factor (CRF) and vasopressin. Neurons of the PVNh project extensively to the median eminence, providing a potent excitatory signal for the synthesis and release of adrenocorticotropin (ACTH) from the anterior pituitary, and thus the means whereby neural activity associated with the stressor is transduced into endocrine signals (de Kloet *et al*., 1998). Elevated ACTH levels in turn increase the synthesis and release of glucocorticoids from the adrenal gland (in primates, the primary glucocorticoid is cortisol, whereas in rodents it is corticosterone).

A second, neural branch of the stress response involves a projection from the PVNh, the central nucleus of the amygdala and the bed nucleus of the stria terminalis to midbrain regions, such as the noradrenergic cell bodies of the locus coeruleus (Valentino *et al*., 1998). Interestingly, this projection uses CRF as its primary neurotransmitter signal, resulting in the activation of noradrenergic neurons and the release of noradrenaline in multiple target regions in the cortex and limbic system. The locus coeruleus as well as neighboring CRF target sites such as the nucleus of the solitary tract are part of the midbrain structures that have classically been associated with emotional arousal. Thus activation of this amygdala–locus coeruleus pathway corresponds to the emotional states of fear and apprehension. Not surprisingly, increased activity in this pathway is associated with anxiety disorders, and consequently these structures have emerged as primary targets for anxiolytic drugs. Activation of these noradrenergic projections to regions such as the amygdala and hippocampus also influences learning and memory.

STRESS HORMONES AND COGNITIVE FUNCTION: THE AMYGDALA

Learning and memory are enhanced under conditions where an event is perceived by the organism as being highly aversive. The emotional state which occurs in response to the cognitive–perceptual processing of the stressor then influences the nature of subsequent cognitive events, including the strength of learning and memory, as well as the nature of what is learned. For example, rats acquire a conditioned fear response to a cue that predicts footshock after only one trial. The neural processes that underlie such emotional learning and memory and the relevant actions of stress hormones have been well described in the rat using various models of fear conditioning (Davis *et al*., 1994; Cahill and McGaugh, 1998; LeDoux, 2000). Information about the aversive properties of stimuli is transmitted from sensory pathways through cortical and subcortical structures such as the thalamus and the parabrachial nucleus to the basolateral complex of the amygdala, which consists of the basolateral, lateral and anterior basal nuclei. While debate continues as to whether the amygdala is the actual locus for the relevant memory, there is unanimity about the importance of the amygdala in the acquisition and expression of fear conditioning. Activity within the basolateral complex of the amygdala is essential for learning/memory that concerns aversive events, while activity within the medial portion of the central nucleus of the amygdala is associated with the expression of fear-like states, but not with learning/memory *per se*.

These sites are also primary targets for the enhancing effects of noradrenergic and dopaminergic projections on fear conditioning. Glutamatergic inputs into the basolateral complex of the amygdala are a critical feature of amygdaloid learning/memory, possibly involving long-term potentiation processes comparable to those of the hippocampus (see below). Thus glutamate-receptor antagonists block fear conditioning as well as fear-potentiated startle. Interestingly, glucocorticoids block glutamate reuptake by glial cells and thus increase the magnitude of an extracellular glutamate signal. In Pavlovian fear conditioning, the sensory information that is associated with the predictive stimuli as well as the nocioceptive information associated with pain converge on the basolateral complex of the amygdala to produce fear conditioning to the predictive stimuli. In contextual fear conditioning, information about context emerges from the ventral hippocampus and involves projections to the basolateral complex of the amygdala. Projections from the basolateral complex to the cental nucleus activate the expression of the fear response through the extensive projection from the central nucleus, including CRF projections, to regions that mediate autonomic, emotional and behavioral responses to stress, such as the locus coeruleus.

Similarly, in humans, emotionally adverse stimuli activate the human amygdala (Cahill and

McGaugh, 1998). Indeed, the degree of amygdaloid activation is highly correlated ($r = +0.93$) with recall of emotionally disturbing (but not neutral) material. With high-temporal-resolution functional magnetic resonance imaging (fMRI) techniques, LaBar *et al*. (1998) found increased amygdala activity during the acquisition phase of fear conditioning, and patients with amygdala damage show profound deficits in fear conditioning. Interestingly, individual differences in the degree of amygdaloid activation during fear conditioning were highly correlated with the strength of a conditioned autonomic fear response.

The enhanced learning and memory that are associated with a stressor are a result of the effects of the stress hormones on the amygdaloid complex. The infusion of a glucocorticoid-receptor agonist directly into the basolateral region of the amygdala facilitates learning. Drugs that activate β-adrenoreceptors in the amygdala have the same effect. In both human and non-human subjects, β-adrenoreceptor blockers attenuate emotional learning and memory. The glucocorticoids act in synergy with noradrenaline to activate hepatic and cardiovascular systems during periods of stress. Essentially the primary effect of the glucocorticoids is to increase tissue sensitivity to noradrenaline. A comparable pharmacological interaction appears to underlie the effects of the stress hormones on the amygdala. Thus the infusion of a β-adrenoreceptor blocker into the amygdala blocks the memory-enhancing effects of glucocorticoids. It is critically important to note here that in each case these effects are obtained with pharmacological treatments that follow learning and memory training. This protocol maps on to the real world, since the elevated levels of stress hormones that facilitate learning and memory occur during and after the stressful event. Increases in the levels of stress hormones prior to an experience have the opposite effect, namely to impair learning and memory (see below).

In general, treatments that decrease the excitability of neurons in the basolateral complex of the amygdala, such as increased γ-aminobutyric acid (GABA) transmission, retard fear conditioning. Drugs such as the benzodiazepines, which serve to dampen the emotional and endocrine response to a stressor, impair learning and memory for information associated with the stressor. The effects of the benzodiazepines are mediated by the ability of these compounds to enhance the effects of GABA at the $GABA_A$-receptor site. There are a variety of intrinsic and extrinsic GABA-ergic amygdaloid projections which can constrain the acquisition (i.e. an 'amnesic' effect) and expression (i.e. an 'anxiolytic' effect) of conditioned fear responses. Thus infusion of $GABA_A$-receptor agonists into the amygdala impairs emotional learning and memory, while $GABA_A$-receptor blockers have exactly the opposite effect.

Whether or not one views the amygdala as the actual location of the memory trace is a matter of debate. What is not contested is the idea that amygdaloid stimulation enhances information storage. This appears to be the case even in learning paradigms that do not normally involve amygdaloid activation. Thus under normal circumstances spatial learning and memory are not dependent on amygdaloid function. However, amygdaloid stimulation facilitates spatial as well as procedural learning. Probably the best way to understand these effects is to assume that stressors activate the amygdala and that this activation can then serve to mediate certain forms of learning and memory, such as fear conditioning, as well as to facilitate other forms of learning.

CHRONIC STRESS AND AMYGDALOID FUNCTION

It should come as no surprise to learn that chronic stress enhances amygdaloid function. We are probably more familiar with this effect in the form of increased feelings of 'anxiety' and 'irritability'. First, chronic stress increases CRF gene expression in the amygdala, an effect that is mediated by adrenal glucocorticoids. CRF, of course, serves to enhance fear conditioning, and this effect is in part due to effects on ascending noradrenergic systems arising from structures such as the locus coeruleus. Secondly, there is also increased activity or 'sensitization' at the level of the noradrenergic neurons of the locus coeruleus. Repeated exposure to a stressor results in an enhanced noradrenergic response to any new form of stress – an effect that derives from the sensitization of the locus coeruleus neurons. The same is true for the HPA axis, and thus chronic stress substantially increases stress hormone responses to new forms of stress appearing on the horizon. Oddly enough, to the best of our knowledge the effects of chronic stress on emotional learning and memory have never been extensively examined. However, the activation of stress hormones in response to emotionally aversive stimuli is enhanced at every level during chronic stress. It would be surprising, therefore, if chronic stress did not serve to further enhance emotional learning and memory. Clearly this topic merits attention.

STRESS HORMONES AND COGNITIVE FUNCTION: THE HIPPOCAMPUS

In order to appreciate the effects of stress hormones on hippocampal function, it is critical to understand the timing of the stress hormone release and the emotional valence of the information (Figure 1). In the real world, the presence of a stressor diverts attention towards the stressor and away from unrelated information. This is the reality of cognitive function under duress. Two factors contribute to this state – first, the activation of the state of emotional arousal which corresponds to amygdaloid activity, and secondly, the increased levels of noradrenaline and glucocorticoids acting at multiple sites in the brain.

Activation of glucocorticoid receptors as a result of the high levels of glucocorticoids released by a stressful event enhances learning and memory for the stressful event. A common behavioral test for learning and memory in the rat is the Morris water maze, in which the animal is placed in a 'pool' of water and must find an escape platform located about 2 cm beneath the water level. Normally rats are proficient but reluctant swimmers, so the task is mildly aversive. Since the platform is not visible, the animal must rely on spatial cues in order to locate it. The administration of glucocorticoid-receptor antagonists following training in the Morris water maze facilitates the learning of escape routes. Similar findings have been obtained using other tests that are aversively motivated, such as the Porsholt swim test. However, the presence of elevated levels of glucocorticoids before the occurrence of an emotionally neutral or mildly stressful event can disrupt information processing. Thus the infusion of glucocorticoid-receptor agonists into the hippocampus prior to training in the Morris water maze disrupts learning and memory. The precise mechanisms underlying these actions of glucocorticoids are not entirely clear, but both effects involve glucocorticoid action at the level of the hippocampus. It has been proposed that the memory-impairing effects may also involve glucocorticoid action at the level of the frontal cortex, and may be associated with impaired working memory rather than information storage *per se*. Overall, the evidence both for facilitation of learning by stress hormones following an emotional aversive event and for the memory-impairing effects of stress hormones preceding an emotionally mild or neutral event is remarkably strong. The timing of the increase in stress hormone levels in relation to the event is critical (Figure 1).

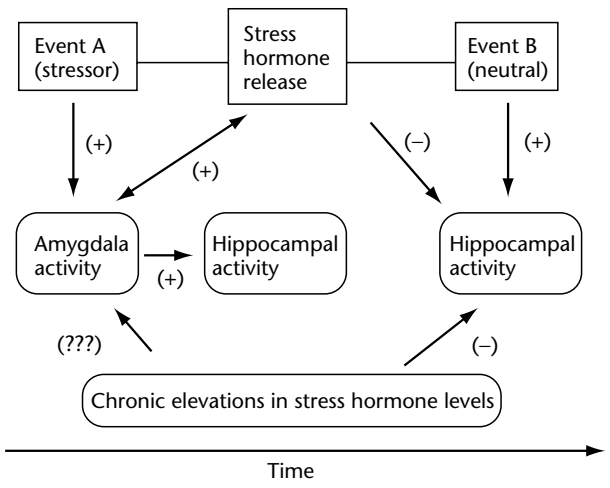


Figure 1. Timing and anatomical distribution of stress hormone release as a function of the emotional valence of stressful events.

The memory-impairing effects of exposure to stress or to elevated levels of glucocorticoids prior to learning and memory are reflected in the results of studies of hippocampal long-term potentiation (LTP). Hippocampal LTP is a model of synaptic plasticity induced by brief tetanic stimulation of glutaminergic afferents to either the CA1 or dentate gyrus region of the hippocampus, and is defined by the enhanced synaptic response to subsequent stimulation. Such LTP is thought to model the processes of synaptic plasticity that mediate learning. Stress or elevated levels of glucocorticoids prior to LTP induction significantly reduce the strength of LTP in the hippocampus. Glucocorticoids regulate the excitability of hippocampal neurons, and more recent studies using models of LTP have clarified the relevant mechanisms of action. High levels of glucocorticoids that result in glucocorticoid-receptor activation suppress LTP. The mechanism underlying this effect involves corticosteroid regulation of intracellular Ca^{2+} concentrations. Following a prolonged sequence of stimulation, hippocampal neurons exhibit a slow, enduring after-hyperpolarization (AHP) that is mediated by a slow Ca^{2+}-dependent K^{+}-conductance, serving to decrease hippocampal excitability. Ca^{2+} influx through voltage-gated channels, and thus the strength of AHP, is substantially increased with the occupation of glucocorticoid receptors. Interestingly, there are also several models of hippocampal synaptic sprouting, and in each case increased levels of glucocorticoids inhibit synapse formation (McEwen, 1999). Unfortunately, in all of these studies the emphasis has been on the effects of glucocorticoids applied prior to stimulation.

As a consequence, we know much more about the mechanisms associated with glucocorticoid-induced impairment of learning and memory than we do about how glucocorticoids act at the level of the hippocampus to enhance learning and memory.

The memory-impairing effects of glucocorticoids administered prior to cognitive testing have been well established in human populations. For example, the oral administration of 10 mg of hydrocortisone leads to a significant decrease in declarative and non-declarative memory performance as tested 60 minutes after hydrocortisone intake. This is of interest in view of the fact that studies report that the hippocampus is essential for a specific type of memory, notably declarative memory, while it is not essential for non-declarative memory. Declarative memory refers to the conscious or voluntary recollection of previous information, whereas non-declarative memory refers to the fact that experience changes the facility for recollection of previous information without affording conscious access to it (priming). Thus this somewhat specialized role of the hippocampus serves as the basis for specific hypotheses about the effects of acute administration of corticosteroids on human cognition (Lupien *et al*., 1998). The results demonstrated that subjects who received hydrocortisone treatment showed impaired performance on the declarative memory task but not on the non-declarative memory task. This suggests that cortisol interacts with hippocampal neurons to induce cognitive deficits.

Although most of the effects of exogenous administration of glucocorticoids on human psychophysical and cognitive processing have been investigated using hydrocortisone, some other studies have used other compounds and reached different conclusions. Poor performance (e.g. errors of commission, incorrectly identifying distractors as targets) on verbal memory tasks occurs in normal volunteers following the administration of prednisone (80 mg/day for 5 days). The general cognitive deficit described in this study involved the relative inability to discriminate previously presented relevant information (target) from irrelevant information (distractors) in a test of recognition memory. The authors concluded that exogenous administration of corticosteroids may reduce the encoding of meaningful stimuli and impair selective attention (i.e. reduce the ability to discriminate between relevant and irrelevant information) (Wolkowitz *et al*., 1990).

It appears that medium to high levels of circulating glucocorticoids may first affect the process of selective attention, thus impairing further explicit acquisition of information, while high or very high circulating levels of corticosteroids may affect explicit memory. This theory has recently been confirmed by a dose–response hydrocortisone infusion study in young normal controls in which infusion of a modest dose of hydrocortisone impaired selective attention capacity, while infusion of a higher dose impaired both selective attention and explicit memory. Interestingly, glucocorticoids also impair the retrieval of information from memory, and indeed these effects may have been confounded in studies where corticosteroids were applied prior to cognitive training, since these steroids will persist in the circulation for periods of up to 1 hour or more. Thus in subjects who had been trained in a declarative memory task, glucocorticoid treatment prior to recall testing impaired performance.

CHRONIC STRESS AND HIPPOCAMPAL FUNCTION

Acute increases in stress hormone levels can serve to enhance or impair learning and memory depending on the timing and emotional impact of the events. The effects of chronic stress have been studied generally within a much more limited set of test conditions. Chronic exposure to elevated levels of stress hormones markedly impairs learning and memory for emotionally neutral information, and the effects can be so severe as to mimic conditions of dementia (Lupien *et al*., 1998; McEwen, 1999).

It has been known for some time that corticosteroid treatment can induce a reversible psychotic condition (so-called 'steroid psychosis') in certain individuals. This condition also includes cognitive disturbances, such as loss of memory, attentional deficits, and problems with logical thinking. A 'dementia-like' syndrome (including attentional deficits and memory impairments) was found to occur frequently in a group of patients who were given high doses of corticosteroids for disorders not related to the central nervous system and who did not show prior evidence of psychosis. These findings are consistent with studies of Cushing's patients, in whom cortisol levels are markedly elevated. Cushing's patients consistently display attentional deficits and memory impairments, and the magnitude of those impairments correlates with plasma cortisol levels. Interestingly, the older subjects (> 45 years) were more seriously affected – they showed significantly greater memory impairments than did younger subjects. Importantly, these deficits were reversible, such that surgical intervention to correct the hypercortisolemic state

resulted in normalized cortisol levels and improved cognitive performance. Moreover, a recent study of Cushing's patients revealed that surgical treatment of hypercortisolism in Cushing's patients leads to a reversal of the glucocorticoid-induced hippocampal atrophy reported to occur in this population, with an average volume increase of 3.2%, and variations up to 10% in some patients (Starkman *et al*., 1999).

In addition, cognitive impairments (including attentional deficits and verbal and visual memory impairments) are often associated with clinical depression and are correlated with increased cortisol levels (Lupien *et al*., 1999). Again, both the increase in plasma cortisol levels and the magnitude of the cognitive disturbances are more severe in elderly patients, often resulting in a condition of so-called 'pseudodementia'. Importantly, in hypercortisolemic patients the cognitive disturbances are more pronounced. It is interesting that depression is also associated with hippocampal atrophy (McEwen, 1999). Considering the evidence for hippocampal atrophy in this population, it is tempting to speculate about the potential involvement of hippocampus dysfunction as a source of the cognitive dysfunction that is associated with the hypercorticoid state.

The results of these studies suggest that in humans, as in rodents, elevated glucocorticoid levels compromise hippocampal function and promote cognitive impairment. In a population of elderly but generally healthy human subjects, cognitive function was related to individual differences in HPA activity and hippocampal integrity (Lupien *et al*., 1998; McEwen, 1999). Cortisol levels were inversely correlated with performance on tests of episodic memory as well as with hippocampal volume. Predictably, the cognitive impairments (including deficits in spatial learning and memory) were associated with the degree of hippocampal atrophy (Lupien *et al*., 1998). These findings are consistent with data obtained for rodent populations, where prolonged exposure to elevated glucocorticoid levels can also compromise hippocampal integrity and certain forms of learning and memory. There is an impressive level of functional equivalence between the rodent and human models of hippocampal aging, and the apparent role of corticosteroids in promoting degeneration.

The assumption that was made on the basis of the earlier rodent studies was that all of this is entirely irreversible. However, more recent studies have emphasized the role of dendritic atrophy as a major component of the loss of hippocampal volume associated with prolonged exposure to elevated corticosteroid levels. Adaptive neuron loss may be a more minor component of the earlier phases of hippocampal degeneration than was originally thought. Instead, the primary effect of prolonged exposure to elevated stress or glucocorticoid levels is that of neuronal atrophy subsequent to dendritic culling. Exposure to chronically elevated glucocorticoid levels compromises hippocampal function to the point of endangering neuronal integrity and survival (McEwen, 1999). Once again the focus is at least in part on corticosteroid effects on Ca^{2+} homeostasis. Chronic stress in rodents and primates produces a reversible regression of apical dendrites. This effect is mimicked by high doses of corticosterone, and is blocked by inhibition of corticosterone synthesis or by Ca^{2+}-channel blockers. Interestingly, these effects parallel changes in the expression of a hippocampal-derived neurotrophic factor, namely brain-derived neurotrophic factor (BDNF) (McEwen, 1999). Chronic stress or chronically elevated corticosterone levels decrease BDNF mRNA expression as well as that of trkB, the BDNF receptor. In essence, the risk is associated with damaging levels of Ca^{2+} as well as with the withdrawal of trophic support, with the decreased production of factors such as BDNF.

SUMMARY

Stressful events impose a very important set of demands on the body, and this is reflected by changes in cognitive function. Although the mechanisms that underlie these effects are still incompletely understood, the existing data appear to provide strong support for the following conclusions. (1) Stress hormones mediate the enhanced learning and memory for emotionally salient information during and following periods of stress. (2) The exact nature of the effects of stress hormones on learning and memory depends on both timing and context. (3) Elevated levels of glucocorticoids following a period of stress can impair learning, consolidation and retrieval of emotionally neutral information, and such effects appear to involve the hippocampus. (4) Chronic stress and thus prolonged exposure to elevated glucocorticoid levels can produce severe impairment in both attention and memory.

References

Cahill L and McGaugh JL (1998) Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neuroscience* **21**: 294–299.

Davis M, Rainnie D and Cassell M (1994) Neurotransmission in the rat amygdala related to fear and anxiety. *Trends in Neuroscience* **17**: 208–214.

de Kloet ER, Vreugdenhil E, Oitzel MS and Joels M (1998) Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* **19**: 269–301.

LaBar KS, Gatenby JC, Gore JC, LeDoux JE and Phelps EA (1998) Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**: 937–945.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* **23**: 155–184.

Lupien S, DeLeon M, DeSanti S *et al*. (1998) Longitudinal increase in cortisol during human aging predicts hippocampal atrophy and memory deficits. *Nature (Neuroscience)* **1**: 59–65.

Lupien SJ, Nair NPV, Briere S *et al*. (1999) Increased cortisol levels and impaired cognition in human aging: implications for depression and dementia in later life. *Reviews in Neuroscience* **10**: 117–139.

McEwen BS (1999) Stress and hippocampal plasticity. *Annual Review of Neuroscience* **22**: 105–122.

Starkman MN, Giordani B, Gebarski SS, Berent S, Schork MA and Schteingart DE (1999) Decrease in cortisol reverses human hippocampal atrophy following treatment of Cushing's disease. *Biology of Psychiatry* **46**: 1595–1602.

Valentino RJ, Curtis AL, Page ME, Pavcovich LA and Florin-Lechner SM (1998) Activation of the locus ceruleus brain noradrenergic system during stress: circuitry, consequences and regulation. *Advances in Pharmacology* **42**: 781–784.

Wolkowitz OM, Reus VI, Weingartner H *et al*. (1990) Cognitive effects of corticosteroids. *American Journal of Psychiatry* **147**: 1297–1303.

Further Reading

Kim JJ and Diamon DM (2002) The stressed hippocampus, synaptic plasticity and lost memories. *Nature Reviews Neuroscience* **3**: 435–462.

Lupien SJ and Lepage M (2001) Stress, memory, and the hippocampus: can't live with it, can't live without it. *Behavioral Brain Research* **127**: 137–158.

McEwen BS (2001) Plasticity of the hippocampus: adaptation to chronic stress and allostatic load. *Annals of the New York Academy of Sciences* **933**: 265–277.

Stroke

Introductory article

Fernando Gonzales-Portillo, University of Arizona, Tucson, Arizona, USA
Bruce M Coull, University of Arizona, Tucson, Arizona, USA

CONTENTS

Stroke is a general term for loss of brain function (persisting for more than 24 h) caused by loss of blood supply to the brain, or by hemorrhage within or around the brain.

INTRODUCTION

Stroke is a broad, generic term that encompasses a vast number of causes for loss of brain function due to vascular dysfunction. It is defined as the abrupt onset of focal or global neurologic symptoms caused by a loss of blood supply to the brain (or spinal cord) or hemorrhage within or around the brain that is a result of diseases of the cerebral blood vessels, the heart or blood elements. Although some strokes are 'silent', stroke is commonly distinguished from other neurologic disorders by the sudden development of symptoms that almost always start with a focal loss of neurological function, such as paralysis, loss of speech, vision, or balance and coordination. Sometimes the symptoms of stroke resolve quickly, and if the focal neurological deficit lasts less than 24 h the episode is termed a 'transient ischemic attack' (TIA). It is important to recognize TIAs since they are a harbinger of future strokes that might be disabling. By clinical convention, if the neurologic symptoms due to vascular causes continue for more than 24 h then the event is diagnosed as a stroke.

Stroke is a major public health problem worldwide. It is the third leading cause of death and the leading cause of major disability in the elderly in the USA. The annual incidence of stroke is estimated to be approximately 750 000. Stroke is the most common neurologic disorder of adults in the USA. More often disabling than fatal, stroke has enormous economic, social and psychological costs, measured in terms of both healthcare expenses and loss of productivity. Major strides have been made in understanding the epidemiology, etiology and pathogenesis of stroke and have led to new approaches to diagnosis and treatment.

Strokes can be classified into two principal categories based upon whether the stroke is caused by a blocked blood vessel (ischemic stroke) or a ruptured blood vessel (hemorrhagic stroke). Hemorrhagic strokes are further divided into categories based on the site and origin of the blood: subarachnoid hemorrhage, when the bleeding originates in the subarachnoid spaces surrounding the brain, and intracerebral hemorrhage, when the bleeding is into the brain parenchyma. Ischemic strokes are classified according to the mechanism or cause of the vascular occlusion.

RISK FACTORS

Although stroke is an illness that mostly affects the elderly, a stroke can occur at any age. The risk of stroke increases with age and in the presence of certain risk factors for stroke. African–Americans are more likely than white Americans to experience stroke. Modifiable risk factors for ischemic stroke include the following:

- hypertension
- cardiac disease (atrial fibrillation, myocardial infarction)
- diabetes mellitus
- cigarette smoking
- hypercholesterolemia
- heavy alcohol use
- TIA
- hyperhomocysteinemia
- carotid stenosis.

Nonmodifiable risk factors include:

- age
- male gender
- hereditary or familial factors

Table 1. Incidence of ischemic stroke in people at high risk

Risk factor	*Stroke rate (%) per year*
General population aged 70 years	0.6
Prior MI	1.5
Asymptomatic bruit	1.5
Asymptomatic carotid stenosis	2.0
Nonvalvular AF	5.0
Nonvalvular AF with TIA or stroke	12.0
TIA	6.0
TIA with > 70% carotid stenosis	13.0
Prior ischemic stroke	12.0 (first year after initial stroke)

AF, atrial fibrillation; MI, myocardial infarction; TIA, transient ischemic attack.

- race or ethnicity (e.g. stroke is more common in African–Americans)
- geographic location (southeastern USA is known as the 'stroke belt').

Risk factors for hemorrhagic stroke differ somewhat from those for ischemic stroke and include female gender, hypertension, cigarette smoking and increasing age.

Some risk factors for stroke increase the chance of stroke more than others (Table 1). Combinations of risk factors have an additive or multiplicative effect on risk. For example, a woman with hypertension and atrial fibrillation who is over 75 years old may have more than 15 times the risk of stroke compared with a woman of the same age without these risk factors.

ISCHEMIC STROKE

Ischemic strokes account for 80–85% of all strokes. A cerebral artery that supplies part of the brain is suddenly occluded, with an immediate loss of the function of the affected brain region. The onset of symptoms is immediate because the brain cells cannot effectively store metabolic energy sources and depend upon a continuous supply of oxygen and glucose in order to function. If the vascular occlusion persists for more than a few minutes the brain tissue begins to die, leading to a permanent neurologic deficit.

Ischemic strokes are usually divided into five subtypes based on the mechanism or cause of the vascular occlusion:

1. Primary thrombotic occlusion of a large cerebral artery occurs in 15–40% of ischemic strokes, usually in a vessel already partially occluded by atherosclerosis, for instance the carotid artery or the basilar artery.
2. Occlusion of a small artery or arteriole (lacunar infarct) occurs in 15–30% of cases. It is characterized by hyaline thickening of small penetrating arteries of the brain (lipohyalinosis) and is most commonly seen in people with diabetes mellitus and hypertension. The occlusion of these vessels results in small, deep, often cystic infarcts, appearing as holes (lacunae). These infarcts are often asymptomatic but may cause certain clinical syndromes, such as pure motor hemiparesis, pure sensory stroke, clumsy hand–dysarthria, ataxia hemiparesis, and others. Not infrequently the arteries may also be the target of embolism, and at times the occlusion is due to atherosclerotic plaque in the large vessel from which it arises.
3. Occlusion of a cerebral vessel by a clot that arises from a distant source, usually the heart (cardioembolism), occurs in 15–30% of cases. The heart is the most common source of the embolic material (mural thrombus formation in the setting of atrial fibrillation or myocardial infarction, patent foramen ovale, prosthetic valves, septic endocarditis, etc.). Less commonly embolism arises from ulcerated rupture of atherosclerotic plaques in the aortic arch or at the origin of the great vessels.
4. Uncommon causes of ischemic stroke include cerebral vein thrombosis, inflammation of the blood vessel (vasculitis), paradoxical embolism, polycythemia vera, subclavian steal, sickle cell anemia, and cocaine abuse, among others.
5. Undetermined cause may account for as many as 40% of cases.

HEMORRHAGIC STROKE

Hemorrhagic stroke (subarachnoid and intracerebral) accounts for 15–20% of all strokes, but this figure depends on race and geography, with greater relative frequencies of hemorrhagic stroke observed in China and Japan than in the USA. Bleeding damages the brain by cutting off connecting pathways and by causing localized or generalized pressure injury to brain tissue.

Subarachnoid hemorrhage is most commonly caused by leakage of blood from a ruptured cerebral (berry) aneurysm. This condition is a common, life-threatening medical and surgical emergency and is potentially curable. The signs and symptoms of aneurysm rupture vary according to the severity of the bleed and location of the aneurysm. The classical presentation is an abrupt onset of the worst headache of one's life followed by altered consciousness, meningismus, photophobia, nausea and vomiting.

Intracerebral hemorrhage is characterized by bleeding into the brain parenchyma, usually from

a small penetrating artery. Hypertension has been implicated as the cause of weakening in the arteriole walls and the formation of microaneurysm. Among elderly, nonhypertensive patients with recurrent lobar hemorrhages, amyloid angiopathy has been implicated as an important cause. Other causes include arteriovenous malformations, bleeding disorders, excessive anticoagulation, trauma, tumors, moyamoya disease, cavernous malformations and illicit drug use. The most common sites are the putamen, caudate, pons, cerebellum, thalamus or deep white matter.

The clinical picture depends on the location and size of the hematoma. Headache, vomiting and evolution of motor or sensory signs characterize the usual clinical presentation over minutes to hours. Consciousness is sometimes impaired at the start, and this often becomes a prominent feature in the first 24–48 h with moderate to large hematomas.

PRIMARY SYMPTOMS OF STROKE

The clinical presentation of stroke is diverse and depends upon the type of stroke and the region of the brain affected. Most patients who suffer a stroke develop an obvious loss of motor or sensory function on one side of the body. The motor symptoms may appear as a profound weakness, paresis, or a loss of coordination termed ataxia. Any sensory system may be involved in stroke, but the somatosensory and visual systems are most commonly affected, with sparing of smell, taste and hearing. People with hemorrhagic strokes frequently complain of severe headache, and may vomit, experience seizures or have an altered mental status, including coma.

COGNITIVE CONSEQUENCES

Cognitive and related higher cortical functions including language are often impaired by stroke, especially when the stroke is large and affects the cerebral cortex. Overall, about 30% of stroke survivors will develop dementia within 3 years. This is more likely to occur in elderly individuals, and in those with previous 'silent strokes', cerebral atrophy or diabetes mellitus. Among individuals with dementia, in about 25% the dementia is due to an underlying vascular cause. The loss in cognitive abilities after a stroke can be obvious, as with aphasia, characterized by an inability to produce and/or understand language including reading and writing. Lesions that disconnect or isolate motor areas can produce apraxias in which voluntary movements are impaired despite an intact motor output function. Such individuals often can perform complex involuntary movements such as scratching the nose, despite an inability to perform the same movement to command. Strokes that affect the prefrontal areas can produce impulsive behavior, poor planning, and disinhibition of emotional control with outbursts of crying or inappropriate laughter. Such behavior can be disruptive to family and caregivers and may impair rehabilitation efforts. Individuals with widespread multifocal small strokes often develop 'sundowning' – the manifestation of agitation, delirium with hallucinations and confusion towards nightfall. Some patients with nondominant parietal lobe lesions deny the existence of illness, paralysis or dysfunction even when severe, or fail to recognize their own body parts: these are forms of agnosia.

One important consequence of stroke that is frequently overlooked by both caregivers and healthcare providers after stroke is depression. The onset of depressed mood and flat affect is often mistaken as a reaction to the neurological deficit produced by the stroke rather than an intrinsic depression. Within 18 months of stroke, up 45% of persons recovering from stroke will experience depression. In about half of these, the depression is considered serious enough to affect outcome or quality of life.

OUTCOME

Outcome after a stroke depends upon the type and severity of stroke. About 25–30% of individuals with ischemic stroke recover with minimal or no deficit and roughly 15% will die within a year. Unfortunately, over 50% of stroke survivors will have a permanent severe deficit. The immediate period after ischemic stroke (the first 30 days) carries the greatest risk of death, with fatality rates ranging from 8% to 12%, and the greatest risk of recurrence, with a range from 3% to 10%. Death is more likely to be caused by cardiopulmonary complications than by the stroke itself. For stroke survivors, recurrent stroke within 2 years is the most frequent occurrence and is responsible for subsequent major stroke morbidity and mortality. The hemorrhagic strokes have worse case fatalities and overall prognosis, mortality ranging from 30% to 80% for intracerebral hemorrhage and 20% to 50% for subarachnoid hemorrhage.

TREATMENT

Stroke prevention is the mainstay of treatment since it is far better to keep a stroke from happening than to effect a cure once it has happened. Most primary

preventive strategies focus on treatment of modifiable risk factors such as hypertension, diabetes and smoking, and on dietary and habit interventions to improve general cardiovascular risk. Anticoagulant drugs such as warfarin are useful in certain high-risk patients with atrial fibrillation. With effective control of stroke risk factors, up to two-thirds of strokes could be prevented annually. The use of low-dose aspirin and related antiaggregate treatments, estrogen replacement in postmenopausal women and the application of carotid endarterectomy to asymptomatic individuals remain somewhat uncertain or controversial.

Ischemic Stroke

Acute medical treatment for stroke is designed to prevent or minimize ischemic brain infarction, to optimize functional recovery and to avert stroke recurrence. The current treatment for acute ischemic stroke is largely based on the results of the National Institute of Neurological Disease and Stroke (NINDS) tissue plasminogen activator (tPA) trial published in 1995, demonstrating a beneficial effect of tPA given within 3 h of symptom onset for some individuals; stroke has thus become a neurological emergency. The NINDS results indicate that patients who receive tPA are 30% more likely to have a full recovery or minimal neurological deficit at 3 months after stroke. However, the administration of this thrombolytic therapy does carry a significant 10-fold increase in risk of brain hemorrhage. In order to decrease the incidence of hemorrhagic conversion after tPA, the treatment protocol has to be followed strictly.

In patients who are not eligible for tPA therapy, a variety of treatments including antithrombotic agents can be considered. Several anticoagulant and antiplatelet drugs have been evaluated in clinical trials. The use of antithrombotic agents in an attempt to reduce the risk of stroke progression is complicated by the existence of different stroke etiologic subtypes. The therapeutic approach to the acute ischemic stroke should take into account the underlying pathophysiologic mechanism, but unfortunately, in the early hours of presentation, this is often unknown.

Platelet antiaggregation agents have been shown to prevent recurrent strokes. Aspirin is the most widely studied of these agents, and until recently was the only drug used broadly for this purpose. Results of clinical trials indicate that ticlopidine, clopidogrel, and dipyridamole in combination with low-dose aspirin, are also effective for stroke prevention. Anticoagulant therapy substantially reduces the risk of cardiac embolism, but the evidence supporting the use of it in patients with acute cardioembolic stroke is based upon limited data. Studies of anticoagulants for acute stroke therapy have failed to show a significant reduction in stroke progression, mortality, risk of stroke recurrence, or favorable outcomes. There have been several studies with neuroprotective agents, but unfortunately none of these drugs has shown any benefit in stroke reduction, morbidity and mortality.

Hemorrhagic Stroke

Hemorrhagic strokes are treated according to the cause of the hemorrhage. For intracerebral hemorrhage the medical treatment is supportive and directed at controlling elevated intracranial pressure and maintaining respiratory and other vital functions. In some instances, such as cerebellar or large intracerebral hemorrhages, surgical decompression is required. Surgical evacuation of a hematoma is considered for a lobar or subcortical white-matter hemorrhage if the patient is neurologically deteriorating. In subarachnoid hemorrhage due to a ruptured aneurysm, surgical extirpation and clipping or endovascular coiling of the aneurysm is indicated in order to prevent recurrent bleeding.

Further Reading

Bossier MG (1997) Aspirin or heparin immediately after a stroke? [see comments]. *Lancet* **349**: 1564–1565.

CAPRIE Steering Committee (1996) A randomized, blinded, trial of clopidogrel versus aspirin in patients at risk of ischemic events (CAPRIE). *Lancet* **348**: 1329–1339.

Carson AJ, MacHale S, Allen K *et al*. (2000) Depression after stroke and lesion location: a systematic review. *Lancet* **356**: 122–126.

Diener H, Cunha L, Forbes C *et al*. (1996) European Stroke Prevention Study 2. Dipyramidole and acetylsalicylic acid in the secondary prevention of stroke. *Journal of Neurological Sciences* **143**: 1–13.

European Atrial Fibrillation Trial Study Group (1995) Optimal oral anticoagulant therapy in patients with nonrheumatic atrial fibrillation and recent cerebral ischemia. *New England Journal of Medicine* **333**: 5–10.

European Carotid Surgery Trial (1998) Randomized trial of endarectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). *Lancet* **351**: 1379–1387.

Feinberg WM (1996) Primary and secondary stroke prevention. *Current Opinion in Neurology* **9**: 46–52.

Gorelick FB (1997) Stroke prevention: windows of opportunity and failed expectations? A discussion of

modifiable cardiovascular risk factors and a prevention proposal. *Neuroepidemiology* **16**: 163–173.

Hebert PR, Gaziano JM, Chan KS and Hennekens CH (1997) Cholesterol lowering with statin drugs, risk of stroke and total mortality. An overview of randomized trials. *JAMA* **278**: 313–321.

Henon H, Durieu I, Guerouaou D *et al*. (2001) Poststroke dementia, incidence and relationship to poststroke cognitive decline. *Neurology* **57**: 1216–1222.

International Stroke Trial Collaborative Group (1997) IST: randomized placebo-controlled trial of early aspirin use in 20 000 patients with acute ischaemic stroke. CAST (Chinese Acute Stroke Trial) Collaborative Group. *Lancet* **349**: 1641–1649.

International Stroke Trial Collaborative Group (1997) The International Stroke Trial (IST): a randomized trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group [comments]. *Lancet* **349**: 1569–1581.

Kreisler A, Godefroy O, Delmaire C *et al*. (2000) The anatomy of aphasia revisited. *Neurology* **54**: 1117–1123.

Singh A, Black SE, Herrmann N *et al*. (2000) Functional and neuroanatomic correlations in poststroke depression: the Sunnybrook Stroke Study. *Stroke* **31**: 637–644.

SPAF Investigators (1996) Adjusted-dose warfarin versus low-intensity, fixed-dose warfarin plus aspirin for high-risk patients with atrial fibrillation: Stroke Prevention in Atrial Fibrillation III randomised clinical trial [see comments]. Comment in: ACP J Club 1997 Jan-Feb;126 (1):1 *Lancet* **348**: 633–638.

Staessen JA, Fagard R, Thijs L *et al*. (1997) Randomised double-blind comparison of placebo and active treatment for older patients with isolated systolic hypertension. The Systolic Hypertension in Europe (Syst-Eur) Trial Investigators. *Lancet* **350**: 757–764.

Welch GN and Loscalzo J (1998) Homocysteine and atherothrombosis. *New England Journal of Medicine* **338**: 1042–1050.

Synapse

Intermediate article

Andrew Matus, Friedrich Miescher Institute, Basel, Switzerland
Michael Frotscher, University of Freiburg, Freiburg, Germany

CONTENTS

Synapses are specialized structures in neuronal circuits where signals are transmitted from one nerve cell to the next. Synaptic plasticity, involving changes in the efficiency of signal transmission and rearrangement of connections between nerve cells, is thought to underlie brain learning and memory.

INTRODUCTION

Information entering the nervous system through the sense organs is processed by vast arrays of neural circuits where it is progressively transformed into the coordinated muscle movements that make up behavior. At the same time, memory traces are laid down within these same circuits allowing animals to modify their behavior based on past experience. The function of these circuits depends on specific patterns of neuronal connections made by synapses and it is their properties and organization that lie at the heart of nervous system function.

The concept of the synapse was introduced at the end of the nineteenth century by the English physiologist Charles Sherrington, and was based on a combination of ideas from physiology and anatomy. Crucial in its inception was Sherrington's observation that reflex activity in the spinal cord always travels from the sensory fibers to the motor nerves and never in the reverse direction. This suggested the existence of a valve-like function somewhere within the reflex circuit that guaranteed the unidirectional transfer of excitation. Another key insight came from the Spanish anatomist Santiago Ramón y Cajal who argued persuasively that neuronal circuits are not continuous, as had been previously widely thought, but are composed of individual nerve cells whose axonal processes terminate in 'free' endings that are in close contact with the dendrites of other neurons. It was the combination of Ramón y Cajal's histological observations and his own physiological data, both pointing to a unidirectional mode of signal transmission in neuronal circuits, that led Sherrington to propose a specific structure at the point of contact – the synapse – that embodied these properties. It took more than fifty years, and the development of the electron microscope with the power to resolve structures on the scale of nanometers, to confirm the existence of such structures, but when they were finally visualized, they were seen to embody precisely the principles of operation that Sherrington had predicted (Figure 1).

FUNCTIONAL PROPERTIES OF SYNAPSES

Synapses of the kind Sherrington envisaged, which transmit signals between neurons using a chemical 'neurotransmitter' substance, dominate signal transmission in vertebrate nervous systems. A second type of 'electrical' synapse, which operates by direct passage of charged ions through minute pores between two neurons, is less common and will not be dealt with here.

Chemical synapses typically consist of two distinct components: a presynaptic terminal or bouton, which releases the neurotransmitter in response to the arrival of a nerve impulse, and a postsynaptic element where specialized receptor proteins on the neuronal surface detect the neurotransmitter and convert it into a signal inside the second cell (Figure 1). It is this differentiation into pre- and postsynaptic elements that provides the chemical synapse with its distinctive property of transmitting information in only one direction: from the presynaptic terminal where the neurotransmitter is released, to the postsynaptic element where it is detected. One important consequence of this arrangement is that it provides for progression, guaranteeing that signals flow from the sense

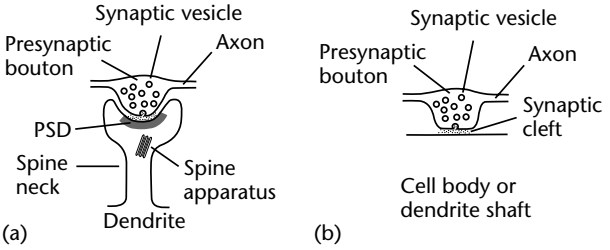


Figure 1. Excitatory and inhibitory synapses. (a) The excitatory synapse is made onto a dendritic spine with a prominent postsynaptic density (PSD) attached to the postsynaptic membrane in the region of the synaptic junction. (b) Inhibitory synapses are typically made directly onto cell bodies or dendrite shafts and lack a prominent postsynaptic membrane specialization. The synaptic cleft is where neurotransmitter molecules (dots) are released from the presynaptic bouton to interact with receptors embedded in the postsynaptic membrane. In both cases presynaptically released neurotransmitter activates postsynaptically located receptors, guaranteeing that signal transmission is unidirectional (from presynaptic to postsynaptic).

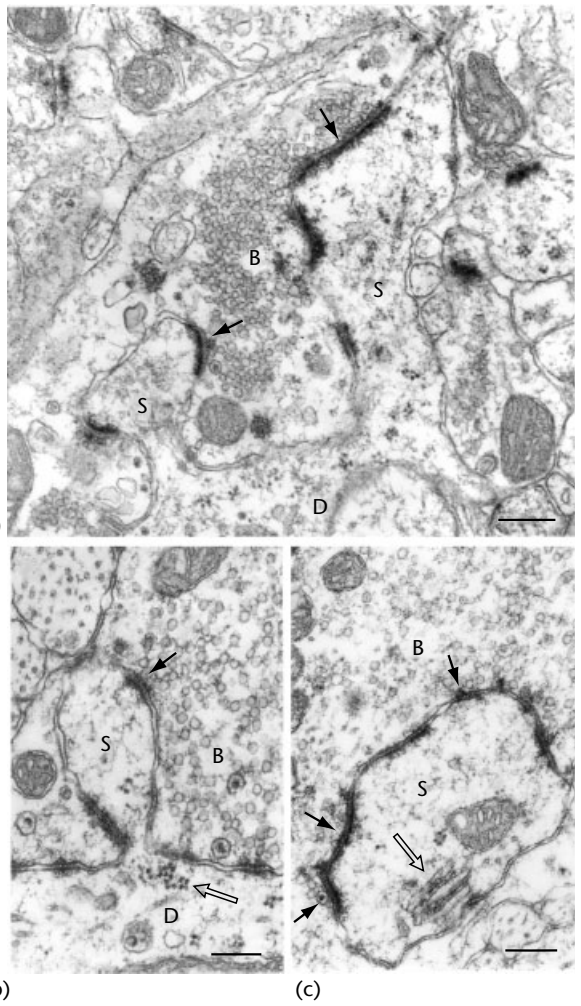


Figure 2. Three examples of giant boutons (B) formed by excitatory mossy fiber axons in the hippocampus which establish asymmetric synaptic contacts (arrows) with large dendritic spines (S). The dendritic shafts (D) are indicated in (a) and (b). The open arrow in (b) indicates ribosomes at the base of the spine, while the open arrow in (c) points to a spine apparatus. Bar, (a), (c) 0.25 μm; (b) 0.2 μm.

organs into the central nervous system and ultimately reemerge from the spinal cord to reach the muscles. It also provides for a wide range of logical operations, such as signal feedback and convergence, which would operate inefficiently or not at all in a bidirectionally transmitting network. These operations are used repeatedly in neuronal circuits throughout the brain and are crucial to their analytical and integrative capabilities.

An important feature of most chemical synapses is that the presynaptic and postsynaptic elements are closely joined by a specialized intercellular junction (marked by arrows in Figures 2 and 3), which provides for specificity of signal transfer within a compact and complex network of axons, dendrites, and synapses. The extent of this problem can be appreciated by considering that 1 mm^3 of gray matter in the cat visual cortex contains in the region of 300 million synapses and estimates for the human cortex range up to one billion synapses per mm^3 (Shepherd and Koch, 1998). With such an enormous density of synapses simultaneously chattering to one another it is crucial that the specificity of transmission be maintained, and it is difficult to imagine a better means of doing so than by 'gluing' the transmitting and receiving elements to one another via an intercellular junction.

Despite this morphological specialization, the small neurotransmitter molecules that convey the signal across the synaptic junction diffuse readily, raising the possibility that they will inadvertently reach and stimulate receptors at other synapses. To avoid this, the neurotransmitter is released into a narrow gap – the synaptic cleft – between the pre- and postsynaptic elements, and its action is rapidly terminated by active reuptake into the neuron or into processes of the glial cells which surround the synapse. Even so, it is thought that spillover of neurotransmitter from one synapse to another may lead to 'cross-talk' between synapses and influence signal processing in the dense central nervous system neuropil (Rusakov *et al*., 1999). Indeed, the presynaptic terminals of axons

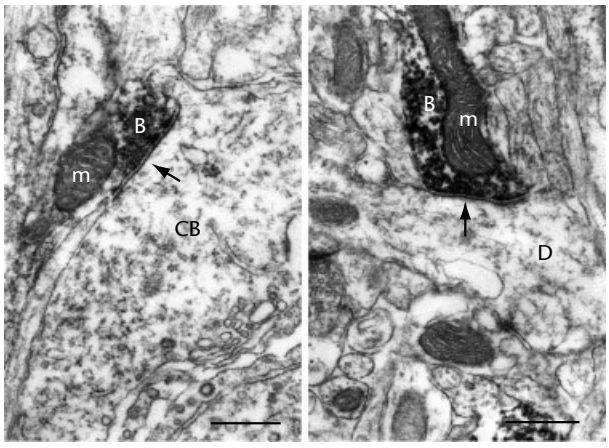


Figure 3. Two examples of inhibitory synapses, identified as such by immunoreactive staining (dark precipitate) in their presynaptic boutons (B) for glutamate decarboxylase, the enzyme synthesizing the inhibitory neurotransmitter γ-aminobutyric acid. The two boutons establish symmetric synaptic contacts (arrows) with a neuronal cell body (CB) in example (a) and a dendritic shaft (D) in (b). Mitochondria (m) are commonly found in axonal boutons. Bar, 0.25 μm.

releasing neurotransmitters such as noradrenaline (norepinephrine), which have a modulatory influence on dendrites spread over wide target areas, seem in general not to make point-to-point contacts.

TYPES OF SYNAPSE

It is also important to recall that while small, focal synaptic contacts of the kind shown in Figure 1 are the rule for the vast majority of neurons in the central nervous system, many other arrangements also occur. Often these are morphologically specialized in ways that are plainly tailored to their function. Examples include giant synapses with multiple independent sites of neurotransmitter release and reception, such as the calyx of Held synapses in the medial nucleus of the trapezoid body, where this reduplication is thought to impart reliability to a circuit whose function is to transmit auditory information with high fidelity. Another example is the giant mossy fiber synapse in the hippocampus, illustrated in Figure 2. Furthermore, whereas in most cases the presynaptic component is part of an axon and the postsynaptic component is part of a dendrite, other arrangements, including axoaxonic and dendodendritic synapses, such as those formed between the dendrites of mitral and granule cell neurons in the olfactory bulb, also occur as solutions to specific processing needs.

Excitatory and Inhibitory Synapses

Major distinctions exist between excitatory synapses, whose activity increases the likelihood of a cell firing by depolarizing the postsynaptic membrane, and inhibitory synapses, which do the opposite. The most salient difference between the two lies in the types of neurotransmitters and neurotransmitter receptors they use. For the vast majority of excitatory synapses in the central nervous system the neurotransmitter is the amino acid glutamate, whereas the corresponding major neurotransmitter for inhibitory synapses is γ-aminobutyric acid (GABA) or, in the spinal cord and brainstem, glycine. At both excitatory and inhibitory synapses, two principal types of neurotransmitter receptors can be distinguished: ligand-gated ion channels serving fast synaptic transmission, and receptors coupled to guanine nucleotide-binding regulatory protein (G protein), whose activation leads to a delayed response in the postsynaptic cell.

The pharmacological distinction between excitatory and inhibitory synapses is accompanied by significant structural differences (Figures 2 and 3). One of these occurs at the synaptic junction, which at excitatory synapses (Figure 1(b)) is marked by a prominent disk-shaped structure, the postsynaptic density (PSD), embedded in the postsynaptic junctional membrane. This postsynaptic specialization is far less prominent at inhibitory synapses, where instead the presynaptic and postsynaptic membrane specializations are of similar thickness (Figure 3). Because of this, inhibitory contacts have been referred to as *symmetric synapses*, whereas excitatory synapses, with their postsynaptic density being more prominent than the presynaptic membrane specialization (Figure 2), have been described as *asymmetric contacts*. The two types are still often referred to as Gray type I (asymmetric synapses) and type II (symmetric synapses), based on their original description by the British electron-microscopist George Gray. The presynaptic boutons illustrated in Figure 3 may be identified as terminals of inhibitory neurons, not only by the symmetric contacts they form but also by their immunostaining for glutamate decarboxylase, the synthesizing enzyme for the inhibitory neurotransmitter GABA. The morphological distinction between asymmetric (excitatory) and symmetric (inhibitory) synapses is correlated with molecular differences, including the virtual absence from

inhibitory synapses of a complex enzyme, calcium–calmodulin kinase II, which is the major protein component of the PSDs of excitatory synapses where it has been postulated to have a role in memory formation (see below).

Advances in molecular biological techniques have led to the identification of a growing number of molecular components that make up these junctional structures, particularly at excitatory synapses. On the presynaptic side, giant modular proteins interact through multiple binding sites with different specificities to build up an active zone which defines the site where synaptic vesicles dock and fuse with the presynaptic membrane to release neurotransmitter in response to an action potential. Postsynaptically, the PSD structure is assembled from a wide range of modular proteins whose binding sites marshal together neurotransmitter receptors and ion channels and couple them to other active components, including other proteins that modulate their function. Much effort is being put into understanding how these pre- and postsynaptic molecular arrays are assembled during development, and how their arrangement and functional state may be modifed in response to activity to modulate the efficiency of synaptic transmission (Sheng, 2001). Rearrangements within these multimolecular complexes, triggered by alterations in neuronal firing rates, may represent an intermediate step in changes of circuit connectivity that underlie learning and memory.

Levels of Organization and Signal Integration

Excitatory and inhibitory synapses differ also in where and how they contact the dendrite of the postsynaptic cell. Excitatory synaptic contacts are concentrated on dendrites and occur far away from as well as close to the cell body. As a result, their influence on the cell's overall state of excitation, and on the rate at which action potentials are generated at the cell body, may be influenced by their location. In contrast, inhibitory synapses are more prominent close to and on the cell body (Figure 3(a)) where their membrane hyperpolarizing influence can act as a 'choke' on action potential generation. This arrangement is particularly important in functions such as surround inhibition, where collateral inhibitory transmission triggered by an active neuron suppresses the firing of its neighbors, thus increasing signal salience relative to background activity. Even the intimacy of the contact made by the two types differs significantly.

Whereas inhibitory synapses are formed directly onto cell bodies (Figure 3(a)) and shafts of dendrites (Figure 3(b)), the vast majority of excitatory synapses, estimated at approximately 90% of those in the brain, are made onto dendritic spines – tiny protrusions, about 1 μm long, from the dendrite surface (Figure 2). Typically a dendritic spine consists of an expanded head, where contact with the presynaptic axonal bouton is made, connected to the dendrite shaft by a narrow neck (Figure 2(a), (c)). Synaptic transmission occurs at the tip of the head (Figure 1(a)), so that in effect the rest of the spine forms a separate compartment interposed between the axon terminal and the dendrite shaft. This compartment often contains what is called a *spine apparatus* (Figures 1(a) and 2(c)), an organelle thought to be involved in sequestering calcium ions. Imaging studies on living neurons using Ca^{2+}-sensitive fluorescent dyes have shown that postsynaptic calcium ion fluxes elicited by low levels of afferent stimulation remain restricted to the spine itself, whereas the larger ion flows produced by higher levels of stimulation can invade the dendrite shaft and neighboring dendritic spines.

Many neurons in the central nervous system receive thousands of synaptic inputs onto dendritic spines, arising from diverse sources, with the figure increasing to tens of thousands for large pyramidal neurons in the cerebral cortex and ranging up to 200 000 on the dendrites of giant Purkinje cells in the cerebellar cortex. The ability of these large numbers of dendritic spines on a single neuron to regulate the access of postsynaptic signaling molecules to the dendrite shaft is widely thought to play a major part in integrating the multiple synaptic signals impinging on a dendrite from different sources.

COMPUTATION IN DENDRITES

Precisely how synaptic signals are integrated to modulate neuronal firing rates is the subject of intensive research that has benefited from increasingly detailed biophysical analyses of electrical activity in dendrites (Hausser *et al*., 2000). These studies have shown that the active properties of dendrites vary as a function of distance from the cell body, so that distant synapses may have a disproportionately large influence on the level of membrane polarization at the cell body, which determines the rate of cell firing. The source of this enhanced efficacy of distal synapses is not yet clear, but may involve higher densities of postsynaptic

neurotransmitter receptors at these sites compared with synapses closer to the cell body.

Another important emerging concept in dendrite function concerns the possibility that the dendritic trees of large neurons are partitioned into discrete zones that can either integrate synaptic inputs independently or combine to give a different, more global output. This has been investigated in large apical dendrites of pyramidal neurons in layer 5 of the cerebral cortex, where synaptic activity in the terminal tuft of dendritic branches in upper layers of the cortex may operate independently of electrical activity in lower levels proximal to the cell body. In this way synaptic influences impinging on basal and proximal dendrite branches may be processed differently from those arriving distally, suggesting that different forms of neuronal computation may occur in different layers of the cortex. Particularly significant is that upper layers of the cortex are rich in long-distance intrinsic connections between pyramidal neurons. This marks a significant shift in thinking from a traditional model of neuronal processing in which the summation of synaptic inputs acts by determining the firing rates of individual neurons, to one in which data-processing operations are carried out by the integrated activity of vast arrays of synapses scattered over the dendrites of many cells. If shown to be a general property of cortical circuits, such a model of 'subthreshold' processing could have significant implications for the processing of multimodal data between widely separated cortical areas.

A further insight from these new data on dendrite function is that the action of distal and proximal dendritic processing may be coupled by calcium-based action potentials within the dendrite itself. Improved techniques for measuring electrical signals in individual dendrites have shown the existence of back-propagating action potentials that originate in the cell body and invade the dendrite. Whether such coupling occurs depends on the relative timing of distal and proximal synaptic events within a timescale of milliseconds. Because of its precisions and association with long-distance intracortical connections, this time-sensitive coupling between cortical layers provides a potential means of synchronizing activities in different cortical regions, a cellular mechanism with obvious implications for cognitive coupling at the psychological level.

SYNAPTIC PLASTICITY

Perhaps the major outstanding question regarding synaptic function is what events mediate the storage of information underlying learning and memory, about which – despite intense efforts – little is known. The most widely investigated memory-related phenomenon is long-term potentiation (LTP), in which an episode of high-frequency stimulation (a tetanus) or the delivery of paired stimuli leads to a long-lasting increase in the synaptic response to a subsequent test stimulus. Conversely, repetitive stimulation at low frequencies can give rise to a long-lasting lowering of synaptic responses known as long-term depression (LTD). In contrast to transitory forms of synaptic plasticity such as post-tetanic potentiation (a short-lived facilitation of synaptic responsiveness that occurs directly following a tetanus), LTP is accompanied by changes in protein synthesis. It has been suggested that this may occur locally at dendritic spines because groups of ribosomes (polyribosomes), the organelles on which proteins are synthesized, are frequently observed at the base of a spine (Figure 2(b)).

Behavioral testing in several sensory systems has associated LTP and LTD with learning and memory and with the adaptation of animals to novel environments (Martin *et al.*, 2000). Various cellular mechanisms have been discussed that might underlie these long-lasting changes. For example, alterations in the amount of neurotransmitter released from presynaptic terminals in response to a standard action potential is one obvious way in which synaptic efficacy might be modulated in response to prior activity. Alternatively, changes in the number or activity of neurotransmitter receptor molecules at postsynaptic sites could alter the level of current elicited in dendrites during synaptic transmission. A considerable body of evidence now exists supporting the existence of such activity-induced changes in postsynaptic receptors, which is particularly marked during synapse development. Experimental evidence suggests that both the insertion of receptors into synaptic junctions and their biochemical modification may be important in the changes in synaptic strength that accompany LTP.

A second general category of potential memory-related adaptation involves changes in synaptic morphology. The shapes of dendritic spines have excited interest over many years as a possible morphological substrate of functional diversity among excitatory synapses that might be linked to learning and memory. Experiments in which animals as diverse as bees, fish, and mice have been raised under conditions of sensory enrichment or deprivation indicate that spine shape and number can vary as a function of brain activity, encouraging

the belief that anatomical plasticity may provide a means of tailoring synaptic throughput to match circuit activity. A link to human cognition is suggested by the observation in many studies that various forms of learning disability are associated with abnormalities in dendritic spine structure.

Based on morphological criteria, three distinct categories of spines are conventionally recognized: 'mushroom', with a thin neck and expanded head; 'thin', with a long neck and less prominent head; and 'stubby', where the neck is absent. These variations might influence synaptic transmission in a variety of ways. For example, differences in the length of the spine neck appear to influence the spread of postsynaptic calcium fluxes from the spine neck into the dendrite. It is also becoming increasingly clear that the expression of functional molecules such as neurotransmitter receptors (both the number and the type of receptor) is closely coupled to the synaptic geometry. A striking example of this principle has emerged from recent experiments showing that α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) subtype of glutamate receptor is closely associated with mushroom spines (Matsuzaki *et al*., 2001).

Such morphology-dependent functional parameters would be expected to be strongly influenced by changes in synaptic geometry, and the recent advent of methods for visualizing synaptic structure in living neurons now allows this hypothesis to be tested. Results of initial studies not only confirm that dendritic spines can change shape but show them to be remarkably morphologically plastic. Time-lapse recordings show that spines are rapidly motile, undergoing spontaneous changes in shape over a time course of seconds which can be regulated by the activation of postsynaptic neurotransmitter receptors. Over longer periods, a variety of other modifications have been described, ranging from the growth of new spines over periods of the order of an hour following the induction of LTP, to their regression and eventual disappearance when synaptic transmission is blocked for a period of days. These phenomena have obvious implications for activity-dependent modulation of neuronal connectivity in central nervous system circuits which are now being explored using *in vitro* cell culture models and, with the aid of new techniques such as two-photon microscopy, in the living brain, where developmentally regulated changes in dendritic spine motility related to sensory stimulation have been described (Lendvai *et al*., 2000).

This structural plasticity depends on the presence in dendritic spines of a specialized cytoskeletal structure consisting of a dynamic meshwork of actin filaments which drives the morphological changes (Matus, 2000). Two kinds of experimental observations lend support to the hypothesis that actin-based motility is an important factor in cognitive function. First, electrophysiological experiments on brain slices have shown that drugs that block actin dynamics interfere with the development and maintenance of LTP, suggesting that actin-dependent changes in dendritic spine shape might be required for memory formation. Second, volatile anesthetic agents such as chloroform and halothane, when applied to cultured neurons at concentrations that are clinically effective, have been shown to block actin dynamics and changes in dendritic spine shape, suggesting that spine motility may contribute to global brain function.

Many of the different molecular mechanisms described as playing a part in synaptic plasticity are susceptible to genetic techniques that allow them to be manipulated in the brains of mutant mice. Applied to glutamate receptor subtypes, this approach has already revealed unexpected subtleties in the relationship of synaptic plasticity to spatial learning (Zamanillo *et al*., 1999). Future experiments of this type should dramatically improve our understanding of the roles that synaptic molecules play in cognitive function.

References

Hausser M, Spruston N and Stuart GJ (2000) Diversity and dynamics of dendritic signaling. *Science* **290**: 739–744.

Lendvai B, Stern EA, Chen B and Svoboda K (2000) Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex *in vivo*. *Nature* **404**: 876–881.

Martin SJ, Grimwood PD and Morris RG (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience* **23**: 649–711.

Matsuzaki M, Ellis-Davies GC and Nemoto T (2001) Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. *Nature Neuroscience* **4**: 1086–1092.

Matus A (2000) Actin-based plasticity in dendritic spines. *Science* **290**: 754–758.

Rusakov DA, Kullmann DM and Stewart MG (1999) Hippocampal synapses: do they talk to their neighbours? *Trends in Neurosciences* **22**: 382–388.

Sheng M (2001) Molecular organization of the postsynaptic specialization. *Proceedings of the National Academy of Sciences of the USA* **98**: 7058–7061.

Shepherd GM and Koch C (1998) Introduction to synaptic circuits. In: Shepherd GM (ed.) *The Synaptic*

Organization of the Brain, pp. 1–36. Oxford, UK: Oxford University Press.

Zamanillo D, Sprengel R and Hvalby O (1999) Importance of AMPA receptors for hippocampal synaptic plasticity but not for spatial learning. *Science* **284**: 1805–1811.

Further Reading

Peters A, Palay SL and Webster HD (1991) *The Fine Structure of the Nervous System*, 3rd edn. Oxford: Oxford University Press.

Shepherd GM (ed.) (1998) *The Synaptic Organization of the Brain*. Oxford: Oxford University Press.

Synaptic Plasticity, Mechanisms of

Introductory article

Preston E Garraghty, Indiana University, Bloomington, Indiana, USA

CONTENTS

Literally billions upon billions of connections between neurons (synapses) exist in the brain. These connections are not static, but wax and wane in strength as we learn, remember and forget.

INTRODUCTION

The brain is composed of perhaps a trillion (1×10^{12}) individual nerve cells called neurons; some estimates are an order of magnitude lower. Neurons communicate with one another using chemical messengers at synapses. Neurons receive information via synapses on their dendrites and cell bodies, and transmit information to other neurons via their axons. A synapse consists of the terminal portion of a presynaptic axonal process, a postsynaptic specialization, and the tiny space separating these. Any given neuron might form several thousand synapses, so the human brain has trillions of synapses. Since each of these synapses can be active or inactive, the number of possible synaptic states is unimaginably large (all of the atomic particles in the known universe total a smaller number). Just these numbers and considerations demonstrate that the brain has immense computational power, but there is more. While synapses can be active or inactive, the strength of the coupling between pre- and postsynaptic neurons can wax or wane for any of a number of reasons (e.g. learning-related or injury-induced changes in inputs). It is such changes in 'synaptic weights', or correlations between pre- and postsynaptic neuron activity, that are the essence of neural plasticity at a system-wide level.

MANIFESTATIONS OF NEURAL PLASTICITY

Changes in the synaptic weights of connections between neurons arise when the inputs impinging on the circuit are altered. Alterations in the pattern of inputs can be due to experience- or injury-induced processes. The most obvious examples of experience-dependent changes in synaptic weights are learning and memory; the former can be thought to involve the identification of the relevant neural circuit, while the latter involves changes in synaptic weights in the connections between the neurons making up the circuit.

Learning can be defined as a change in an organism's behavior as a function of experience. The maintenance of these modifications of behavior is memory. A given memory might be essentially permanent, with its duration defined by the lifetime of the organism, or brief, reflecting moment-to-moment behavioral adjustments to immediate events. Further, learning can occur in response to a single environmental stimulus, or it can involve the formation of associations between two or more stimuli. In any case, the behavioral changes must have a neural correlate in the central nervous system, the organ of behavior.

Nonassociative learning involves a change in behavior in response to a single stimulus. There are two sorts of nonassociative learning: habituation and sensitization. In habituation, the response to a stimulus attenuates with repeated presentations of the stimulus. Loosely speaking, one could say the organism learns to ignore the stimulus. When a rat is presented with an unexpected loud sound, it will reflexively 'freeze', a primary fear response in rats. If the sound is presented a number of times, the magnitude of the startle response will progressively decline, and eventually disappear. The systematic exposure to the unexpected loud sound might occur during a brief period, leading to short-term habituation; if a long period passes during which the sound is not presented, a subsequent presentation will elicit a startle reflex indistinguishable from that produced from the first presentation. Alternatively, if the sound is made a part of the rat's day-to-day experience, the habituation will be retained. The same process occurs in humans; the smell of a

paper mill is obvious to individuals passing by, but the people who live nearby do not notice it. Similarly, people living under a flight path serving a major airport eventually habituate to the sounds of aeroplanes taking off or landing, while for visitors to these people's homes the same level of noise is quite intrusive.

Sensitization represents a different outcome, in which the delivery of one stimulus potentiates responsiveness to other, unrelated, stimuli. For example, if you walk past a cemetery at 2 a.m., a snapping twig will elicit a response of much greater magnitude that it would at 2 p.m. The first stimulus (the cemetery) has potentiated responsiveness to the second. Like habituation, sensitization can be of short or long duration. Habituation reduces synaptic weights, making a postsynaptic response less likely, whereas sensitization increases synaptic weights, increasing the probability of a postsynaptic response to an input. Clearly, the neural mechanisms responsible for short-term versus long-term habituation or sensitization must differ, as the former reflect transient changes in synaptic weights while the latter reflect more permanent changes.

PUTATIVE MECHANISMS OF NEURAL PLASTICITY

Short-term Habituation

An early, elegant set of experiments aimed at elucidating the neural mechanisms of habituation used the marine mollusk, *Aplysia*. The gills of these animals are located on their backs, and as they move around, they spread their gills and extend a siphon that draws in water so that it can be circulated over the gills. A jet of water directed at the gill, or a touch delivered to the siphon, elicits a reflexive retraction of the gill. Continued presentations of either of these stimuli produces habituation of the gill withdrawal reflex. Might this happen because of muscle fatigue, or a reduction in the strength of the neuromuscular synapse? No: direct stimulation of the motor neuron innervating the muscle responsible for gill withdrawal produces muscle contractions of comparable amplitudes before and after habituation. Might the amplitude of the sensory neuron's response to the stimulus (e.g. touch to the siphon) diminish with repeated stimulus presentations so that its activating influence on the motor neuron declines? No: there is no change in the sensory neuron's response to the stimulus during habituation. The neural change upon which this habituation depends must, therefore, exist at the synaptic interface between the sensory and motor neurons.

To understand the mechanism that was eventually revealed, it is necessary first to discuss synaptic transmission in greater detail. The sensory neuron communicates with the motor neuron chemically. When activated, the sensory neuron discharges neurotransmitter molecules that are recognized by specialized proteins embedded in the membrane of the motor neuron. Prior to their release, these neurochemical molecules are sequestered in small, spherical structures called synaptic vesicles. As this 'packaging' occurs in the sensory neuron, roughly equal numbers of transmitter molecules are stored in each vesicle. The amplitude of the response of the motor neuron to activation of the sensory neuron is directly related to the number of synaptic vesicles that are released by the sensory neuron. Moreover, a motor neuron's response amplitude increases in a stepwise fashion, with each step due to the release of a single additional synaptic vesicle. Thus, there is a 'quantum response' to a 'quantum of neurotransmitter release', with the latter being isomorphic with vesicle number. The researchers found that the amplitude of the postsynaptic response of the motor neuron decreased in a stepwise fashion during habituation training, indicating a stepwise decrease in the number of synaptic vesicles released by the sensory neuron. The number of vesicles released by the sensory neuron is determined, at least in large part, by the magnitude of influx of calcium ions as the nerve impulse invades its axon terminals. This influx of calcium ions is triggered by the voltage transition experienced in the axon terminal as the neural impulse in the sensory neuron arrives. During habituation, the influx of calcium ions into the axon terminals of the sensory neuron declines; thus, the quantal release of synaptic vesicles diminishes, as does the resulting postsynaptic response of the motor neuron (Figure 1).

Short-term Sensitization

Aplysia has also proved to be a particularly useful subject for the study of sensitization. Delivery of a noxious stimulus to the animal's tail strengthens synaptic transmission at several sites in the circuit mediating the gill withdrawal reflex. The stimulus delivered to the tail activates sensory neurons that synapse upon facilitator interneurons; these facilitator interneurons, in turn, form synapses on the terminals of sensory neurons responsive to stimulation of the siphon. When the tail is stimulated, the activation of the facilitator interneurons

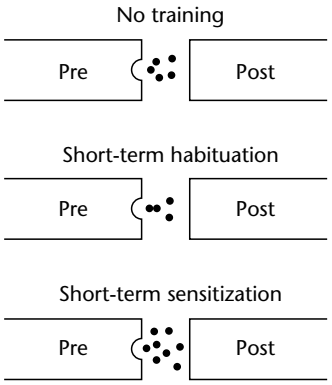


Figure 1. Changes in synaptic weights with short-term habituation and short-term sensitization. Represented are presynaptic terminals of sensory neurons and postsynaptic receptor sites on motor neurons. The black circles in the synaptic space illustrate the level of presynaptic neurotransmitter release. With short-term habituation, the strength of the synaptic connection is decreased owing to a reduction in presynaptic neurotransmitter release. With short-term sensitization the synaptic connection is strengthened owing to an increase in presynaptic neurotransmitter release. Pre, presynaptic terminal of sensory neurons; Post, postsynaptic receptor sites on motor neurons.

potentiates the motor response to siphon stimulation by increasing neurotransmitter release by the siphon sensory neurons. The increased ('sensitized') release of neurotransmitter by the siphon sensory neurons is due to several cascades of events within the presynaptic nerve terminal of the siphon sensory neuron that are initiated with the binding of the neurotransmitter used by the facilitator interneurons with receptors on the siphon sensory neurons. One cascade culminates in a change in potassium ion channels such that potassium ions are impeded in their outward flow across the cell membrane. This serves to prolong the duration of the neural impulse arising from siphon stimulation, and thus, to increase the influx of calcium ions, which in turn increases vesicular release. A second cascade mobilizes mechanisms that act to move vesicles closer to release point, increasing the likelihood that the calcium ion influx will trigger their release. A third cascade leads to the opening of another set of calcium ion channels, and this additional calcium influx also acts to increase neurotransmitter release.

Long-term Habituation and Sensitization

The delivery of 10 tactile stimuli to the siphon in a single training session produces habituation that lasts for minutes. Similarly, a single shock to the tail produces a sensitized response to siphon stimulation that lasts for minutes. On the other hand, four habituation training sessions separated by as much as a day lead to retention of the habituated response that can last several weeks, and five or more shocks to the tail produce sensitization that lasts for days or weeks. These longer-lasting manifestations of habituation and sensitization obviously begin in the same way as their short-term analogs, but just as obviously must involve additional mechanisms that sustain the experience-dependent changes in synaptic weights. Experiments have shown that these additional mechanisms actually alter the morphology of the connections between the siphon sensory neurons and the dendrites of the motor neurons. With long-term habituation, there is a reduction in the number of synaptic connections between sensory and motor neurons in the gill-withdrawal reflex circuit; with long-term sensitization, the number of synaptic connections is increased (Figure 2).

Long-term sensitization has received most of the research attention. Short-term and long-term sensitization both involve a strengthening in connections between sensory and motor neurons due to an increase in neurotransmitter release. Moreover, cyclic adenosine monophosphate (cAMP) and protein kinase A (PKA) are intimately involved in both. The sensitizing stimulus (the tail shock) provokes the release of modulatory neurotransmitters from facilitator interneurons that bind to postsynaptic receptors on sensory neurons. This transmitter–receptor binding initiates the production of cAMP, which in turn activates PKA. Together with some other molecules, PKA acts to phosphorylate (add a phosphate group to, and thus change the function of) several proteins. The phosphorylation

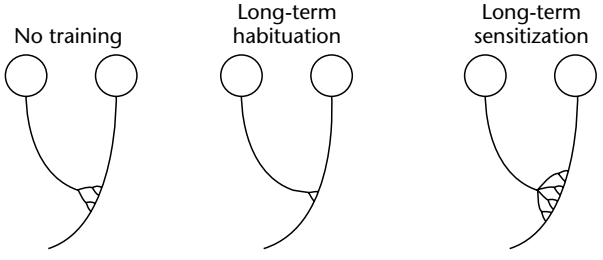


Figure 2. Morphological changes associated with long-term habituation and sensitization. The small triangles represent synaptic connections. Their numbers are decreased with long-term habituation and increased with long-term sensitization. Adapted from Bailey and Chen (1983).

of these proteins leads to an increase in neurotransmitter release from the sensitized sensory neuron (representing the siphon), provoking an enhanced motor response. The transition from short-term to long-term sensitization involves a process referred to as consolidation: the establishment of a more permanent record of experience. When you move to a new town and acquire a new home phone number, some effort is involved in storing that new number in memory for retrieval on demand. Initially one might rehearse the number over and over, and ultimately the number finds its way to long-term memory. Similarly, as one moves from short-term to long-term sensitization, the repeated presentation of the sensitizing stimulus serves the function of rehearsal, and the sensitized responsiveness acquires a less transient representation in the nervous system. With long-term sensitization in *Aplysia* this transition is due to the activation of another kinase by the accumulating PKA. Together, these two kinases migrate to the nucleus of the sensory neuron where they initiate a sequence of genetic processes that result in two significant effects: the activation of molecules that stabilize the PKA (acting to maintain the increase in neurotransmitter release initiated by the initial sensitizing stimulus), and the activation of a gene product required for the growth necessary for the creation of new synaptic connections (thereby augmenting the increased presynaptic neurotransmitter release). Thus, modifications in the efficacy of existing synapses can support the more transient short-term 'memories', while changes in synapse number due to changes in gene expression provide the means to consolidate the short-term memories into longer-lasting memories.

CONCLUSION

While the work discussed here, involving simple forms of learning in an invertebrate model, illustrates nicely how experience can affect synaptic weights, many questions remain. The means by which stimuli reduce calcium ion influx into presynaptic terminals during short-term habituation training remain speculative. The presumptive genetically regulated cascade that results in long-term habituation has received little research attention. Certainly, it is the case that 'positive' changes in synaptic strength (as with sensitization) appear to have greater relevance for learning and memory, but it is reasonable to expect that mechanisms leading to the weakening of connections are equally important. Most significantly, even with these very simple models of learning, the 'decision' mechanisms that act to consolidate short-term to long-term memories of the sensitizing or habituating stimuli are unknown. When is enough enough? How can a presynaptic nerve terminal in the reflex circuit 'know' its history of activation? Perhaps by monitoring free intracellular Ca^{2+} levels. How can it know whether to activate gene-regulated changes in morphology? Perhaps by precisely measuring free intracellular Ca^{2+} levels. Thus, increased (or decreased) intracellular Ca^{2+} levels stimulate locally regulated mechanisms act to increase (or decrease) neurotransmitter release. The larger (or smaller) postsynaptic response then tends the synaptic connection toward sensitization (or habituation).

Perhaps. However, we must not lose sight of the fact that the 'long-term' plasticity discussed here has been shown to last only days to weeks. There are many things that you know that do not require monthly rehearsal regimens – such as an ability to recite the alphabet. Clearly, additional factors (activated by completely unknown mechanisms) are responsible for consolidation as we know it (i.e. the 'permanent' storage of memory). Finally, this review has focused on two forms of nonassociative learning, but associative learning accounts for a much greater proportion of any human's memory store, and neural mechanisms that might support associative forms of learning have been extensively studied. An example of this is long-term potentiation.

Further Reading

Bailey CH and Chen MC (1983) Morphological basis of long-term habituation and sensitization in *Aplysia*. *Science* **220**: 91–93.

Castellucci VF, Carew TJ and Kandel ER (1978) Cellular analysis of long-term habituation of the gill-withdrawal reflex in *Aplysia californica*. *Science* **202**: 1306–1308.

Hawkins RD, Abrams TW, Carew TJ and Kandel ER (1983) A cellular mechanism of classical conditioning in *Aplysia*: activity-dependent amplification of presynaptic facilitation. *Science* **219**: 400–405.

Hawkins RD, Kandel ER and Siegelbaum SA (1993) Learning to modulate transmitter release: themes and variations in synaptic plasticity. *Annual Review of Neuroscience* **16**: 625–665.

Kandel ER (1989) Genes, nerve cells, and the remembrance of things past. *Journal of Neuropsychiatry* **1**: 103–125.

Kandel ER, Schwartz JH and Jessell TM (2000) *Principles of Neural Science*, 4th edn, pp. 1247–1259. New York: McGraw-Hill.

Thompson RF (2000) *The Brain: A Neuroscience Primer*, 3rd edn. New York: Worth.

Thompson RF and Spencer WA (1966) Habituation: a model of phenomenon for study of the neural substrate of behavior. *Psychological Reviews* **173**: 16–43.

Syntax and Semantics, Neural Basis of

Intermediate article

Lewis P Shapiro, San Diego State University, San Diego, California, USA

CONTENTS

Based on evidence from aphasia and neuroimaging studies, elements of syntax and semantics appear to be neurologically isolable and have distinct time courses.

INTRODUCTION

The neurology of syntax and semantics begins with a broad description of these two general areas of language. Syntax includes (a) phrasal geometry – the way in which lexical categories such as noun and verb project hierarchically to form phrasal categories such as noun phrases and verb phrases; (b) the 'dislocation' property of language whereby phrases that are pronounced, heard, or read in one position, are interpreted in a different, nonadjacent position; (c) lexical properties that have structural effects (e.g. the number and type of logical arguments or thematic roles a verb or noun entails); (d) binding relations – the distribution of pronouns, reflexives, and their antecedents (the noun phrases to which they co-refer); and (e) inflectional morphology and agreement. Syntax, then, is essentially about structure-building, from the word to the sentence. Semantics includes (a) the meanings or senses of words (known as lexical-semantics; more formally this may include lexical-conceptual structure); and (b) the meaning of the entire sentence, including its truth-value and logical form. Sometimes semantics also includes how the meaning of a sentence fits into the discourse of multiple sentences. (*See* **Binding Theory**; **Construction Grammar**)

Linguistic theory is concerned with characterizing a mental grammar – the set of abstract categories, rules and principles that underlies language. Accordingly, it has been suggested that a theory of a mentally represented grammar must be accountable to neurological data. The theory can be used as a set of 'discovery procedures', seeking whether elements of syntax and semantics are neurologically isolable. It is also often assumed that linguistic theory is directly relevant to the computational activity of the language processor, the latter being concerned with the study of the mental operations that compute, activate, and integrate linguistic information types. The neurology of language, therefore, involves concepts and methods from linguistics, cognitive psychology, neuropsychology, computer sciences, the neurosciences, and communicative disorders. (*See* **Language and Brain**)

There are two general methods in the study of the neurology of syntax and semantics and in the study of brain–language relations: behavioral studies and functional imaging. Behavioral studies typically use aphasia (that is, lesion studies) as a window into the normal system. Functional imaging techniques could be argued to be more directly aligned with neurology, and are designed to measure the localization of, and in some cases temporal unfolding of, aspects of language. These techniques include positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG), and the measurement of event-related potentials (ERPs) using electroencephalography. Though functional imaging has begun to bear empirical fruit, the study of brain–language relations has, as its bedrock, studies of language pathology subsequent to brain damage. (*See* **Aphasia**; **Electroencephalography (EEG)**)

EARLY LESION STUDIES

Syntactic and semantic analyses of aphasia have evolved from a theory of the localization of 'language as activities' as exemplified by the work of Wernicke, Lichtheim, and Geschwind. This theory focused on the dissociation of production and

comprehension: the former was considered to be disrupted by left inferior frontal lesions involving motor association cortex (Broca aphasia) (Figure 1), the latter by left superior posterior temporal lobe lesions involving auditory association cortex (Wernicke aphasia) (Figure 1). From these findings it was assumed that language production depended upon principles of motor system organization and that comprehension depended upon principles of auditory perception. (*See* **Wernicke–Geschwind Model**; **Broca, Paul**)

Following Chomsky's groundbreaking work in theoretical linguistics, psycholinguists in the 1960s and 1970s began investigating brain–language relations in terms of abstract grammatical categories. In people with Broca aphasia, comprehension limitations were discovered that seemed to parallel production deficits: just as these patients tended to omit grammatical morphemes when speaking, so, too, were they unable to use them for the purpose of comprehension. Also in line with their syntactically simplified output, they were unable to carry out normal syntactic analyses on input strings, their comprehension at the sentence level thereby being abnormally reliant on semantic and plausibility cues. Taken together these findings suggested that damage to left inferior frontal cortex produced an overarching syntactic limitation, even as it spared the capacity to carry out semantic inference (Figure 1). In contrast, people with Wernicke aphasia did not show a normal capacity for semantic inference at the sentence level. Since these patients also tended to produce semantically empty speech in the context of what seemed to be normal syntax, it seemed reasonable to assign this cortical area the role of an amodal semantic center (Figure 1). (*See* **Government–Binding Theory**)

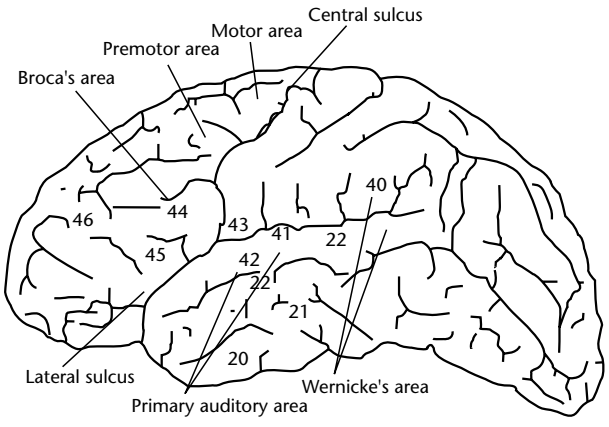


Figure 1. Lateral view of the left cerebral cortex, showing the cytoarchitectural boundaries and Brodmann's areas.

On further analyses in the 1980s and beyond, both functional distinctions (production–comprehension, and syntax–semantics) harvested some support from lesion studies. In what follows, the focus is primarily on comprehension. The language parts and the neurology are considered separately, beginning with an introduction to current psycholinguistic theories of normal language processing, those that are relevant to the study of brain–language relations.

MAJOR PSYCHOLINGUISTIC THEORIES

Just as a linguistic theory should be neurologically defensible, so, too, should hypothesitical processing operations be accountable to data gathered from both normal and language-impaired individuals. Current psycholinguistic theories (mostly concerned with comprehension) can be divided into two competing sets of accounts. Serial, form-driven accounts have a modular architecture and claim that syntactic processing proceeds initially independently from semantic and pragmatic considerations (e.g. Frazier and Clifton, 1996). One such account claims that a first-pass analysis of a sentence includes only placing the incoming lexical items into lexical categories, and, perhaps, activating lexical properties that have immediate structural implications (for example, whether a verb licenses one, two, or three arguments). Lexical categories are then set into phrases via simplicity heuristics: for example, 'build the smallest number of nodes'; 'attach the next lexical item to an already existing phrase'. Thus, only one possible parse is attempted at any given time. The second-pass analysis (or reanalysis if the first parse fails) uses contextual semantic and pragmatic information to identify the correct interpretation for the input. To complicate matters, there are also form-driven accounts that allow multiple syntactic parses in the case of syntactic ambiguities. (*See* **Language Comprehension**; **Psycholinguistics**; **Sentence Processing**; **Sentence Processing: Mechanisms**)

Unlike such form-driven accounts, highly interactive accounts claim that various processes (including syntax and semantics) interact continuously during comprehension. Syntax is considered mostly to be a second-order description for what are really concatenated lexical activation processes (e.g. MacDonald *et al*., 1994). One critical difference between the two accounts, then, is the initial influence of contextual, semantic and pragmatic influences: a form-driven account claims that initial processing is not influenced by such extrasyntactic

information, whereas a highly interactive account claims that processing is continuously affected by multiple sources of information (with different types of information having different degrees of influence). Note that the highly interactive account has some intuitive appeal; after all, it is unarguable that all types of information interact to yield an interpretable sentence. However, the issue is not if information interacts during language processing, but when particular types of information are activated and integrated into the temporal unfolding of language processing. (*See* **Constraint-based Processing**)

To this end, psycholinguistic investigations sometimes make a distinction between offline and online tasks. Offline tasks are useful in studying coarse-grained characteristics of language processing; they are designed to collect data at the endpoints of processing after conscious reflection has occurred. Typical offline methods are untimed, and include sentence–picture matching tasks, grammaticality judgments, answering questions about a sentence, or paraphrasing that sentence. Online tasks collect 'processing time' data moment by moment as input is being analyzed; in particular, they are claimed to be sensitive to rapid, 'reflexive' and unconscious operations. Online tasks include crossmodal lexical priming techniques and word-by-word reading and eye-tracking.

Normal Lexical Processing

The battleground for the form-driven and highly interactive processing accounts has typically involved lexical (and also structural) ambiguities, as well as the dislocation property of sentences. Lexical ambiguities concern the processing of lexical items that have more than one meaning, and the subsequent resolution toward one of those meanings. For example, consider:

> The man saw several spiders, roaches and other bugs [1] in the corner [2] of his room.

The word 'bugs' is lexically ambiguous (ignore for the moment the effect of sentence context); it has (at least) two distinct meanings – 'insect' and 'listening device or hidden microphone'. Given the presentation of a sentence containing a noun (or any lexical item) that has multiple senses attached to one phonological or orthographic form, are one or all of the meanings activated, and when during the time course of sentence comprehension does that activation take place? The answer to these questions seems to indicate that at the point where the phonological/orthographic form of the ambiguous item is recognized, multiple senses are activated. Interestingly, these multiple senses are initially activated even in sentence contexts that appear to be biased toward one particular sense. So, given the example above, at probe position 1 – in the immediate temporal vicinity of the target 'bugs' – the 'insect' sense and 'hidden microphone' sense are both momentarily activated; it is only downstream from the target (position 2, roughly 750 ms later) that the contextually relevant meaning (that is, 'insect') remains active and the inappropriate meaning is suppressed. (*See* **Lexical Ambiguity Resolution**)

These results suggest that lexical access is initially form-driven; context effects and those based on real-world knowledge are seen to take place only subsequent to lexical access. Similar immediate, form-driven effects have been found for the activation of a verb's argument structure (the number of participants/noun phrases entailed by a verb).

Normal Syntactic Processing

The processing of dislocation ('discontinuous dependencies') has likewise reflected on the form-driven nature of initial language operations. Consider:

> The policeman saw the boy who the crowd at the party [1] accused _____ [2] of the crime.

Note that this complex sentence contains two underlying propositions: 'The policeman saw the boy' and 'The crowd at the party accused the boy of the crime'. Each underlying proposition suggests a subject–verb–object order (which is canonical in English). So, in the surface form of the complex sentence, the direct object of the verb 'accused' (that is, 'the boy') occurs prior to the verb, in a noncanonical position. Linguistic theories have various ways for dealing with such dislocation. For example, in a transformational framework the underlying direct object noun phrase (NP), 'the boy', has moved to a noncanonical position, leaving behind a trace or copy of itself at the post-verb position (signified by the gap). In psycholinguistic terminology, the moved NP (or the NP that is heard during auditory presentation) is called the 'filler', and the position from where it moved and where it is interpreted is called the 'gap'.

Presented with such structures, normal listeners indeed appear to interpret the dislocated NP at its canonical, postverb position (position 2 above), yet interpretation does not occur just prior to this position (at 1 above). Thus, the dislocated NP is reaccessed (see, for example, Love and Swinney, 1996). Like the lexical ambiguity work, gap-filling also appears to be a form-driven, automatic, reflexive

operation, even when biasing semantic and contextual information is added. Consider:

> Everyone watched the enormous heavyweight boxer who the small 12-year-old boy on the corner had [1] beaten ____ [2] so brutally.

Even though real-world plausibility strongly suggests that 'the enormous heavyweight boxer' is not a plausible direct object for the verb 'beaten', 'boxer' is activated at the gap position (2), but not at the pregap position (1). This interesting property of language processing is also observed with other complex constructions, and it has been argued that lexical entailments and even probabilistic information do not seem to affect such automatic, initial processing routines.

LESION STUDIES: AN UPDATE

In the 1990s the lesion-localizing value of Broca and Wernicke aphasias continued to dominate the investigative landscape. Though lesion extent is often variable, the neurological damage underlying these two types of aphasia appears to be quite distinct: Broca aphasia involves cortical tissues which include the left frontal operculum (Broca's area) and insula, as well as subjacent white matter; Wernicke aphasia involves tissue that lies more posteriorly, in the left temporoparietal region inferior to the sylvian fissure (Figure 1). What, then, are the behavioral manifestations tied to these profiles that involve linguistic information types and their processing routines?

One influential set of accounts – the trace deletion hypothesis – has claimed that individuals with agrammatic Broca aphasia have a restricted syntactic deficit that disallows the normal representation of traces or copies of moved NPs. This deficit (essentially, a description cast in terms of linguistic theory) behaviorally manifests as an inability to comprehend only certain sentence structures (e.g. passives, object relatives and object clefts), while sparing others (e.g. actives, subject relatives and subject clefts). The empirical support for this account has come primarily from patterns of sparing and loss seen on offline tasks such as sentence–picture matching and grammaticality judgments. Thus, the claim is that the neural tissue implicated in Broca aphasia is used to support the computation underlying dislocation – of transformational relations between moved NPs and their extraction sites – and is not used for any other 'syntactic' or lexical activity (Grodzinsky, 2000). Unlike the interpretable sentence comprehension patterns evinced in Broca aphasia, Wernicke aphasia results in variable, often uninterpretable behavior. When these behaviors are interpretable, however, they suggest a lexical-semantic deficit that does not overlap to any significant extent with syntax or with that of Broca aphasia. There is also considerable online evidence for a distinction between Broca aphasia and Wernicke aphasia, related grossly to the syntax–semantics partition. Zurif and colleagues in a series of online examinations found that people with Broca aphasia do not reaccess fillers at their gap site. Processing a sentence continues without the normal link established between the NP and its extraction site, resulting in offline comprehension difficulties (e.g. Zurif *et al.*, 1993). Individuals with Wernicke aphasia show normal, on-time reaccess effects. Furthermore, though it has also been claimed that lexical deficits cause the syntactic limitations in Broca aphasia, the activation of lexical properties (specifically, verb–argument structure) is intact in Broca – and not Wernicke – aphasia, and therefore independent from observed syntactic limitations (e.g. Shapiro *et al.*, 1993). These online data also suggest that it may not be syntactic processing that is served by Broca's area and surrounding tissue, but rather that this brain region may provide the resources necessary for carrying out fast-acting, reflexive operations that happen to underlie structural processing.

Other Ways to Cut the Language Pie

The interpretation that either (a) a restricted syntactic deficit underlies Broca aphasia while people with Wernicke aphasia evince a lexical-semantic deficit, or (b) reflexive, automatic processing routines are disrupted in Broca aphasia, has not gone unchallenged. The relevant opposing claim is that attempting to generalize sentence comprehension patterns to individuals with Broca aphasia is misguided because of the heterogeneous nature of the behavioral patterns observed in these individuals, and because of the wide-ranging cortical and subcortical damage associated with Broca aphasia. This claim suggests a more 'holistic' approach to language representation and processing in the brain; on this account the apparent linguistic-specific deficits are more likely to be explained by deficits in general cognitive abilities, though these 'abilities' (e.g. working memory, attentional resources) remain relatively undefined.

It would seem, then, that this domain-general account of language deficits is at complete odds with a localizationist, language-specific account. Certainly the underlying assumptions are at odds: the linguistic distinction account suggests that

brain regions are specialized for particular language operations, while the holistic account claims that language emerges from the conjoint activity of spatially discontinuous and widely distributed brain regions. The evidence from neuroimaging might help to resolve some of these issues; indeed, this evidence suggests that the lesion studies have been on the right track – that syntactic and semantic operations recruit different neuroanatomical regions, and, critically, have distinct time courses.

FUNCTIONAL NEUROIMAGING AND EVOKED-POTENTIAL STUDIES

Recent technological advances have demonstrated that neuroimaging and electrophysiological examinations have important potential. Each technique has its limitations: PET has fair spatial resolution (10 mm or more) but poor temporal resolution (approximately 30 s); fMRI has excellent spatial resolution (less than 5 mm) but temporal resolution is moderate (approximately 1 s). Evoked potentials have excellent temporal resolution (in ms), but comparatively poor ability to locate the neural generators of electrical activity in the brain (within about 20 mm).

Positron Emission Tomography

Positron emission tomography involves taking images of regional cerebral blood flow after injection of a radiopharmaceutical tracer such as water labeled with oxygen-15. It thus can be used to see which areas of the brain 'light up' during overall sentence interpretation. Caplan and colleagues have produced a series of such studies, examining sentence structures like right-branching subject gaps (e.g. 'the child spilled the juice that stained the rug') and center-embedded object gaps (e.g. 'the juice that the child spilled stained the rug'). These studies show that the left perisylvian cortex (pars opercularis of Broca's area) is activated in complex (object gap) structures, while Wernicke's area is not (see, for example, Caplan *et al*., 2000).

Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging exploits the local magnetic properties of brain tissue. When the person to be investigated is placed in a magnetic field, the proton nuclei of hydrogen atoms align themselves within the field and move about in random orientation to one another. A brief electromagnetic pulse is introduced which disturbs the nuclei and results in their coherent orientation. This coherence is detected as a radio signal and is formed into an image. Blood can be used as an endogenous contrast agent which allows the measurement of the hemodynamic response to event-related (that is, stimulus presentation) activities. Taken as a whole, fMRI studies indicate that different neural regions appear to support different linguistic information types.

Event-related Brain Potentials

Event-related brain potentials record the electrical activity of the brain time-locked to an external event (in the present case, time-locked to word or sentence input). Like many of the neuroimaging techniques, it is common practice to use the detection of different types of linguistic errors, and the resultant ERPs are measured in terms of their amplitude, latency and (in some cases) distribution. The ERP has been extensively used to study language processing. The N400 – a negative waveform occurring approximately 400 ms after presentation of relevant stimulus – is typically used as a marker for semantic processing, following the groundbreaking work of Kutas, Hillyard and their colleagues (Kutas and Hillyard, 1983). It is likely that the N400 is elicited in response to the semantic or pragmatic expectation of a given word in a particular sentence context, with large N400s found when the word is semantically anomalous. (*See* **Event-related Potentials and Mental Chronometry**)

Syntactic processing has been associated with two distinct waveforms, a late and large positive wave shift (the P600), and a left anterior negative (LAN) wave. The P600 has been found with various types of syntactic violations, such as phrase structure and agreement errors, and appears to be associated with second-pass reanalysis routines. Friederici and colleagues (e.g. Friederici, 1995) have specifically tied their work to models of normal sentence processing. They have suggested a two-stage serial processor whereby first-pass parsing routines are highly automatic and reflexive, and are reflected by an early LAN wave, and that controlled, second-pass reanalysis routines are reflected by the P600.

Magnetoencephalography

A newer electrophysiological technique, magnetoencephalography, has similarly good temporal resolution, and significantly better spatial resolution than ERPs, but is limited by only detecting signals from areas of cortex that lie perpendicular to the scalp. Thus far, MEG has been used mostly to examine phonetic and acoustic variables (the

latter has been directly compared, favorably, to the detailed data obtained from invasive electrode recordings in other animals).

Summary of Aphasia and Neuroimaging Data

Though neuroimaging and electrophysiological techniques have yielded important information about brain–language relations, it could be argued that the fuel for these studies has been investigations of language processing in both normal and brain-damaged individuals. To be sure, there are limitations to both methods. Evidence from brain damage is, by its nature, imprecise. Theoretical differences, stimulus characteristics, task considerations, subject selection, and perhaps individual variation, combine to sometimes yield conflicting data. Neuroimaging has attempted to provide greater precision regarding localization of function, but it turns out that there is considerable variability regarding strict localization of syntactic and semantic processing here as well, and localization of semantic processing is typically computed by reference to semantic anomalies or simple lexical access, which are at best rough attempts at semantics.

Nevertheless, by combining lesion studies with neuroimaging techniques, scientists have begun to unravel the temporal and spatial characteristics of language representation in the brain (Figure 1). These characteristics may include:

- acoustic information that is analyzed in the auditory cortex within 100 ms of stimulus onset;
- identification of phonological word forms and phonological sequencing, which recruits left planum temporale and also involves the upper and posterior parts of Broca's area;
- lexical category formation that occurs within 200 ms of onset, and moves from Wernicke's area (Brodmann's area 40) to the inferior part of Broca's area (Brodmann's area 44), the latter supporting initial structure-building – including dislocation processing;
- lexical-semantic activation and integration, which involves the temporal language region (Brodmann's areas 22 and 21) and occurs within 300–500 ms of onset;
- reanalysis of input (if necessary), which occurs within approximately 600–1000 ms of onset and probably recruits areas involved in working memory (e.g. Brodmann's area 46) and regions including the right hemisphere as well (see Friederici, 1995).

RECOVERY, TREATMENT AND NEUROIMAGING

Recovery from brain damage is just beginning to be seriously studied by scientists interested in brain–language relations. The issues confronting these investigations are particularly thorny, since brain organization, learning or relearning, and several facets of language are involved. In particular, careful attention has to be paid to the research designs and methods that yield and measure learning over time. Although there are many published studies on the treatment of language disorders through the fields of communicative disorders and the neuropsychology of language, very little of this work has been concerned with the sorts of specific syntactic and semantic deficits observed in the aphasias (but see, for example, Thompson *et al*., 1997). Fewer still have exploited neuroimaging techniques to map the purported recovery process. One broad issue is whether perilesional tissue or right hemisphere homologs (of, for example, Broca's and Wernicke's areas) appear to be active during good or poor recovery from aphasia. The evidence is equivocal, and is limited by the large differences in methods and participant selection.

SYNTACTIC PRODUCTION

Most of this discussion has been on how comprehension – normal and disordered – reflects on the neurology of syntax and semantics. There have been few similar efforts in syntactic production, even though the hallmark of Broca aphasia has always been effortful, halting and syntactically impoverished speech. The most recent work in production deficits in Broca aphasia from a linguistic framework suggests that phrasal geometry and its relation to abstract inflection features of language may be involved in a description of the deficit. Within that part of syntactic theory that deals with the phrase structure tree (which describes the structural configuration of sentences), lexical categories such as noun, verb, preposition and so on are said to 'project' upwards in the tree to maximal projections (e.g. noun phrases, verb phrases, prepositional phrases and so on). Careful analyses of agrammatic speech have yielded a circumscribed syntactic deficit: the sentence structures do not project to the topmost part of the tree, resulting in a production impairment of some sentence constructions in English (e.g. *wh*-questions, tenseless verbs), while different constructions that rely on the same part of the tree in other languages are affected (e.g. Friedmann, 2002). These sentence production results highlight the apparent importance of Broca's area for the normal functioning of syntactic abilities that are quite distinct in production versus comprehension; thus far, a description encompassing an overarching syntactic deficit has not been satisfactorily proposed.

CONCLUSION

The study of brain–language relations has benefited greatly from work in theoretical linguistics, psycholinguistics, lesion studies, and functional neuroimaging. Elements of syntax and semantics – the former more specified than the latter – appear to recruit different brain regions as well as distinct time courses, although some overlap has been observed. This localizationist hypothesis in one form or another has been with us in the modern era for at least 150 years (as has been its counterpart, the holist approach); we are now beginning to fill in some of its details.

References

Caplan D, Alpert N, Waters G and Olivieri A (2000) Activation of Broca's area by syntactic processing under conditions of concurrent articulation. *Human Brain Mapping* **9**(2): 65–71.

Frazier L and Clifton C (1996) *Construal*. Cambridge, MA: MIT Press.

Friedmann N (2002) Question production in agrammatism: the tree pruning hypothesis. *Brain and Language* **80**: 160–187.

Friederici A (1995) The time course of syntactic activation during language processing: a model based on neuropsychological and neurophysiological data. *Brain and Language* **50**: 259–281.

Grodzinsky Y (2000) The neurology of syntax: language use without Broca's area. *Behavioral and Brain Sciences* **23**(1): 1–71.

Kutas M and Hillyard SA (1983) Event-related potentials to grammatical errors and semantic anomalies. *Memory and Cognition* **11**: 539–550.

Love T and Swinney D (1996) Co-reference processing and levels of analysis in object-relative constructions: demonstration of antecedent reactivation with the cross-modal priming paradigm. *Journal of Psycholinguistic Research* **25**(1): 5–24.

MacDonald MC, Pearlmutter NJ and Seidenberg MS (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**(4): 676–703.

Shapiro LP, Gordon B, Hack N and Killackey J (1993) Verb-argument structure processing in complex sentences in Broca's and Wernicke's aphasia. *Brain and Language* **45**: 423–447.

Thompson CK, Shapiro LP, Ballard KJ *et al*. (1997) Training and generalized production of wh- and NP-movement structures in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research* **40**: 228–244.

Zurif EB, Swinney D, Prather P, Solomon J and Bushell C (1993) An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language* **45**: 448–464.

Further Reading

Bates E and Goodman JC (1997) On the inseparability of grammar and the lexicon: evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes* **12**(5–6): 507–584.

Caramazza A and Zurif EB (1976) Dissociation of algorithmic and heuristic processes in sentence comprehension: evidence from aphasia. *Brain and Language* **3**: 572–582.

Chomsky N (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Friederici AD, Meyer M and von Cramon DY (2000) Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language* **74**: 289–300.

Geschwind N (1970) The organization of language and the brain. *Science* **170**: 940–944.

Goodglass H (1993) *Understanding Aphasia*. San Diego: Academic Press.

Grodzinsky Y, Shapiro LP and Swinney DA (eds) (2000) *Brain and Language: Representation and Processing*. San Diego: Academic Press.

Shapiro LP, Swinney DA and Borsky S (1998) On-line examination of language performance in normal and neurologically-impaired adults. *American Journal of Speech-Language Pathology* **7**: 49–60.

Swinney D, Zurif EB and Nicol J (1989) The effects of focal brain damage on sentence processing: an examination of the neurological organization of a mental module. *Journal of Cognitive Neuroscience* **1**: 25–37.

Thompson CK, Fix SC, Gitelman DR, Parrish TB and Mesulam MM (2000) fMRI studies of agrammatic sentence comprehension before and after treatment. *Brain and Language* **74**: 387–391.

Thulborn KR, Carpenter PA and Just MA (1999) Plasticity of language-related brain function during recovery from stroke. *Stroke* **30**: 749–754.

Temporal Cortex

Intermediate article

Elisabeth A Murray, National Institute of Mental Health, Bethesda, Maryland, USA

CONTENTS

The temporal cortex is named for the portion of the cranial bone under which it lies. In human and macaque brains, the temporal cortex lies beneath the lateral sulcus and consists of three main parts: the superior temporal cortex, the inferior temporal cortex, and the paralimbic cortex.

INTRODUCTION

Because 'temporal cortex' is a descriptive term, derived from this tissue's physical proximity to the temporal bone of the cranium, the boundaries of this structure are fuzzy and subject to debate. In general, the temporal cortex can be divided into three main parts: the superior temporal cortex, the inferior temporal cortex, and the paralimbic cortex. The superior temporal cortex, including cortex on the superior temporal gyrus as well as cortex lying on the lower bank of the lateral sulcus, processes auditory information, but much of it has a multimodal (polysensory) function involving both auditory and visual inputs. The inferior temporal cortex, including cortex on the middle and inferior temporal gyri, processes mainly visual information. In monkeys the inferior temporal gyrus abuts the paralimbic cortex; indeed, the paralimbic cortex actually extends onto the gyrus. In contrast, humans possess additional cortex on the ventral surface of the brain (mainly on the fusiform gyrus), located between the inferior temporal gyrus and paralimbic cortex, which is also part of the inferior temporal cortex. The paralimbic cortex, which includes much of the cortex on the temporal pole, receives an array of unimodal and multimodal sensory inputs and is thought to play a part in integrating information from the different sensory modalities.

For obvious ethical reasons, human brains are rarely the subject of detailed scientific study at the single-cell level, nor are they available *in vivo* for tracer studies of neuronal connectivity. Accordingly, much of the available information about the temporal cortex comes from the study of nonhuman animals, especially macaque monkeys. For example, insights into temporal cortex function have emerged from the study of activity of neurons in awake, freely moving monkeys. In addition, neuroanatomical studies have been carried out to help define how the various parts of the temporal cortex communicate with each other and with other brain regions. Finally, the behavioral effects of selective cortical lesions have been characterized to a degree that is clearly impossible in humans. These approaches provide complementary sets of information that together provide a picture of the functional organization of the temporal cortex. A noninvasive technique for localizing function in the human brain has now been developed. This functional imaging technique uses sophisticated methods to measure local rates of blood flow and other hemodynamic events in people who are exposed to specific sensory stimuli, or while they engage in a behavioral task. Local physiological activity in a population of cells and synapses contributes to the hemodynamic changes, measures that collectively are called 'activations'.

INFERIOR TEMPORAL CORTEX

The inferior temporal cortex subserves vision, especially what is known as higher-order vision. Higher vision can be distinguished from lower vision along the same lines as perception can be distinguished from sensation. Anatomical and physiological studies have identified at least 30 separate areas with visual functions in the cortex of monkeys, and there could be an even greater number of visual areas in the cortex of humans. One view of the division of labor among the visual areas holds that there are two main processing systems, one devoted to identifying what an object is, and the other devoted to identifying where an object is (Ungerleider and Mishkin, 1982). According this view, the temporal cortex is mainly the province of the former system, known as the ventral stream because of the idea that visual information 'flows',

in some sense, from occipital cortex forwards and ventrally toward the temporal cortex (Figure 1). The other system, known as the dorsal stream because information is held to flow from the occipital to the parietal cortex, is thought to have a role in either spatial perception, as noted above, or in the guidance of movements to spatial targets (Milner and Goodale, 1996). Part of the evidence in support of two separate processing streams is that the inferior temporal areas can continue to function even though the parietal areas have been damaged. In Balint syndrome or optic ataxia, people with damage to the parietal cortex are good at naming and identifying objects, but show profound disabilities in reaching towards those same objects. Conversely, patients with damage to the inferior temporal cortex may be unable to identify objects, but at the same time are able to navigate around them and to reach accurately towards them. (*See* **What and Where/How Systems**)

Anatomy of the Inferior Temporal Cortex

In macaque monkeys, the occipitotemporal or ventral stream pathway begins with the projections from the striate cortex (the primary visual cortex, V1, also known as area 17) to the second and third visual areas, V2 and V3, which in turn project to area V4 (Figure 1). These prestriate visual fields are arranged in adjacent cortical belts that nearly surround the striate cortex. One output of these prestriate regions is to the inferior temporal cortex, comprising areas TE and TEO. (The labels 'TE' and 'TEO' are derived from cytoarchitectonics: the parcellation of cortical fields based on cell size, type and staining intensity, laminar organization, and other structural features that together can be used to differentiate one region from another. The T stands for temporal, the E designates the fifth of a group of fields, classified by anatomists as A to E, and the O in TEO stands for occipital, meaning the subdivision of TE nearest the occipital lobe.) The inferior temporal cortex is in receipt of highly processed visual information arising not only from lower-order visual areas, but also from subcortical structures including the thalamus and basal ganglia.

Inferior Temporal Cortex Function in Monkeys

Neurons in the visual cortex 'see', or represent, pieces of our visual world. The amount of visual space that a neuron represents, or is responsive to, is termed the 'visual receptive field'. In striate and prestriate areas the representation in the cortex is map-like or visuotopic: that is, nearby parts of visual space are represented in nearby parts of a cortical area in a systematic mapping. In area TEO of the inferior temporal cortex, this visuotopic organization seems to be coarser than that in earlier fields; in area TE it may be nonexistent. In area TE, unlike other visual fields, neurons have large

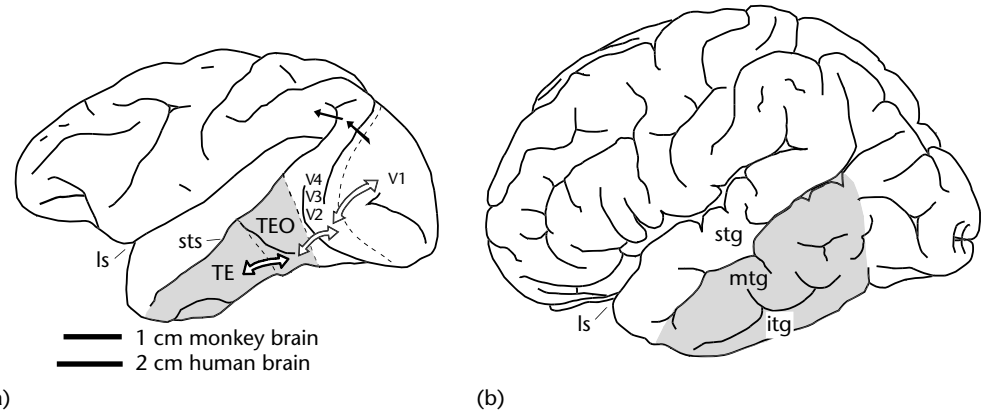


Figure 1. Location of inferior temporal cortex (shaded region) in (a) macaque monkey and (b) human brains, shown on a lateral view of the brain. In monkeys, the inferior temporal cortical fields related to vision are areas TE and TEO, which occupy all of the middle temporal gyrus and much of the inferior temporal gyrus. White arrows indicate flow of information from striate cortex (V1) to prestriate (V2, V3, V4) areas into inferior temporal cortex, together constituting the ventral visual stream. Black arrows suggest flow of information from striate to prestriate to parietal cortex, the dorsal stream. itg, inferior temporal gyrus; ls, lateral sulcus (sylvian fissure); mtg, middle temporal gyrus; stg, superior temporal gyrus; sts, superior temporal sulcus.

receptive fields that often cross the midline. Thus, neurons in area TE can 'see' an object regardless of its position in the visual field.

Neurons in the inferior temporal cortex are sensitive to one or more aspects of the stimulus qualities that we perceive. For example, neurons represent color, texture, direction of motion, size and orientation, and binocular disparity. Because neurons in inferior temporal cortex receive visual information from the earlier stages, it appears that the inferior temporal cortex neuronal responses are built up from those in earlier areas. For example, neurons in caudal visual fields respond well to oriented bars and simple gratings; in contrast, neurons in anterior and ventral portions of inferior temporal cortex respond better to complex visual stimuli. Although single neurons may respond selectively to objects, it seems likely that individual objects are represented by a population of neurons, as opposed to single neurons.

Damage to the inferior temporal cortex leads to deficits in the identification of objects through vision. Behavioral tasks designed to test visual abilities in monkeys indicate that inferior temporal cortical damage produces difficulties in discriminating and in matching objects on the basis of vision. At the same time, objects can be discriminated by touch. Because basic sensory capacities such as acuity and color vision are largely intact after inferior temporal cortical damage, the difficulty is considered to be one of higher-order vision. The nature of the impairment caused by inferior temporal cortex lesions is controversial. One idea is that damage to this region results in a loss of the ability to represent features, in such a way that the greater the damage, the fewer the visual features that can be represented. Another idea is that it is not the ability to represent features that is lost, but rather the ability to represent conjunctions of features. Both these ideas emphasize the role of inferior temporal cortex in visual perception: that is, its role in representing visual features. On either view, damage to inferior temporal cortex will lead to disruption of stored representations of features, resulting in a loss of long-term visual memory. At the same time, this damage would result in an impaired ability to represent features, a perceptual deficit. Thus, it appears that neurons in the inferior temporal cortex contribute to both perception and memory, and that perceptual and mnemonic functions are anatomically inseparable in this part of the brain. (*See* **Object Perception, Neural Basis of**; **Vision: Object Recognition**)

Also in the temporal cortex are two areas involved in processing motion vision, the middle temporal and middle-superior temporal areas, abbreviated as MT and MST, respectively (not illustrated). These areas are located in the depths of the superior temporal sulcus, near the temporal-parietal-occipital junction. Often considered a third visual stream, contrasting motion vision with the object vision of the ventral stream and the spatial vision of the dorsal stream, these visual areas have been the best characterized in terms of their neurophysiological mechanisms. Neurons in MT and MST reflect precisely the judgments reported by monkeys about the direction in which a field of light spots is, on average, moving. Further, exciting cells by passing electrical current through an electrode near them affects the monkey's perceptual judgment in a systematic manner. For example, if a group of neurons represents upward motion, then stimulating these cells during a period when the light spots are tending to move to the left will induce the illusion that they are moving slightly upwards as well as to the left.

Inferior Temporal Cortex Function in Humans

The general pattern of visual information processing appears to be similar in monkeys and humans (Figure 1), and many of the cortical fields identified in monkeys have probable homologues in humans (Wandell, 1999). As in monkeys, there are visual areas devoted to processing object features, and others, buried deep in the posterior portion of the superior temporal sulcus, devoted to processing motion.

Damage to the inferior temporal cortex in humans – which may follow stroke, herpes encephalitis, degenerative disorders such as Alzheimer disease, or surgical intervention to excise a brain tumor – results in a class of deficits termed 'visual agnosia'. Depending on the location and extent of the brain damage, deficits range from severe impairments in shape discrimination, in which patients are impaired in identifying, matching, copying or discriminating simple visual stimuli, to milder deficits characterized by difficulty in naming objects on visual confrontation. Between these extremes are patients who, although able to identify objects presented singly, have difficulty when several objects are present, or when they must describe the contents of complex pictures. Not surprisingly, these patients also have difficulty reading. Yet other patients display remarkably selective visual deficits; for instance, some are unable to identify faces of familiar persons or family members through vision, although they are able

to visually identify other types of objects and to identify familiar people upon hearing their voices (Farah, 1990).

Damage to more rostral portions of the inferior temporal cortex is characterized by intact visual matching abilities but deficient retrieval of information related to familiar objects. The study of patients with semantic dementia, in particular, has helped provide a picture of rostral inferior temporal cortex function. Patients with semantic dementia sustain a progressive loss of neurons in the inferior temporal cortex. The defining characteristic of these patients is a loss of semantic knowledge, that is, a loss of factual information about the world. This loss of knowledge affects both verbal and nonverbal aspects of knowledge. For example, such patients have difficulty naming familiar objects, generating a definition of objects, matching object pictures to their names, selecting appropriate colors for objects, and drawing objects from memory (Hodges *et al*., 1992). As was the case for monkeys with inferior temporal cortex damage, the impairments described above cannot be reduced to simple visual perceptual deficits; in all cases, the deficits appear despite intact acuity, brightness discrimination, and color vision.

Functional imaging studies in humans have been able to identify regions within the inferior temporal cortex that respond selectively when the person views pictures of objects in different categories, including animals, faces, tools, houses, and chairs. In general, the representations of different object categories within the inferior temporal cortex are topographically organized; that is, the spatial arrangement of the locations of the peak activations for tools, houses, etc. are consistent across subjects and tasks. At the same time, the representations are distributed and overlapping. They can be said to be distributed because the region of activation is not focal, but rather is typically a complex pattern spread over an expanse of cortex; they can be said to be overlapping, because even if the locations of the peak activations for two categories differ, there may be overlapping zones of activation as well. This pattern of results has led to the idea that the ventral and lateral temporal cortex contains a kind of feature space in which stored information about objects, especially shape information, is represented (Martin and Chao, 2001).

Similar patterns of activation are found when people are asked either to view or to imagine objects from different object categories. In addition, the same temporal cortical regions are activated when participants recognize, name, read about and answer questions about particular objects. These findings are consistent with the idea that these areas store information about objects, representing object-specific features and attributes.

The regions described so far appear to represent object categories, in that groups of neurons respond not to unique entities, but to almost any item in a given category. This leaves open the question of how individual objects are represented. In humans, the most commonly studied unique item has been photographs of famous faces. Interestingly, such studies usually find activations in the anterior middle temporal gyrus and temporal pole. Based on this and other information, the idea has arisen that more anterior regions of the temporal cortex are important for retrieving information about unique items.

As we have seen, the inferior temporal cortical areas of monkeys and humans share many functional properties. Indeed, taken together, the studies of monkey and human inferior temporal cortex provide converging evidence that this region is important for storing information about the visual world. So far, however, we have ignored potential differences between monkeys and humans in the organization of inferior temporal cortex. First, it should be noted that the regions that are important for representing color and form, while located in the most anterior portions of inferior temporal cortex in monkeys, are located in more caudal and ventral portions of inferior temporal cortex in humans. This leaves relatively uncharted a large expanse of anterior inferior temporal cortex in humans, which may be devoted to language-related processing. Second, in humans there is a division of labor between the left and right temporal lobes such that the left is predominately involved in language-related functions and the right predominately involved in processing complex geometric figures, such as abstract designs. This division is not absolute, as even damage to the right temporal lobe can affect linguistic processing, and furthermore, deficits that occur after damage to inferior temporal cortex in both hemispheres tend to be more severe than those that occur after damage to inferior temporal cortex in one hemisphere.

SUPERIOR TEMPORAL CORTEX

Much of the auditory sensory processing in primates takes place within the superior temporal cortex. Anatomical and physiological evidence from monkeys indicates that there are at least 12 separate cortical fields serving audition. As was the case for visual sensory processing, auditory sensory processing appears to proceed through a

series of stages. In primates, there are three primary or primary-like auditory cortical fields, often referred to collectively as the core region, located on the lower bank of the lateral sulcus (also known as the superior temporal plane), hidden by the overlying parietal cortex (Figure 2). These regions project to a belt region consisting of about seven additional cortical fields. The belt regions, in turn, project to a parabelt area, which occupies much of the superior temporal gyrus (Kaas *et al*., 1999).

Superior Temporal Cortex Function in Monkeys

All of the primary fields and many of the belt fields are cochleotopically organized: that is, different neurons in these regions represent different frequencies of sound, with a systematic shift from low to high frequency as one moves across the cortical field. The activity of neurons in the core and belt regions often reflects not only frequency, but also the temporal features and spatial location of an auditory stimulus. Whereas neurons in the core region are more active in relation to pure tones than to complex sounds (narrow-band noise), neurons in the belt areas show the opposite pattern. Thus, as in visual sensory processing, for auditory sensory processing there are multiple cortical fields containing representations of the auditory world, and there appears to be convergence of inputs from primary areas to higher-order areas.

In addition to primary and nonprimary auditory areas, the superior temporal cortex contains many polysensory fields, primarily related to combined processing of auditory and visual information (not illustrated). These lie largely in the depths of the superior temporal sulcus, in a position between the auditory cortical areas and the visual areas.

Superior Temporal Cortex Function in Humans

Primary auditory cortex in humans is located on the first temporal transverse gyrus (also known as the transverse gyrus of Heschl). There is some evidence to suggest that there is more than one cochleotopically organized field in this region. In addition, preliminary evidence suggests that there are several auditory cortical fields in this general region, many of which might be homologues to the fields identified in monkeys (Melcher *et al*., 1999). It has been proposed that auditory sensory processing, like visual sensory processing, may involve two main pathways, one for interfacing speech with object representations, and another for interfacing sensory with motor processes (Hickok and Poeppel, 2000). This view is in line with a division of ventral (temporal cortex) versus dorsal (parietal cortex) function in terms of perception and visuomotor control, respectively. However, it is possible that the two systems involved in speech processing both interact with parts of the ventral stream,

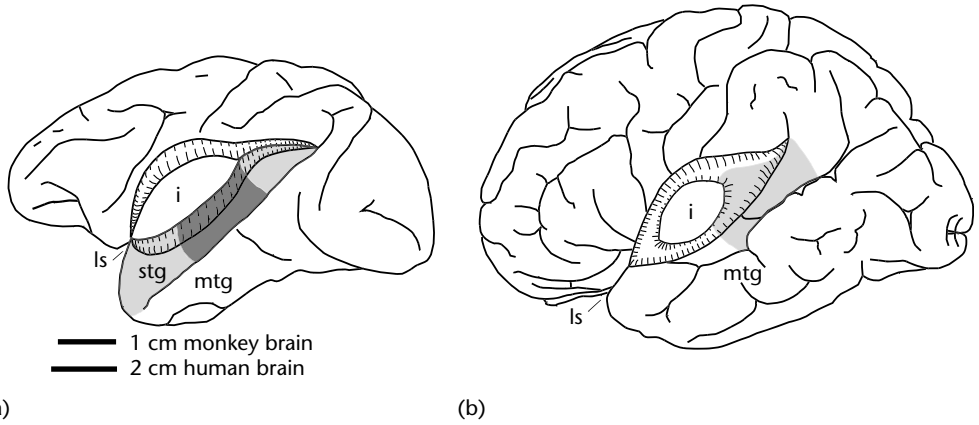


Figure 2. Location of temporal cortical areas (shaded regions) related to auditory processing in (a) macaque monkey and (b) human brains. The lateral sulcus has been opened to reveal cortex on the insula, and on the lower bank of the lateral sulcus, regions that are normally hidden from view. In the monkey brain, the dark shading shows the approximate location of the core, belt, and parabelt fields of auditory cortex. The lighter shading shows other auditory-related areas. In the human brain the shaded region includes both core and other auditory-related areas. Adapted from (a) Kaas *et al*. (1999) and (b) Hickok and Poeppel (2000). i, insula; ls, lateral sulcus; mtg, middle temporal gyrus; stg, superior temporal gyrus.

which, in turn, functions in both visual perception and visuomotor control.

Based on studies of patients with brain damage and on functional imaging studies, sound-based representations of speech are thought to be located bilaterally. After the initial analysis of speech sounds, regions in the temporal-parietal-occipital junction are thought to help interface the sound representations with other semantic and conceptual representations. Patients with bilateral damage in the region of the posterior superior temporal lobe exhibit profound speech perception deficits; although they may display normal pure tone thresholds, they are incapable of understanding auditorily presented words. This constitutes a type of aphasia, usually referred to as Wernicke aphasia. However, numerous subtypes of aphasia have been distinguished, and the term refers generally to a variety of deficits in either expressing or comprehending speech. (*See* **Audition, Neural Basis of**)

There are some differences in left and right hemispheric contributions to speech perception. The left hemisphere is more important than the right for the processing of fine-grained temporal events. Conversely, the right hemisphere may be more important than the left for the processing of spectral information. In addition, evidence from brain-damaged patients and from functional imaging studies suggests that auditory cortex in the left hemisphere may contribute to aspects of speech production.

PARALIMBIC CORTEX

Three cortical fields in the ventromedial aspect of the temporal lobe make up the paralimbic cortex: the entorhinal cortex, the perirhinal cortex, and the parahippocampal cortex (Figure 3). These regions have also been referred to collectively as the 'parahippocampal region', because of their proximity to, and anatomical relations with, the hippocampal formation: all three are six-layered cortical fields that have reciprocal anatomical connections with parts of the hippocampus and with each other.

The paralimbic cortical fields receive not only visual and auditory sensory inputs, but also an array of inputs from other sensory modalities. Indeed, they have long been recognized as a site of convergence of information arising from all the sensory systems. These fields receive anatomical connections from high-order modality-specific fields such as TE (for vision), regions on the superior temporal gyrus (for audition), areas in the caudal insula (for somatic sensation), as well as from areas that process visuospatial information, such as the retrosplenial cortex and posterior parietal cortex. Olfactory inputs also reach the paralimbic cortical

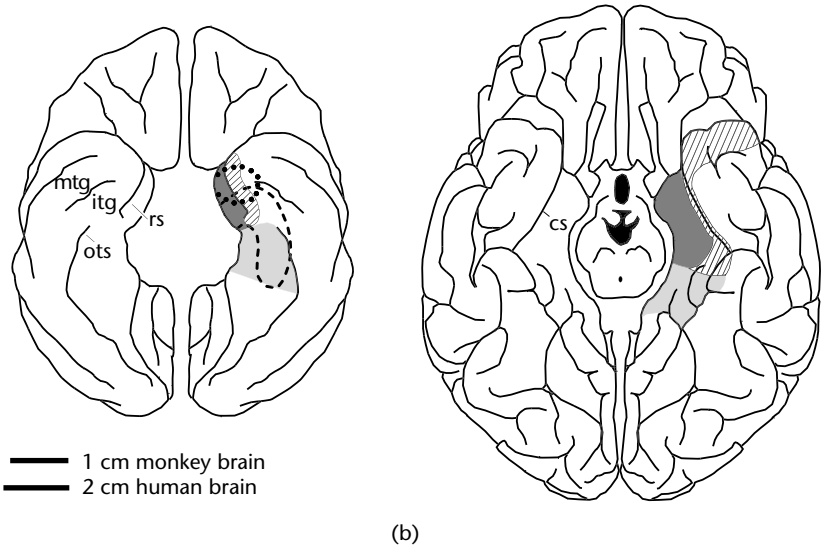


Figure 3. Locations of paralimbic cortex (shaded areas), shown on the ventral surface of (a) macaque monkey and (b) human brains. Dark gray shading represents the entorhinal cortex, oblique hatching shows the perirhinal cortex, and light gray shading represents the parahippocampal cortex. In the schematic of the monkey brain, heavy dotted and dashed lines show approximate locations of the amygdala and hippocampus, respectively. Schematic of human brain adapted from Burwell *et al.* (1996). cs, collateral sulcus; itg, inferior temporal gyrus; mtg, middle temporal gyrus; rs, rhinal sulcus; ots, occipitotemporal sulcus.

fields. Thus, these regions are in a position to bring together disparate pieces of information about objects, including their associated attributes. (*See* **Sensory Integration, Neural Basis of**)

Paralimbic Cortex Function in Monkeys

Although the paralimbic cortical fields have been delineated on anatomical grounds for nearly a century, until recently little was known about the functions of these regions. Historically it has been difficult to separate the functions of the paralimbic cortex from those of the adjacent hippocampus and amygdala, structures that lie deep to the paralimbic region (Figure 3). However, development of methods for making selective lesions has helped resolve this difficulty.

Of the cortical fields composing the paralimbic region, much more is known about the entorhinal and perirhinal cortex than about the more caudally located parahippocampal cortex. Unlike monkeys with damage to the inferior temporal cortex, who have extreme difficulty distinguishing between objects on the basis of vision, monkeys with damage to the paralimbic cortical fields can, for the most part, identify and select objects accurately. Damage to the entorhinal and perirhinal cortical fields produces a characteristic pattern of behavioral impairments including a deficient ability to judge whether an object has been seen or touched before (i.e. a deficit in both visual and tactual recognition memory); an inability to link together in memory items or other sensory inputs that have occurred together repeatedly (stimulus–stimulus associative memory); an impairment in remembering items that were familiar before the brain damage occurred (retrograde memory); and difficulty in certain other types of visual tasks (Murray, 2000). The activity of neurons in the entorhinal and perirhinal cortex mirrors these key functions. For example, single neurons in the paralimbic cortex carry information about whether an item has been seen before, about what other visual items or sensory inputs might have occurred in close proximity (temporally or spatially), and about where those items might have appeared. Preliminary data suggest that the parahippocampal cortex, in particular, may be important for certain aspects of spatial memory.

Although the primary function of the paralimbic region is generally accepted as mnemonic, and although its diverse anatomical relations set it apart from the other types of temporal cortex we have discussed, it is clear that many of the functions of perirhinal cortex are similar to those already described for the inferior temporal cortex. Indeed, neurons in both entorhinal and perirhinal cortex have many of the same properties as neurons in TE. (*See* **Amnesia**; **Neural Basis of Memory: Systems Level**)

For example, neurons in all three regions respond to complex visual stimuli and also carry signals reflecting whether items have been seen before. In addition, the magnitude of impairments on many visual tasks is graded in relation to the extent of combined damage to inferior temporal cortex, perirhinal cortex, and even entorhinal cortex, indicating that all these cortical fields are contributing to performance. For these and other reasons, it seems likely that the paralimbic cortex, especially the perirhinal cortex, is important not only for memory, as described above, but also is operating as an extension of the ventral visual stream. In the latter role, the perirhinal cortex is held to represent complex conjunctions of visual features, thereby extending the notion of the hierarchical organization of visual representations in the ventral visual pathway from inferior temporal cortex into the paralimbic cortex. Thus, the inferior temporal and paralimbic cortical fields together appear to make up a system devoted to the identification of objects.

Paralimbic Cortex Function in Humans

Knowledge of the functions of the paralimbic cortex in humans is sorely lacking. In humans, as in monkeys, damage to the medial temporal lobe that includes the perirhinal cortex yields deficits in visual recognition memory. As already indicated, more extensive damage that involves the inferior temporal cortex disrupts semantic memory as well. Damage to the parahippocampal cortex appears to yield deficits on spatial memory tasks, including difficulty in navigating in large environments.

Functional imaging studies report activations in the ventromedial region in relation to processing of object and word meaning, regardless of how the meaning is accessed (verbally, by Braille, or by pictures). Other studies suggest a role for parahippocampal cortex in representing the spatial layout of scenes and in navigating through space. Precisely how the perirhinal, entorhinal, and parahippocampal fields contribute to these cognitive functions in humans, how the contributions of the regions differ from those of the neighboring visual areas in the inferior temporal cortex, and whether the 'object' and 'spatial' related zones identified in human cortex are homologous with any of the identified paralimbic cortical fields in monkeys, is unknown.

References

Burwell RD, Suzuki WA, Insausti R and Amaral DG (1996) Some observations on the perirhinal and parahippocampal cortices in the rat, monkey, and human brains. In: Ono T (ed.) *Perception, Memory, and Emotion: Frontiers in Neuroscience*, pp. 95–110. New York, NY: Elsevier.

Farah MJ (1990) *Visual Agnosia.* Cambridge, MA: MIT Press.

Hickok G and Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* **4**: 131–138.

Hodges JR, Patterson K, Oxbury S and Funnell E (1992) Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain* **115**: 1783–1806.

Kaas JH, Hackett TA and Tramo MJ (1999) Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology* **9**: 164–170.

Martin A and Chao LL (2001) Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology* **11**: 194–201.

Melcher JR, Talavage TM and Harms MP (1999) Functional MRI of the auditory system. In: Moonen C and Bandettini PA (eds) *Medical Radiology: Diagnostic Imaging and Radiation Oncology*, pp. 393–406. New York, NY: Springer-Verlag.

Milner AD and Goodale MA (1996) *The Visual Brain in Action.* Oxford, UK: Oxford University Press.

Murray EA (2000) Memory for objects in nonhuman primates. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*, 2nd edn, pp. 753–763. Cambridge, MA: MIT Press.

Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA and Mansfield RJW (eds) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Wandell BA (1999) Computational neuroimaging of human visual cortex. *Annual Review of Neuroscience* **22**: 145–173.

Further Reading

Erickson CA, Jagadeesh B and Desimone R (2000) Learning and memory in the inferior temporal cortex of the macaque. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*, 2nd edn, pp. 743–752. Cambridge, MA: MIT Press.

Felleman DJ and Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**: 1–47.

Karnath HO (2001) New insights into the functions of the superior temporal cortex. *Nature Reviews Neuroscience* **2**: 568–576.

Kosslyn SM, Ganis G and Thompson WL (2001) Neural foundations of imagery. *Nature Reviews Neuroscience* **2**: 635–642.

Murray EA and Bussey TJ (1999) Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Sciences* **3**: 142–151.

Newsome WT (1997) The King Solomon Lectures in Neuroethology. Deciding about motion: linking perception to action. *Journal of Comparative Physiology A* **181**: 5–12.

Scharfman HE, Witter MP and Schwarcz R (eds) (2000) *The Parahippocampal Region: Implications for Neurological and Psychiatric Diseases*. Academy New York: New York Academy of Sciences.

Tanaka K (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* **19**: 109–139.

Thalamocortical Interactions and Binding

Introductory article

Antti Revonsuo, University of Turku, Turku, Finland

CONTENTS

Thalamocortical interactions are synchronous neural activities in bidirectional thalamocortical loops. They may serve as the neural mechanism of binding, or perceptual integration and the unity of consciousness.

INTRODUCTION

We experience the world around us as a holistic, subjective reality. The objects we perceive appear to us as coherent bundles of different perceptual features such as color, three-dimensional shape and motion. How is the unity of perception achieved by the cognitive and neural mechanisms in the brain? Elementary sensory features have to be bound together to form coherent perceptual objects, and the individual objects we perceive have to be bound to a common spatial framework so as to appear to us as parts of one globally unified perceptual world. The term 'binding' refers to the process or mechanism of this integration.

It has been suggested that synchronous neural oscillation could be the mechanism that binds together distributed neural elements to form representations of coherent perceptual wholes. The thalamocortical system can be regarded as a unified oscillatory machine, and the abundant bidirectional thalamocortical interactions could therefore have an important role in perceptual binding and the unity of consciousness.

THE BINDING PROBLEM

The binding problem appears in several different formulations in the different branches of cognitive science.

In neuroscience, the binding problem is the problem of the integration of single neuron activity into functional neuronal groups and assemblies. For example, neuroscientists interested in the neural mechanisms of vision formulate the problem in the following way: how do the thousands of anatomically separated neurons, responding to different parts or features of the same visual stimulus, integrate their activity into a neural representation of one single object?

In cognitive psychology, the binding problem is formulated in terms of information processing and representation. According to the modularity hypothesis, the input processing of perceptual information is handled by a multitude of isolated, specialized, mandatory, fast, nonconscious, neurologically specific modules. In a cognitive architecture like this, the binding problem is: how does the information initially processed by a multitude of independent modular input systems become integrated into coherent representations for perception, memory, and action?

The core of the problem is the fact that, at the level of conscious perception, there is an undeniable experiential unity. In subjective visual perception, objects appear as unified perceptual wholes located in one unified perceptual world. This phenomenal unity of subjective experience connects the modern binding problem to the classical philosophical problem of the unity of consciousness.

Immanuel Kant (1724–1804) was preoccupied with this problem. He thought that the world we experience is above all a world of perceived objects, and in order for the mind to produce such a complex and unified representation, the mind has to have some way to relate the different things it experiences to one another. He called it 'synthesis'; we now call it 'binding'. Kant distinguished between two levels of synthesis. Apperception is the process of synthesizing individual objects of awareness, such as binding the form, color and movement of a butterfly to a single coherent representation. In cognitive science, it is called 'feature binding' or 'feature integration'. Transcendental

apperception is the process that constructs the overall unity of consciousness: the representation of one single perceptual world where all the currently perceived objects are simultaneously present.

In experimental psychology, Gestalt psychologists dealt with the binding problem during the first half of the twentieth century. They described sensory organization in terms of perceptual units that are formed in the subjective perceptual field when the contents of particular areas are experienced to 'belong together'. They described these effects in terms of their famous 'Gestalt laws' of perception, and tried to explain them by referring to holistic fields in perception and in the brain. A modern formulation of the binding problem appears in Treisman's feature integration theory, which postulates specialized feature maps for basic visual features, a master map of locations that relates the features to specific locations in the visual field, and a scalable window of attention that scans the master map of locations. Scanning selects all the features that are currently at that location and binds them together into one coherent representation or temporary object file. This theory is formulated at the cognitive level of description and for the time being its neurophysiological basis remains unclear.

In neuroscience, the binding problem is closely related to the problem of neural coding. How is the environment represented in neural firing patterns? There are three competing hypotheses: the 'grandmother cell' theory; population or assembly coding; and temporal coding. The 'grandmother cell' theory assumes that every distinct perceptual element and every possible combination of elements – such as the perception of your grandmother – is represented by a distinct, dedicated neuron; this idea is nowadays regarded as implausible in the light of what we know about the neural basis of perception. The population or assembly coding theory says that the representation of a stimulus such as a visual object consists of a pattern of activity in a large group of neurons. This theory faces the 'superposition' problem: how is more than one stimulus entity coded at a time? The hypothesis of temporal coding addresses this problem by postulating that a cell assembly is distinguished from the other ones by the synchronicity of firing in all the members belonging to the same assembly. Different but simultaneously active assemblies can be distinguished from each other by the timing of their activity. Synchronous activity may constitute a recurrent regular pattern or oscillatory rhythm having a characteristic frequency of simultaneous neuronal discharges.

There is a growing body of theoretical and empirical evidence indicating that the synchronization of neural activity at high frequencies (especially gamma-band or 20–80 Hz synchronization) is associated with the processing of stimulus unity, such as the overall shape or the coherent gestalt of the stimulus. Francis Crick and Christof Koch were the first to explicitly connect neural synchronization with a theory of binding and consciousness. Their hypothesis is that binding involves an attentional mechanism that temporarily binds the relevant neurons together by synchronizing their spikes in 40 Hz oscillations. Subsequently, Wolfgang Singer proposed that only neural activation patterns that are sufficiently organized or coherent can reach the threshold of consciousness, and Andreas Engel further developed this coherence hypothesis by postulating that neural synchronization is necessary for access to visual awareness.

There is now ample evidence that, at the cortical level, synchronous neural activity is related to stimulus unity and to the unity of conscious visual perception. However, these studies are concerned with neural synchronization as a cortical phenomenon. Nevertheless, thalamocortical interaction probably has an important role in the generation of the unity of conscious perception.

PUTATIVE ROLE OF THE THALAMUS IN BINDING

The thalamus is a group of subcortical nuclei anatomically located in the very center of the brain. The thalamus has remarkable connections with virtually every region of the cortex. It has been traditionally regarded as the 'gateway to the cortex' or the brain's 'relay station', because all sensory inputs (apart from olfactory stimuli) make synaptic relays in the thalamus before reaching the sensory receiving areas in the cortex.

This traditional concept is an oversimplification, for the cerebral cortex receives input not only from the sense-specific nuclei, but also from the nonspecific thalamic nuclei that have multimodal connections to the cortex and other parts of the brain. The most remarkable feature of thalamocortical connectivity, however, is its bidirectionality. Thalamic nuclei receive reciprocal pathways from the same cortical areas that they project to. In fact, the number of corticothalamic fibers is significantly greater than the number of thalamocortical axons. These reciprocal thalamocortical connections create bidirectional neuronal loops between the thalamus and the cortex. Hence, the distributed neural representations of simultaneous perceptual features or

events could be related to each other within the thalamocortical system so as to bind input from different sensory modalities into a single perceptual event. Therefore, the thalamocortical system is a plausible candidate for a role in the binding or integration of multiple distributed representations to a coherent perceptual world.

THALAMOCORTICAL INTERACTIONS IN BINDING

The most detailed model of thalamocortical interactions and binding has been presented by the neuroscientist Rodolfo Llinás and his group. They base their model on two facts: first, there are abundant reciprocal thalamocortical connections that establish constant large-scale reverberating activity between the thalamus and cortex; and second, some cortical and thalamic neurons are capable of generating intrinsic 40 Hz oscillations. This leads to the view that the thalamocortical network can generate global oscillatory states on its own, even in the absence of sensory input. When we perceive an external stimulus, the intrinsic activity of the thalamocortical network is modulated by sensory input which thereby becomes incorporated into the functional state of the brain. In accordance with this view, magnetoencephalographic recordings have revealed continuous 40 Hz oscillations over large areas of the surface of the head both in the awake state and during rapid eye movement (REM) sleep. In the awake state, sensory stimulation is consciously perceived and causes a reset of the 40 Hz activity, but in REM sleep the stimulus neither resets the 40 Hz activity nor enters consciousness.

Llinás and Paré argue that there must be a mechanism that is capable of producing and maintaining the synchronized pattern of activity in remote groups of neurons and that this mechanism must have extensive projections to the cerebral cortex. They propose that the reticular thalamic nucleus cells could be responsible for the synchronization of the 40 Hz oscillations in distant thalamic and cortical territories. This mechanism involves two thalamocortical resonant loops: the specific thalamocortical loop and the intralaminar, nonspecific thalamocortical loop. The 40 Hz oscillation in the specific thalamic nuclei can establish a resonating thalamocortical loop through the projections to the cortical layer IV. The oscillation can be fed back to the thalamus via cortical layer VI pyramidal cells. The second thalamocortical loop is between the intralaminar nonspecific nuclei which project to cortical layer I and reenter via cortical layers V and VI pyramidal systems to the thalamus. The reticular nucleus of the thalamus is in interaction with both of these loops and could thus synchronize the activity in both of them.

This model of thalamocortical interaction can be directly related to two different types of binding. The specific thalamocortical loop is assumed to be responsible for the binding of distributed sensory fragments into single coherent objects of awareness (Kant's 'apperception' and Treisman's 'feature integration'). The nonspecific thalamocortical loop is assumed to provide the overall context or functional conscious state where the individual objects of awareness are related to each other within one globally coherent representation (Kant's 'transcendental unity of apperception'). Consistent with this model, lesions of specific thalamic nuclei abolish modality-specific contents of consciousness whereas lesions of nonspecific thalamus (especially the intralaminar nuclei) abolish the global background state of consciousness, resulting in coma. Thus, the interaction of these two thalamocortical loops through synchronous oscillatory activity around 40 Hz is hypothesized to take care of the binding of perceptual content into a single coherent experience.

Other theories of the mechanisms of binding have been less explicit about the role of thalamocortical interactions. The theory presented by Wolfgang Singer's group describes the cortical mechanisms of synchronization in detail, but only mentions in passing that attention could function as an internal synchronizing mechanism through thalamocortical connections. Empirical evidence suggests that the neocortex and thalamus engage in a permanent dialogue and form a unified oscillatory machine, but direct empirical evidence relating thalamocortical interactions to the unity of conscious perception is hard to come by. However, corticothalamic feedback connections from the visual cortex to the lateral geniculate nucleus (the modality-specific visual nucleus in the thalamus) in fact synchronize the responses of the thalamic cells projecting to the visual cortex. Further experimental evidence will be necessary to confirm the hypothesis about thalamocortical resonating loops and their role in binding.

CONCLUSION

The binding problem is a set of related problems that can be formulated at different levels of description, as concerning the integration of neural activity, the integration of modular cognitive processing, or the classical problem of the unity of consciousness. Empirical evidence supports the

view that synchronization could be the neural mechanism of at least some types of perceptual binding. Evidence also suggests that thalamocortical loops are involved in at least some of the synchronous neural activities related to perception. Furthermore, the physiological properties of thalamocortical connections make them intriguing candidates for having a central role in many types of binding.

The most detailed model describing the thalamocortical system as a possible binding mechanism is by Rodolfo Llinás and his group. The correctness of their model remains to be confirmed by direct experimental evidence. Most experiments on humans are capable of measuring cortical synchrony only. For methodological reasons, it will be difficult to design experiments that directly test the role of thalamocortical loops in the integration of perception in humans.

It may be that the unity of consciousness is created in the thalamocortical system, as Llinás and Paré speculate, but for the time being it remains a fascinating hypothesis rather than an established empirical fact.

Further Reading

Brook A (1994) *Kant and the Mind*. New York, NY: Cambridge University Press.

Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2**: 263–275.

Engel AK, Fries P, König P, Brecht M and Singer W (1999) Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* **8**(2): 128–151.

Fodor JA (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Gray CM (1999) The temporal correlation hypothesis of visual feature integration: still alive and well. *Neuron* **24**: 31–47.

Joliot M, Ribary U and Llinás R (1994) Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences of the USA* **91**: 11748–11751.

Jones EG (2001) The thalamic matrix and thalamocortical synchrony. *Trends in Neurosciences* **24**: 595–601.

Köhler W (1947) *Gestalt Psychology*. New York, NY: Liveright.

Köhler W (1957) Psychologie und Naturwissenschaft. In: Henle M (1971) (ed.) *The Selected Papers of Wolfgang Köhler*, pp. 252–273. New York, NY: Liveright.

Llinás R and Paré D (1991) Of dreaming and wakefulness. *Neuroscience* **44**: 521–535.

Llinás R and Paré D (1996) The brain as a closed system modulated by the senses. In: Llinas R and Churchland PS (eds) *The Mind-Brain Continuum*, pp. 1–18. Cambridge, MA: MIT Press.

Llinás R and Ribary U (1993) Coherent 40 Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences of the USA* **90**: 2078–2081.

Llinás R and Ribary U (1994) Perception as an oneiric-like state modulated by the senses. In: Koch C and Davis JL (eds) *Large-Scale Neuronal Theories of the Brain*, pp. 111–124. Cambridge, MA: MIT Press.

Llinás R, Ribary U, Contreras D and Pedroarena C (1998) The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **353**: 1841–1849.

Newman J (1995) Thalamic contributions to attention and consciousness. *Consciousness and Cognition* **4**(2): 172–193.

Revonsuo A (1999) Binding and the phenomenal unity of consciousness. *Consciousness and Cognition* **8**(2): 173–185.

Revonsuo A, Wilenius-Emet M, Kuusela J and Lehto M (1997) The neural generation of a unified illusion in human vision. *NeuroReport* **8**: 3867–3870.

Roelfsema PR and Singer W (1998) Detecting connectedness. *Cerebral Cortex* **8**(5): 385–396.

Roelfsema PR, Engel AK, König P and Singer W (1996) The role of neuronal synchronization in response selection: a biologically plausible theory of structured representations in the visual cortex. *Journal of Cognitive Neuroscience* **8**: 603–625.

Roskies AL (ed.) (1999) Reviews on the binding problem. *Neuron* **24**: 7–125.

Sillito AM, Jones HE, Gerstein GL and West DC (1994) Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature* **369**: 479–482.

Singer W (1994) The organization of sensory motor representations in the neocortex: a hypothesis based on temporal coding. In: Umilta C and Moscovitch M (eds) *Attention and Performance XV, Conscious and Nonconscious Information Processing*, pp. 77–107. Cambridge, MA: MIT Press.

Singer W (1999) Neuronal synchrony: a versatile code for the definition of relations? *Neuron* **24**: 49–65.

Steriade M (1999) Coherent oscillations and short-term plasticity in corticothalamic networks. *Trends in Neurosciences* **22**: 337–345.

Tallon-Baudry C and Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences* **3**(4): 151–162.

Treisman A (1988) Features and objects: the Fourteenth Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology* **40A**: 201–237.

Treisman A (1996) The binding problem. *Current Opinion in Neurobiology* **6**: 171–178.

Zeki S (1992) *A Vision of the Brain*. Oxford, UK: Blackwell.

Time Perception and Timing, Neural Basis of

Introductory article

John Gibbon, Columbia University and New York State Psychiatric Institute, New York, USA
Chariklia Malapani, Columbia University and New York State Psychiatric Institute, New York, USA

CONTENTS

INTRODUCTION

Timing is everything – whether in making shots, in making love or in making dinner. Indeed, it is difficult to conceive of an action that does not require temporal control. Modern humans have invented mechanical clocks that tell them the time accurately, but our hunter-gatherer forebears did not enjoy this temporal precision, and relied on their endogenous sense of elapsed time. In this respect, our forebears were in exactly the same position as animal foragers, estimating the rate of return in a given patch in order to know when to travel to another area. For mobile organisms, virtually any movement at all is timed, often sequenced in complex ways and temporally coordinated (e.g. in the approach-and-capture sequences of predators and the mating rituals displayed by many species). Exquisite temporal coordination is exemplified in the performances of skilled athletes and musicians, but even walking, driving a car or washing one's hands involves temporal sequencing of coordinated activities.

In addition to coordinating complex sequences, timing serves a very basic function that has long been recognized as a fundamental aspect of the learning process in animals, namely anticipation or prediction. Sometimes important events are associated with exteroceptive cues (e.g. a rustle in the grass, the cast of a shadow, or a whiff of smoke) that function as conditioned stimuli, triggering sympathetic reactions. In other cases the only signal is the temporal regularity of the event itself, as occurs in temporal conditioning. Evolution has selected for neurobehavioral mechanisms that anticipate events as well as react to them, making animals do the right thing at the right time. Whereas the 'origin' for time may be in the future (anticipatory timing), or it may be set by a pattern of responses (coordinated timing), animals are often under the control of the time that has elapsed since some prior marker, known as interval timings as opposed to periodic-oscillatory timing.

The last decade has seen a growing interest in quantitative analysis of timing behavior, and also in new theoretical developments that integrate models of timing with insights from areas of behavioral analysis and neurobiology. In previous decades, the idea that animals can represent time was resisted, partly because it was difficult to imagine a plausible physiological basis for this representational capacity. However, research revealed that multicellular organisms possess endogenous oscillators whose periods range from less than a second to minutes, hours, days, weeks or even a year. When a system contains within itself oscillations with widely differing periods, it can represent the time when an event occurred by recording the momentary phases of these oscillatory processes. The endogenous oscillators provide a physiological foundation for the capacity to represent time – the so-called *internal clock*.

PROPERTIES OF TIMING SYSTEMS

There are two basic forms of mechanisms or processes that make possible the recording of moments in time and the determination of temporal intervals, namely oscillatory processes and non-oscillatory decay or accumulation processes.

Periodic-oscillatory Timing

Life evolved in a cyclic environment caused by the relative movements of the earth, sun and moon. Biological clocks are generally thought to have

evolved as adaptations to these geophysical cycles. When studying biological rhythms, researchers actually look at the rhythmic processes (daily, lunar, monthly or annually periodic lengths) and make inferences about the biological clock itself. However, the biological clock is separate from the behavioral rhythms that it drives. Processes become rhythmic when they are coupled to the biological clock.

A defining property of clock-controlled rhythms is that cycles continue in the absence of environmental cues such as light–dark and temperature cycles. However, in the constant conditions of the laboratory, the period of the rhythm is rarely exactly what it was in nature – that is, it becomes slightly longer or shorter. The assumption is made that the period length is a reflection of the rate at which the clock is running. This point is emphasized by describing the circadian (close to 24-hour) period length in constant conditions as free-running, implying that it is no longer manipulated by environmental cues. Although the period length of biological rhythms is typically 'circa' in the constant conditions of the laboratory, in nature each period length matches that of its geophysical cycle, because the clock is 'entrained' to (i.e. locked on to) its corresponding environmental cycle. The 'entrainment' principle is illustrated by the response of a marcher to the beat of a drum. If the marcher hears the timing signal after his foot has struck the ground, he slows his step, so that his next footfall coincides more closely with the timing signal. If he hears the timing signal before his foot has struck the ground, he accelerates his next step, which again makes the next footfall coincide more closely with the drumbeat. The marcher's response to the timing signal – whether it involves decelerating or accelerating his step – depends on where in his stepping cycle he hears the drum (before or after the fall of his foot).

Interval Timing

As life flows inexorably onward, an animal's hold on its past requires memory for events, their times and the relationships between them. Only timing systems that are capable of starting and stopping at will and covering a very broad range of times can manage these key functions. Such interval timing systems must be extremely flexible, and this flexibility is achieved at the cost of precision (the precision of interval timing systems can vary from 5% to about 60% of the interval being timed, in contrast to circadian timing, which show variability as low as 1% for the 24-hour cycle).

During the past two decades, different theories have emerged aimed at elucidating the psychophysical properties of interval timing and/or providing better tools for the quantitative analysis of mechanisms that living organisms use to time intervals. The *scalar expectancy theory* (SET) remains the most prominent of the theoretical accounts. It originated from reliable findings of cross-species similarities in the psychophysical properties of timing from a variety of different paradigms utilizing the time sense. Figure 1 illustrates two main properties of interval timing, namely *flexible accuracy* and *scalar variability*, which are evident in a variety of time perception as well as production tasks, and in different time ranges.

Figure 1a shows the mean time between taps by subjects synchronizing tapping with a metronome at various inter-stimulus intervals. Subjects are very nearly veridical in their average tapping rate over an order of magnitude, often within 1 millisecond of the target interval. In the same graph, the standard deviation of the inter-tap intervals (right ordinate), although small, increases almost proportionally with the size of the target interval. This finding is known as the *scalar property*. Figure 1b shows data from college students estimating three intervals. Subjects were asked to remember a target time presented as the duration of a visual signal. In subsequent trials, the signal was turned on, and subjects reproduced the time by responding on the space bar just before they thought the target time had elapsed, and stopping just after it. What is shown is the relative frequency distribution of the estimates around the target intervals. The three distributions (shown in the left panel, Figure 1b) superpose when rescaled relative to their target size (right panel, Figure 1b), providing a more complete example of the scalar property. Both flexible accuracy and scalar variability are also evident in time perception tasks. Figure 1c shows data from a perceptual timing task in which subjects are asked to remember two anchor times – a short time and a long time. They are then probed at intermediate values and asked to respond 'short' or 'long', depending on their perceived relationship to the anchors. The judgments are quite accurate. The proportion of 'long' reports increases from about 0 at the short anchor to about 1.0 at the long one. As the short–long pairs are further separated, the functions show decreasing slopes. This decrease again reflects the scalar property, illustrated when the data are replotted (right panel) relative to the point of subjective equality (PSE) (i.e. the point where subjects are maximally uncertain whether to categorize the stimulus as closer to the long or short anchor).

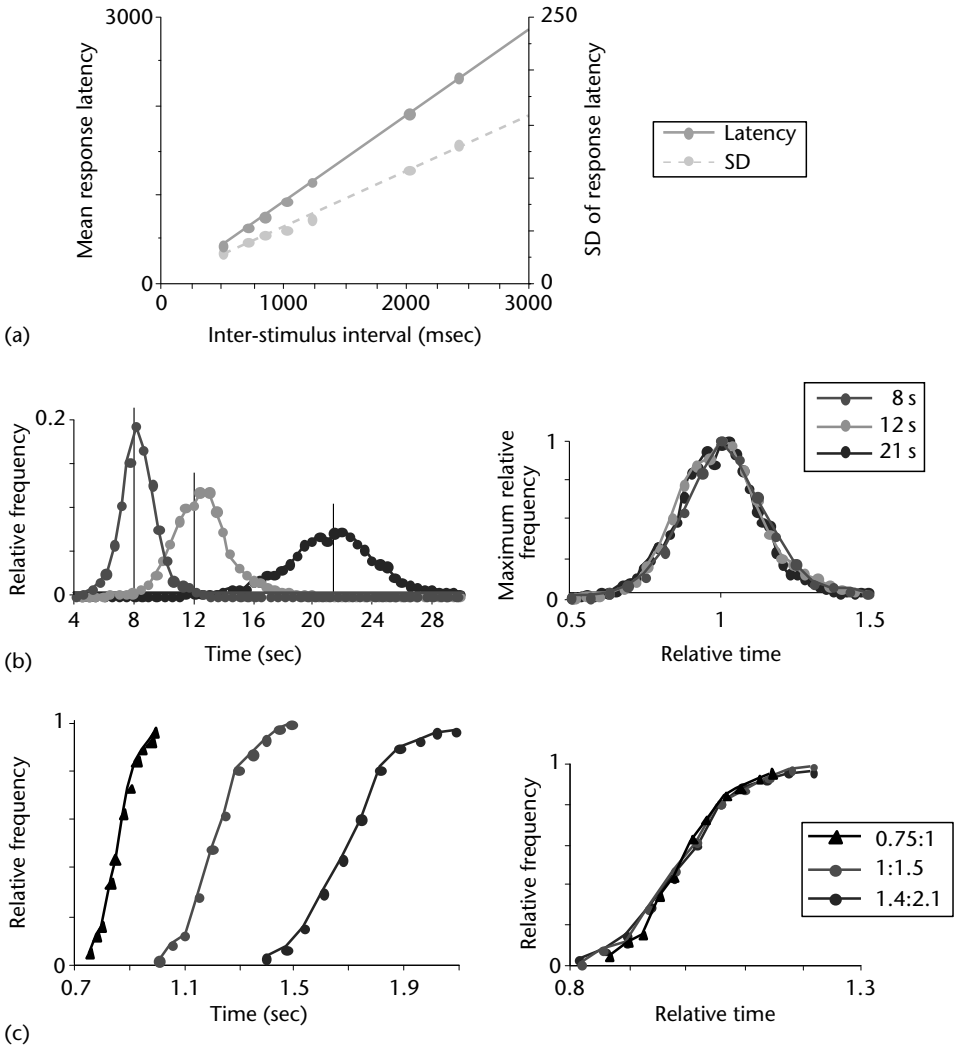


Figure 1. Flexible accuracy and scalar variability of time estimates. Flexible accuracy: subjects are very accurate in their mean reproduction of a known time, which may be an arbitrary value over several orders of magnitude. Scalar variability: subjects in a variety of temporally controlled tasks show error rates that are proportional to the absolute value of the times being estimated. This multiplicative increase implies that essentially one timing distribution is used which is then re-scaled to any arbitrary target time. (a) Tapping data. (b) Peak interval data. (c) Bisection data.

The scalar expectancy theory deals elegantly with yet another psychophysical property of interval timing, which together with flexible accuracy and scalar variability puts important constraints on any mechanism that is purported to underlie timing behavior. That is the *ratio decision*. The comparison process for generating responses involves a ratio of remembered time to current time. Consider for example, the classical conditioning that permits prediction of important events – the unconditioned stimulus (US) – with a formerly 'neutral' stimulus, such as a tone or a light – the conditioned stimulus (CS). After several such pairings, subjects respond to the neutral signal in a similar way to that in which they respond to the important event itself. Significantly, for our purposes, the CS (the neutral signal) is a better predictor than anything else in the environment for the delivery of the US (the important event). That is, the signal must be proximal to the event and not, for example, present continuously or for a long time prior to the event. Pavlov noted that after training his dogs by pairing the sound of a bell and the delivery of meat powder, they would salivate when he entered the laboratory. However, they did not respond for long. It was only his sudden appearance that induced this effect. Thus a very long CS only appeared to have an effect at its onset, and its continued presence no longer elicited the conditioned response. Recently, attention has focused on the relationship between

the duration of the signal and the duration of the time between signaled episodes. This relationship is a relative one. It is the ratio of the trial duration to the inter-trial duration that matters. The ratio – not the difference – controls the speed with which subjects learn about the CS. Subjects acquire similarly to CS of 10 seconds spaced 100 seconds apart and CS of 5 seconds spaced 50 seconds apart. Moreover, the variability of acquisition scores is approximately constant at a given ratio, and decreases roughly proportionally to increasing ratios (another example of the scalar property).

Using the scalar expectancy theory, a quantitative description of the patterns of temporal data is linked to different psychological constructs (clock, memory and comparison/decision mechanisms), as illustrated in Figure 2.

NEURAL BASIS OF TIMING SYSTEMS

Despite the major advances that have been made during the last decade, there is continuing debate about the exact nature of the neural systems and/or brain networks that underlie the functions of timing systems.

Circadian Timing

There is compelling evidence that biological clocks exist in single cells, and that there may be many such clocks in a single individual. The suprachiasmatic nucleus (SCN) in the hypothalamus is widely recognized as the primary clock in circadian timing, but its 'master' role in regulating different rhythms is still debated. Its role may depend on whether information is transmitted through chemical or neural influx within the nucleus itself or its different targets. Figure 3 summarizes the current consensus about circadian organization in mammals.

Time Perception and Timing

Some advances have been made with regard to specifying the brain systems involved in interval timing systems (mainly the basal ganglia and the cerebellum), and links have been established between the cognitive processes underlying timing and specific pharmacological and physiological manipulations. However, there is still controversy about the way in which neural networks involving these structures, either independently or in conjunction, influence timing capacities.

Animal research has mainly emphasized the role of dopaminergic brain systems and the striatofrontal circuitry in timing behavior. Specifically, dopaminergic neurons that originate in the substantia nigra pars compacta (SNc) and project to the striatum are thought to represent the 'pacemaker'

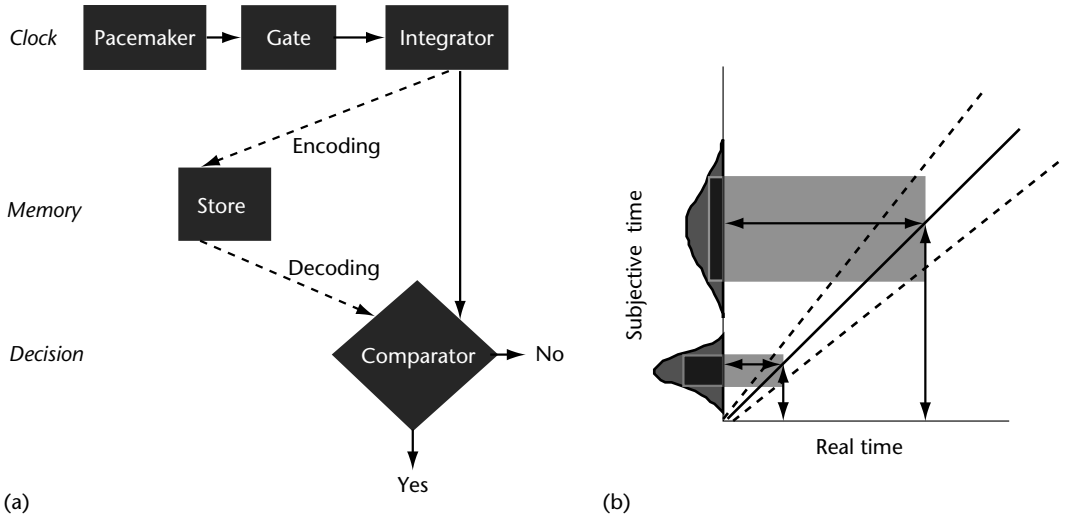


Figure 2. (a) Information-processing components of the scalar timing model. An internal pacemaker/integrator monitors the passage of time (clock stage). Accumulated subjective time is occasionally an important (reinforced) time, and these time records are transferred (encoded) to memory (memory stage) for later comparison (decision stage) with the passage of current time. The decision stage compares current time with time retrieved from memory and identifies an appropriate response outcome. Scalar variance may enter in the clock or memory stage, or through threshold variance in decision. (b) Subjective time increasing linearly with real time. Variation in the rate of the subjective integration (broken lines) generates subjective time distributions for real time (on the ordinate), which are scale transforms of each other (the scalar property shown on the abscissa).

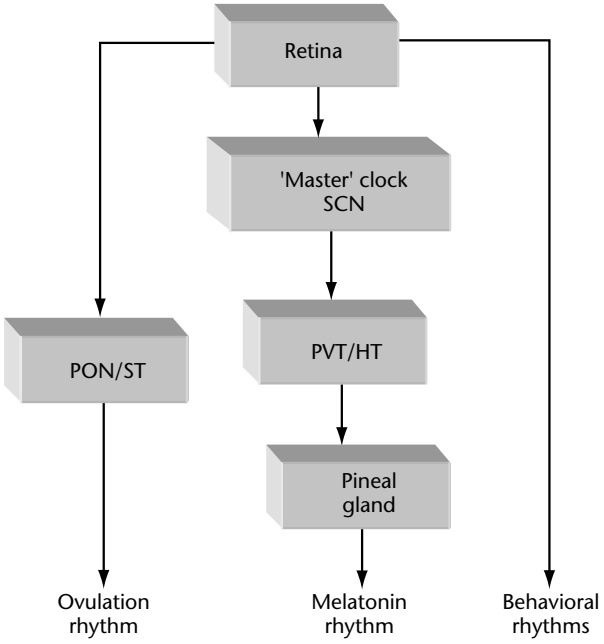


Figure 3. Neural correlates schematic of circadian organization in mammals. The circadian system in birds seems to contain three interacting clocks: the pineal gland, the suprachiasmatic nuclei (SCN) of the hypothalamus, and the eyes. This system of interacting clocks controls other clocks in the body through the pineal gland's rhythmic output of the hormone melatonin. The photoreceptors for entrainment in mammals are the eyes. The information reaches the SCN via the retinohypothalamic neural tract. Whereas some signals from the SCN are sent via neural pathways, others are sent by small diffusible molecules. The assumption is that the SCN is the 'master' clock for certain rhythms, but not for others. The SCN has at least two neural output pathways that affect rhythms – one to the preoptic nucleus of the hypothalamus (rhythm in ovulation): and a second pathway that leads to the paraventricular nucleus in the hypothalamus and then to the pineal gland. Finally activity rhythms are thought to be caused by the SCN's release of a chemical signal to other parts of the brain without the help of neural connections. Chemical control of rhythms by the SCN seems to be unique to behavioral rhythms such as activity and drinking. Reproductive responses to daylength, which depend on melatonin from the pineal gland, and endocrine rhythms require neural connections. PON/ST, paraventricular nucleus in the hypothalamus; PVT/HT, paraventricular nucleus in the hypothalamus.

(clock) in interval timing. However, the striatum is also believed to serve as the 'accumulator' that integrates the action potentials of dopaminergic pacemaker cells. Consistent with this hypothesis, striatal lesions eliminate timing in rats, and timing does not recover after administration of L-Dopa to the lesioned striata. However, L-Dopa does restore timing in animals with damage to the SNc, which suggests that pacemaker-dopaminergic cells that survive the lesion can act effectively again under L-Dopa supplementation. Research also suggests that whereas the pacemaker's speed in emitting pulses would depend on the integrity of the striato-nigral dopaminergic system and the striatum itself, the acetylcholine systems modulate both short- and long-term memory for time. Moreover, short-term memory specifically depends on hippocampal integrity, while long-term memory depends on the integrity of the frontal cortex. A few studies suggest that lesions in the cerebellar cortex impair animal timing behavior in both perceptual timing tasks and temporal conditioning.

Most human timing research has also focused on the basal ganglia, the frontal cortex and the cerebellum as the brain areas that are most likely to be involved in timing and time perception. For example, imaging studies have found activity during both perception and production timing tasks in the basal ganglia and the cerebellum, as well as in their cortical target areas. Moreover, patients with damage to the basal ganglia or the cerebellum are impaired with regard to both time perception and production tasks. Taken together, these findings provide clear evidence that lesions of the basal ganglia and frontal cortex – but not of the cerebellum – impair accuracy, while dysfunction in either area increases variability in timing and time perception. Although increased variability is the norm, it is still unclear whether this effect stems from dysfunction(s) in clock and/or memory, or decision mechanisms.

More recently, clinical studies have provided compelling evidence that distinct neural brain mechanisms are related to specific distortions of the three psychophysical properties of timing defined above. For example, increased overall variability that remains scalar without accuracy distortions is associated with lesions in the lateral cerebellar cortex and the cerebellar nuclei. However, basal ganglia damage is associated with distortions in accuracy, in the form of either under- or over-estimation, together with violation of the scalar rules on variability during both perceptual and reproduction timing tasks. Moreover, dopaminergic deficiency in Parkinson's disease results in accuracy distortions which differ depending on whether information is being encoded to or decoded from temporal memory. While encoding different time intervals in Parkinson's disease, distortions of subjective time reveal linear (unidirectional) overestimation of all intervals. In contrast, while retrieving the trace of two or

more different time intervals normally encoded in previous training, patients with Parkinson's disease show a 'coupling' of remembered values such that the values appear to be more similar to each other. In addition to this coupling phenomenon, decoding deficits in Parkinson's disease are accompanied by a violation of the scalar variability rules, whereas this psychophysical property holds with the unidirectional shifts (i.e. lengthening of subjective time) that are observed during learning.

Although similarities between the findings of animal and clinical studies point to specific brain areas as being involved in timing and time perception, important differences between animal and human data must be noted. For instance, timing distortions in human basal ganglia diseases involve primarily temporal memory and/or decision processes, whereas animal research has provided evidence that timing deficits caused by damage to the striatum result from dysfunctional monitoring of current time. Moreover, bidirectional shifts in accuracy and non-scalar variability induced by dopamine-dependent distortions in the retrieval of traces from temporal memory have not been previously reported in animals. The question of whether these discrepancies reflect species differences can only be answered by future research.

SUMMARY

Our current understanding of the psychophysics of temporal processing far exceeds our understanding of the neural substrates involved. There are also major psychophysical findings that are not well understood, in particular the sources of non-scalar variability and plausible differential transformations that may occur while encoding and/or decoding temporal memories. These unresolved issues await new theoretical developments in the modeling of interval timing and time perception.

Further Reading

Aschoff J (1981) *Biological Rythms*. New York, NY: Plenum.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: Bradford Books/MIT Press.

Gallistel CR and Gibbon J (2000) Time, rate and conditioning. *Psychological Review* **107**: 289–344.

Gibbon J, Malapani C, Dale C and Gallistel CR (1997) Toward a neurobiology of temporal cognition: advances and challenges. *Current Opinion in Neurobiology* **7**: 170.

Ivry RB (1996) The representation of temporal information in perception and motor control. *Current Opinion in Neurobiology* **6**: 851.

Malapani C, Deweer B and Gibbon J (2002) Separating storage from retrieval dysfunction of temporal memory in Parkinson's disease. *Journal of Cognitive Neuroscience* **14**: 311.

Meck WH (1996) Neuropharmacology of timing and time perception. *Cognitive Brain Research* **3**: 227–242.

Zakay D and Block RA (1997) Temporal cognition. *Current Directions in Psychological Science* **6**: 12–16.

Tourette Syndrome

Introductory article

Thomas M Hyde, National Institute of Mental Health, Bethesda, Maryland, USA

CONTENTS

Tourette syndrome is a movement disorder that begins in childhood and is characterized by chronic motor and vocal tics. Clinical features span the disciplines of neurology and psychiatry. Effective pharmacological therapy is now available.

INTRODUCTION

Gilles de la Tourette syndrome (TS) is a movement disorder that crosses the boundary separating neurology from psychiatry. It is defined by chronic motor and vocal tics that begin in childhood, but psychiatric comorbidity also is an important feature. Motor tics – repetitive, involuntary stereotyped movements – most often involve the head and neck, but may involve any muscle group. Motor tics can be simple, such as forceful eye blinking, or complex, such as tugging on clothing. Vocal tics – repetitive, involuntary stereotyped vocalizations – are usually unintelligible sounds, such as sniffing or grunting, but can also be complex, involving words or even whole phrases. Other interesting complex phenomena, occurring less commonly, include coprolalia (involuntary and affectively inappropriate swearing), copropraxia (involuntary and affectively inappropriate use of obscene gestures), echolalia (involuntary repetition of the speech of others), echopraxia (involuntary imitation of the actions of others) and palilalia (involuntary repetition of parts of the individual's own speech). Tics come and go, even remitting completely for extended periods. Among psychiatric symptoms, TS is often associated with obsessive–compulsive traits such as repetitive hand-washing or checking rituals, and attention deficit hyperactivity disorder (ADHD), although these are not part of the current diagnostic criteria. While tics are the hallmark of TS, they also can be seen with other neurological disorders and as a side effect of medications.

The first description of a syndrome of motor and vocal tics was reported by the nineteenth-century French neurologist Jean-Marc Itard. His patient, the Marquise de Dampierre, a French noblewoman, developed motor tics in childhood and shortly thereafter developed involuntary vocalizations consisting of screams and piercing cries. Several years later she developed coprolalia and echolalia. On account of these problems this unfortunate woman was forced to live in seclusion, continuing her involuntary cursing until death at age 85. Some fifty years after Itard's report, Georges Gilles de la Tourette produced a detailed account of several patients with a similar condition, among them the Marquise in her later years. This account was the first case series that clearly established tic disorders as an important and distinct clinical entity. Jean Martin Charcot, a leading French neurologist of the nineteenth century and Gilles de la Tourette's mentor at the Salpêtrière, attached his pupil's name to this syndrome.

For most of the past century, TS was classified as a psychiatric disorder and treated by psychiatrists with a variety of therapies. The tics were thought to represent underlying psychopathology. This confusion was understandable, as the tics could be suppressed voluntarily, at least transiently, stress often led to an acute exacerbation, and many tics looked extremely bizarre. More recently, as the biological basis of TS and allied conditions has become apparent, including the efficacy of medications and the heritability of the disorder, TS appropriately has been reclassified within neurology as a movement disorder. With this change in perspective, the psychiatric comorbidity common in TS patients has been relegated to the sidelines, despite its importance. The syndrome has prominent behavioral as well as motor manifestations, and occupies a niche that spans neurology and psychiatry.

Tourette syndrome usually develops in childhood, with a median age at onset of 7 years. For many patients the disorder is lifelong. The prevalence is 0.3–0.5 per 1000. The syndrome is three to

four times more common in men than in women, and up to nine times more common in boys than in girls. Although in most cases the tics persist throughout life, there is often a decline in the severity of symptoms after puberty.

GENETICS

Tourette syndrome is the most severe form of the tic disorders, disorders that probably share a common genetic basis. There is a spectrum of severity, ranging from transient motor and vocal tic disorders to full-blown TS. The differentiation of these subtypes of tic disorders primarily depends upon the duration, number and type of tics (Table 1). The character of the tics themselves, however, is similar throughout the spectrum of tic disorders. Many of the genetic studies of tic disorders also have suggested a common genetic aetiology for tics and obsessive–compulsive disorder (OCD), which is common in many TS family trees. Obsessive–compulsive disorder is much less common in sporadic cases of TS than in cases with a strong family history.

In some pedigrees, TS and chronic tic disorders follow an autosomal dominant pattern of inheritance, with variable penetrance. Some have contested this mode of transmission. Segregation analysis in large families has shown that penetrance is gender-related. For males, penetrance is nearly complete when all tic disorders are included. For females, penetrance increases from 56% for tic disorders to 70% when OCD is included as part of the phenotypic expression of the TS gene or genes. This observation, and the familial association of OCD and TS, have led to the theory that OCD is a clinical variant in the expression of the TS genetic defect. In fact, obsessive–compulsive symptoms are common in TS patients, although only a minority meet the full diagnostic criteria for OCD. Despite intensive investigation with a number of techniques and patient cohorts, no specific gene has yet been linked to TS.

Nature Versus Nurture: Interactions Between Genes and Environment in TS

Genes and environment appear to interact in determining the clinical expression of the TS gene or genes. Two separate twin studies have reported concordance rates of about 50% for TS in monozygotic (identical) twins but less than 10% in

Table 1. Diagnostic categories for tic disorders

Tourette syndrome
- Multiple motor AND one or more vocal tics occur concurrently at some time during the course of the illness
- Tics occur daily, or nearly daily, intermittently over the course of at least a year
- Anatomic location, number, frequency, complexity or severity of tics changes over time
- Onset is before age 21 years
- There is no concurrent medical or neurologic condition that could explain the tics (e.g. stimulant exposure or Huntington disease)

Chronic multiple motor tic or phonic tic disorder
- Either multiple motor OR vocal tics, but NOT both, have been present at some time during the illness
- Tics occur daily, or nearly daily, intermittently over the course of at least a year
- Anatomic location, number, frequency, complexity or severity of tics changes over time
- Onset is before age 21 years
- There is no concurrent medical or neurologic conditions that could explain the tics (e.g. stimulant exposure or Huntington disease)

Chronic single tic disorder
- Criteria are the same as above except the patient has only a single motor or vocal tic for at least a year

Transient tic disorder
- There are single or multiple motor AND/OR vocal tics
- Tics occur nearly every day for at least 2 weeks but for no longer than 12 consecutive months
- There is no previous history of Tourette syndrome or chronic motor or vocal tic disorder
- Anatomic location, number, frequency, complexity or severity of tics changes over time
- Onset is before age 21 years
- There is no concurrent medical or neurologic condition that could explain the tics (e.g. stimulant exposure or Huntington disease)

Adapted from the Tourette Syndrome Classification Study Group.

dizygotic (fraternal) twins. When all tic disorders are considered, the concordance rate rises to nearly 80% in monozygotic (MZ) twins and about 25% in dizygotic (DZ) twins. While these results suggest a strong genetic contribution to TS, they also indicate that nongenetic factors must also have an important role, since only about half of the MZ twins are concordant for full-blown TS. This fact is illustrated by studies of identical twins, where the twins exhibit significant differences in the frequency, intensity and bodily distribution of tics, despite their identical genetic endowment.

Since TS is much more common in males than in females, sex-specific hormones probably play an important role. Testosterone, present in higher levels in males than females, might enhance tic production; alternatively, oestrogen and/or progesterone, present in higher levels in females than males, might suppress tics. In addition, a number of medications, including stimulants (such as amphetamines) and neuroleptics (such as haloperidol), can produce tics in people with no known genetic predisposition. Stimulants in particular may also aggravate preexisting tic disorders.

Since MZ twins have an identical genes and similar but not identical environmental experiences, twin studies are ideal for examining the relative contribution of genetic and environmental factors. One study of six MZ twin pairs discordant for TS found that the twin with TS consistently had a lower birthweight than the unaffected co-twin. A second study found that in 12 of 13 MZ twin pairs differing in birthweight, the twin with the lower birthweight had a more severe tic disorder than the other. Moreover, the size of the difference in birthweight strongly predicted the difference in the intrapair tic score difference in late childhood.

These findings suggest that crucial environmental events occurring during fetal development subsequently affect the clinical expression of TS. Possible factors *in utero* include such elements as the quality and location of the placenta, the position of the fetus within the uterus and (in the case of twins) intrauterine crowding. Abnormalities in fetal and placental blood vessels can produce decreased blood flow to one twin, resulting in differing degrees of oxygen and nutrient delivery to the developing brain. The basal ganglia are a group of structures deep within the brain. Abnormalities in the basal ganglia may be the cause of TS as well as most other movement disorders. Interestingly, the basal ganglia are especially vulnerable to injury induced by a lack of oxygen and nutrients.

Performance on neuropsychologic tests (nonmotor measures of cerebral function) also illustrates gene–environment interactions in studies of MZ twins with tic disorders. Tic severity and neuropsychologic performance tend to covary. Overall performance on a battery of neuropsychological tests was found to be significantly worse in the MZ twin with the more severe tics. These differences are most notable on tests of attention, visuospatial perception and motor function. This pattern is consistent with the neuropsychologic profile of TS observed in singletons as well as in MZ twins. Interestingly, the overall level of performance on neuropsychological tests in patients with TS is usually within normal limits. Nevertheless, the data on MZ twins illustrate that nongenetic factors influence not only tic severity, but also cognitive function. The behavioral associations and cognitive profile of TS have aroused suspicion that neural systems beyond the basal ganglia are abnormal in individuals with TS. Nevertheless, because of the pivotal role of the basal ganglia in integrating and relaying information from the cortex, it is not necessary to invoke widespread cerebral dysfunction in TS.

PATHOLOGICAL BASIS OF TS

The neuropathologic basis of TS is unknown. Traditionally in clinical neurology, involuntary movement disorders have been associated with basal ganglia dysfunction. Moreover, tics often occur in well-defined basal ganglia disorders such as Huntington disease and postencephalitic parkinsonism (encephalitis lethargica). Medications that stimulate the dopamine system, such as amphetamines, can aggravate tic disorders. Finally, medications that block the dopamine type 2 receptor are effective tic suppressants, and this receptor is most densely expressed in the basal ganglia. Taken together, these observations have led most investigators to focus on the basal ganglia as a likely site of pathology in TS. However, relatively few brains have been donated for research, so comparatively little postmortem analysis has been done. Classical neuropathologic analyses have been published on only two brains, with conflicting findings of basal ganglia (striatal) pathology. A small number of cases revealed possible abnormalities in dopamine and endogenous opioid neurotransmitter systems within the basal ganglia. As more brains are collected from individuals with well-characterized TS, more investigations will be undertaken.

Neuroimaging Studies

In vivo neuroimaging studies have provided some clues about the brain regions involved in TS.

Focusing on the basal ganglia, such studies have found evidence of subtle structural abnormalities in this region. Three quantitative analyses of magnetic resonance imaging scans have been conducted on separate patient populations with TS. The first, in a cohort of singleton TS patients, demonstrated reductions in regional basal ganglia volumes in these patients. The second, also in single TS patients, was more equivocal. In a study of 10 sets of MZ twins, a statistically significant 7% reduction in right caudate nucleus volume was observed in the more severely affected twins compared with the less severely affected ones. More consistent deviations in normal cerebral asymmetries of ventricles and basal ganglia have been reported in studies of both single patients and twins.

Positron emission tomography (PET) and single photon emission computed tomography (SPECT) imaging, which offer insight into regional neurochemistry and cerebral function, have provided evidence of changes in activity of basal ganglia and prefrontal cortex, but the data are not consistent. These imaging techniques can be used to define regional cerebral blood flow and hence the activity level of brain structures. Such studies have revealed a widely distributed network of abnormalities, primarily in the frontal and cingulate regions. *In vivo* neurochemistry data can be derived from SPECT and PET studies as well. One group has found abnormalities in TS patients in dopamine-mediated neurotransmission in the caudate nucleus. Other studies also have found abnormalities in basal ganglia dopamine-mediated neurotransmission. Additional studies are needed to more clearly define the neurochemical basis of TS and related disorders.

TREATMENT

Clinical management depends upon the identification of target symptoms. While often the most obvious and most dramatic manifestation of pathology in TS, tics may not be the most disabling symptom. Concurrent OCD, ADHD, impulse control disorders and depression may be more problematic than the tics. Tics respond most consistently to low doses of dopamine type 2 receptor antagonists such as haloperidol or risperidone. Clonidine and guanfacine, which are agonists at the α_2-adrenergic receptor, offer some benefit, although they are less effective than dopamine antagonists. Some patients respond to benzodiazepines. Particularly as they get older, many patients learn to voluntarily suppress their tics without medication. Family counselling and psychotherapy may help with adjustment problems associated with the social stigma common in more severe cases. In addition, the physician may have to interact with the school system to explain the nature of the tic disorder, and occasionally to clarify the special needs of such children, such as untimed testing. Psychiatric comorbidity can be more disabling than the tics themselves. For those individuals, pharmacotherapy, such as the use of serotonin reuptake inhibitors (e.g. fluoxetine) for OCD or depression, can be an important intervention. For most patients with TS or tic disorders, the clinical course is benign and the outcome is good.

CONCLUSION

Tourette Syndrome is a disorder in which genetic and environmental factors interact to produce a complex neuropsychiatric syndrome. The clinical presentation and course of TS, like many disorders with a primary genetic basis, are influenced by extragenetic factors. Defining the genetic basis of this disorder, however, may lead to more definitive treatments and more focused pharmacological therapies.

Further Reading

Bornstein RA (1990) Neuropsychological performance in children with Tourette's syndrome. *Psychiatry Research* **33**: 73–81.

Bruun RD (1984) Gilles de la Tourette's syndrome: an overview of clinical experience. *Journal of the American Academy of Child Psychiatry* **23**: 126–133.

Bruun RD and Budman CL (1992) The natural history of Tourette syndrome. In: Chase TN, Friedhoff AJ and Cohen DJ (eds) *Advances in Neurology*, pp. 1–6. New York: Raven Press.

Caine ED, McBride MC, Chiverton P *et al*. (1988) Tourette's syndrome in Monroe County school children. *Neurology* **38**: 472–475.

Comings DE and Comings BG (1985) Tourette's syndrome: clinical and psychological aspects of 250 cases. *American Journal of Human Genetics* **37**: 435–450.

Jankovic J (1992) Diagnosis and classification of tics and Tourette syndrome. In: Chase TN, Friedhoff AJ and Cohen DJ (eds) *Advances in Neurology*, pp. 7–14. New York: Raven Press.

Kidd KK, Prusoff BA and Cohen DJ (1980) Familial pattern of Gilles de la Tourette Syndrome. *Archives of General Psychiatry* **37**: 1336–1339.

Pauls DL and Leckman JF (1986) The inheritance of Gilles de la Tourette's syndrome and associated behaviors: evidence for autosomal dominant transmission. *New England Journal of Medicine* **315**: 993–997.

Shapiro AK and Shapiro E (1982) Tourette syndrome: clinical aspects, treatment and etiology. *Seminars in Neurology* **2**: 373–385.

Tourette Syndrome Classification Study Group (1993) Definitions and classifications of tic disorders. *Archives of Neurology* **50**: 1013–1016.

Visual Imagery, Neural Basis of

Intermediate article

Sharon L Thompson-Schill, University of Pennsylvania, Philadelphia, Pennsylvania, USA

CONTENTS

Visual imagery is the process of seeing with the 'mind's eye'. The neural bases of visual imagery have been investigated using physiological recordings of normal brain activity and behavioral measurements in people with brain damage.

INTRODUCTION

Try to answer each of these questions: Do bears have rounded ears? Is a bath towel square in shape? Are roses smaller than a fist? Is a saw's blade narrower near the handle? As you think about the answer to each question, you may feel as if you are seeing each object in your 'mind's eye'. Cognitive psychologists call this phenomenon 'mental imagery'. Although one could speak of imagery in any modality – how does sandpaper feel? how does a ringing telephone sound? – most of the research into the neural bases of mental imagery has focused on the visual modality. Thus, the term 'imagery' will be used here to refer specifically to visual imagery. In addition, this discussion is limited to studies of explicit image generation; related topics such as illusory contour perception are discussed elsewhere. (*See* **Vision: Occlusion, Illusory Contours and 'Filling-in'**)

SIMILARITY BETWEEN VISUAL IMAGERY AND VISUAL PERCEPTION

How similar are the processes of thinking about the shape of a bear's ears and actually looking at a bear's ears? In other words, is visual imagery more closely related to verbal thought, or to perception? This question has driven research in cognitive neuroscience for decades. On one side of the imagery debate are those who believe that mental images have a spatial format and share representations with those used during perception (e.g., Kosslyn, 1980); on the other side are those who maintain that mental images are propositional or symbolic, like language, and therefore do not share representations with perception (e.g., Pylyshyn, 1981). Despite all the behavioral experiments that have been conducted to address this question, there is no unequivocal conclusion. For example, one could point to data illustrating a relationship between the amount of time it takes a person to scan a mental image and the amount of time it takes to scan an actual percept; however, these data have been attributed to the person's response to the experimenter's expectations. Thus, after a quarter-century of active research by psychologists on this topic, the debate is still far from settled.

An alternative way to investigate the extent to which imagery and perception share common representations or processes is to turn to the brain. An enormous amount is known about the neural substrates of perception and the structures of the brain that are responsible for visual perception. Considerably less is known about the neural substrates of mental imagery, because investigators are limited to neuroscientific methods that can be safely used with humans who can be instructed to form mental images. However, there are some physiological data from normal individuals and behavioral data from people with brain damage that provide a useful basis for the comparison between perception and imagery.

Physiological Responses to Visual Imagery

Event-related potentials (ERPs) have been used to address the relationship between visual imagery and visual perception in several ways. One approach is to try to isolate the evoked response to mental imagery and examine its topography, in relation to what is known about visual processing areas. In one such study, imagery and nonimagery conditions of a task were compared in order to yield a relatively pure measure of the evoked response to mental imagery (e.g., Farah *et al.*, 1989): the responses to mental imagery were located

over areas of cortex associated with vision. Furthermore, the magnitude of the ERP during imagery was related to the degree of vividness that the person reported in the images. A second approach is to examine the effects of imagery on early perceptual processing. Mental imagery, performed concurrently with perception, alters an early visual evoked response, termed the N1 component, that is observed 200 ms after a visual stimulus is presented (Farah *et al*., 1988). This response is thought to originate from extrastriate cortex (i.e., secondary visual processing areas). These findings indicate that mental imagery involves – at least in part – neural structures crucial for early visual perception.

The spatial resolution of ERP recordings is limited, so with the advent of new metabolic and hemodynamic imaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), the ability to examine the similarity between the neural structures engaged by visual perception and those engaged by mental imagery has substantially improved. A number of neuroimaging studies have confirmed that mental imagery activates visual processing areas. For example, imagining a walk through a familiar neighborhood, imagining block letters, imagining flashing patterns, and imagining objects all activate regions of visual cortex. Furthermore, images of different sizes produce different patterns of activation, consistent with what is known about the retinotopic mapping of visual cortex (Kosslyn *et al*., 1993). Researchers disagree as to whether the activation is confined to higher-order visual association cortex or whether it includes primary visual cortex. However, there is almost unanimous agreement that visual imagery activates some retinotopically organized cortical regions, which supports the hypothesis that mental images have a spatial format and that imagery and perception have common representations. (*See* **Pattern Vision, Neural Basis of**)

Effects of Brain Damage on Mental Imagery

The strongest test of the assertion that imagery and perception have a shared neural substrate comes from the examination of people who, as a result of brain damage, have impairments in one or both of these processes. If mental imagery requires the same representations and neural structures as does perception, then it follows that a person who has selective impairments in visual perception should also exhibit impairments in visual imagery. Parallel deficits in perception and imagery have been reported in many cases. For example, people with impairments in color vision are unable to report the color of common objects from memory, and people with impairments in object recognition (visual agnosia) are unable to describe the appearance of objects from memory. An agnosic patient studied by Farah *et al*. (1988) was unable to report which animals had long or short tails, what color common objects were, and which US states were similar in shape.

People with damage to small regions of primary visual cortex do not have a complete loss of vision; rather, some portions of their visual field are normal and other portions are impaired. For example, a right occipital lobe lesion will produce a left visual field cut called a hemianopia. In such a person visual processing can still occur, but it takes place in a more restricted field of view. If visual imagery shares representations with visual perception, one might expect to see the effects of a restricted field of view in the 'mind's eye' in people with hemianopia. Farah and colleagues investigated this in a patient scheduled for a right occipital lobe resection as a treatment for her epilepsy (Farah *et al*., 1992). The patient was instructed to imagine walking towards a familiar object and to report how far away she would have to be from the object to 'see' it in its entirety. For example, prior to the operation she might have said that she could be 5 m away from a horse before the image overflowed the visual field of her mind's eye. After the surgery, her estimation would have changed to 10 m – that is, after her occipital resection, the visual field of her mind's eye was restricted just as was her visual field in perception.

DIFFERENCES BETWEEN VISUAL IMAGERY AND VISUAL PERCEPTION

Despite the evidence that visual imagery and visual perception share representations, there are data that point to differences between these processes: in other words, imagery and perception are not exactly the same thing. As briefly mentioned above, there is some disagreement as to whether mental imagery activates primary visual cortex in neuroimaging studies. The earliest neuroimaging study of mental imagery used a 'mental walk' paradigm, and reported activation in visual regions of the parietal and temporal lobes but not the occipital lobe. This result could illustrate a difference between the neural substrates of imagery and perception. However, many subsequent studies have reported early visual cortex activation. Kosslyn *et al*. (2000) suggested that activation of the earliest

visual areas might be observed only in studies that required the generation of high-resolution images. The neuropsychological literature also contains data that suggest differences between imagery and perception. For example, patients with occipital lesions have been described who are unable to perform some, but not all, imagery tasks. Likewise, some agnosic patients with clear impairments in perception of objects demonstrate normal imagery of objects. The implications of these dissociations are discussed by Farah (2000).

MULTIPLE TYPES OF IMAGERY

Another question in study of mental imagery is whether there is a difference between visual imagery and spatial imagery. Spatial representations can be based on visual information but can also be based on information derived from other modalities, such as touch. In contrast, some visual representations, such as color, can only be derived from the visual modality. Many of the tasks used to study mental imagery do not distinguish between visual representations and spatial representations. For example, one of the classic visual scanning tasks requires the participant to image a capital letter F, and then to scan around the letter and report the location of each corner as it is encountered. This visual task clearly requires spatial knowledge, and could be performed without any visual information (e.g., a blind person could use tactile imagery to traverse the letter). Kosslyn *et al*. (2000) have proposed three distinct types of imagery: (1) spatial imagery, which may not require visual representations at all; (2) figural imagery, which may require only low-resolution visual representations, such as those found in the temporal lobe which subserve object recognition; and (3) depictive imagery, which would require the highest-resolution images and which most directly corresponds to the notion of 'seeing with the mind's eye'. Some evidence for these distinctions comes from the neuroimaging literature: imagining a mental walk results in parietal activation, thinking about familiar objects results in inferior temporal activation, and comparing the width of stripes in memory results in medial occipital activation.

Some people with a disorder of spatial processing called 'hemispatial neglect' may have impairments in spatial imagery that parallel their perceptual deficits. These patients show an inability to attend to objects in the left side of space as a result of a right parietal lesion. A group of Milanese patients with hemispatial neglect were asked to imagine the well-known Piazza del Duomo in Milan (Bisiach and Luzzatti, 1978) – first, they were asked to imagine that they standing at the south end of the square, facing north towards the cathedral. When asked to report all the landmarks in the piazza, the patients tended to name landmarks on the east (to the patient's imagined right) side of the square and omit landmarks on the west (to the patient's imagined left) side. Later, the patients were asked to imagine that they were standing at the north end of the square, with their backs to the cathedral. Now, the patients tended to name landmarks on the west (to the patient's imagined right) side and omit landmarks on the east (to the patient's imagined left) side. In other words, they observed hemispatial neglect of imagined scenes.

The strongest evidence for multiple forms of imagery is the double dissociation between spatial and figural imagery impairments. In perception, there is considerable evidence for a dissociation between perception of object location and perception of object identity (i.e., the ventral 'what' pathway and the dorsal 'where' pathway). A parallel dissociation has also been observed during mental imagery (Levine *et al*., 1985). One patient with a dorsal lesion was unable to localize visual stimuli in space but could identify objects; this is the classic 'where' pathway deficit. This patient was able to describe the appearance of familiar objects from memory but was unable to describe the locations of familiar landmarks from memory. In contrast, a second patient with a ventral ('what' pathway) lesion was unable to identify objects or to describe the appearance of familiar objects from memory; this patient was unimpaired on spatial localization, from both perception and memory. In other words, these two patients demonstrate a double dissociation between imagery of spatial location and imagery of object identity that parallels their selective deficits in perception.

References

Bisiach E and Luzzatti C (1978) Unilateral neglect of representational space. *Cortex* **14**: 129–133.

Farah MJ (2000) The neural bases of mental imagery. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, pp. 965–976. Cambridge, MA: MIT Press.

Farah MJ, Hammond KL, Levine DN and Calvanio R (1988a). Visual and spatial mental imagery: dissociable systems of representation. *Cognitive Psychology* **20**: 439–462.

Farah MJ, Peronnet F, Gonon MA and Giard MH (1988b). Electrophysiological evidence for a shared representational medium for visual images and

percepts. *Journal of Experimental Psychology* **117**: 248–257.
Farah MJ, Peronnet F, Weisberg LL and Monheit MA (1989) Brain activity underlying mental imagery: event-related potentials during image generation. *Journal of Cognitive Neuroscience* **1**: 302–316.
Farah MJ, Soso MJ and Dasheiff RM (1992) The visual angle of the mind's eye before and after unilateral occipital lobectomy. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 241–246.
Kosslyn SM (1980) *Image and Mind*. Cambridge, MA: Harvard University Press.
Kosslyn SM, Alpert NM, Thompson WL *et al*. (1993) Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience* **5**: 263–287.
Levine DN, Warach J and Farah MJ (1985) Two visual systems in mental imagery: dissociation of 'what' and 'where' in imagery disorders due to bilateral posterior cerebral lesions. *Neurology* **35**: 1010–1018.
Pylyshyn ZW (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **88**: 16–45.

Further Reading

Farah MJ (2000) *The Cognitive Neuroscience of Vision*, pp. 252–289. Oxford, UK: Blackwell.
Kosslyn SM and Thompson WL (2000) Shared mechanisms in visual imagery and visual perception: insights from cognitive neuroscience. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, pp. 975–986. Cambridge, MA: MIT Press.

Wernicke–Geschwind model

Intermediate article

Andrew Kertesz, St Joseph's Hospital, University of Western Ontario, London, Ontario, Canada

CONTENTS

The Wernicke–Geschwind model is the basis for an aphasia classification system that is favored by many clinicians. It is also a model for language organization of the brain.

INTRODUCTION

Carl Wernicke described *der aphasische Symptomenkomplex* in a series of case reports emphasizing the disturbance of comprehension of language (Wernicke, 1874). These cases ranged from sensory aphasia – what we now call Wernicke aphasia – to global aphasia, where both comprehension and expression were severely disturbed. Wernicke's classification was solidly based on the input–output dichotomy and the division of the brain into sensory and motor systems. It was Wernicke's mentor, Meynert, professor of neurology in Vienna, who advanced this dichotomy based on anatomical and lesion studies, at that time representing a revolutionary advance in our knowledge of the brain. Wernicke extended these observations to language. Although Wernicke's original diagrams contained centers for sensory images of the words connected to motor images, he is very clear in his description of the results of disconnection between these centers. He predicted conduction aphasia would result from sensorimotor disconnection, even though he did not describe such a case himself.

Many of Wernicke's ideas were systematically summarized by Lichtheim (1885), based on the associationist theoretical framework of the nineteenth century. The Wernicke–Lichtheim model distinguished between subcortical motor, motor, subcortical sensory, sensory, conduction, transcortical motor and transcortical sensory aphasias (Figure 1). This diagram, dubbed 'Lichtheim's house', contains the essential pathways and their disruptions, each of which could result in a specific type of aphasia. The schema became accepted and used by many clinicians. Aphasia is not only caused by a focal lesion in an area considered the center for a particular function, but can also result from the disconnection between the centers. Conduction aphasia remains one of the prime examples of the disconnection syndromes, although it is still debated vigorously whether or not it is the insular cortex lesion that produces the deficits, or the disconnection of the temporal lobe from the frontal lobe, known as the arcuate fasciculus lesion.

Eighty years later, Norman Geschwind described the disconnection syndromes and resurrected the Wernicke–Lichtheim model of aphasias (Geschwind, 1965). This has become the basis of a number of current clinical classifications. Geschwind added a considerable degree of sophistication to the model, based on more modern physiological and anatomical knowledge. He also revived some of the syndromes outside the oral language system, specifically the one that was originally described by Dejerine involving the dissociation of visual recognition of words and the writing system: this results in alexia without agraphia, one of the best-known examples of disconnection syndrome.

THE WERNICKE–GESCHWIND MODEL AS IT STANDS TODAY

Sound stimulus is the basis of much animal and human communication. The human brain is particularly sensitive to human speech frequencies; much of this acoustic discrimination occurs in the temporal operculum, or lip of the temporal lobe in the depth of the sylvian fissure separating the temporal from the parietal lobes. Although elementary sound discrimination is processed tonotopically in Heschl's gyrus, the discrimination of speech sounds is elaborated in a much wider association area bilaterally. Geschwind believed that the planum temporale or the area behind Heschl's gyrus played a particular role in language. He used anatomical studies, cutting the brain in the plane of extension of the sylvian fissure, to show that the planum is normally larger on the left side, in the language-dominant hemisphere. The auditory

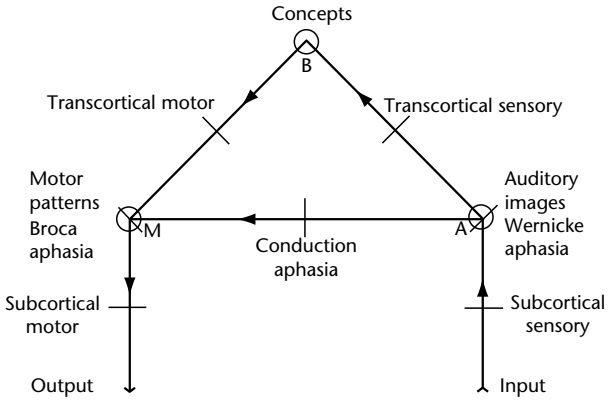


Figure 1. Lichtheim's house. (A = Auditory center, M = Motor center, B = Begriffen)

association area occupies the superior temporal gyrus on the outside, as well as the planum temporale in the inside of the lip of the sylvian fissure, and is distinguished cytoarchitectonically. The columnar organization of this neocortex resembles raindrops on a windowpane. Wernicke himself thought that auditory word images resided in the neurons of this area, now called Wernicke's area. The exact area is debated, although most people agree that both the outside posterior third of the superior temporal gyrus and the planum temporale are essential for the processing of verbal input, and lesions of this area if sufficiently large result in serious and persisting difficulties in comprehension. Wernicke's area also includes the inferior parietal lobule surrounding the end of the sylvian fissure.

Wernicke thought that if subcortical neural elements were damaged then impulses entering this area would be cut off, resulting in a subcortical sensory aphasia, better known as 'pure word deafness'. He reasoned that if the auditory association area itself were damaged then the result would be not only a comprehension deficit, as seen in pure word deafness, but also a disturbed output, since the auditory verbal images are not available to generate appropriate output. Wernicke also anticipated the idea of audiology feedback or monitoring the output system from the auditory area, long before computer systems were applied to theoretical modeling. His concept was that the auditory image area would perceive 'inner speech' before it was vocalized. During this process appropriate corrections could be made, and if damage was present, speech errors would occur.

Destruction of auditory engrams could result in several kinds of deficit and a variety of abnormal output. When a word is substituted, verbal paraphasia, when a phonemic error occurs phonological paraphasia results. When a word was completely replaced by erroneous phonology, unintelligible sequences or jargon would manifest. A combination of these deficits, generally a significant comprehension deficit with phonological and semantic paraphasias, is called 'Wernicke aphasia'.

Once auditory images are appropriately processed and selected, afferent or incoming phonological components would be associated with (or mapped onto) motor images of phonology. Those in Wernicke's time were called motor engrams, and in more recent terminology 'logogens'. Motor engrams would presumably be localized in or around Broca's area, since destruction of the posterior frontal inferior gyrus or the foot of F3, also called Broca's area, often results in output disturbances with well-preserved comprehension, because the auditory system is still operational. The phonological output system (the phonological buffer in the computer analogy) would be impaired, bringing about phonological articulatory errors. Impaired articulation, repetitive stuttering utterances, and distinctive phonological paraphasias affecting initial consonants, or literal paraphasias where only one phoneme or letter would be exchanged, would predominate. In some cases the motor errors are so prominent that Liepmann, one of Wernicke's students, suggested the term 'apraxia of speech', which was later added to the model. Others used the term 'verbal apraxia', which has become popular, especially in North America.

Wernicke and Lichtheim thought if the disturbance was peripheral or distal in the flow of processing motor images, then subcortical motor aphasia or pure motor aphasia would result with articular disturbance as the predominant feature, but relatively preserved comprehension, grammar, and writing, also known as 'pure word dumbness' or aphemia. This latter term has been resurrected in many contexts, and at times it is separated from aphasic disturbances as anathria, a pure motor phenomenon.

Broca aphasia is characterized by effortful speech output, a great deal of word-finding difficulty, and in some languages agrammatism with lack of grammatical words or small words. The idea of agrammatism and paragrammatism was first discussed by Pick, and its modern formulation was developed by Goodglass and his many students; it was not part of the original Wernicke–Lichtheim model. Syntax is much studied in modern aphasiology. It is closely associated with phonology experimentally and physiologically and it can be

dissociated from semantic and articulatory processing in clinical syndromes.

Wernicke predicted there would be cases where auditory images and motor engrams would be preserved, but connections between them would be disrupted, resulting in failed repetition and faulty paraphasic speech. *Leitungsaphasie* was translated as 'conduction aphasia', and cases conforming to this were described subsequently by Kleist and by Goldstein (1948). These patients have relatively well-preserved comprehension and fairly fluent output, which is at times disturbed by phonological paraphasias and at times by repetitive phonemic approximations of the target (*conduit d'approche*). The hallmark of conduction aphasia is the inability to repeat because of the disconnection of auditory images and motor mechanisms. Spontaneous speech therefore is still produced, but the otherwise automatic process of repetition cannot be carried out because of disconnection. A repetition deficit in conduction aphasia has also been viewed as a disturbance in short-term verbal memory or a disorganized execution of a phonological encoding program related to the auditory feedback, going beyond the simple idea of disconnection.

The concept of transcortical aphasias also originated from the Wernicke–Lichtheim model. Transcortical motor aphasia is the disruption of the pathway between word concepts and motor engrams; it is characterized by poor output and poor initiation of speech, but excellent comprehension and, above all, preserved repetition – in other words, the automatic transfer of language from auditory images is undisturbed, but the interrupted transcortical transfer from concepts and nonverbal association prevents the patient from initiating output. Transcortical sensory aphasia, on the other hand, represents a disconnection between sensory input and concepts, and shares preserved repetition with the other transcortical pattern. These patients can repeat, but they do not understand because the sensory images are not connected with their meaningful representation. Mixed transcortical aphasia results from both sensory and motor engrams being disconnected from transcortical associations and the only language function that remains is repetition without comprehension or spontaneous speech, a condition also called 'isolation of the speech area'.

Clinicians have devised systems of examination using the Wernicke–Geschwind model to allow the scoring of auditory comprehension of words, sentences, and grammar, and to assess output mechanisms through the scoring of spontaneous speech, repetition and naming. Repetition is useful, for instance to test the connection of the input and output systems, and it is essential for diagnosing conduction aphasia where the two systems are disconnected. It is also important to demonstrate the preservation of the central core system of auditory input and verbal output in case of transcortical aphasias where transcortical associations are impaired. Without testing repetition, the model cannot be validated. The model also allows the explanation of lexical access and disturbances of naming and the integration of auditory and visual images to account for naming on visual stimulation. Subsequent interpretation of the model suggests that temporal auditory associations and their integration with tactile and visual mechanism are important for lexical storage, selection and access.

The emphasis placed on comprehension by Wernicke was adapted by many subsequent investigators, even those who were detractors of the model. Most aphasiologists consider impaired comprehension the *sine qua non* of aphasic disturbance, while others claim that lexical access disturbance may be present even in cases where comprehension is not impaired. Most agree both are necessary to explain aphasic phenomena, and disturbance of either occurs to a certain degree in clinical aphasia. An account for both processes is necessary to explain normal language processing.

Modality-specific language disorders were also part of the original Wernicke–Geschwind model and were considered to be input and output problems. However, this was further developed by looking at the linguistic aspects of the phenomena. In France, 'aphemia' or 'pure motor aphasia' became known as 'phonetic disintegration disturbance'. Another group of investigators emphasized the complex motor aspects of speech, conceptualized as apraxia by Liepmann. Verbal apraxia is used for many of the motor and articulatory deficits, and older terms for this, such as anathria and aphemia, are resurrected from time to time. Subcortical sensory aphasia was further elaborated as pure word deafness to auditory agnosia for nonverbal sounds, and both were also called cortical deafness. Other subdivisions included sentence deafness, auditory disturbance for syntax, and word meaning deafness, which overlaps semantic aphasia or the loss of meaning, representing a two-way disturbance of comprehension and naming of words. These elaborations on the model have been succeeded by linguistic models, fractionated at a different theoretical level. Modality-specific anomia is exemplified by the specific difficulty of naming visually while other modalities such as tactile naming were preserved, and was called

'optic aphasia'. Other recognized examples are difficulty in naming fingers and discriminating right from left in the Gerstmann syndrome, associated with agraphia.

The Wernicke–Geschwind model provides the first account of the nature of internal representations of concepts. Wernicke's postulation of auditory verbal images does not explain the phenomena of mental imagery, the superior memory performance for pictures rather than for words. However, Wernicke's concept center was not explicitly stated to be restricted to words and was not meant to be located to any particular part of the brain, even though Wernicke's diagram placed it somewhere in the parietal area of the right hemisphere.

Many classifications of aphasia try to fit one model or another. A purely modality-oriented model fails to account for all the clinical phenomena. Adding the transcortical portion of the model represented a significant contribution and provided an adequate explanation for the majority of the cases. When an aphasia test is used that covers input and output modalities as well as repetition, most cases can be classified according to the Wernicke–Geschwind model. Although many clinicians are satisfied with impressions of unmeasured performance and aphasic profiles are constructed on the basis of such impressions, efforts have been made to support the model with classification based on test scores. Numerical taxonomy and discriminant function statistics have been used in attempts to achieve objective classification (Kertesz, 1979).

More recently category-specific naming deficits, showing dissociation between categories of items

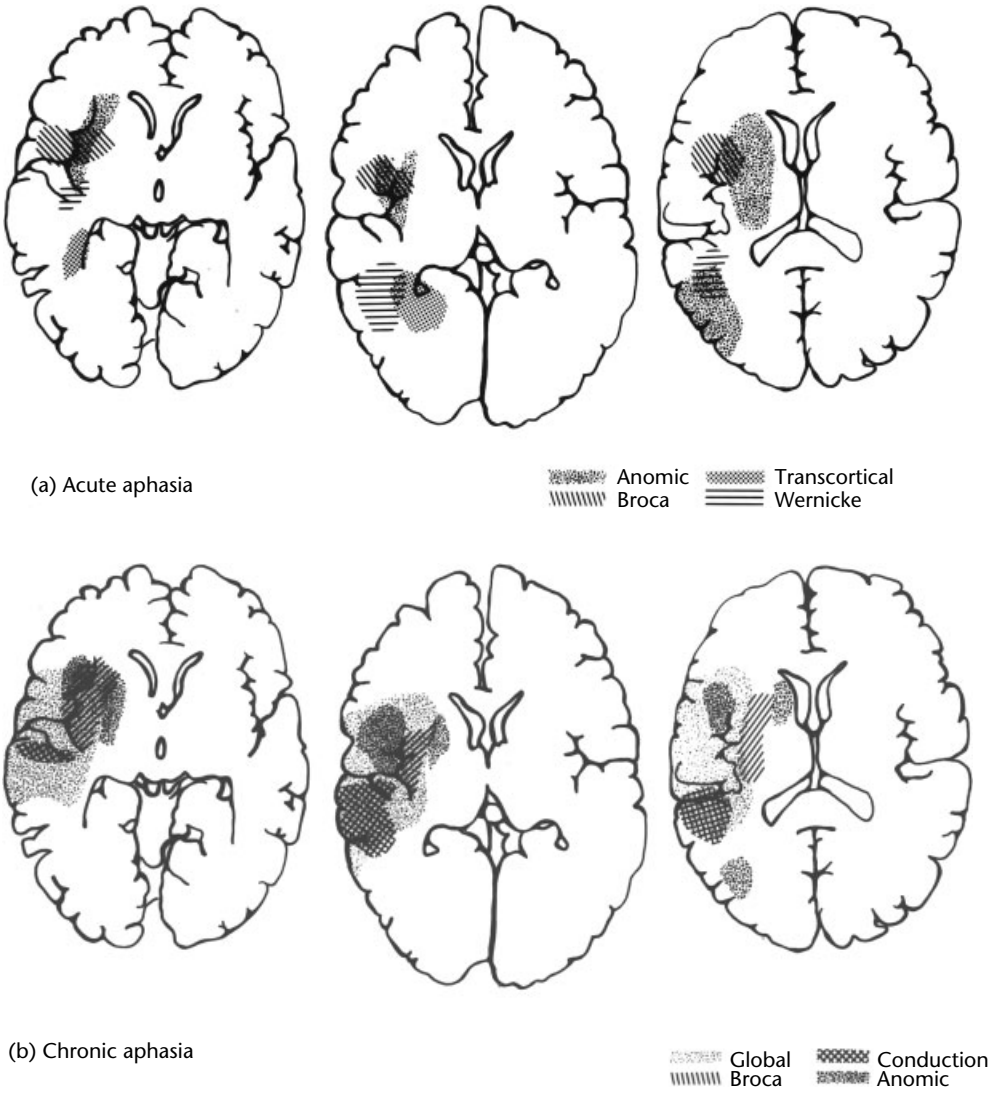


Figure 2. Brain changes following stroke. (a) Acute aphasia; (b) Chronic aphasia.

such as fruits versus vegetables or animate versus inanimate, have been described. None of these can be fitted into the Wernicke–Geschwind model and represent modern developments in aphasiology and linguistics. They tend to occur in temporal lesions secondary to herpes encephalitis or progressive semantic aphasia with Pick disease. The Wernicke–Geschwind model accounts best for discrete stroke and neoplastic lesions affecting nodes of processing or the connectivity between them.

The durability and usefulness of the Wernicke–Geschwind model is related to its reproducible anatomical basis. The model originated from anatomical observations of lesions in patients in combination with the theoretical consideration of the function of the structures involved. It has been confirmed by successive generations of localization techniques beginning with a compilation of autopsy records associated with clinical observations that continues to this day to be an important contribution. Recovery, reorganization and compensation complicates the relationship of the deficit to the structural change, and limits the conclusions about brain function. Changing or unstable lesions, such as growing tumors with all too brief cross-sectional observations, are even more challenging. In stroke lesions scanning images of the brain over a period show different overlaps corresponding to changes in the syndrome as the patient recovers (Figure 2) (Kertesz, 1979).

Functional localization has produced new pieces of information that were not predicted by the Wernicke–Geschwind model, although these are few in number compared with those that were confirmatory. One of these is the persistent appearance of activation in the dominant prefrontal region when semantic associations are made in normal individuals. Although this area is known to be part of the executive system including short-term memory and integration of external and internal stimuli, even planning and judgment, it was connected with language only in an indirect fashion. However, functional language studies persistently show activation of this area. The prefrontal area is active in many experimental paradigms, representing the executive function required for instance in the effort of producing verbs in associations.

When methods other than subtracting elementary auditory from language processes are used, it is evident that the structures involved in the Wernicke–Geschwind model are also activated. The superior temporal gyrus, the posterior frontal region, inferior parietal lobule and the supplementary motor area, which are involved in the lesions producing aphasia, are also activated in various experiments with functional activation of language (Petersen *et al*., 1988; Demonet *et al*., 1992).

References

Demonet JF, Chollet F, Ramsay S *et al*. (1992) The anatomy of phonological and semantic processing in normal subjects. *Brain* **115**: 1753–1768.

Geschwind N (1965) Disconnections syndromes in animals and man. *Brain* **88**: 237–294.

Goldstein K (1948) *Language and Language Disturbances*. New York, NY: Grune & Stratton.

Kertesz A (1979) *Aphasia and Associated Disorders: Taxonomy, Localization and Recovery*. New York, NY: Grune & Stratton.

Lichtheim L (1885) On aphasia. *Brain* **7**: 443.

Petersen SE, Fox PT, Posner MI, Mintun M and Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* **331**: 585–589.

Wernicke C (1874) *Der Aphasische Symptomenkomplex*. Breslav: Cohn & Weigart.

Further Reading

Eggert GH (1977) *Wernicke's Works on Aphasia*. The Hague, Netherlands: Mouton.

Geschwind N (1974) *Selected Papers on Language and The Brain*. Boston, MA: D. Reidel.

Price CJ (2000) The functional anatomy of language. Contributions from neuroimaging. *Journal of Anatomy* **197**: 335–359.

What and Where/How Systems

Intermediate article

Hendrik Christiaan Dijkerman, University of Utrecht, Utrecht, The Netherlands

CONTENTS

What and where/how systems are the two main processing streams along which visual information is processed in the primate cerebral cortex.

INTRODUCTION

A large part of the primate cerebral cortex is involved in processing visual input. Indeed, within the monkey brain, more than 30 visual cortical areas have been identified so far. The work of Ungerleider and Mishkin has been highly influential in establishing the neuroanatomical structure of the visual areas and their putative function. In their seminal paper 'Two cortical visual systems', they argue that, after reaching the primary visual cortex (V1), input is further processed along two different routes (Ungerleider and Mishkin, 1982). One route is situated more dorsally and terminates in the posterior parietal lobe, whereas the other projects more ventrally to the inferotemporal cortex (Figure 1). We now know that the ventral stream proceeds from V1 through the more lateral and anterior visual cortical areas V2, V3 and V4 into the posterior part of the inferotemporal cortex. The dorsal stream also involves V2 and V3, which project to the middle temporal area, V3A and the parieto-occipital area, terminating in several areas including the medial superior temporal area, ventral intraparietal area, lateral intraparietal area, 7a, and the anterior intraparietal area. The anatomical separation in ventral and dorsal streams has been confirmed, although there may be more interconnections between the two streams than originally was thought.

EXPERIMENTAL EVIDENCE

Functional Characteristics of the Visual Ventral Stream

Many of the early studies on the functional characteristics of the two visual streams were performed with monkeys. Early lesion studies suggested impairments in visual discrimination performance after bilateral inferotemporal lesions (Gross, 1973). The visual abilities affected included not only hue and brightness, but also two-dimensional pattern and three-dimensional shape discrimination. In addition, single-cell recordings in monkeys showed that neurons in the ventral extrastriate cortex responded to intrinsic object properties such as shape, color, texture and orientation. Input into the ventral stream comes almost exclusively from V1, as evidenced by the lack of responsiveness in ventral neurons after ablation of this area. Evidence for the involvement of temporal areas in visual perception and object recognition in humans came originally from studies of people with brain damage. Impairments in the recognition of objects, in the absence of elementary visual deficits, are usually associated with temporal lobe lesions. More recent functional imaging studies have provided further evidence that the human inferotemporal cortex is involved in visual perception and recognition.

Functional Characteristics of the Visual Dorsal Stream

The posterior parietal lobe has been linked with many different functions, only two of which are discussed here: visuospatial localization, and visual control of movements. Important evidence for the involvement of the dorsal stream in visuospatial function again comes from monkey ablation studies. After lesions to the parietal lobe, monkeys were impaired in their ability to choose which of two food wells was closer to a visual landmark (Pohl, 1973). Other ablation studies found that monkeys had problems on a variety of other visual spatial tasks including the stylus maze task, cage finding and route following (Milner and Dijkerman, 1998). Furthermore, humans with right parietal lesions often show impairments on a

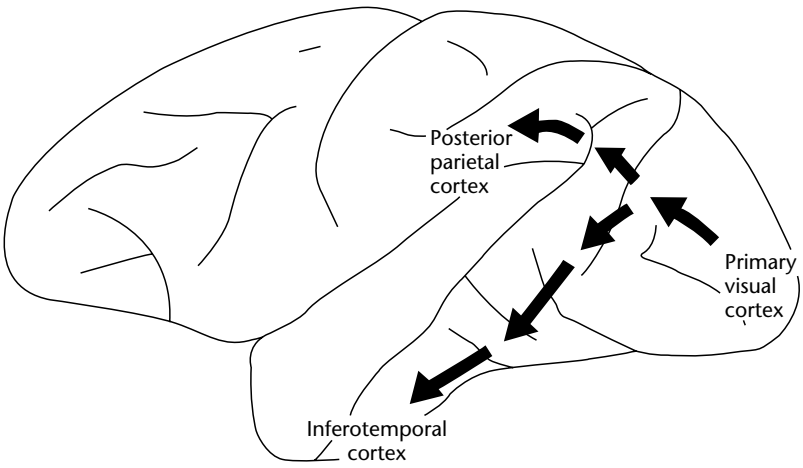


Figure 1. The visual dorsal and ventral streams in the monkey cortex. Reproduced from Milner and Goodale (1995) *The Visual Brain in Action* by permission of Oxford University Press.

variety of spatial tasks including matching of line orientation, estimating distance and discrimination of position (Von Cramon and Kerkhoff, 1993). However, visuospatial deficits do not exclusively follow parietal lesions. Temporal lesions in humans also result in impairments on several spatial tasks, suggesting that at least some spatial processes require an intact ventral stream (Maguire *et al*., 1996). These tasks included distance judgments, route knowledge and map reading, suggesting possible ventral stream involvement in environment-relative coding. Functional imaging studies also indicate activation of the parietal lobe during visuospatial processing. Activation of the posterior parietal lobe has been found in a range of tasks including spatial working memory, spatial long-term memory and spatial matching.

It is well established that lesions of the parietal lobe result in impairments in the visual guidance of movements in monkeys (Ettlinger and Kalsbeck, 1962). The impairment often affected only one arm, ruling out a purely visual deficit, and the monkey was generally able to put food with the affected arm into its mouth, excluding a purely motor impairment. In humans, a similar visuomotor impairment as a result of parietal lesions was first described by Balint (1909). Although the deficits of Balint's patient were assumed to reflect primarily a spatial disorder, it is now clear that nonspatial aspects of visuomotor behavior, such as adjusting the hand opening to the size of the object to be grasped, can also be impaired after parietal lesions. Indeed, there are probably several separate visuomotor channels within the dorsal stream. For example, single-cell recordings show that the responsiveness of neurons in the monkey posterior parietal lobe to visual stimuli may depend on the type of motor response required (Mountcastle *et al*., 1975). Thus, some neurons may only be active when a saccade occurs, while others respond only during the grasp of a certain object. Functional imaging studies also indicate a separation between the posterior parietal areas involved in visually guided saccades and reaching movements (Kawashima *et al*., 1996).

DISCUSSION

Ungerleider and Mishkin proposed not only an anatomical subdivision but also a functional separation between the two processing streams. They suggested that the ventral stream was concerned with visual processing of stimulus characteristics important for object identification (what a stimulus is), whereas the parietal lobe subserved spatial vision (where a stimulus is). This idea has been influential not only within the neuroscience of vision research, but also with respect to philosophical, cognitive, and computational models of the organization of the visual system.

Modifications to this proposal suggest a division between perceptual identification 'what' and visual guidance of goal-directed action 'how' (Milner and Goodale, 1995). Rather than focusing on the visual input characteristics, this view emphasizes the task requirements (what the visual information is processed for). Support for this modification comes from two sources: first, single-cell recordings showed that certain neurons in the monkey parietal cortex responded to intrinsic object characteristics such as shape of the stimulus and/or a grasping movement towards it, and not to its spatial location

(Sakata *et al*., 1995); second, neuropsychological studies of an agnosic patient, D. F., found that while she was unable to recognize the size or orientation of an object she could adjust her visuomotor responses towards this object on the basis of these stimulus characteristics. Although her brain damage was diffuse, it most severely affected the lateral prestriate areas, resulting in a selective disconnection between V1 and the inferotemporal cortex (seen on functional imaging).

A related model was proposed by Jeannerod (1997), who argued for separate object representations within the brain. Activation of a particular representation would depend on the task requirements. Two main representations were distinguished: the pragmatic representation contains parameters relevant for action and could be used to generate the corresponding motor response; in contrast, a semantic mode operates for identification purposes, during which object attributes are bound together. A consequence of this distinction is that object attributes should not be classified according to their putative anatomical channel, but to the requirement of the object-oriented behavior. Jeannerod emphasized that both representations can be used to guide different types of goal-directed action.

Although there is general agreement on the object recognition role for the ventral stream, the visual functions attributed to the dorsal stream differ. There may be several possible explanations for this apparent disagreement. First, many motor responses are based on visuospatial information, while many visuospatial tasks can assumed to require more oculomotor activity than does perceptual identification of foveated objects (Milner and Goodale, 1995). Thus, the requirements for visuomotor and visuospatial tasks may overlap considerably. However, there are some visuomotor tasks for which impairments are found after posterior parietal lesions which may not have a spatial component (e.g. grasping). Similarly, not all visuospatial impairments can be accounted for by deficits in visually guided movements. A second possible explanation for the discrepancy in the hypothetical dorsal stream function is that visuomotor processing and visuospatial representations may depend on different structures within the parietal lobe. The primate posterior parietal lobe can be divided into superior and inferior regions. Evidence from patient studies suggests that the superior part of the posterior parietal lobe is involved in immediate goal-directed visuomotor action, whereas the inferior parietal cortex is involved in spatial cognition. Thus, the common site of lesions in people with optic ataxia is superior to the area typically associated with visuospatial neglect (Perenin, 1997). The inferior parietal lobe regions that are commonly associated with neglect seem to be involved in a supramodal representation of space and receive input predominantly from the ventral stream.

INTERRELATIONS BETWEEN THE TWO VISUAL CORTICAL STREAMS

Whether a distinction is made between 'what' and 'where', or between 'what' and 'how', it is clear that the two visual streams need to work closely together to produce effective and coherent behavior in everyday life. Neuroanatomical studies in monkeys show that there are considerable interconnections between the two systems (Felleman and Van Essen, 1991). The superior temporal sulcus may be especially important in linking the two processing streams. Furthermore, the pattern of ventral and dorsal stream projections to the frontal lobe suggest that the final motor output area, the primary motor cortex (M1), can be influenced by either stream (Rossetti and Pisella, 2002), although its connections with the parietal lobe are more direct.

At a behavioral level, several interactions between the two visual streams are possible, depending on the functions ascribed to them. One important dorsal–ventral interaction is binding the object's spatial position with its identity. Evidence from neurological patients suggests that the medial temporal lobe is involved in retrieval of object–location associations (Milner *et al*., 1997), whereas the parietal lobe may have a role in encoding these associations. The interrelation between the ventral and dorsal streams is also important for interactions between conscious perceptual (semantic) representations and visuomotor (pragmatic) action. An example of such interaction comes from studies of people with selective lesions to the dorsal stream. Such people have been found to compensate for their impaired visual guidance of movements by relying on stored semantic or perceptual information processed within the intact ventral stream (Jeannerod, 1997; Milner *et al*., 2001). The opposite pattern has been observed in patient D. F., whose ability to use visual information for the guidance of arm and eye movements deteriorated when asked to use memorized perceptual target information (Milner and Goodale, 1995).

RELATION TO CONSCIOUSNESS

Studies with both healthy and neurologically impaired human participants indicate that

considerable neural processing of visual input occurs without it reaching awareness. Within the framework of the two visual systems, it has been suggested that only ventral stream processing could lead to conscious visual experience (Milner and Goodale, 1995). Indeed, perturbation studies in normal participants and evidence from neurological patients suggest that the visual guidance of action can occur without conscious awareness of the relevant stimulus features. However, ventral stream processing alone may not be sufficient for visual awareness. The signal additionally needs to be amplified by the kind of neuronal gating associated with selective attention mechanisms (Milner and Goodale, 1995). Neuroimaging studies confirm that although activity in the ventral visual cortex is necessary, it is not sufficient for conscious perceptual experience. Experiments on binocular rivalry and bistable figures (such as the Necker cube) in both healthy and neurologically impaired people indicate that superior parietal areas, in addition to frontal areas, make important contributions (Rees, 2001). These parietal areas are similar to those involved in directing spatial attention.

References

Balint R (1909) Seelenlähmung des 'Schauens', optische Ataxie, räumliche Störung der Aufmerksamheit. *Monatsschrift für Psychiatrie und Neurologie* **25**: 51–81.

Ettlinger G and Kalsbeck J (1962) Changes in tactile discrimination and in visual reaching after successive and simultaneous bilateral posterior parietal ablations in the monkey. *Journal of Neurology, Neurosurgery and Psychiatry* **25**: 256–268.

Felleman DJ and Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**: 1–47.

Gross CG (1973) Visual functions of inferotemporal cortex. In: Jung R (ed.) *Handbook of Sensory Physiology*, vol. VII part 3, pp. 451–482. Berlin: Springer.

Jeannerod M (1997) *The Cognitive Neuroscience of Action*. Oxford: Blackwell.

Kawashima R, Naitoh E, Matsumura M *et al*. (1996) Topographic representation in human intraparietal sulcus of reaching and saccade. *NeuroReport* **7**: 1253–1256.

Maguire EA, Burke T, Phillips J and Staunton H (1996) Topographical disorientation following unilateral temporal lobe lesions in humans. *Neuropsychologia* **34**: 993–1001.

Milner AD and Dijkerman HC (1998) Visual processing in the primate parietal lobe. In: Milner AD (ed.) *Comparative Neuropsychology*, pp. 70–94. Oxford, UK: Oxford University Press.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford: Oxford University Press.

Milner B, Johnsrude I and Crane J (1997) Right medial temporal lobe contributions to object-location memory. *Philosophical Transactions of the Royal Society of London Series B* **352**: 1469–1474.

Milner AD, Dijkerman HC, Pisella L *et al*. (2001) Grasping the past: delay can improve visuomotor performance. *Current Biology* **11**: 1896–1901.

Mountcastle VB, Lynch JC, Georgopoulos A, Sakata H and Acuna C (1975) Posterior parietal association cortex of the monkey: command functions for operations within extra personal space. *Journal of Neurophysiology* **38**: 871–908.

Perenin MT (1997) Optic ataxia and unilateral neglect: clinical evidence for dissociable spatial functions in posterior parietal cortex. In: Thier P and Karnath HO (eds) *Parietal Lobe Contributions to Orientation in 3D Space*, pp. 289–308. Berlin: Springer.

Pohl W (1973) Dissociations of spatial discrimination deficits following frontal and parietal lesions in monkeys. *Journal of Comparative and Physiological Psychology* **82**: 227–239.

Rees G (2001) Neuroimaging of visual awareness in patients and normal subjects. *Current Opinion in Neurobiology* **11**: 150–156.

Rossetti Y and Pisella L (2002) Several 'vision for action' systems. A guide to dissociating and integrating dorsal and ventral stream functions. In: Prinz W and Hommel B (eds) *Common Mechanisms in Perception and Action. Attention and Performance*, vol. XIX, pp. 62–119. Oxford, UK: Oxford University Press.

Sakata H, Taira M, Murata A and Mine S (1995) Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex* **5**: 429–438.

Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA and Mansfield RJW (eds) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Von Cramon DY and Kerkhoff G (1993) On the cerebral organisation of elementary visuospatial perception. In: Gulyás B, Ottoson D and Roland PE (eds) *Functional Organisation of the Human Visual Cortex*, pp. 211–231. Oxford, UK: Pergamon Press.

Further Reading

Creem S and Proffitt D (2001) Defining the cortical visual systems: 'What', 'Where', and 'How'. *Acta Psychologica* **107**: 43–68.

Goodale MA, Milner AD, Jakobson LS and Carey DP (1991) A neurological dissociation between perceiving objects and grasping them. *Nature* **349**: 154–156.

Goodale MA, Jakobson LS and Keillor JM (1994) Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia* **32**: 1159–1178.

Jeannerod M and Rossetti Y (1993) Visuomotor coordination as a dissociable visual function: experimental and clinical evidence. In: Kennard C (ed.) *Visual Perceptual Defects*, pp. 439–460. London, UK: Baillière Tindall.

Jeannerod M, Decety J and Michel F (1994) Impairment of grasping movements following posterior parietal lesion. *Neuropsychologia* **32**: 369–380.

Newcombe F and Russell WR (1969) Dissociated visual perceptual and spatial deficits in focal lesions of the right hemisphere. *Journal of Neurology Neurosurgery and Psychiatry* **32**: 73–81.

Ungerleider LG and Haxby J (1994) 'What' and 'where' in the human brain. *Current Opinion in Neurobiology* **4**: 157–165.

Working Memory, Neural Basis of

Introductory article

John Jonides, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

Working memory is the memory system used by humans and other animals to store small amounts of information for brief periods in the service of reasoning, thinking, problem-solving and the like.

INTRODUCTION

Consider this: you are driving, and stop to ask a passer-by for directions to a store. She tells you to proceed three blocks, make a left at the drugstore, make a right at the first light, go one mile, then find the store on the right. You do not write down the directions but are able to negotiate the route and find the store with little difficulty. Clearly, this task requires what has been called 'working memory', which has several important characteristics, all of which are illustrated by this example.

One is that it cannot store a great deal of information. Had the directions been much lengthier, the chances are that you would have forgotten a step. Another characteristic is that it remains active for only a short period, even if there is no interference. If you were distracted by a conversation with a passenger, you might have forgotten part of the directions, even if they were not complicated. A third characteristic is that the information stored in working memory is easy to retrieve. Finally, working memory is not a passive storage system; you can manipulate the contents of working memory in the service of some task. For example, you might transform the information from the verbal directions of the helpful passer-by to a visual image of a route map and then trace your way mentally along the map to reach your destination. Limited capacity, limited duration, easy accessibility, and provision for manipulating the contents are the hallmarks of working memory in humans.

Working memory is fundamental to many cognitive processes. It is used routinely in complicated problem-solving or reasoning tasks because any sort of mental problem requires information to be stored and manipulated. Also, it is claimed that working memory is essential to the comprehension of spoken language; storing the beginning of a sentence in working memory permits you to interpret the words at the end of that sentence properly. In fact, working memory ability correlates with overall performance on many measures of intelligence, showing that it is a fundamental component of cognition. For these reasons, it is important to understand the psychological and neurological mechanisms of working memory in humans.

A BRIEF HISTORY OF WORKING MEMORY

Working memory is not a new concept in psychological research. Over a century ago William James included a cogent discussion of working memory (which he called 'primary memory') in his classic work, *Principles of Psychology*. Like working memory, James's primary memory had a small capacity, short duration, and immediacy, in the sense of conscious experience. In fact, the alliance of working memory with consciousness has a long history in the study of this topic, a history which James helped to instigate.

Beyond James's introspective reflections, substantial behavioral research has helped to characterize working memory, largely carried out in the second half of the last century. Much of this research concentrated on the short duration and limited capacity of the memory system (then referred to as 'short-term' memory, a concept that focuses more on its storage characteristics and less on its usefulness in conducting the work of cognition; hence the change to the current term, 'working' memory). Based on some of this research, the capacity of working memory is often cited as 7 ± 2 items, where the definition of an

item can vary: it might be a word in a list of words, a letter in a list of letters, and so on. Also, many experiments were devoted to investigating the causes of forgetting from working memory: do new items interfere, or do memory representations simply decay with time? While the answer to this is still unresolved, there appears to be good reason to suspect that both decay and interference play a substantial role. Another issue that has been of concern is whether working memory can be differentiated from long-term memory behaviorally, as James differentiated it from what he called 'secondary memory'. Indeed, there are several reasons to suppose that it can be so differentiated. Of interest also has been research on how material is retrieved from working memory. Initial results suggested that items of information were retrieved one at a time, but more recent data suggest that there may be simultaneous access to several items.

The earlier concept of short-term memory gradually evolved into the concept of working memory because of the inclusion of the important idea captured by the term 'working'. Working memory includes a short-term storage system, one that is used in the service of other cognitive tasks. So, working memory studies have gone beyond the duration of the memory trace and the limited capacity of storage, to focus on how stored information is used by higher cognitive functions (such as comprehending directions, as illustrated above).

COMPONENTS OF WORKING MEMORY

Working memory is not singular in character; rather, it appears to have several components. Perhaps the most influential theory of the componential nature of working memory is due to Alan Baddeley and his colleagues. Baddeley has proposed a model that includes two fundamentally different components. One is a set of storage buffers that are responsible for holding a limited amount of information for a brief period. What differentiates the buffers is the kind of information they store. Baddeley's original conception of them included two types of information: verbal and visuospatial. We now know that the types of information that can be differentiated in working memory are more numerous than this, including working memory for objects that cannot easily be named; there may be other types as well. We also know that the different types of working memory storage are not simply a function of the modality of input. For example, working memory for verbal information appears to make use of the same neural machinery whether the words are presented by ear or by eye.

Because information in the storage buffers decays with time and interfering information, there is a need for a rehearsal mechanism that refreshes the contents of the buffers to keep the contents active. Rehearsal is conceptualized as a kind of inner speech in the case of verbal information, and as an inner spotlight of attention that moves among mental locations for spatial information. Studies of rehearsal for other types of information are lacking.

The second component of Baddeley's model of working memory has to do with the manipulation of information stored in the buffers. Baddeley called this the 'central executive' component, but it is perhaps more apt to refer to a set of 'executive processes'. Discussions of executive processes have focused on three sorts. There are attentional processes that allow one to focus on some particular information in the buffers. There are also processes that inhibit attention from irrelevant information, so that one can concentrate on what is relevant. The third type are processes that allow one to switch successfully between two tasks being performed at the same time (e.g., driving and participating in a conversation with a passenger). This may not exhaust the list of executive processes, but it gives a flavor of what the functions of executive processes are.

DIFFERENT VIEWS OF WORKING MEMORY

Baddeley's view of executive and storage components of working memory has been the most productive in stimulating research on the topic. However, it is not the only view available. One influential alternative comes from the work of Nelson Cowan. According to this view, there are three sorts of memory representations that need to be considered.

The largest and most comprehensive is long-term memory, the enduring repository of information that a person might store. At any moment, most of this information resides in a passive state, ready for retrieval. A small portion of the information, however, can be in an activated state because it has been the recent focus of attention, because it has been recently presented, because it is novel, or for other reasons. Within the temporarily activated information, furthermore, a subset of it might be the focus of attention, and this would correspond to the information that might be in consciousness at any time. All information that

resides in long-term memory is potentially available for activation. The information that is currently activated is akin to what others have called working memory in that it is limited in amount, it is readily accessible, its active state will wane over time or as other information is activated, and it is available to participate in other cognitive processes. Within the body of activated information, a small subset will be the subject of consciousness by virtue of its having been attended to.

Perhaps an intermediate position between that of Baddeley and that of Cowan is reflected in the research of Randall Engle and his colleagues. Like Cowan, they see working memory as the temporarily activated portion of long-term memory. However, they emphasize the importance of mental procedures such as rehearsal that are specific to working memory. In addition, like Baddeley and Cowan, they emphasize the importance of selective attention processes that control the contents of working memory. Unlike Cowan but like Baddeley, Engle and his colleagues also recognize the distinction among different codes of information, such as phonological or spatial. Indeed, the neuroanatomical data on this issue, reviewed below, make this distinction quite persuasive. What Engle emphasizes in his view more than the others is the importance of individual differences in attentional control over the contents of working memory. Such differences lead to differences in fluid intelligence which is important in reasoning and problem-solving. Furthermore, these individual differences are also reflected in how effectively people can maintain information in working memory in the face of interference and distraction. Indeed, research by Lynn Hasher, Rose Zacks and their colleagues has pinpointed the importance of attentional control in excluding unwanted information from working memory while allowing the person to concentrate on information of interest.

While there is no current resolution among the different views of working memory, it is worth highlighting some important features that have been contributed by each.

One is the idea that the contents of working memory are temporarily activated representations in long-term memory. Another is that the information in working memory needs to be refreshed and attended to in order to remain there, and this is controlled by executive processes that turn attention to relevant information and turn it away from irrelevant information. Yet another is that different sorts of information may have different representations in working memory. Add to this that there are now well-documented differences among individuals in working memory, and that these differences have important implications for intelligence and reasoning, and we see that the working memory system, however conceived, has a central role in cognition.

NEUROANATOMICAL BASIS OF WORKING MEMORY

The study of cognition has been greatly aided by studies of brain anatomy and function in both humans and other animals. This is true of working memory as it is of many other cognitive functions. One important set of findings about working memory has to do with which parts of the brain mediate the various components of working memory and what networks of brain areas are involved in working memory tasks. Knowledge about the localization of processes in the brain has permitted us to predict the nature of working memory impairment that a person would suffer with damage to one or another of these brain structures. Another important contribution of this research has been to test various psychological theories of working memory. For example, the view that working memory is not unitary in character, but is composed of subsystems responsible for different types of information, has been confirmed by testing patients with damage to different regions of the brain and by scanning normal individuals who are engaged in working memory tasks with different materials.

There are three important sources of information about the neuroanatomical basis of working memory. One comes from the study of animals who are challenged with working memory tasks when they are given selective lesions or when the activity of individual neurons is recorded. A second comes from the study of patients with localized lesions who offer the opportunity to study what aspects of working memory are compromised with damage to a specific region of the brain. The third comes from the recent use of neuroimaging tools, especially positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), to study normal individuals performing working memory tasks.

Studies of Working Memory in Animals

Beginning with the seminal work of Jacobsen in the 1930s, there has been a succession of important studies of working memory in animals, most notably monkeys. Many of these studies have used various versions of the delayed response

task, in which a stimulus is presented to an animal and must be retained in the animal's memory for a short interval, typically several seconds; the animal is then given choices among stimuli and must produce a response that indicates it has retained a memory of the original stimulus.

A spatial version of the delayed response task has been popular for studying working memory in animals, especially in recent studies by Goldman-Rakic and her colleagues. In this task an animal is trained to fixate a central location on a screen while a brief stimulus is illuminated in the periphery. The animal maintains fixation during the delay interval, and afterwards it is given a cue to shift its gaze to the location that it was storing in memory. Many experiments have shown that lesions of the dorsolateral prefrontal cortex produce a marked deficit of performance on this task. Interestingly, if the animal's task is to maintain not spatial information but information about an object's shape, the crucial lesion that produces a deficit is in a region just inferior to the one that is implicated in spatial working memory, suggesting a difference in working memory representations depending on the type of information stored.

Studies of animals performing a delayed response task while neural responses are recorded from individual cells in dorsolateral prefrontal cortex confirm the studies using lesions in animals: many of these cells are responsive to specific stimuli (e.g., specific spatial locations or specific object shapes) during the delay interval of the delayed response task. It therefore appears certain that even individual cells of the dorsolateral prefrontal cortex have working memory capability. Other studies suggest that the superior parietal cortex, certain subcortical structures, and the hippocampal region in monkeys are involved in the delayed response task as well. There is persuasive evidence of a network of highly interconnected posterior and anterior structures that mediate working memory.

Disorders of Working Memory Following Brain Lesions

Studies of patients with brain damage have also offered important insights about working memory. One of the most salient findings is a clear dissociation between spatial and verbal working memory based on the side of the brain that is damaged. This is supported by numerous studies of patients with deficits in verbal working memory, and by selected cases of patients with deficits in spatial working memory. In one review of patients with verbal working memory deficits, 17 of the 18 patients who were described had lesions of the left hemisphere. These lesions included inferior parts of the parietal lobe. Patients with lesions in this region, for example, have very poor digit spans, a standard test of the storage component of working memory. One well-studied case of this sort is patient K. F., described by Warrington and Shallice, who had a head injury affecting primarily the left parietal-occipital cortex, in the region of the inferior parietal lobe. This patient had great difficulty with repetition tasks, and his digit span was reduced to two. Similar cases confirm that posterior, left-hemisphere lesions seem to be selective for verbal as opposed to spatial working memory.

In contrast, the few cases of deficits in spatial working memory that have been reported suggest that it is the right hemisphere that is involved, not the left. The clearest case of this is patient E. L. D, described by Hanley and colleagues, a right-handed woman who had suffered an aneurysm in the right hemisphere which damaged tissue around the sylvian fissure. When tested on the Corsi Blocks Test, a spatial analog to digit span, she scored noticeably worse than normal; yet her performance on verbal working memory tasks was comparable with that of normal controls. The contrast between patient E. L. D. and patients with left hemisphere lesions suggests that different circuitries are responsible for verbal and spatial working memory.

Studies of patients with brain lesions have revealed yet more about the nature of verbal and spatial working memory systems. For example, the model of verbal working memory proposed by Baddeley and colleagues includes a storage buffer to retain material and a rehearsal system to recycle the material in the storage buffer so that it does not decay quickly. The rehearsal subcomponent has been likened to the activation of the speech system, albeit silently. Yet studies of patients with working memory deficits have often revealed that they have normal or nearly normal speech patterns. So, if the speech system is indeed involved in rehearsal, it must be an aspect of that system that is tied to internal recycling of information and not to overt speech.

The storage of spatial information in working memory has been thought to use the same neural machinery as that for the generation and manipulation of visual images. This idea has been tested by examining patients such as M. G., who had clear deficits on tests of mental imagery that required her to rotate or scan an imagined object, yet her spatial working memory was largely intact. Beyond this, researchers have found that asking people to

engage in a secondary task such as spatial tapping or arm movement while completing a spatial working memory task hinders performance on the memory task, but has little if any effect on visual imagery tasks. It thus appears that spatial working memory and mental imagery involve somewhat different mental machinery.

Studies of patients with brain lesions have furthered our understanding of the neural basis of working memory. First, they have uncovered the regions of the human brain that are critical to working memory. Second, they have allowed us to compare the neuroanatomy of human working memory with working memory in other animals to reveal the importance of certain regions such as the prefrontal cortex, as reviewed by D'Esposito and Postle. However, these studies are compromised in at least two ways. First, naturally occurring lesions that produce working memory deficits are often large, making it difficult to localize precisely the damage that disrupts processing. Second, given the brain's capacity for plasticity and compensation in the face of damage, the study of patients with brain injury may not accurately reveal neural processes in normal individuals.

Functional Neuroimaging Studies of Working Memory

Neuroimaging techniques are an important source of information about the neuroanatomical basis of working memory and are free from the weaknesses of behavioral studies with patients. The use of PET and fMRI especially has provided important insights into circuitry and in so doing has tested some of the psychological claims about working memory.

It would be disappointing if the findings from neuroimaging studies of working memory were in conflict with those from the study of patients. Indeed, there is consistency. A prominent result that confirms this has to do with the circuitry that is activated by different types of materials.

For verbal material, activations in neuroimaging studies have been found predominantly in the inferior frontal gyrus, premotor cortex, supplementary motor cortex, superior parietal cortex, and inferior parietal cortex, most prominently in the left hemisphere. The sites in the inferior frontal gyrus and premotor cortex are similar to ones involved in the production of speech, and evidence suggests that these are involved in rehearsal of verbal material. The site in the inferior parietal cortex is consistent with studies of patients with deficits in verbal working memory who most often have damage there. The site in the superior parietal cortex is probably involved in controlling attention to different items of verbal material, and is consistent with the frequent identification of this site in studies of attentional allocation in other contexts. Some verbal working memory studies have also found activation in the dorsolateral prefrontal cortex, frequently in the right hemisphere, and this may be a result of the operation of executive processes on material that is stored when some manipulation of that material is required. For example, one test that often shows dorsolateral prefrontal activation is the 'two back' task. In this task, the participant sees a succession of letters presented one at a time, once every 3 s. The task is to decide whether each letter matches the one presented two letters back in the sequence. This clearly requires storage of verbal information but also requires continual reconfiguring of what is stored, so the information that is retained is not static. It may be this additional requirement that recruits dorsolateral structures.

Studies of spatial working memory have revealed a quite different neural architecture. Spatial tasks recruit structures in extrastriate occipital cortex, superior and inferior parietal cortex, premotor cortex, supplementary motor cortex, and sometimes inferior frontal gyrus, all largely in the right hemisphere. It has been suggested that the superior parietal site together with the superior frontal sites are involved in a rehearsal circuit for spatial material that is stored in regions of the inferior parietal and occipital cortex. While it is not clear what the homology is between humans and monkeys for different regions of the brain, the fact that there is activation of frontal and parietal sites in human studies is at least qualitatively consistent with studies of spatial working memory in monkeys. Some spatial tasks have also resulted in activation of the dorsolateral prefrontal cortex, as with verbal tasks, and this activation may once again be an indication of processes involved in the executive manipulation of spatial material.

One might think that the important distinction between verbal and spatial circuitry has to do with auditory versus visual material. This, however, does not appear to be so. First, the circuitry for verbal material appears to be the same whether that material is delivered by eye or by ear. Second, the circuitry for visual working memory when the stimuli are shapes is quite different from that for spatial material: it appears to recruit structures of temporal and frontal cortex, possibly largely in the left hemisphere. In this way, the spatial and shape working memory circuitries seem to honor the

distinction between a visual processing pathway specialized for spatial information (involving dorsal structures) and a pathway specialized for information about shape and form (involving ventral structures).

How do these findings about neural circuitry inform us about psychological theories? One issue to which they are certainly germane is whether there are different working memories for different types of material. The answer appears to be that there are, at least for spatial, verbal, and shape materials. Another issue has to do with whether working memory is the same as long-term memory, with the difference merely being levels of activation. In that working memory tasks seem to activate some associative cortex that may be the seat of long-term memory, there appears to be some currency to this idea. However, in that working memory tasks also recruit structures that are not routinely involved in long-term memory tasks (e.g. the superior parietal cortex), the evidence suggests that there may be processes specific to working memory *per se*.

Yet another issue has to do with executive processes. At present there is still too little information about these processes to be sure what brain structures they recruit; yet evidence emerging from studies of working memory is leading to the view that there are prefrontal and superior parietal structures involved in allocating attention, in inhibiting irrelevant material, in manipulating information stored in working memory, and in keeping track of multiple goals in working memory when the individual is engaged in multiple tasks.

The archive of data from neuroimaging studies of working memory is still small, but is already revealing a network of structures which mediate a complex function essential to complex cognitive life.

CONCLUSION

Much has been learned about working memory that goes beyond James's important introspections. We know about its capacity, duration, and function from behavioral experiments; we know what happens to it with damage to selected parts of the brain; and we know what regions of the brain are activated when humans and other animals are engaged in working memory tasks. These three sources of evidence, together with other techniques such as studies of receptor binding for neurotransmitters, promise to reveal a detailed picture of the architecture of working memory in the next epoch of research on the topic.

Further Reading

Baddeley AD (1986) *Working Memory*. New York, NY: Oxford University Press.

Cowan N (1988) Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin* **104**: 163–191.

D'Esposito M, Postle BR and Rypma B (2000) Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Experimental Brain Research* **133**: 3–11.

Jonides J (1995) Working memory and thinking. In: Smith EE and Osherson DN (eds) *Invitation to Cognitive Science: Thinking*, vol. 3, 2nd edn, pp. 215–265. Cambridge, MA: MIT Press.

Jonides J and Smith EE (1997) The architecture of working memory. In: Rugg MD (ed.) *Cognitive Neuroscience*, pp. 243–276. Hove, UK: Psychology Press.

Jonides J, Reuter-Lorenz P, Smith EE *et al*. (1996) Verbal and spatial working memory. In: Medin D (ed.) *The Psychology of Learning and Motivation*, pp. 43–88. New York, NY: Academic Press.

Logie RH (1995) *Visuo-spatial Working Memory*. Hove, UK: Lawrence Erlbaum.

Miyake A and Shah P (eds) (1999) *Models of Working Memory*. Cambridge, UK: Cambridge University Press.

Shallice T (1988) *From Neuropsychology to Mental Structure*. Cambridge, UK: Cambridge University Press.

Smith EE and Jonides J (1999) Storage and executive processes in the frontal lobes. *Science* **283**: 1657–1661.

Action, Philosophical Issues about

Intermediate article

Alfred R Mele, Florida State University, Tallahassee, Florida, USA

CONTENTS

The philosophy of action concerns theories about what actions are, how they are to be explained, and the mental events and states associated with intentional action.

INTRODUCTION

In striving to analyze, understand, and explain actions, philosophers of action are concerned primarily with intentional actions. In discussions of freedom of action, intentional action also naturally occupies center stage. And although people are morally accountable for some unintentional actions, as in cases of negligence, moral assessment of actions is focused primarily on intentional actions. What is the nature of intentional action and how are intentional actions to be explained?

CENTRAL PHILOSOPHICAL ISSUES ABOUT ACTION

There are two main philosophical questions about actions: What is an action? How are actions to be explained? The first question directly raises two others: How are actions different from nonactions? (For example, how does an ordinary instance of running hard differ from an ordinary unintentional instance of breathing hard?) How do actions differ from one another? (For example, if I turn on my computer by pressing a button, are my turning it on and my pressing the button different actions or the same action?) If not all actions are intentional, the first question also raises another: What is it for an action to be intentional? The question about the explanation of actions is also a question for cognitive science. The challenge is to produce a theory of the springs of action in light of which we can, in principle, explain particular intentional actions – in light of which we can explain, for example, why you are reading this article, why I wrote it, and why the editor decided to publish it. If proper explanations of actions are causal explanations, part of what we would like to understand is how the events or states that explain actions help to produce actions.

PHILOSOPHICAL VIEWS AND THEORIES OF ACTION

According to a popular answer to the question how actions differ from nonactions (Brand, 1984; Davidson, 1980; Mele, 1992), actions are like sunburns in an important respect. The burn on Al's back is a sunburn partly in virtue of its having been caused by exposure to the sun's rays; a burn that looks and feels just the same is not a sunburn if it was caused by a heat lamp. Similarly, a certain event is Al's raising his left hand – an action – partly in virtue of its having been appropriately caused by mental items. An influential version of this view claims that reasons, understood as combinations of beliefs and desires, are causes of actions and that an event counts as an action partly in virtue of its having been suitably caused by a reason (Davidson, 1980). Alternative conceptions of action include an 'internalist' position according to which actions differ experientially from other events in a way that does not depend on how, or whether, they are caused; a conception of actions as composites of nonactional mental events or states (e.g. intentions) and pertinent nonactional effects (e.g. an arm's rising); and views identifying an action with the causing of a suitable nonactional product by an appropriate nonactional mental event or state – or, instead, by an agent (for discussion of these alternatives, see Davis, 1979; Ginet, 1990).

Promising theories about how actions differ from one another include a fine-grained theory (Goldman, 1970), a coarse-grained theory (Davidson,

1980), and componential theories (Ginet, 1990). According to a fine-grained theory of actions, *A* and *B* are different actions if, in performing them, the agent exemplifies different act-properties. For example, if Ann starts her car by turning a key, her starting the car and her turning the key are two different actions, since the act-properties at issue are distinct. A coarse-grained theory asserts that Ann's turning the key and her starting the car are the same action described in two different ways. A componential theory claims that Ann's starting her car is an action having various components, including her moving her arm, her turning the key, and the car's starting. Where the first two theories claim to find, alternatively, a collection of related actions, or a single action under different descriptions, component theories assert that there is a 'larger' action having 'smaller' actions among its parts.

Most philosophers agree that at least a sketch of an explanation of an intentional action can be provided by identifying the reasons for which the agent performed it. Whether reasons can have a place in *causal* explanations of actions is controversial. In 1963, Donald Davidson challenged anti-causalists about 'reasons-explanations' to provide an account of the reasons for which we act that does not treat (our having) those reasons as causes of relevant actions. Imagine that Ann has a pair of reasons for using her leaf blower this morning. First, she wants to blow the leaves off her lawn today and she regards this morning as a very convenient time. Second, she has a desire to repay her neighbor for awakening her yesterday with his leaf blower and she believes that blowing the leaves off her lawn this morning would do the trick. As it happens, Ann uses her leaf blower this morning only for one of these reasons. In virtue of what is it true that she uses it for this reason, and not for the other, if not that this reason (or her having it), and not the other, makes an appropriate causal contribution to her using it? Detailed attempts to meet this challenge have been revealingly problematic (for discussion, see Mele, 1992, chap. 13).

IMPACT OF COGNITIVE SCIENCE ON ISSUES ABOUT ACTION

Philosophers of action have explored the nature of psychological states and events thought to play important causal/explanatory roles in intentional action, including beliefs, desires, and intentions. Some philosophers appeal to intentions in an effort to avoid the problems 'deviant causal chains' pose for attempted analyses of intentional action featuring reasons as causes (Brand, 1984; Mele, 1992). The alleged problem is that whatever psychological causes are offered as necessary and sufficient for a resultant action's being intentional, scenarios can be constructed in which, owing to an atypical causal connection between the favored psychological antecedents and a pertinent resultant action, that action is not intentional. For example, Ann wants to awaken her husband and she believes that she may do so by making a loud noise. Motivated (causally) by this desire and belief, Ann may search in the dark for a suitable noise-maker. In her search, she may accidentally knock over a lamp, producing a loud crash. By so doing, she may awaken her husband, but her awakening him in this way is not an intentional action. The task for those who wish to analyze intentional action causally is to specify not only the psychological causes of actions associated with their being intentional but also the pertinent roles played by these causes. Some philosophers have sought help from cognitive science in undertaking this task (Brand, 1984; Mele, 1992).

Presumably, intentions play important roles in the production of intentional actions. Functions plausibly attributed to intention in the philosophical and psychological literature include initiating and motivationally sustaining intentional actions, guiding intentional behavior, helping to coordinate agents' behavior over time and their interaction with other agents, and prompting and appropriately terminating practical reasoning (see Brand, 1984; Bratman, 1987, 1999; Mele, 1992). The initiating, sustaining, and guiding roles are relevant to the problem of deviant causal chains.

Intentions, like many psychological states, have both a representational and an attitudinal dimension. The representational content of an intention may be understood as a *plan*. The intending *attitude* towards plans may be termed an *executive* attitude. Plans, on one conception, are purely representational and have no motivational power of their own (Brand, 1984; Mele, 1992). People may have any number of attitudes towards plans, in this sense. They may believe that a plan is elegant, admire it, hope that it is never executed, and so on.

To understand the executive dimension of intention, compare an intention to attend a concert with a *desire* to attend a concert. Both encompass motivation to attend a concert, and the content of each is or includes a representation of the prospective course of action. But although one can have a desire to attend a concert without being at all *settled* on doing so, intending to attend a concert is partially constituted by being settled on so doing. This is compatible with intentions' being

revocable and revisable. Though Al is now settled on meeting a friend for dinner, he would cancel the arrangement were a pressing problem to arise at home.

An important motivational difference between desires and intentions may lie in their *access* to the mechanisms of intentional action. This difference coheres with the claim that intending to *A* entails being settled on *A*-ing while desiring to *A* does not. Whereas our becoming settled on *A*-ing straightaway is normally sufficient to initiate an *A*-ing at once, this is not true of the acquisition of desires to *A* straightaway. To be sure, someone's being settled now on *A*-ing later normally will not initiate an *A*-ing now. But if the intention is still present at the later time and the agent recognizes that the designated time has arrived, an attempt at *A*-ing will normally be immediately forthcoming. This is not true of someone who still has a mere desire at the later time to *A*; such a person may simply choose not to *A* and behave accordingly.

If intentions initiate actions, it is *proximal* intentions that do so – roughly, intentions to do something straightaway. (More precisely, it is the *acquisition* of a proximal intention that plays this role.) But why do proximal intentions initiate and sustain the actions that they do? Why, for example, does an intention to drive home tend to initiate and sustain one's driving home rather than one's cycling home or one's driving to a friend's house? Return to the representational side of intentions. An intention to *A* incorporates a plan for *A*-ing, and *which* intentional action or actions an intention generates is a partial function of the intention-embedded plan. In the limiting case, the plan is a simple representation of one's *A*-ing. Often, intention-embedded plans are more complex. For example, Al's proximal intention to check the oil in his car incorporates a plan that includes his first unlatching the hood, then opening the hood, then unscrewing the oil cap, and so on. An agent who successfully executes an intention is *guided* by the intention-embedded plan.

An intention-embedded plan identifies a goal and (in non-limiting cases) provides action-directions, as it were. Exactly how deep the representational content of intentions runs is a partly empirical and partly conceptual question. Even when what is intended is routine and very simple behavior for the agent – a doctor's signing a prescription, for example – a great deal is going on representationally. Some psychologists take the representational content of motor schemata to run quite deep, suggesting, for example, that motor schemata involved in handwriting include representations of the neuromuscular activity required to achieve the movement represented by their higher-level components. Standard philosophical conceptions of intention seem not to countenance such representations as parts of the representational content of a normal agent's intention to write her name, probably because of the apparent inaccessibility of these representations to consciousness. On standard conceptions, however, intentions guide behavior in a way that depends on their representational content. If plans embedded in standard writing intentions do not incorporate representations of low-level neuromuscular activity, they can provide guidance at a higher level, with the assistance of motor schemata that are external to intentions. Perhaps a solution to some problems posed by deviant causal chains can be produced by careful attention to the guiding function of proximal intentions.

RELEVANCE OF ACTION THEORY TO COGNITIVE SCIENCE

The philosophy of action has produced theories about what actions are and about how actions are to be explained. In both connections, philosophers have speculated about how mental events and states figure in the production of intentional actions. Theories of the former kind are useful to cognitive scientists concerned to explain human actions, insofar as these theories provide conceptions of what is to be explained. Some philosophical theories of the explanation of action are fertile ground for cognitive science. These theories have elements that are testable, or suggestive of empirical hypotheses, and the theories reveal interesting patterns in our own common-sense explanations of why we do what we do.

References

Brand M (1984) *Intending and Acting*. Cambridge, MA: MIT Press.

Bratman M (1987) *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman M (1999) *Faces of Intention*. Cambridge, UK: Cambridge University Press.

Davidson D (1963) Actions, Reasons, and Causes. *Journal of Philosophy* **60**: 685–700.

Davidson D (1980) *Essays on Actions and Events*. Oxford, UK: Oxford University Press.

Davis L (1979) *Theory of Action*. Englewood Cliffs, NJ: Prentice-Hall.

Ginet C (1990) *On Action*. Cambridge, UK: Cambridge University Press.

Goldman A (1970) *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Mele A (1992) *Springs of Action*. New York, NY: Oxford University Press.

Further Reading

Audi R (1993) *Action, Intention, and Reason*. Ithaca, NY: Cornell University Press.
Bishop J (1989) *Natural Agency*. Cambridge, UK: Cambridge University Press.
Hornsby J (1980) *Actions*. London: Routledge & Kegan Paul.
McCann H (1998) *The Works of Agency*. Ithaca, NY: Cornell University Press.
Mele A (2003) *Motivation and Agency*. New York, NY: Oxford University Press.
Searle J (2001) *Rationality in Action*. Cambridge, MA: MIT Press.
Velleman JD (2000) *The Possibility of Practical Reason*. Oxford, UK: Oxford University Press.
Wilson G (1989) *The Intentionality of Human Action*. Stanford, CA: Stanford University Press.

Aesthetics

Intermediate article

Jerrold Levinson, University of Maryland, College Park, Maryland, USA

CONTENTS

Aesthetics is that branch of philosophy devoted to conceptual and theoretical inquiry into art and aesthetic experience.

THE DOMAIN OF AESTHETICS

We may usefully distinguish three conceptions of the domain of aesthetics, according to what is taken as the focus of attention:

- The practice of making and appreciating works of art.
- Aesthetic properties, features, or aspects of things.
- Aesthetic attitudes, perceptions, or experiences.

There are intimate relations among these three conceptions. Thus, art might be conceived as a practice in which people aim to make objects possessing valuable aesthetic properties, or that are apt to give subjects valuable aesthetic experiences; aesthetic properties might be conceived as those properties saliently possessed by works of art, or those on which aesthetic experience is centrally directed; and aesthetic perception might be conceived as the sort of perception that is central to the appreciation either of works of art, or of the aesthetic properties of things, whether natural or man-made. Finally, it can be argued that art, in its creative and receptive aspects, provides the richest and most varied arena for the exploration of aesthetic properties and the enjoyment of aesthetic experiences.

The aesthetics of nature may be included in the second or third of these conceptions, if it is understood as the study of certain distinctive properties of natural phenomena that can be classified as aesthetic (e.g. beauty, sublimity, grandeur), or of certain kinds of experience provoked by nature, or of certain kinds of attitudes to nature. The theory of criticism may be included in the first conception, if it is understood as the study of that part of the practice of art concerned with the reception of artworks, including their description, interpretation, and evaluation. Craft, too, can be understood as an art-related or quasi-artistic activity, and hence may be included in the first conception.

Art

One conception of art sees it as specially concerned with the exploration and contemplation of perceptible form for its own sake. This view has roots in the work of the eighteenth-century German philosopher Immanuel Kant, who thought that the beauty of objects and phenomena, whether natural or man-made, consisted in their ability to stimulate the free play of the cognitive faculties in virtue of their pure forms, both spatial and temporal, and without the mediation of concepts. In the early twentieth century, the English art theorist Clive Bell took a similar line, holding that spatial form was the only artistically relevant aspect of visual art, and that possessing 'significant form' was the necessary and sufficient condition of a work of art.

Another conception of art sees it as essentially a vehicle of expression or communication, especially of states of mind or nonpropositional contents. In the early twentieth century, the Italian philosopher Benedetto Croce claimed that the essence of art is in the expression of emotion. He emphasized the indissociability, even identity, of content and vehicle in art. The English philosopher R. G. Collingwood developed this line further, observing that making works of art was a way for the artist to articulate or make clear the nature of his or her emotional condition. The Russian novelist Leo Tolstoy identified art with emotional communication from one person to another by indirect means, namely, a structure of signs in an external medium.

A third conception of art sees it as concerned with the imitation or representation of the external world, perhaps in distinctive ways or by distinctive means. This conception can be found in the earliest works in the canon of aesthetics, the *Republic* of Plato and the *Poetics* of Aristotle. Modified so as to allow for representation of matters beyond the visible, it finds expression among later thinkers in the aesthetic theories of Lessing, Hegel, and

Schopenhauer. Some modern discussions of art as representation regard it broadly as semiotic or symbolic in nature.

Art has also been conceived as an activity aimed explicitly at the creation of beautiful objects, including representations of natural and human beauty; as an arena for the exhibition of skill, particularly in fashioning or manipulating objects capable of exciting admiration (Sparshott, 1982); as a development of play, stressing the structured and serious aspects of play (Gadamer, 1986); and as the sphere of experience per se, in which attention is drawn to the interplay of active (creative) and passive (receptive) phases in engagement with the external world (Dewey, 1934).

More recently, art has been conceived as the production of objects intended to afford aesthetic experience (Beardsley, 1981); as the investing of objects with 'aboutness' in the context of a specific cultural framework (Danto, 1981); as a particular social institution identified by its constituent rules and roles (Dickie, 1997; Davies, 1991); and as an activity identifiable only historically through a connection to earlier activities or objects whose art status is assumed (Wollheim, 1980; Levinson, 1990, 1996; Carroll, 2001).

Aesthetic Property

It is generally agreed that aesthetic properties are perceptual or observable, experienced in a fairly direct manner, and relevant to the aesthetic value of the objects that possess them. Various further characteristics of aesthetic properties have been proposed, including: having gestalt character; requiring a certain sensitivity for discernment; having an evaluative aspect; affording pleasure or displeasure in contemplation; not being governed by conditions, or applicable by rule; supervenience on lower-level perceptual properties; requiring imagination for attribution; requiring metaphorical thought for attribution; being notably revealed in aesthetic experience; and being notably present in works of art.

Although the relative status of the characteristics is debated, there is substantial intuitive agreement as to which perceivable properties of things are aesthetic. These include, for example: beauty, sublimity, grace, elegance, delicacy, harmony, balance, unity, power, anguish, sadness, tranquility, serenity, and melancholy. It is evident that expressive properties, which arguably belong only to works of art and not to natural objects, constitute a significant subset of aesthetic properties.

Aesthetic Experience

Among the characteristics that have been proposed as distinguishing aesthetic states of mind (whether attitudes, perceptions, emotions, or acts of attention) from others are: disinterestedness, or detachment from desires, needs and practical concerns; non-instrumentality, or being undertaken or sustained for their own sake; contemplation or absorbtion, with consequent effacement of the subject; focus on an object's form; focus on the relation between an object's form and its content or character; focus on the aesthetic features of an object; and centrality in the appreciation of works of art. It is still a matter of debate whether these criteria, either individually or in some combination, adequately define aesthetic experience.

PROBLEMS IN AESTHETICS

Evidently, among the problems aesthetics addresses are the interrelated characterizations of art, aesthetic properties, and aesthetic experience. These broad problems engender many more specific ones.

The issue of the definition of art leads naturally to many further issues: the ontology of art; the process of artistic creation; the demands of artistic appreciation; the concept of form in art; the role of media in art; the analysis of representation and expression in art; the nature of artistic style; the meaning of authenticity in art; and the principles of artistic interpretation and evaluation. The philosophy of art is, in fact, sometimes conceived of as metacriticism, or the theory of art criticism (Beardsley, 1981).

The ontology of art concerns the question of exactly what sort of object a work of art is, and how this might vary between different art forms. Philosophers have asked whether a work of art is physical or mental, abstract or concrete, singular or multiple, created or discovered, notationally definable or only culturally specifiable; and they have asked what authenticity of a work of art consists in.

Interest in the creative process in art concerns the question of whether the creative process can be characterized in any general way, and the relevance of knowledge of the creative process (and more generally of the historical context of creation) to appreciation of works of art.

Issues about artistic form include the status of formalism as a theory of art, the different kinds of form manifested in different art forms, and the relation of form to content and of form to medium.

Among the modes of meaning that inhere in works of art, perhaps the most important are representation and expression. Goodman (1976) argues that exemplification is an equally important mode.

Accounts of representation (usually with special reference to pictorial representation) have been proposed in terms of resemblance between object and representation; perceptual illusion (Gombrich, 1960); symbolic conventions (Goodman, 1976); 'seeing-in' (Wollheim, 1987); world-projection (Wolterstorff, 1980); make-believe (Walton, 1990); recognitional capacities (Schier, 1986); resemblance between experience of object and experience of representation (Budd, 1995; Hopkins, 1998); and information content (Lopes, 1996).

Accounts of artistic expression (usually with special reference to the expression of emotion) have been proposed in terms of personal expression by the artist; induced empathy with the artist; metaphorical exemplification; correspondence (Wollheim, 1987); evocation (Matravers, 1998); imaginative projection (Scruton, 1997); expressive appearance (Kivy, 1989; Davies, 1994); and imagined personal expression (Vermazen, 1986; Levinson, 1996).

Concerning artistic style, attention has focused on the distinction between individual and period style, on the psychological reality of style, on the interplay between style and representational objective, and on the role that cognizance of style plays in aesthetic appreciation.

Concerning the interpretation of art, attention has focused on the relevance of artists' intentions; on the diversity of interpretative aims; on the debate between critical monism and critical pluralism; on the similarities and differences between critical and performative interpretation; and on the relationship between interpretation and maximization of value.

Concerning the evaluation of art, attention has focused on the question of its objectivity or subjectivity; on the relation between artistic value and pleasurability; on the relation between the value of art as a whole and the value of individual works of art; on the existence of general criteria of value across art forms; and on the relevance of a work's historical influence, ethical import, emotional power, and cognitive reward to its evaluation as art.

Certain concepts are relevant to the understanding of many, if not all, works of art: for example, the concepts of intention, fiction, metaphor, genre, narrative, genius, forgery, performance, and tragedy.

There are other questions concerning the relationships between art and other domains or aspects of human life, such as emotion, knowledge, and morality. For example, there are the questions of how we can coherently have emotions for characters whom we know to be fictional; whether art can be a vehicle of knowledge; and whether art can contribute to moral education.

There are also questions relating to particular art forms: for example, whether photography is an inherently realistic medium; whether poetry can be usefully paraphrased; whether the basic form of music is local or global; and whether narration operates similarly in novels and films.

The question of the nature of aesthetic properties leads naturally to questions about realism in relation to such properties; the supervenience relation between aesthetic properties and the non-aesthetic properties on which they depend; the range of aesthetic properties to be found in the natural world; the special status of beauty among aesthetic properties; the difference between the beautiful and the sublime; the degree of objectivity of judgments of beauty; the relations between artistic, natural, and human beauty; and the relationship between the aesthetic properties of artworks and their artistic properties, such as originality or seminality (which, although appreciatively relevant, are not directly perceivable as are aesthetic properties).

Finally, discussions of aesthetic experience open into discussions of the nature of perception, reason, imagination, feeling, memory, and mood, in relation to art or nature.

UNDERSTANDING WORKS OF ART

Categorial Perception

Walton (1970) follows Beardsley and Sibley in taking aesthetic properties to be perceptual, gestalt-like, non-rule-governed, and dependent on an object's lower-level perceptual properties. But Walton insists that aesthetic properties are dependent as well on the artistic categories (for example, of style, genre or medium) into which works of art may be said to fall. The category of a work is partly a matter of art-historical context, including factors such as the artist's intention, the artist's previous work, the artistic traditions in which the artist worked, and the artistic problems to which the artist is responding. If, as Walton argues, a work's aesthetic properties do not reside in perceivable structure alone, it is even more evident that its artistic properties depend on non-perceivable factors.

Artistic Expression

For Goodman (1976), artistic expression is a matter of an artwork exemplifying or drawing attention to some property it metaphorically possesses, in virtue of its general symbolic functioning.

For Tormey (1971), artistic expression is a matter of an artwork's possessing expressive properties that are related to intentional states, and which are ambiguously constituted by the non-expressive structural features underlying them.

Wollheim (1987) suggests that expressiveness in painting is a matter of intuitive correspondence or fit between the appearances that works of art present and feeling states of the subject, which are then projected onto those works in complex ways.

Davies (1994) offers a theory of musical expressiveness in terms of resemblances between musical patterns and human emotional behavior, and explores the variety of responses that listeners have to perceived expressiveness.

Scruton (1997) locates the perception of musical expression in our ability to inhabit 'from the inside' the gestures that music appears to embody, and thus to adequately imagine the inner states correlative with such gestures.

Levinson (1996) accounts for musical expressiveness in terms of music's ready hearability as the personal expression of an indefinite agent or persona.

Pictorial Representation

There are various accounts of our capacity to see what pictures depict and then respond to those depictions in aesthetically relevant ways. Currently the two most influential theories are Wollheim's (1987) 'seeing-in' theory and Walton's (1990) 'make-believe' theory.

Wollheim's theory is a development of Wittgenstein's idea of aspect perception, or 'seeing-as', for example, seeing a gnarled tree as an old woman. 'Seeing-in' applies to the parts of a picture as well as to the picture as a whole; and it involves simultaneous ('twofold') awareness of the picture's surface and the depicted content.

For Wollheim, seeing-in is a primitive visual capacity, at first exercised on natural phenomena, like stained rock faces, and later deliberately harnessed for making images. A large part of the aesthetic interest in pictures is due to the twofoldness of seeing-in, whereby we appreciate what is depicted, in a virtual three-dimensional space, in relation to the real two-dimensional pattern of marks before us.

Walton understands pictures as props in visual games of guided imagining, or make-believe. The configuration of marks that constitutes a picture prompts us to imagine we are seeing an object, and we imagine that our seeing those marks is a seeing of the object. Pictures generate fictional worlds; and what it is correct to imagine seeing in a picture is determined by implicit rules and conventions. In addition, in interacting with a picture visually, the viewer generates transient fictional worlds specific to him or her.

It is not yet clear whether Wollheim's and Walton's proposals are reconcilable. For Walton, Wollheim's seeing-in is to be analyzed in terms of imagined seeing, whereas for Wollheim, seeing-in is an activity prior to and more fundamental than imagined seeing, however important such seeing is in later phases of pictorial appreciation.

AESTHETICS AND COGNITIVE SCIENCE

Underlying aesthetic experience are certain mental states and processes: those involved in creating, perceiving, understanding and appreciating works of art. Accordingly, much recent work in aesthetics has taken into account empirical research on the human mind.

Pictorial Perception

Schier (1986) appeals directly to facts about ordinary visual processing in support of a theory of pictures. He proposes that a representation is pictorial just insofar as it recruits the visual recognitional capacities subjects already possess for familiar objects, so that a picture represents an object *O* if it triggers in subjects who view it the same capacities for recognition that would be triggered by the sight of *O* in the world. Schier emphasizes that pictorial competence, unlike language learning, is characterized by 'natural generativity', whereby once a subject can decipher a few pictures of a given sort, he or she can generally decipher any number of such pictures, however novel.

Lopes (1996) maintains that the essence of pictorial representation is the furnishing of similar visual information by picture and object. He proposes an 'aspect-recognition' theory of depiction, according to which successful pictures embody nonconceptual aspectual information sufficient to trigger recognition of their objects in suitable perceivers. Developing a theme of Gombrich (1960), Lopes proposes that the essence of depiction as a mode of representation is its inevitable selectivity, so that

a picture of whatever style (unlike a description) is explicitly noncommittal about certain represented properties of its object, precisely in virtue of being explicitly committal about others.

Lopes (1997) argues for the possibility of purely tactile pictures (though he overlooks certain experiential asymmetries between tactile and visual pictures (Hopkins, 2000)). And Lopes (1999) draws on color perception theory to demonstrate how pictures depict the colors of the world without actually replicating them.

Musical Comprehension

Raffman (1993) investigates aspects of the apparent ineffability of music in terms of facts about the mental processing of music. She sketches a cognitivist account of music perception that draws on the work of Jackendoff and Lerdahl, whereby an experienced listener unconsciously assigns a structural description to heard music in accordance with internalized rules governing musical parameters. Though a subject may become aware in an inarticulate way of how he or she is parsing the music, the representations involved elude verbal grasp. Raffman calls this 'structural ineffability'.

'Feeling ineffability' is a result of the fact that knowledge of music is sensory-perceptual: to know a piece is to know how it sounds. Knowledge of music depends on knowing what, say, a minor third actually sounds like.

Raffman also sketches a psychology of musical nuances – those highly specific values of pitch, rhythm and timbre that characterize any musical event – and shows how this underpins what she calls 'nuance ineffability'. Nuance ineffability arises from the fact that in aural experience we are conscious of differences, say, in pitch more subtle than we can inwardly label or classify: we are unable to remember and judge of such nuances. The basic problem is one of memory: reporting a perception requires retention of information in a manner that allows for stable association with verbal labels, but our ability to register musical nuances exceeds the mental schemata we seem to have available for storing them. Nuance ineffability in the reception of musical events poses a problem for those accounts of consciousness that identify it with sentential episodes: if we consciously experience aural nuances but cannot represent them propositionally, then there must be more to consciousness than sentential-type representation can allow.

DeBellis (1995) discusses statements of the form '*S* hears *x* as *F*'. He takes the meaning of such ascriptions to be given by the content of the mental states that ground or justify them; and he takes these mental states to be ones in which passages of music are represented, correctly or incorrectly, as having certain properties.

DeBellis argues that the music-hearing of an ordinary (experienced but untrained) listener is both weakly and strongly nonconceptual. Weak nonconceptuality is the claim that the ordinary listener's hearing of music does not involve those concepts in terms of which an analyst might describe that hearing. Strong nonconceptuality is the claim that the ordinary listener's hearing of music does not involve concepts of any sort, even narrowly perceptual ones. A consequence of both theses is that ordinary musical perception is not a process of acquiring beliefs, since beliefs presuppose concepts. Rather, the comprehending ordinary listener represents the music being heard as having some qualities or features, without thereby believing that it has those qualities or features.

DeBellis's argument for strong nonconceptuality is as follows. Current psychological theory suggests that ordinary listeners represent all heard sound events of a given kind *K* in the same way. Yet they generally prove unable to discriminate between *K* and non-*K* events, and generally fail to judge two *K* events to be similar. This suggests that such listeners lack even a perceptual concept of *K*, and that the conversion of ordinary listeners into expert listeners is in large part the acquisition of perceptual concepts for musical features which ordinary and expert listeners register alike.

Robinson (1994) explores the relevance of recent research on emotions to theories of musical expression.

Jackendoff (1991) proposes an explanation of musical affect in terms of discrepancies between conscious knowledge of musical progression and unconscious states of a postulated musical parser.

Levinson (1998a) questions the extent to which basic musical understanding requires a grasp of large-scale form.

Fictional Appreciation

Feagin (1996) uses simulation theory to help understand what responding appropriately to a work of literary fiction might involve. She proposes that appreciation of a literary text typically involves mental shifts in response to the flow of the text. Such mental shifting is a prerequisite to empathizing with fictional characters. Empathizing involves simulating another's mentality, in effect conforming

one's own mind to that of one's target, by putting one's mind 'offline' and then 'inputting' what one takes to be the experiences of one's target, thus generating an affective 'output' in oneself. This account might explain how, in responding empathically to a work of fiction, one may thereby be acquiring real knowledge – knowledge of 'what it is like' to be a certain person in a certain situation – which is often said to be one of the rewards of reading imaginative literature.

Currie (1995a) criticizes meta-representational theories of pretence, according to which pretending involves decoupling inner symbols from their normal semantic implications or flagging such symbols with special 'pretence' markers. He suggests that such views confuse mental contents and psychological attitudes towards them, and make it hard to explain the specific character of individual pretendings.

Currie questions whether, as has been claimed, empirical studies of autism and related cognitive disorders support meta-representational theories of pretense. He argues instead that such studies support the identification of imaginative pretense with simulation. He suggests that imagination is an 'internal simulator', a part of our mental equipment that evolved for the purpose of strategy-testing. This hypothesis can explain some aspects of appreciation of literary fiction, in particular our capacity to be affected by the plights of fictional characters: empathizing with characters involves the same process as empathizing with real people, taking on their beliefs and desires in imagination. The only additional requirement is that we first imagine them to exist.

Currie (1995b) considers a number of central issues in film theory from a cognitivist perspective, such as how films represent, what cinematic content consists in, and how we interpret cinematic narratives. He argues that the essence of cinematic experience is not seeming to see, or even imagining seeing, the objects and events represented in a film, but rather 'impersonally visually imagining' those objects and events. Since imagining is construed as a form of simulation – whether of others' states or of one's own states on other occasions – the essence of cinematic experience is thus simulated perceptual belief.

Currie accounts for the special 'realism' of film in terms not of its capacity to induce perceptual illusion, but of its mode of representation, whereby temporal properties are represented by temporal ones and spatial properties by spatial ones, which makes cinematic experience of objects similar to experience of those objects in the world.

References

Beardsley M (1981) *Aesthetics: Problems in the Philosophy of Criticism*. Indianapolis, IN: Hackett. [First published 1958.]

Budd M (1995) *Values of Art*. London, UK: Penguin.

Carroll N (2001) *Beyond Aesthetics*. Cambridge, UK: Cambridge University Press.

Currie G (1995a) Imagination and simulation: aesthetics meets cognitive science. In: Davies M and Stone T (eds) *Mental Simulation*, pp. 151–169. Oxford, UK: Blackwell.

Currie G (1995b) *Image and Mind: Film, Philosophy, and Cognitive Science*. Cambridge, UK: Cambridge University Press.

Danto A (1981) *The Transfiguration of the Commonplace*. Cambridge, MA: Harvard University Press.

Davies S (1991) *The Definition of Art*. Ithaca, NY: Cornell University Press.

Davies S (1994) *Musical Meaning and Expression*. Ithaca, NY: Cornell University Press.

DeBellis M (1995) *Music and Conceptualization*. Cambridge, UK: Cambridge University Press.

Dewey J (1934) *Art as Experience*. New York, NY: G. P. Putnam.

Dickie G (1997) *The Art Circle*. Chicago, IL: Spectrum Press. [First published 1984.]

Feagin S (1996) *Reading With Feeling*. Ithaca, NY: Cornell University Press.

Gadamer H (1986) *The Relevance of the Beautiful and Other Essays*. Cambridge, UK: Cambridge University Press.

Goldman A (1995) *Aesthetic Value*. Boulder, CO: Westview Press.

Gombrich E (1960) *Art and Illusion*. Princeton, NJ: Princeton University Press.

Goodman N (1976) *Languages of Art*, 2nd edn. Indianapolis, IN: Hackett. [First published 1968.]

Hopkins R (1998) *Picture, Image and Experience*. Cambridge, UK: Cambridge University Press.

Hopkins R (2000) Touching pictures. *British Journal of Aesthetics* **40**: 149–167.

Jackendoff R (1991) Musical parsing and musical affect. *Music Perception* **9**: 199–230.

Kivy P (1989) *Sound Sentiment*. Philadelphia, PA: Temple University Press.

Levinson J (1990) *Music, Art, and Metaphysics*. Ithaca, NY: Cornell University Press.

Levinson J (1996) *The Pleasures of Aesthetics*. Ithaca, NY: Cornell University Press.

Levinson J (1998a) *Music in the Moment*. Ithaca, NY: Cornell University Press.

Lopes D (1996) *Understanding Pictures*. Oxford, UK: Oxford University Press.

Lopes D (1997) Art media and the sense modalities: tactile pictures. *Philosophical Quarterly* **47**: 425–440.

Lopes D (1999) Pictorial color: aesthetics and cognitive science. *Philosophical Psychology* **12**: 415–428.

Matravers D (1998) *Art and Emotion*. Oxford: Clarendon Press.

Raffman D (1993) *Language, Music, and Mind*. Cambridge, MA: MIT Press.

Robinson J (1994) The expression and arousal of emotion in music. *Journal of Aesthetics and Art Criticism* **52**: 13–22.
Schier F (1986) *Deeper Into Pictures*. Cambridge, UK: Cambridge University Press.
Scruton R (1997) *The Aesthetics of Music*. Oxford: Oxford University Press.
Sparshott F (1982) *Theory of the Arts*. Princeton, NJ: Princeton University Press.
Tormey A (1971) *The Concept of Expression*. Princeton, NJ: Princeton University Press.
Vermazen B (1986) Expression as expression. *Pacific Philosophical Quarterly* **67**: 196–224.
Walton K (1970) Categories of art. *Philosophical Review* **79**: 334–367.
Walton K (1990) *Mimesis as Make-Believe*. Cambridge, MA: Harvard University Press.
Wollheim R (1980) *Art and Its Objects*, 2nd edn. Cambridge, UK: Cambridge University Press. [First published 1968.]
Wollheim R (1987) *Painting as an Art*. Princeton, NJ: Princeton University Press.
Wolterstorff N (1980) *Worlds and Works of Art*. Oxford: Oxford University Press.

Further Reading

Beardsley M (1982) *The Aesthetic Point of View*. Ithaca, NY: Cornell University Press.
Currie G (1989) *An Ontology of Art*. London, UK: Macmillan.
Currie G (1990) *The Nature of Fiction*. Cambridge, UK: Cambridge University Press.
Kivy P (1990) *Music Alone*. Ithaca, NY: Cornell University Press.
Lamarque P (1996) *Fictional Points of View*. Ithaca, NY: Cornell University Press.
Levinson J (1998b) Wollheim on pictorial perception. *Journal of Aesthetics and Art Criticism* **56**: 227–233.
Scruton R (1974) *Art and Imagination*. London, UK: Methuen.
Sibley F (2001) *Approach to Aesthetics*. Oxford, UK: Oxford University Press.
Stecker R (1997) *ArtWorks: Definition, Meaning, Value*. University Park, PA: Penn State University Press.
Walton K (1987) Style and the products and processes of art. In: Lang B (ed.) *The Concept of Style*, pp. 72–103. Ithaca, NY: Cornell University Press.

Anomalous Monism

Intermediate article

Louise M Antony, Ohio State University, Columbus, Ohio, USA

CONTENTS

According to the theory of anomalous monism, every individual mental event is identical with some physical event, but there are no strict laws relating physical kinds to mental kinds.

INTRODUCTION

Anomalous monism (AM) is a theory of the mind–body relationship developed by Donald Davidson (1970, 1974, 1980). The theory states that while every individual mental event is identical with some physical event, there are no strict laws relating physical kinds to mental kinds.

HISTORY

AM grew out of Davidson's attempt to defend a causal theory of intentional action. In Davidson (1963) he considered the question of how appeal to an agent's reasons could serve as an explanation of the agent's actions. Earlier accounts of these 'rational explanations' (or, as Davidson referred to them, 'rationalizations') had it that the explanatory power of such appeals lay entirely in their revealing the action performed to be one that was reasonable for the agent. Advocates of such accounts argued that rational explanations could not be causal in character, for two reasons: (1) the normative element in rational explanation is not present in nonteleological causal explanation; and (2) rationalizations involve mere redescriptions of the action explained, instead of citing prior events contingently connected to the action explained.

Davidson's main objection to such views was that they could not adequately account for the force of the 'because' in rational explanation. A belief and desire of mine may make it reasonable for me to perform a certain action, but something more is needed for it to be true that it is because of the belief and desire that I do it. This 'something more', Davidson argued, could only be a causal connection.

It remained for Davidson to answer objections to the suggestion that reasons were causes. In addition to the objections noted above, there was the following: causation involves laws, and there are no laws relating beliefs and desires to the behavior they rationalize. Davidson's response contained the germs of the theory of anomalous monism: for a law to back a singular causal claim, it is not necessary that the law describe the subsumed events in the same vocabulary as is used in the causal claim. This set the stage for anomalous monism: Davidson would accept his opponents' contention that there was an irreducibly normative element in rational explanations, but would argue that this counted only against there being psychological laws, not against psychological explanation's being causal.

AM is one version of what is now known as 'nonreductive materialism', a view of mind that emerged in the early 1970s, partially in reaction to the 'strong' or 'type–type' reductionism of such philosophers as U. T. Place (1956) and J. J. C. Smart (1962), who argued that mental state types could be identified with, and hence reduced to, neurophysiological state types. Davidson is widely credited with one set of arguments (to be discussed below) against the possibility of such strong reductions for at least one important class of mental states, namely propositional attitude states. A different kind of argument against strong reductionism for mental states came from Hilary Putnam (1967), who argued that, because mental states were functional states, they could be 'multiply realized', that is, realized in a variety of physically different systems: pain might be a matter of c-fibres firing in humans, but of something altogether different in a Martian. (See also Fodor, 1974.) Putnam's arguments were designed to show that there could be no physical necessary conditions for the instantiation of a mental state type; Davidson's were meant to show that there could be no

physical sufficient conditions. Both views are properly classed as 'token-reductionist' materialist theories, for according to both, it makes sense to identify individual mental events with physical events.

ARGUMENTS FOR ANOMALOUS MONISM

Davidson's (1970) argument for AM was that the theory provided an attractive solution to an important puzzle, namely, how the following three principles could all be true together:

> Principle of causal interaction (PCI): At least some mental events interact causally with other events.
> Principle of the nomological character of causality (NCC): Events related as cause and effect fall under strict deterministic laws.
> Principle of the anomalism of the mental (AOM): There are no strict deterministic laws on the basis of which mental events can be predicted and explained.

According to anomalous monism, every mental event is identical to some physical event. In fact, all it is for an event to be a 'mental event' is for there to be a true description of the event in mentalistic terms: the same event could be just as accurately described using purely physicalistic terminology. Thus, every interaction between a mental event and a physical event is equally an interaction between two physical events. This secures PCI. To reconcile PCI with NCC and AOM, Davidson argues that there is a crucial difference between causation and subsumption under laws. The former is an extensional relation, holding of events however described, whereas the latter is an intensional relation, sensitive to the way in which events are picked out. AOM can then be understood as saying that there are no strict deterministic laws covering events described mentalistically, or that there are no such laws utilizing mentalistic predicates. Interpreted in this way, AOM can be seen to be consistent with NCC, which simply states that any pair of causally related events is subsumed under some law or other. If we assume, as Davidson seems to, that all physical predicates are apt for use in the statement of strict, deterministic laws, then the token-identity of every mental event with some physical event suffices to ensure that whenever a mental event interacts causally with a physical event, there is a way of describing the events so that they are covered by an appropriate law.

Perhaps the most controversial of the three principles is AOM (although NCC is also open to challenge – see below). Davidson offers brief and somewhat enigmatic remarks in its defence. Broadly, there are two arguments. The first is that psychological ascriptions answer to different kinds of evidence, and are constrained by different kinds of norms, from nonpsychological claims. As a result, Davidson (1970) says, 'mental and physical predicates are not made for each other'. Since laws can only 'bring together' predicates that are made for each other, there can be no psychophysical laws. The second argument is based on the 'holistic character of the cognitive field' (Davidson, 1974). According to Davidson, the content of a belief or desire depends on its role in the overall pattern of the agent's psychology and behavior. This makes it impossible for there to be local sufficient conditions for the correct ascription of a propositional attitude.

Possibly Davidson has something like the following arguments in mind. The norms that govern mental predicates are the norms of cogency and rationality: we are constrained to attribute beliefs and desires to agents in such a way as to make agents' behavior appear (as much as possible) rational, or else we will fail in ascribing psychological states at all – we will be 'changing the subject' (Davidson, 1970). Since physical predicates are not subject to this normative constraint, the possibility of laws connecting the mental to the physical entails the possibility of competition between the evidence relevant to the physical attributions and the considerations of 'overall cogency' appropriate to the psychological realm. If we could know, for example, that all x-neuron firings were, as a matter of law, beliefs that it is raining, then the physical evidence that someone's x-neurons were firing might warrant the attribution of this belief even if that attribution was not supported by the kinds of behavioral evidence we otherwise rely on, and even if the belief did not cohere properly with others of the agent's mental states. Moreover, if such a law were available, then we would have a localistic basis for the attribution of a belief with the content 'it is raining', in violation of the supposition that what constitutes a mental state's having a particular content is its playing a certain role in the agent's overall patterns of thought and action. Such a situation, Davidson thinks, would amount to the abandonment of our conception of ourselves as rational agents: 'we must conclude that nomological slack between the mental and the physical is essential as long as we conceive of man as a rational animal' (Davidson, 1970).

CRITICISMS OF ANOMALOUS MONISM

There are three major objections to Davidson's theory. The first concerns the conception of causation on which Davidson's original puzzle depends. NCC expresses a commitment to a regularity theory of causation. But there are competing accounts, according to which singular causal claims need not be 'backed' by general laws. If one adopts one of these alternative accounts of causation, then the original problem that was supposed to motivate anomalous monism is dissolved.

The second objection is that the theory does not, in fact, solve the problem of accounting for the explanatory force of rational explanations, because it leaves the rationalizing part of the explanations detached from the causal part. That is, because, according to AM, there is no possibility of relating, in a systematic way, the physical properties of a mental event with its mental properties, and because it is only the physical properties of the mental event that figure in causal laws, the mental properties appear to be causally inert, or 'epiphenomenal'. Davidson's response to these objections can be found in Davidson (1993). (That essay and many others on the issue of the causal relevance of mental properties can be found in Heil and Mele, 1993.) The issues raised in this connection led to a general debate as to whether the mental properties of mental events are causally relevant to the production of action.

The third objection challenges the adequacy of the arguments for AOM. Critics contend that if agents are in fact rational – and the degree to which they are is, arguably, an empirical matter – then the constraints of 'overall cogency' do in fact come into play in determining the adequacy of a proposed physical account of propositional attitude states. For suppose that we have some candidate reductive account of beliefs and desires. If we have independent justification for thinking that agents' behavior is usually reasonable given their beliefs and desires, then there will be a constraint on the adequacy of our candidate reduction theory that it have the consequence that the beliefs and desires we attribute on the basis of purely physical evidence coincide with the ones that we would attribute on the basis of considerations of overall cogency. There is thus no need to fear a competition between norms and evidence pertinent to the respective domains. (See Antony, 1989 and Van Gulick, 1980.)

A number of essays containing critical discussion of AM and the arguments for it can be found in McLaughlin and LePore (1985).

ANOMALOUS MONISM AND COGNITIVE SCIENCE

An important philosophical question in cognitive science is whether the concepts and generalizations of our ordinary mentalistic talk – 'folk psychology' – can be preserved and explained within a scientific psychological theory. If Davidson's theory is correct, then the prospects for such a scientific vindication of our pretheoretic notions seem poor.

Consider, for example, the computational model of mind. According to this view, psychological states are functional states of the brain involving physically realized mental representations. Cognitive processes like rational deliberation are computations defined over these representations. In this way, the causal transactions within the agent's brain can be shown to mirror rational relations among the agent's mental states. If this computational model is accurate, then there would be lawful connections between physical state-types of the brain and mental state-types, of the sort that would support inferences, at least in principle, from local physical evidence to conclusions about an agent's beliefs and desires. And this, it appears, would be incompatible with AOM.

If one rejects the computational model of the mind in favor of, say, a connectionist architecture, then the relevance of AM will depend on whether one thinks that the alternative architecture permits a type reconstruction of propositional attitude states.

However, Davidson is not altogether clear about whether AOM is supposed to rule out any kind of nomic connection involving mental types, or only strict (i.e., exceptionless) nomic connections. The former claim seems to be what the arguments in (Davidson, 1970) mean to establish; if the claim is weakened in the latter way, then the thesis seems considerably less interesting, and AM itself would appear to amount to little more than the claim that psychology is not a basic science. See Davidson (1993) for further discussion.

References

Antony L (1989) Anomalous monism and the problem of explanatory force. *Philosophical Review* **98**: 153–187.

Davidson D (1963) Actions, reasons, and causes. *Journal of Philosophy*, **60**: 685–700. [Reprinted in Davidson, 1980.]

Davidson D (1970) Mental events. In: Foster L and Swanson J (eds) *Experience and Theory*. Amherst, MA: University of Massachusetts Press/Duckworth Press. [Reprinted in Davidson, 1980.]

Davidson D (1974) Psychology as philosophy. In: Brown S (ed.) *Philosophy of Psychology*. New York, NY:

Macmillan Press/Barnes, Noble. [Reprinted in Davidson, 1980.]
Davidson D (1980) *Essays on Actions and Events*. Oxford: Oxford University Press.
Davidson D (1993) *Thinking causes*. In: Heil J and Mele A (eds) *Mental Causation*. Oxford: Oxford University Press.
Fodor JA (1974) Special sciences. *Synthese* **28**: 97–115.
Heil J and Mele A (eds) (1993) *Mental Causation*. Oxford: Oxford University Press.
McLaughlin BP and LePore E (eds) (1985) *Action and Events*. Oxford: Blackwell.
Place UT (1956) Is consciousness a brain process? *British Journal of Psychology* **47**(1): 44–50.
Putnam H (1967) The nature of mental states. In: Capitan WH and Merrill DD (eds) *Art, Mind, and Religion*. Pittsburgh, PA: University of Pittsburgh Press.
Smart JJC (1962) Sensations and brain processes. In: Chappell VC (ed.) *The Philosophy of Mind*. Englewood Cliffs, NJ: Prentice Hall.
Van Gulick R (1980) Rationality and the anomalous nature of the mental. *Philosophy Research Archives* **7**: 1404.

Further Reading

Antony L (1994) The inadequacy of anomalous monism as a realist theory of mind. In: Preyer G *et al*. (eds) *Language, Mind and Epistemology*, pp. 223–253. Dordrecht: Kluwer.
Campbell N (1997) The standard objection to anomalous monism. *Australasian Journal of Philosophy* **75**: 373–382.
Davidson D (1973) The material mind. In: Suppes P, Henkin L, Mosil GC and Joja A (eds) *Logic, Methodology and the Philosophy of Science*. Amsterdam: North-Holland. [Reprinted in Davidson, 1980.]
Evnine S (1991) *Donald Davidson*. Stanford, CA: Stanford University Press.
Honderich T (1982) The argument for anomalous monism. *Analysis* **42**: 59–64.
Johnston M (1985) Why having a mind matters. In: McLaughlin and LePore, 1985.
Kim J (1985) Psychophysical laws. In: McLaughlin and LePore, 1985.
Latham N (1999) Davidson and Kim on psychophysical laws. *Synthese* **118**: 121–144.
LePore E and Loewer B (1987) Mind matters. *Journal of Philosophy* **84**: 630–642.
Lycan WG (1981) Psychological laws. *Philosophical Topics* **12**: 9–38.
McDowell J (1985) Functionalism and anomalous monism. In: McLaughlin and LePore, 1985.
McLaughlin BP (1985) Anomalous monism and the irreducibility of the mental. In: McLaughlin and LePore, 1985.
Patterson SA (1996) The anomalism of psychology. *Proceedings of the Aristotelian Society* **96**: 37–52.

Artificial Intelligence, Gödelian Arguments against

Intermediate article

Peter Slezak, University of New South Wales, Sydney, Australia

CONTENTS

Gödel's incompleteness theorem says that there are arithmetical truths expressible in certain formal systems which cannot be proven within those systems. This result reveals certain inherent limitations on what is computable, which, according to some theorists, show that minds are not machines because they are not subject to the same limitations.

WHAT ARE GÖDELIAN ARGUMENTS?

Gödelian arguments attempt to exploit the famous incompleteness theorem of Kurt Gödel (1931) to show that the mind cannot be a machine. In slightly different forms, such arguments have been advanced by J. R. Lucas (1961) and R. Penrose (1989, 1995). Gödel himself expressed sympathy for such views, but there are significant differences among these accounts regarding the relevance of the mathematics to the nature of the mind.

In general terms, Gödelian arguments conclude that minds cannot be machines because there is something that minds can do but no machine can do. A similar kind of argument was made by René Descartes (1637), who claimed that no purely mechanical contrivance could show the infinite, creative novelty revealed in human language or action guided by knowledge. Descartes' argument concerned machines as understood at that time – namely, clockwork devices with a finite repertoire of behaviors. It was only with the work of Alan Turing (1936) that we arrived at a fundamentally different conception of machines that are not subject to the limitations recognized by Descartes.

Gödel's incompleteness result holds for formal, axiomatic systems such as that described in the *Principia Mathematica* of Russell and Whitehead (1910–1913), but it may be stated as a limitation on any computing machine that realizes a formal system. Gödel's theorem shows that there are true arithmetical propositions that are neither provable nor disprovable; that is, they are undecidable within their own system. Gödel's proof specifies such an arithmetical statement, but the claim that it is undecidable is not itself an arithmetical proposition, but a metamathematical one, since it is concerned with the notion of provability. That is, Gödel's theorem concerns statements about mathematical systems rather than statements of, or within, such systems. The undecidability of propositions within arithmetic is a syntactical claim of such metamathematical proof theory. However, since the undecidable proposition is, in fact, true, the formal axiomatic system is said to be 'incomplete', meaning that there are truths expressible in the system that are unprovable in it.

Gödel himself favored an anti-mechanist position. However, Hao Wang (1987, p. 146) says that Gödel recognized that 'his theorem does not settle the question of mind surpassing matter' and 'unlike certain ignorant philosophers, G[ödel] realizes that his incompleteness theorem does not by itself imply that the human mind surpasses all machines' (1987, p. 197). Wang cites the additional premises that Gödel required, including (1) that the mind is separate from matter, and (2) that the mind is not static, but constantly developing and therefore there is reason to believe the mind's states might converge to infinity (see Wang, 1974, p. 325).

Regarding the first of the above premises Gödel believes that materialism is a 'prejudice of our time' which will be disproved with the progress of science. His skepticism about the thesis of mechanism depends essentially on a belief in dualism, indeed a vitalism (Wang, 1974, p. 326; 1996, p. 193) and not directly on the implications of his incompleteness theorem.

The second of the above premises is the more significant. It is further explained by Wang (1974, p. 325; 1996, p. 194), who cites Gödel's conviction that 'Turing's argument for the adequacy of his definition [of computation] includes an erroneous proof of the stronger conclusion that minds and

machines are equivalent'. Gödel takes the argument of Turing's (1936) landmark paper as supporting the mechanist thesis, but Gödel suggests that the case is inconclusive because it depends on the assumption that a finite mind is capable of only a finite number of distinguishable states. Nevertheless, Gödel recognizes that 'the incompleteness results do not rule out the possibility that there is a theorem-proving computer which is in fact equivalent to mathematical intuition' (Wang, 1996, p. 186). In Wang's opinion, 'clearly he himself realized that such a refutation of mental computabilism is not convincing, as we can infer from his continued efforts to find other ways to achieve the desired refutation' (1996, p. 185). While inclining towards the anti-mechanist position, Gödel says that his incompleteness results imply only that if we are machines then 'either we do not know the exact specification of the computer or we do not know that it works correctly' (Wang, 1996, p. 186). Related conclusions about the limitations of self-knowledge have been drawn from Gödelian arguments by Whiteley (1962), Benacerraf (1967) and Slezak (1982, 1984).

J. R. Lucas (1961) articulated the Gödelian argument against mechanism in a detailed form which has given rise to a considerable critical literature. Lucas states the fundamental difference between minds and machines by saying that it follows from Gödel's theorem 'that given any machine which is consistent and capable of doing arithmetic, there is a formula which it is incapable of producing as being true – i.e., the formula is unprovable-in-the-system – but which we can see to be true' (1968, p. 44). From this alleged difference between the abilities of minds and machines, Lucas concludes 'that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines' (1968, p. 44).

Against critics, Lucas (1999) has continued to insist that his argument must be understood as having a special, essentially 'dialectical' form in which he does not claim the superiority of minds in all respects over all machines, but only that a mind can surpass any particular machine that might be proposed as a model in each case. Lucas claims that he can refute each particular instance of a machine offered by the 'mechanist' as a model of Lucas's mind. However complicated a machine we construct, it will always be vulnerable to the Gödel procedure for finding a formula unprovable-in-the-system that corresponds to the machine. 'This formula the machine will be unable to produce as being true, although a mind can see that it is true.' Lucas concludes that, since this procedure may be repeated endlessly with any machine proposed by the mechanist, 'the mind always has the last word' (1961, p. 48).

RESPONSES TO GÖDELIAN ARGUMENTS

Alan Turing (1950) considers the 'mathematical objection' to intelligent machines specifically arising from Gödel's theorem. In his brief treatment, Turing acknowledges that Gödel's theorem shows that there are certain questions which any machine must fail to answer correctly. However, Turing asks whether this 'proves a disability of machines to which the human intellect is not subject … it has only been stated, without any sort of proof, that no such limitations apply to the human intellect' (1950, p. 16). Nevertheless, Turing concedes that 'a certain feeling of superiority' over machines is 'no doubt quite genuine' and seems to acknowledge the force of the anti-mechanist argument, while discounting its importance.

Lucas (1961, p. 49) notes the irrelevance of such an argument to the central question at issue, which is not whether minds are superior to machines, but whether they are the same. Conceding that machines may be superior to minds in many, perhaps even most, respects, Lucas notes that it is enough to show that they are unable to do one thing, however trivial, that a mind is able to do, to establish that minds and machines are essentially different.

Putnam (1960) has argued that Lucas's argument rests on a misapplication of Gödel's theorem, which in fact asserts the conditional 'if T is consistent, then G is true', where T is the formal system and G is the undecidable sentence. Putnam's point is that Lucas cannot assert the truth of the Gödel sentence G alone, but only the entire conditional of which it is the consequence. Lucas (1968, p. 158) has conceded that the question of consistency is a matter of 'faith' and not susceptible to formal proof. However, Lucas points to the normative character of consistency in regulating our deliberately asserted statements. In any case, Putnam's objection is not decisive, because he says that the consistency of T is only 'unlikely if T is very complicated'.

Pertinently, Benacerraf (1967, p. 19) asks what exactly it is that Gödel's theorem precludes the machine from doing. The impossibility is clearly to prove G from the machine's own axioms according to the machine's own rules. Benacerraf asks: 'but can Lucas do that?' Evidently not. Chihara (1972), too, has noted a certain unclarity in Lucas's claimed ability to 'produce as true' the

Gödel sentence. Lucas uses the expression 'produce as true' to capture both what a machine is precluded from doing and also what a mind can do. Slezak (1982) suggests that Lucas's phrase appears to conflate the notions of proof and truth, which Gödel showed must be distinguished. What minds can do in following Gödel's theorem is to establish the truth of *G* at the level of a metamathematical argument, whereas what the machine cannot do is to generate the sentence *G* from the system's axioms. Thus, the notion of 'produce as true' is being used in two different senses, to describe both what minds can do and what machines cannot do. In spite of the crucial difference between truth and provability, Lucas incorporates them both in his notion of 'produce as true', and relies on the resulting equivocation to establish his desired conclusion. Whatever may be the admitted difficulty in understanding how minds can determine the truth of sentences such as *G* (Lucas, 1968, p. 148), Gödel's theorem has no bearing on this, nor does Gödel's theorem suggest any reason why machines might be incapable of doing it too.

Lucas insists that his argument depends on the mechanist proferring a specific machine which purports to model Lucas's mind before he can refute it by 'producing as true' its Gödel sentence. Lucas (1968, p. 146) says: 'If the mechanist maintains any specific thesis, I show that a contradiction ensues. But only *if*. It depends on the mechanist making the first move and putting forward his claim for inspection.' However, the relevant facts for assessing the mechanist thesis are objective ones concerning the relative abilities of minds and machines. These could not be inherently dependent on whether the mechanist chooses to make the first move by proposing a specific model. Lucas's dialectical strategy relies on the fact that he is able to determine the truth of a machine's unprovable sentence only by virtue of being outside the system; that is, by operating in the metalanguage. However, this amounts to merely restating Gödel's result, and not to demonstrating any essential difference between minds and machines. As J. Webb (1980, p. 230) has pointed out, 'the real source of Lucas's feeling of superiority here is just the *effectiveness* with which he can find the Achilles' heel of any machine'. Gödel's diagonal argument is itself a formal procedure which can be carried out by a machine.

Benacerraf's (1967) reconstruction of Lucas's argument leads him to conclude that, if we are Turing machines, then it seems we may be precluded from obeying the injunction 'know thyself'. Benacerraf construes this as a limitation on empirical psychology. However, Slezak (1982, 1984) has suggested that, if the indexical, self-referential features of the relevant statements are properly taken into account, Benacerraf's own account suggests a limitation for self-knowledge only in the sense of first-person, introspective knowledge. As we saw earlier, this was precisely the alternative to an anti-mechanist conclusion that Gödel himself suggested. The appropriate question for Lucas is not what he can know about the machine, but rather what he can know about himself. Slezak (1982) suggests that the close affinities of Gödel's result with familiar paradoxes of self-knowledge (Popper, 1950; Gunderson 1970) are highly suggestive, so that, far from refuting mechanism, Gödel's theorem may even provide persuasive support for it. As Webb (1980, p. vii) suggests in his detailed study, quite the reverse of Lucas's claims, 'Gödel's work was perhaps the best thing that ever happened to both mechanism and formalism'.

A foundational tenet of cognitive science is the computational character of the mind. The 'functionalist' view of so-called 'strong AI' (Searle, 1980) asserts that minds may be not only simulated on machines, but embodied or realized in them. Accordingly, as Penrose (1995) asserts, if Gödelian arguments are sound, then the ambitions of cognitive science and artificial intelligence are unattainable in principle. In a slight variant of Lucas's argument, Penrose (1995, p.75) takes the computational insolubility of the 'halting problem' to imply that mathematical intuition cannot be formalized in any algorithm. In particular, Penrose sees Gödelian arguments as showing that certain features of minds, such as consciousness, free will, creativity, and insight, are beyond the realm of ordinary rule-governed – and therefore psychological or even physical – explanation. However, Dennett (1989) suggests that such arguments constitute a *reductio ad absurdum* which counts against the Gödelian arguments themselves. Moreover, Dennett points out that it is a non sequitur to conclude that no algorithm can characterize a human ability such as playing chess, or understanding mathematics, just because there is no rule for checkmate, or mathematical insight. These tasks may be performed at any level, including our own, by heuristic problem-solving methods, which are algorithms in the sense of being rule-governed computer programs, but are not recipes that guarantee success.

References

Benacerraf P (1967) God, the Devil, and Gödel. *The Monist* **51**: 9–32.

Chihara C (1972) On alleged refutations of mechanism using Gödel's incompleteness results. *Journal of Philosophy* **64**: 507–526.

Dennett D (1989) Murmurs in the Cathedral. *Times Literary Supplement* **4513**: 1055–1056.

Descartes R (1637/1985) *The Philosophical Writings of Descartes*, vol. I: *Discourse and Essays*, translated by J. Cottingham, R. Stoothoff and D. Murdoch. Cambridge, UK: Cambridge University Press.

Gödel K (1931) Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik* **38**: 173–198. [Translated as 'On formally undecidable propositions of the *Principia Mathematica* and related systems I'. In: Davis M (ed.) (1965) *The Undecidable.* New York, NY: Raven Press.]

Gunderson K (1970) Asymmetries and mind–body perplexities. In: Radner M and Winokur S (eds) *Minnesota Studies in the Philosophy of Science*, vol. IV, pp. 273–309. Minneapolis, MN: University of Minnesota Press.

Lucas JR (1961) Minds, machines and Gödel. *Philosophy* **36**: 112–127. [Reprinted in: Anderson AR (ed.) (1964) *Minds and Machines*, pp. 43–59. Englewood Cliffs, NJ: Prentice-Hall.]

Lucas JR (1968) Satan stultified: a rejoinder to Paul Benacerraf. *The Monist* **52**: 145–158.

Lucas JR (1999) *The Gödelian Argument: Turn Over the Page*. http://users.ox.ac.uk/~jrlucas/turn.html.

Penrose R (1989) *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*. Oxford, UK: Oxford Unversity Press.

Penrose R (1995) *Shadows of the Mind: A Search for the Missing Science of Consciousness*. London, UK: Vintage.

Popper K (1950) Indeterminism in quantum physics and in classical physics. *British Journal for the Philosophy of Science* **1**: 117–133.

Putnam H (1960) Minds and machines. In: Hook S (ed.) *Dimensions of Mind*, pp. 138–164. New York, NY: New York University Press. [Reprinted in: Anderson AR (ed.) (1964) *Minds and Machines*, pp. 72–97. Englewood Cliffs, NJ: Prentice-Hall.]

Russell B and Whitehead AN (1910, 1912, 1913) *Principia Mathematica*, 3 vols. Cambridge, UK: Cambridge University Press.

Searle J (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**: 417–424.

Slezak P (1982) Gödel's theorem and the mind. *British Journal for the Philosophy of Science* **33**: 41–52.

Slezak P (1984) Minds, machines and self-reference. *Dialectica* **38**(1): 17–34.

Turing AM (1936) On computable numbers with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2* **42**: 230–265. [Reprinted in: Davis M (ed.) (1965) *The Undecidable*. New York, NY: Raven Press.]

Turing AM (1950) Computing machinery and intelligence. *Mind* **59**: 433–460. [Reprinted in: Anderson AR (ed.) (1964) *Minds and Machines*, pp. 4–30. Englewood Cliffs, NJ: Prentice-Hall.]

Wang H (1974) *From Mathematics to Philosophy*. New York, NY: Humanities Press.

Wang H (1987) *Reflections on Kurt Gödel*. Cambridge, MA: Bradford/MIT Press.

Wang H (1996) *A Logical Journey: From Gödel to Philosophy*. Cambridge, MA: MIT Press.

Webb JC (1980) *Mechanism, Mentalism and Metamathematics: An Essay on Finitism*. Dordrecht, Netherlands: Reidel.

Whiteley CH (1962) Minds, machines and Gödel: a reply to Mr Lucas. *Philosophy* **37**: 61–62.

Further Reading

Dennett D (1990) Betting your life on an algorithm. *Behavioral and Brain Sciences* **13**: 660–661.

Hofstadter DR (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic Books.

Nagel N and Newman JR (1958) *Gödel's Proof*. New York, NY: New York University Press.

Penrose R (1997) *The Large, the Small and the Human Mind*. Cambridge, UK: Cambridge University Press.

Smullyan R (1987) *Forever Undecided: A Puzzle Guide to Gödel*. New York, NY: Alfred A. Knopf.

Artificial Intelligence, Philosophy of

Introductory article

Eric Dietrich, Binghamton University, Binghamton, New York, USA

CONTENTS

The philosophy of artificial intelligence examines the foundational assumptions, methodologies, and consequences of the field of artificial intelligence.

WHAT IS THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE?

The philosophy of artificial intelligence is a collection of issues primarily concerned with whether or not artificial intelligence (AI) is possible; that is, with whether it is possible to build an intelligent thinking machine. Of ancillary concern is the issue of whether humans and other animals are best thought of as machines themselves.

The most important of the 'whether-possible' problems lie at the intersection of theories of the semantic contents of thought and the nature of computation. One view of human thinking is that it is the manipulation of contentful thoughts. When we engage in such cognitive processes as making inferences, recognizing patterns, planning and executing activities, etc., the thoughts we have and manipulate pick out or refer to various things in our world. This view seems so plausible that it is difficult to imagine how it could be false; but when it is translated to computers, its truth is much easier to doubt, for it entails that machine cognition would be just algorithmic manipulation (computing) of certain kinds of contentful data structures. Here is where the philosophical problem enters: while thoughts obviously have content, it is far from clear that the computations and data structures do. When a computer adds 1 to 1, do its internal states actually denote, or refer to, the number 1? If a computer derives 'Fido will chase cats' from 'Dogs chase cats' and 'Fido is a dog,' do any of its internal states actually pick out Fido, dogs, cats, and the act of chasing? If computations are radically different from thoughts in that they cannot have semantic content, then it is unlikely that computers can think.

Such problems are usually referred to in AI and cognitive science as 'the problem of mental content' or 'the problem of representational content.' In philosophy, the problem of mental content is sometimes called 'the problem of intentionality.' Often this latter term is used when problems of content are combined with the problems of epistemology and consciousness. The Chinese Room problem fits in here (see later).

A second set of 'whether-possible' problems surrounds the nature of rationality. Humans constantly evaluate ways of achieving goals and rank them according to various measures such as the probability of success, efficiency, and consequences. Humans also evaluate the goals themselves, as well as the means needed to achieve the goals. In so doing, humans constantly gauge the relevance of one piece of information to another, the relevance of one goal to another, and the relevance of evidence to achieving a goal or holding a belief. Humans do this evaluation with varying degrees of success, but often are quite successful at it. However, some philosophers believe that computers cannot do such evaluation at all; or that if they can, their methods of evaluation are too brittle to capture anything as robust as human rationality. But if human-level rationality is not obtainable, then the project of building an intelligent machine will fail.

A third set of problems revolves around the seemingly transcendent reasoning powers of the human mind. All these problems begin with Kurt Gödel's famous 'incompleteness theorem'. This theorem states that logic together with number theory (and consistent extensions) contains true statements that are unprovable. Gödel's theorem basically states that any suitably robust, formal

system is incomplete; but all computer systems precisely instantiate formal systems, hence all computer systems are incomplete. Moreover, from within the given formal system, the incompleteness cannot be proved; to do that, one has to step outside the formal system and prove a metamathematical result (this is exactly what Gödel did). Computer systems, being instantiated formal systems, cannot step outside of themselves. Hence humans can do something computers cannot, so humans aren't computers. And since humans' ability to step outside formal systems is a crucial part of their intelligence, it seems to follow that computers cannot be intelligent, at least not in the same way as humans.

A final collection of 'whether-possible' problems concerns the architecture of an intelligent machine. Currently there are four main positions in this debate, each claiming that its approach to building intelligent machines is the best (or the only) way to achieve success. Accordingly, the philosophical issues here are methodological: What are the real differences between the four main approaches? Is it likely that one could succeed where the others could fail? And if so, why?

There are many other important philosophical questions apart from the whether-possible type. Can a computer be conscious? Can a computer have a moral sense? Is it moral for humans to attempt to build an intelligent machine? If we did build such a machine, would turning it off be the equivalent of murder? If we had a race of such machines, would it be immoral to force them to work for us as slaves?

EARLY WORK ON AI

AI research is dedicated to the proposition that it is possible to build a machine, a computer, that is as intelligent as, if not more intelligent than, a human. Why would anyone believe that such an endeavor is possible? Computers are not even alive. A brief examination of the historical beginnings of AI will help to put into context the doubts about AI that philosophers raise.

The roots of AI are often said to lie in the work of Alan Turing and Alonzo Church whose development of the mathematical theory of computation was one of the most important developments in the entire history of mathematics. However, the relevance to AI of the theory of computation was established after AI had already emerged. The primary influence on the early AI researchers was the work of cyberneticists and neurophysiologists.

The first idea was that the computer was not just capable of numerical calculations; it was a general symbol manipulator capable of performing virtually any task having to do with information. The symbols manipulated could represent anything: words, propositions, concepts, dogs, cats, rocks, etc. The idea that computers could be more than 'number crunchers' was revolutionary. ENIAC, one of the first computers, was intended solely for numerical calculations. And Howard Aiken, who built the first electromechanical computer in the United States, once said: 'If it should ever turn out that the basic logics of a machine designed for the numerical calculation of differential equations coincide with the logics of a machine intended to make bills for a department store, I would regard this as the most amazing coincidence that I have ever encountered.' Aiken, of course, was amazed.

From the work of the neurophysiologists, around the middle of the twentieth century, came the idea that at least some neural activity was information processing, and hence that some kinds of thinking could be explained in terms of processing information.

Then, late in 1955, Alan Newell and Herbert Simon (and others) brought these two ideas together. Computers manipulate information; thinking is manipulating information; therefore, thinking might be computing. This insight resulted in a profound hypothesis that is the heart of AI. Newell and Simon also provided researchers with the fundamental ingredient of information manipulation, namely, the *symbol*. A symbol, or what is now called a *representation*, is considered to be a fundamental constituent of both thinking and computing.

THE PROBLEM OF REPRESENTATIONAL CONTENT

Philosophers have been intensely interested in symbols (words, primarily) and their meanings since before Socrates. Essentially, the puzzle is: how can we think about things in the world? It is clear that our thoughts about the things around us (such as a pet dog) are really quite different from the things themselves, if for no other reason than thoughts abstract and leave out information relative to what is being thought about (dog thoughts are not dogs). How then do these thoughts manage to represent what they do if they leave out information? How can one's thought, or mental representation, of one's pet dog be about *this* dog rather than some other dog if the representation leaves out information? Moreover, we can think about things that are not immediately present (the dark side of the moon), things that do not exist

(unicorns), and things that cannot exist (round squares). How is this possible?

However it is that human thoughts have content, many doubt that the computations going on inside a computer are about anything at all. When a computer adds 1 and 1 to get 2, its data structures do not seem to refer to numbers and adding. It is noted sometimes that computers do not seem to *know* anything about the numbers 1 and 2 and the function 'plus'. When a computer adds, a cascade of causal processes occurs which implements an algorithm which in turn, if followed exactly, guarantees that two numbers will be added. Consider an analogy: When a mousetrap catches a mouse, the mousetrap does not represent the mouse; it merely catches it (the mousetrap clearly does not know anything about mice or even the mouse it is catching). Why is the situation any different for computers and their processes? If there is no difference, perhaps computations are contentless. But thoughts are not. Hence thinking cannot be computing. Hence a computer cannot think: it can merely compute. Yet we saw above that there are good reasons for hypothesizing that thinking is computing, and that a computer can think.

One way out of this dilemma is to attempt to develop a philosophical theory of mental content that clearly explains how thoughts get the content that they do. Then we could just check to see whether computations could get content in the same way. If they can, then AI is on firm ground; if they cannot, then AI is without hope.

Much work is going into this project. For many years, there were two general strands, depending on how one views the nature of semantics. One strand investigated world–mind relations. It saw semantics as essentially associated with truth, causation, and getting along in the world. The second strand investigated mind–mind relations. It saw semantics as essentially associated with being able to draw certain inferences, construct plans, and in general determine how one thought or representation relates to another. Philosophers are now tying these two strands together in an effort to develop a complete theory of representational content. Various camps emphasize the different strands, but most agree that both strands are needed.

Neither strand seems beyond the reach of computers. Computers can be causally connected to their environments – such machines are called *robots* or *agents*, depending on whether the relevant environments are virtual or not. And computers can easily implement vast networks of representations and determine how the representations relate to one another. This strand, in particular, receives much attention in AI. So, to many AI scientists and AI-friendly philosophers, the future looks bright: on the best proto-theories of semantics they have, nothing about content seems beyond the capacities of computers.

THE FRAME PROBLEM AND HUMAN RATIONALITY

The human mind does more than merely think about things. When it thinks, the mind is governed, at least in part, by the rules of logic and inductive reasoning, and by analyses of the strength of evidence for and against certain beliefs, desires, and actions. In short, the mind is rational, at least from time to time. It is plausible that computers would be quite good at calculating evidential strength, using various logics. So, on the face of it, computers could be rational, maybe even more rational than humans. But again, closer analysis reveals serious difficulties.

Suppose a human is given a new piece of information: it is snowing outside. Such new knowledge occasions some new inferences and updating of older information, such as: 'If I go outside, I'll need my coat,' 'I will have to shovel the sidewalk in a few hours,' 'The roads will be slick when I drive, so I should drive slowly and hence leave earlier for my meeting,' etc. However, given the information that it is snowing outside, it would not occur to a human to check to make sure her phone number is still the same, or to check whether her bank account is still active. Snowing is not the sort of thing of that affects phone numbers or bank accounts. This is so obvious that it seems silly even to wonder about it. But when it comes to programming an intelligent machine, this problem is not silly; indeed it is quite serious.

When a computer is given a new piece of information – it is snowing outside – it *does* have to check everything else in its store of knowledge to see what has to be updated and what does not. In order to know that a phone number does not have to be updated given that it is snowing outside, the computer has to check the phone number in order to determine that it is not the sort of thing that has to be updated because of the snow. But what is true for snow is true in general. Hence, any time a new piece of information is input, an intelligent computer has to check everything it knows in order to find out what has to be updated. Such checking takes a lot of time because, typically, the computer's knowledge base is large. So an intelligent

computer will spend most of its time checking its knowledge base finding out that most things do not need updating.

This problem is known as the *frame problem*. It was first formulated by John McCarthy and Pat Hayes in 1969. The alleged implications of the frame problem for machine rationality were first discussed by Jerry Fodor in 1983; he revisited the issue in 2000. Since then, narrow aspects of the problem have been solved using nonmonotonic logics, but the general philosophical problem remains. We know why humans are not subject to the frame problem. Humans use the following rational heuristic to update their knowledge: given a piece of new information, update the knowledge that is *obviously relevant* to the new knowledge and leave everything else alone. This is only a heuristic because it does not guarantee that *all and only* the relevant knowledge will be updated, but as a heuristic it works quite well. The problem is that AI researchers do not have, but philosophers insist on, a universal, precise definition of 'obviously relevant.' The answer to the philosopher, however, is that AI does not need a universal, precise definition of 'obviously relevant.'

What gives the 'obviously relevant' heuristic its power in humans is that humans can alter what they consider relevant information; their definition of 'obviously relevant' can change. For example, we can learn that smoking is relevant to lung cancer, that driving is relevant to global weather changes, that computers are relevant to understanding the human mind, etc. We can see analogies such as the one between the structure of the solar system and the structure of the atom, which may suggest that the structure of the solar system is relevant to the structure of the atom. We can construct new categories 'on the fly', relevant to new situations: given that the house is on fire, we can construct the new category 'things to save from a burning house' immediately. Understanding how humans do these things requires understanding how humans do their science, how analogical thinking works, and how humans form new categories and concepts. All of these are difficult questions, but they do appear tractable. Computers, too, successfully use the 'update the information that is obviously relevant' heuristic. It is relatively easy to program simple versions of this heuristic; and, using machine learning techniques, what counts as 'obviously relevant' can change over time. AI researchers have not yet implemented a machine that is as flexible as a human in defining what counts as 'obviously relevant', but there are reasons to believe that they might.

Many philosophers want to know what 'obviously relevant' means across all cases and uses of the heuristic. They want, for example, *one metric* that tells them, for any question and any domain, which information in that domain is relevant to that question. Other philosophers insist that such a thing cannot be obtained. They point out that in-principle solutions are rare in science, and are almost completely unknown outside of basic theoretical physics. Many AI researchers now believe that 'obviously relevant' is going to have only local, pragmatic definitions that change constantly. The AI research task now is to implement the capability to pragmatically use and alter such definitions.

GÖDEL AND METAMATHEMATICS

The logical problems surrounding self-reference and the incompleteness of certain logical systems seem to some to present insurmountable problems to building an intelligent machine. In 1931, Kurt Gödel proved that every consistent, formal, logical system which includes some number theory contains true but undecidable propositions; that is, propositions that cannot be proved true nor proved false *within the given formal system*. Gödel constructed a proposition in logic, which rendered in English is: 'The proposition with the Gödel number G cannot be proven.' (This proposition, often referred to as a *Gödel sentence* for the given formal system, was actually written in first-order logic with identity, the natural numbers, and the arithmetic operations added in.) In Gödel's construction, the sentence with Gödel number G is the proposition 'The proposition with the Gödel number G cannot be proven.' When coupled with the assumption that the formal system, logic with number theory, is consistent, Gödel's theorem establishes that the formal system is necessarily incomplete: there are true propositions that cannot be proven, namely G itself.

This stunning result in logic is thought by some to present problems for AI and intelligent machines because it is assumed that computers are formal systems, no stronger than logic with number theory and hence susceptible to Gödel's results. Hence, it seems, humans know something the machine cannot know, namely that G is unprovable. Hence, humans can do something which computers cannot. To the extent that this kind of insight is crucial to intelligence, it seems that computers cannot be intelligent.

The most common reply to this argument is that it misses the role of the assumption of consistency. What we humans actually know, and what Gödel

proved, is that *if* a suitably powerful formal system is consistent, *then* it contains an undecidable proposition like G, above. And for complex, robust formal systems like logic with number theory, we cannot know if they are consistent (and their consistency cannot be proved within the system, either), so we merely assume consistency. Gödel's famous theorem is not that formal systems are incomplete; it is that if the given system is consistent, then it is incomplete. An intelligent computer can prove Gödel's theorem, it seems. All that is required is the concept of consistency (given the rest of the formal apparatus). Let T be a computer and *G* its Gödel sentence (i.e., *G* is the Gödel sentence of the formal system describing T). It is not the case that we humans know that *G* is true, but T cannot know *G* is true. We know that if the system describing T is consistent, *G* is true. But T can know this, too, provided that it has the concept of consistency. And, it seems, there is nothing inherently noncomputational about this concept.

In spite of this reply, the Gödelian whether-possible argument continues to strike many as a serious problem for AI.

COGNITIVE ARCHITECTURE AND AI METHODOLOGY

A fourth major problem for philosophers is what is the best architecture for an intelligent machine. Deciding on an architecture goes hand in hand with which methodology to use to build an intelligent machine. Currently, there are four major architectures that AI researchers and cognitive scientists are using:

1. Symbolic architectures based on high-level computational representations and algorithms. High-level representations are those that are relatively easy to translate into sentences in a natural language, such as English.
2. Connectionist architectures (sometimes called artificial neural networks) based on some level of neural organization within the brain.
3. Dynamic system architectures based not on the 'hardware' of the brain, but on the dynamical properties of neurons.These architectures are based on the theory of nonlinear dynamics.
4. Embodied robotic architectures based on layering on more and more abstract kinds of robotic control until abstract thought is achieved.

The issues surrounding these four approaches to building an intelligent machine are fascinating. Briefly they are:

- To what extent can intelligence emerge out of non-intelligent processing?
- What is the relationship between perception and cognition?
- What is the role of high-level representations in intelligent thought?
- What is the role of representations, at any level, in intelligent thought?

We can sum up the fourth philosophical problem as follows. The last three architectural approaches discussed above are closely tied to what is known about human cognitive architecture. The idea is that since humans are intelligent, perhaps AI ought to model them. But the first architectural approach is committed to the view that the details of human cognitive architecture (e.g. that we have brains) is merely an implementational detail and not relevant (at least not in theory, though it may be practically). Is this view plausible?

THE CHINESE ROOM ARGUMENT

Probably the most famous 'whether-possible' argument is the anti-AI position called the 'Chinese Room argument'. This does not fall neatly into any of the four categories discussed so far. Instead, it is an amalgam of philosophical issues concerning representational content, a system (machine or human) knowing and understanding its environment, and consciousness.

Imagine someone locked in a room. Chinese texts written on sheets of paper are slipped through a slot in the door. The person's task is to take the sheets, write further Chinese characters in response to them, and pass the responses back out through the slot. Imagine also that the person knows no Chinese at all, either written or spoken. Fortunately, the room includes a large book written in English (or the language in which the person is fluent) which stipulates what to write in response to certain Chinese characters in certain sequences. The person goes about laboriously reading and following the rules in the book, completely ignorant of what the characters mean. Unbeknownst to the person in the room, outside are several Chinese scholars discussing, say, Chinese history. Because of the apparent insightfulness of the comments coming out through the slot, they believe that the person in the room is an expert on China's history.

It has been argued that the person in the room is precisely analogous to a computer, engaging in computations by following the rules. Since there is no difference in this respect between the person in the room and a computer, and since the person knows nothing about Chinese history, it must be the case that no computer can know anything about Chinese history, either. Since there is nothing

special about Chinese history in this situation, it must be the case that computers in general can know nothing.

The replies to this argument would fill a library. One of the most interesting concedes that the person in the room knows no Chinese history, but claims that the *virtual machine* consisting of the person together with the book of rules (and perhaps the interlocutors outside the room) constitutes a genuine understander or knower. Other replies introduce robotic connections to the external environment, or try to pick apart knowing (or understanding), representational semantics, and consciousness. The current consensus is that the Chinese room argument is flawed, but beyond that there is little agreement as to why.

CONSCIOUSNESS AND MORAL ISSUES IN AI

Beyond the whether-possible problems, there are moral issues surrounding the project to build an intelligent machine. Is this something we should be doing at all? For example, computers are not alive, so currently there is nothing morally wrong with tossing one out of an airplane. But would this still be true of an intelligent computer? To answer this, we would have to know whether an intelligent computer would have emotions, dreams, hopes, and plans. Consciousness is crucial to any notion of morality. Would an intelligent computer be conscious? How could we tell? Would an intelligent computer know and care what happened to it? An embodied robotic architecture might well know and care, in a very real way. So would it not be immoral to push such a robot off a cliff on a whim?

Further Reading

Aiken H (1956) The future of automatic computing machinery. In: *Elektronische Rechenmaschinen und Informationsverarbeitung*, pp. 32–34. Braunschweig: Vieweg. [Proceedings of a symposium published in *Nachrichtentechnische Fachberichte* 4.]

Church A (1936) A note on the entscheidungsproblem. *Journal of Symbolic Logic* **1**: 40–41, 101–102.

Crevier D (1993) *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York, NY: Basic Books.

Dietrich E (ed.) (1994) *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines*. San Diego, CA: Academic Press.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor J (1987) Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In: Pylyshyn Z (ed.) *The Robot's Dilemma*, pp. 139–149. Norwood, NJ: Ablex.

Fodor J (2000) *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Gödel K (1931) On formally undecidable propositions of *Principia Mathematica* and related systems I. In: van Heijenoort J (ed.) (1967) *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, pp. 596–616. Cambridge, MA: Harvard University Press.

McCarthy J and Hayes P (1969) Some philosophical problems considered from the standpoint of artificial intelligence. In: Metzer B and Michie D (eds) *Machine Intelligence*, vol. IV, pp. 463–502. New York, NY: Elsevier.

McCorduck P (1979) *Machines Who Think*. San Francisco, CA: WH Freeman.

Morgenstern L (1996) The problem with solutions to the frame problem. In: Ford K and Pylyshyn Z (eds) *The Robot's Dilemma Revisited*. Norwood, NJ: Ablex.

Partridge D and Wilks Y (eds) *The Foundations of Artificial Intelligence: A Source Book*. Cambridge, UK: Cambridge University Press.

Penrose R (1994) *Shadows of the Mind*. New York, NY: Oxford University Press.

Putnam H (1960) Minds and machines. In: Hook S (ed.) *Dimensions of Mind: A Symposium*, pp. 138–164. New York, NY: New York University Press.

Searle JR (1980) Minds, brains, and programs. *The Behavioral and Brain Sciences* **3**.

Turing A (1936) On computable numbers with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, **42**: 230–265 and **43**: 544–546.

Behaviorism, Philosophical

Intermediate article

Max Hocutt, University of Alabama, Tuscaloosa, Alabama, USA

CONTENTS

Philosophical behaviorism, the belief that states and traits of mind are behavioral dispositions, was the dominant philosophy of mind for much of the twentieth century.

INTRODUCTION

Reduced to a slogan, behaviorism is the belief that states and traits of mind are behavioral dispositions. Behaviorism in psychology began with the recommendation that introspective analysis of consciousness be replaced by systematic observation of behavior. Behaviorism in philosophy was the attempt to formulate a theory of mind and personality that would justify this new methodology. According to philosophical behaviorism, thoughts and feelings are not private states of invisible but introspectible minds; they are observable dispositions of visible animals. Thus, fear is a disposition to flee; anger a disposition to destroy; desire a disposition to prefer; belief a disposition to assent; and so on.

HISTORY

Although it had roots in the seventeenth-century empiricism of Thomas Hobbes and the nineteenth-century pragmatism of Charles Peirce, behaviorism as an explicit doctrine began in the first quarter of the twentieth century at the University of Chicago with J. B. Watson, whose colleagues James Angell and John Dewey advocated a kindred idea which they called functionalism. Watson criticized the dependence of psychologists on introspection because its claims could not be independently confirmed (Carnap, 1931; Watson, 1913). As the philosopher of science Arthur Pap would later explain:

> Scientific propositions must be intersubjectively verifiable; the verdicts of introspection, however, are not intersubjectively testable; therefore a science based on introspection is not really a science. (Pap, 1962)

Watson concluded that psychology should cease being a first-person study of private consciousness and become a third-person study of public behavior.

An early formulation of behaviorist philosophy was by the logical positivist Rudolph Carnap, who argued that, if consciousness cannot be studied scientifically, the reason must be that minds and their contents are unreal: the reality is bodies and their behavior. Carnap concluded, to quote Pap again, that 'there is but behavior, dispositions to respond in specific ways to specific stimuli, and neurophysiological processes within the human and animal body'. 'Behaviorism' was the name of the first half of this conjunction; Carnap named the second half 'physicalism'.

To Carnap, physicalism meant the belief that everything is describable in the language of physics. An advocate of the unity of science, Carnap thought that psychology could be reduced to physics, in two stages. First, statements about the mind would be translated into statements about behavior; then these would be translated in their turn into statements about the motions of bodies in space. Thus, mentalistic statements such as 'Sam feels hungry' would give way to behaviorist remarks such as 'Sam is disposed to eat'; then these remarks would themselves be reformulated using the more austere terminology of the physicist.

This programme of analysis – which has been called logical, or analytic, behaviorism – encountered difficulties at the very first step. A glance at our behaviorist analysis of hunger will show why. 'Mary feels hungry' is clearly not synonymous with 'Mary is disposed to eat'. The two statements do not even have the same truth conditions. Mary might feel hungry without being disposed to eat, perhaps because she is fasting; or be disposed to eat without feeling hungry, perhaps because she wants to gain weight.

The usual behaviorist reply to this problem was that it demonstrates not the inadequacy of

behavioral analysis as such but only the inadequacy of a particular analysis: in this case, the analysis of hunger, which could be improved by defining hunger as a disposition to eat that usually results from deprivation of food and will usually manifest itself in eating unless there is some stronger disposition to the contrary. In short, a hungry person is one who has not eaten recently and is, as a result, disposed to eat, other things being equal. Since *ceteris paribus* clauses such as this are common in the physical sciences, behaviorists saw no reason why they should not exist in psychology too; but critics protested that the resulting analyses were illegitimately ad hoc. Worse, they sounded tautological, like saying that people who are disposed to eat are disposed to eat unless they are not. Because the logical behaviorists were seeking analyses that are true by definition, they were not greatly perturbed by this complaint, but it seemed to their critics to be damning.

Perhaps seeking to deflect this criticism, Ludwig Wittgenstein formulated behaviorism more cautiously, by claiming that inner states require outward criteria (Wittgenstein, 1953). What Wittgenstein meant by this was that differences in states of mind must manifest themselves in some way, if not always in a particular way; the mind cannot be a fifth wheel, which spins idly, making no difference to the progress of the vehicle. To see how this Wittgensteinian observation applies to a particular case, consider pain, which a logical behaviorist might define as a disposition to moan and groan. Hilary Putnam's observation that there might be Spartans who grit their teeth and grimace but do not moan may refute Carnap's logical behaviorism, but it does not refute Wittgenstein's criterial behaviorism. To refute criterial behaviorism, one would have to show that there can be a state of mind that manifests itself in no observable response, overt or covert; and it is not obvious how this could be so.

After Wittgenstein had made this point in Cambridge, another blow for behaviorism was struck in Oxford by Gilbert Ryle. Influenced by Wittgenstein, Ryle mounted an assault on behaviorism's *bête noire*: Cartesian dualism, the belief that a human being is a mind, which thinks, within a body, which takes up room. Ryle (1949) argued that this Cartesian belief is the result of a 'category mistake'. What Ryle had in mind can be brought out by using an analogy. Suppose that Jones is a fast runner. Then she may be said to 'have speed'. It would, however, be an error to ask: 'Where does Jones keep her speed?' This would mistake an attribute for a thing; a predicate adjective for a noun. A runner's speed is not one of her possessions; it is one of her characteristics. It is not a thing that she owns; it is a way that she is – namely, able to run fast.

Ryle had the same view of statements about the mind. To say that visible people have minds does not mean that they possess invisible and intangible things that do their thinking and feeling for them. It means that they have visible capacities for visible feeling and thought. Thus, saying that Smith has great intelligence means not that Smith possesses a large quantity of some mysterious stuff, but that Smith can usually be observed to behave in ways that are well adapted to his ends. As Watson's colleague Dewey had put the point: mind is not substantival; it is adjectival. In other words, it is not our minds that do our thinking. It is not even we who do our thinking with our minds. Rather, it is simply we – i.e., we bodily beings – who do our thinking.

Most behaviorists agreed; but there was a problem. Could not someone think silently, without anybody else being able to discover the fact just by watching her? Indeed, would it not make sense to suppose that intelligence never manifested itself in visible behavior? Faced with these questions, behaviorists divided into two groups. Some, reverting to Carnap's physicalism, took the position that mental abilities are rooted not in an invisible mind but in visible brains, glands and muscles. Thus, Ullin Place and Jack Smart held that sensations are events in the brain (Place, 1956; Smart, 1959). As such, they are observable in principle, if not always in practice (because they take place under the skin, which is opaque).

Ryle avoided this conciliatory line because it seemed to him to entail too mechanistic a view of human beings. An Aristotelian, Ryle did not share the physicalist hope of reducing psychology, which concerns purposive action, to physics, which recognizes only mechanical motion. Unaware that B. F. Skinner had undertaken to explain purposive behavior mechanistically by invoking Thorndike's 'law of effect' – which says that the likelihood of future behavior is modified by the effects of past behavior of the same sort – Ryle protested that human beings are animals, not machines. Noting that the physiologist may have an explanation for the patellar tendon reflex, Ryle doubted that he could explain the act of kicking a football. Here Ryle disagreed with Carnap.

In defence of his scepticism, Ryle argued that statements about psychological states and traits are 'mongrel categoricals': conditional statements in categorical disguise. To see what he meant, consider again the statement 'Mary is hungry'. This

resembles 'Mary is short', but if Ryle was right, the similarity of grammar is misleading. In Ryle's view, there is nothing 'iffy' about the statement that Mary is short; but, analysed behaviorally, the statement that she is hungry means that she will eat if food is available. Believing that conditional statements could not be reduced to categorical statements, Ryle did not believe that psychology could be reduced to physiology.

Few philosophers have been convinced by this argument. To see why not, consider the statement: 'The bar is magnetic.' It implies that the bar will attract iron filings if any are present, which is a conditional remark. But this is true only because the constituent molecules of magnets are arranged in dipole, a fact that is in no way 'iffy'. In this case, the conditional fact presupposes the categorical fact; it does not rule it out. Similarly, although '*x* is hungry' implies the conditional statement '*x* is likely to eat if offered food', it also implies the categorical statement '*x*'s blood sugar level is low'. In opposing the conditional to the categorical, Ryle had posed a false dichotomy.

Later, Daniel Dennett, a disciple of Ryle, proposed a variant on behaviorism that would correct Ryle's error. Dennett argued that states of mind are best regarded as states of the body having certain typical causes and effects (Dennett, 1978). Thus, hunger is that state of the body that usually results from food deprivation and usually gives rise to eating; pain is that state of the body that is usually caused by tissue damage and usually elicits moaning; and so on. This showed how categorical facts about the body might be not only connected with but even equivalent to conditional facts about behavior.

Hilary Putnam proposed a somewhat different view that he called functionalism. According to this view, functionally equivalent states of mind might have both different physical embodiments and different behavioral manifestations (Putnam, 1978). Thus, suppose that a Martian and I both feel pain. Because the Martian will have a different nervous system, his pain will take a different form in his brain; and because he is brave while I am a coward, he will suffer in silence while I moan and groan. This possibility – the multiple realizability of mental states – had apparently been suggested to Putnam by the fact that two computers might be constructed with different chips, use different operating systems, and display their result in different ways, yet be performing the very same operation – say, adding two and two.

By taking this line, Putnam changed behaviorism, which had grown out of one kind of functionalism, into a different kind of functionalism, thereby allowing cognitive science – the study of the mental workings of the physical brain – to mature.

ARGUMENTS FOR PHILOSOPHICAL BEHAVIORISM

As noted above, the distinctive feature of behaviorism was its resolve to make psychology scientific by confining it to observable responses and their observable causes. This resolve, which required the psychologist to abstain from discussing what could not be observed, worked well for several decades, turning the analysis of behavior into a flourishing science with useful applications in psychotherapy, advertising, education, and industry. The variety and utility of this technology is evidence of the great value of behavioral psychology and the legitimacy of the philosophy behind it.

Nevertheless, many people believe that behaviorism has been decisively refuted by, among others, the linguist Noam Chomsky in his review (1959) of B. F. Skinner's *Verbal Behavior*. Challenging Skinner's account of speech as reinforced operant, Chomsky argued that the rules of grammar, which are innate, give us the capacity to formulate an unlimited number and variety of sentences. Chomsky's criticism is valid if the rules of grammar are innately known and if Skinner's theory implies that our verbal facility consists of a limited and predetermined repertoire; but the first of these claims has not been proved and the second is not obviously true. Besides, even if Chomsky is right, it is one thing to refute Skinner, another to refute behaviorism, which is not committed to Skinner's analysis of language.

A second criticism of behaviorism, one emphasized most recently by John Searle but stated earlier by Roderick Chisholm, is that it gives no satisfactory account of the mind's intentionality: the directedness of desires towards ends and thoughts towards objects. This criticism is accurate, but in fact intentionality is an unsolved puzzle for every philosophy, behaviorist and nonbehaviorist alike. Thus, Searle declares it a biological given. The case may be comparable to that which once existed in biology. Intentionality is closely related to purpose. Until Darwin showed how natural selection could mimic deliberate design, there was no explanation of the teleology, or purposiveness, that seems to be present in nature.

Recently, behaviorism has been the object of two further charges. Psychologists have charged that it discourages the scientific study of the brain; and

philosophers have charged that it discourages the introspective study of consciousness.

To begin with the first charge: emphasis on studying overt behavior has indeed distracted behaviorists' attention from the brain, the workings of which are observable in principle but not, until recently, easily observed in practice. Thus, the psychologist B. F. Skinner actively discouraged speculation about the brain. However, not even Skinner doubted that understanding behavior would eventually require understanding its biological, including its neurophysiological, underpinnings. Watson emphasized biology from the beginning, and, as noted earlier in this article, Carnap endorsed Watson's approach. Among important behaviorist philosophers, only Ryle can be accused of believing that biology is irrelevant, and his error was soon corrected by Dennett. In short, the first charge is aimed at a straw man.

In fact, so is the second: the charge that, by disallowing introspection, behaviorism entails neglect of consciousness. The problem with this claim is that talk of introspection is ambiguous. No behaviorist has ever denied the uses of introspection if by this is meant the familiar practice of discerning one's feelings, sizing up one's motives, giving voice to one's opinions, noticing what one perceives, or observing one's conduct and appraising one's character. What behaviorists have emphatically denied is that introspection yields knowledge that is either privileged or infallible. The behaviorist acknowledges that you can know whether you are angry, but he insists that other people can know it too – by observing what you say and do. Furthermore, against the Cartesian he adds that you can also fail to know that you are angry – for example, by being so angry that you do not pause to recognize the fact.

Now consider consciousness. Again there is ambiguity. No behaviorist has ever denied that there is a difference between someone who is awake and someone who is asleep, or between someone who is alert and someone who is drugged. Nor, to come back to introspection again, has any behaviorist denied the reality of self-consciousness, if that means being aware of one's own mental condition. What behaviorists have denied are the claims made for consciousness by followers of Descartes, who held that consciousness has features that can be known only by means of introspection and cannot fail to be known by it.

Followers of Descartes are legion. The commonplace view that nobody can know what another person feels is supported both by Thomas Nagel (1974), who argues that no one can know 'what it is like to be a bat', and by Frank Jackson, who argues that there is something which a color-blind physicist could not know – namely, what colors look like.

Behaviorists believe that these Cartesians have been victimized by an ambiguity in the concept of knowledge. Grant that I cannot have, or share, your (or a bat's) feelings. It does not follow that I cannot recognize or identify them. On the contrary, if the behaviorist is right, your feelings (or the bat's) can be known in two ways: by studying your (or its) behavior and by examining your (or its) nervous system, glands, and musculature. As to the color-blind physicist Mary, it is true that she lacks knowledge that other people have, but if the functionalist David Lewis (1980) is right, her problem is not that she is ignorant of some facts; it is that she cannot make certain discriminations, which is a behavioral disability.

Cartesians often buttress their position by arguing that, since only I can know my thoughts and feelings from the inside, only I can know their phenomenal properties. To see what this metaphor and this piece of philosophical jargon mean, suppose that Smith and Jones both look at a red fire engine. David Chalmers and Galen Strawson maintain that each of them will know what sensations he is having but that neither can know what the other's sensations are like, because although every person can look into his or her own mind, no person can look into the mind of another. Therefore, even if Smith and Jones agree that they are both looking at a red fire engine, they may not be having the same sensation. Let us call this the argument from introspection.

U. T. Place said that this argument is based on the phenomenological fallacy: treating the appearances of things as if they were themselves things that appear, only to a limited audience of one person. The idea seems to be that each of us has a personal moving picture screen, which nobody else can view. Objects in the world project pictures onto that screen for the viewer's private edification. Each viewer knows what his or her own pictures look like, but can have no idea whether the pictures on other screens resemble them.

The fallacy lies in the belief that people see not things but their appearances. What we call the thing's appearance is not the thing we see; it is how we see it. To quote Wittgenstein, it is 'what we see the thing as'. Thus, Smith and Jones see a red fire engine; they do not see its appearances. Seeing something requires you to have your eyeballs aimed at it, and by hypothesis, Smith and Jones are aiming their eyeballs at the red fire engine, which is in front of their eyes, not at their

sensations, which, being events in their brains, are behind their eyes.

What about the redness of the sensation of red and the painfulness of the sensation of pain? There are no such things. It is the fire engine, not the sensation of red, that is red; it is the hot poker, not the sensation of pain, that is painful.

To say so is not to deny that the fire engine might present different appearances to different people. For example, it might look blurry to Smith, who has astigmatism, and small to Jones, who is far away. To explain this, however, we do not need to suppose that Smith and Jones are seeing their private sensations rather than public objects. On the contrary, if sensations are events in their brains, their intrinsic nature and character will be better known to the neurophysiologist than to Smith and Jones. It is only if sensations are events occurring in invisible and intangible minds that they become inscrutable, and in that case we have good reason to regard them as unreal. What is real can be detected by more than one person, in more than one way.

PHILOSOPHICAL BEHAVIORISM AND COGNITIVE SCIENCE

Confusion about this may have been augmented in recent years by an idea that has become the basis of much cognitive science: the comparison of the human mind to an electronic computer. According to the usual accounts, coded information is input by peripherals such as a keyboard through a cable to the central processing unit (CPU). This message, a representation of something in the external world, exists in the CPU as a pattern of electrical impulses and, after interpretation by the computer's program, is converted into appropriate outputs to peripherals such as a printer, or a robot welding a plate to an automobile.

As Jerry Fodor has noted, this way of describing the workings of a computer closely resembles the Cartesian account of the mind. According to this account, our sense organs send physically coded messages to the brain, where our mind reads and interprets them as appearances of things external. Having done this, the mind then orders the brain to activate the appropriate muscles or glands in order to produce suitable behavior. The mind is thus a calculating machine.

The behaviorist objection to this account is not that it seeks to understand the brain by comparing it to a computer, but that it seeks to understand the computer by comparing it to a mind. We do this when, speaking metaphorically about neurophysiological processes that are still largely unknown to us, we describe the brain as an 'interpreter' of 'information'. Having set out to mechanize man, we end up anthropomorphizing the machine, thus coming full circle back to Cartesianism, the very outcome Skinner warned against. Dennett aptly calls this Cartesian materialism.

What is needed, if Dennett is right, is not to reject the computational theory of the mind but to purify it of Place's internal cinema screen. As Place himself liked to point out, this can be done by regarding the brain as a connectionist machine, or parallel distributed processor, of the sort described by Rummelhart and McClellan. Instead of having what might be thought of as internal representations, these machines convert inputs (stimuli) to outputs (responses) in accordance with connection weights that are gradually altered using an algorithm designed to respond to feedback by eliminating unsuccessful responses in favor of successful ones – providing a working model of the workings of the behavioral law of effect.

The behaviorist likes this approach, for three main reasons. First, it has been purified of those anthropomorphic notions that the behaviorist finds objectionable in taking literally the idea of the computer as a mind engaged in processing information. Second, it is in better accord with our present knowledge of the brain, which is almost certainly a parallel processor. Third, and most importantly, the connectionist understanding of cognition as adaptation to external stimuli retains the externalist emphasis that is missing from Cartesianism, which is essentially solipsistic. Let us close by saying something about this third point.

Behaviorism came into being shortly after the publication of Darwin's theory of biological evolution, and was inspired by it. According to this theory, nature weeds out organisms that are ill-adapted to cope with their environments. This implies that cognition is best regarded not just as something going on in our heads but also as something that enables us to cope with our surroundings. So what matters in cognition is that internal processing manifests itself in behavior that is appropriate to the external world. What Wallace Matson calls an 'outside-in' philosophy sits better with this than the 'inside-out' philosophy of Cartesianism, and behaviorism is the pre-eminent outside-in philosophy.

References

Carnap R (1931) Psychology in the language of physics. In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 43–79. London: J. M. Dent.

Chomsky N (1959) A review of B. F. Skinner's *Verbal Behavior*. In: Geirsson H and Losonsky M (eds) *Readings in Language and Mind*, pp. 413–441. Oxford: Blackwell.
Dennett D (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.
Lewis D (1980) Mad pain and Martian pain, and knowing what it's like. In: Rosenthal D (ed.) *The Nature of Mind*, pp 229–235. New York, NY: Oxford University Press.
Nagel T (1974) What is it like to be a bat? In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 159–174. London: J. M. Dent.
Pap A (1962) Mind and behaviorism. In: *Introduction to the Philosophy of Science*, pp. 374–409. New York, NY: The Free Press of Glencoe.
Place UT (1956) Is consciousness a brain process? In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 106–116. London: J. M. Dent.
Putnam H (1973) Philosophy and our mental life. In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 133–147. London: J. M. Dent.
Ryle G (1949) *The Concept of Mind*. London: Hutchinson and Company.
Smart JJC (1995) Sensations and brain processes. In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 117–132. London: J. M. Dent.
Watson JB (1931) Psychology as the behaviorist views it. In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 24–42. London: J. M. Dent.
Wittgenstein L (1953) *Philosophical Investigations*. Oxford: Blackwell.

Further Reading

Armstrong DM (1963) Is introspective knowledge incorrigible? *Philosophical Review* **72**(4): 417–432.
Hocutt MO (1996) Behaviorism as opposition to Cartesianism. In: O'Dononue and Kitchner (eds) *Psychology and Philosophy: Interdisciplinary Problems and Responses*, pp. 81–95. London: Sage Press.
Quine WV (1985) States of mind. *Journal of Philosophy* **82**(1): 5–8.
Zuriff GE (1985) *Behaviorism: A Conceptual Reconstruction*. New York, NY: Columbia University Press.

Blindsight

Introductory article

Robert W Kentridge, University of Durham, Tyneside, UK

CONTENTS

Patients with damage to primary visual cortex or its afferents report that they are blind in the area of the visual field corresponding to this damage. Blindsight refers to the ability demonstrated by some of these patients to perform a variety of visual tasks despite denying awareness of the stimuli to which they are responding – a dissociation between performance and consciousness.

WHAT IS BLINDSIGHT?

Blindsight is the term given to the remarkable abilities found in a small number of neurological patients who have damage affecting striate cortex, the first cortical area of the brain which normally processes visual information. Despite its rarity, the condition has profound implications for our understanding of consciousness. As a consequence of its rarity and the importance of its implications it is a condition surrounded by controversy. (*See* **Blindsight, Neural Basis of**)

As a result of their brain damage patients with blindsight deny being aware of visual stimuli in the area corresponding to their damage. For example, a patient with damage to the left side of striate cortex reports that he cannot see stimuli presented to the right of his direction of gaze. When tested using standard procedures these patients are classified as clinically blind in the area corresponding to their damage (that is, they have a scotoma). However, if the patients are tested in a way which forces them to make decisions about stimuli presented in their scotoma then, even though the patients deny seeing anything and maintain that their decisions are simply guesses, they usually make the correct response to the unseen stimuli on a variety of visual tasks.

Blindsight, then, is the dissociation between awareness of visual stimuli and the ability to respond appropriately to them found in patients with damage to striate cortex or the neural connections leading directly to it. It is clear that blindsight subjects can detect whether a spot of light within their scotoma accompanies an auditory signal, whereabouts it is, and, if it is moving, in which direction and how fast it is going. The evidence for more complex residual abilities is less strong.

HISTORY

Striate cortex gets its name from a fine white line identifiable near its surface in slices of the brain. This 'stripe of Gennari', discovered in 1782, was the first evidence that the anatomy of the cortex was not uniform and hence that different areas of cortex may be specialized to serve particular functions. Striate cortex lies at the occipital pole of the brain; in humans much of it is hidden on the adjoining lateral surfaces of the cerebral hemispheres (Figure 1).

In addition to being the first identified anatomically specialized cortical area, it was also the focus of the earliest work on functional specialization. Observations of stroke patients dating back to the 1850s suggested that damage to the brain's occipital pole had specific effects on vision. Towards the end of the nineteenth century, experiments on monkeys showed that lesions of occipital pole large enough to include all of striate cortex rendered animals blind, and it was generally agreed that the occipital lobes were indispensable for vision. From the mid-1930s, however, it became apparent that animals with lesions restricted to striate cortex and not impinging upon other parts of the occipital lobes retained some visual abilities – they could be conditioned to respond to flashes of light and could follow moving spots of light with their eyes.

Starting in the mid-1960s, Nicholas Humphrey studied a single monkey, named Helen, who had bilateral striate cortex lesions. On the basis of many years of observation, Humphrey concluded that Helen retained many (but not all) visual abilities, despite her lesion. For example, she would

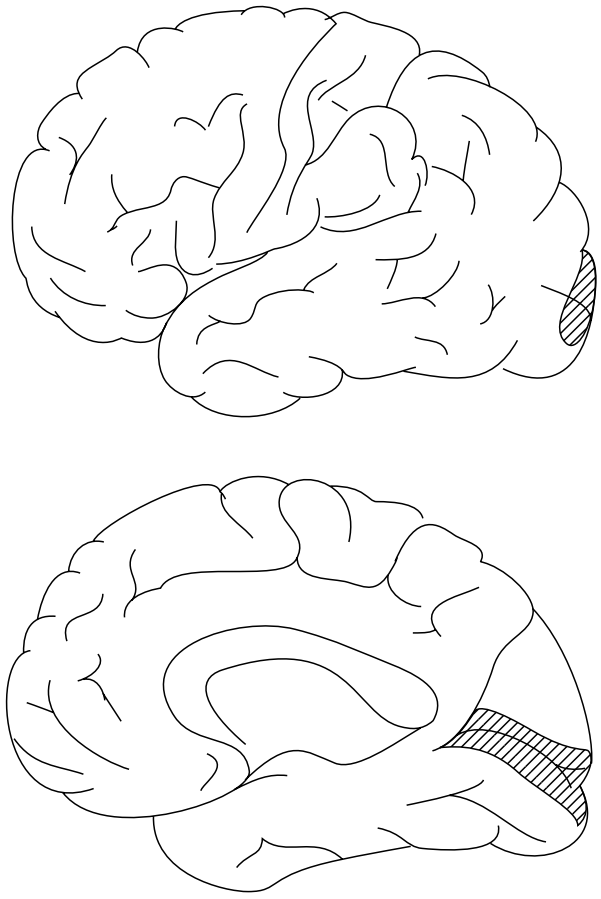


Figure 1. Lateral (upper panel) and medial (lower panel) views of the human cerebral cortex showing primary visual cortex (hatched). Note how little of primary visual cortex is exposed on the surface of the brain. Most of primary visual cortex lies on the medial surface of the brain and is therefore hidden between the two cerebral hemispheres.

routinely pick up very small objects with great precision; however, it was clear that she could not identify what these objects were until she explored them with her mouth. Helen apparently retained the ability to detect and locate visual stimuli despite her lesion, but she could no longer identify them.

Although the animal studies of Humphrey allowed the visual abilities remaining after striate cortex lesions to be identified, they could not provide any insight into the subjective nature of visual experience without striate cortex. To do so one must be able to ask a human patient lacking striate cortex to describe what they see. Such patients had been studied for many years and reported that they saw nothing in the region corresponding to their brain damage. One exception, to which we will return later, was the perception of movement. During the First World War, George Riddoch found that soldiers with injuries to the occipital cortex, although blind to stationary stimuli, reported that vigorously moving stimuli did elicit visual experience. Studies of wounded soldiers feature prominently in the history of visual neuropsychology. Careful collation of the locations of gunshot wounds and areas of lost vision in soldiers from both World Wars provided the evidence for maps of the representation of the visual field in striate cortex.

In 1973 the team of Ernst Pöppel, Richard Held, and Douglas Frost, working at Massachusetts Institute of Technology, decided to test whether soldiers (and one stroke patient) with visual scotomata as a result of damage to the visual cortex could, nevertheless, move their eyes so as to direct their gaze at spots of light presented in their regions of blindness. Pöppel, Held, and Frost were prompted to attempt this experiment by earlier work which, amongst other things, had shown intact responses of the pupil and intact optokinetic nystagmus (a slow drift of eye-gaze in one direction, interrupted by occasional flicks back in the opposite direction, induced by presentation of a continually moving pattern) in patients with occipital lesions. Since both of these responses are mediated by midbrain structures, it might be the case that neural pathways transmitting information directly from the retina to the midbrain without passing through striate cortex could support a range of simple visual abilities in these patients. As at least one circuit used in the control of eye movements is entirely subcortical, eye-movement control was a clear candidate for such a potentially spared function. Although the patients found the task puzzling, one remarking 'how can I look at something I haven't seen', there was a consistent relationship between the location of visual targets and the eye movements the patients produced when asked to look at the locations where they 'guessed' these targets had been presented. The appropriate behavioral response of patients to visual targets shown in this task, coupled with their complete denial of awareness of those targets, is acknowledged as the first systematic experimental demonstration of blindsight.

The term was not, however, coined until a year later when Lawrence Weiskrantz described similar work he had carried out on a patient who, as a result of surgery to alleviate pain caused by abnormalities in the blood supply to the occipital pole of the brain, had lost most of the striate cortex on one side of his brain. Weiskrantz found that not only

did this patient (known as DB) move his eyes appropriately towards unseen targets, but he could also point towards target locations accurately with his finger, detect the presence of a luminance grating (a smoothly varying pattern of light and dark stripes), discriminate the orientation of lines and discriminate between the shapes 'X' and 'O' in his blind field, all while denying any visual experience. Blindsight was clearly a complex phenomenon requiring considerable work, both to evaluate the range of visual functions spared after damage to striate cortex, to determine the extent of the dissociation between behavior and visual consciousness, and to test models of the anatomical basis of residual function.

EXPERIMENTAL WORK ON BLINDSIGHT

Four questions need to be addressed in the experimental study of blindsight. Apart from evaluating the anatomical basis of blindsight, the range of spared functions, and the dissociation between behavior and awareness, it is crucial to demonstrate that blindsight is a real phenomenon and that the results obtained cannot be explained by experimental artefacts which allow subjects to perform tasks using the intact portion of their visual field or in some other unintended manner.

Artefacts

Blindsight patients are quite rare. Moreover, in virtually all reported cases, visual field loss is not total. These patients therefore retain normal conscious vision in part of their visual field. The residual visual abilities of interest in blindsight are those used in response to stimuli presented in the blind portion of the visual field. If, however, visual targets presented to a patient's scotoma also illuminate their intact visual field, then any response they make is not truly indicative of blindsight. Light from a target presented within the scotoma may reach intact areas of the visual field as it is scattered from objects in the room where testing is being conducted or as it is scattered by the internal structures of the eye.

The first of these potential artefacts is relatively easy to detect and control, the second much harder. One approach that has been taken is to use the area of visual field within the scotoma corresponding to the blind-spot in a control condition. The blind-spot is the small area of retina where photoreceptors are absent as nerve fibers from receptors throughout the rest of the eye converge to leave the eye as the optic nerve. A target presented exactly within the blind-spot could not therefore directly activate any pathway, cortical or subcortical. One would therefore expect that the subject's ability to respond appropriately to a target will be eliminated if the target is presented in the blind-spot, whether the subject has blindsight or has an undamaged cortex. If, however, the subject's response to a target depends upon light scattered to remote (and intact) portions of the visual field, it should not matter whether the target is presented over the blind-spot or an adjoining area of retina – the presence of receptors at the target location is neither here nor there. The performance of blindsight subjects does indeed fall to chance when targets are presented to the blind-spot, suggesting that residual performance in blindsight does not depend upon a scattered light artefact. This does not, however, mean that scattered light can be ignored. It may still provide cues to a subject unless steps are taken to control it. The most common of these is to use dark targets against a bright background wherever possible, and to flood the subject's intact visual field with bright light.

Light-scatter is not the only means by which information from stimuli intended to reach the scotoma alone can travel to intact regions of the visual field. The most common method of presenting stimuli to patients is with a computer display screen. Stimuli presented in one part of a computer display can produce unintended but visible effects in other parts of the display. Presentation of a bright spot, for example, can cause a small brightening in a narrow horizontal band at the same height as the spot across the entire width of the screen. Care must, therefore, be taken to mask portions of the screen visible outside a patient's scotoma when using such stimuli.

Anatomical Bases of Blindsight

The processing of visual stimuli starts in the array of interconnected photoreceptors of the retina at the back of the eye (Figure 2). The most prominent output from the retina projects to a midbrain structure called the lateral geniculate nucleus (LGN) and from there to striate cortex. This is not, however, the only output from the retina which projects to many other structures. Initially it was supposed that blindsight was mediated by such structures which controlled basic responses to light without any cortical involvement. For example, the superior colliculus can control reflexive eye movements which direct gaze towards a visual target without involving cortex. Although subcortical circuits

mediate very specialized responses, blindsight patients might learn to monitor these specialized responses in the course of performing more general tasks. It might, for example, be possible to monitor the location towards which one is about to move one's eyes and use this information to choose whether or not to press a button even if the eye movement itself is suppressed. According to this scenario, blindsight may be mediated by subcortical visual pathways.

Although the bulk of visual input to the cortex passes through the striate cortex, there are ways in which visual information can reach the cortex while bypassing the geniculo-striate route. The superior colliculus sends projections, via the pulvinar, to a number of cortical areas involved in vision (V2, V3, V4, and MT). These are parts of cortex involved in visual processing which normally receive their major input via the striate cortex. Since these areas can receive visual input in the absence of the geniculo-striate projection, it is possible that blindsight may be mediated by visual pathways outside the striate cortex.

In addition to mediation by subcortical or extrastriate cortical routes, there remains the possibility that damage to striate cortex in blindsight patients is not, in fact, complete. Rather than demonstrating that circuits other than the major geniculo-striate route support visual function but do not give rise to visual awareness, residual visual function in blindsight would then essentially be a demonstration that the magnitude of stimulation required to evoke awareness from the geniculo-striate system is greater than that required to support simple behavioral responses. Blindsight would not, under these circumstances, be a particularly special phenomenon since it would differ little from the abilities of normal subjects when presented with stimuli near the limits of their visual abilities (e.g., very faint or very short duration stimuli).

A number of studies have been made of patients who have had small spared regions of striate cortex surrounded by damage. These patients did not experience stimuli falling in these spared regions, yet, as in blindsight, they could perform simple visual discriminations in these regions. Can such spared cortex explain the apparently extensive region of

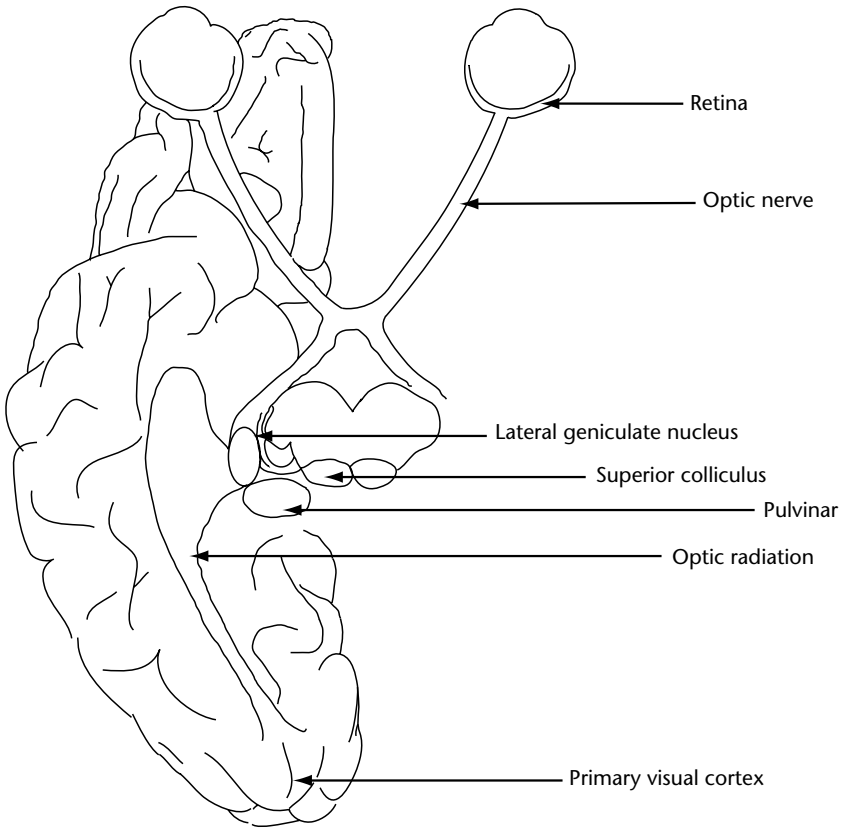


Figure 2. A basal view of the human brain showing major components of the visual system which may be involved in the mediation of blindsight.

blindsight found in other blindsight subjects? Patches of residual vision surrounded by areas of complete blindness might not be revealed in most studies if random eye movements fortuitously brought stimuli into a region of the retina which activated a patch of spared cortex. If, however, one ensures that eye movements cannot affect the location in the cortex which a visual stimulus potentially activates, then any patchiness should become apparent. It is possible to do this by using eye-movement measurements to yoke stimulus position to the direction of gaze. A study using this technique in a patient with blindsight covering a large proportion of one visual field did not reveal patches of residual vision surrounded by blindness, suggesting that an explanation of all blindsight in terms of islands of spared cortex is untenable.

Diffuse, as opposed to patchy, subtotal damage is harder to detect behaviorally. The undamaged neurons in a diffusely damaged region of cortex should, however, still be metabolically active. Functional neuroimaging, which detects changes in blood flow or blood oxygen levels indicative of metabolic activity, has not revealed activity in the striate cortex of blindsight patients when a visual stimulus was presented, although changes did occur in extrastriate cortex.

If blindsight relies on a visual pathway used in normal vision, albeit seriously damaged, the implication is that blindsight should be like very poor normal vision. The apparent dissociation between the abilities of blindsight patients and their reports of awareness may be explained in terms of a change in their willingness to report that they have seen a stimulus – not a surprising change given their knowledge that they have a serious visual impairment. It is, however, possible to disentangle the effects of such biases from the underlying visual sensitivity. The results of such experiments indicate that, for normal subjects presented with stimuli near the limits of visual ability, there is no difference between mechanisms which serve conscious report and those which serve the 'forced-choice' discrimination tasks typically used in assessing blindsight. A similar comparison in a blindsight patient showed quite different properties for conscious report and forced-choice discrimination, indicating behaviorally that blindsight is not simply near-threshold normal vision.

Ingenious experiments have been devised which show that monkeys with unilateral visual cortex lesions treat stimuli in their 'blind' and normal visual fields quite differently, even though they are quite capable of making behavioral responses to those blind-field stimuli. The monkeys were first trained to point at visual targets presented in either their blind or normal hemifields and the minimum brightness contrast required was measured for each hemifield. The target contrasts were then adjusted so that they easily exceeded these thresholds for the rest of the experiment. The monkeys now learned a new task in which they had to make different responses depending on whether one or two stimuli were presented. They performed accurately when both stimuli were presented in the intact hemifield. However, when two targets were presented but one of them fell in the lesioned hemifield, the animals made the 'one target' response. They behaved as if they had seen only one target even though the target they ignored was easily bright enough for them to point at accurately.

There is no question that these monkeys had no spared cortex – striate cortex was surgically removed and the completeness of the damage verified at the end of the experiment. Unless one accepts that there are fundamental differences in the anatomy of vision and awareness between monkeys and man, these results suggest that blindsight cannot rely on spared striate cortex.

Blindsight and Awareness

Although blindsight is the dissociation between awareness of visual stimuli and the ability to respond appropriately to them, it is not the case that blindsight subjects are unaware of all visual stimuli presented in their scotoma. We have already seen that injuries to the occipital cortex leave patients able to report conscious experience of vigorously moving stimuli, as Riddoch discovered at the end of the First World War. Blindsight subjects also report some experience of rapidly moving stimuli or stimuli with sudden onsets or offsets. It is not clear whether these experiences are anything like *visual* sensations. Blindsight subjects differ in the descriptions they give of these experiences, ranging from a feeling that the response they are making is not quite a guess, to descriptions of movement being like a black hand moving across a black background.

Some authors have argued that the fact that blindsight subjects sometimes have an experience induced by visual stimuli, even if it is quite dissimilar to a normal visual experience, invalidates the contention that visual processing and visual consciousness are dissociated in blindsight. Weiskrantz has suggested that blindsight be divided into two subtypes:

- Type 1 blindsight conforms to the 'classical' definition and is residual visual function in the absence of any acknowledged awareness.
- Type 2 blindsight is defined as residual vision accompanied by an acknowledged experience of events in the blind field but in the absence of acknowledged 'seeing'.

It can be hard to draw broad conclusions about the nature of awareness from type 2 blindsight, as distinguishing between visual and nonvisual experience involves a difficult subjective decision about the nature of experience. Interesting results have, however, been obtained by comparing brain activation in blindsights patient when they do and do not report this type 2 nonvisual experience. These results suggest that frontal areas of the brain are activated during the experience of knowing but not seeing, whereas subcortical structures are primarily active during trials in which there is no report of experience whatsoever.

It is important to point out that the distinction between type 1 and type 2 blindsight is not based on performance. It is possible to show that the ability to perform a task and awareness of the stimuli involved are quite dissociated in type 1 blindsight. For example, as task difficulty is varied the performance of blindsight subjects can increase from chance to being near 100 percent correct without any change in their reported absence of awareness. With appropriate stimuli the dissociation in blindsight between awareness and performance remains unequivocal.

Residual Abilities in Blindsight

The early work of Weiskrantz with patient DB showed that a range of visual functions were spared in blindsight. Since then some controversial new claims have been made about the abilities of blindsight patients. First we shall look at some uncontroversial findings.

There is little doubt that blindsight patients can localize single bright or dark visual targets in their blind fields. Similarly, they can discriminate when such targets appear in a task where the subject is required to indicate in which of two time intervals a target is presented (a temporal two-alternate forced choice task). There is evidence from a number of sources that blindsight patients retain some ability to discriminate the color of stimuli presented in their blind fields, although they are impaired in comparison with normal subjects. Blindsight patients can also detect the presence of a pattern of alternating bright and dark stripes even if the average brightness of the pattern does not differ from the background. Their ability to detect these patterns is much poorer than normal in their blind field – they are unable to detect very fine or faint patterns of stripes. The ability to discriminate between stimuli composed of lines with different orientations is also preserved, albeit in a severely impaired guise and with some variations between patients. GY, for example, can discriminate the orientation of single lines but not patches of stripes. His performance becomes poorer than normal as the lines get shorter than 10 degrees of visual angle.

The ability to discriminate the orientation of lines is one of the basic building blocks of form perception. The extent to which blindsight patients can discriminate between complex forms is, however, a vexed question. Weiskrantz found that his patient DB could discriminate reliably between circles and crosses. As he showed, however, this discrimination may be based on discrimination of differences in the components of these shapes, such as the orientation of the line segments that make them up, rather than discrimination of the shapes *per se*. This is borne out by findings that blindsight subjects fail to discriminate between different shapes constructed from the same line segments, for example equilateral triangles with the point either at the top or at the bottom (Δ versus ∇) and are poor at discriminating between rectangles differing in the ratio of side lengths but not orientation.

Early results indicated that form discrimination is absent or severely impaired in blindsight. More recent studies which have tested form-processing abilities indirectly appear to tell a different story. Studies of the manual responses of blindsight subjects to objects placed wholly or partially in their blind fields indicate that shape, orientation, or size properties which could not elicit appropriate verbal or forced choice discriminations nevertheless influenced hand movement and grasp. Other studies have sought to identify whether shapes presented in the blind field influence subsequent responses to stimuli presented to the conscious good field. Although a number of groups have apparently failed to find any such effects, there have been at least two reports of positive results. In one case words presented to the blind field were reported to influence the interpretation of ambiguous words in the good field. For example, if the word 'money' was presented to the blind field then the subject was more likely to describe the word 'bank' in the good field as a financial institution than as the edge of a river. Unfortunately, relatively few short ambiguous words could be used in this study and so the result, which is of great interest given the weakness of simple form processing in blindsight, is based

on relatively few observations. Further evidence derives from a study in which the similarity between the shape of stimuli (in this case single letters) presented in the blind and good fields influenced reaction time to the good field stimulus in a letter discrimination task. The most dramatic evidence supporting the existence of complex shape discrimination without awareness comes from a study on the perception of emotion in blindsight. The blindsight patient GY correctly attributed one of four emotions (happiness, sadness, fear, or anger) to video clips presented to his blind field of an actress expressing one of these emotions.

What are we to make of the apparent contradiction between the limited shape-processing abilities indicated by studies of simple geometric shapes and the abilities necessary to make the complex discriminations required in order to be influenced by letters, words, and facial emotions presented in the blind field? One possibility is that most of the latter tasks did not involve the subject in responding directly to the stimulus in the blind field. By assessing blind-field shape-processing through its effects on seen targets, subjects are relieved of the problem of making decisions about stimuli they do not believe they can see. Perhaps removing the conflict for the subjects between their conscious blindness and the demands of a task in which they must respond to stimuli they cannot see uncovers abilities hidden in direct tasks. It may also be the case that certain properties of stimuli and methods of response are mediated by specialized neural circuits. Perhaps the processing of emotion is of such basic evolutionary importance that facial cues to emotion are processed by systems independent of the brain's general shape identification system. These are open questions; at present there is insufficient evidence to come to a firm conclusion about why and whether blindsight subjects can discriminate complex shapes without awareness.

The basis of another residual ability in blindsight is also controversial. The ability to detect the direction or speed of moving stimuli has been studied in blindsight for many years, and there is good evidence that such discriminations can be made both with and without an accompanying experience (rapidly moving high-contrast stimuli are particularly likely to elicit reports of awareness). There are, however, two ways in which motion can be inferred from the stimuli typically used in these experiments. One of these is not strictly a matter of motion perception. One can infer the direction and speed of a moving dot or line by noting its position at one instant and comparing this with its position some time later. Unfortunately there are stimuli with which such a positional comparison method will not work. For example, one can construct a stimulus comprising many dots, in which each dot is displayed for only a short time before it disappears and another dot appears at a different place. If each of these dots moves in a different direction, but on average the dots move more in one direction than any other, then a normal observer will easily be able to report the average direction and speed of the pattern (this is an example of a random dot kinematogram). In some experiments (but not all) the blindsight subject GY failed to discriminate the direction of motion when stimuli which precluded the use of position comparison were used. He could, however, still distinguish moving from stationary stimuli. It is therefore not safe to assume that motion processing is fully preserved in blindsight, even though some forms of motion can be discriminated by blindsight patients.

Some recent studies indicate that residual abilities in the blind field can be modulated by processes of alerting and spatially selective attention. GY's ability to perform a spatial localization task is enhanced if the visual stimuli are immediately preceded by an auditory warning. He is also faster at responding to a visual target if it appears in the location indicated by a preceding cue. This effect can be found even when the cue itself is also presented in the blind field. Such results may have profound implications for our understanding of the relationship between consciousness and attention.

RELEVANCE OF BLINDSIGHT FOR CONSCIOUSNESS AND COGNITIVE SCIENCE

For years consciousness was a taboo word in psychology. If blindsight has done one thing for psychology and cognitive science it is to make the scientific study of consciousness respectable once again. As well as providing insights about the relationship between processing visual stimuli and visual awareness, blindsight offers some insight into the modularity of psychological processes and the extent to which apparently complex processes can, in fact, occur essentially automatically, without awareness.

Blindsight is often used in philosophical arguments about the nature of consciousness. In particular, the apparent dissociation between access to visual information and visual experience in blindsight has been used to explore the role, and even

existence, of experiences as something distinct from the properties of stimuli in the outside world, our knowledge of them, and our responses to them. One of the attractions of blindsight to philosophers is that it appears to offer a real, albeit partial, example of a favorite of the philosophical thought experiment – the zombie.

The philosophical zombie is a being whose behavior is indistinguishable from that of real people, but who is supposed to have no inner experience at all of the world in which it is behaving. These inner experiences are often referred to as 'qualia'. Thought experiments about zombies sometimes hinge on a *reductio ad absurdum*, purporting to show that presupposing the existence of zombies leads to some paradoxical difference between our observation of the world of real people and that of zombies. It is argued that if zombies and beings with inner experience differ behaviorally, zombies who lack inner experience and yet are indistinguishable from us behaviorally must be an impossibility. Since the only difference between ourselves and zombies is the presence or absence of qualia and zombies cannot exist, then qualia have no explanatory power and hence no existence outside an individual's mind. On the other hand, one might argue that inner experiences are real (it makes sense to discuss them, as I am doing here, for example); perhaps they just do not have causal consequences for the physical world. (*See* **Zombies**)

Blindsight appears to make zombiehood concrete. Blindsight has been used to argue that inner mental states are real and correspond to physical, that is neural, states. In fact blindsight adds an extra twist to zombiehood – well-tested blindsight subjects come to know consciously that they respond appropriately to visual stimuli even though they do not know what they see and have no inner experience of seeing. Of course, it might be the case that blindsight people do have inner experiences, it is just that they do not know they have them. Unfortunately, many of these arguments are weakened when they either ignore some abilities of real blindsight people or use thought experiments which go far beyond the actual abilities of blindsight subjects. Philosophical consideration of the real properties of blindsight (as opposed to those of nonexistent super-blindsighters) does suggest that inner experiences are real, can be investigated scientifically, and make a difference in the real world, even if they do not solve the problem of telling us what such inner experiences are and why they feel the way they do.

Further Reading

Cowey A and Stoerig P (1991) The neurobiology of blindsight. *Trends in Neurosciences* **29**: 65–80.

Cowey A and Stoerig P (1995) Blindsight in monkeys. *Nature* **373**: 247–249.

Holt J (1999) Blindsight in debates about qualia. *Journal of Consciousness Studies* **6**: 54–71.

Kentridge RW, Heywood CA and Weiskrantz L (1997) Residual vision in multiple retinal locations within a scotoma: implications for blindsight. *Journal of Cognitive Neuroscience* **9**: 191–202.

Marcel AJ (1998) Blindsight and shape perception: deficit of visual consciousness or of visual function? *Brain* **121**: 1565–1588.

Morland AB, Jones SR, Finlay AL, Deyzac E, Le S and Kemp S (1999) Visual perception of motion, luminance and colour in a human hemianope. *Brain* **122**: 1183–1198.

Sahraie A, Weiskrantz L, Barbur JL, Simmons A, Williams JCR and Brammer ML (1997) Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences USA* **94**: 9406–9411.

Weiskrantz L (1986) *Blindsight: A Case Study and Implications*. Oxford, UK: Oxford University Press.

Weiskrantz L (1997) *Consciousness Lost and Found*. Oxford, UK: Oxford University Press.

Change Blindness

Intermediate article

J Kevin O'Regan, Centre National de Recherche Scientifique, Université Paris 5, France

CONTENTS

Change blindness is a phenomenon in visual perception in which very large changes occurring in full view in a visual scene are not noticed.

WHAT IS CHANGE BLINDNESS?

A number of studies have shown that under certain circumstances, very large changes can be made in a picture without observers noticing them. What characterizes the experiments showing such 'change blindness' in visual scenes is the fact that the changes are arranged to occur simultaneously with some kind of extraneous, brief disruption in visual continuity, such as the large retinal disturbance produced by an eye movement, a shift of the picture, a brief flicker, five or six small, localized disturbances flashed briefly on the picture, an eye blink, or a film cut in a motion picture sequence. These phenomena are attracting an increasing amount of attention from experimental psychologists and from philosophers, because they suggest that humans' internal representation of the visual world is much sparser than usually thought.

EXPERIMENTAL WORK ON CHANGE BLINDNESS

In the first experiments that triggered interest in change blindness (McConkie and Currie, 1996), observers viewed high-resolution, full-color everyday visual scenes presented on a computer monitor, while their eye movements were being measured. The computer could make changes in the scene as a function of where the observer looked. For example, when the observer looked from the door of a house to the window, the window (or some other element of the scene such as the sky, or the car parked in front of the house) changed in some way: it could disappear, be replaced by a different element, change color, change position, etc. It was found that when the change occurred *during* an eye movement, surprisingly large changes could be made without observers noticing them. Elements of the picture that occupied as much as a fifth of the picture area would not be seen. At first, the explanation of the phenomenon was assumed to have something to do with the mechanisms the brain uses to combine information from successive eye fixations to form a unified view of the visual world. In particular, every time the eye moves, the retinal image shifts. Some mechanism in the brain may correct for such shifts in order to create a stable view of the world. However, the mechanism could be imperfect and not take into account certain differences in the visual content across the shift, thereby explaining why changes made during saccades might sometimes go unnoticed.

But a subsequent set of experiments showed that, in fact, the change blindness phenomenon was not specifically related to eye movements. Rensink *et al*. (1997) used what they called the 'flicker' technique, in which, instead of an eye movement, a brief flicker was introduced between successive images. A first picture (picture A) would be shown for, say, 250 ms, followed by a modified picture (picture B). In between A and B, a brief blank screen (bl) would be shown. This would cause a flicker, lasting about 80 ms, that is, a duration similar to that of an eye movement. The cycle A-bl-B-bl-… was then repeated. Observers were told that something was changing in the picture every time the flicker occurred, and they were asked to search for it.

Under conditions where no flicker was inserted in between the pictures (A-B-A-B-…) the change was immediately visible and totally obvious (animated gif (158 kb): http://nivea.psycho.univ-paris5.fr/ECS/bagchangeNoflick.gif). However, with the flicker, it was often extremely difficult to locate the change. This was particularly true for changes which concerned aspects of the scene which were not of 'central interest'. For example, the reflection

of houses in a lake scene, though occupying a very large part of the picture, would not be considered to be what the picture was about. Observers sometimes were unable to see such changes at all, even after searching actively for as long as one minute. On the other hand, the changes were perfectly visible once they were pointed out to observers (animated gif (158 kb): http://nivea.psycho.univ-paris5.fr/ECS/kayakflick.gif).

The flicker technique is very easy to implement using widely available computer software (for example, software to make video presentations), and so lends itself to easy experimentation. Pictures as well as symbolic or text material can be used. The timing of the flicker between original and modified images is not critical. What is important about the flicker technique is that it shows that change blindness can be obtained without the change being synchronized with eye movements. This shows that in the earlier experiments where changes *were* synchronized with eye movements, the inability to detect the change was probably not specifically related to the eye movement and to the mechanisms that the brain uses to combine images of the world during eye explorations.

Following the discovery that change blindness was not specifically related to eye movements, but to the brief disruption that is inserted between the two versions of the picture, considerable interest in the phenomenon developed, and a large number of further experiments have been performed. These can be classified according to the nature of the disruption that is used between successive images: global disruptions, local disruptions, and progressive changes. (A review of different change blindness experiments can be found in Simons and Levin (1997).)

Global disruptions are ones in which the picture change is accompanied by a disruption which covers the whole area of the picture. The experiments where the changes occurred during eye movements were global disruption experiments, since the whole retinal image is completely smeared during the time of approximately 20 to 80 ms that it takes for the eye to move from one fixation point to the next. The flicker experiments are also global disruption experiments, since the blank displayed briefly between the original and modified images covers the whole picture. Other examples of experiments with global disruptions are experiments involving eye blinks, picture shifts, and film cuts. In the blink experiments, observers' eye blinks, registered by online computer monitoring, are used to trigger the picture change. The blink produces a global disruption similar, though somewhat longer in duration, to the disruption caused by an eye movement. In the picture shift experiments, a picture is suddenly shifted in position, and a change made at the same time. Here a global disruption is caused by the retinal smearing that accompanies the eye movement that observers make to refixate the shifted picture. In film cut experiments, observers view motion picture extracts, and at the moment when the camera 'cuts' from one view to another, a large change is made – for example, an actor is replaced by a different actor. The camera cut produces a global disruption similar to the blank in the flicker experiments.

An additional, amusing, variant of the experiments with global disruptions are experiments in which the change occurs in real life. In a typical scenario described by Simons and Levin (1998), the experimenter stops a person in the street and asks for directions. While the person is speaking to the experimenter, workers carrying a door pass between the experimenter and the person, and an accomplice takes the place of the experimenter. The person usually goes on giving directions after the interruption, and very often does not notice that the experimenter has been replaced by the accomplice.

Change blindness experiments with *local disruptions* are experiments in which, at the moment of the change, five or six small, localized disturbances are superimposed on the picture, like mudsplashes on a car windscreen (O'Regan *et al*., 1999). The disturbances can be small in comparison to the size of the change and they need not coincide with the location of the change: the change takes place in full view. As for change blindness with global interruptions, changes are very often not noticed (animated gif (378 kb): http://nivea.psycho.univ-paris5.fr/ECS/dottedline.gif).

Experiments with *slow changes* are experiments in which there is no local or global disruption at all. Instead, the change is made so slowly that the attention-grabbing processes that would normally cause attention to be attracted to the change location can no longer operate. Again, it is found that in many cases, changes are hard to detect (Quicktime video (1.4 Mb): http://nivea.psycho.univ-paris5.fr/ECS/sol_Mil_cinepack.avi).

RELATED PARADIGMS

The change blindness phenomenon is strongly related to a well-established line of research in experimental psychology that started in the 1970s and concerns visual short-term memory (for a review, see Haber, 1983).

In this literature, experiments analogous to the change blindness experiments had been performed using briefly displayed arrays of simple elements such as letters. These experiments showed that although observers have the impression of seeing all the letters in, say, a 12-element array, in fact they notice changes to or report the identity of only about four or five letters. It appears that there is a kind of attentional 'bottleneck' which limits information transfer into memory: only a fraction of the information available in a scene is transferred into visual storage for later report or comparison. Further work additionally showed that the code in which the information is stored in visual short-term memory is not a visual code, but a code in which only the category or identity of the elements is available. This work was also coherent with another line of research showing that information from successive eye fixations is combined only in categorical form, and not as a picture-like composite image (for a review, see Irwin and Andrews, 1996).

We shall see in the next section that the conclusion from these experiments, showing that visual storage is sparse and categorical, is also applicable to the change blindness results. Because in the case of change blindness, natural, highly detailed visual scenes are used as stimuli, the conclusion is more striking than it was in the older literature using simple stimuli.

Change blindness, in addition to links with research on visual short-term memory, also has relations with several more recent lines of research showing that attentional capacity in short-term visual processing is severely limited, both in spatial extent, and in the way it extends over time. Thus in 'inattentional blindness' (Mack and Rock, 1998), observers do an attention-demanding visual task. At a given moment, a large, unexpected visual event takes place. Even though such an event would be totally obvious under normal circumstances, and even though the event takes place in full view, it is often not noticed. For example, Simons and Chabris (1999) used a task in which observers look at a film of two groups of players, a black-clad and a white-clad group, each playing with their own ball in the same small room. The observer's task is to try to track the number of times one group exchanges the ball. While the observer is doing this task a woman with an umbrella walks through the room, in full view. Observers often fail to notice this totally obvious event.

An example of a temporal restriction on the deployment of attention is the 'attentional blink' (for a review, see Shapiro and Terry, 1998): in this, an observer is required to identify a target letter in a stream of rapidly presented letters. It is found that the observer often fails to report the occurrence of a second target letter if the second target follows the first by less than about 450 ms: it is as though attention had to recover for a brief period after having been solicited. In 'repetition blindness' (Kanwisher, 1987), a visual stimulus such as a letter, symbol, picture, or word tends not to be noticed if it is the second of two identical occurrences of the item in a rapidly presented series.

Another field of research that has connections to change blindness is the extensive literature on memory and cognitive descriptions (for a review, see Pani, 2000). Part of the explanation for change blindness may reside in memory limitations rather than in perceptual limitations. If this is so, then we expect that change blindness will be affected by factors similar to those that affect memory. This is compatible with the finding that changes made to elements in a scene which are of 'central interest' will in general be easier to detect than 'marginal interest' changes. Other work has shown that variables such as semantic coherence, observer familiarity, and task to be achieved, affect change blindness in a way similar to how they affect normal memory.

THEORIES OF CHANGE BLINDNESS

The currently accepted explanation of change blindness owes much to the work done in the 1960s and 1970s showing how visual information is transferred via an attentional 'bottleneck' to a very low capacity short-term visual storage (e.g. Gegenfurtner and Sperling, 1993). Within this context, the explanation of change blindness involves two components: a component related to what is called 'visual transients', and a component related to the way a scene is encoded in memory.

Visual transients are fast changes in luminance or color in the retinal image, such as would be produced by a sudden appearance or disappearance, or through motion of an element of the scene. It is known that such transients are detected in the first levels of the visual system, and that attention is automatically attracted to the location where they occur. Under normal viewing conditions, therefore, when a change occurs, it produces a visual transient which attracts attention to the change location. The transient thus provides information *that* a change has occurred, and it says *where* it occurred, but it does not provide information about *what* the change was.

In order for an observer to be able to determine *what* the change was, he or she will have had to have encoded into visual memory what was at the change location before the change occurred, and compare it to what is there after the change. There are thus two things that can go wrong in change detection: either the transient that attracts attention to the change location may be interfered with (thereby causing a deficit in detection *that* or *where* the change has occurred), or the encoding and comparison process may be interfered with (causing a deficit in determining *what* the change was).

Both these mechanisms may be at work in the change blindness experiments. In the paradigms using global disruptions, such as the flicker, blink, and film cut experiments, the global disruption presumably creates a large number of transients all over the picture, which mask or compete with the local transient corresponding to the sought-for change, and which prevent attention being automatically drawn to it. The change will be immediately noticed only if an observer happens to have been attending to the changing element at the moment it changes. Failing this, in order to find a change, the observer must search through the scene looking for an element which is different from what was previously encoded about the scene. However, because of the limitations in short-term visual memory, very little of the scene is likely to have been previously encoded, and the chances of success are very limited.

In change blindness paradigms using local disruptions like 'mudsplashes', the situation is very similar, with the difference that the local transient corresponding to the change location is missed by observers, not because it is swamped by a global transient, but because the mudsplashes act as 'decoys', attracting attention to locations other than the true change location.

In change blindness paradigms with slow changes, the change occurs so slowly that no local transient is generated. Attention is thus not attracted to the change location, and again, the observer must rely on the very sparse information that he or she has encoded about the scene in order to locate the change.

Whereas researchers working in change blindness will broadly agree on the explanation just outlined of the phenomenon, further work is necessary to ascertain the relative roles of the different component mechanisms involved. To what extent does the flicker in the flicker paradigm act to mask or 'wipe out' the internal representation? Or does it act essentially like the mudsplashes in the mudsplash paradigm to create local transients that act as decoys? Exactly how much information is encoded concerning the initial and final views of the scene? Is the overall 'gist' of a scene coded in some way? Does what is encoded depend on the observer's attentional state, on the task, on viewing strategies, on the semantic relation between the gist of the scene and the element that is changing? Are certain aspects of elements (their layout? their color?) automatically encoded and easier to detect when they change? Even if little information is available to make conscious judgments about display changes, could it be that some information is retained unconsciously? A number of recent lines of research are investigating these issues.

RELEVANCE OF CHANGE BLINDNESS FOR CONSCIOUSNESS AND COGNITIVE SCIENCE

Change blindness raises an important question: if the information that is encoded about a visual scene is so sparse, how is it that we have the subjective impression of visual richness, that is, of seeing everything there is to see in our field of view, so to speak in 'glorious technicolor and cinemascope'?

Perhaps the most natural view to take is to suppose that what we have the subjective impression of seeing is not the very sparse, more semantically coded, content of visual memory, but the content of a shorter-lived but higher quality, image-like replica or 'icon' of the visual scene. The impression of richness that we have from the world would derive from this high-quality icon. On the other hand, only a small portion of the icon's contents, namely the parts that have been attended to, would at any moment be transferred into memory and be available for doing such tasks as change detection – the rest would be forgotten. This view of visual processing has been called 'inattentional amnesia' (Wolfe, 1999): the idea is that we see everything, but forget most of it immediately.

The notion that what underlies the richness of vision is a high-quality internal replica of the outside world is the basis of some of the current work in neurophysiology and neuroanatomy, where cortical sites are being sought which provide the 'neural correlate of consciousness'. Indeed, area V1 of the visual cortex contains a distorted map of the visual field which might be a plausible locus for visual consciousness, possibly in relation to other brain structures.

A more radical answer to the question of why we have the impression of continuously seeing everything in our visual field has also been suggested

(O'Regan and Noë, 2001). The idea is that in fact the experience of seeing does not derive from the activation, inside the brain, of an 'icon' of the outside world. Rather, the experience of seeing is somewhat like the temporally extended, multifaceted experience of driving a car, involving a kind of 'give and take' between the observer and the environment, a kind of attunement to the laws that link the observer's actions to the changes in sensory input.

Under this view, the outside world serves as a form of 'external memory'. Only those aspects of the environment that are currently being 'visually manipulated' are actually available for conscious processing at a given moment. We have the impression of seeing everything because we know we have access to everything, even though without actually accessing something, no detailed information is available about it. This explains the apparent paradox between the feeling of richness we have of our visual environments, and our striking inability, in change blindness experiments, of knowing what has changed.

References

Gegenfurtner KR and Sperling G (1993) Information transfer in iconic memory experiments. *Journal of Experimental Psychology: Human Perception & Performance* **19**(4): 845–866.

Haber RN (1983) The impending demise of the icon: a critique of the concept of iconic storage in visual information processing. *Behavioral and Brain Sciences* **6**: 1–54.

Irwin DE and Andrews RV (1996) Integration and accumulation of information across saccadic eye movements. In: Inui T and McClelland JL (eds) *Attention and Performance XVI: Information Integration in Perception and Communication*, pp. 125–155. Cambridge, MA: MIT Press.

Kanwisher NG (1987) Repetition blindness: type recognition without token individuation. *Cognition* **27**(2): 117–143.

Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

McConkie GW and Currie CB (1996) Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception & Performance* **22**(3): 563–581.

O'Regan JK and Noë A (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* **24**(5): 883–917.

O'Regan JK, Rensink RA and Clark JJ (1999) Change-blindness as a result of 'mudsplashes'. *Nature* **398**: 34.

Pani JR (2000) Cognitive description and change blindness. *Visual Cognition* **7**(1/2/3): 107–126.

Rensink RA, O'Regan JK and Clark J (1997) To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* **8**(5): 368–373.

Shapiro K and Terry K (1998) The attentional blink: the eyes have it (but so does the brain). In: Wright RD (ed.) *Visual Attention*, pp. 306–329. New York, NY: Oxford University Press.

Simons DJ and Chabris CF (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* **28**(9): 1059–1074.

Simons DJ and Levin DT (1997) Change blindness. *Trends in Cognitive Sciences* **1**(7): 261–267.

Simons DJ and Levin DT (1998) Failure to detect changes to people in a real-world interaction. *Psychonomic Bulletin and Review* **5**(4): 644–649.

Wolfe JM (1999) Inattentional amnesia. In: Coltheart V (ed.) *Fleeting Memories*. Cambridge, MA: MIT Press.

Further Reading

Coltheart V (ed.) (1999) *Fleeting Memories: Cognition of Brief Visual Stimuli*. Cambridge, MA: MIT Press.

O'Regan JK (2001) Thoughts on change blindness. In: Harris LR and Jenkin M (eds) *Vision and Attention*, pp. 281–302. New York, NY: Springer.

Pashler HE (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Special issue on Change Blindness in *Visual Cognition* (2000) **7**:(1/2/3). DJ Simons' change detection database: http://viscog.beckman.uiuc.edu/change

Chinese Room Argument, The

Intermediate article

John Searle, University of California, Berkeley, California, USA

CONTENTS

The Chinese room argument is a refutation of 'strong artificial intelligence' (strong AI), the view that an appropriately programmed digital computer capable of passing the Turing test would thereby have mental states and a mind in the same sense in which human beings have mental states and a mind. Strong AI is distinguished from weak AI, which is the view that the computer is a useful tool in studying the mind, just as it is a useful tool in other disciplines ranging from molecular biology to weather prediction.

SUMMARY OF THE ARGUMENT

Strong AI is often expressed in the formula: 'Mind is to brain as program is to hardware.' On this view, the human mind is a program running in the hardware, or 'wetware', of the brain. The Chinese room argument against strong AI proceeds by a thought experiment. If strong AI were true, then one could acquire any cognitive capacity that one does not have by simply implementing the program for that cognitive capacity in a way that would enable one to pass the Turing test.

Imagine that I, who am a native English speaker, unable to speak any Chinese at all, am locked in a room containing several boxes of Chinese symbols (the database). Imagine that I have in the room a set of instructions for manipulating Chinese symbols (the program). I receive, through a window in the room, Chinese symbols which, unknown to me, are in the form of questions. I follow the instructions in the program, and give back through the window Chinese symbols which, unknown to me, are answers to the questions. For the purposes of the thought experiment we may suppose that the programmers get so good at writing the programs, and I get so good at shuffling the symbols, that after a time my answers are indistinguishable from those of the native Chinese speaker. I pass the Turing test for understanding Chinese, and I do so by implementing the program. But I do not understand a word of Chinese. This is the point of the thought experiment: if I do not understand Chinese by virtue of implementing the Chinese-understanding program, then neither does any other digital computer by virtue of doing so.

Why is it that I do not understand Chinese? The answer seems obvious. Though I manipulate the symbols, I have no knowledge of what any of the symbols mean. One can see this by contrasting my manipulation of Chinese symbols with my answering questions in English. Suppose that the people on the outside of the room also submit written questions in English and I submit written answers to the questions. My answers to the questions in Chinese are as good as those of a native Chinese speaker because I have been appropriately programmed. My answers to the questions in English are as good as a native English speaker because I am a native English speaker. From the outside, from the third-person behavioral point of view, my behavior is equally good in Chinese and in English. From the inside it is obviously quite different: in English I understand perfectly both the questions and my answers; while in Chinese I understand nothing – I am just a digital computer.

Construed as a deductive argument, the Chinese room argument has three steps and a conclusion. We may formulate these as follows.

> Computer programs are defined entirely in terms of symbolic or syntactic operations. (1)

The implemented program consists entirely of symbol manipulations. To put this somewhat more precisely: the notion 'same implemented program' defines an equivalence class that is specified entirely in symbolic or syntactic terms, and independently of the physics of the underlying medium. There is nothing more to the implemented program, qua implemented program, than symbol manipulations.

> Minds – actual human minds such as yours and mine – have mental contents or semantics. (2)

For example, when I understand a sentence in English I have more than just symbols going through my head: I know what the symbols mean.

> By themselves, the implemented syntactic steps of the program are neither constitutive of mental content nor sufficient to guarantee the presence of mental content. (3)

This is what was shown by the Chinese room thought experiment. I went through the appropriate syntactic steps, but I had no Chinese thought content, no Chinese semantics, associated with them.

> Conclusion: the implemented computer program is insufficient by itself to constitute or to guarantee the presence of the appropriate mental states. (4)

I went through the right steps of the program, I had the right behavior, but I did not have the appropriate mental states. Therefore, strong AI is false.

The argument rests on two fundamental logical principles: firstly, syntax is not semantics, and secondly, simulation is not duplication. Any problem-solving process that can be described as an effective procedure, that is, a procedure going through a finite number of exactly specifiable discrete steps, can be programmed on a computer. That is why the computer is so powerful: we can represent any domain that we can describe precisely. Thus we can represent the stages of the weather, the flow of money in the economy, or the understanding of Chinese sentences. The syntax of the program states can be used to represent anything. They can be used to represent weather changes, economic developments, and even semantics. But the simulation of the process, whether it be atmospheric, economic, or semantic, is not a duplication of the process. You do not produce a rainstorm by doing a computer simulation of a rainstorm. You do not produce wealth by doing a computer simulation of the production of wealth. And you do not produce understanding and thought processes by doing a computer simulation of understanding and thought processes.

RESPONSES TO THE ARGUMENT

A number of responses have been presented against the Chinese room argument. We will consider four of these.

The Systems Reply

Perhaps the most commonly presented answer is this: the person in the room does not understand, but the person is only an element in a larger system. The system consists of the room, the program, the database, etc. So the understanding should be found in the entire system, not in the person, because the person is only the central processing unit.

Just as we would not say of a single neuron in the brain that it understood English, so we should not say of a single element, the person, in the whole system that that person understands Chinese.

Answer to the systems reply

The answer to this reply is that the reason the person does not understand is that he has no way to attach any meaning to the symbols. But if he has no way to attach meaning to the symbols, neither does the whole room. The whole room has no way to get from the syntax to the semantics any more than the person does. To see this, simply imagine that the person internalizes the entire system. Imagine that the person memorizes the database, memorizes the program, does all of the calculations in his head, and works outdoors in the middle of an open field. In this variation there is nothing in the room that is not in the person, and still there is no understanding in the person.

The Robot Reply

The robot reply is based on a variation of strong AI whereby the unit of understanding is not the computer, but the computer within a motorized system that will be able to process sensory inputs and produce motor outputs computationally. The robot would move about, with video cameras attached to its head; it would take in information from the video cameras, and adjust its movements accordingly.

According to the robot reply, the computer by itself does not have semantic content, but the causal relations between a robot and the external world would be sufficient to give semantic content to the symbols processed by the robot.

Answer to the robot reply

The robot reply tacitly abandons the thesis of strong AI, which is that the implemented computer program by itself is sufficient to guarantee or constitute understanding. The idea behind the robot reply is that the addition of causal relations between the system and the external world would

be sufficient to produce semantic content or understanding.

The answer to the robot reply is that even this amendment to the strong AI thesis will not be sufficient to produce understanding. Imagine that the robot has a very large cranium, and inside the cranium is a room, and I am inside the room. I receive inputs in the form of Chinese symbols. I process them according to the program, and I produce outputs in the form of Chinese symbols. Unknown to me, the input symbols are the product of video cameras and other sensors attached to the outside of the robot. The input stimuli are converted by transduction into Chinese symbols, and the output that I provide is converted into motor output of the entire robot system. But I have no way of understanding what is going on because I have no way of attaching any meaning to any of the symbols, or to anything else that is going on in the robot's brain. I am the robot's homunculus, but unlike the usual homunculi of philosophical literature, I understand nothing, because I have no way of attaching any meaning to any of the symbols that I process.

The robot reply tries to defeat the Chinese room argument by adding causal relations. But the causal relations will produce semantic content only if there is some conscious agent who can become aware of the causal relations. I, as a human being, can become aware of the causal relations between the Chinese symbol for chicken chow mein and the actual food type of chicken chow mein if I can *see* chicken chow mein associated with this symbol. But in the robot, I am just a computer and, like any other computer, I function by processing meaningless symbols. The symbols in the computer brain are meaningless in a way that is quite different from the symbols passing through my mind when I think in English. When I think in English, symbols do indeed go through my mind, but I know what they mean.

The Brain Simulator Reply

Suppose we simulated the actual operations of a Chinese person's brain when that person understands sentences in Chinese. Suppose we produced a perfect computer simulation of all of the synaptic transmissions in the Chinese brain. Then we would have to say that the system understood, otherwise we would have to deny that the Chinese person understood. Since the brain operates, like a computer, with a series of state transitions, there is no reason why we could not produce a perfect replica of these state transitions on a digital computer.

Answer to the brain simulator reply

The computer simulation of the brain is not duplicating the relevant features of the brain. It is merely duplicating the formal pattern. The actual human brain, like any other organ, is a causal mechanism, and it causes consciousness and intentionality by quite specific neurobiological processes. The computer merely produces a model or representation of these processes, but the model or representation lacks the causal features of the original.

To see this, compare the brain to any other organ. We can do a perfect simulation of the digestive processes in the stomach on a digital computer. But even if we have a perfect simulation on the computer, to any degree of accuracy, we do not produce actual digestion. When we run the digestion program, we cannot put a pizza into the computer and expect the computer to digest it. The computational simulation is merely a matter of zeros and ones, not a matter of the enzymes and other chemicals that actually carry out digestion. The situation in the brain is similar. Specific biochemical processes cause consciousness and intentionality. We cannot reproduce those by doing a simulation with zeros and ones, any more than we can reproduce digestion with zeros and ones.

The Parallel Distributed Processing Reply

The Chinese room argument works against the von Neumann symbolic digital computer, but recent developments in computer technology have created new types of computational systems which are immune to the Chinese room argument. These new types of systems are known variously as 'parallel distributed processing' (PDP), 'neural net modeling, 'connectionism' or 'new connectionism'.

PDP systems function in a way that is quite different from the traditional von Neumann system. They have a series of computational processes going on in parallel, distributed over a network. Whereas the traditional von Neumann machine works in a series of discrete steps, PDP systems do massively parallel distributed processing.

Answer to the parallel distributed processing reply

There is an ambiguity in the PDP reply. It is not clear which of the differences between the connectionist machine and the von Neumann machine are being appealed to in order to claim that the connectionist machine is not subject to the Chinese room argument. Either what is claimed is that there is some computational power of the connectionist

machine lacking in the von Neumann machine, or it is claimed that there is some hardware feature of the connectionist architecture which is superior to the von Neumann architecture. But neither of these approaches is successful in answering the Chinese room argument.

According to Church's thesis, there is no computation that can be performed on a connectionist machine that cannot be performed on a von Neumann machine. According to this thesis, any computable function whatever, any problem that can be solved algorithmically, can be computed on a Turing machine. All effective computability is Turing computability. Church's thesis is one of the foundational principles of the modern theory of computation and is universally accepted by the parties to this dispute. It has the consequence that there cannot be any computational power possessed by a PDP system that is not possessed by a von Neumann machine.

The other possibility is that something is being claimed for the connectionist architecture: for the actual structure of the wiring and the hardware. But if this is so, it is no longer strong AI. Strong AI is a thesis about the powers of computation. If it is claimed that the particular hardware of the connectionist machines can duplicate the powers of the brain to cause mental content, then the thesis is no longer strong AI, but is rather a form of speculative neurobiology. The Chinese room argument is not intended to answer any claims in speculative neurobiology, but is intended as a logical thesis about the distinction between the syntax of the implemented program and the semantics of actual human minds.

Either we are to think of the essential feature of the system as being its computational power, or we are to think of it as some causal property of the specific hardware in which the computation is implemented. If it is a matter of computational power, no new power is added by the connectionist architecture. If we are to think of it as an architectural feature, then it is no longer the thesis of strong AI. Actual human brains cause consciousness and other mental phenomena by way of specific neurobiological processes operating in a 'bottom-up' fashion. That is, processes at the level of neurons and synapses cause consciousness and other mental phenomena that are features of much larger elements of the brain system.

The thesis of the PDP reply, if followed to its logical conclusion, would have to be that the connectionist architecture is capable of duplicating and not merely simulating the causal powers of the brain to cause higher-level consciousness, etc., by way of bottom-up causation. Nothing in the neurobiological literature would tend to support this thesis. In any case, it is not strong AI, and in consequence, is irrelevant to answering the Chinese room argument.

COMMON MISUNDERSTANDINGS OF THE CHINESE ROOM ARGUMENT

The Chinese room argument is sometimes misinterpreted, and several of these misinterpretations are common in the literature. Firstly, it is sometimes supposed that the argument is intended to show that 'machines cannot think'. But that is not the point of the argument. The argument assumes that the brain is a machine. The problem with computation is that in the relevant sense it does not name a machine process. It names an abstract mathematical process that we can implement on machines, but computation is not defined in terms of machine processes such as energy transfer. Thus, on the view implicit in the Chinese room argument, the brain is a machine, and brain processes are machine processes. 'Computers' of the ordinary kind are indeed machines, but computation is not essentially a machine process.

Another misinterpretation of the Chinese room argument is that it is attempting to show that only human brains have the power of thinking. But that is not the point of the argument. Whether or not we can build an artifact out of some other type of material capable of producing consciousness is an empirical scientific question. In principle there is no more serious logical obstacle to building an artificial brain than there is to building an artificial heart. The point of the argument is that we do not produce the same causal powers by simply duplicating the formal pattern. The computer gives us a picture, or a model, of thought processes, but it does not actually produce thought processes.

CONCLUSION

In the early days of cognitive science, the computationalist model of cognition was the dominant paradigm. At present there is a gradual paradigm shift away from computational cognitive science towards cognitive neuroscience. As we learn more about the brain we see that cognition is essentially a matter of a certain sort of brain processing. We may be able to simulate this on a digital computer, and we may eventually be able to duplicate it in some other medium. But the Chinese room argument shows that simulation by itself does not guarantee duplication. To guarantee duplication, the artificial

creation of a real mind, we would have to duplicate, and not merely simulate, the powers that actual brains have to cause consciousness and cognition.

Further Reading

Dietrich E (ed.) (1994) *Thinking Computers and Virtual Persons*. San Diego, CA: Academic Press.

Preston J and Bishop M (eds) (2002) *Views Into the Chinese Room: New Essays on John Searle's Arguments Against 'Strong AI'*. Oxford, UK: Oxford University Press.

Searle JR (1980) 'Minds, brains, and programs'. *Behavior and Brain Sciences* **3**: 417–457.

Searle JR (1982) The Chinese room revisited. *Behavior and Brain Sciences* **5**: 345–348.

Cognitive Science: Philosophical Issues

Introductory article

Barbara Von Eckardt, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

CONTENTS

Numerous philosophical questions can be raised about cognitive science, including what cognitive science is, what counts as representation, computation, and explanation in cognitive science, and what the limits of cognitive science are.

INTRODUCTION

Cognitive science is a multidisciplinary approach to the study of cognition and intelligence. It emerged in the late 1950s and early 1960s when researchers in psychology, linguistics and computer science worked together to develop an alternative to behaviorist approaches to mind and language. At its core, it holds that the human mind is a kind of computer which processes information in the form of mental representations (in this article, references to 'mind' are in the sense of a functional description of the brain). The major disciplinary participants in the cognitive science enterprise are psychology, linguistics, neuroscience, computer science, and philosophy. Other fields that are sometimes included are anthropology, education, mathematics, biology, and sociology.

Broadly speaking, there are two kinds of philosophical issue associated with cognitive science: theoretical and conceptual issues raised within or about cognitive science to which philosophers have made a contribution; and traditional philosophical issues to which cognitive science research is relevant. Examples of the first would be: By virtue of what do mental representations have semantic properties? Are 'symbolic' or connectionist models more likely to account for our cognitive capacities? Examples of the second would be: Are there innate ideas? Is color an objective or a subjective property?

This distinction is somewhat blurred because theoretical and conceptual issues in cognitive psychology are sometimes very similar to those that have historically been discussed by philosophers. This article deals only with issues of the first type.

WHAT IS COGNITIVE SCIENCE?

The Core Research Framework of Cognitive Science

Because cognitive science is immature, complex, and continually changing, the question 'What is cognitive science?' is not easy to answer. One way to start is to identify a relatively stable 'core' and then describe various ways in which some cognitive science research deviates from that core.

The aim of core cognitive science is to explain the human cognitive capacities. These include our capacity to use language (perceive it, comprehend it, produce it, translate it, communicate with it, and so on), to perceive visually, to apprehend music, to learn, to solve problems (reason, draw inferences), to plan actions, to act intentionally, to remember, and to imagine. Each of these capacities can be explored in several ways. Consider, for example, the capacity for language. Cognitive scientists study language in normal adults; language development; language variation between cultures, groups (e.g. men and women, first and second language learners), and individuals; pathologies of language (speech disorders, aphasia); and how language is realized in the brain.

In studying the cognitive capacities, cognitive scientists make three basic assumptions: that the human mind is a kind of computer, that it has mental representations, and that it exhibits both conscious and unconscious processing. A popular way of describing cognitive science research is in terms of the three levels proposed by the vision scientist David Marr: a computational level, an algorithmic level, and an implementation level. At the

computational level, we ask what precisely any given capacity is as a (mathematical) function from inputs to outputs; at the algorithmic level, we attempt to explain how a person executes the capacity characterized at the computational level; and at the implementation level, we ask how the capacity is implemented in the human brain.

Core cognitive science is also characterized by methodological assumptions. The most important are: that the research methods of cognitive science are scientific; that a complete theory of human cognition will not be possible without a substantial contribution from each of the contributing disciplines of cognitive science; that human cognition can be successfully studied by focusing on the individual cognizer and his or her place in the natural environment; and that answers to the basic questions of cognitive science in terms of information processing are constrained by the findings of human neuroscience.

The research framework of core cognitive science is, in principle, a framework to which each of the contributing disciplines of cognitive science conforms. However, each of these disciplines contributes to the research program built on this framework in its own distinctive way (see list below). (Note that not all subdisciplines of the contributing disciplines are part of cognitive science. For example, philosophy encompasses ethics and political philosophy, which are not considered part of cognitive science.)

Psychology

Enhances our understanding of the nature and limits of our cognitive capacities.
Proposes hypotheses concerning normal adult cognition, including hypotheses about the mental representations and computational processes involved in any given cognitive capacity, and devises experiments for testing these hypotheses.
Studies the acquisition of cognition in children and the development of cognition throughout our lifetime.
Studies individual and group differences in cognition.
Studies the psychopathology of cognition.

Computer science

Proposes computationally detailed models of cognition and tests these models against the data provided by psychology and neuroscience.
Develops hypotheses regarding the computational nature of the representation-bearers for the human system of mental representation.

Neuroscience

Studies the realization of our capacities in the brain.
Studies what happens when our cognitive capacities are exercised abnormally due to some neural abnormality or lesion.
Develops computational and representational hypotheses concerning cognition in normal adults, children, and abnormal individuals on the basis of low-level information about the brain.

Linguistics

Provides a theory of ideal capacity (a competence model) for language comprehension, production, and acquisition.
Proposes hypotheses regarding the representations involved in language comprehension and production.

Anthropology

Studies cognition across cultures.

Philosophy

Articulates the foundations of the field.
Explores the viability of the cognitive science research program.
Contributes to the development of a theory of content determination for the system of representation posited by cognitive science theories.
Contributes to the discussion of controversial theoretical and empirical issues in cognitive science, often involving adjudication between competing claims.
Helps to develop theories of ideal capacity (competence models) for reasoning and language use.

Areas of Disagreement and Evolution Within Cognitive Science

Within core cognitive science, the primary area of disagreement and change has been the conception of a computer. Initially, computers were thought to be 'physical symbol systems', a view proposed by Allen Newell and Herbert Simon in 1976. Research based on this assumption has been given various labels: 'classical', 'conventional', 'symbolic', and 'rules- and representation-based'. In the mid-1980s a new class of 'connectionist' or 'parallel distributed processing' computers began to attract the attention of cognitive scientists, resulting in heated disagreements about whether symbolic or connectionist computers were the best model of the human mind. Recently, two further views have emerged. The 'dynamic' approach models cognition in terms of a set of quantitative variables that change over time in accordance with dynamical mathematical laws (generally expressed by differential or difference equations). And many neuroscientists hold that the brain is not a physical

symbol system nor a connectionist device nor a dynamic system; rather, it is a fourth, specifically biological kind of computational device, still poorly understood.

Deviations from core cognitive science can be found with respect to its domain, its approach, and the role it assigns to neuroscience. Since the late 1970s and early 1980s, when cognitive scientists started making statements about what cognitive science is, there has been a split (largely along disciplinary lines) over whether the domain of cognitive science is human cognition, or intelligence in general, including human, machine, and possibly animal intelligence. A further area of disagreement concerns whether cognitive science is concerned only with cognition or whether it includes all aspects of the mind, including touch, taste and smell, emotion, mood, motivation, and personality.

Although there is fairly widespread agreement that the core working assumption of the cognitive science 'approach' is that the mind is both a computational and a representational device, there are researchers, who consider themselves to be under the cognitive science umbrella, who have rejected either one of these assumptions: There are those who hold that the mind performs computations without representations; and there are also those who posit representations without computations.

A final area of disagreement concerns the role of neuroscience. Because cognitive science originally emerged from cognitive psychology, artificial intelligence research, and generative linguistics, in the early years neuroscience was often relegated to a secondary role. Citing the fact that descriptions of a system in purely functional terms can always be 'multiply realized', some cognitive scientists even declared that neuroscience, as the science of the physical realization of the functional mind, was irrelevant to cognitive science. Most researchers adopted a more moderate position: that the particular neural realization of the human mind imposes important constraints on what functions the mind can compute. But even such moderates sometimes held that research on the mind could (and should) proceed in a 'top-down' fashion. Most cognitive scientists, however, now believe that research on the mind should proceed in an interactive way: simultaneously top-down and bottom-up. The prevailing view in the field of cognitive neuroscience seems to be that neuroscience will be at the centre of efforts to develop a computational or information-processing theory of the mind driven primarily by bottom-up considerations.

Cognitive Science and Folk Psychology

The relationship between the conceptual frameworks of cognitive science and folk psychology is complex. As one might expect of an immature science, cognitive science has strong roots in common sense, but is also striving to develop its own empirically-based theoretical categories. Thus, cognitive science has clearly taken on board the ordinary notion of a cognitive capacity and the ordinary taxonomy of the specific cognitive capacities we take ourselves to have, including many of their associated states (such as states of perceiving, understanding and intending). However, since folk psychology has little to say about the unconscious processing that goes on when we exercise those capacities, it has been necessary for cognitive science to introduce a variety of subpersonal unconscious information-processing states to describe that processing. Furthermore, to complete the description at the subpersonal level, subpersonal information-processing states are also posited as 'underlying' the conscious states described in folk psychology. Thus, for example, recognizing that a given string of letters is a word of English becomes, in the theoretical language of cognitive science, successfully accessing the appropriate entry in the mental lexicon on the basis of a graphemic representation of the word.

REPRESENTATION IN COGNITIVE SCIENCE

Mental Representations

A mental representation is a structure or state in the mind that has semantic properties, such as referring to phenomena in the world, expressing a proposition, or predicating some property of something. In addition to having semantic properties (content), cognitive scientists generally assume that mental representations are carried by some representation bearer, that their content is 'grounded' in naturalistic properties and relations, and that they have significance for the individual that has them.

Given the assumption that the mind is a kind of computer, the bearers of mental representations are hypothesized to be computational structures or states. The specific nature of these structures or states depends on the kind of computer one takes the mind to be. The representation bearers of classical ('symbolic') computers are typically data structures; in contrast, the representation bearers of connectionist computers are activation

states of nodes or sets of nodes (corresponding to occurrent mental states), or states consisting of links having certain weights (corresponding to dispositional mental states).

In attempting to explain cognition and intelligence, cognitive scientists have posited many kinds of mental representation. There is no neat taxonomy of these kinds. Sometimes representations are grouped together based on what they represent (phonological, lexical, syntactic and semantic representations in psycholinguistics), sometimes on their computational characteristics (local and distributed representations in connectionist systems), and sometimes on both (sentences, frames, schemas and scripts).

Theories of Content Determination

Philosophers interested in mental representation have focused primarily on the problem of what determines the semantic content of mental representations. Such theories are sometimes misleadingly referred to as theories of 'semantics' (e.g. 'informational semantics', 'functional role semantics'). It is more accurate to call them theories of 'content determination'. It is generally assumed that mental content is not a basic fact about the world, hence, that it comes about because the bearers of our mental representations have other naturalistic (non-intentional, non-semantic) properties. These content-determining properties can be considered the 'ground' of representational content.

There are currently a variety of proposals about this ground. They appeal either exclusively or in some combination to: the structure of the representation bearer (Palmer); actual historical or counterfactual causal relations between the representation bearer and phenomena in the world (Fodor, Devitt, Dretske); actual and counterfactual (causal, computational, inferential) relations between the state of the representation bearer and other states in the mind (Block, Cummins); the information-carrying or other functions of the state of the representation bearer and associated components (based on the qualities for which they were selected in evolution or learning) (Millikan); and the phenomenal properties of the representation. All current theories of content determination have problems. A major problem is how to account for the fact that we can falsely represent. For example, we can have a perceptual experience of a state of affairs that does not actually exist, or entertain a thought that is false.

COMPUTATION IN COGNITIVE SCIENCE

Philosophers have discussed two issues concerning the computational assumption of cognitive science. Cognitive science assumes that the mind is a computer. But what exactly is a computer? This question becomes particularly important given that alternative 'machine' models have been proposed (e.g. physical symbol system, connectionist device). Are these all different kinds of computer, as suggested above, or did cognitive science move away from the computational assumption when it became interested in connectionism? To answer these questions we need to know what a computer is.

The second issue concerns the dispute between the 'symbolic' and connectionist approaches. Which is the most promising approach? And can we decide this question on the basis of currently available evidence about cognition?

Computation and Computers

The term 'computer' is used in many different ways. Arguably, the ideal conception of a computer for cognitive science would be one that: (1) builds on the mathematical theory of computation; (2) encompasses classical 'symbolic' models, connectionist devices, and talk of a 'computational' brain; (3) isn't vacuous, and (4) is sufficiently informative to provide some guidance for construction of a theory. Point (3) is important: both Putnam and Searle have argued that the current conception of a computer is so vague as to be vacuous, in the sense that we can interpret any physical system as instantiating any computational characterization.

The theory of computation in mathematics defines a variety of different kinds of abstract machines capable of 'computing' functions in the mathematical sense. Although, in principle, functions can be defined over any domain (e.g. 'father of' maps people to people), mathematicians generally study functions defined over the natural numbers; theorems concerning such functions can then be generalized to functions over other domains. An important discovery has been that different kinds of automata (abstract machines) are associated with different sets of functions distinguished by various interesting mathematical properties.

The kind of machine of greatest interest to cognitive science is the Turing machine, named after the British mathematician Alan Turing. Such machines are important because it is believed that any

function that can be computed by an *effective* method, that is, a method specified by a finite number of exact instructions, in a finite number of steps, to be carried out by a human unaided by any machinery except paper and pencil, and demanding no insight or ingenuity, can be computed by a Turing machine. (This claim is called the 'Church–Turing thesis'.) The so-called 'universal Turing machine' can compute all such Turing-computable functions.

Many cognitive scientists think that the notion of a computer relevant to cognitive science can simply be borrowed from the mathematical theory of computation. Unfortunately, the matter is not so straightforward. To see why, we must first distinguish between a physical system P being *equivalent* to an abstract automaton A and its being an *implementation* of A. Assuming we have some way of systematically mapping the inputs and outputs of A onto the inputs and outputs of P, we take P to be equivalent to A if and only if it can compute the same functions as A. To determine what counts as a legitimate implementation is much more difficult. Basically, the idea is that P counts as an implementation of A if and only if P is equivalent to A and there is a mapping from the formal states of A onto the physical states of P such that the formal state transitions of A are mirrored, both actually and counterfactually, by causal sequences in P.

There are many problems with defining a computer in terms of equivalence to some kind of abstract automaton: principally that such a definition fails to satisfy criterion (4) above, since it tells us nothing about the internal structure of the system. In contrast, while a definition in terms of implementation can be theoretically informative, saying that the mind is an implementation of a Turing machine seems false: at least, on any 'natural' way of doing the mapping. A Turing machine, even a universal Turing Machine, is in only one state at a time and uses a very small number of basic operations. In contrast, the causal structure of the brain relevant to cognition is extremely complex, and brains can change their structure in response to the information being processed. Some researchers, such as Copeland, have even suggested that the brain may have primitive operations that can compute functions that are not Turing-machine computable.

An alternative to defining a computer in terms of one or another class of abstract automaton is to provide a characterization that generalizes over different architectures but also, possibly, excludes some. One such characterization (due to Von Eckardt) is that a computer is a device capable of automatically receiving, storing, manipulating, and outputting information, by virtue of receiving, storing, manipulating and outputting representations of that information in accordance with a finite, effective set of rules that are, in some sense, in the machine itself. Another characterization (due to Copeland), specifically designed to encompass non-classical computation, is that a computer is a device, capable of solving a class of problems, that can represent both the problems and their solutions, contains some primitive operations (which may include operations that are not Turing-computable), can sequence these operations in some predetermined way, and has provision for feedback.

Computational Models and Human Cognition

A major dispute within cognitive science, to which philosophers have contributed in important ways, has been whether the mind is a classical or 'symbolic' computer ('classicism') or a connectionist computer ('connectionism'). Much of the discussion has centered around a challenge to connectionism put forward in 1988 by two proponents of classicism, Jerry Fodor and Zenon Pylyshyn. Fodor and Pylyshyn argued as follows:

1. An adequate theory of cognition must explain the fact that our cognitive capacities exhibit systematicity, that is, roughly, that some capacities are intrinsically connected to others. (If a native speaker of English knows how to say 'John loves the girl' she will know how to say 'The girl loves John'.)
2. Classical models can explain this property of cognition.
3. But it is not clear that connectionist models can, unless they are implementations of a classical model – in which case connectionism wouldn't constitute an alternative to classicism.
4. Therefore, classicism is to be preferred to connectionism.

Their reason for advocating premises 2 and 3 is that it is 'characteristic' of classical systems but not of connectionist systems to be 'symbol processers', that is, systems that posit mental representations with constituent structure (parts ordered in a certain way) and process these representations in a way that is sensitive to that structure.

Since 1988 there has been a flood of responses to this argument; in addition, the argument itself has evolved in two significant ways. First, it has been emphasized that to be a counterexample to premise 2, a connectionist model must not only exhibit systematicity; it must also explain it. Second, it has

been claimed that what needs explaining is not just that human cognitive capacities are systematic; it is that they are *necessarily* (on the basis of scientific law) systematic.

Every aspect of the Fodor–Pylyshyn argument has been subjected to scrutiny. Numerous authors have noted the need for, and proposed, a more precise conception of systematicity. There has also been discussion of what it means to model cognition at the cognitive level, under what conditions one model counts as an implementation of another, and what is required for a model to explain (rather than merely model) either the fact or the necessity of systematicity.

In addition, each of the premises has been questioned. Several computer scientists and philosophers have sought to demonstrate that connectionist systems can model systematicity without being classical implementations, thus refuting premise 3. Some have accepted the Fodor–Pylyshyn conceptualization of the challenge whereby any connectionist model that processes symbols is, *ipso facto*, a classical implementation, and have sought to demonstrate that systematicity can be modelled by a system that is not a symbol processor. Others have resisted the Fodor–Pylyshyn dichotomy, arguing that a system can be connectionist, and have structured representations and structure-sensitive processes, without also being classical. Although proponents of classicism have not explicitly conceded defeat on this point, recent defences of the Fodor–Pylyshyn argument have concentrated on the claim that connectionists will never have the resources to explain either the fact or the necessity of systematicity, rather than that they will not be able to simply model systematicity.

Premise 1 has been questioned by researchers who have variously argued: that our cognitive capacities aren't, in fact, systematic; that that systematicity is a matter of conceptual necessity and, hence, not something that an empirical theory needs to explain; or that the explanation of their systematicity need not come from a theory of cognition. Another point is that the empirical facts about systematicity may be more complicated than Fodor and Pylyshyn assume.

The questions regarding premise 2 are especially important. The Fodor–Pylyshyn argument rests on classical systems being able to do something that connectionist systems (at a cognitive, non-implementational level) cannot. However, it has been pointed out that the fact that a system is a symbol processer (and, hence, classical) does not by itself explain systematicity, much less the lawfulness of systematicity: additional assumptions must be made about the system's computational resources. Thus, if it is true that certain kinds of connectionist system can also explain systematicity, then the explanatory asymmetry between classicism and connectionism will no longer hold (that is, both classical and connectionist systems may or may not explain systematicity when appropriately supplemented) and the Fodor–Pylyshyn argument will be unsound.

EXPLANATION IN COGNITIVE SCIENCE

From the late 1960s onwards, several philosophers have noted that explanation in cognitive science has certain distinctive features. Scientific explanation had commonly been taken to be an answer to a 'why' question ('Why did this event occur?' 'Why do these regularities hold?'), and to involve subsumption of the phenomenon to be explained under laws. In contrast, it was noted, cognitive scientists seek to explain capacities, rather than events or regularities – either by virtue of what we have them (Haugeland's 'systematic explanation') or how we exercise them (Cummins's 'functional analysis') – by appeal to the organized interaction of subcapacities.

While these early discussions succeeded in drawing philosophical attention to some important distinctive features of explanation in cognitive science, the full story turns out to be more complicated. If one adopts the view that explanations are answers to certain kinds of questions and that kinds of explanation are distinguished ('individuated') both by what is being explained and by the kind of explaining being done, then, as Barbara von Eckardt has emphasized, there are many different kinds of explanation to be found in cognitive science. Each of the various questions associated with each type of cognitive capacity has its own kind of explanation. For any given capacity *C* (say, the capacity to perceive visually), cognitive scientists attempt to explain how a normal adult generally exercises *C*, how a child generally acquires *C*, in what ways *C* breaks down under various conditions of psychopathy, what kinds of cultural variation there are in the exercise of *C*, and so on. Further, insofar as cognitive science entertains alternative ways of answering questions – for example, by means of a classical AI program, connectionist model, or dynamical system – even more forms of explanation can be distinguished. Classically explaining how a child acquires the capacity to speak a natural language involves a different kind of explanation than explaining the same phenomenon in a connectionist way.

Do the accounts offered by cognitive science succeed in being explanatory? Do they at least provide possible explanations for what they are intended to explain? Philosophers have raised two sorts of concern. First, the human cognitive capacities that cognitive science seeks to explain are ultimately rooted in our commonsense conception of ourselves. Thus, among the properties of these capacities that require an explanation is their intentionality, that is, the fact that they have content (we perceive or understand that something is the case). The explanatory strategy adopted by cognitive science is to explain the intentionality of mental states, as ordinarily construed, by appealing to the existence of mental representations with content at the subpersonal level. The concern is that this strategy will not work because none of the possible ways of construing subpersonal representational content gives this content the required explanatory role.

A second concern arises from the fact that although most forms of explanation in cognitive science involve 'how' or 'what' questions, cognitive scientists do sometimes seek to explain why people behave in certain ways. (The ability to explain a subject's behavior under experimental conditions constitutes the evidence for proposed answers to the 'how' and 'what' questions.) Typically, of course, these explanations purport both to be causal and to appeal to individual mental representations. But, it has been argued, there are reasons for thinking that it is not a representational state's having a certain content that is doing the causal work; rather, it is the computational or neural structure which 'has' that content (just as what causes me to perceive a word on the printed page is not the word's meaning but its physical embodiment in ink). So there really isn't any genuinely mental causal explanation.

There is one final worry. Granted that explanations that appeal to mental representations could provide possible explanations, are such explanations really needed? For example, couldn't cognitive science simply posit processing that is purely formal, or purely low-level computational (number crunching), or purely neurophysiological?

LIMITS OF COGNITIVE SCIENCE

For more than twenty years, critics of cognitive science, such as John Haugeland, have put together lists of mental phenomena that, they claim, cognitive science will never be able to explain. These include: that people 'make sense' in their actions and speech; that they have sensations, emotions, and moods; that they have a self and a sense of self; that they have consciousness; that they are capable of insight and creativity; that they can possess highly developed intellectual and artistic skills; and that they interact closely and directly with the world in which they live. When cognitive science consisted simply of a top-down, classical ('symbolic') approach, it was quite plausible to view some (or even all) of these phenomena as representing limits of the field. However, given the increasing integration of the non-neural cognitive sciences with neuroscience and the apparent flexibility of the notions of representation and computation, it is now far less clear whether these phenomena are really 'limiting'. Perhaps the most difficult aspect of mentality for cognitive science to explain is what philosophers have called 'phenomenal consciousness', that is, the 'feeling' or experience we have in connection with various kinds of mental state. In particular, as Joseph Levine has emphasized, there seems to be an 'explanatory gap' here: even if we can determine what the neural basis of a certain kind of experience is, cognitive science doesn't seem able to explain why that neural state gives rise to this kind of experience. However, even if this problem of phenomenal consciousness does represent an absolute limit to what cognitive science can explain, it is not obviously a limit to the *scientific* research program of cognitive science, that is, to the programme of developing a scientific theory of the mind.

Further Reading

Albright TD and Neville HJ (1999) Neurosciences. In: Wilson RA and Keil FC (eds) *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Bechtel W and Abrahamsen A (1991) *Connectionism and the Mind*. Cambridge, MA: Blackwell.

Block N (1986) Advertisement for a semantics for psychology. In: French PA, Uehling TE and Wettstein HK (eds) *Midwest Studies in Philosophy, Studies in the Philosophy of Mind*, vol. X. Minneapolis, MN: University of Minnesota Press.

Churchland PS and Sejnowski TJ (1994) *The Computational Brain*. Cambridge, MA: MIT Press.

Copeland BJ (1997) The broad conception of computation. *American Behavioral Scientist* **40**: 690–716.

Cummins R (1983) *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Cummins R (1989) *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Cummins R and Cummins DD (eds) (2000) *Minds, Brains, and Computers: The Foundations of Cognitive Science*. Malden, MA: Blackwell.

Dawson MRW (1998) *Understanding Cognitive Science*. Malden, MA: Blackwell.
Devitt M (1981) *Designation*. New York, NY: Columbia University Press.
Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Dretske F (1986) Misrepresentation. In: Bogdan R (ed.) *Belief*. Oxford, UK: Oxford University Press.
Fodor JA (1975) *The Language of Thought*. Cambridge, MA: MIT Press.
Fodor JA (1987) *Psychosemantics*. Cambridge, MA: MIT Press.
Gazzaniga MS (ed) (2000) *Cognitive Neuroscience*. Malden, MA: Blackwell.
van Gelder T (1995) What might cognition be, if not computation? *Journal of Philosophy* **92**: 345–381.
Haugeland J (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* **2**: 215–260.
Haugeland J (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
Horst SW (1996) *Symbols, Computation, and Intentionality*. Berkeley, CA: University of California Press.
Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.
Macdonald C and Macdonald G (1995) *Connectionism: Debates on Psychological Explanation*. Oxford: Blackwell.
Mathews RJ (1994) Three-concept monte: explanation, implementation and systematicity. *Synthese* **101**: 347–363.
Millikan R (1989) Biosemantics. *Journal of Philosophy* **86**: 281–297.
Newell A (1980) Physical symbol systems. *Cognitive Science* **4**: 135–183.
Newell A and Simon HA (1976) Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery* **19**: 113–126.
Osherson DN (1990) *Invitation to Cognitive Science*, 3 vols. Cambridge, MA: MIT Press.
Palmer SE (1999) *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.
Putnam H (1988) *Representation and Reality*. Cambridge, MA: MIT Press.
Rumelhart DE, McClelland JL and the PDP Research Group (eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, vol. I. Cambridge, MA: MIT Press.
Searle J (1992) The critique of cognitive reason. In: *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
Von Eckardt B (1993) *What Is Cognitive Science?* Cambridge, MA: MIT Press.

Color Vision, Philosophical Issues about

Intermediate article

Alex Byrne, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
David R Hilbert, University of Illinois, Chicago, Illinois, USA

CONTENTS

The main philosophical issues about color concern whether objects have colors, and what sorts of properties colors are. Some philosophers hold that nothing is colored, others that colors are powers to affect perceivers, and others that colors are physical properties.

INTRODUCTION

According to our everyday experience, many things are colored. Roses are red and violets are blue. On the other hand, according to physical science, roses and violets are composed of colorless particles (or if not particles, something equally colorless). These two pictures of the world are not obviously compatible, and some thinkers have considered them to be plainly incompatible. Galileo, for example, thought that physical science had shown that objects are not really colored, and that colors are 'in the mind'. Philosophical theories of color in modern times have attempted either to reconcile the two pictures, or else to explain why one of them should be rejected.

Until recently, philosophers drew most of their data about color and color perception from their own experience of color. Although personal experience is a valuable source, a good deal of information relevant to abstract philosophical questions about color and the world as revealed by science is to be found in the work of color scientists. Many contemporary philosophers take the view that the physical, biological, and behavioral sciences place significant constraints on philosophical theories of color.

CENTRAL PHILOSOPHICAL ISSUES ABOUT COLOR

The Problem of Color Realism

If someone with normal color vision looks at a lemon in good light, the lemon will appear to have a distinctive property – a property that bananas and grapefruit also appear to have, and which we call 'yellow' in English. However, it does not follow from the fact that an object visually appears to have a certain property that the object has that property. To use an example known to the ancient Greeks, a straight oar half-immersed in water appears bent, but of course it does not have the property of being bent. Ordinarily we take for granted that lemons and other objects are as they appear, but in a philosophical mood we naturally wonder whether we are right to do so.

Such reflections give rise to the central philosophical problem concerning color, the problem of color realism. It is posed by the following two related questions. First, do objects like lemons, bananas and grapefruit really have the distinctive property that they appear to have? That is, are any objects yellow? Second, what is this property? Is it a physical property of some sort (for example, a certain way of reflecting light)? Or is it some kind of property that is specified partly psychologically (for example, a power to produce certain sorts of sensations)? One can ask similar questions for other color properties that objects appear to have.

These questions seem especially important when we consider the physically diverse nature of objects that look colored, and the way the visual system processes color information. As noted above, color is not a property attributed to objects in fundamental physics. In addition, the perception of color is a complex process and the relation between the physical properties of an object and the color it appears to have is far from straightforward. For example, it is not true that yellow-looking objects are those that reflect 'yellow light': there is no simple relationship between the light reaching the eye from an object and the color that object is perceived to have.

The problem of color realism is fundamentally a problem about perception – vision, in particular. The questions concern the nature of certain properties that objects visually appear to have, and whether reality is in these respects as it appears to be. The problem is not primarily about how we talk and think, although of course facts about color language and color concepts may be relevant. It is plausible to assume that the way we use color language is closely connected to the character of human color vision, but the problem of color realism itself concerns perception and not language. These points have to be stressed because it is always tempting to think that philosophical questions are basically questions about how words should be defined, and so are of little relevance to science. The problem of color realism is no more about the definitions of words than is the problem of why the dinosaurs disappeared.

The Representational Content of Visual Experience

It is helpful to put the problem of color realism in terms of what visual experience represents, its representational content. When someone has a visual experience, the scene before his or her eyes visually appears a certain way: for example, it might visually appear to someone that there is a yellow ovoid object in the bowl, or that two lines are the same length. In other words, the experience represents that there is a yellow ovoid object in the bowl, or that two lines are the same length. The experience represents the world correctly if there is a yellow ovoid object in the bowl, or if the two lines are the same length; otherwise, the experience represents the world incorrectly. In the former case the subject's experience will be veridical, and in the latter case illusory. For example, if the subject is looking at lines of the same length with appropriately oriented arrowheads on the ends (the Müller–Lyer illusion), the subject's experience will represent the lines to be of different lengths (even if the subject believes that they are of the same length). The experience represents the world incorrectly and is therefore an illusion.

In this terminology, the problem of color realism concerns certain properties (the colors) represented by visual experience, and whether such experiences represent the world correctly. There is a large philosophical (and, of course, psychological) literature on mental representation, and some philosophical discussions of color draw heavily on it (e.g. Boghossian and Velleman, 1991).

Why Color Matters to Philosophy

Although color is of interest in its own right, in philosophy it mainly serves as a tractable example that can be used to investigate problems of more general scope. One reason why color is particularly suitable for these purposes is that a great deal is known about the relevant physical properties of objects, and the way in which color information is extracted and processed.

One of these more general problems concerns the relation between appearance and reality – whether, or to what extent, the world is as it appears. This problem may be investigated in a reasonably manageable way by restricting attention to a specific instance of it, namely the problem of color realism.

There are a number of other philosophical problems that can be usefully addressed by focusing (not necessarily exclusively) on color. Examples include many central issues in the philosophy of perception: how to distinguish the various sensory modalities; the relationship between perception, thought, and action; and whether we see objects like lemons 'directly', or via our awareness of mental intermediaries. And the famous 'inverted spectrum' thought experiment, which (with some qualifications) supposes that objects that look green to me look red to you, and vice versa, has been used to illuminate a variety of philosophical topics, from the nature of consciousness (Block, 1990) to our knowledge of others' minds (Palmer, 1999).

PHILOSOPHICAL VIEWS AND THEORIES OF COLOR

Eliminativism

Eliminativism is the view that nothing is colored – at least, not ordinary physical objects like lemons. The eliminativist therefore regards much ordinary experience as erroneous. Historically, eliminativism has been the dominant philosophical view; it has its roots in the ancient Greek atomists.

Some eliminativists are 'projectivists': they hold that we 'project' colors that are 'in us' onto objects in our environment. According to the projectivist, some things are colored (for example, neural events, or mental entities like sensations or visual experiences), which we then mistakenly take for properties of objects like lemons. Projectivism is often found in psychology textbooks: Stephen Palmer (1999, p. 95), for instance, writes that 'color is a *psychological* property of our visual experiences when we look at objects and lights, not a *physical* property of those objects and lights'.

An obvious problem with projectivism is that the 'inner' things the projectivist says are colored do not have the right colors, if indeed they have any color at all. Nothing inside the brain becomes yellow when one is looking at a lemon, and it is hard to imagine how a visual experience could be colored.

But an eliminativist need not be a projectivist. One may simply take the view that nothing is colored, not even sensations or visual experiences.

The main line of argument for eliminativism proceeds by claiming that science has shown that objects like lemons do not in fact have colors. The surface of a lemon has a reflectance, various microphysical properties, and is disposed to affect perceivers in certain ways. No other properties are required to explain causally our perceptions of color. But the color properties, whatever they are, do causally explain our perceptions of color. So there is no reason to suppose that the lemon is yellow.

This argument represents a challenge to those who think that lemons are yellow but that this property is not to be identified with a reflectance (the percentage of light reflected by a surface), a microphysical property, or a disposition to affect perceivers (see the discussion of primitivism below). But it begs the question if one simply identifies the property yellow with (say) a reflectance.

The case for eliminativism therefore depends on showing that colors cannot be identified with properties that do causally explain our perceptions of color.

Dispositionalism

Dispositionalism is the view that colors are dispositions (powers, tendencies) to cause certain visual experiences in certain perceivers in certain conditions; that is, colors are psychological dispositions. (Strictly speaking we should add that, according to dispositionalism, at least sometimes our perceptions of color are veridical. This qualification should also be added to the three other views discussed below.) A simple version of dispositionalism is this: the property yellow is just the disposition to look yellow to typical human beings in daylight (for other versions, see Byrne and Hilbert, 1997).

Dispositionalism is often associated with the seventeenth-century English philosopher John Locke who, incidentally, also invented the 'inverted spectrum' thought experiment mentioned above. Locke, like other seventeenth-century philosophers, drew a distinction between 'primary' and 'secondary' qualities. Primary qualities have been characterized in a number of different (and often incompatible) ways, but the essential idea is that they comprise a set of fundamental properties in terms of which all material phenomena can be explained. For Locke, the primary qualities included shape, size, motion, and solidity. Because objects have certain primary qualities, they will be disposed to affect perceivers in certain ways: these dispositional properties are the secondary qualities. In this terminology, dispositionalism is the view that colors are secondary qualities. It is not clear whether Locke was himself a dispositionalist: his view sometimes interpreted as projectivism.

Poisonousness is a straightforward example of a dispositional property. To be poisonous is to be disposed to cause bodily harm if ingested or otherwise taken into the body. According to dispositionalism, yellowness is like poisonousness: to be yellow is to be disposed to cause certain visual experiences if placed in certain viewing conditions. The comparison with poisonousness reveals the relational nature of dispositionalism. Just as a substance may be poisonous for certain organisms and harmless to others, many dispositionalists hold that lemons are only yellow 'for us', and might even be blue relative to some other class of perceivers.

One objection to dispositionalism is that 'certain perceivers' and 'certain conditions' cannot be specified in a principled way. Perhaps a more fundamental problem is that it is not obviously well motivated. It is certainly plausible that yellow objects are disposed to look yellow (at least, once various qualifications are made). However, it is equally plausible that square objects are disposed to look square. It is not very tempting to conclude from this that squareness is a disposition to look square – as noted above, shape properties were supposed by Locke to be paradigmatic examples of properties that are not secondary qualities. It is not clear why color should be treated differently. The dispositionalist needs to explain why perceivers should be mentioned in the account of color, but not in the account of shape. Here the arguments are varied and complex. Dispositionalists often draw their arguments from Locke's own discussion of primary and secondary qualities.

The Ecological View

Thompson *et al*. (1992) have recently developed an 'ecological' view of color, inspired by J. J. Gibson. The ecological view rejects the orthodox account of

vision as 'inverse optics' – the process of extracting information about the scene before the eyes from retinal stimulation together with built-in assumptions about the environment. On the other hand, the ecological view stresses the connection between perception and action, and insists that the animal and environment should not be treated as 'fundamentally separate systems'; environmental properties are supposed to be partly 'constituted' by visual perception. Colors, in particular, 'are not already labelled properties in the world which the perceiving animal must simply recover. … Rather, colours are properties of the world that result from animal–environment codetermination' (Thompson *et al.*, 1992, p. 21).

The ecological view is perhaps best seen as a version of dispositionalism, identifying the colors with 'ecological-level dispositions' to affect perceivers (Thompson, 1995, p. 751). However, it must be emphasized that the ecological view's proponents see the comparison to traditional dispositionalism as somewhat superficial.

One obvious criticism of the ecological view is that it relies on controversial claims about perception. But there is a more basic difficulty, namely that crucial components of the theory are hard to understand. In particular, the meaning of the claim that colors are 'codetermined' by the perceiver and its environment is unclear.

Physicalism

Physicalism is the view that colors are physical properties of some kind, for example, microphysical properties, or reflectances.

Physicalism has been disputed on a number of grounds. First, it is argued that physicalism cannot account for the apparent similarities and differences between colors. In other words, the physicalist cannot explain the structure of phenomenal color space (Boghossian and Velleman, 1991).

Second, and relatedly, it is argued that physicalism cannot account for important observations about the way colors appear to us. For example, it is argued that physicalism cannot explain why orange is a 'binary' hue (every shade of orange is seen as reddish and yellowish), while yellow is a 'unique' hue (there is a shade of yellow that is neither reddish nor greenish) (Hardin, 1993).

Thirdly, it is argued that studies of color vision in animals show that there is no single kind of physical property (e.g. a reflectance) detected by color vision. So, since colors are whatever is detected by color vision, it is concluded that colors are not physical properties (Thompson, 1995).

Primitivism

According to primitivism, objects are colored, but the colors are neither dispositions to affect perceivers, nor physical properties (Yablo, 1995). In fact, the primitivist claims that there is no specially informative account of the nature of the colors. If primitivism is correct, the colors are analogous to fundamental physical properties, like the property of being electrically charged. Given the reductive cast of mind in cognitive science, it is not surprising that few cognitive scientists subscribe to primitivism.

Like eliminativism, primitivism is unmotivated if there are already good candidates to be the colors, for instance, physical properties of some sort, or psychological dispositions. The basic argument for primitivism, then, is similar to the argument for eliminativism: alternative explanations must first be eliminated.

There is little consensus on the best approach to the problem of color realism. The arguments for and against particular theories are rarely conclusive. But it would be wrong to think that no progress has been made. Philosophy is often advanced by showing that apparently unrelated theses are after all closely related, thereby forcing a proponent of, say, dispositionalism to take on a particular burden of commitments. Much recent work in the philosophy of color is of this kind.

INFLUENCE OF COGNITIVE SCIENCE ON ISSUES ABOUT COLOR

One notable feature of recent philosophical work on color is the attempt to integrate philosophical concerns with what is known empirically about color and color vision. Below we discuss a few areas in which this interaction either has been or has the potential to be especially significant.

Color Spaces and Opponent Processes

The colors stand in a complex web of similarity relations to each other. For instance, purple is more similar to blue than to green; and shades of red can be more or less similar to each other. Relations of color similarity also have an opponent structure. Red is opposed to green in the sense that no reddish shade is greenish, and vice versa; similarly for yellow and blue. Further, there is a shade of red ('unique red') that is neither yellowish nor bluish, and similarly for the three other 'unique' hues yellow, green and blue. Thus, in experiments summarized by Hurvich (1981), a

normal observer looking at a stimulus produced by two monochromators (light sources that emit in a narrow band of wavelengths) is able to adjust one of them until he reports seeing a yellow stimulus that is not at all reddish or greenish. In contrast, every shade of purple is both reddish and bluish, and similarly for the other three 'binary' hues orange, olive and turquoise. The binary hues are sometimes said to be 'perceptual mixtures' of the unique hues. These sorts of observations, supplemented with physiological data, form the basis of the opponent process theory of color vision. The main idea of this theory is that color perceptions are the result of two opponent processes (red–green and yellow–blue) and one non-opponent process (light–dark).

As mentioned above, two objections to physicalism start from these facts. We can now elaborate somewhat on the second of these. If physicalism is true, the objection runs, then the difference between the unique and binary hues must be explained in terms of 'unique' and 'binary' physical properties of objects like lemons and oranges. However, the correct explanation is in terms of neural opponent processes, and does not involve the physical properties of objects at all. This objection is controversial, but at least the opponent process theory has led philosophers to a renewed interest in understanding relations of similarity and difference among the colors.

Animal Color Vision

One active area of empirical research is comparative color vision: the study of color vision in non-human animals. Color vision is very widely distributed among animals: some degree of it appears to exist in all the major groups of vertebrates, and it is also common among invertebrates. But there is great variation in the precise type of color vision and the purposes for which it is used. Traditionally, philosophers have restricted their attention to human color vision, although this seems now to be changing. As mentioned above, the kinds of variation in color vision found in animals have been used to argue against physicalism, and also to support the ecological view.

Variation in Perceived Color

The perceived color of an object depends in complicated ways on both the illumination and the other objects in the scene before the eyes. As many people have discovered, even lightness relationships can be reversed by sodium lighting of the sort often found in parking lots. And interior decorators know that the perceived color of an object can depend on the color of its surroundings.

The perceived color also depends on the perceiver: color vision in human beings is surprisingly variable for a basic perceptual ability. Approximately 10 percent of men suffer from a substantial deficit of red–green color vision, and complete red–green color blindness is not rare. There is also substantial variation, among people classified as having 'normal' color vision, in, for example, the spectral location of unique green – ranging over 30 nanometers, nearly 10 percent of the visible spectrum. In addition to variation between subjects, there is variation within subjects. For example, the optical media of the eye, in particular the macula and lens, tend to yellow with age, producing shifts in perceived hues.

These facts have implications for philosophical theories of color. For example, dispositionalism appeals to a notion of 'certain perceivers'. Given the degree of normal variation, slightly different specifications of the privileged class of perceivers will lead to big differences in the resulting dispositionalist theories. The dispositionalist needs to give a principled reason for selecting one group rather than another to be the 'certain perceivers'. Similar remarks apply to the dispositionalist's 'certain viewing conditions'.

Color Constancy

As mentioned above, perceived color depends on the illumination. It is less well known that in many circumstances perceived color is relatively insensitive to changes in illumination. This feature of human (and animal) color vision is known as approximate color constancy. It implies that in many circumstances the perceived color of an object is more closely dependent on its (illumination-independent) surface properties than on the spectral character of the light reaching the eye. Some cognitive scientists have attempted to explain color constancy by treating color vision as a system that extracts information about the surface properties of objects, in particular their reflectances, from the light reaching the eye. This provides one of the inspirations for physicalist theories of color (Hilbert, 1987).

The Inverted Spectrum

It is often supposed that there is no reason to believe that people are 'spectrally inverted' with respect to each other. Indeed, under the influence of

the logical positivists of the early twentieth century, it used to be a common philosophical opinion that spectrum inversion makes no sense at all, because it is not 'verifiable'. However, with improved understanding of the genetic and physiological basis of color blindness, since the 1970s some color scientists have speculated that there might be actual cases of spectrum inversion in the human population. Red–green color blindness is caused by genetic defects that result in two of the three types of cones containing pigments that are very similar in their spectral sensitivity (their readiness to absorb light of different wavelengths). There are two types of red–green color blindness, depending on whether the spectral sensitivities of the two relevant cone types match that of the normal middle-wavelength-sensitive pigment or that of the long-wavelength-sensitive pigment. A person who has inherited the genes for both forms of red–green color blindness would have the two pigments switched from the normal condition. If the development of the neural circuitry for color vision depends only on the cell type and not on the pigment it contains, then this scenario should result in spectrum inversion. In fact, though, there does not seem to be the required independence between circuitry and pigment. Still, the fact that spectrum inversion is treated as a testable empirical hypothesis shows that it cannot be dismissed as a philosopher's fiction.

Metamers

Lights with quite different spectra, and surfaces with quite different reflecting characteristics, can appear to be identical in color. This phenomenon, known as metamerism, is a consequence of the fact that all the information available for perception of color derives from just three cone types with broad spectral sensitivity. If the light reaching the eye from two objects produces the same response in each of these three cone types then they will appear to have exactly the same color, no matter how their spectra or reflectances differ. This fact about (human) color vision has been significant in philosophical discussions of color since the 1970s. The central question is whether metamerism is incompatible with taking colors to be physical properties, because the phenomenon seems to show that there is no single spectrum (or reflectance) that all objects with a particular color share. For a variant of this argument see (Hardin, 1993, pp. 63–64); for a response see (Jackson, 1996).

References

Block N (1990) Inverted earth. *Philosophical Perspectives* **4**: 53–79.

Boghossian PA and Velleman JD (1991) Physicalist theories of color. *Philosophical Review* **100**: 67–106.

Byrne A and Hilbert DR (1997) Editors' introduction. In: Byrne A and Hilbert DR (eds) *Readings on Color*, vol. I 'The Philosophy of Color', pp. xi–xxviii. Cambridge, MA: MIT Press.

Hardin CL (1993) *Color for Philosophers: Unweaving the Rainbow (Expanded Edition)*. Indianapolis, IN: Hackett.

Hilbert DR (1987) *Color and Color Perception: A Study in Anthropocentric Realism*. Stanford, CA: CSLI.

Hurvich LM (1981) *Color Vision*. Sunderland, MA: Sinauer Associates.

Jackson F (1996) The primary quality view of color. In: Tomberlin J (ed.) *Philosophical Perspectives* vol. x, pp. 199–219. Cambridge, MA: Blackwell.

Palmer SE (1999) Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences* **22**: 923–943.

Thompson E (1995) *Colour Vision*. New York, NY: Routledge.

Thompson E, Palacios A and Varela FJ (1992) Ways of coloring: comparative color vision as a case study for cognitive science. *Behavioral and Brain Sciences* **15**: 1–74.

Yablo S (1995) Singling out properties. In: Tomberlin J (ed.) *Philosophical Perspectives* vol. IX, pp. 477–502. Atascadero, CA: Ridgeview.

Further Reading

Byrne A and Hilbert DR (1997a) *Readings on Color*, vol. I 'The Philosophy of Color'. Cambridge, MA: MIT Press.

Byrne A and Hilbert DR (1997b) *Readings on Color*, vol. II 'The Science of Color'. Cambridge, MA: MIT Press.

Hilbert DR (1992) What is color vision? *Philosophical Studies* **68**: 351–370.

Jackson F (1977) *Perception: A Representative Theory*. Cambridge, UK: Cambridge University Press.

Lewis DK (1997) Naming the colors. *Australasian Journal of Philosophy* **75**: 325–342.

Maund JB (1995) *Colours: Their Nature and Representation*. Cambridge, UK: Cambridge University Press.

McGinn C (1983) *The Subjective View: Secondary Qualities and Indexical Thoughts*. Oxford: Oxford University Press.

McGinn C (1996) Another look at color. *Journal of Philosophy* **93**: 537–553.

Stroud B (2000) *The Quest for Reality: Subjectivism and the Metaphysics of Colour*. New York, NY: Oxford University Press.

Tye M (2000) *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Westphal J (1991) *Colour: A Philosophical Introduction*. Oxford: Blackwell.

Computation, Philosophical Issues about

Intermediate article

Matthias Scheutz, University of Notre Dame, Notre Dame, Indiana, USA

CONTENTS

'Computation' is a cluster concept and has been characterized in many different ways (e.g. 'the execution of algorithms'). It underwrites philosophical analyses of what can be done in principle by a mechanism, and is intrinsically connected to the idea of manipulating symbols or representations by formal rules.

INTRODUCTION

The notion of computation is undoubtedly one of the very central, increasingly influential notions of our time. It has captured the attention of researchers from many disciplines for different reasons. In cognitive science it was the capacity of computers to process information that inspired cognitive psychologists to think of cognitive functions in terms of programs and of the brain as a computer running these programs. To be able to appreciate this view of cognition and the central role of computation within it, one needs a clear understanding of what 'computation' means and what computations are.

WHAT IS COMPUTATION?

An Intuitive Perspective

Like many widely used notions 'computation' does not have a single, clear-cut meaning, but rather, *qua* cluster concept, takes on different meanings depending on the context in which it is used. A glance in Webster's dictionary reveals the ordinary language conception of 'to compute': derived from the Latin 'com + putare' – to consider, it means something like 'to determine or to calculate especially by mathematical means'. However, this definition is rather vague and furthermore too restrictive to do justice to the variety of uses to which the notion of computation is put in computer science alone.

More to the point is defining computation as 'the execution of algorithms', which, in turn, puts the burden on the notion of *algorithm* and what 'executing an algorithm' means. Roughly speaking, an algorithm consists of a finite set of instructions, which operate on certain entities (symbols, representations of numbers, etc.) and can be *implemented* in some mechanism. To execute an algorithm then intuitively means to have the mechanism carry out the instructions for any given input in a deterministic, discrete, stepwise fashion (without resorting to random or analog methods and devices). The mechanism goes through a sequence of atomic steps in such a way that (one or more of) these steps correspond to some instruction, for all the instructions specified by the algorithm. Note that nothing is said about the nature of the mechanism yet: it could be concrete or abstract, natural or artificial. Depending on the kind of mechanism, the algorithmic specification will take different forms: in the case of computers, it is expressed in a *programming language*; in the case of humans, instructions may be given in ordinary language (as long as the individual steps are clearly distinguishable and described at a sufficient level of precision) – just think of cooking recipes or the instructions on public phones for making phone calls.

Computation defined as the execution of algorithms does not commit one as to what the computation is about or what it is supposed to achieve. Rather, it ties algorithmic descriptions to *mechanically realizable processes*. This leaves two issues to be addressed: first, it needs to be made clear what a mechanism is, and second, a precise specification of the notion of algorithm is required. The following brief historical overview reveals the origin of the idea of mechanism as well as that of using representations for calculations.

A Historical Perspective

The history of computation traces back to Leibniz and before, when daring philosophers pondered mechanical systems that could aid humans in performing calculations, and possibly even calculate by themselves without any assistance. The first functioning mechanical calculators were built in the seventeenth century and were composed of various mechanical parts (such as gears, cogs, etc.). Leibniz, having constructed calculators himself, was one of the first to envision an application quite different from their typical commercial and military use, that of 'mechanical reasoners'. His view that calculations, in particular, and logical reasoning (i.e. thinking), in general, could be *mechanized* lies at the heart of the notion of computation as used in cognitive science today.

Another crucial contribution to the modern notion of computation is also a product of that time (due to Descartes, Hobbes, Locke, and others), namely the idea that reasoning or, more generally, thinking involves *representations*. The mathematical practice of using marks and signs as representations in calculations became a paradigm for thought itself, as expressed by Hobbes' famous dictum that everything done by our mind is a computation (Pratt, 1987).

Computation was, therefore, already very much tied to the idea of mechanically manipulating representations, and prototypical manipulators were found in the mechanical calculators of those days. While many attempts were made at building mechanical calculators up to the end of the nineteenth century (with varying success: e.g. see Williams, 1997), the computing capabilities of these machines remained very modest. It was only the twentieth century that witnessed major progress in the construction of computers and the conception of computing. This was largely due to two quite independent developments: (1) the thorough logical analysis of the notions 'formal system' and 'formal proof' (leading to further studies of notions such as 'effectively computable function' and 'algorithm'), and (2) the rapid progression in the engineering of electronic components (from vacuum tubes, to transistors, to integrated circuits, and beyond).

A Logico-philosophical Perspective

In the 1930s, logicians laid the main philosophical groundwork for a well-defined formal notion of computation in their attempt to make the intuitive notion of computation, then called 'effective calculability', formally precise. Being logicians, they were solely concerned with the class of functions (over the positive integers) that can be effectively calculated *in principle* – besides, digital computers did not exist yet. Church (1936) was the first to give this class of effective calculable functions a formal characterization through a definition postulate, which later came to be known as 'Church's Thesis' (CT): 'We now define the notion … of an effectively calculable function of positive integers by identifying it with the notion of a recursive function on positive integers (or of a λ-definable function of positive integers)' (Church, 1936, p. 356). Note that by virtue of relating an intuitive notion and a formal notion, CT cannot be proved in principle, as mentioned by Church himself: 'This definition is thought to be justified by the considerations which follow, so far as positive justification can ever be obtained for the selection of a formal definition to correspond to an intuitive notion' (Church, 1936, p. 356).

While this was a first step to capture the meaning of 'computable', it was not quite satisfactory, as CT is silent about what 'effectiveness' of a calculation means. As it stands, the notion of 'effectively calculable function' implies that two ingredients are needed to understand computation: a notion of 'effective procedure or algorithm' and a notion of 'function computed by an algorithm'. The latter can be straightforwardly explicated: it is the mapping obtained by pairing all possible inputs with the corresponding outputs resulting from applying the algorithm to them. The former, however, received a satisfactory account only after Turing (1936) had introduced his machine model of a 'computer', which resulted from his analysis of the possible processes a human – what he then called 'the computer' – can go through while performing a calculation using paper and pencil applying rules from a given finite set. It was crucial to Turing's conception of computation that the human computer follow the rules 'blindly', that is, without using insight or ingenuity. In his analysis of the limitations of the human sensory and mental apparatus five major constraints for doing 'automatic computations' crystallize: (1) only a finite number of symbols can be written down and used in any computation; (2) there is a fixed bound on the amount of scratch paper (and the symbols on it) that a human can 'take in' at a time in order to decide what to do next; (3) at any time a symbol can be written down or erased (in a certain area on the scratch paper called 'cell'); (4) there is an upper limit to the distance between cells that can be considered in two consecutive computational steps;

(5) there is an upper bound to the number of 'states of mind' a human can be in, and the current state of mind together with the last symbol written or erased determine what to do next.

Turing then defined a mathematical model of an 'imagined mechanical device' that satisfies all of the above, later referred to as a 'Turing machine' (TM) by Church. A TM consists of an unbounded tape divided into squares, each of which can hold exactly one symbol, a tape head for reading and writing symbols from a given alphabet on the squares, and a controller, which is in exactly one of finitely many states at any given time. Each computational step of the machine first involves reading the symbol under the tape head and then, depending on the current state of the controller, writing a new symbol on the square, possibly switching to another state and possibly moving the tape head one square to the left or to the right. 'The computation proceeds by discrete steps and produces a record consisting of a finite (but unbounded) number of cells, each of which is blank or contains a symbol from a finite alphabet. At each step the action is local and is locally determined, according to a finite table of instructions' (Gandy, 1988, p. 81). This way the TM became a model of human computing, an *idealized* model to be precise, since it could process and store *arbitrarily long, finite sequences of symbols*. The TM model is also a very abstract model, for it only captures high-level processes that take place in humans when they compute (as opposed to low-level neuronal processes, for example).

Turing intended his analysis to show that *any function computable by a human being following fixed rules can be computed by a TM*. And, furthermore, he also believed the converse, that every function computed by a Turing machine could (in principle) be computed by a human computer. Note that this equivalence *per se* does not preclude humans from being able to find answers to problems (expressed in terms of functions) which no TM can compute (e.g. using intuition).

PHILOSOPHICAL VIEWS OF COMPUTATION

Turing Computability and Beyond

The logico-philosophical analyses of the intuitive notion of computation led to the crucial insight that different attempts to characterize it can all be proven extensionally equivalent: recursive functions, λ-definable functions, and TM-computable functions all define the same class of functions. These equivalence results are possible, because what 'computing' means with respect to any of the suggested formalisms is expressed in terms of functions from inputs to outputs, which are used as mediators in the comparison of the various classes of functions defined by the different formalisms. Later, other formalisms such as Markov algorithms, Post systems, universal grammars, PASCAL programs, as well as various kinds of automata, were also shown to give rise to the same class of functions. Hence, by CT, any of the above mentioned formalisms captures our intuitive notion of computation, that is, *what it means to compute*. (Some disagree with this conclusion, arguing that the equivalence results capture only a restricted notion of computation as shared by certain philosophers of mathematics and logicians, e.g. Sloman (1996).)

Common to all the above computational formalisms (besides their attempts to specify formally the intuitive notion of 'computation') is their property of being independent from the physical: computations in any of these formalisms are defined *without* recourse to the nature of physical systems that (potentially) realize them. Even the TM model, which is often considered the prototype of a 'mechanical device', does not incorporate physical descriptions of its inner workings, but abstracts from the mechanical details of a *physical realization*. The first to incorporate physically motivated mathematical constraints into a formal model of computation was Gandy (1980) in his attempt to define a notion of computation for any discrete, deterministic, physical machine. He formulated five conditions to determine whether any system qualifies as a 'mechanical machine' and proved that any function computable by a discrete deterministic device (in his sense) is effectively computable and vice versa. Hence, TM-computability (i.e. effective computability) and computability by mechanical devices are equivalent notions. Some even extend the claim by suggesting that the behavior of any finitely realizable physical system can be 'computed' (in the sense of 'perfectly simulated') by a TM (e.g. see Deutsch, 1985).

It is not clear, however, whether computation should be equated with 'effective computability', since there are, at least in principle, imaginable computing devices that give rise to 'Super Turing computability' (i.e. compute functions that no TM can compute). An example of such a device is Turing's 'oracle machine' (O-TM), which is a TM with additional atomic operations to query an 'oracle'. The oracle itself is a device that somehow produces values of a particular (possibly TM-uncomputable)

function – how the results are obtained is left unspecified. It is easy to see that any O-TM with an oracle for any uncomputable function can compute more functions than any TM. Whether such a machine could be physically realized is an open question (maybe there are physical quantities that happen to encode some TM-uncomputable function). The interesting point is simply that an O-TM would be perfectly mechanistic in the classical sense without being effective as it uses some noneffective device, namely the oracle. O-TMs, hence, drive a wedge between the notions of 'effectiveness' and 'mechanism' (e.g. see Copeland, 2000). A similar point can be made with respect to the notions of 'effectiveness' and 'algorithm'.

There are other suggestions along the same lines coming from neural network research: it can be shown, for example, that certain neural networks (consisting of about 1000 neurones) with rational-valued connection weights between neurones can compute any TM-computable function (Siegelmann and Sontag, 1995). And if real-valued weights are allowed, they can compute any function whatsoever.

Other Construals of Computation

Although TMs have become the canonical models of computation and permeate various academic disciplines in that role, there are other construals targeted more towards possible philosophical merits and potential practical applications of computation. Following Smith (forthcoming), for example, the following views should all be distinguished as they emphasize and capture different aspects of computation:

1. *formal symbol manipulation*: the manipulation of symbols by virtue of their formal properties (without regard to possible interpretations or semantic content);
2. *effective computability*: what can be done effectively by a mechanism;
3. *rule-following* or *execution of an algorithm*: what is involved in following rules or instructions;
4. *finite (digital) state machines*: automata with a finite set of internal states;
5. *information processing*: what is involved in storing, manipulating, and displaying information;
6. *interactive systems*: computation as interaction and communication embedded in an environment;
7. *dynamical systems*: computation expressed in the language of dynamic systems (using concepts like state space, trajectory, attractors, etc.).

To some extent all of the above notions play a role in various disciplines (especially in computer science), but some of them are more dominant in specific intellectual areas: (1) figures mainly in philosophical debates and meta-mathematics, where (2) and (3) are tied to logical investigations; (4) is largely an engineering concept, while (5)–(7) have become increasingly important in cognitive science, the theory of complex systems, and, of course, computer science.

While the above list is far from exhaustive, it is intended to give a flavor of the wealth of different aspects the notion of computation has acquired, especially in the course of the last century. For that very reason, it is argued, we are still lacking the 'grand unified theory' (similar to physics) that can accommodate all these multiple facets, if such a theory is possible in the first place.

Real-life Computation

Despite the theoretical success of TM-computability, computer science *qua* practice is concerned not so much with the limits of what can be computed in theory, but rather with the more modest, mundane question of what can be computed within reasonable limits (using given resources). A whole new discipline within computer science called 'complexity theory' – an offspring of the classical investigations of effective computability – is dedicated to the study of what is computationally feasible. Still other issues arise from computational practice with which the TM model, for example, can hardly cope, in particular, the need for computational systems (embedded in various kinds of devices) to continually interact with their environments: what function does an operating system compute (or the world wide web, for that matter)? According to the classical view, such questions cannot be answered easily as the underlying functions are simply not defined for inputs on which computational processes run forever. (A simple example of a program that loops forever on all of its inputs is the following control code of a 'router' for the internet, which simply copies messages from its input to its output port.) Yet, there is a strong intuition that computational processes as they occur in operating systems or web browsers do have a purpose, can accomplish certain tasks or fail at achieving them. As a consequence the notion of 'computation of a function', and with it the classical notion of algorithm, had to make room for the notion of interaction:

> Interaction is shown to be more powerful than rule-based algorithms for computer problem-solving, overturning the prevalent view that all computing is expressible as algorithms. The radical notion that

> interactive systems are more powerful problem-solving engines than algorithms is the basis for a new paradigm for computing technology built around the unifying concept of interaction. (Wegner, 1997)

This paradigm shift from programs to processes renders many of the old reservations to the notion of computation obsolete, which were a consequence of taking computation to be defined solely in abstract syntactic terms thereby abstracting over physical realization, real-world interaction, and semantics. The new approach reveals computation, contrary to standard orthodoxy, as interactive and embodied, hence very much concerned with the constraints imposed on computational processes by the real world.

ROLE OF COMPUTATION IN COGNITIVE SCIENCE

The Midwife: Computation and the Birth of Cognitive Science

The independence of computations (in the sense of TM-computations) from their physical realizers was one major source of attraction for cognitive psychologists in the late 1950s. The information-processing capabilities of computers, an ability thought to underlie human cognition, and the potential of computer programs to specify exactly *how* information is processed was another. Together they led to the thought that cognition, viewed as 'the processing of information', could be completely understood and explained in terms of computations: if cognitive functions *are* computations, then explanations of mental processes in terms of programs are scientifically justifiable without having to take the 'implementing' neurological mechanisms into account, similar to computers, where it is the programs implemented on the computer hardware, not the hardware itself, that explain (if not entirely, then at least for the most part) what the computer does. The *computer metaphor* implicit in this view has been summarized as the claim that 'the mind is to the brain as the program is to the hardware' (Johnson-Laird, 1988) (note that this should really read 'the mind is to the brain as *computational processes* are to the hardware' to avoid conflating the program–process distinction). Its guiding ideas eventually became so prominent (originally in psychology, later in artificial intelligence) as to assist in the birth of cognitive science and establish *the computational claim about mind*, also called *computationalism*, as a genuine research paradigm.

The Paradigm: Computation and the Computational Claim about Mind

As with the notion of computation, computationalism is not a unified view, but construed differently by philosophers, psychologists, or neuroscientists. Various condensed slogan-like phrases such as 'the brain is a computer', 'the mind is the program of the brain', or 'cognition is computation' can be found in the literature, to name just a few. Yet, they cannot be taken at face value, for if they were read together, they would equivocate essentially distinct notions (such as program and process, mind and cognition, etc.). Furthermore, depending on their subdiscipline within cognitive science, researchers stress different aspects of computations: their information-processing capabilities, their formal nature, their control functions, their potential to have semantics, and so on.

Common to different views of computationalism are the assumptions that (1) mental processes are computational processes and (2) the same kind of relation that obtains between programs and computer hardware (i.e. the *implementation* relation) obtains between mental descriptions and brains too. It follows that cognitive functions can be described by and explained in terms of programs, and that the right level of abstraction at which to understand cognition is the computational level and not the level of the implementing mechanism (i.e. the brain), even though it might be helpful to know the functional organization and role of certain brain areas in determining what they implement.

Computationalism has many appealing facets, especially when it comes to high-level cognition: many features related to logic and language (such as systematicity, productivity, compositionality, and interpretability of syntax or the compositionality of meaning, e.g. see Fodor and Pylyshyn, 1988) are supported by computations 'almost for free', and many mental operations on various kinds of representations such as rotating three-dimensional images, searching for a good move in a chess game, reasoning about other people's behavior, planning a route through a city avoiding construction sites, etc. can be described computationally and implemented on computers. After all, this is what computers do: they manipulate symbol tokens (e.g. strings of bits), some of which are representations of the subject matter the computation is about. These representations, in turn, have both formal and semantic properties, of which the former are causally efficacious. Computational processes then manipulate symbols by virtue of their formal and not their semantic properties (e.g. Fodor, 1981).

While computationalists take this to be a virtue of their approach, it is a major shortcoming for others and various arguments have been advanced to establish that formal symbol manipulation is not sufficient for human intentionality and semantics (e.g. Searle's Chinese Room, 1980) or that minds are not TM-computable (e.g. Lucas's Gödelian argument, 1961). More recently, connectionists and dynamicists have tried to replace the notion of computation with alternatives, arguing that the representational level of description of a cognitive system so crucial to computationalism cannot be taken for granted. In fact, most dynamicists find the symbolic/representational level of description superfluous altogether and argue instead for an explanation of cognition in terms of *dynamic systems* (e.g. Port and van Gelder, 1995).

The Method: Computation and the Simulation of Cognition

While there are undoubtedly tendencies in cognitive science to replace the classical notions of computation, either by dynamic systems or by more adequate notions of computation (e.g. notions based on interaction, real-time constraints, etc.), even those opposed to computationalism agree at least that computation is still a valuable tool in the study of cognition (regardless of its explanatory success). In particular, computer simulations and computational models (of aspects) of cognition have become increasingly important in cognitive science. While computational models, at least to some extent, presuppose that whatever is modeled is computational, simulation models do not have to make such an assumption. Rather, they implement a computational approximation of the mathematical description of the phenomenon under scrutiny, and as long as any resultant error is within predetermined bounds the simulations are considered 'models'. In particular, they might elucidate complex dynamical relations between various parts of the simulated model, which are difficult to see (and often not 'visible' at all) from the formal, mathematical description. From trajectories through complex state spaces of dynamic systems to evolutionary processes in artificial environments, computer simulations provide a testbed for cognitive scientists to evaluate their hypotheses without always having to study them in 'real systems'. Furthermore, simulations can focus on different aspects of cognitive systems at different levels of description, they can be reproduced, slowed down, sped up, and modified in various other ways (e.g. simulating damage, disease, and various other disorders), in which no real cognitive system could be manipulated while preserving its normal functionality (obviously, ethical considerations would enter the picture here as well).

A crucial difference between simulation and computational models is, however, that the former usually does not share all the relevant causal properties with the modeled system, whereas the latter, being a computational model of computational processes, can in principle have the right causal structure (depending on various constraints on inputs, outputs, real-time, etc.).

SUMMARY

'Computation' is a multifarious notion, which defies a single, simple characterization. Yet, it is often explicated as 'executing an algorithm', presupposing some sort of mechanism able to 'execute' instructions as specified by the algorithm. For many logicians and philosophers it was the notion of Turing machine computability that for the first time gave precise meaning to the intuitive notion of computation understood as 'blindly following rules or instructions', thereby answering the question of what 'effective calculability' is supposed to mean. The connection of 'effectiveness' and computation goes back at least to the seventeenth century, when 'calculation' was very much tied to mechanical devices. Only in the twentieth century did effectiveness, mechanism, and computation become separated, when alternative models of computations such as interactive systems were considered. Various construals of the notion of computation (such as 'formal symbol manipulation' or 'information processing') emphasize different aspects of computation, although none of them seems to capture what computation may signify in its entirety. In cognitive science, computation played a crucial role right from the start. It figured prominently in the emergence of the discipline and became the basis of computationalism, the paradigmatic view that mental processes are computational, leading to the development of computational models of cognitive functions. Even for researchers objecting to computationalism, computations can be of great utility when used to simulate cognitive processes.

References

Church A (1936) An unsolvable problem of elementary number theory. *American Journal of Mathematics* **58**: 345–363.

Copeland BJ (2000) Wide vs. narrow mechanism. *Journal of Philosophy* **97**: 5–32.

Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. *Proceedings of the Royal Society*, Series A, **400**: 97–117.
Fodor JA (1981) *Representations*. Cambridge, MA: MIT Press.
Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.
Gandy R (1980) Church's thesis and principles for mechanism. In: Barwise J, Keisler HJ and Kunen K (eds) *Proceedings of the Kleene Symposium*, pp. 123–148. New York, NY: North-Holland Publishing Company.
Gandy R (1988) The confluence of ideas in 1936. In: Herken R (ed.) *The Universal Turing Machine: A Half-Century Survey*, pp. 55–111. Berlin: Kammerer & Unverzagt.
Johnson-Laird PN (1988) *The Computer and the Mind*. Cambridge, MA: Harvard University Press.
Lucas JR (1961) Minds, machines, and Gödel. *Philosophy* **36**: 122–127.
Port R and Van Gelder T (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.
Pratt V (1987) *Thinking Machines – The Evolution of Artificial Intelligence*. Oxford, UK: Basil Blackwell.
Searle J (1980) Minds, brains and programs. *The Behavioral and Brain Sciences* **3**: 417–424.
Sloman A (1996) Beyond Turing equivalence. In: Millican PJR and Clark A (eds) *Machines and Thought: The Legacy of Alan Turing*, vol. I, pp. 179–219. Oxford, UK: Clarendon Press.
Siegelmann HT and Sontag ED (1995) On the computational powers of neural nets. *Journal of Computer System Sciences* **50**: 132–150.
Smith BC (forthcoming) *The Age of Significance. An Essay on the Foundations of Computation and Intentionality*, vols I–VII. Cambridge, MA: MIT Press.
Turing AM (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, Series 2, **42**: 230–265.
Wegner P (1997) The paradigm shift from algorithms to interaction. *Communications of the Association for Computing Machinery* 1997 **40**(5).
Williams MR (1997) *A History of Computing Technology*, 2nd edn. Los Alamitos, CA: IEEE Computer Society Press.

Further Reading

Cleland CE (1993) Is the Church–Turing thesis true? *Minds and Machines* **3**: 283–312.
Copeland BJ (1996) What is computation? *Synthese* **8**(3): 335–359.
Davis M (1958) *Computability and Unsolvability*. New York: McGraw-Hill Book Company.
Dietrich E (1990) Computationalism. *Social Epistemology* **4**(2): 135–154.
Gardner H (1985) *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.
Haugeland J (1985) *Mind Design I*. Cambridge, MA: MIT Press.
Haugeland J (1996) *Mind Design II*. Cambridge, MA: MIT Press.
Herken R (ed.) (1988) *The Universal Turing Machine: A Half-Century Survey*. Berlin: Kammerer & Unverzagt.
Hopcroft JE and Ullman JD (1979) *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley Publishing Company.
Searle J (1992) *The Rediscovery of Mind*. Cambridge, MA: MIT Press.
Smith BC (1996) *The Origin of Objects*. Cambridge, MA: MIT Press.
Sterelny K (1990) *The Representational Theory of Mind*. Oxford, UK: Blackwell.
Van Gelder TJ (1998) The dynamical hypothesis in cognitive science. *The Behavioral and Brain Sciences* **21**: 615–665.
Webb J (1980) *Mechanism, Mentalism, and Mathematics: An Essay on Finitism*. Boston, MA: Reidel.

Concepts, Philosophical Issues about

Intermediate article

Jesse Prinz, Washington University, St Louis, Missouri, USA

CONTENTS

Concepts are generally defined as representations that allow us to think about properties or categories. Philosophers have debated the nature of these representations and the cognitive functions that they serve.

INTRODUCTION

Concepts (sometimes called 'ideas') are the tools by which we think about the world. They have been an object of philosophical scrutiny since Plato's time. New theories of concepts have been developed in recent decades, arising from interactions between philosophers and researchers in other disciplines. There is no consensus about which theory of concepts is correct, but there is wide agreement on the challenges that an adequate theory of concepts must meet.

FUNCTIONS OF CONCEPTS

Concepts are theoretical posits, introduced to play a variety of explanatory roles. This article will present the functions that concepts are most often alleged to serve. Some theorists doubt whether a single kind of entity can serve all of these functions (Rey, 1983), but others are more optimistic (Prinz, 2002). If all these different functions are served by different kinds of entities, there may be no non-arbitrary way to determine which of those entities deserve to be called concepts and the utility of the construct may be called into question.

One function that is often emphasized in philosophy involves reference: concepts are said to represent categories or properties. A category is a class of things consisting of zero or more members (e.g. the class of all actual and possible elephants); and a property can be characterized as that by virtue of which things form a cohesive class (e.g. elephanthood). Some philosophers think that concepts must represent properties rather than categories, because we can have distinct concepts corresponding to distinct properties of the same class (e.g. the class of triangles is the same as the class of trilaterals, but we can distinguish between these conceptually).

Concepts are also widely (though not universally) alleged to serve an epistemic function: they embody the information by which we understand categories. That information is often believed to take the form of 'features'. Features are usually construed as concepts themselves, designating properties possessed by category members. For example, an *elephant* concept may encompass the features *has a trunk* and *is an animal*. If concepts are associated with features, they can serve a categorization function. I determine that Jumbo is an elephant because he is an animal, has a trunk, and so on. Psychologists emphasize the categorization function above all others.

There are two main views about how concepts embody knowledge or information. According to 'decomposition' views, concepts are literally parts of other concepts (e.g. Smith, 1989). For example, an *elephant* concept might be construed as a data structure containing *trunk*. According to 'conceptual role' views, concepts are related to each other by inputs and outputs (e.g. Block, 1986). For example, an *elephant* concept might be construed as a mental predicate ('*X* is an elephant') that licenses certain inferences ('*X* has a trunk').

Decomposition and conceptual role views may be behaviorally indistinguishable. Defenders of decomposition views generally assume that constituent features can be used to draw inferences. If my *elephant* concept contains the feature *trunk*, and I believe that Jumbo is an elephant, I may infer that Jumbo has a trunk. On either approach, then, concepts can be said to serve an inference function.

Concepts also have a combination function: they combine to form compound concepts and thoughts. We often form thoughts that have never been entertained before, even when we have not acquired any new concepts. This suggests that concept combination is a compositional process. Concepts are said to combine compositionally when the content of a compound concept is a function of the concepts that comprise it together with the rules of combination. For example, one can form a concept of a clumsy elephant if one has an *elephant* concept and a concept of clumsiness, even if one has never considered or encountered clumsy elephants before. A compositional system can generate boundless novel compounds from a finite set of more basic concepts.

Concepts are also said to serve functions related to language. Firstly, they contribute to communication. Two people successfully communicate when they associate the same (or similar) concepts with their words. The communication function is related to a meaning function. The simplest version of this thesis says that concepts constitute the meanings of words. Philosophers have developed different theories of meaning, which lead to different interpretations of this hypothesis. 'Meaning-as-use' theories, for example, associate meanings with abilities. According to one version, the meaning of a word is determined by how that word is used in various conversational contexts (Wittgenstein, 1953). If concepts are meanings, it would follow that concepts are verbal abilities; and if concepts are verbal abilities, then one cannot possess a concept without language. On this view, creatures lacking language lack concepts.

Many philosophers reject that conclusion. It makes it more difficult to talk about apparent cognitive similarities between humans and non-human animals. It also makes it difficult to explain how language is acquired in the first place. How can one learn what our first words mean if understanding a meaning requires mastery of verbal abilities (Fodor, 1975)? Not all meaning-as-use theories face these difficulties. According to another meaning-as-use theory, the meaning of a word is determined by how a concept underlying that word is used in thought. This preserves the idea that knowing meanings involves mastering abilities, while denying that those abilities are necessarily verbal.

Proponents of a different class of theories equate the meaning of a term with a 'mode of presentation' (Frege, 1893). A mode of presentation can be a representation of features possessed by whatever the term represents. For example, the meaning of the word 'zebra' might be comprised of a representation of striped, horse-like animals roaming the African savannah. If concepts decompose into features, they are ideally suited to serve as modes of presentation and, thus, meanings. Even conceptual roles can be regarded as modes of presentation if the latter term is very broadly construed.

According to another theory, the meaning of a word is exhausted by its referent. The word 'zebra' simply means the class of all actual and possible zebras or the property of being a zebra. If this theory is correct, it makes little sense to identify concepts with meanings. It would be better to say that concepts are the bearers of meanings. More specifically, one might say that the meaning of a word is exhausted by the referent of the concept with which it associated. If meanings are exhausted by referents, then the features associated with a given concept are not part of linguistic meaning, as mode of presentation theories and meaning-as-use theories imply.

THEORIES OF CONCEPTS

Philosophers and psychologists have proposed a number of theories of concepts. One point of disagreement concerns the question of which concepts are 'primitive' (Fodor, 1981). As noted above, some theorists assume that many concepts are comprised of other concepts. On pain of regress, decomposition must end somewhere. A primitive concept is one whose identity conditions do not depend on any other concept. The British empiricists claimed that primitive concepts are sensory. This view has fallen out of favor, because it is very difficult to analyze abstract concepts into sensory features (but see Prinz, 2002). The issue of primitive concepts has received relatively little attention in recent times.

Another point of disagreement concerns the ontological status of concepts (Peacocke, 1992). According to some philosophers, concepts are abstract entities; while according to others they are mental tokens that reside inside our heads. There may be room for reconciliation here, however. If concepts are mental tokens, and if concept sharing is possible, it must still be possible to talk about them belonging to common types: two tokens of an *elephant* concept must be capable of belonging to a common type. Conversely, those who favor the view that concepts are abstract entities must admit that individuals grasp concepts, and, to the extent that grasping is a psychological process, token mental states must be involved.

A more divisive issue concerns the kind of information that concepts embody. One terminological

caveat must be made before we survey these disputes: the word 'concepts' is sometimes used as shorthand for 'the majority of lexical concepts'. A lexical concept is a concept expressed by a single word in a natural language, such as English. A theory of concepts would thus be better described as a theory of the kind of information embodied in most of our lexical concepts.

The Definition Theory

Overview

The dominant theory in the history of philosophy claims that most lexical concepts are definitions (e.g. Katz and Fodor, 1963). A definition is a collection of features that are jointly sufficient and individually necessary for membership in a category. The concept *bachelor*, for example, is said to entail or decompose into the features *unmarried* and *male*.

Adherents of the definition theory (also called the 'classical' theory) draw a sharp distinction between those features that define a concept and those that are merely known to apply to the category designated by the concept. This generates two kinds of true sentences, those that are true by definition (e.g. 'bachelors are unmarried') and those that are true by virtue of how the world is (e.g. 'bachelors tend to like martinis'). The former are called 'analytic' and the latter 'synthetic'.

There are several different versions of the definition theory. According to the tradition deriving from Plato, each of us implicitly knows how to define the concepts we possess, but that information is not easily accessed. One must engage in arduous philosophical reflection (facilitated by dialogue) to reveal definitions. This process of discovering definitions is said to be *a priori*, because it relies on conceptual intuition rather than observation of experience. *A priori* conceptual analysis is one of the predominant methods used in philosophy.

According to a second version, definitions are discovered *a posteriori*, not by intuition and reflection (Rey, 1983). This account is thought to be most applicable to concepts that designate natural kinds (such as concepts of animals or substances found in nature). The definition of the concept *gnu* might be a description of the gnu genome, or some other scientifically determined conditions that are necessary and sufficient for being a gnu. One might hold a mixed view, according to which some definitions are discovered *a posteriori* and others are discovered *a priori*.

A third version of the definition theory has it that definitions are not discovered at all, but rather invented (Carnap, 1934). Defenders of this version recognize that many of our concepts are unclear, or variable between individuals. Communication is greatly facilitated, they argue, by stipulating definitions. These may roughly coincide with prior intuitions, but they are more precise. The primary motivation for this program of 'precisification' is to facilitate science. If different scientists agree on how their terms are defined, any remaining debates between them must be substantive (e.g., pertaining to data) rather than verbal.

Assessment

Despite its long philosophical pedigree, faith in definitions began to wane in the middle part of the twentieth century. One attack derives from Quine's (1951) famous critique of the distinction between the analytic and the synthetic. Believers in that distinction traditionally hold that analytic truths are known *a priori* and are invulnerable to empirical refutation.

Quine presents an alternative account of how sentences are confirmed, which contradicts this assumption. According to Quine, no sentence is empirically verified in isolation. When we encounter evidence that conflicts with a sentence held to be true, we have the option of either revising that sentence or revising various background assumptions, including those that may appear independent of experience. What we revise will depend on what we observe, what we take to be true, and a variety of pragmatic principles by which we strive to minimize disruption as we revise our theories. Because sentences face the 'tribunal of experience' collectively, any sentence is vulnerable to empirical refutation.

If Quine's 'confirmation holism' is correct, the traditional notion of analyticity is undermined. This has implications for the definition theory of concepts. Proponents of that theory say that some of the features associated with a concept define it and some do not. But which are the definitional features? They cannot be the features that are analytically related to the concept, because that notion lacks foundation. Without a principled way to distinguish defining features from collateral information, the definition theory is in jeopardy of being usurped by 'concept holism' (see below). This is damaging to Carnap's 'precisification' view. Science can stipulate definitions, but these will remain sensitive to the pressures of observation and discovery. No claim is purely verbal.

Another challenge to the definition theory was articulated by Wittgenstein (1953). Wittgenstein notices that certain terms cannot be captured by a

single set of necessary and sufficient conditions. His famous example was the word 'game'. Every plausible defining condition on games has obvious counterexamples. For instance, games do not always have winners (consider catch), and they do not always have two or more sides (consider solitaire). This point may generalize. Every time one philosopher publishes an analysis of a concept, another publishes a convincing example of a case that the analysis fails to subsume. Even concepts that seem to have incontrovertible definitions are vulnerable. A *bachelor* is widely defined as an unmarried male, but there is a strong intuition that Catholic priests are not bachelors even though they satisfy the definition. Conditions thought to be defining often capture a salient range of cases, but they rarely capture all cases.

Psychologists have also criticised the definition theory (see Hampton, 1993, for a review). Experimental evidence has overwhelmingly shown that definitions do not figure prominently in conceptual tasks. People categorize on the basis of family resemblance, associate non-defining features with concepts, have difficulty learning definitions, and fail to rely on definitions once they have been learned. If concepts are mental representations, then it seems unlikely that concepts are definitions.

Defenders of the *a posteriori* version of the definition theory have a response to such arguments. On their view, definitions are scientifically discoverable facts unknown to most concept users. The fact that people have difficulty formulating adequate definitions only shows that most of us do not understand completed science. This might rescue the view that concepts are definitions, but it would not satisfy many cognitive scientists. If definitions are largely unknown, they will play little role in behavior. Even if concept users believe that definitions will someday be discovered, that article of faith cannot distinguish my *elephant* concept from my *walrus* concept. If a theory of concepts is to explain how different concepts affect behavior, then we would be better off identifying concepts with the actual knowledge that ordinary people use to categorize elephants and walruses.

Clusters, Stereotypes, and Prototypes

Overview

The arguments against the definition theory have spawned several alternative theories. Wittgenstein came to see many concepts as revolving around large clusters of features. Rather than having a universal essence, concepts like *game* are applied by determining whether something has a preponderance of the features in a cluster. All games exhibit a subset of features from the same cluster, but few if any cluster features are exhibited by all games. Wittgenstein calls *game* a 'family resemblance' concept, because each of its instances shares features with each other, but no features are universally shared.

Putnam (1975) explores a related idea. He argues that people think about categories by means of stereotypes. A stereotype is a collection of features that are thought to be highly characteristic of category instances. Such features are typically widespread, salient, and diagnostic. For example a *dog* stereotype might include such features as *barks*, *has a tail*, and *is a pet*. Stereotype features are also presumed to be contingent. There are wild dogs that lack tails and never utter a sound. Stereotypes can even contain features that are erroneously associated with categories (e.g., gorillas are stereotyped as ferocious). Putnam observes that when people explain the meaning of a word, they often list stereotypical features, rather than provide a definition. Putnam thinks that stereotypes are generally concise, in contrast to Wittgenstein's clusters. As with clusters, however, members of a category are often presumed to exhibit some sufficient proportion of the stereotype rather than the whole of it.

Stereotypes can be compared to what psychologists call prototypes. According to prototype theory, most lexical concepts are identified with representations of prototypical category instances (Hampton, 1993). Sometimes, a prototype is thought to be a representation of the actual instance that best captures the category's central tendency. More commonly, prototypes are regarded as weighted lists of features corresponding to the properties most frequently recognized in category instances. Features that are more frequently recognized or more diagnostic may be given higher weights. Categorization depends on passing a critical threshold of similarity, computed by comparing a target instance to the features contained in the weighted list. Prototype theorists emphasize the graded nature of category judgments. People regard certain category instances as more typical than others. These are said to reflect increased similarity to the prototype.

Assessment

There is ample evidence that people often categorize on the basis of judged similarity to collections of non-defining features. The psychological reality of

something like clusters, stereotypes, or prototypes is rarely challenged. However, some researchers are reluctant to identify such cognitive structures with concepts. To see why, consider some of the objections to stereotypes (all of which apply to prototypes as well).

Putnam points out that stereotypes are insufficient to determine reference. To demonstrate, he has us imagine two very similar individuals living on different planets. One lives on Earth, where the thirst-quenching, clear liquid in rivers and streams is H_2O. His doppelgänger lives on Twin Earth, where the liquid that has those properties is a different chemical compound, XYZ. These individuals (who are ignorant of chemistry) have the same stereotypes associated with the word 'water', but apparently their concepts refer to different substances. Some researchers believe that concepts with different referents cannot be identical. One is a *water* concept, and the other is a *twin water* concept. If this is the case, then stereotypes can, at best, be one component of our concepts. In addition, we must individuate concepts by their referents. (Putnam draws this conclusion about meanings rather than concepts.)

The insufficiency of stereotypes can also be demonstrated by considering categorization judgments. While we often categorize on the basis of superficial similarity, we recognize that appearances can deceive. A wolf in sheep's clothing is still a wolf even if its disguise is good enough to resemble a stereotypical sheep. This is a psychological analogue of Putnam's observation; concept users know that similarity to their stereotypes is insufficient for reference. Stereotypes are good for rough and ready categorization, but we must be capable of transcending them.

This last observation is connected with a general concern. Stereotypes comprise only a small portion of the knowledge we have about categories. They generally capture superficial appearances. We also know a great deal about features hidden from view, the relations between superficial features, how category instances function, how they relate to instances of other categories, and so on. All of this information can potentially influence our judgments concerning categories. There may be no reason to privilege superficial features from this rich reservoir of information.

One of the most serious concerns about stereotypes is that they do not combine compositionally. For example, the stereotype for *pet bird* includes the feature *being caged*, which is not included in the stereotype for *pet* or the stereotype for *bird* (see Fodor, 1998, for a review). If concepts combine compositionally (as argued above), they cannot be identified with stereotypes.

The Theory Theory and Concept Holism

Overview

Psychologists have developed an alternative to the prototype theory, called the theory theory. Theory theorists maintain that most lexical concepts are similar to scientific theories in a number of ways.

Firstly, theory theorists say that concepts encode knowledge about causal and explanatory principles just as scientific theories encode knowledge about laws and mechanisms (Murphy and Medin, 1985). In addition to knowing that birds fly and that birds have wings (two stereotypical features), people know that wings enable flight. Features that enter into such explanatory relations are more likely to be included in concepts and highly weighted.

Secondly, theory theorists claim that concepts encode the knowledge that the features essential to category membership may be unobservable, like the postulates of many scientific theories. Something is identified as being a horse by virtue of its appearance, but it really counts as being a horse only by virtue of having a particular microstructure. Echoing the *a posteriori* definition theory view, theory theorists recognize that concept users have faith in such essences without knowing what they are in detail. Keil (1989) demonstrates this experimentally. He asked subjects to consider an animal that began looking like a horse but was painted to look just like a zebra, and acted like one. Young children are deceived by the transformation, but the rest of us recognize that such a creature would still count as a horse.

Keil's experiments also illustrate a third claim of theory theorists. In contrast to the horse case, mature concept users believe that artefacts can change their identity when they are superficially transformed. When a coffee pot is modified to look and function like a bird feeder, subjects say it has become a bird feeder. This suggests that we treat different kinds of concepts differently. Theory theorists speculate that concepts are parceled into distinct cognitive domains, each dedicated to knowledge about a distinct class of entities and driven by distinct principles. For example, we may have naive theories of artefacts, biological kinds, intentional agency, and physical mechanics.

The fourth claim of the theory theory involves conceptual change. Children often use words in very different ways from adults. For example, a

young child might extend the word 'alive' to include cars and other inanimate objects (Carey, 1985). Such observations lead Carey to conclude that children's concepts may differ significantly from adults'. She argues that these differences lead to a certain degree of incommensurability, an idea she borrows from Kuhn's (1962) philosophy of science. Sometimes, a child's terms can be translated into our own without loss of meaning. This echoes the conclusion that Kuhn draws about terms used within different scientific theories.

The theory theory has been most explicitly defended by psychologists, but it is related to a class of theories that some philosophers defend. Quine emphasizes the continuity between theoretical knowledge and the totality of ordinary beliefs about a category. There is no obvious boundary between the knowledge that comprises our informal theory of a category and all other information associated with that category (recall Quine's confirmation holism). If concepts are identified via intuitive theories, then any given concept will be identified by a vast collection of beliefs comprised of other concepts, which will be identified by vast belief sets of their own. Following this logic, the identity conditions of any particular concept may depend on just about every other concept known to its possessor. This may be called 'concept holism'.

Without a clear boundary between theoretical knowledge and collateral information, theory theorists may be forced to embrace concept holism. Concept holism is often associated with meaning holism. If concepts are meanings, and concepts are identified by their place in a vast network of beliefs, then it is natural to conclude that the meaning of a term depends on a vast network of beliefs as well. As we saw earlier, concept holism is an unwelcome idea for defenders of the definition theory, who would wish to restrict concepts to discrete sets of necessary features. But it enjoys considerable support in its own right. Concept holists do not always emphasize the theoretical nature of concepts, and theory theorists do not always emphasize concept holism, but these two approaches may have a tendency to coincide.

Assessment

The theory theory and concept holism are open to objections. Like the stereotype theory, they face difficulties with concept combination: how do vast bodies of information combine compositionally? But there is a more pressing problem. If concepts are comprised of everything we know about a category, then no two people would share concepts, because different people's knowledge varies to some degree (Fodor and LePore, 1992). One might try to avoid this problem by saying that people have similar concepts rather than identical concepts. But Fodor and LePore reply that holism offers no way of quantifying similarity between people's concepts. We cannot say that concepts are similar by virtue of having some of the same constituent features, because each feature is itself a concept that, for the theory theorist, must be identified with an entire body of knowledge. If concepts cannot be shared, it is difficult to explain how communication ever occurs. To avoid such difficulties, theory theorists can try to distance themselves from concept holism by identifying a boundary between theoretical knowledge and collateral information.

Informational Atomism

Overview

All of the accounts considered thus far assume that most lexical concepts embody knowledge of the categories they represent. Fodor (1998) argues that any proposal of this kind is doomed to failure. Concepts cannot be definitions, because definitions are too scarce; they cannot be stereotypes, because stereotypes are not compositional; and they cannot be theories, because theories are not shared. The only way to avoid all of these problems is to deny that concepts embody knowledge.

Fodor calls his theory 'informational atomism'. It is called atomism, because it asserts that most of our lexical concepts are primitive. They neither entail nor decompose into any features. Instead, concepts are individuated by the properties to which they refer. They come to refer by means of 'informational semantics'. Informational semantics says that a concept refers to a property by virtue of being lawfully caused by that property. A *dog* concept refers to doghood because it becomes active when one encounters dogs. Fodor does not deny that we associate a considerable amount of knowledge with our *dog* concepts (including dog prototypes and dog theories), but he denies that this knowledge is conceptually constitutive. The *dog* concept can be regarded as an inner label in a language of thought. In principle, one could have the very same label even if one's dog prototypes and theories changed or disappeared entirely.

Inner labels can be compositionally combined, because their content is exhausted by their referents. There is no problem of compounds having features not found in their components. Inner labels can also be readily shared. Two people have the same concept by virtue of having labels

that are lawfully caused by the same objects, even if their beliefs about those objects differ radically.

Assessment

Fodor's concepts are well suited to serve reference, combination, and communication functions, but these benefits come at a price. If concepts do not decompose into features, they cannot satisfy the knowledge, inference, or categorization functions: the very functions that motivate most psychological work on concepts. Fodor also faces the challenge of explaining concept acquisition. Traditionally, primitive concepts are presumed to be innate. If most lexical concepts are primitive, then most lexical concepts are innate (Fodor, 1975; but cf. Fodor, 1998). Many find this consequence unsettling.

CONCLUSION

Concepts are postulated to serve a variety of functions. These include roles in reference, knowledge, inference, categorization, combination, communication, and meaning. A number of theories have been proposed, but each seems to have serious limitations. Some researchers argue that we must reduce the list of functions that concepts ought to serve, while others hope for a more encompassing theory.

References

Block N (1986) Advertisement for a semantics for psychology. In: French PA, Uehling TE and Wettstein HK (eds) *Midwest Studies in Philosophy*, vol. X, *Philosophy of Mind*. Minneapolis, MN: University of Minnesota Press.

Carey S (1985) *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carnap R (1934/1959) *The Logical Syntax of Language*. London, UK: Routledge & Kegan Paul.

Fodor JA (1975) *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor JA (1981) The current status of the innateness controversy. In: *Representations*. Cambridge, MA: MIT Press.

Fodor JA (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford, UK: Oxford University Press.

Fodor JA and LePore E (1992) *Holism: A Shopper's Guide*. Oxford, UK: Blackwell.

Frege G (1893) On Sinn and Bedeutung. In: Beaney M (ed.) (1997) *The Frege Reader*, pp. 151–171. Oxford, UK: Blackwell.

Hampton A (1993) Prototype models of concept representation. In: van Mechelen I, Hampton J, Michalski RS and Theuns P (eds) *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, pp. 67–95. New York, NY: Academic Press.

Katz J and Fodor JA (1963) The structure of a semantic theory. *Language* **39**: 170–210.

Keil FC (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kuhn T (1962) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Murphy GL and Medin DL (1985) The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316.

Peacocke C (1992) *A Study of Concepts*. Cambridge, MA: MIT Press.

Prinz JJ (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Putnam H (1975) The meaning of 'meaning'. In: Gunderson K (ed.) *Language, Mind and Knowledge*, pp. 131–193. Minneapolis, MN: University of Minnesota Press.

Quine WVO (1951) Two dogmas of empiricism. *Philosophical Review* **60**: 20–43.

Rey G (1983) Concepts and stereotypes. *Cognition* **15**: 237–262.

Smith EE (1989) Concepts and induction. In: Posner K (ed.) *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Wittgenstein L (1953) *Philosophical Investigations*. New York, NY: Macmillan.

Further Reading

Armstrong SL, Gleitman LR and Gleitman H (1983) What some concepts might not be. *Cognition* **13**: 263–308.

Barsalou LW (1987) The instability of graded structure: implications for the nature of concepts. In: Neisser U (ed.) *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pp. 101–140. Cambridge, UK: Cambridge University Press.

Clark A (1993) *Associative Engines*. Cambridge, MA: MIT Press.

Gopnik A and Melzoff A (1997) *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Locke J (1690/1989) *An Essay Concerning Human Understanding*. Oxford, UK: Clarendon Press.

Margolis S and Laurence S (eds) (1999) *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Millikan R (2000) *On Clear and Confused Ideas: An Essay About Substance Concepts*. Cambridge, UK: Cambridge University Press.

Smith EE and Medin D (1981) *Concepts and Categories*. Cambridge, MA: Harvard University Press.

Rosch E and Mervis C (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.

Connectionism and Systematicity

Intermediate article

Robert J Matthews, Rutgers University, New Brunswick, New Jersey, USA

CONTENTS

Connectionists claim that human cognitive computational architecture is connectionist. Proponents of classical computational architectures have challenged this claim, arguing that a pervasive feature of human cognition, its 'systematicity', cannot be explained in connectionist terms.

INTRODUCTION

Connectionists claim that human cognitive computational architecture is connectionist, specifically that connectionist computational architectures provide an empirically more plausible foundation for computational theories of cognition than do more familiar classical computational architectures. Classicists, who conceive of cognition as implemented by a classical computational architecture, specifically as the structure-sensitive processing of syntactically structured mental representations, argue that the prospects for a connectionist theory of cognition are remote. Connectionist models, they contend, cannot explain, or at least have not shown that they can explain, any of the fundamental aspects of human cognition, most notably its seemingly pervasive 'systematicity', namely: the fact that many cognitive capacities are systematically related, so that as a matter of psychological law one possesses one capacity if and only if one possesses the other.

THE FODOR–PYLYSHYN CHALLENGE TO CONNECTIONISM

As first issued by Fodor and Pylyshyn (1988), the challenge to explain systematicity was one that connectionists purportedly could not meet because connectionist architectures lacked an essential feature found only in classical computational architectures, namely, representational states with a compositional constituent structure. Classical architectures, Fodor and Pylyshyn argued, postulate certain syntactically structured mental representations that, like the sentences of a natural language, exhibit a compositional syntax and semantics, such that the semantic content of a complex representation is a function of the semantic content of its constituents and of its syntax structure. (The sentence-like character of the postulated representations explains why classicists sometimes call these 'language of thought' architectures.) The computational processes that operate over these compositionally structured representations apply by virtue of their syntactic properties. Connectionist representations, by contrast, lack such compositional constituent structure; hence, the computational processes postulated in connectionist architectures cannot, Fodor and Pylyshyn argued, be causally sensitive to constituent structure. But such sensitivity, they claimed, is essential to explaining systematicity. Fodor and Pylyshyn concluded not only that connectionist architectures could not explain systematicity, but that they were unable even to exhibit systematicity: 'the architecture of the mind is not a connectionist network'.

Fodor and McLaughlin (1990) put the challenge to connectionists this way: how can you 'explain the existence of systematic relations among cognitive capacities without assuming that cognitive processes are causally sensitive to the constituent structure of mental representations'? Fodor and McLaughlin argue that connectionists are faced with a dilemma: 'if connectionism can't account for systematicity, it thereby fails to provide an adequate basis for a theory of cognition; but if its account of systematicity requires mental processes that are sensitive to the constituent structure of mental representations, then the theory of cognition that it offers will be, at best, an implementation architecture for a "classical" (language of thought) model.' Given that a classical architecture could be

implemented on a connectionist architecture (just as connectionist architectures can be, and often are, implemented on a classical architecture), the challenge to connectionists, Fodor and McLaughlin argue, is to demonstrate that connectionism can explain systematicity without implementing a classical architecture.

Fodor and Pylyshyn (1988) offer little by way of a characterization of the property – systematicity – that connectionists are challenged to explain. Instead they offer numerous examples of systematically related cognitive capacities. Many of their examples are drawn from language. They note, for example, that you do not find subjects who know how to say in English that John loves the girl, but who do not know how to say in English that the girl loves John; or who can understand the English sentence 'the monkey bit the lab technician', but who cannot understand the systematically related English sentence 'the lab technician bit the monkey'. In Fodor and McLaughlin (1990), and especially in McLaughlin (1993), systematicity is described more broadly, as a capacity for *intentional*, more specifically 'propositional attitude', states whose contents are systematically related, such that one has the capacity for one such intentional state (e.g. believing that the circle is above the square) just in case one has the capacity for systematically related states (e.g. believing that the square is above the circle).

PHILOSOPHICAL RESPONSES TO THE CHALLENGE

Some connectionists have been willing to accept the classicists' challenge, undertaking to demonstrate that their networks can, at least in principle, explain systematicity (and furthermore do so without implementing a classical architecture). Others have rejected the challenge, arguing that as posed the challenge is impossible to meet. They argue that the classicists' notion of what it would be to explain systematicity, indeed their notion of systematicity itself, at least as regards thought, is such as to insure that there can be no non-classical explanation of systematicity. Some also question the classicists claim to have an explanation themselves, thus challenging the presumption that connectionists are worse off in this respect than classicists.

The classicists' notion of systematicity in language processing capacity is relatively uncontentious: a language processor can be said to exhibit systematicity if when it can process a sentence s it can also process systematic variants of s, where systematic variation is understood in terms of permuting syntactic constituents or substituting constituents of the same syntactic category. Their notion of systematicity of thought is considerably more contentious. Fodor defines it in terms of being able to think thoughts of one form (e.g. *aRb*) just in case one can think other thoughts whose forms are systematically related to the first (e.g. *bRa*). But this characterization, as Cummins (1996) points out, begs the question, since it assumes that thoughts have the form of their classical linguistic representations. This objection can be avoided by reformulating the systematicity claim as follows:

> Anyone who can think a thought with the content c can think a thought with the content c^*, where c^* is a systematic variant of c. (1)

But this reformulation, Cummins notes, relativizes the systematicity of thought to the choice of some representational scheme for thought contents. Yet surely the systematicity of thought, which connectionists are challenged to explain, should not depend on the representational scheme that we, as theorists, choose to employ. For without some way of picking out systematically related thought contents (and hence systematically related thoughts), a way that is neutral regarding the representation scheme, it is not clear that any real property of thought has been identified for connectionists to explain. Nor is there any non-question-begging way of arguing from the systematicity of thought to the conclusion that human cognitive architecture is exclusively classical.

The challenge to connectionists is to *explain* systematicity, not simply to construct a connectionist network that exhibits it. Classicists have a narrow notion of what would count as an explanation of systematicity, one that connectionists might have difficulty satisfying even if they were successful in constructing a network that exhibited systematicity. McLaughlin (1993), for example, requires that explanations of cognitive capacities, including systematically related cognitive capacities, explain what the capacity 'consists in'. More precisely, he requires that the explanation take the form of a 'functional analysis'; i.e. an explanation of a complex cognitive capacity such as the capacity for systematic thought in terms of the cooperative interaction of certain more primitive cognitive capacities that are constitutive of the complex capacity. Systematically related cognitive capacities, McLaughlin assumes, are so related by virtue of certain shared constitutive capacities, so that to explain the former is to specify and describe these shared constitutive capacities and their interaction.

Functional analysis offers a plausible form of explanation if classicists are right in their assumptions about the nature of systematically related cognitive capacities. But these are not assumptions that connectionists accept; moreover, it is not clear, as Matthews (1997) argues, that any cognitive explanations that connectionists might provide could take this form. Connectionist networks do not have constitutive capacities of the sort that functional analysis envisions. Of course, the individual units that compose connectionist networks are constituents of those networks, but arguably they are not cognitive constituents of the networks in the classicists' sense of that term; i.e., they are not amenable to cognitive or intentional interpretation. And even if they were constituents in the requisite sense, there would still be a problem of explanatory 'grain': it would in most cases be very difficult, if not impossible, to grasp how the molar capacity of the network comprises ('consists in') the capacities of the individual units that constitute it. The specific contributions of individual units would typically be so diffuse as to preclude any claims about which units are responsible for which aspects of the networks' molar capacity; moreover, the number of units would typically be so large, and their interaction so complex, that it would be beyond our cognitive capacity to grasp how the molar capacity of the network could 'consist in' the capacities of the individual constitutive units.

To concede that connectionist explanations of cognitive capacities are unlikely to take the form of a functional analysis is not to concede that connectionists are unable to meet the classicists' challenge to explain systematicity (assuming that we can define in a non-question-begging way just what it is for thought to be systematic). Rather, it is to point out that *a priori* assumptions about the appropriate form of cognitive explanation are vulnerable to empirical discoveries about the nature of human cognitive architecture. Empirical discoveries about the bases of cognition may entail corresponding changes in what we take to be the appropriate form that explanations of cognition should take.

EMPIRICAL RESPONSES TO THE CHALLENGE

The empirical response to the classicist challenge has thus far been directed primarily towards demonstrating that connectionist networks can, in principle at least, *exhibit* systematicity in domains such as language processing. Thus, for example, Chalmers (1990) describes a connectionist network that is, he says, a 'direct counterexample' to the argument of Fodor and McLaughlin (1990) that to support structure-sensitive processing, representations of constituent structure must contain explicit tokens of the constituents. In part, this response has been prompted by suggestions to the effect that connectionist architectures are unable to explain systematicity because they cannot perform certain computational tasks that classical architectures can. No one challenges the well-known formal proof that connectionist architectures can compute (or at least approximate to any arbitrary degree) any computable function (Hornik *et al*., 1989), but some have suggested that connectionist architectures may be limited computationally in ways that classical architectures are not. McLaughlin (1993), for example, speaks of connectionist architectures being able only to 'respect' certain syntactic operations that classical architectures can actually execute. Connectionists would naturally wish to dispel any suggestion that connectionist architectures are computationally limited in ways that would prevent them from exhibiting, and hence from explaining, systematicity.

Smolensky has argued that connectionists can explain systematicity, and furthermore can do so without implementing a classical architecture (Smolensky, 1991, 1995; Smolensky *et al*., 1992). He argues that: (1) connectionist networks are naturally viewed as computing functions defined over vector product representations (i.e. over representations of a vector algebraic form); (2) such representations are adequate to express all constituency relations expressible by means of classical representations; (3) all computational operations definable over classical representations can be effected by connectionist operations defined over vector product representations; and hence (4) connectionists can explain systematicity in terms of such representations. Smolensky presents what he calls an 'integrated connectionist/symbolic architecture' that provides an algorithmic encoding scheme for moving between these two forms of representation. Fodor and McLaughlin (1990) criticize Smolensky's explanatory claims, largely on the grounds that vector product representations only encode but do not actually possess the constituent structure that a classicist explanation of systematicity presumes.

Whatever the merits of these criticisms, Smolensky's argument is clearly a non sequitur. It takes non-epistemological premises regarding the computational capacity of connectionist devices and the availability of a vector product representation of their operations, and reaches an epistemological conclusion regarding the availability of an

explanation of the systematicity that connectionist networks so described can exhibit. But the argument nonetheless draws attention to both the known capacity of connectionist networks to compute arbitrary computable functions and the availability of a principled non-classical, specifically vector product, description of the computations of such networks.

CONCLUSION

Much of the empirical response of connectionists to the classicists' challenge has focused on establishing that connectionist networks can do the sorts of things that classicists have claimed that they cannot do. Connectionists have made a good case for the claim that connectionist networks can in principle exhibit systematicity, at least in domains such as language processing where it is tolerably clear what systematicity amounts to. But arguably connectionists have yet to demonstrate that what is in principle possible can in fact be accomplished. It also remains to be seen whether connectionists can actually explain systematicity, especially the systematicity of thought, assuming that classicists and connectionists can agree on what is to be explained and what is to count as an explanation. Ultimately, these are matters for empirical investigation, and cannot be decided *a priori*.

References

Chalmers D (1990) Syntactic transformations on distributed representations. *Connection Science* **2**: 53–62.

Cummins R (1996) Systematicity. *Journal of Philosophy* **93**: 591–614.

Fodor J and McLaughlin B (1990) Connectionism and the problem of systematicity: why Smolensky's solution won't work. *Cognition* **35**: 183–204.

Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Hornik K, Stinchcombe M and White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359–366.

Matthews R (1997) Can connectionists explain systematicity? *Mind and Language* **12**: 154–177.

McLaughlin B (1993) The connectionism/classicism battle to win souls. *Philosophical Studies* **71**: 163–190.

Smolensky P (1991) Connectionism, constituency, and the language of thought. In: Loewer B and Rey G (eds) *Meaning in Mind: Fodor and His Critics*, pp. 201–227. London, UK: Blackwell.

Smolensky P (1995) Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In: MacDonald C and MacDonald G (eds) *Connectionism: Debates on Psychological Explanation*, pp. 223–290. London, UK: Blackwell.

Smolensky P, Legendre G and Miyata Y (1992) Principles for an integrated connectionist/symbolic theory of higher cognition. Technical Report 92–08, Institute of Cognitive Science, University of Colorado, Boulder, CO, USA.

Further Reading

Hadley R (1994) Systematicity in connectionist language learning. *Mind and Language* **9**: 247–272.

Hadley R (1997) Cognition, systematicity, and nomic necessity. *Mind and Language* **12**: 137–153.

Matthews R (1994) Three-concept monte: explanation, implementation, and systematicity. *Synthese* **101**: 347–363.

Niklasson L and van Gelder T (1994) On being systematically connectionist. *Mind and Language* **9**: 288–302.

Consciousness and Attention

Gregory J DiGirolamo, University of Cambridge, Cambridge, UK
Harry J Griffin, University of Cambridge, Cambridge, UK

CONTENTS

The concepts of consciousness and attention have been used in many ways and the processes that constitute them are not well agreed upon. Attention may be considered as an agency for bringing a stimulus into conscious awareness.

INTRODUCTION

It is hard to define consciousness or attention because these concepts have been used in many ways and the processes that fall under these umbrella terms are not well agreed upon. It is even more difficult to specify the functions and mechanisms of consciousness that lead to coherent behavior. Our strongest indication of consciousness remains the subjective experience. Nevertheless, advances in the scientific study of consciousness such as neuroimaging studies of perceptual awareness have given an anatomical and functional reality to both attention and consciousness in the human brain. This article looks at models of attention in order to explore the cognitive processes and neural substrates that may be shared between attention and consciousness. Evidence from cognitive neuroscience (the study of how the human brain carries out psychological processes) is applied to elucidate the complex relationship between consciousness and attention, and to enhance our understanding of both concepts.

THE RELATIONSHIP BETWEEN CONSCIOUSNESS AND ATTENTION

As early as the nineteenth century, psychologists and philosophers were suggesting a close relationship between attention and consciousness. William James (1890) succinctly captured this intuitive connection: 'Everyone knows what attention is … Focalization, concentration of consciousness are of its essence … My experience is what I agree to attend to.'

Attention and consciousness share many features, perhaps the most striking commonality being that at any given moment, one object or thought seems to predominate in the focus of attention and hence in our conscious awareness. In most situations in everyday life we are constantly bombarded by a variety of external stimuli from all sensory modalities (e.g. tactile, visual, and auditory) as well as by our own internal thoughts and memories. It would be difficult for an organism to achieve coherent, goal-directed behavior if all stimuli in the environment were processed and responded to in turn without prioritization. One can bring attention to bear on any of these objects (such as the background noises while you are reading this page), and suddenly this stimulus becomes the primary sensation or attribute in conscious awareness. It is, of course, possible to switch quickly between disparate thoughts or different objects in the environment; yet, the subjective experience is that only a single object is in attention or consciousness. Note that the perception of other objects in the environment remains, but when not the focus of your concentration the unattended objects have a vague, indistinct quality. (*See* **Consciousness, Unity of**)

With advances in neuroimaging it is now possible to investigate noninvasively the workings of the human brain as the subject attends to and becomes consciously aware of a stimulus. In addition, disorders of conscious awareness and attention following injury or psychosis have furthered our understanding of how these processes are carried out by the human brain as well as the relationship between attention and consciousness. (*See* **Consciousness, Disorders of**)

EXPERIMENTAL WORK ON ATTENTION AND CONSCIOUS AWARENESS

An important issue in the field of attention is how attention influences stimulus processing. Does attention manifest itself as alterations in the beta parameter (response biases), of a signal detection analysis or as changes in the d′ parameter (perceptual sensitivity)? In a typical experiment, participants were asked to keep their gaze fixed at a central point of a screen with four peripheral locations marked by small boxes. One location was cued (either by a brightening of that location's box, or by an arrow pointing to that location) so that participants could move their attention (but not their eyes!) to that location ahead of the actual target. A near-threshold target then appeared either at the cued location (valid) on 75% of trials, or at one of the uncued locations (invalid) on 25% of the trials. Reaction times to targets at the attended location were significantly faster and more accurate than targets at the uncued location. In addition, perceptual sensitivity changed at the location that was validly cued. Such studies demonstrate that attention to a location changes the sensitivity to incoming stimuli and makes these stimuli more perceivable. Targets at uncued locations were sometimes completely missed; that is, near-threshold targets at unattended locations did not receive sufficient attentional processing to enter conscious awareness. This simple behavioral paradigm suggests that items that are attended are processed more efficiently and enter consciousness; without attention, the individuals may not be aware of the item at all. (*See* **Attention, Neural Basis of**)

CONSCIOUSNESS AND ATTENTION IN NEUROPSYCHOLOGY

Neuropsychological evidence also elucidates the relationship between attention and awareness. One of the most striking findings is that lesions of the parietal lobe (particularly the right parietal lobe) produce specific deficits in attention. In the immediate aftermath of damage to the right parietal lobe, these patients will 'neglect' (not attend to or be aware of) information coming from the contralesional side of space (the side opposite to the hemisphere with damage). The side of space is important, as information from the left side of space is processed first by the right hemisphere, and vice versa; in addition, each hemisphere controls the opposite side of the body. In its severest form, patients with neglect fail to comb one side of their hair (the contralesional side) and eat off only one side of their plate (the ipsilesional side). If approached by a person on each side, these patients will look at the person on their ipsilesional side, even if the person on the contralesional side is the one who is speaking. They seem to fail to be consciously aware of information coming from the side of space processed by the damaged hemisphere. In fact, this deficit is so severe that, following a stroke producing neglect in the German painter Anton Räderscheidt, a self-portrait shows that, in the painting, he fails to represent the information from the contralateral side of space, including one side of his face. Fortunately, this deficit often resolves over time. (*See* **Attention, Neural Basis of**; **Attention**)

This deficit is not perceptual but rather attentional. For example, if patients with neglect are asked to create a visual mental image (e.g., the central square of their home town), they fail to report items on the left side of their mental image. However, if they are then asked to imagine walking to the opposite end of the square and turning around (so that the left and right sides are reversed), and to report what they see in their image, they will then report all of the items they failed to report from the previous view, and fail to report all the items they have just described. In this case, the patients are neglecting not incoming sensory information, but the represented visual image. Since the information is clearly present (as they report the entire representation between the two perspective shifts), these results suggest that the impairment is nonsensory in nature.

This deficit is not a general impairment, but specific to aspects of spatial attention. In people with neglect, their ability to shift attention to the contralesional side is severely impaired. Stimuli from the contralesional side appeared to have greater difficulty in summoning attention under conditions when the person is already processing something on the ipsilesional side. Thus, while normal people showed only a small deficit in reaction time when the target appeared in the opposite visual field from an attentional cue (invalid trials), people with neglect were often simply unaware of targets presented in the neglected field. Since these people were unable to report the mere presence of a target if their attention had been previously summoned to the good visual field, this finding suggests that attention is necessary for objects to come into conscious awareness. (*See* **Attention**)

Additional studies indicate that stimuli are processed in the neglected visual field even if unconsciously, and awareness is still possible under the

appropriate conditions. For example, while experimental subjects might miss a contralesional target (a circle) in isolation, they might report the stimulus if it integrated with material on the ipsilesional side to make a single form, such as a dumbbell. Moreover, information in the neglected field can sometimes affect processing of items in both fields. Some patients will fail to be aware of an item in the neglected visual field if the item is also present in the good visual field (extinction). This deficit is ameliorated if the two items are different objects (e.g. a spoon and fork) rather than identical. Interestingly, these patients can tell if two objects (one in each visual field) are different, without being able to identify the object in the neglected field; this suggests that patients can process some information without attention, to the extent of telling differences between objects, while identity remains unavailable to conscious report. In addition, semantic information in the neglected visual field will affect processing in the good visual field – that is, the patient will respond faster to the word 'cat' in the good visual field if it has been preceded by the word 'dog' in the neglected visual field, than if it had been preceded by a neutral word in the neglected field. Hence, the word in the neglected visual field is processed to the level of meaning in the absence of attention and without conscious awareness.

These neuropsychological studies have outlined the depth of processing applied to an attended stimulus and implicated a strong relationship between spatial attention to and awareness of a stimulus. While most of these studies have dealt with deficits in spatial attention particular to disengaging and shifting attention in external space, we turn to one final syndrome following brain damage that suggests that attention also works on the object level to help bring stimuli into conscious awareness.

Damage to both parietal lobes (and the occipital lobes) can produce a condition known as Balint's syndrome in which patients can perceive only one object at a time. The visual system of these patients can 'see' the objects; however, attention can only be directed to one object at a time. Hence, these patients are only consciously aware of a single object. If presented with a comb, pen and fork, they will see only the comb, and then only the pen or only the fork. These results suggest that competition for conscious recognition is resolved through attentional processes, and without these processes objects cannot come into consciousness. Balint's syndrome demonstrates that objects excluded from awareness in favor of others need not lie within a specific region of external space. Rather, the simultanagnosia (inability to perceive two objects at the same time) indicates that a lack of awareness can be object-based. A single object may remain outside awareness despite being moved into a previously attended region of space. The return of this object to awareness may occur spontaneously; that is, the patient goes from seeing the fork to the pen. However, if the experimenter moves the object that is not currently in consciousness (the fork), the change in the attentional salience causes the patient to become aware of the moving object. This external change in the stimulus causes a shift in attentional bias from one internal representation (the pen) to the other (the fork). Likewise, a spontaneous shift in the object perceived is likely to be caused by a similar change in attentional bias between representations, but now in the absence of an external cue. (*See* **Attention**; **Attention, Neural Basis of**)

The lack of awareness in both neglect and simultanagnosia illustrates how the human brain chooses, from the representations offered to it, which objects enter into conscious awareness. When the neural areas underlying attention are damaged the number and breadth of these conscious interpretations becomes limited; these patients do not have even a vague, indistinct awareness of other objects in the environment (as normal people do even for objects outside their attentional focus). We now turn to studies in normal individuals that help to clarify the relationship between multiple representations, neural mechanisms of attention, and the entrance into conscious awareness.

DIFFERENT THEORIES OF CONSCIOUSNESS AND ATTENTION

Studies following brain damage have certain disadvantages as the lesions have ill-defined boundaries and may spread over areas that perform different functions. In addition, one must always be aware of the possibility of plasticity of function and the importance of patient strategies. However, a phenomenon exists that allows the study in normal people of the neural mechanisms of attention and its link to visual awareness: binocular rivalry.

Binocular rivalry occurs when dissimilar images (e.g. a face and a house) are presented to each eye. Instead of seeing a permanent mixture of the two monocular images, a multi-stable percept occurs in which the percept switches rapidly and involuntarily between the objects presented to each eye. Short periods of transition may also occur during

which the percept is a fusion of the two images. The more different the two stimuli are in orientation, color, contrast or movement, the less prevalent are periods of piecemeal perception. Also, increasing the contrast of one of the stimuli increases its superiority, leading to it being perceived for longer periods. Despite being out of conscious awareness when suppressed, the subordinate stimulus can still influence cognitive processing (e.g. adaptation or priming).

Usually, functional magnetic resonance imaging (fMRI) studies compare brain response under equivalent rivalry and nonrivalry conditions. The rivalry condition is achieved by presentation of two images (one to each eye). The nonrivalry condition consists either of alternating binocular presentations of each of the two stimuli, or of alternating dichoptic presentation of an image to the relevant eye and a uniform gray field to the other. The rate of alternations is matched to the participant's pattern of awareness in the rivalry condition. In the nonrivalry condition, changes in awareness are caused by an alteration in the visual stimulation, whereas in the rivalry condition, changes in awareness are caused by involuntary shifts from the internal representation of one stimulus (the house) to the representation of the other (the face) in the absence of an external cue. By comparing these two conditions, we are able to study the neural mechanisms of shifts of visual awareness.

Binocular rivalry provides an opportunity to study shifts in conscious perception in the absence of any changes to the external stimuli. Using functional imaging with binocular rivalry, we can measure neural activity in visual awareness associated with both stable perception and perceptual transitions. It is also possible to distinguish neural areas of the visual system in which activation corresponds to the changing percept from areas in which activation corresponds to the fixed retinal stimulation. If the activation of a neural area corresponds to the reported percept, it suggests that the visual scene has been resolved at that stage of processing in the visual system (or, that feedback from attention is modulating the neuronal response in order to coherently resolve and stabilize the visual percept). (*See* **Attention, Neural Basis of**)

Activation of the right frontal and parietal cortices has been noted in perceptual transitions during the rivalry condition. The association of right frontal and parietal areas with tasks requiring shifts of spatial attention (both voluntary and involuntary) is well established. As we have discussed above, damage to the parietal cortex leads to disorders of visual awareness, and indeed, the similarity between binocular rivalry and extinction is worth noting. In both cases, two objects are presented in the environment and impinge upon the retina, but one of the objects is suppressed from visual awareness. Although a perceptual shift in binocular rivalry does not involve a shift in external space, it does involve a shift in object-based attention. The reason for perceptual shifts may be differential activity in the neurons representing each object caused by selective habituation to the perceived stimulus. When sufficient habituation has occurred, attention may shift to the object that now has greater neuronal activity. Once this shift has taken place and attention is now directed to the other object's representation, attention acts to stabilize the percept by increasing the neuronal activity associated with the previously suppressed object in lower visual levels and to push the visual system into a different but equally stable perceptual state. The threshold for the attentional shift is likely to be smaller than the threshold for a change in awareness; hence, the activation of neural areas involved in attention (e.g. the parietal cortex) amplifies the representation prior to conscious awareness. As suggested in the section on extinction, one gateway into conscious awareness is the relative amplification of the representation of an object through attention.

This cyclic habituation and attention capture could produce a multi-stable state, which would be beneficial as it allows accurate perception of at least one part of the visual scene rather than a confused interpretation of the entirety of the visual input. The constant shifts may be indicative of the tendency of the visual system to shift attention automatically towards a salient or novel stimulus in order to orient towards possibly important events (either external objects or internal representations). For example, movement of an object in the environment is salient and produces a capture of attention. As discussed above, patients with Balint's syndrome will become aware of an object currently out of awareness if it is moved about (as attention is drawn to the movement, and then the object enters consciousness). In binocular rivalry, the salience of a stimulus is not associated with changes in the external stimulus, but with its neural representation reaching a threshold level of activation that is sufficiently greater than the representation of the other monocular input. The function of parietal structures could be to disengage attention from the previously perceived object representation and shift it to the now more active representation.

Traditionally it was thought that the resolution of the ambiguous visual scene in binocular rivalry was carried out at a relatively late stage in the visual stream after the processing of each monocular image. Indeed single-cell studies have demonstrated that the activity of the majority of neurons in the striate cortex is unaffected across perceptual changes during rivalry; that is, activity in primary visual cortex correlates with the retinal stimulation whereas activity in later visual areas corresponds significantly to the changing percept. In the inferotemporal cortex, the activity of the majority of neurons follows the percept rather than the retinal stimulation. Functional MRI studies have shown that areas linked to perception of specific object categories increase their activity when an exemplar is perceived. For example, activation in part of the fusiform gyrus correlates with the presentation of a face and is relatively specific for faces or face-like stimuli. During a binocular rivalry condition in which a face and a house are presented to each eye, activation in this fusiform face area is seen only when the percept is that of a face. The changes in activity of these later visual areas, in contrast to that of V1, are as large in the rivalry condition as they are in the nonrivalry condition where the stimulus is presented binocularly (e.g. a face is presented to each eye). These results suggest that the awareness is resolved in a somewhat gradual fashion throughout the visual system. However, at least in later visual areas, neural activity correlates with conscious awareness, not the retinal stimulation. (*See* **Attention, Neural Basis of**)

CAN THERE BE ATTENTION WITHOUT CONSCIOUS AWARENESS?

Finally, we turn to one other neuropsychological disorder to help constrain the relationship between attention and consciousness. As previously explained, people who have sustained brain damage often acquire disorders of thought, perception, or even consciousness. People with damage to the primary visual cortex may deny any conscious sensation of visual stimulation presented in their blind visual field; yet, their 'guesses' of whether an object is in this part of the visual field yield accuracy rates that are often over 90%. This astonishing effect, 'blindsight', has been observed in both monkeys and humans with either unilateral or bilateral damage to the primary visual cortex. (*See* **Blindsight**)

The phenomenon of blindsight suggests that the disruption of primary visual cortex results in a disturbance of visual consciousness; hence, intact primary visual cortex seems necessary for visual consciousness. Although lacking conscious representations of visual stimuli, blindsight patients seem to be 'aware without being aware', which bears upon how attention and consciousness are related. For detailed information into the experiments on patients with blindsight, we refer the reader to the excellent summary by Weiskrantz (1997). (*See* **Consciousness, Function of**)

Although the incidence of damage confined to primary visual cortex is rare, the research that has been performed on blindsight patients is consistent with experimental work on nonhuman primates. Two types of tasks have been used with human blindsight patients. One relies on the patient's conscious report. In this method, one of two possible stimuli is presented on each trial: on half the trials the stimuli are presented to the blind visual field, and the person is instructed to decide which of the two stimuli were presented. The blindsight participants always report seeing nothing in the blind visual field; however, they are able to report some information about or respond to the stimulus in the blind visual field (although they report that there is nothing there, they will eventually guess). These patients can orient their eyes or hands to the approximate position of a stimulus; they can also discriminate stimulus orientation or stationary versus moving objects, as well as the direction of a moving stimulus.

The other method of testing blindsight patients involves measuring whether information presented to the patient's blind visual field influences (i.e. primes) the subsequent material presented to their intact hemifield. Results suggest that blindsight patients often experience implicit processing of stimuli presented in their blind hemifield. For example, information presented to a participant's blind field (e.g. the word 'river') significantly changes the way processing occurs for stimuli presented in the intact hemifield (e.g. the word 'bank'). In this case, when an ambiguous word ('bank') was presented in the good visual field, its conscious meaning was significantly prejudiced by a previous presentation of a semantic relative in the bad visual field. This finding suggests that the primary visual cortex may be unnecessary for some types of unconscious processing (such as priming) that influences our conscious experience.

These results suggest that much of the information that is sensed in the world can be processed without conscious awareness. Does the processing of this information rely on attentional processes even without conscious awareness? Alternatively,

can attention work in the absence of conscious awareness? One telling experiment on a blindsight patient addressed this very issue (Kentridge *et al*., 1999). The researchers asked whether orienting of attention would occur in the blind visual field. Using cuing experiments, blindsight patients were given an arrow cue that either correctly or incorrectly predicted the location of a subsequent target. The cue and the target could occur in either the good or blind visual field. As with normal participants, there was a benefit in performance for targets that occurred in the validly cued location, even when the target occurred in the blind visual field and the patient reported no target but 'guessed' that a target was present. Even more revealing, there was a benefit for the validly cued target location even when the cue itself occurred in the blind visual field and was not 'seen'. These results suggest that attention and consciousness are not absolutely linked. Even in the presence of attentional benefits (speeded and more accurate response to a validly cued target), conscious awareness of the stimulus is not guaranteed. Nor is conscious awareness of the cue necessary for attentional benefits. Hence, attention is not a sufficient conduit to ensure conscious awareness, and awareness is clearly not necessary for attentional processing. (*See* **Attention, Neural Basis of**)

In our view, attention is one agency for conscious awareness of each stimulus. Attention works to increase the neural response of salient, task-relevant stimuli. A shift of attention (and the increased neural activity) brings a stimulus into conscious awareness; however, attention can act on a stimulus without bringing that stimulus into consciousness. As suggested in the section on binocular rivalry, the level of neural activity to attract attention is likely to be below that necessary for conscious awareness. While it is clear that attention can be sufficient to bring a stimulus into consciousness, further research is required to determine what the necessary conditions are for conscious awareness of a stimulus.

References

James W (1890) *Principles of Psychology*. New York, NY: Holt.

Kentridge RW, Heywood CA and Weiskrantz L (1999) Attention without awareness in blindsight. *Proceedings of the Royal Society of London Series B* **266**: 1805–1811.

Weiskrantz L (1997) *Consciousness Lost and Found*. Oxford, UK: Oxford University Press.

Further Reading

Allport A (1988) What concept of consciousness? In: Marcel AJ and Bisiach E (eds) *Consciousness in Contemporary Science*, pp. 159–182. New York, NY: Oxford University Press.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Farber IB and Churchland PS (1995) Consciousness and the neurosciences: philosophical and theoretical issues. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 1295–1306. Cambridge, MA: MIT Press.

Lumer ED (2000) Binocular rivalry and human visual awareness. In: Metzinger T (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, pp. 231–240. Cambridge, MA: MIT Press.

Marcel AJ (1983) Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237.

Norman DA and Shallice T (1986) Attention to action: willed and automatic control of behavior. In: Davidson RJ, Schwartz GE and Shapiro D (eds) *Consciousness and Self-regulation*, vol. 4, pp. 1–18. New York, NY: Plenum Press.

Posner MI (1994) Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Sciences of the USA* **91**: 7398–7403.

Posner MI and Rothbart MK (1991) Attentional mechanisms and conscious experience. In: Milner AD and Rugg MD (eds) *The Neuropsychology of Consciousness*, pp. 91–111. London, UK: Academic Press.

Consciousness and Higher-order Thought

Intermediate article

David M Rosenthal, City University of New York Graduate School, New York, New York, USA

CONTENTS

The higher-order-thought hypothesis is a proposed explanation of what it is for a mental state to be a conscious state and hence of how conscious mental states differ from mental states that are not conscious.

INTRODUCTORY

Any satisfactory theoretical treatment of consciousness must begin by distinguishing several phenomena to which the term 'consciousness' applies. We describe people and other animals as being conscious when they are awake and responsive to sensory stimulation. What it is for a creature to be conscious in this sense is primarily a biological matter and peripheral to cognitive science and related theory.

We also describe creatures as being *conscious of* various things, for example, when they sense those things or think about them as being present. Sensing and thinking are central to cognitive functioning, but their nature is not what theorists typically have in mind in discussing consciousness. Rather, theorists have in mind primarily a third application of the term 'consciousness', by which we describe thinking and sensing itself as being conscious or not. It is this third use which dominates theoretical discussion about consciousness. The central issue is what it is for a mental state, such as thinking, sensing, and feeling, to be conscious, and more specifically what distinguishes the conscious cases from those which are not.

THEORIES OF CONSCIOUSNESS

It is fundamental to a mental state's being conscious that the individual in the state is aware of being in it. This is clear from consideration of mental states an individual is unaware of. If somebody is altogether unaware of thinking, feeling, or sensing something, that thinking, feeling or sensing does not count as conscious. Part of what it is for a state to be conscious is that one is conscious of being in that state.

Some theorists deny this, arguing that we are never conscious of our conscious states (Searle, 1992), or at least that conscious states occur without one's being conscious of them (Dretske, 1995). Thus Dretske, for example, urges that a state's being conscious consists not in one's being conscious of it, but in one's being conscious of something by virtue of being in that state.

This view has a disadvantage. Since sensing and thinking about things typically make one conscious of them, such states could not, on this view, occur without being conscious. Such theorists accordingly argue that the usual examples given of mental states that are not conscious are unconvincing. One especially common type of example does seem vulnerable to this charge. Armstrong (1978/1980) and others have appealed to the case of the long-distance driver who seems for a time not to notice the road consciously. But it may be that the driver notices the road consciously but simply does not at all remember doing so.

There are, however, other examples of mental states that more indisputably occur without being conscious. People often act in ways that betray some feeling, or belief, or desire of which they are wholly unaware until it is pointed out to them; this even happens with pains that are revealed by gestures or bodily movements. And people sometimes respond in a very fine-grained way to things that

occur so far in the periphery of their visual field that they have no conscious perception of them.

Many experimental results confirm these commonsense observations (Merikle *et al*., 2001). In masked-priming experiments, subjects presented very briefly with two successive stimuli report not being aware of the first at all, even though that stimulus has a demonstrable effect on mental processing (e.g. Marcel, 1983a, 1983b). And blindsight subjects, in whom part of the primary visual cortex has been destroyed, deny seeing visual stimuli in the relevant area of the visual field, though they can be prompted to guess the visible characteristics of such stimuli with startlingly high accuracy (Weiskrantz, 1997). Though conscious sensing is absent in these cases, subjects' behavior indicate the occurrence of sensing that is not conscious.

Some theorists have argued that it is circular to explain a mental state's being conscious in terms of an individual's being conscious of that state, since that would be to explain consciousness by appeal to consciousness (e.g. Goldman, 1993). But that explanation is not circular. Being conscious of something is sensing it or thinking about it as present. And, since we understand what it is to sense something or think about it even when that sensing or thinking is not conscious, we understand what it is to be conscious *of* something independently of knowing what it is for mental states to be conscious.

Even if a state's being conscious consists in one's being conscious of that state, theories divide about just how one is conscious of one's conscious states. The traditional and most widespread view is that one senses or perceives one's conscious states. But thinking about something can also make one conscious of that thing, and an alternative theory has been developed on which we do not sense our conscious states, but instead are conscious of them by having thoughts about them. It is useful to refer to the thoughts or sensations in virtue of which one is conscious of one's mental states as 'higher-order thoughts' or 'higher-order sensations'.

THE INNER-SENSE MODEL

The idea that we sense our conscious states has a long history. Locke (1700/1975) speaks of an 'internal sense' by which we are conscious of our mental states, and Kant (1787/1998) speaks of 'inner sense.' More recently, the idea has been defended by Armstrong (1978/1980) and by Lycan (1996).

Several factors suggest an account in terms of such higher-order sensing. For one thing, nothing seems to mediate between the the things we sense and our sensing them. And this intuitively unmediated character of sensing might explain why the way we are aware of our conscious states seems to be direct and immediate.

Another factor has to do with the qualitative character of conscious sensory experience. That qualitative character enters our mental lives through sensing; thinking has no qualitative character. So it may seem that the only way we could be conscious of this qualitative aspect of experience is by sensing it. A third source of the idea that we sense our conscious states is the sense we have that we are regularly and reliably conscious of many of our own mental states. And the best explanation for this may be that we monitor our mental states in the way that our exteroceptive senses monitor the environment (Armstrong, 1978/1980; Lycan, 1996).

But these considerations are far from decisive. Although nothing seems to mediate between our mental states and our consciousness of them, we need not appeal to higher-order sensing to explain that appearance of immediacy. Having thoughts about our mental states would also make us conscious of those states, and if we aware of nothing mediating between those thoughts and the states they are about, our consciousness of those states would also seem to be unmediated.

Perhaps some monitoring mechanism in the brain does subserve our being conscious of many of our mental states, but monitoring need not be sensory. The brain monitors many bodily functions in ways that do not at all resemble sensing. What differentiates sensing from other processes are the distinctive qualitative properties that occur when we sense. When sensing is conscious, we are conscious of these distinguishing qualities, qualities that vary with what is sensed, though these qualities also occur without our being at all aware of them.

The third consideration that seemed to support the inner-sense model, namely, the qualitative character of sensing, actually provides a compelling reason to reject the model (Rosenthal, 1997). Although sensations and perceptual states exhibit distinguishing qualitative properties, the way we are conscious of our own mental states does not. This is evident when the states we are conscious of are thoughts, beliefs, desires, and other so-called intentional states; these states have no qualitative properties, and there is no qualitative character to the way we are conscious of them. But even when the states we are conscious of are qualitative, as with our sensations and emotions, the qualities

belong to the states we are conscious *of*, not to the way we are conscious of them.

Some theorists describe the inner-sense model in terms of higher-order perceiving of mental states (Güzeldere, 1995). Since perceiving, like sensing, has qualitative character, a higher-order perception view faces the difficulty that no higher-order qualities occur. But perceiving not only has qualitative character, but also resembles thinking in having conceptual content. So, if we had higher-order perceptions of our mental states, we would still need to determine whether the qualities or conceptual content of the perceptions were responsible for our being conscious of our mental states. Compare Güzeldere (1995), who argues that the higher-order-perception model collapses into a model that invokes higher-order thoughts.

THE HIGHER-ORDER-THOUGHT MODEL

The two ways of being conscious of things are sensing them and having thoughts about them as being present. Since we are not conscious of our mental states by sensing them, the best explanation of how we are conscious of some of our mental states is that we have higher-order thoughts (HOTs) about them (Rosenthal, 1986, 1993; in press a, b). It would be explanatorily empty to insist that we are conscious of them in some third way unless we have an independent grasp of what that third way consists in.

Difficulties with the inner-sense view actually suggest the HOT model. The higher-order states in virtue of which we are conscious of our mental states lack qualitative properties, and a thought that something is present makes one conscious of that thing in a way that involves no higher-order qualities. If there is a brain mechanism that monitors mental states, it might well make one conscious of those states by producing HOTs about them. And, if those HOTs seemed to arise independently of any inference, it would seem subjectively as though one is conscious of one's mental states in a way that is direct and unmediated.

It is important to distinguish between the ordinary way in which mental states are conscious and the focused, reflective way in which states can become conscious when we introspect them. The HOT model affords a natural explanation of this difference. The HOTs in virtue of which we are conscious of our mental states in ordinary, nonintrospective cases are not, themselves, conscious thoughts; HOTs make one aware of various mental states, but without one's being conscious also of the HOTs themselves. When one introspects a state, one deliberately focuses attention on it. One thereby becomes aware not only of the introspected state, but also of one's being conscious of it. So in these cases the relevant HOTs are themselves conscious thoughts (Rosenthal, 2000a).

Because HOTs need not be conscious, and indeed usually are not, people will normally be unaware of their presence. The occurrence of HOTs is not established by our being aware of them, since we are conscious of them only in the special case of introspection. HOTs are theoretical posits whose occurrence is established by theoretical considerations of the sort sketched above.

These considerations help dispel a certain misunderstanding. It is sometimes held (e.g. Block, 1995a; Chalmers, 1996) that the HOT model explains only introspective consciousness. That would be so if the model appealed only to conscious HOTs, establishing their occurrence by way of subjects' reports. But the HOTs the model invokes typically are not conscious, and they are intended to explain ordinary, nonintrospective consciousness.

A HOT is a thought to the effect that one is in a particular state, and so makes reference to oneself. Such reference to the self does not require any sophisticated concept of the self, but only a concept strong enough to distinguish oneself from everything else and to form thoughts about the self thus distinguished (Rosenthal, in press c). Nor do HOTs need to describe their target states in terms of some concept of the mind; they can describe those states simply in terms of their role in perceiving, or thinking or in information-processing terms.

VARIANT HIGHER-ORDER-THOUGHT THEORIES

A number of variants HOT theories have been put forth. Some differ in only minimal ways from the hypothesis just described. For example, Mellor (1977–78) appeals to second-order beliefs to explain only what it is for beliefs to be conscious, and denies that such an explanation works for any other types of mental state. But the foregoing arguments apply equally to all types of mental state. And Rolls (1998) has argued for a model on which the HOTs must be linguistic in character. Since Rolls construes being linguistic to cover any syntactically composite mode of representation, this again is at most a slight modification of the model.

Brentano (1874/1973) argued that the higher-order state in virtue of which one is conscious of a

mental state is internal to the target state in question. Brentano's examples are largely perceptual, which leads Brentano to see the higher-order state as being perceptual as well; so his view may be best construed as a variant of the higher-order sensing model. Others have advanced views, however, that posit HOTs that are internal to their targets (Kobes, 1996; Gennaro, 1996).

But no view on which HOTs are internal to their targets is likely to succeed. Intentional states are individuated not only by their content but also by mental attitudes, such as mental affirmation, doubt, wonder, hope, and the like. Just as no single state can have two distinct contents, so no single state can exhibit two distinct mental attitudes. The mental attitude of HOTs is always assertoric; HOTs affirm that one is in a particular state. But, if HOTs were internal to their targets, then a conscious case of wondering something would exhibit the mental attitudes both of wondering and of affirming. So HOTs cannot in general be part of their targets. Similar considerations apply to Brentano's perceptual variant, since every perceptual state belongs to some sensory modality, and none of those standard modalities is suitable for making one conscious of one's mental states.

On another variant of the model developed by Carruthers (1996, 2000), a state need not be the object of an actual HOT to be conscious; it is enough that the state simply be *disposed* to cause a HOT about it. One main motive for this variant is that it avoids the high cost, both in computational capacity and cognitive space, of having actual HOTs for each of one's conscious states.

But that consideration is not all that compelling. HOTs very likely take less to implement cortically than their targets, since a HOT simply represents one as being in a particular state; so their causal connections will likely be far less complex than those of the perceptual and cognitive states the HOTs are about. And, since cortical capacity is known to be far from fully utilized, the cost of implementing actual HOTs is unlikely to be significant. Introspection makes this objection seem more pressing than it is. Because we are never conscious of many thoughts at once, it seems that we could not have many HOTs. But HOTs are seldom conscious, and we could have many at once that are not conscious. And introspection cannot be a reliable guide to the mind's nonconscious operations.

The principal reason for a higher-order account is to explain how we are conscious of all our conscious states. But being disposed to have a thought about something does not make one conscious of that thing. So the question arises about how a state's being disposed to cause a HOT could result in one's being conscious of that state.

Carruthers's answer appeals to a particular theory of mental content. On that theory, the content a state has is a matter of what it is disposed to cause. So Carruthers argues that a state's simply being disposed to cause a HOT can confer suitable higher-order content on the state itself. Both teleological (e.g. Millikan, 1984) and inferential-role (e.g. Block, 1986; Peacocke, 1992) theories of content might allow for this result. A state's having such higher-order content directed upon itself would then explain why one is conscious of being in that state.

This reply faces several difficulties. For one thing, these theories of content are far from uncontroversial, and it is preferable to have one's theory of consciousness committed to as little as possible that is not widely accepted. Moreover, since the higher-order content in virtue of which one is conscious of the state is internal to the state itself, the dispositional theory would face the difficulty about mental attitudes that faces any model on which the higher-order state is internal to its target.

Most important, any state with suitable first-order content would, on this model, have dispositional properties that result in its having higher-order content. So one would be conscious of any state that had that first-order content. Since a state's being conscious or not would depend on its first-order content, the model seems unable to explain how states of a given type can sometimes be conscious and sometimes not.

HIGHER-ORDER THOUGHTS AND SPEECH

It is widely accepted that, given a creature with suitable linguistic capacities, a mental state's being conscious coincides with that creature's ability to report noninferentially that it is in that state. Indeed, it is likely that this ability to report mental states noninferentially is what underlies the traditional intuition that we have special access to our mental states (Sellars, 1963).

This fits well with the HOT hypothesis. A noninferential report that one is in some mental state expresses one's thought that one is in that state, a thought that seems subjectively to rely on no inference. And it is arguable that the best explanation of this ability to report one's mental states noninferentially is that one actually has the HOTs that those reports would express (Rosenthal, 1993).

Some theorists hold that we cannot introspectively seem to be in a state that we are not in (e.g.

Nelkin, 1996). Moreover, the seemingly noninferential character of our HOTs may suggest that they reflect some special access we have to our mental states. But thoughts need not be accurate, and thoughts about one's own mental states are no exception. Our consciousness even of what states we are in can be erroneous.

This is evident from compelling experimental findings in which subjects report thoughts and desires that they do not actually have. As with reports of thoughts and desires that do occur, these confabulations tend to make *ex post facto* sense of subjects' behavior, by rationalizing that behavior or by conforming to expectations or preconceived ideas. But in these cases evidence exists that subjects do not actually have the thoughts and desires they report (Nisbett and Wilson, 1977). Such confabulation appears to happen even with qualitative states, such as bodily or perceptual sensations (Staats *et al*., 1998; Holmes and Frost, 1976).

These findings again fit well with the HOT model. When one confabulates being in some mental state, one is conscious *of oneself as being in that state*. And consciousness is a matter of how one appears to oneself. So, if one has a HOT that represents one as being in some state, there is nothing subjectively, from the point of view of consciousness, that could enable one to tell whether any such state actually occurs.

When one thinks that something has a certain property, one in effect interprets that thing as having the property. So having a noninferential HOT that one is in some mental state amounts to spontaneously interpreting oneself as being in that state. This echoes Dennett's (1991) interpretivist account of consciousness. But Dennett (1987, 1991) holds that one's being in a mental state at all, independent of whether that state is conscious, is a matter of one's being subject to some appropriate interpretation. The HOT model does not endorse that more general view.

Because conscious states are sometimes confabulated, the states one is conscious of oneself as being in do not always exist. So we cannot describe a conscious state as a state that bears some actual relation to a HOT. Rather, a state's being conscious must consist in its being the *intentional object* of a HOT, the object that the thought seems to be about. And, because the state may not actually occur, we also cannot require that it cause the HOT.

OBJECTIONS

The HOT hypothesis is sometimes seen as a claim about the *concept* of a mental state's being conscious (Goldman, 2000). Construed as a hypothesis about conceptual analysis the hypothesis is implausible, since it seems *conceivable* that a state accompanied by a HOT could fail to be conscious (Balog, 2000; Rey, 2000). But the HOT model is best taken not as a conceptual claim, but as an empirical hypothesis about the nature of consciousness. On that construal, though we can conceive of a state's being accompanied by a HOT without being conscious, it turns out empirically that this never happens. One might also object that any specification of the nature of consciousness purports to state a metaphysical necessity, and the HOT hypothesis is not metaphysically necessary. But, even apart from the difficulty of determining what is metaphysically necessary in a way that is not question begging, it is arguable that the HOT hypothesis is necessary if at all only in the way in which truths of natural science are.

It is sometimes argued that the stipulation that HOTs be noninferential is arbitrary, since it should not matter to a state's being conscious whether the accompanying HOT is caused by an inference (Byrne, 1997; Seager, 1999). But the aetiology of the HOT does not matter, only the appearance of aetiology. A state is conscious only if we are conscious of it in a way that *seems* spontaneous and noninferential. As long as it seems that way, it does not matter how it is caused. Nor is there any problem about establishing a causal or other connection between HOTs and their targets, as Natsoulas (1993) argues, since the targets are simply whatever states the HOTs are about.

Having a thought about something normally has no effect on it, and in particular does not make that thing conscious. So it may be objected that having a thought about a mental state could not result in that state's changing from not being conscious to being conscious (Block, 1995b; Byrne, 1997; Rey, 2000). But when a state becomes conscious that is not a change in the state itself, but only in whether one is noninferentially conscious of it; being conscious is not an intrinsic property of mental states.

Still, an objector might persist, since having a noninferential thought about a physical object does not result in that object or state's being conscious, why should having a HOT about a mental state result in that state's being conscious? But the only way objects might be conscious is the way a creature can be, by being awake and responsive to sensory input; objects cannot be conscious in the way mental states are. Still, having noninferential thoughts about states of one's liver presumably would not make those states conscious (Block, 1995b). But not every state can count as conscious.

A state can be conscious only if being in it, even when the state is not conscious, results in one's being conscious *of* something, and states of the liver do not qualify (Rosenthal, 2000b).

Dretske has argued that a mental state's being conscious cannot consist in one's having a HOT about it, since there are cases in which a state is conscious without one's being conscious of it. Dretske offers the example of consciously seeing two scenes that differ in some single way without one's consciously noticing that they differ at all. Since one does not notice that the scenes differ, one also does not notice the difference between one's conscious visual experiences of the scenes. But every part of the two experiences is presumably conscious. So, if one is not conscious of that part of the experiences in respect of which they differ, that part is a conscious experience of which one is not conscious (Dretske, 1995). But all that Dretske's case shows is that one need not be conscious of that part *as* the part that makes a difference between the two experiences, not that one is not conscious of that part in some other way (Seager, 1999; Byrne, 1997; Rosenthal, 1999). It may well be that we are conscious of all our conscious experiences.

Conscious states presumably occur not just in humans, but in other animals as well. So perhaps conscious states occur even in animals whose mental functioning is too primitive to accommodate HOTs (Block, 1995a; Dretske, 1995; Byrne, 1997). Indeed, Carruthers (2000) actually argues that few if any nonhuman animals have HOTs. And he concludes that they lack conscious states, though many will resist that conclusion.

In any case, the reasons for thinking that few nonhuman animals have HOTs are not fully convincing. Carruthers (2000) argues that animals with HOTs would also have thoughts about the mental states of others. And he holds that having thoughts about the mental states of others would express itself in deceptive behavior, which he urges nonhuman animals do not engage in (cf. Povinelli, 1996). But it is arguable that many nonhuman animals do engage in deceptive behavior (Whiten, 1996; Whiten and Byrne, 1997). Nor, in any case, is it obvious that creatures would not have HOTs unless they had thoughts about the mental states of others (Ridge, 2001).

It might seem that nonhuman animals lack the conceptual resources needed to have HOTs. But HOTs do not require the elaborate conceptual apparatus characteristic of humans; they are simply thoughts that one is in states of particular types, states which we humans classify as mental.

At the same time, it is not obvious which nonhuman species do have mental states that are conscious. Though many such species plainly do sense and think, that does not show that their thinking and sensing are conscious; states can exhibit the characteristic causal roles of mental states without a creature's being conscious of being in those states. Some way independent of human subjective impressions is needed to establish which species do have mental states that are conscious.

As noted above, it is one thing for a creature to be conscious and another for its mental states to be conscious. Still, it might seem that a creature cannot be conscious unless at least some of its mental states are conscious; whenever humans are awake, after all, they are in some conscious states. But this may not hold generally. For a creature to be conscious it must function in characteristic mental ways, but it can do that without its mental states being conscious.

It is natural to think that a mental state's being conscious serves some useful function, such as enhancing the rationality of thinking and planning (Nelkin, 1996). But the function a mental state serves is a matter of its causal role, and the causal role a state has may well be largely unaffected by being accompanied by a HOT (Dretske, 1995).

Accompanying HOTs might, however, actually alter a state's causal role, and a state together with a HOT will in any case have a different combined role from the state without any HOT. Indeed, it has been argued that HOTs enable the correction of plans that result from first-order processing (Rolls, 1998).

There are also experimental findings that subjects sometimes perform tasks better when stimuli are consciously perceived than when perceived nonconsciously (Merikle and Daneman, 1998). Since the relevant tasks require conscious thought, the difference may be due to operation of HOTs in the conscious cases.

There is a compelling intuition that our conscious states constitute some kind of unity, and one might object that a theory on which mental states are conscious in virtue of many distinct HOTs cannot do justice to that intuition (Shoemaker, in press). But since the content of each HOT is that one is, oneself, in some state, such reference to oneself will give rise to a conscious sense of unity (Rosenthal, in press c).

QUALITATIVE CONSCIOUSNESS

Perhaps the most important objection has to do with qualitative consciousness. It has been argued

that HOTs cannot capture the enormous detail characteristic of conscious qualitative states (Byrne, 1997). And some have argued also that HOTs, which are nonqualitative, could not result in there being something it's like for one to be in various qualitative states (Byrne, 1997; Siewert, 1998; Balog, 2000).

It is easy to exaggerate the qualitative detail we are conscious of at any moment. It is well known that Parafoveal vision yields scant detail. More dramatically, recent work on change blindness shows that we often fail consciously to notice significant changes in a scene we are attentively looking at (Grimes, 1996; Rensink, 2000; Simons, 2000), which suggests that our impression of great conscious qualitative detail is erroneous (Dennett, 1991). And HOTs could presumably capture the detail present in any relatively small area of a sensory field on which one consciously focused.

We do not have concepts for all the individual qualities we are conscious of, but we have concepts for the ways those qualities vary. So HOTs can represent individual qualities comparatively. This may explain why we can judge whether qualities are the same far better when they are all present than when we must rely on memory (Raffman 1995). And, though concepts may be ill suited to capture the way qualitative states represent things, we are typically conscious of the relevant qualities in a way that lends itself to conceptualized description.

It is important not to place excessive demands on an explanation of qualitative consciousness. Very likely no explanation will reveal a conceptual or rational connection between nonconscious resources and conscious qualities (Levine, 2001), but scientific explanation seldom does that. Nor should we expect to discover an introspectible connection between conscious qualities and nonconscious resources, since nothing that is not conscious is available to introspection.

In any event, there is reason to think that HOTs do figure in there being something it's like to be in conscious qualitative states. We often come to be conscious of qualitative differences only when we come to have concepts fine-grained enough to draw those qualitative distinctions: for example as between similar musical instruments or tastes of wine. Such concepts would matter to how those experiences are conscious only if our thoughts about the experiences made a difference to how we are conscious of them (Rosenthal, in press a).

THE SCIENCE OF CONSCIOUSNESS

Although the foregoing arguments in support of the HOT hypothesis do not rely on empirical investigation, the hypothesis meshes fruitfully with scientific findings. Two examples already noted are change blindness and confabulated mental states. But there are others as well. As Weiskrantz (1997) has urged, a HOT model helps explain the phenomena of blindsight. Rolls (1998) argues for his linguistic version of the HOT hypothesis by appeal to different neural pathways that seem to subserve conscious and nonconscious stimulation. Dienes and Perner (2001) have appealed to the HOT model in distinguishing implicit from explicit knowledge and representation, and Dienes (in press) has applied the model in connection with implicit learning and subliminal perception.

The HOT model is particularly useful in explaining the finding by Libet (Libet, 1985) that the neural readiness potentials identified with subjects' decisions occur measurably in advance of subjects' awareness of these decisions, findings recently replicated and extended (Haggard and Eimer, 1999). It is natural to explain this result by supposing that the HOTs in virtue of which subjects become aware of their decisions occur measurably later than those decisions (Gomes, 1999; Rosenthal, in press d).

Frith and Frith (1999) report a number of studies in which functional brain imaging reveals neural activation in subjects who were asked to report their mental states. Strikingly, conscious monitoring of states results in activation of a single brain area, medial frontal cortex, even when the states monitored are as disparate as pain, tickles, emotions aroused by pictures, and spontaneous thoughts. This activation does not occur cortically where the monitored states occur; so a single, independent brain mechanism seems to subserve the monitoring that makes possible the reporting of mental states. Since reports of one's mental states express one's thoughts about those states, it is inviting to construe that activation as indicating the occurrence of HOTs.

References

Armstrong DM (1978/1980) 'What is consciousness?'. *Proceedings of the Russellian Society* **3**(1978): 65–76; reprinted in expanded form in Armstrong, *The Nature of Mind*, St. Lucia, Queensland, Australia: University of Queensland Press, pp. 55–67, 1980.

Balog K (2000) Comments on David Rosenthal's 'consciousness, content, and metacognitive

judgments'. *Consciousness and Cognition* **9**(2) Part 1: 215–219.
Block N (1986) Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* **X**: 615–678.
Block N (1995a) 'On a confusion about a function of consciousness'. *The Behavioral and Brain Sciences* **18**(2): 227–247.
Block N (1995b) How many concepts of consciousness? *The Behavioral and Brain Sciences* **18**(2): 272–287.
Brentano F (1874/1973) *Psychology from an Empirical Standpoint* edited by Kraus O, English edn edited by McAlister LL, translated by Rancurello AC, Terrell DB and McAlister LL. London, UK: Routledge & Kegan Paul, 1973.
Byrne A (1997) Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies* **86**(2): 103–129.
Carruthers P (1996) *Language, Thought, and Consciousness: An Essay in Philosophical Psychology*. Cambridge, UK: Cambridge University Press.
Carruthers P (2000) *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge, UK: Cambridge University Press.
Chalmers DJ (1996) *The Conscious Mind*: *In Search of a Fundamental Theory*. New York, NY: Oxford University Press.
Dennett DC (1987) *The Intentional Stance*. Cambridge, MA: MIT Press/Bradford Books.
Dennett DC (1991) *Consciousness Explained*. Boston: Little, Brown and Company.
Dienes Z and Perner J (2001) When knowledge is unconscious because of conscious knowledge and vice versa. In: Moore JD and Stenning K (eds) *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, pp. 255–260. Mahwah, NJ: Lawrence Erlbaum Associates.
Dretske F (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press/Bradford Books.
Frith CD and Frith U (1999) 'Interacting minds – a biological basis'. *Science* **286**(i5445): 1692ff.
Gennaro RJ (1996) *Consciousness and Self-Consciousness: A Defense of the Higher-Order-Thought Theory of Consciousness*. Amsterdam and Philadelphia: John Benjamins.
Goldman AI (1993) Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition* **2**(4): 364–382.
Goldman AI (2000) Can science know when you're conscious? epistemological foundations of consciousness research. *Journal of Consciousness Studies* **7**(5): 3–22.
Gomes G (1999) Volition and the readiness potential. *Journal of Consciousness Studies* **6**(8–9): 59–76.
Grimes J (1996) On the failure to detect changes in scenes across Saccades. In: Akins K (ed.) *Perception*, pp. 89–110. New York, NY: Oxford University Press.
Güzeldere G (1995) Is consciousness the perception of what passes in one's own mind?. In: Metzinger T (ed.) *Conscious Experience*, pp. 335–357. Exeter: Imprint Academic. Reprinted in: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 789–805. Cambridge, MA: MIT Press/Bradford Books, 1997.
Haggard P and Eimer M (1999) On the relation between brain potentials and awareness of voluntary movements. *Experimental Brain Research* **126**(1): 128–133.
Holmes DS and Frost RO (1976) Effect of false autonomic feedback on self-reported anxiety, pain perception, and pulse rate. *Behavior Therapy* **7**(3): 330–334.
Kant I (1787/1998) *Critique of Pure Reason*, translated and edited by P Guyer and AW Wood. Cambridge, UK: Cambridge University Press, 1998.
Kobes BW (1996) Mental content and hot self-knowledge. *Philosophical Topics* **24**(1): 71–99.
Levine J (2001) *Purple Haze: The Puzzle of Consciousness*. New York, NY: Oxford University Press.
Libet B (1985) 'Unconscious cerebral initiative and the role of conscious will in voluntary action'. *The Behavioral and Brain Sciences* **8**(4): 529–539.
Locke J (1700/1975) *An Essay Concerning Human Understanding*, edited from the 4th edn. by PH Nidditch. Oxford, UK: Clarendon Press.
Lycan W (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press/Bradford Books.
Marcel AJ (1983a) Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237.
Marcel AJ (1983b) Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology* **15**: 238–300.
Mellor DH (1977–78) Conscious belief. *Proceedings of the Aristotelian Society*, New Series, LXXXVIII: 87–101.
Merikle PM, Smilek D and Eastwood JD (2001) Perception without awareness: perspectives from cognitive psychology. *Cognition* **79**(1–2): 115–134.
Merikle PM and Daneman M (1998) Psychological investigations of unconscious perception. *Journal of Consciousness Studies* **5**(1): 5–18.
Millikan RG (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press/Bradford Books.
Natsoulas T (1993) What is wrong with the appendage theory of consciousness? *Philosophical Psychology* **6**(2): 137–154.
Nelkin N (1996) *Consciousness and the Origins of Thought*. Cambridge, UK: Cambridge University Press.
Nisbett RE and Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review* LXXXIV(3): 231–259.
Peacocke C (1992) *A Study of Concepts*. Cambridge, MA: MIT Press/Bradford Books.
Povinelli DJ (1996) Chimpanzee theory of mind?: the long road to strong inference. In: Carruthers P and Smith PK (eds) *Theories of Theories of Mind*, pp. 293–329. Cambridge, UK: Cambridge University Press, 1996.
Raffman D (1995) On the persistence of phenomenology. In: Metzinger T (ed.) *Conscious Experience*, pp. 293–308. Exeter: Imprint Academic.

Rensink RA (2000) The dynamic representation of scenes. *Visual Cognition* **7**(1/2/3): 17–42.
Rey G (2000) Role, not content: comments on David Rosenthal's 'Consciousness, content, and metacognitive judgments.' *Consciousness and Cognition* **9**(2): 224–230.
Ridge M (2001) Taking solipsism seriously: nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies* **103**(3): 315–340.
Rolls ET (1998) *The Brain and Emotion*. Oxford, UK: Clarendon Press.
Rosenthal DM (1986) Two concepts of consciousness. *Philosophical Studies* XLIX(3): 329–359.
Rosenthal DM (1993) Thinking that one thinks. In: Davies M and Humphreys GW (eds) *Consciousness: Psychological and Philosophical Essays*, pp. 197–223. Oxford, UK: Basil Blackwell.
Rosenthal DM (1997) Perceptual and cognitive models of consciousness. *Journal of the American Psychoanalytic Association* **45**(3): 740–746.
Rosenthal DM (1999) Sensory quality and the relocation story. *Philosophical Topics* **26**(1 and 2): 321–350.
Rosenthal DM (2000a) Introspection and self-interpretation. *Philosophical Topics* **28**(2): 201–233.
Rosenthal DM (2000b) Metacognition and higher-order thoughts. *Consciousness and Cognition* **9**(2): 231–242.
Rosenthal DM (in press a) *Consciousness and Mind*. Oxford, UK: Clarendon Press, 2003.
Rosenthal DM (in press b) Explaining consciousness. In: Chalmers DJ (ed.) *Philosophy of Mind: Contemporary and Classical Readings*. New York, NY: Oxford University Press, 2002.
Rosenthal DM (in press c) Unity of consciousness and the self. *Proceedings of the Aristotelian Society* **103**(3) (2003).
Rosenthal DM (in press d) The timing of conscious states. *Consciousness and Cognition* **11**(2) (2002).
Seager W (1999) *Theories of Consciousness: An Introduction and Assessment*. London and New York: Routledge.
Searle JR (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press/Bradford Books.
Sellars W (1963) Empiricism and the philosophy of mind. In: *Science, Perception and Reality*, pp. 127–196. London, UK: Routledge & Kegan Paul.
Shoemaker S (forthcoming) Consciousness and co-consciousness. In: Cleeremans A (ed.) *The Unity of Consciousness: Binding, Integration, and Dissociation*. Oxford: Clarendon Press.
Siewert CP (1998) *The Significance of Consciousness*. Princeton: Princeton University Press.
Simons DJ (2000) Current approaches to change blindness. *Visual Cognition* **7**: 1–16.
Staats PS, Hekmat H and Staats AW (1998) Suggestion/placebo effects on pain: negative as well as positive. *Journal of Pain and Symptom Management* **15**(4): 235–243.
Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford, UK: Clarendon Press.
Whiten A (1996) When does smart behaviour-reading become mind-reading? In: Carruthers P and Smith PK (eds) *Theories of Theories of Mind*, pp. 277–292. Cambridge, UK: Cambridge University Press.
Whiten A and Byrne RW (1997) *Machiavellian Intelligence, II: Extensions and Evaluations*. Cambridge, UK: Cambridge University Press.

Further Reading

Armstrong DM (1968/1993) *A Materialist Theory of the Mind*. New York: Humanities Press; 2nd revised edn. London, UK: Routledge & Kegan Paul, 1993.
Carruthers P (1989) Brute experience. *The Journal of Philosophy* LXXXVI(5): 258–269.
Dienes Z and Perner J (2001) The metacognitive implications of the implicit–explicit distinction. In: Chambres P, Izaute M and Marescaux P-J (eds) *Metacognition: Process, Function, and Use*, pp. 241–268. Dordrecht, Germany: Kluwer.
Dretske F (1993) Conscious experience. *Mind* **102**(406): 263–283; reprinted in Dretske, *Perception, Knowledge, and Belief*, pp. 113–137. Cambridge, UK: Cambridge University Press, 2000.
Haggard P (1999) Perceived timing of self-initiated actions. In: Aschersleben G, Bachmann T and Musseler J (eds) *Cognitive Contributions to the Perception of Spatial and Temporal Events*, pp. 215–231. Amsterdam, Netherlands: Elsevier.
Kobes BW (1995) Telic higher-order thoughts and Moore's paradox. *Philosophical Perspectives* **9**: 291–312.
Levine J (1993) On leaving our what it's like. In: Davies M and Humphreys GW (eds) *Consciousness: Psychological and Philosophical Essays*, pp. 121–136. Oxford, UK: Basil Blackwell.
Libet B, Gleason CA, Wright EW and Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness potential). *Brain* **106**(Part III): 623–642.
Lurz RW (in press) Advancing the debate between HOT and FO theories of consciousness. *Journal of Philosophical Research* **28** (2003).
Mellor DH (1980) Consciousness and degrees of belief. In: Mellor DH (ed.) *Prospects for Pragmatism*, pp. 139–173 Cambridge, UK: Cambridge University Press.
Perner J and Dienes Z (in press) Developmental aspects of consciousness: How much theory of mind do you need to be consciously aware? *Consciousness and Cognition*.
Rensink RA (2000) Seeing, sensing, and scrutinizing. *Vision Research* **40**(10–12): 1469–1487.
Rosenthal DM (2000) Consciousness and metacognition. In: Sperber D (ed.) *Metarepresentation: A Multidisciplinary Perspective*, pp. 265–295. New York, NY: Oxford University Press.
Rosenthal DM (2000) Content, interpretation, and consciousness. In: Ross D, Brook A and Thompson DL

(eds) *Dennett's Philosophy: A Comprehensive Assessment*, pp. 287–308. Cambridge, MA: MIT Press/Bradford Books.

Rosenthal DM (in press e) Why are verbally expressed thoughts conscious?

Seager W (1994) Dretske on HOT theories of consciousness. *Analysis* **54**(1): 270–276.

Weiskrantz L (1986) *Blindsight: A Case Study and Implications*. Oxford, UK: Clarendon Press.

White PA (1988) Knowing more than we can tell: 'Introspective access' and causal report accuracy 10 years later. *British Journal of Psychology* **79**(1): 13–45.

Consciousness and Representationalism

Intermediate article

Benj Hellie, Sage School of Philosophy, Cornell University, Ithaca, New York, USA

CONTENTS

The representationalist theory of consciousness is the view that consciousness reduces to mental representation. This view comes in several variants which must explain introspective awareness of conscious mental states.

INTRODUCTION

Some mental states and processes are like something to their subjects; others are not. For instance, the states of seeing a stop sign, of hearing a screech, and of smelling gasoline are like something; as are the states of feeling fear, elation, or pain; as is the process of talking oneself through a problem. In contrast, states and processes that are not like anything to their subjects are accepted by both scientific and common sense psychology. Chomskian linguistic theories and Marrian theories of vision posit complex subpersonal operations, which make a difference to what one's mind is like to one only by their effects; common sense recognizes states of believing and intending that persist through dreamless sleep. States and processes that are like something to their subjects are conscious; otherwise not.

Among conscious states, what they are like to their subjects can differ: what seeing a red thing is like is standardly different from what seeing a green thing is like; what both are like differs from what smelling gasoline is like. A state has a 'phenomenal character' just in case it is conscious, or like something to its subject; two states have the same phenomenal character just in case what one is like to its subject is the same as what the other is like to *its* subject.

Phenomenal characters pose special problems for a fully naturalistic theory of the mind, for it may seem baffling how these properties can arise ultimately from interactions of particles and fields, or from processes in the brain. Wittgenstein famously wondered how *this* – his then current headache – could be a brain state; such bafflement is a proper reaction to the great difference in the ways in which phenomenal characters present themselves when thought of as phenomenal characters from the ways in which brain properties present themselves when though of as brain properties.

Representationalism is a view that attempts to naturalize phenomenal character without generating such bafflement by adopting a two-stage naturalistic reduction. The representationalist hopes that an intermediate reduction to certain representational properties will not generate bafflement; and that these representational properties may be reduced in turn, through one of the many ambitious projects for naturalizing mental representation.

WHAT IS REPRESENTATIONALISM?

Representationalism is the view that phenomenal characters somehow reduce to representational properties. The notion of a representational property deserves some expansion.

A state has a representational property when, to put it intuitively, it has a meaning or somehow stands in in some process for something else, such as an object, or a 'proposition' – a putative fact. Paradigmatic mental representational states are beliefs: one who believes that snow is green is in a state which means that snow is green, and which stands in for the putative fact that snow is green in a subject's reasoning. Another example is the state of thinking of Vienna: such a state means, or is about, Vienna; and stands in for Vienna itself in the subject's reasoning. Belief and thought-about are known to common sense psychology; scientific psychology also posits representational states: in some linguistic theories, for instance, in a many stage process of linguistic comprehension, a

language-processing module goes into states which represent phonological, syntactic, and semantic properties of heard sentences.

Hoping that snow is green differs from believing that snow is green, although both are representational states and concern the proposition that snow is green. Standard philosophical theories consequently take representational states to involve a relation between a subject and a 'content' – what is meant – via an attitude or the relation borne to that meaning. When one believes that snow is green, the attitude is belief; when one hopes that snow is green, the attitude is hope. A representational property of a representational state may thus be characterized as a pair composed of an attitude and a content.

Representational states have correctness, or satisfaction, conditions partly determined by the correctness conditions for their contents. A proposition is correct just in case it is true; correctness conditions for other sorts of contents, such as Vienna, are less well understood by philosophy. So, for instance, a belief is correct just in case its content is; a desire or hope is satisfied just in case its content is correct.

VARIETIES OF REPRESENTATIONALISM

This crude formulation allows for a good deal of variation, along at least three dimensions.

What Sort of Reduction?

Any attempt at reduction may be more or less ambitious. This ambition influences the relation taken to hold between the reduced entity and the reducing entity.

The weakest interesting relation for reductive purposes is 'supervenience': the reduced entity cannot vary without variation in the reducing entity. Supervenience seems to be a necessary condition for reduction of any sort; whether it is sufficient is hotly debated.

A stronger thesis brings about an ontological reduction by identifying particular phenomenal characters with particular representational properties: for each phenomenal character ϕ there is a representational property ρ such that for a state to be ϕ is for it to be ρ; moreover, the property's status as representational is somehow fundamental, whereas its status as phenomenal is somehow more superficial.

A still stronger thesis brings about an epistemological or explanatory reduction by claiming the relevant identities to be *a priori* (under canonical ways of conceptualizing the properties in question).

Which Phenomenal Characters are Reduced?

The many different sorts of conscious states canvassed in the introduction have a wide variety of phenomenal characters: perceiving is unlike imagining; feeling sad is unlike feeling a physical pain. A representationalist may attempt to reduce all phenomenal characters, or only some privileged set of them, such as experiences of visually perceiving color.

There may be some purposes for which a limited theory would be interesting, such as that of avoiding a perceptual epistemology of sense-data (Russell, 1912; Harman, 1990). However, if the main purpose of representationalism is bringing consciousness under a unified naturalistic umbrella, less ambitious theories with narrower scope are less interesting; moreover, less ambitious theorists are forced to explain what it is about those special phenomenal characters which makes them susceptible to representationalist treatment when others are not.

First-Order and Higher-Order Representationalism

Not all representational states give rise to consciousness: e.g. sound sleepers continue to store memories. Which do?

Some mental states represent other mental states: I can think about my thinking about Vienna. Here the thought about Vienna is 'first order'; the thought about the thought is 'higher order'.

According to first-order representationalism (Harman, 1990; Tye, 1995) (sometimes called 'intentionalism'), some representational states that do not concern other mental states, such as seeing a green tree, are by their nature sufficient to give rise to phenomenal character. First-order representationalists identify phenomenal characters with a pair of a content and an attitude. First-order representation historically developed with the partial intent of avoiding a sense-datum epistemology (Harman, 1990), so that advocates of the view are often concerned to show that any conscious content must concern nonmental reality; but there is no obvious reason why a naturalist must hold this nonmentalist position: represented mental states might be themselves natural, and themselves represented as instancing natural properties.

By contrast, according to higher-order representationalism self-representation is necessary for consciousness, so that first-order states cannot by themselves give rise to consciousness (unless the first-order state is essentially such as to be self-representing; more below). There are several dimensions of variation in higher-order theory.

Perhaps the higher-order attitude is belief (Rosenthal, 1997); perhaps it is perception (Lycan, 1997; Lormand, 1994); perhaps it is a form of Russellian 'acquaintance' (Russell, 1912), which could be thought of as a relation which grounds the meanings of demonstrative concepts such as 'this' and 'thus'; further options are certainly available.

Moreover, there is room for variation in the causal relation the representing state bears to the represented state: perhaps no constraint is required, or perhaps a tight constraint, along the lines of that necessary for veridical perceiving is required. Or, alternatively, perhaps the representing and the represented states are in a tighter metaphysical relation of partial constitution.

Finally, though it would seem natural for the higher-order representationalist to take the phenomenal character of a state to be determined by how it is represented by the higher-order state, the fact that some conscious states are themselves representational gives rise to a choice here. What if the content of the lower-order state is misrepresented by the higher-order state – so that, for instance, an experience of seeing a red thing is misrepresented as an experience of seeing a green thing? Neither option is happy: if the higher-order content determines phenomenal character, although the subject would say 'that's red', he would seem irrational to himself in doing so; if the lower-order content determines phenomenal character, the higher-order content seems otiose. This dilemma can be dissolved if either the higher-order attitude is infallible, or if only noncontent properties of the lower-order state are represented.

A Mixed View

Finally, first- and higher-order views can be combined, if the 'special' first-order attitude is one which essentially involves self-representation: one bears this special attitude A to a content c only if one bears some further attitude A' to one's bearing A to c. If $A = A'$, an infinite hierarchy of bearings of A result (more on this point follows in the subsection: 'A Russellian view').

ARGUMENTS FOR REPRESENTATIONALISM

Higher-Order Representationalism

Higher-order representationalism can seem truistic. Intuitively, the phenomenal character of a mental state makes some impact on the subject's awareness: the idea of a state which has phenomenal character, but of which the subject is not in any way aware, is bizarre in the extreme. The impact need not consist in the presence of an occurrent opinion about which phenomenal character one is currently enjoying: a daydreamer might fail to notice subtle shifts in visual experience resulting from the gradual descent of the sun. However, in subjects with the conceptual capacity for such thoughts, the ground of such thoughts must be present (more on subjects without such capacity follows in the subsection: 'A Russellian view').

More must be done, of course, to specify what such grounding amounts to, and what it is to have an opinion about which phenomenal character one is currently enjoying. However, an analogy to perception may prove a fruitful source of investigation: just as perception provides the ground for occurrent thoughts about which colors and shapes are before one by serving as a stock of representations distinct from occurrent thought, so may awareness of phenomenal character do so by serving as a stock of representations distinct from occurrent thought.

Higher-order representationalism seems to be a commitment of the common idioms of consciousness. We say that a conscious state is 'like something to its subject'; under analysis, this predicate is revealed to apply to a state just in case the subject represents the subject is acquainted with certain features of the state. On a slightly less common, but still natural, way of speaking of phenomenal character, we say that a state 'feels a certain way to its subject': here analysis is not needed to reveal that language draws an analogy between consciousness and perceptual representation. With a suitable theory of the link between truth-conditions and metaphysics, these observations could be extended to a proper argument for higher-order representationalism.

Phenomenology and First-Order Representationalism

The first source of support for first-order representationalism is an effect observed in

phenomenological investigation, commonly known as 'transparency' (Harman, 1990; Tye, 1995): allegedly, when one sees a blue bead, one cannot detect any 'intrinsic' or nonrepresentational property (aside from the bead's apparent property of blueness) making a difference to the phenomenal character of this experience of seeing. If phenomenal characters are as they seem, no nonrepresentational property does make a difference; and what applies for this experience is held to apply for all experiences.

This argument does not show anything deep about consciousness, however. Even if transparency holds for states of seeing, the phenomenal character of one's total experience is complex, and there are contributions made by further mental states one is in: transparency may fail for these. There is nothing incoherent about the idea of a nonrepresentational property of a mental state or process: a mental process might procede at a certain rate, or be subject to voluntary control with a certain number of degrees of freedom. Nor is there anything incoherent about the idea of such a property being introspectively detectable, and thereby influencing phenomenal character. Consequently, if transparency holds, it does so at best contingently. Moreover, it does not even seem to hold generally for actual human phenomenal character (see subsection: 'Straightforward counterexamples to the first-order view').

Epistemology and First-Order Representationalism

The second source of support for the first-order view appeals to a 'recycling' theory of concepts of phenomenal character. Allegedly, when one forms an introspective judgment about which phenomenal character one is enjoying, one singles out that phenomenal character only by reusing discriminative capacities already conferred by undergoing an experience with a certain first-order content, together perhaps with a highly general concept of mental states of a certain sort (Evans, 1982). So for instance, when one sees a blue bead, one singles out the phenomenal character of the experience of seeing the blue bead roughly by taking it to be that phenomenal character had by experiences which represent things as thus, where one's grasp of 'thus' is grounded in the experience itself: here, the material before the 'thus' is responsible for the concept's application to one among the phenomenal characters; 'thus' is responsible for distinguishing the phenomenal character from all others. The end result is to distinguish phenomenal characters in general as representational properties. Hence, if our introspective concepts of phenomenal characters are true to and exhaustive of the natures of phenomenal characters, phenomenal characters just are properties involving representing nonmental reality as a certain way. Some terminology: the perceptual state responsible for grasp of the concept 'thus' contributes 'novel' content; whichever state is reponsible for the concept of phenomenal character contributes 'recycled' content.

The recycling argument has the same flaw as the transparency argument. Even if some discriminations of phenomenal characters recycle novel content concerning nonmental reality, this is compatible with there being introspective concepts with novel content concerning mental states and processes.

Mentalism and Nonmentalism

These objections only concern nonmentalist formulations of first-order representationalism. The recycling and transparency arguments do seem to go through once they have been weakened to allow for the sorts of phenomena to be discussed in the next section.

ARGUMENTS AGAINST REPRESENTATIONALISM

Spectral Inversion and the First-Order View

The first-order view must explain the data that suport the higher-order view: perhaps this can be done by adopting the mixed view described earlier. Once this challenge has been met, two other objections arise.

Consider first a sample inversion argument (Block, 1990). Perhaps there are three possible subjects s_1, s_2, and s_3, such that s_1 and s_2 are alike phenomenally and differ from s_3, and s_2 and s_3 are alike representationally and differ from s_1. For the first condition, suppose that s_1 and s_2 are alike intrinsically in those respects which matter for phenomenal character, while s_3 differs intrinsically enough to make phenomenal character differ.

In support of the possibility of such a trio, many have felt a powerful intuition that phenomenal character supervenes on a subject's intrinsic nature. For the second condition, suppose that internal constitution does not much matter for

representational properties (the sign is arbitrary), so that the difference between s_2 and s_3 does not prevent them from being the same representationally – perhaps as a result of compensating divergences in their environments. A number of ways of establishing that there could be such compensating divergences in the presence of intrinsic likeness have been described in the literature; a typical attempt appeals to s_2 and s_3 being spectrally inverted with respect to one another but nonetheless deferring to the same experts in forming opinions about the colors of things. If thought content is deferential, and people standardly believe things are as they perceptually represent them to be, then s_2 and s_3 standardly perceptually represent things to be of the same color, violating supervenience.

The difficulty with this argument is that the premises both that thought content is deferential and that people standardly believe things are the way they perceptually represent them to be yield together the odd conclusion that perceptual representation is deferential: if perceptual content is nonconceptual, as many take it to be, it is unclear how deference could influence it. Moreover, however plausible deference may be for the concept 'red', it is less so for indexical predicate concepts such as 'thus'. Conversely, while 'thus' seems essentially tied to perceptual content, 'red', if subject to deference, is potentially less so.

The difficulties with this particular argument hold more generally: opinions about the narrowness of phenomenal character and the breadth of content may individually seem initially plausible, but the methodology for establishing such results has come under heavy attack in recent years. Moreover, even if color content is universally wide, the first-order representationalist may still retreat to the view that the phenomenal character-fixing first-order contents concern mental qualities, such as the degree to which certain retinal circuits are stimulated.

Straightforward Counterexamples to the First-Order View

Consider then a putative counterexample (for others see Peacocke, 1983): perhaps the most striking is double vision of the sort that results when one pokes one's eye with a finger. If one were to attempt to describe this experience, the most natural description would apply nonrepresentational predicates to one's own mental processes: one would say that one's visual field fragmented into two portions, which came to transparently overlay one another and move with respect to one another. If this description is correct, we seem to be introspectively aware of nonrepresentational properties of experience.

Mentalist first-order theorists may happily accept such examples. Anti-mentalist first-order theorists standardly reply to such counterexamples by redescribing the experience as one involving only ascription of properties to nonmental entities (Tye, 1995). For example: everything before me was suddenly replaced by a pair of transparent ghostly replicas of the scene before my eyes which then proceeded to move with respect to one another. Unfortunately, this description is implausible. The first-order representationalist should be concerned about this, for the transparency and recycling arguments both rely on a fairly high degree of privileged access to the nature of conscious mental states. Moreover, it is unclear what constrains such a strategy of redescription. Once initial plausibility has been set aside as a constraint, the position quickly threatens to become vacuous.

The Higher-Order View

Now consider the higher-order view. First, note that inversion-style objections can be readily modified to attack higher-order theory with more or less success.

Objections in the literature peculiar to the higher-order view have tended to focus less on the general higher-order approach than on particular hypotheses concerning the higher-order attitude. Of these, those receiving the most attention have been that the attitude is thought (the 'higher-order thought' view) and that it is perception (the 'inner-sense' view). A gamut of objections have been raised against each, too many to cover in detail. I will focus on the predictions these positions make concerning the nature of introspective knowledge of phenomenal character.

Epistemology and Higher-Order Thoughts

According to the higher-order thought view, a state is conscious just in case one has a belief about it; presumably the content of the belief concerns which phenomenal character the state has. Since one can open one's eyes without being flooded with an infinite hierarchy of conscious thoughts about one's perceptual states, the view must thus permit the higher-order beliefs to be themselves not conscious. The motivating idea behind higher-order representationalism was that a state has

phenomenal character only if its subject is in some sense aware of it, in a way that grounds conscious introspective thought about which phenomenal character it has. The higher-order thought theorist should thus regard the possession of the nonconscious higher-order thought as the sort of awareness which provides such a ground. Forming such a conscious thought should then be a matter of bringing the unconscious thought to consciousness.

Compare standard cases in which one brings a nonconscious thought to consciousness: one example is simply bringing a belief to consciousness; another is straining to recall someone's name. Both processes seem distinct from the processes by which one forms conscious thoughts about the phenomenal characters of one's conscious states. As discussed above, one can do so either by recycling content, or by a perceptionlike process. Either case is, intuitively, a matter of making a discovery – perhaps a relatively banal discovery but a discovery still – whereas according to higher-order thought theory, it is a matter of mere rethinking something one already knew.

Recycling and Inner Sense

According to the inner sense view, conscious states are themselves perceived. As noted above, the attractiveness of the idea that conscious states are perceived is encoded in idioms for discussing consciousness ('states that feel a certain way'). Since the concept of perception is not wholly clear, however, nor is the adequacy of the view. Suffice it to say that while the idea that we perceive double vision and other distortions has some plausibility, the analogy becomes rather strained when applied to experiences whose phenomenal characters are known by recycling.

A Russellian View

Thought and perception do not exhaust the space of possible attitudes one might bear to one's conscious states. A view according to which distortions are perceived, and perceptual states are known by recycling, would evade the concerns raised against higher-order thought and inner-sense views. Such a view could avoid being *ad hoc*, if perceiving, and that grasp of concepts of phenomenal character which underlies the capacity for recycling perceptual contents, can both be plausibly taken as determinates of a more inclusive notion of Russellian 'knowledge by acquaintance' (Russell, 1912). There are in fact substantial cognitive similarities between perceptual knowledge and knowledge by recycling that justify so treating them: both are sorted by content and modality, both seem to ground demonstrative and recognitional indexical concepts, and so forth. Then Russell's observation that when one stands in a relation of acquaintance one is in addition acquainted with this relation generates an infinite hierarchy of relations of acquaintance.

The existence of such a hierarchy is plausible: we can introspect, and introspect our introspection, and introspect our introspection of our introspection, and so forth. Nor does this proposal face the concern that scotched the higher-order conscious thought proposal; relations of acquaintance serve as grounds of potential conscious thought, and need not give rise to actual or occurrent conscious thought.

An objection to this approach is that if concepts of phenomenal character are required to get consciousness off the ground, the experiences of dogs and children are not conscious. The correct reply is to bite the bullet: after all, do we have any way of knowing that they are?

Zombies and Nonreductive Representationalism

A final objection to representationalism appeals to zombies: one could perhaps conceive of creatures functionally like us but which lack phenomenal characters; together with suitable principles passing from conceivability to possibility, supervenience would fail. Whatever one might think of conceivability–possibility principles, it is important to note that this argument at best threatens reductive versions of representationalism: if one regards it as inconceivable that there could be consciousness without awareness of phenomenal character, one might do better to accept a nonnaturalistic conception of representation than to give up the consciousness–representation link.

REPRESENTATION IN THE COGNITIVE SCIENCES

It would not be far-fetched to say that the cognitive sciences are only the study of computational operations on mental representation. The Chomskian revolution in linguistics began with the recognition that the (or at least a central) goal of linguistics is the study of the means by which the mind determines the semantic, syntactic, and phonological properties of sentences by running computational operations on mental representations of the

properties of those sentences. Marr's influential work on vision regards as the goal of the study of vision determining which operations must be performed on representations stemming ultimately from retinal stimulation, in order to generate representations of the properties of seen objects which have enough features to enable vision to do what it seems to do.

References

Block N (1990) Inverted earth. In: Tomberlin J (ed.) *Action Theory and the Philosophy of Mind*, vol. 4, *Philosophical Perspectives*, pp. 53–79. Atascadero: Ridgeview.

Byrne A (2001) Intentionalism defended. *The Philosophical Review* **110**: 49–90.

Evans G (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Harman G (1990) The intrinsic quality of experience. In: Tomberlin J (ed.) *Action Theory and the Philosophy of Mind*, vol. 4, *Philosophical Perspectives*, pp. 31–52. Atascadero: Ridgeview.

Lormand E (1994) Qualia! Now showing at a theater near you. In: Hill C (ed.) *The Philosophy of Daniel Dennett*, vol. 22, *Philosophical Topics*, pp. 127–156. Fayetteville, AR: University of Arkansas Press.

Lycan WG (1997) Consciousness as internal monitoring. In: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 755–771. Cambridge, MA: The MIT Press.

Peacocke C (1983) *Sense and Content: Experience, Thought, and Their Relations*. Oxford: Clarendon Press.

Rosenthal DM (1997) A theory of consciousness. In: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 729–753. Cambridge, MA: The MIT Press.

Russell B (1912) *The Problems of Philosophy*. Philadelphia: Hackett.

Tye M (1995) *Ten Problems of Consciousness*. Cambridge, MA: The MIT Press.

Further Reading

Armstrong DM (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block N (1978) Troubles with functionalism. In: Block N (ed.) *Readings in the Philosophy of Psychology*, vol. i. Minneapolis: University of Minnesota Press.

Byrne A (1997) Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies* **86**: 103–129.

Carruthers P (2000) *Phenomenal Consciousness*. Oxford: Oxford University Press.

Dretske FI (1995) *Naturalizing the Mind*. Cambridge, MA: The MIT Press.

Hilbert DR and Kalderon MK (2000) Color and the inverted spectrum. In: Davis S (ed.) *Color Perception: Philosophical, Psychological, Artistic, and Computational Perspectives*. Oxford: Oxford University Press.

Shoemaker S (1982) The inverted specturum. *Journal of Philosophy* **79**: 357–381.

Shoemaker S (1994*a*) Phenomenal character. *NOÛS* **28**: 21–38.

Shoemaker S (1994*b*) Self-knowledge and 'inner-sense': the Royce lectures. *Philosophy and Phenomenological Research* **54**: 249–314.

Thau MA (2002) *Cognition and Consciousness*. Oxford: Oxford University Press.

Consciousness, Animal

Intermediate article

Colin Allen, Texas A & M University, College Station, Texas, USA

CONTENTS

Observable similarities between humans and other animals with respect to behavior and neurology, as well as considerations related to evolutionary continuity between species, underlie most opinions that animals have conscious experiences. The attempt to understand other forms of consciousness – other species of mind – may help us to understand the evolutionary roots of our own.

INTRODUCTION

In the last 30 years many innovative experiments by comparative psychologists and ethologists, both in the laboratory and in the field, have improved our understanding of the cognitive capacities of animals (see the edited collection by Bekoff *et al*., 2002, for contributions by more than 50 leading researchers). As a result of this work we have acquired much knowledge about memory, learning, spatial navigation, social communication, and other cognitive capacities in a wide variety of species. Much of the work has concentrated on primates, and particularly on the great apes, reflecting a certain anthropocentric bias. But there has also been notable progress in understanding the cognitive capacities of other mammals, some birds, and even a few invertebrate species. Nonetheless there are large gaps in our understanding of the distribution of cognitive capacities across all taxonomic groups which limits our ability to understand the evolution of cognitive and mental abilities.

Opinions vary on the relevance of the work in cognition for the topic of consciousness *per se*. Some researchers regard consciousness as in internal, subjective state that is entirely beyond the range of scientific methodology. Others believe that existing scientific investigations shed light on conscious reasoning, self-awareness, and qualitative experience. By raising the issue of animal consciousness in a series of books, Donald Griffin (1976, 1984, 1992) deserves credit for having inspired the field that he named 'cognitive ethology'. Griffin promoted the idea, also to be found in Charles Darwin's work (1881; see Crist, 2002), that careful naturalistic observation and experiments under natural conditions can reveal the operation of mental processes in nonhuman animals, including invertebrates. Griffin's agenda has been strongly criticized, and his methodological suggestions often dismissed as anthropomorphic (see Bekoff and Allen, 1997, for a survey). But such criticisms may have overestimated the dangers of anthropomorphism (Fisher, 1990), and Griffin's suggestions have certainly acted as a catalyst for sophisticated work in cognitive ethology (Cheney and Seyfarth, 1990; Allen and Bekoff, 1997; Bekoff *et al*., 2002). Although many scientists remain skeptical of Griffin's approach, and the question of whether other animals are conscious remains controversial among them, there are indications that the topic is no longer entirely taboo (see the edited collection by Bekoff, 2000).

There are two senses of consciousness that cause particular controversy when applied to animals: phenomenal consciousness and self-consciousness. Phenomenal consciousness refers to the qualitative, subjective, experiential, or phenomenological aspects of conscious experience, sometimes identified with qualia. To contemplate animal consciousness in this sense is to consider the possibility that, in Nagel's (1974) phrase, there might be 'something it is like' to be a member of another species. Self-consciousness refers to an organism's capacity to represent its own mental states, and is often related to the question of whether the organism possesses a theory of mind, can reason about its situation, and plan accordingly.

PHENOMENAL CONSCIOUSNESS

Observable similarities between humans and other animals with respect to behavior and neurology, as well as considerations related to evolutionary continuity between species, underlie most opinions that animals have conscious experiences. For

instance, the behavior, neurology, and evolution of responses to noxious stimuli all appear to point to similar experiences of pain in humans and other animals. Noxious stimuli that humans would report as painful produce responses in animals that are easily mapped onto those of humans in similar circumstances. High-pitched vocalizations, fear responses, nursing of injuries, and learned avoidance are among the responses to noxious stimuli that are all part of the common mammalian heritage. The neural systems underlying these responses are virtually identical to those in humans, and animals respond to the same pain-relieving drugs, such as opiates. These responses are visible to some degree or other in organisms from several taxonomic groups.

Also in the realm of behavioral evidence, but less accessible to casual observation, are scientific demonstrations that members of other species, even of other phyla, are susceptible to the same visual illusions as we are (e.g. Fujita *et al*., 1991) suggesting that their visual experiences are similar. Correspondingly, much of the basic research that is of direct relevance to understanding human visual consciousness has been conducted on the very similar visual systems of monkeys. Monkeys whose primary visual cortex is damaged even show impairments analogous to those of human blindsight patients (Stoerig and Cowey, 1997) suggesting that the visual consciousness of intact monkeys is similar to that of intact humans (see Carruthers, 2000 for dissent).

Such similarity arguments are somewhat weak for it is always open to critics to exploit disanalogies between animals and humans to argue that the similarities do not entail the conclusion that both are conscious (Allen, 1998). Even when bolstered by evolutionary considerations of continuity between the species, the arguments are vulnerable, for the mere fact that humans have a trait does not entail that our closest relatives must have that trait too. Thus, for instance, Povinelli and Giambrone (2000) argue that even quite similar behaviors in humans and chimpanzees are due to different underlying mechanisms, a point that Povinelli believes is demonstrated by his research into how chimpanzees use cues to track visual attention (Povinelli, 1996; but see Hare *et al*., 2000, 2001 for a different view).

Despite the relevance of empirical research to similarity arguments, direct research into the phenomenological aspects of animal consciousness is hampered by at least two factors. One is the general problem that we lack a good theory of phenomenal consciousness, even in the human case. There is great uncertainty about the ontological status of phenomenal consciousness, whether in humans or in other animals. Accounts of consciousness in terms of basic neurophysiological properties, the quantum-mechanical properties of neurons, or *sui generis* properties of the universe are incomplete and controversial. More 'functionalist' accounts which attempt to explain the nature of experience in terms of the complex interactions between various cognitive subsystems fare no better. Consequently, no current theory of consciousness is secure enough to hang a decisive endorsement or denial of animal consciousness upon it.

A second factor hampering progress is that we lack a substantial account of what biological functions might be served by phenomenal consciousness. This point can be put another way: given any putative function of conscious experience, it seems we can imagine the same function being carried out without being accompanied by conscious experience. Thus, for example, while it is commonly asserted that the function of conscious pain is to 'tell' the organism when damage is occurring and thus promote behaviors which protect against further injury, it is also known that spinal reflexes alone, which are presumably quite unconscious, are sufficient to promote quite sophisticated kinds of withdrawal behavior and even associative learning (Grau, 2002). This raises the specter that phenomenal consciousness is epiphenomenal – completely devoid of physical effects – as some philosophers have asserted. If this were correct, then a search for the functions of consciousness would be doomed to futility. In fact, if consciousness is completely epiphenomenal then it cannot have evolved by natural selection, for selection can only operate on the effects of a trait on organismic fitness. On the assumption that phenomenal consciousness is an evolved characteristic of human minds, and therefore that epiphenomenalism is false, an attempt to understand the biological functions of consciousness would be useful for identifying its occurrence in different species by allowing us to search for organisms with the relevant functional capacities.

Many scientists remain convinced that no amount of empirical research can provide access to the subjective states of nonhuman animals. This remains true even among many scientists who are willing to invoke cognitive explanations of animal behavior that advert to internal representations. Opposition to dealing with consciousness can be understood as a legacy of behavioristic psychology: first, because of the behaviorists' rejection of terms for unobservables unless they could be formally

defined; and second, because of the strong association in many behaviorists' minds between the use of mentalistic terms and the twin bugaboos of Cartesian dualism and introspectionist psychology (Bekoff and Allen, 1997). In some cases these scientists are even dualists themselves, but they are strongly committed to denying the possibility of scientifically investigating consciousness, and remain skeptical of all attempts to bring it into the scientific mainstream.

Because consciousness is assumed to be private or subjective, it is often taken to be beyond the reach of objective scientific methods (see Nagel, 1974). This claim might be taken in either of two ways. It might be taken to bear on the possibility of answering the question of whether members of another taxonomic group (e.g. bats) have conscious states. Or it might be taken to bear on the possibility of answering the question of what it's like to be a member of another species. The difference between believing with justification that a bat is conscious and knowing what it is like to be a bat is important because, at best, the privacy of conscious experience supports a negative conclusion only about the latter (Bekoff and Allen, 1997). To support a negative conclusion about the former one must also assume that consciousness has absolutely no measurable effects on behavior, i.e. one must accept epiphenomenalism. But such an assumption leads to the implausible conclusion that phenomenal consciousness did not evolve by natural selection.

REASONING AND SELF-CONSCIOUSNESS

René Descartes argued against animal minds on the grounds that animals do not use language conversationally or reason generally. While he was aware of the capacity of parrots to pronounce human words, he dismissed this as unintelligent, meaningless repetition. This judgment may have been appropriate for the few parrots he encountered, but it was not based on a systematic, scientific investigation of the capacities of parrots. Some would argue that Irene Pepperberg's multiple studies of the African Grey parrot 'Alex' (Pepperberg, 1999) should lay the Cartesian viewpoint to rest. This work, along with research into the acquisition of linguistic competence by chimpanzees (e.g. Gardner *et al*., 1989; Savage-Rumbaugh, 1996; Fouts *et al*., 2002), might be interpreted as undermining Descartes' assertions about lack of intelligent language use and general reasoning abilities in animals. There are now some serious questions about the interpretation of this work, however. Cartesians have pointed out the limitations shown by animals in such studies, and they are often joined by linguists who protest that the subjects of animal-language studies have not fully mastered the recursive syntax of natural human languages (e.g. Pinker, 1994).

Teaching human languages to animals tests their capacities on a task that is well outside the natural repertoire for the species. The same is true of testing animals on the capacity for mirror self-recognition (Gallup, 1970; Gallup *et al*., 2002). It was long known that chimpanzees would use mirrors to inspect their images, but Gallup developed a protocol that appears to allow a scientific determination of whether it is merely the mirror image *per se* that is the object of interest to the animal inspecting it, or whether it is the image qua proxy for the animal itself that is the object of interest. Using chimpanzees with extensive prior familiarity with mirrors, Gallup anesthetized his subjects and marked their foreheads with a distinctive dye. Upon waking, marked animals who were allowed to see themselves in a mirror touched their own foreheads in the region of the mark significantly more frequently than controls who were either unmarked or not allowed to look into a mirror. Gallup's protocol has been repeated with other great apes and some monkey species, but besides chimpanzees only orangutans consistently 'pass' the test. (Shumaker and Swartz, 2002, cite preliminary evidence that the failure of gorillas may be one of motivation rather than basic intellectual capacity.) Gallup interprets this procedure as showing whether animals are self-aware and can infer the states of mind of another individual. (See Heyes, 1998, for a different opinion.)

The capacity to attribute mental states to others is often described as possession of a theory of mind. (But not always – see simulation theory.) The theory of mind debate has origins in the hypothesis that primate intelligence in general, and human intelligence in particular, is specially adapted for social cognition (see Jolly, 1966; Humphrey, 1976; Byrne and Whiten, 1988). Evidence for the theory of mind in great apes beside humans is mixed. Povinelli (1996) argues that in interactions with human food providers, chimpanzees apparently fail to understand the role of eyes in providing visual information to the humans, despite their outwardly similar behavior to humans in attending to cues such as facial orientation. The interpretation of Povinelli's work remains controversial. Hare *et al*. (2000) conducted experiments in which dominant and subordinate animals competed with each other for food, and concluded that 'at least in some

situations chimpanzees know what conspecifics do and do not see and, furthermore, that they use this knowledge to formulate their behavioral strategies in food competition situations.' They suggest that Povinelli may have obtained negative results because his experiments involved cooperative chimp–human interactions. Hare and Wrangham (2002) argue that situations where chimpanzees are competing with each other for resources may be more ecologically relevant than cooperative tasks.

It is also likely that the mirror test is not an appropriate test for theory of mind in most species because of its specific dependence on the ability to match motor to visual information, a skill for which there may have been little selectional pressure. Consequently, it has been argued that evidence for the ability to attribute mental states in a wide range of species might be better sought in natural activities such as social play, rather than in laboratory-designed experiments which place the animals in artificial situations (Allen and Bekoff, 1997; see especially chapter 6; see also Hare *et al*., 2000, 2001; Hare and Wrangham, 2001). Early attempts to find strong evidence of theory of mind in nonhuman animals under natural conditions generally failed to produce such evidence (see, e. g. Cheney and Seyfarth 1990). But anecdotal evidence (Byrne and Whiten 1988) and the more recent experimental results mentioned here tantalizingly suggest that researchers still have not managed to devise the right experiments.

CONCLUSION

Where does this leave questions about animal consciousness? While it may seem natural to think that we must know more about human consciousness before we try to determine whether other animals have it, such an approach may not be the most effective. The attempt to understand other forms of consciousness – other species of mind – may help us to understand the evolutionary roots of our own. Although research on primates is attractive because of their close evolutionary relationship to humans, much more needs to be learned about the cognitive abilities and forms of consciousness in less closely related species. It is important not to stifle research on consciousness because of worries about how to define the relevant notions. In the early stages of the scientific investigation of any phenomenon, putative samples are identified by rough rules of thumb (or working definitions) rather than complete theories. Early scientists identified gold by contingent characteristics rather than its atomic essence, knowledge of which had to await thorough investigation of many putative examples – some of which turned out to be gold and some not. Likewise, at this stage of the game, the study of animal consciousness may boldly investigate interesting cognitive capacities with no firm commitment to the idea that all these examples will involve conscious experience but absent of prejudice that none of them will.

References

Allen C (1998) The discovery of animal consciousness: an optimistic assessment. *Journal of Agricultural and Environmental Ethics* **10**: 217–225.

Allen C (2000) Animal consciousness. In: Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2000 Edition), [<http://plato.stanford.edu/archives/winter2000/entries/consciousness-animal/>]

Allen C and Bekoff M (1997) *Species of Mind*. Cambridge, MA: MIT Press. [See especially chapter 8.]

Bekoff M (2000) *The Smile of a Dolphin*. New York: Discovery Books.

Bekoff M and Allen C (1997) Cognitive ethology: slayers, skeptics, and proponents. In: Mitchell R *et al*. (eds) *Anthropomorphism, Anecdote, and Animals*, New York: SUNY Press.

Bekoff M, Allen C and Burghardt GM (eds) (2002) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Byrne RW and Whiten A (eds) (1988) *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford: Oxford University Press.

Carruthers P (2000) *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge, UK: Cambridge University Press.

Cheney DL and Seyfarth RM (1990) *How Monkeys See the World: Inside the Mind of Another Species*. Chicago, IL: University of Chicago Press.

Crist E (2002) The inner life of earthworms: Darwin's argument and its implications. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Darwin C (1881/1985) *The Formation of Vegetable Mould, through the Action of Worms with Observations on Their Habits*. Chicago: Chicago University Press.

Fisher JA (1990) The myth of anthropomorphism. In: Bekoff M and Jamieson D (eds) *Interpretation and Explanation in the Study of Animal Behavior:* vol. 1, *Interpretation, Intentionality, and Communication*. Boulder, CO: Westview Press. [Reprinted in Bekoff M and Jamieson D (eds.) (1996) *Readings in Animal Cognition*. Cambridge, MA: MIT Press.]

Fouts R, Jensvold ML and Fouts D (2002) Chimpanzee signing: Darwinian realities and Cartesian delusions. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Fujita K, Blough DS and Blough PM (1991) Pigeons see the Ponzo illusion. *Animal Learning & Behavior* **19**: 283–293.

Gallup GG Jr (1970) Chimpanzees: self-recognition. *Science* **167**: 86–87.

Gallup GG Jr, Anderson JR and Shillito DJ (2002) The Mirror Test. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Gardner RA, Gardner BT and Van Cantfort TE (1989) *Teaching Sign Language to Chimpanzees*. Albany, NY: SUNY Press.

Grau J (2002) Learning and memory without a brain. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Griffin DR (1976) *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. New York: Rockefeller University Press. [2nd edn, 1981.]

Griffin DR (1984) *Animal Thinking*. Cambridge, MA: Harvard University Press.

Griffin DR (1992) *Animal Minds*. Chicago: University of Chicago Press.

Hare B, Call J, Agnetta B and Tomasello M (2000) Chimpanzees know what conspecifics do and do not see. *Animal Behavior* **59**: 771–785.

Hare B, Call J and Tomasello M (2001) Do chimpanzees know what conspecifics know? *Animal Behavior* **61**: 139–151.

Hare B and Wrangham R (2002) The evolution of social cognition: comparative tests of the adapted cognition hypothesis. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Heyes C (1998) Theory of mind in nonhuman primates. *Behavioral and Brain Sciences* **21**: 101–148.

Humphrey N (1976) The social function of intellect. In: Bateson P and Hinde R (eds) *Growing Points in Ethology*. Cambridge, UK: Cambridge University Press. [Reprinted in Byrne and Whiten, 1988.]

Jolly A (1966) Lemur social behavior and primate intelligence. *Science* **153**: 501–506. [Reprinted in Byrne and Whiten, 1988.]

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **83**: 435–450.

Pepperberg IM (1999) *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Cambridge, MA: Harvard University Press.

Pinker S (1994) *The Language Instinct*. New York: William Morrow and Company.

Povinelli DJ (1996) Chimpanzee theory of mind? In: Carruthers P and Smith P (eds) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Povinelli DJ and Giambrone SJ (2000) Inferring other minds: failure of the argument by analogy. *Philosophical Topics* **27**: 161–201.

Savage-Rumbaugh S (1996) *Kanzi: The Ape at the Brink of the Human Mind*. New York: John Wiley and Sons.

Shumaker RW and Swartz KB (2002) When traditional methodologies fail: cognitive studies of great apes. In: Bekoff M *et al*. (eds) *The Cognitive Animal*. Cambridge, MA: MIT Press.

Stoerig P and Cowey A (1997) Blindsight in man and monkey. *Brain* **120**: 535–559.

Consciousness, Cognitive Theories of

Intermediate article

Bernard J Baars, Neurosciences Institute, San Diego, California, USA

CONTENTS

'Consciousness' operationally consists of all the things human beings report experiencing, from perception to mental images, inner speech, recalled memories, semantics, dreams, hallucinations, emotional feelings, and aspects of cognitive and motor control. In cognitive theory, consciousness appears to be a global access function, presenting an endless variety of focal contents to executive control and decision-making.

THE SCIENTIFIC REDISCOVERY OF CONSCIOUSNESS

Until recently consciousness was considered a scientifically intractable problem. That deep skepticism has faded with remarkable speed since the 1990s due to a flow of findings about the brain basis of conscious experience in perception, imagery, alertness, selective attention, working memory, episodic memory, and executive control (see Baars *et al*., in press, for 70 scientific articles on the topic). Today a scientific race to consciousness is well under way. While the traditional philosophical paradoxes of mind and body are not resolved, many scientists now believe that significant progress can be made in the empirical study of conscious experience.

EVIDENCE TO BE EXPLAINED

To study anything in science we need to treat it as a variable. The concept of gravitational force would have been useless had Newton been unable to imagine zero gravity. Likewise, to understand consciousness we need to compare at least its presence to its absence; consciousness without unconsciousness is meaningless. In cases such as brain damage we can observe several degrees of consciousness, from full alertness to massive coma. But we need to compare at least two levels of consciousness to be able to ask the question 'what is the difference between them?' Only then can we deal with the issue of consciousness *as such*.

G. A. Mandler (1984) has made the penetrating observation that science is obliged to treat consciousness not as an observable datum but as an inferred concept based on public evidence. In science we can observe only the public *reports* people make about their conscious experience. Often we can make very reliable inferences about human conscious experience based on such reports. Almost 200 years of perceptual studies show that such reports correspond with exquisite sensitivity to the sensory stimulus array. Entire domains of research depend upon this well-established methodology.

Unconscious representations can also be inferred from public observations, though people cannot report them accurately. The simplest example is the vast multitude of memories that are currently unconscious. You may recall this morning's breakfast – but what happened to that memory before it was brought to mind? Apparently it was still somehow represented in the brain, though not consciously. Yet it has been known since Ebbinghaus that unconscious memories can shape mental processes without ever coming to mind; for example, it is easier to memorize material that was previously learned, even if you do not explicitly remember having learned it before. A case can be made for unconscious knowledge of many things: habituated stimuli, memories before and after recall, automatic skills, implicit learning and memory, the rules of syntax and semantics, unattended stimulation, presupposed knowledge, pre-conscious input processing, visual recognition of facial affect, and more.

To study consciousness as a variable, we can for example compare the reader's currently conscious contents, such as *these printed words*, to previous

words in this sentence which are no longer conscious at the moment you are reading this. Notice that those previously conscious words must still be actively represented in memory to allow currently conscious words to be interpreted correctly. Thus the study of immediate memory presents many opportunities for treating consciousness as a variable. As in the case of Newtonian gravity, we can compare an event with its absence by keeping everything constant while varying the dimension of interest. This kind of contrastive study has now become routine, having been applied to waking consciousness (contrasted with sleep, coma, and general anesthesia), visual consciousness (compared to cortical blindness), consciousness of attended input (compared to unattended input), and much else (Baars *et al.*, in press). Taken together this sizable body of evidence serves to constrain theory in a very exacting way.

Even a simple set of contrasts brings out some important facts. The most prominent of these involve the remarkable limitations of conscious contents at any given moment, compared to the vastness and complexity of unconscious processes taking place at the same time. Consciousness is associated with limited capacity, seriality, and integration of multiple sources of information, while comparable unconscious processes involve much greater capacity, parallelism, and distributed autonomy. While conscious contents are limited at any given moment, they appear to facilitate access to multiple unconscious knowledge sources in the brain. Several current theories propose that a conscious event is made widely available to multiple brain mechanisms of memory, skill control, decision-making, semantics, anomaly detection, and the like. Thus consciousness may have a broad, architectural role as a gateway to multiple knowledge sources in the brain (e.g. Crick, 1984; Baars, 1988, 1997, 1998; Schacter, 1990; Chalmers, 1996; Damasio, 1989; John *et al.*, 1997; Edelman, 1989; Tononi and Edelman, 1999).

Fringe Conscious Events

William James (1890/1983) thought that vague or 'fringe conscious' events were at least as important as focally conscious ones. Fringe conscious phenomena include feelings of rightness, familiarity, beauty, coherence, anomaly, tip-of-the-tongue, attraction, repulsion, and the like. These phenomena can be operationally defined by a combination of high subjective certainty and high accuracy, but low experienced detail. Mangan (1993) has developed James's ideas about the fringe in modern terms, suggesting that fringe contents may not be subject to the classical capacity limitations of conscious experiences. Since focal conscious capacity is limited to one internally consistent experience at a time, Mangan sees fringe experiences as a very useful way of circumventing that limitation when needed.

COGNITIVE ARCHITECTURE THEORIES

A small cluster of theories has emerged to account for some of these aspects of conscious experience. Early theories were primarily cognitive, but quite sophisticated brain-based 'neuronal resonance' theories are now available (see below). Not surprisingly, the cognitive theories generally do a better job with psychological evidence, while neuronal resonance theories are better with the rapidly emerging evidence on the brain basis of conscious processes. A major theoretical aim for the future is to combine these approaches in a single, unified framework.

Cognitive theories can be divided into those that try to account for the evidence for consciousness as described above, and those that focus mainly on the various roles of consciousness in memory or executive functions. Baars' Global Workspace theory is the best-known example of the first class, while Schacter (1990) has developed a closely compatible approach to memory systems; Hilgard (1977), Johnson-Laird (1988), and Shallice (1988) focus on the role of consciousness in executive control.

Baars' Global Workspace (GW) theory extends the concepts of cognitive architectures to the problem of consciousness, in the tradition of A. Newell, H. A. Simon, and J. R. Anderson (see Baars, 1988, 1997, 1998). Specific mechanisms are proposed to deal with a sizeable array of psychological processes, from perception to imagery, spontaneous problem-solving, memory retrieval, goals and action control, and self. GW theory appears to be the most thoroughly worked out cognitive theory of conscious processes today, at a high level of description. Consciousness is associated with a global 'broadcasting system' that disseminates information widely throughout the brain. If this is true, then conscious capacity limits may be the biological price to be paid for the ability to make single momentary messages available to the entire system for purposes of coordination and control. Since at any moment there is only one 'whole system', a global dissemination facility must be limited to one momentary content.

GW theory relies on three theoretical constructs: unconscious specialized processors, a global workspace, and contexts. The first construct, the specialized unconscious processor, is an 'expert network' of which there are assumed to be many, working in parallel. There is direct evidence for many types of specialized systems in the brain. There are single cells, such as cortical feature neurons for color, line orientation, or faces, but also entire networks and arrays of neurons, such as cortical columns, cortical areas like Broca's or Wernicke's, large nuclei such as the thalamic relay nuclei, and so on. Like human experts, unconscious specialized processors may be quite limited in scope. They are extremely efficient in specific task domains, able to act autonomously or in coalitions with each other. They can receive global messages, and by mobilizing other experts by way of the global workspace they may be able to control mental or muscular activities. In routine tasks they may work autonomously, without conscious involvement, or they may display their output in the conscious global workspace. Answering a question such as 'What is your mother's maiden name?' requires a mission-specific coalition of unconscious experts, which return their answer to consciousness.

The second construct is the global workspace (GW) itself. A GW is an architectural capability for system-wide integration and dissemination of information. A global workspace is much like the bright spot of light on a theatre stage, cast by a spotlight in a darkened auditorium. Specialized unconscious processors are comparable to members of the audience sitting in the darkened theater. Groups of experts may interact locally, but in order to effect system-wide change they must compete for access to the bright spot on the theatre stage, perhaps supported by a coalition of other audience members. Once an expert reaches the bright spot on stage, it can broadcast a global message to the system as a whole. New links between unconscious experts are made possible by global interaction via the bright spot, and can then spin off to become new autonomous processors. The stage allows novel expert coalitions to form, to work on new or difficult problems which cannot be solved by existing experts and coalitions. Tentative solutions to new problems can then be globally disseminated, scrutinized, and modified. Since conscious experience seems to have a great perceptual bias, it is convenient to imagine that perceptual regions of cortex – visual, auditory, or multimodal – can compete for access to a brain version of a GW.

Theater models of consciousness have been criticized by Daniel Dennett and Marcel Kinsbourne on the ground that they are 'Cartesian' and conceptually flawed (Dennett and Kinsbourne, 1992). However, global workspace architectures have been implemented in artificial intelligence simulations for decades, and are not vulnerable to the Dennett–Kinsbourne critique. They do not assume a Cartesian point centre, for example, and both Dennett and Kinsbourne have refrained from applying their critique to global workspace theory.

'Context', the third construct in GW theory, refers to the powers behind the scenes of the theater of mind. Contexts are coalitions of expert processors that may function like a theatrical director, playwright, or stagehand who can influence the actors that appear in the spotlight. They can be defined empirically as *unconscious factors that shape conscious contents,* just as a director behind the scenes can influence the words and actions of actors on stage without being visible from the audience. Conceptually, contexts are defined as pre-established expert coalitions that can evoke, shape, and guide global messages without themselves entering the global workspace. Indeed, Dennett himself has recently endorsed a 'neuronal global workspace' approach to consciousnes (Dennett, 2001).

Contexts may be momentary or long-lasting. Momentary contexts may shape the reader's conscious interpretation of a word such as 'set', which has many different meanings. The word 'tennis' before 'set' shapes the interpretation of 'set', even when 'tennis' has already slipped from consciousness. But the word 'tennis' needed to be conscious initially to create the unconscious context that interprets 'set'. Contexts can also be long-lasting. The reader's ideas about consciousness from years ago may influence his or her current experience of this paragraph, even if those memories do not become conscious again. In general, major life experiences appear to set lifelong attitudes or character traits. Such major events typically influence current conscious experiences unconsciously, rather than being brought to mind. It is believed that a shocking or traumatic experience can also set up largely unconscious expectations that can shape subsequent conscious experiences.

Several proposals aim to cast the global workspace framework in a more neurally realistic form, giving special regard to thalamocortical mechanisms. Newman and Baars (1993) and Newman *et al*. (1997) described ways to integrate global workspace theory with the neuronal resonance theories discussed below. Newman *et al*. (1997, p. 1195) propose that 'One would expect the neural mechanism for global attention to be complex, and

widely distributed. … But the basic circuitry can be described, to a first approximation, in terms of repeating, parallel loops of thalamo-cortico-thalamic axons, passing through a thin sheet of neurons known as the nucleus reticularis thalami.' The overall framework suggests a neurocognitive model in which consciousness is viewed as a global integration and dissemination system operating in a large-scale, distributed array of specialized bio-processors, which controls the allocation of processing resources in the central nervous system.

Almost all current theories agree that consciousness involves system-wide functions, rather than local ones that are mostly unconscious, and that specialized 'expert' systems tend to be unconscious and relatively isolated. Some detailed implications have been worked out by way of computer simulations (Franklin and Graesser, in press).

Consciousness, Memory, and Self Functions

Other cognitive theories focus on functions that are associated with consciousness, such as working memory, episodic memory, and executive or 'self' functions. Hilgard's framework (e.g. 1977) is based on several decades of research on hypnotic dissociation, in which people under conditions of suggestion appear to lack conscious access to such things as sensory input, memories, normal voluntary control, or aspects of their own identity. Under experimental conditions it is often possible to show that these responses are not merely faked or simulated. Suggestible states are not merely oddities: a fifth of the normal population is highly suggestible, and all humans are suggestible under some circumstances. Hilgard points out a number of implications for normal executive control and access to self.

Johnson-Laird's (1988) operating system model of consciousness emphasized control functions such as directing attention, planning and triggering action and thought, and purposeful self-reflection. Johnson-Laird's cognitive architecture consists of a parallel processing system dominated by a control hierarchy. The system is a collection of largely independent processors (finite state automata), which cannot modify each other but which can receive messages from each other; each starts to compute when it receives appropriate input from any source. Each passes messages up through a hierarchy to the operating system, which sets goals for the subsystems. The operating system does not have access to the detailed operations of the subsystems – it receives only their output. Likewise, the operating system does not need to specify the details of the goals it transmits to the processors – these take the goal, abstractly specified, and elaborate it in terms of their own capabilities.

In this model conscious contents reside in the operating system or its working memory. Johnson-Laird believes his model can account for aspects of self-reflection, intentional decision-making, and action control.

Daniel Schacter has also proposed a compatible approach, to integrate evidence on neuropsychological disconnections from consciousness, particularly implicit memory and anosognosia, called the Dissociable Interactions and Conscious Experience (DICE) model. 'The basic idea motivating the DICE model … is that the processes that mediate conscious identification and recognition – that is, phenomenal awareness in different domains – should be sharply distinguished from modular systems that operate on linguistic, perceptual, and other kinds of information' (1990, pp. 160–1).

Like Johnson-Laird, Schacter's DICE model assumes independent memory modules and a lack of conscious access to details of skilled, procedural knowledge. It is primarily designed to account for memory dissociations in normal and damaged brains. Schacter makes two main observations. First, with the exception of coma and stupor, failures of awareness in brain damage are usually restricted to the domain of the impairment; patients do not have difficulty generally in gaining conscious access to other knowledge. For example, amnesic patients do not necessarily have trouble reading words, while alexic individuals do not necessarily have memory problems.

However, implicit (unconscious) memory for lost conscious functions has been demonstrated in many conditions. For example, name recognition is facilitated in prosopagnosia (face-blind) patients when the name is accompanied by a matching face – even though the patient does not consciously recognize the face. Numerous examples are known of implicit knowledge in patients who do not have deliberate, conscious access to the information. These findings suggest an architecture in which various sources of knowledge function somewhat separately, since they can be selectively lost. These separable knowledge sources are not accessible to consciousness, even though they continue to shape voluntary action.

In DICE Schacter gives additional support to the idea of a system of separable knowledge sources, specifically to explain spared explicit knowledge in patients with brain damage. DICE does not try to explain the limited capacity of consciousness or the problem of selecting among potential inputs. In

agreement with Shallice (below) the DICE model suggests that the primary role of consciousness is to mediate voluntary action under the control of an executive.

Tim Shallice's 1978 theory focused on conscious selection of a *dominant action system*, a set of current goals that work together to control thought and action. More recently Shallice (1988) modified and refined the theory to accommodate a broader range of conscious functions. Shallice's information-processing system also consists of a very large set of specialized processors, like Johnson-Laird's subsystems and Baars' (1988) specialized unconscious processors. A large set of action and thought schemata can 'run' on these modules. The schemata are well-learned, highly specific programs for routine activities, such as eating with a spoon, driving to work, etc. Competition and interference between currently activated schemata is resolved by *contention scheduling*, which selects among the schemata based on activation and lateral inhibition. Contention scheduling acts only during routine operations. A *supervisory system* modulates the operation of contention scheduling. It has access to representations of operations, of the individual's goals, and of the environment. It comes into play when operation of routinely selected schemata does not meet the system's goals, that is, when a novel or unpredicted situation is encountered or when an error has occurred. Finally, a *language system* can function either to activate schemata or to represent the operations of the supervisory system or specialist systems. More recently an *episodic memory* component with event-specific memory has been added to the set of control processes.

Shallice claims that consciousness cannot reside in any of these control systems taken individually. No single system is either necessary or sufficient to account for conscious events. Consciousness remains even when one of these control systems is damaged or disabled. And the individual control systems can all operate autonomously and unconsciously. Shallice suggests that consciousness may arise when there is concurrent and coherent operation of several control systems on representations of a single activity.

NEURONAL RESONANCE THEORIES

Several brain theories have now been developed. They have so much overlap that we will consider only a few in detail. All brain theories of conscious functions can be broadly characterized as 'neuronal resonance theories'. That is, they propose that in waking consciousness the thalamocortical core of the human brain is continuously cycling neuronal activity between the major thalamic relay nuclei and corresponding regions of cortex, supplemented by corticocortical and cortical-subcortical cycles of activity. When consciousness is at a very low level, as in deep sleep, some comas, deep general anaesthesia, and epileptic 'states of absence', the rapid and irregular electrical field activity characteristic of waking consciousness is replaced by slow, correlated, and high-amplitude waves. At the level of single neurons waking consciousness involves temporally uncorrelated firing, while deep sleep shows repetitive, highly correlated burst-pause firing in large cell populations. These basic facts and their underlying neurophysiology, neuroanatomy, and neurochemistry are so well established that they arouse little fundamental disagreement (e.g. Singer and Gray, 1995; John *et al*., 1997; Edelman, 1989; Tononi and Edelman, 1999; Damasio, 1989). (*See* **Neural Correlates of Visual Consciousness**; **Thalamocortical Interactions and Binding**)

The development of parallel distributed processor (PDP) models is consistent with such large-scale brain models. Grossberg (in press) and Taylor (1992) have applied PDP concepts to aspects of consciousness. Grossberg's Adaptive Resonance Theory (ART) has been applied to specific brain functions including perception, recognition, attention, reinforcement, recall, and memory search. Grossberg writes that 'it is suggested that all conscious states are resonant states', especially those that occur 'between bottom-up and top-down processes as they reach an attentive consensus between what is expected and what is there in the outside world'.

Taylor (1992) also believes that conscious contents are determined by the intermingling of past and present. The reticular nucleus of the thalamus (nRt) appears to control the major sensory highways to cortex, and Taylor has provided a model for intersensory competition based on this evidence. Rather than a simple winner-take-all competition, he suggests that consciousness corresponds to a wave of activity of the coupled thalamic-nRt-cortical system, a multidimensional 'bubble' of neuronal firing patterns. Such a wave will show many regions over cortex that have nonzero activity. Recently Taylor has advocated the possibility that parietal cortex may act as a global workspace in the sense advocated by Baars (above).

Antonio Damasio (1989) suggests a role for consciousness in recognition and recall. Like other authors, Damasio's theory involves looping

feedforward and feedback circuits of neurons, a massive resonant assembly of cells. To this mechanism he adds a specific role for sensory projection areas of the cortex (local convergence zones) and their neighboring higher-order association areas (nonlocal convergence zones). Temporal synchrony is proposed to account for retrieval of information from memory when neuron ensembles are activated in a time-locked fashion in the local and nonlocal convergence zones of cortex. Memories are stored as distributed sets of fragmentary features in large populations of neurons, and are retrieved by means of synchronous activation of related firing patterns in a subset of the same cell population.

CONCLUSION

Although consciousness has only recently returned as a central focus of the brain and cognitive sciences, current proposals capture a good deal of the evidence in a broad way. A critical mass of scientists is now collecting relevant evidence and developing theory. Conscious experience seems to create access to multiple, independent knowledge sources. While there are distinct pros and cons to each theoretical perspective, the general impression is of a surprising degree of convergence. A major goal for the future is to show how neuronal resonance could support the global cognitive functions that require consciousness.

Acknowledgements

This work was supported by the Neurosciences Research Foundation.

References

Baars BJ (1988) *A Cognitive Theory of Consciousness*. New York, NY: Cambridge University Press.

Baars BJ (1997) *In the Theater of Consciousness: The Workspace of the Mind*. New York, NY: Oxford University Press.

Baars BJ (1998) Metaphors of consciousness and attention in the brain. *Trends in Neurosciences* **21**(2): 58–62.

Baars BJ (2002) The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences* **6**(1): 47–52.

Baars BJ (in press) Working memory requires conscious processes, not vice versa: a global workspace account. In: Osaka N (ed.) *The Neural Basis of Consciousness*. Amsterdam: Benjamins.

Baars BJ, Banks WP and Newman J (in press) *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press/Bradford Books.

Chalmers D (1996) *The Conscious Mind*. New York, NY: Oxford University Press.

Crick F (1984) Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Science of the USA* **81**: 4586–4590.

Damasio AR (1989) Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* **33**: 25–62.

Dennett D and Kinsbourne M (1992) Time and the observer: the where and when of consciousness in the brain. *Brain and Behavioral Sciences* **15**(2): 183–247.

Dennett D (2001) Are we explaining consciousness yet? *Cognition* **79**: 221–237.

Edelman GM (1989) *The Remembered Present: A Biological Theory of Consciousness*. New York, NY: Basic Books.

Franklin S and Graesser A (in press) A software agent model of consciousness. In: Baars BJ, Banks WP and Newman J (eds) *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press/Bradford Books.

Grossberg S (in press) Brain learning, attention, and consciousness. In: Baars BJ, Banks WP and Newman J (eds) *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press/Bradford Books.

Hilgard ER (1977) *Divided Consciousness: Multiple Controls in Human Thought and Attention*. New York: Wiley.

James W (1890/1983) *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

John ER, Easton P and Isenhart R (1997) Consciousness and cognition may be mediated by multiple independent coherent ensembles. *Consciousness & Cognition* **6**: 3–39.

Johnson-Laird PN (1988) A computational analysis of consciousness. In: Marcel AJ and Bisiach E (eds) *Consciousness in Contemporary Science*, pp. 357–368. Oxford, UK: Clarendon Press.

Mandler GA (1984) *Mind and Body*. New York: Basic Books.

Mangan B (1993) Taking phenomenology seriously: the 'fringe' and its implications for cognitive research. *Consciousness & Cognition* **2**(2): 89–108.

Newman J and Baars BJ (1993) A neural attentional model for access to consciousness: a Global Workspace perspective. *Concepts in Neuroscience* **2**(3): 255–290.

Newman JB, Baars BJ and Cho S-B (1997) A neural Global Workspace model for conscious attention. *Neural Networks* (Special Issue) **10**(2): 1195–1206.

Schacter DL (1990) Toward a cognitive neuropsychology of awareness: implicit knowledge and anosognosia. *Journal of Clinical and Experimental Neuropsychology* **12**(1): 155–178.

Shallice T (1988) Information-processing models of consciousness: possibilities and problems. In: Marcel AJ and Bisiach E (eds) *Consciousness in Contemporary Science*, pp. 305–333. Oxford, UK: Clarendon Press.

Singer W and Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–586.

Taylor J (1992) Towards a neural network model of the mind. *Neural Network World* **2**: 797–812.
Tononi G and Edelman GM (1999) Consciousness and complexity. *Science* **282**: 1846–1851.

Further Reading

Crick F (1995) *The Astonishing Hypothesis*. New York, NY: Touchstone Books.
Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in Neuroscience* **2**: 263–275.
Edelman GM and Tononi G (2000) *A Universe of Consciousness*. New York, NY: Basic Books.
Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press.
Hilgard ER (1986) *Divided Consciousness: Multiple Controls in Human Thought and Action*. New York, NY: Wiley-Interscience.
Llinás R and Paré D (1991) Commentary: of dreaming and wakefulness. *Neuroscience* **44**(3): 512–535.
Milner AD and Rugg MD (eds) (1992) *The Neurophysiology of Consciousness*. London: Academic Press.

Consciousness, Function of

Intermediate article

Thomas W Polger, University of Cincinnati, Ohio, USA

CONTENTS

To inquire about the function of consciousness is to ask what consciousness does, what it enables us to do that we might not be capable of otherwise, or why some creatures came to be conscious.

INTRODUCTION

Consciousness is perhaps the most salient feature of human mental life. The experience of tasting a red wine differs from the experience of tasting coffee, and from that of reading the label on a wine bottle. Whatever else can be said about these differences, they are manifested in us by different conscious experiences. And this seems to be good for us: different experiences are evidently important in our abilities to discriminate among foods, to avoid injury, to identify potential mates, and so on.

But upon reflection it is less obvious what, if anything, consciousness does, what it allows us to do that we might not be capable of otherwise, or why some creatures – like human beings – have come to be conscious. For it seems that conscious experience is not necessary for the ability to distinguish objects in the world, or avoid injury, or seek a mate. Even if the experience of pain, for example, is important to the way that humans detect and avoid noxious elements in the external environment, it seems that we or other creatures could avoid hazards without the experience of pain, or any conscious experience at all. There is little doubt that mindless mechanical devices can be constructed to detect heat, classify wavelengths of reflected light, or distinguish chemical substances. We do not feel compelled to say that such devices feel pain, see colors, or taste wines. And we need not think only of mechanical devices and the automata of science fiction, for there is ample evidence that biological creatures can evolve fairly sophisticated sensory and motor capacities without having conscious experiences. The natural world is rife with creatures (microorganisms, molluscs, insects, and so on) that interact with their environments effectively, at least some of which may lack conscious experiences altogether.

QUESTIONS ABOUT THE FUNCTION OF CONSCIOUSNESS

The question of the function of consciousness does not have an obvious answer. We are faced not with a single question but with a handful of more or less related inquiries about what consciousness (as a matter of fact) does for human beings, about what if anything consciousness enables us to do that we could not (possibly) do otherwise, and about what capacities consciousness allows that would explain why it should be favored by natural selection: why it would evolve. The question of the function of consciousness is ambiguous, and the ambiguity owes as much to the idea of function as it does to any special considerations having to do with consciousness. Questions about the functions of mechanical artifacts and biological organs meet many of the same problems. (*See* **Consciousness, Philosophical Issues about**)

What Does Conscious Experience Do?

We might ask what abilities consciousness in fact mediates in human beings. In this case we are treating consciousness as a mechanism with certain effects, and we are inquiring about those effects in the same way that we might ask about the function of a carburetor or a heart. Carburetors regulate and mix air and fuel in some combustion engines. Hearts pump blood. Conscious experience allows us to discriminate and identify objects in the environment, to avoid hazards, and so forth. This was the answer that made it at first seem obvious what the function of consciousness is. Such explanations tell us what the 'causal role functions' of carburetors, hearts, and sensations are. A causal role function of a trait or mechanism is an effect of that trait

or mechanism that figures in an explanation of the overall capacities of the system of which it is a part. To explain the capacities of a system in terms of the causal role functions of its parts is to provide a 'functional analysis' of the system (Cummins, 1975). A special subset of causal role functions are those that can be characterized in terms of a computational device, such as a Turing machine.

Is Conscious Experience Necessary?

Even if we have a good explanation for what consciousness happens to do for us, we may still ask what it is that conscious experiences (likewise, carburetors or hearts) allow that could not be accomplished without them. This is not a question only about how human beings and cars are put together and what they are now capable of. It is also a question about how they might have been put together differently and what capacities they would have had under those circumstances. Could there be a car that does not have a carburetor? Certainly. Most automobiles these days use fuel injectors to mix air and fuel, rather than carburetors. Could there be a creature that does not have a heart? Mechanical devices are regularly used to circulate the blood of patients in the operating room. There is no reason to deny that some creature, however unlikely, could circulate its own blood without engaging a heart. So it may be with consciousness. The thesis that consciousness, though it may be crucial to our distinctively human way of interacting with the world, is not necessary for any of our capacities is called 'conscious inessentialism' (Flanagan, 1992).

There are interesting empirical phenomena that have seemed, at least to some, to support conscious inessentialism. Consider, for example, Bejamin Libet's (1985) experiments on the timing of conscious intentions to produce behavior. Libet's results purport to show that the muscular action potential that initiates movement occurs temporally prior to conscious awareness of an intention to move. These results have been interpreted as showing that consciousness does not play a role – or at least not the role traditionally envisioned – in the initiation of behavior. Daniel Dennett (e.g. 1991) has made much of the Libet experiments.

Blindsight is another phenomenon that has seemed, to some, to support some version of conscious inessentialism (e.g. Block, 1995). Lawrence Weiskrantz (1986) aroused the interest of many philosophers with his studies of patients with neural injuries who report no conscious visual experience in parts of their visual fields. Nevertheless, some of these patients perform much better than chance when they are forced to 'guess' about stimuli presented to the blind field. It seems that blindsight patients have residual information processing capacities despite lacking visual consciousness in the area of the scotoma, apparently supporting conscious inessentialism. This sort of phenomenon has led theorists to emphasize the importance of nonconscious visual processing (e.g. Milner and Goodale, 1995). But Weiskranz's results can also weigh against conscious inessentialism. After all, the tasks that blindsight patients perform better than chance – however remarkable that may be – are performed unerringly by normal subjects; and blindsight patients never initiate action based on the stimuli presented to the blind field (Van Gulick, 1985). These considerations suggest that consciousness does play an important role. (*See* **Blindsight**)

If conscious inessentialism is true, then there could be creatures that negotiate the world just as human beings do, but that nevertheless lack consciousness. Despite lacking conscious experience, such creatures (called 'zombies') make the same bodily movements as we do: they avoid fire, behave discriminately towards various wavelengths of light and chemical substances, etc. Sometimes it is claimed that conscious inessentialism entails that consciousness is epiphenomenal – that consciousness does not have any causal powers at all. (*See* **Zombies**; **Epiphenomenalism**)

But this is a mistake: from the fact that a carburetor is inessential to the operation of a car (because it could be replaced by a fuel injector) it does not follow that carburetors have no effects in those cars where they are found. Carburetors mix air and fuel in some cars; fuel injectors mix air and fuel in other cars. Neither a carburetor nor a fuel injector is necessary to the operation of automobiles in general; but carburetors and fuel injectors are not thereby epiphenomenal. (And just as there are reasons for generally preferring fuel injectors over carburetors, visual or painful experience may be better or worse ways of engaging with the world.)

Why Did Conscious Experience Arise?

In asking about the function of consciousness we might want to know not what conscious experiences enable us to do currently, but rather why we have come to be conscious at all. That is, we might be asking about the teleology of consciousness, about the purpose that it serves. If we understand teleology in terms of evolutionary history, then we are asking what the etiological function

of consciousness is. The etiological function of a trait is the effect that the trait had in the ancestors of a creature that provided an adaptive advantage to creatures of that kind, so that evolutionary pressures favored creatures with the trait. The etiological or 'selected effect' function of a trait explains why the trait came to be present or maintained in creatures of a kind, why it was naturally selected (Millikan, 1989; Neander, 1991). (*See* **Evolutionary Psychology: Theoretical Foundations**)

Consider again the possibility of conscious inessentialism. If conscious inessentialism is false – if consciousness is necessary for some human capacities (e.g. detecting wavelengths, avoiding injury) – then it is easy to answer the question of why we are conscious: we are conscious because it is adaptively advantageous (e.g. to detect wavelengths or avoid injury). That is to say, consciousness evolved; the evolutionary history of conscious experience is just the same as that of the capacities that conscious experiences mediate. That history need not be obvious or simple; it may not even be knowable by us: we do not assume that we will know the evolutionary history of every (or perhaps any) biological trait. Further, we should not assume that every trait has adaptive value. Some traits are the result of chance alone – though assuming that particularly complex traits are products of evolution by natural selection may be a reasonable methodological stance (Brandon, 1990; Grantham and Nichols, 1999). Of course the contingencies of organism and environment are such that discrimination of wavelengths and chemicals, avoidance of flames, and so on are not themselves compulsory. But insofar as we could explain why a creature should avoid injury, we would be able to explain why it experiences pain. The research program known as evolutionary psychology proceeds on the assumption that most or all psychological traits are adaptations by natural selection that are required for capacities that would have conferred an advantage on hominids living in the late Pleistocene era (Barkow *et al*., 1992).

On the other hand, if conscious inessentialism is true then the question of why consciousness has come to be is somewhat different. In that case we would need to explain not only why the capacity to avoid injury came to be enabled but also why, in some creatures, conscious experience mediates avoidance of injury. The answer might be that the presence of consciousness is a result of mere chance. That would perhaps be disappointing, but it would not undermine our belief that consciousness in fact has important effects in our lives. In particular, it would not force us to adopt epiphenomenalism. Just as an automobile might be built with a carburetor or a fuel injector, so creatures might evolve conscious experiences or some other mechanisms. Perhaps some forms of conscious experience have interesting evolutionary histories while others arose only by chance. We need not take consciousness to be a single phenomenon in order to meaningfully ask about its function.

CONSCIOUSNESS AND FUNCTIONAL KINDS

One might believe that whatever mixes air and fuel is a carburetor. That is to say, one might adopt a sort of metaphysical functionalism regarding carburetors. On this view, carburetors are functional kinds that are constituted by their capacity to mix air and fuel; thus fuel injectors are carburetors. Likewise one could think of hearts as blood pumps, and one could think of whatever mediates injury avoidance as pain experience. One must then regard the thesis of conscious inessentialism (likewise, carburetor inessentialism) as incoherent. According to a functionalist it would not even make sense to talk about something that did all the things (causal role functions) that pain does in human beings but that does not *ipso facto* have pain experiences. Two popular variations of functionalism about consciousness take conscious mental states to be a subset of representational states (e.g. Dretske, 1995) or to be meta-representations of first-order mental states – the higher-order thought theory championed by David Rosenthal (e.g. 1986). (*See* **Functionalism**)

The theory that consciousness is a functional kind is closely aligned with the view that all the facts about consciousness can be explained by reference to its functional role or roles (e.g. its causal role functions). One reason for holding such a view is general commitment to functionalist explanations, at least with respect to psychology; a widely-held theory is that all properties are causal role functional properties, and thus all explanations are functional explanations (Shoemaker, 1984). But in that case, if we cannot secure functional explanation of consciousness then we will be obliged to abandon the belief that consciousness is a physical phenomenon at all (Chalmers, 1996).

References

Barkow J, Cosmides L and Tooby J (eds) (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Block N (1995) On a confusion about the function of consciousness. *Behavioral and Brain Sciences* **18**: 227–247.

Brandon R (1990) *Adaptation and Environment*. Princeton, NJ: Princeton University Press.
Chalmers D (1996) *The Conscious Mind: In Search of a Fundamental Theory*. New York, NY: Oxford University Press.
Cummins R (1975) Functional analysis. *The Journal of Philosophy* **72**: 741–765.
Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.
Dretske F (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.
Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press.
Grantham T and Nichols S (1999) Evolutionary psychology: Ultimate explanations and Panglossian predictions. In: Hardcastle V (ed.) (1999) *Where Biology Meets Psychology: Philosophical Essays*, pp. 47–66. Cambridge, MA: MIT Press.
Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* **8**: 529–566.
Millikan R (1989) In defense of proper functions. *Philosophy of Science* **56**: 288–302. [Reprinted in: Allen C, Bekoff M and Lauder G (eds) (1998) *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge, MA: MIT Press.]
Milner B and Goodale MA (1995) *The Visual Brain in Action*. New York, NY: Oxford University Press.
Neander K (1991) Functions as selected effects: the Conceptual analyst's defense. *Philosophy of Science* **58**: 168–184. [Reprinted in: Allen C, Bekoff M, and Lauder G (eds) (1998) Nature's Purposes: Analyses of Function and Design in Biology. Cambridge, MA: MIT Press.]
Rosenthal D (1986) Two concepts of consciousness. *Philosophical Studies* **49**: 329–359.
Shoemaker S (1984) *Identity, Cause, and Mind*. New York, NY: Cambridge University Press.
Van Gulick R (1985) What difference does consciousness make? *Philosophical Topics* **17**: 211–230.
Weiskrantz L (1986) *Blindsight: A Case Study and Implications*. New York, NY: Oxford University Press.

Further Reading

Allen C, Bekoff M and Lauder G (eds) (1998) *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge, MA: MIT Press.
Amundson R and Lauder G (1994) Function without purpose: the uses of causal role function in evolutionary biology. *Biology and Philosophy* **9**: 443–469. [Reprinted in Allen *et al*. (1998).]
Cowey A and Stoerig P (1991) The neurobiology of blindsight. *Trends in Neuroscience* **29**: 65–80.
Cummins R (1983) *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press. [Reprinted in Allen *et al*. (1998).]
Davies M and Humphreys GW (1993) *Consciousness: Psychological and Philosophical Essays*. Oxford, UK: Blackwell.
Fetzer J (ed.) (2002) *Evolving Consciousness*. Amsterdam, Netherlands: John Benjamins.
Flanagan O (2000) *Dreaming Souls*. New York, NY: Oxford University Press.
Flanagan O and Polger T (1995) Zombies and the function of consciousness. *Journal of Consciousness Studies* **2**(4): 313–321.
Güzeldere G, Flanagan O and Hardcastle V (1999) The nature and function of consciousness: Lessons from blindsight. In: Gazzaniga M (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 1277–1284. Cambridge, MA: MIT Press.
Hardcastle V (ed.) (1999) *Where Biology Meets Psychology: Philosophical Essays*. Cambridge, MA: MIT Press.
Ito M, Miyashita Y and Rolls ET (eds) (1997) *Cognition, Computation, and Consciousness*. New York, NY: Oxford University Press.
Lycan W (1987) *Consciousness*. Cambridge, MA: MIT Press.
Lycan W (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.
Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.
Marcel A and Bisiach E (eds) (1988) *Consciousness in Contemporary Science*. New York, NY: Oxford University Press.
Polger T (2000) Zombies explained. In: Ross D, Brook A and Thompson D (eds) *Dennett's Philosophy: A Comprehensive Assessment*. Cambridge, MA: MIT Press.
Polger T and Flanagan O (1999) Natural answers to natural questions. In: Hardcastle (1999, pp. 221–247).
Polger T and Flanagan O (2002) Consciousness, adaptation, and epiphenomenalism. In: Fetzer (2002).
Sober E (1985) Panglossian functionalism and the philosophy of mind. *Synthese* **64**: 165–193.
Tye M (1996) The function of consciousness. *Noûs* **30**(3): 287–305.
Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Exploration*. New York, NY: Oxford University Press.

Consciousness, Machine

Introductory article

Keith Gunderson, University of Minnesota, Minneapolis, Minnesota, USA

CONTENTS

The widespread use of computers and robots within research programs in cognitive psychology has stimulated interest in the possibility of machine consciousness. As a result, many unsettled issues in the philosophy of mind and theory of knowledge have been reformulated in terms of, and in turn raised questions about, machine-oriented modeling.

INTRODUCTION

Consciousness is one of the most perplexing topics in the study of mind. There are many controversies surrounding its exact nature and relationship with the physical. For centuries, philosophers and others have argued about whether a machine could be conscious, partly so as to reach a better understanding of human and animal minds.

During the second half of the twentieth century, with the development of the digital computer, the topic of machine consciousness became intertwined with questions about artificial intelligence (AI). Many tasks that once seemed to require human conscious intelligence are now performed by computers. Machines with various kinds of programs and related robotic capabilities have been used to help explain some of the more baffling aspects of human mentality. (*See* **Artificial Intelligence, Philosophy of**)

Such approaches face a variety of problems. What counts as an example of consciousness? Some argue that consciousness is primarily a private (subjective) phenomenon. Nevertheless, attempts to construct machine models that would objectify it continue. Even when such models seem flawed or limited, an understanding of their shortcomings can provide fresh perspectives and stimulate lively debate on thinking, perception, awareness, purposive behavior, and the relationship between the mind and the body. (*See* **Mind–Body Problem**)

HISTORY

Could a machine be conscious? This question was being asked at least as long ago as the seventeenth century. Machines then (as now) were generally assumed to consist wholly of matter. To ask whether we might someday be able to build a conscious machine was one way of asking whether consciousness (animal or human) was made out of matter. The view that it was was called materialism. The view that it was not, and that mind was essentially different from matter, was called dualism. (*See* **Dualism**)

Descartes was dualism's most influential proponent. He regarded the human mind (or soul) as connected to a physical body. This connection, in the case of human beings, made the use of language, and many other activities, possible. A great many of these activities, Descartes felt, could not be explained in a purely mechanical way as matter in motion. On the other hand, he believed that the activities of all nonhuman animals could be explained in a purely mechanical manner. So dualism was not true of animals. Their behaviors, upon close inspection, could be shown to be made up of reflexes, and were instinctual, not guided by reason. Animals were organic machines. They had no thoughts or sensations. This view was called 'animal automatism'. (*See* **Descartes, René**)

Most philosophers found the doctrine of animal automatism too extreme. But some of the arguments in support of it were challenging. Perhaps living organisms were more machine-like than had been previously thought. In the eighteenth century, the simile of the mind being like a machine became popular. The French philosopher La Mettrie tried to turn the tables on Descartes. In his book *The Man Machine*, he first applauded Descartes for being a pure materialist about animals, but then tried to show that Descartes' arguments for animal

automatism could be used to show that human beings were just machines too.

Leibniz held the unusual view that both humans and animals were machines, but immaterial ones. Our bodies were said to be 'divine machines' or 'natural automata'. Human automata possessed the powers of perception and self-conscious memory. Animals had only a limited version of these abilities. Leibniz' doctrine, though odd and ambiguous, illustrates how flexible and abstract the idea of a machine can be. Almost anything, it seemed, might be a machine if it could be broken down into parts that functioned together in organized and predictable ways. (Somewhat surprisingly, Leibniz also claimed that we could never explain our perception in mechanical terms.) Leibniz agreed with the Aristotelean view that living created organisms, unlike windmills or clocks, could initiate purposive movement from within themselves, by virtue of a 'vital force' or 'entelechy'. These movements or activities were sometimes teleological, i.e., directed at ends or goals.

Long after Leibniz, philosophers, psychologists, biologists, and physiologists debated whether such behavior could be explained in purely materialistic mechanistic ways. Those who claimed that it could were called 'mechanists'. Those that claimed that it could not were often called 'vitalists'.

With the development of genetic theory, as well as discoveries about instinct and behavioristic psychology, proponents of vitalism nearly vanished from the intellectual scene. By the 1940s and 1950s, the prevailing view was wholly naturalistic.

The goal of the new field of 'cybernetics' (from the Greek word for 'steersman') was to understand in detail how control and communication worked in human and animal organisms as well as in self-regulating machines. The field was explicitly described as both behavioristic and functional. This meant that the focus of study was on inputs or stimuli from the environment to the animal or machine, together with states in between, leading to behavior. (*See* **Functionalism**; **Behaviorism, Philosophical**)

The centuries-old dream of being able to build a conscious machine seemed closer than ever. And it was at this time that the digital computer began to attract attention. Calculating machines had existed in both design and physical reality since the seventeenth century; but machines that could repeatably make use of their own computations in the production of further ones did not appear until the middle of the twentieth. This 'recursive' power had for some time been of intense logical and mathematical interest. It hinted at the possibility of self-reflection, and represented a new way of thinking about how a machine might be designed to imitate human thought processes.

PHILOSOPHICAL ISSUES ABOUT MACHINE CONSCIOUSNESS

Varieties of Consciousness

There is no tidy or uncontroversial concept of consciousness. But when trying to imagine a conscious machine, at least three (sometimes overlapping) aspects of consciousness should be considered. Firstly, there is the awareness involved in responding with purpose to stimuli from an external environment. Secondly (and often subsumed in the first aspect), there are simple perceptual experiences, such as seeing and touching, and inner thoughts, beliefs, intentions, feelings, emotions, moods, desires, and so on. Some of these mental states are 'about' other things (for example, a thought or belief about a cat). This 'aboutness' is sometimes called 'intentionality'. Other mental states are not about anything: they just are (for example, a visual experience of redness, or a bitter taste, or a sharp pain). Thirdly, there is self-consciousness, or the capacity of minds to reflect on their own conscious states, which might be any of the aforementioned ones. This 'second-order' consciousness is sometimes called 'introspective awareness'. (*See* **Consciousness, Philosophical Issues about**; **Introspection**; **Self-consciousness**)

Many of the aforementioned conscious states can be described as having a character or quality or 'feel' to them, such as 'what it's like' to see something red. These qualities are called 'qualia'. (*See* **Qualia**)

Many philosophers believe that, to some degree, some of these aspects of consciousness already exist in various self-regulating cybernetic mechanisms including computers. But it was the digital computer, with its many different forms of programming, that for many seemed to be the most promising candidate for machine mentality.

The Turing Test

What sorts of beings might be capable of thinking? It is often assumed that only beings with some degree of consciousness can think. It is also often assumed that the best evidence for conscious thinking existing in beings other than ourselves lies in various complex behaviors. The computer has shown impressive potential for performing complex tasks of which only thinking intelligent agents

such as ourselves had previously seemed capable. Thus the question of whether machines can think has attracted much interest since the mid-twentieth century, and with it the question of whether machines could be conscious.

In 1950, the mathematician A. M. Turing, in an interesting 'thought experiment', proposed a test for whether a machine can think. He imagined a game being played (called the 'imitation game'), in which an interrogator asks two concealed participants questions and uses their answers as a means for discovering a hidden aspect of their identities (for example, which one is the man and which the woman). The aim of one participant (A) is to fool the interrogator (C). The aim of the other (B) is to help the interrogator guess the right answer. Turing asked whether a computer might do as well as a human participant A in fooling C. He argued that it could, in the foreseeable future, and by virtue of this ability he felt that such a machine should be credited with thinking. This test later became known as the Turing test. (*See* **Turing Test**; **Turing, Alan**)

Some philosophers, as well as scientists involved in devising programs for modeling psychological processes, gave at least qualified support to Turing's position. Others disagreed. One argument was that a computer might perform a task previously only accomplished by intelligent people, without necessarily being intelligent. Perhaps the computer's performance would simply illustrate that there can be a variety of processes – some conscious and intelligent, others not – that result in the same achievement. Electric eyes, like doormen, may open doors for people, but they don't rely on seeing anyone coming, nor are they polite or concerned in their doing it.

In spite of these objections, Turing's test remained popular. Debates concerning the test's merits and demerits are still continuing.

Early Experiments in AI and some Philosophical Reactions to Them

Some of the most influential work in AI during the 1960s and 1970s concerned computer simulation of cognitive processes (CS). Allen Newell and Herbert Simon and others adopted the following strategy. A simulation of intelligent human mental processes, such as proving a theorem or playing chess, was constructed on the basis of observations of a human problem-solver's behavior. These observations might include verbal reports, jottings on paper, and so on. Attempts would then be made to represent these in the programming vocabulary used for the simulation. When the program was run, a trace of its 'thinking' activity would be printed. This trace would then be compared to some of the general features assumed to be present in the mind or behavior of the human problem-solver. (*See* **Newell, Allen**; **Simon, Herbert A.**)

One would not know with precision beforehand what would result from running the computer program. Nor can one know beforehand exactly what will go on in a human problem-solver's mind. But certain parallels between program and person were pursued. First, the tasks performed by a person were compared to the computer's results (deciding whether it passed the Turing test). Other levels of comparison involved much more guesswork. Although the structure and processes of the computer simulation could be described, it was often far from obvious how much of that description could apply to the human subject. Furthermore, it seemed that whatever objections might be raised to the Turing test in its general form would also be relevant to the specific use of that test in CS contexts.

Nevertheless, there was one important principle that seemed to underly AI research in general. This was what Newell and Simon called the 'physical symbol system hypothesis'. It stated simply that a physical symbol system has all the necessary means for general intelligent action. Whether this hypothesis was plausible or not, it had a bearing not only on CS but on many other related AI projects.

SHRDLU was an ingenious AI project which also incorporated robotic features. SHRDLU was a simulation of a robot that could respond to human commands concerning the construction and manipulation of (imagined) blocks of varying sizes and shapes (cubes, pyramids) and colors in a limited environment, say on a table top. When asked to do something with a block with respect to some present arrangement, not only could SHRDLU comply with the explicitly requested move, but it could calculate and carry out on its own whatever moves had to be made in order to carry out the command. (*See* **SHRDLU**)

SHRDLU could also produce replies to questions about what it was doing. And if a new word was contained in a command to SHRDLU, the program might indicate its unfamiliarity, and acquire the new word.

The environment within which SHRDLU functioned consisted entirely of data structures. These structures were symbolically represented within the computer. This program contained much of psychological and philosophical interest. Here was a buildable mechanical being operating in

three-dimensional space. (Another example of such a system, SHAKEY, was a mobile robot which negotiated its way through 'rooms' towards a goal.) In some sense, SHRDLU seemed to be sensitive to a variety of objects and able to interact with them. Its linguistic competence included an expandable vocabulary. But to what extent did SHRDLU, in its limited little world, reflect and illuminate our real world of perception, problem-solving, and discourse? Hubert Dreyfus has discussed the kinds of question–answer and command 'conversations' that SHRDLU had concerning its own blocky universe. He points out that SHRDLU's test for something being 'owned' is simply whether it is tagged as 'owned'. Dreyfus reminds us that SHRDLU couldn't own anything since it doesn't belong to a community in which ownership makes sense. This is just one example among many of how institutional and cultural facts could restrict how a subject like SHRDLU might act or what it could be 'conscious' of. Some form of embodiment seems to be required. (*See* **Embodiment**)

But questions remain about what SHRDLU can tell us about intelligent understanding, and perhaps about consciousness. At the heart of SHRDLU, as well as of other CS projects, is a program. But how plausible is Newell and Simon's physical symbol hypothesis?

Searle's Chinese Room Argument, and Other Problems for AI

Various AI projects durings the 1960s and 1970s seemed to exhibit some degree of understanding or comprehension (of, for example, language). In 1980, the philosopher John Searle published his famous 'Chinese room' argument, which tried to refute this. Searle claims he can imagine himself functioning just like an AI program. His mimicry consists of using a set of rules (written in his native language, English) whereby he manipulates a set of symbols in a language he doesn't understand at all (Chinese). The rules lead him to produce meaningful sentences in Chinese. Questions in Chinese are passed to him through a hole in the wall and he uses the rules to generate answers, which he passes back. To those outside the room where he is confined it will look as if something in the room is understanding questions in Chinese (about stories in Chinese) and answering them in Chinese. But no such understanding is taking place in the Chinese room. Nor, therefore, is it taking place in any AI program. (*See* **Chinese Room Argument, The**)

Symbol-based AI projects faced other serious problems. Conscious abilities such as recognizing patterns (like a face, or a house) or grouping things into kinds (like cats, dogs, or trees), or quickly sizing up situations, or performing complicated tasks, are often quickly done by people, and may involve many processes going on at once. Symbolic AI systems generally compute things step by step, in a serial way. Perhaps some model of consciousness based on the brain, which seems to carry out many different processes at the same time, would provide a better point of departure. This is called parallel processing. A number of 'connectionist' models have tried to capture this kind of processing. There has been considerable debate in recent years between researchers who favor symbolic AI models and those who favor connectionist ones. Serious philosophical questions arise for either approach to machine consciousness. (*See* **Connectionism**; **Computation, Philosophical Issues about**)

Qualia, Functionalism, Eliminative Materialism, and the Mind–Body Problem

Leibniz imagined a vast machine capable of perceptions, which we could go inside and walk around. No matter how hard we looked, he argued, we would never find anything in its mechanical make-up that would explain or disclose to us its perceptions. Leibniz saw this inevitable lack of disclosure as an important fact about the mind. The very same point could be made about consciousness and the brain. We could never 'see' consciousness displayed in a brain, or in any physical thing.

The philosopher Thomas Nagel has made a similar point by asking: What is it like to be a bat? Bats are very different from us. They are presumably conscious. But we don't know what it is like to be one. Nor would we know the answer to this question even if we knew everything that we possibly could about a bat's physical make-up.

Similarly, Frank Jackson has argued that a person brought up in a black-and-white environment, and thus deprived of any color experiences, might know everything there is to know about the material and functional facts underlying color vision. Nevertheless, when entering a multicolored environment for the first time, that person would be treated to an altogether new fact. The new fact would be what the experience of seeing red was like.

This character, or quality, or 'qualia' of our experiences, or 'what it's like' to have them, seems quite different from anything physical having to do with our brains or behavior. There seems to be a

gap between brains and qualia, between machines and minds. The difficulty of trying to bridge this gap so that we can fit consciousness into a scientific picture of the world is sometimes called 'the hard problem'. In essence it is the mind–body problem itself. (*See* **Explanatory Gap**)

Some say we will never be able to bridge this gap. Others have suggested that it may be possible by some as-yet-unknown means. Or perhaps qualia have a special kind of physical nature, irreducible to other physical things such as brain processes. And some philosophers are not troubled by qualia or subjectivity at all.

All these questions arise in connection with the influential doctrine in philosophy and cognitive psychology known as 'functionalism'. Functionalism claims that what makes a mental state the kind of mental state it is is its function or role in a pattern of causal relations. A pain, for example, is viewed as something caused by an input from the environment, say a brick falling on one's toe. The pain in turn causes other mental states, such as wanting to ease the pain, which then cause certain behavior (the output), such as taking an aspirin, grimacing, moaning, and so on. Anything with this role will be a pain. In conscious human organisms, certain neural states of the brain are thought by many to be what pains are. In other kinds of creatures, physical events other than neural processes may have that same role. For example, some future robot may have a 'brain' made of silicon chips that can produce pain. It is not any specific kind of matter that defines pain: multiple realizations are possible. It is the causal role that determines which mental states are present or absent in the creature under consideration. (*See* **Multiple Realizability**)

If so, then qualia don't seem to matter much. Imagine two different people who have two different color experiences when they look at a fire engine. One sees red; the other sees green. But they talk in exactly the same way about it. They both call it 'red'. They have each learned to use the words 'red' and 'green' in the same way. Their behaviors with respect to fire engines have the same causal relations. (*See* **Inverted Spectrum**)

One would think that the experience of seeing red was a different mental state from the experience of seeing green. But for the functionalist it seems that this is not so. Similarly, if there were a robot that behaved in exactly the same way as a human in the presence of fire engines, it would be, according to the functionalist, in the same mental state as the person experiencing green and the person experiencing red – even if it had no color 'experiences' at all. (*See* **Zombies**)

How do qualia fit into the functionalist's theory of mind? A 'quick fix' to these problems is provided by the doctrine of 'eliminative materialism'. The eliminative materialist suggests that many or most of our ordinary terms for talking about minds are misleading. Terms such as 'thought', 'belief', 'desire', and 'mental image' have been passed on to us as part of our prescientific 'folk psychology'. Perhaps these terms will become obsolete, in the same way that the term 'vital force' became obsolete and was essentially eliminated from the biological vocabulary used to explain goal-directed behaviors. In its most extreme, and least popular, version, eliminative materialism may suggest that the question of whether there is or could be machine consciousness is settled by default, since there is no such thing as consciousness. (*See* **Folk Psychology**)

INFLUENCE OF COGNITIVE SCIENCE ON ISSUES ABOUT MACHINE CONSCIOUSNESS

Since the time of Descartes and La Mettrie, machines have 'evolved'. And in the wake of this robotic 'evolution' have come various bold claims, many made by cognitive scientists, about the potential of physical mechanisms to reflect or even share human mentality and that of other animals. Some philosophers have supported these claims; others have criticized them. The lively dialogue that has ensued has been good for both cognitive science and philosophy. We now have new contexts for asking detailed questions about both mental functions and qualia (the 'hard problem').

Functional problems in modeling constitute what David Chalmers has called the 'easy' problems of consciousness. They include aspects of self-reporting, a system's access to its own internal states, reactions to an environment, categorization, and so on. All these, he and others believe, can be directly approached by the computational and connectionist strategies of cognitive science. Some, however, would claim that the so-called 'easy' problems are not so easy. Certain functional, computational activities that seem at first to be free of qualia may not be so, at least when carried out by human beings. For example, on hearing an ambiguous remark, we may have an awareness, or sense, or 'feel' for what the right meaning for a word in a sentence is. We hear the word 'bank' and we know (or 'compute') that the speaker meant a river bank, not a financial organization. Consciousness (with qualia) may be present in virtually every intentional human activity. So the basic nature – the

metaphysics – of qualia no doubt awaits further clarification.

Even if the 'easy' problems really are easy, we are left with the 'hard' questions – unless we adopt the extreme position of eliminative materialism. Cognitive scientists and philosophers alike want to know how 'outside' things (such as mechanisms, or observed machine behaviors) could ever depict subjective, 'inside' ('what it's like'), nonfunctional, mental features. If we imagine trying to build a machine with perspectives and perceptions step by step from scratch, at what point, if any, might we detect conscious experience entering the picture? Any interesting or useful machine model of consciousness should at least aim to incorporate whatever those features are that make the human mind, or any mind, so puzzling in the first place. Thus the model must be complex enough in what it contains that we are assured that something like the mind–body problem could arise for it. And such a model also needs to be instructive (or transparent) enough to give us some way of seeing through that problem. For suppose the machine model were complex enough to give rise to an analog of the mind–body problem, but it did not make consciousness more transparent than it now is. Then all we would learn from the construction (or imagined construction) of a conscious machine or robot would be that we could create things we didn't understand. We would not know if the machine model was itself a purely physicalistic model, or a dualistic one.

Ongoing attempts by cognitive scientists to model machine consciousness will, hopefully, contribute both to recreating the mind's complexity and to understanding its attendant perplexities. Of course, we cannot say what or 'who' the enhanced or combined 'offspring' of today's programs, robots, and neural networks are likely to be. But they will surely 'embody', as it were, various new research goals and presuppositions in cognitive science concerning mental function, qualia, and consciousness, which philosphers will want to assess.

Further Reading

Anderson A (ed.) (1964) *Minds and Machines*. Englewood Cliffs, NJ: Prentice Hall.

Block N, Flanagan O and Güzeldere G (eds) (1997) *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Chalmers D (1996) *The Conscious Mind*. New York, NY: Oxford University Press.

Chomsky N (1966) *Cartesian Linguistics*. New York, NY: Harper & Row.

Dennett D (1991) *Consciousness Explained*. Boston, Toronto and London: Little, Brown.

Descartes R (1960) *Discourse on Method and Meditations*, translated by L. Lafleur. New York, NY: Bobbs-Merrill.

Leibniz G (1898) The monadology. In: *The Monadology and Other Philosophical Writings*, translated by R. Latta, pp. 215–277. London, UK: Oxford University Press.

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Shear J (1998) *Explaining Consciousness: The Hard Problem*. Cambridge, MA: MIT Press.

Shieber S (forthcoming) *The Turing Test*. Cambridge, MA: MIT Press.

Consciousness, Philosophical Issues about

Advanced article

Ned Block, New York University, New York, USA

CONTENTS

There are many problems of consciousness, but the most significant of them is: how is it possible to explain consciousness in terms of its neural basis?

THE 'HARD PROBLEM'

There are a number of different matters that come under the heading of 'consciousness'. One of them is phenomenality, the feeling of, say, a sensation of red, or a pain, that is, what it is like to have such a sensation or other experience. Another is reflection on phenomenality. Imagine two infants, both of which have pain, but only one of which has a thought about that pain. Both would have phenomenal states, but only one would have a state of reflexive consciousness. This article will first discuss phenomenality, reflexivity, and one other kind of consciousness, global availability.

The 'hard problem' of consciousness is how to explain a state of consciousness in terms of its neurological basis. If a neural state *N* is the neural basis of the sensation of red, why is *N* the basis of that experience rather than of some other experience, or of none at all? Chalmers (1996) distinguishes between the hard problem and 'easy' problems that concern the function of consciousness. The hard problem (though not by that name) was identified by Nagel (1974) and further analyzed in Levine (1983).

There are two reasons for thinking that the hard problem has no solution. The first is '*actual failure*': the fact that no one has been able to think of even a speculative answer. The second is '*principled failure*': the materials we have available seem ill suited to providing an answer – as Nagel says, an answer to this question would seem to require an objective account that necessarily leaves out the subjectivity of what it is trying to explain; and we do not even know what would count as such an explanation.

PERSPECTIVES ON THE HARD PROBLEM

Of the many perspectives on the hard problem, we will discuss four, all of which comport with a naturalistic framework.

Eliminativism

Eliminativism is the view that consciousness, as understood above, simply does not exist (Dennett, 1979; Rey, 1997). So there is nothing for the hard problem to be about.

Deflationism

Deflationists, also called philosophical reductionists (e.g. Dennett, 1991), move closer to 'common sense' by allowing that consciousness exists, but they 'deflate' this commitment – again on philosophical grounds – taking it to amount to less than 'meets the eye'. One prominent form of deflationism makes a conceptual reductionist claim: that consciousness can be conceptually analyzed in nonphenomenal terms. The main varieties of analysis are behavioral, functional, representational and cognitive.

Pitcher (1971) and Armstrong (1968) can be interpreted as analyzing consciousness in terms of beliefs. One type of prototypical conscious experience, for example of blue, is a matter of an inclination (perhaps suppressed) to believe that there is a blue object in plain view. (See Jackson (1977) for a convincing refutation.) A different kind of analysis appeals to higher-order thought or higher-order perception. Such analyses take the concept of a conscious pain to be the concept of a pain that is accompanied by another state that is about that pain. A pain that is not so accompanied is not a

conscious state (Armstrong, 1968; Carruthers, 1992; Lycan, 1990). (Rosenthal (1997) advocates a higher-order thought view as an empirical identity rather than as a conceptual analysis.) Another deflationist view, compatible with analyses in terms of beliefs, concerns not the states themselves but their contents. Representationism holds that it can be established philosophically that the phenomenal character of experience is its representational content. Many representationists reject conceptual analysis, but still their accounts do not depend on any details of psychology or neuroscience (Harman, 1990, 1996; Dretske, 1995; Lycan, 1996; McDowell, 1994, Tye, 1995). (Shoemaker (1994) mixes phenomenal realism with representationism in an interesting way.) Conceptual functionalists say that the concept of consciousness is analyzable functionally (Lewis, 1994).

According to deflationism, there is such a thing as consciousness, but there is no hard problem, that is, there are no mysteries concerning the physical basis of consciousness that differ in kind from scientific problems about, for example, the physical or functional basis of liquidity, inheritance, or computation.

Inflationism

Inflationism, also called phenomenal realism, is the view that consciousness is a substantial property that cannot be conceptually reduced or otherwise reduced on armchair grounds in nonphenomenal terms. Logical behaviorists think that we can analyze the concept of pain in terms of certain kinds of behavior, but inflationists reject all such analyses of phenomenal concepts in nonphenomenal terms. According to most contemporary inflationists, consciousness plays a causal role and its nature may be found empirically as the sciences of consciousness advance. Inflationism is compatible with the empirical scientific reduction of consciousness to neurological or computational properties of the brain – just as heat was scientifically, but not philosophically, reduced to molecular kinetic energy. (It is not a conceptual truth that heat is molecular kinetic energy.) Inflationism accepts the hard problem but aims for an empirical solution to it (Block, 1995; Flanagan, 1992; Loar, 1997; Nagel, 1974). McGinn (1991) argues that an empirical reduction is possible but that we can't find or understand it. Searle (1992) endorses a roughly naturalistic point of view and rejects armchair reduction of phenomenality, but he also rejects any empirical reduction of phenomenal properties.

The inflationist regards all the deflationist accounts described above as leaving out the phenomenon. Phenomenality has a function and represents the world, but it is something over and above that function or representation. Something might function like a phenomenal state but be an ersatz phenomenal state with no real phenomenal character. The phenomenal character that represents red in you might represent green in me. Phenomenal character does represent but it also goes beyond what it represents. Pain may represent damage but that is not what makes pain painful.

Dualistic Naturalism

Dualistic naturalism is a broad category which includes Chalmers' (1996) view that standard materialism is false but that there are naturalistic alternatives to Cartesian dualism, such as panpsychism. Nagel (2000) proposes that there is a deeper level of reality that is the naturalistic basis both of consciousness and of neuroscience.

AN ATTEMPT AT A DISSOLUTION OF THE HARD PROBLEM

Suppose following a theory of visual experience proposed by Crick and Koch (1990) that corticothalamic oscillation (of a certain frequency) is the neural basis of an experience with phenomenal quality *Q*. There is a plausible argument that seems to solve, or rather dissolve, the hard problem from a physicalist point of view. According to this argument, the hard problem is illusory. One might as well ask why H_2O is the chemical basis of water rather than gasoline or nothing at all. Just as water just is its chemical basis, so *Q* just is its neural basis, and so the original question is vacuous.

This point is correct as far as it goes, but it does not go far enough. It begs another question: how could one property be both phenomenal property *Q* and corticothalamic oscillation? How is it possible for something subjective to be something objective, or for something first-personal to be something third-personal? The problem is not that we cannot find an explanation for this identity; indeed, in a sense there are no explanations for any identities (Block, 1978a; Block and Stalnaker, 1999). The problem is that the claim that a phenomenal property is a neural property seems just as mysterious as – maybe even more mysterious than – the claim that a phenomenal property has a certain neural basis. Explaining an identity is different from explaining how an identity can be true,

where the latter involves removing a sense of puzzlement.

We can even see the same two obstacles stated above reappearing: 'actual failure' and 'principled failure'. In the first place, no one has even a speculative answer to the question of how something objective can be something subjective. In the second place, actual failure does not seem accidental. The objective seems to necessarily exclude the subjective. The third-personal seems to necessarily exclude the first-personal. Further, as McGinn (1991) notes, neural phenomena are spatial, but the phenomenal is prima facie nonspatial.

Thus, the reasons mentioned above for thinking that the explanatory gap resists closing seem to surface in a slightly different form. However, as we shall see, in this form they are more tractable.

AN APPROACH TO THE HARD PROBLEM

We now discuss a possible approach to the hard problem, whose main element is a distinction which is widely appealed to in discussions of Jackson's (1982) famous 'Mary' example. Jackson imagined a neuroscientist of the distant future who is raised in a black and white room but who knows everything scientific there is to know about color and the experience of it. When she steps outside the room for the first time, she learns what it is like to see red. Jackson argued that since the scientific facts do not encompass the new fact that Mary learns, dualism is true.

The most convincing response to Jackson's, argument (which derives from Loar (1990/1997)) involves an appeal to the distinction between a property and a concept of that property (Churchland, 1989; Loar, 1990/1997; Lycan, 1990; Van Gulick, 1993; Sturgeon, 1994; Tye, 1999; Perry, 2001).

A concept is a thought element in a way of thinking, a kind of representation. For our purposes, concepts can be interpreted as symbols in a 'language of thought'. This usage is different from another common philosophical usage in which a concept is something like a meaning. Concepts in our sense are individuated in part by meanings: x and y are instances of the same concept if and only if they are instances of the same representation and have the same meaning. 'Water' and 'H_2O' are instances of different representations, so the concept of water is distinct from the concept of H_2O.

Someone could believe that 'this color' is useful for painting pots but that red is not, even if 'this color' is red. Our experiential concept of red differs from our linguistic concept of red. An experiential concept involves a phenomenal element, a phenomenal way of thinking, for example a mental image that is in some sense of red, or an ability to produce such a mental image, or at least an ability to recognize red – which arguably could not be done without some phenomenal mental element.

Importantly, we can have an experiential concept of an experience (which we can call a phenomenal concept, phenomenal concepts being a subclass of experiential concepts), as well as an experiential concept of a color. And the very same mental image may be involved in both concepts. The difference between the phenomenal concept of the experience and the experiential concept of the color lies in the rest of the concept – in particular, in the way the phenomenal element functions in the concept. This can be a matter of further concepts explicitly invoked in the concept – the concept of a color in one case and the concept of an experience in the other. One type of experiential concept (of a color or of an experience) involves a demonstrative together with a mental image and a language-like representation, e.g. 'that [attention to a mental image] color' or 'that [attention to a mental image] experience' where the brackets indicate the use of attention to a nondescriptive element, a mental image, in fixing the demonstrative reference. Loar (1990/1997) gives an example involving two concepts in which something like a mental image of a cramp feeling is used to pick out a muscular knot in one concept, and in the other concept, the cramp experience itself. In our notation, the two concepts would be 'that [attention to a mental image] cramp' and 'that [attention to a mental image] cramp experience'.

An experiential concept uses a phenomenal property to pick out a related property. For example, an experiential concept of a color can use an aspect of the experience of that color to pick out the color. A phenomenal concept uses a phenomenal property to pick out a phenomenal property. The phenomenal property used in the concept need not be the same as the one picked out. For example, one could have a phenomenal concept of the experience of the missing shade of blue whose phenomenal elements are the flanking color experiences. Or, a phenomenal element involved in one's perception of a color could be used to pick out the experience of the complementary color. Importantly, the phenomenal element in a phenomenal concept need not be, and in general cannot be, conceptualized, at least if the conceptualization is meant to be itself phenomenal. For if a phenomenal concept had to make use of a

phenomenal element via a distinct phenomenal concept of that element, there could be no phenomenal concepts. Thus we can define a phenomenal concept as a concept of a phenomenal property that uses a phenomenal property to pick out a phenomenal property, but not necessarily under a concept of the phenomenal property used.

In these terms, then, Mary acquired a new concept of a property she was already acquainted with via a different concept. In the room, Mary knew about the subjective experience of red via physical concepts. After she left the room, she acquired a phenomenal concept of the same property. So Mary did not learn a new fact: she learned a new concept of an old fact. She already had a third-person understanding of the fact of what it is like to see red. What she gained was a first-person understanding of the very same fact. She knew already that corticothalamic oscillation of a certain frequency is what it is like to see red. What she learned is that 'this [attention to a mental image] experience is what it is like to see red'. So the case does not demonstrate that there are facts that go beyond physical facts.

Recall that there is a principled reason why mind–body identity seemed impossible: that a first-person subjective property could not be identical to a third-person objective property. But the distinction between concepts and properties allows us to see that the above distinction between subjective and objective, and the distinction between first person and third person, are distinctions between kinds of concepts, not kinds of properties. There is no reason why a subjective concept and an objective concept cannot pick out the same property. Thus we can substitute a dualism of concepts for a dualism of properties.

There is another way in which the concept–property distinction helps with the hard problem. We can blame the explanatory gap and the hard problem on our inadequate concepts rather than on dualism. To take a variant on Nagel's (1974) example, we are like pre-Socratics who have no way of understanding how it is possible that heat equals mean molecular kinetic energy, lacking the concepts required to frame both sides of the equation. (Heat was not clearly distinguished from temperature until the seventeenth century.) What is needed is a concept of heat and a concept of kinetic energy that make it conceivable that there is a causal chain of the referential sort leading from the one magnitude to each concept. Or rather, since the phenomenal concept includes a sample of the relevant phenomenal property (on the Humean simplification we are using), there is no mystery about the mental side of the equation. The mystery is how the physical concept picks out that phenomenal property. This is the remaining part of the explanatory gap, which will be closed, if at all, by science. Is there a principled reason to think it cannot be? The hard problem itself does not contain such a reason. Perhaps our conceptual inadequacy is temporary, as Nagel sometimes appears to suppose, or perhaps it is permanent, as McGinn (1991) supposes.

EMPIRICAL FINDINGS ABOUT CONSCIOUSNESS

We will now turn to a discussion of recent empirical findings on consciousness. The most interesting line of experimental investigation of consciousness in recent years uses phenomena in which perception changes independently of the stimulus. One such paradigm uses binocular rivalry. If two stimuli – for example, horizontal and vertical stripes – are presented, a different stimulus to each eye, one does not see a blend, but rather, first horizontal stripes that fill the whole visual field, and then vertical stripes that fill the whole field. Logothetis (1998) trained monkeys to pull different levers for different patterns. They then presented different patterns to the monkeys' two eyes, and observed that, as with people, the monkeys switched back and forth between the two levers even though the sensory input remained the same. Logothetis recorded the firings of various neurons in the monkeys' visual systems. In the lower visual areas (e.g. V1), 80 percent of the neurons did not shift with the percept. But further along the occipital–temporal pathway, 90 percent shifted with the percept. So it seems that later areas in the occipital–temporal pathway (let us call it the 'ventral stream') are more dominantly part of the neural basis of (visual) consciousness than earlier areas. Recent work using imaging has extended and refined these findings. Kanwisher (2001) notes that 'neural correlates of perceptual experience, an exotic and elusive quarry just a few years ago, have suddenly become almost commonplace findings'. And she backs this up with impressive correlations between neural activation on the one hand and indications of perceptual experiences of faces, houses, motion, letters, objects, words, and speech on the other. Although the neural correlates of, say, faces and houses, are somewhat different, both are in the ventral stream, mainly in the higher areas. These results represent a major success: identification of the neural basis of visual consciousness in the ventral stream.

Paradoxically, what has also become commonplace is activation of the very same ventral stream pathways without 'awareness'. Damage to the inferior parietal and frontal lobes has long been known to cause visual extinction – in which subjects appear to lose subjective experience of certain stimuli on one side when there are stimuli on both sides. Extinction is associated with visual neglect, in which subjects do not notice stimuli on one side. For example, neglect patients often don't eat the food on the left side of the plate. Although subjects say they do not see the extinguished stimulus, the nature of the stimulus has various visual effects. For example, if the subject's task is to decide whether a letter string (e.g. 'saucer' or 'spiger') is a word, the subject is faster for 'saucer' if there is a picture of a cup or the word 'cup' in the neglected field, even though they perform no better than chance in guessing what the picture depicts (McGlinchey-Berroth *et al*., 1996). So the stimulus, of which the subject is in some sense unaware, is processed semantically.

Driver and Vuilleumier (2001) point out that the ventral stream is activated for extinguished stimuli (i.e. stimuli that the subject claims not to see). Rees *et al*. (2000) report studies of a left-sided neglect and extinction patient on face and house stimuli. Stimuli presented only on the left side are clearly seen by the patient, but when there are stimuli on both sides, the subject says he sees only the stimulus on the right. However, the 'unseen' stimuli produce a pattern of activation of the ventral pathway that is the same in location and temporal course as the seen stimuli (though lower in magnitude). Furthermore, studies in monkeys have shown that a well-known 'blindness' syndrome is caused by massive cortical ablation which spares most of the ventral stream but not the inferior parietal and frontal lobes (Lumer and Rees, 1999). Kanwisher (2001) notes that dynamic visual gratings alternating with a gray field showed activation for V1, V2, V3A, V4v, MT and MST despite the subjects saying they saw only a uniform gray field.

Zeki and Ffytch (1998) hypothesize that the difference between conscious and unconscious activation of the ventral pathway is just a matter of the degree of activation. But Kanwisher (2001) mentions that evidence from event-related potential (ERP) studies using the attentional blink paradigm show that neural activation of meaning is no less when the word is blinked than when it isn't, suggesting that it is not lower neural activation strength that accounts for lack of awareness. Dehaene and Naccache (2001) note that in a study of neglect patients, it was shown that there is the same amount of semantic priming from both hemifields, despite the lack of awareness of stimuli in the left field, again suggesting that it is not activation strength that makes the difference.

The paradox, then, is that our success in identifying the neural correlate of visual experience in normal vision has led to the conclusion that in masking and neglect, that very neural correlate occurs without, apparently, subjective experience.

What is the missing ingredient *X* which, added to ventral activation, constitutes conscious experience? Kanwisher (2001) and Driver and Vuilleumier (2001) offer similar proposals as to the nature of *X*, namely, that the missing ingredient is binding perceptual attributes with a time and a place, a token event. Rees *et al*. (2000) make two suggestions as to what *X* is. One is that the difference between conscious and unconscious activation is a matter of neural synchrony at fine timescales. This idea is supported by the finding that ERP components P1 and N1 revealed differences between left-sided 'unseen' stimuli and left-sided seen stimuli. Their second suggestion is that the difference between seen and 'unseen' stimuli might be a matter of interaction between the 'visual stream' as we know it and the areas of parietal and frontal cortex that control attention.

Whether or not any of these proposals is right, the search for *X* seems to be the most interesting current direction for consciousness research. For the purposes of the discussion below, we will assume that *X* equals neural synchrony.

PHYSICALISM AND FUNCTIONALISM

There is another, very different, approach to the nature of consciousness. Dennett (1994) postulates that consciousness is 'cerebral celebrity'. What it is for a representation to be conscious is for it to be widely available in the brain. Dehaene and Naccache (2001) say consciousness is being broadcast in a global neuronal workspace.

The theory that consciousness is ventral stream activation combined with neural synchrony and the theory that consciousness is broadcasting in the global neuronal workspace are instances of the two major approaches to consciousness in the philosophical literature: physicalism and functionalism. The difference is that the functionalist says that consciousness is a role, whereas the physicalist says that consciousness is a physical or biological state that implements that role. For example, red may play the role of warning of danger – but green

might also have played that role. The picture of consciousness as role could be characterized as computational in contrast to the biological picture of consciousness as implementer of the role.

Although functionalists are free to add restrictions, functionalism in its pure form is implementation-independent. Consciousness is defined as global accessibility, and although its human implementation depends on human biochemistry, silicon-based creatures without our biochemistry could implement the same computational relations. Functionalism and physicalism are incompatible doctrines, since a nonbiological implementation of the functional organization of consciousness would be regarded as uncontroversially conscious by the functionalist but not by the physicalist. The big question for functionalists is: 'how do we know that it is broadcasting in the global workspace that makes a representation conscious, as opposed to something about the human biological realization of that broadcasting?

The problem for functionalists could be put like this: the specifically human biochemical realization of global availability may be necessary to consciousness – other realizations of global availability being 'ersatz' realizations. The typical response to this 'ersatz realization problem' (Lycan, 1981) is that we can preserve functionalism by simply bringing lower-level causes and effects into our functional characterizations: for example, causes and effects at the level of biochemistry. But this response is inadequate because one can descend the hierarchy of sciences to the lowest level of all, that of basic physics, and find oneself vulnerable to the same point. Putting the point for simplicity in terms of the physics of the 1960s, the causal role of electrons is the same as that of anti-electrons. If you formulate a functional role for an electron, an anti-electron will realize it. Thus an anti-electron is an ersatz realizer of the functional definition of 'electron'. Physics is characterized by symmetries that allow ersatz realizations (Block, 1978b).

We have been talking about the two approaches of functionalism and physicalism as rivals, but we can instead see them as answers to different questions. The question that motivates the physicalist proposal of 'ventral activation plus *X*' is: what is the neural basis of experience? The question that motivates the 'global broadcasting' proposal is: what makes neuronal representations available for thought, decision, reporting and action? The former is a theory of phenomenal consciousness ('phenomenality'), and the latter of 'access consciousness'. (Theorists will differ on whether access consciousness is really a type of consciousness (Burge, 1997).) We can try to force a unity by postulating that it is a condition on *X* that it promote access, but that is merely a verbal maneuver, and only disguises the difference between the concepts and questions involved. Alternatively, we could hypothesize, rather than postulate, that *X* as a matter of fact is the neural basis of global neuronal broadcasting. Note, however, that the neural basis of global neuronal broadcasting might obtain but the normal channels of broadcasting none the less be blocked or cut, and this would again reveal the distinction between phenomenality and access, showing that we cannot think of the two as one. (As an analogy, rest mass and relativistic mass are importantly different from a theoretical point of view despite coinciding for all practical purposes at terrestrial velocities. Failure of coincidence, even if rare, is theoretically critical to the scientific nature of consciousness.)

Many of us have had the experience of suddenly noticing a sound (say, a drilling noise during an intense conversation), at the same time realizing that the sound has been going on for some time even though we were not attending to it. If there was a phenomenal representation of the sound before it was noticed, that is, if the sound was experienced first and noticed second, that state was not broadcast in the global neuronal workspace until it was noticed: there was a period of phenomenality without broadcasting. Of course, this is anecdotal evidence. But the starting point for work on consciousness is introspection, and we would be foolish to ignore it.

If we take seriously the idea of phenomenality without global accessibility, one theoretical option that we should consider is that ventral stream activation is visual phenomenality and the search for *X* is the search for the neural basis of what makes visual phenomenality accessible. The idea would be that the claims of extinction patients not to see extinguished stimuli are in a sense wrong: they really do have phenomenal experience of these stimuli without knowing it. A similar issue will arise in our discussion of the relation between phenomenality and a special case of global accessibility, reflexive or introspective consciousness, in which the subject not only has a phenomenal state but also has another state that is about the phenomenal state, say, a thought to the effect that he has the phenomenal state.

We have drawn a distinction between two concepts of consciousness, phenomenality and global accessibility. We will now add a third, reflexivity.

PHENOMENALITY AND REFLEXIVITY

Consider the 'false recognition' paradigm of Jacoby and Whitehouse (1989). Subjects are given a study list of 126 words presented for half a second each. They are then presented with a masked word, w_1 and an unmasked word w_2. Their task is to report whether w_2 was old (i.e. on the study list) or new (not on the study list). The variable was whether w_1 was lightly or heavily masked, the former presentations being thought of as 'conscious' and the latter as 'unconscious'. Confining our attention just to cases in which $w_1 = w_2$, subjects were much more likely to mistakenly report w_2 as old when w_1 was unconsciously presented than when w_1 was consciously presented. The explanation would appear to be that when w_1 was consciously presented, the subjects were able to use an internal monologue of the following sort (though perhaps not quite as explicit): 'the reason w_2 looks familiar is that I just saw it (as w_1)'. Thus they 'explained away' the familiarity of w_2. But when w_1 was unconsciously presented, the subjects were not able to indulge in this monologue and consequently attributed the familiarity of w_2 to its appearance in the study list.

Any monologue that can reasonably be attributed to the subject in this paradigm concerns why a word (w_2) 'looks familiar' to the subject. For it is only by 'explaining away' the familiarity of w_2 that the subject is able to decide that w_2 was not on the study list. Thus, in the 'conscious' case, the subject must have a state that is about the subject's own perceptual experience (looking familiar), and thus conscious in what might be termed a 'reflexive' sense. An experience is conscious in this sense just in case it is the object of another of the subject's states: for example, one has a thought to the effect that one has that experience. The reflexive sense of 'consciousness' contrasts with phenomenality, which may apply to some states that are not the objects of other mental states. Reflexive consciousness might better be called 'awareness' than 'consciousness'. Reflexivity is phenomenality together with something else (reflection), and there is the possibility in principle of phenomenality without reflection.

What is the relation between reflexivity and the notion of global accessibility discussed in the last section? Global accessibility does not logically require reflexivity, since global accessibility only requires access to all response modes that the organism actually has. (Perhaps a dog or a cat does not have the capacity for reflection.) Reflexivity is a special kind of access, one that requires intellectual resources that may not be available to every being that can have conscious experience.

There is another aspect of the experimental paradigm just discussed which motivates taking seriously the hypothesis that the reflexively unconscious case might possibly be phenomenally conscious. Consider a variant of the exclusion paradigm reported by Debner and Jacoby (1994). Subjects were briefly presented with pairs of words flanked by digits (e.g. '4reason5'), and then given stems consisting of the first three letters of the word ('rea___') to complete. Subjects were instructed to complete the stem, but not with the word that was briefly presented and flanked by digits. There were two conditions. In the 'conscious' condition, they were told to ignore the digits. In the 'unconscious' condition, they were told to report the sum of the digits before completing the stem. The results were that in the 'conscious' condition, the subjects were more likely than baseline to follow the instructions and complete the stem with a word other than 'reason', whereas in the 'unconscious' condition, subjects were more likely than baseline to violate the exclusion instructions, completing the stem with 'reason'. Merikle and Joordens (1997) report corresponding results for the false recognition paradigm with divided attention substituted for heavy masking.

Consider the hypothesis that there was a fleeting phenomenal consciousness of 'reason' as the subject's eyes moved from the '4' to the '5' in '4reason5'. There are two theoretical options that deserve serious consideration: either (1) the 'unconscious perceptions' are both phenomenally and reflexively unconscious (in this case, the exclusion and false recognition paradigms are about consciousness in both senses); or (2) they are fleetingly phenomenally conscious but reflexively unconscious. A third option, that (3) they are phenomenally unconscious but 'reflexively conscious' seems less likely because the reflexive consciousness would be 'false': subjects would have a state 'about' a phenomenal state without the phenomenal state itself. That hypothesis would require some extra causal factor that produced the false recognition, and would thus be less simple. One argument in favor of (2) is that subjects in experiments with near-threshold stimuli often report a 'mess' of partial perceptions that they 'can't hang on to'. Some critics (e.g. Dennett, 1991) have disparaged the idea of fleeting phenomenal consciousness. But they still do not provide a positive reason to think that (1) is the correct view.

It might seem that there is a principled argument that we could never find out about phenomenality

in the absence of reflexive consciousness, for we require the subject's testimony about phenomenality, which requires the subject to have a state that is about the phenomenal state. We can see what is wrong with this reasoning by attention to some potential lines of evidence for phenomenality in the absence of reflexive consciousness.

Liss (1968) contrasted subjects' responses to brief unmasked stimuli (one to four letters) with their responses to longer lightly masked stimuli. He asked for judgments of brightness, sharpness and contrast as well as what letters they saw. He found that lightly masked 40 ms stimuli were judged as brighter and sharper than unmasked 9 ms stimuli, even though the subjects could report three out of four letters in the unmasked stimuli and only one out of four in the masked cases. Liss writes: 'The subjects commented spontaneously that, despite the high contrast of the letters presented under backward masking, they seemed to appear for such brief duration that there was very little time to identify them before the mask appeared. Although letters presented for only 7 ms with no masking appeared weak and fuzzy, their duration seemed longer than letters presented for 70 ms followed by a mask.'

Perhaps the subjects were phenomenally conscious of all the masked letter shapes, but could not apply the letter concepts required for reflexive consciousness of them. (The subjects could apply the concepts of sharpness, brightness and contrast to the letters, so they did have reflexive consciousness of those features, even if they did not have reflexive consciousness of the shapes themselves. There are two kinds of shape concepts that could have provided – but apparently did not provide – reflexive consciousness of the letters: the letter concepts that we all learn in school, and shape concepts of the kind we have for unfamiliar shapes.) In other words, perhaps phenomenal experience of shapes does not require shape concepts but reflexive consciousness, being an intentional state, does require shape concepts, concepts that the subjects seem unable to access in these difficult attentional circumstances. Alternatively, perhaps the phenomenal experience of shapes does involve shape concepts of some sort but the use of those shape representations in reflexive consciousness requires more attentional resources than were available to these subjects.

There is another hypothesis: that the contents of both the subjects' phenomenal states and their reflective states are the same, and include the features 'sharp', 'high contrast', 'bright' and 'letter-like' without any specific shape representation. On this hypothesis, there is no gap between phenomenal and reflexive consciousness. Both hypotheses have to be taken seriously, but the first is superior in at least one respect: anyone who has been a subject in this or in Sperling's (1960) similar experiment will feel that the last hypothesis does not really capture the experience, which is an experience of being able to see more letters than one can categorize.

Rosenthal (1997) defines reflexive consciousness as follows: *S* is a reflexively conscious state of mine if and only if *S* is accompanied by a thought – arrived at non-inferentially and non-observationally – to the effect that I am in *S*. He offers this 'higher-order thought' (HOT) theory as a theory of phenomenal consciousness. Both phenomenal consciousness without HOT and HOT without phenomenal consciousness are conceptually possible. For example, perhaps dogs and infants have phenomenally conscious pains without higher-order thoughts about them. Conversely, imagine that by biofeedback and imaging techniques of the distant future I learn to detect the state in myself of having the Freudian unconscious thought that it would be nice to kill my father and marry my mother. I could come to know – non-inferentially and non-observationally – that I have this Freudian thought even though the thought is not phenomenally conscious. Since there are conceptually possible counterexamples in both directions, the question again is whether reflexivity and phenomenality come to the same thing empirically.

If there are no actual counterexamples, the question arises of why. Is it supposed to be a basic law of nature that phenomenality and reflexivity co-occur? That would be a very adventurous claim that no one is in a position to make. Is it a contingent fact about us but not other phenomenally conscious creatures? Then reflexivity would not provide a basic account of phenomenality. Well then, is it supposed to be a contingent fact that phenomenality and reflexivity are coextensive in all creatures? What would be the evidence for such a far-reaching claim? Further, if phenomenality and reflexivity are correlated, then there must be a mechanism that explains the correlation, as the fact that both heat and electricity are carried by free electrons explains the correlation of electrical and thermal conductivity. But any mechanism breaks down under extreme conditions – as does the correlation of electrical and thermal conductivity at extremely high temperatures. So the correlation between phenomenality and reflexivity would break down too, showing that reflexivity does not yield the fundamental scientific nature of phenomenality.

Alternatively, it might be said that 'consciousness' is a 'natural kind' term, like 'water' or 'heat' or 'light'. We know that water is H_2O, as a matter of empirical fact without having to ask whether there might be a substance in another solar system that has a very different chemical constitution but nonetheless behaves exactly like water. The reason is that as a matter of the semantics of the word 'water', the question of whether there could be a substance that behaved exactly like water but that had a completely different chemical constitution does not matter since it would be wrong to call it 'water'. But this is a poor analogy for two reasons. First, if there are beings in another solar system who have something that feels like phenomenality but without reflexivity, then reflexivity does not capture the fundamental nature of phenomenality, whatever we call what the aliens have. Second, we call anything that feels like phenomenality 'phenomenality', so phenomenality is not a natural kind concept in the sense that 'water' is.

Rosenthal's definition of reflexivity has a number of *ad hoc* features. Non-observationality is required to rule out, for example, a case in which I know about a thought I have repressed by observing my own behavior. Non-inferentiality is required to avoid a somewhat different case in which I appreciate (non-observationally) my own psychic pain and infer a repressed thought from it. But why should the consciousness of a state depend on its causal history? Furthermore, Rosenthal's definition involves a stipulation that the possessor of the reflexively conscious state is the same as the thinker of the thought – otherwise my thinking about your pain would make it a conscious pain. All these *ad hoc* features can be eliminated by moving to the following definition of reflexivity: *S* is a reflexively conscious state if and only if *S* is phenomenally presented in a thought about *S*. This definition uses the notion of phenomenality; but this is no disadvantage unless one holds that there is no such thing apart from reflexivity itself. The new definition of reflexivity, requiring phenomenality as it does, has the additional advantage of making it clear that reflexivity is a kind of consciousness. (See Burge's (1997) critique of our notion of 'access consciousness' as constituting a kind of consciousness.)

We have encountered three concepts of consciousness: phenomenality, reflexivity, and global accessibility. The hard problem arises only for phenomenality. Imaging experiments on consciousness engage phenomenality and accessibility. But many psychological experimental paradigms mainly engage reflexivity. So empirical investigators of 'consciousness' may sometimes be talking at cross purposes.

References

Armstrong DM (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block N (1978a) Reductionism. In: *Encyclopedia of Bioethics*, pp. 1419–1424. New York, NY: Macmillan.

Block N (1978b) Troubles with functionalism. In: Savage CW (ed.) *Minnesota Studies in the Philosophy of Science*, vol. IX, pp. 261–325. [Reprinted abridged in: Rosenthal DM (ed.) (1991) *The Nature of Mind*, pp. 211–229. Oxford: Oxford University Press.]

Block N (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**(2): 651–726.

Block N, Flanagan O and Güzeldere G (eds) (1997) *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

Block N and Stalnaker R (1999) Conceptual analysis, dualism and the explanatory gap. *Philosophical Review* **108**: 1–46.

Burge T (1997) Two kinds of consciousness. In: Block N, Flanagar O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 427–434. Cambridge, MA: MIT Press.

Carruthers P (1992) Consciousness and concepts. *Proceedings of the Aristotelian Society* **66**: 41–59.

Chalmers D (1996) *The Conscious Mind*. New York, NY: Oxford University Press.

Churchland P (1989) Knowing qualia: a reply to Jackson. In: *A Neurocomputational Perspective*, pp. 67–76. Cambridge, MA: MIT Press.

Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2**: 263–275.

Debner JA and Jacoby LL (1994) Unconscious perception: attention, awareness and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 304–317.

Dehaene S and Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**(1–2): 1–37.

Dennett D (1979) On the absence of phenomenology. In: Gustafson D and Tapscott B (eds) *Body, Mind and Method: Essays in Honor of Virgil Aldrich*, pp. 93–113. Dordrecht: Reidel.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Dennett D (1994) Get Real. In: *Philosophical Topics* **22** (1–2): 505–568.

Dretske F (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Driver J and Vuilleumier P (2000) Perceptual awareness and its loss is unilateral neglect and extinction. *Cognition* **79**(1–2): 39–89.

Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Harman G (1990) The intrinsic quality of experience. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. IV

'*Action Theory and Philosophy of Mind*', pp. 31–52. Atascadero, CA: Ridgeview.
Harman G (1996) Explaining objective color in terms of subjective reactions. In: Villanueva E (ed.) *Philosophical Issues 7: Perception*. Atascadero, CA: Ridgeview.
Jackson F (1977) *Perception*. Cambridge, UK: Cambridge University Press.
Jackson F (1982) Epiphenomenal qualia. *Philosophical Studies* **32**: 127–136.
Jacoby LL and Whitehouse K (1989) An illusion of memory: false recognition influenced by unconscious perception. *Journal of Experimental Psychology: General* **118**: 126–135.
Kanwisher N (2001) Neural events and perceptual awareness. *Cognition* **79**(1–2): 89–113.
Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.
Lewis D (1994) Reduction of mind. In: Guttenplan S (ed.) *Blackwell's Companion to Philosophy of Mind*, pp. 412–431. Oxford: Blackwell.
Liss P (1968) Does backward masking by visual noise stop stimulus processing? *Perception and Psychophysics* **4**: 328–330.
Loar B (1990/1997) Phenomenal states. In: Tomberlin J (ed) *Philosophical Perspectives*, vol. IV '*Action Theory and Philosophy of Mind*', pp. 81–108. Atascadero, CA: Ridgeview. [A much-revised version of this paper is to be found in (Block *et al*., 1997, pp. 597–616).]
Logothetis NK (1998) Single units and conscious vision. *Proceedings of the Royal Society of London, Series B* **353**: 1801–1818.
Lumer E and Rees G (1999) Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proceedings of the National Academy of Sciences* **96**: 1669–1673.
Lycan WG (1981) Form, function and feel. *Journal of Philosophy* **78**: 24–50.
Lycan WG (1990) Consciousness as internal monitoring, In: Tomberlin J (ed) *Philosophical Perspectives*, vol. IX, pp. 1–14. Atascadero, CA: Ridgeview. [Reprinted in (Block *et al*., 1997).]
Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.
McDowell J (1994) The content of perceptual experience. *Philosophical Quarterly* **44**: 190–205.
McGinn C (1991) *The Problem of Consciousness*. Oxford: Blackwell.
McGlinchey-Berroth R, Milberg WP, Verfaellie M *et al*. (1996) Semantic processing and orthographic specificity in hemispatial neglect. *Journal of Cognitive Neuroscience* **8**: 291–304.
Merikle P and Joordens S (1997) Parallels between perception without attention and perception without awareness. *Consciousness and Cognition* **6**: 219–236.
Nagel T (1974) What is it like to be a bat? *Philosophical Review* **83**: 435–450.
Nagel T (2000) The psychophysical nexus. In: Boghossian P and Peacocke C (eds) *New Essays on the A Priori*, pp. 434–471. Oxford: Oxford University Press.
Perry J (2001) *Knowledge, Possibility and Consciousness*. Cambridge, MA: MIT Press.
Pitcher G (1971) *A Theory of Perception.* Princeton, NJ: Princeton University Press.
Rees G, Wojciulik E, Clarke K *et al*. (2000) Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain* **123**: 1624–1633.
Rey G (1997) *Contemporary Philosophy of Mind*. Oxford: Blackwell.
Rosenthal DM (1997) A theory of consciousness. In: Block N, Flanagar O and Güzeldere G (eds) (1997) *The Nature of Consciousness: Philosophical Debates*, pp. 729–754. Cambridge, MA: MIT Press.
Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
Shoemaker S (1994) Self-knowledge and 'inner sense'. Lecture III: The phenomenal character of experience. *Philosophy and Phenomenological Research* **54**(2): 291–314.
Sperling G (1960) The information available in brief visual presentations. *Psychological Monographs* **74**(11): 1–29.
Sturgeon S (1994) The epistemic view of subjectivity. *Journal of Philosophy* **91**: 221–235.
Tye M (1995) *Ten Problems of Consciousness.* Cambridge, MA: MIT Press.
Tye M (1999) Phenomenal consciousness: the explanatory gap as a cognitive illusion. *Mind* **108**: 706–725.
Van Gulick R (1993) Understanding the phenomenal mind: are we all just armadillos? In: Davies M and Humphreys G (eds) *Consciousness: Psychological and Philosophical Essays*, pp. 137–149. Oxford: Blackwell. [Reprinted in (Block *et al*., 1997).]
Zeki S and Ffytch DH (1998) The Riddoch syndrome: insights into the neurobiology of conscious vision. *Brain* **121**: 25–45.

Further Reading

Block N, Flanagan O and Güzeldere G (eds) (1997) *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press. [A collection of papers on consciousness.]
Dehaene S (ed.) (2001) *Cognition* **79**(1–2). [A special issue on the cognitive neuroscience of consciousness.]
Güzeldere G (1997) The many faces of consciousness: *a field guide*. In: Block N, Flanagar O and Güzeldere G (eds) (1997) *The Nature of Consciousness: Philosophical Debates*, pp. 1–67. Cambridge, MA: MIT Press.
Huxley TH (1866) *Lessons in Elementary Physiology*. London: Macmillan.
Jackson F (1986) What Mary didn't know. *Journal of Philosophy* **83**: 291–295.
Jackson F (1993) Armchair metaphysics. In: O'Leary-Hawthorne J and Michael M (eds) *Philosophy in Mind*, pp. 23–42. Dordrecht: Kluwer.
Putnam H (1967) *Psychological predicates*. In: Capitan WH and Merrill DD (eds) *Art, Mind and Religion*. Pittsburgh, PA: Pittsburgh University Press. [Later entitled 'The

nature of mental states.' Reprinted in: Putnam M (1975) *Mind, Language, and Reality*, pp. 429–440. Cambridge, UK: Cambridge University Press.]

Smart JJC (1959) Sensations and brain processes. *Philosophical Review* **68**: 141–156. [This paper has been reprinted many times, starting in 1962 in a somewhat revised form. See, for example: Rosenthal DM (ed.) (1971) *Materialism and the Mind–Body Problem*. Englewood Cliffs, NJ: Prentice-Hall.]

Strawson G (1994) *Mental Reality*. Cambridge, MA: MIT Press.

Consciousness, Sleep, and Dreaming

Intermediate article

J Allan Hobson, Harvard Medical School, Boston, Massachusetts, USA

CONTENTS

Advances in, and links between, neurobiology and psychology as applied in the study of sleep and dreaming compared to the waking state are increasingly revealing significant evidence for the brain basis of consciousness.

CONSCIOUS STATES AND BRAIN–MIND ISOMORPHISM

How can a material structure, like the brain, possibly possess awareness? Philosophers refer to this conundrum as 'the hard problem' and many are deeply pessimistic about its resolution. Countering such pessimism are recent developments in the cognitive neuroscience of waking, sleeping, and dreaming which have made it possible to understand – in considerable detail – how the brain changes its conscious state every day – and every night – of our lives. Even if we cannot yet detail how the brain becomes conscious in the first place, specification of the brain mechanisms responsible for the dramatic changes in consciousness that contrast waking and dreaming constitutes a partial solution of the brain–mind question and points the way to its more complete resolution in the foreseeable future (Hobson, 1999).

That dreaming might provide a privileged access to understanding the relationship of brain and mind has long been recognized by such diverse thinkers as the Enlightenment philosopher David Hartley and the Romantic poet Samuel Taylor Coleridge. Wilhelm Wundt, the father of experimental psychology, astutely speculated that to account for the difference between dreaming and waking consciousness, those brain functions supporting critical thought and self-reflective awareness must be in abeyance (in dreaming) while those supporting internal perceptions and emotions must be enhanced.

In declaring that dreaming was the royal road to the unconscious mind, Sigmund Freud was strongly influenced by his earlier attempt to create a scientific psychology based upon the structures and functions of the 100 billion nerve cells that constitute the brain. Until recently none of these theories could advance because so little was known about the activity of the brain during sleep.

Since the late 1920s brain science has seen dramatic advances that now make a solid neurology of dreaming possible. The first key step in this direction was the discovery of the electroencephalogram (EEG) by Adolf Berger in 1928 and the recognition by him that the pattern of brain waves changed when subjects became inattentive or drowsy. Figure 1 illustrates the normal human sleep cycle as it is understood today.

The idea that the brain waves of the cerebral cortex had to be kept electrically activated to support waking consciousness was crystallized in the discovery of the importance to arousal of the brain stem reticular formation (Moruzzi and Magoun, 1949). Shortly thereafter, it was discovered that the brain was periodically reactivated during sleep and that this activation was associated with rapid eye movements (or REMs) and with intense and sustained dream consciousness (Aserinsky and Kleitman, 1953; Dement and Kleitman, 1957).

It remained to give an account of the mechanisms of electrical activation of the cortex by the brain stem and to distinguish between the brain processes associated with the electrical activation that occurred in waking and in sleep in such a way as to distinguish between dreaming and waking consciousness. These details emerged from studying the animal model of REM sleep provided by the discovery that REM sleep also occurred periodically in the sleep of cats (Dement, 1958).

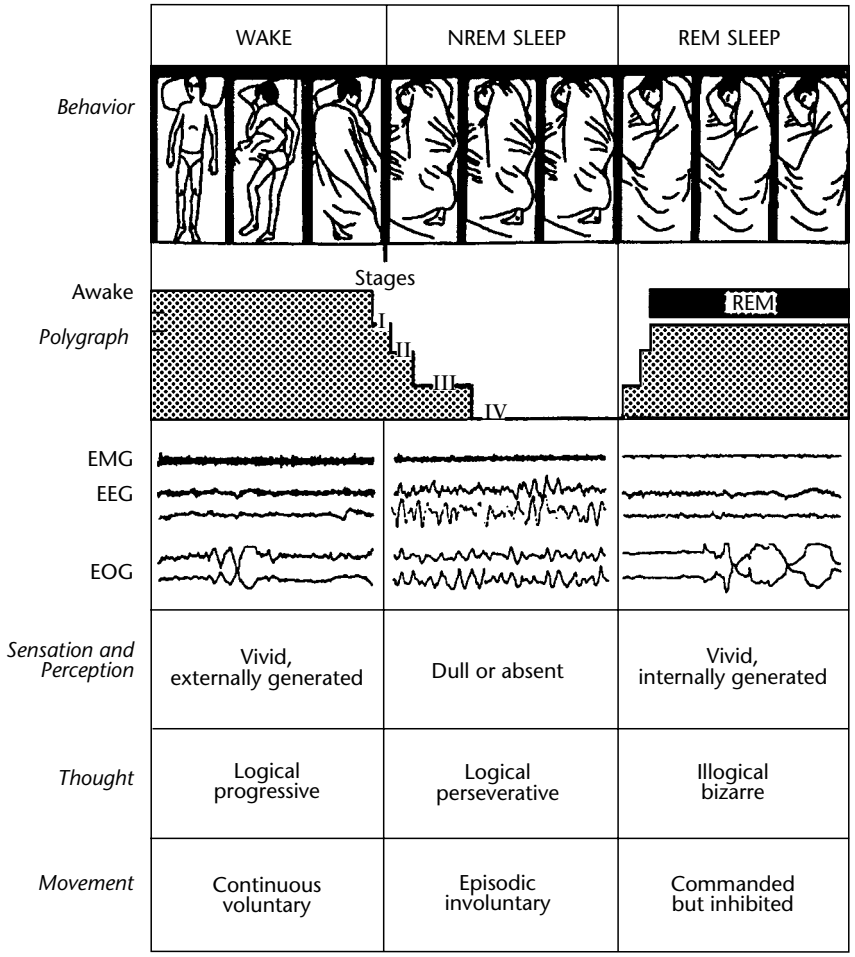


Figure 1. Behavioral states in humans. States of waking, rapid eye movement (REM) sleep, and non-REM (NREM) sleep have behavioral, polygraphic, and psychological manifestations. In the behavior channel, posture shifts (detectable by time-lapse photography or video) can occur during waking and in concert with phase changes of the sleep cycle. Two different mechanisms account for sleep immobility: disfacilitation (during stages I–IV of NREM sleep) and inhibition (during REM sleep). In dreams we imagine that we move but we do not. Sequence of these stages is represented in the polygraph channel. Sample tracings of three variables used to distinguish state are also shown: electromyogram (EMG), which is highest in waking, intermediate in NREM sleep, and lowest in REM sleep; and electroencephalogram (EEG) and electrooculogram (EOG), which are both activated in waking and REM sleep and inactivated in NREM sleep. Each sample record is 20 s. Three lower channels describe other subjective and objective state variables. (From Hobson and Steriade, 1986.)

Working with cats, Michel Jouvet had, by 1959, demonstrated that the brain activation of REM sleep (like that of waking) depended upon the brain stem. That the mechanisms of brain activation were quite different in REM sleep and waking became apparent from Jouvet's discovery that the postural tone of the muscles necessary to support the motor activity of waking was actively abolished in REM. Jouvet also showed that the cerebral cortex was not only activated in REM but received internal stimuli from the brain stem regarding the movement of the eyes. Because these internal visual signals are much stronger in REM sleep than in waking it was possible, for the first time, to imagine how the dreaming brain could produce visual imagery of great intensity entirely on its own (Jouvet, 1962).

All these important discoveries set the stage for the analysis of the cellular and molecular brain processes that mediated the differences between dreaming and waking consciousness. Using the movable microelectrode technique pioneered by the vision specialist David Hubel and the motor expert Edward Evarts, it was possible for Robert McCarley and Allan Hobson to propose a model of reciprocal interaction between two chemically

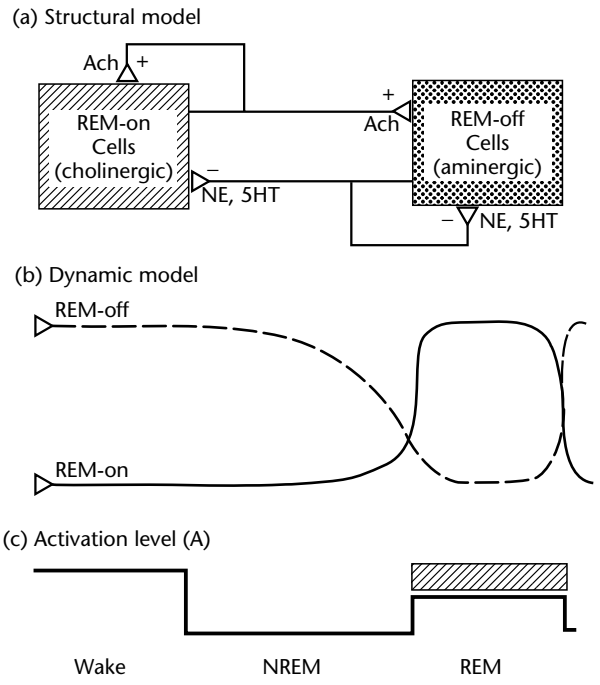


Figure 2. The original reciprocal interaction model of physiological mechanisms determining alterations in activation level. (a) Structural model of reciprocal interaction: REM-on cells of the pontine reticular formation are cholinoreceptively excited and/or cholinergically excitatory (ACH+) at their synaptic endings. Pontine REM-off cells are noradrenergically (NE) or serotonergically (5HT) inhibitory (−) at their synapses. (b) Dynamic model: during waking the pontine aminergic system is tonically activated and inhibits the pontine cholinergic system. During NREM sleep aminergic inhibition gradually wanes and cholinergic excitation reciprocally waxes. At REM sleep onset aminergic inhibition is shut off and cholinergic excitation reaches its high point. (c) Activation level: as a consequence of the interplay of the neuronal systems shown in (a) and (b), the net activation level of the brain is at equally high levels in waking and REM sleep and at about half this peak level in NREM sleep. (Taken from Hobson, 1992.)

distinct brain stem cell groups (Hobson *et al*., 1975; McCarley and Hobson, 1975) and to tie that model to a brain-based dream theory, the activation-synthesis hypothesis (Hobson and McCarley, 1977). Figure 2 illustrates the original reciprocal interaction model and its behavioral consequences, while Figure 3 illustrates the activation-synthesis hypothesis. By integrating these two models with new data, it has recently been possible to elaborate AIM, a three-dimensional state space specifying the major physiological determinants of states of consciousness (discussed below).

FORMAL PROPERTIES OF DREAM CONSCIOUSNESS

Previous approaches to dreaming have reflected the natural tendency to regard dreams as story-like experiences conveying and/or concealing hidden or symbolic meaning. This interpretive tradition is ancient and while it remains humanistically honorable, it has always been scientifically unsatisfactory. Its most zealous modern exponent, Sigmund Freud, understood that a brain-based theory of dreaming was required but, lacking the tools to construct one, he instead created the pseudoscience of psychoanalysis.

Whether dreams are stories or scenarios, and whatever their meaning might or might not be, dreaming is a state of consciousness with distinctive properties that can be defined and measured when the emphasis is placed upon the form of all dreams rather than the particular content of each one. These formal properties of dream consciousness can then be compared with those of waking. Any differences that emerge can then be mapped onto differences between REM sleep and waking state neurobiology (Hobson, 1988).

The emphasis upon the analysis of dream form and the definition and measurement of formal dream properties is every bit as important as the recording of individual nerve cells because it brings subjectivity itself into the scientific arena. In so doing it allows us to describe and quantify

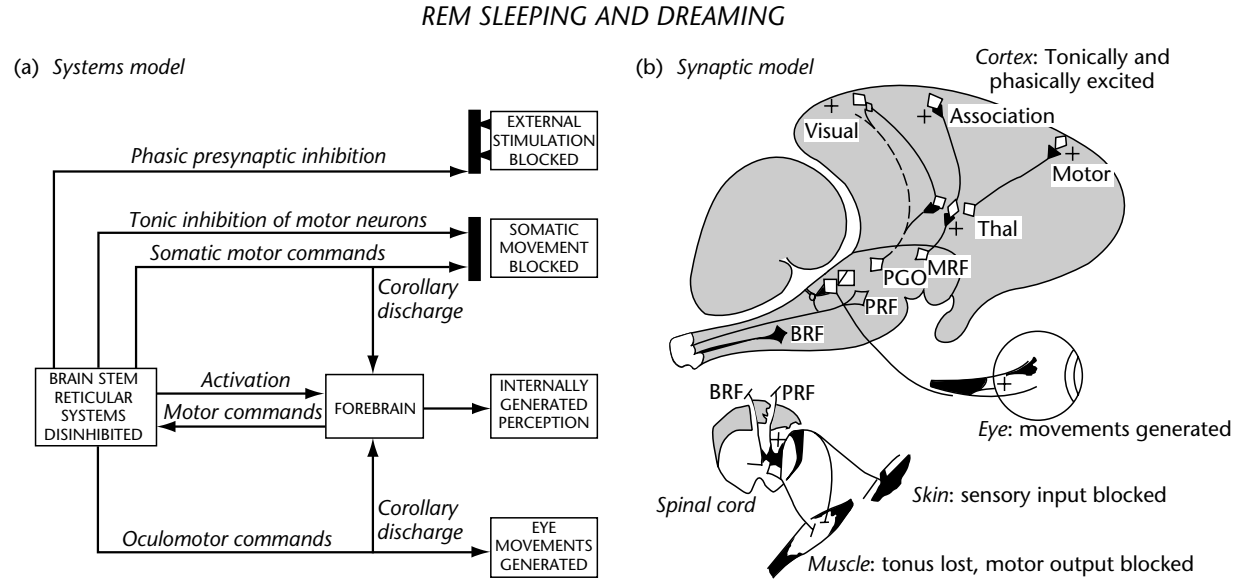


Figure 3. The activation-synthesis model. (a) Systems model: as a result of disinhibition caused by cessation of aminergic neuronal firing, brain stem reticular systems auto-activate. Their outputs have effects including depolarization of afferent terminals causing phasic presynaptic inhibition and blockade of external stimuli, especially during the bursts of REM, and postsynaptic hyperpolarization causing tonic inhibition of motor neurons that effectively counteract concomitant motor commands so that somatic movement is blocked. Only the oculomotor commands are read out as eye movements because these motor neurons are not inhibited. The forebrain, activated by the reticular formation and also aminergically disinhibited, receives efferent copy or corollary discharge information about somatic motor and oculomotor commands from which it may synthesize such internally generated perceptions as visual imagery and the sensation of movement, both of which typify dream mentation. The forebrain may, in turn, generate its own motor commands that help to perpetuate the process via positive feedback to the reticular formation. (b) Synaptic model: the midbrain reticular neurons (MRF) projecting to the thalamus convey tonic and phasic signals rostrally. PGO burst cells in the peribrachial region convey phasic activation and specific eye movement information to geniculate body and cortex (dotted line indicates uncertainty of direct projection). The pontine reticular-formation neurons (PRF) transmit phasic activation signals to oculomotor neurons (V1) and spinal cord which generate eye movements, twitches of extremities, and presynaptic inhibition. The bulbar reticular-formation neurons (BRF) send tonic hyperpolarizing signals to motor neurons in the spinal cord. As a consequence of these descending influences, sensory input and motor output are blocked at level of the spinal cord. At the level of the forebrain, visual association and motor cortex neurons all receive time and phasic activation signals for nonspecific and specific thalamic relays.

universal aspects of conscious experience, leaving aside for the moment individual differences which are more difficult to explain today by a brain-based theory.

The formal psychological feature which has been most extensively analyzed is the bizarreness that makes dreaming so strange, so puzzling, and by turns so amusing and so frightening. To our great surprise, we found that dream bizarreness was reducible to discontinuity and incongruity in the domains of time, place, and person, all three of which are quantitatively more unstable in dreaming than in waking consciousness. And one of them, the tendency of characters to have unstable or hybrid identities, seems to be absolutely unique to dreaming. Only in psychosis do people experience such disturbing phenomena when awake.

Dream consciousness is also characterized by internally generated sensory imagery of hallucinatory clarity and intensity. Dream imagery is predominantly visual, and is associated with the continuous illusion of movement. Dream consciousness is dominated by a false belief that we are awake, which only rarely gives way to the insight (called 'lucidity') that we are actually dreaming. And this delusional belief that we are awake persists despite robust cognitive evidence to the contrary. We simply don't seem to notice that when dreaming we are so disoriented that times, places, and persons change without our notice. Furthermore we cannot direct our thoughts, or our actions, as we do the most improbable or even downright impossible things, such as flying or cycling effortlessly uphill.

It seems probable that one of the most fundamental cognitive features of dreaming is a defect in episodic memory. When we dream we cannot remember prior events or even those that have just occurred in the dream. Furthermore, we cannot remember most of our dreams after we wake up. These perceptual and cognitive peculiarities are most often accompanied by strong negative emotions, such as anxiety, fear, and anger, but pleasantly intense elation may also be prominent. Table 1 summarizes these formal properties of dreaming.

Following Wilhelm Wundt, we agree that any dream theory thus needs to explain two classes of change in the state of consciousness from that of waking: (1) the enhancement of the perceptual and emotional components of consciousness, and (2) the impairment of its orientational, insightful, and memory components. Freud dealt with this dual aspect of dream consciousness by positing a defect in superego and ego functions and a reciprocal increase in instinctual id functions. As will be clear when we discuss new evidence for deactivation of the frontal cortex and a reciprocal increase in subcortical activation during human REM sleep, the new brain-based theory accounts for this kind of duality in a highly specific way.

By demonstrating the automaticity of the REM sleep generator in the brain stem, the activation-synthesis theory (Figure 3) eliminates psychological motives (Freud's unconscious wishes) from the instigation of dreaming. And by demonstrating the neurobiological mechanisms of dream formation, the new theory obviates Freud's disguised censorship hypothesis. It thus greatly weakens the credibility of any interpretative scheme based upon these two erroneous notions.

Before turning to the neurobiology, it is important to emphasize the heuristic value of a brain-based theory of dreams for the scientific understanding of abnormal as well as normal states of consciousness. This is because dream consciousness has so many features of major mental illness. And because it is both hallucinatory and delusional, normal dream consciousness is a psychotic state by definition. But to what abnormal psychosis is dreaming consciousness most similar? Not to schizophrenia where the hallucinations are primarily auditory and not visuomotor. And not to psychotic depression where the most prominent dysphoric emotions are sadness, guilt, and shame – not anxiety, anger, or elation. Because the hallucinations are visual, the thinking undirected, the orientation unstable, and episodic memory badly impaired, dream consciousness is most akin to those delirious states that are associated with toxic conditions which impair the brain organically. Table 2 summarizes these distinct differences between waking and dreaming consciousness.

REM SLEEP NEUROBIOLOGY

We now know that the brain is in an organically altered state in REM sleep. Although it is electrically activated to a level equal to or even exceeding that of waking, the regional pattern of activation is quite different. And because the input–output gates are closed there is no discourse with the outside world. As a result there is no external space–time constancy against which to check the internally generated percepts and emotions. Finally, the percepts and emotions are engendered because of a shift in the chemical balance of the brain which also impairs our memory, our internal

Table 1. The formal features of REM sleep dreaming

Hallucinations – especially visual and motoric, but occasionally in any and all sensory modalities.
Bizarreness – Incongruity (imagery is strange, unusual, or impossible); discontinuity (imagery and plot can change, appear, or disappear rapidly); uncertainty (persons, places, and events often bizarrely uncertain by waking standards).
Delusion – we are consistently duped into believing that we are awake (unless we cultivate lucidity).
Self-reflection absent or greatly reduced relative to waking.
Lack of orientational stability – persons, times, and places are fused, plastic, incongruous, and discontinuous.
Narrative story lines explain and integrate all the dream elements in a confabulatory manner.
Emotions increased, intensified, and predominated by fear anxiety.
Instinctual programs (especially fight–flight) often incorporated.
Volitional control greatly attenuated.
Memory deficits across dream–wake, wake–dream and dream–dream transitions.

Table 2. Cognitive differences between wake and dream states

Function	*Dream compared to wake*
Perception (external)	Diminished
Perception (internal)	Enhanced
Attention	Lost
Memory (recent)	Diminished
Memory (remote)	Enhanced
Emotion	Episodically strong
Orientation	Unstable
Thought	Reasoning *ad hoc*, logical rigor weak, processing hyperassociative
Insight	Self-reflection greatly diminished
Narrative construction	Confabulatory
Volition	Weak

orientational stability, and our ability to think insightfully and critically. Three neurobiological factors affecting consciousness can thus be identified and defined as follows:

1. *Activation (A)*: activating the brain so that consciousness can be at least as intense in REM sleep dreaming as it is in waking is the responsibility of the brain stem reticular formation. This system turns off at sleep onset, allowing the thalamus and cortex to indulge in their own intrinsic rhythmic activity which are seen as the spindles and slow waves of NREM sleep when consciousness abates.
2. *Input–Output Gating (I)*: when the reticular formation turns on again in REM, the brain stem simultaneously inhibits sensory input and blocks motor output so that we do not wake up even though the brain is activated. This input–output blockade is an active process which naturally puts dream consciousness offline, as it were. The eyes move in REM because internal signals are spared the active inhibition that affects the other motor systems of the brain. They arise in the pons whence they are also directed to the upper brain where they stimulate visual and emotional centers in concert with the REMs and thus cause the visuomotor hallucinosis.
3. *Mode (M)*: the internally generated signals emanate from the lateral part of the pontine brain stem where neurons that manufacture acetylcholine are found. Acetylcholine is one of the chemicals involved in learning and memory and it is released in both waking and REM. In waking, the acetylcholine neurons are activated by external stimuli, whereas in REM they become spontaneously active. This is because they are then released from the inhibitory restraint of norepinephrine and serotonin neurons which are active and quell them in waking (see Figure 2). Norepinephrine and serotonin are two other brain chemicals that contribute to attention and memory, and their unavailability in REM contributes to the cognitive deficits in dream consciousness.

The net result of this change in neuronal traffic flow in REM sleep is an electrically activated, offline brain, whose chemical balance has been tipped in the acetylcholine direction. Because the brain is activated it is conscious of its own spontaneous information processing. But because it is offline it is cut off from external reality. Because of the shift in chemical balance, it generates its own signals and interprets them as if they came from the outside world. But the brain is not only deprived of external cues that help to structure waking consciousness. It is also deprived of two of the chemicals that it uses to organize its information in a coherent, logical, and directed manner. This is why dream consciousness is characterized by so much incongruity and discontinuity.

THE AIM MODEL

The perspicacious reader will now realize that the three neurobiological factors we have been discussing collaborate to determine our state of consciousness in the following way:

1. *The Activation level* (A) determines the *intensity* of consciousness. Factor A operates very much in the manner of a power supply. Its value which can be assessed from the EEG, is high in wake and REM but low in non-REM (NREM) sleep.
2. *The Input–Output gating level* (I) determines the source of the information that determines the *content* of consciousness. Its value can be determined by the degree of presynaptic inhibition on the sensory side and by the degree of inhibition of spinal reflex activity on the output side. When the input–output gates are open, as in waking, external information plays a major role in sleeping consciousness. When the input–output gates are closed, as they are in REM, only internal data can be processed.
3. *The Mode* (M) of the brain determines how the internally generated information is processed primarily via its impact upon memory systems. Its value can be determined (so far only in animals) by the ratio of activity of norepinephrine- or serotonin-containing neurons to that of the acetylcholine cells. If the value of M is high, memories will be recorded, but if M is low – as it is in REM – memory will be deficient.

The range of values of A, I, and M can be laid out along the three dimensions of a cube. The resulting cubic space contains and bounds a virtual infinity of possible state points, each of which represents the instantaneous values of A, I, and M. As illustrated in Figure 4(a), the canonical states of waking, NREM sleep, and REM occupy specific

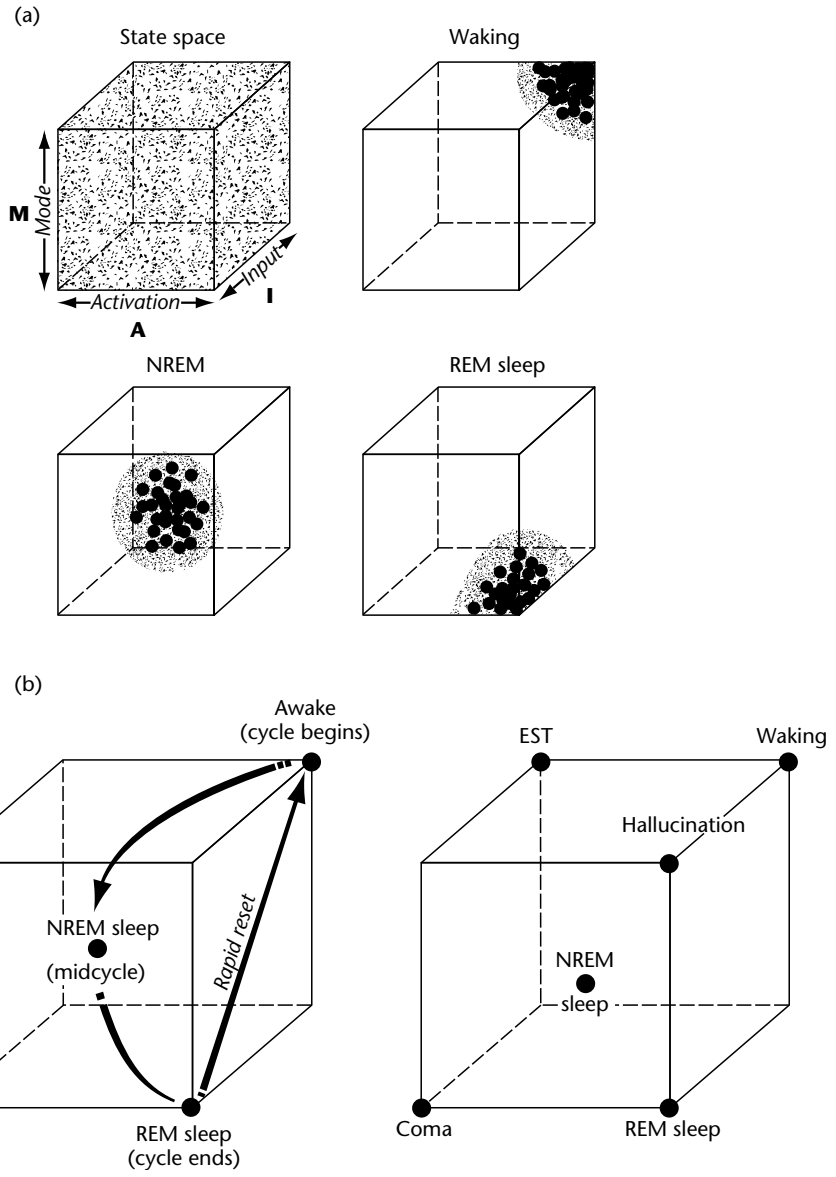


Figure 4. (a) Three-dimensional state space defined by the values for brain activation (A), input source and strength (I), and mode of processing (M). It is theoretically possible for the system to be at any point in the state space and an infinite number of state conditions is conceivable. In practice the system is normally constrained to a boomerang-like path from the back upper right in Waking (high A, I, and M), through the centre in NREM (intermediate A, I, and M) to the front lower right in REM sleep (high A, low I and M). (b) The canonical trajectory of the brain–mind is shown for each cycle of adult human sleep (left box). That this trajectory is an irregular ellipse is in keeping with the cyclical behavior of the brain stem neuronal control system for REM sleep. Besides the canonical states of wake, NREM, and REM, are "forbidden zones" which are only entered under pathological or exceptional normal conditions (right box).

subregions of the consciousness state space. With time as a fourth dimension, the normal diurnal trajectory through this space can be visualized as the sequential points follow a large-diameter elliptical path (representing the circadian rhythm) with smaller dramatic ellipses within it (representing the NREM–REM sleep cycle).

At this early stage, the AIM model is only a heuristic, illustrative construct. But, for the student of consciousness, the three-dimensional model has several advantages over traditional two-dimensional graphic displays of brain–mind state.

The first is its ability to distinguish wake and REM which, despite their identity on the activation

axis, are strongly contrasted along both the I and M axes. This contrast is congruent with the very different formal properties of consciousness in the two states.

The second advantage of AIM is its capacity to accommodate the myriad substates that constitute the conscious vicissitudes of everyday life. These include sleepiness, hyperalertness, fantasy, insomnia, and even creative imagination.

The third advantage of AIM is its recognition of those off-limit sectors of the state space into which consciousness may be pulled – or pushed – when humans suffer from spontaneous dysfunctional states such as epilepsy, schizophrenia, sleep disorders, or the comas that follow head injury. Conceptualization of such abnormal states using the three-dimensional AIM model is illustrated in Figure 4(b).

Finally, the AIM model is at home to the many intentional self-experiments on consciousness performed by humans using and abusing drugs like alcohol, amphetamine, LSD, and cocaine. Its explanatory generosity thus welcomes the vast family of recreational and therapeutic chemicals that play so powerfully – for good and evil – on the modulatory neurons of the brain.

HUMAN NEUROPSYCHOLOGY

Until recently, the experimental study of human REM sleep dreaming has been limited on the physiological side by the poor resolving power of the EEG. Even expensive and cumbersome evoked potential and computer averaging approaches have not helped to analyze and compare REM sleep physiology with that of waking in an effective way. This limitation has probably reinforced the erroneous idea that the brain activation mechanisms of REM sleep and waking are identical, or at least very similar. Fortunately, technological advances in the field of human brain imaging have now made it possible to describe a highly selective regional activation pattern of the brain in REM sleep. At the same time, experiments of nature – in the form of strokes – have allowed the locale of brain lesions to be correlated with the diminution or intensification of various aspects of dream experience in patients (Hobson *et al*., 1998).

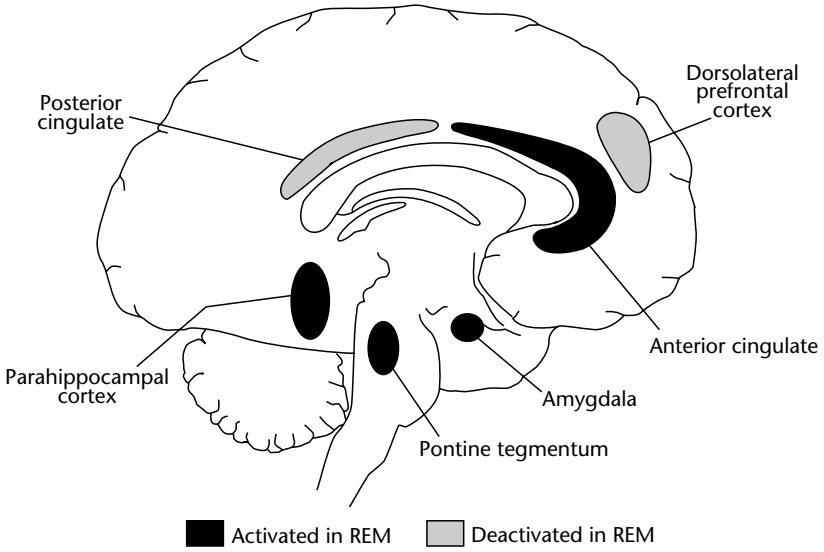


Figure 5. Convergent findings on relative regional brain activation and deactivation in REM compared to waking. A schematic sagittal view of the human brain showing those areas of relative activation and deactivation in REM sleep compared to waking and/or NREM sleep which were reported in *two or more* of the three PET studies published to date (Braun *et al*., 1997; Maquet *et al*., 1996; Nofzinger *et al*., 1997). Only those areas which could be easily matched between two or more studies are schematically illustrated here, and a realistic morphology of the depicted areas is not implied. The depicted areas in this figure are thus most realistically viewed as representative portions of larger central nervous system areas subserving similar functions (e.g. limbic-related cortex, ascending activation pathways, and multimodal association cortex). (Source: Hobson *et al*., 1998.)

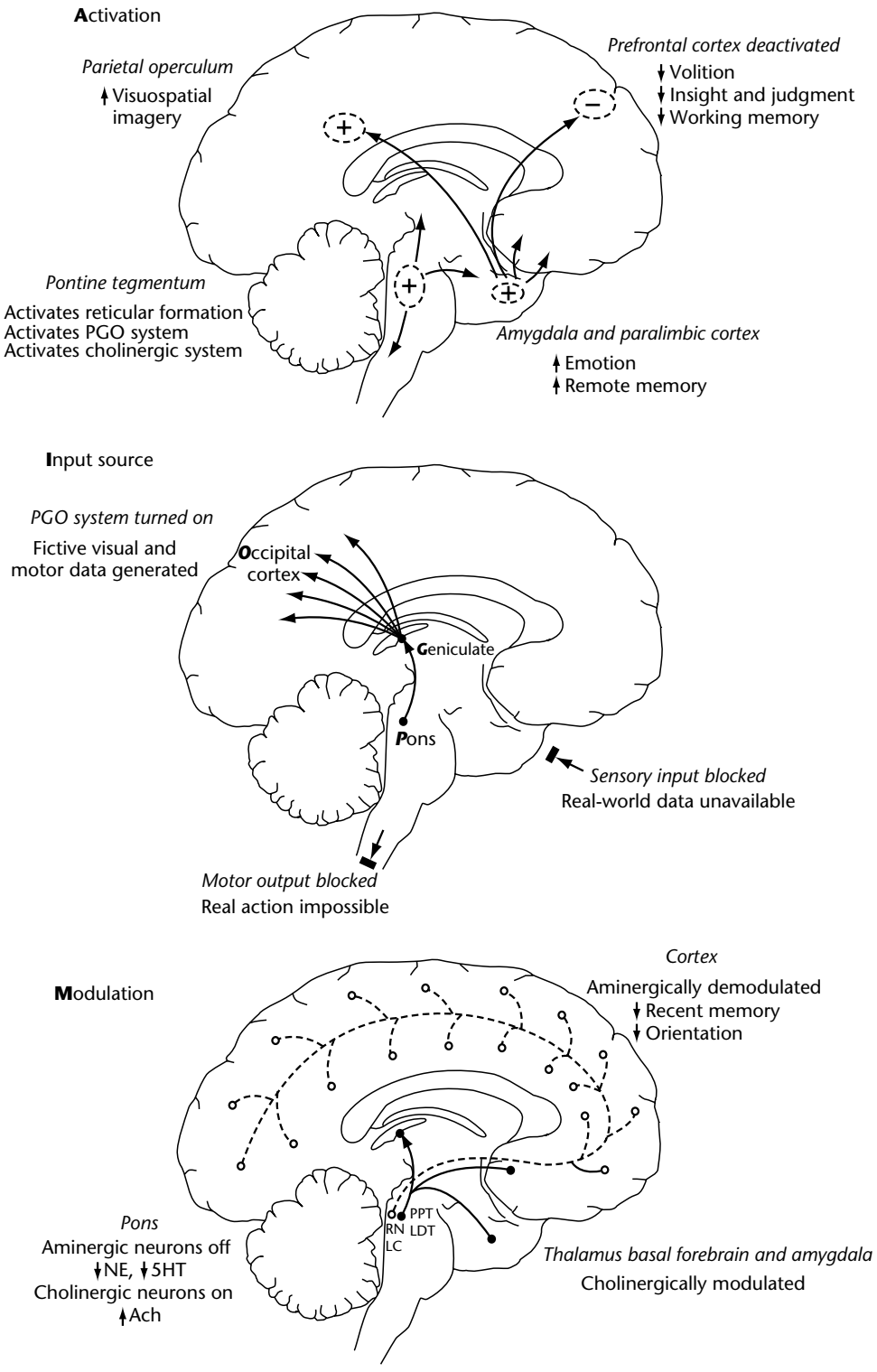


Figure 6. Physiological signs and regional brain mechanisms of REM sleep dreaming separated into the activation (A), input source (I), and modulation (M) functional components of the AIM model. Dynamic changes in A, I, and M during REM sleep dreaming are noted adjacent to each figure. Note that these are highly schematized depictions which illustrate global processes and do not attempt to comprehensively detail all the brain structures and their interactions which may be involved in REM sleep dreaming.

POSITRON EMISSION TOMOGRAPHY IMAGING STUDIES OF REM SLEEP DREAMING

Pierre Maquet and his co-workers at the University of Louvain in Belgium used an $H_2{}^{15}O$ positron source to study REM sleep activation in their subjects, who were subsequently awakened for the solicitation of dream reports (Maquet *et al.*, 1996). They observed a preferential activation of limbic and paralimbic regions of the forebrain in REM sleep (compared to either waking or to NREM sleep). One important implication of this finding is that dream emotion may be a primary shaper of dream plots rather than playing the secondary role in dream plot instigation that was previously hypothesized.

An equally interesting finding, relevant to the cognitive deficits in self-reflective awareness, orientation, and memory during dreaming, is the significant deactivation, in REM, of a vast area of dorsolateral prefrontal cortex. The fact that considerable portions of executive and association cortex are far less active in REM than in waking leads to the idea that in REM sleep there is a specific impairment of executive systems which normally participate in the highest order analysis and integration of neural information. See Figure 5 for results of this study integrated with those of several other recent neuroimaging studies.

LOSS OF DREAMING AFTER CEREBRAL LESIONS

An entirely complementary set of findings and conclusions has been reached following a neuropsychological survey of 332 clinical cases with cerebral lesions by Mark Solms at University College in London, England. The 112 patients who reported a 'global cessation of dreaming' either had damage in the limbic system or the parietal convexity, or suffered disconnections of the mediobasal frontal cortex from the brain stem and diencephalic limbic regions. With respect to the visual imagery aspect, a decrease in the 'vivacity' of dreaming was reported by two patients with damage to the visual associative area in the mediaoccipital-temporal cortex (Solms, 1997).

Taken together, these new neuroimaging and brain lesion studies strongly suggest that the forebrain activation and synthesis processes underlying dreaming are actually very different from those of waking. Not only is REM sleep chemically biased, but the enhanced cholinergic neuromodulation and diminished aminergic modulation are associated with selective activation of the subcortical and cortical limbic structures (which mediate emotion) and with relative inactivation of the frontal cortex (which mediates directed thought). These differences in regional activation obviously force the AIM model to consider still another dimension, that of brain space itself. Figure 6 is an initial mapping of AIM onto patterns of regional brain activation. The regional activation changes could be causally linked to the neuromodulatory dynamics in the following way: those areas which are inactivated in REM are those undergoing aminergic demodulation, while the activated areas are those heavily targeted by cholinergic modulatory neurons.

Whatever the link between the neuromodulatory and regional blood flow data, these findings greatly enrich and inform the integrated picture of REM sleep dreaming as emotion-driven consciousness with deficient memory, disorientation, diminished volition, and impaired analytic thinking. Now that we know that there is a close fit between the animal and human data regarding the mechanism and pattern of brain activation in REM sleep, we are in a much stronger position to strengthen the brain-based theory of dreaming that was first proposed in the early 1980s. And building upon this surprisingly strong base, we can begin to build a general theory of the brain basis of consciousness.

References

Aserinsky E and Kleitman N (1953) Regularly occurring periods of ocular motility and concomitant phenomena during sleep. *Science* **118**: 361–375.

Braun AR, Balkin TJ, Wesensten NJ *et al*. (1997) Regional cerebral blood flow throughout the sleep–wake cycle. *Brain* **120**: 1173–1197.

Dement WC (1958) The occurrence of low voltage fast electroencephalogram patterns during behavioral sleep in the cat. *Electroencephalography and Clinical Neurophysiology* **10**: 291–296.

Dement WC and Kleitman N (1957) Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalography and Clinical Neurophysiology* **9**: 673.

Hobson JA (1988) *The Dreaming Brain*. New York: Basic Books.

Hobson JA (1992) A new model of brain–mind state: activation level, input source, and mode of processing (AIM). In: Antrobus J and Bertini M (eds) *The Neuropsychology of Sleep and Dreaming*, pp. 227–247. Mahwah, NJ: Lawrence Erlbaum.

Hobson JA (1999) *Consciousness*. New York, NY: Scientific American Library. W. H. Freeman Co.

Hobson JA and McCarley RW (1977) The brain as a dream state-state generator: an activation-synthesis

hypothesis of the dream process. *American Journal of Psychiatry* **134**: 1335–1348.
Hobson JA, McCarley RW and Wyzinski PW (1975) Sleep cycle oscillation: reciprocal discharge by two brainstem neuronal groups. *Science* **189**: 55–58.
Hobson JA and Steriade M (1986) Neuronal basis of behavioral state control. In: Mountcastle V and Bloom FE (eds) *Handbook of Physiology: The Nervous System*, vol. 4, pp. 701–823. Washington, DC; American Physiological Society.
Hobson JA, Stickgold R and Pace-Schott EF (1998) The neuropsychology of REM sleep dreaming. *Neuroreport* **9**: R1–R14.
Jouvet M (1962) Récherches sur les structures nerveuses et les mécanismes résponsables des différentes phases du sommeil physiologique. *Archives Italiennes de Biologie* **100**: 125–206.
Maquet P, Peters JM, Aerts J *et al*. (1996) Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature* **383**: 163.
McCarley RW and Hobson JA (1975) Neuronal excitability modulation over the sleep cycle: a structural and mathematical model. *Science* **189**: 58–60.
Moruzzi G and Mogoun HW (1949) Brainstem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* **1**: 455–473.
Nofzinger EA, Mintun MA, Wiseman MB, Kupfer DJ and Moore RY (1997) Forebrain activation in REM sleep: an FDG PET study. *Brain Research* **770**: 192–201.
Solms M (1997) *The Neuropsychology of Dreams: A Clinico-Anatomical Study*. Mahwah, NJ: Lawrence Erlbaum Associates.

Further Reading

Crick F (1995) *The Astonishing Hypothesis*. New York, NY: Touchstone Books.
Damasio A (1995) *Descartes' Error*. New York, NY: Avon Books.
Damasio A (1999) *The Feeling of What Happens*. New York, NY: Harcourt Brace.
Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.
Hobson JA (1999) *Dreaming as Delirium*. Cambridge, MA: MIT Press.
Hobson JA, Pace-Schott E and Stickgold R (2000) Dreaming and the brain: toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences* **23**.

Consciousness, Stream of

Intermediate article

Thomas Natsoulas, University of California, Davis, California, USA

CONTENTS

Except for time-gaps, when the stream comes to a stop and then starts again, a stream of consciousness is constituted, from the Jamesian perspective, of tightly adjacent states of consciousness occurring very briefly one at a time. Unless a second such stream flows in the individual simultaneously, his or her conscious mental life at any instant consists entirely of one integral state of consciousness that typically possesses many intrinsic features.

INTRODUCTION

We commonly speak of someone's being conscious. Also, we often say he or she is conscious *of* something, and we usually specify to some degree the object (or objects) of that consciousness. This object can be an environmental state of affairs or occurrence, the individual himself or herself, or some part of the latter, such as a temporal section of the individual's mental life. Moreover, what one is conscious of can be *merely apparent*: an item, occurrence, or state of affairs having no actual existence, whether present, past, or potentially in the future.

Even if one disbelieves in fire-breathing dragons, it is appropriate to assert, depending on the facts, that someone is undergoing a hallucination and is therein conscious of a fire-breathing dragon. The hallucinatory experiences themselves are certainly real; they are no less real than anything else in the universe. In the example, it is the *object* of the hallucinatory experience that happens to be something unreal, that possesses no kind of existence. Also, hallucinatory experiences may have real objects, such as a long-lost relative. But, of course, the hallucinated presence of real objects in the immediate environment is, with possible exceptions, illusory.

In addition to describing someone as conscious or conscious of something, we may describe certain, or even all, of the mental happenings in an individual as being of the *conscious* kind. Indeed, one sometimes encounters among cognitive scientists the thesis that *all* experiences are conscious in this latter sense. According to this thesis, controversial within cognitive science, to have or undergo any experience is, *ipso facto*, for one to be conscious of it, that is, for it to be an object of 'inner awareness' (Brentano, 1973).

Many cognitive scientists maintain that, in every healthy, intact human being who is functioning normally, there take place both conscious and unconscious mental occurrences. The unconscious mental occurrences are those that cannot be objects of inner awareness; that is, we cannot have first-hand, noninferential apprehension of any instance that transpires of any unconscious mental occurrence. In the case of the conscious mental occurrences, in contrast, all or some of the instances of any one of them are directly apprehended when they occur.

There is scientific disagreement concerning how inner awareness is accomplished. Some cognitive scientists propose that an 'appendage' is required to any mental-occurrence instance that is the object of inner awareness. Thus, inner awareness takes the form of a mental-occurrence instance distinct from the mental-occurrence instance it renders conscious (Natsoulas, 1993b). Other cognitive scientists hold that inner awareness is an *intrinsic* feature of any mental-occurrence instance that is apprehended first-hand. Accordingly, inner awareness is always a dimension of the phenomenological structure of the mental-occurrence instance that is its object (Natsoulas, 1996).

Cognitive scientists are rarely skeptical concerning the reality of conscious mental occurrences as existents in the natural world. They recognize that the pursuit of science itself, among much else in human life and society, *requires* consciousness and necessarily involves the scientist's undergoing mental-occurrence instances with inner awareness. Anyone who may be inclined to cast doubt upon conscious mental occurrences should recall the impossibility of 'mind-blind' science.

Imagine a physical scientist fully capable of perceptual awareness of those molar events in the

environment of interest to his or her science. Now add what is less imaginable: all of the scientist's mental occurrences comprising this perceptual awareness are unconscious. That is, whatever the cause of this condition (of mind-blindness), these mental occurrences cannot be objects of the scientist's inner awareness. A mind-blind scientist could not function as such. It would be, for him or her, as though the environmental events of scientific interest that are objects of his or her perceptual awareness *do not occur*.

Some cognitive scientists attempt to treat of how conscious mental occurrences seem first-hand, to inner awareness, as *illusory* in some respects, although very rarely in respect to their existence. Briefly expressed, here are examples of two such attempts, the second less cogent than the first. (a) Whereas conscious mental occurrences may seem to the person to whom they occur to be mental in the sense of spiritual or nonphysical, they are, in reality, *only occurrences in the brain that*, like the many other kinds of brain occurrences, *possess only physical properties* (Sperry, 1976). (b) Whereas in the large majority of cases, mental-occurrence instances seem to inner awareness to be other than actions or behaviors, they actually are – this includes all of our perceptual experiences and feelings – *forms of self- or other-directed commentary*, either overt, covert, or incipient in any instance (Weiskrantz, 1997). However, cognitive scientists who diverge from the ordinary concept of a conscious mental occurrence do not generally intend thereby to commit a *referential displacement*: that is, to speak of something else, in place of mental occurrences, when they use terms commonly referring to the latter.

Cognitive scientists seldom deny the reality of consciousness, but some argue at length that *unconscious* mental occurrences do not transpire within us at all (e.g. O'Brien and Opie, 1997). In their view, we have brain occurrences that are mental, although not in the spiritual sense; but none of these mental brain occurrences are unconscious in the sense specified above, namely, incapable of being themselves objects of inner awareness. All mental occurrences are *open* to inner awareness, although this does not mean they are actually apprehended on every occasion of their occurrence (Natsoulas, 1995).

The above thesis entails a rejection of Freud's unconscious. According to Freud, an unconscious mental occurrence may 'become' conscious, but conscious only in the sense of evoking a counterpart of itself within a distinct subsystem (called 'perception–consciousness') of the mental apparatus (Natsoulas, 1993a). This counterpart must be of suitable type and instantiate much the same cognitive content as the unconscious mental occurrence. All mental occurrences that take place in the perception–consciousness system are of the conscious variety and, in all instances, are objects of inner awareness. Those mental occurrences that are not conscious transpire in a different subsystem of the mental apparatus in the brain.

The thesis of the nonexistence of unconscious mental occurrences entails, as well, rejection of a large body of contemporary scientific thought. Many hypotheses and theories circulating now inside cognitive science would explain particular instances of behaviors or of conscious mental occurrences by reference to mental happenings to which the person to whom they belong cannot have anything more than inferential access. This is the only kind of access cognitive scientists have to the person's mental life – with the probable exception that, today, some cognitive scientists are applying certain equipment directly to the brain and may be thereby observing mental occurrences by instrument.

The thesis that one's stream of consciousness is the entirety of one's conscious mental life will require modification if people, either normally or under certain conditions, are found to possess more than a single such stream (Puccetti, 1981). Note that the dual-consciousness hypothesis is not equivalent to countenancing the reality of unconscious mental occurrences. One's second conscious stream too would consist of conscious mental occurrences. Admittedly, however, neither of the two streams could have inner awareness of what is transpiring in the other stream and would depend for any information about the latter on inference. For example, the mental occurrences belonging to either stream might ascribe behavior issuing from the one body as causally connected in some instances to a separate consciousness also proceeding in the same body.

HISTORY

Cognitive scientists consider William James's classic contributions to be the most important to date in the history of the concept of the stream of consciousness. If one asks a cognitive scientist about the prevailing concept of the stream, the reply will assuredly contain some mention of James's detailed characterizations. These have a prominent place in *The Principles of Psychology*, in James's famous two-volume masterwork of basic psychology. In many psychologists' opinion, the ninth chapter, 'The Stream of Thought', is an

achievement of great intellectual value and compares favourably to the best analyses of any topic in the scientific psychological literature.

The Principles was published over a century ago and was intended as an introductory textbook for undergraduates. Yet James's phenomenological account of the stream is anything but dated. Indeed, this account would not be out of place if it were included today under 'current work' in a compendium of materials required by anyone seriously interested in consciousness. That James's account of the stream of consciousness is still a central part of our developing scientific thought may be surprising. Its persisting relevance is owed not to cognitive scientists' extending James's empirical observations to any substantial degree. The account itself in its original form still serves to expand the understanding of much about consciousness that lies beyond the merely historical.

This will seem less surprising upon noticing that, soon after James, those who came to control the field of psychology, which James helped found, *refused to acknowledge* the importance of consciousness in psychological functioning. Their refusal to face facts was anti-empirical; their scientific behavior was inconsistent with a true vocation. These academics were motivated by a strong political desire to achieve a broad acceptance for the new psychology as a science among other natural sciences.

To include consciousness as a part of psychology's subject matter would have been, for them, to admit the causal function of something spiritual and to enable thereby religious belief to contaminate their field. The exclusion of consciousness was a joint, systematic, long-term project among people who saw themselves as 'running' the science. And, even with their eventual retirement, their past efforts continued to produce detrimental effects on the science for the remainder of the twentieth century. Beginning in the 1960s, however, it gradually became less difficult to publish scientific articles and books on consciousness and less objectionable to teach the forbidden topic.

The militancy of some psychologists was such that they even denied publicly the existence of consciousness. Others among them, although not expressing such doubts, demanded that all technical concepts be defined in terms only of what could be perceptually observed of the behavior of their experimental or research subjects and the causes of that behavior. Thus, all mental occurrences were to be excluded from playing any role in scientific explanation because they were not in themselves publicly observable. A number of the many psychologists who shared this discredited philosophy of science occupied powerful academic positions and managed to hold back for decades the kinds of research they did not approve of. An indication of their project's success is that, as a whole, James's account of the stream of consciousness has yet to be superseded by a more enlightening account.

THEORETICAL ANALYSES

In describing the stream of consciousness as James conceived of it, one begins traditionally with attention to his metaphor of flowing water. However, a close reading of *The Principles* shows that James's analogy pertains only to how it may well *seem* that one's conscious mental life proceeds. In James's phrase, this life 'does not appear to itself chopped up in bits'. Contrary to some of James's own statements, he did not consistently conceive of the stream of consciousness as a continuous, undivided, internally undemarcated process, albeit subject to stopping and starting. James's more consistent position was, rather, that the stream consists objectively of a succession of *discrete* instances, pulses, or states of consciousness (Natsoulas, 1992–1993).

States of consciousness take place one at a time, and each of them immediately succeeds the one just before it except if there is a 'time-gap' during which the stream goes out of existence sometimes only very briefly. According to James, time-gaps may actually be more numerous than commonly supposed, for they are only inferable, not directly noticeable in his view. However, a temporal section or segment of the conscious mental life that internally involves no time-gaps consists of a succession of pulses of mentality 'with absolutely nothing between'. A state of consciousness lasts very briefly and, thereupon, one's consciousness consists of another such state, and so on. At any moment, one of these states is the entirety of one's consciousness – setting aside the possibility of more than a single stream per person at the same time.

How do the states of consciousness, or basic durational components of the stream, come into existence one after another? Although, according to James, these mental pulses are nonphysical, they are nevertheless the immediate, automatic products of *the total, ongoing brain process*. This said, a major qualification is immediately in order. At different times, the occurrent parts of this completely physical process, which is proceeding in many brain structures, differ in their intensities

and, consequently, are variably determinative of intrinsic features of the pulses that the total process produces.

The stream of consciousness is an accretion of basic durational components, rather than its being an ongoing process of its own whose course is merely influenced by the brain process. In contrast, the latter description is indeed applicable to how pulses of mentality are supposed to affect the total brain process. James's mind–body position expressed in *The Principles* was a dualist interactionism, but the existence of consciousness, of each one of those mental pulses constituting it, was held to depend on the brain. Mind issues from the body rather than the body's having effects on mind. Merely by its occurrence, the total brain process produces as a by-product something new at every point, whereas the pulses of mentality have their direct effects on the brain process as it is already taking place. They can only hinder or further this ongoing process.

Calling the stream's basic durational components 'pulses' of mentality, as James does, may convey the false impression that each of the pulses is proposed to be simple. This is belied by the fact that, soon enough after *The Principles*, James (1899) spoke of the stream of consciousness as consisting of a succession of 'fields' and insisted that these fields are always complex. He stated specifically that most of our concrete states of consciousness contain individually, in different proportions but in some positive degree, all of the following ingredients: sensations from the body, sensations owed to the impact of energies upon sense receptors that are outwardly directed, remembrances of past experiences, thoughts about faraway things, feelings of satisfaction and dissatisfaction, acts of will, desires and aversions, and other emotional conditions. Some of these ingredients are instantiated by a particular state more than by others and seem first-hand to have greater prominence in the state. Thus, although possessing many actual ingredients, one state of consciousness will seem to consist largely of sensation, another largely of remembrance, and so on. James pointed out that, for practical reasons, we tend to classify certain states of consciousness together, calling them states of emotion, sensation, abstract thought, volition, and the like, but a state of consciousness is not equivalent to any of the ingredients mentioned. Rather, these are to be understood as among the intrinsic features of a state, not as parts of it.

Speaking of an image and its 'surrounding' or 'penumbral' content, which is the source of the image's meaning, as though they were two distinct parts belonging to a particular state of consciousness, James in *The Principles* qualified his spatial metaphor in midstream: he described the surround as 'bone of the image's bone and flesh of its flesh'. If the penumbral contents of a state of consciousness had been, in any instance, different, any object of the state would have been taken and understood otherwise. This is not because the penumbral contents are among the causal determinants of a state, but because they are ingredients of the whole unitary, integral state and dimensions of how, specifically, the state apprehends its object or objects.

The ingredients of a single state, pulse, or field of consciousness listed above are not mutually distinct instances of being conscious of something. Such ingredients of a state as sensations, memories, thoughts, feelings, desires, aversions, emotions, and conations are not as they are traditionally considered: separate mental acts. Each state of consciousness is a unitary instance of consciousness. Any object of such a state that may be considered focal is apprehended in relation to all the other objects of the state. The above ingredients are abstractions from the concrete state that contains them. They are features of how the state apprehends its multiple objects together. This makes the ingredients of states no less real, but it does imply that they have no existence outside the states of which they are intrinsic features. For example, no auditory experience exists that is not a feature of one or more states of consciousness.

James found support for his conception of the states of consciousness in an unlikely source: Wilhelm Wundt, the founder of experimental psychology. Wundt assessed his three decades of introspective laboratory work as follows: 'From my inquiry into time-relations, etc. … I attained an insight into the close union of all those psychic functions usually separated by artificial abstractions and names, such as ideation, feeling, will; and I saw the indivisibility and inner homogeneity, in all its phases, of the mental life' (translated and quoted by James, 1899, p. 21). James interpreted the passage in which this sentence appears as a total renunciation of the prevailing conception of the mental life as being made up of 'distinct processes and compartments'.

But James did not also characterize in such terms the neurophysiological source of the stream of consciousness. Although states of consciousness are integral and not compounded of smaller units or acts, the total brain process, which is responsible for the existence of states of consciousness, clearly consists of numerous distinct processes. Indeed,

different sets of brain processes were held by James to be the proximate causes of different states of consciousness. The part-processes that constitute the total brain process of the moment somehow combine together to produce a unitary mental state.

EXPERIMENTAL APPROACHES

Introspection was the empirical ground for James's fundamental conception of consciousness. This conception included: (a) that the whole of one's mental life consists of a stream of consciousness, except insofar as a second stream of consciousness also flows within one; (b) that any stream of consciousness is constituted entirely of a sequence of tightly adjacent states, or pulses, of consciousness, except that the stream does stop and start again, and perhaps very frequently; (c) that each state of consciousness is of an integral character, a unitary awareness albeit often with many objects; and (d) that each such state typically instantiates a complexity of intrinsic features, some of which are traditionally conceived of, wrongly, as distinct mental acts.

The activity of introspecting can be carried out more or less adequately. James sometimes argued against views of consciousness with which he disagreed that these views have their basis in careless or biased introspection. But, in a chapter of *The Principles* devoted to methodology, James stressed that psychologists *always have to rely* on introspection, and that the most basic postulate of psychological science is that people do have inner awareness wherein they distinguish a state of consciousness from what it is about.

When James considered the matter more deeply, he became skeptical concerning our having the ability to apprehend our states of consciousness in a first-hand way. However, this skepticism did not deter him from proceeding on the assumption that the ability to introspect is among our most valuable powers. Indeed, much of what James is known for in psychology involves phenomenological descriptions made possible by inner awareness of his states of consciousness. James came to his skeptical position – which he quickly set aside for evidently practical reasons – as a consequence of introspecting and discovering thereby that inner awareness fails to reveal any spiritual activity at all going on within him.

Before moving on, James contemplated that, in point of fact, we may not have a stream of consciousness as he had been describing it. Better to say, perhaps, that the mental life that we do possess is a stream of 'sciousness' since its components are never objects of inner awareness (cf. Hebb, 1982). Our states of sciousness always have something else as their objects, that is, parts of the environment and body, except insofar as we infer the presence within us of those states from other, observable matters. Thus, states of sciousness may be among the non-immediate objects of some of our states of sciousness.

Although there are inconsistencies in James's reasoning, the only objection to his skeptical view requiring mention here pertains to his implicit notion that one can draw inferences – from perceptual observations to the occurrence of states of sciousness within one – in the complete absence of inner awareness. On James's skeptical view, perceptual awareness transpires not in a stream of consciousness but in a stream of sciousness every one of whose basic durational components are unconscious. The entirety of our mental life is proposed to take place in the 'dark' as we, at most, guess or infer about it. But how can one do any inferring based on the occurrence of a perceptual awareness of which one cannot be aware first-hand? If it is replied that we can know of the perceptual awareness by inference from something we objectively observe, the question is, again, how we can so infer if we cannot have inner awareness of the latter observation either.

More consistently with the skeptical position, which James seems to have preferred, he might have adopted a different methodology than the introspective one that he continued to practice and on whose results he relied for the rest of *The Principles*. This nonintrospective methodology would have him observing his behaviors and their objective context and making inferences about his mental life from what he observed. This is the kind of procedure on which present-day cognitive scientists often rely in studying the mental life of their research subjects, along with putting the latest instruments to use to detect properties of those of their subjects' brain processes that the scientists believe are the mental occurrences of current interest or closely associated with those mental occurrences. Of course, introspection helps them to formulate hypotheses to test experimentally, but they are very restrained in justifying their claims by reference to what they know first-hand by inner awareness.

However, even relying on perceptual observations exclusively – observations of behavior, of its objective context, and (by instrument) of the brain processes that are transpiring at the time – would not mean the cognitive scientist has succeeded in bypassing inner awareness so that inner awareness

forms no part of his or her methodology. The mental processes of a scientist are an essential dimension of any objective methodology; surely, this is an unobjectionable statement. The cognitive scientist cannot proceed to study the states of consciousness of his or her research subjects (or to study anything else for that matter) unless the scientist's mental processes continue to include more than just unconscious mental occurrences. At least some of the scientist's mental states during the experimental researches must be objects of inner awareness; otherwise, the objective observations would be for naught. Although these observations may have unconscious effects upon the researcher, in the absence of all inner awareness of the observations, they cannot be put to use.

In their investigation of states of consciousness, cognitive scientists cannot limit their empirical database to their own introspections. They must seek to determine the properties of states of consciousness more generally, how these states are in other people. For this purpose, cognitive scientists carry out not only programs of objective observation, with some reference to how they find their own states of consciousness to be first-hand. Also, they secure self-reports from experimental subjects concerning their inner awareness of their own respective conscious states, notwithstanding the fact that the causation responsible for self-reports concerning consciousness is not well understood at the present time. Many cognitive scientists remain unwilling to put much evidential weight on such reports, not until cognitive science advances to the point where we know the conditions under which subjects' reports are to be trusted. However, it is important to realize that inference from objective observations to states of consciousness is not any less problematic. The validity of such inference depends on the understanding that we have of the causal relations between these observations and the conscious states that are among the causes of what we are observing about our research subjects.

References

Brentano F (1973) *Psychology from an Empirical Standpoint*. London: Routledge and Kegan Paul. [Original German corresponding edition published in 1911.]

Hebb DO (1982) Elaborations of Hebb's cell assembly theory. In: Orbach J (ed.) *Neuropsychology after Lashley*, pp. 483–496. Hillsdale, NJ: Lawrence Erlbaum.

James W (1899) *Talks to Teachers on Psychology: And to Students on Some of Life's Ideals*. New York, NY: Holt.

Natsoulas T (1992–1993) The stream of consciousness: I. William James's pulses. *Imagination, Cognition and Personality* **12**: 3–21.

Natsoulas T (1993a) Freud and consciousness: VII. Dimensions of an alternative interpretation. *Psychoanalysis and Contemporary Thought* **16**: 67–101.

Natsoulas T (1993b) What is wrong with appendage theory of consciousness. *Philosophical Psychology* **6**: 137–154.

Natsoulas T (1995) A rediscovery of consciousness. *Consciousness and Cognition* **4**: 223–245.

Natsoulas T (1996) The case for intrinsic theory: I. Introduction. *Journal of Mind and Behavior* **17**: 267–286.

O'Brien G and Opie J (1997) Cognitive science and phenomenal consciousness: a dilemma and how to avoid it. *Philosophical Psychology* **10**: 269–286.

Puccetti R (1981) The case for mental duality: evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* **4**: 93–123.

Sperry RW (1976) Mental phenomena as causal determinants of brain function. In: Globus GG, Maxwell G and Savodnik I (eds) *Consciousness and the Brain*, pp. 163–177. New York, NY: Plenum.

Weiskrantz L (1997) *Consciousness Lost and Found: A Neurophysiological Exploration*. Oxford, UK: Oxford University Press.

Further Reading

Armstrong DM and Malcolm N (1984) *Consciousness and Causality: A Debate on the Nature of Mind*. Oxford, UK: Blackwell.

Dulany DE (1997) Consciousness in the explicit (deliberative) and implicit (evocative). In: Cohen JD and Schooler JW (eds) *Scientific Approaches to Consciousness*, pp. 179–212. Mahwah, NJ: Lawrence Erlbaum.

Freud S (1961) The Ego and the Id. In: *Standard Edition*, vol. 19, pp. 12–66. London: Hogarth. [Original work published 1923.]

James W (1950) *The Principles of Psychology*, 2 vols. New York, NY: Dover. [Originally published in 1890.]

Natsoulas T (1977) Consciousness: consideration of an inferential hypothesis. *Journal for the Theory of Social Behaviour* **7**: 29–39.

Natsoulas T (1981) Basic problems of consciousness. *Journal of Personality and Social Psychology* **41**: 132–178.

Natsoulas T (1984) Gustav Bergmann's psychophysiological parallelism. *Behaviorism* **12**: 41–69.

Natsoulas T (1997) Blindsight and consciousness. *American Journal of Psychology* **110**: 1–34.

Natsoulas T (1998) On the intrinsic nature of states of consciousness: James's ubiquitous feeling aspect. *Review of General Psychology* **2**: 123–152.

Searle JR (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Woodruff Smith D (1989) *The Circle of Acquaintance: Perception, Consciousness, and Empathy*. Dordrecht, Netherlands: Kluwer.

Consciousness, Unity of

Intermediate article

Tim Bayne, Macquarie University, Sydney, Australia

CONTENTS

Consciousness is unified in various ways, but the nature and basis of its unity remains a matter of intense debate. This debate promises to illuminate the nature of experience, the self, and various pathologies of consciousness.

INTRODUCTION

At any particular point in time you might be enjoying a number of experiences. For instance, you might have the visual experience of seeing these words, the cognitive experience of understanding what these words mean, and bodily sensations of various kinds, such as the feeling of the ground beneath one's feet. These experiences are unified in various ways. In exactly what ways? How can these unity relations be explained and how are they related to each other? In what sense must experiences be unified? What implications might the unity of consciousness have for our understanding of the nature of consciousness and the nature of the self? These are some of the central questions raised by the unity of consciousness.

VARIETIES OF UNITY WITHIN CONSCIOUSNESS

There is no commonly accepted taxonomy for discussions of the unity of consciousness. The following taxonomy is just one path through what is a rather unstructured debate, and the reader is advised that some of the terms introduced below are used in very different ways by other authors.

Subject Unity

Experiences are *subject-unified* when they are had by the same subject of experience. Your experiences are subject-unified in so far as they are yours, while my experiences are subject-unified in so far as they are mine. It is important to distinguish subject unity from the unity of self-consciousness. Subject-unified experiences are had by a single subject of experience, but the subject in question need not be conscious of these experiences as his or her own; many creatures appear to be subjects of experience without being aware of themselves as subjects of experience. Although most mature human beings can – and often do – ascribe their experiences to themselves, this ability seems to be lost, or at least compromised, in certain pathologies of consciousness, such as depersonalization.

Object Unity

A second unity relation in consciousness is *object unity*. Experiences are object-unified when they are about the same object. When I see my dog barking at my neighbor's cat I unify my visual experience of my dog with my auditory experience of him, thus forming a multimodal percept of a single object. Object unity not only extends across perceptual modalities; it also binds perception and action together, as when one when reaches for something that one can see.

Spatial Unity

The intentional content of experience gives rise to a third unity relation, *spatial unity*. When I see my dog chasing the neighbor's cat I experience the dog and the cat as distinct but spatially related objects. Objects of perception are located in egocentric space – 'egocentric' because the structure of this space is given by the structure of one's own body. The experience of one's own body as a unitary physical object enables one to experience one's perceptual environment as a spatially unified domain.

Epistemic Unity

Sometimes a stream of consciousness is said to be unified to the degree that its contents are coherent, integrated, or comprehensive (Shoemaker, 1996).

We might call this form of unity *epistemic unity*. Shoemaker has epistemic unity in mind when he writes that 'perfect unity of consciousness … would consist of a unified representation of the world accompanied by a unified representation of that representation, the latter including not only information about what the former represents, but also information about the grounds on which the beliefs that make up the former are based, and about what the evidential relations between the parts of that representation are' (1996, p. 186). In a related use of the term, the 'unity of consciousness' is also used to refer to the consistency of consciousness. Although one can see the famous duck–rabbit as either a duck or a rabbit, one cannot simultaneously experience it as both – although priming experiments reveal that both interpretations can be simultaneously active (Baars, 1988).

Phenomenal Unity

There is a unity of consciousness that is arguably more primitive than any of the unity relations considered thus far. Consider again the set of experiences that you are currently enjoying: perceptual experiences, cognitive experiences, emotional experiences, and so on. All of these experiences seem to be contained within a single phenomenal field or stream of consciousness. They seem to be conscious together; they seem to be *phenomenally unified*. It has become common to talk about phenomenal unity in terms of the relation of *co-consciousness*. Experiences are co-conscious when they are experienced together: when they have a conjoint phenomenology. My current visual experiences are co-conscious with my current auditory experiences, but they are not co-conscious with your auditory experiences. (Note that psychologists often describe distinct streams of consciousness that concurrently belong to a single organism as 'co-conscious': a very different use of the term.)

Access Unity

A final type of unity of consciousness concerns the relation between consciousness and reportability or accessability. We can describe experiences that are reportable in exactly the same ways as being *access-unified*. My current auditory and visual experiences seem to be access-unified in that I am able to report my auditory experiences in any way in which I am able to report my visual experiences. Note that experiences can be access-unified without being co-reportable; that is, I may be able to verbally report either my auditory experience or my visual experience without being able to report both experiences together.

How are these various forms of unity related? In particular, how are subject unity, phenomenal unity, and access unity related? It is plausible to suppose that experiences can be access-unified or phenomenally unified only if they are had by the same subject of experience. The controversial questions are whether (simultaneous) co-subjective experiences must be phenomenally unified, and whether phenomenally unified experiences must be access-unified. The study of various pathologies of consciousness might help to answer these questions.

UNITY OF CONSCIOUSNESS IN PHILOSOPHY

History

Philosophical reflection on the unity of consciousness dates back at least as far as Aristotle, who argued that there must be a common sense in addition to the specific senses such as sight and hearing. This common sense was thought to be responsible for perceiving the common sensibles – properties such as motion, rest and number – that are perceivable by multiple sense modalities. It was also assigned the role of integrating the contents of the other senses.

In the seventeenth century, philosophical interest in the unity of consciousness shifted focus from what the unity of consciousness might tell us about the structure of the mind to what it might tell us about the ultimate nature of the mind. G. W. Leibniz argued that the unity of consciousness provides support for substance dualism, the view that the subject of experience is an immaterial substance. He argued that if the self is a material entity it must be spatially extended, and hence the different parts of an experience would be had by different subjects. Consequently, there would be no single subject that had the entire experience, as there obviously seems to be. (A similar argument for substance dualism appears in René Descartes' *Meditations*.) Contemporary functionalists respond to this argument by insisting that the subject of an experience is the entire system within which the experience occurs: although a functional system has parts, none of these parts is a subject of experience in its own right.

Another influential seventeenth-century discussion of the unity of consciousness can be found in John Locke's analysis of personal identity. Locke claimed that the identity of a person extends only so far as their consciousness extends. Although

Locke draws an intimate connection between personal identity and the unity (and continuity) of consciousness, he says very little about what the unity of consciousness involves. Most work on Locke's accounts of personal identity has focused on the connection between personal identity through time and the continuity of consciousness; it is only recently that philosophers have begun to explore the relationship between the identity of a person at a time and the unity of consciousness.

Although the notion played little role in his treatment of the unity of consciousness, Locke held that experiences are had by a substantial self. David Hume, in contrast, was positively hostile to this idea. He claimed that since introspection reveals no sign of the self there is no good reason to posit such an entity. Hume claimed that the self is not something distinct from its experiences, something that has experiences, but is, in fact, nothing but a bundle of experiences. In response to Hume, Kant argued that there must be a transcendental ego that is responsible for synthesizing various representations into a unified consciousness. Kant's discussion of the unity of consciousness is notoriously obscure, and there is disagreement both over what Kant meant by the 'unity of consciousness' and over what he explained the unity (or unities) of consciousness in terms of (Brook, 1994; Hurley, 1998). Kant is variously said to have identified the unity of consciousness with, or explained it in terms of: the unity of concepts, the unity of agency, the unity of self-consciousness, and the unity of an objective world. (*See* **Split Brains, Philosophical Issues about**)

Current Issues

Current philosophical interest in the unity of consciousness is particularly concerned with two major questions.

Is consciousness unified?

The claim that consciousness is unified can be taken in a number of different ways. Some take the claim to mean that there is a single anatomical module or site for consciousness. Others take it to mean that co-conscious experiences are always intentionally integrated; that is, epistemically unified. It is doubtful that consciousness must be unified in either of these ways.

A more plausible sense in which consciousness might be unified is that co-conscious experiences must also be access-unified. We can call this the *accessibility thesis*. The accessibility thesis is attractive; indeed, it seems to be presupposed by certain accounts of consciousness (Baars, 1988). But there is reason to think that it might be false. Certain experimental results and pathologies of consciousness (discussed later in this article) suggest that experiences can be phenomenally unified without being accessible to the same report modalities.

Another conception of the unity of consciousness concerns the relationship between subject unity and phenomenal unity. According to the *unity thesis*, (simultaneous) co-subjective states must be co-conscious; that is, any experiences that a subject has at a time must be contained within a single fully unified stream of consciousness. The plausibility of the unity thesis depends in part on one's account of the subject of experience. The unity thesis seems implausible if, as many philosophers argue, subjects of experience should be individuated in biological terms, for it seems quite possible that a single organism might have two experiences at the same time without those experiences being co-conscious. On the other hand, one could also take the plausibility of the unity thesis as an argument against equating the self with a biological organism.

Finally, we can take the claim that consciousness is unified as a claim about the structure of co-consciousness. It is natural to suppose that the relation 'is co-conscious and simultaneous with' is reflexive, symmetrical, and transitive. (A relation R is transitive if, whenever aRb and bRc, then aRc.) Developing an idea that was inchoate in Nagel (1971), Lockwood (1989) suggested that synchronic co-consciousness might not be transitive, and that as a result some streams of consciousness may be only partially unified. If synchronic co-consciousness is not transitive then it would be possible to have three simultaneous experiences, e_1, e_2, and e_3, such that e_1 and e_2 are co-conscious, and e_2 and e_3 are co-conscious, but e_1 and e_3 are not co-conscious. We can call the claim that synchronic co-consciousness is transitive the *transitivity thesis*.

Philosophical debate over the transitivity thesis has focused on the question of whether partial unity might be possible. Most philosophers admit that it is difficult to conceive of what it would be like to have a partially unified consciousness, but they differ over what this might show. Dainton (2000) defends the transitivity thesis on the basis of the inconceivability of partial unity, while Hurley (2002) argues that this line of argument is mistaken. She claims that the 'what it's like' approach is relevant only when it comes to determining the content-based relations between experiences, and since the transitivity thesis concerns the token identity of states neither it nor its denial

could be supported by appeal to phenomenological considerations.

Philosophers have also been interested in whether various pathologies of consciousness, such as those exhibited by split-brain patients, might yield counterexamples to the unity or transitivity theses (see below).

How can phenomenal unity be explained?

There are many accounts of phenomenal unity (or co-consciousness), but no standard way of categorizing them. One useful distinction is between theories that account for phenomenal unity in terms of factors internal to phenomenology (subjective theories) and theories that appeal to factors outside phenomenology (objective theories).

Some subjective theories appeal to introspection in order to explain phenomenal unity. Such appeals can take at least two forms, depending on the account of introspection offered. Introspection can be conceived of as a nonrepresentational act of awareness that unifies various experiences, or it can be thought of as an experience in its own right, a higher-order experience of first-order experiences.

An alternative subjectivist approach is to posit a primitive unity relation for consciousness. Dainton (2000) takes this approach, holding that co-consciousness itself is a primitive unity relation. Bayne and Chalmers (2002) adopt a similar approach, although they take subsumption as their primitive unity relation. (One experience subsumes another when, roughly, the subsuming experience entails the subsumed experience). Dainton's approach is 'bottom-up' – fully unified streams of consciousness are constructed out of particular experiences and relations of mutual co-consciousness – while Bayne and Chalmers take a 'top-down' approach, according to which particular experiences are phenomenally unified by virtue of being subsumed by a single total experience.

Current versions of objectivism tend to be functional in nature. Shoemaker's (1996, 2002) functionalism is a standard personal-level functionalism involving causal relations between content-bearing states, and is in part motivated by his functionalist account of consciousness. Hurley (1998) defends a sub-personal functionalism according to which the unity of consciousness involves a dynamic singularity centered on an active organism.

A number of approaches to the unity of consciousness can be developed along both objectivist and subjectivist lines. Consider, for instance, subject-based accounts of co-consciousness. It is sometimes suggested that experiences are co-conscious simply by virtue of being possessed by the same subject of experience at the same time. This account of phenomenal unity qualifies as a version of objectivism. However, an account which construes co-consciousness in terms of the phenomenology of self-consciousness – that is, in terms of the sense that certain experiences belong to oneself – is a version of subjectivism, for it locates the binding agent of consciousness within phenomenology itself.

It is important to note that objectivism and subjectivism are not mutually exclusive. Indeed, it seems plausible to suppose that a complete account of phenomenal unity will have both subjective and objective components.

UNITY OF CONSCIOUSNESS IN COGNITIVE SCIENCE

History

Cognitive science has had an interest in the unity of consciousness since its beginnings in the nineteenth century. A central strand of this concern has been to examine the conditions under which the human brain achieves perceptual consistency. In this regard, psychologists have devoted much attention to effects such as the phi phenomenon. Spots separated by a small visual angle (up to 4 degrees) that are briefly lit in rapid succession will produce the phenomenal effect of a single spot in motion. The brain assumes that it must be seeing a single source of light in rapid motion rather than two light sources lit in rapid succession. Closely related to the phenomenon of perceptual constancy are intermodal effects, such as the McGurk effect (Stein and Meredith, 1983). When the sound 'ba-ba' is dubbed onto the video of someone who is actually saying 'ga-ga' participants report hearing 'da-da'. Such intermodal effects are common, and reveal that the operations of the various modalities constrain each other in what are often surprising ways.

Object and spatial unity reveal that the brain binds disparate forms of information together. How does it do this? Many early models of integration attributed binding to specialized neurons – so-called 'grandmother cells' – and multimodal association areas. Although some theorists continue to accord such cells a role in binding, it is generally agreed that 'convergence' models are at best a partial solution to the binding problem. At a theoretical level it is hard to see how there could be a particular cell for each possible object of perception, while at a practical level the search for omnimodal cells and association areas has proved

fruitless. Although there are many multimodal areas in the brain, there seems to be no one anatomical site of consciousness on which all information must converge in order to become conscious. Contemporary approaches to integration generally focus on functional integration rather than anatomical convergence.

While neurophysiologists have explored the unity of consciousness via the mechanisms of neuronal interaction, clinical psychologists have examined the unity of consciousness by studying the effects of brain damage and psychopathology. One class of such disorders can be loosely grouped under the heading of 'dissociative disorders': these include the 'hidden observer' effect in hypnosis, fugue states, and dissociative identity disorder (formerly 'multiple personality disorder') (Hilgard, 1986). It is often claimed that dissociative disorders reveal that consciousness is not unified, although the precise meaning of this claim is often unclear. Dissociation clearly involves a lack of integration in the contents of consciousness, but it is an open question whether this disintegration is accompanied by, or results in, multiple (or partially unified) streams of consciousness within the dissociated individual.

Split-brain research raises many of the same issues as do the dissociative disorders, but it has had a far greater impact on discussion of the unity of consciousness within cognitive science. In the mid-1960s surgeons sectioned the corpus callosum, the bundle of fibers linking the two cerebral hemispheres, of a number of patients in an attempt to alleviate their epilepsy. Although the everyday behavior of these patients was largely unaffected by the operation, under certain laboratory conditions these 'split-brain' patients behaved in ways that suggested to many, notably Sperry (1984), that they had two streams of consciousness. If, for instance, 'key-ring' is briefly projected onto the patient's visual field, the patient claims to see only 'ring', but when asked to pick out the object seen from a range of items, the left hand settles on a key (and rejects a ring). The standard explanation for this behavior proceeds as follows. Information concerning the left side of the visual field is conveyed to the right hemisphere of the brain, and vice versa. Further, in most patients linguistic ability is localized in the left hemisphere, and control over each hand is subserved by the contralateral hemisphere. Thus, it is often suggested that the right hemisphere locates a key with the left hand because it is aware only of 'key', while the patient claims only to have seen the word 'ring' because the left hemisphere was conscious only of the information contained in the right visual field. In other words, it has often been claimed that the best explanation of the patient's behavior is to regard each of the hemispheres as independently conscious. (*See* **Split Brains, Philosophical Issues about**)

Other interpretations of the split-brain data are possible. For instance, it might be suggested that at least some split-brain patients have a partially unified consciousness. For one thing, not all split-brain patients are equally split (Moor, 1982). Patients with complete commissurotomies are both tactually and visually split, while patients with central commissurotomies usually are tactually split but not visually split. Furthermore, even patients with complete commissurotomies are not completely split: they retain the ability to integrate certain types of information (e.g., olfactory, proprioceptive, and emotional) between hemispheres.

It is also noteworthy that the degree of disunity that split-brain patients manifest depends on how they are tested. Consider, for instance, the block design task. In the standard task the patient is asked to manually arrange four patterned cubes to match a sample design. The performance of each hand is timed, and the left hand consistently constructs the design much faster than the right hand. Gazzaniga and LeDoux (1978) found that replacing the standard free manipulospatial response with a response in which the patient was visually presented with three possible answers and asked to point to the correct one removed the asymmetry between left-hand and right-hand responses.

Current Issues

Much current research on the unity of consciousness in cognitive science is concerned with two major questions.

What are the mechanisms of integration and binding?

Although there is intensive discussion in cognitive science about the 'binding problem', there is little agreement about how best to approach it, or even what it is (Revonsuo, 1999). A number of binding problems can be distinguished, of which we will consider two. Both problems begin with the generally accepted claim that the mechanisms of consciousness are widely distributed throughout the brain.

The *object binding* problem is this: given that various visual features – colour, shape, location, etc. – are processed in various parts of the brain, how does the brain bind this information together to form a unified visual percept of a single object?

When we look at a Swiss flag, why is that we see a white cross on a red background, rather than a red cross on a white background? The *global binding* problem is concerned with the unity of the entire field (or stream) of consciousness, rather than the unity of individual phenomenal objects. How are various experiences brought together into a single phenomenal field; by what means are they made mutually co-conscious?

Discussions of binding tend to restrict themselves to object binding (and usually to object binding in visual consciousness). As yet there is no commonly agreed approach to object binding. Some (e.g. Barlow, 1995) argue that single neurons play a significant role in object binding; others (e.g., Triesman, 1999) suggest that object binding might be subserved by selective spatial attention; still others argue that synchronized neural activity plays a role in integrating various features into unitary percepts (Crick and Koch, 1990; Engel *et al*., 1999). It is unclear what implications these accounts of object binding might have for the global binding problem. Proponents of the neurophysiological approach sometimes suggest that neural synchrony might hold the key to all forms of phenomenal unity – indeed, to phenomenology itself – but such suggestions are little more than speculation at present.

It is often suggested that there is an intimate connection between the unity of consciousness and the computational architecture of consciousness. Some claim that because consciousness is unified it must be implemented in a serial or 'von Neumannesque' manner, while others claim that since consciousness is implemented in a parallel connectionist network it cannot be unified. Neither inference is unproblematic: there is no obvious reason why unity at a phenomenal level entails seriality in computational structure.

When and how is the unity of consciousness disrupted?

In addition to dissociative disorders and commissurotomy, a number of other 'pathologies' of consciousness seem to involve some form of disunity in consciousness, although it is often difficult to know exactly how consciousness is disunified in these cases.

Patients with anosagnosia fail to fully appreciate that they have an impairment of some kind. For instance, a patient with a paralysed limb may verbally deny that there is anything wrong. Nevertheless, anosagnosics may behave in ways that indicate that they may have 'dim knowledge' of their condition. They may, for example, agree that if the physician had the same complaint he or she would be unable to get out of bed (Bisiach and Berti, 1995).

Marcel (1993) elicited similar dissociations in response from normal subjects. He asked subjects to respond to the onset of a light in three ways: by blinking, by pushing a button, and by saying 'yes'. Marcel discovered that when subjects were asked to give all three responses simultaneously they were often inconsistent. For instance, a subject might give an affirmative response by blinking and saying 'yes' but a negative response by not pressing the button. Even more surprising, subjects were unaware that they had responded inconsistently. Marcel's subjects clearly suffered from a certain disunity of consciousness, but it is an open question how best to characterize this disunity.

References

Baars BJ (1988) *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.

Barlow H (1995) The neuron doctrine in perception. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Bayne T and Chalmers D (2002) What is the unity of consciousness? In: Cleeremans A (ed.) *The Unity of Consciousness: Binding, Integration, Dissociation*. Oxford, UK: Oxford University Press.

Bisiach E and Berti A (1995) Consciousness in dyschiria. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*, pp. 1331–1340. Cambridge, MA: MIT Press.

Brook A (1994) *Kant and the Mind*. Cambridge, UK: Cambridge University Press.

Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2**: 263–275.

Dainton B (2000) *Stream of Consciousness: Unity and Continuity in Experience*. London, UK: Routledge.

Engel AK, Fried P, König P, Brecht M and Singer W (1999) Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* **8**: 128–151.

Gazzaniga M and LeDoux J (1978) *The Integrated Mind*. New York, NY: Plenum Press.

Hilgard E (1986) *Divided Consciousness*. New York, NY: John Wiley. [Expanded edition.]

Hurley S (1998) *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Hurley S (2002) Action, the unity of consciousness, and vehicle externalism. In: Cleeremans A (ed.) *The Unity of Consciousness: Binding, Integration, Dissociation*. Oxford, UK: Oxford University Press.

Lockwood M (1989) *Mind, Brain and the Quantum*. Oxford, UK: Blackwell.

Marcel A (1993) Slippage in the unity of consciousness. In: Bock GR and Marsh J (eds) *Experimental and Theoretical Studies of Consciousness*, pp. 168–179. Chichester, UK: John Wiley.

Moor J (1982) Split brains and atomic persons. *Philosophy of Science* **49**: 91–106.
Nagel T (1971) Brain bisection and the unity of consciousness. *Synthese* **22**: 396–413.
Revonsuo A (1999) Binding and the phenomenal unity of consciousness. *Consciousness and Cognition* **8**: 173–185.
Rosenthal DM (2002) Persons, minds, and consciousness. In: Hahn LE (ed.) *The Philosophy of Marjorie Grene,* in the Library of Living Philosophers. La Salle, IL: Open Court.
Shoemaker S (1996) Unity of consciousness and consciousness of unity. In: Shoemaker S *The First-Person Perspective and Other Essays*. Cambridge, UK: Cambridge University Press.
Shoemaker S (2002) Consciousness and co-consciousness. In: Cleeremans A (ed.) *The Unity of Consciousness: Binding, Integration, Dissociation*. Oxford, UK: Oxford University Press.
Sperry R (1984) Consciousness, personal identity and the divided brain. *Neuropsychologia* **22**: 661–673.
Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.
Treisman A (1999) Feature binding, attention and object perception. In: Humphreys GW, Duncan J and Treisman A (eds) *Attention, Space and Action*, pp. 91–111. New York, NY: Oxford University Press.

Further Reading

Bertelson P (1998) Starting from the ventriloquist: the perception of multimodal events. In: Michel S and Craik F (eds) *Advances in Psychological Science*, vol. II, *Biological and Cognitive Aspects*, pp. 419–439. Hove, UK: Psychology Press.
Braude S (1995) *First-Person Plural*. Lanham, MD: Rowman and Littlefield.
Hardcastle VG (1994) Psychology's binding problem and possible neurobiological solutions. *Journal of Consciousness Studies* **1**: 66–90.
Hill CS (1991) Unity of consciousness, other minds, and phenomenal space. In: *Sensations: A Defense of Type Materialism*. Cambridge, UK: Cambridge University Press.
James W (1981) The stream of thought. In: *The Principles of Psychology*, vol. I. Cambridge, MA: Harvard University Press.
Marks CE (1981) *Commissurotomy, Consciousness and Unity of Mind*. Cambridge, MA: MIT Press.
O'Brien G and Opie J (1998) The disunity of consciousness. *Australasian Journal of Philosophy* **76**: 378–395.
Radden J (1998) Pathologically divided minds, synchronic unity and models of the self. *Journal of Consciousness Studies* **5**: 658–672.
Rosenthal D (2001) Persons, minds, and consciousness. In: Hahn LE (ed.) *The Philosophy of Margorie Grene*.
Triesman A (1996) The binding problem. *Current Opinion in Neurobiology* **6**: 171–178.
Zaidel E (1995) Interhemispheric transfer in the split brain: long-term status following complete cerebral commissurotomy. In: Davidson RJ and Hugdahl K (eds) *Brain Asymmetry*. Cambridge, MA: MIT Press.

Consciousness

Introductory article

Adam Zeman, University of Edinburgh, Edinburgh, UK

CONTENTS

Consciousness refers both to wakefulness and to the contents of our experience. The subjective aspect of consciousness poses a philosophical problem for objective science.

INTRODUCTION

Since the early 1980s there has been a major effort to make better sense of consciousness. The current fascination with the subject flows from several sources: work by neuroscientists is steadily revealing details of the brain processes which make consciousness possible; psychologists have underlined the existence of a wide range of unconscious brain processes which can be contrasted informatively to conscious ones; computer scientists and engineers are designing sophisticated brain-like systems which can rival human intellectual performance, raising the question of whether such systems are conscious. It is clearly time to work out where consciousness belongs in the scientific scheme of things – and many philosophers are trying hard to do just that.

WHAT DO WE MEAN BY 'CONSCIOUS', 'AWARE', AND 'SELF-CONSCIOUS'?

Consciousness

Defining consciousness is tricky, but it is clearly important to clarify what we have in mind before we try to study it. Its linguistic origins deserve a moment's attention. The Latin source of 'consciousness', as of 'conscience', is the combination of 'cum', meaning 'with', and 'scio', meaning 'to know': in Latin 'conscire' meant to share knowledge, often guilty knowledge, with another person. This use was extended, metaphorically, to the sharing of knowledge with oneself. 'Conscientia' was the knowledge shared. In contemporary use, two senses of consciousness are particularly important.

Consciousness as the waking state

In everyday life, and particularly in medicine, if we ask whether someone is conscious we are generally asking whether he or she is awake – as opposed to asleep, anesthetized, very drunk, or in a coma. We are asking, in other words, about his or her 'state of consciousness'. (*See* **Consciousness, Disorders of**)

We tend to assume that if people are awake, they will also be capable of perceiving their surroundings and their bodies, and of interacting and communicating with others and with the environment. We are usually accurate observers of others' states of consciousness, and of the (normally) linked capacities to perceive, interact, and communicate: these are all 'objective' matters. Indeed, doctors use standardized scales, like the Glasgow Coma Scale, to assess patients' states of consciousness (Figure 1). These scales apply objective criteria to the assessment of consciousness, such as whether a a patient's eyes are open, whether he can speak, and his ability to move his limbs on request.

To be conscious in this sense is to be awake or vigilant. While we can usually come to a firm decision about whether someone is awake, asleep, or comatose, each of these states also admits of degrees: we can be wide awake or drowsy, half-asleep or stuporose.

Consciousness as experience

If someone is conscious in the sense of being awake, the person is usually conscious of something. In its second sense consciousness is the content of experience from moment to moment: what it feels like to be a certain person now, in a sense in which we suppose there is nothing it feels like to be a stone or

Name

Ward

Unit No:

COMA SCALE			
	Eyes Open	Spontaneously	4
		To speech	3
		To pain	2
		None	1
	Best Verbal Response	Orientated	5
		Confused	4
		Inappropriate words	3
		Incomprehensible sounds	2
		None	1
	Best Motor Response	Obey commands	6
		Localise pain	5
		Flexion to pain	4
		Abnormal flexion	3
		Extension to pain	2
		None	1
GCS Total			

Figure 1. Glasgow Coma Scale. This scale is widely used in the clinical assessment of consciousness (it is of course fallible: how would someone who is fully conscious but completely paralyzed score?).

in a coma. We can usually be much less sure of the contents of another person's consciousness than we can be that he or she is conscious. This second sense of consciousness is more inward, more subjective, than the first.

We can make several generalizations about the contents of consciousness in this second sense. They tend to be stable over short periods, from a few hundred milliseconds to a few seconds but changing over time; they have a foreground (at present the words in front of you), a background (the pressure of your clothes or the rumble of traffic), and a limited capacity (you can't simultaneously concentrate on Bach's first *Prelude* or Britney Spears' last album and my article on consciousness); they are usually continuous over time, in the sense that memory allows us to connect the consciousness of the present with the consciousness of the past; they have an immense potential range, including information from our senses, and from all our major psychological processes including thought, emotion, memory, imagination, language, and action planning; and, above all, they are personal, conditioned by the perspective which our particular viewpoint supplies.

We all tend to consider ourselves to be experts on our experience: after all, who could know more than you do about the contents of your own consciousness? But perhaps we can be mistaken about what normally passes through our minds. This is an active area of consciousness research. For example, studies which require subjects to give instantaneous reports of their current experience, when a pocket-held buzzer sounds, produce some surprises: it can be difficult to spell out the contents of consciousness; 'inner thought', rather than sensation, tends to dominate awareness, and these thoughts are often clothed in neither images nor words. Other research, exploring our sensitivity to changes in our visual surroundings, suggests that our visual attention is much more narrowly focused than we normally imagine: we completely miss surprisingly large changes in the scene before our eyes unless our attention is on them at the moment they appear. This kind of study is important: it helps to clarify the data which the science of consciousness needs to explain.

Awareness

'Conscious' and 'aware' are used almost synonymously in ordinary speech, with the difference that 'awareness' tends to imply the occurrence of experience. The two words *can* be used to mark any of several subtly different distinctions: for example, wakefulness versus experience, the contents of experience versus the capacity for it, the objective versus the subjective aspects of experience. But these are all rather technical distinctions which should be explained when they are drawn. The two terms are used interchangeably here.

Self-consciousness

'Self-consciousness' is an even more slippery term than 'consciousness' itself. We can mean at least five different things when we say that someone is 'self-conscious'. (*See* **Self-consciousness**)

Self-consciousness as proneness to embarassment

In colloquial speech, if we say that someone is self-conscious we mean that the person is awkward in the company of others because he or she imagines that they are scrutinizing. In other words, we are

self-conscious in this sense when we are excessively aware of others' awareness of ourselves.

Self-consciousness as self-detection

We sometimes say that a man or animal is 'self-conscious' when he detects a stimulus which impinges directly upon him (like an ant crawling over his hand), or when he behaves in a way that suggests an awareness of his own actions (like a dog with its tail between its legs after eating your supper). But this amounts to little more than perceptual awareness, directed towards events brought about by, or impinging directly upon, the creature in question.

Self-consciousness as self-recognition

This sense of self-consciousness was highlighted by the work of Gordon Gallup, in the 1970s, showing that chimpanzees and orangutans can recognize themselves in mirrors, but monkeys cannot. Human children become able to do so at around the age of 18 months. This ability suggests that apes, like small children, have an 'idea of me', a concept of 'self', although probably a very simple one. It is significant that over the few months after they come to recognize themselves in mirrors, children show a growing interest in self-adornment and master the use of the first-person pronoun, 'I'.

Self-consciousness as awareness of awareness

Between the ages of two and five, most children come to realize that they, like others, gain knowledge of the world from limited points of view, forming beliefs which can be mistaken. Their growing understanding of the nature of belief and the possibility of deception has been described as a developing 'theory of mind'. Once you possess this theory, your 'idea of me' has expanded to take in the notion that *you* are not merely a body, capable of reflection in a mirror, but also a mind, a subject of experience.

Self-consciousness as self-knowledge

In its broadest sense, our self-consciousness includes our knowledge of ourselves as members of particular families, schools, professions, social classes, language groups, and nations. We explore our peculiarly human fascination with ourselves in many forms of art: for example in Rembrant's astonishing life-long series of self-portraits.

THE SCIENCE OF CONSCIOUSNESS

Wakefulness

The electricity of the brain

We have learnt a great deal about the neurology of sleep and wakefulness in the past hundred years. In 1929 Hans Berger, a German psychiatrist, reported the first recordings of the electrical activity of the human brain made from the scalp: the 'electroencephalogram' or EEG. Berger and his followers went on to describe a series of brain rhythms (Figure 2) which correlate with states of consciousness: thus beta rhythm predominates while you read this article; alpha rhythm will become prominent if you relax and close your eyes, with increasing amounts of theta and delta if you let yourself drop off to sleep. (*See* **Consciousness, Sleep, and Dreaming**)

Research in the 1950s revealed that sleep itself has a complex structure: on falling asleep, our brain activity descends through a series of stages of deepening sleep, with progressive slowing of the EEG, but after half an hour or so of deep 'slow wave sleep' (SWS) it reascends through these stages. This ascent culminates in a period of

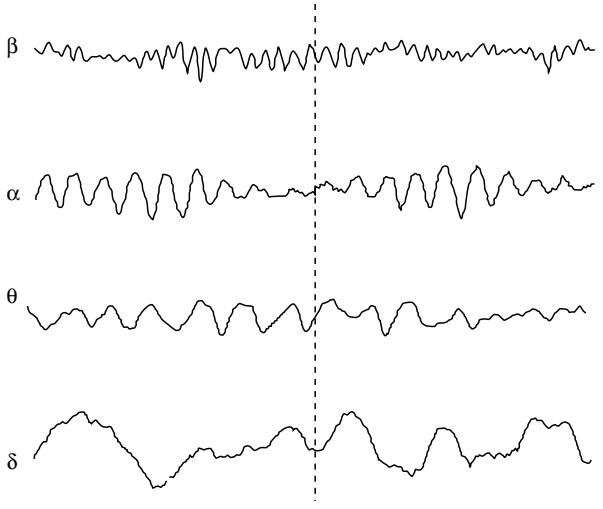


Figure 2. The rhythms of the EEG. This shows the four most commonly recognized EEG rhythms, obtained from four different clinical recordings: beta rhythm (> 14 Hz or cycles/second) characterizes active wakefulness: alpha (8–13 Hz) relaxed wakefulness with the eyes closed; theta (4–7 Hz) and delta (4–7 and < 4 Hz, respectively) occur in sleep and pathological states of depressed consciousness. A two-second period is shown. Reproduced with permission fom Zeman (2001) Consciousness. *Brain* **124**: 1263–1289.

'rapid eye movement sleep' (REM), characterized by EEG appearances similar to those of wakefulness, rapid eye movements, deep relaxation of our muscles, and the experience of dreaming. The cycle is repeated three or four times each night with decreasing amounts of slow wave sleep and increasing amounts of REM in successive cycles (helping to explain why we so often wake in the morning with a dream in the mind's eye).

Since his pioneering work, Berger's technique has been greatly refined. It is now possible to isolate the electrical activity associated with 'mental acts', as he had hoped: electrical correlates of sensation, attention, thought, and intention can all be identified at the scalp. There has been much interest recently in the idea that rapid synchronized activity in the gamma range (25–100 Hz) may be a hallmark of conscious processes.

The control of conscious states

In parallel with the exploration of the electrical correlates of consciousness, a series of discoveries has clarified the brain structures which control our conscious states. Observations of the effects of human brain disease, and experiments with animals, converged on the idea that regions of the brain stem and thalamus contain an 'activating system' that regulates the activity of the cerebral hemispheres (Figure 3).

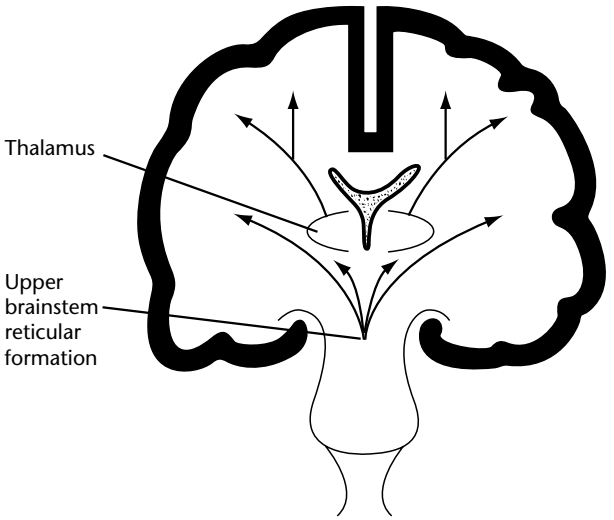


Figure 3. The reticular activating system. This simplified representation makes the points that the upper brainstem and the thalamus play a crucial role in activating the cerebral hemispheres and enabling wakefulness. Reproduced with permission fom Zeman (2001) Consciousness. *Brain* **124**: 1263–1289.

Early models of this 'ascending reticular activating system' supposed that it was a nonspecific mechanism for maintaining wakefulness and alerting the hemispheres to the occurrence of significant events requiring their attention. This picture has been replaced by a much more complex one. It takes account of the existence of several chemical subsystems, employing different neurotransmitters – such as noradrenaline (norepinephrine), acetylcholine, serotonin, dopamine, and histamine – and of regions within the brainstem which serve specific functions, for example inducing REM sleep. But, the broad principle that regions of the brain stem and thalamus orchestrate our conscious states survives within this more sophisticated scheme.

Many details of the signals which switch the brain between wakefulness, SWS, and REM need to be clarified, but neuroscience can now give a plausible account of the neuronal basis of the distinction between wakefulness and sleep. Activating signals from the brainstem to the thalamus fall away as sleep begins, so that the neurons of the thalamus cease to transmit sensory signals faithfully to the cortex (their wakeful 'spike' mode of response). Instead, they enter into a series of rhythmic oscillations, detected at the scalp as the deepening stages of sleep (their 'burst' mode of response). The onset of REM corresponds to a partial reactivation of the thalamus and cortex, giving rise to a 'waking' EEG but with cerebral processing focused on internally generated events. (*See* **Consciousness, Sleep, and Dreaming**)

Pathologies of wakefulness

All the above points to the existence of three principal states of consciousness in health: wakefulness, SWS, and REM. Disease generates a number of further states: these include coma, a state of unresponsiveness resembling SWS but in which the subject is unrousable and the normal cycle of sleep and waking is lost; the vegetative state, a state of 'wakefulness without awareness', in which brainstem mechanisms continue to produce a sleep–wake cycle in the absence of the hemispheric function required to produce experience; and brain death, in which the brainstem is irrevocably destroyed. (*See* **Consciousness, Disorders of**)

Experience

Although brain scientists have sometimes fought shy of consciousness, a great deal of brain research is relevant to the neurology of experience: work exploring brain regions concerned with perception,

attention, memory, language, emotion, and action often reveals correlations between brain events and features of awareness. Some scientists have been working to refine these correlations; others have taken a different, more roundabout, approach to understanding consciousness, by studying unconscious processes. (*See* **Neural Correlates of Consciousness as State and Trait**; **Perception, Unconscious**; **Unconscious Processes**)

Exquisite correlations

Vision is the most intensively studied human brain function and has provided the basis for much of the discussion of the neural basis of consciousness. A series of discoveries this century have revolutionized our picture of the brain events which underly conscious vision. Key findings include the discoveries that: the occipital cortex contains a detailed map of the visual world, in 'area V1'; within this map, cells inspecting each portion of visual space search for the presence of oriented edges; a further 30–40 visual 'maps' surround the primary visual area in the occipital cortex; parallel, though interconnected, streams of visual information flow through these maps, conveying information which defines visual form, color, depth, and motion; two broadly defined pathways fan out from the occipital cortex, an occipito-temporal pathway concerned with identifying objects and an occipito-parietal pathway concerned with visually guided action.

Detailed findings within this program of research have furnished remarkably close correlations between regional brain function and aspects of our experience. For example, human functional brain imaging studies indicate that perception of a colored scene selectively activates a particular region of visual cortex (often called V4); damage in this region can abolish the conscious perception of color. A distinct region plays a comparable role in the conscious perception of movement (V5).

Recently, several scientists have tried to home in on the neural correlate of consciousness (NCC) using a novel strategy. This stems from the thought that correlation need not imply cause: the fact that a brain area becomes active during visual perception does not imply that it causes our conscious experience – it might play some other role in the brain. One partial solution to this problem is to examine changes in brain activity which occur when experience changes without any change in the world. This happens, for example, when we summon up a visual image, or have an hallucination, or switch our attention without moving our eyes, or undergo a switch in the perception of an ambiguous figure. Examination of the neural correlates of these internally driven experiences is at the forefront of the quest for the NCC. (*See* **Neural Correlates of Consciousness as State and Trait**)

Unconscious processes

There is good evidence that our brains can register stimuli which we never consciously perceive, and that these in turn can influence our behavior. Understanding the events in the brain which enable these unconscious processes should help to sharpen the definition of the neurology of awareness. (*See* **Perception, Unconscious**; **Unconscious Processes**)

Conditions under which unconscious perception can be shown to occur include presentation of very brief, faint, or 'masked' stimuli to normal observers; presentation of stimuli to subjects during anesthesia or hypnosis, and neurological syndromes which impair conscious perception, such as blindsight and neglect ('blindsight' is the term given to a range of visually based abilities possessed by subjects who have no conscious vision after damage to area V1; 'neglect' is a disorder in which, most commonly, there is failure to pay attention to the left side of space following damage to the right side of the brain).

The investigation of these unconscious processes is an active area of research. It is too soon to reach firm conclusions about the key differences between conscious and unconscious brain activity. Two main types of explanation for the distinction have been proposed: that unconscious processes result when brain systems which sometimes give rise to consciousness are active at very low levels, and that unconscious processes occur in distinct brain systems, for example subcortical ones, where activity never gives rise to awareness. Recent experiments provide some support for both proposals.

THEORIES OF CONSCIOUSNESS

The data from the developing science of consciousness have spawned a number of overarching theories. They fall into three main types.

Neurobiological Theories

Theories of this type generally assume two broad principles which have emerged from the past century of research: that structures in the upper brainstem and thalamus play a key role in arousal, and that cortical activity supplies much or all of the contents of awareness. They tend, also, to assume that the neural correlate of consciousness will be a more or less extensive network of neurons. These

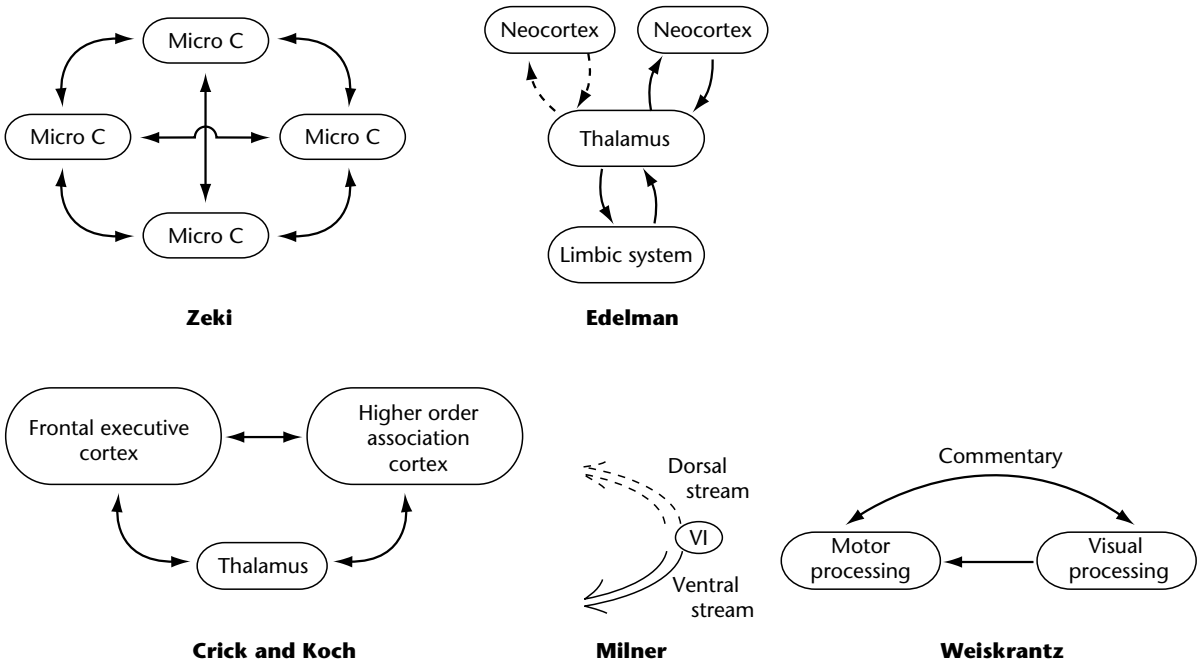


Figure 4. Neurobiological theories of consciousness. This highly schematic figure sketches the outlines of several current theories of consciousness: Zeki's model of interacting 'microconsciousnesses' within the visual system; Edelman's emphasis on interaction between shifting regions of neocortex concerned with perception and the limbic system, via the thalamus; Crick and Koch's proposal that only regions of cortex which 'directly' influence action can participate in conscious processing; Milner's distinction between a conscious 'ventral' and unconscious 'dorsal' stream of visual processing; Weiskrantz's suggestion that consciousness arises from a neural 'commentary' upon otherwise unconscious sensorimotor interactions.

points of agreement leave plenty of scope for disagreement over key details: how large must the network be to give rise to awareness? Need it incorporate particular types of neuron? Need it involve given cortical regions, or possess a particular range of connections with regions elsewhere? Must it engage in any particular pattern or duration of activity? Must it give rise to a certain complexity of interaction?

There is no consensus on these points. Figure 4 sketches a handful of the models currently on offer, underlining their variety. Semir Zeki has suggested that individual visual cortical areas generate their own 'microconsciousness' of color or of movement. David Milner proposes that only the occipitotemporal stream of visual processing is conscious, while the occipitoparietal is involved in action guidance. Gerald Edelman has emphasized the importance of reciprocal interactions between sensory cortical regions, limbic areas concerned with memory and 'value', and the thalamus. Francis Crick and Christof Koch have argued for the role of interactions between sensory areas and motor regions with which they directly interact. Larry Weiskrantz has developed the idea that visual consciousness arises from a secondary 'commentary' on visuomotor processes.

These theories focus mainly on the anatomy of consciousness: it is likely that a certain kind of neural activity is also required for consciousness. Current interest is focused on the role of rapid synchronized gamma-band activity, as there is evidence that this activity is abundant both in states of awareness generally and, specifically, in brain areas which are currently giving rise to experience.

Cognitive and Information-processing Theories

While neurobiological theories consider the nuts and bolts of consciousness, cognitive theories address its functions. Much of what we do, from brushing our teeth to riding a bike, can be achieved with little or no conscious attention. By contrast, novel challenges and unpredictable events force us to mobilize our psychological resources, engaging consciousness.

Cognitive theories take their lead from this everyday observation. Baars, for example, proposes that consciousness allows us to harness the resources

of otherwise independent, unconscious, 'expert systems' in the brain to solve knotty problems as they arise: this sacrifices the high speed and high capacity of automatic 'parallel' processing in the interests of flexible, deliberate behavior. This approach meshes with the widely held idea that consciousness arose in the course of evolution as flexible patterns of learned behavior emerged from the more rigid instinctive patterns of response seen in animals with simpler nervous systems. (*See* **Consciousness, Cognitive Theories of**)

If these theories are correct, they imply that the essence of consciousness lies not in its physical base but in the role it plays in processing information in the brain. If so, it follows that a machine which could reproduce the information flux in the human brain would necessarily be conscious. (*See* **Consciousness, Machine**)

Social Theories of Consciousness

The inspiration for social theories of consciousness is the thought that our awareness of our world and of ourselves is deeply influenced by other human beings: through social interactions from infancy on, through the acquisition of language, the greatest of all our social creations, and through our education in a common culture. In the course of our social development we acquire a 'theory of mind' which, as we have seen, supplies a distinctively human form of self-awareness. Yet while it is clearly necessary to take account of these social facts in giving a full description of human consciousness, it is doubtful whether social theories supply the right level of explanation for the simpler forms of consciousness which we share with animals, for example many of our sensory experiences, desires, emotions, intentions, pleasures, and pains.

THE PHILOSOPHY OF CONSCIOUSNESS

When philosophers discuss the nature of consciousness today, they are continuing an extremely ancient conversation about the relationship between mind and body, subject and object, the realm of experience and the realm of matter. This topic lies at the heart of philosophy: the views philosophers take on it cannot easily be prised apart from their views on a series of other thorny issues, such as the nature of meaning and of knowledge.

Three intuitions lie in the background of much of the philosophical discussion. The first is that conscious experience is 'rich and real': if we are to understand the universe we inhabit fully, we must be able to account for the variety and intensity of our conscious experience, from joy to sadness, from the hues of a sunset to the taste of salt. But, second, experience is clearly bound up with our physical being: everyone knows that fatigue, knocks on the head, too much beer, and countless other physical events can modify the state and contents of our consciousness. Third, we normally assume that awareness makes a difference: if I had not felt hungry, I would not have gone to the fridge. Making sense of the relationship between experience and the brain is difficult because these three intuitions do not square easily with each other.

This becomes clearer on examining three of the more popular philosophical theories of consciousness. One view, 'identity theory', is that conscious events are simply brain events. This idea meshes well with the second and third intuitions: brain events are physical, and well placed to cause our behavior. But does it do justice to the first? Some philosophers think not, arguing that one could know everything there is to know about a brain process and yet lack a full understanding of 'what it is like' to be the creature in whom it takes place. Take the famous example of 'what it is like to be a bat': knowing everything about the bat brain and bat behavior would not tell you what it is like to be a bat engaged in echolocation – or so the argument goes.

A second school of thought, 'functionalism', suggests that the essence of conscious states lies in the functions they serve in our behavior: the essence of vision is that it enables us to control our behavior using our eyes. Understand this control function and you understand visual experience: reproduce this function and you will have created visual consciousness. Once again this approach does justice to the second and third intutions but arguably fails to live up to the first: it is not immediately clear that a 'seeing machine' need be conscious; if it were conscious, it is not clear that its experience would necessarily resemble ours.

The third school of thought, 'dualism', is probably still dominant in our culture. It holds that mental and physical events are of radically different kinds: although the mental and physical realms must be linked in some way, neither can be reduced to the other. This approach appeals to those whose first intuition about consciousness is that it somehow 'goes beyond' the physical: that mental facts are 'further facts' about the world. It can be made compatible with the second intuition, that experience is physical, with the help of 'bridging rules' that link brain events with experience. But it is extremely difficult to see how it can be made

compatible with the third intution, that experience makes a difference to behavior. For if mental events are nonphysical, how can they change the course of physical events?

The difficulty of reconciling these three intuitions calls for radically different ways of thinking about the philosophical problem of consciousness. So, consider two contrasting suggestions to indicate the diversity of current views. Each questions an undeclared assumption of most scientific theories of consciousness.

The first idea is that scientific theories are mistaken in assuming that consciousness *arises* from complexity. Perhaps consciousness is inherent in matter of the simplest kinds, and the complex organization of the brain merely allows this potentiality of matter to emerge and blossom. On this view brain events, in common with all physical events, have inherent mental and physical aspects: any attempt to reduce either aspect to the other would be mistaken. This view, panpsychism, is quite alien to our scientific culture, but it is an understandable response to the problem of explaining how consciousness can be conjured from brain events: on this view there is no need for any magic, as consciousness is present from the start.

The second idea is that we are mistaken in assuming that consciousness arises from the brain. At first sight this idea seems quite bizarre, a denial of the obvious. But it has some powerful backing. The argument runs: when we try to explain consciousness in terms of brain events we are doomed to failure, because we have denied ourselves precisely the resources the explanation requires. We need to expand the limits of our explanation to take in the world we inhabit and the means by which we explore it. For consciousness is a complex kind of interaction with the environment: 'seeing' for example is 'a way of acting'. This approach suggests – and gives reasons for believing – that much of what we take to be 'in our heads' is in fact out in the world, and that our ordinary picture of consciousness, as an internal representation of reality, is mistaken. On this view the brain is not the *source* of consciousness but a device which enables awareness, and awareness is not an invisible process but the use of a set of elaborate skills.

CONCLUSION

There has been huge progress in the scientific study of consciousness over the past century, and there is a growing clarity about the sources of conceptual difficulty in pinning down this elusive prey. The determination of contemporary scientists and philosophers to do justice both to the rich texture of experience and to its intimate relationship to brain events holds out great promise for the years to come.

Further Reading

Chalmers D (1996) *The Conscious Mind*. New York, NY: Oxford University Press.

Churchland PM (1984) *Matter and Consciousness*. Cambridge, MA: MIT Press.

Crick F (1994) *The Astonishing Hypothesis*. London, UK: Simon & Schuster.

Dennett D (1991) *Consciousness Explained*. London, UK: Penguin Press.

Edelman G (1992) *Bright Air, Brilliant Fire*. London, UK: Penguin Books.

Frith CD and Frith U (1999) Interacting minds – a biological basis. *Science* **286**: 1692–1695.

Frith C, Perry R and Lumer E (1999) The neural correlates of conscious experience: an experimental framework. *Trends in Cognitive Sciences* **3**: 105–114.

Thomas Nagel (1979) What is it like to be a bat? In: *Mortal Questions*. Cambridge, UK: Cambridge University Press.

O'Regan K and Noe A (in press) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*.

Rees G, Kreiman G and Koch C (2002) Neural correlates of consciousness in humans. *Nature Reviews* **3**: 261–270.

Schiff ND and Plum F (2000) The neurology of impaired consciousness: global disorders and implied models. http://www.phil.vt.edu/assc/niko.html.

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Weiskrantz L (1997) *Consciousness Lost and Found*. Oxford, UK: Oxford University Press.

Zeki S (1993) *A Vision of the Brain*. Oxford, UK: Blackwell Scientific.

Zeman A (2001) Consciousness. *Brain* **124**: 1263–1289.

Zeman A (2002) *Consciousness: A User's Guide*. London, UK: Yale University Press.

Direct Reference

Intermediate article

David Braun, University of Rochester, Rochester, New York, USA

CONTENTS

The theory of direct reference says that the meaning of a proper name (such as 'Mark Twain') or indexical (such as 'he') is just its referent.

WHAT IS DIRECT REFERENCE?

The theory of direct reference is a theory about the meanings of two sorts of natural language expressions: proper names (such as 'Mark Twain' and 'Paris') and indexicals (such as 'I', 'today', 'you', 'he', and 'that'). It says (roughly) that the meaning of any such expression is just its referent. The traditional rival to this theory is the view that these expressions have descriptive meanings, in addition to their referents.

The theory of direct reference can be divided into negative and positive parts. The negative part says that the meanings of proper names and indexicals are fundamentally different from those of definite descriptions. Definite descriptions are expressions of the form 'the *F*', such as 'the first Postmaster General of the USA' and 'the even prime number'. A definite description expresses a property (for example, being-an-even-prime-number) which is part of the meaning of the definite description; the definite description refers to the object that uniquely has that property. Thus a definite description refers to an object *indirectly*, by expressing a property that determines its referent.

The negative part of the theory of direct reference says that the meanings of proper names and indexicals are not descriptive in this sense. The meanings of these expressions are not properties; their references are not determined by associated properties, in the way that the references of definite descriptions are; instead, these expressions refer *directly*. The positive part of the theory asserts that the meaning of a proper name or indexical is simply its referent.

One can accept the negative part of the theory without accepting the positive part; indeed, sometimes the term 'theory of direct reference' is used to label just the negative part.

The above description of the theory needs some technical refinements, which are discussed below.

HISTORY

The term 'direct reference' was introduced by David Kaplan (1989) in the 1970s, but the theory is almost certainly ancient in origin. John Stuart Mill was perhaps the first modern philosopher to argue that proper names are directly referring (Mill, 1843); thus, the theory of direct reference for proper names is sometimes called 'Millianism'. Gottlob Frege (1893) and Bertrand Russell (1905) formulated seemingly powerful arguments against the view, and presented alternative theories which were widely accepted in the late nineteenth and early twentieth centuries. Ruth Barcan Marcus revived the theory of direct reference in the 1960s (Marcus, 1961); and from the early 1970s, Saul Kripke (1980), Keith Donnellan (1972), David Kaplan (1989), and John Perry (1977) presented arguments for the theory that persuaded many philosophers of its viability. (*See* **Reference, Theories of**)

Frege's and Russell's arguments against direct reference for proper names turn on certain difficulties which are now commonly called *Frege's Puzzles*. The first puzzle is the *puzzle of informative identities*. If the meaning of a proper name is just its referent, then two proper names that refer to the same thing, such as 'Mark Twain' and 'Samuel Clemens', have the same meaning. It seems to follow that the two identity sentences below have the same meaning:

Mark Twain is Mark Twain. (1)

Mark Twain is Samuel Clemens. (2)

But these sentences seem to differ in meaning:

sentence 1 seems to be uninformative, whereas sentence 2 seems to be highly informative.

The second puzzle is the *puzzle of apparent reference to nonexistents*. If the meaning of a proper name is its referent, then names that refer to nothing, such as 'Pegasus', should be meaningless. Therefore, sentences in which 'Pegasus' appears, such as 'Pegasus flies' and 'Pegasus does not exist', should also be meaningless. But these sentences seem to be meaningful; in fact, the second seems to be true.

The third puzzle is the *puzzle of substitution of coreferring names*. If the meaning of a proper name is just its referent, and we have two names that refer to the same thing, then we should be able to substitute one name for the other in any sentence without changing the meaning of the sentence. Moreover, the substitution should not change the truth-value of the sentence (that is, the sentence should not change from true to false or vice versa). But consider the following two sentences:

John believes that Mark Twain wrote *Huckleberry Finn*. (3)

John believes that Samuel Clemens wrote *Huckleberry Finn*. (4)

It seems possible for sentence 3 to be true while sentence 4 is false. So substitution of coreferring names can change the truth-value of a sentence.

Russell concluded from these puzzles that the meaning of a proper name is not its referent, but is instead descriptive in nature; ordinary proper names are just 'abbreviations' for definite descriptions that speakers have in mind when they use names (Russell, 1905). We can determine the description that a speaker associates with a name *N* by asking the speaker 'Who is *N*?' Thus, for some speakers, the meaning of 'Mark Twain' might be identical with the meaning of 'the author of *Huckleberry Finn*', while that of 'Samuel Clemens' might be 'the person who lives next door'. These descriptions express different properties, and so differ in meaning. Thus sentences 1 and 2 also differ in meaning, and one can be informative while the other is not. 'Pegasus' might mean the same as 'the flying horse'; this description is meaningful even if it fails to refer, so sentences containing it can be meaningful. Sentences 3 and 4 differ in meaning because the names 'Mark Twain' and 'Samuel Clemens' differ in meaning; they differ in truth-value because they attribute different beliefs to John.

Like Russell, Frege also attributed descriptive meanings, which he called 'senses', to proper names (Frege, 1893). Frege's solutions to the puzzles are similar to Russell's. There are differences between Frege and Russell which we ignore here. (*See* **Sense and Reference**)

ARGUMENTS FOR DIRECT REFERENCE

Many of the arguments in favor of direct reference are simply arguments against description theories of proper names.

According to Russell, the meaning for a speaker of a name *N* is the same as that of the definite description that the speaker would provide when asked 'Who is *N*?' Suppose that we ask Fred 'Who is Mark Twain?' and he answers 'The author of *Huckleberry Finn*.' Now consider the following two sentences:

If there is exactly one author of *Huckleberry Finn*, then Mark Twain is the author of *Huckleberry Finn*. (5)

If there is exactly one author of *Huckleberry Finn*, then the author of *Huckleberry Finn* is the author of *Huckleberry Finn*. (6)

The only difference between them is that sentence 6 contains the definite description 'the author of *Huckleberry Finn*' where sentence 5 contains the name 'Mark Twain'. Thus, according to Russell, sentences 5 and 6 should have the same meaning for Fred. But Kripke (1980) and Donnellan (1972) present several reasons for thinking that these sentences do not have the same meaning for Fred.

Kripke points out that sentence 6 expresses a necessary truth. So if sentence 5 means the same thing as sentence 6, then sentence 5 should also express a necessary truth. But sentence 5 clearly does not express a necessary truth; after all, Mark Twain might have died before he wrote *Huckleberry Finn* (as Fred would admit). A similar point would hold for just about any definite description that Fred might provide.

The next argument comes from both Donnellan and Kripke. Imagine that *Huckleberry Finn* was written by Samuel Clemens's cousin, Clyde Clemens, who died shortly after completing it; Twain/Clemens stole the manuscript, passed it off as his own, and since then has been commonly thought to be the author. Now if this is the case, then sentence 5 is false, even as spoken by Fred. But the description theory entails that it is true, for according to that theory, sentence 5 (in Fred's mouth) means the same as sentence 6, which is obviously still true in our imaginary situation. Furthermore, according to the description theory, the name 'Mark Twain' (in

Fred's mouth) refers in this situation to the referent of the description 'the author of *Huckleberry Finn*', which is Clyde. But this is incorrect.

Finally, Kripke notes that Russell's objections take for granted that every speaker who uses the name 'Mark Twain' can formulate a definite description that 'identifies' the referent of the name. But this assumption is incorrect. When asked 'Who is Mark Twain?', some people may answer 'A writer'. This description does not determine a unique referent for the name; yet many of these people may be unable to produce a better identifying description.

Kaplan (1989) and Perry (1977) present similar arguments against description theories of indexicals. Consider the hypothesis that the descriptive meaning of an utterance of 'you' is the same as that of an utterance of 'the person I am addressing now' by the same speaker. Then utterances of the following two sentences by the same speaker should mean the same thing:

> If I am addressing exactly one person now, then you are the person I am addressing now. (7)

> If I am addressing exactly one person now, then the person I am addressing now is the person I am addressing now. (8)

However, sentence 8 expresses a necessary truth, whereas sentence 7 does not. To see that sentence 7 does not, imagine that I utter it while addressing Mary. What I say is true, but it could have been false; for, after all, I could have addressed someone other than Mary, or no one at all.

Examples involving indexicals, like sentence 7, show that we need to distinguish between two different sorts of meaning. Suppose that two different people utter sentence 7 while addressing different people. There is a sense in which the two utterances of 'you' share a meaning, which we may call the *linguistic meaning* of 'you'. But there is another sense in which those utterances differ in meaning: let's say that the two utterances of 'you', and the two utterances of sentence 7 as a whole, differ in *content*. Call the content of an utterance of a full sentence a *proposition*. The above arguments show that the contents of utterances of indexicals are not descriptive. Perhaps the linguistic meanings of indexicals are in some way descriptive, but that is an open question. (*See* **Philosophy of Language**; **Meaning**)

It is now possible to state the theory of direct reference more accurately. According to the negative part, the content of an utterance of a proper name or indexical is not descriptive (is not a property). According to the positive part, the content of an utterance of such an expression is simply the referent of the utterance; moreover, an utterance of a sentence containing such an expression expresses a *singular proposition*: a proposition containing the referent of the expression as a constituent.

PROBLEMS WITH DIRECT REFERENCE

The most important problems for the theory of direct reference are those posed by Frege's Puzzles. Advocates of direct reference have proposed solutions to these puzzles, but whether their solutions are successful is a subject of much current debate.

In response to the puzzle of informative identities, Nathan Salmon (1986), Scott Soames (1987), and other direct reference theorists say that sentences 1 and 2 express the same singular proposition, and, in that sense, mean the same thing. But they say that it is possible for a person to grasp that proposition in distinct ways. Different direct reference theorists have different views about what these ways of grasping propositions are. Some think that they are *mental representations*, and so hold that sentences 1 and 2, and the names 'Mark Twain' and 'Samuel Clemens', are connected with different mental representations in most speakers. This may explain why sentences 1 and 2 seem to differ in informativeness, and why many speakers incorrectly think that they differ in meaning. (*See* **Propositional Attitudes**)

In response to the puzzle of apparent reference to nonexistents, Salmon (1998) claims that names like 'Pegasus' refer to mythical or fictional objects that are created by our story-telling activities.

In response to the puzzle of substitution of coreferring names, Salmon (1986) and Soames (1987) claim that sentences 3 and 4 express the same proposition, and must have the same truth-value. But utterances of them differ in what they suggest about John: sentence 3 correctly suggests that John would assent to 'Mark Twain wrote *Huckleberry Finn*', whereas sentence 4 incorrectly suggests that he would assent to 'Samuel Clemens wrote *Huckleberry Finn*'. This might mislead some speakers into thinking that the sentences themselves differ in truth-value.

DIRECT REFERENCE AND COGNITIVE SCIENCE

The theory of direct reference implies that ordinary speakers are less reliable at judging differences in

meaning than some cognitive scientists may think. For example, most speakers would judge that sentences 1 and 2 differ in meaning; but according to the theory of direct reference, such speakers are mistaken.

The theory also implies that two agents who believe and desire the same propositions may nonetheless behave quite differently. Suppose that John assents to 'Mark Twain is the author of *Huckleberry Finn*' and dissents from 'Samuel Clemens is the author of *Huckleberry Finn*'; suppose Mary assents to both. Now according to Salmon and Soames, John also believes that Clemens is the author of *Huckleberry Finn*, just like Mary (whatever he might say to the contrary). But they might behave differently in similar situations. Suppose, for example, that both assent to 'I want Twain to autograph my copy of *Huckleberry Finn*', and both are told 'Clemens is in the next room.' Then Mary will go next door to get her copy of *Huckleberry Finn* autographed, but John will not. The reason they will behave differently is that John does not believe the proposition in the same *way* as Mary: he believes it in a 'Twain' way, but not in a 'Clemens' way, whereas Mary believes it in both ways. Examples like this lead some philosophers to think that, if the theory of direct reference is correct, then ordinary belief attributions do not provide explanations of behavior of the sort we ordinarily expect; full explanations need to mention the ways in which propositions are grasped.

References

Donnellan K (1972) Proper names and identifying descriptions. In: Davidson D and Harman G (eds) *Semantics of Natural Language*, pp. 356–379. New York, NY: Humanities Press.

Frege G (1893) Über Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik* **100**: 25–50. [In German. Translated as: Frege G (1952) On sense and reference. In: Geach P and Black M (eds and trans) *Translations From Philosophical Writings*, pp. 56–78. Oxford, UK: Blackwell.]

Kaplan D (1989) Demonstratives. In: Almog J, Perry J and Wettstein H (eds) *Themes From Kaplan*, pp. 481–614. Oxford, UK: Oxford University Press.

Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Marcus R (1961) Modalities and intensional languages. *Synthese* **13**: 303–322.

Mill JS (1843) *A System of Logic*. London, UK: Longman.

Perry J (1977) Frege on demonstratives. *Philosophical Review* **86**: 474–497.

Russell B (1905) On denoting. *Mind* **14**: 479–493.

Salmon N (1986) *Frege's Puzzle*. Cambridge, MA: MIT Press.

Salmon N (1998) Nonexistence. *Nous* **32**: 277–319.

Soames S (1987) Direct reference, propositional attitudes, and semantic content. *Philosophical Topics* **15**: 47–87.

Further Reading

Salmon N (1989) Reference and information content: names and descriptions. In: Gabbay D and Guenthner F (eds) *Handbook of Philosophical Logic*, vol. IV, pp. 409–461. Dordrecht, The Netherlands: Reidel.

Devitt M (1989) Against direct reference. *Midwest Studies in Philosophy* **14**: 206–240.

Dualism

Intermediate article

David Robb, Davidson College, Davidson, North Carolina, USA

CONTENTS

Dualism is the view that the mental and the physical are distinct and mutually irreducible.

WHAT IS DUALISM?

As a candidate solution to the mind–body problem, dualism is opposed to the two main forms of monism: *materialism*, which reduces the mental to some part or aspect of the physical, and *idealism*, which reduces the physical to some part or aspect of the mental. Dualists regard monists of either sort as making the same mistake: conflating domains that are distinct and autonomous.

Dualism is a metaphysical doctrine, a view about the ultimate nature of reality. One often finds it packaged with other philosophical views, but it need not stand or fall with these. For example, while certain versions of dualism are congenial to theism, there is nothing in dualism *per se* that requires the existence of God, or that is committed to any particular religious tradition. Nor does dualism entail the controversial epistemological doctrines that often accompany it. These include the transparency of the mental – according to which the nature of the mind and its states is fully revealed in introspection – and the incorrigibility of first-person beliefs – according to which I am the final authority on the contents of my own mind. Dualism is compatible with the acceptance or rejection of either of these epistemological doctrines.

Even when isolated from these controversial views, dualism – at least in its most radical forms – is largely rejected by the cognitive science community, where materialistic monism predominates. Nevertheless, dualism has never completely disappeared from the intellectual scene, and it is enjoying something of a renaissance. In fact, as we will see, a moderate form of dualism is today widely accepted among cognitive scientists.

VARIETIES OF DUALISM

So far I have referred broadly to 'the mental' and 'the physical', but when dualists distinguish these realms, they normally have entities of a certain sort in mind. Here dualism divides into several related versions: there are dualisms of substances, properties, events, processes, facts, sentences, and more. This article will discuss only substance and property dualism, though much of the discussion applies to other versions as well.

Substance Dualism

A *substance* is a thing or object, something that has properties and persists through time. A substance dualist conceives of the mind as such an entity, one that, moreover, is distinct from the biological body or any part of it, such as the brain.

The most radical form of substance dualism is *Cartesian dualism*, named after its most famous modern defender, René Descartes. Although it has been out of fashion for sometime, Cartesian dualism has a number of contemporary defenders (e.g. Foster (1991)). A Cartesian dualist says that the mind is a nonphysical 'soul' or 'spirit', a substance entirely lacking in physical properties, including spatial location. As Descartes articulated the view, the mind's intrinsic properties are exclusively mental: they are modes (expressions) of consciousness, the essence of minds. By contrast, material substances (bodies) have only physical properties, modes of spatial extension, the essence of bodies.

This difference of essence entails an ontological independence: one kind of substance could exist without the other. Nevertheless, our minds are in fact 'embodied', albeit contingently and only temporarily. Descartes' own views on the nature of this embodiment are obscure; but the relation is often understood causally: for my mind to be embodied

in a particular material substance is for it to have a privileged causal connection with that substance; only my mind can directly (by means of volitions) affect this biological body, and only this body can directly (by means of sensation) affect my mind.

For the Cartesian dualist, then, mind–body dependence is merely causal. But there are forms of substance dualism in which this dependence is much stronger. *Non-Cartesian substance dualists*, while insisting on the strict numerical distinctness of mind and body, allow that the mind has physical properties. Indeed, they claim, its physical properties are just those of the body or brain, for mind and body spatially coincide. E. J. Lowe is a contemporary defender of this more moderate form of substance dualism. He argues that the mind (or self) is a psychological substance, distinct from the body but sharing many of its physical properties (Lowe, 1996). Here the mind–body relation is analogous to the relation between a statue and the lump of clay composing it. Mind and body spatially coincide, but because they differ with respect to certain properties, they are distinct. In a similar vein, those who believe that the mind 'emerges' from the activities of the brain, and those who think that the brain 'constitutes' the mind, would also count as non-Cartesian substance dualists.

Property Dualism

Unlike substance dualists, property dualists need not postulate immaterial substances. They are willing to grant that the mind is nothing more than a complex physical thing. But they still insist that this substance's mental properties are not physical.

In its strongest form, property dualism says that mental properties are fundamentally different from – though they may be lawfully connected to – physical properties. We might call this *radical property dualism*. It is embraced by, for example, Chalmers (1996), at least about qualia. Chalmers believes that the qualitative, categorical features of our conscious experiences are different in kind from any physical properties, which he takes to be exclusively dispositional and structural. *Emergentists* also fall into this category: they take mental properties to be features of physical systems that have achieved a certain level of complexity. Such properties are at best only nomologically or causally dependent on the physical substrate from which they emerge.

Like substance dualism, property dualism exists in a more moderate form, according to which mental properties are not physical, yet always (perhaps necessarily) are realized in or supervene on the physical. The idea here is that mental properties are instantiated at a higher, more abstract level than their physical counterparts. In ascribing a mental property to someone, we abstract away from the physical details that realize, or implement, the mental property in that particular person. For example, in certain versions of functionalism, for a system S to have a mental property M is just for it to have some physical property P_s that within S plays the causal role definitive of M. Since the same M-defining role can be filled by different physical properties in different systems, we cannot identify M with P_s or with any other physical property. Yet there is a clear sense in which, within S, M is determined by, and in fact is 'nothing over and above', its physical realizer P_s. This realization relation is clearly much stronger than the causal or nomic psychophysical relations allowed by radical property dualism. Indeed, the relation here is so intimate that most adherents of moderate property dualism consider themselves materialists. This view is commonly called *nonreductive materialism*, but it is important to remember that it is, strictly speaking, a form of property dualism, since it denies that mental properties are physical.

ARGUMENTS FOR DUALISM

Although dualists have at times appealed to empirical considerations to support their view, the strongest dualist arguments, and those receiving the most philosophical attention, have been *a priori*. In the typical argument, some state of affairs incompatible with materialism is claimed to be clearly conceivable, and so possible. The dualist then moves from this possible state of affairs to a conclusion about the nature of the mental in the actual world.

Arguments for Substance Dualism

The most famous such argument occurs in Descartes' *Sixth Meditation*. A contemporary version proceeds along these lines (here we follow Descartes and other dualists in assuming that I am the same as my mind):

> I can clearly conceive of my existing in the absence of anything physical. (1)

That is, when I consider the matter carefully and rationally, there seems to be no contradiction in the idea of my existing in an entirely immaterial world. From this it follows that:

It's possible for me to exist the absence of anything physical. (2)

(Since what is clearly conceivable is at least possible.) But now note that:

If I am a physical thing, then I am essentially a physical thing. (3)

That is, any physical substance is physical by nature, and so could not exist as an immaterial thing. But then it follows that:

I am not a physical thing. (4)

From the claim of mere possibility in step 2 we have reached a conclusion about what is actually the case, via the linking premise 3.

In spite of its enormous influence among dualists, this argument has been challenged at almost every step. One line of objection faults the inference from step 1 to step 2. Even if I can clearly conceive of myself existing in the absence of anything physical, at most this shows that for all I know (about my nature) it's possible for me to exist in such a state. It doesn't imply that I really can exist in this way. That is, statement 1 at best establishes the epistemic possibility of my existing in an immaterial world, not its genuine, metaphysical possibility. Another line of objection challenges step 3: perhaps I am a physical thing, but not essentially so. Modal intuitions may differ here, but this objection at least puts the burden on the Cartesian to explain why a physical thing cannot be only contingently physical. Yet another line of objection is that the above argument, even if sound, merely establishes the more moderate, non-Cartesian form of substance dualism. That is, perhaps the argument only shows that I am not identical with any physical substance, not that my nature (and all my properties) are nonphysical. This is a delicate question: it turns in part on whether the non-Cartesian can allow for the possibility of disembodied existence. If so, then step 2 would not seem to be strong enough to support a Cartesian reading of the conclusion. (For further discussion of Descartes' argument, see Yablo, 1990.)

In any case, if a substance dualist wants merely to establish the non-Cartesian version, Descartes' modal argument is not required. Since moderate substance dualists claim only that mind and body are numerically distinct, they need only find some property possessed by one but not the other. Since the brain (say) constitutes the mind, the two substances share many of their properties, such as size, location, and so on. But mind and brain do seem to have different persistence conditions: certain kinds of neurophysiological damage, for example, would destroy the mind, but the brain (as a biological substance) would still exist. This difference in properties, the argument goes, entails a numerical difference – the mind is not the brain (or any other part of the body) – but it stops short of Cartesian dualism.

Arguments for Property Dualism

Philosophical arguments for property dualism, at least in its radical form, have also relied heavily on conceivability arguments. One thought experiment features super-neuroscientists able to examine a functioning brain down to the finest physical detail. They would never, it seems, find anything remotely resembling a thought or an experience. So mental properties are fundamentally different from anything physical. (For quite different versions of this argument, see G. W. Leibniz' *Monadology* and Jackson, 1982.) But a materialist might offer the following explanation for why the scientists fail to 'find' the mind: the mind and its contents can be accessed from two perspectives, the introspective, first-person perspective and the observational, third-person perspective. The scientists are in fact encountering mental properties as they observe the brain, but these mental properties aren't recognized as such because they're being accessed from an unfamiliar perspective. The resulting view would be a 'dualism of perspectives', but a materialism of the mind and its properties.

Another thought experiment marshalled in support of radical property dualism features the *zombie*, a being who, in spite of sharing all of the physical properties of a conscious being, is wholly lacking in conscious states (Chalmers, 1996). Such a being seems to be conceivable, yet if it is so much as possible, then our phenomenal properties (an important class of mental properties) are not the same as any of our physical properties. This argument is subject to some of the same criticisms that apply to Descartes. For example, are zombies really possible, or is their apparent conceivability explained by the limitations of our current knowledge? Just as further investigation may reveal, contra Descartes, that my disembodied existence is impossible, so we may eventually learn that zombies are impossible. At this point, however, the materialist can only offer the hope of such a discovery.

Property dualism in its more moderate form – nonreductive materialism – requires nothing as exotic as zombies. Here an appeal to multiple realizability is thought to suffice (Fodor, 1981). Any given mental property can be, and in fact is,

realized by a variety of different physical properties in different species, different individuals of the same species, and even at different times in the same individual. Given that the same mental property is realized by many different physical properties, it cannot be identified with any one of them; hence mental properties are distinct from, though realized in, physical properties. This argument is often used to support certain versions of functionalism against the psychophysical identity theory. (For an identity theorist's response to the argument, see Kim, 1993.)

PROBLEMS FOR DUALISM

However one evaluates the preceding arguments, dualism faces a number of serious obstacles. These range from empirical objections, to the effect that dualism cannot be integrated into cognitive science, to philosophical and conceptual objections. The former are discussed in the final section of this article. Here I discuss the latter.

Charges of Incoherence and Vacuity

One of the reasons why substance dualism fell out of favour is that it seemed incoherent to a number of philosophers. Ryle (1949), for example, accused substance dualists of an egregious conceptual error, that of placing the mind in the wrong ontological category: having a mind isn't a matter of being (or being related to) a substance, immaterial or otherwise. Rather, talk of 'the mind' is merely talk about a set of capacities. Ryle thought that these capacities were exclusively behavioral, but one needn't be a behaviorist to appreciate his insight. A functionalist, for example, may also claim that talk about 'mind' doesn't refer to a kind of substance, but is just a way of describing the causal organization of an organism's mental and behavioral states.

Ryle's objection applies to any form of substance dualism, but there is another charge of incoherence directed specifically at the Cartesian variety: if minds are not located in space, there seems to be no way to individuate them. It is at least logically possible for there to be two qualitatively identical minds, that is, minds which share all of the same intrinsic properties. But by virtue of what, then, are they two minds and not one? Two qualitatively identical material substances can be distinguished by their different spatial locations, but there is no such medium to individuate Cartesian minds. Coupled with the (controversial) doctrine that there can be no entity without clear conditions of individuation, this objection could render Cartesian dualism incoherent. (See Hoffman and Rosenkrantz, 1991.)

Even if substance dualism is coherent, some have objected that as a proposed solution to the mind–body problem, dualism in any of its forms is devoid of positive theoretical content. It seems that dualists can tell us only that the mental is not physical; they can't say anything informative about the intrinsic nature of the mental. This charge of theoretical vacuity is most threatening to radical forms of dualism. Moderate forms – non-Cartesian substance dualism and nonreductive materialism – allow that minds or their mental properties are realized in the physical world, and the nature of this realization might contain enough resources for moderate dualists to say something positive about the mental. But can Cartesian or radical property dualism say anything more about the intrinsic nature of the mind? Perhaps the most obvious option here is also the most promising: discover the positive theoretical content of dualism by appealing to the direct introspective access we have to our own minds and their contents (Foster, 1991). Those who insist on drawing theoretical content only from objective, third-person sources will balk at this move, but the dualist will reply that first-person data cannot be ignored by any systematic study of the mind, at least any study that aspires to completeness.

The Problem of Interaction

The most serious problem for dualism arises from a seemingly undeniable fact: the mental and the physical causally interact. Such interaction is two-way: for example, my decision to get a drink causes me to walk to the refrigerator (mind-to-body causation); and putting my hand on a hot stove causes me to feel pain (body-to-mind causation). Dualists have had trouble integrating these common-sense facts about mind–body interaction into their ontology.

This problem is perhaps most acute for Cartesians, and again the mind's lack of spatial location is the source of the problem. There seems to be no way for a Cartesian mind to causally link to a location in space. Interacting bodies link by spatial contact; yet this mechanical account of the causal nexus cannot work for Cartesian mind–body interaction, since Cartesian minds cannot come into spatial contact with anything. Similar arguments apply in a more recent version of the problem of interaction, the 'pairing problem' (Kim, 2001). Imagine two qualitatively identical minds M_1 and M_2,

and the bodies with which they (allegedly) causally interact, B_1 and B_2. What makes it the case that M_1 is causally paired with B_1, not B_2, and M_2 paired with B_2, not B_1? If these minds were located in space, we might appeal to their different spatial relations to B_1 and B_2 to resolve this question. But this resource is not available to the Cartesian. (For a Cartesian response to the pairing problem, see (Foster, 1991).)

One might think that the problem of interaction doesn't arise for property dualists, who after all can allow that the mind is a physical substance, causally related to the physical world. Yet property dualists face their own problems of causality, for we still would like mental properties to be causally relevant to bodily behavior. We wish, that is, not merely for minds (or mental events) to cause behavior; we wish them to cause behavior by virtue of their mental properties. And it has seemed to many opponents of property dualism that nonphysical properties cannot have a causal effect on the physical world. One much-discussed version of this objection appeals to the *causal closure* of the physical: the causal history of a physical event includes only physical events and their physical properties. At no point, that is, can a nonphysical event or property break into the network of physical causes. Causal closure raises immediate worries for both versions of property dualism: even for nonreductive materialism, since it is not clear that nonphysical properties, even if they are realized in the physical, can be causally relevant without violating closure (Kim, 1993).

Some property dualists cheerfully accept these consequences, thereby embracing *epiphenomenalism* (Jackson, 1982). Here mental properties are demoted to mere 'nomological danglers', properties caused by what goes on in the physical world, but not themselves having any causal efficacy in return. The mind's operations are, therefore, somewhat like the display on a computer's monitor: the changing patterns of colors reflect the internal operations of the computer without having any reciprocal effect on these operations. Epiphenomenalism is attractive for a number of reasons – for example, it allows us to embrace both the irreducibility of mental properties and the causal closure of the physical – yet it too has some obstacles to overcome. Besides being at odds with common sense, epiphenomenalism seems to rob us of an important kind of knowledge: our knowledge of our own mental states. If my qualia, for example, are causally impotent, how could I ever know about their existence and character? Yet first-person knowledge about conscious states is normally thought to be the most secure knowledge we can have. Either epiphenomenalists must bite the bullet and deny that we have such knowledge, or they have to explain how we can know about features of our minds from which we are causally isolated.

DUALISM AND COGNITIVE SCIENCE

What relevance might dualism have for contemporary cognitive science? Cartesian dualism is no longer considered seriously by most cognitive scientists, in spite of its renewed attention from metaphysicians. If cognitive scientists are substance dualists at all, they typically embrace only the non-Cartesian variety: the mind, if it can even rightly be called a substance, is constituted by the body or brain. In spite of this near-consensus, however, it is worth noting that there is nothing in cognitive science *per se* forcing one to reject Cartesian dualism. Most, if not all, theories in cognitive science are neutral with respect to the ultimate nature of the mind. Even psychological theories that 'locate' mental capacities or processes in the brain could be interpreted by a Cartesian dualist as merely revealing their 'neural correlates'. A scientific study of the mind requires systematic (lawful) correlations between mental states and the empirical states of the body and brain. But whether these states are one and the same, and indeed whether the brain is the mind at all, are questions outside the domain of such a study.

However, while empirical theories about the mind are compatible with Cartesian dualism, the progress of cognitive science is making the view less attractive than it was in the seventeenth century. The more cognitive scientists learn about the capacities of physical systems such as the brain, and the more they learn about the dependencies between mental and neural functioning, the less attractive Descartes' theory has become (Damasio, 1994).

In contrast to substance dualism, property dualism is becoming more popular. Indeed, nonreductive materialism is the dominant view in cognitive science, where it is almost taken for granted that mental phenomena exist at a higher, more abstract level than physical phenomena, even if the former are realized in the latter. Some philosophers question whether the hierarchy of 'levels' cognitive scientists speak of really requires a distinct class of mental properties, but this is how such talk is often understood. And while property dualism in its radical form is still a minority view, it has been earning more respect, especially among theorists

of consciousness. Chalmers (1996), for example, believes that nonphysical qualia can be successfully integrated into a science of the mind. He believes that while no physical theory could ever fully capture the nature of our conscious states, cognitive science can fruitfully investigate the laws connecting consciousness with physical systems such as the brain.

Finally, perhaps the most interesting consequences of dualism for cognitive science are not so much metaphysical as methodological. If dualism in any of its forms is true, then any study of the mind is in an important sense an autonomous discipline: theorizing within cognitive science can proceed by and large independently of the physical sciences. As cognitive scientists formulate psychological laws and explanations, they need not await the perfection of, say, particle physics. Among some dualists, this autonomy takes a rather strong form. Descartes thought that the mind was wholly outside the domain of the emerging mechanistic science of his day. And much more recently Davidson (1980) has argued that the form of explanation in psychology is of a radically different sort than that in the physical sciences. Some nonreductive materialists soften this line on autonomy, allowing psychological explanations to be of the same sort as those in the physical sciences. They might even grant that the lower-level physical sciences can illuminate, and put important constraints on, higher-level theorizing about the mind. Yet because, in their view, mental theorizing occurs at a higher level, it will still retain a degree of autonomy.

References

Chalmers DJ (1996) The Conscious Mind. New York, NY: Oxford University Press.

Damasio A (1994) Descartes' Error. New York, NY: G. P. Putnam.

Davidson D (1980) Essays on Actions and Events. Oxford: Clarendon Press.

Fodor JA (1981) Special sciences. In: *Representations*. Cambridge, MA: MIT Press.

Foster J (1991) The Immaterial Self. London: Routledge.

Hoffman J and Rosenkrantz G (1991) Are souls unintelligible? *Philosophical Perspectives* **5**: 183–212.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Kim J (1993) *Supervenience and Mind*. Cambridge, UK: Cambridge University Press.

Kim J (2001) Lonely souls. In: Corcoran K (ed.) *Soul, Body, and Survival*. Ithaca, NY: Cornell University Press.

Lowe EJ (1996) *Subjects of Experience*. Cambridge, UK: Cambridge University Press.

Ryle G (1949 [Reprinted 1984]) *The Concept of Mind*. Chicago, IL: University of Chicago Press.

Yablo S (1990) The real distinction between mind and body. *Canadian Journal of Philosophy*, Supplement **16**: 149–201.

Further Reading

Baker LR (1993) Metaphysics and mental causation. In: Heil J and Mele A (eds) *Mental Causation*. Oxford: Clarendon Press.

Hasker W (1999) *The Emergent Self*. Ithaca, NY: Cornell University Press.

Hill CS (1997) Imaginability, conceivability, possibility and the mind–body problem. *Philosophical Studies* **87**: 61–85.

Kim J (1999) Making sense of emergence. *Philosophical Studies* **95**: 3–36.

Levine J (2000) *Purple Haze: The Puzzle of Consciousness*. New York, NY: Oxford University Press.

Lowe EJ (2000) *An Introduction to the Philosophy of Mind*. Cambridge, UK: Cambridge University Press.

Merricks T (1994) A new objection to *a priori* arguments for dualism. *American Philosophical Quarterly* **31**: 81–85.

Shoemaker S (1984) *Identity, Cause, and Mind*. Cambridge, UK: Cambridge University Press.

Zimmerman D (1991) Two Cartesian arguments for the simplicity of the soul. *American Philosophical Quarterly* **28**: 217–226.

Dynamical Systems, Philosophical Issues about

Intermediate article

James Garson, University of Houston, Houston, Texas, USA

CONTENTS

Dynamical systems theory provides an alternative to the dominant paradigm in cognitive science that claims human intelligence results from digital computation. The dynamical account of cognition avoids symbolic representation, and stresses the importance of modeling human interactions with the external environment through time.

INTRODUCTION

Dynamical systems theory (DST) has provided a new paradigm for understanding complex systems. The concepts and methods developed to describe nonlinear dynamics have rapidly spread from physics, to chemistry, biology, neurology, ecology, geography, economics, and political science. It is natural to think that the same methods might contribute to cognitive science. The dynamical systems hypothesis (DSH) proposes that the methods and results of DST provide genuine new insights into the nature and explanation of cognition – insights that are missed by the traditional approach in cognitive science, which is to view human intellect as the product of symbolic computation. The differences between the symbolic and dynamical viewpoints raise a number of important philosophical issues. These include the role of symbolic representations, the importance of interaction with the environment, and the way in which time is treated in cognitive models.

WHAT ARE DYNAMICAL SYSTEMS?

Dynamical systems are models of the phenomena of nature that employ the methods of DST. DST is not so much a theory as a body of conceptual and mathematical tools that provide insight into the complex behavior that is generated in nonlinear systems, where solving the equations describing how properties (variables) change through time can be difficult or impossible. DST has developed concepts and graphical techniques for describing the qualitative nature of such systems. Where mathematical solutions for the equations that describe the changes in the variables are unavailable, progress can still be made by simulating a system's behavior on computers and by applying general knowledge drawn from the study of related systems. A main focus of this work is to identify a range of stable and unstable behaviors that the system takes on in responding to forces that affect it. (*See* **Dynamical Systems: Mathematics**)

Phase Space

The concept of phase space (or state space) is a foundational notion in DST. The phase space for a system has a separate dimension devoted to each variable in its equations. For a simple pendulum, for example, the phase space might have two dimensions, one for the position (x) and one for the velocity (v) of the pendulum bob. A phase space describing the brain's neural activity, on the other hand, might include a separate dimension for the firing frequency of each of over 100 billion neurons.

A point in a two-dimensional space indicates a value for each of the two dimensions. For example, the point (x, v) in the phase space for a pendulum might indicate that the position of its bob is x meters away from vertical, and that its velocity is v meters per second. If we give a pendulum bob a push, it will swing back and forth. This oscillating behavior corresponds to a moving point in phase space, with the x and v values of the point changing as the pendulum bob takes on new positions and velocities. The path this point takes is called the *phase trajectory*. Assuming no friction, the trajectory

of a pendulum forms a closed loop. If there is friction, it forms a curve that spirals towards the origin ($x = 0$, $v = 0$), indicating that the pendulum bob is ultimately motionless in a vertical position.

Attractors

Most dynamical systems can be described by 'attractors', which are points, lines or higher dimensional surfaces in the phase space that represent stable motions of the system. A system that is disturbed so that its phase trajectory is moved away from an attractor will soon return to it. There are several fundamentally different kinds of attractors. In the case of a stable attractor, the phase trajectory evolves towards a point in the phase space. The resting position of the pendulum under friction is a good example, for here the system has reached a point of stable equilibrium. A second kind of trajectory is called a limit cycle. In this case, the pathway starting from initial conditions settles into a repeating sequence. For example, the friction-free pendulum traces a limit cycle. A third kind of attractor is the torus, a donut-shaped surface. A phase trajectory on a torus is a curve that may never loop back on itself. 'Strange attractors' are the signature of chaos. These attractors are very complex, and tend to 'fill' large regions of their phase space. Chaotic systems are practically impossible to predict because the slightest deviation in the values of their variables will be quickly magnified into large differences in outcome.

Chaos and Complexity

DST is sometimes called 'chaos theory', but the use of the world 'chaos' is misleading for two reasons. Firstly, it suggests randomness or disorder, yet many chaotic dynamical systems spontaneously create structure. Secondly, chaos is only one of several types of behavior studied by DST. It is not even accurate to refer to a system as chaotic, since chaotic behavior may come and go depending on differences in the system's parameters. Although chaos is not a central feature in most dynamical models of cognition, the possibility of chaotic and near-chaotic behavior has interesting implications.

Dynamical systems described by even very simple equations generally have complex trajectories. Such trajectories display a kind of richness of behavior that defies accurate characterization by any digital program or set of rules small enough to be actually written out or understood by humans. (Although the equations of the system might be easy to state, only an impracticably large and fast digital computer could accurately predict the behavior determined by those equations.) The existence of such chaotic and near-chaotic complex behavior in dynamical systems is an important consideration in understanding the contribution of DST to cognitive science. (*See* **Real-valued Computation and Complexity**)

Self-organization and the Emergence of Order

In nonlinear dynamical systems, an activity that is seemingly disordered at the lowest level can still produce coherent and stable large-scale structures. Chaotic models of Jupiter's famous red spot suggest that such structures can persist autonomously for a long time, and that they do not need any special mechanism to create or maintain them (other than the natural behavior of the system as a whole). Contrary to our intuition that forms of order must always cancel out in a highly energetic and unguided system such as the atmosphere of a gigantic planet, DST shows that structures such as the red spot emerge naturally from the dynamics of gases on the surface of a heavy rotating sphere. It has been proposed that cognition may depend on the same kind of emergent self-organization. (*See* **Emergence**)

WHAT IS THE DYNAMICAL SYSTEMS HYPOTHESIS?

The dynamical systems hypothesis is that cognition is a dynamical system and so is best understood using the concepts and methods of DST. The DSH is often contrasted with the hypothesis that cognition should be modeled on the symbolic operation of a digital computer. The conflict between the two hypotheses centers around the concepts and methods of explanation they advocate. Symbolic modelers attempt to model human knowledge by postulating the existence of symbolic representations in the brain that record information about the external world. They also assume the existence of programs or sets of rules that the brain uses to transform these symbolic data into new forms. These program-guided transformations are thought to account for cognitive processes such as conceptualizing, planning, reasoning, decision making, and eventually action. (*See* **Symbolic versus Subsymbolic**; **Symbol Systems**)

Dynamicists, on the other hand, view the brain as a dynamical system which is continually affected by interaction with its environment. Dynamical representation is not symbolic, but is understood

via the concepts that DST uses to understand structures in phase space, including equation parameters, trajectories or their parts, and attractors.

Levels of Description and the DSH

A common misconception is that the DSH demands that cognitive explanation be carried out in terms of variables for brain features such as the firing rates of neurons, and the strength of the connections between them. Such connectionist models do fall within the DSH. However, dynamical models can also be constructed at higher levels of description. For example, Townsend and Busemeyer (1995) have applied DST to characterize decision making. The variables of their models are clearly cognitive, for they measure the gains or losses expected for the various factors (or dimensions) that are relevant to the decision. (*See* **Connectionism**)

The Role of Representation in the DSH

One of the most important differences between symbolist and DSH approaches concerns the roles that representations play in the two paradigms. The symbolist theory assumes that cognition is correctly described as the execution of programs or subroutines, which are defined over representations. Here representations are directly implicated in a causal explanation of cognitive performance. On the DSH approach, the analogs of programs are the equations that govern the evolution of a system. These equations are not ordinarily defined over representations, but instead over the variables for the system's properties. From the DSH point of view, representations are not essential to an explanation of the mechanisms of cognition, since what matters is the way in which system variables evolve according to the system's equations of motion. Although dynamical explanation may mention representations, these are conceived of as emergent aspects of system activity.

An Illustration of the Difference between Symbolist and DSH Models

An illustration may help to explain the difference between symbolical and dynamical conceptions of representation. Consider two different ways of constructing a thermostat that controls a furnace. The 'symbolic' thermostat would collect temperature information about the outside world with sensors, and convert that information into numbers that are stored in a memory *M* of a small computer. A program running on that computer includes instructions such as 'if the number in *M* is smaller than 20 then turn on the furnace burner'. On the other hand, the 'dynamical' thermostat would consist of a device that connects the furnace to a bimetallic strip in such a way that as air temperature falls, the strip bends, thereby opening a valve that sends more fuel to the furnace's burner.

In the symbolic thermostat, digital representations of the temperature in the room interact with program commands to control the furnace. It would be inappropriate, however, to characterize the dynamical thermostat on the computational model, since no computation over data representing the world is performed. A dynamical model would provide equations describing the interactions between variables for room temperature and valve position. Although one might claim that the amount of bend in the bimetallic strip carries information about (or even represents) the temperature of the room, a satisfactory explanation of how the system works can be made without treating the strip as explicitly representational. The state of the strip is not symbolic data to be processed by a program; it is a part of the mechanism that directly ensures that room temperature and valve opening interact in the right way. Here, in contrast to the symbolist thermostat, we lack any meaningful distinction between the data and the procedures that operate on the data.

Time and Interaction in the DSH

The example of the dynamical and symbolic thermostats helps illustrate other features that distinguish the DSH from the symbolist hypothesis. The DSH promises to provide a richer framework for understanding the evolution of a system through time. The dynamical thermostat can be modeled with a system of equations that specify the relationships between room temperature, the furnace valve position, and its effect on room temperature. These equations specify the rates at which the variables change. A feedback system of this kind displays a rich variety of behavior including oscillations and overshoot when the feedback is too strong or delayed. A symbolist model does not have the well-developed temporal concepts that would explain such dynamic behavior: it does not explicitly treat time as a graded or real-valued quantity. The rate at which quantities change is largely ignored. When time is mentioned at all, it is in description of the sequence of steps a computer performs as it follows its program. Since time is treated as a sequence of discrete moments, rather

than a continuous quantity, symbolist models have difficulty capturing interactions with the environment that depend on how quickly system variables change.

A second difference is that the symbolist has a tendency to view the system as isolated from the environment, being driven by its representations of the world rather than by the world itself. Dynamicists will claim that a better way of understanding cognition is to model the interaction with the world directly.

ARGUMENTS FOR THE DYNAMICAL SYSTEMS HYPOTHESIS

The DSH is in its infancy, and there is not yet conclusive evidence in its favor. However, the available evidence suggests that the DSH is well worth exploring. Firstly, the DSH has a record of success across a wide range of cognitive abilities, including perception, sensorimotor activity, language, attention, decision making, and development (Port and van Gelder, 1995). Secondly, as explained above, the treatment of time and interaction with the environment is richer and more natural in the DSH. Two further arguments for the DSH are worth discussing in detail.

Problems in the Symbolist Paradigm

Certain problems have emerged for the symbolist paradigm. Research programs in artificial intelligence that attempt to symbolically approximate human thought have met with difficulties. The complexity of symbolic programs increases rapidly as rules are elaborated to handle an endless stream of discovered exceptions. Disenchantment with symbolic methods, and interest in alternatives, has arisen from a belief that symbolic methods have failed.

According to the DSH, this failure is a symptom of the fact that the dynamics of the brain embody such complex information processing that even a marginally accurate symbolic approximation would require impossibly complex programs. Although cognition is clearly ordered in many respects, it may be that this order can be expressed in compact form only as 'soft laws' (Horgan and Tienson, 1990). Soft laws incorporate ineliminable exceptions ('*ceteris paribus*' clauses) that reflect the subtlety and flexibility found in human thought. If cognitive regularities are truly soft in this way, then attempts to characterize cognition in the form of programs that embody strict rules are bound to fail.

Even defenders of the symbolist paradigm have voiced doubts about whether symbolist methods can capture higher-level cognition. Given the persistent failure of the symbolist paradigm to present a plausible theory of the flexibility of thought and language, there is every reason to seek an alternative.

The Power of Dynamical Processing

One source of evidence for the DSH comes from the study of dynamical computation (Kauffman, 1993, chapters 5–6). Crutchfield and Young (1990) have examined the computational complexity of high-level system behavior over a range of underlying system dynamics. When a dynamical system exhibits short limit cycles, its computational powers are weak. At the interface between long limit cycles and chaos, the power increases to a maximum, while deeper in the chaotic realm, computational power wanes. Kauffman (1993, p. 221) modeled the evolution of dynamical systems whose survival depends on the ability to solve a simple task. His results lead to a similar conclusion: higher fitness corresponds to complex dynamical behavior at the 'edge of chaos'. Furthermore, he gives evidence that systems evolve more quickly when they operate in this realm.

Work on the superiority of analog over digital computation underscores the potential information processing advantages of dynamical models. Blum *et al*. (1989) have shown that analog computation provides a vast increase in the range of processing available to the brain, and Siegelmann and Sontag (1994) have shown the superiority of analog computation in models of noisy neural networks. Such evidence is far from conclusive, but it suggests that highly complex dynamical processing may account for the intelligent adaptability and avoidance of overly rigid or stereotypical behavior that characterize human cognitive performance.

PROBLEMS WITH THE DYNAMICAL SYSTEMS HYPOTHESIS

Since the DSH challenges an established paradigm, it is not surprising that it faces strong criticism. Some of the main objections are described below.

Problems in Dynamical Modeling

Modeling cognition with DST is not easy. Defining the relevant variables and the dynamical equations that govern them is a daunting task that too often depends on guesswork. Some tools available in

DST allow the researcher to investigate dynamical models even when the actual equations of motion are not known in detail. However, a fundamental problem remains. Even when one has the luck to produce a model that approximates fairly well the behavior to be explained, it can be difficult to determine whether that success is genuine or the result of 'fudge factors' written into the equations.

The 'Wrong Level' Objection

It has been claimed that the DSH attempts to explain cognition at the wrong level. Perhaps DST is useful in giving an account of the behavior of the brain in physical, chemical, or neurological terms, but this does not help us make progress in cognitive science. What is needed is a theory that explains human action in terms of perceptions, memories, beliefs, desires, reasons, plans, and goals. Although some aspects of perception and motor control may be illuminated by DST, the understanding of the central processes that govern human intelligence would seem to require a very different vocabulary from what appears to be available to DST.

One response to this objection has been outlined already, namely that DST is not limited to models framed at the neurological level. Furthermore, work on central processes within DST is in its infancy. Only time can tell whether DST models of a full range of higher cognitive abilities will be successful, either by employing variables for concepts already available in cognitive psychology, or by providing the tools that will allow us to construct more fruitful cognitive-level concepts.

The Systematicity Objection

One of the most widely discussed criticisms of the DSH appeals to the systematicity of higher-level cognitive abilities such as language and reasoning. The systematicity of language means that the ability to produce and understand some sentences is intrinsically connected to the ability to produce and understand others of related form. For example, no one with a command of English who understands 'John loves Mary' can fail to understand 'Mary loves John'.

The systematicity objection proposes that such facts can be explained only by assuming that cognition operates over symbolic strings (such as 'John loves Mary') composed of constituents ('John', 'loves', and 'Mary') that can be combined in different ways. A speaker of English computes the meaning of a string from the meanings of its constituents. If this is so, then understanding 'Mary loves John' can be accounted for as another instance of the same process. Some have claimed that no alternative solution can work, so that symbolic processing of representations with symbolic constituents is required to explain cognition.

It is indeed difficult to explain systematicity without adopting a model that explicitly employs symbolic constituents. Research in DST (van Gelder and Port, 1994) offers some hope that dynamical systems can create the needed combinatorial structures without implementing symbolist structures. However, a convincing response to the systematicity challenge will require more research.

PHILOSOPHICAL ISSUES ABOUT THE DYNAMICAL SYSTEMS HYPOTHESIS

The mission of cognitive science is to explain the highly flexible structure embodied in human intelligence. The DSH requires us to reconsider both the nature of that structure and the methods we use to explain it.

Representation

The DSH undermines common presuppositions about representation in the brain. Symbolists tend to think that having representations of the world is a prerequisite for having knowledge about it. The DSH stresses the idea that intelligence is the result of an ongoing interaction with the world that may not require explicit storage of information. Symbolists tend to think of representations as static entities waiting to be processed by programs. Dynamicists tend to view representation as the product of the system's dispositions to interact with the world as reflected in its attractor 'landscape'. The notion of a phase space in DST provides a framework for understanding the relationships between different representational states. Since representational features are defined in terms of a high-dimensional space, the notion of the distance or similarity between representations is well defined. Within this framework, the features of representations that account for their cognitive roles can be understood in terms of DST concepts that help characterize the similarities and differences to be found in phase space. (*See* **Language, Connectionist and Symbolic Representations of**)

Soft Laws

We tend to expect a science to provide us with the laws of its domain; so that cognitive science should

discover the laws of human thought. But the DSH may require us to revise our understanding of the nature of laws, and hence of the goals of cognitive science. It has been widely noticed that the laws of psychology resist formulation. Some have concluded from this that the study of cognition can never meet the requirements of science; but the DSH suggests that we may need to revise our conception of laws in cognitive science. Soft laws incorporate '*ceteris paribus*' clauses, conceding the existence of exceptions. The DST perspective suggests that such laws may be preferable to hard laws in psychology and other sciences.

The Nature of Explanation in the DSH

The DSH also invites a new conception of explanation in cognitive science. To illustrate this point, let us contrast symbolist and dynamical explanations for how the leopard gets its spots. The symbolist assumes that the genetic code specifies a template for leopard spot patterns that guides the way spots are formed in leopard skin during development. The fetus incorporates a program that reads this template and fixes the colors that fetal skin cells will express. On this paradigm, explaining how the leopard gets its spots amounts to locating the template and the mechanism that reads it. The dynamicist's explanation is very different. The chemistry of fetal leopard skin can be modeled by a nonlinear dynamical system that spontaneously creates spots in leopard-like patterns. So, in a sense, the spot shapes have no explanation apart from the observation that the right dynamical systems make those patterns naturally.

This dynamical explanation may appear vacuous. The symbolist explanation accounts for the mechanism that forces fetal leopard skin to form the right pattern of spots. However, if the dynamicist is right, the project to find a template and program is bound to fail. The lesson of dynamical systems theory is that very complex patterns can emerge spontaneously from the seemingly random interaction of simple units such as skin cells. Patterns may emerge without symbolic guidance.

The same moral may apply to cognition. From the DSH perspective, intelligence is the product of self-organized structures that arise naturally. Searching for programs and data that guide the brain is like searching for a force that keeps objects moving at a constant velocity. No explanation for how cognitive systems behave is necessary beyond merely pointing out that the dynamical systems responsible spontaneously create that behavior.

THE DYNAMICAL SYSTEMS HYPOTHESIS IN COGNITIVE SCIENCE

The relationship between the DSH and alternative paradigms in cognitive science is complex.

Connectionism and the DSH

Connectionism (rather than DST) has so far presented the most popular alternative to the symbolist approach. Connectionists attempt to construct models of interconnected neurons in the brain that are capable of cognitive tasks. While some connectionists accept the thesis that the brain relies on symbolic processing, radical connectionists seek to show that neural networks can produce cognition using non-symbolic methods. Thus both dynamicists and radical connectionists are opposed to symbolism. Furthermore, neural networks are examples of dynamical systems, and some connectionists employ concepts from DST to help understand their behavior.

Symbolism and the DSH

It can be shown that a digital computer can be treated in DST as a special case of a dynamical system. Furthermore, dynamicists typically use digital computers to investigate the behavior of their models, so it may appear that everything they discover also has a symbolist account.

What is Unique about the DSH

However, the fact that symbolist and connectionist models can be treated as dynamical systems does not imply that the DSH is compatible with either paradigm. The main feature that distinguishes the three accounts is found not in *what* they model but in *how* they model. The DSH proposes that the concepts and methods of DST are what is needed to make progress in cognitive science. Symbolism employs entirely different methods and concepts. While some connectionists use tools drawn from DST to obtain insight into the functioning of their neural networks, they differ from dynamicists in their vision of cognition. The DSH views the mind as a structure in constant resonance with changes in the environment. It seeks an explanation for cognition in the complex and often self-organized structures that emerge from systems of nonlinear equations that determine how variables change value smoothly through time. Only time will tell which of these views is the right one or whether

more than one of them may be needed to give a full account of the richness of the human mind.

References

Blum L, Shub S and Smale S (1989) On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society* **21**(1): 1–46.

Crutchfield J and Young K (1990) Computation at the onset of chaos. In: Zurek W (ed.) *Complexity, Entropy, and Physics of Information*, pp. 223–269. Redwood City, CA: Addison-Wesley.

van Gelder T and Port R (1994) Beyond symbolic: prolegomena to a kama-sutra of compositionality. In: Honavaar V and Uhr L (eds) *Aritficial Intelligence and Neural Networks: Steps Towards a Principled Integration*. pp. 107–125. New York, NY: Academic Press.

Horgan T and Tienson J (1990) Soft laws. *Midwest Studies in Philosophy* **15**: 256–279.

Kauffman S (1993) *The Origins of Order*. New York, NY: Oxford University Press.

Port R and van Gelder T (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Siegelmann H and Sontag E (1994) Analog computation via neural networks. *Theoretical Computer Science* **131**: 331–360.

Townsend J and Busemeyer J (1995) Dynamic representation of decision making. In: Port R and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 101–120. Cambridge, MA: MIT Press.

Further Reading

Clark A (1997) *Being There: Putting Brain Body and World Together Again*. Cambridge, MA: MIT Press.

Elman J (1995) Language as a dynamical system. In: Port R and van Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 195–225. Cambridge, MA: MIT Press.

Gleick J (1987) *Chaos: Making the New Science*. New York, NY: Viking.

Gregson R (1995) *Computation, Dynamics and Cognition*. New York, NY: Oxford University Press.

Horgan T and Tienson J (1996) *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.

Kelso J (1995) *Dynamic Patterns: The Self Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Langton C (1991) Computation at the edge of chaos. In: Forrest S (ed.) *Emergent Computation*, pp. 12–37. Cambridge, MA: MIT Press.

Murray J (1988) How the leopard gets its spots. *Scientific American* **258**(3): 80–87.

Port R and van Gelder T (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Skarda C and Freeman W (1987) How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* **10**: 161–195.

Eliminativism

Intermediate article

William Ramsey, University of Notre Dame, Indiana, USA

CONTENTS

Eliminativism is the radical thesis that the common-sense conception of the mind is deeply mistaken and that certain mental states posited by it do not exist.

WHAT IS ELIMINATIVISM?

Eliminativism, or eliminative materialism as it is often called, is the thesis that our common-sense conception of the mind is deeply mistaken and that our ordinary notions of mental states will not belong to a scientifically respectable account of cognitive phenomena. Put more simply, it is the view that certain mental states, such as beliefs and desires, do not exist. Eliminativism is one of the most radical and controversial philosophical positions ever held. In many respects, it goes beyond other forms of scepticism since it challenges deep assumptions about the workings of our own minds. In his *Meditations*, Descartes offered a famously extreme form of doubt, questioning many of our beliefs about ourselves and the world. Nevertheless, he insisted that we can at least know with certainty that our minds contain thoughts and beliefs. Eliminativism claims not only that it is possible to doubt the existence of such mental states, but that it is correct to do so.

Eliminativism involves two central and controversial claims. It is worth examining each of these in some detail. For the sake of brevity, much of this article will focus upon our notion of belief, since it figures so prominently in discussions of eliminativism. Many of the arguments concerning belief are thought to generalize to other mental notions as well, especially other propositional attitudes.

Folk Psychology and the Theory Theory

The first claim of eliminativism is that we employ a theoretical framework to explain and predict intelligent behavior. This position is commonly called the 'theory theory', since it maintains that our 'folk' or 'common-sense' psychology is similar to other folk theories that we use to understand a range of different phenomena. As with most theories, folk psychology is assumed to consist of both universal generalizations (or laws) and theoretical posits, with the former capturing the counterfactual regularities found among the latter. For example, a generalization of folk psychology might be something like:

> If someone has the desire for X and
> the belief that the best way to get X is by
> doing Y, then (barring certain conditions)
> that person will tend to do Y. (1)

According to the theory theory, these generalizations function like the laws of scientific theories, though they are learned and used in a far more informal manner.

The theory theory maintains that the posits of folk psychology are the mental states that figure in our common-sense psychological explanations. As theoretical posits, these states account for observable effects (like statements and behavior), but are not themselves directly observed. If we concentrate on our folk notion of belief, the theory theory claims that common sense assigns two sorts of properties to these states. First, there are various causal properties. We assume beliefs are caused in certain circumstances, interact with desires and other cognitive states in certain ways, and bring about certain kinds of behavior. These causal roles help define and distinguish beliefs from other types of mental states. Second, beliefs are essentially about a wide range of different states of affairs. This inherit 'aboutness', or 'intentionality' (also sometimes called 'meaning', 'content', or 'semantic properties'), is commonly regarded as a unique feature of the mind. While conventional signs and linguistic symbols are meaningful, their meaning is derivative, stemming from the stipulations and interpretations of thinking creatures. Only beliefs and other mental representations have what is

called 'original' or 'intrinsic' intentionality: a type of meaning not derived from other sources.

Eliminative Theory Change

The second claim of eliminativism is that folk psychology is severely mistaken about the actual nature of the mind. Eliminativists argue that the laws and posits of folk psychology radically misdescribe our minds; consequently, they denote nothing that is real. To understand eliminativism, it is important to note the difference between two types of theory change: reductive or retentive theory replacement, on the one hand, and eliminative, or ontologically radical, theory replacement, on the other. In the case of the former, the posits of a rejected theory – be it a folk theory or a former scientific theory – find a new home, perhaps with some modifications, in the superseding theory. For example, the notion of a planet survived the transition from Aristotelian physics to Newtonian physics, though it underwent some changes as Aristotle's notion was clearly mistaken in many respects. Despite problems with the Aristotelian framework, we did not conclude that it was so wrong that the notion of planet has no referent.

This can be contrasted with eliminative theory change, where a theoretical posit of an abandoned theory so misses the mark that we conclude that the notion fails to capture anything real. For example, early explanations of strange behavior posited the existence of malevolent spirits that were thought to possess the souls of unlucky individuals. As the theory of demonology was replaced by more sophisticated accounts of mental and neurological disorders, the notion of a supernatural demon was abandoned altogether. Since there is nothing in the new accounts with which demons can be reasonably identified, we have eliminated demons from our contemporary ontology.

Eliminativists predict that this sort of eliminative change will happen with the theoretical posits of folk psychology. The claim is not that mental states, such as beliefs, currently exist but will somehow be abolished as our scientific understanding of the mind grows. The claim is that mental states such as beliefs have never existed; hence, it is the concept of a belief that will be eventually eliminated from our mental taxonomy. There simply are no such things that have the causal and semantic properties we attribute to beliefs.

Eliminativism shares assumptions with both physicalism and dualism. Eliminativists agree with the physicalist claim that the actual causes of behavior are ultimately physical events, taking place inside the brain. However, eliminativists also agree with dualists that the mental states posited by common-sense psychology are not to be identified with anything inside the brain. Of course, whereas dualists hold this view because they think mental states are nonphysical, eliminativists hold this view because they think these states do not exist.

ARGUMENTS FOR ELIMINATIVISM

As one might expect, arguments for eliminativism typically take the form of arguments against folk psychology. By and large, these arguments fall into two groups. The first group involves general theoretical considerations that are relevant to the evaluation of any theory. The second group focuses upon problematic aspects of folk psychology itself.

Most of the arguments against folk psychology based upon theoretical considerations can be found in the writings of Paul and Patricia Churchland, perhaps the two strongest defenders of eliminativism. One such argument starts with the premise that a promising theory should offer a fertile research programme with considerable explanatory power and range. It then notes that folk psychology appears stagnant, making no real progress throughout much of history. Worse yet, there are many cognitive phenomena that folk psychology does not even begin to explain. Important aspects of the mind such as the nature of consciousness, the oddities of mental illness and the actual mechanisms of learning are all left unexplained by folk psychology. These and similar considerations suggest that folk psychology is ripe for replacement. Another argument looks at the track record of folk theories in general and notes how improbable it would be for this one to turn out true. Folk physics, folk biology, folk epidemiology, and so on, all proved to be radically mistaken. Folk psychology concerns a subject that is far more complex and difficult than these others. Hence, it seems implausible that just this one is right (Churchland, 1981).

In response to these theoretical arguments, many argue that folk psychology actually has stimulated a number of fruitful research programs in scientific psychology. Moreover, some of these do help explain a wide variety of cognitive phenomena not directly explained by folk psychology. Furthermore, it is one thing to claim that a folk theory is incomplete or unproductive, but another thing to claim that it is radically false (Horgan and Woodward, 1985). Defenders of folk psychology claim that these theoretical considerations cannot possibly outweigh the evidence provided by the

everyday, ordinary experience of our own minds, which seems to support the reality of beliefs.

Regarding this last point, eliminativists often warn that we should be deeply suspicious of introspective 'evidence' about the inner workings of the mind. Since all forms of observation – including introspection – are, to some degree, 'theory-laden', our folk-theoretical framework may play a larger role in forming the content of our experienced inner lives than actual mental processes. Indeed, there is considerable empirical evidence that we lack reliable access to the actual workings and nature of cognitive processes. Thus, our inner observations of certain mental states may be like the observations of those who claimed to be able to 'observe' demonic spirits.

The second group of arguments looks at features that are unique to folk-psychological posits and challenges the likelihood that these features will be accommodated by a scientific account of the mind. The most widely discussed features are those associated with the alleged linguistic nature of beliefs and other propositional attitudes. Common sense appears to treat these states as having a form similar to public language sentences, with a compositional structure and syntax. For instance, the belief that John loves Mary appears to be composed of the concepts (or mental 'words') 'John', 'love', and 'Mary', and it differs from the belief that Mary loves John by virtue of something analogous to syntactic arrangement. Along with this quasilinguistic structure, beliefs resemble public sentences in another way: they have semantic properties. Beliefs, like sentences, are individuated by virtue of what they are about.

Both of these quasilinguistic features of propositional attitudes – their alleged sentential structure and their semantic properties – have motivated arguments in favour of eliminativism. With regard to the former, some writers have emphasized the apparent mismatch between the sentential form of propositional attitudes and the actual neurological structures of the brain. Whereas the former involves discrete symbols and a combinatorial syntax, the latter involves action potentials, spiking frequencies and spreading activation. It is hard to see where in the brain we are going to find sentence-like structures (Churchland, 1986).

Of course, this argument depends upon a certain view of folk psychology, namely, that it treats beliefs as having a sentential structure. Those who reject that claim, as many have, will not find the argument compelling. Even for those who find this reading of folk psychology plausible, there is a further difficulty regarding the relevance of neuroscience for determining the status of folk psychology. Many have insisted that just as the physical circuitry of a computer is the wrong place to look for computational symbol structures, so too, the neurological wiring of the brain is the wrong place to look for structures that might qualify as beliefs. Instead, the status of folk posits should be decided at a higher, more abstract level of analysis than the neurophysical. For many, the computer model of the mind – where the mind is regarded as the brain's program, abstracted from the neurological details – provides a more appropriate level at which to seek analogs to the posits of folk psychology.

The second type of argument against beliefs – focusing upon their intentional or semantic properties – concludes that folk-psychological concepts are inappropriate for any scientific theory of the mind, including a computational one. The main proponent of this argument has been Stephen Stich (1983). Stich claims that folk psychology individuates beliefs by virtue of their semantic properties; however, there are several reasons for thinking that a semantic taxonomy is ill-suited to scientific psychology. Firstly, because such a taxonomy depends upon matters outside the head, it will individuate mental states differently from other taxonomies, such as those based upon causal roles. Secondly, Stich has argued that the semantic content of a belief is ascribed on the basis of judgements of similarity. Consequently, these ascriptions are vague, and fail with subjects who are too dissimilar from ourselves, such as the very young and the mentally ill. Rather than adopt a folk-psychological ontology, Stich argues a scientific psychology should employ a syntactic taxonomy: one that individuates mental states by appeal to their purely non-semantic properties.

As Stich himself notes, even if folk posits are inappropriate for a scientific psychology, it need not follow that they do not exist. After all, there are plenty of entities that are vaguely defined or demarcated in ways that are inappropriate for science, but are nevertheless real. If our best scientific account posited states that share many features with beliefs, such as similar causal roles, then even if the two taxonomies differed in certain cases, we may still regard folk psychology as, in some sense, vindicated.

To generate a more robust form of eliminativism, we would need to show that there is nothing in our scientific psychology that shares the central properties we attribute to beliefs. In a more recent paper, Ramsey, Stich and Garon (1990) argue that certain connectionist models of memory and

inference could serve as the basis for this stronger eliminativist claim. These connectionist models store information in a holistic, or distributed, manner; thus, there are no causally discrete, semantically evaluable data structures that represent specific propositions. It is not just that these models lack the sentential, compositional representations assumed in more traditional (or 'language of thought') models. Rather, in these networks there are no causally distinct structures that stand for anything specific. Consequently, there do not appear to be any structures in these models that might serve as candidates for the reduction of propositional attitudes. If these connectionist models should prove accurate, they argue, it will show that there are no such things as beliefs and propositional memories. This argument assumes that in a distributed network, it is impossible to specify which bits of information are causally responsible for various acts of cognition. Some have responded by insisting that, with highly sophisticated forms of analysis, it is possible to pick out casually relevant pieces of stored information.

While most arguments for eliminativism focus upon propositional attitudes, it should be noted that one philosopher, Daniel Dennett (1988), has endorsed an eliminativist stance towards a very different class of mental states: those commonly referred to as 'qualia'. Qualia are mental states picked out by virtue of their intrinsic qualitative or phenomenal character, like pain states. Such states seem intuitively to have a number of special features, such as being purely private, immediately perceived, and intrinsically subjective. Dennett discusses several cases – both actual and imaginary – to expose ways in which these ordinary intuitions about qualia are inconsistent. In suggesting that our concepts of qualia are deeply confused, he attempts to cast doubt on the reality of these states as they are commonly understood by folk psychology.

ARGUMENTS AGAINST ELIMINATIVISM

As might be expected of a theory that challenges our fundamental understanding of ourselves, eliminativism has been subjected to a wide range of criticisms. Here we will discuss four such criticisms which have enjoyed particular prominence.

Incoherence of Eliminativism

The first criticism claims that eliminativism is incoherent or in some way self-refuting. The charge is that eliminativism itself presupposes the existence of the very thing it claims not to exist, namely, belief states. A typical way this charge is made is to insist on some cognitive capacity or theoretical principle that is somehow required by eliminativism and yet requires the existence of beliefs. One popular example is the capacity to make an assertion, and that to assert something one must believe it. Hence, for eliminativism to be asserted as a thesis, the eliminativist must believe that it is true. But if the eliminativist has such a belief, then there are beliefs, and eliminativism is thereby proved false (Baker, 1987).

However, it is important to note that the eliminativist thesis itself – that there are no beliefs – is not conceptually incoherent. The criticism is not that eliminativism is self-contradictory, but that the eliminativist is doing something that goes against his or her own thesis. In the above example, this act was the making of an assertion: we are required, it is claimed, to believe what we assert. But this is exactly the sort of folk-psychological statement that an eliminativist would deny: a central tenet of the eliminativist position is that various capacities that we think involve beliefs actually don't. Thus the critic has merely begged the question, since eliminativists reject the idea that various acts (such as asserting a thesis or formulating a theory) require the existence of beliefs.

Misrepresentation of Folk Psychology

The next criticism of eliminativism challenges the way it characterizes folk psychology. This criticism has two distinct origins. The first origin is, at least partly, in the writings of Wittgenstein (1953) and Ryle (1949). According to one version of this view, common sense does not treat mental states, such as beliefs, as distinct inner causes of behavior. Instead, they are treated as dispositional states, which are used to adopt a certain heuristic 'stance' towards rational agents. Hence, ordinary talk about mental states should be interpreted as talk about abstract things that, although real, are not candidates for straightforward reductions to discrete neurophysical states. They serve in many non-explanatory endeavors – for example, they allow behavior to be interpreted as rational – without assigning any specific inner causal structure to the mind. Consequently, discoveries about the actual structure of the mind are viewed as irrelevant to the status of folk mental states (Dennett, 1987).

The second origin of skepticism regarding the theory theory is in contemporary cognitive science,

and stems from a different model of common-sense psychological explanation and prediction. This view, known as the 'simulation theory', maintains that we predict and explain behavior not by using a theory, but by simulating what we would do in a comparable situation. According to this account, we predict an agent's behavior by taking our own information processing mechanisms 'off-line' and giving them data about the agent's background and circumstances including possible beliefs and desires. We then (subconsciously) use our own decision-making mechanisms to generate output which can thereby serve as predictions (and, in other circumstances, explanations) of the agent's behavior. If this account is correct, then no theory of the mind is used to explain and predict behavior. And if there is no theory of the mind, then there can be no false theory of the mind (Gordon, 1986).

In reply to these objections, defenders of the theory theory have turned to empirical work that supports their position. For example, the developmental psychologists Henry Wellman and Alison Gopnik have argued that in explaining and predicting behavior, children go through phases of development roughly analogous to the phases one would go through when acquiring a theory. Moreover, it turns out that children come to ascribe beliefs to themselves in the same way they ascribe beliefs to others. These considerations lend at least some support to the idea that our notion of belief is employed as the posit of a folk theory rather than as input to a simulation model (Gopnik and Wellman, 1992).

Success of Folk Psychology

Even if we allow that folk psychology is a theory, a third argument against eliminativism is that folk psychology seems to offer a fairly accurate account of cognitive processes. Apart from the way introspection appears to reveal beliefs and desires, we also use folk psychology successfully to predict and control the behavior of others. If folk psychology is so wrong then why are we so good at dealing with one another when we use it? The success of folk psychology suggests an 'inference to the best explanation' argument against eliminativism: the best explanation for the success we enjoy in explaining and predicting each other's behavior is that folk psychology is roughly true.

The eliminativist response to this argument is that any theory – especially one as near and dear to us as folk psychology – may appear successful even when it radically misrepresents reality. As philosophers of science often point out, when we are under the influence of a theory we discount the anomalies with which the theory struggles as insignificant, and attribute more power to the theory than it deserves. The degree to which folk psychology misdescribes mental phenomena may not be fully apparent until we have an alternative theory in hand.

Stich's Criticism of Eliminativism

The final argument against eliminativism comes from the recent writings of a former supporter, Stephen Stich. Stich's argument is complex, but we can outline it here. We noted above that eliminativism is committed to the thesis that the theoretical posits of folk psychology fail to refer to anything. But what exactly does such a claim amount to? Stich argues that this question is far more difficult than eliminativists have assumed. For example, we might think that reference failure occurs as the result of a certain mismatch between reality and the theory in which the posit is embedded. But how much mismatch is necessary before we can say that a given posit doesn't exist? For a variety of reasons, Stich suggests that this question has no definite answer. Consequently, the question of whether there really are such things as beliefs has no definite answer either, contrary to the eliminativist view. Of course, this seems to raise problems not just for the eliminativist, but for the folk-psychological reductionist as well. But Stich has presented compelling reasons for thinking that the usual terms of the debate over eliminativism need to be reconsidered (Stich, 1996).

ELIMINATIVISM AND COGNITIVE SCIENCE

Research in cognitive science has had an important influence on nearly all of the major debates concerning eliminativism. Arguments for eliminativism typically depend upon specific claims about the nature of common-sense psychology. Cognitive research is critical for determining the truth of these claims; and empirical work on our conception of the mind, including our introspective access to our own mental states, has played a central role in debates over the plausibility of the theory theory. As this research continues, we should expect to gain a deeper understanding of the nature of folk psychology.

Eliminativism requires that the transition from folk psychology to scientific psychology be of a certain type, namely, eliminative rather than reductive. This would require psychological theories that

are hostile to the posits of common-sense psychology. Until recently, few psychological theories have been of this nature. The traditional computational paradigm which has dominated cognitive science makes explicit use of data structures that are naturally regarded as the scientific counterparts to folk mental states. However, if Ramsey, Stich and Garon are right, connectionist models may, for the first time, provide us with a plausible account of cognition that supports the denial of belief-like states. The future of eliminativism may depend upon which of these two major accounts of cognition proves correct.

The influence eliminativism has had on cognitive science research is less clear. Because so many theories in cognitive science presuppose the reality of folk-psychological states, both as explanatory posits and as phenomena to be explained, there is no doubt that a commitment to eliminativism would require a radical shift in cognitive science research. Of course, without a widely accepted, well-confirmed theory of cognition, we are not yet in a position to judge the status of folk psychology. Indeed, given the amount of speculative forecasting built into many eliminativist arguments, it is reasonable to ask why cognitive scientists should even consider it. The answer is that eliminativism reminds us to reconsider the constraints we intuitively impose on potential cognitive theories, and not to reject out of hand any psychological theory that failed to incorporate folk-psychological states. Thus, eliminativism is not so much a threat to cognitive science, as a view that potentially admits a range of new theoretical possibilities.

References

Baker L (1987) *Saving Belief*. Princeton, NJ: Princeton University Press.

Churchland PM (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**: 67–90.

Churchland PS (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett D (1988) Quining qualia. In: Marcel A and Bisiach E (eds) *Consciousness in Contemporary Science*, pp. 42–77. New York, NY: Oxford University Press.

Gopnik A and Wellman H (1992) Why the child's theory of mind really *is* a theory. *Mind and Language* **7**: 145–171.

Gordon R (1986) Folk psychology as simulation. *Mind and Language* **1**: 158–171.

Horgan T and Woodward J (1985) Folk psychology is here to stay. *Philosophical Review* **94**: 197–226.

Ramsey W, Stich S and Garon J (1990) Connectionism, eliminativism and the future of folk psychology. *Philosophical Perspectives* **4**: 499–533.

Ryle G (1949) *The Concept of Mind*. London: Hutchison.

Stich S (1983) *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Stich S (1996) *Deconstructing the Mind*. New York, NY: Oxford University Press.

Wittgenstein L (1953) *Philosophical Investigations*. Oxford: Oxford University Press.

Further Reading

Carruthers P and Smith PK (1996) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Christensen SM and Turner DR (1993) *Folk Psychology and the Philosophy of Mind*. Hillsdale, NJ: Erlbaum.

Dennett D (1991) *Two contrasts: folk craft versus folk science, and belief versus opinion*. In: Greenwood J (ed.) *The Future of Folk Psychology*. New York, NY: Cambridge Univeristy Press.

Feyerabend P (1963) Materialism and the mind–body problem. *Review of Metaphysics* **17**: 49–66.

Fodor J (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Forster M and Saidel E (1994) Connectionism and the fate of folk psychology. *Philosophical Psychology* **7**: 437–452.

Goldman A (1992) In defense of the simulation theory. *Mind and Language* **7**: 104–119.

Greenwood J (1991) *The Future of Folk Psychology*. Cambridge, UK: Cambridge University Press.

Horgan T and Graham G (1990) In defense of southern fundamentalism. *Philosophical Studies* **62**: 107–134.

Rorty R (1970) In defense of eliminative materialism. *Review of Metaphysics* **24**: 112–121.

Sellars W (1956) Empiricism and the philosophy of mind. In: Feigl H and Scriven M (eds) *The Foundations of Science and the Concepts of of Psychology and Psychoanalysis*, pp. 253–329. Minneapolis, MN: University of Minnesota Press.

Smolensky P (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–74.

Wellman H (1990) *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Wilkes K (1993) The relationship between scientific and common sense psychology. In: Christensen S and Turner D (eds) *Folk Psychology and the Philosophy of Mind*, pp. 144–187. Hillsdale, NJ: Erlbaum.

Embodiment

Intermediate article

Ronald Chrisley, University of Birmingham, Birmingham, UK
Tom Ziemke, University of Skövde, Skövde, Sweden

CONTENTS

An understanding of how cognition is realized or instantiated in a physical system, especially a body, may require or be required by an account of a system's embedding in its environment, its dynamical properties, its (especially phylogenetic) history and (especially biological) function, and its nonrepresentational or noncomputational properties.

INTRODUCTION

In recent years a number of researchers in cognitive science and artificial intelligence (AI) have criticized many traditional approaches to modeling, building and understanding cognitive systems as not placing sufficient emphasis on the body or physical realization of such systems. Non-embodied approaches to cognitive science typically involve some or all of the following features, to a greater or lesser extent:

- The belief that cognition is computation, and thus can be understood in an implementation-independent way, allowing cognitive science to proceed independently of biology and neuroscience.
- A search for general-purpose cognitive abilities, not relativized to any particular (biological, sensorimotor, physical) context or need.
- A method of analysis, modeling and design that for the most part ignores temporal aspects of cognition, in that it focuses on behaviors (e.g. chess playing) that are evaluated in terms of 'getting the right answer' rather than exhibiting a particular dynamic profile, and sees cognition as a module that mediates between the deliverances of a causally prior perceptual module and the inputs to an autonomous action system.

In contrast, embodied approaches to cognition typically involve some or all of the following features, again to varying extents:

- Acknowledgment of the role that the body and its sensorimotor processes can and do play in cognition. Some aspects of the system that would, on the traditional view, be considered mere matters of implementation, are instead taken to be crucial components.
- Understanding of cognition in the context of its (especially evolutionary) biological function: to support the activities of the body.
- A view of cognition as a real-time, situated activity, typically inseparable from and often fully interwoven with perception and action.

'Embodied' cognitive science or artificial intelligence, then, refers to a range of loosely affiliated philosophies, explanatory frameworks and design methodologies that strive to redress a perceived neglect of the body in cognitive science.

Since the mid-1980s there has been a rapid increase in interest in embodied cognition (and use of the term 'embodiment'), but there are many aspects of cognitive science and artificial intelligence research conducted in the 1960s and 1970s that take embodiment into account.

ISSUES CONCERNING EMBODIMENT

The issue of embodiment is closely related to, though distinct from, several other issues of recent interest in cognitive science.

Embeddedness

Recognizing the role of the body in cognition facilitates an approach that sees cognition as partly constituted by, or in terms of its relationship to, the environment (Clark, 1997). Thus, embodied cognitive science is naturally related to investigations of externalism (the belief that mental states are partially constituted by states external to the cognizer), situatedness (the importance, in cognitive activity, of a cognizer's location in and relations to the environment), offloading (using aspects of the environment, such as numerals when doing long division, to reduce cognitive load), scaffolding (the assistance

a developing infant gains from, e.g. interacting with adults who already have the ability being acquired), and interactivity (cognitive phenomena, such as turn-taking, which depend crucially on the dynamics of interaction between a cognizer and an object or other cognizer).

Noncomputationalism and Nonrepresentationalism

Once one acknowledges the presence of the body, it is possible to use bodily properties, dynamics and configurations to explain some abilities, behaviors and phenomena that previously were thought to require explanation in computational or representational terms (Varela *et al*., 1991). However, it is still unclear whether such explanations which advert to bodily states are themselves noncomputational and nonrepresentational, or whether they instead invoke a new form of computation and/or representation. It certainly seems that an embodied approach need not be anticomputationalist or antirepresentationalist (Clark, 1997).

Dynamics

Many researchers who have taken an embodied approach have found it useful to turn away from discontinuous, nontemporal, logic-based formalisms and instead use the continuous mathematics of change offered by dynamical systems theory as a way to characterize and explain cognitive phenomena (Port and van Gelder, 1995). Again, while there may be natural affinities between embodied cognitive science and these tools, it is certainly possible to have one without the other.

Biology

Perhaps the most obvious connection between the mind and the body is the brain: surely an important part of understanding the mind is to understand the neurophysiology underlying cognition (Churchland, 1986). But there are other connections with biology as well. For example, some researchers have maintained that one can best understand natural cognitive systems in terms of the biological function and purposes that cognitive faculties served in the phylogenetic development of their bodies (Millikan, 1984). This requires understanding not only bodies, past and present, but the evolutionary niches of such bodies as well. An important constraint, then, on models of human cognition will be whether the proposed architecture is an evolvable one: whether it is the kind of architecture that could have been reached through a process of natural selection, given the conditions known to have been in place in our natural history (Sloman, 2001).

VARIETIES OF EMBODIMENT

There are several dimensions of variation in the views and theories of embodiment currently being considered in cognitive science.

Criteria for Embodiment

Perhaps the most important dimension of variation concerns what criteria must be met for something to be an embodied cognizer: what is to count as a body? One can, partly following Ziemke (2001), distinguish a range of views on this question, from the least to the most constraining.

Physical realization

According to one view, to be embodied is just to be realized in some physical substrate. All work in cognitive science is about embodied systems in this sense: even a virtual web agent must be realized in some physical facts at any given time. Only purely mental entities or spirits would lack this kind of embodiment. Thus, even traditional cognitive science is not as disembodied as some have claimed. For example, the influential notion of a physical symbol system (Newell and Simon, 1976) explicitly acknowledges the requirement that a cognitive system be embodied in this (weakest) sense.

Physical embodiment

Physical embodiment requires that the realizing physical system be a coherent, integral system, that to some degree persists over time. This would rule out virtual web agents, the physical realization of which can be radically distributed over the entire planet, but could still include any conventional robot. A tension arises here: if, as some theorists have argued, human cognition extends into the tools and other physical environmental states we exploit, then the localized, biological body is less relevant to cognitive science than the extended, constantly changing, distributed physical system that at any given time includes parts of our environment as well as parts of our bodies. Thus, active externalism (Clark and Chalmers, 1999) may be incompatible with strong notions of embodiment.

Organismoid embodiment

According to another view, the localized physical realization of the system must share some (possibly

superficial) characteristics with the bodies of natural organisms, in terms of form or sensorimotor capacities, but need not be alive in any sense. The most prominent, and perhaps the most complex, examples of organismoid embodiment are humanoid robots such as Cog (Brooks and Stein, 1994) and Kismet (Breazeal and Scassellati, 2000). Work with these robots is based on the view that research in AI and cognitive robotics, in order to be able to investigate human-level cognition, has to deal with systems that have bodies which, although artificial and possibly non-living, have at least some human-like characteristics. For example, Kismet is a humanoid robot that learns how to visually track objects. To do this, a human trainer must move objects of an appropriate size at an appropriate speed at an appropriate distance in front of Kismet's eyes. If the human trainer moves the tracked object too close to Kismet, Kismet responds by raising its eyebrows in a manner which in humans indicates a startle response. This naturally causes the human trainer to startle in return, which prompts a change in the training parameters of speed and distance. Thus, Kismet's organismoid embodiment, in the form of eyebrows and facial expressions, is an integral part of the human–robot training dynamic which should tend towards homeostasis.

Organismal embodiment

The strongest criterion of embodiment is that the body must not only be organism-like, but actually organismal and alive (e.g. Sharkey and Ziemke, 1998). Of course, this raises the question of what is required for something to be alive. Various answers to this question have been proposed, including the ability to metabolize, reproduce, regenerate, or grow; autonomy; and autopoiesis (e.g. Maturana and Varela, 1980; von Uexküll, 1928).

Other Dimensions of Variation

Another dimension of variation is the extent of the domain of cognition that requires an embodied approach. For example, one can ask whether reference to the body is required only for giving an account of low-level, sensorimotor aspects of human cognition, or if it is required for all forms of human cognition, including reasoning, mathematical thought, and language use (cf. the distinction between 'material' and 'full' embodiment in Nuñez (1999)).

Theorists might also disagree on how radical the effect of taking an embodied approach will be on the concepts, theories and methods of cognitive science. Clark (1999) distinguishes between simple and radical embodiment. In simple embodiment, the framework of traditional cognitive science is retained, and facts about embodiment are treated as mere constraints on theories of, for example, 'inner organization and processing'. Radical embodiment is more ambitious, and 'treats such facts as *profoundly altering the subject matter and theoretical framework of cognitive science*' (p. 348, emphasis in original). Radical embodiment is advocated by, for example, Lakoff and Johnson (1999). These authors claim that a central finding of cognitive science is that the mind is inherently embodied, and that this, together with the other two central findings (that thought is largely unconscious, and that abstract concepts are largely metaphorical), force us to reject not just the Western philosophy of mind, but most or all of Western philosophy, including especially Anglo-American analytic philosophy but also including postmodern philosophy. This throws everything into question: the nature of truth, meaning, time, space, language, rationality, and especially the self.

Finally, one can distinguish between epistemological and metaphysical approaches to an embodied cognitive science. An epistemological approach maintains that concepts concerning the body will be required to understand and explain cognition (even if the cognitive system itself is disembodied); a metaphysical approach maintains either that cognitive processes must be realized in a body or that AI should proceed by making embodied robots, but makes no claim as to whether embodiment must be adverted to in explanations of cognitive activity. For example, research involving the robot Shakey (Nilsson, 1984) was metaphysically embodied, but since its design and explanation focused primarily on Shakey's functional, computational aspects (namely, deliberation and planning) and not on Shakey's embodiment, this research was not epistemologically embodied. Conversely, computer simulations are by their very nature not (metaphysically) embodied, in most senses of 'embodiment', yet many researchers use (epistemologically) embodied simulations to model the crucial role a cognizer's body plays in its activity.

PHILOSOPHICAL CONCEPTIONS OF EMBODIMENT

Before the twentieth century, the most influential view of mind in Western thought was dualistic: the mind was regarded as composed of a separate, extensionless, nonphysical substance. This view

led to many insoluble problems, both philosophical and empirical. For example, how do the mental and physical realms interact? How can we scientifically investigate something that is not in the physical world?

Behaviorism rejected dualism, and thereby opened up the possibility of scientific enquiry into an embodied mind, but it left little room for an understanding of the processes underlying much of mentality: it addressed little of what would be called 'cognition' today. Also, it actively avoided explaining or mentioning experience or consciousness (as did many later cognitivists). More promising steps towards an embodied understanding of mind were taken in the first part of the twentieth century: the relevance of von Uexküll's notion of the body has already been mentioned, but there were several other notable thinkers.

For example, in the 1920s Heidegger (1962) developed a phenomenology that understood human activity not as the result of the manipulation of context-free representations of objects, but as the contextualized experience of the body–environment system. Explicit, decontextualized representation of a hammer as an independent object occurs only when there is some kind of breakdown in the system (e.g. the hammer is too heavy). Dreyfus (1992) elaborated this and other aspects of Heidegger's philosophy into a critique of early, disembodied AI work, calling, for example, for systems that react to the particularities of the current perceptual/action situation rather than ones that attempt to create general-purpose, long-term plans. Although heartened by some aspects of neural or connectionist approaches to understanding the mind, Dreyfus provisionally concludes that, like their symbolic predecessors, connectionist models of cognition suffer from their lack of embodiment, in two ways. Firstly, by not being embedded in a real world with actual bodily concerns, most sophisticated connectionist learning must rely on the intervention of a human teacher, which prohibits the connectionist system from developing its own, genuine, meaningful attitude to the world. Secondly, he speculates that connectionist systems will never generalize in a way that we can recognize as being intelligent and meaningful until they have a form of life sufficiently similar to ours – which requires at least a body of some sort, perhaps even a humanoid one.

In the 1930s, Vygotsky (1978) claimed that language is an inherently socially situated activity, and that one can only understand a child's acquisition of language by recognizing this social context. It follows that inasmuch as social activity is embodied, the development and deployment of linguistic faculties will have to be understood as embodied as well.

While placing less emphasis on social embedding, Piaget (1954) was more explicit about the role of the body in the development of cognitive abilities. For example, his accounts typically made essential reference to the notion of a circular reaction, which in its primary form is the repetition of an activity in which the body starts in one configuration, goes through a series of intermediate stages, and eventually arrives at the starting configuration again. Thus, the kinds of abilities that an organism may acquire depend critically on what circular reactions are possible given that organism's body.

In the 1940s, Merleau-Ponty (1962) made the body central to his phenomenology of mind. For example, he claimed that we have intentions that we do not choose to have, by virtue of our bodies being the way they are. Furthermore, the way we perceive an object is determined by the modes of interaction that our bodies, given the nature of the object, allow. (This idea was a precursor of the notion of affordances (Gibson, 1979).) Even more strongly, Merleau-Ponty saw the body as the necessary medium for our having a world at all, with the nature of the activities of the body determining the nature of what could be experienced in our world. Further, the body could be augmented with tools to further develop the elements of our lived world.

Despite the emphasis that these thinkers placed on the body for understanding the mind and behavior, and partly because of their context outside the Anglo-American tradition, it is only recently that the notion of embodiment has had a significant influence on mainstream cognitive science and AI. Instead, these fields were, from their inception in the mid-1950s, dominated by the computer metaphor for mind (not surprising, perhaps, since the notion of computation had for centuries been a concept of the (human) mental activity of symbol and number manipulation). In particular, in the absence of any suggestion as to how they could constrain empirical investigation and modeling, philosophies of embodiment seemed metaphorical or even unintelligible to many cognitive and AI researchers at that time.

COGNITIVE SCIENCE AND EMBODIMENT

The fields of cognitive science and artificial intelligence have played a central role in the development of an embodied concept of mind and

cognition. It was empirical work in cognitive science and artificial intelligence that allowed the development of a more robust and precise notion of embodied cognition to develop.

In particular, work involving mobile robots, at MIT Artifical Intelligence Laboratory and elsewhere, helped to establish the principles and concepts of embodiment and situatedness as the basis of a new approach to artificial intelligence and (later) an embodied cognitive science. For example, Brooks (1991) and his colleagues were able to get robots to perform tasks in the real world in real time that previously could only be done slowly and inflexibly, if at all. They did this by building robot bodies and robot controllers based on a design called 'subsumption architecture'. Rather than trying to graft a domain- and body-independent planning system onto a perceiving and acting robot, the subsumption architecture approach starts with an initial layer of simple perception–action mediation that implements some low-level behavior (e.g. obstacle avoidance). New behaviors (e.g. exploration) are added by adding further layers, which also mediate between perception and action in a simple way, but inhibit ('subsume') the lower layers when necessary to achieve the desired behavior. In such an architecture, there is no central locus of control, no separate planner, and no central model of the world that all processes must write to and read from in order to act appropriately. Communications between processes are not complex symbolic structures, but numerical values. What computation there is in the architecture is distributed, asynchronous, and non-hierarchical.

From an orthodox computationalist perspective, these design features have their disadvantages, but the designs of Brooks and his colleagues exploit the physical properties of the robots to overcome, or bypass, these limitations. For example, although internal communication between processes is limited, the world itself is often used as a medium for communication between the different layers and mechanisms. Much of what traditional thinking would say is required to perform a task is shown to be unnecessary if one takes advantage of regularities and information provided by the body–environment interaction.

The resulting empirical advances in robot engineering served as a springboard for a development and refinement of the notions and philosophies of embodiment. However, the failure of these approaches to quickly scale up to 'higher-level' aspects of cognition has led many to question the ability of the embodied approach to account for conceptual, abstract reasoning and representation. Kirsh (1991) correctly realizes that the issue is not one of representation per se; Brooks concedes that representation is required for some aspects of cognition. The question instead concerns having concepts: 'the ability to find an invariance across a range of concepts', as Kirsh puts it. As we move up a scale of accounting for ever more sophisticated cognitive activities, at what point must we stop limiting ourselves to designs that are tied to the particularities of the body, and begin to use designs that deploy concepts? Brooks (according to Kirsh) says 'almost never'; Kirsh disagrees, saying that much of not only reasoning and abstract thought but even perception and action must be understood in conceptual terms. Perhaps on a strict, rarefied notion of what concepts are, Kirsh is right: the explanation of concept-involving cognition can or must often go beyond what is provided by the body. But perhaps our very concept of concept needs revising (cf. Lakoff and Johnson, 1999); if we can understand how even full concept possession can be the result of being embodied in a particular way, then perhaps embodied robotics can serve as the model for far more of cognition than mere insect-like behavior.

The joint influence of embodied artificial intelligence research and philosophies and concepts of embodiment has prompted researchers to look for and formulate new forms of explanation for natural cognitive phenomena. For example, Thelen and Smith (1994) give an embodied explanation of the development of walking in infants. Rather than attempt to explain changes in gait as the result of changes in plans, rules or representations, Thelen and Smith give an elegant account that emphasizes changes in bodily factors such as the mass of the infant's leg. Such factors are then related to one another in a dynamical-systems framework, the phase transitions of which are used to explain the stage-like developments in infant walking behavior.

While the relevance of embodiment to research in mobile robotics, as discussed earlier, is obvious, some have claimed that concepts of embodiment are required for us to understand non-robotic artificial computational systems as well. Smith (1996) has argued that we will only be able truly to understand what is going on in ordinary desktop computers when we understand how the embodiment (in the sense of being located in space and time, having mass, and so on) of a computer enables it to achieve, for example, various forms of self-reference and even abstract mathematical reference. For example, it is the physically embodied

'two-ness' of a list `L=(a, b)` stored in computer memory that makes it possible for the computer to evaluate expressions such as `length(L)`.

Concepts of embodiment may also be necessary for us to theorize about the representational states of cognitive systems. The traditional means of specifying the content of, say, a belief state, is to provide a 'that' clause: a natural-language sentence that carries the same content as the belief that is being specified, for example, 'the child believes that the toy is within reach'. There are good reasons to believe that many representational contents, such as those of animals, infants, and sub-personal states, cannot be expressed in the conceptual framework of public language. To specify such contents for the purpose of a cognitive-scientific explanation, then, one may have to make essential reference to the body, and in particular the sensorimotor capabilities, of the system being explained (Cussins, 1990; Chrisley, 1995).

References

Breazeal C and Scassellati B (2000) Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior* **8**(1): 49–74.

Brooks RA (1991) Intelligence without representation. *Artificial Intelligence* **47**: 139–159.

Brooks RA and Stein LA (1994) Building brains for bodies. *Autonomous Robots* **1**: 7–25.

Chrisley R (1995) Taking embodiment seriously: nonconceptual content and robotics. In: Ford K, Glymour C and Hayes P (eds) *Android Epistemology*, pp. 141–166. Cambridge, MA: AAAI/MIT Press.

Churchland PS (1986) *Neurophilosophy: Toward a Unified Science of the Mind–Brain*. Cambridge, MA: MIT Press.

Clark A (1997) *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Clark A (1999) An embodied cognitive science? *Trends in Cognitive Science* **3**(9): 345–351.

Clark A and Chalmers D (1998) The extended mind. *Analysis* **58**: 10–23.

Cussins A (1990) The conectionist construction of concepts. In: Boden M (ed.) *The Philosophy of Artificial Intelligence*, pp. 368–440. Oxford, UK: Oxford University Press.

Dreyfus H (1992) *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Heidegger M (1962) *Being and Time*. New York, NY: Harper and Row. [First published in 1927 as *Sein und Zeit*. Tübingen, Germany.]

Kirsh D (1991) Today the earwig, tomorrow man? *Artificial Intelligence* **47**: 161–184.

Lakoff G and Johnson M (1999) *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York, NY: Basic Books.

Maturana HR and Varela FJ (1980) *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Netherlands: Reidel.

Merleau-Ponty M (1962) *Phenomenology of Perception*. London, UK: Routledge and Kegan Paul. [First published in 1945 as *Phénoménologie de la Perception*. Paris, France: Gallimard.]

Millikan R (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Newell A and Simon H (1976) Computer science as empirical enquiry: symbols and search. *Communications of the Association for Computing Machinery* **19**: 105–132.

Nilsson N (1984) *Shakey the Robot*. SRI Technical Note 323. Menlo Park, CA: SRI International.

Nuñez R (1999) Could the future taste purple? *Journal of Consciousness Studies* **6**(11–12): 41–60.

Piaget J (1954) *The Construction of Reality in the Child*. New York, NY: Basic Books. [First published in 1937 as *La Construction du Réel Chez l'Enfant*. Neuchâtel, Switzerland: Delachaux et Niestlé.]

Port R and van Gelder T (eds) (1995) *Mind As Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Sharkey NE and Ziemke T (1998) A consideration of the biological and psychological foundations of autonomous robotics. *Connection Science* **10**(3–4): 361–391.

Sloman A (2001) Evolvable biologically plausible visual architectures. In: Cootes T and Taylor C (eds) *The Proceedings of the British Machine Vision Conference, Manchester, September 2001*, pp. 313–322. Manchester, UK: BMVC Press.

Smith B (1996) *On the Origin of Objects*. Cambridge, MA: MIT Press.

Thelen E and Smith L (1994) A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press.

von Uexküll J (1928) *Theoretische Biologie*. Berlin: Springer.

Varela FJ, Thompson E and Rosch E (1991) *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Vygotsky LS (1978) *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press. [First published 1934 in Russian.]

Ziemke T (2001) Are robots embodied? In: Balkenius C, Zlatev J, Kozima H, Dautenhahn K and Breazeal C (eds) *Proceedings of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pp. 75–83. Lund, Sweden: Lund University Cognitive Studies.

Further Reading

Brooks RA (1990) Elephants don't play chess. *Robotics and Autonomous Systems* **6**(1–2): 1–16.

Chrisley R (2000) *Artificial Intelligence: Critical Concepts*. London, UK: Routledge.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Pfeifer R and Scheier C (1999) *Understanding Intelligence*. Cambridge, MA: MIT Press.

Sheets-Johnstone M (1999) *The Primacy of Movement*. Amsterdam: John Benjamins.

von Uexküll J (1982) The theory of meaning. *Semiotica* **42**(1): 25–82.

Ziemke T (2001) The construction of 'reality' in the robot: constructivist perspectives on situated artificial intelligence and adaptive robotics. *Foundations of Science* **6**(1): 163–233.

Emergence

Intermediate article

Achim Stephan, University of Osnabrück, Osnabrück, Germany

CONTENTS

In ordinary language, to 'emerge' means to 'appear' or 'come into view'; but the technical use of the term is associated with features such as novelty, irreducibility, and unpredictability. The basic idea of emergence is that as systems become increasingly complex during evolution, some of them may exhibit novel properties that are neither predictable nor explainable on the basis of the laws governing the behavior of the systems' parts. Thus, complex wholes can come to have properties that are not reducible to the properties and relations of their constituents.

INTRODUCTION

During the 1990s, the term 'emergence' became widely used in such different fields as the philosophy of mind, self-organization, creativity, artificial life, dynamical systems, and connectionism. The term, however, is not used in a uniform way. It can imply novelty, unpredictability, irreducibility, and the unintended arising of systemic properties, particularly in artificial systems. Thus, it is rather controversial what the criteria are by which 'genuine' emergent phenomena should be distinguished from non-emergent phenomena. Some of the suggested criteria are very demanding, so that few, if any, properties would count as emergent. Others are inflationary, in that they count many, if not all, system properties as emergent. First of all, therefore, one should be clear about the various types of emergence. (*See* **Philosophy of Mind**; **Self-organizing Systems**; **Creativity**; **Artificial Life**; **Dynamical Systems, Philosophical Issues about**; **Connectionism**)

VARIETIES OF EMERGENTISM

Three theories among the different varieties of emergentism deserve particular interest: synchronic emergentism, diachronic emergentism, and a form of weak emergentism. In synchronic emergentism, the relationship between a system's properties and its microstructure (i.e. the arrangement and properties of the system's parts) is at the center of interest. A property of a system is taken to be emergent if it is irreducible, i.e. if it is not reducible to the arrangement and properties of the system's parts. In contrast, diachronic emergentism is mainly interested in predictability of novel properties. Those properties are taken to be emergent that could not have been predicted, in principle, before their first instantiation. Both of these stronger versions of emergentism are based on a common weak theory from which they can be developed by adding further theses.

Weak Emergentism

Physical monism

The first thesis of current theories of emergence – the thesis of physical monism – concerns the nature of systems that have emergent properties or structures. It says that the bearers of emergent features consist of physical entities only. Thus, all substance-dualistic positions are rejected; for they base properties such as being alive or having cognitive states on supernatural bearers, such as an entelechy or a *res cogitans* respectively. (*See* **Dualism**; **Descartes, René**)

Physical monism is the thesis that entities existing or coming into being in the universe consist solely of physical components. Likewise, properties, dispositions, behaviors, or structures classified as emergent are instantiated by systems consisting exclusively of physical entities.

Systemic properties

While the first thesis places emergent properties and structures within the framework of a physicalistic naturalism, the second thesis – the thesis of systemic properties – delimits the types of properties that are possible candidates for emergents. It

is based on the idea that the general properties of a complex system fall into two classes: those that some of the system's parts also have, and those that none of the system's parts has. These latter properties are called systemic or collective properties.

The second thesis is that emergent properties are systemic properties. A property of a system is systemic if and only if the system possesses it but no proper part of the system possesses it.

Both artificial and natural systems with systemic properties exist. Those who would deny their existence would have to claim that all of a system's properties are instantiated already by some of the system's parts. Countless examples refute such a claim, e.g. it is among the properties of a leopard to run, but no part of it (head, heart, nor any cell assembly) can run; and it is among the properties of a connectionist network to recognize patterns, but no single part of it (unit, etc.) has this property.

Synchronic determination

The third thesis specifies the type of relationship that holds between a system's microstructure and its emergent properties. Namely, a system's properties and dispositions to behave depend nomologically on its microstructure, that is to say, on the properties and arrangement of its parts. There can be no difference in a system's systemic properties without some difference in the properties or arrangement of its parts.

Anyone who denies the thesis of synchronic determination has either to admit properties of a system that are not bound to the properties and arrangement of its parts, or to suppose that some other factors, in this case non-natural factors, are responsible for the different dispositions of systems that are identical in their microstructure. One may have to admit, for example, that there may exist objects that have the same parts in the same arrangement as diamonds, but that lack the diamond's hardness. This seems implausible. Equally implausible is the idea that there may exist two physically identical organisms, one viable and the other not. In the case of mental phenomena, opinions may be more divided; but one thing seems to be clear: anyone who believes, for example, that two physically identical creatures could be such that one is colorblind while the other is not, is not a physicalist.

Weak emergentism as sketched so far specifies the minimal criteria for emergent properties. It is the common base for all stronger theories of emergence. Moreover – and this is a reason for distinguishing it as a theory in its own right – it is held not only by some philosophers (e.g. Bunge, 1977), but also by some cognitive scientists (e.g. Varela *et al*., 1991; Rumelhart and McClelland, 1986) in exactly its weak form. Weak emergentism is compatible with current reductionist approaches; and some champions of weak emergentism cite this compatibility as one of its merits compared with stronger versions of emergentism.

Synchronic Emergentism

The essential theses of the two more ambitious theories of emergence are the theses of irreducibility (synchronic emergentism) and of unpredictability (diachronic emergentism). These are closely connected: irreducible systemic properties are *eo ipso* unpredictable before their first appearance. Hence, synchronically emergent properties are also diachronically emergent, but not conversely.

A systemic property is irreducible if it cannot be explained reductively. For a reductive explanation to be successful several conditions must be met: the property to be reduced must be functionally construable or reconstruable; it must be shown that the specified functional role is filled by the system's parts and their mutual interactions; and the behavior of the system's parts must follow from the behavior they show in isolation or in simpler systems than the system in question. (It is an open question whether or not properties exist that demand a construction or reconstruction other than being functional.) If all these conditions are met, the behavior of the system's parts in other contexts reveals what systemic properties the actual system has. (*See* **Reduction**; **Functionalism**)

Since these three conditions are independent of each other, there are three different ways in which systemic properties may be irreducible. Namely, a systemic property is irreducible if: it is not functionally construable (or reconstruable); if it cannot be shown that the interactions between the system's parts fill the systemic property's specified functional role; or if the specific behavior of the system's components, over which the systemic property supervenes, does not follow from the component's behavior in isolation or in simpler configurations. (*See* **Supervenience**)

Thus, we have to distinguish three different types of irreducibility of systemic properties. Their consequences are also different. If a property is irreducible due to the irreducibility of its bearer's parts' behavior we seem to have an instance of 'downward causation'. For, if the parts' behavior is not reducible to their arrangement and the behavior they show in simpler systems, then there

seems to exist some 'downward' causal influence, from the system itself or from its structure, on the behavior of the system's parts.

Such 'downward causation' would not violate the principle of the causal closure of the physical domain. We would just have to accept additional types of causal influences within the physical domain besides the known types of mutual interactions.

Likewise, if it cannot be shown that the interactions between the system's parts fill the specified functional role, it seems that the systemic property has causal powers that the microstructure does not have; hence in this case too there would be some downward causal influence.

In contrast, the occurrence of properties that are not functionally construable does not imply any kind of downward causation. Systems with properties that admit of no functional analysis need not be constituted in such a way that their components' behavior is irreducible. Nor is it implied that the system's structure has a downward causal influence on the system's parts. Thus there is no reason to assume that properties that cannot be analyzed themselves exert a causal influence on the system's parts. Rather, the question is how properties that cannot be functionally analyzed might have any causal role to play at all. And if one cannot see how they might play a causal role, then, it seems, such properties must be epiphenomena. (*See* **Epiphenomenalism**; **Mental Causation**)

Diachronic Emergentism

All diachronic theories of emergence are based on a thesis about the occurrence of genuine novelties in evolution. This thesis excludes all preformationist positions. According to this thesis, in the course of evolution exemplifications of genuine novelties occur again and again. Existing building blocks develop new configurations; new structures are formed that constitute new entities with new properties and behaviors.

However, the thesis of novelty does not by itself turn a weak theory of emergence into a strong one, since reductive physicalism remains compatible with such a variant of emergentism. Only the addition of the thesis of unpredictability, in principle, will lead to stronger forms of diachronic emergentism.

The first occurrence of a systemic property can be unpredictable for at least two different reasons. Firstly, it is unpredictable, in principle, if it is irreducible. This does not mean, however, that further occurrences of the property might not be predicted adequately. Secondly, it can be unpredictable because the microstructure of the system that exemplifies the property for the first time in evolution is unpredictable. Since in the first case the criteria for being unpredictable are identical with those for being irreducible, this notion of unpredictability will offer no theoretical gains beyond those afforded by the notion of irreducibility. Let us focus, therefore, on the second case: unpredictability of structure.

The structure of a newly formed system can itself be unpredictable for two reasons. If the universe is indeterministic, then its novel structures will be unpredictable. However, from an emergentist perspective, it is of no interest if a new structure's appearance is unpredictable only as a result of its indeterminacy – most emergentists claim that the development of new structures is governed by deterministic laws.

Nevertheless, deterministic formings of new structures can be unpredictable in principle if they are governed by laws of deterministic chaos. Against that claim one might argue that a Laplacean calculator could predict even chaotic processes correctly. Whether or not this 'actually' could be the case is not yet settled. It depends mainly on the question of what kind of information we allow such a creature to have. At least we can be sure that creatures of our mental capacities do not have these forecasting abilities, and thus, we can legitimately suppose that where chaos exists, structures exist that are unpredictable in principle.

The thesis of structure unpredictability is that the rise of a novel structure is unpredictable in principle if its formation is governed by laws of deterministic chaos. Likewise, any property that is instantiated by the novel structure is unpredictable in principle.

Although diachronic emergentism with the thesis of structure unpredictability implies the unpredictability of all properties instantiated by systems that emerge from chaotic processes, it does not thereby imply their irreducibility. The unpredictability, in principle, of systemic properties is entirely compatible with their being reducible to the microstructure of the system that instantiates them.

Synopsis

Figure 1 shows the logical relationships that hold between the different versions of emergentism.

Weak diachronic emergentism results from weak emergentism by adding a temporal dimension in the form of the thesis of novelty. Both versions are

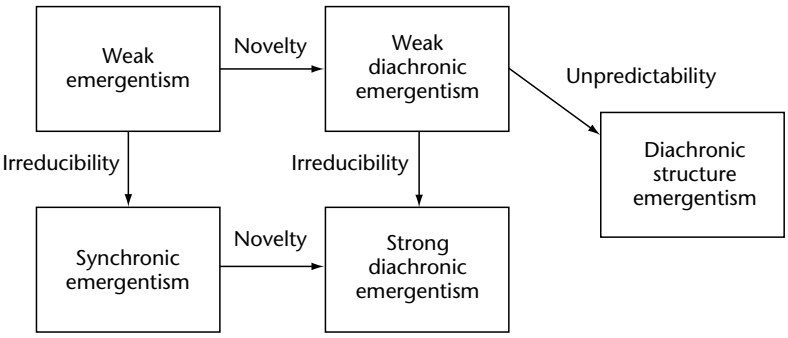


Figure 1. The logical relationships between the varieties of emergentism. Each arrow represents the addition of a thesis to a weaker theory.

compatible with reductive physicalism. Weak theories of emergence are used today mainly in cognitive science, particularly for characterization of systemic properties of connectionist networks, and in theories of self-organization. Synchronic emergentism results from weak emergentism by adding the thesis of irreducibility. This version of emergentism is important for the philosophy of mind, particularly for debating nonreductive physicalism and qualia. It is not compatible with reductive physicalism. Strong diachronic emergentism differs from synchronic emergentism because of the temporal dimension in the thesis of novelty. Structure emergentism is entirely independent of synchronic emergentism. It results from weak diachronic emergentism by adding the thesis of structure unpredictability. Although structure emergentism emphasizes the boundaries of prediction within physicalistic approaches, it is compatible with reductive physicalism, and so it is weaker than synchronic emergentism. Theories of deterministic chaos can be considered as a type of structure emergentism. This perspective is important for evolutionary research. (*See* **Qualia**)

HISTORY

Although some hints of emergentist thinking can be found in the works of Empedocles, Epicurus, and Galen, the proper development of emergentism began in the mid nineteenth century in Britain. George Henry Lewes (1875) introduced the term 'emergent' into philosophy, to distinguish 'emergent' from 'resultant' effects. Here Lewes picked up on John Stuart Mill's distinction between 'homogeneous' and 'heterogeneous' effects: joint effects of causes are called heterogeneous (or emergent) if they are not the 'sum' of their separate effects; otherwise they are called homogeneous (Mill, 1974). Mill's distinction between 'ultimate' and 'derivative' laws was also of great importance for the development of emergentist ideas. Some decades later, C. D. Broad oriented himself by Mill's distinctions and his subsequent considerations about the limits to explanation of psychophysical laws.

In the 1920s, theories of emergence began to attract greater philosophical and scientific interest. In rapid sequence the major works of British and American emergentism appeared: in 1920 Samuel Alexander's *Space, Time, and Deity*, in 1922 Roy Wood Sellars's *Evolutionary Naturalism*, in 1923 Conwy Lloyd Morgan's *Emergent Evolution*, and in 1925 Charles Dunbar Broad's *The Mind and its Place in Nature*.

Most of these philosophers' theories of emergence are reactions to the debate on the nature of life. While vitalists like Hans Driesch and Henri Bergson claimed, for the explanation of vital processes, the existence of supernatural entities such as an 'entelechy' or an *élan vital*, biological mechanists were trying to reduce all phenomena of life to physical and chemical processes without residue. Both positions seem to have implausible consequences: substance-dualistic approaches violate the principle of the causal closure of the physical domain, and it is hard to square them with evolutionary cosmologies; while mechanism does not seem to capture genuine organic and mental processes adequately. The emergentists steered a middle course. They denied both substance-dualistic and reductionist theories: they were nonreductive naturalists.

In the following decades, theories of emergence were much discussed. However, the criticism by Hempel, Oppenheim, and Nagel seemed to put an early end to emergentism, for their analysis led to an uninteresting concept of emergence as meaning nothing but: 'considering all theories we know of, we cannot explain why system *S* has property *P*' (Hempel and Oppenheim, 1948; Nagel, 1961).

With the decline of positivism, interest in metaphysical questions returned. It is the unsettled question about the nature of mental states that has helped emergentism to return to the philosophy of mind. The concept of emergence has also gained ground in such fields as self-organization, artificial life, the philosophy of science, and cognitive science.

EMERGENCE IN THE PHILOSOPHY OF MIND AND COGNITIVE SCIENCE

In different fields of philosophy and cognitive science the idea of 'emergence' has different roles. Thus, within the philosophy of mind, and particularly within the debate about qualia, there is a need for a strong notion of emergence; while within the fields of connectionism and artificial life, weaker notions of emergence suffice.

Emergentism as Nonreductive Physicalism

Within the philosophy of mind, emergentism is the most recent form of what has been called 'nonreductive physicalism' since the 1970s: a doctrine that in one way or other has tried to establish a compromise between physicalist reductionism and various sorts of dualism. First, physicalistic functionalism was seen as a species of nonreductivism because of its violation of biconditional bridge laws of the Nagelian type by its acceptance of multiply realizable mental properties. Subsequently, psychophysical supervenience was thought to be a theory of mind that is essentially both physicalistic and nonreductive. (*See* **Materialism**; **Dualism**; **Functionalism**; **Reduction**; **Supervenience**; **Multiple Realizability**)

Careful analyses, however, particularly by Jaegwon Kim (1993, 1998), revealed that both positions fall short of being what they are supposed to be. Physicalistic functionalism turned out to be reductionistic (it guarantees reductive explanations); and psychophysical supervenience, even in its strong form, turned out to be too weak to establish any specific theory of mind at all. Since even such diverse positions on the mind–body problem as reductive type physicalism and epiphenomenalism entail psychophysical supervenience, theories of supervenience fail to guarantee nonreductivism. In fact, it is synchronic emergentism that comprises the tenets originally associated with supervenience: property covariation and the dependence of supervenient properties on their subvenient bases are captured by the third thesis (synchronic determination) of weak emergentism; irreducibility, of course, is captured by the fourth tenet which is specific to synchronic emergentism. (*See* **Mind–Body Problem**; **Epiphenomenalism**)

Since weak emergentism (like mind–body supervenience) is compatible with both reductionism and nonreductionism, strong emergentism seems to be the only adequate representative of nonreductivism in recent philosophy of mind. An interesting question, however, is whether or not synchronic emergentism really is physicalism. Some philosophers maintain that such a position should be characterized as a kind of dualism, namely property dualism. However, insofar as psychophysical supervenience is regarded as defining minimal physicalism (Kim, 1998), synchronic emergentism can be seen as physicalism, too, and thus be treated as a genuine instance of nonreductive physicalism.

Qualia Emergentism

A case in point for the idea that nonreductive physicalism might be an adequate answer is the problem of phenomenal consciousness. Chalmers, Jackson, and Levine, among others, have argued in various ways that qualitative mental phenomena are not reducible to physical or functional states. If their arguments are sound, they imply strong emergentist positions. Most interesting and powerful seem to be Chalmers's argument for the 'hard problem' of consciousness and Levine's 'explanatory gap' argument. (*See* **Consciousness, Philosophical Issues about**; **Knowledge Argument, The**; **Explanatory Gap**)

According to Levine (1993) and Chalmers (1996), reductive explanations require two stages. The first stage involves the *a priori* process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal or functional role for which we are seeking the underlying mechanisms. The second stage involves the empirical work of discovering just what those underlying mechanisms are.

Since our concepts of phenomenal qualities do not seem to represent – at least in terms of their psychological contents – causal roles, a failure, in principle, of the first task seems to be unavoidable. Thus, to the extent that there is an element in our concept of qualitative character that is not captured by features of its causal role, qualia are irreducible emergent properties.

Emergence in Connectionism

Connectionism gives rise to emergentist considerations in several ways: trained networks show

cognitive features such as 'rule following', 'schema formation', or 'pattern recognition', that their parts do not have. Thus, the systemic properties that a network acquires are weakly emergent. However, they are not irreducible: they are fully deducible from the network's structure, the properties of its units (their activation formulae), and the properties of their links (the distribution of weights, and the formulae for changing the weights). Thus, systemic properties of connectionist networks are not synchronically emergent.

Connectionists often make use of the word 'emergent' in its ordinary sense, sometimes intermingled with a more technical usage. For example, Rumelhart and McClelland (1986) say that a network's high-level properties 'emerge' from low-level interactions: rules and schemata come into being by themselves without being explicitly fed into the system. However, Rumelhart and McClelland do not thereby subscribe to emergentism. They mainly try to differentiate their position from traditional representationalism, accordingly to which all rules and schemata have to be fed in explicitly. (*See* **Representation, Philosophical Issues about**; **Implicit and Explicit Representation**)

Connectionist networks develop their specific distribution of link weights (their soft structures) in a somewhat evolution-like process. Again, this is not a case of genuine emergence. Since the changes of weights are calculable exactly if we know the initial magnitudes, we should not speak of structure emergence in connectionism. On the other hand, regarding their soft structures, networks show great plasticity, compared with other objects. Chemical compounds, for example, have no freedom to change their internal structure: the diamond's property of being hard is always manifest; it does not emerge.

Emergence of Creativity

The notion of emergence is also of interest in the field of creative cognition. There we seem to face a paradox: how is it ever possible to form a truly creative idea? If we could predict it, it would be determined and not creative. If we could not in principle anticipate it, how could we produce such an idea at all? Some psychologists assume that the cognitive structures involved in creative thinking have emergent properties that could be discovered when those structures are explored, at least some of which could not have been anticipated in advance. This seems like a postulation of unpredictable cognitive features as a result of structure emergence. (*See* **Creativity**)

Emergentism and Artificial Life

Within the field of artificial life, the notion of emergence is central. It refers to adaptive features of artificial systems that result both from 'clever' interactions of many simple components and from couplings between agents and their environments (including other agents). However, the term 'emergence' is not used in its strong sense here, since all phenomena studied in artificial life are reductively explainable, at least in principle. Rather, emergence is associated with behavior that is not centrally controlled and that cannot be reduced to the behavior of single components within hierarchical systems, but is seen as the result of the interactions of multiple simple components or as the outcome of the overall dynamics of the agent and its environment. Thus, the notion of emergence used in artificial life is close to that used in connectionism. (*See* **Artificial Life**)

ARGUMENTS FOR AND AGAINST EMERGENCE

Clearly weak emergent properties exist: indeed, one might ask why such properties should be called 'emergent' at all, and not just 'systemic'. Furthermore, since there are chaotic processes of structure formation, structure emergence exists too. Thus, what is really in question is synchronic emergence. What we need is an argument for the existence of properties that are not and will not be reductively explainable. Many natural scientists deny the existence of such properties, since they do not know of any properties that could not be reductively explained, at least in principle. Without exception, all systemic properties studied in the natural sciences are functionally construable, their functional roles are always filled by the interactions of their systems' parts, and the behavior of the parts of any system always seems to follow from their behavior in simpler systems. Therefore, some critics question whether it is useful to develop the notion of synchronic emergence at all. But, even if it should turn out that all systemic properties studied in the natural sciences are reductively explainable, it is useful to have the strong notion of synchronic emergence. More than any other notion, it can be used to clearly formulate nonreductive positions concerning the mind–body problem.

Whether or not synchronically emergent properties actually exist does not seem to depend on empirical, but rather on conceptual grounds. Among others, Broad, Levine, and Chalmers have argued forcefully that properties such as qualia are not

functionally analyzable. If they are right, then phenomenal qualities may be emergent properties in the strong sense.

If mental properties such as qualia are emergent in the strong sense, then new problems arise. Some philosophers have claimed that irreducible properties necessarily exert downward causation. In the case of mental phenomena, however, this would conflict with the principle of the causal closure of the physical domain.

However, as we have seen above, properties that are irreducible for conceptual reasons do not imply downward causation. Rather, they give rise to another objection: how can properties that escape reconstruction via their causal role play any causal role at all? The reply to this objection mainly depends on our concept of causation. If we think that supervenient causation suffices for causation, then irreducible emergent properties can be causally efficacious. If we think that supervenient causation does not suffice, then irreducible emergent properties do not seem able to play any causal role. But these are still open questions (Kim, 1998; Stephan, 1997). (*See* **Mental Causation**)

References

Bunge M (1977) Emergence and the mind. *Neuroscience* **2**: 501–509.

Chalmers DJ (1996) *The Conscious Mind*. Oxford, UK: Oxford University Press.

Hempel CG and Oppenheim P (1948) Studies in the logic of explanation. *Philosophy of Science* **15**: 135–175.

Kim J (1993) *Supervenience and Mind*. Cambridge, UK: Cambridge University Press.

Kim J (1998) *Mind in a Physical World: An Essay on the Mind–Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Levine J (1993) On leaving out what it's like. In: Davies M and Humphreys GW (eds) *Consciousness*, pp. 121–136. Oxford, UK: Blackwell.

Lewes GH (1875) *Problems of Life and Mind*, vol. II. London, UK: Kegan Paul.

Mill JS (1974) *A System of Logic: Ratiocinative and Inductive*. Toronto, Canada: University of Toronto Press [First published 1843.]

Nagel E (1961) *The Structure of Science*. New York, NY: Routledge and Kegan Paul.

Rumelhart DE and McClelland JL (1986) PDP models and general issues in cognitive science. In: Rumelhart DE, McClelland JL and the PDP Research Group (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I, pp. 110–146. Cambridge, MA: MIT Press.

Stephan A (1997) Armchair arguments against emergentism. *Erkenntnis* **46**: 305–314.

Varela FJ, Thompson E and Rosch E (1991) *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Further Reading

Beckermann A, Flohr H and Kim J (eds) (1992) *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin and New York, NY: de Gruyter.

Bedau MA (1997) Weak emergence. *Philosophical Perspectives: Mind, Causation, and World* **11**: 375–399.

Clark A (1996) Happy couplings: emergence and explanatory interlock. In: Boden M (ed.) *The Philosophy of Artificial Life*, pp. 262–281. Oxford, UK: Oxford University Press.

Finke RA, Ward TB and Smith SM (1992) *Creative Cognition*. Cambridge, MA: MIT Press.

Holland JH (2000) *Emergence: From Chaos to Order*. Oxford, UK: Oxford University Press.

Humphreys P (1997) How properties emerge. *Philosophy of Science* **64**: 1–17.

Kim J (1999) Making sense of emergence. *Philosophical Studies* **95**: 3–36.

O'Connor T (1994) Emergent properties. *American Philosophical Quarterly* **31**: 92–104.

Stephan A (1999) *Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation*. Dresden and Munich, Germany: Dresden University Press. [In German.]

Wimsatt W (1997) Aggregativity: reductive heuristics for finding emergence. *Philosophy of Science* **64**: S372–S384.

Emotion, Philosophical Issues about

Intermediate article

Paul E Griffiths, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

CONTENTS

Philosophers have discussed the nature of emotions with particular reference to whether emotions are feelings, whether they are, or involve, cognitive states such as beliefs, whether they are human universals, whether the category of emotion is a unitary one, and the relationship between emotion and rationality.

CENTRAL PHILOSOPHICAL ISSUES ABOUT EMOTION

Feeling and Cognition

Until the twentieth century it was generally believed that emotions are feelings: subjective states of experience. Darwin carried out extensive empirical investigations of the physiological and behavioral components of emotion but interpreted these as merely external manifestations of the emotions themselves. He regarded emotions as closely akin to bodily sensations such as hunger or pain. Feelings also played a central role in the other important theory of the closing years of the nineteenth century, the James/Lange theory of emotion causation. William James provoked a long and productive tradition of research with his suggestion that emotion feelings are caused by the bodily changes associated with emotion, rather than the reverse. Nevertheless, he took the final stage in this causal sequence, the occurrence of emotion feelings, to be the central feature of an emotion considered as a *psychological* phenomenon.

Predictably, the rise of behaviorism led to the decline of the feeling theory of emotion. Behaviorists in psychology had no difficulty in fitting emotional reactions into their theoretical framework. John B. Watson suggested that all adult emotional reactions were conditioned responses based on three unconditioned reactions present in infants that he termed fear, rage, and pleasure. Behaviorist philosophers such as Gilbert Ryle attempted to analyze the meanings of sentences about emotion so as to eliminate any essential reference to subjective states of experience. The 'cognitive revolution' of the 1960s did not rehabilitate the feeling theory. Philosophers assimilated the new turn in the sciences of the mind by taking existing behavioral definitions of mental states and treating them as implicit definitions of underlying mental states which cause behavior. An emotion is an internal state that mediates causally between sensory inputs and behavioral outputs in a characteristic way. Crudely, anger is an internal state that takes slights as inputs and yields aggression as output. No reference to the quality of feeling associated with anger is required.

The consensus that emerged in the philosophy of emotion in the early 1960s and persists to the present day is that emotions are defined by the cognitions they involve. Later theorists have increasingly allowed feelings a role in emotion, but never one that determines the identity of the emotion. Instead, the role of emotion feelings is to add the 'heat' to 'hot cognition'. A leading contemporary philosopher of emotion, Patricia Greenspan, concludes that emotions are feelings of comfort or discomfort directed towards an evaluative thought about an external (or imaginary) stimulus (Greenspan, 1988). It is the evaluative thought that defines the emotion. Different negative emotions, such as anger and fear, are differentiated only by the evaluative thoughts they involve. Philosophers have generally held it to be a conceptual truth that emotions derive their identities from the thoughts associated with them, and so have not seen empirical results showing the differentiation of states of bodily arousal in different emotions as relevant to their research. Philosophers have, however, cited work on the 'cognitive labeling' of states of arousal as evidence that empirical findings converge on

the same conclusion as their conceptual analyses. Many cite a famous 1962 study in which subjects were induced to describe the effects of identical adrenaline injections as either euphoria or anger (Schachter and Singer, 1962). The experimental design is quite complex, but in essence, some subjects were induced to discount the injection as a cause of their aroused state and these subjects described their state of arousal as an emotion appropriate to whichever cues the experimenters provided them with. (*See* **Cognitive Science: Philosophical Issues**)

Universality of Emotion

Emotions are widely believed to be a critical feature of moral agency, and are even more widely believed to be a critical part of aesthetic response. The claim that all healthy people display, recognize, and respond to the same emotions has been used to support the view that moral and aesthetic judgments can have universal validity. Conversely, if human emotions are as diverse as the concepts embodied in different languages, and if humans can understand the expressive repertoire only of their own cultural group, this would seem to support cultural relativism about ethics and aesthetics.

Until the 1970s there was a fairly solid consensus, based on anthropological fieldwork, that the emotions vary widely across cultures. The work of Paul Ekman and his collaborators has produced an equally widespread consensus that certain 'basic emotions' are found in all human cultures (Ekman, 1972). Ekman revived Darwin's experimental work on human facial expressions of emotion and demonstrated that a range of Western facial expressions of emotion could be reliably classified by members of a visually isolated, non-Western culture, and vice versa. He also reconfirmed many of Darwin's claims about the specific muscles used in these pancultural expressions. Other investigators demonstrated the early emergence of these expressions in human infants and established homologies with facial expressions in non-human primates. The widely accepted 'basic emotions' are fear, anger, surprise, sadness, joy, disgust, and perhaps contempt. (Each emotion term in this list refers to an operationally characterized, brief, involuntary response rather than to the full range of cases commonly referred to by the term.)

Cultural relativism about emotions was revived in the 1980s as part of a broader interest in the social construction of mental phenomena. This led to the first real involvement by analytic philosophers in the debate over universality, since the new arguments for social constructionism were as much conceptual as empirical. One influential argument starts from the widely accepted idea that an emotion involves a cognitive evaluation of the stimulus. In that case, it is argued, cultural differences in how stimuli are represented will lead to cultural differences in emotion. If two cultures think differently about danger, then, since fear involves an evaluation of a stimulus as dangerous, fear in these two cultures will be a different emotion. Adherents of the basic emotions view are unimpressed by this argument since they define emotions by their behavioral and physiological characteristics and allow that there is a great deal of variation in what triggers the same emotion in different cultures.

Social constructionists also define the domain of emotion in a way that makes basic emotions research less relevant. The six or seven basic emotions seem to require minimal cognitive evaluation of the stimulus. Social constructionists often refuse to regard these physiological responses as emotions in themselves, reserving that term for the broader cognitive state of a person involved in a social situation in which he or she might be described as, for example, angry or jealous.

It is thus unclear whether the debate between the constructionists and their universalist opponents is more than merely semantic. One side has a preference for tractable, reductive explanations, even if these are of limited scope, and the other is concerned that science may neglect the social and cultural aspects of human emotion. (*See* **Social Cognition**)

Is Emotion a Natural Kind?

The neuroscientist Antonio Damasio recently defined an emotion as 'a specifically caused transition of the organism state' (Damasio, 1999, p. 282). Confronted by similar definitions, Alan Fridlund has remarked: 'Here, the logical question is what *isn't* emotion. Emotion has, in fact, replaced Bergson's *elan vital* and Freud's *libido* as the energetic basis of all human life' (Fridlund, 1994, p. 185). For many theorists emotion has indeed become synonymous with the whole affective life of the organism and perhaps with motivation itself. Damasio is well aware of this situation and is self-consciously using a familiar term for his own purposes in order to facilitate communication in what he sees as a period of conceptual upheaval (Damasio, 1999, p. 341). He does indeed have a very broad conception of emotion, to the extent that he takes it as

axiomatic that a person is always in some emotional state or other.

Paul Griffiths has argued that the scientific investigation of the domain of affective phenomena has been hindered by a continued belief that 'the emotions' are a unitary kind of psychological state (Griffiths, 1997). Science aims to group phenomena into 'natural kinds': categories about which there are many, reliable generalizations to be discovered. The domain of emotion is so large that it is unlikely that all the psychological states in that domain form a natural kind. Hence there will be few if any reliable generalizations about emotion or, in other words, no theory of emotion in general. Scientific progress would be served by dividing up the domain and investigating groups of phenomena that are likely to form natural kinds, as has occurred in research into memory. Basic emotions theorists may well be investigating one such natural kind and social constructionists about emotion may be examining another – perhaps a certain kind of transient social role. New, more specific concepts will be required to replace the emotion concept, and a central role for philosophers of emotion is to facilitate this kind of conceptual revision.

Most philosophers of emotion see no serious problem with the category of emotion, although they admit that it is vague and covers a diverse range of phenomena. Their concern is with the proper analysis of the concept associated with the word 'emotion' in everyday language. Their analyses of the emotion concept are in reasonable agreement with those produced by psychologists studying the use of the term 'emotion' in Western cultures. There are clear paradigms of emotion, such as love, happiness, anger, fear, and sadness, and most philosophers define emotion so as to include these. Their definitions disagree over the same cases that produce disagreement between subjects in empirical studies, cases such as pride, hope, lust, pain, and hunger. Philosophical definitions include features that psychologists have argued are part of the prototype of the emotion concept in Western culture. Emotions are directed onto external states of affairs, are relatively short-lived, and have an evaluative aspect to them, such that their objects are judged to be either attractive or aversive. Most definitions also provide a role for emotion feelings.

Hence philosophers, like ordinary speakers, can achieve a reasonable level of agreement about what counts as an emotion, as opposed to a mood, a desire, or an intention. Whether the psychological states grouped together in this way form a single, productive object of scientific investigation, and whether other cultures conceptualize emotion in the same way, remains to be seen.

PHILOSOPHICAL VIEWS AND THEORIES OF EMOTION

Propositional Attitude Theories of Emotion

Since the early 1960s the cognitive or propositional attitude school has dominated the philosophy of emotion (Deigh, 1994). The basic commitments of this school are twofold. First, emotions are differentiated from one another by the cognitive states that they involve, as discussed above. Second, the cognitive states involved in emotion can be understood in terms of a *propositional attitude* theory of mental content. Mental states are attitudes, such as belief, desire, hope, and intention, to propositions. The nature of propositions is the subject of complex debate, but for present purposes we can treat them as representations of states of affairs.

The simplest propositional attitude theory is the judgmentalist theory, which identifies emotions with evaluative judgments. A person is angry if he or she has the attitude of belief to the proposition 'I have been wronged'. Other prominent varieties of propositional attitude theory are belief/desire theories, hybrid feeling theories, and 'seeing as' theories.

Belief/desire theories analyse emotions as combinations of beliefs and desires. For example, hope is the belief that some state of affairs is possible and the desire that it be actual.

Hybrid feeling theories, like that of Greenspan discussed above, analyze emotions as combinations of propositional attitudes and feelings. The feeling component is used to differentiate cold cognition from hot (emotional) cognition and in some theories to distinguish positive from negative emotions. The specific identity of the emotion is given by the propositional attitude component.

Finally, 'seeing as' theories have become increasingly popular. These theories cope with various anecdotal objections to earlier propositional attitude theories by noting that a subject's beliefs and desires about an object are not sufficient to constitute an emotion unless the subject 'sees' the object in the right way. A typical anecdote involves a mountain climber who is said to retain the same beliefs and desires as she fluctuates between seeing a climb as terrifying and as exhilarating. Earlier versions of this approach were inclined to treat 'seeing as' as a primitive, following some aspects

of the later work of Wittgenstein. Contemporary versions analyze 'seeing as' in terms of attentional phenomena in cognition. Emotions are then biases in cognition that direct attention at some sources of information rather than others or lead to a higher weighting for one consideration than another, and thus lead to actions that would not have eventuated in the absence of the emotion.

The theories just outlined are primarily intended as analyses of the concept of emotion. They are assessed for their ability to correctly predict the author's intuitions about whether an emotion, or some specific emotion, occurs in an imaginary scenario. Some authors draw extensively on literature for these scenarios, others draw on more or less actual cases from the psychoanalytic literature. Some see their work as assisting the scientific investigation of emotion by more clearly defining its subject matter. Others see their work as complementary and parallel to scientific psychology. Philosophical psychology is often distressingly unclear about the identity of its subject population. It is conventional in the literature to refer to 'our emotions' and to 'commonsense' views about emotion, but it is unclear to what extent these locutions are to be read as limiting the claims made to the author's own community.

The Rationality of Emotions

The prime concern of the propositional attitude school in the philosophy of emotion has been with whether emotions are rational. The feeling theory of emotion is condemned for placing emotions outside the realm of rational evaluation. This is seen as part of a wider and invidious tendency to separate the realm of the moral from the realm of the rational. The attempt to bring these realms together is conceived by some authors as a proposal for the reform of moral discourse and by others as an attempt to do justice to one strain of 'our' everyday practice.

The simplest judgmentalist theory brings emotions back into the domain of reason by identifying them with beliefs. An emotion is rational if the beliefs composing it are justified by the evidence available to the subject. More complex propositional attitude theories give more complex accounts of the rationality of emotions. Belief/desire theories face the difficulty that formal accounts of rationality, such as decision theory, are confined to evaluating the suitability of means to ends and take the ends (desires) as given. So these theories must provide an account of what it is rational to desire. Hybrid feeling theories can evaluate the rationality of having one emotion rather than another, since the identity of an emotion is determined by its propositional attitude component. Whether the state is an emotion in the first place, however, relies on the feeling component, and so hybrid feeling theories must give some account of when it is rational to take one's cognition hot rather than cold. An extensive literature canvasses solutions to these and other difficulties with the project of rationally evaluating emotions. 'Seeing as' theories face their own difficulties, but also have new resources to bring to bear on the rationality question. The cognitive biases that constitute emotions can be evaluated for their heuristic value in generating true belief, successful action, and so forth, and judged rational if they are successful in these respects.

IMPACT OF COGNITIVE SCIENCE ON ISSUES ABOUT EMOTION

Evolutionary Psychology and the Universality of Emotion

Two contemporary schools of evolutionary psychology provide arguments for diametrically opposite views on the universality of emotion. John Tooby and Leda Cosmides urge the application of their well-known blueprint for the evolutionary study of the mind to the domain of emotion (Tooby and Cosmides, 1990). The mind is a collection of highly specialized, domain-specific cognitive devices, or modules, each adapted to a specific ecological problem in our evolutionary past. Existing work in basic emotions research is easily assimilated to this model, and these evolutionists see the six or seven pancultural responses confirmed to date as the first of a much larger number yet to be uncovered. A particularly confident prediction is that there will be a specific module for sexual jealousy. The psychology of emotion in modern subjects is conceived as the result of these many evolved, pancultural modules interacting with the environments found in different societies.

In stark contrast to these ideas, evolutionary arguments are used to support a tradition of emotion research that denies that emotions come in discrete types and emphasizes cultural variation in the psychology of emotion. *Transactional* theories of emotion see emotions as acts of social communication: 'nonverbal strategies of identity realignment and relationship reconfiguration which do not easily translate into the official idea of reasoned argument and information exchange' (Parkinson, 1995, p. 295). A central tenet of this approach is

that emotional behaviors do not express emotions. Rather than being expressed in social interactions, an emotion actually *is* a particular kind of social interaction and emotions thus defined do not stand in a one-to-one relation to underlying psychological processes. The empirical research supporting the transactional view is aimed at showing the effects of social context both on the production of emotional behavior given a particular underlying cognitive state and on whether a given behavior is regarded as expressing an emotion. Transactional approaches naturally tend to be associated with the view that there is extensive cultural variation in emotion, since the forms of social interaction and the functions of emotions within those interactions will differ from society to society. The fundamental mechanisms underlying this variety, however, may be pancultural and subject to evolutionary explanation.

According to Alan Fridlund, the transactional view of emotion is strongly supported by the theories of animal communication found in contemporary sociobiology and behavioral ecology (Fridlund, 1994). Like most authors who have discussed the evolution of emotional expression, Fridlund treats these expressions as conveying information about an animal's motivational state. Organisms who transparently express the internal states that guide their future actions would be unlikely to succeed in evolutionary competition. Instead, he argues, selection would bring any existing signs of emotion under increasing voluntary control, enabling organisms to control the flow of information to their own advantage. Fridlund suggests that an evolutionary psychology of emotion will naturally interpret emotional behaviors not as expressions but as signals produced to manipulate the behavior of others. If we can infer other people's emotions when they would prefer us not to, then this must be the outcome of an 'arms race' between organisms seeking to predict the behavior of others and organisms seeking to manipulate their expectations. Fridlund's argument is certainly in line with the fundamental orientation of the game-theoretic literature on animal communication. However, evolutionary theory is notorious for its inability to predict the course of evolutionary change and it would be a mistake to give this theoretical argument much weight in comparison to empirical studies of the reliability, or lack thereof, with which people recognize one another's emotions. (*See* **Game Theory**)

The obvious objection to a transactional theory of emotion is that it cannot explain asocial emotions, such as fear of an asocial stimulus in a solitary subject. Fridlund has labored to demonstrate experimentally that audience effects play a cognitive role even in asocial emotions: 'What would people think if they saw me?' More fundamentally, however, transactional theorists, like the social constructionists before them, are prepared to redefine the domain of emotion in order to capture what they take to be psychological phenomena of a single natural kind and to exclude phenomena that are of a different kind (Parkinson, 1995, p. 303).

The Frame Problem and the Resurgence of the Feeling Theory

Recent work in cognitive neuroscience has shed new light on the relationship between emotion and cognition and has led to a revival of the feeling theory of emotion. Antonio Damasio has argued that effective reasoning is dependent on the capacity to experience emotion. Patients with bilateral lesions to the prefrontal cortex show both reduced emotionality and a diminished ability to allocate cognitive resources in such a way as to solve real-world problems. They do not, however, have deficits in abstract reasoning ability. Damasio interprets these findings as showing that emotion plays an essential role in labeling both data and goals for their relevance to the task in hand (Damasio, 1994).

Damasio's proposal must be regarded as highly tentative, and faces a number of difficulties (Rolls, 1999); but it has aroused considerable interest amongst cognitive scientists who have seen in 'affective computing' a possible solution to the frame problem: the problem of choosing all and only the relevant data without assessing all the available data for possible relevance (Picard, 1997). Damasio's theory bears a resemblance to some of the philosophical 'seeing as' theories which identify emotions with heuristic biases in cognition. In contrast to those theories, however, Damasio sees emotions themselves as feelings. If emotions function cognitively, then his proposal would be that cognitive priorities are assigned by calculating what is most relevant and important. This would not be a solution to the frame problem, but an instance of that problem. Damasio avoids this trap by using emotion *feelings* to prioritize cognition. He describes a class of 'primary emotions' that bear a strong affinity to Ekman's basic emotions. Damasio envisages emotional development as a process in which the feelings associated with the basic emotions become attached to particular cognitive states, giving rise to cognition/feeling composites that he labels 'secondary emotions'.

Damasio has so far given only a suggestive outline of his theory and it remains to be seen whether

this sketch can be developed into a workable model of cognitive processes. Attempts to expand on Damasio's ideas to date resemble traditional behavior conditioning, with thoughts taking the place of behaviors and emotion feelings acting as reinforcers. The limitations of conditioning models as explanations of complex cognitive performances are well known.

Neurological Support for Twin Pathway Models of Emotion

Another important recent development is Joseph LeDoux's detailed mapping of the neural pathways involved in fear conditioning (LeDoux, 1996). Information about the stimulus activates many aspects of emotional response via a fast, 'low road' through subcortical structures, amongst which the amygdala is particularly important. A slower, 'high road' activates cortical structures and is essential for longer-term, planned, and often conscious responses to the same stimulus.

These findings are consistent with Ekman's proposal that an 'automatic appraisal mechanism' is associated with the basic emotions and operates independently of the formation of conscious or reportable judgments about the stimulus situation. LeDoux's findings also help to explain the experimental phenomenon of 'affective primacy', in which emotional associations with stimuli can be conditioned independently from paradigmatically cognitive responses to the same stimuli, such as recognition or recall.

Twin-pathway models suggest that at least for certain basic emotions the idea that an emotion involves a cognitive evaluation of the stimulus needs to be replaced with the idea that it involves two evaluations, which can conflict and which have complementary but independent cognitive functions. Twin-pathway models also provide some support for the many evolutionary accounts that see the basic emotions as 'quick and dirty' solutions to common survival problems.

CONCLUSION

Perhaps the most pressing philosophical question about emotion is the relationship between the philosophical psychology of emotion and the sciences of the mind. Even the most apparently purely philosophical issue, the rationality of emotion, is now the subject of simultaneous scientific study. A closer examination of the work of philosophers who claim to be concerned only with conceptual issues reveals that few of them fail to take some note of relevant empirical findings. Conversely, the exciting recent work of Antonio Damasio has led him to confront the traditional philosophical questions of the nature of mental representation and of conscious experience. These questions are arguably as much conceptual puzzles as empirical questions. The future of the field seems inevitably to be one of closer interdisciplinary cooperation.

References

Damasio AR (1994) *Descartes' Error: Emotion, Reason and the Human Brain*. New York, NY: Grosset/Putnam.

Damasio AR (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt Brace.

Deigh J (1994) Cognitivism in the theory of emotions. *Ethics* **104**: 824–854.

Ekman P (1972) *Emotions in the Human Face*. New York, NY: Pergamon Press.

Fridlund A (1994) *Human Facial Expression: An Evolutionary View*. San Diego, CA: Academic Press.

Greenspan P (1988) *Emotions and Reasons: An Inquiry into Emotional Justification*. New York, NY: Routledge.

Griffiths PE (1997) *What Emotions Really Are: The Problem of Psychological Categories*. Chicago, IL: University of Chicago Press.

LeDoux J (1996) *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon & Schuster.

Parkinson B (1995) *Ideas and Realities of Emotion*. London, UK: Routledge.

Picard R (1997) *Affective Computing*. Cambridge, MA: MIT Press.

Rolls ET (1999) *The Brain and Emotion*. Oxford, UK: Oxford University Press.

Schachter S and Singer JE (1962) Cognitive, social and physiological determinants of emotional state. *Psychological Review* **69**: 379–399.

Tooby J and Cosmides L (1990) The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology* **11**: 375–424.

Further Reading

Calhoun C and Solomon RC (eds) (1984) *What is an Emotion? Classic Readings in Philosophical Psychology*. New York, NY: Oxford University Press.

Darwin C (1872) *The Expression of the Emotions in Man and Animals*. New York, NY: Philosophical Library.

De Sousa R (1991) *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Ekman P (ed.) (1973) *Darwin and Facial Expression: A Century of Research in Review*. New York, NY: Academic Press.

Ekman P and Davidson RJ (eds) (1994) *The Nature of Emotion: Fundamental Questions*. New York, NY: Oxford University Press.

Harré R (ed.) (1986) *The Social Construction of the Emotions*. London, UK: Oxford University Press.

Panksepp J (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York, NY: Oxford University Press.

Epiphenomenalism

Intermediate article

William S Robinson, Iowa State University, Ames, Iowa, USA

CONTENTS

Epiphenomenalism is the view that mind does not affect behavior, or does not affect behavior in the ways it is commonly believed to affect it.

WHAT IS EPIPHENOMENALISM?

Epiphenomenalism, in its strict sense in traditional philosophical contexts, is the view that mental phenomena (or certain classes of mental phenomena) make no causal contribution to behavior. In cognitive science contexts, 'epiphenomenalism' is often used in a limited or relative sense. In this usage, mental phenomena are said to be epiphenomenal if they make no causal contribution to behaviors to which they are generally taken to be causally relevant, even if it is admitted that they have *some* effects.

For example, researchers may hold that certain thoughts have no effect on nonlinguistic behavior, even though they allow that those thoughts causally contribute to subjects' reports that they have had those thoughts. This stance makes the thoughts 'epiphenomenal' relative to the nonlinguistic behaviors of interest, and researchers who take such a stance will be regarded as epiphenomenalists, even though their provision for efficacy upon reports implies that they are not epiphenomenalists in the strict philosophical sense.

An often used analogy for epiphenomenalism, taken from Thomas Huxley (1874), compares mental phenomena to steam whistles, which are sounded by the machinery of locomotives, but have no role in moving a train. This analogy suggests arguments that are relevant to both strict and relative epiphenomenalism but actually illustrates only the latter view. Steam whistles do have effects – they vibrate the air and causally contribute to people's getting out of the way of the train; but they are epiphenomenal relative to the behavior of the locomotive, upon which they have negligible effect.

Although a sharp division cannot be made, discussions of epiphenomenalism in cognitive science contexts are often primarily directed at some form of relative epiphenomenalism. Accordingly, this article will emphasize relative epiphenomenalisms. A full treatment of the strict view, which is discussed primarily by philosophers, can be found in Robinson (1999).

HISTORY

The view now known as epiphenomenalism arose in the nineteenth century, often without a convenient name, and sometimes under Huxley's term 'automatism'. Huxley and others (e.g. Hodgson, 1870; Clifford, 1874; Maudsley, 1886), were impressed with 'automatic' behavior. For example, lesioned frogs that were so unresponsive that their consciousness could be reasonably doubted would nonetheless swim perfectly if thrown into water. A soldier who had suffered a head wound in battle suffered periods in which he would carry out extensive routines such as taking a position behind a tree, aiming, and firing – but with only a cane, and no enemy in sight. During these bouts, he was insensitive to pinpricks and shocks, and to many sensory inputs. Interruption of some activities, e.g. removing a piece of paper on which he was writing, would go unnoticed, and he would simply continue his letter on the next sheet. Cases of this kind suggested that consciousness is not necessary for highly organized activity, and the conclusion was drawn that consciousness floats along with underlying brain processes, which are by themselves sufficient to bring about the behavior.

The term 'epiphenomenalism' began to be used in the 1890s and was probably taken over (most likely by William James) from a medical use, in which 'epiphenomena' are symptoms that appear along with those caused by a serious disease, even though they are not actually an effect of the

main disease being treated. The earliest use of the term in its present meaning known to this author occurs in William James's *Principles of Psychology* (1890).

THE CENTRAL ARGUMENT FOR STRICT EPIPHENOMENALISM

Strict epiphenomenalism rests on three premises, of which the first is the causal closure of the physical. This principle says that each physical event has a set of causes, all of which are physical and which, taken together, are sufficient by themselves to cause the event in question. The second premise is the obvious and noncontroversial claim that behavior consists of physical events. Behavior entails either motion of the body, or the holding still of the body (or some of its parts). In either case, behavior depends on contraction states of muscles, and muscles are plainly physical things.

The third premise that is required to lead to traditional epiphenomenalism is the much more controversial assumption that the mental events to which epiphenomenalism applies are not identical with any physical events. This claim must be understood narrowly. In ordinary life, we may view the eating of a red, ripe, sugar-filled tomato as a single event. The redness of the tomato, however, is a different property from its sugariness, and from a nutritionist's point of view, we must distinguish the eating of a red tomato from the eating of a sugar-filled fruit or vegetable. Analogously, it will not be enough to avoid epiphenomenalism if we merely hold mental properties and physical properties to be different properties of the same events; unless the properties themselves are identical, the causal work can be attributed to the physical properties and the mental properties will be as epiphenomenal with respect to behavior as the redness of the tomato is to nutrition.

With these understandings, we can put the argument for strict epiphenomenalism in the following summary way.

(CC) The set of physical events is causally closed. (causal closure)

(PB) Behavior consists of physical events. (physicality of behavior)

(NI) Mental properties of events are not identical with physical properties. (non-identity)

The first two premises tell us that the causes of behavior are exclusively physical events, and this result, together with (NI), leads to the epiphenomenalist conclusion:

(EC) Mental events cause neither behavior, nor any of the causes of behavior.

A very general principle behind this conclusion is that if {C} is a complete set of causes of an event, E1, i.e. a set that is jointly sufficient to bring about E1, then every event that is not in a strong sense identical with some member of {C} is excluded from being a cause of E1.

Many cognitive scientists would hesitate to assert (NI). They may, however, agree that some particular behavior, B1, has a sufficient set of causes, {C1}, and they may apply the general principle just enunciated. Thus, if some mental event, M, can be shown *not* to be a member of a sufficient set of causes {C1}, M will thereby be excluded from being a cause of B1.

DISADVANTAGES OF EPIPHENOMENALISM

Two important objections apply to strict, but not to relative, epiphenomenalism. (1) It conflicts with the deeply entrenched ideology of materialism. Epiphenomenalists respond by focusing on (NI). They note that even if one assumes materialism, one cannot explain how some mental events (e.g. episodes of feelings, e.g. nausea) could be identical with neural events – see, e.g. Levine (1983) or McGinn (1991). They are thus apt to regard denials of (NI) as doctrinaire. Moreover, some criticisms of (NI) are arguably fallacious – see Chalmers (1996). (*See* **Explanatory Gap**)

(2) Epiphenomenalism is self-stultifying. That is, according to strict epiphenomenalism (but not necessarily to relative epiphenomenalism) a report of a mental event is not caused by a mental event. If it is premised that knowingly reporting X requires the report to be causally influenced by X, then the epiphenomenalist's claim to know occurrences of mental events must be counted as false. So, epiphenomenalists who assert that they have mental events that are not causally contributory to any behavior (including reports) will have to be regarded as asserting something that, according to their view, they cannot know. Epiphenomenalist responses to this objection involve arguments against the key premise that knowingly reporting about X always requires causal influence by X on the reporting.

A third objection applies to both strict and relative forms of epiphenomenalism. This objection

affirms that in many cases we have immediate and virtually certain knowledge that our mental events are causally contributory to our behavior. Denials of causal influence in such cases have the appearance of flying in the face of all common sense and reason. Naturally, the force of this objection depends on the particular kind of mental event involved, and we shall consider several cases below. There is, however, a general response that is available to both species of epiphenomenalist: namely that, as Hume (1739, 1748) taught us, causal connection is never simply an object of our experience. Moreover, it is well understood that common effects A and B of an underlying event, C, often falsely appear to us as cases in which A causes B. This is exactly the appearance that epiphenomenalism alleges may occur in cases where A is a mental event, B is behavior, and C is a neural event causing both.

EPIPHENOMENALISM AND COGNITIVE SCIENCE

There are several types of mental phenomena, and one can be an epiphenomenalist with respect to one, or some, of these types without being an epiphenomenalist for all types. This article distinguishes three types of mental phenomena: phenomenal consciousness, thoughts, and will. Arguments or conclusions appropriate to one of these types may or may not apply to other types.

Phenomenal Consciousness

Under 'phenomenal consciousness' we include bodily sensations (pain, itch, nausea, etc.), qualities in perception (colors, tastes, odors, the feeling of warmth, etc.), imagery, and the 'feeling aspects' of emotional states (the way rage, fear, elation, remorse, etc., feel to us). These qualities are often referred to by the terms 'qualia', '(real) seeming', 'sensory consciousness' and many others. They are focal examples of Block's (1995) term 'phenomenal consciousness' or 'P-consciousness'. (*See* **Qualia**)

From a cognitive science point of view, it is a natural (and vexed) question to ask whether phenomenal consciousness has a function, and if so, what this function might be. Evidently, if one accepts (NI), there can be no behavioral function for qualia. But even if one denies (NI), one may worry that, except for reports of occurrences of qualia, qualia have no effect on behavior. (*See* **Consciousness, Function of**)

Several instances of this worry are discussed in a centrally important (and highly controversial) article by Max Velmans (1991). Here is Velmans's own list of areas in which the role of consciousness has been investigated: 'in the analysis and selection of stimuli, in learning and memory, and in the production of voluntary responses, including those requiring planning and creativity' (Velmans, 1991, p. 666).

We may illustrate the way in which doubts about a function for qualia may arise by considering a recent controversy raised by the experimental work of Weiskrantz (1986, 1988) and discussion of blindsight by Block (1995). Blindsight patients have sustained scotoma, i.e. damage to their occipital cortex. When objects are presented in the (distal) region that corresponds to the area of their scotoma, they say they cannot see anything. Nonetheless, when asked to grasp objects in this region (the 'blindfield'), they often form their hands with the correct shape and orientation for grasping the presented object (Marcel, 1983), and, when forced to 'guess' whether an X or an O is being presented, they exhibit above-chance performance. A natural (but hardly uncontroversial – see Dennett, 1991, and the open peer commentaries on Block, 1995) interpretation of these results is that information processing of a kind normally associated with perception can occur in the absence of phenomenal consciousness. If this interpretation is accepted, we must entertain the possibility that phenomenal consciousness has little or no role to play in the information processing that leads to behavioral reactions to perceived objects, even in normal cases. (*See* **Blindsight**)

In his open peer commentary on Velmans's article, Block (1991) points out that it does not follow that (phenomenal) consciousness plays no role in normal cases, even if it is allowed that information processing leading to useful response to inputs *can* be done without phenomenal consciousness. The general point here is well taken: the fact that X can be accomplished by Y does not entail that, in some other particular case, Z is not causally contributory to X. (The fact that hydraulic lifts exist does not show that pulleys can never have the function of raising weights.) However, if it is known that there is an unconscious mechanism that can perform complex information processing in the human brain, we must at least entertain the possibility that that is the mechanism that regularly performs the information processing function. (*See* **Unconscious Processes**)

The claims of such a possibility to be the actual case are strengthened if phenomenal consciousness arises late in cognitive processing. Many of the cases reviewed by Velmans involve temporal

considerations, and give rise to arguments of this form: consciousness arises only *after* certain other important functions (e.g. stimulus selection) have been completed, therefore consciousness was only an accompaniment, and not a causal contributor, to those functions.

Before turning to some further cases, we must note that although many of Velmans's commentators took him to be supporting epiphenomenalism, Velmans himself disavowed such an aim. The explanation is that Velmans assumed a theory that is not widely accepted, and that was not developed in detail in the original, target article. This theory is that causation is relative. In third-person perspective, neural causes that have only late-occurring conscious results are sufficient for information processing that leads to (nonreporting) behavior, and so there is no room for behavioral function for consciousness (save, possibly, in reporting late results). This point is what led many readers to attribute epiphenomenalism to Velmans. His view, however (see author's responses to commentaries, and Velmans, 2000), was that in first-person perspective conscious states *are* regarded as causes of behavior; and, further, the first-person perspective is as legitimate as the third-person perspective. Thus, Velmans himself did not regard it as correct to say that consciousness does not contribute to behavior – it does do so, from the first-person perspective.

There are many cases of implicit processes which, like blindsight, suggest that events that are sometimes taken to be caused by conscious states may have a causal background that bypasses such states. (But see Dulaney, 1997 for an alternative view.) For example, visual extinction patients cannot identify objects in their left visual field when objects are also presented to the right visual field. Some of these patients regard requests to compare left and right presented line drawings as 'silly', saying that only one object has been presented; other patients are aware of something on the left but cannot say what it is. In both cases, however, forced choice judgments as to whether the objects are the same or different range from 88 to 100 per cent correct (Volpe *et al.*, 1979). Prosopagnosics can show skin conductance responses to faces of family members or famous people whom they are unable to identify by name (Tranel and Damasio, 1985). This evidence suggests that our emotional reactions, even in normal cases, may not depend on first consciously (or explicitly) recognizing the people we are seeing. Zajonc's (1980) familiarity effect likewise suggests that preferences are sometimes formed by processes other than the rational considerations we usually suppose to underlie our preference judgments. Amnesics who perform poorly when asked to name words from a studied list nevertheless show robust effects of previous exposure when tested with less direct, implicit tasks such as word-stem completion. (See Schacter, 1987 for review.) This result suggests that consciousness is epiphenomenal for certain aspects of memory. (*See* **Implicit Cognition**; **Prosopagnosia**)

These phenomena and many others of a similar nature (see Reber, 1997; Köhler and Moscovitch, 1997) suggest that, even in normal subjects, responses that are naively attributed to conscious recognition may occur as a result of unconscious processes.

Thoughts

A salient and cherished part of our self conception is our belief that we know the reasons why we think and act as we do. A widely held analysis, explicitly argued for by Davidson (1963), is that, in nondeviant cases, reasons for action are causes of those actions. Similarly, we usually believe that the reasons we might give for our beliefs reflect the actual causes of our holding those beliefs. If we find that reasons that we believe to be our reasons for acting or holding certain beliefs are not causes of those actions or beliefs, those reasons will be relatively epiphenomenal. Our thoughts about those reasons may indeed still causally contribute to our reporting of those reasons; but the reasons we attribute to ourselves will be epiphenomenal relative to the behavior or beliefs for which they are regarded as reasons, i.e. they will in fact not have the causal role they are conceived to have.

It is now widely accepted that under many circumstances people confabulate, that is, they sincerely offer reasons for facts about themselves that can be demonstrated *not* to reflect the actual circumstances that causally determine those facts. For example, Nisbett and Wilson (1977b) manipulated subjects' views about attractiveness of physical appearance, speech, and mannerisms of a videotaped presenter, by varying the style of the presentation between 'warm' (agreeable responses to questions) and 'cold' (rigid, intolerant responses). The causal direction was demonstrably from style difference to attractiveness-judgment difference in both conditions. Nonetheless, subjects who saw the 'cold' version of the presentation were quite firm in their (false) belief that their negative attractiveness judgments had contributed to their disliking of the presentation. In another experiment, Nisbett and

Wilson (1977a) asked a group of subjects to read a text that presented a novelistic scene with high emotional impact. Three other groups read a text that was the same except for the omission of one passage, or another passage, or both passages. Ratings of emotional impact were not significantly different across groups; nonetheless, when asked about the effect of the manipulated passages on their judgments, subjects in groups that had been exposed to them expressed the definite belief that the passages had increased the impact of the text. In still another experiment, subjects erroneously claimed that extraneous noise (from a nearby power saw) had lowered their rating of a film.

Many of the experiments presented or described in Nisbett and Wilson (1977a) demonstrate only failures of ability to report actual causes of behavior or beliefs. Such failures show that very complex cognitive processing can occur without our being aware of it, and thus indirectly support the *possibility* of epiphenomenal thoughts. The examples cited, and other cases reviewed in Nisbett and Wilson (1977a), go further; they are examples of beliefs that do not have the effects on behavior or judgment that their possessors take them to have. They may cause the reports of those beliefs, but they are epiphenomenal relative to their presumed additional effects. (It should be noted that this attribution is ours; the term 'epiphenomenal' does not occur in the cited papers of Nisbett and Wilson.)

There is no suggestion here that all our beliefs about our reasons are epiphenomenal, i.e. that we are always misguided about our reasons. Nisbett and Wilson (1977a) argue, however, that correct apprehension of our reasons depends on factors that are available to observers in general. They hold that introspection unsupported by publicly available knowledge is not a reliable source of knowledge about our own cognitive processes or the causal relations surrounding our reasons.

Decisions

In many cases, having a reason to act in a certain way cannot be taken to be sufficient for producing the action, because the action is deferred for some time. In some of these cases, the action is to be triggered by an expected external stimulus; in other cases, however, precise timing is left indefinite, and is not tied to any definite occurrence. At some point, we decide that the time to act is *now*, and we begin to carry out the action.

These remarks are not controversial. The last mentioned case, however, is one that may suggest to us that our conscious decisions are the causes of the initiation of our actions, and even that we know by direct experience that conscious decisions cause us to act. With these claims, we move into territory that has seen disputation from at least two lines of experimental work.

The older, and now widely known argument stems from investigations by Libet and coworkers of the time at which decisions are made (see Libet, 1985 for discussion and further references). Libet's subjects were free to make a movement (of finger or wrist) when they formed an urge (or desire or intention) to do so. Subjects watched a rotating spot, and reported its clock position (C) when they formed the intention to make a movement. At the same time, Libet used a scalp electrode to measure the time of a 'readiness potential' (RP), which characteristically precedes bodily movements. In brief, Libet found that the onset of RP was about 345 milliseconds prior to the clock position corresponding to the time of the awareness of intention to act. The suggested conclusion is that spontaneous actions are actually initiated by unconscious brain events, with consciousness of intention to act occurring only after the events that bring about the action are already under way. In that case, the sense that it is a conscious intention that initiates the action must be illusory. (*See* **Free Will**)

Two caveats must be entered here. (1) Libet himself allows for the possibility of conscious control in the form of a 'veto' that would occur between 10 and 50 milliseconds before the action (and thus between 60 and 100 milliseconds after the conscious intention). Several of the commentators on Libet (1985), however, have found this allowance arbitrary, and contrary to the main epiphenomenalistic thrust of Libet's results. (2) At the level of precision at which Libet is working, sensory transduction times, neural transmission times, time necessary to shift attention from one item to another, etc., must all be taken into account. Values of these variables are difficult to establish and not always known with exactness. Interpretation of Libet's data is thus an unusually complex matter and is extremely controversial. (See Dennett, 1991, and the commentators on Libet, 1985. Libet has replied to the latter in the author's response section of Libet, 1985.) (*See* **Event-related Potentials and Mental Chronometry**)

A second line of investigation has recently been undertaken by Wegner and Wheatley (1999), who asked a subject and a confederate of the experimenter to place their fingertips on a small board mounted atop a computer mouse. Movement of this apparatus changed the position of a cursor on

a monitor that displayed pictures of a large number of familiar objects. Participants and confederates wore earphones through which they could receive various instructions, and through which participants sometimes heard a word for one of the depicted objects. On a small number of trials (forced-stop trials), confederates received instructions to force a stop on a certain picture; in most trials (non-forced-stop trials) they let participants roam and stop as they would. Participants were given a musical cue to begin an interval at the end of which a stop was to occur. After each stop, participants rated their contribution on a scale from 'I allowed the stop to happen' to 'I intended to make the stop', and confederates pretended to do the same. Participants were *not* instructed to stop on pictures of named objects, and an ancillary experiment showed that hearing the word for a depicted object did not influence participants toward stopping on its picture. The key finding was that participants rated their own participation higher in forced-stop trials when they heard the word for the picture on which the stop was forced 1 or 5 seconds prior to the stop, as compared with forced-stop trials in which they heard the word 30 seconds before the stop, or 1 second after. Wegner and Wheatly (1999, p. 489) report their conclusion in this way: 'Apparently, the experience of will can be created by the manipulation of thought and action in accord with the principle of priority, and this experience can occur even when the person's thought cannot have created the action'. (The principle of priority is that the experience of will requires thought to precede the action at a proper (i.e. small) interval.)

Wegner and Wheatley's result is silent on the matter of whether the experience of will causes the report of such experience, and is thus silent on the question of strict epiphenomenalism. But it does provide experimental support for a relative epiphenomenalism; that is it supports the view that experienced intent need not be a cause of events for which it is taken to be a cause by the participants in whom the experience occurs. Further, the possibility of the illusion that is produced in Wegner and Wheatley's experiment is explainable by a model in which, in normal cases, the thought of an action about to be taken, and the action itself, are co-effects of a common brain event.

CONCLUSIONS

While even relative epiphenomenalism is a controversial topic in cognitive science, there is evidence from several sources that strongly suggests that commonsense views about the causal relations between conscious occurrences and actions or beliefs are overly simple and misleading.

Relative epiphenomenalism does not imply strict epiphenomenalism, so a finding that there are relatively epiphenomenalistic mental phenomena does not directly support strict epiphenomenalism. Nonetheless, strict epiphenomenalists should be interested in the relatively epiphenomenalistic cases that cognitive science at least seems to provide. The relation can be summarized in the principle that 'the enemy of my enemy is my friend'. Nonmetaphorically, the point is that one of the key arguments against strict epiphenomenalism is the apparent introspective certainty of the existence and direction of causal connection between some of our mental phenomena and other mental phenomena or our actions. To the extent that results from cognitive science weaken the appearance that we have such certainty, they weaken an important objection to (even) the strict form of epiphenomenalism.

References

Block N (1991) Evidence against epiphenomenalism (peer commentary on Velmans (1991)). *Behavioral and Brain Sciences* **14**: 670–672.

Block N (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**: 227–247.

Chalmers DJ (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Clifford WK (1874) Body and mind. Lecture originally given to the Sunday Lecture Society, 1 November 1874. Published in *The Fortnightly Review*, December, 1874, n.s. **16**: 714–736. Reprinted in Stephen L and Pollock F (eds) *Lectures and Essays of the late W. K. Clifford*. London: Macmillan, 1879.

Davidson D (1963) Actions, reasons, and causes. *The Journal of Philosophy* **60**: 685–699.

Dennett DC (1991) *Consciousness Explained*. Boston: Little, Brown & Co.

Dulaney DE (1997) Consciousness in the explicit (deliberative) and implicit (evocative). In: Cohen JD and Schooler JW (eds) *Scientific Approaches to Consciousness*. pp. 179–212, Mahwah, NJ: L. Erlbaum.

Hodgson S (1870) *The Theory of Practice*. London: Longmans, Green, Reader, and Dyer.

Hume D (1739) *A Treatise of Human Nature*.

Hume D (1748) *An Inquiry Concerning Human Understanding*.

Huxley TH (1874) On the hypothesis that animals are automata, and its history. *The Fortnightly Review*, n.s. **16**: 555–580. Reprinted in *Methods and Results: Essays by Thomas H. Huxley*. New York: D. Appleton and Company, 1898.

James W (1890) *The Principles of Psychology*. New York: H. Holt.

Köhler S and Moscovitch M (1997) Unconscious visual processing in neuropsychological syndromes: a survey of the literature and evaluation of models of consciousness. In: Rugg MD (ed.) *Cognitive Neuroscience*, pp. 305–373. Cambridge, MA: MIT Press.
Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.
Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* **8**: 529–566.
Marcel AJ (1983) Conscious and unconscious perception: an approach to relations between phenomenal experience and perceptual processes. *Cognitive Psychology* **15**: 238–300.
Maudsley H (1886) *Body and Mind: An Inquiry into Their Connection and Mutual Influence, Specially in Reference to Mental Disorders*. New York: D. Appleton and Co.
McGinn C (1991) *The Problem of Consciousness*. Oxford: Blackwell.
Nisbett RE and Wilson TD (1977a) Telling more than we can know: verbal reports on mental processes. *Psychological Review* **84**: 231–259.
Nisbett RE and Wilson TD (1977b) The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* **35**: 250–256.
Reber AS (1997) How to differentiate implicit and explicit modes of acquisition. In: Cohen JD and Schooler JW (eds) *Scientific Approaches to Consciousness*, pp. 137–159. Mahwah, NJ: Erlbaum.
Robinson WS (1999) Epiphenomenalism. In: *Stanford Encyclopedia of Philosophy*. Available at [<http://plato.stanford.edu/entries/epiphenomenalism/>].
Schacter DL (1987) Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition* **13**: 501–518.
Tranel D and Damasio AR (1985) Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. *Science* **228**: 1453–1454.
Velmans M (1991) Is human information processing conscious? *Behavioral and Brain Sciences* **14**: 651–726. (Includes open peer commentary by several commentators and author's response.)
Velmans M (2000) *Understanding Consciousness*. London and Philadelphia: Routledge.
Volpe BT, Ledoux JE and Gazzaniga MS (1979) Information processing of visual stimuli in an 'extinguished' field. *Nature* **282**: 722–724.
Wegner DM and Wheatley T (1999) Apparent mental causation: sources of the experience of will. *American Psychologist* **54**: 480–492.
Weiskrantz L (1986) *Blindsight*. Oxford: Oxford University Press.
Weiskrantz L (1988) Some contributions of neurophysiology of vision and memory to the problem of consciousness. In: Marcel AJ and Bisiach E (eds) *Consciousness in Contemporary Society*. Oxford: Oxford University Press.
Zajonc RB (1980) Feeling and thinking: preferences need no inferences. *American Psychologist* **35**: 151–175.

Further Reading

Baars BJ (1988) *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.
Bolton D and Hill J (1996) *Mind, Meaning and Mental Disorder: The Nature of Causal Explanation in Psychology and Psychiatry*. New York: Oxford University Press.
Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press/Bradford.
Heil J and Mele AR (eds) (1993) *Mental Causation*. Oxford: Oxford University Press.
Libet B, Freeman A and Sutherland K (eds) (1999) *The Volitional Brain: Towards a Neuroscience of Free Will*. Thorverton: Imprint Academic.
Mahoney MJ (1995) Cognition and causation in human experience. *Journal of Behavior Therapy and Experimental Psychiatry* **26**: 275–278.
Murphy ST and Zajonc RB (1993) Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology* **64**: 723–739.
Sperber D, Premack D and Premack AJ (eds) (1995) *Causal Cognition: A Multidisciplinary Debate*. New York: Oxford University Press.
Velmans M (ed.) (1996) *The Science of Consciousness: Psychological, Neurophysiological, and Clinical Reviews*. London: Routledge.

Epistemology

Intermediate article

Joseph Cruz, Williams College, Williamstown, Massachusetts, USA

CONTENTS

Epistemology is the philosophical study of what is required in order to have rational beliefs and knowledge. Both traditional a priori *methods of philosophy and* a posteriori *methods of cognitive science have been brought to bear on this question.*

WHAT IS EPISTEMOLOGY?

Epistemology answers to a daunting variety of senses in the humanities and the social sciences. Even when we restrict our attention to epistemology as it is understood in contemporary Anglo-American philosophy, the only uncontroversial claim we can make is that epistemology is an attempt to make sense of the possibility, nature, and limits of human intellectual achievement. Typically, the epistemologist does this by trying to illuminate the difference between knowledge and opinion, or the difference between good reasoning and poor reasoning. This project is distinct from merely giving a descriptive account of what people claim to know or to believe reasonably. Instead, epistemologists try to understand what it is really to know or really to believe reasonably, even if people routinely fail to know or are frequently irrational. Moving beyond the descriptive details of knowledge or belief formation to what people ought to believe is a normative philosophical enterprise.

Construed one way, epistemology aims to understand general and ubiquitous elements of human inquiry, such as perceptual knowledge or inductive inference. This project has sometimes been fueled by skeptical doubts about the veracity of our senses or the trustworthiness of our reasoning. Not all philosophers are persuaded, however, that thoroughgoing skepticism allows for, or requires, a response. As a result we often find in contemporary epistemology the attempt to account for epistemic achievements in a way that does not necessarily offer a reply to the skeptic.

Construed another way, epistemology aims to investigate specific domains of knowledge or rational belief. Some aspects of the philosophy of science may thus be understood as constituting a subfield of epistemology. This kind of research may be narrowed to particular sciences, such as the philosophy of psychology or the philosophy of cognitive science (or physics, or biology, to name two more prominent areas). Efforts to understand the nature of explanation in cognitive science ultimately fall under the umbrella of epistemology, though such efforts lie at the intersection of philosophy of science and philosophy of mind and have thereby taken on a robust theoretical autonomy.

Another distinction may be drawn between epistemology oriented towards individuals and epistemology oriented towards social institutions or practices. Some philosophers claim that there are social practices that positively or negatively influence the formation of knowledge or rational belief. A strong version of this view is where knowledge or rationality is exhausted by socially-mediated factors. This position has currency in some fields of the humanities, but the epistemologists who interact most frequently with the cognitive sciences tend to reject the most radical forms of social epistemology.

We shall concentrate on the general construal of individual epistemology and its relation to cognitive science.

Knowledge

What is knowledge? One sort of knowledge is the kind expressed by 'that' clauses. For instance, we may say 'Carlos knows that Oaxaca is in Mexico'. This kind of knowledge is called propositional knowledge, as it is a proposition – in this case, 'Oaxaca is in Mexico' – that is known. Another kind of knowledge is that expressed by 'how to' clauses. Thus, we may say 'Carlos knows how to ride a bicycle'. This second kind of knowledge is called procedural or nonpropositional knowledge. The bulk of attention in contemporary epistemology has been on propositional

knowledge, but this emphasis should not be viewed as making a claim about which kind is more important in human cognition.

Beliefs are thought to be the primary psychological entities that are candidates for propositional knowledge. Of course, not every belief is an instance of knowledge. Carlos may believe that Oaxaca is in Mexico, but his merely believing it does not seem to be enough to make it an instance of knowledge. Carlos cannot know that Oaxaca is in Mexico if it is false that Oaxaca is in Mexico. So one additional thing required for a belief to be knowledge is that the belief be true.

At least since Plato, epistemologists have thought that true belief, while necessary, is still insufficient for knowledge. One way of appreciating this is to note that beliefs that are accidentally true are not knowledge. Suppose that Carlos is guessing that Oaxaca is in Mexico. In that case, his belief does not seem to count as an instance of knowledge. One prominent and historically important way of pursuing the distinction between accidentally true and nonaccidentally true belief is to rely on the goodness of the reasons for the belief. So, if Carlos' belief that Oaxaca is in Mexico is both true and based on his having visited that state, we are likely to have a case of knowledge (assuming, for this example, that his visiting Oaxaca is a source of good reasons). On this conception of knowledge three central projects of epistemology emerge. First, the epistemologist must determine what constitute reasons. Second, she or he must offer a general account of what makes some reasons good. And third, the epistemologist must illuminate the nature of the relationship between reasons and beliefs.

Unfortunately, the 'goodness of reasons' approach to knowledge seems susceptible to a curious kind of problem that shows that true beliefs for which we have good reasons still might not count as knowledge. I may believe and have very good reasons to believe, for instance, that 'Mary owns a Honda'. I could then, for the purposes of teaching a logic course, create a variety of compound sentences involving my belief that Mary owns a Honda. Suppose I propose that 'Mary owns a Honda OR Sally is in Paris' even though I do not have any idea where Sally is. In the normal case, if I know one half of this disjunctive compound sentence, then I know the entire sentence since arbitrary disjuncts cannot change the truth-value of the entire sentence no matter how improbable they are. Now imagine that, contrary to my excellent evidence, Mary does not own a Honda, and, completely coincidentally, Sally *is* in Paris. It then seems that my belief in the whole expression 'Mary owns a Honda or Sally is in Paris' is true (Sally's being in Paris makes it so) and is something that I have good evidence for (since I have good evidence that Mary owns a Honda even though she does not). The problem is that we would not be inclined to attribute knowledge to me, though the three conditions for knowledge – true belief with good reasons – have been satisfied.

This is called the Gettier problem, after the author of the short article that sparked contemporary interest in it (Gettier, 1963). The Gettier problem has inspired many putative solutions, counterformulations, and modified putative solutions. There is no agreement on how the Gettier problem is to be solved within the goodness of reasons approach, but many epistemologists think that what the problem reveals is that there must be some fourth condition to ensure that truth and good reasons will be tied together in knowledge. Alternatively, the puzzle may show that the concept of knowledge is ill-defined. The Gettier problem has by no means crippled epistemology. Even if knowledge is not a concept that epistemologists can characterize in any simple way, issues in epistemology having to do with the nature of reasons and belief remain.

We shall turn to these issues shortly, but first it should be noted that there are accounts of knowledge that do not rely on the goodness of reasons approach. Some of these are pursued because they seem to offer a way to avoid the Gettier problem, while others seem to hold greater promise in defeating skepticism. We can divide these accounts of knowledge that do not rely on the goodness of reasons into two categories. One is where beliefs are calibrated to the truth in a way that can be characterized by epistemologists even though the relation between belief and truth may not involve reasons (Nozick, 1981). The knower does not have knowledge by virtue of having reasons for a belief; rather, a belief is an instance of knowledge if certain metaphysical or logical facts about the relation between belief and truth are satisfied.

The second category is *contextualism*, which maintains that the truth of attributions of propositional knowledge vary from context to context (DeRose, 1995). For instance, we may properly claim that Carlos knows that Oaxaca is in Mexico in mundane conversational contexts while the same claim would be improper in the rarified context of a university seminar on skepticism. Developing contextualist theories of knowledge requires an elaboration of what makes some contexts more demanding than others as well as a systematic treatment of changes in context.

Justification, Rationality, and Warrant

So far we have been identifying epistemology with the elucidation of knowledge. Being an instance of knowledge, however, is not the only epistemically positive characteristic a belief might have. We may be interested in having good reasons for our beliefs without insisting that such beliefs be instances of knowledge. The possibility of independently exploring epistemically positive beliefs that fall short of knowledge was implicit above when the three central projects regarding reasons (their nature, what makes some good, and what their relationships are) were proposed. In spite of their potential independence from knowledge, though, all three of these projects are typically conceived against the background of treating truth as the fundamental aim of epistemic reasoning. Good reasons are therefore often understood in the first instance to be reasons in favor of taking a belief to be true. Maintaining a close link between the goodness of reasons approach and truth is a deep commitment in epistemology, and explains the confidence in the connection between rationality and knowledge. Still, some epistemologists have offered accounts that sever the link by treating good reasoning as wholly unconnected to the truth.

The most prominent properties of epistemically positive beliefs discussed by philosophers are justification and rationality. These labels carry with them some connotative differences. Except where the distinction is crucial, in this essay we will read 'justification' and 'rationality' as synonyms. *Warrant* also has some currency in describing what a belief must have, in addition to truth, in order to yield knowledge.

How should we undertake evaluations of beliefs in terms of justification or rationality or warrant? The goodness of reasons strategy for investigating knowledge again affords a persuasive framework. We may study the reasons for a belief in order to make some judgment about whether that belief is epistemically positive. Furthermore, if rationality or warrant are graded notions, the strategy may allow us to advance a scale that appeals to the comparative goodness of the reasons for a belief. Thus, we may account for judgments of 'more rational' or 'more warranted' belief.

There seems to be an immediate problem, though, with employing the strategy of reasons for assessing epistemically positive belief along the dimensions of justification or rationality. A belief will inherit the epistemic status of the reasons for it, so we must in turn determine whether the reasons for a belief are themselves rational or justified. This threatens a regress of reasons that must be resolved. Responding to this regress has been instrumental in crystallizing issues in the recent history of epistemology, even if the regress problem no longer occupies the most crucial role in contemporary epistemological research. Deflecting the regress of reasons argument does not by itself answer the question of what makes a reason justified or rational. An answer to the structural question seems like a necessary first step, but it cannot be a complete theory of epistemically positive beliefs.

The structure of the belief corpus: foundationalism and coherentism

Two responses to the regress problem that differ on the structure of the belief corpus are foundationalism and coherentism. The foundationalist claims that there is a set of basic beliefs that do not require reasons to explain their epistemically positive nature because of some special characteristic(s) that they have. The epistemic credentials of beliefs that are not foundational are due to a traceable lineage through reasons, from basic beliefs via a basing relation that must be illuminated by the foundationalist. Thus, the justification relationship in the foundationalist picture is asymmetrical.

Foundationalists have attempted to develop axioms of goodness for foundational beliefs. Candidates for axioms include the claim that beliefs about how things appear are intrinsically epistemically positive. In a like fashion, foundationalists will need to provide enough axioms for foundational beliefs to account for the credentials of all epistemically positive beliefs.

Modulating the claim that the regress of reasons must end in intrinsically epistemically positive beliefs can complicate the foundationalist picture. It is an open question whether reasons for beliefs must themselves be beliefs. If not, then it is possible that the foundational reasons are other cognitive states such as perceptual states or memory states.

In contrast to foundationalists, coherentists maintain that no beliefs are intrinsically epistemically positive (Bonjour, 1985). By their lights, every belief relies on other beliefs for its epistemic status. One (controversial) argument for coherentism is that the foundationalist's putative axioms of goodness require an argument to show that they are good, and any such argument will rely on further beliefs, indicating that the axioms are not foundational after all. On one reading of coherentism, beliefs are epistemically positive based on a lineage of reasons in a structure that may ultimately loop back onto itself. There may be no need, however, to

trace reasons in a way that is circular. One might instead claim that a belief is epistemically positive in the case of its being a member of a coherent belief corpus without pursuing particular reasons in a linear fashion.

Coherentists discharge the task of revealing what makes some reasons good by claiming that cogent arguments can be given for the high epistemic credentials of beliefs, and further arguments can be given in favor of the cogent arguments.

Epistemic internalism

Despite their differences on the structure of the belief corpus, the foundationalist and coherentist strategies traditionally agree on a different issue: reflective, careful agents are able to make assessments of their own beliefs in order to determine whether they are epistemically positive. This is called the *internalist* conception of epistemic justification or rationality.

In order to defend internalism in detail, the epistemologist needs to specify what is meant by the claim that agents can determine whether there are good reasons for their beliefs. A number of interacting but independent issues arise here. One challenge is deciding whether the epistemic agent must be able to determine that reasons are *good*, versus the less strict demand that the agent merely be able to determine what the reasons *are*. Another challenge is whether the agent must be able to determine *what* her or his good reasons are for a belief, versus the much less demanding constraint that the agent must be able to determine *that* she or he has good reasons for a belief. A third challenge is deciding what sort of effort on the part of the agent will be consistent with the claim that agents can access their reasons. Most of the combinations of answers to these challenges have been defended in the internalist literature.

Internalism originates in three related concerns. First, one of the projects that sometimes accompanies an assessment of epistemically positive belief is to illuminate how one might improve one's beliefs. If improvement is to be possible, it needs to be possible to determine which belief among many candidate beliefs is most epistemically positive, and it has seemed that the epistemic agent personally needs to be able to make the judgment. Second, justification (though not rationality) has often been viewed as at least partly a matter of fulfilling a distinctly epistemic duty. Fulfilling a duty seems to require that one be able to do the things that duty requires. In order to secure the means to an intellectual duty, an epistemic agent will need to be able to reflect on her or his condition and on the resources available. Third, recall that the ability to answer the skeptic is sometimes thought to be a crucial component of epistemology. The only answers that the epistemic agent can give, though, are ones that are available on reflection.

Epistemic externalism

Thinking of the epistemic agent as able to determine when beliefs are epistemically positive is not demanded by the overarching goal of epistemology. Though the philosopher may state the conditions that must be met for a belief to be epistemically positive, the epistemic agent may not be able to make assessments of particular beliefs. Thus one might reject internalism and not expect the right philosophical account of epistemically positive belief to enable meliorative, duty-oriented, or skeptic-answering evaluations to take place. This is the *externalist* view in epistemology.

For example, *process reliabilism* – the best-known externalist view – to a first approximation claims that what confers positive epistemic status on a belief is that it be produced by a psychological process that reliably produces true belief (Goldman, 1979). Though the reliability of a psychological process is often opaque to the person employing that process, process reliabilists think that the reliability of belief-forming processes can be uncovered by cognitive science or by other kinds of empirical inquiry.

The motivations for defending an externalist position are diverse. One simple motivation is the impulse to tie justification and rationality directly to truth via an appeal to truth-sensitive properties such as reliability. In this connection, the externalist's commitment to truth as the central goal of justified belief looms large. Externalist theories can be neutral on the question of the structure of the belief corpus, and instead attempt to tackle more directly the issue of epistemically positive belief.

Another motivation that is prominent in discussions of externalism is the role of causal factors in belief formation. It has seemed to many epistemologists that, in order to be justified or rational, a belief has to have both an evidential and a causal relation to the reasons for it. Incorporating this causal element seems to require some of the specialist's insight into the causes of our beliefs, and that is in tension with the internalist's impulse to insist that it is the epistemic agent personally who is in a position to determine the epistemic status of her or his beliefs.

The Methods of Epistemology

In traditional epistemology, the standard of correctness for epistemological questions appeals to intuitions about epistemic methods or particular cases of belief. For example, that beliefs formed under favorable perceptual conditions in a healthy observer are justified is a principle that we may intuitively certify. So, a non-skeptical account of justification should in this view be designed to accommodate intuitions about beliefs formed under favorable perceptual conditions. Alternatively, a theorist may take particular instances of an intuitively epistemically positive belief – the belief that *this* shiny Macintosh apple is red, for instance – and attempt to build a theory that respects this intuition. A successful theory of justification will yield the result that such beliefs are justified unless there is some other overriding consideration that would result in the retraction of the judgment that the belief is justified. Such beliefs are called 'prima facie justified' in order to highlight the fact that new information about the situation might change intuitions about the justifiedness of the belief in question. Many epistemologists have exploited both kinds of intuition by seeking to balance methods and cases.

There are reasons to be dissatisfied with the methodology of epistemology, not the least of which is that the use of intuitions alone can seem to be a precarious or even spurious basis for theorizing. Intuitions might be thought to be too subjective or theory-laden. There was a period in Anglo-American philosophy when philosophy was viewed as wholly conceptual (*a priori*) analysis. The strong commitment to that position has long since passed. Intuitions ground philosophical inquiry in our pre-theoretic understanding of epistemic concepts, and allow a bridge to the long history of philosophical theorizing on the same topics. But intuitions may need to be adjusted in light of other considerations. There is much debate as to what the other sources of constraint may be. It is within a particular answer in this debate that there is the most cross-fertilization between epistemology and cognitive science.

EPISTEMOLOGY AND COGNITIVE SCIENCE

Some epistemologists maintain that a sensible division of labor in understanding justification and rationality is that philosophers investigate the normative elements of belief formation, while cognitive scientists study how we actually form beliefs. To the extent, however, that philosophical intuitions about epistemic concepts need to be constrained by the empirical (*a posteriori*) details of belief formation, cognitive science will play a significant role in epistemology. Theories of rational belief that incorporate a significant empirical component are categorized under the label of 'naturalized epistemology'.

Naturalized Epistemology

Cognitive science gives epistemologists detailed and empirically robust accounts of the origins of belief. This is relevant to epistemology if we are persuaded that a belief is rational only if it bears the right causal relation to the reasons for it. Empirical details will also be important in part because philosophical intuitions are likely to be misleading with respect to the nuances of belief formation. For example, intuitively it seems that inductive reasoning on the basis of small samples would not be a method for forming epistemically positive beliefs. It has been argued, however, that induction on the basis of small samples is epistemically defensible in reasoning involving natural kinds (Kornblith, 1993). It is only through the insight of cognitive scientific accounts of how reasoning interacts with kind concepts that this unintuitive method will appear epistemologically sound. Other insights from cognitive science that have been relevant in epistemology include vision research, research on deductive reasoning, and research on memory.

Process reliabilism advocates a proprietary role for cognitive science in epistemology. Once causal factors are deemed relevant in understanding justification, and once it has been determined that justified beliefs are the product of reliable psychological processes, cognitive science enters into epistemology to provide the details of psychological processes in terms of their reliability (Goldman, 1986).

Naturalized epistemology is often identified with externalism. Depending on how issues of access to reasons are resolved, though, it is possible to maintain an internalist theory that carves out a role for empirical research on belief formation. Artificial intelligence models of cognition have been employed in epistemology to investigate how complex patterns of reasoning interact in order to yield intuitively justified beliefs (Pollock and Cruz, 1999; Thagard, 2000). The rules or norms of reasoning at each step of reasoning may be unintuitive owing to their complexity, but the whole reasoning process may be counted as rational because it yields an intuitively justified conclusion. This position can

be internalist by maintaining that epistemic agents will be able to determine in their own cases whether their reasoning is good, even though the norms of reasoning generally will be inscrutable.

The Cognitive Science of Epistemology

Finally, we consider another way to conceive of the relationship between epistemology and cognitive science. Cognitive science may step back from philosophy and investigate epistemology itself. There are two plausible targets of investigation: concepts and methods. First, cognitive science might investigate how people employ the concepts of justification, knowledge, or good evidence in the same way that nonphilosophical concepts are studied (Goldman, 1992). Drawing conclusions from such research is a delicate matter, however, since it has seemed to philosophers that there is no way to uncover the genuine normative character of epistemological concepts by merely studying their psychological status.

Second, cognitive science might investigate the method of employing *a priori* intuitions to address epistemic questions. A specific instance of this suggestion is that we should think of epistemic intuitions as analogous to grammaticality judgments in linguistics (Pollock and Cruz, 1999). Reasoning on this view is the product of an underlying rational competence, but actual cases of reasoning might go wrong due to resource constraints. If the analogy could be successfully developed, it would help explain the difference between merely describing how people reason (which would always be particular performances of reasoning) and the normative account of how people ought to reason. The explanatory credentials of this account appeal directly to its success in linguistics and related areas of psychology and artificial intelligence research.

References

Bonjour L (1985) *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

DeRose K (1995) Solving the skeptical problem. *The Philosophical Review* **104**: 1–52.

Gettier E (1963) Is justified true belief knowledge? *Analysis* **23**: 121–123.

Goldman A (1979) What is justified belief? In: Pappas G (ed.) *Justification and Knowledge*, pp. 1–14. Boston, MA: D. Reidel.

Goldman A (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman A (1992) Epistemic folkways and scientific epistemology. In: *Liaisons*, pp. 155–175. Cambridge, MA: MIT Press.

Kornblith H (1993) *Inductive Inference and Its Natural Ground*. Cambridge, MA: MIT Press.

Nozick R (1981) *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Pollock J and Cruz J (1999) *Contemporary Theories of Knowledge*, 2nd edn. Lanham, MD: Rowman & Littlefield.

Thagard P (2000) *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Further Reading

Alston W (1989) *Epistemic Justification: Essays in the Theory of Knowledge*. Ithaca, NY: Cornell University Press.

Bernecker S and Dretske F (eds) (2000) *Knowledge: Readings in Contemporary Epistemology*. New York: Oxford University Press. [This anthology collects many of the most important articles in contemporary epistemology.]

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Foley R (1993) *Working Without a Net: A Study of Egocentric Epistemology*. New York, NY: Oxford University Press.

Goldman A (1993) *Philosophical Applications of Cognitive Science,* chaps 1 and 2. Boulder, CO: Westview Press.

Goldman A (1999) *Knowledge in a Social World*. New York, NY: Oxford University Press.

Harman G (1999) *Reasoning, Meaning, and Mind*. New York, NY: Oxford University Press.

Lehrer K (2000) *Theory of Knowledge*, 2nd edn. Boulder, CO: Westview Press.

Plantinga A (1993) *Warrant and Proper Function*. New York, NY: Oxford University Press.

Ethics

Intermediate article

John M Doris, University of California, Santa Cruz, California, USA
Stephen Stich, Rutgers University, New Brunswick, New Jersey, USA

CONTENTS

Ethics is the philosophical study of standards of conduct and value. Such inquiry often encounters questions regarding human motivation, affect, and cognition that can be addressed by empirical and theoretical work in the cognitive sciences.

INTRODUCTION: WHAT IS ETHICS?

Contemporary philosophical ethics is customarily divided into three subdisciplines, here ordered by increasing theoretical abstraction: applied ethics, normative ethics, and meta-ethics (see Darwall, 1998 for a useful survey).

Applied ethics addresses concrete ethical problems in areas such as law, medicine, business, and agriculture. While work in applied ethics may involve the advocacy of particular positions, philosophers often aim only to clarify, through philosophical methods of argumentation and analysis, the issues surrounding such controversial topics as abortion, euthanasia, capital punishment, and the development of genetically modified organisms.

Normative ethics investigates obligation, virtue, and value; very often, work in this area consists in attempts to adduce general claims about the sorts of things that are morally valuable and general principles for the regulation of conduct and choice. In contemporary normative ethics, three approaches are especially prominent. Proponents of virtue ethics, inspired especially by Aristotle, maintain that human excellence is the fundamental evaluative consideration; conduct is to be judged by reference to such virtues as courage and temperance. Consequentialists and Utilitarians, working in the tradition of Bentham and Mill, argue that right actions are those that, of the options available, maximize the total quantity of a valuable end such as pleasure or happiness. Followers of Kant, or Kantians, argue that an act is morally permissible only in so far as it proceeds from a principle that could coherently serve as a law governing the choice of all rational agents.

Meta-ethics concerns the structure of moral reasoning and justification, the sources of moral motivation, and the nature of moral concepts. Perhaps the most persistent issue in meta-ethics concerns the objectivity of ethical inquiry. Is ethical discourse merely the statement or expression of preference or emotion, as various brands of skepticism about moral objectivity maintain? Or is there some method for discovering objectively true answers to ethical questions, or conferring rational justification on ethical judgments, as moral realists have argued?

THE RELATION OF COGNITIVE SCIENCE TO ETHICS

Historically, many moral philosophers have been skeptical about the relevance of the sciences to ethics. In the spirit of Hume (1978/1740, p. 469), it is argued that because scientific inquiry is primarily descriptive or factual and ethical inquiry is primarily prescriptive or normative, there is an unbridgeable 'logical gap' between the *is* of the sciences and the *ought* of ethics. However much the sciences may tell us about human beings, how the beings in question should comport themselves remains – to borrow an influential formulation of Moore's – an open question. For example, contemporary life science has discovered much about the development and functioning of the fetus, yet this has, in the eyes of many, done little to resolve the abortion controversy.

But there have always been those who viewed this gap between *is* and *ought* as illusory, and much recent philosophical work suggests that the barrier between description and prescription may not be as impermeable as followers of Hume and Moore suppose (Railton, 1995). Certainly moral philosophers often offer arguments – such as those

regarding the nature of moral motivation and cognition – whose presuppositions appear amenable to empirical investigation. In so far as this is the case, cognitive science bears a strong prima facie relevance to ethics. Accordingly, we will not attempt to adjudicate the continuing philosophical controversy regarding the relation of scientific to ethical inquiry, but will instead identify interfaces between cognitive science and ethics that merit further empirical study and philosophical discussion.

APPLIED ETHICS

Many issues in applied ethics raise important epistemological questions – questions about what people can and cannot be reasonably expected to know. Other issues raise closely related questions about what people do and do not understand. The cognitive sciences have produced empirical findings relevant to such questions, and these findings often play a central role in debates in applied ethics. In this section we will give two brief illustrations.

One of the first domains in which experimental psychology was applied to an important ethical issue was the debate over reliance on eyewitness testimony in legal proceedings. Traditionally, the identification of a defendant by an eyewitness to a crime has played a central role in Anglo-Saxon law. However, almost 100 years ago, Munsterberg conducted a number of dramatic, though poorly controlled, experiments in which many eyewitnesses to a staged 'crime' gave seriously inaccurate reports of what they had seen. In a book that generated considerable controversy in the legal profession and beyond, Munsterberg (1908, p. 194) argued that it is 'astonishing that the work of justice is ever carried out in the courts without ever consulting the psychologist and asking him for all the aid which the modern study of suggestion can offer'.

More recent work, using much improved methods, has demonstrated that people's memories of events can be affected in dramatic and disquieting ways by information (or *mis*information) that they encounter after the event. In one experiment, conducted in train stations and other naturalistic settings, Loftus (1979) and her students staged a 'robbery' in which a male confederate pulled an object from a bag that two female students had temporarily left unattended and stuffed it under his coat. A moment later, one of the women noticed that her bag had been tampered with and shouted: 'Oh my God, my tape recorder is missing.' She went on to lament that her boss had loaned it to her and that it was very expensive. Bystanders, most of whom were quite cooperative, were asked for their phone numbers in case an account of the incident was needed for insurance purposes. A week later, an 'insurance agent' called the eyewitnesses and asked about details of the theft. Among the questions asked was 'Did you see the tape recorder?' More than half of the eyewitnesses remembered having seen it, and nearly all of these could describe it in detail – this despite the fact that *there was no tape recorder*! On the basis of this and other experiments, Loftus concludes that even casual mention of objects that were not present, or of events that did not take place (for example, in the course of police questioning), can significantly increase the likelihood that the objects or events will be incorporated into people's memories. Other studies have shown that people are often mistaken when they identify someone as a participant in an event that they witnessed, and that both white and black subjects are considerably less accurate at identifying people of the other race (Loftus, 1979; Lloyd-Bostock and Clifford, 1983; Wells and Loftus, 1984).

Findings like these, along with a number of highly visible cases in which people have been convicted of crimes they did not commit on the basis of eyewitness testimony, have led one leading legal scholar to recommend that judges should be required 'to direct the jury that it is not safe to convict upon eyewitness evidence unless the circumstances of the identification are exceptional or the eyewitness evidence is supported by substantial evidence of another sort' (Devlin, 1976). Though this recommendation has not been widely accepted, the recent use of DNA testing to reverse wrongful convictions that were often at least partially based on unreliable eyewitness testimony has once again focused attention on the need for reform. Although some philosophers have questioned the wisdom of relying on empirical studies in this context (Coady, 1992), the research raises an important and troubling question: when, if ever, is it ethically defensible to allow legal decisions to be influenced by such an empirically suspect source of evidence?

Another area of research in cognitive science that has important implications for the legal system, and also for issues about medical and academic decision-making, focuses on the reliability of predictions about an individual's future behavior based on interviews and 'clinical' judgments. In some American states, interviews with parole board members play a central role in deciding whether an offender will be granted a parole. In one state system, for example, interviewers are

asked to make a prediction of the risk that the offender will commit another crime. A study by Carroll *et al*. (1988) showed that psychologists who had never spoken with the prisoners could make much more accurate predictions by using purely statistical 'actuarial' methods and basing the prediction entirely on a few background factors, such as number of previous convictions. This finding is one of well over 100 studies comparing clinical to actuarial methods for predicting a wide range of behavioral phenomena, from academic success to the prognosis for patients admitted to a mental hospital. In just about all of these studies, actuarial methods are at least as good as clinical prediction, and often they are much better. These findings have led to considerable discussion about *why* the human cognitive system does such a bad job on these tasks, despite the often powerful conviction on the part of 'expert' judges that their professional experience enables them to make better predictions (Nisbett and Ross, 1980; Dawes, 1994). Though we do not currently have a good understanding of the mechanisms that give rise to clinical predictions, their unreliability raises serious questions about the moral justification for using such predictions in making important decisions.

NORMATIVE ETHICS

Making defensible moral judgments requires a sensitivity to ethically salient features of the environment: one is not likely to act compassionately if one is oblivious to suffering, nor is one likely to act justly in complex circumstances without some awareness of what justice requires. It appears as though individuals vary on this dimension; just as some are especially attuned to the dictates of etiquette or fashion, others seem especially attuned to the demands of morality.

This sort of observation has been prominent in philosophical thought regarding the virtues, or excellences of character. According to Aristotelians such as McDowell (1979), virtue involves a 'reliable sensitivity' to ethically salient considerations. The requisite sensitivity must be both *reliable* and *robust*; while even morally mediocre people may manifest ethical sensitivity in particular instances, the virtuous person manifests appropriate ethical sensitivity in all circumstances where she can reasonably be expected to do so, regardless of whether this exercise is difficult or easy.

However appealing the foregoing view may be, experimental social psychology suggests that ethical sensitivity is typically far from robust. A large body of research indicates that people's moral judgments and behavior are extraordinarily sensitive to features of the situation in which they are embedded, including many features that most people would agree are morally irrelevant. (See Doris (2002) for details and references.) Experimental studies of helping behavior provide some particularly disconcerting examples. Finding a bit of change or hearing the noise from a lawn mower can make the difference between helping and not – results that likely would have astounded Aristotle. In a classic demonstration by Darley and Batson, subjects performed tasks at two separate sites. The behavior of interest occurred when subjects walked from one site to the other, passing the experimenters' assistant slumped in a doorway, apparently in some distress. Before leaving the first site, subjects were told either that they were running late ('high hurry'), were right on time ('medium hurry'), or were a little early ('low hurry'); thus members of each group experienced a different degree of time pressure while travelling from one site to the next. Helping behavior varied markedly with degree of hurry: 63 per cent offered help in 'low hurry', 45 per cent offered help in 'medium hurry', and only 10 per cent offered help in 'high hurry'. Evidently some subjects wanted to help but reluctantly ceded to the demands of punctuality, while for others (more germane here), time pressures muted their sensitivity to ethically salient stimuli – in some instances subjects stepped unconcernedly over the stricken 'victim', apparently without registering the scene. There is no reason to think that the subjects in these studies were aberrant in their sensitivity to situational factors; it is quite reasonable to suppose that most people exhibit very substantial situational variability in their moral judgments and behavior, whatever we are antecedently inclined to say about their character.

These empirical issues do not by themselves show that a conception of virtue as involving a robust and reliable ethical sensitivity is of no use to ethics. Perhaps virtue is to be understood, as some philosophers have suggested, in terms of a rarely attainable ideal that serves to focus people's ethical aspirations (Blum, 1994, pp. 94–6), or perhaps the empirical findings do not unsettle a suitably sophisticated moral psychology of virtue. Or perhaps, as others have argued, results such as those just recounted suggest major revisions for philosophical thinking on virtue and character (Doris, 1998, 2002; Harman, 1999). It remains to be seen how compelling these alternatives are, but it is obvious that virtue ethics is in a fertile tension with empirical work.

META-ETHICS

Meta-ethics is concerned, in part, with the nature of moral concepts, the structure of moral reasoning, and the nature of moral motivation, and there are many ways in which the findings and theories in cognitive science are relevant to these concerns. In this article we have space for only a single example.

As we noted earlier, the Kantian tradition in moral philosophy emphasizes the role of reason in ethics, and many Kantians have invoked reason to explain why people should be motivated to act morally. The explanation they propose is that it would be irrational not to be motivated to do what one morally ought to do (or what one believes one morally ought to do), in much the same way that it would be irrational not to believe the conclusion of a sound argument (Nagel, 1970, p. 3). One famous obstacle to proposals of this sort is suggested by Hume's (1777/1966, pp. 282ff) hypothetical example of the 'sensible knave' – a person who recognizes that it is morally wrong to be dishonest in a situation where he would benefit from dishonesty, but who is not moved at all by this judgment. More recent writers have urged that Hume's sensible knave is more than a philosophical fiction since there actually are psychopaths who know the difference between right and wrong but simply have no motivation to *do* what is right.

Those who wish to defend the link between reason and moral motivation have adopted two quite different strategies, both of which appear to make substantive empirical assumptions of the sort that cognitive science might test (Nichols, 2002). The first strategy relies on a claim about our ordinary concept of moral judgment: the concept of moral judgment entails that 'agents who make moral judgments are motivated accordingly' (Smith, 1994, p. 66). Philosophers who adopt this strategy recognize that psychopaths may *say* that something is 'morally required' or 'morally wrong' and that they may be quite sincere. But if psychopaths are not motivated in the appropriate way, their words do not mean what non-psychopaths mean by these words and the concepts they express with these words are not the ordinary moral concepts that non-psychopaths use. This strategy only works, of course, if it is true that our ordinary moral concepts *require* that people to whom the concepts apply have the appropriate sort of motivation. And, for two quite different reasons, it is far from clear that this is the case. First, there is considerable disagreement in cognitive science about whether and how concepts are structured, and about how we are to determine when something is built into or entailed by a concept (Margolis and Laurence, 1999). Second, there has been almost no serious empirical work aimed at uncovering the structure of ordinary people's moral concepts. There has, however, been work on ordinary people's epistemic concepts, and that work suggests that people in different ethnic and socioeconomic groups use significantly different epistemic concepts (Weinberg *et al*., forthcoming). If the same is true for moral concepts, it will pose a serious challenge for the conceptual approach to the problem posed by sensible knaves and psychopaths.

A second strategy for defending the connection between reason and moral motivation takes the link to be empirical rather than conceptual. Those who adopt this approach maintain that it is an empirical fact that the faculty of reason, when functioning normally, generates motivation to do what one judges one ought to do, just as it is an empirical fact that the faculty of reason generates an inclination to believe the conclusions of arguments one judges to be sound. If psychopaths fail to have the appropriate moral motivation, it is argued, then there must be something wrong with their reasoning faculty. But is this true? The answer is far from clear.

Blair (1995) has shown that psychopaths *do* exhibit surprising deficits on a cluster of cognitive tasks that have been used frequently by psychologists who study moral development. In these tasks subjects are presented with descriptions of various transgressions such as a child hitting another child, or a child leaving the classroom without the teacher's permission. In the moral development literature, transgressions of the first sort are labeled 'moral' while those of the second sort are labeled 'conventional'. From quite early on in childhood, normal children distinguish moral from conventional transgressions on a number of dimensions: they view moral transgressions as more serious, they explain why the acts are wrong by appeal to different factors (harm and fairness for moral transgressions, social acceptability for conventional transgressions), and they understand conventional transgressions, unlike moral transgressions, to be dependent on authority. If the teacher says there is no rule about leaving without permission, children think it is OK to leave without permission. But if the teacher says there is no rule against hitting other children, they do not think that hitting is acceptable. What Blair found was that incarcerated psychopaths do not draw the moral/conventional distinction. Though normal children and normal adults (including a control group of non-psychopath prisoners) have no

trouble classifying new cases in one category or the other, psychopaths fail to do so.

These results might well give some support to the hypothesis that psychopaths have a reasoning deficit, and thus that they do not pose a problem for those who maintain that a properly functioning reasoning faculty always generates some motivation to do what one believes one ought to do. But, as Nichols (2002) has pointed out, the issue cannot be so easily resolved, because psychopaths have also been shown to have affective responses that are quite different from those of normal subjects. When shown distressing stimuli (e.g. slides of people with dreadful injuries) and threatening stimuli (e.g. slides of an angry man wielding a weapon), normal subjects exhibit much the same suite of physiological responses. Psychopaths, by contrast, exhibit normal physiological responses to threatening stimuli, but abnormally low physiological responses to distressing stimuli (Blair *et al*., 1997). Thus, Nichols argues, it may well be that what is wrong with psychopaths is not that their reasoning system is abnormal, but that their affect system is abnormal, and that it is their affective abnormalities that are responsible for their inability to draw the moral/conventional distinction. If that is the case, then Hume's challenge continues to pose a major problem for those who think there is a link between reason and moral motivation. Resolving this issue will require conceptual clarification on how to draw the boundary between reason and affect, and on what counts as an *abnormality* in each of these domains. It will also require much more empirical work aimed at understanding exactly how psychopaths and non-psychopaths differ. Clearly, this important issue in meta-ethics cannot be addressed responsibly without taking account of the increasingly rich body of empirical findings generated by the cognitive sciences.

THE METHODS OF ETHICS

'Intuition pumps' or 'thought experiments' are among the most prominent elements of philosophical method. The technique is to present a hypothetical example and attempt to elicit some philosophically telling intuition; if the 'experiment' is successful, it may be concluded that competing theories must account for the resulting intuition. In such cases, those forwarding the thought experiment insist that for a theory to be viable, it must be consistent with the intuition elicited by the hypothetical example; if the theory is inconsistent with the intuition, the hypothetical is supposed to be a counterexample to the theory. This is a central rhetorical strategy in ethics, as evidenced by debates over Utilitarianism. Consider, for example, an influential intuition pump proposed by Williams (1973, pp. 97–100), who asks us to imagine the following circumstances: George, a new PhD in chemistry in dire need of a job to support his family, is offered a chance at a job researching chemical and biological weapons (CBW). George is strongly opposed to CBW research, but if he does not take the job, it is likely to go to a man with rather alarming enthusiasm for CBW research. Williams maintains that Utilitarians are bound to say that George should take the job, presumably because doing so will both help his family and hinder CBW research, thus maximizing total available welfare. But Williams argues that 'many of us would certainly wonder whether … that could possibly be the right answer at all'. Williams apparently thinks that in this case people will conclude that the Utilitarian prescription does not properly account for the value of George's 'integrity', a commitment to principle that would be undermined if he took the job.

In response, the Utilitarian may deny that ethical theory should be constrained by intuitions. But this is a rather unappealing expedient; a theory that makes no reference to intuitions risks becoming divorced from the experience of ethical life. On the other hand, it is also perilous to deploy intuitions as a constraint on ethical reflection. Given the ubiquity of moral controversy, cultures and individuals may vary dramatically in their moral intuitions; how is it to be decided which intuitions, and whose should serve as constraints? For example, Haidt *et al*. (1993) found evidence suggesting that Americans' ethical intuitions varied with their socioeconomic status. Returning to George the chemist, is Williams right to suppose that 'many of us' value integrity over food on the table, or is this intuition limited to the 'many of us with sufficiently high socioeconomic status'? A further look at cognitive science presses this question still harder: experimental work suggests that even within a single individual, moral intuitions may vary in quite arbitrary ways.

In an important study, Tversky and Kahneman (1981) presented a group of subjects with the following problem:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

On the first version of the problem most subjects thought that Program A should be adopted. But on the second version most chose Program D, despite the fact that the outcome described in A is identical to the one described in C. The disconcerting implication of this study is that the moral decisions we make are strongly influenced by the manner in which the options are described or *framed*. The general worry is that ethical intuitions may be determined by features of thought experiments that are quite ethically irrelevant. To return once more to George the chemist, are responses to the thought experiment determined by reflection on the ethical substance of the case, or by ethically irrelevant features of the richly textured example? Until we can say with confidence how a given ethical intuition is generated, we should be hesitant to rely on it in argument.

This suggests a dilemma for philosophical ethics: it must either eschew appeal to thought experiments that have not been evaluated empirically, or eschew appeal to intuitions altogether. As we've already said, the abolition of intuitions is a problematic solution: it threatens to distance philosophical ethics from the experience of ethical life, and thereby alienate the study of ethics from the concerns of actual ethical agents (see Williams, 1985, pp. 93–119, esp. 116–119). But continued reliance on intuitions in the face of the difficulties just surveyed presents its own difficulty: intuitions, and the thought experiments generating them, must be subjected to systematic empirical scrutiny before they can be appealed to in adjudicating theoretical disputes. In other words, if practitioners of philosophical ethics wish to rely on intuitions, they must work to develop a cognitive science of ethics. If we are right, the implications of this are radical: philosophers must either depart from the traditional methods of ethics by abandoning the use of intuitions, or depart from the traditional methods of ethics by pursuing experimental work in cognitive science.

References

Blair R (1995) A cognitive developmental approach to morality: investigating the psychopath. *Cognition* **57**: 1–29.

Blair R, Jones L, Clark F and Smith M (1997) The psychopathic individual: a lack of responsiveness to distress cues? *Psychophysiology* **34**: 192–198.

Blum LA (1994) *Moral Perception and Particularity*. Cambridge, UK: Cambridge University Press.

Carroll J, Winer R, Coates D, Galegher J and Alibrio J (1988) Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review* **17**: 199–228.

Coady C (1992) *Testimony: A Philosophical Study*. Oxford, UK: Clarendon Press.

Darwall S (1998) *Philosophical Ethics*. Boulder, CO: Westview.

Dawes R (1994) *House of Cards*. New York: The Free Press.

Devlin P (Chair) (1976) Report to the Secretary of State for the Home Department of the Departmental Committee on Evidence and Identification in Criminal Cases. London, UK: Her Majesty's Stationery Office.

Doris JM (1998) Persons, situations, and virtue ethics. *Noûs* **32**: 504–530.

Doris JM (2002) *Lack of Character: Personality and Moral Behavior*. Cambridge, UK and New York, NY: Cambridge University Press.

Haidt J, Koller S and Dias M (1993) Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* **65**: 613–628.

Harman G (1999) Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* **99**: 315–331.

Hume D (1777/1966) *Enquiry Concerning the Principles of Morals*. New York, NY: Open Court (1966).

Hume D (1740/1978) *A Treatise of Human Nature, 2nd edn*. Oxford, UK: Oxford University Press.

Lloyd-Bostock S and Clifford B (1983) *Evaluating Witness Evidence*. New York, NY: Wiley.

Loftus E (1979) *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Margolis E and Laurence S (1999) *Concepts*. Cambridge, MA: MIT Press.

McDowell J (1979). Virtue and reason. *Monist* **62**: 330–350.

Munsterberg H (1908) *On the Witness Stand*. New York, NY: Doubleday, Page.

Nagel T (1970) *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Nichols S (2002) Is it irrational to be amoral? How psychopaths threaten moral rationalism. *Monist* **85**: 285–304.

Nisbett R and Ross L (1980) *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall.

Railton P (1995) Made in the shade: moral compatibilism and the aims of moral theory. *Canadian Journal of Philosophy Supplementary Volume* **21**: 79–106.

Smith M (1994) *The Moral Problem*. Oxford, UK: Blackwell.

Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* **211**: 453–463.
Weinberg J, Nichols S and Stich S (forthcoming) Normativity and epistemic intuitions. *Philosophical Topics*.
Wells G and Loftus E (eds) (1984) *Eyewitness Testimony: Psychological Perspectives*. Cambridge, UK: Cambridge University Press.
Williams BAO (1973) A critique of utilitarianism In: Smart JJC and Williams BAO, *Utilitarianism: For and Against*. Cambridge, UK: Cambridge University Press.
Williams BAO (1985) *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Further Reading

Alexander R (1987) *The Biology of Moral Systems*. New York, NY: Aldine de Gruyter.
Flanagan O (1991) *Varieties of Moral Personality: Ethics and Psychological Realism*. Cambridge, MA: Harvard University Press.
Gilligan C (1982) *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.
Goldman A (1993) Ethics and cognitive science. *Ethics* **103**: 337–360.
Harman G (2000) *Explaining Value and Other Essays in Moral Philosophy*. Oxford, UK: Oxford University Press.
Kagan J and Lamb S (eds) (1987) *The Emergence of Morality in Young Children*. Chicago, IL: University of Chicago Press.
Kohlberg L (1981) *Essays in Moral Development*, vol. 1: *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. San Francisco, CA: Harper & Row.
Sober E and Wilson DS (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Stich S (1993) Moral philosophy and mental representation. In: Hechter M, Nadel L and Michod RE (eds) *The Origin of Values*, pp. 215–228. New York, NY: Aldine de Gruyter.
Wright R (1994) *The Moral Animal: The New Science of Evolutionary Psychology*. New York, NY: Pantheon Books.

Explanatory Gap

Intermediate article

Joseph Levine, Ohio State University, Columbus, Ohio, USA

CONTENTS

The explanatory gap concerns the problem of explaining the qualitative character of conscious experience by reference to underlying physical and computational processes.

WHAT IS THE EXPLANATORY GAP?

The explanatory gap is a fundamental aspect of the mind–body problem, the problem of determining the relation between physical and mental phenomena.

We commonly explain the behavior of the people around us by appealing to their beliefs and desires, along with their sensory experiences. In our own case we know that there is something particular that it is like to smell a rose, taste coffee, or see a sunset. We have fairly direct knowledge of what we believe, desire, hope and fear. It is almost unimaginable to abandon this conception of ourselves and our fellow human beings (and many higher animals) as possessing mental lives that play a central role in determining behavior.

On the other hand, we also know that we are physical creatures. We have bodies that are subject to the laws of physics, and, more specifically, to the laws of biology. So the question naturally arises: how do these two sets of phenomena relate to each other? Are mental states and events just identical to certain physical states and events – in particular, to states and processes in the brain? Or are mental states and events somehow distinct from, metaphysically independent of, though perhaps causally connected to, processes in the brain?

Materialism is the doctrine that mental phenomena are a species of physical phenomena. Whether the precise relation between mental properties and neurological properties is identity, or instead something like constitution or realization, is a further question. What unites all versions of materialism is the idea that mental phenomena are part of the physical, natural order. This is where the problem of the explanatory gap enters. If we accept this basic tenet of materialism, it seems reasonable that we ought to be able to explain why it is that certain neurological states constitute the mental states they do. We ought to be able to say why having these neurons fire the way they do amounts to believing that $2+2=4$, and why having these other neurons fire the way they do amounts to having a visual experience of red. The problem of the explanatory gap is that, at least for states like visual experiences of red, there seems to be a deep problem about how to explain their character by reference to the physical states that allegedly constitute them.

Mental states fall roughly into two groups: cognitive states and experiential states. Cognitive states involve representing the world in some way and taking an attitude towards the situation represented. Experiential states are those for which it makes sense to talk about 'what it is like' to have them: for example, visual sensations and pains. These are states that have a 'qualitative' character; and their qualitative properties – such as the way a toothache feels, or a red rose looks – are called 'qualia'. While many philosophers agree that progress has been made towards explaining how physical phenomena could represent, there is much less agreement that we have any clear idea how physical phenomena could amount to conscious experience. What is it about certain neurological processes that explains why it is like what it is like to undergo them? Here the explanatory gap seems to many to be as wide as ever.

HISTORY

The term 'explanatory gap' was introduced by Levine (1983), but the idea is older. John Locke (1690/1975) distinguished ideas of primary qualities from ideas of secondary qualities. Primary qualities are essentially the basic physical properties of matter, which Locke thought to be size, shape, and bulk. Secondary qualities are properties like color and taste. On Locke's view, when our

sensory systems are affected by certain physical stimuli, our minds react by forming certain ideas, such as experiences of color and taste, and these are our ideas of secondary qualities. According to Locke, there is no intelligible relation between the configuration of primary qualities in the perceived objects and the ideas of secondary qualities that interaction with them gives rise to. Ideas of red could just as easily have been associated with green objects as with red objects. God just happened to pick one scheme of association rather than the other. Thus, on Locke's view, the reason why there is an explanatory gap between the description of the physical interaction taking place during perception and the description of the experience is that there really is no explanation of the latter in terms of the former. God arbitrarily decreed that these experiences should be the result of those particular physical interactions.

The other important historical antecedent is Leibniz (1840). Leibniz asks us to imagine a brain expanded to the size of a factory. We could enter and look around at all the physical interactions among the various parts of the brain. Where, Leibniz asks, would we see sensations of blue, pains, or thoughts? His point is that we cannot see how physical interactions of this sort could amount to experience.

Among modern philosophers, the most famous expression of the problem of the explanatory gap is by Nagel (1974). Nagel argues that physicalism can never make sense of the relation between physical processes and conscious experience because we can only understand physical phenomena from an objective point of view, whereas our conception of conscious experience is essentially subjective. For example, while we could understand everything there is to know about the workings of the bat's perceptual system of echo location, we could not thereby know what it is like to perceive as a bat does. But if we could really explain what it was like on the basis of the relevant neurology and information processing, we would be able to know what it was like. Thus our ignorance in this case is a direct reflection of the explanatory gap.

ARGUMENTS FOR THE EXPLANATORY GAP

There are three forms of argument that support the existence of the explanatory gap: the conceivability argument, the knowledge argument, and an argument connected with the problem of other minds.

The Conceivability Argument

The conceivability argument is a well-known argument against materialism, due to Descartes but revived recently by Kripke (1980), and even more recently by Chalmers (1996). We begin with the premise that it is at least conceivable that a creature could share all of my physical traits (or, if the argument is employed against functionalism – a particular version of materialism – all of my functional traits) and yet have different qualia or no qualia at all. That is, we can conceive of a creature that is physically like me but for whom there is nothing that it is like to be that creature, or for whom experiences are very different from what they are like for me. We now add the premise that if such a situation is conceivable – we can find no contradiction or incoherence in its supposition – then it is genuinely possible. But if it is possible to have A without B, then it follows that A and B are not identical. Therefore, qualia are not identical to physical (or functional) properties.

One response to the conceivability argument is to deny the second premise: that conceivability entails possibility. But for the advocate of the explanatory gap this doesn't matter. Even if we grant that, merely from the fact that we can conceive of a creature physically like me, but without consciousness, it does not follow that such a creature is indeed possible, still, the fact that such a creature is even conceivable demonstrates the explanatory limits of materialist theory. For if we really could explain why someone had the qualia they had by appeal to their physical (or functional) structure, then, it is argued, the creature in question would not be conceivable. We would see why having this physical structure necessitated having the conscious experiences it has. The fact that we find no problem conceiving a physical duplicate for whom there is no experience, or very different experiences, is a sign of the explanatory gap between the physical and the mental.

The point is best made by contrasting the mental–physical case with other instances of theoretical reductions. We know that water is composed of H_2O molecules. Certainly it is conceivable that this could have not been the case: the world could have been as Aristotle thought it to be, with water a basic element. However, it does not seem conceivable that there are H_2O molecules obeying the laws of chemistry and physics that apply to our world, and yet not manifesting the various superficial, high-level properties by which we normally identify water. In other words, it does not seem conceivable that, with all of our chemistry and physics in

place, masses of such molecules should not freeze and boil at the appropriate temperatures, or not allow light to pass through, or not supply an essential ingredient to plants and animals. We can see how the laws of chemistry and physics necessitate these high-level phenomena.

However, when we turn to the relation between qualia and the neurological states that underlie them, the situation seems different. It is not difficult to imagine how, from a sufficiently detailed description of the underlying neurology, one might be able to infer the structure of the functional, or information processing, relations among mental states. At least such a derivation seems possible in principle. But no similar derivation seems possible from neurological phenomena to conscious, qualitative phenomena. For any proposed neurological correlate of a conscious experience, such as a visual experience of red, it always seems at least conceivable that the neurological state might obtain without the experience: either because we can conceive of a different sort of experience going with that state, such as an experience of green instead, or because we can conceive of there being nothing that it is like to occupy that state.

The Knowledge Argument

The 'knowledge' argument is due to Jackson (1982). He asks us to imagine Mary, a super-scientist of the future who knows everything there is to know about the physical and functional structure of color vision, but who, because she is imprisoned in a black and white environment, has never seen color herself. When she emerges from her black and white world for the first time and sees a ripe tomato, it seems plausible that she will learn something new: namely, what it is like to see red. But if she already knew everything there was to know about seeing red, why would there be anything left for her to learn in this situation? Thus, all the physical and functional facts together fail to encompass all the facts, and therefore materialism is false.

As with the conceivability argument, it is possible to disagree with Jackson's conclusion that materialism is false, but still use the argument to establish the explanatory gap. A standard materialist reply to Jackson is as follows. The fact that Mary will seem to learn something new upon seeing a ripe tomato for the first time does not show that the facts about conscious experience are distinct from the facts about what is going on in her brain. It could be that she is not learning new facts, but rather learning to represent them in a new way, using new concepts.

Nevertheless, one can reply that if the physical and functional theory really explained the qualitative character of visual states, Mary should be able to tell in advance what the experience would be like. Yet it seems plausible that if, instead of a ripe tomato, she were presented with a flat rectangular patch of red immediately after leaving her black and white prison, she would not know what color she was looking at. In other words, all the physical and functional descriptions of color experience do not allow her to predict its qualitative character – and this is precisely the situation we would expect if they do not explain it.

The Problem of Other Minds

Traditionally, the problem of other minds concerns our basis for attributing mental states to anyone but ourselves: how do I know that others have real experiences and are not just automata? We will not address this question here, but a related problem. First, let us distinguish between our functional structure and our physical structure. The former has to do with the causal pattern of interacting states that mediate between environmental stimuli and behavior. In principle, this same pattern could be realized, or implemented, in creatures with significantly different physical structures, provided that their internal physical states followed the requisite pattern of interactions. Functional structure admits of various levels of description. For instance, it is possible for two creatures to both satisfy the same generalizations concerning their ability to discriminate among colors, and to generally agree on their judgments of the relative similarity of colors, but to differ in the details of the algorithms for computing color information that are implemented by their visual systems. So at the more superficial level they would count as functionally identical with respect to color vision, but at a deeper level they would not.

Thus, creatures with a prima facie claim to mental lives can differ from us in two ways: with respect to (at least) deeper-level functional structure, and with respect to physical structure. Here the problem of other minds arises. Just how much, and in which ways, can a creature differ from us and yet still provide us with sufficient reason to attribute conscious experiences to it (or, experiences qualitatively like ours)? Notice that the problem is not one of ignorance, or at least not obviously. Let us imagine that we know all there is to know about the ways in which our candidate creature is functionally and physically like us, and the ways in which the creature differs from us. The

question remains: does this creature have conscious experiences, and, if so, what are they like? It is not clear how we could decide the matter.

However, if we had an adequate explanation of what it is about either our physical or functional structure that is responsible for our experience being like what it is for us (or being like anything at all), then, it seems, we would have a sound basis for making this decision with respect to other types of creature. We could just look for those features, at whatever level of description, that explain the nature of our experience. The fact that we do not know how to solve this problem about alien creatures demonstrates that we do not really know what explains experience in our own case. We do not know what level of functional or physical detail makes the difference.

ARGUMENTS AGAINST THE EXPLANATORY GAP

There are various ways for a materialist to respond to the explanatory gap. The most straightforward way is simply to deny it exists. Denial of the explanatory gap can take three forms: appeal to future science, conceptual reductionism, and eliminativism.

Appeal to Future Science

One might argue as follows. At present we do not have an adequate understanding of how conscious experience arises from the neurological processes of the brain, but this is no different from many other instances of current scientific ignorance. We also do not have an adequate understanding of memory, problem-solving, or many other psychological capacities that do not involve qualia. So those who attribute special urgency to the explanatory gap are just exhibiting unwarranted impatience with the progress of science.

The problem with this line of response is that it is not easy to imagine how future research could bridge this particular gap. In the case of psychological capacities like memory or problem-solving, while it is true that we do not know much about how these tasks are accomplished, still it is clear that the tasks themselves could be understood in terms of the implementation of a causal role. Now, either conscious experience can be conceptualized in similar terms or it cannot. If it can, then it is that analysis that is doing the principal work of bridging the gap (see below). If it cannot, then discovering more details about how the brain implements the various causal roles involved in information processing is not going to solve the problem. But what else can neuroscientific investigation yield except a theory of the implementation of the relevant causal roles?

Conceptual Reductionism

A fairly direct way to argue against the explanatory gap is to adopt a conceptual reductionist position (White, 1986). On this view, our idea of what it is to experience a certain quale is explicated in terms of a state's causal role. For example, we might think that to have a reddish experience is to be in a state that is normally caused by seeing red things, is judged similar to experiences of other red things, and normally leads to verbal reports like 'that's red'. If this is what we have in mind by an experience of red, then there is no problem explaining its qualitative character by reference to neurological processes. As long as we can see how the relevant process plays the requisite role, we have the explanation we are looking for. This position rules out the conceivability of a creature that satisfies the physical but not the mental descriptions true of us; it entails that Mary does not learn anything new when she emerges from the black and white room; and it gives us the principle we need to determine when physically diverse creatures share our experiences.

Eliminativism

To the extent that one finds the arguments for the explanatory gap convincing, one will take them as evidence that the 'causal role' analysis of our concept of conscious experience is inadequate. Another way to deny the explanatory gap, while admitting the inadequacy of a causal role analysis, is to endorse eliminativism (Dennett, 1991). According to the eliminativist it is true that we cannot analyze our idea of qualitative experience in causal role terms, but that is because our pretheoretic conception of conscious experience is confused and does not correspond to any real phenomenon. There is no explanatory gap because the phenomenon allegedly left unexplained by materialist theories – conscious experience – does not really exist.

Put that way, eliminativism seems absurd: feelings of pain, visual experiences of color, itches and tickles somehow do not really exist; they are mere illusions. However, there are two considerations that help to relieve the apparent absurdity. First, eliminativism seems more plausible after one has surveyed all the other attempts to deal with the explanatory gap and other problems posed by

conscious experience. If the advocate of the explanatory gap is right that we have no clear idea even where to look for a solution, it seems appropriate to consider the hypothesis that the reason we cannot find a solution is that there is something seriously wrong with the question.

Second, eliminativism is often associated with a positive account of why it should seem to us that we do have the sort of experience we think we have (Dennett, 1991; Rey, 1995). Something is indeed going on in us when we apparently experience qualia; but it is not what we think. To the extent that such positive accounts successfully explain our tendency to describe our experience in the way we do, the apparent absurdity of denying reality to that description is diminished. However, while some of the adherents of these positive accounts conjoin them with an eliminativist position, most do not.

ACCOUNTS OF THE EXPLANATORY GAP

Non-deflationary Accounts

The most direct account of the explanatory gap is the one that accepts it at face value: we cannot explain qualitative experience in terms of physical or functional mechanisms because conscious experience is not a physical phenomenon. According to Chalmers (1996), there is a basic law of nature – basic in the same way that the fundamental laws of physics are basic – that connects certain functional states with certain conscious states. Since this is a basic law, there is no way to explain the correlation in more fundamental terms. Qualia appear brute and inexplicable because they are.

Another approach is what may be called the 'mysterian' view. McGinn (1991) provides perhaps the clearest account of the mysterian position. On McGinn's view, conscious experience is in fact a material phenomenon, so he departs from dualists like Chalmers. However, there is an explanatory gap because we humans have limited conceptual resources: we cannot form the concepts needed to appreciate the connection between underlying physical processes and conscious experience. Nagel (1974) also thinks that the problem can only be solved with concepts we do not possess, though he is less pessimistic about our ability to develop such concepts.

McGinn and Nagel also offer more specific accounts of where the conceptual gaps lie. For McGinn, the problem largely has to do with our restriction to spatial concepts in scientific explanations, and the unsuitability of a spatial framework for our concepts of qualia. For Nagel, the problem has to do with the objective point of view adopted in scientific explanations and its inability to capture phenomena, like conscious experience, which can only be fully appreciated when apprehended as well from a subjective, first-person point of view.

Deflationary Accounts

A deflationary account of the explanatory gap is one that accepts its existence but does not see it as either threatening to materialism or demonstrating any deep mystery. There is a family of such accounts (e.g. Loar, 1997), all of which share two basic components. First, they appeal to the distinction between concepts and properties. Second, they appeal to some distinctive feature of the way we gain cognitive access to our own conscious experiences. The basic idea is that the explanatory gap is real, but merely reflects certain (explicable) features of our cognitive organization; so that on the proper materialist theory the explanatory gap is to be expected.

The idea behind the concept–property distinction is that properties are objective, mind-independent features of the world, and concepts are mental representations, or modes of presentation, of properties. It is possible to have more than one concept for a given property. For instance, the concept expressed by 'temperature' and the concept expressed by 'mean molecular kinetic energy' are distinct but they pick out the same objective property of objects. Similarly, it may be that our normal concepts for experiences and the scientific concepts derived from psychology and neuroscience, though distinct, nevertheless pick out the same properties.

Merely making the concept–property distinction is not sufficient to deflate the explanatory gap. In other cases of scientific reductions we deal with distinct concepts and yet do not face an explanatory gap. This is where the second component of the deflationary account is needed. The idea is that the concepts involved in our everyday thoughts about our experiences stand in a very special computational relation to those experiences, a relation that makes it difficult to combine them with our concepts for objects and properties other than our mental states. It is this incommensurability between our first-person concepts for our experiences and other concepts that makes the relation between physical and qualitative properties unintelligible to us. Accounts differ as to precisely how first-person

concepts of experience are distinguished from other concepts.

THE EXPLANATORY GAP AND COGNITIVE SCIENCE

For the most part, research in cognitive science can proceed without concern for the explanatory gap. We can investigate the computational basis of language acquisition and processing, visual processing, problem-solving, and the like – indeed, the entire domain of information processing psychology – without encountering the question of what it is like to occupy these various information processing states. Most of the states involved are not conscious anyway, since most of the processing underlying perception and thought occurs below the level of conscious awareness. Furthermore, even those states and processes that are conscious can be investigated to a great extent without regard to their qualitative character.

However, as long as we have not bridged the explanatory gap, or deflated it, cognitive science will face a fundamental limitation on its explanatory reach. There seems to be more to us than how we process information: there is what it is like to be us. Whether theories in cognitive science are likely to successfully bridge the gap in the near future is a matter of controversy.

References

Chalmers D (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Leibniz W (1840) Monadology. In: Erdmann JE (ed.) *Leibniz: Opera Philosophica*. Berlin: Eichler.

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Loar B (1997) Phenomenal states. In: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

Locke J and Nidditch P (ed.) (1690/1975) *An Essay Concerning Human Understanding*. Oxford: Clarendon Press.

McGinn C (1991) *The Problem of Consciousness*. Oxford: Blackwell.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **82**: 435–450.

Rey G (1995) Toward a projectivist account of conscious experience. In: Metzinger T (ed.) *Conscious Experience*. Paderborn, Germany: Ferdinand Schöningh/Imprint Academic.

White S (1986) Curse of the qualia. *Synthese* **68**: 333–368.

Further Reading

Block N (1980) Troubles with functionalism. In: Block N (ed.) *Readings in Philosophy of Psychology*, vol. I, pp. 268–305. Cambridge, MA: Harvard University Press.

Block N and Stalnaker R (1999) Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* **108**, 1–46.

Churchland P (1985) Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy* **82**: 8–28.

Dretske F (1995) *Naturalizing the Mind*. Cambridge, MA: Bradford Books/MIT Press.

Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: Bradford Books/MIT Press.

Hardin CL (1988) *Color for Philosophers: Unweaving the Rainbow*. Indianapolis, IN and Cambridge, MA: Hackett.

Levine J (2001) *Purple Haze: The Puzzle of Consciousness*. New York, NY: Oxford University Press.

Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: Bradford Books/MIT Press.

Rey G (1996) *Contemporary Philosophy of MInd: A Contentiously Classical Approach*. Oxford: Blackwell.

Shoemaker S (1996) *The First-Person Perspective and Other Essays*. Cambridge, UK and New York, NY: Cambridge University Press.

Tye M (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: Bradford Books/MIT Press.

Externalism

Intermediate article

Robert A Wilson, University of Alberta, Edmonton, Canada

CONTENTS

Externalism is the view in the philosophy of mind that an individual's social or physical environment is partly determinative or constitutive of that individual's mental states.

WHAT IS EXTERNALISM?

The debate between individualists (or 'internalists') and anti-individualists (or 'externalists') is sometimes glossed in terms of whether psychological (or mental) states are 'in the head'. At first sight, that is likely to seem a trivial question: of course mental states are in the head ('Where else could they be?', as Robert Stalnaker (1989) asked). So we must first articulate a version of the issue that makes it clearer why externalism is a substantive, and thus potentially controversial, claim about the nature of the mind; we will then show why the (sometimes rarefied) debate over externalism has some important implications for how the mind should be studied with the cognitive sciences.

Consider the question of whether the character of an agent's environment plays some crucial role in determining or fixing the nature of that agent's mind. A natural thought (one, in fact, common to those who disagree about the answer to this question) would be that agents causally interact with their world, gathering information about it through their senses, and so the nature of their minds, in particular what their thoughts are about, is in part determined by the character of their world. Thus, the world is a causal determinant of one's thoughts, and hence one's mind. That is, the world is a contributing cause of the content, or intentionality, of one's mind, that is, what one perceives and thinks about. This is just to say that the content of one's mind is not causally isolated from one's environment. Individualists and externalists in the philosophy of mind agree on this much. What separates them is the question of whether there is some deeper sense in which the nature of the mind is determined by the character of the individual's world. It is this 'deeper sense' of world–mind determination that we need to articulate.

We can approach this task by extending the brief discussion above of the idea that the content of the mind is in part causally determined by the agent's environment to explore the conditions under which a difference in the world implies a difference in the mind. Individualists hold that this is so just in case that difference in the world makes some corresponding change to what occurs inside the boundary of the individual; externalists deny this, thereby allowing for the possibility that individuals who are identical with respect to all of their intrinsic features could none the less have psychological or mental states with different contents. And, assuming that mental states with different contents are *ipso facto* different types or kinds of states, this implies that an individual's intrinsic properties do not determine or fix that individual's mental states.

This provides us with another, more precise way of articulating the difference between externalism and individualism about the mind. Individualists claim, and externalists deny, that what occurs inside the boundary of an individual metaphysically determines the nature of that individual's mental states. The individualistic determination thesis, unlike the causal determination thesis, expresses a view about the nature or essence of mental states, and identifies a way in which, despite their causal determination by states of the world, mental states are autonomous or independent of the character of the world beyond the individual. What individualism implies is that two individuals who are identical in all their intrinsic respects must have the same psychological states. It is the modal aspect to this implication that makes supervenience an appropriate concept to use in stating individualism more precisely: an individual's psychological states must supervene on the intrinsic physical states of that individual.

This implication, and indeed the debate over externalism, is often made more vivid through the fantasy of doppelgangers, pairs of individuals identical at the molecular level, and the corresponding fantasy of 'Twin Earth'. We will return to these fantasies later in this article.

HISTORY OF AND ARGUMENTS FOR EXTERNALISM

Hilary Putnam's paper 'The meaning of "meaning" ' (1975) introduced both of the above fantasies in the context of a discussion of the meaning of natural language terms. Putnam was concerned to show that 'meaning' does not and cannot jointly satisfy two theses that it was often taken to satisfy by then-prevalent theories of natural language reference: the thesis that the meaning of a term is what determines its reference; and the thesis that meanings are 'in the head', where this phrase should be understood as making a claim of the type identified above about the metaphysical determination of meanings. These theses typified descriptive theories of reference – first formulated by Frege and Russell – according to which the reference of a term is determined by the descriptions that a speaker attaches to that term. In attacking this view and its presuppositions Putnam focused on terms denoting natural kinds, such as 'water' and 'tiger', but he intended his attack, and his subsequent alternative theory of natural language reference, the causal theory of reference, to be more general than this.

Consider an ordinary individual, Oscar, who lives on Earth and interacts with water in the ways that most of us do: he drinks it, washes with it, and sees it falling from the sky as rain. Oscar, who has no special chemical knowledge about the nature of water, will associate a range of descriptions with his term 'water': it is a liquid that one can drink, that is used to wash, and that falls from the sky as rain. In a descriptive theory of reference, these descriptions determine the reference of Oscar's term 'water'. That is, Oscar's term 'water' refers to whatever it is in the world that satisfies the set of descriptions he attaches to the term. And since those descriptions are 'in the head', natural language reference on this view is individualistic. But now, to continue Putnam's argument, imagine a molecule-for-molecule doppelganger of Oscar, Oscar*, who lives on a planet just like Earth in all respects but one: the substance that people drink, wash with, and see falling from the sky is not water (i.e., H_2O), but a substance with a different chemical structure, XYZ. Call this planet 'Twin Earth'. This substance, XYZ, is called 'water' on Twin Earth, and Oscar*, as a doppelganger or twin of Oscar, has the same beliefs about it as Oscar has about water on Earth. (Recall that Oscar, and thus Oscar* has no special knowledge of the chemical structure of water.) Twin Earth has what we might call 'twin water' or 'twater' on it, not water, and it is twater that Oscar* interacts with, not water – after all, there is no water on Twin Earth. Given that Oscar's term 'water' refers to or is about water, then Oscar*'s term 'water' refers to or is about twater. That is, they have natural language terms that differ in their meaning, assuming that reference is at least one aspect of meaning. But, by hypothesis, Oscar and Oscar* are doppelgangers, and so are identical in all their intrinsic properties, and so are identical with respect to what is 'in the head'. Thus, Putnam argues, the meanings of the natural language terms that Oscar and Oscar* use are not metaphysically determined by what is in the head of the individual using those terms.

Putnam was attacking a tradition of thinking about language which was, in terms that Putnam appropriated from Carnap's *Aufbau*, 'methodologically solipsistic': it treated the meanings of natural language terms, and language more generally, in ways that supposed that the world beyond the individual language user did not exist. Since Putnam's chief point is one about natural language terms and the relationship of their semantics to what is inside the head, one needs at least to extend his reasoning from language to thought to arrive at a position that denies individualism about the mind itself. But given the tradition to which he was opposed, such an extension might be thought to be relatively easy, since in effect those in the tradition of methodological solipsism – from Descartes to Brentano, to Russell, to Husserl, to Carnap – conceived of natural languages and their use in psychological terms.

The term 'individualism' itself, as well as a series of thought experiments which made a case for externalism and in some ways parallelled Putnam's 'Twin Earth' thought experiment, was introduced by Tyler Burge in his paper 'Individualism and the Mental' (1979). Burge identified individualism as an overall conception of the mind prevalent in modern philosophical thinking at least since Descartes, and argued that our commonsense psychological framework for explaining behavior, our 'folk psychology', was externalist. Burge was explicit in making a case against individualism that did not turn on any controversial claims about the

semantics of natural kind terms: he developed his argument using agents with thoughts about arthritis, sofas, and contracts, so did not presuppose any type of scientific essentialism about natural kinds. Like Putnam's argument, however, Burge's argument does presuppose some views about natural language understanding. The most important of these is that we can and do have incomplete understanding of many of the things that we have thoughts about and for which we have natural language terms.

VARIETIES OF EXTERNALISM

Putnam's externalism is sometimes characterized as a form of 'physical' externalism, in contrast to Burge's 'social' externalism. According to Putnam, the character of the physical world (e.g. the nature of water itself) in part metaphysically determines the content of one's mind; while according to Burge the character of the social world (e.g. the nature of one's linguistic community) does so. While these terms may serve as a useful reminder of one way in which these two views differ, we should also keep in mind the 'social' aspect of Putnam's view of natural language: his linguistic division of labor. Important to both views is the idea that language users and psychological beings depend and rely on one another in ways that are reflected in our everyday, common-sense ways of thinking about language and thought. Thus, on both views, there is a social aspect to the nature of meaning and thought. This is partly what justifies the label anti-*individualism* for both of them.

There is another division between varieties of externalism that turns on their relationship with the cognitive sciences. Taking their cue from the linguistic emphasis of the Putnam–Burge arguments for externalism, some externalists, such as McDowell (1986) and Pettit (1983), have developed their views within a broadly Wittgensteinian tradition of thinking about the mind, one that views externalism as latent in our ordinary language and the 'language games' that we play with mental terms. But other philosophers, such as Segal (1989), Egan (1992, 1995) and Shapiro (1993, 1997), locate the debate between individualists and externalists within the areas of the philosophy of science that concern cognition. These philosophers have probed general notions, such as that of computation and mental representation, central to understanding cognition, as well as exploring specific research programmes within cognitive science for their relationship to the externalism debate.

EXTERNALISM AND COGNITIVE SCIENCE

At around the time when individualism was coming under attack from the externalism of Putnam and Burge, it was also being defended (e.g., by Fodor (1980) and Stich (1983)) as a view of the mind particularly apt for a genuinely scientific approach to understanding the mind, especially of the type that was being articulated within the nascent interdisciplinary field of cognitive science. For these defenders of individualism, there was something suspiciously unnaturalistic about the Putnam–Burge arguments, as well as something about their conclusions that seemed anti-scientific. Part of the defence of individualism and the corresponding attack on externalism turned on what we may call the 'cognitive science gesture': the claim that, as contemporary empirical work on cognition indicated, any truly scientific understanding of the mind would need to be individualistic, and thus could not be externalist.

Common to both Fodor's and Stich's views of cognitive science is the idea that an individual's psychological states should be *bracketed off* from the environments beyond the head in which individuals find themselves. Unlike Putnam and Burge in the papers discussed above, Fodor and Stich have focused on the relevance of individualism to psychological explanation, and have used their respective principles to argue for substantive conclusions about the scope and methodology of psychology and the cognitive sciences. Fodor (1980) contrasted a 'methodologically solipsistic' psychology with what he called a naturalistic psychology, and argued that since the latter (in which he included J. J. Gibson's approach to perception, learning theory, and the naturalism of William James) was unlikely to lead to reliable research in psychology, methodological solipsism was the only fruitful psychological framework for understanding cognition (see also Fodor, 1987). One implication of Fodor's position is that the relevant notion of content for cognitive science is *narrow content* (White, 1991). Stich (1978) argued for a syntactic or computational theory of mind that made no essential use of the notion of intentionality or mental content. Thus he deployed individualism in defence of an eliminativist view of content (see also Stich, 1983).

Although the 'cognitive science gesture' is a gesture rather than a solid argument that appeals to empirical data, it is not an empty gesture. Although Fodor's and Stich's arguments have not won widespread acceptance either by philosophers or by

cognitive scientists, they have struck a chord with the latter, perhaps not surprisingly since the dominant research traditions in cognitive science have been at least implicitly individualistic. Relatively explicit commitments to an individualistic view of aspects of cognitive science include Chomsky's (1986, 1995) use of the distinction between two conceptions of language (the 'I'-language and the 'E'-language, for 'internal' and 'external' respectively), Jackendoff's (1991) related, general distinction between 'psychological' and 'philosophical' conceptions of the mind, and Cosmides and Tooby's (1994) more recent emphasis on the constructive nature of our internal, evolutionary-specialized cognitive modules.

Two further points in brief. First, there is a large literature that discusses Marr's (1982) theory of vision in connection with externalism, beginning with Burge (1986), which raises and discusses many issues relevant to assessing the relevance of externalism for cognitive science (see Wilson, 2002). Second, an important way to develop the externalist view of the mind is to pursue its connection to 'embodied' or 'embedded' approaches to cognition (e.g. Clark, 1997; Hutchins, 1995; McClamrock, 1995). These approaches question business as usual in cognitive science, and their development is likely to be fruitful for the study of cognition.

PROBLEMS FOR EXTERNALISM

Two important problems for externalism, especially for practicing cognitive scientists, are its perceived incompatibility with the insights of the computational and representational theories of mind. The former of these theories holds that mental processing is a form of computation. The latter theory holds that we interact with the world perceptually and behaviorally through internal mental representations of how the world is (as the effects of perceiving) or how the world should be (as instructions to act). Provided that the appropriate internal representational states of the organism remain fixed, the organism's more peripheral causal involvement with its environment is irrelevant to cognition, since that involvement cannot alter the internal mental states that represent that environment. Wide computationalism (Wilson, 1994 and 1995) is one response to the first of these problems; Burge's (1986) interpretation of Marr's theory of vision is one response to the second. (For replies, see Segal, 1989 and 1998, and Matthews, 1988.)

Externalism has also been thought to give rise to various related, but more purely philosophical, problems; for example, by failing to make sense of the notion of mental causation (e.g. Block, 1986), or misconstruing the role of causal powers in psychological taxonomy (Fodor, 1987). Connecting such objections of individualism with the methodological formulations that have influenced cognitive science, it has been claimed that individualism provides a minimal constraint needed to arrive at psychological explanations that locate the mind suitably in the physical world, a psychology that taxonomizes its entities by their causal powers.

A further philosophical problem for externalism concerns its compatibility with a cluster of related epistemological issues: those concerning self-knowledge, a-priori knowledge, and scepticism.

Basic to self-knowledge is knowledge of one's own mind, and traditionally this knowledge has been thought to involve some form of privileged access to one's own mental states. This notion of epistemic privilege has been developed in a number of ways, all of which share the idea that there is an asymmetry between knowledge of one's own mind and knowledge of the minds of others and of other things in the world. (Indicative of the depth of these asymmetries in modern philosophy is the fact that an introduction to epistemology that reflects on skepticism will likely introduce the *problem of other minds* and the *problem of our knowledge of the external world*, but not the corresponding *problem of self-knowledge*.) Skepticism about one's own mind has seemed to be precluded by the very nature of self-knowledge.

Individualistic conceptions of the mind have seemed well suited to making sense of first-person privileged access and the asymmetry between self-knowledge and knowledge of the mental states of others. If mental states are individuated in abstraction from the environment 'beyond the individual', then there seems to be no problem in understanding how the process of introspection – turning our mind's eye inwards, to use a common metaphor – reveals the content of those states. To invoke the Cartesian fantasy in a way that brings out the asymmetry between self-knowledge and other forms of knowledge: even if there were an evil demon who deceived me about the existence of an external world – including the existence of other people with mental states like mine – the one thing that I could be sure about would be that I am having experiences with a certain content. As it is sometimes put, even if I could be deceived about whether there is really a tree in front of me and thus about whether I am actually seeing a tree, I can't be deceived about whether it seems to me that I am

seeing a tree. Thus individualism seems to confer a certain epistemic security on first-person knowledge of one's own mental states which the corresponding third-person knowledge lacks.

Externalism, by contrast, poses a *prima facie* problem for even the more modest forms of first-person privileged access, and has even been thought to call into question the possibility of any form of self-knowledge. For externalism claims that what mental states are is metaphysically determined, in part, by the nature of the world beyond the boundary of the subject of those states. Thus it would seem that in order to know what one is thinking (that is, to know the content of one's mental states), one would have to know something about the world beyond one's self. But this would be to assimilate our first-person knowledge of our own minds to our knowledge of other things, and so to deny any privileged access that self-knowledge might have. It implies that in order to know my own mind I need to know facts (perhaps difficult to discern) about the nature of the physical or social world in which I live; and so also suggests that in a range of ordinary cases where we might unreflectively attribute self-knowledge, I don't actually have self-knowledge.

We can express the problem in another way, which abstracts away from the differences between specific accounts of privileged access and specific accounts of externalism. Knowledge of one's own mental states, whether it be infallible, incorrigible, self-intimating, introspective, or a priori, has a special character. Knowing one's own mental states involves, *inter alia*, knowing their contents. Now, according to externalism, the contents of a subject's mental states are metaphysically determined, in part, by facts about his or her physical or social environment. Knowledge of these facts, however, does not have this special character. How is the special character of self-knowledge compatible with the non-special character of worldly knowledge, given the dependence of the former on the latter?

The problem can be represented as a supposedly inconsistent triad of propositions (see also McKinsey, 1991). Let *P* stand for the contents of our mental states; let *E* stand for facts about the environment; and let 'by introspection' stand for the special character of self-knowledge.

We know *P* by introspection. (Self-Knowledge) (1)

P are metaphysically determined in part by *E*. (Externalism) (2)

E are not known by introspection. (Common Sense) (3)

The claim is that one of these three propositions must be rejected. If we reject Self-Knowledge, then we give up on the idea that we have privileged access to our own minds; if we reject Externalism, then we return to individualism; and if we reject Common Sense, then we make a strange and implausible claim about our knowledge of the physical or social world.

References

Block N (1986) Advertisement for a semantics for psychology. In: French P, Uehling T and Wettstein H (eds) *Midwest Studies in Philosophy*, vol. X 'Philosophy of Mind'. Minneapolis, MN: University of Minnesota Press.

Burge T (1979) Individualism and the mental. In: French P, Uehling T and Wettstein H (eds) *Midwest Studies in Philosophy*, vol. IV 'Metaphysics'. Minneapolis, MN: University of Minnesota Press.

Burge T (1986) Individualism and psychology. *Philosophical Review* **95**: 3–45.

Chomsky N (1986) *Knowledge of Language*. New York, NY: Praeger.

Chomsky N (1995) Language and nature. *Mind* **104**: 1–61.

Clark A (1997) *Being There: Putting Mind, Body and World Together*. Cambridge, MA: MIT Press.

Cosmides L and Tooby J (1994) Foreword to Baron-Cohen S (1994) *Mindblindness*. Cambridge, MA: MIT Press.

Egan F (1992) Individualism, computation, and perceptual content. *Mind* **101**: 443–459.

Egan F (1995) Computation and content. *Philosophical Review* **104**: 181–203.

Fodor JA (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Science* **3**: 63–73. [Reprinted in: Fodor JA (1981) *Representations*. Sussex, UK: Harvester Press.]

Fodor JA (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Jackendoff R (1991) The problem of reality. [Reprinted in: Jackendoff R (1992) *Languages of the Mind*. Cambridge, MA: MIT Press.]

Marr D (1982) *Vision: A Computational Approach*. San Francisco, CA: Freeman.

Matthews R (1988) Comments on Burge. In: Grimm and Merrill (eds) *Contents of Thought*, pp 77–86.

McClamrock R (1995) *Existential Cognition: Computational Minds in the World*. Chicago, IL: University of Chicago Press.

McDowell J (1986) Singular thought and the extent of inner space. In: Pettit P and McDowell J (eds) *Subject, Thought, and Context*. Oxford: Oxford University Press.

McKinsey M (1991) Anti-individualism and privileged access. *Analysis* **51**: 9–16.
Pettit P (1983) Wittgenstein, individualism and the mental. In: *Epistemology and the Philosophy of Science: Proceedings of the Seventh International Symposium*. Vienna: Holder-Pichler-Tempsky.
Putnam H (1975) The meaning of 'meaning'. In: Gunderson K (ed.) *Language, Mind, and Knowledge. Minnesota Studies in the Philosophy of Science*, vol. 7. Minneapolis, MN: University of Minnesota Press. [Reprinted in: Putnam H (1979) *Mind, Language, and Reality*. New York, NY: Cambridge University Press.]
Segal G (1989) Seeing what is not there. *Philosophical Review* **98**: 189–214.
Segal G (1997) Review of R. A. Wilson, Cartesian Psychology and Physical Minds. *British Journal for the Philosophy of Science* **48**: 151–157.
Shapiro L (1993) Content, kinds, and individualism in Marr's theory of vision. *Philosophical Review* **102**: 489–513.
Shapiro L (1997) A clearer vision. *Philosophy of Science* **64**: 131–153.
Stalnaker RC (1989) On what's in the head. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. III. Atascadero, CA: Ridgeview.
Stich S (1978) Autonomous psychology and the belief–desire thesis. *Monist* **61**: 573–591.
Stich S (1983) *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
White S (1991) *The Unity of the Self*. Cambridge, MA: MIT Press.
Wilson RA (1994) Wide computationalism. *Mind* **103**: 351–372.
Wilson RA (1995) *Cartesian Psychology and Physical Minds: Individualism and the Sciences of the Mind*. New York, NY: Cambridge University Press.
Wilson RA (2002) Individualism. In: Warfield T and Stich S (eds) *The Blackwell Guide to the Mind*. Oxford: Blackwell Publishers.

Further Reading

Adams F, Drebushenko D, Fuller G and Stecker R (1990) Narrow content: Fodor's folly. *Mind and Language* **5**: 213–229.
Burge T (1982a) Other bodies. In: Woodfield, 1982, pp. 97–120.
Burge T (1982b) Two thought experiments reviewed. *Notre Dame Journal of Formal Logic* **23**: 284–293.
Burge T (1988a) Cartesian error and the objectivity of perception. In: Grimm and Merrill (eds) *Contents of Thought*, pp. 62–76. Tucson, AZ: University of Arizona Press. Also in: Pettit P and McDowell J (eds) (1986) *Subject, Thought, and Context*, pp. 117–136. Oxford, UK: Oxford University Press.
Burge T (1988b) Individualism and self-knowledge. *Journal of Philosophy* **85**: 649–663.
Burge T (1989) Individuation and causation in psychology. *Pacific Philosophical Quarterly* **70**: 303–322.
Davies M (1991) Individualism and perceptual content. *Mind* **100**: 461–484.
Devitt M (1990) A narrow representational theory of mind. In: Lycan W (ed.) *Mind and Cognition: A Reader*. New York, NY: Blackwell.
Egan F (1999) In defense of narrow mindedness. *Mind and Language* **14**: 177–194.
Fodor JA (1982) Cognitive science and the Twin Earth problem. *Notre Dame Journal of Formal Logic* **23**: 98–118.
van Gulick R (1989) Metaphysical arguments for internalism and why they don't work. In: Silvers S (ed.) *Representation*. Dordrecht, The Netherlands: Kluwer.
Ludlow P and Madin N (eds) (1998) *Externalism and Self-Knowledge*. Palo Alto, CA: CSLI Publishers.
McCulloch G (1995) *The Mind and Its World*. New York, NY: Routledge.
McGinn C (1989) Mental Content. New York, NY: Blackwell.
Millikan R (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.
Patterson S (1991) Individualism and semantic development. *Philosophy of Science* **58**: 15–35.
Rowlands M (1999) *The Body in Mind: Understanding Cognitive Processes*. New York, NY: Cambridge University Press.
Segal G (2000) *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.
Walsh DM (1999) Alternative individualism. *Philosophy of Science* **66**: 628–648.
Wilson RA (2000a) The mind beyond itself. In: Sperber D (ed.) *Metarepresentation*. New York, NY: Oxford University Press.
Wilson RA (2000b) Some problems for 'alternative individualism'. *Philosophy of Science* **67**: 671–679.
Wilson RA (2001) Two views of realization. *Philosophical Studies* **104**: 1–30.
Woodfield A (ed.) (1982) *Thought and Object: Essays on Intentionality*. Oxford, UK: Oxford University Press.

Folk Psychology

Intermediate article

Shaun Nichols, College of Charleston, Charleston, South Carolina, USA

CONTENTS

Folk psychology is the body of information people have about the mind. It is often regarded as the basis for our capacity to attribute mental states and to predict and explain actions.

INTRODUCTION

In its broadest sense, folk psychology is the information that lay people have about the mind. Although the scope of folk psychology is thus vast, contemporary discussion of folk psychology in philosophy and cognitive science focuses mainly on the portion of folk psychology that guides the prediction and explanation of actions. This portion of folk psychology plays a central role in our everyday lives where folk-psychological prediction and explanation abound. We engage in such prediction for mundane tasks, like trying to figure out what the baby wants, what your peers believe about your work, and what your spouse will do if you arrive home late. Folk psychology is also involved in loftier endeavors, like trying to understand Descartes' reasons for thinking that many ideas are innate. So pervasive is the role of folk psychology in our lives that Jerry Fodor has remarked that if folk psychology should turn out to be seriously mistaken, it would be 'the greatest intellectual catastrophe in the history of our species' (Fodor, 1987, p. xii).

HISTORY

The idea that lay people have views about beliefs and desires and that people believe that actions result from beliefs and desires is hardly new. Indeed, Descartes would never have denied that people have such information. However, according to Descartes, the core information about the mind comes from introspection, which he regarded as an infallible source. Since introspection reveals the truth about the mind to each of us, for Descartes, the somewhat pejorative qualifier 'folk' is unnecessary: folk psychology *is* psychology.

In the twentieth century, two developments led to a revolutionary new picture of lay views of the mind. The first development was the challenge to introspection as a source of knowledge about the mind. This challenge occurred in both psychology and philosophy. In the early part of the century, psychologists advocating methodological behaviorism maintained that introspection was scientifically disreputable and could not be a source of knowledge about the mind. In effect, they forswore all talk of mental states in scientific psychology. In philosophy, Ryle (1949) launched a more sweeping attack. In his view, sometimes labeled 'logical behaviorism', it is a mistake to think that there are beliefs and desires inhering in an unobservable mind. Unlike the methodological behaviorists, though, Ryle was not against using terms like 'belief' and 'desire'. Rather, he maintained that such terms refer not to internal mental states, but to publicly observable phenomena, in particular, to dispositions to behave in certain ways under certain conditions. One important consequence of Ryle's view is that, since beliefs and desires are not internal states, they cannot possibly be revealed by introspection. Suspicions about introspection have exercised a powerful hold over psychology and the philosophy of mind ever since, even among those not sympathetic to logical behaviorism.

The second development occurred in the wake of the first. If we cannot rely on introspection to provide us with knowledge of the mind, then we need a new account of the source of our knowledge about the mind. Wilfrid Sellars (1956) developed what has turned out to be the most influential alternative to the introspectionist account of lay knowledge about the mind. Rather than maintain that the mind reveals its secrets to itself through introspection, Sellars suggests that lay people have a *theory* of the mind. Sellars proposes a myth about the origins of our common-sense view. He suggests that in the distant past, our ancestors never spoke of internal mental states like beliefs and desires.

Rather, these 'Rylean' ancestors only spoke of publicly observable phenomena like behavior and dispositions to behave. They even lacked terms for inner mental states. Then one day Jones, a great genius, arose from this group. Jones recognized that positing inner states like *thoughts* as theoretical entities provided a powerful basis for explaining the verbal behavior of his peers, and Jones developed a *theory* according to which such behavior is indeed the expression of internal thoughts. Jones then taught his peers how to use the theory to interpret the behavior of others. We are ultimately the beneficiaries of Jones' genius, since we still use his theory to interpret others' behavior.

Although Sellars explicitly presents this origin story as a myth, the point is that the myth allows us to see clearly a new picture of the nature of our common-sense views about the mind. In this picture, the common-sense view is a theory of mind, and this theory posits inner mental states, like thoughts, that are not publicly observable. So, the myth can be seen as one possible (and surely false) account of the origin of the theory. Sellars has provided us with an alternative account of common-sense psychology that does not rely on introspection. Nor, however, does it adopt the logical-behaviorist view that terms like 'thought' refer to publicly observable phenomena. This idea that folk psychology is a theory, whatever its origins, has come to be known as the 'theory theory' (Morton, 1980). Not only does the theory theory provide a new way to construe commonsense psychology once introspection had been displaced, but Sellars's account provides a new way to construe introspection itself. For Sellars suggests that the common-sense theory of mind is what we use to attribute mental states to ourselves as well as others.

FOLK PSYCHOLOGY AND THE SCIENTIFIC VIEW OF THE MIND

The very possibility that the theory theory might be correct requires us to be explicit about the potential gap between folk psychology and the scientific view of the mind. For if lay views about the mind derive, not from (infallible) introspection but from a common-sense theory, then they may well not cohere with mature scientific views of the mind. It now becomes a question of some moment whether folk psychology adequately characterizes the mind.

In order to address this question, one needs to know much more precisely what folk psychology is. In philosophy, one prominent way of characterizing folk psychology is to say that it consists of ideas or platitudes that everyone accepts, such as 'Persons in pain tend to want to relieve that pain. Persons who feel thirst tend to desire drinkable fluids. Persons who are angry tend to be impatient' (Churchland, 1988, pp. 58–59). According to Churchland, among others, the collection of all such platitudes constitutes the folk theory of mind that guides the prediction, explanation and interpretation of behavior. If one assembled such a list, one might then determine whether individual items on the list are corroborated or refuted by mature science.

Here the most prominent theme in philosophical discussions of folk psychology emerges. For, some suggest, if the folk account diverges widely from the scientific account, then we should conclude that the folk theory is wrong. Indeed, it may turn out that the folk theory is so thoroughly wrong that we must reject the theoretical posits of 'belief' and 'desire' and acknowledge that beliefs and desires don't really exist. This is the position of 'eliminative materialism': the folk theory is so far off the mark that we need to uproot the ontology of folk psychology entirely, just as we have uprooted the ontology of the supernatural. Eliminativist arguments have been developed in two rather different ways. Some, like Stich (1983), maintain that folk psychology will be at odds with a mature scientific psychology and that for this reason we may need to reject the folk ontology of beliefs and desires. Others, like Churchland (1981), see neuroscience as the proper scientific approach to the mind and argue that folk psychology will be at odds with a mature neuroscience; therefore, the folk ontology should be rejected in favour of a neuroscientific ontology.

Evaluating eliminative materialism is a complex undertaking. Eliminativist arguments typically depend on important assumptions about reference, reductionism, and other controversial issues in metaphysics (see, e.g. Lycan (1988), Stich (1996)). The eliminativist claim of primary interest here, though, is that folk psychology is a fundamentally mistaken theory. This has been a subject of much recent debate. Eliminativists like Churchland (1981) bemoan the explanatory failures and limitations of folk psychology, and maintain that these shortcomings indicate that mature science will be at odds with it. Others, however, like Fodor (1987), have celebrated the success of folk psychology. Indeed, Fodor maintains that folk psychology is much better at predicting behavior than contemporary scientific approaches, and that this predictive success suggests that the folk theory is roughly right and hence will be compatible with a mature cognitive science.

FOLK PSYCHOLOGY AS TACIT KNOWLEDGE

While philosophers have debated the continuities between science and folk psychology and the consequences that would follow from various scenarios, cognitive scientists have been concerned with exploring more systematically the nature of the capacity to attribute beliefs, desires, and emotions, and the capacity to predict and explain behavior. It has become increasingly clear that the 'platitude' view of folk psychology does not suffice to explain the lay capacity for psychological attribution, prediction, and explanation.

To take a simple example, people are quite good at inferring others' emotions on the basis of facial expressions. Very small differences in muscular activity in the face determine the emotion we attribute, and people do not seem to be able to articulate the principles behind these attributions. Indeed, some of these processes occur outside conscious awareness. For instance, subjects are more likely to judge a neutral face as sad if they have just been subliminally exposed to a smiling face (Underwood, 1995). Similarly, the attribution of goals from motion cues seems to exceed the informational resources of the available platitudes. In a famous study, Heider and Simmel (1944) showed subjects geometric objects moving around in a two-dimensional scene. Almost all subjects attributed goals to the geometric objects, and there was a good deal of consistency in subjects regarding certain events as chasing, fighting, and trying to get out of a box. What is guiding subjects' judgments here? Certainly we have no platitudes about the likely goals of geometric objects. The judgments seem to be guided instead by low-level motion cues. Determining which motion cues tend to produce which judgments is an area of active research, and the details are far from worked out. But there are a number of cues that seem to contribute to judgments of intention in geometric displays, including relative speed following a position change, orientation of the object relative to direction of motion, and the turning axis of the geometric object (see, e.g., Scholl and Tremoulet (2000)). It seems very unlikely that the final account of the motion properties that elicit judgments of intention will correspond with folk platitudes.

Such experiments suggest that the mechanisms underlying folk-psychological capacities are more intricate than is suggested by platitude accounts. Prediction, explanation and attribution are unlikely to be a matter of applying platitudes to instances. As with other interesting cognitive capacities (such as language comprehension and production and folk physics), we can expect that much of the information underlying the capacity is not consciously accessible. This does not refute the theory theory, of course, since one can simply adopt the view that the folk-psychological theory is at least partly 'tacit'. Indeed, that has long been the prevailing assumption in empirical research on folk psychology.

By far the most extensive empirical research on folk psychology has been focused on charting the development of folk-psychological capacities in children. This research illustrates in a dramatic way the central role that folk psychology plays in our everyday lives. The distressing social deficits of children with autism have been linked to a breakdown in their capacity for folk psychology (Baron-Cohen, 1995). Furthermore, an analysis of everyday speech in normal children indicates that from a very young age, they talk a great deal about beliefs, desires, intentions, and emotions (Bartsch and Wellman, 1995). And this is not just talk. Experimental evidence indicates that children are good at predicting a person's behavior on the basis of that person's beliefs, desires, and emotions (Gopnik and Meltzoff, 1997). Some important capacities emerge at a surprizingly early age. For instance, there is evidence that 18-month-old children can attribute desires on the basis of facial expressions (Repacholi and Gopnik, 1997). Most researchers in the field agree that these capacities depend on a tacit theory of mind. In view of the early emergence of folk psychology in normal children and its breakdown in autism, many researchers maintain further that the tacit folk psychology theory has an innate basis. However, even among those who agree that folk psychology is a tacit theory with an innate basis, there is much disagreement about the nature of the tacit theory. On one account, this body of information is very much like a scientific theory, and the process of development is really a process of theory revision, much like the process of theory revision in science (see, e.g. Gopnik and Meltzoff (1997)). On another account, the tacit theory is not at all like a scientific theory, but rather is represented in an innate module with restricted information flow to other parts of the mind. This module does have to develop, but its development is a constrained process of maturation rather than an open-ended theory revision process (Leslie, 1994).

SIMULATION THEORY

While theory theorists debate among themselves the nature of the folk theory, there is a major

challenge to the theory theory, the 'simulation theory'. According to the simulation theory, one does not use a psychological theory in predicting a person's behavior; rather, one pretends to have the mental states of the person and then runs one's own decision-making mechanisms 'offline' using these inputs. The resulting decision is then used to predict what the person will do (Gordon, 1986; Goldman, 1989). This approach is a departure from the theory theory since it explains important folk-psychological capacities by appealing to something other than a tacit body of knowledge about the mind. In fact, the simulation theory has made the very term 'folk psychology' problematic since this term is usually associated with the theory theory. As a result, 'mindreading' is often used as a more theoretically neutral term for the cluster of capacities we have for attributing mental states and predicting and explaining behavior.

The simulation theory has one overriding virtue. It gives an elegant explanation of the remarkable success of lay prediction of thought and action. Typically, when I am trying to guess what a target person is going to do, it's plausible that the target and I share similar cognitive mechanisms. So if I use my own cognitive mechanism to run a simulation, it's likely that the outcome of that simulation will be very much like what the target's analogous mechanism will actually do. This virtue of the simulation theory is illustrated well by an example from Paul Harris (1992). Harris asks us to imagine a psycholinguistics experiment in which we are to predict the grammaticality judgments of another English speaker on a set of sentences. It seems likely that we would be reasonably accurate at such a task. How is it that we would do so well? A simulation-theoretic explanation is that we use our own mechanisms for judging grammaticality and then we attribute the output to the subject. So, if our own mechanisms produce the judgment that a sentence is grammatical, we attribute that judgment to the subject. Given that the mechanisms that produce judgments of grammaticality are extremely complex, to explain success in this task, a theory-theoretical account would have to appeal to an improbably vast amount of tacit knowledge about how people make grammaticality judgments. The simulation-theoretic alternative is almost certainly a better explanation.

It is important to note that in Harris's example, no pretense is involved; so the example is not entirely parallel to the 'offline simulation' account. But the example succeeds in showing that prediction of behavior will at least sometimes rely on resources that go beyond a tacit theory of psychology, and that in some cases at least, we use a simulation-like process to predict others' behavior.

Thus, simulation (or simulation-like) processes plausibly play some role in mindreading. How much of a role does simulation play? One proposal, 'radical simulation', is that simulation explains everything that the theory theory purported to explain, including all attribution of mental states, prediction of behavior, and explanation of behavior (see, e.g. Gordon (1986 and 1996)). This radical view led some simulation theorists to suggest that there *is* no folk psychological theory. It was noted above that even Descartes would not have challenged the claim that people have a body of information about the mind. But radical simulationists challenge precisely that claim. This would imply that the debate over eliminativism needs to be overhauled. For if there is no folk-psychological theory, it cannot be the case that the theory is false.

However, the radical simulation theory has few advocates. One problem for radical simulation is that there are many cases in which we are very bad at predicting what people will do, and simulation theory has difficulty explaining our failures in these cases (Stich and Nichols, 1992). There is a large body of literature in social psychology documenting situations in which subjects behave in ways that seem surprising to common sense. For example, in an experiment demonstrating the 'endowment effect', Kahneman *et al*. (1990) gave a coffee mug to subjects in one group, the 'endowed' group, and then offered the subjects the opportunity to trade the mug for various amounts of cash. Another group of subjects, the 'unendowed' group, was not given the mug but was allowed to choose between the mug and various amounts of cash. Subjects in the endowed group tended to place a much higher cash value on the mug than subjects in the unendowed group. Most people find this result interesting, and part of the reason it's interesting is because it's surprising. Yet if simulation theory explained all of mindreading, it's puzzling why we are surprised by the result. For we should be able to imagine ourselves in the endowed subject's situation and let our own cognitive mechanisms determine what value we would place on the object. Indeed, recent research confirms that we are not successful at this imaginative exercise. Loewenstein and Adler (1995) explored whether subjects would be able to predict the value they themselves would set on an object if they were endowed with it. They handed subjects a coffee mug and asked the subjects to imagine that they owned the mug and to indicate how much they

would be willing to sell it for if they owned it. After they had completed this part of the task, they were told that they could in fact keep the mug – it was theirs. Then they were once again asked to indicate how much they would be willing to sell it for. It turned out that subjects were bad at predicting their own judgments. They tended to want significantly more money when they were endowed with the object than when they were merely imagining that they were endowed with it. If we made judgments like this using simulation, one would expect subjects to be very good at this sort of task. And this is only one example among many similarly surprising results from social psychology. Indeed, if the simulation theory accounted for all of our mindreading capacities, one would expect social psychology to be a radically different discipline, with few surprising results. The theory theory, by contrast, can provide a natural explanation of these folk mistakes by maintaining that the body of information that guides our judgments in these cases is incomplete. The tacit theory is missing information in crucial places, and the absence of this information leads to the mistakes in our predictions.

A second problem with radical simulation is that for many normal, successful attributions of mental states, it is difficult even to devise a simulation theory that would explain how we arrive at the mental state attributions. This point is perhaps clearest in the case of attribution of perception. You know that if you are sitting at the dinner table, the person opposite you can see your face but not your knees. But if she puts her head under the table, then she may well be able to see your knees. It is not at all clear how pretending to have the mental states of the other person is going to tell you whether she has visual access to your knees. By contrast, the tacit theory approach suggests an obvious answer. You calculate information about the opacity of the table, the line of sight, the amount of ambient light, and so on, and from this information the tacit theory of mind (which would presumably have information about when a subject is likely to see something) produces a perceptual attribution.

Although radical simulation currently has few advocates, the simulation theory has considerably altered the landscape of folk psychology. Researchers from both sides of the simulation debate have mostly converged to the view that a full explanation of the capacities for attribution, prediction and explanation will require a hybrid account, appealing both to simulation processes and to tacit knowledge about mental states (see, e.g. Goldman (2000); Nichols and Stich (forthcoming)). Simulation theory requires a significant revision of the idea that lay understanding of other minds derives from a folk-psychological theory, and this has implications for both cognitive science and philosophy. Firstly, it suggests that the capacity for mindreading cannot be entirely captured by traditional cognitive models. The kinds of mechanisms exploited in mindreading will be very different from the kinds of mechanisms exploited in, say, folk physics and folk biology. Secondly, if simulation processes play an important role in mindreading, much of the eliminativist debate is too crude. For a large part of our mindreading capacity may be insulated from the eliminativist critique. On the other hand, in so far as the success of mindreading depends on simulation-like mechanisms (as opposed to tacit theory), the success of mindreading cannot be casually used to support the claim that the folk-psychological theory is largely true.

INTROSPECTION REVISITED

A second major challenge to the theory theory emerges from a reassessment of the capacity for introspection. One of the historical precursors of the theory theory was the repudiation of introspection as a reliable source of information about the mind. However, introspection fell into disrepute largely under the influence of behaviorism. It is not clear that the reasons for renouncing introspection are still decisive for cognitive scientists who reject behaviourism.

Many cognitive scientists, in fact, still maintain that the available evidence indicates that the only way to access one's own mental states is via the theory of mind (e.g., Nisbett and Wilson (1977); Gopnik (1993)). This view effectively embraces Sellars's early suggestion that the folk theory of mind is essential not only for attributing mental states to others but also for attributing mental states to oneself. Although the theory-theoretical approach to introspection is still very influential, some researchers have recently begun to defend alternative cognitive accounts of introspection according to which access to our own minds does not depend on the theory of mind (Goldman, 1993; Nichols and Stich, 2002). Goldman suggests further that introspective access might provide the basis for our concepts of belief and desire. In this case it might turn out that the theory of mind depends on introspection rather than the reverse.

The question of the nature of introspection and its possible role in generating our understanding of others' minds is far from settled. But whatever the outcome of this debate, there is every reason to

think that our capacity for mindreading is subserved by a diverse and intricate set of mechanisms. Although it is likely that the theory theory explains part of that capacity, it probably cannot provide a complete account. Rather, our capacity for mindreading probably also depends on simulation-like processes and perhaps even introspective mechanisms that are independent of the theory of mind.

References

Baron-Cohen S (1995) *Mindblindness*. Cambridge, MA: MIT Press.

Bartsch K and Wellman H (1995) *Children Talk About the Mind*. New York, NY: Oxford University Press.

Churchland P (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**: 67–90.

Churchland P (1988) *Matter and Consciousness*. Cambridge, MA: MIT Press.

Fodor J (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Goldman A (1989) Interpretation psychologized. *Mind and Language* **4**: 104–119.

Goldman A (1993) The psychology of folk psychology. *Behavioral and Brain Sciences* **16**: 15–28.

Goldman A (2000) The mentalizing folk In: Sperber D (ed) *Metarepresentation*. Oxford: Oxford University Press.

Gopnik A (1993) How do we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**: 1–14.

Gopnik A and Meltzoff A (1997) *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gordon R (1986) Folk psychology as simulation. *Mind and Language* **1**: 158–171.

Gordon R (1996) Radical simulation In: Carruthers P and Smith P (eds) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Harris P (1992) From simulation to folk psychology: the case for development. *Mind and Language* **7**: 120–144.

Heider F and Simmel M (1944) An experimental study of apparent behavior. *American Journal of Psychology* **57**: 243–259.

Kahneman D, Knetsch J and Thaler R (1990) Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* **98**: 1325–1348.

Leslie A (1994) ToMM, ToBY and Agency: core architecture and domain specificity. In: Hirschfeld L and Gelman S (eds) *Mapping the Mind*, pp. 119–148. Cambridge, UK: Cambridge University Press.

Loewenstein G and Adler D (1995) A bias in the prediction of tastes. *The Economic Journal: The Quarterly Journal of the Royal Economic Society* **105**: 929–937.

Lycan W (1988) *Judgment and Justification*. Cambridge, UK: Cambridge University Press.

Morton A (1980) *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Oxford: Clarendon Press.

Nichols S and Stich S (forthcoming) *Mindreading*. Oxford: Oxford University Press.

Nichols S and Stich S (2002) 'How to read your own mind: a cognitive theory of self-consciousness'. In: Smith Q and Jokic C (eds) *Consciousness: New Philosophical Essays*. Oxford: Oxford University Press.

Nisbett R and Wilson T (1977) Telling more than we can know. *Psychological Review* **84**: 231–259.

Repacholi B and Gopnik A (1997) Early understanding of desires: evidence from 14 and 18 month olds. *Developmental Psychology* **33**: 12–21.

Ryle G (1949) *The Concept of Mind*. London: Hutchinson.

Scholl B and Tremoulet P (2000) Perceptual causality and animacy. *Trends in Cognitive Sciences* **4**: 299–309.

Sellars W (1956) Empiricism and the philosophy of mind. In: Feigl H and Scriven M (eds) *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, pp 253–329. Minneapolis, MN: University of Minnesota Press.

Stich S (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Stich S (1996) *Deconstructing the Mind*. New York, NY: Oxford University Press.

Stich S and Nichols S (1992) Folk psychology: simulation or tacit theory? *Mind and Language* **7**: 35–71.

Underwood G (1995) Subliminal perception on TV. *Nature* **370**: 103.

Further Reading

Carruthers P and Smith P (eds) (1996) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Davies M and Stone T (eds) (1995a), *Folk Psychology: The Theory of Mind Debate*. Cambridge, MA: Blackwell.

Davies M and Stone T (eds) (1995b) *Mental Simulation: Evaluations and Applications*. Cambridge, MA: Blackwell.

Stich S and Ravenscroft I (1994) What *is* folk psychology? *Cognition* **50**: 447–468.

Free Will

Intermediate article

David Hodgson, Supreme Court of New South Wales, Sydney, New South Wales, Australia

CONTENTS

The expression 'free will' has been understood as a capacity of human beings to engage in voluntary conduct in accordance with their own unconstrained choices. The questions of how such capacity could relate to causation as understood by the physical sciences, and whether human beings do have such capacity, have not yet been decisively answered.

INTRODUCTION

Problems associated with the correct understanding of the concept of free will, and with the existence and exercise of the human capacity for voluntary conduct, have been debated since the time of the Greek philosophers. Developments in science, particularly during the last four centuries, have tended to support the universal applicability of laws of nature, apparently leaving little scope for any capacity for voluntary conduct that is not itself entirely subject to and determined by the same laws. However, scientific understanding of consciousness is still very limited, and there are reasons to think that consciousness may play a role in human behavior that is not entirely governed by or explicable in terms of the operation of laws of nature upon physical systems. Accordingly, there are still questions concerning free will that are far from settled.

THE CONCEPT OF FREE WILL

The expression 'free will' refers to a capacity or power, supposedly possessed by human beings, to engage in voluntary conduct in accordance with their own unconstrained choices. The questions whether or not human beings do have such a capacity, and if so, how it works and what is its significance, depend in part on what the concept itself is taken to involve.

Most would agree that the concept involves an ability to act consciously, rationally, and without constraint. The word 'will' refers to the capacity for volition, in the sense of the initiation and execution of voluntary conduct; and voluntary conduct is understood as requiring consciousness and control of the conduct in question. The word 'free' carries the notion that the conduct is the result of a choice unconstrained by matters outside the control of the person acting. Control of conduct and choice themselves presuppose some minimum rationality.

Associated with the concept of free will is the concept of *responsibility*: human beings are widely regarded as being responsible for their voluntary actions, precisely because these actions are considered to be under their control. Accordingly, it is widely regarded as appropriate to praise or blame, or reward or punish, a person for voluntary conduct, to an extent that is in some sense proportional to the merit or demerit of the conduct. This is reflected in criminal law, which includes quite elaborate rules for determining whether persons are to be considered responsible for conduct that is objectively in breach of the law.

A central problem concerning the concept of free will is how it relates to theories of causation. One common view of causation is that for an event A to be caused by prior events B, C, D, etc., those prior events must amount to sufficient conditions for the occurrence of event A. For free will to make sense, it would seem that a person's voluntary conduct would have to be caused by the person's exercise of the capacity to engage in such conduct; but what then causes the exercise of the capacity? On this view of causation, there must have been prior sufficient conditions for this, and further prior sufficient conditions for those prior sufficient conditions, and so on.

Associated with this issue is the relationship between free will and *determinism*. If all events are caused by prior sufficient conditions, it would seem that any event that occurs, including any human action, must be determined, indeed

predetermined, by its causal antecedents. Another way of putting much the same idea is to say that any event that occurs is uniquely determined by prior conditions and laws of nature.

One controversy concerning free will is whether the concept of free will should be understood in a basic sense, according to which free will is not excluded even if all choices and exercises of control are themselves predetermined by prior sufficient conditions outside the control of the person acting; or whether it should be understood in a strong sense, as attributing to human beings a power to choose, the exercise of which is not itself entirely predetermined by prior conditions outside the control of the person acting.

The other main controversy concerning free will is whether or not human beings really do have free will, particularly if it is understood in a strong sense.

HISTORY

The history of ideas relevant to the relationship of free will to the cognitive sciences can be seen as one of continuing assaults on the strong sense of free will.

Although there were significant discussions of free will by Greek and medieval philosophers, a convenient starting point for the history of issues relevant today is the seventeenth-century philosopher René Descartes (1986). His famous dualist view of the human mind as an immaterial substance distinct from the material body was significant at the time, not so much because he asserted that there was a free-willing 'ghost' affecting the operation of the human brain, but because he asserted that there were no such ghosts anywhere else. His dualism thus legitimized the application of the physical sciences to everything apart from the human mind: all systems other than the human mind were conceived of as changing or developing over time in accordance with impersonal laws of nature, thus excluding even from animals any conscious control of their actions that was not simply a working out of causation amenable to the physical sciences.

Later in the seventeenth century came the physics of Isaac Newton. Newton's three laws of motion were proposed as applying without exception to all physical matter, and they asserted that, given the mass, position, and motion of any piece of matter at one time and the forces acting on it over a period following that time, its positions and motion during that period were uniquely determined; and Newton's law of gravitation gave rules determining the quantity and direction of the force of gravity at any point. Newton himself did not spell out what that might mean for the physical matter of the human brain or for the dualism of Descartes, but the eighteenth-century French mathematician Pierre Laplace pointed out that, under the Newtonian scheme, initial conditions plus laws of nature determined the future, which could thus be exactly calculated by a being with sufficient information and intelligence; and Laplace made no exception for the human brain, and left no room for any independent efficacy of free will.

Also in the eighteenth century, the British philosopher David Hume (1888) made a further assault on the strong sense of free will with his philosophy of causation and action. Causation, according to Hume, was simply a matter of the regularity with which one type of event followed another type of event; that is, at bottom, the succession or unfolding of events over time conformably with regularities that could be expressed as laws. Hume thereby denied any significant causal role to the efficacy which humans seem to experience in their own voluntary actions and choices.

On the other hand, the German philosopher Immanuel Kant (1956) defended the strong sense of free will by distinguishing between the operation of causation in the realm of phenomena (things-as-they-appear-to-us) and that in the realm of noumena (things-as-they-really-are-in-themselves); and suggesting that, while deterministic law-governed causation operates in the former realm, it does not or need not do so in the latter. Thus, he said, human beings as noumena can be free, although as phenomena they are subject to deterministic causation. Few today find this a satisfying resolution to the problem.

During the nineteenth century, progressive scientific developments gave further support to the view that all matter behaved wholly in accordance with impersonal laws. Two were of particular significance: James Clerk Maxwell showed how the forces associated with electromagnetic radiation, including visible light, could be calculated by reference to mathematical rules, analogous to Newton's law of gravitation; and Charles Darwin's theory of evolution provided an explanation both of how the huge complexity and variety of life could have arisen from simple beginnings by the operation of impersonal laws of nature, and also of how animals and humans could have come to give the appearance of making choices, even if they were in reality only operating as determined by the impersonal laws.

The twentieth century saw further developments challenging the strong sense of free will. Albert

Einstein's theories of relativity appeared to support and advance the view that there are universal laws of nature governing all systems: his special theory, among other things, provided for a modification to Newton's mechanics so as to achieve full consistency with Maxwell's theory; and his general theory offered a deeper explanation of the gravitational force, and also accounted for certain observations which could not be accounted for by Newton's theory. The work of Sigmund Freud directed attention to the fact that much human motivation is unconscious or barely conscious; and since then, it has been impossible to maintain a view of human agency as a matter of wholly conscious decisions based on wholly conscious motives.

The invention of quantum mechanics (QM) in the 1920s challenged the determinism of the classical physics of Newton, Maxwell, and Einstein; but the only indeterminism postulated by QM was *randomness*, so that QM did not directly support the strong sense of free will. In other ways, QM tended to confirm the universal application of laws of nature: it showed how events at the atomic level could be explained by laws that also accounted for the substantial accuracy of classical physics at larger scales; and it also provided the basis for Linus Pauling's explanation of the chemical bond in terms of quantum physics, showing how chemical properties of matter could be explained by basic physics.

In the second half of the century, Crick and Watson's discovery of the structure of DNA showed how life itself could be explained by the laws of physics and chemistry. And the advance of the cognitive sciences has enabled more and more aspects of the working of the human brain to be understood in terms of the operation of laws of nature on physical systems. (I will look in more detail at this below.) However, there is still little understanding of consciousness in general, and its role in voluntary conduct in particular; so that, as considered in the final section, the problems of free will cannot be considered solved.

DIFFERENT VIEWS AND THEORIES OF FREE WILL

Compatibilism

Most people accept that human beings have a capacity to engage in voluntary conduct, in accordance with their own unconstrained choices; but on the other hand, many think it implausible to suggest that human voluntary conduct somehow stands outside the causal order studied by the physical sciences. Accordingly, a popular theory of free will is that free will is compatible with that causal order. In substance, this is taken as meaning that free will is compatible with determinism, because the proponents of this view assume that any indeterminism suggested by QM does not relevantly affect the causation of actions, so that for all practical purposes associated with problems of free will, even QM can be considered deterministic; and compatibilism is sometimes called *soft determinism*.

This theory adopts the basic sense of free will introduced in the opening section, and claims that this sense captures all that is important in the concept (Dennett, 1984). Human beings have freedom and responsibility just because they are free to act in accordance with their own choices and to do whatever it is they most want to do; and it does not matter that their choices and their wants may themselves be determined by prior circumstances and impersonal laws. Furthermore, it is said, our choices can be rationally based on our beliefs and desires, even if they are at the same time causally determined by them or by their neural correlates. Compatibilism is also expressed in terms of one's ability to have done otherwise than one actually did *if* one had chosen or wanted to do otherwise, with this ability or freedom (and its associated responsibility) not being excluded even though it is the case that one *could not in fact* have chosen or wanted to do otherwise.

Compatibilism correctly asserts that determinism is not the same as *fatalism*: while fatalism suggests that what will be will be, no matter how much or how little we deliberate over alternative courses of action or strive to achieve goals, determinism can accept that deliberation and striving do make a difference. And compatibilism correctly asserts that determinism does not entail *predictability* of human conduct.

Some compatibilist thinkers have gone so far as to say that determinism is necessary for free will: following Hume, they have asserted that, if there were any indeterminism in human conduct, this would mean that actions arose from something other than a person's character, beliefs, and desires, and would not be under the person's control (see Hobart, 1934).

Another strand of compatibilist thought, again following Hume, is that the imposition of rewards and punishments makes sense only if they have consistent effects, in accordance with ordinary notions of causation; and that, quite independently of any strong sense of free will, punishment is justified by its consequences, and it is appropriate to limit punishment to voluntary conduct because for

the most part it is only voluntary conduct that can be deterred by threats of punishment (see Hart, 1968).

Hard Determinism

Another theory of free will agrees with compatibilism in holding that human actions are caused deterministically, but departs from compatibilism by asserting that this means that human beings do not have free will in any substantial sense. That is, it says that the basic sense of free will adopted by compatibilism does not capture what is important in the concept, and that determinism is both true and incompatible with free will as properly understood. This theory has been called *hard determinism*.

A central argument for the view that determinism is incompatible with free will is the 'consequence argument', which has been stated as follows by an opponent of determinism:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore the consequences of these things (including our present acts) are not up to us. (van Inwagen, 1983, p. 16.)

Hard determinists can support many of the compatibilist ideas outlined above. The real difference is that they say there can be no such thing as genuine free will and responsibility. One variant of the consequence argument, which highlights the basic dilemma of responsibility, has been put as follows:

> (1) There is a clear and fundamental sense in which no being can be truly self-determining in respect of its character and motivation in such a way as to be truly responsible for how it is in respect of character and motivation.
>
> (2) When we act, at a given time, the way we act is, in some quite straightforward sense, a function of the way we then are, in respect of character and motivation. We act as we act *because of* how we then are, in respect of character and motivation.
>
> (3) It follows that there is a fundamental sense in which we cannot possibly be truly responsible for our actions. For we cannot be truly responsible for the way we are, and we act as we act because of the way we are. (Strawson, 1986, pp. 311–12.)

So according to hard determinists, free will and responsibility are illusions, and it is best to fashion our interpersonal relationships and our legal systems in the light of this reality.

Libertarianism

The third main theory of free will has been called *libertarianism*: it asserts both that free will is incompatible with determinism, and that human beings do have free will, so understood.

Libertarianism meets the dilemma of responsibility by accepting that nothing a person does or can do *at the time of any choice or action* can make the person responsible for the way the person then is, in respect of character and motivation – but claims that the way the person then is, in respect of character and motivation, does not predetermine what the person does. It only predetermines what the alternatives are and how they appeal, and provides the capacity to choose between them; so that the person can be responsible for *the way the person exercises the capacity to choose*. That responsibility may be greater or less, by reason of the nature of the choice posed by the way the person now is – the harder it is, by reason of the person's inclinations, for the person to make the 'right' choice, the less blameworthy will be the 'wrong' choice – but on this view, normal adult human beings always have some responsibility for their choices and voluntary actions. And this means in turn that the person may have some responsibility *through prior choices* for the way the person now is in respect of character and motivation, and thus for presently operating reasons and the way they appeal (see Hodgson, 1999).

One variety of libertarianism supports what is called *agent causation*, the view that the causation of human actions is a special kind of causation; in that while ordinary events are caused by other prior events, human actions are caused not by prior events but by the human agents who perform the actions. This view faces the difficulty that events do seem to have an important role in the causation of human behavior, and so it is very hard to explain just how agent causation could work.

Although libertarians recognize that the randomness of QM cannot directly support the strong sense of free will, some do see three features of QM as providing indirect support: (1) its indeterminism, leaving room within which free rational choices could be made; (2) its nonlocality, suggesting the possibility of global influence from combinations of events in extended regions of the brain; and (3) its reference to observation, suggesting that participation of conscious observers may be as fundamental as laws of nature in the operation of the universe (Hodgson, 1991, 1996, 1999, 2002).

BEARING OF COGNITIVE SCIENCE ON FREE WILL

As noted earlier, the advance of the cognitive sciences has enabled more and more aspects of the

working of the human brain to be understood in terms of the operation of laws of nature on physical systems. It has thereby apparently further reduced the scope for any independent operation of human free will.

As considered in detail in other essays in this encyclopedia, much is now known about the operation of the brain as a physical system. At the level of the individual neurons of the brain, a great deal is now known about what causes them to fire or not fire; how electrical signals are transmitted within neurons and then passed across the synapses to other neurons by means of chemical transmitters; and how all this is affected by the chemistry of the brain. Much is also known about the patterns of connections between the neurons of the brain, and what regions of the brain are involved in particular cognitive functions and behaviour.

Along with physical investigation of the brain's operation, the cognitive sciences have thrown light on its functional organization, giving insights into how the physical processes of the network of neurons can give rise to sensation, perception, memory, problem-solving, emotions, and actions. The brain is considered as an information-processing system, which takes the information from sensory inputs, processes it in various ways, and gives out various kinds of outputs – generally resulting in physical actions, such as walking or speaking. Brain processes are sometimes considered analogous to those of a computing machine. This idea has its basis in the Church–Turing thesis, to the effect that anything that can be computed or calculated can be computed, given enough time, by a general purpose machine. There has been much theoretical and practical work on the idea that human cognitive performance (sensation, perception, reasoning, execution of decisions, etc.) can be reproduced or simulated by computers.

Efforts are being made to identify the brain processes involved in voluntary action, with some success, albeit without much understanding of the role of consciousness in these processes. Three other areas of investigation are of particular relevance to free will, suggesting that our conscious processes may be considered a tip of a cognitive iceberg, arguably having little if any role in determining what it is that we say and do: (1) investigations into the prodigious cognitive abilities of unconscious processes – for example, investigations into the pre-conscious information processing necessary for perception and for language use; into the so-called 'sleep-on-it' phenomenon, where an answer to a problem suddenly occurs to a person some time after ceasing to think consciously about it; and into the calculating and other capabilities of some savants, suggesting that all of us may have similar capabilities inaccessible to consciousness; (2) investigation of the nature and extent of unconscious motivation, not merely confirming that much of our motivation is unconscious, but also suggesting that we are adept at *rationalizing* our conduct by (unconsciously) fabricating, and then believing, plausible but untrue stories to explain why we did what we did (Gazzaniga, 1988; Wegner, 2002); and (3) investigations suggesting that consciousness comes too late for real-time control of actions.

One set of experiments in the third area (Libet *et al*., 1979) has suggested that there is a delay of about half a second between the arrival at the brain of a novel sensory stimulus and its experience in consciousness; and another (Libet *et al*., 1983) has suggested that, when actions are initiated, there are pre-conscious preparations over a similar time before the instant noted by the person acting as the instant of initiating the action. However, these results do not exclude conscious control in *shaping* voluntary conduct: in a concert performance by a pianist, consciousness comes too late to direct fingers to the right keys, but not necessarily too late to shape the performance in response to heard sounds and felt emotions. Also, as Libet himself pointed out, the results do not exclude a conscious veto exercisable right up to the instant of initiating an action. And they say nothing about considered decisions.

RELEVANCE OF FREE WILL ISSUES IN COGNITIVE SCIENCE

The fact remains that human brains seem to be adapted, at some cost in terms of complexity and energy use, to support conscious processes, suggesting that conscious processes must contribute significantly to survival and reproduction. And it is plausible that any such contribution would be by way of determining what voluntary conduct is undertaken – that is, how free will is exercised. The cognitive sciences have yet to explain how such a contribution could be made, in terms of the operation of laws of nature on physical systems.

Libertarians see an analogy here with the inability of logicians to explain human rationality in terms of the application of rules to premisses or data. 'Plausible reasoning' has yet to be justified by reference to rules of logic or probability or mathematics or any other rules of reasoning or information processing that human beings have discovered or devised: there is an element of

judgment in plausible reasoning that has not yet been, and perhaps cannot be, reduced to compliance with rules (Hodgson, 1991, 1995). One of the challenges to the cognitive sciences is to explain plausible reasoning in terms of rules, in a manner that is not self-defeating or circular, or alternatively to explain it in some other way – and any other way might well be consistent with strong free will.

Consciousness involves two tricks that do not appear to be performed by computers, and that have no known function in information processing as presently understood: I call them the *qualia trick* and the *chunking trick*. The former is the trick of associating types of physical processes in the brain with types of qualitative experience (or 'qualia') such as seeing blue or feeling pain; and the latter is the trick of chunking particular instances of these general types of experience into particular whole experiences of a person. This second trick is associated with what is called the 'binding problem' of consciousness, concerning how information carried in distinct parts of the brain (such as 'seeing red' and 'seeing circle') is brought together into a unitary conscious experience (such as 'seeing a red circle'). The cognitive sciences have had some success in identifying physical brain processes involved in both these tricks, but virtually none in explaining their functional role. And the understanding of these tricks at present falls far short of enabling the creation of an artificial system that could perform the tricks, or of stating what would distinguish such a system from one of similar information-processing power that could not perform them.

As one who prefers the libertarian view of free will, I believe that the two tricks may possibly be understood as having a function in combination with a third trick that they make possible, what I call the *selection trick*, involving the capacity to make selections between alternatives thrown up by unconscious information processing. On this view, unconscious information processing gives rise to alternatives available for selection, provides reasons on the basis of which a selection between them can be made, and determines the intensity with which such reasons are felt; and also gives rise to whole experiences having unique qualities with which general laws of nature cannot engage, but to which the person having the experiences can respond rationally in making the selection (see Hodgson, 2001).

This is a vision of free will that I believe could be usefully addressed by the cognitive sciences, particularly as they continue to investigate the unsolved problems associated with voluntary action, plausible reasoning, the qualia trick, and the chunking trick.

References

Dennett D (1984) *Elbow Room*. Oxford, UK: Oxford University Press.

Descartes R (1986) *Meditations on First Philosophy*, translated by J Cottingham. Cambridge, UK: Cambridge University Press.

Gazzaniga M (1988) *Mind Matters*. Boston, MA: Houghton Mifflin.

Hart HLA (1968) *Punishment and Responsibility*. Oxford, UK: Oxford University Press.

Hobart RE (1934) Free will as involving determinism and inconceivable without it. *Mind* **43**: 1–27.

Hodgson D (1991) *The Mind Matters*. Oxford, UK: Oxford University Press.

Hodgson D (1995) Probability: the logic of the law – a response. *Oxford Journal of Legal Studies* **14**: 51–68.

Hodgson D (1996) Nonlocality, local indeterminism, and consciousness. *Ratio* **9**: 1–22.

Hodgson D (1999) Hume's mistake. In: Libet B, Freeman A and Sutherland K (eds) *The Volitional Brain*, pp. 201–224. Thorverton, UK: Imprint Academic.

Hodgson D (2001) Constraint, empowerment, and guidance: a conjectural classification of laws of nature. *Philosophy* **76**: 341–370.

Hodgson D (2002) Quantum physics, consciousness, and free will. In: Kane R (ed.) *Oxford Handbook of Free Will*, pp. 85–110. New York: Oxford University Press.

Hume D (1888) *Treatise of Human Nature*, ed. LA Selby-Bigge. Oxford, UK: Oxford University Press.

Inwagen P van (1983) *An Essay on Free Will*. Oxford, UK: Oxford University Press.

Kant I (1956) *Critique of Practical Reason*, translated by LW Beck. New York: Macmillan.

Libet B, Gleason C, Wright E and Pearl D (1983) Time of conscious intention to act in relation to onset of cerebral activities (readiness potential). *Brain* **106**: 623–642.

Libet B, Wright E, Feinstein B and Pearl D (1979) Subjective referral of the timing for a conscious sensory experience. *Brain* **102**: 191–222.

Strawson G (1986) *Freedom and Belief*. Oxford, UK: Oxford University Press.

Wegner D (2002) *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Further Reading

Blakemore C (1988) *The Mind Machine*. London: BBC.

Claxton G (1994) *Noises from the Darkroom*. London: Aquarian.

Crick F (1994) *The Astonishing Hypothesis*. London: Simon & Schuster.

Double R (1991) *The Non-Reality of Free Will*. Oxford, UK: Oxford University Press.

Flew A and Vesey G (1987) *Agency and Necessity*. Oxford, UK: Blackwell.

Hameroff S, Scott A and Kaszniak A (eds) *Toward a Science of Consciousness II*. Cambridge, MA: MIT Press.
Honderich T (1993) *How Free Are You?* Oxford, UK: Oxford University Press.
Inwagen P van (1993) *Metaphysics*. Oxford, UK: Oxford University Press.
Kane R (1996) *The Significance of Free Will*. New York: Oxford University Press.
Kane R (ed.) (2001) *Free Will* (Blackwell Readings in Philosophy). Malden, MA: Blackwell.
Kane R (ed.) (2002) *Oxford Handbook of Free Will*. New York: Oxford University Press.
Libet B, Freeman A and Sutherland K (eds) (1999) *The Volitional Brain*. Thorverton, UK: Imprint Academic.
Lucas J (1993) *Responsibility*. Oxford, UK: Oxford University Press.
Nagel T (1986) *The View From Nowhere*. New York: Oxford University Press.
Nozick R (1981) *Philosophical Explanations*. Oxford, UK: Oxford University Press.
O'Connor T (ed.) (1995) *Agents, Causes, and Events*. New York: Oxford University Press.
Popper KR and Eccles JC (1977) *The Self and its Brain*. Berlin: Springer Verlag.
Stapp H (1993) *Mind, Matter and Quantum Mechanics*. Berlin: Springer Verlag.
Strawson G (1998) Luck swallows everything. *Times Literary Supplement*, 26 June, 8–10.
Thorp J (1980) *Free Will*. London: Routledge.
Trusted J (1984) *Free Will and Responsibility*. Oxford, UK: Oxford University Press.
Watson G (ed.) (1982) *Free Will*. Oxford, UK: Oxford University Press.

Frege, Gottlob

Introductory article

Jason Stanley, University of Michigan, Ann Arbor, Michigan, USA

Gottlob Frege (1848–1925) was a German mathematician and philosopher widely credited with discovering modern quantificational logic.

Gottlob Frege (1848–1925), a German mathematician and philosopher, did not receive much recognition during his lifetime. However, Frege is now recognized as one of the greatest logicians since Aristotle. Frege's contributions to mathematical logic did not, like those of, say, Georg Cantor or Kurt Godel, consist of proofs of startling new results. Rather, Frege, in his 1879 work, *Begriffsschrift*, introduced the concept of a formal system, in which mathematical proofs could be carried out rigorously and precisely. The theory developed in the *Begriffsschrift* was what would now be called second-order quantification theory, in which there are quantifiers ranging over objects and quantifiers ranging over properties (Frege's 'concepts'). Indeed, Frege is rightly described as the person who discovered modern quantification theory.

Frege's purpose in carrying out proofs within a formal system with clear axioms and rules of inference was, in the first instance, to demonstrate the truth of logicism. Logicism is the doctrine that arithmetic is a branch of pure logic, a view which had been contested by the philosopher Immanuel Kant. For logicism to be true, each arithmetical theorem must be derivable from axioms that are purely logical, together with rules of inference that are justified by reason alone, without appeal to psychological intuition. By couching the proofs of the axioms of arithmetic within a formal system, Frege believed he could ensure that no nonlogical step entered unnoticed into his proofs. Frege's great work, *The Foundations of Arithmetic*, was devoted to an extended philosophical introduction and defense of the thesis of logicism, as well as an informal sketch of the derivation of the axioms of arithmetic; part II of the *Grundgesetze* contains his actual attempted proofs thereof.

Unfortunately, the theory that Frege took to be logic turned out to be inconsistent, a discovery made by the philosopher and logician Bertrand Russell, and communicated to Frege in June of 1902. The problem lay with the fifth axiom of Frege's theory in the *Grundgesetze*. This axiom entailed that to every predicate, there corresponded the set of all and only those things of which that predicate was true. This principle leads directly to Russell's paradox, which involves the predicate 'is a set that does not contain itself'. Frege's Axiom 5 entails that there exists a set of all and only those things of which this predicate is true. However, the question arises whether this set is a member of itself, and either answer leads to a contradiction. Over the years, various repairs and modifications of Frege's system have been proposed, but the general consensus is that no repair of Frege's system yields a theory the axioms of which are all plausibly logical.

Despite the failure of logicism, Frege's technical writings have borne rich fruit. Frege's *Begriffsschrift*, the formal language in which his theories are couched, despite some notational complexities, clearly reflects the quantifier-variable reasoning that has now become so familiar. In the language, *n*th order universal quantifiers are the quantificational primitives, and they also function as devices for binding variables, as in standard contemporary formulations of the language of the predicate calculus. So the idea of a quantifier as a device for binding variables, so influential in linguistic theory and logic, is clear in Frege's work. But so too is the understanding of quantifiers as denoting 'second-level functions', in particular functions from properties to truth-values. This treatment of quantifiers has proven of great value in the semantics of natural language. Furthermore, his sophisticated discussion of the semantics of his formal system in part I of *Die Grundgesetze der Arithmetic* laid some of the groundwork for formal semantics, both in logic and linguistics.

Frege's defense of logicism raises important issues in epistemology. According to logicism, arithmetic is analytic, its propositions derivable from logic and definitions alone. Adherents to the positivist school of philosophy, such as Rudolf Carnap, took logicism to entail that arithmetic was devoid of substantive content. However, this is not how Frege construed logicism. In *The Foundations of Arithmetic*, Frege argues that analytic judgments can lead to substantive extensions of our knowledge. But analytic judgments, by their

nature, seem empty of content. So, one might think that any consequence of them would lack substantive content. The position that logic together with definitions can lead to substantive extensions of our knowledge is a bold anti-empiricist position in epistemology, one that has attracted considerable attention in recent years.

In addition to his work devoted to defending logicism, Frege made many other significant contributions to philosophy and cognitive science. In his 1892 paper 'On sense and reference', the defining paper of the analytic tradition in philosophy, Frege introduced the distinction between sense and reference which was to have a profound influence on future discussions of linguistic and mental intentionality. Other important notions, such as the discovery of the negation test for semantic presuppositions, appear for the first time in this work. In his 1918 paper, 'The thought', Frege had a sophisticated discussion of the content of indexical expressions such as 'I' which prefigured later work in the philosophy of language and mind on the ineliminable nature of indexicality.

Frege's nontechnical writings are best known for his distinction between sense [*Sinn*] and reference [*Bedeutung*]. In 'On sense and reference', Frege argues that two names that refer to the same object in the world, such as 'Mark Twain' and 'Samuel Clemens', may nevertheless present that object in different manners. So, while it is uninformative to be told that Mark Twain is Mark Twain, it can be quite informative to be told that Mark Twain is Samuel Clemens; similarly, one might believe that Mark Twain was a writer without believing that Samuel Clemens was a writer. According to Frege, this is because the names 'Mark Twain' and 'Samuel Clemens', while having the same reference, are associated with different senses. So, we may think of a term such as 'Mark Twain' as having a reference, namely the famous author, as well as a sense, such as that expressed by 'The author of *Huckleberry Finn*'.

The distinction between sense and reference also helped Frege address a problem about how to individuate mental states. Consider verbs such as 'believes', 'fears', and 'expects', which take sentences (sometimes headed by 'that') as complements. Such verbs appear to express relations between persons and the contents of their sentential complements. So 'John believes that Mark Twain is Samuel Clemens' expresses a relation between John and the content of the sentence 'Mark Twain is Samuel Clemens'. If the only notion of the meaning of a lexical item were the entity in the world it denoted, then it would seem that believing that Mark Twain is Mark Twain is the very same mental state as believing that Mark Twain is Samuel Clemens. For if the only notion of the meaning of a lexical item were its reference, then the content of the sentence 'Mark Twain is Mark Twain' would then be the same as the content of the sentence 'Mark Twain is Samuel Clemens'. However, it seems that someone may believe that Mark Twain is Mark Twain, without thereby believing that Mark Twain is Samuel Clemens. This suggests that the mental state of believing that Mark Twain is Samuel Clemens is distinct from the mental state of believing that Mark Twain is Mark Twain. If, following Frege, we take the content of a sentence (as used in a particular context) to be determined by the senses (rather than the references) of the terms that compose it, then we can explain the intuition that the belief that Mark Twain is Mark Twain is a distinct mental state from the belief that Mark Twain is Samuel Clemens. For on Frege's theory, the belief that Mark Twain is Mark Twain has a different content than the belief that Mark Twain is Samuel Clemens, since 'Mark Twain' and 'Samuel Clemens' have different senses.

Frege's distinction between sense and reference has had a profound influence on subsequent research into linguistic meaning. The idea that a theory of meaning for a natural language needs to have recourse to a more 'fine-grained' notion than that of reference proved to have wide applicability. For example, though Frege himself never took seriously talk about the modal properties of thoughts, later research by Rudolf Carnap and Alonzo Church (among others) into the semantics of modal discourse was influenced by Frege's discussions, since here too a more fine-grained notion than mere reference seems required to address the complexity of the phenomenon. Indeed, the term 'Fregean' is now used in the philosophy of language to refer to any position according to which names are semantically associated with something in addition to their referents.

Frege's view of the nature of senses (and hence of thoughts – the senses of sentences) was that they were abstract, mind-independent entities, much like numbers or sets. In entertaining a thought, one is in a relation to one of these abstract entities. This 'Platonist' conception of thoughts played a role in Frege's arguments against psychologism, the doctrine that thoughts were composed out of subjective, mind-dependent ideas. Frege was led to his Platonist conception of thoughts by his insistence on the publicity of content. If psychologism were true, then thoughts would not be communicable or shareable. Yet two people seem to be able to

share the same thought. Furthermore, by uttering a sentence, one seems to give voice to the thought it expresses. But if thoughts were composed out of subjective ideas, then a hearer could not grasp the thought someone expressed by her utterance of a sentence. The tension Frege recognized between psychologistic accounts of the contents of mental states and the apparent publicity of these contents remains an important issue for cognitive science today.

Further Reading

Boolos G (1998) *Logic, Logic, and Logic*, edited by R Jeffrey. Cambridge, MA: Harvard University Press.

Demopoulos W (ed.) (1995) *Frege's Philosophy of Mathematics*. Cambridge, MA: Harvard University Press.

Dummett M (1973) *Frege: Philosophy of Language*. London: Duckworth.

Dummett M (1981) *The Interpretation of Frege's Philosophy*. Cambridge, MA: Harvard University Press.

Dummett M (1991) *Frege: Philosophy of Mathematics*. London: Duckworth.

Frege G (1952) *Translations from the Philosophical Writings of Gottlob Frege*, edited by P Geach and M Black. Oxford: Blackwell Press.

Frege G (1966) *Grundgesetze der Arithmetik*. Hildesheim: Georg Olms.

Frege G (1980) *The Foundations of Arithmetic*, translated by JL Austin. Evanston: Northwestern University Press.

Frege G (1988) *Begriffsschrift und andere Aufsaetze*. Hildesheim: Georg Olms.

Wright C (1983) *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

Functionalism

Intermediate article

David Braddon-Mitchell, University of Sydney, Sydney, New South Wales, Australia

CONTENTS

Functionalism is the view in the philosophy of mind that that what makes some state of some system a mental state is what that state does in the system.

INTRODUCTION

Functionalism is an influential doctrine in the philosophy of mind, which in various ways provided the philosophical impetus for early work in artificial intelligence and cognitive science. This article gives a characterization of the various forms of functionalism, where they come from and what their strengths and difficulties are. It concludes with a brief discussion of the connections between functionalism and cognitive science.

WHAT IS FUNCTIONALISM?

Functionalism about the mind is the view that what makes some state of some system a mental state is what that state does in the system – not, for example, what it is made of or any other purely intrinsic features of the state. This kind of view is familiar in other contexts. What makes something a thermostat is what it does: it turns heating or cooling on and off in a way that keeps temperature stable. The thermostat is multiply realizable: while an explanation of how a particular thermostat works may appeal to the differential coefficient of expansion between copper and iron strips welded together, there are other ways of building a thermostat (e.g. with an electric charge passed through a mercury column) that would produce something intrinsically different, but functionally the same. The mercury-column device is still a thermostat, despite its intrinsic difference from the bimetallic strip, because of the role it plays in the heating or cooling system as a whole.

Functionalism is compatible with dualism, since logically there might be nonphysical systems that count as minds in virtue of what they do rather than what they are made of. However, in this article we will concentrate on cases where there is something about what a physical system and its physical states do in virtue of which it has mental states. A functionalist about the mind thus looks for a specification of what a system comprising a body and a brain does in virtue of which it counts as having mental states.

Multiple realizibility is a necessary condition shared by anything that can be called functionalist. Once we have specified what a system must do so as to have mental states, realizing the system with different methods or materials will, according to functionalism, make no difference.

Most functionalists agree that the way to characterize what the state of a system does is by specifying the relations between its inputs, outputs and internal roles. Thus, for example, a belief that 'ice cream is near' might be characterized in part as the state of a system which is typically caused by ice cream, makes the system (a person) move towards ice cream in combination with a desire for ice cream, and so on.

There is, however, considerable disagreement between functionalists on a range of issues. These include: how we may discover the theory of inputs, outputs and internal roles that determines whether something is a mind; what kinds of inputs and outputs count; and even whether what something does is a matter of its causal influence on the world.

VARIETIES OF FUNCTIONALISM

At the highest level of abstraction there is a distinction to be drawn between 'analytic' (or 'commonsense') functionalism and 'empirical' functionalism.

Analytic Functionalism

According to analytic functionalism, we already possess the right theory of the inputs, outputs

and internal roles, in virtue of our mastery of mental-state language. We are able to attribute mental states to others from observations of their behaviors and the events happening to them, together with some very thin assumptions about how they work (such as that there are internal states causing the behavior, and they are not just puppets controlled from elsewhere). Because of this mastery we are able to predict and explain behavior in mental-state terms. The theory is said to give the meaning of such terms. What belief, desire, and so on, mean depends on their place in the theory that our predictive and explanatory practices reveal that we know. We cannot write down such a theory: it is very complex and difficult. But nonetheless it is held to be tacitly known, since if it were not then we could not explain the relatively widespread agreement in mental-state attributions given similar behavioral evidence. A useful analogy is with grammar. We are said to know grammar tacitly, as this explains our grammatical competence. Nonetheless, it takes an expert linguist to attempt to write down the theory, and even then there is vast room for disagreement.

Thus, the inputs and outputs that analytic functionalism takes to be definitive of mental states (but not the internal roles) must be environmental inputs and behavioral outputs, since these inputs and outputs are all we can directly observe in everyday life. And behaviors must be understood as typical effects on environments: hailing a cab is to be understood in terms of making cabs more likely to stop, rather than, say, intentions to do so, since the environmental effects are observable but the intentions are not.

There need be no commitment in all of this to the discreteness of mental states. The theory may reveal that a person is held to be in a mental state which believes that *P*, *Q*, *R* and so on, and desires that *A*, *B*, *C* and so on. It does not follow, though, that any of these individual beliefs or desires can be identified with some physical or architectural part of the system. The only commitment is to the possibility in principle of a system undergoing some (possibly small, and possibly holistic) change which would be enough to change just one of the beliefs or desires.

Empirical Functionalism

Empirical functionalism, on the other hand, denies that we already know in any sense the right theory of inputs, outputs and internal states that gives the meaning of mental-state terms. Rather, this theory has to be discovered by empirical investigation. (Nevertheless, something like the tacit theory of analytic functionalism may be used to select the paradigmatic mental beings – other humans in our case – for something must tell us what to empirically investigate; to unleash the cognitive sciences on glaciers would be futile. See Braddon-Mitchell and Jackson, 1996, Ch. 5.)

This leaves open the question of how best to characterize in different ways the level at which to look for inputs, outputs, and internal roles. For example, empirical study might reveal that the outputs, rather than being typical effects on the environment, are classes of motor responses of the body, or neural inputs and outputs at the periphery of the central nervous system or even of the brain. This last conception of the 'right' kinds of inputs and outputs is sometimes called 'psychofunctionalism'. (Block, 1991).

Another kind of empirical functionalism is concerned less with inputs and outputs than with more or less abstract specifications of what goes on inside the brain. While such theories retain the generic specification of functionalism as being concerned with what is done, this is not understood as what the system does to the world and what the world does to it. Rather, it is understood as what is going on internally, in a sufficiently abstract sense that there are various ways of making physically different systems that count as 'doing the same things' internally.

Functionalisms of this sort will tend to specify a kind of cognitive architecture or hypothesis about internal organization (e.g. Fodor's 'language of thought' hypothesis (Fodor, 1976, 1987)) which captures what it takes to have mental states. Extreme versions may specify something as abstract as a machine table, which can in principle be mapped onto many obviously non-mental physical systems. (*See* **Language of Thought**)

HISTORY

Functionalism has two fairly distinct historical roots, and this is partly why there are such different varieties of functionalism. Analytic functionalism is essentially a response to behaviorism, allowing mental states to be the hidden causes of behavior. J. J. C. Smart's materialism (Smart, 1959) is sometimes seen as opposed to functionalism. But in fact Smart emphasizes the 'topic-neutrality' of the analysis of mental states: the meaning of mental-state terms is given in part by the roles invoked by functionalists, and indeed his work foreshadows analytic functionalism. Armstrong's causal theory of the mind (Armstrong, 1968) is near to analytic

functionalism, and David Lewis (Lewis, 1972) is a standard reference. (*See* **Behaviorism, Philosophical**)

Empirical functionalism arose as a criticism of a certain understanding of an unrestricted type–type identity theory of the mind according to which any given type of mental state is unrestrictedly identical to some type of physical state. In Putnam (1975, p. 335), something like a nonreductive functionalism is seen as opposed to reductive physicalism. Sometimes in the work of Putnam, and in the tradition that follows, the sort of similarity that is relevant is thought to be discovered by uncovering empirically what the significant abstract features of the state are. Early work in computational artificial intelligence made it seem natural that the relevant abstract state might be something like 'the program that is being run', regardless of how it is connected to the world. Combined with philosophical interest in the 'necessary *a posteriori*' arising from, for example, Saul Kripke's work (Kripke, 1980), this made it seem plausible to many that the investigation of actual paradigmatic minds might reveal hidden internal features that are essential to having minds. (*See* **Mental Causation**)

Contemporary views, including perhaps Fodor's architectural model of the mind and the very different views of dynamical systems theorists, insist that there are internal, architectural features that account for mentality in a way which is undetermined by actual and counterfactual behavior.

ARGUMENTS FOR FUNCTIONALISM

Functionalism is popular partly because it is consistent with physicalism, while avoiding some of the problems associated with other physicalist views. It avoids the worst problems of behaviorism – especially its failure to provide behavioral analyses of individual beliefs – while allowing that mental states are the causes of behavior. It also avoids the main problem of the unrestricted type–type psychophysical identity theory: functionalism insists that some features of a physical system – even the substance of which it is made – may be irrelevant to mentality.

Functionalism provides the best explanation for the multiple realizability of minds. Furthermore, the different kinds of functionalism give us positive accounts of what the relevant features are, together with arguments to establish their relevance. Analytic functionalism – and the kinds of empirical functionalism that refer to the so-called folk roles – also explain how we are able to attribute mental states to entities from behavioral evidence.

Apart from the positive arguments for functionalism, there is the following negative argument: dualism (for independent reasons) and the remaining physicalist theories are ruled out, leaving functionalism as the remaining contender. The strength of this argument will depend on the strength of the objections to functionalism.

PROBLEMS FOR FUNCTIONALISM

Many objections to functionalism have been raised. Some of these are worse for some varieties of functionalism than for others.

Certain traditional arguments have involved imagining functional duplicates of human minds realized in ways that appear intuitively not to count as mental. These include John Searle's 'Chinese room' (Searle, 1980) and Ned Block's 'China brain' and 'Block head' (Block, 1981).

Searle asks us to imagine that we have written out in English the program that is realized in people who understand Mandarin, and in virtue of which they understand it. It is written down in a series of books, and we suppose that a monolingual English speaker is in a closed room with the books. He first receives a text in Mandarin, and he then receives questions – also in Mandarin – designed to test his comprehension of the text. He uses the books to perform (from his perspective) purely symbolic manipulations. He performs a huge number of line-by-line calculations, running the program written down in the book very slowly in his head, in the way that a computer programmer might try to mentally check a short piece of code. This process is repeated with the comprehension questions, and finally the English speaker, following the English directions, will draw certain Mandarin characters and place them in the 'out' slot. These will be good answers to the test, and might persuade those outside the room that there is some understanding of Mandarin, but in fact he understood nothing of it. In this example, understanding stands in for mentality, but appropriate variations could be devised (e.g. the English speaker has no beliefs about the propositional content of the story).

The severity of this problem depends on the kind of functionalism: the example assumes that the right understanding of functional duplication is at the 'program' level. Analytic functionalism, which specifies inputs and outputs not symbolically but in terms of typical environmental effects, may be unaffected, since the example does not provide a functional duplicate of a Mandarin speaker at that level: the outputs are not the typical effects of

understanding Mandarin. And empirical functionalisms that stress the importance of the mental architecture can reply that although the individual in the room fails to understand, there may be a virtual machine running on his or her wetware that does.

A variation of Ned Block's 'China brain' has much the same structure: we imagine a functional duplicate of someone's brain, made by equipping everyone in China with mobile phones and with instructions to make calls to certain numbers on receiving calls from certain numbers in a pattern that duplicates the brain's neuronal firing pattern. The intuition is supposed to be that there is functional duplication with no mental duplication.

One response is to deny this 'intuition' (no one phone caller has the mental states of the original person, but nor does any one of our own neurons have a mental state: this does not mean that the whole system does not). Another response is to note that, again, analytic functionalism is not affected, unless we connect the whole system to a robot capable of having the right environmental effects, and ensuring that the states in the simulation are suitably affected by environmental influences. But then the 'intuition' becomes even less plausible.

The idea behind the 'Block head' argument is to imagine a being programmed so that its interactions with the world are indistinguishable from those of a real thinker, but in such a way that, once we knew it, we would deny that the being had mental states. The suggestion is, roughly, that this could be achieved by prerecording sensible responses to all possible eventualities in a tree-like structure that would ensure that the responses produced at any time and under any circumstances are consistent with previous behavior.

The example, originally intended as an argument against behaviorism and for psychologism, appears to be a counterexample to analytic functionalism, since empirical functionalism requires that the only correct way to duplicate mentality is to duplicate the way we do it, where the internal roles are specified by more that just what it takes to account for behavior. We do not operate in this way, and presumably we will not deny our own mentality by discovering how we in fact function.

Analytic functionalists can reply that if we are able to recognize on inspection that the way the Block head works is not enough to ensure mentality, then it must be something that we know *a priori*. It is not an empirical claim by Block that such creatures do not have mental states. Block's example simply makes explicit what we tacitly know, and in that case it must already be part of the analytic-functionalist theory of the mind that a mental system cannot be organized in that way. The lesson for analytic functionalism is that there are some constraints on internal structure that are stronger than merely how that structure must affect and be affected by the environment.

Qualia and Consciousness

Some forms of functionalism have difficulty accounting for phenomenal experience – the 'raw feelings' of color, sound and touch, and other more subtle qualitative phenomena. (*See* **Qualia**)

An important objection is the so-called inverted spectrum problem (Shoemaker, 1975). If there were two individuals whose color experience was inverted – one saw yellow when the other saw blue, and so forth – but whose color language and other behavioral effects were the same, then they would be functionally indistinguishable. The invert, who saw yellow as blue, would nonetheless use the word 'yellow' for the blue experiences he would have of the yellow things. Thus he would agree that the sun is yellow, that it is the same color as the Bondi sand, and so forth.

This seems problematic for analytic functionalism because the states that mediate between inputs and outputs in this case play identical mediating roles, and hence must be classified as identical in a functional classification. Yet by hypothesis they are distinct qua qualitative experiences – and hence different mentally. This seems to refute the idea that mental classifications are functional classifications.

One response is to 'bite the bullet' and insist that this is a metaphysically impossible scenario. It appears to be conceivable because we appear to have a grasp of the essential intrinsic nature of color experience, but this is an illusion. A softer response is to acknowledge that a partial inversion might be possible: perhaps all the color language might be the same but other functional effects might be different. These might include, for example, the effects of colors on hunger, anger, and intermodal judgments (how similar colors are to certain sounds or sensations). In such a case there would be functional differences between the so-called invert and a normal individual. This partial inversion is genuinely conceivable, and it generates the illusion of conceivability of total inversion. Many, however, still claim to find total systematic inversion conceivable even after distinguishing it from partial inversion.

Empirical functionalism can avoid this problem more easily: color experiences may be identified

with particular neural states that typically play certain roles. Inversion could be explained by the state that typically plays the 'yellow' role taking on, in a rare case, the 'blue' role, and vice versa. The problem lies in justifying the claim that the state has that experiential quality: this will be untestable, since testing will be by the kinds of functional roles specified by analytic functionalism (for example, verbal outputs if we ask people about their phenomenal experience).

The Knowledge Argument and Conceivability Arguments

Frank Jackson's 'knowledge argument' against physicalism (Jackson, 1982) can also be deployed either as an argument against physicalism, and hence against physicalist functionalism, or (with a minor modification) against functionalism in general.

This style of argument proceeds via a thought experiment in which we are asked to imagine someone who knows all the physical facts about color vision, experience and so forth (in the variation, read 'functional facts' for 'physical facts') but has never actually had an experience of, say, redness. She has not been in the physical (or functional) state that is purportedly the one of experiencing redness, but knows everything about it. We then suppose that she actually sees something red. Intuitively it seems that she now learns a fact about that state which she did not know before: what it is like to see something red. But she knew all the physical (or functional) facts before, so if she has learned something new, then it must be some nonphysical (or nonfunctional) state fact that she learns. Therefore, physicalism (or functionalism) is false.

There is a considerable, and growing, literature on this argument. Most responses try to show either that the person would not learn any new fact (even if she acquires new abilities or know-how) or that if it is a fact, she must have known it all along in some way. Most physicalists and functionalists agree that something is wrong with the argument, but there is little consensus on what. Currently a popular idea is that conceivability and thought experiments are no guide to metaphysical possibility.

There are themes in common between this argument and the conceivability argument of David Chalmers (Chalmers, 1996), which has reinvigorated the discussion. Chalmers asks us to imagine functional duplicates of ourselves, which are in fact qualitatively dead. Allowing conceivability to be a guide to possibility, we discover that there could be (perhaps only if the laws of nature were different) functional duplicates that entirely lacked qualia. If so, then qualia do not supervene on functional states in the way required by functionalism. The legitimacy of the conceivability claim is bolstered by arguments that we fix reference on qualitative states in the first place via acquaintance with their phenomenal nature.

Again, replies often depend on separating conceivability and possibility. One reply is as follows. Even allowing conceivability to be a good guide to possibility, it is not clear what we are conceiving, for it may be part of our concept of experience that the nature of experience depends on what we discover about metaphysics. If we are in fact living in a dualistic world, then our concept connects us to something whose essence we grasp directly through experience; but if we are in fact living in a physicalistic world then the thing picked out by the concept of qualia is functional, and the thought experiment fails. The intuitive appeal of the experiment would be explained by our proper acknowledgement of our uncertainty about our world, and thus of what our concept of qualia picks out.

Chauvinism Arguments

'Chauvinism' arguments are directed against empirical functionalism. The idea is that in specifying the functional roles required for mental states by reference to an examination of how we are in fact built, we thereby rule out the mentality of any beings who are built differently internally. If there were apparently intelligent life elsewhere, with an internal architecture different from the most explanatorily salient features of our own, it would not count as mental – and this, it is argued, is unacceptably chauvinistic. The point can be made more dramatic by supposing that the Alpha Centaurans made contact with Earth before the development of the cognitive sciences. If the subsequent philosophical consensus in philosophy of mind conferences (attended by both species) was for empirical functionalism, then it would be features in common between humans and Centaurans that would be regarded as definitive. So the account of the mental under empirical functionalism depends on accidental features of the available paradigms at the time of investigation. One response to this argument is to look for empirical features which are very abstract and can be expected to obtain in all nomologically possible beings that behave intelligently.

FUNCTIONALISM AND COGNITIVE SCIENCE

Functionalism arose at the same time as early work in artificial intelligence, one of the disciplines that helped form cognitive science. Functionalism seemed to justify the claims of artificial intelligence researchers that the creation of genuine artificial intelligence was possible. Analytic functionalism underwrites cognitive science inasmuch as it justifies the claim that a full cognitive account of how we work, if appropriately embodied so as to enable causal connection with the world, would be sufficient for mentality.

Empirical functionalism makes the further claim that duplicating the most salient cognitive features of how we in fact work is also necessary for mentality.

The difference is in how philosophically deflationary an understanding of completed cognitive science is meant to be. Analytic functionalism says that philosophy is the tool for an understanding of what is required for mentality; whereas cognitive science tells us about the nature of our own version of mentality, gives us explanations about how our own mentality works, and gives us tools to potentially duplicate our mentality.

Empirical functionalism assigns to cognitive science the role of establishing the essential nature of the mental itself. According to this doctrine, what mental states are essentially depends on how they actually are in us. Some versions of empirical functionalism which concentrate on 'architectural' features – the so-called program that is running on the brain – assign to cognitive science an autonomous role studying the nature of the mental at a level less abstract than behaviorism or analytic functionalism, but more abstract than various lower-level ways of studying the neural chemical or physical nature of the brain. Thus, according to most kinds of empirical functionalism, cognitive science will be the final arbiter of whether, should we succeed in making contact with extraterrestrials, we would be communicating with other thinkers, or merely with zombies produced by a freak of alien evolution.

References

Armstrong (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block N (1981) Psychologism and behaviorism. *Philosophical Review* **90**: 5–43.

Block N (1991) Troubles with functionalism. In: Rosenthal D (ed.) *The Nature of Mind*, pp. 211–228. Oxford: Oxford University Press.

Braddon-Mitchell D and Jackson F (1996) *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Chalmers D (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Fodor J (1976) *The Language of Thought*. Brighton, UK: Harvester Press.

Fodor J (1987) Why there still has to be a language of thought. In: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, pp. 135–154. Cambridge, MA: MIT Press.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **34**(135): 127–136.

Kripke S (1980) *Naming and Neccessity*. Oxford: Blackwell.

Lewis D (1972) Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* **50**(3): 249–258.

Putnam H (1975) *Mind Language and Reality: Collected Papers*, vol. I. Cambridge: Cambridge University Press.

Searle J (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**: 417–424.

Shoemaker S (1975) Functionalism and qualia. *Philosophical Studies* **27**(5): 292–315.

Smart JJC (1959) Sensations and brain processes. *Philosophical Review* **68**: 141–156.

Further Reading

Braddon-Mitchell D and Jackson F (1996) *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Campbell K (1984) *Body and Mind*, 2nd edn. Notre Dame, IN: University of Notre Dame Press.

Sterelny K (1990) *The Representational Theory of Mind*. Oxford: Blackwell.

Implicit and Explicit Representation

Intermediate article

David Kirsh, University of California, San Diego, California, USA

CONTENTS

The degree to which information is encoded explicitly in a representation is related to the computational cost of recovering or using the information. Knowledge that is implicit in a system need not be represented at all, even implicitly, if the cost of recovering it is prohibitive.

INTRODUCTION

During the brief history of cognitive science, disputes have often arisen over how explicitly certain types of information (or knowledge) are represented in human cognitive systems, and what it means for information to be explicitly rather than implicitly represented in a system. Every few years it seems that new ways are discovered to build information into architecture, internal dynamics, and agent–environment interaction. Information that was stored in 'software' becomes integrated into 'hardware' in the next generation of systems, and then becomes integrated into a new 'architectural' design in the following generation. This seems to hold whether the software is made out of computer programming languages or malleable neural connections, whether the hardware is silicon or organic, and whether the architecture is a computer system or an anatomical plan that only changes across generations. When information is no longer encoded in distinct data structures with definite location and form, should we say that the information is still there, but represented in a more implicit form?

Questions about how information may be embedded in systems, how it may exercise a causal influence over processes, without being easily identifiable with a recognizable state, structure or process, make problematic the analysis of representation, and especially the analysis of what it means for information to be explicitly, as opposed to implicitly, represented in a cognitive system. There is a temptation to simply dismiss the matter as being of purely philosophical, or even semantic, interest. But the questions remain important because any hope of understanding how information enters into causal accounts of cognition must eventually explain the relation between the information attributed to a system, and appealed to in explanatory models, and the underlying states, structures or processes that serve as the vehicles for that information, and which are closer to the level of purely causal or neurophysiological explanation. Since there are many explanatory models in cognitive science that refer to information that is only implicitly present, it is important to clarify what implicit representation means and how it enters into underlying causal interactions.

EXPLICITNESS AND SYMBOL-PROCESSING MODELS

According to the symbol-processing view, all higher-order cognition, and most lower-order cognition too, is a computational process in which syntactically structured representations – such as sentences in an internal language of thought (Fodor, 1975; Pylyshyn, 1984), algebraic or graph structures (Chomsky, 1980; Minsky, 1975), or matrices of numbers (Marr, 1982) – are systematically transformed in a rule-driven or algorithmic manner. To understand, and therefore explain, a cognitive process, it is necessary to track the trajectory of informational states that the host system follows as it moves towards an explicit answer to a computational problem: for example, determining the meaning of an utterance, the implication of a previous thought, or the shape and visual appearance of an external object.

Since a cognitive system, on this view, is a mechanism for applying rules to structured

representations, it is natural to regard information as explicitly encoded if the structured state, process or form – the symbol structure – representing that information can be interpreted according to a well-behaved theory of content, such as a truth or model theory. We can then point to any given structured representation and say 'that form explicitly encodes this content'.

Although the data structures and representations involved in the various symbol-processing models (lattices, matrices, graph structures, extended first-order predicate calculus) go well beyond the structures we typically see in everyday life, this idea of 'implicit' and 'explicit' representation remains close to the everyday meaning of the terms used when discussing natural language. In ordinary parlance, we regard a fact to be 'explicitly' stated if it is expressed literally and unambiguously in a well-formed sequence of words, a sentence. Something that must be inferred from the sentence – because, for instance, it is presupposed, or because it is a consequence of the meaning of the words – is not explicitly stated; it is implicit in the meaning and context of utterance. Thus, when someone says 'when did you stop being a bachelor?' we take them to be explicitly asking a question about a date or time, but implicitly asserting that we no longer are a bachelor (the major presupposition), that we now have a wife (a consequence of the meaning of 'bachelor'), that we know when our marriage took place (a presupposition of the 'when' question), and so on.

On this account, information is 'explicitly' encoded if it can be read directly off a sentence without more inference than is required to understand the meaning of the words and their structure; while information is 'implicitly' encoded if additional inference or semantic processing is needed to recover it.

Using this intuitive analysis, Dienes and Perner (1999) have presented a theory of implicit and explicit knowledge that attempts to integrate and relate the divergent uses of the implicit–explicit distinction in different research areas, such as implicit and explicit memory, blindsight, automatic and controlled action, and development. They offer an explanation of why discussions of explicit and implicit knowledge are so often tied to concepts like consciousness, volitional control, and verbalizability.

Any theory about the nature of explicit and implicit representation based on our intuitions about what is explicit and implicit in natural language can be generalized to other declarative representations discussed in symbol-manipulation models: matrices, connected graphs, vectors, etc. But our intuitions about explicitness, even in natural language, are not complete. For instance, the sentence 'police police police police police' is grammatical and unambiguous. This sentence may be paraphrased as 'policemen who are policed by policemen also police policemen'. Since it has a unique syntactic and semantic identity it should qualify as an explicit representation. One need only understand the meaning of the words and their structure. Yet few people can actually recover this meaning because too much computation is involved in determining which sense of 'police' is being used in each position. Because of the computational complexity of interpreting the meaning of the sentence, the meaning is not easily recovered, it is not on the surface, and cannot be directly accessed. Consequently, we have conflicting intuitions about whether to say its meaning is explicitly encoded. On the one hand, it satisfies the basic truth-theory intuition 'here is the representation, there is the meaning', since it is well formed and unambiguous, so it is explicit on that criterion. On the other hand, few people can actually recover the information in the representation, so it ought be implicit on that criterion. Thus it seems that the intuitive theory is incomplete.

Such concerns about the computational processes involved in understanding the meaning of a representation suggest that a theory of implicit and explicit representation should be based more on the way representation and computation interrelate.

EXPLICITNESS AS A COMPUTATIONAL PROPERTY OF REPRESENTATIONS

One such theory involves measuring the computational effort required to extract, use, or interpret the information encoded in a representation (Kirsh, 1991). The computational complexity of the process of interpretation determines where on the continuum of explicit to implicit a given representation lies. If the interpretative process implements a constant-time algorithm, or extracts the content quickly and without substantial involvement of the rest of the cognitive system, then the information it extracts is directly available and hence explicitly encoded. For instance, the numeral '5' encodes the number 5 in English more explicitly than $\sqrt[5]{3125}$, even though both designate 5, because the value 5 can be read off directly from '5' by English speakers without performing lengthy computation.

It is assumed here that a representation is a well-defined state, structure or process, in a causal

system; that it encodes a specifiable informational content that can be harnessed by the causal system of which it is a part; and that it is possible to use techniques of computational complexity theory to measure the computational effort involved in recovering the information, providing we have a theory of the processes that use or extract that content.

So to determine how explicitly or implicitly a piece of information is represented, we need a substantive theory about the functioning of the different parts of a cognitive system. This must include how the system identifies the state, structure or process carrying the information, and how it exploits the information.

A plausible consequence of this approach is that individual capacities for memory, learning and other cognitive skills can affect how explicit a representation is. For one person, a certain representation may be identified and grasped directly. For another, the representation cannot be grasped without substantial computation. This conforms to both intuition and science. We now know a good deal about the computational costs associated with different neural-network methods of identifying and individuating states (structures or processes). It is known that there are computational trade-offs between the number of perceptron-like connections reaching out into a region where neural states encode information (the spatial complexity of the identifying process), and the amount of later processing required to correctly identify the state (the time complexity of the identifying process) (Minsky and Papert, 1969). Consequently, a neural state that in one person may be immediately identified – because identified in a highly parallel manner – and so meet the identifiability condition of explicitness, for another person may not be immediately identifiable – because, for example, this person has not yet developed efficient recognition methods, and so will take much longer to identify the information-bearing state and access the information contained therein. For a lengthy discussion of these points see Kirsh (1991).

IMPLICATIONS

By treating explicitness as a computational property of a representation, we have a method for deciding the point at which we can say that information or knowledge is so deeply embedded in hardware, architecture or agent–environment interactions that it is no longer a causally active agency and no longer represented, even implicitly. It is the point where the cost to the cognitive system of recovering the information would be so great that there are no connections or accessing procedures that can reliably make use of the information. Since a representation is a mechanism for reliably carrying information across space or time, it is best to say that the information is so implicitly built into the cognitive system as to no longer be represented in it. At the other end of the continuum, information that is encoded in well-defined states, structures or processes, which the cognitive system can immediately identify and use, is explicit.

A further consequence of tying the implicit–explicit distinction to computation is that we have a method for distinguishing different types of explanation in cognitive science. Not all explanations of behavior and design are mechanistic. We accept this in other design sciences, where discussions of design rationale and historical evolution help us to understand why a particular design is a good one. The same should hold true when we discuss the 'implicit theory of the world' built into a system or process.

Although sometimes such claims point the way to a computational explanation, we should not assume that every implicit theory implies that there are counterpart representations in the system. For instance, a vision module designed to extract a three-dimensional shape from two stereoscopic images works rapidly if it is equipped with an algorithm that differentiates. Such an algorithm will work if the assumption about the world that objects change in shape smoothly and continuously is true. Smoothness is a 'success condition' of the algorithm, and any designer who wishes to determine the algorithm's reliability will need to know how often the assumption is correct. But the algorithm may not implicitly represent the assumption. The assumption will not be recoverable by the system itself if the representational vocabulary of early vision does not include terms such as 'smoothness'. The assumption of smoothness serves as a powerful constraint on the design space of useful algorithms and representations; but it is not implicitly represented.

Thus, theoretical assumptions, such as smoothness, generativity, conformity to a truth theory, adherence to a formal linguistic theory, and so on, may guide the design of algorithms or the architecture of a system; they may figure in discussions of implicit knowledge or of the evolutionary fitness of a creature; but they need not figure directly in discussions of the computational mechanism. Depending on the cost of using those assumptions directly, they may be best understood as nonmechanistic explanations.

References

Chomsky N (1980) *Rules and Representations*. Oxford: Blackwell.

Dienes Z and Perner J (1999) A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* **22**(5): 735–755.

Fodor JA (1975) *The Language of Thought*. New York, NY: Thomas Crowell.

Kirsh D (1991) When is information explicitly represented? In: Hanson PP (ed.) *Information, Language, and Cognition*, pp. 340–365. New York, NY: Oxford University Press.

Marr D (1982) *Vision*. New York, NY: W.H. Freeman.

Minsky M (1975) A framework for representing knowledge. In: Winston PH (ed.) *The Psychology of Computer Vision*, pp. 211–277. New York, NY: McGraw-Hill.

Minsky M and Papert S (1969) *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Further Reading

Clark A (1993) *Associative Engines*. Cambridge, MA: Bradford Books.

Cummins R (1986) Inexplicit representation. In: Brand M and Harnish R (eds) *The Representation of Knowledge and Belief*. Tucson, AZ: University of Arizona Press.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Evans G (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Fodor JA (1981) *Representations*. Cambridge, MA: MIT Press.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. In: Pinker S and Mehler J (eds) *Connections and Symbols*, pp. 3–71. Cambridge, MA: MIT Press.

Goodman N (1976) *Languages of Art*, 2nd edn. Indianapolis, IN: Hackett.

Palmer SE (1978) Fundamental aspects of cognitive representation. In: Rosch E and Lloyd B (eds) *Cognition and Categorization*, pp. 259–303. Mahwah, NJ: Erlbaum.

Implicit Cognition

Intermediate article

Eyal M Reingold, University of Toronto, Ontario, Canada
Colleen A Ray, University of Toronto, Ontario, Canada

CONTENTS

'Implicit cognition' refers to unconscious influences reflecting perception, memory, and learning, without subjective phenomenal awareness.

TERMINOLOGY AND DEFINITIONS

Since the 1970s, the 'implicit–explicit' distinction has become the dominant terminology under which experimental cognitive psychology has investigated unconscious influences on behavior and thought. Historically, the relation between consciousness and cognition has been one of the most controversial areas of investigation in psychology. This controversy can be traced back to the association of the conscious–unconscious distinction with psychoanalytic theory. Many experimental psychologists have regarded the unconscious as an unsuitable topic for scientific inquiry. But related phenomena have been empirically investigated under a variety of alternative terminologies. Terms such as 'incidental', 'pre-attentive', 'inaccessible' and 'covert' were used to avoid the unwanted associations of the terms 'unconscious' or 'unaware'. The term 'implicit' is a more recent, and perhaps more successful, example of such terminological camouflage. The implicit–explicit distinction was introduced in the mid-1980s in the context of the study of unconscious or unaware memory. This shift in terminology is illustrated in Figure 1, which plots the number of abstracts in the PsycINFO database from 1985 to 1999 that employed the term 'implicit' to refer to unconscious perceptual or cognitive processes. Figure 1 also displays the number of abstracts referring specifically to implicit memory. The sharp increase in the popularity of this terminology is particularly apparent in the area of memory research.

Given the dominance of the 'implicit–explicit' terminology, it is instructive to consider how these terms are typically defined. For example, Graf and Schacter (1985, p. 501) stated that 'implicit memory is revealed when performance on a task is facilitated in the absence of conscious recollection; explicit memory is revealed when performance on a task requires conscious recollection of previous experiences'. More recently the distinction has been extended to the study of conscious versus unconscious perception, as exemplified in the following definition by Kihlstrom *et al*. (1992, p. 22): 'Explicit perception refers to the person's conscious perception of some object or event in the current stimulus environment. … By contrast, implicit perception is demonstrated by any change in experience, thought or action that is attributable to some event in the current stimulus field, even in the absence of conscious perception of that event'. Similar definitions have also been proposed for implicit and explicit knowledge. As described by Berry and Dienes (1993, p. 2), 'explicit knowledge is said to be accessible to consciousness, and can be communicated or demonstrated on demand, whereas implicit knowledge is said to be less accessible to consciousness, and cannot be easily communicated or demonstrated on demand'. Finally, in reviewing the implicit learning literature, Shanks and St John (1994, p. 368) pointed out that 'different authors have used a variety of definitions to capture the fine detail of the explicit–implicit learning distinction, but the key factor is the idea that implicit learning occurs without concurrent awareness of what is being learned, and represents a separate system from that which operates in more typical learning situations, where learning does proceed with concurrent awareness (i.e. explicitly)'. Thus, with respect to recollection, perception, knowledge, and learning, implicit processes are invariably defined by the absence of consciousness or awareness. Given this observation, it is unclear what is gained by the 'implicit–explicit' terminology. Such proliferation of terminology is problematic in that it obscures the historical link between related ideas and findings.

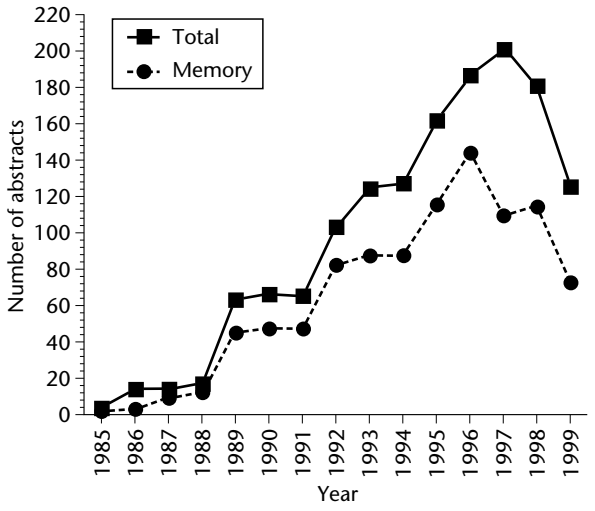


Figure 1. The number of abstracts in the PsycINFO database from 1985 to 1999 that employed the term 'implicit' to refer to unconscious perceptual or cognitive processes. The dotted line shows the number of abstracts referring to implicit memory.

THE DISSOCIATION PARADIGM

Regardless of the terminology used, most studies investigating perception, learning or memory without awareness have employed the 'dissociation' paradigm. This paradigm establishes three prerequisites for demonstrating unconscious cognition:

1. A valid measure C of cognitive information available to consciousness must be selected, and compared with another measure X of cognitive processing.
2. $C = 0$. The measure of conscious awareness indicates null sensitivity, or null awareness.
3. $X > 0$. The second measure of cognitive processing must be shown to have greater than zero sensitivity.

To illustrate the dissociation paradigm, consider the following purported demonstrations of perception and memory without awareness. In all of these examples, the measure of conscious awareness C, was 'claimed awareness' thus $C = 0$ if participants reported an absence of subjective phenomenal awareness.

Perception in the absence of claimed awareness is a very robust phenomenon that can be used as a classroom demonstration. Consider the famous experiments by Sidis (1898). In these experiments, Sidis showed each observer a card containing a single printed digit or letter. 'The subject was placed at such a distance from the card that the character shown was far out of his range of vision. He saw nothing but a dim, blurred spot or dot. … The subjects often complained that they could not see anything at all; that even the black, blurred, dim spot often disappeared from their field of vision'. When Sidis asked the subjects to name the character on a card, their responses were correct considerably more often than would be expected on the basis of purely random guessing, even though many subjects expressed the belief 'that they might as well shut their eyes and guess'. On the basis of these and similar findings, Sidis concluded that his experiments indicated 'the presence within us of a secondary subwaking self that perceives things which the primary waking self is unable to get at'. Thus, in Sidis' experiments, a 'forced choice' identification measure demonstrated the availability of perceptual information concerning stimulus identity (i.e. $X > 0$), whereas subjective report indicated null awareness for such information (i.e. $C = 0$).

An even more dramatic example of perception in the absence of subjective confidence was obtained by studying brain injury patients with lesions in their visual cortex that result in scotomas (blind regions) in their visual field. Case studies of such patients document that they perform perceptual discriminations (e.g. detection, location, orientation) concerning stimuli presented in their blind field at a level well above chance (i.e. $X > 0$), while at the same time claiming to be purely guessing (i.e. $C = 0$). This phenomenon has been called blindsight. For example, Weiskrantz *et al*. (1974, p. 721) described the verbal report of one such patient as follows: 'he was at a loss for words to describe any conscious perception, and repeatedly stressed that he saw nothing at all in the sense of "seeing", and that he was merely guessing'. Thus, just like the participants in Sidis' experiments, blindsight patients report an absence of subjective phenomenal awareness of perceiving, coupled with indirect behavioral evidence of perceiving (i.e. better than chance perceptual discrimination).

When participants' subjective report is used to index their phenomenal awareness of remembering at the time of retrieval, there is ample evidence for unconscious or implicit influences of previously encoded information on cognition and behavior. This was illustrated in the work of one of the founders of the modern empirical study of human memory, Hermann Ebbinghaus. Ebbinghaus (1885) memorized lists of nonsense syllables, and then relearned them at a later time. He quantified the saving in relearning (i.e. the reduction in time or in the number of repetitions required for a perfect recall of the list) as an indirect measure of retention. This paradigm came to be known as the saving method. Ebbinghaus also varied the 'retention interval', that is, the time interval between the

original learning of the list (the study phase) and the relearning of the list (the test phase). Being a subject in his own experiments, Ebbinghaus was able to report that following long retention intervals he observed savings in the relearning of items (i.e. $X > 0$) for which he had no conscious awareness of having studied before (i.e. $C = 0$). In fact, in developing the saving method, Ebbinghaus intended to measure, not only recollection accompanied by the subjective experience of remembering, but also the retention of experiences that are not accessible to introspection, but which nevertheless influence behavior. He stated that 'most of these experiences remain concealed from consciousness and yet produce an effect which is significant and which authenticates their previous experience'. Ebbinghaus advocated a much broader definition of memory than many subsequent memory researchers and lay people who view memory as necessarily involving the subjective experience of remembering.

As Figure 1 shows, there has been a renewed interest in the study of retention in the absence of the subjective phenomenal awareness of remembering. Currently, implicit memory is one of the most actively studied areas of human memory. This is largely due to the development of numerous indirect or implicit memory tasks. In all of these tasks, participants are not instructed to refer back to the original study phase; rather, the experimenter tries to disguise the link between the study and test phases of the experiment. A typical example of an implicit memory task is word stem completion. In this task, participants are given word stems (e.g. *bra_ _*) and are instructed to complete them with the first word that comes to mind. In an earlier phase of the experiment, participants are exposed to words that are potential completions for some of the stems (e.g. *brave*). Typically, stem completion is presented as a new and unrelated task, in order to disguise the link with the earlier study phase of the experiment. Memory for words presented in the study phase is measured as an increase in the tendency to produce these words (as opposed to other possible completions such as *brain*, *braid*, *brake* or *brass*) as responses to the stems. This effect is referred to as 'priming'. Performance on implicit or indirect memory tasks is often compared with performance on explicit memory tasks, such as recognition and recall. In explicit or direct memory tasks, participants are instructed to refer back to the study phase and to retrieve study items.

One of the most important antecedents to the renewed interest in the study of unconscious memory was the observation of memory dissociations in amnesic patients. The amnesic syndrome provides a dramatic demonstration of retention in the absence of reported subjective phenomenal awareness of remembering. In one famous case, HM became amnesic after an operation was performed to alleviate his epilepsy attacks by bilaterally removing parts of his temporal lobes. As a result of his surgery, HM became severely amnesic and seemingly unable to commit new material to memory. For example, after his operation, he was very poor at learning the names or recognising the faces of people he met, or remembering the contents of an article he had read just hours before. However, upon closer examination, HM was found to have retained information for experiences that occurred following his surgery. This preserved learning capacity included, for example, the ability to learn the mirror-drawing task. In this task, the participants must carefully trace the outlines of shapes (e.g. stars) while viewing their hands and the shapes through a mirror. HM performed this task on consecutive sessions. Although at the beginning of each session he denied having performed this task before (i.e. $C = 0$), his performance improved across sessions (i.e. $X > 0$). Similarly, when performing on implicit or indirect memory tasks, such as word stem completion, amnesic patients often demonstrate normal levels of retention ($X > 0$), while reporting no memory for the study episode (i.e. $C = 0$).

OBJECTIONS AND DEBATES

The above examples of dissociation with normal participants and with blindsight and amnesic patients clearly demonstrate that when conscious awareness is defined and measured on the basis of subjective report, there is ample evidence for unconscious or implicit perception and memory. However, despite the intuitive appeal the dissociation paradigm, the study of unconscious or implicit cognition remains highly controversial. We will now outline some of the issues underlying the controversy (for a more thorough discussion, see Reingold and Merikle, 1990 and Reingold and Toth, 1996).

The most important source of controversy is the absence of a general consensus as to what constitutes a valid measure of consciousness or awareness. For example, some critics of the study of the unconscious challenge the validity of the subjective report as a measure of awareness. They argue that requiring a subjective report of awareness of seeing or remembering transfers the responsibility for defining awareness from the experimenter to the

participants, who may vary in the criteria or concepts of awareness they employ as the basis for their report. For example, some participants who report null awareness may discount conscious access to partial or degraded information regarding a stimulus or an event, while other participants may use such information as the basis for reporting awareness of perceiving a stimulus or remembering an event. Furthermore, participants may tend to respond on the basis of what they perceive to be the goals or expectations of the experimenter. This tendency is referred to as 'demand characteristics'. It is therefore likely that subjective report, at least on some occasions, may fail to reflect all of the relevant information that is accessible to consciousness. Thus, the validity of the subjective report as a measure of awareness depends on our ability to distinguish between the participants' response bias, affected by factors such as preconceived notions and demand characteristics, and their subjective phenomenal experience. Given the difficulty of this, many investigators prefer an approach to the measurement of conscious awareness based on objective task performance rather than subjective report. This is referred to as the 'objective' approach, and the use of subjective report has been termed the 'subjective' approach.

Proponents of the objective approach employ tasks such as stimulus detection, identification, and forced choice discrimination to measure perceptual awareness, and explicit memory tasks such as recognition and recall to measure conscious recollection. Null awareness is then defined as chance performance on these discrimination tasks. Unconscious or implicit cognition is inferred when null awareness is coupled with evidence of above-chance performance on indirect or implicit tasks. The most important problem with the objective approach to the measurement of awareness is the implied one-to-one mapping between tasks and processes. It is assumed that performance on explicit or direct tasks reflects explicit or conscious cognition, while performance on implicit or indirect tasks reflects implicit or unconscious cognition. The fact that the 'explicit–implicit' terminology is used to refer to both tasks and processes may foster such an unwarranted assumption. In fact, performance on implicit and explicit tasks may reflect conscious, unconscious, or both conscious and unconscious processing. For example, if the discrimination task used to measure conscious awareness is actually sensitive to both conscious and unconscious information, then requiring above-chance performance as a prerequisite for demonstrating implicit cognition may lead, at best, to underestimating the magnitude of unconscious influences or, at worst, to denying their existence.

Clearly, neither the subjective nor the objective approach to the measurement of awareness can be justified *a priori*. Given that the dissociation paradigm depends critically on the availability of a valid measure of conscious awareness, its use in the absence of such a measure is problematic. Thus, any approach to the measurement of conscious awareness must be validated by converging empirical evidence. Another problem with the dissociation paradigm is that it reflects a preoccupation with trying to prove or disprove the existence of the unconscious. Progress in the study of the relation between consciousness and cognition will require going beyond existence proofs and towards the development of multiple conceptual and methodological frameworks in order to explore this complex and controversial topic in a satisfactory way.

References

Berry DC and Dienes Z (1993) *Implicit Learning: Theoretical and Empirical Issues*. Hove, UK: Erlbaum.

Ebbinghaus H (1885/1964) *Memory: A Contribution to Experimental Psychology*. New York, NY: Dover. [Original work published in German; first English translation 1913.]

Graf P and Schacter DL (1985) Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory and Cognition* **11**: 501–518.

Kihlstrom JF, Barnhardt TM and Tataryn DJ (1992) Implicit perception. In: Bornstein RF and Pittman TS (eds) *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*, pp. 17–54. New York, NY: Guilford Press.

Reingold EM and Merikle PM (1990) On the inter-relatedness of theory and measurement in the study of unconscious processes. *Mind and Language* **5**: 9–28.

Reingold EM and Toth JP (1996) Process dissociations versus task dissociations: a controversy in progress. In: Underwood G (ed) *Implicit Cognition*, pp. 159–202. Oxford: Oxford University Press.

Shanks DR and St John MF (1994) Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* **17**: 367–447.

Sidis B (1898) *The Psychology of Suggestion*. New York, NY: Appleton. [Reprinted by Arno Press, 1973.]

Wieskrantz L, Warrington EK, Sanders MD and Marshall J (1974) Visual capacity in the hemianopic field

following a restricted occipital ablation. *Brain* **97**: 709–728.

Further Reading

Bornstein RFE and Pittman TSE (1992) *Perception Without Awareness: Cognitive, Clinical, and Social Perspectives*. New York, NY: Guilford Press.

Cohen JD and Schooler JW (eds) (1997) *Scientific Approaches to Consciousness. Carnegie Mellon Symposia on Cognition*. Mahwah, NJ: Erlbaum.

Erdelyi MH (1985) *Psychoanalysis: Freud's Cognitive Psychology*. New York, NY: Freeman.

Graf PE and Masson MEJ (1993) *Implicit Memory: New Directions in Cognition, Development, and Neuropsychology*. Hillsdale, NJ: Erlbaum.

Kirsner KE, Speelman CE, Maybery ME *et al*. (1998) *Implicit and Explicit Mental Processes*. Mahwah, NJ: Erlbaum.

Reber AS (1993) *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. New York, NY: Oxford University Press.

Schacter DLE and Tulving EE (1994) *Memory Systems 1994*. Cambridge, MA: MIT Press.

Stadler MA and Frensch PAE (1998) *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage.

Umilta C and Moscovitch M (eds) (1994) *Attention and Performance XV: Conscious and Nonconscious Information Processing*. Cambridge, MA: MIT Press.

Indexicals and Demonstratives

Intermediate article

Maite Ezcurdia, Universidad Nacional Autónoma de México, Mexico City, Mexico

CONTENTS

Indexicals are expressions that vary in reference according to the context in which they are used. They are of two sorts: pure, and impure or demonstrative. Unlike pure indexicals, demonstratives require an extralinguistic element, like a demonstration or an intention of a certain sort, in order to refer.

WHAT ARE INDEXICALS AND DEMONSTRATIVES?

Indexicals are expressions that vary in reference according to the context in which they are used; and they do so in virtue of their semantic rule or rule for referring. For example, 'I' is an indexical. Its semantic rule states that it refers to the utterer in a given context. Hence, for 'I' to vary in reference, the contexts in which it is used must vary with respect to the utterer. Any other variation in the context of use will be irrelevant to the reference of 'I'. In the case of 'today', the relevant variation in context will be in the day of use, because its semantic rule states that it refers to the present day or day of utterance.

'I' and 'today' are examples of pure indexicals. But there are also impure indexicals or demonstratives. These include 'that', 'this', 'that person in the corner', 'these houses', 'you', 'she', and so on. Impure indexicals differ from pure ones in that they require an extralinguistic act or element in order to generate a semantic rule. If I uttered 'that' without producing the extralinguistic element, it would not be possible for it to refer. (Given the contrast between 'that' and 'this', it would have to refer to something that is at a certain distance from me; however, there are a great many properties and objects that are at a certain distance from me.) A semantic rule for 'that' which just stated that it refers in a given context to an object or property if and only if it is at a certain distance from the speaker would be insufficient to determine any object or property as the referent. What is required to complete the semantic rule is an extralinguistic element supplied by the speaker. Three candidates have been offered for such completion: demonstrations, like pointings and overt gestures (Kaplan, 1989a); salience (Kaplan, 1989a, pp. 490 and 527); and certain sorts of intentions (Kaplan, 1989b), namely, intentions to refer to an object that is presented in a certain way to the speaker in a given context.

Not all indexical expressions can be straightforwardly classified as demonstratives or pure indexicals. Such is the case with 'here' and 'now'. In their pure use they refer, respectively, to the place and time of utterance. However, there are also demonstrative uses of them: for example, when I say 'we are here' pointing at a place on a map, or 'we don't do that now' meaning that we no longer do something that was done at the time of the Aztecs.

Furthermore, some personal pronouns, like 'he', 'she' and 'it', do not always behave as demonstratives. They sometimes behave as anaphora, as in 'Every girl thinks she is going to the party' (where 'she' is anaphoric on 'every girl'), or in 'Mark came to the dinner; he enjoyed it' (where 'he' is anaphoric on 'Mark').

Indexicals, like proper names, are referring expressions, or singular terms. Phrases such as 'every man', 'some cats', 'most fences', and even 'the dog in the corner' are general or nonreferring. Referring expressions, unlike non-referring ones, require that for an utterance of a sentence containing a referring term to express a proposition, that term has a referent.

ACCOUNTS OF INDEXICALS AND DEMONSTRATIVES

Indexicals and demonstratives have been the source of much debate among competing semantic theories. There are two main issues: how to give an adequate account of the meaning and reference of indexicals; and whether it is tokens (e.g., inscriptions on a blackboard or piece of paper), utterances of indexicals, or indexical types relative to certain contexts, that do the referring. These issues may be related. For example, Reichenbach (1947) called

indexicals 'token-reflexive expressions', and claimed that their meaning would be given by making reference to a token produced by the speaker. The meaning of 'I' would be given in terms of 'the person who utters this token', and the meaning of 'now' in terms of 'the time at which this token is used'.

Within a Fregean semantics – that is, of a theory that recognizes the semantic levels of sense and reference – Evans's (1981) account of indexicals is the best offered so far. The sense of an expression involves a mode of presentation of its referent. Sense also determines an expression's reference, so that if two expressions have the same sense (e.g., 'cat-like' and 'feline') they must have the same reference. Because an indexical may vary its reference across contexts, it cannot always express the same sense; and because the semantic rule of an indexical is what remains constant across all its uses, senses cannot be the semantic rules in accordance with which an indexical varies in reference. For Evans, sense determines reference only because sense is itself determined by the context of utterance. Context, together with the sense type, determines a sense, which in turn determines the reference. The sense of 'I' when uttered by me differs from the sense of 'I' when uttered by you, although their sense type is the same (and remains constant across all uses of 'I'). If anything, the sense type can be taken to be the semantic rule.

Furthermore, since for the Fregean propositions have only senses as constituents, the account of indexicals as referring expressions depends on the nature of the constituent senses. For Evans, referring expressions, unlike nonreferring ones, express object-dependent senses, senses that entail the existence of their referents. So a referring term that lacks a referent lacks a sense, and the utterance that contains such a term fails to express a proposition. On this account, an utterance of a sentence containing 'you' (unlike an utterance of 'some dogs') requires there to be a referent of 'you' for it to express a sense and, therefore, for a proposition to be expressed.

It is, however, the work of Kaplan (1989a, 1989b) that has been most influential on current research on indexicals and demonstratives. Kaplan argues that if we are to develop a semantics and logic of demonstratives, then the bearers of reference should be the indexical types relative to contexts of use. These contexts are represented as ordered sequences of the following sort.

$$(\text{User},\ \text{time of use},\ \text{place of use},\ \text{actual world},\ (d_1, \cdots, d_n)) \quad (1)$$

where the last member is an ordered n-tuple of possible referents of demonstratives. In terms of these contexts, and of the notion of circumstances of evaluation, Kaplan develops a semantic theory that includes three levels: character, content, and extension. The character of a sentence is both its semantic rule and its linguistic meaning (that is, the part of meaning that remains constant across all uses); its content is the proposition it expresses; and its extension is its truth-value. Character can be represented as a function from contexts of use to content, and content as a function from circumstances of evaluation to extension. A circumstance of evaluation is a possible situation, actual or counterfactual, with respect to which the truth-value of a sentence is evaluated.

Relative to the context of use in which Tony Blair is the user of 'I am in London', the character of 'I' yields Tony Blair as part of the content of that sentence, and it is with respect to Tony Blair that the truth of that sentence in context will be assessed. An indexical may vary in reference with respect to a variation in context, but since relative to a particular context its contribution to the content of a sentence is just its referent, it does not vary in reference with respect to possible circumstances of evaluation. Because the contribution of an indexical to the content of the sentence in which it occurs is (if any) its referent, if the indexical did not have a referent there would be no content or proposition expressed by the sentence containing that indexical. This is as required by the condition for referring expressions namely, that a sentence (in context) containing a referring expression expresses a proposition only if that expression has a referrent. In contrast, a nonreferring expression like a definite description contributes to content a descriptive condition, which in different circumstances of evaluation may determine different extensions or none at all.

INDEXICALS AND DEMONSTRATIVES AND COGNITIVE SCIENCE

We may understand cognitive science roughly as the collection of disciplines whose aim is to study human cognition in any of its forms and at any of its levels. Indexicals, being terms of the languages that we use, are within the scope of linguistics, philosophy, and cognitive psychology. Topics of study include, amongst others, the semantics of indexicals, how they are learned, how they are processed, and how the cognitive abilities which we employ in understanding a use of an indexical or demonstrative involve taking contexts into

account (for a general attempt at this, see Sperber and Wilson (1986)). Moreover, we can obtain special insights into our cognitive abilities from consideration of indexicals; these insights are of particular interest in the philosophy of mind and in cognitive psychology. They concern indexical or demonstrative thoughts – that is, psychological states or processes whose content is most adequately reported in public language with the use of an indexical.

Perry (1979) points out that there are thoughts of certain types that are essential for action and that are reported in public language with the use of 'I', 'now', and 'here'. Suppose I believe that Maite must start writing her paper in Room 105 on the 15th of February at 10 o'clock. This belief, though a true belief about myself, may not give me a reason to act if I am amnesic and do not know I am Maite, or if I do not know that it is 10 o'clock on the 15th of February, or that I am in Room 105. It may, however, give me a reason to act if it is accompanied by the beliefs that I am Maite and that it is now 10 o'clock on the 15th of February and that I am in Room 105. What 'I', 'now', and 'here' thoughts do for the subject is locate the subject, the time, and the place in a way that connects with the subject's abilities to perceive, think, and move. There has been some discussion of the nature of such thoughts, and the necessity and nature of mental representations corresponding to indexicals in public language (Cussins, 1990; Millikan, 1990).

Other abilities associated with the use of indexicals and demonstratives give rise to what Kaplan has called 'cognitive dynamics'. These abilities relate to the sorts of psychological processes involved in tracking a specific object, time or place through space and time. In order to refer to the 15th of February on the 14th, for example, I may use 'tomorrow', but in order to continue to refer to the same day on the 15th I have to use 'today', and on the 16th 'yesterday'. Evans (1981) thought that these uses of different indexicals all expressed the same sense, for underlying such uses was a single ability in the subject to track a particular day. Study of the nature of such tracking abilities with respect to time and with respect to objects throughout space has given rise to discussions on the relation between demonstrative thoughts based on perception and the representational character of the psychological processes underlying those abilities, in particular on whether or not they are conceptual (Cussins, 1990; Evans 1982).

One of the most important topics in the philosophy of mind, and in current research in cognitive psychology, concerns the various kinds of consciousness. The indexical 'I' is of particular importance to studying one such kind, namely, self-consciousness. The focus of such research is on a subject's ability to think of or refer to himself or herself in thought. Upon having an 'I' thought, one is presented to oneself in a special way (in which one is presented to no one else). But if the semantic rule of 'I' is to refer to the speaker, then the semantic rule itself will not give us that special way of being presented. Recent attempts to account for the ability to think of oneself have concerned the way we are aware of our own bodies (proprioception) and of our own thoughts (introspection) (Evans, 1982; Bermúdez, 1998).

References

Bermúdez JL (1998) *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Cussins A (1990) The connectionist construction of concepts. In: Boden M (ed.) *The Philosophy of Artificial Intelligence*, pp. 368–440. Oxford, UK: Oxford University Press.

Evans G (1981) Understanding demonstratives. In: Parret H and Bouveresse J (eds) *Meaning and Understanding*, pp. 280–303. Berlin and New York, NY: DeGruyter.

Evans G (1982) *The Varieties of Reference*. Oxford, UK: Oxford University Press [Edited by J. McDowell.]

Kaplan D (1989a) Demonstratives. In: Almog J, Perry J and Wettstein H (eds) *Themes from Kaplan*, pp. 481–563. Oxford, UK: Oxford University Press [First published 1977.]

Kaplan D (1989b) Afterthoughts. In: Almog J, Perry J and Wettstein H (eds) *Themes from Kaplan*, pp. 565–614. Oxford, UK: Oxford University Press.

Millikan R (1990) The myth of the essential indexical. *Nous* **24**: 723–734.

Perry J (1979) The problem of the essential indexical. *Nous* **13**: 3–21.

Reichenbach H (1947) *Elements of Symbolic Logic*. New York, NY: Free Press.

Sperber D and Wilson D (1986) *Relevance*. Oxford, UK: Blackwell.

Further Reading

Braun D (1996) Demonstratives and their linguistic meanings. *Nous* **30**: 145–173.

Brook A and DeVidi R (2001) *Self-Reference and Self-Awareness*. Amsterdam, Netherlands: John Benjamins.

Burge T (1974) Demonstrative constructions, reference and truth. *Journal of Philosophy* **71**: 205–223.

Castañeda HN (1966) 'He': a study in the logic of self-consciousness. *Ratio* **8**: 130–157.

Davies M (1982) Individuation and the semantics of demonstratives. *Journal of Philosophical Logic* **11**: 287–310.

Dokic J (ed.) (1997) *European Review of Philosophy*, vol. II, *Cognitive Dynamics*. Stanford, CA: CSLI.

García-Carpintero M (1998) Indexicals as token-reflexives. *Mind* **107**: 529–563.

Lewis D (1979) Attitudes *de dicto* and attitudes *de se*. *Philosophical Review* **78**: 513–543.

Perry J (1977) Frege on demonstratives. *Philosophical Review* **86**: 474–497.

Yourgrau P (ed.) (1990) *Demonstratives*. Oxford, UK: Oxford University Press.

Innateness, Philosophical Issues about

Intermediate article

Fiona Cowie, California Institute of Technology, Pasadena, California, USA

CONTENTS

Philosophical problems about innateness cover conceptual and empirical issues regarding the claim that concepts or beliefs are innate, inborn, or genetically determined.

INTRODUCTION

Many different kinds of traits are claimed to be innate in us and other animals. We speak of a termite's innate nest-building behavior, a llama's innate hairiness, a Golden Retriever's innate tendency to hip dysplasia, a person's innate beauty. Philosophical interest in innateness, however, has centered around the innateness or not of various mental properties. Probably because of the issue's traditional connections with epistemological problems like justification and a priori knowledge, philosophical attention has focused on the innateness (or not) of our 'ideas' or representations (e.g. concepts and beliefs) and, to a lesser extent, our cognitive abilities (e.g. mechanisms of learning). This article, too, will focus on the innateness of mental items (ideas or capacities) rather than physical features or behaviors.

The first known philosophical claims about the innateness of concepts and beliefs were made by Plato (428–348 BC) in his dialogues *Phaedo* (74b–75e) and *Meno* (80d–e). (See Grube, 1997.) He argues (in the voice of Socrates) that since some of the concepts and knowledge we possess could not have entered the mind through sense experience or as a result of teaching, they must be innate, that is acquired by our souls prior to our birth. So-called learning, he argues, is really a process of recollection: our experience reminds us of things we already know but have forgotten. Later defenders of the view that certain ideas are innate include the philosophers René Descartes (1596–1650), Gottfried Leibniz (1646–1716), and Immanuel Kant (1724–1804). Notable critics of their 'innateness hypotheses' have included the historical figures Aristotle (384–322 BC), John Locke (1632–1704), and David Hume (1711–1776). Most recently, two cognitive scientists, the linguist Noam Chomsky and the philosopher Jerry Fodor, have taken up arms under the nativist banner: Chomsky defending the notion that our knowledge of language is substantially innate, and Fodor arguing that the vast majority of our concepts are innate (e.g. Chomsky, 1986, 1990; Fodor, 1981, 1998).

CENTRAL PHILOSOPHICAL ISSUES ABOUT INNATENESS

The central philosophical question – indeed, the central question – about innateness is: what, if anything, is innate? To aid in answering this question, philosophers have asked a prior question: what is innateness and why should we care about it? Then, to help in answering this question, they have further inquired: what do particular nativists (i.e. defenders of an innateness hypothesis) mean when they call some idea or capacity 'innate'? Thus, the three main philosophical issues about innateness are:

1. the interpretive issue: what do nativists mean by the claim that such and such mental item is innate?
2. the explanatory issue: what is innateness, and what does an idea's being innate enable us to explain?
3. the factual issue: which ideas (concepts, beliefs, knowledge) and cognitive or behavioral capacities are innate?

The interpretive issue arises because it is frequently very difficult to tell exactly what a person means when she claims that certain ideas are innate. People often appeal to metaphors and analogies in explaining what innateness is, and these often support conflicting interpretations. For example,

Descartes sometimes says that concepts or beliefs are innate in the same way that certain diseases (like gout) are innate in certain families (Descartes, 1985). This argument, together with the fact that many of the concepts and beliefs he discusses are not possessed by babies or very small children, suggests that in his view, innate concepts or beliefs are not present at birth but exist as potentialities or dispositions to acquire ideas during the course of development. (See Stich, 1975.) Chomsky, by contrast, claims that principles of Universal Grammar (UG) are innately represented in the human 'language organ' and are crucially implicated in the process of language acquisition. This assertion suggests that, for Chomsky, innate knowledge of UG is no mere disposition but is encoded in the brain at birth in a robust and causally efficacious form. Fodor (1981), to take another example, has likened innate concepts to imprinted behaviors, suggesting that for him, innate ideas are not learned but are merely 'triggered' by some appropriate releasing stimulus. It is unclear that the same or even a similar notion of innateness is at work in these three authors, and the interpretive problem only increases as the work of additional nativists is examined. (See Cowie, 1999, ch. 1, for more on the interpretive issue.)

While it is, of course, important to understand what theorists (past or present) have thought about innateness, the interpretive issue is primarily of interest to cognitive science because it bears on (2), the explanatory issue. Innateness is most commonly invoked to explain how an idea was acquired: Meno's untutored slave knows Pythagoras's theorem because it is innate; we know that $2+2=4$ because we have innate knowledge of mathematics; people believe in God because of an innate concept of the deity. The fact that innate ideas are often contrasted with those that are learned or otherwise acquired as a result of experience has led to the notorious 'nature versus nurture' debate about whether our genes or our experience are responsible for our mental lives. This debate is generally fruitless for two reasons. First, everyone knows that both experience and the genes are implicated in cognitive development: the opposition between nature and nurture is overstated. Second, participants in nature versus nurture debates often talk past one another, simply because there are so many different understandings of innateness at play in the literature. In order to avoid such unproductive arguments, therefore, is vital to be very clear about what notion of innateness is being used in a specific discussion. For example, it would be a serious mistake to read Fodor's claim that most concepts are innate as asserting (in Cartesian vein) merely that we have dispositions to acquire concepts during the course of development. The latter claim, while true enough, trivializes what is meant to be a substantive – indeed, radical – claim about concept acquisition, because it is quite consistent with the typical empiricist's view that concepts are learned. ('Empiricist' is the name given to opponents of innate ideas.)

The explanatory issue thus arises in part because the interpretive issues are not clear. In discussing the innateness of various concepts, beliefs, or capacities we need to be explicit about what notion of innateness we are invoking. We also need to be clear about why that notion is important: we need to understand what innateness, so understood, explains. For too often it is assumed that the innateness of a trait has implications which, upon examination, it does not have. Herrnstein and Murray (1994), for instance, argue that IQ is substantially innate and that certain ethnic groups are inherently deficient in intelligence, concluding thence that social and educational programs aimed at those groups are ineffective. But as Block (1995) has responded, since 'innate' for Herrnstein and Murray means something like 'highly heritable', this conclusion does not follow even if the facts are as they claim. For the heritability of a trait is a measure of the extent to which variance in the trait in a population is correlated with genetic (as opposed to environmental) variation, and high heritability does not imply imperviousness to (or even difficulty of) change by means of environmental manipulations. In the next section of this article, we will look at a number of different notions of innateness used in the contemporary and historical literatures and examine their explanatory abilities.

Once one has clarified one's explanatory goals and chosen an appropriate notion of innateness, the question arises: what traits, if any, are innate (in that sense)? This is the third, the factual issue, about innateness with which philosophers have been concerned. In the final section, we will look at what sorts of evidence bear on this factual question, focusing in particular on the ways in which cognitive science plays a role in determining which of our ideas are innate.

THEORIES OF INNATENESS

Philosophical investigations of innateness have revealed that there are about as many different things meant by the claim that an idea is innate as there are theorists making such claims. In this section, we will survey some of these different

notions of innateness and see what sorts of explanations they figure in.

'Innate' Means 'Built In' or 'Inborn'

When the person in the street talks of hearts or hands or hemophilia being innate, what she often means to indicate is that those features are 'built in' or present at birth, the intended contrast being with traits that are acquired as a result of our postnatal experiences in the world. Arguably, some mental capacities (e.g. some perceptual and inferential abilities) must be innate in this sense, or we would have no mental lives at all. However, there is room for considerable controversy about what else may be inborn. For example, Locke (1975) famously argued (contra Leibniz, 1981) that no beliefs are inborn, since infants are not conscious of them and cannot make use of them, but must rather learn them as a result of experience. Currently, cognitive scientists are debating the extent to which knowledge of language is inborn (see Pinker, 1994; cf. Cowie, 1999), and neuroscientists are debating the extent to which the organization of the brain is inborn rather than being fixed by experience (e.g. Quartz, 2002).

'Innate' Means 'Genetically Determined'

Since the discovery of genes as the mechanisms of inheritance, talk of innateness has sometimes been code for the genetic determination of traits. When we say that a tendency to violence, or dyslexia, or the capacity for shared attention is innate, we may mean that it is under the control of one or more specific genes. Here, the aim is to explain such facts as why some traits seem to run in families, and/or why they seem to develop reliably under many different circumstances, and/or why they seem independent of other cognitive capacities, and/or why they seem impervious to ordinary environmental interventions (e.g. coaching or therapy). A major problem with the use of 'innate' in this sense, whether in cognitive science or elsewhere, is that the concept of genetic determination on which it is based is itself extremely unclear. If, for instance, a trait counts as genetically determined so long as there are genetic necessary conditions for its development, then arguably all traits are genetically determined, hence innate. If, instead, genetic determination requires that there be genetically sufficient conditions for the trait, then arguably nothing is genetically determined, since the survival of any organism requires certain environmental conditions to obtain. If genetic determination is relativized to a certain class of environments (by saying, e.g. that a trait is genetically determined if, given a normal environment, its development is controlled by the genes), then its meaning changes as the class of relevant environments does. And if genetic determination requires that there be particular genes which are 'the cause of' or are 'responsible for' a given trait, then it is at least unclear that many traits – especially those of interest to cognitive science – would be candidates for innateness. First, many genes are implicated in the development of many different traits – regulatory genes are one example; certain hox genes, which are expressed in both the fore- and hindlimbs of tetrapods are another. Second, the development of most traits requires the coordinated activity of many different genes. So while development can certainly be disrupted by 'knocking out' one or more of these genes (as, e.g. deletion of the ultrabithorax gene in fruitflies causes two copies of the second thoracic segment, including wings, to develop), it cannot be inferred from this that those genes are the 'causes of' the trait in any more robust sense than ambient oxygen is 'the cause of' a wildfire. (See Davidson, 2001; Block, 1995; Kitcher, 1985 for more on genetics and genetic determination.)

'Innate' Means 'Has a Flat Norm of Reaction'

One way of making more precise the idea that innate traits are those which are genetically determined is by invoking the notion of a norm of reaction. A norm of reaction is a function which describes how a phenotypical property, P, varies in response to environmental variation in organisms with a given genotype, G. P is said to have a 'flat' norm of reaction when there is little or no change in P across different environments, and this is of interest because the traits we call 'genetically determined' or 'innate' often display this feature. Thus, for instance, hair is both innate and (pretty much) inevitable in mammalian genotypes, as are feathers in avian ones, and backbones in vertebrate ones. It is not clear, however, that this is a useful sense of innateness when speaking of ideas. There are many ideas which are inevitable, yet not innate, being a result rather of ubiquitous features of the environment. For instance, overwhelmingly, most people who survive at all have the concept 'water' and the belief that night follows day.

'Innate' Means 'Canalized'

A related way of making the notion of innateness more precise involves the biological concept of canalization. A trait is said to be canalized when

its development is buffered against certain environmental and genetic variations (Waddington, 1942). Brains, for example, have this feature: most mutations and most environmental changes do not prevent their development. The philosopher Andre Ariew (1996, 1999) has argued that the best account of innateness holds a trait to be innate (in a given range of environments) to the extent that its development is canalized (in that range). Innateness in this sense, as he notes, enables us to explain why certain traits appear so reliably in a population. What is problematic for this proposal, however, especially as an account of what 'innate' might mean for cognitive scientists, is that many ideas seem to be canalized without being plausibly called innate. Again, the belief that night follows day and the concept 'water' are examples.

'Innate' Means 'Genetically Entrenched'

Another philosophical account of innateness is due to the philosopher William Wimsatt. He argues (e.g. 1986, 1999) that innate traits are the ones that are highly 'generatively entrenched'. Generative entrenchment is a measure of what one might call 'ontogenetic necessity'. The more entrenched a property or process is, the more disruptive to subsequent development is its absence or modification:

> to be innate is to be deeply generatively entrenched in the design of an adaptive structure – to be a functional part of the causal expression of that system…upon which the proper operation of a number of other adaptive features depends. (Wimsatt, 1999, p. 153)

Thus, having a brain is in his view a deeply entrenched or innate property: modifications to that organ cause a cascade of other developmental effects. Some problems with this account of innateness are (i) it does not capture the opposition with learning that is so crucial to many discussions of innate ideas, since learned ideas can be highly entrenched precursors to the acquisition of other adaptive cognitive structures; (ii) it apparently counts environmental features as innate if it turns out that they are necessary to further development. (For example, having edges or lines of different orientations in the visual environment would be an innate trait of Hubel and Wiesel's famous kittens, since normal visual development was found to be contingent on experience of those stimuli.)

'Innate' Means 'Developmentally Constrained'

Elman *et al*. (1996) have proposed that innateness be understood in terms of the notion of developmental constraints. In their view, ideas and behaviors count as innate to the extent that there are constraints on (i) the representational systems they involve, and/or (ii) the architectures that implement them, and/or (iii) the timing of their development. The more constraints there are operating at any of these levels, the more the trait counts as innate. Elman *et al*. maintain that their notion of innateness is of interest to cognitive science because it recognizes the complex interplay of genome and environment during development, thus avoiding spurious arguments about 'nature versus nurture'. However, it is difficult to see what innateness, so understood, explains. Since the development of every trait is multiconstrained, not just by environmental conditions and genes of various sorts but also by myriad other factors (such as the laws of physics and chemistry, the plasticity of the relevant portion of the genome, the existence and fitness of alternative phenotypes), all traits would appear to be to some extent innate, in this view.

'Innate' Means 'Triggered' or 'Not Learned'

The philosopher and cognitive scientist Jerry Fodor seeks to import the notion of triggering (Lorenz, 1965) into discussions of innate ideas. He argues (1981, 1998) that most concepts are innate in us in the same way as imprinting is innate in ducklings: just as a duckling's attachment to its mother is triggered by its first visual and auditory experiences of her, our concepts are triggered by our experiences of their instances. Here, the intended contrast is with learning: just as ducklings do not learn to love their mothers, we do not learn concepts – even concepts like 'train' and 'rose' are unlearned, in Fodor's view. A problem with this proposal is that triggering is in most cases a poor model for the process of concept acquisition. First, the connections between releasing stimuli and triggered behaviors are set up by natural selection, but natural selection has not had time to set up connections between many concepts (e.g. 'train') and the world. Second, whereas there is typically a critical period for the development of triggered responses, there appears to be no critical period for the acquisition of concepts like 'train'. Finally, and because behaviors and releasing stimuli are paired by natural selection, the stimuli for triggered behaviors are arbitrary in a way that the experiences leading to acquisition of a concept are not. There is no intrinsic connection between red spots and sex, yet natural selection has recruited red spots as a trigger for mating behavior in stickleback fish. In contrast,

there is an intrinsic connection between concepts and the experiences that cause their acquisition: 'rose' is acquired as a result of seeing, or hearing talk about, or reading about roses.

IMPACT OF COGNITIVE SCIENCE ON ISSUES ABOUT INNATENESS

It remains to consider the factual issue about innateness, and this is where the various branches of cognitive science play a crucial role.There are three main types of argument used in support of innateness hypotheses. These are:

- Transcendental arguments.
- Poverty of the stimulus arguments
- Impossibility arguments.

The soundness of these arguments clearly depends on what sense of 'innate' is being invoked. Transcendental arguments (Antony, 2001) concern innateness in the sense of 'inborn' or 'present at birth'. They point out that in order for learning from experience to be possible at all, something must be there already. Even the most extreme empiricists accept the conclusion of transcendental arguments; where they differ from nativists is in their accounts of what is inborn. Empiricists like the behaviorist B. F. Skinner think that all that is innate are dispositions to emit certain behaviors as a result of certain environmental stimuli, together with a disposition to modify the strength of these stimulus-response connections in response to reinforcement. In contrast, a nativist like Chomsky thinks that knowledge of UG is inborn. This is a much stronger innateness hypothesis than Skinner's, since it requires that there be inborn syntactic representations. Cognitive science contributes to this debate primarily by clarifying the processes of cognitive development. For it is only when we know what the mechanisms of learning and development are, that we will be in a position to identify their inborn precursors.

Poverty of the stimulus arguments contend that a given idea (a belief or a concept) must be innate (again in the sense of 'inborn') because the environment does not provide enough information to enable it to be learned. For example, Chomsky has argued that since the information about language available to children is so meager, knowledge of language must be substantially innate – in particular, he argues, a representation of Universal Grammar, a theory describing the features shared by all human languages, is an inborn feature of the 'language acquisition device' (e.g. Chomsky, 1965, 1993). Poverty of the stimulus arguments rely on four crucial empirical premisses: (i) people do in fact possess the idea in question (e.g. people have knowledge of a generative grammar); (ii) the available information is of such and such kinds (e.g. children do not get evidence about what *not* to say); (iii) the available learning mechanisms are of such and such types (e.g. language learning, if it occurred, would proceed by hypothesis testing); (iv) the available information is not sufficient to enable the idea to be learned by that learning method. Cognitive science will be critical in verifying all these premisses. For example, psycholinguistic theories of the processes of language production and comprehension will be the ultimate arbiters of Chomsky's claim that language use and understanding involve knowledge of a generative grammar; developmental psychology will tell us what information, as a matter of fact, children learning language have access to; and cognitive psychology or cognitive neuroscience, perhaps, will tell us what are the learning methods available to them. Only then will it be clear whether innate knowledge of UG is required for language acquisition.

A third kind of argument for innate ideas claims that since learning a certain idea is literally impossible, that idea must be innate. Fodor is the contemporary champion of this form of argument, contending that since concept learning is impossible, concepts must be innate. Fodor's impossibility argument starts with a transcendental point. Since learning a concept involves formulating and testing hypotheses about what it means, we must already have some representations available to use in stating our initial semantic hypotheses: unless some concepts are innate, concept learning could never get off the ground at all. Fodor's impossibility argument then continues as follows. Successful learning of a concept involves finding the correct hypothesis about what it means. This requires that we find, in effect, the definition of the concept: we have learned the concept 'dog' when we know that dog applies to a thing if and only if that thing is F, where F specifies the necessary and sufficient conditions for doghood. However, Fodor contends that, since most concepts do not have definitions (in the sense that we cannot specify their necessary and sufficient conditions, except circularly, by using the very concepts at issue), most concepts could not be learned. Concept learning is (in most cases) impossible, so most concepts must be innate.

It is still unclear exactly what 'innate' means in the context of Fodor's impossibility argument – there is still a serious interpretive issue here (see

Cowie, 1999, ch. 4). However it is clear, whatever 'innate' turns out to mean for Fodor, that cognitive science has an important role to play in establishing the soundness this argument. First, by developing alternatives to the hypothesis-testing model of concept acquisition, cognitive science may undermine the transcendental argument with which Fodor's argument begins: maybe concept learning does not require that we already have representations (though it may require that we have, say, a neural network with a certain kind of distribution of initial weights and a certain learning algorithm). Second, the development of alternatives to Fodor's account of concept possession may undermine his argument from the indefinability of concepts to their unlearnability. If having a concept is not necessarily a matter of knowing the necessary and sufficient conditions for its application (as, e.g. Prinz, 2002, argues), then the indefinability of concepts is no bar to their being learned.

In sum: claims about the innateness of ideas are often ambiguous. In order to assess such claims, then, it is necessary to understand what sense of 'innate' is being used and to be clear about what innateness, so construed, can explain. Claims about the innateness of ideas are also subject to empirical test, and this is where cognitive science comes in. When the processes of learning and development are better understood, we will be in a better position to know what is innate, and what implications this might have.

References

Antony LM (2001) Empty heads. *Mind and Language* **16**: 193–214.

Ariew A (1996) Innateness and canalization. *Philosophy of Science* **63**: S19–S27.

Ariew A (1999) Innateness is canalization: in defense of a developmental account of innateness. In: Hardcastle V. (ed.) *Where Biology Meets Psychology: Philosophical Essays*, pp. 117–138. Cambridge, MA: Bradford Books/MIT Press.

Block N (1995) How heritability misleads about race. *Cognition* **56**: 99–128.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky N (1986) *Knowledge of Language, Its Nature, Origin and Use*. New York: Praeger.

Chomsky N (1988) *Language and Problems of Knowledge, The Managua Lectures*. Cambridge, MA: MIT Press.

Chomsky N (1990) On the nature, use and acquisition of language. In: Lycan WG (ed.) *Mind and Cognition: A Reader*, pp. 627–645. Oxford: Blackwell.

Chomsky N (1993) *Language and Thought*. Wakefield, RI and London: Moyer Bell.

Cowie F (1999) *What's Within: Nativism Reconsidered*. New York: Oxford University Press.

Davidson ER (2001) *Genomic Regulatory Systems: Development and Evolution*. San Diego: Academic Press.

Descartes R (1985) Comments on a certain broadsheet. In: Cottingham J, Stoothoff R and Murdoch D (eds and trans.) *The Philosophical Writings of Descartes*, vol. I. Cambridge, UK: Cambridge University Press.

Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: Bradford Books/MIT Press.

Fodor JA (1981) The present status of the innateness controversy. In: *RePresentations: Philosophical Essays on the Foundations of Cognitive science*, pp. 257–316. Cambridge, MA: MIT Press/Bradford Books.

Fodor JA (1998) *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Grube GMA (1997) (trans.): *Meno* and *Phaedo*. In: Cooper JM and Hutchinson DS (eds) *Plato: Complete Works*. Indianapolis: Hackett.

Herrnstein RJ and Murray C (1994) *The Bell Curve*. New York: Free Press.

Kitcher P (1985) *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, MA: MIT Press.

Leibniz WG (1981) *New Essays on Human Understanding*, translated by P Remnant and J Bennett. Cambridge, MA: Cambridge University Press.

Locke J (1975) *An Essay Concerning Human Understanding*, edited by PH Nidditch. Oxford: Oxford University Press.

Lorenz KZ (1965) *Evolution and the Modification of Behavior*. Chicago: University of Chicago Press.

Pinker S (1994) *The Language Instinct: How the Mind Creates Language*. New York: Harper.

Prinz JJ (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Quartz SR (2002) Learning and brain development: a neural constructivist perspective. In: Quinlan P (ed.) *Connectionist Models of Development*. New York: Psychology Press.

Stich SP (1975) Introduction. In: Stich SP (ed.) *Innate Ideas*. Berkeley: University of California Press.

Waddington CH (1942) Canalization of development and the inheritance of acquired characteristics. *Nature* **150**: 563.

Wimsatt WC (1986) Developmental constraints, generative entrenchment and the innate-acquired distinction. In: Bechtel W (ed.) *Integrating Scientific Disciplines*, pp. 185–208. Dordrecht: Martinus-Nijhoff.

Wimsatt WC (1999) Generativity, entrenchment, evolution and innateness: philosophy, evolutionary biology, and conceptual foundations of science. In: Hardcastle V (ed.) *Where Biology Meets Psychology: Philosophical Essays*, pp. 139–179. Cambridge, MA: Bradford Books/MIT Press.

Further Reading

Barkow JH, Cosmides L and Tooby J (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Chomsky N (1988) *Language and Problems of Knowledge, The Managua Lectures*. Cambridge, MA: MIT Press.
Cowie F (1999) *What's Within? Innateness Reconsidered*. New York: Oxford University Press.
Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: Bradford Books/MIT Press.
Fodor JA (1998) *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.
Griffiths P (2002) What is innateness? *Monist* **85**: 70–85.
Kitcher P (1985) *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, MA: MIT Press.
Pinker S (1994) *The Language Instinct: How the Mind Creates Language*. New York: Harper.
Pinker S (1997) *How the Mind Works*. New York: W.W. Norton.
Prinz JJ (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
Quartz SR and Sejnowski TJ (1997) The neural basis of cognitive development: a constructivist manifesto. *Brain and Behavioral Sciences* **20**: 537–596.
Stich SP (ed.) (1975) *Innate Ideas*. Berkeley: University of California Press.

Intention

Intermediate article

Élisabeth Pacherie, Centre National de la Recherche Scientifique, Paris, France

CONTENTS

The concept of intention is used to characterize both actions and states of mind. To a first approximation, we classify actions as intentional if they are both purposeful and voluntary and we say that someone intends or has the intention to do something if he is settled on so acting.

INTRODUCTION

The notion of intention lies at the intersection of the two domains of action theory and philosophy of mind. On the one hand, the concept of intentional action is at the heart of action theory. Providing an analysis of the distinction between intentional and non-intentional action is one of its main aims, with important consequences for work on moral and legal responsibility. On the other hand, the notion of intention is also a central element in the web of concepts used to characterize the mind and the various kinds of states and attitudes that belong to the mental realm.

It is not surprising that intentions figure so prominently in both fields of inquiry. A capacity for intentional behavior is the mark of intelligent agency. Intentional behavior is behavior the correct explanation of which cannot be purely mechanistic but requires us to make reference to the agent's attitudes, beliefs and desires, and reasons for acting in a certain way. There is little doubt, therefore, that there is a close connection between intentions and intentional action. Yet, what exactly an intention is, what qualifies as an intentional action, and what is the connection between the two, are all matters of considerable philosophical debate.

ARE THERE SUCH THINGS AS INTENTIONS?

On one simple and prima-facie plausible view, the connection between intentional actions and intentions is to be construed as follows: someone does something intentionally if and only if he or she has the intention to do it and that intention causes him or her to do it. Yet this simple view can be criticized in a number of ways. One objection that can be raised is that it involves the unnecessary postulation of intentions as distinctive states or events. The philosophers who raise this objection (Anscombe, 1963; Davidson, 1980, essay 1) consider that for an action to be intentional is for it to be explainable by reasons the agent has for acting that way. In other words, an action is intentional if it is explainable in terms of beliefs and desires of the agent, typically a desire for, or more generally, a 'pro' attitude towards, a certain outcome and a belief that acting in such or such a way would promote attainment of this outcome. On this view, to say that someone acts intentionally or with an intention is simply to say that some appropriate relation obtains between the agent's beliefs and desires and the agent's actions. Talk of intentions as distinct states or events is therefore superfluous.

This relational analysis of intentions has in turn been criticized. As several philosophers, including Davidson himself, have pointed out (Davidson, 1980, essay 5; Bratman, 1987; Harman, 1986; Velleman, 1989), this analysis is inapplicable to intentions concerning the future, intentions that, although we may now have them, are not yet acted upon, indeed might never result in action. Acknowledging the existence of future-directed intentions forces one to admit that intentions can be states separate from the intended actions or from the reasons that prompted the action. But, as Davidson notes, once this is admitted there seems to be no reason not to allow that intentions of the same kind are also present in all or at least most cases of intentional actions. Once intentions are acknowledged as separate mental states and not just relational constructs, a new issue arises. Can these states be given a reductive analysis – are they assimilable to complexes of desires and beliefs, to

special kinds of beliefs or judgments – or do they form a *sui generis* and irreducible class of mental states?

BELIEFS, DESIRES AND INTENTIONS

Although they acknowledge the existence of states of intending, a number of philosophers want to resist the idea that intentions constitute some new, irreducible kind of attitude. They see beliefs and desires as the two fundamental kinds of mental state and think they can provide an analysis of intentions that remains within the bounds of the belief–desire framework. These elaborations of the belief–desire model have taken several forms.

One approach sees intentions as reducible to combinations of beliefs and desires. It is commonly admitted that intentions have both cognitive and motivational components. Belief–desire reductionism about intentions suggests that the cognitive component of an intention can be identified with a belief and its motivational component with a desire. In particular, it stresses the link between an intention to perform an action and a belief that one will, and claims not only that intending to do something entails believing that one will do it but that the intention consists in this belief. Yet, not all our beliefs about our future actions constitute intentions: some may simply be predictions. A proponent of this view must therefore introduce further constraints that beliefs should meet to qualify as intentions. Here, the motivational component of intentions becomes highly relevant. Thus, Davis (1984) suggests that an agent's intention to *A* consists in believing that the agent will *A* because he or she desires to *A*, together with believing that this desire will motivate him or her to *A*. Alternatively, Velleman (1989) identifies intentions with self-fulfilling beliefs that are motivated by a desire for their fulfilment and that represent themselves as such.

Another approach, advocated by Davidson (1980, essay 5), rejects this belief–desire reductionism. In particular, Davidson does not accept the claim that intending to *A* entails believing that one will *A*. Yet, his approach maintains a close correspondence between intentions, beliefs and desires. According to Davidson, intentions should be viewed as a special kind of evaluative judgment different from the evaluative judgments that correspond to desires to act. Desires to act are what Davidson calls 'prima-facie' judgments, judgments that actions of a certain kind are desirable insofar as they have a certain attribute. As such, prima-facie judgments are not directly associated with actions, for it is not reasonable to perform an action on the sole ground that it has some desirable feature. By contrast, an intention to act is what Davidson calls an 'all-out' judgment, an unconditional judgment, made in the light of one's beliefs, that a certain action is desirable. In making an all-out judgment as opposed to a prima-facie judgment, we settle on a course of action. Intentions are thus associated with actions in a way that mere desires are not. This analysis of intentions introduces a new element, an all-out judgment in the belief–desire framework, yet it avoids having to postulate a completely *sui generis* kind of mental entity. Intentions, together with other pro attitudes, are deemed to belong to the general class of evaluative judgments.

Each of these proposals has encountered specific objections. For instance, it is not clear how, on Davidson's view, an agent would decide between two courses of action that are equally desirable all things considered but are not mutually compatible. It is also disputed whether intending to *A* need always entail believing that one will, not to mention whether it consists in this belief. A more general criticism of such analyses is that they are too backward-looking. They focus on how intentions are arrived at, and therefore fail to appreciate important aspects of the role of intentions.

THE FUNCTIONS OF INTENTIONS

Opponents of reductive approaches to intentions tend to emphasize a number of functions plausibly attributed to intentions. They argue that intentions have their own complex and distinctive functional role that warrants considering them as forming an irreducible kind of psychological state, on a par with beliefs and desires. Among these functions are some that are most apparent when one considers future-directed intentions and the way they function in practical reasoning. Philosophers pursuing this line of inquiry (Bratman, 1987; Harman, 1986; Mele, 1992) propose to articulate what is at stake in being settled on or committed to some course of action as a result of having formed or acquired an intention. Thus, Bratman stresses three aspects of the roles played by intentions in the period between their initial formation and their eventual execution. First, intentions have a characteristic stability or inertia: once we have formed an intention to *A*, we will not normally continue to deliberate whether to *A* or not; in the absence of relevant new information, the intention will resist reconsideration. Second, during this period between the formation of an intention and action, we will often reason from such an intention to

further intentions, for example, reasoning from intended ends to intended means or preliminary steps. Third, this intention will itself often be an element in a larger plan: it will constrain the other intentions we may form and will help us achieve both intrapersonal and interpersonal coordination.

The role played by intentions in practical reasoning is especially salient when we consider future-directed intentions. The executive aspect comes to the fore when we concentrate on immediate intentions. Here the philosophical focus is on the role of the intention in the production of the corresponding action (Brand, 1984; Mele, 1992). This role is both cognitive and motivational. Many philosophers agree that intentions are motivating causes of actions. Furthermore, this role as motivators is not only to trigger or initiate the intended action but to sustain it until completion. If, say, while on my way to my office, I ceased to intend to go my office, this would bring my action to a halt. Intentions also involve the guidance and monitoring of action. The cognitive component of an immediate intention to *A* incorporates a plan for *A*-ing, a representation or set of representations specifying the goal of the action and how it is to be arrived at. Moreover, intentions may be assigned a monitoring function, a capacity to detect progress towards the goal or to detect and correct deviations from the course of action as laid out in the guiding representation.

An exclusive focus on the practical reasoning aspect of future-directed intentions may give the impression that only agents endowed with sophisticated cognitive abilities may be capable of intentional actions. Yet, some philosophers think that it would be unduly restrictive to limit the sphere of intentional actions to premeditated or planned actions. They are ready to accept that some actions that we perform relatively spontaneously and without prior deliberation, or some actions performed by creatures devoid of the sophisticated capacities necessary for forward planning, still qualify as intentional actions. Concentrating on the executive aspect of immediate intentions may be a way to make sense of this idea. We will return to this question, after considering one further idea related to the executive aspect of intentions.

ARE INTENTIONS SELF-REFERENTIAL?

Reflection on the executive functions of intentions has led some philosophers (Harman, 1986; Searle, 1983) to claim that intentions involve a kind of causal self-referentiality. Put very simply, the self-referentiality thesis is the thesis that an intention to *A* is the intention that this very intention be the cause of the agent's *A*-ing. Two main considerations may be adduced in favor of this thesis.

First, if one accepts a causal analysis of intentional actions according to which an action's being intentional depends on its being appropriately caused by a corresponding intention, one must find a way to exclude cases of deviant causal chains. An example, adapted from Harman, may illustrate the nature of the problem. Mabel may intend to kill Ted by running him over in her car. Ted happens to be walking by when she backs out of her driveway and she runs him down, killing him without even seeing him. Here, although Mabel intended to kill Ted by running him over and killed him in this way, her action is not intentional: although her intention plays a causal role in her running over Ted, it is not the causal role specified in the content of the intention itself. The suggested moral is that for an act of *A*-ing to be intentional, it is not enough that it be caused by an intention to *A* or even an intention to *A* in a certain way, but the intention must be an intention that this very intention lead one to *A* in a certain way.

This conclusion has been challenged by Mele (1992), who argues that what accounts for Mabel's action being unintentional may be a function of the plan component of her non-self-referential intention. For instance, he would say that Mabel's action is not intentional because her intention did not incorporate a plan to run down Ted while backing out of her driveway. Mele insists that it not enough for an action to be intentional that it fits some intended plan; it must be guided by the plan. It is unclear, however, what is involved in the notion of guidance and whether it can be defined so as to exclude causal deviance without invoking the idea of causal self-reference of intentions.

A second kind of consideration supporting the self-referentiality thesis refers to the conditions of satisfaction of intentions. Searle (1983) argues that the conditions of satisfaction of an intention include that the intention itself play a causal role in bringing about the rest of its conditions of satisfaction. Moreover, Searle claims that this causal self-referentiality condition is internal to the content of the intention. Thus, the content of my intention to raise my arm should include that I raise my arm as a result of this very intention.

This construal of the self-referentiality thesis has also been challenged. It has been objected that, even if self-referentiality is part of the conditions of satisfaction of intentions, there is no obvious reason

why this condition of satisfaction should be reflected in the content rather than attached to the psychological mode itself or considered part of the background conditions. It has also been objected that Searle's account of self-referentiality is psychologically implausible in requiring of the intenders unrealistically strong cognitive and conceptual capacities. Recently, attempts have been made to dispel the charge of conceptual oversophistication by showing how self-referentiality can be construed as an implicit element of the content of intentions, reflecting architectural and circumstantial facts about how the intention connects with action (Pacherie, 2000; Roth, 2000).

THE CONNECTION BETWEEN INTENTIONS AND INTENTIONAL ACTIONS

The discussion of self-referentiality above raised the question whether it is sufficient for an action of *A*-ing to be intentional that the agent have the intention to *A* or whether it is furthermore required that the intention be self-referential. We can also ask the converse question, namely whether it is always necessary for an action of *A*-ing to be intentional that the agent intends to *A*. In other words, are our folk-psychological notions of intention and intentional action sufficiently univocal that philosophers may hope to provide a unified analysis of their connection? We will now examine three kinds of cases that seem, in different ways, to shed light on the connection between intentional actions and intentions.

Expected Side-effects and Responsibility

Both Harman (1986) and Bratman (1987) have argued that certain actions we perform are intentional despite not being intended because they are expected side effects of some (larger) action we intend to do. For instance, Bratman claims that if someone intends to run a marathon and believes that he will thereby wear down his running shoes, his action of wearing down his running shoes is intentional despite not being intended. Similarly, Harman suggests that in firing his gun to try to kill a soldier, a sniper who knows that he will thereby alert the enemy to his presence does not intend to alert the enemy, yet does so intentionally. If this analysis is accepted, we are forced to conclude that there is a gap between intentions and intentional actions. One way to try to avoid this conclusion, suggested by Mele and Moser (1994), involves drawing a threefold distinction among intentional actions, unintentional actions and non-intentional actions. Since the sniper does not unknowingly or accidentally alert the enemy, the action is not unintentional, but since, on the other hand, alerting the enemy was not a goal or aim of the agent, nor is it intentional. Non-intentional actions would then constitute a middle ground between intentional and unintentional actions.

However, as both Bratman and Harman note, the tendency to consider an expected side effect as an intentional action depends on the significance of the action and on the reasons one might have not to do it. We can contrast alerting the enemy with another expected side effect of firing the gun, such as thereby heating the barrel. Although we have no inclination to say that the action of heating the barrel was intentional, we may want to say that alerting the enemy was intentional insofar as the sniper had reasons not to want to alert the enemy and yet chose to do so. Here the topic of intentional action becomes entwined with issues of responsibility and moral assessment. Maintaining that an action was intentional may justify holding the agent responsible for it. At the same time, maintaining a distinction between intentional actions that are intended and intentional actions that are not may also make an important difference for moral or legal purposes.

Goals, Plans, and Skills

We have considered cases where one might be said to act intentionally in *A*-ing despite not having *A*-ing as a goal, and thus not having the intention to *A*. Conversely, there are cases where one does *A*, one has *A*-ing as a goal and in this sense an intention to *A*, and yet it is unclear whether the *A*-ing qualifies as an intentional action.

Take the following example adapted from Mele. Tom, who has never handled a gun before but is offered a large prize for hitting the bull's-eye on a distant target, aims carefully, fires and hits the bull's-eye. Does he do it intentionally? Here, intuitions diverge. Some philosophers consider that Tom intentionally hits the target, others deny it. Those (Velleman, 1989; Mele and Moser, 1994) who deny that the action is intentional typically insist that in this and similar cases, successfully *A*-ing is not sufficiently within the control of the agent because he lacks a reliable plan for or sufficient skill at *A*-ing and because there is therefore an important element of luck involved in the action resulting in success of failure. Yet, in Tom's case, even philosophers who would deny that the action

was intentional accept that the action was done for a reason – Tom wanted the prize money and thought that to win it he had to hit the bull's-eye – and that in this sense he acted with the intention to hit the bull's-eye. Velleman sees such cases as proof of a fundamental ambiguity of the term 'intention', which is used to denote both plan states and goal states. It would seem that those who think that Tom's and similar actions qualify as intentional actions have intentions as goal states in mind, and those who deny this think of intentions as plans or representations that guide and control the action.

Yet Velleman's distinction may still not fully capture the complexity of our intuitions. Take the following more extreme example, borrowed from Mele and Moser. Lisa selects a sequence of six numbers to win a fair Florida instant lottery, and, as it happens, she wins the lottery. Here, it seems that we would not want to say that Lisa wins intentionally, and that we would also be reluctant to say that she acted with an intention to win, despite the fact that she chose the numbers out of a desire to win the lottery. Our unwillingness to consider the action as intentional is not motivated by Lisa's lack of skill or the frailty of her plans. Assuming that the lottery is fair and that there is no way to win by cheating, Lisa has done all that was in her power to win the lottery. The difference between the Lisa and Tom cases lies in how much the successes of their actions, considered as types, can in principle depend on the agents' skills or planning abilities.

It seems, therefore, that to account for our intuitions a threefold distinction is needed. At one extreme are actions over which the agent has a sufficient degree of control, and which are therefore uncontroversially intentional. In the middle are actions over which the agent could in principle have a sufficient degree of control (as witnessed by the fact that some other agents do exhibit such control – for instance expert marksmen), and which can be deemed intentional in a weaker sense (Tom's case). At the other extreme are actions over which we could not in principle have sufficient control, because there is some important element of luck on which their outcome depends essentially (Lisa's case).

Sudden, Impulsive, and Subsidiary Actions

Some philosophers (Searle, 1983; Wilson, 1989) argue that sudden and impulsive actions can be intentional despite being performed without the agent's forming, either consciously or unconsciously, a prior intention to do them. Thus, if a pile of books on my desk starts to topple, I might suddenly reach for the pile to keep it from collapsing; or, frightened by an earth tremor while in California I might impulsively throw myself under a table; acting in both cases intentionally. Similarly, Searle claims that even in cases where we have a prior intention to do something, there are often many subsidiary actions that are not represented in the prior intention and yet are intentionally performed. I might have a prior intention to type a certain sentence, without the intention representing the key pressings that must be performed for the sentence to be typed. It may be thought that sudden and impulsive actions are too fast for there to be time to form an intention. It may also be pointed out that such actions do not meet the strong consistency requirements that are distinctive of actions done in conformity with a prior intention. Typically, they are performed with no regard for their potential side effects or for their general coherence with the agent's other beliefs and goals.

On Searle's account, these actions are intentional despite the absence of a prior intention because they involve an 'intention-in-action'. In his terminology, a prior intention is an intention that is formed prior to the action and that represents and causes it as a whole. Thus, both future-directed and immediate intentions are prior intentions. An intention-in-action is the mental component of the action itself; it presents, causes, and is contemporaneous with, bodily movements. The intention-in-action together with the bodily movements it causes constitute the action. According to Searle, all intentional actions have intentions-in-action but not all have prior intentions. And even when both a prior intention and an intention-in-action are present, the intention-in-action will typically be much more determinate than the prior intention.

The claim that some intentional actions need not involve prior intentions can be challenged in two different ways. Some philosophers agree that in the case of sudden or impulsive actions, intentions are absent, but, because of this absence, deny that such actions really qualify as intentional. Bratman (1987) suggests that they are purposeful and voluntary but still not intentional (nor unintentional). Conversely, one might agree that such actions are intentional, but claim that they involve an intention. Thus, Mele (1992) distinguishes between forming an intention and (passively) acquiring one, and suggests that although in the case of sudden and impulsive actions an agent may lack sufficient time to form an intention, he or she may still have time to acquire one. Similarly, in the case of actions

involving many subsidiary actions, one may contend that a prior intention normally includes a detailed plan for acting and thus that intentions-in-action are superfluous.

As Mele and Moser (1994) suggest, highly relevant to this debate is the difficult question of how much of what is going on representationally in the preparation and execution of an action should be included in the content of intentions. Work in the neuroscience of action suggests that there are multiple levels of representation and control of action, from high-level specifications of goals down to low-level specifications of neuromuscular activity. It remains an open question, partly empirical and partly conceptual, which of these representations should be considered part of the content of intentions or, if one countenances them, intentions-in-action, and what role criteria such as accessibility to consciousness or representational format (conceptual or not) should play.

CONCLUSION

Philosophical analyses of the notion of intention reveal a number of aspects – practical reasoning, planning, goal-directedness, control of action – which it may either essentially include or be closely associated with. Presumably, what we take as prototypical cases of intentions and intentional actions are cases where all dimensions are present. In less clear cases, philosophical intuitions differ. The debates over the nature of the connection between intentions and intentional actions bespeak the lack of general agreement on the respective roles and relative importance of these aspects. Different approaches privilege different aspects or weigh them differently when characterizing intentions as states of mind and when characterizing intentional actions. These disagreements may in turn be explained in part by interest in the different issues – rationality, ethics, freedom, mental kinds, motor organization – with which the notion of intention is connected. A number of studies have explored intentions from an intersubjective point of view. Some theorists are interested in collective actions and the types of shared or cooperative intentions that make them possible. Others try to characterize the abilities that underlie our capacity to interpret observed actions as intentional and to attribute intentions to the agents performing them, thus connecting work on intention with the topic of mindreading. These new perspectives may well pay special attention to features of intentions less salient in other approaches.

References

Anscombe GEM (1963) *Intention*, 2nd edn. Ithaca, NY: Cornell University Press.

Brand M (1984) *Intending and Acting*. Cambridge, MA: MIT Press.

Bratman ME (1987) *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Davidson D (1980) *Essays on Actions and Events*. Oxford, UK: Clarendon Press.

Davis WA (1984) A Causal Theory of Intending. *American Quarterly* **21**: 43–54.

Harman G (1986) *Change in View*. Cambridge, MA: MIT Press.

Mele AR (1992) *Springs of Action*. Oxford, UK: Oxford University Press.

Mele AR and Moser PK (1994) Intentional action. *Noûs* **28**: 39–68.

Pacherie E (2000) The content of intentions. *Mind and Language* **15**: 400–432.

Roth AS (2000) The self-referentiality of intentions. *Philosophical Studies* **97**: 11–52.

Searle JR (1983) *Intentionality*. Cambridge, UK: Cambridge University Press.

Velleman JD (1989) *Practical Reflection*. Princeton, NJ: Princeton University Press.

Wilson G (1989) *The Intentionality of Human Action*. New York, NY: Peter Lang.

Further Reading

Audi R (1993) *Action, Intention, and Reason*. Ithaca, NY: Cornell University Press.

Castañeda H-N (1975) *Thinking and Doing*. Dordrecht, Netherlands: Reidel.

Davis L (1979) *Theory of Action*. Englewoods Cliffs, NJ: Prentice-Hall.

Donagan A (1987) *Choice: The Essential Element in Human Action*. London, UK: Routledge.

Gallese V and Goldman A (1998) Mirror neurons and the simulation theory of mind reading. *Trends in Cognitive Science* **2**: 493–501.

Ginet C (1990) *On Action*. Cambridge, UK: Cambridge University Press.

Grice H (1971) Intention and uncertainty. *Proceedings of the British Academy* **57**: 263–279.

Jeannerod M (1997) *The Cognitive Neuroscience of Action*. Oxford, UK: Blackwell.

O'Shaughnessy B (1980) *The Will*. Cambridge, UK: Cambridge University Press.

Taylor C (1964) *The Explanation of Behaviour*. London, UK: Routledge and Kegan Paul.

Intentional Stance

Intermediate article

Don Ross, University of Cape Town, Cape Town, South Africa

CONTENTS

The intentional stance is an attitude to the place of mind and psychology in the world that allows a real scientific role for such concepts as beliefs and desires without forcing us to suppose that they have literal physical (or nonphysical) correlates in people's heads.

WHAT IS THE INTENTIONAL STANCE?

One takes the 'intentional stance' (IS) towards a system – whether person, animal, machine, or group – whenever one explains and predicts its behavior using so-called 'intentional' concepts such as 'belief' and 'desire'. For example, one might explain or predict a delivery-person's leaving a package outside the front gate, instead of on the doorstop, by saying 'she believes that the dog is vicious, and she doesn't wish to get bitten'. Or, to use the original example due to Daniel Dennett (who introduced the IS as a technical idea), one might explain a chess-playing computer's move by saying 'it wants to get its queen out early'. To predict the behavior of the machine, one must ascribe to it the desire to play winning chess, in addition to some beliefs about chess rules, good strategies, and the probable reactions of opponents. In general, to take the IS towards a system is to assume that that system is related to states of the world by a network of intentions, and that these intentions relate to each other in standard ways familiar in our 'folk psychology': for example, if a person desires outcome x, and believes that doing y is a good way to bring about x, and fears other consequences of doing y less than she desires x, and believes that she can do y, then she will do y. (*See* **Folk Psychology**)

It may seem that in the case of the chess-playing computer, the assumption that the machine has beliefs and desires is just a pretense, like assuming that the actor in a film really does fear being eaten by the surrounding aliens. The pretenses may be necessary in both cases – you cannot enjoy the film if you keep reminding yourself that the actor's beliefs are feigned, and you cannot predict what the computer will do if you do not treat it like a real chess player – but, one might think, to take either pretense literally would be to lose touch with reality. Dennett would agree that the case of the actor involves pretense, but he argues that, if the chess-playing computer is complicated enough, we should take the IS towards it just as we take the IS towards people – including ourselves.

We will first briefly explain the reasons for this view, which might initially seem absurd, and then describe the role it has played in cognitive science.

ARGUMENTS FOR THE INTENTIONAL STANCE

Understanding the logical path that leads to the IS requires some reference to the history of attempts to locate the place of mind in nature. Most philosophers before the twentieth century believed that minds were nonphysical entities of some sort. On this view, beliefs and desires could be thought of as real states of 'mental spirit'. However, if mind–body dualism is rejected, then it is necessary to replace the dualist's spiritual states by physical ones. An obvious idea here is that every intentional state – every state, that is, that is 'about' some other state or object p, and so could be described in terms of a belief that p, desire that p, fear that p, or other 'propositional attitude' – must be identical with a brain state, directly identifiable and describable in the language of neuroscience. This idea faces immediate philosophical questions, however. How could a brain state be 'about' an indefinite class of objects (e.g. 'purple things'), or, perhaps worse, an indefinite class of abstract objects (e.g. 'democratic countries')? After all, nothing in a brain is purple or democratic. Moreover, people, who have explained and understood each other in intentional terms for millennia, do not directly observe brain states, even their own. What we observe, at least in the case of others, is just behavior. (*See* **Propositional Attitudes**)

In 1949, the philosopher Gilbert Ryle sought to resolve this problem by arguing that the concept of mind, and its subsidiary concepts, including the intentional concepts, are constructed from observable patterns in behavior. The mental states that we think we directly perceive as 'inner states' in our own cases are in fact, according to Ryle, judgments about our motivations and general dispositions, which, because they happen automatically, 'feel like' inner perceptions.

One of Ryle's foremost students was Daniel Dennett. Around the time (1969) that Dennett published his first book, the view that mental states are identical to brain states was collapsing under the assault of functionalism. If the mind–brain identity theory were true, the functionalists argued, then, by definition, no two creatures with significantly different brains could ever share a common belief. What content a belief has, the functionalists argued, must be a function of that belief's role in a general network of intentional states. What grounds the whole network, Dennett added, are observable patterns in behavior and expectations about behavior (e.g. the belief that *p* is the state that interacts with other beliefs and desires to produce a disposition to say 'yes' when asked whether '*p*'). When one system interprets another, or a system interprets itself, in terms of such patterns of expectation, then the intentional stance is taken. (*See* **Functionalism**)

The integration of work in artificial intelligence (AI) with cognitive science more generally during the 1970s contextualized the idea of the IS. If you are trying to construct a mind, do you need to deliberately build in, one by one, the millions of beliefs and desires you think it will need if it is to be genuinely intelligent? Or, even if you think the system can build these for itself using some learning procedure, must it ultimately come to possess millions of physically distinct symbolic tokens, one for every distinct intentional concept, that it can then internally manipulate? Dennett argued that not only does neuroscience provide no evidence for such a picture, but that it seems too inefficient a kind of design to have been produced by evolution. In real, changing, environments, such serial symbol-processing incurs enormous computational search and retrieval costs. It also fails to exploit the fact that the external world itself stores plenty of information. A system that can just be disposed to react appropriately given various input patterns can achieve complex behavioral capacities using a much less complicated content-storage system. And if a system's reaction patterns can be shaped by its environment over time, so that it acquires new dispositions as it learns, then, if it has a symbolic bookkeeping system like a language, it might try to keep track of these shifting systems of dispositions by labeling them. The labels can be stored in the world – in texts, and in networks of other systems' dispositions. And then the labels themselves can be triggers of dispositions. Thus I can cause you to judge that you believe that Winston Churchill liked brandy.

Many AI programs, especially those running on 'connectionist' architectures, do not store the beliefs by which we can reliably explain and predict their dispositions separately, at distinct physical addresses in their internal circuits. Instead, these dispositions are consequences of the interactions of patterns of informational activity distributed across the system. Some evidence from neuroscience suggests that, at least in their general cognitive operations, brains work in this way too. If this is so, then there is no particular neural state of your brain that codes your belief about Churchill and brandy. Your belief just consists in the fact that your whole system of dispositions is such that your behavior is consistent with that belief, and inconsistent with its denial. When you wonder whether you have that belief, you take the intentional stance towards yourself, and, on the basis of a very simple piece of behavior – asking yourself a question in English – you rightly judge that you do. This is the same sort of procedure by which we judge that the computer believes appropriate things about chess: we put it into a game situation and look for certain sorts of patterns in its behavior. (*See* **Connectionism**)

PROBLEMS WITH THE INTENTIONAL STANCE

There are two different ways in which the intentional stance can be invoked in cognitive science. If one does not wish to be committed to a metaphysical position on the mind, one can merely agree that we can and do assume the IS towards systems for various practical (including scientific) purposes, while remaining agnostic on the question of whether there is more to having a mind than being the sort of system to which the IS is usefully taken. This philosophically cautious use of the IS is generally uncontroversial. Taken this way, the IS is just a rough measuring device for studying interactions between systems with goal-directed behavior patterns and their environments.

Dennett and his followers, however, understand the significance of the IS in a stronger sense. Following Ryle's lead, they argue that no discrete

inner physical state of a system can be identified with a particular intentional state, in the sense of having, by itself, all and only the content picked out by the IS-level description (e.g. the content 'that Churchill liked brandy'). Therefore, they argue, to be an intentional system – that is, a system with a mind – is just to be a system that exhibits behavioral patterns that cannot be adequately explained and predicted without use of the IS (Dennett, 1991b).

Critics of this view argue that it does not explain minds, but rather explains them away. They claim that it amounts to supposing that minds and intentions are merely useful fictions, and that in that case they are not proper scientific objects; and the ultimate goal of cognitive science should be to eliminate them. This view is known as eliminative materialism. (*See* **Eliminativism**)

Dennett and his followers respond to this criticism, first, by noting that the IS may ultimately be indispensable. We may never be able to explain and predict the behavior of a complex system, even a system we have designed and built ourselves, such as a sophisticated chess-playing computer, merely by reference to its internal causal network. We might have to make reference to its web of inferential and causal relations with the world of chess and chess-related patterns and objects: after all, that is what the system is tracking in both its behavior and its learning. If this point is granted, the Dennettian can then argue that patterns of this sort are fully 'real' in the only sense of 'reality' that matters: they encode information in such a way that, if you miss the encoding, there is something about the world you will not be able to know (Ross, 2000).

THE INTENTIONAL STANCE AND COGNITIVE SCIENCE

Even cognitive scientists who do not wish, like Dennett, to explain the very idea of mind by reference to the IS, have found the IS perspective useful in practice.

Firstly, the IS has been helpful to scientists building and theorizing about models of distributed cognition, in which the world-oriented behaviors of the models can only be described at the whole-system level. If described strictly at the level of their internal operations, these systems – even brains themselves – do not seem to exhibit any intentional behavior at all. Their designers must, however, have target 'intentional' behaviors to guide their work. This requires the assumption of the IS towards them (Clark, 1989). (*See* **Distributed Representations**)

Secondly, those who study the minds of non-human animals and prelinguistic human infants must form and test hypotheses about relationships between the creatures' behaviors and their goals and functions, but face the problem that they cannot determine these, as with adult humans, by asking the subjects themselves to describe them. These researchers into what Dennett has called 'cognitive ethology' begin precisely with the rudimentary IS framework – the idea of a psychology based on beliefs and desires – and try to determine its details for other species and for young humans. Work of this sort with children strongly suggests that the capacity for using the IS is a genetically natural part of human development, which arises in systematic and predictable intervals as people mature. (Autism appears to be a syndrome arising from a variety of developmental and genetic mishaps that interfere with the capacity for assuming the IS (Griffin and Baron-Cohen, 2002).) This perspective adds weight to the scientific and empirical 'reality' of the IS, regardless of the philosophical debate. (*See* **Autism**)

Finally, Dennett has himself built an extensive theory of consciousness using the scaffolding of the IS (Dennett, 1991a), and this has played a significant role (along with rival views) in providing a conceptual framework for empirical research.

References

Clark A (1989) *Microcognition*. Cambridge, MA: MIT Press/Bradford.

Dennett D (1991a) *Consciousness Explained*. Boston, MA: Little, Brown.

Dennett D (1991b) Real patterns. *Journal of Philosophy* **88**: 27–51.

Griffin R and Baron-Cohen S (2002) The intentional stance: developmental and neurocognitive perspectives. In: Brook A and Ross D (eds) *Daniel Dennett*, pp. 83–116. Cambridge, UK: Cambridge University Press.

Ross D (2000) Rainforest realism: a Dennettian theory of existence. In: Ross D, Brook A and Thompson D (eds) *Dennett's Philosophy: A Comprehensive Assessment*, pp. 147–168. Cambridge, MA: MIT Press/Bradford.

Further Reading

Allen C, Bekoff M and Lauder G (eds) (1998) *Nature's Purposes*. Cambridge, MA: MIT Press/Bradford.

Bogdan R (ed.) (1991) *Mind and Common Sense*. Cambridge, UK: Cambridge University Press.

Bogdan R (1997) *Interpreting Minds*. Cambridge, MA: MIT Press/Bradford.

Brook A and Ross D (eds) (2002) *Philosophers in Focus: Daniel C. Dennett*. Cambridge, UK: Cambridge University Press.

Dahlbom B (ed.) (1993) *Dennett and His Critics*. Oxford: Blackwell.
Dennett D (1969) *Content and Consciousness*. London: Routledge.
Dennett D (1978) *Brainstorms*. Cambridge, MA: MIT Press/Bradford.
Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press/Bradford.
Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.
Dennett D (1996) *Kinds of Minds*. New York, NY: Basic Books.
Dennett D (1998) *Brainchildren*. Cambridge, MA: MIT Press/Bradford.
Fisette D (ed.) (1992) *Daniel C. Dennett et les Stratégies Intentionelles*. Montréal: Presse du Université du Québec à Montréal.
Millikan R (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press/Bradford.
Millikan R (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.
Ross D, Brook A and Thompson D (eds) (2000) *Dennett's Philosophy: A Comprehensive Assessment*. Cambridge, MA: MIT Press/Bradford.
Ryle G (1949) *The Concept of Mind*. Harmondsworth, UK: Penguin.

Intentionality

Introductory article

Akeel Bilgrami, Columbia University, New York, New York, USA

CONTENTS

'Intentionality' is a term used by philosophers to describe a property which distinguishes a subset of all the mental states there are – the representational states.

WHAT IS INTENTIONALITY?

'Intentionality' is a term used by philosophers to describe a property which distinguishes a subset of all the mental states there are – the representational states. What is meant exactly by 'representational' is not an entirely simple or uncontroversial matter, and moreover, intentionality should not to be confused with the specific representational mental state of 'intentions'. Intentions are but one example of a representational state, while 'intentionality' encompasses the entire class of representational states.

The most general way to describe intentional or representational states has been to say that they are 'about' things. Such states stand in contrast to mental states such as pain and other kinds of sensation that are not (or not obviously) about things. Equally, some emotions, such as certain forms of anxiety, or depression, or feelings of elation are not necessarily 'about' something; they may just be generalized feelings of depression, anxiety, or elation with no particular object towards which these feelings or emotions are directed. Some philosophers, however, think that sensations do involve representation and 'aboutness'. Some even think that pain is representational because it is essentially a form of perception. Thus, a toothache is one way of perceiving a certain tooth, akin to seeing the tooth or touching the tooth. But despite these controversies, most philosophers accept that there are some mental states which are not intentional at all.

Of those mental states which are intentional, which are about things, it is useful to distinguish between two types: those whose 'aboutness' or representationality are captured in clauses specifying propositions or sentences, and those which are not. Beliefs and desires are the most commonly cited examples of intentional states whose aboutness is specified by propositions. Thus, for instance, someone believes that there is a glass of water in front of her, or someone desires that she drink the glass of water in front of her. What follows the 'that' in these examples is the propositionally specified, representational element of these mental states. And philosophers often use the term 'content' or 'intentional content' to describe this sort of representational element provided by these 'that' clauses in our reports of intentional states. Further examples of such 'propositional attitudes' include hopes, wishes, supposals, and conjectures.

Not all intentional states need be specified by propositions in 'that' clauses. Love, for example, is a state of mind that is directed towards an object but is not always specified by a proposition in a 'that' clause. Love may have an object, a person, say, without the state of mind of love being specified in a 'that' clause content. Many also think that not all cases of perception or seeing are cases of 'seeing that'.

While the term 'intentional' has been used for the most part to describe and discuss states that have propositional or 'that' clause content, there has also been discussion of representational states of the other kind, where 'that' clause content is not involved, but such discussion has been carried out under the more general label 'representational'. It probably reduces confusion to follow this tendency and to use 'representational' as the more general term, reserving 'intentional' for identifying those states with propositional content specified in 'that' clauses.

The remainder of this article will focus on this latter kind of state. These states have been the subject of much diverse and interesting philosophical discussion in the last century or more. This article summarizes the two arguments which are perhaps the most current and widely discussed.

IS INTENTIONALITY A NATURALISTIC PHENOMENON?

One of the questions which has produced the most controversial discussion is whether intentionality is a naturalistic phenomenon: is it reducible to the sorts of phenomena that are studied by natural science? Some philosophers have thought that it *is* reducible, others have denied this, while a third group has said that talk of intentional phenomena is merely talk, an idiom used for convenience but with no lasting status in a complete science of the mind. Paul Churchland, following certain ambivalent remarks of Quine, has been an advocate of such 'eliminativism' in recent years, and Daniel Dennett has described our use of the idiom as an intentional 'stance' that we take toward certain sorts of phenomena we study.

Reductionism

Let us look more closely at the first group, the reductionists. Reductionists can broadly be classified into three subgroups. Some have thought that intentionality is reducible to the physical states of the brain ('materialists'), others that it is reducible to tendencies and patterns of behavior ('behaviorists'). The third group has believed that it can be reduced to purely causal or dispositional states ('functionalists').

The last might be described as akin to the dispositional state of solubility of, say, sugar, except that unlike sugar, mental dispositions are not further reducible to physical–chemical states. Behaviorism, which denies any inner reality to mental states, has been out of favor for many decades, though it once had a high degree of popularity. The materialists claim that future research will establish that each intentional state is identical with a neurophysiological state, just as research in physics has demonstrated that heat is identical with mean kinetic molecular energy. But, like the behaviorists, the materialists have also lost ground to the functionalists who favor a weaker form of reduction, one that reduces intentional states to purely causal states. Though weaker, this form of reduction can still claim that it has reduced intentional states to those which are at least naturalistic, that is to say obedient to natural science, and subsumable under causal laws.

Functionalists do not stress the intrinsic properties of intentional states but rather their functional properties or roles, properties derived from the typical causal relationships in which they stand; i.e., the states they are caused by as well as those states and behaviors which they cause. These reductionists, who think that intentional states are reducible only to causal and dispositional states and not to brain states, usually at least admit that intentional states are what they call 'supervenient' on brain states. Supervenience is a weaker form of a dependency relation of such states on physical states than reduction to them. Thus, given supervenience, if there were two subjects with identical states of the brain, there would be no reason to count one as having different mental dispositions from the other and, therefore, as having a different intentional state from the other.

Arguments Against Reductionism

Anti-reductionists deny even a weak reduction to causal or dispositional states, insisting that intentional states are completely 'stand-alone' states. Although they may have various relations to our dispositions, to our behavior, and to the states of the brain, none of these relations can be invoked to provide any form of reduction. When it comes to mental states generally, resistance to reduction is based on arguments invoking the 'multiple realizability' of mental states, invoking the fact that mental states such as pain are realized (recognized) in quite different physical structures in different species.

But when the subject is intentionality, the most interesting argument turns on considerations that are specific to intentional states. Among recent philosophers, Donald Davidson and John McDowell have taken a strong anti-reductionism line about such states. Their chief objection is that these states lack a property that is intrinsic to intentional states, their normativity. Intentional states are states that are essentially governed by normative considerations in ways that the states that are studied by natural science are not. Thus when the reductionist aspires to reduce intentional states to neurophysiological states, or to purely causal roles, or to behavior (which is not itself normatively characterized) with a view to bringing them within the purview of the methods and lawlike explanations of natural science, he is aspiring to something that misses out on this intrinsic feature.

Perhaps the simplest way to put the argument, attributed first to Davidson, is that intentional states have to meet the constraints of rationality – which are normative constraints. Thus for instance a belief, in order to be a belief, has to be understood as rational or irrational depending on whether it meets these constraints or not. And the claim is that no such constraints are placed on neurophysiological

states, or purely causal states, or states described in terms of behavior (where the behavior is itself described non-intentionally). We can always ask whether a belief is rational, but we do not have the conceptual vocabulary to ask the same of purely causal, physical, or behavioral states. Questions of rationality are simply not relevant when it comes to these latter kinds of state, whereas they are central when it comes to intentional states. Therefore, it is claimed, there can be no lawlike correlations between these two radically different kinds of state, that would allow us to see the one as reducible to the other. This is essentially Davidson's argument and it has aroused much controversy.

Eliminativism

Apart from reductionism and anti-reductionism, the third view on the matter, eliminativism, does partly concede something to the anti-reductionist. It concedes that it may not be possible to provide lawlike correlations (or 'bridge laws' as they were once called) between intentional states and the physical and other nonnormative states posited by a scientific theory. But it nevertheless resists the conclusion that the anti-reductionist comes to. It foresees that the time will come when neuroscience will advance sufficiently so as eliminate our present epistemic weakness. Then, there will be no need for these convenient attributions (convenient only while we are epistemically weak) which we presently make in intentional terms. There will be nothing for these attributions to describe which will not be better and more accurately described in neuroscientific terms; and anti-reductionists who claim that intentional states have intrinsic properties will therefore be shown to be wrong.

INTERNALISM VERSUS EXTERNALISM

The second topic that continues to attract controversy has been the question of whether intentional content is determined by the world external to the subjects who possess intentionality. Those who think that content is not beholden to the relations in which agents stand to objects and facts in the external world ('internalists') are committed to saying that content is determined wholly by facts internal to the agent, neurophysiological facts as well as the purely causal/dispositional/functional facts. In a sense, this is a modern version of an essentially Cartesian picture of the mind. In Descartes' picture, one's thoughts were determined only by the interior properties of agents; for Descartes one's thoughts were one's thoughts, even if there were no external world. Thus Descartes was an internalist: nothing external was necessary for the determination of thoughts. But Descartes, being a dualist, gave thought an independence from all things physical, so thoughts were also not dependent on any inner physical facts. Modern internalists think that the contents of our thoughts are determined by inner facts including inner physical facts – neurophysiological facts – and can therefore be understood as retaining Descartes' internalism while discarding his dualism.

By contrast, the 'externalists' think that intentional contents are determined at least partly by the causal relations in which we stand to facts and objects in the external world. Thus they think that the Cartesian method of doubt does not yield what Descartes thinks it yielded: thoughts about the external world that are (possibly all) false because there (possibly) is no external world. Rather it yields a *reductio ad absurdum*: we cannot imagine what the method of doubt asks us to; we cannot imagine that there might not be an external world on the grounds that all our thoughts about the external world are false since such thoughts could not have any content (could not, that is, be thoughts) if there were no external world.

Internalism

There are two quite different sorts of motivation that philosophers have articulated for being internalists. One motivation has been purely metaphysical. It has seemed quite intuitive to some philosophers that intentionality is supervenient on the brain. They are thus committed to a rather specific version of supervenience – not merely that intentional content is supervenient on physical facts but supervenient on internal physical facts about the central nervous system. They think that intentional states are reducible to systematically and holistically linked inner functional and causal roles, which are themselves at least supervenient on, if not reducible to, states of the central nervous system. The externalists who oppose them share no such metaphysical intuition. They do not necessarily oppose a general form of supervenience in their metaphysics whereby the intentional is supervenient on the physical generally, but they resist as bad metaphysics the idea that it should be supervenient on the *inner* physical. They think that it is quite intuitive that language and thought are dependent on the nature of the external environment with which we stand in causal relations.

A quite different motivation for internalism is this. Unless intentional states were internalistically

characterized, they would not be able to do what they are manifestly posited to do, encapsulate the inferences, theoretical and practical, that explain our behavior. This motivation is quite separate from any metaphysical considerations. The idea is that the external elements interfere with intentional content and weaken or destroy their explanatory power and potential. This is by far the more interesting motivation for internalism.

Externalism

What is the external element that determines the content of intentional states? Most externalists view this external element as derived from one of two sources.

One source is a causal account of the references for the terms that express the concepts with which we think and which are the constituents of our intentional states. Thus if someone were to think, for example, that 'London is pretty', her concept of London (whether she uses 'London' or 'Londres' or whatever) is determined by certain causal relations in which she stands to its referent (the city). Some philosophers (such as Saul Kripke) think that these causal relations go all the way back to the occasions when the cities, people, objects were originally given these names. Even our current use of a name like 'London' has its meaning determined by the fact of a chain of causal relations which connects us to that city via that original naming event. Concepts and terms which express or describe natural kinds are given corresponding causal accounts. Our current use of a natural-kind term such as 'water' also derives its meaning from its reference or extension; and that reference is determined by the property that exemplifies the essence of a paradigm sample identified in the original reference-fixing event. Other philosophers (such as Dretske and Fodor) do not appeal to the referents of our terms being determined by causal chains, going back to originary events of this kind, but rather to causal covariances between our token uses of our terms or mental tokens and their referents. Thus our term 'water' gets its meaning via the covariance that holds between instances of a substance that has a certain chemical property (H_2O) and our tokenings of 'water'. These are very different causal accounts of the 'ingredients' (concepts, terms) which compose our thoughts and the sentences which we use to express them. But both preclude the possibility of conceiving the content of our thoughts as being independent of the world around us – as the internalists would like.

The other external source externalists invoke as a determinant of content is social. Michael Dummett was among the first proponents of this, and Tyler Burge has developed the idea. According to social externalists, an agent's concepts – as well as the meanings of the terms which he uses to express those concepts – are determined by the linguistic practices of the agent's community. Thus, someone belonging to an English-speaking community, even if he knows no chemistry, nevertheless has a concept of water which the experts in the community characterize in terms of its chemical properties. The individual agent may not know that the chemical formulation of water is H_2O, but nevertheless his concept of water and the term 'water' itself are determined by his linguistic community. (This view differs from a purely causal externalist view, because, unlike that view, it does not have the consequence that English speakers in 1750, say, had the chemical concept of water we now have. The linguistic community in 1750 never had experts who identified the scientific nature or essence of water as H_2O. Because the concept of water is determined by what the experts of the community think at any given time, it is only now that the concept of water includes the chemical property of H_2O.)

Arguments and Counter-arguments

Why do internalists think that if content is determined by an external source then intentional states cannot explain behavior? With regard to the social version of the externalist notion of content, consider someone who lives in a linguistic community different from our English-speaking community in only one respect: their term 'water' is used not as it is used by English speakers to speak of a substance with the chemical property H_2O, but as it is used to talk of a substance which, though it looks just like water and has other properties of water (e.g., it quenches thirst), nevertheless has a different chemical property called 'XYZ'. According to the social externalist view, anyone who is in this community has a different concept of water than do standard English speakers. So, since concepts compose intentional contents, if someone in that community has the belief which we would quite properly report as the 'belief that water quenches thirst', he nevertheless has a different belief than we have when we have a belief that we report as the 'belief that water quenches thirst'. This strictly follows from this externalist view of content.

The internalist objection to this can be brought out by imagining an English speaker and a speaker

from that alternative linguistic community both, say, drinking a glass of something they both call 'water' in order to quench their thirst after a game of tennis. In explaining their respective actions each would invoke the belief that is quite properly reported in each case as the belief 'that water quenches thirst'. Now if the externalist is right, these are nevertheless two different beliefs. But, the internalist says, this difference should make no difference to the explanation of their actions. It is too intuitive that the explanation should in both cases be exactly the same because the difference in the chemical properties of the two substances does not make a difference to their motivational psychology which prompts them to drink. And if beliefs and desires and intentional states generally have a role to play in our motivational psychologies (which is another way of saying that they have explanatory power), then there is no need to have them determined by elements in the external world that would lead to unintuitive results in our thinking about that role.

There is another closely related point which some internalists raise against externalists. It seems that if one allowed that external elements determine intentional contents in the ways mentioned above, an agent may not know his own intentional contents for highly abstract and philosophical reasons. If some English speaker, call him Jim, knows no chemistry at all and believes that the water in front of him will quench his thirst, he will not, according to the externalist, fully *know* what he believes. He will only know it partially and inexactly. (This is a puzzle akin in many ways to the 'puzzle about belief' which Saul Kripke raised about proper names, not natural terms such as 'water'.) But it is very intuitive that we do know exactly what we believe in ordinary cases such as this. No commitment to Cartesian infallibility underlies this intuition. We may grant that there are lots of beliefs, desires, intentions, and motives of ours about which we are fallible and self-deceived, but in each of those cases there is something in the internal psychology of the agent which accounts for her failing to know her intentional states. No such psychological cause or reason can be given in the more ordinary case of Jim not knowing what exactly he believes, in believing that water quenches thirst. He fails to know what exactly he believes, not because of something in his psychology but because some professors of philosophy have a certain view about reference that they think is relevant to the determination of intentional content. According to their viewpoint, this poor man does not have to overcome psychological blocks and censors before he knows what he believes (as is necessary in the cases of self-deception); he has to learn more chemistry before he *knows* what he believes. This seems to many a very implausible consequence of externalism. And, as said above, there is nothing Cartesian in finding this implausible.

For these reasons many philosophers have thought it best to divide intentional states into two quite different sorts: one which they say possesses 'wide' content, and another which possesses 'narrow' content. Wide content captures the point and rationale that the externalist sees in intentionality, while narrow content captures the explanatory point and role of content. Furthermore, it allows us to retain the intuition that unless there are psychological obstacles of some kind, we know our own intentional states. Narrow content is also intended to fulfill the first motivation for internalism, that content should supervene on the physical states of the brain.

In general, this 'solomonic' way of dealing with the issue of internalism versus externalism by bifurcating content into two is the easier and less ambitious way out of these problems. One of the more challenging tasks in this area of philosophy is to pursue the more ambitious goal of trying to capture both what the externalist and the internalist want from content in a single notion of content. In other words, can a genuinely externalist notion of intentional content also be an explanatory notion? And can a genuinely external notion of content retain our ordinary intuitions about our self-knowledge of our thoughts? It may be that the answer to this question is 'yes', but it is unlikely that it can be 'yes' while retaining externalist notions of the orthodox causal and social referentialist sort employed by Kripke, Putnam, Dummett, Burge, and others discussed above.

Further Reading

Armstrong DM (1968) *A Materialist Theory of Mind*. New York, NY: Humanities Press. [An early and comprehensive causal/dispositional/functionalist account of mind.]

Bilgrami A (1992) *Belief and Meaning*. Oxford, UK: Blackwell. [Stresses the relevance of self-knowledge issues to intentional content, and tries to reconcile externalism with self-knowledge and the explanatory properties of intentional states.]

Brentano F (1960) The distinction between mental and physical phenomena. In: Chisholm RM (ed.) *Realism and the Background of Phenomenology*, Atascadero, CA: Ridgeview Publishing. [A pioneering work identifying intentionality explicitly with 'aboutness'.]

Burge T (1979) Individualism and the mental. In: *Midwest Studies in Philosophy IV*. Minneapolis, MN: University of

Minnesota Press. [A rigorous working out of the social version of externalism.]
Churchland PM (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**. [A prominent contemporary eliminativist.]
Davidson D (1970/1981) Mental events. In: *Essays on Actions and Events*. Oxford, UK: Oxford University Press, 1981. [A very interesting argument for the irreducibility of intentionality on normative grounds.]
Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press. [A central book of essays on the subject of this article.]
Descartes R (1911) *Meditations on First Philosophy. The Philosophical Works of Descartes, vol. 1*, edited by ES Haldane and GRT Ross. Cambridge, UK: Cambridge University Press. [The pre-eminent internalist. For better or worse, the wellspring of much of modern philosophy of mind.]
Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press. [An early version of the causal account of intentional content developed by Fodor (see below).]
Dummett M (1978) The social character of meaning. In: *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press. [The original statement of a social version of externalism.]
Fodor J (1987) *Psychosemantics*. Cambridge, MA: MIT Press. [A rigorous and ingenious naturalist position.]
Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press. [A pioneering statement of the causal theory of reference.]
Kripke S *Wittgenstein On Rules and Private Language*. Cambridge, MA: Harvard University Press. [A far clearer statement of the irreducibly normative character of intentionality than to be found in Wittgenstein himself.]
Lewis D (1972) Psychophysical and theoretical reduction. *Australian Journal of Philosophy* **L**:3. [A rigorous version of functionalism.]
Loar B (1985) Social content and psychological content. In: Grimm R *et al*. (eds) *Contents of Thought*. Tucson, AZ: University of Arizona Press. [A beautifully clear statement of a bifurcated account of intentional content into wide and narrow content.]
Loar B (1981) *Mind and Meaning*. Cambridge, UK: Cambridge University Press.
McDowell J (1986) Functionalism and anomalous monism. In: Lepore E and Mclaughlin B (eds) *Perspectives on Actions and Events*, Oxford, UK: Blackwell.
Putnam H (1978) The meaning of meaning. In: *Mind, Language and Reality: Philosophical Papers* vol. 2, Cambridge, UK: Cambridge University Press. [A pioneer of the causal theory of reference with an explicit statement of its relevance to intentionality.]
Quine WVO (1959) *Word and Object*. Cambridge, MA: MIT Press. [An early and distinguished eliminativist.]
Rosenthal D (ed.) (1988) *The Nature of Mind*. Oxford, UK: Oxford University Press. [The best and most comprehensive anthology of the philosophy of mind, with extensive sections relevant to intentionality.]
Ryle G (1949) *The Concept of Mind*. London, UK: Penguin. [An early, zealous and clear anti-Cartesian work.]
Searle J (1987) *Intentionality*. Cambridge, UK: Cambridge University Press. [A consistently internalist position.]
Skinner BF (1957) *Verbal Behaviour*. New York, NY: Appleton Century Crofts. [A unrepentant behaviourist in the trenches.]
Watson J (1919) *Behaviourism*. Chicago, IL: University of Chicago Press. [A fierce reductionist version of behaviourism.]
Wittgenstein L (1953) *Philosophical Investigations*. Oxford, UK: Blackwell. [A fascinating and profound, if frustrating, work.]

Introspection

Intermediate article

William Lyons, Trinity College, Dublin, Ireland

CONTENTS

Philosophers and psychologists have both appealed to introspection as a privileged and direct method of gaining knowledge about the mind and its activities. What exactly introspection is, and what role it might play in a scientific psychology, are subjects of major debate.

INTRODUCTION

The term 'introspection' might be defined as 'the direct, conscious examination or observation by a subject of his or her own conscious mental processes'. Etymologically the term is derived from two Latin words, *intro*, meaning 'within', and *specere*, meaning 'to look at'. Throughout the long history of philosophy and, more recently, psychology, those investigating the mind have on occasion adopted introspection as an obvious, if subjective, method of gaining immediate knowledge of the mind at work.

HISTORY

It has been claimed that the ancient Greeks, at least up to and including Aristotle (384–322 BC), did not have any conception of introspection; though there is a section in the *De Anima* where Aristotle may be saying that we can reflect on our own mental states (Aristotle, 1986, book 3, chap. 4, 429b–430a). The first indisputable employment of any concept of introspective self-consciousness is probably by the Neoplatonist philosopher Plotinus (*c*. 204–270). In his essays, which were collected after his death by his disciple Porphyry as the *Enneads*, Plotinus discusses whether the One (or Absolute, the source of all existence) is self-conscious or not. Plotinus argues that the One could not be self-conscious in the usual sense because that would imply a dualism of knower and known (Plotinus, 1991, Ennead 5, tract 3). There is also discussion of introspective self-consciousness in the works of another Neoplatonist, the African Bishop of Hippo, Augustine (353–430). In his *De Trinitate*, Augustine (1955, vol. VIII, book 10, section 7) remarks: 'What then can be the purport of the injunction, Know thyself? I suppose it is that the mind can reflect upon itself.' (*See* **Self-consciousness**)

This claim, that the mind can reflect upon itself, was fairly common among the medieval philosophers, particularly among those who, like the Frenchman William of Auvergne (*c*. 1180–1249), were to some degree Augustinian. They often argued for the immateriality of the soul, and for claims about the nature of human knowledge, on the basis of the deliverances of self-consciousness. The most famous of all medieval thinkers, the Italian philosopher and theologian Thomas Aquinas (*c*. 1225–1274), was explicit about our possession of a capacity for self-consciousness. Quoting Augustine in support, he asserts in his *Summa Theologiae* that 'the mind perceives itself' (Aquinas, 1968, Part 1, question 87, article 1).

In the seventeenth century, the English political philosopher Thomas Hobbes (1588–1679), despite a commitment to materialism, laid great stress on the value of self-reflection. Though Hobbes himself did not use the term, it was in the seventeenth century that we find the first use of the word 'introspection' in English. Adopting the maxim of the Delphic oracle, Cicero, and Augustine, *nosce teipsum* ('know thyself'), Hobbes argued that only through self-reflection would humans be able to discover their own psychology and, on that basis, be in a position to 'read and know, what are the thoughts, and Passions of all other men' (Hobbes, 1996). Only when possessed of this knowledge of human psychology could anyone hope to produce an acceptable moral and political program for a civilized society.

Arguably, however, it was the French philosopher, mathematician, and scientist René Descartes (1596–1650) who introduced introspection as a

topic for discussion in modern philosophy and whose ideas ultimately led to introspection being chosen as the method for the new science of psychology in the latter part of the nineteenth century. Descartes not only believed firmly in the capacity of a human subject to gain direct knowledge of his or her own mind via introspection; he placed immense value upon that knowledge. He held that, at least when our introspections produced immediate, clear and distinct knowledge of the contents of our stream of consciousness, that is to say, of our own mind, then that knowledge was, indeed must be, infallible. For that reason introspection should be made the basis of all human knowledge (Descartes, 1954, part 4). (*See* **Descartes, René**)

In the latter half of the seventeenth century, when the English philosopher John Locke (1632–1704) put forward a new philosophy, more consonant with the new observation-based approach to science, namely empiricism, he retained the method of introspection. While claiming that the basis for all our knowledge was the information that came into the human mind via the senses, in his *Essay Concerning Human Understanding,* Locke pointed out that there was an important set of ideas and concepts, such as those of the intellect and will, which could be gained only by 'reflection', his term for introspection (Locke, 1924, book 2, chap. 6). While Locke's conception of psychology was soon challenged, notably by the Irish philosophers Peter Browne (1666–1735) and George Berkeley (1685–1753) and by the German philosopher Gottfried von Leibniz (1646–1716), his *Essay* was the most influential book on human psychology throughout Europe for the next hundred years.

In the late eighteenth and early nineteenth centuries, philosophy appears, at least in retrospect, to have been dominated by the work of the German philosopher Immanuel Kant (1724–1804), who, in his own words, produced a kind of Copernican revolution in epistemology. Copernicus had pointed out that it was in fact the movement of the Earth that produced the apparent movement of the Sun and stars around us, not a movement of Sun and stars around a stationary Earth. In a similar act of reversal, aimed at empiricism, Kant stressed that human knowledge was not a passive reception on the part of the senses of information about objects and events in the world. It was an active creation by the mind of such fundamental categories as 'object' and 'event' out of the material of sensations. For Kant, self-consciousness, or 'apperception', a term he borrowed from Leibniz, had a role to play in our knowledge of these categories. Apperception had two forms, an ordinary empirical form and what he referred to as a transcendental form (Kant, 1993, Transcendental Logic: First Division, book 1, chap. 2). In its empirical form, apperception was more or less Cartesian or Lockean introspection. It gave us an awareness of the ever-changing thoughts that were the contents of our own consciousness. In its transcendental form, however, it gave us an insight into the very nature of thought itself. This insight included an inchoate apprehension of the unity of the self that made connected and so logical thought possible, and thereby created the possibility of human understanding. One of Kant's criticisms of Descartes was that he had coalesced and confused these two quite different forms of apperception. (*See* **Kant, Immanuel**)

In comparison with both Cartesian and empiricist philosophy, introspection, or empirical apperception, had a minor role in Kant's philosophy. He was really interested in transcendental apperception. Thus his influence was mainly on idealist philosophy rather than on empirical psychology. It was the return to a more Cartesian approach, particularly in the work of the German philosopher Franz Brentano (1838–1917), that brought introspection back to the forefront of philosophical discussion and led eventually to the foundation of the new science of psychology.

Though influenced by the work of Aristotle as well as that of Descartes, Brentano thought of himself as an empirical psychologist as well as a philosopher, as the title of his masterpiece, *Psychology from an Empirical Standpoint* (Brentano, 1973), makes clear. For Brentano, there were two sorts of psychology, both empirical. There was physiological psychology – such as was being practiced in Germany by Ernst Weber (1795–1878), Gustav Fechner (1801–1887) and Hermann von Helmholtz (1821–1894) – and there was descriptive psychology. The latter, sometimes also called by Brentano 'descriptive phenomenology', was a program of gaining knowledge about the 'ultimate elements' of the human mind by introspection, much in the way Descartes would have envisaged. One first introspected, and then took as knowledge about the mind itself whatever presented itself introspectively as 'evident' or 'self-evident': this was Brentano's version of Descartes' 'clear and distinct' ideas. Introspection was an empirical process because it was a direct experience of the mind. It was the mind as eyewitness of itself. But Brentano added a signficiant qualification to this Cartesian approach. He distinguished carefully between 'inner perception' (or the passive, 'out of the corner of the mental eye', process of self-consciousness) and

'inner observation' (or the active process of introspection). He felt that only the former was guaranteed not to interfere in what it was observing. He also pointed out that inner perception, although a passive and 'by the way' form of noticing, could be improved by training. The psychological introspectionists were to take this idea very seriously. (*See* **Fechner, Gustav Theodor**)

Besides being a powerful personality, Brentano must have been a remarkable teacher, because many of his students did remarkable work. His pupils included Freud, the founder of psychoanalysis, Tomáš Masaryk, a philosopher who became president of Czechoslovakia, Alexius Meinong, who became an influential philosopher, Carl Stumpf, a gifted musician and psychologist who wrote an important text on the psychology of sound, and Christian von Ehrenfels, one of the founders of Gestalt psychology. (*See* **Freud, Sigmund**)

From the point of view of the history of introspection, Brentano's most important pupil was the German philosopher Edmund Husserl (1859–1938). While Husserl made introspection the basis of his new science of phenomenology, he also pushed introspection further away from the empirical world of ordinary experience. For Husserl, phenomenological investigation, while it might begin with the introspection of the contents of consciousness, was essentially a non-empirical endeavor. It was an *a priori* contemplation of conscious experience as such, in order to understand its essence. In a phenomenological investigation of the experience of seeing a red tomato, for example, you must first set aside and so transcend any knowledge that there is a tomato present, or that it is one that you grew in your vegetable garden, or anything else particular about it. Then you must seek to penetrate to the pure experience of redness itself, the 'universal redness', while at the same time gaining an insight into the nature of the subjectivity involved in having a conscious experience. In effect, Husserlian phenomenological investigation was a return to a Kantian transcendental apperception with the addition of some Platonic contemplation of forms (Husserl, 1999, chap. 19). (*See* **Phenomenology**)

There is some speculation as to why, when the new science of psychology came into being, it became an introspective psychology rather than, say, a physiological psychology in the tradition of Fechner, Weber, and Helmholtz. A plausible reason is that the founders of the new science of psychology, such as Wundt and Stumpf, believed that a truly autonomous psychology must be a study of the mind as such, and so distinct from physiology. They believed that the foundations for this new science of psychology already existed, for example, in the philosophical psychology practiced by Franz Brentano in Germany and John Stuart Mill (1806–1873) in England. To turn the old philosophical psychology into the new science of psychology, one just had to make introspection scientific. This transformation was to be effected by regulating the process of introspection according to good experimental practice. In turn, this was to be achieved by carefully classifying every aspect of the introspective process, by making use of laboratories with specially trained subjects and scientific instrumentation, and by repeating experiments many times. As William James put it, 'these new prism, pendulum and chronograph-philosophers … mean business, not chivalry' (James, 1950, vol. I, pp. 192–193). (*See* **James, William**; **Helmholtz, Hermann von**)

Wilhelm Wundt (1832–1920) is generally regarded as the founder of the new, introspective, science of psychology. Wundt studied medicine and was for a very short time an assistant in physiology to Helmholtz at the University of Heidelberg. However, Wundt was dissatisfied with his physiological studies and, in his first book, on perception (published in 1862), he coined the term 'experimental psychology' for the introspective approach to knowledge of the mind. (*See* **Wundt, Wilhelm**)

After he took the chair of philosophy at the University of Leipzig, the central theme of Wundt's lectures was that the correct method for this new experimental psychology was introspection organized along scientific lines, and that its goal should be the analysis of conscious experience into its elements, and the discovery of the laws of association of these elements. He claimed that his experiments showed that there were two sorts of element, sensational (or pertaining to the senses) and affective (or feeling). Wundt established the first ever psychological laboratory (Wundt, 1896, lectures 1 and 2; Schultz and Schultz, 1996, chap. 4). Through Wundt's teaching, this new introspective psychology was spread around the world. It flourished until behaviorism – which had no use for the notions of introspection or consciousness – dictated a new method for a science of psychology. The ascendance of this new behavioristic method in psychology, and the related behavioristic analyses in the philosophy of mind, was accompanied by severe criticisms of the employment of introspection in psychology (Watson, 1995, 1930). (*See* **Behaviorism, Philosophical**)

Now that behaviorism has been largely superseded and cognitive psychology is widely studied,

there is again much discussion of the use of introspection in psychology. There is a renewed interest in consciousness and an admission of the conspicuous failure of either psychology or the brain sciences to tell us much about it. In 1979, the journal *American Psychologist* published an article by D. A. Lieberman (1979) entitled 'Behaviorism and the mind: a (limited) call for a return to introspection'. Introspection is now sometimes employed in experiments where subjects are asked to relate their feelings or thoughts about something that is going on around them, but there has been no return to the introspective laboratories of Wundt, Titchener, and Kulpe. There is also little interest in renewing discussion of introspection as a psychological method, because it is believed that certain objections to introspective psychology (see below) have never been satisfactorily answered.

There have been similar developments in philosophy. After behaviorism, the majority of philosophers adopted the view that the mind was to be identified with brain states or brain functions. They retained a traditional role for introspection, but translated the two-level Cartesian and Wundtian account of introspection into wholly physical terms. The two-level, Cartesian, process of conscious observer and observed became one part of the brain scanning another, or one part of the brain accessing another, similarly to one part of a computer accessing another. A limited revival of interest in a more orthodox account of introspection came in the wake of a major revival of interest in consciousness itself. This revival came about through the work of several philosophers of mind, notably the Americans Thomas Nagel (1995) and John Searle (1992), who claimed that any theory of mind that tried to reduce the subjectivity of consciousness to non-subjective brain states or functions was misguided.

PHILOSOPHICAL DISCUSSION OF INTROSPECTION

In philosophy, in orthodox Cartesian terms, introspection would have been analyzed in terms of a subject's capacity to observe, in a non-sensuous, private, privileged, incorrigible, subjective and, arguably, comprehensive fashion, events in that person's own stream of consciousness. Until comparatively recently, much of this analysis was still being defended. (*See* **Consciousness, Stream of**)

The process of introspection was held to be non-sensuous because it was not to be considered as an additional sense to be added to those of taste, touch, hearing, smelling, and sight, and it was not to be associated with any sense organ. Although it was a real form of observation, it was held to be one that was stripped of sensory features. Introspection was considered to be private because it was part of a person's inner private life of the mind, as opposed to that person's publicly observable life of behavior, gesture, expression, and speech. Introspection was considered to be privileged because the introspecting subject alone had this access to his or her own stream of consciousness. Introspection was considered to be incorrigible because introspective reports, being private and privileged, were not open to correction by anyone else. Some philosophers went further and claimed that introspection was infallible. Being immediate, simple and direct, there was nothing that could ever go wrong with the introspecting process. The objects of such direct knowledge were therefore held to be 'self-intimating' and the resulting beliefs about them 'self-verifying'. Also, given the requisite command of language, an introspector could not misreport on an introspection. A few philosophers have claimed that introspection was comprehensive, in the sense that all mental events were in principle introspectible. This view made sense, of course, when it was prefaced with the claim that the mind was coextensive with consciousness. Finally, introspection was considered to be subjective. It was subjective epistemologically, that is, as a type of knowledge, because it only provided knowledge to a single subject in respect of that subject's own stream of consciousness. It was subjective ontologically, that is, as regards its mode of existence, because it only existed as a conscious activity of a conscious subject. When a subject was asleep or in a coma, then the subjective point of view, including the introspective one, ceased to exist.

Many of the above claims have been challenged by philosophers, particularly in the latter half of the twentieth century. Many philosophers have pointed out that introspective reports are not incorrigible, especially in relation to reports of complex mental states. For example, at the conference cocktail party, Fred might claim, sincerely, that he must have spilt his drink because he was astonished by anyone in the twenty-first century defending introspectionism in psychology. Everyone else realized that he spilt his drink because he was angry and embarrassed because it was Mary, whom everyone knew had recently divorced him, who was defending introspectionism. Philosophers have also pointed out that introspection cannot be strictly comprehensive in its knowledge of the mind: since the time of Freud, it has become a common belief that many of our strongest desires

and beliefs and motivations exist and operate below the threshold of our consciousness.

The most important of these criticisms, however, have been about the very structure of the orthodox Cartesian account of introspection. This orthodox account is a two-level account. It analyzes introspection into an activity of consciously observing a conscious process. Philosophers have argued that this splitting of consciousness into observer and observed is bound to cause problems in regard to the introspection of complex mental processes. If the introspective task is to observe one's own conscious process of, say, multiplying 2634 by 1292, then, the more one concentrates on the introspecting, the more likely one is to make a mistake with the mental arithmetic.

Philosophers have also argued for a judicious employment of Ockham's razor, or the principle of parsimony, which, in this context, dictates that one should not multiply levels of consciousness beyond theoretical necessity. Neither the introspection of complex mental processes nor that of simple mental states requires a two-level explanation. The introspection of complex mental processes can be adequately explained as a replay, via short-term memory, of just the first-level mental process. To introspect your mental arithmetic is simply to recall the steps you made in arriving at your answer, immediately after they have occurred. The introspection of simple mental states can be explained in terms of focusing one's attention by banishing distractions. To introspect the pain in your stomach, in response to your doctor's query, is to concentrate your attention on your stomach in order to say how it feels.

The above is a 'middle of the road' account of current philosophical views of introspection. In fact there are many variations, and considerable debate about a number of its claims. For example, when rejecting the two-level account of introspection, some, notably Sydney Shoemaker (1996), have argued for a one-level account in terms of a concept of self-intimation. Shoemaker argues that there is no such thing as introspective perception of the self, or indeed of anything else. On the other hand, we do have various forms of self-knowledge. One central form of self-knowledge – indeed, the one that people mistakenly interpret in terms of a two-level account of introspection – occurs in so far as certain very basic mental states, such as being in pain, are self-intimating. By 'self-intimating', Shoemaker means that there is an internal connection between being in a mental state of that kind and having, or at any rate being disposed to have, the introspective belief that one has it. In short, 'introspection' is just the misleading name we give to the fact that some mental states imply others: to be in a certain mental state is *ipso facto*, and necessarily, to generate a meta-belief that one is in that mental state. Shoemaker's argument involves the claim that the proper functioning of human rationality in normal humans requires this to be so.

In contrast to this (by now fairly common) rejection of any two-level account of introspection, David Rosenthal (e.g. 1997) argues robustly for a return to a two-level account of introspection. He argues that, while introspection is not a form of inner perception, it is a form of inner, second-level attention. When one introspects, say, one's mental act of thinking about the conference dinner last night, one is both thinking about the dinner and thinking about one's thinking about the dinner. Furthermore, such second-level thinking is a conscious, occurrent process. Thus introspection consists in conscious, attentive, higher-order thinking about conscious, attentive, ground-level thinking, and introspective reports are therefore real reports about real inner events.

Any detailed interpretation of introspection is bound to be affected by the current debate in the philosophy of mind about 'internalism' (the view that all mental states are wholly identifiable in terms of their inner, nonrelational characteristics) and 'externalism' (the view that mental states, or at least those that involve things like thoughts, are partly identifiable in terms of the meanings of their contents, which in turn involve a relation to some external, communal, states of affairs). In particular, this debate has a bearing on the authority of the introspecting subject's account of his or her own mental states. If some version of externalism about mental states is correct, then a subject can much more easily be wrong about the nature of his or her own inner mental states than if internalism is correct (e.g. Davidson, 1994; Burge, 1994). (*See* **Externalism**)

PSYCHOLOGICAL DISCUSSION OF INTROSPECTION

Whereas most philosophical discussion has been about the theory of introspection, psychological discussion has focused on introspection as a scientific method.

In the first decades of the twentieth century, psychologists questioned whether introspection could or should be part of any scientific psychology. They argued that only objective methods were truly scientific. By this they meant that

scientific experiments must be both observable by two or more people as they are taking place, and repeatable in every important respect. *Ex hypothesi*, introspective experiments could involve only one observer. Strictly speaking, they were also unrepeatable, because it was impossible to ensure that a subject's inner mental acts, which comprise both the observing and observed aspects of the introspective experiments, were similar each time the experiment was repeated.

Psychologists also pointed out that introspection experiments produced unreliable data. There were no agreed results. While there might be almost unanimous agreement about some introspection finding among the psychologists of one laboratory, this might be disputed by those from another introspection laboratory. Notoriously, there was the dispute between those who asserted that their introspection experiments found that conscious episodes of thinking always employed some sensory medium, and so were in the form of heard speech, visualized representations or something similar, and those who asserted the opposite. There were many other disputes along the same lines, which began to weaken the confidence of psychologists in the value of data from introspection experiments.

Some psychologists also conducted experiments about the very possibility of splitting consciousness in the way that seemed to be demanded by the orthodox two-level account of introspection. Although this matter is still disputed, many psychologists believed that their experiments showed that we cannot attend carefully to two cognitive tasks at the same time. This is so, whether the tasks are driving through heavy and volatile traffic on an unfamiliar route while at the same time discussing the theory of introspection with a fellow passenger, or pondering upon Pythagoras' theorem while at the same time introspecting it in such a way as to be able to report on it. The best one can do is to oscillate one's attention, one's focus of consciousness, between the two tasks, or else focus on just one of the tasks and try to accomplish the other automatically.

THE ROLE OF INTROSPECTION IN COGNITIVE SCIENCE

Leaving aside the debates about the correct explanation of what is going on when we introspect, and about the status and reliability of the data of introspection, it is certain that we report on and express our inner conscious states. We do talk about our thoughts, our pains, our hopes, our plans, our beliefs and our desires, and many other inhabitants of our inner life of the mind. Sometimes, too, we display or express our conscious states, say, of pain or disgust, in our gestures or facial expressions or body postures. Both our reports and our expressions are clearly data of a psychological sort, and so cannot be ignored in any reasonable cognitive science.

Few would dispute that part of any complete psychological account of someone's being in pain, or being disgusted, or being angry, or being depressed, or being puzzled, or being astonished, includes an account of the outer expression – in gesture, posture, facial expression, tone of voice, and behavior – of inner conscious states. It is the strictly introspective reports of inner conscious states or processes that have always raised doubts.

The best way of treating an introspective report of an inner conscious state may be to consider it as having more or less the same status as a single independent eyewitness account of some event. Eyewitness acounts are useful but need to be treated with circumspection. Take the case of a single eyewitness report of a road accident. The investigating police officer would want to check it, if he or she could, by reference to more objective evidence. If the eyewitness said that the car skidded and then drove straight into a tree on the left-hand side of the road, the police officer might note both that there were no skid marks on the road and that the area of initial impact was not the front of the car. Again, an introspecting patient might tell the doctor, quite sincerely, that he has a heart pain but, after some tests, the doctor might tell him that it is merely indigestion. Or an introspecting patient in a psychotherapy session might say, again quite sincerely, that it was a lack of interest that made her skip her psychology lectures on motivation, whereas the therapist might realize, from the patient's past history, that the reason was agoraphobia.

Nevertheless the patient in the doctor's surgery and the one in the psychotherapy session have revealed something about themselves. In the first case, the patient directly revealed that he was in pain, and indirectly that he had a fear of heart attack. In the second case, the patient directly revealed that she had a reluctance to go to her psychology lectures, and indirectly that she was in denial as regards her previous history of agoraphobia.

Cognitive science needs to take account of all available data, including introspective reports, when building up a complete picture of a human's cognitive life. (*See* **Metacognition**)

References

Aquinas T (1968) *Summa Theologiae*. In: Durbin PT (ed. and trans.) *Human Intelligence*, vol. XII. New York, NY: McGraw-Hill. [Written *c*. 1265–1273.]

Aristotle (1986) *De Anima*, translated by H Lawson-Tancred. Middlesex, UK: Penguin. [Written in the fourth century BC.]

Augustine (1955) *De Trinitate*. In: Burnaby J (ed. and trans.) *Augustine: Later Works*. London, UK: SCM Press. [Written *c*. 410.]

Brentano F (1973) *Psychology from an Empirical Standpoint*. Kraus O and McAlister L (eds), Rancurello AC, Terrell DB and McAlister L (trans.). London, UK: Routledge and Kegan Paul. [First published 1874.]

Burge T (1994) Individualism and self-knowledge. In: Cassam Q (ed.) *Self-knowledge*, pp. 65–79. Oxford, UK: Oxford University Press.

Davidson D (1994) Knowing one's own mind. In: Cassam Q (ed.) *Self-Knowledge*, pp. 43–64. Oxford, UK: Oxford University Press.

Descartes R (1954) *Discourse on Method*. In: Anscombe E and Geach PT (ed. and trans.) *Descartes: Philosophical Writings*, pp. 31–37. Edinburgh, UK: Nelson.

Hobbes T (1996) The Introduction. In: Gaskin JCA (ed.) *Leviathan*, pp. 7–8. Oxford, UK: Oxford University Press. [First published 1651.]

Husserl E (1999) Phenomenology. In: Welton W (ed.) *The Essential Husserl: Basic Writings in Transcendental Phenomenology*, pp. 322–336. Bloomington, IN and Indianapolis, IN: Indiana University Press. [First published 1927.]

James W (1950) *The Principles of Psychology*. New York, NY: Dover. [First published 1890.]

Kant I (1993) *Critique of Pure Reason*. Politis V (ed.) and Meiklejohn JMD (trans.). London, UK: Dent. [First published 1890.]

Lieberman DA (1979) Behaviorism and the mind: a (limited.) call for a return to introspection. *American Psychologist* **34**: 319–333.

Locke J (1924) *An Essay Concerning Human Understanding*. Pringle-Pattison AS (ed.) Oxford, UK: Clarendon Press. [First published 1690.]

Nagel T (1995) What is it like to be a bat? In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 159–174. London, UK: Dent. [First published 1974.]

Plotinus (1991) *The Enneads*. Dillon J (ed.) and MacKenna S (trans.). London, UK: Penguin. [Written *c*. 305.]

Rosenthal D (1997) A theory of consciousness. In: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*. pp. 729–753. Cambridge, MA: MIT Press.

Schultz DP and Schultz SE (1996) *A History of Modern Psychology*, 6th edn. Fort Worth, TX: Harcourt Brace. [First published 1969.]

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: Bradford/MIT Press.

Shoemaker S (1996) *The First-Person Perspective and Other Essays*. Cambridge, UK: Cambridge University Press.

Watson JB (1930) *Behaviorism*, 2nd edn. Chicago, IL: University of Chicago Press. [First published 1924.]

Watson JB (1995) Psychology as the behaviorist views it. In: Lyons W (ed.) *Modern Philosophy of Mind*, pp. 24–42. London, UK: Dent. [First published 1913.]

Wundt W (1896) *Lectures on Human and Animal Psychology*, trans. JE Creighton and EB Titchener. London, UK: Swan Sonnenschein and New York, NY: Macmillan. [First published 1863.]

Further Reading

Alston WP (1971) Varieties of privileged access. *American Philosophical Quarterly* **8**(3): 223–241.

Armstrong DM (1994) Introspection. In: Cassam Q (ed.) *Self-Knowledge*, pp. 109–117. Oxford, UK: Oxford University Press.

Danziger K (1980) The history of introspection reconsidered. *Journal of the History of the Behavioural Sciences* **16**: 241–262.

Ericsson KA and Simon HA (1980) *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Gopnik A (1993) How do we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**(1): 1–14.

Hebb DO (1969) The mind's eye. *Psychology Today* **2**: 55–68.

Ludlow P and Martin N (eds) (1998) *Externalism and Self-knowledge*. Stanford, CA: CSLI.

Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Lyons W (1986) *The Disappearance of Introspection*. Cambridge, MA: Bradford/MIT Press.

Moran D (2000) *Introduction to Phenomenology*. London and New York, NY: Routledge.

Nisbett R and Wilson T (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review* **84**(3): 231–259.

Ryle G (1949) *The Concept of Mind*, chap. VI. London, UK: Hutchinson.

Shoemaker S (1988) On knowing one's own mind. *Philosophical Perspectives* **2**: 183–209.

Inverted Spectrum

Intermediate article

Martine Nida-Rümelin, University of Fribourg, Fribourg, Switzerland

CONTENTS

If person A sees as green what person B sees as red, and person A sees as red what person B sees as green, and if some other simple conditions are met, A is said to have an inverted spectrum with respect to B.

WHAT IS AN INVERTED SPECTRUM?

Is it possible that one person sees as green what another sees as red, and, equally, that the first sees as red what the second sees as green? If so, then the difference would also apply to mixed colors: surfaces that appear greenish to the first person (e.g. turquoise) would appear reddish to the second person (violet) and vice versa. Both people would have learned to call red things 'red' and green things 'green', so the difference between them might go unnoticed. This is a case of an inverted spectrum (IS). In more precise terms, A has an *inverted spectrum* with respect to B if

(C1) A and B have the same overall set of color experiences, but
(C2) things appear in color systematically different to A and B, and
(C3) the difference in color perception between A and B does not and cannot manifest itself in behavior.

The claim that cases of IS are possible has been called the 'Inverted Spectrum Hypothesis' (ISH). The ISH has received extensive discussion in the philosophical literature but very little discussion within empirical color vision science. There is no commonly accepted view about whether the ISH is true or false (in any of the possible interpretations distinguished below).

The definition above captures the notion of an *inter*subjective IS (the difference exists between different perceivers). *Intra*subjective IS concerns the same individual at different times. An intrasubjective spectrum inversion would normally manifest itself in behavior. In the following, IS without a qualification stands for intersubjective IS.

DIFFERENT VERSIONS OF THE INVERTED SPECTRUM HYPOTHESIS

Sometimes a weaker notion of IS is used in the literature, a notion that only requires the conditions C1 and C2. Most authors require something like condition C3 for a case of fully-fledged IS, but often it is not quite clear precisely what third condition they have in mind. At least three different versions of the third defining condition are frequently assumed and should be distinguished:

(C3a) The difference does not and cannot manifest itself in behavior with respect to normal life situations.
(C3b) The difference does not and cannot manifest itself in behavior of any kind (including behavior in sophisticated psychophysical experiments).
(C3c) There is no relevant functional difference: although person A and person B have inverted color sensations, the inversion leaves the causal roles of color sensation untouched.

If, for example, A and B are red–green inverted with respect to each other, then (according to condition C3c) red experiences occupy the same causal role in A as green experiences occupy in B (they have the same causes, interact in the same way with other mental states and lead to the same kind of behavior). C3c could be further subdivided according to different accounts of causal role and correspondingly different notions of functional difference.

The claim that cases of IS are possible (the ISH) has different content depending on which of these conditions is implicitly presupposed. In addition, the notion that such a possibility is involved has at least three different interpretations: the logical possibility, the nomological possibility, and the metaphysical possibility. IS is logically possible if there is no hidden contradiction in its assumption (if a case of IS can be coherently conceived), it is nomologically possible if the occurrence of a case of IS is compatible with the laws of nature. Metaphysical

possibility is a possibility in a mind-independent way and thus may be called a 'real' possibility as opposed to a merely apparent possibility (but it does not coincide with a nomological possibility; many situations that are excluded by the laws of nature are considered metaphysically possible). According to many philosophers some situations are conceptually possible but metaphysically excluded. A famous example is a situation where the Morning Star is not identical with the Evening Star. This is in some sense conceptually possible, but not 'really' possible since Venus is necessarily identical with Venus. Combining the three different conditions (C3a–C3c) that may be presupposed in the ISH with the three possible interpretations of possibility, we get six variants of the ISH. Each must be judged on the basis of quite different considerations.

PHENOMENAL STRUCTURE AND COLOR INVERSION

The way a red–green inverted person would see the color circle (see Figure 1) can be shown by exchanging F1 with F9 and by rearranging the binary hues accordingly.

Other types of inverted color vision can be represented in the same way by permuting the colors of the circle in Figure 1. But not every arbitrary permutation of colors in the circle represents a possible case of IS. Just exchanging F1 and F2 would represent a color inversion that certainly would manifest itself in verbal behavior (the person would make different color similarity judgments and different judgments about color composition). The transformation that consists in a simple exchange of F1 and F2 does not preserve the relations that obtain in the appearance of the colors; it does not preserve what has been called 'phenomenal structure'. Aspects of phenomenal structure concern the similarity of colors, the relation of being phenomenally composed or mixed of two basic colors, the relation of being complementary to each other, and many other features. (Phenomenal mixture or composition should not be confused with light mixture or pigment mixture. A color is phenomenally composed of two components if it *appears* to contain these two colors). If one restricts the set of permutations that represent possible candidates for an IS to those that preserve phenomenal structure, then only those permutations remain that map each unitary color into a unitary color and rearrange the binary colors accordingly. These permutations are obtained by appropriate rotations (90°, 180° or 270°) or by appropriate reflections (vertical or horizontal axis, axis through F3 and F11, or axis through F15 and F7). All the other permutations violate the above mentioned constraint: they do not preserve phenomenal structure for obvious reasons (similarity relations or relations of phenomenal composition would be changed).

The truth or falsity of the ISH thus depends on whether there are color transformations that preserve phenomenal structure. The latter question is a lot more complicated than the above simple description might suggest. Phenomenal structure

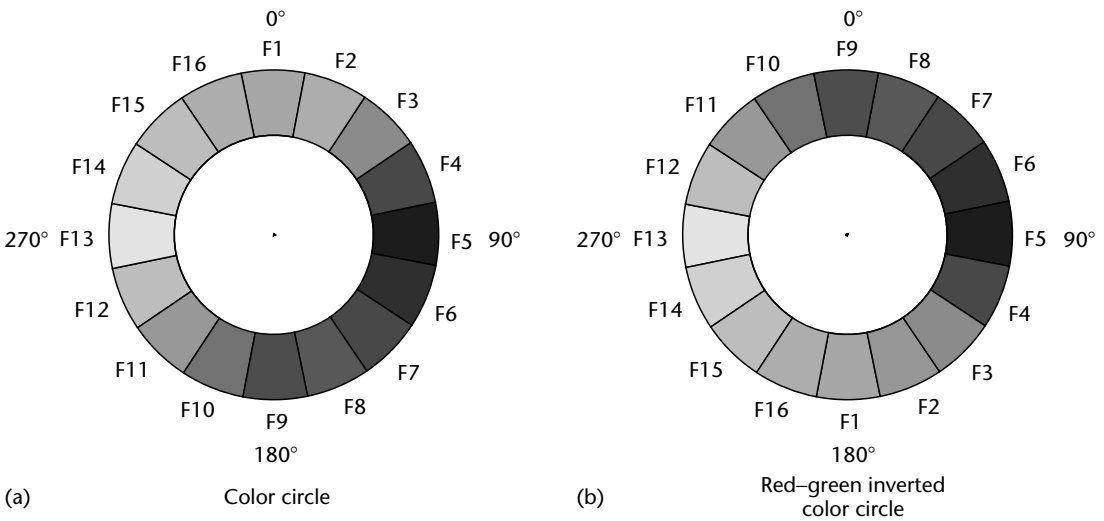


Figure 1. [*Figure is also reproduced in color section.*] (a) Circle representing color spectrum. (b) Circle representing inverted spectrum.

must be described in a way that takes further aspects of colors into account (the third dimension: black–white or brightness), saturation, and maybe further features such as the warmth (coolness) of colors and their role in depth perception.

Some aspects of phenomenal structure may, however, be due to of learning processes (maybe red is experienced as warm because it is associated with fire or experienced as alarming because it is associated with blood) and to sociocultural factors. If these features are not preserved by some color transformation, the transformation may nonetheless represent a metaphysically, nomologically, and logically possible case of IS (in all three senses distinguished above by C3a–C3c).

HISTORY OF THE INVERTED SPECTRUM

The idea of an intersubjective IS occurs in John Locke (1690, book 2, ch. 32, sec. 15). In the first decades of the twentieth century ISH was a favorite example of logical positivists to explain, illustrate, and discuss their verificationist theory of meaning. The idea of intrasubjective IS occurs in the writings of Wittgenstein in the context of his private language argument. In contemporary philosophy IS is of particular interest as a potential counterexample to materialist theories about the nature of phenomenal consciousness.

PHILOSOPHICAL ISSUES ABOUT THE INVERTED SPECTRUM

The ISH plays a prominent role in the debate about what makes an experience an experience of a specific phenomenal (or qualitative) kind, e.g. about what it *is* that red experiences have in common. (Talk of 'red experience' is not meant to imply that the experience is literally red. The expression is used to refer to a specific phenomenal kind of experience.) Phenomenal kinds are distinguished by how it is to be in the state at issue (by their subjective character). The properties that different instances of one phenomenal kind (e.g. the kind specified by 'being a red sensation') have in common are called 'qualia'. According to widespread opinion, qualia represent a serious problem for physicalist (materialist/naturalist) theories of the mind. The idea of an IS has been used in arguments for the claim that some or all materialist theories of the mind fail to capture qualia. In particular it has been used to argue against functionalism.

The Inverted Spectrum and Analytical Functionalism

There would be no qualia problem for physicalism if what it means to be in a state of a particular phenomenal kind could be stated in nonmental terminology. If this were true then the apparent problem of qualia could be resolved on the basis of *a priori* conceptual analysis of terms like 'has a red experience' or 'object O appears red to person P'. Analytical functionalism (as advocated by Lewis (1980a)) proposes a solution of this kind. According to analytical functionalism, to have an experience of red means to be in a state that plays a specific causal role in the functioning of the organism (where the causal role is defined, approximately, by the way the state is normally caused, by its causal interaction with other inner states, and by its causal relevance for behavior). Many have argued that analytical functionalism does not capture the meaning of phenomenal terms like 'red experience'. What it means to have an experience of a particular phenomenal kind (e.g. a red experience) appears to be understandable (for those who have had the experience at issue) without any reference to the causal role of the state at issue (to the way it is caused, to the behavior it typically produces, or to its impact on other mental states). This natural intuition has been justified and illustrated on the basis of the idea of an IS. If it is logically (conceptually) possible that color experiences of one phenomenal kind (e.g. red experiences) occupy in one person the causal role occupied in another person by color sensations of another phenomenal kind (e.g. by green experiences) and vice versa, then at least a particular version of analytical functionalism (a version that defines phenomenal states by their causal role in the individual at issue) must be false. One possible reply has been proposed by David Lewis (1980b), who defines phenomenal kinds by reference to their normal causal role in a given population (according to this view a person can have a red experience relative to one population but a green experience relative to another). Analytical functionalism does not have many adherents today. Most philosophers admit that cases of an IS, even in the strongest sense (presupposing condition C3c), are coherently conceivable.

The Inverted Spectrum and Empirical Functionalism

The conceptual possibility of IS in all three senses can be accepted by empirical functionalists, who

admit that phenomenal concepts (the concept of having a red experience) cannot be defined in nonmental terminology but insist nonetheless that the properties referred to or expressed by phenomenal concepts are functional properties. According to their view, the property of having a red experience consists in being in an inner state that occupies a particular causal role, although *a priori* reasoning about our concept of red experiences does not reveal this truth. In their view, it is the task of empirical science to describe the exact features of the causal roles at issue. According to the empirical functionalists, cases of IS (in the strongest sense; see C3c) appear possible (they are coherently conceivable), but they are not *really* possible (they deny the metaphysical possibility of an IS in the relevant sense). The appearance of such a possibility – they may add – can be explained by our conceptual make-up (our phenomenal concepts and our functional concepts are conceptually independent). It has been argued, however, that in the particular case at issue conceptual possibility implies metaphysical possibility. The issue depends upon one's view about the relation between phenomenal concepts and the corresponding phenomenal properties. Those who think that a person who has the phenomenal concept of a red experience (on the basis of her own red experiences), and thereby knows what is essential for having that kind of experience, will typically deny empirical functionalism and insist on the metaphysical possibility of IS in all three senses.

Some philosophers accept the metaphysical possibility of IS (in all three senses), deny empirical functionalism for the particular case of phenomenal states (while accepting it for other mental states, e.g. beliefs and desires) and propose to conclude that phenomenal states are identical with the particular neurophysiological states that occupy the causal roles at issue (Horgan, 1984). Against this view it has been argued that there is just as good reason to accept the metaphysical possibility of inverted color vision in two subjects that are molecule-by-molecule duplicates as there is to accept the metaphysical possibility of IS in people that are functionally alike. If the former metaphysical possibility must be accepted, then property dualism is the only viable alternative. Property dualism claims that what it is to have an experience of red cannot be captured in physical terminology at all (where physical terminology is meant in a broad sense, including chemical, biological, physiological, and functional descriptions). Property dualism does not have many adherents in contemporary philosophy of mind, but it has gained some ground in recent years (for a dualist conception of phenomenal properties see Chalmers, 1996).

Intrasubjective Color Inversion

Intrasubjective IS is not directly relevant to the debate about functionalism, but the thought experiences of someone who undergoes an intrasubjective color inversion have been used repeatedly to argue for the ISH. The possibility of intrasubjective IS is harder to deny, since it would manifest itself in behavior (the person would surely say that things don't look the same). Suppose a normally sighted person undergoes an intrasubjective spectrum inversion (suddenly grass looks red and sunsets green) but after some years she becomes accustomed to the new look of things and even forgets about the inversion event. She now behaves like a normally sighted person and thus is at least a case of an IS in its weakest sense (C3a). Thus it may seem that those who accept the logical (or metaphysical) possibility of an intrasubjective IS (which is hard to deny) should also accept the logical (or metaphysical) possibility of intersubjective IS (which threatens functionalism). This line of argument has been discussed in Shoemaker (1982), who also develops a possible functionalist response along the following lines: talk of differences or sameness makes sense only within one and the same individual (here difference and sameness can be functionally defined) and does not make sense when applied to experiences of different individuals (here difference and sameness cannot be defined in functional terms).

Inverted Spectrum and the Explanatory Gap

According to the explanatory gap thesis, it is impossible to understand why certain physiological processes are accompanied by experiences of a specific qualitative kind (Levine, 1983). The claim was originally motivated by pointing out the conceptual possibility of an extreme case of IS (where even the physiological details remain the same): even if we knew every physical–functional detail about the physiological state responsible for red sensations, we could still coherently conceive of a situation where this kind of state was accompanied by sensations of another phenomenal kind (e.g. by green experiences). It has been objected that the possibility at issue is only apparent. According to Hardin (1986, 1997), the case will cease to be conceivable once we know enough about the

phenomenal structure of color experience and about the functional structure of the underlying processes. It should be noted, however, that the explanatory gap thesis does not depend on the conceivability claim at issue. Even if Hardin is right, it may still be an open question as to why the physiological processes underlying color sensations lead to phenomenal experiences of a particular given kind: they could have led to hue experiences of a different kind (colors we humans actually do not know) with the same phenomenal structure.

Philosophical Discussion in the Light of Empirical Results

A great part of the philosophical discussion about IS has been led quite independently of color vision science, but some important work about the topic has been done on the basis of detailed knowledge of related empirical results. A prominent example is the work of C. L. Hardin, who assumes that any possible correlation between color sensations and physiological processes is structure preserving in the following sense: the phenomenal structure of color experience must be mirrored in the functional structure of the underlying physiological processes. He then goes on to argue that given the rich phenomenal structure of color experience, it is very plausible (already on the basis of what is known today) that there is just one structure that preserves mapping from brain processes to human color sensations, namely the one actually realized in nature.

Although Hardin's argument uses empirical results, quite unsurprisingly it does not reach its conclusion without additional philosophical assumptions about what is metaphysically possible. For his argument to carry he has to assume that the phenomenal structure of our color experiences is in part metaphysically necessary. Some of these additional assumptions have been questioned (see Levine, 1991).

Alien hues

Several other authors have argued that to attack functionalism one need not argue about the possibility of an IS in humans. Even if the human color space contains asymmetries that render a case of IS metaphysically impossible, this need not be true for other species (Shoemaker, 1982). If there is no reason to deny the metaphysical possibility of such creatures then empirical functionalism is in trouble. This line of argument presupposes that it makes sense to talk of alien colors, colors that we are in principle unable to imagine but that are experienced by other sentient beings. This presupposition may appear strange but becomes quite natural when considering the case of beings who can be shown to have more color dimensions than normal human perceivers. There is empirical evidence that animals with four or five color dimensions exist (see Thompson *et al.*, 1992).

Pseudonormal Vision

The empirical hypothesis that red–green inverted people might be among us (note that this implies that you might yourself be one of them) is a consequence of a model about the inheritance of red–green blindness proposed by Piantanida (1972). In normal humans, red experiences occur if R-cone activity (R-cones are receptor cells on the retina) prevails over G-cone activity and green experiences occur if G-cone activity prevails over R-cone activity. If, however, both kinds of receptor cells are filled with the same photopigment, then their activity in response to light stimuli will always be about the same and the person at issue will never see anything in a color that contains a component of red or green; she will be red–green blind. According to Piantanida's model this condition is caused in two ways: one type of inherited red–green blindness is due to the fact that the person's R-cones and G-cones both contain the photopigment normally contained only in R-cones (it is called 'the red pigment'); another kind of red–green blindness is due to the fact that the R- and G-cones both contain the photopigment normally contained only in G-cones (the green pigment). Now if a person has both genes for red–green blindness the filling of the G- and R-cones will be reversed with respect to normality (Piantanida calls people with this condition 'pseudonormal'). As a consequence of the pigment inversion, red objects will have exactly those effects on a pseudonormal human visual system that green objects have on a normal human color system and vice versa . It has therefore been concluded that pseudonormal people 'would be expected to have normal color vision except that the sensations of red and green would be reversed – something that would be difficult if not impossible to prove'. (Boynton, 1979, p. 356). Note that pseudonormal people (if they are red–green inverted) need not be a case of IS in any of its three senses. Whether they are depends on empirical issues such as the question of whether the coolness of green is due to environmental influences (if so, red might appear cool to pseudonor-

mals; if not, pseudonormal people should be detectable even on the basis of their behavior in normal life situations).

Some philosophical theories about phenomenal consciousness (in particular analytical functionalism and externalist versions of representationalism about qualia) are incompatible with the view that pseudonormal people (if they exist) are red–green inverted. One may be tempted to conclude that these theories have been refuted by empirical science. But the conclusion would be hasty. It can and has been argued that a scientist who predicts red–green inverted vision for people with the pseudonormal condition implicitly presupposes philosophical premises (the falsity, for example of qualia externalism, the supervenience of phenomenal experience on instrinsic properties of the visual system) that cannot be settled by empirical research. But maybe these 'philosophical' premises are nonetheless scientifically justified by certain successful underlying assumptions or methodological principles. It seems quite clear that the issue about whether pseudonormal people if they exist would be red–green inverted requires both the clarification of empirical issues and metatheoretical (philosophical) reflection about the meaning and relevance of phenomenal terminology in color vision science. There has been some discussion about this issue within philosophy (Ross, 1999) and apparently no discussion about it within color vision science. Since 1972 theories about the genetics of red–green blindness have of course become far more complex, and the empirical issue of whether there are pseudonormal people (people with green pigments in their R-cones and red pigments in their G-cones) does not seem to have been settled (for a review article of relevant research see Nathans, 1999). But then the question has not received much discussion. Surprisingly enough, Piantanida's hypothesis that there might be red–green inverted people among us has failed to initiate any vivid debate within empirical color vision science.

Representationalism

The ISH is relevant to the debate about representationalism vis-à-vis qualia. Weak representationalism with respect to qualia claims that they do represent external properties. Strong representationalism says that to represent a specific objective property is all that it is to be in a specific phenomenal state (representationalist views about qualia are proposed by Lycan, 1996; Tye, 1995; and Dretske, 1995). According to the latter view, to have a red experience is to visually represent the property red. What this means is often defined in functional terms while what it means for a given state to represent something is normally defined in a way that requires the state be reliably caused by things with the represented property. Representationalism is naturally combined with objectivism about colors (colors are objective properties of physical things and can be defined without reference to the way they appear to humans) and with externalism about qualia, the view that the kind of phenomenal experience people have can depend on their causal relations to their environment and does not depend on their inner intrinsic properties alone. Most versions of externalist representationalism have to deny the possibility of inverted color vision (a red–green inverted person has reliably red experiences in the presence of green things). Externalist representationalism is incompatible with the plausible assumption that the kind of color experienced (whether it is a red experience or a green experience) depends on intrinsic properties of the brain and not on how the instatiation of these properties is caused by external stimuli. Some representationalists wish to avoid this consequence and propose an account of representationalism that is compatible with the denial of externalism about qualia (see Rey, 1998). In the recent debate about representationalism the ISH has played a minor role compared to the thought experiment involving an inverted earth (introduced by Block, 1990) that is related but different.

Other Philosophical Issues about the Inverted Spectrum

IS is mainly discussed within the philosophy of mind, but it is also relevant to philosophical questions belonging to other domains of philosophy. Any theory of meaning in the spirit of verificationalism needs to be tested intuitively by applying it to versions of the IS. Intrasubjective and intersubjective IS raise the question of whether or in what sense it may be possible to refer to one's own inner (i.e. not publicly accessible) states by employing public language terms and whether the specific quality experienced can be communicated to others (see Strawsen, 1989). Finally, IS raises the question of whether and how we can know about the existence and the contents of other minds. If the ISH is correct, then I cannot deny that you see green where I see red. Can I nonetheless claim to know that we both have green experiences in the presence of green objects? The fear that any acceptable theory of knowledge combined with the

ISH would imply that I cannot (and in this sense leads to skepticism about other minds) is one reason that some philosophers reject the ISH.

EMPIRICAL ISSUES INVOLVING 'IS'

The empirical question of whether there are in the human case color transformations that preserve those aspects of phenomenal structure that can be assumed to remain unchanged under different learning conditions is discussed in detail by Palmer (1999). He argues that all but three transformations can be ruled out from is known about phenomenal structure. The three remaining candidates are: red–green inversion; blue–yellow inversion combined with black–white inversion, and the combination of inversions with respect to the three color dimensions (red–green inversion plus blue–yellow inversion plus black–white inversion). According to Palmer, whether any of these really is structure preserving in the sense just mentioned depends on empirical issues that have not yet been settled.

RELEVANCE OF THE 'IS' TO COGNITIVE SCIENCE

In their study of color vision cognitive scientists are typically interested in the causal structure and the causal mechanisms of those processes that are responsible for the capacity of an organism to distinguish surfaces according to their light reflectance properties. This aspect of color vision can be studied independently of issues related to the idea of an IS. But color vision in the case of humans and many other species also has a subjective side: color experiences have particular phenomenal characters. There are many questions concerning phenomenal character that a science of color vision can legitimately try to answer: e.g. what hues are experienced by color-blind people in its various forms? Would pseudonormal people be color inverted? At what stage of evolution does color experience (in the sense of having a particular phenomenal character) occur? Do other animals know hues that we cannot imagine? Many think that questions about phenomenal character are beyond the reach of empirical science, but this is far from obvious and should be discussed in a metatheoretical study that requires both scientific and philosophical competence. Such a study would have to clarify questions about (a) the meaning of phenomenal terminology in scientific theories of vision and (b) about the possible limitations of scientific approaches to the subjective side of experience. The idea of IS is potentially relevant for this kind of metatheoretical reflection in a way that may not yet have been sufficiently recognized (for a beginning of an interdisciplinary discussion about these issues see Palmer, 1999 and the various comments in the same journal issue).

References

Block N (1990) Inverted earth. *Philosophical Perspectives* **4**: 53–79.

Block N and Fodor JA (1972) What psychological states are not. *The Philosophical Review* **81**: 159–181.

Boynton RM (1979) *Human Color Vision*. Holt, Rinehart and Wilston.

Chalmers DJ (1996) *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Cole, David (1990) Functionalism and inverted spectra. *Synthese* **82**: 207–222.

Dretske F (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Hardin CL (1987) Qualia and materialism: closing the explanatory gap. *Philosophy and Phenomenological Research* **48**: 281–298.

Hardin CL (1988) *Color for Philosophers: Unweaving the Rainbow*. Indianapolis: Hackett.

Hardin CL (1997) Reinverting the spectrum. In: Byrne A and Hilbert DR (eds) *Readings on Color, Vol. 1. The Philosophy of Color*. MIT Press.

Horgan T (1984) Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research* **44**: 453–469.

Lewis D (1980a) Psychophysical and theoretical identifications. In: Block N (ed.) *Readings in the Philosophy of Psychology*, vol. 1, Cambridge, MA: Harvard University Press.

Lewis D (1980b) Mad pain and Martian pain. In: Block N (ed.) *Readings in the Philosophy of Psychology*, vol.1, Cambridge, MA: Harvard University Press.

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Levine J (1991) Cool Red. *Philosophical Psychology* **4**: 27–40.

Locke J (1690/2000) In: Fuller G, Stecker R and Wright JP (eds) *John Locke: An Essay Concerning Human Understanding*. London and New York: Routledge.

Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Nathans J (1999) The evolution and physiology of human color vision: insights from molecular genetic studies of visual pigments. *Neuron* **24**: 299–312.

Nida-Rümelin M (1996) Pseudonormal vision: an actual case of qualia inversion? *Philosophical Studies* **82**: 145–157

Nida-Rümelin M (1999) Pseudonormal vision and color qualia. In: Hameroff S, Kasniak A and Chalmers D (eds) *Toward a Theory of Consciousness*. Cambridge, MA: MIT Press.

Palmer ES (1999) Color consciousness and the isomporphism constraint. *Behavioural and the Brain Sciences* **22**: 923–989.

Piantanida TP (1974) A replacement model of X-linked recessive colour vision defects. *Annals of Human Genetics* **37**: 394–404.

Rey G (1998) A narrow representationalist account of qualitative experience. In: Tomberlin JE (ed.) *Philosophical Perspectives: Language, Mind, and Ontology*, vol. 12. Atascadero, CA: Ridgeview Publishing.

Ross P (1999) Color science and spectrum inversion: a reply to Nida-Rümelin. *Consciousness and Cognition* **8**: 566–570.

Shoemaker S (1975) Functionalism and qualia. *Philosophical Studies* **27**: 291–315.

Shoemaker S (1982) The inverted spectrum. *The Journal of Philosophy* **79**: 357–381.

Shoemaker S (in press) Two cheers for representationalism. *Philosophy and Phenomenological Research*.

Strawson G (1989) Red and 'red'. *Synthese* **78**: 193–232.

Thompson E, Palacois A and Varela FL (1992) Ways of coloring: comparative color vision as a case study for cognitive science. *Behavioral and Brain Sciences* **15**(1): 1–74.

Tye M (1994) Qualia, content, and the inverted spectrum. *Noûs* **28**: 159–83.

Tye M (1995) *Ten Problems of Consciousness: A Representaionalist Theory of the Phenomenal Mind.* Cambridge, MA: MIT Press.

James, William

Introductory article

E Taylor, Saybrook Institute, San Francisco, California, USA and Harvard University, Cambridge, Massachusetts, USA

CONTENTS

William James (1842–1910) made significant contributions to the fields of psychology and philosophy. He is perhaps best known for his monumental work the Principles of Psychology *(1890) which introduced several influential concepts, amongst the most notable being the 'stream of consciousness'.*

SCIENTIFIC LINEAGE

William James (1842–1910), physician and philosopher-psychologist, taught at Harvard University from 1873 to 1907. Weaned on the Swedenborgian and transcendentalist philosophy of his father, Henry James Sr, and his godfather, Ralph Waldo Emerson, James matriculated at Harvard in 1861, where he studied chemistry under Charles William Eliot at the Lawrence Scientific School. He entered Harvard Medical School in 1864 to study comparative anatomy under Jeffries Wyman; microscopy, anatomy, and physiology under Oliver Wendell Holmes Sr; and vivisection under Harvard's first professor of neurology, Charles Edouard Brown-Sequard. James's training in basic science was eclectic but distinguished. He derived his evolutionary theory from the Harvard botanist Asa Gray, a member of Darwin's inner circle. He gleaned reductionstic positivism from the philosopher Chauncey Wright; and his logic of the scientific method he learned from Charles Sanders Peirce, the originator of pragmatism. James's emphasis on clinical applications came from the attitude of *la clinique*, derived from Pierre Marie through Oliver Wendell Holmes Sr, while his expertise in experimental methods came from his training in surgical dissection, this provided an insight into structure and function in a lineage that can be traced through Holmes and Brown-Sequard back to the experimental physiology of Claude Bernard, Francois Magendie, and Xavier Bichat.

According to Walter Bradford Cannon, in the early 1870s James and his medical school colleagues Henry Pickering Bowditch and James Jackson Putnam conducted the first studies in experimental neurophysiology in the United States. These studies, which were conducted in Bowditch's laboratory, involved surgical dissection of dogs' brains and were designed to settle issues involving the localization of function.

James began teaching anatomy and physiology in 1873, taught the first course in physiological psychology in America in 1875, and opened the first laboratory of experimental psychology devoted to student instruction that same year. He awarded the first doctorate in the new experimental psychology to G. Stanley Hall in 1878. At the same time, James continued to involve himself in experimental studies of the nervous system, eventually proposing a cure for seasickness derived from the study of dizziness in deaf mutes. These studies were further corroborated when the New York neurologist George Miller Beard replicated them in 1878.

PSYCHOLOGY AND PHILOSOPHY OF CONCIOUSNESS

Meanwhile, James taught psychology and philosophy throughout the 1870s and 1880s, marching through the categories of abstract philosophy and appropriating them one at a time for the new physiological psychology. His success led to a contract for a textbook in the new psychology in 1878, to be delivered in two years, but it took him 12. Periodic ill health intervened as did the death of his father and mother, and then his godfather, the great Emerson, within months of each other. There was also the intrusion into his psychology of new techniques in experimental psychopathology and psychical research, among them: hypnosis, crystal gazing, and automatic writing. James refitted the experimental laboratory at Harvard to accommodate these

methods and began to investigate dissociation theory as an explanatory mechanism for the altered states of consciousness seen in the mediumistic trance as well as the symptoms of neurasthenia and hysteria observed in patients suffering from functional rather than organic diseases of the nervous system. Hypnosis became for him a research tool for experimental induction as well as extinction of such phenomenon. In this, James proved himself to be a skilled hypnotist.

His work was to have a profound influence on the course of experimental psychopathology and the newly emerging field of psychotherapeutics, as his publications on these subjects from the viewpoint of physiology allowed neurologists such as James Jackson Putnam at the Massachusetts General Hospital and Morton Prince at the Boston City Hospital to introduce psychotherapeutic methods into the outpatient clinics in the treatment of the ambulatory psychoneuroses. His work also influenced Binet, Janet, and Ribot in France, among others. Later, James's studies would also have an impact on the evolution of his thinking about a new metaphysics underlying the way experiments should be conducted in psychological science, if we were to understand the problem of consciousness. He would call this new metaphysics 'radical empiricism'.

COGNITIVE PSYCHOLOGY OF THE OBJECT

Finally, in 1890, ten years late, James's monumental *Principles of Psychology* appeared to international acclaim. It filled two volumes and was over a thousand pages long. Joking but exhausted, James declared that he was finally glad to get that 'dropsical, tumescent mass', off his desk. Its main point was the development of a cognitive psychology of consciousness, focusing on an analysis of objects as they appear at the center of the field of attention. Its epistemological basis was reductionistic positivism, James said, because no other philosophical metaphysics of scientific experimentation superior to it had yet been articulated.

The work opened with an up-to-date survey of brain neurophysiology, followed by a trenchant scientific and philosophical analysis of topics that stand to this day. Instincts are emotions, he said, frozen into the nervous system because of their supreme evolutionary utility. Between the gross and the subtle, whatever emotion we experience, however, is dependent on our perception of the event. Emotions are the great flywheel of society, the habit system being in effect an early version of the cognitive system. Within the cognitive field, all attention is based on interest, reminding us that while what we see is obviously important, equally significant is what we do not see. To see, therefore, means that discriminative attention is what saves us from being constantly inundated by the 'blooming buzzing confusion'. Thus, all culture, James said, begins with rejection. As consciousness is always flowing onward, it is appropriate to speak of a stream of thought and feeling. What we attend to usually takes up the center of the field, however, while the margin of consciousness controls meaning. Every thought is warmed by an emotion, making it our own, and our emotions are implicated in the meaning of each event, although our emotional attachment to an idea remains hidden just below the surface of consciousness, or associated with ideas and stored subconsciously as memory. Our sense of self is made up of a biological, material, and psychological sense of who we are, compounded by our social self, which is made up of as many different selves as we have human relationships. In this regard, the social self mirrors the basic nature of personal consciousness, which is an ultimate plurality of states, not a unity as we would like to believe.

The two-volume 'James' was quickly cut and pasted into the more convenient one-volume 'Jimmy' in 1892, and released as *Psychology: Briefer Course*. The shorter work then became a standard college text for the next 20 years. A recent reassessment of *The Principles* suggests that virtually nothing new has been discovered since the work was first published. It is still considered possibly the greatest general work on psychology in the discipline's history.

SUBCONSCIOUS AND MYSTICAL AWAKENING

After 1890, James began to turn his attention to the scientific study of consciousness beyond the everyday waking state of awareness. Consciousness, he said, was 'a field with a focus and a margin'. He now proposed to study consciousness beyond the periphery: in other words, the penumbra or halo of thought. This took him out of the realm of pure cognition into the domain of the emotional and the imaginal. The problem he faced was an epistemological one, however. The limits of what we would now call 'cognitive neuroscience' were, and still are, clearly defined by the dictates of reductionistic positivism. Experimental science deals only in the rational and the empirical, meaning that its methods are limited to the rational

ordering of sense data. Hence, the new psychology was based on what could be measured, and largely defined by advances in physiology and psychophysics. There were also the new methods of mental testing, which were usually paper-and-pencil tests and therefore similarly constrained by language.

Important new advances in dissociation theory coming from the French experimental psychology of the subconscious around Charcot, Binet, Janet, Bernheim, and others suggested to James, however, the reality of interior states of consciousness. James declared in his presidential address to the American Psychological Association in 1894 that he was no longer going to defend the metaphysics of positivism as the foundation for experimental science, based, he said, on the new evidence coming in from French psychopathology. These states were radically different from the normal everyday waking condition where the cognitivists had exclusively focused their definition of scientific psychology. Moreover, dissociation theory, which was the basis for understanding the effects of hypnosis and the probable mechanisms at work in conditions such as hysteria and multiple personality, had grown out of French neurophysiology, and thus had its own scientific foundation. The principal contribution of the French neurological tradition, corroborated by the replication of its studies through the British psychical researchers, was the psychogenic hypothesis – that ideas can cause physical symptoms. After the early mesmerists had tried and failed to convince the French medical establishment of the scientific validity of hypnosis, and long before the concept of psychogenesis was associated with Freud and psychoanalysis, French, Swiss, British, and American researchers had confirmed the reality of the 'buried idea'. Experiences perceived as traumatic by the individual could be stored in the subconscious. There they existed as unintegrated fragments, floating around in the form of a single powerful but buried idea, collecting similar experiences to themselves until they gained enough energy to appear in the field of consciousness as a physical symptom, or even burst forth into the field looking like a fully developed, independent personality. Here, James was closely following the work of F. W. H. Myers.

After 1890, James, along with others such as James Mark Baldwin, became the principal interpreter of the French dissociation school to American psychologists. In this context, James was also the first to introduce the work of Breuer and Freud to the American psychological public in 1894, in the inaugural issue of James Mark Baldwin and James McKeen Cattell's *Psychological Review*. In that context, he cited Breuer and Freud as providing corroboration for 'Janet's already old findings'.

James went on with this work to pioneer in what we would identify today as the origins of personality, abnormal, and social psychology. After visiting Charcot at the Salpêtrière in 1882, he significantly influenced Ribot's conceptions about the pathology of the emotions; from his background in chemistry he coined the term 'dissociation' as a psychological concept in order to trump the associationists, and his own hypnotic investigations inspired the young Morton Prince to enter the field. In the midst of other classes in psychology and philosophy, from 1893 to 1898, James taught the first graduate course at Harvard in experimental psychopathology, and his previously unpublished 1896 Lowell Lectures on 'Exceptional Mental States' further confirmed the previously loose-knit endeavors of the so-called Boston School of Psychopathology thereafter.

The Lowell Lectures outlined James's conception of a dynamic psychology of the subliminal, or subconscious, as it was variously called, and particularly focused on experimental evidence for the hypnagogic state, the twilight region between waking and sleeping where dreaming naturally occurs and hallucinations appear when there is a fall of the threshold of consciousness in the waking state. Exploration of the subconscious put us in touch with the deepest regions of human nature, James said.

The Exceptional Mental States Lectures investigated the psychopathic region of the subconscious, while James's *Varieties of Religious Experience*, delivered in 1902, presented evidence for a growth-oriented dimension to personality. Within this dimension, James showed, mystical experiences can occur that have ultimately transforming effects on the person. Such philosophical implications of non-ordinary states of consciousness called into question the limitations experimental psychologists put on their subject matter by having decided too quickly that their underlying philosophical metaphysics would be the same as physics and biology.

METAPHYSICS OF EXPERIENCE

But this was a time when scientific psychology was establishing itself in the newly founded American graduate schools, experimental laboratories were being established, and philosophy was being more clearly differentiated from scientific psychology. To all appearances, James seemed to be bringing

philosophy back into psychology. For this, to this day he remains grossly misunderstood. Within the field of philosophy he had just established his doctrine of the will to believe – choosing the good as a commitment to the moral and ethical life even though evil always waited *in potentia* – and in 1898 he formally launched the philosophical movement called 'pragmatism', which soon became international in scope. The true heart of his philosophical metaphysics, however, was radical empiricism, by which he meant pure experience in the immediate moment, before the differentiation between subject and object.

Traditional philosophers had differentiated rationalism from empiricism, the word 'empirical' referring to the data of the senses. James's empiricism was more radical, he said, because he meant for the word to mean not sense perception alone but the full spectrum of a person's experience. This meant that the subjective and the objective were both equally valid points of view within the larger theater of experience. Radical empiricism he conceived as a new metaphysics for experimental psychology and in science generally, because it established the phenomenology of the science-making process as a crucial element, not to be controlled against but to be accounted for, in the conduct of the scientific experiment. Without including this crucial element, he said, a science of consciousness was not possible.

Radical empiricism had far-reaching implications because it purported to replace reductionistic positivism with an observationally grounded phenomenology. It overthrew the doctrine of representation in psychophysics, since there were not two sets of words, the one in front of us at this moment and the one in our mind as we read them. There is only one set, and that exists at the intersection between the history of the text and our immediate autobiography. It also made psychology not derivative but foundational to all the basic sciences, since there could be no measurements or laws without some consciousness somewhere to perceive and interpret them.

But, due to a variety of factors – James's continuing ill health, his first invitation in 1897 to give the Gifford Lectures on Natural Religion, the years of preparation and postponement that ensued, and the international attention afforded pragmatism – the doctrine of radical empiricism was to remain James's great unfinished arch. Nevertheless, when he died in 1910, it was generally acknowledged that an international figure in psychology and philosophy had passed from the scene.

CONTEMPORARY RELEVANCE

Since then, James's ideas have touched almost all aspects of modern psychology. Cognitivists, behaviorists, psychoanalysts, existentialists, phenomenologists, and humanistic and transpersonal psychologists have all variously claimed parts of him. Experimental psychologists will not read him after 1890, however; religious scholars study only his *Varieties of Religious Experience* from 1902; while philosophers continually try to cast him into the mold of the Western analytic tradition, meanwhile ignoring his psychology, his studies in psychical research, and his philosophy of mysticism.

Modern researchers in the cognitive neurosciences also continually go back and take pages from James, mainly from his *Principles of Psychology*, although neurophilosophy and now neurotheology are just beginning to cover the same ground that James had broken in his Exceptional Mental States lectures and in *The Varieties of Religious Experience*. Francis Crick's article on brain neuroscience in *Scientific American* begins with a full portrait of James. Thomas Natsoulas has been developing a cognitive theory of James's stream of consciousness for a number of years, while Bernard Baars has evolved a cognitive workspace theory of consciousness which comes close to James, as does Max Velmans's reflexive theory of consciousness. Since James's time, the mind/brain interface has once again become a central problem in the philosophy of science. What this new revolution in the neurosciences seems to be telling us is that, as science gets closer to understanding the biology of consciousness, the more a phenomenology of the science-making process and the scientific method itself come in for closer scrutiny. James predicted this would happen when, in one of the last statements he made before he died, he advised psychologists to study the fall of the threshold of consciousness in all its manifestations, 'even though we may not understand these phenomena or their far-reaching implications for several generations to come'.

Further Reading

Crick F (1992) The problem of consciousness. *Scientific American* **267**: 152–159.

Donnelly M (ed.) (1992). *Reinterpreting the Legacy of William James. (APA Centennial William James Lectures)*. Washington, DC: American Psychological Association.

Henley T and Johnson M (eds) (1990) *Reflections on The Principles of Psychology: William James after a Century*. New York, NY: Earlbaum.

James W (1890) *Principles of Psychology*. New York, NY: Henry Holt.
James W (1892) *Psychology: Briefer Course*. New York, NY: Henry Holt.
James W (1895) The knowing of things together. *Psychology Review* **2**: 105–124.
James W (1897) *The will to Believe*. New York, NY: Henry Holt.
James W (1902) *The Varieties of Religious Experience*. New York, NY: Longmans, Green.
James W (1904) Does 'consciousness' exist? *Journal of Philosophy* **1**: 477–491.
James W (1907) *Pragmatism*. New York, NY: Longmans.
James W (1910) A suggestion about mysticism. *Journal of Philosophy, Psychology, and Scientific Methods* **7**: 85–92.
Natsoulas T (2000) The stream of consciousness XX: A non-egological concention. *Imagination, Cognition, and Personality* **19**: 79–90.
Perry RB (1935) *The Thought and Character of William James, as Revealed in Unpublished Correspondence and Notes, Together with his Published Writings*. Boston: Little, Brown and Company.
Taylor EI (1982) *William James on Exceptional Mental States: Reconstruction of the 1896 Lowell Lectures*. New York, NY: Charles Scribner's Sons.
Velmans M (ed.) (2000) *Investigating phenomenal consciousness: New methodologies and maps*. Philadelphia: J. Benjamins Publishing Co.

Knowledge Argument, The

Intermediate article

Adam Vinueza, University of Colorado, Boulder, Colorado, USA

CONTENTS

The knowledge argument purports to show that there are nonphysical facts – facts that cannot be expressed in physical terms – because one can know all the physical facts without knowing facts about what it is like to have an experience.

INTRODUCTION

The knowledge argument is against physicalism, the doctrine that all facts are physical facts. Physicalism is a notoriously difficult doctrine to state clearly and plausibly, but we can explicate its main idea by taking a physical fact to be a fact that can be expressed in the language of physics. Thus, to say that physicalism is true is to say that all facts can be expressed in the language of physics.

The knowledge argument purports to show that there are facts about what it is like to have an experience, and that these facts cannot be expressed in the language of physics. If you have never tasted asparagus (or anything relevantly like it), you do not know what it is like to taste asparagus: there is a qualitative aspect to tasting asparagus that you have never experienced. Likewise, if you are an achromat, you have never had color experiences, so you do not know what it is like to have such experiences. These aspects of experience are called *qualia* (singular *quale*): there is a unique quale to tasting asparagus, a distinct unique quale to seeing vermilion, and so on. Thus, the knowledge argument purports to show that certain facts about qualia cannot be expressed in the language of physics.

SIGNIFICANCE AND HISTORICAL CONTEXT

The knowledge argument (Jackson, 1982, 1986) is usually introduced with the following thought experiment. Imagine that some person, Mary, has been confined from birth to a black-and-white room, and has been educated through black-and-white books and through lectures and discussions relayed on black-and-white television. Let us also imagine that Mary, in this room, has been taught all the physical facts – she has learned all the facts one can learn through studying physics, chemistry, neurophysiology, and so on – and that these theories are all complete and correct. Then if physicalism is true, Mary knows everything, because the physical facts are all the facts. But according to Jackson, Mary does not know everything, because if she were to leave her black-and-white environment she would see her first color: she would have her first experiences with color qualia, and would thereby come to know – that is, to learn – what it is like to have color experiences. If she knows all the physical facts before leaving the black-and-white room and learns what it is like to have color experiences upon leaving it, then what it is like to have color experiences is something Mary learns that is not a physical fact. So, according to Jackson, there are things one can know, some facts, that are not physical facts.

If the knowledge argument is sound, physicalism is false. But physicalism is a methodological assumption that guides all scientific inquiry. We reject explanations inconsistent with physical explanations – such as explanations of events that appeal to miracles – on the ground that they are physically impossible, and it is arguable that the only way we can make sense of this dependence of all explanations on physical explanations is by supposing that all the facts, ultimately, are physical facts. Physicalism is also foundational for cognitive science, which is guided by the hypothesis that all features of our mental lives can be explained by appeal to functional (in particular, computational) processes. So if the knowledge argument is sound, the central hypothesis guiding cognitive science is false, and arguably cognitive science is misguided.

That the knowledge argument has such far-reaching consequences can easily lead one to reject it out of hand: one might think that if it entails that physicalism is false, then the argument simply has to be unsound. But this does not explain precisely

where it goes wrong. In this article, we will see the main ways in which philosophers have tried to show that the knowledge argument is unsound, and examine their strengths and weaknesses.

The intuition motivating the knowledge argument goes back at least to Leibniz, who argued that mentality cannot be explained by appeal to physical principles. Leibniz imagined a machine that thinks and perceives; he claimed that if we looked inside, we would never see anything that could explain perceiving (Leibniz, 1981, pp. 66–67).

Leibniz' claim was almost universally accepted by philosophers until the middle of the twentieth century, when philosophical orthodoxy shifted from Cartesian dualism to materialism. The knowledge argument arose during this period of materialist orthodoxy, when many philosophers had begun to feel unsatisfied with materialist explanations of mentality. Nagel (1974) famously expressed this dissatisfaction by arguing that we cannot know what it is like for bats to echolocate, for the simple reason that we cannot have anything relevantly like echo-locatory experiences. This argument, often conflated with the knowledge argument, is weaker: at most it shows only that we cannot form the concepts necessary for recognizing the truth of physicalism. But it is related to the knowledge argument, and may have inspired it.

RESPONSES TO THE KNOWLEDGE ARGUMENT

Let us express the knowledge argument more formally, in order to classify responses to it more easily:

When she leaves her black-and-white environment, Mary comes to know a fact – namely, the fact of what it is like for her to see colors. (1)

Mary did not know what it is like to see colors before leaving her black-and-white environment. (2)

If physicalism is true, then Mary's knowledge of what it is like to see colors is not a fact she did not know before leaving her black-and-white environment. (3)

Therefore, physicalism is false. (4)

Premise 3 follows directly from physicalism and the assumption that Mary knows all the physical facts, so the only premises that can be questioned are 1 and 2. Those who challenge the knowledge argument either deny that what Mary comes to know is a fact (premise 1), or they deny that it is a fact she did not know before leaving her black-and-white environment (premise 2). Let us call a response of the former sort a no-fact response, and a response of the latter sort an old-fact response.

No-fact Responses

No-fact responses may take two forms. The most popular form is known as the ability hypothesis: the hypothesis that knowing what it is like to have an experience is not propositional knowledge (knowledge that something is the case), but instead just know-how, the possession of an ability (Nemirow, 1990; Lewis, 1983, 1988; Churchland, 1985). The other form of no-fact response holds that knowing what it is like to have an experience is what Russell (1910) called knowledge by acquaintance, an intuitively immediate, nonpropositional awareness of a property (Conee, 1990).

The ability hypothesis implies that knowing what it is like to have an experience is knowing how to imagine or re-identify that experience. Of course, knowing how to do many things involves factual knowledge: knowing how to drive a car requires knowing many facts about how the car works, for example. But one can know all the facts about cars without knowing how to drive a car. Mary surely gains abilities when she leaves the black-and-white room – such as the ability to imagine and re-identify new color experiences – and Mary's knowing what it is like to have an experience arguably consists in her having those abilities. Because one cannot plausibly imagine or re-identify experiences one has never had, ability hypothesists can explain why Mary learns something upon leaving her black-and-white environment without denying physicalism, because what Mary acquires is not factual knowledge, but know-how.

The ability hypothesis is controversial, in part because most cognitive theories of know-how in general explain it as the possession of tacit factual knowledge: for example, our ability to produce and understand sentences of a natural language is know-how, but is arguably best explained in terms of tacit knowledge of a grammar for that language, a set of rules determining the grammatical structures of its sentences (Chomsky, 1965). Another objection is a semantic one: the semantics of 'knows *wh*-' constructions in general has them being true in virtue of the knower's having factual knowledge. For example, to know where Tanzania is is to know that it is at such-and-such a place, and to know why Smith murdered Jones is to know that

Smith murdered Jones for such-and-such a reason. Likewise, to know what it is like to have an experience is arguably to know that having that experience is like such-and-such (Lycan, 1996). And if semantics dictates that sentences of the form 'X knows what it is like to have experience E' are true in virtue of the knower's knowing that something is the case, then the ability hypothesis cannot be true.

The acquaintance hypothesis implies that knowing what it is like to have an experience is to be acquainted with it, where knowledge by acquaintance is an intuitively immediate (non-inferential) and nonpropositional form of knowledge. Examples of acquaintance include the awareness we have of our own conscious mental states, and our perceptual awareness of colored patches in the visual field. Nemirow (1990) and Churchland (1985) offer versions of the acquaintance hypothesis, but analyze acquaintance in terms of the possession of abilities; for this reason, many identify the acquaintance hypothesis with the ability hypothesis. Conee (1990) does not analyze acquaintance in this way, but does not say precisely what knowledge by acquaintance is.

The acquaintance hypothesis faces objections analogous to those faced by the ability hypothesis. Acquaintance – at least, understood as perceptual awareness – is arguably best characterized as awareness that something is the case: perceptual awareness is typically characterized by cognitive scientists in computational terms, and computational states typically have propositional content. Also, Lycan's argument about the semantics of 'knows what it is like' applies with equal force to the acquaintance hypothesis.

Note that these objections to the ability and acquaintance hypotheses are not objections against either the view that knowing what it is like is an ability or the view that it is a kind of acquaintance; but they imply that knowing what it is like, whatever it is, is nevertheless a kind of factual knowledge. One might accept that knowing what it is like is a kind of factual knowledge, yet hold that it is, under an unfamiliar guise, knowledge Mary already has. However, this is a kind of old-fact response, not a no-fact response.

Old-fact Responses

Old-fact responses take the form that while Mary does come to know a fact when she leaves her black-and-white environment, this fact is just one of the facts she already knew. Arguably, one can know the same fact in various ways: for example, the fact that Mark Twain is Mark Twain is seemingly identical to the fact that Mark Twain is Samuel Clemens, even though one can know that Mark Twain is Mark Twain without knowing that Mark Twain is Samuel Clemens. Intuitively, knowledge is a relation between a knower and a true proposition, and on certain views different propositions can be true in virtue of the same facts holding. Arguably, then, to learn a fact is to learn a proposition that is true in virtue of that fact holding; hence, one can learn a fact in as many different ways as there are propositions corresponding to that fact.

Old-fact responses are the most popular sort of response to the knowledge argument (e.g. Churchland, 1985; Bigelow and Pargetter, 1990; Loar, 1990; Lycan, 1990, 1995, 1996; Pereboom, 1994; Tye, 1986.) The burden on those who make this sort of response is to make it plausible that what Mary learns is an old fact under the guise of a different proposition. This is usually attempted by drawing analogies between Mary's situation and situations in which it is uncontroversial that someone comes to know a new proposition without learning any new facts. For example, suppose some person, Bill, knows that Mark Twain wrote *Huckleberry Finn*, but does not know that 'Mark Twain' is a pseudonym for Samuel Clemens. When he learns that Samuel Clemens wrote *Huckleberry Finn*, this is plausibly not a new fact for Bill, but a fact that Bill already knows under a different guise. If what Mary learns in learning what it is like (say) to see red is factual knowledge, then we may suppose that what Mary learns is that seeing red is like Q, where Q is the name of the quale Mary has when she sees red things. Therefore, on the old-fact view, Mary already knows that seeing red is like Q, but in virtue of her knowing a different proposition.

The holder of the old-fact view argues in the following way. Plausibly, the propositions that Mark Twain is Samuel Clemens and that Mark Twain is Mark Twain are distinct, yet the fact that Mark Twain is Samuel Clemens is the same fact as the fact that Mark Twain is Mark Twain; analogously, the propositions that seeing red is being in a certain physical state P and that seeing red is like Q are distinct, yet correspond to the same facts, because being in P just is being in a state like Q.

The main challenge for this sort of response derives from the standard explanation of not knowing a proposition while knowing the fact corresponding to it, viz., that one knows some but not all of a thing's properties. If one knows that Mark Twain wrote *Huckleberry Finn* but not that Samuel Clemens did, that is arguably because there are some

properties of Twain that one does not realize he has, viz., being named 'Samuel Clemens'. But Mary is not in this position: if she knows all the physical facts and physicalism is true, then she knows all the properties of every thing. This means that one who maintains an old-fact response must offer some other explanation of not knowing a proposition while knowing the fact corresponding to it.

Old-fact theorists have offered several alternative explanations. The most prominent, due to Loar (1990) is that concepts of qualia are those whose possession requires having certain experiences, and that Mary has not had the experiences requisite for having such concepts of color qualia. These concepts, called phenomenal concepts, enable one to recognize one's own qualia when one has them: they are those concepts in virtue of which one can self-ascribe sensations. (For closely related views see Lycan (1995, 1996) and Tye (1986).) If knowing what it is like to have an experience with quale Q is knowing that the experience is like Q, and the concepts of Q required for having this knowledge are phenomenal concepts, then old-fact theorists can explain why Mary does not know what it is like to have color experiences by pointing out that she lacks the requisite concepts, while denying that knowing what it is like to have experiences is knowledge of distinct facts.

Interestingly, this sort of old-fact response leads to a different, though related, worry. According to Leibniz' law, if any thing X is identical to a thing Y, then whatever is true of X is true of Y. And because X is necessarily identical to X, it follows that X is necessarily identical to Y. Nevertheless, many identities can seem to hold only contingently, because one can coherently conceive them to be false, and we need to do empirical investigation to find out whether they hold: the standard example is the identity of water and H_2O. This would suggest that what we can coherently conceive is not a reliable guide to what is possible: water is necessarily identical to H_2O, but we can coherently conceive that water is not H_2O. But an influential argument from Kripke (1980) seems to show that these sorts of examples do not show that conceivability is no guide to possibility. (For further developments of this argument, see Chalmers (1996) and Jackson (1997).)

Kripke's argument is as follows. Because water is necessarily H_2O, there is no possible situation where water is not H_2O; therefore, when we claim we can conceive that water is not H_2O, we are either not conceiving any situation at all, or what we are conceiving is something that we might legitimately confuse with this impossible situation. When we claim to be able to conceive that water is not H_2O, Kripke argues, we are conceiving a situation in which something that has the properties we normally associate with water – say, being a colorless, odorless potable liquid that flows in our rivers and streams – is not H_2O. This something would surely seem to be water, although it would not be (given that water is H_2O). Such a situation may be possible, but it is not a situation in which water is not H_2O, because that situation is impossible. We mistake one situation for the other; and this explains why it can seem possible that water is not H_2O. We mistake a possibility in which something that seems to be water is not H_2O for a possibility in which water is not H_2O.

Now, physicalists are committed to the view that certain physical properties are identical to qualia, and old-fact theorists claim that phenomenal concepts pick out the very same properties as certain physical concepts. Because these properties are identical, they are necessarily so; but because one can fail to realize that phenomenal concepts and physical concepts pick out the same properties, it may seem to one that the properties picked out are not the same. Hence, it can seem possible that some physical property P is not identical to a quale Q, even if they are identical. Can we explain this 'seeming' in the same way? Kripke argues that we cannot. His reason is that, although there are situations in which something that seems to be water is not water, there is no situation in which something that seems to be a certain quale is not that quale. Whenever it seems to one, say, that one is in pain, Kripke argues, one is in pain; so we cannot say in this case that we are mistaking a possibility in which something that seems to be pain is not such-and-such a physical property for a possibility in which pain is not that physical property. If this is the correct explanation for why identities can seem contingent, old-fact theorists must still explain why the identities of qualia with physical properties can seem contingent. And if they cannot explain this, one may legitimately suppose that the fact that we can coherently deny these identities means that they simply do not hold – that is, that qualia are not physical properties, and physicalism is false.

Of course, this argument is not Jackson's knowledge argument, but an important related one. Anyone who claims that Mary simply learns an old fact under a new guise when she comes to learn what it is like to have color experiences must have some response to it. Current discussions of the knowledge argument tend to focus on what that response should be. Yablo (2000) and others

have argued that we have independent reasons for thinking that conceivability is not a reliable guide to possibility.

CONCLUSION

As yet, there is no decisive refutation of the knowledge argument. This is no reason for theorists in cognitive science to reassess the import of their research; but it should make them pause to think about the philosophical issues their research raises. We cannot simply dismiss the knowledge argument as a piece of fallacious reasoning. That it has survived all attempts at refutation shows that solving the mind–body problem calls not only for a good empirical theory of mind, but also for clear, careful thinking about such philosophical issues as the nature of facts and the epistemology of modal truths.

References

Bigelow J and Pargetter R (1994) Acquaintance with qualia. *Theoria* **61**: 129–147.

Chalmers D (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Chomsky N (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Churchland P (1985) Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy* **82**: 8–28.

Conee E (1990) Phenomenal knowledge. *Australasian Journal of Philosophy* **72**: 136–150.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Jackson F (1986) What Mary didn't know. *Journal of Philosophy* **83**: 291–295.

Jackson F (1997) Finding the mind in the natural world. In: Block H, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness*, pp. 483–491. Cambridge, MA: MIT Press.

Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Leibniz G (1981) *New Essays on Human Understanding*, translated by P Remnant and J Bennett. Cambridge, UK: Cambridge University Press.

Lewis D (1983) Postscript to 'Mad pain and Martian pain'. In: *Philosophical Papers*, vol. I, pp. 130–132. Oxford: Oxford University Press.

Lewis D (1988) What experience teaches. *Proceedings of the Russellian Society*. Sydney: University of Sydney Press.

Loar B (1990) Phenomenal states. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. IV *'Action Theory and the Philosophy of Mind'*, pp. 81–108. Atascadero, CA: Ridgeview.

Lycan W (1990) What is the 'subjectivity' of the mental? In: Tomberlin J (ed) *Philosophical Perspectives*, vol. IV 'Action Theory and the Philosophy of Mind', pp. 109–130. Atascadero, CA: Ridgeview.

Lycan W (1995) A limited defense of phenomenal information. In: Metzinger T (ed.) *Conscious Experience*. Tucson, AZ: University of Arizona Press.

Lycan W (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **83**: 435–450.

Nemirow L (1990) Physicalism and the cognitive role of acquaintance. In: Lycan W (ed.) *Mind and Cognition*, pp. 490–499. Oxford: Blackwell.

Pereboom D (1994) Bats, brain scientists, and the limitations of introspection. *Philosophy and Phenomenological Research* **54**: 315–329.

Russell B (1910) Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society* **11**: 108–128.

Tye M (1986) The subjectivity of experience. *Mind* **95**: 1–17.

Yablo S (2000) Textbook Kripkeanism and the open texture of concepts. *Pacific Philosophical Quarterly* **81**: 98–122.

Further Reading

Churchland P (1985) Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy* **82**: 8–28.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Jackson F (1986) What Mary didn't know. *Journal of Philosophy* **83**: 291–295.

Lewis D (1983) Postscript to 'Mad pain and Martian pain'. In: *Philosophical Papers*, vol. I, pp. 130–132. Oxford: Oxford University Press.

Loar B (1990) Phenomenal states. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. IV *'Action Theory and the Philosophy of Mind'*, pp. 81–108. Atascadero, CA: Ridgeview.

Lycan W (1990) What is the 'subjectivity' of the mental? In: Tomberlin J (ed) *Philosophical Perspectives*, vol. IV 'Action Theory and the Philosophy of Mind', pp. 109–130. Atascadero, CA: Ridgeview.

Nemirow L (1990) Physicalism and the cognitive role of acquaintance. In: Lycan W (ed.) *Mind and Cognition*, pp. 490–499. Oxford: Blackwell.

Language of Thought

Intermediate article

Georges Rey, University of Maryland, College Park, Maryland, USA

CONTENTS

A language of thought is a language that is proposed by philosophers and psychologists as the medium in which people, and probably many animals, actually think.

WHAT IS A 'LANGUAGE OF THOUGHT' (LOT)

The Computational–Representational Theory of Thought (CRTT)

A language of thought (LOT) (sometimes called 'mentalese') is a language that is proposed by philosophers and psychologists as the medium in which people, and probably many animals, actually think. Although conceivably such a language could be encoded in some dualistic, 'immaterial substance', contemporary interest lies in the further view that it is encoded in the brain, rather in the way formal languages are routinely encoded in computers, a view that has come to be called the 'computational–representational theory of thought'.)

'Thinking' here is a generic word for processes involving propositional attitudes, which are mental states that are usually distinguished by words that take a 'that' or 'to' clause as their direct objects; e.g., 'think that', 'hope that', 'want to'. Such states are discernable in at least two ways: by their contents (e.g. believing *God exists* versus believing *God doesn't exist*), or by the different relations the agent may have to the same such objects e.g. (*believing* versus *hoping* versus *fearing* that God exists). According to CRTT, different propositional attitudes of an agent involve different computational relations to LOT sentences that express the content of the attitude. To a first approximation (cf. Field, 1978):

(CRTT-I) For any agent x, time t and propositional attitude, A, there exists some computationally definable relation C_A such that: x A's that p at t iff there is some sentence in a LOT such that

(i) x stands in some computationally specifiable relation C_A to that sentence that is encoded in x's brain.

and

(ii) that sentence means that p.

For example, *Jane notices that the sun has set* would be true iff she stands in some computationally defined relation to a sentence in her brain that means that the sun has set. CRTT then claims that thought processes consist of computations on those sentences.

The two clauses of (CRTT-I) raise very different issues. Clause (i) requires spelling out the computational architecture of the brain, the multitudinous relations between subsystems e.g. (sensations, perceptions, reasoning, judgments, memory, linguistic competence, wishes and intentions) that comprise the organization of a person's mind. CRTT hopes to capture these relations by specifying computer programs and flow charts indicating where and how different LOT sentences are stored and accessed under various conditions. Thus, a *noticing* might be the output of either a perceptual or a reasoning system that serves as the input to a decision-making system (the material from perception and thought that would provide, with an agent's preferences, a basis for action). Some writers (e.g. Schiffer, 1981) think of CRTT as involving a 'belief box', containing the sentences that express the contents of someone's beliefs. Although the metaphor is harmless, it shouldn't be thought to commit CRTT to any specific form of localization in the brain: a 'sentence' is, after all, simply an abstract structure that can be entokened in an indefinite number of ways: pronounced, written down, encoded in a cipher, or stored in diverse media (e.g. magnetically charged particles on tape or in a computer). They might also be coded by

electrical patterns distributed in different regions of the brain, available for some computations and not others. They need not be parts of any natural, spoken language. The only requirement for a sentence is to be a representation with logico-syntactic structure, such as the sort of grammatical structure defined recursively in logic texts in terms of names, predicates ('is bald'), variables (x, y), connectives (and, only if), quantifiers (all, some), and various operators (probably, necessarily), and that can be set out as a kind of 'spelling'. And so the LOT hypothesis could also be expressed as a hypothesis that the brain has elements in it with causally efficacious, logico-syntactic structure, at least some of whose parts are (per clause (ii)) semantically meaningful. (*See* **Language, Connectionist and Symbolic Representations of**)

Clause (ii) of (CRTT-I) raises the fundamental issue of intentionality: namely, what makes it true that some state means (or 'is about') one thing rather than another, or even anything at all? In the case of sentences of spoken language, a standard approach to the answer would involve citing facts about, for example, the intentions and other mental states of the speaker; i.e. it would involve what is called *derived intentionality*. The question raised by clause (ii), however, concerns how these mental states themselves get their meaning (*original intentionality*) in the first place. Various proposals have been made in philosophy, almost any could be adapted to fit the needs of clause (i).

CRTT'S Rival: Radical Connectionism ('RCON')

CRTT arose in reaction to the failures of 'radical behaviorism', which was the modern version of the associationism advocated by the seventeenth-century empiricist David Hume. Associationism itself has made something of a comeback in the form of 'radical connectionism'. Connectionism in general is an approach to computers that treats them as vast networks of interconnected 'nodes', the connections between which are modified in the light of 'training'. Proponents are attracted by the apparent affinity between such networks and the patterns of neural connections in the brain. However, it should be borne in mind that this affinity is only partial (see Smolensky, 1988), and, in any case, may not reveal anything important about the actual mental architecture of the brain, since a connectionist physical architecture could be merely an implementation of a classical LOT architecture just as (ironically enough) most connectionist programs are actually implemented on classical machines. RCONists reject this mere implementational conception, regarding CRTT models as too rule-bound, rigid, and inefficient, in comparison with the 'softer' connectionist networks (Smolensky, 1988). (*See* **Behaviorism, Philosophical**; **Hume, David**; **Connectionism and Systematicity**)

HISTORY

Suggestions about a LOT can be traced back to William of Occam, Hobbes, Descartes, and Leibniz. Its modern appearance occurs, somewhat obliquely, in the work of Wilfred Sellars (1956/97) and then in Newell and Simon (1972) and Gilbert Harman (1972). It has been most systematically and energetically pursued, however, in the work of Jerry Fodor (1975).

The chief inspiration for Fodor's work was, first, the development by Gottlob Frege of formal logic, and then Alan Turing's conception of a Turing machine, or idealized computer, which shows how, among other things, Frege's logic could be realized as a form of mechanically produced computation. (*See* **Turing, Alan**; **Frege, Gottlob**)

These developments suggest the general idea that thinking consists of applying rules to syntactically specified strings of symbols, or strings of letters specified by their spelling, on the model of proving theorems in symbolic logic. The only difference is that, whereas the rules of logic are ordinarily applied by us consciously following the rules, for CRTT the rules are applied by virtue of the causal structure of the brain. For example, where in elementary logic we follow the rule *modus ponens* – e.g. from 'Dogs bark' and 'If dogs bark, then cats meow' derive 'Cats meow' – for CRTT the brain is so constructed that, if it's in a state that represents the premises, then it is (sometimes) compelled to enter a state that represents the conclusion.

This promises to account for people's quite general ability to reason deductively, because the logically relevant syntactic parts of the states (e.g. the 'if…then…') are presumed to be causally efficacious: it is in virtue of the premises and the conclusion having a certain logico-syntactic form that the transition in thought is made. Indeed, CRTT might equally well be regarded as a *causal* representational theory of thought, since what Turing showed us is that anything that is computational could also be mechanically causal. Given the Turing thesis that anything that is computable is computable by a Turing machine, similar hopes arise for formalizations of nondeductive reasoning: induction, abduction, and decision theory.

In a widely read critique, Searle argued that since computations are specified merely syntactically, they lack semantic properties: they are mere manipulations of 'meaningless symbols'. But this is a logical error: that something is specifiable without certain properties doesn't entail that it fails to possess those properties (e.g. bachelors can be bald, even though what it is to be a bachelor can be specified without mentioning baldness). Indeed, even though the computations of Turing machines are specified syntactically, they standardly have a semantics: for example, the symbols are often numerals representing numbers. In any case, clause (ii) explicitly claims that LOT sentences do have a semantics, and Fodor and others have devoted considerable energy to providing a theory about exactly what that semantics might be.

ARGUMENTS FOR A LOT/CRTT

A number of philosophers (e.g. Davies, 1991; Lycan, 1993; Rey, 1995) have advanced *a priori* arguments for a LOT (based, for example, on the Kantian insight that even the simplest thoughts require something analogous to logico-syntactic structure). However, the chief arguments that have been the focus of discussion are *empirical*, according to which a LOT offers the best explanation of the following phenomena – or at least more promising arguments than its rival, RCON:

The Propositional Structure of Attitudes

It is no accident that attitudes such as thought or belief are standardly picked out by sentences: the thoughts themselves seem to possess something very like a sentential structure. Russell (1903: sect. 54) noticed this when he worried about what he called 'the unity of the proposition', or how to distinguish a proposition from a mere list; e.g., how to distinguish between the thought 'Socrates is bald' and a thought involving a list of the ideas, *Socrates, is* and *bald*. Another example might entail different thoughts such as 'Romeo loves Juliet', 'Juliet loves Romeo', and the list in this case 'Juliet', 'love', 'Romeo'. CRTT captures propositional structure by defining different computational roles for different sorts of symbols (predicates, singular terms, variables, quantifiers) to play. Thus, 'Romeo loves Juliet' is subject to different computations than 'Juliet loves Romeo', as is 'someone loves everyone' versus 'everyone loves someone'.

CRTT and Psychology

Fodor (1975) argues that CRTT is presupposed in most current psychological theorizing. For example, many theories of vision account for visual illusions by presuming that the visual system is engaged in computing inferences about the three-dimensional world on the basis of the evidence provided by the surfaces, contours, and textures of objects detected by retinal stimulation; similarly, theories of language acquisition assume that a hearer is representing phonological features, sentence structures, and rules and principles of alternative grammars; and in theories of decision-making it is presumed that an agent is representing alternative states of affairs and courses of action. In order to systematically represent such information (e.g. *that the dog is in front of the fence, that English is a head-first language*, or *if I don't choose red, I'll either receive a shock or another choice*), the representations must be *sentential*.

But perhaps the appeals these psychological theories make to computation and representation are really just a *façon de parler*, a convenient way of presenting these theories informally and not to be taken as part of their literal truth (see Chomsky, 2000, 2002). However, this is to ignore a difficulty first raised by Fodor (1986): many organisms seem sensitive to arbitrary phenomena in a way that is difficult to explain in physical terms alone. For example, there are the sensitivities of even young children to phonological and grammatical categories of natural language, such as to nouns, verbs, main versus dependent clauses and the like (see, e.g. Pinker, 1994). More generally, Rey (1997) has argued that human beings are sensitive to nonlocal and nonphysical properties (e.g. the age of a rock, the authorship of a book), but how can a *local, physical* system such as a brain exhibit such sensitivities? CRTT attempts to answer this question by suggesting that the system performs computations, not on the properties themselves but on their *representations*. It does this in such a way that allows it, for example, to confirm a hypothesis representing those nonlocal, nonphysical properties by mechanically computing on the representations of local, physical properties (e.g. loudness, pitch).

The Productivity of Attitudes

People seem to be able to think a potential infinitude of thoughts. People can (in principle) think indefinitely complex extensions of simple thoughts. Even children can think 'This is the house that Jack built',

then 'This is the rat that lived in the house that Jack built', 'This is the cat that chased the rat …' and so forth, with apparently no upper bound to the nested clauses, except for such 'performance' factors as patience and memory.

Some theorists have objected to the substantial idealization that this involves, and so Fodor (1987: appendix) proposed a related but more modest claim.

The Systematicity of Attitudes

Anyone who can think a thought, *p*, can think any logical permutation of *p*. For example, if someone can think that 'Ann hates Bob only if Charles loves Di', she can also think that 'Charles loves Di only if Ann hates Bob', 'Di loves Charles only if Bob hates Ann' and so forth, for all permissible logical permutations. CRTT captures both productivity and systematicity by presuming that any system in which the logico-syntactic elements of a LOT are causally efficacious is one in which they are readily available for recombination. Without recourse to such a structure, RCON would seem to be unable to explain such regularities in general (although it might be able to handle particular cases).

Rational Relations Among Attitudes

Many forms of reasoning are patently structure sensitive, involving issues of quantifiers, negations, antecedents, and consequents of conditionals, and application of these things. Again, CRTT's insistence on the causal efficacy of logico-syntactic structure affords a systematic way of explaining such sensitivity, where RCON would seem at a loss, since the only account we know of logical relations in general is a 'Fregean-style' description that exploits such structure.

Irrational Relations Among Attitudes

Thoughts can also be irrational, but the patterns in this case nevertheless can be sensitive to syntactic structure. For example, fallacies in reasoning can be due to confusions surrounding negations, antecedents, and consequents; to necessary versus sufficient conditions; and to the limitations of operators (e.g. confusing 'everyone strives for an end' with 'there's an end for which everyone strives'). Obviously any theory that is committed to the existence of causally efficacious logico-syntactic structures (indicating, for example, the scope of quantifiers) stands a better chance of explaining these patterns than a theory such as RCON that does not have such a commitment.

The (Hyper)intensionality of Attitudes

Propositional attitude ascriptions are 'referentially opaque', or 'intensional' (note difference from 'intentional'): *terms that refer to the very same thing cannot be substituted for one another without risking a change in the truth-value of the whole*. For example, there is a difference in thinking that water is wet, or that H_2O is wet, or that the stuff of rain is wet, despite the fact that water = H_2O = the stuff of rain. CRTT makes a distinction among these attitudes by differenciating syntactically between differing symbolic structures to which an agent can be related. (It can do this even when the structures have the same 'meaning', as in hyperintensional cases such as remembering that a fortnight is a fortnight versus remembering that a fortnight is two weeks; or in cases of indexical thought, where for example the thought that Sam might express by 'I am in danger' is different from the thought he'd express by 'Sam is in danger'.) It's difficult to see how a theory like RCON that doesn't exploit such structural differences could capture these distinctions in thought.

The Causal Efficacy of Attitudes

An RCONist might argue that the above rational, irrational, and fine-grained patterns of thought could be regarded simply as patterns of interpretation of an agent (as in Dennett, 1987; see below), not involving the causally efficacious thought structures posited by CRTT. In contrast , a defender of CRTT is impressed by the causal relations among attitudes, not only '(ir)rationalizing' but also causally explaining their changes of state and behavior in virtue of their syntactic and semantic properties. CRTT can thereby explain the possibility of mental–physical interaction, as well as how a reason may also be a cause. Moreover, CRTT can also explain how mental states may have physical effects that they do not 'rationalize'. For example, Sam's thought 'I am in great danger', expressed by a LOT sentence entokened in his brain, might well cause a release of adrenalin (compare: a computer is so wired that typing out the sentence 'Begin print' causes the printer to print).

The Multiple Roles of Attitudes

Different attitudes can focus on the same thoughts. People often *wish* for the very thing that they *believe* does not presently obtain; they often come to think what previously they only *feared*. CRTT captures this by positing different computational relations

to the same internal representation. Since for RCON theories data structures don't exist separately from the associative processes in which they appear, it is unclear how the same structure that expresses the content of a desire could also express the content of a belief.

SOME COMMON OBJECTIONS TO A LOT/CRTT

Introspection

Many people claim not to 'think in words'. But CRTT is not meant to be establishable by introspection; it is a hypothesis intended merely to explain the phenomena mentioned above, but not necessarily to explain how experience appears.

On a related issue, although some philosophers (e.g. Harman, 1972; Carruthers, 1996) have argued that the LOT is largely one's natural, speaking language, such as English or Chomorro, CRTT does not require that this be the case. Indeed, natural languages, with their manifold syntactic and semantic ambiguities, would seem to be ill-suited to the formal needs of the mechanical computations postulated by CRTT. So, even if one doesn't employ words of one's mother tongue when thinking, or even, as in the case of infants and animals, one has no mother tongue at all, one might still be thinking in a LOT.

On the other hand, CRTT might be able to explain certain introspective phenomena. By careful delineation of the different computational roles of specially restricted representations, some philosophers (Lycan, 1996; Rey, 1997) have argued that CRTT is able to capture the phenomena of subjectivity and even intense sensory experience.

Imagery

Many people think that CRTT conflicts with the evidence that people 'think in images'. However, although images may have some role to play in thought, even defenders of such claims acknowledge that purely imagistic systems don't seem adequate to represent all thought. How can images unambiguously represent logically complex thoughts, such as negations ('there are no ghosts'), conditionals ('if there are ghosts, then …'), and nested quantifications ('every ghost loves another')? Moreover, images are multi-ambiguous in ways that formal languages need not be: an image of a Paris street could represent a general category (cities, pollution), or a particular instance (Paris), or just that very specific street itself, depending upon the sentential context in which it is embedded. It is difficult to think of any serious alternatives to sentences, if the mind (and therefore the brain) is to be capable of representing the full range of human thought. (*See* **Mental Imagery, Philosophical Issues about**)

Homunculi

An extremely common objection to CRTT is that it presupposes precisely those processes it purports to explain. In particular, if there's a LOT, don't we need a 'little man' in the brain to read it (Ryle (1949)? CRTT answers this objection by emphasizing Turing's proposal for computation in general, whereby brute causation replaces human calculation. Specifically, a 'cursor' simply determines whether, for example, a '1' or a '0' is inscribed in a single 'cell' on the machine's 'tape', a purely mechanical task for which no intelligence whatsoever is required.

Artificial Intelligence and the Turing Test

The history of work on computers has led to a number of confusions about the commitments of CRTT that should be disentangled:

- CRTT is not committed to just *any* computer having mental states: only a computer with the very specific underlying structures that psychology and (philosophical) reflection ultimately reveal to be essential to specific mental states and processes. And many defenders of CRTT (e.g. Fodor (2001) himself) are quite pessimistic that those structures are likely to be understood in the near future, if ever.
- Consequently, CRTT is not committed to the widespread projects that go under the rubric of 'artificial intelligence' (AI) that, by and large, are concerned not with actually creating a system with genuine mental states, but with creating systems that can do some of the things that systems with mental states can do, whether or not their systems accomplish these things by means of mental states themselves. Of course, if one were a behaviorist there would perhaps be no serious distinction between these two tasks. And many of those interested in AI have been behavioristically inclined. However, most defenders of CRTT subscribe to *functionalism*, or the view that what is essential to most mental states is not merely how a system outwardly *behaves*, but how it is internally *structured*; and there's seldom any reason to think that the programs proposed by proponents of AI remotely capture the internal structures of genuinely thinking systems (consider 'Deep Blue', the computer that recently beat

the world chess champion, patently not through the kind of subtle strategies of chess masters but simply through working at lightening speed through massive numbers of permutations of possible moves).

- The confusion of CRTT with AI has been aggravated by the fact that, in addition to his work in the theory of computation, Turing also proposed what he regarded as an intuitive test of 'intelligence', what has come to be called the 'Turing test'. In brief, a machine is intelligent, according to this view, iff normal people are not able to distinguish teletype exchanges they might have with the machine from those they might have with a normal human being. Needless to say (at least for any functionalist), this could be (and has been!) accomplished by equipping the machine with a mere bag of simple tricks. Turing, like many thinkers of his time, was in the grip of behaviorism. (*See* **Turing Test**)

In contrast to behaviorism, CRTT is most naturally concerned with the characterization of sub*systems* of the mind, in considerable abstraction from actual behavior. For CRTT, the right way to produce genuine artificial intelligence, as opposed to merely fooling people, would be to construct theories about human cognitive competencies, *vis-à-vis*, for example, perception, scientific reasoning, decision-making, and language comprehension and production. Only afterwards should an attempt be made to integrate these systems into one that might conceivably behave with anything like the sophistication of human beings in a natural environment.

Searle's "Chinese Room"

In a widely discussed article, Searle (1980/91) presented what he and others have often regarded as a devastating blow against CRTT. Briefly, he imagines someone entirely ignorant of the Chinese language being placed in a room with a manual for 'understanding Chinese' according to 'correlation rules' that specify which Chinese characters he is to hand out of the room in response to characters that are handed in, so that, were a normal Chinese speaker to view the entire exchange of characters, she would find it indistinguishable from a normal Chinese conversation. The 'room' (or the person in the room, if that person had memorized the manual) would thus pass the Turing test (see above). However, Searle claims, the person in the room carrying out such a program patently doesn't understand a word of Chinese: he is merely following rules relating one set of (for him) meaningless characters to another. Consequently, Searle concludes, no CRTT program can be sufficient for understanding Chinese, or any other natural language. Indeed, 'not the slightest reason has been given to suppose that [programs] are necessary conditions or even that they make a significant contribution' (1980/91a, p. 511).

A number of problems arise when relating Searle's example to CRTT. As stressed above, there is every reason to suppose that CRTT is not committed to the Turing test, much less to a 'conversation manual' account of linguistic competence. More plausibly, CRTT would require that the system be able to access recursive syntactic and semantic rules along the lines of serious contemporary linguistic theory. It might also require that those rules be embedded in a larger system of perception, reasoning, and decision-making that is plausibly essential to something having any sort of mind at all. Once these conditions are met, it is not clear what might be said about the 'mental states' of the resulting 'room', or of the person in it, who (to pass the test) would have to be frantically consulting probably thousands of computer programs simultaneously, and, moreover, would now be only one small part of an enormously complicated system. Further there would be no reason to assume that the mental states of an entire system would be inherited by all its parts.

Searle is right, however, to point to the need for a theory of meaning as part of any AI effort to produce a genuinely intelligent computer, and that, *pace* the suggestions of some CRTTists (e.g., Jackendoff, 1987), a computer program by itself wouldn't be sufficient for all aspects of meaning. However, CRTTists have not been shy in trying to spell out clause (ii) of CRTT-I, providing at least sketches of how a suitably programmed machine could acquire meaning by being appropriately situated in an environment (e.g. Field, 1978; Fodor, 1987; Lycan, 1987; Neander, 1995; Rey, 1997). But few are under any illusion that any adequate theory of meaning has been supplied for CRTT – or, for that matter, for *any* theory of mind; see Rey (1997: chap. 10) for further discussion.

INTERPRETATIONISM

Searle (1993) has argued that computation exists only in relation to someone who consciously interprets a physical process as a computation. Only conscious agents literally follow rules; computers only seem to 'follow rules' because they are created and used by conscious agents. Therefore, CRTT presupposes the existence of mentation and so cannot explain it.

Davidson and Dennett (1987) had actually suggested a variant of this view some years earlier, except for them, unlike Searle, consciousness fails to provide any help. They argue that attitude

ascriptions are not actually determinately true of anything, not even of human beings, but are merely 'interpretations' of behavior that facilitate prediction of behavior. Consequently, such ascriptions do not require corresponding sentences to be in anyone's head. For example, we might say of a chess-playing computer that 'it likes to get its queen out early', even though no corresponding sentence might exist in the computer's program.

Both Searle's and Dennett's views, however, need to explain why ascriptions of rules and attitudes manage to be so explanatorily successful, and why both machines and human beings display their intelligent regularities. Perhaps in some cases (e.g. a thermostat) the explanation can proceed without taking such ascriptions as being literally true. Moreover, a CRTTist could argue that, without an internal sentence, the ascription of an attitude isn't literally true. However, it is by no means obvious that the regularities in the behavior of all modern computers, or the behavior of people and animals, can seriously be explained without taking the ascriptions literally. Again, how are we to explain the sensitivities of people to nonlocal, nonphysical phenomena?

HOLISTIC CONFIRMATION

Interestingly, the very philosopher who has been the most vocal champion of a CRTT has also argued for its severe limitations. Fodor (1983, 2001), following Quine (1956/76; see also Quine and Ullian, 1978), argues that the confirmation of most hypotheses, both in commonsense and science, seems to be both 'Quinean' (or holistic: a hypothesis is assessed by the overall coherence it brings to a system as a whole), and 'isotropic' (any belief in the system is potentially relevant to the (dis)confirmation of any other belief). The very locality feature that is part of the explanatory attractiveness of Turing machines appears to be incompatible with these global features of reasoning. However, whether this apparent incompatibility is a real one should perhaps be regarded as an open question, pending a deeper understanding of the considerable resources of Turing computation.

In any case, as Fodor would be the first to insist, the problems here would at worst show that while CRTT is not sufficient as an account of mental processing, it is still necessary in any such account.

References

Carruthers P (1996) *Language, Thought and Consciousness*. Cambridge, UK: Cambridge University Press.

Chomsky N (2000) *New Horizons in the Study of Mind and Language*. Cambridge, MA: MIT Press.

Chomsky N (2002) Reply to Rey. In: Antony L and Hornstein N (eds) *Chomsky and His Critics*. Oxford, UK: Blackwell.

Davies M (1991) Concepts, connectionism, and the language of thought. In: Ramsey W, Stich S and Rumelhart W (eds) *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT/Bradford Books.

Field H (1978) Mental representation. In: Block NJ (ed.) *Readings in the Philosophy of Psychology*, vol. 2. Cambridge, MA: Harvard University Press.

Fodor J (1975) *The Language of Thought*. New York, NY: Crowell.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor J (1986) Why paramecia don't have mental representations. *Midwest Studies in Philosophy* **X**: 3–24.

Fodor J (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor J (2001) *The Mind Doesn't Work That Way: The Scope and Limits of Cognitive Science*. Cambridge, MA: MIT Press.

Harman G (1972) *Thought*. Princeton, NJ: Princeton University Press.

Jackendoff R (1987) *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Lycan W (1987) *Consciousness*. Cambridge, MA: MIT Press.

Lycan W (1993) A deductive argument for the language of thought. *Mind and Language* **8**(3): 404–422.

Maclaughlin B and Warfield T (1994) The allure of connectionism re-examined. *Synthese* **101**: 365–400.

Neander K (1995) Misrepresenting and malfunctioning *Philosophical Studies* **79**: 109–141.

Newell A and Simon H (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Pinker S (1994) *The Language Instinct*. New York, NY: Harper & Row.

Quine W (1956/76) Carnap and logical truth. In: *Ways of Paradox and Other Essays*, 2nd edn, Cambridge, MA: Harvard University Press.

Quine W and Ullian J (1978) *The Web of Belief*. New York, NY: Random House.

Rey G (1995) A not 'merely empirical' argument for the language of thought'. In: Tomberlin J (ed.) *Philosophical Perspectives*, vol. 9, *AI, Connectionism, and Philosophical Psychology*, pp. 201–222, Atascadero, CA: Ridgeview Press.

Rey G (1997) *Contemporary Philosophy of Mind: a Classical Approach*. Oxford, UK: Blackwell.
Russell B (1904/1938) *Principles of Mathematics*. New York, NY: WW Norton.
Ryle G (1949) *The Concept of Mind*. London, UK: Hutchinson.
Schiffer S (1981) Truth and the theory of content. In: Parret H and Bouvaresse J (eds) *Meaning and Understanding*. Berlin, Germany: Walter de Gruyter.
Searle J (1980/91a), Minds, brains and programs. In: Rosenthal D (ed.) *The Nature of Mind*, pp. 509–519. Oxford, UK: Oxford University Press.
Searle J (1980/91b) Intrinsic intentionality: reply to criticisms of minds, brains and programs. In: Rosenthal D (ed.) *The Nature of Mind*, pp. 521–523. Oxford, UK: Oxford University Press.
Searle J (1993) *The Rediscovery of the Mind*. Cambridge, UK: Cambridge University Press.
Sellars W (1956/97) *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.
Smolensky P (1988) A proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–23.

Levels of Analysis, Philosophical Issues about

Intermediate article

Allen Y Houng, National Yang-Ming University, Taiwan

CONTENTS

The world, including cognitive systems, is organized into hierarchies of levels. Levels of analysis represent different stances from which predictive and explanatory theories are constructed.

WHAT IS A LEVEL OF ANALYSIS?

The term 'level' can be used in several different ways. People talk about levels of existence, levels of organization, levels of complexity, levels of description, levels of analysis, levels of explanation, levels of abstraction, and so on. Levels of existence, levels of organization, and levels of complexity represent the integrated organization of a system. The other usages in the above list have to do with the ways we look at or theorize about the world. These two aspects of the concept of levels – i.e. levels as natural qualities and levels as conceptual or theoretical constructs – are concerned with two different sets of issues.

Organizing the world into hierarchical levels is quite natural to most scientists and philosophers. Many claim that the prevailing 'synthetic' approach to biology is too simple and reductionistic, and that an evolutionary theory must take into account the hierarchies found in nature. In cognitive science, levels of organization arise in the anatomical structure of the brain. Exactly how many levels of organization there are in the brain is an empirical question. The most commonly postulated levels of organization of the brain include, from the bottom up, molecules, synapses, neurons, local networks, layers and columns, topographic maps, and systems (Churchland and Sejnowski, 1992).

Corresponding to each level of organization, there is presumably a level of analysis, which can be taken as a stance or viewpoint, or as a conceptual framework from which theories can be constructed. A level of organization consists of a collection of entities, properties, relations, processes, causal regularities, and other relevant phenomena; a level of analysis consists of a collection of terms, predicates, theories, and so on. A level of analysis can be specified as a set of conditions or principles for classification of theories (or models). It must at least contain conditions specifying what sort of entities and structures are to be considered, what sort of regularities are to be captured, and what phenomena are to be explained or predicted. Hence the specification of a level of analysis resembles a very general specification of a research program, under which several competing or complementary theories can be considered as belonging to the same level.

PHILOSOPHICAL VIEWS OF LEVELS OF ANALYSIS

The central philosophical issue concerning the levels of analysis in cognitive science is the relation between theories at different levels of analysis. Reductionism, for example, claims that higher-level theories such as those of psychology can be reduced to neuroscience. We should note that this refers to 'interlevel' reduction, in which the domains of reducing and reduced theories at different levels are different. The other notion of reduction is 'intralevel' reduction, in which the domains of reducing and reduced theories are the same. Examples of intralevel reduction include the reductions of Newtonian mechanics to special relativity, and of macroscopic thermodynamics to statistical thermodynamics. In the case of intralevel reduction, a reduced theory is regarded as a special case of or an approximation to its reducing theory. On the other hand, in the case of interlevel reduction, a reduced theory is regarded as a derived consequence of its reducing theory, with some bridge laws associating entities in the two domains. Although intralevel reduction has been successful in many domains, successful interlevel reduction is very rare.

Levels in a hierarchically organized system are integrated in a certain way; they are not mutually independent. For example, the cellular level and the molecular level in a biological system are not separate subsystems. Processes and properties at two different levels, especially at two adjacent levels, causally interact with one another. The lower-level processes or properties realize those at higher levels.

A higher-level cognitive function can be realized by a variety of different lower-level substrata. But this does not imply that lower-level theories cannot be used to explain the nature of higher-level cognitive functions. In other words, the multiple realizability of cognitive functions does not imply that neuroscience is irrelevant to cognitive modeling. The lower-level properties and processes do impose severe constraints on the space of possible realizations of higher-level properties or processes. For example: the processing speed in the brain is limited; most synaptic connections are between, not within, cell classes; and the brain is a highly parallel machine. These lower-level properties impose constraints on the realization of higher-level cognitive functions in the brain. If lower-level constraints are not taken into consideration, the result could be very speculative higher-level theories (Sejnowski and Churchland, 1989).

Downward integration takes the opposite direction. Higher-level processes or properties constrain, organize, regulate, or coordinate the otherwise unstructured or unorganized lower-level processes or properties. Some higher-level properties are referred to as 'emergent' because they are not apparently reducible to lower-level properties. These higher-level emergent properties impose structure or organization on lower-level processes or operations. They constrain the lower-level components to be such as to generate certain integral patterns at higher levels.

One purpose of higher-level analysis is to discern the structures or organizations required for lower-level mechanisms to realize higher-level regularities, or law-like behaviors, which would be unlikely without these higher-level constraints namely, the structures or organizations mentioned above. For example, consider Rumelhart and McClelland's model of the acquisition of the English past tense (Rumelhart and McClelland, 1986). This model is considered a successful low-level model for learning the past tense. It has been argued that its success is due to its implicit assumption of a higher-level linguistic theory. The role of higher-level theories is to provide organizational or structural constraints on lower-level theories (Pinker and Prince, 1988).

Not all theorists agree on the necessity for higher-level constraints in developing lower-level theories. Eliminative materialism (represented in the works of P. M. Churchland and P. S. Churchland) claims that psychology should be replaced by neuroscience. Although P. S. Churchland sees the relationship between the theories of psychology and neuroscience as co-evolving rather than strictly reductionistic, she maintains that psychology that commits to propositional attitudes is in principle replaceable by neuroscience, and that co-evolution between psychology and neuroscience, will eventually lead to the elimination of folk-psychological categories (Churchland, 1986).

ROLES OF LEVELS OF ANALYSIS IN COGNITIVE SCIENCE

Researchers in cognitive science employ the concept of levels for various purposes. Pylyshyn (1984) argues that at least three distinct levels of analysis are needed: the physical level, the symbolic (or functional) level, and the semantic (or representational) level. At the physical level, generalizations can be formulated only in the language of physical science. At the symbolic level, generalizations are expressible in terms of predicates referring to the properties of the functional architecture. (The functional architecture of a system is analogous to the operating system of a computer.) At the semantic level, generalizations can be formulated in terms of the contents of mental representations. For example, in a computer system, a physical level of analysis would be sufficient to describe the electronic properties of a computer. But to understand how a program works, we need to appeal to the architecture of the underlying virtual machine. And to understand how a program acts as a model of certain cognitive phenomena, we need to consider the contents of the symbols used in the program.

Marr's three levels of analysis for the science of the mind are also well known. These are the computational, the algorithmic, and the implementational levels. Marr's algorithmic level is similar to Pylyshyn's symbolic level, and the implementational level is similar to the physical level. The computational level is more difficult to understand. According to Marr, it is concerned with abstract and theoretical issues. Marr alludes to Chomsky's distinction between competence and performance and hints that the computational level is somewhat

like a competence theory of the mind. It is a formal analysis of the nature of the mind. With the aid of this level of analysis, Marr wishes to eliminate the ad hoc assumptions and unprincipled restrictions of a model, and a misguided and erroneous level of analysis in cognitive modeling. This is a level of theoretical and methodological consideration (Marr, 1982).

Hofstadter (1979) offers an extensive but somewhat perplexing discussion of levels of analysis. He claims that the explanations of emergent phenomena in our brains should be based on a kind of 'strange loop', a two-way causal interaction between levels. The bottom level in his hierarchy is the neural level. Hofstadter calls it the 'inviolate level' because the way neurons operate in the brain at that level is not changeable by mental activities. Above the inviolate level, there is the 'tangled level', at which the 'tangled hierarchy' resides. The tangled level should be further divided into a hierarchy of levels, since we need a distinct level of analysis to capture the generalizations of each level in the strange loop. But since we use the same kind of language for all cognitive levels, levels become mixed and incorrect theories are articulated. To solve this problem, Hofstadter proposes a multi-level approach to the mind, and suggests that there could be many different undiscovered levels of analysis within the tangled level employing different languages for cognitive modeling.

The debate between symbolism and connectionism in cognitive science since the 1980s is concerned with the cognitive architecture of the mind. The concept of levels of analysis plays a central role in this debate because the choice of level of analysis for cognitive theorizing constrains the choice of cognitive architecture.

Symbolism claims that any adequate cognitive model has to assume a symbolist architecture, i.e. mental representations with constituent structure and mental operations sensitive to that structure. According to this view, cognitive processes operate on symbolic mental representations in accordance with the constituent structure of those representations, and this class of cognitive models defines the proper level of analysis for cognitive modeling. Any model outside this class is either cognitively irrelevant or just an implementation of a model within the class (Fodor and Pylyshyn, 1988).

The level of analysis for distributed connectionist models is neither the neural nor the symbolic level, but rather the subsymbolic level. Smolensky (1988) contends that precise, formal descriptions of the dynamics of cognitive behavior are possible only at the subsymbolic level. The processing units at the subsymbolic level have no semantics: semantic content can be attributed only to patterns or representations residing at the symbolic level. This is a dual-level view of cognitive architecture, in contrast with the single-level view of symbolism.

Fodor and Pylyshyn argue that since connectionist architecture does not allow cognitive processes to operate in accordance with the syntactic structure of mental representations, important cognitive phenomena such as systematicity, compositionality, and inference coherence cannot be explained by connectionism. Therefore, connectionist architecture is not adequate for modeling the mind. Connectionists have argued, with some success, that connectionist architecture can explain compositionality and other important mental phenomena without assuming a symbolic architecture.

Besides symbolism and connectionism, there are two other views of cognitive architecture that deserve mention. The 'hybrid' approach supports a form of division of labor, dividing cognitive functions into higher functions, such as thinking and reasoning, and lower functions, such as perceptions and skills. Symbolist architectures are best suited for the higher functions, and connectionist architectures are best suited for the lower functions. The basic idea of the 'multi-level' approach is to regard a cognitive system as a system organized into a hierarchy of many levels. The discipline of cognitive science involves cultural studies, anthropology, sociology, psychology, artificial intelligence, biology, neuroscience, chemistry, and physics. This diversity may suggest that cognitive science is a special kind of science in that it incorporates theories at many different levels of analysis.

References

Churchland PS (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Churchland PS and Sejnowski JJ (1992) *The Computational Brain*. Cambridge, MA: MIT Press.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Hofstadter DR (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic Books.

Marr D (1982) *Vision*. San Francisco, CA: W. H. Freeman.

Pinker S and Prince A (1988) On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**: 73–193.

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Rumelhart DE and McClelland JL (1986) On learning the past tenses of English verbs. In: Rumelhart DE,

McClelland JL and the PDP Research Group (eds) *Parallel Distributed Processing: Explorations in Microstructures of Cognition*, vol. II 'Psychological and Biological Models', pp. 216–271. Cambridge, MA: MIT Press.
Sejnowski JJ and Churchland PS (1989) Brain and cognition. In: Posner MI (ed.) *Foundations of Cognitive Science*, pp. 301–356. Cambridge, MA: MIT Press.
Smolensky P (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–74.

Further Reading

Bechtel W (1994) Levels of description and explanation in cognitive science. *Mind and Machine* **4**: 1–25.
Campbell DT (1976) 'Downward causation' in hierarchically organized biological systems. In: Ayala FJ and Dobjhansky T (eds) *Studies in the Philosophy of Biology*, pp. 179–186. Berkeley, CA: University of California Press.
Churchland PM (1989) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.
Clark A (1997) *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
Dennett DC (1987) *Intentional Stance*. Cambridge, MA: MIT Press.
Fodor JA (1974) Special sciences. *Synthese* **28**: 97–115.
Friedman K (1982) Is intertheoretic reduction feasible? *British Journal of Philosophy of Science* **33**: 17–40.
Hartfield G (1999) Mental functions as constraints on neurophysiology: biology and psychology of vision. In: Hardeastle VG (ed.) *Where Biology Meets Psychology: Philosophical Essays*, pp. 251–271. Cambridge, MA: MIT Press.
McCauley RN (1996) Explanatory pluralism and the co-evolution of theories in science. In: McCauley RN (ed.) *The Churchlands and Their Critics*, pp. 17–47. Cambridge, MA: Blackwell.
McCauley RN (1998) Levels of explanation and cognitive architecture. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 611–624. Oxford, UK: Blackwell.
Newell A (1982) The knowledge level. *Artificial Intelligence* **18**: 87–127.
Smolensky P (1991) Connectionism, constituency, and the language of thought. In: Loewer B and Rey G (eds) *Meaning in Mind: Fodor and his Critics*, pp. 201–227. Cambridge, MA: Blackwell.
Van Gelder T (1990) Compositionality: a connectionist variation on a classical theme. *Cognitive Science* **14**: 355–384.
Wimsatt WC (1976) Reductionism, levels of organization, and the mind–body problem. In: Globus G *et al*. (eds) *Consciousness and the Brain*, pp. 199–267. New York, NY: Plenum Press.

Materialism

Intermediate article

William G Lycan, University of North Carolina, Chapel Hill, North Carolina, USA

CONTENTS

According to the doctrine of materialism, the mind is no more than a complex arrangement of physical matter. This is maintained by most philosophers of mind and presupposed by cognitive science, but it is difficult to make precise and it faces some formidable objections.

INTRODUCTION

Materialism is, in full, the thesis that only physical matter exists. However, in the context of cognitive science and the philosophy of mind, the term is used to mean the less ambitious claim that nothing psychological or mental exists that is not entirely constituted by physical matter. Traditionally, the main opposing theory has been mind–body dualism of the kind defended by Descartes, according to which minds are purely spiritual and entirely nonspatial, having neither size nor location nor any other physical property. According to the dualist view, a normal living human being or person is an immaterial mind somewhat mysteriously paired with a physical body.

A less radical opposing theory is 'property' dualism, which does not posit immaterial things or beings such as Cartesian minds, but does maintain that human beings and other sentient creatures have some nonphysical or immaterial properties, i.e., properties that are not constituted just by an arrangement of physical matter. Such properties might include the content properties of thoughts, such as my having a thought about the Wife of Bath, or phenomenal properties of sensations, such as the mountains subjectively looking blue to me.

At first sight, these formulations seem fairly clear, but philosophers have had trouble making them precise enough to create a genuine contrast between the materialist and the dualist views.

WHAT IS MATERIALISM?

How are we to understand the phrase 'physical matter'? (This is what Montero (1999) calls the 'body problem'.)

The things that we ordinarily think of as physical objects are made of tiny particles, which are the subject matter of microphysics. One might, therefore, simply define 'physical matter' as matter that is entirely composed of such particles. But this suggestion immediately meets a well-known problem. By 'microphysics', should we mean the present microphysical theory, or should we mean some (unspecified) improved or idealized theory? Suppose the former. Then there is probably no 'physical matter' in the sense defined, since the present microphysical theory is probably incorrect in some respects. But if 'microphysics' does not mean the present microphysical theory, what improved or idealized theory does it mean?

The most common answer is: 'final microphysics' – not in the sense of whatever microphysical theory will in fact be the last to be offered by a living scientist, but of an ideal theory that would be true and would explain everything dynamic and kinematic that needs explaining. One may suspect that the terms 'dynamic' and 'kinematic' already presuppose the concept of physical matter. But the main problem for the appeal to final microphysics is how to assure ourselves that such an ideal theory would contain only things and properties that we now consider physical. It is tempting to think that fundamental physics is largely finished; but that has been thought many times before and always wrongly. Physics has persistently encountered new phenomena, and acquired new theories, in a way that shocks the older generation conceptually: action at a distance; electromagnetism; relativity; Riemannian space-time; quantum indeterminacy; the paradoxes of quantum mechanics;

tachyons; antimatter. Already some quantum physicists have gestured towards panpsychism – the view that every bit of matter, however small, has some irreducibly mental properties – by positing a 'consciousness' that resolves quantum indeterminacies; and some philosophers follow Sellars (1981) in arguing that physicists will ultimately have to recognize some irreducibly mental properties along with their other physical primitives. The result is that if one defines 'matter' by reference to final microphysics, it may follow that 'materialism' is true despite the existence of irreducibly mental entities or properties. Thus would be unacceptable to most current 'materialists'.

A second strategy for maintaining the distinction between materialism and dualism is to follow Descartes in making spatiality the criterion. The physical, we may say, is the spatiotemporal, meaning that to be physical is to be located within the same space-time as are London, North Carolina, the Andromeda galaxy, the Moon, Julius Caesar's left elbow, your body and mine.

But it is not obvious that, necessarily, everything spatiotemporal is physical. Ghosts and disembodied spirits supposedly move about in space; and space-time points themselves are arguably abstract rather than physical objects. Moreover, some dualists have insisted that their posited immaterial items have spatial location.

A third strategy is to abandon the attempt at a direct characterization of 'physical matter', and appeal merely to sameness of composition: materialists hold that creatures with minds are made entirely of the same ultimate components as are ordinary inanimate objects, and that their properties are entirely constituted by the ways in which those components are arranged and related to external things. Descartes would not have accepted that claim.

This strategy faces two objections. The first is that 'materialism' as thus defined is compatible with panpsychism, since if every rock and every subatomic particle has some irreducibly mental properties, sentient creatures are not thus distinguished from rocks; but panpsychism is a form of dualism. The second objection, probably decisive, is that it might turn out, scientifically, that a distinctive kind of basic particle occurs only in the brains of human and other sentient beings. In that case, even if the particle in question were purely physical by any standard, 'materialism' in the present sense would be refuted. This scenario seems unlikely; but even hypothetically it should not refute materialism in its proper sense.

A fourth strategy, suggested by Campbell (1967), is to start with some paradigmatically mental properties, or (better) a list of all the known mental properties, and some paradigmatically physical ones, and then characterize dualism as bluntly separating the two at the level of fundamental entities. 'Materialism' would then be the claim that none of the world's basic components has any of the mental properties; anything with mental properties must be composed only of basic elements that individually do not have them, and its mental properties must consist entirely in an arrangement of these basic elements.

Here there is no appeal to final microphysics, or to spatiality, or to sameness of composition. And if final microphysics should turn out to posit mentalistic properties, then materialism as here characterized would be refuted, as is right.

This strategy does run into a version of the main objection to the first: Even if none of the world's basic components has any of the known mental properties, basic particles might turn out to have strange properties that seem more like the known mental properties than like paradigmatic physical ones. But this does not seem as serious as the objections to the other strategies.

VARIETIES OF MATERIALISM

In the twentieth century, there were three main materialist philosophical theories of the mind: behaviorism (*c.* 1930–1960), the type identity theory (1960–1970), and functionalism (from 1970).

Reacting against what they considered to be scientifically implausible features of dualism, the analytical behaviorists claimed that mental ascriptions simply mean things about behavioral responses to environmental impingements. Thus, 'Roy is in pain' means nothing about Roy's so-called inner life or any episode taking place within Roy, but just that Roy either is actually behaving in a wincing-and-groaning sort of way or is disposed to behave in that way (i.e., he would if something else were not preventing him). It should be noted that a behaviorist need make no claim about the meanings of mental expressions. One might be a reductive behaviorist, and hold that although mental ascriptions do not just mean things about behavioral responses to stimuli, they are, in reality, made true just by properties of actual and counterfactual responses to stimuli. Or one might be an eliminative behaviorist, and hold that there are no mental states or events at all, but only behavioral responses to stimuli, all mental ascriptions being false or meaningless. But to any behaviorist, what

has come to be called the 'Turing test' is decisive: psychological differences cannot exceed behavioral differences. Organisms (as well as computers and other machines) whose actual and counterfactual behavior is alike must be psychologically alike.

Philosophical behaviorism adroitly avoided a number of nasty objections to Cartesian dualism, but it could not allow that any mental states and events are genuinely inner and genuinely episodic. The type identity theorists overcame this shortcoming, denying that mental states are to be identified with outward behavior or even with hypothetical dispositions to behave. But, contrary to the dualists, they claimed that episodic mental items are not ghostly or nonphysical either, but neurophysiological. They are numerically identical with states and events occurring in their owners' central nervous systems. To be in pain is to have (for example) one's c-fibers firing; to believe that erysipelas is nasty is to have one's B_{en}-fibers firing, and so on.

By making the mental entirely physical, the type identity theory, like behaviorism, avoided the serious objections that plagued dualism; but it also accommodated the inner and the episodic. For according to the type identity theory, mental states and events actually occur in their owners' central nervous systems; hence they are inner in an even more literal sense than could be granted by Descartes. The type identity theory also abandoned the Turing test, admitting that organisms could differ mentally despite total behavioral similarity (since clearly organisms can differ neurophysiologically in mediating their outward stimulus–response regularities).

However, Putnam (1960) pointed out a presumptuous implication of the type identity theory understood as a theory of 'types' or kinds of mental items: that a mental state such as pain has always and everywhere the neurophysiological characterization initially assigned to it. For example, if pain is identified with the firings of c-fibers, it follows that a creature of any species (even extraterrestrial) could be in pain only if it had c-fibers and they were firing. But such a constraint on the biology of any being capable of feeling pain is both gratuitous and indefensible: why should we suppose that any organism must have the same biochemistry as we, in order to have what can be accurately recognized as pain? The type identity theorists had overreacted to the behaviorists' difficulties and focused too narrowly on specifically human biology.

Putnam advocated the obvious correction: what was important was not that c-fibers were firing, but what those firings were doing, i.e. what their firing contributed to the operation of the organism as a whole. The role of the c-fibers could have been performed by any mechanically suitable component, without changing the psychology of the containing organism. Thus, to be in pain is not *per se* to have c-fibers that are firing, but merely to be in some state or other, of whatever biological description, that plays the same causal role as did the firings of c-fibers in the human beings we have investigated. We may continue to maintain that pain 'tokens' – individual instances of pain occurring in particular subjects at particular times – are strictly identical with particular neurophysiological states of those subjects at those times, namely, with the states that happen to be playing the appropriate roles. This is the thesis of 'token identity' or 'token physicalism'. But pain itself can be identified only with something more abstract: the causal or functional role that c-fiber firings share with their potential replacements or surrogates. Types of mental state are identified not with types of neurophysiological state but with more abstract functional roles, as specified by state-tokens' causal relations to the organism's sensory inputs, motor outputs, and other psychological states.

Putnam compared mental states to the functionally or computationally characterized states of a computer: just as a computer program can be realized or instantiated by various physically different hardware configurations, so can a psychological 'program' be realized by various organisms of different physiochemical composition, and that is why different physiological states of organisms of different species can realize one and the same type of mental state. Where a type identity theorist's type-identification would take the form 'to be in a mental state of type *M* is to be in the neurophysiological state of type *N*', Putnam's functionalism (as it is called) has it that to be in *M* is to be in some physiological state that plays role *R* in the relevant computer program (i.e., the program that at a suitable level of abstraction mediates the creature's total outputs given its total inputs, and so serves as the creature's global psychology). The physiological state 'plays role *R*' in that it stands in a set of relations to physical inputs and outputs and other inner states that matches (in a one-to-one fashion) the abstract relations between inputs, outputs and computational states codified in the computer program.

There is another version of functionalism that appeals to 'function' not in the computational sense but in the biological sense. In this version the role with which a mental state is to be identified is characterized teleologically, in terms of what the

state, or its containing neurophysiological device, is for: what it contributes to the subject's behavioral capabilities.

Of course, one need not apply the same theory to all kinds of mental state. One might, for example, be a behaviorist about beliefs, a type identity theorist about very specific sorts of sensation, and a functionalist about visual perception.

Two minority materialist views should be mentioned. One is anomalous monism (Davidson, 1970), according to which mental 'types' are irreducible. The anomalous monist grants, indeed insists on, the truth of the 'token identity' thesis, but maintains that mental types are *sui generis*: there is nothing reductively or nontrivially common to various people's beliefs that sugar is sweet, even though each of those tokens is identical with some particular token of a brain state. Mental types form a closed family, and we ascribe them according to their own proprietary epistemology.

Finally, there is eliminative materialism (Churchland, 1981), the view that mental states do not exist. Defenders of this paradoxical thesis argue that the common-sense concept of the mental is vague and incompatible with a scientific view of human beings. They claim that the common-sense or 'folk' theory of human behavior in which our mental concepts are embedded is, on various grounds, a bad theory, and should be replaced by a better and scientifically more successful one; moreover, they contend, the folk theory requires entities of a kind that are unlikely to be found in our biological structure.

ARGUMENTS FOR MATERIALISM

Four main arguments have been put forward.

First, physics presupposes a closed causal order entirely governed by precise (possibly stochastic) mathematical laws. Moreover, if one is trying to explain a physical phenomenon, one is methodologically required to seek a physical explanation; only if the phenomenon is inconceivably strange may one venture to suggest a nonphysical, immaterial or ghostly explanation. Thus, if materialism is not true, something must be wrong with physics as we know it. And it would be surprising if physics could be criticized and corrected by purely philosophical argument.

Second, there is the famous interaction problem, which even Descartes could not solve to his own satisfaction: minds and mental events cause physical behavior, and mental events and states are caused by physical impingements on subjects' sense organs. How is it possible, or even imaginable, for an entirely immaterial entity thus to interact causally with the physical world? The difficulty is not much less acute for the property dualist; for how can nonphysical properties interact with physical ones, given (again) the causally closed nature of the world as described by physics?

Third, philosophers talk of 'supervenience': a type of property *supervenes* on another type of property if the distribution of properties of the first type is entirely determined by the distribution of properties of the second type. Thus, chemical properties supervene on microphysical ones: if two possible worlds are exactly alike microphysically, they must also be alike chemically. Now it seems to most philosophers, even to some dualists, that mental properties supervene on (broadly speaking) physical ones: if two possible worlds are physically exactly the same, then they cannot differ mentally. The obvious explanation (though not the only conceivable explanation) of this supervenience is that mental properties are just complexes of physical properties.

Fourth, Armstrong (1968) maintained that mental terms were defined causally, in terms of mental items' typical causes and effects. For example, the word 'pain' simply means a state that is typically brought about by physical damage and that typically causes withdrawal, complaint, desire for cessation, and so on. Now if by definition pain is whatever state plays a certain causal role, and if, as is likely, scientific research reveals that that particular role is in fact played by some neurophysiological state, it follows that pain is that neurophysiological state.

PROBLEMS FOR MATERIALISM

The difficulties facing the materialist position fall into two main categories, corresponding to the two most distinctive features of mental states and events. Philosophers call these features, respectively, *intentionality* and *qualia*.

The term 'intentionality' is used by philosophers, not as applying primarily to actions, but to mean the property of being about something. Mental states and events are intentional in this technical sense. For example, beliefs and desires and hopes are about things, or have 'intentional objects': I have beliefs about B. F. Skinner; I want a ham sandwich and a cure for my hypertension; and I hope for peace on earth. What is both most distinctive and most troublesome about intentionality is its indifference to reality. An intentional object need not actually exist or obtain: the Greeks worshipped Athena; a friend of mine believes that corks grow on trees; and

even if I get the sandwich, my hope for world peace is probably going to remain unfulfilled.

For materialists, the problem is that it is hard to see how a purely physical or material object could have intentional properties. How could any purely physical entity or state have the property of being about a nonexistent state of affairs? Surely, the argument goes, a brick cannot do that, nor a subatomic particle, nor a collection of subatomic particles, however large.

But materialists are not without resources here. One is the computer or information-processing model of the mind, put forward by Putnam as mentioned above and endorsed by almost all cognitive psychologists. The internal states of a computer make reference to external things, depending on the subject matter of the computing and what program is running. So there can no longer be much doubt that physical states of a purely physical device can be about things, including nonexistent things. Of course, computer states are about things only by virtue of the intentional states of the human beings who design and use the computers; so it remains to be seen whether an adequate materialist account can be given of the intentional contents of human brain states.

The qualitative character of experience also poses a major set of problems for materialism. The *quale* of a mental state or event is that state or event's subjective feel, its introspectible 'phenomenal character'; for example, the distinctive felt quality of a backache, or the vivid green colour of an afterimage (even when one knows that the green afterimage is the phantom by-product of a red light flashing). Many philosophers have objected that no materialist metaphysics can explain, illuminate, acknowledge or even tolerate the notion of what it subjectively feels like to be in a given kind of mental state (Nagel, 1974; Jackson, 1982; Chalmers, 1996). Yet, say these philosophers, the feelings are quintessentially mental; they make the mental states what they are. Something, therefore, must be drastically wrong with materialism.

This suspicion has given rise to a number of distinct objections to materialism, among which are the following.

1. Early critics of the type identity theory argued that the fact that we have immediate mental access to qualia suggests that they are not features of any purely behavioral or neurophysiological item. Surely we do not have immediate, transparent access to (even our own) physical states.
2. Many philosophers have proposed counterexamples to various materialist views: hypothetical examples in which some creature seems to satisfy all the right materialist conditions but whose qualia are of the wrong kind, or which lacks mentality or one of its crucial aspects entirely (Block, 1978; Chalmers, 1996).
3. Nagel (1974) argues for the existence of a special, intrinsically perspectival kind of fact, the fact of 'what it is like' to have a mental experience of a given kind, which intractably and in principle cannot be captured or explained by physical science. Jackson (1982) contends that a brilliant scientist could know all the scientific and other physical facts about color and about people seeing color, yet not personally know what it is like to see green; thus, there is a kind of fact that materialism leaves out.
4. It seems that in consciousness we are presented with mental individuals that themselves have phenomenal, qualitative properties. As before, when a red light flashes in front of you, your visual field contains a green after-image – which is a thing that is really green, has a fairly definite shape and exists for a few seconds before disappearing. If there are such things, they are entirely different from anything physical to be found in the brain of a (healthy) human subject. In fact, if there are such things then materialism is false, since they themselves are immaterial entities. At least, it is hard to see how a purely physical brain could apprehend them.
5. Even if God were to assure us that, say, the Type Identity theory is true, there would still be a problem: Why on earth does such-and-such a firing of fibers going on in my brain feel to me like … *this*?—or for that matter, like anything at all? It seems inexplicable, an insoluble mystery (Levine, 1993).

This is a formidable set of objections, and each seems plausible. Needless to say, materialists have responded at length: there is an enormous literature on objections to materialism based on qualia.

MATERIALISM AND COGNITIVE SCIENCE

Cognitive science is founded on the computer model of the mind, which is also what, on some accounts, allowed materialism to account for intentionality.

It is widely thought that cognitive science, like biology, chemistry and physics, presupposes materialism. Methodologically, that is certainly so: no cognitive scientist is allowed to sidestep a difficulty by appealing to an immaterial entity or mechanism ('at this point a miracle occurs'). But cognitive science does not itself entail materialism. Cognitive science may not address all of what is mental. Some philosophers (such as Chalmers (1996)) hold that although many types of mental event or state, and perhaps all purely cognitive states, are material, the qualia specifically featured in sensory states are not. Furthermore, if functionalism is true, then

even if all psychofunctional roles are in fact played by material states and events, it is still at least conceivable that they might have been played by an immaterial or ghostly substance. The first of these possibilities matters at least slightly to cognitive science itself, the second not at all.

References

Armstrong DM (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block NJ (1978) Troubles with functionalism. In: Savage W (ed.) *Perception and Cognition: Minnesota Studies in the Philosophy of Science*, vol. IX, pp. 261–326. Minneapolis, MN: University of Minnesota Press.

Campbell KKC (1967) Materialism. In: Edwards P (ed.) *Encyclopedia of Philosophy*, pp. 179–188. London: Macmillan.

Chalmers D (1996) *The Conscious Mind*. Oxford, UK: Oxford University Press.

Churchland PM (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**: 67–90.

Davidson D (1970) Mental events. In: Foster L and Swanson JW (eds) *Experience and Theory*, pp. 79–101. Amherst, MA: University of Massachusetts Press.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Levine J (1993) On leaving out what it's like. In: Davies M and Humphreys G (eds) *Consciousness*, pp. 121–136. Oxford, UK: Blackwell.

Montero B (1999) The body problem. *Noûs* **33**: 183–200.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **82**: 435–456.

Putnam H (1960) Minds and machines. In: Hook S (ed.) *Dimensions of Mind*, pp. 138–164. New York, NY: Collier Books.

Sellars W (1981) Foundations for a metaphysics of pure process: the Carus Lectures. *Monist* **64**: 3–90.

Further Reading

Crane T and Mellor D (1990) There is no question of physicalism. *Mind* **99**: 185–206.

Fodor JA (1968) *Psychological Explanation*. New York, NY: Random House.

Fodor JA (1987) *Psychosemantics*. Cambridge, MA: Bradford Books/MIT Press.

Guttenplan S (1994) *A Companion to the Philosophy of Mind*. Oxford, UK: Blackwell.

Kim J (1996) *Philosophy of Mind*. Boulder, CO: Westview Press.

Lewis D (1966) An argument for the identity theory. *Journal of Philosophy* **63**: 17–25.

Place UT (1956) Is consciousness a brain process? *British Journal of Psychology* **47**: 44–50

Poland J (1994) *Physicalism: The Philosophical Foundations*. Oxford: Oxford University Press.

Smart JJC (1963) *Philosophy and Scientific Realism*. London: Routledge and Kegan Paul.

Van Gulick R (1993) Understanding the phenomenal mind: are we all just armadillos? In: Davies M and Humphreys G (eds) *Consciousness*, pp. 137–154. Oxford, UK: Blackwell.

Meaning

Intermediate article

Kent Bach, San Francisco State University, San Francisco, California, USA

CONTENTS

Sentences have meanings, and speakers mean things in using them. The meaning of a sentence is determined by the meanings of its constituents, together with its syntactic structure, but what a speaker means in using it is often not determined by what it means: the speaker may mean something more or something else.

INTRODUCTION

Language is used to express thoughts and to represent aspects of the world. What thought a sentence expresses depends on what the sentence means, and how it represents the world also depends on what it means. Moreover, it is ultimately arbitrary, a matter of convention, that the words of a language mean what they do. So it might seem that what they mean is a matter of how they are used. However, they need not be used in accordance with their literal meanings. One can speak nonliterally, and convey something other than what the sentence means ('The look on his face spoke volumes'), or speak indirectly, and convey something more than what the sentence means ('I wonder if you know the time'). Linguistic communication requires knowledge of linguistic meaning, on the part of both the speaker and the speaker's audience, but it requires extralinguistic knowledge as well.

THEORIES OF MEANING

Words have meanings, and sentences have meanings. Assuming a principle of semantic compositionality, according to which the meaning of a sentence is determined by the meanings of its constituents and its syntactic structure, the aim of a semantic theory is to give an account of how meanings of complex expressions are determined by the meanings of their simplest meaningful constituents. A theory of meaning should explain what meanings are.

Linguistic Meaning

It is plausible to suppose, at least as a first approximation, that what a declarative sentence means is what is asserted by a speaker using it literally, and that what the speaker thus asserts is the belief he or she is thereby expressing. So, for example, if you utter the sentence, 'giraffes do not use periscopes', you are asserting, and expressing the belief, that giraffes do not use periscopes. If you are sincere, you actually believe what you are asserting, and that is the very thing that comprises the meaning of the sentence you uttered. Call this thing a proposition, in this case the proposition that giraffes do not use periscopes.

That analysis leaves open what sort of thing propositions are. Presumably they are abstract entities, independent both of mind and of language. Some philosophers think of them as structured, composed of the semantic values of sentence constituents; others think of them as sets of circumstances or possible worlds with respect to which a given sentence is true. Some are reluctant to speak of propositions at all and prefer to speak of the truth conditions rather than the propositional contents of sentences. To remain neutral on this issue, let us use 'proposition' to mean the truth-conditional content of a (declarative) sentence, whether this is something that has a truth condition or simply is the truth condition. A proposition is something that different people can believe or assert and that different sentences, in the same or different languages, can mean. The truth or falsity of sentences, beliefs, and assertions depends on the truth or falsity of their propositional contents. So, in particular, the truth or falsity of the sentence, 'giraffes do not use periscopes', or of the belief expressed and the assertion made in uttering it, depends on the truth or falsity of the proposition that giraffes do not use periscopes. (*See* **Categorial Grammar and Formal Semantics**; **Possible Worlds Semantics**)

It is widely thought that the meaning of a sentence is determined by the meanings of its constituents and how they are arranged syntactically (in linguistic terms, the meaning, or semantic interpretation, of a sentence is a projection of its syntax at the level of logical form). From this perspective, it is reasonable to identify the meaning of a constituent of a sentence with the contribution it makes to the propositional contents of sentences in which it occurs. Different sorts of expressions make different sorts of contributions. For example, whereas 'giraffes' and 'periscopes' mean things of certain types, 'use' means a relation between agents and instruments, and 'not' is a negation operator.

Some words clearly have meanings and yet their meanings are not clear. With vague words, like 'red', 'old', and 'rich', there is no clear boundary between what they do and do not apply to. They appear to have borderline cases, although it has been argued (most notably by Williamson, 1994) that the boundaries are definite but unknowable. Also, different words can have the same meaning. For example, in one of their senses, 'correct' and 'right' are synonymous. So are 'teacher' and 'instructor', and 'consume' and 'devour'. And, of course, a word can have more than one meaning. Typical examples of ambiguous words are the nouns 'bat', 'club', and 'stroke', the verbs 'cut', 'draw', and 'lie', and the adjectives 'hard', 'high', and 'true'. (*See* **Vagueness**)

The occurrence of an ambiguous word in a sentence, as in 'George never goes near a golf club', renders the sentence ambiguous. But sentences can also be structurally ambiguous, as with 'German history teachers are pedantic', in which 'German history teachers' may mean 'teachers of German history' or 'German teachers of history'. Less apparent is the source of the structural ambiguity in 'The chicken is ready to eat', which can mean that the chicken is ready to do some eating or that the chicken is ready to be eaten. However, if we posit the covert presence of empty categories before and after the infinitive 'to eat', a plausible explanation for the ambiguity is that different empty categories are tied to 'the chicken', as indicated by the different indexing in 'the chicken$_1$ is ready e_1 to eat e_2' and in 'the chicken$_1$ is ready e_2 to eat e_1' (Chomsky, 1986).

Finally, the references of many words that are univocal in meaning can vary from use to use. These include personal pronouns like 'I', 'you', and 'she', temporal terms like 'today', 'tonight', and 'last week' (also tense indicators), demonstratives like 'this', 'those', 'here', and 'there', and even relational nouns like 'enemy', 'neighbor', and 'disciple'. (*See* **Indexicals and Demonstratives**)

Semantics

There are two fundamentally different, though related, conceptions of semantics in the literature. One conception takes semantics to be concerned with the linguistic meanings of expressions (words, phrases, sentences). On this conception, sentence semantics is a component of grammar. It assigns meanings to sentences as a function of the meanings of their semantically simple constituents, as supplied by the constituents' lexical semantics, and their constituent structure, as provided by their syntax. The other conception takes semantics to be concerned with the truth-conditional contents of sentences (or, alternatively, of utterances of sentences) and with the contributions expressions make to the truth-conditional contents of sentences in which they occur. The intuitive idea underlying this conception is that the meaning of a sentence, the information it carries, imposes a condition on what the world must be like in order for the sentence to be true. (*See* **Lexical Semantics**; **Syntax**)

The linguistic and the truth-conditional conceptions of semantics would be equivalent if, in general, the linguistic meanings of sentences determined their truth conditions, and they all had truth conditions. Many sentences, however, are imperative or interrogative rather than declarative. They do not have truth conditions, but compliance or answerhood conditions instead. Even if only declarative sentences are considered, in a great many cases the linguistic meaning of a sentence does not uniquely determine a truth condition. One reason for this is ambiguity, either lexical or structural: the sentence may contain one or more ambiguous words, or it may be structurally ambiguous. Another reason is that the sentence may contain indexical elements. Ambiguity makes it necessary to relativize the truth condition of a declarative sentence to one of its senses, and indexicality requires relativization to a context. Moreover, some sentences, such as 'Jack was ready' and 'Jill had enough', though syntactically well formed, are semantically incomplete. That is, the meaning of such a sentence does not fully determine a truth condition, even after ambiguities are resolved and references are fixed (Bach, 1994; Sperber and Wilson, 1995). Syntactic completeness does not guarantee semantic completeness.

The apparent conflict between these two conceptions of semantics can be resolved if it is

supposed that they have different subject matters: linguistic semantics targets sentences, whereas truth-conditional semantics targets utterances of sentences. But it should not be supposed that an utterance encodes anything that is not encoded by the sentence itself. Information available in a context of utterance may be tied to constituents of the sentence as uttered, but it is not encoded. Such information is combined with the encoded information to produce an interpretation of the utterance. Also, the speaker may convey additional information when speaking indirectly or nonliterally. So we need to distinguish between information encoded by a sentence, information tied to its utterance, and information conveyed in uttering it. The first is the province of linguistic semantics, the second of truth-conditional semantics, and the last of pragmatics.

What Are Meanings?

A traditional theory of meanings, going back to Plato, is the so-called 'idea' theory. It regards meanings as concepts and complexes of concepts. The meaning of a word (or a morpheme) is the concept (or concepts, if it is ambiguous) conventionally associated with it (a full account of what this association involves would be partly philosophical, partly psychological, and partly sociological). The meaning of a phrase is a complex concept made up of the simpler concepts associated with the words it contains, and the meaning of a sentence is the thought made up of the concepts expressed by its constituents, in accordance with how they are arranged syntactically.

An alternative theory is the so-called 'thing' theory: meanings are the references of expressions, as opposed to the concepts they express. 'Things' here include not only particular objects and events but also properties, relations, and other abstract objects. The proposition expressed by a sentence consists of a structured entity made up of the referents of its constituent expressions. In the simplest sort of case, the sentence 'Koko is hungry' expresses the proposition that Koko is hungry, whose constituents are Koko and the property of being hungry.

PHILOSOPHICAL ISSUES OF MEANING

The apparent conflict between the idea theory and the thing theory can be resolved by viewing them as answering two different questions: what confers meanings on words, and what comprises the meanings of words? This resolution distinguishes the cognitive contents of words, the ideas or concepts they express, from their semantic contents, what things (or properties or relations) they stand for. It respects the fact that words are used to talk about things, not ideas, and yet are used to express ideas.

Sense and Reference

The most influential implementation of such a two-tiered approach has been Frege's (1892). He distinguished the sense of an expression from its reference. The sense is the expression's 'cognitive value', but it also determines the expression's reference, if any. The sense of an expression imposes a condition that the reference must satisfy and does not depend on whether or not there is a reference. As Frege explains, 'the thought remains the same whether "Odysseus" has reference or not'. The sense of an expression is its contribution to the thought expressed by a sentence in which it occurs. But words are used to refer to objects of thoughts, not ideas of those objects. Even so, since the same object can be thought of in different ways, under different 'modes of presentation', how the object is thought of, and hence the sense of the expression used to refer to it, enters into the thought expressed. This, Frege suggests, explains how it is possible to think or say that Elton John is a singer and yet sincerely deny that Reginald Dwight is a singer, even though Elton John is Reginald Dwight. (*See* **Sense and Reference**)

By supposing that expressions have both sense and reference, Frege could explain how expressions that are distinct in meaning can have the same reference and how referring expressions without reference, like 'Odysseus', can still be meaningful. Nevertheless, one might ask whether what determines the reference of a referring expression – such as a proper name, an indexical, or a demonstrative – is part of the semantic content of sentences in which the expression occurs. According to the doctrine of direct reference (Kaplan, 1989), such expressions refer directly to particular individuals. This is not to deny that there are conditions that something must meet to be the referent, but only that such conditions are part of semantic content. So, for example, if someone says 'I love you', he is using 'I' to refer to himself and 'you' to refer to his listener. However, the fact that he is the speaker and the fact that the other party is the listener do not enter into what the speaker says. (*See* **Reference, Theories of**; **Direct Reference**)

Analyticity

It seems that some sentences are true (or false) solely by virtue of their meanings. For example, to understand the sentence 'All sofas are couches' is to know that it is true. In contrast, one can understand 'All snails are slow' without knowing it to be true (its truth depends on an empirical fact about snails). The first sentence is evidently analytic, the second synthetic. However, despite the glaring intuitive difference between the two, the analytic–synthetic distinction was vigorously attacked by Quine (1951) and others. Quine argued that this distinction cannot be explained in an illuminating way (i.e. without appeal to equally 'obscure' notions like synonymy, translation and, indeed, meaning itself), and he suggested that it is, at best, a matter of degree. Grice and Strawson (1956) appealed to the obvious intuitive difference between, for example, the two sentences mentioned above, to maintain that the analytic–synthetic distinction, whatever it is, is genuine.

Even if Grice and Strawson are right about this, it doesn't follow that analytic truths are true by definition. For it is arguable that most words do not have definitions, at least of the sort that provide singly necessary and jointly sufficient conditions for their application. Wittgenstein (1953) challenged the Platonic assumption that all the items to which an unambiguous word applies must have something in common. And Fodor *et al*. (1980) marshalled a variety of examples and considerations to show that Wittgenstein's famous example of 'game' is not a special case. This raises the question of what word meaning (and knowledge of it) consists in.

Meaning Skepticism

The notion of analyticity, or of synonymy, is no more puzzling than the notion of meaning itself. If a sentence has a certain semantic content, there is no reason why another sentence cannot have the same semantic content and thereby be synonymous with it. Quine's attack on analyticity is ultimately a form of meaning skepticism. Wittgenstein (1953) and Kripke (1982), developing Wittgenstein's ideas, have offered other reasons for such skepticism (the complex debates they have generated are reviewed in Hale and Wright, 1997, chapters 8, 14–17), arguing that meaning cannot be fixed by facts about individual speakers. They appeal to social facts to explain how meaning can be determinate, but it seems that the same skeptical reasons, if valid at all, apply to this social explanation.

Are Meanings in the Head?

Putnam (1975) used an ingenious thought experiment to argue that at least some linguistic meanings are not 'in the head' (and do not supervene on what is in the head). Imagine a distant planet, Twin Earth, where everything is just as it is here, except that there the clear, tasteless liquid that falls from the skies, fills the seas and quenches thirst is composed not of H_2O but of some other stuff XYZ. It is 1750, and people on either planet could not, if given the opportunity, tell the difference between H_2O and XYZ. There is no psychological difference between the two populations with respect to the word 'water'. Nevertheless, claims Putnam, 'water' means something different on each planet. Accordingly, an Earthian suddenly transported to Twin Earth would incorrectly (relative to what it means on Earth) apply the word 'water' to the common liquid he sees there, which is not water but 'twin water'. Similarly, a Twin Earthian transported to Earth would wrongly apply the word 'water', as used on Twin Earth, to the common liquid he sees here, which is not twin water but water (H_2O). So, Putnam concludes, the reference of 'water' depends on the underlying nature of the clear, plentiful liquid around us and not just on how we think of that liquid.

In another thought experiment (Burge, 1979), an arthritic patient called Art complains to his doctor, 'my arthritis has spread to my thigh'. Nothing in his acquisition of the term 'arthritis' has kept him from supposing that this inflammatory disease can occur in the limbs as well as the joints. Meanwhile, his Twin Earth counterpart Bart registers a similar complaint. There, however, the term 'arthritis' is used to refer to an inflammatory disease of either the joints or the limbs. Bart's exposure to the term 'arthritis' is the same as Art's, but, given how it is used on Twin Earth, he understands it correctly. Now, observes Burge, when Art says that his arthritis spread to his thigh he is speaking falsely, but when Bart says the same thing, using 'arthritis' as it is used on Twin Earth, he is speaking truly. Therefore, what they are saying about themselves is different, even though there is no subjective difference between them. What they mean by 'arthritis' is partly a social matter.

These thought experiments have met with considerable enthusiasm but also with criticism (e.g. Unger, 1984; Bach, 1987, chapter 13). Putnam does not explain why the term 'water' should be (and was in 1750) relevantly different from such terms as 'earth', 'air', and 'fire'. These terms do not denote natural kinds: there is no chemical restriction on the

sort of soil to which 'earth' can apply; 'air' is not restricted to any particular mixtures of nitrogen, oxygen, and carbon dioxide found in Earth's atmosphere; and 'fire' does not apply only to flame-producing chemical reactions that involve oxidation. As for 'water', although it applies (on Earth) to H_2O but not to XYZ, is this because of what 'water' means or because, as a consequence of what we have learned from modern chemistry, we count as water only stuff that is chemically like these samples? It was a discovery that this stuff, unlike earth, air, and fire, is a chemical natural kind. It is not evident why this should show that the term 'water' has a different sort of meaning, with its reference determined in a different way, from 'earth', 'air', and 'fire'.

The arthritis argument depends essentially on the supposition that one can have beliefs with meanings one 'incompletely understands'. It assumes, for example, that Art not only misunderstands the word 'arthritis' but operates with the concept arthritis rather than with some broader concept ('tharthritis') that he mistakenly associates with the word. So, it might be objected, Art understands the term 'arthritis' in precisely the same way as Bart does, and has the very same belief, namely that his tharthritis has spread to his thigh. Whatever evidence there is that he also believes that his arthritis has spread to his thigh is overridden by his idiosyncratic grasp of the term 'arthritis'. We are not tempted to say that he believes that he has inflammation of the joints in his thigh, and this should disincline us to suppose that he means that his arthritis has spread to his thigh. Because of how he misuses the term 'arthritis', he means what Bart means by it.

MEANING AND COMMUNICATION

The distinction between speaker's and linguistic meaning is needed to accommodate the fact that what a speaker means in uttering a sentence can, and often does, diverge from what is meant by the sentence he utters (even if it is neither vague nor ambiguous). A speaker can mean something other than what the sentence means, as in using 'nature abhors a vacuum' nonliterally, or something more, as in using 'it's getting cold in here' indirectly to ask someone to close the window. In general, an utterance can be literal, nonliteral, or indirect (Bach and Harnish, 1979, chapter 4), depending on the relationship between what the sentence means and what the speaker means in uttering it. In so far as the two diverge, which is usually, successful communication requires the listener to fill in the gap inferentially.

Speaker's Meaning and Linguistic Meaning

Grice had two further reasons for invoking the distinction between speaker's and linguistic meanings. Firstly, he thought that linguistic meaning could be reduced to speaker's meaning. This reductive view has not gained wide acceptance, partly because of the extreme complexity of its detailed formulation (e.g., 1989, chapter 6; Schiffer, 1972), and partly because it requires the controversial assumption that language is essentially a vehicle for communicating thoughts and not a medium of thought itself. Still, many philosophers would concede that mental content is more fundamental than linguistic meaning, and perhaps even that semantics reduces to the theory of mental content.

Secondly, Grice invoked the distinction between speaker's and linguistic meaning to counter certain extravagant claims, made by so-called 'ordinary language' philosophers, about various important philosophical terms, such as 'believes' and 'looks'. For example, it was suggested that believing implies not knowing, because to say, for example, 'I believe that dolphins communicate' is to imply that one does not know this, or to say 'Jenny's eyes look blue' is to imply that Jenny's eyes might not actually be blue. However, as Grice (1989, chapter 2) pointed out, what carries such implications is not what one is saying but that one is saying it, rather than the stronger 'I know that dolphins communicate' or 'Jenny's eyes are blue'. He also objected to certain ambiguity claims; for example, that 'or' has an exclusive as well as an inclusive sense, as in 'I would like an apple or an orange', by pointing out that it is the use of 'or', not the word itself, that carries the implication of exclusivity. Grice's 'modified Occam's razor' ('senses are not to be multiplied beyond necessity') helped to clarify the distinction between (linguistic) meaning and use, and has since helped linguists to appreciate the importance of separating, as far as possible, the domains of semantics and pragmatics. (*See* **Implicature**)

Communicative Intentions

Grice's concept of speaker's meaning (1989, chapters 5, 14, 15) was an ingenious refinement of the crude idea that communication is a matter of intentionally affecting another person's psychological states. He discovered that there is a distinctive, rational means by which the effect is achieved: by way of getting one's audience to recognize one's intention to achieve it. The intention includes,

as part of its content, that the audience recognize this very intention by taking into account the fact that they are intended to recognize it. A communicative intention is thus a self-referential, or reflexive, intention. Grice observes that 'this seems to involve a reflexive paradox, but it does not really do so' (1989, p. 219), but he does not explain why not.

Consider his example of deliberately frowning to communicate that one is displeased. Since frowning is a natural sign of displeasure, frowning deliberately might lead another to believe that one is displeased. The simplest scheme would be to misleadingly intend the other person to take one's deliberate frown for a spontaneous one and thereby take it as evidence that one is displeased. But what if one's frowning is obviously deliberate? Then it would lack the evidential value of natural frowning. However, one could exploit the common knowledge that frowning is a sign of displeasure and intend one's frowning to be taken as indicating displeasure, provided the other person supposes one intends it to be so taken. Part of what the other person is to take into account in order to infer that one is displeased is that one intends one's frowning as indicating that. This may seem to be circular, to involve a 'reflexive paradox', but Grice does not mean that one's audience has to already know what one's intention is in order to infer it. He means, rather, that the audience has to take into account that one's intention, whatever it is, is intended to be recognized.

In communicating, linguistically or otherwise, one is expressing some propositional attitude or other psychological state. In performing the speech act of asserting, one is expressing a belief; in requesting one is expressing a desire; in making an offer one is expressing a conditional intention; in apologizing one is expressing regret for something one did; and in thanking one is expressing gratitude for something the listener did. In general, different types of speech acts may be classified by the type of psychological state they express (Bach and Harnish, 1979, chapter 3), although expressing such a state does not imply that one is in it – one could be insincere. (*See* **Propositional Attitudes**; **Speech Acts**)

RELEVANCE TO COGNITIVE SCIENCE

Cognitive science touches often on the subject of meaning. Only a few topics will be briefly addressed here. (*See* **Language Acquisition and Language Change**; **Parsing**)

Language Understanding

Normal speakers understand tens of thousands of words of their language. Understanding a word is knowing its meaning, but what does this involve? A natural suggestion is that it involves two things: possessing a certain concept, and associating that concept with the word. Merely possessing the concept is not enough: the concept must be tied to the word somehow. However, it would be implausible to require that the speaker believe that the concept is the meaning of the word, for that would require the speaker to have concepts of both the concept and the word. A less cognitive and more associative account seems more plausible.

In any case, part of what understanding a word involves is knowing how the word can fit into a sentence: for example, whether it is a noun, a verb, or some other part of speech. Many words are compounds, like 'newspaper', 'hotbed', and 'toadstool', or are otherwise composed of meaningful parts, stems with prefixes or endings, such as 'composed', 'meaningful', 'parts', 'prefix', and 'ending'. Also, lexical knowledge, which includes information about the form and meaning of a word and syntactic constraints on its occurrence, must be combined with knowledge of how words are put together to form sentences. Finally, all of this purely linguistic knowledge must be supplemented with knowledge about how what people mean can go beyond what their words mean. Such knowledge, the subject of pragmatics, is not, strictly speaking, linguistic knowledge. (*See* **Lexicon**)

Pragmatic Processing

Although there is ample research on phonological, lexical, syntactic, and semantic processing, there has been little work on pragmatic processing; that is, on how people manage, when uttering a sentence, to make themselves understood, and how, when hearing a sentence, they manage to understand what the speaker means. Grice's (1989, chapter 2) theory of conversational implicature and the associated maxims of conversation provides a broad framework for identifying the factors that contribute to successful communication, from both the speaker's and the hearer's side, but it does not explain in detail how this process works.

One issue pertaining to pragmatic processing has begun to be addressed. Recanati (1995) has distinguished different models of the hearer's processing of an utterance in cases where what the speaker means is an enriched version of what the

sentence means – for example, 'I've had breakfast (today)', or 'Mary has (exactly) three children' – or where the speaker intends to be implying something. On one model, the hearer computes the proposition strictly and literally (the 'minimal' proposition) expressed by the utterance before arriving at a candidate for what the speaker means; on another model, this proposition is computed, but not necessarily first; on a third 'local processing' model, this proposition is not computed at all. Bezuidenhout and Cutting (2002) have tested these models experimentally. Their findings tend to favor the second model over the third and to refute the first model.

Polysemy

It is simplistic to suppose that knowing the meaning of a word consists in associating the 'right' concept with it. One obvious complication is that a great many words are ambiguous, thereby expressing more than one concept, but this shows only that the relation of concepts to words is many-to-one (actually, it is many-to-many, since many words have synonyms). A further complication is that with a great many words it is unclear how many meanings they have. The diverse uses of common words like 'go', 'get', 'keep', 'put', 'on', 'in', 'from', and 'to' might make it seem that their meanings are not fixed from use to use. Ruhl (1989) maintains that each such word does have a core meaning, but that this meaning is too impoverished to comprise what the speaker can mean in using the word. That is, in any particular use the meaning is enriched somehow. (See Ravin and Leacock (2000) for variations on this sort of view, applied to a variety of lexical items.)

Consider the words 'went', 'from', and 'to' as they occur in the following sentences:

Max went from Wilmington to Washington. (1)

The road went from Wilmington to Washington. (2)

The concert went from eight to ten. (3)

Max went from irritated to outraged. (4)

The house went from Max to his wife. (5)

On a simplistic conception of linguistic meaning, it might seem that only in the first sentence, which involves movement from one place to another, are the words 'went', 'from', and 'to' used literally, and that their uses in the other sentences are in various ways 'extended', hence nonliteral, uses. However, it seems plausible to suppose that the words are used literally in all these sentences, despite the fact that they are used differently from one sentence to the next. The existence of those various uses shows that the meanings of these polysemous terms are more abstract than the simplistic 'movement' model would suggest.

A great many seemingly univocal words turn out to have distinct but related forces when applied to things of different sorts. Consider the adjective 'sad' as it modifies these different nouns:

sad person / sad face / sad song / sad episode (6)

A sad face is not sad in the way that a sad person is. Rather, it expresses the person's sadness. A sad song is not sad in the way that a sad person is, and need not express any particular person's sadness, such as the writer's or the performer's (nor need it make the listener feel sad). And a sad episode is not itself sad – it causes sadness. Similarly, consider the variable import of the adjectives 'fast', 'generous', and 'conscious', as they occur in the following phrases:

fast car / fast track / fast race (7)

generous donor / generous gift (8)

conscious being / conscious state (9)

Obviously the most plausible import of each of these adjectives varies with the noun it modifies. Indeed, only in the first phrase in each case does the adjective express a property that directly belongs to what the noun denotes. Such adjectives have counterparts in other languages with the same different but related uses; therefore the variability of their import is due to polysemy, not ambiguity (linguistic coincidence).

Pustejovsky (1995) proposes that such polysemy involves 'co-compositionality': what varies from case to case is not a term's semantic properties but how those properties interact with those of the term it combines with. Although his ambitious theory, intended primarily as a computational model, is an improvement over what he calls 'sense enumeration lexicons', it seems to conflate pragmatic plausibility with semantic possibility. Although a speaker is not likely to use, say, 'sad song' to mean a song experiencing sadness, 'fast track' to mean a track that is moving fast, or 'generous gift' to mean a gift that gives generously, arguably these improbable interpretations of those phrases are nevertheless generated by the grammar. It does not seem to be a semantic fact (a fact about meanings of words)

that songs do not experience sadness, tracks do not move (much less move fast), and gifts are not generous in their giving.

The above examples of adjectival modification raise questions about the nature of compositionality (Partee, 1995). So do the following noun–noun compounds:

child abuse / drug abuse (10)

election nullification / jury nullification (11)

slalom skiing / snow skiing / heliocopter skiing (12)

jellyfish / goldfish / catfish (13)

clipboard / diving board / bread board / game board / headboard (14)

Such examples make it clear that compositionality is not as straightforward as it might seem, for there are different ways in which the meanings of words can combine. (*See* **Lexical Semantics**)

Concepts and Conceptions

Some cognitive scientists, such as prototype theorists and developmental psychologists, often refer to people's conceptions of various types of things as 'concepts'. This usage is misleading, since conceptions are much richer than concepts and play different roles. Conceptions play a role in how people group and categorize things, judge similarities and differences between things, and form theories of things of different types; but concepts are what people associate with words, by virtue of which words mean what they do. The distinction becomes clear if we consider that people can have different conceptions of a given type of thing and yet use the same word to refer to it. For example, they can associate the same concept with the word 'tree' but have different conceptions of trees.

References

Bach K (1987) *Thought and Reference*. Oxford, UK: Oxford University Press.

Bach K (1994) Conversational impliciture. *Mind and Language* **9**: 124–162.

Bach K and Harnish RM (1979) *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Bezuidenhout AL and Cutting JC (2002) Literal meaning, minimal propositions and pragmatic processing. *Journal of Pragmatics* **34**: 433–456.

Burge T (1979) Individualism and the mental. In: French P, Uehling T and Wettstein H (eds) *Midwest Studies in Philosophy*, vol. IV, pp. 73–121. Minneapolis, MN: University of Minnesota Press.

Chomsky N (1986) *Knowledge of Language*. New York, NY: Praeger.

Fodor JA, Garrett MF, Walker ET and Parkes C (1980) Against definition. *Cognition* **8**: 1–105.

Frege G (1892) On sense and reference. In: Harnish RM (ed.) (1994) *Basic Topics in the Philosophy of Language*, pp. 142–160. Englewood Cliffs, NJ: Prentice-Hall.

Grice HP (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grice HP and Strawson PF (1956) In defense of a dogma. *Philosophical Review* **65**: 141–158.

Hale B and Wright C (eds) (1997) *A Companion to the Philosophy of Language*. Oxford, UK: Blackwell.

Kaplan D (1989) Demonstratives. In: Almog J, Perry J and Wettstein H (eds) *Themes From Kaplan*, pp. 481–563. Oxford, UK: Oxford University Press.

Kripke S (1982) *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Partee B (1995) Lexical semantics and compositionality. In: Gleitman L and Liberman M (eds) *An Invitation to Cognitive Science*, vol. I, Language, 2nd edn, pp. 311–360. Cambridge, MA: MIT Press.

Pustejovsky J (1995) *The Generative Lexicon*. Cambridge, MA: MIT Press.

Putnam H (1975) The meaning of 'meaning'. In: Gunderson K (ed.) *Language, Mind and Knowledge*, pp. 131–193. Minneapolis, MN: University of Minnesota Press.

Quine WV (1951) Two dogmas of empiricism. *Philosophical Review* **60**: 20–43.

Ravin Y and Leacock C (2000) *Polysemy: Theoretical and Computational Approaches*. Oxford, UK: Oxford University Press.

Recanati F (1995) The alleged priority of literal interpretation. *Cognitive Science* **19**: 207–232.

Ruhl C (1989) *On Monosemy: A Study in Linguistic Semantics*. Albany, NY: SUNY Press.

Schiffer S (1972) *Meaning*. Oxford, UK: Oxford University Press.

Sperber D and Wilson D (1995) *Relevance*, 2nd edn. Oxford, UK: Blackwell.

Unger P (1984) *Philosophical Relativity*. Minneapolis, MN: University of Minnesota Press.

Williamson T (1994) *Vagueness*. London, UK: Routledge.

Wittgenstein L (1953) *Philosophical Investigations*. New York, NY: Macmillan.

Further Reading

Bach K (1999) The semantics–pragmatics distinction: what it is and why it matters. In: Turner K (ed.) *The Semantics–Pragmatics Interface from Different Points of View*, pp. 65–84. Oxford, UK: Elsevier.

Cruse DA (1986) *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Davidson D (1984) *Essays on Truth and Interpretation*. Oxford, UK: Oxford University Press.

Kasher A (ed.) (1998) *Pragmatics: Critical Concepts*. London, UK: Routledge.
Keefe R and Smith P (1996) *Vagueness: A Reader*. Cambridge, MA: MIT Press.
Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Larson R and Segal G (1995) *Knowledge of Meaning*. Cambridge, MA: MIT Press.
Levinson S (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
Lyons J (1995) *Linguistic Semantics: An Introduction*. Cambridge, UK: Cambridge University Press.
Quine WV (1960) *Word and Object*. Cambridge, MA: MIT Press.
Searle J (1979) *Expression and Meaning*. Cambridge, UK: Cambridge University Press.

Memory, Philosophical Issues about

Intermediate article

John Sutton, Macquarie University, Sydney, New South Wales, Australia

CONTENTS

Memory is a set of cognitive capacities by which humans and other animals retain information and reconstruct past experiences, usually for present purposes. Philosophical investigation into memory is in part continuous with the development of cognitive scientific theories, but includes related inquiries into metaphysics and personal identity.

INTRODUCTION

Cognitive scientists study an enormous variety of topics under the headings of memory and learning. Events affect humans, other animals, and machines in many ways, and sometimes enduring changes in the respective systems result, altering actual and possible future behavior. Philosophers have been centrally interested in human personal memory for episodes and experiences in the autobiographical past, as manifested in reminiscence, recall, and recognition. This is partly because of a focus on memory as a topic in epistemology, the theory of knowledge, which was in the twentieth century often pursued through methods of linguistic and conceptual analysis that were oddly divorced from psychological inquiry. Recently, however, a naturalistic turn in philosophy of mind has brought theories of memory firmly into productive contact with cognitive science.

This article first sketches central philosophical questions about memory, self, and time. It then focuses on a specific debate which is particularly relevant to cognitive science, about the existence and nature of memory traces, and concludes by outlining recent moves towards a more insistently interdisciplinary approach to memory.

PHILOSOPHICAL ISSUES CONCERNING MEMORY

Memory and Self

Cognitive scientific views of consciousness and self will increasingly influence traditional philosophical discussion about memory and personal identity. To what extent does memory construct and maintain the continuity of personal identity over time? In the history of Western philosophy this question was of pressing concern in religious contexts. In order for it to be truly *me* who is saved or damned at the Day of Judgment, the judged soul has to be numerically identical with the person who committed sinful or praiseworthy acts in this life. For Christian philosophers like John Locke, this meant that we had to retain personal memory in the afterlife.

In our 'materialist' age, the issue remains urgent because many moral and legal practices require a robust notion of continuous responsible agency. Some philosophers argue on conceptual grounds that memory cannot be the basis of personal identity because remembering *presupposes* a self who remembers. But others, notably philosophers who see the self as less unified, stable, and integrated than is acknowledged in traditional philosophical theories of personal identity, urge attention to real case studies of, for example, amnesia and dissociative identity, or to better cognitive theories of the selective and constructive nature of autobiographical memory (Schechtman, 1994).

Memory and Time

What role does memory play in our understanding of time? How is time represented in memory? The philosopher John Campbell (1997) has developed a new picture of the structure of time in autobiographical remembering. For Campbell, the human ability self-consciously to identify personal episodes as having happened at particular past times is bound up with our unique capacities to locate events, and ourselves, in an asymmetric temporal and causal order. Just as our spatial representational skills are not restricted to egocentric models, so our mature temporal orientation

involves an objective conception of time as linear. Drawing on empirical work on the representation of time in humans and other animals, Campbell claims that only humans are genuinely oriented with respect to particular times rather than merely to phase or rhythm. Because we can grasp the temporal relations between temporal cycles or phases, we have a conception of the connectedness of time which gives us the concept of the past. This detached or reflective sense of time is what grounds our understanding of the uniqueness of particular actions.

Campbell's audacious analysis of memory and time suggests, among other things, that there is no true episodic memory in nonhuman animals, and thus offers both empirical and conceptual challenges to his opponents. Comparative ethology and cognitive anthropology are as relevant to the evaluation of his account as is the clinical neuropsychology of amnesia. So this case can serve as a first illustration of the possibility that the immediate future of the cognitive philosophy of memory science will be bewilderingly and excitingly interdisciplinary.

Causation, Representation, and Traces

For me to have a genuine personal episodic memory, my present act of remembering must be causally connected in an appropriate way to the past experience being recollected. Even if it happens to be true that, as a child of four, I got lost in a shopping mall, most of us would deny that I truly personally remember the experience if I had forgotten it, and have only later been told about it by my parents. For this reason, philosophers and psychologists have hypothesized the existence of some kind of 'memory trace' as a continuous physical bridge across the temporal gulf, causally connecting past and present.

When I remember an episode of my personal history, I come into contact with events and experiences which are no longer present. We find it easy to engage in the peculiar sort of 'mental time travel' involved in such autobiographical memory, although we're often aware of significant limits to its reliability. Remembering is an instance of a general, flexible human capacity to think about the absent, so that mental life isn't entirely determined by the current environment and the immediate needs of the organism.

This is one intuitive route to the reliance on mental representations which lies at the foundation of cognitive science. It's natural to think that the physical trace, which is itself the causal result of past experience, somehow 'represents' that experience, or at least carries sufficient information about the past to allow the organism now to reconstruct that experience or something like it. Since we are often able to remember *without* having any such traces in our current external environment (such as photographs or words written in a diary), many philosophers and scientists have postulated memory traces in the brain. There are stronger and weaker versions of this view.

COGNITIVE SCIENCE AND THE PHILOSOPHY OF MEMORY

Localist Models of Memory

Since many historical theorists of memory, from Aristotle to Descartes and on into the twentieth century, explicitly refer to memory traces, it may seem that little progress has been made: 'it is the survival of the memory "trace" concept, some static, permanent, distinct storage form that each experience leaves in the organism, that links together most remarkably the oldest and most modern models' (Colville-Stewart, 1975, p. 402). On this version of the memory trace hypothesis, traces must be independent, 'atomic' items, laid down separately by every experience (or perhaps every part of every experience), and stored at a separate location, until called out again in the reproduction of that experience.

This 'archival' or 'localist' model of information storage has to some extent been physically realized in the design of digital computers since the work of von Neumann in the 1940s. At an abstract level it underlies influential models of human memory in cognitive psychology in which data are clearly separated from processing, or storage system from executive. But it has serious conceptual problems. Followers of the philosopher Ludwig Wittgenstein pursue his point that such static, structural traces are not required by empirical evidence: 'nothing seems more possible to me than that people some day will come to the definite opinion that there is no copy in either the physiological or nervous systems which corresponds to a particular thought, or a particular idea, or memory' (quoted in Stern, 1991, p. 208). Phenomenological and direct realist philosophers, in turn, argue that the view collapses into either incoherence or skepticism. Either, in searching through stored items for the representation of a particular past experience, we must already remember that past experience in checking which is the right representation, in which case the postulation of the trace is redundant; or, if there is

no such independent access to the past, then we are trapped in the present behind a veil of memory ideas, and may not really know the past at all (Wilcox and Katz, 1981).

These anti-representationist criticisms are behind many philosophical doubts about the cognitive science of memory, and sometimes about scientific psychology as a whole. But unfortunately such critics have rarely offered positive alternative approaches, beyond the blunt direct realist claim that memory is an immediate, noninferential awareness of past things themselves. And in fact only strongly localist accounts of fixed and static memory traces are vulnerable to these charges. Alternative, more dynamic conceptions of the memory trace may be more empirically plausible, while yet offering genuinely causal accounts of the mechanisms linking past and present.

Distributed Models of Memory

'Distributed' models of the memory trace build in a degree of plasticity. In connectionist cognitive science, for example, occurrent remembering is the temporary activation of a particular pattern or vector across the units of a neural network. This reconstruction is possible because of the conspiring influences of current input and the history of the network, where this history is sedimented in the particular connection weights between units. Memory traces then are not stored statically between experience and remembering, but are piled together or 'superposed' in the same set of weights (McClelland and Rumelhart, 1986, p. 193):

> We see the traces laid down by the processing of each input as contributing to the composite, superimposed memory representation. Each time a stimulus is processed, it gives rise to a slightly different memory trace – either because the item itself is different or because it occurs in a different context that conditions its representation … the traces are not kept separate. Each trace contributes to the composite, but the characteristics of particular experiences tend nevertheless to be preserved, at least until they are overridden by canceling characteristics of other traces. Also, the traces of one stimulus pattern can coexist with the traces of other stimuli, within the same composite memory trace.

This framework postulates two abstract features: distinct transient patterns of activity, and composite, enduring, but modifiable dispositional states. It is not tied to current computational models, for these two features can be implemented in different physical systems. But how do such distributed models of memory escape the philosophical criticisms of local traces? Firstly, they offer an account of causal continuity which doesn't rely on the permanent storage of independent items, and which thus seems compatible with Wittgenstein's query 'whether the things stored up may not constantly change their nature' (quoted in Stern, 1991, p. 204). Secondly, they suggest a fallibilist realist response to the skeptical worry: while past experience exerts a robust causal influence on the holistic system of traces, veridical recall is not (and need not be) guaranteed. Finally, they can be integrated into a broader consensus in cognitive psychology about the reconstructive nature of memory, to which we now turn.

Constructive Remembering

'A variety of conditions exist', notes the psychologist Daniel Schacter, 'in which subjectively compelling memories are grossly inaccurate' (1995, p. 22). Partly in response to the crisis of the early 1990s over recovered memories and false memories, cognitive psychologists have recently developed a striking consensus about the extent and importance of memory distortions and confusions, assuming that understanding mechanisms of distortion will also elucidate the reliability of memory. Research in a number of areas has focused on the context of retrieval, describing the shaping influences of the remembering situation: many memories are created at the moment they are needed, not simply extracted whole from storage. Related work on source monitoring, interference, and suggestibility has developed fairly independently of connectionist computational modeling. This makes the current state of the sciences of memory a key case study in philosophy of cognitive science.

Because memory is studied in many different disciplines, from neurobiology to narrative theory, there is no obvious unity to either the objects of enquiry or the methods employed. In addition to analyzing evidence for particular psychological classifications of the variety of memory systems, seeking for example the best definition of the notion of episodic memory, a further central task for philosophy is the careful evaluation of relations between different levels of explanation. Memory is thus a test case for the possibility and the pitfalls of genuine interdisciplinarity in cognitive science. The cognitive sciences of memory are as yet immature, yet they harness the vast institutional apparatus of normal science.

Those who value theoretical coevolution across disciplines often optimistically underestimate the difficulty of translating terms and postulated mechanisms; while more skeptical voices doubt

the likelihood of mutually constraining explanations across neighboring levels or disciplines. Yet there are models of interdisciplinarity in the philosophy of science which allow for contact or even local reductive identification between theories at different levels, without unrealistic reliance on the formal unity of science. Valerie Hardcastle, for example, argues (1996, pp. 105–139) that consensus on the existence of dual implicit and explicit memory systems exemplifies a developing 'complicated and cluttered' interdisciplinary theory which relies actively on the methods and underlying assumptions of a number of different research traditions, in this case including developmental psychology, clinical neuropsychology, and experimental cognitive psychology.

The case of autobiographical memory development, to take another example, suggests that this multidisciplinary reach must extend beyond the level of individual psychology to incorporate social and historical factors. Although children talk about the past almost as soon as they start talking, they may first learn some personal narrative forms to organize their explicit, *public* recounting of past experiences, and then use these principles as scaffolding around which to organize their *internal* representations of past experiences. So the development of autobiographical conventions and patterns of coherence is less an automatic unfolding of internal processes than an internalization of cultural schemes (Nelson and Fivush, 2000). The best explanations of the form and content of specific personal memories may often refer not simply to the past episode itself, but (as in the case of failures of source memory) to multiple causes which span internal and external factors. The converse point also applies. This need to bridge the 'personal' level of explanation, the traditional domain of philosophy of mind, and to focus simultaneously on the subpersonal and the social, has implications for the philosophy of the social sciences of memory, in which explanation may need to refer to flexible internal processes of schematization and reconstruction.

Both cognitive anthropologists and philosophers drawing on dynamical and situated approaches to cognition have suggested the need for a general framework for memory science which can make sense of traces both inside and outside the individual. This is not to collapse the distinction between external and internal representational formats: for a connectionist in particular, the kind of 'storage' mechanisms employed by the brain are quite different from those of most external linguistic or digital systems. The point rather is to see brain traces and external traces as parts of temporarily integrated larger systems, used by us so as more successfully to exploit and manipulate information in the environment. As Andy Clark puts it, 'our brains make the world smart so that we can be dumb in peace' (1997, p. 180). This perspective can be given a biological twist, as in Merlin Donald's provocative account (1991) of the crucial changes in cognitive architecture which occurred with the historical development of enduring and transportable external memory hardware or 'exograms'; a psychological twist, as in David Rubin's (1995) ambitious theory of memory for oral traditions; or a metaphysical twist, as in perceptual psychologist Michael Leyton's (1992) general theory of memory as the asymmetric traces left by processes on objects. In each case, the understanding of changing cultural and technological systems becomes an integral part of cognitive science, rather than a humanistic curiosity.

References

Campbell J (1997) The structure of time in autobiographical memory. *European Journal of Philosophy* **5**: 105–118.

Clark A (1997) *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Colville-Stewart SB (1975) *Physico-Chemical Models of the Memory Storage Process: The Historical Role of Argument from Analogy*. PhD thesis, University of London.

Donald M (1991) *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Hardcastle V (1996) *How to Build a Theory in Cognitive Science*. Albany, NY: State University of New York Press.

Leyton M (1992) *Symmetry, Causality, Mind*. Cambridge, MA: MIT Press.

McClelland JL and Rumelhart DE (1986) A distributed model of human learning and memory. In: McClelland JL and Rumelhart DE (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, pp. 170–215. Cambridge, MA: MIT Press.

Nelson K and Fivush R (2000) Socialization of memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 283–295. Oxford, UK: Oxford University Press.

Rubin DC (1995) *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. Oxford, UK: Oxford University Press.

Schacter DL (1995) Memory distortion: history and current status. In: Schacter DL (ed.) *Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past*, pp. 1–43. Cambridge, MA: Harvard University Press.

Schechtman M (1994) The truth about memory. *Philosophical Psychology* **7**: 3–18.

Stern DG (1991) Models of memory: Wittgenstein and cognitive science. *Philosophical Psychology* **4**: 203–218.
Wilcox S and Katz S (1981) A direct realist alternative to the traditional conception of memory. *Behaviorism* **9**: 227–239.

Further Reading

Campbell J (1994) *Past, Space, and Self*. Cambridge, MA: MIT Press.
Casey ES (1987) *Remembering: A Phenomenological Study*. Bloomington, IN: Indiana University Press.
Connerton P (1989) *How Societies Remember*. Cambridge, UK: Cambridge University Press.
Deutscher M (1989) *Remembering 'Remembering'*. In: Heil J (ed.) *Cause, Mind, and Reality*, pp. 53–72. Dordrecht, Netherlands: Kluwer.
Draaisma D (2000) *Metaphors of Memory*. Cambridge, UK: Cambridge University Press.
Engel S (1999) *Context is Everything: The Nature of Memory*. New York, NY: WH Freeman.
Hacking I (1995) *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton, NJ: Princeton University Press.
Hamilton A (1999) False memory syndrome and the authority of personal memory-claims: a philosophical perspective. *Philosophy, Psychiatry, & Psychology* **5**: 283–297.
Hoerl C (1999) Memory, amnesia, and the past. *Mind and Language* **14**: 227–251.
Hoerl C and McCormack T (eds) (2001) *Time and Memory: Philosophical and Psychological Perspectives*. Oxford, UK: Oxford University Press.
Krell DF (1990) *Of Memory, Reminiscence, and Writing: On the Verge*. Bloomington, IN: Indiana University Press.
Malcolm N (1977) *Memory and Mind*. Ithaca, NY: Cornell University Press.
Rowlands M (1999) *The Body in Mind: Understanding Cognitive Processes*. Cambridge, UK: Cambridge University Press.
Sorabji R (1972) *Aristotle on Memory*. London, UK: Duckworth.
Sutton J (1998) *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge, UK: Cambridge University Press.
Sutton J (2002) Representation, reduction, and interdisciplinarity in the sciences of memory. In: Clapin H, Staines P and Slezak P (eds) *Representation in Mind*. Westport, CT: Greenwood Publishers.
Warnock M (1987) *Memory*. London, UK: Faber.

Mental Causation

Intermediate article

Tim Crane, University College London, London, UK

CONTENTS

It is arguably an assumption of both common sense and scientific psychology that mental states and events cause events in the physical world. Yet this fact is problematic for both physicalist and dualist theories of the mind.

INTRODUCTION

Does the mind have effects in the physical world? To believe it does is to believe in mental causation. It can be argued that we are committed to the existence of mental causation when we explain people's actions in terms of their thoughts, beliefs, intentions, desires, and other propositional attitudes. For example, we might say that Jenny drank the whisky because she thought it would calm her nerves. To say that there is mental causation in this case is to say that the 'because' expresses a causal relation between Jenny's thought and her action, just as it does in non-mental cases, as when we say that the bridge collapsed because the bomb exploded beneath it. In other words, the thought, like the explosion of the bomb, is causally efficacious. Understood in this way, mental causation is ubiquitous. Whenever we do something or think something because of something going on in our minds, this is a case of mental causation. But what is the nature of this causation, and why have philosophers found it so problematic?

WHAT IS MENTAL CAUSATION?

Mental causation is when a mental state (like a belief or intention) or a mental event (like an experience) has an effect, either a mental effect (like another thought or experience) or a physical effect, an effect in the physical world. In any case of causation, we can distinguish between the *relata* of causation (what is being related) and the *relation* itself. So, for example, when the explosion caused the bridge to collapse, we can distinguish between the cause (the explosion), the effect (the collapse) and the relation itself (causation) which links these two events. To say that there is mental causation is to say that a cause of some effect is mental; just as to say that there is physical causation is to say that a cause of some effect is physical. It is not necessarily to say that there is a distinctive kind of relation – a distinctive kind of causation – which holds in the cases where a mental entity is a cause. This is a possible position; but it is not required by the idea of mental causation. Hence we should not commit ourselves at the outset to a conception of causation (e.g. that it must involve contact action) that renders mental causation impossible to understand.

What, then, is it that mental and physical causation have in common? What is it that makes them both cases of causation? The answer to this question depends on the correct theory of causation, and it is important to emphasize that few theories of causation entail that causation must be a physical relation. Some theories say that A causes B when there is a law of nature linking A-type and B-type events; others that A causes B when B is counterfactually dependent on A (i.e. if A had not existed, B would not have existed); and others say that A causes B when the probability of B is higher in the presence of A than it would have been otherwise (for all these options, see Sosa and Tooley, 1991). Other theories deny that causation is a relation at all (Mellor, 1995). But however they differ, these analyses can apply equally well to mental as to physical causes and effects.

As well as discussing the nature of the causal relation, theories of causation also discuss what kinds of entity are the relata of causation; that is, what kinds of entity are causes and effects. Some theories say that causes and effects are always events (like the explosion of the bomb or Jenny's drinking the whisky), while others say that they are facts (like the fact that the bomb exploded or the fact that Jenny drank the whisky) or states (the state of Jenny's having drunk the whisky). Others express this distinction as one between events and properties of events (events have many properties,

but only some properties of events are causally efficacious). Theories of 'agent causation', by contrast, claim that the fundamental phenomenon of mental causation is when agents, rather than their states or events involving them, cause things to happen: as when John breaks the window by smashing it. In this article, we will consider only causation by events, or states and properties (where a state is understood as a thing's having a property at a time).

Mental causation is an essential part of some metaphysical theories of mind. So it is with *functionalism*, whose characteristic thesis is that mental states are individuated (distinguished from one another) by the causal roles they play (Block, 1980). A functionalist holds that belief, for example, is the sort of state that is typically caused by perceptions and other beliefs, and is disposed to cause actions in conjunction with desires. Functionalism therefore assumes that mental states are causes and effects – the mind is a causal mechanism – though there are various accounts of what this actually means.

HISTORY

Debates about the causal powers of the mind can be traced back to antiquity; but in their contemporary form they derive from Descartes' influential theory of mind and body. Descartes was a dualist: he thought that mind and body were distinct substances. For Descartes, a substance is a being which is capable of independent existence, one whose existence depends on nothing else. So to say that mind (or soul) and body are distinct substances is to say, among other things, that they are capable of independent existence.

Descartes was criticised in his lifetime for making mental causation hard to understand, most famously by Princess Elisabeth of Bohemia (Descartes, 1985). Princess Elisabeth asked how substances so different as minds and bodies could affect one another; Descartes claimed not to see the difficulty, and the debate between them was left unresolved.

A more effective criticism of Descartes' dualism came from Leibniz. A central thesis of Descartes' physics is that matter is a substance whose characteristic (essential) attribute is extension in space. God has endowed matter with a certain *quantity of motion*, and the total quantity of motion is preserved in all physical interactions: an interaction never diminishes or adds to the total quantity of motion in the world. Thus Descartes believed in the conservation of quantity of motion, but also believed that mental causation was consistent with this law of nature. His reasoning, according to Leibniz, was that the mind causes things to happen in the body by changing the *direction* of motion of the animal spirits (a rarefied form of matter) at the pineal gland in the brain. So the mind can change the direction of motion of matter and not alter the total quantity of motion: mental causation is consistent with the conservation laws, as Descartes understood them.

Leibniz did not challenge the validity of this reasoning, but the correctness of Descartes' conservation laws. According to Leibniz, what is conserved in the physical world of matter is not quantity of motion but quantity of momentum, the product of mass and velocity. Since velocity is a vector of speed and direction, the mind cannot alter the direction of motion of the animal spirits without altering the quantity of momentum in the physical world. Therefore mental causation is inconsistent with the correct conservation law: the conservation of momentum (see Woolhouse, 1993).

Leibniz' alternative to Descartes' dualism was his doctrine of pre-established harmony, sometimes called *parallelism*. This is the view that mind and body do not interact causally, but operate in parallel (in harmony) in accordance with the will of God who initiated (pre-established) the harmony. This doctrine is a form of *epiphenomenalism*: the view that the mind has no effects in the physical world. Another form of epiphenomenalism is the *occasionalism* of Malebranche, which holds that the mind cannot act in the physical world on its own, but needs the help of God's action on each occasion of interaction. Each movement of the body by the mind is, in effect, a miracle. Epiphenomenalism need not deny that there are causal relations between different mental phenomena. But it must deny that there are any causal relations between mental phenomena and matter.

The naturalistic philosophy of the nineteenth and twentieth centuries did not generally see mental causation as a problem. Many naturalists are materialists, and materialists identify the mind with something material, the brain. By identifying the mind with the brain, materialism can allow mental phenomena to cause material phenomena, because mental phenomena are just a species of material phenomenon. In the twentieth century, the term 'physicalism' was sometimes used as a synonym for materialism, while sometimes the term was meant to indicate the special ontological and epistemological authority which physical science has

in telling us about the material world. The supposed difference between materialism and physicalism could then be put like this: materialism holds that everything is matter, whereas physicalism holds that everything is physical, where the physical is the subject matter of physical science. Therefore, if physics talks about various things that are not matter, physicalism can recognise the existence of something that materialism cannot. Since arguably the fields and forces of contemporary physics are not matter in any usual sense, physicalism seems to be the preferable theory. In what follows, therefore, this article will talk of physicalism rather than materialism.

MENTAL CAUSATION AS A PROBLEM FOR DUALISM

The problem of mental causation which originated in the seventeenth century re-emerged in the twentieth century as part of the argument for a specific form of physicalism, the identity theory (Feigl, 1958). Defenders of the identity theory argued that there were no philosophical, *a priori* objections to identifying mental phenomena with states of the brain; the truth of this claim must be established empirically. Identity is here understood literally: the claim is that a mental state is the very same thing as a state of the brain. (The identification of 'pain' with the firing of c-fibres became a common, though empirically false, illustration of the claim.) Later theories went further, and argued that the identity theory could be demonstrated by philosophical argument, rather than simply shown to be coherent (Lewis, 1966; Armstrong, 1968; Davidson, 1970). The general form of their argument is as follows:

1. First premise: mental causes have physical effects.
2. Second premise: the physical world is causally closed; that is, all physical effects have physical causes which are sufficient to bring them about ('the completeness of physics'; see Papineau, 2000).
3. Conclusion: mental causes are physical causes.

Different proponents of the argument elaborate and defend it in different ways, to make it strictly valid. For example, some say that an extra premise denying the existence of mental–physical *causal overdetermination* must be added. (Overdetermination is when an effect has two (or more) causes, each of which is enough to bring the effect about, and each of which would have brought it about if the other (or others) hadn't.) Others say that the second premise must be reformulated to make it compatible with indeterminism, since as it stands it is a deterministic claim. And others say that the first premise must be definitive of the nature of mental states, not just a fact about them (Lewis, 1966). But here we can put to one side these clarifications of detail, and focus on the general form of the argument.

The general form of the argument is that in order to reconcile mental causation with the completeness of physics, we have to identify mental and physical causes. So if all mental phenomena have some physical effects – a widely-held assumption, but an assumption none the less – then all mental phenomena are physical phenomena. The reasoning behind this argument is simple: if there are mental causes of physical effects, then how is this compatible with these effects having adequate physical causes, as the completeness of physics says they must? Either, it seems, the completeness of physics is false or epiphenomenalism is true. In other words, if the completeness of physics is accepted, then mental causation is a deep problem for mind–body dualism. The problem is only resolved, it seems, by identifying the mental and the physical causes.

Can a dualist respond to this problem? Is physicalism the only adequate response? Perhaps the dualist can deny the premises. The first premise of the argument is the existence of mental causation. As we have seen, a dualist could deny this premise by being an epiphenomenalist. But epiphenomenalism is very hard to believe: the view that our minds make our bodies move does not seem to be a theoretical claim, but a datum that theory should account for. Can a dualist deny the completeness of physics? Here matters are more complicated. The completeness of physics is not normally understood as a law of physics (like Newton's laws or the Schrödinger equation) but as a metaphysical speculation based on the laws of physics. A dualist could deny that this speculation is a consequence of the laws of physics. This is widely thought to be contrary to received opinion among philosophers of science; but the issue is still controversial (see Papineau, 2000, and Cartwright, 2000, for opposing perspectives).

The physicalist conclusion is that mental causes are identical with physical causes. So long as all mental phenomena have some physical effect at some point, then physicalists can conclude that each mental phenomenon is identical with some physical phenomenon. This is an *identity theory* of mind and brain. There are two types of identity theory: the 'type identity theory', which identifies mental properties or types, and the 'token identity theory', which identifies mental tokens or particulars.

Which identity theory one accepts might depend on one's views of the relata of causation (see above): if one held that properties or states are causes, for instance, then one would accept the type identity theory (Lewis, 1966), but if one held that events were causes, then one would accept the token identity theory (Davidson, 1970).

MENTAL CAUSATION AS A PROBLEM FOR PHYSICALISM

Since one of the general motivations for a physicalist theory of mind derives from the causal role of the mind, it is surprising to discover that mental causation creates problems for physicalism as well as for dualism. But there is a form of physicalism (called 'nonreductive physicalism') which denies the identity theory. Since the identity theory was what enabled physicalists to solve the problem of mental causation, those physicalists who reject the identity theory encounter that problem in a new form.

Some physicalists deny the identity theory because it entails the thesis that all creatures who are in the same mental state must be in the same physical state too, and this thesis is empirically implausible, given the diversity of organisms. Consider, for example, the variety of creatures who are capable of being in pain, and the variety of their physical constitutions, and then consider how unlikely it is that all these creatures share a physical state or property when they are in the same mental state (Putnam, 1975b). Nonreductive physicalists say that we should not identify mental properties or states with physical properties or states. But they endorse a weaker form of physicalism, to the effect that all particular objects and events are physical, even if not all properties and states are physical. (This is the so-called 'token identity theory'.) The resulting view is called nonreductive because it does not 'reduce' mental states to physical states, as the type identity theory does, by identifying them; but it is still physicalism because it gives an ontological priority to the physical in saying that all particular objects and events are physical. There are no nonphysical objects or events.

How does this affect the question of mental causation? This depends on how nonreductive physicalism regards the relata of causation. If causation is a relation between events, then nonreductive physicalism has no difficulty accounting for mental causation in physicalist terms, since all events are physical, even if not all properties are (Davidson, 1993). But some philosophers argue, for reasons independent of the philosophy of mind, that properties or states are causes, not events. One reason for believing this is from reflection on commonsense examples: if throwing a brick broke a window, then it is not the event of throwing the brick as such that had this effect, but rather the throwing of a brick with certain properties (its weight, its velocity, etc.). If the brick had been made of rubber, or had been thrown with less force, it might not have broken the window. Therefore, it is concluded that strictly speaking, causes are properties or states (i.e. things having properties); or, to put it another way, causes have their effects by virtue of their properties. But if causes are properties or states, then nonreductive physicalists must deny the identity theory of mental and physical causes, and therefore they cannot employ the argument discussed above. If they are not epiphenomenalist, then they must accept the first and second premises and reject the conclusion.

To put it another way: suppose there is mental causation, and the completeness of physics is true. And suppose properties (or states) are causes, and that the identity theory is false. Then it is hard to see how there can be mental causation in the light of the completeness of physics, even if every mental event is a physical event. This is the problem of mental causation for nonreductive physicalists (see Heil and Mele, 1993, for a variety of statements of this problem, and responses to it).

Nonreductive physicalists have tended to respond in one of two ways to this problem: either by developing the notion of causation involved in the debate, or by developing the doctrine of physicalism. Those who wish to develop the notion of causation might say, for example, that mental causes are causally relevant to physical effects, although not causally efficacious. (For similar ideas, see Dretske, 1988, and Jackson and Pettit, 1988.) One difficulty with these approaches is that it is hard to see them as more than ad hoc responses to the problem in hand: it can seem as if a specific notion of mental causation is simply being tailored to solve the problem. Some more ambitious approaches have therefore motivated their solution with detailed independent accounts of causation itself (Yablo, 1992, is a particularly detailed attempt).

The other kind of approach takes causation for granted, but further develops, the idea of nonreductive physicalism (Loewer, 2001). This approach assumes Jackson's definition of physicalism (Jackson, 1998), employing possible worlds: any minimal physical duplicate of our world is a duplicate *simpliciter*. It also assumes that causation is counterfactual dependence between facts or states

of affairs. Jackson's definition yields the metaphysically necessary determination of the mental by the physical: given what the physical facts actually are, the mental facts could not have been otherwise (see also (Lewis, 1993)). It follows that if the mental facts had been different in some way, then the physical facts would have been different, even if the mental and the physical facts are not identical. So, in particular, a mental cause M of a physical effect E causes E even though the completeness of physics guarantees the existence of a physical cause P which is enough for E – because P necessarily determines M, as well as causally sufficing for E. If M had not been the case, then E would not have been the case, since if M had not been the case, P would not have been the case and therefore (arguably) E would not have been the case either. By appealing to this (admittedly problematic) idea of metaphysically necessary determination, physicalists attempt to solve the problem of mental causation without appealing to the identity theory.

MENTAL CAUSATION AND COGNITIVE SCIENCE

In so far as cognitive science is committed to a form of nonreductive physicalism, denies epiphenomenalism, and upholds the completeness of physics, it has to give an account of mental causation. One of the most influential theories of the foundations of cognitive science, Jerry Fodor's 'representational theory of the mind' (RTM), presupposes that mental states involve causally related sequences of mental representations, or symbols in a language of thought. The main argument for RTM is based on the idea that the logical and rational relations between thoughts must have an underlying causal mechanism (Fodor, 1987). The causal mechanism of such thought processes, it is argued, must involve mental representations with a structure that mirrors the logical structure of thoughts; the representations have a semantic and a syntactic (i.e. causal) structure.

Critics have questioned whether RTM renders the content of thought causally idle: since the causal role of mental representations is discharged by the syntactic structure of the representations, what causal role does this leave for the content of thought? And if the content of thought is epiphenomenal, does this make it theoretically dispensable? Defenders of RTM have responded by claiming that the causal efficacy of content is guaranteed by the fact that it *supervenes* on the syntactic structure of the brain, that is, that there is no difference in content without a difference in syntax. But if syntactic structure is an aspect of the local physical structure of the brain, this defence puts RTM in conflict with the widely accepted doctrine of externalism, since according to externalism, the content of our thoughts does not supervene on the local physical structure of our brains (Putnam, 1975a). Fodor (1995) attempts to resolve this apparent contradiction.

References

Armstrong DM (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block N (1980) What is functionalism? In: Block N (ed.) *Readings in the Philosophy of Psychology*. London: Methuen.

Cartwright N (2000) The completability of science. In: Stone MWF and Wolff J (eds) *The Proper Ambition of Science*. London: Routledge.

Davidson D (1970) Mental events. In: Foster L and Swanson J (eds) *Experience and Theory*. London: Duckworth. [Reprinted in: Davidson D (1980) *Essays on Actions and Events*. Oxford: Oxford University Press.]

Davidson D (1993) Thinking causes. In: Heil J and Mele A (eds) *Mental Causation*. Oxford: Oxford University Press.

Descartes R (1985) *The Philosophical Writings of Descartes*, 3 vols., translated by Cottingham J, Stoothof R and Murdoch D. Cambridge, UK: Cambridge University Press.

Dretske F (1988) *Explaining Behavior*. Cambridge, MA: MIT Press.

Feigl H (1958) The 'mental' and the 'physical'. In: Feigl H, Scriven M and Maxwell G (eds) *Minnesota Studies in the Philosophy of Science*. Minneapolis, MN: University of Minnesota Press. [Reprinted as a monograph by the same publisher, 1967.]

Fodor J (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor J (1995) *The Elm and the Expert*. Cambridge, MA: MIT Press.

Heil J and Mele A (eds) (1993) *Mental Causation*. Oxford: Oxford University Press.

Jackson F (1998) *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson F and Pettit P (1988) Functionalism and broad content. *Mind* **97**: 381–400.

Lewis D (1966) An argument for the identity theory. *Journal of Philosophy* **63**: 17–25.

Lewis D (1993) Reduction of mind. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*, pp 412–431. Oxford: Blackwell.

Loewer B (2001) From physics to physicalism. In: Gillett C and Loewer B (eds) *Physicalism and its Discontents*. Cambridge, UK and New York, NY: Cambridge University Press.

Mellor DH (1995) *The Facts of Causation*. London: Routledge.
Papineau D (2000) The rise of physicalism. In: Stone MWF and Wolff J (eds) *The Proper Ambition of Science*. London: Routledge.
Putnam H (1975a) The meaning of 'meaning'. In: *Mind, Language and Reality*. Cambridge, UK: Cambridge University Press.
Putnam H (1975b) The nature of mental states. In: *Mind, Language and Reality*. Cambridge, UK: Cambridge University Press.
Sosa E and Tooley M (eds) (1991) Causation. Oxford: Oxford University Press.
Woolhouse Roger (1993) *Descartes, Spinoza, Leibniz: The Concept of Substance in Seventeenth Century Metaphysics*. London: Routledge.
Yablo S (1992) Mental causation. *Philosophical Review* **101**: 245–280.

Further Reading

Crane T (1995) The mental causation debate. *Proceedings of the Aristotelian Society* Supplementary Volume **69**: 211–236.
Horgan T (1993) From supervenience to superdupervenience: meeting the demands of a material world. *Mind* **102**: 555–586.
Jackson F (1996) Mental causation. *Mind* **105**: 377–413.
Kim J (1993) *Supervenience and Mind*. Cambridge, UK: Cambridge University Press.
Kim J (1998) *Mind in a Natural World*. Cambridge, MA: MIT Press.
McLaughlin B (1992) The rise and fall of British emergentism. In: Beckermann A *et al.* (eds) *Emergence or Reduction?* Berlin: De Gruyter.
Pietroski P (2000) *Causing Actions*. Oxford: Oxford University Press.

Mental Content, Causal Theories of

Intermediate article

Charles Wallis, California State University, Long Beach, California, USA

CONTENTS

Theories of mental content seek to explain how mental states are about the world. Causal theories of mind propose to understand mental content in terms of the causal relationships between brain states and the world.

WHAT ARE CAUSAL THEORIES OF MENTAL CONTENT?

Causal theories of mental content hold that mental states represent the world in virtue of the causal relationships those states have within the mind and/or with the world. Though many theories of mental content are causal theories, not all theories of mental representation are causal in nature. For example, John Locke supposes that ideas of primary qualities (e.g. shapes) represent those qualities in an object because they are similar (share the same properties). Contemporary causal theories have predecessors dating back to Book III of John Locke's *Essay Concerning Human Understanding*, possibly Aristotle's *De Anima*, or even to Plato's *Theatetus*. Locke's notion of secondary qualities (a quality or power of the object which causes particular ideas in us that bear no similarity to the object, e.g. color) looks very much like a contemporary causal theory based upon reliable causally mediated covariation (Cummins, 1989). Aristotle's discussion in Book II of *De Anima* suggests that color in an object, though different from sensations of color, is nevertheless reliably caused by light hitting the object. Plato's analogy of perception as the matching of sensations caused by the world to impressions (knowledge) upon a ball of wax also suggests that Plato entertained the notion that causal connections allowed our minds to represent.

Contemporary causal theories of mental representation have developed in the theoretical context of explanation in cognitive science. Though causal theories can be understood independently, it can be helpful to view these theories in the context of computational explanations in cognitive science.

The computational theory of cognition (CTC) and the representational theory of intentionality (RTI) form the foundation of cognitive science. CTC characterizes cognition as the computation of complex functions on representational states. Computation consists in performing operations defined over representational states. Operations can be rule-based manipulations of syntactic strings as in traditional artificial intelligence, or the weighted passing of activation as in connectionist models. RTI asserts that mental states are about the world (have content) in virtue of a representation relation holding between the world and those states. CTC/RTI portrays cognizers as receiving input through sensory organs and/or memory and generating outputs in the form of memories, inputs to other processes, and/or motor response commands (see Figure 1).

Debate over the concept of computation relevant to the CTC centers around Turing-compatible computation and its development in computer science and mathematics, and dynamically described state transitions (e.g. connectionism). There are several agreed constraints for a theory of representation. First, the representation relation must be consistent with the physicalistic nature of cognitive science. Second, the relation must be present and explanatory in accepted explanations within cognitive science. Nevertheless, a similar dichotomy of theories occurs in the literature on representation. To date, neither theory regarding the representation relation presupposed by the RTI has gained general acceptance.

Varieties of Causal Theories of Mental Content

One theory, nomic covariation, postulates a causal relationship between a mental state and the object

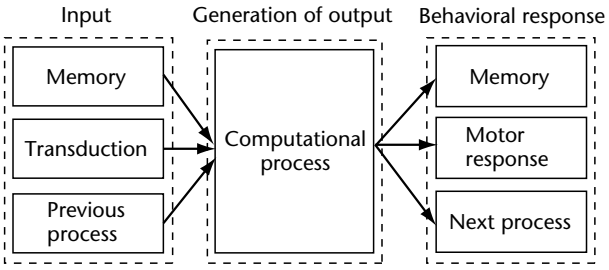


Figure 1. Elements of computational explanations of cognitive capacities in cognitive science.

or property that state represents. The other theory, functional role semantics, hypothesizes that a state has content in virtue of the state occupying a particular position in a complex web of causal relationships characterizing the cognizer's functioning. Candidate content-fixing causal relationships may include causal relationships within the cognizer and/or without. Both functional role and covariation theories satisfy the first, physicalistic constraint, since they suppose a cognizer's states represent the distal environment solely as a result of the specific types of causal connection those states have. The theories diverge in terms of the specific causal relationships each emphasizes.

Nomic Covariation

Nomic covariation theories define content in terms of a causal connection between a given object or property in the world, and the state with which the cognizer represents that object. Contemporary covariance theorists draw their inspiration from information theory as well as work within psychophysics.

> there are circumstances such that red instantiations control 'red' tokenings whenever those circumstances obtain; and it's plausible that 'red' expresses the property red in virtue of the fact that red instantiations cause 'red' tokenings in those circumstances; and the circumstances are nonsemantically, nonteleologically, and nonintentionally specifiable.
>
> In fact, they're psychophysically specifiable.
>
> (Fodor, 1988, p. 112)

Covariation theories seem to satisfy the second constraint upon theories of representation, i.e. that the relation must be present and explanatory in accepted explanations within cognitive science. For example, Hubel and Wiesel (1977) investigated the representational content of cells in the striate (visual) cortex by monitoring the activity of those cells looking for lawlike relationships between the activity of these cells and the presence of properties in the visual field. Covariationists hypothesize that a system's states represent those objects/properties of the distal world with which they covary. More precisely, covariation assigns contents to states via the following definition (Cummins, 1989; Fodor, 1988):

> State, S_b, represents B iff the system tokens S_b when, only when, and because of B.

Advantages of Nomic Covariation

Covariation gains intuitive plausibility in that a cognizer represents an object/property by being causally 'in tune' with that object/property. Covariationists like Jerry Fodor claim their theories assign content to individual states, S_b, independent of the cognizer's potential operations upon those states (inferences it can make) and/or the content of other states with which S_b might interact. Fodor refers to this assignment of content as 'punctate content' or 'atomic content'. Fodor claims that punctate content allows for content identity across individuals having quite different theories regarding the object/property. Similarly, punctate content provides the only theory whereby people can refer to objects even when they have great numbers of false beliefs about them. Fodor and Lepore (1992) note that Aristotle thought and talked about stars even though he falsely believed them to be rotating relatively close around the Earth on glassy spheres.

The systematicity and compositionality arguments suppose that covariation theories are in a better position to explain perceived truths about the combinatorial properties of language and concepts. Compositionality is the theory that the meaning of a complex expression in a language results from the meanings of its constitutive elements. Compositionality plays a central role in many linguistic theories since its supposition for both language and thought provides a fairly straightforward explanation of the human ability to grasp an enormous number of different thoughts of varying complexity and their corresponding linguistic expressions.

Covariation boasts a final perceived advantage in that it provides a clear-cut mechanism through which a system can come to know about an object/property in the world in virtue of its representational capacities. Cognizers represent an object/property insofar as they can reliably detect its presence in the environment.

Problems for Nomic Covariation

Cognitive scientists, particularly philosophers, recognize that covariation theories face a number of standard objections. First, lawful covariation between a state and an object/property cannot straightforwardly explain representation of nonexistent objects/properties. Another objection challenges the ability of covariation to explain the representational capacities of states with regard to objects/properties beyond low-level perceptual properties (e.g. edges), especially very abstract, high-level objects and properties like God or photon. Advocates of covariation suppose that covariation determines content for states representing objects/properties ranging from low-level properties (e.g. edge or blob), to ordinary objects and properties (e.g. dog or chair). Properties/objects associated with thoughts of a higher level of abstraction, like photon, get represented by states in virtue of their being defined in terms of the representational properties of lower level states (Fodor, 1988)

Robert Cummins (1989) complains that covariation of states with objects/properties requires either implicitly or explicitly represented heuristic information in the system. Thus, covariation will explain the representational properties of primitive representational states only. Higher level concepts, e.g. from table to totalitarian, acquire meaning through definition. At first glance, it seems plausible to suppose that states that represent high-level theoretical properties like photon acquire their meaning through definition in terms of lower level properties. After all, such properties and objects often have theoretical definitions. However, the failure of such semantic reductions has shaped the history of Western philosophy and mathematics (Quine, 1953). Worse still, the more ordinary the concept, the less readily one accepts that one knows its meaning because one has an explicit definition. Covariation's critics ask, 'if the state one uses to represent panda acquires meaning from being defined in terms of lower level properties, why does one have difficulty in articulating that definition in any but the most superficial sense?' Psychological research seems to weigh against such semantic reductionism as well. It indicates that object recognition and categorization are strongly influenced by perceptual features, the exact make-up of which is relatively variable.

For Cummins, mental states covary with cats because one already has knowledge about cats (e.g. they are domesticated felines) that one explicitly represents. Further, if covariation is not the simple, unmediated causal relationship between property/object and state, but the result of often complex causal interactions within the system, covariation looks like another instance of its competitor, functional role semantics.

The third, and most widely discussed, problem for covariation is called the 'disjunction problem'. According to covariationists, a state represents an object or property if the system tokens (enters into) that state when confronted with the object or property and only if confronted with that property. Suppose that a cognizer has a state, S_c, that appears to represent cathood. In fact, the cognizer tokens this state in all of its cat interactions. However, one night the cognizer sees a possum in the half light and tokens S_c. The only if clause of the covariationist definition *prima facie* dictates that S_c never really represented cat. Rather, the cognizer has always represented a property one might describe using the disjunction, 'cat or possum'.

The disjunction problem poses difficulties for covariation theories in two ways: first, the theory seems to dictate counterintuitive representational contents for states. If one has beliefs about cats (e.g. cats are domesticated felines having a number of distinct breeds) then covariation would seem to dictate that you in fact have a belief that things having the property of being either a cat *or* a weasel have those qualities. Second, and most widely discussed, the disjunction problem seems to show that covariation cannot account for misrepresentation, since any seeming case of misrepresentation by a cognizer becomes a correct representation of a disjunctive property under nomic covariation. One never mistakenly believes that one sees one's mother at the corner, one always correctly believes that mother/other is at the corner.

Covariationists suggest a number of solutions to the disjunction problem. All of these solutions rely upon some form of idealization. A given solution will separate cases of the tokening (occurrence) of a state into two groups, one in which content is already fixed and representational error can occur, and the second (ideal circumstance) in which content is fixed by perfect covariation. There are several important versions of this solution:

Ideal conditions

Advocates of the ideal conditions solution to the disjunction problem suggest one represents 'mother' and not 'mother/other' despite occasionally mistaking others for mothers, because under ideal perceptual conditions (i.e. in good light, at close distance, etc.) one can always distinguish mothers from others. Difficulties arise for

the ideal conditions solution when one tries to delimit ideal conditions in a precise, nonquestion-begging manner. For instance, in the movie *The Crying Game* an Irishman becomes romantically involved with a woman. However, much to the surprise of the Irishman (and audience) this woman is actually a man. The reaction of the Irishman clearly shows he misrepresented the gender of his romantic partner. Moreover, he saw him in good light, at close distance, etc. The natural suggestion would be that there are conditions under which such a difference would not escape notice. However, in specifying such conditions as ideal conditions for this case, one must be guided in a circular manner by one's knowledge of the property that the state actually represents.

Critics (Wallis 1994a, 1994b) also argue that appealing to idealization to defeat the disjunction problem is strongly and negatively disanalogous with successful scientific uses of idealization.

Learning Periods

Fred Dretske (1981) suggests that a learning period fixes content. In this period, cognizers develop perfect causal connections between states and objects/properties with the help of an instructor who provides examples and corrections. Once the learning period ends, the cognizer's state has a fixed content, and tokenings of the state in cases where the object/property is absent count as misrepresentations.

Fodor (1988) and others criticize this approach on two grounds. First, there seems to be no principled distinction between learning and nonlearning periods. Hence, there seems to be no grounds for calling some tokenings of the state content imbuing and others representing or misrepresenting. Second, even if one could specify a learning period, there seems to be no principled distinction between the univocal and disjunctive causal connections. Critics ask why one ought suppose that causal connections created in the learning period hold between, for example, cats and S_c and not between cat/possum and S_c.

Teleological Accounts

Millikan (1983, 1986) and other advocates of teleological solutions to the disjunction problem suggest that evolutionary history determines content. For example, frogs capture and eat ambient moving dots in their visual field. Consequently, hungry frogs will eat ball bearings rolled in before them. Teleolosemanticists claim frogs misrepresent ball bearings as flies because the frog's visual state has an indicative function in virtue of the state's past co-occurrence with flies. This use of those cells then increases the probability that the frog will propagate its genome.

Fodor (1988) objects that evolutionary selection lacks sufficient precision to account for typical univocal content claims. If frogs represent ambient dots as 'fly or ball bearing' in a fly-rich, ball bearing-poor environment, then that disjunctive content accounts for the frog's ability to propagate its genome. Hence, evolutionary history will not favor fly over fly or ball bearing as the content of frog visual cells.

Furthermore, most artifacts (computers, etc.) were not present during the greater part of human evolution. Teleolosemanticist solutions to the disjunction problem must, therefore, explain a huge percentage of the representational capacities of humans, including the representations of many ordinary objects (e.g. chairs or beer), by definition. Similarly (Cummins, 1989; Davidson, 1987), since teleolosemantics appeals to the evolutionary history of cognizers to explain representational abilities, any seeming representational abilities lacking such a history lack representational content. For example, if a molecule-for-molecule duplicate of, say, Millikan spontaneously appeared, it would seem to have her cognitive/representational abilities. Yet, on Millikan's account, her exact double lacks any representation states. Finally, Cummins argues that one cannot appeal to the adaptational significance of a state to explain its representational content, since in order for the state to have adaptational significance it must play a part in the cognitive capacities of the system, and hence already have representational content.

Asymmetric Dependence

Fodor's (1990) asymmetric dependence theory holds that in addition to a nomic connection between a state, S_c, and an object/property, cathood, representation requires counterfactual connections: if, say, possums cause S_cs, then (1) breaking the possum to S_c connection does not break the cat to S_c connection and (2) breaking the cat to S_c connection does break the possum to S_c connection. As above, a (counterfactual) set of conditions in which subjects can distinguish cats from possums excludes potentially troublesome cases. Critics (Cummins, 1989; Wallis, 1995) argue that asymmetric dependence theories fare no better than other versions of covariation. Since the brain recognizes higher level concepts through the detection of

features, it appears that there is no asymmetric dependence between cases, or worse, it goes the wrong way: from fake to representation. For instance, if I find my car by looking for features x, y, z, I can break the car to S_{car} connection by altering the appearance of my car. But that will not break the look-alike to S_{car} connection (violating 2). If, on the other hand, I break the look-alike to S_{car} connection, then it seems I do break the car to S_{car} connection (violating 1). Similar stories can be told in terms of the normal, albeit somewhat noisy, functioning of cells in the visual system.

Functional Role Semantics

Whereas covariationists focus upon a single causal connection, advocates of functional role semantics (Block, 1986, 1987; Field, 1977, 1978; Harman, 1987) suggest that the overall network of causal relations into which a state can enter fixes its content. Often causal roles are specified functionally/computationally. Versions of functional role semantics include conceptual role semantics, procedural semantics, and inferential role semantics. There are two versions of functional role semantics designated either as (1) 'wide' or 'long-armed' or (2) as 'narrow' or 'short-armed'. Narrow functional role theorists limit content determining causal relations to those occurring between mental states. Wide theorists allow connections to the distal environment and even social contexts as well.

Functional role theories have several attractive features. First, they do not need bifurcated accounts of representational content since all states get their content in the same manner, via their functional role. Hence, functional role theorists avoid explaining the representation of high-level properties via definition in terms of low-level properties. Second, functional role theorists accommodate the observation that changes in beliefs can result in changes in representational content, since changing beliefs will often change the functional roles of states. Finally, functional role theorists can capture the intuition that in trying to understand the representational content of another cognizer's state, one is constrained by the heuristic that the overall set of content ascriptions must 'make sense', i.e. be consistent with the supposition that, overall, the cognizer's interactions with the world are intelligent or rational.

Semantic Holism

Theorists typically raise three distinct but related objections to all versions of functional role semantics. First, identifying particular states in disparate individuals as states having the same representational content *prima facie* requires that the states have identical functional roles. 'Holism', about content (also 'semantic holism') appears to imply, for example, that seemingly both Bill and Bob believe that 'Wallis's article is enlightening'. Each points to the same article, proclaims it enlightening, assigns it to his class, etc. However, the causal roles of their respective states differ in a single respect: Bill believes 'Wallis is a pompous know-it-all', while Bob believes 'Wallis is a precocious windbag'. *Prima facie*, functional role semantics dictates that Bill's belief that 'Wallis's article is enlightening', differs in content from Bob's. Similarly, people with disparate cognitive or perceptual abilities seem to have disparate representational contents. Thus, the theory appears to imply that the concept of traffic light differs in color blind humans and non color blind humans.

Functional role semanticists adopt one of two responses to the above objection. One response prunes the number and/or kind of causal/computational connections necessary for belief/content identity. It thereby allows for belief/content identity across individuals with somewhat different causal/computational roles. Differences in beliefs peripheral to 'Wallis's article is enlightening' would not necessarily constitute the basis for content nonidentity.

While the just-rehearsed response has intuitive appeal, critics (Fodor, 1988) emphasize potential difficulties in distinguishing core (central) beliefs or other causal links from peripheral ones in a manner that is not hopelessly unsystematic and ad hoc. For instance, what causal connections (beliefs, desires, dispositions to take action, etc.) constitute the core of one's belief that the colorless, tasteless, odorless liquid is water? Need one know of the existence of deuterium oxide (heavy water), make appropriate inferences with regard to D_2O, discriminate between D_2O and H_2O, etc.? In the case of beliefs and inferences, one common suggestion includes only analytic beliefs and corresponding inferences in the set of core beliefs and inferences. Analytic truths are conceptual truths, those things that are true solely in virtue of the nature of the concept. Such a suggestion requires a real distinction between analytic beliefs and nonanalytic (synthetic) beliefs. Many philosophers believe that Quine (1953, 1960) has effectively undermined the robustness of such a distinction. Stich (1983) argues that judgments of belief/content identity are more intuitively imprecise than advocates of analyticity predict.

The second response to the *prima facie* difficulty of intuitively identical content across differing functional roles claims that while such beliefs are in fact noncontent identical, they are strongly content similar. The key concept is that belief identity is not a binary notion but ranges from completely nonidentical to completely identical (Cummins, 1989). The strength (or weakness) of this response lies in its ability to accommodate the intuition that Bill's and Bob's beliefs have the same content while simultaneously acknowledging the theoretical constraint that differences in functional roles dictate differences in content.

Critics argue that graded notions of content identity, in addition to being counterintuitive, undermine the formulation of psychological generalizations and subsumption of particular cases under those generalizations. Cognitive science, they claim, would be reduced to the unworkable notion that Bill's and Bob's beliefs are, say, 97 percent content similar to the belief that Wallis's article is enlightening and hence can be, say, 97 percent subsumed under generalizations regarding the belief that Wallis's article is enlightening. Furthermore, Fodor (Fodor and LePore 1992; Fodor 2001) claims that colloquial notions of belief similarity such as, 'His notion of mental representation is similar to mine', as well as their theoretical counterparts, presuppose a notion of belief identity that he claims cannot be provided by holistic theories in any cases where beliefs diverge.

Content Fixing, Error, and Univocal Contents

An objection closely related to holism argues that functional role theories have no nonarbitrary way either of fixing content, distinguishing representational error from veridical representation, and/or they result in nonunivocal content ascriptions or multiple content ascriptions applying simultaneously to a given state. For example, suppose there are two worlds: one, call it $Earth_1$ in which there is no water but in which D_2O (or some chemically different but phenomenally similar substance) is plentiful; and another, call it $Earth_2$, where H_2O is plentiful, but there is no D_2O. Bob grows up on $Earth_1$, where interacting with D_2O results in his forming beliefs, etc. about 'water'. Bill matures on $Earth_2$, developing the exact same set of beliefs, etc. about 'water' based upon interacting with H_2O. Narrow functional role theorists (using only causal connections internal to the cognizer) holds that the two men must have identical belief contents when thinking about 'water'. They either form beliefs, make inferences, etc., using states that represent what we would describe using the disjunction, 'H_2O or D_2O', or their states represent both H_2O and also D_2O (as distinct entities) simultaneously. Most philosophers consider the latter even less intuitively plausible than the former.

Long arm or two factor theories can distinguish Bill's and Bob's beliefs through causal connections to environmental objects/properties. However, suppose that Bob also has the beliefs that 'water is H_2O', that 'I live on $Earth_2$', and '$Earth_2$ has only H_2O on it's surface'. Is Bob representing 'water' as D_2O but forming a false belief about the chemical structure of D_2O? Or, is he representing 'water' as H_2O and forming false beliefs about the D_2O he finds in his environment. Critics assert that functional role semantics lacks the resources to distinguish clearly between such scenarios.

COMPOSITIONALITY AND SYSTEMATIVITY

Finally, Fodor and LePore (1992), Fodor and McLaughlin (1991), as well as Fodor and Plysyhyn (1988), object that functional role semantics seems to violate truths about the structure of language and thought such as compositionality and systemativity. Compositionality is the theory that the meaning of a complex expression in a language results from the meanings of its constitutive elements. Compositionality plays an central role in many linguistic theories, since its supposition for both language and thought provides a fairly straightforward explanation of the human ability to grasp an enormous number of different thoughts of varying complexity and their corresponding linguistic expressions. For instance, because we understand individual notions like 'cup' and 'coffee', we understand the complex expressions 'cup of coffee', etc. We account for our understanding by noting that the meaning of these complex sentences is built up from the meanings of their constitutive elements.

Critics claim that the functional role of a complex, nonidiomatic representation is not always a function of the functional roles of its parts. As a result, functional role theories cannot represent 'cup of coffee' without having the concepts and associated inferences of 'cup' or 'coffee'. This alleged aspect of functional representation schemes results in the possibility that one could represent and think about 'cup of coffee', but could not represent or think about 'iced coffee', 'hot coffee', etc.

CAUSAL THEORIES OF MENTAL CONTENT AND COGNITIVE SCIENCE

Cognitive science progresses in the absence of a resolution of the debate over theories of representation. However, the adoption of either covariance theories or functional role semantics shapes the research of individual cognitive scientists. Many important debates in cognitive science have resulted from allegiance to one or the other causal theory. For example, one aspect of the connectionist versus Turing-compatible approaches is that advocates of connectionist frameworks tend to adopt functional role semantics, while advocates of Turing-compatible frameworks tend to adopt covariation. Similarly, the debate over the nature of mental imagery was drawn along covariance versus functional role lines. It is widely supposed that the adoption of one or the other theory of mental content by the majority of cognitive scientists will have a profound impact on the field.

References

Block N (1986) Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* **9**: 283–328.

Block N (1987) Functional role and truth conditions. *Proceedings of the Aristotelian Society* (supplement 61) 157–181.

Cummins R (1989) *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Davidson D (1987) *Knowing One's Own Mind*. Proceedings of the American Philosophical Association.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Field H (1977) Logic, meaning, and conceptual role. *Journal of Philosophy* **74**: 379–409.

Field H (1978) Mental representation. *Erkenntnis* **13**: 9–61.

Fodor J (1988) *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor J (1990) *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor J (2001) *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Fodor J and LePore E (1992) *Holism: A Shoppers' Guide*. Oxford: Blackwell.

Fodor J and McLaughlin B (1991) Connectionism and the problem of systematicity: why Smolensky's solution doesn't work. *Connectionism and the Philosophy of Mind*. Boston: Kluwer Academic Publishers.

Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Harman G (1987) (Non-solipsistic) conceptual role semantics. In: Lepore E (ed.) *New Directions in Semantics*. London: Academic Press.

Hubel D and Wiesel T (1977) Ferrier lecture: functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society London* **198**: 1–59.

Millikan R (1983) *Language, Thought and other Biological Categories*. Cambridge, MA: MIT Press.

Millikan R (1986) Thought without laws: cognitive science without content. *Philosophical Review* **95**: 47–80.

Quine W (1953) *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Quine W (1960) *Word and Object*. Cambridge, MA: MIT Press.

Stich S (1983) *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Wallis C (1994a) Ceteris paribus laws and psychological explanation. *Philosophy of Science Association 1994* **1**: 388–397.

Wallis C (1994b) Representation and the imperfect ideal. *Philosophy of Science* **61**: 407–428.

Wallis C (1995) Asymmetric dependence, representation and cognitive science. *Southern Journal of Philosophy* **33**: 373–401.

Further Reading

Cummins R (1989) Meaning and Mental Representation. Cambridge, MA: MIT Press.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: The MIT Press.

Fodor JA (1987) *Psychosemantics*, Cambridge. Cambridge, MA: The MIT Press (1987).

Sterelny K (1990) *The Representational Theory of Mind: An Introduction*. Cambridge, UK: Blackwell.

Mental Content, Nonconceptual

Intermediate article

Jose Luis Bermúdez, University of Stirling, Stirling, UK

CONTENTS

It has long been thought that the ways in which a creature can represent the world are determined by its conceptual capacities. Advocates of nonconceptual content have challenged this by suggesting that some mental states can represent the world even though the bearer of those mental states does not possess the concepts required to specify their content.

INTRODUCTION

Recent work in philosophy, psychology and cognitive science has suggested that there might be various types of cognitive state which possess content without necessarily requiring possession of the concepts needed to specify that content. If there are indeed states with this sort of nonconceptual content, then this suggests that the states involving concepts, which have long been thought to be the only means by which cognition can take place, are not really so distinctive. The familiar propositional attitudes and the cognitive processes defined over them might be part of a larger set of cognitive processes which have representational content by virtue of features that do not require the possession of concepts.

WHAT IS NONCONCEPTUAL CONTENT?

The content of a representational state is, to a first approximation, the state of affairs in the world that has to be the case for that state to be correct or true. The notion of content needs to be approached via the notion of truth, whether the representational states in question are sentences or mental states like beliefs. If, therefore, one wants to specify the content of a sentence or a belief one must specify its truth-condition. But not every way of specifying a state of affairs will provide an appropriate specification of the content of the sentence or belief that is true just if that state of affairs holds. It would be incorrect, for example, to characterize the content of my current belief that my pen is resting on a white piece of paper by describing its truth-condition using the concepts of particle physics. This gives rise to an obvious question: how does one specify a truth-condition in a way that reflects the way in which that truth-condition is apprehended by the utterer of the sentence or the thinker of the belief?

Different theories of content offer different answers to this question, but it is natural to think that no theory of content will be able to answer it unless it satisfies the following constraint:

> ***The conceptual constraint***. Specifications of the content of a sentence or propositional attitude state must not employ concepts that are not possessed by the utterer or thinker. (1)

So, if one specifies the content of a belief in the standard way by employing a sentence following a 'that' clause, the conceptual constraint requires that the specifying sentence should only involve concepts that are possessed by the believer. Of course, the fact that a content specification satisfies the conceptual constraint will not guarantee that it reflects the way in which the thinker apprehended the truth-condition. But it has struck many as plausible that it is not possible to capture what a thinker believes unless one restricts oneself to concepts that the thinker possesses.

Certain theories of content and concepts directly entail the conceptual constraint. It is a trivial consequence, for example, of the broadly Fregean view according to which the objects of belief and the meanings of sentences are propositions that are literally composed of concepts. But the conceptual constraint does not depend on any particular theory of content. Its plausibility stems, rather, from the conjunction of two thoughts. The first is that in specifying what a thinker believes or what a speaker is saying by uttering a certain sentence in a

particular context one has to be as faithful as possible to the way in which that thinker or speaker apprehends the world in having beliefs about it or speaking about it. The second is that the way in which a speaker or thinker apprehends the world in speaking about it or having beliefs about it is a function of the concepts he or she possesses. (*See* **Concepts, Philosophical Issues about**)

The basic idea in the theory of nonconceptual content is that there are certain types of representational state for which the conceptual constraint is not applicable. The proposal is that it can be theoretically legitimate to refer to mental states that represent the world but that do not require the thinker of those mental states to possess the concepts required to specify the way in which they represent the world. Alternatively put, a particular content is nonconceptual if (and only if) it can be attributed to a creature without thereby attributing to that creature mastery of the concepts required to specify that content.

The conceptual constraint can be lifted in two different ways. It can be lifted globally, by simply denying that any content specifications need confine themselves to the concepts possessed by the utterer or thinker. The currently popular identification of propositions with functions from possible worlds to truth values involves a global lifting of the conceptual constraint. Possible world semantics is intended to apply to all propositional attitudes, and it is obvious that few believers who are not also professional philosophers will grasp the theoretical concepts of possible worlds semantics. This article will not consider any such global strategies. The conceptual constraint seems extremely plausible for the contents of propositional attitudes.

Alternatively, the conceptual constraint can be retained for the core cognitive domain of the propositional attitudes but lifted locally by identifying particular representational domains for which the conceptual constraint seems inappropriate. There are three different representational domains for which a local lifting of the conceptual constraint has been proposed: perceptual states; representational states at the sub-personal or sub-doxastic level; and the representational states of non-human animals and human infants who do not seem to be concept possessors. Each of these three domains is a potential field of application of the notion of nonconceptual content.

ARGUMENTS FOR NONCONCEPTUAL CONTENT

The notion of nonconceptual content has been motivated within the context of three different explanatory tasks: the task of explaining the content of perceptual experience; the task of explaining the content of sub-personal representational states; and the task of explaining the behavior of certain non-human animals and of pre-conceptual human infants.

Perceptual Experience and Nonconceptual Content

Although some philosophers have suggested that perceptual states should be analyzed as propositional attitudes, usually by treating them as dispositions to form beliefs (Armstrong, 1968), there seem to be several respects in which the content of perception is drastically different from propositional attitude content obeying the conceptual constraint. (*See* **Perception, Philosophical Issues about**)

First, the content of perception seems to be analog in nature, unlike the conceptual content of propositional attitudes which is more plausibly seen as digital. The distinction between analog and digital representations has (for our purposes) been most perspicuously put by Dretske (1981). Let us take a particular fact or state of affairs, say the fact or state of affairs that some object *s* has property *F*. A representation carries the information that *s* is *F* in digital form if and only if it carries no further information about *s* other than that it is *F* (and whatever further facts about it are entailed by the fact that it is *F*). But whenever a representation carries the information that *s* is *F* in analog form, it carries additional information about *s*. It is clear that perceptual states represent the world in analog form and propositional attitudes in digital form.

Second, the content of perception seems to be 'unit-free' (Peacocke, 1986). If I perceptually represent an object as being a certain distance from me I do not usually represent that distance in terms of a particular unit (in inches, say, as opposed to centimetres), even though what I represent is a determinate distance. I simply represent it as being that distance, where the content of my perception specifies the distance. Propositional attitudes, however, can only represent distances (and other comparable quantities) in terms of specific units.

Third, the content of perception is more fine-grained than the content of propositional attitudes. I can see many more colors than I can name, and discriminate many more shapes than I have concepts for. My belief that the grass is green has a single content; but it would be the appropriate response to an enormous variety of perceptual states.

It seems plausible to conclude from these three features of perceptual experiences that their contents, unlike the contents of the beliefs that might be based upon them, are not circumscribed by the concepts we possess.

Sub-personal Computational States and Nonconceptual Content

It is common in various areas of the cognitive science to postulate the existence of representational states at the sub-personal or sub-doxastic levels (for philosophical discussion see Stich, 1978, and Davies, 1989). A good example comes from the representational states implicated in tacit knowledge of the rules of syntax. It is a fundamental tenet of a broadly Chomskyan approach to syntax that speakers have tacit knowledge of a grammar for their language and that this tacit knowledge is deployed in understanding spoken language. Yet when linguists gives theoretical specifications of the syntactic rules contained within the grammar they frequently employ concepts that are not in the conceptual repertoire of the language user. That is, the language user is ascribed knowledge of rules formulated in terms of concepts that he or she does not possess. A similar point holds for the representational states postulated in computational theories of vision like that put forward by Marr (1980). The contents of such states are formulated in terms of concepts (such as the concept of a zero-crossing) that are not possessed by the typical perceiver (Bermúdez, 1995).

Why should it be thought that the language user does not possess the relevant concepts? How could he grasp the rule if he did not possess the concepts required to spell it out? The reason for denying that the conceptual constraint is operative here is not just that language users are not aware of the beliefs in question. Not all unconscious beliefs are nonconceptual. The point, rather, is that their representations of the linguistic rules are completely inferentially insulated from the rest of their beliefs and propositional attitudes, in a way that is fundamentally incompatible with the 'holistic' nature of conceptual contents. Something like this idea seems to have been at the root of Gareth Evans's pioneering discussion of nonconceptual content (Evans, 1982).

Psychological Explanation and Nonconceptual Content

The final set of arguments for the existence of nonconceptual content comes from the need to give adequate explanations of the behavior of nonlinguistic and prelinguistic creatures. In explaining the behavior of such creatures cognitive ethologists and developmental psychologists often appeal to representational states. Arguably they will not be able to provide adequate explanations at all unless they do this. And yet there are ways of understanding what it is to possess a concept on which it seems inappropriate to attribute mastery of the corresponding concepts to the creatures whose behavior is being explained.

One such argument stresses the relation between possessing concepts and being able to justify certain canonical judgments involving that concept (Peacocke, 1992), going on to argue that providing justifications is a paradigmatically linguistic activity: a matter of identifying and articulating the reasons for a given classification, inference or judgment (McDowell, 1994; Bermúdez, 1998). There is a variety of possible responses to this argument. It might be objected, for example, that possessing a given concept simply requires being able to make justified judgments involving that concept rather than being able to justify judgments involving that concept. Or it might be objected that the ability to justify concepts is not necessarily linguistic, since it is possible to identify the justification for a judgment without engaging in communication.

One might see the issues here as being whether intentional assent requires semantic assent – that is, is it possible to have higher-order thoughts about a given thought (such as, for example, the higher-order thought that it has a certain justification) if that thought is not linguistically formed? The dependence theorist might argue that having a higher-order thought requires holding the target thought 'in mind' in a way that is not possible unless that target thought is linguistically formed. Supporters of the language of thought hypothesis will object that the dependence theorist is equivocating between a thought being formed in a public language and its being formed in a private language of thought.

At any rate, the argument from the need to provide psychological explanations of the behavior of non-human animals and human infants to the existence of nonconceptual content does not stand or fall with the thesis that concepts are necessarily linguistic. It seems plausible that there is a distinction between two types of thinking. Most students of the type of cognition engaged in by animals and infants view it as being domain-specific and modular in important respects, best understood in terms of bodies of 'knowledge' focused on particular

aspects of the natural and social worlds. These domain-specific modules have evolved separately and for specific purposes and are not integrated with each other (Hirschfeld and Gelman, 1994). In contrast, many philosophers have suggested that the type of conceptual thought engaged in by language users is essentially domain-independent, systematic and productive. Concept possessors can generate an indefinite number of new thoughts from the concepts they possess, and their thoughts obey what has been termed the generality constraint (that is, any thought-constituent can at least in principle be combined with any other). (*See* **Modularity**)

If these characteristics of conceptual thought are taken to be essential properties (as it is plausible to do) then it seems to follow that many (if not all) nonlinguistic creatures are not capable of engaging in conceptual thought. So, if they are correctly described as representing the world at all, their representations must be nonconceptual.

PROBLEMS FOR NONCONCEPTUAL CONTENT

Some problems for the notion of nonconceptual content arise whatever field the notion is applied to. Others are specific to particular applications. Some of these specific problems are problems more for the premises on which arguments for nonconceptual content are based than for the notion of nonconceptual content itself – such as, for example, the objection that sub-personal computational states do not have contents at all, and a fortiori not nonconceptual contents. Below we will consider only problems with the notion itself.

Determining a Theory of Content

The characterization we have so far given of the notion of nonconceptual content is incomplete. All we know so far are the formal characteristics of the notion. We do not yet have a substantive account of nonconceptual content to compare with, for example, a Fregean approach to the semantics of conceptual thought. The first problem, therefore, is to provide such a substantive account.

The most developed proposal has come from Christopher Peacocke, who has proposed a radically externalist conception of nonconceptual content aimed explicitly at explaining the nonconceptual content of perceptual states. Peacocke suggests that a given perceptual content should be specified in terms of the ways of filling out the space around the perceiver that are consistent with the content's being correct. For each minimally discriminable point within the perceiver's perceptual field (where these are identified relative to an origin and axes centred in the perceiver's body) we need to start by specifying whether it is occupied by a surface and, if so, what the orientation, solidity, hue, brightness and saturation of that surface are. This specification tells us the way in which the perceiver represents the environment. The content of that representation is given by all the ways of filling out the space around the perceiver in which the minimally discriminable points have the appropriate values. The representation is correct just if the space around the perceiver is occupied in one of those ways.

This proposal provides an attractive way of capturing the distinctive features of the phenomenology of perception highlighted above. Scenario content will be analog and unit-free, and will possess the appropriate fineness of grain. It is not clear, however, how it can be applied beyond the domain of perception. It may well be that certain sub-personal states have scenario contents: those associated with the sub-personal underpinnings of vision are obvious candidates. But the representational states implicated in tacit knowledge of syntactic theory do not fit the scenario model.

The natural conclusion to draw is that different specific accounts of nonconceptual content may be appropriate for different theoretical ends. There is no reason to think that there is a single account of nonconceptual content adequate for all applications.

Distinguishing Nonconceptual Content From Conceptual Content

The notion of nonconceptual notion is, of course, a contrastive notion. It depends for its meaning and interest on the correlative notion of conceptual content, and hence on how concepts are understood. The thinner and less demanding the notion of a concept becomes, the less scope there is for the notion of nonconceptual content to have useful meaning. If, for example, possessing the concept of an *F* simply requires being able to discriminate *F*s from the rest of the perceptual environment, or to act on them in a suitable manner, then it is hard to see how any evidence that animals and young infants represent the world will not also be evidence that they represent the world conceptually.

We have already seen some of the resources available to the theorist of nonconceptual content at this point. Such a theorist might, for example,

stress large-scale properties of types of thinking, distinguishing domain-specific thought at the nonconceptual level from domain-independent thought at the conceptual level. Alternatively, they might try to establish some kind of constitutive link between concept possession and language mastery, perhaps via the claim that intentional assent requires semantic assent. (*See* **Concepts, Philosophical Issues about**)

The Autonomy of Nonconceptual Content

One question that arises is whether a thinker can be in states with (nonconceptual) content despite not possessing any concepts at all. That is, can nonconceptual content be completely independent of conceptual content? Many theorists who wish to employ the notion of nonconceptual content to explain the behavior of nonlinguistic and prelinguistic creatures are committed to giving an affirmative answer to this question (depending, of course, on how demanding their notion of a concept is). So too are those theorists who hold both that sub-personal computational states possess nonconceptual concepts and that the relevant modules can exist in creatures who are not capable of conceptual thought.

Peacocke (1992) has offered an argument against this 'autonomy' thesis. His argument is based on a neo-Kantian understanding of the relation between experience of an objective world and self-consciousness. In essence, he suggests that no creature can be in content-bearing states unless it grasps that the surrounding environment has a minimal degree of objectivity and is able successfully to identify and recognize particular locations within it. This minimal grasp of objectivity requires being able to represent both the spatial configuration of the environment and one's own position within that environment – and, he suggests, this would be impossible for a creature that lacked a concept of the first person.

This argument is powerful, but can be challenged (Bermúdez, 1994). Supporters of the autonomy thesis have suggested that the interrelated capacities to represent the spatial configuration of the environment and to represent one's own location within the environment can be understood at the nonconceptual level, perhaps appealing to the notion of a nonconceptual point of view to explain the interdependence of spatial awareness of the surrounding environment and awareness of one's own location within that environment at the nonconceptual level (Bermúdez, 1998).

NONCONCEPTUAL CONTENT AND COGNITIVE SCIENCE

As we have already seen, one important motivation for the theory of conceptual content is to explain the various types of tacit knowledge that are often postulated at the sub-doxastic and sub-personal levels. The conception of tacit knowledge invoked here is at the heart of cognitive science and enters into such canonical theories as Marr's theory of vision and Chomsky's theory of innate syntactic knowledge. The notion of nonconceptual content therefore has a very clear application within the classical, rule-based conception of cognitive science.

But the notion of nonconceptual content is not applicable only within a classical rule-based approach to cognitive science. Nonconceptual content is a theoretical tool that can also be deployed by supporters of a more distributed and dynamical approach to cognition. As Adrian Cussins (1990) has pointed out, it is natural to interpret the sub-symbolic representational elements of a connectionist architecture in nonconceptual terms.

Moving from the sub-personal level to the personal level, the notion of nonconceptual content is potentially important in interpreting some of the recent experimental research on infant cognition. It is becoming clear that the perceptual experience of even the youngest infants is organized and structured in a way that reflects perceptual sensitivity to certain fundamental physical and dynamical properties of objects, as well as to basic principles of causality and number (Gopnik and Metzoff, 1997). It is arguably misleading to suggest, as many do, that this aspect of infant cognition reflects infant mastery of the concept of an object. Information is being processed nonconceptually and in a domain-specific manner. (*See* **Infant Cognition**)

The obvious practical and adaptive advantage of representing the environment nonconceptually is in the control and initiation of motor behavior. Navigating successfully through the environment often does not require identifying and conceptualizing the objects it contains. It is more important to have perceptual sensitivity to information about a limited range of object properties – position, motion, color, relative size, texture, distance, etc. – and to information, both exteroceptive and proprioceptive, specifying the perceiver's own location and movement. Such perceptual sensitivity can feed into motor behavior without any ability to conceptualize the information picked up. Thus an infant can reach out towards an object that it perceives as being within reach even though it has no concept of distance or reaching. In this sense the

theory of nonconceptual content is a natural ally of approaches to perception (Milner and Goodale, 1995; Jeannerod, 1997) that distinguish two distinct pathways in visual perception. It is a plausible hypothesis that we should understand the outputs of the pragmatic visual stream (the dorsal stream) in terms of nonconceptual content and those of the semantic visual stream (the ventral stream) in terms of conceptual content.

Finally, the theory of nonconceptual content provides a useful theoretical framework for presenting and evaluating research in cognitive ethology. Cognitive ethologists, unlike the older generation of comparative psychologists, tend not to try to explain how an animal behaves in terms of non-representational stimulus–response mechanisms or fixed behavior patterns such as innate releasing mechanisms. They start from the assumption that animals have certain desires and certain beliefs about how the world is organized and act on the basis of those beliefs to try to ensure the satisfaction of their desires. Then they look at the natural behaviors of a species, interpreting them as sophisticated strategies for pursuing the desires that members of that species seem to have. The theory of nonconceptual content provides a way of doing justice to the cognitivist methodology of ethological research while at the same time accommodating the differences between animal thinking and the thinking of language-using and concept-possessing humans.

References

Armstrong DM (1968) *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Bermúdez JL (1994) Peacocke's argument against the autonomy of nonconceptual content. *Mind and Language* **9**: 402–418.

Bermúdez JL (1998) *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Cussins A (1990) The connectionist construction of concepts. In: Boden M (ed.) *Artificial Intelligence*. Oxford: Oxford University Press.

Davies M (1989) Tacit knowledge and subdoxastic states. In: George A (ed) *Reflections on Chomsky*. Oxford: Blackwell.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Evans G (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Gopnik A and Metzoff AN (1997) *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Hirschfeld LA and Gelman SA (1994) *Mapping the Mind: Domain-Specificity in Cognition and Culture*. Cambridge, UK: Cambridge University Press.

Jeannerod M (1997) *The Cognitive Neuroscience of Action*. Oxford: Blackwell.

Marr D (1980) *Vision*. San Fransisco, CA: W.H. Freeman.

McDowell J (1994) *Mind and World*. Cambridge, MA: Harvard University Press.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford: Oxford University Press.

Peacocke C (1986) Analogue content. *Proceedings of the Aristotelian Society* **60**: 1–17.

Peacocke C (1992) *A Study of Concepts*. Cambridge, MA: MIT Press.

Stich S (1978) Beliefs and subdoxastic states. *Philosophy of Science* **45**: 499–518.

Further Reading

Bermúdez JL (1995) Nonconceptual content: from perceptual experience to subpersonal computational states. *Mind and Language* **10**: 333–369.

Bermúdez JL (1998) *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Crane T (1992) *The Contents of Experience*. Cambridge, UK: Cambridge University Press.

Cussins A (1990) The connectionist construction of concepts. In: Boden M (ed.) *Artificial Intelligence*. Oxford: Oxford University Press.

Evans G (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Peacocke C (1986) Analogue content. *Proceedings of the Aristotelian Society* **60**: 1–17.

Peacocke C (1992) *A Study of Concepts*. Cambridge, MA: MIT Press.

Stich S (1978) Beliefs and subdoxastic states. *Philosophy of Science* **45**: 499–518.

Mental Content, Teleological Theories of

Intermediate article

Ruth Garrett Millikan, University of Connecticut, Storrs, Connecticut, USA

CONTENTS

Teleological theories of mental content explain how emptiness and falseness can occur in thought. These defects are due to failure of the representation-producing devices to do the jobs they were designed to do. Teleological theories always rest on prior theories of what the jobs of mental representations are and of what their producers are.

INTRODUCTION

To describe the 'mental content' of, say, a belief or desire or intention is to describe what actual or possible state of affairs that thought represents. For example, if you believe that the earth is flat, or if you wish it were, the content of your thought is *that the earth is flat*. According to the nineteenth-century philosopher Franz Brentano, the distinguishing characteristic of mental states that represent or are about other things, is that they can bear real relations to nonexistent things or facts, as when someone believes that the earth is flat or thinks of a golden mountain. Brentano called this peculiar characteristic 'intentionality'. Intentionality poses a paradox for naturalistic theories of mind. How can a natural state of a natural creature bear a natural relation to something nonexistent?

WHAT ARE TELEOLOGICAL THEORIES OF MENTAL CONTENT?

'Telos' means purpose. Naturalistic teleological theories of mental content refer to the biological function or purpose of intentional mental states to explain Brentano's paradoxical relation. These theories generally begin with some more basic theory of the relation between a true thought, taken as embodied in some kind of brain state, and what it represents: for example, with the theory that true mental representations covary with or are lawfully caused by what they represent, or that they are reliable indicators of what they represent, or that they 'picture' or are abstractly isomorphic, in accordance with semantic rules of a certain kind, with what they represent. The teleological part of the theory then adds that the favored relation holds between the mental representation and what it represents when the biological system harboring the mental representation is functioning properly, that is, functioning in accordance with biological design or, perhaps, design through learning; but that when the system fails to perform as designed, false or empty representations may be produced. Then 'what is represented' may indeed be nothing real: to describe 'what is represented' is to describe what would have had to be the case with the represented or the world had the system produced or harbored that same representation when functioning properly.

Teleological theories of content thus sharply separate 'intentional' signs and representations, those capable of displaying Brentano's relation, from natural signs. Even when intentional representations are true, neither the fact that they represent nor what they represent is determined by any current relation they actually bear to what they represent. The representational status and the content of the intentional representation are both determined by reference to its natural purpose or the natural purpose of the biological mechanisms that produced it, and these purposes are determined, it is usually supposed, by history, by what these mechanisms were selected for doing, either during the evolution of the species or through earlier learning by trial and error.

Thus naturalistic teleological theories are 'externalist' theories of mental content. They imply that the content of one's thought is not determined by

anything before one's mind or within one's consciousness or even within one's head. Just as actually remembering something, rather than merely seeming to remember it, does not happen wholly within one's present head but requires that one has previously encountered that thing, thoughts that are about something actual also require the right sort of history. It would be possible for a teleologist to avoid this externalism only with an account of the nature of biological functions that was neither historical nor relative to the environment.

ARGUMENTS FOR TELEOLOGICAL THEORIES OF MENTAL CONTENT

Perhaps the best argument for teleological theories is that no other naturalist theory, taken alone, explains Brentano's relation without sacrificing the determinacy of mental content. For example, causal role theories, picture theories and natural-information theories all map mental content only onto idealizations of actual minds or brains: actual minds can make bad inferences, collect misinformation and fail to correctly identify things perceived. Nor do any of these theories, taken by itself, offer a determinate way to move from an actual mind or brain to just one idealization of it, hence to just one determinate set of contents for its intentional states. But on a teleological theory, if we assume that it is determinate whether a mind or brain is functioning as it was designed to function, the relevant idealization will be determined, no matter which underlying theory of representation turns out to be the correct one.

Nor do teleological theorists need to suppose that the occurrence of false beliefs means there is something wrong with our cognitive systems. Biological systems often fail to perform their more peripheral functions because the environment is wrong rather than because of internal failure. For example, however strong and healthy a bird's wings are, if submerged in water they will not be able to perform their functions properly. Nor does the teleological theory need to imply that most of our beliefs are true or useful. Successful performance of proper functions is often statistically rare. For a mouse, failing to escape from a cat or owl constitutes a failure of its behavioral systems to perform all of the functions for which they were designed. Nevertheless, most mice may eventually be eaten by some predator. If we claim that only states of true belief are biologically proper, this claim has no bearing on the relative statistical frequencies of true and false beliefs.

HISTORY

The teleological theory of mental content arose in the late twentieth century. Dennis Stampe (1977) is usually cited as the first to articulate it. In an effort to explain the possibility of representational 'infidelity' and 'vacuity' (Santa Claus, Macbeth's hallucination of a dagger) in a way that would be consistent with a causal theory of reference, he described the content of a representation as what the representation would probably have been caused by if its producing devices were 'functioning properly'. Gareth Evans (1982, pp. 128–129) described the content of 'information-based thoughts' in a very similar way (though more was required of them), and claimed that their producing mechanisms could 'malfunction', yielding 'informational states' that 'fail to fit their objects' or that are 'of nothing'. Stampe and Evans used the term 'function' without analysis, but David Papineau (1984) and Ruth Millikan (1984) independently proposed teleological theories of mental content which explicitly took 'function' to be defined by reference to natural selection or (for Millikan) learning by trial and error. Fred Dretske (1986, 1988) modified his own purely informational theory of mental content to include a teleological layer to account for 'misinformation', but describing the functions of belief-like states as derived only from a trial-and-error learning history, not from natural selection. At about the same time, Jerry Fodor briefly embraced teleological theory with an essay which he soon repudiated as 'viciously wrong', but was eventually persuaded to publish (Fodor, 1990a). Since then a considerable literature has appeared, criticizing, defending or embellishing teleological theories of content.

VARIETIES OF TELEOLOGICAL THEORIES OF MENTAL CONTENT

It is commonly said that according to teleological theories, the content of a mental state is determined by 'whatever it is the function of the mental state to represent'. This formulation is vacuous, however, unless what it is for one thing to 'represent' another is first specified. Teleological theories all need to rest on more specific underlying theories of the relation between a true representation and what it represents: of the relation that it is the purpose of the perceptual or cognitive systems to produce. We can classify teleological theories accordingly.

Stampe (1977) and Fodor (1990a) took the representing relation to be causal. Fodor claimed that

various kinds of external conditions could in principle be specified under which normal perceptual and cognitive systems would operate optimally in accordance with design, and that what a belief state in the brain represented was whatever would always cause it under these epistemically optimal conditions. Under these conditions, the occurrence of the represented would be sufficient for the occurrence of the representation.

Dretske claimed that the function of a perceptual representation is to 'indicate' or carry 'natural information' about the represented, meaning that if the representation is present, there should be a probability of one, in accordance with natural law, that the represented is also. Occurrence of the representation would thus be sufficient for occurrence of the represented. Theories of this sort might be called 'informational theories', in contrast to 'causal theories' like Stampe's and Fodor's, but the two types have not generally been clearly distinguished in the literature.

Dretske explicitly claims that representations have the 'function' of carrying information because of their situation in some larger system that makes use of the information to guide behavior. Papineau and Millikan claim that it is only the uses to which mental representations are put that are relevant to their content. Millikan claims that a true representation maps onto its represented in accordance with semantic rules determined by the way the systems using the representation are designed to react to it in guiding, perhaps first inference processes, but ultimately behavior. Representations are designed to stand in for aspects of the world outside the organism and, by varying in accordance with these aspects, to control the animal's behavior so as to take account of these aspects. Causal or informational relations between representation and represented play no role in the analysis.

Like Dretske on perceptual representations, Papineau sometimes says that the function of a belief is to be copresent with its represented. But he also says that what a belief represents, its truth condition, is the condition that would guarantee that actions based on that belief and one's other true beliefs will satisfy one's desires, the function of a desire being to produce its satisfaction condition. Again, causal or informational relations are not mentioned. For both Papineau and Millikan, a useful 'correspondence' between representation and represented does indeed occur when the biological system functions properly, but how this correspondence is brought about is not what defines the representing relation.

PROBLEMS FOR TELEOLOGICAL THEORIES OF MENTAL CONTENT

The best-known argument against teleological theories is the 'swampman' argument. Suppose first that teleological theories are right. Then suppose that by some cosmic coincidence, lightning striking a tree near a swamp beside which you are standing destroys you, but at the same time puts a different collection of molecules together out of the swamp to form another creature molecule-for-molecule exactly like you. This creature wouldn't be you, but it would talk like you and behave like you and, presumably, have exactly the experiences you would have had had you survived. But according to the teleological theory, it would have no intentional mental states at all – no beliefs, hopes or desires – because its perceptual and cognitive systems were not designed by evolution or learning. Everyone agrees that this creature wouldn't have memories, but beliefs and desires do not seem, intuitively, to require a history in the way that memories do. They seem to be wholly present-tense occurrences.

Another common objection asks how people's specific beliefs, such as Paul Revere's belief that the British were coming by sea, could each have a selectionist history. It is sometimes suggested that the human mind may not have been designed by natural selection at all, or at least not designed to work as it now works. But even if it has been so designed, surely neither natural selection nor learning by trial and error could have designed, specifically, each belief that each person has. However, the teleologist's position as spelled out by Millikan (1984, 2001) is that the human intellect works in accordance with very general principles of concept formation and belief fixation, applied repeatedly to diverse subject matters: somewhat like a calculator, capable of solving an infinite variety of mathematical problems, but always operating strictly in accordance with the same design. Still, is it plausible that, say, the contemporary theoretical physicist employs no basic principles in thinking except those in use during the evolutionary history of the species?

Other challenges concern whether a history of natural selection yields sufficiently determinate content for representations. Fodor (1990b) has complained that the supposed 'fly detector' that goes off in a frog's optic nerve when it sees a fly also goes off if any other small dark thing crosses its retina, nor has natural selection selected against, say, 'beebees' making it go off. Why then isn't the

detector a 'fly or beebee or …' detector? But it may be countered that what determines the function of a biological mechanism is what caused it to be reproduced more often than competing mechanisms. The fly detector was selected because it caught fies, not because it caught anything small and dark. Neander claims that because biological devices have many functions (the optic nerve fires, causing the tongue to stick out, causing ingestion of the fly, causing … more frogs eggs, and so on) it is not determinate what the function of the 'fly detector' is, hence what it means. But it is only when the fly detector detects a fly that any of these functions gets performed: there is no need to choose among them. Both Neander (1995) and Dretske claim that when the fly detector responds to an irrelevant dark shadow it is not malfunctioning, for there is nothing wrong with it. So detecting, specifically, flies cannot be its function. But this is like saying that opening cans is not the function of my can opener, because if it fails to open cans when I do not use it properly that is not its fault. A suitable environment is often necessary for a device to perform its functions.

TELEOLOGICAL THEORIES OF MENTAL CONTENT AND COGNITIVE SCIENCE

The relevance of the teleological theory to cognitive science is indirect. Cognitive scientists interested in mental representation do, of course, need to understand the relation between a representation and what it represents. But they have mostly been trying to understand and to model systems that are operating properly. Having attempted to model a correctly operating neurological system, they may damage the model in certain ways to see if this simulates certain kinds of damage to real neurological systems, but the point of this exercise is primarily to test the original model. Cognitive scientists are not generally concerned to model systems in environments that these systems are unable to handle. So, for the most part, they have no reason to take an interest in the nature of false representation. Their theories are theories of true representation. Suppose, however, that false beliefs and empty ideas do often derive, not just from faulty processing or from the brain's own inner errors, but from the environment being too uneven and difficult. Suppose that adequate concept formation and true belief fixation are easy or possible only when the environment fits certain common frameworks on which our mental tools are fit to operate. Cognitive systems may turn out to be as deeply embedded in, hence dependent on, their specific environments as are other biological systems. In this case, a strongly ecological cognitive psychology may ultimately be required.

References

Dretske F (1986) Misrepresentation. In: Bogdan R (ed.) *Belief: Form, Content, and Function*, pp. 17–36. New York, NY: Oxford.

Dretske F (1988) *Explaining Behavior*. Cambridge, MA: Bradford/MIT Press.

Evans G (1982) *The Varieties of Reference*. Oxford, UK: Clarendon Press.

Fodor J (1990a) Psychosemantics, or: where do truth conditions come from? In: Lycan W (ed.) *Mind and Cognition: A Reader*, pp. 312–337. Oxford, UK: Blackwell.

Fodor J (1990b) A theory of content. In: Fodor J. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Neander K (1995) Misrepresenting and malfunctioning. *Philosophical Studies* **79**: 109–141.

Millikan RG (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan RG (2001) Biofunctions: two paradigms. In: Cummins R, Ariew A and Perlman M (eds) *Functions in Philosophy of Biology and Philosophy of Psychology*. Oxford, UK: Oxford University Press.

Papineau D (1984) Representation and explanation. *Philosophy of Science* **51**: 550–572.

Papineau D (1993) *Philosophical Naturalism*. Oxford, UK: Blackwell.

Stampe D (1977) Toward a causal theory of linguistic representation. In: French PA, Uehling TE and Wettstein HK (eds) *Midwest Studies in Philosophy: Studies in the Philosophy of Language*, vol. II, pp. 81–102. Minneapolis, MN: University of Minnesota Press.

Further Reading

Antony L, Dennett D, Dretske F *et al*. (1996) Forum. *Mind and Language* **11**(1): 70–130.

Millikan RG (1993) Biosemantics. In: Millikan RG *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Von Eckardt B (1993) Current approaches to content determination. In: Von Eckardt B *What Is Cognitive Science*? Cambridge, MA: MIT Press.

Mental Imagery, Philosophical Issues about

Intermediate article

Nigel JT Thomas, California State University, Los Angeles, California, USA

CONTENTS

Mental imagery, often informally described as 'seeing in the mind's eye', 'visualization', etc., is quasi-perceptual experience: *it significantly resembles perceptual experience, but occurs in the absence of the appropriate perceptual stimuli.*

INTRODUCTION

Imagery is associated not only with fantasy and the imaginary, but also, and perhaps more importantly, with prototypically cognitive functions such as memory, perception, and thought. Although imagery occurs in all sensory modes, most work in philosophy, psychology, and cognitive science has (perhaps regrettably) concentrated upon *visual* imagery. (*See* **Imagery**; **Memory, Philosophical Issues about**; **Memory**; **Perception: Overview**; **Perception, Philosophical Issues about**)

IMAGERY BEFORE COGNITIVE SCIENCE

The Philosophical Tradition

The classical Greek philosophers set the stage for subsequent discussions of imagery. Plato speaks (metaphorically) of an inner artist painting pictures in the soul (*Philebus* 39c), and suggests that memory might be analogous to a block of wax into which our perceptions and thoughts stamp impressions (*Theaetetus* 191c,d). Aristotle endorses this wax impression model of memory, and describes this impression as a sort of picture (*De Memoria* 450a, b). He introduced the notion of a mental faculty of imagination, allied to perception, and responsible for producing and recalling imagery (*De Anima* III. iii). Aristotle was the first systematic cognitive theorist, and he gave imagery a central role in cognition. He asserts that 'The soul never thinks without a mental image' (*De Anima* 431a 15–20), and maintains that the representational power of language is derived from imagery, spoken words being the symbols of the inner images (*De Interpretatione* 16a; *De Anima* 420b). In effect, for Aristotle, images play something very like the role played by the more generic notion of 'mental representation' in modern cognitive science. This was almost universally accepted in the philosophical tradition, even by non-Aristoteleans, up until the twentieth century. With certain qualifications and exceptions (most significantly the 'clear and distinct ideas' of Descartes' epistemology), the 'ideas' that played such a large role in philosophy and cognitive theory from the seventeenth through the nineteenth century are direct descendants of Aristotle's images. Hume, for example, explicitly identifies ideas as images (as does even Descartes in psychophysiological contexts). (*See* **Representation, Philosophical Issues about**; **Descartes, René**; **Hume, David**)

Early Experimental Psychology and 'Imageless Thought'

Imagery played a large role in early experimental psychology (which, especially in Germany where it first flourished, was practiced as a branch of philosophy). Wilhelm Wundt, 'the father of experimental psychology', founded the first psychological research and teaching laboratory in 1876, and imagery played essentially the same pivotal cognitive role in his theories (and those of most of his many students and imitators) that it had played for the philosophers of former ages.

However, from about 1901 Oswald Külpe and his students at the University of Würzburg directly challenged these assumptions. Experimental subjects in Würzburg were asked to provide introspective reports of the contents of their consciousness as they performed specified cognitive tasks. The introspectors (often Külpe himself, or other members of his research team) frequently claimed to experience

not imagery, but rather 'imageless thoughts', conscious contents without any sensory or perceptual quality. In response, Wundt sharply criticized the introspective methodology of the Würzburg experiments, whilst Titchener, Wundt's leading disciple in America, reported that in *his* laboratory, similar introspective experiments *always* evoked imagery (not necessarily visual), producing no evidence whatsoever for conscious imageless thoughts. ('Thinking in words' is plausibly regarded as a form of auditory or vocal-kinaesthetic imagery (Paivio, 1971).) (*See* **Introspection**)

The bitter and irresolvable controversy that arose led to a reaction by which introspective methods became thoroughly discredited amongst the majority of experimental psychologists, and the notion of mental imagery fell out of intellectual favour. J. B. Watson, who inaugurated the very influential Behaviorist movement in psychology, questioned the scientific reality of consciousness in general, and imagery in particular. Between about 1920 and 1960, imagery received minimal scientific attention. The question of the reality of conscious imageless thoughts was left unresolved, and it remains so today (Thomas, 1989; Heil, 1998; Mangan, 2001).

Twentieth-century Philosophy

Amongst philosophers, few questioned the actual occurrence of quasi-perceptual experiences, and imagery continued occasionally to be discussed. However, few twentieth-century philosophers accorded it anything like the theoretically central position it once enjoyed. The analytical philosophy movement (that arose in the early twentieth century, and that still deeply influences most English-speaking philosophers) originated from the hope that philosophical problems could be definitively solved through the analysis of language, using the newly invented tools of formal logic. It thus treated *language* as the fundamental medium of thought, and several of the leading figures of the movement (notably Frege, Wittgenstein, and Schlick) argued strongly against the traditional view that linguistic meaning derives from images in the mind. (For a brief, though critical, summary of the main arguments see Thomas (1997a) and discussion below.) These arguments were widely accepted, and imagery was relegated to the sidelines of philosophy. It no longer seemed to have a vital functional role to play in the workings of the mind. (*See* **Frege, Gottlob**; **Wittgenstein, Ludwig**)

In his seminal *The Concept of Mind* (1949), analytical philosopher Gilbert Ryle set out to refute what he called 'Descartes' Myth': the notion that the mind is somehow a special arena distinct from the physical world, populated by mental (nonphysical) objects. Ryle thus vigorously attacked the notion of a mental image as a 'picture in the mind', and suggested instead that what people call 'imagining', 'picturing in the mind's eye', and so forth, would be better understood as akin to *pretending* (to oneself) to see something. Ryle also questioned whether we really have a coherent, unitary concept of imagination, and it remains controversial whether imagery is really relevant to other notions traditionally associated with imagination, such as creativity (White, 1990; Brann, 1991; Thomas, 1997b, 1999).

Working in the rival phenomenological philosophical tradition, Jean-Paul Sartre (1948) also questioned the cognitive role of imagery and the notion of mental pictures. He argued that an image 'teaches nothing', because any information it contains must have been put there by, and thus have already been in the mind of, the imaginer. Sartre stressed the *intentionality* of imagery, the fact that an image is always an image *of* something (perhaps something nonexistent), but he insisted that an image is not a *thing* in the mind. Although neither Sartre nor Ryle seems to have intended to deny the reality of quasi-perceptual experience, this may not always have been clear to their audience, and their work surely contributed further to the decline of interest in imagery in both analytical and phenomenological traditions. After this, and before the rise of cognitive science, the rare philosopher who wanted to insist on the importance of imagery, perhaps even its very existence, was noticeably on the defensive (Price, 1953; Hannay, 1971; Casey, 1976). (*See* **Phenomenology**; **Intentionality**)

IMAGERY IN COGNITIVE SCIENCE

The Imagery Revival

A revival of research on imagery was an important element of the cognitive revolution of the 1960s and 1970s, contributing greatly to the rising scientific interest in mental representations. Seemingly, this revival initially stemmed largely from applied psychology research on sensory deprivation and on hallucinogenic drugs (Holt, 1964). Another important catalyst was Yates' (1966) seminal historical work on the significance of imagery mnemonics in ancient through Renaissance thought. Once the powerful mnemonic properties of imagery were experimentally confirmed (Paivio, 1971), imagery could no longer be dismissed by psychologists. Interest was only heightened, during the 1970s, by

the stunning 'mental rotation' experiments of Shepard and his students (Shepard and Cooper, 1982), and experiments by Kosslyn (1980) demonstrating 'mental scanning' and related effects. This work was taken to demonstrate that imagery is involved in visuo-spatial reasoning, and has inherently spatial properties. (*See* **Memory Mnemonics**; **Mental Rotation**)

The 'Analog–propositional' Debate

But how could these findings on imagery be reconciled with the functionalist, computational symbol-manipulation approach to cognition that was emerging during the same period? The standard philosophical interpretation of this approach depicts cognition as the computational manipulation of representations expressed in 'mentalese' (the 'Language of Thought') – a hypothetical, essentially language-like, representational system supposedly built into the brain (Fodor, 1975). Two rival approaches arose toward integrating the empirical findings about imagery into computational cognitive science. (*See* **Functionalism**; **Language of Thought**)

Pylyshyn, in a series of influential papers (e.g. 1973, 1981), argued (in effect) that all the genuine phenomena associated with imagery (indeed, *all* truly mental phenomena) can and must be explained entirely in terms of mentalese representations. For Pylyshyn and his allies, the computational paradigm of cognitive science demands that the underlying representational reality of imagery (and of actual perceptual experience) is not picture-like, but rather a detailed mentalese *description* of a scene.

Other cognitive scientists, however, notably Shepard and Kosslyn, argued that the evidence implies that imagery must be a distinct, non-language-like form of representation. Kosslyn, in particular, developed a 'quasi-pictorial' computational theory of visual imagery, based on an analogy with computer graphics (Kosslyn, 1980). Computer graphics files store information in a compressed, non-pictorial form, but when they are displayed they are translated into a mathematical map (bitmap) of the computer monitor screen, that specifies the color at each pixel (tiny dot) on the screen itself. Likewise, suggests Kosslyn, visual information may be *stored* in the brain as compact descriptions, but we experience an image only when this information is used to create a two-dimensional map of visual space in a special, functionally defined memory area he calls the 'visual buffer'. The picture in Kosslyn's theory is merely 'quasi', because there is no equivalent to the monitor screen to display it. What we experience as imagery, and what is available to the cognitive processes that use imagery, is the functional picture, the mathematical map, in the visual buffer. In later work, Kosslyn (1994) identifies this 'visual buffer' with the several retinotopically mapped visual areas of the brain. (*See* **Visual Imagery, Neural Basis of**)

'Description' and 'quasi-pictorial' theorists disagreed sharply over what sorts of computational symbols, or data structures, are acceptable within cognitive theory, and which best capture the empirical properties of imagery. During the 1970s, in particular, this led to a lively and high-profile controversy, commonly, if somewhat misleadingly, known as the 'analog–propositional' debate. ('Picture–description' debate would have been better. 'Proposition' is jargon borrowed from philosophy, where it signifies the underlying *meaning* of a sentence, not, as is intended here, a descriptive 'sentence' of mentalese. Furthermore, the force of 'analog' in this context is hardly clear: Kosslyn, after all, models his quasi-pictures on digitized bitmaps.)

Although the leading combatants in this dispute were psychologists, and experimental evidence was frequently cited, many of the issues raised were conceptual or meta-theoretical in nature – Anderson (1978) questioned whether it was even possible to resolve the debate experimentally – and philosophers soon became involved. The very concept of mental representation seemed to be at stake. Many of the most influential articles from the heyday of this debate, by both psychologists and philosophers, are collected in two volumes edited by Block (1981a, b). Description theory still finds philosophical defenders (e.g. Slezak, 1995), but Tye (1991; see also Rollins, 1989) has undermined much of its appeal with a persuasive defence of the conceptual legitimacy of quasi-pictorial arrays as a distinct form of computational representation. Furthermore, many descriptionist explanations of empirical findings seem worryingly *ad hoc* (Kosslyn and Pomerantz, 1977).

However, Kosslyn's (1994) declaration of victory in 'the imagery debate' may be premature, even though he has certainly developed the venerable picture theory to an unprecedented level of empirical and conceptual sophistication. His (1994) recasting of quasi-pictorialism in neurological terms does little to resolve the significant problems it still faces. Pylyshyn (in press) has now launched a major counter-attack, not only restating his empirically- and conceptual-based objections to

quasi-pictorialism, but arguing forcefully that (despite many claims and some superficial appearances to the contrary) there is no firm evidence whatsoever to support pictorialism. Even the (much disputed) results suggesting that visual imagery experience is correlated with activity in retinotopically-mapped visual cortex (e.g. Kosslyn *et al*., 1995) are quite consistent with non-pictorial theories (Pylyshyn, in press; Thomas, 1999). Furthermore, quasi-pictorial theory does not parsimoniously account for a range of experimental results showing that people have difficulty, in many circumstances, in finding new representational meanings in their imagery, meanings that are relatively easily found in an actual picture (Slezak, 1995; Thomas, 1999). Also, it is unequipped to explain the fact that the experimental effects that it most directly addresses (mental rotation, mental scanning, mnemonic effects, etc.) have since been demonstrated to occur in congenitally blind as well as in sighted subjects. The blind experimental subjects apparently employ haptic (touch) imagery, but any haptic analogue of a quasi-picture would be quite unsuitable to play an equivalent explanatory role (Thomas, 1999). In addition to these empirical shortcomings, quasi-pictorialism fails to address two cardinal characteristics of imagery: its *intentionality* and its *consciousness* (see further below).

Beyond Pictures and Propositions?

Outside the theoretical context of symbolic computationalism, in which the 'analog–propositional debate' first arose, other accounts of imagery, neither pictorial nor descriptional, may become conceivable. Despite a handful of connectionist simulations of versions of quasi-pictorialism, the connectionist movement has had surprisingly little impact on imagery theory. However, more recent alternative approaches to cognition, such as 'dynamical systems theory', and 'situated' or 'embodied' cognition, call into question the basic assumption that mental contents are to be identified with representations in the sense of data structures, embodied as brain states, and manipulated in the cerebral computer. Related work on both robotic and human vision (and perception in general) is converging on the idea that perception is *not* best understood as the processing of sensory input into a detailed inner representation – a description or depiction, of the scene before us – but rather as ongoing, directed exploratory activity (e.g. Landy *et al*., 1996; O'Regan and Noë, 2001). On this view, as our brains direct all the most minute details of our ongoing behavior (including the exploratory perceptual activity itself) they require a constant stream of answers to specific questions about the detailed disposition of the environment, and, instead of seeking this information in a pre-established inner representation, they deploy the sense organs, like measuring instruments, in order to obtain the answers from the environment as and when needed. In order to explain how brains coordinate this activity, we may well need to invoke data structures in the brain. However, they would not encode descriptions or depictions of the environment, but, rather, the procedures that most appropriately direct its exploration. O'Regan and Noë (2001) suggest that perceptual experience is not the experience of having a representation (a percept) in the brain, but, rather, of being engaged in ongoing perceptual exploration. In similar vein, Thomas (1999) proposes that imagery experience (and the empirical data on imagery) is best explained as arising from a sort of abortive, and largely covert, *perceptual activity*: a truncated 'going-through-the-motions' of exploring objects or situations that are not actually there to be explored, under the control of an appropriate procedural representation. (*See* **Connectionism**; **Dynamical Systems Hypothesis in Cognitive Science**; **Dynamical Systems, Philosophical Issues about**; **Situated Cognition**; **Situated Robotics**; **Embodiment**)

Meanwhile, certain philosophers broadly sympathetic to this view of imagery, and with doubts about the 'standard' computational-functionalist view of the mind, have begun to revive the traditional conception of imagery as the vehicle of conscious thought, and the fundamental bearer of mental content, or intentionality (Heil, 1998, chap. 6; Ellis, 1995; Thomas, 1997a). Other recently elaborated approaches to cognition are built around closely related conceptions such as 'image-schemata' (Lakoff and Johnson, 1999) and 'perceptual symbols' (Barsalou, 1999). (*See* **Functionalism**)

KEY PHILOSOPHICAL ISSUES

Meanings of 'Imagery'

Few discussions of imagery draw a clear distinction between the claim that people have quasi-visual experiences and the claim that such experiences are caused by the presence of picture-like objects in the mind. In practice, in the literature, 'mental imagery' (or 'mental images') can mean any or all of at least three things:

(1) quasi-perceptual conscious experience *per se*;
(2) picture-like representations in the mind and/or brain that may be experienced as (1);

(3) any inner representations whatsoever that may be experienced as (1).

Picture theory is so entrenched in our language and our 'folk psychology' that it is only too easy to assume that when people say 'imagery' they mean (2). Far too many confident arguments about what imagery can or cannot do depend on an (often unacknowledged) assumption of pictorialism (Thomas, 1989, 1997a, b).

Block (1981a – Introduction) argues that some confusions could be avoided, without prejudging the 'analog–propositional' issue, if we agreed to define 'imagery' as (3). However, one can perfectly consistently maintain the reality, and perhaps even the cognitive importance, of imagery in sense (1) whilst denying the existence not only of (2) (with description theorists such as Pylyshyn or Slezak), but even of (2) *and* (3) (with Ryle, Sartre, or Thomas: see above). Defining 'imagery' as (1) (as we did initially) seems to beg the fewest questions.

Imagery, Intentionality, and Mental Representation

Nearly all philosophers accept that imagery has *intentionality*: the characteristically mental property of being *of*, *about*, or *directed at* some object (real or imaginary). In this regard, mental imagery proper may be distinguished from superficially similar but nonintentional phenomena such as afterimages and phosphenes. This philosophical concept of intentionality is only distantly related to the notion of doing an action *intentionally* (i.e. on purpose), but it is very closely related to the notion of meaningful representation. To say that an image is *of* a lion is to say that it *represents* a lion. But how is it possible for anything to have this property of intentionality, this power of being able to represent things? This is perhaps the most fundamental problem in the philosophy of cognition, and the answer we give to it will profoundly affect what sorts of cognitive theories we think are workable. (*See* **Intentionality**; **Representation, Philosophical Issues about**)

Most philosophers before the twentieth century held that mental images formed the basis of the mind's power to represent things, and probably assumed that images represent their objects because they *resemble* them: an image of a lion, like a photograph of one, *looks like* a lion. However, consider two photographs of Leo. Each photo resembles the other more than either resembles the lion (both photos are small, rectangular pieces of card, similarly marked, and neither is carnivorous or furry), yet we would normally want to say that they represent Leo, and *not* that they represent each other. Of course, a photograph of Leo does resemble him, when the right aspects of resemblance are considered, but in this regard Leo equally resembles the photograph. We are unlikely, however, to want to say that *he* represents *it*. In order for resemblance to play a role in representation the relevant aspects of resemblance have to be recognized, and the resembling object has to be *used* (or, at least, *taken*) as a representation. But surely, before a cognitive system can recognize or use the relevant aspects of resemblance between a photograph (or an inner quasi-picture) and an object (or a percept), it must already be able to represent the picture and its object, and their various features, to itself. The mind's power to recognize resemblance seemingly depends on its power to represent things, rather than vice versa.

From related arguments, Fodor (1975) concludes (with the analytical philosophy mainstream) that mental images do not possess their intentionality intrinsically. Rather, they derive it from that of another, supposedly more fundamental, form of representation. For Fodor, an image (a quasi-picture, he assumes) of a lion represents a lion not because it resembles a lion, but, in effect, because our minds attach a mentalese caption to it saying 'LION'. It is not that the resemblance is not real, or cognitively useful, but that (contrary to traditional views) mentalese, not imagery, is the fundamental form of representation and the source of intentionality. This line is apparently accepted by Tye (1991) in his philosophical exegesis and defense of quasi-pictorialism. Of course, those like Pylyshyn, who hold that imagery *consists of* mentalese, also ground the intentionality of imagery in that of mentalese.

It is worth noting, however, that, despite strenuous philosophical efforts since the 1970s, no generally acceptable theory of the source of the representational power of mentalese is forthcoming, and, indeed, none may be possible (Cummins, 1997). We might do without the rather extravagant hypothesis of mentalese if the intentionality of imagery could be derived from that of ordinary spoken language (Kaufmann, 1980). However, this would seemingly require an account of the intentionality of ordinary language that avoids reliance upon any appeal to the intentionality of the mental. The prospects for that seem poor. It would also apparently imply that animals and babies have no intentionality: that, in effect, they have no minds.

But just because the resemblance theory of representation fails, it does not necessarily follow that the intentionality of imagery is not intrinsic. After all, resemblance theory was only ever intended to

apply to pictorial theories of imagery. If imagery is conceived of as a form of directed perceptual activity (see 'Beyond Pictures and Propositions?' above), rather than as an inner representation, perhaps its intentionality might be understood as rooted in the inherent goal-directedness of action. Such an approach still awaits detailed articulation, however, and would presumably require an account of the intentionality of action that did not root *that* in the intentionality of representations. (*See* **Action, Philosophical Issues about**)

Imagery and Consciousness

According to most cognitive scientists, mental images are mental representations (pictorial or otherwise) that have their existence as brain states. How could we be conscious of such things (as we clearly can be conscious of imagery)? It seems unlikely to be in virtue of any intrinsic characteristic of the relevant brain state, for brain states are nothing but patterns of excitation of neurons, and excitation of neurons is nothing more than electrochemistry: the movements of certain molecules and ions, polarization and depolarization of membranes, and so on; quite ordinary physical processes that go on outside of brains, and even, quite frequently, within brains, without producing consciousness. (*See* **Consciousness, Philosophical Issues about**; **Neural Correlates of Visual Consciousness**)

Perhaps, then, such representations are conscious not because of any properties they have intrinsically, but, rather, we are conscious of them inasmuch as our minds access and extract information from them. This is a quite traditional idea embodied in the idiom of 'the mind's eye', and it found its clearest philosophical expression in Descartes' (1664) account of imagery, wherein the image is presented as a small physical picture formed deep within the brain (at the pineal gland), from where the immaterial conscious soul is able to apprehend it directly. Nearly all cognitive scientists (and contemporary philosophers) firmly reject such explanatory invocations of the supernatural. Nevertheless, Kosslyn, the leading contemporary quasi-pictorialist, sometimes writes of a 'functional mind's eye' inspecting and interpreting his quasi-pictures. If this is nothing supernatural, however, but just more brain activity, more electrochemistry (as Kosslyn clearly holds), then it is not at all clear that it helps us to understand how we could be *conscious* of a quasi-picture (or, come to that, a mentalese description) in our brain. (Indeed, Kosslyn does not claim that it does.)

Some philosophers do hold that consciousness of mental states arises from their being themselves represented within the brain by 'higher-order thoughts', and a 'mind's eye' account of image consciousness would presumably fit this mold. However, the higher-order thought theory has not gained widespread acceptance for a number of reasons (Lycan, 2000). It probably captures not so much the distinction between conscious and nonconscious states as that between those experiences that flit through consciousness unremarked, and those that *are* remarked, and perhaps thereby become remembered and/or reportable. Furthermore, it relies on the availability of a suitable account of the intentionality of the states in question, which may not be forthcoming. (*See* **Consciousness and Higher-order Thought**)

It might be objected that this is all just a special case of the notorious, unsolved 'hard problem' of consciousness: we just do not know how to explain conscious experience scientifically, so why should it especially concern us that we cannot explain it in the context of imagery? But although we can perhaps legitimately pass over the 'hard problem' in some areas of cognitive theory, we cannot, in good faith, ignore it here. It is of the very nature of imagery to be conscious (or, at the very least, potentially conscious), and, very arguably (Külpe notwithstanding), *all* conscious experiences are imagistic, either perceptual or quasi-perceptual. The apparent intractability of the 'hard problem' just may owe something to entrenched misconceptions about the nature of imagery. (*See* **Consciousness, Philosophical Issues about**)

References

Anderson JR (1978) Arguments concerning representations for mental imagery. *Psychological Review* **85**: 249–277.

Barsalou LW (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* **22**: 577–660.

Block N (ed.) (1981a) *Imagery*. Cambridge, MA: MIT Press.

Block N (ed.) (1981b) *Readings in Philosophy of Psychology*, **vol. 2**. London: Methuen.

Brann ETH (1991) *The World of the Imagination: Sum and Substance*. Savage, MD: Rowman & Littlefield.

Casey ES (1976) *Imagining: A Phenomenological Study*. Bloomington, IN: Indiana University Press.

Cummins R (1997) The LOT of the causal theory of mental content. *Journal of Philosophy* **94**: 535–542.

Descartes R (1664) *L'Homme*. (English translation by TS Hall. Cambridge, MA: Harvard University Press, 1972.

Ellis RD (1995) *Questioning Consciousness: The Interplay of Imagery, Cognition and Emotion in the Human Brain*. Amsterdam: John Benjamins.

Fodor JA (1975) *The Language of Thought*. New York, NY: Crowell.
Hannay A (1971) *Mental Images – A Defense*. London: Allen & Unwin.
Heil J (1998) *Philosophy of Mind*. London: Routledge.
Holt RR (1964) Imagery: the return of the ostracized. *American Psychologist* **19**: 254–266.
Kaufmann G (1980) *Imagery, Language and Cognition*. Oslo: Universitetsforlaget.
Kosslyn SM (1980) *Image and Mind*. Cambridge, MA: Harvard University Press.
Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.
Kosslyn SM and Pomerantz JR (1977) Imagery, propositions and the form of internal representations. *Cognitive Psychology* **9**: 52–76.
Kosslyn SM, Thompson WL, Kim IJ and Alpert NM (1995) Topographical representation of mental images in primary visual cortex. *Nature* **378**: 496–498.
Lakoff G and Johnson M (1999) *Philosophy in the Flesh*. New York, NY: Basic Books.
Landy MS, Maloney LT and Pavel M (eds) (1996) *Exploratory Vision: The Active Eye*. New York: Springer-Verlag.
Lycan W (2000) Representational theories of consciousness. In Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2001 edn). [http://plato.stanford.edu/archives/win2001/entries/consciousness-representational/]
Mangan B (2001) Sensation's ghost: the non-sensory 'fringe' of consciousness. *Psyche* **7** [Online serial: http://psyche.cs.monash.edu.au/v7/psyche-7-18-mangan.html].
O'Regan JK and Noë A (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* **24**.
Paivio A (1971) *Imagery and Verbal Processes*. New York: Holt, Rinehart & Winston.
Price HH (1953) *Thinking and Experience*. London: Hutchinson.
Pylyshyn ZW (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin* **80**: 1–25.
Pylyshyn ZW (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **88**: 16–45.
Pylyshyn ZW (in press) Mental imagery: in search of a theory. *Behavioral and Brain Sciences*.
Rollins M (1989) *Mental Imagery: On the Limits of Cognitive Science*. New Haven, CT: Yale University Press.
Ryle G (1949) *The Concept of Mind*. London: Hutchinson.
Sartre J-P (1948) *The Psychology of Imagination*. New York, NY: Philosophical Library. [original French Publication 1940.]
Shepard RN and Cooper L (1982) *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.
Slezak P (1995) The 'philosophical' case against visual imagery. In: Slezak P, Caelli T and Clark R (eds) *Perspectives on Cognitive Science*, pp. 237–271. Norwood, NJ: Ablex.
Thomas NJT (1989) Experience and theory as determinants of attitudes toward mental representation: the case of Knight Dunlap and the vanishing images of J. B. Watson. *American Journal of Psychology* **102**: 395–412.
Thomas NJT (1997a) A stimulus to the imagination. *Psyche* **3**. [Online serial: http://psyche.cs.monash.edu.au/v3/psyche-3-04-thomas.html]
Thomas NJT (1997b) Imagery and the coherence of imagination: a critique of White. *Journal of Philosophical Research* **22**: 95–127.
Thomas NJT (1999) Are theories of imagery theories of imagination? An *active perception* approach to conscious mental content. *Cognitive Science* **23**: 207–245.
Tye M (1991) *The Imagery Debate*. Cambridge, MA: MIT Press.
White AR (1990) *The Language of Imagination*. Oxford, UK: Blackwell.
Yates FA (1966) *The Art of Memory*. London: Routledge & Kegan Paul.

Further Reading

Kosslyn SM (1983) *Ghosts in the Mind's Machine: Creating and Using Images in the Brain*. New York, NY: Norton.
Morris PE and Hampson PJ (1983) *Imagery and Consciousness*. London, UK: Academic Press.
Neisser U (1976) *Cognition and Reality*. San Francisco, CA: W.H. Freeman.
Newton N (1982) Experience and imagery. *Southern Journal of Philosophy* **21**: 475–487.
Nicholas JM (ed.) (1977) *Images, Perception and Knowledge (Western Ontario Studies in the Philosophy of Science 8)*. Dordrecht/Boston MA: Reidel.
Paivio A (1986) *Mental Representations: A Dual Coding Approach*. New York, NY: Oxford University Press.
Richardson JTE (1999) *Mental Imagery*. Hove, UK: Psychology Press.
Sartre J-P (1962) *Imagination: A Psychological Critique* (trans F. Williams). Ann Arbor, MI: University of Michigan Press. (original French, 1936.)
Sheikh AA (ed.) (1983) *Imagery: Current Theory, Research, and Application*. New York, NY: Wiley.
Thomas NJT (2001) Mental Imagery. In: Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy*. [http://plato.stanford.edu/entries/mental-imagery/]

Metaphysics

Intermediate article

Brie Gertler, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Metaphysics aims to determine the basic nature of entities, states, properties, and events. A metaphysics of cognition will specify the logical relationship between mental and physical entities, and will account for the representational power of brain states.

INTRODUCTION

There are several longstanding metaphysical puzzles about the mind. How can a bit of gray matter be conscious? Are sensations and thoughts physical features of humans? What conditions must a physical state meet in order to be a state of a person? How can brain states represent facts and possibilities, as is required for genuine cognition? No account of the mind will be complete without a resolution of abstract questions such as these. Controversies over the basic nature of the mind and mental representation bear on the broader theories of mind to which research in cognitive science contributes.

WHAT IS METAPHYSICS?

The task of metaphysics is to determine the fundamental nature of reality. What is the best analysis of causation? How can we make sense of the difference between what is impossible and what is possible (but non-actual)? Do ordinary objects genuinely persist through time? If so, what are the identity conditions of objects? Of persons?

We can understand the distinctive purpose of metaphysics by contrasting it with another discipline also concerned with the ultimate nature of reality, namely, physics. Because there is no clear, uncontroversial dividing line between metaphysics and physics, the distinction must be somewhat idealized.

Metaphysical issues transcend the domain of physics in two ways. Firstly, while physics is concerned with the physical basis of reality, metaphysics is concerned with its logical basis. This contrast may be illustrated by the account each discipline gives of shape properties. Consider an ordinary cylindrical coffee cup. Physics will tell us how the cup's physical features – such as the arrangement of its component atoms – determine its shape, solidity, and so on. Metaphysics will specify the factors logically responsible for the cup's having these properties. For instance, some metaphysicians claim that to be cylindrical just is to participate in an abstract entity, 'cylindricality'; others deny that such abstract entities exist, and maintain that to be cylindrical just is to stand in a similarity relation to other concrete (cylindrical) objects; still others deny that there is anything truly cylindrical, since what is present isn't a cup but just a collection of atoms in a particular arrangement. (According to this last view, the belief in cups is merely a 'useful fiction'.) Metaphysics aims to explain what instantiating a property consists in, regardless of the particular physical laws that causally determine property instantiations. For instance, the basic metaphysical account of shape properties will not depend on particular physical laws. In this way, metaphysical claims transcend descriptive claims about the actual state of the universe.

The second way in which metaphysics transcends disputes within physics derives from the first. Since metaphysics is concerned with the logical basis of reality, its methodology is *a priori*. By contrast, the methodology of even theoretical physics is partially empirical or *a posteriori*. (Obviously, physics and other sciences also use nonempirical methods of inquiry, including mathematical reasoning.)

Some philosophers are suspicious of the metaphysical enterprise. Linguistically-minded philosophers maintain that there are no facts of the sort metaphysics seeks, independent of human conventions or conceptual schemes. These philosophers deride putative metaphysical questions as 'pseudo-questions'. Empirically-minded philosophers have questioned the legitimacy of *a priori*

methods of inquiry, and have claimed that genuine knowledge can be gained only through empirical research. Nevertheless, metaphysics continues to be a central and thriving philosophical discipline.

ISSUES IN THE METAPHYSICS OF MIND

Debates in the metaphysics of mind center on three topics: ontology, intentionality, and personhood.

Ontology concerns what sorts of things, properties, states, events, and relations exist or are instantiated, which of these are reducible to others, and which, if any, are irreducible and, hence, basic. Mental ontology thus concerns the mental's fundamental nature and its relation to other ontological categories, most saliently the physical. The famous 'mind–body problem', the problem of specifying the relation between the mental and the physical, is the main problem in mental ontology. (*See* **Mind–Body Problem**)

Theories of intentionality will explain how states have representational power. Since representational power is possessed by a wide variety of items (paintings, thermometers, the dances of bees) this problem does not exclusively concern the mental. A theory of mental intentionality will explain what relation obtains between my current belief that there are nine planets in our solar system, and that actual solar system, by virtue of which the belief is about the solar system; it will explain what relation obtains between my current desire to drink coffee, and the possible event of my drinking coffee, by virtue of which that event satisfies the desire; and so on.

The main questions about personhood are the following. What features of a state renders it the state of a person, as opposed to, say, the state of a mere machine? What features of a process render it something a person does, like making inferences, rather than something that merely occurs within a person, like digestion? Because persons and their actions are usually taken to be morally significant, answers to these questions have consequences for moral theories.

These three topics are interconnected; but for simplicity we will address them separately.

Mental Ontology

The main ontological dispute about the mind is whether mental states and properties are physical, or whether instead the mental constitutes an ontological category distinct from the physical. The first view is called 'physicalism' (or 'materialism'); the second view is called 'dualism', since it is usually further assumed that the physical and the mental are the only basic ontological types. Within each of these views, there are disputes about the precise nature of the mental, and about the relation between particular mental states or properties and particular physical states or properties.

Materialism is often construed as a negative thesis: the denial that there are souls or what one author has dubbed 'spooky mind-stuff'. This denial is not adequate to distinguish materialism from contemporary, naturalistic forms of dualism, which need not rely on a religious conception of the self and which typically deny the existence of 'spooky' things or properties, immune to scientific explanation.

Materialism is an ontological view, not a methodological one. But it is hard to formulate a good definition of physicalism, one which is neither too restrictive nor too liberal. Defining physical entities as those which are similar to entities currently posited by empirical science is too restrictive, as it illegitimately constrains future scientific discoveries. And defining physical entities as those discovered through scientific methods is too liberal as it begs the question against scientific naturalist forms of dualism. This difficulty besets dualism as well. Dualism is the view that there are mental things or properties that are not physical, so dualism can be defined only by reference to a clear notion of the physical. The dispute about materialism and dualism continues to generate lively discussion, and it certainly seems as if there is something at issue here, if only the question of whether the mental is fundamentally, and importantly, dissimilar from the physical. (*See* **Materialism**; **Dualism**)

Another basic ontological dispute about the mind concerns the reducibility of mental states or properties to non-mental states or properties. According to physicalist reductionism, mental states are identical to physical states. While dualists will deny reducibility, not all materialists believe that particular mental properties are reducible to particular physical properties. For example, consider the relation between biological properties, such as having a heart, and physiochemical properties. Is a biological feature identical to a particular collection of physiochemical features? If not, is the latter causally sufficient for the former? Are there interesting explanatory relations between biological and physiochemical features, or are these correlations basic and inexplicable? Reductionism is often advanced on grounds of parsimony. But it also has epistemic advantages, for reductionist

programs seek to understand one class of properties (e.g. thoughts) by another class which is purportedly better understood (e.g. neurobiological properties). It is hoped that a reduction of theories will follow a reduction of properties. For instance, neurobiological theory (or, perhaps, basic physics) may explain the predictive power of psychological theory. (*See* **Reduction**)

Mental Intentionality

The power to represent a variety of things – from particular items, such as a coffee cup, to abstract claims, such as 'justice is blind' – is a remarkable characteristic of the mental. Brentano (1838–1917) went so far as to consider intentionality the mark of the mental, claiming that all genuinely mental states are intentional and all genuinely intentional states are mental. Some present-day philosophers accept one or other of these claims, though few accept both. Still, the persistence of each of these claims, and the fact that their conjunction was once widely accepted, attest to the importance of a theory of mental intentionality in a larger account of the mind.

Most philosophers now believe that mental representation is continuous with other phenomena in the natural world. According to this naturalist viewpoint, the sorts of facts and laws that will explain intentionality will be the same as, or similar to, those that explain non-intentional phenomena. Current theories of intentionality are largely attempts to vindicate this claim by 'naturalizing' intentionality.

The nature of mental representation is highly controversial in several respects. The first of these has to do with the standard distinction between phenomenal and intentional features of states. A state's phenomenal features consist in 'what it's like' to instantiate that state. Pains and tickles are examples of states with phenomenal features: the painfulness or ticklish quality is the state's phenomenal 'content'. A state's intentional features are its representational properties. Beliefs, desires, and perceptual states are examples of states with intentional features. The intentional content of such a state is what it represents. For example, the intentional content of a belief that justice is blind is *that justice is blind*; the intentional content of a desire to bicycle is the activity of bicycling. (*See* **Qualia**)

Historically, some philosophers have treated intentional content as a species of phenomenal content, but this view is now out of favor. A more popular view, 'representationalism', is that phenomenal content is reducible to intentional content. Another view is that these are distinct but that all (nonderivatively) intentional states have phenomenal content, or that all phenomenal states have intentional content. Yet another view is that states with phenomenal content overlap with intentional states only partially, if at all.

The second dispute about the nature of content is more ontologically oriented. Suppose I am wondering whether my coffee cup is in the sink. Is the content of this mental state a (possible or actual) concrete state of affairs (that is, the cup's being in the sink)? Is it an abstract proposition, 'the cup is in the sink', perhaps on a par with 'justice is blind'? Is it a relation between me and a state of affairs? Between me and an abstract proposition? Or is it a hybrid of these – for example, do I represent the concrete state of affairs partially in virtue of standing in an intentional relation to an abstract proposition?

There is currently active debate over a third question concerning the nature of content, namely, whether content logically depends on one's external environment, including social and linguistic facts, or whether external factors are at most causally influential. Content externalism is the view that, holding fixed my 'internal' state, the content of my thought 'Is my coffee cup in the sink?' can vary with variations in physical facts (about the cup or sink), or with variations in social or linguistic facts (e.g. about how my linguistic community uses 'cup'). While content internalists agree that external facts causally influence my thought content, they deny that having a thought with a certain content consists in any such facts. Instead, they hold that all mental content is narrow content; that is, states internal to the subject fix the subject's mental contents. The externalism–internalism dispute is closely tied to the dispute over 'methodological solipsism' (Fodor, 1980), the view that we need make no reference to anything outside the subject's head in order to explain the subject's behavior. (*See* **Narrow Content**)

These disputes take place within a realist view of mental content. But some philosophers deny that mental states are genuinely intentional, and contend that talk of mental content is at best a useful fiction. Among antirealists, 'eliminativists' (e.g. Churchland, 1981) think that talk of beliefs and desires, and their associated contents, is a residue of a naive and mistaken folk theory, and that scientific (specifically, neuroscientific) advances will warrant abandoning intentional notions, just as scientific advances warranted abandoning the notion of phlogiston. By contrast, 'instrumentalists'

(e.g. Dennett, 1987) deny the existence of intentional states but regard intentional notions as pragmatically indispensable. (*See* **Eliminativism**; **Intentional Stance**)

Because of the plurality of views about the nature of mental content, theories of mental content differ in their explanatory aims. Generally speaking, however, the chief task of a theory of content is to explain what having a state with a certain content consists in. Providing a causal etiology of mental content is at most a secondary concern of such theories. (*See* **Intentionality**)

Personhood

Issues about personhood are closely tied to ontology and intentionality. But while ontology concerns the mental and intentionality concerns the representational, the notion of a person is an ethical notion. To be a person is to warrant ethical consideration and to be a possible bearer of moral responsibility. Since moral responsibility is generally taken to apply to free actions, theories of personhood relate to theories of action and free will. (*See* **Action, Philosophical Issues about**; **Free Will**)

What makes a mental state the state of a *person*? There are two aspects to this question. The first concerns how we distinguish persons from non-persons. Is the capacity for thought the essence of personhood? Must one also be capable of free agency? The second aspect concerns the intuitive contrast between what a person does and what processes occur within him or her. For instance, we ordinarily say that persons wonder, or engage in inference, while digestion is something that occurs within them (or their bodies). The former sorts of processes are said to occur at the personal level, while the latter are at the sub-personal level. Which sorts of processes occur at the personal level, and how can we best analyze the difference between the personal and sub-personal levels?

VIEWS ON THE METAPHYSICS OF MIND

Views on Mental Ontology

Materialism is the dominant view among philosophers and cognitive scientists; and it may appear that materialism is the only reasonable position consistent with a scientific outlook. But strong naturalistically-oriented arguments have been made in favor of dualism, and dualism is still taken seriously in the philosophy of mind.

The simplest forms of materialism are 'type identity' theories. These theories identify a type of mental state – say, the type 'coffee desire' – with a type of physical state – say, the release of a certain neurotransmitter *N*. This is a strong claim. The claim is not just that the release of *N* causes coffee desires, or that coffee desires accompany the release of *N*: these are claims of causation or correlation. Type identity theories state that the release of *N* simply *is* the coffee desire: the mental type and the physical type are one and the same.

While identity claims are not merely claims of correlation, they are committed to the necessary, complete overlap of their relata. They are therefore undermined by the possibility that one relatum could be present in the absence of the other. The 'multiple realizability' objection to type identity theories claims that mental state types such as 'coffee desire' can be realized in different physical types. Even if the release of *N* is essential to coffee desires in humans, it is conceivable that organically different beings that lack neurotransmitter *N* (e.g., silicon-based creatures) could desire coffee. (*See* **Multiple Realizability**)

In light of this objection, many philosophers have rejected type identity theories in favor of 'token identity' theories. These theories identify a particular ('token') mental state, such as the coffee desire I have now, with a particular physical state, such as the current release of *N* in my brain. A silicon-based creature's token coffee desire may similarly be identical to a state of her silicon brain.

One influential token identity theory construes mental tokens as states which play a particular functional role in the larger cognitive system. On this view, called 'metaphysical functionalism', my coffee desire and the robot's coffee desire qualify as the same mental type because they share typical causes (e.g. a background fondness for coffee, a sluggish feeling) and effects (coffee-seeking behavior). (*See* **Functionalism**)

Some arguments for dualism make use of the idea that our epistemic access to certain mental states is irreconcilable with a physicalist account of those states. For instance, Nagel (1974) argues that even if we understand the physical workings of a bat's echo-location system, we still do not know the phenomenal aspect of the bat's experiences; that is, we do not know 'what it's like' to be a bat. Other dualist arguments are based directly on the conceivability of certain scenarios. For example, Chalmers (1996) argues that we can conceive of a 'zombie', a creature that shares all of our physical states but is entirely devoid of phenomenal experience. These arguments support the idea that there

is what Levine (1983) has dubbed an 'explanatory gap' between physical and phenomenal features. According to this idea, even a complete account of the world in physical terms will fail to explain the presence of phenomenal features. Most present-day dualists accept a dualism about properties rather than about substances. Property dualism holds that some mental properties – usually, phenomenal properties – are irreducible to any physical properties. (*See* **Knowledge Argument, The**; **Explanatory Gap**)

The deepest difficulty facing dualism concerns mental causation. Mental states obviously causally interact with physical states: a desire for coffee causally contributes to my heading for the kitchen; drinking coffee causes a pleasurable sensation; and so on. But (as Princess Elizabeth of Bohemia complained to Descartes in 1643) it is not clear how the nonphysical could causally interact with the physical (Wilson, 1969, p. 373). Various solutions to this problem have been proposed. Some dualists whose dualism is limited to phenomenality avoid the problem by denying that phenomenal states enter into causal relations. (*See* **Epiphenomenalism**)

Among physicalist views, representationalism is especially relevant to intentionality. Representationalism claims that representational content is what is crucial to a state's identity. Representationalism is most significant as a theory about phenomenal content: it entails that the felt quality of a pain, say, is exhausted by what the pain represents. In general, materialism seems to have more difficulty accommodating phenomenal content than intentional content; by assimilating phenomenal content into intentional content, representationalism seeks to reduce problems about the ontology of qualia to more tractable problems about the ontology of representational content, or intentionality.

Finally, accounts of mental ontology encompass a range of positions with regard to reductionism. Some presume that mental properties are reducible to neurobiological processes; some add that neurobiological processes are themselves ultimately reducible to physical properties. But there are several objections to the claim that the mental is reducible to the physical. (*See* **Anomalous Monism**; **Multiple Realizability**; **Emergence**)

Views on Mental Intentionality

Most currently accepted theories of intentional content are causal theories. One causal theory of intentionality is the 'structural isomorphism' view. On this view, a thinker represents a set of objects if and only if the causal patterns between the thinker's states are structurally isomorphic to the causal patterns among the objects (Cummins, 1996).

The view of content that is most salient for cognitive science is also the most currently influential. This view, 'content functionalism', individuates mental state types by their causal roles *vis-à-vis* other mental states, stimuli, and behavior. For instance, the belief that there is coffee nearby is that state typically caused by an olfactory experience of freshly brewed coffee, or seeing a sign reading 'café open' (and so on), and which – when accompanied by a coffee desire – typically causes one to engage in coffee-seeking behavior. Functionalism is promising for computational models of mind, which construe cognition as information processing and individuate cognitive states by their roles as causal mediators between input (stimuli, other states) and output (other states, behavior). (*See* **Functionalism**)

Views on Personhood

The central question about personhood is: what distinguishes a person from a non-person? Some philosophers believe that this distinction is sharp, while others allow that there may be degrees of personhood. For instance, some causal views of cognition associate personhood with complexity of the causal (cognitive) network, and thus allow that more complex networks have a higher degree of personhood than less complex ones.

Some of the factors that are thought to be essential to personhood are: the ability to think; the capacity for self-consciousness and belief revision; and the capacity for phenomenal states. Each of these is, of course, subject to further scrutiny. The issue of what constitutes thinking is itself difficult. Some humans suffer impairments which arguably leave them incapable of self-reflection, but they are considered morally significant beings nonetheless. And lower animals seem to experience sensations, but are not ordinarily treated as persons.

There is a similar range of views concerning the distinction between personal and sub-personal processes. Some maintain that it is intrinsic to thinking, or to pain, that these occur at the personal level. Others contend that this is a relational feature of these processes: for example, a matter of one's being aware of them, or of their occurring above a certain threshold of consciousness, or of our being free to determine their outcomes. As with the requirements for personhood, the distinction between the personal and the sub-personal levels may be sharp, or it may admit of degrees.

THE METAPHYSICS OF COGNITIVE SCIENCE

Broadly speaking, cognitive science is the attempt to understand the human mind as an information processing system. Cognitive scientists use computational models of cognition to explain how humans acquire, store, and manipulate information, and how this information guides their behavior. The basic assumption of computational models is that humans process information by cognitive operations on representations. For this reason, philosophical accounts of intentionality are particularly important to cognitive science. Ontological concerns are also important, as background motivation for the enterprise of cognitive science. Since artificial computers are physical systems, a computational account of human cognition could solve one aspect of the mind–body problem by showing how thought can be accomplished in a purely physical system.

Cognitive Science and Ontology

One ontological problem about cognition is the difficulty of conceiving how a small bit of gray matter can be responsible for human activities (navigating one's way through unfamiliar terrain, predicting others' actions, carrying out long-range plans, etc.) and human thought (from reflecting on the nature of beauty to speculating about future stock market behavior). Cognitive science approaches this difficulty with a 'divide and conquer' strategy. It seeks to divide complex tasks into simpler ones, and conquer these simpler ones by showing that they are analogous (or, perhaps, identical) to computational processes.

Dividing complex cognitive operations into simpler ones seems to resolve the complexity issue, but it is less clear that an analogy to computational processes resolves the basic ontological issue. Even supposing that cognitive science breaks complex cognitive operations into smaller processes that are relevantly analogous to computational processes, there remains the question: how does this physical and computational process realize a mental process?

Cognitive science is not explicitly concerned with the question of materialism: most cognitive scientists assume materialism from the outset. But ontological issues, including the objections to materialism that focus on qualia, are nonetheless important. There are several positions cognitive scientists can take regarding qualia. Firstly, they might simply deny that qualia are the concern of cognitive science. But since some representational states, like perceptual states, clearly have qualitative properties, this response commits them to the contentious claim that the representational features of such a state can be fully explained independently of its phenomenal features. A second option is to adopt representationalism, that is, to claim that the phenomenality of a state is exhausted by its representational content; representational content would then be explained computationally. This response has the advantage of comprehensiveness, but it also opens the account of qualia to arguments against representationalism, including the anti-materialist arguments described above. A third option is to hold that phenomenal content is irreducible to representational content, and to offer an explanation of how phenomenal content contributes to representational content. This third strategy, perhaps the most ambitious, is especially challenging given that many philosophers take phenomenal content to be an intrinsic property of a state, while computationalist explanations of representational content exploit the relational nature of representational properties.

Cognitive Science and Intentionality

Cognitive scientists generally account for the intentionality of a cognitive process by finding a computational analogue for the process. As with the ontological issue, the method is to resolve questions about how human cognition operates by analogy with the better-understood intentional features of computational processes.

A common objection to this approach to understanding cognitive intentionality is expressed in John Searle's famous 'Chinese room' argument (Searle, 1980), which aims to show that what computational features provide is not sufficient for semantics (meaningfulness). Searle concludes that the central project of cognitive science – understanding cognitive intentionality by analogy with well-understood computational processes – is misguided. (*See* **Chinese Room Argument, The**)

Objections to the computationalist approach to intentionality also arise from connectionism. Computationalism assumes that the entities (symbols) that have intentional content are also the entities involved in the computation. That is, it assumes that intentional content occurs at the computational level. Connectionism denies this, claiming that computation occurs at a subsymbolic level and that intentional content is a higher-order property

of patterns of lower-order computations. One purported advantage of connectionism is that it allows us to treat intentional processes as higher-level algorithms while exploiting lower-level, neuroscientific data in accounts of computational processes. (*See* **Connectionism**)

But this purported advantage is also a potential problem for connectionism. If the causal, computational relations occur at a lower level than the intentional processes, then it seems wrong to say that I conclude that there is coffee in the kitchen *because* I smell coffee; that is, because my cognitive system contains an olfactory representation of coffee. Computationalism uses the meanings of symbols upon which computations operate to partially explain why the system engages in those computations. On the connectionist model, the emergence of higher-order meaningful symbols from lower-order non-intentional computations appears to be inexplicable.

Cognitive Science and Personhood

Cognitive science is strictly neutral about the requirements for ethical personhood. Still, since most believe that ethical status is closely tied to mentality, the cognitive science approach to the mind is likely to encourage a causal view of personhood. Computationalism may lend itself to a straightforward causal view, according to which a particular state's causal role determines whether it is a state of a person. Connectionism may lend itself to a more complicated causal view, according to which personhood is a higher-order (perhaps emergent) property of a larger causal network of states. For parallel reasons, those sympathetic to Searle's objection to computationalism may embrace an intrinsic view of personhood.

Cognitive science's general methodology, of breaking cognitive tasks into progressively simpler tasks, may help to distinguish between processes at the personal level and those at the sub-personal level. For it may be that the level of complexity correlates with the personhood distinction, so that (conscious) thought occurs at the personal level, while each subtask that contributes to thought occurs at a sub-personal level.

References

Chalmers D (1996) *The Conscious Mind*. Oxford, UK: Oxford University Press.

Churchland PM (1981) Eliminative materialism and propositional attitudes. *Journal of Philosophy* **78**: 67–90.

Cummins R (1996) *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Fodor J (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* **3**: 63–72.

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **83**: 435–450.

Searle J (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* **3**: 417–424.

Wilson M (1969) *The Essential Descartes*. New York, NY: New American Library.

Further Reading

Heil J (1998) *Philosophy of Mind: A Contemporary Introduction*. London, UK: Routledge.

Horst S (1996) *Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind*. Berkeley, CA: University of California Press.

Kim J (1998) *Mind in the Physical World: An Essay on the Mind–Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Lycan WG (ed.) (1990) *Mind and Cognition: A Reader*. Oxford, UK: Blackwell.

Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Rosenthal D (ed.) (1991) *The Nature of Mind*. Oxford, UK: Oxford University Press.

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Shoemaker S (1984) *Identity, Cause, and Mind*. Cambridge, UK: Cambridge University Press.

Shoemaker S and Swinburne R (1984) *Personal Identity*. Oxford, UK: Blackwell.

Strawson G (1994) *Mental Reality*. Cambridge, MA: MIT Press.

Tye M (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

Warner R and Szubka T (eds) (1994) *The Mind–Body Problem: A Guide to the Current Debate*. Oxford, UK: Blackwell.

Mind–Body Problem

Introductory article

Robert Van Gulick, Syracuse University, Syracuse, New York, USA
Jason Clark, Syracuse University, Syracuse, New York, USA

CONTENTS

The mind–body problem is the problem of explaining the relation between the mental and the physical, especially the question of whether the mind is something over and above the physical or just a special part of it.

INTRODUCTION

Dualism divides reality into mutually exclusive mental and physical domains. This may involve a duality of substances, as in traditional Cartesian dualism, or one of properties, a more modern perspective. Substance dualists treat minds as thinking things that are distinct from material objects, including the animal bodies with which they may be causally linked; it is an intuitive notion that finds expression in the ancient belief of self as soul. Property dualists regard minds as physical substances (perhaps embodied brains) that none the less have mental properties over and above their physical properties. If the fiery taste of which I am aware when I bite into a chili is not identical with any physical property of my brain, then reality would seem to extend beyond the physical.

Physicalists hold that the physical exhausts reality. They claim that everything real is in some fundamental sense physical, though they differ about just what that sense might be. Thus if minds are real they must exist as parts of the physical world, and they can have no real properties other than those that are physical in the relevant sense. The physicalist must thus address two sets of questions. First, do minds really have such features as consciousness, meaningfulness or free will? Some physicalists would deny the reality of certain of these, just as modern science denies the existence of ghosts and phlogiston and feels no need to fit them within the physical. Second, if some mental feature is accepted as real, how could it be physical (in whatever sense the theory requires, whether that be a matter of identity, material composition or dependence)?

EVOLUTION OF THE PROBLEM

Questions about the place of mind in nature go back to our earliest records of human thought, but the specific problem of the relation between mind and matter came to the fore with the rise of modern physical science in the seventeenth century. The newly-found ability to model quantitative physical relations gave rise to the so called mechanical philosophy with its grand explanatory ambitions. Inevitably the question arose of whether minds could be mechanically explained. This question found its sharpest formulation in the philosophy of René Descartes. Although he was a major contributor to the mathematical and physical theories of his day, he argued that mind and matter are distinct kinds of substance, each with its own essence: minds as thinking things and matter as spatially extended. Despite their different natures, he accepted their mutual causal interaction: what happens in the mind affects the body and vice versa.

Descartes' view quickly attracted criticism. The early materialist Thomas Hobbes argued for a surprisingly computer-like view of minds as physical systems for manipulating symbols. Other critics accepted Descartes' dualism but denied that mind and body interact, despite appearing to do so. Gottfried Leibniz, the German philosopher and co-inventor of the differential calculus, believed in a divinely pre-established harmony that made minds and bodies run in parallel, like the many clocks in a shop striking the hour together.

For the next three centuries the dualists held their own against the materialists, and even at the beginning of the twentieth century the majority of thinkers still had a basically dualist outlook. However, in the twentieth century, physicalism became the dominant viewpoint among scientists and philosophers, in part following advances in neuroscience, computing and biochemistry. If life and reproduction are biochemical and ultimately

physical in nature, should we not expect the same of mind and thought?

CONTEMPORARY POSITIONS

While serious arguments for and about dualism are still heard, contemporary discussion more often concerns how best to formulate physicalism. Physicalists disagree as strongly among themselves as they do with their dualist opponents. Any physicalist theory needs to answer two related but distinct questions: what counts as physical in the strict or primary sense; and how must something be related to the strictly physical to be counted as physical in the extended sense which is supposed to apply to minds? Both questions have received diverse answers.

Regarding the first question, the physicists' inventory of fundamental features has changed again and again over the centuries. Beginning with a simple mechanical view of matter in motion, Newton's gravity was added, and then electromagnetic forces and fields. Physicists have since introduced the strong and weak forces, gluons, and more recently strings in high-dimensional space. Features regarded as 'fundamental' may be added or dropped. Given this theoretical flux, most physicalists leave the exact nature of the strictly physical as an open variable. They claim simply that whatever minimum of basic features suffices to account for nonmental physical reality will account also for minds.

The second question is more controversial. Physicists don't include earthquakes, rockets or cockroaches in their list of basic physical features, yet few would deny that they fall within the physical realm. The physicalist must first explain what relation to the strictly physical makes them count as physical, and then show that minds meet that same criterion.

Contemporary physicalism was first formulated in the 1950s as a claim of strict identity between mental and physical properties. Analogies were drawn with familiar cases of scientifically discovered theoretical identities: lightning is just an atmospheric discharge of static electricity; water is just a collection of H_2O molecules. Similarly, it was proposed, consciousness, beliefs and pains might be just certain sorts of neural processes in the brain. This view is called the type identity theory, since it aims to identify mental properties or types with specific physical counterparts. It does not assert merely that, for example, pains are correlated with certain brains states, or that the latter cause the former, but that the two are one and the same thing viewed from different theoretical perspectives, just as water and H_2O are identical.

A relation of property identity would certainly validate physicalism; but it seems too strong to be true, as one can see from the 'multiple realization' objection. The identity theorist identifies being in pain with some specific neural property, for example, increased activity in the anterior cingulate cortex. But that links the mental property too specifically with the particular neural correlate of pain in humans. If we were to find no such neural processes in birds, monkeys or extraterrestials, it would not follow that they don't feel pain, especially if they respond in pain-like ways to injury. The type identity theory seems even less plausible for more abstract states such as believing, perceiving or wanting. Indeed, it is hard to see why even an electronic robot without neurons might not qualify for having such states.

Thus most physicalists moved away from the type identity theory towards a view of mental properties as higher-level systemic properties based on an underlying physical substrate, more like functional categories such as being an amplifier or being a recessive gene for light eye colour. According to this functionalist view, what makes some state a pain, a belief or a desire is the functional role it plays within the system or organism of which it is a part. Pains, for example, typically indicate bodily damage, cause avoidance of the source of the pain, and interfere with concentration. To validate physicalism one would need to show that the things playing those roles were physical and filled the relevant roles by virtue of their physical causal profiles. To draw a biological analogy, there are many different structures that count as stomachs, but they all perform certain functions and do so because of their physical configuration. In that respect, according to the functionalist, minds are no different from stomachs, and thought no different from digestion.

If mental properties are higher-level functional properties of organized systems, then they are not identical with underlying strictly physical structural properties. The physicalist must find a relation between the two that is weaker than identity but still strong enough to validate the claim that everything real is ultimately physical. Consider again the biological analogy. Many of the things or properties to which biologists refer (e.g. organisms, reproductive fitness, or cellular immunity) are not strictly physical, in the sense that physicists do not refer to them in their theories or models. Yet given our current view of life, we accept such biological features as complex aspects of physical

reality. Can the physicalist specify a relation between mental and physical properties that would support the claim that minds are also just complex aspects of the physical? Various candidate relations have been offered, of which the three most prominent have been physical composition, supervenience and realization.

According to the 'composition' proposal, a thing counts as physical if it is entirely composed of physical parts or constituents. Being a grasshopper may not be a property studied by physicists, but if grasshoppers are composed entirely of parts that have strictly physical properties then grasshoppers would count as physical, as would all their higher-level properties. Being alive or ready to mate might not be strictly physical properties, but they would count as physical in the extended sense that they are properties of physically composed systems.

However, being physical in that sense may be too weak a condition. The fact that all the parts of a system are physical does not guarantee that all its properties are physical. Its parts might have nonphysical (perhaps protopsychic) properties as well as physical properties such as mass and charge: that is, they might be dual in nature. Indeed, some dualists, called panpsychists, believe just that. They claim that reality has both physical and mental aspects down to its lowest level. The view was proposed at the beginning of the dualism–physicalism debate by the Dutch philosopher Baruch Spinoza. Early in the twentieth century, the British philosopher and logician Bertrand Russell and the American philosopher and psychologist William James both proposed similar dual-aspect theories. More recent theorists have cited quantum mechanics as supporting their view that reality embodies a mental–physical duality at its most fundamental level; they appeal especially to those versions of the theory that accord a role to the consciousness of the observer in quantum-mechanical measurements, in which an indeterminate quantum state is collapsed to a determinate value, apparently by the measurement process itself. Although panpsychism and dual-aspect theories remain controversial minority views, they do show one way in which the composition relation might fail to validate physicalism. The requirement might be strengthened to avoid the problem by counting as physical only things that are composed entirely of parts that can be fully described in terms of their strictly physical properties. Such an addition would exclude the panpsychic possibility; but for that reason dual-aspect theorists would deny that everything is physically composed in that sense. They would demand that the physicalist prove any such claim and not merely assume it.

Moreover, even the strengthened requirement might fail to validate physicalism for a second reason. Perhaps a system as a whole could have properties that were emergent in the radical sense: that is, it could be more than the sum of its parts and their modes of combination. 'Emergent' dualists claim just that: they view mental properties as radically emergent nonphysical features of complex systems. Thus, even if all of a system's parts were physical in the strengthened sense, the system itself might still have nonphysical properties.

It is important to be clear about the various senses in which a system's properties might be termed 'emergent'. Some are quite modest and pose no problem for physicalism, while other more radical senses would entail dualism. Modest emergence is just the view that wholes can have properties quite unlike those of their proper parts. Thus a collection of H_2O molecules at room temperature is liquid even though none of the molecules themselves have that property. None the less, the liquidity of the collection is a necessary and understandable consequence of the physical properties of its molecular parts. The higher-level property of liquidity is not identical with any lower-level properties of its parts, but it depends upon them so intimately and necessarily that one can regard it as physical in a very strong sense.

Physicalism would conflict with a more radical form of emergence in which a system had properties that were not only distinct from those of its parts but that could not be explained by appeal to those properties or their modes of combination. In this case, the dualist might well regard the properties of the whole as something over and above the physical. Even if there were law-like connections between the two sets of properties, it would not follow that the system-level properties were themselves purely physical.

Physicalists and emergent dualists disagree about whether there are any cases of radical emergence. The matter can be settled only by the evidence, which is as yet incomplete. If the radical emergence of the mental remains a possibility, however remote, then physical composition, even in the strengthened sense, would not guarantee physicality. A system all of whose parts were physical might still have emergent nonphysical mental properties.

As an alternative to physical composition, some physicalists appeal to a supervenience relation between the mental and its physical substrate. Supervenience is the idea that a set of properties cannot

change without a corresponding change in another set of properties, even if the two are distinct and quite different in nature. Supervenience was first proposed as a way to think about evaluative properties such as moral and aesthetic ones. The beauty of a painting is not identical with any set of its physical properties, such as the distribution of different colored regions of paint on its canvas. None the less, if a second painting had exactly the same physical properties (including the same paint distribution), then the two paintings would have the same aesthetic properties. If the first were beautiful, so would be the second. Supervenience might be summed up in the slogan 'no difference without a physical difference': in particular, no two systems that are the same in all physical respects can differ mentally. The supervenience of the mental on the physical thus seemed to offer a way in which mental properties might be nonidentical with physical properties but still dependent upon them enough to qualify as physical as well.

However, like physical composition, supervenience runs into objections and seems too weak to validate physicalism. In particular, supervenience does not rule out various dualist possibilities. For example, it is consistent with dual-aspect theory as long as the properties of the two domains are reliably and invariably correlated, as they are in Spinoza's version of panpsychic dualism, in which mental and physical properties correspond as a matter of metaphysical necessity. Even if the connection were only a matter of natural law, it would guarantee that there could be no actual mental difference without a physical difference. Moreover, supervenience does not itself entail the sort of one-way dependence needed to support the claim that everything real is ultimately physical. In Spinoza's theory, for example, there is dependence in both directions: neither the physical nor the mental can vary without a corresponding change in the other. Given that supervenience is compatible with such interdependence, it cannot by itself sustain the claim that the physical is primary or ultimate.

A third relation by which physicalists have tried to link the mental and the physical is that of realization, which perhaps best fits the functional view of mental properties. With a functional property, such as being a 20-amp fuse or being a recessive gene for blonde hair, we can specify the required functional role and then show how an underlying physical structure fills it by means of its physical causal profile. For example, it is not that a given DNA sequence causes the gene; rather, that sequence is the gene in so far as it fills the role that qualifies it as a recessive hair color gene and does so because of its the lower-level physical properties. The relation easily accommodates the sort of multiple realization that refuted the earlier type identity theory. A single role might be filled by different physical structures in different people or different organisms; what makes them all pains, for example, need not be a common structure. Realization may thus provide a sufficiently intimate and necessary form of one-way dependence to validate physicalism, but only if all mental properties are in fact physically realized functional properties. Dualists and physicalists disagree on this point.

DIFFERENT MIND–BODY PROBLEMS

As noted above, the physicalist must accommodate as physical each mental feature acknowledged to be real. If, for example, experience really has felt aspects, like the red of which I seem to be aware when I look at a chili, then the physicalist must show how those aspects might be physical or physically realized. Different mind-body problems arise depending on which mental aspect one considers. In particular, important versions of the problem are generated by the difficulty of explaining consciousness, intentionality, and mental causation in physical terms.

Consciousness makes the mind–body problem deeply puzzling. Those aspects of mind that cluster around it (such as qualia (the felt character of experience) the subjectivity of awareness, and the first-person perspective) seem most difficult to fit within the physical domain. How could the conscious taste of a mango be identical with or realized by any physical process in my brain? The two sides of the psycho-physical equation seem so dissimilar that we are left in bafflement as to how they might fit together. It is not so much like identifying liquid water with H_2O as turning water into wine.

The second major variant concerns what philosophers call the intentionality of mind, the fact that mental states have meaning and refer to things in the world outside themselves. One's thought may be about the Eiffel Tower even when one is sitting in Nebraska, and a medical researcher's desire might be for a cure for Alzheimer disease even though none as yet exists. Minds have certain features that are undeniably representational: there could be no beliefs, desires or perceptions if they had no meaningful content. Thus the physicalist must show how brains as physical systems can have such representational powers and intentional content. What might it be about a brain state or process that makes it be about something, perhaps

something distant, or even nonexistent, like children's beliefs about Santa Claus? Many answers to these questions have been offered, and although no consensus has emerged, there are many promising options and active research continues.

A third variant concerns the problem of mental causation, a problem which itself takes many forms. The basic problem, going back at least to Descartes, is to account for the apparent causal interaction of the mental and the physical. I stub my toe and I feel a pain, or I want a sip of coffee so I reach out my arm and bring the cup to my lips. Cartesian dualists who wish to preserve mental causation must explain how two things so different in nature – spatially extended matter and nonspatial conscious mind – could be causally connected. Moreover, if nonphysical mental factors are able to bring about changes in the physical world, the conservation laws which are among the most empirically well confirmed and basic principles in modern science would be violated. The consequences seem equally drastic and unacceptable if the dualist opts to regard the mental as epiphenomenal and causally inert in the physical domain. Seemingly obvious causal facts, as well as moral responsibility, would be put in doubt if our mental states are not among the causes of our actions.

However, the causal status of the mental also poses a problem for many physicalists, as is shown by the so-called exclusion argument. If all physical events are fully determined by prior physical causes or circumstances, then it would seem impossible for mental properties to affect the course of physical events, unless they were identical with physical properties as the type identity theory claims. My desire for a sip of coffee seems to cause my arm to move, but all the real causal work is being done by brain processes and their physical properties. The fact that those properties, in doing so, might also fill an abstract mental role adds nothing to the causal powers they already had based on their physical nature. Thus the physicalist's current higher-level view of mental properties threatens to make them epiphenomenal.

Although these three variants have loomed largest in recent discussions, there are many other forms the mind–body problem might take, if for example one wished to account for the emotive aspect of mind or the alleged freedom of the will. In each case, the dynamic between the dualist and physicalist would be similar. Each would take a stand about the extent to which the relevant feature was real, and then try to show why and how it could or could not fit within the physical framework.

The mind–body problem is sometimes posed as the question of whether the mental reduces to the physical. However, the notion of reduction is ambiguous about what sorts of things are being reduced. It can be either a relation between objective things in the world, such as events and properties, or a relation between theories or ways of describing and conceptualizing the world. On the first (ontological) reading, it roughly corresponds to the core mind-body problem as discussed above. However, on the second (representational) reading, reduction concerns our ability to derive our mental theories or concepts from our physical ones.

Most contemporary physicalists accept ontological reduction but reject representational reduction, and on the basis of that rejection they are classed as nonreductive physicalists. They deny that the representational or conceptual resources of physical theory will suffice for all our legitimate explanatory and descriptive needs. The point seems obvious if one considers the case of a higher-level science such as economics. No one supposes that there is an ontological dualism of the physical and the economic; there is no metaphysical money-matter problem. But economics gives us a way of describing and explaining patterns in the world that could not be described or explained using only the resources of physics. Every act of paying ten dollars for a book, whether with cash, credit card or cheque, is an ontologically physical event, but what those events share is a higher-level feature that can only be understood relative to the whole organized financial system in which they occur. Some philosophers have criticised nonreductivism as 'physicalism on the cheap'; they argue that nothing short of identities or strict derivation will validate physicalism. The nonreductivist replies that it provides everything a reasonable physicalism requires, and that economists and psychologists can remain physicalists in the ontological sense while being strongly nonreductionist in the representational sense and accepting the validity and autonomy of their special sciences and the valuable ways they offer us of understanding our complex reality. Debate on the issue continues.

Further Reading

Block N, Flanagan O and Guzeldere G (eds) (1998) *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Churchland P (1988) *Matter and Consciousness*. Cambridge, MA: MIT Press.

Crumley J (2000) *Problems in Mind*. Mountain View, CA: Mayfield.

Dennett D (1996) *Kinds of Minds*. New York, NY: Basic Books.
Hasker W (1999) *The Emergent Self*. Ithaca, NY: Cornell University Press.
Hiel J and Mele A (eds) (1993) *Essays on Mental Causation*. Oxford, UK: Oxford University Press.
Humphrey N (2000) *How to Solve the Mind–Body Problem*. Thorverton, UK: Imprint Academic.
Kim J (1996) *Philosophy of Mind*. Boulder, CO: Westview Press.
Kim J (1999) *Mind in a Physical World*. Cambridge, MA: MIT Press.
Lycan W (ed.) (1999) *Mind and Cognition*, 2nd edn. Oxford, UK: Blackwell.
McGinn C (1999) *The Mysterious Flame*. New York, NY: Basic Books.

Modularity

Intermediate article

Ron McClamrock, University at Albany, State University of New York, Albany, New York, USA

CONTENTS

Various cognitive and perceptual processes may be separated or isolated from one another. Those processes may be understood as relatively independent subsystems, or modules.

WHAT IS MODULARITY?

A cognitive or perceptual process is said to be modular to the extent that it is an independent sub-process of the overall cognitive architecture. A module is a cognitive or perceptual subsystem whose workings are relatively independent from the rest of the cognitive architecture, and whose functioning can be analyzed and understood relatively independently of the overall system in which it is embedded. This idea is based on the more general engineering notion of 'near-decomposability' (Simon, 1981). According to Simon, a system is nearly decomposable if it is made of components whose behavior is, in most respects, not influenced by anything else going on in the system beyond what is given to the components as input.

Although the modularity of subsystems of perception and cognition has been discussed in various forms throughout the history of the philosophical study of mind, the view that psychological systems are to be understood as systems of interacting modules was particularly popular among the 'faculty psychologists' in the nineteenth century. These theorists typically held (apparently with little empirical justification) not only that there were independent modules governing various skills, abilities, and aspects of intelligence (e.g. a musical faculty, and a mathematical faculty), but that these modules had specific physical locations, and that the strength of a psychological faculty was correlated with the size of bumps on the head in the corresponding location.

The modern notion of modularity is expressed in terms of information processing. Here, the central idea is that various subsystems within the overall cognitive and perceptual system are relatively independent in terms of information processing; that is, they have relatively limited informational interaction with other cognitive subsystems. They are thus taken to be 'informationally encapsulated'.

This informational encapsulation has two 'directional' aspects. Firstly, there may be constraints on what information can get into the module and influence its working. This is sometimes called 'cognitive impenetrability'. A process is cognitively impenetrable to the extent that information that is available to other cognitive and perceptual processes is not available to the modular process, even though it may be relevant to the task being done. This is illustrated clearly by phenomena like the persistence of illusion. In the standard Müller–Lyer illusion (Figure 1), the visual system continues to perceive two lines as of different lengths, even when the viewer (after measuring, say) has available the information that they are in fact the same length. The visual system 'won't listen' to that outside information, and so the illusion persists.

Secondly, there may be constraints on the information apparently available to the module but unavailable to the rest of the cognitive system. This is called 'informational opacity'. For example, although a two-dimensional representation of (at least some of) a visual scene is probably computed along the path of visual processing, that two-dimensional representation seems not to be available for making other kinds of judgments. When a subject looks at a table top, the relative lengths and orientations of the two-dimensional projections (from the subject's perspective) of the edges of the table are not easily cognitively available to the subject to make judgments about. It seems that the visual system gives the rest of cognition some 'proprietary' description of visual input – perhaps as three-dimensional shapes – but reveals little of the

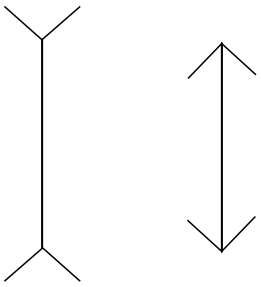


Figure 1. The Müller–Lyer illusion. Even when we know that the vertical lines are the same length, we continue to perceive the line on the left as longer.

methods and assumptions it uses to produce the representation that it gives as output.

Another feature commonly ascribed to modules is 'domain-specificity'. This is the idea that modular processes operate on only very specific domains. It is, however, a difficult notion to define precisely. Firstly, the notion of a 'domain' is not well defined; furthermore, domains, which are concerned with content rather than the form of processing, may change according to how we look at the information processed. A process might be seen as 'domain-specific' by virtue of doing only one particular form of calculation, but those calculations might be applied to representations with different sources and different contents.

Other features commonly ascribed to modules include both information-processing features (such as lack of competition for computational resources with other processes, and speed of operation) and features pertaining to the implementation and ontogeny of modules (that they should be innately specified, neurally specific and localized, hard-wired, and with a characteristic pace and sequencing of development). But there are wide differences of opinion about which of these are critical to the notion, or should even be expected in modules that may be discovered.

VIEWS AND THEORIES OF MODULARITY

The most prominent and influential view on modularity has been that of Fodor (1983). According to Fodor, there is a fundamentally trichotomous architecture to cognition: 'transducers', which just convert stimuli directly into signals to be used in processing; 'input systems', which are informationally encapsulated modules that make inferences about the sources of those inputs; and 'central systems', which are non-modular ('holistic') processors responsible for general inference, reasoning, and the fixation of beliefs. Input modules are constrained and encapsulated subsystems, but, unlike transducers, they engage in real nondemonstrative inference that goes beyond the information given in the stimulus alone: for example, in the case of perception, generating hypotheses about the distant environment.

The linguistic and visual domains were suggested by Fodor as likely to involve significant modularity, and much subsequent research on modularity has focused on these two domains. The views of Marr (1982) and his followers on visual processing have also been influential in encouraging a modular view of the visual system. Marr (for whom the boundary of the visual module is defined by the representation he calls the 'two-and-a-half-dimensional sketch') insists that within the visual module 'the processes can be influenced little or not at all by higher-order considerations' (Marr, 1982, p. 351). Marr further encourages a view of the visual module as largely decomposable into submodules (e.g. for stereopsis, and for constructing the 'raw primal sketch'), each of which is more or less informationally encapsulated even from the rest of the activity in the visual module.

Similarly, in language processing, many theorists have looked for significant modularity. For example, in Jackendoff's analysis, 'syntax … has *no* direct interface with the articulatory–perceptual system; rather, it interfaces with phonological structure, which in turn interfaces with articulation and perception' (Jackendoff, 1997, p. 30).

Much broader and more ambitious accounts of mental modularity have also been proposed. For example, Gardner (1983) suggests that much of higher cognition and intellectual skill is modular. This is in contrast to Fodor's view that higher cognition is generally a holistic and non-modular central system. Other accounts have suggested modules ranging from the perceptually driven (e.g. Lerdahl and Jackendoff's (1983) modular account of musical cognition), to modules in higher cognition specified in much more abstract terms such as the 'theory of mind' module (Scholl and Leslie, 1999), which has been postulated to explain the apparent nonpervasiveness of the cognitive disturbances involved in autism. In the last decade of the twentieth century, some more extreme modularist hypotheses about the mind appeared. Prominent among these is Sperber's (1994) suggestion of 'massive modularity', where a great variety of modules are seen as permeating the structure of the cognitive architecture. Such views have typically been coupled with evolutionary accounts of these allegedly ubiquitous modules (e.g. Cosmides

and Tooby, 1987), and have suggested modules whose 'domains' of operation are much more abstract than domains such as visual processing (e.g. Cosmides and Tooby's suggested 'cheater-detection module').

EMPIRICAL EVIDENCE CONCERNING MODULARITY

There has been significant experimental investigation of the kinds and degrees of informational encapsulation that various cognitive subsystems might exhibit. Much of this work has focused on the cognitive impenetrability of the systems of visual and linguistic perception. But some of the most obvious empirical evidence for at least some degree of modularity in such systems is phenomenological. For example, in the Müller–Lyer illusion, the phenomenological assessment of the stable output of visual processing (roughly, 'how it looks') strongly suggests that the output of visual processing is unaffected by (or cognitively impenetrable by) the information we have about the actual lengths of the lines.

In general, empirical investigation of the possible modularity of these systems has focused on what would appear to be relatively intelligent interpretation of stimuli, and considering whether information for solving such problems is likely to come from generally available (non-encapsulated) cognition about the domain (which would contradict the claim of impenetrability) or from constraints plausibly built into the subsystem itself. Cases in which general background knowledge is actually in conflict with the apparent inference made by the purported module are particularly salient. Thus, with the Müller–Lyer illusion, the background knowledge of a typical subject includes the knowledge that this is a familiar illusion and that the lines are actually the same length. But the apparent lack of penetration of the visual system by this information, combined with the fact that the actual stimuli (both as objects and as retinal stimulations) are lines of the same length, seems to imply that while the visual system is making an inference that goes beyond the stimulus in interpreting the input, it uses only its own limited information in the processing leading to that inference.

Perhaps the best evidence of modularity comes from early processes in vision. Much of this evidence concerns perceptions of illusions, like the Müller–Lyer illusion, and rules and strategies apparently used in visual perception but which are hidden from our notice. Rock (1983) has compiled a catalogue of such data, including facts about non-obvious completions of hidden and illusory visual contours, lightness constancy across illumination changes, and many others. Marr's (1982) analysis of the visual extraction of depth information from stereoscopic images also gives a clear illustration of powerful inferences made within the visual system using rules which the subject would seem to have neither access to nor the ability to override.

Some aspects of language processing also appear to be modular. Competence in learning a language appears to be quite independent of general problem-solving skills – an observation that goes back at least to Descartes, and that has been supported by recent research (e.g. Karmiloff-Smith *et al*., 1997), including brain lesion studies that suggest that localized lesions can affect language skills without significantly affecting other aspects of cognition (e.g. Caramazza *et al*., 1983). And careful chronometric studies of the 'phoneme restoration effect' – one of the most widely-recognized contextual effects on early language processing – have shown it to be plausibly a result of biases introduced by the postperceptual judgments of the subjects, and so perhaps not after all in violation of modularity (Samuel, 1981).

The evidence concerning modularity in other domains is more mixed and controversial; and in all domains, there are some threads of evidence that seem at least to limit the modularity. Farah (1994) has argued that lesion and deficit studies do not support, and sometimes even refute, the suggestion of neurological localization of speech functions. The presence of various sorts of top-down effects on perceptual recognition (e.g. in gestalt shifts) suggests that information flow into perceptual modules is not completely constrained (e.g. McClamrock, 1989). And the 'McGurk effect' (McGurk and MacDonald, 1976) seems to suggest that the process of recognition of even relatively low-level features of language like phonemes is penetrated by information coming from the visual modality, so that the face movements that are seen can influence which phoneme is heard.

ROLE OF MODULARITY IN COGNITIVE SCIENCE

The idea of modularity has had two distinct roles in contemporary cognitive science: one descriptive, the other methodological. In its descriptive role, modularity – and especially its notion of informational encapsulation – is one of many information-processing concepts used for constructing explanatory cognitive models. Empirical results such as those considered above seem to

imply the need for a model of the relative separation of, say, perception of (apparent) length and judgments about the properties of objects. Determining the extent to which such results demand modular models of processing is a difficult but not unusual problem of theory construction and testing.

The methodological role of the notion of modularity is more subtle. Modularity is supposed to provide a kind of 'divide and conquer' strategy for the study of cognition: by defining subsystems whose behavior in relative isolation is similar enough to their behavior in the real, embedded context, we might hope to come gradually to understand the complex working of a system. Conversely, unchecked top-down information flow into perceptual processes would make them difficult to analyze. As Marr puts it, the information-processing approach would be likely to fail with 'systems that are not modular … that is to say, complex interactive systems with many influences that cannot be neglected' (Marr, 1982, p. 356). And as Fodor puts it, 'the limits of modularity are also likely to be the limits of what we are going to be able to understand about the mind … [because] the condition for successful science (in physics, by the way, as well as psychology) is that nature should have joints to carve it at: relatively simple subsystems which can be artificially isolated, and which behave in isolation in something like the way that they behave *in situ*' (Fodor, 1983, pp. 126–128).

As Marr puts it for the case of vision, '[if] we can experimentally isolate a process and show that it can still work well, then it cannot require complex interaction with other parts of vision and can therefore be understood relatively well on its own' (Marr, 1982, p. 101). However, even where this experimental isolation can be achieved, the fact that such an isolable subsystem 'cannot require complex interaction with other parts' does not entail that it can 'be understood relatively well on its own'. The subsystem might still roughly accomplish its task under informationally impoverished conditions, but it might do so in a way that is not a good indication of its normal pattern of working. For example, it might take longer and resort to less efficient means to solve a problem than it would under conditions of normal informational access, where it might work more quickly, with less effort, and with fewer resources. Or it might work as quickly, but with greater error. The case of isolation would then give a misleading view of how the process normally works.

It may be useful to isolate processes in order to study them, but there is a concern that decomposition of the cognitive system has more to do with hopeful simplifying assumptions about cognition than with its actual structure. The accounts of 'massive modularity' mentioned above are perhaps the most optimistic but least empirically grounded of the current modularist views. And Fodor (2000) has recently suggested that assumptions of modularity and related strategies of decomposition have been applied too freely in the cognitive theorizing of the last years of the twentieth century, and are of much more limited applicability than is often suggested.

SUMMARY

Modularity, taken as the decomposability of cognition into components that can be considered in relative isolation from one another, is both a methodological ideal and, sometimes, an assumption that has influenced and guided research in cognitive science. Although there is some evidence to support the idea of modularity, especially in the early stages of processing involved in language and vision, the application of the idea currently exceeds the evidence for it. Modularity is sometimes more an optimistic guess, intended to guide explanation, than a conclusion of experiment and completed theory. Its soundness is a subject of widespread debate, and one of the central metatheoretical issues that cognitive science currently confronts.

References

Caramazza A, Berndt RS and Basili AG (1983) The selective impairment of phonological processing: a case study. *Brain and Language* **18**: 128–174.

Cosmides L and Tooby J (1987) From evolution to behavior: evolutionary psychology as the missing link. In: Dupré J (ed.) *The Latest on the Best: Essays on Evolution and Optimality*, pp. 277–306. Cambridge, MA: MIT Press.

Farah MJ (1994) Neuropsychological inference with a interactive brain: a critique of the 'locality' assumption. *Behavioral and Brain Sciences* **17**(1): 43–104.

Fodor JA (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor JA (2000) *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Gardner H (1983) *Frames of Mind: The Theory of Multiple Intelligences*. London, UK: Heinemann.

Jackendoff R (1997) *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Karmiloff-Smith A, Grant J, Berthoud I *et al*. (1997) Language and Williams syndrome: how intact is 'intact'? *Child Development* **68**: 246–262.

Lerdahl F and Jackendoff R (1983) *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Marr D (1982) *Vision: A Computational Approach*. San Francisco, CA: Freeman.
McClamrock R (1989) Holism without tears: local and global effects in cognitive processes. *Philosophy of Science* **56**: 258–274.
McGurk H and MacDonald J (1976) Hearing lips and seeing voices. *Nature* **264**: 746–748.
Rock I (1983) *The Logic of Perception*. Cambridge, MA: MIT Press.
Samuel A (1981) Phoneme restoration: insights from a new methodology. *Journal of Experimental Psychology: General* **110**(4): 474–494.
Scholl B and Leslie A (1999) Modularity, development and 'theory of mind'. *Mind and Language* **14**(1): 131–153.
Simon H (1981) The architecture of complexity. In: Simon H *The Sciences of the Artificial*, pp. 193–229. Cambridge, MA: MIT Press.
Sperber D (1994) The modularity of thought and the epidemiology of representations. In: Hirschfeld LA and Gelman SA (eds) *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 39–67. New York, NY: Cambridge University Press. [Revised version in: Sperber D (1996) *Explaining Culture: A Naturalistic Approach*. Oxford, UK: Blackwell.]

Further Reading

Carruthers P and Chamerlain A (eds) (2000) *Evolution and the Human Mind: Modularity, Language and Meta-Cognition*. Cambridge, UK: Cambridge University Press.
Garfield J (ed.) (1987) *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, MA: MIT Press.
Karmiloff-Smith A (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.
Samuels R (1998) Evolutionary psychology and the massive modularity hypothesis. *British Journal for the Philosophy of Science* **49**(4): 575–602.

Multiple Realizability

Intermediate article

John Bickle, University of Cincinnati, Cincinnati, Ohio, USA

CONTENTS

Multiple realizability, the claim that a type of mental state is implemented in a variety of distinct types of physical states, is the central premise in the most influential criticism of mind-to-brain reductionist programs. But the validity of this argument is under increasing scrutiny, based both on examples from more mature branches of science and on specific developments in the cognitive and brain sciences.

WHAT IS MULTIPLE REALIZABILITY?

Within the philosophy of mind, 'multiple realizability' is the claim that a type of mental state (property, event) is implemented in a variety of distinct types of physical states (properties, events). A philosopher's fantasy illuminates this claim. Suppose that the long-awaited day arrives and friendly aliens land on earth. Their behavior so closely resembles ours in a variety of circumstances that we soon begin explaining and predicting it using our own repertoire of mentalistic terms. For example, we explain alien Fred's consuming large quantities of beer because of his desire for liquid refreshment and his liking for the taste of earth beer.

At some time during their visit, an alien dies. They give us permission to conduct an autopsy. But when we open the alien's head, we find nothing that resembles our brains on the level of gross anatomy (the alien's head is filled with green slime), nor on the level of cellular anatomy and physiology (the slime contains nothing resembling our neurons), nor in its elemental constitution (the slime is silicon-based, not carbon-based). Clearly, the aliens lack brains like ours. Yet they seem to share our mental states. This fantasy suggests that a given type of mental state (desire for liquid refreshment, or a liking for the taste of beer) can be multiply realized in a variety of physical states that have little in common at any level of physical description.

Arguments appealing to multiple realizability, and critical replies to them, have generated a huge literature. Only the landmarks are discussed in this article. For more comprehensive presentations (written at a more technical level, with extensive bibliographies) see Bickle (1998, 1999).

HISTORY

In a series of papers published in the 1960s (most reprinted in Putnam, 1975), Hilary Putnam introduced multiple realizability into the philosophy of mind. He used it as a premise in an argument against early versions of the mind–brain identity theory. Using 'pain' as an example of a type of mental state, Putnam writes:

> Consider what the brain-state [identity] theorist has to do to make good his claims. He has to specify a physical–chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical–chemical structure; and (b) its brain is in that physical–chemical state. This means that the physical–chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. (Putnam, 1967)

That physical–chemical state must also be 'a state of the brain of any extraterrestrial life that may be found that will be capable of feeling pain' (ibid.). Furthermore, the identity theory makes a similar claim about every type of mental state. So 'if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical–chemical correlate is different in the two cases, the brain-state theory has collapsed' (ibid.). According to Putnam, it is 'overwhelmingly likely' that at least one such predicate can be found.

Putnam also used multiple realizability to defend the functionalist account of mind he was

developing in the same papers. His functionalism identified mental types with machine table states of a 'probabilistic automaton'. The technical details are not important for our purposes. The important point is Putnam's contention that his functionalism is consistent with the facts of multiple realizability, making it an attractive alternative to early identity theories.

Jerry Fodor (1974) broadened Putnam's multiple realizability argument to challenge all forms of 'reductionism'. In particular, he:

- extended the argument to criticize reductionism built upon existing theories of scientific ('intertheoretic') reduction.
- argued that the 'disjunctive strategy' of identifying a type of mental state with the disjunction of all physical–chemical states realizing it was untenable.
- hinted at an important distinction between two senses of multiple realizability.
- extended the argument against reductionism to include not just psychology but all of the 'special sciences' (other than basic physics).
- developed the concept of 'token physicalism' (a weaker doctrine than reductionism), which is consistent with multiple realizability and 'sufficient for any reasonable purpose'.

Fodor's arguments are still influential in their original form and they stand as monuments in the philosophy of mind.

ARGUMENTS APPEALING TO MULTIPLE REALIZABILITY

The most plausible account of scientific (intertheoretic) reduction available in the early 1970s was that of Ernest Nagel (1961). Nagel construed reduction as deduction of each law of the reduced theory from the laws of the reducing theory. However, reducing theories often do not contain predicates of the reduced theory (e.g., laws of statistical mechanics and microphysics do not contain predicates from classical thermodynamics like 'pressure' and 'temperature'). To obtain valid deductions in such cases, 'bridge laws' must be added to the laws of the reducing theory. Each bridge law relates a predicate of the reduced theory to the appropriate predicate of the reducing theory (e.g., 'temperature' to 'mean kinetic energy of the gas's constituent molecules'). The predicates from the reducing theory occurring in these bridge laws must pick out a kind of property (state, event, process) deemed relevant for explanation and prediction by the reducing theory. To do this, the predicate must occur within the laws of the reducing theory. Obviously, any intertheoretic reduction from psychology to physics will require bridge laws, since no physical science contains mentalistic predicates. But if reduction is to establish physicalism, then these bridge laws must be interpreted as (contingent) identity claims (e.g., 'mental state M is identical to physical state P'). Given the extent and variety of multiple realizability of the mental on the physical, identities require the physical predicate P in psychophysical bridge laws to be 'wildly disjunctive', perhaps infinitely so: M is identical to [P_1 (in humans) or P_2 (in octopuses) or … or P_n (in green slimy extraterrestrials) or …]. And even if the disjunction is finite, it is overwhelmingly likely that the resulting predicate will not be one that appears in any law of any physical science. Hence that disjunctive predicate will not denote a kind recognized as relevant for explanation or prediction in any potentially reducing physical science. As Fodor puts it, it would be 'an accident on a cosmic scale' if this requirement of physicalist reduction turned out to be true.

In a famous pair of examples, Fodor (1974) hints at a distinction between two types of multiple realizability. One is the type that Putnam emphasized: 'multiple realizability across structure types', in which different types of physical structures (human brains, octopus brains, alien green slime) realize the same mental state. The other type is more radical: 'multiple realizability within an individual system across times'. A human brain might realize the same mental state in different physical states at different times (depending upon neural changes brought about by learning or upon whatever other tasks it is carrying out). This more radical sense increases the number of disjuncts in the psychophysical bridge laws. Now there must be a predicate in the disjunction denoting every physical state realizing the given mental state in every instance of every type of cognizant creature at every time. This makes reductionism look extremely unlikely.

Although the functionalism spawned by the multiple realizability argument against reductionism is no longer the 'dominant' account of mind in Anglo-American philosophy, its heir – 'nonreductive physicalism' – also adopts Putnam's and (especially) Fodor's multiple realizability arguments (Horgan, 1993).

PROBLEMS FOR ARGUMENTS APPEALING TO MULTIPLE REALIZABILITY

David Lewis (1969) first offered what has come to be the standard reductionist reply to the multiple

realizability argument. Consider the following three claims:

There is only one winning lottery number. (1)

The winning lottery number is 03. (2)

The winning lottery number is 61. (3)

This triad seems inconsistent for the same reason as the following triad, namely, multiple realizability as expressed in claims 2 and 3 (5 and 6 below):

There is only one physical–chemical realization of pain. (4)

The physical–chemical realization of pain is brain state *B*. (5)

The physical–chemical realization of pain is green slime state *G*. (6)

But there is an easy dissolution of the inconsistency in claims 1 to 3: add 'per week' to claim 1, 'this week' to claim 2, and 'last week' to claim 3. Similarly, add 'per creature type' to claim 4, 'in humans' to claim 5, and 'in the friendly aliens' to claim 6, and the inconsistency disappears there. Lewis's general point is that reduction and identity is specific to a domain of phenomena, and hence is consistent with multiple realizability in the sense Putnam characterized.

Following Lewis's lead, a number of reductionists have pointed out examples of domain-specific reductions throughout the history of science. Temperature, for example, is multiply realized. Temperature in a gas is mean molecular kinetic energy. Temperature in a solid is mean maximal molecular kinetic energy (since the molecules are bound together in lattice structures and are restricted to a range of vibratory motions). Even a vacuum, which lacks molecular constituents, can have a 'black body' temperature, depending upon the electromagnetic waves coursing through it. So this paradigm 'reduced' concept is multiply realized at the level of microphysical and statistical mechanical description. So multiple realizability in itself does not appear to preclude reducibility, in actual scientific cases.

But is Lewis's strategy sufficient to handle the more radical, 'same instance of a system across different times' sense of multiple realizability? Wouldn't this sense require us to relativize reductions to descriptions of individual systems at times? And isn't this much 'domain specificity' inconsistent with the generality of scientific (reducing) theories? Are there examples of this more radical sense of multiple realizability from the history of scientific reductions? In answer to this last question, Enç (1983) hints that the example of 'temperature' is multiply realized at the level of microphysical description in this radical sense. A given instance of a gas can realize the same temperature – the same mean molecular kinetic energy – in an infinite number of distinct 'microcanonical ensembles' in which the kinetic energy of each molecule is specified individually. (This argument is developed explicitly in Bickle, 1998.) So multiple realizability in itself, even in this more radical sense, does not appear to preclude reducibility.

The other questions are more difficult to resolve. They appear to require a new account of intertheoretic reduction. But two considerations support the possibility of developing such an account. Firstly, a close examination of scientific examples reveals that multiple realizability in this more radical sense is common to many 'cross-level' reductions (Hooker, 1981). And the beginnings of a general theory of 'token–token' intertheoretic reduction have been suggested (Hooker, 1981), including a precise formulation of the relation using the resources of mathematical set theory and topology (Bickle, 1998). Secondly, there is reason to wonder whether this more radical sense of multiple realizability about mental kinds is actually so widespread. Kim (1993), Bickle (1998), and Bechtel and Mundale (1999) highlight different methodological practices of (obviously successful) contemporary neuroscience that assume commonalities in how psychological states are realized within and across individuals and species.

MULTIPLE REALIZABILITY AND COGNITIVE SCIENCE

Because of its implications for reductionist programs, multiple realizability remains central to cognitive science. Might cognitivist theories, whose generalizations advert to representations and computations over their contents, reduce to developed neuroscientific theories? Does the recent trend in mainstream neuroscience towards cellular and molecular investigations and explanations potentially tell us anything about cognition and mind? Is reduction, as this relation obtains between theories in physics, chemistry, biochemistry, and molecular biology, the appropriate intertheoretic relation to seek and expect across the disciplines comprising the cognitive and brain sciences? Multiple realizability and its perceived consequences are central to these questions. The historical connection between multiple realizabilility, functionalism in the philosophy of mind, and the scientific project of

artificial intelligence (AI) is also important. Is AI research producing actual examples of alternatively realized cognition?

These fundamental questions in contemporary cognitive science are of more than just 'philosophical' interest. The concept, use, and critical scrutiny of multiple realizability appears to be evolving as philosophical issues are continually reformulated into 'foundational' questions within the cognitive and brain sciences.

References

Bechtel W and Mundale J (1999) Multiple realizability revisited: linking cognitive and neural states. *Philosophy of Science* **66**: 175–207.

Bickle J (1998) *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle J (1999) Multiple realizability. In: Zalta E (ed.) *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/multiple-realizability/

Enç B (1983) In defense of identity theory. *Journal of Philosophy* **80**: 279–298.

Fodor JA (1974) Special sciences, or the disunity of science as a working hypothesis. *Synthese* **28**: 97–115.

Hooker CA (1981) Towards a general theory of reduction. III: Cross-categorial reduction. *Dialogue* **20**: 496–529.

Horgan T (1993) Nonreductive materialism and the explanatory autonomy of psychology. In: Wagner S and Warner R (eds) *Naturalism: A Critical Appraisal*. Notre Dame, IN: University of Notre Dame Press.

Kim J (1993) Supervenience and Mind. Cambridge, UK: Cambridge University Press.

Lewis D (1969) Review of *Art, Mind, and Religion*. *Journal of Philosophy* **66**: 23–35.

Nagel E (1961) The Structure of Science. New York, NY: Harcourt, Brace, World.

Putnam H (1967) Psychological predicates. In: Capitan WH and Merrill DD (eds) *Art, Mind, and Religion*, pp. 37–48. Pittsburgh, PA: University of Pittsburgh Press.

Putnam H (1975) Mind, Language, and Reality. *Philosophical Papers*, vol. II. Cambridge, UK: Cambridge University Press.

Further Reading

Block N (ed) (1980) *Readings in the Philosophy of Psychology*, vol. I. Cambridge, MA: MIT Press.

Churchland PM (1982) Is 'thinker' a natural kind? *Dialogue* **21**: 223–238.

Churchland PM (1987) *Matter and Consciousness*. Cambridge, MA: MIT Press. [Revised edition.]

Churchland PS (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.

Crumley JS (ed.) (2000) *Problems in Mind: Readings in Contemporary Philosophy of Mind*. Mountain View, CA: Mayfield.

Fodor JA (1975) *The Language of Thought*. New York, NY: Thomas Crowell.

Fodor JA (1981) *RePresentations*. Cambridge, MA: MIT Press.

Pylyshyn Z (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Richardson R (1979) Functionalism and reduction. *Philosophy of Science* **46**: 535–558.

Narrow Content

Intermediate article

Gabriel Segal, King's College London, London, UK

CONTENTS

Narrow content is the cognitive content of a psychological state that is conceptually and metaphysically independent of relations between the subject of the state and factors in the external environment.

BACKGROUND

Towards the end of the nineteenth century, the philosopher Gottlob Frege distinguished two kinds of representational content that symbols may have: sense and extension, or reference (Frege, 1892). The extension of a symbol is an object or a set of objects to which the symbol applies. For example, the extension of a proper name is its bearer, and the extension of a predicate is the set of things it is true of. Thus the extension of the name 'Gottlob Frege' is the man Gottlob Frege, and the extension of 'red' is the set of red objects. Frege saw that the notion of extension failed to account for important semantic distinctions, since coextensive symbols can have different meanings. For example, 'The Morning Star' and 'The Evening Star' both refer to the planet Venus, but they differ in meaning. Similarly, 'water' and 'H_2O' both extend over the same substance, but differ in meaning. Frege invoked the notion of sense to account for those aspects of literal meaning that go beyond extension.

According to Frege, the sense of a symbol is a mode of presentation of the symbol's extension, a way in which the extension may be presented in thought. Sense explains extension, in that a symbol has the extension it has because it has the sense it has. To understand an expression is to grasp its sense. And sense determines extension, which means that if two symbols have the same sense, they have the same extension.

Sense also accounts for the cognitive content associated with a symbol. For example, 'Hesperus is Hesperus' has different cognitive content from 'Hesperus is Phosphorus': the former is a tautology, while the latter expresses empirical knowledge. This was explained in terms of the sentences having different senses, or, in Frege's terminology, expressing different thoughts. Thoughts are also the contents of propositional attitudes, such as beliefs and desires. For example, to believe that Hesperus is Phosphorus is to stand in a certain cognitive relation to the thought expressed by the sentence 'Hesperus is Phosphorus'.

Hilary Putnam brought to light what he claimed was a common further assumption among philosophers of language who were sympathetic to the essentials of Frege's theory (Putnam, 1975). The assumption was that knowing the sense of a term is a matter of being in a certain 'narrow' psychological state. A narrow psychological state is one that is intrinsic to the subject of the state. It is one that does not require, as a matter of logical or metaphysical necessity, the existence of anything external to the subject. Being jealous of one's brother, for example, is not a narrow psychological state, because it is logically impossible to be in such a state unless one's brother exists.

Putnam presented an argument against the conjunction of the claims that grasping the sense of a term is being in a narrow psychological state and that sense determines extension. The argument was designed to show that narrow psychological states fail to determine extension, so one of the two claims is false. Putnam suggested it was the first one.

Putnam's argument hinges on a thought experiment. We are asked to compare Earth in 1750 with an imaginary planet called 'Twin Earth'. Earth in 1750 and Twin Earth are extremely similar. But they differ in that where Earth has H_2O, Twin Earth has a liquid that is superficially indistinguishable from H_2O but has a different chemical constitution, which we can call 'XYZ'. Since the chemical constitution of water was not known in 1750, nobody on Earth or Twin Earth would have been able to tell the difference between H_2O and XYZ. We then imagine a typical Earth subject, Oscar, and his Twin Earth counterpart, Twin

Oscar. Oscar and Twin Oscar, unlike real twins, are exact duplicates, molecule-for-molecule replicas of each other. They are identical in respect of all their intrinsic properties. (For the purposes of this experiment it is best to ignore the fact that humans are made out of water.)

Putnam then considered the question of whether the word 'water' meant the same thing in the twins' mouths. He argued that it did not. The argument is as follows. Our word 'water' applies to water, and water only. And water, according to our best scientific theory, is H_2O. XYZ is not H_2O, so it is not water. So 'water' is not true of XYZ. Moreover, 'water' had the same extension in 1750 as it has now. So Oscar's word 'water' extends over H_2O only and not over XYZ. By contrast, Twin Oscar's word 'water' extends over what he calls 'water', which is XYZ.

Since the twins are exact duplicates, their intrinsic properties are the same and so their narrow psychological states are the same. They differ only in their relations to their environments. But such purely relational matters are by definition irrelevant to narrow psychological states. So the twins are in the same narrow psychological state, but the word 'water' has different extensions in their mouths. So narrow psychological states do not determine extension.

Putnam argued further that, intuitively, the term 'water' has different meanings on 1750 Earth and Twin Earth. If Oscar and Twin Oscar had both pointed to a sample of XYZ and said 'that is water', then Oscar would have said something false and Twin Oscar would have said something true. And two sentences must differ in meaning, if one would be true and the other false in exactly the same circumstance.

If sense is to accord with the intuitive notion of meaning, then we should hold that their words have different senses. The conclusion is, then, that grasping the sense of a term is not just a matter of being in a certain narrow psychological state. For Oscar and Twin Oscar grasp different senses, but are in the same narrow psychological state.

Putnam was mainly concerned with the meanings of words and not with the content of psychological states. But many philosophers have extended his arguments to the latter, arguing that the twins' psychological states differ in content. For example, the belief that Oscar expresses when he says 'cold water is good to drink' differs in content from the belief that Twin Oscar expresses when he produces the same sounds. Contents that differ across Putnamian twins, which are not intrinsic to their subjects, are called 'broad'. Hence the claim is that at least some contents of psychological states are broad (e.g. Burge, 1982; McGinn, 1982).

Many (but not all) philosophers have felt that the notion of broad content does not capture the kind of content that appears in psychological explanation, and is deployed in the cognitive sciences. They hold that the kind of content that is or should be deployed in scientific psychology is intrinsic; i.e., narrow.

A spectrum of positions has emerged in the debate. At one end there is pure externalism, which holds that the only kind of content required for psychology is broad, and either is just extension, or, like Fregean sense, determines extension. Pure externalists hold further that at least some psychological states have only broad content and that there is no need to attribute to them an additional narrow content (e.g. Fodor, 1994; Burge, 1986). At the other end of the spectrum, pure internalism holds that the fundamental kind of content, and the only kind that matters for the purposes of scientific psychology, is narrow (e.g. Segal, 2000; Cummins, 1989). In the middle of the spectrum are two-factor theories, which accept a need for both broad and narrow contents, holding that all psychological states have a narrow content, and at least some have broad contents as well (e.g. Block, 1986; Loar, 1988; Fodor, 1987).

Each of these positions comes in a variety of different forms, and some philosophers might be classed as holding a borderline view. But a clear and fundamental point of dispute divides pure externalism from two-factor theories and pure internalism: the question of whether every psychological state has a narrow content.

ARGUMENTS FOR NARROW CONTENT

This section outlines three of the major arguments that all cognitive states have a narrow content.

Causal Explanation in Psychology

The first, and most widely discussed, argument for narrow content was offered by Fodor (1987). The gist of the argument is as follows. Psychology offers causal explanations. (1) A science that offers causal explanation should classify the states it describes by their causal powers. It should not classify states as different if they have identical causal powers. (2) The causal powers of twins' psychological states are identical. So, (3) psychology should count the twins' psychological states as the same. (4) Psychology classifies states in terms of their cognitive contents. So, (5) cognitive content is narrow.

The argument may be questioned at various points. For example, step 1 is questionable since the taxonomy of a science that offers causal explanations might be subject to a variety of different constraints, not all directly relevant to the causal powers of the items it describes. For example, historical evolutionary considerations may be relevant to biological taxonomy, even in branches of biology that offer causal explanations. Step 4 would be questioned by, for example, Stich (1983) and Egan (1995), who argue that at least some parts of cognitive science do not taxonomize by content. And Fodor himself has abandoned the argument, on the grounds that the kinds of twin cases in which classification by broad content would violate good taxonomic principles are either nomologically impossible or have other features that make them irrelevant to taxonomic issues (Fodor, 1994).

Interpretationism

A second argument for narrow content derives from a philosophical theory of the kind of cognitive content deployed in computational theories of cognition. This theory is sometimes called 'interpretationism'. It holds that cognitive content arises in a physical system when transitions among physical states of the system, including those of its components, may be systematically interpreted in a way that makes good sense of the activities both of the system as a whole and of its component parts.

Consider, for example, a conventional pocket calculator. If it receives a pair of numerals and a plus sign as inputs, it will produce a numeral as output. And it does so in such a way that if you were to interpret the numerals in the conventional manner, so that for example '2' represents the number two, the output numeral would always represent the sum of the input numerals. Hence the behavior of the system makes sense under this interpretation. Moreover, the same applies to states of the internal components of the calculator: these can be interpreted as performing calculations the results of which are combined to produce the overall result that the calculator delivers.

According to interpretationism, cognitive content just reduces to this sort of interpretability. Any interpretation that would serve to make sufficient sense of a system, in the manner described, would be valid and would be a correct assignment of cognitive content. Any cognitive state of a system will therefore have many cognitive contents, each one being part of a systematic overall interpretation of a set of cognitive states. (Sometimes certain provisos are added to rule out the assignment of cognitive contents to systems that are too simple.) Interpretationism is articulated by Haugeland (1978) and Cummins (1983, 1989).

Under a certain variety of interpretationism, content is narrow because the actual environment of a system is irrelevant to whether a candidate interpretation makes sense of the system's behavior. For example, an interpretation under which Oscar's 'water' concept was assigned the extension XYZ would be just as good as one that assigned it H_2O. Suppose that Oscar is thirsty and reaches for a glass of water. A conventional explanation might be that he believes the glass contains some water and wants to drink some water. But if we regard Oscar as a computational system, then we might just as well interpret the internal states of the system that cause Oscar's arm to move as representing the glass as containing XYZ and representing the goal of drinking some XYZ. And a better interpretation than either of those would be assign the 'water' concept an extension that included both H_2O and XYZ, since nothing in the cognitive system distinguishes between them.

The main problem with this account is the theory of content on which it rests. In particular, many find the very liberal assignment of cognitive contents implausible. It is commonly thought that each cognitive state has a unique cognitive content rather than a plethora of them.

Empty Extensions

A third argument for narrow content focuses on representations with empty extensions, such as representations of ghosts or phlogiston.

The ideas behind the argument are best understood in the context of a specific example. Take as our first subject someone, Ralph, who has never seen a real anaconda, but who has seen a picture of one in a book and read the accompanying description. We can conceive of a Twin Earth on which there are not and never have been any anacondas, nor any twin anacondas, and on which the term 'anaconda' is empty. Twin Ralph has seen a book just like the one Ralph saw. Twin Ralph's book is fictional but he does not know this.

The argument proceeds as follows. The representation Twin Ralph associates with 'anaconda' has a cognitive content. This cognitive content cannot depend on its relation to members of its extension, because there are none. Moreover, whatever accounts for the representation having its cognitive content is present on Earth as well, and applies to Ralph's corresponding representation. So Ralph's representation has that very cognitive content as

well (Segal, 2000). Assuming that the argument generalizes appropriately, the conclusion is that all mental representations have narrow content.

The last step of the argument may be questioned. For it could be that what accounts for the cognitive content of Twin Ralph's representation on the Twin Earth is not present on Earth as well. For it might be that the full account of how Twin Ralph's representation comes to have its cognitive content would include the fact that there are no snakes or other objects that stand in the right relations to the representation to endow it with a non-empty extension.

THEORIES OF NARROW CONTENT

Ordinary, commonsense psychological explanations typically ascribe content using propositional attitude reports; that is, sentences with forms like '*a* believes that *p*' and '*b* desires that *q*'. The content of the attitude is conveyed by the 'that' clauses. According to common intuitions, the contents so ascribed are broad, not narrow: 'Oscar believes that cold water is good to drink' ascribes to Oscar a belief with the content that cold water is good to drink; Twin Oscar does not have a belief with that content, so the content ascribed is broad. A defender of narrow content thus has to provide an account both of what narrow content is and of how it can be deployed in psychological explanation. We now look at three of the major accounts of narrow content, mentioning some of the problems with them.

Functional Role Accounts

The most popular accounts of narrow content are functional role accounts. The functional role of a mental representation is given by some selected aspects of its causal role, where the latter consists of the totality of potential causal relations that the representation enters into.

The basic idea behind functional role accounts is rather simple. Consider the mental representations of the two Oscars that correspond to their word 'water': their *water* representations. The two *water* representations share certain important causal characteristics. For example, visual presentations of typical samples of either H_2O or XYZ would typically cause either subject's *water* representation to appear in a belief state. If either subject is thirsty and has a representation of the form *there's some water in that glass in front of me* appearing in a belief state, then this representation will combine with others to cause the subject to reach out for the glass. More generally, the pattern of potential causal relations – with stimuli, other mental representations, and behavior – that the two representations enter into is the same. The idea is that causal relations of that sort constitute the representations' cognitive contents.

Functional role accounts face two related problems concerning the individuation of functional roles. Firstly, the account needs to distinguish those aspects of a representation's causal role that are relevant to its cognitive content from those that are not. For example, the causal roles of the representations of two subjects who differ in their short-term memory capacities will differ, since memory imposes constraints on the way representations can interact. A defender of a functional role account needs to decide whether this difference is to count as a difference of content, and then find a way of implementing the decision in the specification of functional roles (Segal, 2000).

Secondly, the account needs to provide a way of individuating functional roles that allows for meaningful comparisons across different subjects. The difficulty is that some causal properties of a psychological state depend upon other states with which it is associated. For example, if Ralph believes that anacondas are dangerous while Thelma believes they are not, then the belief that one is being confronted by an anaconda will tend to cause Ralph, but not Thelma, to retreat. Nevertheless, we might want to use the term 'anaconda' to frame psychological generalizations that apply to both subjects. It might be true to say, for example: 'Ralph and Thelma both believe that anacondas are very large snakes, indigenous to South America, which kill their prey by constriction.' The functionalist thus needs to find a way of classifying functional roles that abstracts away from the difference between Ralph's and Thelma's anaconda concepts and *allows for the generalization*. If this is not possible, then functional roles will not provide an appropriate way of grouping psychological states for the purposes of psychological explanation. See Fodor (1987) and Field (1976) for related discussion.

Character

A second account of narrow content has been offered (Fodor, 1987), which can be understood as a development of a kind of content originally proposed by David Kaplan (e.g. Kaplan, 1978), called 'character'. The idea of character arose in response to a problem for Frege's account of sense and extension that is in some ways similar to that illustrated by Putnam's Twin Earth argument. The

problem concerns indexical expressions, such as 'I', 'today', and 'that'. Indexical expressions have variable extensions: for example, utterances of 'today' on different days refer to different days. Since sense determines extension, utterances of indexicals with different extensions must have different senses. But the meaning of an indexical expression remains constant across different occasions of use: 'today' does not alter in meaning from day to day. Frege's account does not capture this constancy of meaning.

Kaplan proposed to capture the constant meaning of indexicals in terms of character. The character of an indexical can be understood as a function from contexts of utterance to extensions. A context can be thought of as a sequence of relevant items, such as speaker, time, and place. The character of 'today' maps any context *c* onto the day encompassing the time of *c*. The character of 'I' maps any context *c* onto the speaker of *c*.

Kaplan (following Perry, 1977) suggested that character might account for cognitive content. For example, suppose that two different people on different days say: 'It is my birthday today.' The utterances of 'my' and 'today' have different senses and different extensions. However, it is plausible that the two utterances express the same cognitive content. Since the words uttered have the same character, character remains a candidate for cognitive content.

Since 'water' is not an indexical expression, its extension does not vary across contexts of utterance, as these are normally construed. As a speaker moves from place to place and persists through time, the extension of his or her term 'water' does not change. Further, according to most thinkers, if Oscar were suddenly transported to Twin Earth, his term 'water' would still extend over H_2O. That is why, if he were confronted with a Twin Earth ocean and said 'that's water', he would be saying something false. Hence, character as originally conceived by Kaplan does not serve as narrow content. However, Fodor developed a generalization of the notion that was designed to meet the difficulty.

Fodor's idea was that if Oscar had been transported to Twin Earth and then stayed there for a long period, then his word 'water' would have come to extend over XYZ, rather than H_2O. This is because, after a sufficient period, he would have become a member of the Twin-English-speaking community and his use of the term 'water' would have become dominated by his interactions with XYZ. And it is those factors that are usually held to determine the extension of a term. The same ideas extend to Oscar's mental representation, *water*.

Fodor extended the idea of a context of utterance, or thought, to include what might be called a 'home environment'. A subject's home environment is the environment in which the subject has become embedded and which fixes the extensions of the subject's mental representations. Fodor's suggestion was that narrow content is a function from home environments to extensions. Oscar's and Twin Oscar's *water* representations have the same narrow content because they instantiate the same function from home environments to extensions. If Oscar had been in Twin Oscar's home environment, then his *water* representation would have extended over XYZ. If Twin Oscar had been in Oscar's home environment, then his *water* representation would have extended over H_2O. Since the two subjects are twins, the result generalizes: for any home environment *e*, Oscar and Twin Oscar's water representations would have the same extension in *e*.

Fodor's account of narrow content, like functional role accounts, faces the problem that it needs to provide a way of individuating narrow contents that allows for meaningful generalizations across subjects. For while it is reasonable to suppose that twins' representations will instantiate the same function from home environments to extensions, it is not clear whether the same would be true of any non-twinned subjects. Subjects whose conceptions of something differ slightly will probably have representations that would extend over different things in certain home environments, hence would have different narrow contents.

A second problem is that we have as yet no vocabulary for talking about these functions from home environments to extensions. The account needs to provide some such vocabulary as well as some explanation of how it would function in psychological generalizations.

Twin Earth Revisited

A third account of narrow content involves less of a departure from the traditional Fregean theory of content than the previous two accounts. This account rejects the standard intuitions about Twin Earth experiments. It denies that the extensions of twins' representations differ. For example, rather than holding that Oscar's water representation extends just over H_2O and that Twin Oscar's extends just over XYZ, it holds that both representations extend over both H_2O and XYZ. The extension of both representations consists in water-like substances generally. This account therefore allows that the cognitive content of a representation, like Fregean sense, determines its extension.

An advantage of this account over the previous ones is that it does not require any sweeping revisions to the vocabulary of psychology. It does require revision of some people's views about particular ascriptions of content that arise in respect of Twin Earth cases. (For example, rather than saying that Oscar believes that water is good to drink, one might say that Oscar believes that the watery stuff is good to drink.) The account does require a revision of certain aspects of commonsense psychology, but not the development of a new psychological vernacular.

One difficulty with this account of narrow content is that it cannot be applied to indexical representations. For different indexicals with the same narrow content have different extensions: when Oscar and Twin Oscar both think *it's hot here*, their thoughts have the same narrow content, but refer to different places. So the account would have to be combined with a separate theory of narrow content for indexicals. Segal (1989a) has attempted to provide such a theory.

NARROW CONTENT AND COGNITIVE SCIENCE

The arguments for narrow content presented in this article have been pitched at a high level of generality. Some of the debate in the philosophy of psychology has been at a more specific level. One can examine specific theories within cognitive science, and address the question of whether the kind of content deployed in these specific theories is narrow. The theory that has received the most attention is the computational theory of vision, as developed particularly by Marr (1982), and which has been discussed by Burge (1986), Segal (1989b, 1991), Egan (1991), Davies (1991), Butler (1996), and Wilson (1995). An advantage of this kind of debate is that different theories within cognitive science have, to an extent, their own particular methodologies and hence their own constraints on the attribution of content. It may therefore be easier to tell whether a specific theory deploys narrow (or broad) content than whether cognitive science in general does. As the various branches of cognitive science develop, issues about the width of content will become clearer.

References

Block N (1986) Advertisement for a semantics for psychology. In: French P, Euhling T and Wettstein H (eds) *Midwest Studies in Philosophy*, vol. X, *Philosophy of Mind*. Minneapolis, MN: University of Minnesota Press.

Burge T (1982) Other bodies. In: Woodfield A (ed.) *Thought and Object*. Oxford, UK: Clarendon Press.

Burge T (1986) Individualism and psychology. *Philosophical Review* **95**(1): 3–45.

Butler K (1996) Individualism and Marr's computational theory of vision. *Mind and Language* **11**: 313–337.

Cummins R (1983) *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Cummins R (1989) *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Davies M (1991) Individualism and perceptual content. *Mind* **100**(4): 461–484.

Egan F (1991) Must psychology be individualistic? *Philosophical Review* **100**(2): 179–203.

Egan F (1995) Computation and content. *Philosophical Review* **104**: 181–203.

Field H (1976) Logic, meaning and conceptual role. *Journal of Philosophy* **74**: 379–409.

Fodor J (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor J (1994) *The Elm and the Expert*. Cambridge, MA: MIT Press.

Frege G (1892) Uber Sinn und Bedeutung. Translated in: Geach P and Black M (eds) (1952) *Translations from the Philosophical Writings of Gottlob Frege*. Oxford, UK: Clarendon Press.

Haugeland J (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* **1**: 215–226.

Kaplan D (1978) Dthat. In: Cole D (ed.) *Pragmatics*, vol. IX, *Syntax and Semantics*, pp. 221–243. New York, NY: Academic Press.

Loar B (1988) Social content and psychological content. In: Grimm R and Merril D (eds) *Contents of Thought: Proceedings of the 1985 Oberlin Colloquium in Philosophy*, pp. 99–139. Tucson, AZ: University of Arizona Press.

Marr D (1982) *Vision*. New York, NY: WH Freeman.

McGinn C (1982) The structure of content. In: Woodfield A (ed.) *Thought and Object*. Oxford, UK: Clarendon Press.

Perry J (1977) Frege on demonstratives. *Philosophical Review* **86**: 474–497.

Putnam H (1975) The meaning of 'meaning'. In: Gunderson K (ed.) *Language, Mind and Knowledge*. Minneapolis, MN: University of Minnesota Press.

Segal G (1989a) The return of the individual. *Mind* **98**: 35–57.

Segal G (1989b) Seeing what is not there. *Philosophical Review* **98**: 189–214.

Segal G (1991) Defence of a reasonable individualism. *Mind* **100**: 485–493.

Segal G (2000) *A Slim Book About Narrow Content*. Cambridge, MA: MIT Press.

Stich S (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Wilson R (1995) *Cartesian Psychology and Physical Minds: Individualism and the Sciences of Mind*. Cambridge, UK: Cambridge University Press.

Further Reading

Burge T (1979) Individualism and the mental. In: French P, Uehling T and Wettetein H (eds) *Studies in Epistemology*, pp. 73–121. Minneapolis, MN: University of Minnesota Press.

LaPorte J (1996) Chemical kind reference and the discovery of essence. *Nous* **30**(1): 111–132.

McGinn C (1989) *Mental Content*, chap. 1. Oxford, UK: Blackwell.

Mercier A (1994) Consumerism and language acquisition. *Linguistics and Philosophy* **17**: 499–519.

Williamson T (1998) The broadness of the mental: some logical considerations. In: Tomberlin J (ed) *Language, Mind, and Ontology*, pp. 389–410. Oxford, UK: Blackwell.

Wilson M (1982) Predicate meets property. *Philosophical Review* **91**: 549–589.

Neural Correlates of Visual Consciousness

Intermediate article

Charles A Heywood, University of Durham, Durham, UK
A David Milner, University of Durham, Durham, UK

CONTENTS

The neural correlates of consciousness are those patterns of neural activity that are associated with subjective experience.

INTRODUCTION

The study of consciousness is thwarted both by the lack of an all-embracing definition and, because it is a private experience, its inaccessibility to objective measurement. Nevertheless, useful distinctions have been drawn between different varieties of consciousness (Block, 1995). Access (A)-consciousness refers to the ability to respond to, report or engage in higher-level activities such as reasoning, on the basis of the contents of a representation that resides in a person's brain. In contrast, phenomenal (P)-consciousness refers to the nature of subjective experience – that is, the quality of the bodily sensation, such as the shrillness of pitch or the glare of sunlight. Monitoring (M)-consciousness and self (S)-consciousness, as their names imply, refer to awareness of bodily sensations and awareness of oneself, respectively.

Scientific accounts of consciousness are mainly concerned with A- and P-consciousness, and fall into two broad classes of proposals with regard to their neural basis. The first class emphasizes the way in which information is represented in assemblies of neurons, and is likely to be more relevant to P-consciousness. The second class is more concerned with functional relationships between representations and the processes that they are engaged in, rather than the contents of the representations themselves, and is likely to include accounts that are more appropriate to A-consciousness.

The philosopher David Chalmers distinguishes between 'easy' and 'hard' problems (Chalmers, 1995). 'Easy' problems include questions about how people tell different sensory stimuli apart, attend to and react to them and produce verbal reports. Such questions, although difficult, are tractable because they concern objective mechanisms of the cognitive system that are explicable in terms of neural and computational mechanisms. For example, we have a tolerably good understanding of the mechanisms whereby different wavelengths of light are differentially absorbed by cone pigments in the retina and produce different neural signals for comparison and discrimination. What is absent is any understanding of why an intense long wavelength of light appears vivid red, or indeed why it looks like anything at all. This is the 'hard' problem – the problem of P-consciousness.

WHAT IS A NEURAL CORRELATE OF CONSCIOUSNESS?

Although it is generally accepted that understanding how physical states of the brain may give rise to conscious states presents philosophical and empirical problems whose solution is currently beyond our ability, that does not prevent the undertaking of a first empirical step in approaching this problem. This first step is to investigate what *correlations* may exist between brain states and mental states. The hoary question of causality would then be left for future investigators to grapple with. Although even this correlational question still has a long way to go before it can be answered fully and convincingly, scientists have begun to address it in recent years.

The most promising strategy so far has been to build on our current neuroscientific knowledge. This means that the question has to be focused on particular types of mental processes, rather than addressed broadly with regard to conscious states

in general. A distinction can readily be drawn between the neural events that underpin the *state*, or *level*, of consciousness and those from which the *content* of consciousness arises. Research pertaining to the former involves establishing which brain states correspond to coarse-grained behavioral states such as sleeping, dreaming, wakefulness or, more generally, levels of arousal. In contrast, the search for the neural correlate of consciousness has been chiefly concerned with the *content* of consciousness, namely the phenomenal properties of conscious experience. Phenomenal experience characteristically has modality-specific representational content (e.g., the vividness of colors or the salience of shapes experienced by the observer of a visual scene). The challenge, then, is to find the neural activity that is necessary and sufficient for our conscious perception of, for example, 'redness' or 'roundness'.

This is a different pursuit from identifying the neural events that determine whether we are visually conscious *per se*. The neural correlates of state consciousness are essentially arbitrary. Any pattern of neural activity may be associated with the state of being conscious, but in the case of the neural determinants of subjective visual experience, there is the additional requirement that some characteristic of the neural state must match the content of consciousness. In the latter case the neural state, mediated by the firing patterns of its constituent neurons, represents attributes of the external world. For example, cells in the visual pathways will preferentially respond to a particular orientation or speed of motion of a line segment. With the exception of the special case of visual illusions, such attributes are perceived veridically and are signaled by the neural responses of cells that are tuned to the particular orientation or speed of the visual stimulus falling in their receptive field.

RECENT RESEARCH ON NEURAL CORRELATES OF CONSCIOUSNESS

The above choice of examples from the visual modality is neither idiosyncratic nor merely fortuitous. The past 25 years have seen enormous breakthroughs in our knowledge of the neuroscience of visual processing, mainly through studies of the monkey, but also increasingly now through application and extension of this knowledge to the human brain. Accordingly, the Nobel laureate Francis Crick and his neurobiologist colleague Christof Koch have argued that the domain of visual perception offers the best way forward at the present time for pinning down consciousness (Crick and Koch, 1998). In other words, the question being addressed is 'What is it about the neural systems that mediate visual processing that makes their activity conscious?'. Needless to say, much visual processing goes on that is not accessible to consciousness. For example, we are not aware of what drives our visual reflexes, such as the pupillary reflex to light, nor are we typically aware (until after the event) of what guides automatic actions such as catching an object that we accidentally knock off a tabletop. In the realm of visual awareness, the rapidly alternating presentation of two equiluminant lights results in a stable fused percept, yet neurons in cortical area V1 respond to the fleeting color changes. A closely related question to the one above is therefore 'What is it about the neural systems that mediate visual processing that makes it conscious under some conditions but unconscious under others?'

One means of answering this question is by studying patients with brain damage who show a mismatch between the contents of visual awareness and the efficiency with which a visual stimulus can elicit an appropriate response. The story begins with the famous investigations of so-called 'blindsight' in patients with hemianopia. The neuropsychologist Larry Weiskrantz was the first to undertake systematic studies of this phenomenon, which refers to residual visual processing in individuals who have suffered damage to the primary visual cortex (usually now termed 'V1') on one side of the brain (Weiskrantz, 1986). Since in cases of this type there is a large area of the person's visual field which is subjectively 'blind', by definition any residual 'blindsight' occurs without concomitant visual awareness. We now know that blindsight can be manifested in a wide range of phenomena, ranging from visuomotor acts such as pointing to or reaching to grasp objects, to guesses about the shape or color of two-dimensional images. For example, a patient may be able to reach, or move his eyes, towards spots of light that are not subjectively 'seen', with surprising (although not normal) accuracy. He may prepare to grasp an object in his 'blind' visual field with approximately correct wrist orientation and handgrip size, despite not knowing anything about the object. The upshot of this field of research is that area V1 appears to be a *sine qua non* for visual awareness, but not for the trivial reason that it removes visual processing altogether, since the latter clearly still proceeds to a remarkable degree.

However, this does not of course mean that visual consciousness is *located* in V1. Indeed,

research on a different type of patient demonstrates that it is not. Patients with visual-form agnosia, who have typically suffered brain damage as a result of anoxia or carbon monoxide poisoning, have a profound loss of shape or form perception – that is, they are quite unable to name, describe or discriminate even simple geometric shapes. Yet despite their loss of awareness of shape, they often have a largely intact area V1, as shown both by magnetic resonance imaging (MRI) and by clinical testing of their visual fields. Clearly, therefore, V1 may be *necessary* for visual consciousness, but it is certainly not *sufficient* for it. It is only a part of the necessary circuitry. Furthermore, in one particular patient with visual-form agnosia, D.F., it has been found that her V1 is able to collaborate with other brain areas well enough to allow her to perform a very wide range of visuomotor acts with normal levels of accuracy (Milner, 1998). For example, she can reach out, form her grasp, avoid obstacles, walk around a cluttered environment, catch moving objects and move her eyes, all with apparent normality. This research shows that the areas underlying this processing beyond V1 (which are almost certainly located in the dorsal-stream areas of the parietal cortex) can operate well without their visual processing reaching awareness. Thus the simple formula 'V1 plus higher visual areas' does not add up to a sufficient recipe for visual awareness.

A clue as to where an ingredient in the recipe *may* be located beyond V1 is provided by the damage that is present in D.F.'s brain. It is located in front of V1, and functional MRI scanning shows that it interrupts the neural traffic between V1 and the temporal lobes at the sides of the brain. For some years these areas have been known to be intimately involved in the neural processing of the complex features of objects that allow us to recognize them. In physiological experiments conducted in awake animals, neurons in the inferior temporal cortex (ITC) of the monkey have been found to respond selectively to images such as faces, colored shapes, and the actions of others. They also send information to brain areas concerned with memory and affect. Nikos Logothetis and his colleagues (Leopold and Logothetis, 1996) have now demonstrated that the activity of these neurons is closely linked to the perceptual experience of the monkey. They have achieved this by training the monkey to 'report' manually which of two alternative images it sees on a screen in front of it. When incompatible images (e.g., a face and a house) are presented to the two eyes, a person does not see a fused combination of the two, but rather either a complete face or a complete house, the two percepts alternating at intervals of a few seconds. When the experimental animal faced such 'binocular rivalry', it reported an alternating awareness of the two images just as would be expected from human experience, and the responses of inferior temporal neurons followed these perceptual reports with a high correlation. In other words, despite the fact that the images on the screen remained unchanged, visual consciousness did change repeatedly, and these changes correlated with the firing of neurons in the ITC. Moreover, and importantly, the responses of cells in the primary visual cortex (V1) do not show such a correlation. When the monkey was presented with horizontal and vertical gratings under conditions of binocular rivalry, responses were recorded equally from those cells with horizontal and those with vertical receptive field preferences in area V1, notwithstanding the fact that the monkey reported only seeing, say, vertical gratings. The neuronal response in area V1 thus covaried with the retinal stimulus, not with the visual percept.

Previous research by Charles Gross and his colleagues had shown that neurons in the inferior temporal cortex depend on area V1 for their visual input. In the absence of V1, the neurons lost their visual responses. This would explain why damage to V1 in humans results in subjective blindness – the neurons in the ITC that seem to determine the contents of visual awareness would no longer receive the information that is needed to activate them. Can we then infer from these data that visual experience has its neural correlate in the firing of neurons in the inferior temporal cortex? The answer is not quite. Our visual experience is certainly severely compromised when this part of the brain is damaged or disconnected, as patient D.F. demonstrates. However, our visual experience is impaired when other parts of the brain are damaged as well.

Visuospatial neglect is a condition typically caused by damage (often following a stroke to the middle cerebral artery) to a region of the right hemisphere around the border between the parietal, occipital, and temporal lobes. The result is impoverishment, or even absence, of any awareness of the left side of visual space. However, unlike damage sustained by area V1, neglect is not limited to one side of the retina's view of the world. When patients with hemianopia move their eyes around, their blind half-field moves as well, so they can still manage to see everything that they want to see. However, the neglect patient *still* fails to see things on the left side, even when the eyes are moved to look there. What is lost from the

person's experience is thus not like any type of blindness, but rather something more abstract than that.

The work of the Italian neuropsychologist Edoardo Bisiach has shown just how abstract this loss can be (Bisiach and Luzzati, 1978). He discovered that many neglect patients in his home city of Milan could describe from memory the view across the Cathedral Square, including the cathedral itself and many of the famous buildings around the piazza. Remarkably, however, they would omit to mention buildings located on the left-hand side of their imagined view. Moreover, when asked to imagine standing with their backs to the cathedral, looking the opposite way, they would omit to mention buildings on the *other* side, now on their imagined left. In addition, Bisiach's compatriot, Guido Gainotti, found that many neglect patients may fail to process the left sides of objects even on the right ('good') side of their perceived visual space. These discoveries of 'imaginal' and 'object-based' neglect suggest that neglect is not something that affects the early stages of visual processing, or even the later stages of processing in the ITC where neurons appear to respond selectively to different categories of object. Those neurons may provide the items of furniture for our visual experience, but there must be a higher level of visual representation where this furniture is arranged and can be rearranged. It is at this stage that neglect seems to exert its effects.

Thus the phenomena of neglect suggest that there is a part of the posterior human right hemisphere where conscious scenes and even multi-component objects are constructed. Just as a television image may offer a long-range view of a landscape or a close-up of a single antique teapot, so this representational system in the brain seems to operate at different spatial scales depending on our needs. Consequently, damage to the system can affect whatever we happen to be looking at (or imagining), however large or small it may be. Why is just the left side of these representations affected? The answer may relate to the notable rarity of neglect phenomena after *left* hemisphere damage. It is suggested that there is a similar representational brain area in the posterior left hemisphere, but that it can represent only the right side of our perceptual experience. When its dominant partner in the right hemisphere has been destroyed, it has to do all of the representational work – and it can handle only the right side of space, and even then not all that well. Its efforts are what the neglect patient experiences as a depleted visual world. However, when the left hemisphere area itself is damaged there is no serious impairment of our perceptual experience, as the dominant partner on the right can handle the whole visual scene sufficiently well alone. (Interestingly, however, other types of mental experience may be affected by such left-hemisphere damage, such as our ability to comprehend or plan spoken sentences, perform mental arithmetic, or think using 'inner speech'.)

So is the parietotemporal cortex the seat of our conscious mental life? It certainly seems to play a critical role, but probably not alone. Many theorists (including Crick and Koch) have argued for the centrality of the prefrontal cortex (whose large size is reflected in the uniquely prominent foreheads that we see in the human species) in our conscious life. However, damage to the prefrontal cortex does not deprive people of their conscious experience. Instead, it causes problems in their ability to plan ahead, to foresee the consequences of their actions, and to abide by rules. In other words, damage to that region disrupts our ability to choose between different courses of action according to criteria that depend on abstract mental representations, as opposed to short-term goals.

So could there be a collaboration between that posterior representational system (which is damaged in neglect) and an anterior 'executive' system that is able to manipulate and interrogate those representations? The posterior system may be able to hold the arranged furniture of our mental imagery, but the anterior system might coordinate the arranging and rearranging of it. The anterior system might also play an essential role in the control of verbal processing in the left hemisphere. This general idea would fit well with the theories of cognitive psychologists such as Alan Baddeley and Tim Shallice.

METHODOLOGIES FOR FINDING NEURAL CORRELATES OF CONSCIOUSNESS

Of course, narrowing down the anatomical substrates of visual awareness is only the first step in the task of finding the neural correlates of visual awareness. It may tell us where to look, but not what we shall find there. For that purpose it will be necessary to use the types of refinements of single-neuron recording pioneered by Logothetis and his colleagues, hand in hand with the progressively more refined techniques of human functional neuroimaging that are now available. These may help us to inch further along the road. Indeed, functional magnetic resonance imaging (fMRI) has

been used to confirm the results of Logothetis in human observers, with the added benefit of monitoring whole-brain activity. Changes in frontal and parietal activity were also shown to correlate with changes in perception when subjects were confronted with a binocularly rivalrous display (Lumer *et al.*, 1998).

A further promising approach to distinguishing between brain regions involved in conscious and nonconscious perception is the use of visual masking. The rapid successive presentation of a visual stimulus and a pattern mask can abolish the subjective experience of the former. Yet it is well known that such unseen stimuli can have behavioral consequences. This provides a convenient method for distinguishing between brain regions involved in conscious and nonconscious processing in normal observers. For example, positive variations in blood flow, measured by positron emission tomography (PET), have been shown in the midbrain and amygdala in response to masked (and unseen) facial expressions of anger which had previously been associated with an aversive stimulus (Morris *et al.*, 1999). Subjects were shown angry faces either alone, or accompanied by an aversively loud noise. During scanning, brief presentation of a face was followed by a visual mask composed of a neutral facial expression. Despite failure to detect the emotional facial expression, the aversive stimulus elicited an autonomic response (an increased skin conductance response) and resulted in increased midbrain and amygdala activity. The emotional valence of a face was therefore conveyed in the absence of visual awareness. This suggests that a comparison of regions of activation during presentation of stimuli that straddle the masking threshold (i.e., those that are just above and just below the threshold for eliciting a conscious percept) would provide a promising means of establishing the neural activity related to visual awareness.

However, it may well be that substantial progress will have to await the development of completely new methodologies. One of the questions that will have to be addressed along the way is that of selective attention. It is a commonplace observation that the brain can process visual information to quite a high level without the products of that processing reaching the threshold of consciousness, simply because one's attention is directed elsewhere. Why does a given assembly of neurons achieve awareness on one occasion but not on another? What distinguishes these two neuronal states? At present we have little idea of the answers to these questions, but there have been several suggestions. For example, it could be that the assembly with the highest level of activation inhibits all others and emerges as the dominant one that determines our awareness at that point. Alternatively, it could be that relative neural 'timing' is at the heart of the matter. There is now quite good evidence for synchronous firing among separate neurons that participate together in response to visual patterns. It may be that the greater the degree of synchronous activity, the higher the probability that the percept will reach the level of consciousness. Of course, neither of these suggested correlations offers any explanation for the process whereby we can choose to attend to one rather than another object in the first place.

The search for neural correlates of consciousness is just beginning, but the results are promising. Recent results suggest that correlates can be found in the firing pattern of individual cells in neural assemblies. Future research will have to address the thornier questions not only of how consciousness emerges from such neural activity, but also of the function of consciousness itself.

References

Bisiach E and Luzzati C (1978) Unilateral neglect of representational space. *Cortex* **14**: 129–133.

Block N (1995) On a confusion about the function of consciousness. *Behavioral and Brain Sciences* **18**: 227–247.

Chalmers DJ (1995) Facing up to the problem of consciousness. *Journal of Consciousness Studies* **2**: 200–219.

Chalmers DJ (1997) Moving forward on the problem of consciousness. *Journal of Consciousness Studies* **4**: 3–46.

Crick F and Koch C (1998) Consciousness and neuroscience. *Cerebral Cortex* **8**: 97–107.

Leopold DA and Logothetis NK (1996) Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**: 549–553.

Lumer ED, Friston KJ and Rees G (1998) Neural correlates of perceptual rivalry in the human brain. *Science* **280**: 1930–1934.

Milner AD (1998) Streams and consciousness: visual awareness and the brain. *Trends in Cognitive Sciences* **2**: 25–30.

Morris J, Ohman A and Dolan RJ (1999) A subcortical pathway to the right amygdala mediating 'unseen' fear. *Proceedings of the National Academy of Sciences of the USA* **96**: 1680–1685.

Weiskrantz L (1986) *Blindsight: a Case Study and Implications*. Oxford, UK: Oxford University Press.

Further Reading

Atkinson AP, Thomas MSC and Cleeremans A (2000) Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences* **4**: 372–382.

Bisiach E (1992) Understanding consciousness: clues from unilateral neglect and related disorders. In: Milner A and Rugg M (eds) *The Neuropsychology of Consciousness*, pp. 113–137. London, UK: Academic Press.

Block N (1996) How can we find the neural correlate of consciousness? *Trends in Neurosciences* **19**: 456–459.

Frith C, Perry R and Lumer E (1999) The neural correlates of conscious experience: an experimental framework. *Trends in Cognitive Sciences* **3**: 105–114.

Kurthen M, Grunwald T and Elger CE (1998) Will there be a neuroscientific theory of consciousness? *Trends in Cognitive Sciences* **2**: 229–234.

Logothetis NK and Leopold DA (1998) Single-neuron activity and visual perception. In: Hameroff S, Kaszniak A and Scott A (eds) *Toward a Science of Consciousness II*. Cambridge, MA: MIT Press.

Milner AD (1995) Cerebral correlates of visual awareness. *Neuropsychologia* **33**: 1117–1130.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford, UK: Oxford University Press.

Other Minds: The Problems

Intermediate article

Alec Hyslop, La Trobe University, Bundoora, Victoria, Australia

CONTENTS

The problem of other minds questions the basis of our knowledge that others have minds, and enquires into how our beliefs about the mental states and inner lives of others can be supported.

INTRODUCTION

Outside of psychopathology, we are all sure that other human beings are very like us. We are sure they think, they reason, they feel pain, feel happy, and so on. But what entitles us to such sureness? Anglo-American philosophy cannot agree on what underpins this most basic of human beliefs. Neither can Continental philosophy.

But this is more than a case of philosophers having some abstractly theoretical, even if deep, disagreement. The different theories affect our view of ourselves and our human situation. Broadly, there is a contest between those who see each individual as initially self-enclosed and needing to reach out to others, somehow, but condemned forever to remain separate; and those who see individuals as essentially social, communal beings, depending on others not merely for their existence, but for their very sense of themselves. Such a difference is fundamental to us as social beings. It has implications for how we live our lives, how we relate to other people, even how we vote.

What Anglo-American philosophy does agree on is what the problem of other minds is concerned with: namely, the basis of our knowledge about the mental states, the inner lives, of other human beings. What is agreed is that it is primarily an epistemological problem, concerned with how our beliefs about those mental states could be supported. Insofar as it is thought that there is a conceptual problem (how is it possible to acquire the concepts we have of mental states other than our own?) that problem is generally raised in the context of solving, or disposing of, the epistemological problem. Even those who, like Wittgenstein (1953), think there is no epistemological problem, agree that the issue is whether or not there is such a problem. This has not been the case within Continental philosophy. There the relevant writings are concerned with our experience of others, our relations with others, how others affect us, are essentially involved in our sense of ourselves, to the extent that often the line between philosophy and psychology seems blurred.

It is remarkable that there is no generally agreed solution to this great human problem. Probably the favored Anglo-American solution takes the form of a (scientific) inference to the best explanation: that other people have experiences is seen as the best explanation of their behavior.

What Generates the Problem?

We have a problem because of the asymmetry that exists between our own case and that of others. We often know directly that we are in a certain mental state. We never have this direct knowledge where others are concerned. The same asymmetry is responsible for the conceptual problem. How can we have the concept of experiences other than our own if our own experiences are the only experiences we know directly to exist?

HISTORY

The remarkable fact about the other minds problem is that it was established as a philosophical problem only as recently as the nineteenth century, when John Stuart Mill (1865) gave us his version of the analogical inference to other minds. Increasing dissatisfaction with the analogical inference culminated within Anglo-American philosophy, strongly influenced by Wittgenstein, in the use of the notion of criterial evidence to deal with the problem. This happened about the middle of the twentieth century. Independently, on the other side of the Channel, Sartre (e.g. 1958) can be seen as engaged in a similar line of thought. Subsequently, and independently again, other elements of Sartre's

thought, stressing the interdependence of the notions of self and other, are mirrored in Anglo-American philosophy, though expressed very differently. More recently, in Anglo-American philosophy, other minds have been regarded as on a par with the theoretical entities of science and supported by means of a scientific inference (Pargetter, 1984). This is probably the currently most favored view in Anglo-American philosophy.

APPROACHES TO THE PROBLEM OF OTHER MINDS

The Analogical Inference to Other Minds

The traditional but now unfashionable solution to the problem of other minds has been an analogical argument. I know that others are very like me in all sorts of ways. I know directly that I have thoughts, feelings, sensations, and the like. So I am enabled to infer that other people also have beliefs, experiences, and emotions. In short, it is with others as it is with me.

This conclusion is, however, impossible to check, not just in fact, but in principle. But where this used to be seen as problematic, it now seems to be regarded as benign, probably because most of the hypotheses so indispensable to science are in practice not such that they can ever be directly verified. However, that my experience is the indispensable centrepiece of this inference, occasions the supposedly fatal objection to this analogical inference, that it is a generalization from one case and therefore unsound, indeed hopeless. This feature is seen as so far from benign that it seems all other responses to the problem of other minds seek to avoid any evidential reliance on our own experience. Approaches to the problem as disparate as applying scientific inference to others, the criterial approach, and virtually all of the responses in Continental philosophy, could be seen as driven by this desideratum.

However, its supporters meet the one-case objection head on, standardly arguing that the requirement for more than one case is to establish a causal link between events. It is claimed that the relevant causal link, involving mental states, can be known to hold from one's own case.

Other Minds as Theoretical Entities

This is most likely, among Anglo-American philosophers, and cognitive scientists, to be the preferred solution to the problem of other minds. A scientific inference is utilized: the best explanation for the way other human beings behave is that they behave as they do because of their mental states. Crucially, no evidence depending on what we know from our own case is used to support this inference.

It has been argued against this approach that the intrinsic content of mental states, how they are experienced, in particular phenomenal properties such as the hurtfulness of pain, cannot be supported by this method. That content can only be filled in by an appeal to one's own case (Hyslop, 1976).

Criteria and Other Minds

Criterialists have attempted to avoid the one-case objection by insisting that the connection between behavior and mental states is neither entailment (as in behaviorism) nor an inductive inference. The connection is claimed to be conceptual and such a connection is characterized as criterial. It has been claimed by some that such a non-inferential connection is required if we are to have any concept of the experiences of others. A claim that itching is conceptually linked to scratching would be an example of such a criterial claim. It is further claimed that scratching is, thereby, known to be evidence of itching, independently of any observed correlations.

Against this conceptual approach it is argued that if there is no entailment directly from the observed behavior to the unobserved inner states to which they are conceptually linked, and there is no appeal to some form of inductive inference, then we are left with the gap. The gap cannot be crossed by *fiat*, as it were.

One way of understanding what has been called the attitudinal approach to other minds is to see it as going beyond other uses of criteria in insisting that our very conception of other human figures is that they have experiences. That is our attitude to them, built in as it were. That is how we perceive them. It is immediate, noninferential, preceding any belief, deeper than mere belief. This view seems, however, not to avoid the criterial gap. Our conceptions might be mistaken. That there are such mistaken attitudes to things and people seems clear (Hyslop, 1995).

PROBLEMS OF OTHER MINDS AND COGNITIVE SCIENCE

Can machines think? What exactly is it to think? Is the Turing test appropriate? Are there significant

differences between different mental states? To what extent, if any, are mental states observable? Are there respects in which, say, emotional states are more inaccessible than beliefs? To what extent is our access to the 'inner' lives of others mediated by their behavior, dependent on what they say, dependent on our capacity to 'read' expressions, dependent on our capacity to understand 'body language', dependent on our imaginative capacities? (*See* **Turing Test**)

These questions are all canvassed inside the other minds problem, to varying degrees. However, some general points can be made. Most discussion is centered on how we are to justify our belief that other human beings have mental lives like our own. Much less attention has been given to whether we have day-to-day practical knowledge about what other people are thinking and feeling in particular cases. Also, the concern is to justify the beliefs we have, not to describe how we acquire them, though this distinction is less honored in Continental philosophy. Continental philosophy differs also in that its focus is on the existence of other human subjects, individuals, persons, rather than (merely) their varied mental states, whereas Anglo-American philosophy confines itself almost entirely to justifying our belief that other human bodies have mental states attaching to them.

A radical response to the problem is to deny, by insisting that we have in some sense direct knowledge of the mental states of others, that the claimed asymmetry between ourselves and others exists. However, some theories of mind have endorsed this, or near enough. Continental philosophy has contested the asymmetry in arguing that awareness of others precedes, or is inseparable from, our having a concept of ourselves.

Even if the asymmetry is accepted, it has been almost universally thought that only a traditional dualist view of the mind has a seriously challenging problem of other minds. Behaviorism is either thought to have no problem at all, or, if it does, nevertheless has no difficulty in solving the problem. Behavior is, after all, observable.

Functionalism is a currently fashionable theory of mind, probably the preferred view of the cognitive scientist, that is thought to have the problem but in a comparatively easy form. Mental states are viewed as internal states differentiated by their various roles, and particularly by the behavior they give rise to, having no other features relevant to their being the mental states they are. It is then comparatively straightforward to infer that such internal states exist, given the appropriate behavioral evidence. Eliminative materialists, denying, heroically, that mental states exist, seem not to have any problem: no minds, no other minds.

However, it has been argued that the other minds problem cannot be so easily evaded or downgraded. Any theory of mind has to be true of one's own mind and other minds, so such a theory cannot, it has been argued, be used to solve the other minds problem. To assume that, say, functionalism is true, in the context of the problem of other minds, is to assume that it holds of minds in general. That would seem to assume there are other minds and the theory fits them. That would seem to ignore the problem, not solve it (Hyslop, 1995).

It has also been insisted that a theory of mind, covering all minds, needs to embrace the theorist's mind as well as other minds. So the evidence (generally implicit) for the theory will include the theorist's experience. That is the only way direct supporting evidence can be obtained.

References

Hyslop A (1976) Other minds as theoretical entities. *Australasian Journal of Philosophy* **54**: 158–161.

Hyslop A (1995) *Other Minds*. Dordrecht, Netherlands: Kluwer.

Mill JS (1865) *An Examination of Sir William Hamilton's Philosophy*. London, UK: Longmans.

Pargetter R (1984) The scientific inference to other minds. *Australasian Journal of Philosophy* **62**: 158–163.

Sartre J-P (1958) *Being and Nothingness*, trans. H Barnes. London, UK: Methuen.

Wittgenstein L (1953) *Philosophical Investigations*. Oxford, UK: Blackwell.

Further Reading

Ayer AJ (1963) *The Concept of a Person*. London, UK: Macmillan.

Buford TO (ed.) (1970) *Essays on Other Minds*. Chicago, IL: University of Illinois Press.

Hill CS (1991) *Sensations: A Defense of Type Materialism*. Cambridge, UK: Cambridge University Press. [Chapter 9, Knowledge of other minds.]

Husserl E (1977) *Cartesian Meditations*, trans. D Cairns. The Hague, Netherlands: Martinus Nijhoff.

McDowell J (1982) Criteria, defeasibility, and knowledge. *Proceedings of the British Academy* **68**: 455–479.

McGinn C (1984) What is the problem of other minds? *Proceedings of the Aristotelian Society* **58** (suppl.): 119–137.

Merleau-Ponty M (1981) *Phenomenology of Perception*, trans. C Smith. London, UK: Routledge.

Schroeder WR (1984) *Sartre and His Predecessors: The Self and the Other*. London, UK: Routledge.

Strawson PF (1959) *Individuals*. London, UK: Methuen.

Pain, Philosophical Issues about

Advanced article

Valerie Gray Hardcastle, Virginia Tech, Blacksburg, Virginia, USA

CONTENTS

There is no consistent philosophical view concerning the nature of pain, how to understand it, or what an understanding of pain might mean for philosophy of mind.

PHILOSOPHICAL VIEWS OF PAIN

Just about every conceivable position concerning the nature of pain is currently held today by some leading thinker or other. Each of these positions has become grist for someone's mill in arguing either that pain is a paradigm instance of a conscious state or that pain is a special case and should not be included in any general theory of consciousness.

Subjective Views

Some philosophers and psychologists hold that pain is completely subjective; it is either essentially private and completely mysterious, or it does not correlate with any biological markers but is completely nonmysterious. The International Association for the Study of Pain (IASP), the formal organization charged with defining pain, has articulated a paradigm subjective view. They write:

> Pain is always subjective … Many people report pain in the absence of tissue damage or any pathophysiological cause; usually this happens for psychological reasons. There is usually no way to distinguish their experience from that due to tissue damage if we take the subjective report … [Pain] … is always a psychological state. (1986)

However, if one holds that pain does not correlate in some way with some sort of bodily twitch, then one becomes a dualist. If pain is merely a private experience, and that experience has no consistent underlying physical cause or correlate, then we lose any interesting connection between the mind and the body over pain.

Philosophers are used to confronting the sorts of issues the IASP comes up against in defining pain. They eschew dualism by retreating to so-called token–token identity theory. Every experience in me is correlated with – identical to – some event or other in my brain. And every experience in you is correlated with – identical to – some event or other in your brain. But, if the subjectivists are right, there is no identifiable neural activity that is the same across all experiences of a migraine, for example. There is no brain correlate for the type 'having a migraine headache'. We can talk about generic headache experiences only from a level of abstraction above brain activity: namely, from the perspective of the mind and its cognitive states.

To uncover the regularities underlying certain brain areas and pain processing, we need repeatable instances of some sort of neural activity being connected to some sort of perceptual experience. However, if we deny type–type identity for larger brain structures across organisms, then at the same time we are denying ourselves any hope of discovering mind–brain connections. For we can generalize to anything from only one instance, and mental event–physical state correlations taken one at a time are all a robust token–token identity theory allows.

At the same time, functionalists of this ilk generally don't want to disconnect the mind entirely from the brain. We believe that there are areas in the brain dedicated largely to pain processing, just as there are other areas dedicated to vision, audition, touch, and so forth. We believe that these areas are basically the same across humans, despite individual variation. Thus, even though a strict type–type identity might fail for particular sensory experiences, it still underlies our views of our sensory systems taken as a whole. Types in science are allowed some play in them. They have to or we would have no mechanism by which to pick out any sort of cognitive processing in the brain at all.

All these lessons are missed by proponents of the subjective view, for they identify pain with the experience of pain and then explicitly deny that that experience has any correlation with any

particular bodily reaction. But since they want to be materialists interested in a scientific understanding of pain, they will have to permit generalizations connecting something in the body with the sensation of pain.

Objective Views

Other philosophers and neurophysiologists argue that pain is completely objective; it is either intrinsic to the injured body part, a functional state, a set of behavioral reactions, or a type of perception. Pain is something we can measure about our bodies or our behavior. As a result, its connection to our mentality, to our sensations of pain, is secondary at best. We might recognize that the pain is there in terms of how it feels to us – the skin burns, for example. For example, according to objective views that take pain as intrinsic to the injured body part, the pain itself is in the tissue. Hence, our beliefs or judgments about the condition of the tissue are derivative. That is, we use the nociceptive or pain information we get from the periphery to infer we are in pain.

Similarly, if pain is understood as a type of perceptual process, then it works no differently than vision or olfaction. We receive some sort of perceptual input on our transducers, manipulate that information in our brains, and then use that manipulated information to alter motor reactions, surrounding mental states, and what have you. Part of the manipulated information might come into conscious awareness, but that sensation would only comprise a subset of what is meant by pain processing. According to this view, our conscious experience of pain, the damaged tissue itself, and our bodily and emotional reactions are fundamental to pain processing. Each is one component in a larger process. Working together, these components take pressure, temperature, and chemical readings of our tissues and use this information to track what is happening in our bodies. They function together as a single complex system to monitor our tissues in order to promote the welfare of our bodies.

In these cases, and most other instances of the objective view, pain is something entirely physical. Prima facie, it appears that the states or processes identified with pain could occur without any awareness of them at all. Most objective views of pain have the unintuitive consequence of divorcing pain from sensations of pain or making the mental events associated with pain processing secondary to and dependent upon the pain processing itself.

Other Views

There are a few objectivist philosophers who hold that pain is not a purely physical event. Instead, it is something like an attitudinal relation. Pain requires both a bodily state and then cognition over that state. Pain itself is the attitude, the belief, we have regarding our bodily condition. This approach gets around the intuitive difficulties of the objective views by identifying pain with the consequent mental state. 'Pain' then just refers to the mental event associated with pain processing. According to this view, we have pain processing and then pain proper.

CENTRAL PHILOSOPHICAL ISSUES ABOUT PAIN

There are three large philosophical difficulties in defending any of the theories about pain processing outlined above: the problem of mental causation, the problem of naturalizing content, and the threat of eliminativism.

The Difficulty with Mental Causation

The difficulty with mental causation is roughly as follows. If I drop a hammer on my foot and subsequently experience pain, that experience is the proximal cause of my writhing, cursing, and gnashing of teeth. Dropping a hammer on my foot leads to pain behavior only if it causes in me the sensation of pain and the belief that I am in pain. If I were unconscious or otherwise oblivious to my surroundings, then I could not sense any pain, nor could I believe that I were in pain. Nor could I manifest any pain behaviors.

On the other hand, if we take a neurophysiological view of the entire hammer-dropping incident, then it seems as though we should be able to explain exactly the same events without appealing to mentality or any sort of psychological entities at all. We might talk about the intense pressure of the hammer head on my foot stimulating various nerve endings which causes action potentials to travel up my leg to my spinal column, where other nerves are then stimulated to fire. These nerves transmit the firing pattern to other nerves, and so it goes until nerves that cause muscles to contract are likewise stimulated and you get the writhing, wincing, and teeth-gnashing behavior. Why doesn't the possibility of this sort of more precise, purely physical explanation rule out the higher level, more general mental account? Or, why doesn't it make the mental account nothing more than a placeholder

until we get the details of our central nervous system figured out? As long as one is persuaded by reductionism, then pain provides an exemplar case for suggesting why psychological explanations appear so tricky.

There is some evidence that depression is related to pain processing. One view is that untreatable chronic pain causes depression, which in turn increases the sensations of pain. This is a (grossly oversimplified) mentalistic explanation of how a mood causally interacts with other psychological states. At the same time, we know that depression is correlated with a decrease in the neurotransmitter serotonin. Persons suffering from chronic pain also show a decrease in their levels of serotonin. But, if depression is (maybe) just an imbalance and a neurotransmitter and sensations of pain are some neural state or other, then why shouldn't we (someday) explain the relation between depression and pain in terms of neurotransmitters affecting neural activity? Why isn't the mentalistic explanation just a stand-in until we have all the more basic neurphysiological details under control?

Mental events causing other mental events seems to be a natural part of our explanatory world. At the same time, accounts of mental causation appear to be nothing over and above a sloppy characterization of more fine-grained and little understood physical details. The difficulty for those who would like to keep the mind intact as an explanatory unit is in explicating how it is that mental causation has a legitimate place in our understanding of the universe, above and beyond being a surrogate for the real causal story. Why shouldn't we claim that our ultimate goal in explaining pain is to redescribe our pain and other related mental states in terms of their neural instantiations, thereby eliminating the former in favor of the latter?

Naturalizing Content

Though most philosophers of mind treat mental causation separately from issues concerning reference, explaining the causal powers of the mind really piggybacks on the problem of naturalizing content. What makes the question of mental causality peculiar is that the content of mental states is relevant to their efficacy. I wince and nurse my foot because my corresponding mental states are about my foot. If they were about something else, then most likely I would be doing something else. To explain exactly how it is that mental events cause other things, we are first going to have to explain how it is that they refer. That is, to justify privileging a mentalistic explanation of sensations and beliefs over a lower-level physicalistic one of neuronal firing patterns or ionic flow in our scientific explanations, we first have to have a clear grasp about what we mean by mental events being contentful, since their content is what is causally relevant to our subsequent behavior.

The matter of the power of the content of beliefs and other mental states is quite important to understanding pain processing. What one is thinking and believing about the world strongly influences how much pain one feels. Athletes intently focusing on their game can break large bones and not even notice. But the same athletes, alone in their living rooms, will writhe on the floor if they stub their toes. Chronic pain patients can be trained to diminish their sensation of pain by changing their focus of attention and their beliefs about death and disease. Those suffering congenital indifference to pain often lead short and unpleasant lives because they can't sense painful stimuli, nor can they form appropriate beliefs about what the vague tinglings that they do feel mean. How pain feels to us depends to a large extent on our current doxastic surround. Hence, understanding pain requires an understanding of what beliefs and desires (and other mental states) are and how they refer.

Eliminativism

One implication of current scientific theories of pain is that our folkways of describing our pains are inadequate and we would be better off eliminating them from our everyday practices. The claim is that our folkways of talking about pain comprise a rough and ready theory of pain. This theory assumes that pains are identical to the sensations of pain and that the word 'pain' can capture the essence of that sensation. From the perspective of some objective views of pain, both assumptions are dubious. Pain processing is enormously complicated and our sensations of pain form only a tiny subset of what these processors do. But even if we focus exclusively on our sensations, since these are most important to our folkways of being, our folk theory is still inadequate. We simply do not have the language to express all the dimensions of our pain experiences. The descriptors we use are either metaphorical or non-existent. Our folk theory of pain needs to be replaced by something commensurate with the phenomenology.

Consider this: not only can we distinguish between the sensory, affective, and cognitive dimensions of pain phenomenologically, we can also

manipulate them independently of one another. We can feel a shooting pain in our leg, but not suffer in the least from it; we can be in agony from pain, without feeling any particular sensation localized to any part of our body. We could simply decide by fiat that 'pain' is going to refer to the localized sensations, or we could just decide that 'pain' is going to refer to the suffering. But either way, we do violence to our folk notion of pain which requires that a single simple sense datum seems to occur in some place and also be unpleasant.

In response to these sorts of claim, some have argued that our folk views of pain do not constitute a theory in any meaningful sense. Some believe that we know certain introspective facts indubitably. Pain is touted as one of those things. Perhaps there are are some sensory states, like pain, about which people have special first-person apprehension. We just know when we are in pain; no inference of judgment is required.

However, it is quite easy to demonstrate that our introspective knowledge of pain can be mistaken. If we burn our hand by touching something hot, we jerk our hand away from the heat source. This is a reflex action; the nociceptive information travels up our arm to the spinal column and then back down again. It takes about 20–40 milliseconds from stimulus to behavior. The information also travels up the spinal column to the brain. We feel the burn as well. Unlike the reflex movement, this processing is more complicated and it takes about 200–500 milliseconds from stimulus to perception, a full order of magnitude longer.

Nevertheless, if we introspectively report on what the incident feels like, we say that we moved our hand away after we felt the pain; feeling pain initiated the motor sequence. For whatever reason, our brains backdate our pain sensations so that they seem causally relevant to our reflex behavior. But clearly we can't cause the effect after it occurs, so our introspective report has to be wrong. We don't have special, first-person knowledge of our pains. Whatever knowledge we do have is embedded and informed by a conceptual framework of our brains' devising. Despite protests to the contrary, our pain experiences have all the earmarks of being at least proto-theoretical in nature.

Other detractors point out that even if a completed science of pain does not use folk terms for pain, that would not entail that those sorts of mental states do not exist; they just would not be referred to in scientific discourse. Our notion of pain would be analogous to our ideas about tables and chairs, germs and gems, and birthday presents and birthday cake. These are perfectly legitimate terms. We just don't use them in science. Being cultural artifacts of one stripe or another, they do not refer to things about which we would have laws. We might not have a mental science or laws about pains, but our folk psychology could still be used as it is now, in our everyday explanations of our behavior.

There is something undoubtedly right about this charge. In many ways, pain experiences are environmentally determined. Puppies raised without ever experiencing pain and without ever seeing any other dog in pain will exhibit no pain behavior. They will repeatedly sniff a lighted match without fear and then show no reaction when burned. Children learn both pain behaviors and the emotional concomitants to pain from the reactions of others around them. Expressions of pain and reports of sensation and experience are significantly different across cultures. Much of our experience of pain and how we react to and express these experiences is socially relative, a cultural artifact of sorts.

However, social relativity is not enough to show that our folkways of understanding pain are adequate. Different cultures have different experiences; they also have different ways of understanding these experiences. Nevertheless, the burden falls on the folk psychologist to demonstrate how our folk theories of pain are actually successful. This work has not yet begun.

THE ETHICS OF PAIN TREATMENT

One of the most hotly debated subjects in pediatric care concerns whether infants are insensitive to pain. The presumption historically has been that because young infants are not conscious, they cannot sense pain. As a result, analgesics and anesthetics are rarely used, even in the most invasive of procedures.

At first, this presumption of insensitivity is curious because infants' reactions to painful stimuli are well documented. Even premature neonates exhibit stress responses, hormonal fluctuations, and slowed recovery to painful interventions. In fact, we know that the afferent nociceptive system is up and running by 29 weeks' gestation, but that pain inhibitory systems don't come into operation until later. If anything, infants should be more sensitive to pain than adults. We can at least say that, by all indications, infants are sensitive to pain in some sense or other.

However, the question for many doctors is whether infants are aware of their pain. Some argue that unless neonates can consciously

apprehend pain, then any sorts of response they give to noxious stimuli are merely reflexes. Hence, we have no reason to treat infants' pain, since the infants can't feel anything.

Suppose they are right, even though there is much that goes on in our brains that is neither conscious nor mere reflex. It is still the case that infants react to pain, both behaviorally and physiologically, that these reactions can be modified with relatively simple treatments, and that treating pain has an impact on recovery. We know that early exposure to pain, whether remembered or not, affects later experiences of and reactions to pain by altering the developmental course of the nervous system. Infants, like other newborn animals, learn to attach particular meanings, or emotions, or importance to particular experiences in virtue of what is associated with those experiences. This sort of behavioral malleability is very important if an organism is going to survive in a complex environment. Consequently, manipulating early experiences can have drastic effects later on, as animal studies show. Merely by changing the smells associated with suckling, scientists can alter adult sexual behavior in male rats, for example. Similar changes occur with pain processing in young infants. Nociceptive stimuli increase the size of the somatic receptive fields for neurons sensitive to pain and help maintain dendritic connections that would otherwise be eliminated over time. Perhaps, as some believe, chronic pain and hypersensitivity can result from early acute pain episodes, in the list of how the neural receptors change. Early pain experiences have been shown to influence later personality and temperament. Something as common as circumcision can have lasting effects on pain sensitivity if done without anesthesia.

We must conclude that given the impact early pain processing can have on later development, we have every reason to prevent infant pain, even if it feels dissimilar to our own, even if it feels like nothing at all to the infant. Whether infants consciously experience pain – and the degree to which they may or may not be aware of some noxious stimulus or their own suffering – is a red herring. Available evidence converges around the idea that infants process pain, though perhaps not in the same way adults do. This processing has an impact on current behavior and later development. Because this influence is generally negative, insofar as we are able to prevent or alleviate some of their pain, we should do so.

Further Reading

Anand KJS and Craig KS (1996) New perspectives on the definition of pain. *Pain* **67**: 3–6.

Deberyshire SWG (1996) Comment on editorial by Anand and Craig. *Pain* **67**: 210–211.

Dennett DC (1978) Why you can't make a computer that feels pain. *Synthese* **38**: 449.

Gamsa A (1994) The role of psychological factors in chronic pain. I and II. *Pain* **57**: 5–29.

Hardcastle VG (1999) *The Myth of Pain*. Cambridge, MA: MIT Press.

International Association for the Study of Pain (IASP) Subcommittee on Classification (1986) Pain terms: A current list with definitions and notes on usage. *Pain Supplement* **3**: 217.

Lawson J (1988) Pain in the neonate and fetus. *New England Journal of Medicine* **318**: 1398.

Wall PD and Melzack R (eds) (1989) *Textbook of Pain*, 2nd edn. New York, NY: Churchill Livingstone.

Pain

Intermediate article

C Richard Chapman, University of Utah, Salt Lake City, Utah, USA

CONTENTS

Pain is an unpleasant sensory and emotional bodily awareness that normally serves a protective function by informing us of tissue damage.

DEFINITIONS OF PAIN AND RELATED CONCEPTS

Although it is familiar to everyone, pain is difficult to define scientifically. The International Association for the Study of Pain (IASP) offered the following definition in 1979, and it remains the current standard for the field: 'An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.'

In referring to pain as an experience, the IASP definition recognizes the private and individual nature of pain, that pain can occur only in a conscious individual, and that pain exists not at the site of injury but as a complex phenomenon occurring at the level of the brain. It avoids the common error of characterizing pain as a simple or primitive sensation. Moreover, it recognizes that the experience of pain has emotional as well as sensory features. Consensus exists among pain researchers and clinicians that pain has both sensory and emotional dimensions and many would add a cognitive dimension to acknowledge that pain involves attention and interpretation. Normally, pain is associated with tissue damage or potential tissue damage, but the definition wisely leaves open the possibility that pain can exist in the absence of tissue damage. A person may experience pain and describe it in terms of tissue damage, whether or not such damage exists.

The IASP definition has stood the test of time, but it will eventually give way to a revised definition. It does not accommodate neonates, infants, and small children, who perceive and react to tissue damage despite their ability to form a complex perception and describe it. Nor can it account adequately for the pain of a cognitively impaired person, who cannot report the pain but emits behaviors consistent with pain, such as grimacing and guarding.

Related Terminology

Several other terms related to pain merit attention and definition. Nociceptors are sensory end organs that are preferentially sensitive to tissue trauma (Figure 1). They respond to mechanical, thermal, electrical, and chemical stimulation that damages, or threatens to damage, tissue. Chemical changes in their immediate environment, such as those that occur with inflammation, can sensitize nociceptors and lower their thresholds for firing. Their activation occasions nonconscious neural signaling, termed 'nociception', which causes reflexes at the level of the spinal cord along with activation of multiple structures in the brainstem, the limbic (or emotional) brain, and the cortex.

Most pain results from nociception, but pain can also arise from damaged neural structures that generate abnormal neural transmission. Pain arising from damaged nerves or nervous structures is neuropathic pain. Patients with this diagnosis often misidentify the origin of the pain, believing

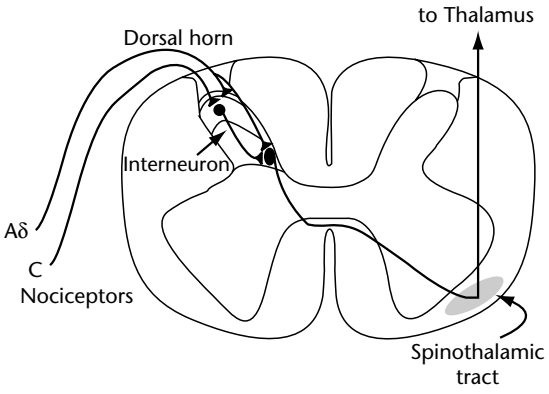


Figure 1. Cross-section of the spinal cord showing nociceptive primary afferents, interneurons, and the spinothalamic tract.

that it comes from a healthy part of the body, although it actually comes from damage to a neural pathway located between a particular body area and the brain. In some cases pain can occur because of damage to the spinal cord or the brain itself. This type of central pain constitutes a central pain state. In the past, physicians sometimes misidentified some central pain states as psychogenic, or arising from psychodynamic origins. True psychogenic pain is extremely rare.

The Relationship of Pain to Nociception

Although nociception is the normal and most common cause of pain, there is a surprisingly poor correspondence in clinical settings between the reported magnitude of pain and observable tissue trauma. This occurs because pain is a complex perception and many factors besides nociception contribute to the quality, expectation, and duration of pain. For example, the correlation between the extent of tissue trauma during surgery and the magnitude of pain that patients report afterwards is weak, and it is difficult to predict pain magnitude well from the depth or area of a burn injury.

When nociceptors become sensitized due to excessive stimulation or chemical irritation, they lower their thresholds for firing and increase their responses to previously painful stimuli. For example, a mild burn injury can cause the affected skin to generate strong pain to mild noxious irritation. This condition, characterized by excessive or exaggerated pain following a painful stimulus is termed 'hyperalgesia'. In some cases, nociceptors and related higher order structures can become so sensitized that they respond to harmless stimuli such as light touch. This is termed 'allodynia'. Hyperalgesia and allodynia distort the relationship between tissue damage itself and the amount of pain that the injured person perceives.

A few persons are born with a congenital absence of the ability to feel pain. Without medical supervision, such people tend to suffer grave injuries without complaint and die prematurely. As children, they may fracture bones to astonish playmates, take significant physical risks, and they may fail to notice and complain of some potentially life threatening disease conditions such as appendicitis. Such cases illustrate the protective function that pain serves in everyday life.

When pain is chronic, or long lasting, the relationship of pain and related disability to a nociceptive origin or neuropathy is often weak or absent. Millions of persons in industrialized countries are partially or wholly disabled by pain that, although vividly real to the patient, has no identifiable cause or is associated with negligible tissue damage. Conversely, various medical diagnostic procedures routinely reveal structural damage or other lesions that should cause nociception in patients who deny suffering pain. These observations indicate that chronic pain is more complex than simple awareness of nociception.

Clearly, pain and nociception are not synonyms. Nociception is nonconscious neural traffic originating with tissue trauma; it may or may not lead to a pain experience. Pain is a complex, unpleasant bodily awareness with sensory and emotional features, to which the person experiencing it ascribes meaning according to immediate context, past history, expectation, and culture. It may, or may not, indicate actual tissue damage, and the magnitude of the pain is rarely a good indicator of the clinical significance of the tissue damage causing the pain.

Analgesia and Pain Relief

Analgesia and pain relief are related but not identical concepts. The term 'analgesia' strictly denotes a complete inability to perceive pain when tissue damage occurs. In common use, however, it typically refers to a reduction in pain. An analgesic drug is one that reduces rather than fully eliminates pain. Pain relief refers to the reduction or elimination of a painful condition. A treatment might cause a person's headache to disappear (pain relief) but fail to cause analgesia for a painful event such as a needle puncture. Conversely, the injection of an analgesic drug might reduce a person's sensitivity to normally painful events like needle puncture but fail to reduce the headache.

ANATOMY AND PHYSIOLOGY OF PAIN

Detection of Tissue Injury

Nociceptors are the free nerve endings of either myelinated, small-diameter (Aδ) or unmyelinated (C) fiber axons (Figure 1). Aδ fibers conduct more rapidly and generate a different quality of sensation than the C fibers. Pinprick at the foot will first produce a fast pain with a bright, sharp quality that lasts no more than 50 milliseconds, followed by a painless interval of approximately one second, and then, a slow pain with a burning quality of roughly one-second duration. Aδ fibers mediate the fast pain and C fibers the slow pain. These nociceptive afferents supply skin, subcutaneous tissue, joints, periosteum, teeth, meninges, viscera, muscles, and blood vessels.

Most nociceptors are polymodal, responding to mechanical stimuli such as crush, extreme thermal changes, and chemical irritation. They encode localization and stimulus intensity. However, visceral nociceptors do not respond to cutting, pinching or burning. Instead, they respond to abnormal distension or contraction of the muscle walls of hollow viscera, the rapid stretching of capsules of solid organs, anoxemia of smooth muscles, traction or compression of vessels or ligaments, and chemical irritation. Pain originating in visceral nociception commonly manifests as a pain on the surface of the body (termed 'referred pain'), and sometimes the location of the pain is distant from the site of pathology. Ischemia in heart muscle, for example, causes pain that patients often experience in the pectoral muscles of the chest or as radiating down the left arm into the little finger and ring fingers.

With repeated noxious stimulation, nociceptors may sensitize, lowering their threshold for firing, decreasing their response latency, and persisting in their firing after a stimulus event terminates. The site of injury itself becomes extraordinarily sensitive, or hyperalgesic, and surrounding, uninjured tissue becomes increasingly sensitive as a result of central nervous system changes. This is called 'peripheral sensitization'.

In most cases of natural injury, inflammation develops at the site of the injury as a protective response. Increased blood flow produces redness and heat, and increased vascular permeability causes local swelling. The production of various mediating substances including bradykinin and certain types of prostaglandins can sensitize, and sometimes directly stimulate, nociceptors. Sensitization of nociceptors is a major factor in many clinical pain states.

The dorsal horn and spinal transmission

At the dorsal horn of the spinal cord, primary nociceptive afferents synapse with the first cells of spinal transmission, that are mainly wide dynamic range (or multireceptive) neurons and nociceptive-specific neurons (Figure 2). The dorsal horn is much more than a relay station. Its complex processing involves selection, modulation, and integration of nociceptive signals. From the dorsal horn, the spinal cord conveys signals to many destinations within the brain: thalamus, hypothalamus, mesencephalon, and reticular formation.

The spinothalamic tract

The thalamus receives and integrates nociceptive and other signals, encoding information about type, temporal pattern, intensity, and localization of tissue trauma (Figure 2). It interacts bidirectionally with both limbic and cortical structures such as the somatosensory cortices. The spinothalamic pathway with its cortical connections makes possible the sensory features of pain.

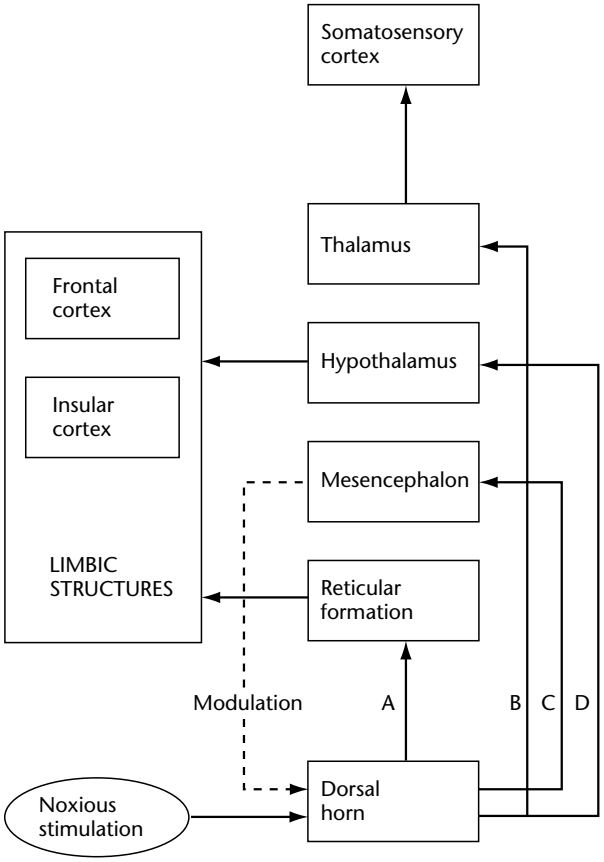


Figure 2. Pathways for the transmission of nociceptive signaling. A, spinoreticular pathway; B, spinothalamic pathway and thalamocortical pathway; C, spinomesencephalic pathway; D, spinohypothalamic pathway. (Based on Willis and Westlund, 1997.)

The spinohypothalamic tract

The spinohypothalamic tract conveys nociceptive messaging to the hypothalamus, which is effectively the control center of the limbic brain (Figure 2). Messages of tissue trauma can provoke the periventricular nucleus of the hypothalamus to initiate a stress response. This is a systematic, adaptive pattern of neural and endocrine activation and behavioral changes orchestrated within the hypothalamic–pituitary–adrenocortical axis. The stress response creates arousal and protective conditions that maximize an individual's ability to respond to threat by flight or fight. A nociceptive stressor can cause the release of multiple hormones that include

adrenocorticotropic hormone, cortisol, the catecholamines (epinephrine, norepinephrine, and dopamine), and renin. The catecholamines produce arousal, cortisol increases blood glucose and decreases inflammation, and the renin-angiotensin system redirects blood flow to the brain and heart during an emergency. In these and other ways, the stress response helps the individual deal with a short-term injurious situation.

Unfortunately, when nociception continues over a long period time, as it can in a patient with a painful cancer, the stress response has numerous negative consequences. Patients develop fatigue and hypersensitive muscles. Sleep is disturbed and rarely restorative. Appetite and interest in sex diminish greatly, as does the ability to sustain mental concentration and attention. Extended exposure to cortisol can damage brain structures and accelerate brain aging. Finally, an extended stress response can suppress immune function and increase vulnerability to adventitious infection. In these respects, prolonged nociceptive signaling exerts an insidious effect.

The spinomesencephalic tract

The mesencephalon contains structures that contribute to descending modulation of nociceptive signaling (Figure 2). These include the nucleus raphe magnus, the periaqueductal gray, and the solitary nucleus. Activation of the pathway descending from these and related structures can result in the attenuation of nociceptive traffic at the dorsal horn of the spinal cord. Direct electrical stimulation of these areas, administration of opioid drugs, and endogenous opioid-like compounds (endorphins) can create analgesia via this pathway. Moreover, increases in blood pressure provoke the carotid baroreceptors, which in turn activate the solitary nucleus, which triggers the descending nociception-inhibiting pathway. The result is reduced nociceptive traffic from the dorsal horn to the brain.

The spinoreticular pathway

The spinoreticular pathway plays a role in the emotional dimension of pain (Figure 2). Nociceptive signals reach multiple structures including the locus coeruleus, which sends extensive noradrenergic projections to diffuse areas of the limbic brain, the hypothalamus, and the cerebellum. One of the projections originating in the locus, the dorsal noradrenergic bundle, extends to many limbic and cortical areas. Activation of this pathway tends to produce hypervigilance, negative emotional arousal, as well as behavior consistent with anxiety and threat.

Pain and Functional Brain Imaging

An extensive literature on functional brain imaging and pain reveals that, during the experience of pain, people show metabolic activity in the thalamus, the lenticular nucleus (a limbic brain capsule surrounding the thalamus), the insular cortex (a cortical area related to emotion), the cingulate cortex, the anterior cingulate cortex, the somatosensory cortex, frontal areas, the vermis of the cerebellum, and other areas related to motor function. Although a few studies have detected activity in hypothalamus, by and large it is difficult for current technology to detect activity in structures at this level and below. Basically, it has become clear that pain emerges from complex, parallel processing in structures that otherwise contribute to emotional arousal, somatosensory awareness, and cognition.

MEASUREMENT OF PAIN

Progress in the scientific study of pain requires sound measurement. However, pain is a private experience, and direct objective measurement is impossible. Moreover, because of the weak correlation between pain and nociception, physiological correlates cannot serve as reliable proxy measures. Researchers and clinicians must rely upon subject or patient introspection for pain measurement. This constraint means that pain measurement is not currently possible in some patient populations.

Simple pain measures treat pain as a single, unidimensional phenomenon and they scale it directly. Subjects rating some feature of pain such as its intensity on a numerical scale ranging from 0 to 10, mark a 10-centimeter line anchored at each end with words such as 'No Pain' and 'Worst Pain Imaginable', or choose one of several descriptive categories that vary in magnitude. Some investigators attempt to scale the sensory and emotional features of pain separately by administering two simple pain measurement tools, one for sensory awareness and the other for emotional awareness.

Many questionnaires exist for measuring pain indirectly and on multiple dimensions. Some approaches ask subjects to choose words describing the qualities of pain from lists that reflect gradations on sensory, emotional, or cognitive dimensions. Others address the ways in which pain interferes with normal function or aspects of daily living. These instruments score pain along several

dimensions and often assess the impact of pain on the activities of daily living.

Pathological Pain

Neuropathic pain

Although nociception has a deleterious impact on health and well-being, it is, in principle, a healthy process if it arises within an intact nervous system. The nervous system is responding as it should to injury. Pain is pathological when it arises from damage to peripheral nerves or central nervous structures, i.e. when it is neuropathic. With rare exceptions, neuropathic pains are chronic, difficult to diagnose, and resistant to most conventional therapies. In many cases, they respond better to drugs otherwise indicated for epilepsy than to conventional analgesic medications such as opioids. Examples of neuropathic pains include postherpetic neuralgia, a pain syndrome that can occur following a shingles attack, and tic douloureux, a facial pain characterized by jolting, painful paroxysms that occur in response to light touch that is normally painless.

Acute and chronic pain

Many clinicians distinguish between acute and chronic pain. Acute pain begins with tissue trauma, diminishes as healing progress, and disappears when healing is complete. Pain after surgery or a broken bone is acute pain. The primary mechanism of most acute pain states is inflammation-induced sensitization of nociceptors.

Chronic pain has two forms. First, when healing never occurs, as in a degenerative disease like arthritis or a progressing disease like cancer, pain becomes chronic because tissue trauma is chronic. Neuropathic pains fall into this category. Second, pain may persist beyond the healing of an injury or disease and continue for an indefinite period, or it may arise from an uncertain or nonspecific origin and continue indefinitely for no apparent reason. Current evidence suggests that pain can become chronic in this sense if persisting nociception brings about long-lasting changes in the function and structure of the central nervous system. This process, called 'neuroplasticity', accounts for the resistance of chronic pain to conventional pain treatments and invasive interventions such as nerve blocks or neurosurgery.

Because persisting pain can alter the neurophysiology of the patient, impose a somatic preoccupation, degrade the patient's lifestyle, and impair general social adjustment, chronic pain requires a different and more thorough clinical evaluation process than acute pain. Many clinics employ multidisciplinary teams that combine psychosocial and medical approaches to evaluation and pain management. The goal of management is not simply to relieve the pain but to restore normal function and reintegrate the patient into a normal lifestyle. Often, it is necessary to include physical rehabilitation and psychological support in the treatment.

SUMMARY

Pain is an unpleasant, subjective, bodily awareness normally experienced as tissue damage or attributed to tissue damage. It serves an important protective function in everyday life. The most common mechanism of pain is nociception, the nonconscious neural signaling of tissue damage, although it may also arise from damage to neural pathways or structures. Various physiological processes can amplify or attenuate nociceptive traffic within the nervous system. In clinical settings, there is a poor relationship between measures of tissue damage and the magnitude of the pain experienced.

Pain is a complex perception with sensory and emotional features, and it depends heavily upon cognitive processes such as attention, expectation, memory, belief, interpretation of the immediate situation, and personal meaning. Functional brain imaging studies of people in pain demonstrate extensive parallel processing in brain areas otherwise associated with somatosensory awareness, negative emotion, and cognition. Measurement of pain is difficult because pain is subjective and highly individual. Because no objective measure of pain exists, introspection is the only window available to researchers and clinicians.

Acute and chronic pain differ, and chronic pain is a costly problem in all industrialized nations because it causes disability. Often chronic pain has a weak or absent relationship to a definable cause and resists invasive treatments that target simple mechanisms. Evidence suggests that persisting nociception can change structure and function within the nervous system to make pain a self-sustaining process.

Further Reading

Casey KL and Bushnell KC (eds) (2000) *Pain Imaging*. Seattle: IASP Press.

Hanssen PT, Fields HR, Hill RG and Marchettini P (eds) (2001) *Neuropathic Pain: Pathophysiology and Treatment*. Seattle: IASP Press.

Loeser JD, Butler SH, Chapman CR and Turk DC (eds) (2001) *Bonica's Management of Pain*, 3rd edn. Philadelphia: Lippencott, Williams and Wilkins.

Turk DC and Melzack R (eds) (1992) *Handbook of Pain Assessment*. New York: Guilford Press.
Wall P (2000) *The Science of Suffering*. New York: Columbia University Press.
Willis WD and Westlund KN (1997) Neuroanatomy of the pain system. *Journal of Clinical Neurophysiology* **14**: 2–31.

Panpsychism

Intermediate article

William Seager, University of Toronto, Toronto, Ontario, Canada

CONTENTS

Panpsychism is the doctrine that mind, *in some sense of the term, is everywhere, in some sense of that term. It regards mind as a fundamental and ubiquitous feature of nature.*

BRIEF HISTORY

While panpsychism's origins precede any records of systematic philosophy and probably spring from very early forms of animism, any substantial doctrine of panpsychism had to await the idea that a naturalistic account of the world is possible, for only from the point of view of some such account can the issue of mind's place within the natural world arise. It is possible to trace debate between panpsychist and emergentist views of mind as early as the proto-scientific accounts of the pre-Socratic philosophers, dating from the fifth century BCE.

The problem of mind faded with the demise of this early proto-naturalism, and the debate between panpsychism and emergentism would not arise again until the scientific revolution. In the sixteenth and seventeenth centuries a plethora of mind–body theories suddenly appeared, as the issue of the place of mind within a world governed by natural law became unavoidable and pressing.

Rejecting the infamous Cartesian dualism, Baruch Spinoza (1632–1677) and Gottfried Wilhelm Leibniz (1646–1716) espoused important and quite distinct forms of panpsychism. Spinoza regarded both mind and matter as simply aspects of the eternal, infinite, and unique substance he identified with God, or nature; there is nothing that does not have a mental aspect – the proper appreciation of matter reveals it to be the other side of a mentalistic coin.

Leibniz's view is sometimes caricatured as: Spinoza with infinitely many substances rather than only one. These substances Leibniz called 'monads', which are characterized as spiritual (i.e. mental) automata. What is of special interest is that unlike Spinoza, Leibniz maintained a difference between things with mental attributes and things without by differentiating between a 'mere aggregate' and 'organisms'. Consider a heap of sand. It corresponds to a set of monads, but no monad represents the 'point of view' of the heap. By contrast, your body also corresponds to a set of monads, but one of these monads – the dominant monad, your mind – represents the point of view of this biological system. This idea is retained in modern versions of panpsychism, and undercuts the rather simpleminded objection that stones do not seem to have minds.

The growth of idealism through the eighteenth and nineteenth centuries meant that panpsychism became, in effect, the default philosophy but with a decided bias resulting from positioning the mental as the primary component of reality. Panpsychism was favored by a surprising number of eminent scientists and philosophers of the time, perhaps most curiously by three men now better known as the founders of scientific psychology – Gustav Fechner (1801–1887), Rudolf Lotze (1817–1881), and Wilhelm Wundt (1832–1920). Other prominent panpsychists of the period include William James (1842–1910) and William Clifford (1845–1879).

The richest development of panpsychism in the twentieth century is that of Alfred North Whitehead (1861–1947). Whitehead proposed radically reforming our metaphysical view of the world, replacing things and matter with events and the ongoing processes of their creation. His panpsychism, which has many similarities to that of Leibniz, arises from regarding these elementary events as possessing mentality in some attenuated sense, expressed in terms of creativity, spontaneity, and perception. Perhaps unfortunately, Whitehead's panpsychism stands or falls with his entire metaphysical system, which entails an even more radical revision of our current scientifically based views than does bare panpsychism.

ARGUMENTS FOR PANPSYCHISM

Arguments for panpsychism can be grouped under three broad headings: analogical, genetic, and

intrinsic nature. The basic analogical argument goes like this: if we look closely we see that even the simplest forms of matter exhibit behavior which is akin to that which we associate with mentality in animals and human beings. Therefore we ought to regard these simple forms as possessing mentalistic attributes. Unfortunately the premise seems quite preposterous. Better hope for an analogical defense of panpsychism springs from the overthrow of determinism in physics. Some modern panpsychists, starting with Whitehead, see this indeterminacy as an expression not of blind chance but of free action, in response to a kind of information-based inclination rather than mechanical causation. The question is whether the source and target phenomena are sufficiently analogous to warrant extending attributes from the one domain to the other. This is doubtful here. The indeterminacy of modern physics seems to be pure randomness quite remote from deliberation, decision, and indecision. Another analogical argument which draws upon quantum physics is more promising. This analogy is between between consciousness and information. It is natural to think that a prime function of consciousness is the integration of information and the monitoring of various external and internal states. Thus, if information monitoring is a fundamental and pervasive feature of the world at even the most basic levels, then perhaps this indicates that consciousness also exists at those levels. It is undeniably suggestive that one of the central features of quantum mechanics is the existence of informational but noncausal relations. These relations are noncausal insofar as they are modulated instantaneously over any distance and do not involve the transfer of energy between the parts of the system. But they are informational in the sense that the changes of state of one part of the system seem in some way to be communicated to the other. While these so-called entangled states are normally susceptible to rapid 'decoherence' caused by environmental disturbance, some systems might resist decoherence and it has been conjectured that these are the physical foundations of consciousness (see Hameroff and Penrose, 1996). Furthermore, the decoherence argument evidently collapses for the universe as a whole which by definition cannot be disturbed by any outside force, so presumably the total universe is in one immensely complex, entangled state. Given a link between consciousness, monitoring, and information exchange, this leads to a view highly reminiscent of Leibniz's, with centers of (perhaps rudimentary) consciousness at the foundation of the world.

Turning to the genetic arguments, Darwin's theory of evolution led in the nineteenth century to a popular empirical genetic argument for panpsychism based upon the idea that no novel property could arise by natural selection from systems entirely devoid of it. Clifford puts the argument thus: 'we cannot suppose that so enormous a jump from one creature to another should have occurred at any point in the process of evolution as the introduction of a fact entirely different and absolutely separate from the physical fact' (1874/1886, p. 266). An *a priori* form has been advanced by Thomas Nagel (1979). Nagel links panpsychism to a necessary failure of emergentism: namely, that emergence lacks the status of a true metaphysical relation. Nagel says: 'there are no truly emergent properties of complex systems. All properties of complex systems that are not relations between it and something else derive from the properties of its constituents and their effects on each other when so combined' (p. 182). Thus the only coherent form of emergentism is a merely epistemological doctrine about the limits to our understanding of complex systems. Panpsychism follows from Nagel's denial of reductionism which precludes simply identifying mental properties with complex physical properties. Then, since we can build an enminded system out of 'any matter', mind must be associated with matter in general and in its most fundamental forms.

Another argument depends on the idea that every fundamental kind of thing must have an intrinsic nature. The objects of physics, it is claimed, are described in purely dispositional terms. That is, while an electron, for example, is said to possess 'spin', all this amounts to is that the electron has certain behavioral dispositions. It is arguable that dispositions must be grounded in some intrinsic, nondispositional attributes, but we have no conception whatsoever of what the intrinsic nature of matter might be. In fact, the only intrinsic nature with which we are familiar is consciousness itself. The qualities of conscious experience (the smell of a rose, the taste of a strawberry, etc.) seem not to be reducible to relations among nonexperiential states, nor entirely specifiable without remainder in terms of their causal powers. They seem instead to possess intrinsic and irreducible characteristics. If this is the only idea of intrinsic nature we possess, and matter must be assigned some intrinsic nature, it seems that matter must be granted a mentalistic intrinsic nature. The core idea of this argument can be traced back to Leibniz who felt forced to ascribe mentalistic attributes to his monads as the only possible feature which could

account for the active forces demanded by an adequate physics. In his discussion of this difficulty, Whitehead describes all 'modern cosmologies' as having to admit a 'mysterious reality in the background, intrinsically unknowable' and notes that Leibniz 'explained what it must be like to be an atom'. Although far from demonstrative this is, in the words of Thomas Sprigge (1999), 'a hypothesis worth exploring as the only alternative to saying that matter is unknowable in its inner essence'.

ARGUMENTS AGAINST PANPSYCHISM

None of the above arguments are demonstrative and arguments against panpsychism are not hard to devise. The most obvious is the lack of evidence that fundamental physical things possess any mental attributes. This does not seem particularly compelling since not every attribute is apparent at every scale; gravitation is fundamental and ubiquitous yet individual elementary particles do not appear to gravitate. The presumed causal completeness and closure of the physical world undercuts panpsychism insofar as it renders superfluous the postulation of mentalistic attributes to fundamental physical entities. This objection is closely associated with the more strictly methodological protest that since there is no apparent need to assign any mentalistic attributes to the basic physical features of the world, panpsychism is a purely metaphysical extravagance with no empirical content whatsoever. While compelling for most modern naturalists, in the absence of a successful scientific account of the emergence of mind in terms of a purely physical substrate these points come close to begging the question. To the extent that one worries that consciousness presents an especially difficult and at present intractable problem for physicalism, one cannot legitimately deploy the complete physicality of the world against a panpsychist position. There are also internal difficulties with panpsychism, of which the most serious and interesting is the 'combination problem': even granting that fundamental physical entities have mental properties, there remains the question of how the more complex minds of composite entities (such as ourselves) emerge out of the protominds of their constituents. This problem reveals that some forms of panpsychism require their own doctrine of emergence. If emergence is required even within panpsychism, then one might be excused for preferring mind to emerge from a nonmental substrate rather than the mentalistic base posited by panpsychism. This objection also reveals that any successful panpsychism will have to grant genuine consciousness to the fundamental constituents of the world; otherwise panpsychism will be claiming that consciousness can emerge from the nonconscious, in which case the advantages of a physicalist emergence seem quite clear. And the claim that the fundamental elements of the world are truly, if minimally, conscious further erodes the plausibility of panpsychism itself.

PANPSYCHISM AND COGNITIVE SCIENCE

Panpsychism is a metaphysical doctrine with no very direct relation to any science. There is no test that could decisively confirm or refute it, no more than physicalism can be proven scientifically. Yet metaphysical views form an indispensable background to all science. They integrate our world view and allow us to situate our scientific endeavors within a larger framework. Science has always informed metaphysical speculation, and in return such speculation helps motivate and pave the way for new science. The primary scientific benefit of panpsychism is to leave open certain pathways of speculation that may yet inform future theories of mind and consciousness.

References

Clifford WK (1874) Body and mind. Reprinted in: Stephen L and Pollock F (eds) *Lectures and Essays*. London: Macmillan, 1886.

Hameroff S and Penrose R (1996) Conscious events as orchestrated spacetime selections. *Journal of Consciousness Studies* **3**(1): 36–53.

James W (1890/1950) *The Principles of Psychology*, vol. 1. New York: Henry Holt and Co. [Reprinted in 1950, New York: Dover.]

Nagel T (1979) *Mortal Questions*. Cambridge: Cambridge University Press.

Seager W (2001) Panpsychism. In: *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu.

Sprigge T (1999) Panpsychism. In: *Routledge Encyclopedia of Philosophy*. London: Routledge.

Further Reading

Chalmers D (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Edwards P (1967) Panpsychism. In: *The Encyclopedia of Philosophy*. New York: The Free Press.

Griffin D (1998) *Unsnarling the World Knot: Consciousness, Freedom and the Mind–Body Problem*. Berkeley: University of California Press.

Hartshorne C (1950) Panpsychism. In: Ferm VTA (ed.) *A History of Philosophical Systems*. New York: Philosophical Library.

Perception, Direct

Intermediate article

Alva Noë, University of California, Santa Cruz, California, USA

CONTENTS

The theory of direct perception holds that perception, in any sensory modality, is a form of non-inferential awareness of the sorts of things we normally take ourselves to perceive. This view rejects the idea, common to many philosophical and scientific approaches to perception, that our only immediate contact is with sensations, impressions, or mere patterns of stimulation of the sensory receptors.

INTRODUCTION: WHAT IS DIRECT PERCEPTION?

The claim that perception is direct has been advanced by a small but distinguished minority of philosophers and psychologists, including Austin, Gibson, McDowell, Neisser, Putnam, Strawson, and perhaps also Aristotle, James, and Reid. Perception, it is argued, is a form of noninferential awareness of the sorts of things that we normally take ourselves to be aware of when we perceive, such as everyday objects and events. Supporters of direct perception reject the idea that in perception we are aware only of mental intermediaries – sense data, impressions, appearances – and that it is only thanks to our direct awareness of these that we can be said to be aware (indirectly) of the world. They accept that perception is a form of direct access to the world, and, therefore, that the world is very much the way it seems to us in perception. For this reason, defenders of direct perception are sometimes known as naive or direct *realists*.

In fact, there is nothing naive about the theory of direct perception. Although it attempts to defendsomething like 'the standpoint of common sense', it is in fact a sophisticated response to the widely held view that perception could not be, inthe relevant sense, direct. For this reason, the best way to understand direct perception is to examine the family of arguments to which it is a response.

ARGUMENTS AGAINST DIRECT PERCEPTION

We can identify a basic line of argument against direct perception which runs as follows. When you see (say) a round plate held out at an angle, you see that it is round. But its roundness is, strictly speaking, inaccessible to your visual point of view. What is really given to you visually, at least as far as shape is concerned, is a glimpse of an elliptical profile. From the fact that the plate presents you with an elliptical profile, however, it cannot be deduced that the plate is round. From this it would seem to follow that perception is, in the first instance, a form of contact not with things as they are, but rather with mere glimpses, or impressions, or appearances of things. Our perceptual contact with the plate itself, so continues this line of thinking, is mediated by our more direct contact with the plate sense datum (or appearance, impression, etc.). Perceptual judgments, then, must go well beyond what is actually given in experience and must be thought of not so much as immediate records of how things are, but as the results of conjecture or speculation.

Philosophers refer to this general line of argument as the *argument from illusion* (Ayer, 1940). It purports to show that we are, in perceptual experience, immediately aware not of what we think we perceive, but of mental intermediaries that in some way stand for or refer to those things. According to the argument from illusion, when you see a red tomato, for example, you are aware not of the tomato itself, but of a red, tomato-like sense datum. The reasoning is as follows. When you hallucinate a red tomato, you are not aware of a red tomato, but merely of a tomato-like sense datum. After all, it is a hallucination. However, the experience of actually seeing a red tomato and merely hallucinating one are qualitatively indistinguishable. If this were not so, then we would never be deceived by our hallucinations. But if the

experiences are qualitatively indistinguishable, then you are aware of one and the same thing when you see a red tomato and hallucinate a red tomato. Hence, when you see a red tomato, you are aware not of a tomato but of a tomato-like sense datum.

That perception is in this way *in*direct appears to gain support from basic facts about the physics of perception. When you see a tomato you do not make direct contact with it. At best you make contact with the tomato only as mediated by a complicated causal process – the tomato affects the light which gives rise to a pattern of stimulation of the receptors in the eyes which in turn produces activity in the optic nerve and brain. At the terminus of this process there is the visual experience as of a tomato. The tomato, it should be clear, enters the process only as a more or less remote cause of the experience one eventually undergoes. Similar points can be made in the case of other sensory modalities, such as touch, which might seem to be direct in precisely the way that vision is not. For example, when you hold an object (a bottle, say) in your hands, you do make contact with the bottle, but all that is recorded in your immediate experience of the bottle is a pattern of nervous stimulation in your hands (or a pattern of sensation). Just that pattern could be produced by something other than a bottle. That it is produced by the bottle is not something your bottle-experience itself can guarantee. The direct object of your tactile experience is not the bottle, then, but a bottle-like tactile impression.

We have been considering a philosophical line of argument, but it is one that has been entrenched in scientific theories of perception. The central problem for perceptual science has traditionally been that of explaining how it is that we perceive what we do – the three-dimensional world of independent objects, etc. – on the basis of the information available in the form of stimulation of the sensory organs. What makes the problem a difficult one is the fact that, strictly speaking, the data for perception, in the form of the pattern of stimulation of the sensory organ, are highly limited and impoverished. In general it is not possible to infer the character of the environmental layout from the pattern of stimulation of the sensory organ. This comes out, for example, in the fact that a small tomato nearby and a large tomato farther away may project a qualitatively indistinguishable retinal image. The dominant strategy for addressing the traditional problem of perception – what Fodor and Pylyshyn (1981) have called the 'Establishment View' – is to suppose that the brain produces the perceptual experience by engaging in a constructive process of inference or conjecture. A perception, in the phrase of Helmholtz, is an 'unconscious inference' (Helmholtz, 1855). Empirical research on perception focuses on understanding the mechanisms, neural and psychological, that make up the brain's ability to perform this constructive feat. (See, for example, the important work of Marr, 1982.) Much work on perception undertaken in cognitive science since the 1970s is constructive in this way and rests squarely on a conception of perception as *indirect*.

Note, the upshot of these criticisms of direct perception is at once epistemological and also metaphysical. Epistemologically, the argument casts perception as enabling us to have immediate knowledge of how things are with us perceptually, but not with how things are outside of us. This approach to perception is thus at one with the so-called Cartesian conception of mind, according to which the domain of the mental, of consciousness, is that domain within which we have a certain kind of immediate self-knowledge. The metaphysical upshot concerns the nature of perceptual experience itself. Perceptual experiences are internal states whose fundamental character is independent of how things are in the world.

DEFENSES OF DIRECT PERCEPTION

The remainder of this article reviews two main lines of defense of direct perception, one philosophical and one scientific.

Criticisms of the Conception of Experience Implicit in the Argument from Illusion

Several authors have attacked the conception of experience implicit in the Argument from Illusion and in related lines of argument against direct perception. These attacks fall into roughly three groupings.

Austin's criticism

Austin (1962) argued that from the fact that we are sometimes deceived by a hallucinatory experience, it does not follow that the hallucinatory experience and its veridical counterpart are qualitatively indistinguishable. So, for example, one might mistakenly judge a straight stick partially submerged in water to be bent. A straight stick in water, after all, may look bent. But it does not follow from this that there is no qualitative difference between the experience of a genuinely bent stick and that of a

bent-looking stick standing partially submerged in water. In the one case, but not the other, for example, there is the difference owing to the presence of water! But if there is no qualitative identity between the experiences, then there is no reason to suppose that the object of awareness is the same for both experiences, and so the Argument from Illusion collapses. Austin further remarked that the assumption that we are aware of anything, when having an outright hallucination, is itself quite gratuitous. So, for example, in the case of the stick standing partially submerged in water, there is nothing bent of which we are aware.

The highest common factor conception

Snowdon (1980–1981) and McDowell (1982, 1986), developing an idea first proposed by Hinton (1973), propound a line of criticism very similar to that of Austin. Like Austin, they challenge the assumption that there is a single experience common to both a perception and its corresponding hallucination. They reject what McDowell has called 'the highest common factor conception' of perceptual experience. It is true that perceivers may not be able to tell, by mere introspection, whether they are perceiving or merely hallucinating. But from this it does not follow that perceivers and hallucinators are in one and the same experiential state. At best all that follows is that, for all they can tell by introspecting, they might be in the same state. In the veridical case, they undergo an experience of something's looking a certain way to them (let's say). But in the hallucinatory case they are not in that state. Rather they are in the state of its merely seeming to them as if something looks a certain way.

In rejecting the highest common factor conception, McDowell and Snowdon reject, like Austin, the grounds for believing that what we are aware of when we perceive is just what we are aware of when we undergo the corresponding hallucination. This line of criticism of the Argument from Illusion is noteworthy because it breaks with the Cartesian idea that the contents of consciousness, including our experiential states, are immediately and certainly available to our introspection. McDowell and Snowdon break with this enduring epistemological notion by embracing a so-called *externalist* metaphysics of mind according to which perceptual experiences (and other mental states) are constituted by relations between perceivers and their environments.

The intentionality of perception

Sellars (1956) and Strawson (1979) have developed a Kantian line of criticism of the Argument from Illusion based on the observation that perceptual experiences are intentional (in the philosopher's sense). Perceptual experience, they reason, is *always* and *essentially* experience as of things being this way or that, and only a creature in possession of the concepts needed to capture in thought how the experience represent things as being could be said to have full-fledged perceptual experience. We have no acquaintance with our experience, they hold, other than as experience as of things being this way or that. Hence, we can make no sense of the idea that experience, properly described, pertains not to the world but only to our sensory impressions.

This Kantian line of criticism attempts to undercut the Argument from Illusion by depriving its proponents of the needed conception of experience as an awareness of something less than the mind-independent world. The criticism rests, at base, on a phenomenological claim. We *mis*describe what our experience is really like if we attempt to describe it in a merely sense-datum idiom. The line of criticism has an important epistemological upshot. It is only by recognizing the conceptual, articulate, propositional character of perceptual experience that we can appreciate the fact that our experiences give us reasons for judgment and belief.

Gibson's 'Ecological' Approach

Gibson's (1966, 1979) defense of direct perception takes as its start the rejection of the way traditional theorists frame their basic problem. Vision, Gibson argued, is not something that takes place in the eye and brain of a perceiver. We misdescribe vision if we think of it as a process whereby the brain builds up an internal model of the environment on the basis of limited sensory stimulation. Such a conception of the nature of vision, Gibson argued, is pitched at the wrong level. The perceiver is not the brain, but rather the whole animal embedded in an environment. The function of perception is not the production of experiences or representations but rather the enabling of the animal to function appropriately in the environment. The seeing, Gibson argued, takes place in the environment thanks to the engagement of the whole animal with its surroundings. The information available to us in vision, then, is not the pattern of irradiation encoded on the surface of the retina. It is the environment itself – the animal's habitat – that is the repository of information about it. Vision, on this proposal, is a way of acquiring information about the environment by coming into direct

contact with the environment thanks to active exploration.

Gibson's bold claim is that if we reformulate our analysis of the basic visual predicament this way, the puzzling character of how we see disappears. Reconsidering our example of the plate, we can notice that although it is true that the round and the elliptical plates might look the same when contemplated from a certain stationary point of view, their different natures will be readily apparent to the active, moving animal.

In this way Gibson believed that many of the great puzzles of visual science – such as that of how we perceive three dimensions – are in fact artefacts of a misdescription of what vision is.

DIRECT PERCEPTION IN CONTEMPORARY PHILOSOPHY AND COGNITIVE SCIENCE

The debate about direct perception is ongoing. It takes place at the foundations of contemporary thought about perception. For science, as we have seen, what is at stake is our basic analysis of what vision is and our account of the level at which perceptual phenomena are to be studied. For philosophy, what is at stake is our understanding of the nature of experience, and of the role experience can play as a source of knowledge. Beyond technicalities, the basic question of direct perception – what do we perceive, really? – is important because it gives expression to an enduring problem for philosophy: how are we to reconcile what science teaches us with what we know, or seem to know, on the basis of experience?

References

Austin JL (1962) *Sense and Sensibilia*. Oxford, UK: Clarendon Press.

Ayer AJ (1940) *The Foundations of Empirical Knowledge*. London, UK: Macmillan.

Fodor JA and Pylyshyn Z (1981) How direct is visual perception? Some reflections on Gibson's 'ecological approach'. *Cognition* **9**: 139–196.

Gibson JJ (1966) *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Helmholtz H (1855) *Über das Sehen das Menschen. Ein popular wissenschafticher Vortrag gehalten zu Königsberg in Prussia zum Besten von Kant's Denkmal am 27. Februar 1855*. Lelpzig: L. Voss.

Hinton JM (1973) *Experiences*. Oxford, UK: Oxford University Press.

Marr D (1982) *Vision*. New York, NY: W.H. Freeman.

McDowell J (1982) Criteria, defeasibility and knowledge. *Proceedings of the British Academy* **68**: 455–479.

McDowell J (1986) Singular thought and the extent of inner space. In: Pettit P and McDowell J (eds) *Subject, Thought and Context*. Oxford, UK: Oxford University Press.

Sellars W (1956) Empiricism and the philosophy of mind. In: Feigl H and Scriven M (eds) *Minnesota Studies in the Philosophy of Science*, vol. 6. Minneapolis, MN: University of Minnesota Press.

Snowdon P (1980–1981) Experience, vision and causation. *Proceedings of the Aristotelian Society* **81**: 175–192.

Strawson PF (1979) Perception and its objects. In: MacDonald GF (ed.) *Perception and Identity: Essays Presented to A. J. Ayer with His Replies*. Ithaca, NY: Cornell University Press.

Further Reading

Anscombe GEM (1965) The intentionality of sensation. In: Butler RJ (ed.) *Analytical Philosophy, Second Series*. Oxford, UK: Oxford University Press.

Bruce V and Green P (1985) *Visual Perception: Physiology, Psychology and Ecology*. London and Hillsdale, NJ: Lawrence Erlbaum Associates.

Michaels CF and Carello C (1981) *Direct Perception*. New York, NY: Prentice Hall.

Neisser U (1967) *Cognitive Psychology*. New York, NY: Prentice Hall.

Putnam H (1999) *The Threefold Cord: Mind, Body, and World*. New York, NY: Columbia University Press.

Turvey MT, Shaw RE, Reed ES and Mace WM (1981) Ecological laws of perceiving and acting: in reply to Fodor and Pylysyhn (1981). *Cognition* **9**: 237–304.

Perception, Philosophical Issues about

Intermediate article

Austen Clark, University of Connecticut, Storrs, Connecticut, USA

CONTENTS

Three major philosophical issues that have arisen from the scientific study of perceptual processes: the variations and limitations of representational systems, the localization of perceptual function in neural tissue, and the relations to verbal outputs of people introspecting.

CENTRAL PHILOSOPHICAL ISSUES ABOUT PERCEPTION

Broadly speaking, philosophers work on three kinds of task. The first consists of all the work one must do to a question before one can say anything sensible about it. The second is to describe the relations between answers to one set of questions and answers to another. The third, and most difficult, is to reconcile the internal tensions that arise when one tries to complete the second task.

The first and third of these kinds of problem are studied almost exclusively within departments of philosophy. Applied to perception, our first kind of task includes questions such as: Should I believe my senses? Can one prove the existence of the external world? Is it fair to treat the existence of the table in front of me as just a plausible hypothesis? And so on. Such questions of epistemology have been studied within Western philosophy more or less continuously since the ancient Greeks. They are of immense importance, and well worth studying. But one must have already made one's peace with them before even starting the enterprise of 'cognitive science' – or any science, for that matter – so a discussion of them does not belong in this article.

The third kind of task is likewise interior to philosophy. In the attempt to come up with some systematic overview of our intellectual landscape, various schools and traditions have arisen over the centuries. They have approached this project in different ways, and between them there are arguments about how to argue, and questions about the enterprise of raising questions. Within perception, this third category of problem includes such questions as: What sort of logic is required in order to describe the things one seems to see in an illusion (Hintikka, 1969)? Do those things actually exist in some other possible world, or merely seem to? Can they cause things to happen in this world? And so on. These are the sorts of questions that greatly preoccupy analytic philosophers, but will appeal to few cognitive scientists.

So this article will confine itself to questions of the second kind: questions about the relations between different portions of the scientific study of perception. Within cognitive science itself there are problems about perception that are philosophical problems – problems that arise as soon as one steps back from some particular research question, and asks instead: how would the answer to this question cohere, or fail to cohere, with our answers to other questions? One strives for a pleasing and coherent overview of how it all hangs together; but one finds places where the landscape is not at all settled, peaceable, or pleasing. Instead one finds fault lines, where vast tectonic plates grind remorselessly, deep underground, producing earthquakes, lava flows, gaseous effusions, heat, and noise. These are the philosophical regions. I will identify three: three obvious zones of tectonic conflict within contemporary cognitive approaches to perception.

RELEVANCE OF COGNITIVE SCIENCE TO THE PHILOSOPHY OF PERCEPTION

Profits and Perils of Representation

Representation is the wondrous elixir that makes cognitive science possible: a theoretical notion that is powerful, ubiquitous, intoxicating, and dangerous.

Consider arranging the various relations of association between classes of events in a hierarchy of increasing orders of logical complexity. At the bottom we find relations of statistical association, of correlation, and of causation. Claims of the form 'x is correlated with y' are common to all branches of cognitive science and require no special theoretical tools. The relation is fully extensional, in the sense that if x is correlated with y and $y = z$ then x is correlated with z. Causal links are one step up the hierarchy, with some additional content. Another step up takes us to information. Like correlation, talk of information describes a kind of association between ensembles of classes of events, but it is a more complicated kind. It requires a rather robust structure of relations of conditional and *a priori* probabilities between ensembles of input events and output events. These relations can help one make discriminations that mere causal or correlational talk cannot. They may, for example, help us to decide which object is the one perceived among all of the causal antecedents of a given perception. Causally, those antecedents are of a piece, but the perceptual state carries much more information about some of them than about others. Talk of information can relate classes of events in ways that causal talk cannot.

The notion of 'representation' adds another order of complexity to our talk. Event x represents y, is about y, says something concerning y, and is more or less accurate or inaccurate in what it says about y. What makes this relation complicated is that it proceeds through a semantics: one must provide a semantic interpretation for events at the 'representing' end of the relation. That is, what it means to say that x is a representation is that it has some content 'about' some putative object y, and that such content can be assessed for correctness or incorrectness, accuracy or inaccuracy, truth or falsehood. In order to understand that content one must understand exactly what x is representing, and in order to understand that one must understand the semantics of the system of which x is a part. Extensionality fails: even though x represents y and $y = z$, x may fail to represent z. It all depends on the semantics, on how x represents y. It might represent its object as y but not as z – for example, as water, but not as H_2O, even though in fact water equals H_2O. The system in question has yet to learn this piece of chemistry.

The notion is intoxicating because it is so powerful. Once we endorse the claim that the objects under investigation themselves employ a system of representation, then suitable adjustments of the details of those systems can explain any behavior one might encounter, or any behavior one pleases. And indeed the cognitive revolution is (arguably) founded on the claim that all mental states are representational; within our theories the elixir can be employed anywhere, at any time, in liberal quantities.

Perception provides a particularly interesting test for this view. Is all perceptual content representational? Is any? What makes perception so interesting in this regard is that its scientific study started at least a century before the cognitive revolution; competitive research traditions, confined to simpler and less powerful relations of association, were already well established by the middle of the twentieth century. Members of the *ancien régime* had accomplished much with the older tools of correlation, causation, and information, and some resisted the claims that were made for the hegemony of representation.

Even today, there is a full range of opinions concerning whether all or any perceptual content is representational. Clearly there are aspects of perceptual experience that do not seem to represent anything. Whatever a sensation of green signals could it seems, just as well have been signaled by a sensation of blue. So the difference between green and blue is not a difference in what the experience represents. Other qualitative variations likewise lack representational content. A penny on the table presents different ellipses at different times, from different perspectives; yet one would never judge the penny to be anything other than circular. All the aspects of object constancy – size, shape, color, and so on – provide similar examples. The variations in the character of perceptual experience that are discounted when one achieves object constancy do not represent variations in the object. They are variations in what philosophers call 'qualitative' content, and they seem not to represent anything.

The most prominent of the researchers who deny that perception has representational content are ecological psychologists and perception–action theorists. Ecological psychologists prefer to cast their theories using relations of association that stop at the level of information pick-up. After all, perception is found in creatures so simple that one hesitates to ascribe them any other cognitive state at all, and perhaps their commerce with the environment can be accurately described without invoking the complexities of semantic interpretation. Such is the promise of perception–action theory.

There is, however, the problem of perceptual illusion. The conventional view about illusions is that they are a kind of misrepresentation: an error of the senses, a case in which the senses deceive.

But error and deception are impossible unless correctness conditions apply. These states have a content, which in such cases is misleading. Other varieties of perception that are less than fully accurate, such as the perception of pictures, of mirror images, or of movies, raise similar issues. All provide cases of 'mere appearance' (intentional inexistence), in which someone perceives something that seems to be *P*, or seems to perceive something that is *P*, though in fact there is nothing in the vicinity – nothing within the optic array – that is *P*. The representational gambit is particularly hard to resist in such cases. The alternative is either to posit some new metaphysical entity – one which is *P* and is somehow directly perceived – or to deny the existence of illusions altogether. Whereas the representational account is straight-forward: the subject in such episodes is representing something to be *P*, but that representation is (for one reason or another) a misrepresentation. What is real is the existence of the representation. What is unreal is the thing represented. We have a real representation of something unreal. This neatly disposes of the ontological problems of mere appearance.

At least, it does if we accept the hegemony of representation, and allow its governance to extend all the way to the earliest stages of the sensory processes of the simplest creatures. This still needs to be empirically substantiated, and doing so is not easy. What are the elementary terms, the morphemes and syntax, of these systems of representation? One needs to specify the primitive elements, the rules by which they are combined, how they refer, what predicates they employ, and what their truth conditions are. Are they simple, isolated 'features' that might be registered in cortical feature maps? Or must even the primitive terms have the relational character of gestalts (or of affordances)? Is the primeval form of sensory reference allocentric or egocentric? Are all the predicates innate? Is comprehension of them modular? And so on. The representation relation adds another degree of freedom and another layer of complexity to one's theory. While this makes the theory more powerful, it also obliges one to determine the empirical details of how representational systems are constituted. There are reasons to doubt our ability to do this: to have warrant for ascribing one system of representation to a creature rather than another, extensionally equivalent, one.

Psychology Versus Physiology

Another region of tension is found in the relations between the top-down, high-level, boss, chief executive psychologists, and the bottom-up, low-level, worker, underling, physiologists. This description is deliberately tendentious, but it reflects the temper of early broadsides in the revolution. The boss executives were supposed to run the show, tell us what jobs needed to be done, and give broad functional specifications of the different subsystems of the mechanism they master. The neurophysiologists, biophysicists, and neuroscientists gratefully receive these job specifications, get to work, and eventually pass upwards the implementation details of how the lowly neurons cooperate so flawlessly, yet mindlessly, to produce the wonders of representation, The 'autonomy' of levels of explanation guaranteed that the offices of the executives would never be invaded by the grubby workers.

This view has been seriously challenged by empirical discoveries. What are we to make of 'feature detectors' in the primary visual cortex, or of the subsequent 'feature maps' in secondary areas (Treisman, 1993)? How are we to understand the distinction between 'what' and 'where' channels in visual processing, or the importance of synchronized 40-hertz oscillations (Crick and Koch, 1990)? Suddenly the physiological underlings seem to be providing job specifications for the psychologists, instead of the other way round. Such developments oblige us to reconsider the relations between sensory contents and physiological implementation.

We need a clearer view of those relations: of relations between perceptual states and neural states. For example, might we someday find a 'color perception center' where perceptions of color are localized? The question poses a dilemma. It remains incredible that we might someday point to a region of Joe's neuroanatomy and say, truthfully, 'process *x*, going on in there, is identical to Joe's sensation of red'. How do processes in that region get promoted to the status of 'Joe's sensation'? And how could we explain the connection between those processes and the fact that the sensation is a sensation of red, and not of green? This is one version of the 'explanatory gap' (Levine, 1983), and it has yet to be bridged. But alternatives to localization are equally hard to fathom. How could the experience of seeing a red patch be distributed across chunks of neural tissue? How are the chunks coordinated? What binds them into one experience, or into an experience of one patch and not two (Crick and Koch, 1990)?

Other questions are equally in need of recalibration. The connected claims for the 'autonomy' of 'levels of explanation', and for psychology as a 'special science', increasingly seem irrelevant to

the scientific study of perception. Can one consistently hold such views while maintaining that all properties are physical properties?

Theories Versus Common Sense

A third area of stress for contemporary cognitive approaches to perception lies on the tense border between it and the legions of ordinary language and common sense. Many of the processes described in scientific theories or perception seem also to be open to direct observation and immediate access by the ordinary Joe. After all, Joe sees, touches, tastes, smells, and hears; is often aware of doing such; can say when he does such; and can describe something of what it's like when he does. What are those processes but the very ones described by theories of perception? Because the theories seem to encroach upon the territory of common sense, theorists are forced to stipulate relations between what they say and what common sense says. This is quite literally a job of public relations, and like any public relations job it is delicate, essential, and potentially explosive.

It is under this heading that one must place all the philosophical arguments relying on ordinary intuitions about what it's like to see red, about the 'qualitative character' or 'qualia' of seeing red, and about the possibility that what it's like for me to see red is qualitatively identical to what it's like for your to see green (see Block, 1980). All such arguments can be entirely dismissed – claims based on common sense might occasionally be true – but simply that the task of figuring out a response to them is similar to the task of negotiating with a public whose intuitions may differ from one's own.

Possible responses range from ignoring ordinary Joe altogether (either denying that perceptual experience has any contact at all with the states and processes hypothesized by these theories, or admitting that there is some contact, but despairing of any useful information arising from introspective methods) to embracing his verbal outputs as constituting the canon for 'heterophenomenology' (Dennett, 1991). At one end of the spectrum we have the public relations triumph of scientific vindication of all the intuitions of dear old Auntie (Fodor, 1985) – the customer is always right! – while at the other end the corporate PR man announces, gleefully, that Auntie and all the other putative customers have no thoughts, no desires, no beliefs, and no existence (Churchland, 1979).

The problem can emerge in the laboratory, for example, in experiments on the cognitive unconscious. What are we to make of the appearances (or lack of appearances!) of first-person introspective access to the states and processes hypothesized by theories of perception? For example, blindsight is problematic largely because one relies on the truthfulness of the subject's testimony that he or she does not see the *X*, is not aware of the *X*, even though pointing (and other behavioral measures) would normally be taken to contradict this (Weiskrantz, 1997). Why should we believe the subject who says, sincerely, that he or she does not see? And why should one adopt the commonsense assumption that seeing something implies consciousness of it, or, alternatively, consciousness of seeing it? These questions lead us back to the old problem of the scope and limits of introspective methods.

Some useful distinctions have gradually become clear in the philosophical literature. For example, the notion of qualia – the paradigm of entities whose essence is fully revealed by introspective consciousness – is ambiguous. In one sense, qualia are the qualities attributed to appearances – the qualities of color, taste, smell, etc., that things seem to have – and in another sense, qualia are the properties of sensations – the properties of sensory states in virtue of which things appear as they do. Neither notion should be confused with 'what it is like' to have a given sensory state. The latter phrase, made popular by Nagel (1979), was meant to pick out all and only the conscious mental states, and if one admits the possibility of sensory states of which one is not conscious, then the qualia of a sensory state may sometimes differ from 'what it is like' to have that state. The term 'conscious' is itself ambiguous. In one sense, it applies to any creature that is awake and sentient; in another sense, it applies only to mental states of which one is introspectively aware. In any case, it is clear that consciousness and sentience are not the same thing. One can study sensory processes and the qualitative character of sensory states without necessarily committing oneself to some verdict about what it is like to have those states.

The logical endpoint of such continuing Balkanization is a continuum of states, starting at some that are clearly unconscious and insentient, and ending with some that are, in the fullest sense, conscious sensory states. In between would be a vast series of intermediate states, of varying orders of organization, encompassing all the transition zones between unconscious and partially conscious states, and capturing all the pathological states and paradoxes. Once that order is empirically described, and all the relations between neighboring points within it are made clear, it becomes a matter

of indifference where one draws the line between cases in which some ordinary language term applies and cases in which it does not. All the facts upon which such a verdict rests would be laid upon the table, and the only remaining questions would be verbal.

RELEVANCE OF THE PHILOSOPHY OF PERCEPTION TO COGNITIVE SCIENCE

The three philosophical issues about perception outlined above are rarely found in 'pure' forms – as isolated and distinct dilemmas about representation, about neural implementation, or about common-sense introspective access. More typically, and more potently, in the outstanding controversies of the day, all three themes are found intermingled, in maximally confounding combination. For example, the puzzles over temporal anomalies of perception and the phi phenomenon (Dennett and Kinsbourne, 1992) involve all three. We need assumptions about how time is represented, how those representations are instantiated in the nervous system, and how ordinary subjects access those representations when they see a moving light. 'Filling in' likewise gains the critical mass of a dilemma only when one imports assumptions from all three of our philosophical regions: assumptions about spatial representations, about their neural implementation, and about introspective access to them (Pessoa, Thompson and Noë, 1998). Blindsight (Weiskrantz, 1997) and the various 'binding problems' (Crick and Koch, 1990; Treisman, 1993) may seem to be almost pure neurophysiological issues, but on closer examination what makes them controversial, what generates all of the heat, is some combination of assumptions crossing our philosophical faultlines. Lots of energy builds up at such places, and when it is let go, the results can be exciting.

References

Block N (1980) Troubles with functionalism. In: Block N (ed.) *Readings in the Philosophy of Psychology*, vol. I, pp. 268–305. Cambridge, MA: Harvard University Press.

Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2**: 263–275.

Dennett DC and Kinsbourne M (1992) Time and the observer: the where and when of consciousness in the brain. *Behavioral and Brain Sciences* **15**: 183–247.

Fodor JA (1985) Fodor's guide to mental representation. *Mind* **94**: 76–100.

Hintikka KJJ (1969) On the logic of perception. In: *Models for Modalities: Selected Essays*, pp. 151–183. Dordrecht: Reidel.

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Nagel T (1979) What is it like to be a bat? In: *Mortal Questions*, pp. 165–180. Cambridge, UK: Cambridge University Press.

Pessoa L, Thompson E and Noë A (1998) Finding out about filling in: a guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences* **21**: 723–802.

Treisman A (1993) The perception of features and objects. In: Baddeley A and Weiskranz L (eds) *Attention: Selection, Awareness, and Control: A Tribute to Donald Broadbent*, pp. 5–35. Oxford: Clarendon Press.

Weiskrantz L (1997) *Consciousness Lost and Found*. Oxford: Oxford University Press.

Further Reading

Bickle J (1998) *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Chisholm R (1957) *Perceiving: A Philosophical Study*. Ithaca, NY: Cornell University Press.

Churchland PM (1979) *Scientific Realism and the Plasticity of Mind*. Cambridge, UK: Cambridge University Press.

Clark A (2000) *A Theory of Sentience*. Oxford: Oxford University Press.

Davidson D (1970) *Mental events*. In: Foster L and Swanson JW (eds) *Experience and Theory*, pp. 79–101. Amherst, MA: University of Massachusetts Press.

Dennett DC (1991) *Consciousness Explained*. Boston, MA: Little Brown.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Fodor JA (1979) *The Language of Thought*. Cambridge, MA: Harvard University Press.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Hill CS (1991) *Sensations: A Defense of Type Materialism*. Cambridge, UK: Cambridge University Press.

Kim J (1993) *Supervenience and Mind: Selected Philosophical Essays*. Cambridge, UK: Cambridge University Press.

Lycan WC (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Peacocke C (1983) *Sense and Content: Experience, Thought, and Their Relations*. Oxford: Clarendon Press.

Rosenthal D (1997) A theory of consciousness. In: Block N, Flanagan O and Güzeldere G (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 729–753. Cambridge, MA: MIT Press.

Warner R and Szubka T (eds) (1994) *The Mind–Body Problem: A Guide to the Current Debate*. Oxford: Blackwell.

Perception, Unconscious

Intermediate article

Jacqueline C Snow, University of Melbourne, Victoria, Australia
Jason B Mattingley, University of Melbourne, Victoria, Australia

CONTENTS

Unconscious perception involves the processing of sensory information by the brain in the absence of subjective awareness. The physical properties of sensory stimuli, their meaning, and their emotional significance, may all be perceived unconsciously.

INTRODUCTION

Many psychological experiments suggest that much of our perception occurs in the absence of conscious experience. In this article we explore the phenomenon of unconscious perception, beginning with a historical account of the concepts of perception and consciousness. We examine the methods adopted by experimental psychologists to try and distinguish between perception with and without consciousness in normal observers. There are different measures of unconscious perception, and different ways to define the boundary between conscious and unconscious processing. Many aspects of unconscious perception are evident in patients with focal brain injury caused by stroke or tumor. These neuropsychological cases provide important clues to the neural substrates of unconscious perception, as do recent brain imaging results. We conclude that considerable perceptual processing can occur in the absence of conscious experience, and that specific brain regions may have a special role in supporting conscious perception.

WHAT IS UNCONSCIOUS PERCEPTION?

'Perception' refers to the automatic and relatively effortless process implemented by the central nervous system to extract information about physical objects and events in the world. Much of what we perceive of the world occurs without us being directly aware or conscious of it, i.e., without any direct or immediate subjective experience. The idea that perception might occur without conscious experience has generated a great deal of interest, because it challenges the conventional idea that perception is an ontologically conscious phenomenon. In this article we focus on unconscious visual perception. Roughly half the primate brain is devoted to processing visual information. Due to the large body of experimental work conducted in the visual domain, we now know more about unconscious perception in vision than in any other modality.

Numerous terms have been used to refer to the phenomenon of perception in the absence of awareness: most commonly 'unconscious' or 'implicit' perception, and sometimes 'automatic', 'covert' or 'tacit' perception. Unconscious perception has proven difficult to define, but it may be broadly characterized as perception that is not accompanied by subjective experience or the ability to act upon relevant information in an intentional fashion. Unconscious perception is generally understood to occur without effort (automatically) and in parallel (simultaneously) across different sensory modalities, and to have a large processing capacity. Unconscious perception can be contrasted with conscious perception, which carries with it a characteristic, though often indefinable, subjective experience of 'what it is like': for example, the smell of coffee, the color of a rose, or the sound of a familiar voice. Conscious perception involves processing of sensory inputs in such a way that we can act upon them in a voluntary, intentional and novel manner. (*See* **Qualia**)

The idea that perception might occur in the absence of consciousness raises an important question: should consciousness be regarded as an 'all or nothing' phenomenon? Or should it be characterized as a graded phenomenon, in which there are varying degrees of conscious perception. Note that the term 'consciousness' is used in many contexts that are not directly related to perception (e.g. consciousness as the waking

state, self-consciousness). In this article we restrict our discussion to the conscious–unconscious dichotomy as it pertains to perceptual awareness. (*See* **Self-consciousness**)

HISTORY AND EPISTEMOLOGY

The question of whether perception occurs in the absence of consciousness has a long history in experimental psychology and philosophy. Nineteenth-century psychologists such as Wilhelm Wundt, William James and Sigmund Freud used introspective measures of awareness (i.e., systematic reporting of conscious perceptual experiences from a personal perspective) to demonstrate the existence of unconscious perceptual processes. Introspectionist techniques fell out of favor with the advent of behaviorism in the first half of the twentieth century. The behaviorist view maintains that only directly and objectively observable behavior is appropriate subject matter for experimental psychology. This led to a rejection of the concept of consciousness, and a denial of mental constructs as explanations of behavior. (*See* **Introspection**; **Behaviorism, Philosophical**)

The failure of behaviorist accounts to adequately characterize cognitive processes that are not overtly observable, such as attention and memory, led many researchers to advocate a cognitive approach instead. The cognitive approach emphasizes the importance of mental representations and computations in providing explanations for behavior. It allows a broader conceptualization of the causation of behavior, thereby again admitting the study of unconscious perception into the domain of scientific investigation.

The study of unconscious perception was advanced in the 1980s through pioneering methods for dissociating conscious from unconscious perception in normal observers. This research was paralleled by studies of unconscious perception in neurological patients with discrete brain lesions. Recent advances in functional brain imaging techniques, such as scalp-recorded event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI), have significantly increased our understanding of the brain areas that mediate conscious and unconscious perception. (*See* **Event-related Potentials and Mental Chronometry**; **Neuroimaging**)

HOW MUCH INFORMATION IS PERCEIVED UNCONSCIOUSLY?

At every moment, our various sensory systems are receiving a vast amount of information about the environment. Receptors sensitive to particular sorts of stimuli encode incoming information according to physical properties, location, intensity, duration, and so forth. This information is carried within distinct neural systems belonging to the relevant sensory pathways, and is ultimately transmitted to specialized regions within the cerebral hemispheres. The information thus conveyed is used to create a representation of the external environment. In the case of vision, it is important to note that perception is not embodied in an exact representation of the world (a 'picture in the head'), but rather reflects a reconstruction based upon the distributed activity of structurally and functionally diverse regions within the brain.

Despite the vast amount of sensory information that is received by the brain, there appear to be limits to the amount of information we are able to consciously process at once. Due to this limit in conscious processing capacity, incoming sensory information must be filtered for importance: relevant stimuli must be given high priority for further processing, while irrelevant information must be filtered out. This filtering is largely achieved by mechanisms of selective attention, which have been likened to an 'attenuator' or 'bottleneck' in information processing. (*See* **Attention, Models of**)

Mechanisms of selective attention form a crucial bridge between unconscious perception and what reaches consciousness. We are able to modulate the content of our conscious perception by choosing what we will attend to. Perception may be influenced by top-down (endogenous) strategies, so that awareness is actively, and with effort, controlled by one's intentions. Because of limited capacity, this necessarily involves the filtering of concurrent sensory events, preventing these from reaching awareness. Conversely, perception may be affected in a bottom-up (exogenous) fashion, by stimuli that automatically grab attention due to their sudden appearance or salience. For example, we can direct our attention to oncoming traffic when waiting to cross the road (top-down control), yet salient stimulus information, such as a car horn, can rapidly modify the current content of our perception (bottom-up control), independently of current goals and intentions. The coexistence of stimulusdriven and goal-driven filtering of information in the control of attention, awareness and behavior has obvious evolutionary advantages: it allows us to act purposefully and intentionally without being oblivious to potentially harmful situations. (*See* **Selective Attention**)

METHODS USED TO ASSESS UNCONSCIOUS PERCEPTION

There are many ways in which awareness of stimulus information can be assessed. There is a continuing debate concerning the best methodology to reveal unconscious perception, how to define a threshold of awareness, and how to establish an unambiguous criterion for awareness. Consequently, many tasks have been employed to demonstrate unconscious perception. It is useful to consider these different methodologies within a framework that illustrates the similarities and differences between them.

Dissociations Between Measures

The most common approach to examining perception without awareness has been to demonstrate a dissociation, or difference in performance, between two measures of perception; one measure is designed to be sensitive to perception with awareness, while the other is assumed to reflect perception in the absence of awareness. The aim of the dissociation technique is to demonstrate that stimuli perceived without awareness (as assessed by the measure of conscious perception) are none the less able to influence performance (as assessed by the measure of unconscious perception). If perception can occur without awareness, then a dissociation between the two measures of perception should be observed.

Manipulating stimulus properties

One approach to detecting and measuring unconscious perception of stimulus information is to systematically reduce the quality of the critical stimulus. This reduces the availability of the critical stimulus to conscious perception (as judged by self-report or forced-choice recognition tasks). The typical approach is to initially ascertain an individual's threshold for awareness of experimental stimuli. These stimuli are then presented under conditions in which they appear below the defined threshold, with the aim of determining whether stimuli that fail to reach awareness are none the less able to influence performance on some task.

Many examples of unconscious perception come from studies using 'priming' paradigms. In a typical word-priming experiment, a single letter string (a 'prime') is presented briefly, followed by another letter string (a 'target') in the same location, to which the participant must make some kind of response. The rationale behind this paradigm is that the priming stimulus can change the efficiency of processes involved in generating the target response, depending upon the physical or semantic relationship between prime and target stimuli. Semantic priming is a phenomenon whereby the time taken to respond to a target (e.g. to classify a letter string as a word or a non-word) is affected by the meaning of the prime. For example, if participants are presented with a prime word such as *BREAD*, followed by a semantically related target word such as *BUTTER*, responses to the target word are facilitated: they are faster than when it is preceded by a semantically unrelated word such as *NURSE*. Presentation of the prime may have either facilitatory or inhibitory effects on a subsequent response to the target, depending on whether the prime makes access to the target response more or less efficient. (*See* **Priming**)

'Masked priming' techniques are designed to limit the accessibility of a prime to awareness. The aim of masked priming is to determine whether prime stimuli are able to influence responses to a subsequent target indirectly, even when presented below an objectively defined threshold of awareness. Several techniques have been used to manipulate awareness of the prime. In 'visual masking', visibility of a prime is reduced by presenting another stimulus (a 'mask') in close temporal or spatial proximity (see Figure 1(a)). (*See* **Masking**)

The sorts of stimuli used in these experiments (as either target or prime) range from numbers, letters and color patches to words, pictures and faces. Results suggest that basic stimulus properties such as shape and orientation are perceived unconsciously, as are more complex stimulus attributes. These include categorical associations (e.g. that 'mouse' and 'horse' both belong to the animal category, whereas 'apple' and 'donkey' belong to different categories); semantic associations (e.g. that 'king' and 'queen' are related in meaning); and emotional valence (e.g. face stimuli with happy or fearful expressions).

Manipulating the distribution or focus of attention

Rather than limit conscious perception using techniques such as visual masking, an alternative approach is to manipulate where, or on what, attention is focused. The aim of this approach is to determine whether there is a dissociation between conscious and unconscious perception of unattended stimuli. In these experiments, stimuli are presented under suprathreshold (readily detectable) conditions, so that when observers are asked to focus attention on the critical stimulus it can be perceived consciously. Arguably, manipulations of

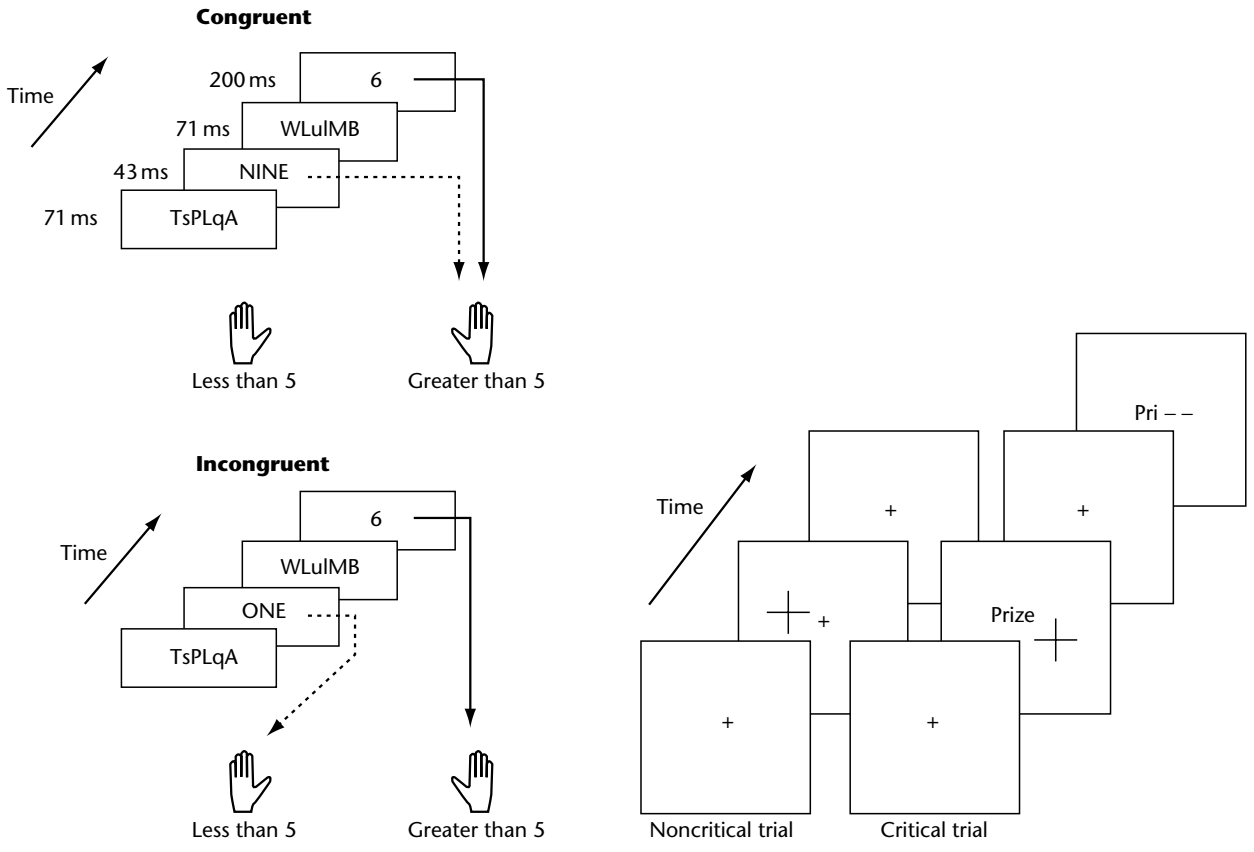


Figure 1. (a) Unconscious perception of a masked prime. Dehaene *et al*. (1998) used a masked priming paradigm, in combination with brain imaging techniques, to demonstrate unconscious perception of masked stimuli. On each trial a number between 1 and 9 (the prime) was presented for a very short duration (43 ms). The prime was forward and backward masked by random letter strings (the first and third displays), and followed by another number (the target). The subject's task was to indicate via a key-press whether the target was greater than or less than 5. The prime could either be 'congruent' with the target (i.e. yielding the same answer), or 'incongruent'. Although subjects were unable to detect the prime, they responded to congruent prime–target pairs faster than to incongruent pairs. Unconscious perception of the masked primes was accompanied by changes in electrical brain activity (as measured by ERPs) and cerebral blood flow (using fMRI) in brain areas responsible for controlling motor responses. (Adapted from Dehaene *et al*., 1998.) (b) Example of displays used in a study of unconscious perception during inattentional blindness (Mack and Rock, 1998). Subjects' awareness of word stimuli was manipulated by having them attend to a distractor task in which they compared the lengths of the vertical and horizontal limbs of a cross, while fixating a central spot. On unexpected 'critical inattention trials', a word (the prime) appeared in the centre of the screen at the same time as the cross. A large proportion of subjects failed to report the word on the critical inattention trial, thus demonstrating inattentional blindness. Unconscious perception of the prime word was subsequently examined using a stem completion task (i.e. making a word from the first three letters presented). Subjects offered the unseen prime as a stem completion on a significantly greater number of trials than would be expected by chance, indicating that the primes had been perceived unconsciously.

attention provide a more realistic simulation of the way in which information is perceived without awareness in everyday life than do manipulations of stimulus quality.

There are numerous cases in which attentional mechanisms influence conscious visual perception. For example, in normal observers the second of two briefly presented visual targets often goes undetected when it occurs within half a second of the first: an effect known as the attentional blink. Failure to detect the second target is due to attentional rather than sensory limits in processing, since it is only observed when the task involves detection of both targets; there is no such deficit when participants are only required to detect the second target. Similarly, in the phenomenon of

change blindness, observers fail to notice salient changes between two visual scenes flashed successively with a blank interval between them. In this case the blank interval removes the visual transient associated with the change, so that attention is not automatically drawn to it. A closely related phenomenon is that of inattentional blindness (Mack and Rock, 1998). When observers have their attention directed to one aspect of a visual display, they frequently miss suprathreshold stimuli presented elsewhere. Note that in this case the failure of conscious perception occurs without a blank interval or mask. Again, the effect is due to having attention devoted to a concurrent task, since if observers are directed to focus their attention on such stimuli, they are capable of consciously perceiving them. (*See* **Change Blindness**; **Change Blindness, Psychology of**)

These examples from experimental psychology have parallels in real-world situations. For instance, Haines (1989) examined experienced pilots as they attempted to land a plane in a flight simulator. As they were landing, the pilots monitored a display of critical flight information that was projected onto the windscreen. While their attention was directed to the flight information, many pilots failed to notice that another aircraft had moved across the runway into their field of view.

What happens to visual information that escapes our awareness due to inattention? Electrophysiological evidence suggests that the meaning of a word stimulus presented during the attentional blink may be accessed unconsciously. In change blindness, above-chance performance in detecting visual changes (e.g. on a forced-choice task) has been observed even though participants are not able to perceive the changes consciously. Similarly, Mack and Rock (1998) found that participants who were inattentionally blind to a prime word tended to offer this word more often in a stem completion task than would be expected by chance (Figure 1(b)). Taken together, these findings provide evidence that engaging attention on a secondary task can limit conscious reports of stimulus events, but leave unconscious perception of them relatively unaffected.

Theoretical Considerations

Subjective versus objective measures of perception

Two general approaches have been used to demonstrate dissociations between measures of perception with and without awareness: 'subjective' and 'objective'. Studies employing subjective measures of awareness consider an observer's direct report to be a valid index of whether or not a given stimulus has been perceived consciously. However, it has long been known that subjective reports can be influenced by factors other than an observer's perceptual sensitivity. Such factors may include confidence, expectations, or simply an affinity for one or another response. For instance, participants may be more or less willing to guess under conditions of uncertainty. This can lead to differences in the measured extent of conscious perception without any differences in sensitivity. It therefore becomes difficult to determine whether subjective measures of awareness actually represent an exhaustive probe of conscious influences on behavior.

Objective measures do not rely on self-report as an indicator of awareness. Rather, conscious perception is inferred using a measurable index of performance (such as forced-choice recognition), and compared with performance on a task designed to reveal any unconscious perception. Objective measures provide a more conservative estimate of the sorts of conditions under which stimuli might be perceived without awareness. However, the objective approach may underestimate the extent of unconscious perception. In contrast, subjective techniques allow a less stringent measure of unconscious perception and thereby increase the likelihood of detecting unconscious influences on performance. Because of the subjective nature of conscious perception, many argue that introspective reports of perceptual experience should be considered valid. In any case, whether assessed by subjective or objective measures, numerous studies have provided unequivocal demonstrations of unconscious perception.

Direct versus indirect tests

The distinction between direct and indirect measures of perception concerns the kind of stimulus information required to complete a task. In direct tests, participants respond to the critical stimulus under investigation. For example, consider a typical priming paradigm in which the word *QUEEN* is presented and masked. A direct test of perceptual awareness might require participants to indicate the identity of the prime via a forced choice. An indirect test might require participants to make a decision as to whether a subsequent letter string (e.g. *KING*) is a word or not. Any difference in the speed of response to the target as a function of prime identity is taken as evidence for unconscious perception of the prime.

Unconscious Perception in Neuropsychological Patients

The phenomenon of unconscious perception in normal individuals may be closely related to the residual abilities exhibited by neuropsychological patients with impaired conscious perception. Understanding the nature of preserved unconscious perception in neuropsychological cases provides important insights into the structural and functional mechanisms subserving conscious perception in the human brain.

Blindsight

Patients with lesions to the primary visual cortex (also known as area V1, or striate cortex) are subjectively blind (i.e. they lack visual experience) for a region of the visual field represented by the lesioned cortex. Despite their subjective visual loss, however, there is evidence to suggest that these patients can perceive aspects of visual stimuli unconsciously. This phenomenon is known as blindsight. Individuals with blindsight may be able to make eye movements towards, or point in the direction of, briefly presented lights flashed in the blind field. Some such patients are also able to discriminate different attributes of visual stimuli, such as color, shape, orientation or direction of movement. Importantly, this residual perception occurs in the absence of the subjective experience of seeing, and so provides a compelling example of unconscious perception. By considering the neural pathways that remain intact in such patients, it is possible to determine the neural basis for these unconscious visual abilities (see Figure 2). (*See* **Blindsight**; **Blindsight, Neural Basis of**)

Visual information from V1 is directed along two anatomically and functionally distinct pathways, or 'streams' (Milner and Goodale, 1995). A dorsal stream, projecting from the occipital cortex to the superior parietal lobe, is involved in the spatial coding of attention and movement; while a ventral stream, projecting to the inferior temporal cortex, is involved in processes of object recognition. The ventral stream relies almost entirely on geniculostriate pathways for visual input. In the dorsal stream, however, retinal information is projected directly to various subcortical sites, including the superior colliculus, inferior pulvinar and suprachiasmatic nucleus, without passing through V1. (*See* **Parietal Cortex**; **Temporal Cortex**)

In cases of blindsight, damage to area V1 disrupts the flow of visual information within the geniculostriate pathway, thereby depriving the ventral stream of visual input. Subcortical structures and their connections to sites within the dorsal stream, however, remain intact. It is these pathways that are likely to underlie the blindsight patient's residual visual abilities in the affected field.

Unilateral neglect and extinction

Unilateral neglect is most commonly observed after damage to the parietal lobe, particularly of the right hemisphere. Neglect patients are typically unaware of stimuli located towards the contralesional side (i.e., the left side after right hemisphere damage), with profound consequences for normal activities of daily living. Patients with left neglect may fail to eat food on the left side of their plate, may have difficulties in dressing the left side of their body, and may collide with objects on the left. Some neglect patients are able to detect isolated stimuli on the contralesional side, but have problems detecting the same stimuli when a simultaneous event is presented on the opposite (ipsilesional) side. Such patients are said to have 'extinction', because the stimulus on the affected side is extinguished from awareness. (*See* **Neglect**)

Conscious perception is lost in neglect and extinction patients even though the primary visual cortex usually remains intact. But, like blindsight patients, these patients demonstrate unconscious perception of visual information in the neglected or extinguished field. For example, although neglect patients are often unaware of left-sided details of objects, they are able to perceive whether such objects are symmetrical in shape – a judgment that requires perception of both sides of the object (Driver *et al*., 1992). This demonstrates that unconscious figure–ground segmentation processes remain intact for neglected information. Other unconscious visual processes, such as perceptual grouping, allow separate stimuli to be bound together to form coherent objects. Contralesional stimuli that would usually be extinguished upon bilateral presentation may suddenly be perceived consciously due to the operation of such perceptual grouping processes (Mattingley *et al*. (1997); see Figure 3(a)). (*See* **Vision: Occlusion, Illusory Contours and 'Filling-in'**)

Unconscious perception in neglect may extend to more complex visual attributes. Visual priming studies have shown that neglected words and pictures may be perceived unconsciously, and in some cases may even activate semantic and emotional associations (McGlinchey-Berroth *et al*. (1993); see Figure 3(b)). Such preserved semantic processing is reminiscent of that observed under conditions of inattentional blindness in normals.

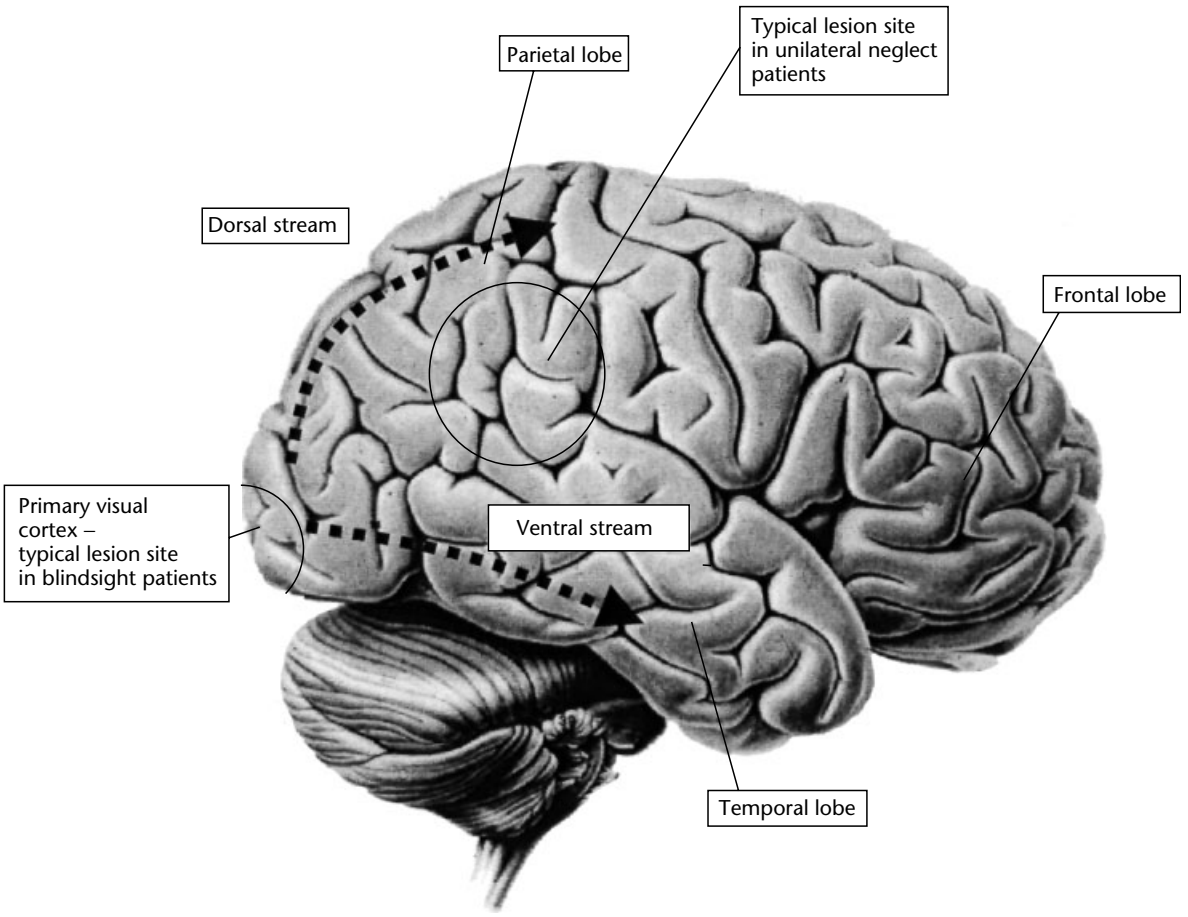


Figure 2. Lateral view of the right hemisphere of the human brain. Temporal, parietal and frontal lobes are shown, as well as the site of primary visual cortex (area V1, or striate cortex). Dorsal and ventral visual pathways are indicated by arrows. Outlined areas indicate typical lesion sites associated with blindsight and unilateral neglect.

Behavioral observations of unconscious perception for neglected stimuli are consistent with the locus of brain damage in these patients (see Figure 2). Although neglect can arise after damage to several different brain areas, it is most severe and persistent following lesions of the inferior parietal lobule, which is located at the interface between the dorsal and ventral visual pathways. Preservation of some contralesional inputs to the ventral pathway following such damage may underlie unconscious perception of object identity and meaning. This hypothesis is supported by recent fMRI and ERP studies with extinction patients (see Figure 4). These studies have shown that extinguished visual stimuli that escape conscious perception still activate early striate and extrastriate areas, as well as subsequent processing stages within the ventral stream, such as face-specific regions in the fusiform gyrus (Rees *et al*., 2000).

The failure of conscious perception for contralesional stimuli in neglect and extinction patients resembles the loss of visual awareness in normals when their attention is engaged elsewhere (as in inattentional blindness), or when it is limited by the rapid succession of target events (as in the attentional blink). Neglect and extinction reflect a pathological bias of spatial attention towards the unaffected field. When patients are cued to attend to stimuli on their neglected side they suddenly perceive them consciously, only to lapse back into their neglectful state shortly afterwards. The modulatory role of attention in conscious perception may be especially relevant in the case of extinction, in which contralesional events are only missed when an ipsilesional stimulus competes for attention. Indeed, extinction patients appear to suffer from an exaggerated form of the normal difficulty in attending to multiple targets at once (Husain *et al*., 1997).

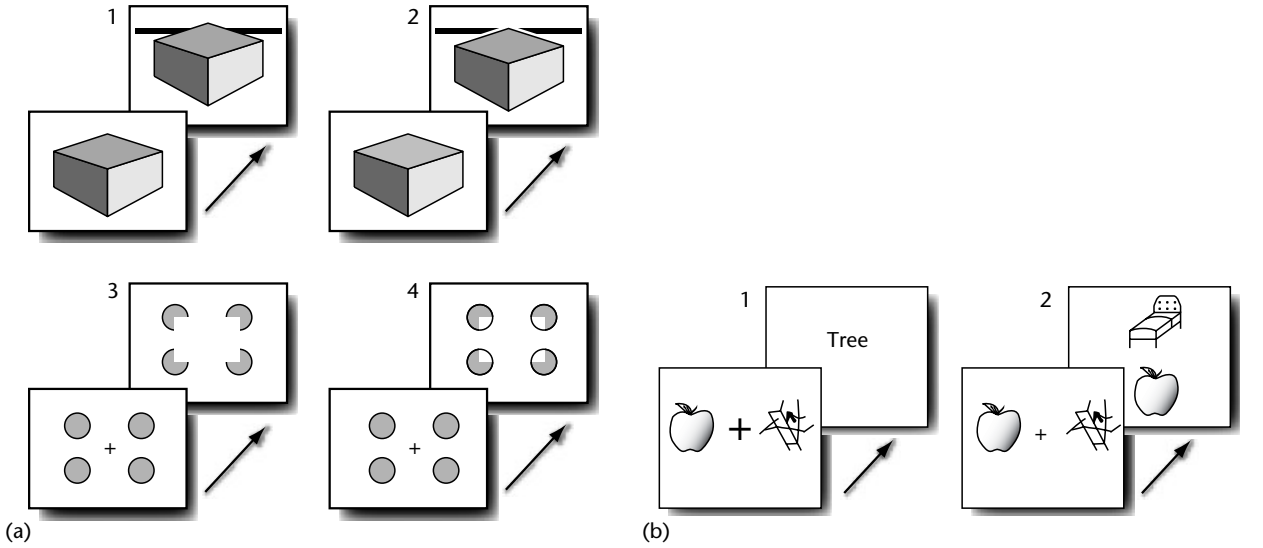


Figure 3. (a) Examples of sequences used to examine unconscious visual completion in a patient with visual extinction. In (1) and (2) the task was to detect the onset of black bars on the left or right alone, or on both sides simultaneously (shown). In (1) the bars appear to form a continuous, but partly occluded, rod, because of image segmentation processes. Marked extinction was observed for the left bar on bilateral trials as pictured in (2), but was significantly reduced in trials like (1), demonstrating the operation of preserved depth perception and image segmentation processes. In (3) and (4) the patient was asked to indicate which quarter-segments had been removed (left, right, both or neither) from four disks. The patient showed less severe left-sided extinction on bilateral trials when the removed portions formed an illusory rectangle, as in (3), than when segments were removed from other parts of the disk, or when arcs prevented the illusion, as in (4). (Adapted from Mattingley *et al.*, 1997.) (b) Example display sequences from a study of unconscious perception of contralesional stimuli in patients with unilateral neglect (McGlinchey-Berroth *et al.*, 1993). Neglect patients were faster to decide whether a target letter string (such as *TREE*) was a word when it was preceded by a semantically related picture (e.g. an apple) to the contralesional visual field, as in (1), than in conditions where the contralesional picture was unrelated to the prime. In a separate control task the same patients were unable to identify the contralesional prime at better than chance levels when presented with the prime and a foil, as in (2). These results suggest that neglected stimuli receive considerable unconscious (semantic) processing. (Adapted from Driver and Mattingley, 1998.)

Neurophysiological and Neuroimaging Contributions to the Study of Unconscious Perception

Recent advances in brain imaging techniques have begun to reveal the neural correlates of conscious and unconscious perception. Given the extent of unconscious perception in normals and in neurological patients, we might ask: what kind of neural activity is needed to yield conscious perception? Is it possible to localize conscious perception to a single area or circuit within the brain, or is it distributed across a network of areas? (*See* **Neural Correlates of Visual Consciousness**; **Neural Correlates of Consciousness as State and Trait**)

One approach has been to use masked priming techniques in combination with fMRI (Dehaene *et al.*, 1998). These studies have demonstrated that unconscious perception of words, objects and faces is associated with increased neural activity in brain areas specialized for processing these different kinds of stimuli. Other studies have investigated the neural correlates of visual perception with changes in selective attention. In an fMRI study of change blindness, observers had to detect changes in face photographs while performing an attention-demanding secondary task (Beck *et al.*, 2001). Brain activity was compared for trials in which there was no change in the face stimuli and trials in which a change had occurred but was not detected. Focal activation was observed in brain areas normally responsive to faces (the fusiform face area), despite the observers' failure to consciously perceive the change. Conversely, conscious perception of change was associated with increased activity in parietal and prefrontal cortices, perhaps indicating a special role for these areas in mediating conscious visual experience. (*See* **Object Perception, Neural Basis of**; **Face Perception, Psychology of**)

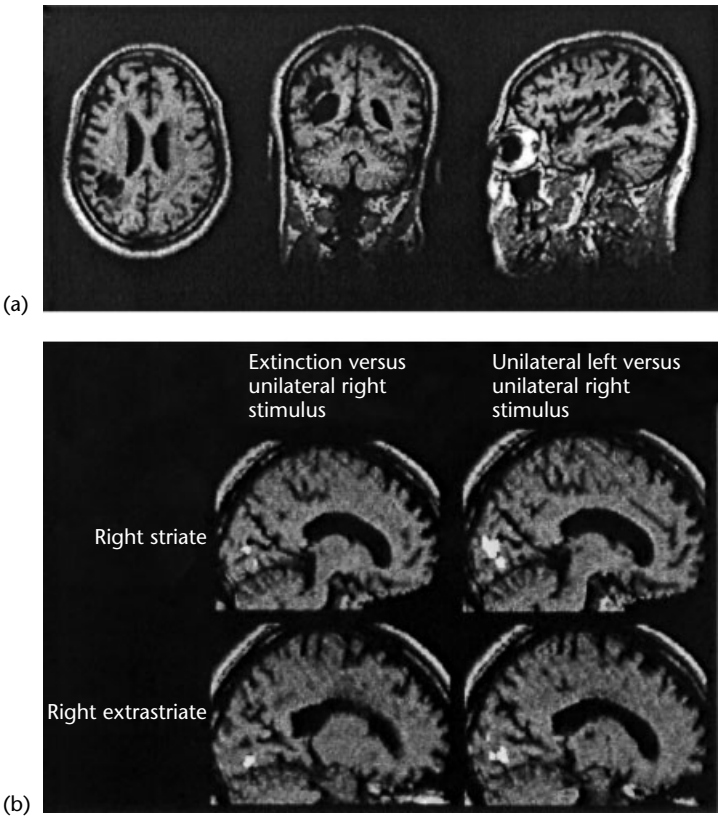


Figure 4. (a) The neural correlates of unconscious perception in extinction, revealed using fMRI. Patient GK suffered damage to the right inferior parietal lobe, as evident in these structural scans (from left to right: axial, coronal and sagittal views). Note that left and right are reversed in these images. (b) In the extinction study, the patient's task was to detect pictures of faces or houses to the left or right visual fields, or to both fields simultaneously. Significant activity was observed in early striate and extrastriate areas on bilateral face trials in which the left stimulus was extinguished and the right stimulus only was detected, compared with trials in which a right stimulus was presented in isolation (slices on left of panel). These regions of significant activity overlap anatomically with those obtained for unilateral left stimuli relative to unilateral right stimuli (slices on right of panel). (Adapted from Rees *et al.*, 2000.)

The results of these and other neuroimaging studies in normal observers are broadly consistent with the finding of substantial unconscious perception in patients with neglect and extinction. They imply that the processing of stimuli perceived unconsciously is correlated with activity in material-specific areas within the ventral pathway; and that conscious visual perception may be especially reliant on areas within the dorsal pathway, including the parietal cortex.

SUMMARY

The phenomenon of unconscious perception has been investigated from a number of theoretical perspectives. Empirically, unconscious perception has been explored using several different techniques, including behavioral investigations in normal observers and neuropsychological patients, and neurophysiological techniques such as scalp-recorded ERPs and fMRI. Although we have focused on evidence for unconscious perception in vision, analogous phenomena have been observed in other sensory modalities.

Taken together, the evidence suggests that significant perceptual processing can occur without awareness. Unconscious perception of objects, words, faces, and many other visual stimuli is associated with activity in specialized brain areas within the ventral pathway, and may also lead to activation of remote areas, such as those responsible for motor responses. Findings in normals and neuropsychological patients suggest that mechanisms of selective attention play a crucial role in conscious perception, while considerable unconscious perception may still proceed in the absence of attention. Activity within parietal cortex is associated with the conscious perception of change

in visual displays, and damage to this area results in contralesional spatial neglect and extinction.

While a comprehensive explanation for the cognitive and neural mechanisms underlying unconscious perception is still a long way off, convergent approaches that combine behavioral and neurophysiological methods are likely to reveal further aspects of it.

References

Beck DM, Rees G, Frith CD and Lavie N (2001) Neural correlates of change detection and change blindness. *Nature Neuroscience* **4**: 645–650.

Dehaene S, Naccache L, Le Clec'H G *et al*. (1998) Imaging unconscious semantic priming. *Nature* **395**: 597–600.

Driver J, Baylis GC and Rafal RD (1992) Preserved figure–ground segregation and symmetry perception in visual neglect. *Nature* **360**: 73–75.

Driver J and Mattingley JB (1998) Parietal neglect and visual awareness. *Nature Neuroscience* **1**: 17–22.

Haines RF (1989) A breakdown in simultaneous information processing. In: Obrecht G and Stark LW (eds) *Presbyopia Research: From Molecular Biology to Visual Adaptation*, pp. 171–175. New York, NY: Plenum.

Husain M, Shapiro K, Martin J and Kennard C (1997) Abnormal temporal dynamics of visual attention in spatial neglect patients. *Nature* **385**: 154–156.

Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

Mattingley JB, Davis G and Driver J (1997) Preattentive filling-in of visual surfaces in parietal extinction. *Science* **275**: 671–674.

McGlinchey-Berroth R, Milberg WP, Verfaellie M, Alexander M and Kilduff PT (1993) Semantic processing in the neglected field: evidence from a lexical decision task. *Cognitive Neuropsychology* **10**: 79–108.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford: Oxford University Press.

Rees G, Wojciulik E, Clarke K *et al*. (2000) Unconscious activation of visual cortex in the damaged right hemisphere of a patient with extinction. *Brain* **123**: 1624–1633.

Further Reading

Driver J and Vuilleumier P (2001) Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* **79**: 39–88.

Kanwisher N (2000) Neural events and perceptual awareness. *Cognition* **79**: 89–113.

Logothetis NK (1998) Single units and conscious vision. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **353**: 1801–1818.

Marcel AJ (1983) Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237.

Merikle PM, Smilek D and Eastwood JD (2001) Perception without awareness: perspectives from cognitive psychology. *Cognition* **79**: 115–134.

Meyer DE and Schvaneveldt RW (1976) Meaning, memory structure, and mental processes. *Science* **192**: 27–33.

Pashler HE (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Rensink RA (2000) Visual search for change: a probe into the nature of attentional processing. *Visual Cognition* **7**: 345–376.

Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford: Oxford University Press.

Personal Identity

Intermediate article

Jennifer Radden, University of Massachusetts, Boston, Massachusetts, USA

CONTENTS

Personal identity is a hotly contested topic in both definition and implication, which questions whether, and in what sense, an individual person can be understood as a uniquely definable entity at any one time (synchronic unity) and as that same entity through time (diachronic unity).

INTRODUCTION

Loosely understood, personal identity is nothing more than the customary unity and integration we expect of lived experience and personality. It is the assumption, presupposed in most social exchange and implicit in most human practices and institutions, that individual persons are one and the same through stretches of time. For all its commonplace and practical side, personal identity understood as a philosophical concept is complex and has generated immense controversy. Among other things, theorists differ over whether personal identity raises a philosophical problem at all, over what the problem is, over what a solution to such a problem might look like, and over the means to reaching such a solution. Recent postmodern theorizing has rejected many of the terms within which these philosophical discussions are conducted, dismissing as myth the unified subject of first-person narratives and the presumption of unified phenomenal experience.

This postmodern skepticism aside, however, so few points of agreement are shared among theorists of personal identity that it is misleading to speak of one philosophical problem or even one philosophical concept here, and an overview of the broad field is better undertaken by characterizing the separate controversies than by focusing on the modest points of agreement.

CENTRAL PHILOSOPHICAL ISSUES

Four different questions allow us to map the complex set of issues which fall under the broad head of philosophical theories of personal identity: (1) whether the reidentification of persons is to be understood on the model of the reidentification of all spatio-temporal particulars; (2) whether person individuation must be dealt with separately from person reidentification; (3) whether personal 'identity', or better 'survival', is a scalar concept, admitting of degree; and (4) whether any questions about personal identity are facts of the matter to be discovered rather than arbitrated.

Person Reidentification

Understood as a composite of bodily and psychological or mental states, a person is one kind of particular, with a unique and uniquely identifying spatio-temporal path in the world. In this respect, persons are not unlike other spatio-temporal particulars, such as chairs. The reidentification of chairs is not always easy in practice, but admits of a simple enough rule: a chair is judged the same between time 1 (t_1) and time 2 (t_2) when the spatio-temporal continuity can be established linking its earlier and later instantiations. This matter is complicated, even with particulars like chairs, inasmuch as we utilize the same language to indicate qualitative as well as numerical identity. 'The same chair' may mean a qualitatively similar chair (same color, style) or the very chair (numerical identity). Context usually prevents us from becoming confused by these potential ambiguities when we are talking about such things as chairs; speakers of English are adroit at distinguishing 'the same x' as bespeaking the spatio-temporal continuity of particulars, and the sameness which rests on similarity of qualities such as color and shape.

Some philosophers have insisted that the spatio-temporal continuity of the person's bodily attributes are similarly the source of personal identity judgments. The body is thus said to be criterial for personal identity (Williams, 1973). On this theory, the continuity of persons is just as strict, no more and no less, and for the same reasons, as the identity of the other material objects we judge to remain

the same through time. This may not be 'identity' in the strict numerical sense. But the workaday spatio-temporal continuity is all we ever need to reidentify particulars, be they chairs or persons, on this view, whether or not we describe the sameness we recognize to be 'identity' properly so called.

The normal person in everyday contexts has a high degree of continuity of several kinds. As well as the continuation of the body through its spatio-temporal path, continuity is also provided by enduring psychological states and dispositions, by memory, by the stream of consciousness, and by persisting capabilities and skills. Rejecting the view that bodily continuity accounts for personal identity, other theorists have identified some of these enduring or relatively enduring attributes as criteria for personal identity. Locke argued that the continuity of memory between earlier and later person or self stages explained personal identity. Several twentieth-century thinkers have pointed to psychological states and dispositions more generally understood to explain the person's sameness or survival through time (Grice, 1941; Parfit, 1984; Shoemaker, 1979). As long as these attributes are taken to be psychological occurrences (particular, datable, introspectible events in consciousness), there will be interruptions in this continuity, such as those resultant from sleep. But by treating these traits as dispositions or capabilities with causal powers (we remain disposed to do, feel, remember, believe, or will, even while asleep) it is not difficult to countenance their providing uninterrupted continuity. On such an analysis, the continuity of psychological states explains the customary presumption that the self or person remains constant through stretches of time, forming a causally related and overlapping series in most people's lives sufficient for us to speak of the same person surviving, if not as maintaining 'identity', strictly understood, through stretches of time.

Person Individuation

The question of whether person individuation must be dealt with separately from person reidentification requires us to consider the person's oneness or unity in two ways. The (synchronic) unity of the person's experiences at a given time and the (diachronic) continuity of the person through time seem to give rise to questions of differentiation, or self-ascription, such as 'On what basis do I know that all these experiences are *my* experiences?' as well as to the questions of reidentification introduced earlier ('Is this the same person between t_1 and t_2?'). But until Hume's famous, fruitless pursuit of the self as an object of experience, these two kinds of question were apparently conflated. The Cartesian subject or self, for which Hume searched in vain, was understood to encompass a unifying principle, linking items in the imagination at a given time, and as well, because it remained unaltered through time, providing the continuity of perfect identity persisting throughout the individual's life.

It was this Cartesian concept of subject or self which Hume believed he exploded. Rather than a simple, united whole which ensured differentiation and synchronic unity as well as reidentification and diachronic unity, Hume discovered an untidy bundle of heterogeneous and ever-changing impressions. Instead of the strict and perfect identity intrinsic to the self associated with its continuation through time, by which it was thought changeless, Hume pictured an imperfect and 'imaginary' or 'fictional' identity, imposed from without for the convenience of the observer.

Hume did not question the self's unity. But he denied the basis for making claims about unity and continuity – the Cartesian notion that underlying these two kinds of identity there stood a metaphysical subject of experiences. If, as Hume concluded, there was no subject of experiences, then the supposed unifying functions of that subject would be lost. Nonetheless, synchronic unity and the consequent differentiation it provided remained – ensured, Hume believed, by the imagination. A loose and relative continuity, misnamed 'identity', based on the relations of contiguity (proximity or closeness), resemblance, and causation linked earlier and later experience stages in a way sufficient for us to speak of the person as remaining the same through time. Hume's is a reductionist theory of the self, and if we speak of selves as unified both at and through time, it must be due to the rough empirical cohesion among and continuities between parts of the self, a cohesion ensured by the imagination.

Some philosophers (Kant was one) have rejected Hume's empirical approach, and appeal is still made to a transcendental subject in explaining the self's synchronic and diachronic unity (Korsgaard, 1989). Moreover, the imagination was hardly sufficient for the task Hume imposed upon it, as others came to recognize. Without the transcendental principle to at once unify the contents of consciousness at any given time and thus provide the means for differentiating one person's experiences from those of another, and to unify the scattered parts of a person's experience through time, permitting

person reidentification, these two kinds of unity required for personal identity were in need of, and received, separate theoretical treatment.

Questions about synchronic unity such as the one noted above ('How do I know these experiences are my experiences?'), and the means of differentiating one person's experiences from another's, concern what is known as self-attribution, or ownership. Different philosophical accounts share the conclusion that self-attribution is grammatically guaranteed: the concept of experience prevents our speaking of unowned experiences, thus 'this experience' entails 'my experience'. Whether this is a trivial or an interesting feature of human experience, however, is disputed.

Some theorists have emphasized the psychological necessity of ownership. As a contingent fact, they suggest, humans have evolved to own all and only our experiences (Dennett, 1991). Others insist that self-ascription is logically guaranteed in any subjective or introspective report. Any first-person account is by its nature an account in which 'this experience' entails 'my experience'. Examining the evidence from psychopathological symptoms of mental alienation, in which a patient appears to deny ownership of his or her experience, others have analyzed normal self-ascription into two parts, one concerning ownership (the assertion that experience *X* occurred within me), and the other agency (the assertion that experience *X* is my mental action, or a thought I think) (Graham and Stephens, 1994; Radden, 1996).This distinction between thoughts which merely occur in us, and thoughts we think, renders self-ascription and the grammar of ownership ambiguous.

Cut loose from the unifying function provided in the transcendental theories of personal identity which are largely rejected today in favor of empirical, psychological continuity analyses, the need for an account of self-attribution becomes more pressing. These efforts to understand self-attribution are important contributions.

Personal 'Identity', Sameness, or Survival as Scalar

If the self is experienced as relatively continuous through time, this may be because of our dispositions to remember, Locke believed. However, we do not remember perfectly. Recognizing this, Grice sketched a memory theory of personal identity which treats memory as an overlapping series (Grice, 1941). A 'total temporary state' (tts) is a set of simultaneous experiences of a single person, and Grice argued that a sequence of tts's comprise parts of the same person (and thus constitute an identity) when certain relations obtain between them: each tts is disposed to remember at least one experience contained in the last, or contains an experience of which the next contains a memory. A single person's tts's thus form an interlocking series, defined by Grice as one in which no subset of members is independent of all the rest. Memory is understood dispositionally here. A tts is 'memorable' when it contains as an element some experience a memory of which would, given certain conditions, occur as an element in some subsequent tts. A tts is 'memorative' when it would, given certain conditions, contain as an element a memory of some experience contained in a previous tts.

Grice's analysis calls for a very modest continuity of memories to ensure the continuation or survival of the same person through time. Though little else were to remain unchanged, one shared memory between two self stages may be sufficient for us to speak of those stages as earlier and later parts of the same person. Grice's account is thus relatively conservative; that is, a (loose) identity or 'singularity' conserving theory. In all but the most extreme cases of discontinuity it allows us to maintain the traditional presumption that just one person inhabits one body for its lifetime (a presumption dubbed the 'one to a customer' rule: Dennett, 1991).

More encompassing than Grice's theory, Parfit's influential empirical continuity argument accommodates not only memory but also other forms of psychological continuity (Parfit, 1984). On this analysis, the 'psychological states' whose continuity explains language suggestive of personal identity include behavioral traits and tendencies as well as occurrent mental items datable to a particular moment – not only memories, dispositions, and personality traits, but also other, more general capabilities, such as how to speak or swim. Like Grice, Parfit uses the notion of an overlapping or interlocking series. Earlier and later self stages housed in a single body at different times have continuity when and to the extent that they overlap. Should psychological states A and B overlap in this way between t_1 and t_3, then there is 'connection' between A and B at t_2. Rather than 'identity' strictly understood, Parfit employs the notion of 'survival'. If person P remains the same between t_1 and t_3 due to such overlapping connection, then P survives between t_1 and t_3. The more traits are continuous, the more survival there will be. So connections admit of, and can hold to, any degree. If between t_1 and t_3 person P's connected traits numbered 100, and person Q's numbered 50, we could conclude

that between t_1 and t_3 person P survived more fully than person Q.

Parfit's psychological continuity analysis permits a broad set of traits to determine survival. Thus, a person in whom brain injury has brought about total amnesia concerning her past life, together with radical personality change, may have retained no trait of memory and personality entitling us to judge her the same person after, as before, her injury. But while there remain some overlapping traits – she can still knit, let us say, or swim – then on Parfit's account we may speak of survival.

The notion of a disposition is a causal one. This means that the continuity provided by interconnected dispositions is 'causally grounded continuity' (Shoemaker, 1984, p. 84). This causal basis of psychological dispositions has been particularly emphasized by Shoemaker, who regards the continuity of persons through time as an instance of the more general continuity of 'continuants', things which we treat as the same through apparent change. Direct psychological connection between earlier and later stages of a person is guaranteed by the later stage's standing in the appropriate relation of causal dependence to a state contained in the earlier stage. (Theorists debate whether, for us to speak of personal 'identity' or survival, these chains of psychological continuity and connectedness must result from normal causal processes. The acquisition of artificially created 'quasi-memories' otherwise bearing the appropriate relation to a person's current states puts pressure on this requirement.)

Because survival admits of degree, the determination that one self has succeeded another is an arbitrary matter calling for decision. How much variance of psychological traits through time ranks as sufficient for us to judge one self to have replaced (or joined) another in a single body will presumably depend on context and on the purposes and interests we bring to the decision.

Personal 'Identity' or Survival as a Matter of Discovery or Adjudication

Some argue that questions about personal identity are facts of the matter – to be discovered, not arbitrated, and this discussion reveals the contrasting approaches and methods found in discussions of personal identity.

In the tradition of Locke's famous request that we imagine the minds of a prince and a cobbler, respectively, transferred to one another's bodies, and allow our intuitions to determine whether the cobbler's or the prince's body housed the person of each, philosophical discussions of personal identity usually employ a thought experiment to establish their claims. As in Locke's model, kinds of continuity are imaginarily eliminated and we are asked to consult intuitive convictions about the survival of the resultant entity and about our moral and other attitudes towards it. Elaborate thought experimental examples of fusion, branching, and brain transplant are constructed in support of these conclusions. For example, if A's brain is transplanted into B's body, we can ask whether reproaching A/B for A's earlier crimes would be right.

Implicit in this methodology is the notion that a metaphysical fact of the matter exists, and can be discovered through patient theorizing. Such a view is fundamentally at odds with the approach which allows that survival is a matter for arbitration.

Some theories of personal identity have been subject to a different kind of challenge. Serious normative costs appear to be associated with the merely empirical self which explains identity language and presuppositions without resort to a numerically identical subject, and admits that several selves might coexist or succeed one another within a bodily lifetime. The effect of Parfit's work, particularly, has been a marshaling of the reasons to value the unified and numerically identical self associated with individualism, and a rehearsal of the normative costs of rejecting it – costs which allegedly entail the loss of central moral and legal conceptions of autonomy, responsibility, agency, blame, praise, and so on. This reasoning takes the form not of discovering a metaphysical truth, but of emphasizing the normative tug of individualism, the several compelling reasons why, even if we could choose to adopt a survival threshold which violated traditional personal identity presuppositions, the costs of doing so in the loss of our most valued concepts, categories, and practices would be unacceptable. Initially raised by those who maintain allegiance to a transcendental subject which serves to ensure strict personal identity, such considerations also affect the position of those accepting an empirical theory of personal survival (Radden, 1996). Conceding the theoretical possibility that several selves or persons might inhabit one body through its lifetime, one might nonetheless argue for a more singularity-conserving survival threshold on the grounds that such normative costs are too high to be tolerated.

For all that they focus on the same issue, these contrasting approaches differ fundamentally. Whether something requires or, alternatively,

invites us to uphold belief in personal identity and/or survival are very different matters, and if we are required to uphold that belief, then we do not need to trouble ourselves over the blandishments of normative individualism.

These four sources of fundamental disagreement provide a rough map of the lines of inquiry found in philosophical discussions of personal identity. As this map indicates, the significance for personal identity of theorizing about brain science will hardly be an uncontested or straightforward matter, but a number of the issues raised above intersect with, and may be affected by, parallel explorations in cognitive science.

THE IMPACT OF COGNITIVE SCIENCE

Speculation about the impact of findings in cognitive science on these traditional philosophical questions of personal identity has been extensive, and may be categorized into four areas.

(1) As the result of surgery, disease, and damage to the brain, a deficiency in specific psychological functions has been correlated with particular brain structure or function. So-called deficit studies in some cases have been germane to issues of personal identity. Thus, for example, experimental data from patients suffering 'split brains' after commissurotomy (severing the connection between the brain hemispheres) has been used to challenge claims about the unity of consciousness (Nagel, 1971). The massive irreversible amnesias resulting from head injury which leave a person without selfhood or normal experience seem to confirm the philosophical claim that personal identity or survival requires memory. In a striking recent example, Damasio implicates the cingulate cortex in generating the 'self in the act of knowing' by showing the way damage to the cingulate cortex affects both the ever-changing consciousness of immediate experience and the extended consciousness in which both known and knower are represented (Damasio, 1999).

(2) So-called functional mental disorders (where no known disease or damage to the brain has yet been identified) have also been linked to philosophical claims about personal identity. In the most vivid of these disorders, multiple personality appears to challenge claims about the unity of the self both synchronically and diachronically understood; and the psychotic symptom of thought insertion in which a person 'disowns' certain mental or experiential content similarly casts doubt, and new light, on usual forms of self-ascription (Stephens and Graham, 2000).

(3) Both theoretical and experimental advances in cognitive science in recent years have provided a range of hypotheses affecting philosophical accounts of personal identity. Certain areas of the brain are apparently implicated in processes like those involved in traditional conceptions of personal identity. Penfield's work has linked the temporal lobe with the familiarity which attaches to feelings and memories (Penfield, 1973); both the limbic system (especially the amygdala) and the prefrontal cortex have been identified as seats of the kind of personal, emotional memory some philosophers regard as constituting personal identity. The complex, iterative nature of memory processes also appears to speak to traditional notions of personal identity (Edelman, 1993). Damasio (1999) has perhaps come closest to philosophical conceptions of personal identity in his speculation about the way brain nuclei, hypothalamus, medial forebrain, and insular and somatosensory cortices create a representation of current bodily states, including a nonconscious synchronic 'core self' modified, moment to moment, by experience, which in turn generates the conscious 'autobiographical self', a feeling of lived past and anticipated future which is diachronic or extended. Closely allied to philosophical questions about the synchronic unity of experience is the so-called binding problem raised by theorists of consciousness: how properties which are represented by spatially separated neurons in the brain come to cohere in the internal structure which allows us to experience the world as objective.

(4) Advances in artificial intelligence offer new ways of construing self-ascription. A self-ascription function which discriminates between self and not-self can be given computational characterization (Perlis, 1999).

In sum, no evidence from cognitive science offers unequivocal proof of claims about personal identity, but significant findings and theories in cognitive science have bearing on the continuing debates between philosophers on these issues.

References

Damasio A (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt Brace.

Dennett DC (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Edelman G (1993) *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York, NY: Basic Books.

Graham G and Stephens L (1994) Mind and Mine. In: Graham G and Stephens L (eds) *Philosophical Psychopathology*. Cambridge, MA: MIT Press.

Grice HP (1941) Personal identity. *Mind* **50**. [Reprinted in Perry J (ed.) *Personal Identity*; Berkeley, CA: University of California Press, 1975.]
Korsgaard C (1989) Personal identity and the unity of agency: a Kantian response to Parfit. *Philosophy and Public Affairs* **18**: 101–132.
Nagel T (1971) Brain bisection and the unity of consciousness. *Synthese* **22**: 396–413.
Parfit D (1984) *Reasons and Persons*. Oxford, UK: Oxford University Press.
Penfield W (1973) *The Mystery of Mind: A Critical Study of Consciousness and the Human Brain*. Princeton, NJ: Princeton University Press.
Perlis D (1999) Consciousness as self-function. In: Gallagher S and Shear J (eds) *Models of the Self*, pp. 131–147. Exeter, UK: Imprint Press.
Radden J (1996) *Divided Minds and Successive Selves: Ethical Issues in Disorders of Identity and Personality*. Cambridge, MA: MIT Press.
Shoemaker S (1979) Identity, properties and causality. In: French P, Uehling TE and Wettstein H (eds) *Studies in Metaphysics*. Midwest Studies in Philosophy, no. 4. Minneapolis, MN: University of Minnesota Press.
Shoemaker S (1984) Personal identity: a materialist's account. In: Shoemaker S and Swinburne R, *Personal Identity*. Oxford, UK: Basil Blackwell.
Stephens GL and Graham G (2000) *When Self Consciousness Breaks*. Cambridge, MA: MIT Press.
Williams B (1973) *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge, UK: Cambridge University Press.

Further Reading

Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press.
Harré R (1998) *The Singular Self: An Introduction to the Psychology of Personhood*. London, UK: Sage.
Hirsch E (1982) *The Concept of Identity*. Oxford, UK: Oxford University Press.
Kolak E and Martin R (eds) (1991) *Self and Identity: Contemporary Philosophical Issues*. New York, NY: Macmillan.
Melzack R (1989) Phantom limbs, the self and the brain. *Canadian Psychology* **30**: 1–16.
Rorty A (ed.) (1976) *The Identities of Persons*. Berkeley, CA: University of California Press.
Wiggins D (1967) *Identity and Spatio-temporal Continuity*. Oxford, UK: Oxford University Press.

Phenomenology

Intermediate article

David Woodruff Smith, University of California, Irvine, California, USA

CONTENTS

Phenomenology is the study of consciousness as experienced from the first-person point of view. Focusing the philosophical theory of mind on intentionality, or mental representation, it lays a foundation for empirical studies of mind in cognitive science.

INTRODUCTION

The theory of mind developed in philosophy from Plato and Aristotle, through Descartes, Locke and Kant, to Brentano and James. Around 1900, drawing crucial distinctions among ideas and their objects, Husserl formulated the basic theory of intentionality that is central to phenomenology. The computational model of mind emerged with cognitive science in the 1970s, and the issue of consciousness became central again in the 1990s. To contemporary cognitive science, phenomenology contributes a developed analysis of conscious intentional experience.

WHAT IS PHENOMENOLOGY?

Phenomenology is the study of consciousness as experienced from the first-person point of view. Its domain is the entire field of conscious experience: including perception, imagination, thought, reasoning, desire, emotion, volition, and embodied action, as well as temporal awareness, awareness of self and personal identity, awareness of others, and practical and social activity. Its focus is on the structure of conscious mental states, or experiences, especially intentionality, that is, the way in which mental states represent or are directed towards various things. (Alternative characterizations of phenomenology are noted below.)

So defined, phenomenology is a discipline: the study of consciousness. Its methodology will be addressed below. There has been controversy as to whether phenomenology should proceed by some form of inner reflection or introspection; or by a more artful form of interpretation of experience akin to textual interpetation; or by an analytic method more like that of logic or linguistics or mathematics; or by some variation on the empirical scientific method of observation, hypothesis, and theory confirmation.

As a program in philosophy, phenomenology would ground or center all philosophy in our own lived experience. As an empirical research program, phenomenology would focus on structures of consciousness as we experience them. The latter program of research intersects with parts of cognitive science. (*See* **Consciousness, Cognitive Theories of**)

HISTORY

Plato spoke of the *psyche* (soul, mind) and the forms or ideas (*eidos*) of things. Then Aristotle proposed that in perception the psyche takes in the form but not the matter of the object perceived. In the Middle Ages Islamic philosophers distinguished form-in-mind from form-in-object, and then the neo-Aristotelian Scholastic philosophers of the fourteenth century dubbed the form-in-mind 'intentio': the mental content that 'aims' at an object.

In the seventeenth century Descartes sharply advanced the conception of mind: firstly, he held, the mind can be known with a kind of certainty while all else is cast in methodological doubt ('I think, therefore I am'); secondly, he held, mind and body are distinct in kind, as a body is extended in space (and time) while a mental state is not. In the eighteenth century, Locke, Hume and Kant further stressed the radically different characters of mind and nature. In the same era Newton's physics laid down mathematical laws of motion, while Locke focused philosophy on the structure of our 'ideas', and so on consciousness and the 'self-consciousness' that, for Locke, distinguishes our conscious mental states. Locke also stressed principles of continuity over time that constitute our personal identity. Hume, however, cast doubt

on our pretensions to knowledge of the existence of the external world. Kant then distinguished 'phenomena', or things as they appear, from 'noumena', or things in themselves, holding that our representations of things were all we could know of things.

In the nineteenth century, following on from these lines of argument, the foundations of phenomenology were laid. First Bernard Bolzano distinguished between subjective ideas (experiences) and objective ideas (including the timeless propositions long studied by logicians). Then Franz Brentano revived the medieval notion of 'intentio': what distinguishes mental from physical phenomena, according to Brentano, is the way in which mental acts are 'directed' towards objects.

By 1900 Edmund Husserl had synthesized these lines of theory – from logic, epistemology, psychology, and ontology – into the discipline he called 'phenomenology' (Husserl, 1900, 1913). The term 'phenomenology' had been used loosely since the seventeenth century, roughly defined as the description of 'phenomena', or appearances, especially sensible qualities of things. Husserl defined phenomenology as the science of the essence of consciousness.

The central thesis of phenomenology, according to Husserl, is that consciousness is always consciousness of something. That is, every act of consciousness is 'intentional' (in Husserl's terminology), or directed towards some object. Alternatively, consciousness represents things in the world. Since the 1970s, cognitive science has stressed the empirical study of mental representation, what Husserl called intentionality. Husserl's great achievement was to analyze the structure of intentionality in general and then to pursue specific forms of intentionality in different forms of experience: in perception, imagination, action, speech, temporal awareness, and intersubjective awareness of other people. Husserl, for the first time, clearly drew the necessary distinctions among subject, act, content, and object of consciousness – though these notions had been developing since Plato and Aristotle. Husserl's account of intentionality, however, was not itself committed to the problematic ontologies of dualism, idealism, or reductive materialism that had dominated philosophy since the seventeenth century – or to the problematic distinction between phenomena and 'things in themselves' that dominated Kantian and post-Kantian philosophy in the eighteenth and nineteenth centuries.

Husserl's work in phenomenology was followed by the writings of Martin Heidegger, Maurice Merleau-Ponty, and Jean-Paul Sartre, often elaborating psychological and social aspects of human existence. Heidegger (1927a,b) stressed the 'being' of human beings in our intentional, practical and cultural activities, downplaying consciousness in favor of our modes of being. Merleau-Ponty (1945) emphasized the role of bodily experience in perception and other forms of human experience. Sartre (1943) stressed our experience of freedom of will, and again our being in the world with others. His account of the 'look' of 'the other' led, through Simone de Beauvoir, to the sociopolitical account of women and minorities being treated and conceived as 'other'.

In the first half of the twentieth century, phenomenologists argued over what is most fundamental in forming intentionality (consciousness or language or social practice), over what is most fundamental in being (the individual self, the act of consciousness, the body or embodied intentional act, the background culture, or 'they'), and over method. Husserl proposed a method of 'bracketing' the object of consciousness in order to focus on the form or content of one's experience, thus describing the object only as it is experienced. Some phenomenologists have pursued this 'transcendental attitude' in a way that resembles Zen meditation, observing the world as we experience it without judgment about the 'natural world' (cf. Sokolowski, 1999). Heidegger practiced a 'hermeneutic' method of interpreting modes of intentional activity within the context of being with others; Merleau-Ponty pursued the description of experience within the context of embodied activity, as even vision is seeing with and by the use of one's body. Phenomenological analyses often went beyond the obviously conscious elements of our experience and into more habitual activities and the background cultural practices that give meaning to our activities. In this way phenomenology spread beyond its original domain, which was the obviously conscious side of our intentional activities.

This work in phenomenology was part of the so-called continental tradition in German and French philosophy, though it developed from the more analytic work of Bolzano, Brentano and Husserl in the Austrian tradition that also produced the positivism of the Vienna Circle in the 1920s and 1930s. Meanwhile, a different philosophy of mind developed in England and then America, including the theories of behaviorism, materialism, and functionalism. Around 1950, logical behaviorism, inspired by Wittgenstein and Ryle, became popular. Ryle had read Husserl and Heidegger sympathetically, but he rejected the Cartesian view that our knowledge of our own mental states is incorrigible,

and he proposed that our language about mental states is logically committed to ascriptions of dispositions to overt behavior. Then materialism (the theory that mind is brain) was revived in the 1950s and 1960s, followed by functionalism (the theory that mind is neural or computational function), which led to the emergence of cognitive science in the 1970s. In these third-person accounts of mind, phenomenology occupied an uneasy position, as we see finally in Dennett (1991), where first-person 'autophenomenology' is rejected in favor of third-person 'heterophenomenology', an analysis of consciousness prompted by neuroscience, which Dennett claims rejects a Cartesian 'theater' where mental events take place in full view of the experiencing subject. (*See* **Materialism**; **Functionalism**)

In the 1970s, however, Dagfinn Føllesdal and others (Dreyfus, 1982; Smith and McIntyre, 1982) stressed the connection between Husserlian phenomenology and modern logical theory. Føllesdal (1969) realigned the theory of intentionality with the theory of reference (and truth) in logical–semantic theory and the philosophy of language.

As cognitive science developed, John Searle (1983, 1992) re-articulated the structure of intentionality. Searle's work was not explicitly phenomenological, but he stressed the first-person ontology of mind, arguing for the irreducibly subjective character of intentionality and consciousness – even though, for Searle, the world remains basically physical, humans basically biological, and the subjective character of mind a natural, biological phenomenon.

As the philosophy of mind developed through the 1980s, consciousness assumed a central position again. Behaviorism had banished consciousness, introspection, and the fruits of phenomenology; but then Nagel (1974) argued that an objective account of the world would necessarily omit the subjective character of mental states, 'what it is like' to experience these states. Gradually, the first-person perspective regained ground. Consciousness was widely studied in the 1990s, and when Chalmers (1996) surveyed the state of the art he declared consciousness – the subjective, first-personal phenomenon – the 'hard problem' for our scientific theory of mind.

PHENOMENOLOGY AS A PHILOSOPHICAL PROGRAM

Phenomenology is not an autonomous discipline designed to study consciousnes for its own sake. For Husserl, Heidegger, and the other pioneers of modern phenomenology, all of philosophy was at stake. Descartes had revolutionized philosophy by turning Aristotelian thought towards pure reason in the light of our own conscious thought. Locke and the empiricists had pursued philosophy through the new 'way of ideas', founding epistemology on sense experience rather than reason. Kant's 'critical' or 'transcendental' philosophy had sought to synthesize rationalism and empiricism, while avoiding the realism–idealism debate by placing all that is knowable within the category of 'phenomena'. Husserl's phenomenology was then designed as a way to approach traditional philosophical issues, by developing a clear method for studying consciousness, and thus knowledge, in an objective, scientific way. Today, phenomenology may be seen to contribute systematically to philosophy through its careful analysis of intentionality. Intentionality is relevant to the philosophy of mind, language, logic, knowledge, reality, and ethical evaluations of right and wrong action. (*See* **Philosophy of Mind**)

The relation between ourselves, our consciousness, and the rest of the world is a crucial theme for philosophy. That relation is called intentionality. According to the phenomenological analysis of intentionality, consciousness consists in a relation between a subject, an act, a content, and an object. The subject is the person, or 'I', who experiences the consciousness; the act is the unit of conscious experience; the content is the image, concept or thought entertained by the subject in experiencing the act; and the object is that which is represented or 'intended' in the act by the content. The overall structure is thus one of representation: consciousness consists in a mental representation of an object by a content in an act of a subject. (*See* **Intentionality**)

We will now discuss these notions in more detail.

Consciousness occurs in units of mental activity called 'acts' (events, states, or processes). Acts of consciousness include particular acts of perception, imagination, thought, desire, will, and so on. Every act of consciousness has a subject, a self or 'I' (*ego* in Latin): the person who has, experiences, or performs the act. Thus, as we say in everyday language, I see or think or will such-and-such. Every act also has a putative object: what I see or think (about) or will. I see a dog or a tree or an automobile: that object, and no other, is the object of my act of consciousness in so seeing. Thus, consciousness has the form of a relation between an act (of a subject) and an object. But this intentional relation is unusual, because in some cases the putative object – the object projected by the act – does not exist. When I see (or seem to see) a snake in the

corner, when there is no such thing present, then my visual experience really has no object; alternatively, its object does not exist. (In one rather unpopular ontology, there is such an object but it lacks existence.)

How do sensations, such as seeing red or feeling pain, fit into the act–object model? The British empiricists isolated 'pure' sensations, and twentieth-century empiricists (such as A. J. Ayer) took sensations either to have special objects called sense data (one sees a patch of red, rather than a tomato) or to have no object at all (one sees 'redly', an experience that is not intentional). Against the strict empiricists, however, Husserl, like Kant, took perception to be a fusion of sensation and conception (one sees, with a visual sensory quality, a red tomato). Recent discussions of sensory 'qualia' have focused on this quality of sensation, as distinguished from the property of intentionality in perception. (*See* **Perception, Philosophical Issues about**)

Different acts of consciousness may represent – 'present' or 'intend' – the same object in different ways. In a popular example, consider my two experiences wherein at dusk I see the evening star and then at dawn I see the morning star, not realizing they are the same. The same object, Venus, is presented in my first experience as 'the evening star' and then in my second experience as 'the morning star'. Accordingly, we must distinguish the content from the object of my experience. The content includes the way the object is represented, while the object itself is what is represented. Alternatively, the content is my concept, image, or percept of the object. The content of an act is sometimes called the 'meaning' of the act (or of the object for the subject in the act). Since different acts (in different times and places, or in different subjects) may share the same content, philosophers say the content is an abstract or ideal entity: something which itself does not have a location in space-time.

Of course, the same person or 'I' may have different experiences, or acts of consciousness. And different acts may have the same or different contents. Furthermore, different contents may represent the same or different objects. Thus, we must distinguish within the structure of intentionality the roles of subject, act, content, and object. Intentionality consists in the relation of representation that obtains between the relevant subject, act, content, and object. (*See* **Representation, Philosophical Issues about**)

No theory of intentionality, or mental representation, can be adequate unless it draws these distinctions and relates these types of entities in this way.

However, at this stage of analysis, there is room to specify further the ontology of the entities that play these roles of subject, act, content, and object. For example, choices among realism, idealism or materialism would have to proceed from here. The program of phenomenological philosophy could thus be pursued within the further assumptions of physicalism, idealism, cultural historicism, etc. One such program (Husserl's, under one interpretation) places phenomenology within a transcendental idealism like Kant's; another (Heidegger's, under one interpretation) pursues phenomenology within a philosophy of cultural practices, or alternatively language games. Another program suspends all metaphysical concerns and practices phenomenology within a reflective way of life – akin perhaps to Buddhist traditions.

In another way, phenomenological philosophy might relate the structure of intentionality to issues in the foundations of logic, mathematics, and science. Husserl (1900) himself developed phenomenology with these concerns in mind.

Finally, ethical and political philosophy may be pursued with phenomenological analyses of relations between self and other. The work of Sartre and his followers pursues this program.

PHENOMENOLOGY AS AN EMPIRICAL RESEARCH PROGRAM

How should we study the essence of consciousness from a first-person perspective? What methodology should phenomenology use? We all experience many states of consciousness and know how to express or articulate them: these methods are built into everyday language, with the psychological verbs, when we say 'I see', 'I hear', 'I think', etc. Yet a scientific formulation of phenomenological method has seemed elusive. Scientific method is regularly taught to students of physics, chemistry, and biology, but phenomenological method is not regularly taught and still seems unfamiliar, especially to scientists.

On the one hand, phenomenology is thoroughly empirical in that it proceeds by observation of our own conscious experiences: by introspection or reflection as we live experience. Yet how do we observe our own experiences? Not by seeing, touching or hearing them. The standard empirical method of noting what we see and hear and touch around us is not the method of observation in phenomenology. Yet we do observe our experiences: we know what it is like to experience familiar forms of consciousness.

On the other hand, phenomenology is highly analytic, like logic or linguistics. Phenomenology analyzes familiar forms of consciousness, somewhat as logic analyzes familiar forms of argument or linguistics analyzes familiar forms of speech.

How should phenomenological analysis proceed? Three different methods are particularly relevant to cognitive science. Broadly, these are: to focus on your own consciousness by 'bracketing' its object; to interpret your experience in terms of its practical context; or to analyze the 'meaning' or form of your consciousness as an instance of a familiar structure (such as logical or linguistic structure in your mother tongue).

Bracketing

In the method of bracketing (used by Husserl, 1913), you bracket the object of your consciousness, that is, put aside the questions of whether the object exists and of what it is really like in itself. Thereby focus on your consciousness of the object: the way the object is given, represented, or intended in your experience. For example, as I see a sheep in a field, I put aside whether it is really there and what it really is (whether a sheep, and so on). Now I can describe my visual experience: 'I see that sheep across the fence, a white sheep with a black face, as I've seen many times before.' This method of bracketing presupposes that we have the ability to focus on our own conscious experiences.

Interpretation

In the method of interpretation (called 'hermeneutics' by Heidegger (1927b) and contextual 'description' by Merleau-Ponty (1945)), you reflect on a particular intentional activity, and interpret its meaning by placing it in its everyday context of significance. For example, as I pick up a hammer and strike a nail, I reflect on the familiar things I encounter and assume while hammering a nail. Thus, I reflect that I am not merely performing the intentional action of hammering the nail; I am engaged in using the hammer to drive in the nail, but not actively thinking about this, or about my hammering style, while I am building a cabinet in my kitchen and selecting tools from my tool chest, and so on. Moreover, I am using my body, engaged in bodily action in hammering, with kinesthetic awareness of what I am doing. This method of interpretation presupposes that we live in the world as embodied, with objects around us, with other people in our community, with extant practices of how we move and act and think and speak.

Analysis

In the method of analysis (used by Smith and McIntyre (1982) with an eye to logical or semantic analysis), you reflect on a particular form of experience you are having (or have recently had), and specify the structure or content of the experience by analyzing this familiar form of experience, somewhat as you may analyze the meaning of a sentence familiar in your own language. The method of analysis presupposes that we have a repertoire of familiar experiences and can reflect on their structure, which we know tacitly from present and prior experience. This requires previous practice, as in logic or linguistics. However, the content of an experience is not usually a linguistic meaning, what philosophers call a 'proposition'; it is usually instead a visual or imaginary image, a concept for which we have no words, or a precept for action for which we have no words. (How do you describe the complex but familiar movement you make as you pick up a hammer and nail and drive the nail into a piece of wood while balancing yourself?)

These three methods are not contradictory (though phenomenologists have argued energetically over them, and their putative assumptions of idealism, dualism, historicism, etc.). Indeed, contemporary phenomenologists might prescribe these three methods in sequence, as a new scientific method for studying consciousness in an empirical way from the first-person perspective. As we move from introspection to contextual interpretation and then to the 'semantics' of experience, we move from observation towards theory. The 'logic' of consciousness is the 'mathematics' of phenomenology. We may then begin to construct, as in any science, a good account or theory of the objects of study: conscious intentional experiences as we know them in our own consciousness. That theory, as it develops, would be tested, and confirmed or refuted, by evaluating particular claims about the structure of consciousness and how these relate to our own experience.

In this way, phenomenology yields an empirical research program that can interact with cognitive science. (Of course, phenomenology can also be developed in other directions, such as literary or cultural criticism, with aims different from those of pure science.)

THE ROLE OF PHENOMENOLOGY IN COGNITIVE SCIENCE

Cognitive science brings the methodology of empirical science (from physics to neuroscience) into

the study of mind, long the province of philosophy. The term 'cognitive science' became widespread in the 1970s in an explicit effort to synthesize the disciplines of psychology, philosophy (of mind), computer science (artificial intelligence), and ultimately neuroscience. The computational model of mind has been a driving force in cognitive science, as algorithms offer mathematical models of certain aspects of mental function. Yet there has been a growing acknowledgement that the problems of consciousness – including qualia, intentionality, and conscious awareness – are not addressed by the computational model, or by any functionalist or reductive physicalist models of the mind.

These problems are the proper domain of phenomenology. Indeed, our understanding of mind begins with our own subjective experience, and phenomenology systematically analyzes our lived conscious experience. Hypotheses of computational or neural function, and experiments designed to test such hypotheses, are another matter. Where cognitive neuroscience seeks to explain how certain patterns of neural activity (implementing certain algorithms) are involved in certain forms of perception, thought, emotion, and so on, the phenomenology of these forms of experience is presupposed (though hardly explicit). It need not be assumed, with Descartes, that we know our own conscious minds with absolute certainty, but only that we experience familiar forms of consciousness, aspects of which are under investigation in neuroscience. (*See* **Neural Correlates of Visual Consciousness**)

In practice, phenomenology and cognitive science meet in many areas, especially in analyses of mental representation or intentionality, as in visual perception. In theory, there are some areas where phenomenology and cognitive science do not meet. Thus, where phenomenology appraises conscious mental states or processes and their character as experienced from the first-person perspective, other parts of cognitive science proceed from a third-person perspective, appraising non-experiential and unexperienced aspects of mental states.

One area where phenomenology and experimental cognitive science clearly interact is in the analysis of blindsight. Our overall theory of mind addresses vision, and phenomenology analyzes familiar cases of conscious visual perception. But then experiments discover the unfamiliar cases of 'blind' sight, or unconscious visual perception, where subjects correctly answer questions about what is before their eyes but insist that they have no awareness of seeing anything at all. These contrasting accounts of conscious vision and blindsight help to factor out of the known structure of perception two important but distinct aspects: intentionality and conscious awareness. It is one thing to see (to have a perception of) something, to take in information visually, but it is something further to be aware of seeing the thing, to consciously see. It is the task of phenomenology to describe and analyze conscious awareness. This description or analysis may then be presupposed in a cognitive theory of blindsight. (*See* **Blindsight**)

There is another way in which experimental cognitive science may interact with phenomenology. Features of conscious experience analyzed in phenomenology may be confirmed and perhaps sharpened by results in neuroscience. For instance, many philosophers have held that what makes a mental state conscious is a kind of awareness the subject has of the state: perhaps in a form of higher-order monitoring of the state. If neuroscience can find a distinctive activity in a certain part of the brain that is operative only when a person is consciously experiencing a given mental state, say, in conscious vision as opposed to blindsight, then the phenomenological analysis of this form of awareness will be confirmed.

Phenomenology and cognitive science intersect in content, and complement and supplement one another in method. None the less, there have been major controversies between the two disciplines, concerning both theory and method. Much work in cognitive science assumes a computational theory of mind. Yet many phenomenologists reject the computational model (and functionalism), concurring with arguments by Searle (1984) and Dreyfus (1979) that intentional content or meaning does not align with input-output rules of computation, and with arguments by Chalmers (1996) that the properties of consciousness are not captured by physicalist or functionalist reduction. (See the many discussions in Petitot *et al.* (1999).) So long as it is assumed that mind is identical with computation, or with neural function (effecting connectionist computation), the phenomenological methods of first-person reflection (cum interpretation and analysis) will seem misguided. However, if it is assumed instead that phenomenological properties of mental states supervene on neural properties of the brain (*pace* Kim, 1998), then the methods of phenomenology will be seen as appropriate for the study of experienced structures of mental activity. In that case phenomenology will take its proper place in relation to cognitive science.

References

Chalmers D (1996) *The Conscious Mind*. Oxford, UK and New York, NY: Oxford University Press.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Dreyfus H (1979) *What Computers Can't Do*. New York, NY: Harper and Row. [Revised edition.]

Dreyfus H (1982) *Husserl, Intentionality and Cognitive Science*. Cambridge, MA: MIT Press.

Føllesdal D (1969) Husserl's notion of Noema. *Journal of Philosophy* **66**: 680–687. [Reprinted in (Dreyfus, 1982).]

Heidegger M (1927a/1982) *The Basic Problems of Phenomenology*, translated by Hofstadter A. Bloomington, IN: Indiana University Press.

Heidegger M (1927b/1968) *Being and Time*, translated by Macquarrie J and Robinson E. New York, NY: Harper and Row.

Husserl E (1900/2001) *Logical Investigations*, vols. I and II, translated by Findlay JN. London: Routledge and Kegan Paul.

Husserl E (1913/1969) *Ideas Pertaining to a Pure Phenomenology and a Phenomenological Philosophy*, vol. I 'General Introduction to Pure Phenomenology', translated by Boyce-Gibson WR. London: George Allen and Unwin, and New York, NY: Humanities Press. [Also translated (1982) by Kersten F. The Hague: Nijhoff.]

Kim J (1998) *Mind in a Physical World*. Cambridge, MA: MIT Press.

Merleau-Ponty M (1945/1962) *Phenomenology of Perception*, translated by Smith C. London: Routledge and Kegan Paul.

Nagel T (1974/1997) What is it like to be a bat? In: Block N *et al*. (eds) *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Petitot J, Varela FJ, Pachoud B and Roy J-M (1999) *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*. Stanford, CA: Stanford University Press/Cambridge University Press.

Sartre J-P (1943/1964) *Being and Nothingness*, translated by Barnes H. New York, NY: Washington Square Press.

Searle J (1983) *Intentionality*. Cambridge, UK: Cambridge University Press.

Searle J (1984) *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Searle J (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Smith DW and McIntyre R (1982) *Husserl and Intentionality*. Dordrecht and Boston, MA: Reidel.

Sokolowski R (1999) *Introduction to Phenomenology*. Cambridge, UK and New York, NY: Cambridge University Press.

Further Reading

Block NJ, Flanagan O and Guzeldere G (eds) (1997) *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Dreyfus H (1991) *Being-In-The-World*. Cambridge, MA: MIT Press.

Mohanty JN (1985) *The Possibility of Transcendental Philosophy*. Dordrecht: Nijhoff.

Smith B and Smith DW (1995) *The Cambridge Companion to Husserl*. Cambridge, UK and New York, NY: Cambridge University Press.

Smith DW (1989) *The Circle of Acquaintance: Perception, Consciousness, and Empathy*. Boston, MA and Dordrecht: Kluwer.

Philosophy of Language

Intermediate article

Kenneth A Taylor, Stanford University, Stanford, California, USA

CONTENTS

Philosophical investigations of language focus on issues about the metaphysical nature of meaning, logical form and compositional semantics, and the pragmatics of conversation.

WHAT IS THE PHILOSOPHY OF LANGUAGE?

Language is a central object of philosophical reflection. Philosophers seek to understand the nature and source of linguistic meaning, the distinctive character of linguistic understanding, how communication works, how language relates to the external world, what the structure of language reveals about the structure of thought and about the nature of cognition more generally. There has been significant cross-fertilization between philosophy and linguistics. This is particularly true of formal semantics and of pragmatics. Nonetheless, there are broad and unmistakable differences of emphasis and methodology. Philosophers often use, but seldom produce, results in syntactic theory. Nor have they been much concerned about linguistic universals, linguistic variation, or linguistic change. Philosophers pay scant attention to phonology and phonetics. On the other hand, linguists have been less engaged than philosophers by certain essentially metaphysical questions about language, such as: how it is possible for expressions in a language to be meaningful at all? What is the mechanism by which words in a language manage to refer to objects in an external world? What is the nature of truth?

A BRIEF HISTORY

Philosophers of antiquity, the medieval period, and early and late modernity all concerned themselves with language and its workings to one degree or another. Beginning near the turn of the twentieth century, the whole of philosophy became reoriented toward the analysis of language. During this linguistic turn, as it has been called, the resolution of issues ostensibly in metaphysics, ethics, the philosophy of mind, and epistemology were taken to depend on the analysis of linguistic meaning. Eventually the linguistic turn gave way to a naturalistic turn. During the naturalistic turn, philosophy of language was demoted from its formerly lofty perch as 'first philosophy' and became merely one philosophical discipline among others. Indeed, naturalistic philosophers tend to view philosophy itself as merely one form of broadly empirical, scientific inquiry among others, possessed of no distinctive methodology. Partly as a consequence of this naturalistic turn, many scientifically minded philosophers of language now see themselves as engaged in a common endeavor with linguists and the cognitive sciences generally. In their view, the borders between these disciplines mark no deep or principled divide.

The linguistic turn had two main sources and two distinct phases. Its first source was the sudden emergence of modern logic, especially in the work of Frege and Whitehead and Russell, with its drastically improved understanding, relative to the systems of classical and medieval antiquity, of the logic of relations and multiple quantification. Though the main advances in logic were directly driven by concerns relating to the foundations of mathematics, those advances had far reaching implications for the philosophical study of language in general. The power and precision afforded by the newly emerging logical calculi, together with the evident distance between those calculi and the resources of traditional grammars, convinced many that the traditional grammars were, from a logical point of view, both unrevealing and misleading. Moreover, they believed that the logically unrevealing and misleading nature of surface grammar was a significant source of philosophical confusion and error. Uncovering the true 'logical grammar' of language and thought was taken to be key both to avoiding error and to positive philosophical

progress. Laying bare the true logical grammar of language and thought would enhance philosophical understanding of the logic of inference and would drastically diminish temptations to ontological extravagance.

The search for logical grammar bears a superficial affinity to the contemporary linguist's search for so-called competence grammars. Like the logical grammarian, many contemporary linguists believe that language has significant 'hidden' structure not revealed by either superficial word order or traditional grammar. But one should not take this line of reasoning too far. Contemporary linguists are engaged in a descriptive, explanatory enterprise. They want to know the psychological basis of our actual linguistic competence. The logical grammarians, on the hand, were really concerned to say what a 'logically perfect' language would be like. Only a logically perfect language could serve, for example, as a perspicuous medium for the expression of thought, or for carrying out rigorous and complete mathematical proofs, or for conducting scientific inquiry. They struggled mightily to say just what a logically perfect language would be like and much of what they said about actual natural languages *per se* was intended to point up their logical imperfections. Because natural languages are shot through with logical imperfections, the logical grammarians took natural languages to be of derivative interest at best. They were hardly concerned at all to describe and explain the psychological basis of our competence in them. (*See* **Performance and Competence**)

If one takes the search for competence grammars to be a search after the language of thought, then the assimilation of logical grammar to competence grammars is more nearly justified. But one should still exercise considerable caution. As true as it was that the logical grammarians were attempting to outline the structure of the medium for thought, they were deeply and avowedly anti-psychologistic. They were more concerned with thinking as it ought to be rather than with thinking as it is. Frege was quite explicit. He distinguished the thought grasped by a thinker from the individual psychological act via which a thinker comes to grasp the relevant thought. The individual act, he held, is a psychological process that takes place over time and within the mind or brain of an individual thinker. Such acts are subject to psychological laws. He professed no interest in such laws. Indeed, he distinguished sharply between the laws of thought and the laws of psychology. The laws of thought govern the unchanging and timeless structure of thought. They are to be articulated not by psychology but by logic. The laws of thought determine that certain thoughts follow from other thoughts. But the *following from* relation he took to be a logical one and not a causal–temporal one at all. (*See* **Language of Thought**)

An enduring landmark of the first wave of the linguistic turn was Russell's theory of descriptions. The so-called definite description 'the present king of France' may appear to be a genuine constituent of the sentence 'the present king of France is bald' and may appear to serve the function of standing for an object. But Russell argued that such expressions are a sort of logical illusion and that they disappear when we execute a proper logical analysis. The eye-opening clarity of Russell's arguments helped to set philosophers off on a relentless search for hidden logical forms. The path-breaking work of Frege played a significant role as well. So too did the work of the early Wittgenstein. Indeed, Frege's work serves still as the point of departure for much contemporary work. This is especially true of his distinction between sense and reference and his explanation, defense, and elaboration of the consequences of the principle of compositionality in semantics. Frege's general approach received systematic and rigorous elaboration at the hands of thinkers like Carnap, Church, and Montague; but until the advent of Quine's wide-ranging attack on the very coherence of notions like sense and reference in the 1950s, the subsequent emergence of Davidson's purely 'extensionalist' approach to semantics, and the still later emergence of the so-called new theory of reference in the work of Kripke, Kaplan, and Putnam in the 1970s, there were no systematic competitors to Frege's main theories.

The second source of the linguistic turn in philosophy was the widespread acceptance of the analytic/synthetic distinction. A statement is analytic just in case its status as true or false depends solely on the meanings of the terms the statement contains and on how those terms are combined. The statement that bachelors are unmarried is paradigmatic. A statement is synthetic if it is not analytic. If a statement is synthetic, fixing the meanings and organization of the terms will not suffice to determine whether the statement is true or false. Though the roots of this distinction reach back to the eighteenth century, the rise of modern logic enabled philosophers to offer a more precise and rigorous characterization of that distinction. (*See* **Sense and Reference**; **Reference, Theories of**)

The analytic/synthetic distinction was central to the logical positivist's defense of logical empiricism. Logical empiricists took the entire fabric of

our knowledge to be a vast logical construction out of the givens of sensation. Once the world order is seen as a vast array of logical complexes, logically constructed out of the logically simple elements of the stream of sensation, logic alone promises to be sufficient to spread warrant upward from statements about the stream of sensations to statements about the vast array of putative 'logical complexes' which the sciences reveal to us. This entire approach demanded, however, that statements about putative logical complexes – middle sized objects, persons, minds, to name but a few – be analytically reducible and thus translatable without loss of meaning into statements about the course of the stream of sensation. Though logical empiricists struggled mightily to provide such reductions, this exercise ultimately proved as futile as it was ambitious.

The second wave of the linguistic turn had the character of a counterrevolution. The main inspiration of that counterrevolution was the work of the later Wittgenstein. Though Wittgenstein was a pioneering logical grammarian, he eventually came to see the very idea of a logically perfect language as misguided. Natural languages, he claimed, are 'perfectly in order' as they stand and are well suited to the multifold and workaday uses to which we put them in ordinary life. Philosophers who seek to understand language and its powers should not seek after some logical ideal; they should seek to understand the hurly-burly workings of living, breathing languages. Wittgenstein had come to believe that the traditional problems of philosophy arise one and all merely from the abuse of language, from language 'gone on holiday'. Philosophy properly done will strive merely to describe the various language games in which certain putatively philosophically problematic expressions have their real homes. The professed goal of philosophy done in this style is not so much to solve philosophical problems but to dissolve them and to cure philosophers of the very urge to theorize beyond the boundaries of ordinary language.

Despite its destructive tendencies, Wittgenstein's work played a decisive role in the rise of so-called ordinary language philosophy which in turn played a significant role in the rise of the pragmatics of communication as an area of major concern within the philosophy of language. The central thought of the ordinary language philosophers was that the key to discovering the nature of some philosophically problematic item – such as goodness, knowledge, or experience – is to determine what it is that we say of a thing when we say that it is good, or of a person that she knows that such and such or has experience as of so and so. This we can do by examining the meanings of relevant expressions. To know the meaning of an expression is to know how it is used in the language game that is its home. With the analyses of meaning in hand, philosophical issues of a more traditional sort can be addressed, sometimes with surprising results. Ordinary language philosophers positively delighted in demonstrating that some philosophically problematic discourse was not, after all, 'fact-stating' discourse but served some heretofore unnoticed communicative purpose – such as expressing our attitudes of approval or disapproval, or issuing 'inference tickets' of various sorts.

At their best, ordinary language philosophers made many penetrating and subtle observations about the multiplicity of things that we do with language. The work of J. L. Austin is a sterling example. But the legitimacy of their methodology depended crucially on there being a principled distinction between matters of meaning and matters of belief. Consider the use of the term 'bankruptcy'. Suppose that it is universally believed that only slovenly people declare bankruptcy. That would not suffice to make it the case that part of what we say when we say that a person is bankrupt is that that person is slovenly. Yet, in a linguistic community in which such a belief was universal, saying that a person is bankrupt might well convey that she is slovenly. Without some principled basis for distinguishing what is strictly literally said by an utterance from what is merely 'conveyed' without being said, the methodology of ordinary language philosophy is hopeless. It was Paul Grice who first and most clearly grasped the need to distinguish what is said by an utterance from what is 'implicated' or 'conveyed' without being said. His masterful attempt to provide such a basis nearly single-handedly laid an enduring foundation for the study of pragmatics.

Though Grice's work dashed some of the methodological pretensions of the ordinary language philosophers, he remained deeply embedded in the linguistic turn. With the advent and widespread acceptance of Quine's attack on the very coherence of the analytic/synthetic distinction, however, the linguistic turn began to unravel entirely and philosophy of language began to lose its status as first philosophy. This unraveling was greatly helped along by Putnam and Kripke. Particularly crucial was Kripke's defense of the claim that there are necessary truths that are neither analytic nor discoverable on any *a priori* basis. Kripke made a convincing case that empirical inquiry has the power to deliver knowledge of so-called

a posteriori necessities. Though Kripke's arguments and Quine's are too complex to recapitulate here, their joint role in reorienting philosophy can hardly be overstated. If there is a principled basis for the analytic/synthetic distinction and if all and only analytic truths are true *a priori*, there can be a distinctively philosophical methodology – linguistic or conceptual analysis – with the power to deliver truths about the constitutive natures of things. If there is no principled basis to that distinction and if, moreover, science itself is in the business of discovering constitutive natures, then the method of conceptual analysis has no legitimacy and no peculiar power to reveal the constitutive natures of things.

Despite unseating philosophy from its privileged perch above the fray of empirical inquiry, the unraveling of the linguistic turn led not to the end of philosophy but to its reorientation. If there are no distinctively philosophical methods, many concluded, then philosophical inquiry is, and ought to be, broadly continuous with empirical scientific inquiry generally. Though philosophers did not rush to become experimentalists as a result of the naturalistic turn, many did come to see their work as being related to work in the empirical cognitive sciences, say, in much the way that work in theoretical physics relates to work in experimental and/or applied physics. Of course, since one can seldom declare complete victory in philosophy, there remain to this day many serious minded philosophers who endorse the analytic/synthetic distinction and with it an a prioristic methodology akin to that practiced by linguistic philosophers of an earlier day.

AREAS OF THE PHILOSOPHY OF LANGUAGE

Contemporary work in the philosophy of language clusters within three broad areas: (1) the metaphysics of meaning and reference; (2) formal semantics; (3) the pragmatics of communication.

Metaphysics of Meaning and Reference

Philosophers concerned with the metaphysics of meaning are preoccupied with two main questions: what endows an expression with meaning? What is the nature of meaning? The endowment question arises because expressions in language are evidently not intrinsically meaningful. Something in nature, human psychology, or human social life has to make it the case that linguistic expressions are meaningful at all. The nature question arises in part because of the manifest variety of linguistic expressions, the variety of different things we do with language, our ability to grasp meaning, and puzzles about the place of meanings in the order of things. Thus the nature question might be put this way: what are linguistic meanings such that: (1) they are potentially graspable by us; (2) they have a place in the natural order of things; (3) they play central roles in explaining the plethora of different things we do with language?

Philosophers of language have proposed a variety of answers to both the endowment question and the nature question. Sentences have been said to have propositions, truth conditions, assertability conditions, or states of affairs as their meanings. And each of these candidate meanings has been construed in a wide variety of ways. Some, for example, take propositions to be abstract, structured mind-independent, extra-linguistic entities. Others are deflationist and claim that propositions are pleonastic entities with no substantial nature. Still others hold that some propositions have concrete constituents.

In connection with the endowment question, philosophers differ over the relative priority of sentence meaning and word meaning. Holists maintain that sentences, or even entire languages, are the primary bearers of linguistic meaning and they take words to have meaning in only a secondary or derivative sense, in so far as they play roles in sentences. Atomists ascribe priority to word meaning and hold that sentence meanings are 'composed' from antecedently given word meanings, in much the way that a house is composed, brick by brick, from antecedently given constituents. Holists tend to think that only sentences can be directly endowed with meaning on the grounds that it is only in the production of sentences that we perform determinate speech acts. In this view, words are endowed with meanings only indirectly in virtue of the speech acts we perform with the sentences in which they are contained. By contrast, atomists believe that words can be directly endowed with meaning and that sentences have their meanings as a consequence of the meanings of the words they contain and the way those words are combined.

An endowment question and a nature question also arise for reference. Atomists tend to think of reference as a real word–world relation, the full fledged nature of which is an open and substantive issue. They take the reference relation to play a substantive role in explaining the power of sentences to express claims about the world. The fact that name 'Smith' stands for or refers to Smith is

supposed to explain, for example, why sentences containing that name can be used to make claims about Smith. But unless the atomists can say what it is for a term to refer to an object, the appeal to the reference relation will be explanatorily idle. Many atomists favor some version or other of the so-called causal theory of reference. A causal theory of reference holds that an expression *e* refers to an object *o*, if *o*-involving events play a distinguished causal role in the production of instances of *e*. The challenge for any such theory is to say just what causal relation is the reference-making relation.

By contrast, holists tend either to deny that there is any such thing as the reference relation, or to deny that it has a substantive nature waiting to be uncovered by either philosophical speculation or empirical investigation. If there is no such relation as reference, or if the reference relation has no substantial nature, talk of reference will play no role in explaining the very possibility of aboutness.

Formal Semantics

Work in formal semantics is closely connected to the nature question. Once it is determined what sorts of meanings are enjoyed by expressions of various sorts, space opens up for a precise and systematic account of how such meanings combine and interact. Formal semanticists seek rigorous, exhaustive, and systematic accounts of the rules and/or principles which determine the meanings of syntactically complex expressions as functions of the meanings and syntactic organization of their parts. The array of alternative approaches and theories here is breathtaking. Merely citing their names may give the reader a feel for the number of extant alternatives: possible worlds semantics, situation semantics, game-theoretic semantics, dynamic semantics, discourse representation theory, model-theoretic semantics, inferential role semantics, categorial grammar. Contemporary formal semanticists are the heirs of the early logical grammarians, but they differ from the logical grammarians in seeking to deploy rigorous formal methods to describe the semantic workings of language as it is and not merely to characterize some abstract logical ideal.

PRAGMATICS

Four kinds of phenomena constitute the core subject matter of pragmatics: speech acts; indexicality and other forms of context-sensitivity; non-truth-conditional aspects of meaning; and conversational implicatures and the like.

Speech Acts

Speech act theory seeks to characterize the totality of so-called illocutionary acts. Such acts come in an apparently dizzying variety: assertions, boasts, promises, threats, requests, commands, vows, greetings, questions, to name just a few. But many have claimed that there is much system and order beneath the variety. The problem of exhaustively and systematically characterizing the topology of illocutionary act space, as we might call it, is a major task for any student of pragmatics.

Context-Sensitivity

Consider the sentence 'I am hungry now'. Taken apart from any particular context, this sentence expresses no determinate proposition. It expresses one proposition in one speaker's mouth and a different proposition in another speaker's mouth. But this sentence does have a certain fixed significance which does not vary from context to context. For example, it is a fact about its fixed significance that for any utterance *u* of it produced by speaker *s* at time *t*, *u* will be true just in case *s* is hungry at *t*. We can represent the situation as follows: Fixed significance + context → proposition. Several questions naturally arise. What is the exact character of the fixed semantic significance of context-sensitive expressions? What is a context? How does context interact with the fixed significance to yield further semantic values? Philosophers interested in pragmatics have proposed a variety of answers to such questions.

Demonstratives like 'this' or 'that', indexicals like 'I', 'here' and 'now', tense, and aspect all introduce elements of contextual variability. With such-constructions, there are relatively precise rules that determine the semantic content of the relevant expressions or constructions as a function of 'objective' features of the context of occurrence, such as who is speaking where or when. Such rules are not always available. The possessive introduces an element of context-sensitivity, but there is no precise rule for determining the semantic significance of a particular occurrence of the possessive as a function of objective features of the context. Context-sensitivity in the absence of precisely specifiable rules has become a subject of considerable philosophical and linguistic investigation.

Non-truth-conditional Ingredients of Meaning

Consider the following pairs:

John went to that wild, raucous party and did not drink. (1)

John went to that wild, raucous party, but did not drink. (2)

The fact that there is no scenario in which (1) and (2) differ in truth value leads many to conclude that 'and' and 'but' are truth conditionally equivalent. Nonetheless, it seems clear that there is a difference in meaning between 'but' and 'and': 'but' is contrastive but 'and' is not. (1) does and (2) does not 'intimate' that there is a contrast between John's going to the wild raucous party and his not drinking. The exact contrast cannot simply be read off of the relevant sentence, but is in some way a function of context. It is a fact about the meaning of 'but' that an occurrence of 'but' in a sentence serves to set up some contextually determined contrast between conjuncts. 'But' is a striking and much studied example of an expression whose meaning contains ingredients which are not potential ingredients of truth conditions. Such non-truth-conditional ingredients are reasonably studied under the rubric of pragmatics because they relate to and give indications of the kind of speech act a speaker performs in using the relevant expression.

Conversational Implicatures and other Pragmatic Externalities

We often convey by our utterances more than we strictly, literally, say. Smith asks whether Jones wants to go out to a movie with him. Jones replies that she has a great deal of work to do. Though Jones has not explicitly said that she has no time for a movie, she has implied as much. Grice called such implications 'conversational implicatures' and developed a groundbreaking theory, based on the notion that conversation is a form of rational cooperation, to explain how they are generated.

Conversational implicatures must be distinguished from garden variety logical entailments. Unlike logical entailments, conversational implicatures exhibit a high degree of cancellability. Jones might have continued the conversation with Smith by saying, 'I have a great deal of work to do, but I really could use a break.' Here Jones forestalls a conversational implicature which might otherwise be generated. We cannot do the same with logical entailments.

Consider the old sexist chestnut 'Have you stopped beating your wife?' One who utters the relevant question in some sense 'implies' that her interlocutor formerly beat his wife. She does not assert or say as much. This kind of implication is commonly known as *presupposition*. A presupposition of a speech act is, roughly, a condition which must be presumed to be satisfied in order for that speech act to be felicitously and non-defectively performed. Suppose that the person to whom this query is addressed has manifestly never once beaten his wife. The very question is illegitimate. Neither the answer 'Yes, I have stopped beating my wife' nor the answer 'No, I have not stopped beating my wife' seems appropriate.

Presupposition remains a topic of considerable debate. Some think of presupposition as a semantic phenomenon. Others take presupposition to be largely pragmatic in nature.

RELEVANCE OF THE PHILOSOPHY OF LANGUAGE TO COGNITIVE SCIENCE

Work in the philosophy of language is relevant to cognitive science in two ways. To the extent that cognitive science is concerned with the nature of language and the plethora of things we do with language, philosophy of language is one more source of insight. Indeed, naturalistically and scientifically minded philosophers of language see themselves as engaged in something of a common enterprise with linguists. Moreover, philosophers of language are expert in clarifying what we mean, for example, by terms like 'representation,' 'consciousness', or 'rationality.' If one endorses the analytic synthetic distinction, one may take this expertise as expertise at the *a priori* specification of the constitutive natures of things. Philosophy thus conceived purports to provide *a priori* specifications of what it is to be a conscious mental state or what it is to be a representation. Even without the analytic/synthetic distinction, one can still afford to give close attention to what we mean by terms expressive of notions of concern to cognitive scientists. Such attention can play an indispensable role both in providing an initial orientation for empirical inquiry and in making interpretive sense of the intermediate and final products of such inquiry.

Further Reading

Austin JL (1975) *How to Do things With Words*, 2nd edn. Cambridge, MA: Harvard University Press.

Carnap R (1967) *The Logical Structure of the World; Pseudoproblems in Philosophy*. Translated by RA George. Berkeley, CA: University of California Press.

Frege G ([1967], 1879) *Begriffschift*. English translation in vam Heijenoort J (ed.) *From Frege to Goedel*. Cambridge, MA: Harvard University Press.

Frege G (1977) *Translations from the Philosophical Writings of Gottlöb Frege*. In: Geach P and Black M (eds) Oxford, UK: Basil Blackwell.

Grice HP (1989) *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.

Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Putnam H (1975) The meaning of 'meaning', in *Philosophical Papers*, vol 2. Cambridge, UK: Cambridge University Press.

Quine WV (1960) *Word and Object*. Cambridge, MA: MIT Press.

Russell B (1905) On denoting. *Mind* **14**: 479–493.

Wittgenstein L (1953) *Philosophical Investigations*. Oxford, UK: Basil Blackwell

Philosophy of Linguistics

Intermediate article

Barbara C Scholz, San José State University, San José, California, USA

CONTENTS

The philosophy of linguistics addresses both general issues in the philosophy of science that occur in the scientific study of language, and philosophical issues that arise in linguistic theorizing. It shares many issues with the philosophies of language and of mind, but is distinct from them in that it focuses directly on questions about linguistic theories, the epistemology and metaphysics of linguistic science, and the psychology of language.

INTRODUCTION

Linguists investigate many aspects of natural language, but current debates in the philosophy of linguistics are generally dominated by issues that arise in syntactic and semantic theory. There are exceptions: there has been sporadic philosophical interest in phonology (Twaddell, 1935; Botha, 1972; Bromberger and Halle, 1992), and occasional attention to biological and social aspects of language (Millikan, 1984; Itkonen, 1983). But Noam Chomsky's success in popularizing an approach to the study of language and languages centered on syntax has dominated discussion.

The following four questions have figured prominently: (1) what natural languages are like; (2) what linguistics is about; (3) how language acquisition can be explained; and (4) whether language use is governed by normative standards. The answers to these questions are conceptually distinct, but they are connected in Chomsky's research program, which claims that: (1) natural languages are generative systems; (2) those grammars are realized in the brain states of competent speakers; (3) they are acquired through the assistance of an innate system of tacit knowledge about universal grammatical principles; and (4) they participate causally in the production of linguistic behavior.

Unlike its competitors, the Chomskian program has articulated a set of fundamental questions and given an integrated response to them. This integrated approach has given the program very broad appeal; and much recent philosophy of linguistics looks like an extended commentary on Chomsky's work.

WHAT ARE NATURAL LANGUAGES?

The currently dominant view about what natural languages are emerged in the 1950s when Chomsky (1957) applied the production systems developed in mathematical logic to the description of natural languages, yielding what is now called 'generative grammar'. A generative grammar is comprised of a lexicon and a set of rules (or parametrized principles extensionally equivalent to a set of rules) that generate (recursively enumerate) explicit structural descriptions. These structural descriptions are the theoretical reconstruction of natural-language expressions. On this view a natural language is a generative grammar that states the characteristic function of a set of expressions in a way analogous with the formal 'languages' of logic and computer science.

This technical conception of a natural language shares little with commonsense notions of language as a shared way of speaking. A generative grammar, by definition, requires specification of a fixed and closed lexicon. The addition of even one new vocabulary item to that grammar produces a new and distinct language.

Advocates of the generative conception of language claim that the commonsense notion has no scientific interest. This view is based, in part, on the controversial claim that only generative grammars can provide fully explicit descriptions of natural-language expressions. It is perhaps also based on the uncontroversial claim that generative grammars state precise individuation conditions for languages, while the commonsense conception of a language has no precise individuation conditions.

Not all linguistic theories take languages as generative grammars. Bloomfield (1933) took a language to be the totality of all the physically possible utterances of a speech community. He took this to

mean that a natural language is a finite set of expressions. Against this, generative linguists have argued that, firstly, natural language use is creative, and secondly, there is no longest sentence in a natural language. Assuming that native speakers have the ability to produce and understand novel, well-formed expressions, and there is no longest expression, a generative grammar defines a denumerably infinite set of expressions. From this, proponents of generative grammar conclude that natural languages are infinite. But the argument is question-begging, since natural languages are infinite only if it is further assumed that a language is a generative grammar – and that is precisely what is at issue.

Evans (1981) argued that the size of the expression set generated by a grammar is irrelevant to linguistic creativity. What is required for creativity is that a language is acquired by learning the systematic recombinability of its linguistic components. In other words, it is systematicity, not size of the generated expression set, that explicates linguistic creativity.

WHAT IS LINGUISTICS ABOUT?

Do the theoretical terms of linguistic theories refer? If so, to what? Early debates in phonology focused on the referent of the term 'phoneme'. Sapir (1925) held that phonemes are mentally real; and Bloomfield (1926) argued that phonemes are physically real. By contrast, Twaddell (1935), citing earlier remarks of Otto Jespersen's, argued that phonemes are a useful fiction.

Hockett (1948), declaring fictionalist views to be misguided, proposed that the structures identified by linguistic theories denote brain states. He took the result of a person's language acquisition to consist in 'a mass of varying synaptic potentials in his central nervous system'. This view was later much elaborated by Chomsky. Katz (1981) refers to it as 'conceptualism'.

Chomsky (1986, 1997) conjoins the conceptualist view that a language is 'a relatively stable element of transitory brain states' with the claim that a language is a generative grammar, and calls the resultant object an 'I-language'. The 'I' of 'I-language' is an allusion to three claims about language: that it is individual, internal, and intensional. (Before 1986, I-languages were usually called 'linguistic competence states'.)

Itkonen (1978) argues against Chomskian conceptualism. Katz (1981) independently takes a similar line to Itkonen, and argues for a kind of Platonism on which the theoretical terms of generative grammars denote abstract objects with no causal efficacy.

The debate between Chomskian conceptualists and Katzian Platonists assumes that the theoretical vocabulary of linguistics denotes only one ontological category of object. This assumption has not gone unchallenged: George (1989) argues that in generative grammar the term 'grammar' is polysemous, denoting abstract objects, the content of mental states, and causally efficacious brain states. If George is right, then conceptualist–Platonist disputes rest on an equivocation.

HOW CAN LANGUAGE LEARNING BE EXPLAINED?

'Linguistic nativism' is the view that language acquisition requires innate information with linguistic content (the so-called 'substantive linguistic universals') utilized by specialized cognitive structures specific to linguistic information processing (e.g. linguistic processing modules). However, the dispute between linguistic nativists and their opponents is not a dispute over whether sensory organs are specialized devices for processing auditory information. It is focused on whether there is innate linguistic content (often misleadingly called 'domain-specific content'). Without this claim, an opponent of linguistic nativism might argue that a complex variety of innate, informationally biased sensory–perceptual modules, with no specifically linguistic content, are required for first-language acquisition.

The most popular argument for linguistic nativism is called the 'argument from the poverty of the stimulus'. There are two forms of this argument: one empirical, the other *a priori*. The empirical versions contain an empirical premise that claims that children (at some particular age) know some rule or generalization about the target language that they could not have learned from their linguistic experience. Since children's experience seems to be too impoverished to support induction of what children actually know, nativists conjecture that language learning is assisted by universal linguistically contentful innate rules (or parameters), called 'universal grammar'.

Any argument for this conjecture needs to articulate: exactly what counts as exposure to and uptake of evidence by the child; an adequate identification of what is learned; and a reason to think that the child does not receive enough information to induce what he or she comes to know.

Statistical analyses of appropriate corpora might help to provide support for the latter. Pullum and

Scholz (2002) present the results of preliminary corpus searches for some of the most famous cases of alleged stimulus impoverishment. These results suggest that the linguistic evidence may not be so impoverished after all.

There are at least two kinds of *a priori* arguments from the poverty of the stimulus. One is based on the simple fact that theories (and in particular, I-languages) are deductively underdetermined by the evidence. The other is based on a result in mathematical learning theory.

The argument from underdetermination assumes that children learn languages by deducing a unique I-language from some finite body of evidence. Since no theory, generative or otherwise, is entailed by any finite number of observations, the nativist concludes that child language learning requires innate knowledge of linguistic universals with specifically linguistic content, which together with the evidence entails an I-language. However, the assumption has been challenged (Scholz and Pullum, 2002).

The *a priori* argument for linguistic nativism based on mathematical learning theory is due to the results of Gold (1967). Gold proved that for all well-known classes of formal languages (regular, context-free, etc.), learning, under one particular definition, is impossible if the evidence is limited to positive evidence, i.e. grammatical strings. However, Gold showed that if negative evidence (about what is not in the language) is supplied, then learning by an algorithmic, inductive method is possible.

Cowie (1999) and others have pointed out that positive evidence can provide implicit negative information: absence of evidence can be evidence of absence. The problem with this line of argument is that there is so much implicit evidence of absence that false hypotheses about the target language would be supported (for example, the absence of utterances in which *papacy* and *wombat* occur together in a clause would spuriously support the hypothesis that their co-occurrence is syntactically forbidden).

Gold's work assumes that learning a natural language amounts to inducing a generative grammar. Thus, his theorem can be seen as a *reductio ad absurdum* of this assumption. A conservative revision of the program might propose that language learners acquire some grammar – perhaps necessary conditions on subsentential expression recombinability – but not a generative grammar. More radically, it could be proposed that learners acquire no grammar at all.

IS LANGUAGE NORMATIVE?

Is an I-language a standard of correct (grammatical, acceptable, interpretable) linguistic performance? Are speakers mistaken when their linguistic performance diverges from what their I-language generates? If so, can having an I-language be explained naturalistically? These questions arise for all who claim that grammatical rules guide linguistic performance.

Structuralist linguists emphasized the difference between explicit description of languages and traditional prescriptive injunctions (e.g. 'don't split infinitives'). Generative grammarians followed suit: I-languages are not normative in the sense of being prescriptive.

However, I-languages generate a set of structural descriptions that are claimed by generative grammarians to be the well-formed expressions for the speaker. This suggests they provide a standard in virtue of which an individual speaker's actual performance is correct or mistaken. Thus the concept of an I-language seems to have normative force because it incorporates a notion of individual error.

An extended argument that there can be no naturalistic or causal explanation of rule-guided behavior was elaborated by Kripke (1982) (from suggestions by Wittgenstein). Chomsky (1986) responded by claiming that I-languages play a purely causal role in linguistic performance and are not a standard of correct individual performance. Wright (1989) argued to the contrary that the Wittgensteinian argument poses a deep challenge to the Chomskian research program. Linguists do take I-languages as standards of correct performance; performance errors are taken to be mistakes, not indications of broken causal mechanisms.

In addition, the hypothesized innate universal grammar suggests that language is normative in two related senses, both biological: one issuing from the claim that innate universal grammar aims to explain what all 'normal' humans beings share; the other from the claim that innate substantive universals are biologically encoded information.

It is a substantive question in the philosophy of biology whether the appeal to what is 'normal' for a species constitutes an illegitimate remnant of folk-biological explanation. If the linguistic content of substantive universals is claimed to be genetically encoded information, the question is whether a non-semantic notion of information is adequate to explicate substantive linguistic universals.

SUMMARY

That a language is a generative grammar is the fundamental thesis of the Chomskian research program. It plays a central role in integrating answers to the main questions of that program. But the resulting integrated answer comprises distinct claims that require separate and independent support.

References

Bloomfield L (1926) A set of postulates for the science of language. *Language* **2**: 153–164.

Bloomfield L (1933) *Language*. New York, NY: Henry Holt.

Botha RP (1972) *Methodological Aspects of Transformational Generative Phonology*. The Hague: Mouton.

Bromberger S and Halle M (1992) The ontology of phonology. In: Bromberger S (ed.) *On What We Know We Don't Know*, pp. 209–228. Chicago, IL: University of Chicago Press.

Chomsky N (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky N (1986) *Knowledge of Language*. New York, NY: Praeger.

Chomsky N (1997) Language from an internalist perspective. In: Johnson DM and Erneling CE (eds) *The Future of the Cognitive Revolution*, pp. 118–135.

Cowie F (1999) *What's Within?* New York, NY: Oxford University Press.

Evans G (1981) Reply: semantic theory and tacit knowledge. In: Holtzman S and Leich C (eds) *Wittgenstein: To Follow a Rule*, pp. 118–137. Cambridge, UK: Cambridge University Press.

George A (1989) How not to become confused about linguistics. In: George A (ed.) *Reflections on Chomsky*, pp. 90–110. Oxford: Blackwell.

Gold EM (1967) Language identification in the limit. *Information and Control* **10**: 447–474.

Hockett CF (1948) A note on 'structure'. *International Journal of American Linguistics* **14**: 269–271.

Itkonen E (1978) *Grammatical Theory and Metascience*. Amsterdam: John Benjamins.

Itkonen E (1983) *Causality in Linguistic Theory*. Bloomington, IN: Indiana University Press.

Katz JJ (1981) *Language and Other Abstract Objects*. Totowa, NJ: Rowman and Littlefield.

Kripke S (1982) *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Millikan RG (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Pullum GK and Scholz BC (2002) The empirical assessment of stimulus poverty arguments. *Linguistic Review* **19**: 9–50.

Sapir E (1925) Sound patterns in language. *Language* vol. 1: 37–51.

Scholz BC and Pullum GK (2002) Searching for an argument to support linguistic nativism. *Linguistic Review* **19**: 185–224.

Twaddell WF (1935) *On Defining the Phoneme*. Baltimore, MD: Linguistic Society of America.

Wright C (1989) Wittgenstein's rule-following considerations and the central project of theoretical linguistics. In: George A (ed.) *Reflections on Chomsky*, pp. 233–264. Oxford: Blackwell.

Further Reading

Chomsky N (1980) *Rules and Representations*. New York, NY: Columbia University Press.

Devitt M and Sterelny K (1987) *Language and Reality: An Introduction to the Philosophy of Language*. Cambridge, MA: MIT Press.

Fodor JA (1975) *The Language of Thought*. New York, NY: Thomas Y. Cromwell.

Fodor JA and Katz JJ (eds) (1964) *The Structure of Language: Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice-Hall.

George A (ed.) (1989) *Reflections on Chomsky*. Oxford: Blackwell.

Harris Z (1951) *Methods in Structural Linguistics*. Chicago, IL: University of Chicago Press.

Jackendoff R (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York, NY: Oxford University Press.

Johnson DM and Erneling CE (eds) (1997) *The Future of the Cognitive Revolution*. New York, NY: Oxford University Press.

Katz JJ (ed.) (1985) *Philosophy of Linguistics*. Oxford: Oxford University Press.

Katz JJ and Postal PM (1991) Realism vs. conceptualism in linguistics. *Linguistics and Philosophy* **14**: 515–554.

Langendoen DT and Postal PM (1984) *The Vastness of Natural Languages*. New York, NY: Blackwell.

Lasnik H (1989) On certain substitutes for negative data. In: Matthews RJ and Demopoulos W (eds) *Learnability and Linguistic Theory*, pp. 89–105. Dordrecht: Foris.

Lyons J (1991) *Chomsky*, 3rd edn. London: Fontana.

Matthews RJ (1984) The plausibility of rationalism. *Journal of Philosophy* **81**: 492–515.

Millikan RG (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Pullum GK (1996) Learnability, hyperlearning, and the poverty of the stimulus. In: *Proceedings of the 22nd Annual Meeting*, pp. 498–513. Berkeley, CA: Berkeley Linguistic Society.

Pullum GK and Scholz BC (1997) Theoretical linguistics and the ontology of linguistic structure. In: *1997 Yearbook of the Linguistic Association of Finland*, pp. 25–47. Turku: Linguistic Association of Finland.

Quine WO (1960) *Word and Object*. Cambridge, MA: MIT Press.

Sampson G (1989) Language acquisition: growth or learning? *Philosophical Papers* **18**: 203–240.

Seuren PAM (1998) *Western Linguistics: An Historical Introduction*. Oxford: Blackwell.

Soames S (1984) Linguistics and psychology. *Linguistics and Philosophy* **81**: 155–179.

Philosophy of Mind

Introductory article

Georges Rey, University of Maryland, College Park, Maryland, USA

CONTENTS

The philosophy of mind is concerned with general questions about the nature of mental phenomena, such as thoughts, feelings, perceptions, consciousness, and sensory experience. Philosophers generally ask these questions where it is not clear how an empirical investigation might settle them.

WHAT IS THE PHILOSOPHY OF MIND?

Philosophical Versus Empirical Issues

Empirical psychologists, who confirm hypotheses by observation, are by and large concerned to discover facts that *happen to be true* of actual people and animals. For example, they might discover that a certain chemical is released in a certain region of the brain whenever someone is in pain. But the philosopher asks whether releasing that chemical is *necessary* or *essential* to being afraid or in pain: would beings lacking that particular chemical or cranial layout be incapable of fear or pain? Answering such questions requires considering not only *actual* cases, but *possible* ones, such as extraterrestrials, ghosts, angels, or human beings with a different embodiment; we cannot observe such cases, but must consider them by some kind of conceptual reflection. The questions become more urgent when the possibilities loom closer, as in the case of computers: what would it take to create a computer that had genuine thoughts, feelings, and sensations? Or can this possibility be somehow ruled out in advance, independently of what any empirical science could discover?

One reason why these questions are so difficult to answer is that there are substantial methodological disagreements about the appropriate way to answer them. In the first place, mental phenomena often seem essentially 'private' and 'subjective', not open to the kind of 'public', 'objective' inspection required of serious science. How, after all, are we to find out what a person's private thoughts and feelings really are? Isn't the person in a special, privileged position with regard to them, a position that no one else could ever occupy? What should we make of the claims of psychotherapists, cognitive psychologists or neurophysiologists about people's thought processes?

A second difficulty is that the *meaning* and *significance* of the products of mind, such as deliberate action, do not seem to be open to the usual kind of causal explanation, by which, for example, one explains why the planets move as they do. Explanations of planetary motion either invoke simple generalizations about regularities in the motion (as in the case of Kepler's Laws) or, more deeply, by appealing to general laws, such as those of universal gravitation. By contrast, to understand why people, say, go the beach in the summer, it is not enough merely to subsume their behavior under a good generalization relating heat, metabolism and beach attendance. Rather, one needs also to understand their *reasons* for going to the beach, what going to the beach *means* to them, what *role* it plays in their lives. Now perhaps this can be understood as simply subsumption under specific laws of psychology concerning how, for example, thoughts and preferences bring about action, but many people have thought that this kind of understanding can only be attained by empathizing with the person, 'putting oneself in their shoes' (a view associated with the work of Max Weber). Alternatively, psychological understanding can seem to involve seeing how the person's actions accord with certain *rational norms* (a view associated with the work of Donald Davidson and Daniel Dennett).

Some Basic Distinctions

Mental phenomena, like phenomena in most other domains, occur in a variety of basic categories, and it is often important which category is being discussed. Thus, in thinking about mental phenomena it is important to bear in mind whether one is talking about a *mental substance*. (Are thoughts and pains *composed* somehow of something fundamentally *nonphysical*, unlike the electrons and

quarks and mass energy of which all physical phenomena seem to be composed?) Or is there only one, perhaps purely physical, substance, with many nonphysical, mental *properties*?

Thoughts and propositions

Thoughts are typically described by *sentential complements*, or sentences prefixed by 'that'. Thus, there is the thought *that broccoli prevents cancer*; or the thought *that Caesar conquered Gaul*. That a thought is different from the sentence that expresses it is entailed by the fact that different sentences can express the same thought: the thought expressed by the English 'snow is white' is also expressed by the German 'der Schnee ist weiss'. Indeed, thoughts are often taken to be the meanings of sentences, and so are also often called 'propositions'. The constituents of propositions are concepts, which correspond roughly to individual words. Note that concepts may or may not correspond to real properties in the world: although there may be the *concept* of 'unicornhood' – which is a constituent of the thought 'there are no unicorns', – there may be no genuine (say, causally efficacious) *property* of unicornhood in the actual world. This is an important distinction to which we will return below. (*See* **Philosophy of Language**)

People, of course, can stand in many different relations to these propositions: they can, for example believe, hope, expect or wish that they are true. These various relations that minds can bear to propositions are called *propositional attitudes*: states that are picked out by mental verbs that take a proposition as their direct object. (The actual grammar of attitudes is quite complex, and can be abbreviated in a number of ways, as in 'Mary wants to fly', which means something like 'Mary desires that she herself fly'.)

Types versus tokens

Thoughts regarded in this way are clearly sharable: two people who don't even speak the same language can have the same thought, for example that snow is white. But these sharable thoughts are to be distinguished from the individual thoughts that each of us has at particular times, which are not sharable: the thought I had on that winter night that snow is white is different from the similar thought I had the next day.

This ambiguity is related to an important ambiguity that also arises in the case of language, where we can, for example, write 'the same word' twice. When we want to talk about words that are located in a specific place for a specific stretch of time, we can talk about *tokens* of the word; when we want to talk about 'the same word' that can appear in different places and times we can talk about word *types*. (Note that only some word tokens are written down; many are pronounced, and others are encoded on magnetic tape in computers.)

Sensations, phenomenal states and 'qualia'

Many mental phenomena do not appear (at least initially) to be propositional attitudes. There are, for example, the *conscious sensations*, such as the pains, itches, tastes and colors that we seem to experience in most of our waking moments. Our talk of sensations is rather loose, in ways that may be crucial: we sometimes talk about 'phenomenal objects', such as *the pains*, itches, tickles and mental images themselves, and sometimes about *the feeling of pain*, or itchiness, or the properties of sounds (e.g. loud, shrill) or images (e.g. red, elliptical). And in these latter cases, where we take our experiences to be reflecting real phenomena in the world, our talk is often ambiguous between the external fact ('the rose is red') and the inner experience ('my mental image is red'). It is this ambiguity that gives rise to the familiar puzzle of whether a tree falling in a lonely forest makes any sound: one might say that it makes a sound in the external sense, but not in the experiental, or phenomenal sense. Phenomenal objects and properties are often called 'qualia'.

The Main Problematic Phenomena

There are many specific mental phenomena with which the philosophy of mind is concerned: for example, free will, intention, introspection, mental causation, personal identity, qualia, reasoning, mental content, and consciousness. This article will discuss in general terms how many of these topics are related both to each other and to what have been some of the central issues in the philosophy of mind since the seventeenth century.

Three phenomena in particular have been the primary focuses of discussion in the philosophy of mind: consciousness, rationality, and intentionality.

Consciousness

Conscious phenomena are the phenomena with which people generally think they are 'directly acquainted': our sensations, pains, itches, joy, pride, and so on. For many people, the existence of these phenomena, at least in their own case, is more obvious and undeniable than anything else in the world. The seventeenth-century mathematician and philosopher René Descartes regarded his immediate conscious thoughts as the basis of all the

rest of his knowledge. Theories that emphasize this first-person immediacy of conscious states have come to be called 'Cartesian'.

However, not all mental phenomena have always been regarded as essentially conscious. Freud and Chomsky have drawn attention to ways in which unconscious attitudes might explain a variety of phenomena, from neurotic syndromes to language acquisition. Still, one might wonder what makes an unconscious mental process mental at all: if a person doesn't have an immediate knowledge of it, why isn't it just part of the physical machinery of the brain? (*See* **Freud, Sigmund**)

Rationality

One common reason for ascribing mental phenomena to other people is to make 'rational sense' of their behavior. There are four standard methods by which we do so: deductive, inductive, and abductive reason, which have to do with increasing the likelihood of believed truth; and practical reason, which has to do with the determination of action (or 'practice'), based on believed truth and the rankings of one's desires. In the twentieth century there was extensive development of formal characterizations of portions of deductive and inductive reason, most dramatically in the case of deduction and with some success in the fields of inductive logic and statistics. Abduction, or 'inference to the best explanation', is less well understood. It is what we engage in when we explain observed phenomena by unobserved phenomena: diseases by microbes, heredity by genes, chemical combination by atoms. (*See* **Deductive Reasoning**; **Inductive Reasoning, Psychology of**; **Philosophy of Science**)

Practical rationality involves proceeding from belief to action. Here, desires become relevant, in addition to truth: successful action is action that satisfies one's desires. One rough definition might be that rational action is what maximizes the probability of successful action: it is what should be done given that the agent has certain desires and otherwise rational beliefs (e.g. beliefs formed by processes likely in general to issue in truth). Suppose, for example, you want some coffee, and think that coffee can best be had from the café at the corner. Other things being equal (you have no other more pressing desires, or beliefs about serious risks of going out) the 'rational' thing for you to do would be to go to the corner café and buy some coffee. Indeed, if we were to hear that this was the 'reason' someone else left their house and did so, we would think that we had found a satisfactory explanation of their actions. Much of life is, of course, much more complex than this, and formal decision theory tries to understand decision making by means of 'cost–benefit–risk analysis', whereby people consider the probabilities of different outcomes and reconcile competing preferences.

None of this focus on rationality is meant to suggest that people are always rational. Many people report being 'weak-willed' and failing to think or do what they deem to be the best or most rational thing. And sometimes rationality does not seem relevant: kicking a stalled car, fondling a lover's locks, and twiddling one's thumbs are all actions that do not seem to be performed for any particular purpose and are neither particularly rational nor 'irrational'. Such actions have been termed 'arational'.

Intentionality

Medieval philosophers noticed a peculiar property of at least a wide class of mental phenomena: that they are *about* things. Thus, a thought that Moriarty is the thief is *about* Moriarty, and a desire for coffee is *about* coffee. By contrast, consider a star or a stone: it does not seem to make much sense to ask what it is 'about'. The medievals called this 'aboutness' property 'intentionality' (not to be confused with the very different use of 'intentional' as meaning, roughly, 'deliberate'). The property has some logical peculiarities: although a hat cannot rest on a head that does not exist, a thought cannot only be about a nonexistent head, but even about an impossible one (say, one that is round and square). Moreover, aboutness seems to be relative to a way of describing something: although a hat resting on Mark Twain is a hat resting on Samuel Clemens, a thought about Mark is not a thought about Samuel (someone might even think that Samuel is different from Mark). (*See* **Intentionality**)

The central cases of intentional states are the propositional attitudes, such as beliefs and desires. Whether all mental states can be regarded as intentional is controversial: it is not obvious what pains or itches or 'free-floating anxiety' are about. There are other mental phenomena that do not obviously fall into either of the two categories of conscious and intentional states: for example, emotions, character traits, and virtues. And many of the peculiar *products* of minds – words, paintings, gestures – are also 'about things', although this 'aboutness' is often *derived* from the minds of their users, and is hence called 'derived' intentionality, as opposed to the 'intrinsic' or 'original' intentionality of mental states. One major controversy is whether computers have genuine intrinsic intentionality, or whether it is merely derived from the intentional states of their creators. (*See* **Propositional Attitudes**)

HISTORY

Although consciousness seems to be what first comes to mind when people nowadays think of the mind, this may not always have been so. The ancient Greeks discussed it very little, concerning themselves with issues of rationality and character (such as being fair or courageous). In the Middle Ages, there was a great deal of development of religious ideas about the mind, especially in the work of Augustine, Maimonides and Aquinas, as well as the discovery of some of the peculiarities of intentionality that we have noted. And in the seventeenth century, particularly in the work of Descartes, it is the issue of rationality that is at the forefront. Descartes argued that matter was incapable of exhibiting the kind of 'universal reason' that he thought was exhibited by human beings. Although resistance to 'materialism' is still influential, it has come to be associated more with consciousness than with rationality.

Historically, the main positions with regard to the nature of mental phenomena have been: 'dualism', whereby mental phenomena are regarded as a fundamental feature of reality different from physical (or 'material') phenomena; and various versions of 'physicalism' (sometimes called 'materialism' or 'reductionism'), whereby mental phenomena are regarded as specific manifestations of the physical.

Dualism

In its most radical, Cartesian form, dualism is committed to the view that mind constitutes a fundamentally 'different substance', one whose functioning cannot be entirely explained by reference to physical phenomena. However, there are other, more modest forms. Many contemporary philosophers who reject Descartes' postulation of some kind of special mental substance are quite comfortable with a dualism regarding special mental properties (or states or events). For example, the property of being in pain may be a particular non-physical property of a nevertheless quite physical human body.

It is important to distinguish such claims about the nature of mind from claims about its causal relations. In Descartes' view, despite their immateriality, mental phenomena can be *both* causes *and* effects of physical phenomena ('dualistic interactionism'). Leibniz, however, claimed that they were neither, but were synchronized with physical phenomena ('parallelism'). An intermediate view, originally advocated by Thomas Huxley and recently resuscitated by F. Jackson, is that mental phenomena are the effects, but not the causes, of physical phenomena ('epiphenomenalism'). Recently, dualism has been energetically defended by D. Chalmers, who claims that there is a law of 'structural isomorphism' between the mental and the physical: our experience and our brains have a certain structure even though they are fundamentally different phenomena. (*See* **Epiphenomenalism**)

There are a number of interesting arguments for dualism. Most of them tend to be negative ones, arising, as we will see shortly, from specific problems with materialism. Perhaps the most striking such problem is what Joseph Levine has called the 'explanatory gap'.

Consider how we explain most ordinary, non-mental phenomena. It is one of the impressive achievements of modern science that it seems to afford, in principle, illuminating explanations of almost every non-mental phenomenon one can think of. For example, if one wants to understand why water expands when it freezes, why the sun shines, continents move, or foetuses grow, most educated people can at least imagine how possible explanations might in principle proceed. They would have to do with the physical properties of many small particles, their spatial and temporal relations, and the physical (e.g. gravitational, electrical) forces between them: put enough of these particles together with just the right forces between them, and it would follow that the water would have to expand, the sun shine, and foetuses grow. As Levine nicely puts it, the microphysical phenomena 'upwardly necessitate' the ordinary macrophysical phenomena.

But it is precisely this upward necessitation that seems very difficult to supply in the case of the mental, particularly in the case of the two phenomena we have discussed, consciousness and intentionality. Consider the question: how do you know that someone else sees colours in the same way you do? To put it in terms of physicalism: what physical facts about a person uniquely determine that it must be 'red' experiences that the person is having when he or she looks at ordinary blood, and not 'green' ones? Or, to take an analogous problem raised by Quine about intentionality: what facts about a person's physical make-up uniquely determine that the person is thinking about rabbits, as opposed to 'rabbithood' or 'undetached rabbit parts'?

Physicalism

For all these conceptual difficulties in relating the mental to the physical, many philosophers are

convinced that the world is nevertheless fundamentally a physical world, and that the mental must be a specific manifestation of it. Interest in this issue is not confined to those with an interest in the centrality of physics. On pain of circularity, if mental phenomena are ultimately to be explained in *any* way, they must be explained in terms of non-mental phenomena. It is a significant fact that the only non-mental phenomena we know of are physical (even the biological has turned out to be physical). Physicalism (also called 'materialism') is the view that all mental phenomena are in some sense just physical phenomena.

The identity theory

Although materialist ideas can be found in the Greek atomists and in Lucretius, Thomas Hobbes is often thought of as the father of modern materialism. However, Hobbes does not clearly distinguish between mental phenomena being the *effects* of physical processes, and their actually *being* physical processes themselves. This latter view, the so-called 'identity theory', whereby the mental is 'nothing but' the physical, emerged only in the twentieth century, in the work of U. T. Place and, most influentially, J. J. C. Smart. Here the proposal is to identify every mental state with a physical state, in the way that, for example, episodes of lightning could be identified with episodes of electrical discharge. The primary argument for this view is a kind of economy about the different kinds of things in the world, and a unification of causal claims: mental events enter into causal relations with physical ones because they are in the end physical events themselves. This view is also called 'reductionism', which unfortunately carries the misleading suggestion that the mental is somehow 'made less' by being physical. This is a mistake: lightning, after all, is not 'less' lightning for turning out to be 'reduced to', or being 'nothing but', electrical discharge.

The analogy with lightning, however, carries what many have thought to be an implausible implication: every instance of lightning is indeed an instance of electrical discharge; but is every instance of, say, *believing that grass grows*, an instance of the very same type of physical state, say, an excitation of specific neurons in the brain? Why couldn't you have silicon implants where I have neurons, without disturbing our shared belief that grass grows?

A way of avoided this unwanted implication was proposed by Nagel, who noticed an ambiguity in identity statements, between identifying type and token mental states. He and others proposed limiting the identity theory to the latter: just as different 'token' occurrences of the same 'type' alphabetic letter need not share any one physical property, so different token mental states might be 'realized' in different physical substances. In this view, all that is important for physicalism is that there are no nonphysical tokens.

Even if we restrict ourselves to token identity claims, however, there are still some problems with the identity theory. One simple problem concerns the relation of many mental phenomena to physical space. It is ordinarily unclear exactly where such things as beliefs and desires are located. We ordinarily say they are 'in your head': but where exactly? Does it make sense to say that my thought that grass grows is 2 mm to the right of my hypothalamus? Or, to take a harder example, many people claim to have vivid 'mental images', and, indeed, experiments suggest that they play a significant role in ordinary reasoning. These images seem to have certain spatial properties: they may be 'oval' and 'vividly colored', and even 'rotate' as we try to solve some problem. But if they too are to be identified with physical things, then it would follow that there should be oval, vividly colored, rotating images in the brains of people who experience such images. This seems absurd: neurosurgeons do not discover such images in brains – and, even if they did, how would such images ordinarily be seen? By an 'inner eye'? So it would seem that a mental image cannot be a physical thing. (*See* **Mental Rotation**)

These difficulties might well lead one back to dualism. However, there is another, even more radical response.

Eliminativism and radical behaviorism

Exasperation with dualism, but also with the difficulties of reductionism, led many psychologists of the twentieth century to the extreme position of eliminativism, or the denial of the existence of any mental phenomena at all. This may seem a preposterous position, but it is worth taking seriously both for the light it sheds on certain issues, and because certain versions of it may be true for specific classes of mental phenomena.

Moreover, there is an interesting argument that can be made on its behalf. This has to do with the evident lack of any causal break in the internal processes of the human brain. So far as we know, there is no movement of anything in anyone's body that fails to have a physical explanation (if it has any explanation at all). If this is true, as many

psychologists believe, then we should ask what reason we have to believe there are any mental phenomena at all: they seem superfluous.

Of course, if one reduced mental phenomena to physical phenomena, then one could perhaps claim that the action of the mind is 'nothing but' the action of the body. Like dualists, eliminativists are impressed by the difficulty of pulling off the reduction. Unlike dualists, however, they are impressed by this lack of a causal break. The point could be put this way: if the difficulties of reductionism force us to regard mental talk dualistically, then the lack of a causal break should further force us to deny the existence of any mental phenomena. They are no more needed to explain the motions of our bodies than are angels to explain the motions of the planets. (*See* **Eliminativism**)

One might retort by asking how any physical theory explains one's own present conscious thoughts and experiences. Why don't the same explanatory gaps that argue against reductionism argue equally against eliminativism? But this begs the question against the eliminativist, who will simply reply: 'What conscious thoughts and experiences? You may as well say that Newton didn't explain the motion of the planets because he didn't explain the beating of the wings of the angels who were pushing them.' What is needed in either case is non-tendentious evidence for the questioned postulation, be it mental or angelic.

At first sight, this seems a difficult challenge. We do not ordinarily think about providing non-tendentious evidence for the mind, and so most of the evidence we might be tempted to provide presupposes mental talk, as when we talk not only about our own thoughts, but even about other people's deliberate actions ('raising an arm') rather than just the motions of their limbs ('an arm rising').

However, there is a source of evidence that may provide at least a good deal of what we need. Consider 'standardized tests' (such as the 'SAT' and 'GRE' regularly given to students in the United States). Here the 'standardization' consists in the fact that both the question sheets and the answer sheets are prepared in such a way as to be in principle physically type-identical: the question sheets consist of identically printed marks on paper, and the rectangles on the answer sheets are supposed to be filled in with a graphite pencil in such a fashion that a machine can mark the test. The correlations between these physical types of 'input' and 'output' is surely no accident. Science is in general concerned with explaining such regularities as these, and, although physics could explain every such *token* event, it is hard to think of any purely physical explanation that stands a chance of explaining such *correlations* between the types. The only explanation that seems plausible is a mentalistic one that talks about the students' thoughts, desires and reasoning abilities.

In the twentieth century, some eliminativists thought they did have an alternative explanation of such intelligent behavior. For about 50 years, they pursued a program, which B. Skinner called 'radical behaviorism', of trying to explain all human and much animal behavior in purely physical terms, specifically in terms of patterns of 'conditioning' among physically specified stimuli, responses and reinforcements. The programme is now of little more than historical interest, partly because its main tenets were refuted by the careful experiments of the behaviorists themselves. They found that rats displayed a variety of navigational skills that defied explanation in terms of conditioning, and that seemed explicable only by reference to 'mental maps' and 'curiosity drives'. K. Lashley provided some theoretical reasons for thinking that many serial behaviors could not be explained by conditioning. Finally, Chomsky provided a devastating refutation of Skinner's efforts to provide a behavioristic account of human language.

Irreferentialism and analytical behaviorism

A more moderate response to the difficulties of both dualism and reductionism was proposed by Wittgenstein. He claimed that philosophers too often presume that words gain their meaning by referring to various phenomena in the world, and that this leads them to think of our inner mental lives using a model too close to the familiar outer one of material objects. He proposed instead that we think of the meaning of a word as its 'use', or role in various 'language games' of which our ordinary talk consists. If we do this, there is no reason to suppose, for example, that our talk of mental images need be taken to refer to objects in a mysterious mental realm; rather, terms like 'mental image' (and, Wittgenstein seems to suggest, most of our mental talk) should be seen more on the model of an expression like 'the average American family', which, of course, does not refer to any actual family (its use is a convenient way of speaking of a ratio).

However, it is not easy to provide a positive account of the meaning of mental talk. Popularizing the then-unpublished views of Wittgenstein, Gilbert Ryle attempted to exorcize what he called the 'ghost in the machine' by showing that mental terms function in language as abbreviations of

dispositions to overt bodily behavior. This programme came to be called 'analytical behaviorism' (a thesis about the *meanings* of mental words, not to be confused with radical behaviorism, which is a thesis about how to do *without* mental words). A famous example of a behaviorist analysis was Alan Turing's proposal that a machine be counted as 'intelligent' if its teletyped answers to questions were indistinguishable from those of a normal human being. (*See* **Turing Test**)

Analytical behaviorism did not meet with great success. It is not hard to think of possible cases of creatures that might act exactly as though they were, say, in pain, but actually were not: consider expert actors, or human bodies wired to be remotely controlled. Nor is it hard to imagine machines passing Turing tests by merely emulating a conversation without having any understanding of it at all.

What all such examples show is that mental states are not tied directly to behavior. Typically, they issue in behavior only in combination with one another: a belief issues in behavior only in conjunction with, for example, desires and attention, and conversely. For example, it is because an actor has different motivations from a normal person that he can behave as though he were in pain without actually being so; and it is because a person believes that she should be stoical that she might be in excruciating pain but not act as though she is. The internal organization responsible for a person's behavior, and not just the behavior itself, determines the identity of a mental state. The Turing test, which ignores such internal organization, is thus a poor test for intelligence. (Wittgenstein had a more subtle view than Ryle: he never claimed that mental expressions could be *analyzed* in behavioral terms, but that 'an inner process stands in need of an outward criterion', a view that is still held by some today.)

Functionalism

The fact that mental terms seem to be applied in ensembles has led some philosophers to think about technical ways to define an entire set of terms together. David Lewis invoked a technique, called 'ramsification', whereby a set of 'new' terms could be defined in terms of their relations to one another and to other 'old' terms already understood. This captured an idea already noted by Hilary Putnam with regard to the set of standard states of computers. Turing's most important work was in the conceptualization of the modern computer, in terms of 'states', each of which is defined in terms of the others and what the machine would do when it receives an input: the machine produces a certain output and passes into another of the states. The states can then be defined together in terms of the overall pattern produced inthis way. (Note that Turing's important thesis that such a machine can compute anything computable is quite independent of his 'test' for intelligence.)

States of computers are not the only things that can be defined in this way: most reasonably complex entities that have parts that function in specific ways will do as well. For example, a carburator in an internal combustion engine can be defined in terms of how it regulates the flow of petrol and oxygen into the cylinders where the mixture is ignited, causing the piston to move. It was such analogies between mental states and the functional parts of complex machines that provided the inspiration for functionalist approaches to mental terms. Such approaches have dominated the philosophy of mind since the 1960s.

There are many different ways in which the functionalist approach can be deployed, depending on the kind of view of the mind that one thinks is constitutive of the meaning of mental terms. Some people take the view reflected in common folk beliefs; some take views informed by philosophical reflection on possible cases; others take views informed by empirical psychological theory. Furthermore, definitions may vary according to whether they are derived from a view of the *whole system at once*, or only of *specific parts* of it, and whether the old terms are confined to observable behavior, or may include references to specific features of our bodies and the environments we inhabit. The most influential form of functionalism is based on the analogy with computers, which, of course, have been independently developed to solve problems that seem to require a kind of intelligence.

The computational–representational theory of thought

The idea that thinking, and mental processes in general, could be treated in computational terms, was inspired by the successes in the formalization of certain portions of reasoning that we mentioned above. It emerges in the work of Newell and Simon, Putnam, Harman, and especially J. Fodor, who has been most explicit in developing the computational–representational theory of thought (CRTT), or the idea that thinking consists in computing upon sentences in a 'language of thought'.

Note that CRTT is not the claim that any computer – even any existing computer – is or has a mind. Rather, it is the claim that a mind is a certain *kind* of

computer, one with specific internal structures and specific relations to the environment, which, together, are responsible for its having certain intentional content. (Hence, *pace* Searle's famous 'Chinese room argument', the Turing test is quite irrelevant to CRTT.) It is, in fact, not so much a claim as a research program: the hope is that by understanding the brain as an elaborate computer – or, more realistically, as a complex assemblage of computers – one could ultimately define mental states in terms of the specific computational roles they play in that assemblage. This research program is more or less the subject matter of cognitive science. (*See* **Artificial Intelligence, Philosophy of**; **Bayesian Belief Network**; **Bayesian and Computational Learning Theory**; **Means–Ends Analysis**; **Reasoning**; **Turing, Alan**; **Unified Theories of Cognition**; **Chinese Room Argument, The**; **Connectionism**; **Artificial Intelligence, Gödelian Arguments against**; **Cognitive Development, Computational Models of**; **Analogy-making, Computational Models of**; **Mental Disorders, Computational Models of**; **Games: Trust and Investment**; **Computational Models: Why Build Them?**)

THEORIES OF CONSCIOUSNESS

Development of computational theories of mind, as well as of more detailed neurophysiological knowledge, has stimulated a renewal of interest in consciousness, which had long been avoided, in philosophical discussions, as a quintessentially subjective phenomenon. However, although much has been written, we are still nowhere near a satisfactory theory.

There has recently been a great deal of investigation of neural correlates of consciousness, e.g., of a 40 Hz oscillation in the primary visual cortex of a cat whenever the cat is having a visual experience. But however robust this finding may turn out to be, a correlation is not an explanation. As we noted at the outset, it is a distinctive concern of philosophy to determine the nature of a phenomenon, and a correlation alone does not provide that. (Would having such an oscillation render a radio conscious, or lacking it render an animal unconscious?)

Executive, Buffer, and Higher-order State Theories

Sometimes 'conscious' is used merely to mean 'awake', as in: 'Is the patient conscious?' But the meaning of interest to philosophers tends to be more specific, involving the availability of certain contents as objects of introspective knowledge. There are three (not necessarily exclusive) sorts of theories of that availability: what might be called 'executive' theories, 'buffer' theories, and higher-order state theories. Executive theories stress the role of conscious states in deliberation and planning. According to buffer theories, consciousness consists in a person standing in specific relations to a specific location in which material is stored for specific introspective purposes, along the lines of the material available to a press secretary, whose claims may or may not match those of an executive. Ray Jackendoff makes the further interesting suggestion that such material is confined to relatively low-level sensory material.

An important family of much more specific proposals are variants of the idea that consciousness involves some kind of state directed at another state. One variant is that it involves some kind of 'internal scanning' or 'perception'; another is that it involves an explicit 'higher-order thought' (HOT), or a thought that one is in a specific mental state. For example, the thought that one wants a beer would be conscious only if one thinks that one wants a beer. This does not mean that the HOT is itself conscious; only that its presence is what renders its target, lower-order thought conscious. D. Rosenthal defends the view that the HOT must actually be occurring at the time of consciousness; Carruthers defends a more modest view that the agent must simply be *disposed to have* the relevant HOT.

'What It's Like'

Ned Block has pointed out an important distinction between two concepts of consciousness that many of the above discussions do not distinguish: 'access' consciousness and 'phenomenal' consciousness. Although it might be defined in a variety of ways, depending on the details of the computational (or other) theory of thought under consideration, access consciousness (of some material) consists in being accessible to various mental processes, particularly, say, introspection. Block points out that the fact that material is accessible to processes does not entail that it actually has a qualitative, phenomenal feel, that there is something 'it's like' to have that material being so accessible.

This latter issue has been made particularly vivid by two influential articles about the very special knowledge we seem to acquire as a result of conscious experience. Nagel pointed out that no matter how much someone might know about the 'objective' facts about the brains and behavior of bats and their peculiar ability to locate objects by means of a kind of built-in 'sonar', that knowledge would not

suffice to convey the 'subjective' facts about 'what it's like to be a bat': indeed, it is unlikely that human beings will ever be able to know what the world seems like to the bat. Frank Jackson made a similar point by imagining a brilliant color scientist, 'Mary', who happens to know all the physical facts about color vision, but has never had an experience of red (either because she was color-blind or because she happened to live in an unusual environment). Suppose that one day (either through surgery, or by leaving her strange environment) she finally does have a 'red' experience. Does not she learn something 'new', something that she didn't know before even though she knew all the physical facts of color vision?

Qualitative States

'Qualiaphilia', as it has come to be called, is the view that no functionalist theory of consciousness can capture phenomenal consciousness: in conscious experience we are aware of 'qualia' that arenot *relational*, in the way suggested by the functionalist, but rather somehow *intrinsic* features ofour experience. It is this sort of consideration that still leads many philosophers to varieties of dualism.

Representationalism

An intruiging alternative to qualiaphilia is a suggestion that has gained a good deal of attention in the last decade whereby qualia are not understood as *real* phenomena in experience, but rather merely *intentional* phenomena: that is, although they are the sorts of 'objects' on which we can direct our intentional states (such as thought and desire), they are no more real than are 'intentional objects' such as Santa Claus or Zeus. We already noted a form of this view when we observed that there is no need to believe in objects such as pains. One of the most interesting suggestions philosophers have made about the mind concerns the apparent properties of our qualitative experience: the look of red, the taste of beer. As noted above, there is no need to believe in objects such as pains, tickles or even mental images: one can just as well speak about experiences of these things, affirming that there are such states, but denying that these further 'objects' actually exist (cf. irreferentialism).

Although this is a widely accepted view in the case of these sorts of phenomenal objects, many people find it harder to accept in the case of phenomenal properties: how can one deny the existence of such properties as say, the intense painfulness of unanaesthetized surgery?

But what makes us think there is an actual property there in experience? As we noted earlier, the concept of 'unicornhood' need not entail the existence of any corresponding property. Perhaps certain stimulations of the nervous system can cause people to possess certain concepts (or, more specifically, certain modes of presentation) of certain phenomenal properties, concepts that could not be possessed by knowledge of physics alone, but there need not be any corresponding properties whose existence is at all problematic. (Think of one's mental life as being like an animated cartoon: does it matter that there are no real things represented by the flickering images on the screen?)

Remaining gaps, and first-person scepticism

What nevertheless continue to bother the qualiaphile are the problems of the 'explanatory gaps' between the physical and various mental phenomena, described above. For all their (potentially elaborate) accounts of the organization of a typical human being, functionalist theories have not yet been shown to 'upwardly necessitate' the specific qualia of experience (red versus green), the specific intentional content of thought (say, thinking of rabbits rather than 'rabbithood'), or even conscious experience: it seems possible to imagine a computer with precisely the same functional organization as a human brain, which is nevertheless a zombie and has no conscious experiences at all. A problem with taking this possibility seriously, however, is that it can leave one wondering about oneself: if there is a possibility that other people may have all the computational organization of a mind but not be conscious, then why shouldn't that be a possibility in one's own case as well? How do you know that you are a conscious agent, and not merely a 'robot' that thinks it is?

Further Reading

Chalmers D (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press. [An influential defence of dualism.]

Jackson F and Braddon-Mitchell D (1997) *Philosophy of Mind and Cognition*. Oxford, UK: Blackwell.

Kim J (1998) *Mind in a Physical World*. Cambridge, MA: MIT Press. [An excellent introduction.]

Levine J (2000) *Purple Haze*. Oxford: Oxford University Press. [A comprehensive discussion of the problems of the 'explanatory gap'.]

Rey G (1997) *Contemporary Philosophy of Mind*. Oxford: Blackwell. [A general, introductory discussion.]

Rosenthal D (2000) *Materialism and the Mind-Body Problem*, 2nd edn. Indianapolis, IN: Hackett. [An excellent collection of historically important discussions, from Hobbes to Fodor.]

Philosophy of Neuroscience

Intermediate article

Ian Gold, Monash University, Clayton, Victoria, Australia

CONTENTS

The philosophy of neuroscience is concerned with the philosophical investigation of neuroscience as a science and is distinguished from neurophilosophy, which is concerned with the application of neuroscience to the philosophy of mind and epistemology.

INTERACTIONS BETWEEN PHILOSOPHY AND NEUROSCIENCE

Philosophy interacts with neuroscience in two different ways, although the boundary between the two is fluid. The first kind of investigation, usually referred to as philosophy of neuroscience, is concerned with philosophical questions raised by neuroscientific data or theory, or by the status of neuroscience as a whole. The second kind of investigation, usually called neurophilosophy, attempts to exploit advances in neuroscience to solve philosophical problems about the mind. The philosophy of neuroscience is thus a branch of the philosophy of science. Neurophilosophy, in contrast, is a branch of the philosophy of mind and is closely related to cognitive science.

Because both of these branches of philosophy are of recent vintage, general principles and clear foci have not yet been identified; nor are there many clearly delineated 'debates' that characterize more established areas of philosophy. For these reasons, no attempt is made here to present a comprehensive picture of the field. Instead, this article discusses some representative illustrations of work in neurophilosophy and philosophy of neuroscience. Many philosophers of neuroscience – like some neuroscientists themselves – are also pursuing research that is better described as theoretical neuroscience. This is not dealt with here.

HISTORY

The modern incarnation of the view that the mind is identical to the brain and its functions was introduced into psychology and philosophy in the late 1950s. This doctrine, known as the identity theory, was eventually supplanted by functionalism, according to which mental states are to be analyzed in terms of their relations to one another, to perceptual input, and behavioral output. Functionalism holds that any material that embodies these relations – the silicon chip of a computer, for example – has a mind. Thus, while the view that the mind is to be explained by science has been the overwhelmingly dominant position in contemporary philosophy, the brain and its properties had been, until recently, as little emphasized in philosophy as in cognitive science and artificial intelligence research.

Against this mainstream position Patricia S. Churchland (1986) articulated a new vision for the philosophy of mind which she called neurophilosophy. With Paul M. Churchland, she has continued to argue that an understanding of the mind requires abandoning the purely functional investigation of philosophy and cognitive science in favor of a biologically based science of mind. Indeed, the Churchlands have defended the view that a successful theory of the mind will be exclusively neuroscientific (see below).

Although philosophical work deserving the name 'neurophilosophy' had existed prior to 1986, Churchland's book baptized new branches of the philosophy of mind and science, and this has produced a new generation of philosophers whose work straddles the boundary between philosophy

and neuroscience. Even if functionalism, broadly speaking, remains the right framework for a theory of the mind, the specific details of biological minds is of independent interest. Further, the facts of how natural minds operate may provide novel ideas for the development of artificial intelligence systems.

NEUROPHILOSOPHY: PHILOSOPHY OF MIND AND EPISTEMOLOGY

Neurophilosophers have addressed a wide variety of traditional problems in philosophy by applying the findings of neuroscience. In what follows, five representative illustrations are presented.

Perception and Intentionality

Perhaps the most significant and striking fact about mental states is that they are about things in the world. This 'aboutness' is usually referred to as the intentionality of mental states, and it is closely related to what is often called perceptual content – that is, how perception represents the world to be. How this is possible, and how it occurs, are two of the most important questions in the philosophy of mind. A paradigm case of the intentionality of the mental is the putative intentionality of perceptual states. When I stare at the ocean from my window, I am in a visual state which is about the ocean. Because this case seems simpler than many others – for example, my ability to imagine what it would be like to visit the Taj Mahal – theories of intentionality often begin with perception as a first step in the development of a comprehensive theory of intentionality.

Kathleen Akins (1996) argues that the relation between sensory states and what they represent may be quite different from the received view about intentionality. According to that view, sensory states represent aspects of the external world which are, typically, informative for the organism. Akins doubts that this is generally true; sometimes sensory states are not about the external world in the familiar sense. An investigation of sensory states, therefore, may not illuminate the 'aboutness' of mental states in the traditional sense.

Akins's argument takes as an illustration the simple sensory system of temperature detection or thermoreception. According to the received view, the purpose of thermoreception is to represent the temperature of the environment to the organism, and this view assumes that temperature receptors function like thermometers. However, the neurophysiology of thermoreception reveals a rather different picture. Temperature receptors do not function like thermometers; they are nonlinear and produce exaggerated responses to temperature changes. The reason for this, according to Akins, is that thermoreception is, as she says, 'narcissistic'. It functions not to give the organism information about the temperature of the environment, but rather to tell the organism how the temperature of the environment affects the organism. Consider, for example, the familiar illusion that comes from running one hand under cold water, one under hot, and then putting both into the same tepid water. One hand will report that the water is hot, the other that the water is cold. On the view that perception is designed to reveal the way the world is, this is a spectacular failure. On the narcissistic view, however, it is a success because each hand is reporting something of relevance, namely, that its environment is rapidly heating or cooling. Because rapid changes in temperature are of great practical importance to organisms, this is important information.

A neurophysiologically inspired conception of sensory function, therefore, produces a rather different picture of perception from the traditional one. If this picture is correct, whatever the lessons of perceptual systems for the nature of intentionality, there is no simple and direct route from the properties of sensory systems to the nature of intentionality.

Pain

The nature of conscious experience and its relation to the external world is a perennial concern of the philosophy of mind and epistemology, and pain experience is a common test case for this investigation. There is, however, considerable controversy over what pain reveals about the mind, and the phenomenon of pain has been used as a way of defending myriad views; for example: that some mental states are entirely subjective; that we have a special access to the states of our own minds; that experience has properties that cannot be explained in purely physical or functional terms; that the mind is something over and above the brain; and that some aspects of our experience are entirely mysterious.

Pain experience is often used as a test case because it is thought to be a 'simple' kind of experience. As against this view, however, Valerie Gray Hardcastle (1997a) argues that philosophical confusions about pain arise as a result of a failure to understand the neural complexity of pain, in particular, the fact that pain perception is an overarching prozcess made up of a number of subsystems.

These include systems that subserve the conscious experience of pain, the sense of suffering typically associated with pain, the affective-motivational aspect of pain, and pain behavior.

Hardcastle argues that philosophers have tended to identify pain with only one of these subsystems and have therefore drawn erroneous conclusions about pain as a whole. She further argues that a correct understanding of pain requires positing two broad pain mechanisms, the first responsible for the experience of pain and the second responsible for the inhibition of pain experience. Each of these putative systems explains many of the features philosophers have thought ascribable to pain as a whole, and together they resolve many philosophical disputes about pain. Disagreement over whether pain is a 'subjective' or an 'objective' phenomenon, for example, results, on Hardcastle's view, from taking a part of the phenomenon for the whole. The existence of a pain sensory system explains many of the features of pain that have led philosophers to argue that pain is an objective phenomenon, whereas the existence of a pain-inhibiting system explains many of the features of pain that support a subjective account of it. Once pain is understood to be composed of both of these functionally distinct subsystems, the disputes over objectivity disappear.

Dreams

The phenomenon of dreams has been a frequent case study in the history of philosophy's attempt to understand consciousness. The causes and function of dreams also, famously, take centre stage in psychoanalysis. Recent work in neurophysiology and neurochemistry has begun to elucidate the biological processes of sleep, including rapid eye movement (REM) sleep, the stage of sleep in which the sleeper dreams. The function of dreams, and what neuroscientific findings say about psychoanalytic theory, remain controversial.

What has recently occupied philosophers most about dreams is the question of whether they are genuine experiences. Dreamers awaken and believe that they have had experiences in dreams of events and adventures of various kinds. However, dream experience is very uncritical. For this reason, the mere memory of a dream is poor evidence for establishing that the dream experience has actually occurred. Other hypotheses – such as that dreams are confabulated upon awakening – are equally plausible and have been fodder for philosophers of mind of a skeptical bent who have denied that dreams are experiences.

Owen Flanagan (1996) argues that neurophysiology can do philosophical work here to refute this skeptical position. The relevant evidence comes from studies showing that the patterns of neural activity during REM sleep are highly similar to the patterns of activity in the waking state. This similarity provides evidence for the accuracy of dreamers' memories of their dreams: dreams are indeed experiences similar in some respects to waking experiences because the state of the brain in REM sleep is much like the state of the brain in wakefulness.

Flanagan (1995) also considers the question of what function dreams perform in mental life. He evaluates the recent neuroscientific theories of dreaming and argues that, strictly speaking, dreams perform *no* function in the cognitive economy of the dreamer. They are, rather, evolutionary 'spandrels' – phenomena that come into existence not because they themselves are adaptive but because they are necessary for, or are a by-product of, some other function that is selected for. Flanagan claims that dreams arise as a necessary, but non-functional, consequence of REM sleep though what the function of REM sleep itself is remains controversial. He suggests, however, that the practice of dream interpretation, ubiquitous both before and after Freud, can be thought of as a technique for self-reflection and understanding. Dreams can, in this sense, be seen to have a cultural, if not a psychological, function.

Rationality

Dreams are sometimes said to be states in which dreamers are irrational. The concepts of rationality and irrationality have also been explored in neurophilosophy by the investigation of delusion. A paradigm case is the Capgras delusion in which sufferers become convinced that intimates (or sometimes significant objects) have been replaced by exactly identical duplicates. As if this is not strange enough, while sufferers are puzzled by this occurrence, they do not seem to be terribly troubled by it.

Other equally bizarre beliefs include the belief that one is dead (Cotard delusion); that one is being followed by familiar people in disguise (Frégoli delusion); that someone else is inserting thoughts into one's mind (delusion of thought insertion); that someone else is controlling one's actions (delusion of control); and that the person one sees in the mirror is someone other than oneself (mirrored-self misidentification). Traditionally, delusional subjects were referred to psychiatrists, and

this practice continues. There is now evidence, however, that some delusions may be brought about by damage to the right hemisphere of the brain. Neuropsychiatrists have, therefore, begun to try to explain delusion in neuropsychological terms.

Two significant questions for neurophilosophy are: in what sense is delusion an instance of irrationality? And what is the correct account of the character of this irrationality? A number of possibilities have been suggested. One is that delusions are not irrational because they are attempts to explain very strange experiences. On this view, the delusional belief is a rational response to that experience. A second view holds that delusions are irrational to the extent that they represent biases in hypothesis-generation or in the evaluation of the plausibility of hypotheses adopted as beliefs. A third view holds that the irrationality of delusion lies in the way experience represents the world. According to this position, certain representations, though not full-blooded beliefs, none the less deserve to be called irrational.

A third question for neurophilosophy is: what, if anything, does the functional anatomy of delusion reveal about the functional organization of belief? Here research is just beginning. Martin Davies and Max Coltheart (2000; Davies *et al.*, forthcoming) are defenders of the view that the source of delusion is to be found in a combination of a strange experience together with a second cognitive factor. They argue that while the strange experiences of delusional subjects could be explained in any number of reasonable ways – including that the subject has a mental or neurological disease! – delusional subjects opt for explanations that are impossibly farfetched. Delusional subjects must, therefore, have some other cognitive dysfunction that does not prevent this extreme hypothesis from establishing itself as part of the subject's store of beliefs. If this view is correct, and if right hemispheric damage is confirmed as the site of damage in delusion, we would have the beginnings of a functional anatomy of belief.

Unity of Consciousness

An early illustration of the relevance of neuroscience to philosophy was provided by Thomas Nagel (1971), who discussed experiments with commissurotomy, or 'split-brain', patients in order to explore the traditional view that consciousness is unified – that our experience is constituted by a single stream of experiences of which we are aware. The neuropsychological phenomenon is well known. Individuals with severe epilepsy can be treated by severing the corpus callosum and anterior commissure (the bundle of fibres that connect the two hemispheres of the brain) in order to prevent the spread of epileptic fits. Despite the fact that subjects who undergo this surgery appear entirely normal in their everyday behaviour, studies of these individuals reveal subtle but dramatic behavioral effects. The absence of anatomical connections between the hemispheres prevents perceptual input from one hemisphere from being transferred to the other, and the behavior elicited from the patients makes it evident that the two hemispheres can function and drive behavior separately and, sometimes, competitively. However, because only one of the hemispheres (usually the left) subserves language, only that hemisphere can express the thoughts of the patient.

Nagel considers the question of what these phenomena say about the apparent unity of consciousness. How many minds exist within the body of the split-brain patient? Nagel considers five possibilities: (1) that there is one mind whose consciousness is located in the linguistic left hemisphere and whose right-hemisphere behavior is caused by a sort of automaton; (2) that there is one mind associated with the left hemisphere, but right-hemisphere behavior is caused by conscious mental activity that is unintegrated into the larger consciousness of the person; (3) that there are two minds only one of which can speak; (4) that there is only one mind, the contents of which come from both hemispheres and are therefore dissociated; and, finally, (5) that there is one mind when the two hemispheres are acting together but two minds when experimental conditions elicit different behavior from each separately.

Nagel argues that there are no principled reasons for choosing any of these options over the others and concludes, therefore, that there is no whole number of minds that can be associated with split-brain patients. The even more radical proposal that follows from this is that the same holds true of normal subjects. The belief that the mind is a single thing is an illusion, Nagel suggests, that arises because every agent experiences his or her own consciousness as unified and imports the assumption of unity into his or her experience of the minds of others. It is rather the case, Nagel argues, that a mind is a conglomeration of diverse functions that operate so harmoniously as to appear, from both the inside and the outside, as a single thing. The falsity of the appearance is revealed when the two hemispheres can no longer operate in concert. (*See* **Split Brains, Philosophical Issues about**)

PHILOSOPHY OF NEUROSCIENCE AND ITS RELEVANCE TO COGNITIVE SCIENCE

Neuroscience and the Cognitive Sciences: Relations among Theories

In contrast to the interests of neurophilosophy, the philosophy of neuroscience investigates philosophical issues within the theories and practice of neuroscience itself. A number of issues have been addressed by philosophers of neuroscience including levels of explanation in cognitive neuroscience; the nature of representation; and the significance of neuroscientific evidence. Perhaps the central question that has been addressed thus far in the philosophy of neuroscience, however, is a very general one: what is the relation between the cognitive sciences and neuroscience? A number of relations are in principle possible, among them that neuroscience and cognitive science will provide separate but equal descriptions of the mind at different theoretical levels; that cognitive science will provide a functional description of the mind for which neurobiology will provide a mechanistic description; and that cognitive science will reduce to neuroscience as genetics reduces to molecular biology. In the process, it is likely that some parts of cognitive theory will be discarded and replaced by neuroscience.

The views of the Churchlands (see especially P. M. Churchland, 1981; P. S. Churchland, 1986; Churchland and Churchland, 1996) have focused debate in this area. They defend a view known as eliminative materialism. According to this view, the cognitive sciences will eventually be entirely replaced by a mature neural theory of the mind. The view is eliminativist because it holds that ordinary psychological concepts, and perhaps the concepts of theoretical cognitive science, will be eliminated and replaced by novel concepts from neuroscience.

Although according to eliminativism neuroscience will provide the final theory of the mind, the historical process of reaching this state of knowledge will, according to the Churchlands, involve a co-evolution of neuroscience and cognitive science. Results from neuroscience will, on this view, constrain cognitive models of mental phenomena which will, in turn, stimulate and direct investigation in neuroscience. Once neuroscience is sufficiently rich in theoretical resources, however, it will be in a position to explain mental phenomena directly. If this view is correct, then the consequences of eliminativism for cognitive science are dramatic. While cognitive science will play an important historical role in the development of the theory of the mind, it is neuroscience alone that will remain once the process of co-evolution is complete.

Eliminative materialism can be supported by a number of lines of argument, but two are particularly important. The first is based on the apparent fact that 'folk psychology' – the quotidian explanations of behavior by appeal to belief and desire – and, to some extent, cognitive science, have, according to the Churchlands, largely failed to provide adequate explanations of mental phenomena. One reason for this failure is that the fundamental psychological concepts are so flawed that no science based on them can get very far. Eliminativism concludes that they must be abandoned in favor of a set of concepts that is more promising.

A second line of argument proposes that psychological notions cannot be unified with the scientific world-view being developed by the other natural sciences. Only a biological science, such as neuroscience, is likely to produce a theory of the mind that will fit naturally into this picture. This argument has two possible conclusions: first, that neuroscience *ought* to be pursued and, second, that the unifiability of neuroscience with the natural sciences provides prima facie evidence that a neural theory of the mind will in fact be successful. The latter conclusion expresses the eliminativist view.

Beyond Eliminativism

Although eliminativism has been repeatedly challenged by philosophers of mind, it continues to represent a mainstream view among neuroscientists themselves. A different view is developed by Ian Gold and Daniel Stoljar (1999). They argue that the view that neuroscience alone will provide the successful theory of the mind embodies two quite distinct pictures of the development of the science of the mind that often fail to be distinguished. According to one view, 'neuroscience' can refer narrowly to cellular and molecular biology of the brain, or it can refer, more broadly, to the interdisciplinary effort to understand the mind and brain usually known as 'cognitive neuroscience'. The eliminativist view that the biological science of the brain alone will provide the successful theory of the mind is a radical view, but it is not as yet supported by anything in neuroscience itself. All of the putative accounts of mental phenomena available in neuroscience – even those that are framed largely in neural terms – turn out to be amalgams of

neurobiology and cognitive science. The present state of the sciences of the mind supports the view that the science of the mind is likely to be an interdisciplinary one.

In contrast, the view that it is cognitive neuroscience alone that will provide the successful theory of the mind is not a radical thesis at all. It is simply the view that some combination of sciences will collectively explain the mind. In particular, this view does not entail, and the radical view does, that cognitive science will form no part of a final theory of the mind. These two quite distinct views are regularly conflated and lead to a distorted picture of the current state of neuroscience. Which sciences will provide the necessary theoretical concepts, and what their relative contributions to successful theory will be, remain open questions.

Interpreting Neuroscience: An Illustration

In addition to addressing the status of neuroscience as a whole, the philosophy of neuroscience is concerned to elucidate particular neuroscientific concepts and their role in theory. One example of the kind of question that is of interest to the philosophy of neuroscience comes from the theory of vision and its relation to the investigation of consciousness. It will be clear that this problem is closely connected to the neurophilosophical discussion of consciousness described above and exemplifies the fluidity of the boundary between neurophilosophy and the philosophy of neuroscience.

Visual experience represents the environment as having a range of different properties such as color, shape, and motion, and the perception of these features are subserved by different cortical regions. These elementary facts raise a family of questions that is usually referred to collectively as the *binding problem:* how do brain areas interact so that the color of a stimulus, for example, is associated with its shape? And how does this neural activity produce perceptual experience that is unified (see Hardcastle, 1997b)?

One recent proposal about how the binding problem is solved at the level of neurons runs as follows. Many neurons exhibit regular physiological patterns of firing called oscillations, and it has been suggested that synchrony of oscillation (especially around the frequency of 40 cycles per second, or 40 Hertz) is a sign of functional connection. If two neurons, one responsive to color and one to edges, were to fire synchronously, this temporal property could signal that the two neurons were responding to the same stimulus. Stimulus identity and perceptual coherence would thus be captured by the temporal features of neural activity. It has further been suggested that synchronous oscillatory firing could be a marker of perceptual consciousness. According to this view, 40-Hz oscillation not only solves the binding problem but brings the stimulus so bound into consciousness.

There has been a good deal of debate about the 40-Hz proposal, and it raises many questions, some empirical, some theoretical. Does 40-Hz oscillation have the right properties to solve the binding problem? What is the relation between neural oscillations and the computational processes posited in cognitive theories of vision? What exactly is the relation between binding and consciousness? Does the 40-Hz theory address both problems equally well? Will the 40-Hz account generalize to other perceptual modalities? If not, how is the binding problem solved across modalities? Some of these questions are being addressed by philosophers together with cognitive and neuroscientists, and it is this kind of investigation that may represent the collaborative and interdisciplinary future of the theory of the mind.

References

Akins K (1996) Of sensory systems and the 'aboutness' of mental states. *Journal of Philosophy* **93**: 337–372.

Churchland PM (1981) Eliminative materialism and propositional attitudes. *Journal of Philosophy* **78**: 67–90.

Churchland PS (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland PM and Churchland PS (1996) Replies from the Churchlands. In: McCauley RN (ed.) *The Churchlands and Their Critics*. Oxford, UK: Blackwell.

Davies M and Coltheart M (2000) Introduction: Pathologies of belief. *Mind and Language* **15**: 1–46.

Davies M, Coltheart M, Langdon R and Breen N (forthcoming) Monothematic delusions: towards a two-factor account. *Philosophy, Psychiatry, Psychology*.

Flanagan OJ (1995) Deconstructing dreams: the spandrels of sleep. *Journal of Philosophy* **92**: 5–27.

Flanagan OJ (1996) Prospects for a unified theory of consciousness. In: Cohen J and Schooler J (eds) *Scientific Approaches to the Study of Consciousness: 25th Carnegie Symposium*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gold IJ and Stoljar D (1999) A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Sciences* **22**(5): 809–830.

Hardcastle VG (1997a) When a pain is not. *Journal of Philosophy* **94**: 381–406.

Hardcastle VG (1997b) Consciousness and the neurobiology of perceptual binding. *Seminars in Neurology* **17**: 163–170.

Nagel T (1971) Brain bisection and the unity of consciousness. *Synthese* **22**: 396–413.

Further Reading

Bechtel W, Mandik P, Mundale J and Stufflebeam RS (eds) (2001) *Philosophy and the Neurosciences: A Reader*. Oxford, UK: Blackwell.

Bickle J (1998) *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Coltheart M and Davies M (eds) (2000) *Pathologies of Belief*. Oxford, UK: Blackwell.

Crick F and Mitchison G (1995) REM sleep and neural nets. *Behavioural Brain Research* **69**: 147–155.

Dennett DC (1976) Are dreams experiences? *Philosophical Review* **85**: 151–171.

Flanagan OJ (2000) *Dreaming Souls: Sleep, Dreams, and the Evolution of the Conscious Mind*. New York: Oxford University Press.

Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71.

Hardcastle VG (1999) *The Myth of Pain*. Cambridge, MA: MIT Press.

Hobson A (1999) *Dreaming as Delirium*. Cambridge, MA: MIT Press.

Munro A (1999) *Delusional Disorder*. Cambridge, UK: Cambridge Univerisity Press.

Philosophy of Science

Intermediate article

Pete Mandik, William Paterson University, Wayne, New Jersey, USA
William Bechtel, University of California, San Diego, California, USA

CONTENTS

Philosophy of science concerns the principles and processes of scientific explanation, including both processes of confirmation and of discovery.

Philosophy of science is primarily concerned to provide accounts of the principles and processes of scientific explanation. Early in the twentieth century, philosophers of science focused on the logical structure of scientific thought, whereas in the later part of the century logic was de-emphasized in favor of other frameworks for conceptualizing scientific reasoning and explanation, and an emphasis on historical and sociological factors that shape scientific thinking. While tracing through the landmarks of this history we note many points of contact between the philosophy of science and the cognitive sciences.

THE LOGICAL STRUCTURE OF SCIENCE

The Deductive-nomological Model of Explanation and Hypothetico-deductive Model of Theory Development

The appeal to logic to articulate the structure of scientific explanation and scientific reasoning was the hallmark of the Logical Positivists, a group of early twentieth-century philosophers and scientists, working initially in Eastern Europe, who sought to provide an explication of science that could explain its high epistemic status. They offered a model of explanation, the Deductive-nomological (D-N) model, which holds that explaining a phenomenon involves deducing its occurrence from laws (Hempel, 1966). Something like this view is thousands of years old and may be discerned in the work of Aristotle. The D-N model was extremely influential in some areas of psychology earlier in the twentieth century. Many behaviorists, for example, sought to discover general laws of learning to characterize how various kinds of experiences (e.g. reinforcement) would change the behavior of organisms.

Laws, which are central to the D-N model, are taken to specify general relations (as in Newton's law that force equals mass times acceleration ($f = ma$)). To apply these general relations to particular events, one must specify conditions holding at a previous time, which are usually called *initial conditions*. Recognizing that multiple laws and initial conditions may be involved in a given explanation, such explanations can then be represented in the following canonical form (where L designates a law, C an initial condition, and E the event to be explained):

$$\begin{array}{l} \text{L1, L2, L3, } \ldots \\ \underline{\text{C1, C2, C3, } \ldots} \\ \therefore \text{ E} \end{array}$$

Advocates of the D-N perspective generally assumed that the Cs and Es were sentences whose truth or falsity could be determined directly through observation. These *observation* sentences also provided the empirical support for the laws. Most advocates of a logical analysis of science disavowed interest in scientific discovery, restricting their focus to justification. Their characterization of the relationship between observations and laws make it clear why they saw little hope for a logic of discovery. (More recently, artificial intelligence (AI) researchers and philosophers sympathetic to AI have proposed discovery tools to employ within this framework: Langley *et al*., 1987; Thagard, 1988; Darden, 1991.) The challenge in discovery is to construct new laws to account for more or new phenomena. Since laws are to be general, not specific, they must go beyond any finite set of data, and thus there inevitably seems to be a major inductive leap to a new hypothesis. Once a hypothesis was put forward, the process of testing could begin by deducing its observational consequences

in accord with the D-N model. This process of hypothesizing general laws and testing their deduced consequences was termed the Hypothetico-deductive model of theory development.

Hempel (1966) illustrates H-D with the example of Semmelweis' work during the 1840s on childbed fever. Semmelweis observed cases of 'puerperal fever' or 'childbed fever' contracted by women who delivered children in his hospital. He noted that cases were especially frequent in groups of women where deliveries were handled by physicians instead of midwives. Semmelweis' key insight into the cause of childbed fever came when he observed that a physician came down with similar symptoms upon injuring himself with an instrument during an autopsy. Semmelweis hypothesized that 'cadaveric material' on the injurious instrument caused the disease, and that, similarly, the physicians associated with outbreaks of childbed fever had cadaveric material on their hands prior to delivering babies. Semmelweis tested this hypothesis by examining its implications. One implication of the hypothesis that cadaveric material is the cause of childbed fever is that its removal from the hands of physicians would result in a decrease in cases of childbed fever. When Semmelweis tested this implication by requiring that physicians wash their hands in chlorinated lime prior to examining patients (which he assumed would remove the cadaveric matter), he observed that groups of women examined by physicians who washed with chlorinated lime had lower incidents of childbed fever than groups of women examined by physicians who did not.

The example of Semmelweis' hypothesis and test conforms to the H-D model in the following way. Semmelweis' observations could be formulated as a series of observation statements, statements of particular states of affairs such as 'Jane Doe was exposed to cadaveric material and contracted childbed fever', 'Mary Smith was exposed to cadaveric material and contracted childbed fever', and so on. His hypothesis took the form of a law-like general statement, 'Any woman exposed to cadaveric material will contract childbed fever'. And, in accordance with D-N, the original observation statements may be deductively inferred from the statement of law. Thus the relation of observation statements to statements of law has a reciprocal structure as depicted by H-D and D-N. (See Figure 1.)

If the only support for a hypothesis were the observation sentences it was intended to explain, the reciprocal relationship between proposing hypotheses and explanation would be circular. Circularity is avoided, though, since a law, being a general statement, deductively entails not just the particular observation statements it was advanced to explain, but an indefinite number of others. The theory is confirmed by demonstrating the truth of some of these additional observation statements.

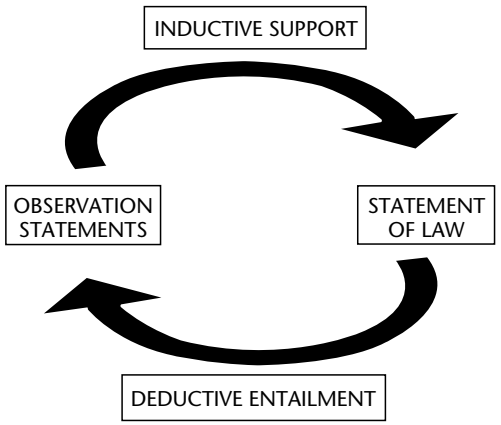


Figure 1. The reciprocal relationship between hypothesizing laws from observations and deriving observations from laws.

Intertheoretic Reduction

Recognizing that one might want to explain why laws held, the proponents also generalized this framework, allowing for the derivation of one or more sets of laws (each comprising a theory) from another set of laws (comprising another theory). The second set of laws will be more general ones from which, under specific boundary conditions, the first set of laws might be derived. (Thus, the boundary conditions replace the initial conditions in the above formalism.) Proponents also suggested that this approach might be extended to relations between laws in one science and those of a more basic science by providing bridge laws relating the vocabularies of the two sciences (by translating the terms of one into the terms of another), giving rise to the following schema:

Laws of the lower-level science
Bridge laws
Boundary conditions

$\therefore$ Laws of the higher-level science

These derivations are known as *reductions*; they figure prominently in discussions about the relation between psychology and neuroscience in which some theorists propose that the theories of

psychology ought to reduce to those of neuroscience (Churchland, 1986). Analogously, reductions are posited, or at least hoped for, for any two adjacent levels, such as from biology to chemistry and from chemistry to physics.

One example of a successful reduction that conforms to the above account is the derivation of the Boyle–Charles Law of classical thermodynamics (specifications of the temperature and pressure relations in an ideal gas) from statistical mechanics. Boyle–Charles Law terms such as *temperature* and *pressure* are translated into the terms of statistical measures of the kinetic properties of molecules in a volume. Equating temperature with mean kinetic energy supplies one of the bridge laws enabling translations of the Boyle–Charles Law into the laws of statistical mechanics. Boundary conditions include specifications of the kinds of molecules, the volume to which their motion is restricted, and the range of temperatures and pressures they are subject to. With translations of terms and boundary conditions in place, the Boyle–Charles Law is derivable from the laws of statistical mechanics.

Just as laws gain their empirical support from the true observation sentences that are derived from them, so reducing laws gain their support from the already confirmed laws that can be derived from them. Reduction, though, can also provide justification for reduced laws. Insofar as the reducing laws are more general, confirmation they receive in some domains can provide indirect support for other laws that can be derived from them.

CHALLENGES TO LOGICAL POSITIVISM

Popper's Critique of Confirmation

By making predictions which turn out true, the Logical Positivists thought we could justify laws. On this claim, however, they were criticized by Karl Popper (Popper, 1935/1959), who noted that such arguments had the invalid form of affirming the consequent:

If L were true, then prediction P would be true
P is true
∴ L is true

This formalism is invalid since it is possible for both premises to be true, but the conclusion false. Instead, Popper argued that the only way evidence could bear upon laws was through the use of *modus tollens* arguments in which failed predictions could be used to falsify a purported law:

If L were true, then prediction P would be true
P is false
∴ L is false

Accordingly, Popper emphasized that the method of science was a method of conjectures and refutations in which scientists proposed explanatory laws and then sought evidence showing that they were false. If a proposed law resisted all attempts at falsification, Popper would speak of it as corroborated, not as true or confirmed, recognizing that future evidence could always reveal it to be false.

Mechanisms instead of Laws

Although the D-N model seems to apply well to a number of scientific domains, especially in physics (for example, to explanations of phenomena that appeal to the laws describing ideal gases or principles of thermodynamics), it does not seem applicable to many domains in the life sciences. A major reason is that explanation does not tend to involve showing that a phenomenon follows a law (Cartwright, 1983; Giere, 1999). Instead, explanation often involves identifying and describing the mechanism that generates the phenomenon (Wimsatt, 1972; Machamer *et al*., 2000). A key component to the idea of mechanism is that of a set of processes that generate a phenomenon, with these processes being performed by different parts of a system. Thus, an explanation consists of functionally decomposing the process of producing the phenomenon into a set of different component processes and localizing these component processes in actual physical parts of a system (Bechtel and Richardson, 1993). (Often actually identifying the component physically is not possible, and researchers settle for indirect evidence that such a component exists.) For example, to explain basic physiological processes such as cellular respiration, biochemists had to identify a set of activities (oxidation of substrates, electron transfer, and phosphorylation of adenosine diphosphate (ADP)) and determine the components of the cell responsible for them (enzymes, co-factors, and membranes with restricted permeability).

An emphasis on mechanisms does not eliminate appeal to laws – sometimes key relations between parts of a system (e.g. between a substrate and an enzyme) are expressed in laws. But laws play a subsidiary role. The emphasis is on differentiating the operations performed in the system, linking

them with physical parts of the system, and then showing how the component parts and processes interact with each other to produce the phenomenon. Some of the individual interactions can be stated in laws, but the specification of the particular components involved and the intricacy of their interactions is generally far too specific to render into laws. Quite often such explanations are presented in diagrams (these are especially useful when the component processes are organized nonlinearly with multiple interactions and feedback loops), not linguistically (although diagrams are typically accompanied by linguistic commentary).

Many of the endeavours of the cognitive sciences can be interpreted as advancing mechanisms. Grammars, for example, are often presented as mechanisms for generating sentences. Psycholinguists who investigate the psychological reality of particular grammars are investigating whether they are actually implemented in language users. Researchers in AI, especially those creating programs to account for human performance, are decomposing an activity into component operations. Implementing the program on a computer provides an existence proof that the hypothesized set of operations is sufficient to generate the phenomenon in question. Behavioral experiments using such tools as reaction times and error analysis are required to demonstrate that the processes actually figure in human information processing.

When the explanatory vehicle is assumed to be a description of a mechanism, scientific inquiry is not restricted to producing data to be subsumed under a hypothesis or to being used in testing a prediction derived from a hypothesis. Early in the process of inquiry, researchers are simply trying to figure out what are the processes that contribute to a particular effect. One strategy is to show that processes are actually separable in the system by showing that one process can be impaired while the other is retained, or that an experimental manipulation produces a crossover interaction between measures of the two processes. When these manipulations are performed after a mechanism has been proposed, they serve the more traditional role of testing the proposed explanation. In such ways, empirical inquiry contributes both to discovery of models of mechanism and to testing them.

Alternative Conceptions of Reduction

The application of the model of mechanistic explanation in the previous paragraphs focused on the traditional disciplines of cognitive science: cognitive psychology, linguistics, and AI. But a very natural place to employ this framework is to relations between more traditional cognitive explanations and neuroscientific ones. Although interest in the neural realization of cognitive mechanisms did not play a critical role in cognitive science until recently, this was largely due to the paucity of techniques that could link cognitive and neural investigations (Bechtel *et al*., 1998). But the emergence of cognitive neuroscience as a major area of scientific collaboration in the 1990s reveals that cognitive and neural modes of investigation can be invoked together. Indeed, neuroimaging experiments, whether with positron emission tomography (PET), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), or evoked response potentials (ERP), require use of cognitive tasks and measures along with measures of neural activity. In terms of the model of mechanistic explanation, what these tools are providing is localization of the functions decomposed in a cognitive analysis of the task (e.g. by showing where various attentional processes are realized in the brain). Like more purely behavioral research, moreover, such studies can not only serve to confirm a cognitive decomposition, but also play a heuristic role in determining the functional decomposition itself (Bechtel *et al*., 2000).

Such connections between cognition and neuroscience are commonly construed as reductionistic. But since one is not starting with laws at the cognitive level, and deriving them from laws of neuroscience, such research does not fit the theory reduction model (above). Rather, relating a functional decomposition to a structural localization provides an alternative conception of reduction, one much closer to Darden and Maull's (1977) conception of an interfield theory. Such an account may be much closer to actual scientific practice. Moreover, it does not raise the specter of either explaining away the higher-level approach (with a successful reduction) or eliminating it (if reduction fails). Rather, it weaves the two approaches closely together; functional decomposition and structural localization are contributors to the common inquiry of understanding the mechanism. Each can provide not only support for the other, but heuristic guidance to the other, and the resulting explanatory account is an integrated one.

There is a further way in which reduction emerges in the context of a mechanistic model of explanation. Once a system has been decomposed into component functions and these localized within the system, a new explanatory task arises – explaining how each component function is performed. Developing this explanation requires

repeating the process: decomposing the component process into its component processes, and localizing these within the subsystem where the process from the first decomposition was localized. Recognizing that this can be done successively, the model of mechanistic explanation provides a framework for reduction through multiple levels (Wimsatt, 1976). There is an important point to recognize about this multilevel conception of reduction – at each level, a different phenomenon is being explained. Moreover, at each level, the explanation involves not only the contributing components, but the interactions that are specified in the account at that level. A process such as seeing is reductively explained in terms of the contributions of the different components of the visual system and their interactions, while it is the activity of a given component which is explained by going into it and identifying its subcomponents and their interaction.

THE SOCIOHISTORICAL STRUCTURE OF SCIENCE

Paradigms and Revolutions

While the Logical Positivists were themselves very interested in the science of their time, their account was grounded primarily in logic, not in the details of scientific practice. (A consequence of this is that they viewed it as a normative model characterizing any possible science.) Kuhn's work drew philosopher's attention (as well as that of historians and sociologists of science) to the specific details of the process of scientific research. Kuhn (1996) challenged the views that science employed a set of general methods that remained constant over time and accumulated a body of truths. In their stead, Kuhn suggested that scientific approaches vary so significantly at different eras of a discipline that the findings and theories at one time cannot be meaningfully related to the findings and theories of other times. Instead of viewing the historical progression of science as the progressive accumulation of truths, Kuhn offered a cyclic model of the stages of scientific activity through history. The cycle involves the five stages of: (1) immature science; (2) normal mature science; (3) crisis science; (4) revolutionary science; and (5) resolution, which is followed by a return to normal science (see Figure 2).

The key notion in understanding Kuhnian philosophy of science is the notion of a *paradigm*. The five stages of the Kuhnian cycle may be unpacked in terms of this notion, since normal science is paradigm-based science, and four of the stages are understood by way of contrast with normal science. For Kuhn, paradigms are 'Universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners'. Further, paradigms serve for a time implicitly to define the legitimate problems and methods of a research field for succeeding generations of practitioners. They do so in virtue of two essential characteristics (p. 10): first, 'Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity.' Second, their achievement 'was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve'. Examples of paradigms include Ptolemaic astronomy, Copernican astronomy, Aristotelian dynamics, Newtonian dynamics, corpuscular optics, and wave optics.

Normal science, then, just is paradigm-based science. Immature science is science studying a domain recognizably the same as that studied by paradigm-based successors, but without the utilization of any paradigms, as in the cases of optics prior to Newton and of electrical research in the first half of the eighteenth century (pp. 12–14). A stage of crisis emerges for a science when the requisite consensus regarding the applicability of a paradigm begins to unravel. Patterns of problem-solving in normal science give way to novel approaches, and where new consensus emerges, a stage of resolution of a new paradigm returns the cycle to normal science.

One of Kuhn's views that garners much attention in the philosophy of science is his claim that different paradigms are incommensurable and thus choice of one over another cannot be subject to rational procedures. One way of thinking of incommensurability of paradigms is by thinking of paradigms as involving different languages, the terms of which cannot be translated into each other. For

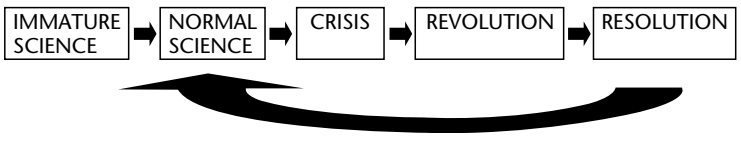


Figure 2. The stages in Kuhn's account of scientific change.

example, the term 'space' as used within a Newtonian physical paradigm cannot be translated as the term 'space' as used within an Einsteinian physical paradigm, and vice versa. Einsteinians differ from Newtonians in holding that space is curved by mass, for example. The meaning of 'space' depends on the theory it is embedded in; where theories diverge, so do the meaning of their terms. Part of Kuhn's argument that paradigms are not open to rational choice hinges on the notion that observation statements cannot serve as neutral points of arbitration since there is no theory-neutral observation language. In other words, observation is theory-laden: how one perceives the world depends on the theory with which one conceives the world. For instance, a newborn baby, if shown a cathode-ray tube, would not see it *as* a cathode-ray tube because the baby would understand an insufficient amount of theory to know what cathode-ray tubes are.

Kuhn's thesis that paradigms are incommensurable leads to another of his theses, namely, that the history of a scientific discipline is noncumulative. Noncumulativity follows from incommensurability in the following way. Since the language of one paradigm cannot be translated into the language of another, the statements held to be true within one paradigm cannot be expressed, let alone judged to be true, within another. The theory-ladenness of perception guarantees that not only can items of theory not be accumulated across paradigms, but neither can observations. Since not even observations are retained from one paradigm to the next, the history of a discipline cannot be viewed as the accumulation of truths or as progressing towards a truer account of the world. This challenges the traditional view of science as an intellectual enterprise that progresses over time. The history of science, as viewed through the Kuhnian lens, is of a series of paradigms and revolutions, none bearing any rational relation to any other.

Challenges to Kuhn

Kuhn's hypothesis of incommensurability has been challenged on several fronts. One denies that the meanings of theoretical terms are determined wholly by factors internal to a paradigm, but holds instead that they may be determined, at least in part, by causal relations between the term and items in the external world. Putnam (1975) suggests that the meaning of certain scientific terms involves causal relations between the terms and things in the world that they denote. For instance, part of the meaning of water is the substance H_2O that was present when the term 'water' was first brought into use to denote that substance. A causal chain leads from current uses of 'water' to the initial dubbing of H_2O as water. These causal chains remain constant regardless of a scientist's theory. Thus, water discourse need not be incommensurable between adherents of divergent theories about water. Appealing to causal relations to fix the content of representations is mirrored in much cognitive scientific practice, especially in the cognitive neurosciences where representations are thought of as carrying information about their environmental causes.

Another challenge to Kuhnian incommensurability arises from theorists who propose that the mind is modular. Fodor (1984), for example, argues that many perceptual processes are modular in the sense of being 'informationally encapsulated' so that their outputs are immune from influence by theoretical and other acquired beliefs. Fodor, therefore, contends that observational reports can be treated as univocal even when theorists hold different theories.

Though many of the issues raised by Kuhn continue to be debated, the Kuhnian spirit is a pervasive force in post-Positivist philosophy of science. It has, for example, led philosophers to focus much more on the diachronic nature of science as well as the actual research processes of science (see, for example, Lakatos, 1970; Laudan, 1977) or to apply Bayes Theorem to model rational theory choice (Howson and Urbach, 1993; Mayo, 1996). None the less, no consensus has emerged in contemporary philosophy of science that is on a par with the status once accorded the Positivists. For cognitive scientists, one of the attractive features of current philosophy of science is the increasing effort to employ ideas from cognitive science in the attempt to understand science (Giere, 1992; Thagard, 1988).

References

Bechtel W, Abrahamsen A and Graham G (1998) The life of cognitive science. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 1–104. Oxford, UK: Blackwell.

Bechtel W, Mandik P and Mundale J (2000) Philosophy meets the neurosciences. In: Bechtel W, Mandik P, Mundale J and Stufflebeam RS (eds) *Philosophy and the Neurosciences: A Reader*. Oxford, UK: Blackwell.

Bechtel W and Richardson RC (1993) *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies*. Princeton, NJ: Princeton University Press.

Cartwright N (1983) *How the Laws of Physics Lie*. Oxford, UK: Clarendon Press.
Churchland PS (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.
Darden L (1991) *Theory Change in Science: Strategies from Mendelian Genetics*. Oxford, UK: Oxford University Press.
Darden L and Maull N (1977) Interfield theories. *Philosophy of Science* **43**: 44–64.
Fodor J (1984) Observation reconsidered. *Philosophy of Science* **51**: 23–43.
Giere R (1992) *Cognitive Models of Science*. Minneapolis, MN: University of Minnesota Press.
Giere R (1999) *Science without Laws*. Chicago: University of Chicago Press.
Hempel CG (1966) *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice-Hall.
Howson C and Urbach P (1993) *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.
Kuhn TS (1996) *The Structure of Scientific Revolutions*, 3rd edn. Chicago: University of Chicago Press.
Lakatos I (1970) Falsification and the methodology of scientific research programmes. In: Lakatos I and Musgrave A (eds) *Criticism and the Growth of Knowledge*. Cambridge, UK: Cambridge University Press
Langley PS, Simon HA, Bradshaw GL and Zytkow JM (1987) *Scientific Discovery: Computational Explorations of the Creative Process*. Cambridge, MA: MIT Press.
Laudan L (1977) *Science and Relativism*. Berkeley, CA: University of California Press.
Machamer P, Darden L and Craver CF (2000) Thinking about mechanisms. *Philosophy of Science* **67**: 1–25.
Mayo D (1996) *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Popper K (1935/1959) *The Logic of Discovery*. London: Hutchinson.
Putnam H (1975) The Meaning of 'Meaning'. In: *Mind, Language, and Reality: Philosophical Papers of Hilary Putnam*. Cambridge, UK: Cambridge University Press.
Thagard P (1988) *Computational Philosophy of Science*. Cambridge, MA: MIT Press.
Wimsatt WC (1972) Complexity and organization. In: Schaffner KF and Cohen RS (eds) *PSA 1972*: *Proceedings of the 1972 Biennial Meeting of the Philosophy of Science Association* pp. 67–86. Dordrecht: Reidel.
Wimsatt WC (1976) Reduction, levels of organization, and the mind–body problem. In: Globus G, Maxwell G and Savodnik I (eds) *Consciousness and the Brain: A Scientific and Philosophical Inquiry*. New York: Plenum Press.

Further Reading

Bechtel W (1988) *Philosophy of Science*. Hillsdale, NJ: Lawrence Erlbaum.
Glymour CN (1980) *Theory and Evidence*. Princeton, NJ: Princeton University Press.
Hacking I (1983) *Representing and Intervening*. Cambridge, UK: Cambridge University Press.
Laudan L (1990) *Science and Relativism: Some Key Controversies in the Philosophy of Science*. Chicago: University of Chicago Press.
Simon HS (1996) *The Sciences of the Artificial*, 3rd edn. Cambridge, MA: MIT Press.
Thagard P (1992) *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Possible Worlds Semantics

Intermediate article

Manuel García-Carpintero, University of Barcelona, Barcelona, Spain

CONTENTS

Possible worlds semantics is a framework for semantics that takes as primitive the notion of possible world. This framework is used to explain how modal expressions creating intensional contexts work, and more generally to characterize propositional content.

WHAT IS POSSIBLE WORLDS SEMANTICS?

We contrast the way the world is with alternative ways that it might be. In reviewing the course of a chess game, we may explore how the game would have gone if we had made alternative moves. As a broad generalization of this notion, we may postulate alternatives to the entire course of the world, in every detail. Possible worlds semantics (PWS) is a framework for semantic theorizing that takes as primitive for explanatory purposes this notion of 'possible world' (counting the actual world as one such), together with less specific logical and set-theoretic notions. The immediate explanatory task of PWS is to characterize the behavior of modal expressions, such as subjunctives, *necessary*, *contingent* and *possible*, *obligatory* and *permissible*, *must* and *may*, which create so-called intensional contexts. The framework is also used to characterize propositional content in general. PWS is applied in accounts of natural languages, and other representational systems, including intentional mental states.

Wittgenstein (1921) introduced PWS, to characterize propositional content, and Carnap (1947) elaborated on Wittgenstein's suggestions. Both Wittgenstein and Carnap focused on declaratives, sentences that count as true or false. Leaving aside 'analytic' sentences (*bachelors are male*), whose truth is constitutive of the meaning of certain expressions, whether a sentence is true or false is not for a linguistic theory to establish. Competent speakers can fully understand *Mallory reached the summit of Everest* without knowing whether in fact he did. As Wittgenstein pointed out, however, if one understands a declarative, one typically knows conditions that might in fact obtain in the actual world, whose obtaining would make the declarative true. Wittgenstein and Carnap explicated this central semantic notion of truth-conditions in terms of possible worlds: to know the truth-conditions of a declarative is to know the class of possible worlds such that, if any of them were actual, the declarative would be true. This does not mean that in understanding *Mallory reached the summit of Everest* one contemplates in every detail each possible world compatible with its truth. It suffices to have generic knowledge, discriminating a subclass of the class of possible worlds including only worlds in which two specific individuals, Mallory and the summit of Everest, stand in a certain relation, that of the former reaching the latter.

PWS is also used to characterize the semantic behavior of modal expressions. Semantic theorists assume that the semantic features of a sentence are compositionally determined by those of its lexical units and some of its syntactic features, constituting the sentence's 'logical form' (LF). Logicians traditionally distinguish two semantic features of expressions: 'extensions' and 'intensions'. Expressions that can occupy the position of the restriction of a determiner (*all … think*) or its scope (*all unicorns …*) are called 'predicates'. The extension of a predicate (relative to a given domain) is the class of entities (in that domain) of which the predicate is true. For example, the predicates *renate* (of an animal with kidneys) and *cordate* (of an animal with a heart) have the same extension.

A position in LF is 'extensional' if and only if replacement of the expression occupying it by another coextensional (of the same LF category) always preserves the original sentence's truth-value: expressions in those positions appear to signify their extensions. It was once debated among logicians whether the restriction and the scope of quantificational determiners are extensional contexts; most now agree with Frege (1892) that they

are. Other positions in LFs are non-extensional in particular, positions governed by modal expressions. Thus, *if that organism were a renate, it would be a renate* is true, but the sentence obtained by substituting *cordate* for the second *renate* can intuitively be taken to be false; similarly with *necessarily, if an organism is renate, it is renate*. Such positions in LFs are said to be 'intensional': expressions occupying them signify their 'intensions', which, in the case of predicates, have traditionally been characterized as concepts.

In PWS, the intension of a predicate is a function that maps a possible world to the extension the predicate would have if the possible world were actual. The concept of extension is extended from predicates to expressions of other LF categories. The extension of a referential expression like a proper name or a demonstrative is its referent, and the extension of a sentence is its truth-value; the extension of a simple sentence consisting of a referential expression and a predicate results compositionally from applying the extension of the predicate to the extension of the referential expression. The concept of intension is generalized accordingly. In PWS, the meaning of a modal expression is defined relative to the intensions of the expressions on which it operates, so that semantic intuitions concerning non-extensional contexts are accounted for. Thus, the truth-conditional import of *necessarily* in *necessarily*, …, with a sentential expression filling in the blank, is such that the complete sentence is true if and only if the intension of the sentence filling in the blank is the function that assigns truth to every possible world. The intuition that *cordate* and *renate* cannot be interchanged *salva veritate* in certain sentences is explained by the fact that they have different intensions, i.e. different extensions in some possible world. Replacing one with the other may result in a sentence with a different truth-value if the sentence included intension-sensitive expressions.

Before the introduction of PWS, modal logicians, including C. I. Lewis, advanced and studied different systems of modal logic characterized syntactically. The system of propositional modal logic *T* results from adding to a system appropriate for propositional logic the inference rule of necessitation – allowing the deduction of any instance of the schema $\Box\phi$ from the null set of premises whenever an instance of ϕ is so deducible – and the axiom schemata $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$ and $\Box\phi \rightarrow \phi$. (In modal logic, the square '$\Box$' is used for 'it is necessary that' or 'it must be that'.) There is also the system *S4*, obtained by adding to *T* the axiom schema $\Box\phi \rightarrow \Box\Box\phi$; and *S5*, obtained by adding to *S4* the axiom schema $\Diamond\Box\phi \rightarrow \phi$. (The diamond, '$\Diamond$', is used for 'it is possible that' or 'it is permissible that', equivalent to '$\neg\Box\neg$'.)

The relationships between the different systems, and their potential applications, became clearly understood through the use of PWS, in particular through the notion of an accessibility relation among worlds with different set-theoretic properties (simultaneously discovered by Hintikka (1957), Kanger (1957) and Kripke (1959)). A modal sentence like $\Box\phi$ is evaluated relative to a given possible world w; it is true relative to w if ϕ is true in all possible worlds accessible from w. The requirement that the accessibility relation be reflexive validates *T*. Thus, if ϕ is true relative to a given world w, it is true relative to any world w' accessible from w; given that the accessibility relation is reflexive, w is itself one of those worlds, and therefore ϕ must also be true in w. Similarly, the requirement that the accessibility relation be transitive validates *S4*, and the requirement that it be symmetric validates *S5* (Hughes and Cresswell, 1995). Together with a measure of 'closeness' or similarity between worlds, PWS has also been used to provide a clear understanding of the semantics of counterfactuals (Lewis, 1973; Stalnaker, 1968). Roughly, according to Lewis's semantics, 'if it were the case that p, then it would be the case that q' is true with respect to possible world w if and only if q is true with respect to all the closest worlds to w in which p is also true.

ADVANTAGES OF PWS

The knowledge that linguistic theories articulate is possessed by speakers in an intuitive form. Intuitive knowledge leaves more matters undecided than is required by the nature of what is thereby known. Correct theoretical articulations of intuitive information help to determine such matters. PWS articulates our modal intuitions, about the truth-values of sentences including modal expressions, and about the modal information encoded by declaratives in general, by explicitly conceptualizing a type of entity, the possible world. This allows for explicit reference to possible worlds and quantification over them. Developments of Montague's (1974a, 1974b) program, the most fully elaborated application of PWS, reap the benefits of this. For example, the nature of scope ambiguities like the one in *Alex will necessarily marry a Catholic* for instance, can be perspicuously explained. Given the PWS explication of truth-conditions, this ambiguity is accounted for as one of the already familiar scope ambiguities induced by the interaction of universal

and existential quantifiers, as in *every congressman hates some journalist*. The two readings can be paraphrased thus: *in all possible worlds, there is a Catholic whom Alex will marry; there is a Catholic whom Alex will marry in all possible worlds*.

Modal expressions are highly context-sensitive: the class of all possible worlds with respect to which the above sentence is true, in either of its readings, is not the class of all possible worlds whatsoever. According to the PWS account, this has to do with the fact that different contexts select different domains of possible worlds: those in which certain moral norms obtain, say, or those that share their physical laws with the actual world. This introduces an additional measure of complexity and intuitive indeterminacy, which PWS helps to resolve.

The context-sensitivity of modal expressions as articulated in PWS also accounts for a remarkable fact about modal logic. Logicians interested in the logical facts essentially involving modal expressions tried to systematize the relevant valid sentences and inference patterns in the form of axiomatic theories, without providing a semantics to justify them. In the case of what we now know as first-order logic, the work of Frege, Russell, Löwenheim, Tarski and others, following this methodology, soon produced a unique system. This was conspicuously not so in the case of modal logic, as illustrated above. On the basis of the semantics that PWS provides for the different systems of modal logic, the explanation can be suggested that, when different domains of possible worlds are at stake, different modal propositions are valid. Thus, instances of the *T* axiom are not valid when the necessity operator is understood deontically. If an instance of *it must be the case that* ϕ is true in the actual world, it is true in all worlds accessible from it, i.e. all worlds in which certain norms obtain; but the actual world itself need not be one of them. If, however, physical necessity is at stake, the axiom is valid.

In support of PWS, Wittgenstein and Carnap pointed out that it provides a clear and intuitively appealing account of one of the central semantic notions, entailment. A sentence *S* is entailed by a class of sentences *K* if and only the class of worlds with respect to which all sentences in *K* are true includes all worlds with respect to which *S* is also true. If we take the relative degrees of informativeness of sentences to increase with the number of possible worlds excluded by their truth, this notion of entailment clarifies the intuitive idea that the information provided by a sentence *S* entailed by a class of sentences *K* is already contained in *K*.

PROBLEMS WITH PWS

Descriptive Problems

The main descriptive limitation of PWS concerns whether the propositional contents expressed by natural-language sentences are fully captured by it. Consider any contingent sentence *p*, i.e. any sentence that is false in some non-actual possible world (*Hesperus shines in the morning*), and any necessary sentence *q* (*two plus two equals four*). Intuitively, *p* and *p* and *q* have different linguistic meanings. However, they appear to be true with respect to the same possible worlds, and thus to have the same intension. PWS also ascribes the same intension to all necessary truths. Stalnaker (1984) has attempted to make this result palatable. Cresswell (1985) pursues another strategy, substituting structured meanings for bare PWS intentions.

A related difficulty is Frege's problem of cognitive significance. This is exacerbated by compelling views advanced by Kripke (1980) and Kaplan (1989) on the truth-conditional import of referential expressions like proper names and indexicals inside the PWS framework. Kripke discusses deep-seated semantic intuitions best explained by taking genuinely referential expressions to be 'rigid designators': when they designate an object *o* they designate it with respect to every possible world in which *o* exists, and do not designate anything else with respect to any possible world. As Kripke points out, this view has the consequence that genuinely referential expressions with the same referent, such as *Hesperus* and *Phosphorus*, have the same truth-conditional import when this is explicated according to PWS. This makes Frege's problem of cognitive significance particularly salient, to the extent that cognitive significance depends on meaning: *Hesperus is Phosphorus* can be informative for a competent speaker, while *Hesperus is Hesperus* is not. To deal with this problem, Stalnaker (1978), Davies and Humberstone (1980) and others have advanced so-called 'two-dimensional' versions of PWS, which ascribe two different intensions to expressions, without apparently departing from the PWS framework.

Foundational Problems

The main foundational problem for PWS is of an ontological nature: are we metaphysically entitled to posit such entities as possible worlds? Defenders of PWS have advanced various views regarding their ontological status. These include Lewis's (1986) view that each possible world has the same

status as the actual world constituted by 'me and my surroundings', Stalnaker's more moderate but still realist view that they are complex properties that the actual world might have, and Carnap's view that they are like fictional stories.

A second foundational objection is Quine's argument that quantified modal logics pragmatically commit their proponents to a discredited view, Aristotelian essentialism: roughly, the view that, among the properties that objects instantiate, some are constitutive of their identities, and thus possessed by them no matter what possibilities may obtain regarding them, and some are accidental. It is, however, unclear whether any commitment to essentialism within modal logic is indefensible. Kripke's (1980) influential discussion of these matters has helped to restore the credibility of essentialism.

PWS AND COGNITIVE SCIENCE

Cognitive science is an interdisciplinary research effort to provide scientific understanding of matters involving the representational states of ordinary people, in particular those states that constitute semantic competence in their native language. A common objection to PWS is that ordinary competent speakers of a language, whose semantic knowledge linguistic theories attempt to characterize, know nothing of such things as possible worlds, and thus that PWS is of no use for cognitive scientists.

This conclusion is unwarranted. A semantic theory need not confine itself to notions that are part of the competent speaker's metasemantic knowledge. We have suggested reasons to believe that a PWS should be included in theoretical accounts of the semantics of natural language. This suggests that competent speakers somehow know about possible worlds. They know about them in virtue of their capacity to distinguish what is necessary from what is merely possible, and also in virtue of their capacity to provide and receive information by means of linguistic sentences – for, as Wittgenstein and Carnap argued, this is already a capacity to discriminate among alternative possibilities. Ordinary speakers may lack a theoretically articulated conception of possible worlds; but this is surely the case for many other theoretically interesting notions required for scientific characterizations of their representational states.

References

Carnap R (1947) *Meaning and Necessity*. Chicago, IL: University of Chicago Press.

Cresswell MJ (1985) *Structured Meanings*. Cambridge, MA: MIT Press.

Davies M and Humberstone L (1980) Two notions of necessity. *Philosophical Studies* **38**: 1–30.

Frege G (1892/1984) On sense and meaning. In: McGuinness B (ed.) *Collected Papers on Mathematics, Logic, and Philosophy*, pp. 157–177. Oxford, UK: Blackwell. [English translation.]

Hintikka J (1957) Quantifiers in Deontic Logic. *Societas Scientiarum Fennica, Commentationes Humanarum Litterarum* **23**(4): 3–23.

Hughes G and Cresswell M (1995) *A New Introduction to Modal Logic*. New York, NY: Routledge.

Kanger S (1957) *Provability in Logic*. Stockholm, Sweden: Almqvist & Wiksell.

Kaplan D (1989) Demonstratives. In: Almog J, Perry J and Wettstein H (eds) *Themes from Kaplan*, pp. 481–563. New York, NY: Oxford University Press.

Kripke SA (1959) A completeness theorem in modal logic. *Journal of Symbolic Logic* **24**: 1–15.

Kripke SA (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis DK (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis DK (1986) *On the Plurality of Worlds*. Oxford, UK: Blackwell.

Montague RM (1974a) English as a formal language. In: Thomason R (ed.) *Formal Philosophy: Selected Papers of Richard Montague*, pp. 188–221. New Haven, CT and London: Yale University Press.

Montague RM (1974b) The proper treatment of quantification in ordinary English. In: Thomason R (ed.) *Formal Philosophy: Selected Papers of Richard Montague*, pp. 247–270. New Haven, CT and London: Yale University Press.

Stalnaker R (1968) A theory of conditionals. In: Rescher N (ed.) *Studies in Logical Theory*, pp. 98–112. Oxford: Blackwell.

Stalnaker RC (1978) Assertion. In: Cole P (ed.) *Syntax and Semantics*, vol. IX 'Pragmatics', pp. 315–332. New York, NY: Academic Press.

Stalnaker R (1984) *Inquiry*. Cambridge, MA: MIT Press.

Wittgenstein L (1921/1992) *Tractatus Logico-Philosophicus*, translated by D. F. Pears and B. F. McGuinness. Atlantic Highlands, NJ: Humanities Press.

Further Reading

Carnap R (1947) *Meaning and Necessity*. Chicago, IL: University of Chicago Press.

Cresswell MJ (1994) *Language in the World*. Cambridge, UK: Cambridge University Press.

Hintikka J (1962) *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Hughes G and Cresswell M (1995) *A New Introduction to Modal Logic*. New York, NY: Routledge.

Kaplan D (1989) Demonstratives. In: Almog J, Perry J and Wettstein H (eds) *Themes from Kaplan*, pp. 481–563. New York, NY: Oxford University Press.

Kripke SA (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lewis DK (1970) General semantics. *Synthese* **22**: 18–67.
Lewis DK (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.
Lewis DK (1986) *On the Plurality of Worlds*. Oxford, UK: Blackwell.
Montague RM (1974a) English as a formal language. In: Thomason R (ed.) *Formal Philosophy: Selected Papers of Richard Montague*, pp. 188–221. New Haven, CT and London: Yale University Press.
Montague RM (1974b) The proper treatment of quantification in ordinary English. In: Thomason R (ed.) *Formal Philosophy: Selected Papers of Richard Montague*, pp. 247–270. New Haven, CT and London: Yale University Press.
Partee B (1989) Possible worlds in model-theoretic semantics: a linguistic perspective. In: Allen S (ed.) *Possible Worlds in Humanities, Arts and Sciences*, pp. 93–123. Berlin and New York, NY: de Gruyter.
Plantinga A (1974) *The Nature of Necessity*. Oxford, UK: Oxford University Press.
Stalnaker R (1968) A theory of conditionals. In: Rescher N (ed.) *Studies in Logical Theory*, pp. 98–112. Oxford, UK: Blackwell.
Stalnaker RC (1978) Assertion. In: Cole P (ed.) *Syntax and Semantics*, vol. IX 'Pragmatics', pp. 315–332. New York, NY: Academic Press.
Stalnaker R (1984) *Inquiry*. Cambridge, MA: MIT Press.

Propositional Attitudes

Intermediate article

Jay L Garfield, Smith College, Northampton, Massachusetts, USA

CONTENTS

Propositional attitudes are psychological states such as belief, desire, hope, and fear that seem to take propositions as their intentional states. Propositional attitudes constitute a central concern for cognitive science, and questions about the attitudes focus on the relations among diverse branches of cognitive science.

INTRODUCTION

The term 'propositional attitude' refers to such psychological states as belief, desire, hope, fear, etc. which, at least *prima facie*, take propositions as their intentional objects. Even this much is controversial but will do as a preliminary characterization. A propositional attitude comprises the subject of the attitude, the attitude, and the proposition towards which the attitude is directed. Albert [subject] may believe [the attitude] *that unicorns eat hay* [the proposition]. But so might Beatrice [subject]. Carlos [subject] might hope [the attitude] *that unicorns eat hay* [the proposition]. Deirdre [subject] might hope [the attitude] *that they eat only clover* [the proposition].

Propositional attitude verbs sometimes do not take propositions as objects. Eric might fear *that a wolf will eat him* (a propositional attitude), but he might simply have an inchoate fear: he fears *wolves*. In English, the word 'that' (sometimes elided as in 'Eric fears a wolf will eat him') marks the propositional attitude construction. Propositional attitudes are those states that are reported by verbs that take 'that clauses' as objects (perhaps minus the verbs of saying, which can also take quoted expressions as objects). Given that clear marking of complementation is not universal, and that 'that clauses' have no obvious syntactic analogues in some languages, this may be a rather parochial characterization, however.

Even where we have a classical propositional attitude construction, there is considerable disagreement among philosophers and cognitive scientists whether the object of the attitude is actually a proposition (and, indeed, exactly what a proposition is). Some argue that it is properly conceived as a sentence (Carnap, 1931); others that it is merely a property and a sequence of objects satisfying that property (Russell, 1912; Quine, 1956); still others maintain that these verbs in fact take no objects (Sellars, 1956; Kiteley, 1964).

The propositional attitudes constitute a central concern for cognitive science, and questions about the attitudes focus on the relations among diverse branches of cognitive science. Psychologists worry about the degree to which propositional attitudes explain behavior, and the mechanisms by which they do so, if indeed they do; neuroscientists are concerned with their physical realization in the brain (if indeed they are realized in the brain); cognitive developmentalists investigate the acquisition of our representation of the attitudes; linguists are concerned with their semantics; and philosophers ask whether there really are propositional attitudes, or indeed propositions, and whether if indeed there are, they can be localized in the brain, and what the relation is between propositional attitudes and the language used to represent their apparent objects. These questions are intertwined, and this entanglement is what makes the exploration of the nature of this (apparent) class of psychological phenomena so focal to cognitive science.

CRUCIAL PROPERTIES AND PHENOMENA

We can identify five crucial properties that sentences ascribing propositional attitudes must explain according to any theory which posits them as genuine phenomena: their opacity; the *de dicto*/*de re* ambiguity; complementation; content-sensitive causality; systematicity.

Propositional attitude contexts are often (though as we will see below, not always) opaque. In propositional attitude contexts we cannot substitute coreferring terms: (1) Frederique believes that the

richest woman in the world lives in Seattle; (2) Frederique kicked the richest woman in the world. Unknown to Frederique, 'the richest woman in the world' and 'Queen Elizabeth II' refer to the same individual. While it follows from (2) that Frederique kicked the queen, it does not follow from (2) on its most obvious reading that Frederique believes that QE II lives in Seattle. We can substitute coreferring terms in transparent contexts ascribing physical attitudes, such as kickings, but not in those ascribing propositional attitudes (*de dicto*), such as believings. Existential generalization over referring terms occurring in propositional attitude contexts also fails. (3) George hopes that Santa will come down the chimney. (4) Hetta sat on Santa's knee at the mall. It does *not* follow from (3) that there is somebody George hopes will come down the chimney. It *does*, however, follow from (4) that there *is* somebody upon whose knee Hetta sat. Transparent contexts permit 'quantification in'; propositional attitudes typically do not.

The second feature of the attitudes complicates this picture. For there are some propositional attitude contexts in which substitution of coreferring expressions and quantification in are permitted. The systematic semantic ambiguity of propositional attitude reports between those in which substitution and quantification is prohibited (*de dicto* reports) and those in which it is permitted (*de re* reports) is another hallmark of the propositional attitudes. So, knowing that Ignatius hopes that he will be the owner of Black Beauty, I might say 'Iggie hopes that he will own a horse', quantifying in to the clause giving the content of the attitude. And even though Jocasta reports her desire with the sentence, 'I want to marry Oedipus,' we can explain the unfolding of the tragedy by the permissible substitution into this *de re* context of the coreferring expression 'her son', even though she would never do so. There are important debates concerning whether the *de dicto/de re* ambiguity marks a distinction between two kinds of psychological states or between two kinds of reports of psychological states (Garfield, 1988).

Propositional attitude reports embed the sentences representing the contents of the attitude reported as complements of the propositional attitude verbs. Syntactically, the contents are reported in clauses which themselves are sentences and from which such operations as wh-extraction are permitted. So the content of Jeremiah's hope in 'Jeremiah hopes that there will be joy to the world' is reported by a complete sentence, and one can follow this utterance with 'What did Jeremy hope would be joyful?' and answer 'The world'. Direct quotation, for instance, ('He said "joy to the world" ') neither requires a sentential object nor permits extraction, even where a sentential object is present. Semantically, the truth-value of the complement sentence does not contribute to the truth-value of the entire sentence. Karl Marx said that the dictatorship of the proletariat will be followed by the withering away of the state. We can truly ascribe false beliefs.

When we use propositional attitudes in psychological explanation, we expect that the causal properties of the attitudes mirror the contents of the attitudes we ascribe and their logical relations to one another. So, we expect that Leila's belief that there is chocolate in the box, and her desire that she have chocolate and her knowledge *that Lottie buried the key*, will generate her hope that Lottie will dig up the key. When we posit nonpropositional attitude states as causes there is no comparable content-sensitivity expected of corresponding causal patterns.

Finally, propositional attitudes exhibit a closely linked triad of properties referred to as 'systematicity', 'compositionality', and 'productivity' (jointly the 'systematicity properties'). That propositional attitudes are systematic means that when someone is able to stand in any one or any set of attitudes, the referent is inevitably capable of standing in a range of closely related attitudes: from the fact that Margaret is able to believe that the Eagles will defeat the Swans, it follows that she is able to believe that the Swans will defeat the Eagles, that she is able to hope that the Swans will defeat the Eagles, etc. The productivity of the attitudes is the fact that from a finite vocabulary or set of concepts we can generate, using logical and syntactic operations, an infinite number of possible objects of attitudes. Since Norbert believes that Noam yet lives, and since he believes that Noam is nativist, he also believes that a nativist yet lives, and that not all nativists are dead, etc. Compositionality is the mirror image of productivity. We can explain Ophelia's capacity to fear that the prince is mad by her grasp of the concept of madness and her possession of a representation of the prince. The fact that she can stand in attitudes to novel propositions is explained by the fact that she can determine their meanings based upon her knowledge of the meanings of their parts and of their structure.

Any theory of the attitudes must explain these phenomena. This constrains psychological theories of the attitudes, semantic theories of attitude ascriptions, and philosophical accounts of the role of the attitudes in cognitive science. If nothing

satisfies at least most of this cluster of properties, there are no propositional attitudes.

BRIEF HISTORY AND SUMMARY OF PRINCIPAL POSITIONS

Frege pioneered research into the attitudes and into propositions in his two essays, *On Sense and Reference* (1892) and *The Thought* (1899), in which he proposes the idea of compositionality and makes the first proposal for an explanation of opacity, suggesting a theory of propositions as objects of thought. Frege distinguishes between the sense and the reference of an expression, arguing that in the context of propositional attitude verbs expressions denote not their customary reference but their sense. This, he argues, explains the failure of substitutivity of coreferential terms (they may have different senses) as well as the failure of existential generalization (some senses fail to be instantiated). Moreover, since the senses of the components of sentences tally to form the sense of an entire sentence (a proposition), and since propositions (what sentences express, as opposed to their truth-values, which Frege argues are the referents of sentences) are precisely what the mind grasps in thought, the fact that the attitudes are attitudes towards *propositions* explains why the attitudes are opaque. Frege's influence on research into the semantics of propositional attitude locutions in the twentieth century cannot be overstated.

Russell (1912) argued that for a proposition to be true just is for a sequence of individuals in the world to have a property, and that the mind is directed towards that state of affairs in a propositional attitude, and not to some intermediate Fregean abstract entity. If Paul believes that elephants can fly, Paul's belief is not, Russell argued, directed on the sense of that English sentence, but rather on elephants, and the property of flying. And the fact that those pachyderms do not possess that property explains the falsity of Paul's belief. Russellian propositions are hence concrete sequences of objects and properties, and a propositional attitude is therefore a relation between a subject and such a sequence. This approach is gaining favor among those disenchanted both with the metaphysics of representations intermediate between subjects and the world and those who see recent attempts to formalize the Fregean approach as encountering intractable problems.

Carnap (1931), suspicious of abstract entities and of properties, thought of propositional attitudes as sentential attitudes, replacing the idea of a proposition with that of an equivalence class of intertranslatable sentences. The attitudes, on this account, are behavioral dispositions towards sentences or their translation-equivalents, or to their truth-conditions. Querida's belief that cows eat grass is her disposition to utter that sentence when asked what cows eat, or, depending on her native language, perhaps to reply 'bzas kyi red' when asked 'ba phyugs kyis tsha bzas kyi red pas?' Her hope that it will not rain tomorrow might be a disposition to say, 'I hope it won't rain tomorrow', or just her disposition to be elated at tomorrow's blue sky. Carnap's account has been criticized both for being overly behavioristic and for failing to expunge abstracta (equivalence classes of sentences may be as problematic as Fregean propositions) but also for its reliance on linguistic items as the direct objects of the propositional attitudes (Fodor, 1978). None the less it has also been influential in part because it has demonstrated a kind of naturalism with respect to the attitudes that harmonizes well with the empirical goals of cognitive science.

Davidson (1968) follows Carnap, taking sentences seriously as objects of attitude ascriptions, and in taking them as functioning as exhibited exemplars of translation-equivalence classes. Davidson argues that propositional attitude locutions are properly understood not as containing sentences embedded as complements in that-clauses but rather as the demonstrata of paratactic constructions. So on Davidson's view, 'Rudolf believes that propositional attitudes are dispositions towards sentences' has the following structure: *Rudolf believes that: Propositional attitudes are dispositions towards sentences*. The lexical specificity of the demonstrated sentence explains opacity. Pragmatic considerations regarding what counts as an equivalent sentence explains the availability of *de re* readings. The linguistic character of the object of belief explains the systematicity phenomena.

Sellars (1956) and Kiteley (1964) have argued for the elimination of the objects of the attitudes entirely, treating the content-clauses of attitude ascription sentences as adverbial, characterizing the manner in which one believes, desires, etc. So on this reading, when we say that Susan believes that Plato was wise, the phrase 'Plato was wise' modifies 'believes' in the same way that 'quickly' modifies 'walks' in 'Susan walks quickly'. This approach eliminates metaphysically problematic entities, but its resources for explaining the fact that we seem to be able to quantify over objects of belief are not obvious, and the explanations this approach suggests for the various properties particular to the attitudes strike some as ad hoc.

The last third of the twentieth century saw an explosion of research into the formal semantics of propositional attitude expressions. Montague's (1974) possible world semantics for English is enormously influential. Though not entirely successful, it suggests a plausible and rigorous strategy for formalizing Frege's insights regarding opacity and compositionality, and predicts many of the central properties of propositional attitude expressions. This research program is still vigorous, prosecuted most notably by Cresswell (1984). Other significant formal approaches include the situation semantics of Barwise and Perry (1983) and the discourse representations semantics of Kamp (1981).

There is also much recent debate concerning the psychological role and realization of the propositional attitudes, much of it inspired by Sellars's (1956) argument for the theoretical character of the attitudes and their essential connection to language. Fodor (1975) argues for the reality of a language of thought and for a model of the attitudes according to which they are relations to internal sentential tokens in that language of thought. Churchland (1979) has argued that there is no such language of thought, and hence that there really are no propositional attitudes, since there are no available sentential representations in the mind. Others (Baker, 1988; Garfield, 1988) have argued that the reality of the propositional attitudes does not depend on the reality of internal representational states and that the attitudes in fact supervene broadly on environmental and linguistic facts. These debates continue to be active, with virtually every imaginable position regarding the ontology of the propositional attitudes finding a champion.

Semantics of Propositional Attitude Locutions

Frege's distinction between sense and reference and his emphases on compositionality and a truth-conditional theory of meaning guided most approaches to the semantics of natural language in the twentieth century. Frege regarded the reference of each sentence to be its truth-value, and the sense to be the proposition it expresses; the reference of each predicate expression to be the set of (sequences of) entities satisfying that predicate and the sense to be the property in virtue of which they do; the reference of every name its bearer, and the sense an individual concept, and so on. The sense of each expression determines its reference. The reference of a sentence, he argues, is determined by composing those of its components; the sense of the sentence is determined by a composition of senses.

In the Montague grammar framework the intentionality of the propositional attitudes is represented by the intensionality of the contexts propositional attitude verbs create, and compositionality is realized by representing the semantic values of all phrases as functions with functional composition as the primary semantic operation. Montague's model for natural language is a structure comprising an infinite set of possible worlds with a universal accessibility relation defined over them, a designated actual world, and an interpretation function assigning nonlogical constants values at each world. The logic is a higher-order intensional logic with modality and tense augmented with lambda-abstraction and with a semantics relying heavily on functional application. Extensional contexts are treated in a standard first-order way through evaluation in the actual world: intensional contexts require evaluation in other worlds, following techniques familiar from Kripke's semantics for modal logic.

The great virtue of Montague's approach is that the semantics provides a rigorous account of how a compositional semantics can preserve all of the crucial features of propositional attitude contexts. Because within the intensional context created by the propositional attitude verb the semantic value of each expression is a function from worlds to extensions, expressions with different meanings will contribute different functions to the meaning of the proposition composed, thus generating distinct propositions, and explaining the failure of substitutivity. So, for instance, let us recall Frederique's belief that the richest woman in the world lives in Seattle, and her failure to believe that the Queen of England lives in Seattle. Without going into too much detail, the intension of 'the richest woman in the world' will be a function from possible worlds to individuals which will yield Elizabeth II in exactly those worlds where she in fact is the richest woman *in that world*. The intension of 'the Queen of England' will be a function from possible worlds to whoever is the Queen of England in each world, yielding Elizabeth II in exactly the worlds where she reigns. Now in the actual world, these two expressions have the same reference, Elizabeth II. If the expressions in these two belief contexts each contributed their reference, there would be no difference between these beliefs and so it would be impossible for Frederique to believe one and not the other. But since there are worlds in which Wendy Gates is wealthier than the queen (e.g. the world where Bill gave all of his

fortune to her), there are worlds in which the function the semantics assigns to one of these propositions yields true and in which that assigned to the other is false. Hence we can see that these propositions are different, and one might be believed by Frederique and the other not believed.

Montague semantics also explains the *de dicto*/*de re* ambiguity very neatly in terms of a contest of scope between quantifiers and the propositional attitude verbs that create intensional contexts. A consequence of the alternative orderings of the quantifiers and verbs in the logic is that *de dicto* and *de re* attitude verbs will have different semantic types, thus explaining the distinct properties of the contexts they create. Suppose that I see Tenzin admiring a yak and I say, 'Tenzin desires to purchase a yak', having that particular yak in mind, thus attributing the attitude *de re*. I see Uma reading a primer on yak care and feeding, and say of her, 'Uma desires to purchase a yak', having no yak in particular in mind, hence attributing the desire *de dicto*. Montague grammar regards each of these sentences as syntactically (and hence semantically) ambiguous, and the ambiguity concerns whether the quantifier phrase 'a yak' applies to the entire phrase 'Uma/Tenzin desires to purchase x', or whether the attitude phrase 'Uma/Tenzin desires' applies to the entire phrase 'Uma/Tenzin purchases a yak'. In the first case we get the reading we want in the Tenzin sentence – true – just in case there is a yak which has the property of being desired by Tenzin. In the second case we get the *de dicto* reading appropriate to Uma – true – just in case she has the property of desiring to own any yak whatsoever.

Montague's system, however, is beset by a number of difficulties. Some of these concern the lack of fit between Montague's syntax and that posited by the most powerful current syntactic theories; some concern the oddness of asserting that the objects of our beliefs are functions from worlds to truth-values; others concern the grain at which Montague grammar can distinguish distinct readings of sentences. Cresswell (1984) has introduced the latest refinement of Montague's system, showing how to explain semantic ambiguities in attitude ascriptions that Montague's formalism cannot clearly represent.

Russell's intuition that objects and properties themselves, as opposed to sets of worlds or functions, are the semantic values of propositions has led to alternative approaches to understanding the semantics of propositional attitude ascriptions. Barwise and Perry's (1983) situation semantics is an excellent example of this approach, as is Jubien's (1997) theory of the attitudes. There is no clear consensus at this point on the correct way to explain the semantic properties of the attitudes.

Propositional Attitudes, the Ontology of Mind, and Methodology of Psychology

Taking propositional attitudes seriously as explanations of human behavior and cognition poses questions about how the attitudes are related to phenomena relevant to other levels of explanation relevant to behavior, such as the biological level. Some (e.g., functionalists or other identity theorists) argue that each propositional attitude is identical to some token neural event or process. Others argue that each type of mental state is identical to some type of physical state. So on a token-identity view, if Van believes that unicorns are mythical, there is some state or event in Van's brain which literally *is* that belief. On a type-identity view, there is some kind of brain state which, wherever it occurs, is the belief that unicorns are mythical. So that if Yolanda and Van share that state, they share that belief.

Identity theories are designed to account for the causal properties and explanatory force of the attitudes. If one adopts the view that all causation is physical and that all explanation at any but the most fundamental physical level must in turn be explained by more fundamental regularities and processes, it is natural to conclude that psychological processes must at bottom be physical processes and that the most plausible candidates are neural processes and states. Neural causation then is only psychological causation.

Moreover, to the extent that propositional attitudes are causally efficacious, representational, and systematic, the causal properties of the representations must track the properties that determine their compositional, productive, and systematic character. A good explanation for these properties is the presence of an underlying recursive system of representations. Thus, Fodor (1975) argues, there must be in the brain a physical system of symbols with a recursive syntax over which operations are performed whose causal properties map onto properties of inference and other forms of thought. In other words, the propositional attitudes are in fact relations to expressions in a language of thought. This also explains how infralingual organisms such as infants and nonhuman animals can stand in propositional attitudes despite having no language in which to express propositions. This hypothesis has proven fertile in cognitive science and has been mobilized to explain

early learning and first language learning as well as the possibility of causation by propositional attitudes.

The language of thought hypothesis has come under attack from a number of directions. Connectionist models of thought often implement superpositional representation in which no representation can be localized in a network, or models in which representations sustaining propositional interpretation are in fact activation vectors with no compositional or syntactic structure. To the extent that these models are correct and to the extent that they vindicate the use of propositional attitudes in psychological explanation, the language of thought hypothesis is dubious.

Philosophers concerned with the ontology of the attitudes have pointed out that the assumption that the reality and causal efficacy of the attitudes demands their neural realization, and hence some form of the identity theory, is not necessarily correct. Anti-individualists (Baker, 1988; Garfield, 1988) point out that the identity conditions of many representational phenomena are relational and that their ontological bases are broad. To be a token of an English word is not to be an ink mark of a particular shape but to play a certain role in a much larger pattern, and one must describe that entire pattern to specify what it is to be a token of a particular word. Similarly, these theorists argue, for Winnie to believe that beliefs are in the head may not require that there be any particular state or process going on in Winnie's head. Instead, it could well be that Winnie bears a complex set of relations to expressions in her own language, to inferences she is likely to make, and to lines of inquiry she is likely to pursue.

Most radically, eliminative materialists (Churchland, 1979) have argued that there really are no propositional attitudes, at least not if our standard of reality is that they actually play a role in the correct characterization and explanation of psychological processes. Eliminative materialists argue that our best models of brain processes make no reference to anything that satisfies the criteria for being a propositional attitude and that the best science of human behavior will be one grounded in brain theory. Moreover, eliminativists argue, a psychology based in the propositional attitudes would be incapable of explaining a large range of human behavior that is plausibly continuous with the remainder of our psychology, and incapable of explaining the behavior of infralinguals with whose psychology ours is also plausibly continuous. Finally, they argue, anti-individualists are probably correct in thinking that no identity theory can possibly be true of the propositional attitudes, but since any biological science of behavior must only advert to individualistically construed states of organisms, this circumstance merely shows that the propositional attitudes are not genuine causal determinants of behavior or cognition.

At present eliminativism is a minority position and has had little effect in empirical cognitive science, where theoretical discourse is still rich in propositional attitude language. If this is an unavoidable situation, we must face the challenge that the propositional attitudes pose for the reduction of psychology and other intentional sciences to sciences like neuroscience whose language is typically physicalistic and extensional. The problem is particularly acute if one thinks that higher-level sciences must be reducible to lower-level sciences *and* that reduction involves the definition of all terms in the higher-level science into the language of the lower-level science, together with the derivation of the laws of the higher-level theory from those of the lower. No such set of definitions will ever be forthcoming with respect to the propositional attitude terms, nor any derivations of the relevant regularities. This forces one either to abandon the view that all true theories must form part of a unified science, with sciences unified by intertheoretic reduction, or to adopt a somewhat weaker notion of reduction to guide the integration of cognitive science. Each approach can be and has been defended (Fodor, 1974; Churchland, 1979; Garfield, 1988).

A second methodological problem is posed by the need to integrate the psychology of language users with that of infralingual organisms. As we have seen, according to many of the plausible analyses of the propositional attitudes they are, *au fond*, relations to linguistic items. If this is the case, then it is hard to see how infralingual organisms could be subjects of propositional attitudes. But it seems equally plausible that there is no catastrophic gap between the psychology of infralinguals and language-users.

There are two principal strategies for resolving this problem. The first is to argue that indeed, appearances to the contrary notwithstanding, there *is* a catastrophic gap between language-users' and infralingual psychology, at least on some dimensions, and in fact part of that gap is to be accounted for by the fact that infralinguals do not have genuine propositional attitudes (Davidson, 1968; Garfield, 1988). This permits one to retain the language-directed analyses of the attitudes and harmonizes with results suggesting a dramatic change

in the quality of thought accompanying language acquisition (de Villiers and de Villiers, 2000). However, this view demands an alternative account of the obviously powerful representational capacities of such infralinguals as human infants and the great apes. The second strategy is to adopt the language of thought hypothesis. This allows one to have a uniform account of the attitudes and psychological processes of infralinguals and language-users. The problems with this strategy include the fact that it relies on a strong, unverified hypothesis about brain function, and it runs afoul of anti-individualist arguments. This debate remains unresolved.

The debate over individualism with respect to the propositional attitudes also raises methodological questions. If each propositional attitude is realized entirely in states within the body of its subject, we have a problem about how to individuate the attitudes. For if Xerxes believes that pigs can fly, there must be something about his state that makes it a belief (as opposed, say, to a fear) about pigs (not birds) and asserting of them that they fly (and not that they dig truffles). And it seems hard to imagine anything simply about Xerxes' brain or indeed the rest of his body, in isolation from his environment, that could fix the content and hence the identity of his belief. But if we have to go beyond Xerxes' body and examine the relation of his brain state to pigs and flying, we have lost individualism. But giving up individualism about the attitudes raises its own problems. How are we to develop a theory adverting to propositional attitudes if their boundaries are so vast and so vague? How, even if we allow them to extend beyond the body to the distal phenomena that determine their content, are we to know to what to turn in order to fix their identity?

These worries suggest another methodological option with regard to the attitudes. Cognitive science could treat them instrumentally and not be realistic about the attitudes. To do this would be to treat propositional attitude talk purely as a useful calculational device. The virtue of this position is that one can simply dismiss all of the metaphysical and methodological problems just identified as pseudo-problems relevant only to a science positing real propositional attitudes. The difficulty is the same as that afflicting eliminativism. It is at least implausible to say that nobody has ever really believed anything, desired anything, or feared anything, and certainly this is not something that a consistent instrumentalist can lead herself to believe!

Acquisition of Propositional Attitude Vocabulary

A great deal of research has been devoted to understanding the mechanisms by which human beings acquire the concepts of the propositional attitudes and the ability to ascribe attitudes to themselves and others and to explain and to predict behavior using the attitudes (Baron-Cohen, 1995; Lewis and Mitchell, 1994). Given that the attitudes are unobservable inner states, it is remarkable that by age 4 most normally developing children are quite competent in this domain. It is noteworthy that very few 3-year-olds are and that this pattern is remarkably consistent across cultures. On the other hand, there are specific disorders, such as autism, in which children never develop this set of abilities or concepts, or if they do, do so very late and without any of the fluency that characterizes normal human attribution of the attitudes. This discovery of a very consistent acquisition pattern, coupled with highly stereotyped disorders specifically disabling the representation of the attitudes, has led some (Baron-Cohen, 1995) to posit an innate cognitive faculty responsible for the representation of and reasoning about the attitudes. This explanation of these phenomena gains additional support from data concerning other aspects of psychological development, including the development of the ability to monitor and to exploit attention.

On the other hand, others (Gopnik and Meltzoff, 1997) have suggested that the development of this knowledge and skill is achieved through the deployment of children's general drive to theorize, with the propositional attitudes posited as theoretical constructs to explain and to predict behavior. It has also been proposed that the development of these capacities is driven by social and linguistic development (Garfield *et al*., 2001; Peterson and Siegal, 1999).

The hypothesis that language acquisition is central to the development of the ability to ascribe propositional attitudes is supported by a range of findings: deaf children of hearing parents who acquire language late, also acquire these abilities late, and not until they acquire linguistic fluency; whereas deaf children of deaf parents with normal language acquisition show normal acquisition of these abilities (Peterson and Siegal, 1999). Moreover, the best predictor of the development of the ability to reason using the attribution of propositional attitudes is the development of the ability to understand and to use complement constructions of the kind necessary to attribute propositional

attitudes (de Villiers and de Villiers, 2000). There is still a great deal of controversy regarding the correct explanation of these phenomena and the nature of our knowledge about the attitudes. But explorations of this question promise to yield a great deal of knowledge about the structure of the mind and the nature of development, and perhaps about the nature of the propositional attitudes themselves, their relation to language, and the degree to which propositional attitude discourse is an essential piece of our cognitive architecture.

References

Baker L (1988) *Saving Belief*. Princeton, NJ: Princeton University Press.

Baron-Cohen S (1995) *Mindblindness*. Cambridge, MA: MIT Press.

Barwise J and Perry J (1983) *Situations and Attitudes*. Cambridge, MA: MIT Press.

Churchland P (1979) *Scientific Realism and the Plasticity of Mind*. Cambridge, UK: Cambridge University Press.

Carnap R (1931) Psychology in a physical language. *Erkenntnis* **II**: 432–465.

Cresswell M (1984) *Structured Meanings*. Cambridge, MA: MIT Press.

Davidson D (1968) On saying that. In: Davidson D *Essays on Truth and Interpretation*. Oxford, UK: Clarendon Press.

de Villiers J and de Villiers P (2000) Linguistic determination and the understanding of false beliefs. In: Mitchell P and Riggs K (eds) *Children's Reasoning and the Mind*. East Sussex, UK: Psychology Press.

Fodor J (1974) Special sciences. Reprinted in Fodor J *Representations* (1981). Cambridge, MA: MIT Press.

Fodor J (1975) *The Language of Thought*. New York, NY: Thomas Crowell.

Fodor J (1978) Propositional attitudes. Reprinted in Fodor J *Representations* (1981). Cambridge, MA: MIT Press.

Frege G (1892) On sense and reference. Reprinted in Geach P and Black M (eds) *Frege's Philosophical Writings*. (1952) Oxford, UK: Blackwell.

Frege G (1899) The thought. Reprinted in Strawson P (ed.) *Philosophical Logic* (1967). Oxford, UK: Oxford University Press.

Garfield J (1988) *Belief in Psychology*. Cambridge, MA: MIT Press.

Garfield J, Peterson P and Perry T (2001) Social cognition, language acquisition and the theory of mind. *Mind and Language* **16**: 494–541.

Gopnik A and Meltzoff A (1997) *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Jubien M (1997) *Contemporary Metaphysics*. Cambridge, UK: Cambridge University Press.

Kamp H (1981) A theory of truth and semantical interpretation. In: Groendijk *et al*. (eds) *Formal Methods in the Study of Natural Languages*: Amsterdam, Netherlands: Centre for Linguistics.

Kiteley M (1964) The grammars of belief. *Journal of Philosophy* **LXI**: 244–259.

Lewis C and Mitchell P (1994) *Children's Early Understanding of Mind*. Hillsdale, NJ: Lawrence Earlbaum.

Montague R (1974) *Formal Philosophy*. New Haven, CT: Yale University Press.

Peterson C and Siegal M (1999) Representing inner worlds: theory of mind in autistic, deaf and normal hearing children. *Psychological Science* **10**: 126–129.

Quine WV (1956) Quantifiers and propositional attitudes. *Journal of Philosophy* **53**: 177–187.

Russell B (1912) *The Problems of Philosophy*. London, UK: Oxford University Press.

Sellars W (1956) Empiricism and the philosophy of mind. In: Feigl H and Scriven M (eds) *Minnesota Studies in the Philosophy of Science I*. Minneapolis, MN: University of Minnesota Press.

Further Reading

Baker L (1995) *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge, UK: Cambridge University Press.

Bilgrami A (1994) *Belief and Meaning*. Oxford, UK: Blackwell.

Carruthers P (1996) *Language, Thought and Consciousness*. Cambridge, UK: Cambridge University Press.

Churchland P (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.

Davies M and Stone T (eds) (1995) *Folk Psychology*. Cambridge, UK: Blackwell.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Dowty M, Wall R and Peters S (1981) *Introduction to Montague Semantics*. Dordrecht, Netherlands: Reidel.

Ludlow P (ed.) (1997) *Readings in the Philosophy of Language*. Cambridge, MA: MIT Press.

MacDonald C and MacDonald G (eds) (1995) *Philosophy of Psychology*. Oxford, UK: Blackwell.

Millikan R (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Richard M (1990) *Propositional Attitudes: An Essay on How We Ascribe Them*. Cambridge, UK: Cambridge University Press.

Qualia

Intermediate article

Torin Alter, University of Alabama, Tuscaloosa, Alabama, USA

CONTENTS

Qualia are the subjective aspects, the characteristic properties, of conscious experience. There is extensive philosophical debate about how qualia relate to the physical world, and in what terms they can be characterized.

INTRODUCTION

Conscious experiences involve neural activity and information processing. They also feel a certain way. Consider your visual experience of reading these words, or the auditory and tactile sensations you had when you turned the previous page. There is something it is like to have those experiences. That is, the experiences have certain properties characterizing what it is like to have them. Those properties are known as *qualia* (singular: *quale*). C. I. Lewis coined the term in 1929. Common synonyms include 'phenomenal properties' and 'phenomenological properties', among others. Phenomenally conscious states, by definition, are states with qualia.

Recent decades have witnessed vigorous philosophical debate about qualia. Most of the controversy concerns whether qualia can be adequately characterized in physical or functional terms. If they cannot, then physicalism and functionalism, two leading theories of mind, are incomplete or false. Other issues include whether qualia exist, which mental states have qualia, how qualia relate to cognition, how physical systems such as brains give rise to qualia, how qualia can be scientifically studied, what the neural correlates of qualia are, which creatures have mental states with qualia, and how qualia are known. All of these issues have empirical components. In some cases, such as the neural correlates issue, scientific investigation is under way.

DO QUALIA EXIST?

'Qualia' is sometimes defined narrowly, in ways that give rise to substantive issues about their existence. For example, some reserve the term for properties that are nonphysical by definition. On that usage, the debate over whether conscious experiences have irreducibly nonphysical properties (discussed below) is a debate over whether qualia exist. And, to take a second example, Daniel Dennett (1988) reserves 'qualia' for properties that are by definition ineffable, intrinsic, private, and immediately apprehensible in consciousness. He attributes belief in their existence to various errors. For instance, he argues that conscious experiences seem to have ineffable properties because they have *practically* ineffable properties. Even the most detailed descriptions one can fathom might fall short of capturing what it is like to, say, taste a pomegranate. Nevertheless, on Dennett's view, a sufficiently detailed, accurate physical/functional description would leave nothing about such an experience unexplained. Thus, Dennett concludes, qualia do not exist.

But if qualia are defined broadly, as the properties characterizing what it is like to have conscious experiences, then their existence is hard to deny. Here 'qualia' should be understood in the broad sense.

QUALIA AND OTHER MENTAL PHENOMENA

Mental states with qualia include bodily sensations such as pains, itches, and orgasms, and perceptual experiences such as seeing, hearing, and hallucinating. Candidates for other states with qualia include at least: emotions such as lust, fear, and grief; moods such as depression, euphoria, and anxiety; thoughts one thinks silently but explicitly; perception of sentences of a language one understands; and cognitive attitudes such as desire, regret, and even belief.

Some use 'qualia' in such a way that by definition only sensory states can have qualia. On that usage,

even if there is something it is like to have a belief, beliefs do not have qualia. But on the most common usage, nonsensory states are not excluded from having qualia by definition. There may be a substantive issue about whether there is something it is like to have a belief, and therefore about whether beliefs have qualia.

Many believe that sensations have their qualia essentially. On that view, for example, no state that lacks pain qualia would count as pain. Opinions vary widely on whether the same should be said of emotions, moods, and other mental states. Disagreements over that issue tend to reflect divergent attitudes towards the overall relationship between qualia and cognition.

The assumption that qualia and cognition are closely linked has a distinguished history. It was more or less standard in seventeenth- and eighteenth-century Western philosophy, including especially (though not only) British empiricism and the Kantian tradition. In a different form, it pervades the writings of Brentano and other phenomenologists. (None of those figures used the term 'qualia', which was introduced in 1929. The preferred term was 'consciousness', and the assumption was usually implicit.)

Attitudes have since changed dramatically, due in part to the influence of behaviorism and the subsequent development of cognitive psychology. Many have come to believe that cognition and intentionality can and should be investigated without paying attention to qualia. Indeed, that opinion predominates in contemporary philosophy of mind and cognitive science, and has done so for more than half a century. But there are prominent dissenting opinions. John Searle argues that intentionality depends essentially on consciousness, which on his view entails qualia by definition. He writes:

> There is a conceptual connection between consciousness and intentionality that has the consequence that a complete theory of intentionality requires an account of consciousness (1992, p. 132).

ARE QUALIA IRREDUCIBLE?

How qualia relate to the physical world is controversial. Some doubt they can be explained in physical terms at all. The discussions usually centre on thought experiments, to which we now turn.

The Knowledge Argument

Perhaps the most widely discussed thought experiment about qualia is Frank Jackson's (1982) case of Mary, the brilliant scientist. Mary is raised in a black-and-white room, but learns all the physical information (all the physical facts) about human color vision by watching lectures on black-and-white television. That includes all the information in completed physics, chemistry, and biology, and everything that follows from that information (including functional information). Then she leaves the room and sees colors for the first time. Intuitively, it would seem that she thereby learns something new. For example, she learns what it is like to see red. So, Jackson concludes, there is nonphysical information about qualia. That is the knowledge argument against physicalism, the view that everything, mental and nonmental alike, is physical.

Physicalists have challenged each of the knowledge argument's assumptions. Some question whether one can learn all the physical information without experiencing color first-hand (e.g. Dennett, 1991). That objection is natural but unpopular; physical information as traditionally conceived is fully explicable in the objective language of science. Others (e.g. Lewis, 1988) argue that what Mary acquires when she leaves the room is not information but rather abilities: abilities to imagine, recognize, and remember color experiences. That objection is also relatively unpopular. As many note, what Mary gains when she finally sees red bears characteristic marks of informational knowledge. For example, the present author (Alter, 2001) argues that she might retain her new knowledge even if she loses the corresponding abilities, which is generally true of informational knowledge. A third objection to the knowledge argument runs as follows: when Mary leaves the room she acquires only new ways to represent information already in her possession. Before leaving the room, she uses physical concepts to represent the facts about color qualia. After leaving the room, she uses phenomenal concepts to represent those same facts (e.g. Loar, 1990). That objection is popular. But many remain unconvinced that phenomenal concepts pick out physical properties rather than distinctive properties of their own.

The knowledge argument has much in common with an argument advanced in Thomas Nagel's classic paper 'What is it like to be a bat?' (1974). Nagel argues that our inability to adopt the subjective viewpoint of echolocating bats prevents us from understanding essential aspects of their mental lives, and that no amount of objective, physical information would render our understanding complete. His reasoning is so similar to Jackson's that the knowledge argument is often attributed to both philosophers.

Absent Qualia

Another familiar thought experiment is usually discussed in connection with functionalism, the view that mental states consist in their causal relations to one another and to sensory stimuli and behavioral responses. The thought experiment is designed to show that the functional organization of a sentient creature could be realized in a system that has no mental states with qualia. In Ned Block's (1978) example, China's population organizes itself in a way that is isomorphic to the functional organization of a human brain. Individual citizens simulate the behavior of individual neurons, radio links correspond to synapses, and the system controls a robotic body. In Block's view, such a system might feel nothing, despite being a functional duplicate of a conscious human being. For example, it might feel no pain. If so, then qualia cannot be explained solely in functional terms.

David Chalmers (1996) argues that the absent-qualia hypothesis challenges not only functionalism but also any version of physicalism. Just as a qualia-free functional duplicate of a conscious human being seems possible, a qualia-free physical duplicate seems possible. Such creatures are known as phenomenal zombies (not to be confused with the Hollywood variety, which may have qualia and are functionally unlike ordinary humans).

Functionalists and physicalists sometimes respond by challenging the coherence of the absent-qualia hypothesis. For example, Shoemaker (1975) argues that a true functional duplicate of a conscious human must have introspective beliefs about its own sensory states, which on his view entails that some of its states have qualia. Another reply is to concede that the absent-qualia hypothesis is coherent, but deny that it undermines functionalism or physicalism. Here many invoke the Kripkean (Kripke, 1972) notion of *a posteriori* necessity, which may be explained as follows. That water is H_2O is a metaphysically necessary truth, which would obtain even if the laws of nature were different. Yet we know that truth only *a posteriori*; conceptual reflection alone cannot reveal the metaphysical impossibility of water existing without H_2O. Likewise, the argument runs, conceptual reflection cannot reveal whether absent-qualia cases are metaphysically possible. And, the argument continues, in fact they are not (e.g. Loar, 1990).

The latter response, though popular, has its problems. The Kripkean reasoning depends on the clear-cut distinction between the ordinary concept of water, which is given by its superficial features, and water itself, the essence of which consists in its molecular structure. Yet there appears to be no analogous distinction between the ordinary concept of pain and pain itself. There might be something, in some possible world, with the superficial appearance of water that is not H_2O. But, on most views, in any possible world if something feels like pain, then it is pain.

Inverted Qualia

A third familiar thought experiment, which Locke mentions (1690, bk II, chap. 32), involves inverted qualia. Imagine that your color experiences are inverted relative to mine. For example, ripe tomatoes look to me the color grass looks to you, and vice versa. We both call ripe tomatoes 'red' and grass 'green', but our qualia are inverted. There is empirical evidence that such cases may actually occur (Nida-Rumelin, 1996). The inverted-qualia hypothesis is that the whole range of one's color qualia could be inverted relative to a functionally identical twin.

The debate over inverted qualia parallels the debate over absent qualia in at least three respects. First, like the absent-qualia hypothesis, the inverted-qualia hypothesis is often used in arguments against functionalism. Second, some generalize the inverted-qualia hypothesis in the same way, arguing that physically indistinguishable creatures could have inverted qualia just as functionally indistinguishable creatures could. Third, reductionists reply in similar ways to the objections based on absent and inverted qualia; they argue that the cases are incoherent or metaphysically impossible.

It does not follow, however, that the absent-qualia and inverted-qualia hypotheses stand or fall together. On Shoemaker's view, the former is incoherent but the latter is not. Also, the two hypotheses raise different problems for reductive explanation. The absent-qualia hypothesis challenges reductive explanations of the existence of qualia (of having mental states with any qualia), whereas the inverted-qualia hypothesis challenges reductive explanations of the nature of specific qualia (of having red qualia as opposed to green qualia, for example).

Inverted Earth and Swampman

Recently, philosophers have been reflecting on two further thought experiments. One is the case of Inverted Earth (Block, 1990). Inverted Earth is just

like Earth, except the sky is yellow, grass is red, and so on. In the middle of the night, kidnappers drug you, transport you to Inverted Earth, and place you in your counterpart's bed. They also change your body pigments and put color-inverting lenses in your eyes, so that you are unaware of any difference (the two inversions cancel each other out).

According to Block, your new linguistic and physical environment will eventually produce changes in the intentional contents of your mental states. In time, your blue experiences will be about yellow things, your red experiences will be about green things, and so on, just like the other inhabitants of Inverted Earth. In Block's view, you will then be both intentionally and functionally inverted with respect to your former self, but your qualia will remain invariant. Inverted Earth thus creates another problem for functionalism.

Inverted Earth also challenges representationalism, the view that qualia are just representational or intentional properties. On that view, blue experiences are equated with perceptual states that represent blue things. Representationalism is popular among reductionists, who combine it with an appropriately reductionist account of mental representation (e.g. Tye, 1995). But if Block is right about Inverted Earth, then qualia can vary independently of, and thus fail to reduce to, intentional properties.

In response, some reductionists argue that the move to Inverted Earth affects qualia in ways that are not subjectively evident. And many deny that the move affects intentional content. The latter reply sometimes involves an appeal to a teleofunctional account of intentionality, on which the intentional content of color qualia is determined by the evolutionary history of one's species. Nature designed a certain type of human perceptual state to track blue things. That remains true even though, after enough time on Inverted Earth, that type of perceptual state is usually produced in you by seeing yellow objects. Therefore, the reply runs, your color experiences' intentional content never switches.

The idea that qualia are teleofunctional-representational properties encounters a difficulty from another thought experiment, devised by Donald Davidson (1986). In Davidson's story, lightning hits a dead tree in a swamp and creates Swampman, a molecule-for-molecule duplicate of a normal human being. Swampman has no evolutionary history. *A fortiori*, its experiences did not evolve for any biological purpose. Therefore, the teleofunctional-representation theory seems to lead to the conclusion that Swampman's states lack qualia, which many find counterintuitive.

QUALIA AND CAUSATION

The Hard Problem and the Explanatory Gap

The thought experiments discussed in the previous section relate closely to what Chalmers calls the hard problem of consciousness: how could a physical system such as the brain give rise to qualia? That problem has long seemed intractable, despite substantial progress in neuroscience and cognitive psychology. We are learning much about how the brain processes sensory stimulation, how it integrates information, and related matters. But why is such processing accompanied by qualia? There appears to be (in Joseph Levine's (1983) phrase) an explanatory gap: it seems that no amount of functional or physical information we might acquire about the brain would explain how it generates conscious experience.

Those who deny that qualia exist dismiss the explanatory gap as illusory. But, as noted earlier, those philosophers characteristically use 'qualia' in a narrow sense in which qualia are irreducible by definition. Even qualia eliminativists recognize that there is a problem, if only an empirical one, of how the brain generates conscious experience. Many qualia reductionists also regard the hard problem as unremarkable, according it the status of other unsolved problems in science. Eliminativists and reductionists sometimes compare the contrary inclination to the mindset of the vitalist, who puzzles over how life could emerge from mere physical processes and insists that nothing we could learn from biology or chemistry would remove the mystery (e.g. Dennett, 1988). Such comparisons may have value, but they can mislead. In the case of life, the phenomena requiring explanation are plainly complex functions, such as how a living system reproduces and how it adapts to its environment. But whether qualia can be explained functionally is a central point of controversy.

Other qualia reductionists agree that there is an explanatory gap, but attribute the problem to the distinctive character of phenomenal concepts. Those philosophers deny that the gap has strong metaphysical consequences. In particular, they argue, it does not undermine physicalism, because qualia can be identified *a posteriori* with physical properties (Loar, 1990).

Some are convinced that the hard problem is insoluble. For example, Colin McGinn (1989) argues

that, although there must be a naturalistic solution, we are constitutionally incapable of comprehending it – not because it is intrinsically difficult to grasp (though it might be), but rather because our distinctive cognitive capacities are ill suited to the task. Others are more optimistic. On Chalmers' view, a complete explanation will require psychophysical principles that connect physical processes and qualia. He proposes a framework for such principles, on which information is treated as basic and has both phenomenal and physical aspects. Nagel (1998) also suggests that qualia and physical properties may be manifestations of some deeper phenomenon, but he thinks our present concepts distort the underlying reality. He proposes that we try to develop new concepts to close the explanatory gap, modeled on Maxwell's development of the concept of a magnetic field, which enabled us to comprehend the relationship between electricity and magnetism.

Epiphenomenalism

The hard problem concerns how qualia are caused. There is also a problem concerning their effects. If qualia reduce to physical properties, then they have the effects of those properties. But what if qualia do not so reduce? How could they affect the physical world? Most agree that the physical world is causally closed, or nearly so; with the exception of some quantum indeterminacy, every physical event has a physical explanation. There would thus seem to be no room for qualia to do any independent causal work. Non-reductionism seems to imply epiphenomenalism, the view that qualia have physical causes but no physical effects. Many find that consequence unpalatable, and some consider it a strong argument for reductionism.

Some non-reductionists respond by defending epiphenomenalism, but more try to block the inference. One way to do that is to reject the assumption that the physical world is causally closed. Some interactionist dualists (e.g. Eccles, 1986) argue that qualia affect brain processes by filling in gaps resulting from quantum indeterminacy. But few contemporary philosophers or scientists accept interactionist dualism.

Many non-reductionists accept the causal closure of the physical, and argue that qualia nevertheless have physical effects. That strategy has been pursued in various ways. For example, some claim that certain physical events are causally overdetermined: that those events have both phenomenal and physical causes. By causal closure, any physical event has a sufficient physical cause. Therefore, any phenomenal cause it might also have would be causally redundant. That is an odd result, which some find as unacceptable as epiphenomenalism. But most non-reductionist strategies for avoiding epiphenomenalism involve substantive, sometimes surprising, views about causation.

Consider a second example. Chalmers describes a view, once proposed by Bertrand Russell (1927), on which the causal powers of qualia derive from intrinsic properties of the physical world. Physical theory characterizes its basic entities relationally. Basic particles, for example, are described in terms of how they interact with other particles and forces. Perhaps fundamental physical entities have intrinsic properties, which ultimately account for their relational properties. If so, then those intrinsic properties might be phenomenal properties. Or perhaps they are protophenomenal properties, from which both physical and phenomenal properties are constructed. Those ideas may sound strange, but either of them would explain how qualia could have physical effects.

Evolution

How did qualia come to exist? Many suspect that they provide organisms with evolutionary advantages. One hypothesis is that they supply an efficient way for organisms to acquire information about their bodies and environments. For example, pain qualia might help creatures capable of locomotion to avoid bodily damage, and olfactory qualia might help them distinguish nutritious food from poison.

Such hypotheses can be instructive (see below), but the extent to which they explain the origins of qualia is controversial. Phenomenal zombies, who lack qualia, have exactly the same informational sensitivities as their conscious counterparts. If phenomenal zombies are possible, then natural selection alone cannot explain why conscious creatures rather than phenomenal zombies evolved. Further principles may therefore be required to explain why qualia exist (Chalmers, 1996). Additionally, until more is understood about qualia and the brain, many will regard speculation on the evolutionary benefits of qualia as premature.

QUALIA AND COGNITIVE SCIENCE

Cognitive science often concerns qualia, or at least subjective reports of qualia, in some way. Consider three examples. First, studies of the splitting of auditory attention tend to rely on first-person reports of what the subject consciously experiences.

Second, consider the search for the neural correlates of consciousness. Some of that research concerns the neural correlates of the general state of being conscious, in a sense of 'conscious' that implies having qualia. And many studies concern the neural correlates of specific qualia. For example, experiments on rhesus macaques, trained to give bar-press reports, indicate strong correlations between specific kinds of visual sensation and activity in the inferior temporal cortex (Logothetis and Schall, 1989).

A third example is psychophysics. The Weber–Fechner law and Stevens's power law, typical results in the field, relate the intensity of percepts, or qualia, to the intensity of corresponding physical stimuli (e.g. luminance). Or consider studies of sensory illusions such as those produced by the Kanisza square, depicted in Figure 1. Normal subjects report seeing a square in the middle of the diagram, with a border as real as if it had been inscribed in ink. They also report perceiving the interior of the square as slightly brighter than the background, although there is no corresponding difference on the printed page. Many regard qualia such as those associated with such illusions as constituting a significant part of what psychophysics seeks to explain.

Studying qualia scientifically presents methodological difficulties. In general, we have only indirect access to a subject's conscious states. We have direct access to our own qualia, but introspective investigation also has well-known problems, which plagued the introspectionist tradition of the late nineteenth and early twentieth century. Further, first-person reports tend to employ coarse-grained and imprecise language such as 'an image of horizontal lines' or 'a high-pitched tone'. Substantial progress will presumably require developing more precise forms of expression. There may be principled limitations to any such endeavor; some claim, for example, that qualia are ineffable. But it may be possible to devise precise languages that capture at least the structural features of qualia. Indeed, there have been attempts along those lines, such as the quantitative techniques used in psychophysics.

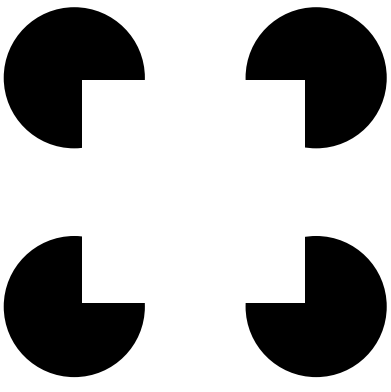


Figure 1. Kanisza square.

Many of the difficulties for the scientific investigation of qualia concern the phenomenal character of particular experiences, rather than the existence of qualia. Therefore, studying whether there is something it is like to be a bat is in certain respects more tractable than studying what it is like to be a bat (Allen and Bekoff, 1997). Here evolutionary hypotheses can be of use. For example, if we assume qualia evolved to allow adaptively flexible behavior, we may infer that adaptively flexible behavior provides evidence of qualia. Philosophical views can also help determine criteria for attributing qualia to other creatures. For example, functionalists might be more inclined than others to base attributions on the presence of certain functional properties.

KNOWLEDGE OF QUALIA

We have direct, first-person knowledge of our own qualia. That much is relatively uncontroversial. But what that knowledge consists in is unclear. Most accept that one can know about one's qualia noninferentially. For example, one need not infer one's qualia from one's behavior. Beyond that, opinions differ. In the early twentieth century, the heyday of sense-data theories, particularly strong epistemic claims about qualia were common. For example, three sentences after C. I. Lewis coins the term 'qualia' he writes: 'The quale is … not the subject of any possible error …' (1929, p. 121). One seldom finds such unqualified claims in contemporary philosophy, though philosophers continue to discuss attenuated variants.

Philosophical arguments sometimes rely on appeals to first-person knowledge of qualia. The knowledge argument (see above) provides one clear example. The literature on representationalism provides another. Some representationalists claim that experience is diaphanous: that when you attend introspectively to your experience of the color patch you perceive, you 'see right through' your qualia to the features of the patch itself, such as blueness and roundness. And that, they argue, suggests that qualia are just representational properties such as representing blue and representing roundness. But the legitimacy of such appeals to introspection is much disputed.

References

Allen C and Bekoff M (1997) *Species of Mind*. Cambridge, MA: MIT Press.

Alter T (2001) Know-how, ability, and the ability hypothesis. *Theoria* **67**(3): 229–239.

Block N (1978) Troubles with functionalism. In: Savage CW (ed.) *Perception and Cognition: Issues in the foundation of Psychology. Minnesota Studies in the Philosophy of Science*, vol. 9, pp. 261–325. Minneapolis, MN: University of Minnesota Press. [Reprinted in Block N (ed.) (1980) *Readings in the Philosophy of Psychology*, vol. 1. Cambridge, MA: Harvard University Press.]

Block N (1990) Inverted Earth. *Philosophical Perspectives* **4**: 53–79.

Chalmers DJ (1996) *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Davison D (1987) Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* **60**: 441–458. [Reprinted in Cassam Q (1994) *Self-knowledge*, pp. 43–64. New York, NY: Oxford University Press.]

Dennett DC (1988) Quining qualia. In: Marcel A and Bisiach E (eds) *Consciousness in Contemporary Science*, pp. 42–77. New York: Oxford University Press. [Reprinted in Lycan WG (1990) *Mind and Cognition: A Reader*, pp. 519–47. Cambridge, MA: Blackwell.]

Dennett DC (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

Eccles JC (1986) Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London* **B227**: 411–438.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Kripke S (1972) Naming and necessity. In: Harman G and Davidson D (eds) *Semantics of Natural Language*, pp. 253–355 and 763–9. Dordrecht: D. Reidel. [Reprinted as *Naming and Necessity* (1980). Cambridge, MA: Harvard University Press.]

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Lewis CI (1929) *Mind and the World Order*. New York: Charles Scribner's Sons.

Lewis D (1988) What experience teaches. In: *Proceedings of the Russellian Society*. Sydney, Australia: University of Sydney. [Reprinted in Lycan WG (1990) *Mind and Cognition: A Reader*, pp. 499–518. Cambridge, MA: Blackwell.]

Loar B (1990) Phenomenal states. In Tomberlin J (ed.) *Philosophical Perspectives IV: Action Theory and Philosophy of Mind*, pp. 81–108. Atascadero, CA: Ridgeview.

Locke J (1690) *An Essay Concerning Human Understanding*, edited by PH Nidditch (1975). Oxford, UK: Oxford University Press.

Logothetis N and Schall JD (1989) Neuronal correlates of subjective visual perception. *Science* **245**: 761–763.

McGinn C (1989) Can we solve the mind–body problem? *Mind* **98**: 349–366.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **4**: 435–450

Nagel T (1998) Conceiving the impossible and the mind–body problem. *Philosophy* **73**: 337–352.

Nida-Rumelin M (1996) Pseudonormal vision: an actual case of qualia inversion? *Philosophical Studies* **82**: 145–157.

Russell B (1927) *The Analysis of Matter*. London: Kegan Paul.

Searle JR (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Shoemaker S (1975) Functionalism and qualia. *Philosophical Studies* **27**: 291–315.

Tye M (1995) *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Further Reading

Block N, Flanagan O and Guzeldere G (eds) (1997) *The Nature of Consciousness: Philosophical Debates*. New York: Oxford University Press.

Churchland PM(1995) *The Engine of Reason, the Seat of the Soul*. Cambridge, MA: MIT Press.

Dretske FI (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Flanagan O (1992) *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Horst S (1995) *Phenomenology and psychophysics*. Manuscript, Wesleyan University: [http://shorst.web.wesleyan.edu/ papers/shorst.psychophysics.spp97.html.]

Kind AL (2001) Qualia realism. *Philosophical Studies* **104**: 143–162.

Levine J (2000) *Purple Haze: The Puzzle of Consciousness*. New York: Oxford University Press.

Lycan WG (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press.

Metzinger T (ed.) (2000) *Neural Correlates of Consciousness: Empirical and Conceptual Issues*. Cambridge, MA: MIT Press.

Savage CW (1970) *The Measurement of Sensation*. Berkeley, CA: University of California Press.

Siewert CP (1998) *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.

Shear J (ed.) (1995–7) *Explaining Consciousness – The Hard Problem*. Cambridge, MA: MIT Press.

Stevens SS (1975/1986) *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New Brunswick, NJ: Transaction Books.

Rationality

Intermediate article

Richard Samuels, King's College London, London, UK
Stephen Stich, Rutgers University, New Brunswick, New Jersey, USA

CONTENTS

Rationality is a normative notion used to evaluate beliefs, inferences, decisions, goals, actions, and people.

INTRODUCTION

One of the most fundamental questions in philosophy concerns the *nature* of rationality: what conditions must be met for an inference or a decision to be rational? Disagreements over this issue have been a perennial theme in philosophy, but they also underlie many of the disputes about rationality in contemporary cognitive science. Another issue with a long philosophical history concerns the *extent* to which humans are rational. Aristotle famously held that rationality is an essential property of human beings, but other theorists have been less sanguine. Starting in the 1960s, a number of investigators have conducted experiments designed to explore how well the reasoning and decision-making abilities of ordinary human subjects accord with what the investigators take to be appropriate normative standards. The results in some of these experiments have been interpreted as showing that, on many sorts of problems, people exploit simple heuristics that violate normative principles and sometimes lead to solutions that are quite irrational. Moreover, some authors have gone on to suggest the rather pessimistic hypothesis that people use these heuristics because they have nothing better available. Because this work focuses on the use of *heuristics* that can lead to mistaken or *biased* results, it is often described as being in the 'heuristics and biases' tradition. In the next section, we sketch some of the more important experimental findings that have emerged from this tradition.

The heuristics and biases program, with its associated view of human rationality, has had a profound impact on cognitive science and related fields. It has, however, been challenged on both empirical and philosophical grounds. We will outline two important empirical challenges to the pessimistic view that, on many sorts of problems, people lack the cognitive competence to reason and decide rationally. One of these challenges comes from investigators influenced by evolutionary psychology; the other comes from investigators who have been looking at individual differences in performance on reasoning tasks.

While much of the contemporary debate over human rationality turns on empirical questions about the sorts of reasoning strategies that people possess, philosophical issues about the nature of rationality and the norms we ought to adopt in evaluating cognition, have also played an important role. These are the issues on which we will focus in the final section. One prominent view in this area is what Edward Stein has called the *Standard Picture*:

> According to this picture, to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth. If the standard picture of reasoning is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with. (Stein, 1996, p. 4)

Although the Standard Picture is widely accepted, we will see that it is far from clear how to apply it to specific cases. We will also review some of the considerations that have recently led a number of writers to challenge the Standard Picture's account of rationality.

EMPIRICAL EVIDENCE ON HUMAN RATIONALITY

The Selection Task

In 1966, Peter Wason published a highly influential study of a cluster of reasoning problems that

became known as the *selection task*. Figure 1 illustrates a typical example of a selection task problem.

Wason found that subjects typically perform very poorly on questions like this. Most subjects respond correctly that the E card must be turned over, but many also judge that the 5 card must be turned over, despite the fact that the 5 card could not falsify the claim no matter what is on the other side. Also, a majority of subjects judge that the 4 card need *not* be turned over, though without doing so there is no way of knowing whether it has a vowel on the other side. It is not the case that subjects do poorly on all selection task problems, however. A wide range of variations on the basic pattern has been tried, and on some versions a much larger percentage of subjects answer correctly. Figure 2 is an example from Griggs and Cox (1982): These results form a bewildering pattern, since there is no obvious feature or cluster of features that separates versions on which subjects do well from those on which they do poorly.

The Conjunction Fallacy

Much of the experimental literature on reasoning has focused on problems that require probabilistic judgment. Among the best-known experiments of this kind are those that involve so-called *conjunction problems*. In one quite famous experiment, Tversky and Kahneman (1982) presented subjects with the following task.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
>
> (a) Linda is a teacher in elementary school.
> (b) Linda works in a bookstore and takes Yoga classes.
> (c) Linda is active in the feminist movement.
> (d) Linda is a psychiatric social worker.
> (e) Linda is a member of the League of Women Voters.
> (f) Linda is a bank teller.
> (g) Linda is an insurance sales person.
> (h) Linda is a bank teller and is active in the feminist movement.

In a group of naive subjects with no background in probability and statistics, 89 per cent judged that statement (h) was more probable than statement (f) despite the fact that one cannot be a *feminist* bank teller unless one is a *bank teller*. When the same question was presented to sophisticated subjects – graduate students in the decision science program of the Stanford Business School – 85 per cent gave the same answer! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Kahneman and Tversky's pioneering studies. This pattern of reasoning has been labeled *the conjunction fallacy*.

Base Rate Neglect

Another well-known group of studies explores the way in which people use base-rate information –

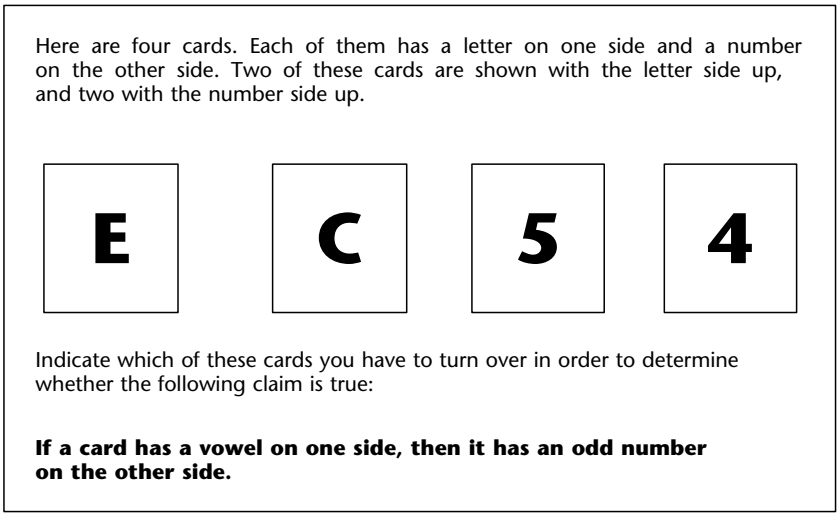


Figure 1.

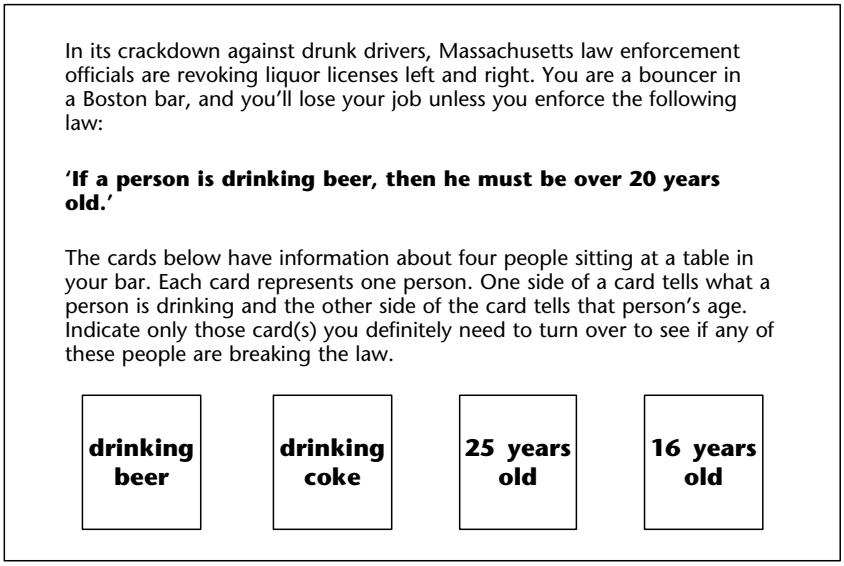


Figure 2.

roughly, background information about the proportion of a population that have a given characteristic – in making probabilistic judgments. According to the familiar Bayesian account, the probability of a hypothesis on a given body of evidence depends, in part, on the prior probability of the hypothesis. However, in a series of influential experiments, Kahneman and Tversky (1973) showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half the subjects with the following 'cover story'.

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The other half of the subjects were presented with the same text, except the 'base rates' were reversed. They were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description was constructed to be totally uninformative with regard to Dick's profession. Thus, the only useful information that subjects had was the base-rate information that the test was administered to 30 engineers and 70 lawyers. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50 per cent.

One further example of base rate neglect will illustrate the way in which the phenomenon might well have serious practical consequences. Here is a problem that was presented to a group

of faculty, staff, and fourth-year students at Harvard Medical School.

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%

Under the most plausible interpretation of the problem, the correct Bayesian answer is 2 per cent. But only 18 per cent of the Harvard audience gave an answer close to 2 per cent. Forty-five per cent of this distinguished group completely ignored the base-rate information and said that the answer was 95 per cent.

Framing

In a study that is widely interpreted as illustrating a deeply irrational feature of human decision-making, Tversky and Kahneman (1981) presented a group of subjects with the following problem:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

On the first version of the problem most subjects chose Program A. But on the second version most chose Program D, despite the fact that the outcome described in A is *identical* to the one described in C. The disconcerting implication of this study is that the decisions we make are strongly influenced by the manner in which the options are described or *framed*.

A Pessimistic View of the Experimental Results

The results we've sketched are a small sample of the enormous literature on human reasoning and decision-making in the heuristics and biases tradition. (For detailed surveys see Nisbett and Ross, 1980; Kahneman *et al*., 1982; Piattelli-Palmarini, 1994; Baron, 2001.) What conclusions about human rationality should we draw from these findings? Though opinions are sharply divided, a cluster of claims often associated with the heuristics and biases tradition suggest a *pessimistic* view about the extent of human rationality. In one often-quoted comment, for example, two leading investigators claimed that the experimental results have 'bleak implications' for the rationality of ordinary people (Nisbett and Borgida, 1975). Other researchers have concluded that individuals are generally affected by 'systematic deviations from rationality' (Bazerman and Neale, 1986). And still others have suggested that the fault may be in our genes, or at least in our evolutionary history: 'people lack the correct programs for many important judgmental tasks … [because we] … have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty' (Slovic *et al.*, 1976). But many other theorists have argued, on both empirical and philosophical grounds, that conclusions like these are unwarranted. In the remaining sections of this article, we consider briefly a number of such challenges. (For a more extensive review of the debate see Samuels *et al*., 2002a.)

EMPIRICAL CHALLENGES TO THE PESSIMISTIC VIEW OF HUMAN RATIONALITY

A Challenge from Evolutionary Psychology

One very influential challenge to the pessimistic interpretation of the heuristics and biases experiments has come from evolutionary psychologists who maintain that the human mind contains a number of specialized reasoning mechanisms that function in a normatively appropriate manner when presented with the sorts of problems that confronted our evolutionary forebears. Some of these theorists propose that the human mind contains a specialized reasoning mechanism designed by natural selection to solve problems in which probabilistic information is explicitly represented as frequencies. The ability to use probabilistic information effectively, they argue, would have been highly advantageous to our hominid forebears, and thus it would be surprising if we had not evolved mental mechanisms capable of using

this information rationally. But we should expect this capacity to manifest itself most clearly when probabilistic information is presented in a format that would have been common in ancestral environments. In those environments, probabilistic information would almost invariably have been presented in the form of information about frequencies, as opposed to single events. Thus we should expect people to do much better on probabilistic reasoning tasks if the problems are presented in a *frequency format*. And it appears that they do. In the 'feminist bank teller' problem, for example, if the description of Linda is followed by a question like this one:

> There are 100 people who fit the description above. How many of them are:
> …
> (f) bank tellers?
> …
> (h) bank tellers and active in the feminist movement?
> …

the number of subjects who commit the conjunction fallacy drops from over 90 per cent to only about 10 per cent.

Further evidence comes from Cosmides and Tooby's (1996) systematic exploration of the 'Harvard Medical School problem', in which they showed that subjects find it much easier to use base-rate information rationally on 'frequentist' versions of the problem, like the one that follows.

> 1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.
>
> Imagine that we have assembled a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
>
> Given the information above:
> on average,
> How many people who test positive for the disease will *actually* have the disease? _____ out of _____.

On this version of the problem, 76 per cent of subjects gave the correct Bayesian answer. Although these results are impressive, the hypothesis that we have evolved mental mechanisms designed to reason rationally about probability when problems are presented in a frequency format remains highly controversial and a number of competing hypotheses exist (Kahneman and Tversky, 1996).

Evolutionary psychologists have also argued that it would have been important for our forebears to have evolved cognitive mechanisms capable of subserving reciprocal exchanges, and that stable reciprocal exchange relationships require the capacity to detect *cheaters* – people who accept the benefits without paying the associated costs. A number of researchers have gone on to argue that the puzzling pattern of successes and failures in selection task problems can be explained by the hypothesis that the problems on which subjects do well are just the ones in which these reciprocal exchange and cheater detection capacities are deployed (Cosmides and Tooby, 1992; Gigerenzer and Hug, 1992). This hypothesis remains very controversial, however, and other authors have offered a variety of alternative explanations (Cheng and Holyoak, 1985; Oaksford and Chater, 1994; Manktelow and Over, 1995; Sperber *et al*., 1995).

A Challenge from the Study of Individual Differences

A different sort of challenge to the pessimistic interpretation of the heuristics and biases experiments comes from the study of individual differences. Stanovich and his colleagues (Stanovich, 1999) have shown that while the average performance on many heuristics and biases problems is quite poor, *some* subjects give the answer that the Standard Picture suggests is normatively correct on *many* of these problems. Moreover, subjects who are successful on these problems also get better scores on widely used measures of cognitive ability such as the Scholastic Aptitude Test (SAT). These researchers have also found significant correlations between success on reasoning tasks and various measures of 'cognitive style' and 'epistemic self regulation' which aim at assessing such things as how open-minded people are, how reflective they are, and how willing they are to consider evidence that contradicts their beliefs.

In light of these findings, a number of researchers have proposed *dual processing* theories which maintain that reasoning and decision-making are subserved by two quite different sorts of systems. One system is fast, holistic, automatic, largely unconscious, and requires relatively little cognitive capacity. The other is relatively slow, rule-based, more readily controlled, and requires significantly more cognitive capacity. Stanovich speculates that the former system is largely innate and that, as evolutionary psychologists suggest, it has been shaped by natural selection to do a good job on

problems like those that would have been important to our hominid forebears. The latter system, by contrast, is more heavily influenced by culture and formal education, and is more adept at dealing with many of the problems posed by a modern, technologically advanced, and highly bureaucratized society. Stanovich also argues that much of the individual variation seen in heuristics and biases tasks can be explained by differences in cognitive capacity (more of which is required for the second system), and by differences in cognitive style which lead to different levels of inclination to *use* the second system.

If Stanovich and the evolutionary psychologists are right, then the more pessimistic interpretations of the heuristics and biases experiments, according to which people typically do not have the correct 'programs' for many important reasoning tasks, are unwarranted, since everyone has the capacity to deal rationally with those reasoning problems that were important in the environment in which we evolved, and some of us also have considerable capacity to deal with a much wider range of problems. The extent to which education can improve the capacity to handle the kinds of problems that are important in a technological society, is, however, very much an open question.

PHILOSOPHICAL AND CONCEPTUAL ISSUES

On the Application and Interpretation of Standard Picture Norms

Though many investigators accept the Standard Picture account of rationality, a number of writers have challenged the way it has been applied to problems like those used in heuristics and biases experiments. So, for example, Gigerenzer (2000) and others have argued that in order to conclude that people are violating Bayesian normative principles, and thus being irrational, in the base rate neglect experiments, one must assume that the prior probability assignments which subjects make are identical to the base rates specified by the experimenters. But as Koehler (1996) observes: 'this assumption may not be reasonable in either the laboratory or the real world. Because they refer to subjective states of belief, prior probabilities may be influenced by base rates and any other information available to the decision-maker prior to the presentation of additional evidence. Thus, prior probabilities may be informed by base rates, but they need not be the same.' If this is right, and we think it is, then it is a genuine empirical possibility that some subjects are not violating Bayes' rule in these experiments but are merely assigning different prior probabilities from those that the experimenters expect.

In a more radical vein, Gigerenzer has noted that according to the influential 'frequentist' interpretation of probability theory, probabilistic statements make no sense unless they are relativized to a specific reference class, and thus claims about the probability of a single event are literally meaningless. Since many of the questions posed in heuristics and biases experiments ask subjects to judge the probabilities of single events (such as the probability that Linda is a bank teller), for a frequentist, no answer could possibly violate probability theory. Frequentism is a hotly contested view, but even if we grant that frequentism is correct, this argument is subject to serious objections. If statements about the probabilities of single events really are meaningless and hence do not violate the probability calculus, subjects are still guilty of making some sort of error in many heuristics and biases experiments. For if the questions are meaningless, then surely the correct response to a problem about the probability of a single event is not some numerical value or rank ordering, but rather: 'Huh?' or 'That's utter nonsense!' or 'What on earth are you talking about?' (For a more extended discussion of this point, see Samuels *et al*., 2002a.)

Objections to the Standard Picture

According to the Standard Picture, *what it is* to be rational – what is *constitutive* of good reasoning – is to reason in accord with rules or principles derived from formal theories, and where we fail to reason in this manner our cognitive processes are, at least to that extent, irrational. This view of rationality is sometimes called a *deontological* account. However, some philosophers have argued that a sharp distinction needs to be drawn between the principles governing *reasoning* and the formal principles of logic and probability theory. Harman (1986), for example, suggests that logic and probability theory are not directly relevant to reasoning at all.

Deontology is not the only conception of rationality that one might endorse. Another prominent view, called *consequentialism*, maintains that *what it is* to reason correctly is to reason in such a way that you are likely to attain certain goals or outcomes. Though the application of rules of reasoning may be a *means* to the attainment of certain ends, what is *constitutive* of being a rational reasoning process, on this view, is being an effective means of achieving some goal or range of goals. According to one

well-known form of consequentialism – *reliabilism* – a good reasoning process is one that tends to lead to true beliefs and the avoidance of false ones. Another form of consequentialism – sometimes called *pragmatism* – maintains that what it is for a reasoning process to be a good one is for it to be an efficient means of attaining the pragmatic objective of satisfying one's personal goals and desires. One reason for adopting a consequentialist rather than a deontological account of rationality is that it explains why it is worth worrying about rationality – why reasoning in a normatively correct fashion *matters.* If the deontological conception of rationality is rejected, however, then it is far from clear whether *any* of the experimental results in the heuristics and biases tradition really have 'bleak implications' for human rationality, since it is not clear which responses are rational on the consequentialist interpretation of rationality.

Another consideration that has led some to reject the Standard Picture is the principle that 'ought implies can'. This principle maintains that, just as in ethical matters where our obligations are constrained by what we can do, so too in the domain of reasoning and decision-making we are not obliged to satisfy standards that are beyond our capacities. But the Standard Picture requires us to perform reasoning tasks that are far beyond our abilities. For instance, it seems to be a principle of the Standard Picture that we ought to preserve the truth-functional consistency of our beliefs. Yet as Cherniak (1986) and others have argued, given even a conservative estimate of the number of beliefs we possess, this is a computationally intractable task – one that we *cannot* perform. Similar arguments have been developed against the claim, often associated with the Standard Picture, that we ought to revise our beliefs in such a way as to ensure *probabilistic coherence*. Here too considerations of computational complexity strongly suggest that we cannot satisfy this standard. And if we cannot satisfy the norms of the Standard Picture, then given that ought implies can, it follows that the Standard Picture is not the correct account of the norms of rationality.

References

Baron J (2001) *Thinking and Deciding*, 3rd edn. Cambridge, UK: Cambridge University Press.

Bazerman M and Neale M (1986) Heuristics in negotiation. In: Arkes H and Hammond K (eds) *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge, UK: Cambridge University Press.

Cheng P and Holyoak K (1985) Pragmatic reasoning schemas. *Cognitive Psychology* **17**: 391–416.

Cherniak C (1986) *Minimal Rationality*. Cambridge, MA: MIT Press.

Cosmides L and Tooby J (1992) Cognitive adaptations for social exchange. In: Barkow J, Cosmides L and Tooby J (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford, UK: Oxford University Press.

Cosmides L and Tooby J (1996) Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **58**(1): 1–73.

Gigerenzer G (2000) *Adaptive Thinking: Rationality in the Real World*. New York, NY: Oxford University Press.

Gigerenzer G and Hug K (1992) Domain-specific reasoning: social contracts, cheating and perspective change. *Cognition* **43**: 127–171.

Griggs R and Cox J (1982) The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology* **73**: 407–420.

Harman G (1986) *Change In View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge, UK: Cambridge University Press.

Kahneman D and Tversky A (1973) On the psychology of prediction. *Psychological Review* **80**: 237–251.

Kahneman D and Tversky A (1996) On the reality of cognitive illusions: a reply to Gigerenzer's critique. *Psychological Review* **103**: 582–591.

Koehler J (1996) The base-rate fallacy reconsidered. *Behavioral and Brain Sciences* **19**: 1–53.

Manktelow K and Over D (1995) Deontic reasoning. In: Newstead S and Evans J St B (eds) *Perspectives on Thinking and Reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nisbett R and Borgida E (1975) Attribution and the social psychology of prediction. *Journal of Personality and Social Psychology* **32**: 932–943.

Nisbett R and Ross L (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford M and Chater N (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* **101**: 608–631.

Piattelli-Palmarini M (1994) *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York, NY: John Wiley & Sons.

Samuels R, Stich S and Faucher L (2002a) Reasoning and rationality. In: Niiniluoto I, Sintonen M and Wolenski J (eds) *Handbook of Epistemology*, pp. 1–50. Dordrecht: Kluwer.

Slovic P, Fischoff B and Lichtenstein S (1976) Cognitive processes and societal risk taking. In: Carol JS and Payne JW (eds) *Cognition and Social Behaviour*. Hillsdale, NJ: Erlbaum.

Sperber D, Cara F and Girotto V (1995) Relevance theory explains the selection task. *Cognition* **57**(1): 31–95.

Stanovich K (1999) *Who Is Rational?* Mahwah, NJ: Lawrence Erlbaum Associates.

Stein E (1996) *Without Good Reason*. Oxford, UK: Clarendon Press.
Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* **211**: 453–458.
Tversky A and Kahneman D (1982) Judgments of and by representativeness. In: Kahneman D, Slovic P and Tversky A (eds) *Judgment Under Uncertainty: Heuristics and Biases*, pp. 84–98. Cambridge, UK: Cambridge University Press.
Wason P (1966) Reasoning. In Foss B (ed.) *New Horizons in Psychology*. Harmonsworth, UK: Penguin.

Further Reading

Dawes R M (1988) *Rational Choice in an Uncertain World*. San Diego, CA: Harcourt.
Evans J and Over D (1996) *Rationality and Reasoning*. Hove, UK: Psychology Press.
Foley R (1987) *The Theory of Epistemic Rationality*. Cambridge, MA: Harvard University Press.
Goldman A (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.
Manktelow K and Over D (eds) (1993) *Rationality: Philosophical and Psychological Perspectives*. London: Routledge.
Nozick R (1993) *The Nature of Rationality*. Princeton, NJ: Princeton University Press.
Samuels R, Stich S and Bishop M (2002b) Ending the rationality wars: how to make disputes about human rationality disappear. In: Elio R (ed.) *Common Sense, Reasoning and Rationality*. New Directions in Cognitive Science, vol. 11. New York, NY: Oxford University Press.
Stich S (1990) *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Reduction

Intermediate article

Robert C Richardson, University of Cincinnati, Cincinnati, Ohio, USA

CONTENTS

Reduction is a relation between two theories, in which one captures or explains the other, usually with some gain in precision or scope. Reduction may involve theories at the same level of organization or at different levels of organization; and reduction may involve one theory explaining another theory or only its empirical domain. Empirical, historical, and methodological issues confront reductionism.

WHAT IS REDUCTION?

Reduction is an explanatory relation between two theories, models, or sciences in which one theory, model, or science captures or explains the other, with some gain in precision or scope. The former is the reducing theory and the latter is the reduced theory. Reduction may involve theories at the same level of organization or at different levels of organization; and reduction may involve one theory explaining another theory or explaining only the empirical phenomena captured by the reduced theory. Insofar as some range of empirical phenomena captured by one theory or model can be explained in terms of another theory or model, the empirical domain associated with the former is reduced within the latter reducing theory. Insofar as a theory or a model, and not merely the associated empirical phenomena, can be explained in terms of another, the reduced theory or model is reduced to the reducing theory or model. The reducing theory or model typically is more general or more exact: it captures a broader range of phenomena and/or explains them with greater precision. The empirical domain of the reduced theory becomes a subset of the empirical domain of the reducing theory. Increased generality or precision in the reducing theory creates an asymmetry between the two theories, giving priority to it over the reduced theory. The reducing theory explains more, or explains it more exactly, than the reduced theory. Reduction is thus an asymmetric relation between different models, theories, or sciences according priority to one, in which the explanatory potential of the reduced theory is captured within the reducing theory, and some explanatory purchase is gained by shifting to the reducing theory.

Examples can be drawn from a variety of domains, though all are controversial in one way or another. Chemical elements are ordered in a periodic table according to ratios in which they combine with other elements in chemical reactions. These chemical bonding properties are explicable in terms of the valence structures of atoms. Hydrogen has a valence of one and oxygen a valence of two, as reflected in the columns they occupy in the periodic table. They therefore combine in a ratio of two to one, resulting in molecules such as H_2O. J. J. Thomson explained the bonding ratios in terms of the number of electrons present in the outermost shells. During the early twentieth century, this was elaborated into the much more detailed theory of the covalent bond. This is a case in which a model of chemical interaction was explained in terms of a physical model at a lower level, concerned with atomic structure. In a similar way, cell biology depends on biochemistry. The behavior of organic systems is a function of constituent organ systems. Group behavior often needs to be understood in terms of local, individual, interactions. These are all commonly represented as cases of reduction at work in the sciences.

The most common paradigm within philosophy of science is the reduction of classical thermodynamics to statistical mechanics (Nagel, 1961; Bickle, 1998). Classical thermodynamics embodied a number of phenomenological laws describing the behavior of gases. The most important were Boyle's law and the law of Charles and Gay-Lussac. Boyle's law says that the product of the pressure and volume for a sample of gas is constant at a constant temperature. The law of Charles and Gay-Lussac says that the ratio of volume to temperature is

constant for a gas held at a constant pressure. The combined law is a straightforward consequence of these two principles. It says that the product of pressure and volume is a constant function of temperature. In a common form, with obvious abbreviations, this is the Boyle–Charles law: $PV = kT$.

Within statistical mechanics, we can derive an analogous result. Under a variety of simplifying assumptions, it is possible to show that the product of pressure and volume is a constant function of the aggregate kinetic energy of a gas. This is the Bernoulli formula: $PV = (2/3)e_t$.

The results bear obvious similarities so expressed. This motivates the conclusion that the thermal energy contained in a quantity of gas (T) is simply the aggregate kinetic energy of the particles constituting that gas (e_t), and that changes in the temperature of the gas are equivalent to changes in the aggregate kinetic energy of these particles. The Bernoulli formula allows us to explain what is explained by the Boyle–Charles law, and even to explain why the Boyle–Charles law holds as nearly as it does. It is also possible to explain why the Boyle–Charles law fails at extremes of temperature and density, though not in as direct a way as we can explain the law itself. The kinetic theory of gases is thus capable of explaining much of what is explained by classical thermodynamics, and can explain the behavior of gases with greater precision. Classical thermodynamics is therefore held to reduce to statistical mechanics.

Reduction can vary considerably in its goals. In some cases, the goal is overtly ontological, appealing to economy or parsimony; that is, the appeal of reduction is that it economizes on ontological commitments and eliminates any need to appeal to nonphysical entities such as vital forces, entelechies, or minds. For example, attempts to reduce number theory to mathematical logic at the beginning of the twentieth century were driven by the desire to undercut claims to any independent domain of mathematical entities, and thereby to synthetic a priori knowledge. In some cases, the goal of reduction is epistemological. For example, it was an important claim of logical positivism that all knowledge can be reduced to empirical knowledge. Early work, such as the work of Carnap on semantics, had this goal at the forefront. In yet other cases, the goal is methodological. It is sometimes easier to understand an effect if it is studied in isolation from other things that affect it. It is also sometimes easier to understand the behavior of a component in relative isolation from the system of which it is a part. The behavior of more complex entities can then, ideally, be understood in terms of component contributions (Bechtel and Richardson, 1993).

Reduction can also vary in form. First, we can distinguish between theoretical and empirical reductions. Theoretical reductions involve the explanation of one theory or model in terms of a second theory or model; by contrast, *empirical* reductions involve the explanation of the phenomena characteristic of one domain in terms of a second theory or model. In either case the predictions of one theory are captured by another, often more general, theory. In both cases, the reducing theory has explanatory priority because of its increased precision and generality. Second, we can distinguish between cases involving a single level of organization from those involving different levels of organization. There is no generally agreed way to distinguish levels of organization, though it is widely agreed that there is a useful distinction to be drawn; levels of organization can usefully be thought of, to a first approximation, as levels of aggregation, so that social groups constitute a level of organization higher than that of individuals, individuals a level of organization higher than that of organs, organs higher than that of cells, and so on (Wimsatt, 1976). Since these two distinctions are independent of one another, this leaves us with four general cases, as illustrated in Table 1.

In the first quadrant (on the upper left), the empirical phenomena captured by a theory or model at one level of organization are explained in terms of a theory or model at the same level. Geocentric models of the solar system explained with reasonable precision the observable behavior of stars and planets, including retrograde motion. Heliocentric models explained the same phenomena with quite different principles. The theories are nonetheless inconsistent, so one cannot be derived from the other. In the second quadrant (on the upper right), the phenomena captured by a theory or model at one level of organization are explained in terms of a theory or model at a deeper level of organization. Group selection was designed at one time to explain altruistic behavior, including cooperative behavior and the inclination sometimes shown to

Table 1. Four types of reduction

	Same level	*Different levels*
Empirical	Copernican astronomy and geocentric models	Group selection and kin (genic) selection
Theoretical	Newton mechanics and relativistic mechanics	Snell's Laws and electromagnetism

limit reproductive behavior to the apparent detriment of the individual. Models for kin selection and reciprocal altruism explained much of this apparently altruistic behavior in terms of selection acting below the level of the group. The behavior is explained at a lower level, without the benefits of group selection. Again, the theories are inconsistent, so kin selection explains the phenomena supposedly explained by group selection but does not reduce the theory of group selection. The third quadrant (at the lower left) includes explanations of theories or models in terms of theories or models at the same level of organization, generally with an increase in explanatory scope. Newtonian mechanics embodies laws that are limiting cases of relativistic mechanics, reasonable approximations at relatively low velocities. The laws characteristic of one theory are explained as limiting cases of laws characteristic of a later theory.

In the fourth quadrant (at the lower right), a theory or model at one level is explained in terms of a theory or model at a different level or organization. These are theoretical microreductions. They are the most common focus for discussions of reduction, and are often thought to be paradigmatic reductions, though there are few unproblematic examples. Snell's law describes the angles of light refraction in different media. Huygens showed that this could be explained in terms of a wave theory of light, and thus is an electromagnetic effect. Mendelian genetics provided a broad description of the patterns of transmission of genes. Some aspects of this theory were subsequently explained in terms of molecular mechanisms: for example, Mendelism during the first decades of the twentieth century involved linkage between genes and understood it as the probability of joint transmission. Molecular genetics gives us a mechanism for understanding linkage and why linkage is a function of distance between the genes coding for various characters. The case already outlined of thermodynamics and statistical mechanics is generally thought to be of this sort, though that characterization is problematic. It seems more likely that this is a theoretical reduction at the same level, explaining the macroscopic behavior of gases in terms of the statistical behavior of aggregates.

Any of the four reductive categories may be more or less conservative. The empirical phenomena within the scope of a reduced theory or its theoretical principles may be conserved to varying degrees. It is now common to treat the theoretical cases as ranging from cases of complete elimination to cases of nearly complete retention (Churchland, 1979; Bickle, 1998). At one extreme, Snell's law of refraction is almost wholly explained by Maxwell's equations describing electrical and magnetic fields, given suitable boundary conditions. Newtonian mechanics is at least approximately true, while phlogiston is not appealed to at all within modern chemistry. Empirical phenomena also are conserved to varying degrees. Conservatism is often high. Ptolemaic, Galilean, and Newtonian theories all needed to explain planetary motions. Some phenomena may be saved only to a limited degree. Independent assortment was an important part of early Mendelism, though it is a special case within molecular genetics. Dominance, likewise, was important for early Mendelism and is a limiting case for molecular genetics. At the other extreme, retention is minimal. Autonomy, or modularity, of development was an important part of early Mendelism, only to be abandoned subsequently in a wide range of cases; in fact, developmental phenomena that had been regarded as important were largely disregarded within the evolutionary synthesis for the first half of the twentieth century.

HISTORY

Empirical reduction has a long history within empiricism. There is a long tradition emphasizing the foundational nature of empirical knowledge, from Locke and Hume to Carnap. This involved reducing all knowledge to experiential knowledge, and all concepts to empirical concepts. British empiricists claimed that all ideas have their origins in the senses or impressions. Early in the twentieth century, this often meant showing that all concepts could be resolved as complexes of empirical concepts. The construction of more complex concepts based on simpler empirical ones was intended to support a radical empiricism. It was soon recognized that the project could not succeed, and this epistemological reductionism generally has been abandoned.

Ontological or methodological motivations for reduction have been more prominent in recent discussions. The most influential analysis is the synthesis by Ernst Nagel in *The Structure of Science* (1961), though many of the general ideas also can be found in other sources at about the same time. Nagel's account assumes that theories can be captured as axiomatic formal systems, with some finite number of primitive predicates expressing theoretical terms. Nagel then imposes two general conditions on reductions: (1) all the primitive terms appearing in the secondary (reduced) theory either appear in the primary (reducing) theory, or are

associated with terms in the primary theory by suitable 'reduction functions'; and (2) all the fundamental 'laws' of the secondary science can be explained from the primary science together with the 'bridge laws' given in the first condition. Many people assume the reduction functions under (1) are assumed to express identities, though that is not part of Nagel's account and is not typical of actual reductions. These two conditions have been modified in a number of ways. K. Schaffner (1967) offered a more general model according to which reduction requires only that what can be derived within the reducing science are laws 'strongly analogous' to those of the reduced science, thereby allowing for significant corrections to the reduced science. Presumably, the principles derived would be more accurate than those they supplant. More recently, others, including P. M. Churchland (1979) and J. Bickle (1998), have embraced similarly weakened conditions, requiring the reducing theory to incorporate principles analogous, or 'relevantly isomorphic', under limiting conditions to the principles in the reduced theory; insofar as the analogy fails, reduction fails and elimination is the result. (*See* **Eliminativism**)

The example of thermodynamics and statistical mechanics fits these weakened expectations reasonably well. The Bernoulli formula is formally analogous to the original Boyle–Charles law, at least in the form above, and the identification of heat with mean molecular motion serves as the relevant 'bridge law'. An equally common example, though involving intralevel reduction, comes from the special theory of relativity (STR). Newton's laws of motion hold not only from a standpoint of rest but for any inertial frames. Only relative motion is observable. The STR generalized and replaced the Galilean transformations to accommodate electromagnetic phenomena, and consequently required a reformulation of the laws of mechanics. Within this reformulation the values obtained from Newton's laws are approximately correct for velocities much below that of the speed of light. What is obtained within the STR is not precisely what can be derived within Newtonian mechanics, but it is a reasonable approximation under limiting assumptions and is arguably 'analogous' or 'relevantly isomorphic'.

One unresolved question confronting the weaker and more generalized models of reduction is whether they are not too generous. There is an 'analogy' to be had between the spread of rumors and the spread of diseases, as there is between economic growth and models within population genetics. Since these are not plausibly cases of reduction, some further constraints seem necessary. There is no broad consensus on additional constraints.

ARGUMENTS FOR REDUCTION

There are a variety of arguments for the importance of reduction. One general appeal is theoretical: nature appears to be hierarchically organized, and if our theories are somehow to reflect the structure we find in nature they too must be hierarchically organized. There are various levels of organization. Social groups consist of individuals. Individual organisms consist of a variety of organ systems, and those in turn of organs. These organs consist of specialized cells. This sort of decomposition can be continued to simpler and simpler entities until we reach fundamental particles. The order can be reversed. If we begin with a theory adequate for entities at any one level, then the behavior of more complex objects consisting of the simpler entities must be a function of the behavior of the simpler entities. If entities at higher levels are composed entirely of entities described at lower levels, and the behavior of these simpler entities can be adequately captured by theories tailored for those lower levels, then higher order behavior must depend only on the behavior of the entities at the lower level; and a theory adequate at the lower level should, in principle, be adequate for higher levels as well.

There is a general empirical appeal to reduction. The history of science shows the power of reduction, and progress in science often is the result of consolidation and unification, however these might be understood. Scientific progress is a matter of developing more and more general theories, and theories which unify more and more phenomena are preferable to those of narrower scope. Galilean mechanics found a more general and empirically more adequate formulation in Newton, just as Newtonian mechanics found a more general and empirically adequate formulation in Einstein. Similarly, the history of other sciences shows a succession of theories of increasing scope and comprehension, and when earlier theories were not wholly misguided, it is not surprising that the earlier theories are reasonable approximations to the truth. Likewise, history exhibits the development of theories of greater and greater scope. The recognition that electrical and magnetic forces are fundamentally the same was a triumph of reductionism, and the discovery of a unified theory encompassing electromagnetism and gravity would also be a reductionist triumph. We should therefore

expect that successful science leads to theories of increasing scope and power; theories at an earlier stage of development, or theories that are more specialized, should be reducible to those that are more developed and more fundamental.

There are also methodological appeals for reductionism. Much of the world is tremendously complex, and involves a variety of interactive processes. It is easier to understand a phenomenon we are studying if we can isolate it from other effects and causes, especially those with which it interacts (Bechtel and Richardson, 1993). If we want to predict the orbit of a planet, it is sufficient to know the several forces acting on it, though in some cases the analytical solutions may be prohibitively complex. If we can determine the effect an enzyme has in isolation from others in the protoplasm or the effect of a gene independently of other genes, then we should be able to use that information in explaining cellular behavior or organismic traits. If we can determine the effect of disabling a specific gene, then that gives us information concerning the role that gene plays in the larger cellular context. Observing deficits can provide the resources for understanding component function; and, correspondingly, component functions contribute to understanding systemic behavior. Reductionist methodology simplifies what would otherwise be an intractably complicated set of interactions, allowing an understanding of complex behavior in terms of individual components.

PROBLEMS FOR REDUCTION

There is a corresponding array of arguments against reduction. We can begin with the theoretical arguments. If nature is hierarchically organized, then our theories should reflect multiple levels of organization. Though there is no wholly satisfying account of what constitutes a level of organization, or therefore a level of explanation, there are clear reasons for recognizing hierarchical organization (Simon, 1969; Wimsatt, 1976). Hierarchies arise because the forces governing interactions between objects do not form an equally distributed continuum. The strongest forces govern interactions at the lowest level and give rise to reasonably stable units at a middle level. The forces responsible for atomic structure are an order of magnitude stronger than intermolecular forces, and those in turn are an order of magnitude stronger than those responsible for the three-dimensional structure of macromolecules. These lower-level forces may not determine which among a variety of complexes at the higher level will be realized. Moreover, there is often no explanatory gain in moving to a lower level and in some cases a substantial explanatory loss. In even the paradigm case for reduction, there would be no point in attempting to explain the behavior of a gas by turning to the behavior of individual molecules. In point of fact, we gain explanatory power by shifting to a higher level and considering the statistical properties of the aggregate without considering the details of structure.

The dominant arguments against reduction depend on multiplicity of realization and multiplicity of function. The thought is that Nagel's first condition on reduction above requires definitions of concepts at higher levels in terms of concepts at lower levels, and this minimally requires a one-to-one relation between levels. A one-to-many or many-to-one relation would violate this condition. The existence of one-to-many and many-to-one relations is widely thought to show that reduction is indefensible. That is, multiplicity of realization and multiplicity of function are both barriers to reduction. A gene, within Mendelian genetics, is a functional unit responsible for the production of phenotypic effects; it is also the case that a given phenotypic effect can be, and is, the effect of different molecular mechanisms (Wimsatt, 1976). The genetic code, to begin with, is redundant. Different molecular sequences code for the same set of amino acids. Different sequences of amino acids, furthermore, can be functionally equivalent, forming enzymes that catalyze the same reactions. The genetic code is also contingent. Though the code is universal, or nearly so, chemistry does not determine the code. The conclusion is that many molecular mechanisms result in the same amino acid sequence, many amino acid sequences constitute similar enzymes, and many enzymes catalyze the same reactions. This multiplicity of realization means that any bridge laws fall short of identities. This multiplicity of realization is common. There are, for example, isotopic variants of oxygen, differing only in the number of neutrons in the nucleus. They are indistinguishable except for a few reactions highly sensitive to atomic mass. Lacking the necessary identifications in the bridge laws, reductions cannot even get started. (*See* **Multiple Realizability**)

There are also historical problems for reduction. An argument nearly as pervasive as the more theoretical ones above is based on dramatic conceptual changes, and the resulting incommensurability (Kuhn, 1962; Feyerabend, 1962). Scientific change is often revolutionary. The ancients thought of planets as 'wanderers' moving with the heavenly

spheres. The spheres within which the planets were embedded were real, and circled about the Earth with definite periods. It was no small change to think of the planets as miniscule pieces of rock or gas orbiting about a minor star in the Milky Way, set no longer in a finite universe but in a largely empty space of vast proportions. The planets of the ancients are not our planets. Later changes are no less dramatic. Newton conceived of a universe infinite in space and time, with an absolute reference frame. Einstein changed our vision again. Within relativistic physics, mass is no longer conserved as it was for Newton, the universe is finite in extent, and space bends under gravitational distortions. There is, strictly, no concept in the theory of relativity capturing Newtonian mass, space, or time. The history shows radical conceptual change, rather than the reduction of one theory to another. Scientific change consists not in consolidation and unification but replacement and revolution.

From a methodological perspective, there also are flaws to reductionism. Though experimental manipulation can often isolate causes, the artificiality of the control can also induce artifacts. One experimental technique used to determine component function involves disabling or destroying some structure. The loss of function provides evidence concerning the functions of parts. A mutation in a gene, or a 'defective' gene, is often correlated with some loss of function, as, for example, a single gene is 'responsible for' the darkening of urine characteristic of alkaptonuria. The relatively simple change is induced by a recessive gene, and indicates a gene with a relatively specific catalytic effect. Likewise, the loss of a neural structure is often correlated with some loss of function, as in the classic results of Broca and Wernicke. The loss of a specific capacity indicates there is a neural 'organ' with a relatively specific cognitive function. However, this technique can also introduce artifacts. The behavior of an enzyme within a cell is not always the same as its behavior in a test tube. Genetic regulatory networks are known to be highly interactive: single genes typically have multiple effects, and multiple genes conspire to produce phenotypic effects. And as Hughlings-Jackson observed in the nineteenth century, loss of an organ does not explain any positive symptoms; what is important is how the person compensates for losses suffered. An analysis that ignores the systemic interactions is doomed to failure. Powerful computational techniques have only recently made it practical to deal with multiple variables simultaneously and to model multivariate systems. The interpretation of such results is controversial, but it is a method diverging significantly from reductionist methods. (*See* **Language and Brain**)

REDUCTION AND COGNITIVE SCIENCE

Within cognitive science, reduction has had an ambivalent status, for reasons analogous to those already outlined. From the mid-1950s through the mid-1980s, there was wide agreement concerning the general architecture of cognition, around a symbolic model or information processing accounts. Symbolic structures encoded representations within memory which were subject to various transformations. AI models were designed to simulate human intelligence, relying on rules and representations analogous to those characteristic of human cognition. These symbolic models constitute abstract models of cognition, with abstract representations. As such, they are compatible with varied physical realizations (Putnam, 1975; Fodor, 1975).

Multiplicity of physical realization has recently come under attack. However, plasticity of neural realization, and the variability of individual development, insure that there will be some differences in neural realization from person to person. Artificial intelligence promised even more radical prospects for differences in physical realization. Functional models abstract from physical details, offering an account of cognition which cuts across variability in physical realization. The abstractness promised a level of autonomy for cognitive psychology that moved it closer to work in AI and psycholinguistics than to neuroscience. Multiplicity of realization was thought to compromise any promise of reduction because it entailed that there would not be the required reduction functions to sustain a reduction of cognitive psychology to neuroscience; that is, it failed Nagel's conditions for reduction and especially the first condition. (*See* **Functionalism**; **Artificial Intelligence, Philosophy of**)

By the 1980s, theorists increasingly explored models of cognition that exploited alternative architectures. Neural networks, or connectionist/PDP models, yielded accounts of learning which differ significantly from symbolic models and promise a more realistic model of cognition. These models had a number of advantages, including enhanced flexibility. Mental representations came to be thought of as distributed rather than discrete. The appeal to distributed representations motivated some to reject cognitive models altogether, and many suggested that elimination rather than

reduction would be the fate of cognitive models (Churchland, 1979). (*See* **Connectionism**)

The emergence of cognitive neuroscience has since provided more support for a reductionist vision, fueled in large part by striking technological advances. Research on visual processing, for example, has revealed a number of processing pathways leading from the occipital lobe forward to temporal and parietal regions. Single cell responses show that cells respond preferentially to different types of information, a result confirmed in a wide range of cases. Coupled with lesion studies, and neuroimaging techniques (PET and fMRI), we are gaining an increasingly detailed understanding of visual processing. One of the classic results in neuropsychology is that language deficits follow from damage to regions anterior to the sylvan fissure known as Broca's area. Neuroimaging confirms the implication of Broca's area in speech production, and has led to a variety of more elaborate models involving other cortical areas. There are also now specific models of associative learning, based on a detailed knowledge of the structure and mechanisms in Aplysia. We also understand the involvement of the hippocampus in memory. It is far from clear how well any of this work fits traditional models of reduction, though it certainly does portend greater integration of neuroscience and psychology. (*See* **Neuroimaging**; **Golgi Staining**, **Aphasia**; **Hippocampus**)

References

Bechtel W and Richardson RC (1993) *Discovering Complexity*. Princeton, NJ: Princeton University Press.

Bickle J (1998) *Psychoneural Reduction: the New Wave*. Cambridge, MA: MIT Press

Churchland PM (1979) *Scientific Realism and the Plasticity of Mind*. Cambridge, UK: Cambridge University Press.

Feyerabend P (1962) Explanation, reduction and empiricism. In: Feigl H and Maxwell G (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, pp. 28–97. Minneapolis, MN: University of Minnesota Press.

Fodor JA (1975) *The Language of Thought*. New York, NY: Thomas Y. Crowell.

Kuhn TS (1962) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Nagel E (1961) *The Structure of Science*. New York, NY: Harcourt, Brace and World.

Putnam H (1975) The nature of mental states. In: Putnam H, *Mind, Language and Reality*, pp. 429–440. Cambridge, MA: Cambridge University Press.

Schaffner K (1967) Approaches to reduction. *Philosophy of Science* **34**: 137–147.

Simon H (1969) *The Sciences of the Artificial*, 2nd edn. Cambridge, MA: MIT Press.

Wimsatt WC (1976) Reductionism, levels of organization, and the mind–body problem. In: Globus G, Maxwell G and Savodnik I (eds) *Consciousness and the Brain*, pp. 205–267. New York, NY: Plenum Press.

Further Reading

Causey R (1977) *Unity of Science*. Dordrecht, Netherlands: D. Reidel.

Churchland PM and Churchland PS (1990) Intertheoretic reduction: a neuroscientist's field guide. *Seminars in the Neurosciences* **2**: 249–256.

Churchland PS (1986) *Neurophilosophy: Toward a Unified Theory of the Mind/Brain*. Cambridge, MA: MIT Press/ Bradford Books.

Hooker CA (1981) Towards a general theory of reduction. *Dialogue* **20**: 38–59, 201–236, 496–529.

Kemeny J and Oppenheim P (1967) On reduction. *Philosophical Studies* **19**: 6–17.

Kitcher P (1993) *The Advancement of Science*. Oxford, UK: Oxford University Press.

Oppenheim P and Putnam H (1958). Unity of science as a working hypothesis. In Feigl H, Scriven M and Maxwell G (eds) *Minnesota Studies in Philosophy of Science*, vol. II, pp. 3–36. Minneapolis, MN: University of Minnesota Press.

Richardson RC (1979) Functionalism and reductionism. *Philosophy of Science* **46**: 533–558.

Simon H (1969) *The Sciences of the Artificial*, 2nd edn. Cambridge, MA: MIT Press.

Reference, Theories of

Intermediate article

Michael Devitt, City University of New York, New York, USA

CONTENTS

Theories of reference are about the relations between representations and the world that are thought to determine (at least partly) the meaning or content of those representations.

WHAT IS A THEORY OF REFERENCE?

Referential relations hold between representations and the world: in particular, between parts of sentences – words – and the world and between parts of thoughts – concepts – and the world. The most obvious example of such a relation is the naming relation, the sort that holds between 'Winston Churchill' and the famous statesman. However, it is usual to think of reference as covering a range of semantically significant relations: for example, between the word 'witty' and wittiness, and between the concept 'bachelor' and all bachelors. These relations are variously described by the terms 'designate', 'denote', 'signify', 'apply', 'satisfy', 'instantiate', 'fall under', and 'about'.

Reference is important because it is thought to be the core of meaning and content. Thus, the fact that 'Winston Churchill' refers to that famous statesman is the core of its meaning, and hence of its contribution to the meaning of any sentence – for example, 'Winston Churchill is witty' – that contains it. And the fact that the concept 'bachelor' refers to all bachelors is the core of its content and hence of its contribution to the content of any thought – for example, the thought 'Winston Churchill is not a bachelor' – that contains it. The usual view is that meaning determines reference.

The first question that arises about the reference of a term is: what does the term refer to? Sometimes the answer seems obvious. For example, 'Winston Churchill' refers to the famous statesman. Even this apparently obvious answer is rejected by those who take words to refer to ideas in the mind or to mental representations. But the received view in contemporary philosophy is that words refer to the world. This leaves room for different opinions about the reference of predicates. Some take 'witty' to refer to the property of wittiness, some to the set of all witty things, and some to each witty thing separately.

The central question about reference is: by virtue of what does a representation have its reference? A theory of reference answers this question by explaining the relation of the representation to its referent. It has proved surprisingly hard to provide these explanations. There was a surge of interest in theories of reference in the twentieth century.

VARIETIES OF THEORIES OF REFERENCE

According to 'description' theories of reference, the reference of a representation is determined by certain descriptions – other representations – inferentially associated with it by competent speakers: these descriptions identify the referent. The simplest form of description theory, derived from the work of Frege (1893) and Russell (1912), specifies a set of descriptions each of which is necessary and all of which are sufficient for reference determination: for example, the references of 'adult', 'unmarried' and 'male' might be jointly sufficient and individually necessary for the reference of 'bachelor'. According to another form of description theory, the reference is whatever is picked out by a (weighted) majority of certain descriptions associated with the representation. On this 'cluster' theory, no one description is necessary for reference fixing (Searle, 1958).

Around 1970, several criticisms were made of description theories of proper names, such as 'Winston Churchill' (Kripke, 1980; Donnellan, 1972), and 'natural kind' words, such as 'gold' and 'tiger' (Kripke, 1980; Putnam, 1975). One important criticism is that the theories yield unwanted necessities. The description we inferentially associate with 'tiger' is along the lines of 'large carnivorous quadrupedal feline, tawny yellow in color with blackish transverse stripes and white belly'. So the objects that 'tiger' refers to must be four-legged and

striped. Yet tigers need not be thus: a tiger might lose a leg; in a different environment tigers might not be striped. Another important criticism is that people who seem perfectly able to use words to refer are too ignorant to provide descriptions adequate to identify the referents. Thus, some who use 'elm' and 'beech' cannot supply descriptions that distinguish elms from beeches; many who use 'gold' cannot distinguish gold from fool's gold. Worse, speakers are often so wrong about the referent that the descriptions they would provide apply not to the referent but to other entities or to nothing at all. Sometimes the whole speech community is ignorant or wrong about the referent. Thus, it was once common to associate 'fish' with 'whale'. Description theories of these words seem to require too much knowledge, placing too great an epistemic burden on speakers.

Putnam added a further argument, built around the following fantasy. Imagine that somewhere in the galaxy there is a planet called Twin Earth. Twin Earth, as its name suggests, is very like Earth. In particular, each Earthian has a doppelgänger on Twin Earth who is a cell-for-cell duplicate of the Earthian. Twin Earth differs from Earth in one respect, however: the liquid that the Twin Earthians who appear to speak English call 'water' – liquid that is superficially indistinguishable from what we call 'water' – is not H_2O but a very different compound XYZ. So Oscar on Earth and Twin Oscar on Twin Earth refer to different liquids by 'water'. Yet Oscar and Twin Oscar are doppelgängers, associating exactly the same descriptions with 'water' (which is more plausible if we place Oscar and Twin Oscar in 1750, before the chemical composition of water was known). So those associations are not sufficient to determine reference, and the description theory must be wrong. Indeed, nothing going on in the head is sufficient to determine reference. As Putnam put it, 'meanings just ain't in the head': meanings require a relation to something external to the thinker.

This is not to say that description theories fail for all representations. They still seem plausible for 'bachelor', and perhaps even for 'pencil' and 'pediatrician' (cf. Putnam, 1975). But even where they work, description theories have a problem: they are essentially incomplete. Suppose that a theory claims that the reference of 'bachelor' is determined by the references of 'adult', 'unmarried' and 'male'. We then need to explain the references of those words to complete the explanation of the reference of 'bachelor'. Description theories might be offered again. But then the explanation will still be incomplete. At some point we must offer a theory of reference that does not make the reference of one word dependent on those of others. We need an 'ultimate' explanation of reference that relates some words directly to the world. Description theories 'pass the referential buck'. The buck must stop somewhere if there is to be any reference at all.

'Verificationist' theories of reference implicitly acknowledge this point. They take a broader view than description theories of the required identification: speakers refer to whatever objects they would identify as the referents, whether by description or by recognition. Speakers recognize a referent by pointing it out in a crowd saying, for example, 'that person'. But these theories still seem to place too great an epistemic burden on speakers: we can only dimly call to mind the appearances of many objects we refer to, and are often mistaken.

There has been disagreement over whether reference can be explained in non-semantic or non-mentalistic terms. Many think not: reference is, in a sense, irreducible. From a naturalistic perspective, this view is unacceptable: reference must be explained in scientifically acceptable terms, ultimately in physical terms. Attempted explanations have appealed to one or more of three causal relations between representations and reality. First, there is the historical cause of a particular token, a causal chain going back to the dubbing of the token's referent. Theorists interested in this have emphasized the 'reference borrowing' links in the chain: in acquiring a word or concept from others we borrow their capacity to refer, even if we are ignorant of the referent (Kripke, 1980; Donnellan, 1972; Putnam, 1975; Burge, 1979). Second, there is the reliable cause of tokens of that type: a token refers to objects of a certain sort because tokens of that type are reliably correlated with the presence of those objects. The token 'carries the information' that a certain situation holds, in much the same way that tree rings carry information about the age of a tree (Dretske, 1981; Fodor, 1990). Third, there is the teleological cause, or function, of tokens of that type, where the function is explained along Darwinian lines: the function is what tokens of that type do that explains why they exist, what the type has been 'selected for' (Millikan, 1984; Papineau, 1987).

What all these recent developments suggest is that the reference of any word or concept will be explained by some such 'ultimate' theory, or by a description theory, or by a theory that combines elements of both. And different sorts of representations may have different sorts of theories.

Finally, it should be noted that some have denied, in one way or another, the reality of reference (and hence the need for, or even possibility of, an explanation of it). Perhaps the most influential of these denials is the 'deflationary' theory of reference, which accompanies the deflationary theory of truth (Horwich, 1998).

THEORIES OF REFERENCE AND COGNITIVE SCIENCE

Cognitive science holds that humans (and some other things) have mental representations – concepts – the contents (meanings) of which determine behavior (Margolis and Laurence, 1999). The common view is that a concept (and the word that expresses it) typically refers, and its reference is central to its content and nature. So, on this view, theories of reference should bear importantly on theories of concepts. However, it is often not clear how they bear on a particular theory.

Many cognitive scientists take the inferential associations of a concept to constitute its content and nature. This could involve a rejection of the centrality of reference, but it usually does not. Most seem to hold the appealing Fregean view that a concept's inferential associations determine its reference, thus endorsing a description theory of reference. The bearing of theories of reference is then obvious. However, some may prefer a 'two-factor' theory of content according to which reference constitutes one factor and inferential associations constitute an independent factor which does not determine reference. Most of what is said about concepts would then concern that independent factor and so would have nothing to do with theories of reference.

Finally, although psychological research on concepts is usually presented as if it concerned the nature of concepts, it may really concern something very different: the 'structure of knowledge'. On this view, the research is not throwing light on what concepts are – on what constitutes their contents – but on what they do – on their cognitive role, in particular, their role in recognitive skills. Theories of reference have little bearing on this.

We can distinguish three particular implications of theories of reference for cognitive science.

First, it is common to have a holistic view of concepts: their contents are given by all or most of their inferential associations. When combined with the appealing Fregean view, this yields a holistic description theory of reference. A consequence of such a theory is that the mistaken views we tend to have about the putative referents of our concepts would lead to widespread reference failure: most of our concepts would not fit anything in the world. So the Fregean view counts against holism; and the holistic view encourages a two-factor theory of concepts.

Second, verificationism is common in psychology: a concept's content is thought to be constituted by those features that lead us to apply it to the world. This application is naturally taken as determining reference. So the concept 'bird' refers to the objects that those who have the concept identify as birds. The recent developments in the theory of reference, particularly the view that reference can be borrowed, cast considerable doubt on verificationism: a person may have the concept 'elm', not by virtue of any ability to identify elms, but by virtue of being appropriately linked with a linguistic community.

Third, the simple Frege–Russell description theory calls to mind the 'classical' theory of concepts in psychology, and the more modern cluster theory calls to mind 'family resemblance', 'prototype' and 'exemplar' theories (Smith and Medin, 1981). In so far as the psychological theories are rightly construed as analogous to description theories, the recent developments cast doubt on them, particularly where they concern natural kind concepts like 'bird'.

Still, this may well be the wrong construal of some of these psychological theories. Thus, prototype theory usually seems to claim, not that a concept refers to an object that has most of the prototypical features (the cluster theory), but rather that it refers to an object to the degree that it has those features. Thus, the concept 'bird' is said to refer to penguins to a lesser degree than to robins. But this could be so only if penguins were birds to a lesser degree than robins, not 'fully' birds. For clearly the concept 'bird' fully refers to birds, and so if penguins are fully birds then the concept 'bird' must 'fully' refer to them. So prototype theory clashes disastrously with a biology which tells us that penguins are indeed 'fully' birds, as much so as robins. In the light of this, perhaps prototype theory should not be taken as a theory of the nature and reference of concepts but rather of their roles in recognition.

In conclusion, it is proving hard to discover fully satisfactory theories of reference. This has led some to pessimism about reference. This pessimism has a price: either we must abandon the view that reference is central to meaning (content), or, more radically still, we must abandon the idea that we have meaningful representations at all.

References

Burge T (1979) Individualism and the mental. In: French PA, Uehling TE and Wettstein HK (eds) *Midwest Studies in Philosophy*, vol. X 'Philosophy of Mind', pp. 73–121. Minneapolis, MN: University of Minnesota Press.

Donnellan KS (1972) Proper names and identifying descriptions. In: Davidson D and Harman G (eds) *The Semantics of Natural Language*, pp. 356–379. Dordrecht: Reidel.

Dretske FI (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Fodor JA (1990) *A Theory of Content and Other Essays*, pp. 51–136. Cambridge, MA: MIT Press.

Frege G (1893) On sense and reference. In: Geach P and Black M (eds) (1952) *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell.

Horwich P (1998) *Truth*, 2nd edn. Oxford: Clarendon Press.

Kripke SA (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Margolis E and Laurence S (eds) (1999) *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Millikan R (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Papineau D (1987) *Reality and Representation*. Oxford: Blackwell.

Putnam H (1975) *Mind, Language and Reality*, pp. 196–290. Cambridge, UK: Cambridge University Press.

Russell B (1912/1959) *The Problems of Philosophy*, pp. 46–59. London, UK: Oxford University Press.

Searle JR (1958) Proper names. *Mind* **67**: 166–173.

Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Further Reading

Devitt M and Sterelny K (1999) *Language and Reality: An Introduction to the Philosophy of Language*, 2nd edn, pp. 45–113. Oxford: Blackwell.

Donnellan KS (1966) Reference and definite descriptions. *Philosophical Review* **75**: 281–304.

Dummett M (1973) *Frege: Philosophy of Language*, pp. 110–151. London, UK: Duckworth.

Evans G (1982) *The Varieties of Reference*, McDowell J (ed.) Oxford: Clarendon Press.

Godfrey-Smith P (1992) Indication and adaptation. *Synthese* **92**: 283–312.

Kaplan D (1989) Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. Afterthoughts. In: Almog J, Perry J and Wettstein H (eds) *Themes from Kaplan*, pp. 481–614. Oxford: Oxford University Press.

Neale S (1990) *Descriptions*. Cambridge, MA: MIT Press.

Neander K (1995). Misrepresenting and malfunctioning. *Philosophical Studies* **79**: 109–141.

Searle JR (1983) *Intentionality: An Essay in the Philosophy of Mind*. Cambridge, UK: Cambridge University Press.

Representation, Philosophical Issues about

Intermediate article

Dan Lloyd, Trinity College, Hartford, Connecticut, USA

CONTENTS

Representation, the relation between a symbol, sign, or other vehicle and its content, has historically been analyzed through resemblance and causality; both accounts are flawed. The connectionist idea of distributed representation opens a window into the algebraic structure of representational vehicles, improving the prospects for a scientific theory of representation.

INTRODUCTION

Humans are representing animals, and we have built a world filled with representations of many kinds. Consider, for example, the number and variety of pictorial representations: paintings, photographs, moving pictures, line drawings, caricatures, diagrams, icons, charts, graphs, and maps. Add the variety of linguistic representations, in signs, titles, texts of all kinds, and especially spoken words and sentences. Multiply by all of the forms of recording and storage, especially computer media with their many kinds of codes. Note that all of those external forms of representation can themselves be re-represented as 'meta-representations' (e.g. a book cover featuring a reproduction of a painting of an artist at work on a canvas depicting a musical score). Finally, add the representations in one's head, including sensation or perception, and one's store of beliefs, opinions, hopes, and fears, expressed in forms similar to words, images, or sensations. Human life is largely a cycle of making and interpreting representations. And yet, it is not clear that we have an adequate theory of representation: what they are (at the most general level) and how they work (again, at the most general level).

We also lack full theories of other essential parts of life, like love or justice, but representation plays an essential role not only in life but also in science. Along with the related notion of computation, representation is a foundation of cognitive science. According to cognitive science, the mind computes: it is a symbol manipulator, where symbols are physical tokens that are moved or manipulated according to their formal (syntactic) properties. The mind also represents: the symbols manipulated have meaning. Without computation, the brain would be inert, at best an archive of information. Without representation, the brain's machinations would be empty, without content or connection to the world. Even in broad outline, it is clear that ideas of computation and representation interact. However, it will be useful to look at representation in isolation first, and then examine its relations to ideas of computation.

Representation is usually understood as a relationship between something that represents, a 'vehicle' (e.g. a picture, sentence, or thought), and what it represents, its 'content' (e.g. an object, scene, or situation). The list of examples above begins to suggest the variety of vehicles. Potentially, the list is infinite. So is the potential range of content. Notably, content is not confined to present reality. It can reach out in time and space, and also embrace metaphor, pure fiction, paradox, and outright nonsense (as in Noam Chomsky's example, 'colorless green ideas sleep furiously'). But even at long physical and metaphysical range, representations nonetheless pinpoint their targets. Collecting the second biggest rock on the third largest planet in the Andromeda galaxy may be difficult, but representing it (as in this sentence) is easy.

CONSTRAINTS ON A THEORY OF REPRESENTATION

Representation is a potent multiform relationship. Any general theory of representation should accommodate the examples above, and more. From

the examples we can also abstract some basic properties of the relationship, pointing toward a theory. A theory of representation must account for the following:

- Concerning vehicles:
  Sign and medium. Signs are components of the vehicle which, if they change, lead to a change in the content of the representation. Media are changeable without changing the signs (content) they carry.
  Atomic and molecular signs. In many representational systems, signs combine to form complex signs whose content is not simply an aggregate of the contents of component signs.
  Adaptability. Any set of objects may be transformed, by various means, into a set of representations.
- Concerning content:
  Range. Representations can represent mundane present facts, but with equal ease they can represent counterfactual content. Erroneous representations are still representations, as are representations of mere possibilities, future and past events, and impossibilities.
  Focus. Representations have unlimited range, but express only certain specific contents, omitting a range of ancillary facts from representation. At one extreme, a representation may have no factual content at all. But even the most accurate representation is a narrow and partial window onto a potentially infinite set of factual conditions.
- Concerning the relationship:
  Asymmetry. The relation of representation is one-way, from vehicle to contents. Contents do not represent their vehicles.

MAIN TYPES OF THEORIES

Historically, there are two classes of theories of representation: those that hold that the relation is dyadic, between just the two main elements of vehicle and content; and those that hold that the relation is polyadic, involving third parties in addition to vehicles and content. But any valid theory must account for the constraints above. Several well-known theories fail to do so.

Dyadic Theories

Resemblance

Plato and Aristotle construed representation as *mimesis* or resemblance between vehicle and content, using pictorial art and tragedy as their main examples. In the *Republic*, for example, Socrates suggests that mirror images are the most perfect representations, and in *Cratylus* Plato raises the possibility that perfect representations, being perfect likenesses, might simply duplicate their objects. Resemblance theories have persisted in many forms ever since, sometimes covertly. Modern examples often restrict the domain and range of the resemblance relation, usually to certain formal properties of vehicle and content. For example, Shepard and Chipman (1970) propose that representations resemble objects by duplicating only the relations between them, rather than 'intrinsic' object properties (like color). (See also O'Brien and Opie, 2002.)

The problems with resemblance as the root of representation are well known. Resemblance fails to satisfy several of the listed constraints on representation. With respect to vehicles, it is too general, as any object is already similar to many other objects, in many respects. As a result, with respect to content, resemblance fails to focus on specific content but rather spans every resembling object or situation. At the same time, resemblance seems not general enough, in that counterfactual content seems fundamentally dissimilar from any real vehicle of representation. Moreover, for many representational systems, including spoken and written language, it is unclear how resemblance could hold at all between vehicle and content. Except in unusual circumstances, a sentence generally bears no resemblance to its content. Finally, resemblance, unlike representation, is symmetric: if A resembles B, then to the same extent B resembles A. Although restricting the domain of resemblance mitigates these problems, it does not dissolve them.

Causality

A second major type of dyadic theory of representation is the causal theory, which holds that the relation of content to vehicle is one of cause and effect. Whereas resemblance theories sprang from reflections on art, causal theories often begin with an account of perception. Descartes, for example, imagined the path from worldly stimulus to mental response as a causal one. Modern versions have usually acknowledged the primary problem with causal theories, strikingly like the problems faced by resemblance. Causality is everywhere, holding between many relata that are not at the same time related as vehicle and content. But at the same time, there is no obvious way that a counterfactual situation can cause anything at all. So causal theory entails representational vehicles that represent both too much (all their causal antecedents) and too little (omitting erroneous or other counterfactual content).

A modern variant on causal theory employs the concept of information, which originated in the work of Shannon and Weaver (1949) and was harnessed for cognitive science by Dretske (1981). The

information carried by a channel, in the traditional theory, is proportional to the degree to which the signals transmitted increase the receiver's confidence that particular conditions obtain at the signal's source. For Dretske, a sign carries the 'informational content' that (for example) the ball is red if it is certain that the ball is red, given that the sign has appeared (and also given certain background conditions). The strong accuracy demanded by Dretske's theory helps limit the generality of the standard informational relation (which can hold in some cases even where causality does not). But even as amended, the theory nonetheless requires additional apparatus to accommodate misrepresentation and focus. That is, it still labors to show how a representation picks out its content among a number of antecedent or intermediate 'sources' other than its proper object, and again when there is no real source at all.

Polyadic Theories

Dyadic theories acknowledge that between content and vehicle there will be diverse mediating processes, but they place no restrictions on these processes. Polyadic theories, in contrast, place restrictions on the processes that produce representations, or follow their production. While the dyadic theories faced their greatest difficulty in trimming a profusion of vehicle and content, the polyadic theories in general face the problem of subtly reintroducing 'representation' into their own explanations.

Intention

Since many representations are artifacts, one obvious theory of representation would simply govern the relation by the intentions of either the producers or receivers of representations. Thus, for example, to understand what an author means to represent, we appeal to the author's intentions. Similarly, a 3-year-old can elevate a smear of paint to a picture of Daddy with a simple declaration. These examples fix representationality and content with an appeal to the producer's intention. One could also assign these representational powers to receivers or interpreters.

In routine representational practice, both producers and receivers do seem to play a role in determining content. However, as a general theory of representation, intention theory faces an immediate problem, which is that the mental states of producers and receivers are themselves representations. Those mental states cannot derive their content from other producers or receivers, without leading to an infinite regress of intention bearers. Therefore, at best, intention theories help with artifactual representations, leaving untouched the problem of mental representations.

Some thinkers (e.g. Putnam, 1988) believe that the covert appeal to intention is inescapable in any theory of representation. Another response, however, is to seek polyadic theories that do not depend on an intelligence, in a way that begs the question, for conferring content. There are several possibilities.

Convention

A short step from the intentional is the conventional. Many familiar representational systems arose and evolved through convention, and several students of representation maintain that even the practices of pictorial mimesis are in fact conventions (Gombrich, 1969; Goodman, 1976). A convention, conceived as a set of orchestrated behaviors arising over time within a group, might escape the regress of intentions if 'orchestrated behavior' can be limited strictly to overt physical movement. Behaviorism itself labored for half a century to purge the mental from behavior, without ultimate success. But even were that problem to be solved, convention theory shares with intention theory the fundamental limitation to public representational systems. It is unclear how a convention could extend to psychological representations, without some appeal to a prior, non-conventional, form of native representation.

Functional role

'Native' or 'mental' representation thus emerges as the central issue not just for psychology but for every polyadic theory of representation. Here ideas of computation engage with ideas of representation. 'Functional role semantics' is one name for the representational theory arising within the general philosophy of functionalism. (It is also called conceptual role semantics, or causal role semantics.) Just as functionalism is the view that psychological states are to be differentiated and identified by their functions within a system of such states, so functional role semantics posits that the identity of a representation and its content depends on its functional role in a system of representations, each deriving its content from its role in the system.

The pocket calculator is a favorite example (Haugeland, 1985). To a Martian engineer, a calculator manipulates tokens whose meaning is unknown. Nonetheless, the tokens (shapes on the screen, or patterns of electron flow in the circuits) shift according to patterns. Looking over those

patterns, the Martian tries out various interpretations. Are the shapes pictures of herding animals? Is the flow a model of a river system? Eventually the engineer hits on the interpretation that the shapes are numerals, and the flow allows the numerals to evolve according to arithmetic laws. What makes the arithmetic interpretation unique is that it is truth-preserving. That is, the symbol sequence $2 + 3 =$ is reliably followed by a symbol which properly fits into many such sentences, but in general does not appear in places where the number 5 would be in error. Thus, the representational system attaches itself to a systematic interpretation all at once, and the most favored interpretation is the one that makes statements in the system most often true. Though devised as a theory of mental representation, this theory could apply to public representational systems as well.

Functional role semantics is not dependent on interpreting Martians or humans: their presence in the example above is merely illustrative, since 'interpretability' is a property of the system itself, in relation to the world. But the theory is dependent on another concept, truth, which is rather similar to our target, representation. The difficulties surrounding the concept of 'truth', even if they are different difficulties, do not encourage one to lean on it in a theory of representation. Nonetheless, functional role semantics has opened a path by regarding representation as a systemic property, rather than a property of individual tokens. This approach deserves closer examination, to see whether it can be developed in a way that avoids the problems faced by other theories.

One way to think of a representing system is as a single complex representation, further ramified by its orderly transformations from moment to moment. Of this vast and intricate vehicle we then ask, what relation could it hold to its equally intricate content? This takes us back to the beginning, looking at the dyadic possibilities of resemblance and causation. Now, however, the many components of the big vehicle all constrain the possibilities for its resembling object. It resembles (in some sense) the world overall. Misrepresentation, fiction, and other counterfactuals are then assigned content through their place in the system, rather than through any relation they may have to their content, since there is nothing real for them to be related to. The last task for the theory, then, would be to identify, somehow, the direction of representation, preserving the asymmetry of the relation.

This relation of complex resemblance has been tested by several thought experiments in which the complex vehicle and the world are held constant while the causal relations between system and world are altered. We will look at just two such experiments.

- *The Swampman* (Davidson, 1987). Imagine that some miraculous implosion of organic chemicals occurs in a chemical dump, resulting in the creation of something indiscernible from an ordinary adult human. So thorough is the coincidence that the Swampman has a full store of apparent beliefs and desires, comparable to those of any human. Are those internal states (its 'beliefs' and 'desires') real mental representations, or are they merely lucky simulations?
- *The brain in a vat* (Harrison, 1967; Putnam, 1981). Imagine that evil neuroscientists steal the sleeping brain of an ordinary human and install it in an aquarium, hooking it up to life support and duplicating all its neural connections. However, its previous connections via sense organs to the real world have now been replaced by an elaborate computer simulation concocted by the scientists. Would the mental states formed by the brain upon waking (for example, the computer-simulated percept of a morning cup of coffee), bear content as ordinary representations do?

The two fantasies are complementary. The Swampman gets to the present moment by an abnormal causal route, so its normal-seeming representational states are not generated by the normal causal processes. Henceforth, however, this individual will be in normal interactions with the world. The brain in a vat has the normal causal history, but at the present moment all of those normal relations have been terminated. Henceforth his or her interactions with the world will not be normal. Yet at every moment in both individuals there is something physically very similar to the representational system in our normal brains (we who presume that we are neither swamp people nor brains in vats). If either or both characters lack 'real' representations, then the thought experiments have revealed a commitment to an essential causal or informational link required for any representing system.

Philosophers have been divided in their intuitive responses to the two fantasies. Both characters, for example, have human moral standing, nor would we regard them as mentally impaired despite the myriad delusions that permeate their thought. This, along with the nagging awareness that we cannot tell from introspection that we are neither swamp people nor brains in vats, suggests that the right internal configuration may be all that matters, as suggested by functional role semantics. However, regardless of our intuitions, it seems that the systems approach has not fully answered the basic question. In part, this is the effect of the term 'complex' in 'complex internal configuration' and

'complex causal relations', which only evades certain questions.

CONTRIBUTIONS FROM COGNITIVE SCIENCE

Complexity, especially complexity evolving over time, has been the special expertise of cognitive science. If philosophy is at an impasse in understanding the intricate scenarios sketched above (and many others like them), then it may be useful to look for more concrete guidance from science. In this case, there is a promising lead in connectionism, variously known as parallel distributed processing or neural network modeling. The vehicle of distributed processing is widely termed 'distributed representation'.

Distributed representation refers to representations 'spread out' through a processing system (conceived as a network of neuron-like processors). Thus, no component of the system is dedicated to representing just one piece of content. Rather, each component contributes to many different representational states to many different degrees. It is the pattern of all the components that comprises a single representation. Different overall patterns involving those same components represent different contents.

This is not just an interesting idea. Many working models of distributed representations have been built. These models demonstrate that for a wide range of interesting human tasks, distributed processors could be doing the work. This 'existence proof' offered by parallel processing is similar to the existence proof offered by the digital computer in the early days of cognitive science, in its demonstration of the possibility of purely mechanical symbol processing. The possibility of parallel processing in the human brain has opened new research avenues in many areas of psychology and neuroscience.

In the appeal to distributed representation, connectionists have appropriated without analysis the notion of representation itself. Nonetheless, the properties of distributed representations offer a new view into the phenomena of 'complex resemblance' and 'complex causation'. With respect to resemblance, distributed representations offer principled methods for assessing similarity among sets of distributed patterns. These methods, borrowed from multivariate statistics, produce maps of the structure of a distributed representing system. They reveal that in various working systems, distributed processors generate internal states that track the external environment. At the same time (and with the same representations) they identify the conceptual or categorical membership of external stimuli, as well as tracking the temporal context of the stimuli and coding their precursors and anticipated consequents (e.g. Elman, 1990). All of this is possible, even easy, for distributed processors, because the many components of each representation allow for subtle variations of content. 'Same but slightly different' representations play specific roles in the processing system, with large or small differences in effect, according to the task at hand.

Complex causation emerges through a similar analysis. As connectionism opens a window onto complex patterns of activity in processing components, so also it opens a window onto the varying connections between the components. Although their analysis is in its infancy, these also afford a principled, mathematical treatment. Thus, through the connectionist lens, both resemblance and causation ultimately become matters of linear algebra.

A deep problem remains, however. In connectionism, both complex resemblance and complex causation can be mathematically defined between components of the representational vehicle (the system). But representation is a relationship between a system and the world. Extending the algebraic treatment to vehicle–world connections requires a step of 'preprocessing' the world to offer a systematic input, amenable to mathematical notation. Will there be a coding scheme for reality which does not ultimately trace (once again) to some intelligent interpreter or producer, whose own unanalyzed representations guide the parsing of the world? This uncertainty suggests that the question of representation is still unanswered. Complexity can be treated in systematic ways (by connectionists), but what kind of complexity is the mark of representation, and what kinds of complexity are 'mere' complexity?

IS REPRESENTATION MONADIC?

Every attempt to express the relation of representation in nonrepresentational terms has failed to meet the constraints of content, being either too short in range or too wide in focus. If neither dyadic nor polyadic relations are adequate for the task, then it is perhaps time for a proposal of last resort: what if representation is not a relational property at all, but rather a monadic, nonrelational property? One could retain the ideas of complex distributed representations discussed above, while escaping the difficulties of understanding the 'right' resemblance or the 'right' causation. What would remain is the

idea of content internal to the representation, content as a complex property without external reference of any sort. The analytical techniques applied to distributed representations would then afford direct displays of contents, again without the need for a further interpretive connection of vehicle to world.

The main advantage of this strategy is that it makes tractable the domain of representation, capturing it in the mathematics of distributed processing and distributed representation. So located, it may be possible to answer the question, what is the right sort of complexity for representation? For example, multivariate analyses generally abstract or simplify the complexity of the distributed representations under analysis. In doing so, these analyses compute how amenable the whole system of distributed patterns is to the analysis itself. A random pattern or random system cannot be analyzed smoothly: any attempt to describe it with fewer terms than there are components to the system results in distortion or inaccuracy (Dennett, 1991). Organized distributed systems, like those emerging from neural network research, on the other hand, analyze crisply into fewer terms. So, for all distributed systems, there is an objective measure of organization. This, or something like it, could be the new measure of representationality.

The main disadvantage of this strategy may be that it seems to abandon the fundamental idea of representation. Representation does seem to be a relation: it is the mind's way of making contact with the world and with other minds. Ignoring both the 'outside' and the 'other' seems tantamount to solipsism. At the end of this train of reasoning might be metaphysical dead-ends like Leibniz' *Monadology*.

Consider, however, that if the randomized network described above were turned loose in the world (even in a simulated 'toy' world), it could not meaningfully interact with the world. Brains are continually adjusted by pragmatic consequences, both on the evolutionary scale and in daily human cognizing. Thus, regardless of the metaphysical 'reach' of representation, something will always influence us from without. Complex systems that survive in life will as a matter of practical necessity develop an internal guidance system, which will be intricately responsive to environmental conditions. But this process (and all the real causal relations contained in it) need not be considered definitive of or necessary for representation. Limiting representation to a monadic property in no way removes the mind from nature.

THE FUTURE OF 'REPRESENTATION'

Cognitive science thus inherits a concept of representation that has resisted definitive analysis. Presumably, either a new analysis will clarify the concept, or 'representation' will disappear from the formal explanatory apparatus of cognitive science, to be replaced by a more direct language, perhaps from cognitive neuroscience. Either development would enhance and clarify the science of the mind. Such developments may possibly also oblige us to revise our way of thinking about human life.

References

Davidson D (1987) Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* **60**: 441–458.

Dennett D (1991) Real patterns. *Journal of Philosophy* **88**: 27–51.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Elman J (1990) Finding structure in time. *Cognitive Science* **14**: 179–211.

Gombrich E (1969) *Art and Illusion*. Princeton, NJ: Princeton University Press.

Goodman N (1976) *Languages of Art*. Indianapolis, IN: Hackett.

Harrison J (1967) A philosopher's nightmare, or the ghost not laid. *Proceedings of the Aristotelian Society (New Series)* **67**: 179–188.

Haugeland J (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

O'Brien G and Opie J (2002) Notes towards a structuralist theory of mental representation. In: Clapin H, Staines P and Slezak P (eds) *Representation in Mind: New Approaches to Mental Representation*. Westport, CT: Greenwood.

Putnam H (1981) *Reason, Truth, and History*. Cambridge, UK: Cambridge University Press.

Putnam H (1988) *Representation and Reality*. Cambridge, MA: MIT Press.

Shannon C and Weaver W (1949) *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Shepard R and Chipman S (1970) Second-order isomorphism of internal representations: shapes of states. *Cognitive Psychology* **1**: 1–17.

Further Reading

Danto A (1981) *The Transfiguration of the Commonplace*. Cambridge, MA: Harvard University Press.

Devitt M and Sterelny K (1987) *Language and Reality*. Cambridge, MA: MIT Press.

Fodor J (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Gadamer H-G (1975) *Truth and Method*. New York, NY: Seabury.

Gibson J (1971) The information available in pictures. *Leonardo* **4**: 27–35.
Smolensky P (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–23.
Von Eckardt B (1993) *What is Cognitive Science?* Cambridge, MA: MIT Press.
Watson RA (1995) *Representational Ideas: From Plato to Patricia Churchland*. Dordrecht, Netherlands: Kluwer.
Wittgenstein L (1921) *Tractatus Logico-Philosophicus*. London, UK: Routledge.
Wittgenstein L (1953) *Philosophical Investigations*. Oxford, UK: Blackwell.

Self, Philosophical Issues about

Intermediate article

Shaun Gallagher, Canisius College, Buffalo, New York, USA

CONTENTS

Multiple concepts of the self present a variety of philosophical problems, including questions about ontological and epistemological status, relations between self and consciousness, self-consciousness, memory, and body. These issues are related to the question of determining the proper methodological approach to the self.

THE CONCEPT OF THE SELF

Currently there is no philosophical definition of the self that has drawn consensus. The concept is characterized in a variety of ways throughout the literature. The disparity of conceptions can be glimpsed by considering an incomplete inventory of terms that have come to proliferate in philosophical and psychological accounts:

- material self, social self, spiritual self (James, 1890);
- ecological self, interpersonal self, extended self, private self, conceptual self (Neisser, 1988) (*See* **Perception: The Ecological Approach**); and
- autobiographical self, cognitive self, conceptual self, contextualized self, core self, dialogical self, embodied self, empirical self, fictional self, minimal self, neural self (see, e.g. Damasio, 1999; Strawson, 1999).

This disparity, which is both problematic and productive, is directly related to the variety of methodological approaches taken within philosophy and in related interdisciplinary studies of the self. These include introspection, phenomenological analysis, linguistic analysis, the use of thought experiments, empirical research in cognitive and brain sciences, and studies of exceptional and pathological behavior. In this light, one problem is whether different characterizations of self signify diverse aspects of a unitary concept of selfhood, or whether they identify different and unrelated concepts. This problem of 'inter-theoretical coherency' is addressed below. Regardless of how one responds to this problem, however, the variety of approaches and definitions found in studies of the self productively reinforces the idea that human cognition involves complex and varied aspects that are not easily reducible to one set of principles. (*See* **Self**)

HISTORICAL BACKGROUND

The modern set of problems concerning the self emerged in a definitive way in the seventeenth century from a background that involved theological controversies concerning immortality and the concept of multiple persons in a trinitarian god. The question of personal immortality, which stretches back to ancient times, is no less contentiously debated in contemporary, and mostly secular, discussions about the survival of the self. The question of the unity of the self, even across multiple persons, has been transformed in modern investigations of personal identity and psychopathology. The *locus classicus* of this transformation is to be found in the pages of John Locke's *Essay on Human Understanding*, where, for the first time, the problem of personal identity was defined. Locke reframed the problem by moving away from traditional ideas of a substantial self. Theorists from Plato to Descartes had answered questions about the self in terms of a soul or mind constituted as a separate or separable, non-physical substance. (*See* **Dualism**)

Although Locke's definition of the self moves away from the idea that it is a spiritual substance, it nonetheless remains perfectly consistent with the idea that the self is not essentially embodied. He considers it to be something that depends on cognitive function. The person or self is that which 'can consider itself as itself, the same thinking thing, in different times and places; which it does only by that consciousness which is inseparable from thinking, and, as it seems to me, essential to it … and by this every one is to himself that which he calls self …' (Locke, 1690) The self, then, is a psychological entity. Locke distinguishes this concept of self or person from the concept of human being

('man') defined as an entity that has a body characterized by a continuity of organization over time. In Locke's definition of the person as a psychologically continuous entity, memory plays the most important role. To the extent that I can remember my past actions, those actions constitute part of what I call my self.

One can see, already in Locke, a nest of related philosophical problems pertaining to the self. First, there are ontological problems about what constitutes self-identity in any present moment, and what constitutes the self's identity over time. Second, there is a complex set of epistemological problems related to the role of consciousness, reflection, memory, and other cognitive faculties in providing experiential or conceptual access to the self. Third, cutting across these problems, there are unresolved questions concerning what role the body might play in either constituting or providing epistemological access to the self.

Locke's account is often viewed as reducing ontological problems to epistemological ones. That is, self-identity is constituted precisely through the epistemological access we have to the self. The self is constituted by a psychological continuity that is more or less transparent to self-reflection. To the extent that we fail to be conscious of ourselves, or to remember past events in our life, such events are not part of who we are. This interpretation is the basis of the famous criticisms launched against Locke by Butler (1736) and Reid (1785), theorists who championed a return to the substantialist view of the self. The identity of the person, they argued, is presupposed by memory, not made by it. On the other hand, Locke's view leads directly to David Hume and to a basic question about the ontological status of the self. Is the self something real? The influential Humean answer is 'No'. (*See* **Hume, David**)

Hume supposes that what we mean by the term 'self' is some constant and invariable impression of which we are continually conscious. The problem is that he can find no such thing in his own experience. What Hume introspects is a succession of ever changing impressions, in effect, 'a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement' (Hume, 1739). According to his 'bundle' theory of the self, there is no real self in the manner defined. Nonetheless, we seem to be under the impression that a self is just what we are. To account for this common (folk psychological) conviction Hume proposes that the self is a false impression, a fiction produced by the imagination which in memory mistakes resemblance for identity, and extends that identity beyond what we can remember. To the extent that we imagine there to be causal relations among our mental or intentional states, Hume suggests, the self has the same kind of ontological status as a republic or nation. We attribute identity to such an entity only in a formal way. The reality is a diverse set of changing individuals (perceptions) and institutions. Furthermore, philosophical disputes about the self are merely grammatical or verbal problems. The question is not so much about identity or continuity of the self over time but about how we should talk about the fictional entity that we call 'the self'. (*See* **Self-consciousness**)

One way to deal with the Humean problem is to ask about the unity of the imaginative faculty. If imagination or memory produces a consistent representation of the empirical self, such consistency is likely due to the unity of the cognitive subject. Thus Immanuel Kant (1781) argues for a transcendental (that is, nonempirical) unity of apperception which he calls the transcendental ego – a self that, in principle, is not open to natural scientific investigation. (*See* **Consciousness, Unity of**)

A more naturalistic way to deal with the Humean problem is to accept the definition of the self as a bundle of impressions and to identity the proper criteria necessary to judge which impressions are to be bundled together. William James (1890), setting aside the ontological issues about what actually makes the self, proposes criteria that account for the sense of personal identity. He finds in the stream of consciousness an apparent constancy, characterized as a feeling of 'warmth and intimacy'. This is an impression produced by the physical processes of our own body, by 'faint physiological adjustments' including changes in heart rate, breathing, and so on. This uniform feeling, to whatever extent it pervades our experiences over time, provides an epistemological criterion that we may use to judge the continuity of self. An important objection to using something like a felt impression as the criterion for judging one's identity, however, is that it is regressive. In other words, use of the criterion depends on a subjective awareness of sameness or resemblance from one instance of the impression to another. It thus presupposes the identity of the experiencing subject of each individual state of awareness, which is precisely what one aims to establish through the use of the criterion. (*See* **James, William**)

CURRENT PHILOSOPHICAL THEORIES OF THE SELF

In contemporary philosophical theory there are two main approaches to the question of the self. They extend the traditional answers that distinguish between psychological and bodily continuity.

The Psychological Approach

The psychological approach is closely associated with philosophical functionalism and is often developed through the use of thought experiments that involve brain transplants or the transfer of neurocognitive information from one subject to another. Our informed intuitions concerning a hypothetical brain transplant from body A to body B suggest that personal identity follows the brain and does not remain with the original body. Split-brain experiments, and thought experiments based on these, suggest the possibility that the survival of the self does not depend on the entire brain but on only a sufficient part of the brain. Functionalist theory takes this one step further. What is important about the brain is not physical continuity, not even of some minimal part of the brain; rather, the only thing about the brain relevant to the survival of the self is the psychological information (or the syntactical functions) it instantiates. Even if information (or function) requires physical instantiation, it is still possible to think that it can be moved around from one physical system to another. Thus, the possibility that, in principle, memory information can be removed from one storage device (the brain) and stored in another (another brain or a machine) suggests that psychological rather than physical continuity is what counts in the constitution of the self. (*See* **Self, Psychology of**; **Thought Experiments**)

Locke's emphasis on memory does not provide an adequate concept of psychological continuity for several reasons. First, it is not just that episodic or autobiographical memory is untrustworthy. Rather, psychologically, I am more than just my memories. The concept of psychological continuity includes broader aspects of intentionality – my other cognitive states and dispositions, my behavioral peculiarities, my habitual ways of thinking, my intentions and plans etc. (*See* **False Memory**)

Second, the reliance on memory is further complicated by the possibility that, according to the terms of another thought experiment, I could end up with memories that actually belong to someone else. If one part of brain A (or information from brain A) is transferred to body B, and another part transferred to body C (and assuming that B and C end up as persons of sound mind), then it would be quite possible for B and C to remember certain events as having happened to each of them respectively, when in fact the events happened to their neurological or psychological ancestor, A. In such cases (referred to as cases of 'fission') the memories had by B and C are called 'quasi-memories' and are not sufficient to guarantee the identity of one self across time (Parfit, 1984, Shoemaker, 1984). (*See* **Memory, Philosophical Issues about**)

Third, solving the problem of personal identity presupposes some account of how self is discriminated from non-self. In this regard, a question arises when one considers whether memory can operate as a criterion of first-person identity. The use of a criterion depends on a comparative judgment. To judge whether some object X is in fact Y I need to compare what I know about X to whatever criterion I have for defining Y. If it satisfies the criterion, then I am justified in saying that X is Y. Is this how it works with memory? If I remember visiting a friend yesterday, and I want to use that memory as a criterion that allows me to judge that the person who visited my friend was in fact the same person as myself, then the memory that I use cannot be the memory '*I* visited my friend yesterday.' It would have to be something like the memory '*Someone* visited my friend yesterday.' I would then have to use some appropriate criterion (to compare some knowledge about myself with some knowledge about the 'someone') that would allow me to judge that the 'someone' was actually myself. Only on that basis, it seems, would I be justified in making the statement '*I* visited my friend yesterday.' In that case, however, the statement 'I visited my friend yesterday' would not be a memory statement but a conclusion drawn from the application of a criterion (Shoemaker, 1959). Clearly, if identification and memory worked this way, there would be no such thing as autobiographical memory. Memory would tell me only that someone did X and I would have to make sure that that person satisfied some criterion before I could say that it was I who did X. Such considerations lead Shoemaker to conclude that memory does not act as a criterion for first-person identity, but not because we use a different criterion. Rather, first-person ability to identify the self does not require the use of a criterion at all. Indeed, my ability to say 'I', or to be self-conscious in certain ways, does not depend on a process of identification in which I make a criterial judgment with the possibility that I could get it wrong.

One's first-person access to the self by introspective means is characterized by a direct and sure immediacy that is immune to misidentification (Shoemaker, 1984; Strawson, 1994). (*See* **Autobiographical Memory**)

The Bodily or Biological Approach

There are clear ways in which bodily continuity does *not* enter into the constitution of the self. If one loses a limb, for example, one's self is not necessarily diminished. It also seems obvious that a dead body is still a body in some sense but is no longer a self. Some thinkers have argued that a working brain, or at least some part of the brain, seems a necessary if not a sufficient condition for the existence of a self. Yet functionalist-inspired thought experiments about body (brain) exchange and information transfer have suggested the possibility, in principle, that the self could exist independently of the human body. To the extent that psychological information could be kept intact indefinitely (for example in a machine) this possibility has even suggested a new sense of immortality. Those who pursue the psychological approach have therefore found reason to dismiss the body as an important element for explaining the self.

Towards the end of the twentieth century, however, a variety of theorists in philosophy of mind, phenomenology, and neuroscience revived the notion that embodiment contributes in essential ways to the constitution of the self (e.g. Clark, 1997; Damasio, 1999; Varela *et al*., 1991). The strongest versions of this approach do not dismiss the importance of psychological continuity but provide a way to understand how embodiment shapes the constitution of the self as a psychological entity. Following this line of thought, the following points can be mentioned. (*See* **Embodiment**)

Aspects of embodiment directly shape mental experience and thereby contribute to the constitution of the self. Many thought experiments cited in support of the psychological approach ignore the fact that a person's character, mannerisms, and emotions, which in some way contribute to his or her identity, are embodied. An individual's disposition is instantiated in the performances of embodied processes; mood and emotional states, and appetitive states, are not divorced from but can dramatically shape cognitive states. If two persons are physically unlike one another – for example, if they were of the opposite sex – it would be difficult to imagine one individual's disposition or expressive mannerisms being instantiated in the performances of the second individual's body as the result of a memory transfer or brain transplant. Such dispositions are not simply the result of brain function. They depend both on *cultural factors* – for example those having to do with gender differences – which come to be inscribed in bodily mannerisms, posture and movement style, and *biological factors* such as autonomic regulations, or the effects of hormonal levels, the mix of neurotransmitters, and so forth. Most psychological-continuity theorists ignore these complications.

The embodied approach argues that the body's operations help to constitute affective and cognitive experience. The physical nature of the body, and how the body operates, completely conditions conscious activities. Disembodied experience is an impossible experience. With regard to perception, for example, the egocentric spatial framework of our experience is in fact a body-centric framework tied to the operations of the body. To see is not only *to see something* but also to see *from somewhere*, and under conditions defined by the position and postural situation of the perceiving body and by certain physical capacities of the body and its sense organs. The perceiving body provides a coherency to consciousness across individual perceptual events. That is, a certain coherency of experience is produced by the fact that it is one body doing the perceiving. Biological and emotional, as well as the sensory-motor aspects of embodiment, resonate in the disposition of experience and help to individuate that experience as one's own. Proprioceptive awareness (already operative in very young infants), for example, provides a mode of first-person access to the embodied self that cannot be mistaken in regard to identifying whose body it is that is being experienced (Bermúdez, 1998; Cassam, 1995). A proprioceptive sense of self, a form of bodily self-awareness, plays an important role in the differentiation between self and non-self that informs ecological perception and shapes our psychological identity on a very basic level.

OTHER PHILOSOPHICAL ISSUES ABOUT THE SELF

Other issues concerning the self are often discounted in the mainstream philosophical theories outlined above. Psychological and biological approaches frequently focus their analyses in ways that ignore or set aside the social dimensions that play an important role in the constitution of the self. Social construction theories understand the self as produced by its interpersonal relationships and the cultural practices and institutions that

surround it. The individual self is defined as the cross-section of the various roles it plays in its social milieu. This idea is further developed by narrative theories of the self that understand self-identity to be determined by the stories that are generated by individuals or groups. Dennett's (1988) concept of the self as a 'center of narrative gravity,' that is, an abstract intersection of the various stories that can be told about the individual, is one version of this theory. Other versions emphasize the notion of a self that may be relatively unified or disunified across the sum total of its life narratives and so include within itself equivocations, contradictions, and latent interpretations that find expression in different behavioral situations (Ricoeur, 1992). The view of the self extended across various social roles and across a rich set of diversified narratives permits an account of conflict, moral indecision and self-deception that would be difficult to work out in terms of either an abstract center or a perfectly unified self concept. (*See* **Culture and Cognitive Development**)

Methodological issues with respect to the self are also of concern to philosophers and cognitive scientists. Within philosophy there are debates about how best to approach the problem of the self. Since the time of Locke, for example, philosophers have appealed to thought experiments to outline the logical possibilities pertaining to the self. A variety of limitations and difficulties with this approach have motivated some theorists to turn to the study of pathological cases (e.g. Wilkes, 1988). Understanding the pathologies of the self involved in depersonalization, dissociation disorders, schizophrenia etc. may help to throw more realistic light on the self-structure of normal experience.

The problem of intertheoretical coherency may be more critical for the cognitive sciences. One important strategy for addressing problems as complex as consciousness and the self is to admit that no one method is adequate for a complete explanation. In the process of triangulation several methods are employed and attempts are made to discover a coherent explanation in the results. One issue that seems especially relevant in this case is whether the disparate concepts of self (listed at the beginning) are simply different but ultimately consistent portraits of one unitary thing, or whether different methods capture entirely different things that in each case we label 'self'. Does a neurological approach to the self explain the same thing that narrative theorists call 'the self'? Is it the same self that is material, social, spiritual, autobiographical, cognitive, contextualized, embodied, fictional, and neural? Or are we dealing with a range of phenomena that ultimately cannot be made consistent with each other?

RELEVANCE TO COGNITIVE SCIENCE

Whether the self is real or fictional, one thing or many, it is clearly something that needs explanation, for at least two reasons.

First, there is an undeniable sense of self that accompanies experience and action. One normally has both a sense of ownership and a sense of agency for one's body, actions, and thoughts, and in certain pathologies one or another of these senses can be disrupted (Gallagher, 2000). If this phenomenology points to something real, then a full explanation of cognition would require some account of the self. Even if the self has a fictional status, however, one needs to explain why the fiction arises. In this case we might appeal to psychology, as Hume did, or to accounts involving the neuromechanisms responsible for the implicit (ecological and proprioceptive) structures of perception and action, for autobiographical memory, and for the generation of narrative. It is possible, for example, that our sense of self is generated in a neural signature that originates in subcortical brain areas, and reiterates throughout various cortical processes so that it structures our experience and accompanies all of our thoughts (Panksepp, 1999). (*See* **Personal Identity**)

Second, self-structures play an important role in the way intelligent systems work. At very basic levels of sensory-motor processes it would be difficult to explain how an organism copes with its environment without employing a differentiation between self and non-self. If this distinction is built into intelligence from the very start, it is essential for the cognitive sciences to design a model to account for it, and to put it to work in fields such as artificial intelligence and robotics (see, e.g., Tani, 1998). (*See* **Consciousness, Machine**)

Interdisciplinary approaches that focus on a single aspect or dimension of the complex problems of self are more likely to mitigate problems of intertheoretical coherency. Neuroscientists, psychologists, psychiatrists and roboticists, as much as philosophers, have an interest in developing a neurocognitive model of the self. The neuropsychology of dissociative disorders and the neuroscience of the split brain, as well as the cognitive linguistics of narrative structure, may throw light on how narrative generates a seemingly unified self in normal humans. A variety of approaches, including the neuroscience of motor action, animal studies, and developmental

psychology, are needed to understand aspects of self-awareness, self-recognition, agency, and social interaction, and how such things contribute to the generation of self-identity. In the end, if good explanations of these various aspects of experience are developed, the cognitive sciences have the potential to recast the central philosophical questions about the self. (*See* **Split-brain Research**)

References

Bermúdez JL (1998) *The Paradox of Self-Consciousness*. Cambridge MA: MIT Press.

Butler J (1736) Of personal identity. In: Bernard JH (ed.) *The Works of Bishop Butler*, vol. II. London: 1900. [Reprinted in Perry J (ed.) *Personal Identity*, pp. 99–105. Berkeley: University of California Press, 1975.]

Cassam Q (1995) Introspection and bodily self-ascription. In: Bermúdez JL, Marcel AJ and Eilan N (eds) *The Body and the Self*, pp. 311–336. Cambridge, MA: MIT Press.

Clark A (1997) *Being There: Putting Brain, Body, and World Together Again*. Cambridge MA: MIT Press.

Damasio A (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt Brace.

Dennett D (1988) Why everyone is a novelist. *Times Literary Supplement* **4459** (Sept.16–22, 1988): 1016, 1028–1029. [Reprinted as: The self as the center of narrative gravity. In: Kessel F, Cole P and Johnson D (eds) *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Lawrence Erlbaum, 1992.]

Gallagher S (2000) Philosophical concepts of the self: implications for cognitive science. *Trends in Cognitive Science* **4**: 14–21.

Hume D (1739) *Treatise of Human Nature*. Selby Bigge LA (ed.) Oxford, UK: Clarendon Press, 1975.

James W (1890) *The Principles of Psychology*. New York, NY: Dover, 1950.

Kant I (1781) *Critique of Pure Reason*, translated by NK Smith. London: Macmillan, 1929.

Locke J (1690) *An Essay Concerning Human Understanding*. New York, NY: Dover, 1959.

Neisser U (1988) Five kinds of self-knowledge. *Philosophical Psychology* **1**: 35–59.

Panksepp J (1999) The periconscious substrates of consciousness: affective states and the evolutionary origins of the self. In: Gallagher S and Shear J (eds) *Models of the Self*, pp. 113–130. Exeter, UK: Imprint Academic.

Parfit D (1984) *Reasons and Persons*. Oxford, UK: Clarendon Press.

Reid T (1785) *Essays on the Intellectual Powers of Man*. [Reprinted as Woozley AD (ed.) (1941) London: Macmillan.]

Ricoeur P (1992) *Oneself as Another*. Chicago, IL: University of Chicago Press.

Shoemaker S (1959) Personal identity and memory. *Journal of Philosophy* **56**: 868–882.

Shoemaker S (1984) *Identity, Cause and Mind*. Cambridge, UK: Cambridge University Press.

Strawson G (1999) The self and the SESMET. *Journal of Consciousness Studies* **6**(4): 99–135.

Strawson PF (1994) The first person – and others. In: Cassam Q (ed.) *Self-Knowledge*, pp. 210–215. Oxford, UK: Oxford University Press.

Tani J (1998) An interpretation of the 'self' from the dynamical systems perspective: a constructionist approach. *Journal of Consciousness Studies* **5**(5/6): 516–542.

Varela FJ, Thompson E and Rosch E (1991) *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Wilkes KV (1988) *Real People: Personal Identity without Thought Experiments*. Oxford, UK: Clarendon Press.

Further Reading

Bermúdez JL, Marcel AJ and Eilan N (1995) *The Body and the Self*. Cambridge, MA: MIT Press.

Cassam Q (1994) *Self-Knowledge*. Oxford, UK: Oxford University Press.

Gallagher S and Shear J (1999) *Models of the Self*. Exeter, UK: Imprint Academic.

Hacking I (1995) *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton, NJ: Princeton University Press.

Martin R and Barresi J (1999) *Naturalization of the Soul: Self and Personal Identity in the Birth of Modern Psychology*. London and New York: Routledge.

Noonan HW (1989) *Personal Identity*. London, UK: Routledge.

Olson E (1997) *The Human Animal: Personal Identity without Psychology*. Oxford, UK: Oxford University Press.

Parfit D (1984) *Reasons and Persons*. Oxford, UK: Oxford University Press.

Strawson PF (1959) *Individuals: an Essay in Descriptive Metaphysics*. Oxford, UK: Oxford University Press.

Taylor C (1989) *Sources of the Self: the Making of the Modern Identity*. Cambridge, MA: Harvard University Press.

Williams B (1973) *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge, UK: Cambridge University Press.

Self

Intermediate article

JF Kihlstrom, University of California, Berkeley, California, USA
Stanley B Klein, University of California, Santa Barbara, California, USA

CONTENTS

Viewed from the perspective of cognitive science, the self is a knowledge structure representing one's declarative knowledge of oneself. This knowledge, in turn, may be classified as semantic knowledge of one's physical, personal, and sociocultural characteristics, episodic knowledge of one's past actions and experiences (and, perhaps, prospective memory of one's future actions and experiences as well), and metaknowledge of one's repertoire of mental and behavioral skills.

INTRODUCTION

Usually we think of cognition in third-person terms, in terms of how people (and other systems) acquire, represent, store, and use knowledge about the world outside themselves. This is also true for social cognition, which has to do with how we perceive, remember, and think about other people, their behaviors, and the situations in which we encounter them. At least for humans, however, cognition also turns inward, representing people's knowledge about themselves. A sense of self is critical to our status as persons. In fact, philosophers often use the terms *self* and *person* interchangeably: a capacity for self-awareness is necessary for full personhood. One has a sense of self if one is able to entertain first-person thoughts, and if one possesses first-person knowledge. The eye cannot see itself, but the self somehow knows itself: the simultaneous status of self as subject and object of awareness is one of the enduring problems of philosophy. (*See* **Social Cognition**)

In philosophy, self-awareness is often taken as a primitive, uninformed by reflection or any other form of conceptual thought. Self-awareness is also often taken as privileged. According to Sydney Shoemaker, first-person statements are immune to error through misidentification: they cannot be mistaken, and they cannot be contradicted; we cannot tell someone that he or she doesn't believe in God, feel depressed, or want a hamburger. Shaun Gallagher notes that there is a difference between the minimal self of immediate experience, and a more coherent narrative self that extends in time from the past through the present into the future. For example, amnesic patients will have a sense of their experiences in the immediate here and now, but no sense of their past experiences. In either case, the self is at the root of our sense of agency (i.e., that we are the causes of our actions) and our sense of ownership (i.e., that we are the ones who are having an experience). Further, as Kihlstrom has noted, the self is critical to both the monitoring and controlling functions of consciousness. In cognitive social psychology, the self is construed simply as one's mental representation of oneself – that is, one's idea or mental picture of one's physical, psychological, and sociocultural attributes, of one's own cognitive, affective, and conative states, and of one's own behavior. (*See* **Self, Psychology of**; **Consciousness, Philosophical Issues about**; **Consciousness and Higher-order Thought**; **Implicit Cognition**; **Consciousness, Function of**; **Consciousness and Attention**; **Consciousness, Cognitive Theories of**; **Consciousness, Sleep, and Dreaming**; **Consciousness, Stream of**; **Consciousness, Animal**; **Consciousness, Unity of**; **Self, Philosophical Issues About**; **Self-consciousness**)

MENTAL REPRESENTATIONS OF SELF

A central problem in cognitive science is knowledge representation. In general, we may distinguish between meaning-based knowledge representations, which store propositional knowledge about the semantic relations among objects, features, and events, and perception-based knowledge representations, which take the form of mental images representing the physical appear-

ance of objects, and the configuration of objects and features in space. Self-knowledge can be construed in both terms: the cognitive view of the self as a knowledge structure is anticipated in ordinary language when we refer to the self-concept and the self-image, and there is some value in taking such terms literally, and exploring their ramifications in cognitive theory. What does it mean to say that the self is a concept? That it is an image? (*See* **Imagery**; **Spatial Cognition, Psychology of**)

The Self as Concept

A concept is a mental representation of a category, a set of objects whose members share some features in common that are somehow distinct from objects in other categories. In the classical Aristotelian view, concepts are proper sets, defined by a list of features that are both singly necessary and jointly sufficient to identify an object as an instance of a category. From the classical point of view, then, the self-concept is identified by a set of features that are singly necessary and jointly sufficient to identify oneself as different from all others. The classical set view of the self as a set with only a single instance aptly recognizes our experience of ourselves as unique – that we are not the same as anyone else. Research by William McGuire and his colleagues has found that people who are in the minority with respect to age, birthplace, gender, ethnicity, and other physical, social, and demographic features are more likely to mention them when asked to describe themselves. Apparently, people notice aspects of themselves, and incorporate these attributes into their self-concepts, to the extent that these features render them distinctive. Along the same lines, Hazel Markus and her colleagues have suggested that the self-schema incorporates those features that are important to one's self-concept, not merely those that are descriptive of the self.

On the other hand, philosophers and cognitive scientists have identified a number of problems with the classical view of concepts as proper sets that have led to the progressive elaboration of a number of revisionist views. Chief among these alternatives is the probabilistic view of concepts as fuzzy sets represented by summary prototypes whose features are only imperfectly correlated with category membership. Instead of sharing some set of singly necessary and jointly sufficient defining features, instances of a concept are related to each other by a principle of family resemblance. This view of the self as a prototype has won wide acceptance within social cognition, but the notion of family resemblance suggests that there must be more than one self. Actually, the multiplicity of self makes sense from a social-psychological point of view. Despite our tendency to describe each other in terms of stable traits, human social behavior is widely variable across time and place, and our self-knowledge must represent this kind of variability. Perhaps, then, the self-as-prototype is abstracted from multiple, context-specific, mental representations of self. Clinical cases of multiple personality disorder (also known as dissociative identity disorder) bring the multiplicity of self into bold relief. In normals, autobiographical memory creates a continuity between the mental representation of self-in-one-situation and self-in-another that is destroyed by the amnesic barrier between multiple personalities. (*See* **Conceptual Representations in Psychology**; **Prototype Representations**)

Just as problems with the classical view of concepts led to the prototype view, so problems with the prototype view have led to further revisionist views. While the classical and prototype views construe concepts as some kind of summary of the features of category members, the exemplar view denies that we have any such summary at all. Instead, we represent concepts as a collection of instances. While the classical, prototype, and exemplar views are all based on a principle of similarity, the theory-based view holds that concepts are based on a theory of the domain in question. The instances of a category are not related by any kind of similarity but only through some theoretical explanation. Applied to the self-concept, the exemplar view would imply that there is no unitary self at any level of representation: all we have are context-specific selves. The theory view, in turn, would imply that our self-concept is a theory about ourselves – how we became what we are, and why we do what we do, rather than a list of features or instances. As yet, however, neither of these views of conceptual organization has been applied systematically to the self-concept.

The Self as Image

Perception-based representations have not been much studied in social cognition, but this is what the notion of 'self-image', taken literally, would entail. The rudiments of the self-image may be found in the body schema postulated by Henry Head to account for the ability of animals to maintain stability of posture and adjust to our physical surroundings. In addition to verbal knowledge about our characteristic features, then, we appear to possess analog representations of our own bodies

and their parts, independent of immediate sensory stimulation. Distortions and other aberrations of the self-image are often observed in cases of acute schizophrenia, anorexia, bulimia and other eating disorders, body dysmorphic disorder, and phantom limb pain. The neurological syndrome known as autotopagnosia is characterized by an inability to localize the parts of one's own body on demand; it is associated with focal lesions in the left parietal lobe. Outside the psychiatric and neurological clinic, experimental subjects prefer left–right reversals of photos of themselves, while they prefer unreversed photos of others. The fact that our picture preferences match the way we view ourselves in the mirror indicates that the self-image preserves both spatial relations and visual detail. (*See* **Schemas in Psychology**)

The Self as Memory

John Locke (1632–1704), the English philosopher, famously identified the self with memory: a person's identity, which is to say selfhood, extends to whatever of a person's past he or she can remember. Modern cognitive theories often distinguish between two forms of knowledge stored in memory. Declarative knowledge is our fund of factual knowledge about the world; it can be represented as sentence-like propositions. Procedural knowledge is our repertoire of rules and skills by which we manipulate and transform declarative knowledge; it can be represented as productions specifying the actions that will achieve some goal under specified conditions. Viewed as a memory, the self is usually thought of as declarative in nature, although certainly we can have declarative metaknowledge about our cognitive skills and abilities. Declarative memory, in turn, takes two basic forms. Episodic memory is autobiographical memory for the events and experiences of one's past, and its relevance to Locke's concept of the self is obvious. Our sense of self is very much tied up with the 'story' of how what we have experienced has made us who we are, and how who we are has led us to do what we have done. By contrast, semantic memory is more generic, context-free knowledge about the world. With respect to the self, semantic memory is tantamount to the self-concept and the self-image. (*See* **Skill Learning**; **Metacognition**)

Cognitive psychologists have only recently begun to study autobiographical memory, in the sense of people's memories for events and experiences occurring in the real world outside the laboratory. The simplest proposal for the organization of autobiographical memory is as a temporal sequence, running from birth to the present, with retrieval based on a serial backwards search. However, the serial record of autobiographical memory is more likely broken up into segments corresponding to the major phases of life: elementary school, high school, and college; first job, promotion, and retirement; first marriage and second. The fact that these phases overlap, and are defined subjectively and redefined retrospectively, makes the organization of autobiographical memory difficult to study except on an individual basis. The earliest years of life are covered by infantile and childhood amnesia: one's earliest recollection is typically dated between the third and fourth birthdays, and autobiographical memory does not typically achieve any kind of continuity before the 'five to seven shift' prominent in studies of child development. 'Flashbulb' memories occur where private and public history meet. Autobiographical memory is reconstructive in that our personal histories are shaped by our current theories of who we are and how we came to be that way. It is our 'story so far', a narrative that is constantly subject to revision.

In a generic associative network model of memory, the self (or each of a multiplicity of context-specific selves) can be represented as a node linked to other nodes representing corresponding knowledge about oneself. Considerable research has addressed the question of how episodic memory for one's past behaviors and semantic memory for one's characteristic traits can be represented in such a scheme. According to a conventional hierarchical model, nodes representing traits fan off the central node representing the self, and nodes representing behaviors fan off the traits they exemplify from nodes. An alternative model, related to 'self-perception' theories of social cognition, holds that memory contains only behavioral knowledge with traits known only by inference. A third model is that items of trait and behavioral information are represented by nodes that fan off independently from the 'self' node.

These models can be tested experimentally by means of a priming methodology. If the hierarchical model is correct, asking people questions about their traits should facilitate their answers to questions about their behaviors. If the self-perception model is correct, asking people questions about their behaviors should facilitate their answers to questions about their traits. A series of studies by Klein and his colleagues has revealed priming of neither sort, a consistent null result supporting the third model, of independence. The independence

model is also supported by neuropsychological evidence summarized below.

IS THE SELF A PERSON LIKE ANYONE ELSE?

Viewed as a concept, the self is a fuzzy set of context-specific selves, or perhaps a theory about oneself. Viewed as an image, the self represents both our perceptible features and their spatial configuration. Viewed as a memory, the self represents propositional knowledge about our abstract traits and specific behaviors, including a narrative record of our personal history. Lying behind all these models is the general idea that oneself is a person like anyone else and represented accordingly.

However, social psychological research has revealed a number of differences between cognition about oneself and cognition about others. While we tend to attribute others' behaviors to their internal traits, we are more likely to attribute our own actions to the demands of the situation (the self–other difference in causal attribution). We tend to perceive ourselves as more central to events than we really are (egocentricity), especially if the outcomes were positive (beneffectance). On the other hand, it is not yet clear that these biases are intrinsic to self-knowledge. Perhaps they apply to knowledge about others, as well, so long as we like them (as we tend to like ourselves) and/or know them well (as we think we know ourselves).

THE SELF AND ITS BRAIN

As philosophers and psychologists became interested in the biological substrates of mental life, and brain-imaging techniques have permitted us to watch the brain in action, many cognitive scientists have evolved into cognitive neuroscientists. Taking cognitive neuropsychology as a model, Klein and Kihlstrom have argued that neuropsychological studies of brain-injured patients, and brain-imaging studies of normal subjects, may provide new solutions to old problems and afford new theoretical insights for personality and social psychologists as well. Consider, for example, the relation between self and memory. If, as Locke argued, our sense of self is intimately tied up with our recollection of our past, what is the sense of self for an amnesic patient? H.M., the famous patient with the amnesic syndrome, cannot consciously remember anything that he did or experienced since the operation that destroyed his medial temporal lobes. H.M.'s anterograde amnesia is virtually complete, and hence his sense of self may be confined to whatever memories he has from before his surgery. Moreover, Locke did not fully appreciate the distinction between episodic and semantic memory. Amnesic patients retain some ability to acquire new semantic knowledge, and this dissociation may permit their self-concepts to be based on 'updated' semantic knowledge, even if they are lacking a complete record of autobiographical memory. (*See* **Amnesia**)

Such questions have not been asked of H.M. himself, but they have been asked of other patients. For example, the patient known as K.C., who suffered a severe head injury as a result of a motorcycle accident, has both a complete anterograde amnesia covering events since the accident, and a complete retrograde amnesia covering his life before the accident. K.C. has no autobiographical memory at all, but research by Endel Tulving reveals that he has a fairly accurate self-concept. The same accident that caused his amnesia also resulted in a profound personality change: the premorbid K.C. was quite extraverted, while the postmorbid K.C. is rather introverted. When asked to rate himself as he is now, K.C. rates himself as introverted, in agreement with his mother's ratings of him. Interestingly, his ratings of his premorbid personality do not agree with his mother's. K.C. has acquired semantic knowledge about himself, but he has not retained in episodic memory the experiences on which this self-knowledge is based; and his newly acquired semantic self-knowledge has effectively replaced that which he possessed before the accident.

Similar results were obtained by Klein and his colleagues in a study of W.J., a college freshman who suffered a temporary retrograde amnesia, covering the period since her high-school graduation, as a result of a concussive blow to the head. Asked to describe herself, W.J. showed a good appreciation of how she had changed since matriculating, as corroborated by her boyfriend's ratings of her. Findings such as these lend strength to the conclusion, based on experimental studies of priming, that semantic (trait) knowledge of the self is encoded independently of episodic (behavioral) knowledge.

Amnesic patients typically suffer damage to the hippocampus and related structures in the medial temporal lobes, leading to the conclusion that these structures constitute a module, or system, for encoding consciously accessible autobiographical memories. Is there a similar structure responsible for the sense of self? Recently, F. I. M. Craik and his

colleagues used positron emission tomography (PET) to image the brain while subjects rated themselves on a list of trait adjectives. As comparison tasks, subjects rated the prime minister of Canada on the same traits; they also judged the social desirability of each trait, and the number of syllables in each word. One analytic technique, statistical parametric mapping, revealed no differences in brain activation between the self- and other-ratings tasks. While this finding would be consistent with the proposition that the self is a person like any other, a partial least squares analysis showed significant self–other differences in the right and left medial frontal lobes, and the middle and inferior frontal gyri of the right hemisphere. Further studies of this sort are obviously in order. (*See* **Neuroimaging**)

So is the self in the right hemisphere? Probably not. Self-referent processing may be performed by a module or system localized in the right frontal lobe, but control is critical in these conditions, and it may well be that other-referent processing is performed by the same system, provided that the other is well liked and/or well known. Although cognitive neuroscience has generally embraced a doctrine of modularity, the neural representation of individual items of declarative knowledge is distributed widely across the cerebral cortex. Self-reference may be localized, but self-knowledge is widely distributed over the same neural structures that represent knowledge of others. (*See* **Modularity**; **Modularity in Neural Systems and Localization of Function**; **Learning and Memory, Models of**; **Memory Models**)

THE DEVELOPMENT OF SELFHOOD

Development can be viewed in two ways: ontogenetically, in terms of changes in individual organisms across the life cycle from birth to death; and phylogenetically, in terms of changes in species across evolutionary time.

Locke viewed a sense of self as essential for personhood, but nonhuman animals may also have a sense of self. In a classic study, Gordon Gallup painted an odorless, nontoxic red mark on the foreheads of anesthetized chimpanzees. In the absence of a mirror, the chimps showed no awareness that their appearance had been altered. When exposed to their reflections in a mirror, however, the animals often examined the spot in the mirror, touched the spot on their foreheads, and then inspected and smelled their fingers. They appeared to recognize a discrepancy between what they thought they looked like, and what they actually looked like – suggesting, in the process, that they possessed at least a rudimentary self-image. The same effect has been found in some orangutans and bonobos, but not in gorillas (except perhaps for the famous Koko), monkeys, and other primates, or in nonprimate species. However, it should be noted that not all chimpanzees pass the self-recognition test, and alternative means of testing may well reveal self-recognition in other species.

By the time they are 18–24 months old, most human infants also pass the mirror-recognition test. However, if the infants are shown a videotape of themselves after a delay as short as three minutes, most fail to recognize themselves on the monitor; most four-year-olds pass this more difficult test. By the age of two, then, human infants have at least a minimal sense of self, but it takes a while longer for them to develop a narrative sense of themselves as extended in time – that they are the same person now that they were a while ago. Similarly, children younger than four years old seem unable to recognize that their current knowledge and beliefs differ from those they held in the past. Interestingly, age four is about the time that children achieve a capacity for episodic memory – the ability to recognize that a current mental state is in fact a representation of a past mental state. (*See* **Theory of Mind**)

PATHOLOGIES OF SELFHOOD

Whatever the findings in infants and animals, a sense of self is part and parcel of the conscious experience of all normal human adults. However, a number of pathological conditions appear to involve disruptions in self-recognition and self-awareness. For example, some prosopagnosic patients fail to recognize their own faces as well as others, while some individuals with Capgras syndrome will identify themselves, as well as others, as imposters. Patients with frontal lobe damage often show a reduced capacity for self-reflection or impaired feelings of self-continuity. Frontal lobe damage can also impair episodic memory, in which self-reference is critical. (*See* **Prosopagnosia**; **Face Perception, Psychology of**; **Frontal Cortex**)

Within psychiatry, some theorists have suggested that schizophrenia involves a breakdown in the neural substrates of self-reflection and self-monitoring, while some dissociative disorders appear to involve a disruption of self and identity. While patients with psychogenic amnesia retain a sense of self (they simply cannot remember a period of their lives), patients with psychogenic

fugue lose their identity as well as their memories. In multiple personality disorder, both identity and autobiographical memory shift back and forth from one 'alter ego' to another. Children with autism, a pervasive developmental disorder, appear to have a limited capacity for self-reflection, personal agency, and personal ownership. Whether these deficits in social cognition are limited to the sense of self, or extend to other people as well, is a topic of much current investigation. (*See* **Disorders of Body Image**; **Autism**; **Autism, Psychology of**; **Williams Syndrome**)

SELF-KNOWLEDGE INTO ACTION

While cognitive psychology tends to study mind in the abstract, social psychology studies mind in action. Mental representations of self, others, and situations do not exist for themselves but as guides to social behavior. How we behave towards others depends not only on how we perceive them but also on how we perceive ourselves. Erving Goffman, E. E. Jones, and others have argued that people often engage in strategic self-presentation to shape others' impressions of them in an attempt to gain or retain control over the social situation. Many social interactions are characterized by what Robert K. Merton would call a 'self-fulfilling prophecy' – in which, for example, a person who believes that someone is aggressive may treat that person in a manner that evokes aggressive behavior that may not have occurred otherwise. A strong sense of self may promote strategic self-presentation, but it may also militate against others' self-fulfilling prophecies concerning oneself. If people do not define themselves as aggressive, perhaps they will be less likely to act aggressively, regardless of how they are treated. From a social-psychological perspective, then, the self is not just something that knows and is known; it is also something that one does. That is to say, the self is actively constructed and reconstructed, maintained, tested, and revised, presented and represented to others, in the course of ongoing social interaction.

Acknowledgments

The point of view represented in this article is based on research supported by NIMH Grant MH-35856 and Academic Senate Research Grants from the University of California, Santa Barbara. We thank Jennifer Beer for comments during the preparation of this article.

Further Reading

Gallagher S and Shear J (eds) (1999) *Models of the Self*. London: Imprint Academic.

Kihlstrom JF (1997) Consciousness and me-ness. In: Cohen J and Schooler J (eds) *Scientific Approaches to Consciousness*, pp. 451–468. Mahwah, NJ: Lawrence Erlbaum.

Kihlstrom JF, Beer JS and Klein SB (2002) Self and identity as memory. In: Leary MR and Tangney J (eds) *Handbook of Self and Identity* (in press). New York, NY: Guilford.

Kihlstrom JF and Klein SB (1994) The self as a knowledge structure. In: *Handbook of Social Cognition, vol. 1: Basic Processes*, 2nd edn, pp. 153–208. Hillsdale, NJ: Lawrence Erlbaum.

Klein SB (2000) A self to remember: a cognitive neuropsychological perspective on how self creates memory and memory creates self. In: Sedikides C and Brewer MB (eds) *Individual Self, Relational Self, and Collective Self*. Philadelphia, PA: Psychology Press.

Klein SB and Kihlstrom JF (1998) On bridging the gap between social-personality psychology and neuropsychology. *Personality & Social Psychology Review* **2**(4): 228–242.

Neisser U (ed.) (1994) *The Perceived Self: Ecological and Interpersonal Sources of Self-Knowledge*. New York, NY: Cambridge University Press.

Neisser U and Fivush R (eds) (1994) *The Remembering Self: Construction and Accuracy in the Self-Narrative*. New York, NY: Cambridge University Press.

Neisser U and Jopling D (eds) (1997) *The Conceptual Self in Context: Culture, Experience, Self-Understanding*. New York, NY: Cambridge University Press.

Snodgrass JG and Thompson RL (eds) (1997) *The Self Across Psychology: Self-Recognition, Self-Awareness, and the Self Concept*. New York, NY: New York Academy of Sciences.

Suls JM (1982) *Psychological Perspectives on the Self*. Hillsdale, NJ: Lawrence Erlbaum.

Suls JM (1993) *Psychological Perspectives on the Self: The Self in Social Perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Suls J and Greenwald AG (eds) (1983) *Psychological Perspectives on the Self*, vol. 2. Hillsdale, NJ: Lawrence Erlbaum.

Suls J and Greenwald AG (eds) (1986) *Psychological Perspectives on the Self*, vol. 3. Hillsdale, NJ: Lawrence Erlbaum.

Tesser A, Felson RB and Suls JM (2000) *Psychological Perspectives on Self and Identity*. Washington, DC: American Psychological Association.

Self-consciousness

Intermediate article

Jose Luis Bermúdez, University of Stirling, Stirling, UK

CONTENTS

Self-consciousness is deeply implicated in many aspects of cognition. To be self-conscious is to be aware of oneself through information that one cannot fail to recognize is about oneself.

INTRODUCTION

A creature is self-conscious to the extent that it is aware of its embodied self in such a way that it cannot fail to realize that it is itself the object of awareness. Philosophers have often discussed self-consciousness as if it were the unique property of language-using humans. There are powerful reasons for thinking that this is not the case, but whether self-consciousness is understood broadly or narrowly it remains a crucial cognitive capacity of central concern to cognitive science.

WHAT IS SELF-CONSCIOUSNESS?

Self-consciousness is primarily a cognitive, rather than an affective, state. Although the term 'self-consciousness' is often used in ordinary language to describe a particular state of hypersensitivity about certain features of one's character or appearance, in philosophy and cognitive science the expression is best reserved for a form of awareness of one's self. But what does it mean to be aware of one's self? Unsurprisingly everything depends on what one means by 'awareness' and what one takes the self to be. Let us take these two dimensions of variation in reverse order.

Historically philosophers have understood the nature of the self in many different ways. According to the various versions of dualism, the self is a purely psychological entity that is connected to a particular body but that could exist without that body. For dualists, therefore, self-consciousness consists in awareness of that psychological entity. Most contemporary philosophers and cognitive scientists are agreed that the self is an embodied person with both psychological and physical properties, although there are disagreements about the appropriate criteria of personal identity. Does the survival of the self require psychological continuity, physical continuity, or both? These disputes cut across the debate between dualists and anti-dualists, because one can hold that the survival of the self does not require physical continuity (the continued existence of the same body) without being committed to the dualist claim that the self need not be embodied at all. (*See* **Personal Identity**)

Turning now to the awareness component, we should start by distinguishing two different types of awareness – *direct awareness* and *propositional awareness*. One can be aware of something (as when I catch sight of someone walking up the garden path) or one can be aware that a particular state of affairs is the case (as when the sound of the doorbell alerts me to the fact that a visitor is at the door). In *direct awareness* the object of awareness is a particular thing. In *propositional awareness* the direct object of awareness is a proposition or state of affairs (a complex of particular things, properties, and/or relations). A further difference between direct awareness and propositional awareness is that the former is *extensional* while the latter is *intensional*. If 'JLB is directly aware of *x*' is a true report, then it will remain true whatever name referring to the same object is substituted for '*x*' in the report. Direct awareness is not sensitive to the mode of presentation of the object of awareness. Propositional awareness, however, *is* sensitive to the mode of presentation of the state of affairs that is the object of awareness. I can be aware that a state of affairs holds when it is conceptualized in one way, but be unaware that it holds under a different conceptualization. If 'JLB is propositionally aware that *x* is *F*' is a true report, it will not necessarily remain true if a co-referential name is substituted for '*x*' and/or a predicate true of the same objects for '*F*'.

Self-consciousness, therefore, can be understood either in terms of direct awareness of the self or in terms of propositional awareness that the self

has such-and-such a property, or stands in such-and-such relations. Some philosophers have maintained that there can be no such thing as self-consciousness at the level of direct awareness. David Hume maintained that the self could never be encountered in introspection:

> For my part, when I enter most intimately into what I call *myself*, I always stumble on some perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch *myself* at any time without a perception, and never can observe anything but the perception. (Hume, 1739–1740/1978, p. 252)

It is clear that this introspective report (Hume's thesis of the *elusiveness of the self*) will be telling if the self were a purely psychological entity, since introspection would then be the only possible source of direct awareness of the self. But philosophers who think that the self is essentially embodied will be untroubled by Hume's point. They will rightly point out that we do have direct awareness of the body through somatic proprioception.

In any case, it is clear that propositional awareness *about* the self does not require direct awareness *of* the self. I can be aware that a particular state of affairs holds without being directly aware of one of the constituent objects in that state of affairs, and so I can be aware that the self has certain properties without being directly aware of the self. So the truth or otherwise of the elusiveness thesis is not directly relevant to the possibility of propositional awareness of the self. Nor is it likely that propositional awareness concerning the self will be analyzable in terms of, or reducible to, direct awareness of the self. The intensional and cognitively mediated character of propositional awareness seems an insuperable obstacle to any such reductive or analytic project.

Propositional awareness of the self is in many ways more interesting than direct awareness of the self. Self-consciousness is important because of the role it plays in the cognitive economy. Self-conscious subjects think about, and react to, the world in distinctive and characteristic ways that are not available to non-self-conscious subjects. Self-consciousness makes possible certain types of inference and reflection, and it does this because of the distinctive types of proposition that it makes available.

But what are these distinctive types of proposition? Self-consciousness makes available to the subject thoughts that are about the thinker of that thought, but not all thoughts that are about the person thinking them qualify as self-conscious. A thought might be about its thinker without the thinker being aware of that fact. So, for example, I might think that the last person arriving at the party is ill-mannered without realizing that I am that person. A genuinely self-conscious thought is about the thinker of that thought in a way that does not leave any room for the thinker to fail to recognize that the thought concerns him. This is part and parcel of the distinctive functional role of self-consciousness. It is mirrored by the linguistic fact that any token of the first person pronoun 'I' always refers to its producer. Self-conscious thoughts would naturally be expressed with sentences involving the first person pronoun.

Self-conscious thoughts can be based on a range of different sources of information. Some of these sources can provide information *either* about the self *or* about other people. Testimony is a case in point. I can learn facts about myself by being told them by others, in the same way as I learn facts about anything else. But there are other sources of information about the self that provide information purely about the self. These sources of information are such that, if we know from them that somebody has a particular property, we *ipso facto* know that we ourselves have that property. Introspection is an example. If I know through introspection that someone is currently thinking about self-consciousness then I know that I myself am thinking about self-consciousness. Introspection cannot provide information about anybody other than me. This does not mean that introspection (and other comparable sources of information) cannot be mistaken. They certainly can, but they do not permit a certain type of error. Judgments made on the basis of them cannot be mistaken about who it is that has the property in question. Such judgments are *immune to error through misidentification relative to the first person pronoun* (Shoemaker, 1968), an epistemological property that they inherit from the information sources on which they are based.

Self-conscious thoughts that are immune to error in this sense (such as the thought that I am in pain, where this is based on information from pain receptors) are clearly more fundamental than those that are not. They reflect ways of finding out about ourselves that are exclusively about the self and that do not require identifying an object *as* the self. Self-conscious thoughts that are not immune to error through misidentification must be analyzed in terms of those that are immune, because they will involve identifying an object as the self, and any such identification must be immune to error through misidentification on pain of an

infinite regress. For this reason influential accounts of self-consciousness, such as those of Shoemaker (1963, 1968) and Evans (1982), have attributed a fundamental role to the phenomenon of immunity to error through misidentification.

PHILOSOPHICAL ISSUES CONCERNING SELF-CONSCIOUSNESS

The Scope of Awareness

There are four different ways in which a cognitive state properly describable as awareness can emerge. One can be aware of something (whether an object or a state of affairs): (1) by having information about it and acting accordingly; (2) by perceiving it; (3) by having beliefs about it; or (4) by having knowledge of it. How do these differ?

One can have information about something without perceiving it. There are many examples of this among neuropsychological disorders, including blindsight. The performance of blindsight patients on certain matching and other tasks shows that they are capable of performing certain perceptual discriminations in their blindfield, and hence that at some level they are picking up visual information about a portion of the distal environment that they are not perceiving. Similarly, one can perceive something without having beliefs about it. Most simply, one might not believe the content of one's perception. Finally, one can have beliefs about something without having knowledge of it. The belief might be mistaken, or fail to be securely enough grounded to qualify as knowledge. (*See* **Blindsight**; **Epistemology**)

Each of these different ways of understanding 'awareness' will yield a different conception of self-consciousness. The strictest and narrowest conception identifies it with self-knowledge (the knowledge of propositions about the self whose natural linguistic expression would involve the first person pronoun). Only slightly less strict would be the view that self-consciousness involves believing the appropriate sort of propositions about the self. Both of these views have the consequence that self-consciousness is only available to creatures who can have beliefs, and on many conceptions of belief this significantly narrows the field of self-conscious subjects. This has struck some philosophers as undesirable.

Understanding awareness in terms of either information pick-up or perception significantly broadens the scope of self-consciousness with respect to its emergence in human development and to where we might expect to find it in the animal kingdom. There are important questions, however, about how these types of awareness of the self should be understood. It is easy to see how they can support a form of direct awareness of the self. But do they admit propositional awareness of the self? If so, what is the content of that awareness? It presumably differs from the content of beliefs and/or knowledge about the self. But in what ways? At this point appeal might be made to the notion of *nonconceptual content*. Whereas the content of beliefs, and propositional attitudes in general, is conceptual (in the sense that it can only be attributed to creatures possessing the concepts required to specify it), some authors have found it helpful to postulate the possibility of representing the world in ways that are not constrained by the conceptual repertoire in this sense (Evans, 1982; Cussins, 1990; Peacocke, 1992; Bermúdez, 1998). (*See* **Mental Content, Nonconceptual**)

As a way of motivating a more inclusive conception of self-consciousness, consider that self-conscious thoughts play a distinctive functional role in the cognitive economy. They have, for example, immediate implications for action (Castañeda, 1969; Perry, 1979). Whereas I may contemplate with equanimity the thought that the worst-performing philosopher in the department will shortly be ejected from the department, as soon as I realize that *I* am the person whose job is on the line I will be galvanized into action. One might wonder, therefore, whether all motivated action might not require some form of self-conscious thought. Let us call this the *thesis of essential self-consciousness*. If motivated action really does require some form of self-conscious thought, then *either* only creatures who are capable of knowledge and/or belief are capable of motivated action, *or* the domain of self-consciousness must be extended until it is co-extensive with the domain of agents. The perceived need to accommodate the second of these options is a prime motivation for taking one of the broad readings of self-consciousness outlined above.

It is natural to point out that what is shown by the above example and the many others like it is that it is not possible for me to act on the basis of knowledge or belief that 'a is F' where 'a' is a term or expression that refers to me but when I do not know that it does. But it does not follow from this that all motivated action requires knowledge or belief that would be expressed in the (genuinely self-conscious) form 'I am F'. If I am hungry and see food then I will act accordingly. But it is hard to see where the self-conscious thought comes in. What is see is what I desire, namely, the food over there.

Defenders of the thesis of essential self-consciousness will suggest that this misrepresents the nature of perception, since perception of the external world has an irreducible first-person component (Gibson, 1979; Bermúdez, 1998). Perception is essentially perspectival and egocentric, most obviously in vision but also in the sense of touch. The world is not presented in perception as an abstract arrangement of objects, but rather as an array of objects that stand in certain spatial relations to the perceiver. The world is perceived from a point of view, where a perceiver's point of view is tied to his possibilities for acting upon the distal environment.

It is important for both the philosophy of self-consciousness and the philosophy of perception to provide an account of the content of visual perception that reflects its perspectival and egocentric nature. Such an account will stress the differences between the content of perception and the content of propositional attitudes such as belief, such as: the unit-free way in which distances are presented in visual perception; the fact that the spatiality of the distal environment is perceived on an egocentric frame of reference; and the way in which the embodied self actually appears in the field of vision and the content of tactile perception. Nonetheless, although we should not expect the contents of visual perception to be propositions of the sort that serve as contents of propositional attitudes, the status of exteroceptive perception as a form of propositional awareness of the self emerges when one remembers that it essentially provides information about the relations between the embodied self and objects in the distal environment. (*See* **Mental Content, Nonconceptual**)

Can we broaden the scope of self-consciousness even further? Can the appropriate type of awareness be derived from nonperceptual information pick-up? Somatic proprioception seems the most plausible place to look. Many of the sources of proprioception, such as the information about limb position and movement provided by joint position sense and the information about the body's orientation and state of balance derived from the vestibular system, have seemed to many to be better described at the level of information pick-up, rather than at the level of perceptual awareness – not least because there does not seem to be a sensory dimension to the information they provide. Yet they also seem to be providing information about properties of the embodied self – intrinsic properties in the case of joint position sense and relational in the case of the vestibular system. It may be, however, that these two information sources are best understood, not in isolation, but rather as embedded within somatic proprioception as a whole, and it has been argued that somatic proprioception does indeed provide a form of perceptual awareness of the body (see the essays in Bermúdez *et al*., 1995 for various perspectives on this issue).

Awareness of Self and Awareness of the World

As has emerged above, the functional role of self-conscious thoughts is both distinctive and important. We have so far concentrated on just two aspects of that functional role – the role of information sources that are immune to error through misidentification in generating first-person thoughts, and the immediate implications that such thoughts have for action. But the functional role of a given type of thought also includes its relations to other types of thought. They will be the subject of this section.

Self-consciousness is essentially a contrastive notion. Subjects are aware of themselves relative to, and as distinct from, other members of a contrast class of either other physical objects or other psychological subjects. In view of this it is natural to adopt what I shall term the Interdependence Thesis, according to which a creature's capacity for self-consciousness is directly proportional to his capacity to represent the external world.

A classic expression of the Interdependence Thesis is Kant's claim, defended in the section of the *Critique of Pure Reason* entitled 'The Transcendental Deduction of the Categories', that self-consciousness both depends upon and makes possible the perception of a spatio-temporal world composed of continuously existing objects causally interacting in law-like ways. The form of self-consciousness he is discussing (the unity of apperception that he describes in terms of the 'I think' being able to accompany all my representations) is largely formal – essentially the awareness, with respect to each member of a series of thoughts and experiences, that it is one's own. The interdependence emerges from the two-way links between the unity of apperception and the possibility of applying the categorial concepts whose applicability Kant took to define the objectivity of the world. In this sense, Kant's version of the Interdependence Thesis is closely linked to his distinctive version of transcendental idealism. (*See* **Kant, Immanuel**)

Philosophers such as P. F. Strawson (1966) and Gareth Evans (1982) have more recently attempted

to defend a version of the Interdependence Thesis that is not committed to a Kantian transcendental idealism. Like Kant, however, they adopt what I earlier described as a narrow conception of self-consciousness – that is, the thesis that self-consciousness involves and requires beliefs and/or knowledge about the self. For Strawson and Evans the Interdependence Thesis holds because the capacity to have a suitably generalized understanding of the first person pronoun (Evans) or to conceptualize the distinction between experience and what it is experience of (Strawson) requires the ability to formulate judgments reflecting a conception of the embodied self as located within an objective world possessing certain very general features. For Evans, possessing a mastery of the first-person concept that is integrated with thought about the rest of the world in a suitably productive and systematic way requires the ability to conceive of oneself 'from the third-person point of view' as an objective particular in a unified spatio-temporal world. For Strawson, the ability to distinguish appearance from reality within the realm of experience requires the ability to ascribe experiences to oneself as a continuously existing particular.

Whether these arguments are sound or not, an important question emerges for those who have sought to defend a broader conception of self-consciousness. How, if at all, can a version of the Interdependence Thesis be motivated once we move below the level of self-knowledge and beliefs/knowledge about the world? Clearly, at the level of perceptual awareness and information pick-up it is not appropriate to construe the Interdependence Thesis in terms of connections between *judgments* about the self and *judgments* about the world. Yet unless some version of the Interdependence Thesis holds at these more primitive levels, it is unclear how they can support genuine forms of self-consciousness at all, given that the Interdependence Thesis reflects the essentially contrastive nature of self-consciousness.

Some materials for answering this challenge are offered in Bermúdez (1998), where it is shown how primitive nonconceptual and prelinguistic forms of self-consciousness can be appropriately contrastive. Analyzing visual perception following the ecological approach of J. J. Gibson (1979) reveals the exterospecific and propriospecific dimensions of visual perception and how the dynamism of visual perception emerges from their interaction. Similarly, somatic proprioception provides a broadly perceptual awareness of the limits of the body as a physical object responsive to the will, and hence as clearly demarcated from all other physical objects. By the same token, it is possible for a creature to have a sense of itself as following a single path through space-time, and hence to possess a (nonconceptual) point of view on the world, as manifested in its memories and navigational understanding of space, rather than in high-level beliefs and judgments.

SELF-CONSCIOUSNESS AND THE COGNITIVE SCIENCES

Many discussions of cognitive science make a distinction between central and peripheral processes. This is often tied to the distinction between modular and nonmodular processes. Central processes are not encapsulated, domain-specific, mandatory, and fast in the manner of peripheral processes such as those responsible for early vision and phonological analysis. On what I have been calling the narrow conception of self-consciousness, according to which self-consciousness is at root a matter of having certain distinctive types of thought about oneself, self-consciousness will have to be analyzed at the level of central processing. Those who follow Fodor (1985) in thinking that cognitive science can have little to say about central processes will draw appropriately pessimistic conclusions. More plausibly, it seems natural to think that the capacity to entertain self-conscious beliefs will depend upon some form of metarepresentational capacity, and hence should be analysed as an element in what is often termed theory of mind. (*See* **Modularity**)

There exists a rich cognitive scientific literature in this area, which one might expect to be highly informative on the ontogenesis of higher-level self-consciousness, as well as on what happens when the mechanisms that subserve higher-level self-consciousness break down. With respect to the former of these, one might expect the capacity for higher-level self-consciousness to emerge as part of the overall 'theory of mind' package at more or less the age of four. As far as pathology is concerned, Christopher Frith's influential analysis suggests that schizophrenia should be understood as a breakdown in the mechanisms that permit awareness of oneself as the author of one's thoughts and beliefs – and hence that it is, at least in part, a deficit of higher-order self-consciousness (Frith, 1992).

The scope for dialogue between philosophy and the cognitive sciences with respect to self-consciousness becomes much greater on what I have termed the broader conception of self-consciousness. Most straightforwardly, the different areas of cognitive science can be expected to

play a crucial role in exploring the more primitive dimensions of self-consciousness (Neisser, 1993). We have already briefly seen how Gibson's ecological approach to visual perception can be of great help in understanding the way in which visual perception incorporates an element of self-awareness. One would also expect the cognitive sciences to provide much-needed clarification of the different information systems and channels subserving somatic proprioception.

There is also scope for influence in the other direction. Accounts within cognitive science need to be sensitive to phenomenological factors revealed by philosophical accounts of the nature of self-consciousness. A single example will make the point. I suggested above that one way in which the Interdependence Thesis might be understood at the level of perceptual awareness and information pick-up is through a creature's possessing a (non-conceptual) point of view on the world in virtue of having certain navigational abilities. It seems plausible to stress the role of the following capacities in underwriting a creature's grasp of the spatial organization of its environment and its location within that environment: (1) the capacity to think about different routes to the same place; (2) the capacity to keep track of changes in spatial relations between objects caused by its own movements relative to those objects; and (3) the capacity to think about places independently of the objects or features located at those places. Recognizing the centrality of these cognitive capacities influences how we interpret some important recent work on animal representations of space and their neurophysiological coding. Chapters 5 and 6 of Gallistel (1990) defend the thesis that all animals from insects upwards deploy cognitive maps with the same formal characteristics in navigating around the environment. Gallistel argues that the cognitive maps that control movement in animals all preserve the same set of geometric relations within a system of earth-centred (*geocentric*) coordinates. These relations are metric relations. The distinctive feature of a metric geometry is that it preserves all the geometric relations between the points in the coordinate system. Gallistel's thesis is that, although the cognitive maps of lower animals have far fewer places on them, they record the same geometrical relations between those points as do humans and other higher animals.

Without, of course, wishing to challenge Gallistel's central thesis that all animal cognitive maps from insects up preserve geometric relations, it none the less seems wrong to draw the conclusion that all animals represent space in the same way. Just as important as how animals represent spatial relations between objects is how they represent their own position within the object-space thus defined. And it is here, in what we should think of as not just their awareness of space but also their self-conscious awareness of themselves as spatially located entities, that we see the major variations and the scale of gradations that the theorists whom Gallistel is criticizing have previously located at the level of the cognitive map.

References

Bermúdez JL (1998) *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Bermúdez JL, Marcel AJ and Eilan N (eds) (1995) *The Body and the Self*. Cambridge, MA: MIT Press.

Castañeda H-N (1969) The phenomeno-logic of the I. Reprinted (1994) in Cassam Q (ed.) *Self-Knowledge*. Oxford, UK: Oxford University Press.

Cussins A (1990) The connectionist construction of concepts. In: Boden M (ed.) *The Philosophy of Artificial Intelligence*. Oxford, UK: Oxford University Press.

Evans G (1982) *The Varieties of Reference*. Oxford, UK: Oxford University Press.

Fodor J (1985) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Frith C (1992) *The Cognitive Neuropsychology of Schizophrenia*. Brighton, UK: Lawrence Erlbaum.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Hume D (1739–1740/1978) *A Treatise of Human Nature*, edited by LA Selby-Bigge, revised by PH Nidditch. Oxford, UK: Clarendon Press.

Neisser U (ed.) (1993) *The Perceived Self*. Cambridge, UK: Cambridge University Press.

Peacocke C (1992) *A Study of Concepts*. Cambridge, MA: MIT Press.

Perry J (1979) The problem of the essential indexical. *The Problem of the Essential Indexical and Other Essays*. Oxford, UK: Oxford University Press.

Shoemaker S (1963) *Self-Knowledge and Self-Identity*. Ithaca, NY: Cornell University Press.

Shoemaker S (1968) Self-reference and self-awareness. *Journal of Philosophy* **65**: 555–567.

Strawson PF (1966) *The Bounds of Sense*. London: Methuen.

Further Reading

Campbell J (1994) *Past, Space and Self*. Cambridge, MA: MIT Press.

Cassam Q (ed.) (1994) *Self-Knowledge*. Oxford, UK: Oxford University Press.

Shoemaker S (1986) Introspection and the self. *Midwest Studies in Philosophy* **10**: 101–120.

Shoemaker S (1996) *The First-Person Perspective and Other Essays*. Cambridge, UK: Cambridge University Press.

Sense and Reference

Intermediate article

Margaret Reimer, University of Arizona, Tucson, Arizona, USA

CONTENTS

Many philosophers of language believe that sense and reference are two essential components of linguistic meaning. The reference of an expression is (roughly) that which the expression is used to talk about; the sense of an expression is (roughly) the criterion an entity must satisfy in order to be the reference of that expression. The intuitive, nontechnical notion coming closest to the notion of 'sense' is that of meaning.

WHAT ARE SENSE AND REFERENCE?

What, if anything, can we say about words that is uncontroversial? Probably nothing – at least if one is to judge by the relevant philosophical literature. Virtually everyone, however, would agree that words have meanings – in *some* sense of 'meanings'. But in what does the meaning of a word consist? Is the meaning of a word nothing more than what it refers to? Is the meaning of the expression 'water' simply water? Or is there something more to linguistic meaning? Those who answer the latter question affirmatively, often claim that what Frege (1892) called 'sense' is this 'something more'. Because reference is generally considered to be the more fundamental of the two notions, let us begin our discussion of sense and reference with a few introductory remarks on reference.

It seems clear that words, at least some of them, *refer*: they somehow attach to the particular objects and individuals that we use them to talk about. Indeed, how else would talk about the world be possible? When a word refers, what it refers to is called its 'reference'. Reference is thus a property of words, at least of certain words. Thus, for example, the reference of the proper name 'George W. Bush' is George W. Bush; the reference of the definite description 'the tallest person ever to live' is that very person – whoever he or she may be; the reference of the natural kind term 'tiger' is the class of all tigers; the reference of the artifactual kind term 'pencil' is the class of all pencils. The basic idea is not particularly deep or subtle, but there are some tricky points to bear in mind.

First, there are words that seem to refer *when we use them*, yet fail to refer when considered in the abstract, as it were. What, for instance, does the demonstrative 'that' refer to? What about the pronoun 'she'? What about the expressions 'here' and 'there'? These sorts of expressions, known in the literature as 'indexicals', are expressions whose reference depends upon the particular context in which they are used, where context is construed as including speaker, hearer, time, and place. (See Kaplan, 1989.) Thus, I may, while pointing to my sister, say 'She is a vet' and, while pointing to my daughter, say 'She is my daughter'. In both cases, the pronoun 'she' refers – but to whom it refers to depends upon the particular context in which it is used. (Some have argued that natural kind terms are also indexical expressions. See Putnam, 1975.)

Second, one must be careful to distinguish between what a word(s) refers to and what a speaker uses a word(s) to refer to. Often the two coincide, but sometimes they come apart. Suppose, for instance, that I see in the distance what I take to be a man in a gray jacket, though in fact the individual I perceive is a woman in a mauve jacket, behind whom is a man in a gray jacket. Suppose further that I an unable to see the man in the gray jacket. Then, when I say to you 'The man in the gray jacket is coming this way', I am referring to the woman (provided my intention is to draw your attention to her). My words, however, arguably refer to the man if indeed they refer to anything at all. (See Kripke, 1977.) The distinction between speaker reference and word reference leads to the question: which kind of reference is primary? Do words refer primarily and speakers only derivatively – or is it the other way around? To ask this question is to ask whether speaker reference is to be analyzed in terms of word reference, or the other

way around. While many philosophers of language appear to have assumed the primacy of word reference, others (Strawson, 1950; Donnellan, 1966) have argued for the primacy of speaker reference.

At any rate, if there are words that refer (even if only derivatively), we are going to need an explanation of how this is possible. One might initially think: that's easy. A word, at least a word of the referring sort, refers to whatever speakers in the relevant linguistic community use that word to refer to. Thus, for instance, the word 'telephone' refers to … yes, *telephones*, and it refers to such things because that is what speakers of the relevant linguistic community (here, English) use that word to refer to. But there is a problem with this view: although correct, it is either circular or woefully incomplete. The view is circular if the notion of reference is the same in the definition as in the term to be defined. And, in that case, it is no more informative than the claim that calculators are things we use to do calculations. The view is incomplete if the notion of reference in the definition is distinct from the notion of reference being defined; for although the definition appears to offer an explanation of word reference in terms of speaker reference, it does not say in what speaker reference consists. One question is thus answered at the expense of generating another, equally interesting and important, question that remains unanswered.

This brings us to the notion of sense. Sense corresponds at least roughly to the intuitive, nontechnical notion of meaning, according to which competent speakers of the language know the meanings of the words in that language. So construed, sense is thought by some philosophers to play a variety of different theoretical roles. Indeed, according to Frege (1892) and his followers, the notion of sense plays four distinct roles. First, it serves as the bearer of 'cognitive significance', and thus explains the difference in 'cognitive value' between sentences like (1) and (2):

Hesperus is Hesperus. (1)

Hesperus is Phosphorus. (2)

(The names 'Hesperus' and 'Phosphorus' are co-referring, both referring to the planet Venus.) Intuitively, while (1) is trivial, (2) is informative. The informativeness of (2) is to be explained by the hypothesis that the expressions 'Hesperus' and 'Phosphorus', though identical in reference, are different in sense or 'cognitive significance'. (The sense of the former might be something like brightest heavenly body in the evening sky; the sense of the latter might be something like brightest heavenly body in the morning sky.) Second, the sense of an expression determines its reference. An expression refers to whatever somehow satisfies its sense. The sense of an expression might thus be thought of as providing a criterion for reference: a set of conditions that an object/individual must satisfy in order to emerge as the expression's referent. Third, senses are the objects of thought: the sense of a declarative sentence simply is the thought it expresses. And fourth, senses are identified as the referents of embedded sentences. Thus, in a sentence like 'Fred believes that Bush is president', the referent of the embedded sentence 'Bush is president' is identical to the sense of the unembedded sentence. Obviously, then, Frege expected senses to do a great deal of theoretical work – according to some, more than could reasonably be expected by a single notion. (See Schiffer, 1990)

Because the topic at hand is sense *and reference*, let us look more closely at the distinction between these two notions.

THE SENSE/REFERENCE DISTINCTION

Frege (1892) is certainly the most famous exponent of the sense/reference distinction. Indeed, he is arguably its original proponent – though others before him proposed similar distinctions. Mill (1843), for instance, argued for a distinction between expressions that merely denote and those that both denote and connote. What an expression denotes is what it refers to, what it signifies; what an expression connotes is what it means, what it implies. But Mill was certainly not the first to appreciate this basic distinction between what a word refers to and what it means. Indeed, Mill attributes the original distinction to the schoolmen of the Middle Ages.

To understand the sense/reference distinction as conceived by Frege, one must understand its initial motivation. It is relatively uncontroversial that words, at least some words, refer to things, and that the reference of a word is somehow relevant to its meaning. Indeed, according to what is often called the 'naive' theory of meaning, the meaning of a word *just is* what it refers to. But there appear to be some fairly obvious problems with this view. Consider the following two sentences:

Jack the Ripper is Jack the Ripper. (3)

Jack the Ripper is Prince Albert Victor. (4)

Suppose that we adopt the view that meaning is nothing more than reference. On such a view, an

expression gets its meaning from the fact that it refers. Then, given that 'Jack the Ripper' and 'Prince Albert Victor' are co-referring, they mean the same thing, and we thus seem forced to conclude that sentences (3) and (4) mean the same thing – that they express the same 'proposition'. But do they? Surely, they do not appear to. After all, while the first is trivially true (analytic or tautologous), the second is anything but. While Queen Victoria would have been unfazed by the putative truth of (3), she might have vehemently denied the putative truth of (4). How can we explain the intuitive difference between such sentences – sentences of the form $a = a$ and those of the form $a = b$?

Such considerations, considerations having to do with the difference in cognitive significance (roughly, informativeness) between sentence pairs like (1)/(2) and (3)/(4), led Frege to the view that all so-called singular terms, terms that refer – or purport to refer to definite objects – invariably have senses. (This is true of 'Santa Claus' and 'Sasquatch', both of which arguably only purport to refer.) For Frege, the distinction between sense and reference is crucial. Thus, although the senses associated with the names 'Jack the Ripper' and 'Prince Albert Victor' may determine the same reference (Queen Victoria's eldest son), they have different senses: they present (or represent) that individual in distinct ways. While the sense of the first might be expressed by the description 'slayer of assorted London prostitutes in the late 1880s', that of the second might be expressed by the description 'the eldest son of Queen Victoria'. This is not, however, generally thought to be an essential feature of senses – they that be expressible via some sort of linguistic expression, like a definite description. It is just that sometimes they are so expressible, and in such cases it is especially easy to see that the senses are indeed distinct. (See, for instance, Evans, 1983 and Searle, 1983.) At any rate, since it is the sense of an expression that enters into the proposition expressed, the hypothesis that 'Jack the Ripper' and 'Prince Albert Victor' have different senses explains the difference in cognitive significance between the propositions expressed by sentences (3) and (4). For senses are generally thought to be compositional, in so far as the sense of a complex expression (such as a sentence) is going to be determined by the senses of the simple expressions out of which it is formed. Thus, since the subject terms of (3) and (4) have different senses, the senses of the sentences themselves will be different – those sentences will (in other words) express different propositions; they will express (in Fregean terms) different thoughts.

Frege's sense/reference distinction enables us to solve the puzzle posed by sentences pairs like (1)/(2) and (3)/(4): that is, it explains why these sentences differ in terms of cognitive significance. It also allows us to solve puzzles associated with other sorts of sentences, as we are about to see.

ARGUMENTS FOR THE DISTINCTION

The usual arguments for the sense/reference distinction are of two basic sorts. According to one, words refer and so we need something to explain how this happens. Sense provides such an answer – an expression refers to whatever satisfies the sense that it expresses. I suppose this might be construed by some as a sort of 'argument to the best explanation'. (Although of course not everyone thinks that senses *are* the best explanation – see below.) A second type of argument is what might be called a 'puzzle' argument. Various logical puzzles arise on the 'naive' view that reference – which seems crucial to meaning – is *all* that there is to meaning. By positing senses, the puzzles are effectively solved. In addition to the problem posed by pairs of identity sentences like (1)/(2) and (3)/(4), there are others, typically divided into three groups: (i) propositional attitude ascriptions, (ii) sentences containing empty names, and (iii) true negative existentials, which constitute a special subclass of (ii). Consider the following three sentences:

Fred believes Cicero, but not Tully, was Roman. (5)

Sasquatch was last sighted in Tucson, Arizona. (6)

The Easter Bunny does not exist. (7)

Intuitively, (5) seems to attribute noncontradictory beliefs to Fred, (6) seems to be meaningful, and (7) seems to be true. But suppose that meaning is no more than reference – that the meaning of an expression is simply what it refers to. Then, all of these intuitions are misleading. For on the meaning = reference view, (5) does ascribe contradictory beliefs to Fred – he believes of one and the same individual that he is both Roman and not Roman; (6) is meaningless, assuming that a sentence with a meaningless part cannot itself be meaningful, and (7) – if meaningful – must be false. For in order to be meaningful, the subject term must refer; but if it refers, then presumably its referent (the Easter Bunny) does exist. If, on the other hand, we suppose that expressions like 'Cicero', 'Tully', 'Sasquatch' and 'the Easter Bunny' have senses that are distinct from their references (if any), then we

can explain the intuitions without much difficulty. For it is the sense of an expression – as opposed to its reference – that enters into the proposition expressed, into what is said.

PROBLEMS WITH THE DISTINCTION

Although the sense/reference distinction seems to be an intuitive one, and the arguments for it have considerable force, some have argued that it is an ill-conceived distinction. The best-known objections to the distinction are those of Kripke (1980) and Putnam (1975).

Both Kripke and Putnam take issue with the idea that sense determines reference. Kripke's objections focus on proper names, while Putnam's focus on natural kind terms. Kripke argues, in effect, that if the sense of a proper name is identical to the property or properties the user of the name associates with that name, then sense is not what determines reference. Thus, for instance, if I declare that Einstein was a brilliant man then I refer to the famous German physicist, even if the (only) property I associate with that name – say, the inventor of the atomic bomb – picks out someone else (Oppenheimer). Kripke of course realizes that he is going to need to provide some alternative explanation of how proper names refer, and he does: he proposes his so-called Causal Theory of Reference, according to which users of a name are able to refer to the name's referent in virtue of their membership in a causal chain of communication, originating in an act of reference-fixing – a baptism of sorts. On this view, conceptual information (or misinformation) about a name's referent is irrelevant to the determination of nominal reference. (For some criticisms of this view, see Evans, 1973 and Searle, 1983.)

Putnam makes a similar point but applies it to natural kind terms. Suppose that sense determines reference, and that senses are (as Frege supposed) abstract conceptual entities grasped by competent speakers of the language. Now consider the natural kind terms 'puma' and 'jaguar'. I can use both of these terms to refer and, in so doing, I refer to pumas and jaguars respectively. But not knowing much about such felines, the concepts that I associate with each of the two expressions are not sufficient to distinguish pumas from jaguars. Indeed, perhaps I associate with both terms the same concept – something like *ferocious, sleek-bodied feline with a short glossy coat*. Thus, if senses are construed à la Frege as abstract conceptual entities grasped by competent speakers of the language, then they are not what determine reference. For if that were so, then 'puma' and 'jaguar', as I use those terms, would not refer (as they in fact do) to different kinds of feline; they would refer to the same kind of feline, if they referred at all. What, then, does determine reference – the reference of natural kind terms in particular? Putnam claims that two factors are relevant to such determination: the way the world is and social features captured by what he calls a 'division of linguistic labor', according to which, in using a natural kind term, speakers defer to the relevant experts (zoologists, in the case of felines), saying in effect that the term refers to whatever the experts would say that it refers to.

RELATION OF SENSE AND REFERENCE TO COGNITIVE SCIENCE

Sense is clearly going to be relevant to issues in cognitive science, while reference is not. The reasons for this are fairly obvious. Sense is supposed to explain cognitive significance; in particular, it is supposed to explain why it is that certain sentences express the particular thoughts they appear to express. Sense thus explains, at least indirectly, our linguistic behavior as well as our reactions to the linguistic behavior of others. Thus, for instance, sense might be invoked to explain why Queen Victoria would have been unfazed (though perhaps perplexed) by an utterance of (3), though outraged by an utterance of (4). And sense might also be invoked to explain why an utterance of (3) would, in most circumstances, seem pointless. Any philosopher of mind who thinks of thought as language-like is going to have to face the same puzzles that Frege did, and is going to have to respond to the challenges that Kripke and Putnam have posed to the Fregean distinction between sense and reference. (See Fodor, 1994 for an idea of how these puzzles might be treated by a philosopher of mind who thinks that thought is linguistic.)

References

Donnellan K (1966) Reference and definite descriptions. *Philosophical Review* **75**: 281–304.

Evans G (1973) The causal theory of names. *Aristotelian Society: Supplementary Volume* **47**: 187–208.

Fodor G (1994) *The Elm and the Expert*. Cambridge, MA: MIT Press.

Frege G (1892/1952) On sense and reference. In: Black M and Geach P (eds) (1952) *Translations from the Philosophical Writings of Gottlob Frege*, pp. 56–78. Oxford, UK: Basil Blackwell.

Kaplan D (1989) Demonstratives. In: Almog J, Perry J, and Wettstein H (eds) *Themes from Kaplan*, pp. 481–614. Oxford: Oxford University Press.

Katz J (1994) Names without bearers. *Philosophical Review* **103**: 1–39.
Kripke S (1977) Speaker's reference and semantic reference. In: French P, Uehling T, and Wettstein H (eds) *Contemporary Perspectives in the Philosophy of Language*, pp. 6–27. Minneapolis: University of Minnesota Press.
Kripke S (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Mill JS (1843) *A System of Logic*. London, UK: Longmans.
Putnam H (1975) Meaning and reference. *Journal of Philosophy* **70**: 699–711.
Russell B (1905) On denoting. *Mind* **14**: 479–493.
Schiffer S (1990) The mode of presentation problem. In: Anderson C and Owens J (eds) *Propositional Attitudes*, pp. 249–268. Stanford: CSLI.
Searle J (1983) *Intentionality*. Cambridge, UK: Cambridge University Press.
Strawson P (1950) On referring. *Mind* **59**: 320–344.

Further Reading

Bach K (1987) *Thought and Reference*. Oxford, UK: Oxford University Press.
Devitt M (1982) *Designation*. Cambridge, MA: Cambridge University Press.
Dummett M (1973) *Frege: Philosophy of Language*. London, UK: Duckworth.
Evans G (1982) *The Varieties of Reference*. Oxford, UK: Oxford University Press.
Recanati F (1992) *Direct Reference*. London, UK: Blackwell.
Salmon N (1986) *Frege's Puzzle*. Cambridge, MA: MIT Press.

Simulation Theory

Intermediate article

Joe Cruz, Williams College, Williamstown, Massachusetts, USA
Robert M Gordon, University of Missouri, St Louis, Missouri, USA

CONTENTS

The simulation theory is an account of our everyday ability to attribute mental states and predict and explain human behavior. It has been developed both as an empirical hypothesis in cognitive science and as an account of mental concepts in the philosophy of mind.

WHAT IS THE SIMULATION THEORY?

The simulation theory (ST) is an account of our everyday ability to make sense of the behavior of others. One crucial element of this ability is the identification and attribution of inner mental states that generate action, especially propositional attitudes such as beliefs or desires. The successful 'mindreading' of mental states allows us to predict and to explain what others do, and makes possible the rich social dynamic that pervades human life.

Conceived most broadly, ST maintains that one represents the mental activities and processes of others by mental simulation, i.e., by generating similar activities and processes in oneself. For example, one anticipates the product of another's theoretical or practical inferences from given premises by making inferences from the same premises oneself. In more complex simulations, one imaginatively adopts the circumstances of the target and then uses one's own mental apparatus to generate mental states and decisions. Computationally, this exercise of imagination is usually represented as feeding pretend inputs into one's own decision-making processes, taking these processes 'offline' so that they do not issue forth in real behaviors.

Some proponents of ST go further and claim that many of the concepts of mental states that we deploy in understanding other human beings are fundamentally linked to our possession of those same mental states. Some prominent accounts attempt to shed light on the conceptual transformation involved in refashioning our first-person concepts in such a way that they can be deployed in the third person.

While ST is related to the empathetic or *verstehen* approaches to explanation in the social sciences that were prominent in the twentieth century, most researchers who currently work on ST do not operate directly within that theoretical framework. In its contemporary forms, ST has often been developed with its principal rival, the theory theory (TT), as an explicit foil. TT maintains that the mental terms and concepts used in understanding, predicting and explaining human behavior derive from a folk theory of the mind. A closer historical source for contemporary ST was the debate in the philosophy of mind over the status of this putative theory. According to one view, known as eliminativism: (1) mental states like beliefs and desires are the posits of a folk theory of the mind; and (2) this theory is radically false. The conclusion drawn by the eliminativist is that the faulty folk theory ought to be rejected in favour of some more scientifically respectable theory such as one derived from neuroscience.

Before the advent of ST, most critics of eliminativism focused on the defensibility of its second tenet. Important articles by Jane Heal (1986), Robert Gordon (1986) and Alvin Goldman (1989) challenged the first tenet by setting out the ST alternative to TT. If it could be shown that mental terms did not derive their intelligibility from their role in a folk theory, then the eliminativist's conclusion would become suspect. Since this important impetus, ST has attracted interest in its own right and has become somewhat independent of concerns over eliminativism.

Some philosophers believe that ST sheds light on traditional topics such as the problem of other minds, referential opacity, broad and narrow content, and the peculiarities of self-knowledge. ST has

had a substantial impact on research into 'theory of mind' in developmental psychology, as well as on branches of philosophy outside the philosophy of mind, especially aesthetics and the philosophy of the social sciences.

SIMULATION THEORY VERSUS THEORY THEORY

One of the dominant explanatory patterns within cognitive science and the philosophy of mind has been to construe a mental capacity as subsumed by a theory of the domain of the capacity. The idea is that domains such as folk physics, folk biology, and intuitive statistics are treated as areas of knowledge in which the layperson's judgments are the results of applying a theory. The results of the application of the theory constitute our spontaneous conscious judgments about cases that seem amenable to the theory. Such theories are usually thought to be representations of law-like generalizations. In some domains, at least, they may be characterized as a set of 'platitudes' tacitly deployed in thinking about problems and circumstances in our world. The layperson need not be aware of using a theory to make a judgment, but, presumably, could generate at least some of the platitudes that inform his or her judgments.

In keeping with this widely endorsed strategy in cognitive science, theory theorists claim that we possess a body of tacit knowledge that governs our judgments about the mental states of others. The theoretical posits of this theory will be mental states like beliefs and desires, and the transitions between the mental states will be described by the theory as mental processes such as inference.

There is a diversity of opinion among theory theorists on the nature and origin of the putative common-sense theory. Some claim that the theory is learned; others that it is innate. Theory theorists also face a question regarding the modularity of the common-sense 'theory of mind'. Some claim that the theory is a distinct cognitive module; others that it is continuous with the system of representations that constitute theories of other, non-mental domains. What unifies theory theorists is the view that attributing inner states and making sense of the behavior of others is carried out by a capacity that deploys knowledge encoded in a theory.

The most straightforward sense in which ST is opposed to TT is that simulation theorists deny that our capacity to attribute mental states is subsumed by a body of knowledge about the minds of others. Rather, our own mental processes are treated as a manipulable model of other minds. Such simulation would typically require indexical adjustments, such as shifts in spatial, temporal, and personal points of view, to place oneself in the other's physical and epistemic situation in so far as it differs from one's own. One may also compensate for the other's reasoning capacity and level of expertise, if possible, or modify one's character and outlook as an actor might, to fit the other's background and behavioral history. With these adjustments, the attributer might enter mental states that differ from those he or she would have in the target's situation. Even when simulation is insufficient for making decisions in the role of the other, it might allow one to discriminate between those options likely to be attractive to the target and those likely to be unattractive. Accordingly one would be prepared for the former actions and surprised by the latter.

Moreover, most simulationists are happy to grant that, in some cases, we will develop general rules of thumb or heuristics for attributing mental states. These may be called on to generalize the results of a simulation to cover the same target at future times, or a class of targets, such as those who share the conventions of a particular culture, who 'as a rule' behave in a certain way in certain circumstances. Still, simulationists deny that general, 'theoretical' considerations play a fundamental part in attributing mental states. ST is often characterized as 'process driven', because it is a cognitive process that is generating the output of the simulation, with little or no influence from general information about minds.

An analogy may be helpful here. If we wished to predict a future state of the solar system, we might appeal to a theory that expresses law-like generalizations about the motions of the planets. So, we might appeal to the theories articulated in a contemporary textbook on astronomy, or, more in keeping with our interest here in folk theories, we might appeal to a set of platitudes about the motions of celestial objects of the kind articulated by the ancients. This would be a sensible approach, but it is not the only way we could successfully carry out the prediction. If we could build a reasonably accurate physical model (an orrery), we could advance this model the correct number of cycles and read from it a future state of the solar system. Depending on how versatile our model was, we might even be able to experiment with counterfactual starting states; or, with a still more sophisticated model, or a digital simulation, we might even adjust the orbits of planets, and their number, to model a range of planetary systems quite unlike our own.

This analogy gives us some insight into the difference between theory-driven and process-driven accounts of a domain. ST can be viewed as the proposal that we use our own mental states and processes – our perceptual, cognitive, motivational and emotional systems – as a model like an orrery.

VERSIONS OF SIMULATION THEORY

It is obviously desirable that any relevant disparities between simulator and target should be removed or offset in some way. If the behavior to be predicted or explained is crucially dependent on beliefs, desires or emotions not shared by the simulator, then the simulator must either adopt the appropriate pretend-beliefs, pretend-desires or pretend-emotions, or compensate by a heuristic rule. However, this proviso says nothing about the nature of belief, desire or emotion. For instance, it is consistent with (but certainly does not imply) a functionalist account of belief. In a well-known box diagram, Stich and Nichols (1992) portray ST as the empirical hypothesis that the same belief–desire system that generates one's own decisions and actions also generates our predictions of the decisions and actions of others, adding for this purpose a pretend-belief generator and a pretend-desire generator. Some simulationists find this portrayal too restrictive, claiming that it commits ST to a questionable conception of mental states and possibly also a mistaken understanding of the dependence of actions on these states. These proponents of ST conceive our everyday ascriptions of belief and other mental states as part of an explanatory enterprise quite unlike the attempt to fill in the 'boxes' of a functional theory like those commonly developed in cognitive science. Our ascriptions specify the 'internal states' of a system only in the sense of attempting an essentially first-person glimpse into a subject looking out on the world.

There is also disagreement among simulationists on another front. Some hold that to ascribe mental states to others by simulation, one must already be able to ascribe mental states to oneself by introspection, and that to do this one must already possess the relevant mental state concepts. On this view, simulation is understood as essentially an application of the argument from analogy. Others attempt to build on the 'subject looking out on the world' idea of mental state ascription. They hold that in such ascriptions, whether concerning oneself or another, one is saying something about the world, albeit in a way that is relativized to a particular 'point of view'. Rather than resting on an analogy between what lies 'inside' two individuals, this account assumes that, unless there is evidence to the contrary, all subjects look out on one and the same world.

ARGUMENTS FOR SIMULATION THEORY

A number of arguments have been put forward in favour of the simulation theory, three of which are outlined below.

Parsimony

The most important distinguishing feature of folk psychology, according to many writers, is the central and essential role it gives to the semantic content of the states it posits, particularly the propositional or sentential 'objects' of propositional attitudes such as beliefs, desires, and intentions. Most theory theorists try to accommodate this feature with the hypothesis that folk psychology comprises laws or principles that quantify over this content, connecting, for example, what someone believes and desires to what that person chooses to do. Moreover, the connections are said generally to mirror the semantic relations that hold among these contents, particularly relations that can be represented abstractly by rules of logic and rational argument such as modus ponens and the practical syllogism. Thus the theory theory posits an internal store of causal laws corresponding to these rules.

However, in so far as the store of causal generalizations mirrors the set of rules to which our own thinking typically conforms, ST appears to render it otiose. For whatever those rules are, our thinking continues to conform to them within the context of simulation, unless, of course, adjustments are made to accommodate evident differences. In short, we can use our own reasoning as a model of the reasoning of beings that reason the way we do. In the light of this alternative, it is argued, the hypothesis that people must be endowed with a special stock of laws corresponding to rules of logic and reason appears unmotivated and unparsimonious.

Other Uses of Simulation

The procedure that ST posits as essential to the common-sense methodology for predicting and explaining behavior also appears to be – with modifications – essential to emotional empathy, and important if not essential to ethical evaluation. Even if one didn't think simulation essential to common-sense explanation and prediction, one would probably have to posit such a procedure

anyway to account for empathy and ethical evaluation.

Explanation of Children's Errors in Predicting Behavior in 'False Belief' Situations

According to ST, the mature capacity for explaining and predicting the behavior of others requires capacities for imaginative pretense of at least three kinds: counterindexical pretending, which recenters the egocentric map (I am spatially or temporally somewhere else in the world, or I am someone else); counterfactual or propositional pretending, in which the world itself is altered in imagination (for example, this banana is a telephone, or dinosaurs roam the boulevards of Paris); and what might be called purposive pretending, in which alternative goals are adopted (for example, the putative goals of a mother, or the goals of an opponent in a game). These capacities are clearly evident in most children before their third birthday, and they are typically combined in role play.

To explain and predict a great deal of the human behavior they are likely to encounter in real life or in stories, young children can probably get by with a relatively simple employment of these and similar imaginative abilities. For example: while Sally and Anne are playing together, Anne grabs the marble in Sally's box and places it in her basket. Taking the role of Sally in this simple scenario, a child should have no trouble deciding where to look for her marble: in Anne's basket. However, suppose that in the story (or in a scene witnessed in real life) Sally is away when Anne takes her marble. When she returns, where does she look for her marble? Taking the role of Sally and deciding where to look, the child would be misled by a simple use of pretense. To get the correct prediction (she will look in her box, where she left it), and at the same time to maintain 'objectivity' (for example, to recognize that she won't actually find the marble there), the child would have to feed contrary premises into two distinct lines of reasoning: 'objectively', the marble has been moved to Anne's basket and therefore can be found there, not in Sally's box; 'subjectively' – and for the purpose of deciding what to do in the role of Sally – it hasn't been moved from Sally's box, and therefore can still be found there. According to ST, until children are capable of such compartmentalized reasoning – and of knowing when to use it – they can be expected to make incorrect predictions in complex scenarios in which behavior is likely to be based on a false belief.

Numerous experimental studies (beginning with Wimmer and Perner (1983)) have confirmed that in fact children do generally make these incorrect predictions until about the age of four. Although ST is not unique in offering an explanation of such errors, its explanation appears more compelling than typical 'theory' explanations, such as the hypothesis that children lack the 'belief' part of the theory of mind until the age of four (Gopnik, 1993; Wellman, 1990; Gopnik and Meltzoff, 1997).

SIMULATION THEORY AND COGNITIVE SCIENCE

The simulation theory has a bearing on a range of disciplines and methods in the cognitive sciences. Consequently, there has been significant research on ST within artificial intelligence (AI), neuroscience, and cognitive psychology. In order to show how ST has been pursued in these fields, it is useful to develop further the 'process driven' conception of ST introduced earlier. One way of refining the claim that simulations are process-driven is to take ST as the hypothesis that essentially the same set of mechanisms is called upon to provide two different competences. One of these competences is the intelligent control of behavior. This would include, among other things, the capacity to make inferences from beliefs to new beliefs and the capacity to make decisions on the basis of beliefs. The second competence is the anticipation and comprehension of intelligently controlled behavior, by predicting the underlying inferential and decision-making processes.

There are thought to be testable consequences of this 'double-duty mechanism' construal of ST. Researchers in AI have claimed that the same models that provide an account of practical reasoning (and of other dynamics between propositional attitudes) can efficiently be adapted to perform mental state attribution. These computational efficiency arguments have in turn sometimes been used as arguments in favour of ST. Both traditional AI programming methods and neural network techniques have been employed in research on ST along these lines.

In AI research, the double-duty hypothesis is a functional claim. The idea is that the same mental program is implicated both in practical reasoning and in mental state attribution, with changes allowed for taking the system 'offline' or feeding in pretend states. So, even if it turned out that different parts of the physical computational system were responsible for practical reasoning and mental

state attribution, the AI approach to ST would remain valid.

If, on the other hand, the double-duty conception of ST is not just understood functionally, but includes a commitment to the commonality of the underlying neural substrate, we arrive at another avenue of cognitive science research on ST. Such research would seek to show that there was a shared neuronal mechanism that is responsible for practical reasoning and mental state attribution. This approach is analogous to a fruitful line of research on vision and visual imagery. These two capacities appear to share substantial portions of the underlying neural substrate.

There is emerging evidence for the existence of 'mirror neurons' in humans and other primates (Fadiga *et al*., 1995; Rizzolatti *et al*., 1996). Single-cell recordings in macaque monkeys and magnetic stimulation techniques in humans show that these mirror neurons exhibit increased activity when another primate is observed performing some characteristic action, such as grasping an object. This is relevant for ST because these are the same neurons that show increased activity in the first-person performance of that action. This research is far from showing that sophisticated cognition is subsumed by double-duty mechanisms, but it is suggestive. Some simulation theorists maintain that this research reveals an evolutionary precursor to the capacity of assuming a different perspective that ST requires (Gallese and Goldman, 1998).

Evolutionary considerations have been lurking in the background of much empirical research in this domain. Thus, elements of cognitive ethology have been thought to bear on ST, and vice versa. Indeed, it was in their research on primate behavior that Premack and Woodruff (1978) introduced the term 'theory of mind'. Though efforts to develop ST within cognitive ethology have been limited, it does present another potential source of data.

A POSSIBLE TEST OF THE SIMULATION THEORY

If it can be shown that incorrect mental state attributions are in some cases best explained by claiming that the attributer lacks some specific and perhaps surprising background information about mental agents, then it seems less plausible to think that the attributer is simulating. Some critics of ST claim to have uncovered experimental evidence for just this.

The logic of such experiments is thought to be as follows. In order for a simulation to be successful, the operation of the attributer's mental processes must be substantially similar to the operation of the target's mental processes. The attributer does not need to know about the vagaries of human psychology because, by hypothesis, his or her own mental mechanism is subject to those same vagaries. One test of ST, according to this line of reasoning, would be to select some surprising feature of our mental life to see whether subjects can successfully attribute mental states to others where the mental processes in question exploit the surprising feature. If the attribution fails, this suggests that either the attributer's mind does not share the surprising feature, or that it is the specific lack of information about the surprising feature that is generating the failure to attribute a correct mental state. Neither of these results would be friendly to ST. The first possibility is against the very spirit of ST, while the second possibility implicates a theory of mental states.

In one attempt to pursue this line of criticism, researchers explored the counterintuitive effect (reported by Langer (1975)) whereby subjects demand more money in exchange for a lottery ticket that they have chosen than for a ticket that has just been given to them. Although there has been some concern about the reproducibility of Langer's initial results, the experiment is thought to show something unexpected about human psychology, namely, that subjects are more attached to items that they have chosen than they are to identical items that they have not chosen. Nichols *et al*. (1996) asked naive subjects to predict the outcome of the Langer experiment without actually putting them under the experimental conditions themselves. It was assumed that if, in making the prediction, the subjects simulated making a decision themselves under each of the two experimental conditions, they would be prone to the same surprising effect, valuing the ticket they had chosen themselves more than the one they had not. Thus, they would predict correctly. However, the subjects did not predict correctly, and Nichols *et al*. concluded that they were not simulating. Several problems have been noted with the experiment conducted by Nichols *et al*. A more discriminating set of experiments has been reported, with mixed results for ST (Perner *et al*., 1999).

One general objection concerns the logic of such tests. The account presented above neglects the following possibility: the experimental effect in question, such as the higher value placed on items one has chosen, may be partially due to aspects of processing that merely imagining does not, or perhaps even cannot, capture. Consider the two lines in the Mueller–Lyer illusion: remove the

arrowheads, and you are likely to judge the lines to be of equal length; merely *imagine* the arrowheads removed, and you are still likely to judge them unequal. Analogously, one may predict what people will do in situation S by simulation, in the sense of imagining being in S and deciding what to do, and yet fail to replicate all the processing that would occur if one actually were in S. If the analogy holds, then tests of ST should take account of a possible gap between imagining and replicating at a subpersonal level.

References

Fadiga L, Fogassi L, Pavesi G and Rizzolatti G (1995) Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* **73**: 2608–2611.

Gallese V and Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* **2**: 493–501.

Goldman A (1989) Interpretation psychologized. *Mind and Language* **4**: 104–119.

Gopnik A (1993) How we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**: 1–14.

Gopnik A and Meltzoff A (1997) *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gordon R (1986) Folk psychology as simulation. *Mind and Language* **1**: 158–171.

Heal J (1986) Replication and functionalism. In: Butterfield J (ed.) *Language, Mind and Logic*, pp. 135–150. Cambridge, UK: Cambridge University Press.

Langer E (1975) The illusion of control. *Journal of Personality and Social Psychology* **32**: 311–328.

Nichols S, Stich S, Leslie A and Klein D (1996) Varieties of off-line simulation. In: Carruthers D and Smith E (eds) *Theories of Theories of Mind*, pp. 39–74. Cambridge, UK: Cambridge University Press.

Perner J, Gschaider A, Kühberger A and Schrofner S (1999) Predicting others through simulation or by theory? A method to decide. *Mind and Language* **14**: 57–79.

Premack D and Woodruff G (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* **4**: 515–526.

Rizzolatti G, Fadiga L, Matelli M *et al*. (1996) Localization of grasp representations in humans by PET: 1. Observation vs. execution. *Experimental Brain Research* **111**: 246–252.

Stich S and Nichols S (1992) Folk psychology: simulation or tacit theory? *Mind and Language* **7**: 35–71.

Wellman H (1990) *The Child's Theory of Mind*. Cambridge, MA: Bradford/MIT Press.

Wimmer H and Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**: 103–128.

Further Reading

Carruthers P and Smith P (eds) (1996) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Davies M and Stone T (eds) (1995a) *Folk Psychology: The Theory of Mind Debate*. Cambridge, MA: Blackwell.

Davies M and Stone T (eds) (1995b) *Mental Simulation: Evaluations and Applications*. Cambridge, MA: Blackwell.

Goldman AI (1995) Simulation and interpersonal utility. *Ethics* **105**: 709–726.

Gordon R and Barker J (1994) Autism and the 'theory of mind' debate. In: Graham G and Stephens G (eds) *Philosophical Psychopathology*, pp. 163–181. Cambridge, MA: Bradford.

Heal J (1998) Co-cognition and off-line simulation: two ways of understanding the simulation approach. *Mind and Language* **13**: 477–498.

Speech Acts

Intermediate article

Robert M Harnish, University of Arizona, Tucson, Arizona, USA

CONTENTS

Speech acts are acts performed in or by speaking (or more generally, in or by using language).

INTRODUCTION

The theory of speech acts lies at the intersection of the theory of action and the theory of language, and historically the developers of speech act theory have tried to relate these two aspects. More recently, cognitive science has provided an opportunity, if not a framework, for unifying some of this work. Cognitive science is commonly viewed as the interdisciplinary study of cognition, the relevant disciplines being neuroscience, psychology, computer science, philosophy, linguistics and anthropology. Speech acts have been studied from the perspective of each of these disciplines, especially the last five.

PHILOSOPHY

The modern theory of speech acts begins with the work of J. L. Austin in the 1950s (Austin, 1962), although before this time important and interesting (albeit fragmentary) work had been done by Frege and Wittgenstein. Frege argued that sentences of natural language contain two kinds of information: information about what he called the 'force' of the sentence, and information about what he called the 'sense' of the sentence. The sense of a sentence is (roughly) the condition the world must meet for the sentence to be true. The sense of a complex expression is derived from the senses of its constituents by fitting argument names to function names – in the simplest case, singular terms to predicates. Frege did not give a general characterization of force, but he connected it with grammatical mood, and to speaking 'with requisite seriousness'. The three kinds of force he focused on were: assertive, interrogative (questions), and directive (requests and commands) (Harnish, 2001). One consequence of this view is that sense cannot be identified with linguistic meaning, because assertions, such as 'It is raining', and interrogatives, such as 'Is it raining?', have the same sense, but differ in force (and so meaning). To understand a sentence it is necessary to grasp both its sense and its force. Because Russell and the early Wittgenstein were almost exclusively concerned with logical form, they tended to neglect the non-assertive uses of language. A good example of this was Wittgenstein's early 'picture' theory of meaning. Later Wittgenstein exchanged the picture metaphor for the tool metaphor: like tools, words have many different uses, and these uses are organized in systems of what he called 'language games' (Wittgenstein, 1958). Speaking a language is like making moves in, and so following the rules of, language games. Frege's non-assertive uses of language had been rediscovered, but not systematically investigated or connected to the structure of sentences. Wittgenstein's antitheoretical stance left the study of language use fragmented until the 1950s when Austin noted that there were many uses of language which had the superficial appearance of stating facts, but really did something different: 'I do', 'I quit', 'I promise I will be there' (Austin, 1961). These are 'performatives' and their job in the language is to do something, not just to say something. As acts they have characteristic ways of succeeding and failing. Austin's famous 'doctrine of infelicities' distinguished two kinds of failure 'misfire' (the act is purported, but void, i.e. no act gets performed – this may be due, for example, to misinvocation, misexecution, misapplication of procedures, or flaws and hitches in the procedures) and 'abuse' (the act is performed, but defective – this may be due, for example, to insincerity or breaches of procedure).

Austin soon came to believe, in the face of examples such as 'I state that flounders snore', that the distinction between saying and doing

could not be rigidly maintained (but see (Recanati, 1987)), and that every sentential utterance has a performative aspect, and in the William James Lectures of 1955–1956 he moved from the special theory of performatives to the general theory of speech acts (Austin, 1962). In this theory, 'locutionary' acts are acts of 'saying something in the full normal sense', i.e. with a certain sense and reference. 'Perlocutionary' acts are acts of affecting the hearer's thought or action in some way (persuading). Austin only characterized 'illocutionary' acts, the third and most important category (which includes acts of stating, questioning, commanding, promising, etc.) by examples and by citing certain of their general features, such as being governed by conventions beyond those of grammar, sense and reference, and being capable of being made explicit by a performative prefix. He did, however, develop a preliminary taxonomy of illocutionary acts which has served as the basis for subsequent taxonomies in philosophy:

Verdictive: An act that delivers a finding based on evidence (e.g. an acquittal or estimate).
Exercitive: A decision for or against something, or a decision that something is to be so (e.g. an order or recommendation).
Commissive: An act that commits the speaker to a certain course of action (e.g. a promise or vow).
Expositive: An exposition of views (e.g. an affirmation or emphasis).
Behabitive: A reaction to other people's behavior (e.g. an apology or expression of thanks).

A theoretical breakthrough came with J. R. Searle's constitutive rule theory of speech acts (Searle, 1969), in which Austin's doctrine of infelicities was integrated with Wittgenstein's notion of language games and rule following and Frege's distinctions between force, reference and predication. Searle criticized Austin's notion of a locutionary act, and replaced it with a more abstract 'propositional' act, which allowed him to factor illocutionary acts with content into an illocutionary force (*F*) and a propositional content (*P*). Searle proposed that the illocutionary force component of a (nondefective) illocutionary act, for example a promise, can be analyzed into four kinds of condition:

1. Propositional content: A future act of the promiser.
2. Preparatory: The promiser can do the act, and the promisee would prefer the promiser to do it than not to do it.
3. Sincerity: The promiser intends to do the future act.
4. Essential: The utterance declares an obligation to do the future act.

These act-specific conditions were supplemented by general background ('input-output') conditions for linguistic communication, as well as by a more controversial Gricean condition (see below) on saying something and meaning it, thereby also capturing Frege's idea that force is connected with 'requisite seriousness'. If these conditions are met in uttering an expression, such as 'I promise I will be there', then the speaker has promised. And if the speaker has (nondefectively) promised, then he or she must have met these conditions. Moreover, sentences of a natural language can typically be semantically analyzed into a device which indicates the illocutionary force of the sentence and a device which indicates the (propositional) content of the sentence. In the above example, 'I promise' is the device for indicating a promise (*Pr*), and 'I will be there' is the device for indicating what is promised (*P*). The constitutive rules for the use of the device *Pr* are derived from the above conditions. For instance, the associated rules say: (1) utter *Pr* only if conditions 1–3 are met (these are 'regulative' rules, which regulate preexisting forms of behavior), and (2) the utterance of *Pr* counts as meeting condition 4 (this is a 'counts-as' rule, which defines a new form of behavior). Searle also follows Frege in viewing the semantics of sentences as compositionally determined by the semantics of their constituents, and proposes propositional acts of referring and predicating governed by constitutive rules similar to 1–4 above. We have arrived full circle at Frege's idea that sentences are analyzable into force and content; and we now have an articulated theory of illocutionary force, requisite seriousness, reference and predication. Searle criticized Austin's taxonomy of illocutionary acts for being unprincipled and failing to distinguish illocutionary acts from illocutionary verbs (see (Wierzbicka, 1987)), and he offered his own alternative taxonomy of illocutionary acts (Searle, 1979):

Assertive: An act that commits the speaker to something being the case (e.g. a statement or claim).
Directive: An attempt to get the hearer to do something (e.g. a request or command).
Commissive: An undertaking to do something (e.g. a promise or vow).
Expressive: An expression of a psychological state of the speaker (e.g. a congratulation or apology).
Declaration: An act that brings about the state of affairs mentioned in the propositional content (e.g. adjourning a meeting or resigning a position).

Unlike Austin, but following Grice (see below), Searle also systematically investigated indirect and nonliteral uses of language, such as: 'Want to go to the movies tonight?' 'I have an exam

tomorrow' (indirect: 'No, I can't'); or 'Sally is a block of ice' (nonliteral: 'Sally is cold and unresponsive'). Searle and Vanderveken have also investigated aspects of illocutionary logic (Searle and Vanderveken, 1985) (see also Vanderveken, 1990). Although Searle describes the above rules as 'semantic' rules, he did not formulate a general theory of how they could confer meaning on expressions. W. Alston has modified Searle's theory of illocutionary acts using the notion of 'taking responsibility for satisfying a condition' as central in such a way that it can form the foundation for a use theory of meaning (Alston, 2000).

At the same time as Austin and Searle were developing the theory of speech acts in terms of conventions and rules of language use, H. P. Grice was developing a theory of language based primarily on the notions of a speaker's beliefs, desires, and reflexive communicative intentions (Grice, 1989). Grice analyzed 'meaning something' (by uttering something) as intending to produce some effect in a hearer by means of the hearer's recognition of that intention. Furthermore, what the speaker means is specified by the intended effect: either a belief, for transfers of information (e.g. assertions), or an intention to act, for attempts to get people to do things (e.g. commands). Finally, what an expression means in the language is related to what speakers mean in uttering it. Grice briefly discussed how these two central categories of meaning and speech acts might be related to sentential mood (declarative, imperative, interrogative), but his main interest was in analyzing the 'total signification of an utterance', and here he drew the distinction between what is 'said' (related to Austin's 'locutionary' act and Searle's 'propositional' act) and what is conventionally and conversationally 'implicated':

> Conventional: 'She's poor but she's honest.' (Implicature: poverty contrasts with honesty.)
> Conversational (particular to the situation): 'He likes his colleagues at the bank and hasn't been to prison yet.' (Implicature: he might be tempted to take some money.)
> Conversational (general, usually implied by the form of words): 'He's meeting a woman.' (Implicature: she's not his mother, sister or close 'Platonic' friend.)

Conversational implicatures depend on spoken exchanges being 'cooperative' ventures between speaker and hearer, where cooperation often amounts to conforming to various 'maxims of conversation', such as: *quantity* (make your contribution just as informative as is required for the current purposes of the exchange); *quality* (do not say what you believe to be false or what you lack adequate evidence for); *relation* (be relevant); and *manner* (be perspicuous, brief and orderly, and avoid obscurity and ambiguity).

Speakers conversationally implicate something (*Q*) in saying something (*P*) by 'flouting' a maxim, i.e. by causing the hearer to reason as follows: the speaker is obeying the maxims, but in order for this to be true something must be being communicated that is not what is being said (*P*), and the likely candidate for this is *Q*. Davis (1998) has examined the explanatory adequacy of Grice's theory of implicature. Others have questioned Grice's suggestion that what is meant consists only of what is said and what is implicated. The utterance 'I've had breakfast' generally communicates that I've had breakfast *today*, but that information does not seem to be a part of the semantics of the sentence, nor does it seem to be implied by any flouting of the maxims. So where does it come from and what is its status? Sperber and Wilson (Sperber and Wilson, 1986) and Recanati (Recanati, 1989) opt for including it in an expanded conception of what is said (and an expanded role for pragmatics in determining it), whereas Bach opts for a narrower Gricean conception of what is said (Bach, 1994). All agree that mechanisms of 'completion' and 'expansion' are required to supplement Grice's account. Finally, Grice's theory of generalized implicature has recently been refined (Levinson, 2000). Grice envisaged a unified theory of meaning, saying, speech acts and communication, but, like Austin, did not live to complete it himself.

In one direction, Grice's ideas about meaning and speech acts were modified and developed by S. Schiffer, who first argued for an alternative analysis of speaker meaning, then, modifying D. Lewis (Lewis, 1969), worked out an analysis of convention, and connected it to speaker meaning by analyzing what it is to be the conventional meaning of a linguistic expression. Finally, building on his characterization of speaker meaning, as well as ideas of P. Strawson (Strawson, 1964), he proposed a taxonomy of illocutionary acts which divided them into an 'assertive' class and an 'imperative' class. Each of these classes was divided into three subclasses which were defined in terms of the form of the belief or desire, and the reasons being offered to form that belief or desire.

In another direction, Grice's ideas about speech acts and communication were developed by Bach and Harnish, who extended and schematized the inference pattern behind conversational implicatures and extended it to speech acts in general (Bach and Harnish, 1979). They studied inferential

strategies for direct, literal, nonliteral and indirect communication, along with the presumptions which drive such inferences. They proposed a taxonomy which divided illocutionary acts into two major categories, conventional (acts performed by satisfying a convention) and communicative (acts performed by having a reflexive intention recognized). Each of these categories was further subdivided as follows:

Conventional

Effective: An act that effects changes in institutional status (e.g. vetoing a bill).

Verdictive: A judgment that has official consequences (e.g. calling a runner out).

Communicative

Constative: An expression of belief (e.g. an assertion).

Directive: An expression of desire (e.g. a request).

Commissive: An expression of intention to undertake an obligation (e.g. a promise).

Acknowledgment: An expression of feelings or compliance with social expectations (e.g. an apology).

LINGUISTICS

In the 1970s linguistics and philosophy were closely related: philosophers of language were making regular appeals to the work of Chomsky and transformational grammar; while linguists, especially those involved in the 'generative semantics' movement, were appealing to work in logic and the philosophy of language, in particular to Austin's notions of performatives and felicity conditions. J. Katz and P. Postal argued, on the basis of linguistic data, for the Fregean idea that information concerning (illocutionary) force is coded into the mood structure of sentences (Katz and Postal, 1964). They proposed that this coding takes the form of abstract morphemes, with semantic interpretations, which trigger syntactic transformations and related lexical and phonological changes characteristic of moods. Declarative sentences are the unmarked case, but interrogative sentences share an abstract question morpheme *Q* which indicates a question, and imperative sentences share an abstract imperatival morpheme *I* with roughly the sense of 'the speaker requests'. J. R. Ross argued, in effect, that these markers are actually higher embedding clauses in the deepest structure of the sentence, which usually remain unspoken and unwritten, and have the form of Austin's performative prefixes (such as 'I promise') (Ross, 1970). Consequently, the seemingly simple sentence 'Floyd sucks eggs' actually has the structure of a complex sentence: 'I assert to you that Floyd sucks eggs.' Similar proposals were made for interrogatives and imperatives (Sadock, 1974); but the 'higher performative' analysis of force had serious problems (Gazdar, 1979), and was abandoned with generative semantics in the 1980s, though much of their data remains unaccounted for to this date.

Katz proposed to incorporate force into the alternative 'interpretive semantics' framework (Katz, 1977). Like Frege, he thought that sentences contain speech act information (which he called 'propositional type') and propositional content. The propositional content is divided into referential information and predicative information (which Katz calls the 'condition'). When the type of the sentence is assertive, it converts the condition into a truth condition, thus correctly connecting assertion with truth. When the type of the sentence is requestive, it converts the condition into a compliance condition, thereby correctly connecting request with compliance. Katz proposes a provisional alternative categorization of illocutionary forces as encoded into the meanings of sentences. These include: requestives, advisives, expressives, permissives, obligatives, expositives and stipulatives. In the 1980s, linguistics, inspired by Chomsky's 'government-binding' framework, turned to comparative studies, especially in syntax. The search for linguistic universals spread naturally to illocutionary force, and the question arose: how similar are languages in the way they encode speech act information? J. Sadock and A. Zwicky, working with a conception of mood that involves the conventional pairing of sentence structures with speech acts, surveyed 23 languages from a variety of families and found a number of general similarities for the major moods (declarative, imperative, interrogative), but very few for the minor moods (tags, exclamations, optatives, etc.) (Sadock and Zwicky, 1985). Generalizing on their work, Harnish proposed a theory of the major moods in terms of sentence structure (including intonation), direction of fit, and speech act potential (Harnish, 1994).

PSYCHOLOGY

Psycholinguistics blossomed with the ascendancy of Chomsky's idea that a linguist's grammar records facts of linguistic 'competence', which itself plays a role in linguistic 'performance', but until the 1970s most of this work was concentrated on speech, words and syntax. With the rise of interest in semantics and speech act theory, some psychologists began to investigate these aspects of language processing and acquisition. Psycholinguistic methodology typically uses reaction times and error

measures to study underlying psychological mechanisms and representations, and the aspects of the theory of speech acts most amenable to these methods were the inferences underlying indirect and nonliteral communication, and conversational implicature. H. Clark and his colleagues have conducted a number of experiments to test whether the stages of inference proposed by speech act theorists are psychologically realistic. In one study, Clark and Schunk (1980) noted, following Brown and Levinson (Brown and Levinson, 1978), that one reason for speaking indirectly is politeness, and they proposed to treat requests as polite to the extent that the cost to the hearer of complying with the request diminishes or the benefit to the hearer increases. On the hearer's side, Clark and Schunk suggested the 'attentiveness hypothesis' (AH): The more attentive the hearer is to all aspects of the speaker's remark, within limits, the more polite it is. In a pair of experiments it was found that AH could account for a significant amount of the correlation in these rankings, and to that extent these experiments support the view that the literal meaning is being processed in such cases. Much of this research is reviewed in (Clark, 1992). In another pair of experiments, Gibbs gave subjects sentences such as 'Must you open the window?' embedded in two different contexts: one that biased their understanding towards the literal and direct interpretation, and one that biased their understanding towards the indirect message (Gibbs, 1979). Puzzlingly, it was found that subjects took less or equal time to judge the indirect interpretations in context compared with the time they took to judge the literal ones. Gibbs has also conducted experiments on figurative language and metaphor (Gibbs, 1994), and others to determine how pragmatic processes determine what is said versus what is implicated (Gibbs and Moise, 1997).

COMPUTER SCIENCE

Winograd's work was a landmark not only in artificial intelligence, but in computational linguistics (Winograd, 1972). Although limited to a microworld of blocks and surfaces, it successfully processed fluent, flexible natural language dialogue in all the major moods. This work led to a bifurcation in the goals of natural language processing studies: building a natural language interface to communicate with computational systems, and building a computational model of human language processing. To achieve either of these goals would require speech act information to be expressed, and recognized, in language (Allen, 1987). The most common and influential approaches to this have been 'plan based'. Cohen and Perrault proposed that a theory of speech acts might be achieved by modeling requests, assertions and questions in a general planning system using operators with associated preconditions and effects defined in terms of the speaker's and hearer's beliefs and wants (goals) (Cohen and Perrault, 1979). Plans are sequences of operators for achieving goals, given certain preconditions, and speakers (agents) achieve goals by constructing and executing these plans. Searle's analysis of illocutionary acts (see above) was the starting point of their investigation, but constraints on compositionality and point of view analysis eventually moved the theory in a more Gricean direction (see above). Perrault and Allen modified and extended this analysis of (direct) illocutionary acts to indirect acts of requesting, including questions as a special case, using Gricean principles of cooperation (plan inference) and rationality (plan construction) (Perrault and Allen, 1980). (*See* **SHRDLU**)

ANTHROPOLOGY

One of the distinctive contributions of anthropology (and other social sciences) to cognitive science is the cross-cultural perspective. One interesting and ambitious research program was the Cross-cultural Speech Act Realization Project (Blum-Kulka, House and Casper, 1988). In these studies a 'discourse completion test' was given to speakers of a variety of languages (Danish, English, French, German, Hebrew, Spanish), belonging to a variety of cultures (American, Argentinian, Australian, Canadian, Danish, English, Israeli), who had to complete a dialogue involving characters of varying social distance (e.g. friends, strangers) and relative status (e.g. professor and student) either requesting or apologizing. These completed dialogues were then scored and analyzed to assess the connections between, for example, the indirectness and politeness of the request, the social roles of the characters, and the language or culture of the person completing the dialogue. In general, politeness was correlated with indirection, but different languages and cultures responded to the dialogue situations quite differently, in ways that are difficult at present to explain.

EXTENSIONS AND CRITIQUES OF SPEECH ACT THEORY

It is characteristic of, though not required by, speech act theory that individual speech acts,

however analyzed or categorized, are regarded as being performed in the utterance of single sentences (or fragments of sentences functioning as sentences). But we normally speak in connected series of sentences which form discourses or conversations. Speech act theory has been extended to include discourse and conversation as structured series of individual speech acts (Labov and Fanshel, 1977). Greetings invite greetings; questions invite answers; apologies, requests and offers invite acceptance (or denial); statements invite agreement (or disagreement); accusations invite denial (or acceptance); goodbyes invite goodbyes. But the idea that speech acts are useful categories of analysis for naturally occurring speech, discourse and conversation has been seriously questioned (Levinson, 1981; Schegloff, 1988). There are two main objections, which are related. First, speech act categories are too coarse and rigid to capture the fluid context-dependence of most conversational moves. Second, there is no principled correspondence between speech acts and individual sentences. The alternative proposal is that naturally occurring speech must be analyzed at a lower level, using a different vocabulary from that of high-level speech act theory. Some speech act theorists think that these objections are not essential to the speech act perspective, which was never intended to be exhaustive (Searle, Parret and Verschueren, 1992). Geis has attempted to combine speech act theory with conversational analysis and related work on computational speech acts (Geis, 1995).

References

Allen J (1987) *Natural Language Understanding*. Menlo Park, CA: Benjamin/Cummings.

Alston W (2000) *Illocutionary Acts and Sentence Meaning*. Ithaca, NY: Cornell University Press.

Austin JL (1961) *Performative Utterances: Philosophical Papers*. Oxford: Oxford University Press.

Austin JL (1962) *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Bach K (1994) Conversational impliciture. *Mind and Language* **9**: 124–162.

Bach K and Harnish R (1979) *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Blum-Kulka S, House J and Casper G (1988) *Cross Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.

Brown F and Levinson S (1978) Universals in language use: politeness phenomena. In: Goody E (ed.) *Questions and Politeness*. Cambridge, UK: Cambridge University Press.

Clark H (1992) *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Clark H and Schunk (1980) Polite responses to polite requests. *Cognition* **8**: 111–143.

Cohen P and Perrault C (1979) Elements of a plan-based theory of speech acts. *Cognitive Science* **3**: 177–212.

Davis W (1998) *Implicature*. Cambridge, UK: Cambridge University Press.

Gazdar G (1979) *Pragmatics: Implicature, Presupposition and Logical Form*. New York, NY: Academic Press.

Geis M (1995) *Speech Acts and Conversational Interaction*. Cambridge, UK: Cambridge University Press.

Gibbs R (1979) Contextual effects in understanding indirect requests. *Discourse Processes* **2**: 1–10.

Gibbs R (1994) *The Poetics of Mind*. Cambridge, UK: Cambridge University Press.

Gibbs R and Moise J (1997) pragmatics in understanding what is said. *Cognition* **62**: 51–74.

Grice HP (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Harnish R (1994) Mood, meaning and speech acts. In: Tsohatzidis SL (ed.) *The Foundations of Speech Act Theory*. London: Routledge.

Harnish R (2001) Frege on mood and face. In: Kenesei I and Harnish R (eds) *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam, the Netherlands: Benjamins.

Katz J (1977) *Propositional Structure and Illocutionary Force*. Cambridge, MA: Harvard University Press.

Katz J and Postal P (1964) *An Integrated Theory of Linguistic Descriptions*. Cambridge, MA: MIT Press.

Labov W and Fanshel D (1977) *Therapeutic Discourse*. New York, NY: Academic Press.

Levinson S (1981) The essential inadequacies of speech act models of dialogue. In: Parret H, Sbisa M and Verschueren J (eds) *Possibilities and Limitations of Pragmatics*. Amsterdam, the Netherlands: Benjamins.

Levinson S (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: Bradford/MIT Press.

Lewis D (1969) *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Recanati F (1987) *Meaning and Force: The Pragmatics of Performative Utterances*. Cambridge, UK: Cambridge University Press.

Recanati F (1989) The pragmatics of what is said. *Mind and Language* **4**: 295–329.

Perrault C and Allen J (1980) A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* **6**(3,4): 167–182.

Ross JR (1970) On declarative sentences. In: Jacobs R and Rosenbaum P (eds) *Readings in English Transformational Grammar*. Waltham, MA: Ginn and Co.

Sadock J (1974) *Towards a Linguistic Theory of Speech Acts*. New York, NY: Academic Press.

Sadock J and Zwicky A (1985) Speech act distinctions in syntax. In: Shopen T (ed.) *Language Typology and Syntactic Description*, vol. I. Cambridge, UK: Cambridge University Press.

Schegloff E (1988) Presequences and indirection: applying speech act theory to ordinary conversation. *Journal of Pragmatics* **12**: 55–62.

Schiffer S (1972) *Meaning*. Oxford, UK: Oxford University Press.
Searle JR (1969) *Speech Acts*. Cambridge, UK: Cambridge University Press.
Searle JR (1979) *Expression and Meaning*. Cambridge, UK: Cambridge University Press.
Searle JR and Vanderveken D (1985) *Foundations of Illocutionary Logic*. Cambridge, UK: Cambridge University Press.
Searle JR, Parret H and Verschueren J (eds) (1992) *(On) Searle on Conversation*. Amsterdam: Benjamins.
Sperber D and Wilson D (1986) *Relevance*. Cambridge, MA: Harvard University Press. [2nd edition (1995) Oxford: Blackwell.]
Strawson P (1964) Intention and convention in speech acts. *The Philosophical Review* **73**: 439–460.
Vanderveken D (1990) *Meaning and Speech Acts*, vol. I. Cambridge, UK: Cambridge University Press.
Wierzbicka A (1987) *English Speech Act Verbs*. New York, NY: Academic Press.
Winograd T (1972) *Understanding Natural Language*. New York, NY: Academic Press.
Wittgenstein L (1958) *The Blue and the Brown Books*. Oxford: Blackwell. [Dictated 1933–1994.]

Further Reading

Arnovick L (1999) *Diachronic Pragmatics: Seven Case Studies in English Illocutionary Development*. Amsterdam: Benjamins. [History of speech acts.]
Cohen P, Morgan J, Pollack M (eds) (1990) *Intentions in Communication*. Cambridge, MA: Bradford/MIT. [Computational speech acts.]
Kasher A (ed) (1998) *Pragmatics: Critical Concepts*, vol. II. New York: Routledge. [Reference source on speech acts.]
Sachs H (1992) *Lectures on Conversation*, 2 vols. Oxford: Blackwell. [Conversational analysis.]

Split Brains, Philosophical Issues about

Intermediate article

Grant Gillett, University of Otago, Dunedin, New Zealand

CONTENTS

A split brain is a brain in which the large bundle of fibres called the corpus callosum is divided. In such brains, information flows separately into the right and left hemispheres.

CLINICAL FINDINGS

The split brain syndrome results from a commissurotomy: an operation in which the large bundle of fibres connecting the two cerebral hemispheres is divided. A number of such operations were done in the 1960s and 1970s to relieve intractable epilepsy (Gazzaniga, 1970). There were two reasons why it was thought acceptable to undertake such a major procedure: first, it was an intervention of last resort for some severely affected epileptic patients; and second, it was difficult to detect any neuropsychological deficits following the operation. So subtle were the effects that it was widely believed that a section of the corpus callosum had no adverse effects on behavior (Parkin, 1996, p. 111). Indeed, most patients, after recovery from surgery, are untroubled in their everyday lives, and even now the literature on split brain patients confirms that 'a typical medical examination would not reveal anything unusual in their behavior, and their scores on standard tests are normal' (Kolb and Whishaw, 1990, p. 808). The most usual indication for this operation – the presence of intractable epilepsy arising from one or other hemisphere which can be limited in its effects by preventing the other hemisphere from being recruited in the course of a seizure – usually entailed that one side or other of the brain was abnormal because of the damage causing the epilepsy. Although such patients often had pre-existing damage in one hemisphere, they typically experienced no adverse events from the surgery itself. However, interesting patterns of behavior emerged on neuropsychological testing for cognitive disconnection.

Most of the experimental work used tachistoscopic exposure of visual stimuli so that they were confined to one or other half of the visual field (Gazzaniga, 1970). A number of different tests showed that a commissurotomy patient's response depended on whether it was controlled by the right hemisphere (e.g. picking out an object with the left hand) or the left hemisphere (e.g. verbal report). The typical disconnection syndrome resulting from commissurotomy is a combination of right-handed apraxia (defective visuospatial abilities at least in relation to complex tasks) and left-handed agraphia (inability to write). Such findings lead some to claim that commissurotomy patients develop different streams of consciousness, one located in each hemisphere: 'Each hemisphere can be shown to have its own sensations, thoughts, percepts, and memories that are not accessible to the other hemisphere' (Kolb and Whishaw, 1990, p. 808). And the problems are not confined to the experimental setting.

Disconnection is also revealed by confabulation when anomalies occur (Gardner, 1974, p. 361). For instance, a picture of a door may be shown 'to the left hemisphere', and a picture of a hammer 'to the right hemisphere'; when asked why the left hand is picking out a nail, the subject may say that you build doors with nails. Here each hemisphere deals with the information it has as best it can and the subject deals with discrepancies by using a technique – confabulation – often seen in amnesic or other cognitively impaired subjects. A further puzzling and bizarre manifestation of commissurotomy is 'manual conflict' or 'alien hand', where one hand interferes with the actions of the

other: 'the patient once grabbed his wife with his left hand and shook her violently while, with his right hand, he sought to rescue her and bring the violent left hand under control' (Gardner, 1974, p. 359). This is deeply problematic for traditional approaches to the philosophy of mind.

The problems become still more complex when we look at the way that subjects admonish and correct themselves, and also cheat, in order to overcome their problems (Weiscrantz, 1997). For instance, if information is delivered to one hemisphere and the parts of the body controlled by the other have to respond (e.g. by report, when the object is projected in the left visual field, or by picking out the object from an array with the hand ipsilateral to the visual field in which the picture of the object was presented), a common form of 'cheating' is to use facial expressions such as smiles and frowns to let the 'ignorant hemisphere' know what the right answer should be. This works because emotional expressions are bilaterally innervated (by sensorimotor systems).

PHILOSOPHICAL INTERPRETATIONS

A number of philosophers, including Nagel, Wiggins, and Parfit, have developed theses about the nature of conscious mental life and its unity or otherwise. These divide into three groups:

- There is a normal unity of conscious identity which is mental rather than physical but which is subserved by the hardware (or wetware) of the brain.
- There is no more to mental unity than a unity of diverse functions, each supervenient on localized brain processes (Nagel, 1979).
- There is no essential unity even in the mental life of a person. What matters to any of us are the relations of continuity and connectedness between mental states, which are the real basis of survival (rather than an enduring locus of personal identity) (Parfit, 1984).

It is problematic for dualists that certain mental acts are subject to profound disruption by an operation on the brain. Robinson (1976) calls such effects 'epistemic contradictions' and so places them among those functions associated with a 'mind, or intelligence, or intellect, or reason' (to quote Descartes in his *Fifth Meditation*). But Descartes also remarked, in his *Sixth Meditation*, that an essential feature of the mind as distinct from the body was its indivisibility. Thus if the mind is divided in split brain cases then Cartesian metaphysics is wrong. The remaining – monist – views are nonreductive only if we can suggest some reading of 'mental identity' that does not amount to the view that the mind is merely a set of functions supervenient on (or reducible to) neural functions.

Monist supervenient or reductive views clearly receive support from the split brain cases. It looks as if properties of mind such as the capacity to reason, form beliefs, perceive what is going on in the world, and act with intention are dissociated by commissurotomy into two sets, one per hemisphere. That is the conclusion reached by Nagel (1979), who claims that 'the numerically single subject is an illusion'. Parfit (1984, p. 251) remarks that 'the existence of a thinker just involves the existence of his brain and body, the doing of his deeds, the thinking of his thoughts and the occurrence of other physical and mental events'. Parfit denies that there is anything corresponding to Descartes' 'necessary unity of consciousness'. He concludes that all there is to mental unity is a set of relations (of connectedness and continuity) between mental states and events.

However, there is a difficulty with this position. In the experiments themselves, the subjects attempted to overcome their epistemic problems. They did this in a variety of ways, including self-correction and self-cueing (for instance by facial expression), and with such success that, even in the test situations, after a time the original neuropsychological problems cannot be demonstrated (Weiscrantz, 1997).

There are two important philosophical implications of this finding. First, each subject noted something amiss in his or her cognitive processes and interpreted it as an internal disruption of information flow. This led to a coordinated behavioral strategy aimed at correcting a perceived disability or departure from normal cognitive function. Thus, to all intents and purposes, the subject exhibits (verbally and practically) the belief (or attitude) that he or she is a single subject with cognitive problems.

Second, some argue that this attitude, and the coordinated strategy it calls forth, entails that there is something which the split brain patient is still 'one of'. This 'something' clearly exhibits reason, perception, and (perhaps impaired, but gradually reintegrated) action. Thus the subject, although temporarily 'fragmented' by cognitive defects, works, during his recovery, on the project of reintegration.

We therefore need a philosophical account of identity and consciousness that accommodates these phenomena.

IDENTITY AND CONSCIOUSNESS

The reductive view would be undermined if we could show that the constitution of thoughts,

perceptions and actions as 'contentful mental acts' requires an underlying unity of consciousness. Such an argument would have widespread implications for ascriptions of mental content in the context of brain damage.

Kant (1789) argued that every representation results from an act of judgment and implies a unitary 'I think' or a unified locus of conscious activity comprising the requisite cognitive skills. The thesis is that, in forming any mental representation, the subject judges that a number of presentations can be unified by being subsumed under a concept which then confers content on that representation. Thus a thinker confronting a small grey furry animal might judge that it is a possum. That judgment assimilates the present set of conditions to those conditions (experienced at other times) that justify the judgment 'possum'. This logically entails a mental unity between the epistemic subject now making the judgment and the subject who has mastered the requisite technique of judgment. This purely transcendental requirement (of Kant and the phenomenological tradition) can be developed further in a naturalistic approach to the mind.

The mental subject who judges in accordance with rules governing the use of concepts learns from those who impart the concept to him. His activity must be manifest so that others can converge in their judgments and participate in the application of consistent norms. The norms require the teaching of correct intentional responses to conditions which are public and are thereby incorporated into the structure of the subject's epistemic repertoire (Macdonald, 1998). On the basis of this activity, the subject develops as a locus of mental activity.

This practice is the basis of what we might call, after Wittgenstein, the grammar of self-ascription and self-reference: the application to oneself of mental predicates. Thus the mental subject is constituted as a unitary locus of those mental properties comprising consciousness. This creates a prerequisite for coherent mental content and therefore a normative influence favoring the reintegration of self after any disruption to it. These considerations lead to the philosophical position on mental ascriptions where the brain is damaged summarized below.

1. One's own conceptual relation to oneself shares truth conditions with judgments embedding mental concepts as used by coreferential others.
2. The relevant truth conditions are one's own manifest intentional activity.
3. One's own access to that activity is not the same as any other person's, and neither is one's access to the mental content informing it.
4. First- and third-person content attributions can fail because of imperfect knowledge on the part of the third or less commonly the first, person.
5. One situation in which first-person content attributions can fail is when a subject's broken brain affects knowledge and may disrupt the pattern of intentional activity that forms the basis of mental ascriptions.
6. One always tries to make sense of a single person's behavior as a tolerably coherent stream of intentional activity. This normatively driven exercise plays a central role in human activities. It underpins many of the things we do and think and it helps to bind the social enterprise of human survival.
7. The narrative self is a norm that cannot be discarded, but its application may be problematic if the brain is broken.
8. A subject always tries to get a narrative coherence into his or her conscious life. Hence even a fragmented brain will attempt to construct a unitary self as the emergent locus of mental activity on the basis of a holistic pattern of behavior.

Although this argument is based on the idea of mental content, it might be argued that what we have in the split brain case is not a single conscious subject but rather a pair of subsystems each exposed to the relevant norms but only partially equipped with the cognitive equipment to comply with them. It remains to be seen which is the best way to account for the clinical and neuropsychological data.

EMERGENT IDENTITY

The argument about the necessary unity of consciousness suggests that the split brain data reveal a single subject who is driven to give a unified account of his or her conscious experience, and sees himself or herself as laboring under epistemic difficulties arising from a commissurotomy. However, some would argue that the conscious subject is just a bundle of mental states and events loosely held together by a material substrate, or even a set of cognitive subsystems working together to provide a relatively coherent behavioral output. Reductionism, in either one of its forms, seems like a strong contender for the metaphysics of people where the mind is seen as being supervenient on a base physical level of reality. To some philosophers this has seemed like the only viable monist alternative, but the possibility of emergence suggests another naturalistic possibility.

The person, according to the anti-reductive or emergent view, is a single embodied subjective

and rational being who does the best he or she can using cognitive mechanisms which are designed to pick up information from the environment and produce a coherent stream of behavior. The person, on this view, holistically fashions an individual story based in a shared natural history, and usually has the cognitive resources to meet the demand that one should lead a unified conscious life as a being among others. This view, traceable back to the metaphysics of philosophers such as Strawson (1959) and Hampshire (1959), identifies the mind with a facet of a living, functioning human being taken as an individual. The naturalism inherent in this anti-reductive approach is appealing when we look at recent findings in subjects with a commissurotomy.

HOW FRAGMENTED NEUROCOGNITIVE FUNCTION MIGHT BE INTEGRATED

Traditionally, 'how' questions have been regarded as the province of science rather than philosophy. But there are areas of cognitive science where philosophical approaches to the mind can be illuminated by empirical data.

The reintegration of the mind after the splitting of its neural substrate is an area where philosophers can both learn and contribute. Neuropsychologists have tended to concentrate on internal neurocognitive mechanisms and the direct transfer of information between hemispheres using existing brain pathways, but there may be other mechanisms involved.

The ongoing activity shaping a human brain as an organ of cognitive adaptation to the environment is bodily activity. The behavior of any subject is the activity of an intentional agent with continuity of identity over time and a (more or less) coherent narrative about that activity. Therefore if there is no stable representation without neurocognitive change, it follows that a normative set of expectations that influence a subject's representation of self and world will induce the plastic brain to realize certain processing patterns. One would therefore expect the brain to find means of cognitive reintegration after commissurotomy. The two hemispheres do not have many ways of communicating; and yet the task is so central in human adaptation that any pathway or potential pathway will surely be used, even if they do not function as well as the neural pathways that are provided by our evolutionary history and strengthened by a lifetime of experience.

Weiscrantz (1997, p. 34) reported: 'Indeed those split-brain patients that I later studied no longer denied seeing stimuli projecting to the right hemisphere – they could give accurate descriptions and approximate dimensions of them.' Weiscrantz mentions neural plasticity and 'indirect bodily signaling stratagems' to account for this recovery. Few pathways are available for hemispheric reintegration once the major commissure with its 200 to 800 million fibres is severed (Kolb and Wishaw, 1990, pp. 504–506). But an interesting possibility is suggested by the fact that some of the patients cheated (one of Weiscrantz' 'indirect bodily signaling stratagems'). Not only does this indicate that the patient still behaves as if he or she is one self, albeit prone to certain kinds of errors, but also that direct transfer of information in the brain is not the only way to reintegrate cognition. The persistent psychological tendency to act as one subject – something that should not happen if the fragmentation of the system entails a fragmenting of the self (because of a strong supervenience of the mental on the physical) – might arise from public norms operating on the subject. But how can the brain achieve this reintegration given the poor resources available after commissurotomy?

According to the holistic view (and considering that a person resorts to indirect strategies), we can remark that the question is flawed – it is the person who achieves this and overcomes the fracture in the brain. We have noted the arguments based on mental content and the demand for reintegration in discussing what we might call the logical requirements of self- and other-ascription. On the view of mental ascription outlined above, a subject requires a body to manifest mental content and to provide a focus for mental ascriptions. Therefore the body is another resource to use in solving cognitive problems. This raises the possibility that, just as the body (or more correctly the embodied person) is the focus of the normative practices which impel towards reunification of the cognitive apparatus, so it might be the medium of neurocognitive reintegration.

All thinking is accompanied by, and even facilitated by, covert neuromuscular activity, involving either the movements that would be involved in the situation being thought about or the subvocal utterance of the words that would be said if the representation were articulated. McGuigan (1997) and others have provided persuasive evidence that cognitions are 'covert reactions' forming 'components of neuromuscular circuits governed by

cybernetic principles' and that 'where the striated musculature is totally inactive, cognitions are inactive'. If this is true, then the whole body is a massive display which exhibits cognitive processes; and it is subject to real-time proprioceptive monitoring. We know that connections between the brain and the body are very rich, two-way, and significantly bilateral. Ordinarily, the dominant traffic is between either half of the brain and the opposite side of the body – though there is also extensive ipsilateral information flow. Thus it is possible that cognitive reintegration is achieved by the brain using the body to talk to its other half rather than via the surviving direct connections. This would be a subliminal version of what often goes on in the so-called 'indirect cueing' of split brain patients during neuropsychological tests.

The implication is that, in cognition, it may be that 'meaning is use'. In other words, cognitive representation may be, at least in part, subliminally 'acting out' the responses that the subject has developed for dealing with a particular set of conditions. Furthermore, the subject might use such 'acting out' to determine what he or she thinks. We may need to reconsider the problem of mental ascriptions in split brain patients and the related ideas concerning the unity of consciousness and the nature of the self.

The dynamic nature of the problems following commissurotomy suggest a reassessment of the mental ascriptions to split brain patients, in much the same way as the phenomenon of blindsight has forced us to reassess visual perception. It is difficult to ascribe conscious intentional content in a split brain case. Our common practices of mental ascription do not apply, because the subjects have only partial cognitive access to the contents in question. The work of Block (1995) and others on theories of consciousness focusing on cognitive access seems to be highly relevant to the split brain data. It is clear that the split brain subject can only do certain things with the information gathered from the environment, and, to the extent that the subject has disconnected conscious access to that information, mental ascriptions are made for which the normal causes do not exist and which do not carry their normal entailments. Thus it is unclear the extent to which we can say that the subject is conscious (*simpliciter*) of, say, the object presented to the left visual field, despite the fact that some of the intentional responses to such data would, in a normal subject, justify ascriptions of conscious mental states.

This is consistent with a Wittgensteinian or even Dennettian view of conscious cognitive life (Gillett, 1999, pp. 102 ff.). Wittgenstein locates consciousness and thought in general in our normal practices of dealing with one another using language as a set of tools to understand and facilitate that interaction. Dennett regards conscious mental life as a 'lived narrative' or discursive articulation of what is happening in a seamless neuro-environmental interaction (1991). The narrative so constructed is jointly shaped by automatic responses and information gathering patterns, habits, well-rehearsed strategies for representation, the links between words and their truth conditions, and the intentional or voluntary direction of one's activity (using discursive techniques). A constant part of that neuro-environmental stream is, it seems from McGuigan's work, covert neuromuscular responses which mirror or reproduce (as it were in miniature) the subject's ways of responding to things. The conscious self is a holistic product of this stream as articulated using both linguistic and nonlinguistic techniques developed during the engagement of a person in his or her complex (natural and cultural) environment.

Therefore to try and find a 'self' in some part of the brain is as futile as Hume believed it to be on the basis of introspection and Kant argued that it was on the basis of transcendental analysis of consciousness and reason. The person who comes to think of himself or herself as a unified center of consciousness is a holistic or emergent unity born of environmental pressure on a physical focus (the body). The self as subject of thought and experience (as the reference of an integrated set of self-ascriptions) is a locus of purposive activity with intentional relations to its environment through the body. Thus the self, we might say, is necessarily embodied and therefore necessarily subject to disruptions of its internal causal structure in certain ways. But that is not to say that the subject can be reduced to a collection of physical processes or functions.

SUMMARY

We might view the philosophical implications of the split brain as follows.

1. The brain's functions can be split by commissurotomy into a set subserved by the left brain and a different set subserved by the right brain.
2. This does not bring about a fragmentation of the self: in fact, the self tries to overcome the problems caused by the damaged brain.
3. The self as integrated subject of mental ascriptions is a precondition for the ascription of 'conscious mental content' as we typically understand it.

4. The grasp of mental content is rooted in a public or shared world in which an embodied subject adapts to norms governing the conditions in which the content is properly exhibited.
5. The norms governing the causes and entailments of mental content are violated in the case of a split brain patient.
6. The relevant truth conditions for the judgments we normally make about mental content therefore rest on the bodily integrity of the subject and his or her more or less coherent human activity in relation to objects in the environment (so that the subject's states are intentional in the phenomenological sense).
7. A subject is not an immaterial Cartesian being. Events in the brain importantly affect mental ascriptions (and self-ascriptions).
8. First- and third-person content ascriptions can fail to meet their normal criteria in disconnection syndromes.
9. A broken brain affects both first-person and third-person knowledge of mental activity.
10. We try to make sense of a person's behavior as a coherent stream of intentional activity, so that the unity of the self is a socially enforced norm (based on normal embodiment), which may not be easily applied if the brain is broken.
11. The subject always tries to give his or her activity narrative coherence. Even the fragmented brain constructs a tract of emergent and more or less coherent activity.
12. The resources for reintegration are available to an embodied neurocognitive organism through social forces and interaction with a natural environment.
13. The subject is best thought of nonreductively, i.e., not just as a set of states and events produced by a brain but as a unified discursive subject inhabiting a shared world with others. On this basis we can understand the processes that overcome the defects caused by a broken brain.

A human being seems to operate as a unitary conscious self drawing on conceptual techniques that are learnt discursively. The idea that one configures oneself by the use of such techniques explains most of the findings about the fragmented self that are often held to be damaging to the folk conception of personal identity – or what makes a person the same person at different times and places. On this view, a person is not reducible to a set of states and events and the relations between them (as implied by the reductive view), and can, as Luria (1973) colorfully puts it, 'fight with the tenacity of the damned to recover the use of his damaged brain'.

References

Block N (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**(2): 227–247.

Dennett D (1991) *Consciousness Explained*. London, UK: Penguin.

Gardner H (1974) *The Shattered Mind*. New York, NY: Random House.

Gazzaniga M (1970) *The Bisected Brain*. New York, NY: Appleton Century Crofts.

Gillett G (1999) *The Mind and its Discontents*. Oxford: Oxford University Press.

Hampshire S (1959) *Thought and Action*. London: Chatto and Windus.

Kant I (1789/1929) *The Critique of Pure Reason*, translated by N. Kemp Smith. London: Macmillan.

Kolb B and Whishaw IQ (1990) *Fundamentals of Human Neuropsychology*. New York, NY: W. H. Freeman.

Luria AR (1973) *The Working Brain*. London: Penguin.

Macdonald C (1998) Externalism and norms. In: O'Hear (ed.) *Current Issues in Philosophy of Mind*. Cambridge, UK: Cambridge University Press.

McGuigan FJ (1997) A neuromuscular model of mind with clinical and educational applications. *Journal of Mind and Behavior* **18**(4): 351–370.

Nagel T (1979) Brain bisection and the unity of consciousness. In: *Mortal Questions*. Cambridge, UK: Cambridge University Press.

Parfit D (1984) *Reasons and Persons*. Oxford: Clarendon Press.

Parkin AJ (1996) *Explorations in Cognitive Neuropsychology*. Oxford: Blackwell.

Robinson D (1976) What sort of persons are hemispheres? Another look at the 'split brain man'. *British Journal for the Philosophy of Science* **24**: 339–355.

Strawson P (1959) *Individuals*. London: Methuen.

Weiscrantz L (1997) *Consciousness Lost and Found*. Oxford: Oxford University Press.

Further Reading

Gazzaniga M (1970) *The Bisected Brain*. New York, NY: Appleton Century Crofts.

Parkin AJ (1996) *Explorations in Cognitive Neuropsychology*. Oxford: Blackwell.

Robinson D (1998) Cerebral plurality and the unity of the self. In: Robinson D (ed.) *The Mind*, pp. 344–363. Oxford: Oxford University Press.

Rorty AO (1976) *The Identities of Persons*. Berkeley, CA: University of California Press.

Sperry RW, Gazzaniga MS and Bogen JE (1969) Interhemispheric relationships: the neocortical commissures; syndromes of hemispheric disconnection. In: Vinken and Bruyn (eds) *Handbook of Clinical Neurology*, vol. IV, pp. 273–290. Amsterdam: Elsevier.

Symbol Systems

Intermediate article

Michael L Anderson, University of Maryland, College Park, Maryland, USA
Donald R Perlis, University of Maryland, College Park, Maryland, USA

CONTENTS

A symbol is a pattern (for example, of physical marks, or electromagnetic energy) that denotes, designates, or otherwise has meaning. The notion that intelligence requires the use and manipulation of symbols, and that humans are therefore symbol systems, has been very influential in artificial intelligence.

INTRODUCTION

We begin this article with a presentation and discussion of the idea of a physical symbol system (PSS), as formulated by Newell and Simon. This notion – and the associated physical symbol system hypothesis (PSSH) – was first presented, under a somewhat different name, in Newell and Simon (1972), with later fuller formulations in Newell and Simon (1976) and Newell (1980), and still later elaborations in Newell (1990).

We then discuss various objections to PSSH, and replies to those objections, especially with reference to the themes of symbol grounding, situated cognition, embodiment, and situated robotics.

PHYSICAL SYMBOL SYSTEMS

In 1972 Allen Newell and Herbert Simon published their seminal book *Human Problem Solving*, in which, among many other things, they described 'information processing systems'. In 1975, Newell and Simon were jointly awarded the ACM Turing Award, and on that occasion they presented a paper in which they offered a more succinct description of these systems, under the now-standard name of 'physical symbol systems' (Newell and Simon, 1976, pp. 114–116):

> One of the fundamental contributions of computer science has been to explain, at a rather basic level, what symbols are … Symbols lie at the root of intelligent action … One [structural] requirement [for intelligence] is the ability to store and manipulate symbols. …
>
> A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another) … A physical symbol system is a machine that produces through time an evolving collection of symbol structures.

So defined, PSSs are very broadly conceived, largely because symbols themselves are conceived in very broad terms. Indeed, it would be difficult to specify any physical entity that would not count as a 'symbol' by this definition.

Associated with the notion of a PSS are various further notions. Thus, an expression in a PSS 'designates' an object if the PSS's behavior depends on the object (for instance by affecting the object); and a PSS can 'interpret' one of its expressions E if E designates a process that the system can carry out.

Designation and interpretation are intended to connect the PSS to the world. That is, the PSS must be able, somehow, to ground its symbols in real referents, and to act upon them. Moreover, in allowing for the double role of object (designatee) and process in a single expression (so that 'pull the chain' is both a sentence and a process), a PSS appears able to play the role of a self-modeling agent. This is crucial to the development of Newell and Simon's criteria for general intelligence. (*See* **Symbol-grounding Problem**)

Newell and Simon (p. 116) then assert the PSSH:

> A physical symbol system has the necessary and sufficient means for general intelligent action … This is an empirical hypothesis.

Given how broadly a PSS is defined, human brains would seem to count as PSSs. Our brains appear to manipulate symbol structures and carry out processes on the basis of (some of) those structures,

which in turn affect objects in the world (as well as other symbol structures). Thus the PSSH might seem obviously true, but Newell and Simon do not see it as obvious. Their hope is that ongoing research in artificial intelligence will succeed in 'bringing forth empirical evidence' in favor of PSSH. In the remainder of their 1976 paper they discuss evidence for PSSH, especially in terms of heuristic search via the 'heuristic search hypothesis' (p. 120):

> The solutions to problems are represented as symbol structures. A physical symbol system exercises its intelligence in problem solving by search – that is, by generating and progressively modifying symbol structures until it produces a solution structure. …
>
> Physical symbol systems must use heuristic search to solve problems because such systems have limited processing resources. (*See* **Search**)

Later, Newell (1980) explored the nature of PSSs in more detail, relating them to standard theoretical machinery. For instance, he defined PSSs simply as universal Turing machines: machines that can be programmed so as to simulate any computationally possible procedure whatsoever. Such a machine must be encodable as an expression that can serve as data, so universality entails symbolic expressions.

Although universal machines seem a far cry from the 1976 definition of PSSs, Newell argues convincingly that they are equivalent. He further argues that universal machines provide the first satisfactory definition of what constitutes a symbol: namely, something that can designate something, via a given machine (or symbol system). Thus Newell turns the definitional process backwards, defining symbol in terms of symbol system. A symbol, then, is directly tied to its use in a physical context, rather than having a prior existence. As a corollary, almost anything whatsoever can be a symbol. This has provoked objections to the PSSH, which we discuss below.

Given the equivalence between a PSS and a universal machine, the PSSH implies that universality is essential for intelligent behavior. Implicit in this, and made clearer in other work (on production systems and especially Soar) is the idea that it is the ability to represent, with symbolic expressions, one's own behavior that allows for behavioral changes and hence learning.

There are many other capacities of PSSs besides designation and interpretation, some of which are discussed by Newell in some detail. They have to do with internal manipulation of expressions, as well as input and output, to make the two basic capacities of designation and interpretation as powerful as possible.

Newell also outlines a series of levels of description of a physical computer: the device level (the electronics); the circuit level (electrical processes); the logic level (memory values and operations on them); the program level (the PSS level for a computer); and the processor-memory-switch level (the level of description of the various large-scale computer units, such as memory devices, processors, and input–output devices). He argues that there must exist a neural (biological) level of organization that supports a symbol structure: an organization he calls an 'architecture'. He regards this as an empirical hypothesis on a par with the PSSH.

This hypothesis is a forerunner to his definition of the 'knowledge level' (Newell, 1982), which is an addition to the above standard hierarchy for computers. Suppose, at the program (PSS) level, one were to implement an additional level of description via a special 'intelligent' program. This program would have stored knowledge (at the symbol level) and would bring that knowledge to bear on whatever problem it encountered. It would thus be a kind of reasoning engine. Newell calls this new level of behavior the knowledge level, and he formulates a principle of rationality for it: if an agent knows that one of its possible actions will achieve one of its goals, then the agent will perform that action. This is clearly an idealization, since it typically is not possible to bring all available knowledge to bear (because of resource limitations). Newell suggests that human cognition is at best an approximation to a knowledge-level system.

DEBATES CONCERNING THE PHYSICAL SYMBOL SYSTEM HYPOTHESIS

The computational model of human intelligence, as expressed for example in PSSH, has been extremely influential in the development of artificial intelligence, resulting in many techniques for simulating – and many systems that display – intelligent, if limited, behavior. These include techniques for problem solving and planning, especially various search techniques (Russell and Norvig, 1995), architectures like SNePS (Shapiro, 1979), and systems like Soar (Newell, 1990; Laird *et al*., 1987) and Hilare II (Giralt *et al*., 1991), as well as logic-based production systems that are time-situated and able to handle uncertainty and contradictory information (Bhatia *et al*., 2001). (*See* **Soar**)

However, this approach to understanding and reproducing intelligent behavior has been increasingly criticized by those who stress the importance of interaction with and utilization of the external environment, and the practical orientation of real-world agents, not just in shaping and guiding particular instances of cognition but in determining the nature of cognition in general (Anderson, forthcoming). We will provide only a brief account of the debate, primarily with an eye to better understanding PSSH. (*See* **Situated Cognition**; **Symbol-grounding Problem**; **Situated Robotics**; **Embodiment**)

To proponents of the situated or embodied approach to understanding intelligence, PSSH suggests a picture of cognition along the following lines. Firstly, a symbol system is characterized by a distinctive kind of decomposition of cognitive functions, such that the sensorimotor system is independent of, and functions primarily as the source of inputs to and the target of outputs from, the reasoning system. Problem solving proceeds in temporally and conceptually distinct steps: the world is sensed (input is received from the sensory system); a model of the world is built; a plan of action is formulated via computation on the model; and an action is taken (output is sent to the motor system). Secondly, the symbols, in terms of which the world is modeled, and by the transformations of which cognition proceeds, are meaningful primarily in terms of their internal relations to other symbols, and not in virtue of their physical relations to external objects, the behavioral dispositions of the cognitive system, or the particularities of their physical instantiation. Symbols 'denote' objects or aspects of the environment, and their representative function rests on this relation of denotation.

Proponents of embodied or situated cognition (SC) question whether intelligent organisms, humans included, fit the above descriptions. They do not deny that cognition involves abstract reasoning and planning, nor that humans employ symbols when engaging in this sort of reasoning. Rather, they suggest that abstract reasoning is only the tip of the cognitive iceberg; that it rests on and requires many other substantive cognitive capacities; and that these other capacities are not symbolic in nature, but rather involve states and processes that are tightly coupled to, and proceed via interaction with, the environment of the agent in question.

For example, whereas symbolic representation suggests an abstract relation of denotation, many internal representational states are in fact directly causally coupled to objects in, or aspects of, the environment. Thus a visual representation may cause a very particular pattern in the sight centers of the brain, which changes with changes in environment or in the relationship between the environment and the perceiver. This inner state is representational only in virtue of this continuing causal coupling, which suggests that the information about the world it contains should be understood not on a grammatical model, which involves abstract denotation of objects and their relations, but rather in terms analogous to the Watt governor, which carries information about the speed of an engine (in the angle of its arms) only in virtue of its direct coupling with that engine.

Whereas PSSH suggests that the functioning of the sensorimotor and reasoning components of an intelligent system can be understood largely in isolation from one another, SC maintains that the process of sensing and representing in fact involves the continual cooperation of these components, which should perhaps not therefore be presented as functionally decoupled. In general, what is worth paying perceptual attention to, and what concepts and categories are appropriate to bring to bear in representing the world, depend upon what one is doing (Clancey, 1993). Further (and partly for this reason), representations, tend to be cast in functional terms – 'the-bee-that-is-chasing-me' rather than 'bee12' (Agre and Chapman, 1987; Bickhard, 1993) – and the world is understood, in part, in terms of the actions it invites or 'affords': a chair is perceived not in terms of abstract qualitative descriptions, but as affording sitting (Gibson, 1979). This suggests that the content and meaning of inner mental states, and the processing they undergo, cannot be understood in isolation from the ongoing activity of the representing agent. Likewise, whereas PSSH suggests that cognition should be understood in terms of the four-step sense–model–plan–act cycle, and that the calculation on symbols that occurs after sensing and before acting is the meat of thinking, SC claims that cognition can (and often does) involve interaction with the environment at any stage – for instance, rotating a puzzle piece to make it easier to visualize, or writing down the intermediate results in a complex mathematical calculation. Thus, rather than being just the result of cognition, a given action can be part of the cognitive process.

Whereas PSSH suggests that cognition should be understood in terms of abstract rules for manipulating abstract symbols, SC maintains that cognition is in fact rooted in basic coping strategies and the embodied experience of thinking agents. Thus, even the apparently abstract rules of logic are best

understood in terms of more basic experience. Lakoff (1987), for instance, suggests that the 'exclusive or' operator is rooted in and derived from our basic experience with physical objects and containers: an item can be in one box or the other, but not both.

One important response to this challenge can be found in Vera and Simon (1993). There the authors argue that proponents of SC have interpreted the notion of a 'symbol' in a restricted sense, much more narrow than originally intended, and that when a symbol system is understood more broadly, there is no necessary antithesis or tension between SC and PSSH. We quote their characterization of symbol systems at length (Vera and Simon, 1993, pp. 8–9):

> A physical symbol system is built from a set of elements, called symbols, which may be formed into symbol structures by means of a set of relations. A symbol system has a memory capable of storing and retaining symbols and symbol structures, and has a set of information processes that form symbol structures as a function of sensory stimuli, which produce symbol structures that cause motor actions and modify symbol structures in memory in a variety of ways.
>
> A physical symbol system interacts with its external environment in two ways: (1) it receives sensory stimuli from the environment that it converts into symbol structures in memory; and (2) it acts upon the environment in ways determined by symbol structures (motor symbols) that it produces. Its behavior can be influenced both by its current environment through its sensory inputs, and by previous environments through the information it has stored in memory from its experiences.
>
> Henceforth, we will usually refer to both symbols and symbol structures simply as 'symbols'. Symbols are patterns. In a computer, they are typically patterns of electromagnetism, but their physical nature is radically different in different computers (compare the vacuum tubes of the 1940s with integrated circuits of today). And, in any event, their physical nature is irrelevant to their role in behavior. The way in which symbols are represented in the brain is not known; presumably, they are patterns of neuronal arrangement of some kind.
>
> When we say that symbols are patterns, we mean that pairs of them can be compared (by one of the system's processes) and pronounced alike or different, and that the system can behave differently, depending on the same/different decision.
>
> We call patterns symbols when they can designate or denote. An information system can take a symbol token as input and use it to gain access to a referenced object in order to affect it or be affected by it in some way. Symbols may designate other symbols, but they may also designate patterns of sensory stimuli, and they may designate motor actions. Thus, the receipt of certain patterns of sensory stimulation may cause the creation in memory of the symbol (say, CAT) that designates a cat (not the word 'cat', but the animal). Of course, this does not guarantee that there is really a cat out there: that depends on the veridicality of the processes that encode the stimulus into the symbol designating a cat. Similarly, a motor symbol may designate the act of 'petting' (with some parameters to assure that the cat will be the object of the petting).

It is not clear that this characterization of PSSs is very different from that offered in the name of SC, above. In particular, Vera and Simon explicitly maintain that symbols are 'patterns that designate or denote', and define such symbols separately from the patterns of sensory stimuli, in response to which symbols may be generated. It also appears to have the same general 'flavor', suggesting as it does the sense–model–plan–act cycle to which SC objects. A physical symbol system operates as follows: stimuli are received; symbols that denote the environment are generated; these symbols are processed to produce more symbols, some of which designate actions, which are then sent to the motor system to cause a motor response.

However, Vera and Simon's discussion of SC suggests that they may have in mind a somewhat broader definition of symbol (pp. 38–39):

> In some situations, an actor's internal representations can be extremely simple, but no one has described a system capable of intelligent action that does not employ at least rudimentary representations. Perhaps the barest representation encompasses only goals and some symbolization of a relation between goal and situation, on the one hand, and action in the other. But some representation of these is unavoidable if action is to be purposive. … In systems like Pengi [Agre and Chapman, 1987] and the creatures of Brooks [Brooks, 1999], often taken as paradigmatic examples of applied [SC], there are substantial internal representations, some of them used to symbolize the current focus of attention and the locations of relevant nearby objects, others used to characterize the objects themselves in terms of their current functions.

Or (p. 37):

> That the symbols in question are both goal-dependent and situation-dependent does not change their status. They are genuine symbols in the traditional information-processing sense. …'The-bee-that-is-chasing-me' is a perfectly good symbol; it denotes a distinct class of object in the world (i.e. any bee that is engaging in the activity of chasing me).

This claim seems convincing in reference to Pengi. Pengi's inner states are symbolic because they do in fact denote, albeit in terms of a functional characterization. It is much less clear that the inner states in Brooks' creatures, are of this sort, or that we should accept this apparent broadening of the term 'symbol' to include any internal state. Brooks's six-legged robot Genghis achieves walking behavior with a very simple set of controllers (Brooks, 1999). The position of each leg is represented in terms of two numbers, α and β, which give, respectively, its horizontal and vertical position. An α-balance machine continually sums the six α values and sends the sum to each leg, so that if one leg is forward, all legs will be sent a signal casuing them to move back slightly to compensate. There is also a simple controller attached to each leg, such that if the β is positive (the leg is up), it increases α by suppressing the signal from the α-balance machine. In addition there is a controller which decreases β (puts the leg down) whenever β is positive, and an up-leg trigger which can cause the leg to go up by suppressing the leg-down command. Finally, there is a walk trigger, which sends a timed signal to the up-leg triggers (e.g. to cause three of the legs to go up every 2.4 seconds). When the legs go up, they move forward and down (because of the leg-forward and leg-down controllers), while the three legs on the ground move backwards (because of the α-balance machine). Then the other three legs go up, and so on. This walk trigger can, in turn, be connected to sensors to cause it to run only under certain conditions (e.g. when the infrared sensors register heat). Although there is certainly representation here, it is difficult to see this activity in terms of denotation and symbol processing; it seems much more natural to understand the representations of leg position as instances of causal coupling, and Genghis' walking behavior, in reactive terms. Although one can stretch the definitions of 'symbol' and 'denotation' to cover this case, doing so would not thereby erase the difference between the 'symbols' in Genghis and their relation to the things they 'denote', and those employed in Pengi. Every environmentally reactive system has internal states: surely some of these states are not symbolic – they do not denote, nor can they be systematically combined with other inner states to form symbol structures.

Vera and Simon approach the issue of affordance – the perceived invitation by an environment to take certain actions – in a similar fashion, accepting the importance of affordances to guiding agency, but insisting that these affordances be understood symbolically, i.e. in terms of internally encoded states (p. 41):

> We have already seen that when people are dealing with familiar situations, using habitual actions, their internal representations, at the conscious level, may be almost wholly functional, without any details of the mechanisms that carry out these functions. The 'affordances' of the environment, represented internally, trigger actions.
>
> Of course, the absence of consciousness of mechanisms implies neither that mechanisms are absent nor that they are non-symbolic. To acquire an internal representation of an affordance, a person must carry out a complex encoding of sensory stimuli that impinge upon eye and ear. And to take the corresponding action, he or she must decode the encoded symbol representing the action into signals to the muscles.

Vera and Simon may be on more solid ground here, for affordances are not the same as causes: they represent the behavioral options offered by an environment, and only act to trigger a given action under particular cognitive circumstances. A chair affords sitting, but triggers sitting only if I want to sit down. Thus, it seems that it must be possible for these representations to be related to other complex cognitive states like desires, which may in turn play roles in longer-term plans and activities. Still, it is worth pointing out that there are important distinctions to be drawn between affordances, which are often unconscious, action-oriented inner states that are perceptually active (in that they directly inform the content of perception) and serve primarily to allow substantial and inexpensive coordination between sensory input and motor responses, and abstract concepts like `CAT`, which, although certainly grounded in sensory experience, also have substantial logical, hierarchical and lexical relations, important to their meaning, the full extent of which is not perceptually active or involved in sensorimotor coordination. Thus, while it may be plausible in some circumstances to understand the role of the latter symbol in the linear terms suggested by PSSH – whereby one senses something, which triggers the symbol `CAT`, which symbol is processed within a network of beliefs and desires (I like cats and like to pet them), thereby producing the symbol structure `Pet(CAT)`, which, when sent to the motor system, causes my petting of the cat – it is not immediately obvious that these simple coordinations between the sensory, motor, and conceptual systems are sufficient, or of the right form, to support the deployment of affordances.

This is relevant to the larger question of how best to understand human problem-solving behavior,

which, as Vera and Simon accept, often involves interactions with the environment whose role is epistemic (Kirsh and Maglio, 1995), i.e. part of the problem solving itself (aimed, for instance, at simplifying the problem at hand, as when one rotates a puzzle piece to better 'see' where it might fit) rather than an implementation of a solution reached by cognition alone. They are certainly right to insist that continual interaction with the environment is fully compatible with PSSH (p. 10).

> [S]equences of actions can be executed with constant interchange among (a) receipt of information about the current state of the environment (perception), (b) internal processing of information (thinking), and (c) response to the environment (motor activity). These sequences may or may not be guided by long-term plans (or strategies that adapt to feedback of perceptual information).

This way of modeling environmental interaction in problem solving treats our interaction with the puzzle piece (sense model, look for fit, rotate model, look for fit, rotate model, see fit, place) as just a variation of the sense–symbolize–plan–act cycle with more frequent attention to incoming sensory information and its changes. At a certain level of abstraction, this is no doubt the case, for surely epistemic actions involve internal processing just as do pragmatic actions. Yet it seems that, when the object of an action is to simplify a calculation or a search, or otherwise change the epistemic parameters of a task, the coordinations required between the sensory, symbolic, and motor systems are somewhat different from those required in the case of pragmatic action.

This is not to say that the execution of epistemic actions does not require symbols and symbol processing. Indeed, it may be that epistemic actions are even more symbol-based than pragmatic ones, for they may involve modeling not just the environment but also the self and its cognitive capacities. Alternately it could be that a different decomposition of cognitive functions would diminish the need for symbolic coordination between the parts, even in the case of epistemic actions. Proponents of SC hope for the latter, while PSSH expects the former.

When the debate between SC and PSSH is cast in terms of whether cognition depends on symbols, or whether it is primarily reactive and interactive, it has a tendency to devolve into semantic questions: what is the meaning of 'symbol'; and is a sense–symbolize–plan–act system really interactive? This is not the most productive impasse; for both positions have room for both symbols and interaction.

There are nevertheless substantive differences between the two positions, which we suggest come down to three main issues:

- *The nature and role of concepts*. We define a concept as a structured, contentful inner object (constituent of thought) that is semantically evaluable, redeployable and largely stable, and that has hierarchical and logical connections to other concepts. Given this definition, there is a disagreement about when such inner structures are needed to explain intelligent behavior. This definition of a concept is similar to, although somewhat narrower than, the PSS definition of a symbol, suggesting that PSSH expects to discover a very central role for concepts at nearly all stages of cognition. In contrast, SC expects a larger role for nonconceptual content, which is best specified in terms of the abilities, skills and dispositions of the agent, or in terms of significant, nonverbalizable bodily or perceptual experience (Chrisley, 1995).
- *The details of cognitive decomposition, and of the relations and coordinations between the parts*. Although PSSH is not committed to any particular implementation of intelligent agency, it has tended in practice to isolate perception and action from reasoning components, and to handle coordination between subsystems or agents in terms of the distribution and interpretation of symbolically encoded information. In contrast, SC is committed to systems that have much more interpenetration between perception, action and reasoning systems (as in the case of affordances), and in which coordinations between subsystems are indirect (as when all systems have access to the same sensor stream, but do not directly exchange information), mediated by the environment (as when Genghis, in lifting one leg, naturally increases the weight on the other five legs, allowing the 'information' that a leg has gone up to be transmitted to the other leg controllers with no direct information exchange), or subsymbolic (Braitenberg, 1984; Edelman, 1992).
- *The nature and origin of higher-order cognition*. Although Vera and Simon do not directly address the SC claim that the contents and rules of higher-order cognition are rooted in more basic experience (for example, the claim that 'exclusive or' is derived from experience with objects and containers), it seems that such claims, insofar as they primarily pertain to the question of how such laws can be understood or learned by humans, are fully compatible with PSSH. Still, the overall project of naturalism, of which SC is one instance, faces difficulties in accounting for certain formal properties of systems like logic and arithmetic (e.g. completeness), and the apparent necessity of the truths they express (one would expect a system based on contingent experience to be likewise contingent, yet the fact that $2 + 2 = 4$ does not appear to be a contingent truth). And questions remain about the nature of abstraction (and how to implement an abstracting agent), and self-modeling, which seem necessary to high-level cognition no matter how one accounts for its origin.

References

Agre PE and Chapman D (1987) Pengi: an implementation of a theory of activity. In: *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pp. 268–272.

Anderson ML (forthcoming) Embodied cognition: a field guide. *Artificial Intelligence*.

Bhatia M, Chi P, Chong W *et al*. (2001) Handling uncertainty with active logic. In: *Proceedings of the AAAI Fall Symposium on Uncertainty in Computation*, pp. 1–9. Menlo Park, CA: AAAI Press.

Bickhard HM (1993) Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence* **5**: 285–333.

Braitenburg V (1984) *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Brooks RA (1999) *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

Chrisley RL (1995) Taking embodiment seriously: nonconceptual content and robotics. In: Ford KM, Glymour C and Hayes PJ (eds) *Android Epistemology*, pp. 141–166. Menlo Park, CA: AAAI Press.

Clancey WJ (1993) Situated action: a neuropsychological interpretation (response to Vera and Simon). *Cognitive Science* **17**(1): 87–107.

Edelman GM (1992) *Bright Air, Brilliant Fire*. New York, NY: Basic Books.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. New York, NY: Houghton Mifflin.

Giralt G, Alami R, Chatila R and Freedman P (1991) Remote operated autonomous robots. In: *Intelligent Robotics: Proceedings of the International Symposium* **1571**: 416–427. Bangalore, India: International Society for Optical Engineering (SPIE).

Kirsh D and Maglio P (1995) On distinguishing epistemic from pragmatic actions. *Cognitive Science* **18**: 513–549.

Laird JE, Newell A and Rosenbloom PS (1987) SOAR: an architecture for general intelligence. *Artificial Intelligence* **33**(1–2): 113–129.

Lakoff G (1987) *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

Newell A (1980) Physical symbol systems. *Cognitive Science* **4**: 135–183.

Newell A (1982) The knowledge level. *Artificial Intelligence* **18**: 87–127.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell A and Simon HA (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell A and Simon HA (1976) Computer science as empirical enquiry: symbols and search. *Communications of the Association for Computing Machinery* **19**(3): 113–126.

Russell S and Norvig P (1995) *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice-Hall.

Shapiro SC (1979) The SNePS semantic network processing system. In: Findler NV (ed.) *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 179–203. New York, NY: Academic Press.

Vera AH and Simon HA (1993) Situated action: a symbolic interpretation. *Cognitive Science* **17**: 7–48.

Synaesthesia

Intermediate article

Simon Baron-Cohen, University of Cambridge, Cambridge, UK
John Harrison, Cambridge Psychometric Consultants, Cambridge, UK

CONTENTS

Synaesthesia is experienced when stimulation of one sensory modality gives rise to a perception in a second modality, without that second modality having received any direct stimulation.

WHAT IS SYNAESTHESIA?

Synaesthesia occurs when stimulation of one sensory modality automatically triggers a perception in a second modality, in the absence of any direct stimulation to this second modality (Vernon, 1930; Marks, 1975; Cytowic, 1989, 1993; Motluk, 1994). So, for example, a sound might automatically and instantly trigger the perception of vivid color, or vice versa. The woman EP, the subject of Baron-Cohen *et al.*'s (1987) study, provided a number of descriptions of color–word correspondences, describing the sound of the word MOSCOW as 'Darkish grey, with spinach green and pale blue'. Those with the condition describe the percept not as part of their external visual experience, nor in their 'mind's eye', but somewhere else, phenomenologically.

Many combinations of synaesthesia are reported to occur naturally, including sound giving rise to visual percepts ('colored-hearing') and smell giving rise to tactile sensation, as in Cytowic's (1993) subject MW. Colored-hearing synaesthesia appears to be the most common form. Certain combinations of synaesthesia almost never occur (e.g. touch to hearing). Synaesthesia is also sometimes reported by those who have used hallucinogenic drugs, such as Lysergic Acid Diethylamide (LSD) or mescaline. This article will focus on the naturally occurring form of synaesthesia, whilst acknowledging that there may be a connection to be found with the drug-induced form.

Terminology

'Developmental' synaesthesia

'Developmental synaesthesia' is distinct from acquired synaesthesia (of which there are at least two forms) and pseudosynaesthesia. Developmental synaesthesia in most cases has several characteristics: (1) it has a childhood onset, in all cases before four years of age; (2) it is different from hallucination, delusion, or other psychotic phenomena; (3) it is different from imagery arising from imagination; (4) it is not induced by drug use; it is (5) vivid; (6) automatic/involuntary; and (7) unlearnt.

Synaesthesia caused by neurological dysfunction

A variety of neuropathological conditions can give rise to acquired synaesthesia. A full account of the varieties of acquired synaesthesia is given in Critchley (1994). Note that the resultant synaesthesic percepts are often much simpler than the complex forms seen in developmental synaesthesia.

Synaesthesia as a consequence of psychoactive drug use

Synaesthesia can result from the use of psychoactive drugs (Cytowic, 1989). The mechanisms by which drug-induced synaesthesia occur are not well understood, though the use of LSD, mescaline, and psilocin are all reported to cause confusion between the sensory modalities, so that sounds are perceived as visual (Rang and Dale, 1987). Neurophysiological studies reported by Aghajanian (1981) suggest different sites of action for LSD and mescaline, with LSD seeming to work by

inhibiting the serotonin-containing neurons of the raphe nuclei, and mescaline by acting upon the noradrenergic system. Drug-induced synaesthesia differs from developmental synaesthesia in several ways: (1) it is often accompanied by hallucinations and loss of reality-monitoring; (2) it is transient; (3) it usually has an onset only in adult life (or whenever the drug was used); and (4) it can produce sensory combinations which do not otherwise occur naturally.

What Synaesthesia Is Not

Metaphor

Almost all writers on the topic of synaesthesia have been drawn into discussion of the possibility that a number of authors, poets, artists, and musicians may have had synaesthesia. A typical list of these individuals would include the composers Liszt, Rimsky-Korsakov, Messiäen, and Scriabin; the poets Basho, Rimbaud, and Baudelaire; the artists Kandinsky and Hockney; and, finally, the novelist Nabokov. There is no evidence of these individuals having been tested formally for synaesthesia. Much of the literature may instead reflect a form of metaphor or analogy. Thus, Charles Baudelaire appears to have believed in the unity of sensation (as implied by his poem 'Correspondances'). Metaphor is widespread in language and provides ripe conditions for confusion with developmental synaesthesia. Distinguishing the 'metaphor as pseudosynaesthesia' from developmental synaesthesia relies on objective tests. However, the key differences are that in metaphor: (1) no percept is *necessarily* triggered; (2) the subject will often acknowledge that the description is only an analogy; and (3) it is voluntary.

Association

A second form of pseudosynaesthesia includes individuals who have simply learnt to pair words/letters with colors (e.g. alphabet books in which letters are depicted in a variety of colors). Detailed examination of the color-letter alphabets of individuals with developmental synaesthesia often yields the finding that successive letters have very similar colors. This is in marked to contrast to colored alphabet books in which successive letters have markedly different color representations.

HISTORY

The closing decades of the nineteenth century saw a considerable number of accounts of synaesthesia, most notable amongst which was Galton's (1883) *Inquiries into Human Faculty and Its Development*. Scientific interest in the condition declined with the rise of behaviorism and very little on the topic appears in the literature from the late 1920s onward. This was probably because behaviorism banished reference to mental states from scientific language. As synaesthesia could only be defined by self-report and reference to mental states, it was not considered 'scientific'.

Since the 1990s synaesthesia research has enjoyed something of a renaissance. Various disciplines within cognitive neuroscience have contributed both new data and theory. Such developments have led to the condition being widely recognized as having a neurological reality. This new acceptance of the condition is in part due to objective approaches to studying it now being available.

EXPERIMENTAL WORK ON SYNAESTHESIA IN PSYCHOLOGY AND NEUROSCIENCE

The conventional test for the colored-hearing synaesthesia involves assessing a subject's *consistency* in reporting color descriptions for words across two or more occasions, when the subject has no prior warning of the retest. This consistency should be irrespective of the length of interval between testing sessions (Baron-Cohen *et al*, 1987; Baron-Cohen *et al*., 1993). Using this method, consistency is typically as high as 90 per cent, even when retested over years and even when stringent criteria are set for retest descriptions.

The advent of neuroimaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) has also provided an opportunity to image the brains of individuals with synaesthesia *in vivo*. Cytowic used the xenon inhalation technique to image the brain of a single subject (Cytowic and Wood, 1982) and others have used PET (Paulesu *et al*., 1995) and more recently fMRI (Gray *et al*., 1997). Given the marked consistency of patterns of activation in synaesthetes studied in the Paulesu *et al*. study, it might ultimately prove possible to determine the presence of synaesthesia objectively using functional brain imaging techniques. The neuroimaging data is discussed later in relation to theories of synaesthesia.

THEORIES OF SYNAESTHESIA

Over the past 200 years a number of hypotheses have been put forward to explain the cause of synaesthesia.

Preserved Neural Connectivity

The normal adult human brain does not contain direct neural connections between auditory and visual areas. However, the early developing brain in many species does. This first theory holds that, probably for genetic reasons, in individuals with synaesthesia, pathways between auditory and visual areas in the brain continue to exist beyond neoteny, such that when words, or sounds, give rise to activation in auditory areas, visual cortex is also stimulated.

There is evidence for the presence of such connective pathways in other species (see Kennedy *et al*., 1996). Kennedy and others in a number of studies (Dehay *et al*., 1984; Kennedy *et al*., 1989), have found that connections between auditory and visual areas exist in the brain structure of species such as the macaque monkey (*Macaca irus*) and the domestic cat (*Felis domesticus*). These projections appear to be transient, typically disappearing approximately three months *post partum*.

There is also some evidence that these transitory pathways exist in human neonates, and may, as in cats and macaques, get 'pruned' as part of the biological maturation of the brain. Much of this evidence is reviewed by Maurer (1993). Maurer's hypothesis is that human babies mix the input from different senses and that this gives rise to normal 'synaesthesia'. We know from the work by Meltzoff and Borton (1979) that babies who suck on either a 'nubby' or a 'smooth' pacifier (dummy) will prefer to look at a picture of the pacifier they sucked on, thereby showing a match between touch and vision. The Meltzoff and Borton study is usually taken as evidence for cross-modal transfer. Maurer goes one step further in suggesting that synaesthesia might be a normal stage of perceptual experience in addition to cross-modal transfer.

Maurer's evidence in support of this view comes from other studies of neonates. One such study is that reported by Lewkowicz and Turkewicz (1980). In this experiment one-month-old children, who had seen a patch of white light for 20 trials, were presented with bursts of 'white noise' presented at different intensities. During the noise presentation, the patch of light that they had been trained on was interspersed repeatedly and the children's heart rate was measured. Normally heart rate increases as a function of noise intensity, but Lewkowicz and Turkewicz found that the heart rate recorded at a noise intensity of 74 dB showed the lowest heart rate change and that for values greater or less than this value heart rate increased. Lewkowicz and Turkewicz's interpretation of this finding was that 'infants were responding to the auditory stimuli in terms of their similarity to the previously presented visual stimulus' (1980, p. 597), or, as Maurer put it, 'the children responded least to the "familiar" intensity' (p. 110). Maurer also cites evidence from electrophysiological studies of neonates showing that the amplitude of somatosensory evoked potentials increases when they are played white noise. Normally, these potentials only increase as a consequence of tactile stimulation. Finally, she cites the work of Neville (1993) that in early infancy auditorily evoked potentials to language evoke a potential in the occipital cortex, whereas in older individuals these stimuli yield potentials only in auditory areas such as the temporal lobes.

The evidence in this section is consistent with the notion that synaesthesia might be due to the persistence of neural information passing from auditory to visual brain areas, beyond the neonatal stage. Taken in the context of development, it also suggests the intriguing possibility that we might *all* be colored-hearing synaesthetes until we lose connections between these two areas somewhere about three months of age, at which point cortical maturation gives rise to sensory differentiation. This is consistent with Cytowic's (1989) view of synaesthetes being 'cognitive fossils'.

Sensory Leakage Theory

Jacobs *et al*. (1981) proposed what can be called the 'Sensory Leakage Theory'. This is an account of how simple photisms arise in cases of acquired synaesthesia, though it could in principle be extended to account for developmental synaesthesia. As mentioned earlier, most cases of *acquired* synaesthesia arise in individuals who suffer brain damage to anterior portions of the brain, often the optic nerve. Close examination of the nine patients reported in Jacobs *et al*. reveals that four of these patients (cases 1, 2, 4, 7) also experienced photisms in the *absence* of auditory stimulation, casting doubt on whether these instances should be described as cases of auditory-visual synaesthesia at all. It is also worth observing that seven patients always experienced their photisms when they were 'relaxed, drowsy or dozing' (p. 214), circumstances in which hypnagogic hallucinations are possible.

The essence of Jacobs *et al*.'s theory is that auditory information 'leaks' into pathways and areas in the brain that ordinarily deal with visual information. Jacobs *et al*. expand this 'leakage' theory by suggesting that there are 'numerous regions of the brain where visual and auditory pathways lie in

close anatomic proximity' (p. 216) and that at these points post-synaptic fibres might converge to cause the synaesthesia seen in a range of pathological states such as congenital blindness and drug intoxication.

Evidence to support leakage between areas subserving different forms of sensory information is sparse, causing some difficulties for Jacobs *et al.*'s theory. However, recent work has suggested that rather than posit the need for leakage, it is possible to find at certain locations in the brain classes of neurons that are responsive to stimulation from more than one sensory modality. For example, in a study of nonhuman primates carried out by Graziano *et al*. (1994), recordings were made from neurons (N = 141) in the ventral portion of the premotor cortex. Of these neurons, 27 to 31 per cent were found to be bimodally responsive, firing as a result of either, or both, visual and somasthetic stimulation.

Sadato *et al*. (1996) showed that congenitally blind subjects show increased bloodflow to primary visual areas when reading Braille, a finding the authors account for by suggesting that in these subjects 'cortical areas normally reserved for vision may be activated by other sensory modalities' (p. 526). Such an account might also explain the case of acquired synaesthesia reported by Rizzo and Eslinger (1989). Their subject, a 17-year-old who had developed retrolental fibroplasia as the consequence of perinatal difficulties, exhibited a florid form of colored hearing for musical tones. Rizzo and Eslinger failed to find evidence to suggest visual area activation as a consequence of auditory stimulation, but limited themselves to the use of electroencephalography as a means of detecting such activity. The Sadato *et al*. finding suggests that functional neuroimaging might prove to be a useful technique for investigating cases of acquired synaesthesia.

Cytowic's Theory of Synaesthesia

The most controversial theoretical account of the cause of synaesthesia is that most recently advanced by Cytowic (1993) in his book *The Man Who Tasted Shapes*. Cytowic proposes that synaesthesia occurs because 'parts of the brain get disconnected from one another … causing the normal processes of the limbic system to be released, bared to consciousness, and experienced as synaesthesia' (p. 163). His assertion that the limbic system is the critical brain locus can and has been tested. In the final chapter, Cytowic concedes that whilst he has no direct evidence to implicate a particular neural structure, given the 'stunning shut-down of the cortex' (p. 152) observed in the ^{133}Xenon studies of blood flow in the subject MW's brain, he points to the limbic areas as being 'the seat of synaesthesia'. Direct evidence of the involvement of the limbic system would have been provided by evidence of blood flow changes in this brain region, though unfortunately neuroimaging using ^{133}Xenon inhalation does not permit such deep structures to be imaged.

This is not a limitation shared by PET and so the importance of the limbic system in synaesthesia can be evaluated using this technique. This was one of the questions addressed in the study of colored-hearing synaesthesia reported by Paulesu *et al*. (1995). This study compared brain activity in synaesthetes and control subjects whilst listening to either words or pure tones. The synaesthetes reported color percepts for words but not for nonword sounds, and so a comparison of brain activation on synaesthetes listening to words as compared with tones, and with control subjects, should yield clues to the neural basis of colored-hearing synaesthesia. From this analysis two areas of particular interest emerged, posterior infertemporal cortex and the parietal-occipital junction, both of which have known involvement in color perception. However, in neither the between-groups nor the within-groups comparisons was there any suggestion of limbic system involvement. Of course it might be that Cytowic's subject MW is different in *kind* from the subjects scanned by Paulesu *et al*. Given MW's grossly abnormal resting blood flow levels, together with his polymodal synaesthesia, this remains a strong possibility.

The Learned Association Theory

This theoretical proposition was originally suggested as an explanation of synaesthesia by Calkins (1893) and holds that in colored-hearing synaesthesia, the color–word and/or sound correspondences reported are due entirely to learned association. The idea is that the color–letter associations are derived from colored alphabet books, or colored letters, that the individual saw as a child. Whilst this is a plausible account of the acquisition of pseudosynaesthesia, we suspect, for a number of reasons, that it is an unsatisfactory account of developmental synaesthesia. Reasons include the following:

1. The sex ratio. The sex ratio in synaesthesia is 6:1 (f:m). Why should so many more women, as compared to men, form such associations? A socialization account

which would lead to this sex ratio is not immediately obvious, though transmission from mothers to daughters via modeling may be a possibility (though a tenuous one).

2. Consecutive letters. Careful scrutiny of the 'colored alphabets' of many synaesthetes, yields the finding that often consecutive letters are closely described in color terms (e.g. 'M' = olive green, 'N' = emerald green, 'O' = washed-out pale green). When compared to colored alphabet books, it is found that publishers logically go to great lengths to ensure that consecutive letters are printed in very different colors. Learned association therefore cannot account for the specific colors of particular letters or phonemes.
3. Synaesthesic twins. A comparison of the colored alphabets of twins so far has yielded substantial variation in the color–letter correspondences made by each of the pair. The same variation is also seen among siblings and by mothers and daughters in the same family. It is surprising that there is not greater similarity in the color–letter correspondences of family members if colored alphabets are acquired as learned associations.
4. Lack of recollection. Most people with synaesthesia are unable to report *knowing* that their letter–color associations were learnt either purposefully or incidentally via exposure to colored alphabet letters or books.

The learned association theory of synaesthesia has not yet provided satisfactory explanations of these anomalies.

The Genetic Theory of Synaesthesia

The possibility that synaesthesia might be an inherited trait seems to have first been put forward by Galton (1883). Genetic mechanisms might cause the preserved neural connectivity described above. Earlier we reviewed the evidence for transitory connections between auditory and visual brain areas in other mammalian species. Assuming that such connections are also to be found in our species, one explanation for synaesthesia is that in individuals with the condition these neonatal pathways persist due to inherited mechanisms. A recent study (Baron-Cohen *et al*., 1996) has provided evidence to support the notion that synaesthesia might be an inherited trait. In that study, the pedigrees of seven families of probands suggested that the condition is inherited.

If the genetic theory is supported, this begs the question of by what mechanism such a biological inheritance has its effect. A candidate mechanism would be the expression of genes that regulate the migration and maturation of neurons within the developing brain. A second candidate mechanism is 'neuronal pruning' (apoptosis). On this account synaesthesia can be best explained not by positive forces creating neural pathways that in non-synaesthetes do not exist, but by maturational effects that lead to neonatal pathways being left active. This would be consistent with Maurer's observations regarding the emergence of modality-specific responses in three-month-old human neonates.

The Cross-modal Matching Theory

This is based on evidence of cross-modal matching in normal subjects, in addition to those found by Lewkowicz and Turkewicz described earlier. Much of the work looking at cross-modal analogs of characteristics such as brightness/loudness, etc. has been carried out by Marks (Marks, 1982a, b, 1987).

Marks (1982a) showed that normal subjects exhibited remarkable consistency when asked to rate a selection of auditory-visual synaesthetic metaphors using scaled ratings of loudness, pitch, and brightness. For example, 'sunlight' was rated as louder than 'glow', which was in turn rated as louder than 'moonlight'. A second study reported by Marks (1982b) required subjects to set the loudness of a 1000-Hz tone and the brightness of white light for 15 cases of visual-auditory metaphor taken from works of poetry. Again, marked consistency characterized the performance of these subjects, leading Marks to propose that intensity might be a common sensory dimension.

The Modularity Theory

In order for us to 'know' that a percept is visual, auditory, olfactory, etc. we must have a method of identifying information as being of one sensory kind or another. We may achieve this via a *modular* structure to sensation (Fodor, 1983). The modularity theory holds that whereas in non-synaesthetes, audition and vision are functionally discrete, in individuals with synaesthesia a breakdown in modularity has occurred (Baron-Cohen *et al*., 1993). The consequence of this, in the case of colored-hearing synaesthesia, is that sounds have visual attributes. Testing the modularity theory is a challenge for future research.

RELEVANCE OF SYNAESTHESIA FOR COGNITIVE SCIENCE

Investigations of colored-hearing synaesthesia suggest that individuals with the condition are

consistent in their descriptions of word–color correspondence and report similar phenomenological accounts of the condition. Further, synaesthetes appear to show different patterns of brain activation when listening to color-evoking sound stimuli. The existence of unusual neural connections between auditory and visual areas has been postulated to explain synaesthetic experience, perhaps as the result of a failure of apoptosis. Recent accounts of familiality of the condition suggest that genetic factors may sustain neonatal auditory-visual pathways. If this proves to be the case, synaesthesia may teach us how unusual wiring in the brain can lead to altered perception, and how genes may affect subjective experience.

References

Aghajanian GK (1981) In: Hoffmeister R and Stille G (eds) *Handbook of Experimental Pharmacology* **55**(2): 89–110.

Baron-Cohen S, Burt L, Laittan-Smith F, Harrison JE and Bolton P (1996) Synaesthesia: prevalence and familiality. *Perception* **25**(9): 1073–1079.

Baron-Cohen S, Harrison J, Goldstein L and Wyke M (1993) Coloured speech perception: is synaesthesia what happens when modularity breaks down? *Perception* **22**: 419–426.

Baron-Cohen S, Wyke M and Binnie C (1987) Hearing words and seeing colours: an experimental investigation of synaesthesia. *Perception* **16**: 761–767.

Calkins MW (1893) A statistical study of pseudo-chromesthesia and of mental-forms. *American Journal of Psychology* **5**: 439–466.

Critchley EMR (1994) Synaesthesia. *The Neurological Boundaries of Reality*. London, UK: Farrand Press.

Cytowic RE (1989) *Synaesthesia: A Union of the Senses*. New York, NY: Springer-Verlag.

Cytowic RE (1993) *The Man Who Tasted Shapes*. New York, NY: Putnam.

Cytowic RE and Wood F (1982) Synaesthesia: a review of major theories and their brain basis. *Brain and Cognition* **1**: 23–35.

Dehay C, Bullier J and Kennedy H (1984) Transient projections from the fronto-parietal and temporal cortex to areas 17, 18 and 19 in the kitten. *Experimental Brain Research* **57**: 208–212.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Galton F (1883) *Inquiries into Human Faculty and Its Development*. London, UK: Dent & Sons.

Gray J, Williams S, Nunn J and Baron-Cohen S (1997) Possible implications of synaesthesia for the hard question of consciousness. In: Baron-Cohen S and Harrison JE (eds) *Synaesthesia: Classic and Contemporary Readings*. Oxford, UK: Blackwell.

Graziano MSA, Yap GS and Gross CG (1994) Coding of visual space by premotor neurons. *Science* **266**(5187): 1054–1057.

Jacobs L, Karpik A, Bozian D and Gøthgen S (1981) Auditory-visual synesthesia: sound induced photisms. *Archives of Neurology* **38**: 211–216.

Kennedy H, Batardiere A, Dehay C and Barone P (1997) Synaesthesia: implications for developmental neurobiology. In: Baron-Cohen S and Harrison JE (eds) *Synaesthesia: Classic and Contemporary Readings*. Oxford, UK: Blackwell.

Kennedy H, Bullier J and Dehay C (1989) Transient projection from the superior temporal sulcus to area 17 in the newborn macaque monkey. *Proceedings of the National Academy of Science* **86**: 8093–8097.

Lewkowicz DJ and Turkewicz G (1980) Cross-modal equivalence in early infancy: auditory-visual intensity matching. *Developmental Psychology* **16**(6): 597–607.

Marks L (1975) On colored-hearing synesthesia: cross-modal translations of sensory dimensions. *Psychological Bulletin* **82**(3): 303–331.

Marks LE (1982a) Bright sneezes and dark coughs, loud sunlight and soft moonlight. *Journal of Experimental Psychology: Human Perception and Performance* **8**(2): 177–193.

Marks LE (1982b) Synesthetic perception and poetic metaphor. *Journal of Experimental Psychology: Human Perception and Performance* **8**(1): 15–23.

Marks LE (1987) On cross-modal similarity: auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology* **13**(3): 384–394.

Maurer D (1993) Neonatal synaesthesia: implications for the processing of speech and faces. In: de Boysson-Bardies B, de Schonen S, Jusczyk P, McNeilage P and Morton J (eds) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Meltzoff AN and Borton RW (1979) Intermodal matching by human neonates. *Nature* **282**: 403–404.

Motluk A (1994) The sweet smell of purple. *New Scientist* **143**: 32–37.

Neville HB (1993) In: de Boysson-Bardies B, de Schonen S, Jusczyk P, McNeilage P and Morton J (eds) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Paulesu E, Harrison J, Baron-Cohen S *et al*. (1995) The physiology of coloured hearing. *Brain* **118**: 661–676.

Rang HP and Dale MM (1987) *Pharmacology*. Edinburgh, UK: Churchill Livingstone.

Rizzo M and Eslinger PJ (1989) Colored hearing synaesthesia: an investigation of neural factors. *Neurology* **39**: 781–784.

Sadato N, Pascual-Leone A, Grafman J, Ibanez V and Deiber MP (1996) Activation of the primary visual-cortex by Braille reading in blind subjects. *Nature* **380** (6574): 526–528.

Vernon PE (1930) Synaesthesia in music. *Psyche* **10**: 22–40.

Further Reading

Baron-Cohen S and Harrison JE (eds) (1997) *Synaesthesia: Classic and Contemporary Readings*. Oxford, UK: Blackwell.

Halligan PW, Hunt M, Marshall JC and Wade DT (1996) When seeing is feeling – acquired synaesthesia or phantom touch. *Neurocase* **2**(1): 21–29.

Honig MG and Hume RI (1989) Dil and dio – versatile fluorescent dyes for neuronal labeling and pathway tracing. *Trends In Neurosciences* **12**(9): 333.

Krohn WO (1893) Pseudo-chromesthesia, or the association of colors with words, letters and sounds. *American Journal of Psychology* **5**: 20–39.

Paulesu E and Frith C (1997) *The neurobiology of synaesthesia*. In: Baron-Cohen S and Harrison JE (eds) *Synaesthesia: Classic and Contemporary Readings*. Oxford, UK: Blackwell.

Ramachandran VS and Rogers-Ramachandran D (1996) Synaesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society of London Series B – Biological Sciences* **263**(1369): 377–386.

Starr F (1893) Note on color-hearing. *American Journal of Psychology* **51**: 416–418.

Zellner DA and Kautz MA (1990) Color affects perceived odor intensity. *Journal of Experimental Psychology* **16**(2): 391–397.

Thought Experiments

Intermediate article

Tamar Szabó Gendler, Syracuse University, Syracuse, New York, USA

CONTENTS

Thought experiment: to perform a thought experiment is to reason about an imaginary scenario with the aim of confirming or disconfirming some hypothesis or theory.

WHAT ARE THOUGHT EXPERIMENTS?

To perform a thought experiment is to reason about an imaginary scenario with the aim of confirming or disconfirming some hypothesis or theory. In its original usage, the expression was reserved for cases intended to evoke intuitions about the physical world; more recently, it has also been used to refer to cases intended to evoke intuitions concerning the proper application of nearly any descriptive or evaluative concept.

So, for instance, Galileo's famous refutation of the Aristotelian view that heavy bodies fall faster than light ones is a paradigmatic example of a scientific thought experiment concerning the physical world. Galileo asks his reader to imagine a heavier body strapped to a lighter one, and shows that the Aristotelian is committed to saying that the joined object will fall both faster and more slowly than the heavier body alone. By contrast, John Searle's (1980, 1984) case of the Chinese Room is a classic example of a philosophical thought experiment concerning the application of our concepts. In an effort to undermine the thesis that a suitably programmed computer might manifest understanding, Searle asks his reader to consider whether a person locked in a room with a sheaf of Chinese characters and a set of instructions enabling her to select certain batches of characters ('answers') when prompted by certain other batches of characters ('questions') would be properly credited with understanding Chinese. (Searle expects his reader to give a negative answer.) Other examples are presented and discussed below.

Although Ernst Mach is generally credited with having coined the expression *Gedankenexperiment* in his 1897 essay of the same name, and although contemporary German, English, and French usage can be traced to Mach's writings, the expression *Gedankenexperiment* appears in the Danish Kantian Hans Christian Örsted's 1811 'Prolegomenon to the General Theory of Nature', and a term for experiment with thoughts – *mit Gedanken experimentieren* – can be found in a 1793 entry to German polymath Georg Christoph Lichtenberg's 'Common Place Book' (cf. Lichtenberg, 1793/1983; Mach, 1897; Mach, 1905/1976; Schildknecht, 1990, pp. 147ff; Witt-Hansen, 1976).

In any case, use of the method antedated its labeling by several thousand years, having been employed by ancient and medieval philosophers and natural philosophers, and by scientists and philosophers in the early modern and contemporary periods (for representative discussions, see Rescher, 1991; King, 1991; and other papers collected in Horowitz and Massey, 1991). After the publication of Mach's 1897 essay, the term itself seems to have taken roughly four decades to become widespread in scientific circles. (Despite his extensive reading of Mach, for instance, Einstein appears not to have used the expression in his own writings. In general, however, it is difficult to trace reliably the term's history, as later editions of works often interpolate it where it was not originally used.) Employment of the expression 'thought experiment' in its philosophical sense seems to have begun sometime in the 1970s, and it was only in the last decade of the twentieth century that philosophical reference works began to include entries for the term. (For an extensive

bibliography of the philosophical literature on thought experiment, see Gendler, 2000, pp. 229–250.)

Given how broadly the term is used, it seems that nearly any imaginary example might reasonably be termed a 'thought experiment'. As a matter of sociological fact, however, the expression tends to be reserved for cases involving a certain degree of visualization, complexity, or novelty. So, for instance, although they describe imaginary scenarios whose consideration may play some role in confirming or disconfirming some hypothesis or theory, simple examples in physics books ('a car travelling at 65 miles per hour strikes a concrete wall …') are rarely considered material for thought experiments, nor are their equally austere analogs in philosophy, psychology, linguistics, law, and so on.

KINDS OF THOUGHT EXPERIMENTS

Although a number of taxonomies for thought experiments have been proposed, none has become canonical. Perhaps the most widely accepted distinction is between scientific and philosophical thought experiments, though these categories are rarely made precise: scientific thought experiments are simply those concerning scientific subject matter, philosophical thought experiments those concerning nonscientific subject matter (cf., for instance, Horowitz and Massey, 1991; Sorensen, 1992).

A more sharply focused version of the scientific/nonscientific distinction is made by George Bealer (1998, pp. 207–208), who distinguishes imaginary cases that are used to evoke physical intuitions from those used to evoke intuitions about the application of nonphysical concepts. The former involve asking the reader to determine what would happen in a given imaginary scenario assuming that natural laws are held constant; the latter involve asking the reader to decide whether a particular scenario is logically or metaphysically possible, or whether a given concept applies to such a scenario. Bealer maintains that the term 'thought experiment' should be reserved for cases of the former sort, roughly the class generally referred to as 'scientific thought experiments'. Tamar Szabó Gendler (2000, pp. 25–27) suggests a slightly different taxonomy, distinguishing between factive and conceptual/valuational thought experiments. Factive thought experiments are those where the question asked is naturally described as 'what would happen?'; conceptual/valuational thought experiments are those where the question asked is naturally described as 'how should we describe or evaluate this outcome?' Thought experiments that are factive tend to be those involving scientific subject matter; thought experiments that are conceptual/valuational tend to be those involving philosophical subject matter.

James Robert Brown (1991) provides a taxonomy of scientific thought experiments that has gained some currency in certain philosophy of science circles. Brown distinguishes between destructive and constructive thought experiments, subdividing the latter category into mediative, conjectural, and direct. Destructive thought experiments are those involving imaginary examples designed to raise difficulties for a particular theory; constructive thought experiments are those aimed at establishing a positive result. Within the class of constructive thought experiments, mediative thought experiments are those which facilitate the drawing of a conclusion from a specific, well-articulated theory; conjectural thought experiments are those where thinking about an imaginary scenario causes us to consider a phenomenon for which we then provide some sort of theoretical explanation; direct thought experiments are those that directly yield a well-established theory. Thought experiments that are simultaneously destructive and direct-constructive Brown calls platonic, since, he claims, they give us *a priori* knowledge of nature.

Other taxonomies have also been proposed, though like those described above, none has gained canonical status. Nicholas Rescher (1991), for instance, distinguishes between thought experiments that are explanatory and those that are refutatory, offering further subdivisions into six more precisely articulated methods. Sarah Thomason (1991) divides thought experiments in linguistics into two categories: those that identify what sort of evidence might be conclusive in testing a particular theory, and those that test linguistic hypotheses by providing introspective data (these might be called 'experiments-in-thought'). D. A. Anapolitanos (1991) offers a six-celled taxonomy of thought experiments in mathematics; Richard Gale (1991) distinguishes thought experiments that yield clear-cut counterexamples from those that result in undecideable cases; Allen Janis (1991) distinguishes three ways in which thought experiments in physics might fail; Roy Sorensen (1992, pp. 197–202) classifies thought experiments on the basis of whether the corresponding actual experiment is gratuitous, unaffordable, or impossible; Pierre Duhem (1914/1954, p. 202) similarly distinguishes merely unperformed experiments, experiments which could not be performed

with precision, physically unperformable experiments, and absurd experiments; and Sören Häggqvist (1996, pp. 136–159) and Kathleen Wilkes (1988, chap. 1) each present principles for distinguishing successful from unsuccessful thought experiments.

USES OF THOUGHT EXPERIMENTS IN PHILOSOPHICAL COGNITIVE SCIENCE

In the cognitive science literature, the term 'thought experiment' is generally used to refer to some widely discussed imaginary case designed to evoke intuitions about the proper application of a concept such as 'meaning' or 'consciousness'. So, for instance, among the cases generally referred to as 'thought experiments' are Frank Jackson's case of Mary the Color Scientist, Derek Parfit's case of fission, Hilary Putnam's case of Twin Earth, and John Searle's case of the Chinese Room. For whatever reason, discussions of zombies and inverted spectra are less commonly referred to as 'thought experiments', though slight variations on them, such as Ned Block's case of inverted Earth, generally are. Each of these cases is described briefly below, followed by a discussion of some of their common features.

Jackson's Mary

In an effort to undermine the view that all facts are physical facts, Frank Jackson (1982) presents the example of Mary, a person who has never had color experiences, having been confined all her life to a black and white room and denied all access to color-involving visual stimuli. Mary is also a brilliant scientist who specializes in the neurophysiology of vision, and who knows all physical facts (including all neurological facts) about color vision. Jackson asks what would happen if Mary were released from her confinement and shown a red object: would Mary learn anything new? Jackson (1982) expects his reader to agree that the answer is 'yes', and concludes that what Mary has learned when she has learned what it is like to see red is a nonphysical fact.

Parfit's Fission Case

In an effort to undermine the view that personal identity is what properly underlies our concern for our future continuants, Derek Parfit (1984/1987) discusses a pair of cases involving brain transplants from an individual in whom all cognitively relevant features are realized in duplicate – once in the left half of the brain, and once in the right. In the first scenario, the left half of the original person's brain is transplanted into the body of his decerebrated identical triplet, resulting in an individual qualitatively identical to the original in all bodily and psychological characteristics, while the right half of the original brain is destroyed. In the second scenario, both the left and right halves of the brain of the original individual are transplanted, each into the decerebrated body of one of his identical triplets, resulting in two individuals each qualitatively identical to the original in all bodily and psychological characteristics. Parfit suggests that the relation between the original individual and his successor in the first case is a relation of personal identity, and *a fortiori* is sufficient to render his prudential concern for that continuer rational. In the second case, the relation between the original individual and each of his two continuers is intrinsically identical to that in the first case; hence, contends Parfit, it is sufficient to render prudential concern for each of them rational. But a relation of identity does not hold between the original individual and both of his two continuers (since identity is a one–one relation). So, concludes Parfit, identity is not what matters in making prudential concern rational.

Putnam's Twin Earth

In an effort to show that an individual's social or physical environment is partly determinative or constitutive of that individual's mental states, Hilary Putnam (1975) presents the example of Twin Earth, a planet identical to Earth in all respects but one: the substance that plays the macro-role of Earthly water is not H_2O, but a substance with a different chemical structure that Putnam calls XYZ. Putnam imagines two individuals: one on Earth named Oscar, and his molecule-for-molecule Twin-Earth duplicate Twin-Oscar. Putnam holds that when Oscar says *water* he refers to water (that is, H_2O), but that when Twin-Oscar says *water* he refers to twin-water (that is, XYZ). So, concludes Putnam, reference is at least partly determined by physical environment. (See also Burge, 1979 for a number of parallel cases). (*See* **Externalism**)

Searle's Chinese Room

See the description in opening section.

Zombies

In an effort to bring out certain issues related to the nature of conscious experience and the plausibility

of physicalism, numerous philosophers have discussed the case of zombies, beings molecule-for-molecule identical to human beings but who lack all conscious experience (cf. Kirk, 1974; Dennett, 1991; Chalmers, 1996). On the basis of such cases, some have concluded that consciousness cannot be fully explained in physical terms.

Inverted Spectrum and Inverted Earth

In an effort to illuminate various issues relating to the status of qualia, materialism, behaviorism and consciousness, numerous philosophers have employed a case first introduced by John Locke (1689/1975 at II:XXXII:15). In its simplest form, the Inverted Spectrum example hypothesizes an individual whose visual experience on seeing, say, yellow is qualitatively identical to the visual experience of a normal person seeing, say, blue. Variations on the case abound. For instance, in arguing against certain representationalist and functionalist accounts of qualia, Ned Block (1990) introduces the example of Inverted Earth, a planet whose colors are inverted, so that grass on Inverted Earth is red and the sky on Inverted Earth is yellow. A person is transported to Inverted Earth, and given color-inverting contact lenses that cause everything on Inverted Earth to appear to her to be normally colored (cf. also Shoemaker, 1982; Chalmers, 1996). (*See* **Functionalism**; **Materialism**; **Qualia**)

Discussion of Common Features of the Above

Appellation notwithstanding, such cases tend to share the following features: (a) a scenario is described; (b) an intuition concerning the scenario is presented with the assumption that it will be endorsed, or some argument is presented for why a particular evaluation of the scenario is correct; and (c) this intuition or evaluation is then taken as a datum in understanding something about cases beyond the scenario. So, for instance, in the case of Twin Earth, the imaginary scenario described posits the existence of the planet on which something qualitatively identical to water has the chemical structure *XYZ*; the intuition Putnam expects the scenario to evoke is that speakers of English and speakers of Twin-English refer to something different by their use of the word *water*; and the larger lesson is that 'meanings [or at least reference] ain't just in the head' (Putnam, 1975, p. 227, italics omitted). In the case of fission, the imaginary scenario posits a pair of cases where the relations between the earlier and later individual(s) are qualitatively indistinguishable, but differ in their identity properties; Parfit's arguments aim to show that this gives us a case where prudential concern for a nonidentical continuer is rational; the larger lesson is that 'personal identity is not what matters' (Parfit 1984/1987, p. 255, italics omitted).

Challenges to particular thought experiments may come at any of these three levels: (a′) incoherence criticisms: the scenario described is in some sense incoherent; (b′) misleading intuition/unsound argument criticisms: although the scenario described is coherent, the intuition it generates is unreliable or the argument establishing the correct evaluation of the scenario is unsound; or (c′) inapplicability criticisms: although the scenario described is coherent and the evaluation of the scenario correct, the conclusion drawn on its basis is mistaken. So, for example, some have argued (a′) that fission is biologically or physically or conceptually impossible; others (b′) that though the scenario described is coherent, it does not present us with a case where someone would bear a relation of rational prudential concern to a nonidentical continuer; and others (c′) that though the scenario presents a case where someone would bear a relation of rational prudential concern to a nonidentical continuer, this does not show that identity is not what matters for rational prudential concern in ordinary cases.

PHILOSOPHICAL AND EMPIRICAL THEORIES OF THOUGHT EXPERIMENTAL COGNITION

Perhaps the most perplexing question raised by the technique of thought experiment is the epistemic puzzle articulated sharply by Thomas Kuhn: 'How, relying exclusively on familiar data, can a thought experiment lead to new knowledge?' (Kuhn, 1964/1977, p. 241). The question can be broken in two: (a) how can thought experiments lead to beliefs that are properly classified as *new*? (b) how can thought experiments lead to beliefs that are properly classified as *knowledge*?

Classic rationalist discussions answer both questions simultaneously by suggesting that in certain cases, thought experimental reasoning can lead to rational insight and thereby give access to *a priori* truths (for a modern defence see Brown, 1991). Classic empiricist answers, such as Mach's, suggest that thought experiments provide access to unsystematized empirical knowledge itself acquired through experience or evolution. The justification for beliefs formed thereby is thus

parasitic on the basic knowledge; their novelty is a consequence of its having been previously unavailable in propositional form. Kuhn's own answer is that thought experiments work by forcing a simultaneous rethinking of conceptual structures and the information they contain, and in this way are able to yield beliefs that are both novel and justified.

Recent discussions of thought experimental cognition have tended to focus on whether the structured contemplation of imaginary examples produces distinctive sorts of cognitive access, rendering thought experiment epistemically indispensable. In a series of articles, John Norton (1991, 1996) has argued against this position, defending instead the view that thought experiments are arguments of a certain sort. Norton's view has been widely discussed and criticized by those who, following Mach (1905/1976, 1933/1960), hold that at least some knowledge accessed by thought experiment is nonpropositional or nonconceptual, and that contemplation of imaginary cases gives us access to that knowledge in a way that argument alone cannot (cf. Arthur, 1999; Brown, 1991; Gendler, 2000, chap. 2). Some, such as Nancy Nersessian (1993) and Nenad Miscevic (1992), have tried to make Mach's notion more precise by assimilating the technique of thought experiment to recent psychological work on mental modeling. Others, such as Daniel Dennett (1984), have suggested that many philosophical thought experiments are best understood as 'intuition pumps'. Yet others stress the parallels between thought experiments and actual experiments, contending that similar explanations can be offered for the utility of each (cf. Sorensen, 1992; Gooding, 1992).

CONTROVERSIES AND ISSUES REGARDING THE USE OF THOUGHT EXPERIMENTS

Controversies concerning thought experiments can be divided into two main categories: controversies about the standard interpretations of particular thought experiments, and controversies about the utility of the methodology itself. Even those whose concern is with the methodology itself, however, tend to be opposed to the use of far-fetched examples, rather than to the technique of reasoning about imaginary cases as such.

Controversies about particular thought experiments are generally expressions of substantive philosophical disagreements. For instance, debates about what, if anything, Mary learns when she leaves the black-and-white room; about whether the person locked in the Chinese Room understands Chinese and if not, what that shows; about whether zombies are negatively conceivable (not *a priori* incoherent) or positively conceivable (verified by a clearly and distinctly conceivable scenario), and if so what that implies about the status of physicalism or the nature of consciousness – each involves conducting a significant philosophical debate primarily through discussion of a particular imaginary case. Similarly, debates about the proper understanding of particular scientific thought experiments – for instance, Einstein and Bohr's 1930 debate concerning the clock-in-the-box – can also be understood along these lines (cf. Bohr, 1949). Occasionally, however, disagreements about a particular case are better understood as disputes about the methodology of thought experiment; this is particularly striking in debates about whether the concept of personal identity is sufficiently far-reaching to deliver reliable intuitions about fission cases.

In general, uneasiness with the methodology of thought experiment tends to be focused on thought experiments involving far-fetched cases, though there are certain strands of Marxist thought that stress the importance of focusing on the actual rather than the hypothetical, and strains of moral particularism that suggest that no situation may stand in as surrogate for another (cf. Dancy, 1985). Far more typical, however, are discussions such as those of Kathleen Wilkes (1988), who expresses misgivings about the use of wildly fantastic imaginary cases in discussions of personal identity on the grounds that the intuitions they evoke are unreliable as guides to our actual conceptual commitments. W. V. O. Quine (1972) expresses similar reservations, claiming that our concepts are indeterminate in their application when we consider such bizarre cases. Others have offered parallel arguments from a Wittgensteinian perspective (e.g. Gale, 1991).

In recent years, two other areas of related interest have begun to be explored, both of which raise concerns for the reliability of thought experiment as a methodology. As before, the implications of these investigations will need to be considered on a case-by-case basis. Following the work of psychologists such as Daniel Kahneman and Amos Tverksy, demonstrating that human reasoning is, in a wide range of cases, subject to apparently intractable cognitive illusions, a number of philosophers and psychologists have begun to consider whether intuition itself is reliable in the ways that thought experimental reasoning seems to presuppose (see,

for instance, the papers collected in DePaul and Ramsey, 1998). In a related vein, though for reasons arising from a concern with the nature of modality and the relation between epistemology and metaphysics, a number of philosophers have begun to rethink the relation between conceivability and possibility (see, for instance, papers collected in Gendler and Hawthorne, forthcoming). If, as some suggest, what we can conceive (or fail to conceive) is unreliable as a guide to what is genuinely possible, or if we lack a reliable sense of what we are capable of conceiving, then reasoning about imaginary scenarios may be an ineffective means of confirming or disconfirming certain hypotheses or theories.

References

Anapolitanos DA (1991) Thought experiments and conceivability conditions in mathematics. In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 87–97. Savage, MD: Rowman and Littlefield.

Arthur R (1999) On thought experiments as *a priori* science. *International Studies in the Philosophy of Science* **13**(3): 215–229.

Bealer G (1998) Intuition and the autonomy of philosophy. In: DePaul M and Ramsey W (eds) *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. New York: Rowman and Littlefield.

Block N (1990) Inverted Earth. *Philosophical Perspectives* **4**: 53–79.

Bohr N (1949) Discussions with Einstein on epistemological problems in atomic physics. In: Schilpp PA (ed.) *Albert Einstein: Philosopher-Scientist*, pp. 199–242. La Salle, IL: Open Court.

Brown JR (1991) *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. New York and London: Routledge.

Burge T (1979) Individualism and the mental. In: French P *et al*. (eds) *Midwest Studies in Philosophy: Studies in Metaphysics*, pp. 73–122. Minneapolis, MN: University of Minnesota Press.

Chalmers D (1996) *The Conscious Mind*. New York and Oxford, UK: Oxford University Press.

Dancy J (1985) The role of imaginary cases in ethics. *Pacific Philosophical Quarterly* **66**: 141–153.

Dennett D (1984) *Elbow Room*. Cambridge, MA: MIT Press.

Dennett D (1991) *Consciousness Explained*. Boston, MA: Little, Brown.

DePaul MR and Ramsey W(eds) (1998) *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield.

Duhem P (1914/1954) *The Aim and Structure of Physical Theory*, translated by P Wiener. Princeton, NJ: Princeton University Press.

Gale R (1991) On some pernicious thought experiments. In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 297–304. Savage, MD: Rowman and Littlefield.

Gendler TS (2000) *Thought Experiment: On the Powers and Limits of Imaginary Cases*. New York, NY: Garland Press.

Gendler TS and Hawthorne JP (eds) (forthcoming) *Imagination, Conceivability, and Possibility*. Oxford, UK: Oxford University Press.

Gooding DC (1992) The cognitive turn, or, why do thought experiments work? In: Giere R (ed.) *Cognitive Models of Science*, pp. 45–76. Minneapolis, MN: University of Minnesota Press.

Häggqvist S (1996) *Thought Experiments in Philosophy*. Stockholm, Sweden: Almqvist and Wiksell.

Horowitz T and Massey G (eds) (1991) *Thought Experiments in Science and Philosophy*. Savage, MD: Rowman and Littlefield.

Janis AI (1991) *Can thought experiments fail?* In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 113–118. Savage, MD: Rowman and Littlefield.

King P (1991) Mediaeval thought-experiments: the metamethodology of mediaeval science. In: Horowitz and Massey (eds), pp. 43–64.

Kirk R (1974) Zombies vs. materialists. *Proceedings of the Aristotelian Society* **48** (supplement): pp. 135–152.

Kuhn T (1964/1977) A function for thought experiments. *Reprinted in The Essential Tension*. Chicago, IL: University of Chicago Press.

Jackson F (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–136.

Lichtenberg GC (1983) *Schriften und Briefe: Sudelbücher, Fragmente, Fabeln, Verse* (Erster Band) (ed.) FH Mautner. Frankfurt, Germany: Insel Verlag.

Locke J (1689/1975) *An Essay Concerning Human Understanding* (ed.) PH Nidditch. New York, NY: Oxford University Press.

Mach E (1897) Über Gedankenexperimente. *Poskes Zeitschrift für den physikalischen und chemischen Unterricht*, January 1897, pp. 1–5.

Mach E (1905/1976) Über Gedankenexperimente. *Erkenntnis und Irrtum,* Leipzig, 1905, pp. 183–199. Reprinted as: On thought experiment. *Knowledge and Error* (translation of 1926 edition of *Erkenntnis und Irrtum* by TJ McCormack and P Foulkes). Dordrecht, the Netherlands: Reidel, 1976, pp. 134–147.

Mach E (1933/1960) *The Science of Mechanics*, 9th edn, translated by T McCormack. La Salle, IL: Open Court Publishers.

Miscevic N (1992) Mental models and thought experiments. *International Studies in the Philosophy of Science* **6**(3): 215–226.

Nersessian N (1993) In the theoretician's laboratory: thought experiment as mental modeling. *Proceedings of the Philosophy of Science Association*, vol. 2, pp. 291–301.

Norton J (1991) Thought experiments in Einstein's work. In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 129–148. Savage, MD: Rowman and Littlefield.

Norton J (1996) Are thought experiments just what you thought? *Canadian Journal of Philosophy* **26**(3): 333–366.
Parfit D (1984/1987) *Reasons and Persons*. Oxford, UK: Oxford University Press.
Putnam H (1975) The Meaning of 'Meaning'. Reprinted in *Mind, Language and Reality: Collected Papers Volume 2*, pp. 215–271. Cambridge, UK: Cambridge University Press.
Quine WVO (1972) Review of *Identity and Individuation*. *Journal of Philosophy* **69**(16): 488–497.
Rescher N (1991) Thought experiments in pre-Socratic philosophy. In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 31–41. Savage, MD: Rowman and Littlefield.
Schildknecht C (1990) *Philosophische Masken: Literarische Formen der Philosophie bei Platon, Descartes, Wolff und Lichtenberg*. Stuttgart, Germany: Metzler.
Searle J (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**: pp. 417–424. [Peer commentary, pp. 425–449; reply by Searle, pp. 450–456.
Searle J (1984) *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.
Shoemaker S (1982) The inverted spectrum. *Journal of Philosophy* **79**(7): 357–381.
Sorensen R (1992) *Thought Experiments*. New York and London: Oxford University Press.
Thomason S (1991) Thought experiments in linguistics. In: Horowitz and Massey (eds) *Thought Experiments in Science and Philosophy*, pp. 247–257. Savage, MD: Rowman and Littlefield.
Wilkes K (1988) *Real People: Personal Identity without Thought Experiments*. Oxford, UK: Clarendon Press.
Witt-Hansen J (1976) H.C. Örsted, Immanuel Kant, and the thought experiment. *Danish Yearbook of Philosophy* **13**: 48–65.

Further Reading

Bunzl M (1996) The logic of thought experiments. *Synthese* **106**(2): 227–240.
Cargile J (1987) Definitions and counterexamples. *Philosophy* **62**: 179–193.
Fodor JA (1971) On knowing what we would say. In: Rosenberg JF and Travis C (eds) *Readings in the Philosophy of Language*, pp. 198–212. Englewood Cliffs, NJ: Prentice-Hall.
Gentner D and Stevens AL (eds) (1983) *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Giere R (ed.) (1992) *Cognitive Models of Science*. Minneapolis, MN: University of Minnesota Press.
Hintikka J (1999) The emperor's new intuitions. *Journal of Philosophy* **96**(3): 127–147.
Miller FD and Smith N (eds) (1989) *Thought Probes: Philosophy through Science Fiction Literature*. Englewood Cliffs, NJ: Prentice-Hall.
Popper K (1959) On the use and misuse of imaginary experiments, especially in quantum theory. *The Logic of Scientific Discovery*. New York: Basic Books.

Turing Test

Introductory article

David J Cole, University of Minnesota, Duluth, Minnesota, USA

CONTENTS

Alan Turing proposed a test for deciding whether a computer can think: if a computer could converse well enough to pass for human, it should be recognized as intelligent. This behavioral test is a natural extension to machines of one way we assess each other's intelligence; but it has been subjected to many criticisms, including the Chinese room argument, the frame problem, and claims that Gödel's results show that computers are limited in ways that human intelligence is not.

TURING'S PROPOSED TEST FOR MACHINE INTELLIGENCE

Can computers think? This is a question about what is possible, a metaphysical question concerning the nature of reality. Answering it requires exploring the nature of digital computation and of thought.

How could one tell if a computer was thinking? This is a very different question, an epistemic question about how one could know. The Turing test is a method for answering this second question. It was proposed in 1950 by the English mathematician Alan Turing (1912–1954) in a journal article entitled 'Computing machinery and intelligence'. Turing argued that the epistemic question was easier to address than the metaphysical question.

Turing's basic idea is this: if a person communicating online with a machine cannot tell by asking questions whether they are communicating with a human or a machine, then the machine is intelligent. It can do as well as a human in a very demanding task, conversation.

Turing was the leading twentieth-century theoretician of computers until his suicide at the age of 42 after he had been arrested for homosexual conduct and sentenced by the British authorities to receive estrogen injections. When he was just 24, Turing had created the idea of a simple abstract universal computer, now known as a Turing machine. This conceptual device helped provide a theoretical basis for understanding digital computers and their capabilities and limits.

Turing believed that suitably programmed computers could think intelligently, and would do so by the end of the twentieth century. He also believed that humans were prejudiced against the possibility of an intelligent machine. Furthermore, he thought that the question whether machines could think would embroil us in endless difficulties about the definition of 'thinking'.

Turing first describes a game he calls the 'imitation game'. The game is played by three people: a man, a woman, and an interrogator. The man and woman (A and B) cannot be seen or heard by the interrogator, who asks each of them written questions, either on slips of paper or via a remote keyboard. The object of the game is for the interrogator to determine which is the man, and which is the woman, while A and B both present themselves as women. Turing then asks what would happen if we replaced the man with a computer, and changed the game to one in which the interrogator must determine which is the human and which is the machine? Turing held that this imitation game test should replace the original question about whether machines can think. If human interrogators cannot tell the difference between machine and human in conversation, then it is reasonable to say that the machine can think: the machine passes the Turing test.

HISTORY OF THE TURING TEST

Two years before his 1950 article, Turing reported that he ran a similar test using a 'paper machine' for playing chess. A paper machine is a set of program instructions that are executed by a human using pen and paper. The human operator of the machine need not know how to play chess: the program determines the moves. Turing reports that a mediocre chess player found it 'quite difficult' to tell

whether a remote opponent was another human or a paper machine. Since Turing knew that an electronic computer could in principle execute the program of the paper machine, he had experimental evidence by 1948 that a computer might pass for a mediocre human chess player.

Turing was optimistic about the future success of computers passing the full version of his test (and he would no doubt have been pleased that a computer, Deep Blue, beat the world chess champion in 1997). The machines Turing knew and helped design were huge, rare, and very expensive. These were the size of a room, based on vacuum tubes, hand-wired with acoustic mercury delay lines for memory. Even with such machines in mind, Turing predicted that advances in programming would be rapid and that by the year 2000 a computer would be able to pass for a human 30 percent of the time after 5 minutes of questioning.

In the second half of the twentieth century, computer hardware advanced more rapidly than anyone, including Turing, could have foreseen. But progress towards passing the Turing test was much slower. By the 1960s, Joseph Weizenbaum had created a program, Eliza, that simulated a Rogerian psychotherapist. In conversation, the program took the initiative, asking questions that elicited responses in the subject, then using keywords in the subject's response to generate new questions (or, failing that, moving on to a new topic). Some subjects preferred Eliza to human therapists. Given the relative simplicity of the Eliza program (one version in BASIC is only 256 lines of code), Weizenbaum drew pessimistic conclusions about our ability to use computers wisely.

By the 1970s, more complicated programs had been developed, using stored background information and more sophisticated language parsers to generate answers to questions posed in natural language. Some researchers in artificial intelligence (AI) claimed that these programs could understand the questions posed in English.

In 1988, the New Jersey industrialist Hugh Loebner and the Cambridge Center for Behavioral Studies set up an annual contest to implement the Turing test. Loebner funds prizes for the best entrants. Since 1990, a Loebner Prize of $2000 has been awarded each year for the best entry, and a grand prize of $100 000 will be awarded to a program that produces responses indistinguishable from a human.

Many of the Loebner entrants have been related in design to the simple ELIZA program. In response both to the tricks that can be used to pass restricted Turing tests and to the charge that computers merely simulate understanding, various proposals have been made for more comprehensive tests. For example, one could require that the machine have sensors and actuators and so be able to identify objects that are sent to it by the interrogator. Such a machine would not only generate the verbal report that it had seen a hamburger, it would have been connected by sensors to the real world that it talks about during the Turing test.

ARGUMENTS FOR THE TURING TEST

The main justification for the Turing test is that it naturally extends ordinary ways of detecting intelligence. Turing does not propose a definition of intelligence. Instead, the test makes use of conversation: a familiar means by which people ordinarily appreciate and assess the intelligence of others. To determine how intelligent someone is, we typically ask questions. The questions may be very deliberate and systematic, as in an IQ test, or they may be more informal, as when one meets someone at a party, or interviews a candidate for a scholarship. Knowledge, understanding, and the ability to make inferences is quickly revealed in linguistic production.

An additional consideration concerns the concept of intelligence itself. One influential approach to understanding intelligence and other mental properties was forcefully described by the British philosopher Gilbert Ryle in his book *The Concept of Mind*, published the year before Turing published his paper. In Ryle's analysis, to be intelligent is just to do things intelligently. Intelligence is not a thing, a hidden faculty or force; intelligence is a way of handling problems (including the problem of how to respond to questions). Like the property of fragility, intelligence is a dispositional property: it is manifest under certain conditions. One sees that someone is intelligent by observing how he or she behaves. If this analysis is correct and intelligence is manifest in behavior, then an appropriately designed behavioral test such as the Turing test should be adequate for detecting intelligence.

Furthermore, Turing and others have remarked on the analogies between human brains and computing machinery. Brains are composed of billions of very similar elements whose primary operation appears to be activating or inhibiting other similar elements. Cognitive psychologists have noted that complex behavior is generated by the interaction of many simple subsystems in the brain. Many cognitive scientists believe that the sub-processes that ultimately produce intelligent behavior can be described as computations on information from

memory and the senses. If we ourselves are machines running computational processes, then wouldn't machines employing electronic technology running computationally equivalent processes also be intelligent? And if these same processes are in practice necessary to produce intelligent behavior, then wouldn't a behavioral test such as that envisioned by Turing reveal intelligence?

Turing believed that humans were prejudiced against the possibility of an intelligent machine. He thought that many had theological reasons for believing that thinking required a soul, or would be threatened by the implication that thought was not exclusively human. Turing therefore designed the test conditions to conceal the identity of the test subject.

Daniel Dennett has defended the Turing test. He participated in early administering of the Loebner version of the Turing test, although he later distanced himself from the project. Dennett points out how very difficult it would be for a machine to pass the Turing test with unrestricted subject domains under reasonably astute questioning. Vast amounts of knowledge would be required, including much background knowledge that is not explicit in books: for example, that lawyers usually know how to tie their shoes, that lawyers are unlikely to be found in the company of lumberjacks, that peanut butter is a poor ingredient for cocktails, or that people look silly with their shirt label sticking up on the back of their neck. Dennett also reminds us that the Turing test is not meant to define intelligence, just to test for it. Nor is the test meant to be so absolute that it might not be made to give an incorrect response with luck and clever programming. Rather, it is like other tests – of gasoline octane, of human competence, of water purity – that draw general conclusions from a sample. Dennett concludes that with able questioners asking subtle questions that draw on life's many experiences, the Turing test is adequate for what Turing intended, namely to reveal an intelligent computer.

PROBLEMS WITH THE TURING TEST

The main problem with the Turing Test is that it is behavioristic. The test examines only observable behavior, not how it is produced. This raises the possibility that something unintelligent might appear to be an intelligent being, by merely simulating intelligence. In principle, programmers might merely produce a huge list of likely test questions, paired with reasonable answers. Then the machine need have no intelligence; it could just be an automated look-up machine. In practice it is impossible to cover every possible question in such a list. However, there could be simple routines for handling unknown subject matter, such as responding 'I don't know' or 'I don't like to talk about …'. Indeed, just such methods have enabled the simple Eliza-type programs to pass the Turing test and to pose as human beings on the internet. But these programs are not intelligent. They do not understand the language they use, and whatever intelligence they may seem to display is entirely that of their programmers. If such a program produces a clever response to a question, we should admire the programmer, not the machine that looked up the response in a table. Computers running such programs have won the Loebner contest, but they do not think.

A quite different criticism of the Turing test is based on Gödel's incompleteness theorem. In 1931 Kurt Gödel proved that certain truths of arithmetic were not provable in a formal logic system. Some thinkers, notably J. R. Lucas in the 1960s and Roger Penrose in the 1990s, have taken this to show that human intelligence exceeds that of any possible formal symbol manipulator. If the Gödelian objection is sound, a computer could be unmasked as a result of limitations to which humans are not subject.

It is highly controversial whether Gödel's theorem implies that there is a difference between human and machine capabilities. Turing himself was very familiar with Gödel's theorem, and proved related results of his own. But he (along with many others) did not believe that these results showed that there was an important difference between human and machine potential for intelligence. Turing argued that intelligence is compatible with making mistakes, and in fact that under certain circumstances intelligent people will make mistakes where a less intelligent person would not. Therefore results about what is provable without mistake do not show that machines cannot be intelligent. Indeed, Gödel's theorem establishes a limitation on what anyone could prove using certain techniques; it does not establish specific limitations on computers, or that humans have access to proof techniques that machines cannot have.

Another criticism of the Turing test, and probably the best known, is John Searle's Chinese room argument. Searle argues that computers, no matter how they behave, cannot really understand language: they can merely simulate understanding. This problem is not confined to the look-up programs described above: Searle argues that it is a problem with any kind of program. Searle argues

that the Turing test cannot show that computers think or understand the questions that are put to them. His argument is simple: a person could do exactly what a computer does in producing answers to questions, yet not understand the language in which the questions are posed. An example is Turing's own paper machine mentioned above. The human operator of a paper chess machine need not know anything about chess. The operator just manipulates some symbols on paper, such as 'KP2–KP4', in accordance with thousands of instructions, yet the result may appear to be a proficient chess player. Similarly, the human operator of a paper program that passes the Turing test need not know the language in which the test was conducted.

Searle imagines a human who only speaks English, working in a room operating a paper machine. Questions posed in Chinese are slipped under the door of the room. The operator inside the room might match the Chinese symbols with those on numbered cards and thereby convert the Chinese input to strings of numbers. Using this numerical input the operator of the paper machine runs the program, and finally, after another conversion, produces strings of Chinese symbols and passes them back out under the door. If the program of the paper machine is a good one, it will appear to those outside the room submitting the questions and reading the answers that the occupant of the room understands Chinese. However, the occupant of the room merely runs a cleverly designed program. And so, computers running similar programs may appear to understand language, but they don't. Passing the Turing test can never show that a computer has intelligent understanding.

Many accept Searle's conclusion, but many do not. One reply to Searle is that whether or not the person running the paper machine understands the language of the inputs and outputs is irrelevant to whether the inputs and outputs are understood – the mind doing the understanding is not the paper machine operator's. The personality, knowledge, and intelligence displayed in answering the questions will not be those of the paper machine operator. Another mind is created by running the program. The understander exists as a result of the processes in the machine, but this mind is not identical with the machine, any more than a person is identical with his or her body (which may still exist even if the person is dead).

Searle draws a larger conclusion from his argument, namely that computers are confined to syntactic manipulation of symbols, and that one cannot get semantics or meaning from syntactic transformations. A computer can manipulate symbols, but it does not understand what the strings of symbols mean. Searle does not deny that a physical system can understand – he holds that we are physical systems that understand meaning and are intelligent. His negative conclusion applies just to digital computers. Just as it is intrinsic to the nature of 'talking' clocks and dolls to not understand, no matter how appropriate their responses, so it is intrinsic to the nature of digital computers that they do not actually understand language, no matter how appropriate their responses on some occasions. Though a computer may simulate intelligent conversation by syntactic manipulations of strings of symbols, the computer will never understand what any of the symbols mean. Computers are inherently simulators: their realities are always virtual.

Against this claim it can be maintained that computers are not inherently either simulators or syntax manipulators. Computers are complex causal systems. Humans attribute syntactic significance to the voltages in these electronic systems because that suits human purposes and aptitudes. A computer that alters the spark advance in an automobile engine in response to engine load does not simulate a mechanical spark advance system; it replaces the mechanical analog system with a more flexible digital one.

Another way of responding to Searle's argument against computer intelligence is to claim that it involves a confusion of levels. Thus, while the digital spark advance system uses digital techniques at the level of implementation, its output varies continuously as a function of the input: its overall behavior is analog. Similarly, a microchip that replaced a neuron in someone's brain might use digital techniques to calculate when to fire, while at the level of its overall function it might behave exactly as the neuron it replaced. From the standpoint of other neurons, it is simply another causal element. That it achieves its high-level causal role by low-level digital processes is irrelevant to the character of the larger system. On this view, a digital computer can realize a system of another type – even a system with the causal architecture of a mind. The digital realization may be at a low level, as it is with a neural network, or it may be at a higher functional level.

Other defenders of the possibility of computer intelligence appeal to particular theories of meaning to respond to the claim that computers only simulate and only perform meaningless syntactic manipulations. Many philosophers believe that meaning is not inside the head (or the machine),

but that meaning exists because of connections between symbols and the world. Since those connections exist whether the symbols themselves are located inside a human head or a computer, a computer can operate on symbols that are intrinsically meaningful. On this view, meaning is not dependent on any operation of interpretation, whether by users of the computer or by the computer itself.

Besides the Chinese room argument, there are other arguments against the possibility of intelligence in computers. For example, Hubert Dreyfus has argued that computers lack other essential characteristics of human intelligence. Human intelligence involves unconscious abilities, and appreciation of context, which cannot be captured in the explicit rule-following to which computers are confined. Related to this argument is the 'frame problem'. The frame problem is defined differently by different researchers. Generally, to behave intelligently, a system must have a current representation of many features of the world – in particular, the relevant features. What should be the form and extent of its representational system? Information used by the system must be readily available and must be updatable. One cannot know everything, so one must be selective, and this appears to entail that one must know what information is relevant to solving the problem at hand. Humans do this intuitively. But computers must be given explicit rules for dealing with every aspect of a situation. A mouse, for example, can scurry across the floor and dodge a cat. But a program that could analyze the rapidly changing visual field of a robot mouse, as well as the position and pressure information from joints and touch sensors, and then go on to calculate the torques and movements needed at all the mouse joints to produce such behavior, would be enormously complicated, and perhaps impossible. And so even our best robots still move slowly and crash into things.

Defenders of traditional AI approaches argue that the problem could be solved using huge databases of information, parallel processing, and faster hardware. Others have held that while traditional programming approaches are open to the objection that computers only follow rules, or cannot handle the frame problem, connectionist approaches avoid these objections. In connectionist approaches, there are no rules or explicit representations in databases; these are replaced by networks, an idealization of the networks of neurons found in biological brains. These networks have shown success in pattern recognition and other areas where traditional programming has produced disappointing results.

Still others have held that the solution lies in decentralized systems that are highly adapted to their environments, and that such 'situated intelligence' does not need to model the world centrally or update the model constantly.

Other new approaches to AI are being developed, which may, according to their proponents, provide the means for producing truly intelligent machines.

Although Dennett has been a defender of the Turing test, he and others have been critical of the conduct of actual tests, including the Loebner Prize tests. In addition, he and others have argued that aiming to pass the Turing test has not been good for AI. Passing the test requires the ability to produce an illusion of humanness. Winning programs use special effects, not the reality of good artificial intelligence and a massive database of information reflecting experience in the world. To build such a database of common-sense 'world knowledge' has for many years been the aim of the CYC project in Texas. Critics of this project argue that it is still book learning, not real experience, and also that it is impossible to codify human experience with the world. Dennett has argued that in order to pass a reasonable Turing test, a computer would need to interact with the real world, with sensors and manipulators. Turing himself thought that to be intelligent, a computer would need to learn from experience.

Unfortunately, at this time we do not understand the nature of concepts, how learning takes place in humans, how humans solve the frame problem, how hypotheses are formed or stored, or the nature of mental representations. So our lack of understanding of thought and intelligence returns to haunt us: in order to build a machine that can pass the Turing test, we may have to confront those metaphysical questions that Turing sought to avoid.

SIGNIFICANCE OF THE TURING TEST FOR COGNITIVE SCIENCE

Turing believed that by the beginning of the twenty-first century, the use of the word 'intelligence' and educated opinion would have changed to the extent that it would be acceptable to speak of machines that think. Perhaps this change has come about. If so, it may be more a product of popular culture than of actual results in AI. A thinking clockwork machine appeared in L. Frank Baum's science fiction story *TikTok of Oz* in 1904. In the mid-1920s, robots were on stage and screen, in the

Czech playwright Karel Čapek's play *R.U.R.* (where the word 'robot' first appeared), and in Fritz Lang's film *Metropolis*, in which a (female) robot was so human that it could lead a human workers' rebellion. Since 1950, the year of Turing's article, robots have appeared frequently in film. In Stanley Kubrick's 1968 adaptation of Arthur C. Clarke's story *2001: A Space Odyssey*, a thinking computer, HAL, played a leading role. In 1977, an android robot in George Lucas's film *Star Wars* was talkative and emotional, while another robot was intelligent, but not humanoid in form and unable to speak human language. An android robot was an important character in the popular television series *Star Trek: The New Generation*. The possibility that machines might think is surely in the mind of everyone who has grown up with these characters and stories, not only in the minds of philosophers and theoreticians.

Yet there is no consensus that thinking machines are a future reality. The unrestricted Turing test is very difficult, and no machine has come close to passing it. There is no general agreement as to whether it is possible for a machine to pass the Turing test, whether it is worth attempting to build such a machine, or whether such a machine would necessarily be able to think. Perhaps the most enduring result of Turing's proposal is the amount of thought about thinking that it has generated.

Further Reading

Cole D (1991) Artificial intelligence and personal identity. *Synthese* **88**: 399–417.

Dennett D (1998) *Brainchildren: Essays on Designing Minds*. Cambridge, MA: MIT Press.

Dietrich E (ed.) (1994) *Thinking Computers and Virtual Persons*. New York, NY: Academic Press.

Dreyfus H (1972) *What Computers Cannot Do: A Critique of Artificial Reason*. New York, NY: Harper and Row.

Ince D (ed.) (1992) *The Collected Works of A. M. Turing*, vol. III 'Mechanical Intelligence'. Amsterdam and New York, NY: North-Holland.

Millican P and Clark A (eds) (1996) *Machines and Thought*. Oxford, UK: Oxford University Press.

Penrose R (1989) *The Emperor's New Mind*. Oxford, UK: Oxford University Press.

Ryle G (1949) *The Concept of Mind*. London: Hutchinson's.

Searle J (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**: 417–457.

Searle J (1984) *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Turing AM (1948/1969) *Intelligent machinery*. In: Meltzer B and Michie D (eds) (1969) *Machine Intelligence*, vol. V, pp. 3–23. Edinburgh, UK: Edinburgh University Press. [Reprinted in (Ince, 1992).]

Turing AM (1950) Computing machinery and intelligence. *Mind* **59**: 433–460. [Reprinted in Ince, 1992.]

Weizenbaum J (1966) ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM* **9**: 36–45.

Unconscious Processes

Intermediate article

Howard Shevrin, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

The existence and nature of unconscious processes have been subjects of philosophical and scientific controversy for centuries. Only recently have scientific methods been developed that may resolve the controversy.

INTRODUCTION

To attempt to define unconscious processes is already to enter into philosophical, psychological, physiological and even theological controversies that have endured for centuries. Part of the problem is reflected in the fact that the very term 'unconscious' is defined by what it is not – not conscious – rather than by what it might be. Yet the scientific stakes are high, because what is at issue is how the mind itself will be conceived and understood. (*See* **Consciousness, Philosophical Issues about**; **Implicit Cognition**)

WHAT ARE UNCONSCIOUS PROCESSES?

The most intense controversy in recent times has been over whether unconscious processes of a genuinely mental, rather than a purely physiological kind, exist at all. Among those who believe that unconscious processes are mental, there are further differences: some believe that unconscious mental processes are dispositional in nature, like a bias or prejudice that remains latent in the mind until an occasion arises to act upon it; others believe that unconscious mental processes are continuously active and influencing our conscious minds and actions in various ways, often indirect and covert. Those who believe that unconscious processes are intrinsically physiological in nature assume, either explicitly or implicitly, that 'mental' and 'conscious' are equivalent terms: nothing mental exists beyond conscious experience. Often proponents of this view prefer to talk about 'nonconscious' processes, perhaps as a way to stress their non-mental nature. (*See* **Consciousness and Representationalism**; **Perception, Unconscious**; **Blindsight**)

Only since the 1950s have unconscious processes emerged as a field of empirical inquiry that can be conducted in the psychological and neurophysiological laboratory – not, however, without critics and skeptics. This period of empirical investigation has been made possible by the development of a method, subliminal stimulation, which permits the objective study of unconscious processes. 'Subliminal stimulation' refers to the presentation of stimuli so quickly or so faintly that they cannot be seen or heard consciously, but whose effects can be detected in various ways.

HISTORY

Throughout much of the history of the study of unconscious processes, two main themes are discernible: on the one hand, unconscious processes have been associated with cognition and normal functioning; on the other hand, with desire, emotion, and psychopathology. These two themes have yet to be harmoniously integrated.

Cognition and Unconscious Processes

The idea that stimuli that were not conscious could somehow register unconsciously and influence consciousness 'imperceptibly' can be traced back to Aristotle (384–322 BC). He observed that one could divide a grain of millet into small particles each of which singly would not be perceptible, and yet in the undivided grain each particle contributes to the conscious perception of the grain (Aristotle, 1902).

Centuries later, Leibniz (1646–1716) proposed a similar idea: minute perceptions, each of which escapes consciousness but which in the aggregate

contribute to a conscious perception (or in his terminology, an apperception). Each minute perception could be said to be below the threshold of consciousness, but if enough of them were aggregated they would become sufficiently strong that they could cross the threshold (Leibniz, 1908).

Herbart (1776–1841) was the first to exploit more fully the notion of a threshold implicit in Leibniz' thinking. He invested ideas (the German word refers to both perceptions and thoughts) with properties of strength and quality, resulting in a continual dynamic, competitive interplay of ideas, some of which would become conscious and others of which would remain in a state of inhibition, that is, unconscious (Herbart, 1891). An important notion in Herbart's theory is that ideas become unconscious not because they are too weak, but because they are inhibited by other ideas. This notion would reappear in a very different context in Freudian psychoanalysis.

For Helmholtz (1821–1894), unconscious inference played a vital role in perception. According to Helmholtz, our perceptions can be modified by repeated exposure to the same object, or by the laws of association (Helmholtz, 1962). At first these modifications are conscious and involve conscious inferences; ultimately these conscious inferences become unconscious, and can in some instances produce illusions or misperceptions. For example, we would tend to perceive a circle with a small segment missing as a complete circle because of an unconscious inference based on multiple conscious experiences with circles. Helmhotz' concept of unconscious inference is a forerunner of such current ideas as unconscious automaticity and procedural memory. (*See* **Automaticity**; **Memory, Philosophical Issues about**; **Memory: Implicit versus Explicit**)

The first cognitive subliminal experiments were conducted in the late nineteenth century. Generally, these experiments sought to demonstrate that, even though subjects were guessing at faintly presented stimuli such as letters or words and claimed that they were not consciously able to discriminate among the stimuli, their guesses were consistently better than expected by chance (e.g. Sidis, 1898). Many cognitive subliminal experiments have been modeled after this paradigm.

In this long line of theory and emerging experimentation extending to the present day, only cognitive processes, mainly perception, have been of interest, and only the normally functioning mind was the intended subject of explanation.

Desire, Emotion, and Unconscious Processes

Aristotle also observed how faint noises might appear in dreams as thunder and lightning. On the face of it, this would not seem to fit with Helmholtz' idea of unconscious inference, because faint noises are not ordinarily associated with thunder and lightning. Something different must be happening unconsciously for the laws of association to be violated. The difference, some have argued, has to do with the influence of desire and emotion. Many have seen these as of paramount importance in understanding unconscious processes and how they influence consciousness and action.

This view of the unconscious was forcefully expounded by Schopenhauer (1788–1860), who saw the individual as driven by an entirely unconscious impetus, which he called the will, originating from the same driving force that animates the universe (Schopenhauer, 1992). In people, the will is most manifest in the sexual instinct, for whose gratification, according to Schopenhauer, people invent deceptions and rationalizations. (*See* **Sexual Arousal**)

Following in Schopenhauer's footsteps, Nietzsche (1844–1900) redefined the will as the will to power, or the elemental unconscious force that was the one true impetus for experience and action (Nietzsche, 1967). Nietzsche was perhaps the first to identify what Freud would later call 'defenses'. Nietzsche talked about inhibitions that kept immoral but powerful desires from consciousness. In this way, he added a motivational and emotional dimension to Herbart's purely cognitive view of inhibition.

Meanwhile, there was a growing interest in people who were considered possessed, or who were in one way or another driven by uncontrollable urges, or suffered from strange physical torments and lapses. Such behavior seemed to be beyond conscious control. Cures were invented that ranged from exorcism, to the laying on of hands, to hypnotism.

With hypnotism, we arrive at the threshold of modern approaches to dealing with those unconscious emotional and motivational processes that can produce 'symptoms', or indicators of underlying psychological disorder. (*See* **Hypnosis and Suggestion**)

Throughout the twentieth century, these two perspectives on unconscious processes – cognitive, and motivational/emotional – existed side by

side, with very little exchange between them. (*See* **Emotion, Philosophical Issues about**; **Motivation**)

UNCONSCIOUS PROCESSES IN PSYCHOANALYSIS

Freud (1856–1939) was introducted to unconscious processes when he observed Charcot (1835–1893) hypnotize hysterical patients at the Salpêtrière. Among Freud's foremost contributions to the debate on unconscious processes was his elevation of unconscious processes to the status of the mental (Freud, 1915). For Freud, the extent of consciousness was tiny compared with the vast depth and extent of unconscious processes. He thus reversed the relationship between conscious and unconscious processes that had previously been advanced in psychology, in which unconscious processes were essentially tributaries to conscious processes. Freud placed motivation – in the form of powerful drives or instincts – at the heart of unconscious processes, and in this respect he aligned himself with Schopenhauer and Nietzsche. More importantly from a scientific standpoint, he initiated the development of a complex method of inquiry and treatment, which he called psychoanalysis, that no longer depended on hypnosis to reveal what was going on unconsciously and did not depend on suggestion as a curative agent. It was through the psychoanalytic method that Freud sought to discover the nature of unconscious processes and how they affected consciousness. (*See* **Motivation**)

The psychoanalytic method was based on two premises: unconscious processes existed and were both mental and causative; and unconscious processes influenced consciousness and actions in indirect and covert ways. The way to gain access to these unconscious processes was through free associations. The patient was encouraged to relax, assisted by lying on a couch and not facing the psychoanalyst, and to say everything that came to mind. Since it was assumed that unconscious processes were active and influencing consciousness but would only reveal themselves indirectly and in covert form, the psychoanalyst paid particular attention to discontinuities in the flow of free associations: omissions, elisions, slips, and illogical jumps.

Freud claimed to have discovered that patients unconsciously fight against, or resist, revealing what is at work unconsciously, and that this accounts for the indirect and covert ways in which unconscious processes influence consciousness. Freud called the various psychological mechanisms creating these distortions 'defenses', a generalization of what Nietzsche had called inhibitions. Repression, or unconscious motivated forgetting, would produce omissions in the free association process. Other defense mechanisms took the form, for example, of displacements, or endowing a trivial idea with undue importance as in an obsession. By shifting the emotional weight to a trivial idea the patient could avoid becoming aware of what was really bothersome unconsciously. An obsession contradicts the principle of saliency, one of the laws of association, according to which importance is determined by objective relevance.

Defense mechanisms did not appear to follow ordinary logical rules of thought, often violating the laws of association to such an extent that Freud hypothesized that unconscious processes were organized on the basis of entirely different principles from the normal waking conscious state (Freud, 1911). Because he further believed that this mode of thought appeared in childhood, he called this organizing principle the 'primary process', to be followed in time by the development of the 'secondary process', which referred to the logical, rational mode of adult thought. With maturity, the primary process more and more characterizes unconscious processes, and the secondary process conscious processes, although no exact equivalence emerges. Freud hypothesized that unconscious processes obeying the principles of the primary process tended towards immediate gratification, regardless of considerations of time or the identity of the gratifying object; thus unconscious processes were characterized by Freud as timeless and violating the principle of identity. The unconscious intent was to make some past pleasure immediate, no matter what constraints prevailed; thus it was necessary to develop defenses against this intent. According to Freud, the common experience in which all these contending forces are most evident is in dreams, which he conceptualized as primarily wish-fulfilling (Freud, 1900).

This powerful tendency for past gratifications to impose themselves on the present was nowhere clearer for Freud than in the phenomenon of transference. According to his observations, patients would often respond to him not as the helping physician he tried to be, but more in terms of some model transferred from the past, usually that of a parent. Often this unconscious press for immediate gratification of past desires in the present runs into obstacles or moral constraints, thus creating conflict and the symptoms for which the patient comes for treatment.

Freud divided unconscious processes into the dynamic unconscious and the preconscious. The dynamic unconscious was the powerful instinctual unconscious which was being defended against, and making itself felt through covert means. The preconscious is the repository of all those unconscious processes that can readily become conscious. In that respect it is similar to the unconscious studied by cognitive psychologists.

One could say that in psychoanalysis unconscious processes came to be recognized as powerful mental determiners of conscious experience. But since the theory was entirely based on anecdotal clinical evidence rather than systematic experimentation, many psychologists and other scientists have been skeptical of Freud's theories.

UNCONSCIOUS PROCESSES IN COGNITIVE SCIENCE

With the advent of behaviorism in the early twentieth century, psychology turned away from the study of unconscious processes, and indeed of consciousness itself. The discipline of psychology was narrowed to behavior, or what could be observed objectively rather than reported on introspectively. In time, however, a new approach emerged which returned psychology first to its older interest in consciousness itself and then to the study of unconscious processes. This new approach was called cognitive science because of its primary interest in cognition: perception, memory, judgment, and thought. (*See* **Cognitive Science: Philosophical Issues**; **Behaviorism, Philosophical**)

The underlying model in cognitive science was the computer (Simon, 1979). Once it had been programmed, quite complex processes could occur inside the computer, and these processes would eventuate in a useful print out. The mind could be conceptualized as a computer in which mental processes were the counterpart of the computer's programmed operations, and consciousness was the counterpart of the printout. And since only the outcome, not the operations themselves, were 'printed out', the counterpart of unconscious processes would appear to be the computer operations themselves. Advocates of what has been called strong artificial intelligence (AI) believed that computers would eventually be designed that would in essence be minds. The workings of the mind would then be equivalent to computer operations, or computations. This approach led many cognitive scientists to recast mental processes as computations which were set into operation by programs autonomously existing in the mind. (*See* **Turing Test**; **Artificial Intelligence, Philosophy of**; **Mental Disorders, Computational Models of**; **Social Processes, Computational Models of**; **Computation, Philosophical Issues about**)

The conception of unconscious processes inherent in this computer-based approach makes them coextensive with everything that occurs prior to consciousness. The function of these unconscious computational operations is to subserve consciousness by producing its contents in the form of solutions to tasks assigned to the mind. Strangely, consciousness thus becomes epiphenomenal, having no inherent purpose of its own. (*See* **Epiphenomenalism**)

Another cognitive approach to unconscious processes derived from earlier ideas about habit and voluntary action. Helmholtz' 'unconscious inferences', forming as a result of many consciously performed inferences, were in effect habits, while conscious inferences necessitated voluntary choices or actions. In contemporary cognitive science the notions of automatic and controlled processes replace 'habit' and 'voluntary action'. (Shiffrin and Schneider, 1977). Automatic processes are defined as involuntary, effortless (i.e. not requiring attention), unconscious, and unintentional. Controlled processes, by contrast, are voluntary, necessitating effort (i.e. requiring attention), conscious, and intentional. To complete the picture one needs to add the concept of semantic networks, the contemporary version of the 'associations' of earlier psychology (Rumelhart, 1999). A semantic network is a reticulation of nodes and weighted links built up between mental representations of stimuli based on experience with the stimuli. When a stimulus is processed automatically (unconsciously), activation spreads through the network solely as a function of the existing nodes and their weighted links. This automatic process is referred to as spreading activation, and is different from the more selective activation that is under conscious control, as in remembering a name, or performing some task.

With this kind of division between controlled (conscious) and automatic (unconscious), there could be no place for motivated unconscious processes. Motivation could only operate consciously. The division between the cognitive and motivational/emotional conceptions of unconscious processes thus persists to this day. Fortunately, we now have available experimental tools with which to test hypotheses derivable from each of these two different views of unconscious processes.

EXPERIMENTAL WORK ON UNCONSCIOUS PROCESSES

Although there had been a few sporadic experimental attempts in the first half of the twentieth century it was with the advent of research on subception (Lazarus and McCleary, 1951) in cognitive psychology and the Poetzl effect (Poetzl, 1960) in psychoanalytic research that experimental investigations of unconscious processes began in earnest. More recently there has been an increasing number of neuroscience-based investigations of unconscious processes.

Cognitive Studies

Lazarus and McCleary (1951) investigated the effect of obscene words on recognition thresholds. They hypothesized that if unconscious processing of words existed then most people would delay conscious recognition of obscene words, so that the recognition thresholds for these words would be higher compared with non-obscene controls. This unconscious perception was named subception. These studies were criticized because often differentials in word frequency were not taken into account, and also because even if people were aware of the obscene words, they might hesitate to say them.

Dichotic listening experiments, which involve presenting stimuli to both ears but requiring the subject to pay attention to one ear, were conducted (Lewis, 1970). It could be demonstrated that stimuli presented to the unattended ear influenced the perception of the stimuli in the attended ear. But the most widely used procedure was based on the masking effect. When a briefly presented stimulus is immediately followed by a longer-lasting stimulus, consciousness of the first stimulus is 'masked', that is, not reported as seen and presumably unconscious. Many experiments indicated that the masked stimulus, although unconscious, was registered and influenced subsequent conscious perception (Marcel, 1983). (*See* **Masking**)

Social cognitive researchers have demonstrated that subliminal emotional facial expressions can influence the way a neutral masking stimulus is consciously perceived (Winkelman *et al*., 1997). Others have shown that subliminal stimuli can reveal unconscious prejudices that are consciously disavowed (Banaji and Greenwald, 1994). (*See* **Emotion**; **Prejudice**)

More recently, a method has been devised called the process dissociation paradigm, whereby subjects are instructed in one condition to exclude stimuli previously masked, and in another condition to include them (Jacoby, 1991). The measure of unconscious perception is based on the failure to exclude the masked stimuli when instructed to do so. Positive results have been obtained from such experiments. (*See* **Dissociation Methodology**)

Psychoanalytic Studies

The Poetzl effect was named after a Viennese contemporary of Freud who devised a means for investigating how dreams were affected by stimuli that had not been present in consciousness (Poetzl, 1960). Poetzl flashed pictures for a duration of 10 ms, asked his subjects to describe what they saw, noted what had not been reported, and then discovered that the unreported and presumably unseen elements of the picture were present in dreams reported by subjects the next day. Freud was impressed by this means of experimentally investigating dream formation. (*See* **Sleep: Polyphasic**)

The Poetzl effect was replicated by a number of psychoanalytically oriented researchers who developed a variety of measures in addition to dreams, such as images and free associations (Fisher, 1954). These early Poetzl effect studies were often anecdotal and lacking in adequate controls, but more rigorous experiments were conducted later. On the whole, results were promising. For the first time it could be demonstrated that a stimulus not in consciousness had measurable effects on subsequent conscious experiences.

In addition to Poetzl effect studies, other methods were devised to more directly test psychoanalytic propositions. The subliminal psychodynamic activation method was developed to test quite specific psychodynamic hypotheses about the nature of unconscious conflict related to various disorders. In several studies the presence of unconscious oedipal conflict in young men was investigated by subliminally presenting the words 'Beat Dad' (Silverman *et al*., 1978). It was hypothesized that the 'Beat Dad' stimulus would activate an unconscious oedipal conflict that would lower dart-throwing scores in a subsequent competition. Positive results were obtained in some but not all studies.

Another approach sought to incorporate a brain response indicator of unconscious conflict (Shevrin *et al*., 1996). In these studies, unlike the subliminal psychodynamic activation investigations, stimuli were individualized for each subject based on intensive interviews and tests. Words were selected by a panel of psychoanalysts to reflect the subjects'

unconscious conflict, and were compared with words reflecting the subjects' experience of their symptoms. These words, and suitable controls, were flashed subliminally as well as supraliminally. It was found that the brain responses grouped the unconscious conflict words together only when they were presented subliminally. This fact was interpreted as supporting two conclusions: the existence of unconscious conflict (a defining characteristic of the dynamic unconscious); and the operation of some kind of inhibitory or defensive process so that the unconscious conflict words were no longer grouped together when presented supraliminally. This experimental effect correlated with a measure of repressiveness, providing additional convergent evidence that defensive behavior was involved. (*See* **Neural Correlates of Visual Consciousness**; **Event-related Potentials and Mental Chronometry**; **Electroencephalography (EEG)**)

Neuroscience Studies

We have already mentioned one study involving brain responses in subliminal research on unconscious conflict. Other brain response studies have demonstrated that emotional meanings can be processed entirely unconsciously, resulting in physiological arousal and appropriate expressive signs (Bernat *et al*., 2001; Bunce *et al*., 1999). The invention of neuroimaging, enabling us to picture neural activity in different parts of the brain, has made it possible to localize the response to a subliminal fear stimulus in the amygdala, a small brain structure known to be important in emotion (Morris *et al*., 1998; Whalen *et al*., 1998). Neuroimaging research is still in its infancy and there are complex methodological issues that remain to be resolved. It is, however, likely that a method combining brain responses with neuroimaging will offer exciting possibilities to future investigators of unconscious processes. (*See* **Emotion, Philosophical Issues about**; **Neuroimaging**; **Amygdala**)

FOUNDATIONAL ISSUES CONCERNING UNCONSCIOUS PROCESSES

The first conclusion one can reach on the basis of research in cognitive science, psychoanalysis, and neuroscience is that unconscious processes exist and must be taken into account in any theory of the mind and how the brain functions. The focus of research can now shift to determining the nature of unconscious processes. Some cognitive scientists believe that unconscious processes are relatively unimportant in the larger economy of the mind. Evidence is advanced, for example, that the effects of subliminal stimuli last only briefly and leave no memory trace. Other evidence suggests that the effects can last for hours, if not days. Most cognitive psychologists are not concerned with motivation and emotion. Research inspired by psychoanalysis, on the other hand, demonstrates that unconscious motivational and emotional processes play an essential role in conflict and symptom formation. From this standpoint, subliminal stimuli can leave memory traces that can last a lifetime and exercise a powerful influence on consciousness and personality.

Motivation is a particularly contested issue in the study of unconscious processes. According to the 'automatic versus control' theory, motivation can only be conscious. Psychoanalytic research (Shevrin *et al*., 1996) and social cognitive research (Bargh and Barndollar, 1996) have begun to cast doubt on this position. Other research suggests that complex motivational and emotional factors may be at work even in straightforward cognitive tasks. Snodgrass *et al*. (1993) and Van Selst and Merikle (1993), for example, have found that individual strategy preferences play an important role in cognitive subliminal effects. In their experiments, subjects were asked to guess which of four words had just been flashed subliminally under two conditions, one in which they were instructed to simply let one of the four words pop into mind, and another in which they were encouraged to base their guesses on whatever they could see. Importantly, they were asked which strategy they preferred. It was found that when the subjects who preferred the 'look' strategy were asked to follow the 'pop' strategy their performance in guessing the correct word was consistently below chance. In order to explain this finding, some form of unconscious inhibition of correct responses had to be inferred. (*See* **Social Cognition**)

How motivation works unconsciously, and how unconscious processes are qualitatively different from conscious processes, remain challenging questions for future research. Psychoanalytic theory would predict that different principles of organization govern unconscious and unconscious processes. Brakel *et al*. (2000) have provided some evidence in support of this hypothesis by demonstrating that similarity judgments based on subliminal categorization follow primary process rules.

In addition to issues bearing on the roles of cognition, motivation, and emotion, there are also issues bearing on whether unconscious processes

are to be considered mental or physiological, dispositional or active. Searle (1992) has taken the position that unconscious processes are physiological and dispositional in nature. Only consciousness is mental in the sense of being about things, or, in more technical terms, as possessing intentionality. Unconscious processes, insofar as they can become conscious, have what Searle called derived intentionality. Unconscious processes are dispositional in that they exist latently until such time as they become conscious. (*See* **Intentionality**)

It could be argued, however, that the evidence drawn from both cognitive and psychoanalytic subliminal studies suggests that unconscious processes are just as mental as conscious processes and are independently active. It would be difficult to account otherwise for unconscious registration of word meanings, which are inherently referential and result in spreading activation of semantic networks independently of any conscious task.

SUMMARY

For centuries, unconscious processes have been thought of either as purely cognitive and subserving consciousness, or as motivational/emotional and interfering with or independent of consciousness. We are now beginning to see some integration of these two seemingly distinct viewpoints.

References

Aristotle (1902) *Aristotle's Psychology*, translated by WA Hammond. London: Sonnenschein.

Banaji M and Greenwald AG (1994) Implicit stereotyping and prejudice. In: Zanna MP (ed.) *The Psychology of Prejudice: The Ontario Symposium* **7**: 55–76. Hillsdale, NJ: Erlbaum.

Bargh J and Barndollar K (1996) Automaticity in action: the unconscious as repository of chronic goals and motives. In: Gollwitzer P (ed.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*, pp. 457–481. New York, NY: Guilford Press.

Bernat E, Bunce S and Shevrin H (2001) Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International Journal of Psychophysiology* **42**: 11–34.

Brakel LW, Kleinsorge S, Snodgrass M and Shevrin H (2000) The primary process and the unconscious: experimental evidence supporting two psychoanalytic presuppositions. *International Journal of Psychoanalysis* **81**(3): 553–569.

Bunce S, Bernat E, Wong PS and Shevrin H (1999) Further evidence for unconscious learning: preliminary support for the conditioning of facial EMG to subliminal stimuli. *Journal of Psychiatric Research* **33**(4): 341–347.

Fisher C (1954) Dreams and perception: the role of preconscious and primary modes of perception in dream formation. *Journal of the American Psychoanalytic Association* **2**: 389–445.

Freud S (1900) The interpretation of dreams. In: Strachey J (ed.) (1953) *The Standard Edition of the Complete Psychological Works of Freud*, vol. V, pp. 1–630. London: Hogarth Press.

Freud S (1911) Formulations on the two principles of mental functioning. In: Strachey J (ed.) (1958) *The Standard Edition of the Complete Psychological Works of Freud*, vol. XII, pp. 213–226. London: Hogarth Press.

Freud S (1915) The unconscious. In: Strachey J (ed.) (1957) *The Standard Edition of the Complete Psychological Works of Freud*, vol. XIV, pp. 150–216. London: Hogarth Press.

Helmholtz H and Southall JPC (ed. and trans.) (1962) *Helmholtz's Treatise on Physiological Optics*. New York, NY: Dover. [Translated from the 3rd German edition].

Herbart JF (1891) *A Textbook in Psychology. An Attempt to Found the Science of Psychology on Experience, Metaphysics, and Mathematics*, translated by MK Smith. New York, NY: Appleton.

Jacoby L (1991) A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language* **30**: 513–541.

Lazarus RS and McCleary RA (1951) Autonomic discrimination without awareness: a study of subception. *Psychological Review* **58**: 113–122.

Leibniz GW (1908) *The Philosophical Works of Leibniz*, 2nd edn, translated by GM Duncan. New Haven, CT: Tuttle, Morehouse and Taylor.

Lewis JL (1970) Semantic processing of unattended messages during dichotic listening. *Journal of Experimental Psychology* **85**: 225–228.

Marcel A (1983) Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual process. *Cognitive Psychology* **15**: 238–300.

Morris JS, Friston KJ, Buechel C *et al*. (1998) A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* **121**: 47–57.

Nietzsche FW and Kaufman W (ed.) (1967) *The Will to Power*, translated by W Kaufmann and RJ Hollingdale New York, NY: Random House.

Poetzl O (1960) The relationship between experimentally induced dream images and indirect images. *Psychological Issues* **2**: 41–120.

Rumelhart DE and Bly BM (ed.) (1999) *Cognitive Science*. San Diego, CA: Academic Press.

Schopenhauer A and Cartwright DE (ed.) (1992) *On the Will in Nature: A Discussion of the Corroborations From the Empirical Sciences that the Author's Philosophy has Received Since its First Appearance*, translated by EFJ Payne. New York, NY: St Martin's Press.

Searle JR (1992) *The Rediscovery of Mind*. Cambridge, MA: MIT Press.

Shevrin H, Bond JA, Brakel LAW, Hertel RK and Williams WJ (1996) *Conscious and Unconscious Processes:*

Psychodynamic, Cognitive, and Neurophysiological Convergences. New York, NY: Guilford Press.
Shiffrin R and Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review* **84**: 127–190.
Sidis B (1898) *The Psychology of Suggestion*. New York, NY: Appleton.
Silverman LH, Ross DL, Adler JM and Lustig DA (1978) Simple research paradigm for demonstrating subliminal psychodynamic activation: effects of Oedipal stimuli on dart throwing accuracy in college males. *Journal of Abnormal Psychology* **87**: 341–367.
Simon HA (1979) *Models of Thought*. New Haven, CT: Yale University Press.
Snodgrass M, Shevrin H and Kopka M (1993) The mediation of intentional judgments by unconscious perceptions: the influences of task strategy, task preference, word meaning, and motivation. *Consciousness and Cognition* **2**: 169–193.
Van Selst M and Merikle P (1993) Perception below the objective threshold? *Consciousness and Cognition* **2**: 194–203.
Whalen PJ, Rauch SL, Etcoff NL *et al*. (1998) Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience* **18**: 411–418.
Winkelman P, Zajonc RB and Schwarz N (1997) Subliminal affective priming resists attributional interventions. *Cognition and Emotion* **11**: 433–465.

Further Reading

Boring EG (1929) *A History of Experimental Psychology*. New York, NY: Appleton-Century.
Dixon NF (1971) *Subliminal Perception: The Nature of a Controversy*. London: McGraw-Hill.
Dixon NF (1981) *Preconscious Processing*. Chichester, UK: Wiley.
Ellenberger HF (1970) *The Discovery of the Unconscious*. New York, NY: Basic Books.
Muller-Freienfels R (1935) *The Evolution of Modern Psychology*. New Haven, CT: Yale University Press.

Vagueness

Intermediate article

Achille C Varzi, Columbia University, New York, New York, USA

CONTENTS

Vagueness arises whenever a concept or word admits of borderline cases of application. It can be viewed as a feature of human thought and language (conceptual or linguistic vagueness) or as a characteristic of the world itself (ontic vagueness).

WHAT IS VAGUENESS?

The standard view is that vagueness arises whenever a concept or linguistic expression admits of borderline cases of application. The predicate 'bald', for example, is vague because there can be situations in which it is indeterminate whether or not it applies to (a name of) a certain object. Some people are clearly bald (Picasso), some are clearly not bald (the count of Montecristo), and some are borderline cases – our concept of baldness and our linguistic practices do not specify any exact number of hairs that marks the boundary between the bald and the non-bald. Similarly, a singular term such as 'Mount Everest' is vague because there is no determinate way of tracing the geographical limits of its referent. Some rocks are clearly part of Everest and some are clearly not, but some rocks enjoy a borderline status.

There is, however, dispute concerning this way of characterizing vagueness. For the statement that a term t admits of borderline cases of application – that it is indeterminate whether or not such and such objects fall within the boundaries of the entity designated by t – can be given a *de re* reading, as in statement 1 below, or a *de dicto* reading, as in statement 2:

> The term t designates an entity x such that it is indeterminate whether such and such objects fall within the boundaries of x. (1)

> It is indeterminate whether the term t designates an entity x such that such and such objects fall within the boundaries of x. (2)

On the first reading the indeterminacy is ontological. The predicate 'bald' is vague, on this reading, because it stands for a vague set: there is no objective, determinate fact of the matter about whether the borderline cases are included in that set (or about whether they enjoy the corresponding property). Likewise, on this reading 'Mount Everest' is vague because it stands for a genuinely vague thing: there is no objective, determinate fact of the matter about whether the borderline rocks are part of the mountain. There may also be no determinate fact of the matter about when the mountain itself came into being, for the temporal boundaries of an object may be vague too.

By contrast, the *de dicto* reading corresponds to a purely linguistic (or conceptual) notion of vagueness. On this view the set of bald people is not vague at all. There are exactly 2^n sets of people (where n is the number of people at the present time), each with a precise membership function; yet our linguistic stipulations do not fully specify which of those sets can serve for the extension of the predicate 'bald'. There is, similarly, no vague mountain on this view: instead there are plenty of aggregates of matter, each with a precise spatiotemporal boundary, and when we say 'Mount Everest' we are just being vague as to which aggregate we are referring to.

The two views are not strictly incompatible, at least in so far as one may be willing to treat some vagueness as ontological and some as linguistic. One could also construe some terms as involving both sorts of vagueness: it would be indeterminate which particular sets or objects those terms designate, and the relevant candidates would include vague specimens along with sharp ones. However, these ways of combining ontological and linguistic vagueness have attracted little attention and current views on vagueness divide rather clearly between one approach and the other. (See Tye, 1990 and Lewis, 1993, for two representative position statements, and Evans, 1978, for a much-debated way of setting up the issue.)

PROBLEMS AND PARADOXES

In so far as vagueness involves borderline cases, whether *de re* or *de dicto*, it manifests itself semantically in the generation of truth-value gaps. If Jones is a baldish person, then the statement that he is bald appears to lack a definite truth-value; if it is indeterminate whether this rock is part of Everest, then the statement that it is part of Everest is likewise neither true nor false. This has been a natural source of concern for philosophers and logicians since Frege, for the admission of truth-value gaps amounts to a failure of the classical principle of bivalence.

The main source of concern, however, is that vagueness generates a deep puzzle. For not only do vague terms involve borderline cases: they also seem to involve borderline borderline cases, or borderline borderline borderline cases. For example, just as there is no sharp line between the bald and the non-bald there does not seem to be any sharp line between the bald and the baldish (or the baldish-ish). In neither case can a single hair make a difference. Intuitively, this means that our notion of baldness satisfies the following principle:

> For every n: if a man with n hairs on his head is bald, then a man with $n+1$ hairs on his head is also bald. (3)

(Let us suppose that baldness supervenes exclusively on the number of hairs.) However, it is enough to combine this principle with

> A man with no hairs on his head is bald. (4)

to reach the paradoxical conclusion that

> A man with $500,000$ hairs on his head is bald. (5)

(This can be shown by 500,000 repeated applications of the rules of universal instantiation and modus ponens.) In other words, the intuition that the applicability of 'bald' cannot be a matter of a single hair seems to force us to reason from the true premise that Picasso is bald to the false conclusion that the count of Montecristo is also bald. A corresponding point can be made about the intuition that the applicability of 'Everest' is not a matter of millimeters. In both cases it is hard to reach a diagnosis, but the clash between logic and intuition is deep.

In its oldest form, this problem is known as the *phalakros* paradox (from the Greek work for 'bald') and is attributed to Eubulides of Miletus, a contemporary of Aristotle. Eubulides is also credited with the formulation of the *sorites* paradox, which builds in a similar way on the vagueness of 'heap.' There are also versions of the paradox that rely on a different way of expressing the inductive principle 3. For example, the Stoics considered replacing the embedded conditional 'if … then' with a negated conjunction (it is not the case that a man with n hairs on his head is bald and a man with $n+1$ hairs is not bald). This makes the paradox even harder, since one cannot just blame the material conditional for the problem. Another common variant involves replacing statement 3 with a long chain of conditionals (or negated conjunctions), one for each relevant n. Again this makes the paradox more robust, since one cannot blame the universal quantifier for the problem. All such paradoxes are collectively referred to as sorites paradoxes. The problem with the notion of vagueness is that it generates such paradoxes, in some form or other. And theories of vagueness are naturally compared on the basis of how successful they are in providing a systematic resolution.

THEORIES OF VAGUENESS

Broadly speaking, there are two strategies for dealing with the sorites paradox. Focusing on the version exemplified by statements 3 to 5, one can either reject the argument as invalid, or reject it as valid but unsound. (One could also bite the bullet and accept the conclusion, but few would be willing to go that far.)

There are two main varieties of the first strategy. On the one hand, one can insist that logically valid reasoning can only be formulated in a precise language. This was, for example, the response advocated by Russell (1923) in the first full paper devoted to the topic of vagueness. Today this is not a popular position because it enforces an intolerable restriction on the scope of logic: since many of the words that we use in ordinary discourse (and even in much scientific discourse) are vague, logic would be of little practical use.

On the other hand, one can question the validity of the sorites argument by questioning the adequacy of classical logic. A popular approach is to adopt some kind of many-valued logic in which statements are allowed to take intermediate truth-values and in which the validity of the inference of statement 5 from statement 4 decreases as the number of applications of modus ponens increases.

In fact, because the notion of a borderline case is itself vague, a natural implementation of this strategy allows for a continuum of intermediate truth-values. The result is a fuzzy logic in which sentential connectives, for example, are represented by operations on the real numbers in the

interval [0,1] rather than on the two-valued truth set {0,1} (e.g. Machina, 1976). If vagueness is thought of as an ontological phenomenon, this account is naturally combined with a fuzzy semantics in which a predicate, for example, is assigned an extension whose membership function is itself continuum-valued. The closer to 1 the value is, the more the argument is a member of the set (Zadeh, 1965). This approach allows one to resolve the paradox as follows. First, the connectives are characterized in such a way that a conditional of the form

> If a man with n hairs on his head is bald, then a man with $n+1$ hairs on his head is also bald. (6)

is sure to be true or nearly true. For example, the truth-value of a conditional may be set to 1 minus the surplus of the antecedent over the consequent (if any). Secondly, there will be values of n such that the truth-value of the antecedent of statement 6 is slightly higher than that of the consequent. The underlying intuition is that one hair does make some difference after all, albeit a very small and generally negligible difference. Thus, as long as validity is defined in such a way that the conclusion of a valid argument must be at least as true as each of the premises, the relevant instances of modus ponens will be invalid whenever the truth-value of the antecedent is less than or equal to the truth-value of the conditional but (slightly) greater than the truth-value of the consequent. The paradox arises because the error is so small as to be undetectable, and yet it increases each time modus ponens is invoked.

This account has a certain prima-facie appeal but it is open to a number of objections. First, the fuzzy-theoretic machinery appears to replace vagueness with extremely refined precision. To what degree, exactly, is it true that Jones is bald? To degree 0.6? Perhaps to degree 0.59? Or maybe 0.5999? Second, the assumption of a totally ordered set of truth-values is itself problematic. How does the degree to which Jones is bald compare to the degree to which Smith is tall? How does it compare to the degree to which a certain borderline rock is part of Everest? Thirdly, there is the embarrassing presupposition that a point must still exist where one goes from fully-fledged truth to partial truth, or from partial truth to fully-fledged falsehood. What is the maximum value of n such that a person with n hairs is truly bald, i.e., bald to degree 1? What is the last rock, along a continuous path descending from the peak of Everest, that is definitely part of the mountain? These questions may not have practical relevance, but they appear to undermine the theoretical force of the account.

Turning to the second strategy for dealing with the sorites paradox – to accept the validity of the argument but to reject one of its premises as false – one can again distinguish two main approaches. One can either reject the 'base step' expressed by premise 4, or one can reject the 'inductive step' expressed by premise 3. (In the version of the paradox where premise 3 is replaced by a chain of conditionals, this amounts to rejecting one of the conditionals. We shall not discuss this variant here.) A rejection of premise 4 amounts to a radical response to the effect that a vague term such as 'bald' is ultimately incoherent (e.g., Unger, 1979). Given the pervasiveness of such terms in natural language, this line of response seems to have few advantages over Russell's version of the first strategy. A rejection of premise 3, on the other hand, amounts to asserting the existence of a precise number n of hairs separating the bald from the non-bald. This appears to contradict the intuition that 'bald' is vague and, by generalization, that there are any vague words at all. Indeed, this problem is real and ineliminable if vagueness is understood entirely in ontological terms, for then the existence of a relevant cut-off value of n amounts to the existence of a sharp boundary around the relevant set or object. However, if vagueness is understood in linguistic terms there is a popular way of resolving this intuitive problem, using an approach that has come to be known as supervaluationism (Fine, 1975).

The basic idea underlying supervaluationism is that a vague term is one that admits of various alternative 'precisifications'. A vague predicate such as 'bald', for instance, could be made precise by deciding that a man is bald if and only if he has at most 10,000 hairs. Or it could be made precise by deciding that a man is bald if and only if he has at most 9,999 hairs. And so on. The predicate is vague precisely because there is indeterminacy between these various ways of picking out a precise cut-off value. Likewise, a vague singular term such as 'Everest' could be made precise by drawing a precise boundary around its referent, but there are many ways of doing this and all of them are compatible with the way we use the name. Given this understanding of vagueness, supervaluationism says that the truth-value of a statement involving vague terms is a function of its truth-values under the various admissible precisifications of those terms. If the statement is true under all such precisifications, then it is true *simpliciter*: the unmade linguistic stipulations do not matter. In other words, it makes no difference to suppose that the meaning of those expressions could be defined

more precisely: what the statement says is true regardless ('super-true'). Likewise, if the statement comes out false under every precisification then we may regard it as false (or super-false) in spite of its vagueness. This explains, for example, why we can confidently assert sentence 4 and deny sentence 5. On the other hand, when a statement comes out true under some precisifications and false under others, the unmade linguistic stipulations become relevant. In such cases, the statement falls into a truth-value gap. This is why, for example, we must suspend judgment when it comes to statements of the form

A man with n hairs on his head is bald. (7)

for various intermediate values of n: the truth-value of such statements depends crucially on how we imagine the extension of 'bald' to be precisified.

This account preserves all theorems of classical logic even though it violates some of its fundamental semantic presuppositions, such as bivalence and truth-functionality. For example, an instance of the law of the excluded middle, such as

Either a man with n hairs on his head is bald, or he is not bald. (8)

is sure to be true even when both disjuncts fall into a truth-value gap. It is precisely this sort of property that allows supervaluationism to resolve the sorites paradox. Supervaluationally the inductive premise 3 is false, because it is false on every precisification. However, contrary to the standard semantics for the quantifiers, its falsity does not imply the existence of a specific n for which the corresponding conditional (statement 6) is false, and this is what allows a supervaluationist to preserve the intuition that 'bald' is vague. Supervaluationally it is true that there is a number of hairs that marks the boundary between bald and non-bald, but there is no number of hairs such that it is true of that number that it marks the boundary. (The same account applies, *mutatis mutandis*, to a sorites for a singular term such as 'Everest'.)

This account is attractive because it reflects a deep, preanalytical intuition concerning vagueness as it arises in ordinary language: we speak vaguely because in ordinary circumstances the vagueness of our words does not matter. Still, various objections have been raised. For example, some critics consider the supervaluationist account of the logical operators in statements 8 and 3 unacceptable. Moreover, supervaluationists have been pressed to provide an account of the phenomenon of higher-order vagueness. This difficulty manifests itself not only in the supposition that there is a clear demarcation between the clear cases and the borderline cases (as in fuzzy logic) but also in the supposition that each vague term is associated with a precise set of precisifications. Presumably, if 'bald' could be made precise by deciding that a man is bald if and only if he has at most n hairs, then it could be made precise by deciding that a man is bald if and only if he has at most $n+1$ hairs – and this yields immediately a sorites paradox for the semantic predicate 'could be made precise'. For a supervaluationist this only shows that the metalanguage within which the semantics is formulated is itself vague, but some critics find this line of response unsatisfactory. Lastly, the assumption that every vague expression can in principle be precisified, or that any number of vague expressions can in principle be simultaneously precisified, has sometimes been regarded with suspicion.

VAGUENESS AND COGNITIVE SCIENCE

To the extent that vagueness is not entirely a matter of ontology, it falls naturally within the range of interest of the cognitive sciences. Supervaluationism, for example, may be viewed as implementing a certain view about how ordinary speakers manage to communicate and reason even in the absence of a precise language. We speak vaguely because in normal circumstances the vagueness of our words does not matter. In normal circumstances what we say is true under all the admissible interpretations of our words, hence we do not need to be more precise (Lewis, 1993).

More generally, the linguistic conception of vagueness has often been associated with the idea that language is but one of many different representation systems. Thoughts and mental images, for some authors, can likewise be vague, and so can every privately or publicly accessible representation. Russell combined his conservative views on logic with the view that all vagueness is analogous to the vagueness that may exist even in a photograph, let alone an impressionist painting.

It is not clear, however, whether a single account can indeed be made to fit all these different cases (Dummett, 1975). Compare the vagueness of 'bald' with that of 'looks bald'. If Jones is a borderline case of the latter predicate, a linguistic account would have to say that on certain precisifications Jones will look bald (to me) even though his identical twin, who has just one more hair on his head, will not look bald. Since the two men look exactly alike to me, this seems to contradict the idea that the

predicate 'looks bald' is entirely observational, i.e., that it applies only by virtue of appearances. A similar point can be made for observational predicates such as 'looks square' (where 'square' is non-vague), or for any other predicate expressing properties whose reality, as some say, *is* their appearance – such as color predicates.

For another example, if there is such a thing as the language of thought, or 'Mentalese', then it would seem to suffer from a different sort of vagueness from that of public languages, at least to the extent that the meaning of Mentalese expressions does not depend on their use. A supervaluationist account would therefore seem unjustified in this case. A fuzzy-theoretic semantics would also be inadequate because of the psychologically unrealistic fineness of definition in the underlying space of truth-values. In the case of public languages one may try to base a fuzzy truth-value assignment on statistical measurements, but Mentalese would defy this way of proceeding. (There is a tradition of psychological studies aimed at measuring the degree to which people are inclined to classify a penguin as a bird, say, but this is irrelevant here: something may fail to be a typical P without being a borderline case of P, just as a perfectly clear case of prime number may fall short of typicality (Armstrong *et al*., 1983).)

Sorensen (1991) has suggested that cases such as these are more amenable to an epistemic account whereby vagueness is a kind of ignorance. On this account, the indeterminacy associated with a vague expression stems primarily from our inability to determine its exact reference (extension). More generally, the epistemic account has been proposed as an alternative to all the theories mentioned above because it provides a straightforward resolution of all sorts of sorites paradoxes. If the vagueness of 'bald' is a matter of ignorance, then a critical cut-off value of n does exist which separates the bald from the non-bald, but it is unknown to us. Moreover, the relevant value cannot be known to us. This would explain our inclination to regard statement 3 as true when it is, in fact, false. In this sense, epistemicism can be viewed as an alternative to supervaluationism in providing an implementation of the second of the two strategies for resolving the sorites paradox. Like supervaluationism, epistemicism validates all theorems of classical logic; unlike supervaluationism, however, it also validates all classical semantic presuppositions, including the principle of bivalence.

The epistemic account of vagueness is generally met with astonishment. How can there be a sharp boundary demarcating the extension of 'bald' if nobody has ever made the necessary semantic stipulations? What could possibly be the explanation of our general ignorance? One response, articulated in some detail by Williamson (1992), is that the boundaries associated with vague terms are unknowable because they violate a general principle that characterizes reliable knowledge. Briefly, this is a principle to the effect that our beliefs are reliable only if we leave a margin for error. For example, the belief that a general condition obtains in a particular case can be reliably true only if that condition obtains in every similar case (the relevant notion of similarity depending on context and cognitive capacities). In the case of baldness this would mean that we cannot *know* that a certain person is bald if people with just one more hair on their head *are not* bald. The vagueness of a predicate such as 'bald' would then be captured, intuitively, not by the inductive principle 3 but rather by a margin-of-error principle such as:

> For every n: if a man with n hairs on his head is known to be bald, then a man with $n+1$ hairs on his head is bald. (9)

And this principle does not combine with statements 4 and 5 to generate a paradox even if classical logic is retained altogether.

Some support for the epistemic conception of vagueness derives from recent experimental data. Notably, Bonini *et al*. (1999) have found that ordinary speakers react to questions about vague predicates as if they were not sure about their boundaries, which leads to the hypothesis that vague predicates are mentally represented like sharp predicates with crisp true–false boundaries of whose location one is uncertain. On the other hand, such findings seem compatible also with the view that vagueness is a phenomenon that reflects the fluid judgmental spreadings involved in human categorization. According to Raffman (1994), ordinary subjects are always likely to break the slippery slope of a sorites series precisely because a sharp category shift is likely to occur at some point on each run of judgments. The point of shift varies with the judgments of different speakers, and those judgments in turn vary with the contexts in which they are made. Rather than explaining this phenomenon in epistemic terms, however, Raffman conjectures that the point of shift is determined by a set of psychological factors, such as the strength of the judgmental inertia induced by the anchoring heuristics (Tversky and Kahneman, 1974) employed by the subjects as they proceed along the series. (One will categorize a greater number of people as bald if one begins

from the hairless side of a corresponding sorites series than if one begins from the hairy side.) In other words, the relevant category shifts are not to be viewed as boundary crossings but as Gestalt-like changes of perspective. If this is right, then it is also plausible to suppose that a subject's judgments may vary depending on whether the items in a sorites series are considered individually or in pairs. Both the basic premise and the conclusion of a sorites argument derive their plausibility from individual judgments. But only the second, pairwise, type of judgment satisfies the inductive premise of the sorites paradox. This means that statement 3 would have to be rewritten as

> For every n: if a man with n hairs on his head is bald then a man with $n+1$ hairs on his head is also bald, in so far as the two men are judged pairwise. (10)

With premise 3 replaced by statement 10, the paradox would dissolve into a fallacy of equivocation.

It is indeed regretful that the available experimental data are still too scarce to throw light on these conjectures. A psychologically plausible account can hardly fail to include some hypothesis about the mental representations that underlie our usage of vague words. Still, few theorists seem inclined to believe that the paradox can be resolved by purely empirical considerations, just as few are willing to accept a purely epistemic account. For the majority, the paradox is a genuine one. Vagueness remains a deep philosophical conundrum.

References

Armstrong SL, Gleitman LR and Gleitman H (1983) What some concepts might not be. *Cognition* **17**: 263–308.

Bonini N, Osherson D, Viale R and Williamson T (1999) On the psychology of vague predicates. *Mind and Language* **14**: 377–393.

Dummett M (1975) Wang's paradox. *Synthese* **30**: 301–324.

Evans G (1978) Can there be vague objects? *Analysis* **38**: 208.

Fine K (1975) Vagueness, truth and logic. *Synthese* **30**: 265–300.

Lewis DK (1993) Many, but almost one. In: Bacon J, Campbell K and Reinhardt L (eds) *Ontology, Causality, and Mind*, pp. 23–38. Cambridge, UK: Cambridge University Press.

Machina K (1976) Truth, belief, and vagueness. *Journal of Philosophical Logic* **5**: 47–78.

Raffman D (1994) Vagueness without paradox. *Philosophical Review* **103**: 41–74.

Russell B (1923) Vagueness. *Australasian Journal of Psychology and Philosophy* **1**: 84–92.

Sorensen RA (1991) Vagueness within the language of thought. *Philosophical Quarterly* **41**: 389–413.

Tversky A and Kahneman D (1974) Judgement under uncertainty: heuristics and biases. *Science* **185**: 1124–1131.

Tye M (1990) Vague objects. *Mind* **99**: 535–557.

Unger P (1979) There are no ordinary things. *Synthese* **41**: 117–154.

Williamson T (1992) Vagueness as ignorance. *Proceedings of the Aristotelian Society, Suppl. volume* **66**: 145–162.

Zadeh L (1965) Fuzzy sets. *Information and Control* **8**: 338–353.

Further Reading

Burns LC (1991) *Vagueness: An Investigation into Natural Languages and the Sorites Paradox*. Dordrecht, Boston and London: Kluwer.

Graff D and Williamson T (eds) (2002) *Vagueness*. Aldershot, UK: Ashgate.

Hill C (ed.) (2000) *Vagueness*. *Monographic issue of Philosophical Topics* **28**.

Horgan T (ed.) (1995) *Vagueness*. *Supplementary issue of Southern Journal of Philosophy* **33**.

Hyde D (2000) A decade of vagueness. *Philosophical Books* **41**: 1–13.

Keefe R and Smith P (eds) (1997) *Vagueness: A Reader*. Cambridge, MA and London, UK: MIT Press.

Keefe R (2000) *Theories of Vagueness*. Cambridge, UK: Cambridge University Press.

Parsons T (2000) *Indeterminate Identity: Metaphysics and Semantics*. Oxford: Clarendon Press.

Sainsbury M and Williamson T (1997) Sorites. In: Hale B and Wright C (eds) *A Companion to the Philosophy of Language*, pp. 458–484. Oxford: Blackwell.

Sorensen RA (1988) *Blindspots*. Oxford, UK: Clarendon Press.

Williamson T (1994) *Vagueness*. London, UK: Routledge.

Williamson T (ed) (1998) *Vagueness*. *Monographic issue of The Monist* **81**.

Zombies

Intermediate article

Güven Güzeldere, Duke University, Durham, North Carolina, USA

CONTENTS

A zombie is a creature that is indistinguishable in behavior as well as in certain physical or physically specifiable respects from a human being, yet which lacks certain mental features that a human being possesses.

INTRODUCTION

The folkloric notion of zombies (which came from the West Indies and has been popularized by Hollywood horror movies) attributes life, or at least lifelike behavior, to them, in a resurrected body, while denying them a soul. The philosophical notion postulates the zombie body as identical to a human body characterized at a particular level of abstraction (molecular or functional), and replaces the 'absence of soul' with the absence of mind, or particular kinds of mental states. As such, zombies are used as an illustrative element in thought experiments in the philosophy of mind that explore the nature and modality of the relation between bodily properties and mental attributes, as well as the relation between different kinds of mental states (propositional attitudes such as beliefs and thoughts versus conscious experiential states such as visual sensations and pains).

HISTORY AND BACKGROUND

Because there have been several different ways in which zombies have been described in the philosophical literature as being indistinguishable from humans, and several different features they have been described as lacking, it is necessary to characterize the idea of zombies in general terms. It is possible to come up with at least nine different philosophically relevant notions of a zombie (see Figure 1). However, in most of the notable discussions, zombies are taken to be identical to human beings either in complete physical make-up or in functionally specifiable internal constitution, and in behavior, and they are taken to lack either conscious qualitative (experiential) mental states or mentality altogether.

Although both the term "zombie" and its core notion had been used in the philosophical literature in a less specific sense earlier (e.g., Martin and Deutscher, 1966; James, 1879, and Campbell, 1970, respectively), it was introduced in its modern form by Robert Kirk (1974) as 'an organism indistinguishable from a normal human being in all anatomical, behavioral and other observable respects, yet insentient'. Although more general conceptions have since been formulated, Kirk's discussion, in its aim to refute materialism, not only led the way but also anticipated much of the discussion on zombies that followed in the next three decades. Kirk's strategy is a familiar one: using a hypothetical argument based on the conceivability of mind, characterized in terms of its cognitive or experiential attributes, and body, characterized in terms of its physical attributes, as having distinct existences. The possibility of body and mind as being thus separable is then used to draw the conclusion that any materialist ontology, however successful in providing an account of the physical world, necessarily falls short of providing a complete account of the mind.

One of the most influential examples of this kind of argument in the history of philosophy is employed in Descartes's *Meditations*. From the clear and distinct conception of mind, characterized as thinking substance, as existing independently of the body, characterized as extended substance, Descartes concludes that body and mind are in fact ontologically separate and independent. The zombie argument employs the converse of the Cartesian argument. From the self-consistent conception of the body and its full behavioral repertoire as existing in the absence of the mind, it is concluded that a complete theory of bodily attributes and behavior, by itself, is silent about the nature of mind, and the relation between body and mind. It then follows, it is argued, that

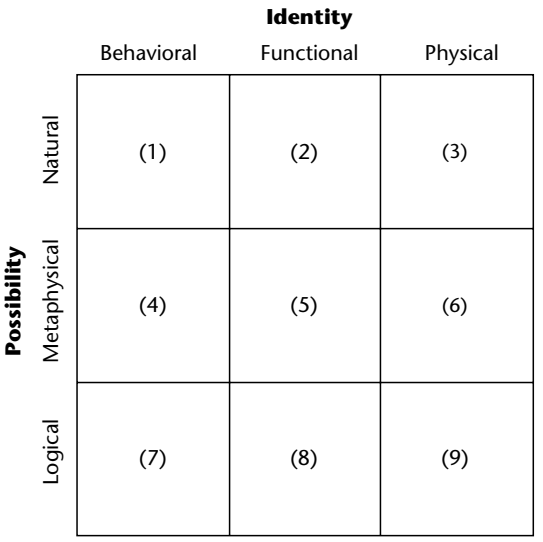


Figure 1. The 'zombie scorecard': nine distinct notions of zombie, classified according to the respects in which the postulated creature is the same as a human being (the 'identity' parameter) and the kind of possible existence the creature is granted (the 'possibility' parameter). (Adapted from Polger (2000).)

the accounts provided by materialist theories at best explain the constitution and nature of our zombie twins, not of complete human beings. Lacking the theoretical machinery to talk about the mind, or about how body and mind are related, they fail to distinguish between us and our zombie replicas.

Notice that these two versions of the conceivability argument can be, and often are, proposed independently of one another. Descartes, for example, never employed the 'zombie argument' (even though he toys with the idea in his discussions on automata and animals), and, despite the fact that he is generally cited in support of it, it is not obvious that his interactionist substance dualism could indeed permit the metaphysical possibility of a zombie. On the other hand, most anti-materialists today shy away from the Cartesian argument, because of well-known difficulties about the nature of mind–body interaction that have plagued Cartesian theory from the very beginning, and use the zombie argument only in favor of a milder version of dualism, 'property dualism'.

A few variants of Kirk's thought experiment have since been proposed as responses to various versions of materialist theories in the philosophy of mind. Most prominently, in the 1980s, the 'absent qualia' thought experiment was defended, primarily by Ned Block, contra functionalist materialism. In the 1990s, David Chalmers brought the notion of zombies to bear against materialism of all stripes, including the identity theory and materialist theories of supervenience. But in terms both of the particulars of the zombie notion employed and of the modal character of the possibility claim, there are important differences between those arguments. In order to better locate these two arguments, I will first sketch a broader framework for discussing zombies.

KINDS OF ZOMBIE

In general terms, zombies can be classified on the basis of two parameters: physical and mental properties, in virtue of which they are identical to humans in certain respects and different from them in certain others, and the modal strength of their possibility of existence. Güzeldere (1995) explored different kinds of zombies in terms of their postulated constituency, and called them 'behavioral', 'functional', and 'physiological' (or 'physical') zombies. Polger (2000) extended the discussion by examining these three kinds of zombies under natural, metaphysical, and logical possibility, producing a 3×3 'zombie scorecard' (see Figure 1).

In the first category of identity are creatures that are behaviorally indistinguishable from human beings, but may be made up of completely different, non-carbon-based stuff, with no bodily mechanism, and no functional or computational internal structure, on the basis of which there could be attributed to them a true psychology. They are candidates for 'behavioral zombies'. It may be that the behavioral zombie goes through the bodily movements that we take to be sophisticated human behavior by a miracle; those movements should not therefore be construed as anything beyond 'as-if behavior'. Nevertheless, a behavioral zombie is so sophisticated in its mimicry of human behavior that it is, by stipulation, impossible to distinguish it from a normal human being solely on the basis of what (it seems as if) it does. This 'as-if behavior' includes, of course, speech acts of the most sophisticated form, which would be sufficient for the behavioral zombie, for example, to pass the Turing Test. However, a behavioral zombie has no internal structure or mechanism that would support a functional description of its psychology, and it would also immediately fail the physical-indistinguishability test once it is internally examined beyond its skin-deep appearance.

In the second category of identity are creatures that are not just indistinguishable from human beings in behavior, but can also be attributed a belief–desire psychology at the right level of

functional characterization. Nonetheless, they may be made not of flesh and bones but of entirely different kinds of matter. This would be the characterization of a 'functional zombie', if it is also postulated that its psychology is incomplete, lacking, in particular, qualitative conscious states. A functional zombie is also a behavioral zombie, but not vice versa.

In the third category of identity are creatures that not only fulfill the criteria of behavioral and functional zombies, but also have the same bodily constituency as human beings, including flesh, blood, bone, nerve cells, and even microtubules – down to the minutest component. We may tentatively regard this kind of creature a candidate for a 'physical zombie'. Of course, a physical zombie is also a functional (and behavioral) zombie. The strongest metaphysical claim based on the possibilities of zombiehood is based on physical zombies.

To recapitulate, at one end of the scale, a behavioral zombie is a creature that is indistinguishable from human beings in its behavior but is unlike a human being in other (physiological and psychological) respects. At the other end, a physical zombie is a replica of a normal human being, identical in all its physical aspects, the psychological attributes of which are, nonetheless, being questioned. A functional zombie lies somewhere in between these two.

A brief examination of which view in the philosophy of mind favors which kind of zombie serves to reveal some of the prior ontological commitments of the various views of the mind. Physicalists, for instance, would need to claim that physical zombies lack nothing at all: that whatever is true of the psychology of humans, including the experiential states and their qualitative phenomenology, will also be true of their physical-zombie counterparts. Functionalists would further assert that functional zombies have a complete mental life, much as we do, because their psychology is functionally equivalent to that of human beings. And metaphysical behaviorists would be committed to the claim that behavioral zombies are just as conscious as any human being, since all mental states are characterizable in purely behavioral and dispositional terms.

Conversely, non-behaviorists, including functionalists and physicalists, may claim that a behavioral zombie would lack crucial aspects of the psychology of a human being; a non-functionalist physicalist may claim that a functional zombie would lack qualia-laden mental states; and property and substance dualists may claim that physical zombies, no matter how perfect molecular replicas they are, can still be 'mindless automata'.

To put the matter differently, for the behaviorist, there are (or can be, in the strongest modal sense) no zombies at all. For the functionalist, the possibility of a behavioral zombie can be admitted, but the possibility of a functional zombie (as well as that of a physical zombie) cannot. And the physicalist has to deny the possibility of a physical zombie.

In addition to the three kinds of zombies distinguished along the identity axis, one can distinguish kinds of zombies along the possibility axis of Figure 1, on the basis of the modal strength of the possibility of their existence. Among the culminating nine elements, I will focus on two particular cells of the zombie scorecard, namely (2) and (9). In recent years, on the basis of the above kind of analysis of zombie kinds vis-à-vis ontological theories, two kinds of zombies in particular, functional and physical, have been used to argue against materialist functionalism under natural possibility and physicalism under logical possibility, respectively.

ZOMBIE ARGUMENTS AGAINST MATERIALISM

Kirk's zombie argument emerged in the early 1970s at a time when similar ideas were in circulation (e.g. Campbell, 1970; Nagel, 1970; Kripke, 1972), in response to the then-dominant thesis of topic-neutral identity between physical (brain) states and mental states. Campbell's 'imitation man' is an early version of a zombie, and Kripke's contention that God would have additional work to do in order to make the mental properties instantiated after having created and set in place all the physical features of the world evokes, in effect, a complete 'zombie world'.

Materialists, by and large, took this identity relation as contingent, subject to empirical *a posteriori* discovery (Smart, 1959; Armstrong, 1968), although some argued that the identity of brain states and mental states was an analytic truth (Lewis, 1966). Kripke's attack on the identity thesis, based on his theory of rigid designators, was closely related to the zombie-based arguments of the 1970s, which were based on the logical or metaphysical possibility of zombies that were physical replicas of humans but lacked minds altogether.

The debate shifted in the 1980s, as the identity theory by and large gave way to functionalism, and zombie arguments became transformed into arguments about 'absent qualia' and 'ersatz pains'. While the notion of zombies was rarely invoked explicitly during this period, the underlying idea of absent-qualia arguments was the same. Critics of

materialism argued that minds (or mental states) could not be fully characterized in terms of their causal or functional roles, on the basis of thought experiments involving functionally equivalent constructs of brains that, intuitively, did not seem capable of underpinning mental events.

One of the best-known examples of such thought experiments is Ned Block's 'China head' argument for the impossibility of mental states emerging from a brain-like, functionally identical but spatially distributed system (Block, 1978). Block asks us to imagine the functional simulation of a human brain by the Chinese nation, by connecting each of the billion inhabitants of China in appropriate ways through radio links, and having them communicate from a distance like neurons in a brain and thereby animate an artificial body for a certain period of time. According to Block, while this system is 'nomologically possible' and 'could be functionally equivalent to [a human being] for a short time', it is doubtful 'whether it has any mental states at all – especially whether it has "qualitative states", "raw feels", or "immediate phenomenological qualities" '.

The system that Block describes is very much like a functional zombie: it can behave in ways similar to a human being, in virtue of having functionally identical but physically very different internal causal states, and it is Block's contention that it will lack mental states, at least qualitative conscious states. Natural, or nomic, possibility is at issue here. That is, the thought experiment aims to show that there can in fact be systems functionally identical to a human being (or the nervous system of a human being) and that these systems would in fact lack qualitative mental states (cell 2 of the scorecard).

A second type of zombie argument is stronger in its claim: it is based on the logical possibility of physical zombies, and it purports to show that all types of materialist theories are bound to fail (Chalmers, 1996; cell 9 of the scorecard).

Materialists, in response, reject the zombie arguments, either on the basis of differing intuitions on what is logically possible, or by resisting metaphysical conclusions drawn from mere logical possibility claims.

EPIPHENOMENALISM AND ZOMBIES

Finally, let us examine epiphenomenalism, a view implied by the possiblity of zombies. The doctrine of epiphenomenalism has a long history. The philosophers and psychologists of the nineteenth century hotly debated the question of whether consciousness was part and parcel of the causal network that was responsible for the decisions we make, actions we take, etc., or whether it was just an 'idle spectator', 'riding on' the causal processes, perhaps being caused by them, but without itself exerting any causal force on those processes. Perhaps, some argued, we are all 'automata', since all of our mental life and behavior seems to be determined by our nervous systems, in a purely mechanical framework, with no respectable place in it for consciousness.

This view does not deny that we are conscious. It comes close, however, in positing that consciousness, in itself, makes no difference. Thus it prepares the way for the concept of zombies.

Thomas Huxley was one of the most influential advocates of such a thesis, known as the 'automaton theory of consciousness'. The thesis was first formulated as applying to animals, in agreement with Cartesian intuitions. Huxley (1902) advanced the claim that 'the consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working, and to be as completely without any power of modifying that working as the steam whistle which accompanies the work of a locomotive engine is without influence upon its machinery'. But the real target was human beings and the nature of human consciousness. This was where Huxley's automata theory differed from Descartes' interactionist dualism. Huxley's account of the 'brutes' was just a lead, to make the same point for humans and maintain that 'in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism'.

For Huxley, consciousness plays no contributory role in the causal chains in the nervous systems, which totally determine the behavior of an organism; consciousness is merely affected by the neural interactions. In contrast, Descartes' idea of consciousness was of a causally efficacious parameter in the formula of mind–body interaction. Just as Descartes is taken to be the founder of interactionism, Huxley laid a clear foundation for epiphenomenalism with respect to the mind.

CONCLUSION

Could there be beings who behave like us in every possible way and yet lack consciousness (or possibly all mental life)? Could such beings be not only behaviorally, but also physically identical to us, on a molecular level, and still not have conscious life? On which modal sense of 'could' are we offering our answers – is this a nomic possibility that

can accommodated within the laws of nature in our world, or is it a metaphysical, or a logical possibility?

The answers one gives to questions of this sort is usually a good indicator of where one stands with respect to a variety of issues regarding consciousness: its ontology, nature, function, evolutionary role, and so on. One's belief in a particular modal possibility of a particular kind of zombie often helps reveal one's implicit metaphysical assumptions, rather than grounding them. As such, the notion of zombies should be considered more of a useful rhetorical tool than the basis of any knockdown argument in philosophy of mind.

References

Armstrong D (1968) *A Materialist Theory of the Mind*. London: Routledge.

Block N (1978) Troubles with Functionalism. In: Savage W (ed.) *Perception and Cognition, Minnesota Studies in the Philosophy of Science*, Vol. 9. Minneapolis: University of Minnesota Press.

Campbell K (1970) *Body and Mind*. New York: Anchor-Books.

Chalmers D (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Güzeldere G (1995) Varieties of Zombies. *Journal of Consciousness Studies* **2**(4): 326–333.

Huxley T (1902) *Methods and Results*. New York: Appleton Co.

James W (1879) Are we automata? *Mind* **4**(13): 1–22.

Kirk R (1974) Zombies v. materialists. *Proceedings of the Aristotelian Society* **48**: 135–152.

Kripke S (1972) *Naming and Necessity*. Cambridge: Harvard University Press.

Lewis D (1966) An argument for the identity thesis. *Journal of Philosophy* **63**(1): 17–25.

Martin CB and Deutscher M (1966) Remembering. *Philosophical Review* **75**: 161–196.

Nagel T (1970) Armstrong on the Mind. *Philosophical Review* **79**(3): 394–403.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* **83**: 435–450.

Polger T (2000) *Natural Minds*. Ph.D. Dissertation, Duke University.

Smart JJC (1959) Sensations and brain processes. *Philosophical Review* **68**: 141–156.

Further Reading

Dennett D (1995) The unimagined preposterousness of zombies. *Journal of Consciousness Studies* **2**(4): 322–326.

Dennett D (1999) *The Zombic Hunch: Extinction of an Intuition*? Royal Institute of Philosophy Millennial Lecture.

Flanagan O and Polger T (1995) Zombies and the function of consciousness. *Journal of Consciousness Studies* **2**(4): 313–321.

Güzeldere G, Flanagan O and Hardcastle V (1999) The nature and function of consciousness: lessons from blindsight. In: *The New Cognitive Neurosciences*. Gazzaniga M (ed.). Cambridge, MA: MIT Press.

Kirk R (1999) Why there couldn't be zombies. *Proceedings of the Aristotelian Society*, Supplementary Volume **73**: 1–16.

Levine J (1998) Conceivability and the metaphysics of Mind. *Nous* **32**: 449–480.

Perry J (2001) *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.

Stalnaker R (2002) What is it like to be a zombie? In: Gendler T and Hawthorne J (eds) *Imagination, Conceivability, and Possibility*. Oxford, UK: Oxford University Press.

Academic Achievement

Introductory article

Harold W Stevenson, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

Worldwide studies of academic achievement in children and adolescents strengthen our understanding of the factors underlying differences in performance.

CONDUCTING CROSS-CULTURAL RESEARCH

This article describes the methodology and findings of several studies of academic achievement of children and adolescents in the US, Taiwan, China, and Japan. In the mid-1980s, when many of these studies were begun, large-scale studies of academic achievement in the US and East Asian countries were still rare, in part because China had not until then opened its doors to foreign researchers. Since then, there has been increasing interest in cross-cultural studies of academic achievement comparing East Asian children with their US peers.

The main purpose of most of these comparative studies is to identify factors that underlie cross-cultural differences in children's performance in academic subjects such as mathematics and science. Such research can generate ideas about what is possible in one's own culture, by drawing attention to those environmental conditions and strategies that have been found to be productive in others. East Asian societies have attracted the interest of many educators and researchers because of the excellence of their students' academic performance. International test results show that the academic performance of American children is consistently below that of their East Asian counterparts.

We begin with a consideration of several important methodological and interpretive issues, and then review some findings that illustrate the importance and relevance of cross-cultural research to educational policy and practice in the United States.

Methodological Issues in Developing Tests

Cross-cultural research is difficult to conduct, for it requires instruments that are culturally specific but are also capable of producing findings that can be applied meaningfully across cultures. Perhaps the greatest problem encountered in cross-cultural research in the social and behavioral sciences has been the use of materials that were initially developed for one culture and then applied inappropriately to other cultures. For example, a scale devised in New York for measuring depression would not necessarily be equally valid in East Asian cultures. Great care must be taken, therefore, in ensuring that research materials and procedures are culturally relevant, reliable, and valid for all cultures being studied.

Another problem often encountered in cross-cultural research is that sample sizes are too small to permit generalization to the larger population being studied. One must ensure that samples are representative of the larger population, and that study materials are prepared in the language of the culture under study. For example, an intelligence test such as the Stanfordu–Binet test developed in the US is not equally useful in China when translated into Chinese. Problems such as these make the research difficult to replicate and to apply.

Measuring Personality and Social Attributes

It is often difficult to translate psychological terms such as 'dependency', 'aggression', or 'anxiety', because of differences in the nuances or meanings of these concepts across different languages. Some researchers measure personality and social attributes through survey research, where large numbers of participants fill out scales designed to record subtle aspects of their reactions to, for example, mathematics instruction. Alternatively, researchers may conduct personal interviews with subjects, making it possible to probe and clarify subjects' responses through questions such as: 'How satisfied were you when you received your last semester's grade in math? Can you tell me

more about why you felt this way?' Another technique for investigating differences in the meanings of terms for personality and social attributes across cultures involves the development of vignettes, prepared by residents of each country, to provide a detailed look at differences in the meanings of the terms across languages.

Studying Classroom Behavior

Studies of the classroom behavior of students and teachers require many hours of observation by researchers and thorough checks of the validity and reliability of their interpretations and conclusions. The advent of videotaping has significantly streamlined the process of observing children and teachers in action, as it has enabled researchers to gather substantial amounts of data without incurring the cost of observation on site. However, it is essential that videographers understand the purposes of the study, and the languages and cultures being studied, to ensure that the tapes are relevant to the research.

CULTURAL DIFFERENCES IN SCHOOLING

Here we briefly describe some results from a series of comparative studies concerned with differences in performance in mathematics and science of students in the US, Taiwan, and Japan.

Main Results

US students generally perform at lower levels at each grade level – from first grade through entrance into college – than do Chinese and Japanese students on a range of tests of cognitive development and academic achievement. Children's academic performance across countries bears little relation to the amount of money the countries spend per child on education. For example, the US spends a greater proportion of its gross national product on education than does China, yet consistently produces lower scores on tests and other indicators of children's cognitive development and academic achievement. Interestingly, the Chinese students who performed better than the American students were less satisfied with their performance than were the American students with theirs. This finding is consistent with the generally higher standards that Chinese and Japanese schools, parents, and students themselves, tend to set for students.

Thinking about Thinking

One explanation of the observed superiority among Chinese and Japanese students on academic tests concerns the manner in which problem-solving is presented. The East Asian teacher guides children through a problem presented in a lesson plan by encouraging them to define the problem in their own words, to determine an acceptable solution to the problem, and then to create additional methods for solving the problem. As the teacher walks around the classroom to observe students' efforts, he or she may offer hints to a confused child. After one acceptable solution is found, the child is asked if another one can be found. Depending on the child's reaction to this challenge, additional solutions may be requested of the student. The child quickly learns that problems can be approached in more than one manner, and that the way to success in solving problems is through close attention, careful thought, creative thinking, and discovery of alternative strategies.

In the US, students are typically expected to learn a prescribed strategy for solving problems and then to apply the strategy in new contexts. Less emphasis is placed on having children define the problems and then develop multiple approaches to their solution.

East Asian students commonly characterize mathematics as something they need to understand, not as something to memorize. The emphasis on problem-solving, rather than rote learning and memorization, is evident in the types of problems included in contemporary curricula in mathematics. Teachers are encouraged to present the questions in a meaningful context so that students become able to understand a rule, to provide a number of problems that are solvable by the same general rule, and to extend the discussion, opening it up for further comments.

Innate Ability

Children's own perceptions of the role that several factors play may also influence their level of academic achievement. These factors include the strength of the teacher, the level of effort students expend, their level of ability, their aspirations for educational achievement, their reasons for working hard, and their dissatisfaction with current performance. A common finding among East Asian participants is the belief that a child's behavior is highly malleable and can be readily changed, depending on the quality, interest, and relevance of the child's experiences. East Asian participants

stressed the importance of the level of effort students apply to their studies, as an important predictor of their level of academic achievement – rather than their natural or innate ability, which US respondents favored as a stronger predictor. Thus, according to East Asian participants, students' test scores may rise or fall according to the quality of their experiences and the effort they apply to their studies.

Development and Expansion of Academies

What happens in the classroom constitutes only a part of the academic training that occurs in schools. A second group of educational institutions provides supplemental training to many students. These are called 'buxiban' in Taiwan, and 'juku' and 'yobiko' in Japan. These academies were set up to ensure that children would be taught aspects of the curriculum that could not be covered during the ordinary school hours. They may be especially relevant now that the school week in East Asia has been reduced from five and a half days to five days.

East Asian educators have developed a variety of extracurricular programs to accommodate differences in students' levels of academic, social, and personality development. Yobiko in Japan, for example, are academies designed specifically to help students who fail the college entrance examination to pass it in subsequent efforts. Buxiban in Taiwan, and yobiko in Japan, provide students with additional practice and review of the content of academic courses, and help them to prepare for college entrance examinations. Buxiban and juku may include activities for students with special abilities who wish to undertake projects of unusual complexity: for example, making their own communications equipment, developing their own pharmaceuticals, or presenting programs of folk dances from various regions of China. Children and young people who wish to become more involved in sports or the arts, or who are interested in filling a void of social experiences, are also often served by these institutions. They provide opportunities for athletics and art, and a meeting place for fostering social interactions among schoolmates in after-school activities.

In the past, admission to colleges and universities in Japan depended solely on students' scores on college entrance tests. The emphasis on test scores as the main basis for college admission guided many aspects of higher education in Japan, including the *Course of Study*, a guidebook that describes the content of the curriculum and what students should know in preparing for the examinations. While this practice may have served the purpose of selecting highly able college students in the past, many Japanese have responded to it unfavorably, claiming that its narrow focus on only one aspect of children's academic achievement is 'elitist'. Following the general practice in the United States, admission procedures have changed to permit grades in high school and letters of recommendation, as well as test results, to be considered as important factors in admission decisions. This 'recommendation' method has resulted in the introduction of personal and social aspects of achievement, such as extracurricular activities and other distinguishing characteristics of children, into the criteria for selecting high school and college students. A negative feature of these changes is that students have less time than was available earlier for traditional academic work. There is increasing interest in transforming the system of education so that lessons place less emphasis on memorization and more on thoughtful, creative problem-solving.

CONCLUSION

It has become increasingly clear that understanding a single culture is not sufficient to create new methods of educating students. Cross-cultural research on cognitive development, conducted with sound methodology, can give researchers and educators substantial insight into methods for improving educational practices and students' academic performance. Recent technological advances, such as fast computers and videography, as well as recent political developments, such as China opening its doors to foreign researchers, have enabled much data to be obtained and analyzed across a wide range of cultures. The findings from cross-cultural research on children's academic achievement have far-reaching implications for US education policy and practice. These findings identify mechanisms to help strengthen children's understanding of academic information and may help to narrow the gap between western and eastern children's performance in mathematics and science.

Further Reading

Cizek GJ (ed.) (1999) *Handbook of Educational Psychology*. San Diego, CA: Academic Press.

Johanek M (ed.) (2001) *A Faithful Mirror: Reflections on the College Board and Education in America*. New York, NY: College Board.

Munro DJ (1996) *The Imperial Style of Inquiry in Twentieth Century China*. Ann Arbor, MI: Center for Chinese Studies.

Paris S and Wellman H (eds) (1998) *Global Prospects for Education: Development, Culture, and Schooling*. Washington, DC: American Psychological Association.

Ravitch D (ed.) (1998) *Brookings Papers on Education Policy*. Washington, DC: Brookings Institution.

Rohlen T and LeTendre GL (eds) (1996) *Teaching and Learning in Japan*. New York, NY: Cambridge University Press.

Sing L (ed.) (1996) *Growing Up the Chinese Way: Chinese Child and Adolescent Development*. Hong Kong: Chinese University Press.

Stevenson HW (2000) *To Sum It Up: The TIMSS Case Studies of Education in Germany, Japan, and the United States*. Philadelphia, PA: Research for Better Schools.

Stevenson HW and Stigler JW (eds) (1992) *The Learning Gap*. New York, NY: Summit Books.

Stigler JW and Hiebert J (1999) *The Teaching Gap*. New York, NY: Free Press.

Action

Intermediate article

Jos J Adam, Maastricht University, Maastricht, Netherlands
Martinus J Buekers, Katholieke Universiteit Leuven, Leuven, Belgium

CONTENTS

Action is the ability to move the body or body parts in a purposeful, coordinated manner in order to physically interact with the environment. It is based on the integration and cooperation of sensory and motor systems.

INTRODUCTION

A characteristic of humans is their ability to perform skilled motor actions. Some of these motor actions are carried out to perform mundane tasks such as walking, picking up a pencil, drinking a glass of water, or buttoning a shirt. Other motor actions accomplish more sophisticated tasks such as painting a picture, flying an airplane, and performing brain surgery. How does the brain generate and control motor actions? This is not a trivial question, as the body contains hundreds of muscle groups that act on scores of joints, thereby introducing a very large number of 'degrees of freedom' in movement. This article discusses some of the variables that shape human motor performance, and describes its main underlying control principles. (*See* **Motor Control and Learning**; **Performance**; **Reaction Time**)

STARTING, STOPPING, AND SEQUENCING ACTIONS

Starting Actions

To start an action, preparatory processes are first needed in order to formulate a movement plan. These preparatory processes include the processing of sensory information in order to perceive the state of the environment and the state of the actor. Based on this information, an action plan is formulated. For instance, when driving a car, one needs to visually scan the environment in order to detect significant environmental changes, such as a traffic light changing to amber. When reaching for a pencil, one first needs to determine its position in space. This kind of perceptual information is then related to the state of the actor (the effector), and a decision is made about what to do and how to do it. These preparatory processes take time, which can be measured by means of the reaction-time paradigm.

In a typical reaction-time task, the actor is first presented with a warning signal in order to increase alertness. Then, after a variable time interval (which can range between 0.2 s and 5 s), the imperative reaction signal (the stimulus) is presented, calling for a physical response. In the laboratory, the response is often a button-press or button-lift movement. The time that elapses between the presentation of the reaction signal and the start of the overt response defines the reaction time. Many variables influence reaction time. Some of the most influential are: the number of response alternatives; the complexity of the response; the spatial compatibility between the possible stimuli and responses; and the performer's subjective preference for speed or accuracy of responding.

Number of stimulus–response alternatives

In some situations, there is only one possible stimulus and only one correct response. The reaction time in this situation is called the 'simple' reaction time. For example, an athlete just before the start of a 100-meter race knows in advance what to expect and what to do, and the reaction time is typically short (less than 200 ms), representing the minimal time needed to perceive and act.

Other situations confront the performer with more uncertainty. For example, a boxer facing an opponent who can attack with the left or right fist must make a fast decision about what to do. In situations like this, the 'choice' reaction time is

substantially longer, mainly reflecting the increased processing demands associated with selecting and programming the appropriate action. In general, as the number of stimulus–response alternatives increases, the choice reaction time increases. The exact relationship between reaction time and the number of stimulus–response alternatives is given by the Hick–Hyman Law, which states that reaction time increases linearly with the logarithm of the number of stimulus–response alternatives.

Response complexity

The nature of the response to be executed is an important determinant of reaction time. A complex action needs more programming time than a simple action, and therefore a longer reaction time. Increased response complexity can be defined in terms of additional movement parts, increased accuracy demands, and longer movement durations (e.g. Klapp, 1996).

Stimulus–response compatibility

The (spatial) relationship between the set of potential stimuli and the set of potential responses is one of the most important determinants of reaction time. For example, choice reaction time in a task involving left and right stimuli and left and right responses is slower when the stimulus and associated response are on opposite sides (i.e. a left stimulus requires a right response and a right stimulus requires a left response) than when they are on the same side (i.e. a left stimulus requires a left response and a right stimulus a right response).

In general, performance tends to be faster when there is a close, natural correspondence or association between the stimulus and its required response. According to recent theories of stimulus–response compatibility, when there is some kind of correspondence or similarity between the stimulus and response sets, each stimulus will automatically activate its natural – or habitually associated – response (e.g. Kornblum *et al.*, 1990). If this response is the correct one, it can be executed immediately and the reaction time is short: this situation represents a compatible or congruent stimulus–response assignment. If, however, the automatically activated response is incorrect – as is the case in a 'crossed' or incompatible stimulus–response assignment – then the automatically activated response must be aborted and inhibited, and consequently reaction time is lengthened.

The notion of automatic response activation has been confirmed by evidence from psychophysiological studies that recorded lateralized event-related brain potentials on the basis of electroencephalogram recordings (e.g. Eimer, 1995). These studies indicate that action control processes may proceed very fast and without conscious awareness.

Speed–accuracy bias

Human task performance can be error-prone. In fact, there is a trade-off between speed of responding and accuracy of responding: improvements in reaction time often co-occur with an increase in the number of errors, and improvements in accuracy with slower responses. Task constraints and subjective biases for speed or accuracy jointly determine the position on the speed–accuracy continuum.

Stopping Actions

Once the motor program has been formulated, it is implemented and executed. 'Movement time' is the time that elapses between the start of the movement and its termination. A movement stop may be incorporated in the motor program before execution. This is the case for extremely fast, ballistic movements that are executed to maximize speed (e.g. a boxer's punch). However, a movement stop may also be planned 'online', that is, during the execution of the movement. This is the case when accuracy is an important task requirement. Under such circumstances, feedback-based modifications of the movement are implemented during the latter parts of the movement – that is, when the movement approaches the target zone – in order to ensure that the movement terminates accurately.

Thus, for ballistic actions movement time is very short and mainly a function of the parameters specified by the motor program; while for accurate actions movement time is typically longer and a function of motor program specifications and feedback-based movement modifications. Note that this distinction describes two modes of motor control: open-loop and closed-loop control. Open-loop control is fast, but inflexible, because it just follows the instructions laid down in the motor program without the possibility of modification. Closed-loop control is more flexible, but slower, because of feedback processing and the implementation of movement corrections.

Research using the manual-aiming paradigm (whereby the hand is moved from a 'start' position to a 'target' position) has identified several important variables that influence movement time: movement distance; target size; the presence

of non-targets (distractors); and the actor's subjective preference for speed or accuracy of moving.

Movement distance and target size

Increasing movement distance and decreasing target size both result in longer movement times. The exact relationship is expressed by Fitts' law. Fitts' law states that movement time increases linearly with the 'index of difficulty', defined as $\log_2(2A/W)$ where A is the movement amplitude (or distance) and W is the width or size of the target (Fitts, 1954). According to the 'optimized submovement' model (Meyer *et al*., 1988), Fitts' law is a consequence of a particular hybrid way of controlling movements: a preprogrammed, initial-impulse phase followed by a feedback-based control phase that allows the implementation of corrective submovements. This model allows actors either to make submovements in order to optimize spatial accuracy or to keep submovements to a minimum in order to maximize speed.

Presence of non-targets

Often, the target for an action does not appear in isolation but is surrounded by other objects (sometimes called non-targets or distractors). An example is the act of picking a particular apple from a basket full of fruit. Research has shown that the presence of additional objects slows down reaching performance (Tipper *et al*., 1992). Moreover, there is an important asymmetry in the spatial nature of this interference effect: distractors close to the starting position of the hand interfere more than distractors further away (the 'proximity-to-hand' effect). According to the visuomotor account of distractor interference (Meegan and Tipper, 1999), both target and distractor automatically trigger the planning of movements towards their locations. Distractor interference, then, reflects the need to suppress or inhibit responses towards the distractor.

Speed–accuracy bias

It is not only external, physical constraints (e.g. size and distance of the target) that influence the control of movements; the intention or goal of the performer is important too. Individual performers may opt to emphasize speed or accuracy. It has been shown that the kinematics of the movements produced under these two strategies are different (Adam, 1992). Whereas a speed strategy results in a more or less symmetric movement profile (i.e. similar durations for the acceleration and deceleration phases), an accuracy strategy results in a much longer deceleration than acceleration phase. The longer deceleration phase associated with an accuracy strategy allows feedback information concerning endpoint accuracy to be monitored and used for movement adjustments.

Sequencing Actions

So far, we have considered discrete or one-element movements which have a single start and a single stop. Often, however, actions involve a series of consecutive movements that occur through time: for example, writing, speaking, piano playing, dialing a phone number, and entering a security code in a bank terminal.

The 'one-target advantage' phenomenon sheds some light on how the brain controls sequences of movements. A rapid aimed hand movement is executed faster when it is performed as a single, isolated movement than when it is followed by an additional movement. The one-target advantage phenomenon may be demonstrated by comparing performance in two conditions: participants are asked either to move as quickly as possible to the first target and stop (the one-tap condition) or to strike the first target and then immediately move on and hit a second target (the two-tap condition). Typically, movement time to the first target is about 20 ms shorter in the one-tap condition than in the two-tap condition, indicating that one of the effects of making a second movement is to slow down the first. This observation implies that the two movements are functionally interdependent.

According to the 'movement integration' account, the one-target advantage results from an anticipatory motor control strategy whereby the two movements in the two-tap condition are planned together in one overall response program before response initiation (Adam *et al*., 2000). This overall response program specifies that the second movement should be implemented during the latter parts of the first movement so that a smooth and quick transition of the first movement into the second can occur. In other words, implementation of the second movement does not await termination of the first movement, but, rather, overlaps with and is superimposed on the execution of that first movement. The movement integration account of the one-target advantage is based on the notion that the serial ordering of movements is controlled by a motor program that represents an integrated series of movements.

Sequences of movements involving more than two elements may be represented hierarchically. This means that the mental representation of a sequence is not just a linear string of event-to-event

associations; rather, elements are organized in groups, and superordinate relationships may exist among them, so that there are distinct (i.e., higher and lower) levels of control mediating the temporal structure of the movement sequence. Consistent with a hierarchical conception of motor control, interresponse times of individual elements in a movement sequence are not identical, but follow closely the hierarchical (or tree-like) structure of a sequence of keyboard responses (Rosenbaum *et al*., 1983).

ACTION AND PRACTICE

As the proverb 'practice makes perfect' implies, high-level performance appears to be a function of repeated rehearsal of motor acts. Indeed, the elegance, speed, and accuracy of actions performed by experts (e.g. athletes, musicians, craftspeople) derive from many years of deliberate practice. The observation that repetition exerts a powerful influence on the quality of the motor act (e.g. Ericsson *et al*., 1993; Helsen *et al*., 1998) has inspired many researchers to concentrate on the mechanisms underlying skill acquisition. (*See* **Expertise**)

The Three-Stage Theory of Practice

Regarding actions as the overt consequences of the interactions of the human neuromuscular system with the environment, one can regard the learning of new motor skills as adaptations of the implemented input–output connections. Apparently the cerebellum plays an important role in this adaptation process, as the repeated soliciting of its circuits establishes, modifies, or strengthens the required input–output connections. The learning of motor skills has a strong biological foundation (this observation is supported by recent research using magnetic resonance imaging). Within these structural boundaries, the learning process can materialize in different ways. Regarding the mental representations of the action to be executed as the controlling and driving forces of movements, the observed practice effects appear to express the enhancement of these 'dynamical' representations. Actually, the adaptability of these representations (which illustrates well the plasticity of the brain functions) is a prerequisite for learning.

What form does this learning process take? The progress made by humans trying to learn new skills does not appear to follow a linear function. Ever since Fitts (1964) formulated his observation that learning follows a triphasic path, it has been clear that learning phases are fundamental features of long-term behavioral adaptations. Indeed, distinct learning phases have been described, each embodying specific characteristics and properties of motor behavior. The amount of mental effort put into the action by the performer is an important determinant of these phases. A strong cognitive involvement is observed in the initial ('cognitive') phase of the learning process, but the need for this cognitive activity gradually decreases as learning progresses. Numerous repetitions are required to pass from the second ('associative') phase, in which performers continually adjust and improve the movement pattern, to the last ('autonomous') phase in which real expertise appears. This final phase, with its smooth, elegant and parsimonious motor behavior, is the automated expression of years of deliberate practice, which has enabled the actor to eliminate cognitive interference and to use higher-order mental processes only to develop and implement strategic elements. A highly skilled performer transforms action into art.

The Importance of Feedback

Various mechanisms are available to facilitate and possibly shorten the route to expertise. Apart from verbal instructions and visual models (which are most effective during the cognitive phase), the most widely used strategy is to inform the actor about his or her performance after completing the action. This feedback can engender positive learning effects; however, neutral or even negative effects have also been reported, showing that the effectiveness of feedback strongly depends on the nature of the task. For example, when a player tries to score a goal, information about the result of his or her kick or throw is redundant as this information is intrinsically available in the task. Moreover, false or conflicting feedback information may engender incorrect behavioral adaptations (Buekers *et al*., 1992). Nevertheless, the crucial element in learning appears to be the availability of (correct) information about the errors produced by the learner. On the basis of this error information the learner can adjust his or her movement in a subsequent trial and better meet the criteria of the task.

How should this feedback information be transmitted to the learner? Knowledge of results is an obvious form of feedback, but it is actually less efficient than one might expect. Knowledge of performance (information about the movement itself) appears to have a more pronounced facilitating effect on the learning of new motor skills. This form of feedback is primarily focused on movement characteristics (form, spatial, and temporal

characteristics, etc.), and aims to change the movement errors that led to performance failure. The term 'transition feedback' has been proposed, referring to how specific elements of the movement must be changed to produce a successful outcome.

MOTOR IMAGERY

Many experiments have shown that performance can improve as a result of purely mental activity, in the absence of the actual action – for example, imagining oneself performing a perfect golf swing, lifting a heavy weight, or throwing a dart at the bullseye. This improvement has been observed for a variety of motor skills, such as tracking in a pursuit task, walking on a balance beam, hitting golf balls to a target, and muscular strength. However, although these effects are substantial, motor imagery cannot replace physical practice.

The observed performance benefits of motor imagery have sometimes been ascribed to motivational factors; however, it has been shown that motor imagery and overt practice stimulate – at least in part – common neural substrates. Apparently, motor imagery and overt practice rely to a large extent on the same cortical structures. For example, qualitatively similar event-related brain potentials have been reported when subjects imagined or executed specific hand movements. More recent studies using positron emission tomography or magnetic resonance imaging confirm this finding and support the hypothesis of functional equivalence, that movement preparation and motor imagery use the same brain processes. Thus, there seem to be relevant mental representations that are accessible to the actor even in the absence of external stimulation and overt action.

ACTION AND CONSCIOUSNESS

Consciousness is very difficult to define, but phenomenologically it refers to experience (Flanagan, 1998). Consciousness often, but not always, plays a role in the control of action. When learning a new motor skill, a substantial amount of conscious attention is typically needed to learn the basic procedures involved. In this early cognitive stage, verbal cues are often used. However, once actions are highly practiced they may no longer require conscious control; indeed, when automated, actions like writing or driving a car can be performed without conscious awareness at all. Thus, skilled actions that have progressed to Fitts' autonomous stage can be performed without consciousness; they are automatic in the sense that they can be carried out concurrent with other mental or motor activities.

The phenomenon of 'blindsight' also illustrates that consciousness is not always necessary for the control of action. Patients with cortical blindness in one hemifield can perform eye movements or pointing gestures towards visual information presented in their blind field even though they report that they are unaware of this information. Furthermore, the Ebbinghaus (or Titchener circles) illusion (a target circle surrounded by smaller circles appears to be larger than a target circle surrounded by larger circles) affects the conscious visual perception of the size of the target but not grasping movements towards it: the grip aperture of the grasping hand is determined by the actual and not by the subjectively perceived size of the target (Agliotti *et al*., 1995). Observations like these indicate that actions are not always controlled by the conscious experience of objects. Indeed, Milner and Goodale (1995) argue that (conscious) visual perception and visuomotor control are separate functions, sensitive to different constraints, and mediated by different neural pathways. According to this view, visual information may affect action directly, without mediation by consciousness. (*See* **Blindsight**; **Visual Scene Perception**)

CONCLUSION

Motor actions are the principal means by which we interact with the world. Brain processes compute an abstract representation of the intended action. The time to prepare a motor representation or motor program is reflected in the reaction time, which is sensitive to factors such as the number of response choices, the complexity of the response, the compatibility between the potential stimuli and responses, and the performer's bias towards speed or accuracy. The time needed to complete a movement depends on variables such as movement distance, target size, the presence and spatial nature of non-targets, and, again, the actor's bias for speed or accuracy.

Movements can be fully preprogrammed and executed as such (open-loop control); the advantage is speed, though sometimes at the cost of accuracy. Accurate (and slower) movements are constrained by feedback processes that allow corrective submovements (closed-loop control).

Motor skill acquisition progresses via three phases: in the first, cognitive phase, conscious attention is required to understand and practice the basic procedures and movements involved; in the second, associative phase, feedback in the form of knowledge of performance is particularly effective

in shaping successful movement patterns and eliminating errors; in the last, autonomous phase, motor performance is automated and freed from the involvement of conscious control, allowing the concurrent engagement in other tasks. Mental practice – especially in combination with physical practice – can improve motor performance. But even though consciousness is necessary for mental practice, and is very involved in the first two phases of skill acquisition, skilled motor action can, and often does, proceed unconsciously.

References

Adam JJ (1992) The effects of objectives and constraints on motor control strategy in reciprocal aiming movements. *Journal of Motor Behavior* **24**: 173–185.

Adam JJ, Nieuwenstein J, Huys R *et al*. (2000) Control of rapid aimed hand movements: the one-target advantage. *Journal of Experimental Psychology: Human Perception and Performance* **26**: 295–312.

Agliotti S, DeSouza JFX and Goodale MA (1995) Size-contrast illusions deceive the eye but not the hand. *Current Biology* **5**: 679–685.

Buekers MJ, Magill RA and Hall KG (1992) The effect of erroneous knowledge of results on skill acquisition when augmented information is redundant. *Quarterly Journal of Experimental Psychology* **44A**: 105–117.

Eimer M (1995) Stimulus–response compatibility and automatic response activation: evidence from psychophysiological studies. *Journal of Experimental Psychology: Human Perception and Performance* **21**: 837–854.

Ericsson KA, Krampe RT and Teschromer C (1993) The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406.

Fitts PM (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* **47**: 381–391.

Fitts PM (1964) Perceptual–motor skill learning. In: Melton AW (ed.) *Categories of Human Learning*, pp. 243–285. New York, NY: Academic Press.

Flanagan O (1998) Consciousness. In: Bechtel W and Graham G (eds) *A Companion to Cognitive Science*, pp. 176–185. Oxford, UK: Blackwell.

Helsen WF, Starkes JL and Hodges NJ (1998) Team sports and the theory of deliberate practice. *Journal of Sport and Exercise Psychology* **20**: 12–34.

Klapp ST (1996) Reaction time analysis of central motor control. In: Zelaznik HN (ed.) *Advances in Motor Learning and Control*, pp. 13–35. Champaign, IL: Human Kinetics.

Kornblum ST, Hasbroucq T and Osman A (1990) Dimensional overlap: cognitive basis for stimulus–response compatibility – a model and taxonomy. *Psychological Review* **97**: 253–270.

Meegan DV and Tipper SP (1999) Visual search and target-directed action. *Journal of Experimental Psychology: Human Perception and Performance* **25**: 1347–1362.

Meyer DE, Abrams RA, Kornblum S, Wright CE and Smith JEK (1988) Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological Review* **95**: 340–370.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford, UK: Oxford University Press.

Rosenbaum DA, Kenny S and Derr MA (1983) Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance* **9**: 86–102.

Tipper SR, Lortie C and Baylis GC (1992) Selective reaching: evidence for action-centered attention. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 891–905.

Further Reading

Gallistel CR (1999) Coordinate transformations in the genesis of directed action. In: Bly BM and Rumelhart DE (eds) *Cognitive Science*, pp. 1–42. San Diego, CA: Academic Press.

Gazzaniga MS, Ivry RB and Mangun GR (1998) Motor control. In: Gazzaniga MS, Ivry RB and Mangun GR (eds) *Cognitive Neuroscience*, pp. 371–422. New York, NY: Norton.

Jeannerod M (1997) *The Cognitive Neuroscience of Action*. Cambridge, MA: Blackwell.

Keele S (1986) Motor control. In: Boff JK, Kaufman L and Thomas JP (eds) *Handbook of Human Perception and Performance*, vol. II, pp. 1–60. New York, NY: John Wiley.

Rosenbaum DA (1991) *Human Motor Control*. San Diego, CA: Academic Press.

Schmidt RA and Lee TD (1999) *Motor Control and Learning*. Champaign, IL: Human Kinetics.

Shumway-Cook A and Woollacott MH (1995) *Motor Control: Theory and Practical Applications*. Baltimore, MD: Williams & Wilkins.

Addiction

Introductory article

George V Rebec, Indiana University, Bloomington, Indiana, USA

CONTENTS

The essence of addiction is persistent performance of a behavior despite increasingly aversive consequences. This definition is most commonly used to describe the compulsive taking of drugs. An addict endures deteriorating health, social isolation, and other setbacks in maintaining a drug habit. In effect, addiction is the loss of control over drug use.

ADDICTIVE PROPERTIES OF DRUGS

Many substances, whether they occur in nature or are synthesized in laboratories, are potentially addictive. Intravenously administered heroin and the smokable forms of cocaine ('crack') and methamphetamine ('ice') stand out as drugs that pose a high risk of addiction. But the risk also extends to routinely available substances, such as nicotine and alcohol.

What makes drugs addictive? Part of the answer involves rapid entry into the brain. For a drug like heroin, rapid brain entry often elicits a brief, but intense, rush or 'high' that can drive the search for more of the drug. In fact, heroin has a higher potency than morphine even though heroin is converted to morphine once it enters the brain. The difference is that heroin has additional methyl groups that allow it to slip across membrane barriers more readily than morphine. The invention of the hypodermic syringe, which can deposit drugs directly into the bloodstream for ready access to the brain, became a key factor in the spread of heroin addiction at the end of the nineteenth century. Drugs that are inhaled or smoked also gain rapid access to the brain. Alcohol, because of its relatively simple molecular structure, is absorbed across mucus membranes and begins entering the blood and brain even before it reaches the stomach.

But rapid brain entry of a drug is not the only factor contributing to addiction, nor is the drug itself. Not all heroin and cocaine users become addicts. Many tobacco smokers and social drinkers also avoid compulsive use. Addiction develops only after regular, repeated exposure to a drug on a chronic basis. The factors that control such behavior are many, ranging from physiology to family history, and they interact in complex ways that thwart the search for a simple explanation.

There are certain features of addictive drugs, however, that make them more likely to be abused than other substances. One of these features is the ability to elicit pleasure or reward.

The Reward Model

That drug reward might play a role in addiction gained widespread acceptance after the demonstration that animals, equipped with appropriate catheters, would work to obtain drugs given intravenously. Rats, for example, will readily and repeatedly press a lever to self-administer many of the same drugs as humans. This was an important demonstration because it showed that there was nothing uniquely human about compulsive drug use. In fact, the animal experiments showed that drugs shared many similarities with naturally occurring rewards such as food, water, or sex. A hungry rat, for example, will work very hard for food just as it will go to great lengths for an intravenous injection of heroin or cocaine. Thus, drugs came to be viewed as positive reinforcers, subject to the same principles that govern the behavioral response to other positive reinforcers like food.

It should be noted that positive reinforcement is not necessarily the same as pleasure or reward. Reinforcement simply refers to a procedure that strengthens or increases the likelihood of a given behavioral response. It is impossible to know what the rat actually experiences. Although discussions of drug addiction often use 'reinforcement' and 'reward' interchangeably, these terms are not always equivalent.

When animal research revealed the powerful reinforcing properties of addictive drugs, the next step was to investigate the brain mechanisms underlying this behavior. Research revealed that

the brain was equipped with a circuit that appeared to mediate the behavioral response to natural reinforcers as well as drugs. Although the investigation continues, now bolstered by powerful brain imaging techniques applied to human subjects, the reinforcing effects of drugs have been linked to specific neuronal and biochemical changes.

The Reward Circuit

The brain circuit that appeared to signal reinforcement, and that now has become known as the reward circuit, was identified in the 1950s when it was shown that rats would work to stimulate electrodes implanted in the medial forebrain bundle. Although this bundle contained fibers connecting a wide array of structures, attention centered on a group of axons extending from the ventral tegmental area in the midbrain to several forebrain structures in the limbic system and cerebral cortex. The axons were found to release dopamine as a transmitter, and it was dopamine that seemed to play a key role in virtually all forms of motivated behavior. In fact, the mesocorticolimbic dopamine pathway, as it came to be known, also appeared to be highly sensitive to drugs of abuse. Opiates, like heroin and morphine, and psychomotor stimulants, like cocaine and the amphetamines, all increased dopamine transmission. The same was found for nicotine and alcohol. Not only did all these drugs appear to elevate the synaptic level of dopamine, they were most likely to have this effect in the nucleus accumbens, an important target of the mesocorticolimbic pathway known to process autonomic, emotional, and cognitive signals. When blockade of the accumbal dopamine system was found to block drug self-administration behavior in rats, the link between drug reinforcement and accumbal dopamine transmission was firmly established.

Although all the major drugs of abuse increase dopamine transmission, they do so in different ways. Opiates promote the release of dopamine by stimulating receptors that normally respond to opiate peptides found naturally in the brain, the so-called endorphins. Although there are at least three different receptors that respond to endorphins, opiate drugs primarily activate the mu receptor. One effect of this activation is an increase in dopamine release. Nicotine and alcohol promote dopamine release by acting at different receptors: nicotine stimulates a group of receptors that respond to acetylcholine, and alcohol interacts with the major receptor for GABA, an amino acid found throughout the brain. Both acetylcholine and GABA, like many of the endorphins, are transmitters that play a role in dopamine release. When opiates, nicotine, or alcohol are introduced into this system, their receptor effects overwhelm any naturally occurring activity and dopamine is released in abnormally high amounts.

Psychomotor stimulants, on the other hand, increase dopamine transmission by interacting with a protein that transports newly released dopamine back into the neuron for re-release. Cocaine prevents the transporter from working and amphetamine forces it to operate in reverse. Thus, cocaine allows dopamine to accumulate outside the neuron and amphetamine causes dopamine release. The net effect of either drug is an increase in synaptic dopamine.

A drug-induced increase in dopamine transmission, however, is only the beginning of the reinforcement story. The mesocorticolimbic pathway is embedded in a complex network of structures, each of which processes information relevant to drug-induced behavior such as emotional state, environment, past experience, and many other key variables. Investigations at the membrane level, moreover, reveal that rather than exerting a powerful excitatory or inhibitory influence on individual neurons, dopamine modulates the effects of other transmitters, serving more as a filter or gain mechanism than as a conveyor of specific information. In addition, some neurons, such as the endorphin system, may respond to drugs of abuse independently of a change in dopamine. Thus, the drug-induced dopamine signal may be only one of many that contribute to what might become a sense of pleasure or reward. The reinforcing effects of most drugs are likely to involve both dopamine-dependent and dopamine-independent neuronal systems.

Whatever role dopamine, endorphins, and other transmitters play in drug reward, the reward model itself is simply a starting point for understanding addiction. Compulsive drug use is not driven by euphoria alone. Many addicts actually report a loss of pleasure after chronic drug use, but the habit persists. A case can be made that liking a drug, which may parallel drug-induced reward, is entirely different from wanting a drug, sometimes described as the craving that overwhelms an addict's behavior. Noteworthy in this regard is evidence that the transition to compulsive drug use may reside in the neuronal changes that occur when the brain reward circuit is exposed to drugs on a chronic basis.

Neuroadaptations Underlying Addiction

One effect of chronic administration of drugs of abuse is a change in the responsiveness of dopamine receptors. Although these receptors exist in multiple forms, they can be grouped into what are known as D1 or D2 families. Both families are *metabotropic* in that their activation leads to intracellular metabolic changes that regulate membrane excitability and gene expression. Whereas the D2 group appears to be critical for normal motor behavior, the D1 family may play a greater role in the neural adaptations that accompany learning. These same adaptations may also be involved in the behavioral changes that accompany repeated exposure to drugs of abuse. In fact, addiction itself is likely to involve some type of learning-related change in the behavioral response to a drug. By disrupting the normal flow of dopamine transmission, and thus changing D1 receptors, drugs of abuse may set in motion a series of neural events that can lead to addiction. A drug-induced change in this receptor, such as a down-regulation or loss of sensitivity, can occur within minutes of an amphetamine injection. This may explain why the first administration of psychomotor stimulants, which are typically taken in a series of administrations known as a run or binge, has the greatest euphoric effect. As the run continues, the drug-induced high loses its intensity. Over a series of runs, however, prolonged stimulant exposure appears to up-regulate or enhance the sensitivity of D1 receptors, and this effect may persist for weeks after the last drug administration.

Stimulation of the D1 receptor family also leads to a change in gene expression and ultimately the production of cellular proteins. Up-regulation of these receptors, therefore, could mean not only an increase in synaptic transmission but also long-term structural changes. In fact, changes in dopamine synaptic structure have been reported in rats after amphetamine exposure. By usurping the dopamine system and driving it to excess, addictive drugs may permanently alter the flow of information through limbic and cortical circuits in a way that increases the likelihood of further drug use.

CONTROL AND TREATMENT OF ADDICTION

By the time an addict appears for treatment, the drug habit is likely to have persisted for years. During this time, there has been ample opportunity for the problems that typically accompany addiction – family, legal, medical, occupational, and social troubles – to complicate and confuse the treatment process. The greater these accompanying problems are, the more difficult it will be for treatment to succeed.

The first step towards rehabilitation involves detoxification, the removal of the addictive drug from the patient's system. The next step, preventing a relapse, is far more difficult, but various behavioral approaches can be effective. These include teaching coping skills or attempting to desensitize an addict to the cues associated with drug-taking. But if chronic exposure to addictive drugs causes lasting changes in critical brain circuits, as ample evidence suggests, then medication may occupy an important place in the rehabilitation effort.

Detoxification

For heroin addicts, detoxification is best accomplished in conjunction with a drug such as methadone, which acts as a mild heroin substitute. Gradually decreasing the maintenance dose of methadone is part of the detoxification strategy. For nicotine addicts, detoxification can be achieved relatively simply by gradually reducing the dose of nicotine delivered by skin patch, chewing gum, or nasal spray. In some cases, a low, maintenance dose of nicotine may continue for several months to encourage abstinence. Detoxification is especially important for alcoholics because there is danger of death from overdose. In addition, an alcoholic suffers convulsions and other life-threatening reactions that increase in severity each time the drug is withdrawn. To minimize these effects, alcohol detoxification includes treatment with one or more benzodiazepines, which act as sedatives or minor tranquilizers. For cocaine or amphetamine addicts, detoxification becomes important in the event of an overdose, which can be lethal. Drugs can be used to counteract the racing heart, high blood pressure, and other dangerous effects associated with stimulant overdose.

Medication Strategies

Medication is used in one of three basic strategies to keep an addict off drugs: antagonizing or blocking the effect of the addictive drug; substituting another drug from the same class as the addictive drug but with less powerful effects; or the use of medication specifically designed to reduce craving. Medication is now available for patients dependent on opiates, nicotine, and alcohol.

In the treatment of opiate addiction, the antagonist strategy involves administering a drug that has a high affinity for the mu receptor but does not cause the same cellular reaction as heroin or morphine. One such drug is naltrexone. Rather than activating the mu receptor, naltrexone acts as a mu antagonist, binding to the mu receptor but preventing it from working. When given to heroin addicts, for example, naltrexone blocks the effects of a subsequent heroin injection. The problem with such treatment, however, is that all the subjective effects of heroin that an addict has come to expect are prevented. Another problem is that if heroin is still present in the body, naltrexone creates an immediate withdrawal syndrome. Thus, unless an addict is firmly committed to overcoming the habit, the antagonist strategy is rarely successful.

Most heroin addicts prefer the substitution strategy. In this case, the drug substituted for heroin acts like heroin itself in that it also stimulates the mu receptor. Unlike heroin, however, the substituted drug has a relatively slow onset of effects and stays attached to the receptor for prolonged periods of time. Thus, the addict experiences some heroin-like effects but in relatively mild form. Moreover, because the substituted drug stays attached to the receptor for a day or more, a subsequent injection of heroin during this time will not elicit the intense rush or high likely to trigger more craving. Methadone was introduced in the 1960s as the first such substitution drug for heroin addicts. Now, even longer-lasting substitutes, such as buprenorphine and L-acetylmethadol (LAAM), have been developed, which may attach to the mu receptor for up to three days. More than 100,000 heroin addicts in the USA are currently being treated with methadone or long-acting opiate substitutes.

Naltrexone is sometimes used in the treatment of alcoholism to reduce craving. This strategy emerged from evidence that endorphins play a role in reward and that blockade of mu opiate receptors decreased responding for alcohol in animals. It may be that the GABA system, which is sensitive to alcohol, interacts with endorphins in the forebrain reward circuit. That naltrexone can reduce craving in human alcoholics supports this view. Another anti-craving compound is acamprosate, which is used in Europe to reduce relapse in detoxified alcoholics. Although some of its brain actions are still obscure, acamprosate is known to modulate neuronal excitability in the nucleus accumbens by interacting with specific groups of glutamate and GABA receptors. Interestingly, nicotine craving can be reduced by bupropion, a drug commonly used to treat mild to moderate cases of clinical depression. Bupropion has a wide range of effects, but its ability to block nicotine receptors and modulate dopamine transmission may play key roles in reducing nicotine craving.

A fourth medication strategy, the use of nausea-inducing drugs, is available for treating alcoholics. They can obtain a prescription for disulfiram (antabuse), a drug that interferes with the normal metabolism of alcohol, creating an abundance of acetaldehyde. A high level of acetaldehyde causes headache, vomiting, disorientation, and other signs of nausea that are so repulsive that the impulse to drink disappears. The problem with this treatment strategy is compliance; an alcoholic who wants to drink again can simply stop taking disulfiram.

The development of medication for treating an addiction to cocaine or other stimulants is still in its infancy, but several possibilities are being explored. One is akin to the methadone strategy for heroin addiction and involves developing a drug that activates dopamine receptors to produce a partial stimulant-like effect with the hope that the addict could be gradually weaned away from the medication without a reinstatement of craving. Another approach is to interfere with the ability of cocaine or amphetamine to reach their main site of action, the dopamine transporter protein. In this case, the drugs would lose the dopamine-enhancing effect that may underlie addiction. No dramatic successes have emerged with either approach in clinical trials, but there is still much to be learned about the neuropharmacology of the dopamine system and how this system interacts with other transmitters. New medications continue to be developed, and further study of their mechanism of action is likely to lead to new strategies for treatment.

PHYSICAL AND PSYCHOLOGICAL ADDICTION

For most of the twentieth century, addiction was explained in terms of how the body responded when the drug was no longer available. The more severe the response, the stronger the addiction. Consider heroin, a drug that causes a wide range of effects, including analgesia, muscle relaxation, suppressed gag reflex, low core body temperature, and constipation. When the drug is stopped or withdrawn, an addict experiences exactly the opposite sensations: pain, tension, retching, fever, and diarrhea. These withdrawal symptoms, which could last for days or weeks depending on how much and for how long the drug was used, were

considered a sign of addiction or, more appropriately, physical dependence. Without the drug, the body became physically sick. To avoid this condition, more of the drug was required, thus perpetuating addiction.

Physical dependence was also used to explain addiction to alcohol and nicotine because of the unpleasant physical reactions that occur when these drugs are withdrawn. In fact, a drug was not considered addicting unless there were clear and profound signs of physical dependence. The problem with this model is that it cannot explain addiction to many drugs, including stimulants like cocaine and the amphetamines. Apart from fatigue or depression, there are no profound withdrawal symptoms associated with these drugs, yet they are strongly addicting. To deal with this issue, some theorists proposed the concept of psychological dependence, the notion that drug withdrawal could lead to a mental or psychological state that triggered addiction. This model was even more difficult to accept because the concept of a mental or psychological addiction was impossible to define and thus impossible to study empirically. It never caught on as a useful model of addiction.

Remnants of the psychological dependence model, however, are apparent in the use of such terms as 'reward' or 'positive reinforcement' to explain a drug habit. Such terms are invoked when the compulsion to take drugs cannot be explained by physical dependence. In fact, even the physical dependence model cannot explain why former heroin addicts or alcoholics can relapse months or years later, long after the withdrawal syndrome has passed. Interesting in this regard is evidence that animals will work very hard to obtain addictive drugs, even at doses that fail to elicit signs of physical dependence. Thus, although physical dependence is a legitimate reason for why some addicts maintain a drug habit, the model has limited applications. As brain research has revealed, drug craving is a long-term process that most probably involves a dysfunction of the neural circuitry underlying motivational behavior. An overriding question is why only some persons exposed to addictive drugs become addicts.

PREDISPOSITION TO ADDICTION

There appear to be certain risk factors that make some individuals especially vulnerable to addiction. These include age, mental state, and personality type, as well as a variety of environmental and genetic conditions. No single factor by itself is a good predictor of addiction but as the number of risk factors increases, the likelihood of addiction also increases.

The highest rates of illicit drug use occur among 18- to 25-year-olds. Experimenting with drugs of abuse at earlier ages, including the pre-teen years, increases the chance of addiction. Some addicts also have co-existing psychiatric conditions, such as depression or schizophrenia, that may impair judgment. Although the psychiatric problem sometimes precedes the addiction, it is most often the case that the addiction develops first. Personality traits may also contribute to addiction, but the relationship is difficult to specify because, as with psychiatric problems, it is not clear if abnormalities in personality are a cause or a consequence of the addiction. In some cases, however, personality influences on drug-taking behavior have been studied in controlled laboratory settings, and the data suggest a problem for adventurous and antisocial personalities, the so-called novelty-seekers or risk-takers. Their lack of impulse control not only puts them in vulnerable situations with respect to drug use but may also complicate rehabilitation. The antisocial personality, for example, correlates with poor outcome in methadone maintenance programs.

Environmental influences on drug-taking behavior take many forms and range from parenting practices to the prevalence of public education efforts. In fact, efforts at informing people about the dangers of drugs are at their peak during times of peak drug abuse and then decline as drug abuse declines, which may help pave the way for the start of another peak period of drug abuse. Wide swings in drug use, such as a sixfold variation in alcohol consumption in the United States over the course of the twentieth century, largely reflect changes in social attitudes and public policy. The context in which drugs are taken is also important. This point was nicely illustrated during the Vietnam war in the 1960s and 1970s. Many United States servicemen became addicted to heroin while serving in Vietnam but either stopped the habit completely or dramatically lowered their heroin use when they returned home. In Vietnam, high-quality opiates were readily available at a cheap price, and any disapproving family members were likely to be thousands of miles away. These factors, combined with the high stress of battle conditions, made it easy to justify drug use. Many servicemen, moreover, disconnected their one-year tour of duty in Vietnam from the rest of their lives. Most, if not all, of these

contributing factors disappeared when the tour of duty ended.

Genetics also may contribute to drug addiction, but genes alone are not destiny. In fact, only one in five of those genetically at risk for alcoholism actually become alcoholic. Another consideration is that even if the genetic influence to addiction is critical, there is likely to be more than one genetic factor involved. Genetics, for example, can contribute to the rate and efficiency at which a drug is metabolized as well as to how receptor proteins in the brain respond to a drug. Both of these factors will influence how a drug alters behavior. The strength of that influence, moreover, will depend on the influence of all the other risk factors that contribute to addiction.

Some attempts to specify the potential influence of genetic and nongenetic risk factors involve work with animal models. In this type of research, mice and rats are strategically bred to establish genomic correlations with drug-induced behavioral responses, to identify neurobiological mechanisms, and to localize the chromosome associated with specific drug-induced behavioral traits. In many cases, the identification of drug-response genes in mice indicates the appropriate gene location on human chromosomes. This work is being carried out for all major drugs of abuse and holds promise for assessing the genetic and environmental risk factors underlying addiction.

Further Reading

Berke JD and Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**: 515–532.

Berridge KC and Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* **28**: 309–369.

Koob GF and LeMoal M (2001) Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology* **24**: 97–129.

Leshner AI (1997) Addiction is a brain disease, and it matters. *Science* **278**: 45–47.

Nestler EJ and Aghajanian GK (1997) Molecular and cellular basis of addiction. *Science* **278**: 58–63.

O'Brien CP (1997) A range of research-based pharmacotherapies for addiction. *Science* **278**: 66–70.

Pickens RW, Elmer GI, LaBuda MC and Uhl GR (1996) Genetic vulnerability to substance abuse. In: Schuster CR and Kuhar MJ (eds) *Pharmacological Aspects of Drug Dependence*, pp. 3–52. [*Handbook of Experimental Pharmacology*, vol. 118.] Berlin, Germany: Springer-Verlag.

Porrino LJ and Lyons D (2000) Orbital and medial prefrontal cortex and psychostimulant abuse: studies in animal models. *Cerebral Cortex* **10**: 326–333.

Tarter RE, Ammerman RT and Ott PJ (eds) (1998) *Handbook of Substance Abuse: Neurobehavioral Pharmacology*. New York, NY: Plenum.

White FJ and Kalivas PW (1998) Neuroadaptations involved in amphetamine and cocaine addiction. *Drug and Alcohol Dependence* **51**: 141–153.

Affective Disorders: Depression and Mania

Introductory article

Michael T Compton, Emory University School of Medicine, Atlanta, Georgia, USA
Charles L Raison, Emory University School of Medicine, Atlanta, Georgia, USA
Charles B Nemeroff, Emory University School of Medicine, Atlanta, Georgia, USA

CONTENTS

Affective disorders are psychiatric illnesses characterized by disturbances of affect or mood, which cause emotional, cognitive, and behavioral symptoms that significantly impair functioning. They are classified according to polarity (the degree of mood elevation or depression), and are treated with pharmacological and/or psychotherapeutic interventions accordingly.

INTRODUCTION

Affective disorders have been described and treated since the age of ancient Greco-Roman medicine. Hippocrates described melancholia (an excess of 'black bile') around 2400 years ago. Affective excitement, or mania, was also described by ancient physicians. Aretaeus of Cappadocia recognized that mania and depression were usually connected in the same individuals, indicating that the concept of bipolar disorder was anticipated in antiquity. Affective disorders continue to be subdivided according to polarity: unipolar disorders are characterized by depressive episodes only, whereas bipolar disorders are marked by the presence of hypomanic, manic or mixed episodes, with or without intervening depressive episodes.

CLASSES OF AFFECTIVE DISORDERS

Affective disorders are currently diagnosed using four specific mood episodes as building blocks: major depressive episode, manic episode, hypomanic episode, and mixed episode.

Mood Episodes

A major depressive episode consists of a period of at least 2 weeks during which five or more specific symptoms are experienced, representing a change from previous functioning. At least one of these symptoms must be either depressed mood, or loss of interest or pleasure (anhedonia). Other potential symptoms include significant change in appetite or weight, insomnia or excessive sleeping, psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness or inappropriate guilt, poor concentration or indecisiveness, and suicidal thoughts. Patients suffering from a major depressive episode often experience negative cognitive patterns, including helplessness, hopelessness, and preoccupation with inadequacy. Low self-esteem is commonly present. Psychotic symptoms accompany approximately 15% of depressive episodes.

A manic episode is a distinct period of abnormally elevated, expansive, or irritable mood lasting at least a week or requiring hospitalization. During this period, three or more specific symptoms (four or more if the mood is only irritable) typically occur: inflated self-esteem or grandiosity; decreased need for sleep; being more talkative than usual or feeling under pressure to keep talking; flight of ideas, or subjective experience that thoughts are racing; distractibility; increase in goal-directed activity or psychomotor agitation; and excessive involvement in pleasurable activities that have a high potential for painful consequences. Psychotic symptoms develop in approximately 25% of manic episodes and typically consist of grandiose delusions, hallucinations of deities, or paranoia born of a delusional sense of importance.

A hypomanic episode, which is less severe than mania, is a distinct period of elevated, expansive or irritable mood lasting for at least 4 days, during which time three or more symptoms of a manic

episode are experienced (four if the mood is only irritable). The changes represent an unequivocal change in functioning, which is observable by others, but is not severe enough to cause marked impairment in functioning or to necessitate hospitalization, and there are no psychotic features present.

A mixed episode is a period of at least 1 week during which the above criteria for both a major depressive episode and a manic episode are experienced. Few patients have episodes that meet the full criteria for a mixed state, but dysphoric manias, in which a depressed and/or anxious mood coexists with manic symptoms, are frequent in patients with bipolar disorder.

For all four of the above mood episodes, the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) specifies that the symptoms should not be due to the direct physiological effect of a substance (e.g. drug of abuse or medication) or a general medical condition.

Mood Disorders

Major depressive disorder is defined by the lifetime presence of at least one episode of major depression that is not precipitated by a medical illness, a medication, or a substance of abuse and is not better accounted for by bereavement. Dysthymic disorder is a mild form of depression that often has an early age of onset (in childhood, adolescence, or early adult life) and usually lasts for protracted periods (at least 2 years in adults, according to DSM-IV). Patients with dysthymia do not meet full criteria for major depression during the course of their illness. Dysthymia tends to be characterized more by emotional and cognitive symptoms, such as depressed mood, pessimism and poor self-esteem, than by the types of neurovegetative symptoms seen in major depression (i.e. changes in sleep, appetite, and physical activity level).

Bipolar I disorder is defined by the lifetime presence of a manic episode not judged to be caused by a medical illness or ingestion of recreational drugs or medications. Bipolar II disorder is characterized by the presence of major depressive episodes and periods of hypomania. Although hypomanic episodes tend, by definition, not to produce the degree of functional impairment caused by full manias, bipolar II disorder patients may actually have more overall impairment in their lives than many patients with bipolar I disorder, owing to the greater length and treatment resistance of depressive episodes in bipolar II disorder, as well as to the tendency of patients with bipolar II disorder to develop rapid cycling, a condition characterized by rapid progression from one mood episode to the next with little or no period of normal mood between episodes.

Cyclothymia, like dysthymia, is a chronic condition characterized by the presence of mood symptoms insufficient to meet the criteria for a major affective disorder. In the case of cyclothymia, these symptoms take the form of a repeated alternation of subsyndromal hypomanic and depressive symptoms. It appears that the presence of cyclothymia is a risk factor for the later development of both bipolar I and II disorders, although some patients never experience a full mania or major depression.

Affective disorders are distinguished from other psychiatric conditions not so much by specific symptoms as by the fact that alterations in mood and/or a loss in the ability to find pleasure in life are considered to be the primary derangements in the disease course of the individual. Beyond this central focus, however, patients with affective disorders can evince symptoms seen in a number of other psychiatric conditions. In addition, comorbidity is the rule rather than the exception in psychiatry. Other psychiatric conditions that are highly comorbid with affective disorders include anxiety disorders (generalized anxiety disorder, social phobia, panic disorder, and obsessive–compulsive disorder) and various alcohol and drug abuse and dependence problems.

ETIOLOGY OF AFFECTIVE DISORDERS

Genetics and Biological Factors

Affective disorders, especially bipolar disorder because of its greater degree of familiality (stronger evidence for a genetic basis), have been the subject of much research into their genetics and heredity. Despite methodological and interpretive limitations, several reproducible findings have emerged.

The risk of bipolar disorder in first-degree relatives of patients with bipolar disorder ranges between 3% and 8%, compared with a 1–1.5% rate in the general population. The risk of depressive disorders among first-degree relatives of patients with depression is two to three times that of the general population. If one parent has an affective disorder, a child's risk of an affective disorder is between 10% and 25%; if both parents are affected, the risk roughly doubles. Having more affected family members confers increased risk, especially when family members have bipolar disorder.

In addition to genetic factors it is becoming increasingly clear that, as a group, people with

affective disorders differ physiologically from well-matched controls without histories of depression or mania. Evidence suggests that patients with affective disorders have consistent changes in brain structure and functioning. These changes include decreased volume in prefrontal and temporal cortex, as well as in the basal ganglia. Consistent with these structural changes, many patients with major depression demonstrate decreased functioning in these same areas when they are studied by modern imaging techniques. Postmortem studies suggest that changes in brain volume and function may reflect cell loss and/or atrophy as well as loss of functional connections (synapses) between nerve cells in these brain areas. Major depression is associated with changes in sleep architecture. Some of these changes remit when affective symptoms resolve, but others remain, and are even found in unaffected family members, suggesting that such sleep abnormalities may be a risk factor for the development of affective disorders.

Psychological and Social Factors

Major depression frequently develops in a psychological context of anxiety and/or neurosis, as well as in individuals with a long-term tendency toward excessive social inhibition. Many patients with bipolar disorders, especially bipolar II disorder, demonstrate premorbid psychological traits of moodiness, impulsivity, and acting-out behavior. The likelihood that a person will develop an affective disorder is also influenced by a number of environmental factors, many of which are psychosocial in nature. For example, it is known that childhood conditions such as the death of a parent, or exposure to neglect or to physical, sexual or emotional trauma, pose a significant risk for the development of affective disorders later in life. Although traumatic events early in life appear to be especially potent in fostering later depression, probably as a result – at least in part – of negatively influencing postnatal brain development, actual or perceived stress at any time in the life cycle is strongly associated with the development of depression.

Investigations into psychological and environmental contributions to the affective disorders tend to be empirical in nature, using biological and ethological approaches in an attempt to map out pathways through which psychological and environmental factors influence mood. However, many of these research interests have been anticipated by the more theoretically driven approaches to the issue that dominated the field of psychiatry until well into the last half of the twentieth century. Many such theories derive from psychoanalytic perspectives and tend to privilege early life experiences. Other schools of thought have advanced behavioral and cognitive factors as primary causes of affective disorders. Fewer psychological theories have tried to explain the experience of mania, and genetic vulnerability has been more clearly implicated. Nonetheless, psychological models have been proposed for manic reactions. For example, some theorists have described mania as a stance defensive against an underlying depression (for example, self-deprecation and excessive guilt of depression become replaced by grandiosity and elevated self-esteem characteristic of hypomania or mania). Key figures in the formulation of these models include Karl Abraham and Melanie Klein.

ONSET AND COURSE OF AFFECTIVE DISORDERS

Affective disorders are highly prevalent illnesses that significantly impair functioning, interpersonal relationships, and quality of life. In the USA, lifetime prevalence rates are 17.1% for major depressive episode, 6.4% for dysthymia, and 1.6% for manic episode. Epidemiological studies reveal that affective disorders are more prevalent among individuals under the age of 45 years (the average age of onset is 20–40 years). Women suffer from unipolar depressive disorders about twice as often as men. Certain points in the human life cycle confer increased risk for the development of depression: these include puberty, as well as childbirth in women and the passage through middle age in men (after middle age, rates of depression are equal between the genders). Menopause may be a risk factor for mood disturbance in certain women, but does not pose anything like the threat of the postpartum period, a time in which 20% of new mothers will develop a major depression. Rates of bipolar disorder do not appear to be affected by gender. The prevalence of affective disorders does not vary significantly by race or ethnicity.

It is increasingly recognized that both unipolar depressive disorder and bipolar disorder are frequently recurrent illnesses that lead to increased functional impairment with the passage of time and the accumulation of episodes. On average, people diagnosed with unipolar major depression in youth or early adulthood can expect five or six episodes over their life span, compared with an average of eight to nine major lifetime mood episodes in people with bipolar disorder. Twenty-five percent of patients with unipolar major depression

demonstrate a chronic course. Similarly, 5% of patients with bipolar I disorder remain chronically manic, often providing a diagnostic conundrum to clinicians used to assigning a diagnosis of schizophrenia to anyone with a chronic psychotic illness. One-third of patients with unipolar depression will have only a single episode. These patients tend to be older at the time of onset and are less likely to have a history of affective disorders in their families.

Affective disorders underlie 50–70% of all cases of suicide, and individuals with serious depression (i.e. requiring hospitalization) have a 15% suicide rate. Rates of suicide in bipolar disorder are probably higher, approximately 20–25%.

TREATMENT

Pharmacologic and Somatic Treatments

Pharmacologic and somatic treatments for depression compare favorably with the pharmacologic treatment of other chronic medical disorders in terms of efficacy. A broad spectrum of effective antidepressant strategies are available, from electroconvulsive therapy through a wide array of antidepressant medications. In addition, psychotherapy and medications have a synergistic effect when used in combination.

The modern era of psychopharmacologic treatment was initiated in the mid-twentieth century with the discovery that lithium, monoamine oxidase inhibitors, and tricyclic antidepressants effectively treated affective disorders. Monoamine oxidase inhibitors (MAOIs) block the enzyme that metabolizes biogenic amines, increasing the availability of these neurotransmitters. Though especially efficacious for atypical depression, the use of MAOIs is greatly limited by their adverse effects, by their potential to induce possibly lethal hypertensive episodes in combination with foods that contain tyramine, and by a propensity to induce a frequently fatal condition known as 'serotonin syndrome' when taken in combination with medications that affect serotonin functioning. Tricyclic antidepressants (TCAs) are thought to work by blocking the reuptake of noradrenaline (norepinephrine) and serotonin into presynaptic nerve cells. However, in addition to these therapeutic actions at reuptake transporter sites, TCAs block other receptors leading to a host of unwanted side effects, including blurred vision, dry mouth, tachycardia, constipation, urinary retention, cognitive dysfunction, postural hypotension, dizziness, sedation, weight gain, and sexual effects. Of even more concern is that fact that TCAs are highly lethal in overdose as a result of direct effects on cardiac conduction.

As suggested by their name, selective serotonin reuptake inhibitors (SSRIs) work by blockade of the serotonin reuptake pump. Unlike the MAOIs or TCAs, these newer agents were specifically designed to diminish or abolish activity at other receptors known to mediate many side effects of the older drugs. As a result, SSRIs are associated with a decreased side-effect burden and significantly enhanced safety in overdose. However, these agents are not without their own limitations, including a high rate of sexual dysfunction, and a tendency, as a class, to block the metabolism of other medications, leading to potentially problematic interactions with other drugs. Antidepressants developed since the arrival of the first SSRIs show a trend toward reversing the specificity of action that was the goal in SSRI development. Examples of newer medications that use a combined action strategy include venlafaxine, which works by blocking the reuptake of both serotonin and noradrenaline, and mirtazapine, which blocks presynaptic noradrenaline autoreceptors and postsynaptic serotonin receptors.

Finally, although typically reserved for patients who have failed to respond to other treatments, electroconvulsive therapy (ECT) remains a highly effective therapy for depression. It is especially effective in melancholic and/or psychotic forms of depression and for patients whose illness has catatonic features. Nonetheless, ECT has long been hampered by the striking short-term memory loss associated with the treatment, as well as by the risks of undergoing anesthesia and the general difficulties inherent in implementing this typically hospital-based procedure.

Because bipolar disorder is a recurrent, severely impairing psychiatric disorder, the search for efficacious pharmacological treatments continues to be at the forefront of psychiatric research. Lithium is a naturally occurring salt that was found to have antimanic properties in 1949. In the experience of many experts, lithium remains the 'gold standard' for treatment of both the manic and depressive phases of bipolar disorder. In addition to its acute effects in curtailing mania and depression, lithium when taken chronically is known to protect against the advent of future mood episodes. Lithium treatment also decreases the lifetime risk of suicide in people with affective disorders, a finding that has been difficult to replicate with other pharmacological agents.

Several anticonvulsant medications have attracted attention as treatments for bipolar disorder. The use of valproic acid for the treatment of acute mania is supported by data suggesting that (like lithium) it is effective in protecting patients from future mood episodes. Valproic acid appears to be more effective than lithium in treating dysphoric manias and may give better results in patients who have a disease course characterized by manias that follow depressions, rather than the other way around. Lamotrigine may be effective for bipolar depression, with a minimal risk of the induction of mania or hypomania that accompanies the use of traditional antidepressants. Carbamazepine was the first anticonvulsant used in the treatment of bipolar disorder, but its use has largely been eclipsed by valproic acid, which has a more tolerable side-effect profile in many patients.

In addition to mood stabilizers such as lithium and the anticonvulsants, antipsychotic agents and benzodiazepines have been shown to be effective in the acute treatment of mania, both as additions to mood stabilizers and as single agents. Finally, ECT is an effective treatment in both the manic and depressed phases of bipolar disorder, and is especially effective when patients demonstrate catatonic symptoms.

Psychotherapeutic Treatments

Several forms of psychotherapy are roughly as effective as antidepressant drugs in the treatment of most cases of major depression. Current treatment of bipolar disorders favors the use of pharmacological and somatic interventions, but studies suggest that psychotherapy has a valuable adjunctive role in these conditions. Two commonly employed psychotherapeutic techniques that have received considerable empirical validation are interpersonal therapy and cognitive behavioral therapy.

Interpersonal therapy is descended from a long-held and valuable insight from the psychoanalytic tradition that interpersonal relationships are crucial to the maintenance of a normal mood state (euthymia) for most people, and that major depression is frequently associated with disturbances in the realm of interpersonal relations. The therapist directly explores and works with the patient's interpersonal difficulties, operating under the assumption that these difficulties are frequently causative of depression and that their resolution promotes a return to euthymia. Cognitive behavioral therapy is a highly structured, effective short-term psychotherapy that aims to correct negative thought patterns, specific distorted schemas, and cognitive errors. Unlike older psychoanalytic traditions, cognitive behavioral therapy specifically and directly confronts maladaptive cognitions through a process of gentle challenging, but also through the use of educational components and 'homework' aimed at encouraging the patient to practice more positive behavior in daily life. Interestingly, although these two psychotherapies are based on different – though not mutually exclusive – theoretical underpinnings, most studies suggest they are equally effective in the treatment of depression.

NEURAL CORRELATES AND THEORIES OF AFFECTIVE DISORDERS

No single physiological abnormality has been found to account for affective disorders, but many years of research have established a number of abnormalities that are present in many affectively ill patients. The current biological theories of affective disorders began to emerge in the 1950s coincident with the discovery of antidepressants and mood stabilizers. Based on the suspected mechanisms of action of antidepressants, early theories focused on putative deficiencies in single biogenic amine neurotransmitter systems, especially noradrenaline, and later, serotonin. While these early (and relatively simplistic) proposals have given way to far more complicated heuristic models, it remains true that affective disorders as a whole continue to be characterized by evidence of altered functioning in catecholamine (noradrenaline and dopamine) and indolamine (serotonin) neurotransmitter systems, as well as in other related systems such as the hypothalamic–pituitary–adrenal axis that, together with the sympathetic nervous system, constitutes the mammalian stress response system.

Biogenic Amines

Based on existing technologies, early psychobiological investigations into the role of biogenic amines in the pathogenesis of affective disorders focused on the concentration of these neurotransmitters (or their metabolites) in blood, urine, and cerebrospinal fluid. Taken as a whole these studies suggested a decrease in serotonergic activity in people with unmedicated depression. The relationship between noradrenaline production and depression appears more complex. Early studies reported decreased urinary levels of catecholamine metabolites in depression, and animal models of stress suggest that depleted noradrenaline in selected brain areas might contribute to the

diminished energy, inability to feel pleasure, and decreased libido that frequently accompany depression. However, later research suggested that people with melancholic major depression have a markedly increased release of catecholamines into the cerebrospinal fluid.

Despite the initial attractiveness of these neurotransmitter models, significant evidence has accumulated countering the notion that affective disorders are caused by a simple neurotransmitter deficiency or excess. Current formulations of the role of biogenic amines in depression tend to focus on the potential restorative role that increased levels of these transmitters might have for 'downstream' neural functioning. Chronic alterations in aminergic availability fostered by antidepressant medication may stimulate intracellular second-messenger pathways to favor the production of neurotrophic factors that lead to enhanced functioning of neurons in brain areas believed to be centrally involved in mediating mood abnormalities.

Neuroendocrine Axes and Neuropeptide Systems

The hypothalamic–pituitary–adrenal (HPA) axis serves as the principal mammalian stress response system. Abnormalities of this axis are amongst the most replicable biological findings associated with affective disorders, especially major depression. These abnormalities in major depression include evidence for increased corticotrophin releasing hormone (CRH) production in the central nervous system, a blunted adrenocorticotrophin response to CRH stimulation, enlargement of pituitary and adrenal glands, increased blood levels of cortisol, and decreased tissue sensitivity to glucocorticoids. Evidence for heightened production of CRH in the brain is especially intriguing, given that this neuropeptide serves as a neurotransmitter in limbic brain regions linked to depression, in addition to its role in the hypothalamus as the primary activator of the HPA axis.

Hypothyroidism is frequently associated with a markedly depressed mood. Studies have documented alterations in thyroid axis activity in patients with depression, including increased thyrotrophin releasing hormone (TRH) concentrations in cerebrospinal fluid, altered thyrotrophin response to TRH stimulation, decreased nocturnal plasma thyrotrophin concentrations, and presence of antimicrosomal thyroid or antithyroglobulin antibodies. Similar abnormalities have been associated with bipolar disorder, the course of which is worsened by hypothyroidism.

CONCLUSION

Affective disorders, including unipolar depressive disorder and bipolar disorders, are prevalent psychiatric illnesses, affecting approximately 25% of individuals during the course of a lifetime. They are classified by the presence of discrete or chronic mood episodes characterized by abnormal depression or elevation of mood and mood-related functions below or above the individual's baseline. The causes of affective disorders are multifactorial, with both genetic and biological factors involved, and modified by psychosocial variables. Treatment consists of psychiatric management with pharmacological agents, ECT, and psychotherapy. Contemporary psychiatry and neuroscience are increasingly concerned with the neurobiological correlates of affective disorders, predictors of response to particular therapies, and achieving return to complete euthymia (remission).

Further Reading

American Psychiatric Association (1994) *Practice Guideline for the Treatment of Patients with Bipolar Disorder*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (2000) *Practice Guideline for the Treatment of Patients with Major Depressive Disorder* (revision). Washington, DC: American Psychiatric Association.

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn, revised. Washington, DC: American Psychiatric Association.

Goodwin F and Jamison K (eds) (1990) *Manic-Depressive Illness*. New York, NY: Oxford University Press.

Kaplan HI and Sadock BJ (eds) (2000) *Comprehensive Textbook of Psychiatry*, 7th edn. Baltimore, MD: Williams & Wilkins.

Aging and Cognition

Introductory article

Pat Rabbitt, University of Manchester, Manchester, UK

CONTENTS

The study of aging and cognition aims to identify the mental changes that occur in old age and relate them to corresponding neurophysiological processes.

INTRODUCTION

Cognitive aging cannot be equated with passage of calendar time because it is a complex of processes that proceed at different rates in different people and in different parts of the brains and central nervous systems (CNS) of individual persons. A main goal of cognitive gerontology is to identify the mental changes that occur in old age and to relate them to these neurophysiological changes. Demographics, socioeconomic status, and lifestyle are also important because factors such as gender, prolonged education, lengthy marriage to an intelligent spouse, complexity of workplace and social environments, level of income, and personality type all affect longevity and also maintenance of cognitive functioning in old age. Because the influences on cognitive aging are so numerous, and their interactions are so complex, individuals have strikingly different trajectories of aging. Consequently, as a population ages, the widening gap in competence between its most and least able members is more striking and informative than is the average decline in ability.

SENSORY CHANGES AFFECT, AND PREDICT, COGNITIVE ABILITY

Loss of efficiency of the sense organs is among the most obvious and widespread changes with age. Besides curtailing efficiency in obvious ways this increases the effort necessary to resolve degraded visual and auditory information, and this distracting difficulty makes it more difficult to remember, or to make useful associations to, material that has been correctly heard or seen. Losses of visual and auditory acuity, and also of balance, muscle strength, and lung capacity, are good markers for levels of cognitive efficiency in old age, probably because they reflect changes in the brain CNS resulting from cerebrovascular inefficiency and other causes.

Effects of 'Primary' and 'Secondary' Aging

For some investigators, the growing dependence of mental ability on physiological efficiency as old age advances begs the question of the relative extents to which cognitive changes are caused by 'primary' processes of 'normal' or 'usual' aging that may be genetically programmed to occur at different rates in different individuals, or by the 'secondary' effects of lifetime accumulations of pathologies and biological damage. Unsurprisingly there is clear evidence that pathologies that become common in later life, such as late onset diabetes, hypertension, or loss of cardiovascular and respiratory efficiency, can significantly reduce cognitive ability.

However, the boundaries between 'primary' and 'secondary' aging are seldom clear-cut or physiologically meaningful. For example, 'primary' heritable differences in immune system efficiency determine longevity and so also maintenance of cognitive efficiency, by protecting against the 'secondary' effects of pathology. Consequently, the effects of 'primary' aging have been defined by default after pathologies have been identified and taken into account. Alternatively, the relative impacts of primary and secondary effects have been estimated by taking individuals' calendar ages as a rough proxy for the progress of 'primary' aging and using multivariate statistics to estimate the relative amounts of variance in mental abilities between individuals that are associated with differences in their calendar ages and with their

self-reported, or diagnosed, health status. Such analyses typically find that while differences in individuals' calendar ages account for only 15 to 20 percent of variance between them, differences in their self-reported health accounts for 2 percent or less.

These estimates can be misleading because while individuals' cognitive performance sharply declines with the number of different pathologies from which they suffer, people who volunteer for studies of cognition in old age are, typically, unusually fit, active, capable, and highly motivated members of their age groups. The variance associated with pathology is thus largely contributed by a few who, as individuals, have suffered severe losses. Another difficulty is that older people's self-reports of their health and cognitive efficiency must be cautiously interpreted because they cannot make absolute judgments of their status and so tend to compare themselves with their frail coevals. Perhaps for this reason, though the number of different pathologies that individuals report significantly increases with their ages, their subjective ratings of their general health may alter accordingly.

Pathologies that Directly Affect Cognitive Function

Apart from illnesses that indirectly affect mental ability by causing general physiological changes that have knock-on consequences for the central nervous system there are also pathologies, collectively known as 'dementias', that directly affect the brain. On its own the term 'dementia' is a label of convenience for changes in mental competence, much grosser than those observed in 'normal' aging, resulting from a diversity of different causes. For example, 'multi-infarct' dementia is a condition of diffuse brain damage often caused by small, but relatively numerous, cerebrovascular accidents (such as strokes, aneurisms, and minor narrowings and blockages of blood vessels) while others, such as Pick's disease, are more tightly defined in terms of causes and symptoms. The most common of these conditions, accounting for over 60 percent of all cases of dementias, is Alzheimer's disease (AD). This condition involves death of and changes in neurons resulting in the appearance of its main diagnostic criterion: 'senile plaques' and neurofibrillary tangles in brain cortex. Unfortunately these key signs can still only be ascertained on post mortem so that early diagnosis remains difficult. Other brain changes are neurotransmitter abnormalities, decreased brain volume, and, as is increasingly recognized, the incidence of many, and substantial, brain lesions that bring about marked loss of general cognitive ability and problems with memory, language, and general mental ability. Emotional problems such as anxiety, agitation, and depression, and marked personality changes may also occur. Any of these changes may also be encountered in dementias resulting from other causes.

The prevalence of AD sharply increases with age. 'Early onset' AD may occur in the 40s and 50s, the condition becomes more common in the 60s, and, it has been claimed, more than 30 percent of individuals aged 70 and above may be diagnosed as having cognitive changes consistent with a prognosis of AD. Early onset AD, occurring in middle age, progresses relatively rapidly, but when the condition occurs later in life it usually develops more slowly and patients may survive for a decade or longer.

For this article the issue is whether the marked and accelerating increase in incidence of AD with age means that it is the 'natural' and inevitable end state for all who survive long enough, or whether it should be regarded as a distinct pathology unrelated to 'normal' cognitive aging. The evidence is inconclusive. Although there are clearly genetic predisposing factors for AD, and there have been suggestions that viral or prion infection may be a causal or a triggering factor, all of the brain changes seen in AD, including the key diagnostic signs of neurofibrillary tangles and senile plaques, are found, though to a much less marked extent, in most healthy older individuals. In this respect the definition of 'normal' aging can, pragmatically, be treated as those changes in brain and behavior that are observed when a diagnosis of AD and of other dementias has been eliminated. Since, at the moment, this can only confidently be done at post mortem, the distinction is logically useful, but not always practically helpful.

Accounting for the Increased Variability between Individuals in Aging Populations

A main aim of cognitive gerontology is that by understanding the nature of declines in ability we may find means to delay them. It is therefore encouraging that trajectories of change vary markedly between individuals, because we may hope to discover what makes some more fortunate than others. For all mental skills, except those, such as vocabulary, that have been acquired early in life and practiced continuously ever since, plots of average levels of performance for successive age

groups show a steadily accelerating decline after youth (Figure 1(a)). However, these averages are uninformative and might represent either of two, very different, limiting scenarios. One is that individuals' trajectories of change all have the form illustrated in Figure 1(a) but, because they accelerate at very different rates, they increasingly diverge as a population ages (Figure 1(b)). Another is that individuals attain peak performance early in life, and experience little loss until terminal pathology causes abrupt decline (Figure 1(c)). Either of these limiting cases can equally well account for both the form of observed average trajectories of change and the marked variance between individuals that increases as a population ages. Both indicate a degree variability between individual trajectories that encourages hopes for useful interventions.

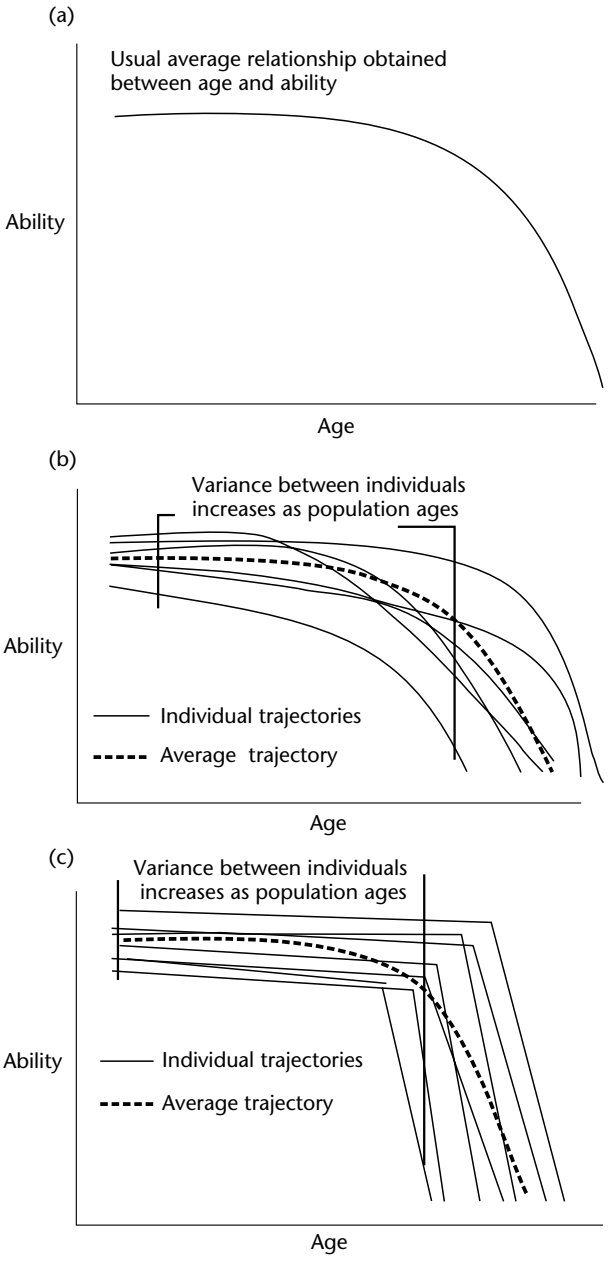


Figure 1. Relationship between age and ability. (a) Average performance levels show steadily accelerating decline after youth; (b) different rates of acceleration lead to increasing divergence in the population; (c) peak performance is maintained until terminal pathology causes abrupt decline.

In practice the best description will depend on the particular population examined. In socioeconomically disadvantaged populations, in which disease or accident sharply and abruptly curtails life expectancy, many individual trajectories of cognitive change will tend to show catastrophic 'terminal drops' rather than continuous declines. In affluent societies in which life expectancy is much longer, and medical interventions postpone death from pathology, gradual declines will become more common. Single cross-sectional comparisons between groups of different ages cannot distinguish true trajectories for individuals from average trajectories for populations. To do this we need longitudinal studies in which the same people are repeatedly assessed over many years.

DO DIFFERENT MENTAL ABILITIES CHANGE AT THE SAME OR AT DIFFERENT RATES?

'Global' Change Models

To relate behavioral to neurophysiological changes we must propose functional models. One limiting case is *global single factor models* which propose that changes in all mental abilities reflect changes in a unique 'master' functional characteristic, most typically 'information processing rate' or 'mental speed' or 'general fluid mental ability'. This would imply that all mental abilities 'age' at similar rates. Another limiting case is *modular aging models*, which are based on evidence that age affects some subsystems in the brain, and so the disparate cognitive abilities that they support, to different extents and at different rates.

If changes in performance on all cognitive tasks directly, and solely, reflect changes in a single, 'master' functional performance characteristic the effects of differences in age should completely disappear when the effects of individual differences in this characteristic have been directly measured and taken into consideration. So, for example, we may test whether age-related declines in 'general fluid intellectual ability' (gf), as assessed by unadjusted

scores on intelligence tests (IT scores), causally determine declines in all other cognitive skills. On nearly all mental tasks individuals' levels of performance correlate negatively with their calendar ages and these correlations are, indeed, abolished or greatly reduced when differences in general mental ability, assessed by unadjusted scores on intelligence tests, are also taken into consideration. Another way to put this is that individual differences in decision speed or IT scores pick up all age-related changes in many simple laboratory tasks.

One explanation might be that IT scores directly reflect some single functional property of the central nervous system that determines levels of performance on all these tasks. A simpler, and more likely, explanation is that intelligence tests predict levels of performance over a very wide range of functionally distinct mental abilities because they include a correspondingly wide range of different problems that, collectively, make demands on all of them.

The same argument cannot dispose of a different theory, that the key performance characteristic on which all mental abilities rely is the maximum rate at which the brain can process information, and that this is directly reflected in the speed with which people can make very simple decisions. It early became apparent that age markedly slows average choice reaction time (CRT) and also that, as tasks become harder, so differences between the average decision times of older and younger groups markedly increase. Subsequent reanalyses of these data suggested that age apparently slows decisions of all kinds, and of all levels of difficulty, by the same proportional amount. That is, on any task, whatever particular demands it makes, decision times for older groups can accurately be estimated by multiplying decision times for younger groups by the same simple constant. This was taken as evidence that 'global slowing of information processing' affects all mental functions, so that changes in information-processing speed are the most sensitive and general indices of cognitive aging. Some authors go much further and suggest that individuals' IT scores, like their performance on all other tasks, directly reflect their maximum information-processing speeds because these, in turn, directly reflect the limiting efficiency of neurophysiological characteristics such as speed of synaptic conduction.

Recent work has questioned both the evidence and the methodological assumptions on which global single factor slowing models are based. It is increasingly recognized that changes in average decision times with age are better described as the consequences of greatly increased moment-to-moment variability in the speed of decisions rather than as the average slowing of all decisions. A direct consequence of this increased moment-to-moment variability is greater average variability from day to day and week to week and that, in both these respects, older people show much greater variability than do the young. When the same individuals are tested on a wide range of tasks, those who show greater moment-to-moment variability on one kind of task are also found to be consistently more variable on others. Findings that age increases variability within individuals at least partly account for findings that, when people are compared on any given occasion, age markedly increases variability between individuals.

There is also evidence that 'global slowing' does not affect all cognitive skills equally. While 'fluid intellectual abilities', indexed by IT scores, decision speed, and rate of learning of novel tasks, all markedly decline with age, language skills, that have become crystallized by practice throughout a lifetime, remain unchanged, or may even slightly improve into the late eighth or even the ninth decade of life. For example, age markedly affects the time taken to respond to simple signals such as lights or tones but has much less effect on speed of decisions about words. Age also does not reduce the speed with which people can carry out highly practised skills such as mental arithmetic. Sustained practice over a lifetime protects skills from age-related decline, and even tasks that have been relatively briefly practiced in the laboratory to the point when they become automatic are then less affected by age.

'Local' or 'Differential' Change Models

Post-mortem and brain imaging studies suggest that age affects frontal cortex more radically and earlier than other parts of the brain in terms of loss of volume, cell loss, reduced cerebral blood flow, and reduction in the concentration, synthesis, and number of receptor sites for dopamine and other neurotransmitters. Consistently, some investigators have found that clinical tests for frontal damage, and laboratory tasks that make demands on functions supported by the frontal and prefrontal cortex, are especially sensitive to age. These functions include the inhibition of unwanted information or of inappropriate responses, the ability rapidly and accurately to generate categories of words, and the ability to switch easily between different criteria for classification of signals. However, there have been inconsistencies of replication

due to problems of measurement, problems of task familiarity, problems of construct validity, and, finally, and probably most basically, much neglected problems of participant selection. On balance, some frontal tasks do seem particularly sensitive to age-related changes, but it is not yet clear whether this is because the incidence of focal brain lesions increases in older populations or because all, or most, people suffer varying degrees of diffuse frontal changes.

MEMORY CHANGES IN OLD AGE

Both older people's subjective complaints and countless laboratory studies confirm that memory efficiency declines in old age and that, in general, the more difficult a memory task is, the greater will be the difference between the average numbers of errors made by young and by elderly adults. This general interaction between age and task difficulty is methodologically inconvenient, because to establish that age affects some functional processes more than others we must show that older people find it especially difficult to cope with a particular *kind* of task demand rather than simply a general increase in task difficulty. Neglect of this point makes brings into question much of the evidence that age affects some particular functional processes in memory, or some 'kinds' of memory, more than it affects others. Consequently, rather than reviewing evidence that age differentially affects speculatively different functional memory processes it may be more helpful to consider the effects of age in terms of the different uses that individuals make of their memories in their everyday lives.

WHAT IS MEMORY USED FOR?

Age changes in efficiency of immediate verbatim memory are slight but consistent, and seem to occur both because age speeds the rates at which memory traces decay and slows the rates at which they can be rehearsed. In everyday life a more common problem is to recover information selectively rather than completely, in some different order from that in which it is presented, or transformed in some way. The ability to meet such demands, particularly in order to schedule decisions and choices, has become known as 'working memory'.

Use of working memory allows people to transcend the limitations of their immediate verbatim recall for novel material ('memory span') by developing and using mnemonics to recode information about items or events. When sufficiently practiced, such recoding techniques seem to allow recall of almost limitless amounts of information but, because old age slows decision speeds, it also makes such recoding less efficient. The extent to which older persons' memories of events are limited by the amounts of information that they can immediately process is illustrated by findings that when older people are asked to recall accounts of actions performed by others, actions they have themselves performed, or actions they have imagined performing, they often correctly recall particular actions but have difficulty remembering whether they have performed, heard about, or imagined them.

Even when they do not deliberately and consciously use mnemonic techniques, people of all ages encode their experiences in terms of their expectations of what is most likely to have occurred. In this sense, video- or tape-recording that passively and unselectively registers information, subject only to failures to register or loss of information, provide poor metaphors for memory which is, rather, a dynamically selective and reconstructive process in which current motivations and lifetime knowledge of the world determine which aspects of new events are attended to, which are ignored, and how previous experience is used to shape interpretation. Because older people can process less information about events in unit time, and lose more of the information that they process, they must increasingly rely on their knowledge of the world to guide economical selection and to reconstruct events from sparse remembered detail. Increased reliance on interpretation and expectation can betray older people when they recreate events that contain unexpected and unfamiliar elements that cannot easily be inferred from context.

This dynamic view of memory highlights the counterintuitive point that humans and other animals typically use their memories to tell them what to do next rather than as archives of past experiences. This use of previously acquired information to anticipate future events and to formulate and execute appropriate plans to cope with them is termed 'prospective memory'. There is evidence that age does impair prospective memory but in everyday life these effects may not be noticeable because most of us, and perhaps especially the elderly, live predictable lives in structured environments that support us by providing contexts that cue us to perform appropriate actions. Such 'environmental support' is especially valuable to the elderly, and may serve them so well that the full

extent of their memory impairment is not apparent until they are deprived of familiar routines and environments.

Once information has been successfully encoded there seems to be no limit on how much of it can be retained, or for how long it remains available. This uncertainty is partly due to the logical difficulty of demonstrating that any information has been permanently lost, rather than having become temporarily inaccessible. Nevertheless, older people's ubiquitous subjective complaints, and many laboratory studies, show that they find it more difficult to access information that they are sure they once knew. For example, when young and older adults are compared on recall of public events that they have both shared, the young do much better. This can be taken to show that the young have forgotten less, but it may also mean that they could encode the events better when they first experienced them so that their memory representations have always been correspondingly more detailed and elaborate.

The complexity of inferences necessary when making such comparisons is illustrated by studies of the relative frequency, and efficiency, with which people of different ages can spontaneously recall events from different times of their lives. Both elderly and young adults recall very recent events more often, and more vividly, than distant events. Apart from losses of information over time a likely explanation is that recent events may be relatively frequently recalled, pondered, and discussed because they tend to have implications for the immediate future. Older people recall relatively fewer events than do the young from intermediate periods of six months or a year previously, but also recall relatively more events from their adolescence and young adult lives than from their middle age. This may partly be because events are better and more elaborately encoded, and so longer remembered, in youth than in middle or in old age, but another factor is that because young adult experiences are often more engaging, and have more long-lasting and memorable consequences than those that occur later in life, they may have since been much more often recalled, reassessed, and discussed. For similar reasons, at whatever age they may occur, vivid experiences during dramatic historical periods, such as wars or marked social change, are better recalled than those in more humdrum epochs. These factors can explain the subjective contrast that older people experience between their vivid recall of remote events and tenuous grasp of the recent past. However, the need for a distinction between older people's subjective impressions of vividness, and the objective completeness and accuracy of their memories, is illustrated when documentation such as school records is available. Older individuals' recall of events in their early lives is often much less reliable than they suppose.

The idea that the accuracy and durability of memories depends on the period of life when they were laid down is supported by recent work showing that individuals who have attained similar standards in university degree examinations forget their hard-won, but seldom used, knowledge more rapidly if they acquired it in early middle age than as young adults.

In contrast to marked age decrements in accuracy of explicit, conscious, recall and recognition of items and events there is evidence that information that is held in memory, but cannot be overtly recalled, can nevertheless influence behavior. For example, people can solve anagrams of words that they have recently learned, but subsequently failed to remember or recognize, more easily than anagrams of words that they have not recently seen, and fail to register as recurring. Loss of conscious, 'explicit' memory, with preservation of unconscious, 'implicit' memory, is found not only in older people but also in patients suffering from amnesias resulting from local damage to the hippocampus and temporal lobes of their brains. This has been taken as evidence that explicit and implicit memories are retained by functionally separate systems that are differentially sensitive both to focal brain damage and to normal aging.

Investigations of the ways in which old age alters mental abilities, particularly memory, have produced extremely detailed descriptions of behavioral changes but, so far, much less information on how these changes relate to changes in the brain and central nervous system. One obvious reason is that our knowledge of relationships between brain and cognition has mainly depended on studies of rare patients with highly localized brain damage and with equally well-defined cognitive impairments. In contrast, old age brings about both diffuse and focal brain changes. Prior to recent advances in brain imaging, diffuse brain changes could not be assessed until death terminated investigations of behavior and post-mortem data became available. Brain imaging now allows detection of global and diffuse, as well as local and specific, brain changes while individuals' cognitive losses can still be investigated. Methodologies for acquisition and interpretation of behavioral data have also greatly improved. It seems likely that the first decade of the twenty-first century will see greater

advances in our understanding of the effects of aging on cognition than have been possible during the entire twentieth century.

Further Reading

Craik FIM and Salthouse T (eds) (1992) *The Handbook of Aging and Cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Kausler DH (1990) *Experimental Psychology, Cognition and Human Aging*. New York: Springer-Verlag.

Rabbitt PMA (2001) Methodology of cognitive gerontology. In: Wixtead J (ed.) *Stevens Handbook of Experimental Psychology*, vol. 4. New York, NY: John Wiley.

Salthouse TA (1991) *Theoretical Perspectives in Cognitive Aging*. Hillsdale, NJ: Lawrence Erlbaum.

Analogical Reasoning, Psychology of

Intermediate article

Dedre Gentner, Northwestern University, Evanston, Illinois, USA

CONTENTS

Analogical reasoning is a kind of reasoning that applies between specific exemplars or cases, in which what is known about one exemplar is used to infer new information about another exemplar. The basic intuition behind analogical reasoning is that when there are substantial parallels across different situations, there are likely to be further parallels.

INTRODUCTION

Analogical thinking is ubiquitous in human cognition. First, analogies are used in explaining new concepts. Domains such as electricity or molecular motion, which cannot directly be perceived, are often taught by analogy to familiar concrete domains such as water flow or billiard-ball collisions. Within cognitive science, mental processes are likened to computer programs (e.g. neural networks; parallel or serial processes), or to searching within a space (e.g. mental distance; close or far associates). Such analogies can then serve as mental models to support reasoning in new domains. Another use of analogy is in making predictions within domains. When the stock market plunged in 2001 after the attack on the World Trade Center, many newswriters made an analogy to the 1929 Wall Street crash, and argued on this basis that the market would be higher after a few years (or that, because key causal conditions are different, the reverse would occur). Analogy is also important in creativity and scientific discovery, as discussed later. Finally, analogy is used in communication and persuasion. For example, environmentalists have compared the earth to Easter Island, where overpopulation and exploitation of the island's once-rich ecology led first to massive loss of species, and then to famine and societal collapse. Such persuasive analogies are meant to invite new inferences: for example, that continued population growth will lead to irreversible ecological decline.

Analogical processing involves several subprocesses. First, given a current topic, *analogical retrieval* is the process of being reminded of a past situation from long-term memory. Once two cases are present in working memory (either because of an analogical retrieval or simply through encountering two cases together), *analogical mapping* can occur. We will begin by discussing the mapping processes.

MAPPING AND USE

History

Important early work on analogical mapping came out of philosophy, notably Hesse's analysis of analogical models in science. Early psychological research on analogy focused on simple four-term analogies of the form a:b::c:d. In the 1970s and 1980s, artificial intelligence researchers introduced a new level of representational complexity and computational specificity. Patrick Winston explored computational algorithms for analogical matching and inference, and Jaime Carbonell modeled the transfer of solution methods from one problem to another. This kind of work inspired psychologists to lay out detailed process models of how analogies are represented and processed. The ensuing period has seen intense computational and psychological research, theory revision, and an expansion of the phenomena studied. The field of analogy continues to be characterized by extremely fruitful interchange between computational models and psychological research. (*See* **Analogy-making, Computational Models of**)

Analogical Mapping

Analogical mapping is the core process in analogy. In a typical instance of analogical mapping, a familiar situation – the *base* or *source* description – is matched with a less familiar situation – the *target* description. The familiar situation suggests ways of viewing the newer situation as well as further inferences about it. Analogical mapping requires *aligning* the two situations – that is, finding the correspondences between the two representations – and *projecting inferences* from the base to the target. Then the reasoner must *evaluate* the analogical match and its inferences. Two further processes that can occur are *re-representation* of one or both analogs to improve the match, and *abstraction* of the structure common to both analogs.

Structure-mapping theory (Gentner, 1983) aims to capture the psychological processes that carry out analogical mapping. According to this theory, the comparison process involves finding an alignment between the base and target representations that reveals common relational structure. On the basis of this alignment, further inferences are projected from base to target. People prefer to find an alignment that is *structurally consistent*: that is, there should be a *one-to-one correspondence* between elements in the base and elements in the target, and the arguments of corresponding predicates must also correspond (*parallel connectivity*). For example, in the analogy below, Timmy in (a) could be put in correspondence with Timmy in (b) (on the basis of a local entity match) or with Fluffy in (b) (on the basis of matching relational roles). People appear to entertain both possibilities during processing, but to settle on one or the other by the end of the process.

(a) Lassie rescued Timmy.
(b) Timmy rescued Fluffy.

Another important early theory was Holyoak's (1985) *pragmatic mapping theory*, which focused on the use of analogy in problem-solving and held that analogical mapping processes are oriented towards attaining goals (such as solutions to problems). According to pragmatic mapping theory, it is goal relevance that determines what is selected in analogy. Holyoak and Thagard (1989) later combined this pragmatic focus with structural factors in their multi-constraint approach to analogy.

Analogical inference projection is a crucial part of the mapping process. Once an alignment is achieved, further inferences can be made by projecting information from the base (or source) domain into the target domain. For example, in the above analogy, suppose we knew more about event (a), such as:

(a) Lassie rescued Timmy because she loves him. She has beatiful brown eyes.
(b) Timmy rescued Fluffy.

In this case, the likely inference in (b) is that Timmy rescued Fluffy because he loves Fluffy. This ability to invite new inferences is central to analogy's role in reasoning. Importantly, analogical inference is rather selective. For example, we are unlikely to make the inference here that Timmy has brown eyes (or that he has four legs, even if we also know this to be true of Lassie).

This illustrates the *selection problem* in theories of analogical inference. If people projected everything known about the base into the target, analogy would be useless in reasoning. Fortunately, people do not do this. Thus a central aim of theories of analogy is to characterize this selection process. At least three factors have been proposed.

Holyoak and his colleagues have emphasized *goal relevance*: the inferences projected are those that fit with the reasoner's current goals in problem-solving.

A second factor, proposed in structure-mapping theory, is relational connectivity – more specifically, *systematicity*: a preference for projecting from matching systems of relations connected by higher-order relations such as *cause*, rather than projecting local matches. In many cases, goal relevance and systematicity will make the same predictions, because problem-solving goals often involve a focus on causal systems.

A third factor in selecting inferences, proposed by Keane, is *adaptability*: the ease with which a possible inference can be modified to fit the target.

There is evidence for all three of these criteria. Spellman and Holyoak (1996) showed that when two possible mappings are available for a given analogy, people will select the mapping whose inferences are relevant to their goals. Evidence for systematicity comes from the finding that when people read analogous passages and make inferences from one to the other, they are more likely to import a fact from the base to the target when it is causally connected to other matching predicates (Clement and Gentner, 1991; Markman, 1997). Finally, Keane (1996) found evidence that the degree of adaptability predicts which inferences are made from an analogy.

There remain many open questions. For example, according to the structure-mapping account, many different higher-order relations can provide inferential selection – including causal relations, deontic

relations such as permission and obligation, and spatial relations such as symmetry and transitive increase. The challenge then is to delineate the set of higher-order relations that can serve this purpose. Another open question is the time course of these constraints. For example, do goals have special priority *during* the analogical mapping process, or do the effects of goals occur through influencing the initial representations of the two analogs (*before* the mapping process) or through selecting among multiple possible interpretations (*after* the mapping process)?

Evaluation

Once the common alignment and the candidate inferences have been discovered, the analogy is evaluated. *Evaluating* an analogy involves at least three kinds of judgment: (1) *structural soundness*: whether the alignment and the projected inferences are structurally consistent; (2) *factual correctness*: whether the projected inferences are false, true, or indeterminate in the target; and (3) *relevance*: whether the analogical inferences are relevant to the current goals. In practice, the relative importance of these factors varies quite a bit. In domains where little is known, or where there is disagreement about the facts – for example, in politics – goal relevance may be more important than factual correctness.

Abstraction

In *analogical abstraction*, the common system that represents the interpretation of an analogy is extracted and stored. This kind of schema abstraction helps to promote transfer to new exemplars. When people are asked to compare two analogous passages, they are better able to later retrieve and use their common structure (given a relationally similar probe) than are people who were given only one of the stories (Gick and Holyoak, 1983). Further studies have shown that actively comparing two analogous passages leads to better subsequent retrieval than reading the two passages separately. These findings are consistent with the claim that analogical alignment promotes the common structure and makes it more available for later use.

Analogy in Real-world Reasoning

Analogy is often used in common-sense reasoning to provide plausible inferences. It must be noted that analogy is not a deductive process. There is no guarantee that the inferences from a given analogy will be true in the target, even if the analogy is carried out perfectly and all of the relevant statements are true in the base. However, the set of implicit constraints described above make analogy a relatively 'tight' form of inductive reasoning. This may be why analogy is heavily used in arenas such as law, where clear reasoning is important but formal principles are often not sufficient to decide issues.

The lack of deductive certainty in analogical reasoning has a positive side. It means that analogy can suggest genuinely new hypotheses, whose truth could not be deduced from current knowledge. One arena in which this kind of analogical inferencing has been extensively studied is scientific reasoning and discovery. Nancy Nersessian has examined the role of analogy and other model-based reasoning processes in the thought processes of Faraday and Maxwell. Paul Thagard has discussed analogy as a contributor to conceptual revolutions in science. Kevin Dunbar has observed scientists in working microbiology laboratories and has found that analogy plays a large role in the discovery process.

Analogy appears to be very important in children's thinking, as Halford, Goswami, and others have argued. Children often use analogies from known domains as a way to fill in gaps in their knowledge of other domains. For example, Inagaki and Hatano (1987) asked five-year-old children hypothetical questions like 'What would happen if a rabbit were continually given more water?' The children often answered by using an analogy to humans: for example, 'I would get sick if I kept drinking water, and I think the rabbit would too.' Interestingly, children's answers tended to be more accurate when they used such analogies than when they did not. Children were most likely to use analogies to humans when the target was somewhat similar to humans, suggesting that for children (as for adults) similar analogs are more likely to be retrieved and are easier to align with the target than dissimilar analogs.

FACTORS THAT INFLUENCE ANALOGICAL MAPPING AND USE

People's fluency in carrying out analogical mappings is influenced by three broad kinds of factors. First are factors internal to the analogical mapping itself, such as *systematicity* – whether the common relational system possesses higher-order connective structure – and *transparency* – the degree to which corresponding elements are similar. The second category includes characteristics of the

reasoner, such as age and expertise. The third includes task factors such as processing load, time pressure, and context.

Transparency and systematicity have been found to be important in analogical problem-solving. The transparency of the mapping between two analogous algebra problems – that is, the similarity between corresponding objects – is a good predictor of people's ability to notice and apply solutions from one problem to the other. For example, Ross (1989) taught people algebra problems and later gave them new problems that followed the same principles. People were better able to map the solution from a prior problem to a current problem when the corresponding objects were highly similar between the two problems: for example, 'How many golf balls per golfer' → 'How many tennis balls per tennis player'. They performed worst in the *cross-mapped* condition, in which similar objects appeared in different roles across the two problems: for example, 'How many golf balls per golfer' → 'How many tennis players per tennis ball'.

The intrinsic factors of transparency and systematicity interact with characteristics of the reasoner, notably age and experience. Gentner and Toupin (1986) gave children a simple story and asked them to re-enact the story with new characters. Both six- and nine-year-olds performed far better when the corresponding characters were highly similar between the two stories than when they were different, and they performed worst when similar characters played different roles across the two stories (the *cross-mapped* condition). Thus both age groups were sensitive to the transparency of the correspondences. In addition, older children (but not younger children) benefited from systematicity – that is, from hearing a summary statement that provided an overarching social or causal moral.

The developmental change found here is an instance of the *relational shift*: a shift from focusing on object matches to focusing on relational matches. Some researchers have suggested that this shift is driven by gains in knowledge (Gentner and Rattermann, 1991), while others propose that it results from a developmental increase in processing capacity (Halford, 1993).

The third class of factors affecting analogical processing concerns task variables such as time pressure, processing load, and immediate context. One generalization that emerges from several studies is that making relational matches requires more time and processing resources than making object–attribute matches. For example, Goldstone and Medin (1994) found that when people are forced to terminate processing early, they are strongly influenced by local attribute matches (such as *A* with *A* in the example below), even in cases where they would choose a relational match (such as *A* with *P*) if given sufficient time:

A above M and *P above A*

Adult performance in mapping tasks is also influenced by immediately preceding experiences. For example, in the *one-shot mapping task* (Markman and Gentner, 1993) subjects are shown a pair of cross-mapped pictures, such as *a robot repairing a car* and *a man repairing a robot*. The experimenter points to the robot in the first picture and the subject indicates which object in the second picture 'goes with' it. Subjects often choose the object match (e.g. the other robot). However, if they have previously rated the similarity of the pair, they are likely to choose the relational correspondence (the repairman). These findings suggest that carrying out a similarity comparison induces a structural alignment.

Kubose, Holyoak, and Hummel used this one-shot mapping task to show that processing load influences analogical processing. The experimenter pointed to the cross-mapped object in the first picture (the robot) and subjects were instructed to point to the relational correspondence (the repairman) in the second picture. Subjects made more object-mapping errors when given an extra processing load, such as having to count backwards. Recent work by Holyoak and his colleagues also suggests that damage to the prefrontal cortex is associated with detriments in analogical tasks, although it is not clear whether this results from specific involvement of the prefrontal cortex in analogical processing or from more general factors such as inhibitory control.

Summary

Research on analogical mapping has revealed a set of basic phenomena that characterize human analogical processing (see Table 1). A striking feature of analogical mapping is the importance of systematic, structurally connected representations. Commonalities that are interconnected into a relational system are considered to be more central to a comparison than are those that are not. Connected systems are easier to map to a new domain than are unconnected sets, leading to better transfer in analogy and problem-solving. Systematicity also governs inferences: inferences are projected from interconnected systems in the base to fill out corresponding structure in the target. Even the

Table 1. Basic phenomena of analogy (adapted from Gentner and Markman, 1995, 1997; see also Hummel and Holyoak, 1997)

1 *Relational similarity*	Analogies involve relational commonalities; object commonalities are optional.
2 *Structural consistency*	Analogical mapping involves one-to-one correspondence and parallel connectivity.
3 *Systematicity*	In interpreting analogy, connected systems of relations are preferred over isolated relations.
4 *Candidate inferences*	Analogical inferences are generated via structural completion.
5 *Alignable differences*	Differences that are connected to the common system are rendered more salient by a comparison.
6 *Interactive mapping*	Analogy interpretation depends on both terms. The same term yields different interpretations in different pairings.
7 *Multiple interpretations*	Analogy allows multiple interpretations of a single comparison.
8 *Cross-mapping*	Analogies are more difficult to process when there are competing object matches.

differences associated with a similarity comparison are influenced by systematicity: the differences that are psychologically salient in a comparison are those that are connected to the common system. In addition, goal relevance may have effects over and above the effects of connected relational structure.

RETRIEVAL OF ANALOGS

So far, we have discussed the processing of an analogy once both analogs are present. When we turn to the issue of what leads people to think of analogies, we see a very different pattern of results. People often fail to retrieve potentially useful analogs, even when there is an excellent structural match, and even when they clearly have retained the material in memory. For example, Gick and Holyoak (1983) gave subjects a thought problem: how to cure an inoperable tumor without using a strong beam of radiation that would kill the surrounding flesh. Only about 10 percent of the participants came up with the ideal solution, which is to converge on the tumor with several weak beams of radiation. If given a prior analogous story in which soldiers converged on a fort, three times as many people (about 30 percent) produced convergence solutions. Yet the majority of participants still failed to think of the convergence solution. Surprisingly, when these people were simply told to think back to the story they had heard, the percentage of convergence solutions again tripled, to 80–90 percent. Thus, the fortress story had been retained in memory, but it was not retrieved by the analogous tumor problem. The implication is clear. Even when a prior experience has been successfully stored in memory, it might not be retrieved when a person encounters a new analogous situation where it would be useful.

When we ask what does facilitate analogical retrieval, one major factor emerges: the similarity between the analogs. As noted earlier, similarity is one of the factors that facilitates analogical mapping; but it has a much larger effect on retrieval. For example, Gentner *et al*. (1993) gave subjects a set of stories to remember and later showed them probe stories that were either surface-similar to their memory item (e.g. similar objects and characters) or structurally similar (i.e. analogous, with similar higher-order causal structure). Surface similarity was the best predictor of whether people would be reminded of the prior stories; people were two to five times more likely to retrieve prior stories with only surface commonalities than with only structural commonalities. However, their judgments of the goodness of the match were completely different. They rated the surface-similar pairs (including their own remindings) as low in inferential value and in similarity, and preferred the structurally similar pairs. A similar dissociation between reminding and use has been found in problem-solving tasks: remindings of prior problems are strongly influenced by surface similarity, even though structural similarity better predicts success in solving current problems (Ross, 1989). This failure to access potentially useful analogs (unless they are highly similar to the target) is an instance of the *inert knowledge* problem in education. One piece of good news is that it appears that domain expertise may improve matters somewhat. For example, Novick (1988) found that people with mathematics training retrieved fewer surface-similar lures in a problem-solving task than did novice mathematicians. Moreover, experts were quicker to reject these false matches than were novices.

One factor that may contribute to experts' success in analogical retrieval is *representational uniformity*: the extent to which the relations in the memory trace are represented similarly to those in the probe. Clement *et al*. (1994) explored the effect of relational predicate similarity on analogical access and mapping between stories. They found that retrieval was more likely when the probe and target

contained synonymous terms (*manifest* similarity) than when they contained loosely similar predicates such as 'munched' and 'consumed' (*latent* similarity). However, unlike retrieval, analogical mapping when both situations were present was relatively unaffected by the latent–manifest distinction. In analogical reminding, with only the current situation in working memory, success depends on the degree of match of the pre-existing representations; whereas during mapping, with both situations present in working memory, there is opportunity for re-representation.

ANALOGICAL LEARNING

Analogical comparison can lead to new learning in at least four ways: analogical abstraction, inference projection, difference detection, and re-representation. The first two we have already discussed. In *analogical abstraction*, the structure common to base and target is noticed and extracted. Sometimes the common system is stored as a separate representation; this is referred to as *schema abstraction* (Gick and Holyoak, 1983). In *inference projection*, a proposition from the base is mapped to the target. If it is retained as part of the target structure, then learning has occurred. In *difference detection*, carrying out a comparison process leads people to notice certain differences – namely, those connected to the common structure. In *re-representation*, two non-identical predicates are aligned and decomposed (or abstracted) to find their commonalities, resulting in a re-representation that contains a common predicate: for example, comparing *chase*(dog, cat) and *follow*(detective, suspect) might result in *pursue* (entity1, entity2). A further kind of knowledge change, hypothesized to take place in scientific discovery, is *restructuring*, in which the target undergoes a radical change in structure.

COMPUTATIONAL MODELS

The interplay between computational models and psychological studies has been extremely productive in analogical research. Current models include Boicho Kokinov's AMBR model, which integrates retrieval and mapping; Keane's IAM model, which utilizes an incremental mapping algorithm; Halford's tensor product model; and the systems of Doug Hofstadter and his colleagues Melanie Mitchell and Robert French, which integrate perceptual processing with analogical matching. (*See* **Analogy-making, Computational Models of**)

Analogical modeling has made great strides, but there are still open questions. At present no model fully captures human analogy processing. Two challenges for analogical models are (1) the *selection problem* discussed above – namely, how to avoid indiscriminate inferencing; and (2) the problem of *representational flexibility* – that is, how to achieve a matching process that does not require absolute identity matches. Falkenhainer *et al.*'s (1989) SME, which uses a local-to-global alignment and inference process over structured symbolic representations, meets the benchmarks in Table 1 and can capture selective matching and inference, as well as schema abstraction. But it is not yet sufficiently flexible in its match process. Another leading model, Hummel and Holyoak's (1997) LISA model, uses a combination of distributed representations of concepts and structured representations of the relational connections (necessary for achieving structural consistency in mapping). It uses a connectionist temporal-binding algorithm and makes its matches in a serial order partly guided by the experimenter. LISA's use of distributed representations allows for flexible matching, and unlike most models of analogy, it attempts to capture working memory limitations. However, it has yet to solve the selectivity problem in inferencing.

THE FUTURE

Of the many research questions that remain, four stand out as particularly interesting and timely. First is the role of analogy in cognitive development: how much of children's rapid learning can be attributed to the processes of comparing and drawing inferences between partially similar situations? Second is tracing the neuropsychology of analogical processes: what areas of the brain are implicated, and what is the course of processing? Third is the exploration of analogy in animal cognition. Comparative research so far indicates that humans excel in analogical ability, yet this ability exists in certain other species as well – for example, in chimpanzees and dolphins. Cross-species comparisons may help us decompose the cognitive components of analogical ability. A final important research frontier is the integration of analogy into larger models of cognition.

References

Clement CA and Gentner D (1991) Systematicity as a selection constraint in analogical mapping. *Cognitive Science* **15**: 89–132.

Clement CA, Mawby R and Giles DE (1994) The effects of manifest relational similarity on analog retrieval. *Journal of Memory and Language* **33**: 396–420.

Falkenhainer B, Forbus KD and Gentner D (1989) The structure-mapping engine: algorithm and examples. *Artificial Intelligence* **41**: 1–63.
Gentner D (1983) Structure-mapping: a theoretical framework for analogy. *Cognitive Science* **7**(2): 155–170.
Gentner D and Markman AB (1995) Analogy-based reasoning in connectionism. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks*, pp. 91–93. Cambridge, MA: MIT Press.
Gentner D and Markman AB (1997) Structure-mapping in analogy and similarity. *American Psychologist* **52**: 45–56.
Gentner D and Rattermann MJ (1991) Language and the career of similarity. In: Gelman SA and Byrnes JP (eds) *Perspectives on Thought and Language: Interrelations in Development*, pp. 225–277. London, UK: Cambridge University Press.
Gentner D, Rattermann MJ and Forbus KD (1993) The roles of similarity in transfer: separating retrievability from inferential soundness. *Cognitive Psychology* **25**: 524–575.
Gentner D and Toupin C (1986) Systematicity and surface similarity in the development of analogy. *Cognitive Science* **10**: 277–300.
Gick ML and Holyoak KJ (1983) Schema induction and analogical transfer. *Cognitive Psychology* **15**(1): 1–38.
Goldstone RL and Medin DL (1994) Time course of comparison. *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**(1): 29–50.
Halford GS (1993) *Children's Understanding: The Development of Mental Models*. Hillsdale, NJ: Lawrence Erlbaum.
Holyoak KJ (1985) The pragmatics of analogical transfer. In: Bower GH (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 19, pp. 59–87. New York, NY: Academic Press.
Holyoak KJ and Thagard PR (1989) Analogical mapping by constraint satisfaction. *Cognitive Science* **13**(3): 295–355.
Hummel JE and Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* **104**(3): 427–466.
Inagaki K and Hatano G (1987) Young children's spontaneous personification as analogy. *Child Development* **58**: 1013–1020.
Keane MT (1996) On adaptation in analogy: tests of pragmatic importance and adaptability in analogical problem solving. *Quarterly Journal of Experimental Psychology* **49**: 1062–1085.
Markman AB (1997) Constraints on analogical inference. *Cognitive Science* **21**(4): 373–418.
Markman AB and Gentner D (1993) Structural alignment during similarity comparisons. *Cognitive Psychology* **25**: 431–467.
Novick LR (1988) Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition* **14**: 510–520.
Ross BH (1989) Distinguishing types of superficial similarities: different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory and Cognition* **15**: 456–468.
Spellman BA and Holyoak KJ (1996) Pragmatics in analogical mapping. *Cognitive Psychology* **31**: 307–346.

Further Reading

Carbonell JG (1983) Learning by analogy: formulating and generalizing plans from past experience. In: Michalski RS, Carbonell JG and Mitchell TM (eds) *Machine Learning: An Artificial Intelligence Approach*, vol. 1, pp. 137–161. Palo Alto, CA: Tioga.
Dunbar K (1995) How scientists really reason: scientific reasoning in real-world laboratories. In: Sternberg RJ and Davidson JE (eds) *The Nature of Insight*, pp. 365–395. Cambridge, MA: MIT Press.
French R (1995) *The Subtlety of Sameness: A Theory and Computer Model of Analogy-making*. Cambridge, MA: MIT Press.
Gentner D, Holyoak KJ and Kokinov BN (eds) (2001) *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.
Goswami U (1992) *Analogical Reasoning in Children*. Hillsdale, NJ: Lawrence Erlbaum.
Halford GS, Wilson WH and Phillips S (1998) Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences* **21**: 803–864.
Hesse MB (1966) *Models and Analogies in Science*. Notre Dame, IN: University of Notre Dame Press.
Hofstadter DR and Mitchell M (1994) The Copycat project: a model of mental fluidity and analogy-making. In: Holyoak KJ and Barnden JA (eds) *Advances in Connectionist and Neural Computation Theory*, vol. 2: *Analogical Connections*, pp. 31–112. Norwood, NJ: Ablex.
Holyoak KJ and Thagard PR (1995) *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.
Keane MT (1990) Incremental analogising: theory and model. In: Gilhooly KJ, Keane MTG, Logie RH and Erdos G (eds) *Lines of Thinking*, vol. 1. Chichester, UK: John Wiley.
Kokinov BN and Petrov AA (2001) Integrating memory and reasoning in analogy-making: the AMBR model. In: Gentner D, Holyoak KJ and Kokinov BN (eds) *The Analogical Mind: Perspectives from Cognitive Science*, pp. 161–196. Cambridge, MA: MIT Press.
Nersessian NJ (1984) *Faraday to Einstein: Constructing Meaning in Scientific Theories*. Dordrecht, Netherlands: Nijhoff.
Thagard P (1992) *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.
Winston PH (1982) Learning new principles from precedents and exercises. *Artificial Intelligence* **19**: 321–350.

Animal Cognition

Introductory article

Valerie A Kuhlmeier, Yale University, New Haven, Connecticut, USA
Sarah T Boysen, Ohio State University, Columbus, Ohio, USA

CONTENTS

The study of how animals learn, behave, and think, often from a comparative perspective, is the crux of animal cognition research. Recent topics explored in this field include understanding of abstract concepts, spatial learning and memory, imitation, representation of social relations, and examining the similarities between nonhuman primates and human children.

INTRODUCTION

The field of animal cognition has emerged since the 1980s as a rich, interdisciplinary area representing animal learning and behavior, comparative psychology, ethology, and behavioral ecology. It is different from the behaviorist tradition in that researchers do not focus primarily on conditioning processes. Interest in animal cognition cuts across numerous academic fields, including psychology, zoology, biology, anthropology, and primatology, among others. In psychology, several topic areas have been particularly important, although they represent only a portion of ongoing studies in animal cognition. In particular, studies of abstract concepts, spatial learning and memory, imitation, representation of social relationships, and developmental comparisons between human children and animals (typically primates) represent a growing literature. These areas encompass research issues and approaches that are currently being studied with different animal species, including (but not limited to) rats, pigeons, primate species such as rhesus monkeys, capuchins, and squirrel monkeys, as well as great apes such as chimpanzees, bonobos, and orangutans.

ABSTRACT CONCEPTS

One of the most basic questions being investigated with animals is whether any species is able to reason abstractly. That is, can an animal understand an abstract concept such as 'same versus different', recognize itself in a mirror (possess a 'concept of self'), or even count using numbers through a process similar to numerical skills shown by young children? Are they able to use their understanding of a concept when presented with new examples? Finally, and most importantly, how might such capacities for abstract thinking and concept formation serve a particular species in the wild when they are confronted with real problem-solving opportunities and/or life-and-death situations?

What is meant by an abstract concept, particularly for an animal, requires careful thought, and has been a contentious issue in animal learning and cognition. If concepts are defined in such a way that only humans can acquire them, as some investigators have proposed, and especially if conceptual understanding can be acquired only if mediated by language, then clearly concepts must be unique to *Homo sapiens*. However, investigators in animal cognition have suggested that an abstract concept should be operationally defined. For example, if an animal is presented with complex stimuli such as photographs of trees and other scenes, learns to respond only to pictures of trees, and subsequently chooses new pictures of trees from among a novel set of photographs, the animal may be said to have an understanding of the 'tree concept'. Similarly, if an animal with training on counting and other number skills is able to correctly assign a number to a novel collection of objects, despite learning to count using only gumdrops, it could be argued that the subject, whether it was a chimpanzee, rat, or pigeon, has a concept of number. In this case, the ability to generalize an understanding of numbers to brand-new counting opportunities would suggest that the animal has

some representation of an abstract concept of number and numerical relationships.

According to the current literature in animal cognition, both types of capacities for abstract concepts have been demonstrated in animal subjects. Such studies include pigeons' abilities to recognize a range of items conceptually, demonstrated by identifying slides depicting trees, people, or even the letter 'a' presented in very different typefaces, when the animals were presented with novel sets of visual stimuli. In the case of number concepts, Sarah Boysen has demonstrated that chimpanzees can assign Arabic numerals to quantities of candies and other collections of objects. More impressively, individual chimpanzees have been shown to invoke spontaneous addition algorithms that enabled them to count different-sized arrays that were hidden in several locations around a test room. The animals were able to report the total number of objects found during their search of the sites when they were given the very first opportunity with this novel task, even though they had no prior training or testing with items separated in both time and space.

In this case, previous associative training with numerals and candy arrays most probably contributed to emergent conceptual skills with numbers and counting which went far beyond the animals' specific training. Thus, this study of rudimentary addition in chimpanzees supported the notion that the subjects had acquired a 'concept of number' and could use it spontaneously with completely novel problems and in a new setting. Such cognitive flexibility and generalizability of skills, above and beyond explicit training, provides evidence for animals' capacities for conceptual understanding.

SPATIAL LEARNING AND MEMORY

As early as the 1930s, an experimental psychologist, Edward Tolman, studied the ability of rats to learn the spatial organization of a maze. As the animals explored the maze, they learned the various turns, blind alleys, and eventually, the location of the goal box. Tolman proposed that, during the course of their exploration, the rats came to represent the spatial configuration of the maze internally as a kind of 'cognitive map'. The term is still used today to refer to a mental representation of territory and the landmarks within it, sites where food has been cached, and a host of other spatially mediated information which may contribute to survival. Since Tolman proposed the idea of a cognitive representation of the external environment within the rat, scientists from numerous disciplines, including comparative psychology, neuroscience, cognitive science, biology, and ethology, have designed experiments to test ideas about such mental representations in a wide range of species, including many types of passerine birds and rodents, nonhuman primates, as well as comparative studies with children and adult humans from differing cultures.

Studies of spatial learning and memory are providing a wealth of clues towards understanding basic mechanisms related to the processing of spatial information. Most animals must be able to learn and remember information about their location within their territory in order to forage for food, seek mates, migrate, select nest or den sites, care for offspring, store food (cache), and avoid predators. For example, hoarding food and relocating it later, sometimes months after the original caching, places exceptional demands on spatial learning and memory. Among the caching birds, several species of jays and nutcrackers store large numbers of nuts and seeds, and are dependent upon finding stored food in order to survive the winter months. During experimental laboratory studies, efforts to control possible odor and visual cues did little to decrease the birds' accuracy in locating cached food. However, displacement of the cached food to a site a short distance away resulted in the birds' inability to locate it. Furthermore, while there have been some suggested alternative explanations for the birds' success, including mere chance encountering of the sites, their ability to locate and retrieve hoarded food has repeatedly been far better than would have been predicted by chance alone. Such findings have been replicated in numerous laboratories, with several different species, with the same results.

Another factor influencing spatial learning and memory is a species' particular social structure and mating system. For example, males and females of a monogamous species (i.e. those in which males and females mate exclusively) live in the same territory together. Consequently, both sexes use similar spatial abilities for getting around their range. On the other hand, in species that are polygynous (i.e. those in which males have access to more than one female), males have additional spatial demands: they must keep track of females' locations within the territory and avoid rival males. These critical behaviors might translate to better spatial processing abilities in polygynous males relative to females of polygynous species and to males and females of monogamous species.

This hypothesis has been explored in studies of sex differences in neural structure and spatial cognition in different rodents, specifically prairie voles, a monogamous species, and meadow voles, which are polygynous. Males and females of these two closely related species were compared on their performance on a series of mazes, in order to test their spatial abilities. As predicted, polygynous males outperformed females and monogamous males. Studies such as these demonstrate how particular spatial abilities may develop in a species and subsequently help to explain observed differences within and between species.

It is likely that the mechanisms and processes that subserve spatial learning and memory in animals will continue to be an active area of research, as a host of critical and intriguing questions remain about how animals acquire and utilize spatial information. For example, new research in animal spatial cognition is examining what types of cues in the environment (e.g. landmarks, configurations of landmarks) animals attend to, learn, and remember in order to effectively navigate and orient. Researchers are even exploring animals' understanding of symbols of their environment, such as maps and scale models. The research continues to be comparative in nature, looking for similarities and differences among animal species and between humans and other animals. Indeed, the study of animal spatial learning and memory promises to have an exciting future in the field of animal cognition.

REPRESENTATION OF SOCIAL RELATIONS

Animals that live in long-lasting social groups interact with the individuals of their group in a complex manner. Required for these interactions are cognitive abilities that may include recognizing others as individuals, remembering which individuals have aided one in the past, and monitoring interactions among the other members of the group. Indeed, some researchers in the 1960s and 1970s proposed that the evolution of general problem-solving abilities was driven by the need for complex cognitive mechanisms for social interactions. But what exactly do animals understand and represent of their social world? Do animals recognize and interact with others in certain ways only because of past experiences with these individuals (i.e. by making associations), or can animals also distinguish classes of relationships and understand concepts such as kinship and dominance (by having mental representations)?

Although the social lives of many animals with long-lasting social groups have been examined (e.g. wolves, lions, elephants, marine mammals, birds), most research in the representation of social relations has focused on primate species. For example, Verena Dasser tested whether Java monkeys can discriminate pairs of animals based on their familial relationship. One monkey subject was trained to choose a photograph of a mother and daughter from the subject's social group and ignore a photograph of an unrelated pair. After training trials using this same mother–daughter pair, the monkey subject was able to correctly choose new photographs of other mother–offspring pairs from the social group. How did the monkey do this? It is possible that the monkey relied on the perceptual similarity between the mother and her offspring to represent the relationship between the members of the pair and solved the task by choosing the pair of animals that looked alike. Results from a recent study by Lisa Parr and Frans deWaal support this possibility. They tested chimpanzees with a similar task; however, their subjects had to match photographs of unfamiliar chimpanzee mothers with their equally unfamiliar offspring, while ignoring a third photograph of an unrelated chimpanzee. The chimpanzees were successful; they correctly selected the photographs of the sons of the sample mothers. Since the subjects had no past experience with the pictured animals, successful matching implied a recognition of physical similarity between individuals.

Dorothy Cheney and Robert Seyfarth have used vocal playback experiments to study how monkeys represent their social world. In one experiment, when an infant vervet monkey's cries were played out of hidden speakers, the other monkeys in the group looked to the direction of the infants' mother. Their behavior indicated that they associate particular infants with the infants' mothers. But can monkeys go beyond simple associations like these and demonstrate the existence of mental representations of their social world?

A second vocal playback experiment with baboons suggested that this might indeed be the case. The experimenters took advantage of natural vocal exchanges between baboons at different levels of the dominance hierarchy. Normally, when a dominant female approaches a subordinate female, the dominant monkey will grunt and the subordinate monkey will fear bark. However, when a subordinate female approaches a dominant female, there is no vocal exchange. The experimental procedure relied on the fact that monkeys tend to look longer at sources of unfamiliar or

strange sounds. The experimenters played sequences of consistent and inconsistent grunt/fear bark exchanges. In the inconsistent exchange, the sound of a subordinate female's grunt was played, followed by a dominant female's fear bark, an unlikely event. However, in the consistent exchange, a subordinate female's grunt was followed by a high-ranking female's fear bark and a higher-ranking female's grunt. This sequence is consistent with natural baboon vocal behavior due to the last part of the sequence; the high-ranking animal's fear bark could have been caused by the higher-ranking female's approach. The baboons looked longer in the direction of the hidden speaker that played the inconsistent sequence, indicating that they recognized the strangeness of the event and had some representation of their dominance hierarchy and the appropriate behavior of the members within it. This understanding of social causation suggests a mental representation of their social environment.

Thus, research has demonstrated that some social animals may interact differentially with the individuals of their group with the aid of associations such as physical similarity and past experience. Furthermore, they might also develop mental representations of the underlying social relations based on these associations. It is possible, then, that the learning of associations between individuals and other individuals, or even associations between individuals and certain behaviors, may help give rise to mental representations such as kinship and dominance status.

IMITATION

Since the time of Darwin, there has been an interest in whether animals can imitate, that is, perform an action after seeing the same action being performed by another animal. Darwin concluded that this ability was one we share with other animals; however, researchers today are not so quick to grant animals this ability. Indeed, many terms other than 'imitation' have been coined to describe behavior that may appear to be truly imitative, but that actually occurs through much simpler cognitive processes. But why the careful attribution of imitation? Many researchers believe true imitation to be a precursor to theory of mind (the ability to attribute mental states to others), and thus involves processes such as self-awareness and perspective-taking. Thus, to grant that a behavior is imitative is to grant the actor the precursors to highly complex cognitive capacities.

For most researchers, demonstrating true imitative behavior typically requires that an animal has a representation of another animal's action and uses the representation to behave in a manner that matches that of the other. Observations of animals doing human-like actions (e.g. a pet opening the back door) are not enough to demonstrate true imitation. Without knowing past training experience, it is difficult to determine how the behavior developed. Also, it may be that an animal is not attending to the actions of another, but attending to the object or location the other animal is acting on. For example, when animal A behaves similarly to nearby animal B (perhaps by digging for food in a certain area), it may be the result of A's attention being drawn to the location of the food, not a direct imitation of B's action. This type of behavior has been called 'stimulus enhancement' or 'local enhancement', and it is often difficult to separate this type of explanation from one implicating true imitation.

One promising paradigm has been the 'two-action task', in which one animal watches the actions (usually a trained, unusual behavior) of another animal and is then placed in a similar situation to examine whether it completes the same actions. For example, Thomas Zentall and his colleagues have trained pigeons and Japanese quail to either step or peck on a lever to receive food. Then, untrained birds were allowed to observe the trained demonstrator as it pecked or stepped on the lever. After the observation period, the untrained observer was placed in the cage with the lever and its behavior was measured. These birds displayed almost 90 percent imitative responses during testing. That is, if they had observed the demonstrator bird step on the lever, they too stepped, but if they observed pecking, they pecked. One criticism that has been raised is that the pecking and stepping behaviors themselves were not necessarily unusual for the birds in natural contexts, and thus, these experiments may not have demonstrated true imitation, but stimulus enhancement to the lever. However, despite this criticism, the results are very suggestive of imitative behavior and point to the value of the two-action experimental design for examining animal imitation.

The study of nonhuman primate imitative ability has consisted of many suggestive anecdotes, but recently, controlled experiments have also been conducted. Many of these have incorporated object manipulation and tool-using, taking advantage of the animals' dexterity and, in some cases, their

natural tool-using abilities. For example, Andrew Whiten and his colleagues developed a two-action task for both chimpanzees and young children. The chimpanzee and human subjects observed an adult human either twist or poke out bolts that secured the opening of a box containing a desired fruit. When given access to the box, all subjects used the action demonstrated, suggesting imitation. However, the chimpanzees exhibited fewer instances of imitation than the oldest children tested (four years old). Indeed, the children seemed to imitate the fine motor details of the demonstrator more closely than the chimpanzees.

Similar observations have been made by Michael Tomasello and his colleagues. In their study, chimpanzees tended to achieve the same goal as a demonstrator (e.g. use a rake to attain food), but the techniques the chimpanzees used differed slightly from those of the demonstrators. Tetsuro Matsuzawa and Masako Myowa-Yamakoshi have interpreted these results and results from their own similar experiments to suggest that apes may have difficulty in transforming the demonstrator's motions into their own matching motor acts. That is, in these imitation experiments, the apes may be less sensitive to the demonstrator's body movements than to his or her underlying goals and the objects used to attain them.

In summary, new methodologies have brought us closer to understanding the extent of animals' imitative capacities. The results so far are highly suggestive of true imitative behavior, yet many skeptics are still not convinced. The two-action task has proven to be an effective method of testing, although experimenters will have to address concerns that performance is due to stimulus enhancement and not true imitation. Furthermore, the creative studies with nonhuman primates, especially those simultaneously testing young human children, are valuable and should be explored further to determine possible species differences and similarities.

COMPARISONS BETWEEN NONHUMAN PRIMATES AND HUMAN CHILDREN

Animal cognition research is often comparative, and very often the comparisons are made between humans, specifically infants and children, and nonhuman primates. These specific comparisons are often made to examine the cognitive capacities that can exist in the absence of language, to examine the possibility of innate modules for certain cognitive or perceptual processes, and to chart the evolution of our cognitive capacities and examine conditions that may have supported their development. Research in animal cognition and human cognitive development has often delved into the same questions, yet direct comparisons on a given topic have sometimes been hard to make owing to different experimental procedures. Of late, however, there has been an increase in experimental studies that directly compare children and nonhuman primates by using very similar methodologies. Some of these studies have been discussed above, including the imitation studies by Andrew Whiten and his colleagues. Two more examples, studies of numerical ability and scale model comprehension, will be mentioned here.

Karen Wynn has used the 'violation of expectation' procedure to examine the numerical abilities of infants. This procedure relies on the fact that infants tend to look longer at events that are unexpected, or involve some sort of violation. In one study, infants watched as an object was placed on a stage. A screen was then raised to hide the object. The infants then saw a second object being placed behind the screen. Now, the screen was lowered, revealing one of two outcomes: one object (the unexpected outcome) or two objects (the expected outcome). Infants looked longer at the unexpected outcome of one object. In fact, in subsequent studies, they looked longer at a $1 + 1 = 3$ event than at $1 + 1 = 2$. This effect was also seen for simple subtraction events, and together, the studies suggest that infants have some understanding of number estimation and quantity. Using similar experimental procedures, Marc Hauser and his colleagues have examined wild rhesus monkeys' numerical abilities. The experiment was unchanged from Wynn's except that large eggplants were used as the items to be counted. The monkeys responded to the events in a manner similar to the infants, looking longer at operations that yielded unexpected numbers of eggplants. Thus, the use of identical experimental procedures has demonstrated that infants and rhesus monkeys seem to have similar numerical estimation abilities.

How and when humans come to understand the correspondence between a physical representation of space, such as a map or a model, and its real-world referent has also been the focus of much research in developmental psychology. Judy DeLoache has approached this question in her studies exploring children's ability to understand the representational nature of a scale model. She and her colleagues have found that after witnessing a

miniature item being hidden in a scale model of a room, three-year-old children can locate a full-size item hidden in the analogous location in the real room. However, slightly younger children, 2.5-year-olds, have difficulty with the task. Their difficulty with the task implies the lack of 'representational insight', or knowledge that the model and room are related as symbol and referent. DeLoache and her colleagues have suggested that many factors can contribute to the development of this understanding, including the perceptual similarity between the model and the room, experience with other symbol systems, instruction on the nature of the model–referent relationship, and the ability to form a 'dual representation' (i.e. understanding the model as an object unto itself as well as a symbol for something else).

Until recently, it was not known if a nonhuman species could understand a physical representation of space such as a scale model and use it as a source of information regarding the environment. Valerie Kuhlmeier and Sarah Boysen found that chimpanzees were able to solve a scale model task that was similar to DeLoache's procedure. After watching an experimenter hide a miniature bottle of juice within a scale model of an outdoor enclosure, three chimpanzees readily found the real juice bottle that was hidden in the analogous location in the actual enclosure. That is, they went to the correct site and retrieved the bottle immediately upon entering the enclosure. These chimpanzees performed similarly to the three-year-old children in DeLoache's task. However, the performance of the other four chimpanzees tested was poor, or at best, varied. These four chimpanzees often relied on a search strategy that consisted of searching the hiding site in the front left corner of the enclosure, and continuing to search each site successively as they circled the room clockwise. The search strategy that was observed with these chimpanzees has not been reported in studies by DeLoache and suggests that, although some chimpanzees solve the task in a manner similar to young children, there may be other factors that influence some chimpanzees' performance. These factors will no doubt be the focus of future study.

Thus, these studies of numerical ability and scale model comprehension illustrate how similar test procedures can be used to examine the similarities and differences between nonhuman primates and human children. They demonstrate that the dialogue between those who study cognitive development and those who study animal cognition is increasing, with many researchers flexibly moving back and forth from one to the other during their careers.

CONCLUSION

The study of animal cognition is a growing one, with increasing dialogue among researchers in different academic fields. The topics discussed above represent only a subset of research in animal cognition but provide evidence for flexible and complex cognitive abilities in nonhuman animals. Future research of these topics, and all areas of animal cognition, is necessary to further examine the manner in which animals learn, behave, and think.

Further Reading

Boysen ST and Berntson GG (1989) Numerical competence in a chimpanzee (*Pan troglodytes*). *Journal of Comparative Psychology* **103**: 23–31.

Cheney DL and Seyfarth RM (1990) *How Monkeys See the World*. Chicago, IL: University of Chicago Press.

Dasser V (1988) A social concept in Java monkeys. *Animal Behaviour* **36**: 225–230.

DeLoache JS (1987) Rapid change in the symbolic functioning of very young children. *Science* **238**: 1556–1557.

Hauser MD, Carey S and Hauser LB (2000) Spontaneous number representation in semi-free ranging rhesus monkeys. *Proceedings of the Royal Society, London*, **267**: 829–833.

Kuhlmeier VA, Boysen ST and Mukobi KL (1999) Scale model comprehension by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology* **113**: 396–402.

Myowa-Yamakoshi M and Matsuzawa T (1999) Factors influencing imitation of manipulatory actions in chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology* **113**: 128–136.

Nagell K, Olguin R and Tomasello M (1993) Processes of social learning in the imitative learning of chimpanzees and human children. *Journal of Comparative Psychology* **107**: 174–186.

Parr LA and deWaal FBM (1999) Visual kin recognition in chimpanzees. *Nature* **399**: 647–648.

Shettleworth S (1998) *Cognition, Evolution, and Behavior*. New York, NY: Oxford University Press.

Whiten A, Custance DM, Gomez J-C, Teixidor P and Bard KA (1996) Imitative learning of artificial fruit processing in children (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology* **110**: 3–14.

Wynn K (1992) Addition and subtraction by human infants. *Nature* **58**: 749–750.

Zentall TR, Sutton JE and Sherburne LM (1996) True imitative learning in pigeons. *Psychological Science* **7**: 343–346.

Animal Language

Introductory article

Duane M Rumbaugh, Georgia State University, Atlanta, Georgia, USA
Michael J Beran, Georgia State University, Atlanta, Georgia, USA

CONTENTS

Animal language references the field in which animals' capabilities for various dimensions and functions of language are researched. Language is here defined as a neurobehavioral system that provides for the construction and use of symbols to enable the conveyance and receipt of information and novel ideas between individuals. The meanings of symbols in this system are basically defined and modulated through social interactions.

INTRODUCTION

Although fascinated for centuries by the prospect of 'talking with animals', humans have only recently begun studying language as a part of the behavioral repertoire of nonhuman animals. Although there are abundant data indicating that many animal species have evolved various forms of both simple and complex communication systems (such as the vocalizations of vervet monkeys), language, as it is defined and promoted by humans, has been a more difficult phenomenon to demonstrate when studying nonhuman animals. Songs of the whales and birds, dances of the honeybee, vocalizations of monkeys, and other such forms of communication all fall short of the requirements of a formal language system in that they lack some or all of the formal components of language as it is defined. For example, the meaning of the dance of the honeybee is disrupted with landmark orientation changes. Also, monkeys give alarm calls to individuals already aware of the danger, and bird songs are predominantly ritualized and do not frequently change to meet new circumstances. However, some experimental attempts to demonstrate that nonhuman animals can learn and use language systems have been successful.

SYNTAX USE AND COMPREHENSION

There are two primary structural components to a language system: its semantics and its syntax. Syntax refers to the rules and guidelines governing how linguistic units (typically words) can be combined and the order in which those units must be used to convey the meaning of the speaker. Syntax is therefore highly tied to productive use of a language. Semantics refers to the word meanings of a language, and is intricately tied to comprehension of a language. Animal language studies have focused on both of these aspects of language.

APES

Numerous experimental investigations of language have been conducted with great apes. In these projects, the focus has shifted across time. Initial attempts were aimed at teaching apes to speak. Later, attempts were made to teach apes the formal syntactic structure of a language. Then, the focus shifted to comprehension of symbol systems. Most recently, apes' acquisition of language has occurred most robustly when infant apes are provided with a linguistic environment in which to mature, and the apes can observe and integrate themselves into this language-rich surrounding. In these attempts, the focus is on the environment and its role in both comprehension and production of language by apes.

It was discovered very early on that apes had a limited ability for producing speech sounds. This handicap would have to be overcome through the use of 'artificial' symbol systems. These systems included sign language, plastic tokens used on magnetic boards, and embossed geometric symbols called lexigrams. The sign language studies were among the first conducted, and they were conducted in four separate research projects. The Gardners worked with the chimpanzee Washoe, Terrace worked with the chimpanzee Nim, Patterson worked with the gorilla Koko, and Miles worked with the orangutan Chantek. In these projects, the apes were taught to make signs, and they

were exposed to human caretakers' use of sign as well. Although initial reports indicated that the apes quickly learned to use and respond to others' use of the signs, doubt was cast on the projects by Terrace himself. Terrace concluded that the apes were simply imitating the sign use of the humans around them, and that there was not an originality or appropriateness (nor any indication of syntax) in the sign language of the apes.

Premack worked with a chimpanzee, Sarah, using plastic tokens that the chimpanzee responded to, based on the 'questions' posed to her. Sarah showed a great affinity for properly answering the questions posed to her using this system, but she rarely used the tokens to communicate her intentions or desires. Her use of the tokens was in response to posed questions only. Thus, although Sarah demonstrated comprehension of the 'linguistic' problems, she did not develop language in the sense of operating in a two-way, give-and-take communicative context.

Rumbaugh worked with the chimpanzee Lana through use of a computerized system. The lexigram keyboard responded to the requests made by Lana, provided those requests were in the proper grammatical order. Lana learned to string together lexigrams into stock sentences that produced desired outcomes. On her own, she learned to finish sentences correctly when those sentences were started by others.

Despite these novel approaches to studying language acquisition in apes, concerns and criticisms remained that, although the apes were learning what was expected of them, they did not have a true language. Rather, claims were made that apes such as Lana learned only to chain sequences of lexigrams appropriately without understanding the meanings of the symbols that were used. To address these concerns, future projects centered on the issues of intentional communication, reference, and semantics, and syntax became of secondary concern. Receptive competence became the important capacity to demonstrate.

Savage-Rumbaugh made several important findings when she began work with two chimpanzees, Sherman and Austin. These chimpanzees were taught not only to use lexigrams (which were no longer linked solely to a computer system but were used in the everyday interactions between humans and apes) but also to comprehend each other's (and human caretakers') use of those symbols. Sherman and Austin displayed a sense of symbolic thought through the novelty of their lexigram use as well as the use of other symbols (such as labels to containers and food items). The two chimpanzees made statements about their future actions, they requested items from each other, and they responded to each other's requests appropriately. Both chimpanzees also learned not only to categorize real-world items into functional categories, but also to categorize the lexigrams for those items as well. Lexigrams were categorized as either foods or tools on the very first trial in the same manner in which the real-world objects represented by those lexigrams were categorized. In other words, the lexigrams had meanings for these apes, and the lexigrams functioned as symbols for these apes.

Fortuitously, the next finding in ape-language research occurred as a result of the failure of an adult female bonobo (*Pan paniscus*, a species closely related to the chimpanzee) to replicate the findings with Sherman and Austin. After repeatedly trying to teach this bonobo to use the lexigram symbols, the attempts were interrupted so that she could be bred with another bonobo. During this time, it was discovered that her son, who had never been instructed in the use of lexigrams but who had constantly been in the area while his mother was instructed, had acquired not only comprehension of some of the symbols, but also comprehension of spoken English. This bonobo, Kanzi, responded appropriately to novel spoken requests of humans as well as to their lexigram use. This finding, of spontaneous language acquisition that occurs when young apes are raised in a language-rich environment in which there is structure in the use of lexigrams and spoken English, has since been replicated with two additional bonobos and a chimpanzee. Additionally, research with bonobos has demonstrated a much greater complexity in both language comprehension and production than had previously been demonstrated in any other study of nonhuman animals. The bonobo Kanzi responded appropriately to sentences in which word order was of vital importance in correct interpretation of the speaker's meaning. Additionally, Kanzi and other bonobos raised in a similar language-enriched environment displayed comprehension of far more spoken English than lexigrams; this comprehension is the result of their living in a linguistically rich environment in which information is available, provided that the apes can glean it from both the context of an interaction and the words spoken by humans in that context.

These findings provided clear evidence that both comprehension (of spoken English and lexigrams) and productive use (of lexigrams) was a part of the cognitive competence of these animals when raised in a structured, language-rich environment in

which language learning was latent (i.e. not taught through discrete trials training). This environment is much like that in which human children are raised, and it is almost certainly this environmental context which provides the necessary stimulation for language to develop in humans and apes.

In ape-language studies, those focused on sign language have been primarily concerned with the extent to which apes could learn the meanings of different signs (the semantic aspects of sign language). Washoe, Nim, and Chantek learned the names of numerous items as well as actions and locations. The work of Premack and Rumbaugh, however, was more focused on the grammatical requirements of language. Lana learned not only to construct grammatically correct sequences of lexigrams that followed the grammar of her language (called 'Yerkish'), but she also finished sentences started by experimenters by retaining the necessary grammar needed for the sentence to be correct. However, if the provided sentence fragment was already grammatically incorrect (as produced by the experimenter), Lana would erase it and start from scratch with a correct sequence. Thus, the emphasis was on Lana's productive language skills.

As the emphasis shifted to the productive and receptive skills demonstrated with lexigrams by the chimpanzees Sherman and Austin, elements of receptive language were more evident. However, a focus on the syntactic understanding of these chimpanzees was absent. In contrast, it became clear with later research that at least one bonobo, Kanzi, responded appropriately to spoken requests dependent not just on the words within that request, but also on the order in which the words were produced. Kanzi, therefore, could respond appropriately to sentences such as 'Pour the Coke in the lemonade' and 'Pour the lemonade in the Coke'. Additionally, Kanzi responded appropriately to sentences with embedded phrases. For example, when presented with the sentence 'Kanzi, get the ball that is outdoors', Kanzi would walk past another ball indoors to retrieve the ball that was outdoors. Kanzi demonstrated syntactic understanding of word-order rules for spoken English, and he responded appropriately based on the referential and relational aspects of spoken sentences.

DOLPHINS

Dolphins also made good candidates for language acquisition research because of their complex use of vocal communication (in the form of whistles) as well as their large brain size. Research with dolphins by Herman has focused on comprehension rather than production. In one experimental paradigm, the dolphin Akeakamai (Ake) was taught to respond to gestural signals produced by humans using their arms and hands (another paradigm used by Herman involved an acoustic language with acoustic signals). Each signal representing an object, an action or a modifier (such as 'left' or 'right') could be combined with other signals so that the dolphin's comprehension could be measured through its response. Initially, the question was whether the dolphins could respond appropriately to requests either asking for an action to be done to a single object or asking that the dolphin perform an action with two objects. Later, semantic categories could be combined according to syntactic rules in such a way that three-, four-, and five-word relational sequences could be directed to the dolphin. The dolphins organized their responses in such a way as to take into account both the syntactical aspect of the sentence and the meanings of all symbols within that sentence. For example, syntactic understanding was demonstrated through correct responding to sentences containing the same words but different word order (such as LEFT HOOP PIPE FETCH, which asked the dolphin to take the pipe to the hoop on her left, versus HOOP LEFT PIPE FETCH which asked the dolphin to take the pipe on her left to the hoop). Semantic understanding was demonstrated through selection of the correct items.

In addition to responding to the requests of humans who were present, the dolphins responded appropriately to video displays of this gestural language even when the clarity of the image was degraded, and to anomalous gestural signals in which the semantic rules and syntactic constraints of the language were violated. These abilities indicate that the dolphins have a referential understanding of the signs used in their language. As with apes, dolphins utilize mental representations when responding to language-mediated tasks. Comprehending degraded images without training suggests that the dolphins have a network of semantic and gestural representations in memory. The dolphins recognized degraded images by comparing those images to a set of representational exemplars in memory rather than through simple stimulus generalization. The degree of generalization in responding to even highly degraded gestures indicates that the dolphins represent the gestures in memory.

The research with dolphins, as well as research by Schusterman with sea lions, has focused heavily

on syntactic aspects of language comprehension as well as semantic understanding of word meaning. As noted, the dolphins respond appropriately not only to the meaning of individual gestures and signals, but also to the relation of those gestures and signals to each other within a sentence. Syntactic understanding was shown through responses to semantic contrasts in reversible sentences, to syntactically anomalous sentences, to structurally novel sentences, to sentences with word modification within the sentence (where one word modified the meaning of another), to interrogative and imperative sentence form contrasts, and to sentences that contained variations in the placement of modifiers.

SUMMARY

We have provided a working definition appropriate in defining the aspects of communication that differentiate language from nonlanguage. This definition allows us to investigate language competence in nonhuman animals. Such a definition is needed to establish the comparative and evolutionary bases of fully elaborated human language. Although humans clearly do more with their language capacities (especially in the area of verbal speaking) than do any other animals, comparative language acquisition research is vital to understanding the basic mechanism of language. Animal competencies for certain dimensions of human language clearly reflect comprehension and use of meanings accrued by symbols.

Further Reading

Gardner RA, Gardner BT and van Cantfort TE (1989) *Teaching Sign Language to Chimpanzees*. Albany, NY: State University of New York Press.

Herman LM (1988) The language of animal language research: reply to Schusterman and Gisiner. *Psychological Record* **38**: 349–362.

Herman LH, Morrel-Samuels P and Pack AA (1990) Bottlenosed dolphin and human recognition of veridical and degraded video displays of an artificial gestural language. *Journal of Experimental Psychology: General* **119**: 215–230.

Herman LM, Richards DG and Wolz JP (1984) Comprehension of sentences by bottlenosed dolphins. *Cognition* **16**: 129–219.

Patterson FL and Linden E (1981) *The Education of Koko*. New York, NY: Holt, Rinehart & Winston.

Premack D and Premack AJ (1983) *The Mind of an Ape*. New York, NY: Norton.

Roitblat HL, Herman LM and Nachtigall PE (1993) *Language and Communication: Comparative Perspectives*. Hillsdale, NJ: Lawrence Erlbaum.

Rumbaugh DM (1977) *Language Learning by a Chimpanzee: The LANA Project*. New York, NY: Academic Press.

Rumbaugh DM and Savage-Rumbaugh ES (1994) Language in a comparative perspective. In: Mackintosh NJ (ed.) *Animal Learning and Cognition*, pp. 307–333. San Diego, CA: Academic Press.

Savage-Rumbaugh ES (1986) *Ape Language: From Conditioned Response to Symbol*. New York, NY: Columbia University Press.

Savage-Rumbaugh ES (1991) Language learning in the bonobo: how and why they learn. In: Krasnegor NA, Rumbaugh DM, Schiefelbusch RL and Studdert-Kennedy M (eds) *Biological and Behavioral Determinants of Language Development*, pp. 209–233. Hillsdale, NJ: Lawrence Erlbaum.

Savage-Rumbaugh ES and Lewin R (1994) *Kanzi: The Ape at the Brink of the Human Mind*. New York, NY: Wiley & Sons.

Savage-Rumbaugh ES, Murphy J, Sevcik RA *et al*. (1993) Language comprehension in ape and child. *Monographs of the Society for Research in Child Development* **1**: 1–221.

Schusterman RJ and Gisiner R (1988) Artificial language comprehension in dolphins and sea lions: the essential cognitive skills. *Psychological Record* **38**: 311–348.

Terrace HS (1979) *Nim*. New York, NY: Knopf.

Animal Learning

Introductory article

Armando Machado, University of Minho, Braga, Portugal
Francisco J Silva, University of Redlands, Redlands, California, USA

CONTENTS

The field of animal learning studies the behavioral mechanisms and processes that animals use to adapt to changes in their environment. Its emphasis is on environment–behavior interactions.

INTRODUCTION

Learning refers to a heterogeneous set of processes that evolved in animals to accommodate changes in their environments. These processes can produce relatively permanent changes in behavior and are brought into play by some form of interaction between the animal and its surroundings. For instance, a moving nematode (a tiny roundworm) that stops momentarily or reverses its motion when it experiences a vibration will cease doing so if the vibration occurs repeatedly. Foraging bees perform an intricate dance, the orientation and speed of which change with the direction and distance of the food source from the hive, such that other bees also can locate the food site. Hungry domestic cats mew in the presence of their owner, who will then give them food. In all of these examples, the organism's behavior is showing the effects of particular interactions between itself and its environment. Classifying distinct types of interactions, identifying their elements, quantifying their static and dynamic properties, and describing how their cumulative effects are expressed is the goal of people who study learning. Before classifying interactions between organisms and their environments, it is important to place the process of learning within an evolutionary context – that is, to understand why learning evolved.

EVOLUTION AND LEARNING

It is widely assumed that learning evolved in specific contexts, such as gathering food, eluding predators, capturing prey, attracting mates and avoiding poisons. Despite this variety in contexts, countless experiments indicate that the same principles of learning hold across many different species, tasks and behaviors. How can we reconcile the assumption that learning evolved in specific contexts with the fact that it occurs similarly in many contexts? The answer is, by conceiving of learning processes as mixtures in varying proportions of both context-specific and general mechanisms. For example, taste aversion learning occurs when animals avoid gustatory and olfactory cues associated with foods that made them ill, even when the flavor is separated from the illness by many hours. That animals can learn which cues predict biologically significant events is commonplace; that animals can learn to avoid cues separated by many hours from a biologically significant event typically happens only when flavor is the predictive cue and illness is the significant event.

To clarify further the relationship between context-specific and general features of learning, consider the following analogy. A house built for living near the Arctic Circle will differ considerably from one built for living in south Florida: the former requires thick insulation, double-paned windows and a furnace; the latter needs mildew-resistant insulation, shaded glass and air conditioning. Despite these differences, there are general features common to both houses, such as the presence of windows, doors, rooms, walls and a roof, and a general function for both houses, such as sheltering and protecting its inhabitants from weather and predators. Because the function of a house is similar in both regions, there is some commonality to the solutions.

The presence of general features in learning is illustrated by the fact that animals from widely separated taxa respond to environmental stimuli in similar ways. For example, they ignore repetitive harmless stimuli, a process called habituation; they

detect correlations among biologically important events, a process called Pavlovian conditioning; and they learn causal relations between their behavior and its consequences, a process called operant conditioning.

Central to these processes is temporal integration, for only by tracking and integrating events across time can animals determine whether an event is repeating, whether it occurs before, during or after other events, or whether it is a reliable consequence of responding or not responding. It should come as no surprise, then, that temporal variables are often critical determinants of learning.

HABITUATION: ADAPTING TO REPETITIVE, HARMLESS EVENTS

When a harmless stimulus occurs repeatedly and there are no other events associated with it, there might be an advantage to ignoring the stimulus. For example, imagine a land snail on a small wooden platform that vibrates briefly while the animal moves. Typically, this stimulus (the vibration) elicits a protective response, contracting the antennas. If these vibrations are repeated, say every 30 s, then the contractions decline. Eventually, the vibrations are ignored in the sense that the antennae are not contracted and the snail keeps moving. This waning of a response to repeated presentations of a stimulus is termed habituation.

To interpret habituation we can appeal to the concept of response threshold, which is defined roughly as the minimum stimulus intensity required to elicit a reflexive response. As the stimulus is repeated, the threshold increases, which makes it more difficult to produce a response. Eventually the threshold is greater than the current stimulus intensity and the response fails to occur.

Properties of Habituation

Habituation is present in virtually every animal species, from single-celled animals such as the ciliate *Vorticella* to humans. It has even been observed in individual motor neurons. This widespread phenomenon deserves attention for two related reasons. First, it introduces some of the key variables and functional relations that psychologists have identified in most learning processes. Second, habituation reveals the amazing complexity of even the simplest of learning processes.

Recovery from habituation

In the example above, if the platform ceases to vibrate after habituation has occurred, then with the passage of time since the last vibration, the snail's antennae are increasingly likely to contract when the platform again vibrates. In other words, habituation seems to 'wear off' when the stimulus that produced it is not presented. This recovery of the response corresponds with a return of the threshold to its initial value.

Stimulus intensity

If the vibration of the platform is sufficiently intense, then the snail may withdraw into its shell. This response might also habituate if the strong stimulus is repeated. Typically, however, the courses of habituation and recovery from habituation are slower for strong than for weak stimuli. When the stimulus is more intense, the rise of the threshold takes longer to surpass the stimulus intensity, and the fall of the threshold in the absence of the stimulus takes longer to return to its initial, baseline value.

Time between stimulus presentations

All else being equal, stimuli closer in time produce faster habituation than stimuli farther apart. This finding is consistent with the threshold account of habituation: longer intervals give the threshold more time to decrease, which partly offsets the effect of the stimulus presentations. Interestingly, high rates of stimulation may also lead to faster recovery from habituation. This result, unlike the preceding ones, does not follow from the view that changes in threshold are responsible for habituation, unless the rate at which the threshold returns to its baseline value depends on the rate of stimulation.

Relearning effect

Imagine the following experiment. After we record the course of habituation on day 1, we stop the vibrations and let recovery occur. On day 2 we vibrate the platform again and record the new course of habituation. Typically, the rate of habituation is faster during the second day. The importance of this finding is that the difference in the rates of habituation shows that the animal's internal state on day 2 is different from its state on day 1 despite the similarity in its initial responses. Again, a simple change in the response threshold cannot accommodate this finding.

Stimulus generalization and stimulus specificity

The contraction of the antennas ceases not only in the presence of the original vibration, but also in the presence of similar stimuli (stimulus generalization). However, it is readily elicited by different

stimuli such as a blast of air (stimulus specificity). These properties are two sides of the same coin; generalization focuses on the fact that habituation to one stimulus extends to some of the other stimuli that can also elicit the response; specificity focuses on the fact that habituation to one stimulus does not extend to all stimuli that can elicit the response. Careful empirical work is needed to identify the stimulus properties (e.g. intensity, duration, rate) along which generalization proceeds.

Functions of Habituation

As noted above, habituation has been observed in almost every animal species. Why is it so prevalent – and why do the properties of habituation hold true across many different species, responses and *a fortiori* physiological mechanisms? Probably, habituation occurs because it is sometimes safe and economical to ignore a repetitive stimulus. An animal that continued to respond to every stimulus impinging on its receptors would be overwhelmed by stimulation and incapable of acting appropriately. However, the animal would pay a high price if the effects of habituation were permanent, for what was once a harmless vibration caused by the running of a distant animal might now be caused by an approaching predator. In the same vein, assuming stronger stimuli are potentially more harmful than weaker stimuli, it makes sense that they should be ignored less quickly than weaker stimuli. Classifying incoming stimuli as 'The same!' 'The same!' 'The same!' also seems safer when the stimuli are close in time.

PAVLOVIAN CONDITIONING: LEARNING ABOUT CORRELATIONS

Food and water, predators and prey, mates and offspring, and escape routes and shelter, are some of the primary determiners of survival and reproduction. As such, it is reasonable to attribute great evolutionary advantage to animals capable of anticipating them. For all animals, specific sounds, sights or odor trails, places or times of occurrence, or more complex sequences and configurations of stimuli might be reliably correlated with biologically important events. If an animal could learn the correlational texture of its world (i.e. the relationships among events), then it would have the advantage of responding one way when a stimulus predicts an important event and in another way when a stimulus does not.

The correlations that an animal can learn depend on the animal and the types of stimuli and events in its environment. In terms of the animal, there might be reliable cues that it cannot detect simply because it has not evolved the required biological machinery (e.g. sensory receptors). In terms of the environment, a stimulus might be detectable but its frequency of occurrence or its reliability as a cue might be too low to support the evolution of an ability to fully exploit it; the cost would outweigh the benefit, as it were. Learning about the cueing function of a stimulus is therefore constrained both by the animal's physiology and the specific arrangements of events in the animal's world.

Constraints notwithstanding, how does an animal learn the correlation between a neutral and a biologically important stimulus? The pioneering work on this question was conducted by Ivan Petrovich Pavlov (1849–1936), the famous Russian physiologist and 1904 Nobel prizewinner. In good scientific fashion, Pavlov reduced the problem to its bare essentials: a tone reliably preceded a bit of meat powder delivered to the mouth of a hungry dog. Of interest was the animal's behavior during the tone. Initially, when the tone was presented the dog pricked up its ears and looked in the direction of the source of the tone, but, critically, it did not salivate. After a few pairings of the tone and food, the orienting response elicited by the tone ceased (habituation had set in) and a new response during the tone began to occur – salivation. Because, 'food in the mouth' elicited copious salivation without any previous training, Pavlov called it the unconditional stimulus (US) and 'salivation in the presence of food' the unconditional response (UR). As the quantity and quality of salivation to the tone depended on the prior predictive history of the tone, Pavlov called the salivation to the tone a conditional response (CR) and the tone a conditional stimulus (CS). The study of how behavior changes when two or more stimuli are paired, as in the preceding example, is known as Pavlovian or classical conditioning.

With this and similar laboratory preparations, Pavlov and many subsequent researchers have tried to understand how animals learn the cueing function of stimuli. Some of their experiments showed the following results, many of which resemble those obtained in studies of habituation.

Extinction

If, after the tone elicits salivation reliably, it is presented without the food, then the dog will eventually stop salivating during the tone. That is, when the CS no longer predicts the US, the CR weakens and may eventually disappear. Through acquisition

and extinction processes, animals adjust to changes in the pattern of events in their environment.

Spontaneous Recovery

If the experimenter allows the dog to rest for, say, 24 h after the extinction training, and then presents the tone again, the animal that had stopped salivating to the tone may again salivate to it; that is, the CR spontaneously recovers. The passage of time undoes some of the effects of extinction. Why spontaneous recovery of the CR happens is still poorly understood.

Stimulus Generalization

Having learned to salivate to a specific tone, the dog also will salivate to similar tones. That is, a CR will be elicited by the original stimulus as well as similar stimuli; however, the more different these other stimuli are from the original CS, the weaker the CR they elicit. Because no stimulus ever recurs in precisely the same way (e.g. the rustling of the leaves announcing a lion is different in different situations), it is advantageous to extend newly acquired responses to similar stimuli.

Stimulus Discrimination

When Pavlov alternated two tones during training and paired one but not the other with food, his dogs eventually salivated only to the tone paired with food. That is, if one stimulus (CS+) is paired with a US, but another stimulus (CS−) is not, then the CR will occur only or mainly in the presence of the CS+. Stimulus discrimination helps ensure that a response occurs in particular environments, rather than indiscriminately across situations and time.

Contingency Effects

Assume that during the original training the food only follows the tone on 50 percent of the trials. On the remaining 50 percent the tone occurs alone. Under this circumstance, the amount of salivation to the tone during training is smaller than when the food always followed the tone. Similarly, if food also occurs occasionally in the absence of the tone, the CR is weaker than when food only follows the tone. In the extreme, if food occurs more often in the absence of the tone than in its presence, the tone will actively suppress salivation instead of eliciting it. In summary, the results of many experiments show that animals are sensitive to the direction (positive or negative) and the strength of the correlation, or contingency, between the CS and the US.

The effect of contingency shows that temporal contiguity between the tone and food is insufficient to ensure that the tone will become a CS. Much depends on what else the animal has been experiencing, both during the presence and the absence of the tone. That is, the animal seems to integrate events that are temporally extended, and to behave according to the actual correlation value between the tone and the food. Both temporal and probability relations between the CS and US, or contiguity and contingency, are important in Pavlovian conditioning.

In fact, the process is even more complex than stated above. Consider an experiment in which a tone is paired with food until it elicits salivation reliably. Next, the tone is presented along with a light, and this compound stimulus is followed with food. Will the light elicit salivation when it is presented alone and without the food? Because food always occurs after the light and never in its absence, the light is maximally (and positively) correlated with food. Moreover, because the food closely follows the light, the two stimuli also are temporally contiguous. Hence, one might predict a strong association between the light and food and, therefore, salivation to the light. However, routinely little or no salivation to the light is found. Control experiments indicate that because the tone already predicted the food at the end of the first part of the experiment, it somehow blocked the association 'light–food'. We could say that the light provided no new information about the food beyond that already provided by the tone and, hence, the light did not help the animal anticipate the US any better than the tone. The important point is that such blocking highlights the fact that an animal's prior experiences can modulate the effects of contiguity and contingency.

Relevance of Pavlovian Conditioning

Since Pavlov's pioneering work, the study of Pavlovian conditioning has revealed many other complex relations among the CR and temporal variables, the sequential arrangements of the various stimuli, the context in which conditioning occurs, and the animal species and the particular response system under consideration. Pavlovian conditioning is fundamental to understanding drug addictions, phobias and a variety of sexual responses in humans and other animals. Its domain of study also has become increasingly quantitative. Real-time, dynamic models of the learning process

have started to replace verbal accounts. However, much remains unknown about the process through which stimuli that are insignificant when considered in isolation become significant when they signal biologically important events.

OPERANT CONDITIONING: LEARNING ABOUT CAUSATION AND CONTROL

The preceding discussion focused on how an animal's behavior is changed by repeated presentations of single stimuli (habituation) or by relationships among stimuli (Pavlovian conditioning). In both of these situations, behavior changes as a result of the stimuli that precede it. However, it is also the case that things happen after a response. For example, a young male cowbird sings one of its song variants and elicits a subtle wing flick from a female. An adult male cowbird sings a variant that stimulates a precopulatory display in a female and a vigorous attack from a dominant male. However, if the same adult cowbird sings a less stimulating song, then it avoids being chased by the dominant male. Operant or instrumental learning results when an animal's behavior causes a stimulus change, which in turn changes the animal's subsequent behavior: the young cowbird is more likely to sing the variant that caused the positive female reaction; the adult cowbird is less likely to sing the song that caused the attack and more likely to sing the one that avoided it. This capacity to change behavior because of its consequences enables animals to learn about control and to exploit the causal texture of their social and physical worlds.

In the examples above, the operant response produced different types of consequences. Psychologists classify these consequences by means of their effect on behavior. Consequences of an action that increase the likelihood of that action recurring are termed 'reinforcers' – positive reinforcers if the consequence is the occurrence of a stimulus (e.g. the wing flick display from the female), and negative reinforcers if the consequence is the cessation or avoidance of a stimulus (e.g. the threat and attack avoided by the adult cowbird when it sang the less stimulating song). Consequences of an action that decrease the probability of that action recurring are 'punishers' (e.g. the attack suffered by the low-ranking cowbird when it sang its most stimulating song). By modifying its behavior to produce reinforcers and eliminate or avoid punishers, an animal shapes its world while its behavior is shaped by its world. This closed feedback system is the hallmark of operant conditioning.

The laboratory study of operant conditioning began with the work of Edward L. Thorndike (1874–1949), who showed that cats placed in a puzzle box (a wooden cage with a door that could be operated by the animal) become quicker at escaping with repeated successes. Later, B. F. Skinner (1904–1990) studied how behavior is shaped by its consequences, and how new response forms emerge when variations in behavior have different consequences. To conduct his experiments, Skinner invented the operant chamber, a box with a lever that hungry rats could press to receive food dispensed into a tray. The operant chamber soon became the microscope of learning psychologists.

What sort of consequences function as reinforcers or punishers? An agreed-upon theory that predicts which stimuli reinforce or punish behavior, the circumstances in which they do so, and why they do so, has eluded experimental psychologists. However, researchers have identified several factors that influence the behavioral effects of reinforcers and punishers.

Contiguity and Contingency

As in Pavlovian conditioning, contiguity and contingency are important. Other things being equal, long intervals between a response and a consequence weaken the strengthening effect of the latter; the more immediately a consequence follows a response, the more likely it is that the response will be affected by the consequence. However, short intervals may have weak effects if the correlation between the response and the consequence is low. This can occur in two ways – when a response is followed by the consequence too infrequently, or when a reinforcer occurs independently of the response that normally produces it. An adult, subordinate cowbird might continue to sing its most stimulating song if that song rarely produces aggression from other males; a young cowbird might spend less time singing if the positive female display occurred in the absence of the song.

Extinction

The behavioral changes brought about with operant conditioning are reversible. When the environment changes and a response that used to be followed by a positive outcome is no longer followed by it, the probability of that response occurring declines. That is, the animal ceases emitting behavior that is no longer functional. However, extinction may be rapid or slow. On some occasions, the animal quickly changes its behavior,

whereas on other occasions it perseveres for long periods of time. Whether extinction is rapid or slow depends largely on whether every instance of a particular response was reinforced (rapid) or only occasionally reinforced (slow).

Schedules of Reinforcement

In the natural environment, it is rare that every instance of a particular response is followed by a consequence. To understand the effects of intermittent reinforcement psychologists have studied different rules specifying when a response will produce a consequence. Two examples of these rules, collectively known as schedules of reinforcement, are ratio and interval schedules. In ratio schedules the reinforcer depends solely on the occurrence of behavior (e.g. a rat receives food each time it completes five lever presses); in interval schedules the reinforcer depends on the occurrence of behavior and the passage of time (e.g. a rat receives food following the first lever press after 15 s since the previous occurrence of food). Because there are no restrictions to when the rat can press the lever, the experimenter can study how the rate of a response changes across time as a function of how it produces a consequence.

Typically, ratio schedules support higher rates of responding than comparable interval schedules. Why? Because the two types of schedule induce different feedback functions, that is, different relations between response rate and reinforcement rate. As an illustration, consider a ratio schedule in which five responses produce one reinforcer. In this schedule, how often reinforcers occur depends exclusively on how rapidly the animal responds; reinforcement rate will always equal one-fifth of the response rate. In contrast, consider an interval schedule in which a response is reinforced if it occurs at least 15 s since the previous reinforcer. In this schedule, a response rate of four responses per minute matches the reinforcement rate. Slower response rates produce proportional changes in reinforcer rates, but faster response rates do not. That is, reinforcement rate ceases to vary with response rate. Differences in the feedback function explain why ratio schedules typically produce higher rates of responding than interval schedules.

Choice

Just as simple processes combine to produce complex phenomena such as weather, geological formations and evolution, so too do basic processes of learning combine to produce more complex behavior. Choice among options is an example that illustrates how reinforcement rates interact to affect behavior.

In the simplest situation, an animal faces two response keys, each of which delivers a reinforcer according to a schedule of reinforcement. For example, a pigeon might peck one key and receive a morsel of food with probability p, or peck another key and receive food with probability q. Another example might consist of a rat that presses one lever that delivers reinforcers with rate r, or another lever that delivers them with rate s. Studies such as these with pigeons, rats, rhesus monkeys and humans, among other species, have yielded a robust empirical finding known as the matching law. The law states that the proportion of choices on one alternative equals the proportion of reinforcers obtained from that alternative. In symbols:

$$x/(x+y) = R_x/(R_x + R_y) \quad (1)$$

where x and y are the total numbers of choices of each alternative, and R_x and R_y are the corresponding total numbers of obtained reinforcers.

Much less understood is how basic behavioral processes combine to yield the matching law. Some researchers propose that the equality is the outcome of the cumulative strengthening effect of individual reinforcers on the two response alternatives. Others suggest that the law results from the animal's sensitivity to global rates of reinforcement and its ability to maximize these rates under constraint. Still others suggest that matching derives from the tracking of the intervals between successive reinforcers on the two alternatives. In these three hypotheses we see, once again, the difficulty of determining the timescale of the learning process. Equally poorly understood is the acquisition of preference and how it relates to the various parameters of the choice situations (e.g. how the values of p and q, or r and s, in the examples above, determine how fast the animal comes to prefer the best alternative).

Stimulus Control

Because no response occurs in a vacuum, a response–consequence relation is always context-specific. In the laboratory, if a pigeon is reinforced for pecking a green key but not a red one, then the bird will restrict its pecking to the green key. This differential responding occurs because the two stimuli are correlated with different response–consequence relations: but, as with habituation and Pavlovian conditioning, the extent that stimuli different from the ones used in training control

operant behavior depends on a variety of factors. For example, the amount of pecking in the presence of stimuli similar to the green key (e.g. a blue key) and stimuli similar to the red key (e.g. an orange key) depends on how much training the pigeon received with the original stimuli. More extensive stimulus discrimination training promotes discrimination, whereas less training promotes stimulus generalization. Also, reinforcing behavior differentially in the presence of different stimuli produces sharper discriminations (less generalization) than reinforcing a response in the presence of a single stimulus.

Moreover, when two or more stimulus elements signal that a response is likely to be reinforced, some form of stimulus competition may ensue. Consider the following experiment: a pigeon is trained to peck a green key with a black horizontal line, but not to peck a red key with a vertical line. The degree of stimulus discrimination and generalization is then tested by recording the amount of pecking at a white key during presentations of the line in various degrees of orientation. During the test, the pigeon pecked all line orientations similarly (i.e. stimulus generalization). However, when tested with keys of different colours but without the line, the pigeon pecked most at the green key and least at the red one (i.e. stimulus discrimination). This example shows that, for reasons that are poorly understood, some features of a stimulus may overshadow others. For the bird in this example, color overshadowed line orientation; but the reverse could have happened. A similar effect occurs in Pavlovian conditioning.

Timing

Temporal variables play a fundamental role in habituation and in Pavlovian and operant conditioning. Time may also be more directly involved in learning, as when animals learn to act according to the temporal attributes of a stimulus. For example, rats, pigeons, monkeys and other vertebrates can learn to behave in one manner after a 2 s stimulus and in another manner after an 8 s stimulus. When a reinforcer such as food is available periodically, say every 30 s (an example of an interval reinforcement schedule), animals learn to pause immediately after food and then, after about 15 s, respond at an increasingly faster rate until reinforcement. The study of the temporal regulation of behavior is one of the most developed areas in the study of animal learning.

A major empirical finding that has emerged from these studies is the scalar property of temporal discrimination, which states that all temporal judgments are relative. Hence, how a rat behaves at 10 s when reinforcement occurs every 30 s is similar to the way it behaves at 20 s when reinforcement occurs every 60 s. As another example, assume that a rat is trained to press a lever on the left after a 2 s signal and a lever on the right after an 8 s signal. Empirical tests show that the rat will be indifferent (i.e. it is just as likely to press the left as the right lever) when presented with a 4 s signal, because 2 is to 4 as 4 is to 8. However, if the two training stimuli were 4 s and 16 s long, then the rat would be indifferent when presented with an 8 s signal. The scalar property derives its name from the fact that when the intervals of a temporal discrimination change, the animal's performance is scaled (stretched or shrunk) by the same factor. Why the scalar property holds remains a matter of controversy.

Avoidance Learning

Although operant and Pavlovian conditioning seem clearly distinguishable, components of each are often present in the procedures of the other. In this sense, it is perhaps better to conceive of operant and Pavlovian conditioning as analogous to elemental hydrogen and oxygen. These two elements are rare in nature, but their combination in the form of water is common. Similarly, most learned behavior is a mixture in varying proportions of operant and Pavlovian conditioning.

The clearest examples of operant–Pavlovian interaction can be seen in situations where animals have to avoid aversive outcomes. A gopher that sees a hawk overhead will not wait to see what the bird will do; upon sensing the hawk, the gopher will retreat to its burrow. In the laboratory, a dog will jump over a hurdle during a tone if this response prevents the delivery of shock. In these and similar circumstances, the sight of the hawk or the sound of the tone predict an aversive outcome (being attacked or shocked) unless a certain response occurs (retreating to a burrow or jumping a hurdle). The relation between the signal and the aversive event is a Pavlovian CS–US relation; but because responding during the CS allows the animal to avoid the aversive event, this response is negatively reinforced (an operant response–consequence relation).

Relevance of Operant Conditioning

Since Thorndike's and Skinner's early work, the study of operant conditioning has been extended

in many different directions. Neuroscientists, pharmacologists and clinical psychologists, for example, have used the techniques and conceptual tools of operant conditioning to understand the functioning of the nervous system, the behavioral effects of drugs, and the intricacies of behavioral disorders such as depression. Artificial intelligence researchers also have borrowed ideas from the domain of operant conditioning to design systems that learn through the consequences of their actions. The study of operant conditioning also has become more quantitative. As in Pavlovian conditioning, real-time, dynamic models of the operant process have started to replace purely verbal accounts. However, much remains unknown. For example, although it is reasonable to assume that a consequence with survival value (or which is closely associated with a stimulus that has survival value) will have a strong effect on the response that produced it, research has yet to yield a general theory of reinforcement and punishment.

CONCLUSION

The processes reviewed above represent only a fraction of the most basic categories of the taxonomy of learning. Much else has been done, but still more remains to be investigated. Despite a century of research, three central questions of learning theory remain largely unanswered. First, how does an animal's evolutionary history constrain the sorts of things that it can learn? Second, how are processes that occur on different timescales integrated? Third, why are seemingly simple processes so complexly organized? These questions are likely to set the research agenda of learning psychologists for the next decades.

Further Reading

Abramson CI (1994) *A Primer of Invertebrate Learning: The Behavioral Perspective*. Washington, DC: American Psychological Association.

Hearst E (1988) Fundamentals of learning and conditioning. In: Atkinson RC, Herrnstein RJ, Lindzey G and Luce RD (eds) *Steven's Handbook of Experimental Psychology*, 2nd edn, vol. 2, Learning and Cognition, pp. 3–109. New York, NY: John Wiley.

Mackintosh NJ (1974) *The Psychology of Animal Learning*. New York, NY: Academic Press.

Mackintosh NJ (1983) *Conditioning and Associative Learning*. Oxford, UK: Oxford University Press.

Mazur J (1998) *Learning and Behavior*, 4th edn. London: Prentice-Hall.

Pavlov IP (1927) *Conditioned Reflexes*, translated by GV Anrep. London, UK: Oxford University Press.

Rescorla RA (1988) Pavlovian conditioning: it's not what you think it is. *American Psychologist* **43**: 151–160.

Skinner BF (1961) Selection by consequences. *Science* **213**: 501–504.

Staddon JER (2001) *Adaptive Dynamics: The Theoretical Analysis of Behavior*. Cambridge, MA: MIT Press.

Thorndike EL (1911) *Animal Intelligence*. New York, NY: Macmillan.

Williams BA (1988) Reinforcement, choice, and response strength. In: Atkinson RC, Herrnstein RJ, Lindzey G and Luce RD (eds) *Steven's Handbook of Experimental Psychology*, 2nd edn, vol. 2, Learning and Cognition, pp. 167–244. New York, NY: John Wiley.

Animal Navigation and Cognitive Maps

Intermediate article

Bruno Poucet, National Centre for Scientific Research, Marseille, France

CONTENTS

Many species have impressive spatial navigation capabilities which seem to rely on the existence of representations, or cognitive maps, of the spatial environment.

INTRODUCTION

One of the most challenging questions for contemporary neuroscience is to understand the mechanisms by which the brain processes, encodes and stores information. In this respect, our understanding of memory has considerably improved in recent decades. Such progress is largely due to the use of animal models which allow us not only to tackle the problem of memory from a comparative perspective, but also to raise empirical issues that cannot be addressed in humans.

Spatial memory is a popular model of memory. One reason is that spatial memory is ubiquitous: almost every action takes place in space and requires some form of spatial memory. Another reason is that it can easily be studied in animals that often have outstanding spatial capabilities. Consider, for instance, the memory capabilities of rats that solve the radial arm maze task (Olton, 1979). In this task, the animal has to gather food at the end of the eight arms of a radial maze (Figure 1). As arms depleted of food are not rebaited, the rat learns to avoid locations that have already been visited. Once the animal is trained, a delay can be inserted between the fourth and fifth choices: the rat is blocked at the central choice point of the apparatus and required to wait for some time before completing the trial. Even with delays lasting up to 30 min, the rat is still able to perform the last four choices without returning to arms visited before the delay. This ability implies that it has stored the memory of the locations of depleted arms and uses this memory to visit the remaining baited arms. The current interpretation of such performance is that the rat's choice is based on a cognitive map; that is, a representation of the food locations relative to the configuration of visual cues within the testing environment.

COGNITIVE MAPS AND PLACE NAVIGATION

Historically, the concept of spatial cognitive maps is important because it suggests that animals do not

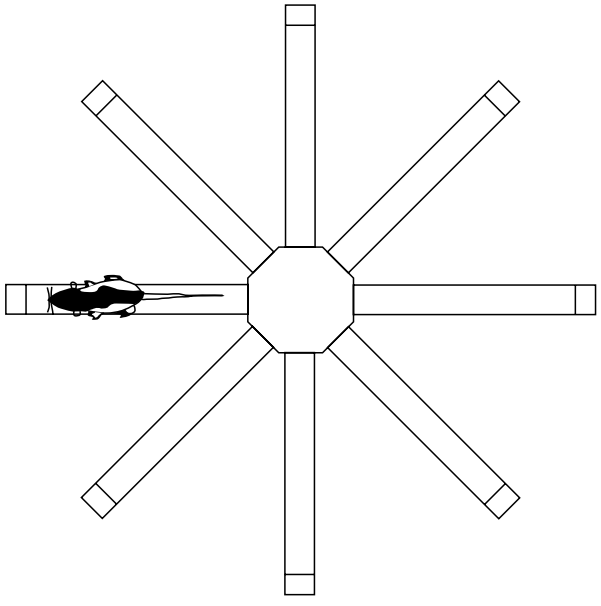


Figure 1. Overhead view of the radial arm maze. Food is located in small containers at the end of each arm and cannot be seen by the rat until it has reached the container. The task for the animal is to visit each arm without returning to an arm previously visited during the trial.

merely base their actions on specific stimulus–response associations (as strongly advocated by the early behaviorist theory in the twentieth century), but also internally reorganize acquired spatial information so as to form cognitive representations of the environment (Tolman, 1948). An important property of such representations is that they allow a reaction to stimuli that are not immediately present, since the relationship of such stimuli to those actually perceived is maintained in a cognitive representation. In other words, animals are aware of the properties of the environment beyond their field of perception. Thus, cognitive maps confer greater flexibility and efficacy to behavior. These properties depart considerably from the rigidity in behavior that results from the gradual acquisition of fixed relationships between stimuli and responses. (*See* **Tolman, Edward C.**)

A prototypical illustration of the behavioral flexibility afforded by cognitive maps is provided by the ability of rats to navigate efficiently in the water maze task (Morris, 1984). In this spatial task, the rat must find a safe platform in a pool filled with water (Figure 2). As the start position is changed from trial to trial and the platform is not visible, the animal cannot apply rigid solutions to the problem. Instead, it must rely on the visual cues located outside the swimming pool so as to infer the platform location. The rat quickly learns to swim directly towards the platform from all starting positions. More importantly, it shows immediate transfer when novel start points are used. Thus, the animal's knowledge of the platform location is independent of its current location.

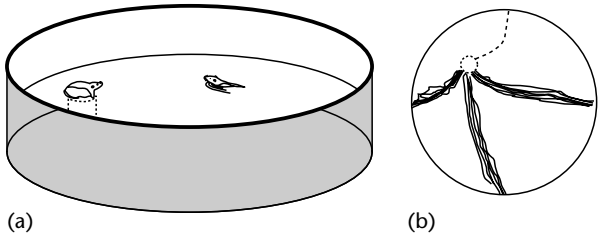


Figure 2. The water maze. (a) The pool is filled with water made opaque by the addition of powdered milk, preventing the rat from seeing the platform which lies under water level. Only distal visuospatial cues can be used to locate the goal platform. After 10 days of training, the rat performs direct trajectories. (b) Superimposed trajectories for the last six trials from the three start points used for training (black lines) and the almost direct path used by the rat when started from a start position never experienced before (dotted line).

ALLOCENTRIC CODING AND PLANNING

The rat's performance in the radial maze and water maze navigation tasks provide nice illustrations of two major functions of spatial cognitive maps. Namely, such maps are useful to memorize the positions of potential goals and to perform place navigation. These capabilities rely on an allocentric coding process, which allows the organism to memorize a location in relation to the spatial layout of surrounding landmarks, independently of its own current position.

The allocentric coding of spatial locations stands in contrast to an 'egocentric' (self-centred) coding process. This process allows an animal to memorize a location in relation to its own position. The goal location is stored in egocentric coordinates as a vector that specifies the distance and head-referred direction of the goal from the animal's current location. The egocentric coding process requires a path integration mechanism which updates the memory of the goal location as the animal moves. Although egocentric coding has the major advantage that the information to be stored is limited to two values (distance and direction of the goal relative to current position), its major drawback is that the computation of these parameters is subject to cumulative error. If the cumulative error is great enough, the animal may miss its target. Although the egocentric coding process is unreliable, its precision can be improved if the path integration information can be recalibrated from visual or other sensory information available from the environment. It is not clear, for rats or humans, if any reliance is placed on path integration except when external sensory information is missing or inadequate.

Place navigation in the water maze cannot be accounted for by an egocentric coding process, because the experimental design precludes the use of route information. So, how can we explain this performance? One possibility is that the rat takes a 'snapshot' of the cue configuration when it is on the goal platform. It could return to the platform by moving in a direction that reduces discrepancies between its current views of environmental landmarks and the views of the same landmarks from the goal. However, analyses of paths suggest that the rat does not behave in this way. The straightness of swimming paths when the rat is started from different locations in the water maze indicates that it is not just continually adjusting its current position in reference to a memorized snapshot taken at the goal; such a solution, which relies on

step by step movements, would result in more erratic movements. Rather, right from the start of its movement, the rat seems to have some knowledge of the distance and direction of its final destination, and sets its course immediately to reach the goal as soon as possible.

This ability to generate a plan requires a stored representation of the spatial relationship between the goal and rat's current position relative to the environment. This representation makes behavior relatively independent from immediately available sensory information, thus allowing adaptive changes in trajectories when the circumstances so require. Rats, cats, dogs as well as many other species are able to take short cuts when a new, less circuitous, path is made available, or to make detours around obstacles in their way to a target. For example, if a previously available path is suddenly blocked, the animal quickly reorganizes its trajectory to select the next most appropriate path leading to the goal. This ability shows that place navigation directed at a specific goal location takes into account the overall connectivity of space. (*See* **Navigation**; **Spatial Cognition, Models of**; **Animal Learning**)

EXPLORATION AND COGNITIVE MAPS

Anyone who watches a rat in a novel environment immediately notices the exploratory responses displayed by the animal. Exploratory behavior is vigorous and is presumably necessary to acquaint the animal with its new environment. As familiarization goes on, exploratory activity decreases, to stabilize at a low asymptotic level. This habituation process suggests that the animal comes to know various aspects of its environment. If the animal is deprived of the opportunity to explore its environment, it is unable to successfully solve spatial problems. Thus, exploration is required for the emergence of place navigation, presumably because cognitive maps are built during this phase through active collection of information about the spatial environment. Eventually, the map comes to match the real environment as closely as possible, therefore providing the animal with spatial invariants which allow it to perform efficient place navigation. Knowing the layout of its current space, the animal is also in a position to detect any changes that might occur. To do so, it performs routine patrolling to check the stability of the environment and to update the contents of the map if a change has been detected.

Thus, exploration may lead either to a new representation or to the updating of a former spatial representation. An illustration of the updating process is provided by studies which show that subtle changes in an otherwise familiar environment induce strong re-exploration. If such changes only involve the spatial arrangement of environmental components, it becomes possible to study the way the animal encoded the initial arrangement.

INFORMATION CONTENT OF SPATIAL COGNITIVE MAPS

Recognizing that animals are able to use representations of their spatial environment does not say much about the nature of such representations. In fact, there is some evidence that these representations might not encompass all aspects of the environment.

Configural Cues and Geometry

Based on the re-exploration technique, it can be demonstrated that changing the spatial configuration of small objects in a previously explored arena induces a renewal of exploratory activity generally aimed at the displaced objects (Thinus-Blanc *et al*., 1987). Further analyses of exploratory responses as a function of the object arrangement show that although the animal keeps a record of the spatial situation, this record is possibly specific to certain classes of spatial relationships. In general, changes that induce the strongest responses are those that affect either the overall geometrical arrangement of the object set, or the topological relationships among the objects. The configuration seems to be privileged over the absolute position of objects.

This conclusion is in agreement with the observation that performance in both the radial maze and the water maze tasks relies on configural cues rather than on individual landmarks. It also agrees with the notion that spatial geometry is an important piece of information contained in the map. In one study, rats were placed in a rectangular chamber and were required to visit the four corners of the chamber. Each corner contained a different amount of food and was associated with a distinctive visual insert. Two opposite corners had the most bait while the remaining two opposite corners had the least bait. After the animals mastered the task (i.e. visited each corner according to a decreasing order from the most baited to the least baited arm), a number of probe tests were conducted based on transformations of the initial distribution of the inserts. These tests revealed that the rats' patterns of visits to the food locations were

remarkably insensitive to the modification brought to the spatial arrangement of the inserts. Rather, the animals were using the rectangular shape of the experimental chamber as a means to locate the various food sources (Gallistel, 1990). Thus, rats ignore obvious landmarks and attend to landscape features instead in reorienting themselves in symmetric environments. Although they are aware of the landmarks, they simply do not appear to use them for certain purposes. The rules governing when landmarks or landscapes are used in behavior remain to be fully explicated. Current thinking suggests that local landmarks are not used to determine directions in space, nor to define a spatial framework within which such directions can be nested. Rather, landscape features are seemingly used for these purposes.

Distal Landmarks

Landscape features cannot be considered to be the sole markers of the spatial arrangement. In fact, distal landmarks provide another major source of information on which spatial navigation can rely. In general, animals will preferentially use such distal information when it is available, to the detriment of local cues. Distal information also predominates when it conflicts with movement-related information such as that provided by path integration mechanisms. While path integration-based homing behavior in hamsters is barely altered by the manipulation of local cues, it can be easily altered by manipulating distal information. To understand why distal cues are so effective in controlling place navigation, it is necessary to realize that only distal cues maintain their reciprocal relationships with respect to the animal during its motion. Thus the perceived reciprocal relationships between distal cues are minimally affected by an animal's movements. In contrast, the perceived reciprocal relationships between local cues are subject to strong changes during movements. Since a cognitive map contains absolute topographical information rather than information about egocentric locations relative to the animal, it becomes evident that distal cues provide a more reliable source of information for localization since they retain relatively stable relationships as the animal moves about the environment.

Path Information

Since the primary function of spatial cognitive maps is to allow efficient navigation in space, a question that arises is whether animals are aware of the structural properties of space, so that their knowledge is based on a representation of possible paths (and therefore not limited to the start and the goal). Studies in cats which had to choose between several paths all leading to the same goal location provide an unambiguous answer to this question. The paths differed in several ways, such as their length or angular deviation (i.e. how much the start of a path deviated from the goal direction). Cats displayed preferences for particular paths primarily on the basis of length (with short paths being preferred over long paths), and secondarily on the basis of their angular deviation (with paths whose starting direction was closer to the goal direction being preferred over less direct paths). Thus, spatial knowledge is not limited to representation of the start and goal, but extends to more indirect relationships provided by the rest of the environment. There appears to be an encoding of properties of the structure of the environment. The validity of the hierarchy of properties depends, however, on the goal location being hidden. If the goal is visible, the animals' choices appear to be constrained by the perceived direction of the goal from the start point. The visible goal seems to act as an anchor, shifting control over behavior from spatially based information to sensory guidance. The animals no longer make optimal choices about path length, but instead tend to take the path that most closely approximates the direction from the start to the goal.

Metrics and Topology

Evidence indicates that rats possess a representation of the geometry of their environment. The representation, however, seems not to be complete, as certain spatial relationships are better handled than others. Also, the representation is not homogeneous: animals process certain locations in a more detailed way during exploration. This is not surprising, because space is not homogeneous. It is more surprising, however, that the topological relationships among such locations (e.g. whether they are in the same vicinity) often have stronger control over spatial behavior than do metric relationships (e.g. their absolute distance in Euclidean space). It is therefore possible that spatial representations can be both topological, therefore affording relatively unstructured information about the connectivity within space (for example, place A is directly connected to place B but not to place C), and metric (for example, place A is a certain distance and direction from place B), affording more detailed information about specific relationships

among places (Poucet, 1993). The advantage of this dual format is that topological information is more rapidly acquired than metric information because metric encoding would largely rely on motion-related signals provided by repeated movements between places. However, direct empirical support for a dissociation between topological and metric encoding of spatial information is still awaited.

NEUROBIOLOGY OF SPATIAL COGNITION

Since rats remember various aspects of their spatial environment, it seems natural to look for the brain processes that underlie this ability. One locus that has received much attention in the recent years is the hippocampal formation, a structure lying below the cortex in mammals. Although other parts of the brain are known to participate in spatial behavior, the role of the hippocampus has been demonstrated on many occasions with lesion experiments. Thus, removal of the hippocampus induces dramatic and permanent deficits in a wide variety of spatial abilities. Rats with such lesions are impaired in the water maze navigation task and have an impaired spatial memory in the radial maze. Their spatial patterns of exploration are also strongly altered following hippocampal damage, and they fail to detect spatial changes in a familiar environment.

Spatial Signals in the Rat Brain

The critical evidence for the spatial function of the hippocampal formation is the existence of cells that carry a spatial signal. Such cells can be classified as 'place cells' and 'head direction cells'. (*See* **Hippocampus**)

Place cells

Place cells were first identified in the hippocampus of rats. With appropriate methods, it is possible to record the firing of a single hippocampal neuron while tracking the position of a rat as it moves freely inside a circular arena. It is then possible to display the spatial firing of the neuron by plotting a map of the number of action potentials fired by the neuron per time unit and in each location in the environment. The activity of many pyramidal cells in the dorsal hippocampus is strongly correlated with the rat's position (O'Keefe, 1976). A given place cell is virtually silent except when the animal is in a region called the firing field (Figure 3). Each place cell has its own specific firing location. Place cell discharge is often independent of the direction faced by the rat and varies only with location.

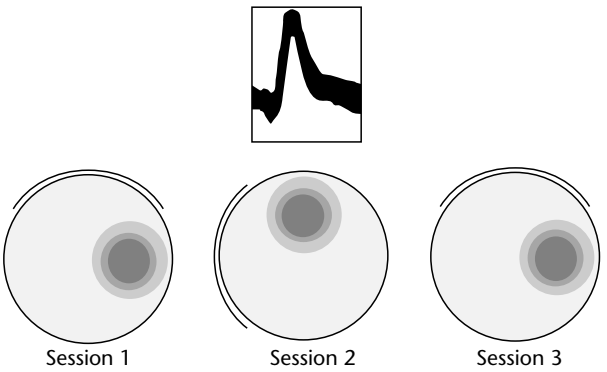


Figure 3. Firing rate maps of a hippocampal place cell during a cue rotation experiment. The cell activity is characterized by the production of action potentials whose waveforms (shown in the inset) are its signature. The cell was recorded for three successive sessions during which the rat freely moved in a cylindrical apparatus one meter in diameter (the outer circle). The firing field of the cell is shown as a set of smaller concentric circles. Darker shading indicates increasing activity of the place cell. The only available visual cue was a white cue card attached to the inner wall of the cylinder (shown as an arc). When the cue card was rotated 90° counterclockwise from the first to the second recording session, the cell firing field also rotated 90°. When the cue card was returned to its original location, the firing field also returned to its initial position. This shows that the cue card has control over the position of the field.

Two properties of place cells are important. First, when rats are placed in new surroundings, place cells become progressively active while the rat is at a given location and, once established, the locations of their firing fields are stationary for weeks and months. Second, place cell firing fields are controlled by environmental visual cues. For example, a rotation of these cues induces a corresponding rotation of firing field. Additional experiments show, however, that this control is more complex than a mere sensory triggering. Thus, if the landmark is removed, place cells usually display firing fields remarkably similar to those observed when the cue is present. This property suggests that these cells encode information about locations in the environment rather than information about sensory views of the environment (Muller, 1996). (*See* **Navigation and Homing, Neural Basis of**; **Place Cells**)

Head direction cells

Head direction cells are primarily found in the postsubiculum. However, cells with similar properties have been found in other brain areas such as the anterior and lateral dorsal nuclei of the thalamus as well as in specific cortical areas. All these

regions have extensive connections with the hippocampus. The firing pattern of head direction cells depends only on the heading of the animal, independently of its location. Each head direction cell has its own specific preferred firing direction. Both head direction cells and place cells share many properties, including being under the control of visual and idiothetic (motor-related) inputs. As for place cells, the activity of head direction cells is maintained when the environmental cues are removed. Their activity is therefore not simply visually triggered. Also, head direction cells do not function independently of each other. Changes in activity of a given cell are accompanied by corresponding changes in activity of other cells, which suggests that they are part of a tightly connected functional neural network. This network includes place cells since their firing fields react in the same manner to environmental manipulations as head direction cells. Thus, the two types of cells have access to the same information, and appear to form a tightly connected functional neural network whose function is to provide the animal with information about both its location and its heading. Their cooperative function is understood to allow the rat to navigate efficiently in the current environment.

Spatial Signals in the Primate Brain

Interestingly, cells that carry somewhat similar spatial signals have been found in the hippocampus of nonhuman primates, suggesting a comparable function in higher species. In addition, evidence based on functional neuroimaging of brain activity during navigation in familiar yet complex virtual-reality environments suggests that the human hippocampus also has a special role in spatial navigation. Activation of the right hippocampus is strongly associated with knowing accurately where places are located and navigating between them (Maguire *et al*., 1998). In addition, specific regions of the hippocampus are found to be enlarged in London taxi drivers, who need to make extensive use of their navigation abilities. This finding bears a direct relation to animal studies and supports the hypothesis that the role of the hippocampal formation in spatial memory should be extended to other mammalian species including humans.

Contribution of the Cerebral Cortex

Clear-cut spatial deficits are found after specific cortical lesions. Rats with parietal cortical lesions are impaired in maze learning, place navigation, spatial working memory and response to spatial novelty, while having no gross visual impairment. However, the magnitude of these deficits is usually smaller than that produced by hippocampal damage. Thus, although the involvement of the parietal cortex in spatial processing is undoubted, its specific contribution is not yet understood. The situation is complicated because other cortical areas are important in spatial processing, as shown by lesion and electrophysiological studies. For example, damage to the posterior cingulate cortex produces a severe spatial deficit in place navigation, even though spatial memory does not seem to be affected. Lesions of the medial frontal cortex result in spatial navigation impairments, although animals are not impaired in their response to spatial novelty. Rather, their deficit is best explained as caused by impaired working memory, which would preclude them from appropriately planning their trajectories. Further work will be necessary, however, to understand how these different cortical areas interact with the hippocampal formation to provide the organism with a spatial representation useful for navigation.

CONCLUSION

Although the navigation capabilities of mammals are impressive, they are not unique since they are found in many other species as well. Consider, for instance, the food-storing birds. During autumn, they can store seeds in hundreds of locations scattered throughout their home range, and yet retrieve them several months after the storing episode. Similarly, the homing ability of pigeons is a well-described behavioral system that allows birds to orient efficiently within huge territories. Even phylogenetically lower species such as bees have outstanding spatial navigation abilities. Do all these organisms rely on a cognitive map to find their way in their environment ? Answering this question would first require that cognitive maps are unambiguously defined. While the literature is replete with the concept, there is still uncertainty about what it means exactly. In fact, while there is little doubt that many species are able to construct internal models of their environment, the spatial extent of such representations is still a questionable issue. For example, are animals able to represent remote portions of space and to use new routes when there is no overlap in the perception of the landmarks available at the origin and goal of the trajectory? So far, this ability seems to be specific to humans. However, uncovering the

premises of such an ability in nonhuman animals is likely to help us understand how our brain creates complex representations of the world.

References

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Maguire EA, Burgess N, Donnett JG *et al* (1998) Knowing where and getting there: a human navigation network. *Science* **280**: 921–924.

Morris RGM (1984) Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods* **11**: 47–60.

Muller RU (1996) A quarter of century of place cells. *Neuron* **17**: 813–822.

O'Keefe J (1976) Place units in the hippocampus of the freely moving rat. *Experimental Neurology* **51**: 78–109.

Olton DS (1979) Mazes, maps and memory. *American Psychologist* **34**: 583–596.

Poucet B (1993) Spatial cognitive maps in animals: new hypotheses on their structure and neural mechanisms. *Psychological Review* **100**: 163–182.

Thinus-Blanc C, Bouzouba L, Chaix C *et al*. (1987) A study of spatial parameters encoded during exploration in hamsters. *Journal of Experimental Psychology: Animal Behavior Processes* **13**: 418–427.

Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* **55**: 189–208.

Further Reading

Clayton NS (1998) Memory and the hippocampus in food-storing birds: a comparative approach. *Neuropharmacology* **37**: 441–452.

Hermer L and Spelke E (1994) A geometric process for spatial orientation in young children. *Nature* **370**: 57–59.

McNaughton BL, Barnes CA, Gerrard JL *et al* (1996) Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *Journal of Experimental Biology* **119**: 173–185.

O'Keefe J and Nadel L (1978) *Hippocampus as a Cognitive Map*. Oxford, UK: Clarendon.

Poucet B, Save E and Lenck-Santini PP (2000) Sensory and memory properties of place cells firing. *Reviews in the Neurosciences* **11**: 95–111.

Roitblat HL (1987) *Introduction to Comparative Cognition*. New York, NY: Freeman.

Taube JS (1998) Head direction cells and the neurophysiological basis for a sense of direction. *Progress in Neurobiology* **55**: 225–256.

Thinus-Blanc C (1996) *Animal Spatial Cognition: Behavioral and Neural Approaches*. Singapore: World Scientific.

Trullier O, Wiener SI, Berthoz A and Meyer JA (1997) Biologically based artificial navigation systems: review and prospects. *Progress in Neurobiology* **51**: 483–544.

Wallraff HG (1996) Seven hypotheses on pigeon homing deduced from empirical findings. *Journal of Experimental Biology* **199**: 105–111.

Antisocial Personality and Psychopathy

Introductory article

Peter R Finn, Indiana University, Bloomington, Indiana, USA

CONTENTS

Antisocial personality is a mental disorder that begins in childhood and continues into adulthood, and involves the persistent violation of the rights of others and social norms in general. Deficits in emotional, motivational, and cognitive processes contribute to the development of the disorder.

INTRODUCTION

Antisocial personality is a personality disorder that affects about 3% of men and 1% of women. Personality disorders are mental disorders that reflect fundamental problems in the early development of personality, resulting in enduring psychological and behavioral problems that begin in childhood or adolescence and continue into adulthood. Antisocial personality is characterized by a range of symptoms that involve difficulties controlling impulses, forming caring interpersonal attachments, learning from experience, and experiencing guilt or remorse. The specific symptoms of antisocial personality disorder include deceitfulness, stealing, destroying property, aggression, impulsivity, engaging in unlawful and other rule-breaking behavior, consistent irresponsibility, and a disregard for others. Conduct disorder is the childhood precursor to adult antisocial personality.

Although antisocial personality is defined as a specific disorder, there are different subtypes, such as primary and secondary psychopathy (or sociopathy). In primary psychopathy antisocial behavior is associated with a deficit in the capacity to experience negative emotions such as fear, anxiety, and guilt, and a superficial and manipulative manner of relating to others. Secondary psychopathy involves high levels of impulsivity and negative affect (guilt and anxiety) combined with a pattern of chronic antisocial behavior. Regardless of the variations in the pattern of symptoms, antisocial individuals demonstrate a failure to inhibit behavior that violates others, transgresses social norms, or increases the risk for generally negative consequences to self or others. Research indicates that emotional, motivational, and cognitive deficits all contribute to the inhibitory problems experienced by those with antisocial personality.

EMOTIONAL PROCESSES

The Role of Emotion in Self-control and Socialization

Because people generally want to avoid experiences that result in fear, anxiety, guilt, or shame, these emotional experiences play an important role in facilitating self-control and socialization. For instance, experiencing anticipatory anxiety when one considers engaging in a specific behavior that might lead to a negative outcome should bolster self-control and lead to an inhibition of that behavior. Also, a common method of socialization involves using the threat of punishment to manipulate fear and anxiety as a means of encouraging or enforcing conformity.

Emotion also has an important role in regulating interpersonal behavior through the experience of sympathy and empathy. The process of experiencing, or sharing, the negative emotion of another contributes to a sense of identification with that person. Since identification involves incorporating aspects of the other into one's own self construct, people are unlikely to do harm to someone with whom they identify because they would be symbolically doing harm to themselves in the process.

Deficits in the ability to experience fear, anticipatory anxiety, or negative emotions in general would compromise the processes that contribute to self-control, socialization, and the regulation of interpersonal behavior.

Emotional Deficits in Psychopathy

A key feature of primary psychopathy is a lack of fear and arousal. Research on the skin conductance response (SCR), a physiological measure of arousal, indicates that people with primary psychopathic disorder have a deficit in the ability to experience anxiety as a response to threatening stimuli. The SCR reflects changes in the level of electrical conductivity of the skin on the palm of the hand resulting from activity of the sweat glands, which are influenced solely by the sympathetic nervous system. In studies where SCRs are classically conditioned to stimuli that signal aversive events, such as electric shock, people with primary psychopathic disorder consistently show smaller SCRs when compared with those with secondary psychopathy and normal volunteers. Research also shows that they have smaller anticipatory SCRs when waiting for a noxious stimulus to be administered (Figure 1).

Another hallmark of primary psychopathy is a deficit in caring, loving, and showing empathy for others. People with this disorder show a peculiar pattern of emotional detachment, in which they can correctly label the emotions of another with words, but do not experience the emotion. Such people show blunted physiological responses to pictures of mutilated accident victims but can correctly describe the pictures as depicting unpleasant scenes.

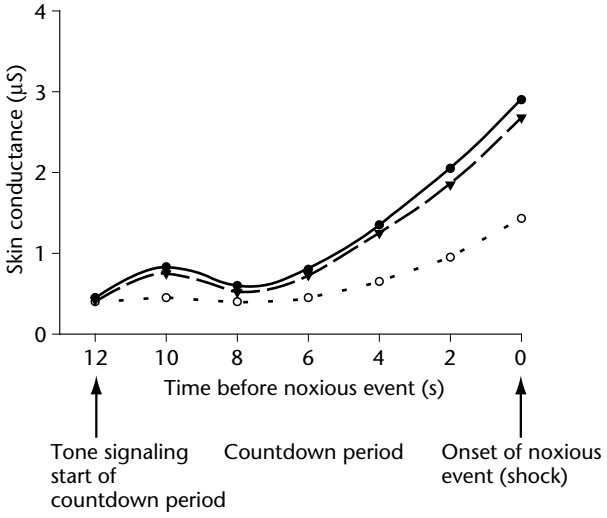


Figure 1. The typical pattern of anticipatory skin conductance while anticipating a noxious stimulus in people with primary psychopathy (open circles), secondary psychopathy (triangles), and normal controls (solid circles).

This same pattern of emotional detachment has been observed in patients who developed psychopathic traits after sustaining damage to the frontal lobes of the cortex of their brain.

IMPULSIVITY

Impulsivity is associated with a wide range of problems, including antisocial personality, substance abuse, suicide, overeating, gambling, and aggressive behavior. Rather than indicating a single trait, impulsivity refers to a range of characteristics that include preferences for immediate versus delayed gratification, acting without thinking, acting quickly, paying more attention to the present than the future, having poor inhibitory control, and being overresponsive to reward. These characteristics promote behavior that often leads to unforeseen negative consequences to self and others, and behavior that disregards the rights, feelings, and safety of others in the service of immediate gratification for self. Both primary and secondary psychopathy are associated with impulsivity, but impulsive behavior is associated with different qualities and probable causes in the two subtypes of antisocial personality. In secondary psychopathy, impulsivity is often associated with negative emotions, cognitive deficits (lack of planning and future orientation), and poor inhibitory control. Impulsive behavior in primary psychopathy is probably due more to a lack of concern about negative outcomes and a greater interest in obtaining immediate gratification of impulses.

Laboratory studies indicate that when given the opportunity to choose between an immediate smaller monetary reward or a larger, but delayed, reward, individuals with antisocial or impulsive traits display a greater preference for the immediate reward and tend to discount the value of future rewards. Other studies indicate that impulsive personality traits and an inability to tolerate delay of gratification are associated with lower levels of brain serotonin, a neurochemical found in areas of the brain associated with emotion and basic motivational functions.

MOTIVATIONAL PROCESSES

Antisocial personality is associated with abnormalities in the motivational processes underlying approach and avoidance behavior. In this broad theoretical perspective, antisocial personality is the result of strong approach tendencies that override the inhibitory influences that affect avoidance

behavior. For example, excessively strong desires for wealth, power, dominance, or sexual gratification (i.e. strong approach motivations) may override the normal inhibitory effects that social sanctions (e.g. arrest, humiliation, rejection) have on attempts to unlawfully gain these rewards. On the other hand, antisocial personality could also result from an insensitivity to punishment that results in weak inhibitory influences. In this scenario, individuals commit antisocial acts because they do not care about – or do not pay attention to – the possibility that their behavior may be punished. Research suggests that both perspectives are true: some with antisocial personality are less sensitive to punishment (weak avoidance), some are more sensitive to reward (strong approach), and still others have deficiencies in processing and attending to information about the potential for reward and punishment.

Response to Reward

The majority of studies of motivational deficits in antisocial personality use go/no go tasks where the person is required to learn when to make a response (go trial) and when to inhibit a response (no go trial). The studies manipulate the rewards and punishments for responses and nonresponses on each type of trial. Typically 'go' responses result in rewards (winning money) and 'no go' responses (failures to inhibit) result in punishment (losing money). In these types of laboratory studies antisocial individuals consistently show deficits in the ability to learn to inhibit their response on no go trials, termed a passive avoidance learning deficit. Figure 2 illustrates the typical pattern of passive avoidance deficits in antisocial personality. Additional research indicates that these passive avoidance learning deficits are only apparent when the antisocial person is trying to obtain a reward, and when the reward stimuli are more salient (i.e. approach behavior is the dominant response). Other research indicates that antisocial individuals are more sensitive to rewards in general. Overall, these studies indicate that when actively seeking out rewards, those with antisocial personality have difficulty inhibiting their behavior to avoid negative consequences. In other words, they are less able to develop an optimum strategy for behavior that maximizes positive outcomes while minimizing negative consequences.

Response to Punishment

Studies that use aversive consequences, such as electric shock as punishments, indicate that people

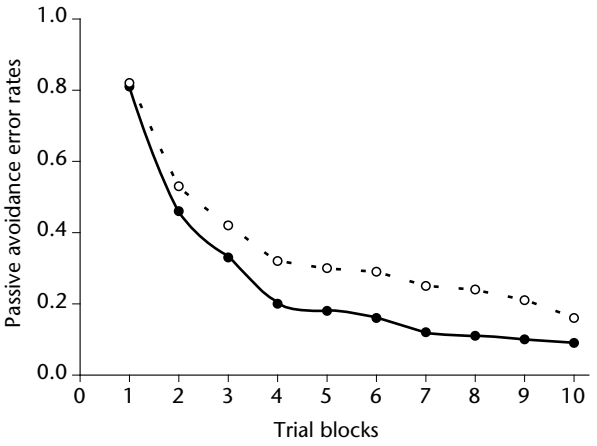


Figure 2. The typical pattern of passive avoidance learning for people with antisocial personality (open circles) and normal participants (solid circles) in a go/no go task where passive avoidance errors (responses on no go trials) result in losing money (e.g. 20 cents) and correct responses on go trials win the same amount of money. The learning curves reflect gradual reductions in the error rates within blocks of four no go trials.

with primary psychopathic disorder are uniquely insensitive to punishment. Although few studies use noxious stimuli as punishments, these studies indicate that such people show substantial deficits in their ability to learn to avoid electric shocks. Rather than being associated with increased response to reward, insensitivity to noxious punishments is associated with a fearless, low harm-avoidant temperament.

COGNITIVE PROCESSES

Cognitive processes such as reflection, planning, and problem-solving have a central role in self-regulation and are important in the motivational deficits and impulsive traits of those with antisocial personality. When abilities to reflect, plan, or solve the problems of strategizing for behavior are deficient, the immediate situation tends to control a person's behavior. In such situations, people tend to act according to their strongest feelings, impulses, and concerns, and fail to consider the effect of their behavior on others or the future. These types of cognitive deficits are associated with the disruptive, socially deviant, and aggressive behavior of many with antisocial personality.

Abnormal, or underdeveloped, frontal cortical lobes are thought to contribute to the cognitive deficits observed in those with antisocial personality. Those with damage to the dorsolateral and orbital areas of the prefrontal portion of the frontal

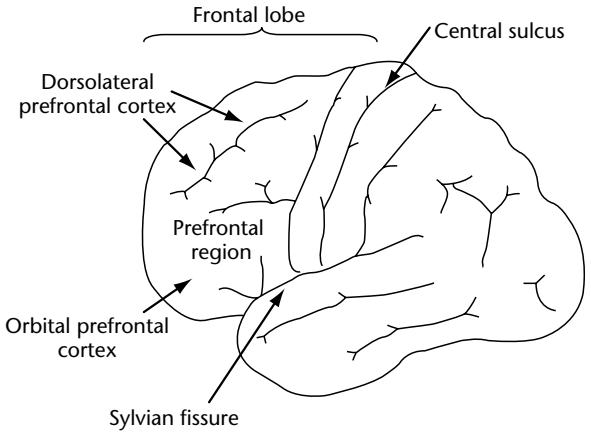


Figure 3. Lateral view of the human brain illustrating the location of specific regions in the frontal lobe of the left hemisphere. The frontal lobe is bounded by the central sulcus and the sylvian fissure.

lobes (Figure 3), display both the antisocial behavior and cognitive deficits observed in antisocial personality. In fact, those with antisocial personality show evidence of cognitive deficits on the neuropsychological tests that were developed to measure the deficits resulting from frontal brain damage. Brain scanning studies also indicate that people with chronic antisocial behavior show evidence of abnormal frontal lobes.

Cognitive Deficits in Antisocial Personality

Deficits in verbal ability and executive cognitive functions are the most commonly reported cognitive abnormalities associated with conduct disorder and antisocial personality. Executive cognitive functions refer to the cognitive abilities involved in the planning and execution of effective goal-oriented behavior. Executive functions include abstract reasoning, planning, associative learning, problem-solving, attentional processes, self-monitoring, set-shifting, and working memory. Deficits in verbal ability and executive functions are associated with a greater severity in antisocial symptoms and poor outcomes in children with a history of antisocial behavior.

Verbal ability

Verbal and language deficits associated with antisocial traits include low levels of fluency (ease of finding words to label experience or concepts), limited vocabulary, poor language expressive skills, problems in understanding and following verbal instructions, reading problems, and poor concept formation.

Executive functions

Antisocial persons perform most poorly on laboratory tests of executive function that tap: (a) the ability to learn arbitrary stimulus–response associations and rules for behavior; (b) the ability to change behavior to adapt to changes in the rules for appropriate behavior (i.e. response set-shifting); and (c) the ability to plan behavior in complex tasks that require the development of strategies for effective performance. In addition, on complex decision-making tasks that tap these executive processes and manipulate rewards and punishments, both antisocial personality and damage to the orbital prefrontal cortex are associated with decisions that favor a larger immediate reward even when it leads to long-term losses. This disadvantageous decision-making style reflects a preference for immediate gratification and a lack of consideration for long-term consequences.

CONCLUSION

Antisocial personality is a disorder associated with a range of symptoms and apparent causes. The factors that contribute to the development of antisocial personality are deficits in the capacity to experience fear and anxiety, preferences for immediate versus long-term rewards, increased sensitivity to reward, deficits in the capacity to inhibit previously rewarded behavior that leads to long-term negative consequences, and cognitive deficits that compromise the ability to reflectively learn and develop effective strategies for goal-directed behavior.

Further Reading

Bechara A, Damasio AR, Damasio H and Anderson SW (1994) Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**: 7–15.

Cleckley H (1982) *The Mask of Sanity*. St Louis, MO: Mosby.

Evenden JL (1999) Varieties of impulsivity. *Psychopharmacology* **146**: 348–361.

Fowles DC (1993) Electrodermal activity and antisocial behavior: empirical findings and theoretical issues. In: Roy JC, Boucsein W, Fowles DC and Gruzelier JH (eds) *Progress in Electrodermal Research*, pp. 223–237. New York, NY: Plenum Press.

Hare RD (1993) *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York, NY: Pocket Books.

Lykken DT (1995) *The Antisocial Personalities*. Hillsdale, NJ: Erlbaum.

Mazas CA, Finn PR and Steinmetz JE (2000) Decision making biases, antisocial personality, and early-onset alcoholism. *Alcoholism: Clinical and Experimental Research* **24:** 1036–1040.

Moffitt TE (1993) The neuropsychology of conduct disorder. *Development and Psychopathology* **5**: 135–151.

Newman JP and Schmitt WA (1998) Passive avoidance in psychopathic offenders: a replication and extension. *Journal of Abnormal Psychology* **107**: 527–532.

Patrick CJ, Bradley MM and Lang PJ (1993) Emotion in the criminal psychopath: startle reflex modulation. *Journal of Abnormal Psychology* **102**: 82–92.

Anxiety Disorders

Introductory article

Patrick A Palmieri, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA
Wendy Heller, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

CONTENTS

Anxiety is an emotional state generally characterized by fear and worry. When anxiety levels become excessively high owing to a combination of biological, psychological and social influences, one or more anxiety disorders may be diagnosed.

INTRODUCTION

Anxiety is one of the most important topics in abnormal psychology. It is an emotional state that all of us experience at some point in our lives. For some individuals anxiety reaches excessive levels, resulting in significant impairment of functioning in several areas of life – occupational, academic and social. Depending on the specific symptoms shown, such psychopathology would probably be diagnosed as one or more of the anxiety disorders listed in the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV). These psychological disorders are described primarily by maladaptive levels of anxiety and associated emotional responses such as fear, panic and worry. Several effective treatment options that alleviate the human suffering due to these conditions are available. These interventions are designed to target the biological, psychological or social aspects of the anxiety disorders. (*See* **Emotion**)

What is Anxiety?

Imagine you are walking across a stage to a podium where you will soon deliver a keynote address to a large audience of respected members of your community. How do you feel? What is going through your mind? In such a situation, most of us would experience one or more of the following: a sense of uneasiness, upset stomach, sweating, trembling, worry about the audience's reaction to your speech, rapid heartbeat, and other similar phenomena. These are all examples of symptoms of anxiety.

Anxiety *per se* is not a bad thing. It is an adaptive response: a low or even moderate level of anxiety would serve to motivate you to prepare for your speech. In evolutionary terms, an appropriate fear response to a threatening situation enhances the likelihood of taking an action that would promote survival. When anxiety levels become excessively high, however, they become maladaptive, often leading to persistent problems in various areas of life. For example, individuals who are deathly afraid of flying (a common phobia) may encounter difficulties in pursuing their career. (*See* **Evolutionary Psychology: Theoretical Foundations**)

Related Constructs

Any discussion of the nature of anxiety must address the related but distinct concepts of fear, panic, worry, obsessions and compulsions. Anxiety can be described in terms of categorical disorders, as defined by the DSM-IV. It can also be described in terms of symptom dimensions such as worry. Worry (anxious apprehension) is a future-oriented emotional response. It involves cognitive activity that is best described as uncontrollable negative thoughts about future potentially threatening events. Obsessions are repetitive, intrusive, anxiety-provoking thoughts. Whereas worry is typically brought on by everyday difficulties, obsessions come more 'out of the blue'. Compulsions are repetitive behaviors engaged in to reduce or avoid anxiety caused by obsessions. Fear (anxious arousal), which involves physiological hyperarousal, is an emotion experienced in the face of immediate danger that serves to prepare the individual to react adaptively to threatening situations. Panic, on the other hand, is essentially a fear response that can happen at an inappropriate time – that is, when no real danger is present. All of these constructs are combined in various ways in the anxiety disorders.

The phenomenon of anxiety is also closely linked to depression. In fact, there is some debate about whether they are distinct constructs, partly because of their high rate of co-occurrence. David Watson and Lee Anna Clark's tripartite model of anxiety and depression posits that negative affect – a general distress dimension – is a component of both anxiety and depression. This general distress may be responsible for the high comorbidity and also may signify a common etiology. Their model also proposes unique aspects of depression (low positive affect) and anxiety (physiological hyperarousal). Jack Nitschke and colleagues have found that negative affect is distinct from anxious arousal and anxious apprehension, both of which, according to their research, were robust factors in their own right. Thus, although depression and anxiety are commonly found together, it appears that each has separable components. (*See* **Depression**)

CLASSES OF ANXIETY DISORDER

History of Anxiety Classification

Sigmund Freud and his followers provided some of the earliest clinical descriptions of anxiety. Because they believed the underlying causes of pathological anxiety were similar, they did not place much value on differentiating types of anxiety. This conceptualization of anxiety persisted through the first two editions of the DSM, in which all forms of anxiety were lumped into one category called 'neuroses'. Over time, however, this method of classification was discontinued, because the psychoanalytic principles on which it was based fell from favor.

The DSM-III marked a major change in the classification of psychopathology. This classification system was based on clinically descriptive symptoms of abnormal psychology, rather than on unproved theories of their etiology. Thus, the category of neuroses was eliminated, and several new categories were formed based on clinical features. Several of these categories were contained under the newly created class of anxiety disorders. Further differentiation of the anxiety disorders is found in the fourth edition of the DSM. Its class of anxiety disorders contains several related but distinguishable psychological disorders.

Specific Phobia

Specific phobia involves a persistent and excessive fear elicited by the presence or anticipation of a particular object or situation. This is the most common anxiety disorder, occurring in approximately 9 percent of the adult population in any given year. Some common phobic objects or situations are spiders, snakes, flying and heights. This diagnosis requires that the individual actively avoid the stimulus, and if avoidance is not possible, that exposure to the object or situation evoke an immediate fear response. Importantly, the fear must be irrational. Fear of truly dangerous situations is adaptive, and therefore not abnormal. The fear and avoidance also must be of sufficient severity to disrupt performance in important areas of the person's life, such as at work or in personal relationships.

Social Phobia

Social phobia involves the fear and avoidance of social situations, such as speaking in public or attending a party. It is similar to specific phobia, but often involves performance (e.g. speaking in front of a group) and the fear of being evaluated negatively, causing embarrassment or humiliation.

Agoraphobia

Another form of phobic anxiety is agoraphobia, a fear of public spaces. Commonly feared situations of the agoraphobic individual are being in a crowded supermarket or theater, or traveling on public transport. The main source of fear is the feeling that one will be unable to escape the situation. People suffering from agoraphobia feel increasingly uncomfortable the farther they are from a safe place, such as home. In its most severe form, the individual might be unable to leave the house.

Panic Disorder

Panic disorder is defined primarily by the presence of recurrent and unexpected panic attacks. A panic attack is an overwhelming experience of fear that develops suddenly 'out of the blue', reaches peak intensity quickly, and is described mostly in terms of physical or somatic symptoms such as rapid heart rate, sweating, dizziness, nausea, trembling, chest pain or discomfort, or chills or hot flashes. Cognitive symptoms also may be present, such as thinking that one is dying or losing one's mind. A misinterpretation of somatic experiences may be largely responsible for panic attacks. For example, a rapid heart rate may be misinterpreted as an impending heart attack (even immediately following aerobic exercise, such as running up

a flight of stairs), and this conclusion can trigger a panic attack. To meet the criteria for panic disorder, panic attacks must be followed by a significant period during which there is worry about the effects of a panic attack or the possibility of future panic attacks, or a change in behavior (e.g. avoidance) designed to prevent future attacks. Often, panic attacks are experienced in settings like those feared in agoraphobia. Because of this, there are two types of panic disorder: with agoraphobia and without agoraphobia.

Obsessive–Compulsive Disorder

Obsessive–compulsive disorder involves the presence of obsessions or compulsions (but usually both) that the individual recognizes as being excessive. The obsessions cause significant levels of distress and anxiety. Two common obsessional themes are contamination (e.g. disease, germs) and loved ones being involved in an automobile accident. Compulsions are repeated behaviors designed to reduce or avoid the anxiety caused by obsessions. One common compulsion is excessive hand-washing; in its most severe form, the individual washes so much that the hands start to bleed. Another common compulsion is excessive checking: for example, an individual might get out of bed dozens of times before falling asleep to check repeatedly that the front door is locked. An individual suffering from this disorder cannot stop engaging in these repetitive behaviors, even though attempts are made to resist them.

Generalized Anxiety Disorder

Generalized anxiety disorder is mainly characterized by excessive worry that is difficult to control and that precipitates impairment in one or more areas of life, such as personal, academic or occupational functioning. The worry must be related to numerous concerns and be present consistently for at least 6 months. In addition, the concerns should not be related solely to events associated with other anxiety disorders, such as having one's performance evaluated negatively, worrying about panic attacks, or fearing embarrassment in a social situation.

Posttraumatic Stress Disorder

Epidemiological studies have shown that most people experience or witness at least one traumatic event in their lifetime. A trauma is a stressful situation involving actual or threatened death or serious injury, or a threat to the physical integrity of self or others; it also must evoke in the individual a sense of intense fear, helplessness, or horror. Examples of such traumatic events are combat experiences, sexual assault and serious motor vehicle accidents. Sometimes such experiences result in a wide variety of physical and psychological symptoms. Posttraumatic stress disorder (PTSD) is diagnosed when those symptoms include, but are not necessarily limited to, reexperiencing or reliving the trauma (flashbacks, nightmares), persistent avoidance of stimuli associated with the trauma (avoiding people, places or activities that evoke memories of the trauma) and increased arousal (difficulty falling or staying asleep, exaggerated startle response). Once the symptoms appear, they must be present for at least 1 month. Interestingly, the symptoms do not always appear immediately after the traumatic experience; instead, onset may be delayed for months or even years. (*See* **Post-traumatic Stress Disorder**)

Acute Stress Disorder

Acute stress disorder is similar to PTSD but with two important differences. First, there is the additional symptom requirement of dissociation during or following the trauma. Dissociative symptoms include such experiences as being unable to recall important aspects of a trauma (dissociative amnesia), feeling detached from or an outside observer of one's own body (depersonalization), perceiving the external world as unreal or dreamlike (derealization), experiencing a reduction in awareness of surroundings, and having a sense of emotional detachment or numbing. Second, symptoms must start within 4 weeks of the traumatic event and must last no longer than 4 weeks. If they persist, then the diagnosis would be converted to PTSD.

ETIOLOGY OF ANXIETY

Anxiety is a complex biopsychosocial phenomenon and thus has no single type of causative factor. In some cases anxiety disorders originate in some stressful life event or series of events that is experienced by the individual as unpredictable or uncontrollable. These events can include social situations that involve fear or perceived physical or emotional threat. The presence of such events interacts with psychological and biological factors to lead to the development and maintenance of anxiety disorders.

Psychological Factors

An abundance of evidence indicates that phobias can develop through classical conditioning, a learning mechanism described by Ivan Pavlov. This is the process in which a previously neutral stimulus, when paired with an unconditioned stimulus (US) that evokes fear (unconditioned response, UR), can acquire the ability (conditioned stimulus, CS) to elicit the fear response (conditioned response, CR) even in the absence of the US. Martin Seligman proposed the idea of preparedness, which states that we are biologically prepared to learn certain CS–US associations very easily, sometimes with only one pairing, because there is evolutionary advantage in doing so. For example, relevant to specific phobias, humans seem hard-wired to develop a conditioned fear response to snakes. (*See* **Conditioning**)

Conditioning also appears to be critical for the development and maintenance of PTSD. Consider the example of a soldier in a fear-evoking combat situation. When stimuli such as loud sounds of machine-gun fire are repeatedly paired with the fear of being in combat, similar loud noises become able to elicit the fear response in noncombat contexts (classical conditioning). This fear is then maintained through operant conditioning, a learning theory introduced by B. F. Skinner, that states that behavior is a function of its consequences (i.e. the frequency of a behavior increases if it is reinforced and decreases if it is punished). Avoidance of fear-evoking situations reduces one's anxiety level. This decrease in anxiety reinforces, or makes more likely, the avoidance behavior, thereby maintaining the post-traumatic stress response. (*See* **Reward, Brain Mechanisms of**)

Information-processing models of anxiety identify cognitive biases as potential etiological and maintenance factors of anxiety disorders. A wealth of research using a variety of cognitive tasks shows that anxious individuals exhibit a selective attention towards threatening stimuli. This attentional bias may result in certain stimuli becoming more salient, thereby affecting the individual's assessment of the level of threat. An increase in perceived threat leads to heightened anxiety which, in turn, can further increase the salience of the stimuli and the anxiety level, in a self-perpetuating cycle. David Clark uses such a cognitive approach to model the etiology of panic disorder. An individual's heart might skip a beat, which might elicit an anxiety response (including the typical physiological symptom of increased heart rate), thereby making the heartbeat more salient, leading to the misinterpretation that a heart attack is imminent. This cycle may continue until it ultimately induces a panic attack. (*See* **Information Processing**; **Selective Attention**)

Genetic Factors

Several twin and family behavioral genetics studies have helped elucidate the role of genes in the development of anxiety disorders. In one such study Russell Noyes found that relatives of patients with panic disorder had higher levels of panic disorder than the general population, suggesting the possibility that a predisposition to this disorder is inherited. Relatives did not have elevated rates of generalized anxiety disorder, however, suggesting that these disorders have some unique etiological factors. A family study of obsessive–compulsive disorder conducted by Donald Black and colleagues suggested that what is inherited is a general predisposition toward developing any anxiety disorder, rather than a specific predisposition for developing OCD. (*See* **Behavior, Genetic Influences on**)

Because family behavioral genetic studies are confounded by shared environment, twin studies are needed to provide converging evidence of the role of genes in anxiety disorders. A large twin study of several anxiety disorders conducted by Kenneth Kendler and colleagues revealed that concordance rates were higher for monozygotic (MZ) or identical twins than for dizygotic (DZ) or fraternal twins but were not exceptionally high, suggesting both moderate genetic and environmental etiological influences. In a separate twin study, William True and colleagues found that concordance rates for experiencing traumatic events and for developing PTSD were higher for MZ than for DZ twins. This genetic effect may be acting through psychological characteristics known to increase risk for trauma and PTSD, such as antisocial personality. In addition to genetic factors, there are other biological factors that may contribute to the etiology of anxiety disorders.

TREATMENT OF ANXIETY

Given the magnitude of anxiety problems as evidenced by epidemiological studies, it is fortunate that several methods of effective treatment exist. These treatments are designed to intervene by affecting the psychological and biological mechanisms responsible for causing and maintaining anxiety. Depending on the nature of the anxiety disorder and possible comorbid conditions, the

optimal treatment might be psychotherapy, pharmacotherapy, or a combination of the two.

Psychotherapy

Behavioral and cognitive-behavioral interventions involving exposure to feared and avoided objects or situations enjoy the strongest empirical support among psychotherapeutic treatments, despite their seemingly counterproductive nature. Joseph Wolpe developed a specific exposure technique called 'systematic desensitization'. In this intervention the first step is to teach relaxation techniques. Then the patient and therapist collaboratively construct a hierarchy of the patient's fears, arranged from least to most anxiety-provoking. Next, the patient imagines the least provoking situation while maintaining the relaxation response. Exposure is repeated or maintained until the imagined situation no longer evokes the fear response. This sequence is repeated with each step up the hierarchy. Empirical evidence indicates that this strategy is very effective in treating phobias, and the treatment gains are maintained after treatment is terminated. In the treatment of obsessive–compulsive disorder, the exposure component is combined with response prevention. Patients are exposed to anxiety-provoking thoughts (their obsessions) and prevented from engaging in the compulsive behavior typically used to reduce their anxiety. This prevention component is necessary for maintaining exposure to the anxiety-provoking situation. (*See* **Imagery**)

Cognitive therapy is another common approach to treating anxiety disorders and is often used in conjunction with the behavioral technique of exposure. The therapist helps the patient identify illogical thoughts (e.g. exaggerated estimates of the likelihood of various negative outcomes of a situation) that may be connected to the anxiety response. For instance, a patient with social phobia might believe the likelihood of a social interaction going poorly is 90 percent. It is possible that this is an overestimate of the actual probability. Having the patient record the outcomes of a number of interactions can be a useful way of gathering evidence to convince the patient that the estimate is exaggerated. Hopefully, the ensuing reduction in the probability estimate will help decrease the fear associated with the situation. Edna Foa and Barbara Olasav Rothbaum have developed an effective cognitive-behavioral treatment for patients with PTSD following rape. It includes, among other things, imaginal exposure to the traumatic event, cognitive restructuring and relaxation training. Another example of an effective cognitive-behavioral intervention is David Barlow's approach to the treatment of panic disorder. In addition to relaxation and exposure it includes a cognitive component to address faulty logic such as jumping to conclusions, overgeneralizing, all-or-none thinking and grossly exaggerating 'worst case' scenarios.

Antianxiety Medications

The benzodiazepines are a commonly prescribed class of medication for alleviating anxiety symptoms. Examples of these tranquilizers are diazepam and alprazolam. They work by binding to receptor sites in the brain for the neurotransmitter γ-aminobutyric acid (GABA) and inhibiting the activity of GABA neurons. They seem to reduce many somatic symptoms of anxiety but are less effective at reducing cognitive symptoms such as worry. There is some empirical evidence for the effectiveness of these drugs for treating social phobia, panic disorder and (ironically) generalized anxiety disorder. On the other hand, they do not seem to provide much relief for individuals suffering from other phobias and obsessive–compulsive disorder. (*See* **Psychoactive Drugs**)

Although there is moderate success for this drug treatment, there are several disadvantages. Common side effects include motor and cognitive impairments. In addition, it is common for gains to be lost once the medication is discontinued. Perhaps most importantly, though, there is a risk of addiction. This risk is highest for individuals who have a history of abusing alcohol and other substances. Unfortunately, such histories are common among people with anxiety disorders, who often self-medicate themselves with these substances to alleviate their anxiety symptoms.

Given the high comorbidity of anxiety and depression it is perhaps not surprising that antidepressant medications show some effectiveness in treating anxiety disorders. This improvement might be due to the alleviation of depressive symptoms, but there also seem to be more direct effects on some anxiety symptoms, which suggests there are some shared etiological factors for depression and anxiety. Traditional (tricyclic) antidepressants such as imipramine and clomipramine can be effective in treating panic disorders and obsessive–compulsive disorder, respectively. However, selective serotonin reuptake inhibitors (SSRIs), a newer class of antidepressants including such drugs as fluoxetine and paroxetine, have shown effectiveness in treating several anxiety disorders and have a more manageable side-effect profile.

ANIMAL MODELS OF ANXIETY

Animal research is an important subset of anxiety research. Such research enables us to investigate questions that are not researchable with humans for ethical reasons. In doing so, much of this work provides a valuable complement to our understanding of the nature of anxiety. (*See* **Neuropsychological Disorders, Animal Models of**)

Controllability and Unpredictability

Early research on experimental neuroses provides a good example of the benefits of animal research. When exposed to inescapable stressful situations or tasks, animals exhibit behavior that resembles anxious behavior in humans. Results of animal experiments show that uncontrollable and unpredictable stressful events reliably elicit animal behavior analogous to intense and persistent fear and physiological arousal in humans. Cognitive theories of anxiety used these findings in stressing the importance of the relation between anxiety and perception of control. Anxiety is less likely to be shown by individuals who feel in control of events than it is by individuals who feel helpless.

Observational Learning, Preparedness and Phobias

Michael Cook and Susan Mineka have conducted animal research on observational learning and preparedness in the development of phobias. Rhesus monkeys raised in their natural habitat are frightened of snakes. Those raised in captivity do not exhibit this fear response when initially exposed to a toy snake. After they observe a monkey, in real life or on videotape, act fearfully in the presence of a snake, however, they do react to the snake with fear. Interestingly, similar results were not found in monkeys observing a model reacting with fear to stimuli not evolutionarily associated with fear. Thus, this animal model of phobic response is consistent with Seligman's idea of preparedness in humans. (*See* **Social Learning in Animals**)

NEURAL BASES OF ANXIETY

Increasing attention is being paid to the neurobiological aspects of the anxiety disorders. Some especially useful research findings deal with particular brain structures, patterns of brain activity associated with anxiety, and the effects of psychological trauma on brain physiology. (*See* **Emotion, Neural Basis of**)

The Amygdala

An abundance of neurobiological research on emotion has focused on the amygdala, a subcortical brain structure involved in processing emotion-related information. Joseph LeDoux and Michael Davis have conducted extensive research on the role of the amygdala within the neural circuitry underlying fear conditioning. This brain structure is thought to be responsible for attaching emotional meaning to incoming sensory stimuli and triggering anxiety responses. Indeed, damage to the amygdala results in a loss of the fear response. (*See* **Amygdala**)

Patterns of Brain Activity Associated with Anxiety

Neuropsychological studies have provided discrepant results regarding the patterns of hemispheric brain activity associated with anxiety. Some suggest there is more right hemisphere activity, others that there is more left hemisphere activity. As it turns out, the studies implicating the left hemisphere tended to include participants characterized primarily by excessive worry, or anxious apprehension. On the other hand, studies implicating the right hemisphere tended to include individuals with more somatic symptoms, or anxious arousal. In a direct test of the patterns of brain activity associated with these proposed subtypes of anxiety, Wendy Heller and colleagues showed that they were distinguishable. These results highlight the need for studies to distinguish between different types of anxiety and depression, since they often occur together, if we hope to identify the precise neurobiological patterns associated with each anxiety disorder. (*See* **Electroencephalography (EEG)**; **Brain Asymmetry**)

Effects of Psychological Trauma on the Brain

The experience of psychological trauma can have profound biological consequences, some of which may be related to the etiology and maintenance of PTSD. Research in this area is growing rapidly with the more widespread availability of technologies for measuring the structure and function of specific brain regions. (*See* **Neuroimaging**; **Neuropsychological Development**; **Neurotransmitters**)

One response to trauma involves an increase in the production of noradrenaline (norepinephrine). It is believed that elevated levels of this neurotransmitter are partly responsible for the emergence of

the hyperarousal symptoms associated with PTSD. In addition, the dysregulation of this neurotransmitter system may elicit a normal stress response that is more intense and more easily triggered. This may be one mechanism by which traumatic events increase the likelihood of future traumatic events leading to full-blown PTSD. There may be a cumulative effect of traumatic experiences.

A complementary response to trauma involves the release of cortisol, a stress hormone that serves as the parasympathetic counterpart to the sympathetic activity of noradrenaline. Whereas noradrenaline serves to mobilize the body's resources in stressful situations, cortisol is released to moderate that response. Cortisol levels are reliably elevated shortly after a stressful event, reflecting the shutting off of the stress response. However, if the trauma is repeated or long-term, such as extended combat or chronic child sexual abuse, cortisol levels are paradoxically low. This may be due to a sensitization of the stress response system, resulting in lower levels of cortisol being needed to shut off the stress response. (*See* **Stress and Cognitive Function, Neuroendocrinology of**)

The hippocampus is a brain structure important for memory. It contains many cortisol receptor sites, and it appears that it may be damaged by elevated levels of cortisol (prior to sensitization) in response to trauma. In fact, some trauma studies show significantly smaller hippocampal volumes in trauma victims with PTSD than in trauma victims without PTSD. This finding is interesting, because one of the symptoms of PTSD is a memory deficit for details of the traumatic event. Owing to the correlational and retrospective nature of these studies, though, it is not yet known whether a small hippocampus is a risk factor for developing PTSD in response to a traumatic experience, or whether the volume reduction is a direct result of experiencing (and possibly reexperiencing) traumatic events. (*See* **Hippocampus**; **Amnesia**; **Hormones, Learning and Memory**; **Memory Distortions and Forgetting**)

CONCLUSION

The anxiety disorders are among the most commonly diagnosed psychological disorders, affecting millions of individuals and their families. Research on humans and nonhumans has provided a great deal of information about the biological, psychological and social factors responsible for the escalationof anxiety from adaptive to maladaptive levels. Several effective psychological and pharmacological treatments have been developed to help ease the human suffering associated with anxiety and its disorders. Further elucidation of the etiology of anxiety will pave the way for even more effective treatments.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Barlow DH (1988) *Anxiety and Its Disorders: The Nature and Treatment of Anxiety and Panic*. New York: Guilford.

Black DW, Noyes R, Goldstein RB and Blum N (1992) A family study of obsessive-compulsive disorder. *Archives of General Psychiatry* **49**: 362–368.

Clark LA and Watson D (1991) Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology* **100**: 316–336.

Davis M (1992) The role of the amygdala in conditioned fear. In: Aggleton JP (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pp. 255–306. New York, NY: Wiley-Liss.

Foa EB and Kozak MJ (1986) Emotional processing of fear: exposure to corrective information. *Psychological Bulletin* **99**: 20–35.

Foa EB and Rothbaum BO (1997) *Treating the Trauma of Rape: Cognitive-Behavioral Therapy for Posttraumatic Stress Disorder*. New York, NY: Guilford.

Heller W, Nitschke JB, Etienne MA and Miller GA (1997) Patterns of regional brain activity differentiate types of anxiety. *Journal of Abnormal Psychology* **106**: 376–385.

Kendler KS, Neale MC, Kessler RC, Heath AC and Eaves LJ (1992) The genetic epidemiology of phobias in women: the interrelationship of agoraphobia, social phobia, situational phobia, and simple phobia. *Archives of General Psychiatry* **49**: 273–281.

LeDoux JE (1996) *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon & Schuster.

Mineka S (1985) Animal models of anxiety-based disorders: their usefulness and limitations. In: Tuma AH and Maser JD (eds) *Anxiety and The Anxiety Disorders*, pp. 199–244. Hillsdale, NJ: Lawrence Erlbaum.

Mineka S, Watson D and Clark LA (1998) Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology* **49**: 377–412.

Nitschke JB, Heller W and Miller GA (2000) Anxiety, stress, and cortical brain function. In: Borod JC (ed.) *The Neuropsychology of Emotion*, pp. 298–319. New York, NY: Oxford University Press.

Noyes R, Clarkson C, Crowe RR, Yates WR and McChesney CM (1987) A family study of generalized anxiety disorder. *American Journal of Psychiatry* **144**: 1019–1024.

Skinner BF (1953) *Science and Human Behavior*. New York, NY: Macmillan.

True WR, Rice J, Eisen SA *et al*. (1993) A twin study of genetic and environmental contributions to liability for

posttraumatic stress symptoms. *Archives of General Psychiatry* **50**: 257–264.
Wolpe J (1958) *Psychotherapy and Reciprocal Inhibition*. Stanford, CA: Stanford University Press.
Yehuda R and McFarlane A, eds (1997) *Psychobiology of Posttraumatic Stress Disorder*. New York, NY: New York Academy of Sciences.
Zinbarg RE, Barlow DH, Brown TA and Hertz RM (1992) Cognitive-behavioral approaches to the nature and treatment of anxiety disorders. *Annual Review of Psychology* **43**: 235–267.

Attention Deficit Hyperactivity Disorder

Introductory article

James M Swanson, University of California, Irvine, California, USA
Nora D Volkow, Brookhaven National Laboratory, Upton, New York, USA
Jeffrey Newcorn, Mount Sinai School of Medicine, New York City, New York, USA
BJ Casey, Sackler Institute, Weill College of Medicine at Cornell University, New York City, New York, USA
Robert Moyzis, University of California, Irvine, California, USA
David Grandy, University of Oregon Health Sciences Center, Portland, Oregon, USA
Michael Posner, University of Oregon, Eugene, Oregon, USA

CONTENTS

Attention deficit hyperactivity disorder is a childhood syndrome characterized by developmentally inappropriate inattention, impulsivity, and hyperactivity which produces impairment at home and school. Long-term outcome is poor, but treatment with stimulant medication and behavior modification is effective. Investigations of the disorder and its treatments suggest a dopamine deficit exists that may be corrected by stimulant medication.

INTRODUCTION

The combination of inattentive, hyperactive, and impulsive behavior in children has been recognized as a syndrome since the start of the twentieth century, dating back to Still's description in 1902 of children with 'marked inability to concentrate and sustain attention' and impaired 'inhibitory volition'. The term now used as a label for this syndrome, attention deficit hyperactivity disorder (ADHD), is defined in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), published by the American Psychiatric Association. The DSM-IV definition lists 18 behaviors as grounds for diagnosis (Table 1), which fall into two domains: inattention, and hyperactivity/impulsivity. These are behaviors of normal childhood when they occur infrequently or at a low intensity, so they represent symptoms of a psychiatric disorder only when they are developmentally inappropriate, severe, and produce significant impairment in multiple settings. For a diagnosis of ADHD–combined type, both symptom domains must be present and contribute to impairment. Partial syndromes (ADHD–inattentive type or ADHD–hyperactive/impulsive type) are diagnosed if symptoms in only one domain are present, and comorbid conditions (such as anxiety and depression) are diagnosed if they are also present. Based on these criteria, about 3–5% of the population of children in American elementary schools are diagnosed and treated for ADHD.

The identification of a specific cognitive deficit unique to ADHD has been elusive. Some researchers have suggested that deficiencies of children with ADHD are due to their inability to control their behavior, rather than a structural deficit of attention. Others have concluded that there is no attentional deficit in ADHD, but instead that the core deficit is in behavioral inhibition. However, ADHD children do clearly have abnormal performance on several tasks such as the Stroop color-word naming task, the Matching Familar Figures test of comparison of almost identical complex figures, the Tower of Hanoi test of planning and stacking colored rings to match a pattern, and the Trails B test of search for characters on a page in the face of distraction.

Advances in the field of cognitive neuroscience led to new concepts of attention linked to specific brain circuitry. For example, Posner and Raichle's neuroanatomical network theory of attention is based on the concepts of alerting (supressing background neural noise by inhibiting ongoing or

Table 1. Alignment of symptom domains, cognitive processes and neural networks

Symptom domain	*Cognitive process*	*Neural network*
Inattentive – Alerting	Sustained attention	Alerting
difficulty sustaining attention	vigilance level/decrement	cortical: right frontal
fails to finish	persistence	midbrain: locus ceruleus
avoids sustained effort	performance	thalamic:?
Inattentive – Orienting	Selective attention	Orienting
distracted by stimuli	visual cueing	cortical: parietal
does not seem to listen	auditory cueing	thalamic: pulvinar
fails to give close attention	visual search	other:?
Inattentive – Memory	Memory/planning	Executive control
has difficulty organizing tasks	planning	cortical: prefrontal
loses things	memory for objects	striatal: basal ganglia
is forgetful	memory for time	other:?
Impulsivity – Executive control	Cognitive regulation	Executive control
blurts out answers	conflict resolution	cortical: anterior cingulate
interrupts or intrudes	behavioral inhibition	striatal: nucleus accumbens
cannot wait	delay aversion	other:?
Hyperactivity – Fine motor	Motor/vocal control	Fine motor control
fidgets	fine motor control	cortical: left frontal
cannot play quietly	nonverbal control	striatal: cerebellar vermis
talks excessively	verbal	other:?
Hyperactivity – Gross motor	Activation level	Gross motor control
leaves seat	gross motor control	cortical: right frontal
runs about and climbs	novelty seeking	striatal: caudate
always on the go	arousal level	other:?

irrevevant activity or mental effort to establish a state of vigilance), orienting (mobilizing specific neural resources toward a source of sensory stimulation), and executive control (coordinating multiple specialized neural processes by detecting targets, starting and stopping mental operations, and resolving conflict among responses), each with a well-defined neural circuitry (anterior cingulated, prefrontal cortex and basal ganglia). This cognitive neuroscience approach can be used to constrain the definition of attention, and it offers modern terminology for describing its components. The application of three levels of analysis (behavioral, cognitive, and neural) may provide some new insights about the cognitive component of this disorder. Each symptom can be classified based on its relationship to alerting, orienting, and executive control (see Table 1). The nine symptoms of inattention listed in Table 1 logically split into three groups when aligned with the three concepts of attention and the underlying neural networks. The three symptoms of impulsivity are behavioral manifestations of deficits in self-regulation, which align with the executive control network, and the six symptoms of hyperactivity fall into two groups based on deficits in fine motor and gross motor control.

ONSET AND COURSE

The DSM-IV diagnostic criteria specify the onset of symptoms by the age of 7 years, but in most cases the symptoms of ADHD are present much earlier. Impairment tends to increase during the elementary-school years in response to the cognitive and behavioral demands of the classroom setting, and this is when most diagnoses are made. At this age, more boys than girls are recognized and treated for ADHD (reported male to female ratios range from 3:1 to 9:1), but this may be due to referral biases related to disruptive behaviors (aggression, opposition, and defiance) that often coexist in boys. In most cases, ADHD symptoms decline with age, especially for the domain of hyperactivity and impulsivity. When symptoms no longer produce impairment, the diagnostic label changes to ADHD–residual type. In about one-third of the cases, the full criteria are still met in adulthood, and in another third symptoms are present but at a subthreshold level.

The subjective nature of the assessment process raises legitimate questions about the validity of the diagnosis of ADHD, but evidence for validity has accumulated from follow-up studies showing that a childhood diagnosis of ADHD is associated with

extremely poor outcome in many areas, including juvenile delinquency. Children identified by the ADHD diagnosis have a serious disorder that demands recognition and deserves treatment.

TREATMENT

Since the 1930s ADHD has been treated with stimulant drugs. The first of these was amphetamine, but over the years this has been superseded by methylphenidate. Immediate-release formulations of these drugs are short-acting and must be given two or three times a day. Newer sustained-release formulations based on 'osmotic pump' and 'coated bead' delivery systems have been developed that have long duration of action and avoid the mid-day dose at school (which is often associated with embarrassment). The stimulant medications are effective in reducing the symptoms of ADHD (about 80% of children with this diagnosis show clinically meaningful benefits) and are safe (despite some common side effects such as decreased appetite and sleep). With the immediate-release formulations, the therapeutic effects emerge within 1–2 h after each oral dose, but then dissipate within 3–6 h. The sustained-release formulations have a duration of action of 8–12 h. These stimulant medications exert a profound cognitive effect, characterized by focused attention to tasks (even those with low intrinsic interest) and maintenance of attention over time (even in the face of repetitive or boring tasks). The behavioral effects are also profound: inappropriately high levels of activity and inattention in the classroom setting are reduced and compliance with typical requests and rules is increased dramatically. Stimulants do not produce a paradoxical response in ADHD children: normal children and adults respond in the same way on most measures (e.g. by decreasing normal levels of activity and increasing normal levels of attention). It is important to note that in individuals who do not have ADHD there is no impairment in these areas, so the response to stimulants does not alleviate impairment, which is the hallmark of clinical response.

In addition to pharmacotherapy, contingency management programs have been developed based on the general principles of behavior modification (reinforcement, punishment, extinction, and stimulus control). Typically, these interventions use token systems in the home and at school to prompt and shape appropriate target behavior (e.g. getting started, staying on task, interacting appropriately with others, completing work, and shifting activities on schedule) and to extinguish inappropriate behaviors (e.g. getting out of seat, talking without permission).

The most recent information on treatment comes from the Multimodality Treatment of ADHD (MTA) study, a large, six-site randomized clinical trial designed to evaluate the long-term effects of pharmacological and psychosocial interventions. Over 500 children with ADHD aged 7–9 years were recruited from a variety of sources and randomly assigned to a treatment group for a period of 14 months. In this study methylphenidate administered three times a day was more effective than psychosocial treatment (intensive behavioral intervention at home and school), and combinations of these two therapies were little better than medication alone. The success rates defined by a reduction of symptoms to a subthreshold level reflected this also: psychosocial 34%, pharmacological 56%, and combination 68%. This study provides empirical evidence of the long-term effectiveness of these two most common treatments for ADHD.

ETIOLOGY

The most prominent current theory about the cause of ADHD implicates dysfunction of brain dopamine (DA), a neurotransmitter involved in the regulation of motoric, attentional and motivational circuits. One variant of this theory suggests that ADHD is the result of a DA deficit at the neural level, which results in inattentiveness and distractibility at the cognitive level. This theory is supported by the mechanism of action of methylphenidate, which blocks DA transporters, the primary mechanism for removing DA from the synapse. Imaging studies in humans have demonstrated that therapeutic doses of stimulants block more than half of the DA transporters, markedly enhancing DA neurotransmission in the brain. Similar findings have been obtained in animal studies, which have shown that stimulants given at therapeutically relevant dosages increase extracellular DA and activate DA-regulated circuits. In animals, gene 'knockout' studies have suggested that the DA transporter gene (*DAT*) located on chromosome 15 and the DA type 4 receptor gene (*DRD4*) located on chromosome 11 are involved in basic underlying processes of activity and attention that may contribute to ADHD.

What might produce a DA deficit? Acquired and inherited factors have been proposed. One suggestion is that bouts of hypoxia and hypotension during fetal development might selectively damage striatal neurons, which are the main target for DA cells. Inherited factors have also

been implicated by molecular genetic studies of ADHD. Initial investigations focused on two candidate genes involved with DA regulation: the *DAT* and *DRD4* genes. Several research groups have documented association of these two genes with ADHD.

NEURAL CORRELATES OF ADHD

In the early 1990s several teams of investigators used imaging techniques to investigate brain anatomy in groups of children with ADHD compared with children free from this disorder. Abnormalities in size of specific brain regions were observed across multiple studies. Even though groups of children with ADHD were recruited from very different clinical settings by independent research teams, research teams showed a moderate reduction in size (about a 10% decrease compared with a normal group) for measures of frontal lobes and basal ganglia (caudate nucleus and globus pallidus). Functional brain imaging studies have provided converging information implicating basal ganglia and frontal lobe abnormalities in ADHD. Imaging studies based on single photon and positron emission tomography and performed during baseline (resting) conditions documented a reduction in blood flow and metabolism in striatal and frontal brain regions. Studies using [^{18}F]-labeled dopa as a marker of DA synthesis in brain showed significant reductions in prefrontal cortex in people with ADHD when compared with normal controls. Functional magnetic resonance imaging has shown hypoactivity of frontal circuits during activation by cognitive tasks, including blunted activation in the anterior cingulate gyrus, a brain region with a central role in executive attention that has been linked to the behavioral symptoms of inattention and impulsivity in ADHD (see Table 1).

CONCLUSION

The phenomenology of ADHD has been refined over the years, and clinical manuals now agree on the specific symptoms of this disorder. At a cognitive level, abnormalities in neuropsychological performance suggest that children with ADHD are inefficient in information processing, resulting in slow and inaccurate performance. Advances in brain imaging and molecular biology have started to reveal functional and biochemical brain abnormalities associated with ADHD, localized predominantly in dopamine pathways and frontal and striatal circuits modulated by DA. Since these circuits regulate attention, executive function, motivation, response inhibition and motor activity, research on ADHD has focused on how their dysfunction could result in the cognitive deficits and behavioral symptoms of this disorder.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: APA.

Barkley RA (1997) Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological Bulletin* **121**: 65–94.

Barkley RA, Fischer M, Edelbrock CS and Smallish L (1990) The adolescent outcome of hyperactive children diagnosed by research criteria 1. An 8-year prospective follow-up study. *Journal of the American Academy of Child and Adolescent Psychiatry* **29**: 546–557.

Bradley C (1937) The behavior of children receiving benzedrine. *American Journal of Psychiatry* **94**: 577–585.

Bush G, Frazier JA, Rauch SL *et al*. (1999) Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fMRI and the Counting Stroop. *Biologicol Psychiatry* **45**: 1542–1552.

Castellanos FX (1997) Toward a pathophysiology of attention-deficit/hyperactivity disorder. *Clinical Pediatrics* **36**: 381–393.

Castellanos FX, Giedd JN, Marsh WL *et al*. (1996) Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. *Archives of General Psychiatry* **53**(7): 607–616.

Collier D, Curran S and Asherson P (2000) Mission: not impossible? Candidate gene studies in child psychiatric disorders. *Molecular Psychiatry* **5**: 457–460.

Ernst M, Zametkin AJ, Matochik JA, Jons PH and Cohen RM (1998) DOPA decarboxylase activity in attention deficit hyperactivity disorder adults. A [fluorine-18] fluorodopa positron emission tomographic study. *Journal of Neuroscience* **18**: 5901–5907.

Faraone S, Doyle A, Mick E and Biederman J (2001) Meta-analysis of the association between the dopamine D4 gene 7-repeat allele and ADHD. *American Journal of Psychiatry* **158**: 1052–1057.

Greenhill LL, Halperin JM and Abikoff H (1999) Stimulant medications. *Journal of the American Academy of Child and Adolescent Psychiatry* **38**(5): 503–512.

Levy F and Swanson JM (2001) Timing, space and ADHD: the dopamine theory revisited. *Australian and New Zealand Journal of Psychiatry* **35**: 504–511.

Lou HC (1996) Etiology and pathogenesis of attention-deficit hyperactivity disorder (ADHD): significance of prematurity and perinatal hypoxic-haemodynamic encephalopathy. *Acta Paediatrica* **85**: 1266–1271.

Lou HC, Henriksen L and Bruhn P (1990) Focal cerebral dysfunction in developmental learning disabilities. *Lancet* **335**(8680): 8–11.

Mannuzza S, Klein RG, Bessler A, Malloy P and LaPadula M (1993) Adult outcome of hyperactive boys:

educational achievement, occupational rank, and psychiatric status. *Archives of General Psychiatry* **50**: 565–576.

MTA Cooperative Group (1999) Multimodal Treatment Study of Children with ADHD. A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry* **56**: 1073–1086.

NIH Consensus Conference (2000) National Institutes of Health Consensus Development Conference statement: diagnosis and treatment of attention-deficit/hyperactivity disorder (ADHD). *Journal of the American Academy of Child and Adolescent Psychiatry* **39**: 182–193.

Pelham WE and Fabiano G (2000) Behavior modification. *Child and Adolescent Psychiatric Clinics of North America* **9**: 671–688.

Pennington BF and Ozonoff S (1996) Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines* **37**(1): 51–87.

Posner MI (2001) Developing brains: the work of the Sackler Institute. *Clinical Neuroscience Research* **1**: 258–266.

Posner MI and Raichle ME (1994) *Images of Mind*. New York, NY: Scientific American Library.

Posner MI, Rothbart MK, Farah M and Bruer J (2001) Human brain development: introduction to the report to the McDonnell Foundation. *Developmental Science* **4**/3 (special issue): 253–384.

Rubia K, Overmeyer S and Taylor E (1999) Hypofrontality in attention deficit hyperactivity disorder in higher order motor control: a study with functional MRI. *American Journal of Psychiatry* **156**: 891–896.

Satterfield J, Swanson JM, Schell A and Lee F (1994) Prediction of antisocial behavior in attention-deficit hyperactivity disorder boys from aggression/defiance scores. *Journal of the American Academy of Child and Adolescent Psychiatry* **33**: 185–190.

Sagvolden T and Sergeant JA (1998) Attention deficit/hyperactivity disorder – from brain dysfunctions to behaviour [editorial]. *Behavioural Brain Research* **94**(1): 1–10.

Still GF (1902) Some abnormal psychical conditions in children. *Lancet* **1**: 1008–1012; 1077–1082; 1163–1168.

Swanson JM, McBurnett K, Wigal T *et al*. (1993) Effect of stimulant medication on children with attention deficit disorder: a 'review of reviews'. *Exceptional Children* **60**: 154–162.

Swanson JM, Sergeant JA, Taylor E *et al*. (1998) Attention-deficit hyperactivity disorder and hyperkinetic disorder. *Lancet* **351**: 429–433.

Swanson J, Castellanos FX, Murias M, LaHoste G and Kennedy J (1998) Cognitive neuroscience of attention deficit hyperactivity disorder and hyperkinetic disorder. *Current Opinion in Neurobiology* **8**: 263–271.

Swanson J, Posner M, Cantwell D *et al*. (1998) Attention-deficit/hyperactivity disorder: symptom domains, cognitive processes and neural networks. In: Parasuraman R (ed.) *The Attentive Brain*, pp. 445–460. Boston, MA: MIT Press.

Swanson JM, Kraemer H, Hinshaw S *et al*. (2001) Clinical relevance of the primary findings of the MTA: success rates based on severity of symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry* **40**: 168–179.

Swanson JM, Deutsch C, Cantwell D *et al*. (2001) Genes and ADHD. *Clinical Neuroscience Research* **1**: 207–216.

Taylor E, Chadwick O, Heptinstall E and Danckaerts M (1996) Hyperactivity and conduct problems as risk factors for adolescent development. *Journal of the American Academy of Child and Adolescent Psychiatry* **35**: 1213–1216.

Volkow ND, Ding YS, Fowler JS *et al*. (1995) Is methylphenidate like cocaine? Studies on their pharmacokinetics and distribution in human brain. *Archives of General Psychiatry* **52**: 456–463.

Volkow ND, Wang G, Fowler JS *et al*. (2001) Therapeutic doses of oral methylphenidate significantly increase extracellular dopamine in the human brain. *Journal of Neuroscience* **21**(2): RC121.

Volkow ND, Wang GJ, Fowler JS *et al*. (2002) Relationship between blockade of dopamine transporters by oral methylphenidate and the increases in extracellular dopamine: therapeutic implications. *Synapse* **43**: 181–187.

Attention, Models of

Introductory article

Rajesh PN Rao, University of Washington, Seattle, Washington, USA

CONTENTS

Attention is the ability of an organism to select and process only the relevant or 'interesting' parts of its sensory inputs, while discarding other potentially irrelevant parts. Models of attention seek to explain the mechanisms and neuronal substrates underlying this evolutionarily important sensory ability.

INTRODUCTION

Animals are confronted with a vast amount of sensory information in their day-to-day interactions with the natural world. Only a fraction of this information can be processed at any given moment in time owing to the brain's limited processing resources. Fortunately, only a fraction of the total information received is typically relevant to any particular task at hand and to the animal's continued survival. Thus, the fundamental problem faced by an animal's perceptual system is to select and process only the relevant or 'interesting' parts of its sensory inputs, discarding the other potentially irrelevant parts. Attention is nature's solution to this problem. (*See* **Attention**; **Selective Attention**)

Traditionally, attention is defined as the phenomenon by which animals preferentially process parts of their input, shifting their focus of processing from one part of their input to another in a serial fashion. The metaphor that is frequently used, especially in psychology, is that of a spotlight that can be moved to different parts of a scene independent of eye movements. Any information lying outside the spotlight is filtered out. This form of attention is generally referred to as 'covert' attention, to distinguish it from the 'overt' attentional shifts due to eye movements. (*See* **Eye Movements**)

Although the spotlight metaphor is not a computational model of attention, it has proved useful in characterizing various properties of attention. For example, events occurring within a pre-cued 'spotlight' region in the visual field can be detected faster and more accurately than events occurring outside the spotlight region. The spotlight metaphor has also been used to distinguish between two forms of visual search: parallel search, corresponding to a search for an object that differs from neighboring objects by a single feature (e.g. an X among O's), and serial search, where the target differs from neighboring objects by a conjunction of features (e.g. a green X among red X's and green O's). It has been suggested that serial search requires sequential analysis of the scene by an attentional spotlight that 'binds' different features of an object together, while parallel search is the result of 'pop-out' of the target item due to differences in a single feature. Although considerable psychophysical data exist, for instance, on the speed and size of the hypothetical attentional spotlight, neurobiological evidence for such a spotlight has remained inconclusive. The notion of a spotlight has, however, been influential in the formulation of several computational models of attention, some of which are discussed below.

GATING MODELS OF ATTENTION

One of the early quantitative models of attention inspired by the spotlight metaphor is the attention gating model, which characterizes how information is processed within the attentional spotlight. It was originally formulated to explain results from rapid serial visual presentation experiments. In these experiments, participants would fixate on a location on a screen and focus attention on a stream of letters displayed rapidly and sequentially to the left of the fixation point. Upon detection of a specified target letter, the participant shifted attention as quickly as possible to a stream of numerals being displayed to the right of the fixation point, and reported (for example) the four earliest occurring numerals after the target. This paradigm is based

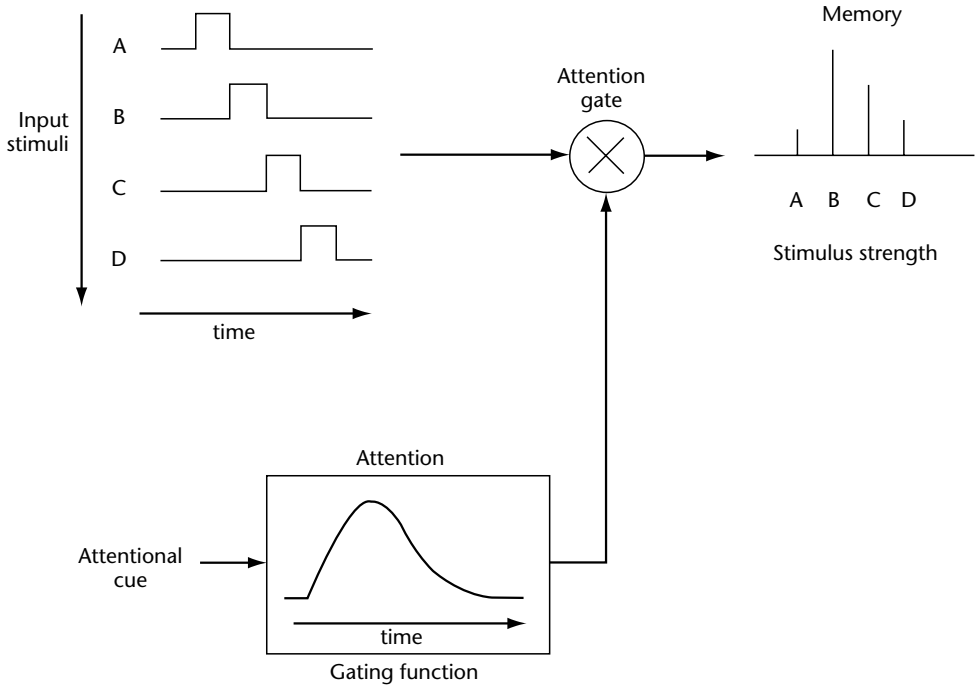


Figure 1. Attention gating model. When an attentional cue is detected, the attention gate is activated with a time course given by the attention gating function (shown within the box). The input stimuli A, B, C and D pass through the attention gate with a strength given by the product of the stimulus with the current value of the gating function. Each input stimulus is stored in memory according to its strength (indicated by the vertical bars on the right). Stimuli from memory are recalled and output as responses in decreasing order of strength.

on the view that the spotlight of attention is first focused on the stream of letters, allowing the target to be detected, and then is moved to the stream of numerals, where the spotlight allows the next four numerals to be registered in short-term memory. Thus, attention acts as a 'gate' into short-term memory, regulating the flow of sensory information into conscious awareness. (*See* **Working Memory, Computational Models of**; **Neural Basis of Memory: Systems Level**; **Working Memory, Neural Basis of**; **Memory Models**; **Working Memory**)

The attention gating model is depicted in Figure 1. In this model, when the attentional cue (for example, the target letter) is detected, the attention 'gate' is opened and the items from the to-be-attended stream of inputs is admitted into memory. To avoid overflow of memory capacity, the gate is automatically closed a short time after opening. This prevents other stimulus items from entering memory. The strength of an item stored in memory depends on the gating function G and the time at which the item occurred with respect to gate opening. The interaction is multiplicative: a stimulus S passes through the attention gate with output strength $G \times S$. The items stored in memory are then reported in decreasing order of strength. The gating function G can be estimated experimentally based on the performance of the subjects and has been shown to be a function with exponential rise and decay (Figure 1).

The attention gating model has been quite successful in quantitatively modeling psychophysical results. However, it is a purely phenomenological model (compare with the description below of Crick's model). For example, the gating model does not address how a spatial region is selected for attentional processing and how information within this region is processed before reaching short-term memory and awareness. These issues have been addressed using saliency maps and dynamic routing circuits respectively.

SALIENCY MAPS

A saliency map is a topographic representation of an input sensory field: each location in a saliency map contains a value that represents how different or salient that location is, compared with other locations in the sensory field. A 'winner takes all' mechanism is then used to select the current most salient location (peak) in the map. Attention is

directed to the selected location by a gating mechanism that allows the stimuli near the selected location to be conveyed to higher processing centers and suppresses stimuli at other locations.

The saliency of a location may be computed based on 'bottom up' and/or 'top down' information. In 'bottom up' calculations the saliency of a location may be defined in terms of the difference between the stimulus at that location and the stimuli in neighboring locations. For example, the presence of an X in a field of O's would cause the location containing the X to attain a higher saliency value than locations containing O's, owing to the large difference in features between the X and the surrounding O's. Thus, 'bottom up' saliency maps provide a computational mechanism for implementing the 'pop-out' inherent in the types of visual search characterized as parallel search. Certain forms of serial search may also be modeled using 'bottom up' saliency maps where saliency is computed based on conjunctions of stimulus features (such as 'red/green' and 'X/O'). In this case, attention is directed sequentially to successively smaller peaks in the saliency map and previously visited locations are suppressed for a short duration, a process known as 'inhibition of return'.

'Top down' saliency maps are based on the differences between sensory stimuli at different locations and a stored prototype target object. The more similar the stimulus at a particular location is to the target object, the higher the saliency value of that location. Such saliency maps provide a mechanism for implementing object-based serial search, where attention is directed sequentially to different objects in a visual scene in the order of their similarity to the prototype target object. This type of search is frequently performed by our visual system, for example when we are looking for a familiar face in a crowd, or for a pen on a cluttered desk. Figure 2 demonstrates how 'top down' saliency maps could be used to perform an object-based serial search during a visual counting task.

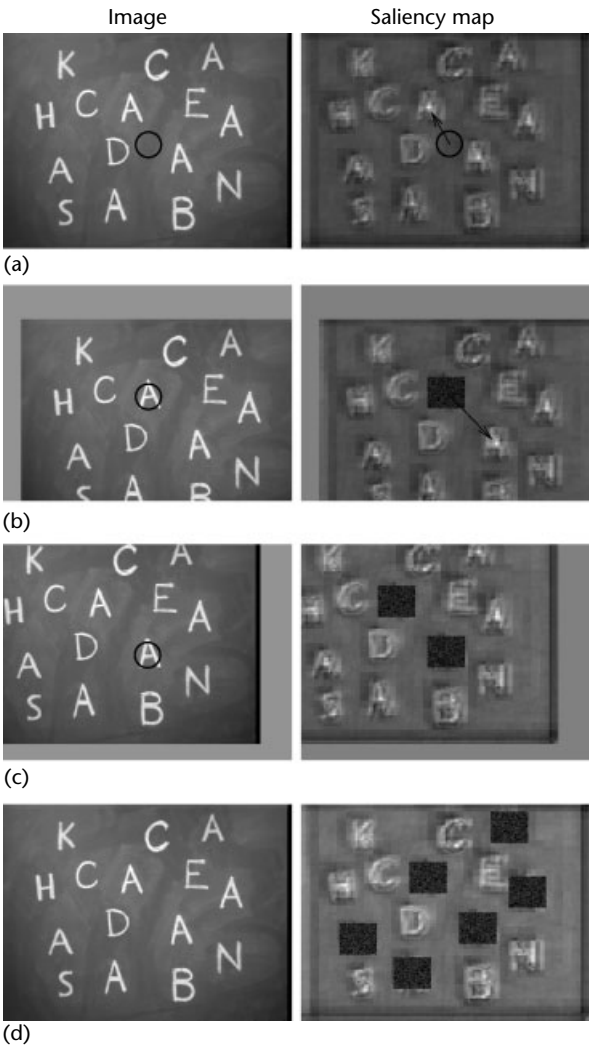


Figure 2. Saliency maps. The use of 'top down' saliency maps in a visual counting task. The goal is to count the number of occurrences of the letter A on a blackboard containing a collection of letters. The image of the board is shown on the left and the corresponding saliency map on the right. Each letter is represented by the vector of outputs of a set of oriented spatial filters modeled after the receptive fields of visual cortical neurons. The brightness at each location in the saliency map indicates the degree of closeness (saliency) of the letter at that location to the letter A, as given by the similarity in their filter-based representations. The brightest spot, indicated by the arrow, is chosen to be the next location to be attended (a). The circle indicates the original focus of attention. (b, c) The result of shifting attention to two different locations containing the letter A. In both cases, once the letter has been attended to and processed, the location containing the letter is inhibited (dark square) to prevent this location from competing again for the focus of attention during the course of the task. This implements the process of 'inhibition of return' during visual search. (d) The final result of counting all six occurrences of the letter A in the input image. Figure adapted from Rao and Ballard (1997).

SHIFTER CIRCUITS AND DYNAMIC ROUTING

Saliency maps offer a solution to the problem of selecting relevant portions of a sensory field for further processing. They do not, however, solve the problem of routing this information to higher centers in a position- and size-invariant way. Shifter circuits were proposed as a computational model for how the brain may accomplish this within the context of visual information processing. These circuits are so named because they allow

information in one location of the visual field to be 'shifted' and fed to a different location in a higher processing module. Such a process leads to the formation of higher-level visual representations that remain stable regardless of actual object position and size at the lowest levels.

The primary components of shifter circuits (also called dynamic routing circuits) are the control neurons, which set the size and position of the attentional window. The control neurons modulate the strength of synaptic connections from the lower level to a higher level such that only the information from the attended region in the lower level is routed to the higher level. The control neurons are driven by a saliency map. Only neurons that correspond to highly salient locations are activated, causing information in only these locations to be routed to higher processing centers.

Position and size invariance is achieved by arranging multiple inputs from a lower level to converge onto a single higher-level unit (Figure 3a). This allows units at higher levels to have access to information from increasingly large areas of the input field. By having the control neurons select appropriate subsets of input connections at each level, one can 'focus attention' on any local region in the input field and route the information in this local region to the same set of processing units at the highest level (Figure 3b). This makes the circuit invariant to the actual position of objects in the input field. Similarly, size invariance can obtained by integrating several lower level inputs into a single higher level output, repeating this strategy at each level until the very highest level (Figure 3c). Once again, the control neurons dictate which inputs are integrated.

As mentioned above, saliency maps control attention in routing circuits. Once a saliency map has been computed, a competitive mechanism is used to activate the control units that correspond to the most salient parts of the input. These control units in turn select the appropriate input connections at each level for routing the selected information to the highest level. The exact equations governing the dynamics of the control neurons can be derived from well-known optimization principles. Information received at the highest level can be fed into an associative memory for recognition. In addition, during the recognition process, the output of the associative memory can be used to readjust the position and scale of the attention window to enhance recognition accuracy. Small-scale dynamic routing circuits of up to three levels have been imulated on desktop computers. Promising results have been obtained for simple pattern recognition tasks such as recognizing letters in binary images, but the performance of routing circuits in more realistic tasks involving natural images has not yet been thoroughly investigated. (*See* **Vision: Object Recognition**)

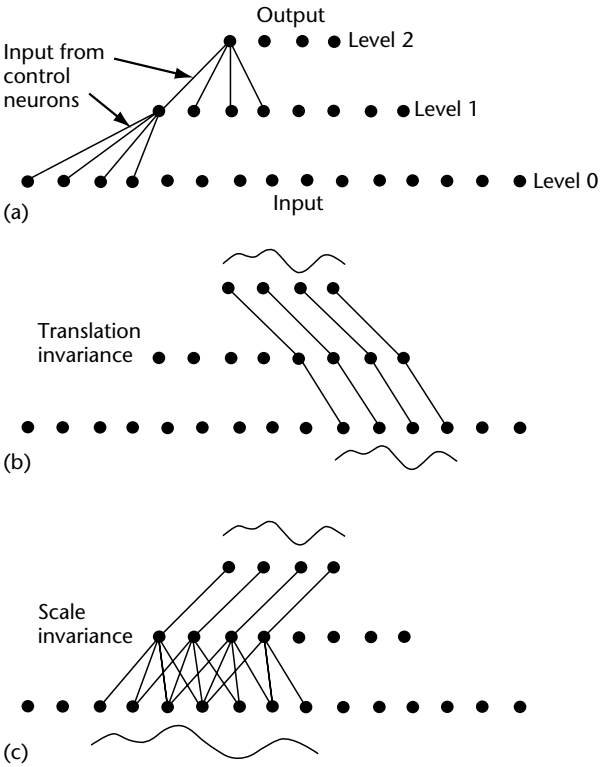


Figure 3. Routing circuits. (a) A dynamic routing circuit with three levels. Each neuron, represented by a black circle, receives inputs from several lower level neurons via dynamically modifiable connections. Only the connections for the leftmost neurons in levels 1 and 2 are shown. The connections for the other neurons are identical except for a rightwards shift. The strength of these connections is determined by the control neurons, which modulate the connections multiplicatively (indicated by the two arrows) and set the position and size of the window of attention. (b) Attention can be focused on a local input region. Information lying within the attended region (depicted as a wavy line) is routed to the output layer by the control neurons. The pattern obtained at the output level is invariant to translations of the pattern at the input level. (c) The window of attention can be enlarged to process large patterns. In this case, the control neurons set the connections for a net convergence from input to output. As a result, the pattern obtained at the output level is invariant to changes in input pattern size.

SPATIAL VERSUS OBJECT-BASED ATTENTION

The models we have considered so far have emphasized how attention may be directed towards

specific spatial locations, and how sensory information from these locations may be routed to higher centers for further processing. Behavioral and single-cell studies have suggested that attention may also select parts of a sensory input based on visual features (such as motion or color) or even whole objects, a phenomenon known as feature-based or object-based attention. (*See* **Visual Attention**; **Selective Attention**)

Evidence for feature- and object-based attention comes from studies showing that participants can reliably track a target object that is superimposed with a distracter object in the same spatial location. For example, if the image of a face is transparently superimposed on the image of a house and the face image is moved back and forth, subjects are able to track the face despite the presence of the image of the house. In addition, brain imaging studies reveal higher neural activity in the face-selective and motion-selective areas of the brain when tracking the face, compared with the case where the house is the target of attention. Other evidence for object-based attention comes from studies involving patients with damage to their right parietal lobe. Many of these patients are unable to attend to the left side of objects, where the definition of an object is dependent on the task at hand. (*See* **Parietal Cortex**)

Computational models of object-based attention are still in their infancy, partly because of the broad range of phenomena that fall into this category. Preliminary attempts at modeling object-based attention have focused on cooperative networks that comprise two complementary subnetworks, one that estimates objects and their features (such as color and shape) and another that estimates object transformations (such as translation, dilation and rotation). In such networks, attention can be focused on a particular object or a particular location by biasing 'top down' signals from higher processing centers or short-term memory. Object-based attention corresponds to keeping the 'top down' signal for the object-estimating network fixed: this causes the transformation network to converge to estimates of the attended object's transformations. Spatial attention corresponds to keeping the 'top down' signal for the transformation network fixed: this causes the object network to converge to the identity of the object in the attended location. The inspiration for such a model comes from neuroanatomical studies showing a rough dichotomy in the mammalian brain between networks specialized for object identification and networks geared towards spatial transformations and action.

NEUROANATOMICAL SUBSTRATE

The spotlight metaphor inspired some of the early attempts at identifying possible neuronal substrates of attention. In 1984, Francis Crick, one of the co-discoverers of the structure of DNA, proposed that the thalamic reticular nucleus (TRN) may have an important role in implementing a spotlight. The thalamus is a nucleus that receives input information from a majority of the senses, including visual, auditory and somatosensory information. This information is conveyed to the neocortex by relay cells in the thalamus. The TRN is a single-layer network of inhibitory neurons strategically located between the thalamus and the cortex. The network receives both ascending inputs from the thalamus as well as descending inputs from the cortex. Its neurons process these inputs and inhibit neighboring neurons as well as corresponding neurons in the thalamus. Thus, the TRN is well suited to regulating the flow of input sensory information from the thalamus to higher cortical centers. In other words, it could implement an attentional spotlight by allowing only selected parts of the sensory inputs to be conveyed to the cortex, inhibiting all other neurons in the thalamus that correspond to other sensory inputs. Despite considerable progress in our knowledge of the TRN and the thalamus, evidence for Crick's spotlight model remains inconclusive. The main stumbling block remains our poor understanding of the feedback loop between the cortex and the thalamus in alert behaving animals, and the role of the TRN in modulating this feedback loop.

Several possibilities exist for biophysically implementing the gating mechanisms that are integral parts of not only the attention gating model, but also other models such as the routing circuit model. The primary requirement here is a mechanism that can inhibit or filter out input channels that are not being attended, allowing only the attended inputs to proceed to the next stage of processing. Presynaptic inhibition (inhibition of an input fiber before it can affect a cell) allows precise gating of inputs to a single neuron, but there is little evidence for its existence in the neocortex. Postsynaptic inhibition (inhibition of a cell after an input has been delivered) can also be used to gate the inputs to a neuron but at coarser level. Various cellular mechanisms have been suggested for this form of gating, most of which are based on particular characteristics of the ionic channels and receptors that are embedded in the membrane of neurons.

The visual cortical areas V1, V2, V4 and inferotemporal cortex (IT) are assumed to be the major

neural substrates of dynamic routing circuits in the visual cortex. These areas are organized in a roughly hierarchical manner, with neurons in each area receiving convergent inputs from up to a thousand neurons in the preceding area. This is consistent with the multiple-level architecture of a routing circuit. The control neurons that direct the flow of information from one level to the next are hypothesized to be located in the pulvinar, a subcortical nucleus of the thalamus. The pulvinar sends and receives connections from each of the areas from V1 to IT, making it a suitable candidate for controlling attentional routing. Neurophysiological and brain imaging studies support the hypothesis that the pulvinar is involved in filtering out unattended stimuli. The competitive interactions between the control neurons are assumed to be mediated by lateral inhibition within the pulvinar and/or through the TRN. Finally, the control neurons in the pulvinar are assumed to be driven by saliency maps, whose neural implementation is discussed below. (*See* **Occipital Cortex**; **Temporal Cortex**)

Several different cortical and subcortical areas have been suggested as possible substrates for implementing saliency maps. The posterior parietal cortex (PP) is one such area. Neurons in PP show either elevated or suppressed activity when attention is directed to a visual target. In addition, damage to PP impairs the ability to disengage attention from a currently attended location. These results are consistent with a saliency map-based interpretation of PP. Another possible neural substrate for a saliency map is the superior colliculus, a subcortical nucleus that plays a major role in targeting eye movements to different parts of a scene. The superior colliculus receives direct input from the retina and can influence activity in higher cortical areas via its connections to the pulvinar. Other possible areas that may encode saliency and/or behavioral relevance of targets include the frontal eye fields and the inferior/lateral divisions of the pulvinar itself. It is currently unclear whether these different areas encode different types of saliency or different types of sensorimotor modalities (such as attentional shifts, or eye or hand movements). (*See* **Parietal Cortex**)

Models that jointly address spatial and object-based attention usually rely on the approximate division of visual processing in the visual cortex into the tasks of 'what' (object identification) and 'where' (motion/spatial reasoning). Object-related processing has traditionally been associated with the so-called 'form' pathway, comprising the hierarchy of areas V1, V2, …, IT in the ventral part of the visual cortex. Spatial reasoning and action-related processing is traditionally ascribed to areas in the dorsal part of the visual cortex, particularly the intraparietal and PP cortex. One model for spatial and object-based attention assumes that the cortical areas in the frontal lobe store task-relevant information and apply either spatial or object-related constraints to the dorsal or ventral networks respectively. A spatial constraint causes the object ('what') networks in the ventral pathway to respond strongly to the objects or features in the input image with that spatial property (such as position or motion). Similarly, an object-based constraint on ventral networks would cause the dorsal networks ('where') to focus upon the spatial properties of the object expressed in the constraint. Such a model, which is based on interactions between dorsal, ventral and frontal areas of the brain, integrates the neuronal and systems-level views of attention. It is, however, hard to validate, given the current difficulty in recording from multiple brain areas simultaneously. (*See* **Temporal Cortex**; **Parietal Cortex**; **Frontal Cortex**)

At the single neuron level, neurophysiological recordings in awake behaving monkeys have provided some important clues to understanding the mechanisms of attention. For example, the response of many neurons in the visual cortical areas V2 and V4 are modulated by attention. In these experiments, a visual stimulus that can activate a recorded neuron is first found and then a second stimulus is placed within the activating region (receptive field) of the neuron. This typically causes the neuron's response to shift significantly from its original response. However, when the monkey is made to focus attention on the original stimulus, the neuron's response becomes almost identical to its original response, as if the attended stimulus appeared all by itself. This provides direct evidence for attention filtering out irrelevant stimuli. A model based on 'biased competition' between two populations of neurons, one representing the attended stimulus and the other representing the distracter stimulus, has been suggested to explain these results. Attention is assumed to increase the strengths of the inputs from the population representing the attended stimulus. Such a model has been shown to account quantitatively for the neurophysiological results. However, the lack of details regarding its neurobiological implementation makes it hard to distinguish this mathematical model from the computationally motivated models discussed above, most of which use biased competition in one form or another to focus attention. (*See* **Attention, Neural Basis of**; **Spatial Attention, Neural Basis of**)

CONCLUSION

Computational models of attention provide useful insights into how the brain selectively processes portions of its inputs based on measures of saliency and task relevance. The process by which parts of a scene are selected and processed without movement of the eyes is often explained by a spotlight metaphor. Saliency maps provide a mechanism for implementing a spotlight by allowing the selection of a single stimulus based on either its 'bottom up' saliency compared with competing stimuli or its 'top down' relevance to the task at hand. Once selected, a stimulus can be routed to higher processing centers and eventually to memory using dynamic routing circuits. These circuits allow information from any portion of the sensory field to be selectively routed from one stage of processing to another in a manner independent of size and position. Once information is routed to memory, attention gating models provide a quantitative characterization of how attended items are transferred to short-term memory. Recent models of attention have stressed global interactions between networks specialized for object identification and spatial reasoning. Such models may provide a unified framework for modeling object-based and spatial attention.

The neural substrates of attention have not yet been completely identified, but the effects of attention on single neurons have been studied in several visual cortical areas, including V2, V4, the frontal eye fields and posterior parietal cortex. The pulvinar nucleus of the thalamus, the thalamic reticular nucleus and the superior colliculus are some of the subcortical nuclei implicated in attentional control. How these different areas interact to produce the emergent phenomenon of attention is currently the subject of both modeling as well as experimental studies.

Further Reading

Crick F (1984) Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Science of the USA* **81**: 4586–4590.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**: 193–222.

Kastner S and Ungerleider LG (2000) Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience* **23**: 315–341.

Koch C and Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* **4**: 219–227.

Olshausen BA, Anderson CH and Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience* **13**: 4700–4719.

Parasuraman R (ed.) (1998) *The Attentive Brain*. Cambridge, MA: MIT Press.

Pashler HE (1998) *The Psychology of Attention*. Hillsdale, NJ: Erlbaum.

Posner MI and Raichle ME (1994) *Images of Mind*. New York, NY: Scientific American Books.

Rao RPN and Ballard DH (1997) A computational model of spatial representations that explains object-centred neglect in parietal patients. In: Bower JM (ed.) *Computational Neuroscience: Trends in Research 1997*, pp. 779–785. New York, NY: Plenum Press.

Reeves A and Sperling G (1986) Attention gating in short-term visual memory. *Psychological Review* **93**: 180–206.

Treisman AM and Gelade G (1980) A feature-integration theory of attention. *Cognitive Psychology* **12**: 97–136.

Attention

Introductory article

Daniel Gopher, Technion Institute of Technology, Haifa, Israel
Cristina Iani, Technion Institute of Technology, Haifa, Israel

CONTENTS

Attention is the scientific term primarily used to describe all processes and mechanisms that govern the subjective constraints imposed by the human organism on the flow and interpretation of external and internal information, and on the organization and selection of responses, in the service of goal-directed behavior. In some cases, attention can also be automatically captured by sudden changes in the situation, or by well-trained stimulus–response tendencies.

WHAT IS ATTENTION?

At any single moment in life, humans perceive, attend to, respond to, and make use of only a very small fraction of the information from the outside world that stimulates their sensory systems (vision, audition, touch, motion, balance, etc.), or from the internal information and skills that are stored in their memory and acquired through past experiences. This fraction is described as the focus of their attention. Humans are not passive subjects of their environment. Rather, they actively seek, bias, and attribute significance to outside and inside events. Attention represents this active process. Like any other physical system, humans are limited in the amount and rate at which they can process information and perform tasks. The study of attention is the study of the processes and mechanisms that govern the 'top-down', subjective constraints, that are imposed by the human organism, on the flow and interpretation of information from the outside world, the use of internal information, and the organization and selection of responses in the service of goal-directed behavior. In some cases, it has been shown that attention can be automatically captured by sudden changes in the environment or by well-trained stimulus–response tendencies ('bottom-up' processes). (*See* **Attention, Neural Basis of**; **Visual Attention**)

To illustrate this process, think for example of a person reading a novel in a train, while travelling back home from work. He has to focus attention on the book page and ignore all other visual information, noises of the train, announcements, or passenger conversations. He also has to stabilize the book in his hands and combat interference from train motion, vibration, and acceleration changes. On the book page he needs to focus on one line, ignore all other lines, and progress continuously with reading. Note also that the task of reading calls upon his knowledge of the language, reading ability, and memory of the content of previously read chapters of the book. All are knowledge bases stored in memory. Reading is one of many alternative tasks in which the person can engage himself. He can hum a song, participate in a conversation, or operate a laptop computer. He can also switch from one task to another. Note the difference in the sources of information and responses that have to be attended to or ignored in each of these tasks. Attending to a task is engaging, and limits giving attention to other tasks. The person should be careful to monitor the train route from time to time, so as not to miss his stop.

The above example illustrates many of the phenomena signifying the act of attending. It is also accompanied by distinct subjective experience. Attended information appears vivid and intense, while unattended and ignored information may fade away completely, or appear attenuated and weak.

The study of attention has been a major topic in scientific psychology since its early days. Research has questioned the nature of the underlying processes; the span and limits of attention; the ease and cost of attending to different environmental features and internal events; the ability to control, mobilize, and divide attention; and the linkage

between attention and consciousness. The knowledge and understanding of human attention have greatly benefited from behavioral and neurophysiological studies using animals – in particular, mammals – where direct invasive measurement of brain activity was possible.

ATTENTION AND CONSCIOUSNESS

The relationship between attention and consciousness is of special significance, because early models in psychology equated attention with the content of consciousness. William James, one of the forefathers of scientific psychology, wrote in 1890:

> Every one knows what attention is. It is taking possession by the mind, in a clear and vivid form, of one of what seems several simultaneously possible objects or trains of thoughts. Focalization, concentration of consciousness are of its essence. It implies withdrawing from some things in order to deal effectively with others …

However, we now know that only a very limited portion of the ongoing processing and response activities is admitted to consciousness. Even if triggered by a voluntary, conscious intention, the influences, consequences, and products of this intention on perception, processing, and response, are mostly not admitted to consciousness. The nature of the relationship between conscious and nonconscious products of attention is still a topic of contemporary theoretical debate. It also has an important methodological implication. The study of attention and its influences cannot limit itself to verbal reports of people. Specific research paradigms, employing behavioral and physiological measures, were developed to complement verbal reports and assess the different aspects of attention. Behavioral measures focus on changes in performance speed and accuracy, resulting from a manipulation of attention demands. These measures are complemented by a variety of neurophysiological indices (e.g. heart rate, pupil dilation, brain-evoked potentials), which have been shown to accompany changes in attention requirements. (*See* **Consciousness and Attention**; **Consciousness, Cognitive Theories of**; **Consciousness, Disorders of**)

ASPECTS OF ATTENTION

The study of attention has followed several major task categories: selective attention, divided attention, intensive aspects of attention, and mobilization of attention. They are briefly reviewed below.

Selective Attention

The study of selective attention has addressed two major questions. One is concerned with factors that influence the ease and efficiency of selection. The second examined the consequences of selection for the processing of and response to the selected and the rejected or unattended stimuli. Most of the experiments in selective attention require subjects to perform some kind of a filtering task: that is, subjects are instructed to process, memorize, judge, listen to, or respond only to stimuli that satisfy a specified criterion, for example, 'listen to female and ignore male voices', 'read blue words and ignore red and yellow', 'vocalize and memorize only animal names in a sequence containing animal, plant, and artifact names'. (*See* **Attention, Models of**; **Selective Attention**)

Overall, humans seem to be able to utilize and selectively focus on any group of stimuli that are consistently distinguished by a signifying attribute. Selection is shown to be efficient and easy when the signifying properties are highly distinguishable physical attributes, such as visual location or auditory pitch. In this case there is no long-term memory and little trace of the ignored information. The ease of selection depends on the discriminability of stimuli, with simple physical dimensions easier than semantic dimensions. Indirect behavioral measures indicate that unattended stimuli are also analyzed to a certain level, but in the majority of cases do not reach the level of full semantic analysis.

An interesting aspect of selection is the influence of a preparatory set. Research has shown that when a single stimulus is presented, processing efficiency can derive considerable benefit from advance information on the nature or likelihood of this stimulus. Similarly, performance is impaired when invalid information is presented. These effects are stronger when the number of possible stimuli and their confusability increase. Selective attention is tuned and biased, in anticipation of the forthcoming information.

Divided Attention

In divided attention tasks individuals are asked to attend to or monitor multiple information sources arriving at the same time, or to perform more than one task concurrently. Under such conditions research demonstrated good, parallel, unlimited-capacity monitoring and search capabilities in two major cases. One case is when targets differ from non-targets on a simple, distinct physical feature,

such as color or size. The other case is when targets belong to a very well learned and frequently used symbolic category, and are perceptually easy to distinguish from non-targets (e.g. letters and digits). In all other cases time-sharing and concurrent performance show considerable deficit, compared to single task performance. An interesting phenomenon has been labeled 'attention blink', in which detection of a difficult target impairs performance for a short period thereafter. (*See* **Selective Attention**)

Capacity limitations in dividing attention have been shown to be more severe when targets are defined by complex discriminations (e.g. conjunction of features), rely heavily on working memory (e.g. mental arithmetic), or require coordinated response. The great sensitivity to attention limitations has made divided attention conditions a popular paradigm in the study of mental workload and the cost of mental operations.

Intensive Aspects of Attention

Selection and division of attention do not manifest themselves only in the sources of information and type of events that are attended to or ignored. They also influence the intensity and invested effort in processing and response. This is the intensive dimension of attention. There is ample evidence to show that selection and division do not operate in an all-or-none fashion. Rather, invested effort, which operates like an amplifier or gain factor, influences the response rate and quality of performance on the attended task. Humans can allocate graded levels of effort to the performance of tasks. When fully concentrated on an interesting and demanding task, there is little attention for anything else. When bored by a task, little attention is invested in attending to it; attention is easily distracted, and captured by new events. In divided attention conditions, more than one task is attended to simultaneously, and effort is allocated according to priorities. Think, for example, of the multiple elements comprising the task of driving a car in heavy traffic and the change in the relative importance of each element in different conditions.

Intensity modulations of attention have an autonomous and a voluntary component. The autonomous component is closely linked with the physiological cycles of metabolism, hormones, body temperature, and rest–activity. They jointly influence the arousal, vigilance, and efficiency of the human processing and response system. Research has shown that time of day, and phase in the circadian rhythm, may lead to a change of up to 30 per cent in the efficiency of performance. We are also well aware of the effects of fatigue, exhaustion, jet lag, and sleep deprivation on attention capabilities. The second intensive component is voluntary effort: this represents intensity modulations attributed to the ability of humans to concentrate, focus, and invest graded effort at will. The influence of intensive aspects is of special significance in sustained attention tasks when performers have to attend to a task for extended periods, such as driving for long distances, control room operations, and studying for exams. Such tasks are extremely sensitive to variation in the autonomous component. In addition, it has been shown that the voluntary component can operate at full force for only a limited period, ranging from a few minutes to about half an hour, depending on the demands of the performed task. Tasks performed for extended periods demonstrate an increased number and longer durations of lapses of attention.

Mobilization of Attention versus Modulations of Processing Intensity

An interesting distinction has emerged in the study of attention, between the act of moving attention from one focal point to another, and the act of modulating the intensity of processing once attention has been locked and engaged in a task. Evidence from neurophysiological research on visual attention has shown that these acts are associated with separate brain mechanisms, localized at different brain areas. Corroborating evidence comes from the study of individual differences in attention. It shows that the ability to focus and divide attention, both representing different levels of attention allocation to a task, is distinct from the ability to switch attention between tasks, which represents disengagement, mobilization, and reconfiguration efforts.

THE LIMITS OF ATTENTION

An important topic in the study of attention has been the attempt to identify the limits of attention. How many things can a person do in parallel? How much information can be processed and responded to concurrently? What are the limits on the rate of processing? Theoretical models of attention have advocated two major views of the source of limitations. 'Bottleneck' models associate the limit with the existence of a central, limited-capacity processor that can deal with only one task at a time. This processor constitutes a bottleneck, because a new task cannot access it before processing of the

previous task has been completed. Two variants of this approach have located the bottleneck either early, when only partial or limited analysis of a stimulus has occurred, or late, at the stage of response selection, after stimuli have been fully identified. (*See* **Selective Attention**)

An alternative view of the limitation is proposed by resource models. According to this approach parallel processing is possible, but there is a limited pool of processing and response resources. Resources are allocated to the performance of tasks, performed singly or in parallel, as long as their total demand does not exceed the limit of resources. One variant of this approach is a single general pool of resources which are utilized by all tasks. Another variant advocates a multiple resource view, according to which there are several types of resources which are distinguished by the types of information, processing activities, or response modes that they serve. Performance of tasks is limited only to the extent that jointly performed tasks compete for and exhaust the limits of the same resource.

Proponents of both models conducted extensive experimentation, and brought ample evidence to support their claims. Can both views coexist? A third, and more recent, concept of the limitation of attention proposes to view it as strategic. According to this view there is no mandatory architecture to the flow of task performance, nor a strict universal central limitation and scarcity of processing and response facilities. Rather, given the nature of the task involved, the characteristics of the environment, the specific abilities and skills of an individual, and his motivations and intentions, the best strategic solution to the task's performance is developed. This solution represents the most advantageous combination of all composites at that moment. Once a strategy is developed, it does have its costs and limitations; however, those may change as the conditions are changed, experience accumulates, or utilities vary. The basic idea of the strategic approach is that there are many redundancies in the performance of every task, many degrees of freedom (elasticity) of the human processing system, and a considerable flexibility in using them. Hence, there are many ways in which a single task may be performed, which in turn may change its demand profile. Repeated or very common experiences associated with the performance of some mundane tasks may create the impression of a more 'hard-wired' limitation. However, even this can be changed with extensive training. This view of limitation shifts the focus from the study of the sources of limitation and universal limits, to the study of the degrees of freedom of the human processing system, and the ability of humans to control and efficiently channel their attention capabilities.

ATTENTION AND TRAINING

There is robust evidence to show that experience and practice lead to a dramatic reduction in the attention requirements of tasks. Think, for example, of the differences in the attention requirements of driving, for expert drivers and novice trainees. The difference is attributed to the increased organization of behavior with practice, and the development of well-organized sequences of behavior (schemas) that can operate as one, by a single command. This mode of behavior is labelled 'automatic'. It is contrasted with a 'controlled' mode of operation, which characterizes early stages of responding to new tasks and unexpected events. Controlled processes impose high demands on attention, and in the majority of cases result in less efficient performance.

This relationship between attention demands and training highlights some of the aforementioned functional and phenomenological aspects of attention. First, it underlines the important role of attention during early stages of training, in focusing on relevant information, linking external information with internal knowledge bases, coordinating and synchronizing activities, and developing the overall architecture of a given task. Second, the development of well-organized sequences of processing and response activities that are activated by a single command hints at the possible break between attention and the content of consciousness. It is possible that only the beginning, but not the entirety, of an organized sequence is admitted to consciousness. A related observation is that automated components can capture attention involuntarily and compete with voluntary intents. This finding is easy to interpret, once it is realized that automatic sequences, which are not conscious and only partially controlled, bias the processing and response flow, and attribute significance to events. Finally, training illustrates the way in which attention and performance strategies are linked together, can be developed, and acquire potency. Such strategies may then compete with other tasks, and require considerable effort to change.

The Skill of Attention Control

Intentions and goals are the major drivers of attention. We have seen that humans can focus, divide,

assign priorities, and invest differential efforts in the performance of tasks. However, research has also shown that there are many difficulties and failures in the application and maintenance of attention policies. Can humans be taught to improve the management of their attentional capabilities? The study of this question has shown that there are two major difficulties in the control of attention. One is that people do not have sufficient knowledge of the efficiency of their invested efforts (this is not surprising in view of the limits of consciousness). The second is an execution problem: some attention policies and priority settings are difficult to establish and maintain. Studies that have targeted the development of these components with training have shown that attention control is a skill that can be developed and improved with proper training protocols. Moreover, the skill can be generalized to other situations in which performers face difficult and demanding tasks. Training of attention control was shown to improve the flight performance of trainee pilots, and to help older people to cope with demanding time-sharing tasks.

INDIVIDUAL DIFFERENCES IN ATTENTION

Are there consistent individual differences in attention? Do some people have better attention capabilities than others? The study of these questions showed that there are indeed consistent individual differences in attention, which are generic over and above the specific properties of the tasks that are performed, the modality, and the mode of information and response. Some individuals are better able to deal with high load than others. These studies served for the development of attention ability tests, which were shown to predict performance in flight tasks and the accident-proneness of bus drivers and others. More recent research distinguished between the attention acts of focusing, dividing, and switching between tasks. Consistent individual differences were revealed on all three dimensions. However, focusing and dividing capabilities were shown to be highly correlated, while attention-switching appears to constitute a separate ability. The research has also shown the existence of a general factor of attention control ability, which is common to all acts of attention.

NEUROPHYSIOLOGY OF ATTENTION

Most of our knowledge of the neurophysiology and brain mechanisms of attention in normal-performing humans derives from two classes of measurement techniques. One is the recording from the scalp of the electrical response of large groups of neurons to specific events (event-related brain potentials or ERPs). The second is the measurement of regional cerebral blood flow (positron emission tomography, PET, and functional magnetic resonance imaging, *f*MRI).

Data from ERP and neuroimaging studies together with the behavioral observation of patients with brain damage suggest that several brain areas are involved in attentional processing (Figure 1). The brain systems that have been associated with attention processes can be described as an interconnected network of cortical and subcortical structures which include the prefrontal and posterior parietal lobes, the cingulate gyrus, the basal ganglia, the pulvinar and reticular nuclei of the thalamus, and the midbrain and superior colliculus. These areas have been shown to be sensitive to attention-related manipulations, and control attention through their inputs to perceptual processing areas. ERP studies have also been very useful in identifying the stages of information processing at which attention begins to exert control.

Recent studies have attempted to connect different brain areas to specific aspects of attention. The posterior attentional system, including the parietal cortex, the superior colliculus and the pulvinar, has been shown to be involved in aspects of attention mobilization (directing attention to what has been selected as the focus of attention, engaging and disengaging attention from locations or objects). The two brain hemispheres seem to be involved in different ways in attention control, with the right hemisphere controlling the mobilization of attention to both sides of space and the left hemisphere controlling only the orienting of attention to the right side of space. The anterior attentional system, including the prefrontal cortex and the cingulate gyrus, is responsible for executive control processes. In particular, the prefrontal cortex exerts top-down influence on early sensory processing,

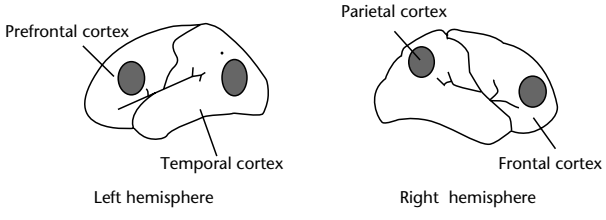


Figure 1. Lateral view of the right and left hemispheres of the human brain. The figure shows the main areas involved in attention control (gray patches).

determining which stimuli entering the perceptual pathway have to be facilitated or suppressed. This area is extensively connected to numerous cortical, limbic, and subcortical areas and seems to play a crucial role not only in attention to external events, but also to the internal mental representations that contribute to working memory, decision-making, and planning. The vigilance system, including the right frontal lobe and the brain stem, is responsible for maintaining readiness during sustained attention and vigilance tasks. (*See* **Attention, Neural Basis of**; **Spatial Attention, Neural Basis of**)

Attention Disorders

As already mentioned, attentional processing seems to involve several brain areas. Research with patients has related deficits in specific aspects of attention to impairment in one or other of these areas. Two major classes of disorder have been documented: impairment in spatial attention, and deficits of attention control. These are briefly described below.

Neglect

Visual neglect is a neurological syndrome that commonly results from lesions of the right cerebral hemisphere, especially from lesions involving the temporoparietal cortex. Although vision is intact, neglect patients are not aware of and fail to report or respond to objects or parts of objects presented controlateral to the lesion (that is, in the left visual field). This disability affects many aspects of their life. For example, patients typically fail to eat the food located on the left side of their plate. They shave or make-up only one half of their face. When asked to copy a picture, they draw only one half of that picture. Patients may no longer recognize the left side of their body as their own. Neglect has been observed in auditory, tactile, and proprioceptive modalities and can also affect mental imagery; for instance, patients may neglect the left side of an imagined mental map. Despite the fact that neglect is considered a disorder of spatial attention, the exact explanation of neglect is still a controversial matter. The combination and severity of symptoms shown by different patients seems to depend mostly on the extent and location of the lesion. (*See* **Consciousness and Attention**; **Attention, Neural Basis of**)

Extinction

This deficit is often associated with unilateral neglect to the extent that some accounts consider it a mild form of neglect. Like neglect, extinction follows unilateral brain damage. Patients suffering from extinction have no difficulty in identifying a single object presented on either side of the visual field, but when two stimuli are presented simultaneously, they do not seem to 'see' the object presented in the visual field controlateral to the lesion.

Balint syndrome

This neurological syndrome is associated with symmetrical lesions to the hemispheres (mostly involving posterior parietal areas or the parietal–occipital junction). It is less common than neglect. It involves three major deficits. First, subjects have difficulties in orienting to visual stimuli or in tracking a moving object. Second, subjects are unable to orient their arm and hand correctly when trying to reach or grasp objects. Third, subjects are able to see only one object at a time (simultagnosia), even when objects overlap. For example, if presented with an apple and a knife and asked to peel the apple, they are unable to follow the instructions because they can see only one of the items. (*See* **Consciousness and Attention**; **Attention, Neural Basis of**)

Frontal lobe syndrome

After lesions of the frontal lobe, patients show disorganized and incoherent behaviour. People suffering from this syndrome, also known as dysexecutive syndrome, are unable to plan actions and to solve problems. They show behavioural rigidity and perseveration, that is they are unable to change their behaviour when required to; increased distractibility accompanied by difficulties in focusing and maintaining concentration; difficulties in inhibiting unwanted behaviours; and difficulty in maintaining goal-directed behaviour. These symptoms seem to be caused by the impairment of areas involved in the top-down control of information processing, such as the prefrontal cortex.

Attention deficit hyperactivity disorder

This disorder appears in childhood; the major features are excessive and impairing levels of activity, inattention, and impulsiveness. The diagnosis is often reached because of the difficulties shown by these children at school. In general, the inattention symptoms include difficulties in remaining seated when required to (for example, at the dinner table or in the classroom); inability to follow instructions; inability to concentrate and focus attention on details; easy distractibility by external stimuli; inability to interrupt an action once initiated; and difficulty in organizing tasks and activity. Some

current accounts of ADHD suggest the possibility that children with ADHD may be impaired in executive control processes. (*See* **Attention Deficit Hyperactivity Disorder**)

Further Reading

Gopher D and Koriat A (eds) (1998) *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*. Cambridge, MA: MIT Press.

Inui T and McClelland L (eds) (1996) *Attention and Performance XVI: Information Integration in Perception and Communication*. Cambridge, MA: MIT Press.

Kahneman D (1973) *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Meyer DE and Kornblum S (eds) (1992) *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Monsell S and Driver J (eds) (2000) *Attention and Performance XVIII: Control of Cognitive Processes*. Cambridge, MA: MIT Press.

Parasuraman R (ed.) (1998) *The Attentive Brain*. Cambridge, MA: MIT Press.

Pashler H (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Posner M, Petersen SE, Fox PT and Raichle ME (1988) Localization of cognitive operations in the human brain. *Science* **240**: 1627–1631.

Styles AE (1997) *The Psychology of Attention*. Hove, UK: Psychology Press.

Attitudes

Introductory article

George Y Bizer, Eastern Illinois University, Charleston, Illinois, USA
Jamie C Barden, Ohio State University, Columbus, Ohio, USA
Richard E Petty, Ohio State University, Columbus, Ohio, USA

CONTENTS

An attitude is a global and relatively enduring evaluation (e.g. good or bad) of a person, object, or issue. Attitudes can be based on affective, cognitive, or behavioral information and can vary in their strength (e.g. how enduring, how resistant to change, and how predictive of behavior they are).

INTRODUCTION

What drives our behavior? When we choose a candy bar at the grocery store or decide for whom to vote in an election, what determines the choices that we make? Attitudes, the mental representations of what we like and dislike in our world, help to explain these choices.

Attitudes are one of psychology's fundamental concepts because they help to explain people's decisions and actions. An attitude is a global and relatively enduring evaluation of a person, object, or issue – a representation of whether we think the target is generally good or bad, desirable or undesirable. We can hold attitudes towards tangible objects such as ice cream or trees, people such as the President or a sister, ideas such as democracy or wealth, and issues such as the death penalty or tax increases. Simply put, the more favorably we evaluate something, the more positive our attitude towards the object is; the more negatively we evaluate something, the more negative our attitude is.

Attitudes serve various functions. As noted by Daniel Katz and others, some attitudes serve a utilitarian function in that they help us to achieve rewards and avoid punishments (e.g. having the correct evaluation of one's mortgage company can save you money). Other attitudes serve an ego-defensive function in that they foster our own self-images (e.g. holding prejudiced attitudes might make some people feel superior). A number of additional functions have also been identified.

MEASUREMENT

Researchers have developed a wide array of tools to measure attitudes. These techniques can be categorized into two broad groups. *Explicit* measures directly ask people to report their attitudes; in contrast, *implicit* (or indirect) measures are assessments that allow inferences about a person's attitude without having to ask him or her directly. The latter method is commonly used in situations in which people either do not want to or are unable to provide their true evaluations of an object. No measure of attitudes is perfect, however, as assessments can be affected by the measurement context. Seemingly innocent influences like the weather or answers to previous questions can have a considerable impact on people's reported attitudes.

Explicit Measures

Two common explicit measures are the *Likert scale* and the *semantic differential*. The Likert procedure presents respondents with series of evaluative statements along with a series of response options for each. For example, an attitude scale on ice cream might contain the statement 'Ice cream tastes good' and choices of various degrees of agreement from which the respondent can choose ('strongly agree', 'agree', 'neither agree nor disagree', 'disagree', and 'strongly disagree'). Participants report the extent to which they agree or disagree with each statement. Likert scales include a wide variety of evaluative statements regarding the object, and scores from all the statements are combined to create a measure of the attitude.

With the semantic differential technique, developed by Charles Osgood and colleagues, respondents are presented with the name of the attitude object and some evaluative adjectives that might describe that object. Participants then rate how well the adjectives describe the object. For example, a series of items might prompt respondents to report the extent to which they think ice cream is beneficial versus harmful, good versus bad, and pleasant versus unpleasant. These scores are combined to form one global attitude measure.

Implicit Measures

Implicit measures come in various shapes and sizes. They range from monitoring simple behaviors from which evaluation can be inferred (e.g. how close one person chooses to sit next to another) to complex physiological techniques. A good example of an implicit measure is Russell Fazio's priming procedure. To assess racial attitudes with this technique, participants are presented with images of Caucasian or African-American faces to make the concept of one or the other race more accessible. Immediately after being shown a face, participants are asked to judge whether a particular concept (e.g. ice cream) is 'good' or 'bad'. Over many pairings of faces and concept words, the amount of time it takes the participant to report 'good' or 'bad' for each word following a face is measured. The pattern of reaction times is used to infer the person's implicit attitude. Since one negative attitude tends to activate or prime others, if a participant dislikes African Americans, showing an African-American face should make the evaluations of other negatively perceived objects (e.g. 'dirt') faster, but make the evaluations of positively perceived objects (e.g. 'ice cream') slower. Thus, if a person needs more time to report that a good word like 'ice cream' is 'good' after seeing an African-American face and less time to report that a bad word like 'dirt' is 'bad' (compared to seeing a Caucasian face), there is evidence that the person holds a more negative attitude towards African Americans than Caucasians.

TRIPARTITE MODEL OF ATTITUDE STRUCTURE

As global and enduring evaluations, attitudes can be based on up to three separate types of input: affective, cognitive, and behavioral. An attitude can be based on any one or a combination of these three information sources. Attitudes, once formed, also guide affective, cognitive, and behavioral reactions to the object.

The *affective* basis of an attitude is made up of feelings, moods, and emotions that have become associated with the attitude object through past or current experience. It is possible to have multiple affective responses to an object based on the same, or different, experiences with it. Each affective response has an evaluation made up of valence (positive to negative) and magnitude (strong to weak). Researchers often measure the affective basis of the attitude by asking to what extent individuals *feel* good or bad about the object, or the extent to which the attitude object makes them feel 'happy', or 'disgusted', or 'angry'.

The *cognitive* basis is made up of particular attributes that are ascribed to the object. An *attribute* is any characteristic, quality, trait, concept, value, or goal associated with the object. The impact of an attribute is determined by the evaluation of whether the attribute is good or bad, and the perceived likelihood that this attribute applies to the object. Thus, if the attitude object is 'ice cream', one attribute associated with this object might be 'fattening'. If the person thought this attribute was negative and highly associated with ice cream, the attribute would contribute to a generally unfavorable evaluation of ice cream. Of course, any one attitude object can be associated with many attributes that contribute to the overall evaluation.

In practice, researchers such as Martin Fishbein and Icek Ajzen suggest obtaining a listing of attributes about an object and then the evaluation and likelihood associated with each attribute. The evaluation and likelihood are multiplied together for each attribute and the products are added across attributes to estimate the cognitive component of the attitude. One implication of this approach is that two individuals endorsing the same attributes can have different attitudes if they evaluate the attributes differently, and individuals believing in different attributes can hold the same attitude.

The *behavioral* basis is made up of two kinds of information, past behaviors and intentions to commit future behaviors. Daryl Bem's *self-perception theory* holds that we sometimes infer our attitudes directly from our past behaviors towards an object. For example, if a person looks back on his or her life and realizes that he or she has never eaten at a Chinese restaurant even though he or she had many chances to do so, the person might infer that he or she does not like Chinese food. This inference occurs as long as there is no memory of external forces compelling the behavior – the behavior needs to be seen as voluntary.

CONSTRUCTED VERSUS STORED ATTITUDES

When an object is encountered for the first time, there is no information about it in memory. An attitude must therefore be *constructed* by making inferences from the behaviors, thoughts, and feelings that occur in the current social environment. Irrelevant features of the current context can bias the constructed attitude even though they have little to do with the attitude object itself. For example, one's attitude towards the economy might be more favorable on a sunny than a rainy day. Norbert Schwarz and colleagues have documented a wide variety of contextual influences on attitude reports.

After information is gathered about an object and an evaluation is formed, the attitude can be stored in memory and subsequently retrieved directly. A *stored attitude* is an evaluation that is linked to the object in the form of a thought (e.g. 'I like candy'). If the attitude object is brought to mind again, and the object–evaluation link is strong enough, the stored attitude is brought to mind as well.

Generally, the attitudes people report fall somewhere in between purely constructed and purely retrieved. That is, even if a person has already formed a global evaluation, the specific evaluation that is reported at any moment in time is dependent on a wide variety of factors. In general, the stored attitude acts as an anchor point and is adjusted based on affective, cognitive, and behavioral information that is currently salient in memory or in the immediate context. That is, when an attitude is retrieved, some information related to that attitude may also be retrieved and pull the attitude report in its direction. For example, on one occasion the 'taste' attribute of ice cream might be especially salient, but on another occasion, the 'fattening' attribute might be more salient. The immediate context can influence which attitude-relevant information is retrieved, providing a source of contextual bias. Because 'strong' attitudes are less likely to be influenced by context effects than are 'weak' attitudes, the study of attitude strength is also important in attitude research.

STRONG AND WEAK ATTITUDES

Attitudes fall along a continuum from weak to strong, such that stronger attitudes are more durable and impactful. A durable attitude is persistent over time, meaning that it does not decay in memory, and is resistant to change when faced with counter-attitudinal information. An attitude has impact when it influences information processing and guides behavior. Attitudes can possess these strength properties to varying degrees. Also, these strength properties can be independent. Thus, it is possible for an attitude to persist over time but not influence behavior, or to greatly influence thought at a given point in time, but not resist attempts to change it.

A number of variables contribute to making attitudes stronger or weaker. First, extreme attitudes (i.e. when people rate the object as intensely good or intensely bad) tend to be stronger than more moderate attitudes. This may be because extreme attitudes tend to have a number of structural properties that contribute to this strength. For example, extreme attitudes may be based on high amounts of consistent knowledge, and they may come to mind more rapidly (i.e. are more accessible) than more moderate attitudes.

Subjective beliefs about our attitudes are also related to strength. For example, strength can result from perceptions of how much knowledge we have on a topic (regardless of actual knowledge), how important the attitude object is to us personally, or how confident we are in the validity of our attitudes. Finally, the manner in which an attitude is formed can contribute to its strength. Most notably, if an attitude is created through extensive thinking and careful scrutiny of information, it tends to be stronger than if it was formed by means requiring less effort.

EXPLICIT AND IMPLICIT ATTITUDES

To this point, we have discussed attitudes as global evaluations we are aware of and can control. These conscious or *explicit attitudes* result from integrating information from one or more components into an evaluation. These attitudes can vary from strong to weak, and from mostly stored to mostly constructed. Retrieval or construction of these explicit attitudes can either be relatively automatic or require considerable cognitive effort.

In addition to these explicit attitudes, researchers such as Anthony Greenwald and Mazarin Banaji have argued that people can also hold *implicit attitudes* – attitudes of which they are generally not aware. Implicit and explicit attitudes can sometimes be in opposition to each other, such that the implicit attitude can lead people to think and behave in ways they do not consciously intend. For example, a person who holds a prejudiced implicit attitude based on negative stereotypes learned as a child, but now consciously rejected,

may also hold an explicit unprejudiced (and conscious) attitude learned later in life. In such situations, conscious attitudes direct behaviors that are generally under constant conscious control (such as deciding the guilt or innocence of a black defendant). However, more automatic behaviors such as one's body language and eye contact can reflect a person's implicit attitudes.

ATTITUDE CHANGE

Although attitudes are generally considered to be relatively stable and enduring, they are subject to change over time. Being exposed to new information and new experiences can lead people to change their attitudes. Numerous studies have demonstrated the processes by which attitudes change.

Central and Peripheral Routes to Change

Much contemporary research is guided by the idea that attitudes are sometimes changed thoughtfully, but sometimes are changed with very little cognitive effort. The *elaboration likelihood model* (ELM) of persuasion developed by Richard Petty and John Cacioppo presents a framework that helps explain the various processes and outcomes of attitude change. Although the amount of thinking involved in attitude change forms a continuum from none to extensive, the model divides the specific processes of attitude change conveniently into two 'routes' to persuasion.

The first or *central route* to attitude change occurs when people are relatively careful in scrutinizing the issue-relevant information available. If, after careful consideration, a person finds the information to be compelling, attitudes change accordingly. If, however, the information is deemed specious, attitudes will not change, or can even change in a direction opposite to that advocated – a boomerang effect. For example, a person following the central route when processing a magazine advertisement for a car will carefully assess the perceived validity of the information presented in the ad. The person might examine the information presented about the horsepower, price, resale value, safety record, and so forth. The person will be influenced if his or her *cognitive responses* – the thoughts generated during message processing – are positive. The person is not likely to be influenced, however, by the beautiful sunset pictured in the background or the cute puppy sitting in the driver's seat because these are peripheral cues that are unrelated to the central merits of the car.

Sometimes, however, people follow the *peripheral route* when exposed to a persuasive communication. In such cases, people are not likely to pay attention to all of the issue-relevant information within the message. Rather, people seek a shortcut way to evaluate the ad. In this case, they might be influenced by the mere number of arguments in the ad (regardless of their quality). Or, the cute puppy in the driver's seat might lead to a positive feeling that becomes associated with the car.

An important consequence of the route to persuasion that a person takes is the strength of the attitudes that result. Specifically, when people change or form an attitude through the central route to persuasion, attitudes tend to be stronger than those created or changed through the peripheral route. Attitude changes that occur because a person has carefully considered issue-relevant information have a substantive backing which contributes to the durability and impact of the attitude. In contrast, attitudes formed under the peripheral route do not have this substantive support. Because they lack supporting cognitions, these attitudes are much less durable and impactful. This does not mean that peripheral route changes are completely unimportant. For example, advertisers can take advantage of the short-term effects of the peripheral route by continual pairing of peripheral cues with their products in repeated messages. Also, in some cases, what starts out as a peripheral cue can become an argument if people subsequently think about the cue in a way that gives it substantive meaning.

Amount of elaboration

What determines whether a person will follow the central or the peripheral route? This depends on whether the person has the *ability* and the *motivation* to think carefully about the message. Variables influencing ability (whether a person is able to think) include distraction and time pressure. If a person is distracted or under great time pressure while exposed to a persuasive communication, it is simply not possible to follow the central route, and thus the peripheral route to persuasion is more likely. Variables influencing motivation (whether a person wants to think) include personal relevance and accountability. For example, if people are told that a message is of low relevance to them (the product advertised is available only in a faraway country), or that they will not be accountable for the attitude they report on the topic (questionnaires will be completed anonymously), it is likely that they will feel little motivation to think carefully. They are likely to follow the less taxing peripheral

route to persuasion instead. All else being equal, people prefer to save their cognitive energy for the most important tasks and decisions in life. People who are high in their *need for cognition* tend to enjoy thinking about a wide variety of topics and thus tend to follow the central route to persuasion. People who are low in this need tend to follow the peripheral route.

Objective and biased processing

In addition to the amount of information processing that takes place, it is also important to consider whether that processing is relatively objective or biased. Objective processing refers to the case in which thinking is guided by the qualities of the information at hand. If the information is cogent, people's thoughts are favorable, but if the information is specious, their thoughts are largely negative. However, people can process messages in a biased manner. For example, people may be forewarned that a message will attempt to change their attitudes. In such cases, people tend to think negatively about all of the arguments – regardless of their actual quality – in an attempt to assert their individual freedom not to be influenced. There are a number of motivations besides asserting freedom that can induce biased processing, such as the motive to be consistent, or to maintain one's self-esteem. Each of these motivations selectively directs people's information-processing activity to favor one attitudinal position over another.

Multiple roles of variables

The ELM highlights the fact that variables can influence attitudes by serving in different roles in different situations. For example, the physical attractiveness of the source of a persuasive message might influence attitudes in a number of ways. First, such a source might serve as a simple peripheral cue when the situation constrains people's motivation or ability to think about the message to be low. For example, when people are distracted, they might go along just because the source is attractive, regardless of the merits of what the source says (peripheral route). On the other hand, if people are highly motivated and able to think about the message, an attractive source might bias that thinking in a favorable way, or the source itself might be scrutinized to see whether it provides information central to the merits of the issue (central route). Finally, if thinking is not already constrained to be high or low, an attractive person might encourage recipients to pay more attention to the message – leading to more agreement if the message is sound, but to less agreement if the message is not. Under this scenario, then, attractiveness would serve as a determinant of elaboration.

Thus, although a variable can serve as a determinant of elaboration in one scenario (unconstrained elaboration), it can serve as a cue in others (low elaboration) or can bias processing or serve as an argument in still other situations (high elaboration). The ELM thus limits the fundamental roles a variable can play in persuasion situations and provides a guiding framework for assessing when variables take on each role.

Mood and Persuasion

One persuasion variable that has been studied extensively in its multiple roles is a person's mood state – whether he or she is feeling good or bad. As with other variables, the effect of mood on persuasion depends on the amount of elaboration taking place during the message presentation. Under low-elaboration conditions, a person's mood can serve as a peripheral cue. In such situations, people may associate their mood with the message's object. The pairing of a good mood with the object can produce positive attitudes towards that object, but bad mood can produce negative attitudes. Second, when elaboration is high, a positive mood can bias people's reactions to the message arguments. In particular, positive mood states make good consequences (e.g. living longer if you stop smoking) seem more likely than when in a neutral or negative mood state, but make negative consequences (e.g. getting cancer if you don't stop smoking) seem less likely. Negative mood states have the opposite consequences. Finally, under moderate elaboration conditions, positive moods affect the amount of thinking people do about the message. If the message appears to be negative or depressing, positive moods decrease information processing compared to negative moods. People in positive moods do not want to think about negative information. On the other hand, if the message appears to be positive or uplifting, positive moods increase thinking over negative moods.

Persuasion from Our Own Behavior

Leon Festinger's theory of *cognitive dissonance* suggests that sometimes our own behavior can lead to attitude change. Specifically, the theory holds that cognitive conflict occurs when people believe that they have behaved in a way that is inconsistent with their attitudes. This cognitive conflict produces tension that people are motivated to reduce in order to restore consistency. Since behavior is

often difficult to undo, one way to restore attitude–behavior consistency is to change one's attitude to be in line with one's behavior. This is not the only way to reduce dissonance (people could reduce the importance of the conflicting attitude or behavior), but it is a common one. Dissonance theory explains such processes as why people come to favor products more after they purchase them, and why people come to like groups more if they have voluntarily exerted considerable effort to join them. Dissonance can also lead people to process information in a biased fashion. That is, they think about attitude objects in a way that restores consistency.

Resisting Persuasion

Although there are many different ways in which people can be persuaded, there are some techniques through which attitudes can be made more resistant to change. One way to create resistant attitudes is simply to make those attitudes stronger. This can be done by increasing issue-relevant thinking prior to the attacking message. For example, if a positive attitude towards obeying the speed limit is weak, a person could spend time thinking and learning about why he or she holds the attitude. This additional thinking and learning should serve to strengthen the once-weak attitude.

Perhaps surprisingly, an attitude can in some situations be made more resistant to persuasion through attempted counter-persuasion! Some attitudes are weak because they have very little substantive basis at all. These attitudes may have been created through peripheral processes or may simply be 'cultural truisms' – attitudes we hold just because we have always been taught to think that way (such as favorable attitudes towards freedom of speech). If cultural truisms are mildly attacked, they can actually become stronger. This process of *inoculation*, as outlined by William McGuire, occurs because although the weak attack may not be enough to change the attitude, it may be strong enough to make the recipient think (often for the first time) about why he or she holds that attitude in the first place. This additional thought can serve to create a basis for holding the attitude and motivate individuals to effectively counter-argue subsequent attacking messages.

ATTITUDE – BEHAVIOR CONSISTENCY

One reason why attitudes are a principal area of research in psychology is that, under the right circumstances, attitudes guide people's behavior, and thus are useful to know in order to predict voting, consumer purchases, jury decisions, and so forth. According to Russell Fazio's model of attitude–behavior consistency, exactly *how* attitudes guide behavior depends on the type of behavior in question – is the behavior one that is engaged in spontaneously or one that elicits reflection prior to action?

Some behaviors in which we engage are not well thought out. When it comes to impulse purchases, such as the candy people buy while waiting at the checkout line, people may not spend much time in making decisions. In such cases, people simply look to their attitudes to make a choice. In such situations, whether our attitudes drive behavior is determined by whether we can recall the attitude easily and quickly (i.e. if the attitude is accessible).

Other behaviors are not as spontaneous. When we have the motivation and opportunity to choose our behaviors more carefully, accessibility alone is less important. Instead, according to Fishbein and Ajzen's theory of reasoned action, behavior is determined by one's behavioral intention, which is in turn determined by several factors. First, intention is determined by one's attitude towards the particular behavior under consideration. Attitude towards the behavior in any given situation (e.g. baking a cake for your spouse's birthday on Wednesday) will depend on the beliefs that come to mind in assessing this action – if an attitude relevant to this behavior is not readily retrievable. Intentions are also determined by subjective norms – what other people we admire would want us to do in the situation. Perceptions of our own abilities to carry out some action also play an important role in determining deliberative behaviors.

CONCLUSION

Attitudes have a profound impact on virtually every aspect of our lives. From fundamental issues, such as how attitudes can be measured, to more complex ones, such as the nuances of how attitudes can be changed, a long and rich array of research has helped us to understand the nature of the attitude construct. Space limitations have allowed us to provide only a sampling of what is known about attitudes; the extensive literature on this topic requires further exploration.

Further Reading

Bem DJ (1972) Self-perception theory. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, vol. 6, pp. 1–62. New York, NY: Academic Press.

Eagly AH and Chaiken S (1993) *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Eagly AH and Chaiken S (1998) Attitude structure and function. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, pp. 269–322. Boston, MA: McGraw-Hill.

Fazio RH (1990) Multiple processes by which attitudes guide behavior: the MODE model as an integrative framework. In: Zanna MP (ed.) *Advances in Experimental Social Psychology*, vol. 23, pp. 75–109. San Diego, CA: Academic Press.

Festinger L (1954) *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fishbein M and Ajzen I (1975) *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Greenwald AG and Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review* **102**: 4–27.

Katz D (1960) The functional approach to the study of attitudes. *Public Opinion Quarterly* **24**: 163–204.

McGuire WJ (1964) Inducing resistance to persuasion: some contemporary approaches. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, vol. 1, pp. 191–229. New York, NY: Academic Press.

Osgood CE, Suci GJ and Tannenbaum PH (1957) *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Petty RE and Cacioppo JT (1986) The Elaboration Likelihood Model of persuasion. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, vol. 19, pp. 123–205. New York, NY: Academic Press.

Petty RE and Krosnick JA (eds) (1995) *Attitude Strength: Antecedents and Consequences*. Hillsdale, NJ: Lawrence Erlbaum.

Petty RE and Wegner DT (1998) Attitude change: multiple roles for persuasion variables. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, pp. 323–390. Boston, MA: McGraw-Hill.

Schwarz N (1999) Self-reports: how the questions shape the answers. *American Psychologist* **54**: 93–105.

Zimbardo PG and Leippe MR (1991) *The Psychology of Attitude Change and Social Influence*. New York, NY: McGraw-Hill.

Auditory Event-related Potentials

Intermediate article

Terence W Picton, Rotman Research Institute, Toronto, Ontario, Canada

CONTENTS

Auditory event-related potentials are electrical changes recorded from the brain in association with auditory stimuli.

INTRODUCTION

Human auditory event-related potentials (ERPs) are evoked by an acoustic stimulus and indicate the cerebral activity occurring as that stimulus is processed in the brain. When recorded from the human scalp, ERPs are intermixed with other electrical potentials. Several techniques may be used to distinguish ERPs from these other activities (Picton *et al*., 1995): one of the most common is averaging – when the responses to multiple stimuli are averaged together, the auditory ERP (which is generally constant from one stimulus to the next) remains the same, whereas the other activity (which varies from one stimulus to the next) attenuates. (*See* **Electroencephalography (EEG)**)

Auditory ERPs can be classified in many ways. In relation to the stimulus, the ERPs may be transient, steady state, or sustained. Transient responses are evoked by rapid changes such as the onset or offset of a sound. Transient evoked potentials are often named according to their polarity and latency, e.g. the N100 is a negative wave recorded with a typical peak latency of 100 ms. Steady state responses are evoked by a regularly changing stimulus such as an amplitude-modulated tone. These can be described in terms of the frequencies at which they are elicited, e.g. the 40 Hz ERP is best recorded when stimuli are presented at rates near 40 Hz. Sustained potentials are recorded through the duration of a stimulus. Some ERPs are characterized by where they are generated in the brain, e.g. the auditory brainstem response, others by where they are maximally recorded, e.g. the vertex potential, and others by their purported function, e.g. the mismatch negativity. The plethora of nomenclatures reflects the paucity of our understanding.

In order for potentials to be recorded at a distance from where they are generated, many neurons must respond, their activation must be synchronous, and the neurons must be oriented in a similar direction so that their fields combine rather than cancel. Much of what goes on in the auditory nervous system does not fulfill these criteria and passes unnoticed in the scalp-recorded ERP. What is recorded, however, can tell us much about the processing of sounds in the brain.

SENSORY PROCESSING

A brief transient stimulus evokes a sequence of potentials that indicate the processing of stimulus information all the way from cochlea to cortex (Figure 1). The upper part of the figure shows an early sequence of vertex-positive waves (numbered with Roman numerals). These waves are usually lumped together as the auditory brainstem response, even though wave I is generated in the auditory nerve rather than the brainstem. Since it can be recorded down to intensities near threshold, wave V can demonstrate that the brain is receiving sounds without the person being tested having to make a subjective response to those sounds. This 'objective' audiometry is important when evaluating the hearing of infants and others who cannot provide reliable behavioral responses.

The middle part of Figure 1 shows a sequence of waves named according to their polarity and alphabetical sequence. The most prominent of these middle-latency waves is Pa, with a peak latency of about 25 ms. The Pa is most likely to be generated in or near the primary auditory cortex. The initial activation of the human primary auditory cortex probably occurs at about 15 ms, but this may not be reliably recorded from the scalp. If stimuli are presented at rapid rates, the middle-latency waves evoked by one stimulus superimpose on those evoked by preceding stimuli to give a periodic response at the frequency of stimulation.

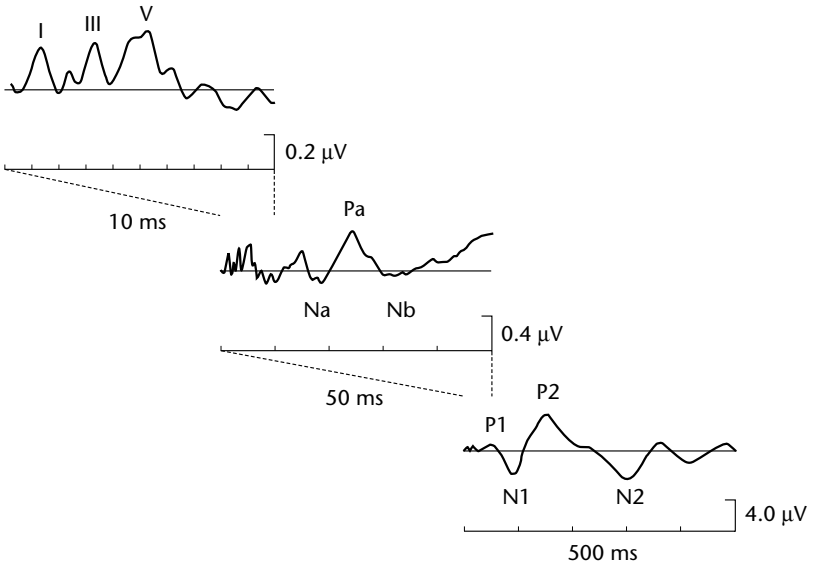


Figure 1. Auditory event-related potentials recorded from the vertex relative to the right mastoid in response to clicks at 60 dB above threshold presented at a rate of 1 s^{-1}. This recording was from a single person showing typical responses. The largest waves occur with latencies between 50 ms and 500 ms, as seen in the lower right waveform. The first 50 ms of this recording can be expanded in time and amplitude to show the middle-latency responses. The first 10 ms of this recording can then be expanded to show the auditory brainstem response.

The most prominent of these steady state responses occurs at 40 Hz (Galambos *et al*., 1981). This response is attenuated by sleep and blocked by anesthesia. The 40 Hz ERP is probably related to bursts of cortical activity that occur at frequencies of 20–50 Hz (the 'gamma band'). These rhythmic responses may reflect the oscillatory transfer of information between sensory areas (Singer, 2000).

The lower part of Figure 1 shows a sequence of waves named according to polarity and numerical sequence. Since these waves are much larger than the earlier waves, they were historically the first to be recognized, and were called 'vertex potentials' because they were maximally recorded from the top of the head. However, since the earlier waves are often also largest at the vertex, the terminology is more nostalgic than definitive. The most prominent of the late waves is N1, which has a peak latency of about 100 ms. Several different intracerebral generators contribute to the scalp-recorded N1 (Picton *et al*., 1999). As well as being evoked by the onset of a stimulus, an N1 also follows the offset of a stimulus or a change in any of its attributes. The N1 wave remains an enigma. Why does such a large response occur so late in the processing of auditory information? The N1 may represent a process whereby the nonspecific detection of a change in the auditory world, perhaps through the brainstem reticular system, initiates a read-out of specific information from the auditory cortex to other regions of the cortex.

SENSORY MEMORY

Memory for a stimulus can be demonstrated by a change in the response to that stimulus when it occurs at a later time. A common demonstration of memory involves the habituation of a response when a stimulus is repeated. This is most efficiently measured in ERPs by recording the responses at different interstimulus intervals. The N1 wave of the auditory ERP decreases quite strikingly with decreasing interstimulus intervals. This decrease is specific to the stimulus attributes; for example, if the tonal frequency of a test stimulus is different from the frequency of a repeating stimulus, the N1 is larger than when the stimuli are the same.

In addition to a change in the N1, a later mismatch negativity (MMN) occurs when a repeating auditory stimulus changes. In this context, the repeating stimulus is usually called the 'standard' and the changed stimulus the 'deviant'. The MMN is typically measured in a difference waveform obtained by subtracting the response to the standard stimulus from the response to the deviant. This subtraction removes elements of the response that are common to both stimuli, leaving the MMN (Figure 2). The MMN differs from the N1 in many ways. First, it has a longer latency, typically

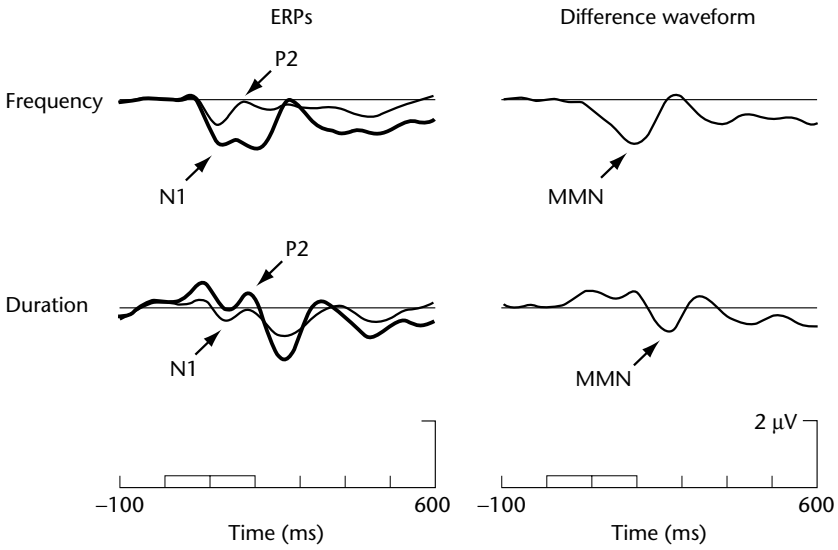


Figure 2. The mismatch negativity (MMN) in response to changes in the frequency or duration of a brief tone. The standard stimuli were 1000 Hz tones with a duration of 200 ms and an intensity 60 dB above threshold. Recordings were obtained from the midfrontal scalp using an average reference. Deviant stimuli occurred with a probability of 0.2. For the upper tracings the deviants had a frequency of 1100 Hz; for the lower tracings the deviants had a duration of 100 ms. Deviant–standard difference waveforms are shown on the right. These event-related potentials (ERPs) are averaged over ten test participants. Deviant ERP, thick line; standard, thin line.

peaking about 50 ms later than the N1. Second, this latency increases with decreasing difference between the deviant and the standard. Third, the latency is determined not by the onset of the stimulus but by the time when the deviant can be distinguished from previous stimuli. When the deviance is a change in duration, the MMN occurs approximately 150 ms after the time of the shorter stimulus, regardless of whether the deviant is the shorter or the longer stimulus. Fourth, the MMN is more related to the size of the deviance than to the parameters of the stimuli. A decrease in the intensity of a stimulus can thus elicit a MMN even though the N1 is smaller for the less intense stimulus. Fifth, the MMN has a different scalp topography, with a maximum amplitude anterior to that of the N1. The intracerebral sources for the MMN may vary with the nature of the deviance, for example being larger in the left hemisphere for changes in speech sounds (Näätänen, 2001).

The MMN is related to sensory (or echoic) memory through the time during which a sensory regularity must be detected. The system generating the MMN must recognize that the deviant does indeed break some perceived regularity in the stimuli. It can only do this when the information is maintained in sensory memory (Picton *et al.*, 2000). Since the MMN occurs whether or not the person is attending to the stimuli, it seems to represent an automatic processing of stimulus changes. However, attending to the stimuli can enhance the wave (Woldorff *et al.*, 1998), particularly when detecting the deviance is more complicated, such as noting a change in the temporal pattern of the stimuli (Alain and Woods, 1997).

The role of the MMN is unknown. It may indicate the increased processing that is evoked by a deviant stimulus, either to detect it as different from the standard or to initiate processing once it is detected. In the latter case, it might serve to alert other parts of the brain, particularly the right frontal regions, that a deviant has occurred.

AUDITORY ATTENTION

The effects of attention on the auditory ERP can be studied by presenting two or more trains of information, like hearing different conversations at a party, and asking the participant to attend to one and ignore the others. Attention can be monitored by measuring how well the participant detects occasional changes in the stimuli ('targets'). The difference between the ERPs evoked by the attended and ignored stimuli then indicates a specific effect of selective attention, independent of nonspecific changes in arousal. Despite some contrary reports, the bulk of the evidence indicates that the early auditory ERPs are unaffected by whether the

participant pays attention to the stimuli or ignores them (Hackley *et al*., 1990). This suggests that the brain automatically processes auditory information to the level of the auditory cortex. The earliest undisputed evidence of an attentional effect on the auditory ERPs is a change in the middle latency potentials between 20 ms and 50 ms after the stimulus (Woldorff and Hillyard, 1991). This change occurs in demanding selective attention tasks and probably reflects some facilitated processing of the attended information in the auditory cortex. (*See* **Attention**)

The most striking effect of auditory attention is an increased negativity beginning at about 50 ms after the onset of a sound and overlapping the N1 wave. This effect can also be demonstrated as a negative difference wave or Nd, obtained by subtracting the response to ignored stimuli from the response to the same stimuli when they are attended. The effect probably indicates both an enhancement of the processing normally represented by the N1 wave and extra processing independent of the N1. The size and duration of the Nd varies with the amount of processing needed for the attentional task and the time pressure under which it occurs. Two parts of the Nd wave have been distinguished: an early wave, which is probably related to processing attended information, and a later wave, which may be related to task monitoring. The early Nd wave is mainly generated in the auditory cortices of the supratemporal plane. The neurons contributing to this wave probably vary with the different attentional tasks, e.g. whether the person is attending to the loudness of the stimulus or its frequency. Differences in Nd between tasks are difficult to demonstrate because neurons specific to different tasks have similar locations on the supratemporal plane. Nevertheless, the Nd for attended stimuli clearly varies with the location of the stimulus in space (Teder-Sälejärvi *et al*., 1999).

WORKING MEMORY

Once the information in an attended stimulus is processed, it is available to working memory for further evaluation. The ERP evoked by the detection of an improbable auditory target (or 'oddball') in an attended train of standard stimuli contains, in addition to the P1–N1–P2 waves, an N2–P3 complex and a later 'slow wave' (Figure 3). The P3, the largest of these attention-dependent waves, typically occurs with a peak latency of about 300 ms, and therefore also goes by the name of P300. It is maximally recorded from the vertex and midparietal regions of the scalp. The amplitude of the wave is

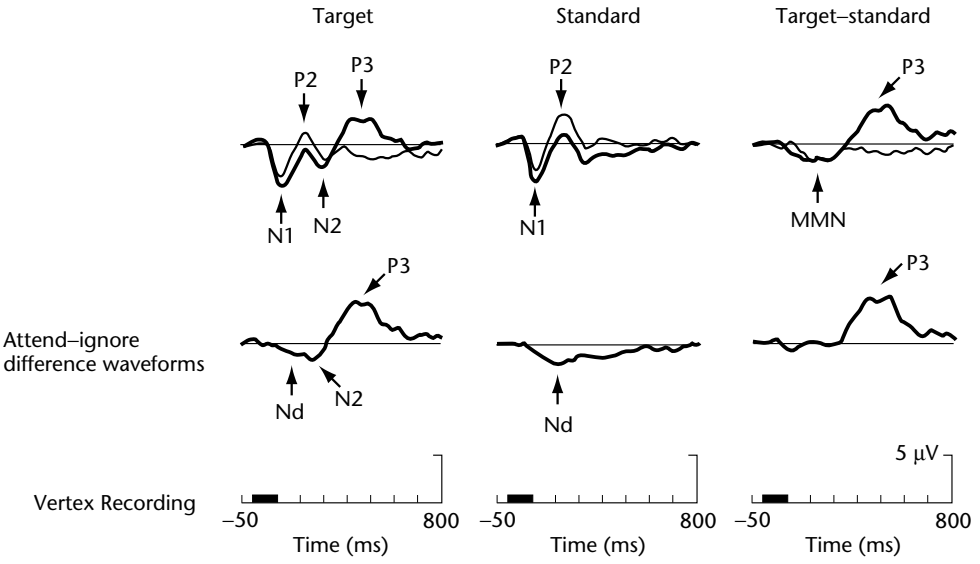


Figure 3. Effects of attention on the auditory event-related potentials (ERPs) to tones presented 65 dB above threshold at a rate of 2 s^{-1}. The participant's task was to detect occasional ($p = 0.2$) targets with a slightly different frequency or to ignore the stimuli and attend to a concurrent set of stimuli in the other ear. The ERPs for ten participants were averaged across the ears according to whether the stimuli were targets versus standards, and attended (thick line) versus ignored (thin line). The attend–ignore difference waveforms show a Nd wave for both targets and standards, and a P3 wave for the targets. The target–standard waveforms show a mismatch negativity (MMN) for the target regardless of attention and a P3 wave during attention.

inversely related to the probability of the stimulus. Its peak latency increases and its amplitude decreases with increasing difficulty in distinguishing the target from the standard. The peak of the P3 may occur before a motor response in difficult tasks, but often occurs at the same latency or later than the motor response in easy tasks. It therefore probably represents cerebral activity that is unrelated to the motor response. In general, the latency is more closely related to the duration of sensory processing and is relatively unaffected by manipulations that alter the subsequent selection of perceptual or motor responses. Several hypotheses have been proposed for the function of the P3 wave, including the updating of working memory, the access of information to conscious processing, and the resetting of perceptual analyzers once their processing has finished. (*See* **Working Memory**; **Event-related Potentials and Mental Chronometry**)

Intracerebral recordings and blood flow studies indicate that many different regions of the brain are active during the P3, most prominently regions of the hippocampus and the parietal lobe (Linden *et al*., 1999). All these regions probably contribute to the scalp-recorded waveform during the P3 and the subsequent slow wave. A network of intracerebral events overlapping in time and in potential occur in relation to recognizing the target, associating it with the required response, initiating the response, and updating memory about the occurrence of both target and response. Working memory is probably manifest in these interactive connections.

CONCLUSION

The auditory ERPs can time the different activities that occur as auditory information is processed in the human brain. Studies of the increased blood flow that occurs with these activations can indicate the locations of the processing. Studies of animals can indicate the underlying neuronal mechanisms. Combining information from all these approaches should elucidate the 'when', 'where' and 'how' of human hearing.

References

Alain C and Woods DL (1997) Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology* **34**: 534–546.

Galambos R, Makeig S and Talmachoff PJ (1981) A 40 Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of Sciences USA* **78**: 2643–2647.

Hackley SA, Woldorff M and Hillyard SA (1990) Cross-modal selective attention effects on retinal, myogenic, brainstem and cerebral evoked potentials. *Psychophysiology* **27**: 195–208.

Linden DEJ, Prvulovic D, Formisano E *et al*. (1999) The functional neuroanatomy of target detection: an fMRI study of visual and auditory oddball tasks. *Cerebral Cortex* **9**: 815–823.

Näätänen R (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent MMNm. *Psychophysiology* **38**: 1–21.

Picton TW, Lins O and Scherg M (1995) The recording and analysis of event-related potentials. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 10, sect. 14, Johnson R (ed.) Event-Related Brain Potentials and Cognition, pp. 3–73. Amsterdam: Elsevier.

Picton TW, Alain C, Woods DL *et al*. (1999) Intracerebral sources of human auditory evoked potentials. *Audiology and Neurootology* **4**: 64–79.

Picton TW, Alain C, Otten L, Ritter W and Achim A (2000) Mismatch negativity: different water in the same river. *Audiology and Neurootology* **5**: 111–139.

Singer W (2000) Response synchronization: a universal coding strategy for the definition of relations. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 325–338. Cambridge, MA: MIT Press.

Teder-Sälejärvi WA, Hillyard SA, Röder B and Neville HJ (1999) Spatial attention to central and peripheral auditory stimuli as indexed by event-related potentials. *Cognitive Brain Research* **8**: 213–227.

Woldorff MG and Hillyard SA (1991) Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology* **79**: 170–191.

Woldorff MG, Hillyard SA, Gallen CC, Hampson SR and Bloom FE (1998) Magnetoencephalographic recordings demonstrate attentional modulation of mismatch-related neural activity in human auditory cortex. *Psychophysiology* **35**: 283–292.

Further Reading

Hillyard SA, Mangun GR, Woldorff MG and Luck SJ (1995) Neural systems mediating selective attention. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 665–681. Cambridge, MA: MIT Press.

Näätänen R (1992) *Attention and Brain Function*. Hillsdale, NJ: Lawrence Erlbaum.

Näätänen R and Picton TW (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* **24**: 375–425.

Picton TW (1990) Auditory evoked potentials. In: Daly DD and Pedley TA (eds) *Current Practice of Clinical Electroencephalography*, 2nd edn, pp. 625–678. New York, NY: Raven Press.

Picton TW (1992) The P300 wave of the human event-related potential. *Journal of Clinical Neurophysiology* **9**: 456–479.

Starr A and Don M (1988) Brain potentials evoked by acoustic stimuli. In: Picton TW (ed.) *Handbook of Electroencephalography and Clinical Neurophysiology*, vol. 3, Human Event-Related Potentials, pp. 97–157. Amsterdam: Elsevier.

Woods DL (1990) The physiological basis of selective attention: implications of event-related potential studies. In: Rohrbaugh JW, Parasuraman R and Johnson R (eds) *Event-related Brain Potentials: Basic Issues and Applications*, pp. 178–209. New York, NY: Oxford University Press.

Woods DL (1995) The component structure of the N1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology* supplement **44**: 102–109.

Auditory Perception, Psychology of

Introductory article

Stephen Handel, University of Tennessee, Knoxville, Tennessee, USA
Mark Hedrick, University of Tennessee, Knoxville, Tennessee, USA

CONTENTS

The production of all sounds creates regularities in the air-pressure sound wave that reaches the listener. The physiological mechanisms and cognitive processes involved in locating and identifying the source of those sounds take advantage of these regularities.

INTRODUCTION

We can identify our friends, rainfall, doorbells, and screeching chalk on a blackboard by sound alone. The key to understanding the diversity of auditory perceptions is to comprehend the complementary nature of the production and subsequent perception of those sounds. The production of all sounds creates regularities in the physical air-pressure wave that reaches the listener. The physiological mechanisms and cognitive processes necessary to locate and identify the source of those sounds take advantage of these regularities.

The first step is to describe the regularities in the sound wave, namely 'what is out there to perceive'. The second step is to describe the physiological adaptations of the auditory system that transform the amplitude of the air-pressure wave into the neural signal which is transmitted to the cortex. The third step is to describe the cognitive heuristics (where a *heuristic* is a rule that usually works but which can lead to a wrong outcome in some instances) that are used to make sense of the neural signal in order to perceive what is happening in the external world.

THE PRODUCTION OF SOUND

Source-filter Model

The basic notion is that a 'source' (e.g., a violin string, or puffs of air coming through the vocal cords) is excited by energy. The source then imposes its vibration pattern on the filter (e.g., the wooden sound body of the violin, or the shape of the mouth), which modifies the vibration of the source into the pressure wave that is radiated to the environment.

The source can vibrate at several component frequencies which are determined by its physical construction, material, and method of excitation. The frequencies of the vibrations usually form a harmonic series (the frequencies of the components are integer multiples of the lowest frequency: 1:2:3:4:5…). The amplitudes of each component vary depending on the above characteristics and, within limits, component vibration starts and ends at the same time. The source vibrations then stimulate the filter. The latter also has specific vibration frequencies (resonances) and is thus capable of being excited when the source vibration frequencies match the filter's vibration frequencies. Owing to the match and mismatch of frequencies, certain source vibrations are transmitted if the source and filter frequencies match, and others are 'stilled' if the frequencies do not match.

Thus the vibrations in the sound pressure wave coming from one source are typically harmonically related (albeit at different amplitudes), start at the same time, evolve slowly across their duration due to frictional decay or changes such as vibrato, and finally end at approximately the same time. All of these regularities can be used to detect and identify the source of the sound.

Transparency of Sound Waves

If only one sound occurred at any particular time, then the problem of auditory perception would be easy. If each unique sound wave represented one

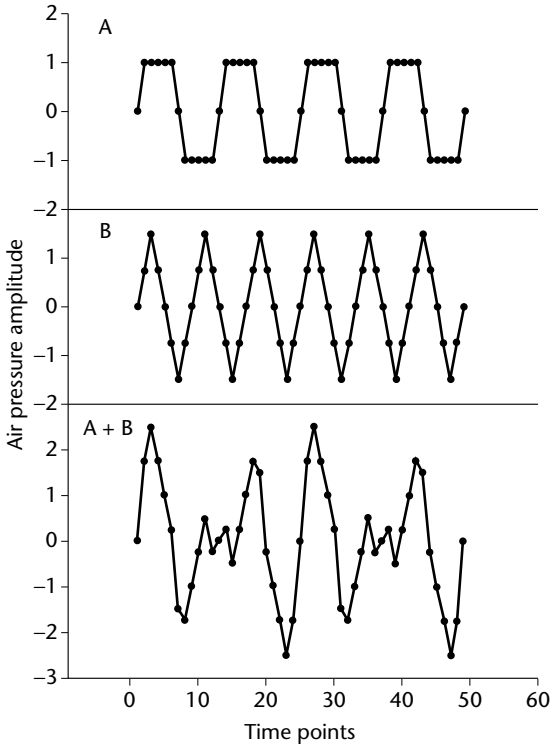


Figure 1. The addition of two waves gives rise to a combination wave that is the sum of the two individual waves. The 'square wave' in A takes 8 time periods to go through one cycle, repeating at 9, 17, 25, and so on. The 'triangle wave' in B takes 12 time periods to go through one cycle, repeating at 13, 25, and so on. The sum of A and B takes 24 time periods to go through one cycle, repeating at 25, 49, and so on.

source, it would merely be a problem of memorization. Unfortunately, the sound waves coming from two or more overlapping sources (e.g., a radio, street noise, and someone talking) are added together to give one wave in which the individual sound waves from each source are lost in the composite (Figure 1). The frequency components from each sound source are completely intermingled. The unique problem for auditory perception is to untangle this mixture in order to recover the individual waves created by each separate source. We hear each source easily without conscious calculation, but this effortlessness belies the difficulty involved. No computer can do this task. In contrast, nearly all visual objects are opaque rather than transparent, so that the light ray from each point in space invariably comes from light reflecting off just one object.

The goal of the auditory system must be to analyze the composite sound pressure wave into its frequency components, and to assign sets of components to different sound sources. The heuristics used to make this assignment are based on the regularities inherent in the production of sound, namely that the frequency components from one sound source tend to be harmonically related, start and end at the same time, change in quality slowly and continuously, and occur at one location in space. Thus we would expect that the physiological mechanisms would be designed to analyze the sound wave into the amplitude pattern of its component frequencies, to maintain the onset and offset timing of the components, and to maintain the location of the components. Essentially, this is what the auditory pathways accomplish.

PHYSIOLOGICAL STRUCTURES

The Ear

The fundamental problem is to convert the changes in air pressure into neural impulses that travel to the auditory cortex. The physical construction of the ear is designed to overcome the large mismatch in physical properties between the air medium of the sound and the fluid-filled medium in the vertebrate inner ear.

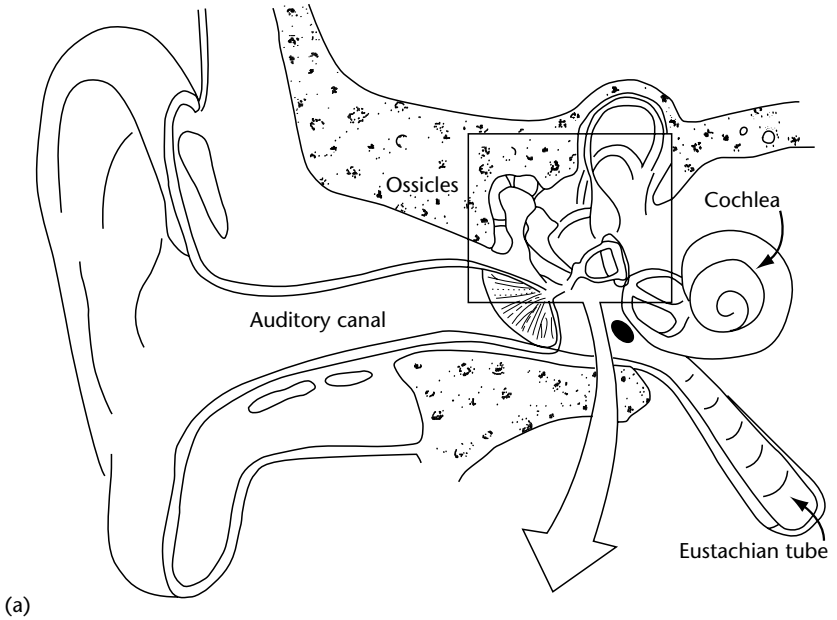
The outer ear consists of the visible pinna and the hollow ear canal terminating at the ear-drum that captures sound energy at frequencies mainly used for speech (Figure 2). The middle ear consists of three tiny bones that provide a bridge between the vibration in air at the ear-drum and the fluid vibration at the base of the inner ear. Together, the motion of the bones and the relative size of the ear-drum compared with the oval window at the base of the inner ear convert most of the air pressure vibration into fluid vibration. The inner ear consists of the cochlea, which is responsible for transforming the vibration pattern into neural firing (Figure 3). The cochlea is a coiled fluid-filled tube, about the size of a small bean, with a complex membrane that divides the tube into two halves. Vibration of the middle ear bones creates fluid pressure waves that travel down the cochlea, causing the center membrane to move, and that in turn distorts the neural cells (known as hair cells) attached to the membrane. This distortion causes the hair cells to stimulate neural firing, which eventually reaches the auditory cortex.

Frequency Analysis

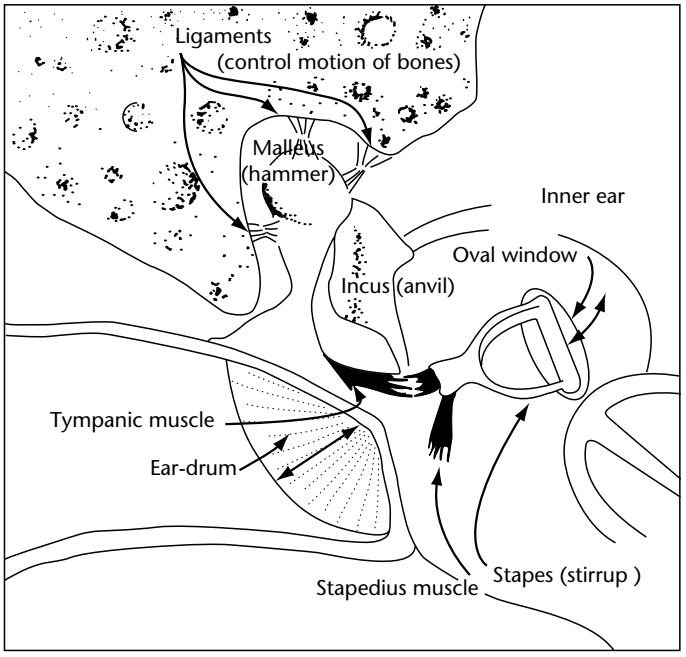
The goal of the auditory system is to isolate the frequency components and to maintain the onset

and offset timing between the components. This is achieved in two complementary ways in the inner ear.

First, the membrane changes its shape and stiffness along its length so that it distorts maximally to higher frequencies of vibration at the base, close to the vibrating bones, and it distorts maximally to lower frequencies at the apex, where the membrane ends. For a complex vibration composed of many frequencies, the membrane would distort at several



(a)



(b)

Figure 2. (a) A schematic view of the outer, middle, and inner ear. (b) A detailed view of the middle ear. The middle ear consists of three connected bones, namely the malleus, incus, and stapes. The stapedius and tympanic muscles control the acoustic reflex that protects the ear from loud prolonged sounds by loosening the connections between the middle ear bones. Reproduced from Handel (1989, p. 466) with permission.

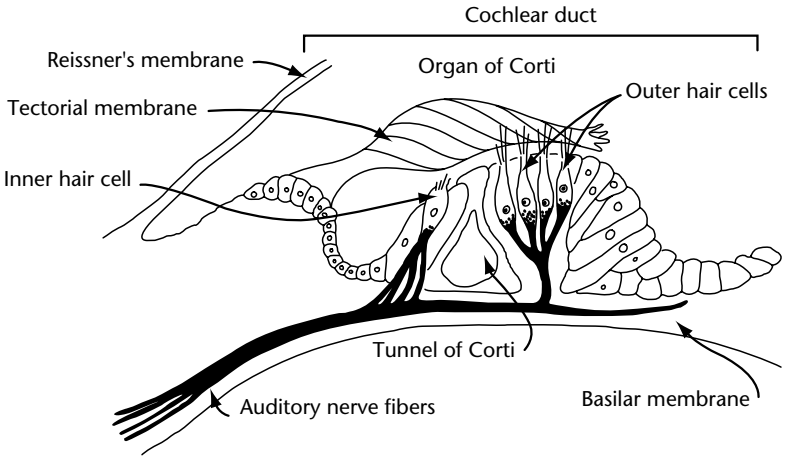


Figure 3. View of the inner ear showing the center segment bounded by Reissner's membrane on the top and the basilar membrane on the bottom. The firing of the inner hair cells creates the neural signal, while the outer hair cells act to change the mechanical properties of the basilar membrane. Reproduced from Handel (1989, p. 471) with permission.

points, each representing one component frequency. This has been termed *place coding*, and a larger amount of distortion will result in a higher rate of neural firing.

Secondly, the membrane essentially makes one 'up-and-down' movement for one cycle of the vibration. For most of the frequencies used for speech and music, the hair cells track that motion and cause the attached neurons to fire once per cycle. This firing pattern has been termed *frequency coding*, and the timing between firings can be used to code frequency. If the hair cells fire at the same point in the movement, this is termed *phase-locking*.

Even though the inner ear seems to be physically rather crude, its frequency, intensity and temporal resolution is remarkably good. In terms of the ability to detect differences, the resolution is about 0.5% for frequency (a difference of between 1000 Hz and 1005 Hz), about 10% for intensity, and about 2 ms for onset. These values are more than sufficient to identify sound objects.

As the neural signal travels to the cortex, it passes through many brain nuclei, where there is much neural elaboration. Although there are only about 2000 hair cells per ear, there are about 1 000 000 cells in each side of the auditory cortex.

PERCEPTION OF LOCALIZATION

Although each ear analyzes the pressure wave only in terms of frequency, the auditory image is one of objects that appear to be fixed in space outside the head, and which do not appear to move despite head and body movements. It is the human body that generates the physical cues to object localization. If we were simply points in space, there would be no way of localizing sounds, and if sound is presented directly to the ears using headphones, objects appear inside the head and not out in space.

The perception of direction depends on two acoustic cues that are produced by the difference in position of the two ears. First, with the exception of sounds exactly in front of or exactly behind the listener, all sounds will reach the near ear before they reach the far one. The time differences are quite small, reaching a maximum of about 0.5 ms if the sound is directly opposite one ear. Secondly, the head casts a sound shadow so that the intensity of the sound reaching the far ear will be less than that of the sound reaching the near ear (Figure 4).

Both time and intensity differences are used to assess direction. Time differences appear to be more important for frequencies lower than 1500 Hz, whereas intensity differences appear to be more important for higher frequencies. The head reflects all frequencies, but at the lower frequencies the sound wave can 'bend' around the head, reducing the intensity difference. At frequencies higher than 2500 Hz the sound waves still bend around the head, but they converge at a distance behind it. Therefore there is a strong intensity shadow that can be used to detect the direction of the sound. There are always ambiguities. For example, a source 45° in front of one ear creates the same time and intensity cues as a source 45° behind the same ear. These types of ambiguities can be resolved by simple head motions.

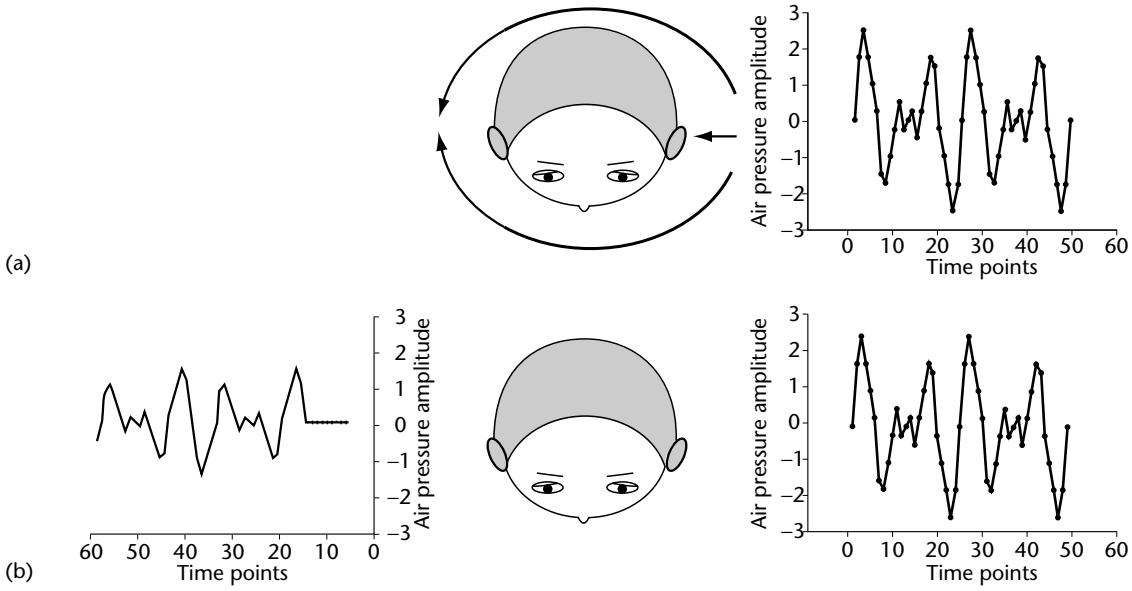


Figure 4. Sound localization. (a) A sound pressure wave is directly opposite the left ear. The wave travels directly to the left ear, and circles the head to travel to the right ear. (b) The wave to the left ear is unchanged, but the wave arriving at the right ear is delayed in time and reduced in amplitude. Note that the right ear wave is reversed in time.

MAKING SENSE OF THE SOUND WAVE: WHAT ARE THE OBJECTS IN THE WORLD?

To hear the singer of a jazz combo, the listener must group together the frequency components produced by the singer separately from those produced by the instruments. Surprisingly, the most important acoustic cue leading to the grouping of the components is synchrony of the onsets. Remember that the auditory system has very fine temporal acuity – it can distinguish differences of less than 2 ms. Several experiments have placed the various cues in conflict and have found that listeners will group frequency components that begin at the same time, even if the components are not harmonically related or they are presented to different ears. If the components start at slightly different time points, listeners do not report that they hear two sounds starting at different time points. Instead, they simply report hearing two sounds, and they are unable to report the order of those sounds. Thus, curiously, short time asynchronies are converted into source information and not order information.

The second most important cue is that of harmonic relatedness. If all of the frequency components are harmonically related, the obvious decision would be that all of the components come from a single source, and the percept is that of one complex tone. If one of the components is mistuned so that its harmonic relationship is lost (e.g., 200, 415, 600, 800 Hz), the percept is divided – the mistuned component is heard as one emergent tone, and the remaining components are heard as a second tone that has changed in sound quality because the mistuned harmonic has been segregated out. If the frequency components are inharmonic (e.g., 100, 125, 200, 250, 300, 375 Hz…), the auditory system partitions the components so that each set is harmonically related – one complex tone based on a fundamental frequency of 100 Hz and a second one based on a fundamental of 125 Hz. For speech, fundamental frequency differences as small as 2% are sufficient to lead to the perception of two vowels or syllables, each spoken by a different voice.

The rank order of synchrony, harmonicity, and location reflects the predictability of the cues. It is extremely unlikely that two different sounds will start at exactly the same time. However, there are sounds in which the components are not harmonic (e.g., any type of static or noise) and, owing to the possibility of multiple reflections off hard smooth surfaces, the timing and intensity cues to direction may be seriously distorted.

The final step is to categorize, identify, and label each set of frequency components. Although this process is not understood, it is clear that the fundamental frequency and the overall amplitude pattern of the frequency components (its *shape*) as

well as short-term variations influence identification. For example, male voices are distinguished from female ones on the basis of fundamental frequency, but female (and male) voices are often distinguished within gender on the basis of the pattern of frequency components.

CONCLUSION

Although there are striking differences in the stimulus energies and physiology of the auditory and visual perceptual systems, both are concerned with the description of objects in the world. The auditory system is maximally tuned to the timing and rhythm of events (after all, we don't dance to lights alone), and the cognitive rules that are used to interpret the neural firings make use of those regularities.

Further Reading

Bregman AS (1990) *Auditory Scene Analysis: the Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Handel S (1989) *Listening: an Introduction to the Perception of Auditory Events*. Cambridge, MA: MIT Press.

Hartmann WH (1996) Pitch, periodicity and auditory organization. *Journal of the Acoustical Society of America* **100**: 3471–3502.

McAdams S and Bigand E (eds) (1993) *Thinking in Sound*. Oxford, UK: Clarendon Press.

Moore BCJ (1997) *An Introduction to the Psychology of Hearing*, 3rd edn. London, UK: Academic Press.

Taylor C (1976) *Sounds of Music*. New York, NY: Scribner's Son.

Warren RM (1999) *Auditory Perception: a New Analysis and Synthesis*. Cambridge, UK: Cambridge University Press.

Yost WA (2000) *Fundamentals of Hearing*, 4th edn. San Diego, CA: Academic Press.

Autism, Psychology of

Introductory article

Christopher Jarrold, University of Bristol, Bristol, UK
Francesca Happé, Institute of Psychiatry, London, UK

CONTENTS

Autism is a disorder that has a biological cause but is diagnosed on the basis of problems in socialization, communication, and imagination. Psychological explanations of the condition aim to account for this pattern of behavioral difficulties, and for the strengths that individuals with autism show in other areas.

BACKGROUND

Autism is a disorder that is most often diagnosed in childhood. It affects five to 10 in every 10,000 individuals, and is about four times more common among males than females. It is clear that autism has a biological origin, rather than being the result of psychogenic factors, but as yet the precise cause or causes of autism are unknown. One thing that is clearer is that the condition has a genetic component. This does not mean that it only occurs among children of people who themselves have autism. However, if a mother has one child with autism then the chances of a subsequent child having autism are increased approximately 50-fold – although of course this is still relatively unlikely.

The absence of a clearly specified biological cause of autism means that the condition is currently diagnosed on the basis of individuals' behavior. The term 'autism', which comes from the Greek word *autos* for 'self', was first used to refer to the condition by an American clinician, Leo Kanner. Writing in 1943, Kanner identified the key feature of 'autistic aloneness' among his patients, describing their apparent reluctance to engage in social interaction with other people. At approximately the same time, and independently of Kanner, an Austrian physician, Hans Asperger, identified 'autistic psychopathy' among a sample of children showing similar patterns of social peculiarity.

Difficulties in socialization are now seen as one of the key behavioral features of autism, along with problems of communication and imagination, and this 'triad' of impairments forms the basis of most current diagnostic schemes. Problems in *socialization* are seen in individuals' reluctance or inability to engage in social interactions with others. Some individuals with autism do attempt to talk and interact with other children and adults, but often do so clumsily, without a real awareness of the social nuances that underpin our normal everyday contact with others. Problems of *communication* often take the form of severe language delay. Those individuals with autism who have good language skills still have difficulties in understanding subtle aspects of speech, such as the use of irony, metaphor, or jokes, and as a result can often appear rude or blunt when they speak to people. Finally, problems of *imagination* are reflected in the inflexibility and repetitive nature of thoughts and actions. Younger individuals with autism tend not to engage in pretend play, and may show repetitive behaviors such as hand-flapping or rocking. Older individuals tend to be inflexible and like to stick to precise routines. Individuals with autism often develop intense, narrow interests, usually based around predictable or repetitive themes, such as an interest in bus timetables.

Two other aspects of the condition are also worth noting. First, the majority of individuals with autism suffer from a degree of intellectual handicap, which in some cases can severely affect day-to-day functioning. However, it is certainly possible to have autism and to be of normal, or above average, intelligence. The children identified by Hans Asperger had a higher level of intelligence than those seen by Leo Kanner, and today the term 'Asperger syndrome' refers to a condition on the autism spectrum, but which is not associated with mental disability. However, it is not entirely clear whether these two conditions are really distinguishable, and the characteristics which might differentiate them (lack of language

or cognitive delay) are a subject of current debate. A second point to emphasize is that autism is characterized by strengths as well as weaknesses. In addition to their problems, individuals with autism perform well in certain (nonsocial) areas, described below.

THEORY OF MIND

Arguably, the most influential psychological explanation of autism at present is the 'theory of mind deficit' hypothesis. The term 'theory of mind' refers to our normal ability to predict and explain what others are doing on the basis of their beliefs and desires. For example, if we see someone leave their house, walk down the road, stop, check the pockets of their coat, turn around, and re-enter their house, then we assume that they have realized that they have forgotten something which they believe to be still inside the house. On the basis of people's behavior we infer what they are thinking, and equally, if we know what people are thinking we can predict what they will do.

Psychologists often assess individuals' theory of mind by asking them to predict what someone will do on the basis of a 'false belief'. For example, if someone doesn't see that an object is moved from one location to another, they will incorrectly believe that it is still in the original location. Typically developing children younger than around four years of age fail to appreciate that someone in this situation will have a false belief, and instead predict that they will look for the object where the child themselves knows it to be. Around the age of four children come to realize that others can have beliefs which differ from their own – an important marker of theory of mind.

Many studies have shown that individuals with autism have severe difficulties on this kind of task, even when they are much older than four years of age or have intellectual abilities well above the four-year-old level. The majority of individuals fail to appreciate that others can have false beliefs, and those that do pass this kind of test show more subtle problems in 'reading other people's minds'. This has led to the claim that autism is associated with a failure to acquire a theory of mind, or at least, is linked to severely delayed acquisition of this ability.

A major strength of this hypothesis is its potential to explain the triad of symptoms seen in autism. Problems of socialization are the result of a failure to appreciate what motivates social interaction, as most of our dealings with others rest on an appreciation of what people know or don't know, and what they want or don't want. Similarly, communication requires an understanding that others intend to communicate ideas to us, that others' beliefs can be changed by what you tell them, and that others may or may not know what you know. Some of the problems in imagination may arise from an inability to make sense of the mental state of 'pretence'. It has also been suggested that repetitive behaviors may reflect an attempt to impose order on what is, to individuals with autism, a confusing and unpredictable social world. However, there is no direct evidence to support this suggestion, and while this may explain a preference for routines among individuals with autism it is not clear how a 'theory of mind' problem would lead to more basic repetitive behaviors such as hand-flapping.

EXECUTIVE FUNCTION

The theory of mind hypothesis can therefore account for many of the features of autism, particularly problems in socialization and communication, but the explanation it offers for problems of imaginative flexibility is less strong. Instead, lack of imagination and inflexibility appear more likely to be related to problems of 'executive functioning' which are also thought to be associated with the condition. The psychological notion of executive function concerns our ability to control our actions. In the same way that a chief executive in a company guides the 'behavior' of the company by instigating procedures, stopping other actions, monitoring progress, and planning for the future, so we need executive control to allow us to consciously control our actions, and to prevent us from making inappropriate automatic responses. This ability appears to be linked to the frontal regions of the brain, as individuals who have suffered damage to these regions become impulsive and show inappropriate repetitive behaviors. Individuals with autism often show similar problems, and make impulsive and inappropriate responses on the kinds of tasks used to assess executive control. There is also some evidence of frontal lobe abnormalities in individuals with autism (see below).

SAVANT ABILITIES AND CENTRAL COHERENCE

The theory of mind account and executive function theory struggle to explain why individuals with autism show strengths in certain areas, as noted above. One example of this is the 'savant abilities' seen in a small minority of individuals with autism.

This term refers to skills that are far superior to what would be expected given an individual's general level of intellectual ability. For example, some individuals with autism can tell you the day of the week that corresponds to any given date – for example, the 3rd of October 1907 was a Thursday – even though they might find simple maths problems difficult. Other individuals with autism have an incredible memory for music and can play a piece note-perfect after hearing it only once. Others have remarkably precise drawing abilities, and can recreate complex scenes accurately from memory.

Savant abilities are seen in only a few individuals with autism, but most, if not all, individuals show strengths in visual and spatial areas, rote memory, or attention to detail. Many people with autism are good at doing jigsaw puzzles, and individuals with autism perform relatively well on psychological tests that require them to look for details in a complicated visual image. These strengths are thought to relate to a particular bias among individuals with autism. When ordinary people are presented with a visual image or a story, they typically remember the overall pattern of the picture, or the gist of the story, at the expense of the details. In other words they focus on the 'whole' and not the 'parts'. This tendency has been termed a 'drive for central coherence'. In contrast, individuals with autism focus on the parts of a stimulus and not the whole, and are said to have 'weak central coherence'. They therefore find it difficult to extract the overall meaning from complex information, but are very good at perceiving and remembering the details of this information.

This bias towards parts rather than wholes may explain some of the savant skills described above. In some of these cases it may be that perceiving information in terms of its constituent parts is the best way to remember it accurately. For example, if an individual perceives and remembers a visual image in terms of every individual line that makes up the picture, rather than by remembering the overall pattern of the image, then they may be able to recreate it in great detail.

NEUROLOGICAL UNDERPINNINGS

The exact brain basis of autism is as yet unclear, despite several decades of investigation. Many brain regions have been proposed as the site of abnormality, and studies in this area have produced somewhat contradictory findings, perhaps due to differences of participants, comparison groups, and techniques employed. As a result, there is as yet little agreement as to what differences might be specific and universal to the brains of people with autism. Current suggestions for the site of key abnormalities include the cerebellum and parietal cortex, which have been linked to attentional control; the amygdala and limbic system, important in processing of emotional information; and the prefrontal cortex, responsible for higher-order control functions such as planning and flexibility. There is also a suggestion of greater cell density and larger, heavier brains in autism, perhaps reflecting a failure of synaptic pruning, thought to be an important part of normal brain maturation. Future progress in this area may be made using functional brain imaging techniques.

Studies of the brain basis of normal theory of mind have begun to suggest specific regions for further investigation in autism. Specifically, medial and orbital regions of the frontal lobes, the amygdala, and temporo-parietal regions have all been shown to be more active during theory of mind tasks than during control tests. Recent research suggests that some of these regions (specifically the amygdala, and the medial-frontal region) are not activated when people with autism attempt theory of mind tasks. Future brain scanning studies with children with autism may clarify the role of these and other brain regions in the development of the psychological characteristics of the condition. At present, however, biological or genetic therapy appears a long way off, and the most effective current interventions are educational and behavioral.

Further Reading

Bailey AJ (1993) The biology of autism. *Psychological Medicine* **23**: 7–11.

Bailey A, Phillips W and Rutter M (1996) Autism: towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *Journal of Child Psychology and Psychiatry* **37**: 89–126.

Baron-Cohen S (1992) The theory of mind hypothesis of autism: history and prospects of the idea. *The Psychologist* **5**: 9–12.

Baron-Cohen S (1995) *Mindblindness*. Cambridge, MA: Bradford Books/MIT Press.

Frith U (1989) *Autism: Explaining the Enigma*. Oxford, UK: Basil Blackwell.

Happé F (1994) *Autism: An Introduction to Psychological Theory*. London: UCL Press.

Happé FGE (1994) Annotation: current psychological theories of autism: the 'theory of mind' account and rival theories. *Journal of Child Psychology and Psychiatry* **35**: 215–229.

Happé F (1999) Autism: cognitive deficit or cognitive style? *Trends in Cognitive Sciences* **3**: 216–222.

Jarrold C, Butler DW, Cottington EM and Jimenez F (2000) Linking theory of mind and central coherence bias in autism and in the general population. *Developmental Psychology* **36**: 126–138.

O'Connor N and Hermelin B (1988) Low intelligence and special abilities. *Journal of Child Psychology and Psychiatry* **29**: 391–396.

Ozonoff S (1995) *Executive functions in autism*. In: Schopler E and Mesibov GB (eds) *Learning and Cognition in Autism*, pp. 199–219. New York, NY: Plenum Press.

Russell J (1997) *Autism as an Executive Disorder*. Oxford, UK: Oxford University Press.

Autism

Introductory article

Karen Pierce, University of California, San Diego, California, USA
Eric Courchesne, University of California, San Diego, California, USA

CONTENTS

Autism is a developmental disorder, five times more common in males, with clinical onset during the first years of life. It is characterized by abnormalities in social behavior, language, and cognition, and is now known to be biological rather than psychogenic in origin.

INTRODUCTION

Autism is a pervasive developmental disorder, affecting males five times as often as females, with clinical onset during the first years of life. The prevalence is approximately 1 out of every 600 live births. This biological disorder is characterized by abnormalities in social behavior, language, cognition, and environmental interests that persist throughout the affected individual's lifetime. Although symptoms are probably mediated by diverse brain defects, no physical abnormalities are apparent in individuals with this disorder.

BEHAVIORAL SYMPTOMS OF AUTISM

Social Behavior

One of the first indicators of social abnormality in autism is a deficit in the ability to engage in joint social attention, which in normal infants is the merging of attention between two people and an object or activity of interest. Joint social attention is normally achieved by 14 months of age, and is an important precursor to both social and language development. The absence of such a skill in autism has thus been implicated as fundamental to the cascade of developmental problems that ensue. As the autistic toddler matures, other social abnormalities become apparent, such as low rates of eye contact and reciprocal social interaction, and difficulties in identifying and interpreting the emotions of others. Social abnormalities in autism have often been referred to as the hallmark of the disorder, probably because such abnormalities are not only severe, but also obvious in affected individuals.

Language

Approximately 50% of all individuals with autism fail to develop functional speech; in those who do, language is characterized by several abnormalities including pronoun reversals (saying 'he went to the market' instead of 'I went to the market'), use of neologisms (nonsensical or made-up words), stereotyped or rigid speech and abnormalities in intonation. The speech of autistic individuals is also characterized by echolalia, which is the repetition of words either immediately after someone has spoken them or after a delay of hours, days, or even months. For example, an autistic individual may repeat the phrase 'How old are you?' hundreds of times in a single day, after hearing the phrase only once.

Cognition

Seventy-five per cent of autistic individuals also suffer from mental retardation, ranging from mild to severe. Since social interactions and language ability are important for learning about the world and developing cognitive skills, it has been difficult to establish whether impaired mental development in the autistic child occurs independently of the main symptoms of autism (social and language impairment) or occurs in part because of the early failure of normal social and language skills. Although it is difficult to disentangle secondary effects of early social and language impairment from mental retardation *per se*, some scientists believe that deficits in higher-order memory abilities, conceptual reasoning (e.g., categorization skills), executive function (e.g., switching between two or

more mental sets) and auditory information processing may exist as important features of this disorder.

One robust finding, however, has been in the domain of attention. Behavioral as well as functional neuroimaging studies have consistently demonstrated dysfunction in three primary attention abilities: disengaging attention from one source of information, orienting attention to a new source, and shifting attention back and forth between two separate sources of stimulus information. Autistic individuals are slow and inefficient in each of these abilities. Also, they are unable to properly adjust their 'spotlight of attention' so that they may have an excessively narrow focus of attention on visual details or an abnormally broad focus. Understanding attention deficits in autism is important because the ability to attend is required for an infant to follow the rapid and unpredictable ebb and flow of human social activity, such as words, gestures, and facial expressions. Dysfunctions in attention thus significantly interfere with the general ability to learn, as well as being likely to amplify other areas of difficulties for autistic individuals, such as language and social behavior.

Restricted and Repetitive Interests

Individuals with autism commonly display restricted, repetitive and stereotyped patterns of interests and activities. This general category of behavior manifests itself in many ways, such as an inflexible adherence to specific routines, stereotyped and repetitive motor mannerisms such as hand-flapping (Figure 1) or a preoccupation with an object or part of an object. In general, autistic individuals display a limited interest in their environment, instead focusing their attention on one specific aspect or obsessive idea (e.g., amassing facts about cars). Further, individuals with autism may insist on sameness and show distress over trivial changes in their surroundings, such as movement of a piece of furniture. Such restricted and repetitive environmental interests are likely to interfere with learning and may have significant developmental implications for autistic children, because they may miss many learning opportunities that fall outside their scope of interest. Combined with attention deficits described above, the autistic child is ill-prepared to learn from the environment.

ETIOLOGY

Autism is among the most heritable of neuropsychiatric disorders. Twin studies report pairwise concordance rates as high as 90% for monozygotic (identical) twins but only 5–10% for dizygotic (fraternal) twins, suggesting the disorder is strongly genetic. Studies of the location of chromosomal abnormalities and break points can be extremely useful in the identification and mapping of genes that predispose an individual to disease. Although chromosomal abnormalities have been reported on many chromosomes in autism, the most consistent site is on chromosome 15. For example, reports indicate a duplication in a specific region on this chromosome (15q11–13) in approximately 2–4% of autism cases. This finding has prompted scientists to look closely at this chromosome for a particular candidate gene or genes that might contribute to autism.

Figure 1. A 7-year-old girl with autism. The left picture illustrates a commonly found repetitive motor behavior in children with this disorder, known as hand-flapping. Other odd hand mannerisms such as finger posturing (right) are also common.

Another common approach in genetic studies of autism is the 'genomic screen', where the whole genome in multiplex families (families with more than one person with autism) is screened in order to identify autism susceptibility loci. As with studies of chromosomal abnormalities, genomic screening studies have reported linkage to specific locations on many chromosomes; however, the strongest evidence seems to point to linkage to chromosome 7 and also chromosome 2. Although such linkages are encouraging, the genes related to autism have not yet been identified.

In addition to genetic inquiries, the search for the causes of autism has led some researchers to suggest a viral, toxic, or teratogenic etiology. Although certain viral or teratogenic agents (e.g., environmental toxins, anticonvulsant medicines)

are known to produce brain defects similar to some of those present in autism, research has not yet shown these agents to be significant in the etiology of most individuals with autism.

CLINICAL ONSET AND COURSE

There is almost complete scientific consensus that autism is a disorder with biological onset prenatally or shortly after birth, but clinical symptoms may not be recognized until late infancy or early childhood. Prompted by their child's failure to achieve normal developmental milestones in speech, language and social behavior, most parents seek professional help when their child is aged 2–3 years. Documentation of behavioral characteristics prior to this age is therefore sparse. There is, however, some evidence based on retrospective analyses (e.g., videotapes of children at their first birthday parties) of defects in attention patterns as well as in motor behaviors such as walking or crawling in infants and toddlers with autism. As the autistic child develops, the symptoms described above become pronounced, and many children enter a specialized treatment program in the home, clinic or school setting. Treatment may take the form of behavioral intervention (e.g., repeated practice of a target skill, such as pointing, followed by reward), occupational therapy (e.g., integrating fine and gross motor skills) or pharmacological intervention (e.g., drugs that facilitate serotonin transmission). Most children with the disorder live at home with their families and attend ordinary schools, although some children (usually those with severe symptoms, including mental retardation) may be placed in residential settings under the exclusive care of professionals. As adults, however, many with the disorder move into assisted living facilities.

NEURAL DEFECTS

Interestingly, when autism was first described by Leo Kanner in 1943, people formulated the idea that autism was a psychological disorder, caused by poor parenting or other environmental factors. By the 1970s, however, scientists began to investigate biological explanations for the bizarre behaviors noted in this disorder.

Among all types of biological abnormalities in autism, evidence for defects in the structure of the brain is the strongest. This evidence comes from two primary sources: magnetic resonance imaging (MRI) and autopsy studies. Magnetic resonance imaging is a relatively new technology that allows scientists to image the living brain. The size of specific brain regions in autism can be easily obtained from MRI scans and compared with the brains of nonautistic people. Autopsy cases of autism are rare, but provide essential information not only about size, but also about more microscopic details, such as the type, density or number of individual neuronal cells, that are not possible to obtain with MRI.

Studies show that in autism most major brain structures may be affected (Figure 2); these include the frontal lobes, parietal lobes, cerebellum, brainstem, corpus callosum, basal ganglia, amygdala, and hippocampus. Within the cerebellum and cerebrum, abnormalities have been found in gray matter (where neuronal cell bodies are located) and also in white matter (which contains axons

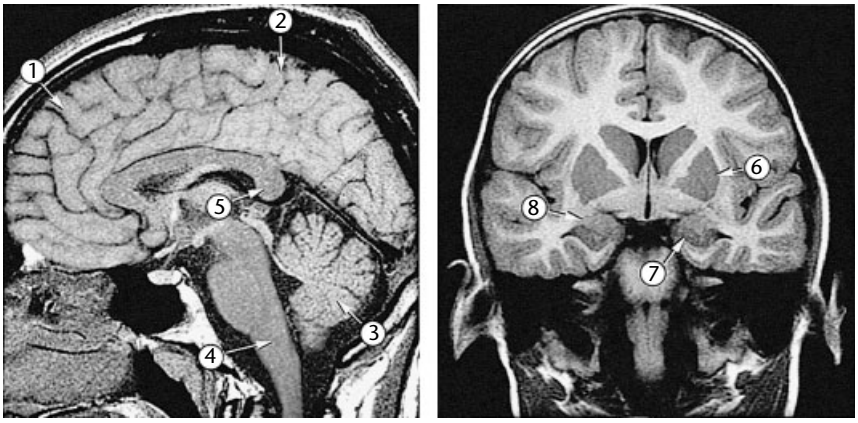


Figure 2. [*Figure is also reproduced in color section*.] Cortical and subcortical structures reported as abnormal in autism from autopsy or magnetic resonance imaging data from laboratories across the USA and Europe. 1, frontal lobes; 2, parietal lobes; 3, cerebellum; 4, brainstem; 5, corpus callosum; 6, basal ganglia; 7, hippocampus; 8, amygdala.

carrying signals from neurons in one location to those in another) in the youngest autistic children. These cerebellar and cerebral growth abnormalities are so commonly found in these patients and so extreme, that quantitative measures of them could potentially be used in clinical settings in conjunction with psychological information to assist in the diagnosis, prognosis and treatment recommendations at the youngest possible ages. Such widespread anatomic abnormality may explain why autism involves pervasive and persistent neurological and behavioral dysfunction. It is important to note, however, that not every autistic person has every neuroanatomic abnormality. For example, MRI studies suggest that approximately half of autistic adults have decreased volume of brain tissue in the parietal lobes, whereas the other half do not. One consistent neuroanatomic finding, however, has been that the cerebellum is abnormal in the majority of individuals with autism, both young and old. For example, in brain autopsy studies, 95% of all autism cases had reduced numbers of cerebellar Purkinje cells, a type of neuron crucial to cerebellar learning functions.

Certain defects in the autistic brain are signs of prenatal maldevelopment, which makes it likely that autism has a biological onset prior to birth. Such signs include, for example, incomplete formation of a group of neurons in the brainstem called the inferior olive, which is part of a learning circuit involving cerebellar Purkinje neurons.

Abnormal brain growth appears to continue after birth. For example, although whole brain volumes appear to be normal at birth, by the time an autistic child is 2–3 years old, the brain is far larger than normal. This pattern of inflated growth early on in the disorder is followed by a period of reduced growth so that eventually the normal brain outgrows the autistic brain. This illustrates the likelihood that neuroanatomic abnormalities in autism both interact and compound over time, and is consistent with the complex symptom profile seen in this disorder.

Two rules should guide interpretation of neuroanatomical findings in autism. First, apparently 'normal' measures of a particular brain structure (e.g., the parietal lobes in about 50% of autistic patients) do not necessarily imply normal function. The results from functional neuroimaging techniques such as functional magnetic resonance imaging (fMRI) suggest that several regions in the autistic brain may be functionally abnormal, in spite of appearing structurally normal. Second, 'normal' macroscopic structure does not necessarily mean normal microscopic structure. That is, given that much of what we know about the neurobiology of autism is obtained from MRI – an excellent technique for macroscopic analysis but less suitable for microscopic investigation – subtle defects in structure, such as dendritic or synaptic density, may be missed. As the number of histological cases examined increases, important and detailed questions about the neurobiology of autism will be answered.

Biochemical markers have also recently been reported in autism. For example, increased levels of several brain proteins, specifically neurotrophic factors, have been reported in the blood samples of newborn babies who were later diagnosed with autism. Neurotrophic factors are known to regulate cell growth and proliferation, and thus this finding is provocative in light of reports of increased brain growth early on in the disorder. The relationship between these two biological excesses (increased brain chemicals and increased brain growth) in autism may afford important insights into biologically based treatments and ultimately prevention of the disorder.

In conclusion, developmental brain defects, abnormal brain protein levels from birth, combined with the strong heritability component as shown by twin studies, together provide clear evidence that autism is a biological disorder, not a psychogenic one as was thought in the past.

Relationship Between Neural Defects and Behavioral Symptoms

Only a few studies have investigated the relationship between the brain and behavior in autism. Given the consistency of anatomical defects noted in the cerebellum, it should not be surprising that much of what is known about brain–behavior relationships in autism relates to this structure. For example, it is known that individuals with autism with one type of cerebellar defect, namely smaller vermis lobule VI–VII area measures, are less likely to explore their environment and more likely to engage in repetitive and stereotyped behavior than those with a more normal area measure. Similar relationships between the cerebellum and other behaviors such as attention have also been reported. For example, individuals with autism with more vermis VI–VII lobule abnormality take longer to orient their attention and make more errors when asked to respond to an attention-orienting stimulus than those with less damage. As another example, autistic patients with parietal defects are abnormally slow in disengaging their attention from a source of visual information in

order to attend to an unexpected different source of information.

Interesting work has also been done investigating the neural basis of social abnormalities in autism using fMRI. One finding has been that when autistic individuals look at the faces of strangers, neural activity in the expected brain region (i.e., the 'fusiform face area') is drastically reduced. However, when autistic subjects look at the faces of the people closest to them (e.g., mother or classmate), many brain regions are active including those essential to emotion processing. This fMRI evidence suggests that autistic individuals may not be as socially detached as once thought, and is one example of how neuroimaging techniques can add insight into our understanding of the behaviors found in autism. Many functional abnormalities, however, have been found in the autistic brain when processing more complex social stimuli, and many believe that social dysfunctions are related to abnormalities in the amygdala.

CONCLUSION

The constellation of symptoms that constitute autism are severe and affect almost every domain of functioning including language, cognition, social behavior, and environmental interests. Such symptoms are likely to be mediated by equally complex and diverse systems of biological abnormalities. Although treatment efforts have been successful at ameliorating some symptoms for some children with the disorder, currently there is no cure. Significant strides have been made, however, in understanding the neurobiology as well as the etiology of this disorder. For example, it is now known that multiple neuroanatomical sites are affected both structurally and functionally in autism, brain chemical levels are abnormal at birth, and some individuals present defects at the chromosomal level. Such insights may lead to biologically based interventions, and ultimately prevention of this disorder.

Further Reading

Bailey A, Luthert P, Dean A *et al*. (1998) A clinicopathological study of autism. *Brain* **121**: 889–905.

Courchesne E, Yeung-Courchesne R, Press GA, Hesselink JR and Jernigan TL (1988) Hypoplasia of cerebellar vermal lobules VI and VII in autism. *New England Journal of Medicine* **318**: 1349–1354.

Courchesne E, Yeung-Courchesne R and Pierce K (1999) Biological and behavioral heterogeneity in autism: role of pleiotropy and epigenesis. In: Broman S and Fletcher J (eds) *The Changing Nervous System: Neurobehavioral Consequences of Early Brain Disorders*, pp. 292–338. New York, NY: Oxford University Press.

Kanner L (1943) Autistic disturbances of affective contact. *Nervous Child* **2**: 217–250.

Lamb JA, Moore J, Bailey A and Monaco AP (2000) Autism: recent molecular genetic advances. *Human Molecular Genetics* **9**: 861–868.

Lewy AL and Dawson G (1992) Social stimulation and joint attention in young autistic children. *Journal of Abnormal Child Psychology* **20**: 555–566.

Lovaas OI (1987) Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology* **55**: 3–9.

Mundy P and Sigman M (1989) Theoretical implications of joint-attention deficits in autism. *Development and Psychopathology* **1**: 173–183.

Pierce K, Müller RA, Ambrose J, Allen G and Courchesne E (2001) People with autism process faces outside the 'fusiform face area': evidence from fMRI. *Brain* **124**: 2059–2073.

Townsend J and Courchesne E (1994) Parietal damage and narrow 'spotlight' spatial attention. *Journal of Cognitive Neuroscience* **6**: 220–232.

Autobiographical Memory

Intermediate article

David Rubin, Duke University, Durham, North Carolina, USA

CONTENTS

Autobiographical memory refers to the store of memories of events that have happened to an individual.

A TAXONOMY: RECOLLECTIVE MEMORY AND AUTOBIOGRAPHICAL FACTS

There is no universally agreed definition of autobiographical memory, but there is a taxonomy based on philosophical and behavioral considerations developed by Brewer (1986, 1996) that makes many issues clear. Two factors underlie the classification: (1) whether what was recalled was a single or a repeated occurrence and (2) whether the memory involves having an image. What is often called an *autobiographical memory*, a *personal memory* or a *recollective memory* occurs when a single event is recalled that involves an image. Thus you may have a memory of typing an email at a computer on one particular occasion in which you have an image, although it may be quite vague, of the email, the computer and your surroundings. A memory of the same event without the image would be an *autobiographical fact*.

In terms of the remember/know distinction commonly used in experiments of laboratory recognition, you know that you typed the letter but you would not remember, or more precisely you do not recollect, doing so. If the memory involves an image but is not for one instance of typing a particular email, but rather of sitting in your usual surroundings typing an email the way you usually do, you would have a *generic personal memory*. If all that you remember is the fact that you generally type emails on a particular computer, then you would have a *self-schema*. The four types of memories seem different to the lay person and the courts, and for a particular event that is one of a series of similar events these four types of memories can be lost independently either with the simple passage of time or as a result of neurological damage. Experimental psychologists who studied memory strived for a long time not to make introspection or phenomenology a part of the definitions of their terms or objects of study, but as I hope Brewer's analysis demonstrates, it appears to be needed.

RECOLLECTIVE MEMORY: MEMORY FOR THE PERSONAL PAST

From the previous description, autobiographical or recollective memory can be regarded as a small subset of memory – memory for an event that comes with an accompanying image. However, when one considers what is needed in order to have an autobiographical memory, the latter becomes a synthesis of many cognitive systems that are often put to other uses. To have a recollection one needs a memory for some details of an event, a sensory image (usually visual, but often in several modalities), a spatial sense of the location of actors and objects in the event, and usually emotional connotations. One does not need a sense of when the event occurred, as this is inferred or remembered independently of the other types of information just listed (Brewer, 1986, 1996; Larsen *et al*., 1996). The sensory, spatial and emotional information must be strong enough to lead to the metacognitive judgment that you recollect, as opposed to just know, that the event occurred. Moreover, there is almost always a belief that the event really happened to you – a belief that can exist in the face of counter evidence (Brewer, 1986, 1996). For many autobiographical memories there is a coherent narrative that links the memory to one's self or life narrative and helps to organize it (Conway and Pleydell-Pearce, 2000; Habermas and Bluck, 2000). The memories may arise as a

result of a conscious search or unbidden as involuntary memories through associations with ongoing thoughts or environmental cues without effort (Berntsen, 1998). Thus much of our cognitive and emotional ability comes into play in an autobiographical memory.

The visual image, sense of reliving or recollection, and belief that are part of autobiographical memories cause problems in the real world. For example, eyewitnesses and people who recover childhood memories of trauma have a strong belief in the accuracy of their memories, but at least on some occasions these beliefs have been shown to be unjustified. Although autobiographical memory is generally accurate, there are well-documented cases in which it is not. In many cognitive tasks, people use visual imagery as a basic form of mental models – objects and their locations in images can be transformed and manipulated at will. For instance, if you see yourself in an autobiographical memory, you have transformed the view seen out of your own eyes to that seen by an outside observer (Nigro and Neisser, 1983). However, in autobiographical memory people tend to view images as permanent, unchanging and accurate photographs. Thus you can manipulate an image yet later believe that it is unchanged. The sense of recollection and the belief that one's memories are generally true add to this problem.

VIVID OR 'FLASHBULB' MEMORIES

Brown and Kulik (1977) coined the phrase *flashbulb memory* to describe vivid memories of important public events, because it suggests surprise, relatively indiscriminate although not necessarily complete illumination, and brevity. According to Brown and Kulik, flashbulb memories are 'memories for the circumstances in which one first learned of a very surprising and consequential (or emotionally arousing) event' (Brown and Kulik, 1977, p. 73). Two issues raised by Brown and Kulik's paper, namely whether flashbulb memories are different in kind from other autobiographical memories, and whether they are more accurate, have fueled much research. There is no strong evidence that flashbulb memories are a different type of memory to other recollective memories; they might be nothing more than the most extreme case of vivid recollection outside the flashbacks discussed in the section below on memory for traumatic events. Nonetheless, the extreme sense of reliving that characterizes them makes them a special and interesting topic of study and a battleground for the issue of whether recollective memories are accurate. Here it appears that flashbulb memories are often accurate, but that major distortions can occur. A common error involves recalling two different times when one 'first learned' of an event; a relatively minor error in noting where one first heard of the event can result in large differences in the details and even major points of what is recalled. For example, in reporting when they first learned of the explosion of the *Challenger* space shuttle, people who initially reported hearing about it from a person later reported, often with great clarity and confidence, that they initially learned about it by watching television (Winograd and Neisser, 1992).

AUTOBIOGRAPHICAL MEMORY FOR EMOTIONAL AND TRAUMATIC EVENTS

The relationship between emotions and autobiographical memory is complex. The literature on this issue, although as rigorous as any on autobiographical memory, produces few generalizations (Christianson and Safer, 1996). For instance, the flashbulb memory literature shows good recall of the circumstances and context in which a shocking event occurred, while studies of similarly shocking events tend to show poor recall of the context and more focus on and better recall of the central events. Whereas in diary studies pleasant events tend to be much better recalled than unpleasant ones, most studies of emotion and autobiographical memory use unpleasant events, and these studies have found that increased intensity of emotion increases recall. At more extreme levels, such as traumatic memories, concepts from clinical psychology (e.g. dissociation and repression) are often invoked.

Traumatic events often violate one's expectations, making them difficult to integrate into a life narrative. They involve strong emotions, and when they return in memories they can bring with them an overpowering and unwanted sense of reliving in the form of intrusive memories. Thus autobiographical memories for traumatic events are at the extremes of the systems for and phenomenology of autobiographical memory noted earlier. In some case, traumas lead to post-traumatic stress disorder, a syndrome that is defined by reliving, avoidance and arousal symptoms (American Psychiatric Association, 1994). The reliving symptoms can occur in the form of intrusive memories, which makes autobiographical memory part of the diagnosis and, for the patient, part of the problem. The intrusive memories may occur as flashbacks – an

extreme form of flashbulb memories which takes the patient back to the time and setting of the original trauma.

THE DISTRIBUTION OF AUTOBIOGRAPHICAL MEMORY OVER THE LIFESPAN

One of the most regular quantitative findings in the literature on autobiographical memory is the distribution of memories over the lifespan. Autobiographical memories for such analyses have been obtained in many different ways, such as having people give their lives in narrative form, just list events, or provide memories of specific types (e.g. important memories, or memories cued by odors). Although the results are similar, most work has been done with word cues, so this approach is the easiest to synthesize over many studies. This method was developed by Galton over a century ago and revived by Crovitz and Schiffman (1974). Individuals give a memory to each of a set of randomly chosen words, and after all autobiographical memories have been obtained, they date them. The distribution of memories over the lifespan of undergraduates can be described as a power function of the time since the event occurred (i.e. $y = at^{-b}$) (Crovitz and Schiffman, 1974). The fits to the curve are surprisingly good, with correlations usually being greater than 0.95. If we assume that undergraduates encode an equal number of events each day of their lives, then the curve is a retention function. Because the power function is a common choice for a retention function, it appears that as a first approximation, laboratory and autobiographical memory have similar patterns of forgetting.

The data for older adults are more complex, requiring three components. The first component is the retention function, which can be described mathematically as a power function, that covers the whole lifespan, but which has its main quantitative contribution over the two most recent decades of life; for periods longer than two decades ago, it is too small to produce a measurable number of memories. The second component is childhood amnesia. This component is very stable over numerous studies, having the same basic shape irrespective of the method used to produce the data, so long as the participants come from the USA (Rubin, 2000). Based on an average of the available data, the percentage of memories from before the age of 8 years for when a person was 0, 1, 2, 3, 4, 5, 6 and 7 years old, respectively, are 0.1, 0.4, 1.7, 5.5, 13.1, 22.1, 27.0 and 30.1% (or only 2.2% before the third birthday). Figure 1 shows this component.

The third component, known as the bump, is an increase in the number of memories from when older adults were between 10 and 30 years of age compared with what would be expected from the other two components or from any monotonically

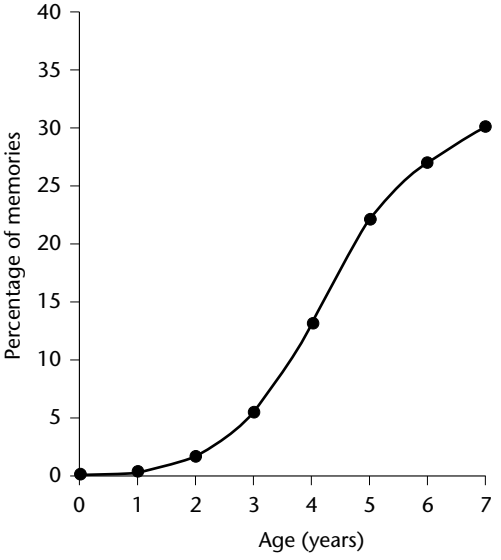


Figure 1. The distribution of 10 118 memories dated as occurring before the age of 8 years from published studies. From Rubin (2002).

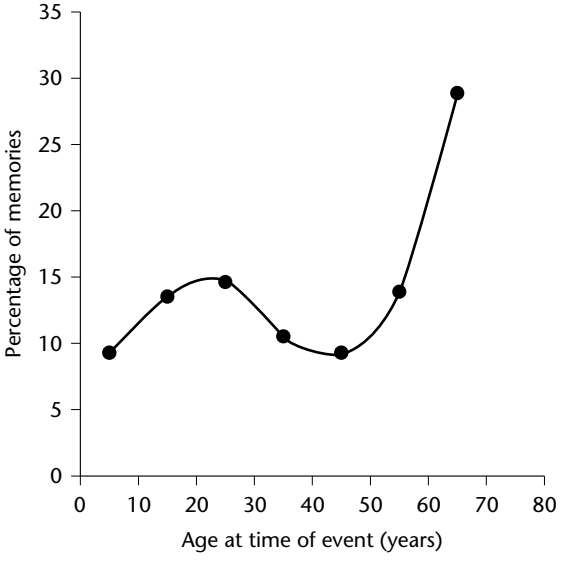


Figure 2. The distribution of autobiographical memories over the lifespan for older adults; events dated as occurring in the most recent year are excluded. From Rubin *et al*. (1986).

decreasing function. The initial description of the bump in 1986 was based on data from several laboratories. Since that time there have been consistent findings using the word cue technique with older adults (Rubin, 2002; Rubin *et al*., 1998). Minor differences exist in the shape of the distribution with changes in procedure, but the bump appears repeatedly, even for individuals. Figure 2 shows this component.

References

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Berntsen D (1998) Voluntary and involuntary access to autobiographical memory. *Memory* **6**: 113–141.

Brewer WF (1986) What is autobiographical memory? In: Rubin DC (ed.) *Autobiographical Memory*, pp. 25–49. Cambridge, UK: Cambridge University Press.

Brewer WF (1996) What is recollective memory? In: Rubin DC (ed.) *Remembering our Past: Studies in Autobiographical Memory*, pp. 19–66. New York, NY: Cambridge University Press.

Brown R and Kulik J (1977) Flashbulb memories. *Cognition* **5**: 73–99.

Christianson S-A and Safer MA (1996) Emotional events and emotions in autobiographical memory. In: Rubin DC (ed.) *Remembering our Past: Studies in Autobiographical Remembering*, pp. 218–243. Cambridge: Cambridge University Press.

Conway MA and Pleydell-Pearce CW (2000) The construction of autobiographical memories in the self-memory system. *Psychological Review* **107**: 261–288.

Crovitz HF and Schiffman H (1974) Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society* **4**: 517–551.

Habermas T and Bluck S (2000) Getting a life: the emergence of the life story in adolescence. *Psychological Bulletin* **126**: 748–769.

Larsen SF, Thompson CP and Hansen T (1996) Time in autobiographical memory. In: Rubin DC (ed.) *Remembering our Past: Studies in Autobiographical Memory*, pp. 129–156. New York, NY: Cambridge University Press.

Nigro G and Neisser U (1983) Point of view in personal memories. *Cognitive Psychology* **15**: 467–482.

Rubin DC (2000) The distribution of early childhood memories. *Memory* **8**: 265–269.

Rubin DC (2002) Autobiographical memory across the lifespan. In: Graf P and Ohta N (eds) *Lifespan Development of Human Memory*, pp. 159–184. Cambridge, MA: MIT Press.

Rubin DC, Rahhal TA and Poon LW (1998) Things learned in early adulthood are remembered best. *Memory and Cognition* **26**: 3–19.

Winograd E and Neisser U (eds) (1992) *Affect and Accuracy in Recall: Studies of 'Flashbulb' Memories*. New York, NY: Cambridge University Press.

Further Reading

Conway MA (1990) *Autobiographical Memory: An Introduction*. Milton Keynes, UK: Open University Press.

Conway MA (1995) *Flashbulb Memories*. Hove: Erlbaum.

Conway MA, Rubin DC, Spinnler H and Wagenaar WA (eds) (1992) *Theoretical Perspectives on Autobiographical Memory*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rubin DC (ed.) (1986) *Autobiographical Memory*. Cambridge, UK: Cambridge University Press.

Rubin DC (ed.) (1996) *Remembering our Past: Studies in Autobiographical Memory*. New York, NY: Cambridge University Press.

Winograd E and Neisser U (eds) (1992) *Affect and Accuracy in Recall: Studies of 'Flashbulb' Memories*. New York, NY: Cambridge University Press.

Automaticity

Introductory article

Thomas J Palmeri, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

Automaticity refers to the way we perform some mental tasks quickly and effortlessly, with little thought or conscious intention. Automatic processes are contrasted with deliberate, attention-demanding, conscious, controlled aspects of cognition.

INTRODUCTION

Try to think back to when you first learned how to drive a car. Your primary aim was to steer the car clear of other vehicles, pedestrians, and trees – a difficult task by itself. But you also had to control the pressure applied to the accelerator pedal to keep within posted speed limits. You needed occasionally to apply the brake to obey traffic signals and to avoid plowing into the car in front of you. Added to this, if you first learned to drive a car with a manual transmission, you had to decide when to change gear and then you needed to coordinate the complex movements involved in doing so – releasing the accelerator pedal, depressing the clutch, shifting to the appropriate gear, carefully releasing the clutch while applying some gas. And you had to do this while continuing to pay attention to the road ahead. On top of that, you probably had to linguistically process the commands, pleas, and screams of the poor soul who (perhaps regrettably) agreed to teach you how to drive. You had to direct all your mental energies to controlling and coordinating the complex sequence of movements involved in safely driving a car. Trying to simultaneously steer, accelerate, brake, shift, and listen was an exceedingly difficult task.

Contrast this scenario with how you may be able to drive after many years of practice. On long trips, you find yourself daydreaming and may not even remember what happened during the last several uneventful miles of highway driving. Shifting gears becomes one smooth continuous action. Indeed, breaking up this complex action into its component parts may require some deliberate thought – in fact, on my initial draft of the previous paragraph, I forgot that the first critical step in shifting was to release the accelerator pedal; this is something I have done thousands of times during my twenty years of driving cars with manual transmissions, but this basic action did not initially come to mind when I tried consciously to decompose the act of shifting gears. Experienced drivers use so few mental 'resources' that some people can drink coffee, talk on a cellular phone, and groom themselves while driving at high speeds on a congested expressway. Things are fine until something unexpected happens – another distracted driver veers into their lane or someone stops very abruptly ahead – now those resources diverted to drinking, talking, and grooming are not available to take immediate action to avert a serious accident.

That effortless way that we perform the various components of skilled actions, such as driving a car, is termed *automaticity*. Many routine daily events become so automatic that we may seem unconscious of them – how many times have I lathered my hair this morning, did I remember to put the freshly ground coffee in the coffee maker, have I checked my mailbox yet this morning? Literate adults read automatically – try not to read the billboards and signs that bombard you when driving through suburban commercial developments. When skilled at playing a musical instrument, reading musical notation, translating notes into finger and hand movements, controlling breathing and embouchure (mouth position), are all automatized procedures, allowing the musician to focus on higher levels of musicality such as style, phrasing, and coordination with the conductor and other musicians. Skilled professionals automatically execute complex tasks that demand years of training. Experienced radiologists may be able to tell automatically, at a glance, whether a patient has a benign growth or a malignant tumor.

Experienced pilots control complex aircraft automatically. Landing a commercial jetliner in good weather is performed with nearly the same fluency as driving to the neighborhood grocery store. This automaticity allows the pilot to monitor for unexpected events – an unauthorized aircraft on the runway, an approaching flock of geese, an engine fire, or wind sheer – and be able to take corrective action immediately to avert potential disaster.

This article describes the properties that distinguish automatic processes from those that require conscious mental control, describes factors necessary for achieving automaticity, illustrates the effects of automaticity with some classic experimental paradigms, and describes some psychological models of the acquisition of automaticity.

CHARACTERISTICS OF AUTOMATIC PROCESSES

A number of characteristics have been emphasized to distinguish *automatic processes* from those that require some kind of overt mental control, what have been referred to as *controlled processes*. Theorists disagree on what particular characteristics are most important for describing a process as being automatic, and disagree on whether some particular properties appropriately characterize automaticity at all. In addition, some theorists have argued that perhaps the concept of automaticity itself should be abandoned entirely, since no cognitive process is ever truly automatic given most lists of critical characteristics. Automaticity is a current topic of active research in the cognitive sciences, and ideas of how best to characterize automatic processing are still evolving.

Table 1. Some proposed characteristics of automatic and controlled processes

Automatic processes	*Controlled processes*
obligatory	intentional
stimulus-driven	executive-driven
stereotypic	reconfigurable
rigid	flexible
no monitoring	monitoring
no dual-task interference	dual-task interference
parallel	serial
well practiced	novel
expert	novice
fast	slow
effortless	effort
unconscious	conscious
no attention	attention

The aim of this section is to survey most of the various characteristics of automaticity that have been proposed. These characteristics, summarized in Table 1, will be elaborated upon below. These characteristics should certainly not be considered independent dimensions of automatic processes because many of them may overlap in some respects.

- Automatic processes are *obligatory*. Given the presence of particular stimuli within particular contexts, automatic processes can execute without the conscious intention of the individual. Automatic processes seem to occur reflexively. Controlled processes require conscious intention to become initiated.
- For this reason, automatic processes are said to be *stimulus-driven*. Given the appropriate triggering conditions, automatic processes execute without intention. Controlled processes are intentionally initiated by the individual, often with the guidance of central executive processes.
- Automatic processes are often *rigid* and *stereotypic*. Controlled processes can be reconfigured to deal with novel events, allowing for a far greater degree of flexibility.
- Once initiated, automatic processes require *no monitoring*. They run to completion without any need for overt executive control. Controlled processes require monitoring, and distractions can lead to breakdowns in performance.
- Automatic processes are *free from dual-task interference*; that is, automatic processes are not influenced by other tasks that are executed concurrently. Controlled processes suffer from dual-task interference. It is often extremely difficult to perform more than one controlled process at the same time.
- Because automatic processes can execute simultaneously, they are said to be processed in *parallel*. Not only can independent automatic processes be executed in parallel, but the various component processes of a complex skill may overlap one another in a parallel manner. Controlled processes execute serially. They are processed one step at a time and cannot be processed simultaneously.
- Many automatic cognitive processes are *well practiced*. Controlled processes may be novel or less practiced.
- Automatic processes often characterize *expert* performance. Controlled processes often characterize novice performance.
- Because automatic processes can be performed in parallel without conscious monitoring, automatic processes are often *fast* compared to controlled processes.
- Automatic processes seem *effortless*. Controlled processes require mental effort.
- Automaticity is often discussed in the context of consciousness. Automatic processes may be *unconscious*. Controlled processes are conscious.
- Automaticity is also often discussed in the context of attention. Automatic processes may require *no attention*. Controlled processes do require attention.

FACTORS NECESSARY FOR AUTOMATIC PROCESSES

Some processes may be automatic because the human brain is equipped with special-purpose neural mechanisms for carrying out certain critical aspects of perception and cognition. Such automatic processes are obligatory because a specialized neural 'module' operates autonomously, triggered by particular stimulus events in the environment. These are hard-wired mechanisms, making them rigid and stereotypic. Because these modules operate independently, they are not influenced by other concurrent processes operating within other parts of the brain, they do not require monitoring, they operate unconsciously, and they require no overt deployment of attentional resources.

Let us illustrate an example of an automatic process that may reflect the operation of one such hard-wired perceptual mechanism. Search for a yellow **X** in each panel of Figure 1. The target automatically 'pops out' from the distractors in the left panel but an active search is required to locate the target among the distractors in the right panel. In the left panel, the yellow **X** differs from the red **X**s by a single feature, but in the right panel, the yellow **X** differs from the yellow **O**s and red **X**s by the particular combination of color and form features. Salient singleton features are thought to pop out automatically because of the way early stages of the visual system process elementary visual information. Indeed, visual search tasks have often been used to distinguish automatic, pre-conscious processing of elementary visual features from the more high-level, attention-demanding processing of conjunctions of multiple visual features necessary for object recognition. Similarly, we may also automatically notice other kinds of perceptual events such as abrupt onsets of visual stimuli (a flash of lightning), auditory stimuli (a clap of thunder), or somatosensory stimuli (a crawling insect), because our perceptual systems may be hard-wired to process sudden unexpected changes in the environment automatically. So some aspects of perception and cognition may be automatic, and truly reflexive in nature, because there exist special-purpose neural mechanisms that operate autonomously, below the level of conscious awareness and control.

Clearly there do not exist innate hard-wired mechanisms for reading a book, driving an automobile, or flying an airplane. Yet people can become automatic at the elements of these tasks with sufficient practice. Therefore, a great deal of automaticity must be learned. How can a process go from being one that requires overt cognitive control to one that is automatic? And are there limitations on what kinds of tasks can become automatized?

For most aspects of human cognition that can become automatized, no one achieves automaticity without a great deal of practice. But some tasks may become automatized more quickly than others. Clearly, a simple task may be automatized more quickly than a complex task. Some real-life tasks may take only a few hours of practice to become automatized. Others require many years of training. But complexity is not the only factor, nor the most critical factor in determining how rapidly a task can become automatized.

To illustrate, let us consider another example of a search task that has been used to study the development of automaticity. The visual display can contain between one and four letters (call this variable the display size, D). You can be asked to search for between one and four possible target letters (call this variable the memory set size, M). The task is to decide whether a target is present or absent in each display as quickly as possible without making any errors. The time to make a correct response will be recorded. A display size of one ($D = 1$) and a memory set size of one ($M = 1$) is a relatively easy search. As shown in the top of Figure 2, if the target is an **X**, the 'search' is simply a matter of deciding whether the single presented letter on each display is an **X** or not. As the number of items in the display is increased, the task gets harder, and as the number of items in the memory set increases, the task gets harder. As shown in the bottom of Figure 2, suppose I tell you that the target memory set is now **T**, **L**, **Z**, and **V** ($M = 4$). Each display will contain four letters ($D = 4$) and you must decide if any of those four letters is one of the four target letters in the memory set. This search is quite hard. To accomplish this task, people generally search through each item in the display one at a time and compare it with each item in the memory set one at a time until a target is found. As such, search times increase systematically as a function of both the display size and the memory set size. This is a slow, deliberate, attention-demanding, serial search process.

Can this controlled search become automatized through training? Imagine that the set of targets and distractors changes throughout training such that a target on one trial may be a distractor on another trial. In such *varied mapping* conditions, it is very difficult, if not impossible, for the search task ever to become automatized, even with

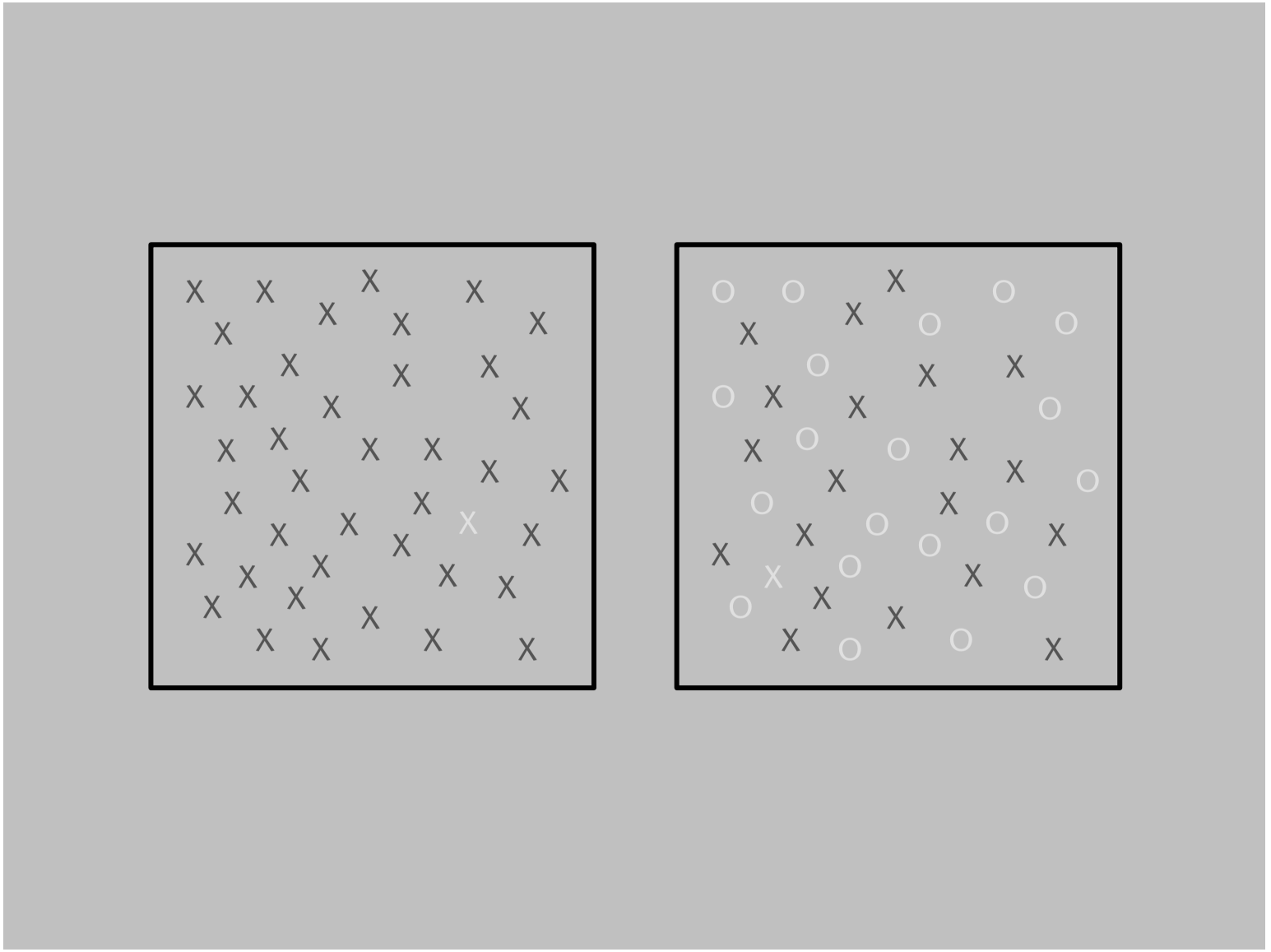


Figure 1. [*Figure is also reproduced in color section.*] Find the yellow **X**. The left panel illustrates a feature search task in which the target automatically 'pops out' from the field of distracters. The right panel illustrates a conjunction search task in which the target must be actively searched for with deliberate shifts of attention.

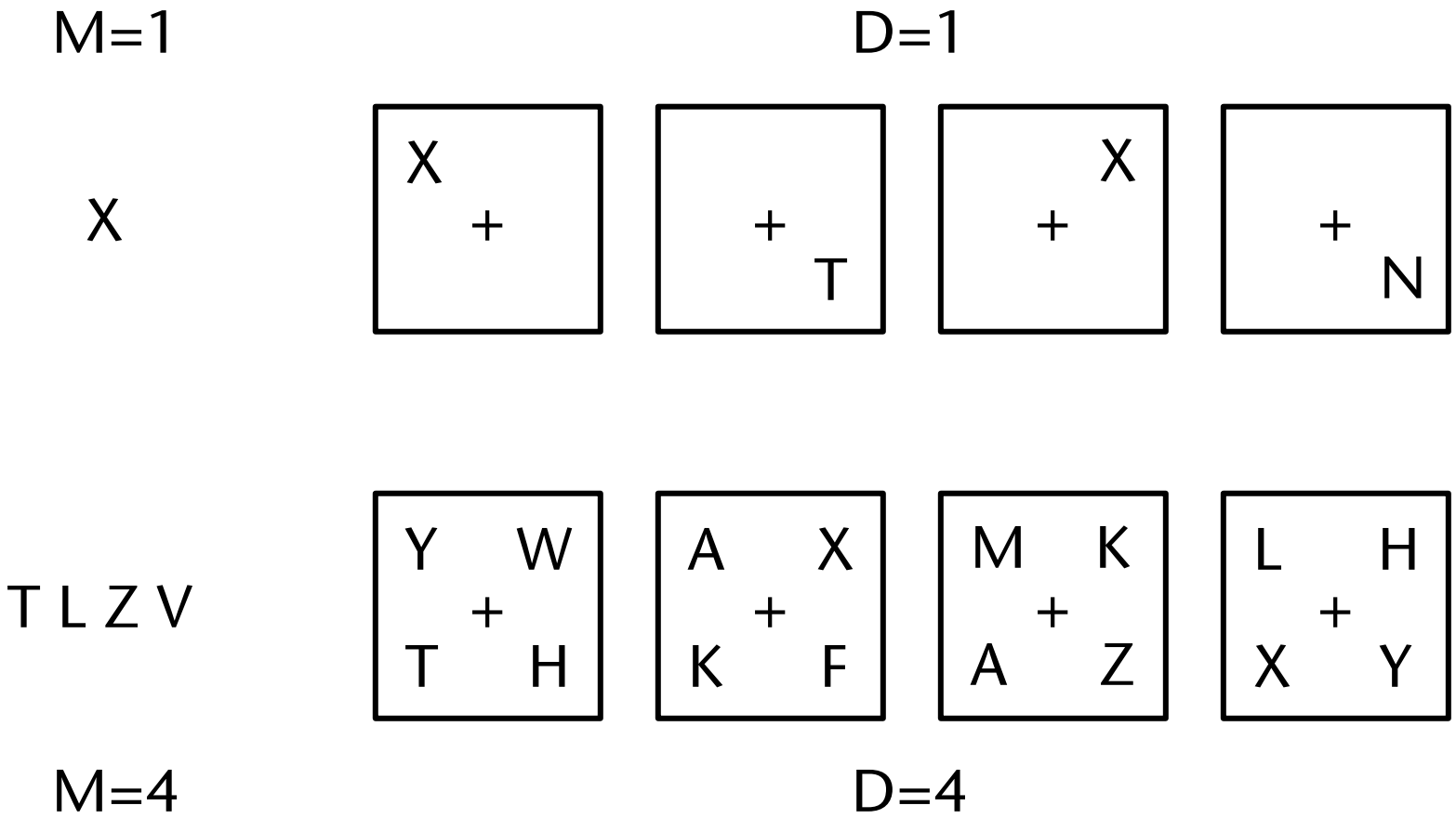


Figure 2. Illustration of search task that manipulates display size (D) and memory set size (M). The top search has a single target (M = 1) and a single display item on every trial (D = 1). The bottom search has four possible targets (M = 4) and four display items on every trial (D = 4). The task is to detect a target as quickly as possible without making errors.

extended practice over several weeks. So practice by itself is not guaranteed to produce an automatic process.

Instead imagine that the set of targets and distractors remains consistent, such that the targets must be drawn from one set of letters and the distractors must be drawn from a different set of letters, and this differentiation is maintained throughout the entire course of training. In such *consistent mapping* conditions, automaticity can be achieved with practice. Indeed, after extended practice, the time taken to search for targets does not vary with display size or memory set size. That is, a target pops out from the display much like the color pop-out shown in the left panel of Figure 1. But this is a learned automaticity, not a hard-wired one. This automaticity is immune to dual-task interference. This automaticity is rigid and inflexible in that switching to a varied mapping condition causes the search to revert back to a slow, deliberate, attention-demanding, serial process. Moreover, switching targets to distracters and distracters to targets causes performance to become even worse than it was before any training whatsoever, and it takes a long time to 'unlearn' the original automatization of target searches. So one important criterion for developing automaticity is that there is a consistent mapping between stimuli and responses. This may be one reason for the stimulus-driven nature of much automatic processing.

STROOP INTERFERENCE AND OTHER RELATED MEASURES

A different manifestation of automaticity can be seen in the classic Stroop effect. Named after John Ridley Stroop, the psychologist who developed it as part of his doctoral dissertation in the 1930s, the Stroop task has been used in thousands of experiments to study automaticity. First, find a stopwatch or a clock with a second hand. Now, time how long it takes you to *name the ink color* of the words in the first column of Figure 3 (i.e. BLUE, RED, PURPLE, etc.) – name the ink color, don't read the words. Next, time how long it takes to name the ink colors in the second column (i.e. RED, BLUE, ORANGE, etc.). And then do the same with the third column (i.e. PINK, RED, YELLOW, etc.). In all cases, try to respond as quickly as possible without making errors.

The classic Stroop interference effect is that the identity of the word can have a large effect on the speed of color naming. In the first column, the words themselves have no color association. In the second column, each word is congruent with its ink color, such as 'red' in RED ink or 'green' in GREEN ink. People are generally a bit faster to name the colors in the second column (congruent condition) than to name the colors in the first column (control condition). In the third column, each word is incongruent with its ink color, such as 'red' in GREEN ink or 'blue' in YELLOW ink. People are generally far slower to name the colors in the third column (incongruent condition) than to name the colors in the other columns. In the original paper by Stroop, subjects took nearly twice as long to name colors in the incongruent condition than in the control condition, a finding that has since been replicated thousands of times across numerous experimental variations. Even without a stopwatch, you surely found naming ink colors in the third column quite difficult and perhaps a bit frustrating. This is the fundamental Stroop interference effect.

Stroop interference is not simply caused by having an incongruency between words and their ink color. Time how long it takes you to *read the words* in the first column of Figure 3 (i.e. TRUCK, HOUSE, GRASS, etc.). Then time how long it takes to read the words in the second column (i.e. RED, BLUE, ORANGE, etc.). Then do the same with the third column (i.e. GREEN, PINK, BLUE, etc.). Ink color has little or no effect on the speed of reading words. Even for color words like 'red', the speed of word reading is not influenced by whether the word 'red' is written in RED ink or GREEN ink.

Stroop interference is asymmetric. In the incongruent condition, words interfere with color naming but colors do not interfere with word reading. One acknowledged explanation for this is that word reading is a more highly automatized process than color naming. Word reading happens rapidly and effortlessly, without conscious intention, and cannot generally be suppressed. Naming colors requires more attention, conscious intention, and effort. Even when the task is to name the colors, and to ignore the words, word reading happens anyway, automatically, and can interfere with color naming.

The Stroop effect is not limited to interference of word reading on color naming. Figure 4 shows incongruent conditions from three variants of the Stroop task. In the first column, the task is either to read the digits (i.e. 4, 3, 5, etc.) or to count the number of digits (i.e. THREE, FIVE, FOUR, etc.). Reading digits is more automatized than counting, so digit identity interferes with counting, but the number of digits does not interfere with digit naming. In the second column, the task is either to

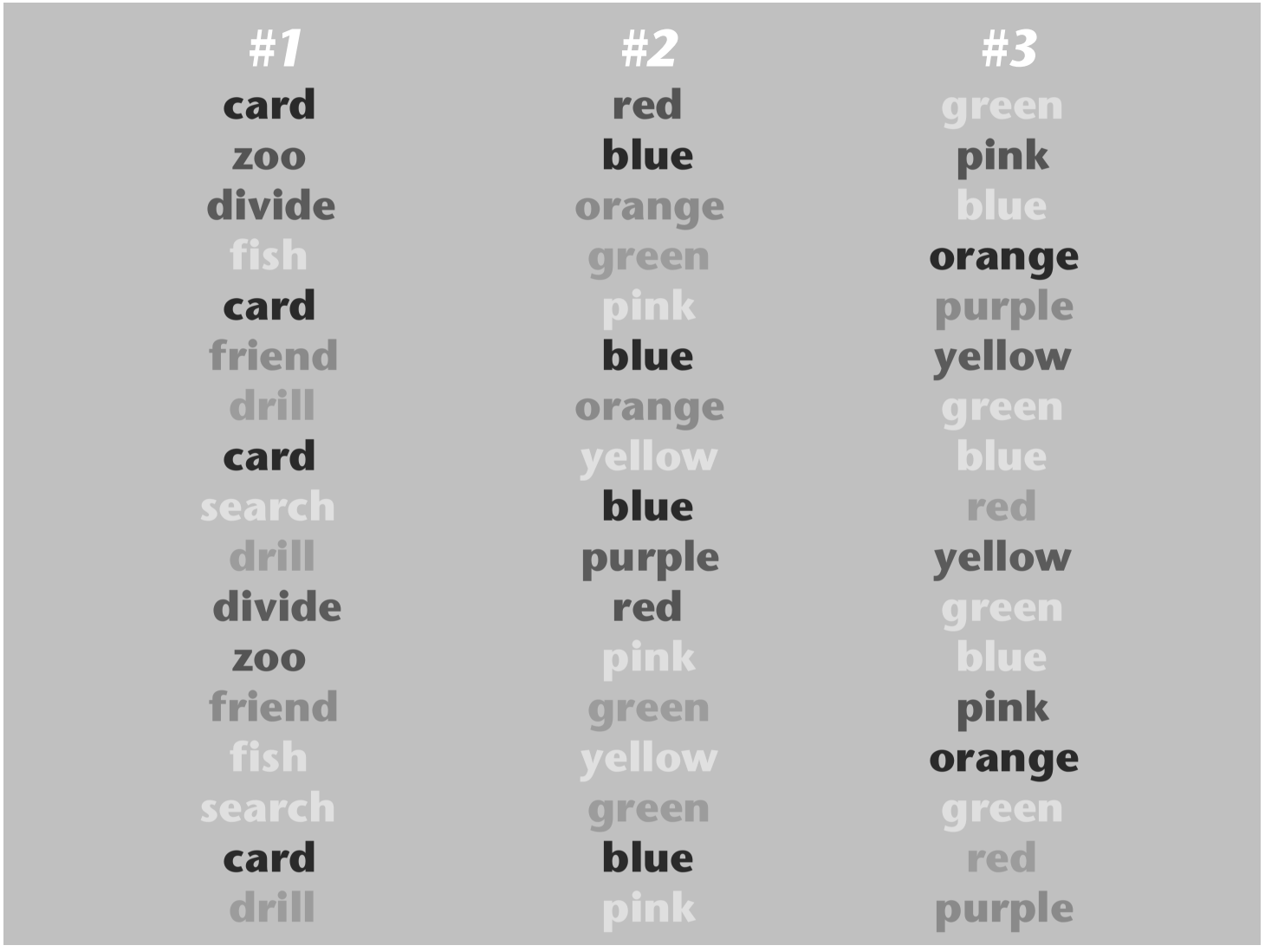


Figure 3. [*Figure is also reproduced in color section*.] Demonstration of the Stroop task. Using a stopwatch, separately time how long it takes to *name the color* of each printed word in column 1, column 2, and column 3. Then separately time how long it takes to *read each word* in column 1, column 2, and column 3. Column 1 is a *control condition* in which the word and the color bear no relationship. Column 2 is a *congruent condition* in which the word and the color match. Column 3 is an *incongruent condition* in which the word and the color mismatch.

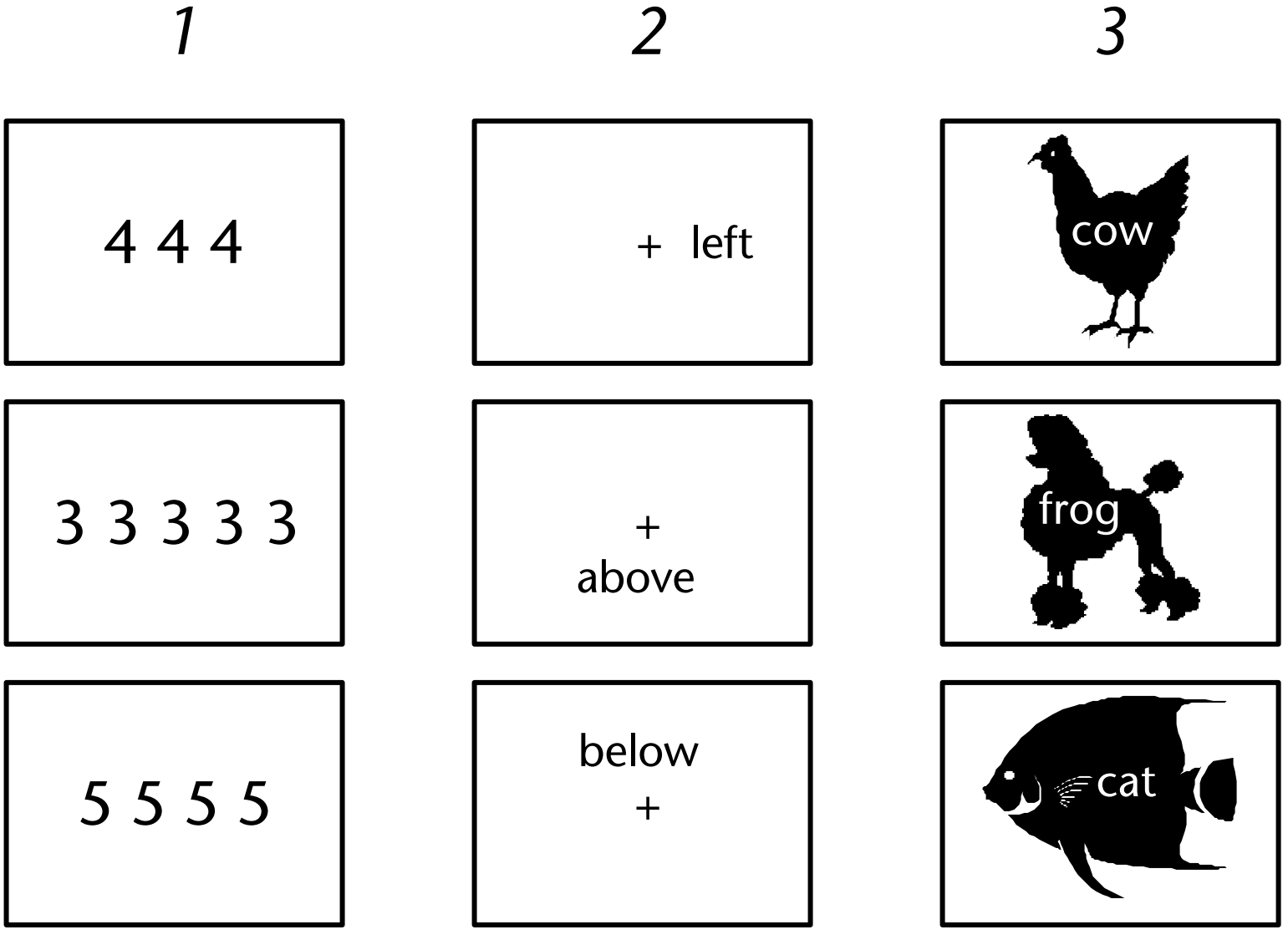


Figure 4. Illustration of incongruent conditions from three variants of the Stroop task. In column 1, you either name the digit or count the number of digits. In column 2, you either read the word or describe the spatial position of the word with respect to the central cross. In column 3, you either read the word or name the picture.

read the spatial terms (i.e. 'left', 'above', 'below', etc.) or to specify the location of the term with respect to the central cross (i.e. RIGHT, BELOW, ABOVE, etc.). Word identity interferes with specifying spatial locations, but not vice versa. Finally, in the third column, the task is either to read the animal name (i.e. 'cow', 'frog', 'cat', etc.) or to name the animal (i.e. BIRD, DOG, FISH, etc.). Word identity interferes with object naming, but not vice versa.

The classic case of Stroop interference is thought to occur because word reading is more automatic than color naming. If automaticity can be achieved through training, might it be possible to influence the direction of Stroop interference by manipulating practice with color naming? In principle, it should be possible to have color names interfere with word reading if color naming has been sufficiently practiced. But even with practice, it is extremely difficult to overcome the great prior advantage of word reading over color naming.

Instead imagine that you have just memorized that the symbols shown in Figure 5 are glyphs in some ancient language for the concepts blue, yellow, green, and red, respectively. The glyphs can be filled with various colors, creating congruent stimuli (e.g. 'blue' glyph in BLUE) or incongruent stimuli (e.g. 'red' glyph in YELLOW), as illustrated in the figure. When asked to name the color of the glyph, color naming is not influenced by the identity of the glyph, but when instead asked to name the glyph, glyph naming is strongly influenced by the color of the glyph. Because color naming is much more automatized than glyph naming, color interferes with glyph naming, but not vice versa. Now imagine that you are trained on glyph naming for several weeks, causing glyph naming to become more automatized than color naming. The direction of Stroop interference now reverses. Glyph identity interferes with color naming, but not vice versa. Results such as these suggest a continuum of automaticity, with the direction of Stroop interference a potential marker for which cognitive process is more automatized.

MODELS OF THE ACQUISITION OF AUTOMATICITY

Resource theories are based on the intuitive notion that people seem to have a limited amount of mental 'energy' that can be allocated to performing various tasks. Controlled processes require a certain amount of these limited mental resources whereas automatic processes do not. Automatic processes are fast because they are not limited by available resources. Automatic processes are effortless because mental 'effort' is proportional to the amount of resources needed to execute a process. Automatic processes are free from dual-task interference because they do not have to compete for the limited pool of resources. Automatic processes are obligatory because they do not need to wait until resources have been specifically allocated for their execution. The development of automaticity is viewed as a fundamental change in a process that makes it go from a resource-demanding controlled process to a resource-free automatic process. One criticism of resource theories has been that the learning mechanism by which processes reduce their resource demands is generally unspecified.

Another problem for resource theories is that complex patterns of interference have sometimes been observed. Some tasks interfere with one another, others do not. To deal with this complexity, some theorists have proposed that there may be multiple pools of mental resources. As an analogy, we could imagine that some processes consume electricity, other processes consume gasoline, and others consume coal. Any time that two tasks interfere with one another, they must be dipping from the same pool of limited resources. While intuitively appealing, *multiple resource theories* have been criticized as being inherently untestable assertions. Any complex pattern of task interference effects is explained *post hoc* by positing multiple pools of resources.

Instead of viewing resources as mental energy that is allocated to different tasks, resources may instead be conceptualized as specific processing components of the cognitive system that different tasks may need to share. As an analogy, we could imagine a mental toolbox, with some tasks requiring a screwdriver, others requiring a hammer, and others a saw. When two tasks need to use the same tool, they have to wait their turn. For example, working memory is limited. To the extent that two tasks both store information in working memory, they may interfere with one another. There is evidence for multiple modality-dependent working memory systems for verbal, spatial, and object information, so complex patterns of interference may be the result of different tasks placing demands on different working memory systems. To the extent that an automatic process is divorced from its reliance on working memory, it will not interfere with other processes that demand those limited processing resources.

In an extreme case, there may be some central process that must be shared by all aspects of

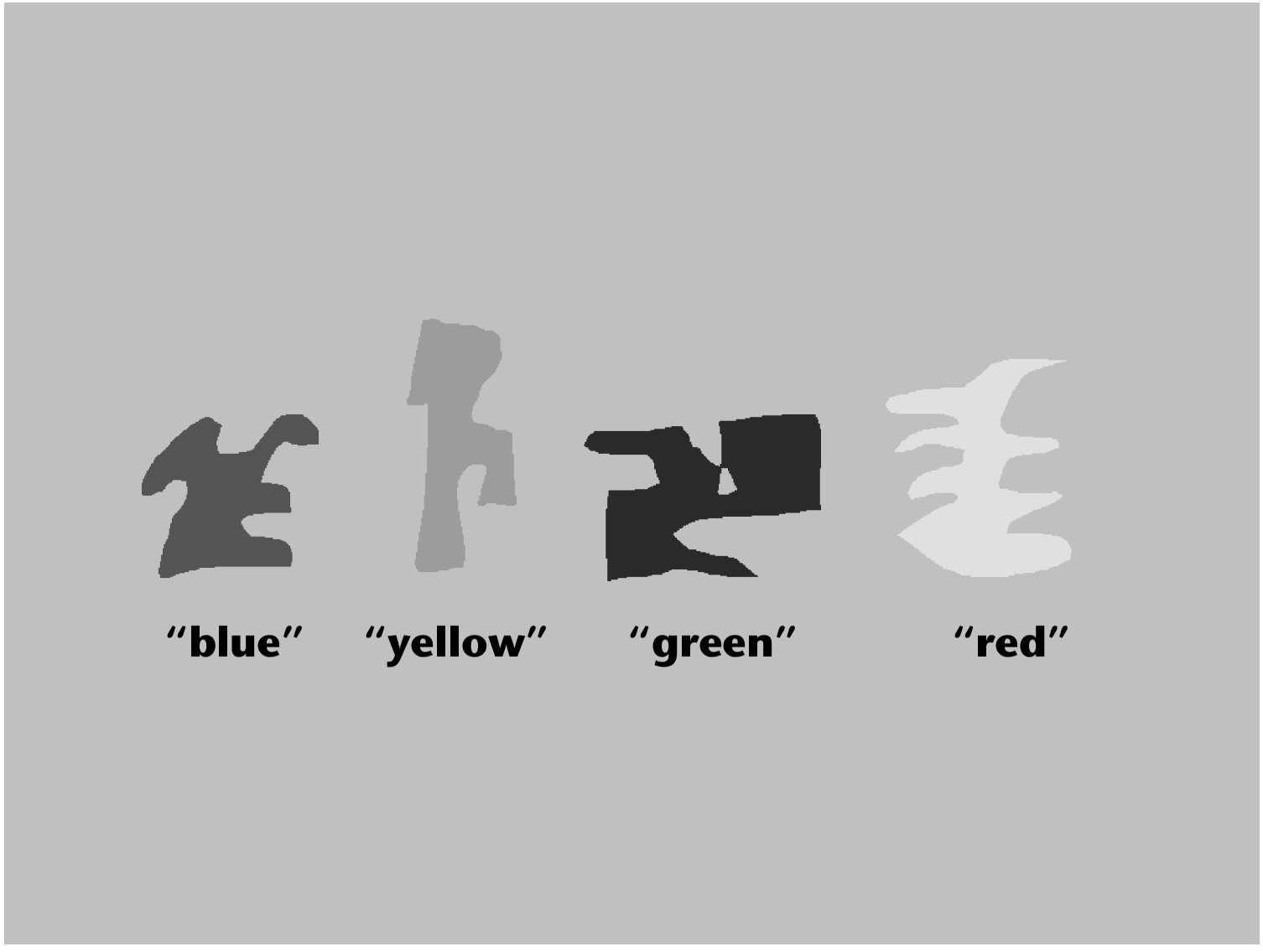


Figure 5. [*Figure is also reproduced in color section*.] Illustration of novel stimuli used to manipulate the direction of Stroop interference through training. Each shape (glyph) is associated with one of four color names that must be learned. Each shape can also be filled with one of four colors. Subjects either name the shape ('blue', 'yellow', 'green', or 'red') or name the color of the shape (RED, GREEN, BLUE, or YELLOW). Early in training, color interferes with shape naming. Later in training, shape interferes with color naming.

cognition that require selection among competing responses, what has been termed a *central bottleneck theory*. All tasks can be decomposed into a series of processing stages that extend from stimulus to response. Bottleneck theory posits that a particular one of these stages, that responsible for selecting among competing responses, can only be dedicated to one task at a time. All other stages prior to and subsequent to the response selection stage may proceed in parallel, but only one process can access response selection – other processes must wait. According to this theory, no cognitive process, no matter how highly practiced, can ever become truly automatic because all processes must share the limited response selection resource.

Our discussion of the Stroop effect should convince you that automaticity is not an all-or-none phenomenon. Word reading interferes with color naming because word reading is more automatic than color naming. But color naming interferes with shape naming because color naming is more automatic than shape naming. But with training, shape naming interferes with color naming because shape naming is now more automatic than color naming. It is not clear how a resource account could explain these asymmetric interference effects, nor how the direction of interference effects can be modulated by training. *Strength theories* represent learning in terms of the strength of association within pathways from particular stimuli to particular responses. Such theories have been implemented within a variety of frameworks from production systems to connectionist networks. The development of automaticity is seen as the strengthening of particular associations – be they production rules or connection weights – as a function of experience. Where these pathways intersect, interference can be observed. Stronger pathways interfere more with weaker pathways, leading to asymmetric interference effects.

Finally, *instance theories* propose a different account of the development of automaticity. Controlled processes are the result of the execution of some explicit algorithm whereas automatic processes are the result of memory retrieval. The development of automaticity is caused by a transition from algorithm to retrieval. When first engaged in some task, people may use an algorithm or rule to execute that behavior. For example, when first learning to add single digits, children typically adopt a strategy of starting with one of the digits and counting the requisite number of additional digits to generate the answer. Instance theory equates automaticity with memory retrieval. With experience, children (and adults) just remember that $2 + 2$ equals 4 without needing to explicitly count. Automatic processes are fast because memory retrieval is fast. Automatic processes are obligatory because memory retrieval is obligatory. Automaticity is effortless because memory retrieval seems effortless, especially compared with the execution of a multistep algorithm. Automatic processes are free from dual-task interference because a single memory retrieval offers less opportunity for interference than a multistep algorithm.

Execution of the algorithm and memory retrieval are assumed to take place concurrently, racing against one another to completion. The winner of the race determines what response is made. Early in learning, the algorithm is used because it completes before memory retrieval can finish (or because no memories can be retrieved). The development of automaticity is caused by the obligatory encoding of stimuli and responses in memory. As more memories of solutions are stored, memories can be retrieved more quickly. Thus, with experience, memory retrieval can eventually complete before the algorithm can complete. Consistent mappings are important because they yield consistent information from memory; varied mappings yield conflicting information from memory.

SUMMARY

Automatic processes are the autopilots of human cognition. They seem to execute outside our awareness and without our conscious control. They seem to execute quickly, and we may be entirely unaware of the steps involved in their execution. They can execute while we are doing other things at the same time. Some processes are automatic because our brains have evolved special-purpose mechanisms that respond without our conscious intention, and even sometimes against those intentions. Other processes can become automatic because of our experiences. Automaticity can be learned. But automaticity can be achieved only under certain circumstances: it may well be that some processes may just never become automatic, regardless of how much experience a person has had. The concept of automaticity is intimately tied with concepts of attention, consciousness, learning, and memory. Some research aims to relate attention and automaticity, with some attempts to relate both to the far more elusive concept of consciousness. Other research aims to relate the development of automaticity to what we know about learning and memory more generally, examining how they all manifest themselves across the full spectrum of human cognition.

Acknowledgements

This work was supported by NIMH Grant MH61370 and NSF Grant BCS-9910756.

Further Reading

Dulany DE and Logan GD (1992) Special Issue: Views and varieties of automaticity. *The American Journal of Psychology*. **105**(2) (Summer).

Groeger JA (2000) *Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task*. Philadelphia, PA: Psychology Press.

Kirsner K, Speelman C, Maybery M *et al*. (1998) *Implicit and Explicit Mental Processes*. Mahwah, NJ: Lawrence Erlbaum Associates.

Logan GD (1988) Toward an instance theory of automatization. *Psychological Review* **95**: 492–527.

MacLeod CM (1991) Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin* **109**: 521–524.

MacLeod CM and Dunbar K (1988) Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition* **10**: 304–315.

Pashler H, Johnson JC and Ruthruff E (2000) Attention and performance. *Annual Review of Psychology* **52**: 629–651.

Schneider W and Shiffrin RM (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* **84**: 1–66.

Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* **84**: 127–190.

Wyer RS (1997) *The Automaticity of Everyday Life: Advances in Social Cognition*, vol. X. Mahwah, NJ: Lawrence Erlbaum Associates.

Behavior, Genetic Influences on

Intermediate article

M Frank Norman, University of Pennsylvania, Philadelphia, Pennsylvania, USA

CONTENTS

Studies of twins have established that individual differences in intelligence, personality, and psychopathology are strongly associated with genetic variations. Environmental variations are also important.

INTRODUCTION

The program of traditional (i.e. nonmolecular) behavioral genetics is to use correlations of relatives (especially twins) to make inferences about genetic and environmental contributions to behavioral variation. There are two major conclusions from such studies. First, a substantial percentage of behavioral variation seems to be due to genetic variation. Second, although environmental variation is also important, it appears that unique, idiosyncratic, experiences are more important for behavioral variation than experiences shared by twins. Both findings represent challenges to traditional social science, the first because social science does not emphasize biological factors, and the second because the environmental factors that are the focus of social science research (parents, homes, socioeconomic status, neighborhoods, and schools) are shared by twins and are thus seen by behavioral genetics as being in a sense less potent than social science supposes. The 'in a sense' qualification is necessary since twins obviously have unique experiences with common environmental factors. A 'strict' or 'cultured' home environment may be experienced in different ways by different children (Maccoby, 2000).

TWIN STUDIES

The following text provides a moderately detailed example of a behavioral genetic analysis that supports the first and second conclusions outlined above; it exposes the weaknesses as well as the strengths of such analyses. The analysis applies to studies that include twins reared apart as well as twins reared together. Thus these studies involve twins separated early in life by adoption. Analysis of twin studies without adopted subjects is considered later.

Fundamental Equations

Behavioral genetic analyses typically lead to equations expressing correlations in terms of genetic and environmental parameters. Consideration of 'identical' (monozygotic, MZ) and 'fraternal' (dizygotic, DZ) twins reared apart (A) and together (T) leads to four equations, the character of which varies somewhat with the precise assumptions that are in force. The following set is representative:

$$r_{\mathrm{MZA}} = h^2 + d^2 + i^2 \quad (1)$$

$$r_{\mathrm{DZA}} = 0.5h^2 + 0.25d^2 \quad (2)$$

$$r_{\mathrm{MZT}} = h^2 + d^2 + i^2 + c^2_{\mathrm{MZ}} \quad (3)$$

$$r_{\mathrm{DZT}} = 0.5h^2 + 0.25d^2 + c^2_{\mathrm{DZ}} \quad (4)$$

where h^2 is the additive genetic variance, d^2 is the dominance genetic variance, i^2 is the epistatic genetic variance, c^2_{MZ} is the common or shared environmental variance for MZ twins raised together, and c^2_{DZ} is the common or shared environmental variance for DZ twins raised together. These equations instantiate the model that is the focus of this section. In the present interpretation there are four population correlations (r) on the left, identified by the subscript abbreviations defined above, and five population parameters on the right. Our objective is to use these equations to develop estimators of the parameters based on sample correlations. Before doing this, however, we must make a long digression to define the parameters and discuss the kinds of assumptions that lead to the equations.

All of these variances are relative to the total phenotypic variance of the trait under discussion. Additive genetic variance is associated, fundamentally, with the sum of effects of the gene derived

from the mother and the gene derived from the father at a single genetic locus. *Dominance* is the interaction of these effects. However, it is clear that most if not all of the traits discussed here are influenced by many genetic loci, and our h^2 and d^2 correspond to the sums of additive and dominance effects over all contributory loci. Variance i^2 associated with multilocus interaction is epistatic. Dominance and epistasis are termed 'nonadditive effects'. Remarkably, it is possible to see the signature of such low-level effects in molar correlational data.

Additive genetic variance, h^2, is called 'heritability' or 'narrow heritability'. The sum of all three genetic variances is termed 'broad heritability' and denoted h_b^2. According to eqn [1], this quantity equals r_{MZA}.

Both genetic variation and environmental variation contribute to behavioral variation. Behavioral genetics usually treats environmental variation as the sum of two uncorrelated components. The first reflects parts of the environment such as parents, home, neighborhood, and schools that are shared by twins reared together. The second represents such things as unique friends and unique experiences that are not shared. The extent of the latter is indexed by yet other parameters, u_{MZ}^2 and u_{DZ}^2. It is assumed that only shared experiences contribute to psychological similarity. That is why only shared environmental variances appear on the right in eqns [3] and [4].

Obviously, different twins may have unshared experiences with common parents, houses, neighborhoods, and schools, and such experiences would be recorded in the unshared column, if it were really necessary to tally them up. Fortunately, it is not. Although effects of particular environmental variables can be studied (e.g. Caspi *et al*., 2000), behavioral genetics can and usually does make inferences about aggregate shared and unshared environmental effects without explicitly measuring their constituents.

Measurement error (m^2) is a source of differences between twins' test scores. It is sometimes implicitly included in unshared environmental variance, but it can be separated from other sources of uniqueness if the reliability of the test is known.

Newcomers to this subject may be puzzled by the 'squares' in these parameters. The unsquared genetic and common environmental parameters are correlations of a single individual's test score with an underlying genetic or environmental component. The squares arise in correlations between relatives because the 'path' between the relatives has two links: from one relative to a common genetic or environmental factor and then on to the other relative.

Now that we have defined the parameters that appear in eqns [1] to [4], we can begin to discuss the equations themselves. The 0.5 coefficient of h^2 in eqns [2] and [4] reflects the fact that DZ twins share half their genes, on average. The 0.25 coefficient of d^2 is explained by the observation that, at a single locus with only two alleles, 0.25 is the probability that two twins receive copies of exactly the same gene from both parents, and thus inherit precisely the same genotype at this locus.

Absence of covariances on the right-hand sides reflects the assumption that all sources of variation are uncorrelated. Also absent by assumption are genotype–environment interaction and effects of assortative mating. It is easy to think of concrete environmental factors (such as parental education, in the case of cognitive ability) that are definitely correlated with children's genotype. This brings out the point that the environmental factors in our equations should not be thought of as composites of concrete, directly measurable, environmental factors, but rather as akin to regression residuals, so that they are, by construction, uncorrelated with genetic 'main effects'. It is possible that the oversimplified model instantiated in our equations may allow interactions and correlations involving concrete environmental factors to be disproportionately absorbed into genetic terms.

Substantial assortative mating is known to occur for some of the traits to which we will apply the model (e.g. cognitive ability). The main defense that can be offered for omitting it, and for our other questionable assumptions, is that the model gives an adequate fit to most of the data to which we will apply it. This is not a strong justification, since MZA twin correlations are based on relatively small numbers of twin pairs, so the corresponding tests of goodness of fit are not very powerful. Thus, successful fits can be regarded only as showing that our assumptions are probably not flagrantly inappropriate.

Fitting the Model

One's first impulse is to replace population correlations in the equations by sample correlations, and solve the resulting linear equations for the quantities on the right. The solutions would then be estimators of the corresponding population parameters. However, this is not feasible because there are five unknowns but only four equations. Fortunately, it is possible to eliminate one unknown by

reparameterization, in such a way that our ability to test interesting hypotheses is not seriously compromised. This involves appending half of d^2 to h^2 and the other half to i^2, yielding the new parameters $h_d^2 = h^2 + 0.5d^2$ and $i_d^2 = i^2 + 0.5d^2$. These parameters can respectively be called 'intermediate heritability' (since it lies between broad and narrow heritability) and 'nonadditivity'. In terms of these quantities, eqns [1] to [4] take the following attractive form:

$$r_{\mathrm{MZA}} = h_d^2 + i_d^2 \quad (5)$$

$$r_{\mathrm{DZA}} = 0.5h_d^2 \quad (6)$$

$$r_{\mathrm{MZT}} = h_d^2 + i_d^2 + c_{\mathrm{MZ}}^2 \quad (7)$$

$$r_{\mathrm{DZT}} = 0.5h_d^2 + c_{\mathrm{DZ}}^2 \quad (8)$$

These equations are easily solved for the four parameters on the right in terms of the correlations on the left, and it is somewhat amusing to apply the resulting formulas to sample correlations to obtain parameter estimates. However, this 'equation-solving' approach is a sterile exercise, since it does not yield a test of the framework within which the estimation takes place. Consequently, one does not know whether the estimates are of any interest. Thus we will proceed directly to a more modern model-fitting approach, in which we explicitly test various hypotheses about the parameters, and implicitly test the underlying framework. The specific hypotheses to be tested are:

- equal environments assumption (EEA): $c_{\mathrm{MZ}}^2 = c_{\mathrm{DZ}}^2$
- no common environmental effects: $c_{\mathrm{MZ}}^2 = c_{\mathrm{DZ}}^2 = 0$
- no nonadditive effects (NNE): $i_d^2 = 0$
- no additive or dominance effects: $h_d^2 = 0$
- no genetic effects: $h_b^2 = 0$

Note that $h_b^2 = h_d^2 + i_d^2$, so the last hypothesis is equivalent to the conjunction of the two preceding it. Note also that, if EEA fails, one expects $c_{\mathrm{MZ}}^2 > c_{\mathrm{DZ}}^2$.

Any of these hypotheses reduces the number of parameters to less than four, so eqns [5] to [8] cannot be solved exactly. Instead, one accepts approximate equality in place of exact equality, and seeks parameter values that yield the best approximate solution. Different estimation methods correspond to different overall measures of the approximation error. We will use the error function and associated tests described by Loehlin (1989) in a slightly different context. The computer programs MX and LISREL provide other approaches to model fitting (Neale and Cardon, 1992).

INTELLIGENCE AND PERSONALITY

The Swedish Adoption/Twin Study of Aging

The Swedish Adoption/Twin Study of Aging (SATSA) is a study of elderly twins. The basic design and many results are summarized by Pedersen *et al*. (1991), and twin correlations and analyses of different types of variables appear in separate papers (Pedersen *et al*., 1992; Bergeman *et al*., 1993). Table 1 summarizes the results of a reanalysis of a few variables using the approach described above.

Cognitive ability is the first principal component of a number of tests of special abilities, and is thus a variant of intelligence quotient (IQ). The next seven variables in the table are standard dimensions of personality. The type A variable is derived from the famous Framingham type A scale, which measures the degree to which an individual is hard-driving, ambitious, and feels as if he or she is under pressure. Variables 'F-Cohesion' and 'F-Control' relate to the twins' recollections of the warmth and strictness of the families in which they were raised. Bear in mind that, for twins reared apart, these were different families, so these variables explore the possibility that twins' perceptions of family warmth and strictness may, to some extent, derive from the twin instead of from the family. Variable BMI is the body mass index, a measure of fatness. This is not a personality variable, but relates to eating habits, which are of great psychological interest in connection with eating disorders.

The χ^2 and p-values in the last two columns correspond to tests of the equal environments assumption. This assumption is rejected only for the last variable, 'F-Control'. This rejection confirms that the tests of EEA are not hopelessly insensitive. The parameter estimates given in the table are optimal assuming EEA, and are thus meaningful for all variables except 'F-Control', which will not be considered further. The c^2 parameter in Table 1 is the common value of c_{MZ}^2 and c_{DZ}^2 under EEA. Asterisks refer to tests of hypotheses that the corresponding parameters are zero, with one, two, and three asterisks indicating $p < 0.05$, 0.01, and 0.001, respectively. These tests are based on increments in the χ^2 goodness of fit index, beyond its value assuming just EEA. The $u^2 + m^2$ values cannot be tested for significance within this framework, but these values are large for all variables except cognitive ability and BMI.

Table 1. Components of variation for variables in the Swedish Adoption/Twin Study of Aging

Variable	h_d^2	i_d^2	h_b^2	c^2	$u^2 + m^2$	χ^2	p
Cognitive ability	0.55***	0.24	0.79***	0.00	0.21	0.54	0.46
Extraversion	−0.02	0.40***	0.38***	0.12	0.50	2.27	0.13
Neuroticism	0.49***	−0.16	0.33***	0.04	0.63	1.81	0.18
Openness	0.38**	0.11	0.49***	−0.01	0.52	1.24	0.26
Agreeableness	−0.06	0.21	0.15	0.26**	0.59	0.00	1.00
Conscientiousness	0.08	0.21	0.30**	0.12	0.58	2.65	0.10
Impulsivity	0.25*	0.18	0.44***	−0.01	0.57	0.56	0.46
Monotony avoidance	0.27*	−0.05	0.22*	0.03	0.75	0.06	0.80
Type A	0.34**	−0.07	0.27**	0.08	0.65	0.40	0.53
BMI (male)	0.41*	0.26	0.67***	0.08	0.25	0.63	0.43
BMI (female)	0.51***	0.14	0.65***	0.01	0.34	0.02	0.90
F-Cohesion	0.61***	−0.16	0.44***	0.14*	0.41	0.31	0.58
F-Control						7.89	0.01**

BMI, body mass index. Probability: *, $p = 0.05$; **, $p = 0.01$; ***, $p = 0.001$. See text for details of variables and parameters.

The table confirms the overall conclusions presented in the introduction, which are, in turn, consistent with those of the original SATSA papers. In 10 out of 12 cases, broad heritability is statistically significant whereas common environmentality is statistically insignificant. The reversal of this pattern for 'Agreeableness' indicates that this pattern is not forced by an artefact of the method. The very high estimate of broad heritability for cognitive ability is consonant with values obtained for IQ in other studies involving adult MZAs (Neisser *et al*., 1996).

Negative estimates arise because the χ^2 minimization routine varied $a = h_d^2$, $b = i_d^2$, etc., without restricting these quantities to positive values. In no case would the fit have been significantly worse if the negative quantity had been assumed to be zero, so the negative values should simply be regarded as negligible.

Of the 11 variables with significant h_b^2, 9 had h_d^2 but not i_d^2 significant, suggesting that genetic variation is mainly additive. One, 'Conscientiousness', had neither h_d^2 nor i_d^2 significant, leaving us little basis for inference about the distribution of broad heritability among its three components. Finally, 'Extraversion' had i_d^2 but not h_d^2 significant, suggesting that genetic variation is mainly epistatic. (Dominance and epistasis do not contribute to parent–child correlation, so they contradict the common misconception that genetic effects are always revealed by parent–child resemblance.)

We close this section with the results of analysis of a variable from the Minnesota study of twins reared apart (Bouchard *et al*., 1990), by the methods used in Table 1. There is a personality scale called 'well-being' that is, roughly speaking, a measure of happiness. For this scale, EEA was not rejected, $c^2 = -0.04$ is negligible, and $h_b^2 = 0.48$. Happiness in adults thus appears to be highly heritable (Lykken and Tellegen, 1996).

Studies Involving Only Twins Reared Together

Monozygotic twins reared apart are uniquely informative, but they are scarce. The analyses reported in Table 1 involved between 44 and 95 pairs of such twins. This leads to high variability of estimates and low power of tests. On the other hand, dizygotic and monozygotic twins reared together are plentiful, so it is not surprising that studies involving only twins reared together are the mainstay of traditional behavioral genetics. Unfortunately, the gain in precision from large samples is matched by a loss of generality due to the necessity of extra assumptions. For a feeling for the difficulties involved, consider

$$h^2_{\text{est}} = 2(r_{\text{MZT}} - r_{\text{DZT}}) \tag{9}$$

the traditional rough-and-ready estimator of heritability using sample correlations of twins reared together. For population correlations,

$$2(r_{\text{MZT}} - r_{\text{DZT}}) = h_b^2 + i_d^2 + 2\left(c^2_{\text{MZ}} - c^2_{\text{DZ}}\right) \tag{10}$$

as a consequence of eqns [7] and [8], so h^2_{est} will tend to overestimate h_b^2 when EEA fails or nonadditive effects are present.

There is a companion formula, $2r_{DZT} - r_{MZT}$, that is traditionally used to estimate common environmental variance. According to eqns [7] and [8],

$$2r_{DZT} - r_{MZT} = c_{DZ}^2 - i_d^2 - \left(c_{MZ}^2 - c_{DZ}^2\right) \qquad (11)$$

so the companion formula has a tendency to underestimate c_{DZ}^2 when EEA or NNE fails. Sizeable negative values of the companion formula strongly suggest that use of h_{est}^2 is inappropriate, but one sees many instances in the literature where this warning has not been heeded.

Modern behavioral genetics uses model-fitting techniques in place of h_{est}^2 (see, for example, Loehlin, 1992), but model-fitting analyses, like their less sophisticated predecessors, typically assume EEA and sometimes also NNE.

MENTAL ILLNESS

A characteristic feature of many studies of mental illness is a categorical 'sick versus well' classification of each patient's condition. In place of twin correlation for (say) IQ, we have concordance for (say) schizophrenia, estimating the likelihood that a second twin is affected given that the first twin is affected. Concordances carry some of the same intuitions as correlations, but they are not the kinds of correlations to which behavioral genetic theory can be directly applied.

The liability threshold model provides a simple bridge from concordances to behavioral genetics. According to this model, there is a normally distributed liability, L, to the condition under consideration, and the condition is manifested if and only if L exceeds a threshold parameter T. Assuming that the distribution of liability has a mean of zero and a standard deviation of 1, one can estimate T from the prevalence of the condition.

Assuming a bivariate normal distribution of twins' liabilities, computer programs like PRELIS can estimate the liability correlations corresponding to concordances. These are sometimes referred to as *tetrachoric* correlations, and their variances are different from those of ordinary, Pearson, correlations. Taking account of this, behavior genetic analyses of liability correlations can be done in a manner analogous to behavior genetic analyses of Pearson correlations. In particular, there are older studies that calculate h_{est}^2 from liability correlations.

Modern studies apply model-fitting techniques via LISREL or MX to twins reared together, usually assuming EEA and, often, NNE. Substantial differences between MZ and DZ liability correlations are invariably associated with substantial heritability in these analyses. Such differences have been found for autism, attention deficit hyperactivity disorder, depression, bipolar disorder, and schizophrenia (McGuffin and Martin, 1999).

There is great uncertainty concerning the extent of possible genetic contributions to bulimia and anorexia (Fairburn *et al*., 1999). Though it is not a mental illness, it is interesting to note that McGue and Lykken (1992) have reported a heritability estimate of 0.525 for liability to divorce.

THE SEARCH FOR DEFINITE GENES

Traditional behavioral genetics operates at a tremendous level of abstraction (some might call it vagueness). It provides information only about aggregate quantities, though these quantities are of considerable interest. Some results are available showing behavioral effects of specific genes. In the future, one expects increasing emphasis on studies seeking to demonstrate such effects.

Specific gene effects on behavior vary greatly in size. One imagines that many genes affect each of the personality and ability dimensions, as well as most of the psychopathologies, considered above. If many genes contribute to a dimension, most of these contributions will be small and thus relatively hard to detect.

There are a number of cases where variation at a single genetic locus has drastic effects on the organism, including mental retardation. An especially interesting case of such a single major gene effect is phenylketonuria, in which neither of the alleles at a certain locus on chromosome 12 supports production of an enzyme that breaks down phenylalanine. The consequent build-up of the latter substance in the brain causes mental retardation. Such build-up and retardation can be prevented by a special diet. Since the genetic deficiency leads to a behavioral deficiency only in the presence of a certain environment (normal diet), this is an example of a genotype–environment interaction. At a more basic level, phenylketonuria illustrates that genetic involvement in a behavioral deficiency does not imply that the deficiency is immutable. One of the major thrusts of modern medical science is to discover environmental compensations (medicines) for genetic conditions.

Definite genes, or small chromosomal regions, have been implicated with various degrees of certainty in the following conditions: early-onset and late-onset Alzheimer disease, autism, bipolar disorder, dyslexia, and schizophrenia (McGuffin and Martin, 1999; Owen and Cardno, 1999). For example, the apolipoprotein E ε4 allele appears to be associated with late-onset Alzheimer disease.

Information from the human genome project will doubtless make great contributions to this rapidly developing area.

EVOLUTION

Although we have considered genetic variability in a number of psychologically relevant dimensions, we have not considered genetic variability in fitness or reproductive success, the additive component of which directly controls evolution. In so far as certain traits are prevalent today, it is natural to think that they might somehow have conferred enhanced fitness long ago. Such evolutionary speculation is among the most powerful heuristics in biological science in general and evolutionary psychology in particular. Although the disciplines that incorporate it are flourishing, it is important to be aware that inference from present predominance to past superior fitness is fallible. This can be shown by examples involving variation controlled by a single genetic locus with only two alleles, A_1 and A_2. Assuming random mating, the heterozygotic genotype A_1A_2 has relative frequency $2p(1-p)$, where p is the relative frequency of the A_1 allele. Thus, regardless of its fitness, the relative frequency of A_1A_2 cannot exceed 1/2. In fact, it is not difficult to construct examples where the homozygote A_1A_1 predominates after many generations of evolution, even though the heterozygote is the fittest genotype.

Variants of such simple genetic examples can be constructed within the frameworks of W. D. Hamilton's kin selection theory and J. Maynard Smith's (1982) evolutionary game theory, two of the pillars of evolutionary psychology (Norman, 1981). A frequently cited part of Hamilton's theory suggests that one can predict the evolutionary fate of altruistic and selfish genotypes just by examining their inclusive fitnesses, but it turns out that superior inclusive fitness is not sufficient for asymptotic predominance of, say, an altruistic genotype, if that genotype is heterozygotic. Similarly, the standard version of Maynard Smith's criterion for an evolutionarily stable strategy (ESS) depends entirely on fitness 'payoffs', irrespective of genetic structure. However, a behavioral strategy exhibited by a heterozygote cannot be evolutionarily stable, regardless of associated payoffs, since a population composed entirely of heterozygotes will give rise to a mixed population in the next generation.

These examples represent relatively minor blemishes on these theories, but the examples do suggest that the theories should be applied with caution.

References

Bergeman CS, Chipuer HM, Plomin R *et al*. (1993) Genetic and environmental effects on openness to experience, agreeableness, and conscientiousness: an adoption/twin study. *Journal of Personality* **61**: 159–179.

Bouchard TJ, Lykken DT, McGue M, Segal NL and Tellegen A (1990) Sources of human psychological differences: the Minnesota study of twins reared apart. *Science* **250**: 223–228.

Caspi A, Taylor A, Moffitt TE and Plomin R (2000) Neighborhood deprivation affects children's mental health: environmental risks identified in a genetic design. *Psychological Science* **11**: 338–342.

Fairburn CG, Cowen PJ and Harrison PJ (1999) Twin studies and the etiology of eating disorders. *International Journal of Eating Disorders* **26**: 349–358.

Loehlin JC (1989) Partitioning environmental and genetic contributions to behavioral development. *American Psychologist* **44**: 1285–1292.

Loehlin JC (1992) *Genes and Environment in Personality Development*. Thousand Oaks, CA: Sage.

Lykken D and Tellegen A (1996) Happiness is a stochastic phenomenon. *Psychological Science* **7**: 186–189.

Maccoby EE (2000) Parenting and its effects on children: on reading and misreading behavior genetics. *Annual Review of Psychology* **51**: 1–27.

Maynard Smith J (1982) *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press.

McGue M and Lykken DT (1992) Genetic influence on risk of divorce. *Psychological Science* **3**: 368–373.

McGuffin P and Martin N (1999) Science, medicine, and the future: behaviour and genes. *British Medical Journal* **319**: 37–40.

Neale MC and Cardon LR (1992) *Methodology for Genetic Studies of Twins and Families*. Dordrecht, Netherlands: Kluwer.

Neisser U, Boodoo G, Bouchard TJ *et al*. (1996) Intelligence: knowns and unknowns. *American Psychologist* **51**: 77–101.

Norman MF (1981) Sociobiological variations on a Mendelian theme. In: S Grossberg (ed.) *Mathematical Psychology and Psychophysiology*, pp. l87–l96. Providence, RI: American Mathematical Society.

Owen MJ and Cardno AG (1999) Psychiatric genetics: progress, problems, and potential. *Lancet* **354**(suppl. 1): 11–14.

Pedersen NL, McClearn GE, Plomin R *et al*. (1991) The Swedish Adoption Twin Study of Aging: an update. *Acta Geneticae Medicae et Gemellologiae: Twin Research* **40**: 7–20.

Pedersen NL, Plomin R, Nesselroade JR and McClearn GE (1992) A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychological Science* **3**: 346–353.

Further Reading

Bailey MJ (1998) Can behavior genetics contribute to evolutionary behavioral science? In: Crawford C and

Krebs DL (eds) *Handbook of Evolutionary Psychology: Ideas, Issues, and Applications*, pp. 211–233. Mahwah, NJ: Lawrence Erlbaum.

Buss DM (1999) *Evolutionary Psychology: The New Science of the Mind*. Boston, MA: Allyn & Bacon.

Kendler KS (1993) Twin studies of psychiatric illness: current status and future directions. *Archives of General Psychiatry* **50**: 905–915.

Lykken DT, McGue M, Tellegen A and Bouchard TJ (1992) Emergenesis: genetic traits that may not run in families. *American Psychologist* **47**: 1565–1577.

Plomin R (1994) *Genetics and Experience: The Interplay Between Nature and Nurture*. Thousand Oaks, CA: Sage.

Plomin R, DeFries JC, McClearn GE and McGuffin P (2001) *Behavioral Genetics*, 4th edn. New York, NY: Worth.

Reiss D, Neiderhiser JM, Hetherington EM and Plomin R (2000) *The Relationship Code: Deciphering Genetic and Social Influences on Adolescent Development*. Cambridge, MA: Harvard University Press.

Rowe DC (1994) *The Limits of Family Influence: Genes, Experience, and Behavior*. New York, NY: Guilford Press.

Wahlsten D (1999) Single-gene influences on brain and behavior. *Annual Review of Psychology* **50**: 599–624.

Categorical Perception

Intermediate article

Stevan Harnad, University of Quebec, Montreal, Canada

CONTENTS

CATEGORIES: CATEGORICAL AND CONTINUOUS

A category, or kind, is a set of things. Membership in the category may be (1) all-or-none, as with 'bird': something either is a bird or it isn't a bird; a penguin is 100 percent bird, a platypus is 100 percent not-bird. In this case we would call the category 'categorical'. Or membership might be (2) a matter of degree, as with 'big': some things are 'more big' and some things are 'less big'. In this case the category is 'continuous' (or rather, degree of membership corresponds to some point along a continuum). There are range or context effects as well: elephants are relatively big in the context of animals, relatively small in the context of bodies in general, if we include planets.

Many categories, however, particularly concrete sensorimotor categories (things we can see and touch), are a mixture of the two: categorical at an everyday level of magnification but continuous at a more microscopic level. Color categories are good examples: central reds are clearly reds, and not shades of yellow. But in the orange region of the spectral continuum, red/yellow is a matter of degree; context and contrast effects can also move these regions around somewhat. Perhaps even with 'bird', an artist or genetic engineer could design intermediate cases in which their 'birdness' was only a matter of degree.

RESOLVING THE 'BLOOMING, BUZZING CONFUSION'

Categories are important because they determine how we see and act upon the world. As William James noted, we do not see a continuum of 'blooming, buzzing confusion' but an orderly world of discrete objects. Some of these categories are 'prepared' in advance by evolution: the frog's brain is born already able to detect 'flies'; it needs only normal exposure rather than any special learning in order to recognize and catch them. Humans have such innate category-detectors too: the human face itself is probably an example. So too are our basic color categories; although, according to the 'Whorf hypothesis' (Whorf 1956; also called the 'linguistic relativity' hypothesis) colors are determined by how our culture and language happens to subdivide the spectrum (we will return to this).

But if one opens up a dictionary at random and picks out a content word, it is probable that it names a category we have learned to detect, rather than one that our brains were innately prepared by evolution to detect. The generic human face may be an innate category for us, but surely all the specific people we know and can name are not. 'Red' and 'yellow' may be inborn, but what about 'scarlet' and 'crimson'?

THE MOTOR THEORY OF SPEECH PERCEPTION

And what about the very building blocks of the language we use to name categories. Are our speech sounds – ba, da, ga – innate or learned? The first question we must answer about them is whether they are categories at all, or merely arbitrary points along a continuum. It turns out that if one analyzes the sound spectrogram of ba and pa, for example, both turn out to lie along an acoustic continuum called 'voice-onset-time'. With a technique similar to the one used in morphing visual images continuously into one another, it is possible to 'morph' a ba gradually into a pa and beyond by gradually increasing the voicing parameter.

Liberman *et al*. (1957) reported that when people listen to sounds that vary along the voicing

continuum, they hear only ba's and pa's, nothing in between. This effect – in which perception jumps abruptly from one category to another at a certain point along a continuum, instead of changing gradually – he called 'categorical perception' (CP). He suggested that CP was unique to speech, that it made speech special, and that its explanation lay in the anatomy of speech production. This came to be called 'the motor theory of speech perception'.

According to the (now abondoned) motor theory, the reason we perceive an abrupt change between ba and pa is that the way we hear speech sounds is mediated by the way we produce them when we speak. What is varying along this continuum is voice-onset-time: the 'b' in ba is voiced and the 'p' in pa is not. Unlike the synthetic morphing apparatus, our natural vocal apparatus is not capable of producing anything in between. So when I hear a sound from the voicing continuum, my brain perceives it by trying to match it with what it would have had to do to produce it. Since the only thing I can produce is ba or pa, I will perceive any of the synthetic stimuli along the continuum as either ba or pa, whichever it is closer to. A similar CP effect is found with ba/da; these too lie along a continuum acoustically, but vocally ba is formed with the two lips, da with the tip of the tongue and the hard palate, and our anatomy does not allow any intermediates.

The motor theory of speech perception explained how speech was special and why speech sounds are perceived categorically: sensory perception is mediated by motor production. Wherever production is categorical, perception will be categorical; where production is continuous, perception will be continuous. And indeed vowel categories like a/u were found to be much less categorical than ba/pa or ba/da (less categorical, but not altogether continuous either; we will return to this).

ACQUIRED DISTINCTIVENESS

If motor production mediates sensory perception, then one assumes that this CP effect is a result of learning to produce speech. Early research, however, found that infants show speech CP before they begin to speak. Perhaps, then, it is an innate effect, evolved to 'prepare' us to learn to speak. But Kuhl found that chinchillas also show 'speech CP' even though they never learn to speak, and presumably did not evolve to do so (Kuhl, 1987). Lane (1965) went on to show that CP effects can be induced by learning alone, with a purely sensory (visual) continuum in which there is no motor production discontinuity to mediate the perceptual discontinuity. He concluded that speech CP is not special after all but merely a special case of Lawrence's (1950) classic demonstration that stimuli to which you learn to make a different response become more distinctive, and stimuli to which you learn to make the same response become more similar.

It also became clear that CP was not quite the all-or-none effect Liberman had originally thought it was: it is not that all pa's are indistinguishable and all ba's are indistinguishable. We can hear the differences, just as we can see the differences between different shades of red. It is just that the within-category differences (pa1/pa2 or red1/red2) seem to be much smaller than the between-category differences (pa2/ba1 or red2/yellow1), even when the size of the underlying physical differences (voicing, wavelength) are the same.

WITHIN-CATEGORY COMPRESSION AND BETWEEN-CATEGORY SEPARATION

This evolved into the contemporary definition of CP that is no longer peculiar to speech or dependent on the motor theory: CP occurs whenever perceived within-category differences are compressed and/or between-category differences are separated, relative to some baseline. The baseline might be the actual size of the physical differences involved, or, in the case of learned CP, it might be the perceived similarity or discriminability within and between categories before the categories were learned.

A typical learned CP experiment would be the following. A set of stimuli is tested for pairwise similarity or discriminability. In the case of similarity, multidimensional scaling might be used to scale the rated pairwise similarity of the set of stimuli. In the case of discriminability, same/different judgments and signal detection analysis might be used to estimate the discriminability of a set of stimuli. Then the same subjects or a different set are trained, using trial and error and corrective feedback, to sort the stimuli into two or more categories. After the categorization has been learned, similarity or discriminability are tested and compared against the untrained data. If there is significant within-category compression and/or between-category separation, this is operationally defined as CP (Harnad, 1987).

THE WHORF HYPOTHESIS

We can now return both to the 'Whorf hypothesis' and the 'weaker' CP for vowels: according to the

Whorf hypothesis (of which Lawrence's acquired similarity/distinctiveness effects would simply be a special case), colors are perceived categorically only because they happen to be named categorically; our subdivisions of the spectrum are arbitrary, learned, and vary across cultures and languages (Whorf, 1964). But Berlin and Kay (1969) showed that this was not so: not only do most cultures and languages subdivide and name the color spectrum the same way, but even for those who don't, the regions of compression and separation are the same. We all see blues as more alike and greens as more alike, with a fuzzy boundary in between, whether or not we have named the difference. So there is no Whorfian learning effect with colors; or is there?

EVOLVED CP

First, let us go back to vowels. The signature of CP is within-category compression and/or between-category separation. The size of the CP effect is merely a scaling factor; it is this compression/separation 'accordion effect' that is CP's distinctive feature. In this respect, the 'weaker' CP effect for vowels, whose motor production is continuous rather than categorical but whose perception is by this criterion categorical, is every bit as much of a CP effect as the ba/pa and ba/da effects. But, as with colors, it looks as if the effect is an innate one. Our sensory category detectors for color and speech sounds are born already 'biased' by evolution: the perceived color and speech sound spectrum is born 'warped' with these compression/separations.

LEARNED CP

Is that all there is to it? Apparently not. There are still the Lane/Lawrence demonstrations, lately replicated and extended by Goldstone (1994, 1999), that CP can be induced by learning alone. And there are also the countless categories cataloged in our dictionaries that could not possibly be inborn (though nativist theorists such as Fodor (e.g. 1983) have sometimes seemed to suggest that all of our categories are inborn). There are even recent demonstrations that although the primary color and speech categories are probably inborn, their boundaries can be modified or even lost as a result of learning, and weaker secondary boundaries can be generated by learning alone (Roberson *et al*., 2000).

Perhaps CP performs some useful function in categorization? In the case of innate CP, our categorically biased sensory detectors pick out their prepared color and speech sound categories far more readily and reliably than if our perception had been continuous. Could something similar be the case for our repertoire of learned categories too?

COMPUTATIONAL AND NEURAL MODELS OF CP

Computational modeling (Tijsseling and Harnad, 1997; Damper and Harnard 2000) has shown that many types of category-learning mechanisms (e.g. both backpropagation and competitive networks) display CP-like effects. In backpropagation nets, the hidden-unit activation patterns that 'represent' an input build up within-category compression and between-category separation as they learn; other kinds of net display similar effects. CP seems to be a means to an end: inputs that differ among themselves are 'compressed' onto similar internal representations if they must all generate the same output; and they become more separate if they must generate different outputs. The network's 'bias' is what filters inputs onto their correct output category. The nets accomplish this by selectively detecting (after much trial and error, guided by error-correcting feedback) the invariant features that are shared by the members of the same category and that reliably distinguish them from members of different categories; the nets learn to treat all other variation as being irrelevant to the categorization.

Very little is known yet about the brain mechanisms of category perception and learning. The computational models are really causal hypotheses about what the brain might be doing. Neural data provide correlates of CP and of learning (Sharma and Dorman, 1999). Differences between event-related potentials recorded from the brain have been found to be correlated with differences in the perceived category of the stimulus viewed by the subject. Neural imaging studies have shown that these effects are localized and even lateralized to certain brain regions in subjects who have successfully learned the category, and are absent in subjects who have not (Seger *et al*., 2000).

LANGUAGE-INDUCED CP

Both innate and learned CP are sensorimotor effects. The compression/separation biases are sensorimotor biases, and presumably had sensorimotor origins, either during the sensorimotor life history of the organism, in the case of learned CP, or during the sensorimotor life history of the

species, in the case of innate CP. The neural net I/O models are also compatible with this fact: their I/O biases derive from their I/O history. But when we look at our repertoire of categories in a dictionary, it is highly unlikely that many of them had a direct sensorimotor history during our lifetimes, and even less likely in our ancestors' lifetimes. How many of us have seen unicorns in real life? We have seen pictures of them, but what had those who first drew those pictures seen? And what about categories I cannot draw or see (or taste or touch)? What about the most abstract categories, such as goodness and truth?

Some of our categories must originate from a source other than direct sensorimotor experience, and here we return to language and the Whorf hypothesis. Can categories, and their accompanying CP, be acquired through language alone? Again, there are some neural net simulation results suggesting that, once a set of category names has been 'grounded' through direct sensorimotor experience, they can be combined into Boolean combinations (man = male and human) and into still higher-order combinations (bachelor = unmarried and man). These combinations not only pick out the more abstract, higher-order categories in much the way the direct sensorimotor detectors do, but also inherit their CP effects, as well as generating some of their own. Bachelor inherits the compression/separation of unmarried and man, and adds a layer of separation/compression of its own (Cangelosi *et al*., 2000; Cangelosi and Harnad 2001).

These language-induced CP-effects remain to be directly demonstrated in human subjects; so far only learned and innate sensorimotor CP have been demonstrated (Pevtzow and Harnad 1997; Livingston *et al*., 1998). The latter show the Whorfian power of naming and categorization, in warping our perception of the world. That is enough to rehabilitate the Whorf hypothesis from its apparent failure on color terms (and perhaps also from its apparent failure on eskimo snow terms – see Pullum, 1989); but to show that it is a full-blown language effect, and not merely a vocabulary effect, it will have to be shown that our perception of the world can also be warped, not just by how things are named but by what we are told about them.

References

Berlin B and Kay P (1969) *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Damper RI and Harnad S (2000) Neural network modeling of categorical perception. *Perception and Psychophysics* **62**(4): 843–867. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/20/index.html.]

Eimas PD, Siqueland ER, Jusczyk PW and Vigorito J (1971) Speech perception in infants. *Science* **171**: 303–6.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Goldstone RL (1994) Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General* **123**: 178–200.

Goldstone RL (1999) Similarity. In: Wilson RA and Keil FC (eds) *MIT Encylopedia of the Cognitive Sciences*, pp. 763–765. Cambridge, MA: MIT Press.

Kuhl PK (1987) The special-mechanisms debate in speech perception: nonhuman species and nonspeech signals. In: Harnad S (ed.) *Categorical Perception: the Groundwork of Cognition*. New York, NY: Cambridge University Press.

Lane H (1965) The motor theory of speech perception: a critical review. *Psychological Review* **72**: 275–309.

Lawrence DH (1950) Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology* **40**: 175–188.

Liberman AM, Harris KS, Hoffman HS and Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**: 358–368.

Pullum GK (1989) The great eskimo vocabulary hoax. *Natural Language and Linguistic Theory* **7**: 275–281.

Seger CA, Poldrack RA, Prabhakaran V, Zhao M, Glover GH and Gabrieli JDE (2000) Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia* **38**(9): 1316–1324.

Sharma A and Dorman MF (1999) Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America* **106**(2): 1078–1083.

Tijsseling A and Harnad S (1997) Warping similarity space in category learning by backprop nets. In: Ramscar M, Hahn U, Cambouropolos E and Pain H (eds) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, pp. 263–269. Edinburgh, UK: Department of Artificial Intelligence, Edinburgh University. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/08/index.html.]

Whorf BL (1964) *Language, Thought and Reality*. Cambridge MA: MIT Press.

Further Reading

Andrews J, Livingston K and Harnad S (1998) Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**(3): 732–753.

Belpaeme T (2002) *Factors Influencing the Origins of Colour Categories*. PhD thesis. Artificial Intelligence Lab, Free University of Brussels. [http://arti.vub.ac.be/~tony/phd/index.htm.]

Bimler D and Kirkland J (2001) Categorical perception of facial expressions of emotion: evidence from multidimensional scaling. *Cognition and Emotion* **15**: 633–658.

Calder AJ, Young AW, Perrett DI, Etcoff NL and Rowland D (1996) Categorical perception of morphed facial expressions. *Visual Cognition* **3**: 81–117.

Campanella S, Quinet O, Bruyer R, Crommelinck M and Guerit JM (2002) Categorical perception of happiness and fear facial expressions : an ERP study. *Journal of Cognitive Neuroscience* **14**(2): 210–227.

Cangelosi A, Greco A and Harnad S (2000) From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science* **12**(2): 143–162. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/47/index.html.]

Cangelosi A and Harnad S (2001) The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. *Evolution of Communication* **4**(1): 117–142. [http://cogprints.soton.ac.uk/documents/disk0/00/00/20/36/index.html.]

Goldstone RL, Lippa Y and Shiffrin RM (2001) Altering object representations through category learning. *Cognition* **78**: 27–43.

Guest S and Van Laar D (2000) The structure of colour naming space. *Vision Research* **40**: 723–734.

Harnad S (ed.) (1987) *Categorical Perception: the Groundwork of Cognition*. New York, NY: Cambridge University Press. [http://cogprints.soton.ac.uk/documents/disk0/00/00/15/71/index.html.]

Harnad S (1990) The symbol grounding problem. *Physica D* **42**: 335–346. [http://cogprints.soton.ac.uk/documents/disk0/00/00/06/15/index.html.]

Kotsoni E, de Haan M and Johnson MH (2001) Categorical perception of facial expressions by 7-month-old infants. *Perception* **30**: 1115–1125.

Pevtzow R and Harnad S (1997) Warping similarity space in category learning by human subjects: the role of task difficulty. In: Ramscar M, Hahn U, Cambouropolos E and Pain H (eds) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, pp. 189–195. Edinburgh, UK: Department of Artificial Intelligence, Edinburgh University. [http://cogprints.soton.ac.uk/documents/disk0/00/00/16/07/index.html.]

Roberson D, Davies I and Davidoff J (2000) Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* **129**: 369–398.

Rossion B, Schiltz C, Robaye L, Pirenne D and Crommelinck M (2001) How does the brain discriminate familiar and unfamiliar faces? A PET study of face categorical perception. *Journal of Cognitive Neuroscience* **13**: 1019–1034.

Schyns PG, Goldstone RL and Thibaut J (1998) Development of features in object concepts. *Behavioral and Brain Sciences* **21**: 1–54.

Steels L (2001) Language games for autonomous robots. *IEEE Intelligent Systems* **16**(5): 16–22.

Steels L and Kaplan F (1999) Bootstrapping grounded word semantics. In: Briscoe T (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge, UK: Cambridge University Press.

Categorization, Development of

Intermediate article

Jean M Mandler, University of California, San Diego, California, USA

CONTENTS

Categorization occurs when individual items that can be discriminated from one another are considered to be the same or belong together because they are alike in some way.

INTRODUCTION

Categorization is an automatic part of perceiving and conceptualizing. Whether perceiving something or thinking about it, the mind tends to subsume individual items into larger classes. The term 'categorization' is also applied to a task that individuals perform. Such tasks require deliberate choice as to the basis of similarity that is used to do the grouping: for example, in a set of red and green dogs and cats, one might choose to group them on the basis of similar color or similar kind. One can categorize in this way on many bases ranging from taxonomic categories to categories made up on the spot.

Behavior on deliberate categorization tasks changes dramatically in the early years, but the reasons for the changes are complex and varied. Children's understanding of the instructions, their construal of what is wanted, and the salience of different aspects of objects vary depending on school and home experiences. These variables influence the developmental change that has been reported on these tasks, from the preference of young children to categorize objects by their overall similarity to the preference of older children and adults to concentrate on dimensions of the objects such as size or shape (Smith and Kemler, 1978). Schooling is also implicated in the tendency of children and young adults to categorize taxonomically on these tasks, compared with preschool children and older adults, who are more likely to categorize objects by their functions or the events in which they take part (Smiley and Brown, 1979).

There is a good deal of research on categorization in infancy using instructionless tasks that rely on infants' spontaneous categorizing behavior, and that therefore provide information about the earliest bases of categorization. A typical experimental procedure familiarizes infants with one category and then measures preferential looking or examining to a member of another category. This article concentrates on the infancy period when the first perceptual and conceptual categories are formed.

MODELS OF CATEGORIZATION

One model of categorization posits that encountering similar stimuli leads to the formation of a prototype of the presented items. For example, upon seeing many dogs, one extracts the central tendency or principal components (as in factor analysis) of 'dogginess' in terms of physical features and presumably in terms of behavior as well. Even 3-month-old babies extract a perceptual prototype from a set of similar stimuli. For example, if shown a number of squares or triangles, each of which is somewhat distorted in shape, they will treat a regular square or triangle as more familiar than any of the distorted figures they have seen (Bomba and Siqueland, 1983).

Another model of categorization is that we accumulate many instances and compare new instances with this accumulated knowledge base. This procedure enables us to recognize a prototypical instance of a category because it is most similar to the largest number of examples. Although much published research differentiates these two theories of categorization, there is little to choose between them at a behavioral level and in terms of development almost no relevant data to decide between them.

PERCEPTUAL VERSUS CONCEPTUAL CATEGORIES

People categorize perceptually ('this dog looks like that dog') and conceptually ('courage and honesty

are both virtues'). Both kinds of categorization can be characterized as involving similarity (e.g. in shape or value system), and so some researchers describe both kinds of categorization in terms of a single process of computing similarity (either in terms of similarity to a prototype or similarity of exemplars). However, it is not always easy to specify the similarities among exemplars of conceptual categories such as virtues (exactly how are courage and kindness similar?), so some researchers find it preferable to say that conceptual categories are based on a definition or rule. On this view, then, there are two different processes involved in categorization: assessing similarity, and applying a rule. (*See* **Concept Learning**; **Conceptual Change**)

It seems plausible that current debate on this topic reflects differences in the way that perceptual and conceptual categories are formed. The items being categorized differ greatly in the two cases, and there appear to be a number of differences in their formation, suggesting that a two-process account may be necessary. Perceptual categories are almost certainly based on similarity, but many conceptual categories may be more rule-like in nature. For example, there is evidence that the perceptual categorization of faces as male or female is an automatic part of the operation of the visual system and is implicit in nature. We have all been able to categorize male and female faces since infancy, but we cannot say accurately what the differences are. Conceptual categories, on the other hand, are based on notions that are open to our inspection and we can often give a rule for why two things belong to different classes. Even a category of objects such as animals can be described as rule-based. For example, for infants an animal might be anything that moves itself and acts on other objects, regardless of what it looks like.

Perceptual categorization happens automatically and does not require conscious attention. The process, which enables generalization from prior instances of a category to new instances, is a fundamental aspect of perception and is present very early in life. For example, using a familiarization/preferential-looking technique, 3-month-old babies have been found not only to extract perceptual prototypes but also to categorize realistic pictures of cats after only a few exposures and differentiate them from pictures of dogs (Quinn *et al*., 1993). If the stimuli are more variable, the process takes longer; hence, it takes more exposure to categorize pictures of dogs than of cats.

Unfortunately, most of the research on perceptual categorization of realistic stimuli at this young age has been conducted with pictures of animal kinds, which may not be representative of category formation in general. It has also tended to contrast objects at what is known as the basic level (Rosch *et al*., 1976). For example, dogs have been contrasted with cats, as opposed to a subordinate comparison of poodles contrasted with terriers, or a superordinate comparison of animals contrasted with vehicles. However, there are some data indicating that babies at 3 months and 6 months of age can also categorize pictures of tables, differentiating them from chairs. They also discriminate pictures of mammals from vehicles at this age, although more trials are required, presumably because of the greater perceptual variability of these global categories. (Superordinate categories are usually termed 'global categories' in the infancy literature, because infants typically make few conceptual subdivisions within such large classes.)

By about 6 months, as infants begin to manipulate objects, a familiarization/preferential-examining technique is used with models of objects to assess categorization. Seven-month-old infants categorize animals as different from vehicles. By 9 months, they categorize animals as different from furniture as well, and by 11 months categorize both these classes as different from plants and utensils. The range of perceptual variation in these global categories is much greater than for basic-level categories and even for four-legged mammals (compare dogs with horses versus dogs with birds). Nevertheless, 9-month-old infants have no difficulty in differentiating models of birds with outstretched wings from airplanes in spite of the high degree of similarity in appearance of the two classes. Because of the apparent lack of difficulty posed by within-category variability in the case of animals, and between-category similarity in the case of birds and airplanes, it has been suggested that perceptual similarity alone cannot account for this kind of categorization, and more conceptual information, as opposed to the physical appearance of the exemplars, is being used (Mandler, 2000).

GLOBAL VERSUS BASIC-LEVEL CATEGORIES

The infants who at age 7–11 months categorize animals, vehicles, and furniture as different kinds do not categorize dogs versus cats or tables versus chairs on object examination tests. Nevertheless, infants aged 3–6 months who categorize pictures of dogs versus cats and tables versus chairs have more difficulty categorizing furniture versus

vehicles. Although the dissociations shown on these various categorization tests do not perfectly differentiate global from basic-level categories, there are enough differences in categorizing behavior to make it likely that different processes are being measured when infants categorize objects rather than pictures. Eventually, of course, pictures and objects produce similar data, but this is not the case in infancy, which makes the study of categorization at this period of life particularly informative.

Perceptual processes are capable of categorizing differently shaped pictures, even if the patterns are novel or meaningless. Hence, 3-month-old infants who have had no experience with dogs and cats nevertheless quickly learn to tell pictures of them apart. However, when perceptual variation is great, as is often the case for superordinate (global) categories, it may not be possible for the perceptual system to extract the principal components of the displayed patterns, and so this kind of perceptual categorization fails.

In contrast to the automatic perceptual processes involved in perceptual categorization, global categorization of animals, vehicles, furniture, and plants appears to be a response to a conceptual difference between these very different classes (Mandler, 2000). It is more selective in nature (and hence more 'rule-like'), acting on those aspects of events that have attracted infants' attention and interest. Young infants are particularly interested in differences in the way that animals and inanimate things move and the kinds of interactions in which they engage. Even relatively superficial analysis of the activities of animate and inanimate objects results in concepts that differentiate these classes. For example, such analysis leads to categorization of animals as things that start themselves and interact with other objects from a distance in a contingent fashion, and categorization of inanimate objects as things that either do not move at all or, if they do, do not start themselves and do not act on other objects from a distance. In the early months of life infants may not be able to conceptualize global categories more finely than this. It is relatively easy to differentiate the behavior of an animal from a piece of furniture, but it requires considerably more analysis to conceptualize how a dog differs from a cat or a car from a truck. (*See* **Infant Cognition**; **Object Perception, Development of**)

TIME COURSE OF CATEGORIZATION

Perceptual categorization begins at or near birth. To date, all the research indicating conceptual categorization has involved manipulating objects, which means that no data are available for infants younger than about 6–7 months. One piece of evidence for the conceptual nature of the global categories found in the second 6-month period of life is that it is these categories that are used for purposes of inductive inference. By 9–11 months of age, infants generalize the behavior of an observed animal, such as a dog drinking, to all animals including fish. They generalize using a key on a car to all vehicles, including airplanes (Mandler and McDonough, 1996). At the same time they do not differentiate between cups and pans as appropriate containers from which to drink, or between beds and bathtubs as appropriate places to sleep. The initially global categories become gradually differentiated during the second year, with current evidence suggesting that (at least in American culture) artefacts are differentiated earlier than natural kinds. The vocabulary that children are learning contributes to this development (Waxman, 1999). By the age of 2 years, children's conceptual categories become increasingly differentiated and so basic-level inductions result. By 4 years of age, the adult tendency to be more certain of inductions made on the basis of smaller classes is already established (Gelman and O'Reilley, 1988).

CONCLUSION

Although more research is needed, the preponderance of evidence indicates that beginning in infancy at least two kinds of categorization occur. The first is an automatic part of perception that computes the perceptual similarity of objects. This leads to implicit categorizing at the basic level. In addition, infants attempt to make sense of what they perceive – to construe the meaning or significance of the events they observe. This leads to explicit conceptual categorization, which tends to be global and overly general at the start.

It is not known for certain whether the two kinds of categorization involve the same or different processes. However, there are some distinctive differences that suggest different processes are at work. First, perceptual categorization occurs automatically and seems to use all the information encoded, whereas conceptual categorization depends on selective attention to only certain kinds of information. Second, perceptual categorization initially makes finer groupings than does conceptual categorization, which begins at a more global level. Third, the functions served by the two kinds of categories differ. Perceptual categorization is used to identify exemplars of conceptual categories.

Infants may respond in a global conceptual way to animals as a class, but must learn to identify them by their features. For example, they learn that 'self-starting objects that do things to other objects' tend to have legs or wings, and so forth. Conceptual categories, on the other hand, are created to understand the world, to characterize the important differences among object kinds. Hence it is these categories that are used for purposes of inductive inferences.

References

Bomba PC and Siqueland ER (1983) The nature and structure of infant form categories. *Journal of Experimental Child Psychology* **35**: 294–328.

Gelman SA and O'Reilley AW (1988) Children's inductive inferences within superordinate categories: the role of language and category structure. *Child Development* **59**: 876–887.

Mandler JM (2000) Perceptual and conceptual processes in infancy. *Journal of Cognition and Development* **1**: 3–36.

Mandler JM and McDonough L (1996) Drinking and driving don't mix: inductive generalization in infancy. *Cognition* **59**: 307–335.

Quinn PC, Eimas PD and Rosenkrantz SL (1993) Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception* **22**: 463–475.

Rosch E, Mervis CB, Gray W, Johnson D and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology* **3**: 382–439.

Smiley SS and Brown AL (1979) Conceptual preference for thematic or taxonomic relations: a nonmontonic age trend from preschool to old age. *Journal of Experimental Child Psychology* **28**: 249–257.

Smith LB and Kemler DG (1978) Levels of experienced dimensionality in children and adults. *Cognitive Psychology* **10**: 502–542.

Waxman SR (1999) Specifying the scope of 13-month-olds' expectations for novel words. *Cognition* **70**: 35–50.

Further Reading

Carey S (1985) *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Gelman SA and Wellman HM (1991) Insides and essences: early understandings of the nonobvious. *Cognition* **38**: 213–244.

Keil FC (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Mandler JM (1998) Representation. In: Kuhn D and Siegler R (eds) *Cognition, Perception, and Language*, vol. 2 of *Handbook of Child Psychology*, pp. 255–308. New York, NY: John Wiley.

Mandler JM (2000) Perceptual and conceptual processes in infancy. *Journal of Cognition and Development* **1**: 3–36.

Markman EM (1989) *Categorization and Naming in Children: Problems of Induction*. Cambridge, MA: MIT Press.

Waxman SR (1999) The dubbing ceremony revisited: object naming and categorization in infancy and early childhood. In: Medin DL and Atran S (eds) *Folkbiology*, pp. 233–284. Cambridge, MA: MIT Press.

Causal Perception, Development of

Intermediate article

Lisa M Oakes, University of Iowa, Iowa City, Iowa, USA

CONTENTS

The perception of causality develops during the first year of life. In general, this development progresses from infants perceiving the individual elements of events (e.g. particular objects, whether or not those objects touch) to their perceiving the relationship between those objects.

INTRODUCTION

Perceiving cause-and-effect relationships is important for understanding how things work, how to produce outcomes and how to link events that co-occur. Philosophers and psychologists have long debated the origins of causal understanding. Hume, for example, argued that real causal connections are unknowable. According to this view, humans determine causality through experience with regularities in the environment and causal perception should develop gradually. Infants may perceive the causality of some events from an early age, but a sophisticated appreciation of causality in a wide range of events develops as they gain experience of the world. Others have argued that the idea of cause and effect is an innate predisposition of the mind. According to this view, causality itself can be perceived without prior experience of events. Therefore young infants should be able to perceive causality in a wide range of events from an early age, and little development should be observed.

PERCEPTION OF LAUNCHING AND COLLISION EVENTS

Psychologists have primarily studied the development of infants' causal perception by assessing their perception of launching events or collisions (e.g. Leslie, 1984; Oakes and Cohen, 1990). At the start of a launching event, an object moves from one side of a display toward a second object sitting at rest in the middle of the display (Figure 1). The first object hits the second object, which begins to move immediately upon contact. Adults report that the first object appears to cause the second object to move (Michotte, 1963). The perception of causality can be interrupted by imposing a delay (i.e. the two objects remain stationary momentarily after they have made contact and before the second object begins to move) or a spatial gap between the two objects (i.e. the second object begins to move before the first object contacts it). Infants' sensitivity to causality is tested using habituation procedures.

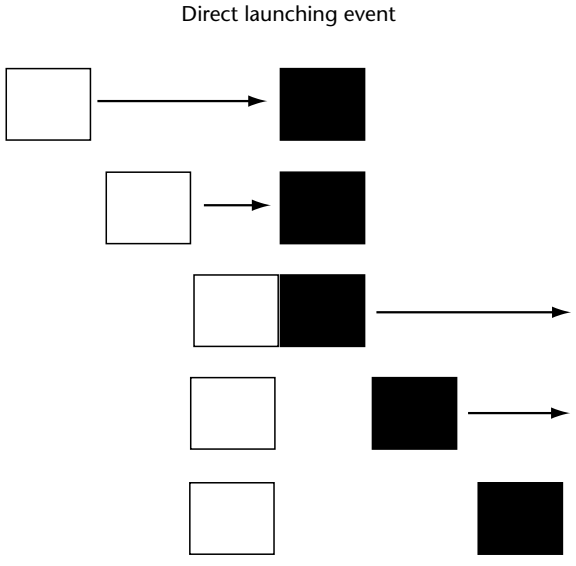


Figure 1. The sequence of actions in a typical launching event. The white object moves from the left of the screen towards the black object, which is initially stationary in the middle of the screen. When the white object makes contact with the black object, the black object immediately begins to move towards the right side of the screen. Reprinted from Oakes LM and Cohen LB (1995) with permission.

They are first shown one event on several trials until their looking time habituates, or decreases to some specified criterion (e.g. 50% of their original level of looking). Then they are shown one or more new events, and if they perceive those events as being different from the familiarization event, they will dishabituate or increase their looking. For example, if infants perceive the causality of events, then infants who are habituated to a delayed launching event, a noncausal event, will dishabituate to a novel causal event but not to a novel no-collision event. However, if infants do not attend to causality, but instead attend to the particular features of the event (e.g. whether or not the two objects touched), then they will dishabituate to novel events that differ in terms of those features regardless of changes in causality.

Using this type of procedure, studies have revealed that young infants are sensitive to causality. Infants aged 6 to 7 months treat the causal launching event as different from noncausal events, and they treat different noncausal events as if they are equivalent (Leslie, 1984; Oakes, 1994; Cohen and Amsel, 1998). Thus we might conclude that some aspects of causal perception are innate. However, causal perception develops considerably during infancy. Infants under 6 months of age do not perceive the causality of launching events. Instead, following habituation with one event, these younger infants dishabituate to changes in other types of features of the event (e.g. whether or not the two objects touched) (Cohen and Amsel, 1998). Even once infants begin to perceive causality, their perception is limited. Infants aged 6 to 7 months only perceive the causality of launching events if the objects are simple (e.g. colored squares). They do not perceive the causality of events if the objects in the event are complex (e.g. multicolored, multi-featured objects). By 10 months of age, infants perceive the causality of launching events involving both complex and simple objects. However, if the objects do not move along the same trajectory, 10-month-old infants fail to perceive causality (Oakes, 1994; Oakes and Cohen, 1995). In general, therefore, the perception of launching events develops over time, and salient perceptual features of the events (or the objects in the events) may overwhelm young infants' ability to perceive the causal relationship in those events.

CAUSALITY VERSUS INDEPENDENT FEATURES MODELS

Clearly, infants over 6 months of age, at least under some conditions, perceive causality (Oakes and Cohen, 1995). The causal event is treated as different from any noncausal event, and different noncausal events are treated as being equivalent. This pattern is consistent with the causality model (Figure 2b). This model is based on the idea that events can be organized in terms of causality, with causal events being perceived as distinct from all noncausal events. According to this model, causal events or spatio-temporally contiguous launching events (events in which the objects touch and there is no delay imposed) are perceived as being different from noncausal events, or launching events with violations of spatio-temporal contiguity. Although this view is called the 'causality model', it is not clear whether causality *per se* is perceived. The same outcome is expected whether observers perceive the causality or are responding to the spatio-temporal contiguity of the events. That is, infants would respond in the same way if they treated a spatio-temporally contiguous event as different from any event that violates spatio-temporal contiguity. What is important is that differences between events are determined along a single dimension that corresponds to causality or spatio-temporal contiguity. As a result, two noncausal events that differ in two ways (e.g. in terms of both the temporal and spatial features) would be treated as the same, and a causal event and a noncausal event that differ in only one respect (e.g. only in terms of temporal features) would be treated as different.

However, infants younger than 6 months do not perceive the causality of events (Cohen and Amsel, 1998). Rather, these younger infants respond to differences in the independent features of the

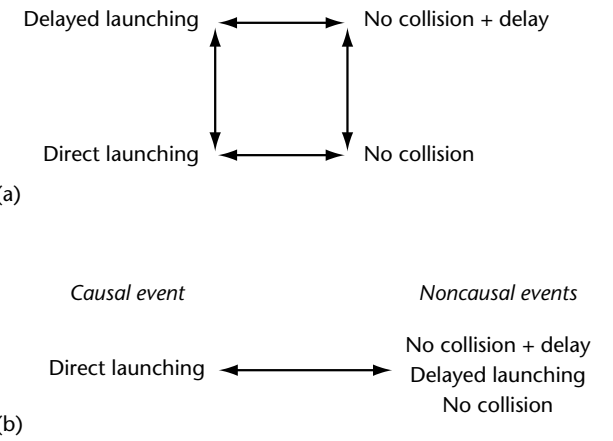


Figure 2. Abstract representations of (a) the independent features model and (b) the causality model of launching events. Reprinted from Oakes LM and Cohen LB (1995) with permission.

event (e.g. whether or not the two objects touch). The perception of launching events by infants younger than 6 months is consistent with the independent features model (Figure 2a). According to this model events are organized in terms of the presence or absence of specific features (e.g. whether or not the two objects touch). If infants process events as sets of independent features, then the perceptual difference between any two events can be represented by an additive combination of the lines shown in the rectangle. Infants would not treat launching events that violate spatio-temporal contiguity as equivalent. Rather, they would respond to differences in the events, such as the presence or absence of a gap or a delay.

In summary, therefore, infants first perceive events according to the independent features model, and in the middle of the first year of life they begin to perceive events according to the causality model. Importantly, this developmental transition is not 'all or none'. Infants can perceive the causality of collisions involving simple objects by approximately 6½ months, but it is not until 10 months of age that they perceive the causality of collisions involving more complex objects.

AGENT VERSUS PATIENT DISTINCTION

Perceiving causality does not simply mean recognizing the difference between causal and noncausal events. Rather, in causal events the two objects have meaningful roles, and a full appreciation of causality requires a recognition of the difference between those roles. Consider the events depicted in Figure 3. For adults, in the initial event the white object appears to cause the black object to move, and it is therefore the causal agent. What makes an object an agent? The agent has the force to cause an outcome, and may be thought of as acting in pursuit of goals (Leslie, 1995). Infants seem to be sensitive to the different roles that objects have in events (Cohen and Oakes, 1993). In this study, 10- to 12-month-old infants were habituated to a causal and a noncausal event in which either the first object was associated with the type of event (e.g. object A was always seen in the agent role of a causal event, and object B was always seen in the agent role of a noncausal event), or the second object was associated with the type of event. The infants were then tested with an event in which the roles were switched (e.g. object A was in the agent role of a noncausal event). Infants dishabituated to this role switch when the agent was associated with the type of event, but not when the patient was

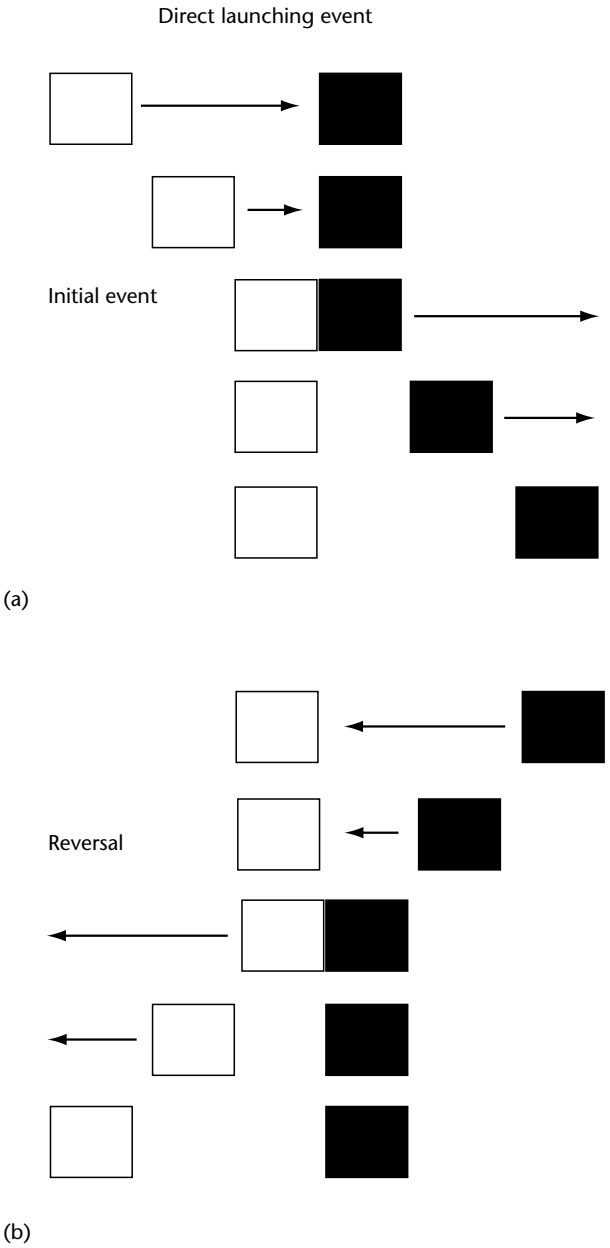


Figure 3. The sequence of actions in (a) a typical launching event and (b) a reversal of that event. Note that when the event is causal, a reversal not only involves a reversal of the direction of movement, but also a change in the roles of the objects in the events.

associated with the type of event. In other words, they linked the agent with whether or not the event was causal, but not the patient. This is an important step in distinguishing between the agent and the patient in the events. However, simply linking the first object with the type of event does not reflect a full understanding of the distinction between the agent and the patient.

The difference in the roles of the two objects is illustrated by the reversal of the event (Figure 3b). In this event, the action occurs in the opposite direction to that in the initial event and, importantly, the roles of the black and white objects have changed. In this reversal, the black object appears to cause the white object to move, and it is now the agent. A reversal of a causal event should therefore be more compelling than a reversal of a noncausal event. Reversals of causal events involve changes in the agent–patient roles, and reversals of noncausal events do not. In fact, infants do find reversals of causal events more interesting than reversals of noncausal ones, which suggests that they are sensitive to the agent–patient distinction (Lesile and Keeble, 1987; Cohen *et al*., 1998). However, their recognition of this distinction develops during infancy. Six-and-a-half-month-old infants notice the agent–patient distinction when the objects involved are simple red and green bricks (Leslie and Keeble, 1987), but it is not until they reach 14 months of age that infants notice the agent–patient distinction when the objects are complex and multi-featured (Cohen *et al*., 1998).

RELATION TO LANGUAGE

The development of general concepts such as causality is believed to be a prerequisite for learning language. Indeed, infants perceive causal relationships long before they learn the corresponding linguistic concepts. Thus we have evidence that the development of cognitive concepts precedes language development. For example, infants perceive the distinction between 'pushing' and 'pulling' at between 10 and 14 months of age. However, they do not associate labels with the type of action until 18 months of age (Cohen *et al*., 1998). In other words, infants first perceive the type of action (pushing and pulling), and only later learn a label for that action. Thus they appear to be able to learn the general concept earlier than they can learn the linguistic concept. Interestingly, by 14 months of age infants can associate labels with objects (Werker *et al*., 1998), which suggests that they can first associate labels with the individual objects in events, and can only later associate labels with the relationships between the objects. In general, therefore, infants are able first to learn words that refer to parts or features of the event, and only later are they able to learn words that refer to the relationships between those parts.

CONCLUSION

In summary, the perception of causality develops gradually during the first years of life. Infants' perception of causality itself develops from initially perceiving the independent features of events, such as the particular objects or some aspects of the relationships between those objects (e.g. whether or not they touch), to distinguishing causal events from noncausal ones. However, the perception of causality does not emerge fully developed. Rather, infants perceive causality earlier in some events than they do in others. They are sensitive to the agent–patient roles in events, but they recognize this distinction earlier in events that involve simple objects. Finally, it is only relatively late in infancy that they become able to associate labels with the actions in the events. Thus, in general, infants first learn about particular objects and only later learn about the relationships between those objects. Therefore an understanding of causality develops gradually with experience.

References

Cohen LB and Oakes LM (1993) How infants perceive a simple causal event. *Developmental Psychology* **29**: 421–433.

Cohen LB and Amsel G (1998) Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development* **21**: 713–732.

Cohen LB, Amsel G, Redford MA and Cassasola M (1998) The development of infant causal perception. In: Slater A (ed.) *Perceptual Development: Visual, Auditory and Speech Perception in Infancy*, pp. 167–209. Psychology Press Ltd.

Leslie AM (1984) Spatiotemporal contiguity and perception of causality in infants. *Perception* **13**: 287–305.

Leslie AM (1995) A theory of agency. In: Sperber D, Premack D and Premack AJ (eds) *Causal Cognition: A Multidisciplinary Debate*, pp. 121–149. Oxford, UK: Clarendon Press.

Leslie AM and Keeble S (1987) Do six-month-olds perceive causality? *Cognition* **25**: 265–288.

Michotte A (1963) *The Perception of Causality*. New York, NY: Basic Books.

Oakes LM (1994) The development of infants' use of continuity cues in their perception of causality. *Developmental Psychology* **30**: 748–756.

Oakes LM and Cohen LB (1990) Infant perception of a causal event. *Cognitive Development* **5**: 193–207.

Oakes LM and Cohen LB (1995) Infant causal perception. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research* vol. 9, pp. 1–54. Norwood, NJ: Ablex.

Werker JF, Cohen LB, Lloyd VL, Cassasola M and Stager CL (1998) Acquisition of word–object associations by

14-month-old infants. *Developmental Psychology* **34**: 1289–1309.

Further Reading

Cohen LB, Amsel G, Redford MA and Cassasola M (1998) The development of infant causal perception. In: Slater A (ed.) *Perceptual Development: Visual, Auditory and Speech Perception in Infancy*, pp. 167–209. Psychology Press Ltd.

Leslie AM (1984) Spatiotemporal contiguity and perception of causality in infants. *Perception* **13**: 287–305.

Leslie AM (1995) A theory of agency. In: Sperber D, Premack D and Premack AJ (eds) *Causal Cognition: A Multidiciplinary Debate*, pp. 121–149. Oxford, UK: Clarendon Press.

Michotte A (1963) *The Perception of Causality*. New York, NY: Basic Books.

Oakes LM and Cohen LB (1990) Infant perception of a causal event. *Cognitive Development* **5**: 193–207.

Oakes LM and Cohen LB (1995) Infant causal perception. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, vol. 9, pp. 1–54. Norwood, NJ: Ablex.

White PA (1988) Causal processing: origins and development. *Psychological Bulletin* **104**: 36–52.

Causal Reasoning, Psychology of

Introductory article

Barbara A Spellman, University of Virginia, Charlottesville, Virginia, USA
David R Mandel, University of Victoria, Victoria, British Columbia, Canada

CONTENTS

Causal reasoning is an important universal human capacity that is useful in explanation, learning, prediction, and control. Causal judgments may rely on the integration of covariation information, pre-existing knowledge about plausible causal mechanisms, and counterfactual reasoning.

INTRODUCTION

The subject of causal reasoning has engaged psychologists of many kinds (e.g. cognitive, social, animal, clinical, developmental), other cognitive scientists (e.g. philosophers, computer scientists, anthropologists), and others outside the cognitive science community (e.g. lawyers). The question that psychologists want to answer is: how do we go from the information that the world provides us in the form of events occurring (seemingly at random sometimes) to our beliefs about what causes what? Sometimes causal judgments are made in formal settings: in the laboratory, scientists try to find out what causes cancer or heart disease; in the legal system, before liability or punishment is imposed, jurors are required to determine who caused the accident or who caused someone's death. But more informally, we all reason about causality daily. Why did I fail this exam (or all of the exams in this course)? Why is my friend unhappy? Why is my computer likely to crash in the next five minutes? To answer these kinds of questions we may rely on repeated observations, pre-existing knowledge, thought experiments, or all of these cues to causality.

THE ROLE OF REPEATED OBSERVATIONS

One kind of information that we use to assess causality is information from repetitions of the same events: watching causes and effects as they repeatedly occur. Sometimes people use the word 'cause' deterministically, so that '*A* causes *B*' means that every time *A* occurs *B* must follow. But people also use the word 'cause' probabilistically, so that '*A* causes *B*' means that *A* increases the chances that *B* will occur. For example, someone might say that a baseball team has a winning record early in the season 'because' they have played most of their games at home. The argument is that playing at home increases the chances of winning; however, it's still possible for the team to lose some games at home and win some away games. Similarly, scientists claim that smoking causes lung cancer because it increases the chance of getting lung cancer. Yet there are people who smoke and don't get lung cancer, and people who get lung cancer without smoking.

In order to determine whether something might be causal when we have repeated observations, we divide our observations into the following categories (see Figure 1): cause present and effect present (cell *A*); cause present and effect absent (cell *B*); cause absent and effect present (cell *C*); cause absent and effect absent (cell *D*). We then use that information to decide whether the cause increases the probability of the effect. We do this by comparing the proportion of times the effect occurs when the cause is present, $A/(A + B)$, with the proportion of times the effect occurs when the cause is absent, $C/(C + D)$. The difference between these proportions is called the contingency (symbolized by Δp – read 'delta *p*'), which is a measure of the strength or effectiveness of a cause. If the proportion of times the effect occurs is greater when the cause is present than when the cause is absent then the difference will be positive, and we speak of a 'generative' or 'facilitative' cause – it makes the effect more likely to happen. If the proportion of times the

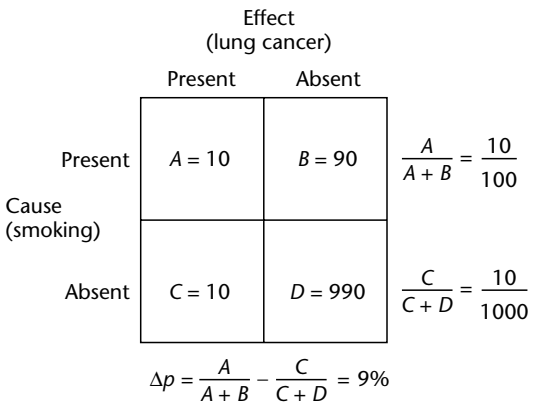


Figure 1. We need four kinds of information to understand the relation between smoking and lung cancer. Note that we do not directly compare the number of people who smoke and get lung cancer with the number of people who don't smoke and get lung cancer. Rather, we look at whether smoking increases the chances of getting lung cancer. For people who smoke, the proportion who get it is 10/100 or 10%, whereas for people who don't smoke, the proportion who get it is 10/1000 or 1%.

effect occurs is smaller when the cause is present than when the cause is absent then the difference will be negative, and we speak of a 'preventive' or 'inhibitory' cause – it makes the effect less likely to happen. (Contingencies can range from -1 to $+1$.) Usually people are concerned with generative causes.

Using all the Information

The above computation of Δp suggests that all of the information should be equally important in evaluating the effectiveness of a cause. However, studies have shown that people tend to overweight the information in cell A – the 'present–present' cell. This overweighting may be one reason why people believe in superstitions or horoscopes. For example, there are people who believe that if they walk under a ladder they will have bad luck. And, in fact, once or twice when they did walk under a ladder they did have bad luck. However, they may fail to remember or consider the times they walked under a ladder and didn't have bad luck (cell B), the (possibly many) times they had bad luck without walking under a ladder (cell C), and the very many times they didn't walk under a ladder and didn't have bad luck (cell D). All of that information is needed before one can correctly evaluate whether walking under a ladder causes bad luck.

Considering the Base Rate of the Effect

Although the Δp computation captures the idea that a cause is something that changes the probability of an effect, the number that results from the computation may be deceptive when one is trying to evaluate the effectiveness of a particular cause. Besides looking at the contingency, people also consider how often an effect occurs in general (called the 'base rate' of the effect).

For example, suppose you have 100 plants in your garden but only 80 of them have flowers. You buy a special plant food, and soon all 100 have flowers. How effective is this product? It may only work 20 percent of the time (and have happened to work on the flowers that hadn't already bloomed); or it may work 100 percent of the time (but you couldn't tell because some flowers had already bloomed anyway).

It appears that people are sensitive to this problem when judging the effectiveness of a cause. For example, compare two plant foods. One was given to 100 plants where there were initially no flowers, and then 20 bloomed ($20/100 - 0/100 = 0.20$). The other, as above, was given to 100 plants where there were initially 80 flowers, and then all 100 bloomed ($100/100 - 80/100 = 0.20$). Both plant foods have a contingency of 0.20, yet the first worked only on 20 percent of the plants that didn't already have a flower whereas the second worked on 100 percent of the plants that didn't already have a flower.

People judge the second plant food as more effective than the first even though the contingencies are equal. Thus, it seems that people take into account the base rate of the effect and adjust their estimates according to how much influence a cause had above and beyond the influence of other causes.

THE ROLE OF PRE-EXISTING KNOWLEDGE

We have seen how we can use statistical covariation information to assess the relation between a potential cause and effect. However, we don't only use statistical information when making causal judgments; our pre-existing knowledge of the world will influence which statistical information we will pick up and use, and how we will limit and interpret the statistical relations we discover.

The most important limitation to note is that even though finding a contingency means that there is a correlation between the two events, that does not mean that the first event causes the second event. As scientists often say: 'correlation

does not prove causation'. For example, every morning, just before dawn, the rooster crows. Then the sun rises. Yet we do not believe that the rooster crowing causes the sun to rise. In many cities, when ice cream sales go up, the murder rate goes up; when ice cream sales go down, the murder rate goes down. Yet we do not believe that eating ice cream causes people to commit murder. Why not? Because we have other knowledge.

Causal Mechanisms

Why don't we believe that the rooster causes the sun to rise even though there is a perfect correlation? One idea is that we don't believe it because we can't conceive of a causal mechanism. How could the noise of a tiny animal affect a powerful celestial object? In fact, belief (or non-belief) in a mechanism can direct, or misdirect, searches for potential causes. For example, in the mid-nineteenth century a physician named Ignaz Semmelweis had a very difficult time convincing physicians to wash their hands after examining a cadaver before turning to deliver a baby (to lessen mortality of the mothers) – a practice that seems obvious and obligatory now – because no one then could imagine a causal mechanism.

So where do our beliefs about causal mechanisms come from? One possibility is that they come from our knowledge of similarities, categories, and other statistical relations. We don't believe that the rooster causes the sun to rise because we know that lions roaring don't cause rain and dogs barking don't cause full moons. There are no statistical relations between one 'kind' of event (animal noise) and the other 'kind' of event (weather), so we never developed the idea that there could be a causal mechanism. On the other hand, these days we are willing to accept that many new ailments can be caused by unseen microorganisms which can be transmitted in many ways (e.g. breathing, touching) because we have noted other similar relations in the past.

Labeling a Cause as a Cause

Given the same combination of events, which gets labeled as a 'cause' of the outcome may differ between situations, individuals, and cultures. For example, suppose a fire breaks out in a nearby warehouse and you are explaining the cause of the fire to a friend. You are likely to mention that there was an arsonist or a stroke of lightning. You are unlikely to mention the presence of combustible material or oxygen, even though both of those are necessary for the fire. Under 'normal' circumstances we just assume that they are present, and so their presence or absence does not covary with the outcome and we do not consider them causes. Now imagine a special furniture factory in which an area is kept free of oxygen so that high-temperature welding can take place. One day there is an oxygen leak, and when the usual welding begins a fire ensues. Under these special circumstances, you would call the oxygen a cause of the fire. Thus, what we point to as being causal is not just something that increases the probability of an effect, but rather something that increases it relative to some background assumption of what is stable or normal.

Which of the many relevant factors a person chooses to pick out from the background and label as a cause may also depend on that individual's beliefs. For example, suppose a young man robs a shop. What caused this behavior? Some people would argue that it was because he was brought up in a bad neighborhood, citing the fact that children brought up in his neighborhood are more likely to go on to commit such a crime than children brought up in better neighborhoods. Other people would argue that it was because he was a 'bad apple', citing the fact that there are many other children brought up in his neighborhood who do not commit such crimes. In such a case, what you label as a 'cause' may influence what you believe is the best treatment for the problem.

Different cultures also tend to pick out different factors as causal. For example, in individualist cultures (such as the United States and Australia) people are more likely to attribute causality for an action to the actor's personality or 'disposition', whereas in collectivist cultures (such as China and India) people are more likely to attribute causality to the situation or circumstances. (*See* **Cultural Differences in Causal Attribution**)

Alternative Causes

Although it seems easy enough to figure out the statistical relationship between one cause and one effect once you know where to look, the world is complicated and it is not always possible to examine one potential causal relation at a time. Sometimes potential causes are independent, so you can evaluate each separately. But often potential causes covary with each other, making it difficult to distinguish the causal contribution of each.

Controlling for alternative causes

When two or more potential causes of an effect act at once, and not independently, we have to control for one cause while evaluating the other. As a simple example, suppose you rush into your favorite coffee shop and assert loudly that drinking coffee must cause lung cancer because people who drink lots of coffee get lung cancer more often than those who do not (a positive Δp). Probably none of the coffee drinkers there would be alarmed: they would point out to you that perhaps people who drink more coffee also smoke more, so although it may look as if coffee drinking causes lung cancer, it is really smoking that is doing the causal work. Here people see an alternative causal mechanism to explain the lung cancer: smoking. To evaluate whether coffee drinking causes lung cancer while controlling for smoking you need to (1) consider all people who don't smoke and ask whether coffee drinking increases their probability of getting lung cancer, and (2) consider all people who do smoke and ask whether coffee drinking increases their probability of getting lung cancer. If the answer is negative in both cases, then coffee drinking is not a cause of lung cancer: it is only because it covaries with smoking that it seems to raise the probability of lung cancer.

When evaluating whether something is a cause of an effect, it is important to control for alternative causes. Obviously, it is never possible to know for certain that one has considered all potential alternative causes, but controlling for known alternative causes is a technique intentionally used by psychologists and other scientists to improve scientific reasoning.

However, controlling for alternative causes is difficult without a theory of what those alternative causes might be. In the coffee example, people don't believe there is a way in which coffee drinking could cause lung cancer, so they seek an alternative causal mechanism. But in the case of ice cream sales and murder rates, it might be plausible to think that increased sugar consumption causes violence, and leave it at that. A mechanism is not necessarily correct just because it is plausible. When temperature is controlled for, there is no correlation between ice cream sales and murders; they only seem related because hot weather leads to more of both. Many experiments have shown that people do control for known alternative causes when judging causal effectiveness. If experimenters tell people about an alternative cause, or if they have a reason to believe that some alternative factor (e.g. smoking) might be causal, then they will think of controlling for that cause. However, in experiments where the alternative factor (like temperature) is not so obvious, people are less good at controlling for it.

Thus, pre-existing knowledge of a causal mechanism may affect what information people acquire from the environment, what they control for, and, ultimately, what they will judge as the causes of other events.

Discounting

Although people do control for known alternative causes, the presence of an alternative cause may lead to a misjudgment of causal strength – known as 'discounting'. Discounting occurs when someone learns about two causes at the same time, and the judgment of the strength of the first cause is affected by the strength of the second cause. For example, suppose you have allergies and your doctor prescribes two medications A and B. Sometimes you take one, sometimes the other, and sometimes both. Medication A works ($\Delta p = 0.33$), but medication B doesn't work ($\Delta p = 0$). You tell the doctor that A is fairly effective in relieving your allergies. Now consider what would have happened had the doctor prescribed medications A and C instead. A still works ($\Delta p = 0.33$), but C works even better ($\Delta p = 0.67$). In your report to the doctor, you are likely to judge A as being less effective when the alternative cause is strong (C) than when the alternative cause is weak (B) – even though A's effectiveness is the same.

Experiments have shown that discounting occurs even in simple cases when the two causes are independent. Discounting might result from a strategic decision or a belief that once we have found a good cause of an effect we need not invest in reliably assessing other potential causes.

JUDGING CAUSALITY IN SINGLE INSTANCES

We have considered how people make causal judgments when they have information about many instances of the cause and effect occurring. But how do we make causal judgments for events that occur only once (e.g. an accident, a crime, or a big promotion)? One theory is that we use counterfactual reasoning to make these judgments.

Counterfactual Reasoning

In many situations, people look back on a past episode and wonder what might have happened if some change had occurred leading up to its conclusion. For instance, suppose you decided to

drive home from work by a scenic route one day because it was particularly beautiful outside, and along the way your car was hit by a reckless driver. Many, if not most, people in this situation would replay the episode in their mind in such a way that the accident is somehow 'undone'. For instance, you might imagine that if only you had taken your usual route home, or if only you had left a few seconds later, the accident would have been avoided. These imaginings of how the past might have been different involve counterfactual (or contrary-to-fact) thinking because the mentally replayed episode differs in some respects from the real episode.

Some philosophers and psychologists have proposed that counterfactual thinking plays an important role in causal reasoning. To explore the possible causes of a particular outcome, a person may mentally change one of the events preceding the outcome (an 'antecedent') and observe whether the outcome still occurs in the mental replay. If it is easy to imagine that the outcome would also be undone, then the antecedent is likely to be seen as one of its necessary causes. If, however, the outcome still seems inevitable, then the antecedent is unlikely to be seen as causally relevant.

Political scientists, historians, and legal scholars sometimes run counterfactual thought experiments to examine the causal implications of a particular change to a complex system. For example, some historians have considered what might have happened if Archduke Franz Ferdinand had not been assassinated in Sarejevo; and some have concluded that if that event had not happened, then the First World War would not have happened either. Although counterfactual thought experiments can be informative, it is usually impossible to verify whether the causal inferences drawn from them are justified – precisely because history cannot literally be replayed.

Similarities and Differences Between Causal and Counterfactual Reasoning

Counterfactual and causal reasoning sometimes focus on different events. For instance, in the car accident scenario, if people are asked directly about the cause of the accident they tend to identify the reckless driver, whereas if they are asked to generate counterfactuals that would undo the accident they tend to mention the route taken home. The counterfactual thoughts that come to mind may represent one's after-the-fact understanding of how the bad outcome could have easily been prevented. Thus, counterfactual thoughts often focus on behaviors that individuals can control. On the other hand, causal explanations often focus on antecedents that our knowledge of the world indicates would covary with the outcome over a set of similar cases. Thus, people say that the reckless driver was the cause of the accident because they realize that reckless driving is predictive of car accidents in general.

Even though the two forms of reasoning sometimes diverge, there is nevertheless a strong interplay between counterfactual and causal reasoning. If one cannot imagine that an antecedent might have been different, then it is unlikely that the antecedent will be identified as a cause. This is why, under normal conditions, oxygen makes a poor causal explanation for fire, even though it is a necessary condition. Similarly, if one cannot imagine an outcome having been different, then it is questionable whether it could be causally explained other than by recourse to concepts such as fate or destiny, which by definition, emphasize the immutability and inevitability of past episodes. (*See* **Counterfactual Thinking**)

DEVELOPMENT OF CAUSAL REASONING

Causal reasoning is necessary for human survival and, not surprisingly, the ability to perform such reasoning develops early. However, it is difficult to study causal reasoning in infants, because researchers cannot ask them direct questions about their judgments. Instead, researchers often use a technique called the 'habituation paradigm'. This technique takes advantage of the fact that when infants see the same events repeatedly (e.g. pictures on a video monitor, animations, objects moving in real life), they gradually get bored and will look at the events for shorter durations. If, however, they see a new or different event, they will 'dishabituate' and look for longer. Researchers can infer what infants count as 'the same thing' or 'a different thing' using this technique.

When using the habituation paradigm to study causal reasoning, researchers may show infants videotapes of collision events. In the 'causal launching' event, object *A* moves across the screen and hits stationary object *B*. When *A* strikes *B*, *A* stops, and *B* immediately begins to move with the same speed and direction as *A* had previously. This event looks natural to an adult, as if *A*'s collision with *B* caused *B* to move. Researchers can then modify these events. For example, a delay may be introduced, so that after *A* hits *B*, *B* remains in place

for a second or two before it begins moving. Or, *B* may begin moving before being hit by *A*. With these modifications, adults will claim that it doesn't look as if *A* caused *B* to move. Do infants treat these modified events as the same as, or different from, the natural-looking causal launching event? Research has shown that by the age of about seven months, infants do perceive a difference between causal launching events and noncausal events. However, they perceive the difference only if the objects involved in the events are simple; for more complicated objects infants may have to be 10 months old to make the distinction.

If it takes infants longer to understand causality regarding more complicated objects, what about more complicated events? Often in life, we don't just say that an outcome was caused by the action that immediately preceded it; rather, we look back in time to see what caused that particular action. For example, when Mum sees milk spilled on the floor and yells at Little Sister to wipe it up because it fell from her tilted glass, if Little Sister says 'Big Brother pushed me' then he, rather than she, is seen as the cause and gets Mum's wrath. It is considered a sign of sophisticated reasoning in adults to be able to look back into the past for causes; this ability also develops over time in infants: at 10 months old they don't look back at earlier causes, whereas at 15 months old they do.

Thus, we see that causal reasoning starts to develop early, but not all at once. Rather, both the complexity of the objects involved in the events and the complexity of the relations between events affect causal understanding. (*See* **Causal Perception, Development of**)

CONCLUSION

Causal reasoning is a pervasive and important form of thinking that begins at an early age. People engage in causal reasoning in order to explain past outcomes, to achieve control over their natural and social environment in the present, and to forecast, plan and prepare for the future. However, reasoning about causality is complicated. It is complicated, in part, because causal judgments depend on multiple cues, such as covariation, spatial and temporal contiguity, and our beliefs about what is normal. It is also complicated because information about such cues may be obtained in a variety of ways, such as by observing new cause–effect sequences, recalling knowledge about the world, and mentally imagining counterfactuals. Moreover, the answer to the question 'What is the cause?' may depend on one's choice of causal background – which may be affected by motivation, knowledge, and culture. Despite the complexity of the concept, the power of such knowledge, when it is accurate, is formidable. Without the ability to understand causality and use causal knowledge, both our internal mental world and the external physical world in which we live would be radically different.

Further Reading

Cheng PW and Wu M (1999) Why causation need not follow from statistical association: boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science* **10**: 92–97.

Cohen LB, Rundell LJ, Spellman BA and Cashon CH (1999) Infants' perception of causal chains. *Psychological Science* **10**: 412–418.

Hart HL and Honoré AM (1985) *Causation in the Law*, 2nd edn. Oxford, UK: Oxford University Press. [First edition published in 1959.]

Hilton DJ (ed.) (1988) *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York, NY: New York University Press.

Mandel DR and Lehman DR (1996) Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology* **71**: 450–463.

Mandel DR and Lehman DR (1998) Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General* **127**: 269–285.

McGill AL (1989) Context effects in judgments of causation. *Journal of Personality and Social Psychology* **57**: 189–200.

Spellman BA (1997) Crediting causality. *Journal of Experimental Psychology: General* **126**: 323–348.

Spellman BA and Mandel DR (1999) When possibility informs reality: counterfactual thinking as a cue to causality. *Current Directions in Psychological Science* **8**: 120–123.

Spellman BA, Price CM and Logan J (2001) How two causes are different from one: the use of (un)conditional information in Simpson's paradox. *Memory and Cognition* **29**: 193–208.

Sperber D, Premack D and Premack AJ (eds) (1995) *Causal Cognition: A Multidisciplinary Debate*. New York, NY: Oxford University Press. [Symposia of the Fyssen Foundation.]

White P (1990) Ideas about causation in philosophy and psychology. *Psychological Bulletin* **108**: 3–18.

Change Blindness, Psychology of

Intermediate article

George McConkie, University of Illinois, Champaign, Illinois, USA
Lester Loschky, University of Illinois, Champaign, Illinois, USA

CONTENTS

Change blindness is the tendency to fail to detect changes in a stimulus array while actively exploring it. This happens when the perception of the motion that typically accompanies stimulus change is prevented or disrupted.

INTRODUCTION

As we look around our visual world, we have a sense that we are continuously perceiving most or all of the available information. Thus, it came as some surprise in 1992 when scientists from the University of Illinois reported research indicating that if parts of a rich photographic image are changed while an observer makes a saccadic eye movement, these changes frequently go unnoticed (Grimes, 1996). Such intrasaccadic changes of the presence, location, size, orientation, color, category, or identity of a prominent object often go undetected as a person looks around a picture. Normally a person notices an image change because of the stimulus motion accompanying it; however, when the change occurs during a saccade, that motion is hidden by saccadic suppression. Thus, people cannot notice the change unless information they acquire following the change conflicts with that obtained prior to the change. Such detection failures have been taken as evidence that making a saccade (lasting only 20–80 ms) destroys the iconic image from the prior eye fixation, thus ruling out integration of iconic retinal images across fixations (Irwin, 1992), and further, that only a small amount of information (e.g., an object's identity or category, or perhaps the scene's gist) is actually retained from each eye fixation. The mental representation of a scene that is built up, incrementally, across eye fixations is apparently quite sparse; although visual memory cannot store the rich retinal image that is present during a fixation, it may not need to, since (as Kevin O'Regan has argued) the brain can easily access that information again by simply making an eye movement, so long as the stimulus array remains unchanged. Failure to detect a change in the visual stimulus was later labeled 'change blindness' by Daniel Simons. A lively research area has arisen, investigating what is, and is not, retained from a complex display during its viewing, with numerous theories about the significance this has for understanding perception, attention and cognition.

REQUIREMENTS FOR CHANGE BLINDNESS

Change blindness occurs when an initial stimulus is presented, followed by a change to a different stimulus, but with the transition being perceptually hidden in some way. Normally a local or global change in a picture is easy to detect, since the change induces the perception of motion, which automatically captures attention. In the original paradigm, making the transition during a saccade hid the change. Studies were conducted in which an image was changed for a single eye fixation, allowing a precise examination of whether the changed information was acquired during that particular fixation (McConkie and Currie, 1996). However, a later study (Rensink *et al*., 1997) showed that a saccade was not necessary to elicit change blindness. Rather, the same effect can be produced by alternating between two versions of a picture, blanking the screen for roughly 100 ms or more between them. Here, the visual disruption produced by the blank screen hides the transition. By continuously 'flickering' between images, the experimenter can measure the time taken by the viewer to discover what is changing. Many such

changes are surprisingly difficult to find, as illustrated in Figure 1. Soon other methods for hiding the transition were being employed: changing the image during a blink, briefly introducing irregular blotches ('mud splashes') that compete for attention simultaneously with the image change, morphing between the two images so slowly that the motion is not detected, using film editing techniques to 'cut' from one camera shot to another in motion pictures, or even physically blocking the person's view in a real-world setting.

(a)

(b)

Figure 1. Two versions of a photograph used to study change blindness. These two versions (a) and (b), are repeatedly presented, one after the other, on a computer screen for about 0.25 s each, with a 0.10 s blank period between. After a period of searching, people typically detect a change in one small area, and only later realize the full extent of the change taking place. This demonstrates that people are able to retain only a small amount of information from each presentation of the picture. (used with permission of Gregory J. Zelinsky).

In a dramatic demonstration of the last method, while an experimenter asked a person on the street for directions, their view of each other was briefly occluded by two men who rudely walked between them carrying a door, and an experimenter of different dress and build quickly replaced the questioner (Figure 2). Roughly half of the people tested failed to notice the change in their interlocutor (Simons and Levin, 1998).

PREVIOUS OBSERVATIONS OF CHANGE BLINDNESS

On reflection, it is obvious that the change blindness phenomenon has long been utilized: magicians developed ways of diverting attention so a change is not noticed, and the familiar 'spot the difference' cartoons require people to make saccades (with the attendant disruptions) between two pictures to find what has been changed. In psychology, from the 1950s on, various studies tested people's memory by asking them to look first at one picture, then at another, and indicate any difference. However, only recently has the change blindness phenomenon elicited particular theoretical interest among those studying perception and cognition.

EXPLANATIONS OF CHANGE BLINDNESS

There are several types of explanations for the failure to detect image changes. First, the critical prechange information might never have been attended to, or it might have been forgotten afterwards. Alternatively, after the change, the critical information might not be attended to. Finally, the information both before and after the change might have been attended to but the discrepancy between them might not be noticed, indicating a lack of integration of information acquired at different times. Thus, researchers try to determine the extent to which change blindness in a given situation results from attentional selectivity (information from before or after the change is not attended and stored), forgetting, or failure to integrate stored information. Explanations of instances of change blindness are given below.

Spatial or Object-based Inattention

Selective visual attention is critical in change blindness. Factors that draw attention to an object increase the likelihood of detecting changes to it,

Figure 2. A real-world example of change blindness. (a) An experimenter stops someone (an unwitting 'subject') on the street and asks for directions. (b) In the middle of their interaction, two men rudely pass between them carrying a door; during this brief interval, the original experimenter is replaced by a second experimenter. After the door passes, the second experimenter continues the conversation with the subject as if nothing had happened (c). About half the time, subjects fail to notice that they are now talking to a different person, even though the two experimenters differ in height, build, hair style, and clothing (d). From Simons and Levin (1998), with permission.

while reducing the detection of changes elsewhere. If a verbal cue indicates the identity of a change before its occurrence, or a visual cue draws attention to the changed object, detection is greatly facilitated (Rensink *et al*., 1997; Scholl, 2000). Having more objects in the display delays detection of a change in one object due to increased attentional scanning (Zelinsky, 2001), and when several objects change location, people usually detect the movement of only one (McConkie and Loschky, 2000). People who have a wider attentional breadth, measured by their ability to detect briefly presented targets in peripheral vision, are better at detecting changes under flicker conditions (Pringle, 2000). Changes occurring in central vision, the region most likely to be attended to during a fixation, are better detected than changes in the visual periphery. Finally, changes to objects of greater interest in a picture are detected faster than changes to objects of lesser interest (Rensink *et al*., 1997). All these phenomena argue for the importance of selective attention, though attending to an object does not necessarily result in detection of a change to it.

Temporal Inattention

During reading, there is evidence of attention to words at only selected times during fixations (Blanchard *et al*., 1984). In one study, at the beginning of each fixation the text being read contained one letter at a specified location. After a short period the text was briefly masked, and then reappeared but with the critical letter being changed, for example changing the word 'leaks' to 'leans'. Thus, one word was present at that location at the beginning of each fixation, and a different word during the latter part of each fixation. Readers were usually unaware of this change, reporting having read only one of the words: the perceived word was sometimes the earlier word and sometimes the later. This suggests temporal inattention. Whether this inattention to a word only occurs because attention is being given to a different word is not known.

Memory

The initial information that will be changed must be retained if the change is to be noticed. Pringle

(2000) found a strong relationship between people's composite scores on a battery of visuospatial working memory (VSWM) tests and their ability to detect changes, providing evidence for the role of memory in change detection. In fact, the VSWM measure predicted people's detection ability better than a test of attentional breadth did.

Expectations

Viewer's expectations and world knowledge also play a part. Changes of task-related aspects of a display (color of a traffic light for a driver) are detected more quickly than changes of unrelated aspects (location of a light pole) (Pringle, 2000). Changes to objects inappropriate in their context are detected faster than when they are appropriate (Hollingworth and Henderson, 2000).

Stimulus Characteristics

The more objects or parts of an object that change, the greater is the likelihood that a change will be detected (McConkie and Loschky, 2000; Williams and Simons, 2000). On the other hand, some features of the stimuli to which the gaze is directed can be changed during a saccade without detection. For example, size changes are often not detected (Grimes, 1996; McConkie and Currie, 1996). When reading text printed in AlTeRnAtInG cAsE (every other letter in upper case), the case of every letter can be changed during a saccade, thus changing the shape of every letter and word, without the readers' awareness or any effect on their eye movements. This indicates that letter and word shapes are not preserved across saccades during reading. Just directing the gaze toward an object does not guarantee that changing it will be detected: there are cases in which the viewer makes two successive eye fixations on an object in a picture, with the object changing substantially between those fixations, with no detection of the change (Grimes, 1996). These examples are all consistent with the proposal that the visual system retains only a limited set of the features of an attended object.

Coordinating Perception Across Saccades

A different mechanism, related to basic processes involved in coordinating perception across saccades, has been proposed for detecting intrasaccadic shifts, or displacements, of the stimulus (McConkie and Currie, 1996). It is proposed that, prior to making a saccade, the visual system chooses a target to send the eyes to (the 'saccade target object') and stores some identifying features (the 'locating features'). On the next fixation, the first task is to search for the locating features within the region of central vision (the 'search region'), to find the retinal location of the saccade target object. Locating that object establishes a mapping function between current retinal information and information obtained during prior fixations, enabling further perceptual activities to occur. If an intrasaccadic displacement of the stimulus moves the saccade target object out of the search region, this disrupts further processing and produces a conscious awareness of the image displacement, or change detection. Similarly, if the locating features are changed, this will interfere with the finding of the saccade target object.

The importance of the saccade target object in perception is confirmed by the fact that spatially displacing only that object is far more detectable than is displacing everything else in a picture except the saccade target (Currie *et al*., 2000). This work suggests that intrasaccadic display changes can be detected on different bases. Changes may be detected by early visual processes that are involved in locating the saccade target at the beginning of a fixation, or later as information acquired in the fixation is recognized as conflicting with information retained from earlier fixations.

CONSCIOUSNESS AND CHANGE BLINDNESS

An important limitation of current research is that it has depended almost exclusively on conscious report: did the viewer detect the change? However, it is possible that image changes produce effects on processing even if not consciously perceived. Studies are needed to search for effects of undetected display change on implicit measures of processing, such as eye movements or brain waves. Results from such studies could challenge current conclusions about the sparsity of information retained from brief stimulus exposures.

CONCLUSION

Although change blindness is an intriguing phenomenon, in some sense it is only incidentally a topic of study. Rather, it is a research paradigm that can be used to study issues regarding the selection, acquisition, retention and integration of information. Its primary contribution has been to call attention to how little information is retained from eye fixations and other brief exposures to

complex stimulus patterns such as photographs or real-world scenes, and how sparse our mental representations actually are.

References

Blanchard HE, McConkie GW, Zola D and Wolverton GS (1984) Time course of visual information utilization during fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance* **10**(1): 75–89.

Currie CB, McConkie GW, Carlson-Radvansky LA and Irwin DE (2000) The role of the saccade target object in the perception of a visually stable world. *Perception and Psychophysics* **62**(4): 673–683.

Grimes J (1996) On the failure to detect changes in scenes across saccades. In: Atkins KA (ed.) *Perception*, vol. 5, pp. 89–110. New York, NY: Oxford University Press.

Hollingworth A and Henderson JM (2000) Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition* **7**(1–3): 213–235.

Irwin DE (1992) Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**(2): 307–317.

McConkie GW and Currie CB (1996) Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance* **22**(3): 563–581.

McConkie GW and Loschky LC (2000) Attending to objects in a complex display. In: Benedict ME (ed.) *Advanced Displays and Interactive Displays Consortium ARL Federated Laboratory Fourth Annual Symposium Proceedings*, pp. 21–25. Adelphi, MD: Army Research Laboratory.

Pringle HL (2000) *The Roles of Scene Characteristics, Memory and Attentional Breadth on the Representation of Complex Real-world Scenes* [unpublished doctoral dissertation]. Urbana, IL: University of Illinois.

Rensink RA, O'Regan JK and Clark JJ (1997) To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* **8**(5): 368–373.

Scholl BJ (2000) Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition* **7**(1–3): 377–396.

Simons DJ and Levin DT (1998) Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review* **5**(4): 644–649.

Williams P and Simons DJ (2000) Detecting changes in novel, complex three-dimensional objects. *Visual Cognition* **7**(1–3): 297–322.

Zelinsky GJ (2001) Eye movements during change detection: implications for search constraints, memory limitations, and scanning strategies. *Perception and Psychophysics* **63**(2): 209–225.

Further Reading

Hayhoe MM, Bensinger DG and Ballard DH (1998) Task constraints in visual working memory. *Vision Research* **38**(1): 125–137.

O'Regan JK, Rensink RA and Clark JJ (1999) Change-blindness as a result of 'mudsplashes'. *Nature* **398**(6722): 34.

O'Regan K (1992) Solving the 'real' mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology* **46**(3): 461–488.

Ryan JD, Althoff RR, Whitlow S and Cohen NJ (2000) Amnesia is a deficit in relational memory. *Psychological Science* **11**(6): 454–461.

Simons DJ (ed.) (2000) Change blindness. *Visual Cognition* **7**(1–3) [special triple issue].

Simons D and Levin D (1997) Change blindness. *Trends in Cognitive Sciences* **1**: 261–267.

Choice Selection

Intermediate article

John W Payne, Duke University, Durham, North Carolina, USA
James R Bettman, Duke University, Durham, North Carolina, USA

CONTENTS

Choice selection involves a set of alternatives, each described by some attributes or consequences for the decision-maker's goals. Decision-makers choose among such options using a variety of different psychological strategies.

INTRODUCTION

Multiattribute Choice Problems

Decision-making is a fundamental cognitive behavior which depends on a person's values and beliefs and can range from mundane problems (such as selection of a menu item for lunch) to more substantial and infrequent decisions (such as which automobile to purchase), and life-or-death decisions (such as a choice between alternative medical treatments). At the heart of the decision-making process is choice or selection among alternative courses of action. A simplified automobile choice task is illustrated in Table 1. Each choice option i (alternative) is described by a vector of attribute values (x_{i1}, x_{i2}, …, x_{in}), reflecting the extent to which each option meets the objectives (goals) of the decision-maker.

A key feature of almost all choice problems is the presence of conflict, since no single alternative is best (most preferred) on all attributes. Conflict is a major source of decision difficulty; negative emotions can arise from the need to accept less of one valued attribute (e.g., safety) in order to achieve more of another valued attribute (e.g., cost savings). The presence of conflict and the fact that a rule for resolving the conflict often cannot be drawn from memory are reasons that even simple choice tasks can lead to tentativeness and the use of novice problem-solving methods rather than the kind of pattern recognition methods that are typically associated with expertise in various domains.

Choice Among Risky Options

Another key feature of choice problems is the degree of certainty of the consequences associated with an attribute value. For example, one might (or might not) be certain about the level of reliability for a particular car in Table 1. Another example of uncertainty and choice is selecting between two gambles, one of which offers a higher probability of winning a lower amount of money while the other offers a lower probability of winning a larger amount of money. A real-world example of a risky choice problem with such trade-offs is selecting among investment options (gambles) such as bonds versus computer stocks for the next year when one is uncertain whether the state of the economy will improve or worsen.

CHOICE MODELS

The study of decision processes has long been of interest to psychologists, economists, and researchers in many other fields. How do people make preferential choice decisions of the types described above? The rational choice view is that people solve all (most) decision problems by obtaining all the relevant information about the decision problem, incorporating uncertainties into their reasoning, making trade-offs where necessary, and eventually selecting the alternative that maximizes their values. That is, people are presumed to be exquisitely rational beings who have, if not perfect, at least clear and voluminous information, who are able and willing to make trade-offs, and who always select the best course of action (Simon, 1955). The view that people are generally rational decision-makers means that one can start by trying to work out the best way a decision problem should be solved, assume that people try to do the best and are capable of calculating the best option, and therefore that the same model that provides a

Table 1. An example of a choice task

Car	Reliability	Price	Safety	Horsepower
A	Worst	Best	Good	Very poor
B	Best	Worst	Worst	Good
C	Poor	Very good	Average	Average
D	Average	Poor	Best	Worst
E	Worst	Very poor	Good	Best

Attributes are scored on seven-point scales ranging from 'best' (the most desirable value for the attribute) to 'worst' (the least desirable value).

normative definition of rational choice also provides a reasonable model of actual decision behavior.

Mathematical (Rational) Decision Models

Two classic rational choice models of decision behavior are the weighted additive value model (WADD) and the subjective expected utility model. The WADD model is often used to represent in a mathematical form the trading-off process in decision-making. A measure of the relative importance (weight) of an attribute is multiplied by the attribute's value for a particular alternative and the products are summed over all attributes to obtain an overall value for that alternative, WADD(X):

$$\text{WADD}(X) = \Sigma\, W_i V(X_i) \text{ for } i = 1 \text{ to } n \qquad (1)$$

where W_i is the weight given to attribute i, $V(X_i)$ is the value of option X on attribute i, and n is the total number of relevant attributes. The WADD model represents a normative procedure for dealing with multiattribute decision problems (see Keeney and Raiffa, 1976) in that it uses all the relevant information, explicitly resolves conflicting values, and selects the option with the highest overall evaluation. Note that it also assumes that the effects of the attributes are independent (i.e., there are no interactions), which may not always be appropriate. The WADD model underlies many of the techniques used by economists, market researchers, and others to assess preferences.

A model for making choices under risk and uncertainty is the subjective expected utility (SEU) model. The subjective expected utility of a risky option (gamble) is given by

$$\text{SEU}(X) = \Sigma\, S(P_i)U(X_i) \text{ for } i = 1 \text{ to } n \qquad (2)$$

where $S(P_i)$ is the subjective probability of occurrence for outcome X_i, and $U(X_i)$ is the utility to the individual of receiving amount X_i, e.g., an amount of money. Note that the SEU model is similar in structure to the WADD model, weighting each possible outcome by its probability of occurrence, summing those products over all possible outcomes, and then selecting the gamble with the highest SEU.

Cognitive Limitations and Heuristic Decision Processes

Although people sometimes make decisions in ways consistent with the WADD and SEU models, it has become obvious over years of decision research that people often make decisions using simpler decision processes (heuristics), more consistent with the idea that people are, at best, only boundedly rational. The bounded rationality view of decision-making emphasizes a decision-maker's limited information processing capabilities and the interaction of those computational limits with the complexity of task environments: 'human rational behavior is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor' (Simon, 1990: p. 7).

In part because of limitations in information processing capacity, preferences for objects are often constructed at the time people are asked to make choices. That is, people may construct preferences on the spot when needed rather than simply retrieving well-defined values for objects when they are asked to make a choice (March, 1978). In addition, because of limited processing capacity and the need to be adaptive to task demands, decisions are often made not by using some invariant rule such as SEU, but rather by using a variety of decision strategies contingent upon the demands of the task environment (Payne *et al*., 1993). Such contingency upon a variety of task and context factors is consistent with and implied by the constructive nature of preferences. Finally, although the concept of bounded rationality does not mean that people are poor decision-makers, the combination of limited cognitive capabilities and difficult decision problems means that people sometimes make systematic errors when facing choice problems. For example, people will sometimes say that they prefer option A to option B, B to option C, and C to A (Tversky, 1969), or that they prefer A to B but would be willing to pay more for B than for A (Tversky *et al*., 1988).

Several commonly used decision-making heuristics have been identified. If a lexicographic strategy (LEX) is used, the alternative with the best value on the most important attribute is simply selected (assuming that there are no ties on this attribute); for example, if a decision-maker faced with the choice problem in Table 1 thought that reliability was the most important attribute for cars, he or she could use a lexicographic strategy, examine reliability (and no other information) for all five cars, and choose car B. If two alternatives have tied values, the second most important attribute is considered, and so on until the tie is broken. The LEX strategy is a choice heuristic, consistent with the bounded rationality notion that limited capacity for processing information implies that people generally cannot process all of the available information in a particular situation and must therefore be selective in what information is used. Even when the amount of available information is within the bounds of processing capacity, the processing of that information imposes cognitive costs. Thus, selective processing of information is generally necessary, and the information that is selected for processing will have a major impact on choice. Under some task conditions, for instance, the LEX choice heuristic can be almost as accurate a decision rule as more information-intensive strategies like WADD (Payne *et al*., 1993). Finally, the LEX strategy is a good example of a conflict-avoiding decision strategy that may minimize the emotional aspects of making a decision, because the focus is on only a single attribute at a time. The LEX strategy is a form of 'one reason' decision-making emphasized by Gigerenzer *et al*. (2001).

Satisficing (SAT) is another classic strategy for coping with bounded rationality. With a satisficing strategy, alternatives are considered sequentially, in the order in which they occur in the choice set. The value of each attribute for the option currently under consideration is compared to a predetermined cutoff level for that attribute. If any attribute fails to meet the cutoff level, the option is rejected and the next option is considered. For example, car A might be eliminated rapidly because it has the worst level of reliability. The first option passing the cutoffs for all attributes is selected. If no option passes all the cutoffs, the levels can be relaxed and the process repeated. Like the LEX strategy, satisficing does not involve explicitly considering tradeoffs and is therefore a noncompensatory model of decision-making – that is, a good value on one attribute cannot compensate for a below cutoff (poor) value on another attribute. One of the key differences across choice processes is the extent to which a compensatory (e.g., WADD) or noncompensatory decision process is utilized. Another property of the SAT strategy is that the option chosen will be a function of the order in which the options are processed. Thus, one can potentially influence choice by structuring the order in which options are considered.

Elimination by aspects (EBA) is a commonly used choice heuristic combining elements of both the LEX and SAT strategies. It eliminates options that do not meet a minimum cutoff value for the most important attribute. This elimination process is repeated for the second most important attribute, with processing continuing until a single option remains (Tversky, 1972). In our car example, suppose that the decision-maker's two most important attributes were reliability and safety, in that order, and that the cutoff for each was an average value. This individual would first process reliability, eliminating any car with a below-average value (cars A, C, and E). Then the person would consider safety for cars B and D, eliminating car B. Hence, car D would be selected. Elimination by aspects focuses on the attributes as the basis for processing information, is noncompensatory, reflects rationality in the ordered use of the attributes, and does not use all potentially relevant information. The extensiveness and selectivity of processing will vary when using EBA depending upon the exact pattern of elimination of options. As a general rule, the preferences expressed when using choice heuristics like LEX, satisficing, and EBA are subject to potentially 'irrelevant' task variables such as the order in which options and/or attributes are considered.

Decision-makers also use combinations of choice strategies. A typical combined strategy has an initial phase in which some alternatives are eliminated, and a second phase where the remaining options are analyzed in more detail. One frequently observed combination is an initial use of EBA to reduce the choice set to two or three options, followed by a compensatory strategy such as weighted adding to select among those. An important implication of the use of combined strategies is that the 'properties' of the choice task itself may change as the result of using a particular strategy first. For example, the use of a process for eliminating dominated alternatives from a choice set, an often advocated procedure, will make the conflict among attribute values more extreme, perhaps then triggering the application of a new strategy on the reduced set of options.

TASK AND CONTEXT EFFECTS

Constructed Preferences

Research supports the idea that people use choice heuristics and the more general constructive view of preferences. For example, people use a variety of different strategies to solve choice problems contingent upon the nature of the task, e.g. how many options are available, and the context of the choice problem, e.g. is the choice among a set of generally good or generally poor options. Task factors are general characteristics of a decision problem, such as number of alternatives available, response mode (e.g. choice or judgment), time pressure, or information format, that do not depend on the particular values of the alternatives. One well-replicated and important task effect is that people use compensatory (trade-off based) types of decision strategies (e.g. WADD) when faced with a decision problem involving only two or three alternatives. However, when faced with a more complex (multialternative) decision task, people tend to use noncompensatory strategies such as EBA. People also tend to use more noncompensatory strategies when asked to make choices rather than judgments, and when choosing under time pressure.

Context factors, such as the similarity of alternatives, are associated with the particular values of the alternatives in a choice set and the relationships among those values. Context variables affect both how decisions are made and which option is chosen from a set of alternatives. For example, the more negative the correlations among the attribute values (i.e., the more one has to give up something on one attribute to obtain more of another attribute), the more likely it is that people will use a WADD (expected value) strategy when choosing among gambles. Interestingly, when the attributes are more emotional, such as safety, greater conflict among attribute values tends to lead to greater use of strategies like the LEX rule (Luce *et al*., 2001).

A context effect that shows how the nature of the choice set can influence which option is chosen is illustrated by the problem in Table 2. Comparing options A and B, one is faced with a trade-off between ride quality (better with A) and fuel consumption (better with B). A basic principle of most choice models, regularity, states that adding a new option to the choice set containing A and B cannot increase the probability of choosing one of the original options. However, imagine that you are faced with the choice of A, B, or C rather than just A or B. Note that C is 'dominated' by option B: B is better than C on one attribute (ride quality) and at least equal to C on the other attribute (fuel economy). Option C is not dominated by A, so that in comparing options A and C there would still be a trade-off to be considered. Thus, there is an asymmetric dominance relationship among the three options A, B, and C. Given that C is dominated by B, it is highly unlikely that a person would select C. However, does the presence of C influence the choice between A and B? The answer is yes. Adding option C to the original choice set of A and B tends to increase the probability of selecting B (Huber *et al*., 1982). This increase in the probability of choosing the dominating option B with the addition of C violates the principle of regularity. Such a context effect indicates that 'people do not maximize a precomputed preference order, but construct their choices in the light of available options' (Tversky, 1996: p. 17).

Reason-based Choice

A number of explanations have been offered for the asymmetric dominance effect. One that has received support is that people use the relations among options as reasons for justifying their choices; that is, one can easily see that B is better than C, and this relationship provides a good reason for choosing B and not A. The size of the asymmetric dominance effect increases when people are asked to explicitly justify their choices (Simonson, 1989). Bettman *et al*. (1998) speculated that the fact that option B dominates option C is a good reason for choosing option B at an outcome level of explanation (it is clearly a better outcome than C). There is no need to refer to a process-level explanation (e.g. I chose B over A because of the trade-offs I prefer between ride quality and fuel economy). Outcomes are likely to be more salient than process in the wake of a decision, so arguments based on outcomes may provide better reasons.

Using easily detected relationships among the options in a choice set as a reason for choice allows a person to avoid cognitively and emotionally difficult trade-offs. For example, selecting the option in a set of three that is between the more extreme

Table 2. An example of an asymmetric dominance task

Car	*Ride quality*	*Fuel consumption (miles per gallon)*
A	83	24
B	73	33
C	70	33

options can be justified as a 'compromise' choice without having to explicitly consider trade-offs. For more on a reason-based view of choice processes, see Shafir *et al*. (1993).

CONCLUSION

The study of choice processes has been a subject of long-standing interest for psychologists, economists, and researchers in many other fields. How people make choices frequently departs from a purely rational decision process, reflecting the interplay of cognitive processing limits and task demands. People use a variety of simplifying choice heuristics and search for easy-to-justify reasons for preferring one option over another. As a consequence, the option people choose can be influenced by a variety of predictable task and context factors. More generally, in many situations the preferences people exhibit are constructed at the time of choice rather than reflecting pre-computed values.

While much has been learned about choice behavior, there is still much to be learned about how decisions are made and how decision-making can be improved.

References

Bettman JR, Luce MF and Payne JW (1998) Constructive consumer choice processes. *Journal of Consumer Research* **25**: 187–217.

Gigerenzer G, Todd PM and the ABC Research Group (2001) *Simple Heuristics That Make Us Smart*. New York, NY: Oxford University Press.

Huber J, Payne JW and Puto CP (1982) Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis. *Journal of Consumer Research* **9**: 90–98.

Keeney RL and Raiffa H (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York, NY: John Wiley.

Luce MF, Bettman JR and Payne JW (2001) Emotional decisions: tradeoff difficulty and coping in consumer choice. *Monographs of the Journal of Consumer Research*.

March JG (1978) Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics* **9**: 587–608.

Payne JW, Bettman JR and Johnson EJ (1993) *The Adaptive Decision Maker*. Cambridge, UK: Cambridge University Press.

Shafir E, Simonson I and Tversky A (1993) Reason-based choice. *Cognition* **49**: 11–36.

Simon HA (1955) A behavioral model of rational choice. *Quarterly Journal of Economics* **69**: 99–118.

Simon HA (1990) Invariants of human behavior. *Annual Review of Psychology* **41**: 1–19.

Simonson I (1989) Choice based on reasons: the case of attraction and compromise effects. *Journal of Consumer Research* **16**: 158–174.

Tversky A (1969) Intransitivity of preferences. *Psychological Review* **76**: 31–48.

Tversky A (1972) Elimination by aspects: a theory of choice. *Psychological Review* **79**: 281–299.

Tversky A (1996) Contrasting rational and psychological principles in choice. In: Zeckhauser RJ, Keeney RL and Sebenius JK (eds) *Wise Choices: Decisions, Games, and Negotiations*, pp. 5–21. Boston, MA: Harvard Business School Press.

Tversky A, Sattath S and Slovic P (1988) Contingent weighting in judgment and choice. *Psychological Review* **95**: 371–384.

Further Reading

Baron J (2001) *Thinking and Deciding*. Cambridge, UK: Cambridge University Press.

Goldstein WM and Hogarth RM (1997) Judgment and decision research: some historical context. In: Goldstein WM and Hogarth RM (eds) *Research on Judgment and Decision Making: Currents, Connections, and Controversies*, pp. 3–68. Cambridge, UK: Cambridge University Press.

Hastie R (2001) Problems for judgment and decision making. *Annual Review of Psychology* **52**: 653–683.

Hastie R and Dawes RM (2001) *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Thousand Oaks, CA: Sage.

Kahneman D and Tversky A (2000) *Choices, Values, and Frames*. New York, NY: Cambridge University Press.

Plous S (2001) *The Psychology of Judgment and Decision Making*. New York, NY: McGraw-Hill.

Cognitive Development, Computational Models of

Intermediate article

Denis Mareschal, Birkbeck College, London, UK

CONTENTS

Computational models of cognitive development are formal models of the information processing, and changes in information processing, that take place during cognitive development. They are generally implemented as running computer simulations. They are tools for exploring mechanisms of transition (development) from one level of competence to the next during the course of cognitive development.

COGNITIVE DEVELOPMENT AND THE COGNITIVE SCIENCES

Jean Piaget, the father of cognitive development research, saw himself as an empirical philosopher. His goal was to answer the fundamental questions of epistemology through rigorous experimentation. He asked how knowledge (especially abstract conceptual knowledge and logic-based reasoning) could emerge from a child's interactions with the world. Piaget produced a vast body of work exploring the development of cognitive components such as the concepts of Space, Time, Number, and Causality. He is generally recognized as having identified the key questions that have set the agenda for cognitive development research since the 1950s.

Piaget was greatly influenced by the philosophies of Kant and Bergson, and by the Cybernetics movement of the early twentieth century. He believed that children constructed an understanding of the world through active engagement with the world and that feedback played a crucial role in learning and development. Unfortunately, he failed to ground many of his proposals because he lacked an appropriate vocabulary with which to express his dynamic and mechanistic ideas. The advent of cognitive science has provided powerful conceptual tools for addressing many of Piaget's original queries.

Contemporary theories of cognitive development lie along two distinct – albeit related – dimensions. One of these is the Nativist vs. Empiricist dimension (i.e. the nature vs. nurture debate). Radical Nativists believe that all knowledge is available to the infant prior to any learning experiences. Radical Empiricists believe that the infant is born with no prior knowledge, and that all knowledge is acquired through some form of experience with the world. Although no developmentalist will admit to holding either of these extreme views, the field is nevertheless polarized into two camps with strong, sometimes extreme, biases towards one or the other of these poles.

A second, more recent, dimension is the symbolic vs. subsymbolic dimension. Those in the symbolic camp believe that cognition is best characterized as a rule-governed physical symbol system. In this view, cognitive development consists in the modification rules. Those in the subsymbolic camp see cognition as a highly interactive dynamic system (e.g. an artificial neural network) that does not operate as a symbol processing system. In this view development consists in the continuous tuning of the underlying parameters of the cognitive system.

The rest of this article presents a number of models that illustrate the different developmental domains in which modeling has been undertaken. It also illustrates fundamental differences in the developmental mechanisms proposed. As will become apparent in this review, most symbolic models have emphasized the tractability of the knowledge representations involved in cognitive development at the expense of implementing explicit transition mechanisms. In contrast most subsymbolic models have emphasized the specification of a developmental mechanism at the expense of the tractability of the knowledge representations.

WHY BUILD COMPUTATIONAL MODELS OF COGNITIVE DEVELOPMENT

The Computer Modeling Methodology

Computer models complement experimental data gathering by placing constraints on the direction of future empirical investigations. First, developing a computer model forces the user to specify precisely what is meant by the terms in his or her underlying theory. Terms such as representations, symbols, and variables must have an exact definition to be implemented within a computer model. The degree of precision required to construct a working computer model avoids the possibility of arguments arising from the misunderstanding of imprecise verbal theories.

Secondly, building a model that implements a theory provides a means of testing the internal self-consistency of the theory. A theory that is in any way inconsistent or incomplete will become immediately obvious when trying to implement it as a computer program. The inconsistencies will lead to conflict situations in which the computer program will not be able to function. Such failures point to a need to reassess the situation and re-evaluate the theory.

A positive corollary of these two points is that the model can be used to work out unexpected implications of a complex theory. Because the world is highly complex with a multitude of information sources constantly interacting, even a simple process theory can lead to uninterpretable behaviors. Here again, the model provides a tool for teasing apart the nature of these interactions and corroborating or falsifying the theory.

Perhaps the main contribution made by computational models of cognitive development is to provide an account of the representations that underlie performance on a task that also incorporates a mechanism for representational change. One of the greatest unanswered questions of cognitive development is the nature of the transition mechanisms that can account for how one level of performance is transformed into the next level of performance at a later age. This is a difficult question because it involves observing how representations evolve over time and tracking the complex interactions between the developing components of a complex cognitive system. Building a model and observing how it evolves over time provides a tangible means of doing this.

Formulating development in computational terms forces the theoretician to be explicit about the mechanisms that underlie information processing. Piaget's own mechanistic theory provides an excellent example of why this is necessary. He described cognitive development in terms of assimilation, accommodation, and equilibration. Assimilation consisted in adapting or filtering incoming information to make it more compatible with existing knowledge representations. In contrast, accommodation consisted in adapting one's knowledge representations to make them more consistent with novel information. Equilibration was the process by which assimilation and accommodation interacted to cause cognitive development. While assimilation and accommodation capture intuitive notions of what might be involved in cognitive development, they are too loosely defined to be of any explanatory value. Some connectionist computational models of cognitive development have tried to solve this problem by providing computational implementations of assimilation and accommodation in terms of activation flow and weight updates respectively.

MODELS OF DEVELOPMENT IN INFANCY

Infancy is an ideal age range to begin modeling because infant behaviors are not complicated by the presence of language and sophisticated metacognitive strategies. Infant abilities are closely tied to their developing sensorimotor skills.

Object-Directed Behaviors

Kant identified objects as a fundamental category of cognition. The ability to represent hidden objects liberates infants from the tyranny of direct perception. It is the first step towards representational thought. Piaget suggested that infants progressed through six stages on the way towards an adult level of understanding of object permanence at the age of two. While many of Piaget's original findings have been replicated, changes in methodology (e.g. relying on visual attention measures rather than manual retrieval measures) have suggested that infants have a far more precocious understanding of hidden objects. These studies have all focused on infant competence at different ages but not on the mechanisms of development from one level of competence to the next.

There are relatively few computational models of infant object-directed behaviors (Mareschal, 2000). Early models took a strong cognitivist stance on behavior and were thus implemented in rule-based production systems (e.g. Luger *et al*., 1983).

Unfortunately, they were basically competence models that described infant behaviors but did not provide a mechanistic account of development. They proposed different sets of rules to describe behavior at different ages but did not explain how new rules could be acquired or how one set of rules was transformed into another set of rules. More recent (cognitivist) models have turned to attention-based accounts of object processing in an attempt to explain infant behaviors (Simon, 1998). Unfortunately, these models still fail (by and large) to implement any account of *how* development might occur.

One mechanistic learning model has implemented a parallel processing version of Piaget's sensorimotor theory of infant development. Drescher (1991) tried to show how the coordination of intra- and intermodal perceptual motor schemas could lead to a single unified representation of an object. Perceptual motor schemas were encoded as 'context-action-result' rules and implemented in a parallel processing machine. Learning consisted in using marginal probabilities to fill in context and results slots in appropriate perceptual motor schemas. Although this system developed an intricate network of intra- and intermodal schemas that mimic the infant's sensorimotor integration, it did not develop according to the pattern described by Piaget.

A number of connectionist models have also been proposed. In one family of models, a partially recurrent autoencoder network learns to predict the reappearance of a stationary object from behind a moving screen that temporarily occludes the object (Munakata *et al*., 1997). Network performance is measured by taking the difference in response of the nodes coding the location of the hidden object when an object should be revealed and subtracting it from the response of the nodes when an object should not be revealed. An increase in this difference is interpreted as increased knowledge of hidden objects. What this model shows is that object representations that guide action can be graded and arise though interactions within an environment.

Mareschal *et al*. (1999) describe an alternative connectionist model that is more closely tied to the neuropsychological finding that visual object information is processed down two separate routes. This model uses a combination of modules to implement dual route processing. One route learns to process spatial-temporal information while the other route learns to process feature information. Finally, a response module recruits and coordinates the representations developed by the other modules as and when required by a response task. The route specializations emerge as a result of the different associative mechanism in each module.

Perceptual Categorization

Because categorization lies at the heart of cognition it is not surprising to find that great effort has been exerted in trying to understand the early roots of category formation. Many infant categorization tasks rely on preferential looking or habituation techniques, based on the finding that infants direct more attention to unfamiliar or unexpected stimuli (Mareschal and Quinn, 2001). Connectionist autoencoder networks have been used to model the relation between sustained attention and representation construction (Mareschal *et al*., 2000). The successive cycles of training in the autoencoder are an iterative process by which a reliable internal representation of the input is developed. This approach assumes that infant looking times are positively correlated with the network error. The greater the error, the longer the looking time, because it takes more training cycles to reduce the error.

The perceptual categories formed by infants are not always the same as the corresponding adult categories. For example, when shown a series of cat photographs three- to four-month-olds will form a category of CAT that includes novel cats and excludes dogs (as will adults). However, when shown a series of dog photographs, the same infants will form a category of DOG that includes novel dogs but also includes cats (in contrast to adults). Many aspects of early infant perceptual categorizations (including the asymmetric exclusivity of CAT and DOG categories) are captured by the connectionist autoencoder model. In contrast to adults who apply top-down schemas when recognizing photographs of cats and dogs, three- to four-month-olds, like the autoencoder networks, simply process the bottom-up information in these images. Hence, their internal category representations are yoked to the distributional properties of features in the images. The model demonstrates that categorical representations can self-organize in a neural system as a result of exposure to the familiarization exemplars encountered within the test session itself.

Early Word Learning

Categorization is equally important in early language acquisition, both in terms of learning which sounds in the environment are relevant to speech

and in terms of learning the domain of applicability of a new word.

Infants are better at discriminating phonemes that do not belong to their linguistic environment at seven months than at 14 months. This has been interpreted as evidence for a shift in lexical processing between these age groups. The suggestion is that seven-month-olds are processing the sounds of the stimuli whereas the 14-month-olds are processing the stimuli as words. The process of word learning itself changes the way the stimuli are processed. In contrast, Schafer and Mareschal (2001) present a connectionist autoencoder model suggesting that these behaviors reflect two stages of the same processing mechanism. The model learns to associate sound representations with image representations (i.e. simple word learning). Early in training, the model is not committed to any meaningful internal representations for sound and is therefore better able to learn novel speech sounds in testing than later in training when it has firmly committed to representations tailored to its native linguistic environment. This model illustrates how discontinuities in behavior can emerge by the slow tuning of continuous parameters.

Once infants have segmented the sound stream into word units, there is the further problem of identifying the extension of the category referred to by the word (the Symbol Grounding Problem). One early model (Schyns, 1991) used a combination of self-organizing (kohonen feature maps) and supervised (backpropagation) connectionist networks to model the interactive process of identifying the category underlying the word, and then associating it with an appropriate label. This model showed many of the same developmental effects as young children learning novel words. In particular, the categories showed prototype effects and mutual exclusivity constraints. It demonstrated how the acquisition of a new word arose as an interaction of bottom-up and top-down effects. A more recent (autoencoder) model (Plunkett *et al*., 1992) shows how over extension and under extension errors, typical of children's early vocabulary, can arise through simple associative mechanisms in a system with shared sound and image internal representations.

MODELS OF DEVELOPMENT IN CHILDHOOD

Language acquisition marks the end of infancy and the beginning of childhood. Reasoning and conceptual development are the hallmark of cognitive development in childhood. The models in this section all focus on some aspect of reasoning development. We begin by reviewing models that have explicitly tried to implement Piagetian ideas. This is followed by a review of work that breaks away from the Piagetian tradition.

Modeling Piagetian Stage Development

There have been several attempts to explain the apparent stage-like growth of competence in children in terms of self-organization in dynamic systems, competition between cognitive growers, and bifurcation theory (e.g. van Geert, 1998). However such accounts have tended to rely only on mathematical descriptions that are either not implemented in running computer models or grounded in measurable information processing components.

As early as the 1960s an effort was made to use neural networks to operationalize the Piagetian notions of assimilation and accommodation. Several interpretations of connectionist learning in terms of assimilation and accommodation have been proposed. One such interpretation suggests that weight changes constitute a form of accommodation whereas the transformation (by the network weights) of input patterns into internal patterns of activation constitutes assimilation (McClelland, 1995). Another interpretation suggests that the adaptation of a network architecture constitutes a form of accommodation whereas the adaptation of weights is a form of assimilative learning (Shultz *et al*., 1995).

Models that try to implement Piagetian notions of development have attempted to model children's performance on key tasks (e.g. conservation: Klahr and Wallace, 1976; Shultz, 1998). One such task is the seriation (or sorting) task. Piaget found that children's ability to order a set of sticks developed through a number of stages. In a first stage, children were unable to sort the sticks. In a second stage, they were able to apply local ordering relations but could not extend the order to the set as a whole. In the third stage, they were able to sort the set of sticks, but only by applying a costly trial and error strategy. Finally, in the fourth stage children were able to sort the set quickly and efficiently by applying a systematic selection strategy.

Young (1976) approached this task from an information processing perspective. He carried out detailed analyses of the actions children carried out at different ages when sorting blocks. Based on the results of protocol analyses, he developed a rule-based production system that captured children's performance at each stage of development.

Progress from one stage to the next was modeled by the (hypothesized) modification of the rules. Although this model provided a good fit to children's behavior at individual stages, it fails to provide a working account of how those rules are modified. A recent connectionist model of the seriation task suggests how development could occur (Mareschal and Shultz, 1999). This model argues that development consists in the gradual tuning of connection weights, and the gradual extension of knowledge about small sets to larger sets. The model not only captures the stage progression described by Piaget, but it also captures the variability in sorting behaviors observed both within and between subjects.

Beyond Piaget: The Balance Scale Task

A recent benchmark of cognitive development, first developed by Inhelder and Piaget and later significantly extended by Robert Siegler, is the balance scale task. Siegler explored children's developing abilities to reason about balance scales. In these problems, children were presented with a symmetric balance scale with five equally spaced pegs on either side of the fulcrum. Weights were then placed on pegs to the left and the right of the fulcrum and children were asked to predict whether the balance scale would tip to the left, to the right, or remain balanced.

Seigler used a rule assessment methodology to infer the rules that govern children's performance. He found that rule one children relied only on a dominant dimension (weight) to predict which side the balance scale would tip. They predicted that the side with the most weights would be the side that the balance scale would tip. Rule two children applied the same rule as the preceding children, but had an additional rule stating that if the number of weights was equal on both sides, the side with the greatest distance would predict the side to which the balance scale would tip. Rule three children behaved like the rule two children with the exception that if the weight and distance cues provided conflicting answers they would guess which side went down. Finally, rule four children had a set of rules that effectively computed the torque on both sides of the balance scale and chose the side with the greatest torque. Seigler suggested that this knowledge was represented in the form of a growing decision tree. Klahr (1992) pointed out the equivalence of this representation to that of an increasing rule; a representation consistent with production system models of cognitive development.

Both connectionist (McClelland, 1995; Shultz *et al*., 1994) and decision tree models (Schmidt and Ling, 1996) of development on the balance scale task have been proposed. The connectionist models construe the problem as one of integrating information from two sources. They capture stage development in terms of microgenetic weight changes and hidden unit recruitment. An assumption of these models is that children have greater experience with weight comparisons than distance comparisons. The decision tree model uses the C4.5 tree-inducing algorithm. Children were hypothesized to have an increasing memory capacity and to care increasingly about the detail of correctness of their answers. While this latter model captures the main features of children's performance, the decision trees developed did not map onto those proposed by Siegler to reflect children's knowledge at different ages.

Modeling of the Development of Reasoning

Many of the successful developmental models described above are connectionist models. Such models process information based on the surface similarity between different exemplars. However, there are cases when children's (and adults') reasoning does not follow surface similarity. Analogical reasoning requires the child to distance his or herself from the surface similarity between the target and vehicle domains. Indeed between six and nine years of age children move from basing their analogies on surface similarity (such as color) between the two domains to structural similarity (such as the function of an object) between the two domains. This reliance on structural similarity is very difficult for connectionist systems to capture.

Gentner has suggested that adults and children solve analogical problems by comparing mental representations via a structure-mapping process of alignment of conceptual representations. A structurally consistent match conforms to a one-to-one mapping constraint between the domains. The process is implemented in the Structural Mapping Engine (SME) model. The SME is used to model the relational shift in children's analogical reasoning in terms of increased domain knowledge. As their knowledge of domain relations increases, children's relation representations within a domain become richer and deeper, increasing the likelihood that their comparisons will focus on matching relations. Thus, what develops between six and nine years is only knowledge and not processing.

Siegler and Shipley present a model of strategy choice. The intention is to provide some account of the range and variability of strategies observed in young children's problem solving. Their model is based on the strategies used by children when adding integers. The strategies are explicitly represented in terms of rules. Strategy choice is probabilistic. The probability of retrieving and executing a strategy depends on the previous association of that strategy with an outcome in conjunction with considerations of cost and efficiency. The strategy pool evolves according to a Darwinian procedure in which infrequently used strategies die off and new strategies enter the pool via random perturbations of existing strategies.

CHALLENGES TO CURRENT MODELS OF COGNITIVE DEVELOPMENT

The 'poverty of the stimulus' argument was most effectively put forward by Chomsky in response to Skinner's account of language acquisition. It is not possible for an unconstrained inductive learner to acquire a particular target grammar within a reasonable time. This argument has been wielded against connectionist models of language and cognitive development in general. However, it is important to understand that the 'poverty of the stimulus' argument holds for all inductive learning systems and in all domains (its application is not unique to connectionism). This is why most contemporary scholars of learning (whether studying learning in children or machines) believe that the key to understanding cognitive development is to identify the nature of the constraints on the learner that will allow knowledge to emerge.

Fodor's paradox claims that an inductive learner can never acquire any truly novel concept (Fodor, 1980). Indeed, in order to test the domain of applicability of a concept, the inductive learner must be able to represent that concept prior to having identified it. Hence, any learning simply involves the recombination of existing representational tokens in a system. While this may be true of inductive learning systems, it is not true of systems that increase their representational power in response to environmental pressures. Such systems include neural networks that construct their own architecture as part of learning and development (Mareschal and Shultz, 1996).

Finally, while computer models allow us to formulate questions about what can possibly cause cognitive development (e.g. processing capacity, processing speed, knowledge, and strategy choice) more constraints are required to identify the actual mechanisms involved in children's cognitive development. Recent advances in neuroimaging techniques have allowed us to place greater constraints on how information is processed in the brain. Many of the models above make little use of these constraints and, in the future, such constraints should be incorporated in any functional models of development. Furthermore, cognitive development does not occur in a social vacuum. Vygotsky has emphasized the role of social interactions in cognitive development. Society provides a kind of cognitive scaffolding that nurtures and aids the child's cognitive development by actively selecting and filtering the type of problems the child is faced with at any age. Future models will need to consider these constraints to reflect the child's learning environment more accurately.

References

Drescher GL (1991) *Made-up Minds. A Constructivist Approach to Artificial Intelligence*. Cambridge, MA: MIT Press.

Fodor J (1980) Fixation of belief and concept acquisition. In: Piatelli-Palmarini M (ed.) *Language and Learning: The Debate between Chomsky and Piaget*, pp. 143–149. Cambridge, MA: Harvard University Press.

van Geert P (1998) A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review* **105**: 634–677.

Klahr D (1992) Information processing approaches to cognitive development. In: Bornstein MH and Lamb ME (eds) *Developmental Psychology: An Advanced Textbook*, 3rd edn, pp. 273–335. Hillsdale, NJ: Lawrence Erlbaum Associates.

Klahr D and Wallace JG (1976) *Cognitive Development: An Information Processing View*. Hillsdale, NJ: Erlbaum.

Luger GF, Bower TGR and Wishart JG (1983) A model of the development of the early object concept. *Perception* **12**: 21–34.

Mareschal D (2000) Infant object knowledge: current trends and controversies. *Trends in Cognitive Science* **4**: 408–416.

Mareschal D and Quinn PC (2001) Categorisation in Infancy. *Trends in Cognitive Science* **5**: 443–450.

Mareschal D and Shultz TR (1996) Generative connectionist networks and constructivist cognitive development. *Cognitive Development* **11**: 571–604.

Mareschal D and Shultz TR (1999) Children's seriation: a connectionist approach. *Connection Science* **11**: 153–188.

Mareschal D, French RM and Quinn P (2000) A connectionist account of asymmetric category learning in infancy. *Developmental Psychology* **36**: 635–645.

Mareschal D, Plunkett K and Harris P (1999) A computational and neuropsychological account of object-oriented behaviors in infancy. *Developmental Science* **2**: 306–317.

McClelland JL (1995) A connectionist perspective on knowledge and development. In: Simon TJ and Halford GS (eds) *Developing Cognitive Competence: New Approaches to Process Modeling*, pp. 278–304. Hillsdale, NJ: Erlbaum.

Munakata Y, McClelland JL, Johnson MH and Siegler RS (1997) Rethinking infant knowledge: towards an adaptive process account of successes and failures in object permanence tasks. *Psychological Review* **104**: 686–713.

Plunkett K, Sinha C, Møller MF and Strandsby O (1992) Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science* **4**: 293–312.

Schafer G and Mareschal D (2001) Modeling infant speech sound discrimination using simple associative networks. *Infancy* **2**: 7–28.

Schmidt WC and Ling CX (1996) A decision-tree model of balance scale development. *Machine Learning* **18**: 1–30.

Schyns P (1991) A modular neural network model of concept acquisition. *Cognitive Science* **15**: 461–508.

Shultz TR (1998) A computational analysis of conservation. *Developmental Science* **1**: 103–126.

Shultz TR, Mareschal D and Schmidt WC (1994) Modeling cognitive development on balance scale phenomena. *Machine Learning* **16**: 59–88.

Shultz TR, Schmidt WC, Buckingham D and Mareschal D (1995) Modeling cognitive development with a generative connectionist algorithm. In: Simon TJ and Halford GS (eds) *Developing Cognitive Competence: New Approaches to Process Modeling*, pp. 205–261. Hillsdale, NJ: Erlbaum.

Simon TJ (1998) Computational evidence for the foundations of numerical competence. *Developmental Science* **1**: 71–78.

Young R (1976) *Seriation by Children: An Artificial Intelligence Analysis of a Piagetian Task*. Basel: Birkhauser.

Further Reading

Boden MA (1995) *Piaget*, 2nd edn. Modern Masters. London, UK: Fontana Press.

Elman JL, Bates EA, Johnson MH *et al*. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Karmiloff-Smith A (1992) *Beyond Modularity*. Cambridge, MA: MIT Press.

Lewandowsky S (1993) The rewards and hazards of computer simulations. *Psychological Science* **4**: 236–243.

Rogoff B (1990) *Apprenticeship in Thinking*. Oxford, UK: Oxford University Press.

Siegler RS (1996) *Emerging Minds*. Cambridge, MA: MIT Press.

Siegler RS (1997) *Children's Thinking*, 3rd edn. Prentice Hall.

Simon TJ and Halford GS (eds) (1995) *Developing Cognitive Competence: New Approaches to Process Modeling*. Hillsdale, NJ: Erlbaum.

Cognitive Science: Experimental methods

Introductory article

Raymond S Nickerson, Tufts University, Medford, Massachusetts, USA

CONTENTS

Cognitive activity can be studied in a variety of ways, including observation, simulation by computer modeling, and controlled experimentation.

INTRODUCTION

Cognitive science is a broad topic. It encompasses cognitive – or cognitive-like – activity wherever it is found, in humans, animals or machines. It is studied in a variety of ways, including observation, simulation (notably by efforts to give computers the ability to do things that when done by people are considered cognitive), and controlled experimentation.

PURPOSES OF EXPERIMENTATION

Scientific experimentation involves the investigation of how the controlled manipulation of one or more (independent) variables affects one or more other (dependent) variables.

Experiments are performed for several purposes, including testing hypotheses, establishing the values of parameters of process models, comparing the predictive power of competing theoretical accounts of specific phenomena, and evaluating the effectiveness of operational devices or procedures in realizing the intents of their designers. Sometimes experiments are done for such purposes as investigating hunches that are not sufficiently precise to be treated as testable hypotheses, looking for relationships or regularities that are worthy of more focused study, checking the adequacy of experimental designs, or fine-tuning setups of experimental equipment in anticipation of their use for more formal purposes. Experiments of these types are sometimes referred to as exploratory, pilot or calibration studies, and they are less likely than more formal studies to be reported in scientific journals – but they are experiments nonetheless. Experiments are also sometimes done to demonstrate already known relationships among variables; although these may replicate the conditions of experiments that have already been conducted for purposes of discovery, their purpose is strictly educational.

All these distinctions pertain to experimentation on cognition as well as to experimentation in other fields.

COMPARISON OF METHODS

Laboratory Versus Field Methods

Most studies of cognition take place in university laboratories or classrooms, or in the research facilities of industrial or government organizations. Some, however, are done in the field. An example of laboratory research is an experiment designed to investigate the effects of the spacing of rehearsal on the memorization and recall of verbal material. An example of field research is a study of the effects of a secondary task, such as carrying on a telephone conversation, on how well an automobile driver performs the driving task.

The choice of laboratory or field research in any particular case is likely to involve consideration of a trade-off between control and realism: between precision and applicability to real-world contexts. Generally speaking, it is possible to exercise much greater control over variables – both those whose effects the investigator is interested in studying, and those that are better thought of as nuisance factors – in a laboratory setting than in the field. However, this greater degree of control is usually bought at the price of making the situation so artificial that generalization of the findings from the laboratory to the real world may be difficult or impossible.

For these reasons it is desirable when possible to verify the generalizability of findings obtained in the laboratory to the real-world situations to which

they are believed to apply before drawing firm conclusions about their applicability. The laboratory findings can serve as tentative conclusions that need to be confirmed in the real-world situation of interest. This approach is illustrated by efforts to determine whether what students have learned in a laboratory context designed to teach specific aspects of flying an airplane transfer to performance in an actual flight situation.

Cross-sectional Versus Longitudinal Studies

Most experiments in psychology involve comparisons between measurements that are made at approximately the same time. To determine the immediate effects on hearing of short-term exposure to noise, for example, one might simply determine the ability of people to detect weak auditory stimuli after exposure to noise of different intensities. Sometimes, however, the interest is in how people's abilities (attitudes, values, beliefs) change over long periods.

One experimental approach to the study of such changes is to investigate people of different ages; another is to study the same group of people over many years. The advantage of the first (cross-sectional) approach is that the study can be done quickly; a disadvantage is that people in the different groups are likely to differ in ways other than age (they were born and grew up in different times), which can complicate the interpretation of results. The advantage of the second (longitudinal) approach is the opportunity to see how people change with respect to abilities and other characteristics of interest with the passing of time; a disadvantage is that such studies take many years to perform, during which participants may leave the study, funding may be lost, and so on.

Both cross-sectional and longitudinal studies have proved useful in the study of aging, the findings of the one type complementing those of the other.

Single-case Studies Versus Averaging Over Multicase Samples

The vast majority of experiments on cognition involve the averaging of experimental data over a group of participants. Case studies, however, can sometimes provide extensive and in-depth information about individuals that is generally not available in data from multicase samples. Often case studies are opportunistic in the sense that they capitalize on the occurrence of a unique event or the chance discovery of an individual with an unusual ability. The highly publicized case of Phineas Gage, the railroad worker who in 1848 suffered a horrendous but nonfatal injury when an iron bar used for tamping explosive powder was propelled by an accidental explosion completely through his head, illustrates the first situation. Gage's experience was unusual, if not unique, and documentation of the long-term cognitive and affective effects of his injury contributed to a better understanding of how certain functions depend on specific parts of the brain.

There are many published examples of case studies of individuals with unusual memories. John Dean's recorded recollection of events as given in testimony before the Watergate committee of the US Senate in 1973 provided the basis for one such study; discovery of a man (Hideaki Tomoyori) who was able to recite from memory the first 40 000 digits of pi provided another. Experiments with Tomoyori showed his ability to recall the details of narrative stories to be quite ordinary.

Comparing and Evaluating Models with Empirical Evidence

As in all experimental sciences, the acid test of the tenability of a theory in cognitive science is its ability to predict the results that would be obtained in carefully controlled experiments. In practice, progress occurs in a cyclic fashion. Predictions are derived from a theory. The predictions are tested by experimentation. Depending on the outcomes of experiments, theories may be strengthened and made more precise, or they may be shown to be false or in need of modification. Sometimes experimental results match theoretically derived predictions closely, in which case the theories from which the predictions were derived are considered to have been corroborated – not to have been proved, but to have gained greater credibility. Sometimes the results match the predictions only marginally or not at all. In such a case, if the experiment has been carefully done, the theory from which the prediction was derived might have to be modified or replaced.

A type of theorizing that has been used to great advantage in science involves the building of models of processes of interest. Attempting to build a working model of a process – an artifact that behaves in the same way as the process of interest – is an especially fruitful way to develop an understanding of that process. It has often been pointed out that centuries of observing birds in flight were not nearly as effective a means of

learning about aerodynamics as attempts to build machines that could fly. Efforts to give computers the ability to do things that human beings do with apparent ease has revealed the hidden complexity underlying many human capabilities. The point is illustrated by the history of efforts to give computers the ability to understand natural language. Despite decades of intensive work on this problem, it remains unsolved in any general sense. It is now clear that the complexity of the problem was grossly underestimated when the quest was first engaged, but the effort has revealed much about human language comprehension, answering many questions while raising others that no one knew enough to ask before the effort was made.

The case for computational modeling has been articulated by several psychologists and cognitive scientists. Mathematical and computer models of human performance have been applied to great advantage to the design and study of complex systems in which humans function as hands-on operators or as supervisors of largely automated processes. Modeling typically involves an iterated series of steps: a model is developed that will accommodate experimental data that have already been collected, that model is then used as a basis for making predictions about the outcomes of additional experiments, and the actual outcomes of those experiments are used to modify the model or adjust its parameters so as to increase its predictive power or accuracy. This process of testing and refining can be continued indefinitely, or at least until the model is capable of predicting outcomes with a desired degree of accuracy or it becomes clear that it should be discarded for a qualitatively different one.

CONVERGING METHODOLOGIES

Many factors are relevant to the selection of an experimental methodology: the nature of the phenomenon of interest; the availability of resources (equipment, experimental participants, time); the practicalities of controlling the variables that must be controlled; the types of measurements that are feasible (brain potentials, skin conductance, pupil size, eye fixation, motor response times, verbal responses); the options one has for analyzing experimental data; the kinds of inferences one wishes to draw; and the nature of the population to which one wants to generalize the results. Often it is necessary to make trade-offs, gaining control, for example, at the expense of limited generalizability. Similarly, it typically is much easier for academic researchers to conduct experiments with students as participants than to conduct them with samples that are more representative of the general population, or of nonstudent populations to which they may wish to generalize results.

In general, experimenters tend to think of variance in data, other than that produced by intentional manipulation of the independent variables involved, as 'noise' that must be treated statistically in order to determine the effects of the experimental manipulations. However, one of the ways of limiting this noise (of controlling within-condition variability) is to use relatively homogeneous samples of participants; although this increases the chances of obtaining statistically significant results, it can also preclude the generalizability of the findings to populations that are less homogeneous than the sample studied.

For these and other reasons, it is good that experimentation can be done in a variety of ways. Generally, advances are made when insights into relationships are confirmed by experimental evidence of more than one type. Seldom, if ever, is any question of more than trivial importance, theoretical or practical, settled decisively with a single experiment. The general rule is one of gradual increase in understanding of phenomena of interest resulting from the convergence of evidence from a variety of sources.

Further Reading

Damasio H, Grabowski T, Frank R, Galaburda AM and Damasio AR (1994) The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science* **264**: 1102–1105.

Gopher D, Weil M and Bareket T (1994) Transfer of a skill from a computer game trainer to flight. *Human Factors* **36**: 1–19.

Meyer DE and Kieras DE (1999) Précis to a practical unified theory of cognition and action: some lessons from EPIC computational models of human multiple-task performance. In: Gopher D and Koriat A (eds) *Attention and Performance*, vol. 17, pp. 17–88. Cambridge, MA: MIT Press.

Neisser U (1981) John Dean's memory. *Cognition* **9**: 1–22.

Recarte MA and Nunes LM (2000) Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied* **6**: 31–43.

Color Perception, Psychology of

Intermediate article

Michael D'Zmura, University of California, Irvine, California, USA

CONTENTS

The psychology of color perception concerns color appearance and the visual processing of light spectral information. Researchers in this area seek to determine the relationships among visual stimuli, activity in the human nervous system and the conscious representation of color.

INTRODUCTION

Color is the psychological representation of light spectral properties. Normal human color vision is served by three classes of retinal photoreceptor, which differ in their abilities to respond to photons of varying wavelength. Comparing the responses of the three kinds of photoreceptor to a light provides an estimate of the light's spectral properties: how its energy varies with wavelength. Color depends also on the perceived cause of light reaching the eye, such as emission by a light source or reflection by a surface. Color thus helps to identify objects and their material properties, to the extent that object color remains constant under varying conditions of viewing and that color is handled appropriately by memory and other cognitive systems.

PHOTORECEPTORAL RESPONSES TO LIGHT

The transduction of light energy into neural responses is accomplished by the approximately 5 million cones and 100 million rods arrayed across the retina. Found primarily within the central 10° of the visual field, cones serve photopic vision under viewing conditions when colors are visible. Color vision depends on the responses of more than one spectral class of cone. The spectral sensitivity of a class of cones is determined by the photopigment molecule responsible for photon absorption. Knowledge of cone spectral sensitivities lets one predict the initial chromatic response of the visual system to a light.

Trichromatic Color Matching

That normal human color vision is served by a system with three spectral classes of sensors may be inferred from the results of color-matching experiments. A prototypical experiment uses quasi-monochromatic lights, each possessing energy in just a narrow band of visible wavelengths. The visible spectrum ranges in wavelength from about 400 nm (blue, with ultraviolet lights at shorter wavelengths) to about 700 nm (red, with infrared lights at longer wavelengths). Results show that a single monochromatic test light can be matched in appearance by combining additively three primary lights R, G and B, of appropriate intensity. It is sometimes necessary for one of the primaries to be added to the test light rather than to the other primaries. By varying the wavelength of the test light, one generates three color-matching functions $\bar{r}(\lambda)$, $\bar{g}(\lambda)$ and $\bar{b}(\lambda)$, which describe the intensity of each primary required to match the test of a particular wavelength (Figure 1a). The dependence of such color-matching functions on the arbitrary choice of primary wavelengths led to the development in 1931 by the Commission Internationale de l'Eclairage (CIE) of the XYZ colorimetric system. The $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ color-matching functions in this system have no negative values, and the $\bar{y}(\lambda)$ function is identical to the photopic luminosity function $V(\lambda)$, which describes the relative brightness of monochromatic lights presented at photopic intensity levels (Figure 1b).

Cone Spectral Sensitivities

The three physiological response systems underlying normal color matching are the long (L), medium (M) and short (S) wavelength-sensitive cones, each class with a different photopigment. Their spectral sensitivities, taken in combination with knowledge of the spectral absorption characteristics of the eye's lens and macular pigment,

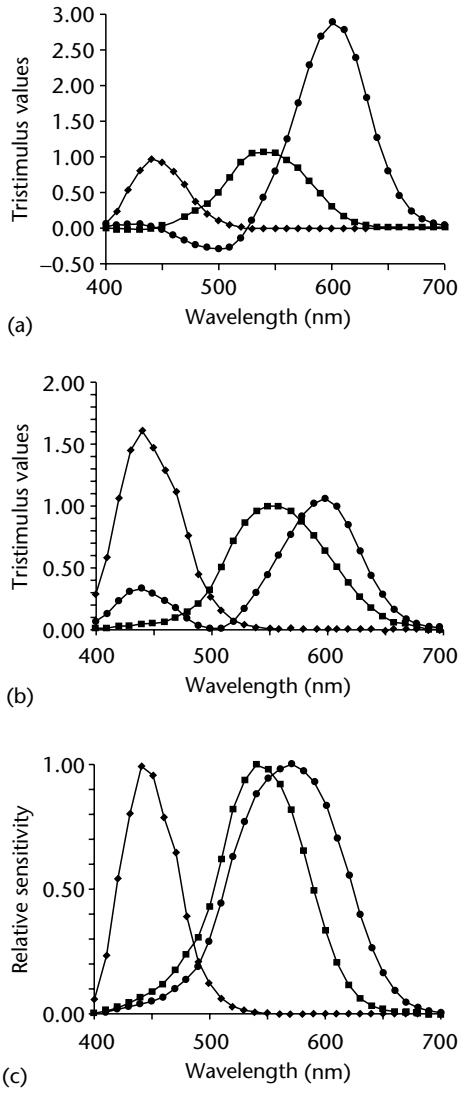


Figure 1. Trichromatic spectral sensitivity. (a) Values of the Stiles 2° color-matching functions $\bar{r}(\lambda)$ (●), $\bar{g}(\lambda)$ (■) and $\bar{b}(\lambda)$ (♦) with monochromatic primaries at 645.2 nm, 526.3 nm and 444.4 nm, plotted at 10 nm intervals. (b) The color-matching functions $\bar{x}(\lambda)$ (●), $\bar{y}(\lambda)$ (■) and $\bar{z}(\lambda)$ (♦) of the Judd-modified CIE 1931 standard 2° observer *XYZ* system. (c) The Stockman and Sharpe (2000) 2° cone fundamentals $L(\lambda)$ (●), $M(\lambda)$ (■) and $S(\lambda)$ (♦). The L and M cones have similar spectral sensitivities; S cones are almost completely insensitive to lights at wavelengths longer than 570 nm.

provide color-matching functions that determine the initial response of the visual system to spectral variation in lights. Rods possess a fourth photopigment, rhodopsin, which differs from those of the three classes of cones, and so potentially support a tetrachromatic response to lights of mesopic intensity.

The first accurate estimates of the cone spectral sensitivities were provided by an analysis of color matching by dichromatic observers (Smith and Pokorny, 1975). A dichromat can match in appearance any light by combining additively just two primaries. A dichromat confuses lights that are distinguishable to trichromatic observers in one of three ways, depending on whether the dichromat lacks L cones (a protanope), M cones (deuteranope) or S cones (tritanope). The color-matching confusions made by the three classes of dichromat let one derive a linear transformation relating cone spectral sensitivities to the CIE system. The psychophysical estimates were soon confirmed by the results of physiological experiments using microspectrophotometric and electrophysiological techniques. Refinements of these estimates based on molecular genetic considerations (Figure 1c) have now been introduced (Stockman and Sharpe, 2000).

Variation Among Individuals

Differences among individuals in chromatic sensitivity can be traced to five primary factors. The first is the pigmentation of the eye's lens, which tends to become more yellow (absorb more light at shorter wavelengths) as one ages and with exposure to ultraviolet light. The second is the amount of macular pigment in an eye. This yellow pigment covers the region of the retina serving central (foveal) vision and varies in density from one person to the next. The third is the numerosity of each class of cone. Anatomical measurements show that S cones form about 7% of the cone population; there are none in the very center of the fovea. There is some controversy about the average ratio of L cones to M cones, which may be about 1.5–2; however, this ratio varies considerably from person to person, and is thought to influence both photopic luminosity, to which S-cone signals do not contribute, and estimates of unique yellow. The fourth factor is photopigment optical density, which influences absolute photoreceptor responsivity. The fifth factor influencing individual chromatic sensitivity is photoreceptor relative spectral sensitivity. There are polymorphisms in the genes lying on the X chromosome which code for the L and M cone opsins (Sharpe *et al*., 1999). There are also hybrid genes which underlie anomalous trichromacy, evident when one of the L or M cone pigments has an abnormal spectral sensitivity. Dichromacy is exhibited when there is only one X-chromosome-linked cone photopigment (that of either the normal L or M cone, or of a hybrid), or when there are polymorphic versions of a single gene. Estimates of the peak spectral sensitivity of the M cone pigment lie in the range 528–532 nm, and

those for the two primary L cone polymorphisms in the ranges 553–558 nm and 557–563 nm. Hybrid genes lead to anomalous M cone peak sensitivities, estimated to lie in the range 529–538 nm and, for anomalous L cones, in the range 545–559 nm. In males of European descent, the behavioral expression of red–green color deficiencies occurs at about a rate of 7.4%; for males of Asian and African descent, the rates are 4.2% and 2.6%, respectively. Owing to the X chromosome placement of the genes for the L and M cones, behavioral expression of red–green color deficiency is much less common among females (about 0.5%).

Models of color perception by dichromats as reduced forms of normal trichromatic perception have been validated by reports of observers who are born with one normal and one color-deficient eye. Rather than perceiving the hues of the visible spectrum in the sequence red, orange, yellow, green, blue, indigo and violet, in passing from long to short wavelengths, both protanopes (1% incidence in males of European descent) and deuteranopes (1.3%) see the sequence yellow, blue, with the transition occurring at a wavelength just shorter than 500 nm. They are able to distinguish lights normally perceived as red and green only in terms of saturation and brightness variations. Tritanopes perceive the sequence red, blue–green, with the transition at about 560 nm; they are unable to discriminate lights at short and middle wavelengths. Inherited tritan defects are rare, because the gene sequence for the opsin component of the S cone photopigment molecule resides on chromosome 7. Acquired defects are more common, because S cone function deteriorates in a variety of medical conditions.

Chromatic Response

The first step in photoreceptoral transduction is the absorption of a photon by a photopigment molecule, which causes a change in the molecule's shape, leading in turn to a change in membrane potential. The likelihood that a photopigment molecule will absorb a photon, as a function of photon wavelength, corresponds to the receptor's spectral sensitivity. The principle of univariance holds that equal numbers of absorbed photons, no matter what their wavelength, generate identical receptoral responses. Univariance underlies the computation of a receptor's response to a light with photons at many wavelengths. In discrete form, one sums over wavelength the product of a photoreceptor's spectral sensitivity $P[\lambda]$ and light energy $E[\lambda]$ to compute the response p:

$$p = \sum_{i=1}^{n} P[\lambda_i]E[\lambda_i] \tag{1}$$

where n is the number of wavelengths at which the light possesses energy. In continuous form, one integrates over wavelength the product of the photoreceptoral spectral sensitivity function $P(\lambda)$ and the function $E(\lambda)$ describing light energy:

$$p = \int P(\lambda)E(\lambda)\mathrm{d}\lambda \tag{2}$$

One often works with spectral sensitivity functions that are tabulated at discrete intervals, for instance, from 400 nm to 700 nm in steps of 10 nm. The 31 numbers describing such a function can be thought of as comprising a 31-dimensional vector $\mathbf{p}$. By describing the light energy in a similar fashion to provide a vector $\mathbf{e}$, one finds that the response p is given by the dot product of the two vectors:

$$p = \mathbf{p}\cdot\mathbf{e} \tag{3}$$

The response of a trichromatic system to a light is represented by three numbers. For instance, one can use the L, M and S cone spectral sensitivities in place of the spectral sensitivity P above to determine the responses l, m and s, respectively. One consequence of the three-dimensionality of trichromatic visual response is that lights that differ physically may produce the same response. Such different but indistinguishable lights are known as metamers, and exist in consequence of the fact that a three-dimensional representation of lights cannot possibly represent accurately variation in the high-dimensional space of lights.

A light's chromaticity describes its chromatic properties independently of its intensive properties. For instance, if X, Y and Z are the responses to a light determined using the CIE $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ functions, then the light's chromaticity (x, y) and its luminance L are given by the following formulae:

$$x = X/(X + Y + Z) \tag{4}$$

$$y = Y/(X + Y + Z) \tag{5}$$

$$L = Y \tag{6}$$

Perceived hue and saturation vary in the two-dimensional color space described by chromaticity; equiluminant lights vary in chromaticity alone.

COLOR OPPONENCY

The Young–Helmholtz theory that the responses of the L, M and S cone systems are tied directly

to the perception of red, green and blue, respectively, is of historical interest alone. Subsequent work has substantiated Hering's idea that color appearance is due to the activity of opponent mechanisms that compare photoreceptoral system responses. Color-opponent processing commences in the retina, as is evident in the responses of bipolar cells and ganglion cells, and continues in the lateral geniculate nucleus (LGN) and in visual cortex. The mismatch between psychologically determined color-opponent functions and physiological opponent-mechanism sensitivities is of current research interest. Higher-order color-opponent mechanisms, although implicated in the results of chromatic detection experiments and organized in a manner consistent with physiological studies of visual cortex, have yet to be linked to appearance.

Standard Mechanisms of Color Appearance

Color-opponent theory is based on the analysis of hue. Under simple viewing conditions, one cannot perceive a light that appears both reddish and greenish, or one that appears both bluish and yellowish. This suggests that the responses of two opponent mechanisms underlie color appearance. The first is a red–green mechanism, the response of which can generate red or green, but not both simultaneously; the second is a similarly organized blue–yellow mechanism. A second critical observation concerns the existence of lights with unique hues: unique red and unique green appear neither yellowish nor bluish, while unique yellow and unique blue appear neither reddish nor greenish. All other hues appear to combine two of the primary hues. For instance, orange appears both reddish and yellowish. One can integrate the observation of unique hues into the opponent model as follows: first, unique red and unique green are seen when the red–green mechanism responds red or green, respectively, and when the response of the blue–yellow mechanism is zero (signals neither blue nor yellow); and second, unique blue and unique yellow are seen when the blue–yellow mechanism responds blue or yellow, respectively, and when the response of the red–green mechanism is zero. Graded responses by the mechanisms can account for perceived color saturation (chroma). For instance, a highly saturated red will be seen if the red–green mechanism produces a strong red response; less-saturated reds correspond to weaker responses.

Hurvich and Jameson used a hue-cancellation technique to derive spectral sensitivities for the two opponent mechanisms (Hurvich and Jameson, 1957). For instance, the relative amounts of red in monochromatic lights of equal energy can be measured by adding a green light of fixed wavelength and varying its intensity to produce a unique yellow light, which appears neither reddish nor greenish. The intensity of the green light needed to cancel the redness found in lights of long and short wavelength provides a measure of the red half of the red–green mechanism's spectral sensitivity function. Similar procedures provide the green half of the red–green mechanism's sensitivity, and the blue and yellow halves of the blue–yellow mechanism's sensitivity (Figure 2a).

Opponent-mechanism sensitivities can be described in terms of cone responses by combining cone spectral sensitivity functions to provide the best fit to the opponent functions. Such fits suggest, first, that the responses of L and S cones are opposed to those of the M cones by the red–green mechanism, and second, the responses of S cones are opposed to those of L and M cones by the blue–yellow mechanism. The contribution by L and S cones to red causes not only lights at long wavelengths to appear red but also lights at very short wavelengths, a feature required to account for the redness in the violet lights at short wavelengths. The L and M cones combine in the blue–yellow mechanism to produce a yellow response that is opposed to the S cone's blue.

The opposition of cone responses results in color-opponent sensitivities that take on both positive and negative values. For instance, if the L and S cones are taken to excite and the M cones to inhibit the red–green mechanism, then the 'red' portion of the red–green spectral sensitivity will take on positive values while the 'green' portion will take on negative values. A positive response by such a mechanism would cause red to be perceived, and a negative response green.

A common emendation to the standard model, consistent with physiological evidence, uses rectification to produce separate red, green, yellow and blue mechanisms (as well as black and white) (Figure 2b). Nonlinear combination of cone inputs by such mechanisms can account for the spectral nonlinearity in blue–yellow processing; stated simply, too much of the spectrum appears blue and too little yellow to be consistent with linear combination of cone responses.

Chromatic Sensitivity

Chromatic sensitivity depends on adaptation by the visual system to viewing conditions. Absolute

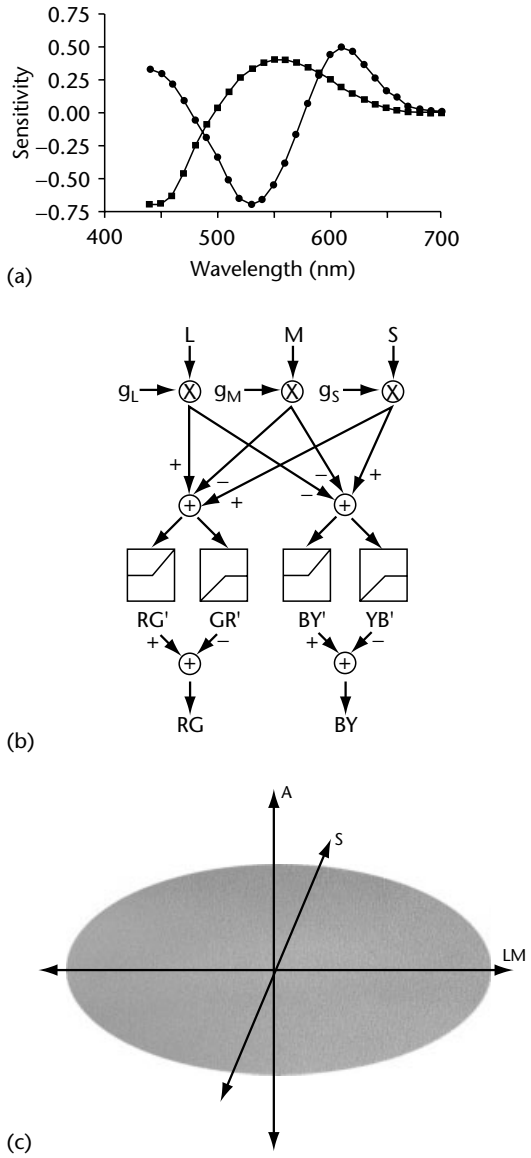


Figure 2. [*Figure is also reproduced in color section.*] Color-opponent transformation of photoreceptoral signals. (a) Hurvich and Jameson hue-cancellation functions. The red–green function (●) codes redness through positive values and greenness through negative values, while the yellow–blue function (■) codes yellowness through positive values and blueness through negative values. (b) Wiring diagram for color-opponent mechanisms. Signals from L, M and S cone photoreceptors with multiplicative gain controls are combined by half-wave-rectified opponent mechanisms to produce mechanisms that signal red (RG′), green (GR′), blue (BY′) and yellow (YB′) which are then combined to form RG and BY opponent channels. (c) Equiluminant colors in a plane defined by LM and S axes. Modulation among lights along LM and S axes changes only L and M cone signals and only S cone signals, respectively. These axes correspond to the peak chromatic sensitivities of the retinogeniculate pathways.

sensitivity to changes in light level about some reference light declines as the intensity of the reference light increases. This change occurs, in part, through the action of multiplicative gain controls within individual cone photoreceptors, although such adaptation of cones is significantly less than that predicted by psychophysical measurements. Chromatic adaptation to a steady background light also occurs at color-opponent sites in the visual pathways and is thought to involve both multiplicative and additive components.

The sensitivity of color-opponent mechanisms to modulations about a steady background light is high. The red–green mechanism supports the detection of red signals with L cone contrasts as small as 0.1%, a sensitivity more than five times greater than that found with stimuli of varying intensity (e.g. black–white). Equiluminant stimuli (Figure 2c) are generally used to probe the sensitivity of chromatic processing, because chromatic mechanisms are more likely to detect equiluminant stimuli than are mechanisms sensitive specifically to variation in light intensity. Equiluminant stimuli provide poor inputs to visual subsystems that handle motion, stereopsis, Vernier acuity and the interpretation of shadows. This is thought to be due, in part, to the strong coloration possible with very small contrast signals.

Chromatic sensitivity depends also on the spatiotemporal pattern of light modulation about a steady background light. For instance, prolonged viewing of a red–green temporal modulation reduces red–green sensitivity but leaves blue–yellow sensitivity relatively unscathed, and vice versa. Such long-term habituation to chromatic modulation is complemented by the activity of rapidly acting gain controls which operate on the spatial pattern of color-opponent contrast signals. Both habituation and color contrast gain control are thought to have a cortical locus.

Higher-order Mechanisms

Two spectral classes of color-opponent neuron are found in macaque retina and LGN. The first opposes L and M cone inputs linearly. Such a neuron does not have the sensitivity of the standard red–green opponent mechanism, which requires S cone input to produce red at short wavelengths. The second opposes S cone inputs to those of L and M cones. This class of neuron is largely insensitive to lights that correspond to unique red, so bearing one of the characteristics of the standard blue–yellow mechanism, but typically

combines cone responses linearly in a steady state of adaptation.

The mismatch between retinogeniculate processing and the standard color-opponent model takes a different form in areas V1, V2 and V3 of the visual cortex. Color-sensitive neurons in cortex have spectral sensitivities which are scattered more uniformly in the color plane (Lennie *et al*., 1990). Many such neurons are most sensitive to hues lying along axes in the color plane that lie between the LM and S axes that characterize retinogeniculate processing (Figure 2). Psychophysical work with habituation, visual search and noise-masking paradigms suggests that these higher-order mechanisms operate in everyday visual detection tasks. For example, observers can deploy a violet-sensitive mechanism when looking for a violet signal rather than looking for simultaneous red and blue signals. How these mechanisms contribute to color appearance, if at all, is an open question.

COLOR CONSTANCY

The utility of color in object recognition depends on the stability of an object's color appearance under change in its viewing conditions. An important aspect of this is the stability of surface color under change in the spectral properties of illumination. Observers can discount illumination spectral change nearly completely under appropriate circumstances (Kraft and Brainard, 1999). Physics-based analyses have identified conditions under which this is possible, and have explored a variety of cues that a trichromatic visual system can use to find descriptors of an object's surface reflectance function from reflected lights (Maloney, 1999). Transparency provides a further avenue into the study of surface color appearance; observers assign colors to surfaces behind a filter in a manner consistent with filter-specific changes in color-opponent signals (D'Zmura *et al*., 2000). The parallel representation of filter and surface chromatic properties in perceived transparency suggests that low-level chromatic signals are shunted into parallel, layered representations of scene color.

Change in Illumination Spectral Properties

A white piece of paper reflects very different lights towards the eye when viewed outdoors under direct sunlight or under a yellow tungsten lightbulb (Figure 3a). Perfect color constancy entails identical colors for the white paper under the two viewing conditions. This is possible if the visual system represents directly the chromatic properties of the paper's surface. These are specified by a reflectance function $R(\lambda)$, which describes the fraction of light reflected by a surface as a function of wavelength (Figure 3b).

Principal components analysis of collections of surface reflectance functions shows that much of their variation (greater than 99% in sets such as the Munsell chips) can be captured by just three components; to a good approximation, surface reflectance functions lie in a three-dimensional space (Figure 3d). A particular surface reflectance is identified by three descriptors which tell how much each principal component contributes to the reflectance. A trichromatic visual system can thus represent surface reflectance properties directly through a linear transformation of cone responses, similar to that by the standard black–white and color-opponent mechanisms, which produces surface reflectance descriptors.

Nevertheless, reflected light confounds illumination and surface spectral properties. The reflected light $E(\lambda)$ is the product of the illumination spectral power distribution $D(\lambda)$ and the surface reflectance function $R(\lambda)$. The linear transformation of cone responses which provides surface reflectance descriptors must change in response to illumination change if color constancy is to be possible.

Principal components analysis of collections of daylight spectral power distributions (SPDs) shows that almost all of their variation can be captured by just three components (Figure 3c). For daylight, then, knowledge of an illuminant's chromaticity lets one reconstruct its SPD, up to a single scale factor.

Potential cues to the chromatic properties of scene illumination are numerous. If one assumes that the brightest surface in a scene is white, with a spectrally flat reflectance function, then the photoreceptoral responses to the white surface correspond to those of the illuminant itself. This knowledge can be used to adjust the linear transformation of photoreceptor responses to provide color-constant reflectance descriptors. This idea works more generally for reference surfaces with known reflectance properties. The gray-world assumption is similar; one assumes that the space-averaged surface reflectance in a scene corresponds to a spectrally neutral gray, so that illuminant chromatic properties can be inferred from the space-averaged photoreceptor responses. Specularity provides a further such cue; the lights from two or more surfaces with highlights provide sufficient information to infer illuminant chromatic

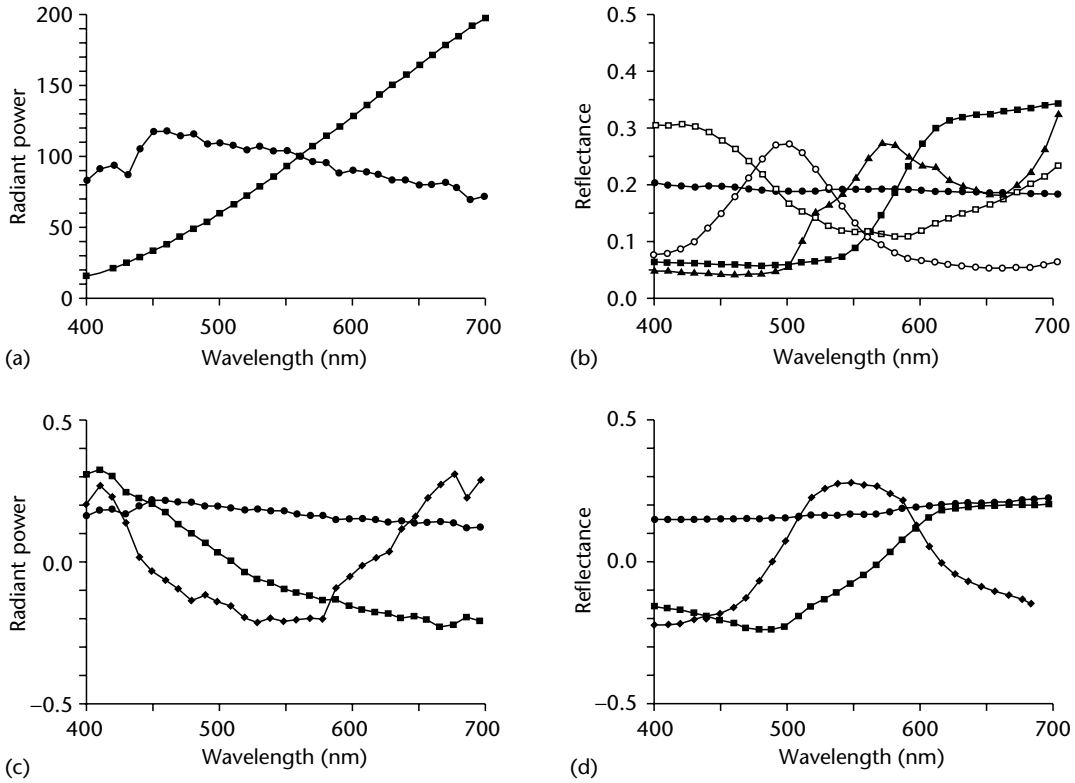


Figure 3. Finite-dimensional linear models for illuminant spectral power distributions (SPDs) and surface reflectance functions. (a) The SPDs of CIE illuminant D65 (●) and illuminant A (■), which match those of average daylight and a yellow tungsten light bulb, respectively. (b) Reflectance functions of five Munsell chips: N 5 (●), 10R 4/6 (■), Y 5/6 (▲), BG 4/6 (○) and 10PB 4/6 (□). (c) First three functions in the CIE daylight illumination model (Gram–Schmidt orthogonalized). The function $D_1(\lambda)$ (●) captures variation in the mean level of illumination; the function $D_2(\lambda)$ (■) captures 'blue–yellow' variation like that found between blue sky and the yellow solar disk, while the function $D_3(\lambda)$ (♦) captures minor 'red–green' variation. Daylight with SPD $D(\lambda)$ can be approximated accurately as a linear combination of these three functions: $D(\lambda) \approx d_1\ D_1(\lambda) + d_2\ D_2(\lambda) + d_3\ D_3(\lambda)$. (d) First three functions in the Cohen principal components analysis of Munsell chip reflectance functions (Gram–Schmidt orthogonalized). The function $R_1(\lambda)$ (●) captures variation in the mean level of reflectance; the function $R_2(\lambda)$ (■) captures 'red–green' variation, while the function $R_3(\lambda)$ (♦) captures 'blue–yellow' variation. A reflectance R can be approximated by a linear combination of these three functions: $R(\lambda) \approx r_1\ R_1(\lambda) + r_2\ R_2(\lambda) + r_3\ R_3(\lambda)$. The coefficients $\{r_1, r_2, r_3\}$ describe the reflectance function and do not vary with illumination, so that a visual system able to recover these descriptors can exhibit color constancy.

properties. Of these three cues, only the gray-world assumption is thought to have perceptual relevance; eye movements let mechanisms with sufficiently long integration times adjust their sensitivity to space-averaged inputs. The scaling of photoreceptoral responses by their average inputs, a form of von Kries adaptation, is considered an important influence on perceived surface color. Yet numerous examples show that von Kries adaptation is insufficient to account for surface color appearance.

The gray-world assumption is a statistical one that may be generalized to assumptions concerning the prior distributions of possible surface reflectances and illuminant SPDs in visual scenes. Bayesian algorithms for color constancy take as data the photoreceptoral responses from a set of surfaces and determine the illuminant and surface chromatic properties most likely to have given rise to the data.

Models of photoreceptor response that depend linearly on both surface reflectance and illumination spectral properties are bilinear models. Analysis of such bilinear models shows that viewing several surfaces under an unknown daylight illuminant does not provide enough information to find three reflectance descriptors per surface. Yet if the same surfaces are seen under first one illuminant and then another, so that two views of the surfaces are provided, sufficient information exists to determine both surface and illuminant chromatic properties. A key assumption in the use of multiple

views is that the visual system must know the correspondence between surfaces viewed under the first and second illuminants.

Transparency

Surface correspondence is readily determined in viewing situations involving transparent filters. Surfaces that lie along the edge of a filter are viewed both directly and through the filter. At the intersections of surface and filter edges are X junctions, where color changes are cues to surface and filter chromatic properties. Psychophysical studies show that a filter of uniform color properties is perceived best when surface chromatic changes are characterized by a single shift (translation) in color space and/or a single change in contrast. The shift in surface responses corresponds to filter color, while the change in contrast corresponds to filter cloudiness, which ranges from clear to opaque.

The perceived separation of a spatiochromatic pattern into surface and filter layers is an example of scission, the layered representation of the visual field. Standard color theory allows for a trichromatic representation of scene spectral properties at every point, and this is seemingly insufficient to account for our ability to perceive two (or more) colors simultaneously at the same point. One possibility is that mechanisms involved in scene segmentation feed back on standard representations to shunt trichromatic signals into multiple, concurrent color representations.

Color Coding

While color provides information about an object's material properties, it also serves as a flexible visual code well suited to investigations in visual attention, category learning, verbal production (e.g. the Stroop effect) and other cognitive tasks. Performance based on color coding depends on memory and language, which depend, in turn, on categorization. All major modern languages have equivalents for the achromatic color names white, gray and black; the primary color names red, yellow, green and blue; and the secondary color names pink, brown, orange and purple. These names correspond to volumes in three-dimensional color space that bound the chromatic stimuli that elicit the names. Stimuli in the center of such a volume typify the color category better than those close to a boundary. Studies of color memory suggest that remembered colors often tend to shift towards their prototypes, revealing perhaps the action of verbal encoding on color memory. Color coding has many important roles in society through semantic binding (e.g. the red, orange and green of traffic lights), perceived emotional content, and through esthetic considerations expressed in art and design.

CONCLUSION

The psychophysical study of color relies on careful distinctions among physical light stimuli, chromatic properties related to photoreceptoral and opponent mechanism responses, and perceived color. Relating perceived color to physical light stimuli is difficult, not only because of the intermediate chromatic mechanisms, but also because of philosophical issues in understanding color and comparing color among individuals. These difficulties notwithstanding, considerable progress has been made in characterizing chromatic processing and in identifying sources of individual differences in chromatic sensitivity.

References

D'Zmura M, Rinner O and Gegenfurter K (2000) The colors seen behind transparent filters. *Perception* **29**: 911–926.

Hurvich LM and Jameson D (1957) An opponent-process theory of color vision. *Psychological Review* **64**: 384–404.

Kraft JM and Brainard DH (1999) Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences USA* **96**: 307–312.

Lennie P, Krauskopf J and Sclar G (1990) Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology (London)* **357**: 649–669.

Maloney LT (1999) Physics-based approaches to modeling surface color perception. In: Gegenfurtner KR and Sharpe LT (eds) *Color Vision: From Genes To Perception*, pp. 387–416. New York: Cambridge University Press.

Sharpe LT, Stockman A, Jaegle H and Nathans J (1999) Opsin genes, cone photopigments, color vision, and color blindness. In: Gegenfurtner KR and Sharpe LT (eds) *Color Vision: From Genes To Perception*, pp. 3–51. New York: Cambridge University Press.

Smith VC and Pokorny J (1975) Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision Research* **15**: 161–171.

Stockman A and Sharpe LT (2000) Spectral sensitivities of the middle- and long-wavelength sensitive cones derived from measurements in observers of known genotype. *Vision Research* **40**: 1711–1737.

Further Reading

Backhaus WGK, Kliegl R and Werner JS (eds) (1998) *Color Vision: Perspectives from Different Disciplines*. New York: Walter de Gruyter.

Byrne A and Hilbert DR (eds) (1997) *Readings on Color*, vol. 1. *The Philosophy of Color*. Cambridge, MA: MIT Press.

Byrne A and Hilbert DR (eds) (1997) *Readings on Color*, vol. 2. *The Science of Color*. Cambridge, MA: MIT Press.

Gegenfurtner KR and Sharpe LT (1999) *Color Vision: From Genes to Perception*. New York: Cambridge University Press.

Kaiser PK and Boynton RM (1996) *Human Color Vision*, 2nd edn. Washington, DC: Optical Society of America.

Katz D (1935) *The World of Colour*, translated by RB MacLeod. London: Kegan Paul.

Nassau K (1983) *The Physics and Chemistry of Color: The Fifteen Causes of Color*. New York: John Wiley.

Stiles WS (1978) *Mechanisms of Colour Vision*. London: Academic Press.

Wandell BA (1995) *Foundations of Vision*. Sunderland, MA: Sinauer.

Wyszecki G and Stiles WS (1982) *Color Science. Concepts and Methods, Quantitative Data and Formulae*. New York: John Wiley.

Comparative Psychology

Introductory article

Edward A Wasserman, University of Iowa, Iowa City, Iowa, USA

CONTENTS

Comparative psychology explores the behaviors of different species of animals with a special interest in any similarities and differences that may reveal the evolutionary origins of those behaviors.

INTRODUCTION

Are animals intelligent? How can we learn about animal intelligence? Do different species differ in intelligence? How can we measure species differences in intelligence? Does animal intelligence resemble human intelligence? What would it mean if animals were indeed intelligent and if their intelligence approached, or even eclipsed, our own? (*See* **Intelligence**; **Animal Cognition**; **Animal Learning**)

Most people, young and old, have asked these intriguing questions. Answering them is the business of comparative psychology, a field that explores similarities and differences in the behavior of human and nonhuman animals. Comparative psychologists are interested in a wide range of behaviors: from mating to migrating, from feeding to fighting, from sleeping to scratching. These scientists have been especially interested in intelligent behaviors: those actions whose acquisition advances an animal's chances of surviving and reproducing in an environment that is fraught with danger and uncertainty.

Historical Foundations

The comparative psychology of intelligence is only a century or so old if we date the origin of the field as an exact experimental science with the 1898 publication of Edward L. Thorndike's pioneering monograph, *Animal Intelligence: An Experimental Study of the Associative Processes in Animals*.

But, weren't people interested in and didn't they know about animal intelligence for much more than the past century? Certainly. Ancient texts and drawings suggest that humans have trained animals for work and amusement for centuries; doing so meant that humans understood and exploited the modifiability of animal behavior. Nevertheless, a practical understanding of animal intelligence is decidedly different from a scientific understanding. Formulating precise laws of learning and developing effective technologies of teaching are crucial to determining the nature of intelligence and to exploring its generality throughout the animal kingdom. These critical additions to our knowledge of animal intelligence came only at the beginning of the twentieth century with the work of Thorndike and other trailblazers in the new science of comparative psychology. (*See* **Learning, Psychology of**)

Even before this modern experimental era, famous scholars had deemed animal intelligence to be central to comprehending the nature and origin of humankind. Two key points in the history of human thought place the study of animal behavior and learning at the very center of philosophical and scientific inquiry: (1) Descartes' distinction between humans and brutes, and (2) Darwin's hypothesis of mental continuity between human beings and nonhuman animals.

The seventeenth-century French philosopher René Descartes believed that human beings were profoundly different from brutes. Animals were mere machines; they had intricate bodily systems that controlled their physiology and behavior, but they lacked what humans alone possessed – a rational soul. The rational soul was divinely created, it was not made of matter, nor did it reside in the human body. (*See* **Descartes, René**)

The operation of the rational soul had two unique behavioral consequences: (1) it allowed us to communicate our private thoughts and feelings to other human beings, and (2) it permitted us to suitably tailor our behaviors to a vast variety of complex and ever-changing environmental situations. Descartes

believed that animals had no thoughts to communicate; they were thus forced to respond as their sensory and motor systems demanded, without the involvement of intelligence.

Against this backdrop of Cartesian thinking, the nineteenth-century English biologist Charles Darwin proposed that the nature and descent of human beings was not a matter for theology or philosophy, but biology. Scores of naturalistic and anecdotal observations convinced Darwin that humans and animals were not fundamentally different from one another, nor did they have different origins; all beings were the products of organic evolution. In contrast to Descartes, Darwin viewed both communication and intelligence from a natural scientific perspective; primitive or even highly advanced forms of each of these abilities were to be found throughout the animal kingdom, thus disclosing what Darwin called 'mental continuity' between human and nonhuman animals. (*See* **Evolutionary Psychology: Theoretical Foundations**)

Darwin's bold evolutionary ideas made the study of animal behavior crucial to understanding human behavior: if humans arose from lower forms of animals, then the study of animal intelligence is essential to elucidating the biological precursors of the human mind.

Thorndike's Contributions

When he began his research, Thorndike accepted Darwin's evolutionary perspective, but he was skeptical of Darwin's evidence of animal intelligence. Indeed, Thorndike devised his famous 'puzzle box' method in order to provide an objective and experimental antidote to the subjective and anecdotal accounts of animal intelligence that were in vogue during the final decades of the nineteenth century, and that Darwin used to make his case for mental continuity between human and nonhuman animals. These anecdotes were tall tales of animal genius that were spun by pet owners, zookeepers, and amateur naturalists. Most of these astounding and amusing anecdotes proved to be of dubious accuracy and reliability; but, they did provoke great interest and debate in popular and scientific circles.

In his innovative research, Thorndike placed animals into small boxes from which they could escape and receive food by solving a simple behavioral puzzle. For instance, a hungry cat might have to pull a loop of string; doing so opened a door through which the feline could exit the box and nibble a tasty titbit of fish. By measuring the time that it took the cat to claw the string after each placement into the box, Thorndike found that this time progressively fell with successive trials.

Thorndike had discovered a basic law of learning – the law of reinforcement. When a response is followed by a reward, an organism is more likely to make that response again in that situation. Further study by Thorndike and later comparative psychologists showed that this law of reinforcement is not limited to cats, to pulling loops of string, or to fish snacks. Pigeons more quickly peck a button when grains of seed ensue. Rats more rapidly rotate a wheel when draughts of water follow. And, human infants kick with added alacrity when a motorized mobile turns afterwards. Such learned behaviors need not continue indefinitely; if the reward is revoked following the response, then the behavior returns to its initial level – the law of extinction.

By 1911, Thorndike's own research and that conducted by several other investigators led him to conclude that most vertebrate animals learn in the same general way: stimulus–response associations are automatically strengthened by reward and weakened by extinction. (*See* **Reinforcement Learning: A Biological Perspective**)

Learning by consequence was positively Darwinian. Natural selection leads to the retention of fit organisms and to the elimination of unfit ones; the laws of reinforcement and extinction lead, respectively, to the retention of effective behaviors and to the elimination of ineffective ones.

Most relevant to the evolution of intelligence, Thorndike believed that species differences in learning are matters of degree, not kind: stimulus–response associations may increase in number, may be formed more quickly, may last longer, and may become more complex. Growth in the number, speed of formation, permanence, and complexity of associations reaches its high point in human beings. Thorndike also suggested that a parallel exists between individual development (ontogeny) and species evolution (phylogeny): the development of the infant's intelligence to the adult condition may be viewed as progressing from animal to human competence.

COMPARATIVE PSYCHOLOGY OF LEARNING

Thorndike's groundbreaking research and theorizing set the stage for the work of succeeding generations of comparative psychologists.

Critical to this later work was the crafting of advanced methods which could sensitively and reliably measure animals' learning of new behaviors. Most famous among these new methods was I. P. Pavlov's conditioned reflex procedure and B. F. Skinner's 'Skinner Box' procedure, a refinement of Thorndike's puzzle box procedure. These new methods have been creatively adapted to the unique behavioral repertoires of different species in the quest to uncover similarities and differences in their intelligence. (*See* **Pavlov, Ivan Petrovich**; **Skinner, Burrhus Frederic**; **Conditioning**)

Quantitative Differences in Learning

Despite the large amount of experimentation into quantitative differences among animal species that has been conducted since Thorndike's investigations, most reviewers have concluded that this line of research has not been productive. Quite simply, animal behavior is affected by so many other factors – differences in sensory and motor capabilities, variations in motivation and reward, differences in daily activity levels – that valid quantitative cross-species comparisons of intelligence have proven to be virtually impossible to document. (*See* **Motivation**)

Consider an illustrative example: is a cat smarter than a dog? To begin, we have to select some common behavioral technique that can properly measure intelligence in these two very different animals. But, which technique? The senses of cats and dogs are obviously not equally keen. Cats have excellent audition (hearing); dogs have excellent olfaction (sense of smell). By requiring the animals to use one sense or the other in the experimental task, we could stack the deck in favor of one species over the other. The two species' motor abilities also differ, with cats being outstanding jumpers and dogs being outstanding runners. Different response requirements could favor one species over the other. What about the reward? How much of which kinds of foods will equate the value of the rewards for the two species? What will be the effect of different levels of food deprivation in the two species? And, at what time of day should the animals be trained? Should we train in the day for the diurnal dog or in the night for the nocturnal cat?

These and many additional considerations make it abundantly clear that, however interesting and well-intentioned the original question may have been, it is practically impossible to say whether the cat or the dog is the more intelligent species of animal.

Does it therefore follow that there are no differences in intelligence among different species to be documented? Not necessarily. We may simply have to adopt a different experimental tactic in order to discover them.

Qualitative Differences in Learning

The comparative psychologist M. E. Bitterman has suggested that we investigate qualitative differences in learning among different species of animals. Bitterman's proposed plan of investigation has been to compare species according to their orderly responses to changes in the conditions of reinforcement.

For example, some species of animals may come to learn discrimination reversals faster than they learned the original problem, whereas others may not show such improvement. To illustrate: the choice of a white circle over a black one might initially lead to reward. Then, the discrimination is reversed: the choice of the black circle over the white one now leads to reward. Later, the discrimination is reversed over and over again from problem to problem. In fact, some species of animals improve in the speed of reversal learning, whereas others do not. Such a dramatic difference suggests that qualitatively different learning processes may be producing these effects.

As another example, some species may exclusively select the better response alternative ('maximizing'), whereas other species may match their choices to the reward probabilities ('matching'). So, if responses to one button produce food with a probability of 0.75, whereas responses to a second button do so with a probability of 0.25, then the receipt of reward will be maximized by choosing the first button 100 per cent of the time ($100(0.75) + 0(0.25) = 75.00$), but not by choosing the first button 75 per cent of the time and the second button 25 per cent of the time ($75(0.75) + 25(0.25) = 62.50$). These two patterns of choice responding may represent qualitatively different decision strategies: a potentially important species difference in learning and behavior. (*See* **Choice Selection**)

COMPARATIVE PSYCHOLOGY OF MEMORY

Beyond the comparative study of learning in animals, other advanced intellectual processes have attracted experimental attention, among them memory. Memory is obviously necessary for

learning to occur. If an organism in Thorndike's puzzle box were to forget what response it last performed in that situation, then there would be no way for food to strengthen that particular stimulus–response bond. (*See* **Memory**)

Given how central memory is to the acquisition of intelligent action, it should not be surprising that comparative psychologists have long been interested in the experimental investigation of memory. Furthermore, given Darwin's hypothesis of mental continuity and Thorndike's suggestion that species differences in learning are primarily quantitative and not qualitative in character, it should not be surprising that researchers have endeavored to determine if some species can remember prior acts and events for different lengths of time.

Delayed Response

The pioneer in the study of animal memory was W. S. Hunter, who in 1913 devised the delayed response paradigm. In this paradigm, an animal might see one of three potential sites baited with food or otherwise marked by a brief stimulus such as a light. Then, after a delay of a few or several seconds, the animal would be required to select the earlier baited or marked site. If the animal succeeded in choosing the correct site and in receiving the food that was to be found there, then this result might constitute evidence for memory of the baited location. The longer the delay over which the animal is able to respond correctly, the better is its memory.

But, a rival interpretation presents itself. Successful performance in the delayed response task might not be the result of some enduring cognitive or neural process; it might instead be due to the animal merely maintaining its bodily orientation to the baited site during the delay period. Postural mediation was an especially obvious possibility for one of Hunter's contingent of different experimental subjects – the dog, an animal that is well known for its pointing behavior.

Hunter and other memory researchers cleverly strove to eliminate this postural possibility by rotating the animal on a turntable or by removing the animal from the apparatus during the delay period. These measures did sometimes succeed in disrupting the animal's delayed discrimination performance, as they should if pointing were all that there was to accurate performance after a delay. But, these measures did not always disrupt the animal's performance. Clearly, then, there are at least some bona fide cases of memory in the delayed response paradigm that cannot be due to mere postural mediation.

Delayed Matching-to-Sample

More recent research has exploited an alternative testing method for measuring animal memory – delayed matching-to-sample. This method is not plagued by the problem of positional mediation; also, delayed matching-to-sample is much more versatile in its application to a wide variety of issues in animal memory and intelligence than is the delayed response paradigm. Further contributing to the popularity of delayed matching-to-sample has been its extremely successful application to the behavior of the pigeon, which – because of its long life, excellent vision, and ready adaptation to laboratory captivity – has become a favorite of experimental psychologists interested in animal cognition.

In this procedure for the pigeon, two simultaneously presented testing stimuli (for example, red and green lights) follow the presentation of the sample stimulus (randomly, a red or a green light on alternate trials). Only one testing stimulus (whose color matches the sample stimulus) is correct and leads to food if it is chosen; the other testing stimulus (whose color does not match the sample stimulus) is incorrect and does not lead to food if it is chosen. With a short delay between the sample and the testing stimuli, pigeons show a strong tendency to choose the correct (matching) testing stimulus. However, as the delay is lengthened, choice accuracy declines towards the indiscriminate selection of the correct (matching) and the incorrect (nonmatching) testing stimuli, thereby documenting the forgetting of the sample stimulus. Not only is there a decline in memory as the sample–test interval is lengthened; choice accuracy is also positively related to the duration of the sample stimulus and negatively related to the time between trials.

Because these simple temporal parameters of delayed matching-to-sample exert such a strong influence on memory performance within a species, any claims about memory differences between species are extremely difficult to prove. The parallel to the case of quantitative species differences in learning is obvious.

An important variant of the delayed matching-to-sample procedure involves sample and testing stimuli that are drawn from different pools of stimuli. For instance, pigeons might be shown different colors as sample stimuli and different forms as testing stimuli. No true matches are possible; only arbitrary or symbolic matches can hold. So, the selection of circle after red or the selection of square after green might be the correct responses,

whereas the selection of square after red or the selection of circle after green might be the incorrect responses. This so-called 'delayed symbolic matching-to-sample' procedure has afforded researchers special opportunities to expand the investigation of animal memory.

With true matching-to-sample procedures, it has been shown that pigeons can remember the color of a sample stimulus, its shape, its orientation, and its spatial location. With the symbolic matching-to-sample procedure, it has also been possible to show that the duration of a stimulus can be remembered. Thus, pigeons remember different durations of a red sample stimulus and report that memory during testing stimuli of differing line orientations.

Beyond the attributes of single sample stimuli, pigeons that have been given embellished versions of the symbolic matching-to-sample paradigm have also been shown to remember the temporal order (e.g. red–green) of two differently colored sample stimuli, the spatial order (e.g. left–right) of two identically colored sample stimuli, and the relative duration (e.g. short–long) of two differently colored sample stimuli. As well, pigeons and rats have successfully been trained to make one of two different responses depending on the number of prior visual or auditory stimuli (for example, two versus four).

Another way in which the memory of complex information has been studied has involved sample stimuli that comprise two or more elements. Pigeons were thus shown two-element sample stimuli that were composed of color (red or green) and line orientation (horizontal or vertical) elements. Tests with just color comparisons or just line comparisons each yielded highly accurate testing performance, indicating that the pigeons discriminated and remembered both the color and the line orientation of the compound sample stimuli. Significantly, however, accuracy on these compound sample (color and line) trials was lower than on other trials involving only single-element samples (color or line); this result suggests that the two sample elements on compound sample trials competed with one another for what in humans is commonly called *attention*. (*See* **Attention**; **Selective Attention**)

All of this evidence shows that animals are amazingly sensitive to the richness and subtlety of their environment. That sensitivity is also enduring, as witnessed by the fact that stimuli can be retained for substantial periods of time, sometimes for durations as long as half a minute.

Rehearsal

Many theorists of human memory have proposed the operation of control processes: means by which memories are changed in accord with the needs of the individual or the demands of the task. One key control process is *rehearsal*: a covert activity that helps to sustain the memory of a prior event. Engaging in rehearsal should aid in retaining earlier information, whereas terminating rehearsal should impair retention.

In order to investigate the role of rehearsal in human memory, researchers have devised the directed-forgetting paradigm. In one version of the directed-forgetting paradigm, people are given one of two cues (e.g. red or green colors) either to remember or to forget a previously presented stimulus. When the remember cue is given the memory test is presented afterwards; when the forget cue is given the memory test is not presented afterwards.

Finally, a trick is played on the individual; a forget cue is given and is followed by a retention test. When these unexpected retention tests are given, researchers generally find that memory performance is worse on forget-cued trials than on remember-cued trials; this result implies that the post-stimulus cues were affecting the rehearsal process and thereby modulating memory.

Researchers have further found that postponing the forget cue in a delay interval of fixed duration leads to a loss in its effectiveness; memory performance for earlier information improves the later into the delay interval the forget cue is given. This result suggests that the spontaneous or uncued rehearsal that occurs prior to the forget cue effectively protects memory from the decremental effect of the forget cue.

Several workers in the area of animal memory have attempted to see whether directed forgetting is uniquely human. Research with both pigeons and monkeys has adapted the delayed matching-to-sample paradigm to this objective by adding brief post-sample cues during the delay interval in order to signal that a test for sample memory either would or would not be given. As in the case of human memory, animal memory proved to be much lower on forget-cued trials than on remember-cued trials. In addition, memory was more markedly reduced if the post-sample forget cue was presented early than if it was presented late in the delay interval.

These results suggest that animals may indeed have active control over memory processing.

Rehearsal may not be uniquely human nor necessarily verbal.

COMPARATIVE PSYCHOLOGY OF CONCEPTUALIZATION

Human beings and other animals are incessantly bombarded by an extraordinarily complex array of external stimuli; yet, they somehow make sense of these varied and varying stimuli. One way to reduce the demands on an organism's sensory and information-processing systems is for it to treat similar stimuli as members of a single class; by doing so, considerable behavioral economy can be achieved, thus freeing its adaptive machinery to deal with other competing demands of survival. In addition, categorical processing permits an organism to identify novel stimuli as members of a particular class and to generalize knowledge about that category to these new members. So, an organism need not be bound to respond only to those stimuli with which it has had prior experience, further enhancing its ability to cope with a continually changing world. (*See* **Generalization**)

Theorists often trumpet these adaptive virtues of categorization and conceptualization; yet, we remain far from understanding precisely how organisms partition the world into classes of related objects and events. Indeed, theorists have historically doubted whether nonhuman animals are even capable of conceptual behavior. More than a century ago, the famous English comparative psychologist C. Lloyd Morgan denied animals the ability to behave conceptually. Morgan believed that only adult humans, and not even children, are capable of conceptualization. Recent research is changing that opinion. (*See* **Categorical Perception**)

Object Concepts

One familiar case of conceptual behavior involves the kinds of open-ended categorization responses that we make when we label different natural (e.g. cat) and artificial (e.g. car) objects with different nouns. Such verbal behaviors are occasioned by specific instances of wide variability and individuality. Indeed, accurate classification even extends to categorical exemplars that we have never seen before. Is it at all possible for nonhuman animals lacking language to engage in this form of conceptual or classificatory behavior?

In order to answer this question, a new technique was devised to train pigeons concurrently to discriminate stimuli from several human language categories. The specific method parallels the technique that parents often use in order to teach their children to label objects in a picture book. When the page is turned, the child is first asked to look at the object and then she is requested to name it. If she is correct, then she is praised. If she is incorrect, then she is told 'no' and is encouraged to try again. If self-correction fails, then she is provided with the correct name. (*See* **Concept Learning**)

In order to implement this method with pigeons, a color snapshot was displayed on a small screen and the pigeon was required to peck a clear plastic key covering the screen in order to guarantee that it was looking at the snapshot. Then, four differently colored keys were illuminated just beyond the corners of the screen. A single choice response was permitted. If the response was to the correct key for reporting the stimulus on the viewing screen, then the pigeon was fed grain; if the response was to any of the three incorrect report keys, then no grain was given and the pigeon had to repeat that trial until it made the correct response. The slides that were shown in each daily session depicted several different examples of cats, flowers, cars, and chairs. The pictures contained one or more instances of the critical stimulus object; the objects were indoors or outdoors, near or far away, centered or off-center, and in different colors, orientations, and backgrounds.

In a representative experiment, pigeons attained a level of discriminative performance that averaged about 75 percent correct at the end of a month of training, after beginning the investigation near the chance level of 25 percent correct. Most important were the results of two later days of testing performance with the original training slides and with brand-new slides of cats, flowers, cars, and chairs. Accuracy to the old slides averaged about 80 percent and accuracy to the new slides averaged about 65 percent.

These results suggest that the pigeons had learned general object concepts that permitted them to categorize both old and new stimuli from four human language classes. Although the pigeons' testing performance was highly discriminative to both old and new stimuli, accuracy was reliably higher to the old pictures than to the new ones, perhaps because the birds memorized some or all of the old slides. (*See* **Concept Learning and Categorization: Models**)

In another project, three groups of different pigeons were also trained to categorize photographic slides. The three groups were given 48 daily training trials comprising: 12 copies of one

example from the categories cat, flower, car, and chair (group 1); three copies of four examples from the same categories (group 4); or one copy of 12 examples from the same categories (group 12). The rate of learning was a negative function of the number of examples per category. Of additional importance were the results of a generalization test with 32 novel stimuli: eight from each category. Now, accuracy was a positive function of the number of training examples.

Although increasing the difficulty of original learning, greater numbers of training examples per category enhanced the accuracy of generalization performance, perhaps because of the increased likelihood that any given test stimulus resembled one or more of the remembered training stimuli. These data are not only orderly, but they neatly correspond with a large body of research on categorization in human adults and children. Empirical parallels like these are unlikely to be coincidental; rather, they suggest a basic behavioral similarity. Whether there are other correspondences in conceptualization by humans and animals is a topic of current research.

Abstract Concepts

One of the empirical hallmarks of conceptualization is that discriminative responding is independent of the specific details of the prevailing stimuli. Note that it was imperative in research on object concepts to show that the discriminative responding that was established to a familiar set of training stimuli also extended to a novel set of testing stimuli. To have conceptualized 'chairs' requires that new chairs occasion the same response as old ones. (*See* **Representations, Abstract and Concrete**)

An even more advanced level of conceptualization may be achieved when organisms respond 'same' or 'different' to several simultaneously or successively presented stimuli. Again, the critical test comes when novel stimuli are given, in order to see whether the organism appropriately responds to them.

Here, too, great doubt has historically been expressed that animals can learn abstract concepts. The seventeenth-century English philosopher John Locke, in particular, believed that abstract conceptualization represented the key intellectual divide between humans and animals.

Despite his firm conviction that nonhuman animals were incapable of abstraction and conceptualization, Locke may have been premature in his assessment. Mounting behavioral evidence suggests that nonhuman animals – even pigeons – can form abstract concepts.

In this research, pigeons received food reward for pecking one button (for example, red) when they were shown any displays that pictured 16 copies of the same computer icon, and for pecking a second button (for example, green) when they were shown any other displays that pictured one copy of 16 different computer icons. Incorrect responses led to nonreward.

After the pigeons had reached a high level of discrimination accuracy (exceeding 80 percent correct choices when the chance score was 50 percent correct), in testing sessions the birds were shown brand-new Same displays and brand-new Different displays that pictured icons that they had never seen before. In various experiments, discrimination accuracy averaged from 83 to 93 percent correct to the Same displays and to the Different displays from the training set; accuracy averaged from 71 to 79 percent correct to the Same displays and to the Different displays from the testing set. These high levels of discrimination accuracy to both familiar and novel displays are consistent with the pigeons' having learned an abstract same–different concept.

Not only has same–different conceptualization been demonstrated for pigeons that were given simultaneous displays of visual items, but also for pigeons that were given successive lists of visual items. In the latter case, the 16-list icons were presented one at a time, thereby requiring the memory of prior items in order to decide whether the just-presented list comprised identical or nonidentical items.

Despite this important procedural change, discrimination accuracy averaged 94 percent correct to the Same lists and to the Different lists from the set of training icons; further, accuracy averaged 72 percent correct to the Same lists and to the Different lists from the set of testing icons. Abstract conceptualization may thus be within the ken of even the pigeon.

DIFFERENT APPROACHES TO COMPARATIVE INTELLIGENCE

The foregoing has described a sampling of investigations whose results suggest that advanced intellectual processes – including learning, memory, and conceptualization – can be comparatively studied. The results of these investigations disclose many salient similarities in the behavior of human and nonhuman animals.

Should we thus agree with Thorndike that intelligence is pretty much the same throughout the

animal kingdom? Perhaps not. But the many similarities that have been uncovered so far are simply too provocative not to merit further investigation. However, although they are surprisingly few in number, species differences in intelligence have been found and many more may ultimately surface if we assiduously seek them.

The Process Approach

Despite the prominence of the above line of research, some authors have questioned the value of the comparative study of such distantly related species as pigeons, rats, cats, dogs, monkeys, and humans. These critics have insisted that any comparisons among species that are not close evolutionary relatives of one another are capricious and of doubtful biological significance.

These critics have also argued that the seemingly expedient selection of distantly related species is an anthropocentric or human-centered strategy that springs directly from Darwin's hypothesis of mental continuity. Indeed, to such anthropocentric theorists, the true vision of comparative psychology is as a distinctly human psychology whose prime purpose is to understand and to specify the rules of human psychological functioning, as well as their degree of zoological generality.

A different and perhaps more accurate picture of this line of research in the comparative psychology of intelligence focuses on the matter of process. Thus, we are interested in establishing whether some intellectual process is involved in human and nonhuman behavior and the degree to which that process reflects common behavioral and biological mechanisms.

For example, the comparative study of the laws of reinforcement and extinction could speak to the nature and generality of these adaptive processes. If these behavioral processes are indeed general, then animal models and subsequent neuroanatomical and neurophysiological research might shed great light on their underlying biology.

Furthermore, if most extant species of animals similarly obey the laws of reinforcement and extinction, then there is good reason to believe that a common ancestor of all of these species must also have been subject to the same behavioral laws, perhaps because of common selection pressures. It would not make much sense for each species to have independently devised the same evolutionary solution to a common survival problem.

The Phylogenetic Approach

It is also possible more directly to develop the comparative psychology of intelligence from a phylogenetic perspective. In this approach, researchers study the behavior of more strategically selected species in order to elucidate the natural and evolutionary histories of the chosen animals. Adopting this more ethological and ecological approach means that one may compare either closely related species that face divergent survival demands or unrelated species that face similar survival demands. Such phylogenetic study may not only help to identify the selection pressures for complex behavior and intelligence, but it might also help to identify any specialized behavioral or intellectual adaptations. (*See* **Learning and Memory, the Ecology of**)

Take, for example, the spatial memory of animals that do not always eat the food that they find while they are foraging; rather, these animals may store the food in small reserves to which they can return for later meals. Such a food-storing strategy makes especially good sense when food is abundant at some times of the year, but scarce at other times.

One famous food-storing animal is Clark's nutcracker, a bird that lives in high mountainous regions of the Southwestern United States. These nutcrackers collect pine seeds in a small pouch under their tongue; the birds later drive those seeds into the soil with their beak. Field observations indicate that the nutcrackers return to small feeding caches many months later in order to retrieve those seeds, even under snow cover; in autumn, 33,000 seeds may be stored in 2500 caches for later recovery in winter and spring.

This is truly a remarkable feat of spatial memory. But what does it imply about the general memory ability of this particular species? In order to find out, laboratory experiments were conducted with this species and with three other bird species that are not proficient at storing and recovering food. Two general types of memory tasks were devised that were variants of delayed matching-to-sample: one task required memory for the location of a prior stimulus and the other task required memory for the color of a prior stimulus.

Clark's nutcrackers easily won the contest for spatial memory; but, they were in the middle of the pack in the contest for color memory. These data suggest that Clark's nutcrackers do not have generally exceptional memory ability; rather, they possess a more highly advanced spatial memory that may be a very special adaptation to their

particular evolutionary niche. Other special intellectual adaptations may include the pigeon's homing ability, the rat's rapid learning of taste–illness associations, and the human's propensity to learn language.

Reconciliation

These two different approaches to the comparative study of intelligence may seem to be at odds with one another. But, there really is no inherent conflict because the two programs ask different questions about animal intelligence. The process approach concentrates on a few taxonomically distant 'focus' animals in order to understand the general processes that are shared among species, whereas the phylogenetic approach concentrates on more closely related species in order to elucidate the intellectual mechanisms of specific adaptive behaviors and their ecological determinants. Together, the parallel use of both strategies should make for a powerful and complementary alliance for the future study of animal intelligence.

CONCLUSION

This brief review of research in the comparative psychology of intelligence suggests that we have merely scratched the surface in our quest to understand the nature of animal intelligence. Are animals capable of even greater cognitive feats? Can intelligence be demonstrated in animals without backbones? Are the same behavioral results the product of common brain mechanisms? (*See* **Learning in Simple Organisms**)

These questions are among the most important for the science of comparative psychology as it embarks upon its second century. Answering them will help us to tackle the most challenging question of all: how did intelligence evolve? This question was famously posed by Darwin, who hypothesized that humans and animals not only share common natural origins, but that they share common emotions and intelligence. Answering these questions with hard scientific evidence, not with beguiling anecdotes, is the business of our field.

Of course, the more we learn about the intelligence of animals, the more similar to us they seem to be. Such similarity often gives rise to empathy, leading some individuals to question whether animals should ever be the objects of scientific investigation. But, continued scientific inquiry is not only critical to answering key theoretical questions, it is also crucial to solving vexing practical problems.

Basic research into behavioral principles is vital to remedying such maladies as obesity, drug abuse, and AIDS – all the result of human behavior. That research may as well help us to contend with other serious societal woes such as crime, pollution, and overpopulation – also the products of human action.

Animals too may profit by such study, as we try to preserve and to expand their natural habitats. Knowledge of animal behavior and intelligence may also enable us to protect natural animal populations or even to increase them through repopulation efforts coordinated through zoos and animal sanctuaries.

Further Reading

Bitterman ME (1975) The comparative analysis of learning. *Science* **188**: 699–709.

Budiansky S (1998) *If a Lion Could Talk*. New York, NY: Free Press.

Domjan M (1987) Comparative psychology and the study of animal learning. *Journal of Comparative Psychology* **101**: 237–241.

Gallistel CR (1989) Animal cognition: the representation of space, time, and number. *Annual Review of Psychology* **40**: 155–189.

Macphail EM (1985) Vertebrate intelligence: the null hypothesis. *Philosophical Transactions of the Royal Society of London* **B308**: 37–51.

Roberts WA (1998) *Principles of Animal Cognition*. New York, NY: McGraw-Hill.

Roitblat HL, Bever TG and Terrace HS (eds) (1984) *Animal Cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Shettleworth SJ (1993) Varieties of learning and memory in animals. *Journal of Experimental Psychology: Animal Behavior Processes* **19**: 5–14.

Spear NE, Miller JS and Jagielo JA (1990) Animal memory and learning. *Annual Review of Psychology* **41**: 169–211.

Wasserman EA (1993) Comparative cognition: beginning the second century of the study of animal intelligence. *Psychological Bulletin* **113**: 211–228.

Wasserman EA (1995) The conceptual abilities of pigeons. *American Scientist* **83**: 246–255.

Wasserman EA (1997) Animal cognition: past, present, and future. *Journal of Experimental Psychology: Animal Behavior Processes* **23**: 123–135.

Weiskrantz L (ed.) (1985) *Animal Intelligence*. New York, NY: Oxford University Press.

Concept Learning

Introductory article

Bradley C Love, University of Texas, Austin, Texas, USA

CONTENTS

Concept learning is the process of acquiring knowledge structures that enable an agent to make predictive inferences.

INTRODUCTION

The human species evolves to meet challenges in the environment. Unfortunately, evolution is a slow 'learning' process. Evolution can only help us address aspects of our environment that are not very variable and that are stable over a long period of time. Of course, many aspects of our environment are constantly undergoing change. Accordingly, many concepts have to be learned *de novo* by each individual. For example, a radiologist is not born knowing how to interpret x-ray images. It is hard to imagine how that particular skill could evolve.

Concept learning is integral to the survival of any agent (e.g. a human, an animal, a robot, etc.) operating in a complex and changing environment. A concept is a mental representation that is often derived from experiences with specific instances. We often develop concepts of categories (i.e. collections of objects) in the world. Without acquired concepts, we would be unable to make sense of the world around us. Every new object encountered would appear completely novel and we would not know how to interact with it. For example, the first time a child encounters a hot stove he may get burned. When the child visits a friend's house and encounters another stove, it is unlikely the child will touch it, even though the new stove may differ in a number of ways from the original stove (e.g. size, color, design, etc.). If the child did not generalize from his experiences and form a concept of stoves, he would go through life with burned hands. (*See* **Categorization, Development of**; **Generalization**)

One basic question is how do we learn new concepts? Philosophers, psychologists, and computer scientists have all pondered this question. In the following sections, three basic views (i.e. models) of concept learning and concept representation (i.e. what is stored as a consequence of learning) will be examined. The first account posits that concepts consist of rules. A more recent account holds that concepts are represented as prototypes. A prototype can be thought of as the average example of a concept. A third account of concepts is the exemplar view. The exemplar view holds that concepts are nothing more than a collection of stored exemplars (i.e. examples of the concept). We will evaluate the relative merits of each of these accounts of human concept learning. All three accounts correctly characterize some aspects of human concept learning. After evaluating these three accounts, we will discuss more modern neural network models of concept learning. Neural network models embody some of the characteristics of rule, prototype, and exemplar approaches. (*See* **Concept Learning and Categorization: Models**; **Classifier Systems**; **Concepts, Philosophical Issues about**; **Conceptual Representations in Psychology**)

RULES

The classical view of concepts holds that categories are defined by logical rules. In Figure 1, any item that is a square is a member of category A. This simple rule determines category membership. According to the rule view, our concept of category A can be represented by this simple rule. Discovering this rule would involve a rational hypothesis-testing procedure. This procedure attempts to discover a rule that is satisfied by all of the positive examples of a concept, but none of the negative examples of the concept (i.e. items that are members of other categories). In trying to come up with such a rule for category A, one might first try the rule 'if dark, then in category A'. After rejecting this rule (because there are counterexamples), other rules would be tested (starting with simple rules

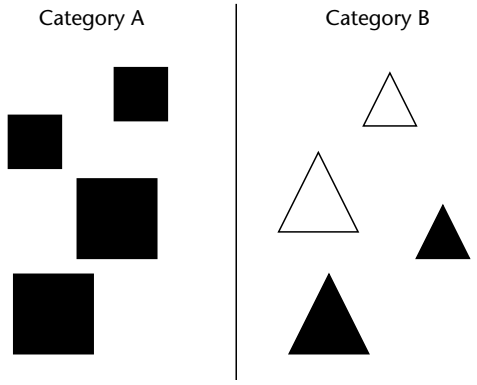


Figure 1. Examples of category A and category B.

and progressing towards more complex rules) until the correct rule is eventually discovered. For example, in learning about birds, one might first try the rule 'if it flies, then it is a bird.' This rule works pretty well, but not perfectly (penguins do not fly and bats do). Another simple rule like 'if it has feathers, then it is a bird' would not work either because a pillow filled with feathers is not a bird. Eventually, a more complex rule might be discovered like 'if it has feathers and lays eggs, then it is a bird'.

Although rules can in principle provide a concise representation of a concept, often more elaborate representations would serve us better. Concept representation needs to be richer than a simple rule because we use concepts for much more than simply classifying objects we encounter. For instance, we often use concepts to support inference (e.g. a child infers members of the category stove can be dangerously hot). Using categories to make inferences is a very important use of concepts. Knowing something is an example of a concept tells us a great deal about the item. For example, if you can classify a politician from the USA as a Republican, you can readily infer the politician's position on a number of issues. The point is that our representations of concepts need to include information beyond what is needed to classify items as examples of the concept. For example, the rule 'if square, then in category A' correctly classifies all members of category A in Figure 1, but it does not capture the knowledge that all category A members are dark. One problem with rule representations of concepts is that potentially useful information is discarded.

The biggest problem with the rule approach to concepts is that most of our everyday categories do not seem to be describable by a tractable rule. To demonstrate this point, Wittgenstein noted that the concept game lacks a defining property. Most games are fun, but Russian roulette is not fun. Most games are competitive, but ring around the roses is not competitive. While most games have characteristics in common, there is not a rule that unifies them all. Rather, we can think of the members of the category game as being organized around a *family resemblance* structure (analogous to how members of your family resemble one another).

A related weakness of the rule account of concepts is that examples of a concept differ in their *typicality*. If all a concept consisted of was a rule that determined membership, then all examples should have equal status. According to the rule account, all that should matter is whether an item satisfies the rule. Our concepts do not seem to have this definitive flavor. For example, some games are better examples of the category game than others. Basketball is a very typical example of the category game. Children play basketball in a playground, it is competitive, there are two teams, each team consists of multiple players, you score points, etc.

Basketball is a typical example of the category of games because it has many characteristics in common with other games. On the other hand, Russian roulette is not a very typical game – it requires a gun and one of the two players dies. Russian roulette does not have many properties in common with other games. In terms of family resemblance structure, we can think of basketball as having a central position and Russian roulette being a distant cousin to the other family members. These findings extend to categories in which a simple classification rule exists. For example, people judge the number three to be a more typical odd number than the number forty-seven even though membership in the category 'odd number' can be defined by a simple rule.

The fact that category membership follows a gradient as opposed to being all or none affords us flexibility in how we apply our concepts. Of course, this flexibility can lead to ambiguity. Consider the concept mother. It is a concept that we are all familiar with that seems straightforward – a mother is a woman who becomes pregnant and gives birth to a child. But what about a woman who adopts a neglected infant and raises it in a nurturing environment? Is the birth mother who neglected the infant a mother? What if a woman is implanted with an embryo from another woman? Court cases over maternity arise because the concept of motherhood is ambiguous. The concept exhibits greater flexibility and productivity than is even indicated above.

For example, is it proper to refer to an architect as the mother of a building? All the above examples of the concept mother share a family resemblance structure (i.e. they are organized around some commonalities), but the concept is not rule based. Some examples of the concept mother are better than others.

PROTOTYPES

The prototype approach to concept learning and representation was developed by Rosch and colleagues to address some of the shortcomings of the rule approach. Prototype models represent information about all the possible properties (i.e. *stimulus dimensions*), instead of focusing on only a few properties like rule models do. The prototype of a category is a summary of all of its members. Mathematically, the prototype is the average or central tendency of all category members. Figure 2 displays the prototypes for two categories, simply named categories A and B. Notice that all the items differ in size and luminance (i.e. there are two stimulus dimensions) and that the prototype is located amidst all of its category members. The prototype for each category has the average value of both the stimulus dimensions of size and luminance for the members of its category. (*See* **Prototype Representations**)

The prototype of a category is used to represent the category. According to the prototype model, a novel item is classified as a member of the category whose prototype it is most similar to. For example, a large bright item would be classified as a member of category B because category B's prototype is large and bright (see Figure 2). The position of the prototype is updated when new examples of the category are encountered. For example, if one encountered a very small and dark item that is a member of category A, then category A's prototype would move slightly towards the bottom left corner in Figure 2. As an outcome of learning, the position of the prototype shifts towards the newest category member in order to take it into account. A prototype can be very useful for determining category membership in domains where there are many stimulus dimensions that each provide information useful for determining category membership, but no dimension is definitive. For example, members of a family may tend to be tall, have large noses, a medium complexion, brown eyes, and good muscle tone, but no family member possesses all of these traits. Matching on some subset of these traits would provide evidence for being a family member. (*See* **Multidimensional Scaling**; **Similarity**)

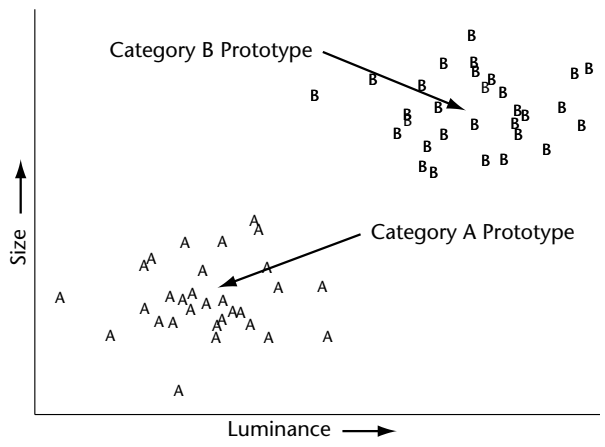


Figure 2. Two categories and their prototypes.

Notice the economy of the prototype approach. Each cloud of examples in Figure 2 can be represented by just the prototype. The prototype is intended to capture the critical structure in the environment without having to encode every detail or example. It is also fairly simple to determine which category a novel item belongs to by determining which category prototype is most similar to the item.

Unlike the rule approach, the prototype model can account for typicality effects. According to the prototype model, the more typical category members should be those members that are most similar to the prototype. In Figure 2, similarity can be viewed in geometric terms – the closer items are together in the plot, the more similar they are. Thus, the most typical items for categories A and B are those that are closest to the appropriate prototype. Accordingly, the prototype approach can explain why robins are more typical birds than penguins. The bird prototype represents the average bird: has wings, has feathers, can fly, can sing, lives in trees, lays eggs, etc. Robins share all of these properties with the prototype, whereas penguins differ in a number of ways (e.g. penguins cannot fly, but do swim). Extending this line of reasoning, the best example of a category should be the prototype, even if the actual prototype has never been viewed (or does not even exist). Indeed, numerous learning studies support this conjecture. After viewing a series of examples of a category, human participants are more likely to categorize the prototype as a category member (even though they never actually viewed the prototype) than they are to categorize an item they have seen before as a category member.

Because the prototype approach does not represent concepts in terms of a logical rule that is either satisfied or not, it can explain how category membership has a graded structure that is not all or none. Some examples of a category are simply better examples than other examples. Also, categories do not need to be defined in terms of logical rules, but are rather defined in terms of family resemblance to the prototype. In other words, members of a category need not share a common defining thread, but rather can have many characteristic threads in common with one another.

The prototype approach, while preferable to the rule approach for the reasons just discussed, does fail to account for important aspects of human concept learning. The main problem with the prototype model is that it does not retain enough information about examples encountered in learning. For instance, prototypes do not store any information about the frequency of each category, yet people are sensitive to frequency information. If an item was about equally similar to the prototype of two different categories and one category was one hundred times larger than the other, people would be more likely to assign the item to the more common category (under most circumstances).

People are also sensitive to the variability along stimulus dimensions. To use Rips' example, a circular object with a 10 cm diameter may be more similar to a US quarter (which is about 2.5 cm in diameter) than to a pizza (which is much larger). Nevertheless, the novel object is more likely to be classified as a pizza than a quarter because quarters display very little variability in their diameters whereas pizzas can vary in size.

Finally, prototypes are not sensitive to the correlations and substructure within a category. For example, a prototype model would not be able to represent that spoons tend to be large and made of wood or small and made of steel. These two subgroups would simply be averaged together into one prototype. This averaging makes some categories unlearnable with a prototype model. One example of such a category structure is shown in Figure 3. Each category consists of two subgroups. Members of category A are either small and dark or they are large and light, whereas members of category B are either large and dark or they are small and light. The prototypes for the two categories are both in the centre of the stimulus space (i.e. medium size and medium luminance). Items cannot be classified correctly by which prototype they are most similar to because the prototypes provide little guidance.

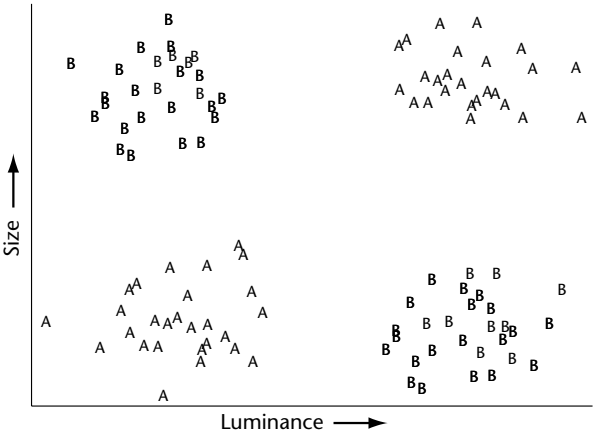


Figure 3. Two categories, each containing two subgroups.

In general, prototype models can only be used to learn category structures that are *linearly separable*. A learning problem involving two categories is linearly separable when a line or plane can be drawn that separates all the members of the two categories. The category structure shown in Figure 2 is linearly separable because a diagonal line can be drawn that separates the category A and B members (i.e. the category A members fall on one side of the line and the category B members fall on the other side of the line). Thus, this category structure can be learned with a prototype model. The category structure illustrated in Figure 3 is nonlinear – no single line can be drawn to segregate the category A and B members. Mathematically, a category structure is linearly separable when there exists a weighting of the feature dimensions that yields an additive rule that correctly indicates one category when the sum is below a chosen threshold and the other category when the sum is above the threshold.

The inability of the prototype model to learn nonlinear category structures detracts from its worth as a model of human concept learning because people are not biased against learning nonlinear category structures. Some nonlinear category structures are actually easier to acquire than linear category structures. For example, it seems quite natural that small birds sing, whereas large birds do not sing. Many categories have subtypes within them that we naturally pick out. One way for the prototype model to address this learnability problem is to include complex features that represent the presence of multiple simple features (e.g. large and blue). Unfortunately, this approach quickly becomes unwieldy as the number of stimulus dimensions increases.

EXEMPLARS

Exemplar models address many of the shortcomings of the prototype model. Exemplar models store every training example in memory instead of just the prototype (i.e. the summary) of each category. By retaining all of the information from training, exemplar models are sensitive to the frequency, the variability, and the correlations among items. For the learning problem illustrated in Figure 2, an exemplar model would store every training example. New items are classified by how similar they are to all items in memory (not just the prototype). For the category structure illustrated in Figure 2, the pairwise similarity of a novel item and every stored item would be calculated. If the novel item tended to be more similar to the category A members (i.e. the item was small and dark) than the category B members, then the novel item would be classified as a member of category A.

One aspect of exemplar models that seems counterintuitive is their lack of any abstraction in category representation. It seems that humans do learn something more abstract about categories than a list of examples. Surprisingly, exemplar models are capable of displaying abstraction. For instance, exemplar models can correctly predict that humans more strongly endorse the underlying prototype (even if it has not been seen) than an actual item that has been studied (a piece of evidence previously cited in favor of the prototype model). How could this be possible without the prototype actually being stored? It would be impossible if exemplar models simply functioned by retrieving the exemplar in memory that was most similar to the current item and classified the current item in the same category as the retrieved exemplar (this is essentially how processing works in a prototype model, except that a prototype is stored in memory instead of a bunch of exemplars).

Instead, exemplar models engage in more sophisticated processing and calculate the similarity between the current item (the item that is to be classified) and every item in memory. Some exemplars in memory will be very similar to the current item, whereas others will not be very similar. The current item is classified in the category in which the sum of its similarities to all the exemplars is greatest. When a previously unseen prototype is presented to an exemplar model it can be endorsed as a category member more strongly than a previously seen item. The prototype (which is the central tendency of the category) will tend to be somewhat similar to every item in the category, whereas any given non-prototype item will tend to be very similar to some items (especially itself!) in memory, but not so similar to other items. Overall, the prototypical item can display an advantage over an item that has actually been studied. Abstraction in an exemplar model is indirect and results from processing (i.e. calculating and summing pairwise similarities), whereas abstraction in a prototype model is rather direct (i.e. prototypes are stored).

The exemplar model does seem to make some questionable assumptions. For example, exemplar models store every training example which seems excessive. Also, every exemplar is retrieved from memory every time an item is classified. In addition to these assumptions, one worries that the exemplar model does not make strong enough theoretical commitments because it retains all information about training and contains a great deal of flexibility in how it processes information. These issues are currently being resolved by researchers. On the whole, exemplar models seem to be a more viable approach to understanding human concept learning than existing prototype or rule-based approaches, but there is still room for further work. (*See* **Computational Models of Cognition: Constraining**)

NEURAL NETWORK MODELS

Neural network models are intended to learn in a manner analogous to how the brain learns. A neural network consists of layers of neuron-like units that connect to units in other layers. Units can excite and inhibit one another across these connections. An item is represented at the input layer (the first layer) and passes activity to more advanced layers in the network until it reaches the output layer which determines the category the item is a member of (e.g. if the unit in the output layer representing category B is the most activated, then the item is classified as a member of category B). Each unit integrates all the activity originating from the layer below via its connections and passes this summed activity through a transfer function to generate its own output which is passed on to the next layer. Figure 4 illustrates a feedforward neural network with an input, hidden, and output layer. (*See* **Connectionism**)

The connections between units are altered as a result of learning in order to minimize the prediction error (i.e. the weights are altered in order to correctly classify items). Sophisticated learning algorithms dictate how the weights should be altered as a result of learning. Neural networks with only an input and output layer share many of

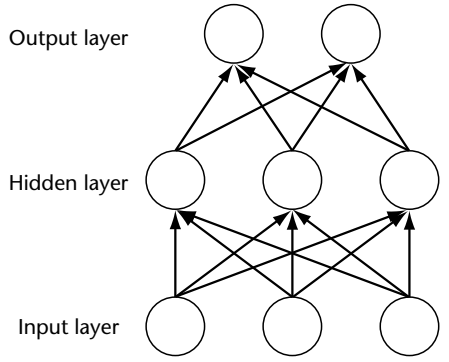


Figure 4. A typical feedforward neural network.

the limitations of the prototype model – they can only learn linearly separable functions (i.e. simple category structures). More complicated neural networks with a hidden layer (and nonlinear transfer functions) can learn just about any category structure. However, neural networks of this variety are not very good models of human concept learning because they tend to learn problems quickly that people learn slowly and vice versa.

Neural network models that are conceptually related to rule, prototype, and exemplar models have been successful as models of human concept learning. For example, the ALCOVE model replaces the hidden layer in Figure 4 with encoded exemplars. In other words, units in the hidden layer are added as exemplars are encountered. This exemplar neural network model, which combines an exemplar representation of concepts with the powerful learning algorithms of neural networks, does a good job of accounting for aspects of human concept learning. The SUSTAIN model is a neural network model that combines aspects of both exemplar and prototype models. SUSTAIN initially begins like a prototype model, but it can store exemplars (which themselves can later evolve into prototypes) when prediction errors occur. For the problem illustrated in Figure 3, SUSTAIN would form four prototypes that correspond to the four clusters of items. The ability to store multiple prototypes per category allows SUSTAIN to avoid the problems that plague prototype models. Both ALCOVE and SUSTAIN also incorporate rule-like dynamics. These models learn to attend to the most relevant stimulus dimensions and neglect the less meaningful dimensions, much like how rule models tend to focus on a limited number of stimulus dimensions (e.g. if it is *large*, then it is in category A).

CONCLUSIONS AND FUTURE DIRECTIONS

From this brief review of concept learning models we saw that the progression from rule models to prototype models to exemplar models was marked by a shift towards more concrete representations (i.e. more information about the training examples is retained), greater fluidity (i.e. category boundaries are not seen as rigid), and more sophisticated processing at decision time (exemplar models are the quintessential case – all abstraction is done after the training examples are encoded). Although all three approaches have their shortcomings, they all reflect some aspects of human concept learning. The successful neural network models of concept learning retain characteristics of all three approaches. Like the rule approach, these neural network models acknowledge the utility of strategically focusing on a subset of stimulus dimensions. If a stimulus dimension is irrelevant to a learning problem, the models will ignore the dimension and not be distracted by it. Like prototype models, some of these neural network models form abstractions which can assist generalization and reduce storage requirements. Like exemplar models, these neural network models are quite fluid, can encode individual exemplars, and engage in sophisticated processing at decision time.

One important aspect of concept learning that these models do not address is the influence of prior knowledge. Our prior knowledge exhibits strong influences on what we learn from a series of examples. For example, even if all the blue cars on a mechanic's lot have transmission problems and none of the red cars do, the mechanic would never predict that blue cars in general have transmission problems. Certainly, the mechanic would not paint a car red in the hope of repairing it. The mechanic's prior knowledge and theories of how cars function preclude this association. Instead, the mechanic is oriented towards more fruitful solutions. One important challenge for concept learning models is to illuminate how prior knowledge affects our interpretation of examples. Conversely, more work is needed in understanding how examples we encounter affect our theories of the world.

Further Reading

Lakoff G (1987) *Women, Fire, and Dangerous Things: What Categories Tell Us About the Nature of Thought*. Chicago, IL: University of Chicago Press.

Medin DL (1998) Concepts and conceptual structure. In: Thagard P (ed.) *Mind Readings*, pp. 93–126. Cambridge, MA: MIT Press.

Mervis CB and Rosch E (1981) Categorization of natural objects. In: Rosenzweig MR and Porter LW (eds) *Annual Review of Psychology* **32**: 89–115.

Rumelhart D (1989) The architecture of mind: a connectionist approach. In: Posner MI (ed.) *Foundations of Cognitive Science*, pp. 133–159. Cambridge, MA: MIT Press.

Wisniewski EJ (in press) Concepts and categorization. In: Medin DL (ed.) *The Steven's Handbook of Experimental Psychology*. New York, NY: John Wiley and Sons.

Conceptual Change

Intermediate article

Paul Thagard, University of Waterloo, Waterloo, Ontario, Canada

CONTENTS

Conceptual change is the creation and alteration of mental representations that correspond to words. It is an important part of learning in science and everyday life.

INTRODUCTION

Concepts are mental representations corresponding to words. For example, the concept 'dog' is a mental structure that corresponds to the word 'dog' and refers to dogs in the world. Conceptual change is produced by mental processes that create and alter such mental representations. Explaining how conceptual change works is important for understanding the growth of scientific knowledge, the development of children's thinking, and the education of students in fields such as science and mathematics. In each of these kinds of learning, a theory of conceptual change is needed that can answer such questions as the following. What is the nature of the concepts that are learned? What kinds of changes do concepts undergo? What are the mental processes that produce different kinds of conceptual change? It is also interesting to inquire whether the processes of conceptual change in scientists, young children, and students are similar or different.

TYPES OF CONCEPTUAL CHANGE

The simplest type of conceptual change is when people learn a new concept. A more challenging type occurs when existing concepts must be adjusted and reorganized to accommodate new information: in such cases, the meaning of concepts changes in relation to other concepts and the world. In radical conceptual change, the development of knowledge involves a shift in which a collection of important concepts undergo alterations in meaning. In such cases, learning is not simply a matter of accumulating new concepts and beliefs; it also requires substantial revision and restructuring of mental representations.

CONCEPTUAL CHANGE IN SCIENTISTS

The problem of conceptual change in science was first highlighted in Thomas Kuhn's famous book, *The Structure of Scientific Revolutions* (Kuhn, 1962). He challenged the prevailing view that scientific knowledge grows cumulatively by progressively adding to the stock of available theories and concepts. Instead, Kuhn proposed that the development of science often involves revolutionary changes in which one theory or paradigm is replaced by a radically different one. For example, the acceptance of the Copernican theory that the earth revolves around the sun required the rejection of the Ptolemaic theory that the sun revolved around the earth. Replacement was not merely a matter of one theory being substituted for another, but also involved shifts in meaning of the concepts used in the theories. In the Copernican revolution, for example, the concept 'planet' shifted to include the earth and exclude the sun and moon. According to Kuhn, radical differences between theories make it difficult to establish rationally that one is better than another.

Kuhn distinguished between normal science, in which a dominant paradigm is taken for granted, and revolutionary science, in which the dominant paradigm is replaced by a radically new one. The main activity in normal science is puzzle solving, which deals with problems within the scope and constraints of the dominant way of thinking. Scientists pursue normal science until there is an accumulation of anomalies, which are problems that the paradigm fails to solve. For example, in the eighteenth century the prevailing theory of combustion based on phlogiston, a substance supposed to be given off by burning objects, encountered the anomaly that objects gain rather than lose weight

during combustion. Scientists attempt to deal with individual anomalies as puzzles to be solved with the tools provided by the paradigm they accept, but the accumulation of anomalies produces a state of crisis in which scientists begin to consider the need for new theories. When a new paradigm is conceived that can solve the problems that were anomalous for the old one, a scientific revolution occurs and a new theory becomes accepted. Kuhn's favorite examples of scientific revolutions include the Copernican revolution, the chemical revolution in which Lavoisier's oxygen theory of combustion replaced the phlogiston theory, and the revolution in physics in which relativity theory was adopted.

Before Kuhn, science was generally viewed as a cumulative process in which new theories built on the successes of previous ones. Kuhn insisted that scientific revolutions are noncumulative episodes in which an older paradigm is replaced by an incompatible new one. He even suggested that the new and old theories are incommensurable with each other, that is, there may be no logical means for objectively choosing between them. A major source of incommensurability is the use by the different paradigms of very different concepts. For example, it might seem that the Newtonian physics and relativity theory both use the concept of mass, but Einsteinian mass can be converted into energy whereas Newtonian mass is conserved. Thus for Kuhn a major aspect of scientific revolutions was radical conceptual change.

In *Conceptual Revolutions*, Thagard (1992) offered a comprehensive account of the kinds of conceptual changes that have occurred in the major revolutions in the history of science. Most scientific revolutions involve the introduction of new concepts, such as Newton's gravitational force, Lavoisier's oxygen, Darwin's natural selection, and Wegener's continental drift. In addition, revolutions usually involve reclassification in which a concept changes its place in the hierarchy of kinds, just as Copernicus reclassified earth as a planet, Darwin reclassified humans as a kind of animal, and the cognitive revolution in psychology reclassified thinking as a kind of computation. Even more radically, the principle of classification sometimes changes, as when Darwin argued that species should be organized into kinds on the basis of evolutionary history rather than similarity. Like many other philosophers of science, Thagard argued that Kuhn had overestimated the conceptual differences between theories, so that conceptual change did not prevent one theory from being rationally preferred to another on the basis of its explanatory power. Nevertheless, he accepted Kuhn's basic contention that new theories often have very different conceptual systems from the ones they replace.

Philosophers and psychologists have discussed the cognitive mechanisms by which new conceptual systems in science are constructed. These include conceptual combination, in which a concept such as 'sound wave' is constructed out of the previously existing concepts 'sound' and 'wave'. New concepts are rarely derived directly from experience, but instead are built up from previously existing concepts. A concept produced by conceptual combination need not be a simple sum of the original concepts, but instead can involve emergent properties. For example, the concept 'blind lawyer' has characteristics not found in either 'blind' or 'lawyer': people use causal reasoning to conclude that a blind lawyer must be courageous.

Another creative mechanism is analogy, in which new scientific concepts are formed by adapting and transforming previous concepts. For example, Darwin's concept of natural selection was based in part on his familiarity with artificial selection practiced by breeders who produced new varieties of plants and animals. Maxwell developed concepts of electromagnetism using mechanical analogies (Nersessian, 1992), and Kepler extensively used analogies to develop new concepts concerning light and motion (Gentner *et al*., 1997).

Once a new conceptual system has been constructed by mechanisms such as combination and analogy, it becomes a contender to replace an existing conceptual system. The major cognitive mechanism for such large-scale conceptual change is explanatory coherence: scientists adopt a new theory along with its conceptual system because it provides a better explanation of the evidence and is more coherent with other beliefs (Thagard, 1992). Of course, most conceptual change in science does not involve such large-scale shifts in which conceptual systems are substantially altered, but rather the introduction of new concepts that fit in with existing conceptual schemes and theories.

CONCEPTUAL CHANGE IN YOUNG CHILDREN

Young children acquire a wealth of new concepts as their knowledge of language and the world increases. The average high-school graduate in the USA knows around 60 000 root words, which must have been acquired at a rate better than 10 per day. Presumably, children have concepts that are mental representations corresponding to all these words, so how can we account for their acquisition in such large numbers? Much conceptual change is

straightforwardly cumulative, as children simply add new concepts such as 'dog' and 'ice cream' to their mental systems. However, some developmental psychologists have argued that conceptual development in children is like conceptual change in science, in that it sometimes requires substantial revisions of existing conceptual schemes.

Susan Carey argued that children's acquisition of biological knowledge between the ages of 4–10 years involves considerable conceptual reorganization (Carey, 1985). In particular, the concepts 'alive' and 'animal' undergo substantial change during those years. Many 4-year-olds have difficulty naming any objects that are not alive, and take objects such as tables and clocks as being alive because they have activities or motions associated with them. By the age of 10 years, however, most children have acquired the adult concept of 'living thing'. Similarly, children under 7 years old often do not count people and insects as animals. According to Carey, children undergo a complete reorganization of knowledge of functions such as eating and sleeping and of organs such as the stomach and heart as the domain of biological knowledge becomes differentiated from the domain of knowledge of human activities. It is not just that the concepts of a 10-year-old have different relations among them than those of a 4-year-old, but more that the concepts themselves have changed as the result of additional biological knowledge. The concepts 'animal' and 'plant' coalesce into the concept 'living thing' by virtue of recognition that they are fundamentally alike. At the same time, children learn to differentiate 'dead' from 'inanimate' as two different senses of 'not alive'. Just as scientists had to learn to differentiate between heat and temperature, so children have to learn to differentiate weight from size and density. Like scientists, children have theory-like conceptual structures, and learning consists in radical alteration of such structures, not just additions to them.

Frank Keil reached similar conclusions from his studies of the development of children's concepts of biological kinds (Keil, 1989). As children gain an increasing appreciation of the biological principles that organize adults' intuitive theories of biology, they increasingly appeal to origins and internal parts in their biological classifications, reducing the impact of visible features. For example, older children are more likely to judge that a pear covered with apple skin is still a pear. In contrast, there was no similar shift for artefacts such as cup and nail, indicating that conceptual change was specific to biological kinds. Keil argues that concepts are part of coherent belief systems, so that conceptual change is closely tied in with theory change in children.

Gopnik and Meltzoff (1997) are even more emphatic in tying conceptual change to theory change. They advocate the 'theory theory', according to which the process of cognitive development in children is similar to and perhaps even identical to the process of theory development in scientists. They describe changes in understanding of objects in infants, who are born assuming a world of three-dimensional objects that have visual, auditory, and tactile features. By 6 months, infants have gained systematic, coherent knowledge about the movements of objects, but they still lack understanding of hidden objects, which develops around 9 months. Later, at around 18 months, infants acquire the ability to represent invisible movements. Gopnik and Meltzoff contend that these shifts are like theory change in science, and that there is a certain incommensurability between the concepts of the old and new theories held by the infants.

These and other studies of learning in children strongly suggest that conceptual development is not simply a matter of accumulating new concepts but also involves important changes in concepts and conceptual systems. However, the evidence is still limited for claims that children's conceptual systems are like those of scientists and that the cognitive mechanisms of change in children are like those that take place in the minds of scientists. It is possible that children's knowledge is much more fragmented than the conceptual systems that make up scientific theories such as relativity and evolution by natural selection. Scientific theories consist of hypotheses that provide unifying explanations of diverse empirical phenomena, but no one knows whether children's beliefs involve the same kind of explanatory hypotheses. Moreover, the process by which scientists come to realize that one theory is better than and should replace a previous one involves a systematic comparison of the explanatory coherence of the two theories. Belief change in children may be much more piecemeal, as isolated fragments of a new theory of objects and kinds are acquired from experience and teaching. It is possible that new ways of looking at things supplant previous ones by a process of gradual build-up of new concepts and progressive disuse of old ones, rather than by a dramatic replacement of the old theories by new ones. The view that conceptual change in children is similar to theory change in scientists has been heuristically useful in stimulating research on children's learning, but much more empirical research is needed before the analogy between children and

scientists can be accepted as showing a common set of cognitive processes.

CONCEPTUAL CHANGE IN STUDENTS

Suppose it is true that learning in children and scientists involves radical conceptual change rather than mere accumulation of new concepts and beliefs; then teaching students cannot be understood as merely providing new material to mesh with what students already know. Rather, education in science and other subjects may require a much more challenging process of dealing with the prior concepts and hypotheses that guide students' thinking. If teachers are not aware that students come to science classes with misconceptions about living things and physical processes, the teachers will not understand many of the difficulties that the students have in learning. From the perspective of conceptual change, teaching requires an active approach in which children must be engaged in building explanations that challenge concepts and beliefs that they previously held. Effective teaching may require the use of the kinds of analogical models and thought experiments that have often facilitated conceptual change in the history of science.

Chi (1992) argues that physics education is often difficult because it requires conceptual change across fundamental ontological categories such as matter, events, and abstractions. For example, naive students start with concepts of force, light, heat and current that class them as kinds of material substances, but physics students must learn to reconceptualize them as fields, which are a complex kind of event. Vosniadou and Brewer (1992) studied the development of children's knowledge of astronomy and found that children have difficulty reconciling the teaching that the earth is round with their other beliefs and observations. Children develop models that reconcile their observation-based belief that the earth is flat with what they are taught about the earth being round. For example, first-graders often believe that there are two earths – a flat one on which we live and a round one up in the sky. Other children think that the earth is a sphere, but we live inside it rather than on top of it. Thus, teaching children that the earth is round is not just a matter of telling them an additional fact, but requires them to revise their basic beliefs about the nature of the earth and other planets.

Science education is thus in part a cognitive process involving conceptual change, but it is also being increasingly recognized as a social, contextual, and emotional process (Guzzetti and Hynd, 1998). Conceptual change is a kind of mental change, but this may come about because of social interactions that students have with teachers and each other, as well as with the physical world. Motivation and emotion can greatly influence conceptual change when students acquire the intention and enthusiasm to adopt new concepts and hypotheses rather than to remain entrenched in their previous frames of mind. Future research on conceptual change will have to find ways to integrate cognitive processes with social and emotional processes that interact with them continuously.

The last section raised the question of whether conceptual change in children is like that found in scientists undergoing major theoretical changes. It is also an open question whether students need to undergo conceptual revolutions, or whether instead they can learn by a more gentle process in which new conceptual systems come to predominate over previous ones without the explanatory conflicts that occur in science. More research is needed to determine whether the cognitive mechanisms of conceptual change and theory evaluation that operate in scientists are also responsible for educational progress in science students.

References

Carey S (1985) *Conceptual Change in Childhood*. Cambridge, MA: MIT Press/Bradford Books.

Chi M (1992) Conceptual change within and across ontological categories: examples from learning and discovery in science. In: Giere R (ed.) *Cognitive Models of Science*, Minnesota Studies in the Philosophy of Science, vol. 15, pp. 129–186. Minneapolis, MN: University of Minnesota Press.

Gentner D, Brem S, Ferguson R *et al*. (1997) Analogy and creativity in the works of Johannes Kepler. In: Ward TB, Smith SM and Vaid J (eds) *Creative Thought: An Investigation of Conceptual Structures and Processes*, pp. 403–459. Washington, DC: American Psychological Association.

Gopnik A and Meltzoff AN (1997) *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Guzzetti B and Hynd C (eds) (1998) *Perspectives on Conceptual Change*. Mahwah, NJ: Lawrence Erlbaum.

Keil F (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press/Bradford Books.

Kuhn T (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Nersessian N (1992) How do scientists think? Capturing the dynamics of conceptual change in science. In: Giere R (ed.) *Cognitive Models of Science*, vol. 15, pp. 3–44. Minneapolis, MN: University of Minnesota Press.

Thagard P (1992) *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Vosniadou S and Brewer WF (1992) Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology* **24**: 535–585.

Further Reading

Ball T, Farr J and Hanson RH (eds) (1989) *Political Innovation and Conceptual Change*. Cambridge, UK: Cambridge University Press.

Carey S (2001) *Science education as conceptual change*. [http://www.house.gov/science/carey_03-04.htm]

Dietrich E and Markman AB (eds) (1999) *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*. Mahwah, NJ: Lawrence Erlbaum.

Feyerabend PK (1981) *Realism, Rationalism and Scientific Method*. Philosophical Papers, vol. 1. Cambridge, UK: Cambridge University Press.

Kunda Z, Miller D and Claire T (1990) Combining social concepts: the role of causal reasoning. *Cognitive Science* **14**: 551–577.

Nersessian N (1989) Conceptual change in science and in science education. *Synthese* **80**: 163–183.

Pearce G and Maynard P (eds) (1973) *Conceptual Change*. Dordrecht: Reidel.

Thagard P (1999) *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.

Conceptual Representations in Psychology

Introductory article

Arthur B Markman, University of Texas, Austin, Texas, USA

CONTENTS

Conceptual representation refers to the way that information about categories is stored and organized.

INTRODUCTION

Concepts are mental representations that are used to divide the world into groups that will be treated as equivalent for some purpose. Concepts may refer to objects, events, or ideas. Concepts may be used for reasoning, prediction, and communication. Some researchers have distinguished between concepts, which are the mental representations of information, and categories, which are sets of objects in the world that are grouped together. Often, however, these terms are used interchangeably.

Psychologists have explored concept representations in detail. This work has examined both within-category representation and between-category structure. Within-category representation refers to the information that describes a particular category such as 'dog'. Between-category structure refers to the relationships among different categories such as that between the categories 'dog', 'cat', and 'animal'.

WITHIN-CATEGORY REPRESENTATION

The central question about within-category representation involves the way people store information about particular concepts that enables them to classify new items (exemplars) as members of a category. Some work has looked at other uses of categories such as making predictive inferences, causal reasoning, and communication, but this discussion will focus on classification. Three broad types of within-category representation are rule-based models, similarity-based models, and theory-based models.

Rule-based Models

The classical approach to concept representation has been to seek a rule that specifies the necessary and sufficient conditions for something to be a member of a category. A property is a necessary condition for being in a category if all members of that category possess the property. A set of necessary conditions is sufficient to specify a category if all exemplars that have that set of properties are members of the category, and no exemplars that have that set of properties are members of any other category. For example, an object is a triangle if it is a three-sided closed figure. This set of features is necessary and sufficient, because all triangles are three-sided and closed. No object that has these properties can be anything but a triangle.

Unfortunately, outside formal domains like geometry, it is difficult (or perhaps impossible) to find a set of necessary and sufficient conditions that specify the members of a category. For example, it might seem at first glance that a bachelor is an unmarried adult man. While this rule correctly classifies most bachelors, there are many dubious cases. For example, Catholic priests and widowers are both unmarried adult men, but one might be hesitant to classify them as bachelors. While it is possible to continue to refine this definition, it is likely that exceptions could be found to any rule that was generated.

One approach that has been tried to save rule-based approaches has been to assume that people generate fairly simple rules that are good for classifying most exemplars, and then store exceptions to the rules separately. In the example above, the rule 'unmarried adult man' would be used to classify

most bachelors, but exceptions such as priests and widowers would be considered separately.

Similarity-based Models

Intuitively, it seems that a new exemplar is classified based on its similarity to the category. For example, an object might be classified as a bird, because it looks like a bird. This intuition has been captured by similarity-based models, which assume that people classify a new exemplar based on its similarity to some stored category representation. Similarity-based models differ from each other primarily in their assumptions about the nature of the stored category representation.

Prototype models assume that people store some average representation of an object. The average need not be identical to any actual exemplar, but rather contains the features most frequently associated with that category. For example, the typical bird might be a small animal that flies, sings, and has feathers. Not all birds have this entire set of properties (e.g. penguins do not fly), but the more of these features an exemplar possesses, the more likely it is to be a bird.

One result often taken as evidence for prototype models is that categories have a graded typicality structure: that is, people have strong intuitions about which members of a category are typical members of that category and which members are atypical. For example, robins and sparrows are generally thought to be typical birds, while chickens and emus are thought to be atypical birds. Generally, the typicality of an exemplar is related to its similarity to the prototype of the category.

A second prominent similarity-based model is the exemplar model, which posits that people store representations of each category member rather than creating a prototype. New exemplars are then classified by comparing them with all of the known exemplars. The more similar a new exemplar is to the known exemplars of a particular category, the more likely it is that the exemplar will be classified as a member of that category. Exemplar models are also able to account for graded typicality structure, because typical exemplars are similar to many members of a category, but atypical exemplars are similar to only a few members of a category.

Theory-based Models

Despite the success of similarity-based models in predicting how people classify new items, there are situations in which people classify items in a manner that violates similarity. For example, at a party, a person might be classified as drunk if he or she dives headfirst into a cake. This person is not being classified on the basis of any similarity to known exemplars of drunk people; rather, common beliefs about drunken behavior are sufficient to classify the person as drunk.

There seems to be a developmental change in people's ability to use theory-based information. When children first learn categories, they often classify on the basis of surface characteristics. For example, if told about a black cat that has a white stripe painted on its back, and a bag of smelly stuff surgically placed inside it, young children will classify it as a skunk. Older children and adults, however, will classify it as a cat, suggesting that they are able to use a theory about biological categories to classify this (rather strange) exemplar.

Which Type of Model is Right?

Of the three types of models, the rule-based models are least often used in conceptual processing. There are some situations in which people must make repeated classifications that involve a rule and exception process. However, empirical studies suggest that even when people are asked to form rules, their ability to apply the rule is influenced by the similarity of a new exemplar to those seen before.

Both similarity-based and theory-based processes are often used in categorization. There are times when people must be able to identify a new item on the basis of the similarity of its properties to those of items seen in the past. In addition, there are cases in which people's theories about a domain influence categorization. Current research is focusing on how to integrate similarity-based and theory-based approaches.

BETWEEN-CATEGORY STRUCTURE

A second important aspect of category representation involves the relationships among categories. In this section, two aspects of between-category structure are examined. First, much research has examined the hierarchical organization of categories. This work is concerned with understanding how people may categorize objects at different levels of abstraction. A second area that has received attention is people's ability to generate categories based on goals. This study of goal-derived categories also provides a window into conceptual processing.

Hierarchical Organization of Categories

If you see a small, curly-haired, four-legged living creature being walked on a leash down the street, you can classify this thing as a poodle, a dog, or an animal. That is, for any given object, there are a variety of categories to which it belongs. Many of these categories differ from each other in their degree of abstraction: 'dog' is a more abstract category than 'poodle', because all poodles are dogs, but not all dogs are poodles. Similarly, 'animal' is a more abstract category than either 'dog' or 'poodle'.

A striking aspect of this category structure is that if you show people a picture of some object, they are most likely to identify it using a category at a middle level of abstraction. For example, shown a picture of the item described in the previous paragraph, people are likely to identify it first as a dog rather than as a poodle or as an animal. This tendency has led psychologists to refer to this middle level of abstraction as the basic level. Categories more abstract than those at the basic level (e.g. animal) are called superordinate categories, and categories more specific than those at the basic level are called subordinate categories (e.g. poodle).

Basic level categories have been shown to have a number of characteristics. First, they tend to be the most abstract categories whose members have a common shape, and whose shape differs from other contrasting basic level categories. For example, dogs tend to be shaped similarly to each other and differently from other animals; in contrast, animals come in many different shapes. Second, basic level categories tend to have shorter labels than either subordinate or superordinate categories; for example, 'car' is the label for a basic level category, and the labels 'vehicle' (for the superordinate) and 'sports car' (for a typical subordinate category) are both longer than the basic level label. Basic level categories are also the most abstract level for which the category members share the same set of parts: cars all have wheels, brakes, and engines, whereas there are many other vehicles (e.g. helicopters and boats) that do not share these parts. Finally, children tend to learn basic level labels for objects before learning the labels for categories at other levels of abstraction. The factors that characterize basic level categories can be summarized as follows:

- Objects are identified first at the basic level.
- Objects are classified fastest at the basic level.
- The basic level is the most abstract level at which the members tend to have the same shape.
- The basic level is the most abstract level at which the members tend to share parts.
- The basic level is the most abstract level at which people interact with the members using similar motor movements.
- The basic level is the most abstract level at which the category members tend to be similar.
- The basic level is the most abstract level at which category members tend to be dissimilar from members of contrasting categories.
- Children often learn basic level labels before labels at other levels of abstraction.
- Basic level labels are shorter than labels for categories at other levels of abstraction.

The hierarchical organization of categories seems to be strongest for object categories. Some research has been done on the between-category structure of abstract concepts such as events and ideas. People also have categories at different levels of abstraction for these concepts, but the basic level does not have as much of an advantage relative to subordinate and superordinate categories.

Goal-derived Categories

The previous section suggested that an object might belong to many different categories, and that these categories generally differ in their level of abstraction. There are some categories, however, that are organized around people's goals rather than around the overall shape and parts of objects. Some of these categories are ones that we use all the time, and their labels have become words. For example, a 'pet' is a domesticated animal that is kept as a companion. Thus, membership in this category is determined by whether an object serves a particular goal.

An important observation is that people can also generate goal-derived categories as they are needed. For example, you may never have considered the category 'Things to take out of a house in the event of a fire'. Now that this category has been suggested, however, it is easy to generate members of the category (e.g. children, jewellery, photographs).

Novel goal-derived categories are called *ad hoc* categories. A striking finding is that, although they are being generated on the fly, they share many characteristics with categories that were previously learned. For example, like regular categories, *ad hoc* categories exhibit a graded typicality structure: that is, people find it easy to determine which members of an *ad hoc* category are typical or atypical. For example, people might agree that old photographs

of family members are good examples of things to take out of the house in the event of a fire, but that an old sofa is a poor example.

One key difference between regular categories and goal-derived categories is in the way that typicality is assessed. For regular categories, an object is typical to the extent that it is similar to the average member (or prototype) of the category. In contrast, for goal-derived categories there is often an ideal member, and items are more typical to the extent they are similar to the ideal. For example, someone might create the goal-derived category 'diet foods'. The ideal member of this category tastes great and has no energy content. A new object will be typical of a category to the extent that it is similar to its ideal.

CONCLUSION

Psychologists have explored the internal (within-category) representation and external (between-category) structure of category representations. Research on within-category representation has focused on the role of rules, similarity, and theory in determining category representation. Research on between-category structure has focused both on relationships among categories at different levels of abstraction and on goal-derived categories.

Further Reading

Barsalou LW (1983) Ad hoc categories. *Memory and Cognition* **11**: 211–227.

Keil FC (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Medin DL, Lynch EB and Solomon KO (2000) Are there kinds of concepts? *Annual Review of Psychology* **51**: 121–147.

Morris MW and Murphy GL (1990) Converging operations on a basic level in event taxonomies. *Memory and Cognition* **18**(4): 407–418.

Murphy GL and Medin DL (1985) The role of theories in conceptual coherence. *Psychological Review* **92**(3): 289–315.

Nosofsky RM (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* **115**(1): 39–57.

Nosofsky RM, Palmeri TJ and McKinley SC (1994) Rule-plus-exception model of classification learning. *Psychological Review* **101**(1): 53–97.

Rosch E and Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.

Rosch E, Mervis CB, Gray WD, Johnson DM and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology* **8**: 382–439.

Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Conditioning

Introductory article

Frances K McSweeney, Washington State University, Pullman, Washington, USA

CONTENTS

Classical and operant conditioning are procedures for changing behavior. Behavior changes when a previously neutral stimulus predicts an important stimulus (classical conditioning). Behavior also changes when a response produces a particular consequence (operant conditioning).

INTRODUCTION

Classical and operant conditioning provide powerful techniques for understanding and controlling behavior. In classical conditioning, the behavior towards a previously neutral stimulus changes when that stimulus predicts the occurrence of a stimulus that already evokes a response. In operant conditioning, the frequency of, or the time spent making, a response is changed by consequences (reinforcers or punishers) that follow that response.

CLASSICAL CONDITIONING

The discovery of classical conditioning is usually attributed to Ivan Pavlov, a Russian physiologist. Pavlov briefly turned on a metronome and then presented food to a dog. After several exposures to the metronome followed by food, the dog salivated when the metronome was presented alone. That is, the dog's behavior toward the metronome changed when the dog learned that the sound of the metronome was followed by food.

In Pavlov's procedure, the metronome or tone is usually referred to as the 'conditioned stimulus' (CS). The CS is originally neutral to the animal in the sense that it does not automatically evoke the response of interest (salivation). The food is referred to as the 'unconditioned stimulus' or US. It is a stimulus that already evokes a response. The response that is evoked by the US is called the 'unconditioned response' (UR). The response that occurs to the CS, as a result of its pairing with the US, is called the 'conditioned response' (CR).

The classical conditioning procedure presented above might be described by saying that when a CS is followed by a US several times, the CS comes to evoke a CR that resembles the UR. Even stated this way, classical conditioning is of some practical interest. For example, it may explain the acquisition of a fear or phobia (CR) when a stimulus (CS, e.g. a snake) precedes a frightening event (US: e.g. someone screams). It may play a part in the development of preferences for or aversions to food, and it may facilitate digestion because stimuli that precede food intake may help to prepare the body. Since the time of Pavlov, however, our understanding of classical conditioning has changed in ways that have greatly increased its practical usefulness.

The Form of the CR

Voluntary behaviors as CRs

Pavlov studied the reflexive response of salivation while his animals were immobilized by suspending them in a hammock. Reflexive responses are behaviors that are automatically evoked by a stimulus (e.g. you blink your eye in response to a threat). Reflexive responses are often distinguished from voluntary behaviors that are emitted rather than evoked by a particular external stimulus (e.g. you read this article). Pavlov's study of reflexive behaviors in restrained animals may have limited peoples interest in classical conditioning. Most people are more interested in the voluntary behavior of freely moving animals, including people.

The principle of 'sign tracking' suggests that classical conditioning does apply to the voluntary behavior of freely moving animals. Sign tracking states that animals approach and contact the stimulus that is the best predictor of a US, and withdraw from stimuli that signal the absence of a US. According to this idea, when a CS predicts a US, the behavior that is learned is movement (approach or withdrawal), not just reflexive behavior (e.g. salivation). For example, I once visited a wildlife park in Australia where a vending machine sold kangaroo chow. The machine made a loud noise

when it operated and that sound (CS) predicted the availability of food (US). As expected from sign tracking, a stampede of kangaroos approached (CR) the food machine as soon as it operated (CS), an undesirable event for those standing by the machine. (Technically, without further tests, this behavior cannot be conclusively identified as classically conditioned. It might also be a discriminated operant, as described later.)

CRs may differ from URs

In Pavlov's experiment, similar responses served as the CR and UR. That is, dogs salivated when food was presented (UR) and they learned to salivate to the sound of a metronome that preceded food (CR). If the CR must be identical to the UR, then classical conditioning cannot be used to train a response unless a US can be found that automatically evokes that response. This is difficult in many cases (e.g. teaching someone to play the piano).

We now know that the CR may differ from the UR. For example, when some types of drugs are used as a US, the CR may be the opposite of the UR. To give one example, morphine is a painkiller (UR), but animals become hypersensitive to pain (CR) in the presence of an arbitrary stimulus (CS), e.g. the sight of a needle, that predicts a morphine injection (US). This means that classical conditioning may contribute to the build-up of tolerance for drugs and to the withdrawal symptoms that are observed when drugs are not available. Think of the UR to morphine as a 'high' (a pleasant state) and the CR to morphine as a 'low' (an unpleasant state). Classically conditioned responses gradually become stronger with each successive pairing of the CS and US. If a conditioned 'low' becomes stronger with each morphine injection, then more and more of the drug will be needed to overcome it – that is, tolerance will develop. If the conditioned stimuli that accompany a drug injection (e.g. time of day, sight of the needle) occur without the drug, then the animal will experience only the low (CR) without the high (UR) produced by the US. This low may contribute to withdrawal symptoms. Although classical conditioning may play a role in drug tolerance and withdrawal, it is undoubtedly only one of many contributors.

Information versus Temporal Contiguity

In the earlier description, classical conditioning occurred when a CS was followed by a US. More recently, it has been argued that a CS must provide information about, or predict, the occurrence of the US for conditioned responding to occur. Consider the following experiment. One group of animals receives a CS followed by a US 10 times. A second group receives the same 10 CS–US pairings, but also receives 10 extra unconditioned stimuli that are interspersed among the CS–US pairings. If conditioned responding develops when a CS is followed by a US, then conditioned responding should be strong in both groups. They both experience the CS followed by the US 10 times. If conditioned responding occurs when the CS provides information about the US, then the first group should respond more than the second group. In the first group, the CS is a perfect predictor of the US; in the second group, it is not. The evidence supports the information view.

Cue Competition Effects

The strength of conditioned responding to a CS depends on how well that CS predicts the US. It also depends on the strength of conditioning to other conditioned stimuli that also predict the US. Because the presence of other predictive conditioned stimuli usually weakens conditioned responding to any one of them, cues are said to compete for conditioning. Overshadowing and blocking are examples of cue competition.

Overshadowing

In most cases, more conditioned responding occurs to CS1 (e.g. a light) when CS1 alone predicts the US (CS1 $\rightarrow$ US) than when CS1 and CS2 (e.g. a light plus a tone) together predict the US (CS1 + CS2 $\rightarrow$ US). The stimulus CS2 is said to 'overshadow', and therefore to reduce conditioned responding to, CS1. Although overshadowing is the usual finding, the opposite, potentiation, is occasionally reported, especially when the stimuli are tastes and odors. Potentiation refers to greater conditioned responding to CS1 when it predicts the US in compound with CS2 than when CS1 predicts the US alone.

Blocking

If an animal learns that CS1 predicts a US (CS1 $\rightarrow$ US), then the animal will not later perform a CR to a second CS (CS2) that also predicts the US when presented in compound with CS1 (CS1 + CS2 $\rightarrow$ US). The prior conditioning to CS1 is said to block conditioned responding to CS2. The occurrence of blocking provides further support for the idea that conditioned responding develops only when a CS provides information about a US. Presumably, a CR does not occur to CS2 because it provides

no information about the US beyond that already provided by CS1.

Again, blocking is the usual finding, but the opposite result of augmentation is sometimes reported. Augmentation refers to a situation in which prior conditioning to CS1 strengthens, rather than weakens, conditioned responding to a redundant CS2. Augmentation, as potentiation, is particularly likely when tastes and odors serve as CSs.

Cue to Consequence Effect

Some combinations of CSs and USs yield stronger CRs than others. For example, a strong CR of aversion will develop when a taste (CS) predicts illness (US), as well as when a light (CS) predicts an electric shock (US). Weaker or no aversion will result when a light (CS) predicts illness (US) and when a taste (CS) predicts an electric shock (US). Although it is not known why this occurs, some have argued that evolution favors animals that quickly learn to avoid foods whose tastes predict illness. Such animals are less likely to die of poisoning.

Significance of Classical Conditioning

Classical conditioning was once described in ways that made it appear to be a simple mechanical transfer of behavior from one stimulus to another. This view was challenged by many findings, including that the CR need not resemble the UR and that blocking may occur. More recently, classical conditioning has been seen as a mechanism through which the predictive relations among stimuli in the environment alter behavior. In many cases, receiving an early warning from an arbitrary stimulus may help an animal deal with a potentially harmful stimulus (e.g. a drug) to come.

OPERANT CONDITIONING

Operant conditioning is a change in behavior that occurs as a result of the consequences of that behavior. Its most prominent student was B. F. Skinner. Because of the power of operant techniques, they form the basis for behavior therapies that are effective in correcting many human behavioral problems. They are used to train nonhuman animals for performances in films or circuses. They are also often used in scientific studies to establish a baseline for assessing the effect of other manipulations (e.g. drug injections, physiological interventions). Operant techniques are useful as baselines because they provide stable and replicable control over behavior.

Positive Reinforcement

The most frequently studied form of operant conditioning is positive reinforcement. According to the principle of positive reinforcement, a response that is followed by a reinforcer will increase in frequency. For example, if lever pressing (response) yields food (reinforcer) for a hungry rat, the rat will press the lever more frequently in the future. Some behaviors are more easily measured in terms of the time devoted to them than in terms of their frequency (e.g. reading). In that case, following the response by a reinforcer will increase the amount of time spent making the response. For example, if practicing the piano (response) yields the opportunity to watch a favorite television program (reinforcer), then the time spent practicing will increase.

Notice that a response cannot be strengthened by reinforcement unless a reinforcer can be found. Over the years, many definitions for the term 'reinforcer' have been rejected. For example, reinforcers were once thought of as substances that are physiologically needed (e.g. food, water), but there are many reinforcers that are not physiologically needed (e.g. watching television, going to the cinema). Reinforcers were once defined as stimuli that reduce tension (e.g. sexual behavior), but many stimuli that increase tension also serve as reinforcers (e.g. watching a scary film, riding a roller coaster).

Because of these failures, a reinforcer is technically defined as any stimulus that increases the frequency of a response that it follows. This is an undesirable definition because it makes the principle of positive reinforcement circular. That is, the principle now states that a response followed by any stimulus that increases the frequency of the response that it follows will increase in frequency. This definition has been accepted because scientists can identify a stimulus as a reinforcer in one situation (e.g. they can show that the stimulus will increase the frequency of one response that it follows); they can then test the principle of positive reinforcement in another situation (e.g. they can ask whether that reinforcer also increases the frequency of other responses that it follows).

Some stimuli will serve as reinforcers for both human and nonhuman animals (e.g. food, water, access to conspecifics for herd animals). Other reinforcers will be more effective with people than with other animals (e.g. praise, money, the opportunity to watch television). Different items will serve as reinforcers for different people, and the

Premack principle provides a way to identify effective reinforcers. The Premack principle states that the opportunity to perform any higher probability response can serve as a reinforcer for any lower probability response. The probability of a response is measured by examining what the animal does when it has free time. According to the Premack principle, if a child often plays electronic games, then the opportunity to play such a game will serve as a reinforcer for less probable behaviors, such as doing homework.

Eliciting the first response

A response cannot be reinforced until that response occurs. When working with people, verbal instructions may be used to elicit the first response (e.g. 'please practice the piano'). For complex responses, such as a golf swing, physical guidance known as 'prompting' may also be needed. Shaping by successive approximations may be useful when dealing with nonhuman animals or with a nonverbal person such as a baby. During shaping, closer and closer approximations to the desired response are reinforced. For example, if you want to teach your dog to sit up, you could begin by following any movement by a reinforcer. Then you might reinforce only movements that involve some transfer of the dog's weight to its back paws. Then you might reinforce only movements that involve weight transfer to the back paws plus lifting the forepaws off the ground. By carefully choosing which behaviors to reinforce and when to alter the reinforced response, you should quickly have your dog sitting up.

The Four Basic Conditioning Procedures

Operant conditioning can be used to either increase (reinforcement) or decrease (punishment) the frequency of a response. The frequency of a response may change when the response produces something (positive) or when it escapes or avoids something (negative). It is called 'positive reinforcement' when a response increases in frequency because it produced something (e.g. you work because you have been paid for working). Negative reinforcement occurs when a response increases because it escaped or avoided something (e.g. you drive safely because safe driving has prevented accidents). Positive punishment occurs when a response decreases in frequency because it produced something (e.g. your cat no longer scratches the furniture because you squirted it with water whenever it scratched). Negative punishment occurs when a response decreases in frequency because it prevented or removed something (e.g. your child stops hitting his little brother because you took away his allowance or confined him to his room when he did so).

Technically, a stimulus is classified as a reinforcer or punisher depending on its effect on behavior. Reinforcers increase behaviors that they follow; punishers decrease behaviors. Terms related to pleasure and pain do not appear in these definitions. Nevertheless, stimuli that serve as positive reinforcers when they are delivered and as negative punishers when they are removed are usually described as pleasant. Stimuli that serve as negative reinforcers when they are removed and positive punishers when they are presented are usually described as aversive or painful.

Schedules of Reinforcement

Reinforcers are delivered according to schedules of reinforcement, which are rules specifying which response will be followed by a reinforcer. During a continuous reinforcement (CRF) procedure, every occurrence of a particular response is followed by a reinforcer. For example, in a perfect world, depositing an appropriate amount of money into a soft-drink machine (response) would yield a soft drink (reinforcer) each time that you did it. Continuous reinforcement procedures have drawbacks that reduce their usefulness. For example, the person administering a CRF procedure must be alert enough to identify and to reinforce every instance of the behavior when it occurs. Therefore, CRF is often used to initially teach a response but it is usually replaced by partial reinforcement as the response becomes stronger.

During partial reinforcement (PRF), some instances of a response do not yield the reinforcer. There are four basic schedules of partial reinforcement. In a fixed ratio schedule (FR x), a reinforcer is delivered after every x occurrences of a response. For example, in a piecework factory, you might be paid (reinforcer) every time you completed 10 widgets (10 responses). This would be an FR 10 schedule. In a variable ratio schedule (VR x), a reinforcer is delivered after every xth occurrence of the response on average. For example, a foraging pigeon probably does not find food (reinforcer) each time it pecks the ground (response), but it does find food after some variable number of pecks.

In a fixed interval (FI x min) schedule, a reinforcer follows the first response emitted after a fixed period of time since the last reinforcer. For

example, if the post is delivered at approximately the same time every day, then the response of checking your mailbox will be followed by the reinforcer of finding mail approximately 24 h after the last time that you received mail. In a variable interval (VI *x* min) schedule, a reinforcer follows the first response emitted after a variable period of time since the last reinforcer. For example, the response of getting through when dialing a telephone number that is busy is probably reinforced on a VI schedule. Notice that that response must be emitted or the reinforcer will not be delivered during FI and VI schedules. Notice also that only one response is required to produce the reinforcer. In spite of this, animals emit many more than one response per reinforcer when responding on these schedules.

Psychologists distinguish between these four basic schedules because the schedules control behavior in different ways. Animals respond at a relatively high and steady rate when responding on VR and VI schedules. In contrast, animals pause after reinforcement when responding on FR and FI schedules. When responding begins, it either continues at a steady rate after the pause (FR schedules) or gradually accelerates to a peak rate just before the next reinforcer is delivered (FI schedule). The length of the pause following reinforcement is directly proportional to the size of the schedule requirement. For example, animals may pause for approximately 30 s when responding on an FI 1 min schedule that makes reinforcers available once per min. They may pause for approximately 5 min when responding on an FI 10 min schedule that makes reinforcers available only once every 10 min. If the requirement becomes too large, the animal may stop responding entirely. This is called 'ratio strain'. Ratio strain can be reversed by returning to a less difficult requirement for reinforcement.

CHARACTERISTICS OF CONDITIONED BEHAVIOR

Acquisition

Conditioned responses gradually gain strength as the reinforcer repeatedly follows the response or the CS repeatedly predicts the US. For example, the amount of saliva that Pavlov collected when he presented the CS (strength of the CR) would be greater after ten pairings of the CS with the US than it was after only two pairings. However, conditioned responses do not gain strength indefinitely. Eventually, a point is reached beyond which further pairings of the CS with the US or the response with the reinforcer do not further increase the strength of the conditioned behavior.

Extinction

The extinction of a classically conditioned response refers to the return of a CR to its baseline strength after the relation between the CS and US is broken. Baseline strength is the strength of the CR before conditioning. Extinction may be accomplished in either of two ways. The US may be removed entirely, or the CS and US may be presented randomly with respect to each other. In the earlier example, kangaroos approached (CR) a feeder because the sound of the feeder (CS) predicted that food (US) was available. Kangaroos would stop approaching the feeder (extinction) if either the feeder was empty so that the sound occurred with no food (US removal) or if the feeder was broken so that the sound occurred randomly with respect to the availability of food (random CS and US presentation).

The extinction of an operantly conditioned response refers to the return of that response to its baseline strength when the relation between the response and the reinforcer is broken. Again, this relation may be broken by removing the reinforcer or by presenting the reinforcer randomly with respect to the response. For example, if you work (response) because you have been paid for working (reinforcer), you would work less (extinction) if you were no longer paid (the reinforcer was removed) or if you won the lottery (received money regardless of whether you worked or not).

Generalization

Generalization of classical conditioning refers to the fact that a CR to one particular CS also occurs in response to other stimuli that resemble that CS. The greater the resemblance between the new stimulus and the CS that was actually paired with the US, the stronger the CR to the new stimulus. For example, if you are stung (US) by a bee (CS), you may learn to fear (CR) other insects, and your fear will be stronger the more closely the insect resembles a bee.

In operant conditioning, a response that has been reinforced in the presence of one stimulus will also occur in the presence of other stimuli that resemble the original stimulus. Again, the response to a new stimulus will be stronger, the more that stimulus resembles the stimulus in the presence of which the response was reinforced. For example, teachers do not have to teach appropriate classroom behavior

at the beginning of each school term. The new classroom resembles the students' classrooms of the previous year. Therefore, behaviors (e.g. sitting quietly) that were reinforced in the old classroom occur in the new one without training.

Discrimination

During a classical conditioning discrimination procedure, one stimulus (CS+) predicts the US and another stimulus (CS−) does not. The CR occurs to CS+ but not to CS−. In our kangaroo example, the kangaroos approach (CR) the sound of the food magazine (CS) because it predicts food (US). They do not approach the sound of the gate opening (CS−) because it does not predict food (US).

During an operant discrimination procedure, a response is reinforced in the presence of one stimulus (S+) and not in the presence of another (S−). The response occurs in the presence of S+, but not in the presence of S−. For example, if a child learns that whining (response) gets him anything that he wants (reinforcer) when his father is around (S+), but not when his mother is around (S−), the child will whine only when his father is present.

Discrimination procedures provide useful techniques for asking questions of nonhuman animals or nonverbal people (e.g. infants). You may have heard that dogs do not see colors and wondered how we know. Part of the answer comes from discrimination training. Suppose you reinforce sitting up by giving the dog a treat in the presence of anything red, but not in the presence of anything green. If the dog can see colors, then you will quickly have a dog that sits up when a red, but not a green, stimulus is presented. This experiment must be done carefully. For example, if red and green stimuli reflect different amounts of light, then the dog will solve the discrimination on the basis of the brightness of the stimuli rather than on the basis of their color. However, when this experiment is done properly, dogs have difficulty forming a discrimination between red and green. As a result of many discrimination experiments, a great deal is known about the sensory and conceptual worlds of infants and many species of nonhuman animals.

Higher-order Conditioning

Some stimuli innately serve as unconditioned stimuli or reinforcers without additional training. These stimuli are called 'primary reinforcers' (or primary US). They include biologically important stimuli, such as food and water. Other stimuli acquire their ability to act as reinforcers through experience. These stimuli are called 'secondary', 'conditioned' or 'higher-order' reinforcers (or secondary US). Money provides the most obvious example of a secondary reinforcer.

Stimuli acquire the ability to act as secondary reinforcers in many ways. Two examples illustrate this. First, secondary reinforcers called 'tokens' are stimuli that can be exchanged for primary reinforcers. For example, money acquires the ability to act as a reinforcer because it can be exchanged for food, drink and other primary reinforcers. Second, classical conditioning pairing of a stimulus with a primary US or reinforcer will produce a secondary reinforcer or US. Therefore, a bell that is used to summon animals for feeding will gain the ability to act as a reinforcer itself.

The ability of stimuli to act as secondary reinforcers will be extinguished if their relation to the primary reinforcer or US is broken. Therefore, money would gradually lose its ability to reinforce if it was no longer exchangeable for goods, and the bell would lose its ability to reinforce if it was presented often without food.

Further Reading

Domjan M (1998) *The Principles of Learning and Behavior*, 4th edn. Pacific Grove, CA: Brooks/Cole Publishing.

Hearst E and Jenkins HM (1974) *Sign-tracking: The Stimulus-reinforcer Relation and Directed Action*. Austin, TX: Psychonomic Society.

Honig WK and Staddon JER (1977) *Handbook of Operant Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Mackintosh NJ (1974) *The Psychology of Animal Learning*. New York, NY: Academic Press.

Pavlov IP (1927) *Conditioned Reflexes*, translated by GV Anrep. London, UK: Oxford University Press.

Pear JJ (2001) *The Science of Learning*. Philadelphia, PA: Psychology Press.

Pearce JM (1997) *Animal Learning and Cognition*. Hove, UK: Psychology Press.

Pierce WD and Epling WF (1999) *Behavior Analysis and Learning*. Upper Saddle River, NJ: Prentice-Hall.

Rescorla RA (1988) Pavlovian conditioning: it's not what you think it is. *American Psychologist* **43**: 151–160.

Skinner BF (1938) *The Behavior of Organisms*. New York, NY: Appleton-Century-Crofts.

Constructivism

Intermediate article

Jacqueline Grennon Brooks, State University of New York, Stony Brook, New York, USA

CONTENTS

Constructivism is a learning theory based on the notion that learners generate meaning through iterative mental formulation and reformulation of theories that satisfy the search for understanding.

INTRODUCTION

Constructivism has roots in various research traditions. It asks the psychological question: how is learning self-regulated? It considers the epistemological query: what is knowledge? It poses the pedagogical problem: how can educators facilitate knowledge construction? It examines the philosophical dilemma: is there an objective truth that we struggle to know, or are their different truths dependent on perception? Amid an array of widely varied responses to these questions is the cohesive focus on the learner's active role in generating meaning.

ACTIVE PROCESSES OF MEANING MAKING

The cognitive processes involved in making meaning are active ones that require the learner to continually evaluate new information and experiences against the learner's current theories, rules or notions. This viewpoint stands in stark contrast to other assertions that the learner's mind is a clean slate ready for inscription through direct teaching. Constructivism states that the learner approaches new experiences with a set of pre-established beliefs and naive theories, and that the learner changes those beliefs and theories only when unable to reconcile new data with previously held conceptions. Often learners will dismiss data that do not fit their present thinking as irrelevant or unrelated, and will assimilate the new information into their previously established theories without any disequilibrium or cognitive conflict. When learners find new data compelling enough to reconsider old theories, however, they experience cognitive conflict. They conquer their disequilibrium by accommodating current theories to include adequate explanations of past and present data (Piaget, 1953, 1970a). Educational programs based on constructivist principles are premised on the idea that it is the learner's responsibility to construct meaning through reflection on experiences with objects, phenomena or people, and that it is the teacher's responsibility to scaffold learner reflection in a manner that may generate learner analysis, synthesis and insight. (*See* **Naive Theories, Development of**)

The Nature of Knowledge

When a learner can construct interrelationships among sets of factual information and apply those understandings of interrelationships in novel contexts, the learner has constructed knowledge. This knowledge is dependent on mental structures. If a learner does not have the precursor mental structures in place, for example, to understand proportions, no amount of repeating the statement 'density is a ratio of mass to volume' will help the learner establish the concept. The learner's search for relationships among variables is the mental activity critical for an understanding of density as a ratio and the comparison of densities as a proportion. The depth to which a learner can construct understandings is predicated not only on the precursor mental structures mentioned previously, but on the information to which the learner has access. For instance, with information about atomic structure, density can be further understood in terms of the degree and nature of the packing of atoms in crystalline form. Within the iterative process of cognitive growth, the learner can create new knowledge from an ever-expanding repertoire of information that gives rise to yet new mental structures and then possibly new mental stages. (*See* **Piagetian Theory, Development of Conceptual Structure**)

Mental Structures

One way of describing constructivism is through an analysis of mental structures.

> We may say that a structure is a system of transformations. Inasmuch as it is a system and not a mere collection of elements and their properties, these transformations involve laws. … In short, the notion of structure is comprised of three key ideas: the idea of wholeness, the idea of transformation, and the idea of self-regulation. (Piaget, 1970b: p. 5)

Wholeness refers to the learner's quest to map the relationships among parts of a set. For instance, knowing the meaning of each word in a sentence is different from knowing the meaning of the entire sentence. It is the learner's ability to derive meaning from the relationships among the words in a sentence that forms the basis of the learner's understanding of the sentence. Transformation refers to the notion that as the learner constructs deeper understandings of the world, a mathematical necessity for grouping those understandings emerges. Thus, fundamental transformations of undergirding logical structures gives rise to new, more inclusive structures. Piaget's own career history provides an example. As a biologist studying the appearance of new structures in plants, Piaget transformed his understanding of how the hereditary aspects of the plant are affected by conditions in the environment into a more inclusive relationship between previous structures and environmental conditions that he could apply in a variety of settings. In the immediate case, he applied this relationship in the realm of psychology. Self-regulation refers to the learner's ability to engage in assimilation and accommodation in order to either maintain cognitive equilibrium or resolve cognitive conflict and reestablish cognitive equilibrium.

Personal knowledge construction is a key element of the constructivist paradigm. The eighteenth-century philosopher Vico was one of the earliest writers to put forth the notion that human beings can only know what their cognitive structures allow them to know. This notion surfaces for teachers in terms of readiness for learning. How far a learner can progress is a function of what the learner currently understands, and what the learner currently understands is a function of the learner's existing mental structures. The teacher plays a part in facilitating the learner's development of new knowledge and new structures through the creation of settings in which the learner may detect discrepancies and in which the teacher fosters discrepancy resolution. This iterative process gives rise to ensuingly more rigorous theory building.

Cognitive Conflict

Logical structures and processes of transformation, wholeness and self-regulation are terms from the philosophical and psychological literature. Constructivism, when described by educators, is associated with another set of correlated terms. Ausubel (1963) explains learning through a process he calls 'subsumption', in which new, more specific knowledge is linked with previous, more inclusive knowledge. His term 'obliterative subsumption', describes a process highly related to the law of transformation set out by Piaget in which concepts are modified so dramatically over time that learners can lose access to some of the specifics of their own earlier thinking. These more robust concepts characterize the superordinate learning that is typically called 'meaningful' learning.

Prior Knowledge

What a learner already knows and how the teacher scaffolds the current learning environment are important determiners of the future understandings the learner will be able to construct. When a learner is investigating a new phenomenon, whether it be how shadows are cast, how an odd number of supplies can be shared among an even number of students, or how one locates an entry in an encyclopedia, the constructivist teacher negotiates the phenomenon or concept with the learner using a subtle yet observable set of practices. The teacher seeks to understand the learner's readiness to generate certain types of knowledge by determining the learner's present hunches, conceptions, beliefs, etc. The teacher then provides opportunities for the learner to confirm or refute those initial thoughts. Unless the confirmations or refutations come from the learner, the likelihood that the learner will generate understanding is compromised. Unveiling the prior knowledge of the learner is an important aspect of constructivist teaching because constructivist theory stresses the importance of the learner's transforming current mental structures. It is the constant replacement of understandings with richer and deeper ones that characterizes the processes of cognitive growth and learning.

ROLE OF SOCIAL INTERACTION IN MEANING MAKING

The theory of constructivism postulates that learners come to know their world by interacting

with it in ways that allow them to build the mental structures that 'explain' what is perceived. Ultimately, this process of meaning making is individually constructed, but it is an outgrowth of social interaction. The social interaction may be immediate, as in the case of discourse with others in a community, or may be contextual, as in the case of learners formulating ideas within historical and cultural moments. Some theorists view knowledge as the mechanism by which learners make sense of their experiences in their environment. Others view it as the outcome of learners making sense of their experiences in their environment; and yet others see knowledge as both a means and an end. Within these divergent views of knowledge, there is much debate as to the nature of the relative components of knowledge, and also over the merit of classifying particular knowledge as true or false (Philips, 1998). However, there is general agreement among constructivists that knowledge is socially constructed and a function of the culturally derived, community-sanctioned perspective of the knower. The goal of education is to foster the development of shared knowledge among community members. Peers play a powerful role in this shared knowledge construction.

Community of Learners

The theory of constructivism gives rise to a 'community of learners' model within educational settings, a model in which learners in search of understanding communicate current thinking with others by formulating and reformulating their thoughts based on peer and expert feedback and by reflecting on that feedback. This model presupposes a definition of knowledge as dynamic and socially constructed and rejects a definition of knowledge as static and 'passed on'. This model also requires a good deal of scaffolding on the part of the teacher to maximize the likelihood that meaningful learning will occur. Some level of dissonance must be established either through peers' sharing diverse perspectives or through teacher prompts to highlight sources of cognitive dissatisfaction. The task must be within the reach of the learners, more advanced than any individual within the group would be likely to complete independently, but not so far advanced that learning shuts down.

Zone of Proximal Development

What is the teacher's role in the community of learners model? As the 'expert', the teacher can guide the learner in intellectual arenas in which the learner could not independently navigate. Vygotsky (1962) referred to this arena as the zone of proximal development. The teacher provides an intellectual framework at the leading edge of the learner's current thinking on a topic. That framework can include questioning designed to help the learner see relationships, can include contradictions designed to help the learner examine subtleties, or can include hypothetical comments to help the learner extend an argument. Although the nature of the scaffolding may be diverse, the scaffolding has a unified purpose: to aid learners in restructuring their current theories. For Vygotsky, language is the basis of cognition. For Piaget, language is a mechanism to express cognition.

Teachers offer learners this guided participation to maximize the likelihood that restructuring will occur. To every learner, his or her present thinking holds a great deal of merit. Therefore, cognitive restructuring is resistant to casual interference. Furthermore, it is even resistant to direct instruction. Much research discusses the inability of direct instruction to dispel the misconceptions widespread over many topics to which learners cling (Clement, 1982). Fostering a learner's restructuring of present conceptions requires an analysis of the learner's current perspective with specific regard to the topic, concept or issue at hand. While a casual observer in such a classroom may not readily see the underlying pedagogy, the pedagogy none the less exists and is powerful. The teacher's pedagogy drives decisions concerning which responses to pursue, which student groupings to establish, which supplies to gather, and which follow-up questions to generate.

CONCEPTUAL CHANGE

The term 'constructivism' holds different meanings in many circles. The radical social constructivists discuss the illusoriness of objective truth (von Glasersfeld, 1998), the cognitive constructivists engage in structural analyses of knowledge generation (Piaget, 1953), and the human constructivists seek a synthesis of epistemological and psychological phenomena (Novak, 1993). Where there is significant intragroup and intergroup variation, a binding construct for all groups is the focus on conceptual change. The goals to which the learner aspires may differ, but the constructivist teacher and researcher are focused on better understanding the learner's conceptual changes over time, the nature of the changes and the contributing variables. (*See* **Conceptual Change**)

References

Ausubel DB (1963) *The Psychology of Meaningful Verbal Learning*. New York, NY: Grune & Stratton.

Clement J (1982) Algebra word problem solutions: thought processes underlying a common misconception. *Journal for Research in Mathematics Education* **13**(1): 16–30.

Novak J (1993) Human constructivism: a unification of the psychological and epistemological phenomena in meaning making. *International Journal of Personal Construct Psychology* **6**: 167–193.

Philips DC (1998) Coming to terms with radical social constructivism. In: Matthews MR (ed.) *Constructivism in Science Education*, pp. 139–158. London, UK: Kluwer.

Piaget J (1953) *Logic and Psychology*. Manchester, UK: Manchester University Press.

Piaget J (1970a) *Genetic Epistemology*. New York, NY: Columbia University Press.

Piaget J (1970b) *Structuralism*. New York, NY: Basic Books.

Von Glasersfeld E (1998) Cognition, construction of knowledge and teaching. In: Matthews MR (ed.) *Constructivism in Science Education*, pp. 11–30. London, UK: Kluwer.

Vygotsky L (1962) *Thought and Language*. Cambridge, MA: MIT Press.

Further Reading

Brooks JG and Brooks MG (1993) *In Search of Understanding: The Case for Constructivist Classrooms*. Alexandria, VA: Association for Supervision and Curriculum Development.

Copple C, Sigel L and Saunders R (1984) *Educating the Young Thinker*. New York, NY: Van Nostrand.

Davis RB, Maher CA and Nodding N (1990) Constructivist views on the teaching and learning of mathematics. *Journal for Research in Mathematics Education*, Monograph No. 4. Reston, VA: National Council of Teachers of Mathematics.

Driver R, Guesne E and Tiberghien A (eds) (1985) *Children's Ideas in Science*. Philadelphia, PA: Open University Press.

Duckworth E (1987) *'The Having of Wonderful Ideas' and Other Essays on Teaching and Learning*. New York, NY: Teachers College Press.

Fosnot CT (ed.) (1996) *Constructivism: Theory, Perspectives, and Practice*. New York, NY: Teachers College Press.

Piaget J and Inhelder B (1971) *Psychology of the Child*. New York, NY: Basic Books.

Sigel IE, Brodzinsky DM and Golinkoff RM (eds) (1981) *New Directions in Piagetian Theory and Practice*. Hillsdale, NJ: Lawrence Erlbaum.

Von Glasersfeld E (1995) A constructivist approach to teaching. In: Steffe L and Gale J (ed.) *Constructivism in Education*, pp. 3–16. Hillsdale, NJ: Lawrence Erlbaum.

Vygotsky LS (1962) *Thought and Language*. Cambridge, MA: MIT Press.

Counterfactual Thinking

Intermediate article

Neal J Roese, University of Illinois, Champaign, Illinois, USA
James M Olson, University of Western Ontario, Ontario, Canada

CONTENTS

Mental constructions of alternatives to facts or events. These thoughts of 'what might have been' are linked to a variety of emotional and judgmental consequences.

INTRODUCTION

The term 'counterfactual' means contrary to established facts or actual events. Counterfactual thinking typically involves imaginative speculation about alternatives to past outcomes: that is, about what might have been. Counterfactuals often (though not always) take the form of conditional propositions, containing the dual components of antecedent and consequent. In everyday cognition, counterfactual thinking usually targets personal goals and desires, such that individuals focus on actions that might have brought about particular desired ends (e.g. 'If I had studied harder, I would have earned a higher grade'). Counterfactuals can also be deployed in everyday speech as arguments ('If not for Gorbachev, the Soviet Union and the Cold War would have persisted into the twenty-first century') or invitations to further speculation and elaboration, e.g., 'What if President Kennedy hadn't been assassinated?' Counterfactuals have intrigued philosophers throughout the twentieth century because of their implications for logic and epistemology, but more recently counterfactual thinking has inspired psychological research because such thought processes influence a wide range of emotional, judgmental, and behavioral outcomes.

The form and content of counterfactuals is limitless, and although they may conjure the bizarre and the fantastic, everyday counterfactual thinking is mundane. Indeed, an essential feature seems to be that counterfactuals preserve the integrity of the world as we know it, altering but one or two specific features, then unfurling immediate consequences against a backdrop that is essentially the same as actuality. Thus, one might wonder how the Second World War might have unfolded had Hitler attacked and defeated the British at Dunkirk rather than allowing them to escape, but background features, such as the previous history of Europe, the power of the respective nations' armaments, and for that matter the laws of physics, remain unchanged. Given this rule of restricted alteration, a key theoretical focus has been to specify which finite features of reality are perceived to be more changeable, or mutable, as opposed to the infinite background features that remain constant within one's mind. The sections below on determinants of counterfactual thought are descriptions of these patterns.

Counterfactual thinking is a rule-bound creative act, and as such has been construed as a principal ingredient of consciousness and language. Hofstadter (1985), for example, argued that a comprehensive attempt to create artificial intelligence must include some facility for production of counterfactuals that operates in a manner similar to that of human cognition. A further elaboration of this theme is that counterfactuals are constrained by reality because they are functional; that is, they often provide useful prescriptions for how a goal might have been achieved in the past, and hence how it might yet be achieved in the future (Roese, 1994).

TYPOLOGY

Counterfactuals have been classified in two main ways: direction and structure. Direction refers to whether the counterfactual specifies a state that is better than actuality (an upward counterfactual) or worse than actuality (a downward counterfactual). Counterfactuals are also described in terms of structure of their phrasing. The counterfactual antecedent may be an addition of some feature not in fact present (an additive counterfactual), or

it may remove a feature that was present (a subtractive counterfactual). These two typologies have proved effective in delineating a variety of theoretical relations, described below.

PSYCHOLOGICAL CONSEQUENCES

Causation

Counterfactuals are intimately related to causal inferences. Causation may be defined as a relation between two variables (objects, states, etc.) in which one produces or generates changes in the other. A counterfactual conditional nearly always implies causation. Counterfactual conditionals denote an antecedent-consequent pair that diverges from a related, factual antecedent-consequent pair, thereby satisfying the logic of J. S. Mill's method of difference for inferring causation. For example, the observation that a match held motionless remains bereft of flame might be followed by the counterfactual supposition that 'if the match had struck a hard surface, it would have ignited.' The mental alteration of but one feature of actuality (striking as opposed to not striking the match), when accompanied by the imagined consequential variation in ignition, provides the basis for inferring that the antecedent of match strike causally influences ignition. The logic of the method of difference is the same as the covariation criterion for causation that forms the theoretical platform for many theories of causal attribution, in that counterfactuals present one datum, albeit imagined, that may be added to a set of divergent background observations. Although absence of covariation may be used to rule out causation, presence of covariation is not in itself sufficient to infer causation. Therefore, the same problems of induction that bedevil formal analyses of causation apply similarly to counterfactual reasoning (Spellman and Mandel, 1999). Nearly all psychological consequences of counterfactual thinking appear to be rooted either in this causal inference mechanism or in a contrast effect mechanism.

Contrast Effects

In comparative judgment, the juxtaposition of one object with a second can render judgments of the features of the latter more extreme. Thus, as demonstrated in classical psychophysics experiments, an object may be judged to be heavier after holding a lighter object, a color may be deemed darker if set against a lighter background, and so on. Counterfactual comparisons may similarly influence emotional appraisals of specific outcomes by making them, in contrast, seem better or worse. Thus, upward counterfactuals make an actual event seem less favorable, whereas downward counterfactuals make an actual event seem more favorable. This contrast effect underlies a variety of effects of counterfactual thinking on social judgment.

Social Judgment

A variety of social judgmental consequences of counterfactual thinking have been mapped; five are detailed here.

First, counterfactuals influence emotion, typically making emotional reactions more extreme (by way of a contrast effect) than would otherwise have been the case. Regret is an affective state predicated on upward counterfactual thinking and is the subject of much research in its own right. Counterfactual-induced affective changes can then influence judgment further. For example, in responses to victimization, inferring that a victim's misfortune could easily have been averted might create greater sympathy for the victim, but also greater recommendations for monetary compensation to the victim (Miller and McFarland, 1986).

Second, counterfactuals influence likelihood estimates in at least two ways, both rooted in the causal inference mechanism. The mental simulation of an alternative antecedent event can make future, similar events seem more likely. This would occur to the extent that the prior action is controllable and presumed to be sufficient to have brought about a favorable outcome. That is, the individual might intend to perform the action in the future to bring about a desired goal, in part because the individual infers that performing it in the past would have brought about that desired goal in the past (Roese, 1994). Counterfactuals can also make past events seem more predictable (the hindsight bias) to the extent that they clarify causal linkages, i.e. specify how an event was brought about and thus how it might have been improved or negated (Roese and Olson, 1996). For example, a student who reacts to a poor grade with the counterfactual, 'If only I had studied harder, I would have performed better', has used the counterfactual to articulate the causal power of studying to influence performance. This causal inference may then form the basis of a behavioral intention to study more thoroughly for the next examination, which then yields beliefs in the heightened probability of future success.

Third, and drawing directly on the previous description of heightened likelihood estimates,

counterfactuals can heighten perceived control, again by way of causal inferences. To the extent that a desired event is seen to be attainable had one only acted in a certain way, it confers a belief in personal control (Nasco and Marsh, 1999). In other words, one may generalize from the specific instance of having been able to effect positive outcomes ('If I had studied harder, I would have performed better') to the beliefs regarding global personal efficacy ('I can accomplish many things with a little extra effort').

Fourth, counterfactuals can influence decision-making. If a decision is made but an alternative decision might have brought about clearly better rewards, the resulting emotion of regret may compel changes in decision-making strategy, altering the course of subsequent behavior. Research on cognitive dissonance theory specifies conditions under which individuals alter appraisals as a function of postdecisional regret, but theory linking dissonance to counterfactuals is underdeveloped.

Fifth, counterfactuals can make observers suspicious. If an event occurs but is surprising because it is easy to imagine it occurring differently, an observer might be more suspicious regarding ulterior goals of the actor than in cases in which it is easy to imagine the event occurring in many similar ways, even if the probability of event occurrence remains constant. Take the example of a child who loves chocolate-chip cookies: the child is permitted to have just one cookie before dinner, but is required to select the cookie with eyes closed from a jar containing one chocolate-chip cookie and nine oatmeal cookies. If the child happens to select the coveted chocolate-chip cookie, an observer might suspect that the child had peeked. If, however, the cookie jar contained ten chocolate-chip cookies and ninety oatmeal cookies, suspicion might be reduced as there are ten similar ways for the coveted cookie to be selected without intent. Even though the probability of selecting the chocolate chip cookie is identical in both cases, ease of generation of alternatives differs and results in variation in suspicion (Miller *et al.*, 1989).

DETERMINANTS OF COUNTERFACTUAL THINKING

Activation

When does counterfactual thinking occur? A principal trigger is negative affect resulting from an undesirable outcome (Sanna and Turley, 1996). When things go wrong, people often ruminate about how the outcome could have been avoided. Thus, thoughts about 'what might have been' are more common following defeats than victories, failures than successes, and penalties than rewards. A second activator of counterfactual thinking is surprise resulting from an unexpected event. Unexpected occurrences violate implicit predictions and thereby attract attention, which induces consideration of why the outcome occurred. A third trigger of counterfactual thinking is a near miss, or an event that almost occurred. When something nearly happens, it seizes the perceiver's imagination. An athlete who finishes second by a hair's breadth in a 100 m race is likely to experience vivid thoughts about the counterfactual outcome of winning, whereas finishing a distant second evokes fewer thoughts of hypothetical victory.

These triggers of counterfactual thinking correspond to situations where this activity is most useful. As noted earlier, counterfactual thinking provides causal information about an outcome. What kinds of outcomes are most important to understand? Negative outcomes demand comprehension for survival reasons (prevention). Unexpected outcomes, by definition, indicate failures of prediction. Outcomes that almost occurred might occur in the future. Thus, these triggers reflect adaptive coping and support a functional view of counterfactual thinking (Roese, 1994).

Content

Of the infinite number of possible alternatives to reality, which does the mind select for consideration? That is, what are the typical contents of mental reconstructions? Researchers have identified several qualities that render events or antecedents more mutable. As Hofstadter (1985: p. 239) argued, there are natural 'fault lines' of the mind along which reality is cognitively cleaved.

One variable influencing the content of counterfactual thoughts is the normality of the antecedents to an event (Kahneman and Miller, 1986). When considering alternative possibilities, perceivers often focus on unusual things preceding an outcome, rather than routine aspects of the situation, with the mental reconstruction transforming the unusual antecedent into a more normal form. For example, a student who spends less time than usual studying for an examination and performs poorly is likely to think, 'If only I had studied more, I would have done better', even though many other mutations are also theoretically possible (e.g. 'If only the test had been easier').

A second feature of antecedents that increases the probability that they will be selected for counterfactual mutation is controllability. Perceivers are more likely to mutate controllable than uncontrollable aspects of a situation. For example, following a car accident at high speed on a slippery winter road, the driver is more likely to think 'If only I had driven more slowly' than 'If only it hadn't been snowing'. Serial position also influences counterfactual content. Typically, the most recent antecedents are mutated. A missed shot at the buzzer of a one-point loss in basketball is more likely to be altered than preceding misses, even though all misses were equally responsible for the outcome. If, however, several antecedents constitute a causal chain, then early events are likely to be mutated. For example, if a truck blows a tire and hits a car, which then runs into a school bus injuring some children, perceivers will think 'If only the truck hadn't blown a tire', rather than 'If only the car hadn't hit the bus'.

CONCLUSION

Counterfactual thinking, or thoughts of alternatives to past outcomes, is a common feature of everyday mental life. It exerts a variety of effects on emotion and judgment, and is thought to do so primarily through underlying mechanisms rooted in causal inference effects or contrast effects.

References

Hofstadter DR (1985) *Metamagical Themas: Questing for the Essence of Mind and Pattern*. New York, NY: Basic Books.

Kahneman D and Miller DT (1986) Norm theory: comparing reality to its alternatives. *Psychological Review* **93**: 136–153.

Miller DT and McFarland C (1986) Counterfactual thinking and victim compensation: a test of norm theory. *Personality and Social Psychology Bulletin* **12**: 513–519.

Miller DT, Turnbull W and McFarland C (1989) When a coincidence is suspicious: the role of mental simulation. *Journal of Personality and Social Psychology* **57**: 581–589.

Nasco SA and Marsh KL (1999) Gaining control through counterfactual thinking. *Personality and Social Psychology Bulletin* **25**: 556–568.

Roese NJ (1994) The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology* **66**: 805–818.

Roese NJ and Olson JM (1996) Counterfactuals, causal attributions, and the hindsight bias: a conceptual integration. *Journal of Experimental Social Psychology* **32**: 197–227.

Sanna LJ and Turley KJ (1996) Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin* **22**: 906–919.

Spellman BA and Mandel DR (1999) When possibility informs reality: counterfactual thinking as a cue to causality. *Current Directions in Psychological Science* **8**: 120–123.

Further Reading

Ferguson N (ed.) (1997) *Virtual History: Alternatives and Counterfactuals*. London: Picador.

Gilovich T and Medvec VH (1995) The experience of regret: what, when, and why. *Psychological Review* **102**: 379–395.

Harris PL, German T and Mills P (1996) Children's use of counterfactual thinking in causal reasoning. *Cognition* **61**: 233–259.

Kahneman D and Tversky A (1982) The simulation heuristic. In: Kahneman D, Slovic P and Tversky A (eds) *Judgment Under Uncertainty: Heuristics and Biases*, pp. 201–208. New York, NY: Cambridge University Press.

Lewis D (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.

Roese NJ and Olson JM (eds) (1995) *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Mahwah, NJ: Erlbaum.

Sanna LJ, Turley-Ames KJ and Meier S (1999) Mood, self-esteem, and simulated alternatives: thought-provoking affective influences on counterfactual direction. *Journal of Personality and Social Psychology* **76**: 543–558.

Tetlock PE and Belkin A (eds) (1996) *Counterfactual Thought Experiments in World Politics*. Princeton, NJ: Princeton University Press.

Creativity

Intermediate article

Thomas B Ward, Texas A&M University, College Station, Texas, USA
Katherine N Saunders, Texas A&M University, College Station, Texas, USA

CONTENTS

Creativity is the result of the convergence of basic cognitive processes, core domain knowledge, and environmental, personal, and motivational factors which allow an individual to produce an object or behavior that is considered both novel and appropriate in a particular context.

INTRODUCTION

One of the most salient features of the human mind is its capacity to generate novel ideas that are useful and appropriate for a given task or problem, that is, to exhibit creativity. Creativity is a complex phenomenon, determined by a wide range of factors, and requiring a multifaceted approach to arrive at even a partially complete understanding of the topic. This article addresses some of the issues that are important to that understanding, including a consideration of whether or not there are different types of creativity, what cognitive processes are most associated with creative outcomes, the extent to which machines can be said to be creative, the factors that limit creativity, the techniques that have been purported to enhance creativity, and the sources of individual differences in creative performance.

TYPES OF CREATIVITY

Although it is possible to describe creativity as the production of novel and useful outcomes from a convergence of skills, processes, knowledge, personal traits, environmental factors, and motivation, this general statement belies potentially important distinctions among types of creativity. These distinctions include contrasts between extraordinary and more mundane instances of creativity, and between general and specific manifestations of creativity.

Examples of the attempt to differentiate between extraordinary and commonplace forms of creativity include Boden's (1992) distinction between psychological (P) and historical (H) creativity, Gardner's (1993) contrast between 'little C' and 'big C' types of creativity, and Czikszentmihalyi's (1988) separation of personally creative and unqualifiedly creative individuals. Ideas that are P-creative are said to be novel in the mind of the individual currently having the idea, although the same ideas may have occurred to many other people before; in contrast, H-creative ideas are novel with respect to all of human history. Similarly, 'little C' creativity is manifested in everyday, small variations on themes, whereas 'big C' creativity occurs rarely and can represent a striking departure from what has come before. Personally creative people can adopt original perspectives, but unqualifiedly creative people radically alter whole domains of endeavor.

Sternberg's (1999) propulsion model introduces still more distinctions among various types of creative contributions. The model views creative work as propelling a field in different ways. These include replications that keep the field where it is, redefinitions that provide a new perspective, incrementations that move the field further in the direction it is already going, and redirections that take it in a new direction.

A question of some debate is how best to account for extraordinary versus everyday manifestations of creative behavior. One approach is to suggest that the minds of those who make notable creative contributions operate according to fundamentally different sets of rules than the minds of those whose generative accomplishments are more mundane. Alternatively, the cognitive processes may be similar, but major breakthroughs may occur only with very special convergences of personal, social, historical, and societal factors. The thought processes presumed to be involved in generating novel ideas (e.g. combining of concepts, analogical

reasoning, imagery) are ones available to most humans, albeit on perhaps a lesser scale for most, but it remains for future research to delineate the ways in which these processes are invoked in everyday and extraordinary creative accomplishments.

Another long-standing issue in the field is whether creativity can best be characterized as domain-specific or domain-general. Do creative individuals, in general, possess some common, core set of traits and abilities that would allow them to function creatively in any of a variety of domains, or do the traits and abilities needed for creative accomplishment differ considerably from one domain to the next? One approach to providing evidence on this question has been to assess the personality traits of creative artists versus those of creative scientists. Data from these types of studies support a position between the extremes of pure domain-specificity and complete domain-generality. Creative artists and creative scientists appear to share some traits and differ on others (Barron and Harrington, 1981; Feist, 1999). For example, creative artists have been reported to be more open to experience, fantasy-oriented, imaginative, driven, and ambitious, and to demonstrate higher levels of anxiety, emotional sensitivity, and independence than non-artists. Creative scientists, on the other hand, have been described as possessing traits of arrogance, drive, introversion, flexibility of thought, ambition, and independence. Thus, in either domain, a basic level of unorthodox thought and behavior is characteristic, but achieving eminence in a scientific frontier may require a greater degree of conscientiousness, responsibility, and emotional stability than that which is found in the creative artist.

Another approach to the question of how different or similar artistic and scientific creativity are is to examine the cognitive processes associated with the production of novel ideas in each of the domains. Although specialized skills would be expected to contribute differentially to success in particular domains (e.g. visuo-spatial ability for art or pitch discrimination for music), many of the most basic generative processes, such as combining previously separate concepts and using analogies, are relevant in virtually all creative domains (e.g. Finke *et al*., 1992).

INCUBATION, INSIGHT, AND OTHER CREATIVE PROCESSES

A widely noted creative phenomenon is incubation, defined as a temporary withdrawal from the problem at hand, which may culminate in an illumination or insight; that is, a sudden realization of a problem solution. Interestingly, there is much less experimental evidence regarding incubation than would be expected from the broad dissemination of the term. Historical anecdotes abound, including Archimedes' purported recognition of the principle of displacement while bathing, and Kekulé's realization regarding the circular structure of benzene while dozing by the fire. What such anecdotes have in common is a solution sequence in which the thinker devotes considerable deliberate effort towards solving a problem, reaches an impasse, withdraws temporarily, and is then struck with a sudden realization for a problem solution.

Although the phenomena of incubation and insight are broadly noted, the mechanisms by which incubation may facilitate insights are not well established. Theoretical mechanisms that have been proposed include conscious work, unconscious work, forgetting of interfering material, recovery from fatigue, and assimilation of cues encountered by chance during the incubation period.

According to the conscious work hypothesis, deliberate effort can continue on the problem while the thinker is engaged in routine tasks, such as bathing, which require only limited cognitive resources. Because the conscious thoughts that led to the solution may be quickly forgotten, the insight may appear to come 'from out of the blue'.

The unconscious work hypothesis also holds that work continues on the problem during the incubation phase, but the work occurs below the level of conscious awareness. That is, the effort is not consciously noted and then forgotten, but rather it is not available to consciousness at all.

The forgetting hypothesis states that inappropriate strategies adopted and ideas considered during initial work on a problem may be forgotten during incubation, which can facilitate the retrieval or generation of more appropriate ideas.

Recovery from fatigue holds that incubation serves as a kind of rest period during which the problem-solver can recover from the debilitating effects of an extended period of deliberate mental effort on the problem.

Finally, according to the opportunistic assimilation view, the problem-solver remains sensitive to cues in the environment that may relate to unsolved problems, even while not engaged in deliberate effort on the task.

There is little experimental research that clearly favors one view of incubation over the others, but at least some laboratory studies by S. Smith and his

collaborators (e.g. Smith, 1995) are consistent with the forgetting hypothesis.

Some models of insight attempt to specify component processes that work in concert to produce the phenomenon. For example, Sternberg and Davidson's (1995) model includes subprocesses of selective encoding of problem-relevant information, selective comparison of new and old information, and selective combination of different pieces of information.

Experimental findings do reveal some differences between insight problem-solving and analytic or logical problem-solving. For example, Metcalfe (1986) has shown that feelings of 'warmth' or progress towards a solution increase gradually as subjects near solutions to analytic problems, whereas they jump dramatically for insight problems. In addition, J. Schooler has shown that verbalization can interfere with insight problem-solving, but not analytic problem-solving (Schooler and Melcher, 1995). Such results suggest that insight may be the result of special processes unlike those involved in noncreative problem-solving. However, Weisberg (1995) has attempted to show that insights, even those described in historical anecdotes, are the result of ordinary cognitive processes applied to existing knowledge. By this view, what appears as a dramatic change in awareness of a solution may well reflect a more incremental building of solution-relevant knowledge.

Although incubation occupies a special historical role in attempts to understand creative functioning, several fundamental cognitive processes have been either theorized or demonstrated to be central to the production of novel and useful ideas (see, e.g. contributions to Ward *et al*., 1997). These include conceptual combination, analogical reasoning, and mental imagery.

In conceptual combination, the thinker merges two concepts that had previously been separate. Anecdotal accounts from creative individuals often include reference to a combining of concepts underlying some important creative advance. In addition, some theorists (e.g. Rothenberg, 1979) suggest that a simultaneous consideration of opposing concepts, termed 'Janusian' thinking, is a particularly important source of emergently creative ideas, and laboratory research on how people interpret novel combinations of concepts is beginning to provide support for this idea.

Analogical reasoning, in which a thinker uses information from a familiar domain to aid in understanding a less familiar domain, is also a central process underlying creative accomplishment. Historical cases abound in science, music, art, and literature. Recent analyses, such as Gentner's (1997) examination of Johannes Kepler's use of analogy in reasoning about the nature of the solar system, have related historical accounts directly to principles from contemporary process theories. That work has helped to establish the validity of claims that analogies between distant knowledge domains can underlie great creative advances. Studies of reasoning among contemporary scientists, such as Dunbar's (1997) look at the ongoing activities of molecular biology laboratories, also reveals that analogies to closely related domains (as opposed to distant domains) often dominate the day-to-day reasoning involved in creative breakthroughs.

MAKING MACHINES CREATIVE

A number of attempts have been made to get computers to function creatively, and Boden (1992) has provided a thorough account of these efforts. An important goal of such computational approaches is to develop a better understanding of human creativity by attempting to simulate it. In that sense, to the extent that creative outcomes spring from fundamental cognitive processes, even computational models of basic processes such as analogy (e.g. Structure Mapping Engine, or SME) are relevant to the issue of making machines creative.

In addition to computational attempts to understand broad processes such as analogy, there are also more direct attempts at simulating specific instances of creativity (e.g. scientific discovery). One of the best-known examples of such an attempt is BACON, which used heuristics to simulate the discovery of scientific laws (Bradshaw *et al*., 1983). BACON was shown to be able to rediscover Kepler's laws of planetary motion from a set of heuristics and data on observations of planetary motion, although it has come under criticism for underrepresenting the complexities involved in real-world instances of discovery. Such programs, along with others concerned with creativity in drawing, literature, and music, still leave much to be desired, but they do represent important first steps.

LIMITS TO CREATIVITY

Both individual and environmental factors can provide limits to creativity. It is clear that below some minimum level of intellectual ability (e.g. an IQ of 85), a person would have a limited capacity to generate and express creative ideas, although studies tend not to examine creative functioning in

those individuals. Studies on individuals with somewhat higher scores have shown that creative performance is linked with intellect in individuals with an IQ below 120, but this link all but disappears in individuals with IQs above 120 (Barron and Harrington, 1981).

Environmental factors also play a role in the expression of creativity. Extensive research by Amabile has found that the use of external rewards or evaluations decreases task motivation for creativity and overall creative performance in both adults and children (Hennessey and Amabile, 1988). This negative effect of reward on creative performance is so strong, in fact, that Amabile found that merely the expectation of some sort of external reward or evaluation diminished creative task motivation and performance in the same way that actually using such external constraints had.

TECHNIQUES FOR IMPROVING CREATIVITY

A wide variety of techniques have been developed with the goal of trying to improve creative performance. Some have emphasized the dynamics of group interaction, others the learning of specific idea-generation techniques, and still others the enhancement of intrinsic motivation.

One of the earliest and best-known techniques is brainstorming. Developed by Osborn (1953), this technique is designed to enhance creativity by encouraging groups (and individuals) to generate as many ideas as possible about a problem without expressing criticism towards those ideas. By eliminating criticism and allowing individuals to 'piggy-back' on ideas suggested by other group members, brainstorming is supposed to result in more ideas being generated, some of which may be extremely creative and provide excellent solutions to the problem being considered. Although there is some support for the usefulness of the procedure, a number of studies have actually shown a productivity loss in groups. That is, groups sometimes produce fewer ideas than the same number of individuals working independently. Thus, the question of where, when, and how brainstorming improves creative performance is yet to be resolved.

Another well-known attempt to enhance creativity is Edward deBono's (1970) lateral thinking approach, which encourages people to engage in thinking that moves off in different directions and to adopt many different perspectives on a problem, rather than thinking along a single narrow path. A major aspect of the approach is to teach people specific techniques designed to facilitate this type of broad attack on a problem, including the 'six hats' approach, in which people 'wear different types of hats', that correspond to different modes of thought (e.g. critical versus generative) (de Bono, 1985). An approach that makes use of idea-generation techniques and principles to facilitate group interaction is Gordon's (1961) synectics procedure, in which group members are coached to generate ideas using analogies and metaphor while also being instructed to suspend criticism of ideas generated by others.

Brainstorming, lateral thinking, synectics, and a host of other procedures have enjoyed a great deal of popular success, but do they make people more creative? To some extent the answer depends on how one defines and measures creativity. Although any given technique may be shown to facilitate performance on a particular task, the extent to which such changes in generative performance last or generalize beyond the immediate situation is less clear. Thus, it may be more appropriate to state that various training procedures can alter patterns of performance on a range of generative tasks, rather than to claim that they make people more creative.

Another avenue of training for improving creativity has been to increase creative motivation. Developed by Amabile and colleagues, this training paradigm, called inoculation training, seeks to increase creativity by training individuals to focus on the intrinsic joy that creative activities bring (Hennessey and Amabile, 1988). Developed as a way to counteract the negative effects of external reward on creative performance, inoculation training involves talking to groups about the internal or intrinsic rewards of behaving creatively. This is done in conjunction with watching videos demonstrating others behaving creatively in the face of external reward and finding pleasure in just engaging in the creative act alone. The use of this type of training has been shown to increase creative performance of both schoolchildren and adults on tasks which have an element of reward or evaluation associated with them.

MODELS OF CREATIVITY

Models of creativity differ in their scope, in the factors they emphasize, and in the tendency to view creativity as stable or malleable. Although historical models of creativity sought to explain creative behaviors as a reflection of differences in individual personalities, beginning with Guilford

(e.g. 1967, 1968) creativity began to be viewed as a set of traits which, though stable, were influenced by motivation and temperament (Brown, 1989). Creativity was seen as the result of a set of traits such as problem sensitivity, fluency, flexibility, complexity, evaluation, the use of novel ideas, the ability to break down existing symbolic structure, and the general tendency to organize ideas into larger patterns. When conceptualized this way, variations in creativity could be measured using a variety of open-ended tests.

Torrance modified Guilford's definition slightly, viewing creativity as the combination of ability, skills, and motivation (Ford and Harris, 1992). By including skills in the account of individual differences, creativity became a teachable entity to the degree that a person's creative skills could be improved.

In a further departure from the focus on creativity as a personality trait, Amabile developed a model in which creativity cannot be simply the result of a single isolated personality trait or process, but rather must be accounted for by a constellation of personal characteristics, cognitive abilities and processes, and social environment factors. By this approach, creativity emerges from the confluence of domain-relevant skills, creativity-relevant skills, and task motivation (Amabile, 1990).

Gruber (1988), in what is known as the evolving systems approach, has also regarded creativity as the merging of personal knowledge, affect, and purpose. According to this developmental approach, creativity is the result of developmental changes in knowledge systems that result from the increasingly different situations that a person encounters over time. In this theory, creativity is an extended process, with a person having more than one insight or metaphor over time, and with multiple changes in thoughts and knowledge systems along the way.

A focus on a concert of factors as the root of creativity can also be seen in the burgeoning of research attempting to explain creativity as a multifaceted concept. Called *componential theories of creativity*, such theories hold that creativity occurs when a variety of biological, cognitive, and social factors merge or interact. As an example of one of the modern componential theories, Czikszentmihalyi (1988) regards creativity as an interaction of components both within and outside the individual. According to this model, creativity results from the interaction of the individual with any given domain of knowledge and those controlling the field of that domain. An individual is creative only to the extent that he or she can use cognitive processes, personality traits, and motivation to alter a particular domain in a way that is acceptable to the field at large. More recently, Sternberg and Lubart (1999) have also pursued this idea of creativity as the convergence of multiple components. In their *investment theory of creativity*, creative people are those who can 'buy low and sell high'; that is, generate or adopt ideas before they become popular, then popularize them, thus becoming associated with novel, impact-producing ideas. Such behavior is thought to require the merging of six resources: intellectual abilities, knowledge, thinking styles, personality traits, motivation, and environment.

In contrast to componential approaches, which provide a global account of the factors that interact to determine the creative impact of novel ideas, cognitive models focus more narrowly on the way in which basic cognitive processes operate on existing knowledge structures to produce those novel ideas. The models acknowledge that social and motivational factors can influence the likelihood or intensity of engaging in particular processes. Similarly, they acknowledge that factors outside the individual's thought processes will determine the extent to which an idea is judged acceptable or has an impact. However, they view cognitive processes as the crucial source of the ideas to be judged, and to some extent of the judgments as well.

Often called *process approaches*, these cognitive models focus on the acts of problem-identification and solution-generation as the keys to creative production. An example of this type of model is the Geneplore model of Finke *et al*. (1992), which characterizes the development of novel and useful ideas as resulting from an interplay between *generative* processes, that produce candidate ideas of varying degrees of creative potential, and *exploratory* processes that expand on that potential. Generative processes such as retrieval, conceptual combination, and analogical reminding are assumed to result in candidate ideas, which vary in their apparent novelty, surprisingness, aesthetic appeal, or other factors that would influence the creative person's perception that they hold promise for solving the current problem. People can use such properties to determine which ideas to develop by way of exploratory processes that modify, elaborate, consider the implications, assess the limitations, or otherwise transform the candidate ideas. The model also assumes that real-world constraints, such as the social acceptability of particular ideas, can influence the form of initially generated ideas, the person's judgment about which ideas to

explore, or the way in which a candidate idea is modified through exploratory processes.

Other models focus on the production and retention of novel ideas, and make use of a Darwinian perspective: many variations on ideas may be developed but only the fittest will be selected and survive. Simonton (1999b) has extended the evolutionary view and claimed that the production of creative ideas should be viewed as akin to blind variation, in which the creator does not have any notion of whether a given generated idea will be successful or not. While others adopt a somewhat similar generation/selection view, they do not necessarily endorse the blind variation notion (e.g. Johnson-Laird, 1988; Perkins, 1998; Sternberg, 1998).

UNDERSTANDING INDIVIDUAL DIFFERENCES IN CREATIVITY

Traditional research concerned with individual differences in creative performance attributes those differences to one of two sources: differences in the ability or tendency to use particular creative thought processes, and differences in personality attributes thought to be related to creative behaviors. The work has made use of psychometric procedures as well as assessments of the historical record of the achievements of eminent creators.

Research concerned with thought processes has focused on individual differences in the ability to identify or recognize problems or solutions that have creative potential, to tap into broad thought networks, and to apply this expanded base of knowledge to the task at hand. Defining differences in creativity as the result of individual differences in the ability to associate or bring together different elements of thought to form new and useful creations is the hallmark of the associative approach to creativity (Brown, 1989). This associative approach is not a new one in psychology, as can be seen in the many introspective studies and historical anecdotes concerning the creative process (Barron and Harrington, 1981; Brown, 1989). Mednick (1962) extended this approach by defining creativity as the forming of associative elements into new combinations, which either meet specific requirements or are in some way useful. By this view, individual differences would be attributable to the ability to access remote associations, which in turn gives rise to the use of the Remote Associates Test as a measurement technique (Mednick and Mednick, 1967).

Individual differences in creative thought have also been explained by variations in divergent thinking ability, including fluency (the tendency to produce many ideas), flexibility (the tendency to produce differing ideas), and originality (the tendency to produce ideas that are normatively uncommon). A classic example of a divergent thinking task used to measure such differences is the Torrance Test of Creative Thinking (Torrance, 1974) in which people generate questions, unusual uses, and/or drawings in response to particular stimuli.

One question that can be raised about paper-and-pencil measures of divergent thinking ability is whether or not performance on those measures is indicative of real-world creative skill. Although various researchers have found a relationship between test performance and real-world indicators, such findings have not been consistent. Measures of divergent thinking, while related to some indices of creative achievement, are often unable to significantly predict creative achievement and behaviors in a real-world setting (Barron and Harrington, 1981; Brown, 1989). In addition, concerns have been raised about the domain-specificity of divergent thinking.

Another type of explanation for individual differences in creativity focuses on personality traits, and assumes that creativity differences are based on variations in personality attributes that are thought to contribute to creative production. The characteristics that have been identified as important to creativity are tolerance for ambiguity, openness, independence, positive sense of self, high energy, general curiosity, wide interests, as well as introversion, attraction to complexity, need for recognition, and a variety of others (Barron and Harrington, 1981). As indicated previously, however, the importance of each of the characteristics may vary according to the domain of creativity being pursued. In fact, the search for the single set of 'creative personality' traits that map onto real-world creative performance has, so far, been unsuccessful.

Variations in intrinsic motivation are another possible source of individual differences in creative performance. For instance, Amabile, in her seminal research on the relationship between intrinsic motivation and creativity, found that creative performance in such areas as writing and art can be both enhanced and hindered by changes in intrinsic interest in a task. Hennessey and Amabile (1988), in a continuation of this line of research, have proposed the intrinsic motivation principle of creativity which says that people will be the most creative when they feel motivated to perform primarily by the interest and enjoyment of the task, and not by external factors such as reward or

punishment. Thus, all individual differences in creative performance are due to those differences in motivation towards the task at hand, and hence, all creativity is, at heart, domain-specific to the interests of the individual.

Although much contemporary work on individual differences relies on tests or laboratory observations of a broad sampling of participants, another approach involves detailed, narrative case studies of a small set of highly creative individuals in history. Somewhere between these extremes is Simonton's (1999a) historiometric approach in which historical data (e.g. number of publications, citations, performances, and so on) is sampled for a large number of contributors to a field, and statistical tests are performed to relate those measures to other indices. The approach can be used to examine a broad range of factors, including intellectual precocity, family background, and propensity towards mental illness. It goes beyond paper-and-pencil measures of the attributes of the many to a detailed look at individual differences among the eminent.

Acknowledgements

This work was supported by the National Science Foundation under Grant No. BCS-9983424.

References

Amabile TM (1990) Within you, without you: the social psychology of creativity and beyond. In: Runco MA and Albert RS (eds) *Theories of Creativity*, pp. 61–91. Newbury Park, CA: Sage.

Barron FX and Harrington DM (1981) Creativity, intelligence, and personality. *Annual Review of Psychology* **32**: 439–476.

Boden M (1992) *The Creative Mind: Myths and Mechanisms*. New York, NY: Basic Books.

Bradshaw GF, Langley PW and Simon HA (1983) Studying scientific discovery by computer simulation. *Science* **222** (4627): 971–975.

Brown RT (1989) Creativity: what are we to measure? In: Glover JA, Ronning RR and Reynolds CR (eds) *Handbook of Creativity*, pp. 3–32. New York, NY: Plenum.

Czikszentmihalyi M (1988) Society, culture, and person: a systems view of creativity. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 325–339. New York, NY: Cambridge University Press.

De Bono E (1970) *Lateral Thinking*. New York, NY: Harper.

De Bono E (1985) *Six Thinking Hats*. Boston, MA: Little, Brown and Co.

Dunbar K (1997) How scientists think: on-line creativity and conceptual change in science. In: Ward TB, Smith SM and Vaid J (eds) *Creative Thought: An Investigation of Conceptual Structures and Processes*, pp. 461–493. Washington, DC: American Psychological Association.

Feist GJ (1999) The influence of personality on artistic and scientific creativity. In: Sternberg RJ (ed.) *Handbook of Creativity*, pp. 273–296. New York, NY: Cambridge University Press.

Finke RA, Ward TB and Smith SM (1992) *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: MIT Press.

Ford D and Harris JJ (1992) The elusive definition of creativity. *Journal of Creative Behavior* **26**: 186–198.

Gardner H (1993) *Creating Minds*. New York, NY: Basic Books.

Gentner D, Brem S, Ferguson RW and Wolff P (1997) Analogy and creativity in the works of Johannes Kepler. In: Ward TB, Smith SM and Vaid J (eds) *Creative Thought: An Investigation of Conceptual Structures and Processes*, pp. 403–459. Washington, DC: American Psychological Association.

Gordon WJ (1961) *Synectics: The Development of Creative Capacity*. New York, NY: Harper and Row.

Gruber HE (1988) The evolving systems approach to creative work. *Creativity Research Journal* **1**: 27–51.

Guilford JP (1967) Creativity: yesterday, today and tomorrow. *Journal of Creative Behavior* **1**(1): 3–14.

Guilford JP (1968) *Intelligence, Creativity, and their Educational Implications*. San Diego, CA: Robert R Knapp.

Hennessey BA and Amabile TA (1988) The conditions of creativity. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 11–38. New York, NY: Cambridge University Press.

Johnson-Laird PN (1988) Freedom and constraint in creativity. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 202–219. New York, NY: Cambridge University Press.

Mednick SA (1962) The associative basis for the creative process. *Psychological Review* **69**: 220–232.

Mednick SA and Mednick MT (1967) *Remote Associates Test, College and Adult, Forms 1 and 2 and Examiner's Manual*. Boston, MA: Houghton Mifflin.

Metcalfe J (1986) Feeling of Knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **12**(2): 288–294.

Osborn AF (1953) *Applied Imagination*. New York, NY: Scribner's.

Perkins DN (1998) Is the country of the blind an appreciation of Donald Campbell's vision of creative thought. *Journal of Creative Behavior* **32**(3): 177–191.

Schooler JW and Melcher J (1995) The ineffability of insight. In: Smith SM, Ward TB and Finke RA (eds) *The Creative Cognition Approach*, pp. 97–134. Cambridge, MA: MIT Press.

Simonton DK (1999a) Creativity from a historiometric perspective. In: Sternberg RJ (ed.) *Handbook of Creativity*, pp. 116–133. Cambridge, UK: Cambridge University Press.

Simonton DK (1999b) *Origins of Genius: Darwinian Perspectives on Creativity*. New York, NY: Oxford University Press.

Smith SM (1995) Getting into and out of mental ruts: A theory of fixation, incubation, and insight. In: Sternberg RJ and Davidson JE (eds) *The Nature of Insight*, pp. 229–251. Cambridge, MA: MIT Press.
Sternberg RJ (1998) Cognitive mechanisms in human creativity: is variation blind or sighted? *Journal of Creative Behavior* **32**(3): 159–176.
Sternberg RJ (1999) A propulsion model of types of creative contributions. *Review of General Psychology* **3**: 83–100.
Sternberg RJ and Davidson JE (eds) (1995) *The Nature of Insight*. Cambridge, MA: MIT Press.
Sternberg RJ and Lubart TI (1999) The concepts of creativity: prospect and paradigms. In: Sternberg RJ (ed.) *Handbook of Creativity*, pp. 3–15. New York, NY: Cambridge University Press.
Torrance EP (1974) *Torrance Tests of Creative Thinking: Norms-technical Manual*. Bensenville, IL: Scholastic Testing Service.
Torrance EP (1988) The nature of creativity as manifest in its testing. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 43–75. New York, NY: Cambridge University Press.
Ward TB, Smith SM and Vaid J (eds) (1997) *Creative Thought: An Investigation of Conceptual Structures and Processes*. Washington, DC: American Psychological Association.
Weisberg RW (1995) Prolegomena to theories of insight in problem solving: A taxonomy of problems. In: Sternberg RJ and Davidson JE (eds) *The Nature of Insight*, pp. 157–196. Cambridge, MA: MIT Press.

Further Reading

Amabile TM (1983) *The Social Psychology of Creativity*. New York, NY: Springer.
Czikszentmihalyi M (1996) *Creativity: Flow and the Psychology of Discovery and Invention*. New York, NY: HarperCollins.
Guilford JP (1950) Creativity. *American Psychologist* **5**: 444–454.
Perkins D (1988) The possibility of invention. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 362–385. New York, NY: Cambridge University Press.
Rothenberg A (1979) *The Emerging Goddess*. Chicago, IL: University of Chicago Press.
Runco MA and Chand I (1995) Cognition and creativity. *Educational Psychology Review* **7**: 243–267.
Sternberg RJ and Lubart TI (1995) *Defying the Crowd: Cultivating Creativity in a Culture of Conformity*. New York, NY: Free Press.
Torrance EP (1988) The nature of creativity as manifest in its testing. In: Sternberg RJ (ed.) *The Nature of Creativity*, pp. 43–75. New York, NY: Cambridge University Press.
Ward TB, Finke RA and Smith SM (1995) *Creativity and the Mind: Discovering the Genius Within*. New York, NY: Plenum.

Cultural Differences in Abstract Thinking

Introductory article

Fons J R van de Vijver, Tilburg University, The Netherlands

CONTENTS

Abstract thinking is a central part of reasoning and the highest cognitive attainment in Piagetian theory. Studies of cross-cultural differences and similarities in abstract thinking show its relationship with culture.

INTRODUCTION

There are two research traditions in cognitive psychology for examining the relationship between culture and abstract thinking: the formal and the informal approach (Table 1). In formal research the scientific approach is the normative model of good problem-solving; there is an emphasis on the application of inductive and deductive reasoning, the solution of formalized problems that are unlikely to be met in everyday life, and the correctness of solutions (e.g., 'continue the following series: 1, 2, 4, 8, 16, …'). In the informal tradition the 'bricoleur' (jack-of-all-trades) is the implicit model of problem-solving. There is an emphasis on problem-solving in everyday life: an example would be, 'What would you do when you are due for promotion at your work and a reliable source tells you that your direct colleague may be promoted instead of you?' These two more or less independent traditions have their own models of cross-cultural differences and similarities of abstract thinking.

FORMAL STUDIES OF ABSTRACT THINKING

Two types of study predominate in the formal tradition, differing in theoretical orientation and assessment procedures: psychometric research (based on Western models of intelligence and using psychological 'paper and pencil' tests) and the Piagetian approach (applying Piagetian theory and tasks).

Psychometric studies of abstract thinking have used a variety of tests (such as Raven's Progressive Matrices, Wechsler's Intelligence Scales and Cattell's Culture Fair Intelligence Test) and investigated a number of cultures. Many cultural comparisons involved participants from different countries; especially in the USA, research often compared the performance of different ethnic groups, mainly African Americans and Anglo-Americans. These studies have shown a remarkable consistency in results. Using advanced statistical techniques (mainly exploratory and confirmatory factor analysis), it has been shown that the structure of intelligence is identical across cultural groups; tests of abstract thinking tend to be related to reasoning and general intelligence. Broad cognitive abilities, such as reasoning, memory and visualization, are universal. Moreover, analyses of test performances have consistently shown that the difficulty order of items tends to be invariant across cultures; within the homogeneous domains used in the tests items that are easy in one culture are likely to be easy in another culture (though not necessarily solved by the same proportion of the population). The psychometric tradition does not support the alleged qualitative difference in abstract thinking between Western and non-Western individuals, historically often associated with Lévy-Bruhl. To the best of our knowledge abstract thinking is a universal attainment that can be found in all cultural groups.

Nevertheless, studies have reported consistent differences in scores on tests of abstract thinking between Western and non-Western groups, with the former groups usually obtaining higher scores. Similarly, comparisons of scores obtained by different ethnic groups in the USA have shown consistent differences in scores, ranking as follows (from high to low): East Asians (e.g. Chinese and Japanese), European Americans, Hispanics and African Americans. Interpretation of these differences has been controversial. Some researchers, such as Jensen and Eysenck, argue that the differences in performance are to be interpreted as cultural

Table 1. Differences between formal and informal research traditions

Formal tradition	*Informal tradition*
Closed problem spaces	Open problem spaces
Deterministic problems with one correct answer	Probabilistic problems with several correct answers
Contrived problems	Problems derived from everyday life
Academic intelligence	Practical intelligence
Focus on correctness of solution (is the solution correct?)	Focus on practical value of the solution (does the solution solve the problem?)
Problems and solutions are context-independent	Problems and solutions are context-dependent
Scientist as normative model of good problem-solver	Bricoleur as normative model of problem-solver
Algorithmic solutions	Heuristic solutions
Product-oriented (psychological tests, Piagetian tasks)	Process-oriented
Solution requires conceptual, theoretical knowledge	Solution requires practical intelligence
Cross-cultural comparison of test performance	Studies within a single culture

differences in cognitive ability. These authors attributed these performance differences to genetic differences between cultures. The current immature status of knowledge in behavior and molecular genetics does not yet allow for precise estimates of the role of genetics in cross-cultural differences in abstract thinking. Critics of Jensen and Eysenck often point to the influence of schooling and potential problems in the tests used to assess abstract thinking as a source of cross-cultural performance differences. Many psychological tests in the formal tradition have a format and test contents that are school-related. Schooling has been found to enhance the performance on tests of abstract thinking but has no formative influence on abstract thinking; however, the occurrence of abstract thinking among illiterate individuals and groups demonstrates that schooling is not a precondition for developing abstract thinking.

In Piagetian theory abstract thinking is part of formal-operational reasoning, which, according to Piaget, is acquired by Western subjects at age 12–15 years. Various measures of formal-operational reasoning are available; these are often based on elementary laws from physics. The person tested has to evaluate the impact of presumably relevant variables by experimental manipulation. One such task deals with the oscillation time of a pendulum. The variables that are presumably relevant are the length of the pendulum string, the weight of the object fastened to the string, the angle of the swing of the weight, and the momentum given to the weight at the start of the swing. The participant is asked to determine experimentally which of the four variables determines the oscillation time.

Many cross-cultural studies have addressed earlier Piagetian stages (e.g. the transition from preoperational to concrete-operational thinking), but formal-operational thinking has not been extensively studied. Studies among unschooled non-Western individuals invariably showed a poor performance, which was sometimes interpreted as evidence that non-Western individuals did not show abstract thinking. Some theoreticians even conjectured that abstract thinking was an achievement of Western, industrialized nation states. Later research redressed the picture. First, cognitive anthropological evidence argues strongly against the cultural specificity of abstract thinking; for example, Kalahari Bushmen show on average low scores on intelligence tests, yet their tracking and navigation skills in the desert far exceed those of Westerners and show great cognitive complexity. Moreover, even in Western societies many participants gave a poor performance on Piagetian tests. It was increasingly appreciated that the method of assessment may affect the test outcome; that individuals are unable to solve physics problems which are remote from their daily reality does not imply that their abstract thinking is undeveloped. A distinction was introduced between competence and performance: whereas the latter refers to actual test performance, the former refers to the performance in optimal conditions dealing with common tasks. The distinction implicitly points to the potential problems of Piagetian tasks in cross-cultural research, but subsequent research has not identified new procedures to assess the competence. It is widely believed now that although the performance on Piagetian tests of formal-operational thinking often points to the absence of abstract thinking, the competence is universal. Abstract thinking is a universal attainment, but the domains of application may vary across individuals and cultures. A car mechanic may be able to use abstract thinking in dealing with cars, but

perform badly when applying laws of logic to other domains. It is paradoxical that abstract thinking, theoretically based on context-independent rules such as the laws of logic, turned out to be domain-specific.

INFORMAL STUDIES OF ABSTRACT THINKING

The label 'informal' is used here as a summary label for various traditions in the psychological literature, such as 'everyday cognition', 'indigenous cognition' and 'practical intelligence'. Common to these studies is their emphasis on the observation of cognitive processes in everyday problem-solving.

Mathematical thinking has been often studied. For example, shoppers were interviewed while they bought groceries in order to determine how they determined 'best buys'. It was consistently found that they seldom relied on the arithmetic competence they acquired in school. Rather, arithmetic in school and in everyday life seem to constitute different competencies. In one experiment a group of women were asked which of two cans of peanuts they would buy on the basis of a comparison of prices: can A weighing 10 oz for 90 cents, or can B weighing 4 oz for 45 cents. In another test the same women were asked to compare the ratios 90/45 and 10/4. From a mathematical perspective the two problems are identical, but from a psychological perspective they are very different. The former problem was correctly solved more often.

Some studies have examined the relationship between skills applied in everyday life and skills to solve tests in the formal tradition. It has been consistently found that, possibly contrary to expectation, there is no relationship between the scores on 'school' tests and 'everyday' tests in the same domain. Individuals who are capable of displaying highly skilled behavior in the context of their professional specialization are not necessarily the individuals with the highest scores on intelligence tests.

Another line of research has examined to what extent complex cognitive skills acquired in the course of learning a profession generalize beyond this professional context. For example, Zinancanteco women in Mexico can weave highly complex patterns. These women showed superior planning skills in a weaving task when they had to reproduce known patterns, but did not outperform nonweavers when the planning involved an unfamiliar task. Planning skills acquired in the context of professional training did not generalize broadly across the cognitive spectrum. This result exemplifies the findings in studies of specialized cognitive skills: the cognitive effects of mastery of a craft often do not go beyond the specific domain in which the craft is applied.

CONCLUSION

Abstract thinking has been studied from two different perspectives in cross-cultural psychology. The first, the formal tradition, focuses on the application of general principles such as the laws of logic and inductive schemes. Studies in this tradition have shown that abstract thinking is a universal attainment, but cultures may well differ in their areas of specialization. Cross-cultural differences in scores on tests of abstract thinking are typically open to multiple interpretations (e.g. valid cultural differences in abstract thinking, familiarity with testing procedures, differential cultural appropriateness of tests, confounding of cross-cultural differences in schooling). The second, the informal tradition, studies abstract thinking in action. This tradition has shed further light on the domain specificity as studied in the formal tradition. Within an area of expertise such as a profession, individuals can display remarkably high levels of performance. From a cognitive perspective these areas are often sharply delineated. The training of professional expertise often does not have a broad impact on cognitive functioning. Virtuoso performance is often restricted to one domain.

Abstract thinking shows both important similarities and differences across cultures. There is ample evidence for the universality of the basic structures of abstract thinking (general reasoning, Piagetian formal-operational thinking). There is no evidence for the existence of qualitatively different types of abstract thinking across cultures. On the other hand, there are massive cross-cultural differences on tests of abstract thinking. Moreover, domains in which individuals are able to use abstract thinking show some variation across cultures.

Further Reading

Dasen PR (ed.) (1977) *Piagetian Psychology. Cross-Cultural Contributions*. New York: Gardner.

Jensen AR (1998) *The G Factor. The Science of Mental Ability.* Westport, CT: Praeger.

Schliemann A, Carraher D and Ceci SJ (1997) Everyday cognition. In: Berry JW, Dasen PR and Saraswathi TS (eds) *Handbook of Cross-Cultural Psychology*, 2nd edn, vol. 2, pp. 177–216. Boston, MA: Allyn & Bacon.

van de Vijver FJR and Willemsen ME (1993) Abstract thinking. In: Altarriba J (ed.) *Culture and Cognition*, pp. 317–342. Amsterdam: North Holland.

Cultural Differences in Causal Attribution

Intermediate article

Douglas S Krull, Northern Kentucky University, Highland Heights, Kentucky, USA
Michael W Morris, Stanford University, Stanford, California, USA

CONTENTS

An attribution is a judgment about why an event (typically another person's behavior) occurred. Research suggests that cultures differ in the types of attributions that their members prefer.

ATTRIBUTION THEORY

An attribution is an explanation, a judgment about the cause of an event. Psychological research on attribution has primarily studied judgments about the cause of another person's behavior. Attributions for behavior are ubiquitous in everyday life (e.g. 'I think Luis achieved the highest score on the calculus exam because he has a talent for mathematics', 'Mariko is sad because her best friend moved away'). Moreover, attributions have important implications for social interaction in that they shape perceivers' expectations about others' future behavior (e.g. 'I helped Robert because I thought he was trying his best to succeed', 'I'm angry at Torsten because he took my car keys on purpose'). The ubiquity and importance of attributions has made them a central topic of social psychological research. (*See* **Judgment**)

A foundational premise of attribution theory is Heider's (1958) contention that perceivers seek to attribute fleeting behavior to stable dispositions in order to learn about the social environment. He wrote: 'It is an important principle of commonsense psychology … that man grasps reality, and can predict and control it, by referring transient and variable behavior and events to relatively unchanging underlying conditions, the so-called dispositional properties of his world' (1958, p. 79). For example, upon noticing the anxious behavior of her new co-worker, Susan would make a judgment about something stable in the environment, either that the co-worker has an anxious personality or that the co-worker's job is stressful. Although most behaviors can reflect either situational or personal influences, Heider suggested that perceivers tend to trace action to dispositions of the actor. The existence and consequences of this tendency were documented by social psychology experiments (see Ross and Nisbett, 1991 for a review). Because of its apparent ubiquity and potentially important consequences, this tendency was designated the *fundamental attribution error* (Ross, 1977).

INDIVIDUALIST AND COLLECTIVIST CULTURES

Until recently, psychologists paid little attention to the possibility that attributions might differ across cultures, but lately the study of cultural differences has captured the attention of attribution researchers. Although the study of culture in causal attribution is in its infancy, research suggests that there are substantial and potentially important differences in how people from different cultures think about behavior, as well as potentially important similarities across cultures. (*See* **Cultural Psychology**; **Cultural Differences in Abstract Thinking**)

However, the attribution researcher who desires to learn something about culture is faced with a dilemma. Given that there are many cultures, how can one hope to reach general conclusions about the role that culture plays in attribution? Without denying the fact that there may be important differences between many different cultures, psychologists have found it useful to divide cultures into those that tend towards individualism and those that tend towards collectivism. In individualist cultures (e.g. Australia, Britain, the United States), personal autonomy is emphasized, and so, for the most part, people are seen as free agents who behave as they choose. In contrast, collectivist cultures (e.g. China, Guatemala, India) tend to emphasize supporting the goals of groups and behaving in a collectively appropriate manner, and so people

are seen as constrained by social forces. As described in the sections that follow, these differences between individualist cultures and collectivist cultures have important implications for explanations of behavior.

CULTURAL DIFFERENCES IN ATTRIBUTION TENDENCIES

Judging Causes of Actions by Persons

Although the tendency to favor dispositional attributions for others' actions may well be fundamental in the sense of having many consequences, it does not seem to be fundamental in the sense of being universal. Ethnographers have long reported that lay people in some collectivist cultures refer to personality dispositions rarely and instead attribute behavior frequently to social roles (for a review, see Shore, 1996). More controlled, quantitative evidence for this claim first came in a study by Miller (1984) which asked participants of various ages from the USA and India to explain everyday actions that they had observed. Young children in both cultures were alike in the proportional frequency of their references to personality traits and to situational factors. However, as age increased, Americans showed an increasing reliance on attributions to personality, and Indians an increasing reliance on situational attributions. Thus, it seems that North Americans learn that behavior is primarily caused by personality, whereas Indians learn that behavior is primarily caused by the circumstances.

Although this initial evidence for cultural differences from everyday explanations of behavior had compelling external validity, it is open to multiple interpretations. It might reflect cultural differences in attributions, but it might reflect merely that everyday behaviors, and the actual causes thereof, differ across cultures. Indeed, there is evidence that personality traits do account for more variance in behavior in individualist cultures, whereas social roles and situations account for more variance in collectivist cultures (e.g. Argyle *et al*., 1978; Triandis, 1995). To clarify the role of culture, Morris and Peng (1994) conducted several studies that examined attributions for the same event, such as a prominent crime covered by American and Chinese newspapers. Results showed that attributions for these events differed across cultures, suggesting that cultural differences indicate a difference in the interpretive tendencies of perceivers, not just in the events they typically explain. Insight about how perceivers differ is accumulating from studies of different kinds of attributions. (*See* **Causal Reasoning, Psychology of**; **Causal Perception, Development of**)

Judging Personal Traits from Behaviors

One of the most important research paradigms within attribution theory focuses on how perceivers judge traits from situationally constrained behavior. Strictly speaking, the task in these studies is not to explain the cause of the actor's behavior but to judge the degree to which an actor's personality corresponds to his or her behavior. A wealth of research with individualist participants suggests that people often infer that personality matches behavior, even when situational forces that are sufficient to explain the behavior are present. For example, although we know that television actors are only playing roles, we may assume that their personalities correspond to the characters that they play. This tendency has been called the *correspondence bias* (see e.g. Gilbert and Malone, 1995 for a review).

The correspondence bias seems to be a multiply determined phenomenon. Gilbert and Malone (1995) have suggested that the bias can arise through at least four distinct mechanisms. People may display correspondence bias because the situation is not immediately apparent (e.g. Bernard infers that Tia's anxiety reflects an anxious personality, because Bernard does not know that Tia is about to give an important speech). People may display correspondence bias because they fail to appreciate the power of the situation (e.g. Anne may infer that Tia has an anxious personality because Anne does not realize that making a speech can be anxiety-provoking). People may display correspondence bias because they are too busy to consider the influence of the situation (e.g. Dolf infers that Tia is an anxious sort of person because he is distracted and so does not fully consider the anxiety-provoking situation). Finally, people may display correspondence bias because their knowledge of the situation inflates their perceptions of the behavior (e.g. Roland knows how anxiety-provoking giving a speech can be, so he perceives greater anxiety in Tia's behavior, and infers that only a dispositionally anxious person would be so anxious).

Given these different mechanisms, one might expect that whether or not the correspondence bias is found across cultures depends on a variety of factors. Research indeed bears out that sometimes the correspondence bias is found across cultures (e.g. Choi and Nisbett, 1998, experiment 1;

Krull *et al*., 1999). However, cultural differences arise under some conditions as a function of these mechanisms. First, although perceivers across cultures are capable of ignoring situational forces, when such forces are made salient, collectivists become less likely than individualists to show the bias (Choi and Nisbett, 1998, experiment 2). Second, collectivists recognize the strength of some situational forces, such as authority pressure, that individualists underestimate, and this is another source of cultural differences (Morris *et al*., 2000). Third, collectivists, perhaps because they are more practiced in thinking about situational constraint, are not hindered by cognitive busyness as much as are individualists (Knowles *et al*., 2001).

Judging Causes of Actions by Groups

The findings described heretofore might lead one to conclude that collectivists generally attribute to the context rather than to dispositions of actors. However, such a conclusion would be incomplete. Recent research suggests that collectivists readily attribute actions by groups to dispositions of the group, and make more reference to group dispositions when attributing events involving a combination of actions by individuals and groups (Menon *et al*., 1999, Studies 1 and 2). In another study, perceivers in separate conditions read about an act of wrongdoing by an individual or a group (Study 3); results showed that Americans showed a stronger tendency towards dispositions in response to the individual actor, and Chinese, in response to the group actor.

Additional research by Chiu and colleagues (Chiu *et al*., 2000) investigated how the need for cognitive closure affects perceivers. Results showed that time pressure increased Chinese perceivers' attributions to dispositions of groups, and increased American perceivers' attributions to dispositions of individuals. Another study that examined individuals who were chronically high or low on the need for cognitive closure found the same pattern. Thus, the desire to identify stable, dispositional properties of the social environment is by no means limited to perceivers in individualist cultures. However, perceivers in collectivist cultures are more likely to look for such properties in groups than in individuals.

TOWARDS A MODEL OF HOW CULTURE INFLUENCES ATTRIBUTION

Because research on culture and attribution is relatively new, there is no consensus on an explanatory model of how culture affects the attribution process. Broadly speaking, one model of cultural differences suggests that culture shapes general thinking principles or cognitive styles (Witkin and Berry, 1975). According to this view, collectivists are more holistic in their thinking and individualists are more analytical. In support of this view, research indicates that collectivists often seem to be more sensitive to context in a variety of perceptual and cognitive tasks (see, e.g. Nisbett *et al*., 2001, for a review). Thus, collectivists may be more aware of the situational forces that influence behavior as part of a general contextual focus of attention.

A second model suggests that differences between individualists and collectivists stem from differences in implicit causal theories. Recent research on individual differences in attribution biases (Dweck *et al*., 1995) has revealed that implicit theories can be proximal determinants of attributional biases. Accordingly, differences between cultures may or may not emerge, depending upon whether the cultures possess similar or dissimilar theories with regard to the specific domain and whether the circumstances foster theory-based processing. In support of this view, some clear boundary conditions on cultural differences have been identified, corresponding to the applicability of implicit theories. For example, Morris and Peng (1994) found that although Chinese individuals were more situational than North Americans in their judgments of behavior induced by social causality (e.g. an individual moving away from the pressure of a group), cultural differences were not obtained in judgments of behavior induced by mechanical causality (e.g. an object moving after being struck by another object). Further evidence that implicit theories are a mechanism through which cultural differences or similarities are produced comes from studies that demonstrate that cultural differences in attribution can be evoked by priming cultural knowledge (Hong *et al*., 2000). (*See* **Implicit Learning**; **Memory: Implicit versus Explicit**)

In sum, attributions may only differ across cultures when different implicit theories are activated, and activation depends on the applicability of theories to the domain and the cognitive dynamics of the perceiver (Morris *et al*., 2001). However, the two models – cognitive style and implicit theories – need not be considered as rivals, in that the former serves as a heuristic framework whereas the latter serves as a middle-range hypothesis-testing model. As described here, holism can be manifested in an implicit theory about the causal role of context in the behavior of individuals, or in an implicit theory

of groups as actors. In sum, cultural differences depend on a variety of factors, but this complexity seems amenable to social cognition models.

References

Argyle M, Shimoda K and Little B (1978) Variance due to persons and situations in England and Japan. *British Journal of Social and Clinical Psychology* **17**: 335–337.

Chiu CY, Morris MW, Hong YY and Menon T (2000) Motivated cultural cognition: the impact of implicit theories on dispositional attribution varies as a function of need for closure. *Journal of Personality and Social Psychology* **78**: 247–259.

Choi I and Nisbett RE (1998) Situational salience and cultural differences in the correspondence bias and actor-observer bias. *Personality and Social Psychology Bulletin* **24**: 949–960.

Dweck CS, Chiu C and Hong Y (1995) Implicit theories and their role in judgments and reactions: a world from two perspectives. *Psychological Inquiry* **6**: 267–285.

Gilbert DT and Malone PS (1995) The correspondence bias. *Psychological Bulletin* **117**: 21–38.

Heider F (1958) *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Hong YY, Morris MW, Chiu CY and Benet-Martinez V (2000) Multicultural minds: a dynamic constructivist approach to culture and cognition. *American Psychologist* **55**(7): 709–720.

Knowles E, Morris MW, Chiu CY and Hong YY (2001) Culture and cognitive-process models of attribution: evidence for automatic situational correction among East Asians. *Personality and Social Psychology Bulletin* **27**: 1344–1356.

Krull DS, Loy MH–M, Lin J *et al*. (1999) The fundamental fundamental attribution error: correspondence bias in individualist and collectivist cultures. *Personality and Social Psychology Bulletin* **25**: 1208–1219.

Menon T, Morris MW, Chiu CY and Hong YY (1999) Culture and the construal of agency: attribution to individual versus group dispositions. *Journal of Personality and Social Psychology* **76**: 701–717.

Miller JG (1984) Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology* **46**: 961–978.

Morris MW, Knowles E, Chiu CY and Hong YY (2000) *Culture and Judgment of Obedient and Disobedient Acts: Perceived Situational Force and Cultural Role Expectations*. Unpublished manuscript, Graduate School of Business, Stanford University.

Morris MW, Menon T and Ames DR (2001) Culturally conferred conceptions of agency: a key to social perception of persons, groups, and other actors. *Personality and Social Psychology Review* **5**: 169–182.

Morris MW and Peng K (1994) Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology* **67**(6): 949–971.

Nisbett RE, Peng K, Choi I and Norenzayan A (2001) Culture and systems of thought: holistic vs. analytic cognition. *Psychological Review* **108**: 291–310.

Ross L (1977) The intuitive psychologist and his shortcomings: distortions in the attribution process. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, vol. 10, pp. 174–221. New York, NY: Academic Press.

Ross L and Nisbett RE (1991) *The Person and the Situation. Perspectives of Social Psychology*. New York, NY: McGraw-Hill.

Shore B (1996) *Culture in Mind: Cognition, Culture, and the Problem of Meaning*. New York, NY: Oxford University Press.

Triandis HC (1995) *Individualism and Collectivism*. Bolder, CO: Westview Press.

Witkin HA and Berry JW (1975) Psychological differentiation in cross-cultural perspective. *Journal of Cross Cultural Psychology* **6**: 4–87.

Further Reading

Ames DR, Knowles ED, Rosati AD *et al*. (2001) The social folk theorist: insights from social and cultural psychology on the contents and contexts of folk theorizing. In: Malle BF, Moses LJ and Baldwin DA (eds) *Intentions and Intentionality: Foundations of Social Cognition*, pp. 307–329. Cambridge, MA: MIT Press.

Choi I, Nisbett RE and Norenzayan A (1999) Causal attribution across cultures: variation and universality. *Psychological Bulletin* **125**: 47–63.

Gilbert DT (1998) Ordinary personology. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*, 4th edn, vol. 2, pp. 89–150. New York, NY: McGraw-Hill.

Jones EE (1990) *Interpersonal Perception*. New York, NY: Freeman.

Krull DS (2001) On partitioning the fundamental attribution error: dispositionalism and the correspondence bias. In: Moskowitz G (ed.) *Cognitive Social Psychology*, pp. 211–227. Mahwah, NJ: Lawrence Erlbaum Associates.

Morris MW, Ames DR and Knowles E (2001) What we theorize when we theorize that we theorize: examining the 'Implicit Theory' construct from a cross-disciplinary perspective. In: Moskowitz G (ed.) *Cognitive Social Psychology*, pp. 143–161. Mahwah, NJ: Lawrence Erlbaum Associates.

Rosati AD, Knowles ED, Gopnik A *et al*. (2001) The rocky road from acts to dispositions: insights for attribution theory from developmental research on theories of mind. In: Malle BF, Moses LJ and Baldwin DA (eds) *Intentions and Intentionality: Foundations of Social Cognition*, pp. 287–303. Cambridge, MA: MIT Press.

Cultural Psychology

Introductory article

Janxin Leu, University of Michigan, Ann Arbor, Michigan, USA
Nicole S Berry, University of Michigan, Ann Arbor, Michigan, USA
Lawrence A Hirschfeld, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

Cultural psychology is an interdisciplinary program of research that explores the relationship between individual minds and the complex environments in which they are deployed. The approach focuses on the contribution that content-rich, complex environments – ranging from workplaces to cultural traditions to nation states – make in shaping basic cognitive processes.

INTRODUCTION

Do cognitive theories predict the mental and behavioral processes of humans in all societies and communities? Is a scientific study of cultural variations in meaning, perceiving, thinking, and feeling across groups of people part of cognitive science? If the answer to the first question is 'no', and the second 'yes', how should cognitive science approach cultural differences? Cultural psychology is a loose amalgam of scholarship in psychology, anthropology, and linguistics that explores these questions. In broad terms, cultural psychology makes the empirical claims that (1) there exist substantial group differences in cognition, emotion, and motivation that cannot be understood without the study of cultural life; (2) aspects of cognition are extra-individual and need to be studied at the population level; and (3) differences in the content of thought result in differences in psychological processes in some, but not all, areas of cognition.

Cognitive scientists typically approach cognition by examining the processes that take place in individual minds or brains. Research in this vein has ordinarily assumed that the processes observed are primary. Higher-level phenomena such as categorization, perception, language use, and the capacity for culture are all supposed to be built upon universal, individual, cognitive processes. For these reasons, the content chosen to test cognitive processes is frequently assumed to be transparent. Stated differently, content is not considered to affect the process under examination.

In contrast, a central argument of much cultural psychological research is that content and process mutually constrain cognition. Myriad scholars from Darwin onwards have proposed that the human mind evolved to solve social problems. Therefore, it is not only plausible but probable that the brain is extremely sensitive to social context. Stated in these terms, much of cultural psychological research can be viewed as an attempt to isolate the effects of social context on cognition. Nevertheless, exactly what constitutes 'social context' is still debatable and evident in the three main methods used in cultural psychology: the comparative, situated/distributed, and epidemiological approaches.

The *comparative approach*, or cross-cultural approach, uses experimental methods to compare and contrast groups in their performance on a range of psychological tasks with the objective of demonstrating cultural difference and commonality on some psychological aspect in the lab and field. For example, Richard Nisbett and colleagues have shown differences among North Americans and members of various East Asian societies in their performance on tasks ranging from perception to categorization. Americans tend to perceive and recognize an object in isolation from its surrounding field and to categorize an object using its traits. In contrast, subjects from several East Asian societies tend to perceive and recognize an object in relation to its surrounding field and to categorize it according to its relationship with other objects. Nisbett and colleagues argue that these differences arise out of socialization in the home, school, and community that emphasize individuality and

agency on the one hand, or harmony and compromise on the other.

The *situated/distributed approach* focuses primarily on cognition in specific, typically coordinated, contexts characterized by common sets of artifacts and habitual activities. This approach uses the distribution and locations of activities and their relation to the material environment as the units of analysis and has the advantage of examining how thought emerges as part of a system of local social practices in a continually evolving process of socialization. For example, Jean Lave studied arithmetic activity in grocery shopping using the supermarket as the arena for cognitive activity. She demonstrated that shoppers who made frequent arithmetic errors in formal testing situations did not make similar errors in best-buy problems in the supermarket. In a situated method, context is conceived as a relation between actors and the setting in which they act, as opposed to a singular 'cultural' unit determining behavior. The situated/distributed approach therefore de-emphasizes the independent study of the innate characteristics of individuals and the particular properties of the context in which action occurs. Instead, it focuses on everyday practices, the locus of interaction between a person and the environment as paramount to understanding cognition.

The *epidemiological approach* explores the processes that underlie the distribution of both mental and public representations, especially those representations that are widespread and enduring. A central question in the epidemiological approach is why some representations, such as myths and folk tales, are stable in a population, while others, such as gossip and rumors, are not. Dan Sperber, who developed the approach, argues that cultural beliefs, including apparently irrational ones, that resonate with common sense are more easily learned and remembered and hence are more likely to become widespread and enduring. In an empirical application of this proposal, Lawrence Hirschfeld has shown that underlying, and substantially constraining, historically recent and socially constructed concepts like 'race' and 'ethnicity' is a special-purpose knowledge structure treating input relevant to collectivities in the socio-political environment.

While the comparative, situated/distributed, and epidemiological approaches make transparent how group life and cultural content (i.e., beliefs, attitudes, domains of knowledge) shape psychological process (i.e., thinking, feeling, perceiving), they differ in how they highlight context. For example, the situated/distributed approach demonstrates not only context-specificity in thinking across everyday events, such as problem solving in the classroom versus the supermarket, but flexibility across everyday actions. The locus of motivating cognitive performance is in the dynamic rituals of daily habits. On the other hand, a comparative or cross-cultural approach centers the discussion of cognition-in-context on cognition in groups that have a set of shared, identifiable, and fairly stable practices. Comparative psychologists complain that the situated/distributed approach has lost sight of psychology and individual minds. Situated/distributed cognitive psychologists argue that comparative approaches mistake groups for daily contexts and fail to account for dynamics in cultural systems. Advocates of the epidemiological approach, in turn, caution that psychologists working in the comparative tradition often overestimate the coherence, stability, and boundedness of culture, hence misattributing causal properties to a fairly fluid environment. They similarly observe that the situated/distributed approach, while intriguing, has no principled way to delimit the context in which a task is situated or distributed.

COMPARATIVE STUDIES OF COGNITION

Language

Cultural psychologists are not the first to suggest that language, culture, and cognition are inextricably linked. Benjamin Lee Whorf and Edward Sapir each proposed, in what became known as the Whorf–Sapir hypothesis, that the cognition of a member of a particular culture is constrained by what can be said in the language the person speaks. For example, Whorf famously proposed that the Hopi have 'no general notion or intuition of TIME as a smooth flowing continuum … [because the Hopi language contains] no words, grammatical forms, constructions or expressions that refer directly to what we call "time," or to past, present, or future'. While this strong version of linguistic relativity is no longer seriously entertained, a number of scholars have explored more modest versions.

Counting

J. A. Lucy, in studies involving English and Yucatec Mayan speakers, has demonstrated that variations in grammatical form governing counting can influence categorization. Yucatec requires the speaker to use both numeric modifiers and classifiers when counting: whereas an English speaker counts, 'one

candle, two candles', a Yucatec speaker counts, 'one, long-thin wax, two, long-thin wax'. Lucy contends that this grammatical difference accounts for differences in English and Yucatec Mayan-speaking participants' similarity judgments when asked to put together objects which could be sorted either by shape or substance. Yucatec speakers grouped substances together based on their construction from like substances, while English speakers grouped objects together based on shape.

Noun extension

Linguistic influences on the nature of categories were similarly demonstrated among English and Japanese speakers: children aged two, two-and-a-half, and four, and adults. As in Yucatec, nouns are not pluralized in Japanese. In an experimental procedure similar to the one described above, American children as early as two years of age extended a noun label to other objects based on shape, whereas Japanese children extended the label based on substance.

Early differences in language socialization and pragmatics may underlie linguistic variation in categorization. For example, Fernald and Morikawa compared Japanese and American mother–infant interactions at home in a study of lexical development. Infants were preverbal: aged six, 12, and 19 months. They found that American mothers labeled objects more often than Japanese mothers. While American mothers reported trying to direct their children's attention to objects and teaching them the proper noun labels, Japanese mothers reported emphasizing polite exchanges of the objects with the infant in fostering social exchange. Comparisons of American mothers with Korean and Chinese provide further evidence of socialization practices that differentially emphasize target objects and noun labels on the one hand, and social interactions and substance on the other.

Categorization

There has been a long tradition in both psychology and anthropology of trying to understand the origins of the categories humans use. Anthropologists long assumed that the systematic variation in lexicalized and covert categories reflected systematic variations in cognition. Cognitive psychologists long assumed that the representations of and computations over contrived unfamiliar categories reflected universal cognitive processes. Pioneering studies by Brent Berlin and Paul Kay, in anthropology, and Elinor Rosch, in psychology, refuted both assumptions.

Previous to Berlin and Kay's research, anthropologists assumed that each culture arbitrarily divided and named different colors on the color spectrum. After surveying color terms from over 80 different languages, both universal and culture-specific patterns were observed. Not all languages identified the same number of basic color terms. The number of color terms in any given language, however, was found to be a function of a closely organized system (in a comprehensive survey of languages Berlin and Kay found that only 11 of over 2000 combinations of basic color terms are actually used). Importantly, although Berlin and Kay found that there was great variation in the boundaries of color terms across (and within) languages, speakers of all languages converged on the same focal colors which best represented the color category.

Eleanor Rosch demonstrated that natural language categories are organized differently from contrived categories. As Berlin and Kay revealed for color terms, natural language categories are internally structured such that (a) some category members are better examples of the category than others, and (b) these prototypical members are linked to other category memberships by a relation of family resemblance, not common necessary and sufficient features.

Recently, it has been demonstrated that exemplar use can also vary across cultures, as can the principles around which objects are organized. For example, in one study, Chinese and American college students were asked to judge similarity among objects in which grouping by category (e.g., notebook and magazine) or relationship (e.g., pencil and notebook) was possible. Chinese were more likely to group objects by relationship and to justify their decision by invoking relationships, whereas Americans were more likely to group objects by shared categories and to refer to category membership as the explanation.

Reasoning

Cultural psychologists have demonstrated group differences in inductive and deductive reasoning. These differences have been explained as differences in expertise, as well as differences in intellectual traditions and resulting styles of reasoning.

Inductive reasoning

Scott Atran and Douglas Medin have carried out experiments comparing conceptualization of and reasoning about living beings across different cultures to demonstrate both universals and cultural

specificity in inductive reasoning. In support of universal principles of folk biological categorization, Atran and Medin found that undergraduates from two Midwestern American universities and Itzaj Maya of Guatemala grouped different animals in ways that reliably correlated with scientific taxonomies. These groupings, organized around taxonomic rank, strongly influenced reasoning: across both groups, one taxonomic level (the folk-generic) was found to be the most inferentially rich despite differences in which level is most 'psychologically' basic in the sense identified by Rosch and her colleagues.

Having shown commonalties in the conceptual domain between the undergraduates and Itzaj Maya, Atran and Medin also examined specificity in category-based induction about living beings. For example, Atran and Medin asked American students and Itzaj Maya to judge which of two arguments is stronger: (1) all mammals will share a trait if it is known that two diverse mammals (e.g., a mouse and a jaguar) share the trait, or (2) all mammals will share a trait is if is known that two similar mammals (e.g., a jaguar and a leopard) share the trait. American students reasoned that (1) is the stronger argument, whereas Itzaj Maya reasoned that (2) is the stronger argument. Atran and Medin speculated that the Itzaj based their judgments on what they knew about the behavior and ecology of the species in the argument rather than reasoning from taxonomic positions. Choi, Nisbett, and Smith also found a cultural effect in inductions based on diversity. Koreans are less likely than their American counterparts to feel that increasing diversity of two given examples strengthens inductions about the higher category of which both examples are members. Nevertheless, when the higher category is made more salient to the Korean subjects through a manipulation, they perform the same as their American counterparts.

Deductive reasoning

Research also demonstrates cultural differences in how people use formal logic versus experiential knowledge in reasoning based on typicality. Sloman demonstrated that the typicality of an exemplar (e.g., eagles versus penguins as birds) influences the persuasiveness of deductive arguments for all subjects (a claim about all birds). Norenzaya found that the effect is stronger among Koreans than European Americans, with Asian American scores falling between the two populations. Further, Koreans, unlike their American counterparts, were less likely to dismiss conclusions as implausible based solely on logic. Rather, they more frequently used empirically derived beliefs about the world than logic to judge the plausibility of the conclusions.

Kaiping Peng and Richard Nisbett suggest that unlike Americans, subjects from several East Asian societies tolerate contradiction and are not as concerned with demands of consistency. Chinese and American students were compared in their preference for argument form and judgments about contradictory propositions. Whereas the Americans preferred logical arguments (i.e., God exists because there had to be a first cause), the Chinese preferred 'dialectical' arguments (i.e., God exists because there must be a truth that rises above all individual perspectives). Likewise, whereas the Americans rejected one proposition in face of a contradicting proposition, the Chinese embraced both propositions, finding merit in the middle ground.

Rather than concluding that individuals from these East Asian societies apply faulty logic, this work suggests that distinct, fully integrated systems of reasoning may exist among different groups. Nisbett and colleagues argue that American participants tend to prefer abstract logic over complex experience in highlighting how the world works. They stress the parts of a system that are coherent and noncontradictory, marginalizing and trying to resolve inconsistencies. It is argued that East Asian participants, on the other hand, tend to rely on empirical experience over logical models to understand their environment, which they view as constantly changing and inextricably interconnected. Within this framework, contradictions and inconsistencies are prevalent, natural, and even necessary.

Moral reasoning

Although a topic that has not traditionally attracted great attention in the cognitive sciences, the interest in evolutionary psychology has brought issues of morality to the forefront. This interest flows from morality's role in explanations of adaptations to social life – including the necessity to attend to and conceptually represent increasingly larger, more disperse, and more internally complex social groupings and the need to track exchange and acts of altruism. Morality is particularly salient in setting out, stabilizing, and justifying sanctions that underlie systems of complex social control.

Developmental psychologists have identified correlates of this salience in children's early reasoning, specifically about the differences between transgressions of social conventions and

moral transgressions. Turiel and his colleagues have proposed that children spontaneously employ a universal moral appraisal involving expectations about harm and welfare. They provide experimental evidence that children everywhere discriminate among social conventions – which they believe vary readily from culture to culture and situation to situation, from moral transgressions – which they identify as harming others and believe to be wrong under any system of thought or in any situation.

For example, Smetana *et al*. presented American preschool children with accounts of both hypothetical and real transgressions. The children were asked to rate how 'bad' each was, whether it would be as equally bad if it occurred elsewhere, and whether it deserved punishment. Some were transgressions of social convention ('this child is not saying "please" when asking for something'); others were moral transgressions ('this child is hitting this child'). Even three-year-olds judged moral transgressions as more serious, more generalizable, and more deserving of punishment than social conventional transgressions. This pattern of reasoning appears to be robust across cultures and across individuals who personally suffered harm and violations of welfare (e.g., among abused, neglected, and maltreated children).

Other cognitive scientists have challenged this view, arguing that moral appraisals are culturally specific. Against Turiel's universalist position, Shweder and his colleagues contend that the distinction between morality and social convention is not universal and that moral appraisal is contingent on a particularly socio-historical context. For instance, when they asked Brahman and American children to rate how harmful various acts were, they found significant disagreement across the two populations. American children, but not Brahman children, accepted that it is wrong to cane a naughty child or eat with one's hands; whereas Brahman children, but not American children, deemed that eating beef or chicken or cutting hair after one's father's death is wrong. In short, in this view content is a function of context. Unlike other areas of reasoning, whether or not content affects the cognitive process is not addressed.

Spatial Cognition

Beginning with the last two decades of the twentieth century, a body of literature exploring the effects of language and culture on spatial reasoning has emerged. Several ways of identifying an object's relation in space to other objects have been identified: reference can be made with respect to the cardinal directions (the bus is to the south of the car), with respect to the position of the speaker (the bus is to the right of the car), and with respect to fixed geographical features (the bus is uphill from the car). Not all languages provide speakers with as ready means to express these relationships; nor do speakers of different languages employ all strategies with the same frequency. For example Steven Levinson and his colleagues found that Guugu Yimmithirr aboriginals are more likely to anchor an array of objects by reference to cardinality than speakers of Dutch, who are more likely to describe spatial arrangement with reference to their own body.

Attention and Visual Perception

All humans appear to be susceptible to most visual illusions. A well-studied exception is the Müller–Lyer illusion, which Herkovits *et al*. found did not affect the perception of members of some African societies. The researchers explained this by citing the absence of a susceptibility to the 'carpentered-world hypothesis'. On this hypothesis, susceptibility to the illusion is a function of experiencing a material world in which 90-degree angles are common (i.e., 'the carpentered world'), but in the African societies studied round shape architecture prevails.

In a study of change blindness, Masuda and Nisbett showed animations of scenes, which included stationary and moving objects, to Japanese and Americans; some animations were altered in later displays. The Japanese were more aware of changes in the background of a picture and in the changing location of objects in it, whereas the Americans were more aware of changes in the characteristics of objects.

Developmental evidence of group differences in attention suggests that differences can be learned. Chavajay and Rogoff demonstrated differences between Guatemalan Mayan and American 14–20 month-olds and their caregivers in attention management. Mayan caregivers and toddlers were more likely to attend simultaneously to competing events, unlike American caregivers and toddlers who were more likely to alternate their attention between competing events.

SITUATED/DISTRIBUTED STUDIES OF COGNITION

Cognitive scientists have typically examined how individual humans or their machine counterparts

process information, traditionally through controlled investigations and laboratory studies. A number of cultural psychologists caution, however, that cognition cannot be fully understood as (1) taking place in individual minds, or (2) apparent place under controlled conditions. On the contrary, researchers who use the situated/distributed approach propose that the interactions between individual minds in their natural environments are mediated by extra-individual cognitive phenomena. Reflecting the influence of Lev Vygotsky's notion of 'the zone of proximal development' (and its variant, scaffolding), considerable work in this approach is developmental. According to Vygotsky, novices gain expertise by participating in contexts that extend their own skills sufficiently to boost performance and to stabilize enhanced competencies rather than acquiring skills through formal teaching.

Edwin Hutchins and others have proposed that cognition may be distributed among individuals in ways that argue for the existence of extra-individual cognitive systems. These systems are comprised of the habitual interactions between individuals and artifacts and, because of their emergent quality, cannot be understood from the perspective of the cognitions of the individuals participating. For example, Hutchins proposes that the speed of an airplane is remembered by the cockpit, arguing that the task is distributed over each of the crew and each of the instruments and manuals that they use, so that the cockpit itself is the level at which the cognitive task is being accomplished.

In a study of situated cognition, Geoffrey Saxe compared the mathematical competencies of three different groups of Brazilian children: urban street vendors, urban nonvendors, and rural nonvendors. None had received any formal schooling, and all three groups did poorly on abstract, orthographic tasks involving large number representation. When essentially the same tasks were repeated using large bills of money, the performance of urban children, who regularly deal with currency exchange, significantly improved, but the performances of both the rural children or nonvender urban children did not.

THE EPIDEMIOLOGICAL APPROACH

Both the comparative and situated/distributed approaches seek to demonstrate that culturally varying content affects underlying cognitive processes. The epidemiological approach, pioneered by Dan Sperber, in contrast understands cultural variation as a function of underlying and invariant cognitive processes. In particular, the approach focuses on the conditions by which beliefs and practices become distributed in populations, much as medical epidemiology focuses on how disease and pathogens are distributed.

Richard Dawkins proposed a similar notion, which he calls the 'meme', the cultural evolutionary analogue to a gene. Dawkins proposes that memes, like genes, replicate. He further suggests that the psychological processes underlying the reproduction of memes, specifically imitation, is fairly low level. In contrast, Sperber argues that cultural reproduction involves both ecological and higher-order psychological processes that in interaction determine the distribution of cultural phenomena in a given population. For Sperber to account for culture 'is to explain why some representations become widely distributed … to explain why some representations are more successful – more contagious – than others'.

The epidemiological approach, accordingly, is closely linked with specific claims about cognitive architecture. The predominant view holds that humans are endowed with a general set of reasoning abilities that are brought to bear on a myriad of cognitive tasks. In the latter decades of the twentieth century researchers in psychology, linguistics, anthropology, and philosophy concluded that many cognitive abilities are specialized to handle specific types of information: the mind is composed of domain-specific or modular devices that, among other things, underlie much of common sense. For example, all normally endowed humans have a folk or naive theory of mind that interprets human behavior to be the result of mental states like belief and desire. Similarly, all humans have a naive physics that interprets the movement of inanimate objects. Those representations whose processing are subsumed by these various modularized dimensions of common sense are, according to Sperber, more likely to become widespread and relatively stable. Culture, then, is the totality of representations with these particular properties of distribution.

Building on the model of symbolism developed by Sperber, Pascal Boyer argues that supernatural entities are easily accepted and remembered not because they can be subsumed under common sense, but because they achieve a balance between commonsense expectations and violations of commonsense expectations. For example, Barrett and Keil have shown that folk notions of the Judeo-Christian concept of God tend to make God more humanlike (more consistent with commonsense expectations about humans), de-emphasizing aspects

of formal theological beliefs that attribute omnipotence *and* omnipresence to God.

CONCLUDING REMARKS

Cultural psychologists and their counterparts in anthropology, cognitive anthropologists, are committed to developing an interdisciplinary program of research that explores the relationship between individuals' minds and the complex environments in which they exist. Much cultural psychological research is now focusing on topics that have tended to receive less attention in the cognitive sciences (e.g., the self and emotions). Cultural psychology is contributing significant insights into higher-order cognitive processes of central concern to cognitive science. Cultural psychology challenges the widely accepted assumption in cognitive science that the relationship between content and process is transparent, that the environment 'merely' provides content for cognitive processes. Instead, cultural psychology proposes that cognition must be understood in relation to population-level phenomena, ranging from local task-specific environments such as work groups to global, comprehensive domains such as cultures.

Further Reading

Barrett J and Keil FC (1996) Conceptualizing a nonnatural entity: anthropomorphism in God concepts. *Cognitive Psychology* **31**: 219–247.

Berlin B and Kay P (1969) *Basic Color Terms; Their Universality and Evolution*. Berkeley: University of California Press.

Boyer P (1994) *The Naturalness of Religious Ideas: A Cognitive Theory of Religion*. Berkeley: University of California Press.

Cole M (1996) *Cultural Psychology: A Once and Future Discipline*. Cambridge, MA: Harvard University Press.

Coley JD, Medin DL, Proffitt JB *et al*. (1999) Inductive reasoning in folkbiological thought. In: Medin DL (ed.) *Folkbiology*, pp. 205–232.Cambridge, MA: MIT Press.

Fernald A and Morikawa H (1993) Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development* **64**: 637–656.

Herskovits M, Campbell D and Segall M (1969) *A Cross-cultural Study of Perception*. Indianapolis: Bobbs-Merrill.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Lave J (1988) *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge, MA: Harvard University Press.

Levinson SC (1996) Language and space. *Annual Review of Anthropology* **25**: 353–382.

Lucy JA (1992) *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge, New York: Cambridge University Press.

Markus HR and Kitayama S (1991) Culture and the self: implications for cognition, emotion, and motivation. *Psychological Review* **98**(2): 224–253.

Nisbett RE, Peng K, Choi I and Norenzayan A (2001) Culture and systems of thought: holistic vs. analytic cognition. *Psychological Review* **108**: 291–310.

Rogoff B and Lave J (1984) *Everyday Cognition: Its Development in Social Context*. Cambridge, MA: Harvard University Press.

Saxe G (1988) The mathematics of child street vendors. *Child Development* **59**: 1415–1425.

Shweder R, Mahapatra M and Miller J (1987) Culture and moral development. In: Kagan J and Lamb S. *The Emergence of Morality in Young Children*, pp. 1–83. Chicago, IL: University of Chicago Press.

Sperber D (1996) *Explaining Culture: A Naturalistic Approach*. Oxford, UK: Blackwell.

Turiel E (1983) *The Development of Social Knowledge: Morality and Convention*. Cambridge, UK: Cambridge University Press.

Vygotsky LS (1978) *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Culture and Cognitive Development

Intermediate article

Michael Cole, University of California at San Diego, La Jolla, California, USA

CONTENTS

Although it is widely believed that culture is fundamental to cognitive development, theoretical and empirical advances are impeded by severe methodological problems which render firm conclusions elusive.

INTRODUCTION

Controversies over the relationship between culture and cognition began well before the formation of contemporary behavioral sciences, and continue to this day (Cole, 1996). Discussion of the topic was, and remains, complicated by confusion over each of the basic terms under discussion as well as by the severe limitations placed on the experimental method when it is used to deal with naturally occurring differences among human beings involving unknown mixtures of phylogenetic, cultural-historical and ontogenetic variations – all of which are involved in seeking to understand the relation of culture to cognitive development.

It is useful to begin by tracing the concept of culture as it has evolved since the term entered the English language from Latin many centuries ago. Modern conceptions of culture originate in terms that refer to the process of helping things to grow. From earliest times, this notion of culture included a general theory of how to promote growth: the creation of an artificial environment in which young organisms would have optimal conditions to develop. Such tending requires tools: both material (e.g. hoes) and mental (e.g. the knowledge of how and when to use a hoe). These tools were perfected over generations and designed for the special tasks to which they were put, and constitute culture in the present.

In contemporary social science writing, the term 'culture' is generally used to refer to the entire body of socially inherited past human accomplishments that serves as the resource for the current life of a social group ordinarily thought of as the inhabitants of a country or region. Although there is evidence of the rudiments of culture in nonhuman species, human beings are unique in their dependence upon the medium of culture and the forms of organism–environment interactions that culture supports to sustain and reproduce themselves.

Combining the historical notion of culture as a process of growing things with the modern conception of culture as social inheritance of prior generations' accomplishments, the study of culture in development can be seen to focus on the way in which biologically immature human beings incorporate and are incorporated into the cultural 'designs for living' that are their social heritage.

The interpenetration of cultural and phylogenetic contributions to human development was driven home several decades ago by Clifford Geertz, who noted the mounting evidence that the human body, and most especially the human brain, underwent a long (perhaps 3 million year) coevolution with the basic ability to create and use culture, and was led to conclude that

> man's nervous system does not merely enable him to acquire culture, it positively demands that he do so if it is going to function at all. Rather than culture acting only to supplement, develop, and extend organically based capacities logically and genetically prior to it, it would seem to be ingredient to those capacities themselves. A cultureless human being would probably turn out to be not an intrinsically talented, though unfulfilled ape, but a wholly mindless and consequently unworkable monstrosity (Geertz, 1973: p. 68).

As a consequence, those who wish to understand the role of culture in cognitive development are faced with the difficult interdisciplinary task of studying development in terms of antecedents that are tightly interwoven.

METHODOLOGICAL PROBLEMS

Culture-free versus Culture-based Measures of Cognitive Development

For most of the history of scholarly interest in the role of culture in development, research been based upon 'cross-cultural' comparisons. This phrase is placed in quotation marks because often the comparisons are cross-national in nature, involving social groups thought to differ in some theoretically interesting way (for example, with respect to language, natural ecology, or social institutions) such that cultural, biological, social and ecological factors all differ simultaneously.

The hazards of restricting such comparisons to only two naturally occurring groups has long been recognized. As a consequence, leading researchers into culture and cognitive development have routinely included a range of societies in their studies to reduce the risk that some factor other than the one under investigation is covertly influencing the results. However, such multisociety research is expensive, and also carries with it the difficulty that although it might reduce the probability of undetected covarying factors it does not eliminate it completely, often leading to such an apparently endless set of possibilities that definitive conclusions elude further research.

The problem of culturally valid cognitive assessment has been most extensively discussed with respect to the possibility of creating culture-free measures of cognition, seemingly a prerequisite for valid cross-cultural comparisons (the same issue applies, but is rarely addressed, when age comparisons within a single homogeneous group are of interest). A variety of cognitive measures used in cross-cultural comparisons have been the object of such analyses. Research using intelligence quotient (IQ) tests and Piagetian tests of conservation can serve as accessible examples.

Long ago, Florence Goodenough identified the crucial shortcoming of the cross-cultural use of IQ tests in a way that has broad – if rarely recognized – implications for all cross-cultural cognitive research, when she wrote that 'the fact can hardly be too strongly emphasized that neither intelligence tests nor the so-called tests of personality and character are measuring devices. They are sampling devices'.

Goodenough argued that when applied in American society, IQ tests may represent a reasonable sampling device because they are 'representative samples of the kind of intellectual tasks that American city dwellers are likely to be called upon to perform'. However, such tests are not representative of life in other cultural circumstances and hence their use as measuring devices for purposes of comparison is inappropriate. This injunction applies as much to variations among subgroups living in the USA as it does to people living in a wide variety of other societies.

It might be thought, on the basis of Goodenough's critique, that attempts to discover the influence of culture on the development of intelligence using IQ tests would have been abandoned long ago. However, the usefulness and theoretical implications of cross-cultural IQ testing continue to be heatedly debated. Logically, the only way to obtain a culture-free test is to construct items and procedures that are equally a part of the experience of all cultures. Following Goodenough's approach, this would require us to sample the valued adult activities in all cultures (or at least two!) and identify activities equivalent in their structure, their valuation and their frequency of occurrence. No one has carried out such a research program.

The same problem can be seen in the widespread use of Piagetian conservation tasks as measures of cognitive development during the late decades of the twentieth century. It was Piaget's initial hypothesis that the development of conservation would be a universal achievement, occurring in an invariant sequence at roughly equal ages across cultures because it represents a universally applicable logical principle (Piaget, 1974). However, application of standardized Piagetian procedures in different societies produced widely varying results, leading some to speculate not only that some cultures promote more cognitive development than others, but that without particular kinds of cultural experience, such as formal schooling, development might cease at the level of concrete operations.

However, this same literature contained within it anomalies that implicated cultural differences in interpretation of the task – not differences in logical development – as the source of cultural differences. The standard procedure, following methods developed in Geneva, was to present the participant with two beakers of equal circumference and height, filled with equal amounts of water. The water from one beaker was then poured into another, taller beaker with a smaller circumference, and the person was asked which of the two beakers containing water had the most water in it. Children and young adults from nonliterate societies of a given age were significantly more likely to assert that the taller, thinner beaker, contained more water; see Cole (1996) for a summary. However, an experiment on the effects of modifying testing

procedures to match local cultural knowledge revealed a different pattern of results. Early research had shown that conservation was much more likely to be achieved if children were allowed to pour the liquid themselves instead of observing the experimenter. It was speculated that this change in procedure reduced the children's tendency to interpret the experiment as something of a magic show, allowing their real competence concerning the conservation of volume to be revealed.

In the revised procedure, participants were asked to solve the conservation task and then to act as informants whose job it was to clarify, for the experimenter, the local terms for resemblance and equivalence with respect to the task. When confronted with the critical test in which one beaker of water was poured into a narrower, taller beaker, the participants asked to play the role of informant gave the wrong response – they said that the beaker with water higher up its sides contained more liquid. However, in the role of linguistic informants these same people went on to explain that while the level of water was 'more', the quantity was the same. Such results provide nice examples of the kinds of performance factors that can block the actualization of competence and mislead researchers about the nature of cultural differences.

THEORIES OF CULTURAL CONTRIBUTIONS TO COGNITIVE DEVELOPMENT

Two major lines of theory have dominated discussions of culture and cognitive development, those of Jean Piaget and Lev Vygotsky. (*See* **Piaget, Jean**; **Vygotsky, Lev**)

Jean Piaget

According to Piaget, cognitive development occurs through a process of equilibration where accommodation to existing environmental circumstances constantly vies with assimilation of environmental structures to existing mental structures. In a widely cited article, Piaget (1974) divided potential influences on cognitive development into four main factors, each of which could be expected to influence the timing of developmental milestones resulting from the interplay of assimilation and accommodation.

- Biological factors: here Piaget mentions nutrition and general health, factors that influence what might be called the rate of physical maturation.
- Coordination of individual actions: this factor refers to equilibration, the active process of self-regulation resulting from the back-and-forth tug and pull of accommodation and assimilation. Equilibration is the proximal mechanism of development. All other factors operate through their influence on equilibration.
- The social factor of interpersonal coordination: the process by which children 'ask questions, share information, work together, argue, object, etc.' (Piaget, 1974: p. 302).
- Educational and cultural transmission. Piaget reasoned that children acquire specific skills and knowledge through interaction in culturally specific social institutions. In so far as some societies provide more overall experience relevant to discovering the nature of the world, true developmental differences would be created in either the rate or the final level of development.

Piaget recognized that the four contributing factors he identified are all tightly interconnected with each other in any given society. However, barring conditions of extreme malnutrition, he assumed that rates and levels of cognitive development would be universal across societies with respect to the first three factors, and would differ only with respect to the fourth. In early writings, he assumed that modern industrial societies would provide the requisite additional experiences to speed cognitive development, but it is interesting that Piaget was skeptical of schooling's development-enhancing properties, since this social institution varies greatly across cultures and in popular thinking ought to have a major influence on cognitive development. He argued that the asymmetrical power relations of teacher and student created an imbalance of equilibration, because the pressure to accommodate to teachers' views far outweighed the pressure for assimilation of instruction to the child's already existing schemas. The result was learning of a superficial kind that was unlikely to create fundamental cognitive change. He believed that fundamental change was more likely to occur in informal actions where the asymmetry of power relations was reduced, allowing for a more equal balance between assimilation and accommodation. Evidence on this matter was too inconclusive in the 1960s to support more than speculation.

Confronted by subsequent research with evidence of cultural variations in performance on his tasks, and particularly evidence that schooling enhanced performance within cultures, Piaget (1972) concluded that all individuals reach a universal stage of formal operational thinking, but that formal operations are attained first (and perhaps only) in fields of adult specialization or as a consequence of school-based training. This view offers an obvious line of reconciliation of Piagetian theory with the facts about cultural variability. The overall

conclusion, however, accords a restricted role to culture in cognitive development, which is more a matter of individual invention through action on the environment than of environmental, and particularly cultural, influences, on the child.

Lev Vygotsky

The Russian psychologist Lev Vygotsky, unlike Piaget, accorded culture and the influence of children's social environments a central role in cognitive development (Vygotsky, 1978). The central thesis upon which his 'cultural-historical' school of psychology was founded is that the structure and development of human psychological processes emerge in the process of humanity's culturally mediated, historically developing, practical activity. Each term in this formulation is tightly interconnected with and in some sense implies the others, making it difficult at times to discern how each contributes to description and analysis of the dynamics of psychological experience.

1. Mediation through artifacts. The initial premise of the cultural-historical school is that the psychological processes of humans emerged simultaneously with a new form of behavior in which material objects were modified by human beings as a means of regulating their interactions with the world and each other.
2. Historical development. In addition to using and making cultural artifacts, human beings arrange for the rediscovery of the already created artifacts in each succeeding generation, which in turn adds its modifications to the culture pool of artifacts. The accumulated artifacts of a group – culture – is then seen as the species-specific medium of human development. Cognitive development, as a consequence, represents the capacity to develop within that medium and to arrange for its reproduction in succeeding generations.
3. Practical activity. The third premise of the cultural-historical approach, adopted from Hegel by way of Marx, is that the analysis of human psychological functions must be grounded in their everyday activities. It was only through such an approach, Marx claimed, that the duality of materialism versus idealism could be superseded, because it is in activity that people experience the ideal/material residue of the activity of prior generations.

Vygotsky's emphasis on the historical accumulation of artifacts and their infusion in activity led him to emphasize the social origins of human thought processes. Vygotsky argued it is only through the mediation of socialized others that the child can come to experience, and hence acquire, the cultural heritage which is the foundation of cognitive development. This view of social origins requires that special attention be paid to the power of adults to arrange the environments of children in a way that optimizes their development according to existing norms. It generates the idea of a 'zone of proximal development' which affords the proximal, relevant environment of experience for development.

CONTEMPORARY RESEARCH ON CULTURE AND COGNITIVE DEVELOPMENT

Because of the methodological problems inherent in cross-cultural approaches to the study of development, the pace of such work has now slackened. At the same time, there has been an increase in research within cultures that take advantage of naturally occurring contrasts associated with differential participation in particular cultural practices in order to show how culture enters into the process of development. This does not mean that cross-cultural research has come to a standstill. For example, faced with the ambiguities inherent in research carried out in different cultures, where differences in language, nutrition and social relations can so easily enter unbidden into comparisons that seek to highlight specific cultural factors, some researchers interested in studying the relation of culture to development have turned to studies within societies that greatly reduce the chances of undetected contamination by uncontrolled factors. One such method is referred to as the 'school cut-off strategy' (Christian *et al*., 2001).

In many countries school boards require that in order to begin attending school, a child be a certain age by a particular date. For example, to enter grade 1 in September of a given year, children in Edmonton in Canada must have passed their sixth birthday by March 1 of that year. Six-year-olds born after that date must attend kindergarten instead, so their formal education is delayed for a year. Such policies allow researchers to assess the impact of early schooling on different domains of cognitive development while holding age virtually constant: they simply compare the intellectual performances of children who turn six in January or February with those who turn six in March or April, testing both groups at the beginning and at the end of the school year.

Researchers who have used this strategy find that the first year of schooling brings about a marked increase in the sophistication of some cognitive processes but not others. Frederick Morrison and his colleagues, for example, compared the ability of children to perform on tasks

ranging from picture recall to word and number manipulation, and Piagetian number conservation tasks (Christian *et al*., 2001). The first-grade children were, on average, only a month older than those in kindergarten, and at the start of the school year the performances of the two groups were virtually identical. At the end of the school year, however, the first-graders could remember twice as many pictures as they did at the beginning, whereas the kindergarten group showed no improvement at all. Significantly, the first-grade children engaged in active rehearsal during the testing, but the kindergarten children did not. Clearly, a year of schooling had brought about marked changes in memory strategies and performance. The same pattern of results was obtained for standardized reading and mathematics tests. However, similar differences were not obtained for a standard Piagetian test of number conservation. This evidence shows clear influences of a cultural factor, schooling, on cognitive development, while supporting Piaget's skepticism about the cognitive benefits of schooling.

Meanwhile, cross-cultural research with respect to a variety of cognitive functions including categorization, logical reasoning and memory continues to show significant effects. In some cases schooling appears to be critical to the results; in others it does not. For example, Alexander Luria (described in Cole, 1996) compared the performance of Central Asian peasants who had been organized into collective farms and exposed to a modicum of Western-style schooling with those who had not. He reported that categorization processes grounded in 'graphic, object-oriented experience' gave way to 'more complex processes which combine what is perceived into a system of abstract, linguistic, categories.' Thought processes grounded in 'practical, situational' thinking gave way to more abstract theory-driven modes of thinking, which Luria referred to as the 'the transition from the sensory to the rational'. While Luria attributed such findings to involvement in modern industrialized practices including schooling, more recent research into cultural influence on the development of categorization has shown that even among people without schooling, both density of experience and cultural variations in world views substantially influence the development of biological thinking in addition to physical similarity (Medin and Atran, 1999). This latter work fits well the idea that cultural variations build upon panhuman, phylogenetically derived perceptual processes to construct local theories appropriate to local circumstances.

A considerable body of work has been devoted to the study of cultural variations in syllogistic reasoning, one of the hallmarks of Piaget's stage of formal operations. According to Luria's data, as well as research conducted in other nonliterate societies in different parts of the world, syllogistic reasoning appears to be closely linked to the sociocultural institutions of Western-style schooling. However, cases where nonliterate peasant workers succeed in responding to formal syllogisms, as well as cases where college-educated adults do not, have been reported.

The tradition of studying the role of participation in culturally organized activities within a single culture has emphasized the many ways in which cultural practices beyond schooling enter into the process of cognitive development. Here we can mention research on the development of mathematical thinking among street-vendor children in Brazil as well as work among Nepalese youth who become engaged in new forms of economic activity (Saxe, 1994; Beach, 1995; Ueno, 1995).

CONCLUSION

Taken as a whole, research on culture and cognitive development persuasively demonstrates the centrality of culture and cultural variation to cognitive development. The current challenge facing researchers is to solve the methodological difficulties (which themselves involve cultural factors) in order to identify more clearly the necessary and sufficient conditions involved.

References

Beach K (1995) Activity as a mediator of sociocultural change and individual development. *Mind, Culture and Activity* **2**(4): 285–302.

Christian K, Bachnan HJ and Morrison FJ (2001) Schooling and cognitive development. In: Sternberg RJ and Grigorenko EL (eds) *Environmental Effects on Cognitive Abilities*. Mahway, NJ: Erlbaum.

Cole M (1996) *Cultural Psychology: A Once and Future Discipline*. Cambridge, MA: Harvard University Press.

Geertz C (1973) *The Interpretation of Cultures: Selected Essays*. New York, NY: Basic Books.

Medin DL and Atran S (eds) (1999) *Folkbiology*. Cambridge, MA: MIT Press.

Piaget J (1974) Need and significance of cross-cultural studies in genetic psychology. In: Berry JW and Dasen PR (eds) *Culture and Cognition: Readings in Cross-cultural Psychology*. London: Methuen.

Piaget J (1972) Intellectual evolution from adolescence to adulthood. *Human Development* **15**: 1–12.

Saxe G (1994) Studying cognitive developments in sociocultural context: the development of a

practice-based approach. *Mind, Culture and Activity* **1**: 135–157.
Ueno N (1995) The social construction of reality in the artifacts of numeracy for distribution and exchange in a Nepalese bazaar. *Mind, Culture and Activity* **2**(4): 240–257.
Vygotsky LS (1978) *Mind in Society*. Cambridge, MA: Harvard University Press.

Further Reading

Berry J, Poortinga Y, Segall M and Dasen P (1992) *Cross-cultural Psychology: Research and Applications*, 2nd edn. Cambridge, UK: Cambridge University Press.
Greenfield PM (1976) Cross-cultural Piagetian research: paradox and progress. In: Riegel KF and Meacham JA (eds) *The Developing Individual in a Changing World: Historical and Cultural Issues* vol. 1. Chicago: Aldine.
Hallpike CP (1979) *The Foundations of Primitive Thought*. Oxford: Clarendon Press.
Irvine J (1978) Wolof 'Magical thinking: culture and conservation revisited.' *Journal of Cross-cultural Psychology* **9**: 300–310.
Jahoda G (1993) *Crossroads Between Culture and Mind*. Cambridge, MA: Harvard University Press.
Luria AR (1976) *Culture and Cognitive Development*. Cambridge, MA: Harvard University Press.

Decision-making

Introductory article

Barbara A Mellers, University of California, Berkeley, California, USA

CONTENTS

The field of decision-making focuses on normative questions – how should we make decisions if we want to do it right? – and description questions – how do we actually make decisions when dealing with uncertainties? Numerous demonstrations of how people violate the axioms of normative theory have given rise to a deeper understanding of the processes underlying actual decision-making.

INTRODUCTION

Expected utility theory is widely accepted as the standard for normative decision-making. It tells us how we 'should' make decisions if we want our choices to be internally consistent and coherent. However, we do not always follow those principles. Researchers in decision-making have constructed puzzles and paradoxes that show how people violate the axioms of expected utility theory. People focus on many other factors, including personal, social and cultural rules, justifiability, and emotional rewards. Despite these numerous other factors, actual choice behavior is often systematic and orderly.

EXPECTED UTILITY THEORY AND ITS VIOLATIONS

Expected utility theory has been, and continues to be, the dominant theoretical framework in the social sciences. The classical theory dates back to 1738 when Daniel Bernoulli suggested that people should make decisions that maximize their expected utilities (subjective values). Utilities reflect the psychological satisfaction of wealth, rather than wealth *per se*. Bernoulli suggested that the utility of wealth increased rapidly at first, then gradually slowed, consistent with a logarithmic function. A function of this shape describes the fact that people are often risk-averse in their preferences, although less risk-averse as wealth increases.

Modern utility theory began in 1947 with a book called *Theory of Games and Economic Behavior*. In this book, von Neumann and Morgenstern provided a mathematical foundation for expected utility theory. Suppose decisions can be represented as choices between gambles. If people can rank order their preferences for gambles, and their preferences are consistent with a small set of axioms, choices can be represented 'as if' they were based on the maximization of expected utilities.

Allais' Paradox

Soon after von Neumann and Morgenstern completed their work, researchers started to devise examples showing how people violate the axioms of expected utility theory. One famous example is Allais' paradox. Consider a choice between options A and B. Option A is \$1 million for sure. Option B is a 10% chance of winning \$2 million, an 89% chance of winning \$1 million, and a 1% chance of winning nothing. When presented with this hypothetical choice, most people prefer A to B. If the utility of nothing is zero, this preference can be expressed in the expected utility framework as:

$$1.0u(\$1\text{m}) > [0.10u(\$2\text{m}) + 0.89u(\$1\text{m})] \qquad (1)$$

where $u(\$1\text{m})$ and $u(\$2\text{m})$ are the utilities of \$1 million and \$2 million respectively. The expression can also be written:

$$0.11u(\$1\text{m}) > 0.10u(\$2\text{m}) \qquad (2)$$

Now consider a choice between options C and D: option C is an 11% chance of \$1 million, and option D is a 10% chance of \$2 million. Most people prefer to D to C, which implies:

$$0.10u(\$2m) > 0.11u(\$1m)$$

which directly contradicts the previous expression. A preference for A implies a preference for C, or a preference for B implies a preference for D; but one

cannot prefer both A and D (or B and C). Examples such as this led researchers to question whether expected utility theory was both a normative and a descriptive theory of choice.

PROSPECT THEORY

In 1979 Kahneman and Tversky proposed a descriptive theory of risk choice called prospect theory. Prospect theory differs from expected utility theory in several respects. In expected utility theory, decision-makers evaluate the utility of total wealth. In prospect theory, utilities (which are referred to as values) are associated with changes in wealth relative to the status quo. Furthermore, losses have greater impact than gains of equal magnitude, an assumption known as loss aversion.

Endowment Effects

Loss aversion provides an explanation for a well-known finding called endowment effects. Consider an experiment that takes place in a college classroom. Half of the students are randomly assigned a gift, such as a university coffee mug. These students are sellers. Those without mugs are buyers. Sellers are asked to report the minimum amount of money they would accept to sell their mug. Buyers report the maximum amount of money they would be willing pay to buy a mug. An experimental market is conducted; if there are transactions, mugs and money are exchanged.

According to expected utility theory, the experimental market will ensure that mug owners are those who value mugs the most. Since mugs were randomly assigned to students, there is no reason to think that students designated as sellers would value the mugs more than those assigned the role of buyers. In order for those students who value mugs most to become mug owners, approximately half of the mugs would, on average, be exchanged. However, that is not what happens. Few mugs are ever traded. Selling prices are typically larger than buying prices by a factor of two or more. The explanation for the effect is loss aversion; the pain of losing the mug is greater in magnitude than the pleasure of gaining the mug.

The way in which an object is endowed also influences its value. People who are rewarded with an object for an exemplary performance tend to value that object more highly than people who obtain the same object through chance or poor performance. Furthermore, windfall gains, such as lottery winnings or inheritances, are spent more readily than other assets, presumably because they are valued less.

Framing Effects

Endowment effects demonstrate how shifts in the status quo can influence the value of objects. Framing effects demonstrate how shifts in the perception of the status quo can lead to preference reversals. Framing effects were initially demonstrated by Tversky and Kahneman in a story called the Asian disease problem. Participants were told, 'Imagine that the USA is preparing for the outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: with program A, 200 people will be saved; with program B, there is a 1/3 chance that 600 people will be saved, otherwise no one will be saved.' Participants were asked to select a program, and the majority chose program A. Another group of participants were told the same story, except descriptions of the programs were presented in terms of lives lost. They were told, 'With program A, 400 people will die. With program B, there is a 1/3 chance that no one will die, otherwise 600 people will die.' The majority selected program B.

Although programs A and B are identical, preferences reverse with framing. The majority of participants preferred the safer option when alternatives were described as gains (lives saved), and the majority preferred the riskier option when alternatives were described as losses (lives lost). Prospect theory predicts these reversals. Both gains and losses have diminishing sensitivity. Saving 200 lives with certainty has greater value that a 1 in 3 chance of saving 600 lives. Furthermore, the certain death of 400 people is more painful than a 2 in 3 chance that 600 will die.

Framing effects go far beyond laboratory demonstrations. Researchers in the USA examined preferences for automobile insurance among drivers in New Jersey and Pennsylvania. Both states offered insurance at similar costs, and both states allowed drivers to cut costs by giving up their right to sue if an accident occurred. In New Jersey, the default coverage offered by insurance companies contained no right to sue, although a driver could buy that right at additional cost. In Pennsylvania, the default coverage contained the right to sue, although drivers could decline the right and reduce costs. The different reference points led to big differences in coverage. Only 20% of New Jersey

drivers purchased the right to sue at additional cost, but as many as 75% of the Pennsylvania drivers purchased the right as part of the package. Drivers tended to accept the default coverage. If the default coverage in Pennsylvania had resembled the default coverage in New Jersey, Pennsylvania drivers would have saved approximately $200 million in annual insurance costs.

Contextual Effects

Expected utility theory implies that the relative preference for one option over another should not change when additional options are added in the choice set. This principle is called 'regularity'. Researchers have shown that relative preferences vary systematically with the context. In one study, decision-makers are asked to choose between two options, A and B, each described in terms of two attributes. At a later point, the same decision-makers choose among A and B, and a new option, C. Option C is worse than A on one attribute, and worse than B on both attributes. Suddenly in the context of C, option B starts to look better, and the relative preference for B over A increases.

Decisions can also be influenced by the global context or the implicit comparisons people make across many choices. A given change in an attribute has a greater effect on choice when the attribute range is narrow than when the attribute range is wide. In one experiment, students made choices between apartments that varied in monthly rent and distance to campus. When monthly rents ranged from $200 to $400, a $50 increase in monthly rent was much more aversive than when monthly rents ranged from $100 to $1000. Such contextual effects are also inconsistent with expected utility theory.

ALTERNATIVE FRAMEWORKS FOR CHOICE

It is clear that the decision strategies people use vary with the individual, the task, the context and the frame. Prospect theory describes some of these phenomena, such as endowment effects and framing effects, but not all of them. Contextual effects, for example, require other explanations. What other psychological processes describe choice processes? A number of researchers have explored alternative frameworks.

Rule Following

Researchers point out that some choices are based on the application of rules or norms to situations. People ask themselves, 'What kind of situation is this?', 'What kind of person am I?', and 'What does a person like me do in a situation like this?'. Rules do not involve the calculus of cost–benefit analysis. They are adopted simply because they seem right and reflect our social or professional identities. Generally, doctors make decisions that coincide with medical guidelines, teachers make decisions that follow academic codes, and lawyers make decisions that build on legal precedents.

Rules may also reflect our personal identities. Prudential rules of thumb are often used to cope with issues of self-control. Some courses of action have minimal effects when done once, but large effects when done repeatedly. Smoking a cigarette or eating a slice of chocolate cake has a small effect in the short run, but done habitually has large effects in the long run. In these cases, the short-term benefits loom large, and the long-term costs are remote. Following a rule minimizes effort and guards the decision-maker against both temptation and potentially painful trade-offs.

Some have argued that interactions among people can be categorized in terms of four fundamental social rules. These rules are communal sharing, authority ranking, equality matching, and market pricing. Communal sharing rules stress the common bonds among members of a group, such as families, lovers or nations. Rules based on authority ranking highlight asymmetries in rank, privilege or prestige, as found in the military or the workplace. Equality matching stresses reciprocity. Decisions to join babysitting cooperatives or car pools imply that one agrees to give back whatever one takes. Finally, market pricing rules are governed by supply and demand, expected utilities, or trade-offs between costs and benefits. People often find these rules intrinsically satisfying for their own sake. They also insist that others adhere to the rules and punish those who do not conform.

When decision-makers apply the wrong rule to a social situation, they may make 'taboo' trade-offs. To attach a monetary value to one's friendship, one's children or one's academic integrity is to demonstrate that one is not really a friend, a parent or a scholar. In fact, the mere mention of selling priceless things can degrade the reputation of the person suggesting it.

Reason-based Choice

When rules conflict or seem irrelevant, people look for reasons to justify their decisions to themselves or others. This approach to decision-making

involves a balance of arguments for and against a course of action. Reasons might be lists of the pros and cons, or they might be stories. Jurors often construct stories to explain the facts. Courtroom evidence presented in the form of stories often leads to stronger decisions and more confident jurors than evidence presented in the form of issues.

Reason-based choices can be peculiar. In one study, the researchers asked students to imagine they were serving on a jury deciding the award of sole custody of an only child following a messy divorce. Parent A was described as having an average income, average health, average working hours, reasonable rapport with the child, and a relatively stable social life. Parent B was described as having an above-average income, a very close relationship with the child, an extremely active social life, lots of work-related travel, and minor health problems.

One group of students decided which parent would be awarded sole custody of the child. The majority selected parent B because of the positive features, such as the close relationship with the child and the good income. Another group decided which parent would be denied custody of the child. These students also selected parent B because of the negative features, such as the extensive travel schedule and health problems. Positive reasons can be more compelling when we select, and negative reasons can be more compelling when we reject. In these cases, we might accept and reject the same option.

Pattern Matching

Although many researchers have adopted the gamble as a template for a decision, other researchers reject this metaphor. They argue that decisions made by experienced people under time pressure in real-world settings are better represented by recognizing patterns. In these cases, decision-makers do not necessarily compare two or more options. Their experiences allow them to see situations as examples of prototypes, and they often know a course of action that is likely to succeed without making comparisons. This form of decision-making involves pattern matching.

Emotion-based Choice

Sometimes we make choices to feel good. Which film to watch, which book to read or which perfume to buy might depend on our view of what seems pleasurable. 'Feeling good' might mean avoiding regret. Researchers have shown that women who anticipated regret about their child dying as the result of a vaccination were less likely to have the child vaccinated, even when the mortality risks of the disease were greater than those of the vaccination. Consumers who imagined purchasing an unfamiliar product that later malfunctioned were more likely to buy a familiar product. Finally, students who were given a lottery ticket and asked if they would trade their ticket for a new one with objectively better odds tended not to trade because of the regret they anticipated if their original ticket were to win.

In other cases, 'feeling good' might mean anticipating the pleasure of positive outcomes and the pain of negative outcomes, and selecting the option that feels better on average. This rule is similar, though not identical, to expected utility theory. Expected utility theory predicts that people will choose the option with greater average utility, not greater average pleasure. Differences between the theories arise when anticipated pleasure deviates from utilities. This is often the case. Unlike utility, anticipated pleasure varies with multiple reference points. Comparisons with one's past performance, with other peoples' performance, with outcomes under different states of the world, and with outcomes under different choices might influence the anticipated pleasure of consequences, but they are typically not included in utilities.

Such comparisons have systematic effects on anticipated pleasure. Exceeding one's expectations adds to the pleasure of an outcome, and falling short of one's expectations detracts from the outcome. Furthermore, the impact of these comparisons is asymmetric. The incremental pain of doing worse than expected is greater in magnitude than incremental pleasure of doing better. In a gambling context, a small win can become more pleasurable if one avoids a big loss; but the same small win can be extremely disappointing if one fails to win an even larger amount. Another way in which anticipated pleasure and utilities differ involves our beliefs about outcomes. Pleasure depends on beliefs, while utilities are independent of beliefs. Unexpected outcomes are associated with more intense pleasure and pain. Utilities should be independent of likelihoods. Surprise effects are strong enough to make a smaller but surprising win more pleasurable than a larger win that was almost certain. In contrast, the utility of the smaller win could never exceed that of the larger win.

If decision-makers base their choices on the anticipated pleasure of future outcomes, the accuracy

of their forecasts is essential. Inaccurate predictions can surely lead to suboptimal choice. Researchers have identified some systematic errors in forecasting. One source of error is our immediate emotions: feelings of joy, anger, and sadness influence our attention, perception, memory and information processing strategies. When happy, we are better at retrieving happy memories, and when sad, we are better at recalling sad events. When happy, we use more flexible and creative problem-solving strategies. Sadness can lead to greater analytical thinking and longer response times, while anger has been linked to faster and less discriminate use of information.

Another source of error is the tendency to focus on whatever is salient at the moment, even when it has little effect later. In one study, students in the American Midwest and in California were asked to assess how happy they were, and how happy students like them in the other region of the country would be. The comparison highlighted the cultural opportunities, better climate and greater natural beauty of California. Both groups thought that students in California were happier; but in fact, the students were equally happy.

Researchers also asked college professors who were coming up for tenure how they expected to feel if they did or did not receive tenure. The professors expected to be happy if given tenure and extremely unhappy otherwise. Later the researchers contacted the professors to ask them what had happened and how they felt about it. Those denied tenure were actually much happier than they expected to be.

People also overestimate the pleasure of future experiences. In a classic study, researchers examined the happiness of lottery winners, matched control subjects, and people who were paraplegic. Although the lottery winners might have been thrilled about their wins in the days or weeks immediately after the event, they were no happier than the control subjects approximately a year later. Furthermore, the control subjects were only mildly happier than the paraplegic subjects. The tendency to focus on a single event can lead to overestimates of both pain and pleasure.

Deciding How to Decide

What determines the strategy for making decisions? Researchers argue that people are limited information processors who have access to many possible choice rules. Holding all else constant, they want to minimize the effort involved in a decision, and make accurate decisions. The strategies they use depend on trade-offs between effort and accuracy. More recently, this framework has been extended to include trade-offs between accuracy and negative emotions. People want to make accurate choices, and simultaneously avoid painful trade-offs. These goals can also determine choice strategies.

THE PSYCHOLOGY OF BELIEFS

Most decisions are associated with uncertainty, and beliefs play an important role. Researchers investigated the extent to which people update their beliefs according to normative principles, such as Bayes' theorem. Some argued that people shift their beliefs in the appropriate direction, but not to the right extent. Others argued that people are not Bayesians at all; instead, they use at least three heuristics. Representativeness is judgment based on the similarity of the event to the category. Availability is judgment based on the ease with which instances come to mind. Anchoring and adjustment are judgment based on insufficient movement from a reference point. These heuristics are rules that influence judgments of probability.

Base Rate Neglect

Researchers have argued that people often neglect base rate information. Evidence comes from a famous story, the 'cab problem'. Participants are told, 'A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city: 85% of the cabs in the city are green and 15% are blue. A witness identified the cab as blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was blue rather than green?' The Bayesian solution is 41%, but the majority of participants say 80%. This judgment is sensitive to the accuracy of the witness, but not the base rate. People estimate probabilities based on the information that seems most salient and compelling. In the cab problem, that information is the accuracy of the witness.

Conjunctive Probabilities

Researchers have also argued that people violate the conjunctive rule, according to which the judged probability of the intersection of two events cannot

exceed the judged probability of either single event. In these studies, participants are told a story about a woman named Linda who is described as 31 years old, single, outspoken, and very bright. She majored in philosophy and cared deeply about issues of discrimination and social justice. She also participated in antinuclear demonstrations. Participants are asked to rank the likelihood of various statements, including 'Linda is a bank teller' and 'Linda is a bank teller and a feminist'. Participants report that the statement, 'Linda is a bank teller and a feminist' is more probable than 'Linda is a bank teller'. Some researchers claim that representativeness, or the similarity of the target description to the category, governs these probabilities, and this heuristic can lead to mistakes.

Is it possible to 'fix' these mistakes? Some researchers have presented information using frequency formats. Participants are told to imagine 100 women who fit the description of Linda. Of those, how many are bank tellers? Feminists? Bank tellers and feminists? With this format, probability judgments are more likely to resemble the correct solutions with both base rate problems and conjunction problems.

Fast and Frugal Heuristics

Not all heuristics lead to errors. Some heuristics work well because they exploit the structure in the environment. Fast and frugal heuristics require a minimum of time, knowledge and computation, and tend to focus on one-reason decision-making. Such heuristics are based on easily computable search and stopping rules. The recognition heuristic involves the use of recognition to draw inferences. In some conditions, the lack of knowledge is beneficial. In one study, students at the University of Chicago were asked to decide which of two US cities had the larger population. University of Chicago students tend to be quite knowledgeable about US cities, yet despite their knowledge, they were less likely than German students to judge correctly whether San Diego or San Antonio had the larger population. Most of the German students had never heard of San Antonio, and using the recognition heuristic, they answered the question correctly.

Another heuristic called 'take the best' applies to predictions, inferences and decisions based on multiple imperfect cues. Suppose decision-makers are asked which of two cities is larger, and both cities are recognized. They compare cities on the basis of the most valid cue, and if that does not discriminate, they take the next best cue, and so on. With this rule they 'take the best, ignore the rest'. This heuristic can do almost as well as statistical rules in many cue environments.

CONCLUSION

Research on decision-making has demonstrated important strengths and limitations in human judgment and choice. These insights have led to descriptive theories that incorporate psychological concerns, such as social pressure, conformity, justifiability, and emotional satisfaction. Behavioral assumptions like these can greatly improve the prediction of human decision-making. Although people violate rational principles, their decisions are still regular and orderly. Uncovering this lawfulness will provide fertile ground for future research.

Further Reading

Dawes RM (1988) *Rational Choice in an Uncertain World*. Fort Worth, TX: Harcourt Brace.

Gigerenzer G, Todd P and the ABC Research Group (1999) *Simple Heuristics That Make Us Smart*. New York, NY: Oxford University Press.

Kagel JH and Roth AE (eds) (1995) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Kahneman D and Tversky A (eds) (2000) *Choices, Values, and Frames*. New York, NY: Cambridge University Press.

Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Klein G (1998) *Sources of Power*. Cambridge, MA: MIT Press.

March JG (1994) *A Primer on Decision Making*. New York, NY: Free Press.

Plous S (993) *The Psychology of Judgment and Decision Making*. New York, NY: McGraw-Hill.

Von Winterfeldt D and Edwards W (1986) *Decision Analysis and Behavioral Research*. New York, NY: Cambridge University Press.

Deductive Reasoning

Introductory article

PN Johnson-Laird, Princeton University, Princeton, New Jersey, USA

CONTENTS

Deductive reasoning yields conclusions that must be true given that the premises are true. The challenge to cognitive scientists is to discover the valid conclusions that human reasoners are competent to draw, and the mental processes underlying deductive performance.

INTRODUCTION

Deductive reasoning is a process yielding valid conclusions from premises, where a conclusion is valid if it must be true given that the premises are true. The premises may be assertions, perceptions, memories or thoughts. The conclusion is in the first instance a thought that may lead on to an assertion or to an action. A typical example of a valid deduction is:

Either government invests or a recession occurs.
Government doesn't invest.
Therefore, a recession occurs.

If the premises are true, then the conclusion must be true. If the premises are false, however, a valid deduction may yield a true or a false conclusion. Invalid inferences, such as inductions from experience, can be useful, but they lack a guarantee of truth even if their premises are true.

The business of life depends on deductive reasoning. Individuals differ in this ability, and those who are better at it also perform better on tests of intelligence. They also appear to be more successful in life. A person who is poor at deductive reasoning is liable to blunder. Conversely, without deductive reasoning, there would be no logic, mathematics, science, law or society. The challenge to cognitive scientists is to discover the valid conclusions that human reasoners are competent to draw, and the mental processes that underlie their performance.

LOGIC

Logic is the science of valid deduction, and logicians have systematized many different logical calculi. They approach the task in two distinct ways. The first way concerns the formal pattern or syntax of symbols, and it is known as 'proof' theory. Consider as an example the sentential calculus, which concerns negation and idealized versions of such connectives as 'if', 'or' and 'and'. The calculus can be formalized by specifying rules of inference that govern it. The following rule, for example, concerns disjunctions:

A or B.
Not-A.
Therefore, B.

This rule states that given a disjunction of the form *A or B*, such as:

Either government invests or a recession occurs.

and an assertion of the form *not-A*:

Government does not invest.

then one can derive a conclusion of the form *B*:

A recession occurs.

When logicians formalize a calculus, they bear in mind the intended meaning of the connectives, but the formal rules themselves make no use of this meaning. They are rules for writing new patterns of symbols, which are sensitive only to the syntax or form of sentences. Hence, formal rules operate like a computer program. When a computer program predicts the economy, for example, the computer has no idea of what an economy is or of what it is doing. It merely slavishly follows its instructions, and displays symbols that economists can interpret as predictions. Indeed, an intimate relation exists between computer programs and formal proofs: logic can be used to prove theorems about programs, and programs can be used to construct logical proofs.

Formalization can be carried out in different, though equivalent, ways. One way, which has been influential in cognitive science, is known as the method of 'natural deduction'. It uses formal

rules of inference for each of the main sentential connectives. Some rules introduce connectives, such as the following three rules:

A B Therefore, A and B	A Therefore, A or B, or both	A|− B Therefore, If A then B

where A|− B signifies that the assumption of A for the sake of argument yields a proof of B. Other rules eliminate connectives, such as:

	A or B, or both	If A then B
A and B	Not-A	A
Therefore, B	Therefore, B	Therefore, B

Natural deduction had a vogue in textbooks, but still more intuitive methods exist for teaching logic.

The second way in which logicians systematize a calculus is semantic; i.e. it is concerned with meaning and truth. It is known as 'model' theory. For the sentential calculus, this method states the truth or falsity of a proposition containing a connective in terms of the truth or falsity of its constituent propositions. The ultimate constituents are atomic propositions, propositions that contain neither negation nor connectives. Consider a disjunction of the form *A or B, or both*, such as:

> Either government invests or a recession occurs, or both.

Logicians assume that it is true if at least one of its two atomic propositions (government invests, a recession occurs) is true, and it is false if both of them are false. These conditions apply to many other disjunctions. Natural language, however, is not always so tidy. Just as there is poetic license, so there is logical license – a need for logicians to make simplifying assumptions about the meanings of logical terms. Logicians lay out the truth conditions for connectives in a truth table, such as that shown in Table 1. Each row in the table shows a possible combination of truth values of A and B, and the resulting truth value of the disjunction. The first row in the table, for instance, represents the case where A is true and B is true, and so the disjunction as a whole is true.

Table 1. Truth table for the disjunction *A or B, or both*

A	*B*	*A or B, or both*
True	True	True
True	False	True
False	True	True
False	False	False

To complete the semantic characterization of the sentential calculus, definitions of each connective need to be made. One tricky connective is 'if'. A conditional assertion such as:

> If government invests then inflation occurs.

is true given that government invests and inflation occurs, and false given that government invests and inflation does not occur. What is its status when its antecedent is false – that is, when government does not invest? This puzzle has generated a vast literature. The simplest semantics, however, may govern usage in daily life. It treats the conditional as true when its antecedent is false. In other words, granted that the conditional is true but that the government does not invest, then inflation may or may not occur.

As logicians have proved, any conclusion that can be derived using the formal rules of the sentential calculus is also valid using truth tables, and vice versa. They have also proved that there is a decision procedure for the calculus; i.e. a method that takes a finite number of steps to decide whether an inference is valid or invalid. However, this happy state of affairs does not apply to all logical calculi. An important extension of the sentential calculus is known as the predicate calculus. It deals with proofs concerning quantifiers, such as 'some' and 'all', as in the following syllogism:

> Some of those economies are inflationary.
> All inflationary economies are unstable.
> Therefore, some of those economies are unstable.

The predicate calculus is only semi-decidable. That is, any valid inference can be derived in a finite number of steps; but if an inference is invalid, no guarantee exists that its invalidity can be established in a finite number of steps. The most important logical discovery, as Kurt Gödel showed, is that logics exist in which not all valid inferences can be proved using formal rules. As he also famously proved, there are truths in arithmetic that cannot be derived in any consistent formal calculus. Some theorists have taken an analogous proof about computer programs to show that human creativity is not computable. Alan Turing, the intellectual father of the programmable digital computer, anticipated and rebutted this argument when he pointed out that human beings are unlikely to be consistent.

DEDUCTIVE COMPETENCE

Once, the task for cognitive scientists seemed to be to isolate the logic (or logics) that people have in their minds. The challenge was the variety of

candidates. For example, there are infinitely many distinct modal logics, which deal with possibility and necessity. Nevertheless, theorists argued throughout the twentieth century that logic is *the* theory of deductive competence. Jean Piaget, the famous investigator of children's mental development, argued that 'reasoning is nothing more than the propositional [i.e. sentential] calculus itself'. Yet there is a flaw in this claim. The sentential calculus yields infinitely many different valid conclusions from any set of premises. For instance, the following premises:

If government invests then inflation occurs.
Government invests.

yield the valid conclusions:

Government invests and government invests.
Government invests and government invests and government invests.

… and so on *ad infinitem*.

Of course such conclusions are preposterous. No sane individual – other than a logician – is likely to draw them. Yet they are all valid. Given, say, the following two premises and asked what follows from them:

A villanelle is a poem.
Ruislip is in London.

most people respond, 'Nothing'. Yet, to repeat, logic allows infinitely many valid conclusions from any premises. So, the response is wrong. However, the conclusions that follow from these premises are of no interest or use:

A villanelle is a poem *and* Ruislip is in London.

People are sensible. They do not draw just any valid conclusion, and sometimes they respond that nothing follows. Logic is therefore not a theory of deductive competence (*pace* Piaget). It captures which conclusions are valid, but it has nothing to say about which particular valid conclusions, if any, individuals draw.

The conclusions that people do draw tend to conform to three general principles. First, the conclusion maintains the semantic information given in the premises; that is, individuals do not throw away semantic information by adding disjunctive alternatives. Thus, the inference:

Government invests.
Therefore, government invests *or* inflation occurs.

is valid. Yet logically untrained individuals baulk at it. They do so for the excellent reason that the premise conveys more information than the conclusion, i.e. the premise is compatible with fewer possibilities than the conclusion. Second, reasoners are parsimonious in their conclusions. They do not spontaneously make deductions such as:

Government invests.
Inflation occurs.
Therefore, government invests *and* inflation occurs.

This conclusion uses more words to say no more than the premises do. Third, reasoners try to draw conclusions that make explicit something that was not stated as such in the premises. For example, given the premises:

If government invests then inflation occurs.
Government invests.

they do not draw the conclusion:

Government invests.

The conclusion is parsimonious, but it does not say anything new. Valid deductions cannot add semantic information to the premises, and so sceptics sometimes suggest that valid deductions serve no useful purpose. In fact, they are wrong. Valid deductions can convey propositions that are not obvious consequences of the premises. They can make explicit a proposition that was not asserted as such by any of the premises. It is these inferences that naive individuals strive to make. Hence, given the preceding premises, they almost always infer:

Inflation occurs.

If there is no conclusion that satisfies these three principles, then people are sensible. They say that nothing follows.

In short, to deduce is to maintain semantic information, to simplify, and to reach a new conclusion. None of these principles can be derived from logic.

THEORIES OF DEDUCTIVE PERFORMANCE

How people make deductions is a matter of controversy. Do their mental processes depend on a single system, or, as evolutionary psychologists suppose, on a set of separate systems shaped by natural selection? Do the processes rely on formal rules like those of a logical calculus, or on rules with a specific content, or, as some researchers in artificial intelligence suggest, on calling to mind past cases of valid reasoning? Psychologists have been struggling with deduction for nearly a century; cognitive scientists have recently homed in on it, and defended each of the preceding positions. The controversy is hot because deduction is an excellent test case: if cognitive scientists cannot

understand it, they are unlikely to understand much about any sort of thinking.

You can make valid deductions about matters of which you have no substantial knowledge. Even if you know nothing about poetry, for instance, you can still grasp the validity of the following inference:

If a poem is a villanelle then it has five tercets followed by a quatrain.
Elizabeth Bishop's *One Art* is a villanelle.
Therefore, it has five tercets followed by a quatrain.

Hence, theories based on knowledge of a domain can at best tell only part of the story of reasoning. Among the theories above, however, are two that can cope with inferences in general. These two theories run in parallel with the distinction in logic between proof theory and model theory. The first sort of theory postulates that your mind is equipped with formal rules of inference akin to those of a 'natural deduction' system. Lance Rips, Daniel Osherson, Martin Braine and others have proposed theories of this sort. You are not aware of these rules when you reason, and so if they exist, then they must be unconscious. One such rule is of the following form (see the earlier section on logic):

If A then B.
A.
Therefore, B.

You can make the preceding inference by matching its premises to the premises of this rule, and then by drawing the conclusion permitted by the rule.

The Selection Task

The British psychologist Peter Wason and his students pioneered the modern experimental study of deductive reasoning, and established that two main factors affect performance: the logical structure of inferences, and their semantic content. These studies showed, for example, that responses in Wason's well-known selection task were qualitatively different depending on the content of problems. In the abstract version of the task, the experimenter laid out four cards bearing letters or numbers as follows in front of the participants:

A B 2 3

The participants knew that each card had a letter on one side and a number on the other side. They were given a conditional assertion:

If a card has the letter 'A' on one side then it has the number '2' on the other side.

Their task was to select those cards that needed to be turned over to discover whether the rule was true or false about the four cards. Most people selected the A card alone, or the A and 2 cards. What was puzzling was their failure to select the 3 card: if it has an A on its other side, then the conditional assertion is false. Indeed, nearly everyone judges it to be false in this case. When the content of the assertion in the selection task was changed to a sensible everyday generalization, such as:

Every time I go to Manchester I travel by train.

many people made the correct selections. Each of four cards had the name of a destination on one side and the mode of transport for getting there on the other side, and the four cards shown to the participants were:

Manchester Sheffield train car

The participants tended to select the 'Manchester' and 'car' cards.

The selection task has been the most popular paradigm in the experimental study of deductive reasoning. Its large literature yields one overwhelming message: certain contents improve performance. Evolutionary psychologists argue that the mind has developed a set of specialized systems (or modules) as adaptations to the lives of our forebears. Thus, they have proposed an innate module for reasoning about cheating, because social exchanges were important to the hunter-gatherers of the Pleistocene epoch. Experiments have shown that selection tasks about potential cheaters do indeed improve performance; but, as the example above shows, such contents are not necessary to elicit insight into the task. Formal rules of inference cannot account for these phenomena either, because manipulations of content that have no bearing on logic can have a large effect on accuracy. These phenomena, however, did not deter subsequent theorists from postulating that reasoning relies on formal rules of inference.

Mental Models

As a reaction to the phenomena of the selection task, a second theory of deduction makes no use of formal rules of inference. This theory postulates that reasoning is not a formal matter (unless you have learned logic), but depends instead on your understanding of the meaning of the premises, on your perception of the situation, and on your general knowledge. You use this information to construct mental models of the situations described by the premises. You formulate a conclusion that

holds in the models but that was not explicitly asserted by any premise. The strength of your inference depends on the proportion of the models in which the conclusion holds. A conclusion that holds in all the models is *necessary* given the premises, and a conclusion that purports to be necessary can be refuted by a counter-example; that is, by a model that satisfies the premises but not the conclusion. A conclusion that holds in most of the models is *probable* given the premises, and a conclusion that holds in at least one model is *possible* given the premises. Models accordingly go beyond formal rule theories to provide a unified account of logical, modal, and probabilistic reasoning.

The theory makes three principal assumptions. The first assumption is that possibilities, not truth values, are fundamental, and that each mental model represents a possibility. Thus, the disjunction:

> Either government invests or a recession occurs, or both.

calls for three mental models, shown here on separate lines, which represent the three possibilities:

```
Government invests
                    recession
Government invests  recession
```

where 'Government invests' denotes a model of the government investing, 'recession' denotes a model of the occurrence of a recession, and the third line denotes their conjunction in a single model. The diagram uses words, but that does not imply that mental models are made up of words. There is a critical difference between the diagram and what it denotes, namely a set of mental models. You understand what it means for government to invest. Your understanding is represented in your mind in a way that is independent of your language. Indeed, if such a thought comes to you, you may have difficulty in putting it into words. Moreover, if you also speak Spanish, then you are likely to build the same mental models from the corresponding assertion in Spanish.

The second assumption of the theory is known as the 'principle of truth'. It postulates that mental models represent only what is true, and not what is false. Hence, the preceding set of models does not represent the case in which the disjunction as a whole is false. In addition, however, each model represents the clauses in the premises only when they are true in a possibility. For instance, the first model in the preceding set represents the possibility that government invests, but it does not make explicit that in this possibility it is false that a recession occurs. People represent what is false in 'mental footnotes', but these footnotes are ephemeral and normally soon forgotten. The failure to consider what might falsify the conditional assertion accounts for the difficulty of the abstract selection task. The failure is mitigated, however, when the content of the task yields more familiar counter-examples.

The third assumption is that the parts of a mental model correspond to the parts of what it represents, and so the structure of the model corresponds to the structure of what it represents. A mental model is therefore like an architect's model of a building. Mental models underlie visual images, which are projections from them, but not all mental models can be visualized. Complex mental models of systems can be built up in long-term memory.

To illustrate reasoning by model, consider again the inference:

> If government invests then inflation occurs.
> Government invests.
> Therefore, inflation occurs.

The conditional calls for a model of the case in which its antecedent and consequent are both true. People realize that there are other possibilities consistent with the conditional – in particular, those in which it is false that government invests. However, they do not represent them explicitly. The conditional therefore elicits two models:

```
Government invests    inflation
                  …
```

where the ellipsis denotes a model that has no explicit content. This implicit model is a 'place holder' for the cases in which it is false that government invests. The premise that government invests eliminates this second model, and the first model yields the conclusion that inflation occurs.

Illusions and Departures from the Norms of Logical Validity

Do individuals reason using formal rules or mental models? A recently discovered phenomenon may answer the question. Consider the following problem:

> Only one of the following assertions is true about a hand of cards:
> There is a king in the hand or an ace, or both.
> There is a queen in the hand or an ace, or both.
> There is a jack in the hand or a ten, or both.
> Is it possible that there is an ace in the hand?

Nearly everyone responds, 'Yes,' and the conclusion seems obvious. Yet the inference is illusory. It

is impossible for an ace to be in the hand. If there were an ace then the first two assertions would both be true, contrary to the claim that only one of the three assertions is true. People succumb to the illusion because they think only about what is true, and the truth of the first assertion suggests that an ace is possible. But, when the first assertion is true, the other two assertions must be false. And the falsity of the second assertion means that there is neither a queen nor an ace in the hand.

Individuals are vulnerable to a variety of illusions in modal and probabilistic reasoning. The rubric 'Only one of the assertions is true' is equivalent to an exclusive disjunction, and a compelling illusion occurs in the following inference about a particular hand of cards:

> If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand.
> There is a king in the hand.

What, if anything, follows? Nearly everyone infers that there is an ace in the hand. It follows from the mental models of the premises. Yet it is a fallacy granted a disjunction between the two conditionals. One or other of the conditionals could therefore be false. And if, say, the first conditional is false, then there is no guarantee that there is an ace in the hand even though there is a king. Of course, skeptics can argue that 'or else' is treated as meaning 'and' in the problem, but that argument fails to explain the illusions based on the rubric 'only one of the assertions is true', and it also fails to explain which sentences yield an interpretation of 'or else' as 'and'. They seem to be precisely those for which the model theory predicts that illusions will occur because individuals neglect what is false.

The illusions are so compelling that they go unnoticed in daily life. For example, a professor cautioned students on his Web page:

> Either a grade of zero will be recorded if your absence [from class] is not excused, or else if your absence is excused other work you do in the course will count.

The mental models of this assertion represent the two possibilities that presumably he had in mind, assuming that a zero grade is incompatible with other work counting:

```
not-excused     zero-grade
excused         other-work-counts
```

The students probably made the same interpretation. Yet it is wrong. What really conveys these two possibilities is a conjunction:

> A grade of zero will be recorded if and only if your absence is not excused, *but* if and only if your absence is excused then other work you do in the course will count.

An inclusive disjunction of the two conditionals is a much weaker assertion compatible with more possibilities than those above. It allows that students with an excuse are no different from those without an excuse: both may get a zero grade or have other work on the course count, or neither. An exclusive disjunction of the two conditionals is even stranger: students with an excuse either get a grade of zero or not, but other work never counts; and students without an excuse never get a grade of zero, and other work may or may not count.

The illusions are a litmus test for mental models, because its principle of truth predicts them, but they jeopardize theories of reasoning based on formal rules. These theories currently postulate only valid rules, which cannot explain the systematic invalidity of illusory inferences. If these theories introduced invalid rules to account for the illusions, then it would follow that human beings are intrinsically irrational. However, the illusions have no such implication: people do understand the explanation of their errors. Given the limitations of human working memory, reasoners cannot cope with truth and falsity, that is, with complete truth tables. The principle of taking into account what is true and forgoing what is false is a sensible compromise. Truth is more useful than falsity. Just occasionally, however, truth alone leads human reasoners into the illusion that they grasp a set of possibilities that is in fact beyond them.

Further Reading

Baron J (1994) *Thinking and Deciding*, 2nd edn. New York, NY: Cambridge University Press.

Braine MDS and O'Brien DP (eds) (1998) *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum.

Brooks M (2000) Fooled again. *New Scientist* **2268**: 24–28.

Davis M (2000) *The Universal Computer: The Road from Leibniz to Turing*. New York, NY: Norton.

Evans JSBT, Newstead SE and Byrne RMJ (1993) *Human Reasoning: The Psychology of Deduction*. Mahwah, NJ: Lawrence Erlbaum.

Garnham A and Oakhill J (1994) *Thinking and Reasoning*. Cambridge, MA: Blackwell.

Jeffrey R (1981) *Formal Logic: Its Scope and Limits*, 2nd edn. New York, NY: McGraw-Hill.

Johnson-Laird PN (1999) Deductive reasoning. *Annual Review of Psychology* **50**: 109–135.

Johnson-Laird PN and Byrne RMJ (1991) *Deduction*. Hillsdale, NJ: Lawrence Erlbaum.

Rips LJ (1994) *The Psychology of Proof*. Cambridge, MA: MIT Press.

Schaeken W, De Vooght G, Vandierendonck A and d'Ydewalle G (2000) *Deductive Reasoning and Strategies*. Mahwah, NJ: Lawrence Erlbaum.

Stanovich KE (1999) *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Delusions

Introductory article

William O'Donohue, University of Nevada, Reno, Nevada, USA
Andrew Lloyd, University of Nevada, Reno, Nevada, USA

CONTENTS

Delusions are strongly held false beliefs about reality. Although there is only moderate agreement as to what counts as a delusion, delusional beliefs and associated behaviors have long been regarded as evidence of insanity.

INTRODUCTION

Delusions are strongly held false beliefs about reality. The belief that one's thoughts are being stolen by a clandestine government agency or that one's internal organs have been replaced with mechanical replicas are good examples of delusional beliefs. Delusions are associated with over 70 psychological conditions and disorders. Schizophrenia, Alzheimer disease, Parkinson disease (along with other dementing diseases), and many substance abuse-related disorders are often marked by delusions. Delusional beliefs, and the behaviors consistent with them, are frequently the most relevant factors leading to the conclusion that a person is mentally ill. Indeed, delusions have been considered to be the defining, or paradigmatic, feature of madness for centuries. However, despite the clear relevance of delusions to psychological, psychiatric, and general cognitive theories and models, only moderate agreement exists as to what counts as a delusion.

SYMPTOMATOLOGY

Standard definitions of delusions involve at least four components. First, a delusional belief is a false belief or error. Such false beliefs can run from the relatively mundane, such as the belief that a close friend is secretly in love with you, to the very bizarre, for example, that one's head is full of bees. Second, a delusional belief is maintained with great conviction or certainty. Such beliefs have the character of being fundamentally true and incapable of being questioned owing to their perceived importance. Third, a delusional belief is maintained despite, or in the face of, incontrovertible evidence to the contrary. Not only do such beliefs have no empirical support (because they are false), they also persist in the face of what most people would consider incontrovertible evidence of their falsity. The belief that the Earth is flat or that the Holocaust never took place are examples of beliefs maintained in the face of strong counter-evidence. Fourth, a delusional belief is the kind of belief that is not typically supported by the individual's culture. This condition, though it is of critical importance to the definition of delusions, admits of a great deal of complexity and is, by definition, relative to many context-dependent variables. Though certainly false, the belief that consuming the brain of a recently deceased family member can protect one from the cause(s) of that person's demise would not properly be considered delusional in the context of certain South American tribes.

These four components may be construed as the necessary conditions for labeling a belief delusional. That is, each of these criteria must be met before a belief can be considered a delusional belief. Because, however, each of these criteria poses potentially difficult diagnostic problems, it is possible that, even taken together, they do not constitute sufficient criteria for the classification of a belief as delusional. That is, even if it is judged that each of these four conditions has been met, it does not guarantee that the belief in question is delusional.

The number, type, and content of delusional beliefs are unlimited. Because of the unlimited complexity of language it is not possible to identify antecedently what particular beliefs might count as delusional. This further complicates the identification of delusional beliefs. Some beliefs, however, are straightforwardly identifiable as delusional. Some of these beliefs are so common in relation to other delusional beliefs that they have been named and studied in some detail. The Capgras delusion

and the Cotard delusion are prominent examples. The Capgras delusion is marked by the individual's belief that his or her family (close acquaintances, co-workers, etc.) have been replaced by imposters. In the Cotard delusion the individual believes that he or she is dead or possesses no 'real' existence. Both of these exemplar delusions satisfy each of the four criteria outlined above. However, though both the Capgras and Cotard delusions are distinguished by their bizarreness, delusional beliefs need not be bizarre in their content.

One of the most prominent difficulties associated with correctly identifying a belief as delusional is encountered when the truth of the belief is assessed. People may express a number of beliefs that, at face value, seem clearly false. Believing one is dead falls squarely in this category. Little difficulty is associated with classifying such beliefs as false. However, unless care is taken to assess the truth of a belief, it would be hasty to classify the belief as delusional. In one classic example of a mistaken judgment of falsity, in the early 1970s Martha Mitchell, wife to the Attorney General of the United States, expressed seemingly false beliefs regarding illegal activities going on in the White House. Mrs Mitchell was believed to be suffering from some sort of psychopathology based on the bizarreness of her beliefs. As it turned out, however, Mrs Mitchell's accusations were accurate. Mrs Mitchell's claims pertained to the Watergate scandal of 1972. There are countless examples of situations similar to this one. A provisional name has been applied to situations in which the truth-value of a belief has been mistaken, the 'Martha Mitchell Effect'.

Delusional beliefs are distinguished from other forms of erroneous beliefs, such as overvalued ideas, obsessions, and confabulations. Overvalued ideas are recognized by their possessors as being potentially false and, as such, are not maintained with the degree of certainly associated with delusions. Obsessions are defined as persistent ideas, thoughts, impulses, or images. Obsessions are often experienced as distressing, and though the obsessional person attempts to ignore or suppress them, they are clearly recognized as his or her own thoughts. Because people attempt to suppress obsessional thoughts, they fail to count as delusional beliefs that are held with certainty. Confabulation is defined as filling in the gaps of memory and knowledge with false information. This information, however, need not involve beliefs that are generally considered questionable or bizarre by the members of the person's community.

COGNITIVE MECHANISMS

There are a number of cognitive theories that attempt to explain the presence of delusional beliefs. The most prominent features of these theories will be outlined. According to some thinkers, delusions may reflect errors in discriminating the origin of information. Mistakenly attributing one's belief to experiences, rather than to the operation of imagination, can lead to false beliefs. Internal imaginary events can generate local effects in the same pathways used by perception itself, thus increasing the difficulty associated with discriminating the original source of the belief. This phenomenon is similar to the one we experience when we are unable to discriminate accurately between reality and fiction while dreaming. The more elaborate imaginary scenarios are in terms of sensory detail, the more likely they are to be conflated with actual experiences. When we imagine a scene or event in elaborate detail (e.g. the color of the sunset and the odor of the ocean) the imagined event takes on more of the properties of our veridical experiences. This issue has been at the core of the debate regarding repressed memories.

Occasional confusion regarding the source of one's beliefs is, in itself, relatively innocuous. Most people have experienced this kind of confusion during their lives: 'Did somebody tell me that, or do I simply think it?' 'Do I really remember doing that when I was three years old, or am I relying on Mom's stories?' As the definition of delusional beliefs suggests, more than simple confusion or error with respect to the source of one's beliefs is required to classify them as delusional. Two other cognitive mechanisms are required to induce the individual to hold such false beliefs with certainty in the face of evidence otherwise: hypervigilance and an unfounded sense of profoundness.

Hypervigilance has been identified by many theories as one of the critical cognitive dispositions associated with delusional beliefs. Hypervigilance is marked by an increase in a person's awareness of the details of his or her environment. Under normal conditions of vigilance the environmental information to be processed is well within the individual's cognitive capacity to organize and contextualize into a manifold whole. States of hypervigilance, however, may induce a cognitive overload resulting in a sense of ambiguity with respect to environmental stimuli. Human cognitive mechanisms simply cannot contextualize and organize all of the environmental stimuli that are present at any given moment. The ambiguity that results from such experiences may lead to an unusually intense

sense of mystery and profoundness. For example, an individual may notice that there are four pennies in his pocket, four steps in his staircase, four clients he must meet during the day, and four children playing tag on the lawn as he leaves for work. These four events might strike him as being in need of some sort of explanation. How often, after all, does one note four sets of four events on the way to work? The search for an explanation of this series of events may result in the belief that these four events are connected in some important way, despite the fact that they are simply chance observations. The meaning of these seemingly connected events may take on a significance that far outweighs any reasonably objective interpretation of the events. Does it mean that the man has only four years to live? Perhaps this means that the man will meet his soulmate at four o'clock that afternoon.

This profoundness of experience is best described as being similar to that which one has at the birth of a child, the death of a close friend, or the intense spiritual experiences common to many religions. In the normal course of events such experiences are rare and often life-altering. For the delusional individual, mundane daily experiences become increasingly entangled in a sense of mystery and profoundness that can seemingly only be explained in a profound manner. Such profundity may be associated with a feeling of certainty. Further complicating the issue is the increased likelihood that the individual will note that others do not react with a sense of awe at what seem to be awesome events. This, in turn, may incline the person to believe that only he or she is meant to understand, or see, these events as they are. This can lead either to isolation or to a sense of righteousness. When isolated, a person is unlikely to come into contact with enough evidence to dissuade him or herself of the meanings that have been mistakenly associated with random events. Similarly, self-righteousness may interfere with a willingness to accept the input of others when such input contradicts what he or she believes.

Thus, a person who takes close notice of his or her relatives may come to see things that were never noticed before. Behavioral irregularities, a small scar, and a slightly different vocal intonation may be attended to and seem to require explanation. For the individual suffering from the Capgras delusion, the explanation may be that his or her close relatives have been replaced by imposters. Such a belief is unlikely to be openly expressed to those who could counter it. As such, the belief may become more and more ingrained. By the time the individual expresses this belief to others it will be so thoroughly tangled in a web of spurious confirmatory evidence (e.g. Mom never used to like cabbage) that it will be difficult to dispel. The foregoing explanation of the predispositions and factors leading to the development of delusional beliefs are consistent with the view of delusional beliefs as unnecessary and errorful explanatory theories of events that required no explanation.

PERCEPTUAL MECHANISMS

Perceptual mechanisms are implicated in the development of delusional beliefs insofar as they are involved in the individual's experience of the world. The most relevant role that perceptions play with respect to delusions is in the area of hallucinations. Both visual (e.g. seeing snakes when no snakes are present) and auditory (e.g. hearing one's name called when no one is present) hallucinations may result from an elaboration of elementary sensations. For example, *muscae volitantes*, commonly known as 'eye floaters' (which are typically overlooked or ignored), have been experienced as rats, snakes, armies struggling for the person's soul, and so on. In such cases, the experience of something in the visual field is accurate, only the interpretation of what it is is inaccurate. Similarly, individuals suffering from alcohol withdrawal may incorrectly identify the sounds associated with intrinsic tinnitus with the buzzing of bees inside his or her head. Auditory hallucinations are more common than visual hallucinations because of the relative difficulty in identifying the source, either internal or external, of sounds as opposed to images.

NEURAL SUBSTRATES

Though little is known in this area, research continues in the identification of the neural substrates of delusional predispositions. Promising research has suggested that schizophrenic delusions and hallucinations may be associated with overactivity of the left hippocampus and ventral striatum.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Berrios GE (1991) Delusions as 'wrong beliefs': a conceptual history. *British Journal of Psychiatry* **159**: 6–13.

Jaspers K (1963) *General Psychopathology*, translated by J Hoenig and MW Hamilton. Chicago, IL: University of Chicago Press.

Oltmanns TF and Maher BA (1988) *Delusional Beliefs*. New York, NY: John Wiley.
Reed G (1988) *The Psychology of Anomalous Experience: A Cognitive Approach*. Buffalo, NY: Prometheus.
Roberts G (1992) The origins of delusion. *British Journal of Psychiatry* **161**: 298–308.

Depression

Introductory article

Lyn Y Abramson, University of Wisconsin-Madison, Madison, Wisconsin, USA
Lauren B Alloy, Temple University, Philadelphia, Pennsylvania, USA
Catherine Panzarella, The Philadelphia Behavioral Health System, Philadelphia, Pennsylvania, USA

CONTENTS

Depressive disorders are characterized by persistent depressed mood or loss of interest (normally for at least two weeks) and at least four other symptoms such as change in eating patterns or appetite, sleep disturbance, psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness or guilt, difficulty in concentration, and suicidal thoughts, plans, or attempts.

IMPORTANT FACTS ABOUT DEPRESSION

Descriptive research has produced some important facts about depression. First, depression is a common disorder. Recent estimates indicate that in many western countries, about 20 percent of the population will experience a clinically significant episode of depression at some point in their lives. Moreover, the rate of depression appears to be on the rise, especially among young people. The increased rate of depression in modern society may be due, in part, to increased focus on the self and its accomplishments. Although focus on the self may motivate achievement, it also may lead to depression when people fail to achieve their goals.

Second, depression is recurrent. Over 80 percent of depressed patients have more than one depressive episode in their lifetime. In fact, over 50 percent of depressed patients experience a relapse within two years of recovery. Such recurrence of depression suggests that some individuals are depression-prone because they exhibit a relatively stable vulnerability factor or diathesis for this disorder.

Third, the rates of depression surge during middle to late adolescence. Although young children can experience depression, rates of this disorder are relatively low during childhood. The surge in depression during middle to late adolescence may be due to the consolidation of vulnerability factors for the disorder such as negative cognitive styles as well as increased rates of negative life events within this age group.

Fourth, gender differences in depression exist among adults with twice as many women experiencing depression as men. This gender difference in depression emerges during adolescence.

Fifth, depression is associated with significant physical, vocational, and interpersonal impairment. Epidemiological studies have shown that depression is associated with poor physical health including cardiac problems and elevated rates of smoking. Moreover, depression lowers productivity in the workplace. In one year, it is estimated that in the United States the costs of depression-related lost productivity can exceed $33 billion. Finally, depression has high interpersonal costs. For example, depressed people have a higher divorce rate than nondepressed people.

Sixth, depression can be lethal as it clearly increases the risk for suicide. Research has suggested that hopelessness is the key factor contributing to suicide among depressed individuals.

Seventh, life events play a role in the development of depression. The occurrence of undesirable, major life events is associated with the onset of depression. However, not all people become clinically depressed when confronted with severe life stressors. Vulnerability–stress models posit that individuals exhibiting vulnerability factors (cognitive, genetic, biological, etc.) for depression are more likely to become depressed when confronted with severe stressors than individuals not exhibiting such vulnerability.

Finally, depression long has been viewed as heterogeneous with multiple causes. Thus, there are likely to be different causal pathways that culminate in depression. (*See* **Affective Disorders: Depression and Mania**)

COGNITION AND DEPRESSION

Although depression long has been recognized as an important form of psychopathology, experimental psychopathologists neglected this disorder until the 1970s. At that time, research on depression burgeoned within clinical psychology, and many investigators began to emphasize cognitive processes in the etiology, maintenance, and treatment of depression. A core idea within this cognitive perspective is that different people can perceive the same event differently. Psychologists focus on the characteristic ways that individuals perceive situations and how their perceptions, in turn, relate to their behaviors and emotions. From the cognitive perspective, emotional reactions, such as depression, to events are determined by a combination of characteristics of the event itself and the cognitive construal processes of the perceiver.

COGNITIVE THEORIES OF DEPRESSION

Why are some people vulnerable to depression whereas others never seem to become depressed at all? According to the two major cognitive theories of depression, the hopelessness theory (Abramson *et al*, 1999, 2002) and Beck's theory (Beck, 1987; Clark *et al*, 1999) the meaning or interpretation that people give to their experiences importantly influences whether they will become depressed and whether they will suffer repeated, severe, or long-duration episodes of depression. Indeed, the demonstrated efficacy of cognitive therapy for depression underscores the powerful clinical implications of a cognitive approach to depression.

According to the hopelessness theory, the expectation that highly desired outcomes will not occur or that highly aversive outcomes will occur and that one cannot change this situation – hopelessness – is a precipitant of depressive symptoms. How does a person become hopeless and, in turn, develop the symptoms of depression? As shown in Figure 1, negative life events (or the nonoccurrence of desired positive life events) are 'occasion setters' for people to become hopeless. However, the relation between negative life events and depression is imperfect; not all people become depressed when confronted with negative life events. According to the hopelessness theory, three kinds of inferences or conclusions that people may make when confronted with negative life events contribute to the development of hopelessness and, in turn, depressive symptoms: causal attributions, inferred consequences, and inferred characteristics about the self. In brief, hopelessness and, in turn, depressive symptoms are likely to occur when negative life events are (1) attributed to stable causes (likely to persist over time) and global causes (likely to affect many areas of life) and viewed as important; (2) viewed as likely to lead to other negative consequences; and (3) construed as implying that the person is unworthy or deficient.

For example, suppose a student fails a test. According to the theory, the student is likely to become depressed if he/she believes that the failure:

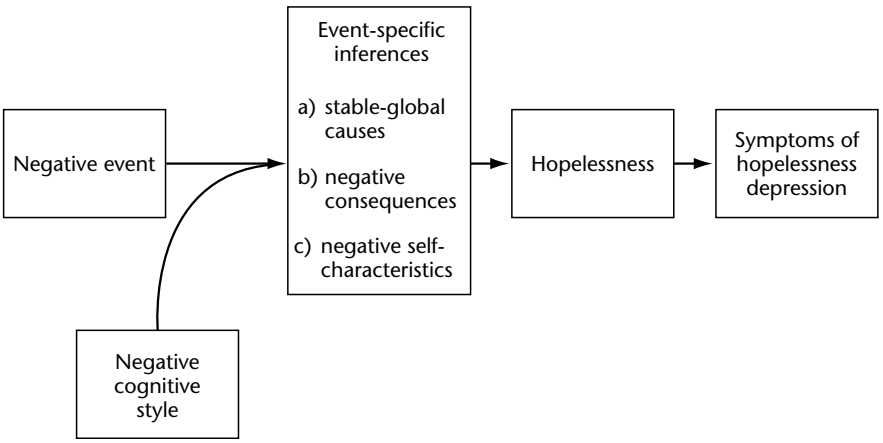


Figure 1. Causal chain in the hopelessness theory.

(1) was due to low intelligence; (2) will prevent him/her from pursuing a particular career; and (3) means that he/she is worthless. In contrast, another student who fails the same test will be protected from becoming depressed if he/she believes that the failure: (1) was due to not studying hard enough; (2) will motivate him/her to do especially well on the next test; and (3) has no implications for his/her self-worth.

In the hopelessness theory, informational cues in the situation as well as individual differences in cognitive style influence the content of people's inferences about cause, consequence, and self when negative life events occur. Individuals who exhibit a general tendency to attribute negative events to stable and global causes, to infer that current negative events will lead to further negative consequences, and to infer that the occurrence of negative events means that they are deficient or unworthy, are more likely to make these depressogenic inferences about a given negative event than individuals who do not exhibit this depressogenic cognitive style. However, in the absence of negative left events, those exhibiting a depressogenic cognitive tendency should be no more likely to develop hopelessness and, in turn, depressive symptoms than people not exhibiting this tendency. This aspect of the theory is a vulnerability–stress component: negative cognitive patterns are the cognitive vulnerability, and negative life events are the stress.

The etiological hypotheses of Beck's cognitive theory of depression are similar to those of the hopelessness theory. To understand Beck's theory, it is useful to describe cognitive schemata. Schemata are generalized core beliefs about the self and world that assist people in processing complex environmental information by: (1) selecting only a fraction of incoming stimuli for processing; (2) abstracting meaning from incoming information and favoring storage of the meaning rather than a veridical representation of the original stimulus; (3) using prior knowledge to assist in processing and interpreting information; and (4) integrating information to favor internal consistency over external accuracy. Schemata pertaining to the self are featured in Beck's theory.

As Figure 2 shows, in Beck's theory, maladaptive self-schemata containing dysfunctional attitudes involving themes of loss, inadequacy, failure, and worthlessness constitute the cognitive vulnerability for depression. Such dysfunctional attitudes often involve the theme that one's happiness and worth depend on being perfect or on other people's approval. Examples of dysfunctional attitudes include, 'If I fail partly, it is as bad as being a complete failure', or 'I am nothing if a person I love doesn't love me'. When these hypothesized depressogenic self-schemata are activated by the occurrence of negative life events (the stress), they generate specific negative cognitions (automatic thoughts) that take the form of overly pessimistic views of oneself, one's world, and one's future (the negative cognitive triad) that, in turn, lead to sadness and other symptoms of depression. In the absence of activation by negative events, however, the depressogenic self-schemata remain latent, less accessible to awareness, and do not directly lead to negative automatic thoughts or depressive mood and symptoms.

Although differing in some specifics, hopelessness and Beck's theories share many important features. At the most basic level, both theories emphasize the role of cognition in the origins and maintenance of depression. In addition, both theories contain a cognitive vulnerability hypothesis in which negative cognitive patterns increase people's vulnerability to depression when they experience

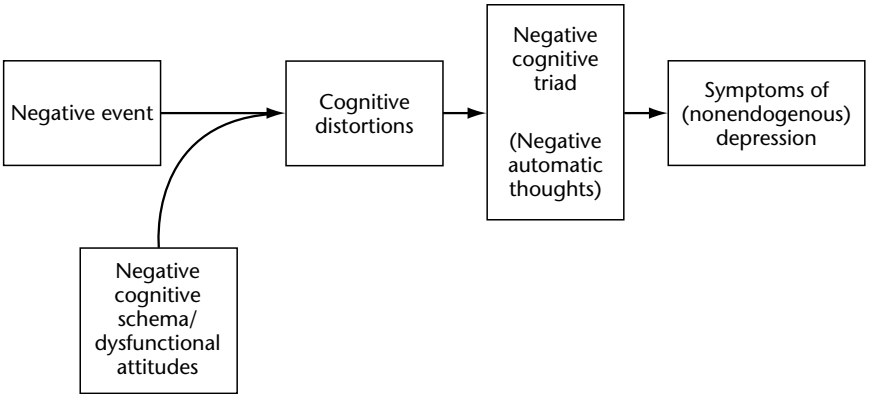


Figure 2. Causal chain in Beck's theory.

negative life events. Moreover, both theories also propose a mediating sequence of negative inferences that influence whether or not negative events will lead to depressive symptoms. Finally, both theories recognize the heterogeneity of depression and acknowledge that other causes of depression may exist.

Despite their similarities, there is one striking difference between hopelessness and Beck's theories. To understand this difference, it is useful to distinguish between cognitive 'processes' and cognitive 'products'. Cognitive processes involve the operations of the cognitive system such as information encoding, retrieval, and attentional allocation. Cognitive products are the end result of the cognitive system's information processing operations and consist of the cognitions and thoughts that the individual experiences. Inferences about cause, consequences, and self, as featured in the hopelessness theory, are examples of cognitive products. According to the hopelessness theory, depressive and nondepressive cognition differ in content (e.g., stable, global versus unstable, specific causal attributions for negative events) but not in process. In contrast, Beck's original theory emphasized that depressive and nondepressive cognition differs not only in content but also in process. Beck suggested that the inference process is 'schema driven' among depressed people and 'data driven' among nondepressed people. Thus, although both theories emphasize that depressed people's inferences are negative, Beck further proposed that depressed people's negative inferences are unwarranted given current information. Specifically, Beck suggested that depressed individuals ignore positive situational information and are unduly influenced by current negative situational information in making their negative inferences. In contrast, Beck hypothesized that nondepressed individuals appropriately utilize current information in making inferences. In short, Beck's original theory emphasized that depressive cognition is distorted whereas hopelessness theory is silent on the distortion issue.

EMPIRICAL FINDINGS ON COGNITION AND DEPRESSION

Supporting the cognitive theories of depression, many cross-sectional studies have established that negative cognitive patterns are associated with depression in adults, children, and adolescents. However, such cross-sectional studies do not provide strong tests of the cognitive vulnerability hypothesis, because they cannot distinguish between the possibility that the cognitive vulnerability came first and contributed to the occurrence of depression, as hypothesized in the cognitive theories of depression, and the alternative possibility that cognitive vulnerability does not contribute to depression and, instead, is a correlate or consequence of depression. Prospective behavioral high-risk studies are needed to establish that cognitive vulnerability actually precedes the occurrence of depression and contributes to its onset.

In the behavioral high-risk design, currently nondepressed people are selected who are at high versus low risk for depression based on the presence versus absence of the hypothesized depressogenic cognitive patterns. These cognitive high and low risk groups would then be followed and compared on their likelihood of developing depression in the future. Although there are some exceptions, in general recent studies with children, adolescents, and adults using the behavioral high-risk design and approximations to it have found that individuals who exhibit the hypothesized cognitive vulnerabilities are more likely to develop depressive moods, symptoms, and clinically significant episodes over time than individuals who do not exhibit cognitive vulnerability. Moreover, individuals exhibiting cognitive vulnerability also show elevated rates of suicidal thoughts when they are followed over time. Thus, work supports the hypothesis that ingrained negative patterns of thinking provide risk for depression.

In addition, some research suggests that attributional style for interpreting positive, rather than negative, events may be important in predicting recovery from depression as well as lower rates of relapse. For example, depressed people who show a tendency to attribute positive events to stable, global causes are more likely to recover from depressive symptoms following the occurrence of positive events than depressed people who attribute positive events to unstable, specific causes. Similarly, depressed patients with an internal, stable, global attributional style for positive events are less likely to relapse in the year following hospital discharge than depressed patients who do not show this attributional style. In other words, depressed people who did *not* show the tendency to minimize or discount positive information were more likely to recover and were less vulnerable to relapse.

In addition to work on cognitive styles as vulnerability and invulnerability factors for depression, much research has focused on documenting the nature of cognitive processes exhibited by depressed persons (Alloy *et al*, 1997). For example,

this research shows that processing negative self-referent information is less effortful and more automatic for depressed than non-depressed people. Depressed people also exhibit more negative memory biases than nondepressed people. Taken together, a vast amount of research unequivocally shows that depressive cognition is more negative than nondepressive cognition. But are depressed people's negative cognitions also more unrealistic than nondepressed people's positive cognitions?

DEPRESSIVE REALISM?

In contrast to Beck's original characterization of 'depressive cognitive distortion' and 'nondepressive accuracy', research has demonstrated pervasive optimistic biases among nondepressed people and a 'depressive realism effect' in which depressed individuals actually are more accurate than nondepressed individuals. For example, nondepressed individuals often exhibit an 'illusion of control' in which they believe that they control outcomes over which they objectively have no control, whereas depressed individuals seem less susceptible to this illusion. Research on optimistic biases among 'normal' people suggests that in formulating theories of depressive cognition, clinical researchers may have been wrong to assume that accuracy is the baseline of normal cognitive functioning. Instead, laboratory work has demonstrated that both depressed and nondepressed people show cognitive biases and illusions that are consistent with their preconceived beliefs or schemas (Dykman *et al*., 1989).

Although work on depressive realism and nondepressive optimistic illusions has not established that depressed people are *always* more accurate than nondepressed people, these studies nevertheless have posed an important challenge to the portrayal of depressed people as either impervious to information in their environments or hopelessly biased by pervasive negative schemata and of nondepressed people as completely data-driven and free from the influence of biasing schemata. Depressive and nondepressive cognition may differ more in content than in process. Thus, the negative cognitive biases that predict risk for depression may be no more distorted than the positive cognitive biases that predict protection from depression. Lively debate continues about the question of who are more accurate in perceiving reality over the long run – depressed people or nondepressed people?

DEVELOPMENTAL ORIGINS OF COGNITIVE VULNERABILITY

Given the growing body of work suggesting that negative cognitive styles may confer vulnerability for depression and suicidality, it is important to understand the developmental origins of these negative patterns of thinking. It is likely that a multitude of factors contributes to negative cognitive patterns such as genetic factors, biologically based temperamental factors, parental 'modeling' of negative cognitive patterns, and parental feedback, to name a few. Recent research has shown that adults who exhibit marked cognitive vulnerability to depression report growing up in environments characterized by maltreatment and neglect. Emotional maltreatment may be an especially potent contributor to cognitive vulnerability. Telling a child that he or she is incompetent, unlovable, or unattractive may 'program' the child to make very depressogenic interpretations of negative events later in life that, in turn, lead to depression. Further research is necessary to establish definitively that emotional maltreatment contributes to cognitive vulnerability to depression.

COGNITIVE-BEHAVIOR THERAPY FOR DEPRESSION

Consistent with the cognitive perspective on depression, cognitive–behavioral therapy (CBT) for depression has been developed that targets the negative cognitions such as hopelessness that are hypothesized to precipitate the onset of depressive symptoms and episodes as well as the negative cognitive patterns such as the tendency to make stable, global, causal attributions for negative events that are hypothesized to provide vulnerability for depression. CBT has been shown to be successful in remediating current depressive episodes and compares favorably with antidepressant medications for all but the most severely depressed patients. Given that depression is a recurrent disorder, it is especially noteworthy that preliminary work suggests that CBT may have an enduring effect that decreases the risk of relapse and recurrence among formerly depressed people. Finally, initial studies suggest that administration

of CBT to currently nondepressed individuals at risk for depression can prevent the onset of first episodes of depression.

INTEGRATION OF COGNITIVE AND NEUROSCIENCE PERSPECTIVES ON DEPRESSION

Given the empirical support for the cognitive theories of depression and the success of cognitive therapy for depression, it is critical to integrate the cognitive approach with other successful approaches to depression. Much important work has been conducted demonstrating the fruitfulness of integrating cognitive and interpersonal approaches to depression. However, little has been done to integrate the cognitive perspective with biological approaches to depression.

Although much important work has been conducted on cognitive and biological vulnerability to depression, these two lines of research have proceeded in relative isolation from each other. Paving the way for an integration of cognitive and biological approaches to depression, recent research has begun to elucidate the neural circuitry involved in implementing the behavioral approach system (BAS) and the behavioral inhibition system (BIS) (Davidson *et al*., 2002). Specifically, activation of the left frontal cortex is a key component of the neural circuitry implementing BAS function, and activation of the right frontal cortex is involved in BIS function. Consistent with this perspective, unipolar depression, which reflects low approach motivation, is associated with relative low left frontal cortical activity.

What is the nature of the relationship between cognitive vulnerability to depression and biological vulnerability to this disorder as indexed by patterns of cerebral asymmetry? To the extent that hopelessness, the expectation to which cognitively vulnerable individuals are predisposed, may be particularly powerful in signaling a shutdown of approach motivation (inactive BAS), cognitively vulnerable individuals may be characterized by relative low left frontal cortical activity. Consistent with this view, there are initial indications that attributionally vulnerable individuals exhibit relative low left frontal hemispheric activation. More generally, pessimism may be associated with low left frontal cortical activity, whereas optimism may show the reverse pattern. An exciting line of work has just begun to document a relationship between cortical activity and pessimism/optimism. For example, the autobiographical memories of patients with left hemisphere lesions, particularly to frontal regions, are more negative than those of patients with right hemisphere lesions. Moreover, activation of the left hemisphere by behavioral methods results in more self-serving attributions and more optimistic expectancies for the future.

Integration of the cognitive and neural approaches to depression promises to be an exciting and important direction for further theory and research.

EVOLUTIONARY APPROACHES TO UNDERSTANDING DEPRESSION

From an evolutionary perspective, depression may appear paradoxical. Depressed people show deficits in many basic human 'instincts' such as the pursuit of pleasure, sexual drive, appetite, sleep, parenting behavior, goal striving, and desire to be with other people. Perhaps most paradoxical of all, in defiance of the 'law of survival', depressed people sometimes actively try to terminate their own lives by suicide.

Despite these paradoxes, depression may be understood from an evolutionary perspective. Drawing on the fact that depression is common, some investigators have suggested that in our evolutionary past, depression may have had adaptational significance (Neese, 2000). For example, depression may facilitate disengagement from the pursuit of an unattainable goal. Alternatively, depression may involve a 'conservation of energy' principle in which a person slows down to reduce depletion of energy in pursuit of an unattainable goal or ceases activity that is likely to be futile or expose the self to some harm, disease, attack, or situation that is even worse than the current one. A complete account of depression is likely to involve an evolutionary perspective.

Further Reading

Abramson LY, Alloy LB, Hankin BL, Haeffel GJ, MacCoon DG and Gibb BE (2002) Cognitive vulnerability–stress models of depression in a self-regulatory and psychobiological context. In: Gotlib IH and Hammen CL (eds) *Handbook of Depression*, pp. 268–294. New York, NY: Guilford Press.

Abramson LY, Alloy LB, Hogan ME, Whitehouse WG, Donovan P, Rose D, Panzarella C and Raniere D (1999) Cognitive vulnerability to depression: theory and evidence. *Journal of Cognitive Psychotherapy: An International Quarterly* **13**: 5–20.

Alloy LB, Abramson LY, Murray LA, Whitehouse WG and Hogan ME (1997) Self-referent information-processing in individuals at high and low cognitive risk for depression. *Cognition and Emotion* **11**: 539–568.

Beck AT (1987) Cognitive models of depression. *Journal of Cognitive Psychotherapy: An International Quarterly* **1**: 5–37.

Clark DA, Beck AT and Alford BA (1999) *Scientific Foundations of Cognitive Theory and Therapy of Depression*. Philadelphia, PA: John Wiley.

Davidson RJ, Pizzagalli D and Nitschke JB (2002) The representation and regulation of emotion in depression: perspectives from affective neuroscience. In: Gotlib IH and Hammen CL (eds) *Handbook of Depression*, pp. 219–244. New York, NY: Guilford Press.

Dykman BM, Abramson LY, Alloy LB and Hartlage S (1989) Processing of ambiguous feedback among depressed and nondepressed college students: schematic biases and their implications for depressive realism. *Journal of Personality and Social Psychology* **56**: 431–445.

Hollon SD, Haman KL and Brown LL (2002) Cognitive-behavioral treatment of depression. In: Gotlib IH and Hammen CL (eds) *Handbook of Depression*, pp. 383–403. New York, NY: Guilford Press.

Nesse RM (2000) Is depression an adaptation? *Archives of General Psychiatry* **57**: 14–20.

Depth Perception

Introductory article

Myron L Braunstein, University of California, Irvine, California, USA

CONTENTS

The three-dimensional world that is immediately and effortlessly perceived is a product of inferential mechanisms that rely on many different types of information. Among these are binocular disparity, motion parallax, texture gradients, linear perspective, shading, interposition, accommodation and convergence.

INTRODUCTION

Depth perception has been described as a paradox. We experience a three-dimensional (3D) world immediately and effortlessly, yet our visual information about the external world comes from light imaged on the two-dimensional retinal surfaces at the backs of our eyes. Our experience of the 3D world is unambiguous, yet the information in these images is inherently ambiguous (Figure 1). Before reviewing specific sources of information (cues) for depth perception, it will be useful to state some general principles underlying our ability to perceive a 3D world:

1. Information is available in the retinal images, and in the mechanisms controlling the muscles that change the convergence angle of the two eyes and the shapes of the lenses, that can be used to recover the 3D structure of the world.
2. This information is inherently ambiguous and must be interpreted using constraints – in a sense, biases based on knowledge about how the 3D environment is structured and about the structure of the visual system itself. In interpreting binocular disparities, for example, the assumption of a constant interocular distance would be a constraint. In interpreting image motion, an assumption of rigid 3D motion is a possible constraint.
3. This process of interpretation can be formally regarded as inductive inference. It is important to understand that the use of inductive inference based on prior knowledge does not mean that the perceiver engages in a conscious or even in an unconscious thought-like process. Instead, the inferential processes in depth perception can be implemented by 'hard-wired' biological mechanisms with 'knowledge' of the structure of the 3D environment built into these mechanisms through evolution. These inferential processes may not match the optimum processes that would be theoretically available to an 'ideal' observer. Instead, the human observer appears to use heuristic processes – efficient shortcuts that usually lead to a correct interpretation of the 3D environment, but which can fail when presented with unusual stimuli, leading to illusions.

Different types of information about the 3D world have been labeled as 'cues', based on the concept that there are a small number of ways in which picture-like images on the retinas are interpreted by the brain. These cues are sometimes thought of as involving separate modules that function independently until their outputs are combined to determine what is perceived. Most of what have been regarded as unitary cues, however, have turned out to involve more than one type of information or more than one type of processing, and important interactions among the cues suggest that they may not function as separate modules. The cue concept is a convenient way to summarize what is known about the information used in depth perception, but it should be recognized that this

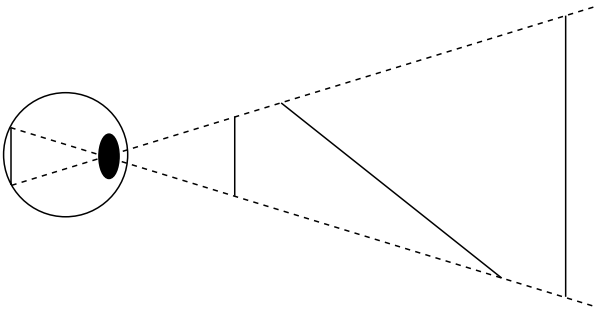


Figure 1. The ambiguity of the retinal image. An edge projected onto the retina is consistent with an infinite number of edges in three dimensions, varying in size, distance and orientation.

concept is a simplification that provides only a rough categorization of types of information used to infer a 3D world.

BINOCULAR DEPTH CUES

The two eyes are separated horizontally by about 6 cm. As a result of this separation the images projected onto the two retinas are different. The primary difference is that the relative positions of points in the images are shifted horizontally; this horizontal shift is a function of the interocular separation and the distances of the points from the eyes (Figure 2). It has been demonstrated using random dot stereograms (Figure 3) that horizontal disparities can lead to perceived variations in depth even in the absence of recognizable features in the individual images. A gradient of disparities, such as that produced by a slanted surface, is not, however, an effective depth stimulus. A discontinuity in disparities is needed, as when a scene contains an edge at which the depth or slant changes. Horizontal disparities appear to provide information about relative depth rather than absolute distance. Research suggests that vertical disparities in the two retinal images, when sufficiently large, can provide absolute distance information. Vertical disparities occur because one eye may be closer to an object, producing a different perspective projection. Another binocular cue is the convergence of the optic axes that occurs when the two eyes fixate a near object (also shown in Figure 2). Convergence is one of the first recognized depth cues, but is effective only at short distances.

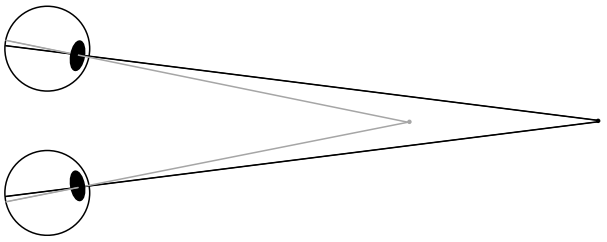


Figure 2. Binocular disparity and convergence. Note that the relative positions of the near and far points are reversed in the two retinal projections. The eyes are fixated on the near point and this determines the convergence angle.

MONOCULAR DEPTH CUES

Accommodation

Along with convergence, accommodation was one of the first recognized depth cues. When we fixate a nearby point the rays of light from this point diverge as they reach the lens of the eye and must be converged by the lens, so that a sharp point is imaged on the retina. This is accomplished by a change in the shape of the lens, which bulges to focus near objects and flattens to focus more distant objects (Figure 4). Information about the changes in the shape of the lens can therefore provide information about the distance of the fixated object. Accommodation is effective at short distances (within 2 m).

Monocular Perspective Cues

Perspective effects occur when the observation point (the eye or a camera) is relatively close to an object. The overall effect of perspective is that 3D distances closer to the observer project to larger

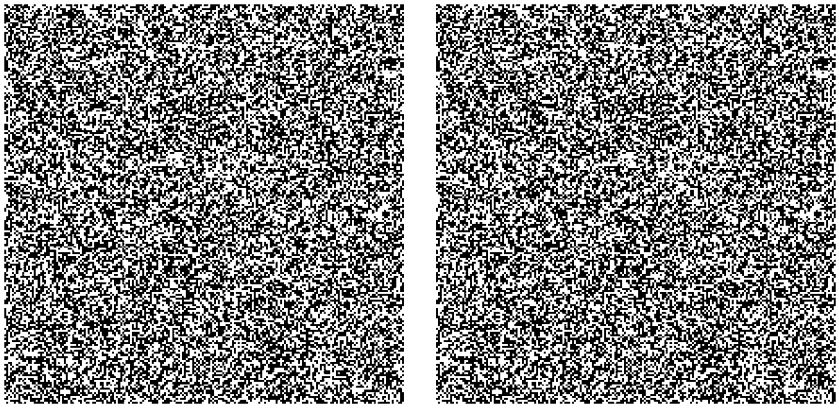


Figure 3. A random dot stereogram of the type introduced by Bela Julesz. If the images are fused by crossing the eyes, a small square should appear in front of the large square.

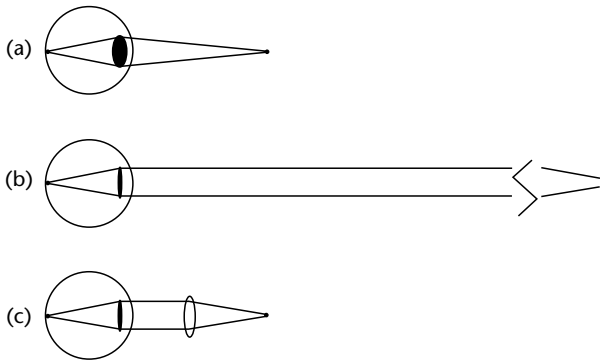


Figure 4. Accommodation to a near point (a) and far point (b). (c) Collimation: a lens is used to place a near point at optical infinity, making it appear to be a distant point.

distances in the retinal image than equivalent 3D distances further from the observer. This is reflected in several ways in a monocular image. Linear perspective, for example, refers to the convergence in the image of parallel lines extending into the distance in the 3D scene. It is a highly effective cue – merely drawing two converging lines can create an impression of a surface extending in depth. Other perspective cues are best described as gradients, as suggested by J. J. Gibson. The projection of a uniform 3D texture extending in depth results in a texture gradient. By itself, a texture gradient is not an especially effective cue unless the texture is regular (for example, a grid pattern), but texture gradients may combine with other information such as shading to provide an effective source of information about 3D layout. The projected sizes of similar objects describes another gradient that may result from perspective projection. This is closely related to the texture gradient. Usually a texture gradient describes spacing between texture elements (like blades of grass), whereas the size of similar objects describes the gradient in the projections of the elements themselves. Figure 5 shows several examples of monocular perspective cues.

Motion

Motion is not a single cue to depth but includes several ways in which changing patterns in the retinal projections over time provide information about 3D relationships. Motion parallax describes the changes that occur in the retinal image when there is relative motion between the eye and an object or a scene that is being observed. This usually occurs when the head moves relative to a stationary scene, but can also occur when an object or surface is moved relative to the head. It is primarily a perspective effect in that motions projected on the retina are more rapid for nearer objects than for more distant objects, assuming that the objects are moving in the same direction and at the same speed in 3D space. Another motion cue, structure from

Figure 5. Examples of perspective, shading and interposition. (a) The road illustrates linear perspective. (b) The rocks form a relative size gradient. (c) The arches form a regular texture gradient; specular highlights are found on the door at the back of the tunnel. Examples of interposition are found in all three photographs. (Images adapted from the Photo Library of the National Oceanic and Atmospheric Administration, US Department of Commerce.)

motion (also called the kinetic depth effect), occurs when an object is rotated relative to the direction of observation. Consider, for example, a cylinder rotating about a vertical axis. As the object rotates, the projections of imaginary lines connecting pairs of points on the cylinder will change in length (Figure 6). If we assume that the 3D distances between the points do not change as the object rotates (a rigidity assumption), the changing distances between points in the image can be used to compute the relative depths of the points on the 3D object. Unlike motion parallax, structure from motion does not depend on viewing distance and can provide information about the shape of a rotating object even when the object is at a great distance – for example, when an object is viewed through a telescope.

Shading

Consider a 3D shape and a light source illuminating that shape. At any point on this shape we can draw a tangent to the surface. The angle between this tangent and a line drawn from that point to the light source will vary over the surface and this will result in variation in the light reflected by points on the surface. The pattern of light reflected by the surface thus provides information about the shape of the surface and this source of depth information is referred to as 'shape from shading'. Note that shape from shading provides relative depth information within a surface, not information about distances from the eye. It is thus a source of object-centered depth information. Cast shadows and specular highlights also provide important information about object depth. Examples of shading cues are also seen in Figure 5. In dynamic scenes, important additional information about depth relationships is provided by changes over time in shading, in the positions of shadows, and in the locations of specular highlights.

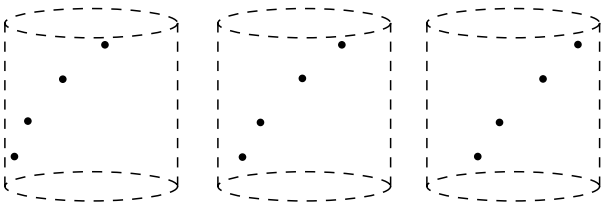


Figure 6. Structure from motion. Shimon Ullman showed that three-dimensional structure can be recovered mathematically from just three views of four non-coplanar points.

Occlusion

Occlusion is an unusual cue in that it provides only ordinal depth information. Occlusion of a far object by a near object, indicated by an interruption in the contours of the far object, is sometimes called interposition. Kinetic occlusion occurs when texture elements on one object disappear as they reach an implicit contour of another object, indicating that the first object is behind the second object.

Atmospheric Attenuation

Atmospheric attenuation refers to the reduced intensity and change in color of light reaching the eye that results from the scattering of light and absorption of light by particles in the atmosphere. Atmospheric effects can produce a gradient of intensity and color in the retinal projections that is informative about relative depth over large distances.

FLATNESS CUES

The retinal images should not be regarded as flat pictures that we can perceive as such in the absence of depth cues. To see a flat image we need information indicating that all points on the image are equally distant; this information is provided by flatness cues. Some flatness cues use the same sources of information as depth cues. For example, absence of differential accommodation to points on a near object indicates that these points are equally distant from the eye. In special applications, such as flight simulators, accommodation is removed as a flatness cue by collimating the image. Collimation, typically with a lens or a parabolic mirror, converts the diverging rays from a near object into parallel rays so that the visual system interprets the near object as a distant object (see Figure 4). Similarly, lack of binocular disparity within a near region suggests that the points within this region are equally distant. An especially interesting cue to flatness is a surrounding frame. Looking at a high-quality motion picture through a tube, which eliminates any explicit or implicit frame surrounding the image (and eliminates disparity if the tube is monocular) makes the picture appear less flat than if the frame were visible. Surrounding a real 3D scene with a frame can make that scene appear more flat, indicating that the frame is indeed a cue to flatness. What happens if there are neither cues to depth nor cues to flatness? This situation can be produced by creating a ganzfeld – a room with uniform illumination and no visible edges. The resulting perception is ambiguous. Usually a 3D fog is perceived.

CONCLUSION

The perceived structure of the 3D world is inferred from information available in the retinal projections and in the oculomotor system. There are many sources of information for depth which have been roughly categorized into 'cues', including binocular disparity, motion parallax, texture gradients, linear perspective, shading, interposition, accommodation and convergence.

Further Reading

Braunstein ML (1994) Decoding principles, heuristics and inference in visual perception. In: Jansson G, Bergström SS and Epstein W (eds) *Perceiving Events and Objects*. Hillsdale, NJ: Erlbaum.

Braunstein ML (1994) Structure from motion. In: Smith AT and Snowden RJ (eds) *Visual Detection of Motion*. New York: Academic Press.

Epstein W and Rogers S (eds) (1995) *Perception of Space and Motion*. San Diego, CA: Academic Press.

Gibson JJ (1950) *The Perception of the Visual World*. Boston, MA: Houghton Mifflin.

Pastore N (1971) *Selective History of Theories of Visual Perception: 1650–1950.* New York: Oxford University Press.

Developmental Disorders of Language

Intermediate article

April A Benasich, Rutgers University, Newark, New Jersey, USA
Jennifer J Thomas, Rutgers University, Newark, New Jersey, USA

CONTENTS

A developmental disorder of language is a significant delay or impairment in the expression and/or comprehension of language in the absence of a known cause.

INTRODUCTION

A developmental disorder of language is a significant delay or impairment in language acquisition (for expression and/or comprehension) in the absence of a known cause. Here we will use the term 'specific language impairment' (SLI) to refer to such language disorders. Specific language impairment (also called developmental dysphasia) is diagnosed on the basis of exclusion, meaning that the child's language difficulties cannot be accounted for by readily identifiable factors such as hearing impairment, neurological disease, psychiatric disability (disorders such as autism or childhood schizophrenia), low intelligence, physical malformation of the vocal apparatus, or severe environmental deprivation. Individuals with SLI exhibit normal nonverbal intelligence and, with the exception of language, have otherwise uncompromised development.

There is no universally accepted criterion for the classification of SLI. The World Health Organization *International Classification of Diseases* (ICD-10) classifies children with language abilities (assessed using a standardized language test) in the lowest 3% of the population as language-impaired (WHO, 1993). The American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) has similar diagnostic criteria to the ICD-10, but it also requires that a child's language difficulties interfere with normal, everyday activities – academic, occupational or social (APA, 1994). In addition, researchers sometimes create their own classification or selection criteria for studies of developmental language problems, resulting in a heterogeneous population with the diagnosis of SLI. There is also no general agreement regarding the classification of children with subtypes of SLI, categorizing different patterns of deficits in language skills. Based on psychometric evaluations of language ability, three broad subtypes have been defined: receptive only, expressive only and expressive-receptive mixed (Tomblin, 1996). These categories, reflective of the DSM-IV and ICD-10 classification systems, have been used in many research studies. However, some researchers use a subtype classification based on the work of Rapin and Allen (1987), in which clinical evaluation of a child's spontaneous language during play (including phonologic, syntactic, semantic and pragmatic skills) leads to assignment to one of six subgroups: verbal auditory agnosia, verbal dyspraxia, phonological programming deficit syndrome, phonological–syntactic deficit syndrome, lexical–syntactic deficit syndrome or semantic–pragmatic deficit syndrome. In sum, there are a number of subtypes of SLI that attempt to define precisely an individual's areas of difficulty.

Although there are advantages to using diagnostic criteria specific to the stated research or clinical objectives, consensus about the defining characteristics of SLI, and the classification of its different subtypes, is essential if there is to be meaningful comparison among the many research efforts investigating the etiological conundrum of whether SLI stems from a single common underlying mechanism, or has multiple genetic and or developmental origins.

The traditional view has been that the symptoms exhibited in SLI can be attributed to delays in the

learning of linguistic-specific semantic and syntactic rules, which are critical to the development of language. In addition, there is substantial support for the role of poor phonological processing (perceiving and discriminating phonemes, which are the smallest units of sound in language that alone can differentiate meaning) in children with both language and reading deficits. However, there is also strong evidence that differences in basic auditory processing abilities of children with SLI may be related to their language deficits: see Leonard (1998) and Tallal (2000) for reviews. Leonard (1998) comments that findings of disordered processing of brief or rapidly presented stimuli in children with SLI are both ubiquitous and consistent across laboratories, tasks, and stimulus variations. Thus, it has been posited that basic difficulties in processing the brief, rapid, successive auditory cues that constitute speech could impair or delay the formation of distinct (categorical) representations of the sounds of language (phonemes). This would have an impact on emerging language through a cascade in which each disordered step (beginning with the most basic sound processing) affects subsequent levels of language processing. According to this hypothesis, auditory perceptual deficits lead to weak representations of phonemes, and these weak phonological representations may lead to oral language disabilities, and subsequent reading, writing and/or spelling deficits due to poor phonographic (spoken) to orthographic (visual/written) transfer (Tallal, 2000). However, there is still much debate about the etiology of SLI – in particular, which of the deficits simply co-occur and which are causative (Mody *et al*., 1997; see also Denenberg, 1999, 2001).

Dyslexia is also considered to be a developmental disorder of language, although dyslexia is classically defined as the failure to develop age-appropriate reading ability in the presence of otherwise normal skills. Nevertheless, research has shown that approximately 80% of children classified as SLI go on to develop dyslexia, providing support for the view that developmental language and reading disorders may have a common etiology (APA, 1994). Individuals with SLI and dyslexia often exhibit similar behavioral deficits in rapid sensory processing, and neuroimaging studies have revealed alterations in neuroanatomical substrates common to these two disorders. Further, measurements of brain activation employing event-related potentials (ERPs) (electrical signals recorded from the scalp in response to external stimuli) of SLI and dyslexic individuals are comparable. Thus, it is likely that children with SLI and dyslexia form an overlapping, though not identical, population. For this reason, behavioral, neuroimaging and neuroanatomical studies of both SLI and dyslexia are discussed here.

NEURAL SUBSTRATES

The classic studies of the language areas in the brain described adult patients suffering from various types of aphasia (disturbances in language expression or comprehension) due to brain lesions sustained in adulthood. Disruption of language abilities in these patients was consistently linked to specific areas of damage in the brain, thus allowing a coarse topographic neural map of language function to be constructed. These landmark studies were the basis for the Wernicke–Geschwind model of language function. In this model, auditory information is first processed by the primary auditory cortex (A1; Brodmann's areas 41 and 42, located on Heschl's gyrus on the dorsal surface of the left temporal lobe), and is then sent to secondary auditory cortical areas: Wernicke's area (the left posterior perisylvian area, including Brodmann's area 22), shown to support language comprehension, and Broca's area (the inferior left temporal gyrus containing Brodmann's areas 44 and 45), which mediates speech production (Figure 1). These areas are linked by a number of subcortical fiber systems. The arcuate fasciculus, a fiber bundle that carries information from Wernicke's area to Broca's area, has been of particular interest as it has been implicated in subtypes of aphasia characterized by impairments of language expression.

These early studies were critical in laying the framework for subsequent research into the neural bases of language processing. This research has shown that there are many additional language centers in the brain and that these interact via complex neural networks. However, the precise mechanisms governing language function continue to be the subject of intense study. Many brain areas, including left hemisphere regions of temporal, parietal and frontal cortex, are the focus of contemporary research into the neural substrates of developmental language disorders.

In contrast to acquired language deficits such as aphasia, for which a specific brain lesion can often be identified, SLI and dyslexia (in which normal language skills fail to emerge in the context of otherwise normal development) present a far less clear picture of which brain areas might be disrupted. The search for the underlying neural substrates of these disorders has involved behavioral, neurophysiologic and neuroanatomical approaches

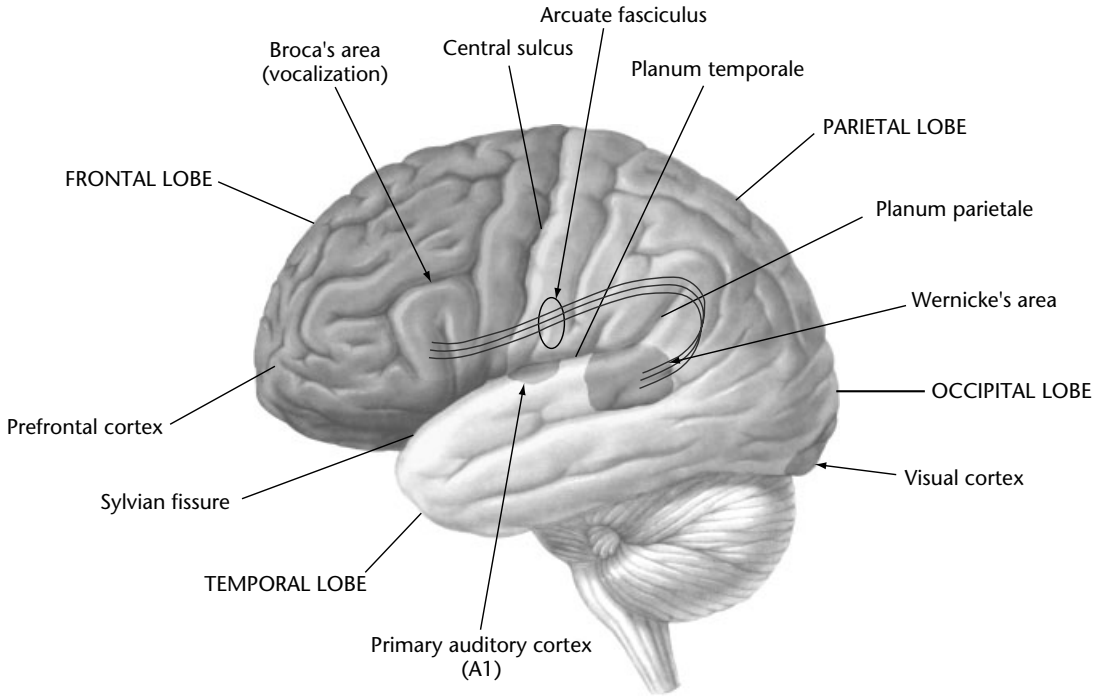


Figure 1. Areas of the human brain that are important for language, including the primary auditory cortex (A1), Wernicke's and Broca's areas, and the arcuate fasciculus.

and has been accelerated by methodological advances in functional neuroimaging and event-related potential (ERP) paradigms.

There is now a substantial literature reporting neurophysiological and neuroanatomic abnormalities in language-impaired children. Although there are generally concordant findings regarding differences in hemispheric asymmetry (see below), a number of studies report inconsistent or conflicting findings concerning other structural brain anomalies. These include alterations in corpus callosum size, extra sulci, white-matter abnormalities, and cortical atrophies. Such findings suggest that a single, focal anatomical abnormality will not be sufficient to explain SLI, and may also account for the variation in the severity, nature and perhaps type of language deficits that individuals with SLI exhibit.

A number of hypotheses have been proposed that attempt to link the structural abnormalities and behavioral deficits found in individuals with SLI and dyslexia. One is that neuroanatomical disruptions in SLI may be bilateral (affecting both hemispheres) rather than unilateral (affecting only one hemisphere), thus reducing the possibility for recovery of language function. Subcortical structures such as the caudate and thalamus might also be involved, and lesions in these areas have been shown to cause lasting and profound language impairments. Another possibility is that a specific disruption of critical language processing systems might be due to delay in brain maturation, induced genetically or by some prenatal or perinatal brain insult. Such a disruption would particularly affect the left hemisphere (given the normal time course of early brain development) and could cause structural brain abnormalities. The resulting deviant brain organization and related alterations in connectivity might disturb critical periods of language development, resulting in delays as well as lasting impairments. Finally, a basic sensory integration deficit in SLI has been posited. This basic deficit would induce dysfunction of neural systems which mediate rapid processing of dynamic sensory information important to rate processing.

LATERALITY

It is generally accepted that, in the normal population, left hemisphere structures are dominant for language function. Evidence for this first came from lesion and psychophysical studies, and with the advent of neuroimaging techniques such as positron emission tomography (PET) and magnetic resonance imaging (MRI) it has been shown that the left hemisphere of the brain is larger than the

right in the vast majority of normal people. Likewise, functional imaging studies (using PET and functional MRI) have revealed that language areas of the left hemisphere show more neural activity (reflected in cerebral blood flow or metabolic activity) during tasks that tap language processes, compared with the right hemisphere.

Interestingly, in individuals with dyslexia and SLI this asymmetry appears to be reversed or altered: right areas are larger or more active than left areas. For instance, differences have been found in the planum temporale, a structure located on the superior surface of the temporal lobe in the posterior perisylvian region (which includes a portion of Wernicke's area) that is important for language comprehension and phonological processing. The planum temporale is larger on the left than on the right in most people; however, in individuals with SLI and dyslexia this is not the case. In studies of dyslexia reversed asymmetry (right larger than left) of the planum temporale has been consistently reported, in both postmortem and neuroimaging studies. Evidence of structural abnormalities in the left hemisphere of individuals with dyslexia is also consistent with ERP studies reporting abnormal brain activation in language regions of the left hemisphere in this population. Other small, focal cellular abnormalities have also been reported in people with dyslexia, including ectopias (the presence of neurons in inappropriate cortical layers) and microgyria (abnormally small gyri). Such anomalies are thought to result from errors in neuronal migration caused by intrinsic genetic factors, or perhaps by prenatal or perinatal brain insult. Examination of the thalamus in dyslexics has also revealed cellular changes, suggesting that the processing of sensory information at this critical subcortical level may be compromised.

There have been few analogous findings in children with SLI owing to the fact that experiments involving imaging techniques are not often feasible in children (imaging studies in dyslexia usually recruit adult participants), and individuals with SLI are not routinely referred to medical professionals for brain scans. Additionally, there have been no postmortem studies of the brains of individuals with SLI, so it is impossible to confirm or rule out the presence of focal cellular abnormalities like those found in the sample of dyslexic individuals examined by Galaburda and colleagues (Galaburda *et al*., 1994). The similar behavioral deficits in SLI and dyslexia suggest a similar neuropathology, but this remains speculative. Despite these limitations, the studies of children with SLI to date are fairly similar to those of adults with SLI and dyslexia in reporting abnormal asymmetry, especially in the perisylvian region. In one MRI study, the volume of the posterior perisylvian region, including the planum temporale, was found to be reduced bilaterally, though more dramatically in the left hemisphere, in language- and learning-impaired participants compared with controls (Jernigan *et al*., 1991). Further, bilateral volumetric reductions were found in subcortical structures, including the caudate and putamen. A number of more recent MRI studies also report atypical asymmetry of perisylvian structures, with areas on the right larger than usual in individuals with SLI. In sum, neuroanatomical abnormalities were most often localized in the perisylvian region, which includes areas important for language processing, and atypical asymmetry in SLI compared with control subjects is consistently reported.

Functional imaging studies of individuals with SLI and dyslexia, using PET and fMRI, also report abnormal patterns of activation. Studies of children with language-related disorders report reduced activation (indexed by cerebral blood flow or glucose metabolism) in the perisylvian region in the left as compared with the right hemisphere during tasks requiring utilization of language areas. Such studies support the idea that the anomalies in anatomic asymmetry described above might be related to the inability of individuals with SLI to use these same areas effectively to process sounds. Brain activation studies of ERPs during visual and auditory sensory processing tasks add to these findings. Across studies, the ERP results support the premise that abnormal patterns of hemispheric activation can be documented in SLI populations.

Examination of results from studies of the brain's functional activity in response to rapidly presented and or brief acoustic cues (both speech and nonspeech) support the relevance of such cues to speech perception. This body of research further suggests that deficits in the underlying ability to process rapid acoustic change are often associated with abnormal patterns of brain activity during processing. It is important to note, however, that such relations are not always found. Some children with SLI do not exhibit abnormal asymmetry, and some individuals with verifiable atypical asymmetry exhibit no language difficulties. Moreover, individuals with SLI and dyslexia have been found to show abnormalities in brain areas outside of language centers. These deviations from the reported pattern of findings reinforce the view that SLI and dyslexia are disorders characterized

by heterogeneous behavioral profiles, and thus may ultimately show variability in the underlying neurobiological substrates.

RATE OF PROCESSING

Individuals with SLI or dyslexia exhibit deficits in the ability to process successive, brief, sensory stimuli which are rapidly presented to the nervous system. Such stimuli may be tactile, visual, somatosensory or acoustic, including linguistic stimuli such as consonant–vowel (CV) syllables, and nonlinguistic stimuli such as tone pairs. The perception of human speech requires the decoding of brief, rapidly changing auditory stimuli that constitute language. Words are made up of phonemes, and phonemes are characterized by formants – frequency patterns created by sound resonating in the vocal tract. Formants provide acoustic cues in the perception of speech and represent sound waveforms across time. Formants of vowels are more or less constant over time, whereas stop consonants (/p, b, t, d, k, g/) are characterized by formant transitions. During a formant transition, frequency position changes rapidly as the stop consonant occlusion (vocal tract is in a closed position specific to a stop consonant) is released and the vocal tract shape changes to form the subsequent vowel (a stop consonant cannot be uttered alone, but must be combined with a vowel). So, in order to discriminate accurately and perceive stop consonants, the rapid formant transitions must be correctly processed by the auditory system. Frequency spectrographs (sound waveforms) for the CV syllables /ba/ and /da/ are shown in Figure 2. Note that the only differences between these CV syllables occur within the first 40 ms of formants 1 and 2. After that point, the spectrographs for /ba/ and /da/ are almost identical. Thus, in order for a listener to distinguish these CV syllables, the rapidly changing frequency information contained in the initial 40 ms of the speech sound must be properly processed by the auditory system. Such brief auditory cues, critical to discrimination of language, enter the central nervous system in rapid succession, and an inability to process such transient, rapid, successive stimuli might ultimately result in deficits symptomatic of SLI and dyslexia. Indeed, in studies of adults with dyslexia, one of the most persistent deficits appears to be in phonological processing, suggesting that underlying auditory processing deficits are extremely long-lasting.

Studies of children with SLI provide further support for an underlying rapid auditory processing

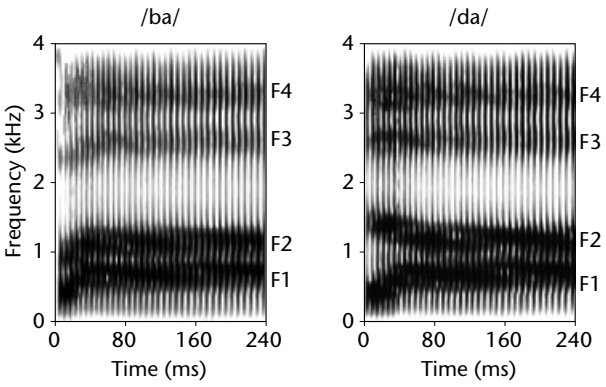


Figure 2. Spectrograph showing the formant transitions for the consonant–vowel (CV) syllables /ba/ and /da/. Notice the difference between /ba/ and /da/ in the 0–1 kHz range during the first 40 ms of formants 1 and 2. This brief sound difference must be detected in order to discriminate /ba/ from /da/. Formants 1 through 4 are noted on the right side of each spectrogram (F1–F4).

(RAP) deficit. Schoolage children with SLI were found to be impaired in processing both verbal and nonverbal stimuli that had brief and rapidly changing auditory components (Tallal *et al*., 1985). The replication and extension of these findings, across many laboratories, supported the hypothesis that an impairment of basic-level RAP hinders the development of normal language and reading abilities. Although children and adults with deficits in RAP and associated language problems hear normally and can sequence sounds, such individuals are selectively impaired in their ability to both perceive and produce speech sounds characterized by brief or rapidly changing temporal cues. These auditory processing limitations may directly interfere with adequate perception of those speech sounds that include rapid acoustic changes – such as stop consonants – and disrupt processing of the speech stream.

Another well-established finding is that SLI occurs within families and that infants born to families with affected close relatives are at greater risk of the disorder. Children from families with a history of SLI are approximately four times more likely to develop SLI than children from control families. These facts, and the insight that the cortical abnormalities implicated in SLI probably occur at an early stage of development as a result of errors in neuronal migration, suggest that different neurophysiological responses in infancy might predict SLI in later childhood. For example, it is thought that the mechanism responsible for causing malformations such as ectopias and microgyria

occurs during prenatal development, approximately between 18 weeks and 24 weeks of gestation. An insult to the prenatal brain or a genetic anomaly could initiate a cascade of events in the developing brain, ultimately resulting in the sensory processing profile characteristic of SLI associated with the underlying abnormal neurobiology described. If this hypothesis is valid, then RAP deficits should be observable in preverbal infants, and would predict subsequent language delay or impairments. Studies suggest that this may indeed be the case.

The links between infants' ability to process brief, rapidly presented stimuli and later language development has been examined in a series of studies by Benasich and colleagues, who found that infants with a family history of SLI performed more poorly on measures of RAP compared with a control group. In a longitudinal follow-up, it was found that the RAP thresholds measured in infancy were strongly related to later language comprehension and production through 36 months of age. Moreover, there are also indications that these basic acoustic processing skills in infancy (e.g. auditory gap detection and RAP) are predictive of later language development even in infants who are not at higher risk for language disorders. Such findings provide support for the notion that poor processing of rapidly presented and/or brief acoustic cues in infancy might be diagnostic of later language delays and impairments (for a review, see Benasich, 1998).

TEMPORAL INTEGRATION

It has been suggested that functional impairments in children with SLI might arise from basic sensory integration deficits, reflecting in turn dysfunctional neural processing of rapidly changing sensory information. So how might temporal integration deficits be propagated within the auditory processing pathways? Within the central auditory system, each neural stage of processing has a unique set of acoustic filtering and temporal integration properties. Thus, the critical linguistic cues of speech (phonemes) are first extracted from the raw acoustic waveforms produced by the vocal apparatus, and then processed by multiple neural stations (Figure 3). At the level of the auditory nerve, which carries sound information from the inner ear to the central nervous system, temporal information is probably encoded millisecond by millisecond. The acoustic information is then processed by the medial geniculate nucleus (MGN) of the thalamus, where the high resolution of the signal carried by the auditory nerve is largely preserved. The first cortical processing area is the primary auditory cortex (A1). At the level of the thalamus and A1, neural responses seem to occur primarily to the onset of temporal change. After A1, the

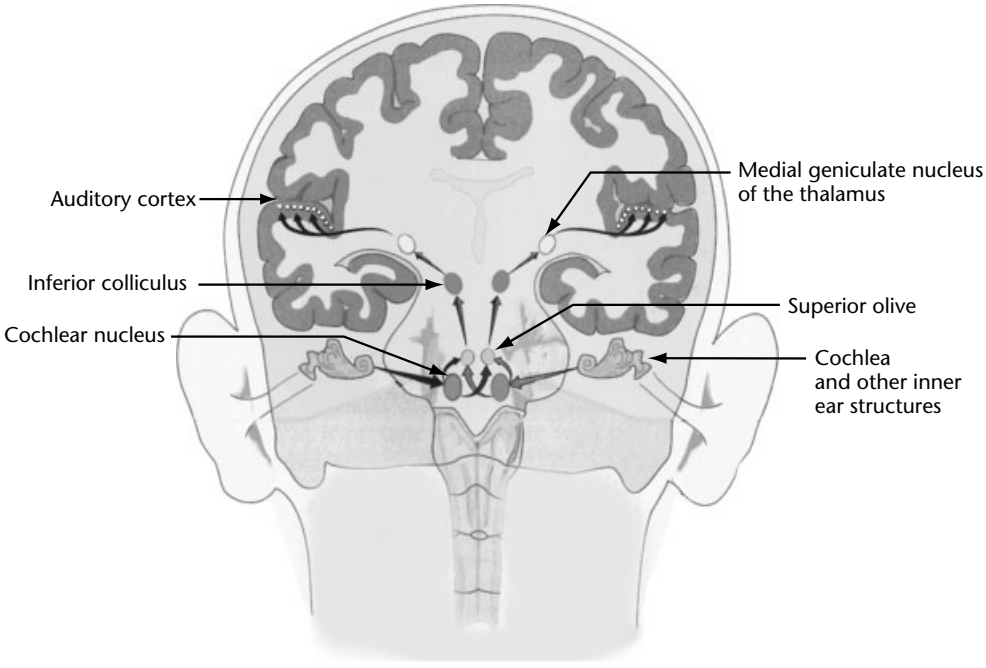


Figure 3. The ascending auditory processing pathways, as seen on a midcoronal section at about a 70° plane.

acoustic information is further processed by secondary and association auditory areas, which appear to respond to increasingly 'segmented' components of the original temporal information. Thus, each area where acoustic information is processed may be specialized for extracting different temporal features from a given sound. This may explain how a deficit at the most basic level of auditory processing could perturb the entire system. The subcortical structures implicated in SLI and dyslexia, such as the thalamus, are likely to be critical for encoding rapid temporal transitions (5–40 ms) embedded within longer acoustic sequences such as the speech stream. Therefore, if the rapid temporal integration capabilities of the thalamus are compromised (information is poorly encoded or incomplete), the cortical representations of the temporal features in a sound would degrade radically. Thus, the key elements of a neurobiological model in which auditory temporal processing defects lead, in a cascading fashion, to deficits in speech and language processing would be in place.

MULTIMODAL VERSUS AMODAL

Although SLI and dyslexia are defined as specific disorders of language in the presence of otherwise normal development (i.e. normal intelligence), there are reasons to question these diagnostic criteria as only applying to the auditory system. Research has shown that individuals with developmental language disorders exhibit processing deficits in other modalities, including motor coordination, and tactile and visual perception; see Tallal *et al*. (1985) for a review. It should be noted, however, that it has been proposed that some of these deficits seen in children with SLI and dyslexia might be the result of poor attention span (not explicitly assessed by traditional intelligence tests), and indeed quite a high proportion of these children meet the criteria for attention deficit disorder. To address this issue, future research must control for or eliminate variations in sustained attention in sample populations. Despite this caveat, there is accumulating evidence in support of the idea of a multimodal rapid sensory processing impairment in individuals with developmental language disorders.

Of particular interest are studies of visual processing that have revealed that the ability to process fast, low-contrast visual stimuli is impaired in people with dyslexia. The system responsible for processing this type of information is called the magnocellular pathway of the visual system. In contrast, no difference was found between participants with dyslexia and control participants in parvocellular visual processing. The parvocellular pathway of the visual system is responsible for processing high-contrast, fine detail, and color information. The magnocellular deficit is evident in behavioral assessment, and is also reflected in neurophysiological measurements: ERPs to rapidly presented visual stimuli have longer latencies in dyslexia, reflecting a slower neural response. Furthermore, postmortem analyses of the brains of individuals with dyslexia have revealed changes in the magnocells (larger cells which have faster conduction velocities than parvocells) in the lateral geniculate nucleus, the visual nucleus of the thalamus. Magnocells in dyslexic brains were abnormally small in comparison with controls. This convergence of behavioral, anatomical and electrophysiological evidence led to suggestions of a visual processing disturbance in dyslexia in the magnocellular system (Lehmkuhle *et al*., 1993; Galaburda *et al*., 1994).

Studies by Talcott, Witton and colleagues have demonstrated that individuals with dyslexia are impaired in processing dynamic visual and auditory events, and also have elevated thresholds for high-frequency tactile stimuli (Talcott *et al*., 2000). These findings, together with the results of physiological investigations (ERP and postmortem studies) support the possibility of a multisensory deficit in rapid processing in individuals with dyslexia. Future experiments with SLI populations may help to characterize potential multimodal temporal processing impairments in developmental language disorders.

CONCLUSION

Much is still unknown about the developmental disorders of language, despite years of research. Debate continues as to whether disorders such as SLI and dyslexia result from deficits in basic sensory processing that affect developing language, or whether the difficulty is primarily one of acquisition of higher-level semantic, syntactic and phonological skills. However, with the focus on atypical or inefficient processing of basic acoustic input, progress has been made in delineating potential causative mechanisms. One scenario that could lead to such effects involves a confluence of genetic and/or environmental factors, which may occur in the perinatal period during key periods of neural development, disrupting the normal cascade of events. Atypical asymmetry, as compared with control subjects, is consistently reported in SLI

and dyslexic groups, as well as a higher incidence of neuroanatomical abnormalities in areas important to language processing. A substantial number of contemporary functional imaging studies report reduced or abnormal patterns of activation in language areas in these populations. Moreover, close examination of the results of such studies supports the idea that processing rapidly presented and/or brief acoustic cues (both speech and nonspeech) is impaired in individuals with developmental language disorders.

Many questions remain unanswered regarding the developmental events and underlying pathological changes that lead to the specific behavioral deficits that are the hallmark of developmental disabilities. Nevertheless, it is possible that the neural characterization of these specific developmental disabilities, in combination with known behavioral profiles, can be used to gain some insight into their etiology. Unfortunately, studies that might link the developmental course of neural anomalies to expressed cognitive deficits in a causal fashion are exceedingly difficult to perform in humans. Furthermore, the nature of higher-order cognitive deficits (e.g. in language and reading) has not lent itself to study in nonhuman models, though such models have been useful in anatomical and basic sensory processing experiments. One promising area of investigation is prospective longitudinal studies of infants at high risk of SLI and dyslexia. This approach may allow the concomitant deficits in sensory processing, phonological processing and attention seen in children with SLI to be parsed. In addition, the emergence of more sensitive and sophisticated neuroimaging technology (both functional and anatomical) promises to provide additional evidence about the deficits implicated in developmental disorders of language. A better understanding of the developmental disorders of language will make the ultimate goal of remediation more attainable.

References

[APA] American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Benasich AA (1998) Temporal integration as an early predictor of speech and language development. In: von Euler C, Lundberg I and Llinás R (eds) *Basic Mechanisms In Cognition And Language – With Special Reference To Phonological Problems In Dyslexia* (Wenner-Gren International Series, vol. 70), pp. 123–142. Oxford, UK: Elsevier Science.

Denenberg VH (1999) A critique of Mody, Studdert-Kennedy, and Brady's 'Speech perception deficits in poor readers: Auditory processing or phonological coding?' *Journal of Learning Disabilities* **32**: 379–383.

Denenberg VH (2001) More power to them – statistically that is: a commentary on Studdert-Kennedy, Mody, and Brady's criticism of a critique. *Journal of Learning Disabilities* **34**: 299–301.

Galaburda AM, Menard MT and Rosen GD (1994) Evidence for aberrant auditory anatomy in developmental dyslexia. *Proceedings of the National Academy of Sciences USA* **91**: 8010–8013.

Jernigan TL, Hesselink JR, Sowell E and Tallal PA (1991) Cerebral structure on magnetic resonance imaging in language- and learning-impaired children. *Archives of Neurology* **48**: 539–545.

Lehmkuhle S, Garzia RP, Turner L, Hash T and Baro JA (1993) A defective visual pathway in children with reading disability. *New England Journal of Medicine* **328**(14): 989–996.

Leonard LB (1998) *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.

Mody M, Studdert-Kennedy M and Brady S (1997) Speech perception deficits in poor readers: auditory processing or phonological coding? *Journal of Experimental Psychology* **64**(2): 199–231.

Rapin I and Allen D (1987) Developmental dysphasia and autism in preschool children: characteristics and subtypes. In: Martin J, Martin P, Fletcher P, Grunwell P and Hall D (eds) *Proceedings of the First International Symposium on Specific Speech and Language Disorders in Children*, pp. 20–35. London, UK: AFASIC.

Talcott JB, Witton C, McClean M *et al*. (2000) Dynamic sensory sensitivity and children's word decoding skills. *Proceedings of the National Academy of Sciences USA* **97**: 2952–2957.

Tallal P (2000) Experimental studies of language learning impairments: from research to remediation. In: Bishop DVM and Leonard LB (eds) *Speech and Language Impairments in Children: Causes, Characteristics, Intervention and Outcome*, pp. 131–155. Philadelphia, PA: Psychology Press.

Tallal P, Stark R and Mellits D (1985) Relationship between auditory temporal analysis and receptive language development: evidence from studies of developmental language disorders. *Neuropsychologia* **23**: 527–536.

Tomblin JB (1996) Genetic and environmental contributions to the risk for specific language impairment. In: Rice ML (ed.) *Toward a Genetics of*

Language, pp. 191–211. Mahwah, NJ: Lawrence Erlbaum.

WHO (1993) *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: World Health Organization.

Further Reading

Fitch RH, Read H and Benasich AA (2001) Neurophysiology of speech perception in normal and impaired systems. In: Jahn A and Santos-Sacchi J (eds) *Physiology of The Ear*, 2nd edn, pp. 651–672. San Diego, CA: Singular Publishing.

Benasich AA and Spitz RV (1998) Insights from infants: temporal processing abilities and genetics contribute to language development. In: Willems G and Whitmore K (eds) *A Neurodevelopmental Approach to Specific Learning Disorders*, pp. 191–210. London, UK: MacKeith Press.

Bishop DVM, Bishop SJ, Bright P *et al*. (1999) Different origin of auditory and phonological processing problems in children with language impairment: evidence from a twin study. *Journal of Speech, Language, and Hearing Research* **42**: 155–168.

Eden GF, VanMeter JW, Rumsey J *et al*. (1996) Abnormal processing of visual motion in dyslexia revealed by functional brain imaging. *Nature* **382**: 66–69.

Fitch RH, Miller S and Tallal P (1997) Neurobiology of speech perception. *Annual Review of Neuroscience* **20**: 331–353.

Kraus N, McGee TJ, Carrell TD *et al*. (1996) Auditory neurophysiologic responses and discrimination deficits in children with learning problems. *Science* **273**: 971–973.

Livingstone MS, Rosen GD, Drislane FW and Galaburda AM (1991) Physiological and anatomical evidence for a magnocellular defect in developmental dyslexia. *Proceedings of the National Academy of Sciences USA* **88**: 7943–7947.

Tallal P, Merzenich MM, Miller S and Jenkins W (1998) Language learning impairments: integrating basic science, technology, and remediation. *Experimental Brain Research* **123**: 210–219.

Tomblin JB (1996) Genetic and environmental contributions to the risk for specific language impairment. In: Rice ML (ed.) *Toward a Genetics of Language*. Mahwah, NJ: Lawrence Erlbaum.

Dissociation Methodology

Intermediate article

Mark G Packard, Yale University, New Haven, Connecticut, USA

CONTENTS

Dissociation methodology is an experimental design tool used to examine functional dichotomies and independence between and within various domains of psychological function.

INTRODUCTION

The brain is capable of perceiving electrical messages transmitted by the optic nerve as visual information, and those by the auditory nerve as sound information. The nineteenth-century German physiologist Johannes Müller (1801–1858) proposed the 'doctrine of specific nerve energies' to account for these observations. This hypothesis essentially holds that independent neural pathways determine the perceived quality of sensory information, and provides a basis for understanding the empirical data indicating that deaf individuals can possess normal eyesight and blind individuals can possess normal hearing. These findings also provide evidence of a dissociation of the neural circuitry underlying these two sensory capabilities, at least at some fundamental level of psychological processing.

In the psychological sciences, dissociation methodology has been used extensively as an experimental design tool to investigate possible functional dichotomies that might shape understanding of fundamental laws in various domains of cognition and behavior, such as language, memory, and emotion. An early example of the use of this methodology is provided by the discovery of the different roles played by Broca's area and Wernicke's area in language expression and reception, respectively.

In the field of cognitive neuroscience, scientists often discover evidence of a functional dichotomy at the neuroanatomical level, and speculate as to the differences in operating principles that best relate this anatomical (neural level) dissociation to behavior or cognition (psychological level).

Several functional dichotomies have been proposed in different areas of psychological research, and each has influenced theoretical debate in these areas. These include, for example, dissociations between object recognition and localization visual pathways, explicit versus implicit memory, conscious versus unconscious perception in 'blindsight', or effortful versus automatic attention. The development of each of these hypotheses, and several others on their respective topics, relies on dissociation methodology.

SINGLE VERSUS DOUBLE DISSOCIATIONS

A description of the types of dissociations that may be observed in a particular experiment can be readily understood using the following scenario. Consider a black box with two levers (X and Y) protruding from the sides, and containing a latched top and a sliding front drawer. When lever X is pulled, the top of the box flips open (box operation A). However, pulling lever X does not effect the operation of the box drawer. In attempting to elucidate the functions of lever X, this manipulation has produced a single dissociation: pulling lever X effects box operation A but not B. Next, when lever Y is pulled, the drawer in the front of the box slides open (box operation B). However, pulling lever Y does not effect the operation of the box top. These manipulations have produced a double dissociation, in which pulling lever X influences box function A but not B, and pulling lever Y influences box function B but not A. If the black box used in this example represents a mammalian brain, and the two levers represent different brain structures, then one might discover that selective damage to lever X impairs box function A but not B, and selective damage to lever Y impairs box function B but not A.

The double dissociation methodology used above illustrates a fundamental approach for examining the hypothesis that independent neural and/or psychological functions exist. Hans-Lukas Teuber (1955) originally coined the term 'double

dissociation', and provided early arguments for the importance of this pattern of results in supporting the hypothesis of functional independence between neural structures. Teuber noted that neuropsychological studies employing dissociation methodology often reveal only a single dissociation (e.g., damage to structure X impairs performance of A but not B). Although it is tempting to conclude that A and B are functionally independent based on a single dissociation, it is also possible that their operation may be arranged in a hierarchical fashion. Clearly, if the primary goal of dissociation methodology is to provide information concerning the functional independence of various psychological and/or neural processes, then a double dissociation is necessary for the strongest version of such a hypothesis to be advanced.

Within the broader field of cognitive science, examples of double dissociations exist from research in traditional neuropsychology (the study of individuals who have suffered brain damage through injury or disease), more recent studies employing brain imaging techniques and interference tasks in normal humans, as well as research in nonhuman experimental animals. Research on the neurobiology of learning and memory can be considered an exemplar of the historical use of dissociation methodology in the psychological sciences, as this experimental tool has been used effectively across four decades of research in this area. Of course, the logic behind dissociation methodology and the use of this approach are not unique to memory research.

DISSOCIATION METHODOLOGY

Examples from Human Neuropsychology

A double dissociation of the roles of two cortical regions in facial identity illustrates the usefulness of dissociation methodology for providing information about the organization of a particular psychological function. Patients with bilateral damage of the ventromedial frontal cortex are able to recognize the identity of familiar faces in a normal fashion, but are unable to generate a discriminatory skin conductance response (SCR) to the same faces. In contrast, patients with bilateral occipitotemporal cortical damage display impaired identity recognition yet can generate discriminatory SCRs to familiar faces. These findings have been interpreted to suggest that a functional independence exists between the neural circuitry that processes somatic-based 'valence' and nonsomatic-based 'factual' information in this process. Note that interpretation of the psychological operating principles that might account for the anatomical double dissociation observed in any behavioral study ultimately involves debate over theoretical constructs such as emotion and memory, and often relies to some degree on *a priori* assumptions that are made at the time the behavioral tasks are developed. Nonetheless, an empirical double dissociation on task performance in humans with brain damage compromising different neuroanatomical structures provides the necessary evidence for postulating the existence of functionally independent systems.

Neuropsychological studies have also revealed double dissociations in performance of brain-damaged humans on pairs of memory tasks; patients with limbic-diencephalic damage acquire a probabilistic learning task normally, but memory for the training episode as assessed in an interview questionnaire is severely impaired. In contrast, patients with Parkinson disease are severely impaired in acquisition of the probabilistic task, yet demonstrate normal memory for the training episode. These findings have been interpreted as evidence of a double dissociation between the role of limbic-diencephalic and neostriatal brain regions in declarative memory and habit learning, respectively (Knowlton *et al*., 1996). Again, separate theoretical accounts of the critical differences in the psychological operating principles that underlie the dissociations observed in performance of different learning and memory tasks have been offered and debated, and several 'dual-memory' theories describing different sets of principles exist. However, it is the use of dissociation methodology that has ultimately driven the development and refinement of these theories.

Imaging the Living Human Brain

The advent of technologies for imaging the living human brain has provided an extraordinary avenue for the use of dissociation methodology to elucidate the organization of various psychological functions. Neuroimaging studies have revealed double dissociations in the patterns of brain activity associated with performance of different tasks. For example, functional magnetic resonance imaging has been used to demonstrate a double dissociation in levels of activation of the left perisylvian cortex and the dorsolateral prefrontal cortex in working memory processes. Based on these findings, the former structure has been proposed to mediate working memory storage

processes, and the latter to mediate executive control processes contributing to accurate working memory (Postle *et al*., 1999). Other brain imaging research using injections of radiolabeled water in combination with positron emission tomography to measure regional cerebral blood flow has revealed dissociable roles of the ventral and dorsal human extrastriate cortex in object and spatial visual processing, respectively, consistent with a prominent dissociation in these visual pathways or processes originally developed in research with nonhuman primates (Haxby *et al*., 1991).

Research in Experimental Animals

Research in experimental animals (e.g. nonhuman primates and rats), in which brain manipulations can be performed in a localized manner, provides further examples of empirical double dissociations. In monkeys, lesions of the medial temporal lobe impair performance of a delayed nonmatch to sample task, but do not affect acquisition of a concurrent visual discrimination (Mishkin and Petri, 1984). In contrast, lesions of the ventrocaudal neostriatum impair concurrent discrimination learning, but do not affect performance of delayed nonmatch to sample behavior (Fernandez-Ruiz *et al.*, 2001). In rats, a double dissociation between the roles of the hippocampal system and caudate nucleus in acquisition of spatial and visual discrimination tasks, respectively, has been observed following lesions of these two structures (Packard *et al.*, 1989).

The most compelling experimental design using dissociation methodology is one in which as many domains of psychological processes as possible can be equated across different tasks. An important feature of the pairs of tasks used in the rat study is that they each involved similar motor (maze running), sensory (use of visual cues), and motivational (appetitive) characteristics, and were hypothesized to differ primarily in mnemonic requirements.

The hypothesis of functional independence between brain areas in these memory tasks is strengthened by converging evidence from manipulations that affect neurochemical mechanisms. For example, localized intracerebral injections of drugs affecting various neurotransmitter systems have been used to doubly dissociate the roles of different brain regions in memory consolidation. Unlike traditional lesion techniques such as the use of irreversible or reversible lesions in which dissociations are revealed as an impairment in performance, drug injections have been used in these same brain regions to produce double dissociations using memory-enhancing agents (Packard and White, 1991).

CONCLUSION

Dissociation methodology has been used extensively to gather evidence supporting the possible existence of functional dichotomies in various psychological domains. The examples of double dissociations provided in this brief article illustrate the use of this methodology in developing theories of memory organization in the mammalian brain. Historically, the use of dissociation methodology has not been limited to this research topic, and important examples exist of dissociations in tasks measuring perceptual, attentional, and motivational processes, as well as complex language and numerical capabilities in humans.

An important goal of dissociation methodology is to provide evidence of functional dichotomies. However, as dissociation methodology has gained widespread use in psychological science, several scientists have stressed that such conclusions must be made cautiously, and others have raised reasonable concern over the use of dissociation methodology to develop a potentially endless taxonomic classification of various psychological functions. As considered earlier, in view of hierarchical relationships a single dissociation can never be interpreted as providing unequivocal evidence of functional independence. Moreover, even a double dissociation may not be an entirely sufficient condition for offering such a conclusion (Weiskrantz, 1968). For example, the observation of an empirical double dissociation of the effects of damage to different brain structures (X and Y) on performance in two different memory tasks (A and B) could conceivably be influenced by manipulation of a particular parametric setting such as task demand (e.g. delay interval or memory load; Olton, 1989). In this case, rather than posit a strong functional independence between the roles of brain structures X and Y in task performance, one might postulate that a third factor influencing performance on both tasks may be interacting differentially with each structure. Analyses of parametric settings in experiments using dissociation methodology are likely to be of particular importance in the psychological sciences, and may provide constraints on the number of fundamental dichotomies that are ultimately deemed dissociable at a functional level.

As dissociation methodology continues to find widespread use in brain imaging research, it should be noted that there are caveats in interpretation of data obtained in studies measuring

localized increases in brain activity. For example, in some imaging studies activation of particular brain regions has been observed in tasks for which there is neuropsychological data indicating that damage to these same brain areas does not impair task performance. Thus, dissociations between the roles of different brain structures in behavior may be revealed in a neuropsychological study when brain damage is present, but not observed when imaging techniques are applied to the intact human brain. Conclusions about the necessary versus sufficient roles of different brain structures implicated in a dissociation study using neuroimaging techniques may be difficult, and converging evidence from neuropsychological testing in individuals with brain damage is likely to be important for interpretation of the ultimate functional significance of the dissociations observed. A similar caution is necessary in interpreting dissociations at the functional level using electrophysiological methodology, or more recently developed molecular science techniques that use gene expression as for a marker of nerve cell activation (e.g. c-*fos* activation). Nonetheless, if the functional dichotomies proposed to underlie the double dissociations observed at the neural and/or psychological level prove predictive across experimental settings using converging techniques, then the promise of dissociation methodology will continue to be the possible discovery of fundamental scientific laws.

References

Fernandez-Ruiz J, Wang J, Aigner TG and Mishkin M (2001) Visual habit formation in monkeys with neurotoxic lesions of the ventrocaudal neostriatum. *Proceedings of the National Academy of Sciences of the USA* **98**: 4196–4201.

Haxby JV, Grady CL, Horwitz B *et al*. (1991) Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the USA* **88**: 1621–1625.

Knowlton BJ, Mangels JA and Squire LR (1996) A neostriatal habit learning system in humans. *Science* **273**: 1399–1402.

Mishkin M and Petri HL (1984) Memories and habits: some implications for the analysis of learning and retention. In: Squire LR and Butters N (eds) *Neuropsychology of Memory*, pp. 287–296. New York, NY: Guilford.

Olton DS (1989) Inferring psychological dissociations from experimental dissociations: the temporal context of episodic memory. In: Roediger HL and Craik FIM (eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*, pp. 161–174. Mahwah, NJ: Lawrence Erlbaum.

Packard MG and White NM (1991) Dissociation of hippocampus and caudate nucleus memory systems by pasttraining intracerebral injection of dopamine agonists. *Behavioral Neuroscience* **105**: 73–84.

Packard MG, Hirsh R and White NM (1989) Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *Journal of Neuroscience* **9**: 1465–1472.

Postle BR, Berger JS and D'Esposito M (1999) Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences of the USA* **96**: 12959–12964.

Teuber HL (1955) Physiological psychology. *Annual Review of Psychology* **6**: 267–296.

Weiskrantz L (1968) Some traps and pontifications. In: Weiskrantz L (ed.) *Analysis of Behavioral Change*, pp. 415–429. New York, NY: Harper & Row.

Further Reading

Crowder RG (1989) Modularity and dissociations in memory systems. In: Roediger HL and Craik FIM (eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*, pp. 271–294. Mahwah, NJ: Lawrence Erlbaum.

Dunn JC and Kirsner K (1988) Discovering functionally independent mental processes: the principle of reversed association. *Psychological Review* **95**: 91–101.

Schallice T (1988) *From Neuropsychology to Mental Structure*. Cambridge, UK: Cambridge University Press.

Weiskrantz L (1991) Dissociations and associates in neuropsychology. In: Lister RG and Weingarter HJ (eds) *Perspectives on Cognitive Neuroscience*, pp. 157–164. New York, NY: Oxford University Press.

Down Syndrome

Introductory article

Ira T Lott, University of California, Irvine, California, USA

CONTENTS

Down syndrome is a chromosomal disorder characterized by specific dysmorphic features, organ malformations, and variable but often severe learning difficulties.

INTRODUCTION

Down syndrome (DS) is the most common known form of developmental disability. First described by Langdon Down in the *London Hospital Reports* in 1886, the disorder was given scientific focus by Lejeune who identified the chromosomal abnormality in 1960. Despite the mental and physical handicap invariably associated with the disorder, people with Down syndrome have been prominently represented in many areas of daily life, including sports and the entertainment industry.

INCIDENCE AND GENETIC BASIS

Down syndrome (DS) is a genetic disorder that occurs once in every 600–1000 live births and affects all races and both sexes equally. The presence of additional chromosomal material has been shown to be responsible for the condition. The vast majority of individuals with DS have received the extra chromosomal material from an error involving the separation of chromosomes during the cell divisional process of meiosis. During normal meiosis an originator cell containing two copies of each of the 23 chromosomes is divided into daughter cells, each containing a single copy of each chromosome. In DS an error in this process, nondisjunction, results in a sperm or egg cell containing two copies of chromosome 21 instead of the usual single copy. When this sperm or egg cell pairs at fertilization with the other normal complement containing a single copy of chromosome 21, the resulting fertilized egg has three copies of chromosome 21 instead of the usual two – hence the alternative name for the syndrome, 'trisomy 21'. Nondisjunction is much more likely to occur in maternal than paternal meiosis. The actual cause of this error is unknown but it is more likely to occur in older women.

In about 4% of cases the extra copy of chromosome 21 is the result of the attachment of chromosome 21 onto another chromosome. This type of DS is called 'translocation DS' and unlike the disjunction error it occurs equally in males or females and is not age-dependent. Interestingly, one cannot determine by appearance or physical examination whether the condition is a result of a nondisjunction or a translocation. The third mechanism discovered for DS involves a genetic error in which there is a coexistence of normal and trisomic cell lines. This type of DS is called 'mosaic'. It has been argued that people with the mosaic form of DS have a less severe expression of the disorder since not all of the cells are trisomic. However, identifying mosaic DS based on its expression is difficult for even the most experienced clinicians.

It is not clear how the extra genetic material in DS accounts for the spectrum of clinical abnormalities in the disorder. The imbalance in genetic material appears to cause a pertubation in the timing of developmental sequences which eventually results in the clinical expression – phenotype – of the condition. Some individuals diagnosed with DS have only a tiny portion of chromosome 21 existing in triplicate. This small portion is enough to cause the phenotypic expression common to DS, and has thus been identified as the Down syndrome critical region.

CLINICAL SIGNS AND SYMPTOMS

Every physical feature of DS can be seen in a small percentage of the normal population. It is only when the features are combined into a single

individual that one has the appearance or the phenotype of DS. Although Langdon Down became confused about the cause of DS, he was the first to describe the physical characteristics of the disorder. The middle portion of the face is flat. There is an upward slant of the eye sockets (or palpebral fissures). The irises are speckled (Brushfield spots). The hands and feet are short and broad. There is an incurving of the fifth finger (clinodactyly) and there is a separation between the first and second toe. The patterns of the skin creases in the palm impart a particular appearance to the hand of a person with DS, and the major crease across the palm harkens back to an earlier evolutionary appearance (simian crease). The tongue is thickened and this is one of the factors contributing to a lack of clarity in the speech of people with DS.

People with DS are subject to certain illnesses involving organ systems other than the brain. There is a high incidence of thyroid gland dysfunction, which can produce thyroid hormone excess or deficiency (the latter is more common). The immune cells of the body react abnormally in DS and this causes a predilection to infections such as pneumonia and hepatitis. The heart is malformed in DS about 50% of the time and the most common lesion is a defect between the atrium and the ventricle. In some cases surgical repair of the lesion is required, but in many instances the effect on cardiac function is minimal and further intervention is not necessary. The immune problem that underlies the tendency to infection may also contribute to an increased incidence of leukemia in DS. Most cases of leukemia in DS occur in children less than 10 years old.

The ear canals in DS are small and misshapen. As a result, children with DS have frequent ear infections and run the risk of hearing loss along with other complications. A hearing loss that would be a mild handicap for a normal child can be devastating to a school-age child with DS who is struggling against the cognitive disorder associated with the condition. Tongue size and enlargement of the tonsils may provoke periodic cessation of breathing during sleep in DS (sleep apnea) and result in fatigue as well as poor mental performance the next day.

Whether people with DS age prematurely has not been decided with certainty. Skin laxity, adult-onset diabetes, cataracts, and orthopedic difficulties would argue for a precocious biological aging. However, people with DS seem to have a low incidence of atherosclerotic heart disease and hypertension – factors that distinguish them from aging people in the general population. The tendency to Alzheimer disease is a special circumstance (see below).

COURSE OF DS

The infant with DS is often in a vulnerable state. Muscle tone is lax and motor development is often delayed. Infection always looms as a possibility. Congenital heart disease, immune dysfunction, thyroid abnormalities, and the tendency to leukemia are always present. Given so many potential obstacles, it is reassuring to see that so many children with DS grow up and thrive. Particular attention to people with DS is required across the life span, and a universal set of guidelines for preventive intervention is increasingly being applied within the medical profession.

NEURAL CORRELATES

The experienced neuropathologist can often make a diagnosis of Down syndrome by the appearance of the brain. Certain neuroimaging procedures afford the same opportunity. In DS, the brain is slightly underweight, foreshortened front to back, and has a steep decline to the posterior regions.

Muscle tone is abnormally low in infants with DS and this is a sentinel sign of the disorder at birth. The ligaments are lax and this provides a problem in stability for younger children with DS. More worrisome is the ligamentous laxity that occurs in the bones at the top of the spinal column. This atlantoaxial junction is subject to more movement in DS than seen in the general population and occasionally the spinal cord will be compressed by the vertebrae (atlantoaxial dislocation). For this reason, people with DS should have an X-ray examination of the spinal area before participating in certain sports.

Although a cognitive deficit is universally seen in DS, there is a broad range of potential abilities in the disorder. This variation suggests that we do not yet understand the real potential of people with this condition or how to account for the spread of abilities. Social abilities often allow the child and adult to function in a more adaptive manner than would be predicted on the basis of tests measuring general intelligence.

One area of challenge for people with DS about which there is universal agreement is the linguistic system. The typical person with DS has more difficulties with language than would be predicted on the basis of overall cognitive functioning. Verbal memory is a particular deficit. People with DS can not only recall fewer digits in a span than can

people in the general population, but they also can recall fewer digits than intelligence-matched control subjects who do not have DS. The rate of acquiring new vocabulary is slower in DS than in children of similar mental age.

Epileptic seizures may occur at two separate points in the life span of an individual with DS. In infancy, seizures often take the form of sudden jerking movements called 'infantile spasms'. In later life the onset of seizures may herald the decline associated with Alzheimer disease. Since the early description of DS, personality characteristics have included a strong power of imitation, good sense of humor, and a tendency towards obstinacy. As a rule, people with DS appear to be outgoing, affectionate, with social quotients exceeding intelligence quotient (IQ) measures between 4 and 17 years of age. Older age in DS often brings bouts of depression.

LINKS WITH ALZHEIMER DISEASE

Alzheimer disease (AD) is the most commonly recognized form of dementia and is manifested by a progressive and ultimately fatal deterioration in brain function and capacity. From the late 1940s, researchers have found a curious association between DS and AD. Almost every brain examined from individuals with DS over the age of 40 years shows microscopic signs of the disorder. Often the primitive lesions of AD can be seen in brain tissue from children and young adults with DS. Despite the ubiquity of the microscopic signs of AD, not every person with DS develops clinical symptoms of the disorder. Prevalence rates vary widely, but average about 25%. The early symptoms of AD include change in personality, loss of ability to carry out complex daily skills, and (as mentioned above) the onset of epileptic seizures. As the decline progresses, awareness of the environment is lost and the ability to acquire recent memories ceases. In DS the disease may run a fatal course within 5 years. One of the factors imparting a predisposition to AD in DS relates to a gene located on chromosome 21 which is triplicated in DS. This gene is responsible for producing an overabundance of the peptide amyloid. Amyloid deposition in brain appears to have a key role in the pathogenesis of AD. (*See* **Alzheimer Disease**)

TREATMENT AND GENETIC COUNSELING

There is no medical cure for DS but there are many opportunities to prevent complications in the disorder and to provide a nurturing environment for the realization of full potential for people with this syndrome. A signal advance in modern society has been the provision of opportunities for people with DS (and other forms of learning difficulties) to live and work within the community. No longer are people with DS sent to institutions to live their lives apart from society. With increasing vigilance among medical professionals and the general willingness to intervene on behalf of the child with DS, the life span for people with DS has been greatly extended.

Genetic counseling for DS shares many of the same principles that are applied to other genetic conditions. The genetic counselor's role includes:

- interpreting recurrence risk for families where a DS birth has occurred
- assisting other medical professionals in explaining the diagnosis of DS, patient management, and treatment options
- facilitating psychological assessments where indicated for both patient and family members
- providing information to families about support groups.

The recurrence risk for DS is dependent on the genetic form of the disorder. The disclosure of full information on genetic risks often involves 'cytogenetic' testing of the patient and possibly family members. The recurrence risk to a couple with the classical form of trisomy 21 is about 1%, but there is a strong maternal age dependency which may increase this risk and requires explanation for couples engaging in family planning. In the translocation form of DS, the risk is dependent on whether the translocation was inherited by a parent or occurred for the first time in that individual child. A parent who can transmit a chromosome 21 translocation is referred to as a 'balanced translocation carrier'. Such a carrier parent is asymptomatic and has the normal chromosomal complement, except that part or all of one copy of chromosome 21 is attached to another chromosome. In this case the genetic risk varies widely and is dependent upon the chromosome to which the translocated segment of 21 is joined. The genetic risk for the mosaic form of DS depends on the percentage of mosaic cells in the individual as well as the percentage of trisomic cells in the gonadal tissues. Counseling for these rare forms of DS is complicated and always requires the services of a professional genetic counselor.

For families who wish to know whether the fetus being carried has DS, several chemical compounds have been identified in maternal blood that have a predictive value for diagnosis in up to 70% of cases.

These are considered risk factors but are not diagnostic by themselves. More precise diagnostic data may be obtained prenatally by examining the amniotic fluid and determining the actual chromosomal complement of fetal cells. Maternal blood screening is best done at 15–17 weeks of gestation, whereas examination of amniotic fluid (amniocentesis) is generally done between 14 and 18 weeks of gestation. A newer procedure at around 10 weeks of gestation avoids amniocentesis and obtains fetal cells for genetic testing by sampling part of the placenta (chorionic villus biopsy). The complexity of prenatal diagnosis requires a close working relationship between professionals in obstetrics and in genetics. Many clinics specializing in these integrated services are available in modern medical systems.

Further Reading

Berg HM, Karlinsky H and Holland AJ (eds) (1993) *Alzheimer Disease, Down Syndrome, and Their Relationship*. Oxford, UK: Oxford Medical.

Beck MN (1962) *Expecting Adam: A True Story of Birth, Rebirth, and Everyday Magic*. New York, NY: Times Books.

Cicchetti D and Beeghly M (eds) (1990) *Children with Down Syndrome: A Developmental Perspective*. Cambridge, UK: Cambridge University Press.

Epstein CJ (1986) *The Consequences of Chromosome Imbalance: Principles, Mechanisms and Models*. Cambridge, UK: Cambridge University Press.

Hassold TJ and Patterson D (eds) (1999) *Down Syndrome: A Promising Future, Together*. New York, NY: Wiley-Liss.

Lott IT and McCoy EE (1992) *Down Syndrome: Advances in Medical Care*. New York, NY: Wiley-Liss.

Miller JF, Leddy M and Leavitt LA (1999) *Improving the Communication of People with Down Syndrome*. Baltimore, MD: Brooks.

Rondal JA (ed.) (1996) *Down's Syndrome: Psychological, Psychobiological and Socioeducational Perspectives*. San Diego, CA: Singular.

Selikowitz M (1990) *Down Syndrome: The Facts*. Oxford, UK: Oxford University Press.

Stratford B (1989) *Down's Syndrome: Past, Present and Future*. London, UK: Penguin.

Early Experience and Cognitive Organization

Introductory article

Barbara Landau, Johns Hopkins University, Baltimore, Maryland, USA

CONTENTS

There is much support for the idea that cognitive development is specialized, with different aspects of knowledge developing under different mechanisms and timelines. Many aspects of mature knowledge can be traced to innate origins. At the same time, variation in early experience due to genetic defects, blindness or deafness can lead to significant reorganization of the developing brain. Thus human cognitive development is constrained by innate potential, but is also affected by experience during early development.

INTRODUCTION

One of the great mysteries of human development is how our knowledge comes to be organized in the way in which it is. Two competing views have quite different answers to this. The nativist view holds that we are born with specific capacities to organize the world in the way that we do. This view can be traced to the thinking of philosophers such as Immanuel Kant and René Descartes. In contrast, the empiricist view holds that experience after birth molds our knowledge. This view is traceable to philosophers such as John Locke and David Hume. The nativist view is supported by the fact that different aspects of our knowledge have quite different forms, and these develop quite early in life without any formal tutoring. Yet some variations in early experience can have profound effects on brain organization, suggesting that there is some flexibility in the way in which the brain and mind organize the world. This article will consider two types of change in early experience. One is due to changes in an individual's genetic endowment, and the other is due to changes in an individual's sensory and perceptual experience.

SPECIALIZATION IN COGNITIVE DOMAINS

When considering the relative roles of innate potential and experience, it is important to understand that different aspects of knowledge have very different types of organization – that is, they are 'specialized' in the sense that the principles which govern one domain of knowledge may be irrelevant or inapplicable to the next domain. For example, our capacity for language allows us to learn language early in life, and to fluently produce and understand even quite complex sentences throughout life. This capacity reflects our knowledge of language, which is a complex, rule-governed system that allows us to produce and comprehend an infinite number of sentences. In contrast, our capacity to move around the world and find our way, to remember the spatial locations of objects, and to read maps, is part of a different system of knowledge. This system allows us to learn the spatial relationships between places in space, and to understand how to get from one place to another, even if we are travelling along novel routes. This is part of our spatial knowledge system, but it is irrelevant to our knowledge of language. Other cognitive domains are specialized as well. For example, our knowledge of number, causality, time and social relationships is governed in all of these cases by rules and representations that are distinctly different from each other.

Examining development under conditions of varying genetic endowment and/or different types of sensory and perceptual experience can shed light on the question of how cognitive specialization emerges. The nativist view would suggest that

there are strong biological constraints on the types of knowledge that humans can develop, and that changes in genetic endowment might have specific, targeted effects on cognition. At the same time, the empiricist view would suggest that changes in sensory or perceptual endowment – such as congenital blindness or deafness – might lead to changes in the way in which the brain's organization supports cognition. Evidence from recent research supports aspects of both views.

WILLIAMS SYNDROME: A GENETIC DISORDER THAT RESULTS IN SELECTIVE COGNITIVE IMPAIRMENT

Recently, cognitive researchers have become interested in Williams syndrome, a relatively rare genetic syndrome (1 in 15 000 live births) which is caused by a microdeletion of material (approximately 20 genes) on chromosome 7. Individuals with Williams syndrome show distinctive physical characteristics and an unusual cognitive profile. In general, these individuals are moderately mentally retarded, with an average IQ of around 60 (where 100 is the average for the general population). However, their retardation is not 'across the board'. Rather, individuals with Williams syndrome show severe impairments in their understanding of spatial relationships, but considerable strength in language ability. Their severe spatial impairments are manifested in a number of ways. One of the clearest examples is that they have extreme difficulty in copying simple figures. Figure 1 shows an example of drawings produced by two 11-year-olds with Williams syndrome, and one drawing produced by a normally developing child who was younger, but who had the same 'mental age' on an intelligence test, and had a roughly average IQ. It can be seen that the individuals with Williams syndrome could easily duplicate the colors in the model, but could not copy the spatial relationships shown there, resulting in a very distorted copy. In contrast, the normally developing child could reproduce both the colors and the spatial organization. The spatial impairment in people with Williams syndrome contrasts strikingly with their strong language capacities, which are demonstrated by a fluent, articulate

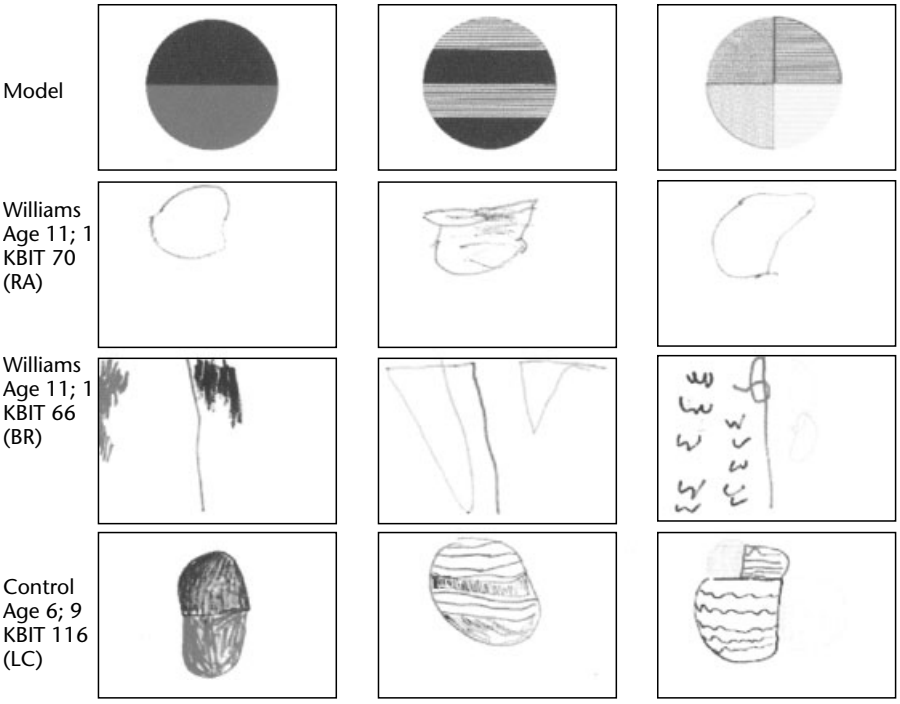


Figure 1. [*Figure is also reproduced in color section.*] Copies made by two 11-year-old individuals with Williams syndrome and a normally developing 6-year-old who was matched for mental age. Each model contains between two and four colors, and these are correctly copied by the children with Williams syndrome as well as by controls. However, there is considerable impairment of the spatial organization of Williams syndrome copies, resulting in unrecognizable configurations. Ages are stated in years and months. KBIT scores are from the Kaufman Brief Intelligence Test, and represent approximate IQ equivalents.

conversational style, relatively high scores on vocabulary tests, and the capacity to produce and understand sentences that are grammatically complex. Individuals with Williams syndrome also tend to be musical, and some people with this syndrome are highly skilled in musical activities such as singing and playing musical instruments. The striking impairment of spatial capacities, coupled with the fact that language appears to be strongly preserved in Williams syndrome, supports the idea that cognitive specialization might ultimately be linked to aspects of genetic endowment.

Researchers are currently asking many questions about the emergence of cognitive specialization in this syndrome. For example, it may be that the spatial impairment is more apparent in some aspects of the spatial knowledge system than in others. Some authors have conjectured that the profile of strengths and weaknesses within spatial cognition may be linked to somewhat unusual organization of the brain in Williams syndrome. One possibility is that the spatial impairment may be limited to certain specific spatial functions that are performed by the brain. The most obvious examples of spatial impairment in people with Williams syndrome are shown when these individuals are asked to copy an existing pattern, either by arranging a set of blocks to copy a pattern, or by drawing a copy of a model, as in Figure 1. However, people with Williams syndrome do not appear to show impaired performance on other spatial tasks. For example, they do not appear to have difficulties in recognizing faces or identifying pictures of objects, even when the objects are presented from very unusual viewpoints, as in Figure 2. In order to identify objects under these types of conditions, it would seem to be necessary to recognize both the parts of the objects and their spatial relationships. Thus this aspect of spatial capacity might be relatively spared in Williams syndrome, even though other aspects of spatial capacity are clearly impaired. Evidence from cases of normal adults who then sustain brain damage (e.g. through strokes) suggests that object recognition and identification may be carried by the brain's 'ventral' stream. The relative strength of object recognition in Williams syndrome may indicate selective sparing of functions carried by this stream. As another example, individuals with Williams syndrome tend to have some difficulty in planning visual–motor acts, such as posting a letter through a mail slot. Yet they have less difficulty when they must simply view the slot and compare its orientation to another sample slot. Again, some authors have conjectured that the relative difficulty in acting on objects, compared with simply perceiving and matching them, could be due to selective impairment of the brain's 'dorsal' stream. Finally, individuals with Williams syndrome perform better than normal children matched for mental age on tests of 'biological motion' perception, in which they perceive a moving animate figure that is shown only by a set of individual dots. Overall, the evidence suggests that there is specialization in the breakdown of spatial cognition in Williams syndrome. This and other hypotheses are currently being actively tested by researchers.

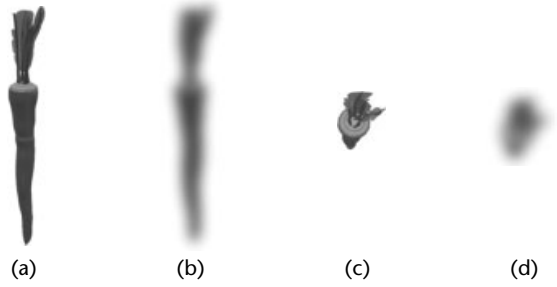


Figure 2. [*Figure is also reproduced in color section.*] These objects are easily recognized and named by children with Williams syndrome, despite the fact that they have profound spatial impairments. The objects are all the same carrot, but vary in whether they are common viewpoints (a and b) or unusual ones (c and d), and also in whether they are clear images (a and c) or blurred ones (b and d).

DOWN SYNDROME: A GENETIC DISORDER THAT SHOWS A MORE GENERAL DECLINE IN COGNITION

In comparison with Williams syndrome, Down syndrome has a very different cognitive profile, again suggesting that cognitive development may be partially under the control of genetic endowment. Down syndrome occurs in roughly 1 in 1000 births, and in the majority of cases it is associated with an additional copy of chromosome 21 (resulting in three rather than two chromosomes at that locus – hence the name 'trisomy 21'). Individuals with Down syndrome tend to have distinctive physical characteristics, although these are different from those in Williams syndrome. Furthermore, there are some gross anatomical differences between the brains of individuals with Down syndrome and those of individuals with Williams syndrome. Finally, the cognitive profile of the two groups is quite different. Individuals with Down syndrome are moderately retarded, but their cognitive profile is relatively even. That is,

they are impaired in various cognitive domains to roughly the same extent, and they do not show the striking profile of strengths and weaknesses that is characteristic of individuals with Williams syndrome.

A number of studies have directly compared the spatial and linguistic capacities of individuals with Down and Williams syndromes. The results of spatial studies show that although both groups are impaired relative to normally developing individuals, people with Williams syndrome perform even more poorly on some spatial tasks than individuals with Down syndrome who have the same IQ. In addition, they may use different types of solutions for the same spatial problems. For example, several reports indicate that individuals with Down syndrome can replicate more global spatial aspects of a model, whereas individuals with Williams syndrome are more likely to replicate accurately the more local aspects of a model – the smaller elements that are put together to make up the overall pattern. It is possible that these different patterns of spatial impairment in Down versus Williams syndrome are due to different types of brain organization and impairment.

The results of language studies show that although neither group performs at a level commensurate with its chronological age (i.e. they are both impaired to some extent), individuals with Williams syndrome outstrip those with Down syndrome who have the same IQ. Moreover, when individuals with Williams syndrome are tested on grammatically complex structures, they tend to do well compared with individuals with Down syndrome. The spontaneous speech of individuals with Down syndrome does not exhibit a large amount of grammatical complexity. For example, longer sentences are primarily produced by stringing together simple phrases with conjunctors such as 'and'. In contrast, individuals with Williams syndrome produce longer sentences that exhibit a considerable degree of grammatical complexity, including relative clauses (e.g. 'The boy that was looking at the man jumped over the fence'). The fact that these can be fluently produced by individuals who are retarded points to the possibility that language capacity may develop rather normally in the case of Williams syndrome. In the case of Down syndrome, the general dampening of linguistic capacity is roughly commensurate with the overall retardation of these individuals. The combination of cases indicates that changes in genetic endowment may have highly specific and targeted consequences for cognitive development.

CONGENITAL BLINDNESS AND DEAFNESS

Sensory deficits such as congenital blindness and deafness may result in reorganization of the brain. There is considerable evidence that knowledge in a variety of domains can develop normally in individuals who are blind or deaf from birth. For example, congenitally blind individuals show the capacity to understand locations in space, to get from one place to another by independent locomotion, and to read haptic or Braille maps. As another example, congenitally deaf individuals acquire manual (signed) languages effortlessly, just as hearing individuals effortlessly acquire spoken languages. In these cases and others, the knowledge that develops is organized in the same manner as in individuals who are sighted or hearing. This indicates that cognitive organization can emerge guided by principles that do not depend on these particular types of sensory or perceptual experience (i.e. seeing or hearing).

However, there is also evidence of flexibility in the way in which the brain accomplishes these cognitive functions. This flexibility is evident from studies of brain activity following sensory or perceptual deprivation. A number of studies have shown that changes in early experience can have a major impact on the organization of the brain's cortex. If an individual is deprived of input from sensory receptors, the cortical areas that are normally activated by those receptors are 'taken over' by other types of input. For example, in adults who have had a limb amputated, the cortical area that used to represent that limb (i.e. be activated by it) now becomes responsive to stimuli from other areas of the body that would normally activate neighboring areas of the cortex. In effect, the areas of cortex that have been 'abandoned' by the amputated limb are commandeered to serve other functions. Other research shows that individuals who have been blind or deaf from birth have brains that are organized somewhat differently from those of individuals who are sighted or hearing. For example, in some research, congenitally blind individuals who perform tasks using the skin and hands show brain activation in areas that are normally 'designated' for vision. In this case, the visual areas are commandeered for touch, and hence the area of cortex dedicated to touch is expanded in these individuals. Other research has shown that congenitally deaf individuals show unusual brain activation patterns when they pay attention to stimuli in their peripheral visual field – that is, outside the central view. Additional evidence

shows that these individuals may be more accurate at detecting these stimuli than hearing individuals. In essence, congenital auditory deprivation appears to lead to a reorganization of the brain in which there are changes in the ways in which visual stimuli are processed.

SUMMARY

Mature knowledge reflects the development of a set of highly complex, differentiated and specialized systems, including language, spatial knowledge, number, causality and other domains. These specialized systems are partly under the control of genetic endowment and partly under the control of variations in experience. The results of genetic deficits are highly specific, with some syndromes targeting specific knowledge systems and other syndromes having more general effects. The results of massive variation in experience, such as congenital blindness or deafness, are some degree of reorganization in the brain, indicating plasticity during early development.

Further Reading

Gallistel CR, Brown A, Carey S, Gelman R and Keil F (1991) Lessons from animal learning for the study of cognitive development. In: Carey S and Gelman R (eds) *The Epigenesis of Mind: Essays on Biology and Cognition*, pp. 3–36. Hillsdale, NJ: Erlbaum.

Gazzaniga MS, Ivry RB and Mangun GR (2000) *Cognitive Neuroscience: The Biology of the Mind*. New York: Norton.

Johnson M (1997) *Developmental Cognitive Neuroscience*. Cambridge, MA: Blackwell.

Jordan H, Reiss J, Hoffman JE and Landau B (2001) Intact perception of biological motion in the face of profound spatial deficits: Williams syndrome. *Psychological Science* **13**(2): 162–167.

Mervis C, Morris CA, Bertrand J and Robinson B (1999) Williams syndrome: findings from an integrated program of research. In: Tager-Flusberg H (ed.) *Neurodevelopmental Disorders: Contributions to a New Framework From the Cognitive Neurosciences*, pp. 65–110. Cambridge, MA: MIT Press.

Uecker A, Mangan PA, Obrzut JE and Nadel L (1993) Down syndrome in neurobiological perspective: an emphasis on spatial cognition. *Journal of Clinical Child Psychology* **22**: 266–276.

Education, Learning in

Introductory article

Stanton Wortham, University of Pennsylvania, Philadelphia, Pennsylvania, USA

CONTENTS

Learning takes place in many settings, but educational institutions foster both breadth and depth of learning. Different types of teaching make very different assumptions about what learning is.

INTRODUCTION

Theories of learning have been applied most often in educational institutions. The relationship between cognitive science and education has benefited both scientists and practitioners. Scientists have used educational settings to develop and test their theories, and practitioners have used new knowledge about learning to design more effective education.

Broadly conceived, education is the process of continuing the human species. All humans are born immature, without the knowledge and skills they will need to function – without language, without knowing how to use complex tools, and so on. The species continues because adults communicate knowledge and skills to the next generation. This intergenerational transfer allows future generations to build on prior accomplishments.

Thus all humans teach. Whether they realize it or not, all teachers act as if some theory of learning is true. Particular ways of teaching make assumptions about what learning is. Furthermore, theories of learning themselves rest on conceptions of human nature. Different accounts of how people learn assume different things about what people are essentially like.

This article describes three broad theories of learning – together with the conceptions of human nature underlying these theories – and the types of educational practice that have been built on these theories. The article has two purposes. First, it is important to recognize the theories of learning and conceptions of human nature that underlie various types of schooling. The article describes how typical teacher and student behavior makes assumptions about how learning happens. Second, as theories of learning have developed, we have learned that earlier theories were too simple. The article describes how more complex accounts of learning and human nature are needed to guide educational practice.

BEHAVIOR

Theories of learning that focus on behavior are called 'behaviorist'. Behaviorists argue that humans should not consider themselves special. Copernicus showed that the earth was not the center of the solar system, and Darwin showed that humans were not qualitatively different from animals. Behaviorists further puncture our sense of superiority, arguing that humans do not have free will to act as they choose. 'A person does not act upon the world', B. F. Skinner said, 'the world acts upon him.' On this theory, the environment shapes people's behavior through reinforcement. Just as Darwin showed that organisms appear designed by a creator to fit their niche, even though adaptation is in fact a result of random variation and natural selection over time, behaviorists show that humans appear to reflect and choose their actions, while in fact their behavior has been shaped by reinforcement.

To learn, then, is to change one's behavior in response to reinforcement. This account of learning contains three central elements: behaviors by the organism, conditions present in the environment, and consequences that follow from various behaviors. People, like other animals, will generate various behaviors in a new situation. Some of these behaviors will result in positive consequences, while others will not. People learn to respond more often with behaviors that were reinforced positively in a given situation.

Behaviorist Education

On a behaviorist account, teaching is the systematic shaping of a student's behavior. The teacher has control and students are raw material to be shaped.

Teachers arrange reinforcements so that students come to behave as teachers want them to. Scientists have successfully taught pigeons to play ping pong, for instance, by designing a long series of intermediate skills that lead from natural pigeon behavior to ping pong. They reinforce the pigeons for performing each of these intermediate skills, in turn, until the pigeons produce the target behavior. Similarly, teachers of human students should define the target behaviors, design a path of intermediate behaviors from what students can already do, then reinforce students at each step until they produce the target behavior.

Behaviorists have designed 'teaching machines' that dispense rewards as students accomplish prespecified tasks. One famous picture shows a small boy playing a piano, with a candy dispenser on top. Although these pictures now look outdated, many practices in today's schools presuppose a behaviorist account of learning. Discipline systems almost always rely on rewards and punishments to shape students' behavior. Grades are used as reinforcers. And many classroom practices, from worksheets to testing, involve teachers rewarding students for producing desired behavior.

Research in cognitive science from the second half of the twentieth century has shown that behaviorism is not an adequate theory of learning. People often act because they value activities intrinsically, not for external reinforcement. As described in the next section, people also develop complex representations of the world and reflect on their actions in a way that behaviorists denied. Why, then, do students and teachers so often act as if behaviorism were true?

Because it works. If you have control over effective reinforcers, you can shape people's behavior. Behaviorism is not false. It is true, but it is not the whole truth. Under certain circumstances, people do learn just like animals. The question is whether we should create more circumstances that encourage people to learn in this way. Cognitive scientists claim that we should not, because humans have the potential to learn in nonbehaviorist ways, and because students can develop deeper knowledge when encouraged to learn differently.

MIND

Theories of learning that focus on mental representations are called 'cognitivist'. Cognitive approaches to learning see humans as actively making sense of the environment. People develop mental models of the world and act on the basis of these models, not simply in response to reinforcements. When people encounter a new situation, they assimilate it to their own pre-existing models of the world. Learning involves expanding those mental models, in order to make them more accurate.

This account of learning distinguishes between genuine understanding and merely producing the right behavior. People often just parrot the right answer without understanding it, just as pigeons can play ping pong without understanding what they are doing. True learning involves a deeper grasp of the subject matter, such that people's mental models line up with the world. Furthermore, people cannot be forced to learn. True learning requires a change in people's internal models, and learners must change these models themselves.

Cognitive scientists have described various structures and processes that underlie learning. There seem to be some universal constraints, which presuppose people to certain broad types of mental models. Particular domains of knowledge are also organized in distinct ways, to facilitate learning. And individuals sometimes vary in the types of structures that they operate most effectively with. For instance, there are different learning styles – some people learn most effectively through verbal explanations, while others learn more effectively through visual diagrams, and so on.

Cognitivist Education

From this perspective, learners need to develop deeper understandings, not just produce the right behaviors. Deeper understandings cannot be imposed on students, because they must construct their own mental models. So teachers do not shape students, nor do they deliver correct answers. Teachers should develop educational environments that push students to broaden and deepen their own models, thus opening up areas of the world that students have not thought about. After teachers have set up rich educational environments, ones that contain puzzles designed to provoke students to reflect, then they must allow students to explore. Teachers can challenge students, by pointing out contradictions in their beliefs, but students themselves must recognize the puzzles and work to solve them. Teachers can explain, but if students can only repeat a teacher's explanation then they have not truly learned. Students themselves must integrate new experience with their own developing mental models.

Assessment is a bigger challenge for cognitivist educators than for behaviorist ones. Behaviorists pre-define the educational objective, and they assess whether students produce the desired behavior.

Genuine cognitive learning, in contrast, takes place internally. Teachers can infer about students' understandings, but they do not want to encourage rote learning by using simple tests. Instead of assessing whether students get the right answers, cognitivist educators try to assess underlying thought processes by examining how students reached certain answers.

Cognitivist theories of learning are more widely accepted than behaviorist ones. Nonetheless, there is less cognitivist teaching than behaviorist teaching in our schools. This happens partly because cognitivist education is difficult for both teachers and students. Because they are responsible for students' learning, it is hard for teachers to let students pursue their own ideas much of the time. Students also find it easier to write down what the teacher says, instead of developing their own accounts. This sort of resistance can be overcome, and many teachers do successfully encourage students to develop their own deeper understandings. But behaviorist practice has been harder to overcome than behaviorist theory.

SOCIETY

Theories of learning that go beyond mental representations to include social practices are called 'social cognitivist'. Cognitivist learners are autonomous, developing models themselves to make sense of the environment. Recent theories present the learner, instead, as a participant in social activities. Learning, on this account, is a transformation of participation in activity, not primarily the creation of mental models. Instead of simply developing their own representations, people become increasingly competent participants in the intellectual lives of those around them.

From this point of view, people learn as they more competently use tools to facilitate thought and action. Adults incorporate learners into their activity by teaching them how to use certain cognitive tools. Some of these tools are mental, such as mnemonic devices. Others are objects, such as maps. But learners do not have to construct them alone, because these tools have already been developed and can be borrowed from others.

Any theory of learning presupposes a 'unit of analysis'. This is the smallest unit that preserves essential behavior of the whole. In order to study the behavior of water, for example, one must understand the molecular level. Studying hydrogen and oxygen atoms separately will not allow one fully to understand the behavior of water. Similarly, one cannot fully understand learning solely by studying individuals' mental representations. Individual cognitions are essential, as hydrogen atoms are essential to water, but learning itself depends on a larger unit: a social activity, which includes individuals' mental representations, various cognitive tools, and others' knowledge and skill, all of which together allow learning.

Unlike behaviorists, and like cognitivists, social cognitivists describe how cognitive structures and processes mediate between the environment and people's actions. But social cognitivists emphasize that these mediating structures go beyond individuals' mental models to include tools and other aspects of social activities. Although some activities (such as conventional tests) do require individuals to think in isolation with limited tools, a full account of learning must analyze social activities in addition to mental representations.

Social Cognitivist Education

In a social cognitivist approach, both teacher and student are active. Instead of relying primarily on students' own exploration and model-building, the teacher acts as a competent practitioner of the activity being taught and brings tools for students to use. Teachers guide students as they begin to participate in the activity. This guidance allows students to do tasks that they would not be able to perform on their own. Students act like apprentices, at first doing minor parts of the task while observing others, then taking on increasing responsibility.

Teachers should design more naturalistic or 'authentic' activities for students to participate in, where the goal is competent participation in real activity. Many medical schools, for instance, now use 'problem-based learning' – in which groups of beginning students are given real, complex cases and asked to diagnose the problem. They must consult more expert practitioners, do research on relevant topics, and develop alternative diagnoses to present in class. Students thus learn how to participate in the practice of medical diagnosis, and they learn the relevant facts along the way.

From this perspective, testing is unnatural. If students must learn to participate competently in real activities, teachers should not test whether they can solve problems by themselves out of context. And because learning most often involves participating with others to accomplish a task, students should not be tested alone. Students should instead be asked to exhibit their mastery by participating competently in naturalistic activities.

Like behaviorism, pure cognitivism is only partly true. Just as people are often manipulated by reinforcements, people often rely primarily on their own mental models to understand the environment. But if our educational goal is to help young people build on the knowledge and skills that have been developed by previous generations, we should treat them neither as animals to be shaped nor as lone thinkers. We must help them grow into and expand the activities that make us human. This will require educational practices based on more complex accounts of learning.

Further Reading

Anderson J, Reder L and Simon H (1996) Situated learning and education. *Educational Researcher* **25**: 5–11.

Duckworth E (1987) *'The Having of Wonderful Ideas' and Other Essays on Teaching and Learning*. New York, NY: Teachers College Press.

Engeström Y, Miettinen R and Puramäki R (1999) *Perspectives on Activity Theory*. Cambridge, UK: Cambridge University Press.

Gardner H (1999) *Intelligence Reframed*. New York, NY: Basic Books.

Greeno J (1997) On claims that answer the wrong questions. *Educational Researcher* **26**: 5–17.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Lave J and Wenger E (1991) *Situated Learning*. Cambridge, UK: Cambridge University Press.

McGilly K (1994) *Classroom Lessons*. Cambridge, MA: MIT Press.

Piaget J (1967) *Six Psychological Studies*. New York, NY: Random House.

Renninger K (1998) Developmental psychology and instruction. In: Siegel I and Renninger K (eds) *Child Psychology in Practice*, pp. 211–274. New York, NY: John Wiley.

Rogoff B, Turkanis C and Bartlett L (2001) *Learning Together*. Oxford, UK: Oxford University Press.

Schwartz B (1985) *The Battle for Human Nature*. New York, NY: WW Norton.

Skinner BF (1968) *The Technology of Teaching*. New York, NY: Appleton-Century-Crofts.

Vygotsky L (1978) *Mind in Society*. Cambridge, MA: Harvard University Press.

Wertsch J (1998) *Mind as Action*. Oxford, UK: Oxford University Press.

Emotion

Introductory article

Paula M Niedenthal, National Centre for Scientific Research, Blaise Pascal University, Clermont-Ferrand, France

CONTENTS

Emotions are sets of processes involved in an organism's response to significant, goal-relevant life events. Such processes include expressive behavior, cognitive appraisals, physiological arousal, action tendencies and subjective feelings.

INTRODUCTION

It is difficult to imagine life without emotions. We would feel no joy at successfully accomplishing a task, no sadness at failing an examination, no anger when we witnessed displays of prejudice and discrimination. We would not feel ashamed upon insulting another individual in social interaction. Nevertheless, emotion as a field of study remained firmly ensconced in the realms of philosophy and literature for many centuries. Now of central interest to psychology, the cognitive sciences and the neurosciences, emotion is finally also a topic of empirical investigation. Neuropsychological and psychological investigation shows that it is indeed hard to imagine life without emotions, because emotions are essential for human functioning. Social interaction, decision-making and judgment would be very poor without the capacity to experience emotion. Normal and pathological emotion states are determined and controlled by almost all of the systems of the human body, and emotional states in turn influence the cognitive functions of attention, perception, categorization, memory and judgment.

EMOTIONS AS MULTICOMPONENT PROCESSES

A precise definition of emotion eludes scientists. This is because no single event or process constitutes an emotion. Emotions are sets of processes that involve different components including subjective feelings, but also expressive motor action, cognitive appraisals, physiological arousal, and tendencies to take particular actions. If you see a bear in the forest, a favorite example of the nineteenth-century psychologist William James, you might experience strong physiological arousal, have an urge to run, open your eyes and mouth wide, and feel something that you label as fear. In this example the components are quite coherent and conform to a common emotional experience. However, the different components of emotion can be decoupled. In different situations and across different cultures, social norms influence the expression and experience of emotion. For example, an adult in an industrialized Western country might feel like laughing at a funeral if a funny joke about a priest or a rabbi suddenly comes to mind; however, the person would probably suppress any laughter, avoid smiling, and display some degree of sadness. Such norms are one force that can decouple the multiple components of emotion. In this example the decoupled components are subjective experience and expressive behavior. (*See* **Cultural Psychology**)

Differentiation of the Component Processes: Are There Discrete Emotions?

Although scientists agree that emotions have many components, debate continues about the extent to which there are unique patterns or levels of each component process that correspond to discrete (basic or fundamental) emotions such as sadness, joy or fear.

Facial expression

The component of initial interest and debate was that of facial expression of emotion. Charles Darwin was an early proponent of the idea that facial expressions of emotions serve adaptive

functions – he called them 'serviceable habits' – and have thus evolved as hardwired expressions of discrete subjective states. For example, disgust, Darwin suggested, is associated with a gesture that represents the expulsion of food from the mouth and avoidance of the intake of an odor through the nose, precisely because the function of disgust is to prevent individuals from ingesting dangerous substances. From this evolutionary perspective, facial expressions of emotion should be both distinctive and universal. Indeed, studies of people of different cultures, even those not exposed to Western media, blind children, and infants, suggest that there exist universal displays and recognition of the expressions of at least joy, sadness, fear, anger, disgust, and perhaps surprise. Thus, such facial gestures may be based in innate neural motor programs. However, such a conclusion is still widely debated because each existing demonstration of universality can be criticized on technical grounds. (*See* **Face Perception, Psychology of**; **Face Perception, Neural Basis of**)

Vocalizations

Somewhat less controversial is the study of vocal expression of emotion, also called emotional prosody. Research supports the idea that vocal expression is biologically based and that there is evolutionary continuity of vocal emotion expression. For example, data from studies conducted by behavioral biologists suggest that there are significant similarities in the vocal expression and communication of emotional states across species: angry states are typically expressed by loud, harsh vocalizations, while fear and anxiety are typically associated with high-pitched, shrill vocalizations. Furthermore, human perceivers show wide agreement about the emotions communicated by utterances characterized by specific patterns of physical parameters such as fundamental frequency and pitch.

Autonomic activity

William James, as well as a Danish contemporary named Lange, originally proposed the idea that there are distinctive patterns of autonomic arousal that correspond to discrete emotions. Although this peripheralist position was attacked for over a century, both by scientists who believed that arousal was nonspecific, and by those who believed that the autonomic nervous system responds too slowly to subserve discrete emotional states, research suggests that some physiological parameters do differentiate some pairs of emotion. Sadness and fear differ in their patterns of heart rate, for example. Other studies show that fear is characterized by increases in heart rate, contractility of the heart musculature and respiration rate, which together mobilize a flight response, while anger is characterized by a rise in diastolic blood pressure and in peripheral resistance, which together mobilize a fight response. (*See* **Autonomic Nervous System**)

Subjective experience

Despite intuition to the contrary, some emotion theorists do not believe that emotional states are differentiated as discrete categories (happiness, sadness, disgust) of subjective experience. Rather, they believe that the conscious experience of emotion can be accurately characterized by two underlying psychological dimensions. Early in the last century these dimensions were referred to as pleasantness–unpleasantness and excitement–depression, although now the dimensions are often termed 'valence' (positive–negative) and 'activation' (high–low). In the dimensional view, for instance, the emotion that is called fear can be described psychologically as a state that is negative in valence and high in activation. One answer to the question of whether subjective experience is differentiated in terms of a few dimensions or whether it is categorical, is to say that it depends on the person. Research suggests that some individuals do indeed experience their emotional states in terms of a few simple dimensions ('I feel good') while others make finer, categorical distinctions among states ('I feel joyful'; 'I feel proud').

THEORIES OF EMOTION

Theories of emotion are precise statements about the causes of an emotion, the order in which emotional processes unfold, and how the different components of emotion interact. Although within each category several different versions can be discerned, there are currently two major categories of emotion theory. These are evolutionary theories and cognitive appraisal theories. Depending upon which approach is adopted by a scientist, the methods for producing and measuring an emotion in the laboratory, the types of emotions under scrutiny, the way in emotions are thought to be elicited and regulated and the application of the research findings may be quite different.

Evolutionary Theories

The evolutionary approach to emotion is based in part on the thinking of Charles Darwin and begins with the assumption that emotions are biologically

based and functional. In considering the evolution of emotion, Darwin focused largely on his idea that facial expressions of emotion are remnants of serviceable habits. In addition, he argued for an adaptive signaling function of facial expression: that is, facial expressions allow members of the same species to know the subjective experience of the expresser, and therefore the emotional significance of the situation, as well as the expresser's likely actions. In a more general way, the evolutionary perspective assumes that emotions motivate adaptive action, such as the tendency to flee when fearful and the tendency to fight when angry. The problems of adaptation thought to be associated with specific emotions are finding and consuming food and drink, locating shelter, seeking support from other members of the same species, being social, satisfying curiosity, appraising sexual partners, nurturing offspring, and escaping dangerous situations. From a biological perspective, then, emotions are responses that have evolved to motivate individuals to successfully pass on their genes to offspring. The three types of evidence cited in support of an evolutionary approach to emotion are: (1) blind children produce recognizable facial expressions of the basic emotions; (2) human facial expressions such as smiling have homologs in chimpanzee expressions; and (3) there are cross-cultural similarities in the antecedents of emotion. (*See* **Aggression and Defense, Neurohormonal Mechanisms of**; **Cultural Differences in Causal Attribution**; **Evolutionary Psychology: Theoretical Foundations**)

The implications of this approach are that there are biologically based systems that subserve emotions, in a discrete way. There also exist biologically relevant 'signal stimuli' that recruit specific emotional reactions and adaptive response tendencies in situations that are significant to the person for phylogenetic or ontogenetic reasons. At the behavioral level, emotional phenomena are displayed as action or motor tendencies. They thus recruit metabolic processes related to arousal and behavioral energetics. Such processes can be subjected to scientific investigation through psychophysiological indices such as cardiovascular and electrodermal responses. (*See* **Emotion, Neural Basis of**)

Cognitive Appraisal Theories

If the evolutionary approach links emotions to biological adaptation in the distant past, appraisal theories of emotion link emotions to 'higher level', cognitive processes of evaluation of meaning, causal attribution, and assessment of coping capabilities. The main principle of appraisal theories is that emotions are elicited and differentiated by individuals' evaluation of the significance or meaning of an object or event for themselves and their current goals on a number of dimensions or criteria. A classic demonstration of the role of appraisal in differentiating (although not eliciting) emotion was a study conducted by psychologists Schachter and Singer in the 1960s. These psychologists proposed that individuals sometimes experience physiological arousal that does not have a known cause. The arousal motivates individuals to explain the cause and nature of the arousal, which as consequence gives rise to a discrete emotional state. In their study the researchers injected participants with adrenaline (epinephrine), which enhances arousal, or with saline, a placebo which causes no physiological change. They then informed some of the participants injected with adrenaline to expect an increase in arousal; other participants were left uninformed about these effects. Later, when the adrenaline had taken effect for those who had received it, all participants were placed in a situation in which they interacted with a confederate of the researcher who acted in either a euphoric or an angry manner. Subsequent assessment of participants' emotions showed that those who had received adrenaline and were uninformed about the accompanying arousal – but not those participants who received placebo or who had an expectation of arousal – reported feeling the same emotions as those expressed by the confederate. That is, participants who experienced unexplained arousal sought an explanation in their environment and this appraisal determined their precise emotional state.

It is no longer widely believed that emotions are the products of unexplained arousal shaped by reference to events in the surrounding environment. However, modern cognitive appraisal theories do hold that discrete emotions result from processes of evaluation of significant events and of attributions of the causes of those events. In particular, whether unconsciously or consciously, individuals assess the degree to which events facilitate or impede their goals, whether they are pleasant or unpleasant, whether they are controllable or not, and whether they are novel or familiar. Depending upon the result of this appraisal, a discrete emotion is evoked. Tables 1 and 2 summarize examples of appraisal theories and the profiles that in theory give rise to certain emotions.

The implications of appraisal theories are that emotions do not unfold in a hardwired way in response to certain situations or objects, but that

Table 1. Comparison of the appraisal criteria postulated by different theorists

Scherer	*Frijda*	*Roseman*	*Smith/Ellsworth*
Novelty	Change		Attentional activity
• Suddenness			
• Familiarity	Familiarity		
• Predictability			
Intrinsic pleasantness	Valence		Pleasantness
Goal significance		Appetitive/aversive	
• Concern relevance	Focality	motives	Importance
• Outcome probability	Certainty	Certainty	Certainty
• Expectation	Presence		
• Conduciveness	Open/closed	Motive consistency	Perceived obstacle/
• Urgency	Urgency		Anticipated effort
Coping potential			
• Cause: agent	Intent/self–other	Agency	Human agency
• Cause: motive			
• Control	Modifiability	Control potential	Situational control
• Power	Controllability		
• Adjustment			
Compatibility standards			
• External	Value relevance		Legitimacy
• Internal			

Table 2. Examples of theoretically postulated appraisal profiles for different emotions

Stimulus evaluation checks	*Anger/rage*	*Fear/panic*	*Sadness*
Novelty			
• Suddenness	High	High	Low
• Familiarity	Low	Open	Low
• Predictability	Low	Low	Open
Intrinsic pleasantness	Open	Open	Open
Goal significance			
• Concern relevance	Order	Body	Open
• Outcome probability	Very high	High	Very high
• Expectation	Dissonant	Dissonant	Open
• Conduciveness	Obstruct	Obstruct	Obstruct
• Urgency	High	Very high	Low
Coping potential			
• Cause: agent	Other	Other/nature	Open
• Cause: motive	Intent	Open	Chance/neg
• Control	High	Open	Very low
• Power	High	Very low	Very low
• Adjustment	High	Low	Medium
Compatibility with standards			
• External	Low	Open	Open
• Internal	Low	Open	Open

Open-different appraisal results are compatible with the respective emotion.

the emotional significance of the events and objects depends upon the goals and the coping capacities of each individual. Thus, appraisal theory can comfortably predict that one person will respond to the same stimulus with fear while a second will respond with anger. Emotions are differentiated and can be associated with different physiological processes and facial expressions in this view, but

the antecedent of the emotion – the specific profile of appraisal – determines which discrete emotion is experienced.

COGNITIVE REPRESENTATION OF EMOTION

Individuals do not only experience emotions, they also verbally label emotions, represent their ideas of what emotions are as concepts, and can reflect on and remember their emotional feelings.

Emotion Labels

Subjective experience of emotion is largely revealed to scientists through emotion words. In the English language there are between 500 and 2000 words that refer more or less to emotions (e.g. shame), affective or valenced states (e.g. tranquility), and cognitive–affective states (e.g. vengeance). In Malay there are about 230 emotion words, and in Ifaluk there are only about 50. Clearly, then, although there is enormous cultural variation in the number of words individuals use to talk about emotions, they talk about emotions in all cultures. The basic categories of words, despite important differences, show general consistency across culture. (*See* **Cultural Psychology**)

The dimensional analysis of the meaning of emotion words usually reveals the same bipolar structure as does an analysis of individuals' ratings of their subjective emotional states. That is, the meaning of the labels are organized around the two dimensions of valence and activation or arousal. More specifically, a circular representation of emotion words, a circumplex structure, is typically observed. Figure 1 illustrates this structure, showing the position of the basic or discrete emotions as well as other affective states. The circumplex shows that individuals make fine-grained distinctions among emotion words in their reference to pleasantness and level of arousal.

Emotion Prototypes

The categories that emotion labels refer to share many properties in common with categories of concrete objects such as animals, vehicles and furniture, as well as more abstract concepts such as personality disorders, negotiation strategies and scientific theories. First, the defining characteristics of the categories vary in their diagnosticity, or probability of being present in any given episode of an emotion. For instance, individuals know that the experience of anger might almost always include the feeling of the face growing hot or the muscles being tense, but only sometimes the action

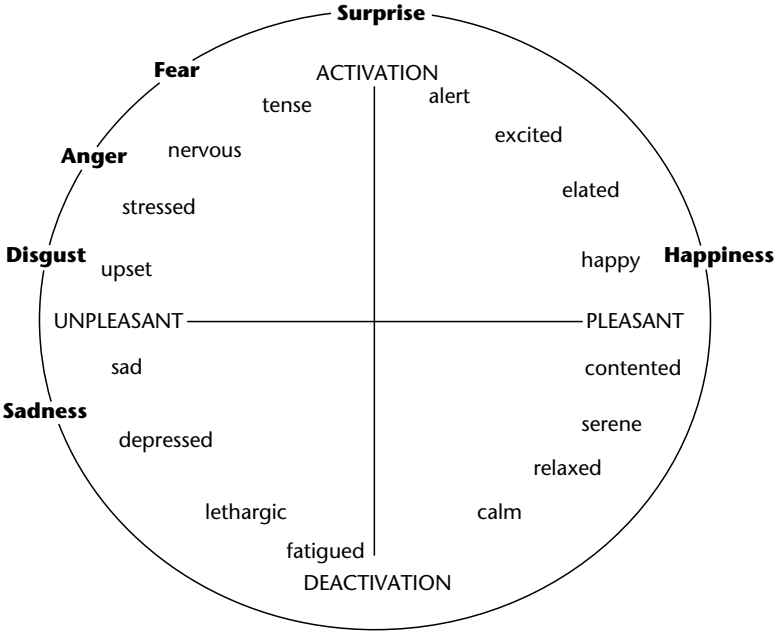


Figure 1. The circumplex structure of affective meaning that results from the multidimensional scaling of judgments of similarity between emotion words. The outer circle shows where several prototypical emotions fall. From Feldman Barrett L and Russell JA (1998) Independence and bipolarity in the structure of affect. *Journal of Personality and Social Psychology* **74**: 967–984.

of striking out. Second, membership in the categories is a matter of degree, and is not 'all or none'. Thus, a good example of feeling fear is a situation in which you see a bear and run away, heart pounding and mouth open to scream. But being on a carnival ride or jumping off a high diving board, with all the physiological and subjective experiences these entail, may be less good examples of fear. The best examples of each emotion are the central or prototypic examples of the categories, while the others are considered borderline or less prototypic examples. Finally, each category of emotion contains a script for the ordering and patterning of events involved in that emotion. Individuals are able to describe the antecedent events, the behavioral reactions to those events and the physiological and expressive responses involved in a way that preserves the temporal and spatial structure of a prototypic emotion episode. (*See* **Conceptual Representations in Psychology**; **Schemas in Psychology**)

Representation of Emotional States

Emotional feelings and experiences themselves are also represented in memory. Individuals can remember what emotions feel like and even reproduce aspects of the feelings if they retrieve memories of past episodes. It has been proposed that emotional states, perhaps the discrete emotions, are represented as informational units in an associative network in memory. In such associative network models the informational units are proposed to be linked to memories of the experience of the emotion and to memories for times when these emotions were induced. The existence of mental associations between a representation of the feeling of emotion and memories and perhaps the semantic categories of knowledge about the emotion has important consequences for understanding the processing of information during emotional states. (*See* **Memory Models**; **Spreading-activation Networks**)

EMOTION–COGNITION INTERACTION

Emotional states, and indeed emotional traits such as depression and anxiety, influence the manner in which individuals process information. Such states and traits can affect the allocation of attention, the efficiency of perceptual encoding, the retrieval of memories from long-term memory, the learning of new information, decision-making and judgment, as well as the organization of conceptual material in memory. (*See* **Affective Disorders: Depression and Mania**; **Depression**)

Emotion and the Content of Cognition

The associative network model makes a specific prediction concerning the influence of emotion in information processing, which is reminiscent of expressions such as, 'She was so happy that she saw the world through rose-colored glasses.' According to the model, emotional states activate the unit of information representing that state in long-term memory. Through passive diffusion of activation, other information that has been associated with or causal to that same emotional state becomes activated too, and then influences the ease and efficiency with which new information is perceived and stored. The consequence of such activation is the facilitated processing of emotion-congruent information. For example, in a happy state an individual will allocate attention to and rapidly encode smiling faces, people he or she likes, and words or phrases with positive meanings. One study also showed that individuals in happy states who were watching a happy expression gradually disappear from a face of another person, perceived the expression to still convey happiness for a significantly longer time – even when almost completely neutral – than did individuals in a sad state. This means that the happy individuals were particularly efficient at detecting signs of a happy expression that was ambiguously neutral. Other research has shown that, while in a happy state, individuals are also better able to learn and retrieve information associated with happiness. Finally, the judgments of individuals in happy states tend to be optimistic rather than pessimistic in nature. In a sense, emotions mobilize the entire cognitive system to process information that is relevant to that state. (*See* **Priming**)

Emotion and Information-processing Strategies

Emotions also influence the actual cognitive processes that are employed in judgment and decision-making: that is, the structure of those processes. When individuals are in positive emotional states they process information less deeply, in a simple, heuristic way. This does not mean that they cannot make careful judgments or that they think poorly, only that they tend to use short cuts for solving problems and may sacrifice attention to detail. The same appears also to be true for individuals who are in states of anger. On the other hand, sadness is associated with a more careful, less schematic type of processing. Individuals who are sad are likely to scrutinize information and to engage in

more systematic processing of that information. Research has shown, for example, that individuals in happy and angry states are more likely to base their judgments of the defendant in a hypothetical legal trial on racial stereotypes, than are individuals in sad or neutral states, who based their judgments more on the presented legal evidence.

There are a number of reasons why emotions are associated with more or less systematic styles of information processing. First, different emotions have different information values to the individuals experiencing them: happiness tends to indicate that the environment is safe and that all is going well; sadness indicates that there is a problem to be solved and that all is not going well; and anger suggests that fast action must be taken to solve a problem. Thus, happiness and anger have different signal values that nevertheless may result in the same tendency to process rapidly and in less detail. Sadness signals that careful attention to details of the environment is required. Second, happy states exert a high degree of cognitive load. Because happy states are associated with more information in memory – probably because individuals are on average happy more often than they are sad, angry or fearful – many different ideas come to mind during those states. The rush of associations can inhibit the ability to process additional incoming information in detail. In contrast, sad states, which are associated with less information stored in memory, at least among nondepressed individuals, tend not to trigger as many associations and tend therefore not to overload the system. Consequently individuals in sad states seem to have more capacity to process information in a systematic way. Third, different emotions have different motivational properties. Individuals who are feeling happy or angry may not be motivated to distract themselves from or otherwise change those states by processing new information in a systematic way. They protect their emotional states from change, albeit for different reasons. Individuals in nonclinical sad states, on the other hand, are often motivated to distract themselves from the sadness or engage in other mood regulation strategies. They are thus more likely to systematically process incoming information.

Emotion and Conceptual Organization

Finally, emotional states influence the way information is organized in memory, and therefore the content and structure of concepts. Natural categories in the environment are revealed by perceptual similarity among their members: it is easy to learn the category of birds because most examples of that category (most birds) have physical properties in common. Concepts, the representation of categories in memory, are therefore grounded in perceptual similarity. They are also organized theories that individuals possess about the origins and functioning of objects and events. However, during emotional states concepts may be organized somewhat differently. Rather than representing categories of things that belong together because they look alike or conform to a common theory of cause and effect, concepts may temporarily represent groups of objects and events that have elicited the same emotional state. For example, individuals are more likely during emotional states, compared with neutral states, to see emotional equivalence among events and objects, and group them together as being 'the same kind of thing', if they evoked the same emotion. (*See* **Conceptual Representations in Psychology**; **Concept Learning and Categorization: Models**)

REGULATION AND SUPPRESSION OF EMOTION

Emotion regulation includes processes that individuals use to influence whether they have an emotion, which emotions they experience and do not experience, the conditions under which they experience emotion, and how they express such states. Although Western cultures seem ambivalent about the need to regulate emotions, as conveyed in the conflicting expressions 'He who keeps a cool head prevails' versus 'Let your feelings be your guide', it is clear that individuals in all cultures attempt to control, alter, augment and suppress at least some emotions. Anthropologists and psychologists have observed specific social and cultural rules for doing so.

Emotion regulatory strategies are important to understand because they are fundamental to personality functioning and adjustment, and to psychological and physical health. Major depressive disorder, for example, is characterized by a deficit in experience of positive emotion and a surplus of processing negative emotions and negative information. Chronic hostility and anger inhibition are associated with hypertension and coronary heart disease. Emotion inhibition or suppression may enhance minor illnesses and accelerate the progression of major illness such as cancer.

Processes of Emotion Regulation

Given the multicomponent nature of emotions, it is not surprising that there are many different

processes involved in the regulation of emotional states, and that these processes operate more or less on physiological, expressive, and experiential aspects of the emotion.

Antecedent-focused regulation involves regulation strategies that are mobilized in the service of avoiding or enhancing emotions in an anticipatory way. Potentially emotion-arousing situations or stimuli can be avoided or approached, aspects of the situations can be selectively attended to or ignored, and, as suggested by appraisal theories of emotions, situations can be (re)evaluated in ways that augment or diminish the probability of specific emotions.

Response-focused regulation involves attempts to alter emotional states once they have been elicited. Representative strategies include the suppression or enhancement of expressive and behavioral components of the emotion, and distraction of conscious attention away from the subjective experience of the emotional state to other events internal or external to the individual, such as distracting thoughts and stimuli, respectively. An example of the regulation of an expressive component of emotion as a means to regulate emotional state is the voluntary manipulation of facial expression of emotion. This is the deployment of the adage for people who are depressed to 'Put on a happy face', for example, in order to feel happier. Putting on a happy face may indeed regulate emotional experience through a number of different mechanisms. First, it may alter the social environment such that more positive experiences are possible. In addition, self-perceptual processes – noticing that one is smiling – may lead individuals to believe that indeed they are happy and to realize this belief. Finally, several facial feedback theories of emotion also hold that facial expression of emotions actually feed back through physiological systems to influence subjective state. Specifically, the strategic use of facial musculature may alter other neurochemical processes involved in emotion.

Effects of Different Regulation Strategies

Emotion regulation strategies are differentially effective and have different costs and benefits over time. Of particular interest has been comparison of reappraisal of negative situations compared with the suppression of negative expressive behavior and feelings. Research has generally demonstrated that reappraisal can diminish the subjective experience and expressive aspects of an emotion. Of course, reappraisal of some negative situations is quite difficult in the first place. Suppression of emotional expression and behavior may often (or even always) be possible. However, suppression is associated with increases in sympathetic nervous system activation, and thus can be a physiologically costly strategy. It is probably for this reason that individuals who habitually suppress negative emotions tend also to suffer both minor and more permanent health consequences.

Further Reading

Eich E, Kihlstrom JF, Bower GH, Forgas JP and Niedenthal PM (2000) *Cognition and Emotion*. Oxford: Oxford University Press.

Ekman P and Davidson RJ (eds) (1994) *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Ekman P and Friesen WV (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* **17**: 124–129.

Feldman Barrett L and Russell JA (1998) Independence and bipolarity in the structure of affect. *Journal of Personality and Social Psychology* **74**: 967–984.

Forgas JP (2000) *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge, UK: Cambridge University Press.

Frijda NH (1986) *The Emotions*. Cambridge, UK: Cambridge University Press.

Gross JJ (1998) The emerging field of emotion regulation: an integrative review. *Review of General Psychology* **2**: 1–29.

Izard CE (1991) *The Psychology of Emotions*. New York: Plenum Press.

Lane RD and Nadel L (eds) (2000) *The Cognitive Neuroscience of Emotion*. Oxford: Oxford University Press.

Lewis M and Haviland JM (1993) *Handbook of Emotion*. New York: Guilford Press.

Niedenthal PM, Halberstadt JB and Innes-Ker AK (1999) Emotional response categorization. *Psychological Review* **106**: 337–361.

Öhman A (1987) The psychophysiology of emotion: an evolutionary-cognitive perspective. *Advances in Psychophysiology* **2**: 79–127.

Ortony A, Clore GL and Foss MA (1987) The referential structure of the affective lexicon. *Cognitive Science* **11**: 361–384.

Russell JA (1991) Culture and the categorization of emotions. *Psychological Bulletin* **110**: 426–450.

Scherer KR (1999) Appraisal theory. In: Dalgleish T and Power M (eds) *Handbook of Cognition and Emotion*. New York: John Wiley.

Environmental Psychology

Introductory article

Stephen Kaplan, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

In the context of evolution and environment, human information processing has many connections with preference and motivation, with far-reaching implications for the possibilities of enhancing human effectiveness.

INTRODUCTION

If we think of an environment as a pattern of information, then environmental psychology and cognitive science have much to learn from each other. Computer scientists have provided a useful concept, the 'software environment', which connects these disciplines. When dealing with a new software program one faces a pattern of information that is only a small piece of the total collection of information that makes up the program. Yet to be comfortable with the program and make good use of its capabilities one has to build some sort of mental map of the whole program. This example gives a fairly good idea of what constitutes an environment from an informational perspective.

The environment that early humans faced was not, of course, a software environment, but a physical environment: if they had not been able to master that environment, their future development as a species would not have been possible. To understand environmental psychology from a cognitive point of view it is necessary to consider these two quite different environments: the one in which humans evolved, and the one that humans now inhabit.

THE ENVIRONMENT AS A SOURCE OF INFORMATION

The ancestors of the human species were originally tree dwellers. As latecomers to the savanna, humans adopted a life at the ecological margins, scavenging where they could, taking advantage of opportunities wherever possible, and getting away from predators before it was too late. The story of human evolution is the story of how a creature with limited physical assets managed to survive by its wits, by developing powerful and ingenious ways of dealing with environmental information.

Information Processing for Survival

The typical early human was home-based yet hunted a vast territory (probably about 250 square kilometers), in which one could easily get lost, so that a good memory for spatial information was essential. In addition to spatial knowledge, several other information-processing capabilities would have been essential to a creature surviving at the ecological margins: recognition, prediction, evaluation, and action.

Quick recognition of important objects in the environment, along with a capacity to discern the essence of what is happening, to get to the gist of things, would have been imperative. Quick recognition of objects is no small challenge (one that is still very difficult for computers). Interpretation of a scene is even more difficult, as it involves both the set of important objects and their spatial configuration.

Still, quick recognition is not sufficient. It is necessary not only to understand the present situation, but to anticipate what might happen next. Thus, prediction is a vital aspect of information processing.

Even recognition and prediction are not enough. One must evaluate whether the current situation and what might come next are likely to have good or bad consequences. And one must do so quickly, so that one can act before one is overtaken by events. Reviewing many similar events that have happened before, and pondering whether or not one would want them to be repeated, are often

unaffordable luxuries. There must be a way to translate many different experiences into a quick and simple mandate for action.

The Cognitive Map

How can an animal with only a modest brain store so much spatial knowledge economically? The 'cognitive map' (some sort of durable mental model of the environment) provides a possible explanation. The capacity to recognize, think about and remember objects has been studied extensively. Considerable evidence suggests that experience leads to internal representations in the human mind, which stand for things in the world. These mental units make possible quick and confident recognition of objects despite the enormous variability in how they are experienced.

There is also evidence that people tend to associate things that occur near to one another in time; in other words, they readily form mental sequences that connect the representations of things that were experienced at about the same time. It is useful to think of these sequences as composed of 'nodes' (the representations of the objects) and 'paths' (the associations between them).

Sequences are necessary to cognitive maps, but not sufficient. What is needed is some way to integrate them into a coherent model of the environment. Suppose, for example, that there is a particular internal representation that stands for a landmark, such as the campus bell tower. Any sequence one experiences during one's walks on campus that include the bell tower will lead to the activation of this representation. Thus various routes will share this common node. Imagine, for example, that one is new on campus and only knows two routes: the 'weekday' pattern, from dorm to bell tower to classroom building, and the 'weekend' pattern, which includes the sequence from stadium to bell tower to cafeteria. With the knowledge of these two sequences, even if one has never traversed it, one also knows the route from dorm to bell tower to cafeteria. The cognitive map, therefore, is not a long list of separate sequences but a network of interlocking paths whose integration is due to their common elements.

Such an integration of nodes and sequences provides an economical way to store relevant information while ignoring much unnecessary detail. This cognitive map or mental model would have been a great help in avoiding getting lost, but it accomplishes much more than that. The capacity to form networks about the structure of the environment is likely to have served our distant ancestors even before the development of language. For example, early hunters, because of their limited weapons, had to attack potential prey from close proximity. In addition to spatial knowledge, they therefore needed considerable knowledge of prey species, including the capacity to recognize important signs and to interpret and predict behavior. The cognitive map is well suited to these challenges as well. The capacity to anticipate possible futures, and to recognize distinctive configurations and what they lead to, is called 'lookahead'. It is what enables people to use mental models to try out possible alternatives and ascertain possible consequences before making a decision. Given its antiquity, as well as its flexibility and generality, the cognitive map may in fact constitute a general-purpose way of storing knowledge that underlies much of what modern humans know and do.

INFLUENCES OF ENVIRONMENT ON AFFECT AND COGNITION

The importance of the cognitive map to the survival of early humans points to another issue, often neglected in cognitive science. Imagine an individual who finds being frequently confused or lost a satisfactory state of affairs. One can be reasonably sure that such an individual could not have been our ancestor. Indeed, although little note is taken of this fact in most current research and theory, modern humans, presumably like our ancestors, have strong negative feelings about being confused and, by contrast, find being knowledgeable and competent most enjoyable. Far from being totally separate domains, cognition and emotion (or affect, as it is sometimes called) are necessarily intimately linked. Strong feelings apply not only to the adequacy of our knowledge: they are also associated with the environments in which we feel competent or confused.

Preference and the Variables that Predict It

Experts trained to design landscapes consider form, line, color, and texture to be the essential elements of beauty in the landscape. However useful these concepts may be to the designer, an important component of what people like in the landscape is, of course, content. Trees, water, and smooth places to walk play an important role. There are, however, other properties of the preferred landscape that are about information rather than content. Two categories of informational factors have been identified: understanding (making sense of a setting

and expecting that one could venture into it without getting lost), and exploration (having much to look at and the possibility of learning more as one ventures into the setting).

In a sense these two informational categories are in opposition to each other. What one already understands, while providing safety, may not offer opportunities for new learning; and a place that offers new learning opportunities might also be a place where one can get lost. Not surprisingly, environments where both learning and staying oriented are possible are strongly preferred.

Together, the often competing forces of understanding and exploration provide a solution to an important problem. As a knowledge-based organism, the human must be pulled towards opportunities for learning, for acquiring new information. Thus, when one is in a highly familiar environment the support for understanding is likely to be high, so one's choice of direction will be influenced by the path that offers greater opportunity to learn. At the same time, it is important that the organism not be pulled too far from what it understands. If its survival depends on information, it is dangerous to move too deeply into unfamiliar territory where it is not clear what to do or how to interpret events. The strong preference for exploration and understanding together leads the organism to seek new information without getting too far from what it knows and can handle.

Affect and the Environment

Pleasure and pain are often triggered by actual situations. The lookahead capability of cognitive maps, however, also enables us to experience pleasure and pain with respect to events that have not yet happened. For example, if one dislikes public speaking, having to make a presentation two weeks from now can create great pain. In this case the painful event is a prediction, not a reality. Similarly, the expectation of a holiday can give one pleasure long before one departs.

What makes the human reaction to environments associated with pain and pleasure particularly interesting is the subtlety of the environmental dimensions that are coded for pleasure or pain. People readily respond with strong negative reactions when their abilities to understand or explore the environment are blocked, when they are made to look foolish or experience harm to their sense of self, or when they feel incompetent or helpless. Human competence depends on the capacity to process information effectively. An environment that makes this difficult therefore lowers competence and, in turn, safety. Thus, it is adaptive to experience pain in such environments, leading one to quickly leave the situation and avoid it in the future. Correspondingly, environments that support people's information-processing needs enhance competence and confidence, and are experienced as pleasurable.

STRESS AND MENTAL FATIGUE

While pleasure and pain are powerful forces for moving individuals away from 'bad' environments and towards 'good' ones, they are less helpful for dealing with information within an environment. The environment is highly complex. Most of what is in one's field of view at any one time is of little or no relevance to one's purposes. To function effectively in a vast sea of stimulation, selection is essential. Humans have a variety of mechanisms for selection; one of the most important of these is attention.

Of the several kinds of attention, two are particularly important to understanding the human–environment relationship. 'Fascination' refers to a kind of attention that is automatic. A fascinating stimulus is one that is hard not to attend to. In evolutionary times, what was fascinating was what was important for human purposes, so that the problem of what to attend to was often solved automatically and effortlessly. In the modern world, many factors, such as advertising and the media in general, have utilized our capacity for fascination to so great a degree that, rather than supporting one's purposes, it often serves the purposes of others. A fascinating stimulus that is irrelevant to one's purposes is a distraction. When distractions are many, it is necessary to employ a different kind of attention, called 'directed attention', to maintain one's focus on what is important.

As we have seen, recognition depends on a process whereby patterns in the environment activate corresponding mental structures. There is, however, a limit to how many such structures can be active at once ($5 +/- 2$, assuming that the patterns are well learned). If what uses up this precious limited capacity is just whatever is biggest or brightest or most fascinating in the environment, what goes on in the mind will be determined by the environment and not by one's purposes. An all-too-frequent consequence is that one's mental activity will be dominated by someone else's purposes. Even when environmental patterns do support a purpose that an individual is committed to, that support is not necessarily enduring. A change in

environmental configuration might present a new opportunity that fits some other purpose better. Thus, if one has no way of fending off distraction, environmental variability can undermine the persistence that is essential to effective goal-directed behavior. Some means of escaping from environmental control is necessary. Directed attention provides the mechanism for achieving this. This achievement, however, comes at a cost: directed attention requires effort and is subject to fatigue. Not surprisingly, the very contexts that require its use – resisting distraction, making sense of confusing situations, maintaining one's focus when what is important is not particularly interesting – are the very contexts that lead to such mental fatigue.

As fatigue increases one may eventually become aware of one's declining effectiveness, and this can cause stress. Stress can itself serve as a severe distraction, resulting in heavy demands on directed attention. Thus, although stress and fatigue are distinct, they often occur at the same time.

Environments that lead to mental fatigue or stress reduce confidence and competence while simultaneously increasing the sense of urgency and concern for self-preservation. One would hardly expect people to be helpful or responsive under such circumstances. In other words, environments that undermine competence and confidence will be unlikely to bring out the best in people.

USING THE ENVIRONMENT TO IMPROVE EFFECTIVENESS

The task of influencing one's own behavior is often achieved more readily by an indirect approach than by a direct one. Since people are responsive to environmental factors, arranging the environment to support appropriate behavior can be a particularly effective strategy. In addition, it does not call upon willpower, which depends on the same resource necessary for directed attention.

The attentional resource is perhaps the most important of the cognitive factors that one can do something about. Since mental fatigue is an ever-present threat and a serious handicap, ways to keep it under control and to recover from it are important interventions. Recent discoveries in environmental psychology are relevant here. Research on 'restorative environments' points to the importance of spending time in settings where fascination is readily available and where directed attention is little needed. Natural environments have been shown to serve this function particularly well. A related strategy is to make the everyday environment less costly. Such strategies as reducing distractions and avoiding noisy settings can protect directed attention from unnecessary demand. The focus on the attentional resource, which recasts the way one looks at one's environment, can have beneficial effects.

Both at the personal level and on a larger scale, strategies for achieving these benefits call for many small experiments to determine what works. In the individual case, 'know thyself' here takes on a more tangible meaning, as well as a greater urgency. The solutions, however, must go beyond the individual. People who are stressed and fatigued are not ideal neighbors. They contribute little to safety, stability, and civility. The pervasiveness of unreasonable behavior is one of the plagues of the modern world. Environments more responsive to human needs might help to temper this unfortunate trend.

Further Reading

Chown E, Kaplan S and Kortenkamp D (1995) Prototypes, location, and associative networks (PLAN): towards a unified theory of cognitive mapping. *Cognitive Science* **19**: 1–51.

Cimprich B (1993) Development of an intervention to restore attention in cancer patients. *Cancer Nursing* **16**: 83–92.

Clark A (1989) *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press.

James W (1892) *Psychology: The Briefer Course*. New York, NY: Holt.

Kaplan R, Kaplan S and Ryan RL (1998) *With People in Mind: Design and Management of Everyday Nature*. Washington, DC: Island Press.

Kaplan S (2001) Meditation, restoration and the management of mental fatigue. *Environment and Behavior* **33**: 480–506.

Kaplan S and Kaplan R (eds) (1978) *Humanscape: Environments for People*. Belmont, CA: Duxbury. [Republished by Ulrich's, Ann Arbor, MI, 1982.]

Kaplan S and Kaplan R (1982) *Cognition and Environment: Functioning in an Uncertain World*. New York, NY: Praeger. [Republished by Ulrich's, Ann Arbor, MI, 1989.]

Kearney AR and Kaplan S (1997) Toward a methodology for the measurement of the knowledge structures of ordinary people: The Conceptual Content Cognitive Map (3CM). *Environment and Behavior* **29**: 579–617.

Pfeiffer JE (1978) *The Emergence of Man*. New York, NY: Harper & Row.

Episodic Memory, Computational Models of

Intermediate article

Kenneth A Norman, University of Colorado, Boulder, Colorado, USA

CONTENTS

Computational models of episodic memory constitute mechanistically explicit theories of how we recall previously experienced events, and how we recognize stimuli as having been encountered previously. Because these models concretely specify the algorithms that govern recall and recognition, researchers can run computer simulations of these models to explore how (according to a particular model) different manipulations will affect recall and recognition performance.

INTRODUCTION

Episodic memory refers to our ability to remember specific, previously experienced events: we can *recall* (i.e. mentally re-create) previously experienced events, and we can *recognize* a stimulus as having been encountered previously. From a computational standpoint, the unifying feature of episodic memory tests is the need to isolate the memory trace corresponding to the to-be-remembered (*target*) event. Recall tests ask subjects to isolate the target event's memory trace in order to retrieve some missing detail, and recognition tests ask subjects to assess whether the target event is actually stored in memory. On both recall and recognition tests, good performance is contingent on the system's ability to screen out the effects of non-target memory traces. (*See* **Memory, Long-term**)

Computational models of episodic memory can be divided into two categories: *abstract* and *biological*. Abstract models make claims about the 'mental algorithms' that support recall and recognition judgments, without addressing how these algorithms might be implemented in the brain. The primary goal of these models is to account for challenging patterns of behavioral recall and recognition data from list learning paradigms (where stimuli are presented as part of a well-defined 'study episode'; then, subjects are asked to recall specific events from the study episode, or to discriminate between stimuli that were and were not presented during that episode). Biological models, like abstract models, make claims about the computations that support recall and recognition judgments; the main difference is that they also make specific claims about how the brain gives rise to these computations. This brain-model mapping provides an extra source of constraints on the model's behavior. (*See* **Memory Models**)

ABSTRACT MODELS

Abstract *global matching* memory models take a unified approach to recognition and recall. There are several different global matching models; this article will focus on a single, representative model, MINERVA 2 (Hintzman, 1988); see Clark and Gronlund (1996) for an explanation of differences between the various models. MINERVA 2 (M2) represents each study list item as a vector where each element equals 1, −1, or 0. Each element corresponds to a particular feature that may or may not be present in that item; a 1 value indicates that the feature is present, a −1 value indicates that the feature is absent, and a 0 value indicates that the feature value is unknown. Thus, the vectors corresponding to different items will overlap (i.e. have the same value for a particular vector element) to the extent that they consist of the same features. M2 posits that the vectors corresponding to different studied items are stored separately in memory (but see Murdock (1993) for an example of an abstract model that posits that memory traces are stored in a composite fashion; Murdock's TODAM model stores items by adding together vectors corresponding to different items). When an item is presented at test, M2 computes how well the test item vector matches all of the different vectors stored in memory. On recognition tests, M2 sums together all of these individual match scores to get a 'global match' (*familiarity*) score; recognition decisions are made by comparing familiarity scores to a criterion

value, i.e. respond 'studied' if an item's familiarity score exceeds the criterion value, respond 'nonstudied' otherwise. M2 implements recall by computing a weighted average of all of the items stored in memory, where each item is weighted by its match to the test probe. Thus, M2 generates a *vector* output on recall tests and a *scalar* output on recognition tests, but both recall and recognition depend on the same underlying match computation. (*See* **Memory Models**)

How Match is Computed

In global match models like M2, the match computation weights multiple matches to the same trace more highly than the same total number of feature matches, spread across multiple memory traces (e.g. a test cue that matches two features of one item yields a higher familiarity signal than a test cue that matches one feature each of two items). Put another way: the match computation is sensitive to whether the features of the test probe were studied together (vs. separately). M2's match computation achieves this *sensitivity to conjunctions* by first computing the dot product of the cue vector and each memory trace vector (this gives the proportion of matching features minus the proportion of mismatching features for each memory trace) and then cubing the dot product score for each trace; finally, M2 adds together the cubed match scores to get the familiarity score for that cue. This algorithm yields sensitivity to conjunctions insofar as matches spread across multiple stored traces are combined in an additive fashion, but – because of the cube rule – multiple matches to a single trace are combined in a positively accelerated fashion.

Sensitivity to conjunctions ensures that one strong match outweighs the effects of several weak matches. Given the fact that different episodes share features to some extent, it is inevitable that test probes will match at least one feature from several memory traces other than the target trace. In models that lack sensitivity to conjunctions, these small matches – in aggregate – would swamp the one large match score associated with the target trace. Figure 1 illustrates how sensitivity to conjunctions (implemented using the cube rule) helps reduce interference in M2.

Modeling Interference Data

List length effects

While sensitivity to conjunctions minimizes interference caused by low amounts of cue-trace overlap, there is no way to completely eliminate interference caused by higher amounts of cue-trace overlap. There will always be non-target memory traces that, by chance, match the test probe strongly; these strong, spurious matches add noise to the global match signal and degrade performance. All global matching models predict that adding new items to the list (increasing *list length*) will impair both recognition and recall, by increasing the odds that a strong (but spurious) match will occur. In keeping with this prediction,

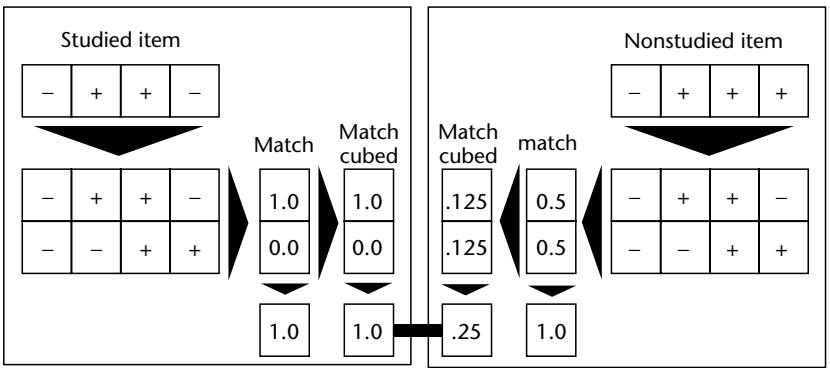


Figure 1. Illustration of how MINERVA 2 (M2) computes global match. For each of the two traces stored in memory, M2 computes a match value = (number of matching features – number of mismatching features)/(total number of nonzero features); then, M2 cubes these match values and adds them together to get a familiarity score for that cue. In this example, if match values are summed prior to cubing, the summed match values are equivalent for the studied item and the nonstudied item. However, if match values are summed after cubing, the studied item generates a much larger summed match score than the nonstudied item. Cubing benefits discrimination by minimizing the effect of weak (partial) matches on the summed match score, relative to the effect of more complete matches.

a very large number of studies have obtained list length effects for recognition and recall, although it is becoming clear that list length effects are not always obtained for recognition (see Dennis and Humphreys (2001) for discussion of this issue). (*See* **Catastrophic Forgetting in Connectionist Networks**)

Modeling the null list strength effect

One finding that global match models initially failed to predict is the null list strength effect (LSE) for recognition: Ratcliff *et al*. (1990) found that strengthening some list items, by presenting them repeatedly or for a longer duration, does not impair recognition of other (non-strengthened) list items. In models like M2, strengthening is operationalized by storing extra copies of an item to memory, or by increasing the probability of successful feature encoding. Both of these manipulations increase the global match score triggered by the strengthened item (thereby allowing the model to accommodate the finding that strengthening an item improves memory for that item). However, strengthening an item's memory trace also increases the mean and variance of the global match signal triggered by *other* items (intuitively: random, spurious match between the test probe and memory trace X has a larger effect on the global match signal when X is *strong* vs. when X is *weak*); this increase in variance leads to decreased recognition performance.

Researchers have been working from 1990 to the present to modify global matching models so they can accommodate the null recognition LSE obtained by Ratcliff *et al*. (1990). One promising approach to modeling the null LSE has been to posit that *differentiation* occurs as a consequence of strengthening (Shiffrin *et al*., 1990); the gist of differentiation is that as participants acquire experience with an item, the item's representation becomes increasingly refined, making it less likely that it will spuriously match some other item at test.

One example of an abstract model that incorporates differentiation is the REM model described by Shiffrin and Steyvers (1997) (see McClelland and Chappell (1998) for a similar model). REM uses a 'match' rule, based on Bayesian statistics, that computes the odds that the test probe and the stored trace are the *exact same item*. With this rule, a small amount of mismatch can have a very large effect; if you are certain that a memory trace contains a particular feature, and you are certain that the test probe does *not* contain that feature, the match value will be zero. In REM, strengthening a memory trace increases the number of encoded features; strong traces are less likely to trigger a spurious match because they are more likely to contain features that mismatch the test probe. Figure 2 provides a simple illustration of how strengthening reduces spurious matches in models like REM.

Single-process vs. dual-process approaches

An important feature of abstract global match models is that they attempt to explain recognition performance entirely in terms of the scalar familiarity signal. This *single-process* approach contrasts with *dual-process* theories of recognition, which posit that both familiarity and recall contribute to recognition performance (i.e. items can be called 'old' because they trigger a nonspecific feeling of familiarity, or because the subject specifically recalls some detail from when the item was studied). Furthermore, the most prevalent dual-process theory (Jacoby *et al*., 1997) posits that the operating characteristics of familiarity and recall are qualitatively distinct; according to this theory, familiarity is a signal-detection process (i.e. studied-item and lure familiarity are both normally distributed, and the two distributions overlap extensively) but recall is a high-threshold process (i.e. recall is all-or-none; studied items are sometimes called 'old' based on recall, but lure items are never called 'old' based on recall). Note that recall can be applied to recognition in abstract models like M2 (e.g. by cuing with the test item and comparing the recalled vector to the test item – if they match above a certain threshold, say 'old'). However, recall-based recognition and familiarity-based recognition have very similar operating characteristics in abstract models, because they are based on the same underlying match computation; as such, adding a recall process typically does not affect these models' recognition predictions. The only time that adding recall affects these models' performance is on recognition tests where some kind of content has to be retrieved (e.g. an exclusion test, where subjects have to say 'old' to studied words from one list and 'new' to studied words from another list). Also, some models derive different predictions for recall and familiarity based on the assumption that subjects cue memory differently when they are trying to recall items, i.e. they are more likely to incorporate context into the retrieval cue (Shiffrin *et al*., 1990); however, this is not a difference between recall and familiarity *per se*.

It would be possible to build an abstract model that uses different match rules for recall and familiarity-based recognition, in keeping with the idea that these systems have distinct operating

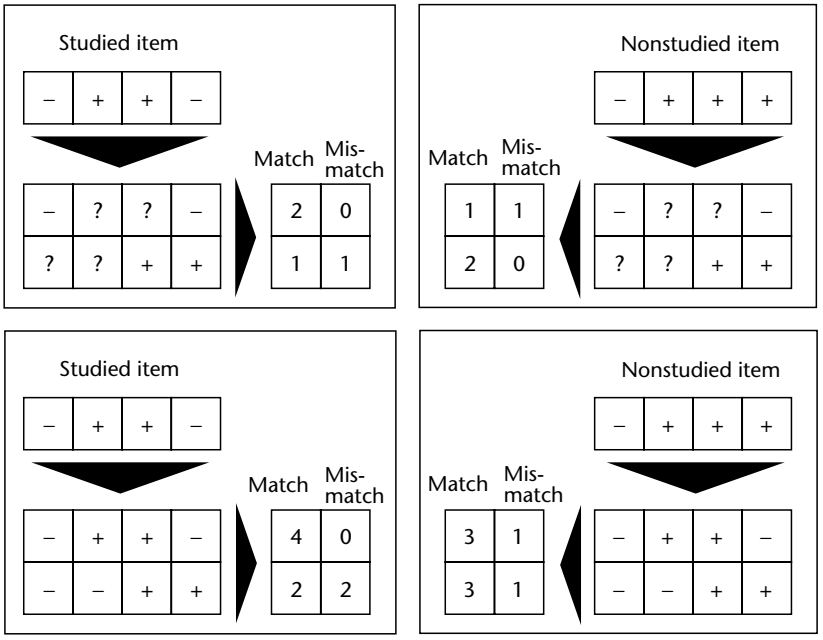


Figure 2. Illustration of how strengthening – operationalized as more complete feature encoding – can reduce spurious matches. Question marks indicate features that were not encoded at study. Prior to strengthening (upper two panels), the studied item and the nonstudied item match stored traces equally well; for both items, there is one trace that appears to match the test cue, i.e. there are some matching features, and no mismatching features. After strengthening (lower two panels), it is apparent that the nonstudied item does not exactly match either of the stored traces.

characteristics. This has not occurred because, once recall and familiarity are allowed to use different match rules, it is unclear how to constrain these (separate) systems based purely on behavioral data. Extant techniques for measuring the separate contributions of recall and familiarity to behavioral recognition performance are controversial because they rely on assumptions about the properties of recall and familiarity that cannot be tested empirically (e.g. that recall and familiarity are stochastically independent; for more on these techniques, see Jacoby *et al.*, 1997).

BIOLOGICAL MODELS

One way to further constrain dual-process models is to incorporate information about how recall and familiarity are computed in the brain. Biological models of episodic memory, like abstract models, try to account for the widest possible range of behavioral findings; however, unlike abstract models, biological models incorporate explicit claims about how the brain gives rise to recognition. Biological models of episodic memory have focused largely on the hippocampus, because neuropsychological data unequivocally indicate that the hippocampus is necessary for recall.

Modeling Hippocampal Contributions to Episodic Memory

Over the past decade, several researchers have developed biologically detailed computational models of the hippocampus, with the goal of explaining how the hippocampus contributes to episodic memory (Norman and O'Reilly, in press; Hasselmo and Wyble, 1997). The aforementioned models all view the hippocampus as a machine that is specialized for rapidly storing patterns of cortical activity ('episodes') in a manner that minimizes interference and allows for *pattern completion*: subsequent recall of entire stored patterns in response to partial cues. Furthermore, these models make similar – albeit not identical – claims about how different hippocampal substructures contribute to this process. This article will focus on the Norman and O'Reilly Complementary Learning Systems (CLS) neural network model, which has a hippocampal component (described in this section) and a cortical component (described in the next section); a schematic diagram of the CLS hippocampal network is shown in Figure 3. (*See* **Hippocampus**)

In the CLS model, the hippocampal network binds together sets of co-occurring neocortical

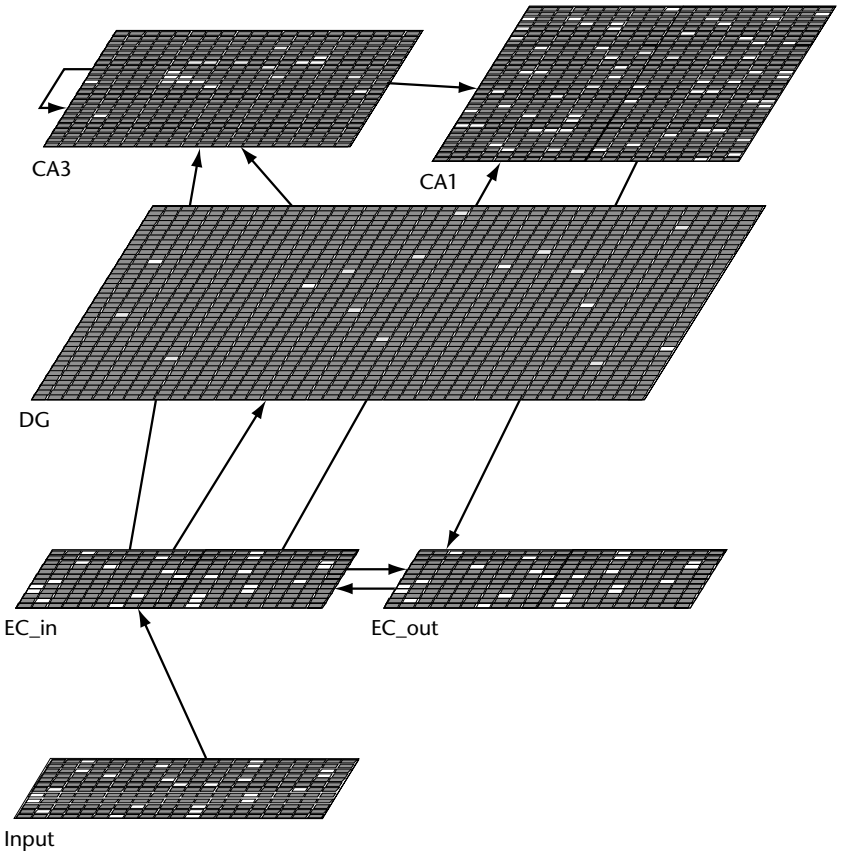


Figure 3. Diagram of the CLS hippocampal network. The hippocampal network links input patterns in entorhinal cortex (EC) to relatively non-overlapping (*pattern-separated*) sets of units in region CA3; recurrent connections in CA3 bind together all of the units involved in representing a particular EC pattern; the CA3 representation is linked back to EC via region CA1. Learning in the CA3 recurrent connections, and in projections linking EC to CA3 and CA3 to CA1, makes it possible to recall entire stored EC patterns based on partial cues. The dentate gyrus (DG) serves to facilitate pattern separation in region CA3; see O'Reilly and McClelland (1994) for details.

features (corresponding to a particular episode) by linking co-active units in entorhinal cortex (EC – the neocortical region that serves as a gateway to the hippocampus) to a cluster of units in region CA3 of the hippocampus; these CA3 units serve as the hippocampal representation of the episode. Recurrent connections between active CA3 units are strengthened. To allow for recall, active CA3 units are linked back to the original pattern of cortical activity via region CA1. Learning in the model occurs according to a Hebbian rule whereby connections between units are strengthened if both the sending and receiving units are active, and connections are weakened if the receiving unit is active but the sending unit is not.

At test, when a partial version of a stored EC pattern is presented to the hippocampal model, the model is capable of reactivating the entire CA3 pattern corresponding to that item because of learning that occurred at study; activation then spreads from the item's CA3 back to the item's EC representation (via CA1). In this manner, the hippocampus manages to retrieve a complete version of the EC pattern in response to a partial cue.

To minimize interference between episodes, the hippocampus has a built-in bias to assign relatively non-overlapping (*pattern separated*) CA3 representations to different episodes. Pattern separation occurs because hippocampal units are sensitive to conjunctions of neocortical features; given two neocortical patterns with 50% feature overlap, the probability that a particular conjunction of features will be present in both patterns is much less than 50% (see O'Reilly and McClelland (1994) for a much more detailed treatment of pattern separation in the hippocampus, and for discussion of the role of the dentate gyrus in facilitating pattern separation). The hippocampal model is sensitive to conjunctions because it uses *sparse* representations (where this sparseness is enforced by inhibitory

competition); in the model, a given input pattern only activates about 4% of the units in CA3. Inhibitory competition forces units to compete to represent input patterns, and units that are sensitive to multiple features of a given input pattern (i.e. feature conjunctions) are more likely to win the competition than units that are only sensitive to single input features.

A key property of neural network models is that some degree of structural interference between memory traces at storage is inevitable, assuming that there is overlap between memory traces (i.e. different items activate the same units). Whenever there is overlap, sensitivity to features that are shared across items increases, but sensitivity to features that are unique to specific items decreases. Pattern separation mechanisms in the hippocampus reduce structural interference (effectively preventing *catastrophic interference*, where studying new items totally wipes out stored memory traces) but do not eliminate interference entirely. The view that degradation is inevitable contrasts strongly with models like MINERVA 2 and REM, which posit that memory traces with overlapping features can be stored separately, with no structural degradation (but do not explain how this could come about). (*See* **Catastrophic Forgetting in Connectionist Networks**)

The raw output of the CLS hippocampal model is a vector comprising recalled information. Norman and O'Reilly (in press) apply the model to recognition by comparing the output vector (recall) to the input vector. If recalled information matches the test probe, this constitutes evidence that the test probe was studied; if recalled information mismatches the test probe, this is evidence that the test probe was not studied; specifically, Norman and O'Reilly compute a recall score equal to the number of matching features minus the number of mismatching features. While the hippocampal model shows good recognition discrimination in standard list-learning paradigms, there are several findings in the recognition literature that the hippocampal model, taken by itself, cannot explain. For example, the hippocampal model tends to underpredict false recognition. Because of pattern separation, test cues have to overlap strongly with studied patterns in order to activate the CA3 representations of these studied patterns (thereby triggering recall); as such, the hippocampal model predicts that nonstudied lure items should not trigger any recall, unless they are highly similar to studied items. This prediction conflicts with the finding that false recognition rates are typically well above zero in list-learning experiments.

Modeling Neocortical Contributions to Episodic Memory

One way to accommodate these issues with the hippocampal model is to argue that the hippocampus is not the only structure that contributes to recognition memory. Consistent with this view, neuropsychological data indicate that *medial temporal neocortex* (MTLC) also contributes to recognition – patients with hippocampal damage but spared MTLC perform well above chance on recognition tests. Norman and O'Reilly (in press) have also constructed a neural network model of MTLC to explore how this structure contributes to recognition memory. In keeping with the complementary learning systems view set forth by McClelland *et al*. (1995), Norman and O'Reilly posit that the primary function of neocortex (including MTLC) is to integrate across episodes to learn about the statistical structure of the environment. In contrast to the hippocampal model, which is biased to assign distinct representations to episodes and uses a large learning rate (thereby allowing it to quickly memorize individual episodes), the MTLC model assigns overlapping representations to similar episodes (thereby allowing it to represent what these episodes have in common) and uses a relatively small learning rate.

A schematic diagram of the MTLC model is shown in Figure 4. Because cortex uses a small

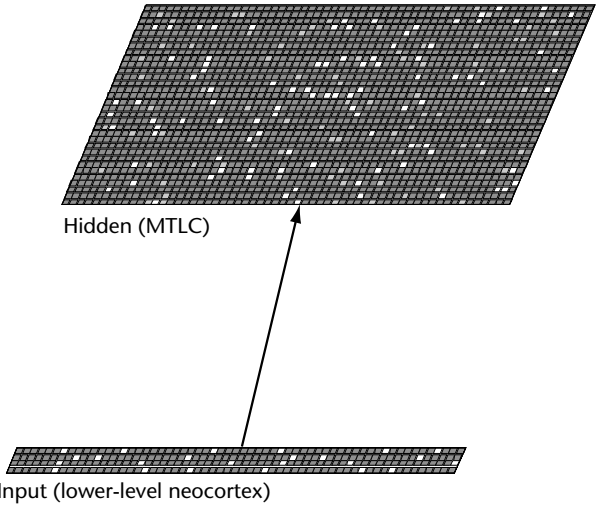


Figure 4. Diagram of the CLS cortical network. The cortical network consists of two layers, an input layer (corresponding to 'lower' cortical regions that represent basic features of input patterns) and a hidden layer (corresponding to MTLC). Units in the hidden layer compete to encode (via Hebbian learning) regularities that are present in the input layer.

learning rate, it is not capable of pattern completion (recall) following limited exposure to a stimulus. However, it is possible to extract a scalar signal from the MTLC model that reflects stimulus familiarity. In the MTLC model, as items are presented repeatedly, their representations in MTLC become *sharper*: novel stimuli weakly activate a large number of MTLC units, whereas familiar (previously presented) stimuli strongly activate a relatively small number of units. Sharpening occurs because Hebbian learning specifically tunes some MTLC units to represent the stimulus. When a stimulus is first presented, some MTLC units, by chance, will respond more strongly to the stimulus than other units; these units get tuned by Hebbian learning to respond even more strongly to the item the next time it is presented; and these strongly active units start to inhibit units that are less strongly active. To index representational sharpness – and through this, stimulus familiarity – we measure the average activity of the MTLC units that win the competition to represent the stimulus. Because there is more overlap between representations in MTLC than in the hippocampus, the MTLC signal has very different operating characteristics than the hippocampal recall signal. Whereas lures rarely trigger hippocampal recall, Norman and O'Reilly (in press) showed that the MTLC signal tracks, in a graded fashion, how similar the test probe is to studied items.

Some Predictions of the CLS Model

Effects of hippocampal lesions

Because the CLS model maps clearly onto the brain, it is possible to use the model to address neuroscientific data in addition to (purely) behavioral data. For example, the model makes predictions about how different kinds of medial temporal lesions will affect episodic memory. One prediction is that hippocampal lesions should impair performance on yes–no recognition tests with *related lures* (i.e. lures that are similar to specific studied items) more so than on tests with *unrelated lures*. When lures are not highly similar to studied items, both systems (MTLC and hippocampus) discriminate well, but when lures are similar to studied items the hippocampus outperforms MTLC because of its ability to assign distinct representations to similar stimuli, and its ability to reject lures when they trigger recall that *mismatches* the test probe. For evidence in support of this prediction see Holdstock *et al*. (in press).

Interference: a challenge for biological models

While the biological approach to episodic memory modeling has led to new insights into episodic memory (and the brain basis thereof), this approach faces several challenges. One major challenge is accounting for the effects of *interference* (e.g. list length, list strength) on recognition and recall. As discussed above, biological models generally predict some degree of structural interference between memory traces at study, i.e. learning about one item degrades the memory traces associated with other items. Several researchers have questioned whether models that posit structural interference at storage could account for the null list strength effect on recognition sensitivity, because of this pervasive tendency towards trace degradation. However, Norman and O'Reilly (in press) showed that biologically realistic neural network models with overlapping representations are, in fact, capable of accommodating the null list strength finding. The CLS cortical model predicts a null list strength effect for recognition sensitivity, given low-to-moderate levels of input pattern overlap, because (initially) the model's responding to lures decreases as much as its responding to studied items as a function of interference; as such, the distance between the studied-item and lure-item familiarity distributions stays relatively constant, and discriminability does not decrease.

Importantly, the CLS model also predicts that list strength effects should be obtained for the hippocampal recall process. In the hippocampal model, interference degrades the model's ability to recall studied items, and recall of lures is typically near zero (and thus can not decrease); because of this floor effect on lure recall, interference has the effect of pushing together the studied-item and lure-item recall distributions, leading to decreased discriminability. For evidence in support of the CLS model's list strength predictions, see Norman (in press).

SUMMARY

Abstract episodic memory models like M2 provide an elegant account, at the algorithmic level, of our ability to recall and recognize specific events from our personal past. These models posit that recall and familiarity rely on the same 'match' rule and thus have similar operating characteristics. A potential weakness of abstract models is that they do not consider the *neural plausibility* of these algorithms; it is very difficult to see how features

of some abstract models (e.g. the total absence of structural interference between traces at study in M2 and REM) could be implemented in the brain.

Recently developed biological episodic memory models seek to remedy this by establishing a clear isomorphism between parts of the model and parts of the brain that have been implicated in episodic memory (e.g. in neuropsychological studies). The Norman and O'Reilly CLS model posits that recall and familiarity have different operating characteristics, insofar as they rely on distinct neural structures – the hippocampus and medial temporal neocortex – that differ in their architecture and connectivity. The hippocampus is more sensitive to feature conjunctions than cortex, which in turn leads to less overlap between representations. Low overlap makes it possible for the hippocampus to rapidly memorize patterns without catastrophic interference (although interference still occurs), and it also decreases the probability of false recognition, relative to what occurs in cortex. (*See* **Neural Basis of Memory: Systems Level**)

References

Clark SE and Gronlund SD (1996) Global matching models of recognition memory: how the models match the data. *Psychonomic Bulletin & Review* **3**: 37–60.

Dennis S and Humphreys MS (2001) A context noise model of episodic word recognition. *Psychological Review* **108**: 452–478.

Hasselmo ME and Wyble BP (1997) Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural Brain Research* **89**: 1–34.

Hintzman D (1988) Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* **95**: 528–551.

Holdstock JS, Mayes AR, Roberts N *et al*. (2002) Under what conditions is recognition relatively spared relative to recall after selective hippocampal lesions? *Hippocampus* **12**: 341–351.

Jacoby LL, Yonelinas AP and Jennings JM (1997) The relation between conscious and unconscious (automatic) influences: a declaration of independence. In: Cohen JD and Schooler JW (eds) *Scientific Approaches to Consciousness*, pp. 13–47. Mahwah, NJ: Erlbaum.

McClelland JL and Chappell M (1998) Familiarity breeds differentiation: a Bayesian approach to the effects of experience in recognition memory. *Psychological Review* **105**: 724–760.

McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex. Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–457.

Murdock BB (1993) TODAM2: a model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review* **100**: 183–203.

Norman KA (in press) Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning and Cognition*.

Norman KA and O'Reilly RC (in press) Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*.

O'Reilly RC and McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a tradeoff. *Hippocampus* **4**: 661–682.

Ratcliff R, Clark SE and Shiffrin RM (1990) The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 163–178.

Shiffrin RM, Ratcliff R and Clark SE (1990) The list strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 179–195.

Shiffrin RM and Steyvers M (1997) A model for recognition memory: REM: retrieving effectively from memory. *Psychonomic Bulletin and Review* **4**: 145–166.

Further Reading

Gillund G and Shiffrin RM (1984) A retrieval model for both recognition and recall. *Psychological Review* **100**: 546–567.

Grossberg S and Stone GO (1986) Neural dynamics of word recognition and recall: attentional priming, learning, and resonance. *Psychological Review* **93**: 46–74.

Hasselmo ME and McClelland JL (1999) Neural models of memory. *Current Opinion in Neurobiology* **9**: 184–188.

Hintzman DL (1990) Human learning and memory: connections and dissociations. *Annual Review of Psychology* **41**: 109–139.

Howard MW and Kahana MJ (2002) A distributed representation of temporal context. *Journal of Mathematical Psychology* **46**: 269–299.

Humphreys MS, Bain JD and Pike R (1989) Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. *Psychological Review* **96**: 208–233.

Mensink G and Raaijmakers JG (1988) A model for interference and forgetting. *Psychological Review* **95**: 434–455.

Murnane K and Shiffrin RM (1991) Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**: 855–874.

Nobel PA and Huber DE (1993) Modeling forced-choice associative recognition through a hybrid of global recognition and cued-recall. Proceedings of the 15th Annual Conference of the Cognitive Science Society, pp. 783–788.

Raaijmakers and Shiffrin RM (1992) Models for recall and recognition. *Annual Review of Psychology* **43**: 205–234.

Ratcliff R (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review* **97**: 285–308.

Ratcliff R, Van Zandt T and McKoon G (1995) Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General* **124**: 352–374.

Ratcliff R and McKoon G (2000) Memory models. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 571–581. New York, NY: Oxford University Press.

Rolls ET and Treves A (1998) *Neural Networks and Brain Function*. New York, NY: Oxford University Press.

Event-related Potentials and Mental Chronometry

Introductory article

Edward L Wilding, Cardiff University, Cardiff, Wales, UK
Jane E Herron, Cardiff University, Cardiff, Wales, UK

CONTENTS

Mental chronometry is the label that has been given to the use of reaction times to identify and characterize the stages of human information processing that occur between a stimulus and a response. Event-related potentials (ERPs) are also suited to this, since they index in real time the brain processes that reflect successive stages of human information processing.

INTRODUCTION

Cognitive scientists study and attempt to explain the workings of a complex information-processing system, namely the human brain. One particularly influential concept in this endeavor has been the computer metaphor, which incorporates the view that human cognition can be understood by comparison with the ways in which computers process information. Consideration of the limitations and operating characteristics of computers yields several features that might also be true of the human information-processing system. For example, information is often processed in a series of stages, at least some of those stages have capacity limitations, and each processing stage takes time to complete.

The challenge for cognitive psychologists is to develop appropriate models that describe these information-processing stages, and then to test the models in experiments. There are several approaches that can be taken in pursuit of this goal. These include (1) the implementation of computational models that mimic the way in which the system might work, (2) the study of individuals with selective brain damage who have deficits at particular information-processing stages, and (3) the use of reaction times, measures of accuracy, and/or online measures of brain activity to test and generate predictions of a model. It is a subset of this third approach – which can be defined as *mental chronometry* – that will be the focus of the rest of this article.

MENTAL CHRONOMETRY

The term *mental chronometry* was introduced in the late 1970s to describe (primarily) the use of reaction-time measures to gain a better understanding of the structure, function, and dynamics of the human information-processing system (often referred to as the *cognitive architecture*). The main idea is that differences between reaction times that are obtained in different tasks can be employed to determine the dynamics of human information processing, and to infer the presence or absence of certain processing stages, as well as the way in which those processing stages work.

One set of tasks that has been widely used in order to investigate human information processing involves systematically manipulating the relationship between the types of stimuli that are presented and the types of responses that are to be made to them. One reason for doing this is to investigate the operation of two stages that intervene between a stimulus and a response. These stages are *stimulus discrimination* and *response selection*. The idea is that by designing tasks which challenge these processing stages in different ways, and by using reaction times as a measure of performance, it will be possible to determine some of the characteristics of these two stages.

Imagine, for example, that three tasks all involve presentation of a series of pictures one at a time, and that each picture is either a blue circle or a red circle. Each task differs in what you know about the picture that will be presented next and/or in the

rules with regard to when you should and should not respond by pressing a key. The different requirements are described briefly below.

Task Structure

- *Task 1 (simple reaction time)*. You know ahead of time that the circle will always be blue. All you have to do is to respond as quickly as possible by pressing a key when the blue circle appears.
- *Task 2 (choice reaction time)*. You do not know what color of circle will be presented in each trial. You must respond as quickly as possible with the appropriate hand (e.g., left for blue, right for red) when the circle appears.
- *Task 3 (go/no-go)*. This is the same as for task 2, except that you must respond to red circles (go trials) and make no response to blue circles (no-go trials).

Comparison of the reaction times in these three tasks allows estimates to be made of the time courses of stimulus discrimination and response selection. Consider the differences between tasks 1 and 2. Only the latter requires you to determine the color of the circle and to map the color on to the appropriate response, because in task 1 you know ahead of time which hand to respond with. Thus any difference in reaction time between these tasks probably reflects a combination of the time taken for stimulus discrimination *plus* the time taken for response selection.

Now consider tasks 1 and 3. In this case, the primary disparity is the requirement for stimulus discrimination in task 3 only, while in both cases the response selection requirement is the same – there is only one response to make. Triangulating the reaction times in the three experiments thus provides a means of estimating separately the times that are taken for stimulus discrimination and for response selection.

LIMITATIONS OF REACTION-TIME MEASURES

The example described in the previous section gives some idea of how reaction times can be used to investigate characteristics of the cognitive architecture. This particular approach has been termed the *subtraction method* because it relies on a comparison of the differences between reaction times that are obtained in tasks which are assumed to differ in terms of the particular information-processing stages that they challenge. The principal limitation of the use of reaction times to study the time course of cognitive processes stems from the fact that it is reasonable to assume that multiple, possibly temporally overlapping, processing stages intervene between a stimulus and a response. Reaction times provide only an estimate of the total time that is taken to progress between stimulus and response. The concern that a particular experimental manipulation does not isolate the stage or stages of interest is one that is difficult to dispense with altogether. Indeed, much of the discussion of the appropriate framework within which to employ reaction times optimally in the study of human information processing reflects these concerns.

The ideal measure of the processing stages that intervene between stimulus and response would be one that can track *in real time* the progression of the cognitive operations that are engaged between these two points. Examining the way in which these measures of particular cognitive operations change as a result of cognitive challenges would then allow more precise claims about the time course and the architecture of the human information-processing system than is available from reaction-time measures alone. The acquisition of event-related potentials (ERPs) is one approach that could potentially bridge the gap between stimulus and response.

EVENT-RELATED POTENTIALS (ERPS) AND MENTAL CHRONOMETRY

Information transmission in the brain is achieved by means of electrical signals. These signals can be recorded from electrodes that are placed on the scalp. There are several types of such recordings, including *event-related potentials (ERPs)*, which are small changes in short segments of the ongoing electrical activity of the brain, revealed by averaging across several segments of the same type. ERPs can track changes in brain activity on a millisecond-by-millisecond basis, and by starting recording precisely when a stimulus (e.g., a red circle) is presented, ERPs may provide insights into the processing stages that commence with stimulus presentation.

ERPs are typically recorded from multiple sites on the head, and the activity reflecting the operation of different cognitive processes is detectable at different sites. Of most relevance to mental chronometry is the fact that an ERP recorded from a single site can be regarded as a graph that plots changes in brain electrical activity over time. Each graph consists of a series of positive and negative deflections, and these can be thought of as reflecting the different processing stages that are set in train by a stimulus. (*See* **Visual Evoked Potentials**)

One line of psychological inquiry involves identification of the cognitive processes that are indexed by particular deflections. For at least some of these there is a degree of consensus as to which information-processing operations they reflect, and this is important if ERPs are to be used as a tool to complement the information about human information processing that is provided by reaction-time studies. The core assumption of the ERP approach to mental chronometry is that inferences about cognitive architecture can be made by observing the way in which these positive and negative deflections vary across different psychological tasks.

There are at least two ways in which ERP deflections can vary across tasks. First, the amplitude (or size) of a deflection can change. This is generally taken to mean that the cognitive process indexed by that deflection is active to a greater extent in the task where the deflection is largest. Second, the time after a stimulus at which a deflection is largest (the *peak latency*) can differ. In this case, it is assumed that the time course of the cognitive process that is indexed by the deflection of interest differs across tasks. In some cases, amplitude and latency changes can occur together, while in others only one of the two will change. The following example gives more substance to this abstract characterization, and is concerned with changes in the amplitude of deflections only.

THE LATERALIZED READINESS POTENTIAL (LRP)

Some ERP deflections are given names that describe, at least in part, the psychological processes that they index, and the lateralized readiness potential (LRP) is one example of this. The LRP is considered to reflect preparatory motor processes, and can therefore be used as an index of covert preparation for movement. The LRP, as the label implies, is larger over the left or the right hemisphere of the brain depending on which hand the participant is preparing to respond with. The LRP is largest over the hemisphere opposite the hand with which the participant is preparing to respond. The LRP is typically evident in the electrical record during the period immediately before a response is made. The presence of a measurable LRP is a very strong indication that a participant is preparing to make a response.

In studies of mental chronometry, measurements of the LRP have been used to determine what sources of information influence the selection of a particular motor response in cognitive tasks. This question is interesting because it is concerned with the way in which the response-selection system works. According to one account, response selection begins only after complete evaluation of a stimulus has occurred. However, according to an alternative account, partial information can be sufficient to influence response selection. The experiment described below provides an example of how it might be possible to determine which of these competing accounts is correct.

Consider a variant of the go/no-go task described earlier. In that task, the stimuli were either red or blue, and the requirement was to respond to red stimuli only. In this new task, the stimuli are again either red or blue, and in addition they are either the letter N or the letter M. Crossing all possible combinations of colors and letters means that there are four different stimuli. In this task, the requirement is to respond to a red M only. The question is whether a larger LRP is observed for the red N than for the two blue stimuli. Observation of the LRP for the red N but not for either of the blue stimuli would indicate that color information has privileged access to the response-selection process, since color provides only partial information about the appropriate response in that trial. It would be almost as if the system was making an educated guess about the likely response, since knowing the color increases the probability that a response is required (since only two of the four stimuli are red), but it does not provide all of the information that is necessary for a response to be made or withheld.

There are a number of studies in which the LRP has been measured and it has been demonstrated that certain types of partial information influence response selection, while others do not. Findings such as this are important for our understanding of cognitive architecture – they tell us something about the way in which the system works – and they have important practical implications as well, since there are circumstances in which it would be undesirable to provide partial information that might influence response selection. Imagine, for example, that the use of partial information will lead to an increased likelihood of a response error (in the above example this would be a response to a red N). In safety-critical environments such as a cockpit, pilots have to distinguish quickly between multiple competing stimuli, some of which are more important than others. It is desirable to use stimuli that minimize errors in this environment, and of course optimal stimulus design will be

guided by fundamental psychological knowledge of the way in which human response selection works.

CONCLUSION

Mental chronometry is the label that has been given to the use of reaction times to study the dynamics and structure of cognitive architecture. ERPs index (with millisecond resolution) some of the cognitive processes that are involved in the gap between the presentation of a stimulus and a motor response. Because of this characteristic they are well suited to providing information about cognitive architecture which complements, and in some cases might extend, that which can be obtained by using reaction times. (*See* **Electroencephalography (EEG)**; **Auditory Event-related Potentials**; **Reaction Time**)

Further Reading

Coles MGH, Smid HGOM, Scheffers MK and Otten LJ (1995) Mental chronometry and the study of human information processing. In: Rugg MD and Coles MGH (eds) *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*, pp. 86–131. Oxford, UK: Oxford University Press.

Donchin E (1979) Event-related brain potentials: a tool in the study of human information processing. In: Begleiter H (ed.) *Evoked Potentials and Behaviour*, pp. 13–75. New York, NY: Plenum.

Donders FC (1969) On the speed of mental processes. *Acta Psychologica* **30**: 412–431.

Eimer M (1998) The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, Instruments and Computers* **30**: 146–156.

Libet B, Wright EW and Gleason CA (1982) Readiness potentials preceding unrestricted spontaneous and preplanned voluntary acts. *Electroencephalography and Clinical Neurophysiology* **54**: 322–325.

Low KA and Miller JO (1999) The usefulness of partial information: effects of Go probability in the choice/nogo task. *Psychophysiology* **36**: 288–297.

Osman A, Bashore TR, Coles MGH, Donchin E and Meyer DE (1992) On the transmission of partial information: inferences from movement-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 217–232.

Posner MI (1978) *Chronometric Explorations of Mind*. Hillsdale, NJ: Lawrence Erlbaum.

Sternberg S (1969) The discovery of processing stages: extensions of Donders' method. *Acta Psychologica* **30**: 276–315.

Executive Function, Models of

Introductory article

Machiko Ohbayashi, University of Tokyo School of Medicine, Tokyo, Japan
Seiki Konishi, University of Tokyo School of Medicine, Tokyo, Japan
Yasushi Miyashita, University of Tokyo School of Medicine, Tokyo, Japan

CONTENTS

Executive control is involved in the operation of processes supporting various prefrontal functions such as set shifting, working memory and long-term memory.

INTRODUCTION

The frontal cortex comprises one-third of the human brain; it is the structure that enables us to engage in higher cognitive functions such as planning and problem-solving. What are the processes that serve as building blocks for these higher cognitive functions, and how are these implemented in the frontal cortex?

Patients with prefrontal damage show a characteristic deficit in the ability to construct mental plans of action ('schemas') and to execute them. This deficit can be objectified by several neuropsychological tests. It is most apparent in tests that require internal programming of new behavior. The ability to plan, which depends on executive function, is severely curtailed in many patients with prefrontal cortical damage.

The Tower of London Test is considered to be a test of the ability to plan and is used on patients with prefrontal cortical damage (Figure 1). The test material consists of a board with three vertical sticks fixed on it and three wooden rings of different colors made to slide up and down on them. The sticks are of different lengths, so that the first can accommodate the three rings on top of one another, the second two rings, and the third only one. From an initial position of the rings on the sticks (for example, red over green on the long stick, blue on the middle stick), the person completing the test is instructed to move one ring at a time from stick to stick in a prescribed number of moves, and achieve a certain order (for example, green over blue over red on the long stick in five moves). The test requires planning a series of subgoals in order to attain the ultimate goal; the participant must anticipate and visualize not only that goal but also the steps leading to it in the proper sequence. Patients with prefrontal cortical damage were found to be severely impaired in performing this test, suggesting that the prefrontal cortex exerts executive function.

ANATOMY OF THE PREFRONTAL CORTEX

The prefrontal cortex is located on the anterior pole of the brain (Figure 2). Its increase in size with phylogenetic development can be inferred from the study of brains of existing animals as well as from paleoneurological data. It is most apparent in the primate order based on cytoarchitectonic calculations, forming 29% of the total cortex in humans, 17% in chimpanzees, 11.5% in gibbons and macaques, and 8.5% in lemurs. For dogs and cats, the figures are 7% and 3.5% respectively. Thus, the greater magnitude of the human prefrontal cortex in relation to other cortical regions has been considered to indicate that this cortex is the substrate for neural activity of higher cognitive functions, which, as a result of phylogenetic differentiation, has become a distinctive part of the evolutionary patrimony of our species.

With regard to studying the neural bases of cognition, a key question is the delineation of the cortical areas in which changes in neuronal activity occur during specific aspects of cognitive processing. The aim of delineating the cortical areas is to achieve an understanding of the cortical and subcortical networks that underlie different cognitive processes, and to specify the contribution of each cerebral area to the functional network in which it is involved.

To divide the prefrontal cortex into areas, the prefrontal cortex of the human and the monkey

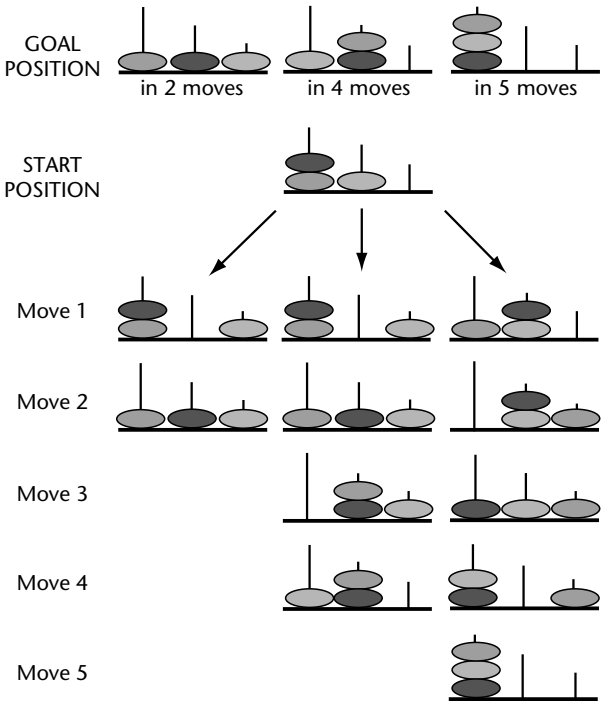


Figure 1. [*Figure is also reproduced in color section.*] The Tower of London Test. Modified from Fuster (1997).

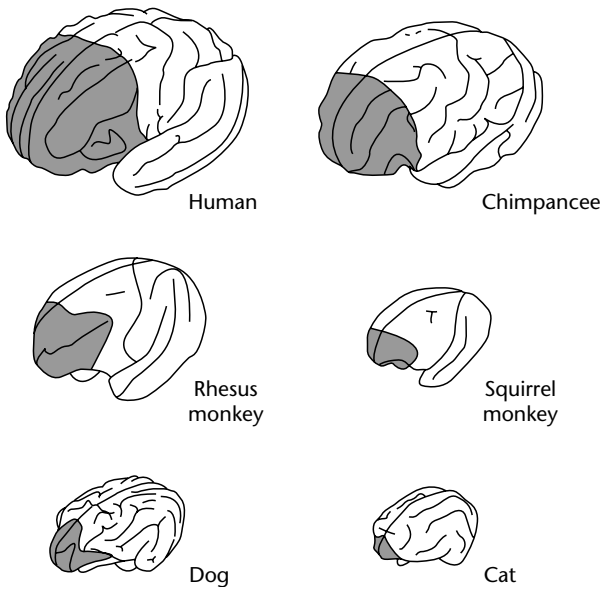


Figure 2. The prefrontal cortex (shaded) in six different species. Modified from Fuster (1997).

has been studied anatomically. The brain was sectioned in the coronal plane and stained for cytoarchitectonic and myeloarchitectonic analysis. One of the well-known cytoarchitectonic maps of the prefrontal cortex is provided in Figure 3. The prefrontal cortex is divided anatomically into three major regions: dorsolateral, orbital and medial. In the case of the human prefrontal cortex, these regions are outlined as follows. The dorsolateral cortex is the lateral aspect of the prefrontal cortex and comprises areas 8, 9, 10, and 46. The orbitofrontal cortex is the ventral aspect of the prefrontal cortex and it mainly comprises areas 11 and 13. The medial prefrontal cortex comprises parts of areas 8 through 10, and areas 12, 24 and 32. Damage in each of these three regions tends to be associated with a characteristic syndrome or cluster of symptoms.

NEUROPSYCHOLOGY

Studies of lesions of the prefrontal cortex caused by disease or trauma provide helpful insights into the functions of this structure, even though the lesions occur unpredictably and without experimental logic. The neuropsychological effects of prefrontal cortical damage vary greatly depending on the location and the extent of that damage.

Damage to the major prefrontal region generally results in one of three different syndromes. The dorsolateral damage syndrome is characterized mainly by difficulties in focusing and sustaining attention, lack of initiative and decision-making, inability to form plans and execute them, poor working memory and low verbal fluency. All these disorders are more common and conspicuous if the left dorsolateral cortex is involved than if the lesion is on the right hemisphere. A common result of dorsolateral damage is the deficit in working memory, i.e. the memory of information used for prospective motor or mental action. The orbital damage syndrome is mainly marked by impulsivity, hypermotility, distractibility, instinctual disinhibition, irritability, euphoria, perseveration and lack of moral restraint. The medial/anterior cingulate syndrome is mainly distinguished by a lack of initiative, hypokinesia or akinesia, apathy and mutism.

A prefrontal cortex deficit can be detected using a variety of tests, including the Wisconsin Card Sorting Test (WCST). This test requires the categorization of sensory (visual) items according to a temporally changing principle (Figure 4). Patients with dorsolateral lesions in the prefrontal cortex are unable to perform the task flexibly. At the start of the test, the participant is shown four 'target' cards on a table, each with a different printed design: one red triangle, two green stars, three yellow crosses and four blue circles. The participant is given a deck of cards differing in the number, color and shape of the items they depict, and instructed to sort the cards and place them, one

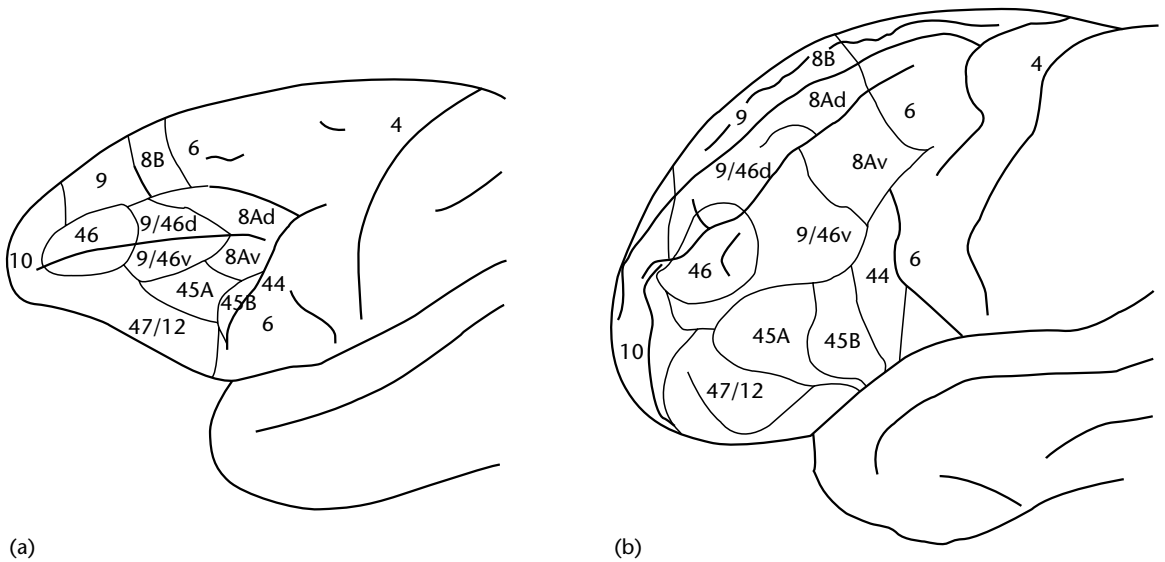


Figure 3. Cytoarchitectonic map of the lateral surface of the frontal lobe of human (a) and macaque monkey (b). d, dorsal; v, ventral. Modified from Petrides and Pandya (1994).

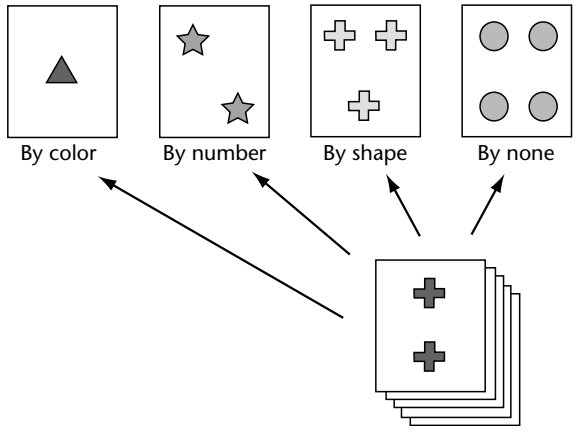


Figure 4. [*Figure is also reproduced in color section.*] The Wisconsin Card Sorting Test. Modified from Milner (1964).

at a time, under the target cards. After each card placement, the tester simply tells the participant whether the choice is correct or not (according to a tacit matching principle – number, color, or shape – determined by the tester). After a succession of correct choices, the tester, again tacitly, changes the matching principle, and the participant must guess that new principle and make subsequent choices according to it. After a succession of correct choices the principle changes again, and so on.

In order to perform the WCST correctly, the person taking the test must not only adapt to a new principle but also reject an old one. Thus, the WCST tests not only short-term memory but also the ability to withstand interference from inopportune memories. It also tests the ability to plan actions and to carry them out. All these functions can be impaired in the prefrontal syndrome, and it is probably for this reason that performance on the WCST may be poor as a result of dorsolateral lesions.

The prefrontal syndrome varies presumably depending on the region most affected. In particular, the dorsolateral syndrome can be summarized as the failure of executive functioning.

NEUROIMAGING

Computerized scanning and tomographic methods allow the visualization of changes in regional blood flow and metabolism related to neuronal activity. Thus, neuroimaging provides indirect records of the global neuron discharge in various regions of the brain simultaneously: that is, a functional map of the brain.

Recent imaging studies of the prefrontal cortex have focused on working memory, a system used for temporary storage and manipulation of information. The system is divided into two general components: short-term storage and a set of 'executive processes'. Short-term storage involves active maintenance of a limited amount of information for a matter of seconds; it is a necessary component of many higher cognitive functions and is mediated in part by the prefrontal cortex. Executive processes

are implemented by the prefrontal cortex as well. Although executive processes often operate on the contents of short-term storage, the two components of working memory can be dissociated: there are neurological patients who have intact short-term storage but have defective executive processes, and vice versa.

Storage Processes and the Prefrontal Cortex

Many neuroimaging studies are founded on Baddeley's model of working memory, which posits separate storage buffers for verbal and visuospatial information. Baddeley further argued that verbal storage could be subclassified into a phonological buffer for short-term maintenance of phonological information and a subvocal rehearsal process that refreshes the contents of the buffer. Here we examine evidence on each aspect of this model with respect to the prefrontal cortex.

Some evidence about storage mechanisms was obtained by positron emission tomography and functional magnetic resonance imaging studies using 'two-back' and 'three-back' tasks. In the two-back task, participants viewed a sequence of single letters presented at intervals of a few seconds; for each letter they had to decide whether it was identical to the letter that appeared two items back in the sequence. The experiment used two different controls. In one, participants were presented with a sequence of letters but only had to decide whether each letter matched a single target letter. Subtracting this control from the two-back condition identified many areas of activation known to be involved in item recognition tasks, including the left frontal speech regions and the parietal area. The second control required participants to rehearse each letter silently. Subtracting this rehearsal control from the two-back task should have removed much of the rehearsal circuitry since rehearsal is needed in both tasks; indeed, in this subtraction, neither Broca's area nor the premotor area remained active. Hence, this experiment isolated a frontal rehearsal circuit. Frontal regions that no doubt evolved for the purpose of spoken language appear to be recruited to keep verbal information active in working memory.

Research on nonverbal working memory has been influenced by physiological work with nonhuman primates. Single-cell recordings performed while monkeys engaged in spatial-storage tasks have identified 'spatial memory' cells in the dorsolateral prefrontal cortex (which is usually considered to include areas 46 and 9). These cells selectively fire during a delay period and are position-specific. Recordings performed while monkeys engaged in object-storage tasks have identified delay-sensitive 'object memory' cells in the prefrontal cortex.

Executive Processes and the Prefrontal Cortex

Most researchers concur that executive processes are mediated by the prefrontal cortex and are involved in the regulation of processes operating on the contents of working memory. Although there is a lack of consensus about the taxonomy of executive processes, there is some agreement that they include:

- focusing attention on relevant information and processes, and inhibiting irrelevant ones (attention and inhibition);
- scheduling processes in complex tasks, which require the switching of focused attention between tasks (task management);
- planning a sequence of subtasks to accomplish some goal (planning);
- updating and checking the contents of working memory to determine the next step in a sequential task (monitoring);
- coding representations in working memory for time and place of appearance (coding).

Performance of tasks manifesting each of these processes is known to be selectively impaired in patients with prefrontal damage. Of these five executive processes noted, the first two appear to be the most elementary and the most interrelated; for these reasons, we focus on task management.

A canonical case of task management arises when patients are required to perform a dual task. For example, they may be presented with a series of numbers and have to add 3 to the first number and subtract 3 from the second, and so on through successive trials. Both tasks require some nonautomatic or 'controlled' processes, and a critical aspect of task management is switching from one controlled process to another.

A magnetic resonance imaging study has examined dual-task performance. In one task, participants had to decide whether each word presented in a series named an instance of the category 'vegetable'; in the other task, participants had to decide whether two visual displays differed only by a matter of rotation; in the dual-task condition, the patients had to perform the categorization and rotation tasks concurrently. Only the dual-task condition activated prefrontal areas, including

dorsolateral prefrontal cortex (BA 46) and the anterior cingulate.

Neuroimaging studies of humans show that storage and executive processes are the major functions of the prefrontal cortex. The distribution between short-term storage and executive processes appears to be a major organizational principle of the prefrontal cortex. Neuroimaging analysis of executive processes is quite recent, and must lead to clear dissociations between processes.

STUDY OF PREFRONTAL CORTEX IN ANIMALS

The role of the prefrontal cortex has also been studied in nonhuman primates in lesional and physiological studies.

Lesions in the prefrontal cortex elicit characteristic behavioral abnormalities. The dorsal and lateral prefrontal surfaces of the cortex are primarily involved in cognitive aspects of behavior; the rest of the prefrontal cortex, medial and ventral, appears to be mostly involved in affective and motivational functions, as well as in the inhibitory control of external and internal influences that interfere with purposive behavior. The ability of monkeys with lesions in the prefrontal cortex to perform an analogue of human neuropsychological tests such as the WCST is also impaired.

Neurophysiological studies of monkeys that have explored the neural basis of cognitive control indicate that the prefrontal cortex has a role in operating nodal points, where neural circuits integrate currently available or memorized information to generate the information necessary to perform an action. Subsequent reports provide insight as to how the processes of information retrieval and integration are actually carried out in the prefrontal cortex. The prefrontal cortex performs the following functions: retrieval of sensory information to meet behavioral demands; executive control of memory retrieval from sites of long-term storage; active maintenance of either sensory or memory information; and integration or manipulation of the retrieved or stored information for subsequent use.

Sensory Information Retrieval

Prefrontal neurons reflect learned associative relationships between goal-relevant elements – they show conjunctive tuning for learned associations between cues, voluntary actions and rewards. Prefrontal neurons even show tuning for complex, behavior-guiding rules. Thus they may help form neuron ensembles that represent the regularities across experiences that describe the principles needed to achieve a particular goal in a particular situation.

Many lateral prefrontal cortex neurons reflect these learned associations. For example, most lateral prefrontal cortex neurons were found to reflect the association between a cue and a reward. A given neuron might be activated by a cue, but only when it signalled 'reward'. In contrast, another neuron might be activated only by a cue that signalled 'no reward'. Similarly, lateral prefrontal cortex neurons can reflect learned associations between a cue and behavior. Monkeys were trained to associate, in different blocks of trials, each of two cue objects with a saccade to the right or left. The activity of lateral prefrontal cortex neurons reflected associations between objects and direction of saccades instructed. Other neurons exhibited activity that reflected the cues or the saccades alone, but these were fewer in number. The prefrontal cortex neurons can reflect learned associations between visual and auditory stimuli.

Storage Processes

Active maintenance of information for subsequent use is an important aspect of the prefrontal cortex function, and its operation has been studied extensively using a variety of working memory tasks. One of the classic signs of the prefrontal cortical damage is increased distractibility: people with such damage seem unable to focus on a task when other, irrelevant events compete for their attention. This may reflect the loss of mechanisms for maintaining goal-relevant information, a process known as 'working memory'. This has been explored in a variety of neurophysiological studies of monkeys.

Many cortical areas seem to have some sort of short-term buffering ability. What sets working memory apart as being more 'cognitive' is that it can retain information over potentially distracting events. The prefrontal cortex neurons particularly have this ability. For example, when monkeys are required to sustain the memory of a sample object for a delay period filled with visual distracters each requiring attention and processing, sustained activity within the prefrontal cortex acts to maintain the sample memory. In contrast, sustained activity in extrastriate visual areas seems to be more easily disrupted by the presence of distracters – following presentation of a distracter, the neural activity in the inferior temporal cortex and posterior parietal cortex no longer reflected the sample object that the monkey has retained in memory.

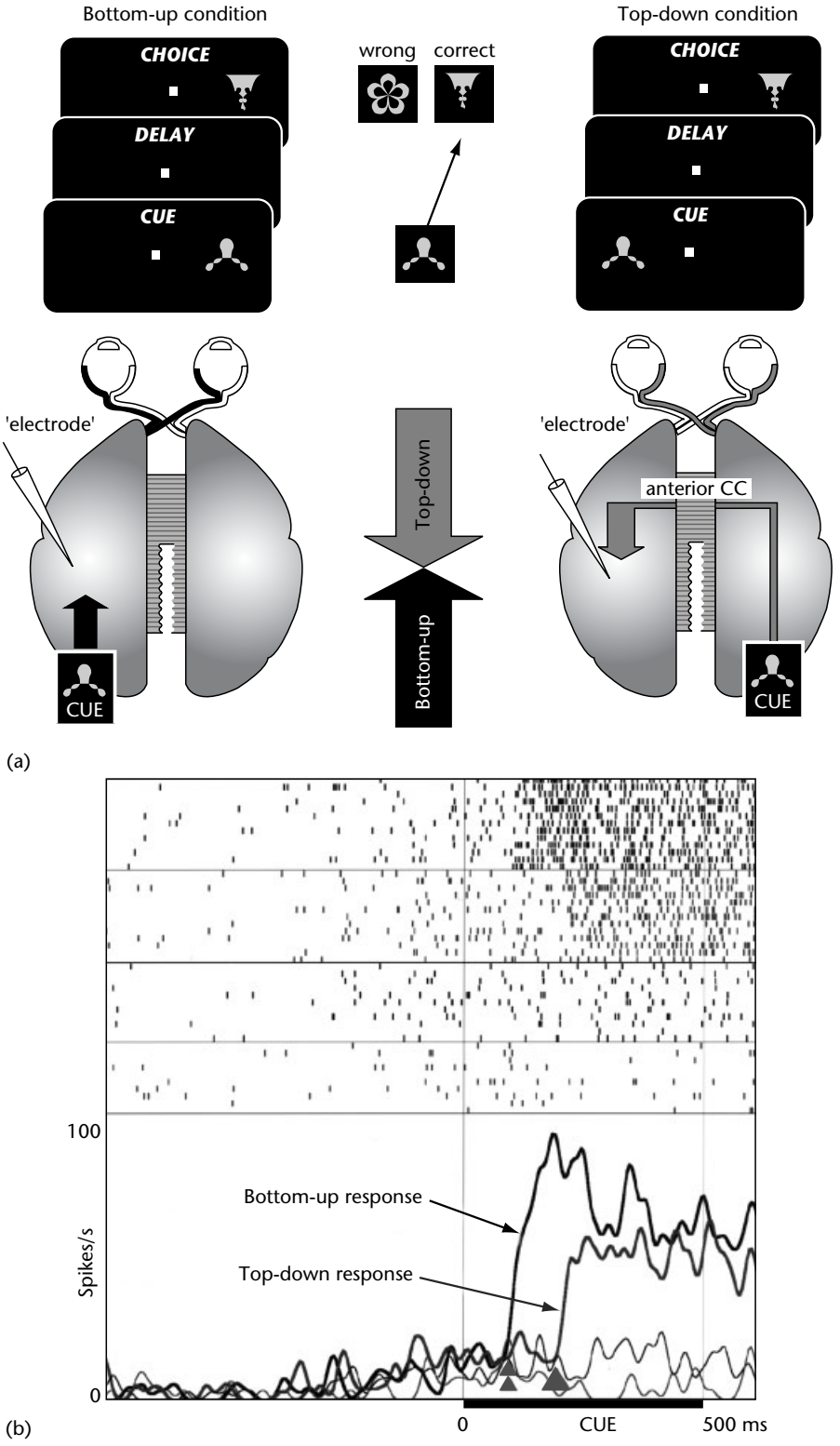


Figure 5. Top-down signal detection. (a) Experimental design; (b) neuronal activity in top-down condition (single inferior temporal cell). Modified from Tomita *et al.* (1999).

Top-down Control

The ability to sustain task information is of little use unless the prefrontal cortex can somehow use it to control processing in other brain systems. The prefrontal cortex activity could exert a 'top down' influence by providing an excitatory signal that biases processing in other brain systems towards task-relevant information. To understand how this may work, let us consider selective visual attention. In the visual system, neurons processing different aspects of the visual scene compete with each other for activation. This is thought to be important for enhancing contrast and separating objects from the background. The neurons that 'win' the competition and remain active are those that incur a higher level of activity. The biased competition model proposes that visual attention exploits this circuitry. In voluntary shifts of attention, a competitive advantage is conferred by excitatory signals that represent the 'to be attended' stimulus. These excitatory signals enhance the activity of neurons in the visual cortex that process that stimulus and, by virtue of mutual inhibition, suppress activity of neurons processing other stimuli. This concept of excitatory bias signals that resolve local competition can be extended from visual attention to cognitive control in general.

Several studies have indicated that the prefrontal cortex exerts a top-down influence over other neocortical regions. Deactivation of the lateral prefrontal cortex attenuates the activity of extrastriate neurons in response to a behaviorally relevant cue. Top-down signals originating from the prefrontal cortex have been shown to be required to activate (recall) a long-term memory stored in the inferior temporal cortex. In the absence of 'bottom up' visual inputs, single inferior temporal neurons were activated by the top-down signal, which conveyed information on semantic categorization imposed by visual stimulus–stimulus association (Figure 5). Performance on a task was severely impaired with the loss of the top-down signal. Thus, feedback projections from the prefrontal cortex to the posterior association cortex appear to have executive control of voluntary recall.

Other indicative evidence was obtained from investigations into the respective roles of the prefrontal cortex and inferior temporal cortex in working memory. Monkeys were trained to hold a sample object 'in mind' while they viewed a sequence of objects. They were required to respond when the sample was repeated and to ignore other irrelevant object repetitions. As noted above, sustained activity in the prefrontal but not in the inferior temporal cortex maintained the sample memory across intervening stimuli. However, many inferior temporal neurons showed an enhancement of their neuronal responses to the sample repetition but not to irrelevant repetitions. This indicated that sustained activity in the prefrontal cortex with respect to the sample might have enhanced responses to its repetition in the inferior temporal cortex.

Because task representations in the prefrontal cortex include disparate information, the excitatory signals from this area could be involved in selecting particular sensory inputs (attention), memories (recall) or motor outputs (response selection).

These neurophysiological studies of nonhuman primates show that the prefrontal cortex may be essential for implementing task information, particularly in situations when familiar sets of behaviors need to be flexibly combined into a coherent sequence. In addition, the prefrontal cortex is required to activate long-term visual memories stored in the temporal lobe. The prefrontal cortex could retain links to stored representations that allow it to bring visual memories and other task knowledge 'online' when required.

SUMMARY

One of the greatest mysteries in cognitive neuroscience is cognitive control and the neuronal mechanism mediating it. The prefrontal cortex is involved in many higher cognitive functions, most typically executive control. Lesions to the prefrontal cortex of humans and nonhuman primates cause the failure of executive functioning. Neuroimaging studies of humans show that storage and executive processes are the major functions of the prefrontal cortex. Neurophysiological studies in monkeys indicate that the prefrontal cortex has a role in the extraction of behavioral significance from sensory cues and the storage and retrieval of mnemonic information.

Further Reading

Fuster JM (1997) *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. New York, NY: Raven Press.

Konishi S, Nakajima K, Uchida I *et al*. (1998) Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience* **1**: 80–84.

Miller EK (2000) The prefrontal cortex and cognitive control. *Nature Reviews* **1**: 59–65.

Milner B (1964) Some effects of frontal lobectomy in man. In: Warren JM and Akert K (eds) *The Frontal Granular Cortex and Behavior*, pp. 313–334. New York, NY: McGraw-Hill.

Petrides M and Pandya DN (1994) Comparative architectonic analysis of the human and the macaque frontal cortex. *Handbook of Neuropsychology* **9**: 17–58.

Smith E and Jonides J (1999) Storage and executive processes in the frontal lobes. *Science* **283**: 1657–1661.

Tanji J and Hoshi E (2000) Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology* **11**: 164–170.

Tomita H, Ohbayashi M, Nakahara K, Hasegawa I and Miyashita Y (1999) Top-down signal originating from the prefrontal cortex for memory retrieval. *Nature* **401**: 699–703.

Expertise

Introductory article

Andreas C Lehmann, School of Music, Würzburg, Germany
K Anders Ericsson, Florida State University, Tallahassee, Florida, USA

CONTENTS

Expertise refers to the cognitive, perceptual-motor, and physiological mechanisms that allow experts to attain consistently superior levels of performance on representative activities in their domains.

EXPERTISE AND EXPERT PERFORMANCE

In everyday terminology, experts are individuals, such as medical doctors, accountants, teachers and scientists, who have first been certified as professionals after extended training and then have gained substantial experience in their specialty. More recently, the term has been expanded to describe any highly skilled performer who consistently exhibits superior achievement after instruction and extended experience in a domain such as one of the arts (e.g. music, painting, writing), a sport (e.g. swimming, soccer, golf) or a game (e.g. chess, Othello, bridge). The general concept of expertise is based on the hypothesis that experts in the different domains will acquire their superior performance according to similar learning principles and that the general structure of the acquired mechanisms will be similar in domains with similar demands on performance.

The main task for researchers of expertise is to explain how some individuals attain the highest levels of achievement in a domain, and why so few reach such levels. Expert performance in many domains looks effortless, and one is thus led to believe that experts excel in general basic characteristics, such as intelligence, memory, speed, and flexibility. It has been assumed traditionally that such characteristics are impossible to train and thus are determined to a large degree by genetic factors (nature). Everyone agrees, however, that experts must acquire at least some necessary domain-specific knowledge and skill (nurture). Without denying the possibility of effects of genetic factors, research in expertise tends to emphasize the importance of knowledge and acquired skills.

Sir Francis Galton, a pioneer in the study of excellence in nineteenth-century England, claimed that instruction and training were beneficial, even necessary, and associated with large initial improvements in performance. However, he also maintained that the upper bound of individuals' performance was fixed by hereditary factors. This view is still common among many researchers and practitioners in some domains. Thus, proponents of this view are still searching for innate talent and measuring basic capacities of memory, perception, and thinking, that would allow them to successfully predict future outstanding performance in children and untrained adults. Today, over a century later, we can safely say that efforts to measure individual differences in basic capacities have not succeeded in predicting which individuals would eventually attain expert performance. For example, when athletes or other experts are tested in the laboratory on how fast they can respond to the onset of a light (simple reaction time), they are not systematically faster than other people. The amazing speed of tennis players' actions is thus not due to superior speed of neural impulses, and we have to assume that returning a fast tennis serve reflects an acquired ability to respond rapidly in specific types of situations.

Domain-specific superiority of experts is very noticeable for cognitive abilities. Chess masters can recall nearly all the 24 chess pieces in a typical chess arrangement after a quick glance, whereas beginners in chess can recall only around four pieces. The experts' ability to recall more chess pieces is, however, restricted to representative (or typical) chess positions in which the expert can meaningfully encode the relations between the chess pieces. When random arrangements of chess pieces are presented, the large recall advantage of the experts over the novices is virtually eliminated.

The major differences between experts and less proficient individuals nearly always reflect specific adaptations acquired by the experts during their lengthy training. This holds true for many anatomical and physiological characteristics of athletes, such as the size of their muscles and bones and the flexibility of joints, and for the increased range of mobility of the limbs in ballet dancers and musicians. Some of these attributes, such as structural changes in the brains of musicians, are even correlated with the early onset of training. Other adaptations – for example, the growth of muscles and thickening of the bones of tennis players (restricted to the arm holding the racket), or the optimization of metabolism of runners (for particular running speeds and different lengths of races) – are so specific that (self)-selection of individuals in those domains appears unlikely. Finally, many physiological adaptations, such as the heart size of endurance runners, have been shown to revert to the normal values once the athletes stop training, which would be expected only for acquired and not for innate characteristics. However, there is at least one exceptional attribute – height – where genetic factors are known to control the development of body size. Above-average height constitutes an advantage in attaining the highest levels in some sports such as basketball, but is a disadvantage in other sports such as gymnastics.

The occurrence of expert achievement in similar domains across several generations in some famous families is frequently cited as proof for the genetic transmission (high heritability) of special talents. However, this can be questioned since the early instruction of children by parents and access to networks and specialized training seem to offer plausible alternative accounts. So far, studies on experts have been unable to document any convincing evidence for significant heritability of expert levels of achievement.

EXPERTISE AS ACQUIRED KNOWLEDGE AND SKILL

Organized Knowledge as a Base for Expertise

Our understanding of expertise was advanced primarily by comparing the thought processes of experts while exhibiting their superior performance with the thoughts of less accomplished individuals (or even novices) performing the same tasks. In his 1940s pioneering study, researcher Adrian de Groot wanted to explain why world-class players generated consistently better chess moves than less skilled players after only a brief exposure to the chess position. He instructed chess players in each group to think aloud while they selected their next moves for a given chess position. The transcriptions of these reports (think-aloud protocols) revealed that in an initial phase all chess players were quickly generating promising moves while examining the organization and structure of the presented chess position. Next, the chess players spent minutes evaluating these potential moves by searching and planning and discovered often even better moves. Neither de Groot nor other researchers found that world-class players differed from less skilled chess players with regard to basic intelligence or to speed of their planning. The rapid speed of move generation indicated that the superior moves were directly retrieved from knowledge of similar chess positions in memory.

According to Simon and Chase's theory of expertise, the essential factor is the accumulation of increasingly complex patterns that allow experts to maintain access to their vast body of knowledge and experience. The same mechanisms were proposed to explain the efficient problem-solving and decision-making of experts in physics, medicine, and business. Hence, the superior ability of experts to reason was found to be specific to domain-related material and thus depended directly on their large body of knowledge and experience. Viewing expertise in terms of superior knowledge led investigators in artificial intelligence to interview experts in order to extract the necessary relevant knowledge and decision rules for building computer programs (expert systems) that were designed to reproduce the experts' behavior.

Challenges to the Knowledge-based View of Expertise

Paradoxically, individuals recognized as domain experts because of their lengthy training and experience do not always exhibit consistently superior performance compared with less experienced individuals and sometimes even complete amateurs. For example, professional stockbrokers do not consistently select better investments than statistical models and amateurs, or even random picks. Impressive dissociations between the level of expertise (indicated by the amount of schooling and experience) and performance have been demonstrated in medical diagnosis of common diseases, treatment with psychotherapy, auditing, and many types of expert decision-making and forecasting.

Why does extended experience not always lead to improvements in performance? It is well known

that when persons are introduced to an activity, their gains in performance are initially large, but once an acceptable level of performance is reached, the behavior of individuals tends to become stable and automated with no further improvements of performance. For example, the vast majority of amateur musicians and recreational athletes in golf and tennis remain at a stable level of performance for years and decades despite their regular involvement in relevant activities. Therefore, it is necessary to distinguish individuals who are customarily labeled 'experts' because of their extended experience, training and reputation, from true 'expert performers' who exhibit consistently superior performance.

Superior Achievement as a Prerequisite for Expert Performance

If we restrict expertise to reproducible expert performance, then the first step is to design tasks that can reliably measure the associated superior performance under standardized conditions. In sports, athletes compete under fair conditions, when they run the 100 m dash or swim 200 m freestyle. Competitions in music and ice skating rely on panels of judges to rate performance. In chess and tennis, the best performers are identified by tournaments where winners advance. In most domains procedures have evolved to evaluate and measure performance in order to identify individuals with superior performance.

The Necessity for Experience, Instruction, and Practice

The development of expert performance requires extensive experience, and it progresses gradually without sudden increases in performance level over extended periods. This is true for the development of adult experts and even for child prodigies. Most expert performers continue to improve their performance long into adulthood, and attain their peak performances in vigorous sports typically at age 25–35 years, and much later for achievements in the arts and sciences. Simon and Chase showed that it takes at least 10 years of training and experience in chess even for the most talented individuals to reach an international level after the start of their engagement in the domain. This 10-year rule has since been shown to apply to most other domains, even in swimming and music where children start practice as early as age 4 or 5 years. Extended experience is thus necessary but not sufficient to attain high levels of performance. If mere experience does not automatically change individuals' behavior, what does?

When our actions and activities run smoothly, a change in the structure of performance is generally not necessary and hence would not be expected. Individuals tend to think that their performance in a given area is 'good enough', and mistakes occur only for very rare or difficult behavior or decisions. For example, a tennis player might have problems with a particular backhand volley, but the need for this stroke during recreational play might occur rarely and unpredictably. In these typical circumstances more experience would not necessarily be associated with improvement of the stroke. However, the problem could be remedied with some precise suggestions by a coach and subsequent practice. The coach would perhaps start with preparatory exercises and work up to difficult and unpredictable exchanges, provide enough challenge and repetitions, and eventually incorporate the strokes in regular training matches.

This optimal type of training activity has been called 'deliberate practice', and its sole purpose is to improve performance effectively. Note that the particular activities that constitute deliberate practice will vary from domain to domain and differ with the level of attained skill. Recreational athletes and amateurs rarely engage in deliberate practice, although they realize that deliberate practice would improve their performance. Since engaging in a domain activity is usually motivated by inherent enjoyment (play) or external rewards (work), most individuals shy away from activities that are more effortful or less enjoyable than the regular activity. Consequently, regular recreational activities and work lack the prerequisites of deliberate practice that are essential for efficient improvement, namely specific goals for training, feedback, and opportunities for gradual improvement through repetition. Engagement in deliberate practice appears to be imperative for individuals to develop the mechanisms for reaching the highest levels of performance.

Over time, experts in domains of expertise have extracted and accumulated a body of organized experience in the form of knowledge and skills that can be effectively transmitted by teachers to students with developed practice techniques. For example, in the Middle Ages it was believed that 50 years were required to master the known forms of mathematics, whereas today students acquire comparable knowledge in highly organized form during their high-school years. Similarly, improved training allows serious amateurs in sports events such as running or swimming to reach levels of performance attained only by elite performers

earlier in the twentieth century. Instead of having to rediscover items of knowledge and ways of practicing or learning, individuals can be given instruction to master and even surpass the established levels of attainment.

The need for specialized instruction and deliberate practice has been recognized. In many domains master teachers and coaches instruct and supervise promising individuals and design their practice from a very young age. The core assumption of deliberate practice is that expert performance is acquired gradually and that effective improvement of the student's performance depends on the teacher's ability to isolate a sequence of simple training tasks that the student can then successively master by repetition with feedback and instruction. The individual training tasks have to be difficult enough to lie slightly outside the student's current range of skills so that the student concentrates on critical aspects and gradually refines performance through repetition in response to feedback. This requirement of focused attention on individual task components differentiates deliberate practice from both mindless drill and playful engagement.

Successful expert performers in different domains share some biographical characteristics, and Bloom and his colleagues have developed a schematic model of the phases of skill acquisition (Figure 1). Often, future experts start training with teachers at young ages after a brief period of playful interaction with the domain (phase I). Their initial formal training (phase II) consists of short training units of about 15–20 min per day, and often a parent supervises practice and helps the child to concentrate. With increasing age, domain-related activities, especially deliberate practice, occupy more and more room in the daily lives of future expert performers, until at the end of adolescence their commitment to the domain is essentially full-time (phase III). During this phase, increases in performance go hand in hand with the acquisition of the cognitive mechanisms that mediate the superior performance (see below). Bloom found that to reach an international level of achievement, the performers – almost without exception – had been trained by excellent teachers who either were international level experts themselves or had successfully trained students to reach international levels.

At some point the experts start their professional career (phase IV). During this last phase the elite performers strive to make eminent contributions to the domain, which could constitute the setting of new world records in sports, publishing scientific discoveries, or making major creative contributions to the arts.

By using diaries and other methods to study how expert musicians spent their daily lives throughout the different phases, Ericsson and his colleagues demonstrated the importance of deliberate practice for attaining expert performance. They investigated three groups of experts differing in their level of attained musical performance. Based on retrospective estimates of past practice times, the researchers calculated the number of hours of deliberate practice accumulated by the different groups of musicians (Figure 2), and found that the better musicians had spent more time in deliberate practice during their development. By the age of 20 years, the best piano players had spent over 10 000 h in practice compared with only around 2000 h for typical amateur pianists of the same age. A number of subsequent studies in chess, sports, and music have confirmed the relationship between the level of attained performance and amount or quality of deliberate practice.

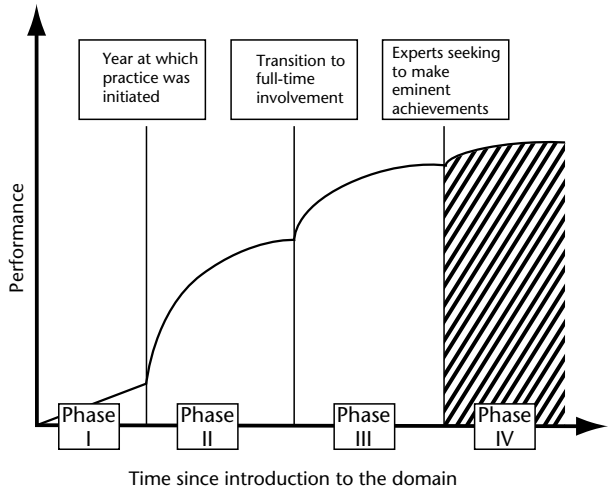


Figure 1. Bloom's three phases of acquisition of expert performance, followed by a qualitatively different fourth phase during which experts attempt to go beyond the available knowledge in the domain. From Ericsson KA, Krampe RT and Heizmann S (1993) *The Origins and Development of High Ability*, pp. 222–249. Chichester, UK: John Wiley. Copyright 1993 by CIBA Foundation. Adapted with permission.

EXPERT PERFORMANCE

The extended period of deliberate practice enables expert performers to acquire complex mental mechanisms that mediate superior performance in their domains. These same mental mechanisms allow them to become their own teachers, which in turn permits them to continue learning and

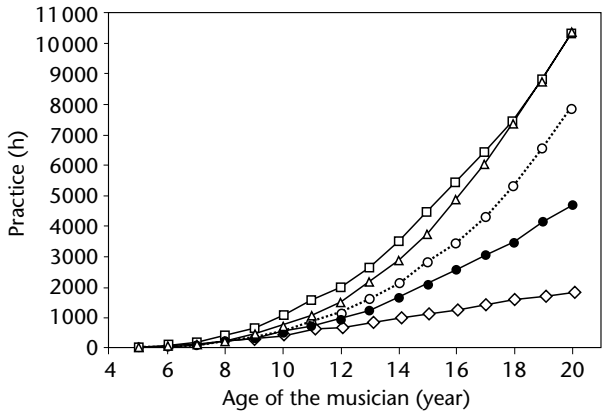


Figure 2. Amount of time spent in solitary practice as a function of age for the middle-aged professional violinists (triangles), the best expert violinists (squares), good expert violinists (open circles), the least accomplished expert violinists (solid circles) and amateur pianists (diamonds). From Ericsson KA, Krampe RT and Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**(3): 379, 384. Copyright 1993 by American Psychological Association. Adapted with permission.

improving their performance for their entire professional career.

Cognitive Mechanisms Underlying Captured Expert Performance

In most domains it is possible to identify the representative tasks and cognitive mechanisms that capture the essence of that expertise. For example, the essence of chess expertise is the consistent ability to select the best move on the chessboard. Hence, following de Groot's paradigm, we can present different chess players with the same series of unfamiliar chess positions and ask them to think aloud as they select the best move (Figure 3(a)). By analyzing the think-aloud protocols, we can scientifically investigate the cognitive processes that differentiate the experts' performances from those of less skilled players.

In chess, analyses of such think-aloud protocols reveal that chess experts rapidly encode the chess position and access adequate moves from memory. However, chess experts will then spend several minutes evaluating these potential moves by planning out the consequences for each of them. The ability to predict the consequences a particular chess move would have several moves ahead requires extensive working memory. Chess players acquire memory skills that allow them to store information about planned chess positions in long-term working memory (LTWM). The proficiency of LTWM increases slowly as a function of chess skill and can account for increases in the ability to plan, as well as the previously discussed increased memory for briefly presented chess positions. Although chess players can often select appropriate moves very early in the decision process, with additional time the quality of the selected moves will increase for all levels of chess players. Thus, at higher levels of performance, experts have acquired the ability to mentally represent relevant information. Chess masters, for instance, have perfected their mental representations of the chess positions to the point where they can play chess blindfolded.

Mental representations have been shown to play a central role in all domains where it has been possible to capture expert performance with laboratory tasks. For example, medical expertise can be captured by presenting doctors with information about challenging medical cases and asking them to reach a diagnosis (Figure 3(b)). From analyses of protocols of medical experts and students thinking aloud while diagnosing patients, it is known that the experts have acquired a superior LTWM with a better representation of the relevant clinical information. As a result, to arrive at the correct diagnosis medical experts are able to plan and reason about the many relevant symptoms.

In domains of perceptual-motor activity, acquired mental representations allow experts to anticipate upcoming events. For example, the best predictor of individuals' keyboard typing speed is not the basic speed of their finger movements but how far they look ahead. By planning ahead, expert typists can anticipate future keystrokes and thus move fingers to their positions ahead of time rather than rush their fingers. Similarly, the rapid reactions of athletes, such as hockey goalkeepers, baseball hitters, and tennis and football players, have also been found to reflect the ability to anticipate events. The resulting shorter reaction times of experts (compared with less accomplished individuals or novices) for representative tasks are thus not due to an innate speed advantage but to superior anticipation, preparation, and improved perceptual skills.

In many domains, such as music, dance and some sports, the key to expert performance is the acquisition of representations that enable the performer to produce the same movement repeatedly. For example, superior golf players can execute the same putt several times with fewer variations than less skilled golfers; similarly, expert pianists are better able to play the same musical piece several

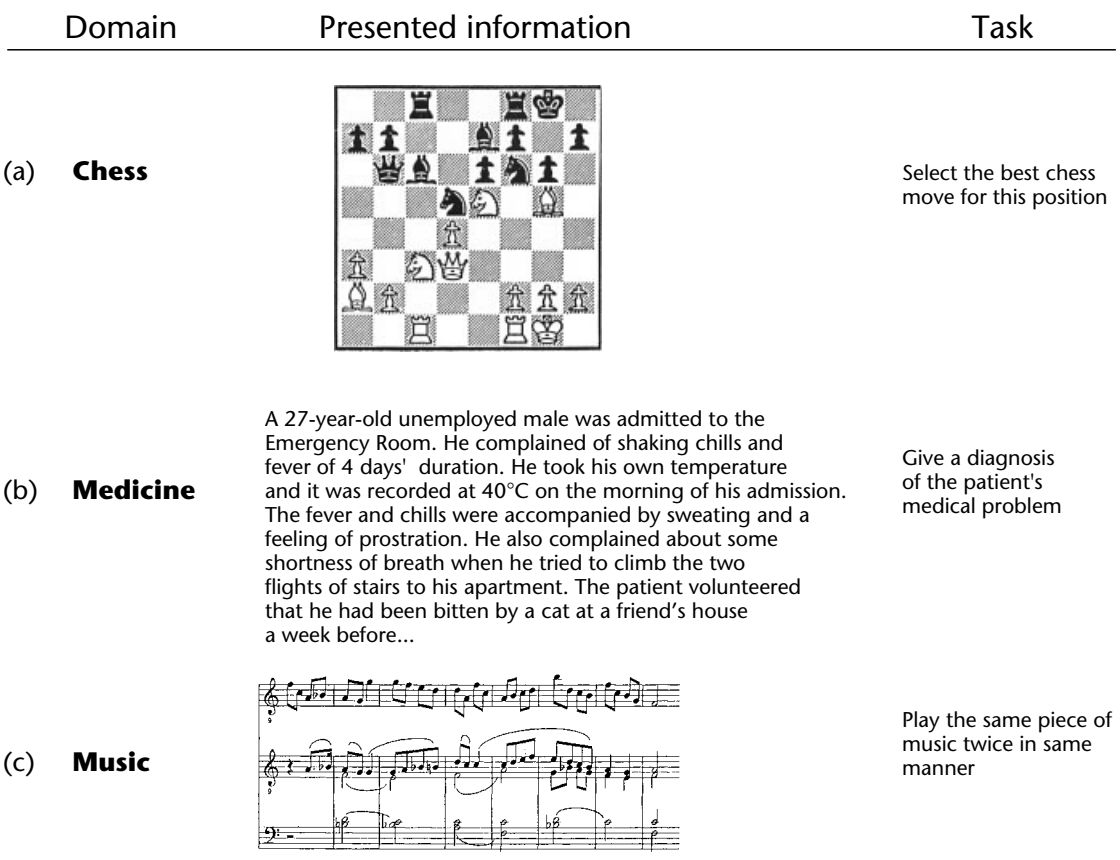


Figure 3. Three examples of laboratory tasks that capture the consistently superior performance of domain experts in chess (a), medicine (b), and music (c). The description of the medical case was taken from the first part of a case report of acute bacterial endocarditis, from 'Inferences in clinical case comprehension' by LD Coughlin and VL Patel, in *Technical Report in Cognitive Research Series*, Center of Medical Education, McGill University, Montreal, Canada, p. 21, 1987. Adapted from Ericsson KA (1998) The scientific study of expert levels of performance: general implications for optimal learning and creativity. *High Ability Studies* **9**: 92. Copyright 1998 by European Council for High Ability.

times in a consistent manner than less skilled pianists (Figure 3(c)).

In sum, the experts' performance is not characterized by reduced cognitive processing and automaticity; rather, experts acquire and refine their mental representations to increase control over their performance.

Importance of Mental Representation for Continued Learning

The experts' ability to mentally represent and manipulate relevant information endows them with the level of control that they need to respond flexibly to changing conditions in real life. Imagine a pianist who has to play a well-entrenched piece under changing acoustical conditions and on different instruments every night, or a tennis player who needs to adapt to different opponents' strengths and weaknesses. In order to maintain a high level of performance despite changes in the environment, the expert needs a flexible and generalizable skill.

Reaching expert performance in a domain does not simply consist of gradually increasing the level of performance, but rather requires the acquisition and the refinement of underlying mental representations, such as being able to image and plan a desired performance, and monitor and evaluate one's own ongoing performance. These are the ultimate goals of instruction, because they make the learner independent of teachers and coaches. In fact, the learner becomes his or her own teacher or coach. This ability to become one's own teacher is important for the development of domain-specific skills at a societal level. As mentioned earlier, the primary objective of experts in later stages of their lives is to make a personal creative contribution to the domain. Before making an innovative contribution to a respective domain, the

performer has to master all the available knowledge and skill; otherwise, he or she would not know when and how to go beyond the current knowledge of performance.

How do experts use their highly developed representations to organize their own training and further improvement in performance? A common method involves studying and analyzing performances and achievements of masters in their respective domains. For example, expert chess players collect books and magazines with published games of chess masters and spend several hours every day playing through those games move by move, trying to predict the next best move the master could have chosen. Any inconsistency between their own prediction and the chess master's actual move is carefully analyzed for oversights in planning and evaluation. In general, this form of study is theoretically interesting because attempting to copy the model behavior of established masters allows performers to gradually refine their own independent representations and expand their body of knowledge. Other domains are likely to require other types of activities that promote further improvements and innovations.

CONCLUSION

The public performance of expert musicians and dancers often appears to be effortless and natural. However, research on expert performance shows that even the most 'talented' individuals need around 10 years and thousands of hours of deliberate practice to attain an international level of performance in their respective domains. Research that has successfully captured expert performance with representative tasks in the laboratory has revealed the complexity and domain specificity of the mental representations and mechanisms that enable superior performance. Once expert performance is understood as the acquisition of an integrated structure of cognitive and physiological adaptations, then the main challenge is to account for how individual domain experts gradually develop their superior performance during decades of deliberate practice.

Further Reading

Ericsson KA (ed.) (1996) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games*. Mahwah, NJ: Lawrence Erlbaum.

Ericsson KA and Kintsch W (1995) Long-term working memory. *Psychological Review* **102**(2): 211–245.

Ericsson KA and Lehmann AC (1996) Expert and exceptional performance: evidence of maximal adaptations to task constraints. *Annual Review of Psychology* **47**: 273–305.

Ericsson KA and Smith J (eds) (1991) *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge, UK: Cambridge University Press.

Starkes JL and Allard F (eds) (1993) *Cognitive Issues in Motor Expertise*. Amsterdam, Netherlands: North Holland.

Bloom BS (ed.) (1985) *Developing Talent in Young People*. New York, NY: Ballantine Books.

Chi MTH, Glaser R and Farr MJ (eds) (1988) *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum.

Hoffman RR (ed.) (1992) *The Psychology of Expertise: Cognitive Research and Empirical AI*. New York, NY: Springer-Verlag.

Simon HA and Chase WG (1973) Skill in chess. *American Scientist* **61**: 394–403.

VanLehn K (1996) Cognitive skill acquisition. *Annual Review of Psychology* **47**: 513–539.

Eye Movements

Introductory article

Michael K Tanenhaus, University of Rochester, New York, USA

CONTENTS

The organization of the retina requires the eyes to be able to move to alter the fixation of the gaze. These eye movements are closely linked to shifts in attention and have an important role in visual search, perception of scenes, and complex visual tasks.

INTRODUCTION

During everyday tasks involving vision, such as reading a newspaper, looking for the car keys, making a cup of coffee and conversing about objects in the immediate environment, people rapidly shift their gaze to bring task-relevant regions of the visual field into the central area of the fovea, where visual acuity is greatest. These gaze shifts are accomplished by rapid jumping eye movements known as 'saccades'. The pattern and timing of saccades, and the resulting fixations, are one of the most widely used response measures in the cognitive sciences, providing important insights into the mechanisms underlying attention, visual perception, reading, memory, and spoken language processing. Within a fixation, smooth eye movements maintain the stability of the retinal image, compensating for motion on the retina caused by movements of the head, body and moving visual targets.

SACCADES

Saccades are extremely rapid ballistic eye movements. During a saccade, the eye is in motion for 30–60 ms, with the duration of the saccade related to the distance that the eye travels. At peak velocity, the eye can be moving at a rate of 500–1000° per second. During a saccade, sensitivity to visual information is dramatically reduced. Suppression of visual information occurs in part because of masking, and in part because of central inhibition. A saccade is followed by a fixation that typically lasts for 200 ms or more depending upon the task. The minimal latency for planning and executing a saccade is approximately 150 ms when there is no uncertainty about target location. In reading, in visual search, and in other tasks in which there are multiple target locations, saccade latencies are somewhat slower, typically about 200–300 ms.

The brain mechanisms involved in controlling saccades have been intensively studied in recent years. Saccadic control is subserved by both a 'what' system, generated by brainstem cells that show a drop in sustained activity before and during a saccade, and a 'where' system based on spatial coding in neighboring regions. These cells receive inputs from the superior colliculus, which in turn receives primary input from the frontal eye field and posterior parietal cortex.

MEASURING EYE MOVEMENTS

Currently popular methods for monitoring eye movements include measuring infrared corneal and retinal reflections (Purkinje images), video-based pupil monitoring, and search coils attached to a contact lens. Accurate measurement of the 'point of regard' must distinguish between rotation and translation, which is difficult for many systems. Changes introduced by rotation of the eye in its orbit due to a saccade or a smooth eye movement systematically affect the position of the retinal image. However, the position of the image varies with target distance, for the translations introduced by movement of the head. Eye movements differ depending upon whether the head is fixed or not, with more accurate gaze shifts, and faster and more accurate saccades.

Studies of reading with normal-sized text require monitoring eye position with a high degree of accuracy (20 minutes of arc or less). These studies frequently use Purkinje image tracking with the head fixed on a bite bar to minimize translational movement. Studies with action-based tasks increasingly use video-based trackers that monitor

the pupil and the cornea, with independent tracking of the head or compensation for head movement, when stimuli are presented on a screen. Measuring head movement can be bypassed by superimposing fixations on a head-based video record, although this limits analysis to video rates (60 Hz) and requires hand-coding of video records.

EYE MOVEMENTS IN PERCEPTION, COGNITION, AND ACTION

Eye movements are necessary because visual sensitivity differs across the retina. Acuity is greatest in the central portion of the fovea, then markedly declines. The organization of the retina can be viewed as a compromise between the need to maintain sensitivity to visual stimuli across a broad range of the visual field and the requirement for detailed spatial resolution in task-relevant aspects of the visual field. This division of labor also helps restrict most processing to a relevant subset of the visual field, thus reducing the amount of information being made available from the visual environment. However, it also requires an eye movement system to quickly bring new regions of the field into the fovea, and to maintain fixation on the most relevant region of the visual field, for visual guidance in fine motor tasks such as reaching for and grasping an object. These demands suggest a close interleaving of eye movements and attention.

Attention and Eye Movements

Smooth pursuit eye movements are controlled by stimulus motion on the retina; they cannot be voluntarily generated or suppressed. Nonetheless, people cannot easily dissociate attention from the target of a pursuit movement. Participants also make anticipatory and predictive pursuit movements, which can be modulated by cues about the likely direction of an upcoming change in the direction of motion, further suggesting a link between attention and the systems. Dual task studies demonstrate that attention shifts precede a saccade. However, the attentional requirements of saccades are quite modest. People can allocate some resources to nontarget areas without delaying saccades or reducing their accuracy. This makes functional sense: some attentional control over saccades is necessary to prevent irrelevant information from continuously triggering saccades, yet it is important to be able to distribute attention across the visual field, and for saccadic eye movement to be a relatively low-threshold, low-overhead action. (*See* **Visual Attention**)

The claim that shifts of attention and saccadic eye movements are closely coupled also receives strong support from neurophysiological studies in monkeys. Cells in the parietal cortex that are active prior to a saccadic eye movement are also active prior to a shift in attention, even when the monkey is required to maintain fixation. (*See* **Attention, Neural Basis of**)

Visual Search

Influential studies of visual search manipulated the number of distracters in the set and the relationship of the target to the distracters. Detection time was relatively unaffected by set size when a single visual feature, such as color, distinguished the target from the distracters. However, search time was a linear function of set size when the target was distinguished from the distracters by a conjunction of features (e.g., color and orientation). Treisman and colleagues proposed that attention is required to bind features together for object recognition. On the assumption that a linear function is diagnostic of a serial search, the resulting rate of search would implicate several shifts of attention within a single fixation. However, this idea seems increasingly untenable. Parallel search can result in linear effects of set size for conjunction searches; moreover, there is a strong correlation between the number of fixations and target detection time in visual search. It seems increasingly likely that visual search involves parallel evaluation of targets within a fixated region.

Close links between fixation and attention are further supported by studies investigating detection of changes in scenes. (*See* **Visual Scene Perception**; **Change Blindness, Psychology of**)

Perceivers are often remarkably insensitive to changes in scenes, a phenomenon known as 'change blindness'. However, detection of changes to attended objects is quite accurate. In a 'flicker' paradigm in which scenes with changes alternate, changes are detected more reliably if the changed region is fixated prior to a change. When 'saccade-contingent' changes are made to a display during a saccade, perceivers accurately detect a change to a recently fixated target, especially the most recently fixated target. Detection of a change to the target of an ongoing saccade is extremely accurate, suggesting that the fixation has followed a shift in attention to the target object.

Eye Movements in Natural Tasks

The close link between attention and fixations suggests that eye movements should provide an

important window into cognitive processes during complex visual tasks that involve both search and memory. Yarbus showed that the scanning patterns for paintings varied depending upon the viewer's task, and appeared systematically related to the content of the paintings. However, trying to infer underlying mental processes from scan-paths alone is extremely difficult.

Researchers have begun to make considerable progress by measuring eye movements in natural tasks with a well-defined goal structure, such as copying a block pattern or making a cup of tea. Several striking results emerge from these studies. First, gaze is almost exclusively directed to task-relevant objects, such as fixation on an object to provide visual guidance for reaching. Second, fixation is closely time-locked with – and nearly always restricted to – objects relevant to the aspect of the task currently being executed. Third, the information maintained in memory from a fixation appears to be limited and task-specific.

For example, one study monitored eye movements as participants used blocks from a resource area to copy a block pattern from a model into a workspace. Participants typically made two fixations to the model area for each block. They first fixated on a block in the model area, then retrieved a block of the same color from the workspace, and then made a second fixation to the model area before placing the block in its correct location in the workspace. The pattern of fixations suggests that participants were retrieving only color information on the first fixation on the block in the model, returning a second time to retrieve information about its relative position. This strategy is consistent with the hypothesis that fixations are used to gather information 'on the fly', as it is needed for the task at hand, thus reducing the need to build and maintain rich internal memory representations. When task demands require use of richer memory representations, participants make fewer saccades and performance slows.

However, it is important to distinguish between explicit memory representations and implicit measures. (*See* **Memory: Implicit versus Explicit**)

In the block copying task, multiple changes to unattended areas of a display increased fixation duration. Moreover, changes in natural scenes that are not explicitly detected receive longer duration fixations than unchanged regions, suggesting an implicit memory representation. Indeed, some such memory is necessary for coordinating complex actions, including making saccadic eye movements to remembered locations.

Eye Movements in Memory, Imagery, and Other Cognitive Tasks

Eye movement measures have been applied in ingenious ways to shed new empirical light on some classic results and claims. For example, after briefly seeing a chessboard with pieces taken from an actual expert game, chess experts show far better recall than novices. However, the expert advantage is eliminated when the target is a board with randomly placed pieces. Chase and Simon suggested that much of this advantage comes from early perceptual processing. Reingold and colleagues confirmed this hypothesis using a gaze-contingent window paradigm, in which pieces outside a moving window are obscured. During a fixation, chess experts distribute attention across larger spatial regions than novices for chess configurations but not for random piece assignments, confirming Chase and Simon's hypothesis. Spivey and Geng confirmed Hebb's hypothesis that constructing images involves generating saccades. When listening to a story, participants generate patterns of eye movements that reflect the spatial content of the story (e.g., looking upwards when the event takes place high up in a building).

Eye Movements in Reading

Much of our knowledge about linguistic processes in reading comes from the study of eye movements. (*See* **Reading, Psychology of**)

Reading has also served as an important empirical test-bed for evaluating models of oculomotor control and models of perceptual processing during fixation. To a first approximation, readers fixate each word in turn, with each saccade averaging seven to nine characters, and the average fixation taking about 250 ms. The probability that a word will be skipped depends largely on its length. Short function words, such as 'the', 'a', and 'of', are skipped 80% of the time. In a language written from left to right, such as English, more than 80% of fixations are to the right of the previous fixation. Contrary to predictions made by some influential early models of reading, even highly predictable words are typically fixated, though statistically less frequently than less predictable words. Moreover, good readers are not more likely to skip predictable words than poor readers are.

The region of text in which there is some uptake of information during a fixation, sometimes called the 'perceptual span', is asymmetrically centered on the fixation. In English, the span extends

approximately four characters to the left of the character at the center of fixation and 15 characters to the right. In Hebrew, which is written from right to left, the direction of the perceptual span reverses. Estimates of the perceptual span come from studies using different saccade contingent changes of the display. These include creating gaze-contingent windows of various sizes, centered around the fixation, with the text altered outside the window, and making changes to a letter contingent on the saccade crossing a predefined (invisible) boundary. The perceptual span varies with the difficulty of a text, becoming smaller for more difficult texts. Processing difficulty localized to individual words also affects the perceptual span.

Although the perceptual span can include several words, processing seems primarily localized to the word currently being fixated and the upcoming word, with somewhat limited uptake of information from peripheral areas of the span. Fixation duration is strongly affected by linguistic variables. Less common words such as 'viola' receive longer-duration fixations than more common words such as 'piano', with effects spilling over to the next word. Ambiguous words such as 'organ' are fixated for longer than unambiguous words of similar frequency. The ambiguity effect is greatest when alternative meanings are equally frequent.

Fixation duration is also modulated by the preceding context. Increased contextual constraint reduces fixation duration. Moreover, the ambiguity effect for homographs with equally common meanings is eliminated when the preceding context strongly biases one sense. However, there are negligible effects of information from the adjacent right context on the current fixation. Taken together, these results suggest that processing is strongly localized to the word currently being fixated and the upcoming word. As in nonlinguistic visual processing, fixations in reading reflect ongoing cognitive processes and cognitive processes appear to center on the word being fixated. Some information is clearly gleaned from text in the periphery, including information about word length, letter shape, and some phonological information. A striking demonstration of the importance of fixation-based processing in reading comes from contrasting the relatively large disruption that comes from perturbing only the word currently being fixated (i.e., replacing it with a random string of letters) with the less disruptive effect of leaving only that word intact.

As in complex visual processing, the inferences that one can draw about the cognitive processes in reading from patterns of fixations depend strongly on understanding the task. A crucial issue is how the uptake and use of information in the periphery guides upcoming saccades and how use of peripheral information is modulated by the attentional demands of the word that is currently being fixated. Unresolved questions include: (1) to what extent does a shift in attention precede an upcoming saccade? (2) Is the distribution of attention to peripheral information influenced by ongoing processing? (3) To what degree do lexical processes modulate the timing of upcoming saccades?

Eye movements are widely used by psycholinguists investigating sentence processing in reading, especially in testing models of syntactic processing. Locally ambiguous sentences that are eventually resolved in favor of the initially less-preferred structure (e.g., 'From his window, John spied on the girl with the telescope in her lap') result in longer fixations in the disambiguating region ('in her lap') and trigger more regressive eye movements. These 'garden path' effects are modulated by a variety of variables, and are the subject of current theoretical debate. Eye movements in reading also provide insights into the interpretation of referring expression such as the pronouns 'his' and 'her' and the timing of inferences in the comprehension of text. (*See* **Sentence Processing: Mechanisms**; **Sentence Processing**)

Eye Movements and Spoken Language

A rapidly expanding community of psycholinguists are now using eye movements to study both spoken language comprehension and language production. (*See* **Language Comprehension**; **Speech Production**)

The use of eye movements in spoken language comprehension was pioneered by Cooper and later extended by Tanenhaus and colleagues using a task in which participants followed spoken instructions to manipulate objects in a visual workspace. Crucially, eye movements are remarkably time-locked to the unfolding input, with participants looking at objects as they become relevant during the instruction.

Figure 1, summarizing results by Eberhard and her colleagues, shows the pattern of fixations of eye movements to a display containing 8 cards, including 2 five of hearts, one of which is below the eight of clubs and the other below a card of another denomination. In the instruction 'Put the five of hearts that is below the eight of clubs above the two of diamonds,' the referent is disambiguated at the word 'eight'. The pattern of fixations as the instruction unfolds shows that listeners looked at

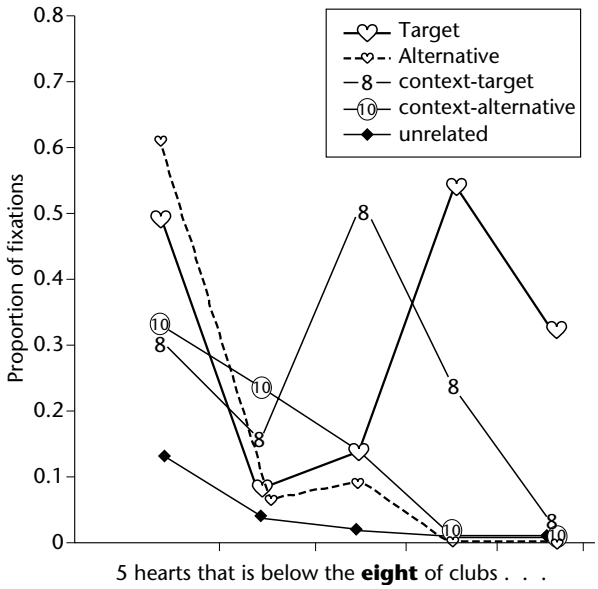


Figure 1.

cards as they became relevant. For example, after 'five of hearts' people were looking at the two potential referents (the five of hearts); at 'below' they were looking at the cards that the fives are beneath (the eight of clubs and/or ten of spades); and at 'eight' they were looking at the eight of clubs.

Fixation patterns are also closely time-locked to utterance generation in language production when speakers describe scenes. Speakers fixate on a referent for approximately 800 ms before beginning to generate the phrase containing its name, shifting their gaze to the next entity to be described in the middle of the previous phrase.

Eye movements are now being used to trace the time course of lexical access, reference resolution, sentence processing in preliterate children, and utterance planning in production, as well as issues in higher-level interactive conversation.

CONCLUSION

The organization of the retina, in particular the presence of the fovea, necessitates that the eyes be able to move to maintain central fixation and shift fixation. These eye movements are closely linked to shifts in attention and have an important role in visual search, perception of scenes, and complex visual tasks. The close coupling of fixations to perception and action allows cognitive scientists to use eye movements as windows in moment-by-moment cognitive processes in vision, problem-solving and in language. Eye movement research on reading and on oculomotor control has flourished for decades. More recently there has been a dramatic increase in the use of eye movements for research in other domains within the cognitive sciences. This trend seems likely to continue as cognitive scientists increasingly turn towards the use of natural tasks that combine perception and action, including language and vision.

Further Reading

Ballard D, Hayhoe M, Pook P and Rao R (1998) Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* **20**: 723–767.

Chase WG and Simon HA (1973) Perception in chess. *Cognitive Psychology* **4**: 55–81.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 97–136.

Cooper RM (1974) The control of fixation by spoken language: a new methodology for real time investigations of speech perception, memory and language processing. *Cognitive Psychology* **6**: 84–107.

Eberhard KM, Spivey-Knowlton MJ, Sedivy JC and Tanenhaus MK (l995) Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research* **24**: 409–436.

Eckstein MP (1998) The lower visual search difficulty for conjunctions is due to noise and not serial attentive processing. *Psychological Science* **9**: 111–118.

Findlay JM and Gilchrist ID (1998) Eye guidance and visual search. In: Underwood G (ed.) *Eye Guidance in Reading and Scene Perception*, pp. 295–312. Oxford, UK: Elsevier.

Findlay JM and Walker R (1999) A framework for saccadic control based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences* **22**: 661–721.

Frazier L and Rayner K (1987) Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* **14**: 178–210.

Griffin ZM and Bock K (2000) What the eyes say about speaking. *Psychological Science* **11**: 274–279.

Hayhoe M (2000) Vision using routines: a functional account of vision. *Visual Cognition* **7**: 43–64.

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: John Wiley.

Henderson J and Hollingworth A (1998) Eye movements during scene viewing; an overview. In: Underwood G (ed.) *Eye Guidance in Reading and Scene Perception*, pp. 269–294. Oxford, UK: Elsevier.

Khurana B and Kowler E (1987) Shared attentional control of smooth eye movement and perception. *Vision Research* **27**: 1603–1618.

Kowler E (1995) Eye movements. In: Kosslyn SM and Osherson DN (eds) *An Invitation to Cognitive Science*, 2nd edn, vol. 2, *Visual Cognition*. Cambridge, MA: MIT Press.

Kowler E (1999) Eye movements and visual attention. In: Wilson RA and Keil FC (eds) *The MIT Encyclopedia of the Cognitive Sciences*, pp. 306–309. Cambridge, MA: MIT Press.

Kowler E, Anderson B, Dosher B and Blaser E (1995) The role of attention in the programming of saccades. *Vision Research* **35**: 1897–1916.

Kustov AA and Robinson DL (1997) Shared neural control of attentional shifts and eye movements. *Nature* **384**: 74–77.

Liversedge S and Findlay J (2001) Saccadic eye movements and cognition. *Trends in Cognitive Sciences* **4**: 6–14.

McConkie GW and Rayner K (1975) The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics* **17**: 578–586.

Meyer AS, Sleiderink AM and Levelt WJM (1998) Viewing and naming objects: eye movements during noun phrase production. *Cognition* **66**: B25–B33.

O'Regan JK (1992) Sharing the 'real' mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology* **46**: 461–488.

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* **124**: 372–422.

Reichle ED, Pollatsek A, Fisher DL and Rayner K (1998) Toward a model of eye movement control in reading. *Psychological Review* **105**: 125–157.

Reingold EM, Charness N, Pomplun M and Stampe DM (2001) Visual span in chess players: evidence from eye movements. *Psychological Science* **12**: 48–55.

Schall JD and Thompson KG (1999) Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience* **48**: 269–297.

Simons DJ and Levin DT (1997) Change blindness. *Trends in Cognitive Sciences* **1**: 261–267.

Spivey MJ and Geng J (2001) Ocolumotor mechanisms triggered by images and memories: spontaneous eye movements to objects that aren't there. *Psychological Research*

Steinman RM, Kowler E and Collewijn (1990) New directions for oculomotor research. *Vision Research* **30**: 1845–1864.

Treisman A and Galdade G (1980) A feature integration theory of attention. *Cognitive Psychology* **12**: 97–136.

Vivianni P (1990) Eye movements in visual search: cognitive, perceptual and motor control aspects. In: Kowler E (ed.) *Eye Movements and Their Role in Vision and Cognitive Processes*. Amsterdam, Netherlands: Elsevier.

Yarbus AL (1967) *Eye Movements and Vision*. New York, NY: Plenum Press.

Zelinksy GJ, Rao RPN, Hayhoe MM and Ballard DH (1997) Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science* **8**: 448–453.

Face Perception, Psychology of

Intermediate article

Alice J O'Toole, University of Texas, Richardson, Texas, USA

CONTENTS

Perceptual information from the human face can be used to determine the identity, sex, race, age, and current mood of an individual. It may serve also as a guide for social interaction, providing us with continually changing feedback about the emotional state of the people with whom we interact.

INTRODUCTION

The number of different faces we must recognize as individuals makes the challenges associated with remembering faces nearly unique in the realm of visual memory. No other class of objects places such stringent requirements on visual memory, and simultaneously calls on human abilities to integrate information into a social context. To recognize and categorize faces, the facial 'features' useful for accomplishing these tasks must be extracted from a moving three-dimensional surface, encoded visually, and represented in memory. Evidence of the importance of face perception to human survival can be seen in the complex network of brain regions specialized for these functions (Haxby *et al*., 2000). (*See* **Face Perception, Neural Basis of**; **Face Cells**)

FACES AS CONFIGURAL OR HOLISTIC UNITS

Human faces can be defined by the universally recognizable configuration of facial features they share. Indeed, newborn infants only a few hours old respond to faces and face-like configurations more than to other visual patterns (Nelson, 2001). Though all faces share the same 'features' arranged in the same basic configuration, we are able to recognize hundreds, if not thousands, of people by their faces. These impressive abilities are thought to drive from our expertise at perceiving and remembering subtle variations in the configuration of the facial features. The reliance of the human perceptual system on configural rather than feature-based information in perceiving faces has been demonstrated using various experimental manipulations aimed at perturbing the configuration of a face or at disrupting our ability to process the configural information in a face. These manipulations include distortion of the relative positions of the mouth, eyes, and nose (Bartlett and Searcy, 1993), inverting a face (Yin, 1969), and altering the vertical alignment of the contours (Young *et al*., 1987). The results of such studies have supported the view that face perception and recognition rely heavily on processing the configural information in faces. The perceptual importance of the facial configuration can be powerfully illustrated using a variation of the face inversion technique. The 'Margaret Thatcher illusion' (Thompson, 1980) illustrates that inverting a face interferes with our ability to detect even an extreme distortion of the configural information in a face (see Figure 1).

FEATURES OF FACES

What are the features of the human face? Although we generally think of the eyes, nose, and mouth as facial features, these elementary facial parts are neither a psychologically valid nor a computationally adequate description of a face. From both a psychological and a computational perspective, the primary function of a facial feature set is to provide a quantification of the information in faces. An effective feature set must therefore retain the information that makes a face uniquely recognizable as an individual and that identifies it categorically (e.g. as male or female). There is still disagreement on what constitutes a good facial feature set. However, computational studies of face recognition, in combination with what is known about the factors that affect human performance, have suggested that the features we use to encode faces may be derived from the statistical structure of the faces we experience.

Figure 1. The 'Margaret Thatcher illusion', so named for the use of the British Prime Minister's face in the original demonstration (Thompson, 1980). The illusion shows that distortions in the configuration of the face, which appear grotesque when the face is viewed upright, are barely noticed when the face is viewed upside down.

FACE SPACE

The theoretical construct of an abstract multidimensional 'face space' has been proposed as a model of human memory for faces (Valentine, 1991). This construct assumes that the human face memory system can be considered metaphorically as a multidimensional space. Each axis of the space represents a feature and individual faces are represented by points in the space. The coordinates of a face in the space, therefore, indicate the feature values of the face. Similarity between pairs of faces is modeled as the distance between the faces in the multidimensional space.

Face space theory was proposed originally to give insight into certain phenomena in human face perception that relate to the importance of facial distinctiveness for recognition. Faces judged as 'distinct' are recognized more accurately than faces judged as 'typical' (Light *et al.*, 1979). In the context of face space theory, the distinctiveness of a face can be measured as its distance from the mean or 'prototype' of the space, or alternatively, as the density of faces around its position in the space. Faces close to the prototype are thought to be more similar to most other faces than faces far from the prototype. These measures have been hypothesized to be indicative of the confusability of the face with other known faces in the space. The face space theory thus provides a useful framework for conceptualizing the similarity relationships among faces and their importance in predicting recognition accuracy.

EIGENVECTOR REPRESENTATIONS

The face space theory of face recognition can be modeled computationally using principal components analysis (PCA) applied to sets of faces. This analysis provides a concrete model of a face space. Principal components (eigenvectors) are derived from the statistical structure of the input faces, and form the 'feature' axes of the space. Eigenvectors can be ordered according to their 'eigenvalues', which measure the proportions of variance that the eigenvectors explain in the data set. An individual face can be expressed as a linear combination of the eigenvectors, where the coefficients are the coordinates that define the position of the face in the space. This interpretation of PCA makes it analogous to the abstract face space theory proposed by Valentine (1991). The eigenvectors comprise a feature set for describing the faces. The coordinates, which define the position of a face in the space, specify feature values for individual faces. It is worth noting that the eigenvalues add to the generic face space model an inherent measure of the importance of individual 'features' for describing faces. (*See* **Pattern Recognition, Statistical**)

PCA has been applied directly to a number of relatively unprocessed image- and surface-based representations of faces. These include raw images, three-dimensional surface representations from laser scanners, and several 'corresponded' or aligned image- and surface-based representations. It is important to note that the eigenvectors that emerge from PCA represent faces according to the same representation as was analyzed. For example, when two-dimensional images of faces are analyzed, the resulting eigenvectors are also two-dimensional images. Therefore, eigenfeatures, or 'eigenfaces', can be displayed visually, alone, or in combination. Recently, corresponded facial codes have emerged as powerful tools for facial synthesis (Blanz and Vetter, 1999). This is because the resulting face space supports continuous 'morphing' between individual faces via simple linear transformations of the coordinates in the face space.

PCA-based codes have been used to model the effects of distinctiveness on recognition memory for faces and have been applied to the analysis of the categorical information specifying the sex, race,

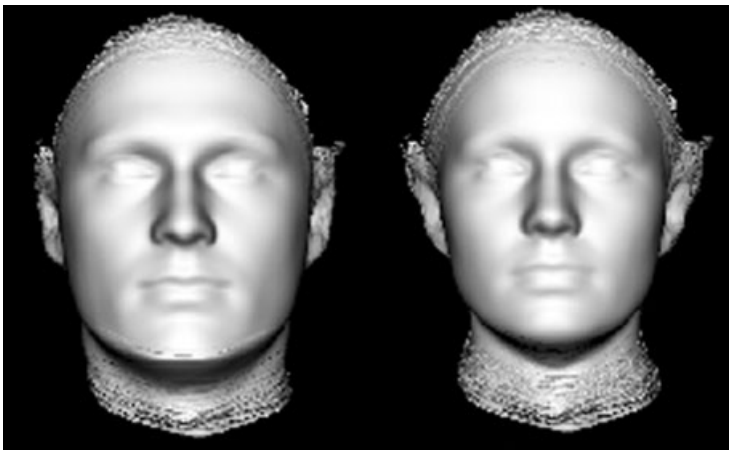


Figure 2. An illustration of the localization of the information that specifies the sex of the face in the PCA code. The first eigenvector captures a global configural contrast between male and female faces. On the left, the first eigenvector is added to the average head. On the right, the first eigenvector is subtracted from the average head.

and expression of a face. The importance of individual eigenvectors for categorization has been assessed by simple neural network algorithms. These networks learn to predict the categorical status of faces using their coordinates in the face space. Several studies have shown that eigenvectors with large eigenvalues tend to encode the categorical information in faces, whereas eigenvectors with smaller eigenvalues tend to encode information about the identity of a face. An example of this principle for sex classification appears in Figure 2. Global configural information about the sex of faces emerges in the eigenvector that explains the largest proportion of variance in the face set (O'Toole *et al*., 1997).

Finally, it is worth noting that the feature sets derived from PCA are dependent on the composition of the set of faces analyzed. This dependence allows the model to simulate specific aspects of human experience and learning in the acquisition of facial features. We will return to this point below when we consider the other-race effect. (*See* **Perceptual Learning**)

MDS REPRESENTATIONS

The abstract face space provides a theoretical framework for understanding human memory for faces, while PCA provides a computational model of this theoretical construct. Multidimensional scaling (MDS) analysis provides data about the psychological validity of the face space theory and about the adequacy of computational implementations of it. Though computationally analogous to PCA, MDS generally refers to the analysis of data produced by human observers. These data comprise similarity judgments between all possible pairs of faces in a set. The resulting human-generated 'distance' measures are used to create a multidimensional map that preserves the psychologically-based similarity relationships among the faces. In this psychological face space, faces are again points in a multidimensional space. The axes are again ordered according to the proportion of variance they explain in the human similarity data; but for most psychological applications, only a small number of axes (usually fewer than four) are retained. Determining the psychological 'features' represented by the axes, however, is a speculative endeavor that involves comparison of faces at opposite ends of the axes.

For many years, MDS has been applied to the analysis of human judgments of face similarity (see Shepherd *et al*. (1981) for the early history of these studies). The early studies were aimed at revealing the basic psychological dimensions underlying human face perception. However, the dimensions of variation that emerged from the analysis were found to depend strongly on the homogeneity of the face set judged by observers. More recently, MDS has been used to focus on more specific questions about the organization of face memory and about the general accord between the assumptions of face space models and the psychological similarity data. MDS analysis of human similarity data has also been used to provide a benchmark for assessing the validity of the computationally-derived face spaces that emerge from different kinds of face input encodings (e.g., two-dimensional image versus three-dimensional surface data).

THE OTHER-RACE EFFECT

The other-race effect for face recognition is the well-known fact that people recognize faces of their own race more accurately than faces of other races (Malpass and Kravitz, 1969). A number of hypotheses have been advanced to explain the other-race effect, including contact with faces of other races, prejudice, and race as a 'feature'. Of these, the 'contact hypothesis' has been investigated most thoroughly. The contact hypothesis assumes that we have more experience with faces of our own race than with faces of other races, and that this experience differential affects the quality of the encoding we make for own- versus other-race faces. An indication of our inability to create distinctive codings for other-race faces is the well-known anecdote 'they all look alike to me'. Studies of the contact hypothesis have related contact with members of other races to the magnitude of the other-race effect. Although these studies have yielded inconsistent results with adult participants, the few studies done with children have yielded more consistent results (Shepherd, 1981). This pattern of findings may indicate the importance of early childhood contact with other-race faces in the acquisition of feature sets, but further experimentation is needed before a firm conclusion can be reached. (*See* **Expertise**)

CONCLUSION

Human face processing relies heavily on the configural information in the face. The face space theory provides a theoretical construct for understanding the organization of human memory for faces. This construct can be modeled computationally with PCA and can be related to empirical data from human observers using MDS. Facial features are unspecified in the original model of Valentine (1991). In the computational framework, features are derived from the statistical structure of the input faces. This kind of analysis can accommodate differences in the experience profiles that individuals have with faces (e.g. experiencing faces of different races in different proportions), and can yield feature sets that are in some senses optimal for the task.

References

Bartlett JC and Searcy J (1993) Inversion and configuration of faces. *Cognitive Psychology* **25**: 281–316.

Blanz V and Vetter T (1999) A morphable model for the synthesis of 3D faces. In: *SIGGRAPH'99 Conference Proceedings*, pp. 187–194. Computer Society Press.

Haxby JV, Hoffmann EA and Gobbini MI (2000) The distributed human neural system for face perception. *Trends in Cognitive Sciences* **6**: 223–233.

Light L, Kayra-Stuart F and Hollander S (1979) Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory* **5**: 212–228.

Malpass RS and Kravitz J (1969) Recognition for faces of own and other race faces. *Journal of Personality and Social Psychology* **13**: 330–334.

Nelson CA (2001) The development and neural basis of face recognition. *Infant and Child Development* **10**: 3–18.

O'Toole AJ, Vetter T, Troje NF and Bülthoff HH (1997) Sex classification is better with three-dimensional head structure than with image intensity information. *Perception* **26**: 75–84.

Shepherd J (1981) Social factors in face recognition. In: Davies G, Ellis H and Shepherd J (eds) *Perceiving and Remembering Faces*, pp. 55–79. London, UK: Academic Press.

Shepherd J, Davies G and Ellis H (1981) Studies of cue saliency. In: Davies G, Ellis H and Shepherd J (eds) *Perceiving and Remembering Faces*, pp. 105–131. London, UK: Academic Press.

Thompson P (1980) Margaret Thatcher: a new illusion. *Perception* **9**: 483–484.

Valentine T (1991) A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology* **43A**: 161–204.

Young AW, Hellawell D and Hay DC (1987) Configurational information in face perception. *Perception* **16**: 747–759.

Yin RK (1969) Looking at upside-down faces. *Journal of Experimental Psychology* **81**: 141–145.

Further Reading

Bruce V, Young A and Young AW (1998) *In the Eye of the Beholder: The Science of Face Perception*. Oxford, UK: Oxford University Press.

Levin DT (2000) Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General* **129**: 559–574.

O'Toole AJ, Abdi H, Deffenbacher KA and Valentin D (1993) Low dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America* **10**: 405–411.

Turk M and Pentland A (1991) Eigenfaces for recognition. *Journal of Cognitive Neuroscience* **3**: 71–86.

False Memory

Intermediate article

Elizabeth F Loftus, University of Washington, Seattle, Washington, USA

CONTENTS

False memory refers to the experience of thinking that we remember something from our past that did not really happen.

INTRODUCTION

Although memory is one of the most fundamental activities of the human mind, and generally serves us well, it is also true that memory distortion is a fact of life. Portions of what we think we remember are contrary to truth. We sometimes remember the essence of events from the past, but get some of the details wrong. At other times, we remember entire events that were never really experienced by us. Work on memory distortion has not only captured the attention of cognitive psychologists, but has also fascinated physicists and philosophers throughout recorded history. Understanding how we can be tricked by memory is essential for developing ways of avoiding errors, and for resolving memory conflicts in everyday life.

EYEWITNESS TESTIMONY

Thousands of studies have been conducted to determine how people remember accidents, crimes, and other significant events from the past. One popular method of conducting these studies involves simulating what real eyewitnesses might experience. Typically, people are shown simulated crimes or accidents, and are then tested for their memory of the details of these events. So many different factors can affect the accuracy of what a witness remembers, that it has been convenient to divide the factors into three major categories: (1) factors that affect the initial perception or encoding of the event; (2) factors that affect the retention of information in memory; and (3) factors that affect the retrieval of information from the memory system (Loftus, 1979/1996; Wells *et al*., 2000).

Factors Affecting Encoding

When a witness sees a crime, such as an armed robbery, there are some obvious factors that influence the accuracy of the initial perception, such as how good the lighting is or how far away the witness is standing. But there are also some not-so-obvious factors. While it may seem evident that the more time a person has to see the robbery, or the robber's face, the better the memory would be. That's true. But what is not so obvious is that when witnesses try to report how much time they had to look at something, they invariably stretch out those times in their mind. They think they had longer to look than they actually did. In one study people who saw a simulated bank robbery lasting 30 seconds sometimes claimed it lasted for 3 minutes, 5 minutes, or even longer (Loftus *et al*., 1987).

The stress or fright that a witness feels during the robbery can also be a factor. While many people believe that extreme stress leads to an indelible fixation in the mind, in fact the opposite can be true. With highly severe stress, there can be impairments in memory, particularly for the peripheral details of the event. People can focus in on the frightening weapon, and have difficulty remembering the face of the person holding the weapon.

There are a variety of other factors that affect the accuracy of the memory right at the moment a key event occurs. Some have to do with characteristics of the event itself, such as the lighting or duration of the event, or whether a weapon is present. For example, there is a phenomenon called 'weapon focus': when people see a crime involving a weapon, the weapon captures some of their processing attention, and this can result in good memory for the weapon and poorer memory for other details.

Another set of factors concerns characteristics of the witness (such as age, or whether the witness was

under the influence of alcohol, drugs, or other substances). For example, while it might seem obvious that being under the influence of a good deal of alcohol might affect perception and memory, in fact research suggests that as few as two or three drinks can significantly impair the formation of memory. As for the age of a witness, there is now ample evidence that young children have poorer memories than older children and adults, and their memories are more vulnerable to suggestive interviewing and other bad influences (Ceci and Bruck, 1993). At the other end of the age spectrum, research also shows that the elderly sometimes have more trouble forming new memories than do young adults. A selection of further examples are documented in many books devoted solely to the topic of eyewitness testimony (e.g. Cutler and Penrod, 1995; Loftus and Doyle, 1997). In addition, thousands of scientific articles have been published, each devoted to one or more of the various factors.

Factors Affecting Retention

After the event has been witnessed, the memory does not simply sit in the mind waiting to be plucked out. Rather, people are often exposed to new information about the event, and when this postevent information is misleading in some way, people make errors when they later report what they saw. The new information can become incorporated into the recollection, supplementing or altering it, sometimes in dramatic ways.

To show the power of postevent information, researchers have used a simple procedure where witnesses see a simulated accident that, say, involves a car going through a stop sign at an intersection. Later, half receive new misleading information about the event, for example that it was a yield sign. Sometimes the misinformation is presented in the form of a leading question (e.g. 'Did another car pass the red Datsun while it was at the intersection with the yield sign?'); or the misinformation might be embedded in another witness's report to which the original witness is exposed. A separate group of witnesses would not be exposed to any misinformation. Finally, all participants try to recall the original event. When asked about the key details, those given the phony postevent information tend to adopt it as their memory. In the example involving the traffic signs, they claimed that they saw a yield sign. People who had not received the phony information gave much more accurate reports. In some studies, providing misinformation has led to people being half as accurate as they would have been without the misinformation. Distortions in memory from inaccurate postevent information have been found with a wide range of materials. People have recalled nonexistent broken glass and tape recorders, a clean shaven man as having a moustache, and even something as large and conspicuous as a barn in a country scene that contained no buildings at all. The change in report after people are exposed to misinformation is often referred to as the 'misinformation effect'.

A prolonged intellectual debate has occurred over the last few decades about the misinformation effect. Namely, when suggestive information contaminated a person's memory report, did it actually alter the underlying memory traces? Or, did the suggestive information coexist with the original memory traces and under the right circumstances the original might still be accessible? This became known as the debate about the 'permanence of memory' (Loftus and Loftus, 1980). Although this debate continues, what is accepted by virtually all scientists who work in this area is that misinformation can produce profound changes in people's reports of their past experiences.

Factors Affecting Retrieval

A critical stage for an eyewitness is when the person conveys to others what was seen at the original event. This can occur when the witness gives a statement to the police, or tries to identify a culprit from a lineup, or testifies in court. A new set of factors comes into play at this stage. If the witness is asked biased or leading questions, the report can be distorted. Even a simple question like 'Did you see the broken glass?' can lead people to false reporting if there was no broken glass. The use of the definite article 'the' (which implies that there *was* broken glass) is enough to influence what people say. Leading questions can clearly distort what people report. Moreover, there is some evidence that misleading questions can also influence the actual memories.

When witnesses make a memory report, they are often asked for their level of confidence. 'I'm positive that's the guy. I'll never forget that face.' As it turns out, the confidence that a witness expresses is only weakly related to the person's accuracy level. People can be confident, but wrong. Moreover, the level of confidence that a witness expresses can be influenced by feedback that the witness gets (Wells and Bradfield, 1999). If a witness is told 'You picked our suspect', or 'Another witness picked the same guy', or 'The guy's fingerprints were found at the scene', the witness's confidence in her or his memory rises. Confidence can rise, even

when the feedback given is completely inaccurate. Thus confidence itself is malleable, subject to distortion – a kind of false memory of its own.

PLANTING FALSE MEMORIES

It is one thing to see that small details about an event from the past can be distorted, but quite another thing to consider that entire memories can be 'planted' into the mind of people for things that never occurred. Two types of false memory research have been done. One type reveals the creation of a false memory for a word that was never presented, and hundreds of studies of this type have been conducted in part because of the ease of gathering large amounts of data. A second type reveals the creation of false memories of childhood experiences that are far more complex, such as being lost, or being hospitalized, or being injured.

False Memories for Never-seen Words

The paradigm for creating false memories of words is a straightforward laboratory task in which people study lists of related words (e.g. *thread*, *pin*, *eye*, *sewing*, *sharp*, *point*, *prick*, *thimble*, *haystack*, *thorn*, *hurt*, *injection*, *syringe*, *cloth*, and *knitting*). They are all related to a critical lure word that is not presented (*needle*). At a later time, when people have to remember the words that were presented, they frequently falsely recall the critical word. The false recall of the critical lure occurs frequently – as often as half the time or more, and is often at a rate even greater than the recall of actual words that were presented on the list (Roediger and McDermott, 1995).

Because the false word memory studies are so easy to conduct, and large amounts of data can be collected so readily, and the false memory is so reliably produced, the paradigm has been a favorite of memory researchers. Numerous investigators have gone on to learn what they can about memory distortion by ploughing this fertile ground. They have shown, for example, that when people are warned about the problem that a critical nonpresented lure should be avoided, they can reduce the false memory effect somewhat, but the warning does not begin to come close to eliminating the effect.

Individual differences have also been observed in the false word memory paradigm. Older, as opposed to younger, adults have shown a somewhat greater tendency to falsely recall the critical lure. Alzheimer patients show increased false recall. Self-reported episodes of dissociative experiences (e.g. forgetting whether you said something or intended to say it) are connected to increased false recall. And women who have experienced a traumatic event and are suffering from posttraumatic stress disorder (PTSD) have revealed greater false memory for critical lures (Bremner *et al.*, 2000).

Why do people report hearing a word that never occurred? A widely embraced explanation is that while listening to words that are presented (e.g. *thread*, *pin*, *sewing*), the critical lure (*needle*) may come to mind. This might occur consciously, and the person is actually aware of thinking of the lure. Or it might occur unconsciously when an underlying mental representation of the lure is activated. Later, when tested on the lure, the person is confused about whether he or she actually heard it or whether it seems familiar for another reason.

The false memory for words paradigm has been criticized as being artificial, and so different from the kinds of false memories that occur in real-world conditions. Although many researchers believe that the paradigm is a highly valuable one for studying the mental processes that are involved in creating false memories in both laboratory and real-world conditions, it is still useful to explore other memory distortion paradigms and to see whether people can be led to false memories of a more realistic nature.

Planting Childhood Memories

Studies have even shown that more realistic false events can be planted in memory. One example is the research showing that people could be convinced that they had been lost in a shopping mall as children, that they had been upset by the experience, but eventually rescued by an elderly person and reunited with their families. A fairly strong form of suggestion was used in these studies: subjects were 'told' that their parents or older relatives remembered the experience, and the subjects were asked to also remember. In other research using 'parental suggestion', people were led to believe and remember that they had been hospitalized with ear pain, that they had knocked over a punchbowl at a family wedding and spilled punch over the parents of the bride, or that they had been victims of a serious animal attack (Hyman *et al.*, 1995; Loftus, 1997). Real memories and the planted memories differed statistically on a number of dimensions, such as confidence or vividness. Despite the statistical differences when a group of real memories is compared to a group of planted false ones, it was still the case that many of the false memories were vivid and held with confidence and virtually indistinguishable from the real memories

(Porter *et al.*, 1999). Moreover, when people are asked to repeatedly rehearse their false memories, they come to increasingly possess the characteristics of true memories, making the job of distinguishing true from false all that much harder.

The sad fact is that without independent corroboration there is no reliable way at the moment to tell the difference between real memories and ones that are a product of suggestion.

Imagination and False Beliefs

Telling people that a family member has reported a past event is a powerful form of suggestion, so powerful that it has even led people to confess to things that they did not do. But lesser forms of suggestion can also affect past beliefs and memories. When people are led to simply imagine false events they show increased confidence that those events were personally experienced. This confidence-enhancing phenomenon is called 'imagination inflation'. In these imagination inflation studies, people are asked to spend a minute imagining that they had an experience as a child that would have been upsetting if it had occurred. For example, they imagined that they were playing near a window and tripped on something, fell towards the window and broke it with their hand, rendering them cut and bloody. Later on, those who underwent the imagination activity showed enhanced confidence that they had personally experienced breaking a window with their hand. And writing about the contrary-to-truth event can have a similar confidence-boosting effect.

These imagination-type activities can not only make people falsely believe that they had experiences as children, but they can also make people believe that they performed actions a few weeks earlier that they did not perform (Goff and Roediger, 1998). People who imagined doing something multiple times, such as playing drums with toothpicks, frequently falsely claimed that they had actually performed these actions on a specified occasion two weeks earlier.

Some people are more susceptible to imagination inflation than others. For example, people who have self-reported lapses in memory and attention, or people who show better imagery abilities, display more imagination inflation (Heaps and Nash, 1999; Horselenberg *et al.*, 2000).

Why does imagination affect our memory for the past? One explanation is that a simple source confusion error has occurred. When people imagine that they had an experience, there is later on a separation between the content of the information and its source. People remember the content but mistakenly attribute it to the wrong source – they attribute it to their own experience rather than to the imagination activity. The idea that people routinely make source-monitoring errors has a long tradition of study in cognitive psychology outside the area of imagination. People sometimes remember hearing a fact, but misremember who told them that fact. They sometimes remember that they said something to their spouse (e.g. 'remember to buy milk'), but it turns out they only thought about saying that directive, and the spouse comes home empty-handed. (A good discussion of source-monitoring research can be found in Johnson *et al.*, 1993.)

Another explanation for how imagination affects memory is that the imagination activity makes the information seem more familiar. At the time people are asked about an experience, they judge whether it happened to them on the basis of familiarity. Enhanced feelings of familiarity lead them to enhanced confidence that the event was a personal experience (Garry and Polaschek, 2000).

Dreams and False Memory

One commonly used technique in psychotherapy is dream interpretation. Some therapists believe that dreams convey some innermost truth about the past. However, exactly what that truth is appears to depend on the clinician who is doing the interpreting.

Some recent studies have shown that dream interpretation is a powerful way to make people believe that they had experiences in the past that they probably did not have. In this research, subjects reported on various childhood experiences on two separate occasions. Between these two sessions, they went to a clinician who analyzed a dream report. No matter what the person dreamed, the clinician steered them into believing that the dream revealed that they had experienced, before age three, a critical event such as being lost in a public place, or abandoned by their parents, or that they faced a threat to their lives and had to be rescued from that threat. A typical finding is that the dream interpretation led people to believe that they had had the suggested experience. Some of them also went on to develop concrete narrative descriptions of these made-up experiences (Mazzoni *et al.*, 1999).

One of the reasons why this type of suggestion is so effective is that the information is ostensibly coming from an authority figure and it is highly personalized. The clinician is saying, in essence,

'I'm an expert, I've examined your dream, and here's what it means.'

IMPLICATIONS OF FALSE MEMORY RESEARCH

Relevance to the Study of Memory

That memory is constructive is something that psychologists have known for nearly a century. That our memories can fail us because of difficulties that occur at the time information is laid down in memory, or at later times, has been well documented. More recently, the power of suggestion to go beyond simple tinkering with memory, and to lead people to entirely false beliefs about their past, has been shown using a variety of experimental paradigms, such as parental suggestion, imagination, or dream interpretation.

These findings teach us about the rather flimsy curtain that sometimes separates genuine memory from imagination and other processes. We are just beginning to explore how brain activity might help us differentiate real from suggested memories. Studies that explore the activity of the hippocampal and other regions of the brain – using neuroimaging or evoked potentials, for example – are in their infancy but eventually have the potential to help us understand how false memories develop and why they are so often experienced as genuine memories.

Relevance to Legal and Clinical Settings

There are practical implications of memory distortion research to police interrogations, to clinical settings, and to other domains of real-life experience. Police interrogations often involve the strategy of supplying information ostensibly produced by other witnesses. 'Mrs Jones said it happened this way, what do you think?' 'Mrs Jones picked the same person in the lineup.' These are just the kinds of statements that can distort the content of a person's memory, or the confidence with which a memory is expressed. Since judges and juries are influenced by the detail and the confidence of witness testimony, these interventions can improperly affect the outcome of cases. A number of cases of wrongful convictions have recently come to light – individuals who served lengthy prison sentences but who were eventually exonerated by DNA evidence. These cases reveal that faulty memory was a major cause of the wrongful conviction. Fortunately, various government agencies are beginning to appreciate the problems of memory, and to implement procedures that will reduce the mistakes made by witnesses in actual cases (Technical Working Group, 1999).

The finding that memories can be distorted, and even planted entirely, has relevance to the 'repressed memory controversy'. In recent times, numerous psychotherapists have acknowledged the use of hypnosis, guided imagery, imagination exercises, and other memory-focused techniques designed to help patients recover allegedly repressed memories. Self-help books with imagination and writing exercises that encourage counterfactual expression can be having enormous unintended side effects. These techniques have been met with skepticism, precisely because of their ability to lead people to develop false memories. In therapy, there is often pressure brought to bear on patients to comply with a therapist's conceptualization of the patient's difficulties. Good clinical practice demands that therapists appreciate how their treatments work, and what possible side effects those 'treatments' might be having.

References

Bremner JD, Shoebe KK and Kihlstrom JF (2000) False memories in women with self-reported childhood sexual abuse. *Psychological Science* **11**: 333–337.

Ceci SJ and Bruck M (1993) Suggestibility of the child witness: a historical review and synthesis. *Psychological Bulletin* **113**: 403–439.

Cutler BL and Penrod SD (1995) *Mistaken Identification: The Eyewitness, Psychology, and the Law*. New York, NY: Cambridge University Press.

Garry M and Polaschek DLL (2000) Imagination and memory. *Current Directions in Psychological Science* **9**: 6–10.

Goff LM and Roediger HL, III (1998) Imagination inflation for action events: repeated imaginings lead to illusory recollections. *Memory & Cognition* **6**: 20–33.

Heaps C and Nash M (1999) Individual differences in imagination inflation. *Psychonomic Bulletin and Review* **6**: 313–318.

Horselenberg R, Merckelbach H, Muris P, Rassin E, Sijsenaar M and Spaan V (2000) Imagining fictitious childhood events. *Clinical Psychology and Psychotherapy* **7**: 128–137.

Hyman IE, Husband TH and Billings FJ (1995) False memories of childhood experiences. *Applied Cognitive Psychology* **9**: 181–197.

Johnson MK, Hastroudi S and Lindsay SD (1993) Source monitoring. *Psychological Bulletin* **114**(1): 3–28.

Loftus EF (1979/l996) *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Loftus EF and Doyle JM (1997) *Eyewitness Testimony: Civil and Criminal*. Charlottesville, VA: Lexis Law Publishing.

Loftus EF and Loftus GR (1980) On the permanence of stored information in the human brain. *American Psychologist* **35**: 409–420.

Loftus EF, Schooler JW, Boone SM and Kline D (l987) Time went by so slowly: overestimation of event duration by males and females. *Applied Cognitive Psychology* **1**: 3–13.

Mazzoni GAL, Loftus EF, Seitz A and Lynn SJ (1999) Creating a new childhood. *Applied Cognitive Psychology* **13**: 125–144.

Porter S, Yuille JC and Lehman DR (1999) The nature of real, implanted, and fabricated memories for emotional events. *Law and Human Behavior* **23**: 517–537.

Roediger HL and McDermott KB (1995) Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–814.

Technical Working Group for Eyewitness Evidence (1999) *Eyewitness Evidence: A Guide for Law Enforcement*. Washington, DC: United States Department of Justice, Office of Justice Programs.

Wells GL and Bradfield AL (1999) Distortions in eyewitnesses' recollections: can the postidentification feedback effect be moderated *Psychological Science* **10**: 138–144.

Wells GL, Malpass RS, Lindsay RCL, Fisher RP, Turtle JW and Fulero SM (2000) From the lab to the police station: a successful application of eyewitness research. *American Psychologist* **55**(6): 581–598.

Further Reading

Leippe MR (1995) The case for expert testimony about eyewitness memory. *Psychology, Public Policy, and Law* **1**: 909–959.

Roediger HL and McDermott KB (2000) Tricks of memory. *Current Directions in Psychological Science* **9**: 123–127.

Ross DF, Read JD and Toglia MP (eds) (1994) *Adult Eyewitness Testimony*. New York, NY: Cambridge University Press.

Game Theory

Introductory article

David K Levine, University of California, Los Angeles, California, USA

CONTENTS

Game theory is the mathematical study of human interactions described by rules of play and alternative choices.

INTRODUCTION

Situations that economists and mathematicians call 'games' psychologists call 'social situations'. While game theory has applications to 'games' such as poker and chess, it is the social situations that are the core of modern research in game theory. Game theory has two main branches: noncooperative game theory models a social situation by specifying the options, incentives, and information of the 'players' and attempts to determine how they will play; cooperative game theory focuses on the formation of coalitions and studies social situations axiomatically. This article will focus on noncooperative game theory.

Game theory starts from a description of the game. There are two distinct but related ways of describing a game mathematically. The *extensive form* is the most detailed way of describing a game. It describes play by means of a *game tree* that explicitly indicates when players move, which moves are available, and what they know about the moves of other players and nature when they move. Most important, it specifies the *payoffs* that players receive at the end of the game.

STRATEGIES

Fundamental to game theory is the notion of a *strategy*. A strategy is a set of instructions that a player could give to a friend or program on a computer so that the friend or computer could play the game on her behalf. Generally, strategies are contingent responses: in the game of chess, for example, a strategy should specify how to play for every possible arrangement of pieces on the board.

An alternative to the extensive form of game description is the *normal* or *strategic* form. This is less detailed than the extensive form, specifying only the list of strategies available to each player. Since the strategies specify how each player is to play in each circumstance, we can work out from the *strategy profile* specifying each player's strategy what payoff is received by each player. This map from strategy profiles to payoffs is called the normal or strategic form. It is perhaps the most familiar form of a game, and is frequently given in the form of a game matrix.

The matrix shown in Figure 1 is the celebrated *Prisoner's Dilemma* game. In this game the two players are partners in a crime who have been captured by the police. Each suspect is placed in a separate cell, and offered the opportunity to confess to the crime. The rows of the matrix correspond to strategies of the first player; the columns are strategies of the second player. The numbers in the matrix are the payoffs: the first number in each pair is the payoff to the first player, the second the payoff to the second player. Notice that the total payoff to both players is highest if neither confesses, so each receives 5. However, game theory predicts that this will not be the outcome of the game (hence the dilemma). Each player reasons as follows: if the other player does not confess, it is best for me to confess (9 instead of 5). If the other player does confess, it is also best for me to confess (1 instead of 0). So no matter what I think the other player will do, it is best to confess. The theory predicts, therefore, that each player following her own self-interest will result in confessions by both players.

	Player 2		
Player 1		not confess	confess
	not confess	5,5	0,9
	confess	9,0	1,1

Figure 1. Game matrix: 'The Prisoner's Dilemma'.

EQUILIBRIUM

The previous example illustrates the central concept in game theory, that of an *equilibrium*. This is an example of a *dominant strategy* equilibrium: the incentive of each player to confess does not depend on how the other player plays. Dominant strategy is the most persuasive notion of equilibrium known to game theorists. In the experimental laboratory, however, players who play the prisoner's dilemma sometimes cooperate. The view of game theorists is that this does not contradict the theory, so much as reflect the fact that players in the laboratory have concerns besides monetary payoffs. An important current topic of research in game theory is the study of the relationship between monetary payoffs and the *utility* payoffs that reflect players' real incentives for making decisions.

By way of contrast to the prisoner's dilemma, consider the game matrix in Figure 2. This is known as the *Battle of the Sexes* game. The story goes that a husband and wife must agree on how to spend the evening. The husband (player 1) prefers to go to the ballgame (2 instead of 1), and the wife (player 2) to the opera (also 2 instead of 1). However, they prefer agreement to disagreement, so if they disagree both get 0. This game does not admit a dominant strategy equilibrium. If the husband thinks the wife's strategy is to choose the opera, his *best response* is to choose the opera rather than the ballgame (1 instead of 0). Conversely, if he thinks the wife's strategy is to choose the ballgame, his best response is the ballgame (2 instead of 0). While in the prisoner's dilemma the best response does not depend on what the other player is thought to be doing, in the battle of the sexes the best response depends entirely on what the other player is thought to be doing. This is sometimes called a *coordination game* to reflect the fact that each player wants to coordinate with the other player.

For games without dominant strategies the equilibrium notion most widely used by game theorists is that of *Nash equilibrium*. In a Nash equilibrium, each player plays a best response, and correctly anticipates that her opponent will do the same. The battle of the sexes game has two Nash equilibria: both go to the opera, or both go to the ballgame: if each expects the other to go to the opera (ballgame) the best response is to go to the opera (ballgame). By way of contrast, one going to the opera and one to the ballgame is not a Nash equilibrium: since each correctly anticipates that the other is doing the opposite, neither one is playing a best response.

Games with more than one equilibrium pose a dilemma for game theory: how do we or the players know which equilibrium to choose? This question has been a focal point for research in game theory since its inception. Modern theorists incline to the view that equilibrium is arrived at through learning: people have many opportunities to play various games, and through experience learn which is the 'right' equilibrium.

MIXED STRATEGIES

While the battle of the sexes has too many equilibria, what about the game in Figure 3? You may recognize this game as the *Matching Pennies* game. There is, however, a more colorful rendition from Conan Doyle's Sherlock Holmes story *The Last Problem*. Moriarty (player 2) is pursuing Holmes (player 1) by train in order to kill Holmes and save himself. The train stops at Canterbury on the way to Paris. If both get off at Canterbury, Moriarty catches Holmes and wins the game (−1 for Holmes, 1 for Moriarty). Similarly, if both get off at Paris. Conversely, if they get off at different places, Holmes escapes (1 for Holmes and −1 for Moriarty). This is an example of a zero sum game: one player's loss is another player's gain. In the story, Holmes gets off at Canterbury, while Moriarty continues on to Paris. But it is easy to see that this is not a Nash equilibrium: Moriarty should have anticipated that Holmes would get off at Canterbury, and so his best response was to get off also at Canterbury. As Holmes says: 'There are limits, you see, to our friend's intelligence. It would have been a coup-de-maître had he deduced what I would deduce and acted accordingly.' However,

	Player 2		
Player 1		opera	ballgame
	opera	1,2	0,0
	ballgame	0,0	2,1

Figure 2. Game matrix: 'The Battle of the Sexes'.

	Player 2		
Player 1		Canterbury	Paris
	Canterbury	–1,1	1,–1
	Paris	1,–1	–1,1

Figure 3. Game matrix: 'Matching Pennies'.

this game does not have *any* Nash equilibrium: whichever player loses should anticipate losing, and so choose a different strategy.

What do game theorists make of a game without a Nash equilibrium? The answer is that there are more ways to play the game than are represented in the matrix. Instead of simply choosing Canterbury or Paris, a player can flip a coin to decide what to do. This is an example of a random or *mixed strategy*, which simply means a particular way of choosing randomly among the different strategies. It is a mathematical fact, although not an easy one to prove, that every game with a finite number of players and a finite number of strategies has at least one mixed strategy Nash equilibrium. The mixed strategy equilibrium of the matching pennies game is well known: each player should randomize 50–50 between the two alternatives. If Moriarty randomizes 50–50 between Canterbury and Paris, then Holmes has a 50 percent chance of winning and a 50 percent chance of losing, regardless of whether he chooses to get off at Canterbury or at Paris. Since he is indifferent between the two choices, he does not mind flipping a coin to decide between the two, and so there is no better choice than for him to randomize 50–50 himself. Similarly, when Holmes is randomizing 50–50, there is no better choice for Moriarty than to do the same. Each player, correctly anticipating that his opponent will randomize 50–50, can do no better than to do the same. So perhaps Holmes (or Conan Doyle) is not such a clever game theorist after all.

Mixed strategy equilibrium points out an aspect of Nash equilibrium that is often confusing for beginners. Nash equilibrium does not require a positive reason for playing the equilibrium strategy. In matching pennies, Holmes and Moriarty are indifferent: they have no positive reason to randomize 50–50 rather than doing something else. However, it is only an equilibrium if they both happen to randomize 50–50. The central thing to keep in mind is that Nash equilibrium does not attempt to explain why players play the way they do. It merely proposes a way of playing so that no player would have an incentive to play differently. As with the issue of multiple equilibria, theories that provide a positive reason for players to be at equilibrium have been one of the staples of game theory research, and the notion of players learning over time has played a central role in this research.

Further Reading

Bierman H and Fernandez L (1993) *Game Theory with Economic Applications*. Reading, MA: Addison-Wesley.

Binmore K (1992) *Fun and Games: A Text on Game Theory*. Lexington, MA: DC Heath.

Dixit AK and Nalebuff B (1991) *Thinking Strategically*. New York, NY: Norton.

Dixit AK and Skeath S (1999) *Games of Strategy*. New York, NY: Norton.

Fudenberg D and Levine DK (1998) *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Fudenberg D and Tirole J (1991) *Game Theory*. Cambridge, MA: MIT Press.

Kreps D (1990) *A Course in Microeconomic Theory*. Princeton, NJ: Princeton University Press.

Luce R and Raiffa H (1957) *Games and Decisions*. New York, NY: John Wiley.

Myerson R (1991) *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Osborne MJ and Rubinstein A (1994) *A Course in Game Theory*. Cambridge, MA: MIT Press.

Gamma Oscillations in Humans

Intermediate article

Catherine Tallon-Baudry, Centre National de la Recherche Scientifique (CNRS), Paris, France
Olivier Bertrand, INSERM, Lyon, France

CONTENTS

Gamma oscillations in human electroencephalographic signals are defined by their frequency range, from 20 Hz up to 100 Hz. Different types of oscillations are related to arousal states, early sensory processes, and neural mechanisms of sensory or cognitive binding by oscillatory synchronization.

INTRODUCTION

The existence of oscillations in the gamma range (20–100 Hz) in human electroencephalographic (EEG) or magnetoencephalographic (MEG) signals was first reported in the 1950s. In the 1990s, interest in these high-frequency activities increased following the proposal that frequency tagging could solve the feature-binding problem (Gray *et al.*, 1989). This problem becomes obvious if one considers the situation when the visual system is confronted with two objects – e.g. a green bar moving upwards and a red one moving downwards. How is the information 'green', coded by one set of neurons, combined with the direction of motion 'up' encoded in a different neural set? How can the system correctly assign the correct color and motion attributes to an object? A possible solution to this problem is that neurons responding to different features of the same object synchronize their discharge on an oscillatory mode. Using different frequencies, distinct cell assemblies could coexist to code simultaneously for different objects (Singer and Gray, 1995). More generally, oscillatory synchrony could provide the necessary link between the multiple areas that are activated in any task, as exemplified in numerous brain imaging studies in humans. It would thus not be dedicated to a particular cognitive process, but could rather be considered as a general neural mechanism that binds together the sensory and cognitive properties of any perceived or recalled object into the experienced entity. (*See* **Binding Problem**; **Neural Oscillations**; **Neural Correlates of Consciousness as State and Trait**; **Neural Correlates of Visual Consciousness**; **Perception, Gestalt Principles of**)

However, very different types of activities having in common a peak in their power spectrum between 20 Hz and 100 Hz have been gathered under the same general appellation of 'gamma oscillations'. They have been studied in various experimental contexts and probably reflect very different mechanisms. So far, only one type of gamma activity ('induced' oscillations) has been repeatedly related to the binding problem.

TYPES OF GAMMA OSCILLATIONS

The different types of gamma oscillations listed below have only one common property: they reflect the synchronization of large neural populations in the gamma frequency range (20–100 Hz). Indeed, scalp EEG or MEG signals always correspond to the collective behavior of neural ensembles. However, the precise anatomical extent of a synchronous activity measured at the scalp level is not readily accessible. (*See* **Electroencephalography (EEG)**)

Spontaneous Gamma Activity and Arousal

The frequency content of the EEG during the sleep–wake cycle changes markedly and indeed is used as a criteria to distinguish between sleep stages. During wakefulness and rapid eye movement (REM) sleep, the EEG power spectrum is shifted towards higher frequencies. In keeping with this classical observation, randomly occurring, spontaneous bursts of 40 Hz activity were shown to appear in the awake and REM sleep states, but not during slow-wave sleep (Llinas and Ribary, 1993). This 'desynchronized' EEG in wakefulness and REM sleep corresponds to a global state of

arousal of the whole brain. It can thus be observed at any recording site, as opposed to gamma responses obtained in response to a stimulus, which show up at the scalp level with a modality-dependent topographical distribution.

The Phase-locked 40 Hz Response: Evoked and Steady-state Responses

Transient averaged evoked potentials (EPs) are obtained by averaging the EEG or MEG signals obtained in response to a stimulus in a series of trials (Figure 1a). Any signal that does not have a strict time relationship with stimulus onset is considered to be 'noise' and tends to disappear in the average. Evoked responses, as they appear in the EPs, are in contrast strictly phase-locked to stimulus onset; they appear in each single trial at the same latency.

Galambos *et al*. (1981) were the first to notice that the series of waves appearing in the first 100 ms following the delivery of an acoustic click could be considered as a sequence of positive and negative peaks separated by approximately 25 ms, thus resembling a sinusoid wave at 40 Hz. This observation led to the definition of the early evoked 40 Hz response, which has been repeatedly observed in both the auditory and visual modalities, in EEG and MEG signals. Like many other EP components, the evoked 40 Hz response is sensitive to the physical parameters of the stimulus, and can be modulated by attention. No conclusive evidence is available yet regarding the nature of this evoked 40 Hz response: it could be a truly oscillatory phenomenon, or reflect a succession of activations of different neural structures that appear by chance at latencies separated by 25 ms at the scalp level. Whether the early gamma response is distinct from the evoked potentials appearing in the same latency range remains unclear.

Steady-state evoked potentials are obtained in response to stimuli presented at a high-frequency rate. They were first observed in the visual modality (Regan, 1968). When a stimulus is presented at frequencies of 30–60 Hz, the evoked response is a nearly pure sinusoid at the stimulus frequency. Large steady-state responses are obtained with a stimulus frequency between 40 Hz and 50 Hz, in both visual and auditory modalities. This finding was first interpreted as the existence of natural resonance frequencies in the brain at around 40 Hz. However, the current prevailing view is that steady-state potentials simply reflect a linear overlap of the early evoked responses to consecutive stimuli (Figure 1c). The steady-state response is a large and stable signal; it can be acquired in a single recording session and readily analyzed in the frequency domain at the driving stimulus frequency. It is therefore considered to be a useful indicator of the functioning of early sensory processes, with applications in both clinical and cognitive research.

Gamma Activities Induced by a Stimulus

As opposed to transient or steady-state evoked responses, induced activities are characterized by a loose temporal relationship to the stimulus (Figure 1a). Induced gamma responses received much attention following the observation of gamma oscillatory synchronization in animals, not phase-locked to the stimulus (Gray *et al*., 1989). Since an induced activity appears with a jitter in latency from one trial to the other, it tends to disappear in the averaged evoked potential. Such activities thus require specific methods to be extracted from EEG or MEG signals. All the methods used so far have in common the transformation of the potential or magnetic field into spectral power prior averaging (Figure 1b). Different techniques, with different time and frequency resolutions, have been successfully applied, such as filtering and rectifying, short-term Fourier transform, or wavelet-based time–frequency decomposition. Using such methods, any noise present in the data (as produced by power supply or screen radiations) tends to show up in the final power average. Other artefacts have physiological sources: muscle activity is picked up in the EEG and can show up in the gamma frequency range. Particular care has thus to be taken to obtain high-quality data. Although a growing number of research groups have been able to record these induced gamma activities, some authors still deny their existence at the scalp level.

INDUCED GAMMA ACTIVITIES

Existence and Characteristics

Induced gamma activities have been observed in different sensory modalities (auditory, somatosensory, visual), before and during movement execution, in various experimental conditions, ranging from a simple auditory detection task to semantic decision tasks. In response to a transient stimulus, an induced gamma-band response usually appears 200–400 ms after stimulus onset. Its frequency is most often between 30 Hz and 40 Hz but it may

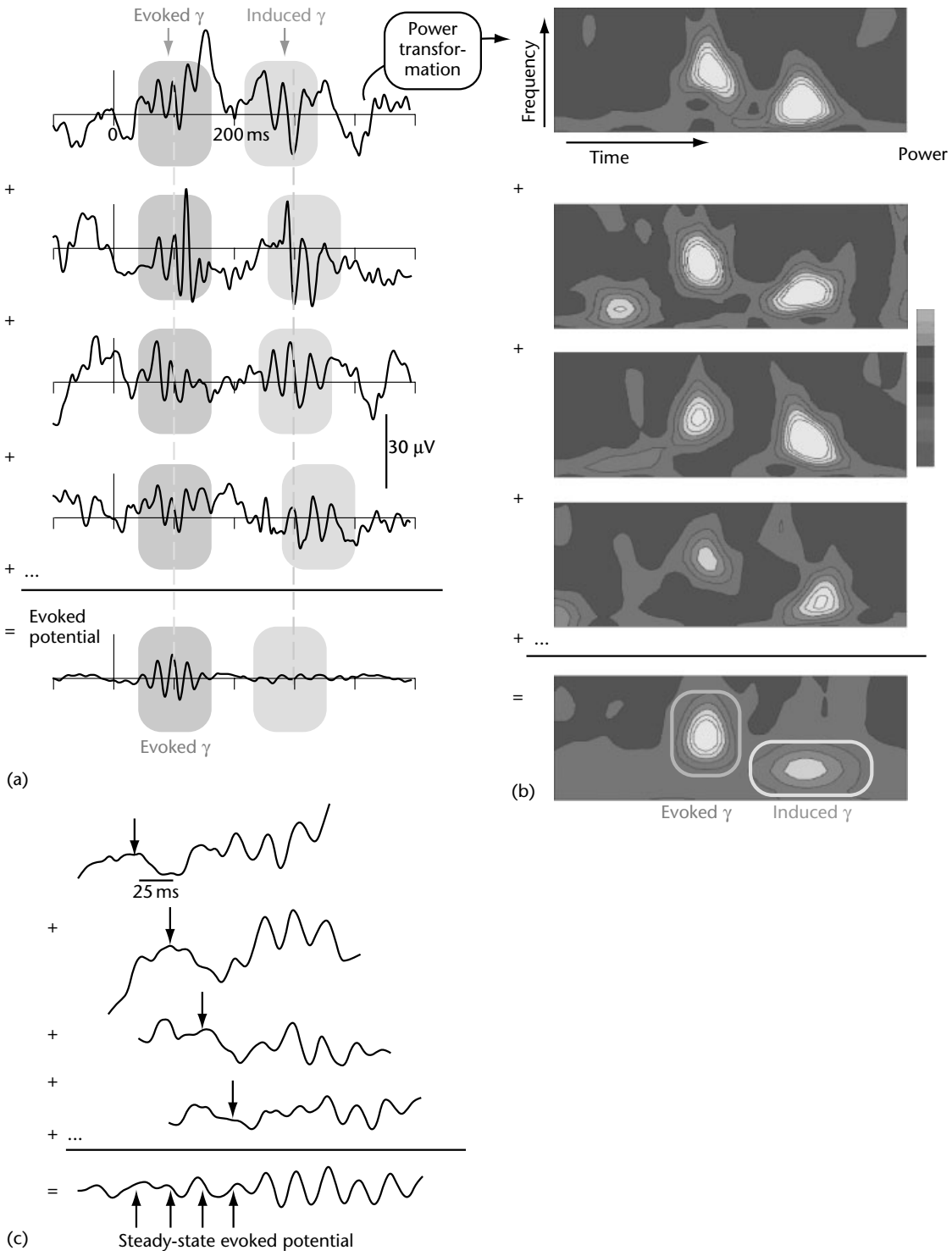


Figure 1. [*Figure is also reproduced in color section.*] The different types of gamma oscillations observed in humans (simulated data). (a) Two bursts of oscillatory activity (shaded) appear in successive single trials obtained in response to a stimulus (stimulus onset at zero latency). The first is evoked: it appears always at the same latency, and thus shows up in the averaged evoked potential (bottom). The second is induced by the stimulus: it appears with a jitter in latency and thus tends to cancel out in the average. (b) Transformation of single trials into power prior averaging. The method depicted here is a wavelet-based time–frequency decomposition of the signal. Both evoked and induced activities show up in the average of these time–frequency plots across single trials (bottom). (c) The steady-state evoked potential could reflect the linear combination of successive early sensory responses, when the stimulus (arrow) is delivered at 40 Hz (Galambos *et al*., 1981).

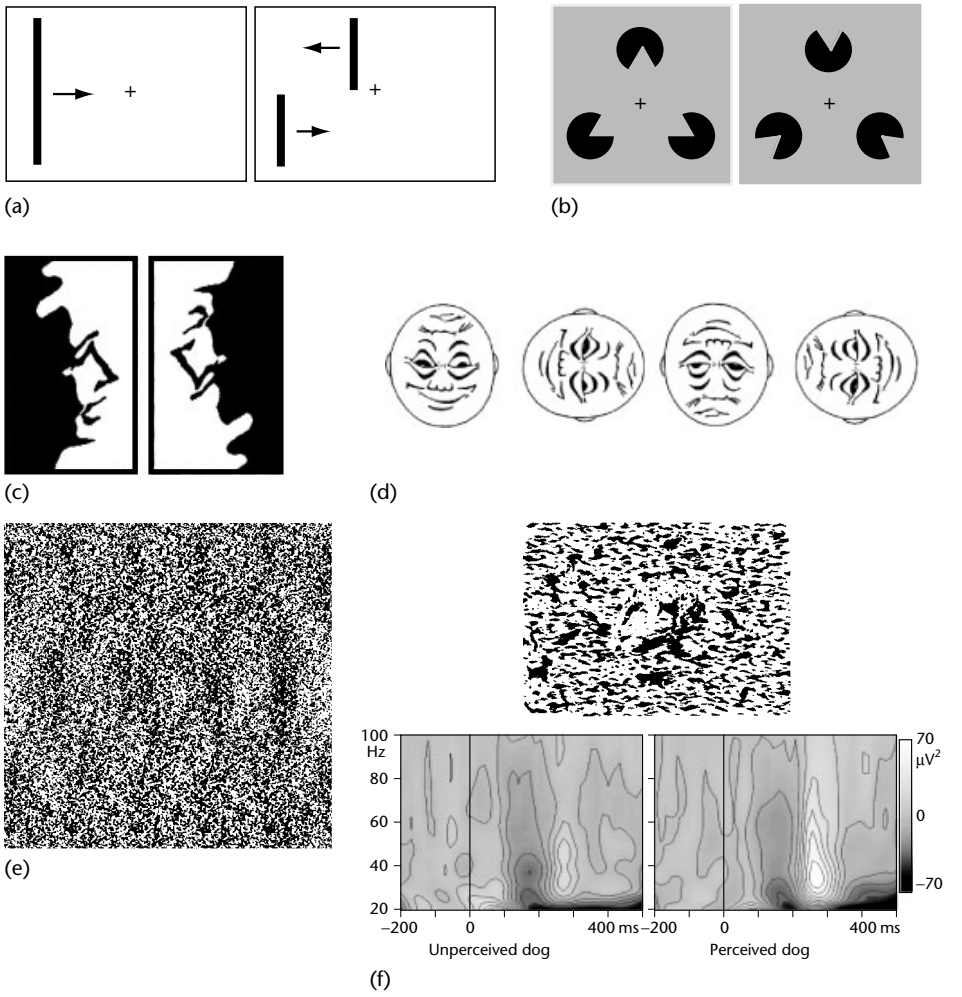


Figure 2. Coherent stimuli giving rise to an occipital increase of induced gamma oscillations. (a,b,c) Coherent object (left) and control, 'non-object' stimulus (right). Multiple objects can sometimes be perceived from the control stimulus. However, if multiple oscillatory assemblies, corresponding to each of these objects, coexist at different frequencies, they should tend to cancel each other in the resulting signal recorded at the scalp level. (d) A continuously rotating stimulus can be perceived as a happy or sad face (vertical orientation). Peaks in gamma activity are observed at these positions only. (e) An occipital increase of gamma oscillations is observed prior the fusion of this autostereogram into a three-dimensional percept. (f) Modified version of the Dalmatian dog picture. Naive observers perceive this stimulus as meaningless blobs on a gray background, and have only a very weak induced gamma response; once trained to perceive the dog hidden in this picture, however, they show a prominent occipital gamma activity. Reproduced with permission from the following sources: (a) Müller MM, BoschJ, Elbert T *et al*. (1996) Visually induced gamma-based responses in human electroencephalographic activity – a link to animal studies. *Experimental Brain Research* **112**: 96–102. (Copyright 1996 Springer-Verlag.) (b) Tallon-Baudry C, Bertrand O, Delpuech C and Pernier J (1996) Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *Journal of Neuroscience* **16**: 4240–4249. (Copyright 1996 Society for Neuroscience.) (c) Rodriguez *et al*. (1999). (d) Keil A, Müller MM, Ray WJ, Gruber T and Elbert T (1999) Human gamma band activity and perception of a Gestalt. *Journal of Neuroscience* **19**(16): 7152–7161. (Copyright 1999 Society for Neuroscience.) (e) Revonsuo A, Wilenius-Emet M, Kuusela J and Lehto M (1997) The neural generation of a unified illusion in human vision. *NeuroReport* **8**: 3867–3870. (f) Tallon-Baudry C, Bertrand O, Delpuech C and Pernier J (1997) Oscillatory gamma-band (30–70 Hz) activity induced by a visual search task in humans. *Journal of Neuroscience* **17**: 722–734 (Copyright 1997 Society for Neuroscience).

vary from 25 Hz up to 90 Hz. The topography of induced gamma activities is both modality- and task-dependent. This suggests that induced gamma responses reflect the oscillatory synchronization of stimulus- and task-related areas, rather than an unspecific arousal process. For instance, the induced gamma activities that were observed prior to voluntary movement have a scalp distribution

that follows the somatotopic organization of the motor cortex (Pfurtscheller *et al*., 1994). In many of these studies, a relevance of induced gamma activities for sensory or cognitive processes is suggested. However, a more precise functional role of these induced gamma oscillations emerges from studies in the visual modality.

Induced Gamma Oscillations and Coherent Visual Representations

In a series of EEG studies from different groups, it was shown that when the stimulus leads to a coherent percept, a transient increase in the gamma band is observed at occipital leads, compared with meaningless stimuli (Figure 2). In addition, in several of these studies, the coherence of the stimulus was not reflected directly in the classical EPs, nor in the the alpha band (8–12 Hz). Last, variations of gamma power at particular recording sites can be accompanied by episodes of gamma synchronization between electrodes, suggesting that areas wide apart can become synchronized (Rodriguez *et al*., 1999). Altogether, these results provide arguments in favor of a specific role of induced neural synchronization in the gamma range in the integration of visual information into a coherent object representation. Because induced gamma oscillations and evoked potentials in the same latency range do not show the same functional variations, it is likely that these two signals reflect different neural mechanisms.

In addition, induced gamma oscillations also increase when the object representation is not directly triggered by the physical characteristics of the stimulus. A transient increase in gamma power appears when a person has to activate an internal representation of the target to detect it in a picture where it is hidden (Figure 2). Sustained gamma and beta (15–20 Hz) oscillations have also been shown to be present during the rehearsal of an object in short-term memory, both at occipital and frontal sites (Tallon-Baudry *et al*., 1998). This suggests a role of oscillatory neural synchrony in bringing together pieces of information, probably encoded in distinct functional areas, into the meaningful and coherent representation that is perceptually experienced or rehearsed in short-term memory. (*See* **Hebbian Cell Assemblies**; **Neural Development, Models of**)

Finally, following Hebb's concept of cell assembly, neural synchronization in the gamma range could be involved in associative learning: Miltner *et al*. (1999) observed an increase of coherence in the gamma range between occipital and central scalp electrodes in a task involving a conditioned association between a visual and a somatosensory stimulus. (*See* **Hebbian Cell Assemblies**; **Hebb, Donald Olding**)

CONCLUSION

Despite a rapidly growing amount of research, this field of investigation remains controversial. Scalp-recorded gamma oscillations are still sometimes considered to be artefacts. Indeed, EEG signals may be contaminated by muscle activity. However, intracerebral recordings in epileptic patients indisputably show the existence of both within-area and between-area oscillatory synchrony (Fell *et al*., 2001; Tallon-Baudry *et al*., 2001). Finally, the functional interpretation of gamma oscillations suffers from the frequent confusion between the different types of gamma responses.

However, there is converging evidence that induced gamma oscillations are involved in sensory processes, memory rehearsal, visuomotor transformation and learning. This could mean that they are a mere epiphenomenon accompanying any mental activity. Conversely, the variety of experimental situations in which they are encountered suggest that they cannot be correlated with a unique cognitive process, but rather reflect a general neural mechanism used for dynamic binding of activities from the different functional areas activated in sensory and cognitive tasks. This interpretation is in keeping with the theoretical hypothesis of cell assembly synchronization, since scalp-recorded oscillations necessarily reflect a massive neural synchrony. An alternative interpretation is that oscillatory synchrony may serve as an attentional gating mechanism (Fries *et al*., 2001). (*See* **Attention, Neural Basis of**)

To confirm the functional role of induced gamma oscillations, it would be necessary to show that a selective disruption of gamma activity modifies perception or behavior. This could probably be achieved only in studies of freely moving animals. However, it is not clear yet how to relate the gamma oscillations recorded at the scalp level in humans to those observed at a microscopic scale in animal studies.

References

Fell J, Kläver P, Lehnertz K *et al*. (2001) Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. *Nature Neuroscience* **4**: 1259–1264.

Fries P, Reynolds JH, Rorie AE and Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**: 1560–1563.

Galambos RS, Makeig S and Talmachoff PJ (1981) A 40-Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of Sciences USA* **78**: 2643–2647.

Gray CM, König P, Engel AK and Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**: 334–337.

Llinas R and Ribary U (1993) Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences USA* **90**: 2078–2081.

Miltner WHR, Braun C, Arnold M, Witte H and Taub E (1999) Coherence of gamma-band EEG activity as a basis for associative learning. *Nature* **397**: 434–436.

Pfurtscheller G, Flotzinger D and Neuper C (1994) Differentiation between finger, toe and tongue movement in man based on 40 Hz EEG. *Electroencephalography and Clinical Neurophysiology* **90**: 456–460.

Regan D (1968) A high-frequency mechanism which underlies visual evoked potentials. *Electroencephalography and Clinical Neurophysiology* **25**: 231–237.

Rodriguez E, George N, Lachaux JP, Martinerie J, Renault B and Varela FJ (1999) Perception's shadow: long-distance synchronization of human brain activity. *Nature* **397**: 430–433.

Singer W and Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–586.

Tallon-Baudry C, Bertrand O, Peronnet F and Pernier J (1998) Induced gamma-band activity during the delay of a visual short-term memory task in humans. *Journal of Neuroscience* **18**: 4244–4254.

Tallon-Baudry C, Bertrand O and Fischer C (2001) Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance. *Journal of Neuroscience* **21**: 1–5.

Further Reading

Chatrian GE, Bickford RG and Uihlein A (1960) Depth electrographic study of a fast rhythm evoked from the human calcarine region by steady illumination. *Electroencephalography and Clinical Neurophysiology* **12**: 167–176.

Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2**: 263–275.

Ghose GM and Maunsell J (1999) Specialized representations in visual cortex: a role for binding? *Neuron* **24**: 79–85.

Milner PM (1996) Neural representations: some old problems revisited. *Journal of Cognitive Neuroscience* **8**: 69–77.

Singer W (1999) Neuronal synchrony: a versatile code for the definition of relations? *Neuron* **24**: 49–65.

Tallon-Baudry C and Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences* **3**: 151–162.

Treisman AM (1996) The binding problem. *Current Opinion in Neurobiology* **6**: 171–178.

Gender Differences in Cognition and Educational Performance

Introductory article

Marcia C Linn, University of California, Berkeley, California, USA
Cathy Kessel, University of California, Berkeley, California, USA

CONTENTS

Gender differences in cognitive and educational performance refer to the apparent differences in the cognitive abilities of men and women, especially with respect to spatial and verbal tasks. They have been ascribed to both cultural and biological influences.

HISTORICAL AND CULTURAL INFLUENCES

Meaningful interpretations of gender differences in cognitive and educational performance require a knowledge of the relevant cultural, societal, and historical contexts. In the United States, as in other countries, women's participation in education has indeed been a function of culture and history, as well as economic need and social justice. Women began to gain access to selected institutions of higher education by the late 1800s, although many still believed that study might overtax females by drawing blood away from their ovaries to their brains. Research discredited this belief, but it resonated with popular opinion, patterns of participation in intellectual pursuits, and the images of intellectual women as spinsters. When women started to enroll as undergraduate and graduate students, they tended to earn higher grades than men – a trend that has continued to the present.

Almost a century later, when Title IX came into effect in the United States, most colleges lifted their quotas for women. Now almost every institution of higher education admits both men and women; few professional programs retain gender quotas. As a result, over half of undergraduates are women, and graduation rates of women from high school and college have soared and surpassed those of men. Females comprise an increasing proportion of students in professional schools including medicine, law, and business. National Science Foundation statistics show that almost half of all PhDs are earned by women, and over a third of the doctorates in science and engineering go to women.

The picture is less rosy with regard to employment. Cultural expectations concerning who can succeed, especially in positions of leadership, impact on selection decisions. However, exceptions are widespread. Men have excelled in fields traditionally viewed as feminine, including nursing and office administration, and women have succeeded in endeavors traditionally viewed as male, including conducting orchestras, leading technology companies, and winning intellectual prizes. Nevertheless, progress in these areas has been slow for nontraditional participants, and, in some cases such as computer science, initial gains have subsided.

The subtle but powerful pressures of cultural expectations, giving the work of females less value, are apparent, for example, even at the prestigious Massachusetts Institute of Technology. Female faculty jointly assessed the space they were given, the treatment of their requests for research funds, the assignment of courses, and the assignment of committee responsibilities. Comparing practices across departments revealed systematic and substantial differences in treatment for female compared with male faculty. MIT publicized the discrepancies and remedies in an effort to increase equitable opportunities for women at all institutions. Despite this publicity most other institutions have resisted similar analyses and possible remedies.

Individuals in power may be slow to welcome those who do not resemble themselves. As described, for example by Virginia Valian, individuals who are different from their peers frequently experience difficulties in succeeding and often are more closely scrutinized than their more traditional

peers. These conditions are exacerbated by expectations that tend to place greater responsibility for family and children on females than on males, reducing available female human capital.

Cultural expectations may affect performance as well as selection. The psychological research of Claude Steele and his colleagues has demonstrated how context may make these expectations salient for particular groups. When students of a subgroup that traditionally performs poorly in a given domain are reminded of this fact, they tend to perform poorly on difficult tests; without this reminder, they perform similarly to their peers. These and other findings suggest that cultural expectations permeate all social interactions, including those in research laboratory settings, and may influence outcomes of experimental and classroom studies in subtle and unexpected ways.

Cultural expectations may affect educational participation and performance as well. For example, in 1980 Geoffrey Driver compared the academic performances of students who were white and those of West Indian origin. He found that although they experienced the same general academic program, when tested at age 16, West Indians tended to outperform their English counterparts, English boys tended to outperform English girls – but West Indian girls tended to outperform West Indian boys. A possible explanation is that the West Indian girls assumed a family structure more prevalent in the West Indies in which women are responsible for family subsistence – and were preparing for their futures as primary wage-earners.

Cultural expectations may explain differences in participation found by a Public Policy of California study. In California, middle school girls' course enrollment was greater than boys' for English, foreign language, mathematics, science (except that required for college), social science, but not for computer science courses. The situation was similar for college preparatory high school courses. Female enrollment was greater than that of males – again, except for computer science. This enrollment pattern reverses the earlier gender gap in mathematics and science courses. Moreover, in each ethnic group in the study categories, girls enrolled in more mathematics and science courses than boys. However, gaps were far larger for Filipinos, Hispanics, and African Americans than for whites and non-Filipino Asians.

Cultural expectations and instructional practices may explain the findings of David Byrnes and his colleagues. In contrast with long-standing experience in the United States, no gender gap was found when the SAT was administered to a sample of Chinese high school students. Instructional practices, but perhaps not cultural expectations, may be responsible for a similar result for high school students enrolled in an innovative mathematics course in the United States.

In the past, cultural interpretations of gender differences often followed a deficit model. From this perspective, male behavior was considered normative and desirable, and deviations from this norm were considered deficits. This model also influenced research. Studies looked for differences between males and females and accumulated less successful performances of females to paint a picture of females as having 'deficits'. This deficit model was sometimes elaborated into a compensation model, in which typical male behavior was the norm in 'male' activities, including academic, intellectual, and athletic performance, and female behavior was the norm in activities typed as female, including child care, family relations, and homemaking. These models drew attention away from the large overlap in the distributions of male and female performance in every area, and neglected the rapid improvements that result when barriers to participation or success disappear.

SPATIAL AND VERBAL ABILITIES

Historically, accounts and definitions of human abilities have varied substantially. In the United States at the beginning of the twentieth century, research views of human ability as unitary were reinforced by measures of general ability that yielded a single score. As researchers performed more detailed analyses, they identified more specific and often contextualized abilities. A substantial commercial enterprise in the design, sale, and scoring of tests for these abilities developed.

Some of this proliferation in abilities stemmed from newly available statistical techniques. As factor analysis developed, individual researchers used this technique to distinguish more and more abilities. At the same time, a research program concerning the nature of ability constructs took shape. The identification, statistical properties, and labeling of human ability constructs has continued as an active research area. Researchers seek to distinguish 'traits' that are innate or unresponsive to instruction from achievements that respond to instruction. Good methods for determining whether an ability is inherent or responsive to instruction have been difficult to develop. Sorting out the effects of cultural expectations on human abilities has proven problematic. When cultural

expectations limit opportunity to learn, constructs often appear innate. For example, many assumed there are innate differences in spatial ability only to discover that performance responds to instruction and varies across educational systems. Similar differences in mathematical attainment have narrowed, as Rosenthal observed, 'faster than the gene can travel'. A serious debate about the inherent nature of abilities plays out in each of the areas highlighted in this review.

Spatial Performance

Research on spatial ability has its origins in Francis Galton's work on imagery in the 1880s. Findings concerning gender differences soon followed. Galton found that 'scientific men as a class have feeble powers of visual representation. There is no doubt whatever on the latter point, however it may be accounted for'. G. Stanley Hall reported in his 1891 study, 'The contents of children's minds on entering school', that girls excelled in space concepts and boys in number. Systematic research on this topic began in the 1920s.

Tests of spatial reasoning such as Raven's progressive matrices were viewed as independent of verbal skill. However, using factor analysis to distinguish verbal and nonverbal constructs proved difficult. Raymond Cattell argued that fluid and crystallized ability, although distinct constructs, were correlated factors. Despite these difficulties, by World War II spatial ability measurement played an important role in assessing and assigning army recruits, and measures of spatial ability were used in other career decisions. Eventually, the belief that females had greater difficulty in spatial reasoning was associated with their underrepresentation in mathematics and science.

In 1974, Eleanor Maccoby and Carol Jacklin reviewed all the research on gender differences and concluded that, although performance across tests used to measure spatial ability varied considerably, in general males excelled in spatial tasks. These tests included Raven's progressive matrices, the embedded figures test, and a newly developed measure of speed in mental rotation.

In the late 1970s, Marcia Linn and Anne Petersen extended Maccoby and Jacklin's findings on spatial performance. They sought to distinguish the varied constructs captured under the broad category of spatial ability. They used meta-analysis to group tasks in categories with uniform characteristics and identified three constructs.

First, syntheses of performances on tasks such as embedded figures, paper folding, and Raven's progressive matrices that asked students to reason about figural rather than verbal information revealed virtually no gender differences.

Second, tasks measuring what is often referred to as the *water level task* revealed large gender differences. This task, originally used by Jean Piaget, asks students to predict the level of water in a tilted glass when given information about the water level in an upright glass. Many factors can influence performance, including whether the glass is square or curved, and whether the respondent believes that the glass has just been moved to a new location or has been resting in that new location for some time.

Third, tasks requiring students to mentally rotate two- and three-dimensional objects to a new orientation revealed large differences, primarily in speed. Most respondents obtained the right answer; males responded more rapidly. Some respondents took over twice as long as others, and the studies of strategies for mental rotation showed that some participants mentally rotated an entire object and chose an answer, while others rotated features and repeated the process until the answer was chosen. In these studies, women were more likely than men to use a feature matching approach, and therefore to take longer to respond to mental rotation items. Changing the instructions to encourage rapid responding reduces differences, supporting the idea that a strategy preference is a factor in these differences.

Subsequent research has shown that spatial ability, rather than being an enduring trait, appears quite responsive to instruction and, for instance, multiple strategies can be used to solve most spatial reasoning tasks. For example, Sherry Hsi and Alice Agogino studied spatial reasoning performance among first-year engineering students at a competitive university, a group likely to have been successful in high school mathematics classes. They found large initial differences between males and females on measures of spatial reasoning. Two Saturday morning voluntary training sessions reduced or eliminated these gender differences.

Moreover, in-depth interviews of successful engineers revealed that they rarely used the rotation of whole objects in their work. Instead, most reported that they were likely to rely on descriptive geometry, feature matching, and the rotation of a subset of familiar shapes and forms, rather than using their mental capacities to rotate whole objects. This suggests that success in engineering may not require the use of mental rotation.

Another area where spatial reasoning ability is frequently implicated concerns the interpretation

of maps. Here again, cultural expectations and anecdotal evidence suggest that males are more successful at way-finding, map reading, and map interpretation. Nevertheless, extensive study of young children's ability to find their way in complex settings reveals a fairly systematic result of males being more successful than females. A study of adults found that males were more successful at way-finding and, as with the engineering students, that females improved with training.

All these results underscore the importance of instruction promoting varied spatial abilities. Studies attempting to link spatial ability to activities requiring spatial reasoning such as sports and hobbies have primarily illustrated the complexity of the distinction between opportunity to learn and performance. Scrutiny of the US curriculum, however, reveals little opportunity to learn spatial reasoning. Until recently, elementary mathematics in the United States was almost synonymous with arithmetic. This is followed by a sudden jump to Euclidean geometry in high school and later (often in college) to the geometry concerned in calculus. Curriculum designers consciously chose to reduce or eliminate emphasis on spatial reasoning to be fair to all. This intriguing policy illustrates the role of culture in decision making. It appears that exactly the opposite decision would have been more sensible. By equalizing opportunity to learn, gender differences might have been reduced. Limited instruction and experience with spatial reasoning in the curriculum may exacerbate gender differences.

In summary, in the area of spatial performance, the definition of the construct 'spatial reasoning' has varied substantially. Research shows that the definition of the construct determines the existence and magnitude of gender differences in performance. Moreover, for tasks involving mental rotation where the largest gender differences are found, opportunity to learn as well as instructions to respond rapidly are both important factors in reducing or even eliminating gender differences in performance. Although the connection of the ability to rotate a whole object mentally with performance in mathematics, science, or engineering remains unclear, mental rotation continues to be an object of study.

Verbal Performance

Measures of ability, starting with the Binet measure of intelligence, have generally relied on verbal skills. The Binet test includes vocabulary items as well as comprehension and interpretation items. Tasks such as verbal analogies, opposites, and sentence completion and paragraph comprehension characterize a broad range of items and predominate in widely used tests, including the SAT verbal measure, the Graduate Record Examination verbal measure, and a multitude of achievement tests administered in schools.

Cultural expectations typically include the view that women excel in 'verbal ability', variously defined. This perception is consistent with established findings that on average girls learn to speak and to read earlier than boys. In addition, serious problems with reading are more common among boys than girls; however, individual reports of these difficulties may be affected by cultural expectations. Maccoby and Jacklin reported that differences in verbal ability were well established in their 1974 summary of gender differences. Subsequent meta-analyses conducted by Janet Hyde and Marcia Linn suggest that these differences are very small and vary by the selectivity of the sample as well as the discipline represented in the item.

Differential opportunities to develop specific forms of verbal ability influence performance on verbal measures. Vocabulary items that require the interpretation of word meanings without supplying a context often require students to rely far more on broad verbal experience and test-taking expertise than do more contextualized items such as sentence completion or verbal comprehension tasks. Furthermore, performance on verbal comprehension tasks that strongly rely on information from a specific discipline often display patterns of gender differences compatible with participation in courses from these fields. Because more males than females participate in engineering and physical science courses, passages with physics or engineering content tend to favor males. Similarly, because more females participate in humanities and social sciences, passages from these domains tend to favor females.

Another important consideration in analyzing gender differences in verbal ability concerns the selectivity of the sample. In general, differences in performance between males and females are relatively small when the general population is used but are amplified when selective populations are studied. A very large group of college-bound students take the SAT but this is sample is selective. An interesting pattern of results has emerged over the years in SAT performance. Early measures of verbal ability for this test tended to favor females. However, in the late 1980s, this measure started to favor males. This change in gender differences coincided with an increased reliance on science

passages in the comprehension section. Hyde and Linn demonstrated that when the selective SAT sample was considered, gender differences favoring males were significant, but when a representative group of students were used, these differences disappeared. One might hypothesize that among the selective sample the advantage of specific knowledge about the topics might have a bigger effect on overall differences between males and females. This effect, while still operative, would be very small in a population of individuals when most lacked specific knowledge of the topics.

Recent research calls for viewing verbal ability, verbal performance, and verbal communication skill as nested in the discipline and context where it is used. Ability to write news articles, poetry, short stories, novels, technical manuals, scientific papers, or law briefs depends far more on opportunity to learn, diligence, and motivation than on gender. Viewed through this lens, gender differences play a minor role in verbal performance. By far, the larger differences occur in the age of first speech.

BIOLOGICAL BASIS FOR ABILITY

Belief in a biological basis for gender differences in intellectual performance has motivated numerous research programs. Many studies have examined differences in distributions of ability for males and females. The notion that males are more variable dates back at least to Charles Darwin, who made this generalization from observations and studies of physical characteristics. Scientists sometimes assumed that mental traits were inherited, and would therefore show distributions similar to those of physical traits. Edward Thorndike used this supposition to argue that because 'men differ in intelligence and energy by wider extremes than do women', women should be not educated for professions that required giftedness rather than average ability.

Leta Hollingworth noted that psychologists had variable meanings for 'variability', including: a wider range in distribution of a trait, or the same range with greater frequency in the extremes. Moreover, her large-scale study of neonate measurements did not support any interpretation of 'greater male variability'. Subsequent large-scale studies have reinforced Hollingworth's view that variability between the genders reflects cultural rather than biological factors, and the hypothesis of greater variability has lost currency as an explanation of why 'leadership in the world's affairs … will inevitably belong oftener to men'.

Researchers now use a variety of new, and rapidly advancing methods, for studying the brains of males and females. These include non-invasive measurement of neurotransmitter levels, synaptic strength, number of synapses, and receptor levels as well as postmortem studies comparing the size of various brain areas. Studies often reveal gender differences but no consistent pattern has emerged and new methods often yield findings that contradict prior work. Both small sample sizes and difficulty associating brain functioning with complex performance limit the generalizability of current research. Recent findings offer some insight into the diagnosis and treatment of problems related to brain functioning. For example, when reading performance is measured, men show a higher rate of dyslexia. However, Sally Shaywitz and her colleagues have found that the incidence of dyslexia in brain function appears to be equally distributed by gender. To explain this apparent contradiction, Shaywitz reports evidence that women and men process language in different parts of their brains when reading (men use the right inferior frontal gyrus, women use the left gyrus as well).

ACADEMIC ACHIEVEMENT

As the discussions of spatial and verbal abilities suggest, the distinction between educational attainments in mathematics, science, language and other domains, and spatial and verbal ability has blurred over the past decade. However, it is still a long way from the laboratory to the classroom. Experimental research suggests constraints and draws attention to certain capabilities but does not often yield direct instructional implications. Educational accomplishments depend on opportunity to learn; cultural expectations have a profound impact on who participates and who persists in a variety of fields. Initially, access to education varied by gender, and today gender patterns in enrollment and persistence in courses remain.

The most profound differences in gender, access, and participation are in mathematics and science, where men predominate. In the past 20 years, fields like computer science have gained prominence and importance both in the economy and in academia. After an initial surge, to approximately one-third of the student body, the proportion of women has declined, and today women receive about 12 percent of the computer science PhDs in the United States.

Participation in the natural sciences varies both by field and by nationality. For example, in the United States, equal numbers of women and men

enroll in calculus courses and select mathematics as a college major. By graduate school, according to counts in 1999, women earned over one-third of the mathematics PhDs granted to US citizens by US universities. Since the United States educates a large percentage of graduate students from other countries, when all nationalities are considered only 28 percent of those receiving mathematics PhDs from US universities were women. In contrast to the situation for computer science, both the number and proportion of women receiving PhDs in mathematics has increased. Proportions differ considerably in other fields: women currently earn about 20 percent of PhDs in physics but 40 percent of those granted in biology.

Opportunity to learn effects vary by institution of higher education as well. For example, Marcia Linn and Cathy Kessel found that proportions of bachelor's degrees granted to women from the 'top ten' mathematics departments ranged from 9 to 47 percent. This variability is consistent with research done by Elaine Seymour and Nancy Hewitt. Their analysis of a national sample of 800 000 undergraduates showed that about two-thirds of students who enter college intending to study mathematics or statistics switch to another field. Moreover, 72 percent of the females and 60 percent of the males switched to other fields, but half of the males and two-thirds of the females switched to a field outside of mathematics, science, and technology.

Seymour and Hewitt's more in-depth study of 335 undergraduates in science and engineering at seven institutions suggests an explanation for why men and women switched fields – and moreover, that explanations differed for men and women. Although these undergraduates were considered 'highly qualified' (their SAT-M scores were at least 650), the primary concern of both switchers and persisters was poor teaching: courses were fast-paced, instructors were often unclear, unavailable, and uninspiring, even lecturing by reading from the textbook. Women more than men, however, were repelled by poor instruction. This study suggests that context of instruction rather than 'lack of ability' is a better explanation for students' lack of persistence in science and engineering fields. Possibly reflecting their status as nontraditional students, other motivations for participation – lack of acceptance by male peers, hence lack of opportunity to learn from peers outside of class – contributed to switching.

Other evidence suggests that women may be more often affected by poor instruction, but that improvements advantage all. For example, Marcia Linn and Michael Clancy demonstrated that new computer science courses tend to favor males, contrary to the usual experience of higher female grades. When courses are iteratively refined and improved, they tend to be equally successful in improving the learning and understanding of males and females. In general, when instructors use information from past courses to redesign their instruction to be more responsive to their students, individuals who have been at risk for failure, and groups that have traditionally performed less well, gain more than individuals who are traditional course participants. This finding resonates with the comments of switchers in the Seymour and Hewitt study and that students tend to persist more in nontraditional fields with effective instruction. The pattern of success and persistence also coincides with accounts of success of women who attended all-female colleges in the past.

CONCLUSION

The history of views of gender differences in intellectual ability has responded to changes in research methods. As methods for identifying and assessing gender differences have developed, new interpretations have emerged. Both more powerful tests and new methods for representing results, such as path analysis and meta-analysis, have influenced thinking. Analysis of aspects often categorized as 'noise' in early studies has strengthened the view that opportunity and inclination to learn play a powerful role in measured differences.

The nature of research findings depends on the methods used. The majority of psychological investigations of gender differences tests whether score differences for males and females significantly differ from chance. As the sample size increases and the power of the test increases, the likelihood of concluding that there are statistically significant differences also increases. Measures that tap aspects of broad constructs, such as verbal ability, spatial ability, and school achievement, often yield contradictory results. Synthesis techniques such as meta-analysis make the often unrealistic assumption that the studies replicate each other. In contrast, neurobiological studies with small numbers of women and men offer hints of results to come but are limited by sample size and by complex relationships between indicators and the potential constructs they measure.

In conclusion, the investigation of differences between males and females over the past decade reflects advances in methodological, conceptual, and cultural views of the field. Advances in the understanding of complex behavior have

undermined the once prevalent deficit model and strengthened the belief in opportunity to learn. Changing cultural views of who can succeed in intellectual endeavors have enhanced opportunity and increased the choices available to all students. Most importantly, research on gender differences in intellectual attainment reinforces the importance of effective educational programs and establishes the value of research designed to enhance the learning of students with disparate prior experiences and opportunities.

Acknowledgments

This material is based upon research supported by the National Science Foundation under grants REC 98-73160 and REC 98-05420. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors are grateful for input from members of the Web-based Integrated Science Environment project, as well as Susan Klein.

Further Reading

Caplan P, Crawford M, Hyde JS and Richardson JTE (eds) (1997) *Gender Differences in Human Cognition*. New York, NY: Oxford University Press.

Gould SJ (1981) *The Mismeasure of Man*. New York, NY: W. W. Norton.

Linn MC and Petersen A (1985) Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Development* **56**: 1479–1498.

Linn MC and Kessel C (1996) Success in mathematics: increasing talent and gender diversity. In: Schoenfeld A, Dubinsky E and Kaput J (eds) *Research in Collegiate Mathematics Education II*, pp. 101–144. Providence, RI: American Mathematical Society.

Maccoby E and Jacklin C (1974) *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.

National Center for Education Statistics (2000) *Trends in Educational Equity of Girls and Women* (NCES 2000–030). Washington, DC: US Department of Education, Office of Educational Research and Improvement. Available at: http://nces.ed.gov/spider/webspider/2000030.shtml

Rossiter M (1982; 1995) *Women Scientists in America*. Baltimore, MD: Johns Hopkins University Press.

Science (1994) **263**: 1345–1532. [Special issue: Women in science: Comparisons across cultures.]

Seymour E and Hewitt N (1997) *Talking About Leaving: Why Undergraduates Leave the Sciences*. Boulder, CO: Westview Press.

Shaywitz S (1996) Dyslexia. *Scientific American*. Available at http://www.sciam.com/1196issue/1196shaywitz.html

Shields SA (1982) The variability hypothesis: the history of a biological model of sex differences in intelligence. *Signs* **7**(4): 769–797.

Steele C (1997) A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist* **52**(6): 613–629.

Sternberg R (2000) The Holey Grail of General Intelligence. *Science* **289**: 399, 401.

Tavris C (1992) *Mismeasure of Woman*. New York, NY: Simon & Schuster.

Valian V (1998) *Why So Slow? The Advancement of Women*. Cambridge, MA: MIT Press.

Generalization

Introductory article

Peter J Urcuioli, Purdue University, West Lafayette, Indiana, USA

CONTENTS

Generalization is the finding that behavior explicitly conditioned to a stimulus also occurs without additional training to other, similar stimuli.

GENERALIZATION IN ANIMAL LEARNING

An adaptive psychological characteristic shared by humans and other animals is that behavior they explicitly learn to certain stimuli or in certain situations will also occur ('generalize') to other stimuli or situations resembling those involved in original learning. This produces considerable behavioral economy by forgoing the necessity of having to learn the same behavior to each and every member in a class of similar stimuli. Thus, learning to speak to certain individuals early in our lives (e.g., our parents) readily generalizes to other people. Likewise, an animal that learns that food is to be found among certain flowers or plants will later search with good effect in different areas containing similar flowers or plants.

Interest in stimulus generalization dates back to Pavlov's pioneering work on classical conditioning. He reported that dogs explicitly conditioned to salivate to a tactile stimulus applied to a certain area of their skin also salivated when that stimulus was applied to other skin areas. When generalization is studied in the context of operant conditioning as popularized by B. F. Skinner, animals are first trained to respond to obtain a reinforcer in the presence of some well-defined stimulus. For instance, a hungry pigeon might be taught to peck a key illuminated by certain wavelength in order to receive food. Afterwards, responding to other, similar stimuli is recorded – for example, pecking is measured to other wavelengths projected onto the key. The typical finding from such a generalization test is that the animal responds to these other ('test') stimuli, but that the frequency or probability with which it responds declines the more different a test stimulus is from the training stimulus. Plotting the frequency or probability of responding as a function of the test stimulus yields a generalization gradient. Typically, the gradient shows most responding occurring to the training stimulus and progressively less responding to stimuli whose characteristics are increasingly discrepant from those of the training stimulus.

Because operant behavior usually occurs in the presence of so-called discriminative stimuli, the generalization test provides a means to assess what features of those stimuli control that behavior. In short, generalization is a measure of stimulus control. Thus, when pigeons learn to peck a key of a certain color, we can determine whether color actually exerts stimulus control over pecking by varying the wavelength of light, and seeing whether or not the rate of pecking changes accordingly. Likewise, since the colored light also has a particular luminance, we can determine whether that characteristic controls pecking by holding wavelength constant and varying its intensity. This procedure of systematically varying some characteristic of a discriminative stimulus during the generalization test has certain advantages over other stimulus control assessments, such as simply removing the feature in question. For instance, if the test stimulus can't be seen (i.e. if its intensity is lowered to zero), observing a lack of responding tells us very little, if anything, about whether behavior is controlled by how bright the stimulus is when it can be seen. The generalization test thus provides relatively more discriminating and comprehensive information than other tests of stimulus control. Over the years, it has been used to demonstrate an impressive range of effective stimuli for animal behavior, from relatively simple stimuli such as color and tonal frequency, to extended stimuli such as the temporal duration of an event, and to relatively complex stimuli such as numerosity and 3-D visual orientation.

Methods for Obtaining and Plotting Generalization Gradients

There are two general methods for assessing generalization of responding around a stimulus signaling reinforcement (an S+). One method involves presenting a single test stimulus to each subject and testing different subjects with different stimuli. This between-subjects assessment obviously requires a large number of subjects in order to measure generalization along an entire stimulus continuum. Consequently, it has been less popular than an alternative, within-subject procedure developed in the 1950s. Here, each subject is exposed to all test stimuli during the generalization test. This procedure has the advantage of requiring fewer subjects to obtain the desired behavioral measures but the disadvantage of requiring a relatively long period of testing in order to obtain response measures to all stimuli in the test range.

With either method, the generalization test is usually conducted in extinction (that is, with responding not reinforced to any stimulus). Testing in extinction avoids possible contamination of the generalization measure that might occur if subjects were to learn (via reinforcement) to respond to every stimulus during the test itself. In addition, the procedure ensures a comparable assessment condition for *all* stimuli, including the formerly reinforced training stimulus. However, testing in extinction runs the risk that the animal will stop responding altogether, especially in a relatively long test, so the animal must be trained to persist despite prolonged periods of nonreinforcement. Persistence is conditioned during training by arranging intermittent reinforcement for responding. For instance, pigeons pecking a particular wavelength of light in training might receive food at variable times averaging only once or twice per minute. Such intermittent reinforcement is a standard feature of within-subject generalization studies. It establishes a steady rate of responding to the training stimulus and, more importantly, allows for a prolonged assessment of generalized responding.

An absolute generalization gradient plots the total number of responses made by a subject or a group of subjects to each test stimulus. The filled symbols in the left panel of Figure 1 show such a gradient as it would appear around a reinforced training (T) stimulus (an S+). Relative gradients, on the other hand, express responding to each test stimulus either as a proportion of responding to the training stimulus or as a proportion of total responding across all test stimuli. The latter measure is plotted in the right panel of Figure 1 using the data from the left panel. Relative gradients are desirable when there are large differences in overall responding across different training conditions or subjects that the researcher wishes to compare.

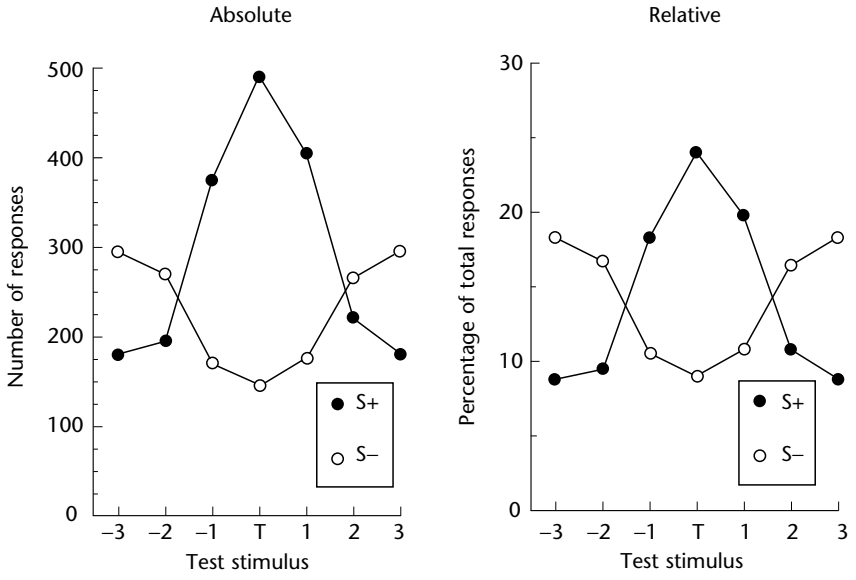


Figure 1. Absolute (left panel) and relative (right panel) generalization gradients around a training (T) stimulus that previously signaled reinforcement for responding (S+; filled symbols) or nonreinforcement for responding (S−; open symbols). Values for the test stimuli represent changes in equal increments above and below the training (T) value.

The Form of the Generalization Gradient

There is no single form to the generalization gradient. Its shape depends upon a variety of factors, such as the nature of training preceding the generalization test, how easy or difficult it is for subjects to distinguish each test stimulus from the training stimulus, whether or not the stimulus characteristics varied during the test do control the animal's behavior, and whether or not other stimuli are present during testing that also control responding.

Excitatory generalization gradients are those obtained by varying some aspect of a stimulus that has been established as a signal for reinforcement (an S+). These gradients generally appear as inverted U-shaped functions with the maximum level of responding at or near the S+, as illustrated by the filled-symbol gradients in Figure 1. Excitatory gradients are also called 'decremental gradients' because of the progressive decrease in responding to stimuli further and further removed from the S+.

Inhibitory generalization gradients are obtained by varying some aspect of a training stimulus that has been established as a signal for nonreinforcement (an S−). Assuming that responding has been concurrently established to an S+, the S− can be viewed as a stimulus that controls 'not responding'. After certain types of training procedures (described more fully later on), varying one of the characteristics of an S− during the generalization test yields a U-shaped gradient with a minimum at the S− value. In other words, an incremental gradient may appear, with the least amount of responding occurring to the stimulus signaling nonreinforcement in training and increasingly greater amounts of responding to test stimuli further and further removed from the S−. Absolute and relative inhibitory gradients are plotted by the open symbols in Figure 1.

Sometimes, the generalization gradient is flat: subjects respond with equal frequency or probability to all values along the test dimension. Flat gradients are difficult to interpret because they can arise for a variety of reasons. For instance, subjects may be perceptually insensitive to the stimulus feature varied during testing. Alternatively, the varied feature may not control behavior even though it is perceptually distinctive because the 'functional' stimulus, the feature that actually controls an animal's behavior, is not what the researcher varies during the generalization test. Finally, a flat gradient may occur despite variation in a functional stimulus because some other 'incidental' feature(s) common to all generalization test stimuli exerts greater control over responding than the varied feature.

An example of the latter situation occurs when pigeons are trained to peck a lighted key for food in the presence of an auditory S+ (e.g. a 1000-Hz tone). When generalization to different tone frequencies is subsequently measured, pigeons peck the lighted key at the same rate independently of the frequency of the tone accompanying it. This flat gradient does not mean that pigeons are deaf or that they cannot distinguish one frequency from another. Rather, the complete generalization of responding across different frequencies occurs because key pecking is controlled to a much greater degree by the lighted key, the stimulus to which pecking is directed. Since the key light accompanies every auditory stimulus presentation during training, it predicts the availability of food just as well as the tone. Moreover, it accompanies every auditory stimulus in the generalization test. Thus, the flat gradient simply reflects the fact that pecking is controlled more by the key light than by the tone.

Explaining Stimulus Generalization and Generalization Gradients

Pavlov believed that generalization from an explicitly trained stimulus to other, similar stimuli reflected a neurophysiological 'spread of effect'. He believed that the training stimulus established a cortical point or region of maximum neural activity and that this focal activity spread decrementally to adjacent cortical regions. Later, researchers using less physiologically oriented language suggested that generalized responding to a test stimulus reflected the degree to which the subject is unable to discriminate that stimulus from the training value. In other words, generalization was thought to be the inverse of discrimination: behavior generalizes from one stimulus to another to the extent that subjects do not distinguish between them. This proposition was not meant to suggest that subjects *could* not make such a differentiation if forced to do so. Rather, the argument was that they *did* not differentiate between certain stimuli in the generalization test.

Contemporary analyses of generalization have combined and refined these ideas into the notion that generalization reflects the number and the nature of the elements that the test stimuli share in common with the training stimulus. According to this view, the training stimulus is an amalgam of

many potentially conditionable elements. Thus, a pigeon that pecks a key of a particular color for food may be controlled in its behavior not only by the dominant wavelength of the S+ but also by its luminance, size, location, etc. Likewise, pecking a lighted key for food in the presence of a particular auditory frequency may establish control by both the tone and the key. During the generalization test, responding to each test stimulus is postulated to be a direct function of the number of conditioned elements it shares with the training stimulus. The greater the number of shared elements, the greater the amount of generalized responding. From this perspective, then, 'failure to discriminate' means that a test stimulus shares many conditioned elements, or a particularly potent element, with the training stimulus, thus generating considerable responding. Conversely, 'discriminating' means that a test stimulus shares few, if any, elements with the S+, thus yielding very little responding.

To illustrate this idea, consider that reinforced responding has been conditioned to an S+ consisting of elements [*a, b, c, d*], and that two generalization test stimuli consist of elements [*b, c, d, e*] and [*c, d, e, f*], respectively. Assuming that each element of the training stimulus gains control over responding, more responding should occur to the first test stimulus than to the second because the first has three elements in common with the training stimulus (viz. [*b, c, d*]) whereas the latter shares only two (viz. [*c, d*]). Similarly, if one element of the S+ is particularly potent – for example, [*a, b, c,* ***D***] – then any test stimulus containing '***D***', for example [*b, c,* ***D***, e] or [*c,* ***D***, *e, f*]), should generate considerable responding. Indeed, if every test stimulus contains a potent [***D***] element, then a flat generalization gradient would be a likely outcome.

DISCRIMINATION LEARNING AND STIMULUS CONTROL

In discrimination training, a stimulus in the presence of which responding is reinforced (an S+) is randomly alternated with another stimulus during which responding is either nonreinforced (an S−) or reinforced less often. In most studies of stimulus generalization, training involves at least an implicit discrimination. For example, pigeons trained to peck a lighted key for food usually encounter brief periods of time during which the key is dark. During these blank intervals separating successive presentations of the key light, food is unavailable (i.e. responding is nonreinforced).

Whether training involves implicit discrimination contingencies such as the periodic absence of S+ or, more commonly, explicit contingencies in which a different stimulus serves as an S−, the effect is a sharpening of the generalization gradient. In other words, the 'fall off' in responding accompanying a change in some feature of S+ is larger than it would be without discrimination training. This means that the animal responds even less to a given test stimulus than it would otherwise. Thus, the peak of an excitatory gradient is more 'pointed' and its slope is steeper after discrimination training than after training with just the S+.

Discrimination training sharpens the generalization gradient because it ensures that the relevant stimulus feature(s) of S+ and S− will exert greater control over responding than irrelevant or incidental features. The strengthening of stimulus control by the relevant features occurs because the discrimination contingencies establish them as better predictors of reinforcement and nonreinforcement than other elements common to both stimuli.

To illustrate the impact of this on the generalization gradient, consider an S+ consisting of elements [*a, b, c, d*] and an S− consisting of elements [*a, d*]. During discrimination training, responding in the presence of the [*a, d*] elements will sometimes yield reinforcement (when S+ is present) and sometimes will not (when S− is present). In other words, these two elements are relatively poor predictors of reinforcement. By contrast, elements [*b, c*] are perfectly correlated with reinforcement and nonreinforcement: when they are present, reinforcement is available; when they are absent, reinforcement is unavailable. Consequently, responding will be more strongly conditioned to the predictive [*b, c*] elements than to the less predictive [*a, d*] elements. A number of studies have shown that stimulus control by less predictive stimuli is relatively weak compared to that of more predictive stimuli. In addition, the control exerted by poorly predictive elements such as [*a, d*] following discrimination training is demonstrably weaker than the control they would exert if they were just as predictive of reinforcement as anything else, as would occur after training with only the S+.

Consider now a single generalization test stimulus consisting of elements [*c, d, e, f*]. In the absence of discrimination training (that is, after training with S+ alone), responding to this test stimulus will occur by virtue of the conditioning accruing to both the [*c*] and [*d*] elements. By contrast, after discrimination training of the sort described above, control by the [*d*] element should be weakened or 'neutralized' leaving only the [*c*] element to control responding. Consequently, less responding will

occur to this test stimulus after discrimination training than after training with only the S+; stated otherwise, the gradient between these two values will be steeper.

Interdimensional Discrimination Training

Interdimensional discrimination training involves reinforcing responding in the presence of a particular stimulus feature but nonreinforcing it when the feature is absent (or vice versa). Such training yields generalization gradients that are much steeper than those obtained when subjects are trained only with the stimulus containing the feature of interest. For example, although pigeons produce flat frequency-generalization gradients when their training consists solely of food-reinforced pecking in the presence of a particular tone, they provide decremental gradients when their training consists of reinforced presentations of the tone (S+) randomly alternated with nonreinforced presentations of the key light without the tone (S−).

With interdimensional discrimination training between tone present (S+) and tone absent (S−), the lighted key alone signals reinforcement only 50 percent of the time. By contrast, the presence versus absence of the tone is perfectly correlated with reinforcement versus nonreinforcement. Not surprisingly, then, responding is predominantly controlled by the auditory stimulus, and this will be evident in the finding that varying one of its attributes (frequency) produces an orderly decrement in responding that would not otherwise occur. Thus, the appearance of a decremental gradient following such training arises in part because the incidental feature common to all of the test stimuli (i.e. the lighted key) has been 'neutralized'.

Interdimensional discrimination training is also used to obtain inhibitory (incremental) gradients. Here, the feature varied during the generalization test is initially established as a signal for *non*reinforcement (an S−); a stimulus minus that feature is concurrently established as a signal for reinforcement (an S+). Because subjects learn to respond only when the feature is absent, the feature-present stimulus (the S−) can be viewed as a stimulus controlling not-responding (or, more precisely, some behavior other than that conditioned to the S+). If 'not-responding' generalizes to stimuli similar to S−, then progressively less 'not-responding' should occur to test stimuli increasingly different from the S−. Generalization of not-responding often translates into a progressive increase in the responses separately conditioned to S+. The result is a U-shaped gradient with a minimum at S−, shown by the open-symbol gradients in Figure 1. The slopes of incremental gradients are often not as steep as those of decremental gradients because not all of the generalized 'not-responding' ends up as the measured responses conditioned to the S+. Some of the generalization ends up in other behavior not recorded by the researcher.

Inhibitory generalization gradients can also be obtained by intermittently reinforcing responding to every stimulus in the generalization test (as opposed to testing in extinction). This resistance-to-reinforcement test is useful because it counteracts any tendency of subjects not to respond to the S− and to any test stimulus containing its negative feature(s). In addition, it raises the overall level of responding during the generalization test, thus avoiding the possibility of flat gradient that could easily arise if subjects rarely respond to any stimulus along the 'negative' test dimension. The incremental or U-shaped gradient obtained in a resistance-to-reinforcement test arises because the S− stimulus is the most 'negative' and, thus, the most resistant to the response-enhancing effects of reinforcement. Test stimuli further removed from S− are progressively less 'negative', and so are progressively less resistant to the effects of reinforcement.

Intradimensional Discrimination Training

Intradimensional discrimination training involves two stimuli both of which lie on the generalization test dimension – for example, two different colors, two different auditory frequencies, or two different line orientations. Both stimuli, in other words, contain the feature varied during the generalization test. During training, responding is typically reinforced in the presence of one stimulus (the S+) and nonreinforced in the presence of the other (the S−). As the preceding discussion suggests, generalization gradients around the S+ from an intradimensional discrimination are sharper than those around the same stimulus trained in isolation (i.e. when subjects learn to respond only to an S+). Sharpening of the gradient under these conditions is hardly surprising considering that subjects have explicitly learned *not* to respond to one value along the test dimension (the S−). Moreover, intradimensional discrimination training should reduce stimulus control by any incidental features common to the S+ and S−. Nonetheless, the post-discrimination gradient (PDG) obtained following

such training frequently exhibits a unique characteristic not anticipated by an analysis of stimulus generalization in terms of common elements or incidental features – a peak shift.

THE PEAK SHIFT

A generalization test following intradimensional discrimination training often reveals a shift in the peak level of responding from S+ to a value further removed from S−. Figure 2 shows an example of a peak-shifted PDG and, for comparison purposes, a gradient obtained after training only with the S+. The peak shift has been observed in many different species: pigeons, rats, horses, goldfish, and humans.

The peak shift was discovered in the course of testing a theoretical account of the phenomenon of transposition. Transposition refers to the finding that after reinforcing responding to, say, the larger of two discriminative stimuli, subjects given a choice between the S+ and an even larger stimulus frequently choose the latter. One theoretical account claims that transposition reflects the algebraic summation of two opposing response tendencies: an excitatory (response-producing) tendency to S+ and an inhibitory tendency to S−. Given that each tendency will generalize to other similar stimuli, yielding an excitatory gradient around S+ and an inhibitory gradient around S−, response strength to any stimulus simply equals the net excitatory and inhibitory tendencies, respectively, at that point. By this account, there is more inhibitory strength at S+ than at stimulus values beyond it because S+ is closer to the S− than more remote values, so more inhibition will generalize to it. This yields a lower net response strength at S+ than at immediately adjacent but more extreme values and, hence, transposition. In addition, it should yield a peak shift – greater responding to values further removed from the S− than the S+ itself in a generalization test.

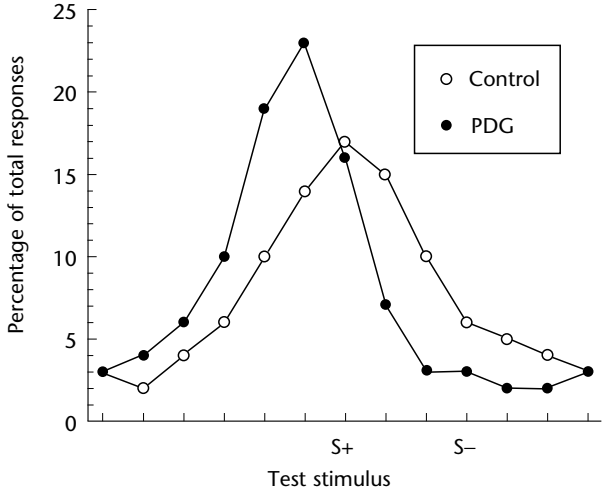


Figure 2. Relative generalization gradients following training with just a reinforced stimulus (Control) and following intradimensional discrimination training between a reinforced (S+) and a nonreinforced (S−) stimulus (PDG). The postdiscrimination gradient (PDG) exhibits a peak shift.

This gradient-interaction account of the peak shift predicts, and empirical findings have confirmed, that the size of the shift, or the likelihood of obtaining it, is inversely proportional to the S+/S− separation during training. In other words, the closer the two discriminative stimuli are, the larger the shift or the greater the likelihood of obtaining it. Unfortunately, a full evaluation of this account is difficult when predictions are derived from *hypothetical* excitatory and inhibitory gradients because the predicted PDG crucially depends on the form and steepness of each gradient. For instance, no peak shift is anticipated if the hypothesized inhibitory gradient is especially steep or especially flat. To avoid this difficulty, researchers have sometimes used actual excitatory and inhibitory gradients obtained after interdimensional discrimination training (see Figure 1) to predict the PDG following intradimensional discrimination training. The predicted PDGs sometimes resemble the empirically obtained PDGs, but sometimes do not.

Despite these theoretical difficulties, the increased potency of stimuli beyond the S+ is a well-established, reproducible finding. Researchers have also shown that the PDG will show a peak shift even when the 'negative' stimulus is not an S−. Specifically, a peak-shifted gradient also occurs when the stimulus alternating with the S+ is one that signals a lowered rate of reinforcement than that experienced during the S+. Apparently, such a stimulus is relatively 'negative' or aversive in comparison to the S+, and so controls a response tendency opposite to that controlled by the S+. Indeed, after interdimensional training, varying a reinforced stimulus that also signals a drop in the rate of reinforcement yields an inhibitory or incremental gradient.

CONCLUSION

Generalization tests are used to assess stimulus control, the degree to which particular stimulus features control responding. Gradients around a previously reinforced stimulus are typically

decremental, with the greatest amount or highest probability of responding at the former S+ and increasingly less responding to values further and further removed from it. Gradients around a previously nonreinforced stimulus (S−) are often incremental if responding is separately reinforced to another stimulus that does not contain the S− feature. Such interdimensional discrimination training also sharpens the generalization gradient around the feature when it serves as an S+. The sharpening occurs because discrimination training increases the relative control exerted by the predictive or relevant aspects of the training stimuli and reduces control by the irrelevant or incidental aspects. Intradimensional training involves discrimination between two stimulus values on a continuum and can often cause other stimuli to have a greater effect on generalized responding than the reinforced stimulus (S+) itself. This effect, called the peak shift, has been traditionally interpreted in terms of the combination of two opposing tendencies: a generalized tendency to respond to S+ and a generalized tendency not to respond to S− or to a stimulus with a lower reinforcing value than S+.

Further Reading

Guttman N and Kalish HI (1956) Discriminability and stimulus generalization. *Journal of Experimental Psychology* **51**: 79–88.

Honig WK and Urcuioli PJ (1981) The legacy of Guttman and Kalish (1956): 25 years of research on stimulus generalization. *Journal of the Experimental Analysis of Behavior* **36**: 405–445.

Mackintosh NJ (1974) *The Psychology of Animal Learning* (chap. 9: Generalization). New York, NY: Academic Press.

Purtle RB (1973) Peak shift: a review. *Psychological Bulletin* **80**: 408–421.

Rilling M (1977) Stimulus control and inhibitory processes. In: Honig WK and Staddon JER (eds) *Handbook of Operant Behavior*, pp. 432–480. New York, NY: Prentice-Hall.

Spence KW (1937) The differential response in animals to stimuli varying within a single dimension. *Psychological Review* **44**: 430–444.

Thomas DR (1993) A model for adaptation-level effects on stimulus generalization. *Psychological Review* **100**: 658–673.

Gesture

Intermediate article

Susan Goldin-Meadow, University of Chicago, Chicago, Illinois, USA

CONTENTS

The hand movements that people of all ages and all cultures produce while talking are called gestures. These gestures can reflect thoughts that are not expressed in the speech they accompany and, as a result, can play a role in problem-solving and language acquisition.

INTRODUCTION

Gesture is frequently equated with nonverbal communication and it is indeed nonverbal – performed with the hands, arms, head, legs, and body but not with the apparatus specialized for speech. However, the study of nonverbal communication is concerned with how the behavior of others is regulated and how inner states are expressed (Mueller, 1998) and takes as its focus all parts of the body. In contrast, the current study of gesture focuses on how thought is represented and focuses primarily on the hands.

Gesture is comprised of a continuum of communicative actions which differ on a number of dimensions (Kendon, 1980) – whether they must accompany speech; whether their form is governed by convention; whether they are characterized by linguistic properties (e.g. are segmented and analytic). At one end of the continuum, sign languages invented by deaf communities are produced without speech, are conventionalized, and have the same repertoire of linguistic properties as spoken languages. At the other end, the spontaneous hand movements that are generated when people talk are always produced with speech, are idiosyncratic without recognized standards of form, and are characterized by global and synthetic rather than analytic and segmented forms of representation. In the middle, occupying less extreme positions on each of the dimensions, are gestures such as emblems (e.g. thumbs-up for OK) and pantomime (pretending to play a violin).

Because they are not bounded by convention, the spontaneous hand movements that accompany speech can offer a unique window into thought. As a result, they are the focus here.

THE RELATION BETWEEN GESTURE AND SPEECH

When people move their hands as they talk – they gesture. Gesture is a widespread phenomenon, occurring across cultures, ages, and tasks. Even individuals who are blind from birth gesture as they speak, despite the fact that they have never seen gesture (Iverson and Goldin-Meadow, 1998). Gesture thus appears to be an integral part of the speaking process.

Gesture and speech are tightly intertwined in time and meaning (McNeill and Duncan, 2000). For example, a speaker raises her hand just as she says, 'and he climbs up the pipe', with the upward gesture overlapping in time with the phrase 'up the pipe'. Typically, gesture slightly precedes its co-expressive speech, and it does so systematically – the more unfamiliar the word, the longer the interval between the onset of gesture and the onset of speech, and the longer the duration of the gesture itself (Morrel-Samuels and Krauss, 1992). Moreover, the processing unit that gesture and speech form resists division. Artificially delaying the feedback speakers receive from their own voices grossly disrupts the timing of speech, but gesture–speech synchrony remains intact (McNeill, 1992). Clinical stuttering also disrupts speech but has no effect on gesture–speech synchrony (Mayberry *et al*., 1998).

Producing gesture and speech together brings the imagery conveyed in gesture into the system of categories provided by the spoken language; perhaps not surprisingly, gesture is influenced by those categories. For example, English-speakers typically mention the manner by which a motion is carried out in the verb itself (e.g. *flies* out), while Spanish-speakers mention manner, if at all, in a separate construction, often a gerund (sale *volando* = exits *flying*). The way in which English- and

Spanish-speakers produce gestures for manner differs accordingly (McNeill and Duncan, 2000). English-speakers produce manner gestures in synchrony with their verbs; they use their manner gestures when they want to draw attention to manner (e.g. a flapping gesture to indicate how the flying was done), but use path gestures when they are forced by convention to mention manner in their verbs but do not want to focus on it (e.g. a forward-moving gesture to focus attention on the path of the flight rather than the flapping). In contrast, Spanish-speakers often omit manner from their speech but produce it in gesture nonetheless; their manner gestures do not synchronize with a single grammatical category but extend through multiple clauses. Thus, although gesture has a systematic relation to speech in all languages, the nature of that relation varies as a function of the particular language.

THE ROLE OF GESTURE IN PROBLEM-SOLVING

When produced within a single utterance, gesture and speech express the same underlying idea unit. However, the two modalities often highlight different aspects of that idea. For example, a speaker says 'I climbed the stairs' while spiraling his index finger upward. The speaker's gestures provide the *only* clue that the staircase is a spiral. By looking at gesture and speech as a unit, we gain access to thoughts that speakers have but do not express in their speech.

When in a problem-solving situation, a speaker's gestures often express thoughts that pertain to the same topic as speech but are not always easy to integrate with the aspects of that topic highlighted in speech. For example, consider a child explaining her judgment that the amount of water in a tall, thin glass changed when it was poured into a short, wide dish. The child says, 'they're different because that one's wider than that one', while at the same time indicating with her hands the height of the dish and then the glass. The child focused on the containers' widths in speech but their heights in gesture. Such *gesture–speech mismatches* (Church and Goldin-Meadow, 1986) are found in a variety of tasks and over a large age range: toddlers going through a vocabulary spurt; preschoolers explaining a game; elementary school children explaining mathematical equations and seasonal change; children and adults discussing moral dilemmas; adolescents explaining mechanical tasks; and adults explaining how gears work and problems involving constant change.

The gestures that speakers produce when describing a problem reflect their representations of that problem. Importantly, those representations play a role in predicting how speakers will then go about solving the problem (Alibali *et al*., 1999). For example, speakers often incorporate information about manner of change into their gestures when asked to describe algebra word problems involving continuous change (e.g. change in the amount of air pressed per minute into a hot air balloon over a 30-minute period) or discrete change (e.g. change in the number of books on each shelf of a six-shelf bookcase). The representation reflected in gesture is not always the same as the representation reflected in speech – a speaker may talk about the problem as one of continuous change while producing discrete, steplike gestures. When gesture reinforces the representation displayed in speech (e.g. both reflect a continuous representation), adults are very likely to solve the problem using a strategy compatible with that representation (a continuous strategy); this is much more likely than when gesture does *not* reinforce the representation reflected in speech (i.e. when gesture reflects a discrete representation, but speech reflects a continuous representation). Thus, gesture and speech together provide a better index of mental representation than speech alone.

Taken together, gesture and speech also provide an excellent index of a speaker's readiness to learn. Children who produce many gesture–speech mismatches when explaining their incorrect answers to math problems, or to reasoning problems about quantity, are particularly likely to benefit from instruction on those problems – much more so than children who produce few gesture–speech mismatches in their explanations of the same problems (Church and Goldin-Meadow, 1986; Perry *et al*., 1988). Gesture considered in relation to speech thus reflects mental processes. Does gesture also affect those processes? There are two different, though not mutually exclusive, ways in which gesture can bring about cognitive change (Goldin-Meadow, 2000).

The first is an outgrowth of the fact that gesture is 'out there', occurring routinely in naturalistic talk and reflecting aspects of a speaker's cognitive state. If communication partners are able to read the signals contained in a speaker's gestures, they may alter their interactions with that speaker as a function of the information they glean from gesture. Gesture, by influencing the input speakers receive from others, would then be part of the process of change itself. In fact, adults, and even children, are able to interpret the spontaneous gestures that

speakers produce in problem-solving situations (Goldin-Meadow and Sandhofer, 1999). Whether communication partners then act on the information they glean from a speaker's gestures, however, remains an open question.

The second asks whether gesture has effects on speakers themselves and thus brings about cognitive change more directly. Gesture externalizes ideas differently and draws on different resources than speech uses. Using gesture along with speech may therefore divide the cognitive burden across several systems. For example, if speakers are asked to remember a list of words while explaining how they solved a mathematics problem, they remember significantly more words when they gesture during their explanations than when they do not (Goldin-Meadow *et al*., 2001). The components of the verbal explanation that can be encoded in gesture are presumably shifted out of a verbal store and into a visuospatial store. Gesturing during the mathematics explanations thus reduces the load on verbal working memory, freeing capacity to be used on another verbal task (the word–memory task). In this sense, gesturing not only reflects a speaker's cognitive state but may assist in the allocation of mental resources and, in this way, plays a role in shaping that state.

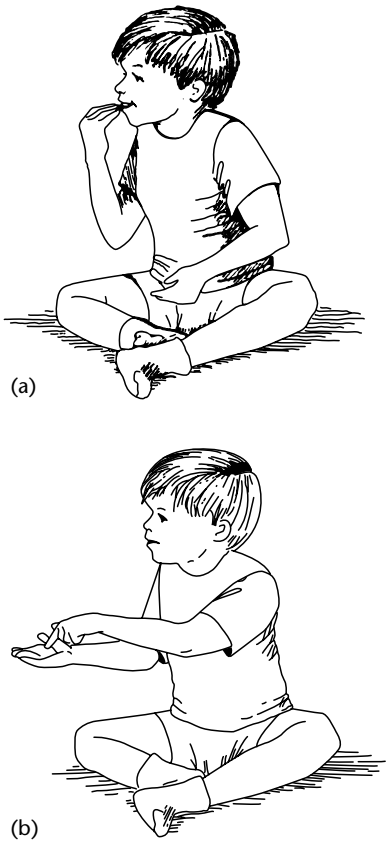


Figure 1. Self-made signs in a deaf boy never exposed to sign language. A two-sign sequence: (a) the first sign means 'eat' or 'food' (in this case, given immediately before the boy had pointed to a grape); (b) the second sign means 'give'. The total sequence presumably means 'give me the food'. From Goldin-Meadow (1981); drawing courtesy Noel Yovovich.

THE ROLE OF GESTURE IN LANGUAGE ACQUISITION

Gesture is very often a young child's first way of communicating with others. At a time when children are limited in what they can say, gesture can extend the range of ideas they are able to express. The earliest gestures children use, typically beginning around 10 months, are deictics, gestures whose referential meaning is given entirely by the context and not by their form: e.g. holding up an object to draw an adult's attention to that object or, later in development, pointing at the object (Bates *et al*., 1979). In addition to deictic gestures, children also use iconics. Unlike deictics, the form of an iconic gesture captures aspects of its intended referent and thus its meaning is less dependent on context, e.g. opening and closing the mouth to represent a fish. These iconic gestures are rare in some children, frequent in others. If parents encourage their children to use iconic gestures, these gestures become more frequent, facilitating, at least temporarily, the child's production of words (Goodwyn *et al*., 2000). The remaining types of gestures that adults produce – metaphorics (gestures whose pictorial content presents an abstract idea rather than a concrete object or event) and beats (small baton-like movements that move along with the rhythmical pulsation of speech) – are not produced routinely until relatively late in development.

Combining gesture and speech within a single utterance can also increase the communicative range available to the child. Most of the young child's gesture–speech combinations contain gestures that convey information redundant with the information conveyed in speech, e.g. pointing at a cookie plus 'cookie'. However, young children also produce combinations in which gesture conveys information that is different from the information conveyed in speech, e.g. pointing at a cookie plus 'mine'. This second type of combination allows a child to express two elements of a sentence (one in gesture and one in speech) at a time when the child may not be able to express those elements within a single spoken utterance. Interestingly, the onset of

this type of gesture–speech combination reliably predicts the onset of two-word speech. Children who produce combinations such as pointing at cookie plus 'mine' early in development are the first to produce combinations such as 'cookie mine' (Butcher and Goldin-Meadow, 2000). Note that it is the relation between gesture and speech that predicts change in the child's language, providing further evidence that gesture and speech form an integrated system.

All children who are learning a spoken language use gesture. But some children – deaf children with profound hearing losses, for example – are unable to learn the spoken language that surrounds them. If exposed to a conventional sign language, these deaf children would acquire that language. If, however, they are not exposed to sign, deaf children rely on gesture (Figure 1). Interestingly, their gestures do not have the global and synthetic form of their hearing parents' gestures but take on characteristics of systems at the other end of the continuum – self-created standards of form and the analytic and segmented properties of a linguistic system (Goldin-Meadow and Mylander, 1998); these systems are called 'home sign'. Thus, gesture is an adaptable modality, able to form with speech a single communication system or, as the need arises, assume on its own the properties of a linguistic system.

References

Alibali MW, Bassok M, Solomon KO, Syc SE and Goldin-Meadow S (1999) Illuminating mental representations through speech and gesture. *Psychological Science* **10**: 327–333.

Bates E, Benigni L, Bretherton I, Camaioni L and Volterra V (1979) *The Emergence of Symbols: Cognition and Communication in Infancy*. New York: Academic Press.

Butcher C and Goldin-Meadow S (2000) Gesture and the transition from one- to two-word speech: when hand and mouth come together. In: McNeill D (ed.) *Language and Gesture*, pp. 235–257. New York: Cambridge University Press.

Church RB and Goldin-Meadow S (1986) The mismatch between gesture and speech as an index of transitional knowledge. *Cognition* **23**: 43–71.

Goldin-Meadow S (1981) In: Gleitman H (ed.) *Psychology*. New York: WW Norton.

Goldin-Meadow S (2000) Giving the mind a hand: the role of gesture in cognitive change. In: McClelland J and Siegler RS (eds) *Mechanisms of Cognitive Development: Behavioral and Neural Perspectives*, pp. 5–31. Mahwah, NJ: Erlbaum Associates.

Goldin-Meadow S and Mylander C (1998) Spontaneous sign systems created by deaf children in two cultures. *Nature* **391**: 279–281.

Goldin-Meadow S, Nusbaum N, Kelly S and Wagner S (2001) Explaining math: gesturing lightens the load. *Psychological Science* **12**: 516–522.

Goldin-Meadow S and Sandhofer CM (1999) Gesture conveys substantive information about a child's thoughts to ordinary listeners. *Developmental Science* **2**: 67–74.

Goodwyn SW, Acredolo LP and Brown CA (2000) Impact of symbolic gesturing on early language development. *Journal of Nonverbal Behavior* **24**: 81–103.

Iverson JM and Goldin-Meadow S (1998) Why people gesture as they speak. *Nature* **396**: 228.

Kendon A (1980) Gesticulation and speech: two aspects of the process of utterance. In: Key MR (ed.) *The Relation between Verbal and Nonverbal Communication*, pp. 207–227. The Hague: Mouton.

Mayberry RI, Jaques J and DeDe G (1998) What stuttering reveals about the development of the gesture–speech relationship. In: Iverson JM and Goldin-Meadow S (eds) *New Directions for Child Development, no. 79. The Nature and Functions of Gesture in Children's Communication*, pp. 77–87. San Francisco: Jossey-Bass.

McNeill D (1992) *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill D and Duncan S (2000) Growth points in thinking-for-speaking. In: McNeill D (ed.) *Language and Gesture*, pp. 141–161. New York: Cambridge University Press.

Morrel-Samuels P and Krauss RM (1992) Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition* **18**: 615–622.

Mueller C (1998) *Redebegleitende Gesten: Kulturgeschichte – Theorie – Sprachbvergleich*. Berlin: Verlag Arno Spitz.

Perry M, Church RB and Goldin-Meadow S (1988) Transitional knowledge in the acquisition of concepts. *Cognitive Development* **3**: 359–400.

Further Reading

Acredolo LP and Goodwyn S (1996) *Baby Signs: How to Talk with Your Baby Before Your Baby Can Talk*. Chicago: NTB/Contemporary Books.

Emmorey K and Reilly J (eds) (1995) *Language, Gesture, and Space*. Hillsdale, NJ: Erlbaum Associates.

Goldin-Meadow S (1999) The role of gesture in communication and thinking. *Trends in Cognitive Science* **3**: 419–429.

Goldin-Meadow S, Alibali MW and Church RB (1993) Transitions in concept acquisition: using the hand to read the mind. *Psychological Review* **100**: 279–297.

Goldin-Meadow S, McNeill D and Singleton J (1996) Silence is liberating: removing the handcuffs on grammatical expression in the manual modality. *Psychological Review* **103**: 34–55.

Iverson JM and Goldin-Meadow S (eds) (1998) *New Directions for Child Development, no. 79. The Nature and Functions of Gesture in Children's Communication*. San Francisco: Jossey-Bass.

Kendon A (1994) Do gestures communicate?: a review. *Research on Language and Social Interaction* **27**: 175–200.
Krauss RM (1998) Why do we gesture when we speak? *Current Directions in Psychological Science* **7**: 54–60
McNeill D (ed.) (2000) *Language and Gesture*. New York: Cambridge University Press.
Messing LS and Campbell R (eds) (1999) *Gesture, Speech, and Sign*. Oxford: Oxford University Press.
Volterra V and Erting CJ (eds) (1990) *From Gesture to Language in Hearing and Deaf Children*. New York: Springer-Verlag.

Group Behavior

Introductory article

Bibb Latané, Latané Center for Human Science, Highland Beach, Florida, USA

CONTENTS

By influencing one another in proportion to their strength, immediacy, and number, individuals create spatially clustered groups of people similar in their beliefs, values, and behavior.

The behavior of individuals interacting in social networks leads to the emergence of groups of like-minded individuals, and the interaction of these groups, in turn, leads to the temporal evolution and regional differentiation of the patterns of beliefs, values, and habits that characterize human culture. This broad generalization, based on the detailed modeling of such interactions, can provide a basis for understanding the reality of social life. The underlying models of the dynamics of social interaction may also help us organize ideas about the role of social groups and networks in shaping the contents of individual cognition.

SOCIAL INFLUENCE AND GROUP BEHAVIOR

We live in a physical and cognitive world constructed and modified by people following natural law operating through biological evolution. Assisted by large and small ideas and by harnessing animal and other sources of energy, humans working alone and together have transformed the surface of our planet. For better or worse, the world we have made is far different from that experienced by our earliest ancestors. Little of this could have happened without the human capacity and propensity for social influence.

Social influence is a pervasive component of human experience, permeating every aspect of our lives. Individuals affect one another in many different ways. We may be aroused by the presence of other people, frightened by their attention, touched by their plight, guided by their example, persuaded by their arguments, coerced by their power. Such forms of social influence can operate singly or simultaneously (the failure of bystanders in groups to intervene in an emergency, for example, can result from performance apprehension, social influence from seeing others not respond, and/or diffusion of responsibility). We learn by example and by instruction from our parents, from our peers, and from many generations of predecessors.

These types of individual influence come about from different mechanisms and can have different consequences. Sometimes the effect of the presence or actions of others is to reduce pressures on an individual to work hard in a group or to intervene in an emergency; sometimes the effect is to increase pressures to act or to change attitudes. Yet, each type of influence seems to follow certain general laws – the laws of social impact. Because individuals differ in their power, intelligence, and expressiveness, a major factor governing social influence is the strength of an influence source – just as a brighter bulb casts more light, some people are more influential than others. Because most influence takes place through direct contact, immediacy facilitates impact: strangers have less influence than neighbors. Because influence can be a gradual, cumulative process, it grows in proportion to the number of people who are the source of influence: a larger group can be more influential than a single individual. Research has shown that these effects tend to be multiplicative, so that the amount of impact or pressure to change experienced by an individual will tend to be proportional to the strength, immediacy, and number of people who are the source of pressures to change relative to the strength, immediacy, and number of people who provide support.

DYNAMIC SOCIAL IMPACT – THE EMERGENCE OF GROUPS FROM INTERACTING INDIVIDUALS

These laws may be useful in predicting and explaining the reactions of individual humans to their social environments, but they also raise a fascinating question: what should we expect from a network of people each responding to and being influenced by the others? For example, if persons *n* and *o* both affect *p*, the resulting changes in *p* will in turn affect the social pressures experienced by *n* and *o*.

In fact, analytic theory, computer simulation, and experimentation identify four group phenomena that should result from the interaction of individuals influencing one another:

1. Because members of minorities are exposed to more adverse influence than people in the majority, they are more likely to be converted, resulting in a *consolidation* of positions on any group topic.
2. Because people are more influenced by their neighbors than by more distant people, they are likely to become more similar to them, resulting in a *clustering* of positions on any given topic.
3. Because neighbors can influence one another on several different topics, a *correlation* of positions will emerge over time.
4. Because clustering produces neighborhoods of like-minded people, minority neighbors will shield one another from influence by the global majority, resulting in a *continuing diversity* of beliefs.

CONTAGIOUS THOUGHTS – AN EXAMPLE OF DYNAMIC SOCIAL IMPACT

To understand these phenomena, consider the results of some recent experiments in which 456 participants were organized into 24-person networks, each consisting of six four-person subgroups or 'families'. During five computer sessions spread over a three-week period, participants exchanged electronic messages with the other three members of their family as well as with one member of a neighboring family. People were randomly assigned to positions in the communication network and had no prior or concurrent contact with one another aside from the electronic message exchange. Message exchange was asynchronous, as when people correspond by e-mail. Messages sent one session were read the next.

Norm Detection

Participants were asked to guess the majority preference in their 24-person group on a series of bipolar choices such as 'Which will be the preferred colour, red or green?' or '…the preferred artist, Klee or Kandinski?' At the first session, each person guessed the positions thought to be held by the majority of the full group on each of six issues. At each succeeding session, these guesses were read and reacted to by the three other family members as well as by the outside contact. In order to motivate them, each person earned $1 each time their final answer corresponded to that of the majority of the 24 participants. Obviously, the best strategy for predicting the group norm would be to combine one's own best guess with the information from others, and indeed individuals seemed to do this, changing their own answers 75 percent of the time when three to four people disagreed with them, but only 5 percent when fewer did so. Under these conditions, beliefs about norms can be considered 'contagious' with people 'catching' whichever view is most common in their neighborhood. The four C's of social impact are exemplified in the emergent pattern of beliefs in the group.

Consolidation is shown by the fact that the number of people in the minority was reduced for each of the six issues by an average of 20 percent per item. For example, before discussion, 11 people thought Elgar would be more popular than Handel, while 8 thought so afterwards.

Clustering is represented by the fact that there is less disagreement within subgroups after message exchange than before. In this case, the chance of two people within the same family disagreeing about the group norm was only 6 percent, compared to the 50 percent to be expected by chance and found before discussion and for any two people not in the same group after discussion. The remarkable degree of within-group similarity is not a result of self-segregation, because people could not move from one group to another; in other situations in which they can move then clustering would be expected to be even greater.

Correlation is shown by the increased number of significant correlations across the 24 people between all possible pairs of issues: from 5 percent before message exchange to 35 percent after. Initially unrelated beliefs became related through a reduction in independence from social influence, not because of any inherent association. In real-world groups, such correlation can be expected to provide a basis for a perceived group ideology.

Continuing diversity is maintained because further messages should lead to little further change in individual norm estimates, since every person is now in a local majority. Even if the task were to continue indefinitely, there is no reason to expect further convergence.

EVOLUTION, DISSEMINATION, AND PRESERVATION OF IDEAS

These four phenomena are not restricted to beliefs about norms, or to situations where people are bribed to agree. Further research has shown that the same 'four C's' emerge when people are discussing political and social issues, giving reasons for their beliefs. In fact, across a wide range of tasks, 'the four C's' of social impact emerge in proportion to the degree to which people are influenced by one another.

Participants in this experiment had no knowledge of who was exchanging messages with whom, and thus could not see the 'family' resemblances that came about through communication. If we imagine that people did have a chance to see the similarities within their own neighborhood and how their group differed from others, we could expect them to form social identities, to make invidious comparisons, to develop group solidarity, and to engage in in-group favoritism. By such processes, groups can create an identity for themselves, and this identity can remain distinct over long periods of time and change. The 'four C's' of group dynamics can be understood as the nonlinear dynamics of spatially distributed individuals influencing each other in proportion to their strength, immediacy, and number. Like other self-organizing systems, groups of all sizes respond in complex, nonintuitive ways to external and structural change but often produce emergent factions or subcultures with an apparent life of their own. Such entities can even become institutionalized as they recruit and indoctrinate new members and codify their beliefs.

These subcultures can be seen as temporally evolving, regionally clustered, partially correlated sets of socially influenced beliefs, values, and behaviors held by individuals in a spatial network. They can be seen as social representations as characterized by the French social psychologist Serge Moscovici – shared beliefs that evolve from and form the basis for ordinary conversation. As evidence that such processes may have helped shape human history, consider the changing maps of language and dialect, of religious belief and political ideology, and how, on a global scale, they mirror the local neighborhoods described in the experiment above.

This article provides a basis for understanding how social interaction, even in the absence of social mobility, can lead populations to divide into groups of people who share distinctive patterns of beliefs and norms. Focusing on intra-group processes, the 'four C's' provide a foundation for understanding inter-group processes that are beyond the scope of the present article – social identity and inter-group conflict. Although they result from the interactions of individuals, these phenomena characterize groups as a whole, and can only be detected by knowing how individuals are located in a social group or network. They show that group norms can be created from the bottom up as well as imposed from above.

Further Reading

Latané B (1996) Dynamic social impact: the creation of culture by communication. *Journal of Communication* **46**(4): 13–25.

Latané B (1997) Dynamic social impact: the societal consequences of human interaction. In: McGarty C and Haslam A (eds) *The Message of Social Psychology: Perspectives on Mind and Society*, pp. 200–220. Oxford: Blackwell.

Latané B and Bourgeois M (2002) The emergent group mind: how communication creates social representations. In: Forgas JP and Williams KD (eds) *Social Influence: Direct and Indirect Processes*. Philadelphia: Psychology Press.

Human Altruism

Advanced article

C Daniel Batson, University of Kansas, Lawrence, Kansas, USA
David A Lishner, University of Kansas, Lawrence, Kansas, USA
EL Stocks, University of Kansas, Lawrence, Kansas, USA

CONTENTS

Altruism is a motivational state with the ultimate goal of increasing another's welfare. It is contrasted with egoism, which is a motivational state with the ultimate goal of increasing one's own welfare.

INTRODUCTION

Altruism refers to the motivation of one organism, usually human, for benefiting another. Although some biologists and psychologists speak of altruistic behavior, meaning behavior that benefits another, this use of the term altruism is not recommended. It fails to consider motivation for the behavior, and motivation is of primary concern in discussions of altruism. If the ultimate goal in benefiting another is to increase the other's welfare, then the motivation is altruistic. If the ultimate goal is to increase the organism's own welfare, then the motivation is egoistic. Motivation for benefiting another might be altruistic, egoistic, both, or neither.

THE BASIC QUESTION: IS ALTRUISM PART OF HUMAN NATURE?

The question of whether humans have the capacity to be altruistically motivated has been debated for centuries. The majority view among Renaissance and post-Renaissance philosophers, and more recently among biologists, psychologists, and social scientists, is that we humans are, at heart, purely egoistic, that we care for others only to the extent that their welfare affects ours (Mansbridge, 1990). A persistent minority has, however, insisted that altruism exists, that in some circumstances and to some degree, we can act to benefit others for their sakes, and not simply for our own.

In recent years, questions about the existence of altruism have been addressed at two distinct levels, at the level of evolutionary biology and at the level of human psychology. The questions addressed at these two levels differ because what is meant by altruism differs. Evolutionary altruism refers to behavior by one organism that diminishes its reproductive fitness relative to the reproductive fitness of one or more other organisms. Psychological altruism refers to a motivational state with the ultimate goal of increasing another's welfare. Evolutionary altruism is neither necessary nor sufficient to produce psychological altruism. The question of whether altruism is part of human nature concerns psychological altruism, not evolutionary altruism. Yet it is evolutionary altruism that has received the most attention in recent years.

EVOLUTIONARY ALTRUISM

Any behavior that diminishes the probability of an organism producing offspring may, at first glance, seem to decrease the likelihood of the genes carried by that organism appearing in the next generation. If such behavior is associated with a particular genetic alternative, or allele, then over successive generations this allele should decrease in the population and so should the behavior. This logic suggests that a genetically based inclination toward evolutionary altruism is contrary to the theory of natural selection, leading many biologists to conclude that evolutionary altruism cannot exist. Yet people do at times help in ways that reduce their reproductive potential.

In the 1960s and 1970s some evolutionary biologists (often called sociobiologists because of their interest in the genetic basis of social behavior) sought to explain how helping that diminishes the reproductive fitness of the helper, and thus seems to violate the theory of natural selection, really does not. Edward Wilson (1975), Richard Dawkins

(1976), and others pointed out that although biological reproduction occurs at the level of the individual organism, the unit of natural selection is not the individual but the gene. Under certain circumstances genes can enhance their own survival by leading the individual carrying them to risk personal survival to benefit another.

One such circumstance is *kin selection*. Kin selection occurs when the benefactor and the benefited share the same genes, and as a result of the benefactor's help, the benefited has a greater chance of passing these genes to the next generation. In such cases, it may be to the advantage of the shared genes for the benefactor to promote survival of the benefited even at expense to survival of self. Specifically, it will be advantageous when the degree of genetic overlap (relatedness) multiplied by the increased probability of the other placing his or her genes in the next generation outweighs the decreased probability of the benefactor placing his or her genes in the next generation. William Hamilton (1964) used this logic of *inclusive fitness* to explain the self-sacrificial behavior of sterile worker castes among social insects, including bees and wasps. This logic may also explain help given to siblings and other close kin.

A second circumstance is *reciprocal benefit*. Robert Trivers (1971) suggested that acts benefiting nonkin can be genetically based and consistent with the theory of natural selection if they are cases of what he called 'reciprocal altruism'. (We prefer the term 'reciprocal benefit' because, as noted, altruism refers to motivation not simply behavior, and behavior is all Trivers meant.) Helping another survive can be in the best interest of the benefactor's genes if the other is likely to return the benefit, either to the benefactor or to another organism carrying the same genes as the benefactor. As Trivers (1971) explained, 'Reciprocal altruism can also be viewed as a symbiosis, each partner helping the other while he helps himself. The symbiosis has a time lag, however; one partner helps the other and must then wait a period of time before he is helped in turn' (p. 39). For reciprocal benefit to be advantageous to the genes promoting it, three conditions must be met: (1) need situations must arise frequently (providing opportunities to give and receive benefits), (2) the proximity of potential helpers must be high, and (3) cheating (receiving help without reciprocating) must not be viable. One way to discourage cheating is for potential helpers to refuse future benefit to any individual who does not reciprocate. Another way is to inflict harm on cheaters. Consistent with this reasoning, Axelrod and Hamilton (1981) found that a form of reciprocal benefit and retribution, a tit-for-tat strategy, proved stable and effective in competitive encounters (iterated prisoner's dilemma games).

Group selection is the view that an allele promoting individually disadvantageous behavior such as helping can be selected for if this behavior enhances the survival of the group or species. Sociobiologists often took pains to distinguish their emphasis on the gene as the unit of natural selection from group selection. They noted that selection among individuals within the group would allow a self-benefit allele to drive out a group-benefit allele, even if the latter allele gave the group a selective advantage relative to other groups.

Sober and Wilson (1998) recently pointed out, however, that to juxtapose gene selection and group selection inappropriately contrasts the unit of selection (the gene) with a level of selection (the group). They suggested that gene selection can actually occur at a range of levels – at the level of the gene, the individual, and the group – leading them to propose a *multilevel* selection theory. Selection often occurs among individuals within groups, but there are cases of strong intra-group cohesion, coordination, and cooperation (e.g. social insect colonies) in which selection also occurs among groups. At times, the reproductive benefit of acting to benefit the group can outweigh the reproductive cost relative to other group members of this action. Viewed from the perspective of multilevel selection theory, kin selection and reciprocal benefit can be considered special cases of group selection. Whether human groups ever have the degree of cohesion necessary for genetically based group selection to evolve is, at present, unclear. Selection among human groups may be a product of cultural, rather than genetic, evolution (Campbell, 1975).

PSYCHOLOGICAL ALTRUISM

Over the centuries, the most frequently proposed source of altruistic motivation has been an other-oriented emotional response congruent with the perceived welfare of another person – what is today often called 'empathy'. If another person is in need, then empathic emotions include sympathy, compassion, tenderness, and the like. The empathy–altruism hypothesis claims that empathic emotion produces motivation with an ultimate goal of increasing the welfare of the person for whom the empathy is felt – that is, altruistic motivation.

Experimental research has provided evidence that feeling increased empathy for a person in

need can cause increased helping of that person (see Eisenberg and Miller, 1987, for a review). To observe an empathy-helping relationship, however, does not reveal the nature of the motivation that underlies this relationship. Benefiting the other person could be (a) an ultimate goal, producing self-benefits as unintended consequences, (b) an instrumental goal on the way to the ultimate goal of gaining one or more self-benefits, (c) both, or (d) neither. In other words, the motivation could be altruistic, egoistic, both, or neither. To know whether empathy produces altruistic motivation, it is necessary to determine whether benefiting the person for whom empathy is felt is an ultimate goal.

Three general classes of self-benefits have been proposed to result from helping a person for whom empathy is felt: (1) helping reduces one's empathic arousal, which may be experienced as aversive; (2) helping permits one to avoid possible social and self-punishments for failing to help (e.g. blame, guilt); and (3) helping offers one social and self-rewards for doing what is good and right (e.g. praise, pride). Advocates of the empathy–altruism hypothesis do not deny that these self-benefits of empathy-induced helping exist. They claim, however, that these self-benefits are unintended consequences of reaching the ultimate goal of reducing the other's suffering. Advocates of the egoistic alternatives to the empathy–altruism hypothesis disagree. They claim that one or more of these self-benefits is the ultimate goal of empathy-induced helping.

EVIDENCE FOR THE EXISTENCE OF ALTRUISM

To test these claims, it is necessary to employ experiments, not self-reports. People may not know, or report, their true motives. In the past 25 years, more than 30 experiments have tested the three egoistic alternatives against the empathy–altruism hypothesis (see Batson and Shaw, 1991, for a partial review). The basic research strategy has been to provide individuals induced to feel either low or high empathy for a given person in need with an opportunity to help that person. A cross-cutting variable is also included, making it possible for some of these individuals to obtain one or more of the self-benefits proposed as the source of empathy-induced helping without having to help. If this modification eliminates the increase in helping produced by empathy, then one has evidence that the ultimate goal of the empathy-induced helping was to obtain the self-benefit. If the modification does not eliminate the increase, then one has evidence that this self-benefit was not the ultimate goal. Only after all plausible self-benefits – and other motives – have been ruled out in this way does one have grounds for concluding that the ultimate goal is to benefit the person in need, that the motivation is altruistic.

Results of these experiments have effectively ruled out all plausible self-benefits and other motives proposed to date and have provided remarkably consistent support for the empathy–altruism hypothesis. Because these results contradict the dominant assumption in Western thought that all human motivation is egoistic, they have proved controversial. Attempts to find a plausible egoistic explanation for them continue. At this point, however, it seems likely that empathy-induced altruism is part of human nature. Whether there are sources of altruism other than empathy, such as a personal disposition to be altruistic (an 'altruistic personality'), is not yet clear.

One important contribution of the recent psychological research has been to provide experimental methods for differentiating altruistic motivation from various egoistic alternatives. These methods make it possible to address the question of the existence of human altruism directly by empirical observation of the behavior of humans in controlled laboratory conditions – rather than indirectly through deduction, extrapolation, anecdotes, and self-reports.

CONCLUSION

Evolutionary altruism refers to behavior by one organism that diminishes its reproductive fitness relative to that of one or more other organisms. Evolutionary altruism can exist in cases of kin selection and reciprocal benefit, both of which can be considered special cases of group selection in a multilevel selection theory. Psychological altruism refers to a motivational state with the ultimate goal of increasing another's welfare. The basic philosophical question about whether altruism is part of human nature concerns psychological altruism, not evolutionary altruism. Recent experimental evidence suggests that empathy-induced psychological altruism is indeed part of human nature.

References

Axelrod R and Hamilton WD (1981) The evolution of cooperation. *Science* **211**: 1390–1396.

Batson CD and Shaw LL (1991) Evidence for altruism: toward a pluralism of prosocial motives. *Psychological Inquiry* **2**: 107–122.
Campbell DT (1975) On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist* **30**: 1103–1126.
Dawkins R (1976) *The Selfish Gene*. New York, NY: Oxford University Press.
Eisenberg N and Miller P (1987) Empathy and prosocial behavior. *Psychological Bulletin* **101**: 91–119.
Hamilton WD (1964) The genetical theory of social behavior (I, II). *Journal of Theoretical Biology* **7**: 1–52.
Mansbridge JJ (1990) *Beyond Self-Interest*. Chicago, IL: University of Chicago Press.
Sober E and Wilson DS (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Trivers RL (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57.
Wilson EO (1975) *Sociobiology: The New Synthesis*. Cambridge, MA: Belkap Press of Harvard University Press.

Further Reading

Batson CD (1991) *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Erlbaum Associates.
Batson CD (1998) Altruism and prosocial behavior. In: Gilbert D, Fiske S, and Lindzey G (eds) *The Handbook of Social Psychology*, vol. 2, pp. 282–316. New York, NY: McGraw-Hill.
Hardin G (1977) *The Limits of Altruism: An Ecologist's View of Survival*. Bloomington, IN: Indiana University Press.
Hoffman ML (1981) Is altruism part of human nature? *Journal of Personality and Social Psychology* **40**: 121–137.
MacIntyre A (1967) Egoism and altruism. In: Edwards P (ed.) *The Encyclopedia of Philosophy*, vol. 2, pp. 462–466. New York, NY: Macmillan.
Oliner SP and Oliner PM (1988) *The Altruistic Personality: Rescuers of Jews in Nazi Europe*. New York, NY: Free Press.
Piliavin JA and Charng H-W (1990) Altruism: A review of recent theory and research. *American Sociological Review* **16**: 27–65.
Wilson DS and Sober E (1994) Re-introducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences* **17**: 585–608.

Human Cognition

Intermediate article

H Plotkin, University College London, London, UK

CONTENTS

Human cognition is the set of processes and mechanisms by which we come to know the world, whether it be through learning, perception, or thinking and problem solving. Studies of cognition as an evolved set of traits which must be understood within a developmental framework are beginning to resolve some of the oldest philosophical problems of how we come to have knowledge.

INTRODUCTION

Epistemology is concerned with how knowledge is possible, and the certainty that we can attribute to that knowledge. How do we come to know anything; and what confidence can we have that what we know relates with any accuracy to the state of the world? Within classical philosophy, there is no more fundamental distinction and dispute than that between rationalist and empiricist approaches to epistemology. Philosophically, it is the oldest and most important difference of view that we have about how humans come to have knowledge. It is not surprising, therefore, that an almost exact distinction and dispute has been evident in psychological approaches to the understanding of human knowledge. It is equally unsurprising that cognitive science, the science of knowing, now offers the possibility of both theoretical and empirical resolution of this oldest of philosophical issues.

THE RATIONALIST–EMPIRICIST DISTINCTION

In its strongest form, rationalism argues that reason is the only, the unique, path to knowledge. In less strong form, the claim is that reason has precedence over other ways of acquiring knowledge. The oldest, and certainly most famous, of rationalists was Plato. The world to be known was for him that which existed as abstract objects that he called 'forms' or 'ideas', which are independent of thought. The forms are changeless and incorporeal, and the only way that we can come to have true knowledge of them is through thought. Knowledge that comes to us through our senses is only ever partial and always uncertain. True knowledge is an imminent property of every human child at birth – knowledge is innate – and it becomes accessible through long and patient study of subjects such as logic and mathematics. More recent philosophical exponents of rationalism have been Descartes, Spinoza, and Leibnitz.

An almost exactly opposite position is that of the British empiricists such as Locke, Berkeley, and Hume. Empiricists rejected the notion of innate knowledge. For them knowledge begins at our sensory surfaces and is based on what we experience of the world. In Hume's words, there is 'no idea without an antecedent impression'. It was John Locke in particular who attacked the notion of innate ideas and knowledge. The mind at birth, he argued, is a *tabula rasa*, a blank slate upon which experience writes, and experience begins with sensation. The empiricists were much influenced by the science of their day, especially that of Newton, and, of course, science begins with observation. Hence the empiricists developed Newtonian conceptions of the knowing mind, a mechanics of mind, in which knowledge is particulate and ruled by temporal and spatial proximity. The empiricists were, in effect, associationists seeking for laws of knowledge through an understanding of the lawful interaction of ideas whose origins are in sensory impressions.

RATIONALISM AND EMPIRICISM IN TWENTIETH-CENTURY PSYCHOLOGY

The founders of psychology as a natural science in the mid-nineteenth century were strongly

influenced by empiricist philosophy. The earliest science of knowledge, therefore, was inclined towards an empiricist skepticism as to the existence of innate knowledge. While Darwin's theory of evolution was published and widely discussed at about the same time that a scientific psychology was being established, evolution had remarkably little direct impact upon the new science of mind. It did, though, give rise to two movements that appeared at the start of the twentieth century, both originating in the proposed evolutionary continuity between species. The first led to the seeking of instincts in humans, instincts whose origins were assumed to lie in the history of natural selection pressures acting upon the minds of the ancestors of modern humans; and the second resulted in attempts to establish the existence of, previously largely unrecognized, rational powers of learning and thought in nonhuman animals.

In pursuit of the first of these movements, literally thousands of instincts were claimed to shape almost every aspect of human psychological functioning, such instincts comprising a form of innate knowledge that causes the human mind to function in specific ways. Much of this work was feeble speculative description without emprirical support which collapsed in the face of two developments. One of these, the outcome of the movement seeking rationality of some form in nonhumans, was the laboratory studies of animals by Thorndike and Pavlov, both presented within the framework of empiricist associationism, and convincingly supported by solid experimental evidence. If animal behavior is driven by general learning processes based on laws of association, why, it was asked, should this not apply also to humans? The other, related, development was the onslaught of a fierce environmentalism in all the social sciences that in psychology took the calamitous form of behaviorism. The legacy is what Tooby and Cosmides (1992) called the 'standard social science model' (SSSM) that came to dominate psychology for well over half a century, and which in Fodor's (1998) words 'takes a form of empiricism for granted: human nature is abritrarily plastic'.

Initially, the onset of the cognitive revolution in the late 1950s and early 1960s did not change the SSSM. It provided psychology with the powerful conceptual tool of invoking directly unobservable psychological constructs as causes of human behavior, constructs such as attention and different forms of memory. But under the influence of Piaget's powerful form of cognitivism, which had never conformed to the absurd strictures of behaviorism, cognitive theory remained focused upon general processes that wrote upon the blank slates of the human mind. It was only in the 1960s and onwards, and especially in the last two decades of the twentieth century, that the tide began to turn, with rationalism, and its basic tenet of innate knowledge, making a slow return to center stage in theories of human cognition. Three principal developments have been responsible for this. One is the ethological account of animal learning. The second concerns human language. Finally, recent studies of cognitive development in children have been profoundly important in changing the weight of opinion.

THE CHANGE IN VIEW

One group of scientists studying behavior had never relinquished the notion of instinct or innate behavioral tendencies. These were the ethologists, biologists who studied mainly animal behavior in natural settings and whose principal conceptual tool in explaining behavior, unlike psychologists, was evolutionary theory. Through the 1950s and in to the 1960s, ethologists, with their emphasis upon evolved behavioral dispositions, debated fiercely with psychologists, whose concepts inclined to center on flexible, learned, behaviors. This was but one more phase in the 'nature–nurture' argument which had been debated for almost a hundred years.

Attempting to resolve the issue, one of the founders of ethology, Konrad Lorenz, wrote a seminal monograph (Lorenz, 1965) in which he argued that, unnoticed by psychologists, one of the characteristics of learning is that it is almost always adaptive in outcome. This, he argued, can be explained only by learning itself being an adaptation, a product of evolution vested in evolved and innate neurological mechanisms, which he called 'innate teaching mechanisms'. This collapsed the dichotomy of nature and nurture into a singularity, an instinct (or instincts) for learning about specific features of the world.

At the time Lorenz had no hard evidence that learning in any species is tuned to acquiring specific forms of knowledge. Within a few years, however, laboratory-based evidence supporting this view began to accumulate. For example, from studies through the 1970s and beyond, it became clear that learning of song in songbirds is tightly constrained, resulting in powerful biasing towards species-specific song acquisition; hummingbirds were found to acquire a win-shift strategy more easily than a win-stay strategy that fits well with their natural foraging behavior; and sex differences

in spatial learning ability in different species of vole matched their reproductive strategies. Such findings gave rise to the general notion of learning by instinct (Gould and Marler, 1987), learning as evolved, adaptive, cognitive specializations constrained to operate within specific domains of information.

To echo the question about general learning processes of a half-centry earlier, but with an inverted meaning, if domain-specific learning, bearing all the hallmarks of evolved adaptive mechanisms for cognition, occurs in nonhumans, might this not apply to human cognition? Part of the answer came from studies of language learning in humans.

Language is a rule-based system of communication in humans that is defined by the quality of generativity: namely, the capacity to generate a virtually infinite number of messages from a finite, usually quite limited, number of component symbols. Despite claims that following intensive training a few chimpanzees show glimmerings of a language facility, it is widely accepted that such behavior has never been observed in nonhuman apes under natural settings, and that language is a human-specific cognitive characteristic. Earlier accounts of language learning, based either on associative learning principles or more general developmental processes, were eclipsed from the late 1950s onwards by the linguistic theory of Noam Chomsky. Though Chomsky's views have altered somewhat over the decades (Chomsky, 2000, represents his most recent position), his broad view as it affects theories of human cognition in general has remained constant: language is an innate organ of mind, the acquisition of which is driven by language-specific cognitive mechanisms. The evidence for this is that language acquisition, whichever of the world's five thousand plus languages is being learned, is remarkably constant across children both in terms of content (vocabulary) and structure (grammar and syntax); that this applies equally to profoundly deaf children who are raised within a signing environment and who communicate through a language based on hands and eyes rather than tongues and ears; and that rich linguistic structures invariably develop, even within linguistically impoverished environments. Recent brain imaging studies have shown that the same areas of the brain are activated by seeing sign language in deaf individuals expert in sign language as are activated when hearing people are listening to spoken language.

Nobody doubts, of course, that language is *learned* in the sense that children raised in an environment which has German as its native language do not end up speaking Japanese. But the strong claim from the evidence is that all children come into this world with innate knowledge of a single, universal, language which then unfolds within specific linguistic environments into particular versions of that language. The similarity to Lorenz's innate teaching mechanisms is striking. Two points, though, should be noted. One is that the Chomskian account is not universally accepted. Deacon (1997), for example, offers an alternative, more general-process view, which nonetheless is compatible with evolutionary approaches to language. The second point is that Chomsky himself, while always taking an unbending nativist (rationalist) stance, does not associate himself with an evolutionary account of language. Others (for example Pinker, 1994) have taken the view that an innate organ of mind like language must be a product of evolutionary processes.

What then of more general studies of cognitive development in infants and children? What follows is necessarily the briefest of sketches. Detailed reviews can be found in Hirschfeld and Gelman (1994) and Sperber *et al*. (1995). The methodological advance that resulted in much of this work came with the refinement of the dishabituation paradigm in the early 1980s. Habituation is a reduction in response to a familiar event, while dishabituation is an arousal of attention to an unexpected event.

Take the paradigmatic case of an infant placed on the lap of one of its caretakers and facing a screen. At the center of the screen is a ball, and after a few seconds a second ball appears at the edge of the screen, moves towards the stationary ball at the center, strikes it, and comes to a stop while the previously still ball is propelled off the screen. Such a launching display elicits little interest from the infant. But if the second ball comes to a stop some distance from the centrally placed ball, and the stationary ball then moves off the screen – an example of action at a distance which is never observed in inanimate objects – the infant stares long and hard at the screen. These so-called looking-time experiments have yielded a wealth of data taken at ages previously inaccessible to cognitive developmentalists. Other examples involving infants entail blocks being placed on other blocks which support them (the unsurprising experience), contrasted with blocks placed next to other blocks, unsupportd by the latter but remaining suspended in mid-air (the surprising experience), or occluding bodies with holes in them that do or do not reveal the objects over which they pass. The general finding has been that infants as young as 10 or 12 weeks, who have

spent much of their postnatal life sleeping, have a knowledge of 'intuitive' physics. That is, they are suprised by physical events when those events are contrary to the macrobehavioural events of the inanimate world perceived directly by our senses.

There are two opposed interpretations of such findings. One is that infants have had insufficient experience to have learned about the behavior of inanimate objects in so short a period, and that these experiments demonstrate the existence of innate knowledge of intuitive physics. The other interpretation is that two to three months, even with limited sensory capacities and periods of wakefulness, is a sufficient period for infants to learn, using general learning mechanisms, about the properties of the physical world. The former is of course, a form of rationalism; the latter is the empiricist stance.

Before evaluating these alternatives, consider some other findings of the last 20 years of the twentieth century, the most important of which concerns 'theory of mind' (ToM). ToM is generally held to underly knowledge of social causation, as opposed to the physical causation of intuitive physics. ToM is the means by which each of us, pathology apart, come to understand that others have minds made up of intentional mental states such as knowing, feeling, and wanting. The sequence by which this occurs appears to be invariant, though the contents of ToM are clearly culture-specific. During the first eight months after birth the infant appears to inhabit an egocentric sensory-motor world. From about 9 months, the previously dyadic interactions of self–other or self–object gives way to the triadic interaction of self–other–object. The infant begins to share attention with others, follows the gaze of others, and checks their gaze to ensure they are looking at the same thing. By about 15 months, declarative pointing appears; between the ages of 18 and 24 months pretend play becomes evident, and vocabulary begins to embrace mental state terms such as 'want' and 'know'. The defining point in the completion of ToM occurs at between 40 and 50 months when children come to understand that others can have false beliefs, which among other things means that others may have intentional mental states that are different from their own. Impairment to the normal development of ToM is thought to result in serious pathology, including the spectrum of autistic disorders.

The comparative literature indicates that while chimpanzees are able to act on the basis of what other chimpanzees know, this cannot be taken to mean that these animals are able to attribute mental states to others. Until there is evidence to the contrary, the assumption is made that ToM is a uniquely human adaptation essential for the human capacity for culture. Both functional imaging and brain lesion studies point to specific regions in the frontal and temporal lobes as being a part of the neurological substrates of ToM.

Other domain-specific human cognitive capacities that have been postulated to exist include various aspects of vision, including face recognition, hearing, the detection of cheaters, responses to sexual stimuli, comprehension of living forms (intuitive biology), and perhaps even the understanding of supernatural beings. Duchaine *et al*. (2001) provide a recent review.

ADDITIONAL CONSIDERATIONS

It is a truism of comparative neuroanatomy that the brains of vertebrates share a common ground plan. It is equally widely held that the organization of the forebrain, specifically the neocortex, in terms of major processing areas for sensory inputs and motor outputs, is both conserved across all mammals yet shows a degree of specialization for particular taxonomic groups (de Winter and Oxnard, 2001). This is strongly supported by recent genetic studies on phylogenetically widespread genes coding for the structure of brains, including those of invertebrate species. The general conclusion is that the brains of vertebrates, and especially the neocortex of mammals, are organized for some degree of specialized, localized function, such specialization being genetically part caused and partly emerging through normal developmental sequences, including activation of neural pathways. The human brain is expected to display at least a similar degree of structural and functional specialization. As indicated above with regard to ToM, though the general finding goes much wider than mental state attribution, localization of function across individuals is becoming a common finding across a range of cognitive functions. The brain is not equipotent under conditions of normal development, though this does not deny the extraordinary capacity of the brain for flexibly adjusting to abnormal conditions, whether these be the result of pathology or developmental circumstances.

INTERPRETATIONS OF MODULARITY

The modularity thesis asserts that the brain comprises a set of computationally specialized, informationally relatively encapsulated, automatic and rapid processors of specific domains of

information. When the thesis was first developed by the philosopher of mind Jerry Fodor in 1983, the emphasis was placed on modules as restricted processors of input: for example the conversion of a two-dimensional retinal array into three-dimensional visual images. In the years since then, the modularity conception has been widened, often to include whole psychological functional domains (such as reproductive strategies or reciprocity of social interactions in the sharing of resources), as well as more restricted sensory problems (like face recognition). This is sometimes referred to as 'the massive modularity thesis', and it has found special favor with evolutionary psychologists who subscribe to the view that the human mind, including cognition, is a set of specialized, evolved adaptations. Another variant of the modularity concept is that modularity is a product of development.

Whether it is the original, limited, Fodorian thesis of modularity, the more recent massive modularity espoused especially by evolutionists, or the developmental variation of the modularity conception, almost all cognitive psychologists now subscribe to one or other version of the modularity thesis. Where the deep divisions remain is in the interpretation of modularity and its causes. As indicated, there are three schools of thought, with the distinctions drawn usually being a matter of emphasis and degree, rather than absolute difference and theoretical exclusion. For example, no one denies that what humans are now is a result of the evolutionary history of our species. But just what those evolutionary processes were, and how specific have been their effects on human cognition, are the issues of contention. In general, these, often greatly exaggerated, differences revolve around the meaning attached to the concept of innateness, to a commitment to more rigid or more flexible developmental approaches, and adherence to specific evolutionary theory.

The first two schools of thought are both strongly nativist, and hence rationalist. The one, the strong evolutionary psychology school, maintains that human cognition comprises a set of specialized information-processing modules each with specific, innate computational features whose origins lie in past selection for specific cognitive functions which evolved through neo-Darwinian evolutionary processes. Pinker (1994) exemplifies this view. The second, characterized as the new rationalism (Fodor, 1998), espouses a strong form of innatism for some cognitive processes and mechanisms with tacit acceptance of genetic part causes of these; but is neutral, if not hostile, towards neo-Darwinist processes as the evolutionary origins of cognitive traits, that is of the manner such processes and mechanisms become manifest as specific forms of cognition.

The third position challenges innatist arguments, especially with regard to representational nativism, adopts a strong developmental (ontogenetic) stance, and offers connectionist neural network models as potential general processes underlying cognition. It is best described as a form of developmental empiricism. The mechanisms may be innate, and hence evolved, but the cognitive functions that they mediate are not based on specific contents, such as linguistic structure or facial configuration. Modularity emerges through a complex cascade of developmental events acting upon connectionist networks. Such networks, initially relatively untuned, act as holistic, parallel processors of information responding to specific informational inputs and correction procedures to the outputs of the networks, resulting eventually in networks trained to process specific forms of information.

SUMMARY: WAITING FOR EMPIRICAL RESOLUTION

The comparative, genetic, and neurological evidence all points to humans not being blank slates at birth. The brain, including the neocortex, does show a degree of specialized structure and function at birth, if only in terms of gross connectivity. The impasse in the argument as to whether such young infants come to have so much knowledge so early in their lives, and pass through relatively invariant developmental stages, because of relatively fixed innate causes or because of more flexible ontogenetic processes, will not be broken by cognitive developmental studies alone, unless there is a further methodological breakthrough in cognitive developmental procedures. Exactly how specialized in function the brain is at birth, just what extent of neural structure can be traced to specific genetic causes, how much flexibility there is in developmental pathways, and how much postnatal cognitive specialization can be attributed to more general developmental processes, are questions that all await empirical resolution. But it will be outside of the strictly cognitive experimental realm. There are, though, already signs that a resolution will be reached. For example, there is recent evidence that visual experience during the first few months of life is necessary for establishing the neural architecture that specializes in the expert processing of faces over the next ten or more years.

The answers will come in the next decades with increasing knowledge of neurogenetics, of developmental neuroscience of real neural networks, and the inevitable unfolding of understanding of the connections between brain structure and function and psychological mechanisms. What is becoming clear is that cognitive science is telling us that neither rationalists nor empiricists are entirely correct. We do come into the world with the slates of our knowing minds written upon; but how much is written and where that writing is coming from remains to be determined.

References

Chomsky N (2000) *New Horizons in the Study of Language and Mind*. Cambridge, UK: Cambridge University Press.

Duchaine B, Cosmides L and Tooby J (2001) Evolutionary psychology and the brain. *Current Opinion in Neurobiology* **11**: 225–230.

Fodor J (1998) *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. Cambridge, MA: MIT Press.

Gould JM and Marler P (1987) Learning by instinct. *Scientific American* **256**: 62–73.

Hirschfeld LA and Gelman SA (eds) (1994) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge, UK: Cambridge University Press.

Lorenz K (1965) *Evolution and Modification of Behaviour*. Chicago, IL: University of Chicago Press.

Pinker S (1994) *The Language Instinct*. London, UK: Allen Lane.

Sperber D, Premack D and Premack AJ (eds) (1995) *Causal Cognition*. Oxford, UK: Clarendon Press.

Tooby J and Cosmides L (1992) The psychological foundations of culture. In: Barkow JH, Cosmides L and Tooby J (eds) *The Adapted Mind*, pp. 19–136. Oxford, UK: Oxford University Press.

de Winter W and Oxnard CE (2001) Evolutionary radiations and convergences in the structural organization of mammalian brains. *Nature* **409**: 710–714.

Further Reading

Baron-Cohen S, Tager-Flusberg H and Cohen DJ (eds) (2000) *Understanding Other Minds*. Oxford, UK: Oxford University Press.

Deacon T (1997) *The Symbolic Species*. London, UK: Allen Lane.

Elman JL, Bates JA, Johnson MH, Karmiloff-Smith A, Parisi D and Plunkett K (1996) *Rethinking Innateness*. Cambridge, MA: MIT Press.

Fodor J (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Hickock G, Bellugi U and Klima S (1998) The neural organization of language: evidence from sign language aphasia. *Trends in Cognitive Sciences* **2**: 129–136.

Lillard A (1998) Ethnopsychologies: cultural variations in theories of mind. *Psychological Bulletin* **123**: 3–32.

Petitto LA, Zatorre RJ, Gauna K, Nikelski EJ, Dostie D and Evans AC (2000) Speech-like cerebral activity in profoundly deaf people processing signed language. *Proceedings of the National Academy of Sciences of the USA* **97**: 13961–13966.

Plotkin H (2002) *The Imagined World Made Real: Towards a Natural Science of Culture*, London, UK: Allen Lane.

Stuss DT, Gallup GG and Alexander MP (2001) The frontal lobes are necessary for theory of mind. *Brain* **124**: 279–286.

Tomasello M (1999) *The Cultural Origins of Human Cognition*. Camridge, MA: Harvard University Press.

Human Factors and Ergonomics

Intermediate article

Peter A Hancock, University of Central Florida, Orlando, Florida, USA
Raja Parasuraman, Catholic University of America, Washington, DC, USA

CONTENTS

Human factors and ergonomics is the study and practice of designing technology around the capabilities, both physical and mental of the user. Included in the chapter are brief descriptions of neuroergonomics, the combining of neuroscience and ergonomics, and human factors and ergonomics in the twenty-first century.

INTRODUCTION

Most individuals reading an *Encyclopedia of Cognitive Science* will have more than a passing familiarity with the research areas and issues that they encounter in the text. Yet this may not be true for *human factors* and *ergonomics* (*HF/E*), since these terms appear to lie outside the mainstream of cognitive science. However, we hope to persuade the reader that HF/E represents a central concern of any cognitive science that aspires to be relevant to the real world. We support this claim by beginning with a reference to the origins of cognitive psychology and its historical antecedents.

Cognitive Psychology and Human Factors and Ergonomics

Most modern cognitive scientists are unconcerned with, or not fully cognizant of, the goals, issues and findings of HF/E. This relative neglect is understandable, but paradoxical, since the post-Second World War origin of modern cognitive psychology owes much to HF/E. The 'cognitive revolution' of the late 1950s supplanted the behaviorist tradition that dominated American (and, to a lesser extent, European) psychology. This paradigm shift is rightly seen as the beginnings of the cognitive science movement, accompanied by developments in artificial intelligence (Newell and Simon, 1956) and linguistics (Chomsky, 1957). Two of the leading figures of this cognitive revolution were George Miller in the USA and Donald Broadbent in the UK. The latter can be seen in action at the prime of his career in Figure 1.

Broadbent's classic information-processing model of the human cognitive system (Broadbent, 1958) inspired many hundreds of cognitive psychologists, and still remains influential today. Both Broadbent and Miller were especially interested in applications of cognitive psychology to real-world problems, and Broadbent in particular was a leader in the early development of HF/E. Many younger cognitive scientists who may be aware of Broadbent's contribution to cognitive psychology may nevertheless be surprised to learn that he also made fundamental contributions to such areas of HF/E as evaluation of the effects of noise, vigilance, stress, and industrial health and productivity. His classic book *Decision and Stress* (Broadbent, 1971) summarized this work and made an eloquent case for continued interaction between basic cognitive psychology and HF/E, a collaboration that we support and seek to sustain for the continued vitality and well-being of each area.

What Are Human Factors and Ergonomics?

Although human factors and ergonomics have different historical antecedents (see Jastrzebowski, 1857), they can now be regarded as essentially synonymous (see Dempsey *et al*., 2000). HF/E concerns the way in which people work with technology, where the latter is a broadly encompassing term that embraces artifacts ranging from something as simple as a paper clip to those as complex as a nuclear power station, as well as environments at work, at home and while in motion. HF/E is both a science and an engineering discipline. The science of HF/E consists of developing theories, empirical

Figure 1. Donald Broadbent the motorcyclist at the height of his powers. Many cognitive scientists remain unaware of Broadbent's fundamental contributions to human factors and ergonomics.

databases and quantitative models of human physical and cognitive capabilities and limitations in relation to the use of technology. The goal of HF/E as an engineering discipline is to design technologies and environments for effective and safe human use. Unfortunately, many products and processes are designed with little if any attention to such human capabilities and limitations. Any cognitive scientist who has struggled to learn to use a new software interface, or any individual who has been fooled by the complexity of action modes of a TV remote control, or any older person who is puzzled by the mystery of 'butterfly' voting ballots will recognize this problem immediately. More often than not we find ourselves unable to use technologies effectively because of poor HF/E design. Consequently, one definition of HF/E is 'the branch of science which seeks to turn human–machine antagonism to human–machine synergy' (Hancock, 1997).

One of the notable qualities of HF/E is the range of activities that it embraces. At one level, HF/E is represented by a real-world practice concerned only with engineering and design decisions regarding a manufactured product. Foundation-level information about human performance is taken as the premise for new system design. Often this foundation may be lacking in substance and sophistication, so that many elements of design (e.g. those evident in interface development) may represent more of an art than a logical scientific process. Development is often contingent on 'user preferences', and usability analysis has to tread a fine line between the use of objective measures and subjective response to create new and effective systems. Thus, in the practical world, HF/E lies at the confluence of science and art, where the preponderance of the contribution of each is very much contingent upon the type and training of involved personnel. In this article, however, we shall focus on the *science* of HF/E.

Overview

In what follows, we shall first articulate the science of HF/E through an examination of the relationship between the psychology of action and the application of cognitive science in HF/E. Secondly, we shall examine the process of interaction with particular reference to the control of 'virtual' systems in which the laws of psychology rather than the laws of physics hold primary sway. Thirdly, we shall demonstrate how inherent human abilities and limitations are magnified in designed worlds, and finally we shall promulgate a new interdisciplinary endeavor – neuroergonomics, which seeks a marriage of ergonomics and neuroscience. We shall conclude with an affirmation of the centrality of HF/E to any cognitive science concerned with behavior in real-world settings and to an even greater extent in envisioned worlds.

COGNITIVE SCIENCE, HF/E AND THE PSYCHOLOGY OF ACTION

How can the study of cognition be linked to issues surrounding the human use of technology in the real world? One possibility is to develop modular theories of parts of cognition, and to apply these modules to the solution of real-world HF/E problems. However, since all action is situated, the context of the behavior exerts a crucial influence on the outcome (Flach *et al.*, 1995; Hancock *et al.*, 1995). Thus the strategy of trying to understand behavior in complex worlds through the concatenation of multiple theories of simple behaviors is crucially flawed, as the component interaction of each simple theory and the complexity of the circumstances in which behavior occurs both act to defeat this constructivist aspiration.

An alternative strategy is to begin not with individual components of behavioral capability, but with a more fundamental understanding of the actual environment in which behavior occurs. We are, of course, aware of the direct link of this notion to ecological psychology, from which this persuasion emanates (Gibson, 1966, 1979). Indeed, it is of

more than passing interest to note that some of Gibson's earliest work focused on the very practical HF/E issue of vehicle control, and that his work on driving (Gibson and Crooks, 1938), with its construct of the 'field of safe travel', represents one of the earliest but most crucial contributions to practical driving research even today (Figure 2).

Unlike some more radical ecological theorists, we see technology and the manufactured world as the contemporary ecology to be explained. Technology certainly adds a layer of complexity to understanding, since the world as we experience it is largely a palimpsest of previous design decisions by those who have preceded us. Furthermore, despite the fact that most living humans exist to a considerable extent in a designed environment, their resident capabilities are derived from many millennia of evolution and are not, in and of themselves, intrinsically amenable to sudden change or development. The fact that technology provides magnifications of these perception/action capacities, often incorporating their faults as well as their virtues, is a major design issue (Hancock and Chignell, 1995; Hancock, 1997). Thus anyone who wishes to understand cognitive capabilities cannot ignore the externalization of those abilities into the real world, first as a formative influence and secondly as a clear clue to functional significance. It is in this respect that HF/E is central to cognitive science and in crucial ways expresses the same essential goal. That this facet of science should also express opinions about what should be (i.e. *purpose*) as well as what is (i.e. *process*) is a polemical contention but one that we directly support. If the goal is to understand human behavior, individual and collective purpose cannot be omitted from the equation.

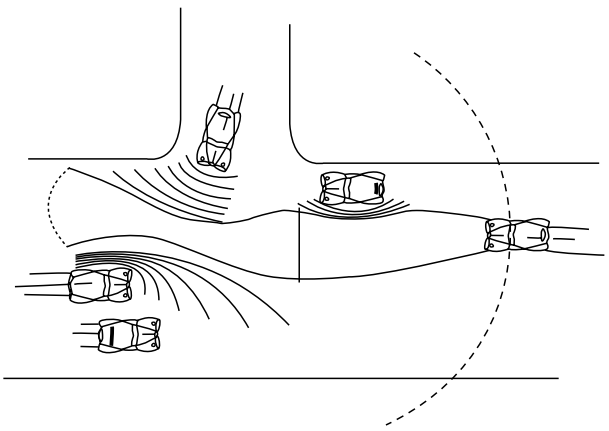


Figure 2. Illustration of the 'field of safe travel' and the 'minimum stopping zone' from the seminal paper by Gibson and Crooks (1938). Like other psychological theorists, Gibson's questions are firmly founded on concern to understand real-world experience.

Cognitive science seeks to understand mind and brain. It is our contention that one method by which such a complex activity can be understood is to observe its outcome. We suggest that tools (or more generally technology) represent collective extensions to cognitive capability, and that the fabricated world around us represents the summed action of collective human cognition. Thus, in much the same way as sophisticated, modern brain-imaging techniques allow us to window the structure and function of individual brains, so examination of the modern world allows a similar glimpse of brains' collective output. Furthermore, since technology represents the most powerful formative force in the modern world, contemporary interaction with technology acts to shape who we will be in the future. Since all behavior is situated in and crucially dependent upon context, and since that context is largely dictated by technology, the way in which humans design, develop, fabricate, interact, perform and occasionally recover from the effects of technology – that is, ergonomics – is central to a dynamic understanding of cognitive science.

The notion that cognition should be considered in relation to action in the world has many antecedents. Piaget's work on cognitive development in the infant and its dependence on exploration of the environment certainly anticipated the concept of situated or embodied cognition. A modern statement of this thesis can be found in the philosopher Andy Clark's engaging book, *Being There*: *Putting Brain, Body and World Together Again* (Clark, 1997). Clark goes beyond the old distinction between mind and matter to examine the characteristics of an embodied mind that is shaped by and helps to shape action in a physical world. If a true cognitive science therefore cannot study minds in isolation, but must examine them in interaction with the physical world, then it is a natural second step to ask how to design artefacts in the world which best facilitate that interaction. This is the domain of HF/E.

HUMAN PERFORMANCE CAPABILITIES AND LIMITATIONS

To illustrate the intimate link between HF/E and cognitive science, we only have to consider how the study of certain processes is common to each. Consider, for example, short-term or working memory. While those in cognitive science seek to explore this

facility and its limitations, those in HF/E take such knowledge and seek ways in which memory limits can be obviated in the real world. A specific example may help to explain this transfer process. During the take-off sequence, an aircraft attains two crucial velocities, labeled V1 and V2 in the aviation realm. Although these are critical speeds, they are not constant even for the same aircraft, since weather conditions, load and other variables act to dictate their exact value on a particular take-off on any given day. Thus V1 and V2 must be calculated during the preflight procedure and held in working memory by the pilot. Clearly, take-off is a critical time and it is best not to overload the pilot at this juncture, since during take-off emergencies things can happen quickly, and briefly memorized but crucial information can be forgotten. One design answer to support pilot memory was 'speed-bugs.' These small markers were attached to the airspeed indicator so that the pilot could position them during preflight and immediately 'see' the velocity values, rather than having to refer to short-term memory. Unfortunately, as with many good ideas, it was thought that if two were good, more would be better, and soon the proliferation of 'bugs' meant that the memory load was returned to the operator, who now had to remember which of the specific attached markers related to which crucial speed value. This is a design situation that is constantly encountered in HF/E, where good original conceptions are distorted by well-meaning but uninformed designers. This was highlighted by the ballot failure experienced in specific counties in Florida during the US Presidential election of 2001 (see Woods and Hancock,2001).

The moment of take-off is also an example of the problem of cognitive or mental workload (Hancock and Meshkati, 1988). An especially difficult construct to define unequivocally, nevertheless mental workload retains high face validity and indeed utility in the practical realm since, like fatigue, many individuals can identify and empathize with the experience. The fundamental question, and certainly one related to the theoretical notions of attentional capacity, concerns the maximum tolerable level of cognitive activity that an individual can sustain, and for how long they can do this. In addition to questions of acute and chronic mental overload and underload, fundamental developments in naturalistic decision-making (see Klein, 1999) have been stimulated by the problem of understanding the real-world choices of novices and experts in different application domains. Similarly, many cognitive scientists might well know Fitts' name because of its association with a fundamental law of movement relating speed to accuracy, without being aware of his central role in HF/E in relation to air-traffic control (Fitts, 1951) or pilot error (Fitts and Jones, 1947). Indeed, as well as the early referenced work on car driving, Gibson's fundamental theories originated from his experiences with practical aviation problems during the Second World War. More recent examples abound. The multiple resource theory of attention promulgated by Wickens has been used extensively as a design heuristic, especially in aviation psychology, a field to which Wickens himself has made a number of fundamental contributions. Finally, the present authors are each intimately concerned with both theoretical aspects of cognition and their application in diverse applied settings (Parasuraman and Mouloua, 1996; Hancock, 1997). In summary, there is no part of cognitive science which does not have its counterpart and leave its mark in the applied world, and we submit that this information flow is strongly bidirectional, and should be even more so.

The study of human performance capabilities and limitations must also be considered in relation to the ecological validity of the environment (e.g. the laboratory) in which performance is elicited. Cognitive science methods, such as laboratory experiments, computational modeling, etc., do not necessarily make intimate contact with human performance in real settings. We recognize the value of such basic research, but suggest that it must be supplemented by research which examines cognition in relation to the work setting and technology, not just individual cognition in isolation. Of course, there has been an ecological validity movement in cognitive science for some years now (Hutchins, 1995), and situated cognition theorists do consider the environment, but even then, if the environment is simulated in the laboratory at a very basic level, the theoretical principles that emerge may not apply to the real world. This necessarily means that some of the artificial laboratory tasks of basic cognitive science may not be generalizable to complex real-world systems.

An example will illustrate the value of a combined basic and applied research agenda. Broadbent's (1958) classic theory of short-term memory and attention was developed following observations of the fallibility of air-traffic controllers in handling multiple messages. He studied the real-world problem, abstracted it in a laboratory task and developed a theory. Irrespective of the current validity of Broadbent's theory, its heuristic power for cognitive psychology research is undeniable.

Thus real-world problems can advance empirical studies and theory in cognitive science. The theories in turn, if applied judiciously, can impact on real-world problems. This interactive relationship between basic cognitive science and HF/E was present in the early days of both disciplines, but has perhaps been lost over the years and needs to be revived.

HUMAN INTERACTION WITH TECHNOLOGY

One of the major issues in HF/E continues to be human interaction with complex systems. In this context, 'complex' usually refers to large-scale process control systems such as nuclear power or advanced computer-based systems (e.g. those found in aerospace and aviation operations). Spectacular failure in these areas has always provided an impetus for gaining a greater understanding of what are often conceived of as human failures. In recent decades, these systems have achieved such complexity that some of them simply cannot be operated without computer support, and their momentary control implies the presence of computer mediation. In HF/E, the question of function allocation has been a continual concern (Hancock and Scallen, 1998). In a human–machine or human–computer system, who does what when is a central design and operational issue. As more automation has been introduced, and as the human operator has become progressively more divorced from momentary, hands-on control, the question of interaction with such semi-automated systems has become of paramount importance (Parasuraman and Riley, 1997).

Automation can be defined as the execution by machine (usually a computer) of a function previously performed by a human (Parasuraman and Riley, 1997). Given their rapid growth in speed, capacity and 'intelligence', computers are increasingly being assigned functions that at one time could only be performed by humans, including complex cognitive activities such as decision-making and planning. Automation has also been extended to functions that humans cannot perform as accurately or reliably as machines. Consequently, automated systems are now common in many areas, including air, ground and maritime transportation, medical systems, manufacturing and process control (Parasuraman and Mouloua, 1996). Further reductions in the size and cost of computers will result in far-reaching applications of automation to virtually all aspects of life (Rawlins, 1996; Satchell, 1998).

The advent of these technologies has been accompanied by growing research on human interaction with automation. An important finding to have emerged from this research is that automation can fundamentally change the nature of the cognitive demands and responsibilities of the human operators of systems, often in ways that were unintended or unanticipated by the designers (Wiener and Curry, 1980; Bainbridge, 1983; Woods, 1996; Parasuraman and Riley, 1997). Much of this work has examined the factors that promote or limit effective use of automation by humans, and the consequences for system efficiency and safety. These findings have raised the issue of the degree to which automation should be implemented in a given system. Given the capabilities of current and projected automation technologies, which system functions should be automated and to what extent? For example, should a computer be responsible for virtually all aspects of decision-making, allowing the operator only a limited veto? Or should the computer only be allowed to generate potential decision options, leaving the choice to the human? Specific aspects of automation have definite consequences – both good and bad – for human performance. Understanding these consequences is critical to the design decision to automate. Several other factors, including risk, reliability, safety, ease of system integration, implementation cost and liability, are also important (Parasuraman *et al*., 2000).

The developing knowledge base of human performance research can be used to design more effective automation. All too often automation has been designed with purely technical considerations in mind. This might have been justified at one time, given the relative paucity of human performance data on the use of automation, but that is no longer the case. There is now an extensive body of empirical work on human performance in automated systems. Details of this growing knowledge base, which includes cognitive theory, laboratory experiments, simulator studies, field studies, incident analyses and accident investigations, can be found in original sources and more detailed expositions (Rasmussen, 1986; Parasuraman, 1987; Wiener, 1988; Sheridan, 1992; Billings, 1997; Parasuraman and Riley, 1997; Sarter *et al*., 1997; Wickens *et al*., 1998).

Research on automation also illustrates how commonalities between basic and applied research can be revealed to their mutual benefit, even though the researchers involved may be ignorant of each other's work. The problems with high-level automation are that human operators tend not to

monitor the automation (exhibiting so-called 'complacency'), they may lose their overall 'picture' (or 'situation awareness') of the system, and their manual cognitive skills may degrade over time (Parasuraman and Riley, 1997). One solution to this problem that has been advocated by applied researchers is adaptive function allocation, in which the automated task is returned to the operator for manual performance for a brief period of time, a procedure that has several benefits (Parasuraman *et al*., 1996). The explanation is that people are more aware of their environment and can respond better to anomalies when they are actively involved in the system than when they are simply passively monitoring the system. This phenomenon has a parallel in the basic cognitive science literature on memory, as pointed out by Farrell and Lewandowsky (2000), namely the so-called generation effect (Slamecka and Graf, 1978). This refers to the improvement in memory performance when subjects have to self-generate the second half of a word pair (e.g. 'Short-T----?') compared with when they read the word pair 'passively' (e.g. 'Short-Tall') and do not have to generate a word. The reading condition, which can be regarded as a form of 'automation' because the second word is 'produced' for the subject without him or her having to do anything, leads to poorer memory performance for the word pairs compared with the generation condition, just as automation does in comparison with manual performance. Farrell and Lewandowsky (2000) accordingly developed a connectionist model of complacency and adaptive automation that had as its basis the improvement in memory strength that self-generation (or manual performance) provides. We would venture that most basic memory researchers in cognitive science are unaware of HF/E research on human–automation interaction, and perhaps vice versa. However, as this example illustrates, the two groups might do well to be cognizant of each other's work.

COGNITION AND THE PHYSICAL ENVIRONMENT

If our central thesis concerning the primacy of context for understanding behavior is accepted, then one of the obvious sequelae is that, contingent upon the present state of knowledge, one can design and fabricate environments that foster optimal performance. Up to the present time, and largely in the area of ergonomics, this idea has only been expressed as a form of defense (i.e. how to protect the individual from harmful circumstances). One of the more entertaining ways to understand this protection was suggested by Haddon (1970), who referred to dangerous environments as 'ecological tigers'. Haddon went on to describe such tigers as the uncontrolled escape of ranges of energy. In particular, kinetic, thermal, radiative and chemical forms of energy were a major concern. In typical HF/E handbooks, these are dealt with as individual concerns, such as heat stress (Hancock and Vasmatzidis, 1999), with some general overviews of system safety (Karwowski and Marras, 1999). As the currency of work has evolved from the erg to the byte, concern about major forms of physiological stress in HF/E has declined somewhat, although modern problems of manual materials handling are unfortunately alive and well, and cost the modern world many billions of dollars each year (Karwowski and Marras, 1999). Although these physical issues have been partly superseded by concerns in *cognitive ergonomics*, in the developing world ergonomics is still fundamentally considered to be the science of physical work.

If, in the past, protection in the industrial world has been concerned with physical sources of threat, contemporary concerns are much more directed towards factors that interfere with optimal productivity in more sedentary, computer-mediated work. Thus most recent ergonomic concerns have been expressed in relation to office design, such as repetitive strain trauma, seating issues, glare, lighting and the many other environmental concerns that relate to the modern workstation environment (Leuder and Noro, 1994). Those still concerned with physiological rather than psychological conditions have focused on issues of performance in exotic and threatening conditions, such as productivity in outer space and in the deep ocean. Despite our advances and sophistication, many people still have to perform hard, dirty, unpleasant physical tasks round the clock, and scientists involved in the traditional aspects of ergonomics still do a tremendous service in seeking to improve the work environment of such individuals. If protection from physical harm is the history of ergonomics, direct interaction with specific brain function is the future, and it is to this that we shall now turn.

NEUROERGONOMICS

If the view that cognition occurs in specific environments and is therefore 'situated' should not be surprising, then the realization that cognition is *embodied* should also not ruffle any feathers, and of course the brain provides the controller for the

embodiment. Until the 1980s cognitive psychology did not pay much attention to the brain. In the 1950s and 1960s, many cognitive scientists pinned all their hopes on understanding human cognition on artificial-intelligence-inspired models. Their rallying cry was that the mind is software which must be understood, irrespective of the hardware that implements it. Accordingly, the rules of mind could be as easily studied in a computer as they could in a human. The actual hardware – the structures and mechanisms of the brain – was unimportant. Similarly, HF/E has been blissfully unconcerned with the brain, even as cognitive psychology has abandoned its previous hostility to brain research and embraced the new paradigm of cognitive neuroscience (Parasuraman, 1998b; Gazzaniga, 2000).

We believe that this neglect is shortsighted. One of us (Parasuraman, 1998a, 2002) has therefore coined the term 'neuroergonomics' to refer to the inclusion of neuroscience in HF/E. Neuroergonomics can be defined as the study of brain and behavior at work. Traditionally, ergonomics has not paid much attention to neuroscience or to the results of studies of the brain mechanisms underlying human perceptual, cognitive, affective and motor processes. To paraphrase the philosopher Mario Bunge (1980), until recently psychology (and HF/E) has been 'brainless', whereas neuroscience has been 'mindless'. At the same time, neuroscience and its more recent offshoot, cognitive neuroscience, have been only partially concerned with whether their findings bear any relation to human functioning in real (as opposed to laboratory) settings, with the exception of applications to clinical disorders. Neuroergonomics is a response to this twin disregard.

To the extent that cognitive neuroscience advances theoretical knowledge of human functioning, it can influence the application of that knowledge to the design of systems. At the same time, HF/E may provide an avenue for examining the practical utility of basic findings generated by cognitive neuroscientists. There are several examples of work that can be characterized as falling within the rubric of neuroergonomics. For example, knowledge of the brain mechanisms and neurochemical systems that control circadian rhythms could be used to devise optimal schedules for shift work, or to minimize circadian disruption due to travel across time zones. This is a pressing practical problem, given that we are moving towards a 24-hour society in which up to one-third of the labor force is involved in some form of shift work.

A second example is the use of functional brain-imaging measures of cerebral blood flow to index human operator mental workload. Event-related, functional magnetic resonance imaging (fMRI) studies of visual attention (Corbetta *et al*., 2000) and working memory (Jiang *et al*., 2000) have revealed the temporal dynamics (to a resolution of the order of a few seconds) of neural activity in localized cortical regions that mediate the component operations associated with these domains of cognition. Such techniques are highly suited to the problem of evaluating mental workload during complex, multitask performance. A recent study by Peres *et al*. (2000), in which fMRI was used to evaluate the cognitive strategies of expert pilots during a simulated aviation task, provides an example of this type of research. We anticipate that there will be many more such studies in the near future. Functional brain-imaging studies of workload can considerably aid the HF/E goal of optimizing the design of human–machine systems for safe and efficient use by human operators.

HF/E IN THE TWENTY-FIRST CENTURY

Neuroergonomics represents an emergent subdiscipline of HF/E that we envisage will be prominent in the new millennium, and will take advantage of the startling pace of new developments in cognitive neuroscience. More generally, we anticipate a growing realization that brain and behavior must be examined both in relation to technology and in context. Given that computer technologies are likely to continue to predominate in the future, work on human–computer interaction will remain at the forefront of research and practice. As computers continue to increase in power and decrease in size, their omnipresence – in the daily artifacts of everyday use, in our clothes, in our modes of transportation and at home – will necessitate new conceptualizations of computers as information appliances (Norman, 2000), and new ways of examining how best to design them for effective use by humans.

SUMMARY

Through the themes that we have developed and prosecuted here, we hope to have persuaded even the most theoretical of cognitive scientists that an excursion into the world of HF/E will repay itself. Cognition is an activity that occurs in context, and the contemporary context is technology. The fact that technology itself is an outflow of cognition

completes the circle, and surely shows why these two pursuits are synergistic in their fundamental aims and processes. Furthermore, we have advocated that HF/E must be concerned with what should be, as well as with what is. This being so, cognitive science must embrace a larger and societally more influential role, since the study of the brain, as the quintessential science, is the foundation of all human aspiration. It is this knowledge and power that must be used wisely, for the true marriage of HF/E and cognitive science will change the world.

References

Bainbridge L (1983) Ironies of automation. *Automatica* **19**: 775–779.

Billings CE (1997) *Aviation Automation: The Search for a Human-Centered Approach*. Mahwah, NJ: Erlbaum.

Broadbent DE (1958) *Perception and Communication*. London: Pergamon.

Broadbent DE (1971) *Decision and Stress*. London: Academic Press.

Bunge M (1980) *From Mindless Neuroscience to Brainless Psychology*. Paper presented at the Annual Winter Conference for Brain Research, January 1980, Keystone, CO.

Chomsky (1957) *Syntactic Structures*. The Hague: Mouton.

Clark A (1997) *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Corbetta M, Kincade J, Ollinger JM, McAvoy M and Shulman GL (2000) Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience* **3**: 292–297.

Dempsey PG, Wogalter MS and Hancock PA (2000) What's in a name? Using terms from definitions to examine the fundamental foundation of human factors and ergonomics science. *Theoretical Issues in Ergonomic Science* **1**: 3–10.

Farrell S and Lewandowsky S (2000) A connectionist model of complacency and adaptive recovery under automation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**: 395–410.

Fitts PM (ed.) (1951) *Human Engineering for an Effective Air Navigation and Traffic Control System*. Washington, DC: National Research Council.

Fitts PM and Jones RE (1947) *Analysis of Factors Contributing to 460 'Pilot Error' Experiences in Operating Aircraft Controls*. Report TSEAA-694-12. Wright-Patterson Air Force Base, OH: Air Materiel Command, Aeromedical Laboratory.

Flach J, Hancock PA, Caird JK and Vicente K (eds) (1995) *Global Perspectives on the Ecology of Human–Machine Systems*. Mahwah, NJ: Lawrence Erlbaum.

Gazzaniga M (2000) *The New Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Gibson JJ (1966) *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton-Mifflin.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Gibson JJ and Crooks LE (1938) A theoretical field analysis of automobile driving. *American Journal of Psychology* **51**: 453–471.

Haddon W (1970) On the escape of tigers: an ecologic note. *Technology Review* **72**: 44–47.

Hancock PA (1997) *Essays on the Future of Human–Machine Systems*. Eden Prairie, MN: Banta.

Hancock PA and Meshkati N (eds) (1988) *Human Mental Workload*. Amsterdam: North-Holland.

Hancock PA and Chignell MH (1995) On human factors. In: Flach J, Hancock PA, Caird JK and Vicente K (eds) *The Ecology of Human–Machine Systems, I. Global Perspectives* pp. 14–53. Mahwah, NJ: Lawrence Erlbaum.

Hancock PA and Scallen SF (1998) Allocating functions in human–machine systems. In: Hoffman RR, Sherrick MF and Warm JS (eds) *Viewing Psychology as a Whole: The Integrative Science of William N. Dember*, pp. 509–539. Washington, DC: American Psychological Association.

Hancock PA and Vasmatzidis I (1999) On the behavioral basis for stress exposure limits: the foundational case of thermal stress. In: Karwowski W and Marras W (eds) *The Occupational Ergonomics Handbook*, pp. 1707–1739. Boca Raton, FL: CRC Press.

Hancock PA, Flach J, Caird JK and Vicente K (eds) (1995) *Local Applications in the Ecology of Human–Machine Systems*. Mahwah, NJ: Lawrence Erlbaum.

Hutchins E (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.

Jastrzebowski W (1857) *An Outline of Ergonomics, or the Science of Work Based Upon the Truths Drawn From the Science of Nature*, commemorative edn, 2000. Warsaw: Central Institute for Labour Protection.

Jiang Y, Haxby JV, Martin A, Ungerleider LG and Parasuraman R (2000) Complementary neural mechanisms for tracking items in human working memory. *Science* **287**: 643–646.

Karwowski W and Marras W (eds) (1999) *The Occupational Ergonomics Handbook*. Boca Raton, FL: CRC Press.

Klein G (1999) *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Leuder R and Noro R (1994) Hard facts about soft machines. London, UK: Taylor & Francis.

Newell A and Simon H (1956) The logic theory machine. *IRE Transactions on Information Theory* **2**: 61–79.

Norman DA (2000) *The Invisible Computer*. Cambridge, MA: MIT Press.

Parasuraman R (1987) Human–computer monitoring. *Human Factors* **29**: 695–706.

Parasuraman R (1998a) *Neuroergonomics: The Study of Brain and Behavior at Work*. Washington, DC: Cognitive Science Laboratory. www.psychology.cua.edu/csl/neuroerg.html.

Parasuraman R (1998b) *The Attentive Brain*. Cambridge, MA: MIT Press.

Parasuraman R (2002) Neuroergonomics: research and practice. *Theoretical Issues in Ergonomics Science* **3**: (in press).

Parasuraman R and Mouloua M (1996) (eds) *Automation and Human Performance: Theory and Applications*. Mahwah, NJ: Lawrence Erlbaum.

Parasuraman R and Riley VA (1997) Humans and automation: use, misuse, disuse, abuse. *Human Factors* **39**: 230–253.

Parasuraman R, Mouloua M and Molloy R (1996) Effects of adaptive task allocation on monitoring of automated systems. *Human Factors* **38**: 665–679.

Parasuraman R, Sheridan TB and Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics* **30**: 286–297.

Peres M, Van der Moortele P, Peirard C *et al*. (2000) Functional magnetic resonance imaging of mental strategy in a simulated aviation performance task. *Aviation, Space and Environmental Medicine* **71**: 1218–1231.

Rasmussen J (1986) *Information Processing and Human–Machine Interaction*. Amsterdam: North-Holland.

Rawlins GJE (1996) *Moths to the Flame: The Seductions of Computer Technology*. Cambridge, MA: MIT Press.

Sarter N, Woods DD and Billings CE (1997) Automation surprises. In: Salvendy G (ed.) *Handbook of Human Factors and Ergonomics*, 2nd edn, pp. 1926–1943. New York: Wiley.

Satchell P (1998) *Innovation and Automation*. Aldershot: Ashgate.

Sheridan TB (1992) *Telerobotics, Automation and Supervisory Control*. Cambridge, MA: MIT Press.

Slamecka NJ and Graf P (1978) The generation effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory and Cognition* **4**: 592–604.

Wickens CD, Mavor A, Parasuraman R and McGee J (1998) *The Future of Air Traffic Control: Human Operators and Automation*. Washington, DC: National Academy Press.

Wiener EL (1988) Cockpit automation. In: Wiener EL and Nagel DC (eds) *Human Factors in Aviation*, pp. 433–461. San Diego, CA: Academic Press.

Wiener EL and Curry RE (1980) Flight-deck automation: promises and problems. *Ergonomics* **23**: 995–1011.

Woods DD (1996) Decomposing automation: apparent simplicity, real complexity. In: Parasuraman R and Mouloua M (eds) *Automation and Human Performance: Theory and Applications*, pp. 1–17. Mahwah, NJ: Lawrence Erlbaum.

Woods DD and Hancock PA (2001) Webpage editorial (cited in Ballot reform gets an airing. *Human Factors Society Bulletin* **44**: 1).

Human–Computer Interaction

Introductory article

John M Carroll, Virginia Tech, Blacksburg, Virginia, USA

CONTENTS

Human–computer interaction is an area of applied cognitive science and engineering design. It is concerned both with understanding how people make use of devices and systems that incorporate computation, and with designing new devices and systems that enhance human performance and experience.

HUMAN–COMPUTER INTERACTION AND COGNITIVE SCIENCE

Three touchstone themes are frequently sounded in cognitive science: the importance of investigating and analyzing real domains of cognition; the possible synergies between cognitive science and cognitive engineering; and the need for broader and better-integrated models and theories of cognition. These themes were heard in many of the talks at the inaugural meeting of the Cognitive Science Society in 1979, and have guided the vision of cognitive science ever since. As a scientific endeavor, the discipline of human–computer interaction (HCI) was created – largely by cognitive scientists – to address these three themes.

Investigating Real Domains of Cognition

The first decade of cognitive science research investigated the nature and role of areas of domain knowledge such as algebra, mechanics, radiology, and HCI. During the 1970s, a set of challenging cognitive issues in HCI had been identified, under the banner of the 'software crisis': programming languages were difficult to learn, programming activity was poorly supported by tools and methods, and the software produced was unreliable and hard to correct and maintain. The rapid expansion of computing into business organizations made the problems more urgent: if professional programmers could not deal effectively with computers, then who could?

HCI research in the 1980s took these problems as a realistic test-bed for assessing and developing concepts, models, and techniques from cognitive science: software design was analyzed as a type of problem solving. Learning and performance with programming languages and command interfaces were analyzed as psycholinguistic tasks. Interactions with text editors and programming tools were analyzed as cognitive skills.

This strategy proved effective. Cognitive science provided HCI with a research framework from its inception. This guided HCI researchers in planning empirical projects and making sense of results. It raised the intellectual level of early HCI research from studies of assorted problematic phenomena to studies of general issues. Throughout the 1980s, HCI research on routine expertise, ill-structured problems, concept names, mental models of hypermedia information structures, and a host of instructional technologies, from 'discovery learning' to intelligent tutoring systems, was articulated and pursued as mainstream cognitive science. HCI, in turn, provided a very high-profile case study of how real domains of cognition could be investigated. By the mid-1980s, in the wake of the personal computer 'revolution', HCI was the most visible application of cognitive science.

Synergies Between Cognitive Science and Cognitive Engineering

HCI proved to be a particularly suitable domain in which to explore the synergies between cognitive science and cognitive engineering. Software systems are highly malleable, affording powerful experimental approaches that are not possible in traditional areas of cognitive science like natural language and naive physics. Thus, the hypothesis that a command language with given linguistic properties might evoke certain patterns of confusion could quickly be tested by specifying the command mappings in a real software system. Integrating specific cognitive science research with real HCI contexts narrows the gap between building knowledge and building products, and

increases the likelihood that research outcomes will have practical import. Indeed, HCI is a paradigmatic example of Simon's 'sciences of the artificial' (sciences that investigate artifacts and systems produced by humans).

Originally, it was hoped that cognitive engineering work would contribute creatively to cognitive science. This is a sophisticated and ambitious objective: even a reliable one-way relationship between basic and applied science is fairly rare. However, direct manipulation and user interface metaphors are two examples of scientifically significant engineering developments in HCI.

Direct manipulation is a style of computer interaction whereby a person manipulates data and functions through gestures with display objects – for example, pointing and clicking in windows with a mouse – rather than by referring to data and functions by name in typed command strings. Direct manipulation was developed because it seemed to simplify and streamline human–computer interaction. Subsequently, this innovation in cognitive engineering guided the emergence of new cognitive theories of human action.

During the 1970s, user interfaces became more elaborate, developing from single input lines on a teletype to graphical displays incorporating direct manipulation. As user interfaces became more elaborate, users and designers interpreted them using more elaborate analogies: the user interface as a typewriter, as a stack of papers, as a desktop, as a virtual world. There is a considerable body of research and theory on analogy in cognitive science, and this general theory guided early design work in HCI. However, cognitive science research tends to focus on the role of analogies in attaining concepts in natural science domains. Because HCI is a design domain, it allows a far greater variety of analogies. Today there are hundreds of analogies (called user interface metaphors) employed in user interface designs.

Broader and Better-Integrated Models and Theories

Cognitive science seeks to integrate the variety of perspectives on cognition from different disciplines. It seeks to balance the tendency of researchers to focus ever deeper on problems that are ever smaller by encouraging interdisciplinary approaches to problems of broader scope. HCI is a case in point. Even the earliest cognitive modeling efforts in HCI were unusually comprehensive by the standards of the time. They had to be: in order to generate predictions about user performance in realistic contexts, the models had to make assumptions about perception, attention, short-term memory operations, planning, and motor behavior. Interestingly, these models were sometimes criticized within HCI for being incomplete and too narrow. For example, even fairly successful models of error-free expert performance were criticized for not describing error diagnosis and recovery, skill acquisition, and user affect. Over time, HCI models have become increasingly complete and comprehensive, and have had a continuing influence on modeling research throughout cognitive science.

Because HCI is defined in part with respect to computer technology, it is in some respects broader and intellectually more eclectic than cognitive science itself. HCI incorporates all ideas and technologies that facilitate understanding and enhancing the use of devices and systems that incorporate computation. Some of the ideas appropriated by HCI have pushed back the boundaries of cognitive science. An example of this is the incorporation of ethnomethodology into HCI, which occurred during the late 1980s following the publication of Suchman's book *Plans and Situated Actions*. This work had a huge influence on HCI because it demonstrated very concretely how work activity depends on coordination among participants, and how it can go awry when interactions are not sufficiently intelligible to all participants – including machines. Ethnomethodology contributed to a significant broadening in HCI theory which eventually also incorporated (Russian) activity theory, (British) sociotechnical design, (Scandinavian) participatory design, and ethnography.

In this respect, HCI appears to be able to serve as a conduit for broader theory in cognitive science. Suchman dismissed the representational theory of mind, and characterized her work as an empirical refutation of the concept of planning as developed in cognitive science and artificial intelligence. Without HCI, her work might have been ignored by mainstream cognitive science (like much other work in ethnomethodology). But in 1993, a special issue of the journal *Cognitive Science* was devoted to reconsideration of her book.

CURRENT CHALLENGES

Even the most fundamental issues in HCI are constrained by computer technology and the computing industry. It is likely that the basic scenario of human–computer interaction (namely, a person working with a keyboard, mouse, and graphic display) will change in the early twenty-first century,

perhaps radically. This will raise new challenges for HCI and for cognitive science.

Ubiquitous Computing

Many of the technologies underlying HCI improved steadily in the last quarter of the twentieth century. Increased processor capacity and memory density, and low-power displays, allowed the development of portable and handheld computers. Wireless applications are being developed and deployed rapidly, raising the possibility of totally mobile networked computing in the future. Other technologies are more specific to the user interface: examples include high-quality synthetic speech and robust speech recognition, and devices like sensing gloves, which permit finely articulated gesture input with no input surface, and heads-up displays incorporated into glasses.

These technologies could make computing invisible and ubiquitous – seamlessly, pervasively, and intimately incorporated into every facet of daily activity and experience. Ubiquitous computing raises many questions in cognitive science. How will people perceive and perform in computer-augmented realities? How will they think about and manage personal information processing when they, in effect, wear it? How will people conceive of themselves, if most human activity becomes computer-mediated?

Computer-Supported Cooperative Work

In the 1980s, the focus of HCI was on individual users: word processing and spreadsheets were HCI paradigms. In the 1990s, the focus shifted to computer-supported cooperative work (CSCW): email, instant messaging, and the World Wide Web became paradigmatic. By the end of the 1990s, these technologies had transformed many everyday activities, including shopping, information seeking, interactions with family and friends, and workplace collaboration.

The emergence of CSCW was closely linked with cognitive research in HCI, drawing upon concepts like 'common ground' from psycholinguistics, input–process–output models of group functioning from social psychology, and ethnographic and ethnomethodological models from sociology. For example, the design and investigation of relatively large-scale infrastructures like scientific collaboratories and community networks has allowed real and long-term collaborative processes to be investigated as never before. Such mediated collaborations can be more effective than face-to-face interaction when the technology makes task-irrelevant factors (like culture, age, and sex) less salient, and facilitates sharing of resources.

CSCW is also a tool for investigating traditional cognitive science issues. For example, anonymity is an important factor in problem solving activities like brainstorming, but it is difficult to study. The adoption of CSCW technology has been so rapid that studies can now be carried out in real task contexts instead of in contrived laboratory simulations. Many studies have described how anonymity can attenuate the effects of various sorts of power imbalance and evaluation anxiety, but at the same time evoke social 'loafing' and certain antisocial behaviors.

In the near future, it is likely that comprehensive collaborative environments will become widely available; tools like email, messaging and the Web will be integrated and supplemented by video conferencing and joint authoring and editing. As these CSCW technologies are deployed throughout society, further cognitive science issues will need to be addressed. For example many social practices depend upon physical or ephemeral acts and artifacts: what will happen as all data and interaction becomes digital and permanent?

Integrated Models and Theories

HCI is a good illustration of what cognitive science is about and how it operates with respect to the themes of understanding real domains of cognition, capitalizing on possible synergies between science and engineering, and developing better-integrated theories of cognition. In HCI, as in cognitive science itself, every success raises new challenges. HCI is valued and well utilized by the computer industry, but there is often tension between research and product development. HCI is a far more diverse area of cognitive theory now than it was in 1980. And the people who work in HCI now are more diverse than those who founded the discipline: they include graphic designers, electrical engineers, computer scientists, and business administrators.

The challenge now is to create an integrated cognitive science of HCI, one that encompasses the considerable range of models and theories currently in use, that defines effective tools and methods in a manner accessible to all HCI practitioners, and that anticipates likely future challenges, such as ubiquitous computing and computer-supported cooperative work.

Further Reading

Card SK, Moran TP and Newell A (1983) *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

Carroll JM (2000) *Making Use: Scenario-Based Design of Human–Computer Interactions*. Cambridge, MA: MIT Press.

Carroll JM (ed.) (2002) *Human–Computer Interaction in the New Millennium*. Reading, MA: Addison-Wesley.

Hutchins E, Hollan J and Norman DA (1986) Direct manipulation interfaces. In: Norman DA and Draper SW (eds) *User Centered System Design*, pp. 87–124. Hillsdale, NJ: Erlbaum.

Kyng M and Mathiassen L (eds) (1997) *Computers and Design in Context*. Cambridge, MA: MIT Press.

Landauer TK (1995) *The Trouble with Computers: Usefulness, Usability, and Productivity*. Cambridge, MA: MIT Press.

Mullet K and Sano D (1995) *Designing Visual Interfaces: Communication Oriented Techniques*. Englewood Cliffs, NJ: Prentice-Hall.

Neale DC and Carroll JM (1997) The role of metaphors in user interface design. In: Helander M, Landauer TK and Prabhu PV (eds) *Handbook of Human–Computer Interaction*, 2nd edn, pp. 441–462. Amsterdam: North-Holland.

Norman DA (1999) *The Invisible Computer*. Cambridge, MA: MIT Press.

Rosson MB and Carroll JM (2001) *Usability Engineering: Scenario-Based Development of Human–Computer Interaction*. San Francisco, CA: Morgan Kaufmann.

Suchman LA (1987) *Plans and Situated Actions: The Problem of Human–Machine Communication*. New York, NY: Cambridge University Press.

Hypnosis and Suggestion

Introductory article

Steven Jay Lynn, State University of New York, Binghamton, New York, USA
Irving Kirsch, University of Connecticut, Storrs, Connecticut, USA
Judith W Rhue, Ohio University, Athens, Ohio, USA
Joseph P Green, Ohio State University, Lima, Ohio, USA

CONTENTS

Hypnosis has been defined as a procedure in which a person designated as a hypnotist gives suggestions to a person designated as a subject, client, or patient for changes in sensations, perceptions, thoughts, feelings, and behaviors.

INTRODUCTION

Shrouded for centuries in mystery and myth, hypnosis has inspired equal measures of suspicion and fascination. During hypnosis, many people appear to lose control over normally voluntary behavior; some exhibit temporary, selective amnesia; they report seeing and hearing things that are not present and not seeing or hearing things that are present; while still other people report dramatic changes in sensations such as pain. Although the term 'hypnosis' was coined by the nineteenth-century British physician James Braid (1843), the dramatic changes in hypnotized subjects' appearance, experiences, and behaviors that the term encompasses had been well known for at least a half-century earlier under the names 'animal magnetism' or 'mesmerism'.

By the later eighteenth century, the idea that effects of hypnosis were due to magnetism was refuted by scientific evidence, whereas other outrageous claims – such as that hypnosis enabled the subject to see without the use of the eyes and to detect disease by seeing through the skin – were debunked by the mid-nineteenth century. Fortunately, such hokum did not dissuade the most renowned scholars of human behavior, including Sigmund Freud, Alfred Binet, William James, Wilhelm Wundt, Clark Hull, and Ernest R. Hilgard, from turning their attention to hypnosis. At the present time hypnosis is a subject of intensive investigation in psychological laboratories around the world, and is widely used as a catalyst for psychological and medical treatments.

HYPNOTIC SUGGESTIBILITY AND INDIVIDUAL DIFFERENCES

A consensus has begun to emerge about how to define hypnosis, across theorists and practitioners with widely divergent views about hypnosis. The American Psychological Association, Division of Psychological Hypnosis, has adopted a definition of hypnosis as a procedure during which changes in sensations, perceptions, thoughts, feelings, or behavior are suggested.

There are considerable differences in the extent to which people respond to hypnotic suggestions. Approximately 15–20 percent of the population is minimally suggestible, 15–20 percent of the population is highly suggestible, and the remainder of the population scores in the medium range of hypnotic suggestibility. People who are likely to respond to hypnotic suggestions exhibit the ability to respond to waking imaginative suggestions, almost to the same degree that they do during hypnosis. In addition, responding to suggestion is correlated with positive beliefs, attitudes, expectancies, and motivation to respond to the hypnotist's suggestions. Some studies have shown that imaginative abilities, fantasy proneness, dissociative experiences, a tendency to become absorbed in experiences, and openness to experience also contribute to hypnotic suggestibility. However, questions have been raised about whether the correlations between hypnotic suggestibility and measures of these constructs are inflated simply because they are measured in the same test context. Studies have shown that when hypnotic suggestibility and these

measures are tested in independent contexts, the correlations are either not statistically significant or very small in magnitude. Nevertheless, research into individual differences in hypnotic suggestibility remains a fruitful and important area of inquiry.

FACTS ABOUT HYPNOSIS

Many misconceptions about hypnosis come from movies, television, books, and stage hypnosis. Individuals who hold the popular yet mistaken belief that only gullible people respond to hypnosis, or that hypnosis induces a sleep-like trance state marked by loss of control, are not likely to respond fully to hypnotic suggestions. Contrary to popular misconceptions, hypnosis is neither sleep nor relaxation; suggestions that emphasize alertness are as effective as suggestions that emphasize drowsiness and relaxation. Additionally, participants retain control during hypnosis and can refuse to respond to suggestions if they wish.

Research has also documented the following points:

- Most hypnotized people are neither faking nor merely complying with suggestions.
- Suggestions can be responded to with or without hypnosis. The function of a formal induction is primarily to increase suggestibility to a small degree.
- A wide variety of hypnotic inductions can be effective. Authoritative, traditionally worded hypnotic techniques are as effective as permissive, open-ended, suggestions.
- All of the behaviors and experiences occurring in hypnosis can also be produced by suggestions given without the prior induction of hypnosis.
- Hypnotic suggestibility is relatively stable, with test–retest correlations reported as high as 0.71 over a 25-year follow-up. This stability has been variously described as a reflection of the trait-like aspects of hypnotic suggestibility or a reflection of attitudes and beliefs about hypnosis, and interpretations of hypnotic suggestions that remain stable over time.
- Hypnotic suggestibility can be substantially modified by training programs that: (a) inculcate positive attitudes, beliefs, and expectancies about hypnosis; (b) instruct participants to imagine and get involved in test suggestions; and (c) train individuals in how to interpret correctly the implicit requirements of test suggestions (e.g. not passively wait for a response to occur to a suggestion for hand levitation).
- Hypnosis depends more on the efforts and abilities of the subject than on the skill of the hypnotist.
- Subjects typically remain aware of their surroundings and recall what transpired during hypnosis.
- Hypnosis is not a dangerous procedure when practised by qualified clinicians and researchers.
- Some studies suggest that hypnotic suggestibility is related to dissociative disorders, posttraumatic stress disorder, and phobias.
- Reliable physiological indicators of hypnosis, independent of the effects of administered suggestions, have not been identified in well controlled, consistently replicated studies.
- Hypnosis does not permit subjects to literally re-experience the events of childhood or to function in a truly child-like fashion.
- Hypnosis increases inaccurate as well as accurate memories and tends to inflate confidence in what is remembered, regardless of accuracy.

HYPNOSIS IN THE LEGAL ARENA

Concerns about the reliability of hypnotically enhanced recall have led many psychotherapists to eschew the use of hypnosis to recover memories of purportedly repressed traumatic events. Relatedly, many courts in the USA have expressed serious concerns about the admissibility of hypnotically elicited testimony and have banned the entire testimony of a witness subjected to hypnosis for memory retrieval. Some courts have stipulated that hypnotically augmented testimony can be admitted only when certain procedural guidelines and safeguards are implemented in the way hypnosis is conducted, while a number of states and federal courts allow exceptions to the 'admissibility with safeguards' rule and, instead, apply a 'totality of circumstances' test. When the latter test is applied, the case is judged not so much in terms of whether specific guidelines were followed or not, but in terms of the 'totality of circumstances' that attended the hypnosis session. However, the majority of states (25 of 30 state supreme courts that have ruled on admissibility of hypnotically elicited testimony) have rejected the idea of admitting hypnotic testimony on a case-by-case basis and considering the totality of circumstances; they have, instead, opted for excluding hypnotically elicited testimony.

THEORETICAL ACCOUNTS

Different explanations of hypnosis have been as vigorously debated as the use of hypnosis in the courtroom. Hypnosis was traditionally regarded as an altered state of consciousness (or trance), and the effects of hypnotic suggestion were hypothesized to be due to the induction of that hypothesized state. However, beginning in the 1950s, the socio-cognitive theorists Theodore Sarbin and T. X. Barber rejected the idea that hypnotic responses were due to an altered state, and maintained that

all the phenomena of hypnosis, including behavioral responses to suggestion, and even the subjective experience of a trance state, can be accounted for without postulating any special state or condition. From this perspective, and the vantage point of other socio-cognitive theorists such as Spanos and Chares, Kirsch, Lynn and Rhue, and Wagstaff, hypnotic behaviors are like other complex social behaviors; they are a product of such factors as ability, attitude, belief, expectancy, attribution, and interpretation of the situation.

Neodissociation theories have emerged as strong competitors of sociocognitive models. Hilgard's neodissociation theory is based on the idea that there exist multiple cognitive systems or cognitive structures in hierarchical arrangement under some measure of control by an 'executive ego'. The executive ego or 'central control' structure is responsible for planning and monitoring functions of the personality. During hypnosis the hypnotist's suggestions take much of the normal control away from the subject. An amnesic barrier prevents conscious awareness of particular ideas, imaginings, and fantasies, thereby producing the subjective impression of nonvolition that typically accompanies hypnotic responses.

Whereas neodissociation theorists contend that an altered state is a characteristic of the person at a descriptive level, they do not claim that a particular altered state of consciousness explains or causes hypnotic phenomena. In fact, when the term 'hypnotic state' is used by researchers today, it is usually used merely in a descriptive sense to denote the subjective changes that hypnotized subjects report experiencing. It is not used to explain those changes. Instead, most hypnosis researchers agree that the impressive effects of hypnosis stem from social influence and personal abilities, not from a trance-like state of altered consciousness.

A number of other theories of hypnosis have been proposed. These include psychoanalytic theories, psychobiological models, interactive-phenomenological models of hypnosis, the theory of dissociated control, and derivatives of traditional trance theory. Despite considerable agreement on the basic empirical findings, there is no consensus on how those findings are to be explained.

CLINICAL HYPNOSIS

There is considerable consensus regarding the fact that hypnosis can have great clinical value. Hypnosis is not a type of therapy, like psychoanalysis or behavior therapy. Instead, it is a procedure that can be used to facilitate therapy. As it is practised today, clinical hypnosis can be defined as the addition of hypnosis to accepted psychological or medical treatment.

Hypnosis can have clinical utility for a number of reasons. Hypnosis can foster a positive collaborative working alliance with the client, provide a structure for therapeutic suggestions and activities, and facilitate a focus on selected thoughts, feelings, images, and experiences that could never occur in reality. Hypnosis also provides a context in which clinicians can do the following: (a) talk with clients in metaphoric, rather than strictly literal terms; (b) use suggestions and imagery to desensitize fears and rehearse coping responses to stressful tasks and situations; (c) interrupt negative patterns of thoughts, feelings, and behaviors; and (d) enhance positive affect and feelings of self-control. Hypnotic procedures can be defined as 'self-hypnosis' to enhance perceptions of treatment success and the likelihood that what is learned during the therapy hour will be implemented in everyday life. Additionally, posthypnotic suggestions (e.g. relaxation, calmness) can be used to generalize treatment gains to everyday life. Because many clients have positive attitudes about hypnosis, the hypnotic context may enhance their confidence in the effectiveness of therapy and thereby produce a placebo effect without the deception that is generally associated with placebos.

Meta-analyses have demonstrated that the addition of hypnosis to cognitive–behavioral and psychodynamic treatments substantially enhances their efficacy. Positive effects of hypnosis were associated with the treatment of hypertension, anxiety, phobias, and duodenal ulcers. But the most impressive effect was found for weight reduction treatments, and this was particularly impressive at long-term follow-up.

Recent reviews that have examined the empirical support for hypnosis as an adjunctive intervention revealed that: (a) suggestion can be a very effective means of alleviating severe and persistent pain and that this effect is not restricted to highly suggestible clients; (b) hypnotic procedures generally yield higher rates of smoking abstinence relative to deferred treatment (i.e. wait list) and no treatment conditions; (c) hypnosis can successfully treat a variety of medical conditions such as irritable bowel syndrome, dermatological disorders, and postchemotherapy nausea and emesis, and can assist in the preoperative preparation of surgical patients; and (d) there are indications that hypnosis can be useful in the treatment of trauma and the treatment of children with a variety of conditions

including nocturnal enuresis and pain associated with medical operations.

CONCLUSION

Hypnosis has fired the imagination of scholars, researchers, and clinicians for more than 200 years. However, it is only in the past half-century or so that hypnosis has received the empirical attention and status in the scientific community that it deserves. The study of hypnosis has the potential to shed light on the way that interpersonal communications can be transformed into deeply felt subjective experiences, with vital power to relieve human suffering. Hypnosis researchers have contributed to our knowledge of the determinants of profound alterations in consciousness, the role of imagination and fantasy in everyday life, the link between responses to suggestion and physiological processes, the genesis of false memories, and the way that expectancies shape myriad subjective and behavioral responses. The study of hypnosis will contribute as much to our understanding of basic cognitive and interpersonal processes as insights from social, cognitive, and biological psychology will contribute to our understanding of hypnosis.

Further Reading

Barber TX (1969) *Hypnosis: A Scientific Approach*. New York, NY: Van Nostrand Reinhold.

Fromm E and Nash M (1992) *Contemporary Hypnosis Research*. New York, NY: Guilford.

Gauld A (1992) *A History of Hypnotism*. Cambridge, UK: Cambridge University Press.

Hilgard ER (1986) *Divided Consciousness: Multiple Controls in Thought and Action* (expanded edn). New York, NY: John Wiley.

Kirsch I, Capafons A, Cardena-Bulena E and Amigo S (1998) *Clinical Hypnosis and Self-Regulation: A Cognitive-Behavioral Perspective*. Washington, DC: American Psychological Association.

Kirsch I and Lynn SJ (1995) The altered state of hypnosis: changes in the theoretical landscape. *American Psychologist* **50**: 846–858.

Lynn SJ and Rhue JW (1991) *Theories of Hypnosis*. New York, NY: Guilford.

Rhue JR, Lynn SJ and Kirsch I (eds) (1993) *Handbook of Clinical Hypnosis*. Washington, DC: American Psychological Association.

Sheehan PW and McConkey KM (1982) *Hypnosis and Experience: The Exploration of Phenomena and Process*. Hillsdale, NJ: Lawrence Erlbaum.

Spanos NP (1986) Hypnotic behavior: a social psychological interpretation of amnesia, analgesia, and trance logic. *Behavioral and Brain Sciences* **9**: 449–467.

Idioms, Comprehension of

Intermediate article

Patrizia Tabossi, Università degli Studi di Trieste, Trieste, Italy

CONTENTS

Idioms are fixed expressions whose meaning is not a direct function of the meanings of their parts. Their comprehension involves both recognition and parsing processes, and hinges upon people's pragmatic as well as syntactic and semantic competence.

IDIOMS: A CHALLENGE TO CURRENT MODELS OF LANGUAGE COMPREHENSION

Idioms are usually characterized as strings of words whose meaning is not a direct function of the meanings of their parts. For example, the meanings of the words *kick*, *the*, and *bucket*, composed according to the syntactic relations among them, do not produce the meaning 'die suddenly'. This is conventionally associated to the string and a person must know the convention in order to correctly understand it.

Thus, idiomatic expressions defy the standard view of language comprehension according to which understanding a sentence requires at least recognizing the words in the sentence, retrieving from the mental lexicon the different types of information associated with them, and combining their meanings according to their grammatical relations. (*See* **Language Comprehension**; **Parsing: Overview**; **Sentence Processing**; **Sentence Processing: Mechanisms**)

Idioms are multifaceted objects with properties along several distinct dimensions, including the following:

Informality
Idioms are used more frequently in spoken than written language, typically in colloquial and informal situations. They, along with other tropes such as metaphor, irony, or metonymy, convey meaning in a vivid fashion, and usually possess an affective connotation.

Figurativeness
We often perceive that some degree of figurativeness is involved in the meanings of many idioms When we are able to recover the relation between their literal and figurative meanings, the idioms are said to be transparent (e.g. *sweep under the carpet*); otherwise they are opaque (e.g. *kick the bucket*).

Frozeness
Frozeness refers to the fact that typically idioms occur only in limited syntactic constructions. Idioms are very different with respect to the syntactic operations that they can undergo. While some are almost completely unconstrained (e.g. *spill the beans*, *let the cat out of the bag*), others allow very few operations, if any (e.g. *by and large*, *trip the light fantastic*).

These characteristics are not necessary features of idioms. *Pay attention to* or *care for*, for example, can easily be used in formal occasions, and figuration is hardly involved in idioms like *by and large* or *try for*. However, they tend to be related to one another and to other characteristics in interesting ways. Frozeness, for example, often co-varies with opacity and ill-formedness, i.e. an idiom's lack of a literal interpretation. In fact, ill-formed idioms tend to be less syntactically flexible than well-formed ones, and typically involve little or no figuration.

In general, idioms constitute a very heterogeneous class of expressions. Although these expressions do not appear to serve a logical function in language, they are very common and reflect our tendency to say things in conventionalized ways. How are these elusive expressions understood?

THE 'CLASSIC' VIEW

Traditionally, idioms have been considered as one of the many forms of figurative language, and have been conceived as 'dead metaphors', i.e. expressions which were once innovative but are now conventionalized. Within this framework, idiomatic understanding occurs only after the literal analysis of the string has failed or its outcome does not fit in the context. At this point, the comprehension system goes into a special mode to retrieve the

string's conventional interpretation from an idiom list (Bobrow and Bell, 1973). (*See* **Metaphor**)

The idiom list hypothesis predicts that comprehending an idiomatic expression should take longer in its figurative than in its literal meaning. Empirical evidence, however, fails to support this claim. Understanding the figurative meaning of an idiomatic expression is a fast process; even faster than understanding the expression's literal meaning. To cope with these findings, Swinney and Cutler (1979) proposed the lexical representation hypothesis. According to this influential view, idioms are stored in the mental lexicon as long, morphologically complex words, and are recognized like any other lexical item. The process of idiom identification starts at the beginning of a string, and runs in parallel with the computation of its literal meaning. However, computing is a longer process than retrieving; hence, the idiomatic meaning of the string becomes available before its literal one. (*See* **Lexicon**; **Lexical Semantics**; **Morphology**; **Lexical Access**; **Word Recognition**)

This view correctly accounts for the basic fact that idioms are easy to understand. Several lines of investigation, however, have recently challenged its claims.

NOT JUST COMPLEX WORDS

Compositionality

Some years ago, Nunberg *et al*. (1994) argued that although idioms have conventional meanings, at least some of them are combining expressions. There are many idioms, in fact, whose constituents 'carry identifiable parts of their idiomatic meanings'. In *spill the beans*, for example, there is a clear, albeit motivationally opaque, correspondence between *spill* and *beans* and the relevant parts of its figurative meaning 'divulge information'. This analysis fits well with the intuitions of people who are relatively consistent in their ratings of idioms' compositionality and seem to be faster at understanding decomposable than nondecomposable idioms (Gibbs *et al*., 1989).

Syntactic Parsing

Peterson *et al*. (2001) investigated the relationship between syntactic processing and availability of figurative meaning during the comprehension of well-formed idioms. They found that the syntactic parsing of an idiomatic expression continues even after its figurative meaning has been retrieved and the computation of its compositional semantics terminated. These findings suggest that a full syntactic analysis of the idioms is always computed, regardless of the status of their literal interpretation. (*See* **Sentence Processing**; **Sentence Processing: Mechanisms**)

Flexibility and Productivity

Idioms may be not only syntactically but also lexically flexible: a word in an idiom can be substituted by a synonym, leaving the meaning and the acceptability of the string virtually unchanged (e.g. *hit the hay* and *hit the sack*) (Gibbs *et al*., 1989). Furthermore, idioms can be altered productively to create new variants. *Shatter the ice*, for example, modifies the meaning of *break the ice*, adding a striking alternative to the original meaning of breaking down an uncomfortable social situation. McGlone *et al*. (1994) compared the reading times of sentences containing idiomatic strings in their canonical form (e.g. *He had two left feet*), in their variant form (e.g. *He had three left feet*), or in their respective literal paraphrases (e.g. *He was extremely clumsy*, *He was incredibly clumsy*). The canonical form was read faster than either variants or paraphrases. However, the latter two did not differ from each other, suggesting that they are processed in much the same way, i.e. compositionally.

Recognition

Scrape the bottom immediately calls to mind *of the barrel*, whereas *hit the nail* does not makes one think of its idiomatic conclusion *on the head*. Tabossi and colleagues referred to the two types of idioms as predictable and nonpredictable, and named as key the word following which idioms come to mind. In a series of studies, they explored the time course of activation of the figurative meaning of early and late key idioms. Idiomatic meanings were found to activate sooner in early than in late key expressions, but never at the beginning of the strings.

This suggests that the recognition of idioms and words may differ in important ways. It is generally agreed that spoken word recognition is a continuous process involving the initial activation of multiple lexical candidates. However, the activation of idiomatic meanings is rather slow. Moreover, it is influenced by factors different from those relevant in word identification. In particular, an idiom key,

unlike a word's uniqueness point, indicates a point at which a string can still have multiple completions, and, unlike the recognition point, does not depend on the context of the occurrence of an idiomatic string (Cacciari and Tabossi, 1988; Tabossi and Zardon, 1993). (*See* **Word Recognition**)

A NEW PERSPECTIVE

In light of these findings, Cacciari and Tabossi (1988) proposed that idioms are mentally represented as configurations of words whose meanings become activated whenever sufficient input has rendered the configuration recognizable. A configuration is made up of the same lexical items that need to be activated during the comprehension of literal discourse; at what point in the string the idiomatic meaning is activated depends on the position of the key in that string.

This hypothesis explains how idioms are recognized in their canonical form and is compatible with the fact that they are fully parsed. It also complements current views on idiom comprehension (Glucksberg, 2001).

But at present, there is no full-blown theory that can account for the various limitations that apply to the different idioms, and explain, for example, why people are reluctant to interpret the *bucket was kicked* idiomatically, even though they cope perfectly well with *the beans were spilled*. However, a promising working hypothesis is that the variations an idiom can undergo without losing its figurative interpretation are constrained by (at least) the following factors:

The syntactic and semantic properties of its parts. *Kick*, for example, denotes a discrete action and its progressive form can only refer to the repetition of the action. However, one cannot repeatedly die, and hence *kicking the bucket* is not usually acceptable in its idiomatic meaning.

The idiomatic meaning of the string. Dying can occur tragically, but not sharply. Accordingly, *He tragically kicked the bucket* is acceptable, but *He sharply kicked the bucket* is not.

The pragmatic function of the variation. It is true in general that variations are introduced for communicative reasons, without which they sound odd. *The bucket was kicked by John* sounds odd because no entity in the discourse corresponds to the bucket, and the expression is therefore contextually inappropriate.

Depending on whether all, some, or none of these constraints are satisfied, a variation will result in an expression which will be perceived as ranging from perfectly appropriate to totally unacceptable.

IMPLICATIONS FOR THEORIES OF LANGUAGE COMPREHENSION

In current psycholinguistics, idioms are relegated to the sidelines. Research on language comprehension tends to focus on the lexicon: how individual words are mentally represented and recognized. It also concentrates on the parsing processes which lead people to compositionally reconstruct the semantic interpretation of ever-new sentences. No doubt, this approach captures the crucial fact that people use language creatively. However, in its rigid distinction between what is creative – the syntax – and what is not – the lexicon – it may incur the risk of underestimating the relevance of people's use of conventional language. (*See* **Lexical Access**; **Lexical Ambiguity Resolution**; **Means–Ends Analysis**; **Sentence Processing**; **Representations Using Formal Logics**)

Idioms constitute an important part of the vast family of fixed expressions which includes compounds (e.g. *frequent flyer program*), names (e.g. *Boston Pops Orchestra*), clichés (e.g. *gimme a break*), titles of songs, books, movies (e.g. *All You Need is Love*), quotations (e.g. *May the force be with you*), foreign phrases (e.g. *C'est la vie*). The number of fixed expressions that we have available for use is comparable with the number of individual words. It is a reasonable assumption that these expressions, which can hardly be considered marginal, are represented in the lexicon (Jackendoff, 1995). Hence, no theory of the lexicon may aspire to be psychologically adequate unless it provides an acceptable account of how fixed expressions are mentally represented and processed.

The complex nature of idioms with respect to compositionality and syntax makes them a rather special type of fixed expression, and lexical processes are unlikely to account for how people understand the variants that each idiom can take. The study of idiom comprehension suggests that much of their syntactic behavior may be explained outside the syntactic realm. This factor highlights the need for a greater integration of the syntactic, semantic, and pragmatic aspects of people's competence to appropriately understand their actual use of language. (*See* **Modularity**)

References

Bobrow S and Bell S (1973) On catching on to idiomatic expressions. *Memory and Cognition* **1**: 343–346.

Cacciari C and Tabossi P (1988) The comprehension of idioms. *Journal of Memory and Language* **27**: 668–683.

Gibbs RW Jr, Nayak NP and Cutting C (1989) How to kick the bucket and not decompose: analyzability and

idiom processing. *Journal of Memory and Language* **28**: 576–593.
Glucksberg S (2001) *Understanding figurative language: From metaphysics to idioms*. Oxford: Oxford Psychology Press.
Jackendoff R (1995) The boundaries of the lexicon. In: Everaert M, van der Linden E-J, Schenk A and Schreuder R (eds) *Idioms: Structural and Psychological Perspectives*, pp. 133–166. Hillsdale: Erlbaum.
McGlone MS, Glucksberg S and Cacciari C (1994) Semantic productivity and idiom comprehension. *Discourse Processes* **17**: 167–190.
Nunberg G, Sag IA and Wasow T (1994) Idioms. *Language* **70**: 491–538.
Peterson RR, Burgess C, Dell GS and Eberhard KM (2001) Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**: 1223–1237.
Swinney D and Cutler A (1979) The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior* **14**: 523–534.
Tabossi P and Zardon F (1993) The activation of idiomatic meaning in spoken language comprehension. In: Cacciari C and Tabossi P (eds) *Idioms: Processing, Structure and Interpretation*, pp. 145–162. Hillsdale: Erlbaum.

Further Reading

Cacciari C and Glucksberg S (1991) Understanding idiomatic expressions: the contribution of word meanings. In: Simpson GB (ed.) *Understanding Word and Sentence*, pp. 217–240. Amsterdam: North-Holland.
Cacciari C and Tabossi P (eds) (1993) *Idioms: Processing, Structure and Interpretation*. Hillsdale: Erlbaum.
Cutler A (1982) Idioms: the older the colder. *Linguistic Inquiry* **13**: 317–320.
Everaert M, van der Linden E-J, Schenk A and Schreuder R (eds) (1995) *Idioms: Structural and Psychological Perspectives*. Hillsdale: Erlbaum.
Fraser B (1970) Idioms within a transformational grammar. *Foundations of Language* **6**: 22–42.
Gibbs RW Jr (1986) Skating on thin ice: literal meaning and understanding meaning in conversation. *Discourse Processes* **9**: 17–30.
Gibbs RW Jr and Nayak N (1989) Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology* **21**: 100–138.
Glucksberg S, Brown M and McGlone M (1993) Conceptual metaphors are not automatically accessed during idiom comprehension. *Memory and Cognition* **21**: 711–719.
Katz J (1973) Compositionality, idiomaticity, and lexical substitution. In: Anderson S and Kiparsky P (eds) *A Festschrift for Morris Halle*, pp. 357–376. New York: Holt, Rinehart, & Winston.
Nunberg G (1978) *The Pragmatics of Reference*. Bloomington: Indiana University Linguistic Club.

Illusions

Introductory article

Richard L Gregory, University of Bristol, Bristol, UK

CONTENTS

The senses we rely on for knowing the world about us, such as sight and hearing, can suffer dramatic errors known as illusions. Illusions of sight are the best-known and have been used to elucidate the nature of visual perception.

INTRODUCTION

We rely on our sense of seeing – and hearing, touching, tasting, smelling and so on – for surviving, and knowing the world of objects. However, our senses are not completely reliable. They can suffer dramatic errors known as 'illusions'. These may occur when the sense organs (the eyes, ears, and so on) are physiologically disturbed; or, very differently, illusions can occur when we perceive special kinds of objects, even though the physiological function is normal. Some distortion illusions occur for pictures, though not for normal objects. Illusions of sight have been most studied, and will be mainly considered here. Theories are controversial: the following text presents one view of illusions.

There are many kinds of illusions, and there are many causes. Some are dangerous; some are useful, especially for artists. Understanding the phenomena of illusions is useful in science – for finding ways of avoiding them, and for discovering how perception works. They have interested people for centuries, including the ancient Greek philosophers, although the latter's knowledge of perception was limited. They did not even know that eyes have optical images. Illusions can be dangerous for flying or driving: but are useful for special effects in the cinema.

It is surprisingly hard to define the concept of illusion. Illusions are departures from truths of object reality; but how should we define, or indeed recognize, truths of what objects really are? This is not a 'merely academic' question, for there are very different accounts of reality in science, in art and religions, and the many flavors of metaphysics. Science's accepted realities grow ever further from how things appear to common sense. Objects do not look at all like the descriptions of physicists, and (as John Locke realized in the seventeenth century) much that appears inherent in objects is created by the mind or brain, including colors.

Science's accepted reality changed from the 'mechanistic' physics of the nineteenth century, to the quantum physics and relativity of the twentieth. As science departs ever further from commonsense appearances, we may be tempted to say that all experience is illusion; but this is no more helpful than saying that reality is a dream. If we say this, 'dream' simply loses any meaning. The problem is to decide which, of various candidates, we should accept as the reference reality for recognizing departures of illusion.

The familiar visual illusions figures, in children's books and texts of psychology, are departures from simple measurements of lengths, curvatures of lines or edges, and so on. There are also simply measured illusions of speed, temperature, and nonvisual illusions of weight and time. The reference we accept for judging illusions seems to be the world of 'kitchen physics' – simple measurements with rulers, scales, clocks, and thermometers. It is indeed because perceptions of length, weight, time and so on are subject to illusion, that these instruments have been so important for crafts and science.

There are also 'qualitative' illusions – especially mistaking one object, or kind of object, for another. There can be spontaneous alternations – the brain never making up its mind. Such ambiguities are important for showing that perceptions are guesses – perceptual hypotheses – of what might be out there in the world of objects. Some illusions cannot be matched against 'reality'; for there are perceptual fictions, and also paradoxes, which do not exist in the physical world, yet are seen.

KINDS OF ILLUSIONS

There are four principal phenomena of visual illusion: ambiguities, distortions, paradoxes, and

fictions. It may be no accident that these correspond to errors of language, for language may have developed from ancient, prehuman, perceptual classification of the world of objects and of actions. This could explain how language developed so fast in humans.

Ambiguities

A retinal image – like any picture – is infinitely ambiguous. It could correspond to, or represent, an infinity of shapes and sizes and distances. Thus, an ellipse in a drawing could represent a tilted circle, which might be small and near, or distant and large. The possibilities are endless; yet amazingly, we generally see just one of the infinite possibilities. How the visual brain usually selects a single possibility is not fully understood.

The term 'ambiguity' is itself ambiguous. It can mean several possibilities though but one or a few are seen or chosen. Alternatively, it can mean spontaneous changes of perception, as occurs with the images in Figures 1 and 2. The first is passive confusion of different stimuli or patterns; the second is active generation of alternative perceptions from the same stimulus or pattern.

Figures that flip between a few possibilities (ambiguous figures) show the dynamics of perception, as guesses of what might be out there; for perception creates hypotheses of object reality. Ambiguities are spontaneous changes between visual hypotheses, though there is no change in the object. The primary ambiguity is between something and nothing: objects, and spaces between objects (Figure 1). Objects can also 'flip' in orientation; a well-known example is the Necker cube (Figure 2). A well-known example of change of object is the old woman/young woman image in Figure 3.

Dynamic ambiguities are useful for research into how perception works, as perceptions change though the objects remains unchanged. This shows

Figure 1. Ambiguity: Spontaneous changes of perception occur in which the figure is either present or absent.

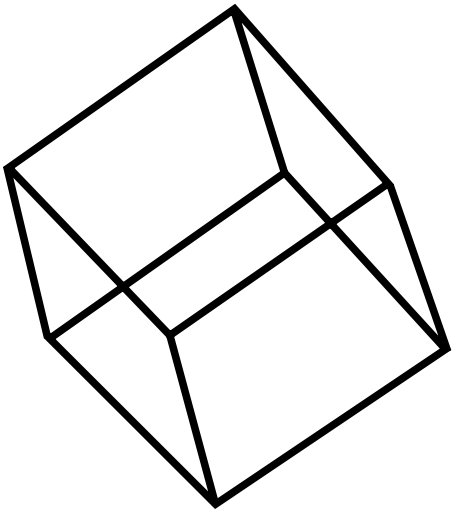


Figure 2. The Necker cube, which 'flips' in orientation. This was discovered by the Swiss crystallographer L. A. Necker in 1832 while drawing rhomboid crystals seen through a microscope.

Figure 3. Old woman or young woman?

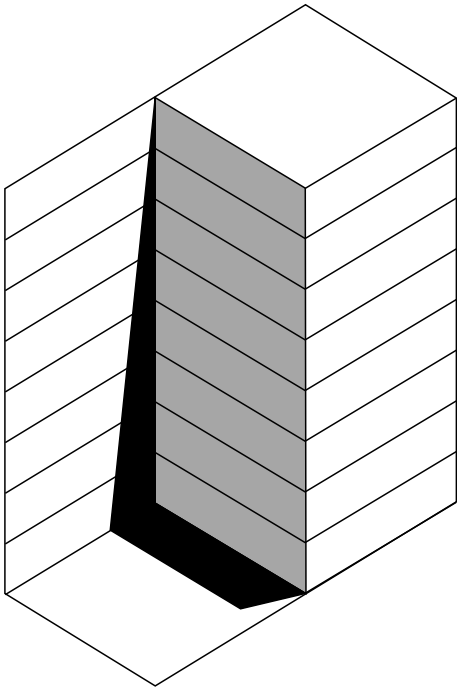


Figure 4. Mach corner (or card). A shadow (or painted shadow) on a card bent to this shape changes brightness each time the corner 'flips' in depth. When 'in' it is darker than when seen as 'out': this is because when 'in' it is accepted as a shadow, and shadows are minimized as they are not behaviorally significant objects.

Figure 5. The hollow side of the mask appears as a normal face because the hollow form is too unlikely to be seen.

that perceptions are not directly linked to objects. More technically, they allow us to separate 'bottom up' signals from the eyes, from 'top down' knowledge of objects, stored in the brain. Much of what is seen is psychologically projected into the world, especially colors, which are created in the brain although they appear to lie on the surface of external objects. A revealing demonstration by the

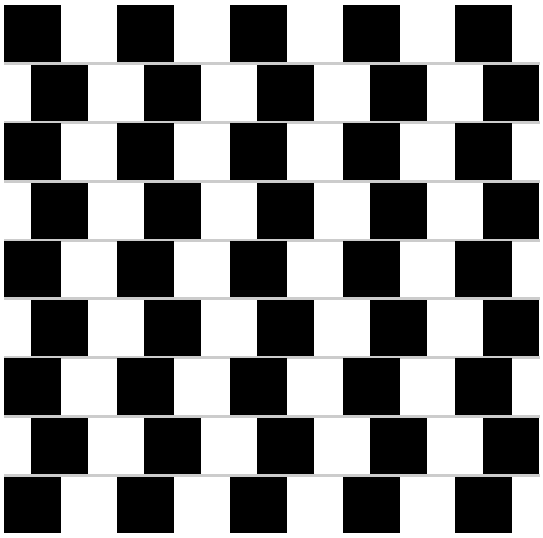


Figure 6. The 'café wall' distortion illusion. The alternating light and dark 'tiles' are rectangles, and the horizontal 'mortar' lines are parallel, although they appear as wedges.

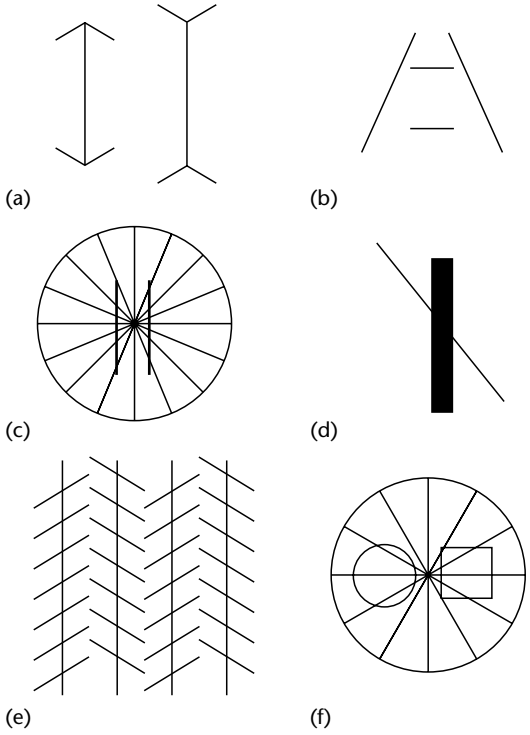


Figure 7. Cognitive distortions. The illusions of (a) Müller-Lyer, (b) Ponzo, (c) Hering, (d) Ponzo, (e) Zollner and (f) Orbison (two combined).

Austrian physicist Ernst Mach (1838–1916) shows how simple sensations such as brightness are affected by how a figure or an object is seen in different ambiguous states (Figure 4). This illusion

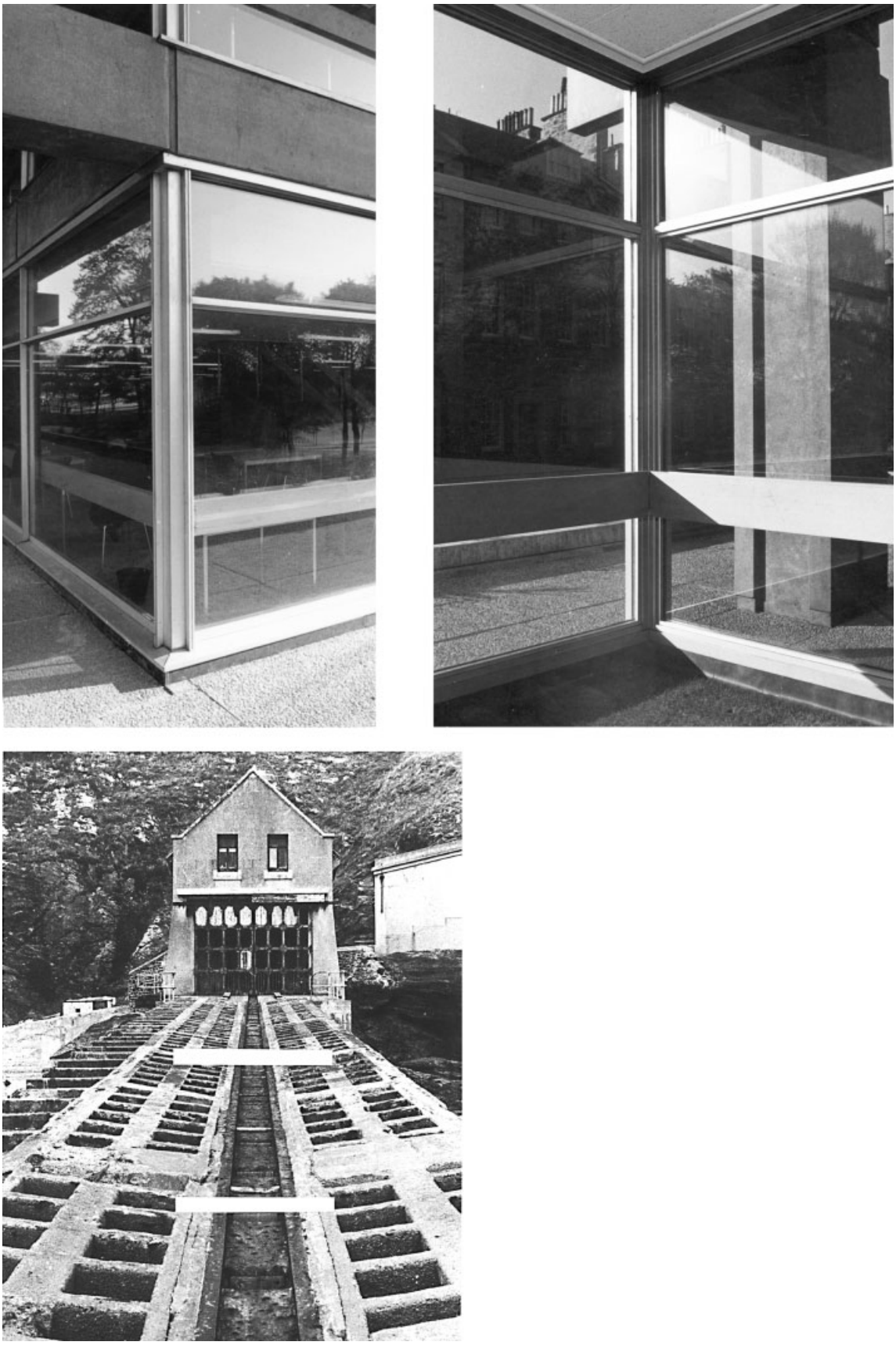


Figure 8. Distortion figures correspond to typical perspective depth.

shows something of the power of 'top down' knowledge in visual perception. Anatomically there are more descending nerve fibers from the cortex than ascending nerve fibers from the eyes. Vision therefore seems to be more 'top down' than 'bottom up': it depends more on knowledge from the past than on signals from the present. So, although not all of perception is in the mind, most of it is! This is surprising, and not all authorities would agree.

The most dramatic demonstration that knowledge is important for perception is the 'hollow face' (Figure 5). Although hollow, it appears to be a normal face with the nose sticking out – simply because a hollow face is too unlikely to be seen. This is strong evidence of the power of 'top down' knowledge to challenge and beat 'bottom up' signals from the eyes, including considerable stereopsis. It works for any object that is very unlikely to be hollow, but is most effective for a typical human face; like Jekyll and Hyde, it morphs from outside to inside face and back again, without becoming (as it truly is) hollow.

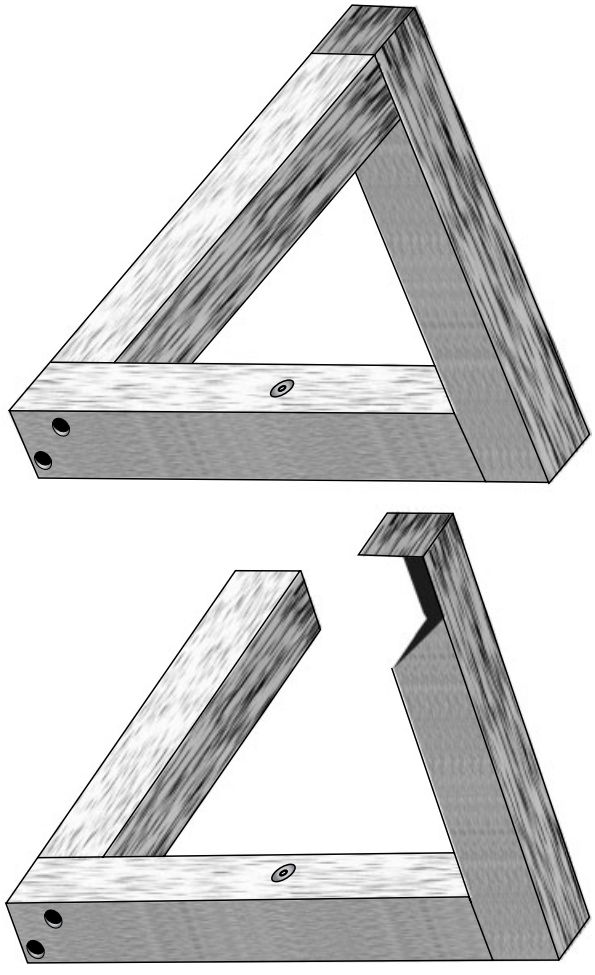


Figure 9. Paradoxical illusion: the 'impossible triangle'. It appears to be impossible when viewed from one position (top), but when seen from a different viewpoint (bottom) it becomes a realistic object.

Distortions

It is well known that size and shape may be distorted in simple illusion figures. Many of these were discovered by physicists and astronomers in the nineteenth century, when designing hairlines for eyepieces of optical instruments – to make their measurements more accurate! There is a longstanding controversy as to whether these are due to distortions of signals from the eyes, or whether they are cognitive in origin. It turns out that there are these two very different kinds of distortions, which may be hard to distinguish experimentally. The 'café wall' illusion (Figure 6) is almost certainly due to physiological signal distortion.

Figure 7 gives well-known examples of what are almost certainly cognitive distortion illusions. Among the best known, and easiest to measure, is the Müller-Lyer image (Figure 7(a)). The line between the outward-pointing arrows looks longer, even though both lines are the same length, as may be checked with a ruler – though the ruler may appear distorted! The illusions in Figure 7 are associated with perspective, signaling depth in flat pictures. The outgoing fins of the Müller-Lyer image may represent an inside corner of a room, the ingoing ones an outside corner of a building.

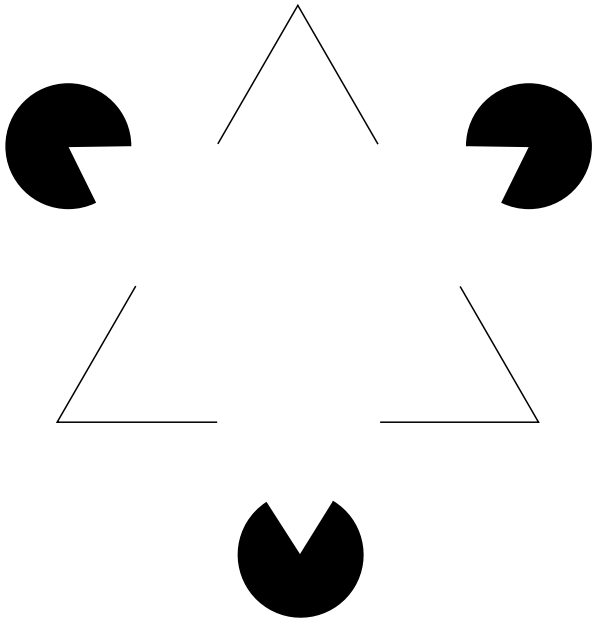


Figure 10. Fiction illusion: the 'ghostly triangle'.

The Ponzo illusion (Figure 7(b)) is a perspective of parallel lines, such as a road or railway. In all cases, features represented 'further' are seen as 'larger'. The idea is that size scaling is set by these perspective distance cues. Objects are normally seen as much the same size over a wide range of distance – although the retinal image shrinks with object distance. This shrinking with distance is normally compensated by active 'size scaling'. The idea, here, is that the perspective in flat pictures sets size scaling almost as for normal scenes (Figure 8), but as the pictures are flat, this normally useful compensation produces distortion – expanding features that are signaled as distant in the flat picture.

It is important to note that in pictures, depth cues such as perspective set constancy scaling even though the figure is seen as flat. When seen with appropriate depth (with a stereoscope) the distortions disappear. This shows they are cognitive phenomena. This is very different from the café wall illusion in Figure 6, in which the illusory wedge distortion occurs with no perspective or other depth cue. It is caused by the brightness contrast across the neutral 'mortar' lines pulling the edges together for half of each 'tile', forming small wedges, which integrate to give the long wedges across the figure. These 'physiological' processes are very different from the 'cognitive' distortions, given by depth cues mis-setting size scaling.

Paradoxes

Many perceptions are unlikely – some are impossible. Because probable alternatives of ambiguous figures are favored this is surprising, demanding explanation. It fits the view that perceptions are hypotheses generated by rules. Similarly, hypotheses of science can be paradoxical when generated by rules from incorrect assumptions. The best-known perceptual paradox is the 'impossible triangle', devised by the English geneticist Lionel Penrose and his son the cosmologist Roger Penrose (Figure 9). It can be an object, made of wood, but is paradoxical when seen from a certain position. This paradox arises from the visual rule that touching features are most likely to be at the same distance. Here this assumption is false, as the ends do not touch in the third dimension, though they do at the retina. This false visual assumption generates the paradox.

Table 1. Classification of illusions

Kinds	*Physical*		*Cognitive*	
	Optical	*Physiological*	*Rules*	*Knowledge*
Ambiguous	Cataract Mist and fog Some shadows	Retinal rivalry	Figure–ground Ames window 'Twin peaks' (Patric Hughes)	Vase–face Duck–rabbit Necker cube Hollow face Johansson dots
Distortion	Astigmatism (shape)	Café wall	Müller-Lyer	Size–weight (smaller object feels heavier)
	Mirage (of place)	Tilt after-effect	Ponzo	Mystery spots (tilted room)
		Color contrast	Poggendorff	Induced movement?
	Mirror reversal (physics: not optical)	Pulfrich pendulum	Zollner Sander's parallelogram Horizontal–vertical?	
Paradox	Looking-glass (oneself doubled)	Rotating spiral (expands without getting larger)	Impossible triangle (figure and model) Devil's fork Shepard's elephant	Magritte painting of man in mirror – the face appearing instead of the back of the head
Fiction	Rainbow	After-images (brightness and color)	Ilusory edges and surfaces (Schumann, Kanizsa)	'Faces in the fire'
	Moiré patterns	Mach bands Autokinetic movement Phi movement Bidwell's ghost		Ink blots

Fictions

The most famous fiction illusion is the 'ghostly triangle' of the Italian artist–psychologist Gaetano Kanizsa (Figure 10). Gaps may be actually gaps, or may be due to some nearer eclipsing or occluding surface, hiding part of the object. When vision postulates a nonexistent nearer object to 'explain' the gap, an illusory surface is seen. This process is generally useful: small blind regions of the eye (including the blind spots, where the optic nerves leave the retinas) are usefully filled in, to be invisible and so not distracting.

CLASSIFYING ILLUSIONS

Classifying phenomena is important in science. Illusions may be classified by appearances and causes. Some are caused by physical (optical) effects, between objects and the eyes; others by physiological disturbance of neural signals; others by cognitive misreading of neural signals. A classification in these terms is given in Table 1.

WHAT CAN ILLUSIONS TELL US?

- Perceptions are indirectly related to the object world.
- Perceptions are guesses – predictive hypotheses – which may be wrong (illusory) in characteristic ways.
- The main kinds of illusions are: ambiguities, distortions, paradoxes, and fictions.
- Phenomena of illusions can be used to discover principles of perception: adaptations (especially distortions) can tease out physiological channels; ambiguities can separate 'bottom up' signals from 'top down' knowledge; paradoxes can reveal cognitive rules and assumptions; and fictions can show rules of creativity.
- Some illusions are useful, especially for artists; some can be dangerous.
- Realizing that we are all subject to illusions might make us generally more tolerant and less dogmatic.

Further Reading

Gregory RL (1963) Distortion of visual as inappropriate constancy scaling. *Nature* **199**: 678–691.

Gregory RL (1970) *The Intelligent Eye*. London, UK: Weidenfeld.

Gregory RL and Harris JP (1975) Illusion destruction by appropriate scaling. *Perception* **4**: 203–220.

Gregory RL and Heard P (1979) Border locking and the café wall illusion. *Perception* **8**: 365–380.

Hoffman D (1998) *Visual Intelligence: How We Create What We See*. New York, NY: WW Norton.

Petry S and Meyer G (1987) *The Perception of Illusory Contours*. New York, NY: Springer.

Penrose LS and Penrose R (1958) Impossible objects: a special type of illusion. *British Journal of Psychology* **49**: 31.

Imagery

Introductory article

Maryjane Wraga, Smith College, Northampton, Massachusetts, USA
Stephen M Kosslyn, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

A mental image is an internal representation that produces the experience of perception in the absence of the appropriate sensory input.

INTRODUCTION

The term 'mental imagery' historically has been used in two ways. The first refers to the subjective experience of 'seeing with the mind's eye', 'hearing with the mind's ear', and so on. The second refers to the current usage in cognitive science, that defines mental imagery as an internal representation that gives rise to the experience of perception in the absence of the appropriate sensory input. This scientific sense of the term construes imagery as a perceptual representation, stored in short-term memory and corresponding to a particular brain state. The conscious experience of imagery is thought to be a byproduct of the brain state.

Imagery traditionally has been studied via several methodologies, ranging from observations of behavior (e.g. recording response times and accuracy levels of imagery task performance) to neuropsychological techniques (e.g. observing imagery deficits following brain damage; recording brain areas activated during imagery tasks). Because research on mental imagery has focused primarily on visual imagery, we will focus on this type of imagery in this article.

USES OF IMAGERY

Mental imagery is a versatile cognitive ability. In this section we examine its functions, as well as situations in which it appears to be used spontaneously.

Predicting the Outcomes of Actions

One of the main functions of imagery is to help people reason and solve problems by allowing them to predict the outcome of a given action. Such actions can involve the imagined movement of objects or the observer. For example, when packing up a moving van, individuals can visualize the most efficient arrangement of their belongings before actually placing any of the items. In attempting to cross a rock-strewn stream, the safest route can be plotted by imagining jumping along the path of rocks that best affords a dry, sturdy crossing. In each case, the observer creates an imagined scenario and 'watches' what happens in order to solve the problem.

Creating Mental Models

Imagery can also be used to reason on a more abstract level. For example, consider the following problem: Rich is older than Pete, and Paul is younger than Pete. Who is the oldest? One way to solve the problem is to imagine each person as a dot appearing sequentially on a vertical line. If the dots are ordered from bottom to top according to chronological age, then solving the problem becomes the simple matter of observing which dot is at the top. Such strategies may aid in language comprehension by allowing one to create a mental model of what is said.

Visualizing and Retrieving Memories

Another purpose of mental imagery is to help one remember information that may not have been noticed in detail at the time a memory was encoded. For example, when asked a question such as 'What geometric shape most closely resembles a cat's ear?' most observers report using imagery to arrive at the answer of 'triangle'. Such information is not likely to have been considered explicitly before. Thus, the answer is not readily accessible: it has

not been stored verbally or in any other explicit manner. Instead, the information is available implicitly in a mental image of a cat, and it is retrieved from memory via imagery. One forms the image and then 'inspects' it to find the answer.

Learning

As discovered by the ancient Greeks, mental imagery can help one memorize a list of objects. For example, to remember what items to purchase at the grocery store, one can employ the 'method of loci', in which the individual items are imagined as if they were placed along a familiar route. In this case, let's say that the route is the path from your home to the grocery store. Using items from the grocery list, you might imagine a loaf of bread at a letter box, a carton of eggs near a lamppost two blocks farther down, and so on. In order later to recall the entire list, you would need simply to visualize the route, imagine walking along it, and 'see' what is stored at each landmark. In general, items visualized in an interactive fashion are remembered best (e.g. imaging the loaf of bread stuck in the slot of the letter box rather than sitting beside it). Creating interactive images allows one to utilize the associations among objects in addition to their individual names and images. Such associations have been shown to aid in later memory retrieval.

Imagery can also help one to improve skills. In trying to improve a tennis serve, for example, you can watch a professional player and then mentally replay his or her actions, 'observing' them in order to improve your own performance. Or you can replay and 'observe' your own tennis serve, tuning the stored information that guides the physical performance of that action.

Insights from Daydreaming

Kosslyn and colleagues asked people to keep track of their imagery on an hourly basis, and found that most reported imagery occurred in daydreams. However, for some subjects, even such seemingly aimless images proved useful. The subjects occasionally noted that their daydreaming suddenly gave them a new idea or revealed an important oversight in a previous situation. Such insights may occur because mental images contain information that is stored implicitly in memory. Once objects and events are visualized, one may observe features and relationships that previously went unnoticed.

FUNCTIONAL PROPERTIES OF IMAGES

Most uses of imagery rely on the ability to transform or manipulate images. In this section we discuss three ways of manipulating images that cognitive scientists have studied.

Scanning

Just as one can scan a physical scene with one's eyes, it appears that images can be scanned with the 'mind's eye'. For example, Kosslyn and colleagues showed subjects a drawing of a map and had them memorize the locations of landmarks such as a hut, palm trees, and a well. Subjects then formed an image of the island and mentally focused on a specific landmark. The subjects then heard the name of a second, target landmark, which may or may not have appeared on the actual map. The task was to locate the target on the image. The farther away targets were from the initial landmarks, the longer it took subjects to locate them. This finding makes sense if images are depictive, (i.e. they preserve spatial extent), and thus requiring more time to shift attention over larger distances. Not all researchers agree with such an interpretation, however. In fact, the findings of scanning experiments have been a focus of contention in a long-standing debate on the nature of image representation. In the section on rotation, we discuss an alternative theory to the depictive-imagery view.

Zooming

In order to 'see' an imaged object in finer detail observers can 'zoom in' on it. For example, suppose you visualize a bird perched high up in a distant tree. If you are asked to indicate whether the bird has a crest on its head, you can 'zoom in' closer to the image to see this feature more clearly. Experiments have shown that the amount of time it takes to answer such questions is dependent on the initial 'viewing distance' of the observer from the visualized object. The farther away the object appears in the image, the longer it takes to report the correct answer. Further evidence comes from Malstrom and colleagues, who demonstrated that when participants imagine objects at different distances, their eyes actually change shape in accordance with distance, much as they do when observers look at physical objects.

Rotation

Just as observers require time to scan or zoom in on an imaged object, they also require time to transform imaged objects. For example, Roger Shepard and colleagues presented subjects with depictions of pairs of three-dimensional multiarmed objects, one of which was at a different orientation with respect to the other. The task was to decide whether the objects were the same or different. The time to make a response increased linearly with the angle of displacement between the two objects. Shepard and colleagues interpreted this finding as evidence that subjects had mentally rotated one object into alignment with the other. (*See* **Mental Rotation**)

Perhaps the most intriguing aspect of the findings on image transformations is that the internal processes are analogous to the corresponding physical process. In the real world, physical properties such as distance and the fact that objects cannot change position instantaneously constrain what we can do – but such physical properties are not actually present in mental images. For example, mental images are not rigid objects that must pass through a trajectory when rotated. So why do images mimic actual objects? Although evolutionary and information-processing theories have been posited to explain this correspondence, no one theory has compelling support.

An alternative account is that mental images do not mimic actual objects at all. According to this view, proposed by Pylyshyn and others, the image transformation findings that show a linear relationship between imaged and physical object movement are an artifact of observers' reliance on tacit knowledge of physical laws and object properties. For example, observers know through experience that it takes less time to rotate an object a short distance than a longer distance. Pylyshyn proposed that such knowledge is stored in the mind in an abstract, languagelike form (in representations termed 'propositions') rather than as picturelike images. Access to propositional information allows observers to treat the corresponding imagery task as a kind of pretend version of real rotation. In this case, they simply wait longer before responding when they think it would have taken longer to perform the corresponding actual operation (such as rotating a figure a greater amount). Pylyshyn's theory was important in that it inspired imagery researchers to construct experiments aimed at defining the precise structure of mental images. However, most researchers in the field today tend to hold the view that images involve both depictive and propositional forms of representation. (*See* **Representations, Abstract and Concrete**; **Representation Formats in Psychology**)

DISSOCIATIONS BETWEEN IMAGERY AND LANGUAGE

Although nonhuman animals may rely on imagery as the major vehicle of thinking, we humans have another means: language. How are imagery and language related? Each type of representation characterizes different types of information. For example, visualize a dog sitting under a tree. Depending on your degree of imagery skill, you will be able to 'see' concrete characteristics of the object, such as the shape of the dog, the shape of the tree, and the location and size of the dog with respect to the tree. Notice that the latter relationship is inherent in the image; nothing about the spatial characteristics of the dog relative to the tree must be inferred. Now consider the sentence, 'The dog is sitting under the tree.' Nothing about the manner in which the words 'dog' and 'tree' look or sound resembles their physical counterparts. In addition, the spatial relationship of the dog to the tree does not exist without the explicit specification provided by the preposition 'under'. The relationships between the words and their meanings are completely arbitrary.

What is the functional significance of having two types of representation? You may rely upon one or the other form, depending on what information is needed. For example, in answering the question, 'Are dogs animals?' one would resort to the linguistic representation of information associated with dogs; a mental image of a dog would not be helpful. On the other hand, the question 'Which are bigger, the dog's front or hind legs?' might be best answered by relying on an image of the dog. Such distinctions have led Paivio and colleagues to hypothesize that information is encoded in two forms, as images and as words. Empirical support for this idea is found in studies where verbal tasks are shown to interfere with verbal encoding, whereas pictorial tasks interfere with visual imagery encoding.

Despite the differences between imagery and language, the two may interact with each other. For example, Barbara Tversky and colleagues, as well as Michel Denis and his colleagues, have shown that representations of scenes constructed from verbal descriptions tend to preserve the spatial relationships occurring in the scene. The time it took subjects to indicate the relative

locations of objects described in a text was directly related to their imagined position in the scene. Imagery is also used to answer certain types of questions, particularly about subtle visual properties that cannot be inferred from other information – such as the shape of a cat's ears, the number of windows in your living room, or the precise color of a cantaloupe.

SIMILARITIES/DIFFERENCES BETWEEN IMAGERY AND PERCEPTION

Whereas imagery and language are largely distinct, there is ample empirical evidence that visual imagery shares many of the same mechanisms as visual perception. For example, visual images interfere with visual perception more than with auditory perception; the opposite is true for auditory images. Indeed, visual images can sometimes be mistaken for physical stimuli. Moreover, visual illusions sometimes occur in imagery. For example, we tend to have difficulty seeing oblique lines, compared to vertical or horizontal ones – and the same thing occurs in imagery.

However, the correspondence between imagery and perception is not perfect. Experiments using reversible figures, which perceptually can be construed as more than one object (such as the classic duck/rabbit; see Figure 1), have demonstrated limitations to the imagery–perception analogy. Subjects were briefly shown a reversible figure and were asked to arrive at an alternative interpretation from that initially perceived. Most subjects had difficulty with this task. It was only after they were allowed to draw the figures from memory that they were able to arrive at the alternative interpretation. Later studies showed that specific hints provided by the experimenter, such as directing the subjects to think of the object's facial profile as the back of the new object's head, helped them to discover the alternative interpretation. However, this ability did not reach the degree of spontaneous reinterpretation that occurs with an actual visual image. Images fade quickly and have an inherent organization; if the organization must be changed substantially, it can be difficult to retain the image long enough to complete this process.

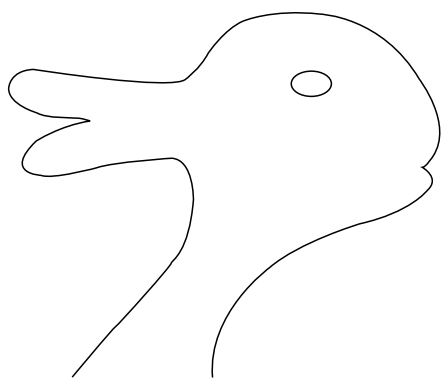


Figure 1. A classic reversible figure, which can be seen either as a duck or a rabbit.

THE NEUROPHYSIOLOGICAL BASIS OF IMAGERY

Both studies of brain-damaged patients and neuroimaging experiments have shown that imagery shares most of the same neural machinery with perception. Thus, it is not surprising that brain damage often has corresponding effects in imagery and perception. For example, patients with damage to the right parietal lobe of the brain often ignore objects in the left half of the visual field. Thus, a patient standing in front of a building will not notice its left side. These patients will have the same difficulty if asked to close their eyes and visualize the building. However, if the patient is asked to turn so that the previously unattended side of the building is now in the perceivable half of space, he or she will now be able to 'see' it, both in actual perception and in forming a mental image.

In the past two decades, the advent of neuroimaging technology has led to much progress in our understanding of the neural underpinnings of mental images. One finding indicates that about two-thirds of the brain areas that are activated by either perception or imagery are activated in common by the two functions. Of particular interest is the finding that the primary visual cortex, which is the first part of the cortex to receive input from the eyes, can also be activated during mental imagery – even when the eyes are closed. This finding is important in part because it suggests that imagery can modulate perceptual processing in the brain from the start, thus determining not only what we see but what we remember seeing. Thus, it is not surprising that imaged objects are often mistakenly remembered as actual objects. We can 'see' objects in visual images precisely because they are generated as patterns of activation within the areas of the brain used by visual perception.

CONCLUSIONS

Images are important for two major reasons. First, they make use of the brain machinery used in

perception. Thus, we can re-present objects and events, and notice characteristics that previously escaped notice. Second, because images can stand in for objects, we can mentally transform them and observe the likely consequences of performing the corresponding actual transformation. Mental imagery is a world within, and can be used to help us navigate, learn about, and master our interactions with the world itself.

Further Reading

Bisiach E and Luzzatti C (1978) Unilateral neglect of representational space. *Cortex* **14**: 129–133.

Farah MH (1988) Is visual imagery really visual? Overlooked evidence from neuropsychology. *Psychological Review* **95**: 301–317.

Finke RA (1989) *Principles of Mental Imagery*. Cambridge, MA: MIT Press.

Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.

Kosslyn SM, Cave CB, Provost D and Von Gierke S (1988) Sequential processes in image generation. *Cognitive Psychology* **20**: 319–343.

Kosslyn SM, Segar C, Pani J and Hillger LA (1990) When is imagery used? A diary study. *Journal of Mental Imagery* **14**: 131–152.

Kosslyn SM, Thompson WL, Kim JJ and Alpert NM (1995) Topographical representations of mental images in primary visual cortex. *Nature* **378**: 496–498.

Paivio A (1971) *Imagery and Verbal Processes*. New York: Holt, Rinehart and Winston.

Shepard RN and Cooper LA (1982) *Mental Images and their Transformations*. Cambridge, MA: MIT Press.

Implicit Learning

Introductory article

AS Reber, Brooklyn College of the City University of New York, New York, USA

CONTENTS

Implicit learning is essentially learning without awareness. Knowledge that has been acquired implicitly is knowledge that has been acquired and held largely without conscious effort.

INTRODUCTION

Implicit learning is the capacity to pick up information about complex stimulus displays largely without awareness of either the process or the products of learning. People generally display implicitly acquired knowledge by improved performance of various tasks independently of whether they can report knowing the material that has enabled them to perform effectively. In many ways, it is akin to the layperson's notion of intuition. When people make what they regard as 'intuitive' judgments or guesses they are, in fact, using considerable knowledge about the particular situation within which the judgments or guesses are made, and typically find themselves unable to articulate the nature of that knowledge.

In the normal course of events, implicit acquisition mechanisms can be seen operating in virtually every interesting thing that human beings do. Natural languages are acquired with substantial contributions from implicit acquisitional mechanisms. Children do not learn to speak the languages of their social community by testing hypotheses or carrying out 'top-down' analytical procedures. Rather, they gradually induce the underlying structure of the various structured components of the rich and diverse communicative network – specifically the semantic, syntactic, pragmatic, gestural, and metalinguistic elements of communication. Similar arguments can be made for other complex acquisitional frameworks such as socialization and acculturation.

In principle, the mechanism that underlies implicit learning is the detection of patterns of covariation expressed in the stimulus displays. Relatively simple patterns of covariation detection are seen virtually everywhere. The most commonly studied example of implicit learning is classical conditioning, where an organism acquires an associative link between a once-neutral conditioned stimulus (the CS) and another unconditioned stimulus that reliably elicits a response (the US). Importantly, even in this relatively simple 'bottom-up' system, the acquisition is easily shown to be the detection of genuine covariation and not mere co-occurrence. As has been often and famously noted, Pavlov's dogs only learned to respond to the CS, not to Pavlov. In the original study the bell became a reliable predictor of the arrival of food because it covaried with food. Pavlov, because of his presence whether or not food was forthcoming, merely co-occurred with food. Even at this simple level the fundamental associative process can be seen operating.

As this example suggests, it is important that implicit learning be seen in an evolutionary perspective. Mechanisms that operate largely independently of consciousness, as implicit learning appears to, are probably supported by neurological systems that are old evolutionarily and antedate those 'top-down' systems that are dependent on complex encoding and conscious control, the hallmarks of explicit systems. This evolutionary argument has a number of important implications that provide insight into how implicit and explicit acquisition and representational systems can be dissociated from each other; these will be explored in more detail below.

WHAT INFORMATION CAN BE LEARNED IMPLICITLY?

As the above discussion suggests, implicit mechanisms operate in a wide variety of domains, from the distinctly primitive (classical conditioning) to the compellingly complex (some aspects of natural language acquisition). Exactly what kinds of information can be acquired using implicit mechanisms is

unknown – although the current literature gives some hints.

Basically, implicit learning mechanisms are 'bottom-up' systems. They function by picking up patterns of covariation in environmental displays. Such mechanisms are rather easily simulated by connectionist architectures that capture and represent these covariations. These acquisitional mechanisms are not particularly effective in settings that involve problem solving, hypothesis testing, or creative extrapolation. In short, they are mechanisms typically not associated with processes normally thought of as 'smart'. Yet, there are good reasons for suspecting that implicit mechanisms are not 'stupid'. Considerable evidence suggests that they are powerful and capable of carrying out sophisticated inductions, resulting in representations that seem to have at least some abstract and symbolic features. This issue is the focus of rather intense debate and will be discussed in more detail below.

TESTS OF IMPLICIT LEARNING

Various tasks have been used in the study of implicit learning. Perhaps the simplest are those based on the 'hidden covariation' procedure of Lewicki and his colleagues. In these studies simple elements of a complex display covary, and implicit learning is assessed by participants acquiring a sensitivity to these covariants. Most other procedures use more complex settings. A good bit of early work was done by Broadbent and his colleagues, using a 'production control' task. Here, subjects had to adjust complex relationships between such factors as worker satisfaction, pay, and productivity in order to maximize the efficiency of a mythical production plant. Other procedures have used processes as diverse as the detection of phonetic patterns in auditory displays, the development of preferences for structured visual and melodic sequences, responding rapidly to repeating sequences of stimuli, and artificial grammar learning where the stimuli consist of strings of symbols whose order is determined by complex probabilistic rule systems.

While diverse in their superficial forms, all of these procedures share common features. Stimulus displays always contain some degree of regularity or lawfulness. Subjects approach the tasks in a relatively neutral mode in the sense that they are not informed about the structured nature of the displays, usually treating them as memorization tests or motor learning tasks. Participants learn or become sensitive to the environmental regularities largely without conscious knowledge of either the process or the products of acquisition. Of the many procedures in use, two have dominated the attention of workers in the field, the sequence learning procedure and the artificial grammar learning experiment. The following sections take a closer look at these and outline the general findings.

Sequence Learning

The most commonly used procedure here is the sequential (or serial) reaction time (SRT) task introduced by Nissen and her colleagues. The task, at least on the surface, is simplicity itself. Participants sit in front of a monitor on which a number of locations are marked (usually four but occasionally five or six). A series of target stimuli (typically an asterisk or other symbol) appears at the locations. The participant's task is to press the key that corresponds spatially with the location of each target as quickly and accurately as possible.

In the canonical procedure the sequence of target locations displays some degree of regularity. In Nissen and Bullemer's original study and in most of those run since, a repeating sequence with particular statistical properties is used, although several investigators have used a nonrepeating sequence where a complex set of rules dictates successive locations. The classic finding is straightforward and remarkably robust. Reaction times (RTs) decrease as subjects learn to exploit the structure in the target sequence. To show that this is more than just the learning of a motor task, the introduction of a block of trials where the sequence is modified in some fashion produces an increase in RTs.

A common variation uses a secondary task. In this 'dual-task' experiment, somewhere during the interval between the key press and the next target, either a high- or low-pitched tone sounds. Participants must also keep an accurate running count of how many, say, high-pitched tones occurred in each block of trials. The extent to which this tone-counting task does (or does not) compromise learning is evaluated. The dual-task protocol has given the SRT experiment a special cachet as a forum for evaluating models of such diverse phenomena as implicit learning, divided attention, working memory, and the form of memory representations.

The reasons for this attention are easy to specify. First, the basic procedure is a model environment for the study of implicit learning. In these studies participants generally proceed unaware of the structured nature of the sequence. In cases where explicit knowledge can be shown to exist, it is

typically insufficient to account for the full range of behavior observed. Second, by manipulating aspects of the tone-counting task the role of attention in learning can be readily explored. Third, by manipulating the statistical characteristics of the sequence, the procedure becomes a forum for examining the impact of complexity and the contributions of working memory on learning. Fourth, the reliability of the basic findings makes it a useful vehicle for exploring the neuroanatomic underpinnings of implicit learning. Finally, because of the large and reliable data base, it is an excellent platform for testing theoretical models of learning and representation.

While there are dissenting voices, the general consensus is that the learning manifested in these tasks is classic implicit learning. Subjects are generally unaware of the nature of the sequence, particularly in experiments where the sequences are generated by complex rules. The acquisition of knowledge is accompanied by other markers of implicit functioning such as the development of preferences for structured displays over those that violate the rule. And intact functioning is observed in a variety of patient populations who show severe deficits in more controlled, top-down cognitive tasks.

Artificial Grammar Learning

The typical artificial grammar (AG) task involves the use of a complex set of rules that dictates the order of a string of arbitrary symbols. The most commonly used AGs are based on conditionalized rules such as 'strings may begin with either a T or a V', 'if a T occurs in the first position, it can only be followed by another T or a Q', and so forth. Although these AGs can, like natural languages, produce strings of unlimited length, most experiments have subjects work with items that are between three and eight symbols long.

In the standard AG learning experiment, subjects are presented with a series of exemplary strings from the grammar and asked to memorize them. Following this learning phase they are informed (for the first time) that the letter sequences were, in fact, rule-governed and are asked to carry out a grammaticality task. Here, they are presented with a series of novel letter strings, some of which follow the rules of the AG and some of which contain a violation, and asked to judge how well formed these items are. Evidence of learning is successful classification of novel strings at rates above chance and above what control subjects who never experienced the learning phase accomplish.

The data from literally hundreds of studies using AGs look remarkably like that from the SRT studies. Participants learn a considerable amount of knowledge about the underlying structure of the symbol strings. The learning appears to be taking place largely independent of awareness, and subjects find it particularly difficult to communicate to others the knowledge they are applying. Developmental level appears to be no barrier since young children, adults, and aged populations all perform at roughly the same levels. And, as with the SRT task, participants develop a marked preference for items that follow the rules over those that do not – including items that are completely novel.

Perhaps not surprisingly, the same fundamental arguments can be found. Despite the converging lines of evidence here, dissenting views exist. These alternative perspectives tend to argue either that the tests of subjects' awareness of the underlying rules have not been carried out with sufficient care, or that the learning is trivial and does not reflect any particularly sophisticated cognitive processing. The next sections will examine these issues and attempt to put them into a framework based on basic principles of evolutionary biology.

CONTINUUM OF IMPLICIT–EXPLICIT REPRESENTATION

There has been an unfortunate tendency to view the implicit–explicit issue in an either/or framework: a particular experiment taps either implicit mechanisms or explicit ones; a memory representation is either available to consciousness or it is not. This polarization tendency is almost certainly a mistake and has clouded issues in unhappy ways. Virtually everything cognitively interesting that people do is a complex blend of consciously controlled processing that is declarative in nature and implicit, automatic functioning that lies largely outside of awareness. Task purity is going to be hard to find.

Moreover, these systems and the underlying neuroanatomical mechanisms that subsume them have a distinct evolutionary history, and any theory of implicit and explicit functioning must cohere with basic principles of evolutionary biology. As has been argued in several places, explicit systems, linked as they are with conscious control, are relatively recent arrivals on the evolutionary scene. Implicit systems, which function procedurally and largely independent of top-down control, are much older. In fact, they are almost certainly as old as the original neural nets since organisms as primitive as *Aplysia californica*, a sea slug of

most modest intellectual achievements, can, nevertheless, detect patterns of covariation in its environment and exhibit differential Pavlovian conditioning.

It is virtually certain that as neural systems grew in complexity and natural advantages accrued to systems that exhibited modulatory control, what we like to call 'consciousness' emerged. But this self-reflective, top-down system with all of its cognitive richness does not neutralize or diminish the importance of the older, bottom-up mechanisms. They are still very much present and, by virtue of the hierarchical nature of evolutionary processes, form the foundation upon which the more recently emerged systems function. In short, we expect to see a blending of the implicit and the explicit in virtually all complex human behaviors.

However, this line of argument still leaves unanswered the question of representation. There is no doubt (except, perhaps, in the minds of radical behaviorists) that human consciousness forms representations that are abstract and symbolic. However, no such consensus exists when it comes to the cognitive unconscious. Perruchet and Pacteau, in a recent critique of the representation issue in implicit learning, put the matter quite starkly: 'We do not question human abstraction ability, no more than we question the existence of unconscious processes. What we do question is the joint possibility of unconscious abstraction'. There are, to be sure, difficulties with this point of view.

The term 'abstract' comes from the Latin for drawn away and usage turns on the notion that for qualities of objects, events, or phenomena to be abstract is to regard them as separate from the specific objects, events, or phenomena themselves. 'Symbolic' has a more checkered etymology, but the essential notion is similar. An action, an idea, an utterance, a mental state can be regarded as symbolic if it signifies actions, ideas, utterances, or mental states beyond or distinct from the original instance.

But neither of these notions has firm definitional boundaries. Being abstract is not an either/or situation. Many implicit learning experiments find subjects forming representations with a measure of abstractness or symbolic content. This is commonly seen in the experiments on transfer where subjects learn an AG instantiated with one set of letters and then make judgments on the grammar of strings composed of a different set of letters. In fact, as Manza and Reber showed, abstract representations can be encouraged simply by presenting the learning stimuli in two different instantiations. When subjects memorize strings like XXRTRXV and QQWMWQP where individual letters have shared privileges of occurrence, they tend to set up representations that are not tied to the physical form of the inputs.

In short, there is no 'default' form of mental representation. Implicit representations, like the explicit ones, can be either abstract and contain symbolic content, or they can be rigid, inflexible, and tied to the form of the stimulus inputs. It is likely that the implicit tend more toward the inflexible simply because they are dependent on older and more primitive neuroanatomical systems. But it would be a mistake to conclude that because of this evolutionary antiquity that the systems must be ineluctably tied to the stimulus inputs. The patterns of covariation among the elements of the environments that we live in are under constant flux. Any representational system, even the most basic and cognitively unadorned, must be sensitive to these variations. And since implicit learning is essentially the detection of patterns of covariation among elements in stimulus displays, to be viable the representational mechanism must be able to capture random adjustments in form and structure.

ARE IMPLICIT AND EXPLICIT MEMORY SYSTEMS SEPARATE?

This question contains two different elements. First (echoing the preceding), are there reasons for suspecting that implicitly formed representations are different in kind from explicitly formed? Second, are there distinct anatomical systems for the implicit and the explicit?

Differences between Implicit and Explicit Memories

There is a tendency to think that implicit representations are inflexible and tied to the form of the stimulus input, while explicit representations are abstract and flexible. And, as noted, while on occasions they may take these forms, it is unlikely that they must. The literature on implicit memory strongly suggests that unconsciously held representations are rigid and tied to the form of the stimulus environment. The literature on implicit learning suggests that such representations are often rather flexible and abstract. These seemingly conflicting findings are unhappily confounded with methodological difficulties.

In the typical implicit learning experiment something new is learned. Subjects acquire knowledge about a structured sequence of events, a set of rules

for symbol ordering, an obscure or hidden pattern of covariations and, for the most part, this knowledge is held without awareness of its content. As noted above, the form of this knowledge may be flexible or inflexible, abstract or concrete, symbolic or tied to the form of the input. Because material is learned and knowledge is acquired, task demands, encoding constraints, context effects, and the like will play a large role in determining the nature of the representation.

On the other hand, in the canonical implicit memory experiment *nothing new is learned*. The typical study consists of the presentation of knowledge subjects already possess – a list of words, a set of pictures – all of which were part of the subjects' knowledge base when they entered the experiment. Under these conditions, subjects tend to form episodic memories based on mnemonics or other encoding tags that emphasize each item's uniqueness. The results of such a process are memory representations that are concrete in nature and close to the physical form of the input stimulus. Hence, it is not surprising that degrading or modifying superficial features of probe stimuli reduces evidence of implicit memory. Findings such as these have led to the conclusion that implicit memory representations are relatively inflexible.

In short, the question of whether implicit and explicit memories can be regarded as fundamentally distinct based on the nature of the encoding and the resulting representations is, to date, unanswerable.

Differences in Anatomical Structures for Implicit and Explicit Systems

The other line of attack on this issue, however, has proven more fruitful. One of the corollaries of the standard model of evolutionary biology is that older systems, because they form the foundation for later developments, should be more robust in the face of disorders and dysfunctions than more recently evolved systems. The natural implication of this argument is that we can expect to see a pattern in the lost and preserved functions that accompany particular disorders. There is an enormous literature here virtually all of which points to exactly this picture. Specific neural correlates of explicit processing are known, and damage to them typically produces little or no impairment of related implicit functioning. Implicit learning and implicit memory processes remain intact in a large number of psychiatric and neurological disorders where explicit cognitive functions are severely impaired.

Patients with anterograde amnesia who show an inability to acquire new explicit memories typically have damage to one or more areas in the medial temporal lobes. Nevertheless, these patients exhibit the same artificial grammar learning ability as control participants, even when the test is given using a novel letter set applied to the same underlying grammar. These patients are also unaware of the underlying rule structure of the AGs and are impaired at recognizing the training strings compared with control participants. Amnesic patients also show normal learning on an SRT task and, not surprisingly, are impaired in recognizing or reporting the sequence.

In a similar fashion, amnesic patients show virtually normal implicit memory. Indeed, the first, critical insight into the neural basis of consciousness in learning and memory came from the assessment of the famous patient H.M. This patient became profoundly amnesic following surgery (to control otherwise intractable epilepsy) that removed, bilaterally, most of the medial-temporal lobe including the hippocampus, amygdala, and surrounding cortical regions. The striking result of this surgery was a profound anterograde amnesia that essentially eliminated H.M.'s ability to acquire new explicit knowledge of everyday facts and events while leaving his general cognitive functions largely intact. For example, his memory impairment is such that one half-hour after eating lunch, he cannot remember a single item that he ate or, in fact, that he ate at all. But, his memory impairment is selective for encoding new explicit, long-term memories. His immediate memory (e.g., digit span) is normal, he is able to retrieve explicit memories for childhood, and his post-surgery IQ is above normal and his implicit learning and memory systems function well within normal limits.

Psychotic patients whose overall functioning is so poor that they can barely solve simple numerical problems perform at virtually normal levels on AG learning tasks. And, while not exactly in the category of neurologically impaired, infants show intact implicit learning of phonetic patterns and are virtually indistinguishable from adults in their ability. In aged populations, implicit processing is almost invariably intact even as consciously controlled cognitive functions begin to decline. Performance on the SRT task doesn't begin to diminish until extreme old age and, even then, the elderly display performance levels much closer to that of younger adults than then do on explicit tasks such as problem solving, reasoning, and long-term memory.

One hypothesis is that implicit learning and memory occur within the cortical areas involved in processing the stimuli for which the learning is occurring, and hence will be found virtually throughout the brain. If this model proves correct, only with diffuse lesions would one expect global impairments in implicit learning. Focal lesions would result in impairment only for those implicit learning tasks supported by the affected area, with other implicit learning functions still being operative. As suggested above, one of the hallmarks of implicit learning is that it is less vulnerable than explicit learning to neurological damage. If implicit processing is supported by widely divergent and diffuse neural structures, then this robustness is to be expected. In short, the neurocognitive literature strongly suggests that the explicit and implicit systems are distinct.

CONCLUSION

The evidence that has accumulated over the past several decades of research supports the notion of an evolutionarily old acquisition and representation system that operates largely independent of consciousness and of top-down control mechanisms. This implicit system operates primarily by picking up and encoding patterns of covariation in the stimulus environment. It is a relatively robust mechanism that shows considerable resilience to psychological and neurological disorders, displays little variation across developmental level, and enables the establishment of memorial representations that, under the proper circumstances, can capture rather sophisticated features.

Acknowledgment

This article was prepared while the author was supported by Grant 0113025 from the National Science Foundation.

Further Reading

Berry DC (ed.) (1997) *How Implicit is Implicit Learning?* New York, NY: Oxford University Press.

Berry DC and Dienes Z (eds) (1993) *Implicit Learning: Theoretical and Empirical Issues*. Mahwah, NJ: LEA Press.

Cleeremans A and French R (eds) (2001) *The Role of Implicit Learning in Representing the World*. London, UK: Psychology Press.

Reber AS (1993) *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. New York, NY: Oxford University Press.

Stadler MA and Frensch PA (1998) *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage.

Individual Differences

Introductory article

David Lubinski, Vanderbilt University, Nashville, Tennessee, USA
Rose Mary Webb, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

Variation among individuals is observed on various behavioral attributes, including global dimensions of human cognitive ability, interest, and personality, as well as more specific attributes. The scientific study of the nature and causes of human variation is known as the study of individual differences or differential psychology.

INTRODUCTION

Individual differences research focuses on the systematic assessment of relatively stable attributes measured for the purpose of allowing longitudinal forecasts (behavioral predictions over time) pertaining to broad behavioral patterns and proclivities. Dimensions of human abilities, interests, and personality play critical roles in structuring a variety of important behaviors and outcomes (e.g. educational and vocational choice, academic achievement, occupational performance, income, health-risk behavior, crime and delinquency). Since particular attention is devoted to socially relevant phenomena, individual difference dimensions are important variables to consider in various areas of social and medical research, especially those that examine people at risk for negative outcomes or those showing promise for positive outcomes.

The preferred method of assessing individual differences has been to build scales based on the aggregation of many multiple-choice items. This method is based upon the observation that, although an individual item on a scale possesses only a sliver of genuine information about an attribute, a series of lightly correlated items can combine to yield an informative measure of the attribute under analysis. This aggregation is analogous to the information one gleans from an individual's grade in a single college course, as compared with knowledge of that individual's college grade point average over a four-year period.

DISPOSITIONS

Organization of Cognitive Abilities

The dominant scientific conceptualization of intelligence views cognitive abilities as being organized hierarchically. This organizational scheme is clearly revealed by Carroll's factor-analytic examination of decades of ability research. In this framework, the attribute of general intelligence (g) represents what is common to cognitive tests. That is, g is what is common in the secondary level of more content-specific abilities, such as quantitative reasoning, spatial visualization, and verbal ability. These are in turn supported by more circumscribed abilities closely associated with specific tasks, such as numerical facility, reading speed, memory span, and reading comprehension. Such molecular skills are most closely associated with the manifest content of cognitive tests.

The attribute g represents approximately half of what is common among many different kinds of cognitive (aptitude and achievement) tests. It is a distillate of the elements common to each cognitive test. In terms of its external connections, g reflects the largest dimension of psychological diversity revealed by differential psychology. Since the early twentieth century, large-scale studies have documented the utility of g in forecasting educational outcomes, occupational training, and work performance. To a lesser degree, g is related to a number of diverse phenomena: for example, it is positively correlated with altruism, sense of humor, practical knowledge, response to psychotherapy, social skills, and supermarket shopping ability, and negatively

correlated with impulsiveness, accident-proneness, delinquency, smoking, racial prejudice, and obesity. (*See* **Academic Achievement**)

There are important abilities beyond the general factor. Mathematical, spatial–mechanical, and verbal reasoning abilities have all been shown to have psychological import beyond *g*. This is especially relevant in predicting individual differences in performance across educational and vocational domains, and also for predicting educational and career paths that people self-select (different learning niches). For example, a recent study compared three different types of profoundly gifted individuals identified as belonging to the top 0.01% on either mathematical or verbal reasoning ability. Initially, they were assessed at age 12 on the Scholastic Assessment Test (SAT), and scored either 700 or more on SAT-Mathematics or 630 or more on SAT-Verbal (several participants met both criteria). These participants were tracked over 10 years, and assigned to the following groups for analysis: those with highly advanced mathematical reasoning ability, relative to their verbal ability (high-math); those with highly advanced verbal reasoning ability, relative to their mathematical ability (high-verbal); and those whose mathematical and verbal reasoning abilities were more uniformly advanced (high-flat). Differential interests among these groups were apparent in their choice of favorite courses in high school and college (see Figure 1). High-math individuals tended to prefer mathematics and science courses, whereas high-verbal individuals tended to prefer humanities courses. Course preferences for high-flat individuals were intermediate. (*See* **Quantitative Reasoning**)

Awards and other special accomplishments of these groups, defined by differing ability profiles, demonstrated a similar pattern: high-math individuals tended to succeed in areas of science and technology, whereas high-verbal individuals tended to succeed in the humanities and arts (see Table 1). Again, high-flat individuals were intermediate. These findings further emphasize the importance of specific abilities beyond general intelligence. (*See* **Intelligence**)

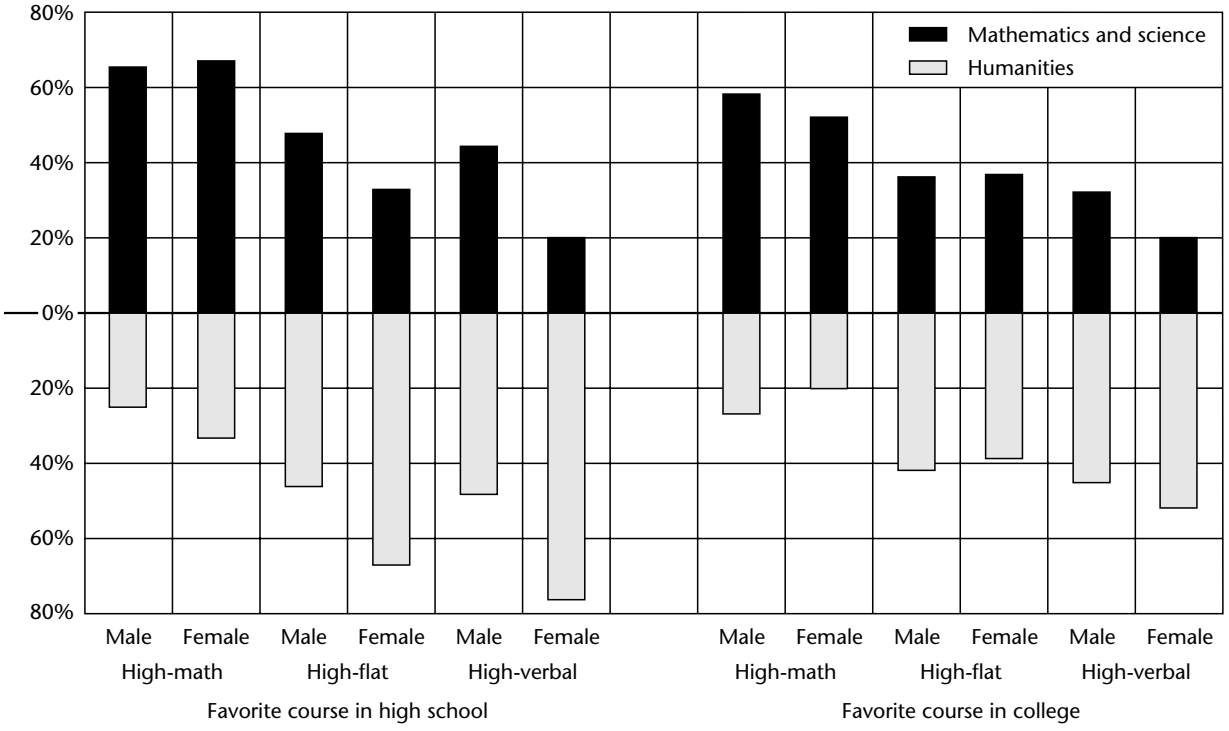


Figure 1. Favorite courses in high school and in college of participants in a study of individuals who at the age of 12 scored very highly in SAT-Mathematics ('high-math'), SAT-Verbal ('high-verbal'), or both ('high-flat'). Percentages in a column do not necessarily sum to 100% because only participants indicating either mathematics and sciences or humanities courses are displayed. Significance tests for differences among groups for favorite course are as follows: high school mathematics and science χ^2 (2, $N = 320$) = 20.7, $p < 0.0001$; college mathematics and science χ^2 (2, $N = 320$) = 18.2, $p < 0.0001$; high school humanities χ^2 (2, $N = 320$) = 36.6, $p < 0.0001$; college humanities χ^2 (2, $N = 320$) = 30.2, $p < 0.0001$.

Table 1. Special accomplishments as listed by the participants. Numbers in parentheses after an accomplishment represent the number of participants associated with the accomplishment. All other entries are for a single individual

	Sciences and technology	*Humanities and arts*	*Other*
Accomplishments	• Scientific publications (11) • Software development (8) • Inventions (4) • National Science Foundation fellowship (2) • Designed image correlation system for navigation for Mars Landing Program • The American Physical Society's Apker Award • Graduated from Massachusetts Institute of Technology in three years at age 19 (entered at 16) with perfect (5.0) grade point average and graduated from Harvard Medical School with MD at age 23 • Teaching award for 'Order of Magnitude Physics'	• Creative writing (7) • Creation of art or music (6) • Fulbright award (2) • Wrote proposal for a voting system for new South African Constitution • Solo violin debut (age 13) with Cincinnati Symphony Orchestra • Mellon Fellow in the Humanities • Presidential Scholar for Creative Writing • Hopwood writing award • Creative Anachronisms Award of Arms • First place in midreal-medieval poetry • Foreign language study fellowship • International predissertation award	• Phi Beta Kappa (71) • Tau Beta Pi (30) • Phi Kappa Phi (14) • Entrepreneurial enterprises (2) • Omicron Delta Kappa • Olympiad silver medal • Finished bachelor's and master's in four years • Received private pilot's license in one month at age 17
Total high-math	16	5	
Total high-flat	6	6	
Total high-verbal	7	13	

Interests

A general model of interest dimensions has emerged which is helpful for understanding how people approach and operate within learning and work environments. This model is a hexagonal structure known as RIASEC. It is defined by six general interest themes, with adjacent themes most correlated and opposite themes least correlated (see Figure 2). Briefly, the six themes are: 'realistic' (working with things and tools), 'investigative' (scientific pursuits), 'artistic' (aesthetic pursuits and self-expression), 'social' (contact with and helping people), 'enterprising' (buying, marketing, and selling), and 'conventional' (office practices and well-structured tasks).

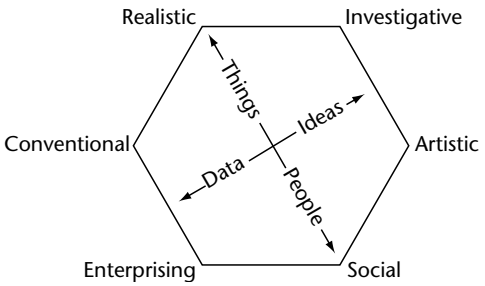


Figure 2. The RIASEC model of the structure of interest. Adapted from: Holland JL (1996) Exploring careers with a typology. *American Psychologist* **53**: 728–736.

While RIASEC has emerged repeatedly in large samples and across cultures, it is not embraced by everyone. It has been argued that the hexagonal model can be reduced to two relatively independent bipolar dimensions: people versus things, and data versus ideas. The former runs from 'social' (people) to 'realistic' (things) while the latter runs from the midpoint of 'enterprising' and 'conventional' (data) to the midpoint of 'artistic' and 'investigative' (ideas) (see Figure 2). Although there are certainly finer gradations of interest that carry psychological importance, RIASEC constitutes a useful framework for the study of interests. It also contains one of the largest sex differences discovered by psychological science on a continuous dimension, namely, people versus things. Females score much higher on the former, males on the latter. This difference has important implications for the kinds of work males and females prefer doing, other things being equal.

Research on educational and vocational interests comprises both temporal stability analyses (reliability) and forecasts of occupational group membership (validity). These investigations have established interest measures as among the most important in applied psychology. Research indicates that interests begin to crystallize during adolescence; they can forecast antecedents to occupational choice (e.g. college major) and, as such, serve as important tools in counseling and educational contexts. These are scientifically significant tools, which (like cognitive abilities) are predictive of a broad spectrum of criteria ranging from educational and vocational settings to activities in everyday life (hobbies and pastimes). Individual differences in interests are estimated to be approximately 50% heritable.

The most commonly used interest inventory is the 'Strong Vocational Interest Inventory' (SVII), which assesses an individual's interests in different occupational fields by asking them to indicate their preferences for various activities. Since people in a given occupation tend to share common interests that differentiate them from people in other occupations, the SVII compares an individual's unique combination of interests with average interest profiles of individuals in many different occupations. Occupational interests are classified according to the hexagonal structure of interest outlined above, and individuals are characterized by the relative strengths of these six occupational themes.

Personality

Although some consensus has emerged on the major personality dimensions, it is less clear-cut than for cognitive abilities and interests. Most studies of personality have followed the 'lexical' approach, which is based on the idea that important dimensions of human behavior are encoded in natural language for economy of thought. Hence, the dictionary, when systematically examined, proves an invaluable source for identifying personality characteristics. A working model of descriptors from the dictionary is available, commonly referred to as the 'big five'. Labels for each of the five factors have varied, but they include: 'extraversion' (joy, positive emotionality), 'agreeableness' (the opposite of antagonism), 'conscientiousness' (will to achieve), 'neuroticism' (anxiety, negative emotionality), and 'openness' (culture, or an intellectual orientation).

Other investigators argue that the big five need to be augmented by two additional dimensions: 'positive valence' and 'negative valence'. Positive valence is represented by terms such as 'outstanding', 'excellent', and 'remarkable', which form a continuum from ordinary to exceptional, or common to impressive. Negative valence is represented by terms such as 'cruel', 'evil', and 'sickening', which form a continuum from worthy to evil, or decent to awful.

Some of the best contemporary evidence for the scientific significance of broad dimensions of personality is found in predictions of vocational criteria. Several studies have shown that individual differences in personality dimensions, like those of vocational interests, are around 50% heritable.

Perhaps the most widely utilized personality inventory is the 'Minnesota Multiphasic Personality Inventory' (MMPI), which asks respondents to indicate which items are like or not like them. Although the MMPI has extensive clinical applications (in the identification of various forms of psychopathology), it also includes scales of several dimensions of personality, including social introversion, ego strength, and dominance. The 'California Psychological Inventory', which shares many of its items with the MMPI, eliminates those scales of the MMPI intended to identify psychopathology and concentrates on personality dimensions only. Another widely used personality scale is the NEO Personality Inventory, which yields assessments of each of the big five. Several other measures of individual differences in personality are available on a collaborative public website designed to help scientists to improve these inventories (see Further Reading).

CONSTELLATIONS OF ATTRIBUTES

Examining dispositional attributes individually can be challenging because the manner in which each operates depends on the full constellation of personal characteristics. Similar interest and ability patterns often produce markedly different behavioral patterns as a result of differences on dimensions from other classes. The paths traveled by two mathematically gifted students, for example, are likely to be very different if the two students occupy very different positions on the 'people versus things' interest dimension (other things being equal). Assuming that more comprehensive assessments will enhance psychological theory and practice, an approach that combines abilities, interests, and personality dimensions is generally considered to be the ideal.

One model of adult intellectual development, embraced by P. L. Ackerman, uses cognitive

abilities, personality, and interest dimensions simultaneously to describe developmental changes in cognitive content and processes throughout the lifespan. This is the 'process, personality, interests, and knowledge' theory of intellectual development. Interest and personality attributes channel the development of intelligence-as-knowledge down different paths, while intelligence-as-process determines the complexity and density of the knowledge assimilated. In other words, intelligence-as-process, through interactions with interests and personality, fosters adult intellect (intelligence-as-knowledge). Over time, personality and interests direct the application of intelligence-as-process towards the development of particular knowledge structures. This model provides a useful way of revealing why individuals with similar cognitive profiles can, and frequently do, differ widely in their knowledge base.

Analysis of the relationships among ability, interest, and personality attributes has revealed that these dispositions are not entirely independent. Four across-attribute (ability, interest, personality) trait complexes have emerged: social, clerical–conventional, scientific–mathematical, and intellectual–cultural. The social trait complex includes social and enterprising interests, and the personality traits of extraversion, social potency, and well-being, but does not include any ability traits. The clerical–conventional trait complex includes perceptual speed (an ability trait), conventional interest, and the personality traits of control, conscientiousness, and traditionalism. The scientific–mathematical trait complex includes two cognitive abilities, mathematical reasoning and visual perception, and realistic and investigative interests, but does not include any personality traits. The intellectual–cultural trait complex includes crystallized intelligence and ideation fluency (ability traits), artistic and investigative interests, and the personality traits of typical intellectual engagement and openness to experience. Note that these trait complexes are not mutually exclusive.

The constellations formed by the salient features of these personal attributes differentially tune people to opportunities in education and the world of work, thereby serving as critical components in determining lifestyle and influencing important life outcomes. A large part of what goes into understanding oneself, from a psychological point of view, involves coming to terms with one's abilities, interests, and personality, and using this information to make life choices and structure opportunities for personally relevant psychological growth throughout the lifespan.

STYLES

Psychological 'styles' refer to the temporal dynamics of behavior, rather than its content or structure. 'Capacity to work', 'industriousness', 'will', and 'zeal' are descriptors frequently found in the scientific literature on psychological 'tempo' or style. One of the most systematic accounts of style is found in Dawis and Lofquist's four dimensions of personality style: celerity, pace, rhythm, and endurance. Celerity refers to the speed at which one reacts to the environment. Pace is defined by the individual's level of activity. Rhythm refers to the stability of that activity level. Endurance represents the duration of an individual's interaction with the environment.

Although measures of these important psychological attributes are not readily available, there is little doubt that large individual differences exist in the temporal dynamics of behavioral patterns. These attributes are certainly important to consider when large differences are observed in performances and outcomes between two individuals with comparable ability, interest, and personality structures.

GENETIC AND ENVIRONMENTAL CONTRIBUTIONS TO INTELLIGENCE

There has been a great deal of discussion regarding the relative contributions of genetics and environment to intelligence. Historically, this debate has been referred to as the question of 'nature versus nurture', but more recent discussions have recognized the importance of both nature (genetics) and nurture (environment). Clearly, both contribute to the individual differences observed in intelligence. However, the question of their relative contributions remains a debated issue in psychology.

The genetic contribution to individual differences in intelligence refers to the effects of one's genes, which are inherited from one's parents. Half of an individual's genetic make-up is inherited from the individual's mother, and half is inherited from the individual's father. The specific combination of genes inherited from one's parents are unique to an individual, that is, no other person will have exactly the same combination of genes (except in the case of identical twins, which will be examined in further detail below). The degree to which individual differences in a trait are genetically influenced is represented by an estimate of heritability, the proportion of the observed individual differences on a given trait that are attributable to genetic differences among the individuals.

The environmental contribution to individual differences in intelligence is less clearly defined than the genetic contribution. Environmental contributions are broadly defined as all non-genetic influences. These influences include obvious components such as the individual's educational opportunities and family circumstances (e.g. parental education, family income, social standing), but they also include less obvious components of the environment such as pre- and post-natal nutrition, hormonal influences, and illness. Environmental contributions are further parsed into shared and non-shared components, which are, respectively, those elements of the environment that are common to compared individuals and those elements of the environment that are unique to compared individuals.

Genetic and environmental contributions to intelligence are certainly intertwined and may even interact, but there are research methods that allow the disentanglement of genetic and environmental sources of variation in intelligence, as well as other personological traits. The heritability of intelligence, which provides an estimate of the proportion of the observed individual differences in intelligence that are attributable to genetic differences among the individuals, may be gleaned by studying the degree of similarity on that trait among family members. These studies often include examinations of parent–offspring, sibling, identical and fraternal twin, and adoptive relationships. The premise underpinning these behavioral genetic research methodologies is that more closely (genetically) related family members should be more alike in intelligence than less closely (genetically) related family members, if indeed the trait is at least partially determined by genetic factors. For example, to the extent that individual differences in intelligence are genetically determined, monozygotic (identical) twins (who are genetically identical) should be more similar to each other than dizygotic (fraternal) twins (who share only half of their genes).

Quantitative estimates of heritability vary somewhat across studies, but typically the heritability of intelligence is estimated at between 0.4 and 0.8 (on a scale from 0 to 1). Monozygotic twins who have been reared apart (in different environments) provide a unique opportunity to directly measure the contribution of genetics to individual differences in intelligence; these studies indicate that the heritability of intelligence is about 0.70, meaning that genes account for approximately 70% of the individual differences observed in intelligence. Other familial relationships suggest similar, or slightly lower, estimates.

Adoptive relationships are also informative because one set of parents provides the genetic make-up of the child and a different set of parents provides the environment in which the child is raised, thereby allowing an examination of the unique contribution of the environment to the individual differences in intelligence. Studies of various adoptive relationships suggest that rearing environments account for no more than 30% of the individual differences in intelligence observed in children, and account for practically none of the variance in adults.

Recent behavioral genetic research has moved beyond the simplistic 'nature versus nurture' question towards an understanding of how nature and nurture interact in the development of intelligence. For example, it has been observed that estimates of heritability of intelligence tend to increase across the lifespan, from less than 40% in childhood to more than 60% in adulthood. This pattern is thought to be a reflection of the tendency for individuals to self-select environments congruent to their own proclivities as they grow older and become less constrained by familial demands and expectations. Conversely, the influence of rearing environments diminishes to practically zero in adulthood, as individuals rely increasingly upon themselves for their selection of environments.

However, the fact that intelligence is highly heritable does not mean that individual intelligence is not malleable through environmental intervention. The environmental (or non-genetic) contribution to the individual differences observed in intelligence is certainly important, but it is still unknown which aspects of the environment may be manipulated, and which interventions, if any, might increase an individual's intelligence.

Modern scientific advances are taking genetic research to the next step: molecular genetics. Scientists are now searching the genetic make-up of humans for genetic markers of intelligence. For example, a study of the genetic make-up of a group of precocious adolescents (the same group who were examined phenotypically in Figure 1 and Table 1) and their biological parents is currently under way. The DNA from these children, some of whom have differential intellectual strengths (e.g. exceptional verbal abilities relative to mathematical abilities), is being examined for genes relevant to general and specific intellectual abilities. Indeed, this work appears to suggest the identification of a number of genes related to general intellectual functioning. (*See* **Behavior, Genetic Influences on**)

EMOTIONAL INTELLIGENCE AND MULTIPLE INTELLIGENCES

Innovative measures and theories periodically surface in psychology which purport to explain psychological phenomena or to at least to put them in a clearer light. Emotional intelligence and multiple intelligences are two examples of such constructs.

Briefly, the theory of emotional intelligence involves many aspects of social intelligence. These aspects began to be discussed shortly after the first tests of general intelligence began to appear. Emotional intelligence is seen as an amalgam of interpersonal skills and social judgment. However, valid measures of this complex construct have yet to be developed.

The theory of multiple intelligences (MI) postulates seven distinct intelligences: musical, bodily–kinesthetic, logical–mathematical, spatial, linguistic, interpersonal, and intrapersonal. It asserts that these seven intelligences are mutually independent. Moreover, it assumes that every individual has the potential to excel in one of the seven areas, and proposes that society should recognize and value each in equal measure.

MI has attracted much attention, particularly within educational circles. Given its 'something for everyone' basis, it is certainly attractive. However, there is little scientific evidence to support the structure of intelligence it suggests, beyond the meager evidence that led to its initial formulation.

Both emotional intelligence and MI are intuitively appealing, especially given that virtually everyone agrees that there is much more to effective functioning than the dimensions outlined above. Nevertheless, as S. Messick has pointed out, very similar concepts have been proposed before, but when measures have actually been developed to assess them they have been found to be unable to improve scientific forecasts beyond the existing measures, and the existing measures are almost always superior predictors. The situation is similar with 'moral reasoning' measures, which were used for decades in hundreds of studies. When they were finally compared with verbal ability measures, it was found that they had little more to offer. The problem of attaching different names to scales that measure essentially the same attribute is so common in psychology that Truman Kelley has named it the 'jangle fallacy'.

This is not to say that we should not try to improve the measurement of individual differences in psychological science: there is still much room for improvement in the prediction process. However, innovative measures need to be evaluated against existing measures of ability, interest, and personality before it can be claimed that they capture something new, and, especially, before they can serve as a foundation for more comprehensive models of human functioning.

Further Reading

Ackerman PL (1996) A theory of adult intellectual development: process, personality, interests, and knowledge. *Intelligence* **22**: 227–257.

Carroll JB (1993) *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, UK: Cambridge University Press.

Dawis RV and Lofquist LH (1984) *A Psychological Theory of Work Adjustment*. Minneapolis, MN: University of Minnesota Press.

Day SX and Rounds J (1998) The universality of vocational interest structure among racial/ethnic minorities. *American Psychologist* **53**: 728–736.

Holland JL (1996) Exploring careers with a typology. *American Psychologist* **51**: 397–406.

Gottfredson LS (1997) Intelligence and social policy. *Intelligence* **24**: 1–320. [Special issue.]

International Personality Item Pool. http://ipip.ori.org. [A scientific 'collaboratory' for the development of advanced measures of personality traits and other individual differences.]

Jensen AR (1998) *The g Factor*. Westport, CT: Praeger.

Lubinski D (1996) Applied individual differences research: Its quantitative methods and its policy relevance. *Psychology, Public Policy, and Law* **2**: 187–392. [Special issue.]

Lubinski D (2000) Assessing individual differences in human behavior: 'Sinking shafts at a few critical points'. *Annual Review of Psychology* **51**: 405–444.

Lubinski D (2000) Intelligence: success and fitness. In: Goody J (ed.) *The Nature of Intelligence*, pp. 6–36. New York, NY: John Wiley.

Lubinski D and Benbow CP (1995) An opportunity for empiricism: review of Howard Gardner's *Multiple Intelligences: The Theory in Practice*. *Contemporary Psychology* **40**: 935–938.

Lubinski D and Benbow CP (2000) States of excellence. *American Psychologist* **55**: 137–150.

Lubinski D, Webb RM, Morelock MJ and Benbow CP (2001) Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology* **86**: 718–729.

Messick S (1992) Multiple intelligences or multilevel intelligence? Selective emphasis on distinctive properties of hierarchy: on Gardner's *Frames of Mind* and Sternberg's *Beyond IQ* in the context of theory and research on the structure of human abilities. *Psychological Inquiry* **3**: 365–384.

Plomin R (1999) Genetics and general cognitive ability. *Nature* **402**: C25–C29.

Rowe D (1994) *The Limits of Family Influence*. New York, NY: Guilford.
Waller NG (1999) Evaluating the structure of personality. In: Cloninger CR (ed.) *Personality and Psychopathology*, pp. 155–197. Washington, DC: American Psychiatric Press.

Inductive Reasoning, Psychology of

Introductory article

Steven Sloman, Brown University, Providence, Rhode Island, USA

CONTENTS

Inductive reasoning is the capacity to draw uncertain conclusions from known facts. Despite lacking a sound justification, people make inductions in systematic ways.

THE PROBLEM OF INDUCTION

You have just encountered your first Rhode Island native, and have discovered that this person knows a lot about shellfish. Should you infer that most Rhode Islanders are likely to know a lot about shellfish? This is a question of inductive reasoning, and it has no definitive answer. Your new knowledge about one particular Rhode Islander does not allow any certain conclusions about all Rhode Islanders, although it may provide more or less evidence about the general statement; it may increase your degree of belief in the conclusion even though it does not make the conclusion logically necessary.

We can express the question as a request for a judgment about the strength of an argument:

One Rhode Islander knows a lot about shellfish.
Therefore, most Rhode Islanders know a lot about shellfish.

where the first statement is a premise and the second a conclusion. Each statement can be broken down into a predicate (knows a lot about shellfish) and a category (Rhode Islanders). In this case, the premise category is particular, it has only one member. The problem of induction has traditionally concerned categorical arguments, like this one, in which a predicate is projected from a category to a more general superordinate one (in the extreme case, from a particular to a universal).

Most work in cognitive science has focused on a more general class of arguments, those that project a predicate from one category to another that may or may not be superordinate. This includes the specific to general case (e.g. dogs to mammals), as well as arguments in which premise and conclusion categories share a superordinate (e.g. dogs and horses to cats). The class of inductive arguments excludes those in which the conclusion or its negation is logically derived from the premise (e.g. dogs and cats have fleas, therefore dogs have fleas), causal arguments (e.g. the man ate greasy food, therefore the man had a heart attack), and abductive arguments in which inferences are made to the best explanation of a known fact (e.g. the DNA test shows her blood at the scene of the crime, therefore she's guilty of murder). Nevertheless, the class of inductive arguments is large enough that logic, causal reasoning, and explanation each enter into the determination of the strength of many inductive arguments.

The philosopher Nelson Goodman stated the problem of induction in terms of how to determine the projectibility of a predicate. Presumably you found the argument above about shellfish to be less than completely convincing. Nevertheless, it is stronger than:

One Rhode Islander is related to the Governor.
Therefore, most Rhode Islanders are related to the Governor.

In Goodman's terminology, the predicate 'knows a lot about shellfish' is more projectible than the predicate 'is related to the Governor'. The problem of induction is to develop an objective means of assessing the projectibility of a predicate, and thus the strength of an inductive argument. David Hume argued long ago that no such objective method is possible because no number of observations can justify a law and, in fact, no proposed inductive system has achieved a consensus.

This is not to say that there are no generally accepted constraints on how people judge argument strength. For example, an argument with a more specific conclusion cannot be weaker than an argument with a more general one. The premise 'All A are P' provides more support for the conclusion 'All B are P' than for the conclusion 'All B and

C are P', because if Bs and Cs are P, then Bs are P. Such constraints are captured by probability theory; in this case, by the conjunction rule, that – given A – the probability of B is not less than the probability of B and C. Probability theory has emerged as the most frequent formal model of inductive strength, but it is not an inductive logic because it does not provide a general means of determining the believability of a conclusion from knowledge of premises. Nevertheless, probability theory is a normative theory of induction in that argument strength judgments that violate the constraints of probability theory are usually considered fallacious. For example, the probabilities that X occurred and that X did not occur cannot both be high, so two arguments with identical premises but contradictory conclusions cannot both be strong.

Partial solutions to the problem of induction have been offered. For example, when sample spaces are clear and well-defined, statistical models can provide great leverage in determining the probability of a conclusion conditional on some premises. Knowing that 999 new, randomly chosen light bulbs work and one doesn't provides good reason to believe that the probability is 999/1000 that the next light bulb chosen will work.

INDUCTION PHENOMENA

The general problem of induction has not been solved and may never be. But whether or not the procedures that people use for making inductive inferences are always justified, people make them in very reasonable ways. We quickly learn such things as which mushrooms to make soup with. In fact, the evidence shows that even infants of less than a year of age make inferences. The following systematicities in inductive reasoning have been experimentally demonstrated. They reflect general tendencies in how people judge argument strength, tendencies that are sometimes overriden by extreme cases or by other factors.

Similarity

Arguments are strong to the extent that categories in the premises are similar to the conclusion category. For example,

Robins have sesamoid bones.
Therefore, sparrows have sesamoid bones.

is judged stronger than

Robins have sesamoid bones.
Therefore, ostriches have sesamoid bones.

because robins are more similar to sparrows than to ostriches.

Inclusion Similarity

Similarity relations can even override transparent class inclusion relations. If you believe that all lakes are bodies of water, then you should judge

Every individual body of water has a high number of seiches.
Every individual lake has a high number of seiches.

to be a perfectly strong argument because all members of the conclusion category must have the property if every individual member of its superordinate does. But people do not always judge this argument perfectly strong even when they agree that a lake is a body of water. In fact, people often judge

Every individual body of water has a high number of seiches.
Every individual reservoir has a high number of seiches.

to be an even weaker argument, presumably because reservoirs are less typical bodies of water than lakes.

Typicality

The more typical premise categories are of the conclusion category, the stronger is the argument. For example, people are more willing to project a predicate from robins to birds than from penguins to birds because robins are more typical birds than penguins.

Asymmetry

Switching premise and conclusion categories can lead to arguments of different strength:

Tigers have 38 *chromosomes.*
Therefore, buffaloes have 38 chromosomes.

is judged stronger than

Buffaloes have 38 *chromosomes.*
Buffaloes have 38 chromosomes.

either because tigers are more typical mammals than buffaloes or because tigers are more familiar than buffaloes.

Diversity

The less similar premises are to each other, the stronger the argument tends to be. People are

more willing to draw the conclusion that all mammals love onions after being told that hippos and hamsters love onions, than after being told that hippos and rhinos do, because hippos and rhinos are more similar to each other than hippos and hamsters.

Monotonicity

When premise categories are sufficiently similar, adding a premise will increase the strength of an argument. Telling people that elephants also love onions increases their willingness to believe that mammals do.

Nonmonotonicity

A counterexample to monotonicity occurs when a premise with a category dissimilar to all other categories is introduced:

Crows have strong sternums.
Peacocks have strong sternums.
Rabbits have strong sternums.
Therefore, birds have strong sternums.

is weaker than

Crows have strong sternums.
Peacocks have strong sternums.
Therefore, birds have strong sternums.

Variability/Centrality

All the phenomena mentioned thus far reflect how people treat the relations amongst the categories of an argument. But the nature of the predicate matters too. People are more willing to project predicates that tend to be invariant across category instances than variable predicates. For example, people who are told that one Pacific island native is overweight tend to think it is unlikely that all natives of the island are overweight because weight tends to vary between people. In contrast, if told the native has dark skin, they are more likely to generalize to all natives because skin color tends to be more uniform within a race. Having dark skin may be seen as less variable by virtue of being more central, having more apparent causal links to other features of people.

Relevance

People's willingness to project a predicate from one category to another depends on what else the two categories have in common. For example, people are more likely to project 'has a liver with two chambers' from chickens to hawks than from tigers to hawks, but more likely to project 'prefers to feed at night' from tigers to hawks than from chickens to hawks. More generally, argument strength depends on how people explain why the category has the predicate. If the premise and conclusion of an argument are explained in the same way, the argument will be judged stronger than if they are explained differently.

Preferred Level of Induction

Both adults and children find arguments stronger the more specific the categories involved. If told that dalmations have an ulnar artery, people are more willing to generalize ulnar arteries to dogs than to animals. The judged strength of arguments drops off sharply when the premise and conclusion categories are not members of a common basic-level or more specific category.

Human Bias

Small children prefer to project a property from people than from other animals. Four-year-olds are more likely to agree that a bug has a spleen if told that a person does than if told that a bee does. Adults show the opposite preference.

Naming Effect

Children prefer to project predicates between objects that look similar than between objects that look dissimilar. However, this preference is overriden when the dissimilar objects are given similar labels.

Gap Effect

Inductions are stronger from categories that one would not expect to support the predicate than from categories whose support is less surprising. Learning that baby giraffes can survive for several days without water makes it seem likely that adult giraffes can. But learning the fact about adult giraffes doesn't support the inference about baby giraffes as strongly.

MODELS OF INDUCTION

Many of these phenomena are clearly sensible. For example, in favor of the diversity phenomenon, philosophers of science have argued that diverse evidence supporting a hypothesis should be more convincing than similar evidence which is

more redundant. To the degree that people's inductions have a rational justification, they can be described using probability theory, constrained as it is by possible relations amongst sets. But some phenomena, such as inclusion similarity, are not rational, implying that human induction is not entirely consistent with probability theory. Analogously, a theory of induction could appeal to general principles of good scientific practice, but such a theory is only as good as scientists' inductions are justified.

One psychological theory of induction assumes that people make categorical inductions on the basis of two principles, similarity and category coverage. Arguments are deemed strong to the degree that premise and conclusion categories are similar and to the degree that premises 'cover' the lowest-level category that includes both premise and conclusion categories. The notion of category coverage is intended to capture phenomena such as nonmonotonicity, by positing that adding a premise category unrelated to other categories makes a higher-level category relevant, thus reducing coverage. For example, by adding a premise about rabbits to ones about crows and peacocks, the superordinate 'animals' becomes relevant even if birds had previously been the lowest-level relevant category. And 'crows and peacocks' covers the category of birds better than 'crows, peacocks, and rabbits' covers the more general category of animals.

Another theory of induction reduces the two principles of similarity and category coverage into a single principle of feature coverage. Instead of appealing to class inclusion hierarchies, this theory appeals to relations amongst the properties of categories. Predicates are projected from premise categories to a conclusion category to the degree that the conclusion category's properties are covered by the premise categories. Feature coverage can explain the inclusion similarity phenomenon. Lakes don't inherit with certainty the properties of bodies of water because lakes' properties are not covered by the properties of bodies of water. Reservoirs show even less inheritance because they have even more distinctive features. The feature coverage view is not as comfortable with nonmonotonicities, because adding a premise category should not decrease feature coverage and therefore argument strength.

Neither category nor feature coverage can explain phenomena that depend on the nature of the predicate, because these theories focus exclusively on the relations amongst categories. Some models of induction focus on the features of a category that are made relevant by a predicate. If told 'Rhode Island has a beautiful autumn', you are likely to project the predicate only to regions close to Rhode Island, not to those that are shaped like it. Determining relevant features requires explaining why the category would support the predicate, and this explanation therefore determines the induction.

CONCLUSION

People use a variety of cognitive frameworks to evaluate inductive argument strength. Sometimes people use logic. For example, if a premise category is observed to include the conclusion category, the conclusion is usually correctly considered certain. Often people reason causally, as when they generate explanations and let those explanations mediate their judgment. The notion of similarity is at the heart of much inductive reasoning; people rely more on their sense of similarity than on anything else to make inductive judgments. But similarity provides only limited explanatory value if a judgment of similarity changes with the prevailing context.

People are very good at generating explanations, and these explanations quickly become part of the cognitive schema we use to understand our experience. The relative success of simple computational models in accounting for the phenomena of inductive argument strength suggests that we are able to make very good inductive guesses using these sophisticated schema in fairly coarse ways. The evidence that our usage is coarse is the systematic errors we make. We often ignore logical or categorical structure in favor of similarity. But when that more fine-grained structure is attended to, people can use it.

Further Reading

Goodman N (1955) *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Heit E (2000) Properties of inductive reasoning. *Psychonomic Bulletin and Review* **7**: 569–592.

Nisbett RE (ed.) (1993) *Rules for Reasoning*. Hillsdale, NJ: Lawrence Erlbaum.

Osherson DN, Smith EE, Wilkie O, Lopez A and Shafir E (1990) Category-based induction. *Psychological Review* **97**: 185–200.

Sloman SA (1993) Feature based induction. *Cognitive Psychology* **25**: 231–280.

Sloman SA (1998) Categorical inference is not a tree: the myth of inheritance hierarchies. *Cognitive Psychology* **35**: 1–33.

Infant Cognition

Introductory article

Amanda L Woodward, University of Chicago, Chicago, Illinois, USA

CONTENTS

Cognition comprises the mind's processes for acquiring, interpreting, and storing information, as well as processes for reasoning and thinking abstractly. The period of infancy in humans is generally agreed to be the first two years of life, a time during which children are largely preverbal.

INTRODUCTION

Since ancient times, those interested in the nature of the human mind have turned to infants when asking fundamental questions. What is essential and immutable in the human mind? What is given by the particulars of experience? The motivation for seeking answers to these questions from infants is that they have little in terms of experience and (presumably) all of what is innately given. It was not until the twentieth century that the scientific study of infant cognition began, with most of the foundational research occurring in the final three decades. This research has shown that the traditional nature/nurture dichotomy is misleading. Development is a process of organism–environment interaction from the beginning, even before birth. It is extremely difficult to isolate the individual contributions of nature and nurture because they do not act in isolation. Nevertheless, it is possible to ask which human capacities emerge early and therefore may serve as foundations for subsequent development. Current studies of infant cognition focus on the nature of infants' abilities at particular ages, and the process of development during the first years of life.

PERCEPTION

Perceptual systems are the mind's point of contact with the world. Human sensory organs take in a limited range of information. The problem of perception is to take this limited sensory data and yield the experience of a coherent, multidimensional environment. William James expressed a long and widely held view when he wrote that the infant's world was a 'blooming, buzzing confusion'. Early theorists assumed that infants lacked the abilities necessary to construct a coherent world from sensory data. Current work contradicts this view. Perceptual systems function from before birth, and constraints present at birth provide even newborns with a basis for making sense of sensory data. During the first year, experience and maturation tune and enrich the infant's perceptual capabilities. Vision and audition are the most well-studied types (or modalities) of perception in humans of any age, and they are the focus here. Smell, taste, touch, and proprioception also function from the beginning of life.

Vision

How can we know what infants see? Beginning in the 1950s, researchers began to devise methodological innovations to probe the perceptual processes and experiences of infants. Behavioral measures, such as patterns of visual attention, as well as brain activity are currently used to investigate infant perception. To illustrate, one measure of acuity exploits infants' preference to look at a pattern rather than a blank field. Researchers present infants with two pictures, one a gray field and the other a grating of black and white lines. If infants look longer at the latter, this is taken to indicate that they can perceive the grating pattern. By varying the size of the grating, researchers can test the limits of the infant's acuity. Studies of this sort have revealed that newborns can see, but their acuity seems very limited. Acuity may approach adult levels by 8 months of age.

Beyond the capacity to take in visual information, infants possess systems for interpreting this information – that is, perceiving a three-dimensional world populated by distinct objects. The data available to the eyes are two-dimensional, yet we perceive a three-dimensional world. Adults draw on many cues to depth relations, including the relative

motions of objects with respect to one another and the observer's own motion, the different viewpoints provided by the two eyes (stereopsis), the physical activity required to focus on objects, and pictorial cues such as shading, overlap, and perspective. When and how do infants become sensitive to this information as specifying depth?

Early theorists proposed that experience reaching for and manipulating objects at different distances was essential to the ability to interpret visual cues to depth. Modern experimental work has challenged this hypothesis. From birth, well before they are able to reach for objects, infants seem to perceive in three dimensions. A striking demonstration of this is that newborns track an object's size as constant as it moves nearer and farther from them. Because the size of an object's projection on the retina varies as a function of its distance from the observer, perceiving depth is important for tracking size constancy. In one experiment, newborns were familiarized to a cube at various distances. Then they were shown the original cube at a new distance and a similar second cube of a different size. The cubes were positioned so that both projected the same sized image on the retina. Infants stared longer at the novel sized cube, suggesting that they could distinguish between the two sizes and that they recognized the original cube as familiar, despite its new position. Other studies have suggested that infants are sensitive to motion-based depth cues from the first month of life. Infants are not sensitive to pictorial depth cues until late in the first year, suggesting that experience may contribute to their development. Stereopsis is not functional at birth, but matures at between 2 and 4 months of age.

A second challenge in visual perception is to determine which pieces of sensory data comprise a single object. Objects are often partially occluded by others (as when a mother's hand holds a bottle), and boundaries between contiguous objects (books on a shelf, or pencils in a jar) may not be obvious. Adults succeed in perceiving discrete objects by drawing on many cues, including patterns of motion (parts that comprise an object move together), and regularities in the shape and texture of objects. Young infants rely heavily on motion as a cue to object unity. By 2 months of age, infants use common motion as a basis for treating two segments emerging from opposite side of an occluder as part of the same object. The ability to use featural information (shape, color, texture) as a cue to object unity emerges at around 4 months, and may be related to developments in the infant's ability to manually explore objects.

Audition

Fetuses respond to sounds prenatally, and remember what they have heard after birth. Prenatal listening may be the source of newborns' preferences for their mothers' voice and for their native language. Speech is central to normal human cognition and development, and therefore much of the research on infant audition has focused on speech sounds. The minimal difference in sound that carries a difference in meaning is called a *phoneme*. For example, in most spoken English *bit* and *pit* differ by a single phoneme. The physical difference between any two phonemes can be described using a continuous metric; for example, it is possible to artificially produce a sound that is halfway between a *bih* and a *dih*. However, people do not perceive speech sounds as varying continuously. Human perception of speech sounds is categorical; that is, physically different sounds that fall within a category are not easily distinguished, sounds that fall in two different categories are readily distinguished, and no 'middle ground' is perceived. Infants seem to perceive speech sounds categorically from birth. Although this might suggest an innate specialization for language, in fact other species, including quail and dogs, also have categorical perception of speech sounds. Language may have evolved to exploit a pre-existing property of the auditory system. Language experience tunes infants' innate categorical perception abilities. Languages differ in the particular sound contrasts they use as phonemic. For example, *r-* and *l-* are different phonemes in English but not in Japanese. Adults find it difficult to perceive non-native phonemic contrasts. Young infants, however, perceive non-native contrasts as well as native ones. By the end of the first year, infants become insensitive to non-native contrasts. Beyond perceiving the elements of language, by 8 months infants are able to learn recurring patterns in language sounds, an ability that probably contributes to their extraction of word units and grammatical patterns.

Speech carries many different kinds of information simultaneously, and infants are sensitive to much of this information. Beyond the level of phonemes, speech has prosodic characteristics including rate, loudness, rhythm, pitch contours, and pauses. The speech that is specifically directed at infants exaggerates these prosodic elements. Infants prefer infant-directed speech and are more sensitive to the exaggerated prosodic cues it provides. By 6 months, infants seem to respond to prosodic features which correlate with clause

boundaries. During the second half of the first year, infants seem to become sensitive to the particular prosodic cues to grammatical boundaries in their native language. Prosody also conveys emotion – revealing, for example, whether the speaker is excited or frightened, or wishes to praise or scold. Infants respond to these emotional messages appropriately, even when they are given in an unfamiliar language.

Perceptual integration

Even at the beginning of life, different perceptual modalities speak to one another, yielding the experience of a coherent world in which sights, sounds, and tactile experiences are integrated. Newborn infants turn in the direction of a sound, a response that can provide a basis for connecting sounds with the visual properties of the objects that produce them. Studies have suggested that young infants spontaneously match information across modalities. When infants see two images and hear a soundtrack, they will look at the image that matches the soundtrack. Young infants can match novel images and sounds based on temporal synchrony, as well as matching familiar stimuli based on knowledge about the noises emitted by different kinds of objects (e.g., colliding sponges versus colliding blocks) and correspondences between speech sounds and the facial configurations that accompany them. Infants seem able to match tactile and visual information. For example, in one experiment, 1-month-old infants who were given either a bumpy or a smooth pacifier to suck looked longer at a picture that matched the pacifier in their mouth.

ATTENTION

For adults, perception is an active process. People actively seek out certain kinds of information, directing their attention to do so. From birth, infants strategically deploy attention in ways that are likely to facilitate information extraction and learning. Infants, like many other organisms, habituate to a repeated stimulus, attending less to it over repeated presentations, and dishabituate, or increase their attention, when a novel stimulus is presented. These responses aid the efficient use of attention, by reducing attentional resources for old information, and increasing them for new information. In addition, studies have suggested that infants seek out information-rich stimuli. They scan for edges and contours, prefer moderately complex patterns to less complex ones, and, as just reviewed, seek to integrate information across modalities. These attentional responses provide researchers with tools for probing other cognitive processes in infants.

For adults, it is important not only to allocate attention effectively to the current environment, but also to anticipate future events and direct attention accordingly. Some researchers have found that, as young as 3 months, infants direct their attention in anticipation of an event. For example, an infant might see pictures appear in alternation on the right and left side of a screen. After exposure to this pattern, infants look to where the next picture will appear before it actually appears. Infants show anticipatory looking for more complicated patterns as well, with performance on these improving during the first year.

The ability to coordinate one's own attention with the attention of a social partner (joint attention) is a critical contributor to cognitive development. Infants depend on social partners not only for survival, but also for instruction in key domains including language and culture. Systematic attention to social partners starts early. Newborns prefer pictures of faces to other patterns, and research suggests they form preferences for familiar voices and faces. Young infants attend preferentially to eyes over other facial features, respond to shifts in gaze direction, and direct their own attention in accord with another at person's gaze. Gaze-following becomes more robust at between 6 and 18 months of age. By 12–18 months, infants use an adult's gaze direction to interpret the reference of the adult's utterances and emotional expressions.

MEMORY

Development depends on the ability to retain information in memory. Many of the findings discussed so far imply memorial abilities in infants. For example, in order for infants to respond differently to familiar and novel stimuli they must have formed memories of the familiar stimulus. Moreover, since infants seem to prefer sounds heard prenatally, memory may begin before birth. Infant memory has been assessed by many other procedures including retention of conditioned responses, search for hidden objects after delays, response to novel and familiar items after delays, and deferred imitation. Infants remember for long periods as well as short ones. Infants in the first few months of life can retain memories for days or even weeks, and studies with older infants indicate retention of new information for many months.

A central debate concerns the nature of infant memories. Do they include declarative representations, or are they limited to procedural representations? The latter claim concerning infants derives, in part, from Piaget's theory that infant intelligence is limited to sensory-motor abilities (that is, organized ways of acting), and does not include abstract, conceptual representations (but see the section on conceptual structure, below). Much of the evidence for memory in young infants seems to be procedural in nature. For example, 3-month-old infants can learn to kick their feet in order to activate a mobile, and retain this response in memory for long periods. While this act of learning clearly involves procedural memory, it is not clear whether infants also form a more explicit memory of the task. The clearest evidence for declarative memory would be verbal report, which is not available from infants. Researchers have used deferred imitation as a measure of declarative memory in infants. Infants are shown a novel toy, on which an experimenter models a novel action. Infants are not allowed to touch the object until they return to the laboratory after a delay. At that point, if they are able to reproduce the modeled actions, this indicates that they could be recalling the prior event, and representing information that is not strictly procedural. Infants as young as 6 months succeed at producing deferred imitations when the task is not very demanding.

CONCEPTUAL STRUCTURE

In the minds of adults, knowledge is organized around a set of basic concepts, which comprise the building blocks of everyday reasoning. We form mental models of objects in the world, their physical properties, and their causal interactions. We distinguish between inanimate objects and intentional agents. We extract and manipulate information about number. Underlying these systems of knowledge is a general ability to form conceptual categories. How is infants' knowledge organized? Does it reflect the basic concepts that organize adult knowledge? Working in the early part of the twentieth century, Jean Piaget was among the first scientists to frame and address these questions, which continue to focus much of the current research on infant cognition.

Recent experiments have yielded evidence for structured knowledge much earlier in infancy than Piaget's observations indicated. There is currently debate concerning the implications of these findings for the nature of the infant mind. Some theorists have taken these findings as evidence for innate concepts or processing modules dedicated to particular core notions in human cognition. An alternative proposal is that innate constraints on information acquisition enable rapid learning in key domains. Others have argued that specific innate concepts are both neurologically implausible and unnecessary to account for infants' abilities. In considering these alternatives, it is relevant to keep in mind that recent experiments also indicate that infants' knowledge representations differ in significant ways from those of adults. Adult concepts are multifaceted because they are embedded in webs of knowledge, or *folk theories*. Infants' knowledge representations may contain parts of or precursors to mature knowledge. The challenge is to specify the nature of these partial representations.

Object permanence

A basis for much of cognition is the ability to form mental models of objects which can be held in mind and used when the object is no longer available to the senses. Mental models are most useful if they represent the physical properties of objects, such as their solidity, location in space, and continuous existence over time, as well as their particular features. Piaget argued that the ability to form such mental models (which he termed *object permanence*) was not achieved until the end of infancy, at around 18 months, based on his robust observations that young infants fail to search for an object that has been hidden from view and that difficulties on search tasks persist well into the second year of life.

However, search tasks impose demands that may mask infants' representational abilities. In particular, they require means–end abilities, which have been linked to prefrontal cortex, an area of the brain that undergoes critical development toward the end of the first year. More sensitive experimental techniques have revealed that young infants form mental representations of objects and their positions in space. To illustrate, in one experiment, 3-month-old infants were shown a screen rotating through 180 degrees until their attention to it had declined (that is, until they had habituated to it). Then a block was placed on the far side of the screen and infants saw one of two events. In one, the screen moved 120 degrees and then stopped. This was a novel stopping position for the screen, but one that was physically possible given the presence of the now hidden block. In the other, the screen continued to move through 180 degrees. Because this rotation was familiar from the habituation event, if infants did not

remember the object hidden behind the screen, it should have been uninteresting to them. However, if infants represented the hidden block, this event should have been surprising because the screen seemed to move through the space occupied by the block. Infants looked longer on the apparently impossible trials than on the possible trials, suggesting that they represented the existence of the block behind the screen. Further evidence for infants' representational abilities comes from studies of reaching: infants reach for an object in the dark, so long as they have been shown the object before the lights go out. Moreover in these studies, infants reach differently depending on the size of the object, indicating that they hold in mind the object's size as well as its continued existence and location.

Physics

During the first few months of life infants are sensitive to two principles that are deeply intuitive to adults – the principles of solidity (an object cannot pass through the space occupied by another object), and continuity (objects exist continuously in time and space). In the experiment described above, for example, infants not only tracked the block's continuous existence behind the screen, but also responded as if both the block and screen were solid. The principle of continuity also provides information about the number of individual objects present in a scene. By 4 months of age, infants seem to be sensitive to this aspect of continuity. In addition, young infants seem to be sensitive to the causal properties of events. When one object moves toward another, contacts it, and then the contacted object immediately moves off, adults perceive the first object as causing the second to move. Infants differentiate between these apparently causal events and similar events that do not appear causal to adults because of a gap between the objects or a delay before the second object is 'launched'.

Infants' representations of physical reality differ significantly from those of adults. Infants seem to be less sensitive to the constraints imposed on object motion by gravity and inertia than they are to the constraints imposed by solidity and continuity. Across a range of physical phenomena, infants seem to begin with an initial all-or-none representation, only later changing their perception to reflect the relative properties of objects. Although young infants use spatiotemporal information to determine the number of individual objects that are present in a scene, they are less able than older infants to use featural information to do this.

Infants also differ strikingly from older children and adults in that their ability to express their physical knowledge is extremely limited. Visual habituation procedures reveal abilities in very young infants, but other procedures reveal apparent deficits in physical reasoning in much older children. This discontinuity has led to a number of hypotheses concerning the nature of infant cognition and its development. It has been suggested that infants' mental representations are inaccessible to general reasoning systems or action systems because they are encapsulated in particular processing modules, are implicit, or are weak. Current work is directed at elucidating and distinguishing between these alternatives.

Intentionality

The abilities to distinguish between inanimate objects and intentional agents, and to interpret the actions of the latter, are critical to human functioning. Adult folk psychology explains human action in terms of the actor's underlying psychological states. Recent experiments have explored the infant precursors to this system of knowledge.

Infants possess several propensities that may facilitate learning about human action. Newborns attend selectively to faces and voices, and are highly responsive to the contingent response patterns that are typical of social partners. They also spontaneously imitate some of the behaviors they observe, including facial gestures and arm movements. By 6 months, infants respond to shifts in other people's gaze direction by directing their own attention in the same direction. This sets the stage for later developments in joint attention.

Beyond these sensitivities, do infants analyze actions in terms of desires, perceptions, and intentions? A number of studies indicate that 18-month-olds interpret the behavior of other people in terms of their intentions and attentional states. To illustrate, on observing a person attempt and fail to complete a novel action, 18-month-olds infer the person's goals and reproduce the intended action. Recent findings suggest that 6- to 9-month-olds understand certain actions as being goal-directed. It has been hypothesized that imitation provides a basis for infants' interpreting other's actions in terms of underlying psychological states by allowing infants to connect their own internal experiences and actions with the actions of other people. This possibility is consistent with recent findings indicating common neural substrates for the perception and production of actions.

Number

Adults extract and mentally manipulate information about number. Observing flowers in a vase, we can represent the fact that there are precisely six flowers. Moreover, we can predict the resulting number of flowers for transformations such as removing two flowers or adding four.

Habituation experiments have revealed infant abilities that may be related to mature number knowledge. First, young infants are sensitive to features associated with exact small (1–3) numbers. Infants habituated to one number of items (e.g. three dots) dishabituate to a new number (e.g. two dots), but not to a novel display with the same number of items. This sensitivity extends to sounds and events as well as visual stimuli. In addition, when 5-month-old infants observe objects added to or subtracted from a hidden display, they look longer when an incorrect number of items is revealed than when a correct number is revealed. Moreover, recent studies suggest that infants form approximate representations for large numbers that allow them to distinguish between sets with large proportionate differences (8 versus 16), but not small proportionate differences (8 versus 12).

Whether these findings reflect sensitivity to number *per se* has been debated. Number is often correlated with other perceptual dimensions, including density and overall amount, and studies with infants do not always successfully control for these features. Moreover, some of the findings concerning small sets could be accounted for by infants' establishing representations of each individual object involved, rather than extracting the number of items in the set.

Categorization

Adults group together perceptually distinct individuals as being the same kind of thing. The resulting conceptual categories provide a means for efficient information storage and retrieval as well as a basis for inferring the properties of new exemplars.

From early in life, infants seem to be sensitive to category structure. Older infants spontaneously sort objects into categories. Young infants manifest a sensitivity to category structure in their patterns of attention. For example, in one experiment, having been shown a series of pictures of cats, 3-month-old infants generalized habituation to new cats and dishabituated to dogs. In addition, like adults, infants seem to structure their categories around the prototype, or most central member, of the category. Having been familiarized to members of a category, infants respond to the category's prototype as if it is highly familiar, even if they have not seen that item before.

Infants categorize artificial stimuli (e.g. dot patterns) as well as real ones (e.g. photographs of animals), and they attend not only to individual features but also to correlations between features in doing so. During the first year of life, infants can group items into relatively broad categories, such as *vehicle* or *animal*, as well as more narrow, basic level categories, such as *dog* or *car*. Different kinds of experiments tap different levels of category sensitivity in infants. In visual attention tasks, infants are sensitive to basic level categories, but on tasks that involve manual manipulation infants sometimes perform better with broad categories than with basic level categories. Infants, like adults, categorize more efficiently when given the opportunity to compare members of a category to one another or to members of a contrasting category.

Adult category knowledge includes not only the perceptual attributes shared by members of a category but also the underlying, often unobservable properties that unite a class. We understand that dogs not only tend to look, smell, and sound alike, but also that they share deeper, important properties including species-typical behaviors and internal structure. A question of long-standing debate is whether infants' categories are exclusively based on perceptual features or instead include more abstract knowledge. Investigating this question has proven difficult, because deep properties are strongly correlated with perceptual attributes, and therefore infants' propensity to group items may not provide unambiguous evidence concerning their knowledge about deep properties *per se*. Nevertheless, as young as 7 months, infants group together items from broad categories which differ from one another on many perceptual dimensions (e.g. a dog, a bird, and a rabbit), and they distinguish between members of different categories which are similar on several dimensions (e.g. a bird and an airplane), suggesting that they base category judgments on more than raw perceptual similarity. Older infants actively seek commonalities beyond the level of surface similarity. By the end of the first year, infants apparently infer that members of a kind share properties that are not immediately observable. By 13 months, infants may begin to understand the link between conceptual categories and language: hearing diverse members of a kind given the same name leads infants to seek out commonalities between them. Finally, by 18 months, infants selectively learn

feature correlations that are meaningful (e. g. having wheels and rolling versus having wheels and squeaking), suggesting that they seek the underlying causal explanations for these correlations.

Further Reading

Baillargeon R (1995) A model of physical reasoning in infancy. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, vol. 9, pp. 305–371. Norwood, NJ: Ablex.

Bauer P (1996) What do infants recall of their lives? *American Psychologist* **51**(1): 29–41.

Gopnik A, Meltzoff AN and Kuhl P (1999) *The Scientist in the Crib: Minds, Brains and how Children Learn*. New York, NY: William Morrow.

Haith MM and Benson J (1998) Infant cognition. In: Damon W, Kuhn D and Siegler RS (eds) *Handbook of Child Psychology*, vol. 2: *Cognition, Perception and Language*, pp. 199–254. New York, NY: John Wiley.

Kellman P and Arterberry ME (1998) *The Cradle of Knowledge: Development of Perception during Infancy*. Cambridge, MA: MIT Press.

Kuhl PK (2000) A new view of language acquisition. *Annals of the National Academy of Sciences* **97**: 11850–11857.

Mandler JM (1998) Representation. In: Damon W, Kuhn D and Siegler RS (eds) *Handbook of Child Psychology*, vol. 2: *Cognition, Perception and Language*, pp. 255–308. New York, NY: John Wiley.

Mehler J and Dupoux E (1994) *What Infants Know*. Cambridge, MA: Blackwell.

Piaget J (1954) *The Construction of Reality in the Child*. New York, NY: Basic Books.

Quinn PC and Eimas PD (1996) Perceptual organization and categorization in young infants. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, vol. 10, pp. 1–36. Norwood, NJ: Ablex.

Rovee-Collier C, Hayne H, and Colombo M (2000) *The Development of Implicit and Explicit Memory*. Amsterdam, Netherlands: John Benjamins Publishing.

Spelke ES (1991) Physical knowledge in infancy: reflections on Piaget's theory. In: Carey S and Gelman R (eds) *The Epigenesis of Mind: Essays on Biology and Cognition*, pp. 133–170. Hillsdale, NJ: Lawrence Erlbaum.

Spelke ES and Newport EL (1998) Nativism, empiricism, and the development of knowledge. In: Damon W and Learner RM (eds) *Handbook of Child Psychology*, vol. 1: *Theoretical Models of Human Development*, pp. 275–340. New York, NY: John Wiley.

Information Processing

Intermediate article

James T Townsend, Indiana University, Bloomington, Indiana, USA
Kan Torii, Indiana University, Bloomington, Indiana, USA

CONTENTS

The information-processing approach in cognitive science assumes mental architectures with various components that interact in order to manipulate incoming information.

DEFINITION AND HISTORY

Any field of study, in this case cognitive science, always has as a backdrop a kind of super-theory or metatheory, which may itself be rather vague but which provides a philosophy or way of thinking about problems and helps to motivate the theoretical and experimental investigations that are carried out. The information-processing approach is a good candidate for the central metatheory of cognitive science. Cognitive science began to emerge, slowly at first, in the 1950s, and developed rapidly in the 1960s and 1970s, becoming the dominant paradigm in psychology and attracting researchers in linguistics, computer science and philosophy.

A popular introductory text on cognitive science (Ashcraft, 2002) defines the information-processing approach as 'the coordinated operation of active mental processes within a multicomponent memory system'. The essential element of this definition is the specification of 'active mental processes' indicating dynamic ongoing activity. The term 'processes' emphasizes this dynamic aspect, and its early and continued use led to the term 'process model', again suggesting a model or theory that tells what happens cognitively across time. A 'model' is a theory constructed to explain and make predictions about a relatively restricted set of phenomena, while a 'theory' may aim to cover a much wider range. However, the terms 'model' and 'theory' are often used interchangeably. Much of psychology has traditionally been oriented more towards static descriptions (e.g., most personality trait models are static in nature), so the idea of being composed of active processes was a major innovation in the study of human thought and behavior. The words 'coordinated operation' also suggest that the model builder is paying close attention to how the operations of the various components involved in a mental task might start, continue, and stop relative to each other. A notable omission in the above definition of information processing is the arrangement, or architecture, of the processes. Architecture has been the subject of many studies.

In the 1950s, a number of well-known experimental (and early cognitive) psychologists employed information theory (e.g., Shannon and Weaver, 1949) as a theoretical and methodological tool. Claude Shannon, an engineer, invented information theory to measure how much information is contained in messages and how much is transmitted from one place to another. He went on to define the notion of 'channel capacity', the average amount of information transmitted through a communication channel as the time allowed for transmission becomes very large and when an optimal code for messages is constructed. All these notions have entered the domain of perceptual and cognitive psychology, and they strongly influenced the development of the information-processing approach (e.g., Attneave, 1959; Garner, 1962; Luce, 1960). Nevertheless, Shannon's theory of information processing is rarely employed in cognitive science (perhaps unfortunately); and it should not be confused with the more vague, but more globally applicable, information-processing approach.

Other lines of research by engineers, economists, computer scientists, and mathematicians also influenced the early development of cognitive science, and especially the information-processing approach. Cybernetics, the science of feedback control systems and adaptive systems (systems that can evolve their own behaviors and, to some extent, intelligence (Wiener, 1948)), was one of these. Another was the explosion of research in theoretical computer science, particularly with the work of John von Neumann. The fact that computers – in theory, and to a growing extent in practice – could carry out the mental operations about which philosophers and psychologists had long argued, convinced most psychologists and early researchers in artificial intelligence (computer scientists, engineers, biologists, and psychologists who study how to make machines that can think and act like humans) that important progress could be made in the study of human perception, thought, and action simply by basing their models on computer-like concepts, information theory, and similar structures and ideas.

SERIAL AND PARALLEL PROCESSING, FLOW DIAGRAMS, AND CONTINUOUS-FLOW SYSTEMS

Many early computers operated in a sequential, or 'serial', way on the tasks to be accomplished. This means that each task or part of a task is operated on and completed before the next is started. Hence, it was natural for cognitive investigators to hypothesize that certain simple mental activities might be carried out in this way.

Such an arrangement is called an 'architecture'. An obvious alternative architecture is a 'parallel' architecture, whereby all tasks or processes are carried out simultaneously. Parallel and serial systems are probably the simplest nontrivial kinds of architectures, but much more complicated architectures, many of them composed of combinations of serial and parallel processes, can be designed and tested against laboratory data (e.g., Schweickert, 1978; Schweickert and Townsend, 1989).

Many of the early models proposed in the information-processing approach were serial in nature. This trend was probably due to the serial character of many computer operations. Such analogies have often been made in psychological theorizing over the decades (e.g., Roediger, 1980).

The most popular device employed in the information-processing approach is that of the flow diagram. This is a schematic picture that shows as boxes the hypothetical processes thought to be involved in the mental behavior under study. Figure 1(a) shows a typical flow diagram for memory scanning. It represents a serial interpretation of short-term memory search along the lines suggested by Sternberg (1966). Figure 1(b) shows a parallel network supported by facial-feature detection (Wenger and Townsend, 2001). Figure 1(c) shows a hybrid network called a 'bare Wheatstone bridge' (Schweickert, 1978). Note that there exist parallel channels, but that the presence of the 'bridge' allows serial-like activity as well. Schweickert has found that the Wheatstone-bridge structures are required to handle certain kinds of mental processing.

While many cognitive studies have associated quantitative models with the architecture indicated by their flow diagrams, many more have simply relied on intuitive reasoning allied with experimental results to support their conclusions. Needless to say, verbal reasoning in scientific contexts can be hazardous.

All of the kinds of models discussed so far contain the important assumption of 'discrete flow'. This means that each subtask (e.g., identifying an item, performing a calculation) must be entirely completed before beginning the next. Thus, in serial processing, there is no overlap of mental activities taking place in succession. In parallel systems, of course, different subtasks can take place simultaneously, but anything happening next must await a discrete event, such as all of the preceding parallel items being completed.

It is easy to conceive of alternative systems. Even a task whose subtasks take place sequentially might allow overlap in their actual processing. A hybrid system (neither strictly serial nor strictly parallel) could have, say, the second item to be processed begin 50 ms after the first begins, even though the first is uncompleted (e.g., Taylor, 1976; Townsend and Ashby, 1983, pp. 370–371).

We refer to a system that allows total overlap, even though the structure is sequential, as exhibiting 'continuous flow'. Total overlap means that all items or channels in a sequence begin simultaneously even though they are not processed in parallel. This concept may seem strange at first. Actually, many practical systems – for instance, those used to describe electronic networks – closely approximate this situation. As the first process begins its work, it immediately begins sending output to the second process, which immediately sends its output to the third, and so on. McClelland (1979) developed a model obeying this principle and fit it to data usually interpreted as supporting strict serial models. (See also Ashby (1982).) He

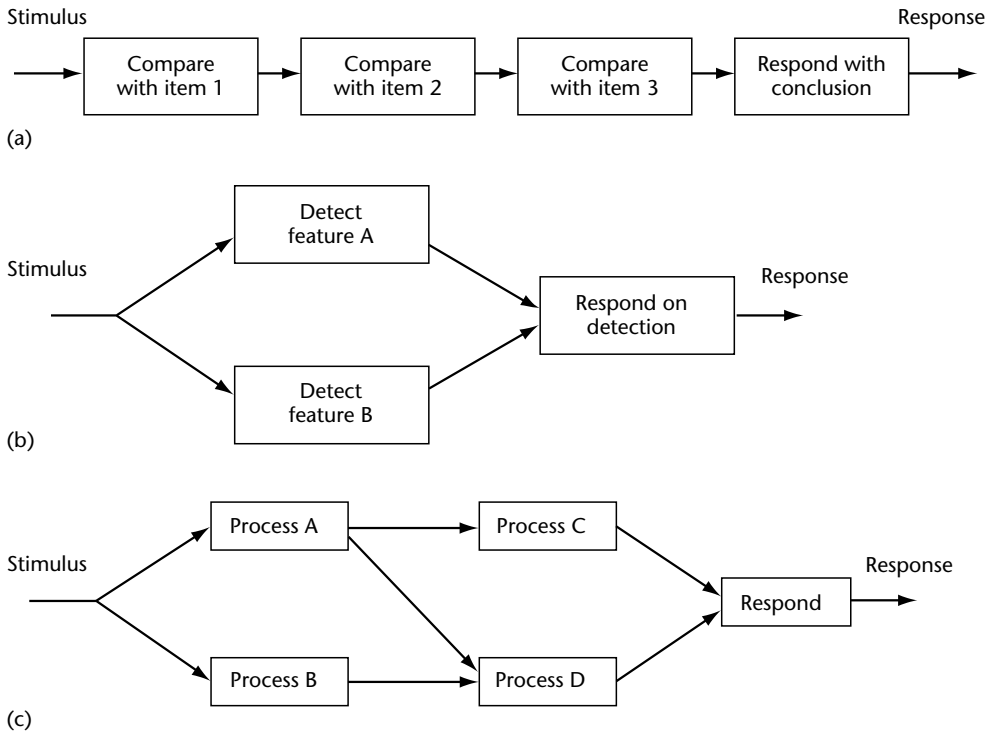


Figure 1. Flow diagrams. (a) A flow diagram of the short-term memory model discussed by Sternberg (1966). It shows a model that compares the stimulus with each item in memory. Each process in the diagram starts only when the previous one is finished, and all processes must be finished before a response is made. At the end, a response is made based on the results of the preceding processes. Note that this diagram does not specify the implementation. For instance, a computer with one central processing unit can perform each process in sequence; or several processing units may be connected in line to perform these processes. (b) A flow diagram of a simplified version of a parallel network investigated by Wenger and Townsend (2001). A 'yes' response is made when one or two features of the stimulus match those of a face in memory. The diagram shows that detection of the two features can be processed simultaneously. A response can be made when either one of the two processes is finished. (c) A flow diagram of a Wheatstone-bridge structure analyzed by Schweickert (1978). In this model, process C cannot start processing until process A is finished; process D cannot start until processes A and B are finished; and a response cannot be made until processes C and D are finished. Schweickert found that these dependencies can be investigated by a generalization of Sternberg's additive-factor method.

called his model the 'cascade model', borrowing terminology from engineering literature.

Naturally, there can exist parallel subsystems, each of which consists of a cascaded (i.e., continuous-flow) sequence of operations. Although such systems are likely to become increasingly popular (e.g., Usher and McClelland, 2001), their formal properties are more difficult to establish in a general manner than those of discrete-flow systems (Schweickert and Townsend, 1989; Schweickert, 1989).

INDEPENDENCE, CAPACITY, AND STOPPING RULES

Besides the architecture issue, the issue of independence versus dependence among the processing times of the items is also important (e.g., Townsend, 1974). Whether two items or channels operate independently is important in terms of both accuracy (Townsend *et al*., 1984; Ashby and Townsend, 1986) and response time (e.g., Townsend and Ashby, 1983). For instance, it is natural, in parallel systems, to assume that the parallel channels are independent of one another – just as, when two different people toss their own coins, we are justified in supposing that the two tosses are independent. It is usually assumed in serial processing that any two successive processing times are independent. Of course, in either parallel or serial systems independence will depend on the particular setting and task.

Another important concept is that of 'capacity', that is, the efficiency, speed and accuracy of processing. It is most often understood in terms of the effect on processing times of the number of

things to be worked on (Fisher, 1982; Townsend and Ashby, 1978, 1983). In parallel processing, it is clearly reasonable to suppose that processing slows down when the number of items to be processed increases (capacity is limited). However, limitations in capacity that are indirect, even with serial processing, can be conceived. For instance, a serial processor might speed up as it goes through the items, due to warm-up effects, or slow down, due to inertia or fatigue of the processor. Even though capacity and independence are logically separate notions, they can interact. For instance, an important type of parallel system, one that can mimic serial processing, assumes that as each item is completed its processing capacity is reallocated to other remaining items (Townsend and Ashby, 1983). This obviously affects the overall 'reaction time' (RT), but it also creates a positive dependency (i.e., facilitation) among the item processing times. And different types of dependency can affect capacity measures in different ways. Thus, it has been shown that positive dependencies across two parallel channels can speed things up.

Another important notion is that of the 'stopping rule'. Depending on the task, it may or may not be necessary for the participant to process all of the items in order to make a correct response. For example, in the memory-search paradigm, if a probe is present, the processing can cease before everything has been processed: an event known as 'self-termination'. However, if no probe is present, it is necessary to process all of the memory items in order to be sure of correctly making a 'no' response: that is, exhaustive processing must occur. In some experimental designs, all the items meet the probe's criterion (for instance, the 'probe' might be specified as 'any vowel' and all memory items might be vowels). In this case there is the possibility of 'first-terminating' or 'minimum time' processing.

Of course, it is an empirical question whether any kind of self-termination can actually take place in high-speed perceptual or cognitive operations. It has been proved that some of the early data from the 1960s that did not distinguish certain classes of parallel and serial models provides strong evidence for or against self-termination (van Zandt and Townsend, 1993; Townsend and Van Zandt, 1990; Townsend and Colonius, 1997).

MATHEMATICAL MODELS OF HUMAN INFORMATION PROCESSING AND COGNITION

The use of mathematical models is a powerful alternative to a flow diagram together with verbal reasoning. Although mathematical models have a long history in psychology, it is only since the 1950s that they have begun to exert a consistent and strong influence on psychology in general, and cognitive psychology in particular. They have been particularly effective within the information-processing approach, since the investigator can immediately begin to work out what a flow diagram purports to do in terms of mathematical expressions. Mathematical models express, in terms of formulae and equations, assumptions and explanations and predictions of psychological ideas.

A large number of mathematical models have been developed within the information-processing approach. Typically, a mathematical model will have some parameters, which are given as symbols that can assume values as numbers. For instance, consider the rather trivial model given by the equation $y = ax + b$. In this model, the constants a and b are parameters, which, in general, can take any numerical value. The term x is the main independent variable (usually not called a parameter, although usage varies), which is manipulated in the model while a and b are held constant. Mathematical models, whether in cognitive science, physics or biology, can be highly complex, but have been of inestimable value.

Mathematical models must be tested against real data. The usual way of testing a model requires us first to find the parameter values that allow the model to come as close as possible to a set of experimental data (the so-called 'fit procedure'), and then to examine how well the model does indeed reproduce the data. Some measures of fit permit statistical testing of the model, to learn whether it differs from the data more than would be expected on the basis of chance. These are called 'tests of significance'. Since the use of sparse or noisy data can sometimes make a model seem to fit data better than it deserves (i.e., where a large representative set of data might falsify the model), many theorists recommend comparisons of models obeying different and often opposing principles (e.g., parallel versus serial principles). This is a survival-of-the-fittest strategy. It has much to recommend it, especially since it is a very human tendency, even among scientists, to want to 'save' one's own model or theory.

However, a rather different strategy has emerged over the past few decades: to attempt to discover broad, qualitative kinds of predictions that a model, or a whole set of models (e.g., the set of all parallel models, that is all models that predict simultaneous processing of the task under study), make for a set of phenomena. For instance, one

model might predict that a certain dependent variable is a linear function of a certain independent variable, while another predicts a concave function. Then graphing the data can falsify one model and support another, even though the investigator does not actually fit either model to the data. A more complex, but much more powerful, qualitative approach is found in 'systems-factorial methodologies', which we describe later.

In recent years, mathematical models have become a part of everyday cognitive and experimental psychology. In the 1950s it was rare to find a mathematical model in a journal of experimental psychology, and most were found in specialist quantitatively-oriented journals. Nowadays, any issue of an experimental journal will typically contain several articles that employ mathematical models.

CODES IN INFORMATION PROCESSING

Since the term 'information processing' explicitly refers to information, it is natural to inquire about the nature and measurement of information, that is, the 'code' used, in various perceptual and cognitive operations by the individual.

The traditional unit of information, namely, the 'bits' (short for 'binary unit', basically equal to the logarithm to base 2 of the number of message possibilities), has rarely been employed in recent models. But there are exceptions. In studies related to perception, such as pattern recognition, where the observer must identify which of a set of patterns (letters, words, faces, etc.) has been presented, or in areas where the coding is an integral part of the science, such as psycholinguistics, researchers do explicitly employ, and sometimes test, the code used in the pertinent operations. An example is the study of perception of letters and letter-like forms, where hypothetical sets of features making up the visually-presented letters are subjected to study (e.g., Townsend and Ashby, 1982). Things like lines and curves, as well as more complicated features such as whether a letter is closed (like an 'O') or open (like a 'C'), have been examined. However, other approaches have sought explanation in terms of the wave forms of light and dark of which any picture is composed (i.e., Fourier analysis): such an approach is akin to the analysis of sounds into frequencies that are related to pitch sensation (e.g., Ginsburg, 1986).

Often, in the absence of any knowledge or hypotheses about the form of the code, one can simply measure the information, or important aspects of the information, by means of parameters in mathematical models. A parameter in a model can represent something that relates to a code, without actually being a code. For example, consider a model of pattern identification that assumes that the likelihood of confusing two items is a function of the similarity between the two items. For instance, in the 'overlap model', the probability of confusing, say, the letter K with the letter R would be given by the parameter c_{KR} (e.g., Townsend, 1971). A very successful model of identification called the 'similarity choice model' (Luce, 1963) also uses similarity parameters.

Note that these models do not try to give the perceptual code of the stimulus patterns. A response-time model may possess a parameter that actually measures the rate of processing of, say, the visual features that make up a letter. However, often the model will simply include a term that represents the likelihood that a sub-process is completed by a certain time, without clear specification of a code or how fast that code is employed or processed.

SYSTEMS-FACTORIAL METHODOLOGIES FOR DETERMINING ARCHITECTURE

One promising general qualitative approach to testing different ideas about mental architecture is based on the notion of 'selective influence' of experimental factors: a notion first employed in tests of strict seriality by Sternberg (1969) in his well-known 'additive-factor method'. All factorial methodologies, like the original Sternberg strategy, depend on the assumption (powerful, if true) that distinct experimental factors affect distinct processing components (i.e., subsystems). This is the assumption of selective influence. The expression $\overline{\mathrm{RT}}(X + \Delta X, Y)$ refers to the mean RT where the X factor has prolonged RT but Y is at base level; and similarly for $\overline{\mathrm{RT}}(X, Y + \Delta Y)$. The fundamental statistic for the original method, and most extensions, is the 'mean interaction contrast', defined as $\mathrm{MIC} = [\overline{\mathrm{RT}}(X + \Delta X, Y + \Delta Y) - \overline{\mathrm{RT}}(X + \Delta X, Y)] - [\overline{\mathrm{RT}}(X, Y + \Delta Y) - \overline{\mathrm{RT}}(X, Y)]$. Schweickert (1978) formulated the first major extension of the additive-factor method involving more complex architectures under the assumption of selective influence, in his 'latent mental network theory'. Townsend and Ashby (1983) found that the mean interaction contrast distinguished parallel and serial stochastic models, when selective influence was assumed, and Schweickert and Townsend (1989) produced general theorems for Schweickert's

latent networks, within a stochastic setting, assuming exhaustive processing. Note that factorial strategies do not demand that the investigator provide a specification of the information codes that the underlying system is using.

Although the early theorems were established in the context of exhaustive processing, analogous results can be found in the case of self-terminating and first-terminating processing times (e.g., Schweickert and Giorgini, 1999; Townsend and Nozawa, 1995). Because Sternberg's original ideas have been extended in so many new directions, it has been suggested that the general approach be referred to as 'systems-factorial technology' (Townsend and Thomas, 1994). For instance, one interesting strategy has been to enlist entire RT distributions in providing more powerful tests of parallel versus serial processing or other related architectures (Dzhafarov and Schweickert, 1995; Roberts and Sternberg, 1993; Townsend, 1990; Townsend and Ashby, 1983; Balakrishnan, 1994).

The assumption of selective influence is critical to the legitimacy of systems-factorial technology, and much is now known about its foundational underpinnings and what may go awry if it is violated (Dzhafarov, 1997; Townsend, 1984; Townsend and Schweickert, 1989; Townsend and Thomas, 1994).

Goldstein and Fisher (1991) have produced a theory and methodology for stochastic networks and processing rules even more general than those based on directed graphs. Miller (1993) and Liu (1996) have proposed theoretical approaches based on the theory of 'queuing processes'. Queuing theory is an approach of great value in engineering. It studies or describes how customers in lines waiting for service in various arrangements are served, and how long it takes with various arrangements. For instance, queuing theory may try to find an optimal architecture to enable lines of partially-finished products to make their way through a factory. The tools of queuing theory, and the general networks of Goldstein and Fisher (1991), may prove of considerable value to the study of human information processing, when appropriately adapted.

CONCLUSION

The information-processing approach has grown in breadth and depth, far beyond the sometimes simplistic early use of computer analogies, flow diagrams, and other concepts and terminologies that were not always backed up by rigorous theory and methodology. With the aid of modern mathematical modelling, theory-driven methodology, and adept experimentation, we can undoubtedly look forward to continued and cumulative progress in the understanding of human information processing.

References

Ashby FG (1982) Deriving exact predictions from the cascade model. *Psychological Review* **89**: 599–607.

Ashby FG and Townsend JT (1986) Varieties of perceptual independence. *Psychological Review* **93**: 154–179.

Ashcraft MH (2002) *Cognition*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall.

Attneave F (1959) *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. New York, NY: Holt, Rinehart & Winston.

Balakrishnan JD (1994) Simple additivity of stochastic psychological processes: tests and measures. *Psychometrika* **59**: 217–240.

Dzhafarov EN (1997) Process representations and decompositions of response times. In: Marley AAJ (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*, pp. 255–277. Mahwah, NJ: Lawrence Erlbaum.

Dzhafarov EN and Schweickert R (1995) Decompositions of response times: an almost general theory. *Journal of Mathematical Psychology* **39**: 285–314.

Fisher DL (1982) Limited-channel models of automatic detection: capacity and scanning in visual search. *Psychological Review* **89**: 662–692.

Garner WR (1962) *Uncertainty and Structure as Psychological Concepts*. New York, NY: John Wiley.

Ginsburg AP (1986) Spatial filtering and visual form perception. In: Boff KR and Kaufman L (eds) *Handbook of Perception and Human Performance*, vol. II, *Cognitive Processes and Performance*, pp. 1–41. New York, NY: John Wiley.

Goldstein WM and Fisher DL (1991) Stochastic networks as models of cognition: derivation of response time distributions using the order-of-processing method. *Journal of Mathematical Psychology* **35**: 214–241.

Liu Y (1996) Queueing network modeling of elementary mental processes. *Psychological Review* **103**: 116–136.

Luce RD (1960) *Developments in Mathematical Psychology*. New York, NY: Free Press.

Luce RD (1963) Detection and recognition. In: Luce RD, Bush RR and Galanter E (eds) *Handbook of Mathematical Psychology*, vol. I, pp. 103–189. New York, NY: John Wiley.

McClelland JL (1979) On the time relations of mental processes: an examination of systems of processes in cascade. *Psychological Review* **86**: 287–330.

Miller JO (1993) A queue-series model for reaction time, with discrete-stage and continuous-flow models as special cases. *Psychological Review* **100**: 702–715.

Roberts S and Sternberg S (1993) The meaning of additive reaction-time effects: tests of three alternatives. In: Meyer DE and Kornblum S (eds) *Attention and*

Performance, vol. XIV, *Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, pp. 611–653. Cambridge, MA: MIT Press.
Roediger HL (1980) Memory metaphors in cognitive psychology. *Memory and Cognition* **8**: 231–246.
Schweickert R (1978) A critical path generalization of the additive factor method: analysis of a Stroop task. *Journal of Mathematical Psychology* **18**: 105–139.
Schweickert R (1989) Separable effects of factors on activation functions in discrete and continuous models: *d'* and evoked potentials. *Psychological Bulletin* **106**: 318–328.
Schweickert R and Giorgini M (1999) Response time distributions: some simple effects of factors selectively influencing mental processes. *Psychonomic Bulletin and Review* **6**: 269–288.
Schweickert R and Townsend JT (1989) A trichotomy: interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. *Journal of Mathematical Psychology* **33**: 328–347.
Shannon CE and Weaver W (1949) *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.
Sternberg S (1966) High-speed scanning in human memory. *Science* **153**(3736): 652–654.
Sternberg S (1969) The discovery of processing stages: extensions of Donders' method. *Acta Psychologica* **30**: 276–315.
Taylor DA (1976) Stage analysis of reaction time. *Psychological Bulletin* **83**: 161–191.
Townsend JT (1971) Theoretical analysis of an alphabet confusion matrix. *Perception and Psychophysics* **9**: 40–50.
Townsend JT (1974) Issues and models concerning the processing of a finite number of inputs. In: Kantowitz BH (ed.) *Human Information Processing: Tutorials in Performance and Cognition*, pp. 133–168. Hillsdale, NJ: Lawrence Erlbaum.
Townsend JT (1984) Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology* **28**: 363–400.
Townsend JT (1990) Serial vs. parallel processing: sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science* **1**: 46–54.
Townsend JT and Ashby FG (1978) Methods of modeling capacity in simple processing systems. In: Castellan NJ and Restle F (eds) *Cognitive theory*, vol. III, pp. 200–239. Hillsdale, NJ: Lawrence Erlbaum.
Townsend JT and Ashby FG (1982) Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology: Human Perception and Performance* **8**: 834–864.
Townsend JT and Ashby FG (1983) *The Stochastic Modeling of Elementary Psychological Processes* Cambridge, UK: Cambridge University Press.
Townsend JT and Colonius H (1997) Parallel processing response times and experimental determination of the stopping rule. *Journal of Mathematical Psychology* **41**: 392–397.
Townsend JT and Nozawa G (1995) Spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology* **39**: 321–359.
Townsend JT and Schweickert R (1989) Toward the trichotomy method of reaction times: laying the foundation of stochastic mental networks. *Journal of Mathematical Psychology* **33**: 309–327.
Townsend JT and Thomas RD (1994) Stochastic dependencies in parallel and serial models: effects on systems factorial interactions. *Journal of Mathematical Psychology* **38**: 1–34.
Townsend JT and van Zandt T (1990) New theoretical results on testing self-terminating vs exhaustive processing in rapid search experiments. In: Geissler H-G and Miller MH (eds) *Psychophysical Explorations of Mental Structures*, pp. 469–489. Kirkland, WA: Hogrefe & Huber.
Townsend JT, Hu GG and Evans RJ (1984) Modeling feature perception in brief displays with evidence for positive interdependencies. *Perception and Psychophysics* **36**: 35–49.
Usher M and McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review* **108**: 550–592.
Wenger MJ and Townsend JT (2001) Faces as Gestalt stimuli: process characteristics. In: Wenger MJ and Townsend JT (eds) *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, pp. 229–284. Mahwah, NJ: Lawrence Erlbaum.
Wiener N (1948) *Cybernetics*. New York, NY: MIT Press.
van Zandt T and Townsend JT (1993) Self-terminating versus exhaustive processes in rapid visual and memory search: an evaluative review. *Perception and Psychophysics* **53**: 563–580.

Further Reading

Ashby FG and Townsend (1986) Varieties of perceptual independence. *Psychological Review* **93**: 154–179.
Batchelder WH and Riefer DM (1990) Multinomial processing models of source monitoring. *Psychological Review* **4**: 548–564.
Busemeyer JB and Townsend JT (1993) Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* **100**: 432–459
Dosher BA and Liu ZL (1999) Mechanisms of perceptual learning. *Vision Research* **39**: 3197–3221.
Fisher DL and Glaser R (1996) Molar and latent models of cognitive slowing: implications for aging, dementia, depression, development and intelligence. *Psychonomic Bulletin and Review* **3**: 458–480.
O'Toole AJ, Wenger MJ and Townsend JT (2001) Quantitative models of perceiving and remembering

faces: precedents and possibilitis. In: Wenger MJ and Townsend JT (eds) *Computational, Geometric and Process Perspectives on Facial Cognition: Contexts and Challenges*, pp. 1–38. Mahwah, NJ: Erlbaum.

Smith PL (2000) Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology* **44**: 408–463.

Townsend JT and Nozawa G (1995) On the spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology* **39**: 321–359.

Information Theory

Introductory article

Zoubin Ghahramani, University College London, UK

CONTENTS

Information theory is a mathematical theory defining the limits and possibilities of communication. It provides a quantitative measure of the information content of a message, which is independent of the meaning of the message, in terms of the reduction of uncertainty resulting from receiving the message.

INTRODUCTION

Information is the reduction of uncertainty. Imagine your friend invites you to dinner for the first time. When you arrive at the building where he lives you find that you have misplaced his apartment number. He lives in a building with 4 floors and 8 apartments on each floor. If a neighbor passing by tells you that your friend lives on the top floor, your uncertainty about where he lives is reduced from 32 choices to 8. By reducing your uncertainty, the neighbor has conveyed *information* to you. How can we quantify the amount of information?

Information theory is the branch of mathematics that describes how uncertainty should be quantified, manipulated, and represented. Ever since the fundamental premises of information theory were laid down by Claude Shannon in 1949, it has had far-reaching implications for almost every field of science and technology. Information theory has also played an important role in shaping theories of perception, cognition, and neural computation. This article will discuss some of the basic concepts in information theory and how they relate to cognitive science and neuroscience.

ENTROPY AND MUTUAL INFORMATION

The most fundamental concept in information theory is entropy. Shannon borrowed the concept of entropy from thermodynamics, where it describes the amount of disorder of a system. In information theory, entropy measures the amount of uncertainty of an unknown or random quantity. The entropy of a random variable X is defined to be

$$H(X) = -\sum_x P(x)\log_2 P(x) \quad (1)$$

where the sum is over all values x that the variable X can take, and $P(x)$ is the probability of the value x occurring. Entropy is measured in *bits*, and can be generalized to continuous variables, although care must be taken to specify the precision level at which we would like to represent the continuous variable. Returning to our example, if X is the random variable that describes which apartment your friend lives in, initially it can take on 32 values with equal probability $P(x) = 1/32$. Since $\log_2(1/32) = -5$, the entropy of X is 5 bits. After the neighbor tells you that he lives on the top floor, the probability of X drops to 0 for 24 of the 32 values and becomes $1/8$ for the other 8 equally probable values. The entropy of X thus drops to 3 bits (by convention $0 \log 0 = 0$). The neighbor has therefore conveyed 2 bits of information to you.

This fundamental definition of entropy as a measure of uncertainty can be derived from a small set of axioms. Entropy is the average amount of 'surprise' associated with a set of events. The amount of 'surprise' of a particular event x is a function of the probability of that event: the less probable an event (e.g. a moose walking down Wall Street), the more surprising it is. The amount of surprise of two independent events (e.g. the moose and a solar eclipse) should be the sum of the amounts of surprise of the events. These two constraints imply that the surprise of an event is proportional to $-\log P(x)$, with the proportionality constant determining what base logarithms are taken in (e.g. base 2 for bits). Averaging over all events according to their respective probabilities, we get the expression for $H(X)$.

Entropy in information theory has deep connections with the thermodynamic concept of entropy and, as we will see, it can be related to the smallest

number of bits it would take on average to communicate X from one location (the sender) to another (the receiver). On the one hand, the concepts of entropy and information are universal, in the sense that a bit of information can refer to the answer to any yes–no question where the two options are equally probable. A megabit is a megabit (the answers to about a million yes–no questions, which can potentially distinguish between $2^{1\,000\,000}$ possibilities) regardless of whether it is used to encode a picture, music, or large quantities of text. On the other hand, entropy is always measured relative to a probability distribution, and for many situations it is not possible to consider the 'true' probability of an event. For example, I may have high uncertainty about the weather tomorrow, but a meteorologist might not. Hence there may be different entropies for the same set of events, defined relative to the subjective beliefs of the entities whose uncertainty we are measuring. This subjective, or Bayesian, view of probabilities is useful when we are considering how information communicated between different (biological or artificial) agents changes their beliefs.

While entropy is useful in determining the uncertainty in a single variable, it does not tell us how much uncertainty we have in one variable given knowledge of another. For this we need to define the *conditional entropy* of X given Y:

$$H(X|Y) = -\sum_{x,y} P(x \,\&\, y)\log_2 P(x|y) \tag{2}$$

where $P(x|y)$ denotes the probability of x given that we have observed y. Building on this definition, the *mutual information* between two variables is the reduction in uncertainty in one variable given another variable. Mutual information can be written in three different ways:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X\&Y) \end{aligned} \tag{3}$$

The mutual information between two variables is symmetric: $I(X;Y) = I(Y;X)$. If the random variables X and Y are independent, that is, if the probability of their taking on values x and y is $P(x \,\&\, y) = P(x)\,P(y)$ for all x and y, then they have zero mutual information. Similarly, if one can determine X exactly from Y and vice versa, the mutual information is equal to the entropy of either of the two variables.

SOURCE CODING

Consider the problem of transmitting a sequence of symbols across a communication line using a binary representation. We assume that the symbols come from a finite alphabet (e.g. letters and punctuation marks of text, or levels from 0 to 255 of grayscale image patches) and, to begin with, that the communication line is noise-free. We further assume that the symbols are produced by a source which emits each symbol x randomly with some known probability $P(x)$. How many bits do we need to transmit per symbol so that the receiver can perfectly decode the sequence of symbols?

If there are N symbols in the alphabet then we could assign a distinct binary string (called a *codeword*) of length L to each symbol, provided $2^L > N$. This suggests that we would need at most L bits. But we can do much better than this by assigning shorter codewords to more probable symbols and longer codewords to less probable ones. Shannon's '*noiseless source coding theorem*' states that if the source has entropy $H(X)$ then there exists a decodable prefix code having an average length L per symbol such that

$$H(X) \le L < H(X) + 1 \tag{4}$$

Moreover, no uniquely decodable code exists having a smaller average length. In a *prefix* code, no codeword starts with another codeword, so the message can be decoded unambiguously as it comes in. This result places a lower bound on how many bits are required to compress a sequence of symbols losslessly. A closely related concept is the Kolmogorov complexity of a finite string, defined as the length in bits of the shortest program which when run on a universal computer will cause the string to be output.

Huffman Codes

We can achieve the code length described by Shannon's noiseless coding theorem using a very simple algorithm. The idea is to create a prefix code which uses shorter codewords for more frequent symbols and longer codewords for less frequent ones. First we combine the two least frequent symbols, summing their frequencies, into a new symbol. We do this repeatedly until we have only one symbol. The result is a tree with the original symbols at the leaves. This is illustrated in Figure 1 using an alphabet of 7 symbols {a,b,c,d,e,o,k} with differing probabilities. The codeword for each symbol is the sequences of left (0) and right (1) moves required to reach that symbol from the top of the tree.

In this example we have 7 symbols, so the naive fixed-length code would require 3 bits per symbol ($2^3 = 8 > 7$). The Huffman code (which is a variable-length code) requires on average 2.48 bits; the

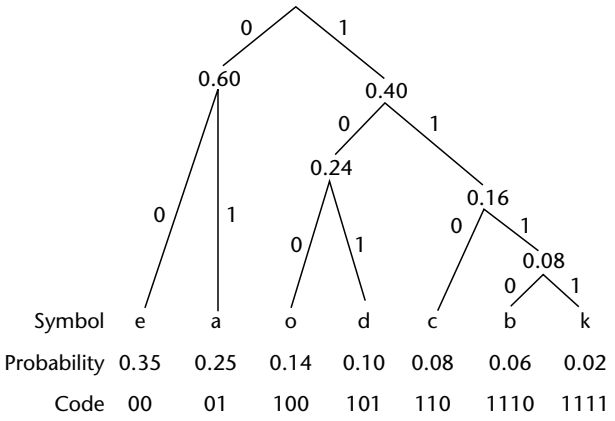


Figure 1. Huffman coding for an alphabet of seven symbols with differing probabilities.

entropy gives a lower bound of 2.41 bits. The fact that it is a prefix code makes it easy to decode a string symbol by symbol, by starting from the top of the tree and moving down left or right every time a new bit arrives. For example, try decoding: 1010011010010100.

If we want to improve on this to get closer to the entropy bound, we can code blocks of several symbols at a time. Many practical coding schemes work by forming blocks of symbols and coding each block separately. Using blocks also makes it possible to correct for errors introduced by noise in the communication channel.

Information Transmission Along a Noisy Channel

In the real world, communication channels suffer from noise. When transmitting data to a mobile phone, listening to a person in a crowded room, or playing a DVD movie, there are random fluctuations in signal quality, background noise, or disk rotation speed, which we cannot control. The noise on a channel can be simply characterized by the conditional probability of the received symbols given the transmitted symbol: $P(r|t)$. This noise limits the information capacity of the channel, which is defined to be the maximum, over all possible distributions over the transmitted symbols T, of the mutual information between the transmitted symbol and the received symbol R:

$$C = \max_{p} I(T;R) \tag{5}$$

For example, if the symbols are binary and the channel has no noise, then the channel capacity is 1 bit per symbol (corresponding to transmitting 0 and 1 with equal probability). However, if 10% of the time a 0 transmitted is received as a 1, and 10% of the time a 1 transmitted is received as a 0, then the channel capacity is only 0.53 bits per symbol.

This probability of error could not be tolerated in most real applications. Does this mean that the channel is unusable? Not if one uses the trick of building redundancy into the transmitted signal in the form of an 'error-correcting code' so that the receiver can still decode the intended message (Figure 2). One simple scheme is a 'repetition code'. For example, encode the symbols by transmitting three repetitions of each; decode them by taking blocks of three and outputting the majority vote. This reduces the error probability from 10% to 2.8%, at the cost of reducing the rate at which the original symbols are transmitted by 1/3.

If we want to achieve an error probability approaching zero, do we need to transmit at a rate approaching zero? Remarkably the answer is negative, as Shannon proved in his '*channel coding theorem*'. This states that all rates below channel capacity are achievable; i.e., that there are codes which transmit at a given rate and have a maximum probability of error approaching zero. Conversely, if a code has probability of error approaching zero, it must have a rate less than or equal to channel capacity. Unfortunately Shannon's channel coding theorem does not say how to design codes that approach zero error probability near the channel capacity. Of course, codes with this property are more sophisticated than the repetition code, and finding good error-correcting codes that can be decoded in reasonable time is an active area of research. Shannon's result is of immense practical significance since it shows that we can have essentially perfect communication over a noisy channel.

INFORMATION THEORY AND LEARNING SYSTEMS

Information theory has played an important role in the study of learning systems. Just as information theory seeks to quantify information regardless of its physical medium of transmission, so learning theory seeks to understand systems that learn regardless of whether they are biological or artificial.

Learning systems can be broadly categorized by the amount of information they receive from the environment in their supervision signal. In 'unsupervised learning', the goal of the system is to learn from sensory data with no supervision. This can be achieved by casting the unsupervised learning problem as one of discovering a code for the system's sensory data which is as efficient as

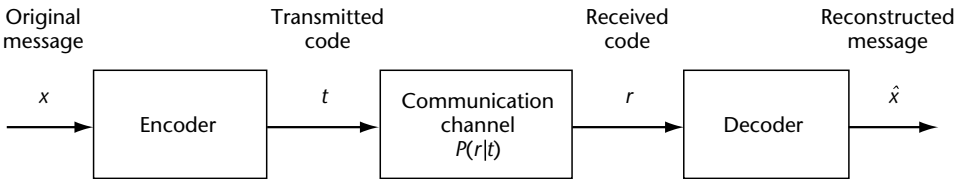


Figure 2. A noisy communication channel.

possible. Thus the concepts of entropy, Kolmogorov complexity, and description length can be used to formalize unsupervised learning problems.

We know from the source coding theorem that the most efficient code for a data source is one that uses $-\log_2 P(x)$ bits per symbol x. Therefore, the problem of discovering the optimal coding scheme for a set of sensory data is equivalent to the problem of learning what the true probability distribution $P(x)$ of the data is. If at some stage we have an estimate $Q(x)$ of this distribution, we can use this estimate instead of the true probabilities to code the data. However, we incur a loss in efficiency measured by the *relative entropy* between the two probability distributions P and Q:

$$D(P\|Q) = \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (6)$$

which is also known as the Kullback–Leibler divergence. This measure is the inefficiency in bits of coding messages with respect to a probability distribution Q instead of the true probability distribution P, and is zero if and only if $P = Q$. Many unsupervised learning systems can be designed from the principle of minimizing the relative entropy.

INFORMATION THEORY IN COGNITIVE SCIENCE AND NEUROSCIENCE

The term 'information processing system' has often been used to describe the brain. Indeed, information theory can be used to understand a variety of functions of the brain. A few examples are mentioned here.

In neurophysiological experiments where a sensory stimulus is varied and the spiking activity of a neuron is recorded, mutual information can be used to infer what the neuron is coding for. The mutual information for different coding schemes can be compared, for example, to test whether the exact spike timing is used for information transmission. (*See* **Decoding Neural Population Activity**; **Rate versus Temporal Coding Models**)

Information theory has been used to study both perceptual phenomena and the neural substrate of early visual processing. It has been argued that the representations found in the visual cortex arise from principles of redundancy reduction and optimal coding.

Communication via natural language occurs over a channel with limited capacity. Estimates of the entropy of natural language can be used to determine how much ambiguity or surprise there is in the next word following a stream of previous words, and learning methods based on entropy can be used to model language. (*See* **Natural Language Processing, Statistical Approaches to**)

Redundant information arrives from multiple sensory sources (e.g. vision and audition), and over time (e.g. a series of frames of a movie). Decoding theory can be used to determine how this information should be combined optimally and whether the human system does so.

The human movement control system must cope with noise in motor neurons and in the muscles. Different ways of coding the motor command result in more or less variability in the movement.

Information theory lies at the heart of our understanding of computing, communication, knowledge representation, and action. As in many other fields of science, the basic concepts of information theory have played, and will continue to play, an important role in cognitive science and neuroscience.

Further Reading

Attneave F (1954) Informational aspects of visual perception. *Psychological Review* **61**: 183–193.

Barlow HB (1961) The coding of sensory messages. In: Thorpe and Zangwill (eds) *Current Problems in Animal Behaviour*, pp. 330–360. Cambridge, UK: Cambridge University Press.

Berger A, Della Pietra S and Della Pietra V (1996) A maximum entropy approach to natural language processing. *Computational Linguistics* **22**(1): 39–71.

Cover TM and Thomas JA (1991) *Elements of Information Theory*. New York, NY: John Wiley [General reference.]

Ghahramani Z (1995) *Computation and Psychophysics of Sensorimotor Integration*. PhD thesis, Massachusetts Institute of Technology.

Harris CM and Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* **394**: 780–784.

MacKay DJC (2001) *Information Theory, Inference and Learning Algorithms*. http://wol.ra.phy.cam.ac.uk/mackay/itprnn/book.html. [General reference.]

Olshausen BA and Field DJ (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* **381**: 607–609.

Rieke F, Warland D, de Ruyter van Steveninck R and Bialek W (1999) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Shannon CE and Weaver WW (1949) *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press. [General reference.]

Intelligence

Introductory article

Robert J Sternberg, Yale University, New Haven, Connecticut, USA

CONTENTS

Intelligence has been defined variously, across a range of cultures and populations, as encompassing such mental and social phenomena as learning and understanding, reasoning and problem-solving, perception and adaptation.

INTRODUCTION

Intelligence, according to *Webster's New World College Dictionary* (3rd edn), is 'the ability to learn or understand from experience, ability to acquire and retain knowledge; mental ability…' Such a definition captures many facets of the nature of intelligence, but not necessarily those believed by experts to be key.

Two symposia, one in 1921 and the other in 1986, sought to ascertain the key features of intelligence according to experts in the field such as Lewis Terman and Edward Thorndike (in the 1921 symposium) and Ulric Neisser and Douglas Detterman in the 1986 symposium. Critical elements of the definition of intelligence, according to experts, are (a) adaptation in order to meet the demands of the environment effectively, (b) elementary processes of perception and attention, (c) higher level processes of abstract reasoning, mental representation, problem-solving, decision-making, (d) ability to learn, and (e) effective behavior in response to problem situations.

Some experts, however, have been content to define intelligence operationally, simply as what intelligence tests test. This definition, first proposed by Edwin Boring in 1923, relies on tests to define intelligence and ultimately is circular.

Laypeople also can be asked to define intelligence, and it turns out that their definitions differ from expert definitions in placing somewhat greater emphasis on social competence skills. In one study, for example, laypeople defined intelligence in terms of three broad classes of skills: (a) practical problem-solving, (b) verbal ability, and (c) social competence. But how people define intelligence varies across occupations. For example, one study found that philosophy professors tend to stress critical and logical thinking very heavily, whereas physicists tend to place more value on precise mathematical thinking, the ability to relate physical phenomena to concepts of physics, and the ability to grasp quickly the laws of nature.

Definitions of intelligence also vary across cultures. For example, in a study by Yang and Sternberg of Chinese people's conceptions of intelligence in Taiwan, it was found that in addition to emphasizing cognitive skills in their conceptions of intelligence, the Taiwanese also emphasized interpersonal skills, intrapersonal (self-understanding) skills, knowing when to show one's intelligence, and knowing when *not* to show one's intelligence.

FACTOR-ANALYTIC THEORIES

One of the oldest approaches to understanding intelligence is through psychometric theories. These theories seek understanding of intelligence through its measurement.

Among these theories, the earliest major one is that of Spearman, who proposed that intelligence comprises a general factor (*g*) of intelligence common to all intellectual tasks, as well as specific factors (*s*), each of which is unique to a given test of intelligence. His proposal was based on his finding of a 'positive manifold' among intelligence tests: all tests seemed to be positively intercorrelated, suggesting the existence of a general factor. Spearman's theory still has many proponents today, such as Arthur Jensen.

Louis Thurstone disagreed with Spearman, arguing that the general factor was an artefact of the way Spearman analyzed his data. Thurstone suggested that seven primary mental abilities underlie intelligence: verbal comprehension, verbal fluency, number, spatial visualization, inductive reasoning, memory, and perceptual speed. More modern theorists, such as Raymond Cattell and John

Carroll, have attempted to integrate these two kinds of views, suggesting that intelligence is best understood hierarchically, with a general factor at the top of the hierarchy and narrower factors under it. Cattell proposed two such factors: fluid intelligence, which is involved in reasoning with novel kinds of stimuli; and crystallized intelligence, or stored knowledge base.

J. P. Guilford had a very different kind of model. He suggested that intelligence could be understood in terms of 150 separate factors that represented different combinations of operations, products, and contents. Operations included cognition (knowing), memory, divergent production (generation of alternatives), convergent production, and evaluation. Products included units, classes, relations, systems, transformations, and implications. Finally, contents included the figural, symbolic, semantic, and behavioral. For example, solving a verbal analogy such as 'LAWYER is to DOCTOR as CLIENT is to ?' would require cognition of semantic relations. Although this theory was popular at one time, the empirical evidence supporting it was shown by John Horn to be flawed. In particular, using Guilford's methods, Horn found that random data yielded as good a fit to Guilford's model as did real data!

Not all psychometric work on intelligence has been theoretically based. One of the earlier theories of intelligence was proposed by Sir Francis Galton, an Englishman, in 1883. Galton believed that intelligence is the capacity for labor and sensitivity to physical stimuli. Galton and his follower in the United States, James McKeen Cattell, measured intelligence using tests to assess skills such as the rate of arm movement over a distance of 50 cm, the threshold for distance on the skin by which two points need to be separated for them to be felt separately, and the span of letters that could be recalled from memory.

A more plausible inroad into intelligence testing was made by Alfred Binet and Theodore Simon in France. In 1904, the Minister of Public Instruction in Paris appointed a commission to find a means to differentiate truly mentally 'defective' children from those who were unsuccessful in school for other reasons. The commission was to ensure that no child suspected of mental retardation be placed in a special class without first being given an examination to measure the child's intelligence. Binet and Simon created the test, which measured judgment skills. A later test, created by David Wechsler, also measured primarily judgment skills.

Both the Binet and the Wechsler tests are still used in modern versions. What is actually found on such tests? The Wechsler tests (which cover three age ranges from infancy through adulthood) are divided into two major scales, verbal and performance. The verbal scale includes items measuring (1) comprehension, which requires examinees to answer questions of social knowledge; (2) vocabulary, which requires individuals to define words; (3) information, which requires people to supply generally known information; (4) similarities, which requires examinees to say in what ways two concepts are similar; (5) arithmetic, which requires people to solve simple arithmetical word problems; and (6) digit span, which requires individuals to repeat back number series, forward or backward. The performance scale includes items requiring test-takers to solve tasks. The subtests are (1) object assembly, which requires the participant to put together a puzzle by combining pieces to form a particular common object; (2) block design, which requires the individual to use patterned blocks to form a design that looks identical to a design shown by the experimenter; (3) picture completion, which requires an individual to say what is missing from each of a set of pictures; (4) picture arrangement, which requires the individual to put a set of cartoon-like pictures into a chronological order, so they tell a coherent story; and (5) digit symbol, which requires the individual to use a key matching particular symbols to particular numerals to copy a sequence of symbols based on the key of correspondence to the numerals.

KINDS OF INTELLIGENCE

In recent years, scholars seeking to understand intelligence have sought to understand not just factors of intelligence, but also kinds of intelligence. Three major theories have received special attention.

Howard Gardner has proposed a theory of multiple intelligences, according to which intelligence is not just a single entity, but eight separate entities. These are (1) linguistic intelligence – used to read a book, write a paper, or understand speech; (2) logical–mathematical intelligence, used to solve mathematical problems and balance a checkbook; (3) spatial intelligence, used to get from one place to another and to read a map; (4) musical intelligence, used to sing a song or compose a sonata; (5) bodily–kinesthetic intelligence, used to dance or play basketball; (6) naturalist intelligence, used to understand patterns in the natural world; (7) interpersonal intelligence, used to relate to other people; and (8) intrapersonal intelligence, used to understand oneself. So far, there have been no

psychometric tests designed formally to measure the different intelligences in this theory, so it is in need of empirical validation.

Robert Sternberg has proposed that intelligence comprises three aspects. Analytical abilities are used to analyze, evaluate, and critique: for example, they are used to decide whether a certain argument you or someone else has made is a logical argument. Creative abilities are used to create, discover, and invent: for example, you use creative intelligence when you come up with new ideas for a paper topic, a scientific experiment, or a work of art. Practical abilities are used to apply, implement, and utilize: for example, they are used when you decide how you should best say something to someone to convince that person to do what you wish them to do.

According to Sternberg's triarchic theory of human intelligence, intelligence draws on three kinds of information processes. Metacomponents are used to plan, monitor, and evaluate problem-solving. Performance components are used to implement the commands of the metacomponents. Knowledge-acquisition components are used to learn how to solve particular kinds of problems in the first place.

A third theory is the theory of emotional intelligence, originally proposed by Peter Salovey and John Mayer and then popularized by Daniel Goleman. Emotional intelligence is the ability to perceive accurately, appraise, and express emotions; the ability to access and/or generate feelings when they facilitate thought; the ability to understand emotion and emotional knowledge; and the ability to regulate emotions to promote emotional and intellectual growth. Although emotional intelligence is a relatively new construct, there is already some solid evidence suggesting its validity as a psychological construct. Mayer and Salovey have devised tests of emotional intelligence which predict behavior independently of the behavior's relations to conventional cognitive intelligence.

GENETIC AND ENVIRONMENTAL FACTORS

The ancient nature–nurture controversy continues in regard to intelligence. However, today, the large majority of psychologists and behavior geneticists – those who study the effects of genes on behavior – believe that differences in intelligence result from a combination of hereditary and environmental factors. The degree to which heredity contributes to intelligence is often expressed in terms of a heritability coefficient, a number on a scale from 0 to 1, such that a coefficient of 0 means that heredity has no influence on variation among people, whereas a coefficient of 1 means that heredity is the only influence on such variation. This coefficient can be applied to intelligence or to any other trait, such as height or weight.

It is important to remember that the coefficient indicates variation in measured intelligence. The heritability coefficient can tell us only about genetic effects that result in individual differences among people. It tells us nothing about genetic effects when there are no, or only trivial, differences. For example, both how tall you are and how many fingers you have at birth are in large part genetically preprogrammed. But we can use the coefficient of heritability only to assess genetic effects on height, where there are large individual differences. We cannot use the coefficient to understand number of fingers at birth because there is so little variation across people.

It is also important to realize that heritability tells us nothing about the modifiability of intelligence. A trait can be heritable and yet modifiable. For example, height is highly heritable, with a heritability coefficient greater than 0.9 in most populations. Yet heights of Europeans and North Americans increased by over 5 cm between 1920 and 1970. Consider, as another example, attributes of corn. Many attributes of corn, including height, are highly heritable. But if one batch of corn seeds were planted in the fertile fields of Iowa, and another similar batch were planted in the Mojave desert, the batch planted in Iowa undoubtedly would grow taller and thrive better, regardless of the heritability of the attributes of the corn. In this case, environment would largely determine how well the corn grew.

Current estimates of the heritability coefficient of intelligence are based almost exclusively on performance on standard tests of intelligence. The estimates can be no better than the tests and we have already seen that the tests define intelligence somewhat narrowly. How can we estimate the heritability of intelligence (at least that portion of it measured by the conventional tests)? Several methods have been used. The main ones are studies of separated identical twins, studies of identical versus fraternal twins, and studies of adopted children.

Many psychologists who have studied intelligence as measured by IQ believe the heritability of intelligence to be about 0.5 in children and somewhat higher in adults, for whom the early effects of the child-rearing environment have receded. However, there probably is no one coefficient of

heritability that applies to all populations under all circumstances. Indeed, changes in distributions of genes or in environments can change the estimates. Moreover, even if a trait shows a high heritability, we could not say that the trait cannot be developed. For example, as we have seen, the heritability of height is very high – about 0.9 – yet we know that over the past several generations, heights have been increasing. We can thus see how better environments can lead to growth, physical as well as intellectual.

At one time, it was believed that intelligence was fixed, and that we are stuck with whatever level of intelligence we have at birth. Today, many researchers believe that intelligence and the thinking skills associated with it are malleable, that these skills can be shaped and even increased through various kinds of interventions. For example, the Head Start program was initiated in the 1960s in the USA as a way of giving preschoolers an edge on intellectual abilities and accomplishments when they started school. Long-term follow-ups have indicated that by mid-adolescence, children who participated in the program were more than a grade ahead of matched controls who were not in the program. Children in the program also scored higher on a variety of tests of scholastic achievement, were less likely to need remedial attention, and were less likely to show behavioral problems. Although such measures are not truly measures of intelligence, they show strong positive correlations with intelligence tests. A number of newer programs have also shown some success in environments outside of the family home.

Perhaps the best evidence for the modifiability of intelligence comes from research by James Flynn. This research suggests that ever since record-keeping began early in the twentieth century, IQ (intelligence quotient) scores have been increasing roughly 9 points per generation (every 30 years). This result is sometimes referred to as the Flynn effect. From any point of view, this increase is large. No one knows exactly why such large increases have occurred, although the explanation must be environmental, because the period of time involved is too brief for genetic changes to have had an effect. If psychologists were able to understand the cause of the increase, they might be able to apply what they learned to increasing the intellectual skills of individuals within a given generation.

Altogether, evidence now indicates that environment, motivation, and training can profoundly affect intellectual skills. Heredity may set some kind of upper limit on how intelligent a person can become. However, we now know that for any attribute that is partly genetic, there is a reaction range – the broad limits within which a particular attribute can be expressed in various possible ways, given the inherited potential for expression of the attribute in a particular individual. Thus, each person's intelligence can be developed further within this broad range of potential intelligence. We have no reason to believe that people now reach the upper limits in the development of their intellectual skills. To the contrary, the evidence suggests that, although we cannot work miracles, we can do quite a bit to help people become more intelligent.

INDIVIDUAL AND GROUP DIFFERENCES

People in different cultural groups may have quite different ideas of how to behave intelligently. One of the more interesting cross-cultural studies of intelligence was performed by Michael Cole and his colleagues. These investigators asked adult members of the Kpelle tribe in Africa to sort a selection of terms. In Western culture, when adults are given a sorting task on an intelligence test, more intelligent people will typically sort hierarchically. For example, they may sort names of different kinds of fish together, and then the word 'fish' over that, with the name 'animal' over 'fish' and 'birds', and so on. Less intelligent Westerners will typically sort functionally. They might sort 'fish' with 'eat', for example, because we eat fish, or 'clothes' with 'wear', because we wear clothes. Members of the Kpelle tribe generally sorted functionally – even after investigators tried indirectly to encourage the Kpelle to sort hierarchically.

Finally, in desperation, one of the experimenters directly asked one of the Kpelle how a foolish person would do the task. When asked to sort in this way, the Kpelle had no trouble at all sorting hierarchically. He and the others had been able to sort this way all along; they just had not done so because they viewed it as foolish – and they probably considered the questioners rather unintelligent for asking such foolish questions. Why would they view functional sorting as intelligent? Simple. In ordinary life, we normally think functionally. When we think of a fish, we think of catching or eating it; when we think of clothes, we think of wearing them. However, in Western schooling, we learn what is expected of us on tests. The Kpelle did not have Western schooling and had not been exposed to intelligence testing. As a result, they solved the problems the way Western adults might do in their everyday lives, but not on an intelligence test. The Kpelle people are not the

only ones who might question Western understandings of intelligence. Work by Robert Serpell in Zambia shows that Zambians also have conceptions of intelligence quite different from those of North Americans, and research shows that there are many other such differences around the world.

A study by Seymour Sarason and John Doris provides a closer-to-home example regarding the effects of cultural differences on intelligence, particularly on intelligence tests. These researchers tracked the IQ scores of an immigrant population: Italian Americans. Less than a century ago, first-generation Italian American children showed a median IQ of 87, which is considered to be in the low average range, even when nonverbal measures were used and when so-called mainstream American attitudes were tested. Some social commentators and intelligence researchers of the day pointed to heredity and other nonenvironmental factors as the basis for the low IQs – much as they do today for other minority groups. For example, a leading researcher of the day, Henry Goddard, pronounced that 79 percent of immigrant Italians were 'feebleminded'. He also asserted that about 80 percent of immigrant Hungarians and Russians were similarly unendowed with intelligence. Goddard also asserted in 1917 that moral decadence was associated with this deficit in intelligence; he recommended that the intelligence tests he used be administered to all immigrants and that all those with low scores be selectively excluded from entering the United States.

Yet, Italian American students who take IQ tests today show slightly above-average IQs; other immigrant groups that Goddard denigrated have shown similar 'amazing' increases. Even the most fervent hereditarians would be unlikely to attribute such remarkable gains in so few generations to heredity. Cultural assimilation, including integrated education and adoption of American definitions of intelligence, seems a much more plausible explanation. At various times, and in various places, not all children have been encouraged to pursue an education.

Cultural and societal analyses of the concept of intelligence render it particularly important to consider carefully the meaning of group differences in measured IQ. For example, on average, African Americans score somewhat lower than Caucasians on conventional standardized tests of intelligence; but remember, Italian American scores used to be considerably lower than they are now. Scores of African Americans have been showing an increasing pattern over time, just as have scores for other groups. Available evidence to date suggests an environmental explanation for these group differences. Moreover, differences between groups in societal outcomes, such as likelihood of graduating from high school or going on welfare, cannot really be attributed simply to differences in IQ, as some people have tried to do, because after equating for IQ (and thus removing IQ as a source of group differences), African Americans are still considerably more likely than Caucasians to be born out of wedlock, born into poverty, and be underweight at birth. Group differences may thus originate from a number of factors, many of which change over time. The result is that group differences are not immutable: a group that scores, on average, lower than another group at one given time may score, on average, lower, the same, or even higher at another time.

An example of a change in the nature of group differences is that with regard to sex. Overall, males and females do about the same on cognitive ability tests, although differences have been noted on specific ability tests. Analyses of trends over time suggest that sex differences in scores on these cognitive ability tests have been shrinking over the years. Nevertheless, there do appear to be some differences that remain. In particular, males, on average, tend to score higher on tasks that require visual and spatial working memory, motor skills that are involved in aiming, and certain aspects of mathematical performance. Females tend to score higher on tasks that require rapid access to and use of phonological and semantic information in long-term memory, production and comprehension of complex prose, fine motor skills, and perceptual speed. These differences refer only to averages, and there are many individuals of one sex who do better than individuals of the other sex, regardless of the particular skill measured by a given test. In any case, these score differences are not easily interpretable. Claude Steele, for example, has found that when boys and girls matched for general mathematical abilities take very difficult mathematical tests, boys often do better. But when the two groups are told in advance that a particular test will show no difference, on average, scores of boys and girls converge, with girls' scores increasing and boys' scores actually decreasing.

Another group difference is between African Americans and whites. As mentioned earlier, African Americans tend to score lower than do white Americans on conventional tests of intelligence. The available evidence is largely consistent with an environmental explanation of this difference. For example, in one study, offspring of American servicemen born to German women during the

Allied occupation of Germany after World War II revealed no significant difference between IQs of children of African American versus white servicemen. This result suggests that given similar environments, the children of the two groups (African American and white) of servicemen performed equally on tests of intelligence. Another study found that children adopted by white families obtained higher IQ scores than did children adopted by African American families, again suggesting environmental factors contributing to the difference between the two groups. Another way of studying group differences has been through transracial adoption studies, in which white parents have adopted African American children. The results of these studies have been somewhat difficult to interpret, in that both white and African American children who were adopted in the study showed decreased IQ in a 10-year follow-up on their performance.

There are a number of mechanisms by which environmental factors such as poverty, undernutrition, and illness might affect intelligence. One mechanism is through resources. Children who are poor often do not have the resources in the home and school that children from more affluent environments have. Another mechanism is through attention to, and concentration on, the skills being taught in school. Children who are undernourished or ill may find it hard to concentrate in school, so they may profit less from the instruction they receive. A third mechanism is the system of rewards in the environment. Children who grow up in economically deprived environments may note that the individuals who are most rewarded are not those who do well in school, but rather those who find ways of earning the money they need to survive, whatever those ways may be. It is unlikely that there is any one mechanism that fully explains the effects of these various variables. It is also important to realize that whatever these mechanisms are, they can start *in utero*, not just after birth. For example, fetal alcohol syndrome results in reduced IQ and has its initial effects prenatally, before the child even enters the world outside the mother's womb. Thus, couples can help maximize their future children's intelligence by giving those children good prenatal care.

Further Reading

Carroll JB (1993) *Human Cognitive Abilities: A Survey of Factor-analytic Studies*. New York, NY: Cambridge University Press.

Gardner H (1983) *Frames of Mind: The Theory of Multiple Intelligences*. New York, NY: Basic Books.

Gardner H (1993) *Multiple Intelligences: The Theory in Practice*. New York, NY: Basic Books.

Gardner H (1999) *Reframing Intelligence*. New York, NY: Basic Books.

Goleman D (1995) *Emotional Intelligence*. New York, NY: Bantam.

Jensen AR (1998) *The g Factor*. Westport, CT: Praeger-Greenwood.

Neisser U (ed.) (1998) *The Rising Curve*. Washington, DC: American Psychological Association.

Sternberg RJ (ed.) (1994) *Encyclopedia of Intelligence* (2 vols). New York, NY: Macmillan.

Sternberg RJ (1997) *Successful Intelligence*. New York, NY: Plume.

Sternberg RJ (ed.) (2000) *Handbook of Intelligence*. New York, NY: Cambridge University Press.

Yang S-Y and Sternberg RJ (1997) Taiwanese Chinese people's conceptions of intelligence. *Intelligence* **25**: 21–36.

Intermodal Perception, Development of

Intermediate article

Lorraine E Bahrick, Florida International University, Miami, Florida, USA

CONTENTS

Intermodal perception is the perception of unitary objects and events through spatially and temporally coordinated stimulation from multiple sense modalities. Research suggests that the senses are united in early infancy, fostering the rapid development of intermodal perception.

INTRODUCTION

Intermodal perception is the perception of an object or event that makes information available to two or more sensory systems simultaneously. Most objects and events are multimodal in that they can be experienced through multiple sense modalities. For example, a person talking, a fire, or a bouncing ball can all be seen as well as heard and felt. Intermodal perception is thus one of the most fundamental human capabilities and forms the basis for most of what we perceive, learn, and remember. One of the questions developmental psychologists have asked is how and when the child comes to perceive multimodal events as single, unitary events, in the way adults do. For example, without prior experience with objects and events, how does the infant learn that certain patterns of auditory and visual stimulation, such as the sight of the mother's face and the sound of her voice, belong together and constitute a unitary event, whereas other concurrent patterns of sensory stimulation are unrelated? How does the child acquire intermodal knowledge such that the sound of footsteps in the hallway will elicit the expectation of seeing a person in the doorway?

INTEGRATION VERSUS DIFFERENTIATION

Researchers have discovered that intermodal perception develops rapidly during infancy. Infants are intrinsically motivated to pick up new information. Some researchers (e.g., Piaget, 1954) have characterized the development of intermodal perception as a process of integration. According to this view, the senses are separate at birth and the infant must gradually learn to put together or 'integrate' stimulation from the different sense modalities in order to perceive a unitary multimodal event. This 'integration' may occur through associating concurrent information across different modalities. Thus, before integration takes place, infants would perceive only unrelated streams of light, sound, or tactile impressions. A contrasting position is the 'differentiation' view of development (e.g., Gibson, 1969). According to this view, the senses are unified at birth, and perceptual development is characterized by a progressive process of 'differentiation' of increasingly finer levels of stimulation. Thus, in early infancy, information from the different senses must be gradually separated from the global, undifferentiated perceptual array. From this perspective, intermodal perception of some kinds of information is possible at birth and infants continue to show perceptual learning of more complex multimodal relations throughout infancy and early childhood.

Recent evidence has demonstrated that young infants are adept at perceiving a wide array of multimodal objects and events and they do so by detecting information that is common, or invariant, across the senses. This body of research has thus weakened the integration position, especially when intermodal abilities are discovered in very young infants who have had little opportunity to learn to associate or integrate information across the senses. Much infant research has provided support for the differentiation view, particularly the large body of research on young infants' detection of 'amodal

invariants', suggesting that the senses are unified in early infancy.

AMODAL INVARIANT RELATIONS

Amodal information is information that is not specific to a particular sense modality, but is completely redundant or invariant across two or more senses. For example, the sights and sounds of hands clapping share a synchrony relation, a common tempo of action, and a common rhythm. The same rhythm and tempo can be detected by watching or hearing the hands clap. Thus, synchrony, rhythm, and tempo are 'amodal invariant relations' in that this information can be perceived across different sense modalities. Most information that is amodal characterizes how events are distributed in space and time, two of the most fundamental dimensions of our experience. According to the differentiation view, detection of amodal relations focuses attention on meaningful, unitary events and buffers infants from making incongruent, inappropriate associations (e.g., Bahrick and Pickens, 1994). For example, if the infant detects synchrony, shared rhythm, and common tempo between the sight of a person's moving face and the sound of the person's voice, the infant would necessarily be attending to a unitary event: the person talking. In this way, unrelated sounds and movements would not be merged with the event. In support of the differentiation view, research has found that young infants detect a wide array of amodal invariant relations in multimodal events.

In contrast to amodal relations, information can also be nonredundant and arbitrarily related across the sense modalities (e.g., speech sounds and the objects they refer to; particular faces and voices). Information such as color, pattern, timbre, or pitch is 'modality-specific' and can be perceived only through a single sense modality. Research suggests that infants detect amodal relations (such as temporal synchrony) developmentally prior to arbitrary relations, and detection of amodal relations can then guide and constrain learning about arbitrary relations.

AUDITORY–VISUAL CORRESPONDENCE

To assess intermodal perception of auditory–visual relations, sometimes an intermodal preference method is used. In this method, infants view two filmed events simultaneously, along with the soundtrack to one of them coming from a centralized speaker. It is expected that if the infant detects the intermodal relations, he or she will look longer at the film that belongs with the soundtrack played.

Research using these and similar procedures has demonstrated that young infants display a wide array of intersensory abilities in the area of audio-visual perception (see Gibson and Pick, 2000; Lewkowicz, 2000; Lewkowicz and Lickliter, 1994). Neonates turn their eyes in the direction of a sound, demonstrating a basic coordination of audio-visual space. In the first month of life, infants detect the temporal synchrony between sights and sounds of an object striking a surface, and the spatial location common to the sights and sounds of a moving object. By three to five months, infants can match films and soundtracks of moving objects on the basis of their substance (rigid versus elastic) or their composition (single versus multiple objects), as well as the rhythm and tempo of their impact sounds. Further, by four to six months, infants can match faces and voices on the basis of affective expressions, including happy, sad, neutral, and angry (Walker-Andrews, 1997). They can also match faces and voices on the basis of age (adults versus children) and gender of speaker. All these relations are amodal and invariant across vision and audition.

BIMODAL PERCEPTION OF SPEECH

The perception of speech, an auditory–visual event, has traditionally been studied as a unimodal, auditory event. However, speech is produced by a speaker who can be heard and seen, and who typically uses gesture as well. It turns out that the multimodal nature of speech is salient to infants and facilitates its perception (e.g., Meltzoff and Kuhl, 1994). By the age of at least two months, infants are sensitive to voice–lip synchrony during speech. By four months, infants are able to detect the voice–lip correspondence between speech sounds such as '*a*' and '*i*'. When one of these speech sounds is played in synchrony with two films side by side of a speaker's face intoning each sound, infants look more to the face with the matching lip movements. The McGurk effect, an auditory–visual illusion, also illustrates how infants and adults merge information for speech across the senses. When we view the face of a person speaking one speech sound such as '*ga*', while hearing a different speech sound, for example '*ba*', we perceive another sound, '*da*', a blend between the two. Infants show evidence of this effect in the first half-year of life. Visual input appears to have significant auditory consequences.

Amodal information during speech is also important for learning the arbitrary relation between

speech sounds and the objects they denote (Gogate *et al*., 2001). By 14 months of age, infants are able to learn to pair a speech sound and an object during a brief familiarization. However, if amodal synchrony unites the sounds and object movements, for example in showing and naming the object simultaneously, infants can learn the relation as early as seven months of age. Adults even match their teaching style to the infant's needs. They use more synchronous movement with labeling, to highlight object–sound relations, when they are first teaching the names of new objects to their young infants. Their use of synchrony decreases as infants become more linguistically competent. Further evidence for the importance of visual information for perceiving speech lies in the success of teaching speech to deaf individuals using a visual depiction of the lip and tongue movements involved in different speech sounds.

VISUAL–TACTILE CORRESPONDENCE

Amodal invariant relations also unite perception across vision and touch. Information for shape, texture, substance, and size are invariant across visual and tactile stimulation (Rose and Ruff, 1987). One method for investigating perception of visual–tactile correspondence is the cross-modal transfer method. An object is presented to one sense modality alone, and a preference test is then given in another sense modality to determine whether the information transfers across modalities. Using this method, research has shown that, by the age of one month, infants can perceive the correspondence between an object they experienced tactually (on the back or a pacifier) and a visual replica of the object. Infants looked more to the object of the shape and texture that they had previously experienced orally. Infants are also able to transfer information about the substance of an object (rigid versus deforming) across touch and sight.

Evidence also shows that infants can transfer information obtained through manual exploration to vision, and this develops across the first year. One factor determining the extent to which manual information is perceived is whether exploration is active or passive. Tactile exploration develops over the first year. Young infants tend to grasp objects, whereas older infants become more adept at obtaining tactile feedback by moving their hand relative to the object's surface. By four months, infants can perceive whether two parts of objects are connected or separate, by the type of motion they produce during haptic exploration. By six months, infants can recognize the shape of an object visually that they have manually explored, as long as exploration is active.

VISUAL–MOTOR CORRESPONDENCE AND THE SELF

Infants are also able to perceive information specifying the self by detecting amodal invariant relations (Rochat, 1995). Even in the first weeks of life, infants can imitate facial expressions. In order to do this, they must relate the visual appearance of the adult's facial expression with their own production of the expression. This is most probably guided by proprioception: proprioception is information about self-movement based on feedback from the muscles, joints, and vestibular system. Facial imitation reveals evidence of early intermodal coordination between visual information and motor behavior, and this coordination continues to develop over the first year (Meltzoff and Moore, 1995).

Infants also show evidence of self-perception by detecting amodal invariant relations in a procedure where they view their own body moving live in a video display (Bahrick, 1995). By three to five months, infants can distinguish between a live video of their own legs kicking and a video of another infant's legs kicking, a pre-recorded video of their own legs, or a spatially incongruent video of their own legs. They do this by detecting the amodal temporal synchrony and spatial relations common to the visual display of their motion and the proprioceptive experience of their motion.

Infants also demonstrate a phenomenon called 'visually guided reaching', which develops rapidly during the first year. That is, they show continuous adjustments in their reaching and manual behavior as a function of visual input about the size, shape, and position of objects. Infants are even able to contact a moving object by aiming their reach ahead of the object and taking into account the speed and direction of its movement as well as that of their arm motion. Later, infants show an ability to adapt their crawling and exploratory behavior as a function of visual information about the slant and solidity of the surface. These examples illustrate a close coupling between vision and motor behavior and an understanding of self in relation to objects (Gibson and Pick, 2000).

NEURAL BASES OF INTERMODAL PERCEPTION

Behavioral research on the rapid development of intermodal perception during infancy is consistent

with research findings from the neurosciences. Some areas of the brain (cortex, superior colliculus) contain 'multimodal neurons' that respond to inputs from multiple sense modalities, providing a biological basis for the early integration of the senses (e.g., Stein and Meredith, 1993). Further, some cells of the superior colliculus (devoted to attention and orienting) are activated much more by simultaneous auditory and visual inputs than by either auditory or visual information alone. Other cells, however, are modality-specific but can have receptive fields that are spatially coordinated across the sense modalities as a result of experience with multimodal events. Thus, auditory and visual input from the same spatial location can be related. Neurophysiological findings suggest that if input to one modality is somehow modified, the receptive field of cells in the superior colliculus can compensate and realign with those of the other modality to maintain a coherent multimodal spatial mapping. The early plasticity of the brain, its sensitivity to multimodal inputs, and its reliance on experience in the multimodal world to guide neuronal development, appears well tailored to the behavioral findings of the early development of intermodal perception.

CONCLUSION

Infants demonstrate a diverse array of intermodal abilities. These abilities illustrate the close connection between the senses during early development and the rapid growth in intersensory abilities across the first year of life. Development appears to be guided by the detection of amodal, invariant relations, and this promotes accurate and unitary perception of multimodal events.

References

Bahrick LE (1995) Intermodal origins of self-perception. In: Rochat P (ed.) *The Self in Infancy: Theory and Research* pp. 349–373. New York, NY: Elsevier.

Bahrick LE and Pickens JN (1994) Amodal relations: the basis for intermodal perception and learning. In: Lewkowicz D and Lickliter R (eds) *The Development of Intersensory Perception: Comparative Perspectives*, pp. 205–233. Hillsdale, NJ: Lawrence Erlbaum.

Gibson EJ (1969) *Principles of Perceptual Learning and Development*. New York: Appleton-Century-Crofts.

Gibson EJ and Pick AD (2000) *An Ecological Approach to Perceptual Learning and Development*. New York, NY: Oxford University Press.

Gogate LJ, Walker-Andrews AS and Bahrick LE (2001) The intersensory origins of word comprehension: an ecological-dynamic systems view. *Developmental Science* **4**: 1–37.

Lewkowicz DJ (2000) The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological Bulletin* **126**: 281–308.

Lewkowicz DJ and Lickliter R (eds) (1994) *The Development of Intersensory Perception: Comparative Perspectives*. Hillsdale, NJ: Lawrence Erlbaum.

Meltzoff AN and Kuhl PK (1994) Faces and speech: intermodal processing of biologically relevant signals in infants and adults. In: Lewkowicz D and Lickliter R (eds) *The Development of Intersensory Perception: Comparative Perspectives*, pp. 335–369. Hillsdale, NJ: Lawrence Erlbaum.

Meltzoff AN and Moore KM (1995) A theory of the role of imitation in the emergence of self. In: Rochat P (ed.) *The Self in Infancy: Theory and Research*, pp. 73–93. New York, NY: Elsevier.

Piaget J (1954) *The Construction of Reality in the Child*. New York, NY: Basic Books.

Rochat P (ed.) (1995) *The Self in Infancy: Theory and Research*. New York, NY: Elsevier.

Rose SA and Ruff HA (1987) Cross-modal abilities in human infants. In: Osofsky J (ed.) *Handbook of Infant Development*, 2nd edn, pp. 338–362. New York, NY: John Wiley.

Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.

Walker-Andrews A (1997) Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychological Bulletin* **121**: 437–456.

Further Reading

Kellman PJ and Arterberry ME (1998) *The Cradle of Knowledge: Development of Perception in Infancy*. Cambridge, MA: MIT Press.

Lickliter R and Bahrick LE (2000) The development of infant intersensory perception: advantages of a comparative convergent-operations approach. *Psychological Bulletin* **126**: 260–280.

Masarro DW (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

Thelen E and Smith LB (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Intuitive Physics

Intermediate article

Dennis R Proffitt, University of Virginia, Charlottesville, Virginia, USA
Mary K Kaiser, NASA Ames Research Center, Moffett Field, California, USA

CONTENTS

Intuitive physics describes the common-sense beliefs that people hold about how the world works. The study of intuitive physics has focused on people's beliefs about biology and physical mechanics.

INTRODUCTION

Intuitions about the natural world are based upon implicit, common-sense beliefs that people hold about the way the world works. It is useful to distinguish between individual and culturally shared beliefs. Intuitive (or naive) beliefs develop within each individual, in the absence of formal training or instruction. Folk knowledge is a belief system reflecting a culture's cumulative understanding of a class of physical phenomena. These understandings may consist of simple compilations of observations, or they may evoke systematic belief systems. In some regards, folk knowledge can resemble aspects of scientific inquiry (Kuhn, 1970), in which practitioners share common concepts and techniques. However, folk knowledge lacks the paradigmatic rigor of science, and its core belief structure is often constrained or defined by a society's religious tenets.

Like intuitive beliefs, folk knowledge tends to explain only a subset of the relevant phenomena in a physical domain. Models tend to be simplistic and are often animistic. Observations that are inconsistent with the prevailing model are either discounted, or explained by *ad hoc* elaborations.

FOLK BIOLOGY: KEEPING THE BODY'S SYSTEMS IN BALANCE AND CLASSIFYING NATURE

Biology is one of the most extensively developed domains of folk knowledge. People are naturally interested in understanding the workings of their own bodies, and in organizing the other living things in their environment into meaningful groups and clusters. Hence, two of the primary foci of folk biology are medicine and taxonomy.

Folk Medicine

Of necessity, virtually every culture has individuals trained to serve as healers. Even today, many societies reject the therapies of modern Western medicine in favor of traditional methods of treatment. Typically, such treatment involves herbal medicines. The therapeutic efficacy of many of these remedies has been scientifically validated (Fontanarosa, 2000). However, folk medicine also involves rituals and totems. In some cultures, these are used to appeal to deities for restorative intervention. Other cultures view these exercises as a means to restore the body to a healthy state of balance. Thus, ancient Greeks and medieval physiologists attributed illness to an excess or deficiency of one of the body's four humors (blood, phlegm, choler, and bile). Similarly, Hispanic folk medicine holds that disease is caused by an imbalance between hot and cold principles (Buhner, 1996).

The practice of folk medicine can be quite sophisticated, with hierarchies of healers and specialists who advise patients on preventative as well as curative measures. The 'four-humors' model emphasized a balanced diet for health maintenance, whereas the 'hot/cold' model stresses the avoidance of exposure to extreme temperature. Should illness occur, treatment is given to restore the body's natural balance (e.g. 'cold' diseases such as pneumonia or colic are treated with 'hot' remedies).

Folk Taxonomy

The manner in which cultures categorize animals and plants reflects their societal needs and priorities. Still, there are some notable commonalities in the taxonomies derived (Atran, 1995). First, in

establishing the defining criteria for a species, certain features are attended to while others are ignored. Usually, the features deemed relevant appeal to what Atran terms the 'underlying essence' of the species. Thus, category structures are teleological, and irrelevant surface features (such as fur color) are ignored.

Second, as Darwin noted (1859) and contemporary researchers have confirmed (Hays, 1983; Brown, 1984), virtually every culture's folk taxonomy is hierarchical in nature: 'groups under groups' (Darwin, 1859, p. 431). As with their scientific counterpart, the categories at each level of these folk taxonomies are exhaustive (every plant or animal can be assigned to a class) and exclusive (no plant or animal belongs to more than one class). Folk taxonomies differ both from one another and from scientific taxonomies in their hierarchical groupings (which species of animals are seen as related) and 'terminal contrast' (the lowest level of grouping recognized) (Lévi-Strauss, 1966). Thus, one folk taxonomy may have a terminal contrast of 'bat' while another groups bats into a terminal contrast of 'flying animals'. Modern science, of course, recognizes 'bat' as corresponding to diverse families, genera, and species in the order Chiroptera.

FOLK NOTIONS OF THE PHYSICAL WORLD: GENESIS MYTHS

The most pervasive folk notions of physics deal with a fundamental cosmological issue: the origin of the earth and universe. The realm of creation accounts is quite diverse. Many are mythic in nature, positing a divine agent (or agents) as creator. In some cultures, the model is closely woven with the society's religion, and there is little effort to reconcile the model with physical evidence. One example of how theology constrained cosmological models can be found in the debate between geocentric/heliocentric models of the solar system. Despite the difficulty of reconciling the geocentric model with astronomical data, religious forces retarded the acceptance of the more parsimonious heliocentric view (Casper and Noer, 1972).

The tendency to maintain a belief in the face of contrary evidence can be observed within individuals as well as cultures. This bias is quite evident when one examines people's intuitive physics.

INTUITIVE PHYSICS: SURPRISING PHENOMENA

The literature on intuitive physics is filled with accounts of people's erroneous predictions about basic mechanical systems. For example, in the problem depicted in Figure 1, people are asked to indicate, by drawing a line, what the orientation of a liquid's surface would be given that it intersects the glass at the spot indicated (Piaget and Inhelder, 1948/1956). About 40 percent of the adult population gets this problem wrong by drawing lines that are inclined more than 5 degrees away from the horizontal in the same direction as the tilt of the container (McAfee and Proffitt, 1991). This is odd and surprising given that people observe liquids in containers every day. Why should people be systematically wrong about something with which they are so familiar? Systematic error in the face of familiarity is the hallmark of the phenomena that define the intuitive physics literature.

One could pick problems in physics that are so complex that almost no one could predict the systems' behavior. Predicting the behavior of spinning-tops is an example (Proffitt and Gilden, 1989). One could also pick physical systems that are so simple that everyone would answer correctly, such as predicting the path of a ball exiting a straight tube. What has captured the imagination of researchers in this field are those problems that garner systematic error as opposed to on-the-spot guessing or correct responses.

Many studies have looked at people's ability to predict trajectories in simple dynamical systems (Champagne *et al.*, 1980; Clement, 1982; Kaiser

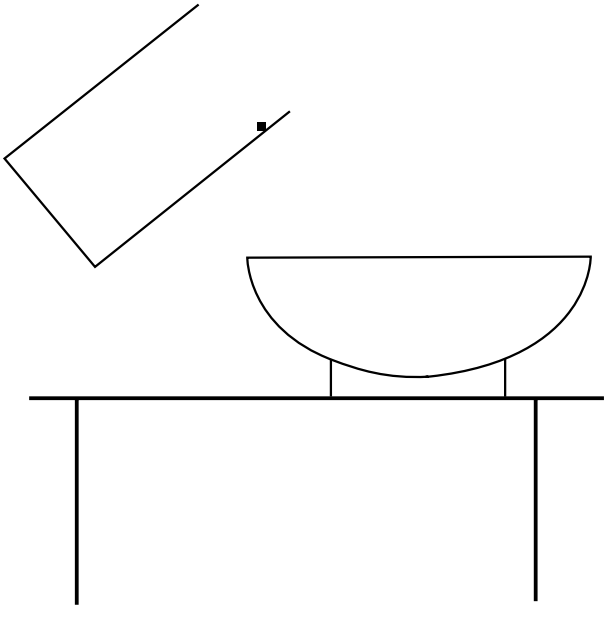


Figure 1. The water-level problem. The participant's task is to draw the surface of water in the tilted glass so that it touches the point on the right side of the glass.

et al., 1986; McCloskey, 1983; McCloskey *et al*., 1980; Shanon, 1976). One of the most extensively investigated problems is depicted in Figure 2 (McCloskey *et al*., 1980). In this problem, a C-shaped tube is shown lying flat on a horizontal surface. Participants are asked to imagine that a ball is made to roll through the tube and then they are required to draw the trajectory that it would take upon exiting the tube. A correct drawing would show the ball following a straight trajectory tangential to the tube's curvature at the point of exit. It has been found that about 40 percent of college students get this problem wrong and draw a trajectory in which the ball continues to curve after exiting the tube. What is interesting about this problem is not simply that people are wrong, but rather that those who make errors produce very similar predictions.

Among the trajectory problems, predicting how an object will fall has been found to elicit systematic errors. Figure 3 shows the pendulum problem. Participants are asked to predict the path that the pendulum bob will take following a break of its tether at one of two phases in its swing. The figure shows one correct trajectory along with some of the common errors. Another problem is that shown in Figure 4, in which a moving carrier drops an object. In this situation, many people predict that the object will fall straight down as opposed to following the parabolic curve that combines its forward movement with the downward acceleration of gravity. Interestingly, if a ball is shown rolling

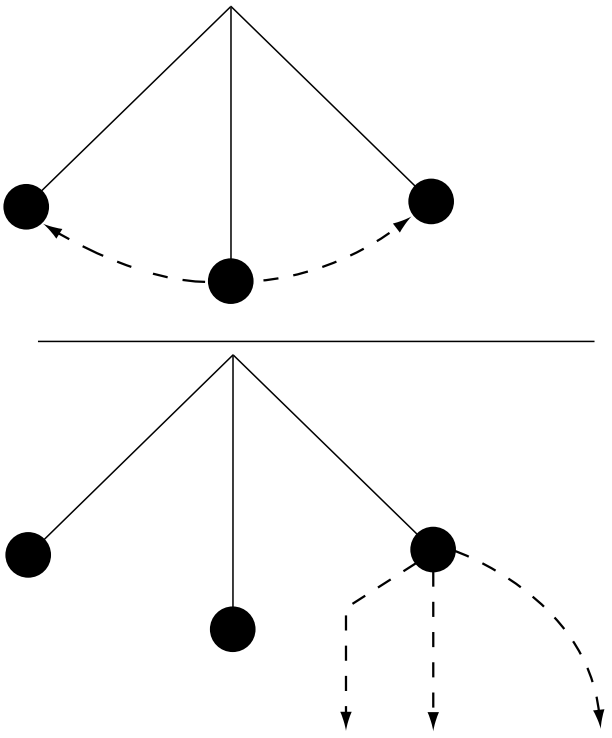


Figure 3. The pendulum problem. The top panel depicts a swinging pendulum. Participants are asked to draw the trajectory that the bob would take if its tether broke at either the apex or the nadir of its swing. The bottom panel shows the correct straight-down path for the apex condition along with two incorrect paths that are frequently drawn.

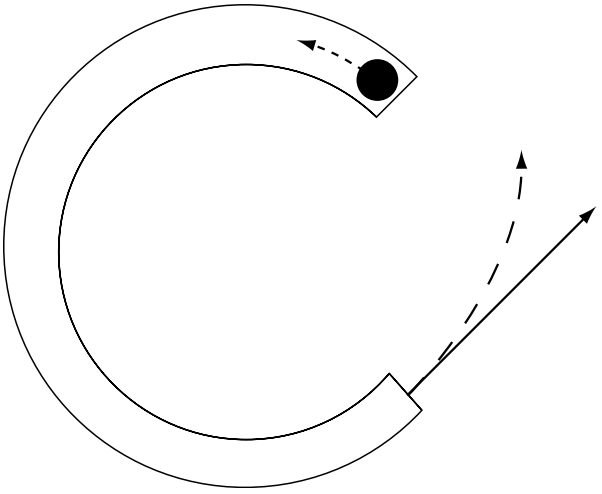


Figure 2. The C-shaped tube problem. The curved tube is to be construed as lying flat on a table and participants are to draw the trajectory of the ball after it exits the tube. (In the problem, the trajectories are not shown.) Although the correct path is straight (solid line), many people draw paths that continue to curve (dotted line).

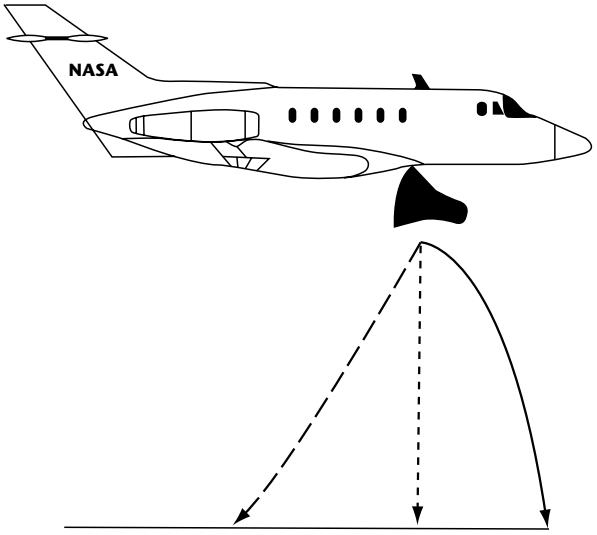


Figure 4. The object dropped from a moving carrier problem. The figure depicts a space capsule being dropped by an airplane. Participants are asked to draw the capsule's falling trajectory. The correct trajectory is a forward parabolic path (solid line). The incorrect paths shown with dotted lines are often drawn.

off a table as in Figure 5, then adults are far more likely to take into account its forward motion. People are also prone to err when making judgments about falling rate. Shanon (1976) found that many people report that objects will fall at a constant velocity proportional to their mass.

Hecht and Bertamini (2000) found that misconceptions about projectile motions are not limited to predictions about the shape of their trajectories. In their studies, people were asked to indicate where in the trajectory of a thrown ball the velocity was greatest. Many participants judged the speed of the ball to be greatest some distance after it was released by the thrower, implying that the ball continued to accelerate after its release.

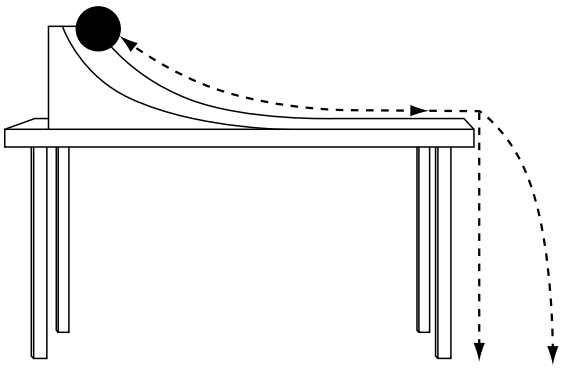


Figure 5. The ball rolling off a table problem. Participants are asked to predict the path of a ball that rolls off a table. Most adults correctly draw parabolic paths, whereas young children tend to draw straight-down paths.

People's intuitions about the dynamics of rotation seem to be especially error-prone. Proffitt *et al*. (1990) asked people to predict whether size (radius), mass, or shape (moment of inertia) would affect the time that it would take a wheel to roll down a ramp (Figure 6). Among these variables, they found that people were most certain that shape would not be an effective variable; in fact, it is the only factor that matters. 'Moment of inertia' describes the distribution of an object's mass relative to its axis of rotation. In the case of the rim-like shape in Figure 6(c), a greater proportion of the wheel's mass is located away from the axis as compared to the standard. Assessments of university and high school physics teachers showed only minimal improvement so long as these individuals were forced to answer quickly and were prohibited from writing down the relevant equations.

Once it is recognized that people predict motions that defy the laws of physics, a natural question follows as to what these people would experience should they see animated events that conformed to their predictions. Would people who erroneously predict that a ball exiting a C-shaped tube continues on a curved path be surprised or accepting when viewing this event? Studies have shown that some problems are penetrated by perception but others are not. When shown animations of C-shaped tube, pendulum, and dropped object problems, people who make erroneous predictions on paper-and-pencil tests view their predicted paths

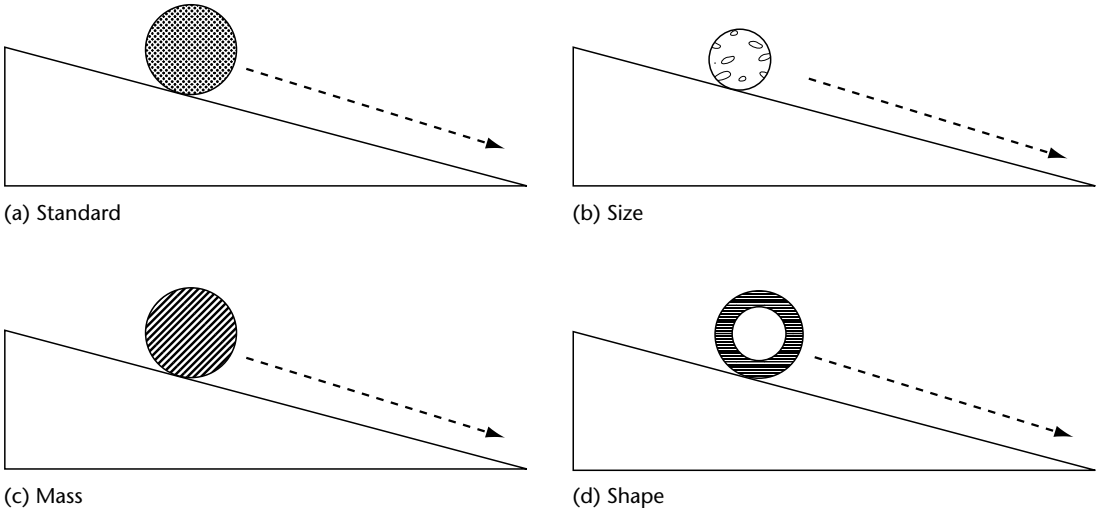


Figure 6. The wheels rolling down a ramp problem. Relative to a standard wheel, participants are asked to judge which of two wheels will arrive at the bottom of the ramp first. The comparison wheels are identical to the standard except for one of the following variables: size (radius), weight (mass), or shape (moment of inertia).

as being anomalous and correctly choose the appropriate trajectories as being natural (Kaiser *et al*., 1985; Kaiser *et al*., 1992). However, animations of the water-level problem shown in Figure 1 (Howard, 1978; McAfee and Proffitt, 1991) and of the moment of inertia problem (Proffitt *et al*., 1990) do not evoke more accurate responses than their static paper-and-pencil versions.

HEURISTICS IN COMMON-SENSE REASONING

The list of surprising phenomena in the intuitive physics literature is long, varied, and could probably be extended indefinitely. What is needed, of course, is a theoretical account capable of predicting when and how people are likely to make accurate or biased judgments and when they are likely to be totally befuddled. The classes of explanation fall into at least three non-mutually exclusive categories.

Implicit Theories

It has been proposed that people develop implicit theories about how the world works. These theories then become evident in the heuristics that people use when attempting to solve intuitive physics problems (Chi and Slotta, 1993; Ranney, 1994; Smith and Casati, 1994). Shanon (1976) and, in his earlier writings, diSessa (1982) proposed that people's reasoning reflected an Aristotelian conception of dynamics. McCloskey (1983) suggested that people's judgments were more consistent with medieval impetus theory. By this account, objects are put into motion by an impetus that dissipates over time. Since both linear and curvilinear impetus persist after their initial application, impetus theory nicely accounts for the common curvilinear predictions made on such problems as the C-shaped tube.

There are a number of problems with implicit theory accounts. First, people are often inconsistent in their judgments from one problem to another (Cooke and Breedin, 1994; diSessa, 1982; Kaiser *et al*., 1992; Ranney and Thagard, 1988; Shanon, 1976). This could mean that people's judgments are not grounded in implicit theory or that they possess multiple theories. A second problem is that people are highly influenced by the surface structure of the problem. For example, no one predicts that water exiting a C-shaped hose will continue to curve upon exit (Kaiser *et al*., 1986). Finally, the implicit theories that have been proposed have limited scope in accounting for the wide variety of intuitive physics problems. Impetus theory, for example, explains why curvilinear trajectories are predicted but is mute with respect to the water-level problem or the belief that a thrown ball will continue to accelerate after it leaves the thrower's hand.

Perceptual Biases

What makes erroneous judgments about nature surprising is that people's intuitions about the way that the world works ought to have been informed by manifest regularities in everyday experience. A possible explanation for this conundrum is that judgmental biases reflect biases that are inherent in perception itself. In the case of the object being dropped by a moving carrier, Kaiser *et al*. (1992) proposed that the bias to predict a straight-down motion is due to the perceptual bias to notice and attend to relative motions. Relative to the carrier, the object does indeed move straight down. A similar frame of reference argument has been applied to the water-level problem (Hecht and Proffitt, 1995; McAfee and Proffitt, 1991). If the water level's orientation is related to its container instead of the environment, then a perceptual bias is evoked in which the horizontal is misperceived to be inclined slightly in the direction of the container's tilt. As with implicit theories, perceptual bias accounts suffer from a lack of scope. They account for a few phenomena well but not the others.

Problem Complexity

As previously noted, physics problems range in their difficulty from laughably easy to stupefyingly difficult. As a first approximation towards defining problem complexity, a distinction has been made between particle (easy) and extended body (difficult) motions (Proffitt and Gilden, 1989). Particle motions are those that can be described adequately by treating the object as if it were a point particle located at the object's center of mass. Freefall is a particle motion if air resistance is ignored. Extended body motions make relevant other object properties such as shape. A wheel rolling down a ramp is an extended body motion because its shape (moment of inertia) is of dynamic relevance. This distinction predicts that particle motions will be easier to construe than extended body ones, and this prediction is fairly well borne out by research findings (Proffitt and Gilden, 1989; Proffitt *et al*., 1990). For example, all rotations are extended

body motions – a point particle cannot rotate – and as noted above, the dynamics of rotations are especially difficult to appreciate.

It would be desirable to have a general account that could predict performance across the whole range of intuitive physics problems. That no account currently meets this goal may reflect the current state of theory development, or the plurality of means by which people attempt to solve the different sorts of problems that comprise the intuitive physics literature. The latter possibility is probably more likely.

References

Atran S (1995) Classifying nature across cultures. In: Smith EE and Osherson DN (eds) *An Invitation to Cognitive Science*, vol. 3. Cambridge, MA: MIT Press.

Brown C (1984) *Language and Living Things: Uniformities in Folk Classification and Naming*. New Brunswick, NJ: Rutgers University Press.

Buhner SH (1996) *Sacred Plant Medicine: Explorations in the Practice of Indigenous Herbalism*. New York, NY: Rinehart.

Casper BM and Noer RJ (1972) *Revolutions in Physics*. New York, NY: WW Norton.

Champagne AB, Klopher LE and Anderson JH (1980) Factors influencing the learning of classical mechanics. *American Journal of Physics* **48**: 1074–1079.

Chi MTH and Slotta JD (1993) The ontological coherence of intuitive physics. *Cognition and Instruction* **10**: 249–260.

Clement J (1982) Students' preconceptions in introductory mechanics. *American Journal of Physics* **50**: 66–71.

Cooke NJ and Breedin SD (1994) Constructing naive theories of motion on the fly. *Memory & Cognition* **22**: 474–493.

Darwin C (1859) *On the Origin of the Species by Natural Selection*. London, UK: Murray.

diSessa A (1982) Unlearning Aristotelian physics: a study of knowledge-based learning. *Cognitive Science* **6**: 37–75.

Fontanarosa PB (ed.) (2000) *Alternative Medicine: An Objective Assessment*. Washington, DC: AMA Press.

Hays T (1983) Ndumba folk biology and the general principles of ethnobotanical classification and nomenclature. *American Anthropologist* **85**: 489–507.

Hecht H and Bertamini M (2000) Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance* **26**: 730–746.

Hecht H and Proffitt DR (1995) The price of expertise: effects of experience on the water-level task. *Psychological Science* **6**: 90–95.

Howard I (1978) Recognition and knowledge of the water-level problem. *Perception* **7**: 151–160.

Kaiser MK, Jonides J and Alexander J (1986) Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition* **14**: 308–312.

Kaiser MK, Proffitt DR and Anderson KA (1985) Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Human Perception and Performance* **11**: 795–803.

Kaiser MK, Proffitt DR, Whelan SM and Hecht H (1992) Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 384–393.

Kuhn TS (1970) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lévi-Strauss C (1966) *The Savage Mind*. Chicago, IL: University of Chicago Press.

McAfee EA and Proffitt DR (1991) Understanding the surface orientation of liquids. *Cognitive Psychology* **23**: 669–690.

McCloskey M (1983) Intuitive physics. *Scientific American* **248**: 122–130.

McCloskey M, Caramazza A and Green B (1980) Curvilinear motion in the absence of external forces: naive beliefs about the motion of objects. *Science* **210**: 1139–1141.

Piaget J and Inhelder B (1956) *The Child's Conception of Space*. London, UK: Routledge & Kegan Paul. (Original work published 1948.)

Proffitt DR and Gilden DL (1989) Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance* **15**: 384–393.

Proffitt DR, Kaiser MK and Whelan SM (1990) Understanding wheel dynamics. *Cognitive Psychology* **22**: 342–373.

Ranney M (1994) Relative consistency and subjects' 'theories' in domains such as naive physics: common research difficulties illustrated by Cooke and Breedin. *Memory & Cognition* **22**: 494–502.

Ranney M and Thagard P (1988) Explanatory coherence and belief revision in naive physics. In: *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pp. 11–117. Hillsdale, NJ: Lawrence Erlbaum.

Shanon B (1976) Aristotelianism, Newtonianism, and the physics of the layman. *Perception* **5**: 241–243.

Smith B and Casati R (1994) Naive physics. *Philosophical Psychology* **7**: 227–247.

Further Reading

Caramazza A, McCloskey M and Green B (1981) Naive beliefs in 'sophisticated' subjects: misconceptions about trajectories of objects. *Cognition* **9**: 117–123.

diSessa A (1993) Toward an epistemology of physics. *Cognition and Instruction* **10**: 105–225.

Gilden DL (1991) On the origins of dynamical awareness. *Psychological Review* **98**: 554–568.

Kaiser MK, Proffitt DR and McCloskey M (1985) The development of beliefs about falling objects. *Perception & Psychophysics* **38**: 533–539.

Larkin JH (1983) The role of problem representation in physics. In: Gentner D and Stevens AL (eds) *Mental Models*, pp. 75–98. Hillsdale, NJ: Erlbaum.

McCloskey M and Kohl D (1983) Naive physics: the curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory and Cognition* **9**: 146–156.

Mann J (2000) *Murder, Magic and Medicine*, (2nd edn). Oxford, UK: Oxford University Press.

Moss KK (1999) *Southern Folk Medicine: 1750–1820*. Columbia, SC: University of South Carolina Press.

Spelke ES, Breinlinger K, Macomber J and Jacobson K (1992) Origins of knowledge. *Psychological Review* **99**: 605–632.

Judgment

Introductory article

Jonathan Baron, University of Pennsylvania, Philadelphia, Pennsylvania, USA

CONTENTS

A judgment is an evaluative response to a stimulus. Judgments are often based on rules of thumb, or heuristics, which may cause systematic errors. Often judgments can be improved with the use of algebraic formulas.

INTRODUCTION

We make judgments when we evaluate stimuli with respect to some set of criteria. Classic examples of judgments are the ratings given in Olympic sports events such as diving and gymnastics. The stimulus might be a dive, a half gainer, say, and the criteria are the list of properties that a good half gainer should have, such as coming close to the board, entering the water at a 90-degree angle, and so on. Other examples are grades given to students' examinations or papers, ratings of movies, and ratings of stock offerings (buy, hold, sell, etc.). Judgments need not be expressed numerically, but they are always expressed in terms of some ordered scale along a better–worse continuum.

Judgments are not the same as decisions, but decisions may depend on judgments. In making consumer purchases, we may first make judgments of price and quality. Judgments about the future are predictions. Weather forecasters make judgments about tomorrow's temperature.

Psychology researchers learn about judgments by presenting stimuli to subjects and asking for judgments as responses. By varying features of the stimuli, we can learn what affects the response. For example, in a typical experiment on judgment, a psychologist might present a set of stimuli like that shown in Table 1, one row at a time. The stimuli are in the first three columns.

The judgment in this case is a rating of suitability of each of these eight students for graduate school in psychology. Can you tell from this table what the judge is doing? (We will come back to this example.)

NORMATIVE AND DESCRIPTIVE ACCOUNTS

One way in which researchers study judgments is to compare them to a standard for how the judgments should be made. This is called a *normative* standard, or a normative model. A simple normative model of predictions of tomorrow's temperature is what that temperature turns out to be. Often we cannot specify the normative model exactly because we have no obvious right answer to the question of what the judgment should be. But we can often specify normative criteria. For example, in the case of giving grades to essays, it is good, other things being equal, to give the same grade to the same essay presented at another time. If a grader gives different grades to the same essay, we can use the difference to estimate a certain kind of error, an error of inconsistency.

Often we find that judgments show some sort of systematic error. They tend to depart in a predictable way from some normative standard. Such a systematic error is called a bias. When a bias is discovered, researchers often try to explain it. The explanation is called 'descriptive' as opposed to normative.

Explanations are typically of two types, which are not incompatible. One is in terms of a mathematical model of the judgment process. The other is a verbal description of the rules that the judge is using. These rules are often *heuristics*. A heuristic is a weak method.

Interestingly, the term 'heuristic' was coined (by the mathematician George Polya) to explain how mathematicians solve problems when they cannot rely on procedures (algorithms, such as the procedure for long division). Heuristics were thus seen as good things. The psychologists Daniel Kahneman and Amos Tversky adopted this term to explain biases, which produce departures from normative models. Did Kahneman and Tversky completely turn the meaning of the term on its head, so that

Table 1. Hypothetical predictors of suitability for graduate school

Grade average in psychology	*Grade average in science*	*Letters of recommendation*	*Judgment response*
4.0	4.0	4.0	20
4.0	4.0	3.0	19
4.0	3.0	4.0	17
4.0	3.0	3.0	16
3.0	4.0	4.0	10
3.0	4.0	3.0	9
3.0	3.0	4.0	7
3.0	3.0	3.0	6

heuristics were now seen as bad rather than good? Not really. Kahneman and Tversky, and others who follow their heuristics-and-biases approach, still think of heuristics as fundamentally useful short-cuts. We use heuristics exactly because they are usually useful. We have trouble when they are less useful than normal.

REPRESENTATIVENESS

An example is the representativeness heuristic. When people make judgments of the probability that a case belongs to a category, they often base these judgments on similarity of the case description to their concept of the category. For example, Kahneman and Tversky presented subjects with a paragraph description of 'Tom W.', a graduate student, which characterized Tom as intelligent but uncreative, with a need for order, with a writing style that was 'dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type', among other features. The subjects judged the probability that Tom was in several different fields of study.

Subjects judged the description to be more typical of the category of computer science students than of social science students, and other subjects judged Tom as more likely to be a computer science student. Usually, such use of similarity would be a good heuristic for a judgment. But a normative model of probability judgments also says that the judgments should take into account the sizes of the categories. When this study was done, the category 'social science students' was many times larger than 'computer science students'. Yet the subjects completely ignored the relative sizes of the categories.

In this example, the normative model was the theory of probability, and the descriptive model was that subjects use the representativeness heuristic, judging probability on the basis of similarity alone, and ignoring the base rates of the categories. The judgments were unaffected by the sizes of the categories, and, in this way, they were biased. The representativeness heuristic has been found in many other situations.

ANCHORING AND ADJUSTMENT

Another well-studied heuristic is found in numerical judgments. Tversky and Kahneman asked subjects to estimate certain quantities, such as the percentage of African countries in the United Nations. For each quantity, a number between zero and 100 was determined by spinning a wheel. The subject was instructed to indicate whether that number was higher or lower than the value of the quantity, and then to estimate the value of the quantity. The number the subject started with, determined solely by the spin of a wheel, strongly affected the final estimate. Subjects who were given high numbers gave higher estimates than those given low numbers.

Descriptively, subjects use a heuristic of anchoring their judgment on the number they are given and then adjusting it, but the adjustment is insufficient. Normatively, the judgment should be unaffected by the number they are given. The anchoring-and-adjustment heuristic can distort a great variety of real judgments and predictions. We can reduce this effect by avoiding the use of anchors, even though they often make the judgment easier. We cannot stop people from making up their own anchors, however. If you are asked the proportion of African countries in the United Nations, you might well start by asking yourself whether it is more or less than 50 percent. As a result, your final judgment will be too close to 50 percent.

RATIONALITY

Are biased judgments irrational? The concept of 'rational' is much more difficult to define than that of 'normative', and the two are not the same. One view of rationality is that it involves doing the best we can, all things considered. Among the things to be considered are the time and effort involved in using methods that will reduce biases and errors. Often, in daily life, our biased judgments are much better than nothing and good enough for the purpose at hand. It would not be worth the time and effort to try to avoid biases, so the biases would not, in these cases, be irrational.

In other cases, much more hinges on our judgments, and it is worth making an effort to improve

them. An example is the use of the anchoring-and-adjustment heuristic in the estimation of how long it will take to complete a project. Typically, people imagine how long it would take to carry out each step if nothing goes wrong, and then they adjust for the possibility that something will go wrong. Typically, they underadjust, so they underestimate completion times. If the time in question is that for cooking dinner, the dinner may be a little late, but that is not a grave matter. However, the same problems seem to occur in estimating the cost and time of larger projects such as the Sydney (Australia) Opera House or the tunnel under the English Channel.

HUMAN JUDGES VERSUS FORMULAS

One way to improve judgments is to use algebraic formulas. This is especially useful in the case of prediction, when formulas can be discovered by examining past cases. Such formulas have been discovered for predicting the future price of a wine crop on the basis of the weather conditions under which the grapes were grown, suicide attempts by depressed patients, college student grades, and defaults on bank loans. In principle, formulas of this sort ought to be useful in predicting the success of employees, so they could be used in hiring decisions, but they are rarely used in this way.

Many useful formulas are simply a weighted sum. We have various attributes of the object in question, such as the student's high-school grade average, high-school class rank, and test scores, to predict college grades. We multiply each of these numbers by its weight, which depends on which kind of number it is (grade average, class rank, etc.), and then we add up the results. This is called a linear model because a graph of the result as a function of any of the attributes is a straight line. Linear models make sense when each attribute is better at one end of its scale and worse at the other end. For wine, 'total rainfall' might not work this way, because there can be too much rain as well as too little. But there can't be too high a test score, in the same sense.

Researchers have compared linear models with human judges. In situations where linear models make sense, the judges are never more accurate than the linear models. Why is this? The answer is that the models always give the same answer to the same item presented twice, but human judges give different answers. Human judgments are variable. We can compensate for this variability by using the average of several judgments made by different judges, but usually it is less costly to use a formula.

Notice that the variability in judgment will occur even when we have no way of determining the right answer. In the example in the table, we have no final right answer to a student's 'suitability for graduate school'. This is a subjective judgment. Still, we can remove the variability in judgment by finding a formula for the judgments themselves. In the example in the table, the judgments are well fit by the formula: $10(psychology) + 3(science) + letters - 36$.

Ordinarily, the fit is not so good, because the judgments are higher or lower than those predicted by the best-fitting formula. Still, we can find the formula with the smallest errors in predicting the judgments, and we can use that formula. We can get a few hundred judgments, then dismiss the judges and use a formula based on their judgments.

When we have a right answer (such as employee performance), formulas derived in this way have never been found to do worse than the judges themselves. This implies that the cost of judges' errors is greater than any gains that could be made from the ability of judges to consider special patterns or exceptions to the rules. We have reason, then, to use formulas even without an objective normative criterion. They will at least lead to consistent judgments.

The errors in judgments that cause this effect are not always biases. Sometimes they are random, occurring equally often in both directions. The bias is in our overconfidence, thinking that we can do better by considering individual cases. In fact, in every case so far, the linear model of the judge does better than the judge.

Further Reading

Baron J (2000) *Thinking and Deciding*, 3rd edn. New York, NY: Cambridge University Press.

Buehler R, Griffin D and Ross M (1994) Exploring the 'planning fallacy': why people underestimate their task completion times. *Journal of Personality and Social Psychology* **67**: 366–381.

Connolly T, Arkes HR and Hammond KR (2000) *Judgement and Decision Making: An Interdisciplinary Reader*. New York, NY: Cambridge University Press.

Dawes RM, Faust D and Meehl PE (1989) Clinical versus actuarial judgment. *Science* **243**: 1668–1674.

Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgment under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Tversky A and Kahneman D (1974) Judgment under uncertainty: heuristics and biases. *Science* **185**: 1124–1131.

Knowledge Representation, Psychology of

Introductory article

Arthur B Markman, University of Texas, Austin, Texas, USA

CONTENTS

'Knowledge representation' is an umbrella term for the methods by which information is stored in the mind for later use.

THE BASICS OF REPRESENTATION

From the beginning of the study of psychology, philosophers and psychologists have been interested in the way information is stored in the mind. In his *Theaetetus*, the Greek philosopher Plato described memory as a wax tablet, in which information is stored as impressions in the wax. In this proposal, the information remained in memory for as long as the impression remained in the wax. Later in the same work, Plato suggested that memory is like an aviary with birds flying around it. Retrieving information was like grabbing a bird from the aviary.

The philosopher David Hume realized that one problem with proposals about the nature of representation was that they required someone (or something) to look at them to interpret the information in them. He searched for some form of self-interpreting representation, which led him to speculate about the various ways that ideas can be related to each other. The idea that there must be some kind of person in the head to interpret representations – called a homunculus ('little man') – baffled psychologists and philosophers until the development of computers. Computers process information by defining processes (or algorithms) that manipulate the structure (or syntax) of the information stored in representations. Computers are able to process information without needing a homunculus, because they need to attend only to the structure of the information presented to them, without needing to understand the content. (*See* **Knowledge Representation**)

Since the late 1950s, psychologists have operated with a computational view of mind. This view assumes that the mind is like a computer in that it carries out algorithms using representations. It is important to bear in mind that this view does not assume that the mind is constructed using the same components as a computer (such as having a central processing unit). The mind is like a computer only in the way that information is processed. (*See* **Knowledge Representation**)

Based on the computational view of mind, a definition of representation can be developed that has four parts: (1) a represented world, (2) a representing world, (3) a set of representing rules, and (4) processes for using the information in the representation. These ideas are illustrated in Figure 1. The represented world is the situation that is going to be represented. For example, the represented world might be information about the outside world. In Figure 1(a), the represented world is the heights of three people. In complex systems, the represented world for one representation might be another representation inside the system. (*See* **Symbol-grounding problem**)

The representing world is the system used as a representation. In Figure 1(b), the representing world is a set of lines that vary in length. The represented and representing worlds are related by a set of representing rules that specify how aspects of the represented world are encoded in the representing world. These relationships determine how information is carried by the representation. For example, the rules might state that larger heights (from the represented world in Figure 1(a)) are represented by longer lines (from the representing world in Figure 1(b)). The representing rules may be arbitrary. For example, Figure 1c shows a symbolic representation for heights using familiar Arabic numerals. In order for these symbols to represent heights, a set of rules relating these symbols to quantities must be established.

Finally, in order for something to be used as a representation, there must be some process that makes use of the information. This last part of

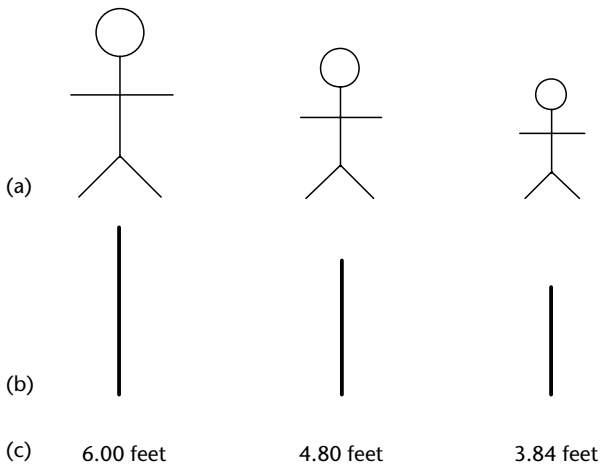


Figure 1. Sample representations: (a) a represented world; (b) a representing world in which the height of the line represents the height of the people; (c) a symbolic representing world in which numbers are used to represent the height of the people.

the definition is quite important. Often, proposals about psychological representations are accompanied by figures that depict the representation. When someone looks at the picture, it seems obvious what information is there. However, if no process is specified that can use the representation, then it is as if the information were not part of the representing world at all.

As an example of a very simple system that uses representations, consider a common thermostat attached to a home heating system. The thermostat has a bimetal strip that curls more as the temperature rises. Attached to the end of the bimetal strip is a glass bulb filled with mercury. When the bimetal strip loses its curvature (because of a decrease in room temperature), the mercury collects in the bottom of the tube, connecting a pair of electrical contacts that completes a circuit and turns on the heater. When the temperature rises, the strip curves more, the mercury pools in the end of the bulb, the circuit is broken, and the heater shuts off. In this system, the represented world is the temperature in the room. The representing world is the shape of the bimetal strip. These worlds are linked by physical laws that determine how changes in temperature affect the shape of the strip. Finally, the bulb of mercury allows the thermostat to use the information to turn the heater on and off.

A thermostat does not have a deep psychology. Nonetheless, according to the definition given earlier, it does have simple representations (though some philosophers reserve the term 'representation' for more complex forms of information-carrying systems). What separates such simple systems from more complex ones like people is the additional properties that can be added to the representations that enable them to carry out more complicated processes. The following sections will describe some important types of representations used in cognitive science.

TYPES OF REPRESENTATIONS

There are a number of different types of representations that have been proposed in cognitive science. These representations differ from each other in two ways. First, representations can be distinguished by virtue of the kinds of information that can be stored in them. Second, each type of representation has its own set of processes that can naturally be applied to it. The processes that can be applied to a representation determine what can be done easily and what can be done with difficulty. This section looks at four types of representations: (1) spatial representations, (2) featural representations, (3) semantic networks, and (4) structured representations.

Spatial Representations

Spatial representations use a space as a representing world. For example, Figure 2 shows a diagram of a basketball play. The positions of the players in this diagram are meant to represent the positions of the players in space on a real basketball court. Thus, in this example, the representing world is a two-dimensional space, and the represented world is the two-dimensional surface of a basketball court. In this case, one space is being used to represent another. However, it is possible to use spatial representations for a variety of different kinds of represented worlds.

A space is a geometric entity that satisfies certain conditions. It can be described by a set of dimensions, where each dimension is a line that points in an independent (or orthogonal) direction. For example, in Figure 2, the space has two dimensions. To see that it has two dimensions, look at the top left corner of the box surrounding the diagram. There are two perpendicular lines that emerge from this corner. If you move along the line moving downward, then you are not moving at all in the left–right direction. Similarly, if you move along the line emerging to the right, you are not moving at all in the up–down direction. Thus, these lines are orthogonal.

Any other line that you draw in the plane of the paper will move at least partially in the direction of

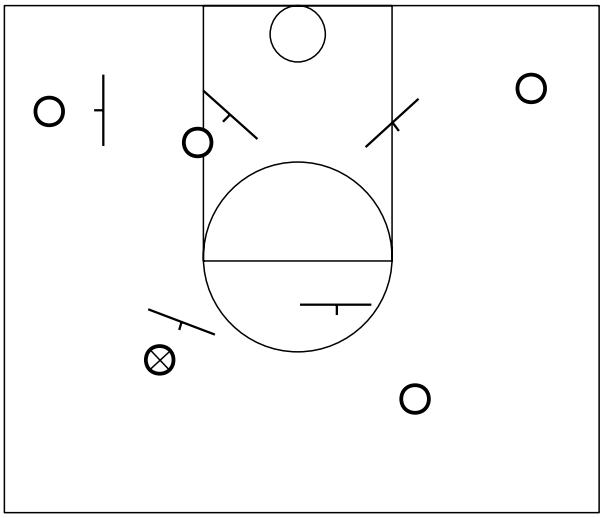


Figure 2. An example of using a spatial representation to represent a space. In this case, the basketball diagram represents the positions of players on the two-dimensional surface of the court using their position within a two-dimensional rectangle. The circles represent offensive players, the circle with the x is the player with the ball, and the lines represent defensive players where the small line indicates the front of the player.

one or both of the lines. The space has two dimensions, because you can put at most two orthogonal dimensions into the space. Physical space has three dimensions, but a psychological space can have any number of dimensions. While it might be difficult to imagine a space with 1000 dimensions, the mathematics that defines the concept of a space can still be used. (*See* **Spatial Representation and Reasoning**)

Distance measured in a space must obey the metric axioms. The three metric axioms are minimality, symmetry, and the triangle inequality. Minimality states that the distance between any point and itself is zero. The symmetry axiom states that the distance between any two points is the same regardless of which point you start from. Finally, the triangle inequality states that the distance between two points is less than or equal to the sum of the distances between those two points and a third point. These metric axioms are important, because they provide a set of characteristics that the representing world must have, which in turn places constraints on what a spatial representation can be used for.

Finally, in a spatial representation, distance can be measured in many ways. The measure that is most familiar from our everyday experience is the Euclidean metric. In a Euclidean space, the distance between any two points is simply the length of the line between the two points. Another way of measuring distance that is often used in spatial representations is the 'city block metric', in which the distance between two points is equal to the distance you would have to travel to get from one point to the other moving only in the direction of the dimensions of the space. For example, if you were in New York City, and wanted to get from 10th Avenue and 50th Street to 8th Avenue and 63rd Street, you could not simply walk straight from one point to the other. Instead, you would have to follow 10th Avenue to 63rd Street and then walk along 63rd Street to 8th Avenue. Often, the Euclidean metric is used for spaces that represent psychological dimensions that are integral (such as the saturation and brightness of a color). Integral dimensions are ones for which people have difficulty attending selectively to one but not the other. In contrast, the city block metric is often used for spaces that represent psychological dimensions that are separable (such as the color and size of a shape).

The easiest process to carry out in a space is the determination of distance, and so many psychological models that use spatial representations focus on cases in which the distance in the representing world corresponds to a salient property of the represented world. For example, a concept space can be developed in which the points in the space represent categories of objects (such as different animals). The distance between points is then inversely proportional to the psychological similarity of those animals. Thus, the closer a pair of objects, the more similar they are.

Spaces have been popular as the basis for models of mental representation, because there are many ways to create spatial representations. One popular technique is multidimensional scaling (MDS). Multidimensional scaling takes a set of distances between points and creates a spatial map of those points. For example, if an MDS program is given the flying distances between 10 European cities, it generates a two-dimensional map of the relative locations of those cities. MDS programs can also be given conceptual distances. For example, a program could be presented with pairwise similarity ratings among a set of concepts obtained from a group of subjects, in which case the output would be a map of the conceptual space. (*See* **Multidimensional Scaling**)

Multidimensional spaces can also be derived from analyses of text. Techniques such as latent semantic analysis (LSA) and the hyperspace analog to language (HAL) take large corpora of text (sometimes with millions of words) and perform

statistical analyses on the co-occurrences among words. Using these techniques, a space is generated in which words that often appear in the same language contexts are represented by points that are near to each other in space. These representations often have hundreds or thousands of dimensions. Analyses of these spaces suggest that they capture a number of aspects of word meanings, like the degree of association among words and their grammatical class. (*See* **Lexicon, Computational Models of**; **Language, Connectionist and Symbolic Representations of**)

Finally, many distributed connectionist models use spatial representations. In a connectionist model, there are many simple units that are interconnected. Each unit, which is roughly analogous to a neuron in the brain, has some degree of activation. The pattern of activation across a set of units is assumed to represent information. These patterns of activation are often written mathematically as vectors. A vector may also be interpreted as a list of coordinates of a point in a multidimensional space. The processes that allow input units in a connectionist model to give rise to patterns of activation on output units can also be interpreted as spatial operations. (*See* **Connectionism**; **Distrubuted Representations**)

In sum, in a spatial representation, items are represented as points (or vectors) in space. The primary operation carried out on pairs of points is a determination of distance between them. The distance between points often has a psychological interpretation. In many cases, distance is assumed to be inversely proportional to psychological similarity.

Spatial representations are not suitable for all kinds of cognitive models. In particular, there is no way to refer to specific items in a space, because a space has no symbols in it that can serve as labels. Thus, whenever a cognitive process requires access to symbols, a purely spatial representation is a poor candidate as an underlying representation. In those cases, one of the representations described in the following sections should be used.

Featural Representations

As mentioned in the previous section, it is often important to be able to refer to a particular aspect of a mental representation, as opposed to merely knowing the similarity of a pair. For example, language involves using labels for objects and for properties. These labels are symbols. (*See* **Symbol Systems**)

The philosopher Pierce distinguished between icons, indexes, and symbols. An icon represents something by physical resemblance. For example, a portrait of Thomas Jefferson is an iconic representation. An index represents something through a causal relation. For example, smoke indexes a fire, because smoke is caused by fires. A symbol has an arbitrary relation to the thing it represents, where that relation is fixed by convention. For example, the word 'hat' is related to the concept 'hat' by arbitrary convention (as evidenced by the fact that other languages may use very different words for the same concept).

In a featural representation, the represented world is a collection of features that are used to describe concepts. In the simplest featural representations, the features are treated as independent of each other. This assumption of independence eases the processes that can be carried out on the representation.

For example, a featural model of similarity (as shown in Figure 3) assumes that each object is represented by a set of features. Then, elementary set operations can be used to determine the common and distinctive features of the objects. The common features are those in the intersection of the feature sets representing the two objects. The distinctive features are those that are not in the intersection. Studies of similarity have demonstrated that the rated similarity of a pair increases with the number of common features of the pair and decreases with the number of distinctive features of the pair. (*See* **Analogical Reasoning, Pscychology of**; **Similarity**)

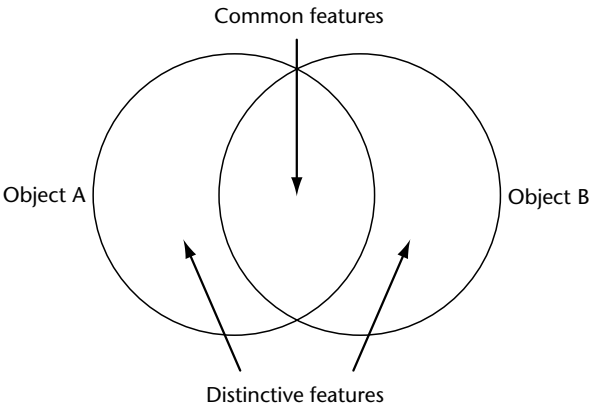


Figure 3. Sample feature sets for two objects, illustrating that the intersection of the feature sets is the set of common features, and that the remaining features are distinctive features.

For featural models, as well as more complex kinds of representations, it is important to consider what the basic (primitive) elements of the representations are going to be. Attempts to generate the primitive elements for representations have been made in research on object recognition, word learning, and text comprehension. Each of these theories tried to posit a small set of basic elements that could be used to construct the remaining concepts in the representational system. For example, Schank's conceptual dependency theory of text comprehension proposed that there are 11 primitive actions, and that verb meanings involve combinations of these actions. Attempts to specify primitives have failed because, for any set of primitives posited, there are many exceptions that require additions and modifications to the initial set. (*See* **Distinctive Feature Theory**; **Prototype Representations**)

Featural representations are the simplest form of symbolic representation. They are simple, because they represent only a set of properties that an object might possess. For example, the concept 'bird' might have the features {wings, flies}. However, this simple representation does not have any information about how the features are related. For example, this representation could not capture a belief that having wings enables a bird to fly. A number of extensions of featural representations have been developed, however. These extensions are designed to account for the way people represent and process relations among aspects of mental representations. In the following sections, two such representations are discussed: semantic networks and structured representations. (*See* **Conceptual Representations in Psychology**)

Semantic Networks

Semantic networks capture relationships among concepts. A small piece of a semantic network is shown in Figure 4. This network consists of nodes (shown as ovals in Figure 4) and links (shown as arrows). The nodes represent concepts in the domain, and the links represent relations among the concepts. If you were to state the relation as a simple sentence, then the arrow points from the subject of the sentence to the object (e.g. an animal has the part 'cells'). Many different kinds of relations among concepts can be represented. Even in the simple example in Figure 4, there are three different kinds of relations. (*See* **Semantic Networks**; **Representations, Abstract and Concrete**)

There are two kinds of processes that are typically associated with semantic networks: marker passing and spreading activation. Marker passing is used by networks for language processing. When interpreting a sentence, a scout is sent to the nodes for concepts mentioned in the sentence. At the first time step, the scout walks along every link leaving a node (i.e., for which the arrow points out of the node), and marks the node. At each subsequent time step, the scouts move further away from the original nodes down the links from the nodes they visited. When scouts starting from two different places mark the same node, an intersection has been found, and an interpretation is formed of the path between the nodes where the scouts originated. (*See* **Implicit and Explicit Representation**)

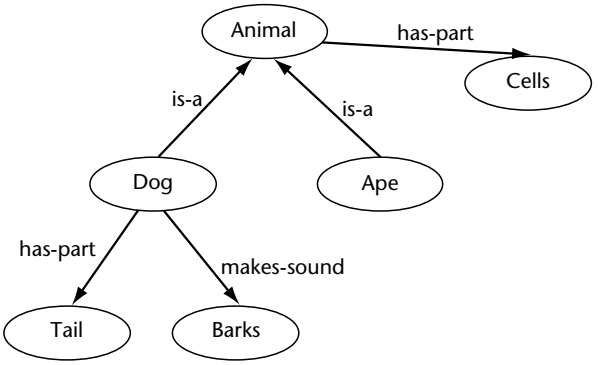


Figure 4. A piece of a semantic network. The ovals are nodes, which represent concepts, and the arrows are links, which represent relations between pairs of concepts.

For example, if a model were trying to verify the sentence 'dogs have cells', then it would send scouts to the 'dog' node and the 'cells' node. Scouts would leave the 'dog' node and walk to the 'tail', 'barks', and 'animal' nodes. At the next time step, the scout would leave the 'animal' node and reach the 'cells' node. At this point, this scout would intersect with a marker from the other scout, and the marker passing process would stop. Analyzing the path between these nodes suggests that dogs do have cells, because dogs are animals and animals are known to have cells.

The second prominent process carried out on semantic networks is spreading activation. Models such as Anderson's ACT (Adaptive Control of Thought) assume that each node can have a degree of activation that corresponds to its current prominence in working memory. In addition, each link has a weight that determines the influence of activation of one node on the activation of another. If a link has a positive weight, then activation of one node will increase the activation of the node to

which it is connected. If a link has a negative weight, then activation of one node will decrease the activation of the node to which it is connected. (*See* **ACT**)

Spreading activation models are used to account for patterns of associative relatedness in memory. For example, a classic finding is that the time taken to determine that a string of letters is a word (i.e. to make a lexical decision) is faster if the string of letters is preceded by a related word than if it is preceded by a neutral string of asterisks. In the present example, if the string of letters DOG was preceded by the word 'barks', activation from 'barks' would spread to the concept 'dog', which would in turn speed the recognition of the word 'dog'.

Semantic networks are quite useful for modeling simple comprehension processes and effects of semantic relatedness. They have difficulty with more complex processes, however. For example, in sentence verification, it is difficult to determine when a property is not true of an object, because the scouts would not intersect for properties that are not possessed by an object. In addition, in semantic networks, the links can connect only pairs of objects. In contrast, there may be times when a cognitive process must relate two or more relations to each other. In this circumstance, more complex structured representations are needed.

Structured Representations

Semantic networks are a constrained form of a more general type of representation called a structured representation. The key aspect of a structured representation is the concept of binding in which the scope of a relationship is determined by items it connects (or binds). For example, we could rewrite the relationship that a dog is an animal from Figure 4 using the following structured representation:

$$\text{is-a}(\text{dog}, \text{animal}) \quad (1)$$

In this simple example, the term 'is-a(*?x*,*?y*)' is a predicate, and the items 'dog' and 'animal' are constants. The *?x* and *?y* in the predicate are variables that can be filled by constants or other predicates. These variables specify the scope of the relation. In the example, when the variables are bound to constants, the predicate specifies that the first item is a type of the second. Variables that specify the scope of a relation are called 'arguments'. When the arguments to a predicate are specified, then the statement is called a 'proposition', because it can be evaluated as being true or false about the represented world. (*See* **Binding Problem**)

Structured representations permit efficient representations of relational information. In general, if we are able to represent a relation like 'kiss (John, Mary)', then we can also represent the same relation with the arguments reversed, 'kiss(Mary, John)'. Thus, we can use the same three elements (kiss(*?x*,*?y*), Mary, and John) to represent two very different concepts. It would be quite difficult to represent these relations with a featural representation, because there is no way to bind features to each other. (*See* **Representations Using Formal Logics**; **Syntax**)

Structured representations are more general than semantic networks in two ways. First, in semantic networks, all links relate two items (i.e., there are only predicates with two arguments). In contrast, in structured representations, it is possible for a predicate to have only one argument or for it to have more than two arguments. For example, predicates with one argument are often used to represent specific attributes of an object such as:

$$\text{red}(?x) \quad (2)$$

Predicates with more than two arguments may be used to represent complex actions, such as:

$$\text{give}(\text{?giver}, \text{?recipient}, \text{?item-given}) \quad (3)$$

The second way in which structured representations are more general than semantic networks is that semantic networks relate only concepts specified by nodes, but structured representations permit predicates to take other whole propositions as arguments. These more complex predicates are often used to represent causal relations, such as:

$$\text{cause}\,[\text{love}(\text{John}, \text{Mary}), \text{kiss}(\text{John}, \text{Mary})] \quad (4)$$

In this example, the fact that John loves Mary causes him to kiss her, where the predicates 'love(*?x*,*?y*)' and 'kiss(*?x*,*?y*)' serve as arguments to the cause relation. (*See* **Causal Reasoning, Psychology of**)

The processes that act on structured representations can be quite complex, because they must be sensitive to the way predicates take arguments. For example, when a process that calculates the similarity of a pair of structured representations finds a relation in each representation that corresponds, this process must then ensure that the matching relations have matching arguments.

Making use of this argument structure increases the complexity of the processes that act on structured representations relative to processes that act on featural representations.

USING REPRESENTATIONS

The diversity of approaches to representation in psychology reflects that there are a variety of goals that cognitive processes must satisfy. Some processes must run quickly and efficiently. These processes are likely to use spatial or featural representations or perhaps a semantic network, because procedures such as distance calculation, feature set comparison, or spreading activation can be carried out quickly. In contrast, the procedures that operate on structured representations are often computationally intensive, and thus require time to complete. (*See* **Computablity and Computational Complexity**; **Computation, Philosophical Issues about**)

There are also situations in which a cognitive process requires the ability to represent complex information (such as language comprehension or causal reasoning). In these cases, efficiency may not be an issue, as the process may have seconds or minutes to be carried out. In this case, the greater expressive power of structured representations outweighs the complexity of the processing.

This analysis leads to an important topic in current research on knowledge representation, which is the way that different types of representations are integrated during processing. As an example, the ACT model mentioned above uses a spreading activation process to allow concepts to be retrieved from memory. Because memory retrieval requires the potential consideration of many different pieces of information, it is important to have processes that run efficiently. The ACT model also uses structured representations for carrying out reasoning processes where there is sufficient time to permit the use of complex representations. Work that tries to integrate representations has also been done in the areas of object representation and analogical reasoning. There is much still to be learned, however, about the way different representational systems can be incorporated into a single model. (*See* **Connectionist Implementations and Hybrid Systems**)

Further Reading

Anderson JR (1978) Arguments concerning representations for mental imagery. *Psychological Review* **85**(4): 249–277.

Anderson JR (1993) *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gärdenfors P (2000) *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Kruskal JB and Wish M (1978) *Multidimensional Scaling*. Newbury Park, CA: Sage.

Landauer TK and Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**(2): 211–240.

Markman AB (1999) *Knowledge Representation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Markman AB and Dietrich E (2000) In defense of representation. *Cognitive Psychology* **40**(2): 138–171.

Palmer SE (1978) Fundamental aspects of cognitive representation. In: Rosch E and Lloyd BB (eds) *Cognition and Categorization*, pp. 259–302. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank RC (1972) Conceptual dependency: a theory of natural language understanding. *Cognitive Psychology* **3**: 552–631.

Tversky A (1977) Features of similarity. *Psychological Review* **84**(4): 327–352.

Language and Cognition

Introductory article

Catherine L Harris, Boston University, Boston, Massachusetts, USA

CONTENTS

Cognitive scientists have long debated whether language and cognition are separate mental faculties, or whether language emerges from general cognitive abilities.

INTRODUCTION

What is the relationship between language and cognition? Do people who speak different languages think differently? Is a certain level of cognitive development required for language acquisition? These questions were of keen interest to thinkers in the early twentieth century and remain important in anthropology, linguistics and psychology. However, the cognitive revolution of the 1950s raised a new question about the relationship between language and cognition: is language the same type of mental entity as other cognitive abilities, or is it fundamentally different?

A hallmark of modern cognitive science is the goal of developing a theory of cognition powerful enough to encompass all human mental abilities, including language abilities. A long-standing controversy concerns two ways of conceptualizing the architecture (or basic design) of cognition. One approach proposes that general-purpose processes and mechanisms provide a foundation for all varieties of human intelligence. We can refer to this as 'general purpose' cognition. Examples of possible universal processes are the ability to induce a category from exposure to examples (category induction), and the ability to mentally complete a known pattern when confronted with a piece of it (pattern completion). Cognitive scientists frequently attempt to precisely specify their proposed mechanisms by implementing them as computer algorithms which can be tested in artificial intelligence (AI) programs. Researchers have tried to use AI programs to show that the same principles that can explain general problem solving can also explain aspects of language acquisition and processing.

The second way of conceptualizing human cognition emphasizes the differences between language and other abilities. A key idea is that many distinct domains of cognition exist and must be learned separately, using different mental mechanisms. This approach is referred to as the 'modularity of cognition' or 'mental modules' approach. At first glance it may seem contrary to the interdisciplinary spirit of cognitive science and to the possibility of a unified theory of cognition. However, the unifying theory is the thesis of distinct mental modules, which are believed to have evolved to accomplish specific tasks relevant to mammalian evolution, such as visual exploration, or relevant to human evolution, such as language use. Much of the appeal of this approach comes from findings in neuropsychology showing that distinct areas of the brain serve distinct functions such as vision, language processing, motor coordination, memory, and face recognition. The interdisciplinary spirit is maintained because advocates of this approach reach out to biological scientists and evolutionary theorists. Those favoring modularity embrace the principle of converging methodologies: a theory must have explanatory power in the distinct academic disciplines that compose the cognitive sciences. These two approaches to the architecture of cognition developed out of different philosophical traditions, and have evolved considerably during the half-century history of cognitive science.

CONCEPTS OF COGNITION AND LANGUAGE

Why are there two different views on the relationship between language and cognition? At the dawn of the cognitive revolution, in the late 1950s, there were two distinct ideas about the nature of mind.

Discussed first are the views of linguist Noam Chomsky and the field of generative linguistics which he developed, because the underlying philosophy has remained fairly constant over the intervening years.

Chomsky's major innovation was to conceive of language abilities as akin to a mental organ. According to this view, children are born with a 'language acquisition device' and with specific linguistic knowledge. This knowledge is thought to include the concepts of noun, verb, grammatical subject, and structures that constrain possible grammatical rules. In contrast to the views of the dominant psychological paradigm of the 1950s, behaviorism, Chomsky argued that children do not learn to speak by imitating adults. His key evidence was that children spontaneously use incorrect forms they could not have heard, like 'goed' and 'breaked'. Linguistic overregularizations like these suggest that children are extracting rules from the language they hear, not merely imitating. Theorists at that time found it noteworthy that parents do not generally tell children that their utterances are ungrammatical. Because the language input to children is full of mistakes, stops and restarts, Chomsky felt that children could not learn language using general purpose problem-solving or regularity-extraction skills. They needed to come to the task with a rich set of expectations about the nature of language. These expectations were believed to be specific to language, and thus did not share commonalities with other aspects of cognition. This set of language-specific abilities has been variously called the 'language acquisition device' (the historically early term) and 'universal grammar' (a more recent term). Chomsky's approach to linguistics is called 'generative linguistics' because its early goal was to describe mental structures that can generate all the grammatically valid sentences of a language.

Chomsky felt that the aspect of language that is unique is syntactic ability. An example of specifically syntactic knowledge is illustrated by the sentence, 'Colorless green ideas sleep furiously'. Although the words in this sentence contradict each other and do not correspond to a possible reality (green is not colorless, ideas cannot sleep furiously), speakers nevertheless recognize the sentence as having a correct grammatical structure. Chomsky used this sentence as an example of how syntactic structure represents information independently from the meaning of the words in the sentence. He argued that syntax is a unique, independent human capacity and not derivative from other abilities. The proposal that syntax is not influenced by the meaning of the words in the sentence or speakers' communicative goals came to be called the 'autonomy of syntax' hypothesis.

Chomsky's innovations developed in tandem with the dawn of the cognitive revolution. With the birth of computer science a new way of conceptualizing human cognition arose, using the metaphor of the brain as a computer and the mind as software. Much of Chomsky's early work depended on these metaphors: he conceived of grammar as a set of rules for generating novel combinations of words, just as a computer program could generate a string of symbols according to a formula. Thinking of mental operations as akin to steps in a computer program allowed psychologists and workers in the new field of artificial intelligence to begin to subdivide mental tasks, such as arithmetic or language comprehension, into a series of steps in a computer program. Their research evolved in a different direction from Chomsky's. Computer scientists and psychologists such as Alan Newell, John Anderson, Roger Schank and Patrick Winston began to describe a range of human abilities – from visual object recognition and general problem solving, to metaphor use and story understanding – in terms of a set of internal representations and processes that transform those representations. A key aspect of this approach to cognition was that the theorist could write computer programs that would provide a formalizable theory of mental operations. A rigorous test of the theory could be performed by running the program and seeing if output matched human output.

Many of the early successes of this field involved language, including sentence comprehension, story understanding and metaphor use. An important aspect of the language and cognition relationship is that the AI models of language did not draw on language-special algorithms or knowledge structures. The new AI tradition and the information processing movement within psychology emphasized learning, particularly general-purpose learning, and was thus opposed to the emphasis on innate knowledge structures that was part of Chomsky's new linguistics. Psychologists drawn to the information processing movement were frequently inspired by the work of Swiss psychologist Jean Piaget, whose works began to be translated into English in the 1960s. Piaget also emphasized the commonalities between language and cognition, and proposed that language emerged out of the same broad cognitive changes that transform the sensorimotor processing of infants into the formal and logical mind of adults.

Table 1. Theoretical perspectives on the language–cognition relationship

Timeline	*Movement*	*Main source of constraints*	*Language/cognition*
1957–present	Chomskyan linguistics	Innate	Language unique, unlike cognition
1960s–1990	Artificial intelligence	Learned	Subject to same principles
1980s–1990	Connectionism	Learned	Subject to same principles
1980s–present	Modularity of mind	Innate	Language unique, unlike cognition
1990s–present	Cognitive neuroscience	Dynamical interaction	Complex similarities and differences

In the 1970s and 1980s, among those researchers who aligned themselves with the interdisciplinary field of cognitive science, there was a tendency for linguists to emphasize the specialness of language and cognition, and for psychologists to emphasize commonalities between language and cognition. However, there was considerable variation in viewpoint within psychology and linguistics, which continues to this day. Three scientific developments during the 1980s and the 1990s had implications for theories of the language–cognition relationship: connectionism, cognitive linguistics, and the cognitive neuroscience movement. The development of these movements is summarized in Table 1.

CONNECTIONISM

The 'connectionist revolution' was launched in the mid 1980s. A new computational metaphor emerged to explain both language and cognition, based not on the Von Neumann computer, but on the idea that sophisticated computations emerge from massive networks of simple processing units, in which the units are akin to idealized neurons.

In the early 1980s several groups of cognitive scientists became dissatisfied with the AI systems being used to model cognition. These rule-governed systems employed databases containing knowledge of common situations, such as what one does at a restaurant; 'if–then' rules specified the action that an expert would take in a specific situation. This expert knowledge thus constituted intelligent reasoning. Critics noted that these expert systems were brittle and frequently failed to generalize beyond the bounds of their circumscribed database. Real human cognition seemed intuitively to involve not the application of fixed rules to a specific situation, but the satisfaction of many soft constraints, some of which could be ignored. Categories were not logical sets of necessary and sufficient conditions, but were graded groupings which shared 'family resemblance', as had been argued earlier in the century by Wittgenstein. To the mathematical and cognitive psychologist David Rumelhart, flexible and creative reasoning seemed to be similar to the motor process involved in reaching for a cup behind a pencil-holder on one's desk – or at least, more similar to this type of planning than to rule-governed algebraic reasoning.

Rumelhart and other cognitive scientists such as Terrence Sejnowski and Stephen Grossberg agreed that mental functioning involves computation, but asked what type of computation would be carried out by an organic structure like the brain, composed of massive numbers of simple processing units (neurons), linked together in a complex network of fiber tracts, with many units firing simultaneously. This movement was labeled 'connectionism' because intelligent behavior was posited to emerge from large numbers of neuron-like processing units, connected together into networks in ways that fostered parallel processing. One early success story was a connectionist model that could learn English past tense forms given the present tense of a verb. The domain of English past tense was provocative because it contained both rule-like behavior (past tense forms are generated by adding '-ed' to a present tense form), and exceptions which themselves congregated into patterns, such as that emerging from considering 'grow/grew', 'blow/blew', 'know/knew'. Both connectionist networks and adult second-language learners are likely to make the error of generating 'glew' for the past tense of 'glow', presumably through analogy with this subregularity within the past tense system. Other connection language models explored how ambiguous words are understood in their sentence context (such as 'bat' in 'The boy hit the bat'), and how abstract categories such as noun and verb can emerge from distributional regularities in text (such as the fact that nouns tend to occur in similar sentence positions). Because the basic system of processing units under these connectionist language models was the same as those used to model visual and motor behavior, successes like the past tense model were taken as support for a

common computational architecture underlying both language and cognition.

COGNITIVE LINGUISTICS

A subset of linguists disagreed with Chomsky's emphasis on the uniqueness and specialness of language, then generally accepted. The field of cognitive linguistics emerged in the late 1980s and helped initiate a flood of work connecting language and cognition. One source for the cognitive linguistics movement was an older tradition within linguistics called 'functionalist' linguistics. This held that constraints on the form of language (where 'form' means the range of allowable grammatical rules) derive from the function of language. The goals of using language include serving efficient communication. Thus the function of diverse syntactic forms, such as subject versus object and main clause versus subordinate clause distinctions, serve communicatory goals such as conveying which information is most important and which is background or context. However, the goals of conveying the relevance and background/foreground structure of the message are outside the language system. The proposal that the form of grammatical rules is influenced by communication is thus inconsistent with the autonomy of syntax hypothesis. That hypothesis specified that syntax was its own system, not shaped by the need for efficient processing or other exigencies of communication.

Functionalist linguists as well as other linguists were dissatisfied with the range of linguistic phenomena excluded by generative linguistics. They rejected the Chomskyan view that the most important aspect of language was a mechanical device for generating only legitimate grammatical sentences. They wanted to understand language in all its diversity, including narrative, discourse, dialects, sociocultural influences on language use, and metaphor. Some functionalist linguists even proposed that rule use is just a minor aspect of language. Linguists such as Dwight Bolinger and Charles Fillmore argued that speech utterances, not rules for generating utterances, are what is mentally stored. They noted that every language speaker has memorized huge numbers of odd coinages, colloquialisms, idioms and collocations, many of which share patterns. An example is the phrase 'know by heart' which has the variation 'learn by heart'. Rules, partial regularities and rule exceptions appear to differ in degree, occupying a continuum from fully idiosyncratic, to partially regular, to fully rule-governed. This idea made researchers in this nascent linguistic movement sympathetic to the connectionist movement, although the commonalities have not yet been fully explored.

Although trained in the generative grammar tradition, linguists such as George Lakoff and Ronald Langacker noted that one could not describe which sentences are syntactically valid and which are invalid without reference to nonlinguistic concepts. They pointed to the tendency for the same words describing movement in space to be used to describe movement in time ('This meeting runs until 3 o'clock'), and emphasized the necessity of incorporating the cognitive psychology of human category formation into linguistics. This approach to linguistics came to be called 'cognitive' linguistics because aspects of general cognition – such as how we construe the meaning of a grammatical construction – were proposed to be important for describing linguistic structure. For example, one descriptive problem of grammar is to account for why some sentences, but not others, can undergo the 'passivization' transformation. The sentence 'John was hit by Mary' sounds fine, but 'John was known by Mary' does not. To fully describe which transitive sentences can undergo passivization, one needs to invoke the notion that the subject and direct object are in dynamic interaction with each other. Thus, one cannot passivize a sentence like 'John left the auditorium' because John is not acting on the auditorium in a way that has consequences for it; but the sentence 'John left the auditorium unguarded' can be transformed into 'The auditorium was left unguarded by John', presumably because John's action affects the status of the auditorium.

By 2000 the cognitive linguistics movement had grown into an enduring subfield, but it has remained outside the mainstream of linguistics. While some cognitive linguists have remained focused on specific linguistic questions, others have addressed questions in an interdisciplinary manner, drawing on experimental psychology, brain science, and category induction performed by artificial neural networks.

THE COGNITIVE NEUROSCIENCE MOVEMENT

The field of cognitive neuroscience emerged from work in neuroscience and cognitive science. Cognitive neuroscience differs from basic neuroscience by having the goal of explaining complex cognitive abilities, but rejects the tradition of artificial intelligence (and much of cognitive science) that one can understand cognition abstractly, without reference to its neural underpinnings.

In the 1990s some cognitive neuroscientists argued that basic aspects of the language–cognition relationship, such as the autonomy of syntax hypothesis and the innateness and modularity of language, could be evaluated from the neuroscientific point of view. Neurobiologists have noted that developing neural tissue is very plastic. For example, the auditory association areas of the brain frequently represent visual and gestural language in individuals who are born deaf. The regions of the brain that mediate language use appear to be especially malleable. Like other aspects of cognition, language acquisition is heavily dependent on experience.

Ralph-Axel Mueller has remarked that regional specialization in the brain is beyond doubt, but modularity of cognitive functions, including language, is highly debatable from the view of neurobiology and evolution. Functional specialization of brain areas most probably emerges because some brain areas are near to the site of sensory input, such as sensory systems for vision and audition. Scientists such as Jeffrey Elman, Elizabeth Bates and their colleagues note that however closely one looks at the anatomy and physiology of the brain, there is no evidence of cortical structures unique to language or unique to humans. These researchers argue that language has an 'epigenetic' not a 'genetic' origin. Epigenetic development is the proposal that behavior results from a complex dynamic evolution of genes and environmental forces during prenatal and postnatal development. The concept of epigenesis dates back to psychologist Jean Piaget, who argued that cognitive abilities emerge from a biological structure which evolves, both before and after birth, in tandem with environmental forces. Contemporary researchers who embrace the epigenetic view point out that there are too few genes in the human genome to code directly for outcomes such as the ability to use language. Like other brain regions, the language areas in the adult brain are the end product of complex chains of interactions with internal and external environments. These sequences of events are based probabilistically on genes rather than being rigidly determined by the genome.

The neurobiological evidence thus may run counter to what would be expected under the autonomy of syntax hypothesis. There is no known way that genes could encode for concepts like 'subject' and 'verb'. Thus the most parsimonious perspective is that language is similar to other aspects of cognition in terms of emerging out of a brain which evolved to have an oversized frontal cortex (relative to other primates) and an elongated period of childhood, which privilege the role of learning.

Cognitive neuroscientists share a view of language that resonates with the cognitive linguists: they emphasize the joint development of language and perceptuomotor processes, with language acquisition understood to be semantically driven and embodied. The neurological representation of grammar is continuous with the representation of other language 'components' and the neural substrate.

CONCLUSION

In the first half of the twentieth century the main question about the relationship between language and cognition was whether the grammatical structure or vocabulary of our language influenced thought processes. Cognitive science introduced a new question: are language and cognition similar or distinct human abilities? The last 50 years have seen considerable controversy on this question, mirroring the development within cognitive science of two fundamentally different conceptions of the cognitive architecture. The tradition of artificial intelligence emphasized general-purpose problem-solving abilities, while the tradition of linguistics and philosophy led to an emphasis on distinct mental modules.

The different theoretical perspectives on the language–cognition relationship are summarized in Table 1. The view at the beginning of the twenty-first century appears to be best captured by the idea that cognition and language have complex similarities and differences, and both develop over the human life span from genetic factors constrained by environmental input and cultural learning. New possibilities for synthesis continue to emerge, especially as cognitive scientists pay more attention to evolutionary, neurobiological and cultural factors. It may be possible to set aside the question of whether language is distinct from cognition and whether the brain is composed of distinct mental modules. The theorist Howard Gardner has noted a growing consensus about the importance of a new set of questions about how to divide up the grand areas of mind and brain. Scientists are emphasizing the distinction between areas of human ability that are available to all humans and played a part in the evolution of our species (such as language and basic number use), and areas requiring cultural elaboration (such as algebra and the ability to play musical instruments). The era of simplistic statements about the language–cognition relationship is drawing to a close, as cognitive scientists begin to deliver on the promise of a truly interdisciplinary approach to understanding the mind–brain.

Further Reading

Baars BJ (1986) *The Cognitive Revolution in Psychology*. New York, NY: Guilford Press.

Dromi E (ed.) (1993) *Language and Cognition: A Developmental Perspective*. Norwood, NJ: Ablex.

Edelman GM (1992) *Bright Air, Brilliant Fire*. New York, NY: Basic Books.

Elman JL, Bates E, Johnson MH *et al*. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Gardner H (1985) *The Mind's New Science: A History of the Cognitive Revolution*. New York, NY: Basic Books.

Gumperz JJ and Levinson SC (1996) *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.

Harris CL (1990) Connectionism and cognitive linguistics. *Connection Science* **2**: 7–34.

Lakoff G and Johnson M (1999) *Philosophy in The Flesh*. New York, NY: Basic Books.

Langacker RW (1987) *Foundations of Cognitive Grammar*, vol. 1, Theoretical Prerequisites. Stanford, CA: Stanford University Press.

MacWhinney B (ed.) (1999) *The Emergence of Language*. Mahwah, NJ: Erlbaum.

Morris RGM (ed.) (1989) *Parallel Distributed Processing: Implications for Psychology and Neuroscience*. Oxford, UK: Oxford University Press.

Muller RA (1996) Innateness, autonomy, universality? Neurobiological approaches to language. *Behavioral and Brain Sciences* **19**: 561–610.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pinker S (1994) *The Language Instinct*. Cambridge, MA: MIT Press.

Rumelhart D and McClelland J (1986) *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, vols 1 and 2. Cambridge, MA: MIT Press.

Language Comprehension and Verbal Working Memory

Introductory article

Gloria S Waters, Boston University, Boston, Massachusetts, USA
David Caplan, Massachusetts General Hospital, Boston, Massachusetts, USA

CONTENTS

Working memory is considered to be a specialized component of memory that is responsible for the temporary storage and manipulation of information necessary to accomplish a cognitive task. Language processing is an example par excellence *of a task that requires temporary storage and manipulation of information, and that therefore provides a domain in which ideas about working memory can be explored.*

OVERVIEW OF WORKING MEMORY

Alan Baddeley and his colleagues are generally credited with introducing the concept of 'working memory' into contemporary cognitive psychology and neuropsychology. They argued that the concept of 'short-term memory' (STM) as a system that retains information for short periods of time was too restrictive. They suggested that a combination of a short-duration storage function and a computational capacity was necessary for performing many mental functions. Thus was born 'working memory' (WM). Baddeley and his colleagues went on to elaborate a model of working memory that postulated the existence of a 'central executive' responsible for both computations and information storage, and 'slave systems' responsible for information storage when the capacity of the central executive was exceeded (see Figure 1).

It is now commonly accepted on the basis of both behavioral and neurological evidence that there are different working memory systems for verbal and nonverbal (e.g. visuospatial) tasks. There is considerable controversy concerning whether there are further subdivisions within each of these systems.

WORKING MEMORY AND LANGUAGE COMPREHENSION

The concept of a limited resource system that determines computational capacity and efficiency is a powerful one with many potential applications, and it has been widely utilized. Language is a natural domain in which this concept might apply, because virtually all language processing – from recognizing spoken words on the basis of their phonemic content to participating in a conversation – requires storage of representations and performing computations on these representations over some period of time.

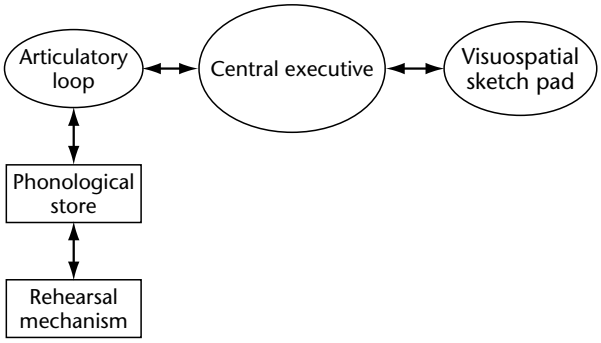


Figure 1. A schematic of Baddeley's model of working memory. Working memory is composed of the central executive, which is responsible for computations and information storage, and slave systems, such as the articulatory loop and the visuospatial sketch pad, which are responsible for information storage when the capacity of the central executive is exceeded.

Measurement of Working Memory Capacity

To relate WM to language, we have to be able to measure both the capacity of the WM system and the speed and accuracy of various language functions. A fair number of WM tests have been developed (see Figure 2).

Probably the most widely used is the 'reading span' measure. In this task, subjects read aloud increasingly longer sequences of sentences and

Daneman and Carpenter Reading Span Task

Subject reads aloud the following:
I imagine that you have a shrewd suspicion of the object of my early visit.
I'm not certain what went wrong but I think it was my cruel and bad temper.
Filled with these dreary forebodings, I fearfully opened the heavy wooden door.

Subject recalls sentence-final words:
visit, temper, door

Waters and Caplan Sentence Span Task

Subject reads the following sentences silently, making an acceptability judgment after each. Reaction times for judgments are measured.

It was the gangsters that broke into the warehouse.	*Acceptable*
It was the pillow that clenched the man.	*Unacceptable*
It was the cat that climbed the tree.	*Acceptable*

Subject recalls sentence-final words:
warehouse, man, tree

Alphabet Span Task
Experimenter reads aloud the following:
tree, arm, vest

Subject recalls the following:
arm, tree, vest

Subtract 2 Span Task
Experimenter reads aloud the following:
3-7-4

Subject recalls the following:
1-5-2

Running Item Span Task
Experimenter reads aloud the following:
1-7-9-3-4-5-6-2-8
'*Now tell me the last three numbers.*'

Subject recalls the following:
6-2-8

Figure 2. Sample items from several tests of verbal working memory (span size = 3).

recall the final word of all the sentences in each sequence. A subject's working memory capacity is defined as the longest list length at which he or she can recall the sentence-final words on the majority of trials. This task is thought to involve both a processing and a storage component, and therefore is thought to be a better measure of WM capacity than STM tasks, such as digit span, that simply require the recall of a list of items and so do not have a processing component.

There are close to a dozen WM tasks to choose from today (see Figure 2 for examples). Many of these tasks are similar to the reading span task in that the processing operation involves sentences. However, other types of WM tasks have been developed that require quite different processing operations, for example, to arrange a list of words alphabetically. Thus, while these different tasks share the characteristic of requiring both storage of some information and some sort of computation, they differ quite radically in their specifics. These differences might lead to different estimates of WM capacity. Indeed, studies of these tasks show that they are not always well correlated with one another. In addition, one cannot always be confident in the stability of a measurement of WM over time in an individual subject.

Thus we are at the uncomfortable point of having many WM tasks to choose from, which correlate only modestly with one another and are often unstable over time, and on which it is often unclear what cognitive functions determine a subject's performance.

CONTROVERSY OVER HOW/WHETHER WORKING MEMORY CONSTRAINS LANGUAGE PROCESSING

Researchers have nonetheless studied the relationship of WM to performances on many language

tasks and to language impairments, and found that measures of WM capacity have predictive value both for performances of normal subjects and for the presence of abnormalities of language functions. The simplest, and most common, account of these relationships is that WM is involved in language functions and that differences in WM capacity give rise to differences in performances on language tasks.

Given the above concerns about what tests of WM measure, however, it is advisable to examine the relationship between performance on WM tests and performance on tests of language functions more closely, and to ask whether any such relationships reflect the reliance of a language function on the WM system measured by these WM tests.

Measurement of Language Functions

One starting point for this examination is to deconstruct tests of language functioning, as we have those of WM. There are two ways to do this. One is to consider language tests in relationship to the nature of the linguistic representations that they require. This approach looks to linguistic analyses that describe the structure of language as a multilevel code that relates meanings to forms, where the levels include lexical, morphological, sentential, and discourse representations. The second is to consider these tests in relationship to the psychological processes they involve. For example, one could describe the processing requirements of language tasks in terms of their computational and memory requirements, and relate WM capacity to task performance as a function of these measures.

Relationship between Language Functions and Working Memory Capacity: Evidence from Individual Differences and Brain-damaged Populations

Beginning with representations, the strongest case for a relationship between a specific type of linguistic representation and WM has been made by Gathercole and her colleagues in the area of lexical phonological representation. They have argued that the 'articulatory loop' – the part of the verbal WM 'slave system' responsible for rehearsing phonological representations – is involved in lexical phonological processing. More specifically, these researchers have made the case that the articulatory loop is involved in learning new words.

On the surface, this pattern is surprising, since one might expect *a priori* that processing lexical phonological representations, which are made up of phonemes and distinctive features that are adjacent to one another and that extend over a relatively small temporal interval, would be expected to require relatively little WM capacity. This invites a deeper look at this finding.

Such a look shows that the evidence for a relationship between the articulatory loop and new word learning consists of high correlations between subjects' performances in repeating nonwords and in learning new words. It is not surprising that the ability to repeat nonwords is predictive of the ability to learn new words since, at the time they are first presented, new words are like nonwords with respect to their phonological familiarity. It is also unsurprising that the ability to repeat nonwords is related to the ability to rehearse. This suggests that it is the considerable overlap between the representations that need to be processed in a particular task (repetition) that is the basis of the relationship between the articulatory loop and new word learning.

This analysis suggests that the relationship between this one component of the WM system and this language function is not that the capacity of this part of the WM system determines the efficiency of an aspect of language functioning, but rather goes in the opposite direction. The capacity to repeat – a language function that partially underlies learning new phonological representations – is utilized by the WM system. WM recruits language, rather than vice versa.

Where WM tasks and measures of language function do not both emphasize processing the same linguistic representations in the same task, the relationship between WM and language processing is hard to demonstrate. Processing syntactic representations is an interesting case, because syntactic relationships are often established over elements that are much more discontinuous than the phonological units that constitute words, and their processing requires considerable computational activity. For example, working memory would seem to be needed to relate 'the boy' not 'the girl' to 'fell' in a sentence such as 'The boy who chased the girl fell'. Nonetheless, in many studies performance on WM tests does not correlate with performance on tests of the ability to process complex syntactic structures. For example, patients with dementia of the Alzheimer type who have severely reduced WM capacity have in many cases been found to have little difficulty in structuring syntactically complex sentences. The same can be said for populations with less extreme reductions in WM capacity – such as elderly individuals and college students with low WM spans.

Looking at the relationship between WM and language functions in psycholinguistic terms gives us another perspective. As noted above, there are many studies that document significant correlations between performance on a WM task and language functions. In many studies moderate correlations are found between measures of verbal WM and 'global, standardized' measures of comprehension (such as the verbal Scholastic Achievement Test (SAT) or the Nelson–Denny Reading test). Moderate correlations have also been found between measures of verbal WM and more specific tests of integration such as making inferences or abstracting the main theme. However, although it is clear from this literature that there are reliable correlations between many psycholinguistic functions and WM performance, it is not clear that WM capacity will predict performance on many of these tasks as a function of their storage and computational requirements.

Finally, we can look at psycholinguistic tasks in a qualitative manner. What is interesting about the language tasks on which performance bears a relationship to a measurement of WM capacity is the extent to which these tasks require conscious, controlled processing of language. It may be important to draw a distinction between those processes that are responsible for the recovery of linguistic forms and meanings from the acoustic signal to yield a meaning for an utterance, and those processes that use the products of language comprehension. The first set of processes, which is sometimes referred to as 'language interpretation', is largely unconscious, obligatory once initiated, fast, and accurate. The second set, which is sometimes referred to as 'post-interpretive processing', is more accessible to conscious awareness, not obligatory, and often slower than language processing. In general, performance on WM tasks is poorly correlated with the efficiency of interpretive processing and better correlated with that of post-interpretive processing.

The lack of a correlation between performance on WM tests and the efficiency of interpretive processing is not due to the fact that interpretive processing is so automatic that it does not require processing resources. Nor is it the case that interpretive processing does not require temporary storage and manipulation of information. Rather, the information storage and manipulation requirements of interpretive processing appear to be supported by a system that is not well measured by WM tests. This suggests that there may be several systems devoted to the temporary storage and manipulation of linguistic representations, one of which is responsible for language interpretation.

A specialization within verbal WM may have developed to support language interpretation because of a combination of factors. One is that language interpretation involves a set of related operations that always compute items within the same restricted set of representational types. A second is that these operations have the qualities of obligatoriness, unconscious operation, and speed noted above. A third is that these operations are among the most highly practiced of human cognitive functions. Because of their integration and degree of over-practice, one system for storage and manipulation of linguistic representations may be utilized by all the processes that constitute the interpretation process.

Performance on post-interpretive processing and WM tasks correlates much better than performance on interpretive processing and WM tasks. Post-interpretive processing and WM tasks share the basic psychological characteristics of being conscious, controlled functions. Such functions involve a large number of cognitive operations, including the ability to sustain, focus, and shift attention, to retrieve relevant information from long-term semantic and episodic memory, to reason, to assess likelihood, and others. The ability to store and manipulate linguistic and other symbolic representations – at least to do so consciously – is one functional ability that is found in both WM tests and these tasks; but the extent to which it, rather than other shared capacities, is responsible for the relationship between performance on the two types of tasks remains to be explored. Certainly, the simplest view, that differences in WM – the capacity to store and manipulate linguistic and other symbolic representations – are responsible for differences in performance on language tasks, needs to be replaced by a much more nuanced and detailed analysis of the determinants of performance on both these types of tasks.

Further Reading

Baddeley AD (1986) *Working Memory*. New York, NY: Oxford University Press.

Baddeley AD, Gathercole S and Papagno C (1998) The phonological loop as a language learning device. *Psychological Review* **105**: 158–173.

Caplan D and Waters GS (1990) The role of short-term memory in language comprehension: a critique of the neuropsychological literature. In: Shallice T and Vallar G (eds) *Neuropsychological Impairments of Short-term Memory*, pp. 337–389. Cambridge, UK: Cambridge University Press.

Caplan D and Waters GS (2000) Sentence comprehension in Alzheimer's disease. In: Connor L and Menn LK

(eds) *Neurobehavior of Language and Cognition: Studies of Normal Aging and Brain Damage*, pp. 61–76. Boston, MA: Kluwer Academic.

Daneman M and Carpenter P (1980) Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* **19**: 450–466.

Daneman M and Merikle PM (1996) Working memory and language comprehension: a meta-analysis. *Psychological Bulletin and Review* **3**: 422–433.

Gathercole SE and Baddeley AD (1993) *Working Memory and Language*. Hillsdale, NJ: Lawrence Erlbaum.

Just MA and Carpenter PA (1992) A capacity theory of comprehension: individual differences in working memory. *Psychological Review* **99**(1): 122–149.

Waters GS and Caplan D (1999) Verbal working memory capacity and on-line sentence processing efficiency in the elderly. In: Kemper S and Kliegl R (eds) *Constraints on Language: Aging, Grammar, and Memory*, pp. 107–136. Boston/Dordrecht/London: Kluwer Academic.

Wingfield A, Waters GS and Tun P (1997) Does working memory work in language comprehension? Evidence from cognitive neuroscience. In: Raz N (ed.) *The Other Side of the Error Term: Aging and Development as Model Systems in Cognitive Neuroscience*, pp. 319–394. Amsterdam, Netherlands: North-Holland.

Language Development, Critical Periods in

Intermediate article

EL Newport, University of Rochester, Rochester, New York, USA

CONTENTS

First language acquisition typically occurs in infancy and early childhood. An important question concerns whether the acquisition of a first or a second language shows a critical or sensitive period: that is, whether acquisition displays a normal course and leads to full proficiency in the language only when it begins early in life.

INTRODUCTION

In many species, including humans, important and species-typical behaviors develop through an intricate combination of innate and experiential factors. One hallmark of such systems is the appearance of a critical or sensitive period for normal development.

A critical period is a maturational time period during which some crucial experience will have its peak effect on development or learning, resulting in normal behavior attuned to the particular environment to which the organism has been exposed. If the organism is not exposed to this experience until after this time period, the same experience will have only a reduced effect, or in extreme cases may have no effect at all. Well-studied examples of species-typical behaviors showing peak plasticity within a critical or sensitive period include the identification of a species member as an attachment object (called 'imprinting') in ducks and birds, the acquisition of the species mating song by finches and sparrows, and the spatial tuning of auditory localization in barn owls. In contrast, in other domains and systems, there may be plasticity uniformly throughout life (open-ended learning), or plasticity may increase with age as experience or higher-level cognitive skills increase.

In his seminal book *Biological Foundations of Language*, Eric Lenneberg (1967) hypothesized that human language acquisition was an example of biologically constrained learning, and that it was normally acquired during a critical period, beginning early in life and ending at puberty. Outside of this time period, he suggested, language could be acquired only with difficulty or by a different learning process. He also suggested a neural mechanism for this developmental change: he hypothesized that the critical period for language acquisition ended with the establishment of cortical lateralization of function, as the brain reached its mature organization in late puberty.

Since the time of Lenneberg's book, an extensive research literature has asked whether there is indeed a critical or sensitive period for human language acquisition. These studies have provided strong support for the existence of such a critical or sensitive period (particularly for acquiring the phonology and grammar of language), though not for Lenneberg's specific hypothesis about the relationship between lateralization and the end of the critical period.

The term 'critical period' is sometimes used when there is an abrupt decline in plasticity and no residual plasticity after this period is over, whereas the term 'sensitive period' is used when there is a more gradual decline and some (reduced) plasticity remaining throughout life. However, recent research has shown that most critical periods show more gradual offsets and more complex interactions between maturational and experiential factors than the original concept of a critical period had anticipated. The terms are therefore often used interchangeably, as will be done in the present article.

EVIDENCE FOR A CRITICAL OR SENSITIVE PERIOD FOR LANGUAGE ACQUISITION

A number of lines of research, both behavioral and neural, suggest that there is a critical or sensitive period for language acquisition. Case studies of

individual feral or abused children, isolated from exposure to their first language until after puberty, have shown extreme deficits in phonology, morphology, and syntax resulting from this deprivation. The best studied of these cases is a girl named Genie, who was followed closely for a number of years after her discovery and placement in a normal linguistic environment at age 13 (Curtiss, 1977). While Genie did successfully acquire some English after puberty, her phonology was abnormal, and her control over English syntax and morphology was limited to only the simplest aspects of the language.

However, in cases of isolated children, general physical and cognitive status may be a concern. In studies of populations of normal individuals, one can systematically examine proficiency in relation to age of linguistic exposure without concern about the physical status of the learning brain. These studies show a strong relationship between the age of exposure to a language and the ultimate proficiency achieved in that language (Johnson and Newport, 1989; Krashen *et al*., 1982; Long, 1990; Newport, 1990), though typically with many fewer extreme deficits in adult learning than those found in the case studies of isolated children. Learning during the first months or year of exposure may show an advantage for adult learners, particularly in the acquisition of vocabulary and the speed of using certain complex sentence forms; however, long-term outcome clearly favors those who start learning the language during childhood. Peak proficiency in the language, in control over the sound system as well as the grammatical structure, is displayed by those whose exposure to that language begins in infancy or very early childhood. Such early learners show not only flawless control over the accent and rhythm of the language but also full and productive control over the syntax and morphology. With increasing ages of exposure there is a decline in average proficiency, beginning as early as ages 4 to 6 and continuing until proficiency plateaus for adult learners (Johnson and Newport, 1989; Newport, 1990). Learners exposed to the language in adulthood show, on average, a lowered level of performance in many aspects of the language, though individual variation also increases with age (Johnson and Newport, 1989), and some individuals may approach the proficiency of early learners (Birdsong, 1992).

These effects have been shown for both first and second languages, and for measures of proficiency including degree of accent, production and comprehension of morphology and syntax, grammaticality judgments for morphology and syntax, and syntactic processing speed and accuracy. For example, Johnson and Newport (1989) have shown that Chinese or Korean immigrants who move to the United States and become exposed to English as a second language show strong effects of their age of exposure to the language on their ability to judge its grammatical structure many years later, even when the number of years of exposure is matched. These effects are not due merely to interference of the first language on the learner's ability to acquire the second language: deaf adults, acquiring American Sign Language as their primary language, show effects of age of exposure on their grammatical skills in ASL as much as 50 years later, even though they may not control any other language with great proficiency (Newport, 1990; Mayberry and Eichen, 1991).

While there are effects of age of acquisition on both first and second languages and on both spoken and signed languages, an important question is how these effects compare. Does the acquisition of a language early in life reduce the effects of age on later language learning? This question has been examined by comparing hearing and deaf individuals' acquisition of English or American Sign Language as either a first or a second language, and (if as a second language) after early exposure to either a spoken or a signed language (Mayberry *et al*., 2002). The results show that age of first language onset has a significant effect, while language modality does not: late first language acquisition results in lower performance than does late second language acquisition, regardless of whether the languages in question were spoken or signed. According to one recent finding, even overhearing a language during early childhood, without producing it or hearing it again for many years, can result in learning to pronounce that language with a more native accent as an adult (Au *et al*., 2002).

However, age of exposure does not affect all aspects of language learning equally. The acquisition of vocabulary and semantic processing occur relatively normally in late learners. Critical period effects thus appear to focus on the formal properties of language (phonology, morphology, and syntax) and not the processing of meaning. Even within the formal properties of language, though, various aspects of the language may be more and less dependent on age of language exposure. For example, late learners acquire the basic word order of a language relatively well, but more complex aspects of grammar show strong effects of late

acquisition (Johnson and Newport, 1989; Newport, 1990). Further research is needed to characterize the structures that do and do not show strong effects of age of learning.

Age of exposure also affects the way language is represented in the brain, with similarities between the behavioral and neural effects. PET (Positron Emission Tomography), fMRI (functional magnetic resonance imaging), and ERP (event-related potential) studies all show strong left hemisphere activation for processing the native language, in bilinguals as well as monolinguals. However, when second languages are learned after age seven, the regions and patterns of activation are partially or completely nonoverlapping with those for the native language. Neural organization for late-learned languages is less lateralized and, like proficiency itself, displays a high degree of variability from individual to individual (Perani *et al*., 1996; Weber-Fox and Neville, 1996; Kim *et al*., 1997). The few studies that have observed early bilinguals or highly proficient late bilinguals report congruent results for native and second languages (Perani *et al*., 1998), though more refined techniques in the future might be expected to show neural differences whenever there are behavioral differences.

As with linguistic behavior, there is considerable specificity in these neural effects. In particular, age of acquisition appears to have more pronounced effects on grammatical processing and its representation in the brain than on semantic processing (Weber-Fox and Neville, 1996). When native speakers of English, respond to the appropriateness of open-class content words, ERP components distributed over the posterior regions of both hemispheres; and these same patterns appear in Chinese–English bilinguals who have acquired English as late as age 16. In contrast, when judging English syntactic constructions or responding to the placement of closed class function words in sentences, only early learners show the characteristic anterior left hemisphere ERP components; learners with delays of even 4 years show significantly more bilateral activation (Weber-Fox and Neville, 1996). Similar effects appear for signed languages (Neville *et al*., 1997).

Taken together, these results provide fairly strong evidence for a critical or sensitive period in acquiring the phonological and grammatical patterns of the language and in organizing the neural mechanisms for handling these structures in a proficient way. Nonetheless, the question of whether there is a critical period for language acquisition continues to be controversial.

QUESTIONS CONCERNING A CRITICAL OR SENSITIVE PERIOD FOR LANGUAGE ACQUISITION

Several questions have been raised about whether these age effects represent the outcome of a critical or sensitive period, or whether they might arise from variables correlated with age but not with maturation. One set of questions concerns whether the behavioral function has the correct shape for a critical or sensitive period. Must a critical period involve an abrupt decline and a total loss of plasticity at the end? Some investigators have argued that, in order to support a critical period hypothesis, age effects must coincide with the onset of puberty (though neural maturation continues throughout the teenage years and does not cease at ages 12 to 13). Other investigators have suggested that, if there were a critical or sensitive period for acquisition, no adult learners should achieve native proficiency. Finally, investigators have noted that it is difficult to distinguish a critical or sensitive period for learning from an interference effect.

However, many of the strong or absolute characteristics expected or demanded by these investigators are not true of critical or sensitive periods in other domains. Critical or sensitive periods in most behavioral domains involve gradual declines in learning, with some (reduced but not absent) ability to learn, and greater individual variation, in mature organisms. Critical periods in other domains also exhibit more learning during the waning portion of the critical period if the organism is presented with extremely salient or strongly preferred stimuli, or with learning problems similar to those experienced early in life. It should therefore not be surprising that a critical period for language in humans would show some continuing ability to learn, with individual variation, during adulthood. If such complex phenomena are routinely found within critical periods in other domains, they should also be expected for language learning.

References

Au TK, Knightly LM, Jun S-A and Oh JS (2002) Overhearing a language during childhood. *Psychological Science* **13**: 238–243.

Birdsong D (1992) Ultimate attainment in second language acquisition. *Language* **68**: 706–755.

Curtiss S (1977) *Genie: a Psycholinguistic Study of a Modern-Day 'Wild Child'*. New York, NY: Academic Press.

Johnson JS and Newport EL (1989) Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* **21**: 60–99.

Kim KHS, Relkin NR, Lee K-M and Hirsch J (1997) Distinct cortical areas associated with native and second languages. *Nature* **388**: 171–174.

Krashen SD, Long MH and Scarcella RC (1982) Age, rate, and eventual attainment in second language acquisition. In: Krashen S, Scarcella RC and Long M (eds) *Child–Adult Differences in Second Language Acquisition*, pp. 161–172. Rowley, MA: Newbury House.

Lenneberg EH (1967) *Biological Foundations of Language*. New York, NY: John Wiley.

Long M (1990) Maturational constraints on language development. *Studies in Second Language Acquisition* **12**: 251–285.

Mayberry RI and Eichen E (1991) The long-lasting advantage of learning sign language in childhood: another look at the critical period for language acquisition. *Journal of Memory and Language* **30**: 486–512.

Mayberry RI, Lock E and Kazmi H (2002) Linguistic ability and early language exposure. *Nature* **417**: 38.

Newport EL (1990) Maturational constraints on language learning. *Cognitive Science* **14**: 11–28.

Neville HJ, Coffey SA, Lawson DS, Fischer A, Emmorey K and Bellugi U (1997) Neural systems mediating American Sign Language: effects of sensory experience and age of acquisition. *Brain and Language* **57**: 285–308.

Perani D, Dehaene S, Grassi F, Cohen L, Cappa SF, Dupoux E, Fazio F and Mehler J (1996) Brain processing of native and foreign languages. *Neuroreport* **7**(15–17): 2439–2444.

Perani D, Paulesu E, Galles NS, Dupoux E, Dehaene S, Bettinardi V, Cappa SF, Fazio F and Mehler J (1998) The bilingual brain: proficiency and age of acquisition of the second language. *Brain* **121**: 1841–1852.

Weber-Fox C and Neville HJ (1996) Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience* **8**: 231–256.

Further Reading

Birdsong D (ed.) (1999) *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum.

Knudsen EI (1999) Early experience and critical periods. In: Zigmond MJ *et al*., *Fundamental Neuroscience*, pp. 637–654. San Diego, CA: Academic Press.

Newport EL, Bavelier D and Neville HJ (2001) Critical thinking about critical periods: perspectives on a critical period for language acquisition. In: Dupoux E (ed.) *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*. Cambridge, MA: MIT Press.

Neville HJ and Bavelier D (2000) Specificity and plasticity in neurocognitive development in humans. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, pp. 83–98. Cambridge, MA: MIT Press.

Latent Semantic Analysis

Advanced article

Thomas K Landauer, University of Colorado, Boulder, Colorado, USA

CONTENTS

Latent semantic analysis (LSA) approximates human understanding of relations between word and passage meanings in a wide variety of ways.

DERIVING SEMANTIC RELATIONS FROM LARGE TEXT CORPORA

Literate adults know the meaning of many tens of thousands of words. It is clear that much of this knowledge comes from reading, if only because a large fraction of the words that adults read are rarely if ever used in oral communication. However, before the mid-1990s, no artificial method could simulate the human ability to extract word meaning directly from text. The first successful model, latent semantic analysis (LSA), was made possible by large electronic text corpora, advanced computational power, and newly sophisticated algorithms (Landauer and Dumais, 1997). (LSA is also known as latent semantic indexing (LSI) in information retrieval applications.) LSA is not a complete model of language: it ignores word order within passages, and cannot produce sentences. Nevertheless, it approximates human understanding of relations between word and passage meanings remarkably well and in a wide variety of ways.

Computational Basis of LSA

The meaning of a passage must be a function of the meanings of its words and context. If m is meaning:

$$m(\text{passage}_i) = f(m(\text{word}_{i_1}), m(\text{word}_{i_2}), \ldots m(\text{word}_{i_n}), m(\text{context}_i)) \quad (1)$$

LSA divides a corpus into passages and treats them as a system of simultaneous equations, which it solves for the meanings of the words. To make this feasible, LSA ignores context and assumes that passage meaning is a linear sum of word meanings:

$$m(\text{passage}_i) \approx m(\text{word}_{i_1}) + m(\text{word}_{i_2}) + \ldots + \text{m}(\text{word}_{i_n}) \quad (2)$$

It is interesting to note that 'composionality', a property of language often posited and debated in linguistics and computational linguistics, is expressed in a very strong and simple form in eqn 2. The representation of a passage is composed by the sum of representations of its components.

Computing LSA by Singular Value Decomposition

LSA is given a corpus of text as close as possible in size and kind to language sources from which simulated humans commonly learn large parts of their vocabulary. The resulting huge and 'ill-conditioned' system of equations is solved by singular value decomposition (SVD), a form of factor analysis. After SVD, each word or passage is represented by a vector in a 'semantic space' defined by hundreds of abstract dimensions. Any old or new passage is represented as the vector sum of its words, as in eqn 2 and Figure 1. Similarity of meaning between two words or passages is usually estimated by the cosine of the angle between their vectors. The representation of a word specifies its average effect on passages, in a way reminiscent of Wittgenstein's (1953) conjecture that linguistic meaning lies in the relations of all words to all others. LSA can also be formulated as a neural net model, but training is more difficult. A related model is HAL (Lund and Burgess, 1996). LSA is fundamentally different from semantic network models based on human judgments (e.g. Lenat and Guha, 1991) rather than machine learning.

DATA COMPRESSION

One effect of LSA is representational economy. LSA dimensions are orthogonal, so there is no

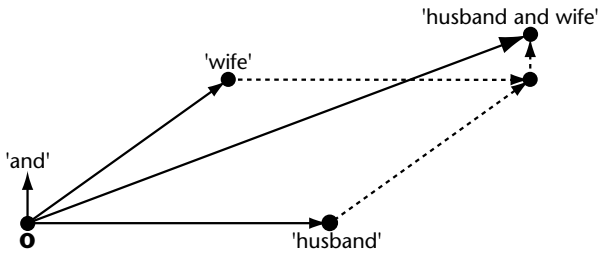


Figure 1. In LSA, passages are composed by the vector addition of words.

redundancy. Ignoring word order gains about another 20 percent. But the most important economy comes from dropping most of the dimensions: usually fewer than 500 are kept, rather than the tens of thousands that might be required to solve the original equations exactly. A typical paragraph requires only about a quarter of the bits of the original text.

The Value of Dimension Reduction

However, LSA is not used primarily for data compression, but for its powerful capacities for language representation. Through dimension reduction, similarities of the (roughly 98 percent of) word–word pairs that never co-occur in the same passage in a typical corpus are estimated. In one experiment, removing all such co-occurrences of tested words produced only modest decrements on a standard vocabulary test taken by LSA. The common belief that associative mechanisms cannot account for human learning of word meaning (e.g. Bloom, 2000; Chomsky, 1987) assumes a mechanism that depends entirely on local sequential or contextual co-occurrence. LSA extracts qualitatively different and quantitatively much greater amounts of associative knowledge.

The Quality of LSA Meaning Representation

Several of LSA's interesting properties are illustrated below. Table 1 shows cosine similarities ('cos') between representative pairs of words based on LSA learning from a 12.6 million word representative corpus of general English. Random word pairs have cos values in the range 0.02 to 0.03 with standard deviation in the range 0.06 to 0.12.

In LSA, different senses are not separately represented, and word forms are usually significantly related to the texts of all their dictionary definitions. Thus 'swallow' has a cos value of 0.57 with 'the process of taking food into the body through the mouth by eating', and a cos value of 0.30 with 'small long winged songbird noted for swift graceful flight and the regularity of its migrations'.

Table 1. Cosine similarities between pairs of words based on LSA learning from a 12.6 million word representative corpus of general English

Word pair		*cos*
love	life	0.35
man	woman	0.37
should	ought	0.51
chemistry	physics	0.65
sugar	sucrose	0.69
sugar	sweet	0.42
mouse	mice	0.79
come	came	0.71
doctor	physician	0.61
doctor	doctors	0.79
man	men	0.41
go	went	0.71
office	self	0.03

Because words combine their effects with their linguistic contexts, significant ambiguity does not necessarily arise from the fact that a single representation carries a complex of meanings. For example, adding the word 'swallow' itself to the two definitions above does not appreciably alter the LSA-meaning of either. (For the two definitions above, the cosine values are 0.99 and 1.00 between the definition and the definition prefixed with 'swallow'.)

Even short phrases containing a word with one of two apparently unrelated common meanings may show no appreciable distortion by the alternative meaning, as shown in Table 2.

In the first three pairs of phrases in Table 2, 'lead' has almost exactly the same effect as 'direct' when combining with 'army'; in the second three pairs of phrases it has very nearly the same effect as 'heavy' when combining with 'weight'. The last pair shows that the two contextual meanings involved are far from the same. Thus, the multiple meaning components of 'lead', which are not first 'disambiguated' by sense, do not interfere with its appropriate influence on meaning in the different contexts.

Another interesting property is that phrases sharing no words can sometimes have high similarities, while ones with many words in common can be dissimilar. Thus, 'the radius of spheres' has a cos value of 0.55 with 'a circle's diameter', but a cos value of 0.01 with 'the music of spheres'.

Table 2. Cosine similarities between pairs of phrases. The word 'lead' has two apparently unrelated common meanings

Phrase pair		*cos*
lead the army	direct the army	0.87
lead the army	desert the army	0.91
direct the army	desert the army	0.89
a lead weight	a heavy weight	0.92
a lead weight	a light weight	0.57
a heavy weight	a light weight	0.61
a heavy weight	direct the army	0.14

Correspondence with intuition is usually good for words and paragraphs, but is often poor for phrases and sentences, especially where local syntactic effects are large. The examples above serve to illustrate phenomena that can occur in LSA representations, but those phenomena do not always occur.

Quantitative evidence of how well LSA represents human meaning comes from simulations of human performance. For example, after training on general English, LSA matched the scores of successful college applicants from foreign countries on multiple-choice questions from the Test of English as a Foreign Language. After learning from an introductory psychology textbook it passed the same multiple-choice final exams as university students. Differences in knowledge between before and after reading a technical article, and between undergraduates and graduate students, were reflected more sensitively by measures based on LSA than by grades assigned by professional readers.

Of course, LSA must be applied judiciously. Modeled human performance must match its capabilities; and words and knowledge in the training corpus must match the human task. For example, LSA does not represent idioms correctly if they are not treated as single 'lexemes' in the input, and it will not represent text in one domain correctly if trained only on another domain. Misuse aside, it is clear that LSA has limited ability to simulate language comprehension: it is best suited to simulating semantic relations among single words and among paragraphs or longer texts, where syntactic and local context effects are minimized.

LSA AS A RESEARCH AND APPLICATION TOOL

LSA's approximation to human semantics is sufficiently good for many scientific and language engineering purposes.

Research Uses

Some researchers use LSA instead of human judgments of semantic similarity, free associations, or experimenter intuition in cognitive models and for equating verbal materials for experiments; others to measure recall accuracy or differences in semantic content of written responses. These uses have included, for example, models of sentential predication and metaphor (Kintsch, 2001).

Practical Applications

LSA is used in artificial intelligence systems to replace or augment 'handmade' lexicons, ontologies, and semantic rules with automatic evaluation of conceptual similarity between texts. For example, systems using LSA score essay content as accurately as professional readers, link people with jobs and training, support assessment and interactive dialogue in intelligent tutors (Hearst, 2000; Graesser *et al*., 2000), and form the basis of comprehension measures. LSA is used in content analysis and information retrieval. In the latter, it finds documents with similar meanings to a query despite containing different words, and is especially useful for routing and topic spotting tasks. LSA applications are language-independent, and with special forms of training can match documents in one language with those in another. Successful matching of paragraphs in Chinese characters with English translations, for example, gives additional evidence that LSA's additive model yields a good approximation of passage meaning.

LSA AS A THEORY OF COGNITION

LSA offers a computational model for investigating certain long-standing puzzles of cognition. One of these, called Plato's problem or the poverty of the stimulus, is that people have greater knowledge, in particular about language, than appears to be learnable from experience. Another is acquired similarity: what makes stimuli psychologically equivalent? These phenomena are often attributed to instinctive knowledge, rules, or primitive features. As exemplified in word meaning, which cannot be innate, LSA strongly suggests that an alternative explanation can be found in the form of a sufficiently powerful learning mechanism.

In LSA, similarity arises from experience, not from biologically defined relations among predetermined features. Inductive mechanisms like LSA's could apply in all kinds of recognition and

generalization: elementary variables could be neural events at any level and related in highly complex, perhaps holistic, ways. The psychological properties of LSA exhibit their power best in full-scale simulation of human learning from human-like quantities of experience. LSA's inferential mechanism evades Goodman's (1972) accusation that 'similarity is an imposter' by showing how it arises rather than using it as a primitive in explanations of other phenomena.

LSA has many human-like cognitive properties. It can learn nearly the same meaning or represent nearly the same idea (a point in semantic space, which may or may not correspond closely to any verbal statement) in an unlimited number of different ways. Meanings are slightly different for each person and in each context, yet are sufficiently stable and sharable to support communication and mutual comprehension. LSA's knowledge is implicit, resembling intuition rather than declarative propositions.

Philosophers and psychologists may be concerned that LSA has no contact with perception or action, the so-called 'grounding' that has often been thought to be essential to learning word meanings (Bloom, 2000). Given LSA's ability to mimic a variety of human linguistic performances, it can no longer be sensibly argued that nothing about language can be learned without such sources. Nonetheless, LSA certainly must miss something by its complete reliance on vicarious experience. An interesting question is whether the same basic mechanisms applied to experience with words in natural perceptual contexts would close the gap, or whether additional factors unique to human cognition make important contributions. Also, LSA does not model important aspects of meaning that depend on order, matters of syntax which have long been a primary focus of linguistics.

Overcoming these limitations is a major scientific challenge. However, even with these limitations, LSA's success, by showing that much human use of language can be mimicked without sources and processes that had been believed to be necessary, raises deep theoretical questions about the nature of language.

References

Bloom P (2000) *How Children Learn the Meaning of Words*. Cambridge, MA: MIT Press.

Chomsky N (1987) Language in a psychological setting. *Sophia Linguistica* **22**: 1–73.

Goodman N (1972) *Problems and Projects*. Indianapolis, IN: Bobbs-Merrill.

Graesser A, Wiemer-Hastings K, Wiemer-Hastings P, Kreuz R and the Tutoring Research Group (2000) AutoTutor: a simulation of a human tutor. *Journal of Cognitive Systems Research* **1**: 35–51.

Hearst MA (2000) The debate on automated essay grading. *IEEE Intelligent Systems and Applications* September/October: 22–37.

Kintsch W (2001) Predication. *Cognitive Science* **25**: 173–203.

Landauer TK and Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 11–140.

Lenat DB and Guha RV (1991) *Building Large Knowledge-Based Systems: Representation and Inference in CYC*. Reading, MA: Addison-Wesley.

Lund K and Burgess C (1996) Producing high dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers* **28**: 203–208.

Wittgenstein L (1953) *Philosophical Investigations*. Oxford, UK: Blackwell.

Further Reading

Christiansen MH and Chater N (1999) Connectionist natural language processing: the state of the art. *Cognitive Science* **23**: 417–437.

Collins AM and Loftus EF (1975) A spreading activation theory of semantic processing. *Psychological Review* **82**: 407–428.

Foltz P, Kintsch W and Landauer TK (1998) The measurement of textual coherence with latent semantic analysis. *Discourse Processes* **25**: 285–307.

Kintsch W (2001) *Comprehension, a Paradigm for Cognition*. New York, NY: Cambridge University Press.

Kitajima M and Polson PG (1997) A comprehension-based model of exploration. *Human-Computer Interaction* **12**: 345–389.

Landauer TK (1999) Learning and representing verbal meaning: the latent semantic analysis theory. *Current Directions in Psychological Science* **7**: 161–164.

Pustejovsky J (1995) *The Generative Lexicon*. Cambridge, MA: MIT Press.

Seidenberg MS (1997) Language acquisition and use: learning and applying probabilistic constraints. *Science* **275**: 1599–1603.

Learning and Memory, Models of

Introductory article

David M Eagleman, Salk Institute, La Jolla, California, USA
P Read Montague, Baylor College of Medicine, Houston, Texas, USA

CONTENTS

Learning is the process of acquiring knowledge, and memory is the retention of knowledge. Both have their roots in the biochemistry and anatomy of the brain, and are commonly studied by investigating the connection between brain cells, the synapse.

INTRODUCTION

Animals interact appropriately with the world through their ability to learn and remember the specifics of their environment. Learning and memory are possible only because of the changeable properties of the vast networks of cells in the brain. In this way, neural activity can dynamically modify the brain's organization. Learning and memory can be studied at many levels – biochemical, cellular and systems – and even by investigation of perturbations that decrease learning and memory, from genetic (Down syndrome) to pathological (Alzheimer disease).

What do we mean by learning and memory, and how do we detect when it happens? The answer seems complicated by the fact that memory is not a monolithic entity, but comprises many different types. Memory can be divided into declarative learning (names, facts) and nondeclarative learning (riding a bicycle), and within these categories are scores of subtypes. What neural processes could underlie such a variety of different types of memory?

Specialized Brain Areas

Examples of brain damage demonstrate that different areas of the brain are involved in different subtypes of learning and memory. For example, injury to the medial temporal lobe of the cerebral cortex affects declarative memory but not nondeclarative memory. The cerebellum, on the other hand, seems to be important in learning certain motor skills, especially those involving balance and coordination. The basal ganglia are evidently important in learning that links rewards with motor activity. The list of brain structures and their involvement with learning and memory is large and ever-increasing, and it is interesting to note that the integrity of a particular subsystem is not always essential to the functioning of others. That is, one can lose the ability to learn dates and facts (as in amnesia), but this has no bearing on the ability to learn and remember new motor skills. Much of what we will discuss here will be distilled from data concerning two main areas: the cortex and hippocampus.

The hippocampus (and its surrounding regions) seems to be a central organ of learning, and its structure makes its physiology very amenable to laboratory study; it is perhaps the best-studied area of the brain. In 1953, a 27-year-old patient referred to as HM had his hippocampus and surrounding areas surgically removed to relieve intractable epilepsy. Thereafter, HM lost his ability to form new memories or learn new facts, and although he could acquire new skills, he had no memory of having acquired them.

General Cellular Principles?

While learning and memory can be different in specialized structures, it is possible (although by no means proved) that general learning mechanisms underlie all these different types of learning. It is notable that the different brain areas have many properties in common: short-term and long-term memory, one-trial and multiple-trial learning, similar cellular mechanisms and biochemical pathways, and activation of particular memory genes. This suggests the possibility of general

learning principles at the cellular and subcellular levels.

Almost all current theories of learning and memory involve some variant of the idea that efficacy of the connections between cells can be modified based on their previous activity. Such theories urge us to seek biophysical – as well as computational – descriptions of what happens at the synapse. Although the synapse has received the most experimental attention, we should begin with the caveat that the full picture of learning and memory is largely unknown (the final section of this article reviews other forms of change in the brain that could have a role in learning and memory).

The Synapse

Over a century ago, the idea that neural tissue is a continuous network, or reticulum, was challenged by the proposal that the nervous system is an intricate network of discrete cells. The great Spanish neuroscientist, Ramón y Cajal, proposed this 'neuron doctrine', which ushered in an important new idea: separate cells influence each other primarily through specialized connections called synapses. Rámon y Cajal is credited as the first to suggest that learning and memory might occur by changes in the connections between neurons.

FORMS OF SYNAPTIC PLASTICITY

What is plasticity? A plastic system is one that is changeable, and able to retain change (hence, when a manufacturer moulds a cup out of a lump of plastic, it is useful only because it retains its new shape). A nonplastic system would be unable to store memories of anything, being unchanged by its experiences.

The brain comprises both plastic and nonplastic components. The brain's control of respiration and heartbeat is not thought to be plastic, just as a bird's knowledge of how to build a nest is not something learned, but hard-wired. However, many parts of the brain are plastic, and this is what allows an animal to learn about and interact with its environment. Born in one country, you might learn to forage for food from particular broad-leafed shrubs; in another, you might learn to satisfy your cravings from a refrigerator. This learning of the proper location for the food could only be accomplished because parts of your brain were plastic. In general, plasticity is much greater in the developing animal, and decreases with age.

Associative Learning

Many classic experiments in psychology demonstrate the role of association. Every student knows about Pavlov's dogs, who salivated when they heard the bell that signaled the presentation of meat powder. The behavioral psychologist B. F. Skinner found that particular stimuli caused an organism to repeat an act more frequently. He called stimuli with this effect 'reinforcers'. Other psychologists found that by providing reinforcement in a systematic way one could shape an animal's behavior in desired directions.

What underlies these fundamental learning paradigms is the notion of 'association' – in these cases, the association of the bell and the meat, or the association between a stimulus and reward, or between a particular behavior and a punishment. An appealing idea in neuroscience has been that if an association is established between two stimuli, perhaps there is a neural substrate that should directly reflect this; but what cellular mechanisms could possibly underlie association?

In 1949, the neuroscientist Donald Hebb outlined the following hypothesis:

> When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

In other words, if a presynaptic cell (A) consistently participates in driving a postsynaptic cell (B), the connection between them is strengthened (potentiated) (Figure 1). Hebb's hypothesis goes on to prescribe that if A consistently fails to drive B, the connection is weakened (depressed). Systems that display this behaviour are said to follow a Hebbian, or correlational, learning rule. This rule can be written as

$$\Delta w(t) = \lambda x(t) y(t) \qquad (1)$$

where, at time t, $w(t)$ represents the efficacy (weight) between two neurons, and $x(t)$ and $y(t)$ are measures of pre- and postsynaptic activity (a common measure is the rate of action potential generation, or 'firing rate'); λ is the learning rate, as it specifies how quickly the weights will react to changes in $x(t)$ and $y(t)$. The rule states that if the presynaptic and postsynaptic cells are both firing more than normal (i.e. are co-active; assume $x(t)$ and $y(t)$ represent firing rates above baseline, and hence are both positive numbers in this example), then the weight change will be positive (potentiation). If the activity $x(t)$ is high and $y(t)$ is low,

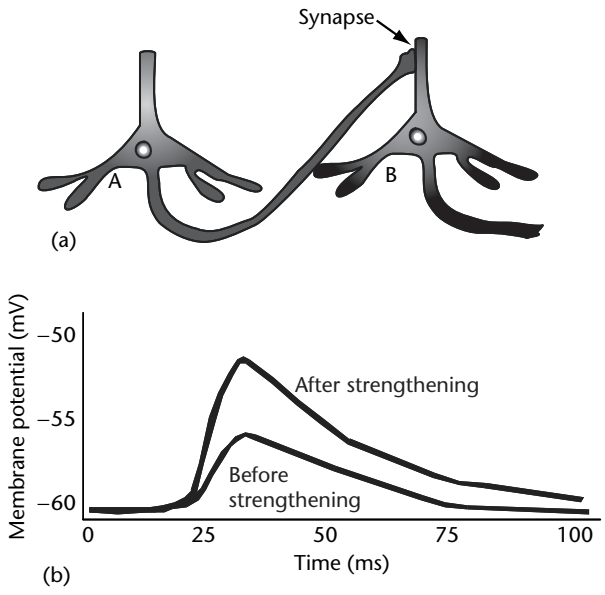


Figure 1. Learning and memory at a synapse. (a) According to Hebb's postulate, if neuron A consistently causes neuron B to fire, then the connection between them will strengthen. (b) When neuron A releases neurotransmitter on B, an excitatory postsynaptic potential (EPSP) can be measured in B. If long-term potentiation is induced by tetanic stimulation of A, the EPSP measured in B is now larger, so the connection has been strengthened.

then the connection between them will weaken. To date, most models of neural function employ such a rule.

At the time Hebb proposed his hypothesis, there was no direct experimental demonstration that could be marshaled for support. Then, in a famous experiment in 1973, researchers stimulated a bundle of nerve fibers in a rabbit's hippocampus. Trains of electrical stimulation were administered at 15 Hz for 10–15 s, and it was shown that one or two such exposures were enough to 'condition' (potentiate) an increased electrical response from the postsynaptic cell for up to 10 h. In other words, this was a direct demonstration that connections could be modified based on the history of the activity of the cells involved.

Although this experiment galvanized thousands of researchers in the ensuing decades, it is still unclear whether the results of this experiment completely and accurately represent the neural changes that take place during learning and memory. Our current understanding suggests that some of the stimuli employed in such experiments are unlikely to be seen in the real animal *in vivo*. However, other stimuli do appear to be biologically feasible; for example, it was later discovered that an optimal stimulus to induce this sort of modification is at a frequency of 5 Hz – which is the frequency of electrical rhythms seen in the hippocampus when animals explore a novel environment.

In general, the biological relevance of the correlational learning rule has been supported by decades of research into the detailed biophysical properties of synapses. It is clear from developmental neuroscience that appropriate neural activity is required for establishing the proper connections between brain areas. This developmental self-organization is consistent with the hypothesis that changes in synaptic efficacy are controlled by processes resembling Hebbian rules. However, such activity-dependent properties of cells are not limited to the developing animal: many experiments have now revealed the reorganizational plasticity of the adult brain. It is widely hypothesized that the cellular rules that account for large-scale self-organization and reorganization (in both the developing and adult brain) may be the same rules that account for learning and memory.

Different Timescales of Synaptic Plasticity

Animals display learning and memory over many different timescales; it may be significant that changes in synaptic efficacy, or strength, similarly occur over different timescales.

Short-term changes

Sometimes the modification of synaptic efficacy is short-term, lasting on the scale of seconds to minutes. The two most commonly studied examples of this are short-term potentiation and short-term depression (also known as fast synaptic depression). The former can be induced by increased levels of intracellular calcium in an axon terminal due to recent activity. This leads to a higher probability of neurotransmitter release with successive activity, and thus the connection between the cells is considered potentiated. Short-term depression can come about by a depletion of the readily releasable pool of neurotransmitter vesicles, which take time to be repackaged and docked. When high activity levels cause increased vesicle release, the terminal becomes temporarily less able to respond to future activity, and is thus considered depressed.

Such fast modifications in synaptic activity can cause rapid dynamic changes in the behavior of networks, modifying their function as a result of recent activity.

Long-term changes

Long-term potentiation

In the 1973 synaptic modification experiment mentioned above, the changes lasted for many hours, and thus the phenomenon was labelled 'long-term potentiation' (LTP). We now know that the phenomenon of LTP is common to many brain areas. It is studied most extensively in the neocortex and hippocampus, the latter because of its crucial role in memory formation.

In most cases, LTP is induced only when the activity in the postsynaptic cell (depolarization) is associated with activity in the presynaptic cell. Depolarization alone or presynaptic activity alone is ineffective. Additionally, LTP is synapse specific, which means that each individual synapse on a cell could, in principle, strengthen or weaken according to its own personal history (although this view is questioned by some).

What goes up must come down: long-term depression

If a connection is able to potentiate, it also needs the ability to depress, otherwise the system will become saturated, and be unable to store anything new. Long-term depression (LTD) is obtained by using low-frequency repetitive stimulation (e.g. 1 Hz instead of 15 Hz). After this conditioning, the connection between two cells is weakened. There are other ways to achieve LTD, as will be seen in the section on timing, below. Long-term potentiation and depression are part of the same phenomenon; LTD is found at the same synapses as LTP and also depends on the same mechanisms.

In general, it is suggested that LTP and LTD could mediate one-trial associative learning, since pairing of two events (presynaptic and postsynaptic activity) creates a long-term change in synapses. The biophysical mechanisms underlying long-term changes are still a subject of intense investigation, but the differences between LTP and LTD are mainly thought to involve differences in the concentration (and temporal dynamics) of postsynaptic calcium ions. Although it may turn out to be inadequate, a current hypothesis is that lower calcium concentrations lead to depression, whereas higher concentrations lead to potentiation.

The NMDA Receptor

For the induction of memory, one of the most important biophysical mechanisms is a subtype of glutamate receptor called the NMDA receptor (NMDA-R), so named because it is selectively stimulated by *N*-methyl-D-aspartate. How does this receptor work, and why does it enjoy such prominence in the research literature?

In most systems, the NMDA-R is crucial for induction of LTP. An animal can be taught a behavioral task, but with the infusion of NMDA-R antagonists, the ability to remember the specifics of the task disappears.

At resting potentials, the NMDA-R is blocked by magnesium. Depolarization of the postsynaptic cell expels the magnesium ions and opens up the channel. Many postsynaptic membranes contain NMDA as well as non-NMDA glutamate receptors. During normal low-frequency stimulation, only the non-NMDA channels will open, owing to Mg^{2+} blockage of the NMDA ion channels. In contrast, high-frequency presynaptic input resulting in depolarization of the postsynaptic membrane displaces the magnesium ions, making the NMDA receptors sensitive to subsequent release of glutamate. In this way, the NMDA-R can act as a coincidence detector, sensing coincidence of presynaptic and postsynaptic activity. Thus, NMDA synapses are the quintessential biological Hebbian synapses, and thus may be crucial to the storage of associations.

The fact that NMDA receptors have a particularly high permeability for calcium allows them to stimulate a second-messenger system that results in long-term structural changes to the postsynaptic cell. Interestingly, one cannot bypass the NMDA channel by depolarizing the cell to allow calcium influx. It is not simply the amount of calcium influx that matters, but also the exact spatial location: the influx must occur in close proximity to the NMDA receptors at the synapse. This highlights the specificity of computations that are performed on spatial scales smaller than we can resolve.

Traditionally, it has been thought that strong presynaptic input is sufficient for local depolarization of the membrane in which the NMDA-R sits, but we now know that when a cell generates an action potential, this potential can (under the right circumstances) propagate back into the dendritic tree. Thus, a back-propagating action potential could act as a global dendritic signal, depolarizing thousands of synapses (and their NMDA-Rs) at once.

Certain forms of LTP induction are known to depend on postsynaptic burst firing. Hippocampal burst firing is presumed to be associated with memory induction *in vivo* (it occurs during active exploration of novel environments).

Note that the NMDA-R is only necessary for the induction of most forms of LTP and LTD; other

mechanisms underlie the maintenance of the changes – most generally, new protein synthesis is required at the nucleus of the cell. An animal can be trained to associate two stimuli (say, pairing an electric shock with a bright light) in the short term, but if protein synthesis is blocked, no long-term memory develops.

How Do Synaptic Efficacy Changes Take Place?

The common observation with plasticity studies is that after conditioning, the presynaptic cell gives a stronger or weakened electrical input to the postsynaptic cell. Mechanistically, how does the strength of an individual synapse change?

A synapse, like most elements of a cell, has many modifiable parameters. Since we know the induction of most forms of conditioning to be dependent upon the NMDA-R, and since the NMDA-R is postsynaptic, this hints at postsynaptic changes. It is known that calcium ions flow in and trigger a postsynaptic biochemical cascade involving protein kinases, and that blocking specific kinases stops LTP: some kinases that have been implicated are calcium–calmodulin kinase II (CaMK-II) and protein kinase C (PKC). These cascades eventually lead to the genome, and to the synthesis of new proteins which solidify the changes – for example, by the expression of more postsynaptic neurotransmitter receptors. Recently, it has been shown that certain receptors appear to increase in concentration after LTP. Other structural changes (such as the formation of new dendritic spines) also appear to occur minutes after LTP induction.

Presynaptic changes and retrograde messengers

Changes at the presynaptic terminal could also lead to greater synaptic efficacy; for example, an increase in probability of transmitter release. There is a raging debate as to whether the changes take place pre- or postsynaptically, and as in most great debates in biology, the answer will probably turn out to be both.

Whatever the case, any presynaptic changes will of necessity require a signal passing back from the postsynaptic to the presynaptic side. Such a signal has been labeled a retrograde messenger. There are several good candidates for biologically realistic second messengers. For example, the influx of calcium (and the subsequent activation of CaMK-II) induces synthesis of nitric oxide (NO), the molecules of which are so small that they effortlessly diffuse through cell membranes to presynaptic terminals (thus acting as a retrograde messenger) to induce the presynaptic neuron to enhance transmitter release.

Very long-term storage (compression)

It is speculated that associations formed in a subset of hippocampal neurons (area CA3) are built into more economical form in another part of the hippocampus (CA1 neurons). The hippocampus itself may function by building more economical storage (long-term memory) in the cerebral cortex. In this view, the hippocampus functions as a transient intermediary in the formation of long-term memories in the association areas of the cerebral cortex. From the point of view of psychology, such long-term storage may go hand-in-hand with the observation that we improve our learning and perform tasks faster as they become proceduralized (i.e. with practice, a skill becomes faster until it is virtually automatic).

IMPORTANCE OF TEMPORAL ORDER IN SYNAPTIC PLASTICITY

Timing: More than Coincidence?

For a long time Hebb's rule was summarized as 'neurons that fire together, wire together'. In other words, if two cells fire within some small window of coincidence, the connection between them is strengthened. However, this rule turns out to be insufficient. Hebbian rules are good for forming associations, but one theoretical shortcoming is that such rules are insensitive to the order of events. Experiments have long shown that animals are strictly sensitive to the order of sensory inputs – such that, for example, Pavlov's dog will not learn an association if the meat is presented before the bell. Similarly, many animals develop strong aversion to a tasty food following a single experience of nausea after eating it, but reversing the order (nausea and then the food) does not lead to an aversion.

It is now becoming appreciated how the relative timing of presynaptic and postsynaptic activity may be crucial at the cellular level. An interesting rule has emerged about the timing of synaptic change, at least in some systems. If an input from cell A contributes to driving cell B, then the synapse is strengthened. If an input from A comes after cell B has fired, the synapse is weakened. The importance of this timing can be seen in Figure 2. This learning rule is commonly called a 'temporally asymmetric Hebbian rule', and has expanded our view of the importance of exact spike timing.

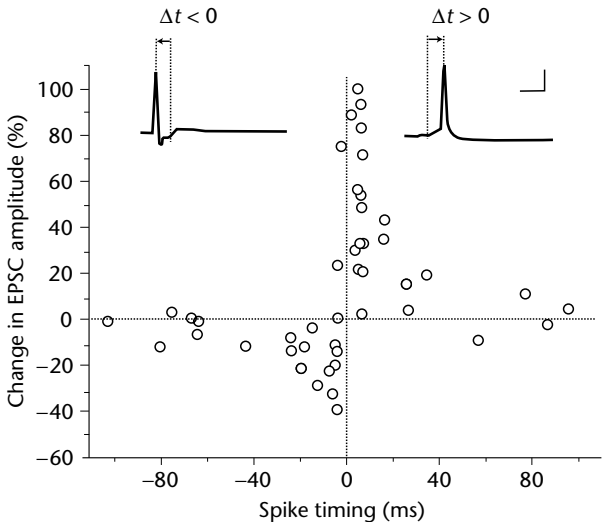


Figure 2. Timing for synaptic potentiation and depression. In this plot, reproduced from Bi and Poo (1998, *Journal of Neuroscience*), $\Delta t > 0$ means that the postsynaptic spike occurred after the presynaptic spike. In such cases the presynaptic spike has a role in the activation of the postsynaptic spike and the connection between the two neurons strengthens, as measured by the percentage change in the excitatory postsynaptic current (EPSC). If the postsynaptic spike arrives before the presynaptic spikes ($\Delta t < 0$), the connection between the two neurons weakens.

Most generally, the temporally asymmetric rule strengthens connections that are predictive; if A consistently fires before B, it can be viewed as a successful prediction, and will be strengthened. (As a caveat, it should be noted that the data in Figure 2 were obtained under specialized and perhaps biologically unrealistic circumstances. It is not yet known what would happen at a synapse with a more realistic barrage *in vivo* of pre- and postsynaptic spikes.)

THE ROLE OF CONTEXT IN LEARNING AND MEMORY

No discussion of learning and memory can take place in the absence of considerations of context – both the context into which new information can be stored, and the situational context of the animal.

Informational Context

Much of what we learn is in terms of what we already know. When you read this article, the points made here would be meaningless unless you were already equipped with an idea of how to operate a book, how to read this language, and what a brain is. Two people might look at a list of important dates in Mongolian history; if one of them already commands a richly developed cognitive model of Mongolia, the new facts are much more readily incorporated into that person's network of knowledge. At the simplest levels, this reflects associational mechanisms, as discussed above. By associating one new concept with others already known, it can be more readily stored and retrieved. Associative neural networks display this sort of property: items can prime webs of association, in the same way that the smell of coffee can immediately conjure associations with the sight of the black liquid, the sensation of warmth on the hands, the bitter taste, and so on.

Situational Context

Neural wiring can dictate which experiences result in learning and which do not. Many birds learn to form a strong emotional bonding at birth to any nearby distinctive and animate object – a process known as imprinting (this is a good example of one-trial learning). Many neural network models fail at such learning, requiring many thousands of trials for the learning of even simple tasks. More recent neural network models have therefore begun to include context as an important variable. The notion that particular types of learning are triggered to occur at certain times – 'schematic' learning – is an expression of an important philosophy that has emerged in neuroscience: the brain is not a *tabula rasa*, a blank slate upon which the world imprints itself. Instead, the brain comes preequipped for certain types of learning, and learning in particular situations. Experience is more likely to result in learning when it has relevance to the life of the organism – especially when it is connected to pleasure or pain, fear or satisfaction.

One way that context is likely to be expressed, biophysically, is through neuromodulatory systems, which are generally global neural systems that signal reward, punishment, alertness, and so on. Neuromodulatory systems have a crucial role not only in developmental plasticity, but also in the synaptic plasticity of learning and memory in the adult. Among other functions, neuromodulators can turn synaptic plasticity on and off, and influence the presynaptic/postsynaptic communication pathways. In this way the plasticity of synapses can be gated, so that learning takes place only at the appropriate time instead of each time activity passes through the cell – which could, in theory, overwrite previous learning. It has been demonstrated in the adult animal that reorganization of

parts of the cortex can occur only when paired with the release of particular neuromodulators.

Lastly, there has been a recent surge of physiological experiments in awake, behaving animals to study the role of attention. It was known by the ancient Greeks that learning and memory are most reliable when a student is paying attention. Recent experiments have shown that the firing rates of individual cells can be highly modulated by the attentional state of the animal. Since the number and timing of spikes seem to be important, it is easy to see how changes in firing rate could modify the dynamic changes in synaptic modification in the networks of cells.

OTHER FORMS OF PLASTICITY

Although synaptic transmission has been favored as the major means of communication between nerve cells, it is almost certainly an incomplete description. Signals of synaptic or nonsynaptic origin can diffuse through large volumes of neural tissue to affect signal-readers, even at distant sites. This mode of communication, 'volume signaling', can function between neurons and also between neurons and glial cells. The consideration of signaling that extends beyond the synapse allows the three-dimensional arrangement of neural elements to play a part in information processing. Such a hypothesis would predict that if one could reproduce the synaptic connections of the brain on a two-dimensional circuit board, the resulting machine would be insufficient for the functional simulation of neural tissue. This is increasingly becoming appreciated as researchers attempt to form models with explicit representations of the three-dimensional composition of neural tissue.

Although this discussion has centered on changes at chemical synapses, certain types of synapses are electrical. Electrical synapses, or 'gap junctions', are increasingly being discovered in the mammalian brain, especially within networks of inhibitory cells. In essence, they are ion channels running through one cell membrane to the cell membrane of an adjoining cell. Gap junctions permit rapid and bidirectional flow of ions between cells, and it is possible that they could exhibit plasticity or other properties that give synapses a central role in learning and memory.

There are many other possible substrates in which to store activity-dependent changes. Researchers are now studying what they call 'intrinsic' (or nonsynaptic) changes in cells: changes in the excitability of the cell, changes in the distribution of ion channels, changes in the shape of dendritic trees, changes in the phosphorylation states of intracellular proteins, and so on. With so many degrees of freedom in biological systems, the possibilities are vast for discovering different storage strategies for learning and memory.

Additionally, in many cases memory can be stored (at least short-term) in reverberating circuits of activity. For example, a form of short-term memory referred to as 'working memory' is associated with the ability to store contemporary representations of the outside world. The apparent locus for working memory is in the prefrontal area of the cerebral cortex. Monkeys shown that food is hidden behind an obscuring object, but who are restrained from immediately taking the food, maintain their knowledge of the whereabouts of the food, and will find it after a delay, once restraints have been removed. Monkeys who have had areas of their prefrontal cortex surgically removed will forget about the food as soon as they can no longer see it ('out of sight, out of mind'). The mechanisms that underlie this working memory may be in circuits of reverberating activity, which keep the information temporarily stored without the need for synaptic change.

CONCLUSION

Many features of neural plasticity in learning and memory seem consistent with Hebbian rules. This convergence of experimental and theoretical approaches provides a powerful example of the modern approach to neurobiology. Future experiments will continue to elucidate the extent to which synaptic learning rules account for the properties of learning and memory, and will contribute to expanding neural models to encompass issues of timing and context.

Further Reading

Bi GQ and Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience* **18**: 10464–10472.

Churchland PS and Sejnowski TJ (1995) *The Computational Brain*. Cambridge, MA: MIT Press.

Koch C (1998) *The Biophysics of Computation*. New York, NY: Oxford University Press.

Montague PR and Sejnowski TJ (1994) The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. *Learning and Memory* **1**: 1–33.

Learning and Memory, the Ecology of

Introductory article

Bennett G Galef Jr, McMaster University, Hamilton, Ontario, Canada

CONTENTS

Specialized learning and memory mechanisms have sometimes evolved in response to the particular demands of the ecological niches occupied by different species.

INTRODUCTION

For most of the hundred or so years that scientists have studied how animals learn and remember, investigations of these topics have been carried out without attention to differences in the environments in which different species evolved. For example, one of the twentieth century's foremost investigators of animal learning, B. F. Skinner, when discussing the behavior of animals given food rewards, expressed unambiguously the prevailing view of how learning should be studied: 'Pigeon, rat, monkey, which is which? It doesn't matter.'

Skinner, of course, realized that different animals behave in different ways. However, for half a century and more behavioral scientists focused on the many features of animal learning in laboratory situations (for example, in mazes or in boxes where levers had to be pressed to obtain food) that were similar in all species. Such an approach to the study of animal learning, emphasizing similarities across species, is often referred to as a 'general process' approach, because it is based on the assumption that there are one or two basic learning mechanisms (or general processes) that are responsible for all learning by all animals.

A different view of animal learning and memory proposes that specialized learning and memory processes have evolved in response to differences in the selective pressures acting on different species; after all, differences in the physical characteristics and sensory systems of animals are known to be products of natural selection acting in different ecological situations. There is no reason why learning and memory mechanisms should not also have evolved to respond to the different environmental demands faced by members of different species. According to this view, animals would not only learn and remember in general, but also learn and remember particularly well things that are especially important to their survival and reproduction in the natural world.

TASTE-AVERSION LEARNING

In 1966, John Garcia discovered that learning not to eat substances that are associated with illness might be very different from learning in other situations. Garcia was looking at the effects of radiation on feeding behavior, an issue of potential importance both for patients receiving radiation therapy and to the military whose members might have to survive in radioactive areas in the event of war.

Garcia was studying the feeding behavior of rats exposed to X-irradiation while eating a type of food that they had never before eaten. He discovered, quite unexpectedly, that even though the illness resulting from X-irradiation did not start until some time after the rats had finished eating, the rats later refused to eat that type of food a second time. In most situations, in order for an animal to learn to associate two stimuli (in this case the taste of the unfamiliar food and the illness resulting from exposure to radiation), the stimuli have to occur within tenths of a second of one another. Many pairings of stimuli are often needed before learning is observed. Garcia had discovered a situation in which learning of an association between two stimuli occurred in a single trial, and despite the fact that the stimuli to be associated were separated by many minutes. Indeed, later experiments were to show that learning of an aversion to an unfamiliar taste followed by illness could occur in one trial even when taste and illness were separated by several hours.

In a now-classic second experiment, Garcia demonstrated that, although rats would learn to avoid the taste of a food associated with illness, they would not learn to avoid either visual or auditory properties of a food that had been associated with illness. In this second experiment, rats drank a sweetened solution from drinking tubes wired so that each time the rat's tongue contacted the sweet solution (and the rat experienced a sweet taste) a bell rang and a light flashed. Members of one group of rats received X-irradiation whenever they drank this 'sweet, bright, noisy' water; members of a second group received a mild electric shock to their feet whenever they drank it. A day later, half of each group of rats (one that had been X-irradiated and the other with shocked feet) were tested to determine whether they had learned to avoid drinking a sweet solution. The other halves of each group were tested to determine if they had learned to avoid drinking plain water when licking caused a bell to ring and a light to flash.

Garcia found that rats that had been exposed to X-irradiation after drinking 'sweet, bright, noisy' water avoided ingesting sweet water, but did not avoid 'bright, noisy' water. On the other hand, rats that had received foot shock after drinking 'sweet, bright, noisy' water drank sweet water, but avoided 'bright, noisy' water.

These results were a surprise, because in most other situations learning proceeds equally well regardless of what stimuli are paired with one another. Garcia's rats, on the contrary, associated only taste with illness and only audiovisual cues with shock. Even worse, from the general process point of view, it was soon discovered that birds such as quail more readily learned to avoid the visual properties than the tastes of foods associated with illness.

The general process view of animal learning was faced with a serious challenge because learning about the consequences of eating foods seemed to be different from other kinds of learning: it occurred faster, it occurred with longer delays between the stimuli to be associated, and different species seemed to learn to use different cues to avoid potentially dangerous foods. Pigeon, rat, monkey, which is which? It did seem to make a difference.

It was soon pointed out that there was some biological sense to animals being able both to learn to avoid a potentially dangerous new food after a single pairing of that food with illness and to tolerate long delays between eating a food and becoming ill. After all, eating spoiled food or poisonous substances can result in illness delayed by many hours, and repeated ingestion of toxic substances can have fatal consequences. So, if animals are to be able to learn to avoid ingesting poisons in nature, they would have to be able to learn rapidly to associate properties of substances they ate with consequences of ingesting those substances, even if the consequences of ingestion were long delayed.

It also seemed to make some biological sense for rats to depend on taste cues and birds to depend on visual cues to identify potential poisons. Birds select foods largely on the basis of the food's visual properties, whereas rats tend to eat at night, and use their senses of taste and smell to select things to eat. So, if animals preferentially learn to associate with illness only stimuli in the sensory modality that they use when choosing foods (taste for rats and sight for birds), one might expect the differences among species found in learning associations to illness.

The results of studies of taste-aversion learning clearly suggested that all animal learning might not reflect one or two basic processes. Rather, learning might be in some way modular, with evolution producing a variety of specialized learning and memory systems each of which facilitated learning about biologically important relationships in the natural environment. If so, there should be special processes for learning things other than poison avoidance.

LANDMARK LEARNING

Bee-hunting Wasps

Early students of animal behavior had already shown that some animals whose general ability to learn did not seem particularly impressive could learn surprisingly well those few things most important to the animals' survival and reproduction. Niko Tinbergen, who was later to win a Nobel prize for his work on animal behavior, conducted extensive studies of bee-hunting wasps of the genus *Philanthus*, which lived in Tinbergen's native Holland.

Philanthus is a solitary wasp that lives in small burrows excavated in sandy soil. After stinging and paralyzing a honeybee, a female *Philanthus* returns with her paralyzed prey to her burrow where she stores the bee along with her maturing larvae. The paralyzed honeybees serve as food for the developing young wasps.

The problem that a *Philanthus* female faces after capturing prey, often thousands of meters from home, is how to find her nest entrance, a hole less

than a centimeter in diameter. Tinbergen, in his best-known experiment, waited until rainy weather kept wasps in their nests for a couple of days, and early in the morning of the first fair day, just before a wasp emerged from her nest to go hunting, placed a ring of pine cones around her burrow entrance. When the wasp first emerged from her burrow, she circled above her nest entrance for 6–12 s before flying off to hunt for honeybees. While the wasp was gone, Tinbergen moved the circle of pine cones, which the female had seen only once in her life, a few tens of centimeters from its original position around the nest entrance. He then waited for the wasp to return with a paralyzed bee for her young.

If, during the brief flight she made near the nest entrance before leaving to hunt honeybees, the wasp learned the location of the nest entrance with respect to the pine cones, then she should have landed inside the displaced circle of pine cones when she returned with a honeybee to provision her young. In fact, wasps were four times as likely to land in the ring of pine cones (and at a distance from the true nest entrance) than at the nest entrance itself. Clearly, the wasps had learned about the pine cones in the few seconds between coming to the surface and flying off to hunt for bees.

Of course, it is just possible that wasps are more intelligent than is generally suspected. That turns out not to be the case. This same wasp, *Philanthus*, hunts bees by first approaching any bee-sized moving object, and then flying downwind of it. If the wasp detects honeybee scent while hovering downwind of the object it is inspecting, the wasp lands on the object. If the object feels like a honeybee, the wasp stings it and takes it back to its burrow.

Tinbergen conducted another experiment in which he tethered both a dead honeybee and a honeybee-sized piece of wood on separate threads suspended from a clothes line. The piece of wood was hung a few centimeters downwind of the bee. A wasp would, as usual, approach the objects and then fly downwind of them. Because the smell of the scent of a honeybee was on the wind, the wasp then landed. However, it landed on the piece of wood downwind from the suspended, dead honeybee, not on the honeybee itself. Because the piece of wood did not feel like a bee, the wasp then ended its attack without stinging. The wasps never learned to recognize wooden dummies by sight and avoid attacking them. Instead, the wasps repeatedly attacked the piece of wood, rather than the honeybee just a few centimeters upwind of the wooden decoy, and would land on the decoy dozens of times. So here, in a single animal, one sees both a striking ability to learn about landmarks around a nest and a striking inability to learn to use visual cues to distinguish bees from sticks.

Of course, in the natural world, there are rarely if ever inanimate objects hovering in midair between a bee and a hunting wasp. On the other hand, all bee-hunting wasps have to learn the location of their burrow entrances, if they are to raise their young successfully. *Philanthus* wasps appear to be specialized to learn just those things that they need to learn in the natural environment.

Clark's Nutcracker

Many species of bird and mammal create hoards of food to eat during times of food shortage. Some, like chipmunks or dormice, create a single large cache of food. Others, like squirrels or chickadees, called 'scatter hoarders', create a number of food caches in different locations.

Clark's nutcracker (a middle-sized bird about the size of a blue jay but without a crest, and colored gray and black with white wing and tail patches), is probably the champion among scatter-hoarding birds. In late summer, a single Clark's nutcracker will place twenty to thirty thousand pine seeds in six to eight thousand separate caches. During the next winter and early spring, when relatively little food is available on the mountainsides, each nutcracker recovers the seeds it has cached. The cached seeds, rich in protein and fat, enable nutcrackers to breed far earlier in the spring than other birds that live in the same area but either do not cache seeds or cache far fewer seeds than do the nutcrackers. Nutcrackers are also different in having special pouches that open under their tongues (sublingual pouches) where they can place ninety seeds or more, thus easing transport of seeds to caching sites.

The scrub jay is phylogenetically closely related to and lives in the same area as Clark's nutcracker. However, scrub jays are considerably less dependent than nutcrackers on cached food, and do not have specialized pouches for carrying seeds.

The fact that nutcrackers must remember the locations of thousands of seed caches for weeks or even months suggests that, along with a physical structure for transporting food to caches, nutcrackers might have evolved a specialized system of learning and memory to keep track of the cache sites they have created. Indeed, in natural circumstances, after a nutcracker lands and begins to dig in the ground, more than 70 percent of the time it

recovers pine seeds cached there. This level of accuracy is truly remarkable considering that a nutcracker spends only about 30 s hiding each cache, has to remember thousands of caches, returns to harvest its caches months after creating them, and recovers caches from areas which may have changed considerably in appearance since caching took place: nutcrackers cache seeds in the late summer, when the ground is free of snow, but retrieve them in winter and early spring, when the ground is often snow-covered.

It is, of course, possible that nutcrackers do not really remember where they have cached seeds at all. Perhaps they simply locate caches by their smell, or make marks near caches that they use to guide them to hidden pine seeds. In the laboratory, smells can be removed, as can any marks made by the birds. Laboratory studies, in which cache recovery depends entirely on learning and remembering landmarks that identify cache sites, have shown repeatedly that both nutcrackers and scrub jays can use memories of landmarks to recover cached seeds. However, nutcrackers, the caching specialists, are significantly better than are scrub jays at recovering seed caches a week after creating them. Such findings, and there are a number of them, suggest a specialization of learning and memory for caches in birds that cache extensively in nature.

If there is indeed specialization in nutcrackers for learning and remembering landmarks associated with caches, you might expect to find areas of the brain involved in learning and remembering landmarks better developed in nutcrackers than in scrub jays. Further, one might predict that, in caching bird species in general, brain areas involved in landmark learning would be larger than the same brain areas in noncaching bird species.

The hippocampus, a part of the cortex of the brain, is extensively involved in memory for cache sites. We know that the hippocampus is involved in cache recovery because, although caching birds with lesions in this structure show normal feeding and caching behavior, they are unable to remember where they have cached seeds when they later look for them.

In comparison with noncaching birds, caching birds have large hippocampi for their body size. It is important to be sure that the different body size of caching and noncaching birds is taken into account, because otherwise if caching species were just generally larger than noncaching species and bigger birds tended to have bigger brains, it would look as if the hippocampus of food-storing birds was especially large, even though that was not true.

Homing Pigeons

Caching birds are not the only birds with especially large hippocampi. Lesions of the hippocampus disrupt the ability of homing pigeons to use local landmarks to return to their lofts, and the hippocampi of homing pigeons are larger than those of breeds of pigeon that do not home.

Meadow and Pine Voles

Meaningful relationships have also been found between the need for navigational skill and hippocampus size in mammals, although the best-studied relationship between spatial learning and brain size in a mammal involves differences between the sexes, as well as differences between species.

Various species of vole (small, plump, short-tailed rodents) are to be found in grasslands throughout North America, and different vole species differ markedly in their mating patterns. For example, male meadow voles mate with several different females, and during the breeding season each male meadow vole moves about an area that overlaps the territories of several female meadow voles. Male pine voles, on the other hand, are relatively faithful to a single female, and the territories of male and female pine voles are of roughly the same size throughout the year.

Because male meadow voles travel greater distances than do female meadow voles, whereas male and female pine voles travel equal distances, male meadow voles (but not male pine voles) would seem to need greater proficiency in navigation than would females of their respective species. Indeed, in the laboratory, male meadow voles (but not male pine voles) perform better than do females of their species on tests of spatial learning. As you might expect, male meadow voles (but not male pine voles) have larger hippocampi than do females of their species.

BIRDSONG LEARNING AND IMPRINTING

Learning

Males of many bird species produce a series of notes, trills and pauses (a song) that is used during the mating season both to attract females and to defend territory against intrusion by other males. Males of each species sing a different song, and in some species birds from different geographical areas sing local ‘dialects’, not unlike the different

dialects of native English speakers coming from different parts of the world.

It has long been known that birds learn to sing the song typical of their species. However, the special properties of birdsong learning were clearly demonstrated only in the 1960s by Peter Marler and his associates, who studied song learning in a common North American species, the white-crowned sparrow.

Marler took young sparrows from the nest and reared them by hand in the laboratory under conditions where they could not hear other sparrows sing. When hand-reared sparrows that had never heard a sparrow song grew to adulthood and began to sing, they sang abnormal, simplified songs. Marler reared other white-crowned sparrows in the laboratory and allowed them to listen to tape recordings of adult male white-crowned sparrows. When adult, these sparrows sang not just normal song but also the same dialect as the male whose song was recorded on the tape. Clearly, learning was important for development of song in this bird species.

Marler also found that, although hand-reared white-crowned sparrows would learn white-crowned sparrow song from tape recordings, they would not learn the songs of other species of birds from such tutor tapes. Indeed, white-crowned sparrows reared listening to the songs of other bird species sang simplified songs, just like white-crowned sparrows reared in total auditory isolation. Further, tapes of white-crowned sparrow song played to sparrows when they were between 10 days and 50 days old saved them from singing simplified song as adults, while the same tapes played to the sparrows later in life had little or no effect. So, song learning in white-crowned sparrows was restricted both to certain songs and to certain times of life. The fact that sparrows learn song when young, but do not use this information until they are adult, also differentiated song learning from other types of learning.

Song learning is obviously very different from learning to traverse a maze or to press a lever for food, and like landmark learning has its own special physical basis in the brain. Nottebohm and his colleagues removed various areas from the brains of canaries and recorded their songs both before and after these operations. As a result, the researchers were able to describe a series of clusters of nervous tissue and their connections that control both song learning and song production. The more songs a male sang, the larger were his brain areas concerned with song. Males (which sing) had larger structures than did females (which do not sing). The relevant brain areas, but not others, also grew in the spring and summer, when males sing, and shrank in the fall and winter, when singing ceases.

Imprinting

Imprinting is a term used to describe two kinds of effects of early social experience on later social behaviour. Filial imprinting refers to the learned tendency of young precocial birds to become attracted to and follow their parents (precocial birds are those that hatch in a relatively mature state, like ducks and chickens). Sexual imprinting refers to the effects of early social experience on adult mate preference.

Similar features distinguish imprinting and birdsong learning from the usual types of learning:

- there is restriction on the stimuli which a young bird will learn to follow or to respond to sexually
- there is a restricted period during life when imprinting will occur
- there is (in sexual imprinting) a long interval between the time of imprinting and expression of the imprinted behavior
- there are identifiable neural structures that support imprinting.

CONCLUSION

Immediately following discovery of the special properties of taste-aversion learning, there was a reasonable expectation that many similar cases of adaptively specialized or domain-specific learning and memory processes would soon be discovered, and that a new era would dawn in the study of animal learning and memory. It was an expectation that was to prove difficult to fulfill. Although a few apparently novel learning and memory systems have been discovered, particularly those concerned with landmark learning, progress has been slow.

Nevertheless, the search for adaptively specialized learning processes has led both biologists and psychologists to look to the behavior of animals in their natural environments to identify instances in which animals in nature appear to need to learn. Such instances are not hard to find. Animals have to learn to recognize predators and prey; they have to learn to recognize both kin and other members of their social groups; they have to learn their way around their home ranges; they often have to learn mate preferences or vocalizations appropriate to their species. As we have seen, members of some species have to remember where they have stored caches of food.

However, field data pointing to instances in which animals need to learn provide little information on how learning occurs. Is the learning of biologically important relationships similar to learning in artificial situations, or is such learning special? Such questions can be answered only under the controlled conditions of the laboratory.

Although the search for adapatively specialized learning mechanisms in the laboratory has been in progress since the 1970s, it is not yet clear just how common such domain-specific cognitive processes are. In some cases, such as those described above and a very few others, learning does seem to reflect information-processing systems evolved to respond to particular environmental demands. However, more frequently than was anticipated, general process learning seems to be all that is needed to get the job done.

Further Reading

Balda RP, Kamil AC and Bednekoff PA (1996) Predicting cognitive capacity from natural history. *Current Ornithology* **13**: 33–66.

Balda RP, Pepperberg IM and Kamil AC (1998) *Animal Cognition in Nature: The Convergence of Psychology and Biology in Laboratory and Field*. San Diego, CA: Academic Press.

Bateson PPG, Rose EP and Horn G (1973) Imprinting: lasting effects on uracil incorporation into chicken brain. *Science* **181**: 576–578.

Immelmann K (1972) Sexual and other long-term aspects of imprinting in birds and other species. *Advances in the Study of Behavior* **4**: 147–174.

Jacobs LF, Gaulin SJ, Sherry DF and Hoffman GE (1990) Evolution of spatial cognition. *Proceedings of the National Academy of Sciences of the USA* **87**: 6349–6352.

Konishi M (1965) The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift für Tierpsychologie* **22**: 770–783.

Marler P (1970) A comparative approach to vocal learning: song development in white-crowned sparrows. *Journal of Comparative and Physiological Psychology* **71**: 1–25.

Marler P (1991) Song-learning behavior: the interface with neuroethology. *Trends in Neurosciences* **14**: 199–206.

Marler P and Terrace HS (1984) *The Biology of Learning*. Berlin, Germany: Springer-Verlag.

Nottebohm F (1991) Redressing the mechanisms and origins of vocal learning in birds. *Trends in Neurosciences* **14**: 206–211.

Rozin P and Kalat J (1971) Specific hunger and poison avoidance as adaptive specializations of learning. *Psychological Review* **78**: 459–486.

Sherry DF and Schacter DL (1987) The evolution of multiple memory systems. *Psychological Review* **94**: 439–454.

Sherry DF, Jacobs LF and Gaulin SJ (1992) Spatial memory and adaptive specialization of the hippocampus. *Trends in Neurosciences* **15**: 298–303.

Shettleworth SJ (1998) *Cognition, Evolution and Behaviour*. Oxford, UK: Oxford University Press.

Skinner BF (1956) A case history in scientific method. *American Psychologist* **11**: 221–233.

Learning in Simple Organisms

Introductory article

Kenneth W Eng, University of British Columbia, Vancouver, British Columbia, Canada
Jacqueline K Rose, University of British Columbia, Vancouver, British Columbia, Canada
Catharine H Rankin, University of British Columbia, Vancouver, British Columbia, Canada

CONTENTS

Invertebrate animals with small nervous systems have been studied to investigate the cellular basis of learning and memory. Through behavioral, physiological, and genetic studies many of the mechanisms of habituation, sensitization, and classical conditioning have been discovered.

INTRODUCTION

The principal objectives of researchers using a simple systems approach to learning and memory are to establish the neural circuits involved in a specific form of learning and to identify the cellular changes that occur within those neurons as a result of experience. Simple systems approaches to learning have revealed that modifications in behavior are the result of alterations in precise synaptic connections.

The use of simple organisms offers researchers the opportunity to analyze the neuronal and cellular mechanisms responsible for learning and memory. Because the nervous systems of many invertebrates are composed of only several thousand or tens of thousands of neurons, an investigator can establish with some certainty that any biochemical or neurophysiological changes in the neural system subsequent to learning are the result of that learning. The majority of advances made in cellular and molecular research on learning and memory using simple systems has come from investigating forms of involuntary learning: habituation, dishabituation, sensitization, Pavlovian conditioning (otherwise known as classical conditioning) and instrumental (or operant) conditioning. These simple forms of learning show the same behavioral characteristics and rules in all organisms studied, from the simplest invertebrates to humans. Thus studying them in a simple system may lead to insights unavailable in larger, more complex vertebrates.

ASSOCIATIVE AND NONASSOCIATIVE LEARNING

Behavior is altered by experience through two processes: learning, which refers to a change in behavior that results from an animal's experience with the environment, and memory, which is the storage of information from previous experience that can be retrieved to affect later behavior. This behavioral modification must be reflected in an alteration at the cellular and molecular level of the organism's nervous system; it must remain separate from the animal's natural development and biological maturation, and must not simply reflect sensory adaptation or motor fatigue. Memory is the ability of an animal to retain information from previous experience and the nervous system's ability to store and make use of that information.

Traditional learning theorists have divided simple forms of learning into two types: nonassociative learning and associative learning. The three forms of nonassociative learning studied are habituation, dishabituation, and sensitization. Habituation is defined as a decrement in behavioral response resulting from repeated presentation of the same stimulus. However, this decrement in response is not due to fatigue or sensory adaptation. Habituation can be rapidly reversed by the introduction of a novel or noxious stimulus, causing a return of the habituated response to almost baseline levels, a process referred to as 'dishabituation'. Sensitization refers to an increase of a reflex response above baseline levels due to the presentation of a strong or noxious stimulus.

The two forms of associative learning studied in simple systems are classical conditioning and instrumental conditioning. Classical conditioning, also known as Pavlovian conditioning, is the method by which an animal learns about the predictive relationships between stimuli and events in the environment. Classical conditioning is best illustrated by Pavlov's experiments with dogs. In these experiments Pavlov paired a tone with the presence of food in a dog's mouth. The food is an unconditioned stimulus (US) because it elicits the reflexive response of salivation, an unconditioned response (UR). With repeated exposure to a previously neutral stimulus (the tone) and food together, the dog eventually learns to salivate to the tone, now a conditioned stimulus (CS), before the onset of the food. When the tone alone leads to salivation, a conditioned response (CR) is said to have occurred.

Instrumental conditioning (also known as operant conditioning) is a change in behavior due to the positive or negative consequences of that behavior. Animals are more likely to repeat behaviors that are rewarded and to discontinue behaviors that fail to elicit rewards or that result in punishment.

APLYSIA CALIFORNICA

Aplysia californica is a salt-water mollusk that is about the size of a large rat. Members of this species have been useful in studies of the cellular basis of learning and memory because their large neurons are relatively easy to record from. Learning and memory have been extensively studied using the gill and siphon withdrawal reflex in *Aplysia*. The neurons that control this reflex are found in a cluster of neurons called the abdominal ganglion. Using this model system, researchers have shown that short-term forms of learning are produced through modification of neurotransmitter release, while the persisting memory for that learning involves activation of molecular cascades that can result in gene transcription.

Studies of nonassociative learning in *Aplysia* have focused mainly on defensive withdrawal reflexes: the gill and siphon withdrawal reflexes induced by a tactile stimulus applied to the siphon or mantle shelf, and the tail withdrawal response, which is a contraction of the tail musculature due to tail stimulation. Sensitization is observed when a noxious stimulus administered to the tail of an animal leads to an increase in reflex withdrawal responses. Habituation occurs as a decrement in withdrawal responses as a result of repeated exposure to a tactile stimulus to the siphon. Dishabituation can be observed when a novel or noxious stimulus is applied to the tail and reestablishes a previously habituated withdrawal response to a normal level. Repeated stimulation of the gill and siphon withdrawal reflex results in the formation of long-term memory that can last several weeks.

The best-understood form of learning in *Aplysia* is sensitization: following tail shock, a tactile stimulus to the siphon evokes a greater gill withdrawal response than it would have done before the shock. The cellular correlate of behavioral sensitization is heterosynaptic facilitation. Tail shock causes facilitatory interneurons to release serotonin onto the axon terminals of the siphon sensory neuron. This serotonin activates a biochemical process in the sensory neuron leading to the closing of an ion channel called the S potassium channel. Closing potassium channels causes the action potential to last longer, which results in calcium channels remaining open longer and allowing more calcium into the sensory neuron terminal. A higher level of calcium in the terminal leads to the release of a greater amount of neurotransmitter in response to each stimulus. This increase in transmitter release produces the strengthened behavioral response seen in short-term sensitization. For long-term memory of sensitization, Kandel and colleagues have shown that the serotonin receptors on the sensory neuron also activate a second metabolic pathway that activates genes involved in long-term memory formation. One gene that has been shown to be important in memory formation in *Aplysia*, *Drosophila*, mice, and humans is *CREB*, which encodes cyclic AMP reaction element binding protein, a transcription component linked to long-term memory formation.

The mechanisms underlying long-term memory for both habituation and sensitization in *Aplysia* have been studied in some detail. Memory for both of these forms of learning is reflected in changes in the number and the morphology of the sensory neuron synapses (Figure 1). Long-term sensitization results in more synapses per neuron, larger synapses, and more neurotransmitter vesicles in the area of the synapse. Long-term habituation results in fewer synapses per neuron, smaller synapses, and fewer neurotransmitter vesicles in the area of the synapse. Thus, one role of gene activation in long-term memory is to alter the shape and size of the synapse. This means that memory is reflected in an actual change in anatomical structure of the neurons.

In addition to nonassociative forms of learning, *Aplysia* also shows classical conditioning of the gill

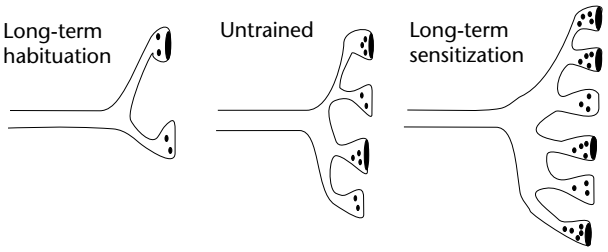


Figure 1. A physical change in the neurons reflects long-term memory. Morphological changes take place in the presynaptic terminals of sensory neurons following long-term nonassociative training in *Aplysia*. Dark ovals represent neurotransmitter vesicles, larger dark areas represent active zones. Long-term habituation training leads to a decrease in the number of synaptic terminals, transmitter vesicles, and active zones; long-term sensitization leads to an increase in their numbers when compared with untrained controls. Based on data from Bailey and Chen (1983).

and siphon withdrawal reflex: a pairing of weak siphon or mantle stimulation with tail shock leads to a pairing-specific enhanced response to the weak stimulus. During conditioning a cellular process known as long-term potentiation (LTP) takes place in the postsynaptic cell (the motor neuron) and involves activity of a special type of glutamate receptor, the *N*-methyl-D-aspartate (NMDA) receptor. When paired stimuli (siphon stimulation and tail shock) are received, increasing calcium influx through NMDA receptors activates a cellular cascade leading to changes in gene expression and memory for the classical conditioning. Interestingly, LTP is believed to underlie classical conditioning in mammals.

CAENORHABDITIS ELEGANS

Caenorhabditis elegans is a 1 mm long, free-living, nonparasitic nematode which has become a powerful model system for genetic studies of development and modern neurobiology. The *C. elegans* genome contains approximately 19 000 genes, more than 40% of which can also be found in humans. The deoxyribonucleic acid (DNA) of *C. elegans* has been completely sequenced and many mutations have been studied. With an understanding of the worm's general physiology, genetic composition, and neuroanatomical development, investigators are well equipped to study the underlying mechanisms of learning and memory. Of the 959 cells of the worm, some 302 are neurons. All 302 neurons and their synaptic connections have been anatomically mapped using electronmicroscopy to generate a complete wiring diagram of the worm's nervous system. This organism shows both nonassociative (habituation, dishabituation, sensitization) and associative (classical conditioning) forms of learning as well as short-term and long-term memory. Thus the learning ability of this well-understood model system allows researchers to isolate specific neuronal synapses where modulations due to learning may be occurring, in addition to identifying specific genes involved in that learning.

Studies using a mechanical stimulus produced by tapping the dish holding the worm have been used to investigate the behavioral, neural, and genetic aspects of habituation and dishabituation of the tap withdrawal response in *C. elegans*. Normally, a tap will elicit a reversal response in which the worm swims some distance backwards. With repeated mechanical taps, reversals become progressively smaller until the worm eventually stops responding, or habituates (Figure 2). Both habituation and spontaneous recovery from habituation are contingent upon the interstimulus interval (ISI) of the taps. Worms habituate more rapidly and more completely to stimuli delivered with short ISIs (10 s) than to long ISIs (60 s). Interestingly, spontaneous recovery is also more rapid at short ISIs than long ISIs. The worm also shows long-term

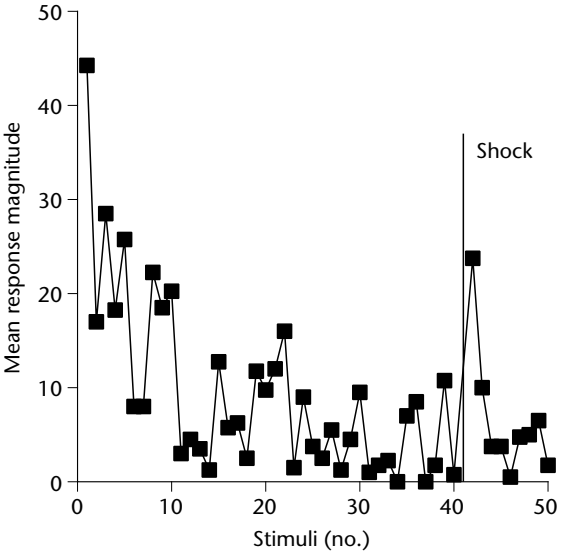


Figure 2. Habituation and dishabituation to tap in *Caenorhabditis elegans*. Mean response magnitude is shown for ten worms stimulated with 40 taps at 10 s intervals, followed by a 60 V shock, then nine more taps at 10 s intervals. The worms show habituation to the first 40 taps, and after the shock, show dishabituation and then rehabituation to tap.

habituation, which involves maintaining memory for habituation training for at least 24 h. Dishabituation (the immediate recovery of a habituated response) is seen in the worm by applying an electrical stimulus to the agar on either side of a habituated animal (Figure 2). *Caenorhabditis elegans* also shows context conditioning in habituation. In context conditioning the worm learns to associate specific environmental cues (e.g. a distinct chemical taste) with habituation to tap. In later tests the worm shows better memory for habituation if tested in the presence of the chemical cue that was present during training.

In order to study the neural circuit involved in the response to tap, a laser was used to kill candidate neurons and the resulting behavioral effects were recorded. The neural circuit for tap consists of five sensory neurons and 11 interneurons which activate a large number of motor neurons (Figure 3). The response to tap is produced by the activation of two competing circuits, one in the head of the worm producing backward movement, and one in the tail of the worm producing forward movement. The observed behavior is an integration of the two competing responses. Laser ablation studies have also shown that the two competing responses (backward movement following head stimulation, and forward movement following tail stimulation) habituate at different rates and with different properties.

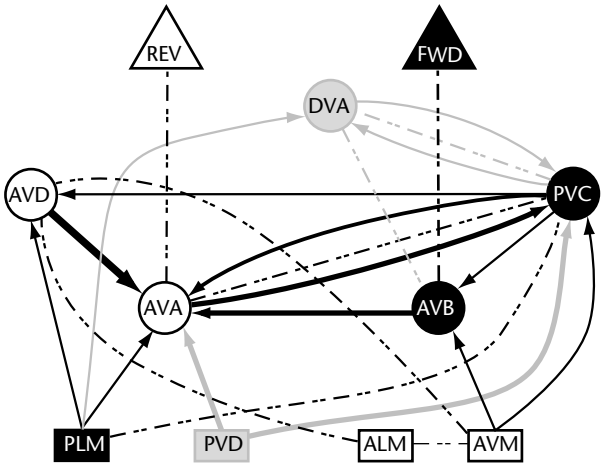


Figure 3. The neural circuit underlying response to tap. The squares represent sensory neurons, the circles represent interneurons, and the triangles represent large pools of motor neurons. Solid lines represent chemical connections with the width of the line corresponding to the number of synapses between the two cells connected. Dashed lines represent electrical connections. Cells shaded in black represent the tail touch circuit to produce forward movement (FWD), white cells represent the head touch circuit to provide backward movement (REV), and the gray cells appear to be involved in both forward and backward movement. Based on data from Wicks and Rankin (1995).

Since habituation to tap does not affect any other behavior in the worm, it is most likely that the learning is occurring in the sensory neurons or in the synapses from the sensory neurons onto the interneurons. These neurons are thought to use the neurotransmitter glutamate to activate the interneurons. A mutant, *eat-4*, thought to have decreased glutamate transmission, has been studied using the habituation to tap assay. At all ISIs *eat-4* worms habituate more quickly and recover more slowly than their wild-type counterparts. In addition, *eat-4* worms do not show dishabituation. Together, these findings suggest a possible role for glutamate transmission in normal habituation to tap in *C. elegans*.

HERMISSENDA CRASSICORNIS

The nudibranch *Hermissenda crassicornis* is a small marine mollusk with no shell which has been used to study memory for classically conditioned pairings of light and rotation. In nature these organisms receive strong vestibular stimulation in their intertidal environment; to avoid injury they respond to this stimulation by clinging to underwater rocks. To mimic this stimulation in the laboratory, researchers subject *Hermissenda* to rotation by placing individual animals in seawater-filled tubes clipped to a turntable. The animals are naturally attracted to light (positive phototaxis), as light directs the animal towards its surface plankton food source. In the classical conditioning paradigm researchers fix a light at the center of the rotating turntable to show that when rotation (US) is repeatedly paired with the light (CS), animals will eventually respond to light alone with increased latency of phototaxis and shortening of the foot. Using simultaneous CS–US presentations it has been demonstrated that the increased phototactic response latency persists during repeated training, lasts for more than 7 days, and is dependent upon the temporal CS–US association, since trials in which the CS and US occurred randomly showed no change in response latency.

The neuroanatomy of *Hermissenda* is well studied, giving researchers the opportunity to investigate specific locations along the sensory input pathways where neural changes take place during classical conditioning of paired light and rotation. Each eye comprises five photoreceptors (two type

A and three type B). In classical conditioning, the CS (light) is mediated specifically by the type B photoreceptors. When rotation (US) is coupled with light presentation, the sensory cells for rotation (statocyst hair cells) are thought to release serotonin which then acts on the type B (CS) photoreceptors. In response to serotonin the type B photoreceptors show a prolonged increase in excitability which produces a suppression of the phototactic response. This prolonged excitation in the photoreceptors is due to decreased potassium channel conductance. Photoreceptors also show increased calcium influx following activation that may also contribute to the reduction in potassium conductance. Long-term retention of type B photoreceptor excitability is blocked by protein synthesis inhibition, suggesting that protein synthesis is responsible for long-term memory for conditioning. Interestingly, application of serotonin followed by conditioning seems to facilitate suppression of phototaxis resulting from conditioning, lending support to the findings of the role of serotonin in learning in *Aplysia* described above.

Researchers have also demonstrated that *Hermissenda* shows inhibitory conditioning, in which the occurrence of one stimulus predicts the absence of another stimulus, and that *Hermissenda* is capable of context conditioning, where animals receive rotation stimulation in either a light or dark environment – animals who experience rotation in the dark prefer the light environment, while animals who experience rotation in the light prefer the dark environment. Changes in the activity of the type B photoreceptors were also reported following context conditioning.

DROSOPHILA MELANOGASTER

Drosophila melanogaster (fruit flies) are tiny flies that have been used for many years in genetic studies. The genome of *Drosophila* has an estimated 12 000 genes. Owing to its complex (many hundreds of thousands of very small neurons) and inaccessible nervous system, cellular examinations are difficult in this species. However, genetic investigation of learning and memory can be done with relative ease in these flies.

Although many different paradigms have been used to study learning in fruit flies, one of the most popular approaches uses a classical conditioning method that pairs an olfactory signal (CS) with an electric shock (US). Researchers measure the change in learned behavior by giving the flies a choice between the CS odor and a control odor not previously paired with shock; trained flies avoid the CS odor but not the control odor. This method has identified a number of mutations that alter flies' ability to learn or remember the training, leading to the hypothesis that *Drosophila* memory has four stages or components (Figure 4(a)). First, there is short-term memory (mutant flies missing this stage are called *dunce* and *rutabaga*), which lasts for about 2 h subsequent to training. Second, there is middle-term memory (mutant flies missing this stage are called *PKA* and *amnesiac*, which lasts approximately 1–7 h and is thought to be contingent upon short-term memory. Third, there is anesthesia-resistant memory (mutant flies missing this stage are called *radish*), which is dependent upon middle-term memory and lasts between 2 h and 7 days, is produced by massed training and is not disrupted by placing the flies in cold, which acts as an anesthetic. The last and final memory phase is termed long-term memory (mutant flies missing this stage are called *dCREB2*), which occurs about 24 h subsequent to spaced or distributed training and is maintained for more than 7 days. Long-term memory is dependent upon middle-term memory formation as well, but is reduced by low-temperature anesthesia and requires new protein synthesis to occur.

APIS MELLIFERA

Honeybees (*Apis mellifera*) are of interest to researchers owing to their well-developed yet compact brains (approximately 950 000 neurons within a cubic millimeter) and their complex behavioral repertoire which includes flying, nest building, defense and attack as well as communication. On the practical side, honeybees are easy to maintain in large numbers in a small space, since queen bees can lay up to 1500 eggs per day. In terms of developmental analysis, worker honeybees only take 21 days to develop from egg to adult, making the examination of several generations possible. Further, honeybees have been shown to be capable of several forms of learning and memory.

Since the early years of the twentieth century investigators have examined the prosboscis extension response (PER), whereby sucrose stimulation of the antennae or the proboscis of bees (analogous to the tongue) results in extension of the proboscis. When a novel odor stimulus (CS) is paired with sucrose stimulation (US) honeybees will show the PER when presented with the CS odor alone. Honeybees show retention of this olfactory conditioning following simultaneous CS–US presentation and forward conditioning; however, backward conditioning (where the US precedes the CS)

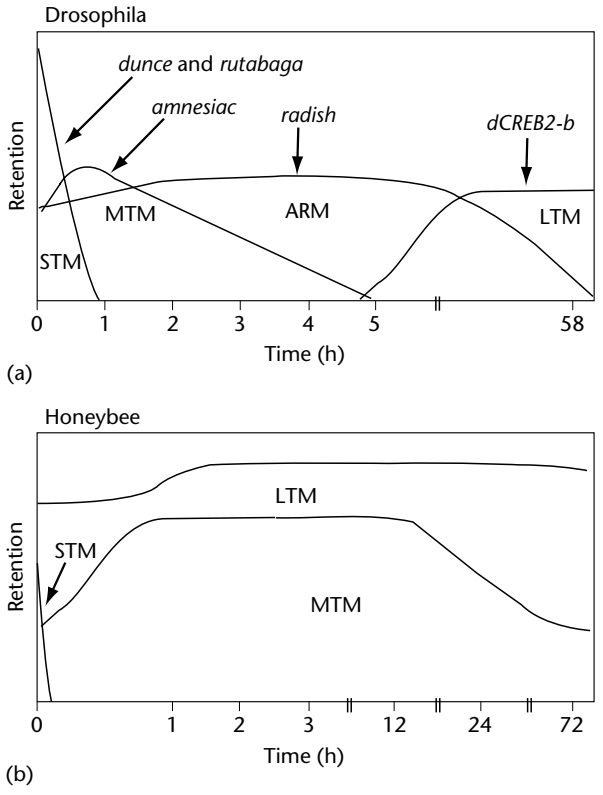


Figure 4. Time course of stages of memory for *Drosophila* and honeybees, with increasing values on the retention (x) axis representing better memory at certain time points following training. (a) *Drosophila* shows short-term memory (STM) which decays within 1 h of training. Single-gene mutants *dunce* and *rutabaga* are defective in this memory stage. Median-term memory (MTM) is present for approximately 5 h after training with *amnesiac* mutants showing a deficiency in this stage of memory. Anesthesia-resistant memory (ARM) and long-term memory (LTM) are longer-lasting forms of memory: ARM decays sooner than LTM (approximately 3–4 days) and is not interrupted by cold shock anesthesia. *Radish* mutants show deficiencies in ARM. Long-term memory lasts for more than 7 days and its formation can be blocked by cold shock anesthesia. The *dCREB2-b* mutants show LTM deficiencies (based on data from Dezazzo and Tully, 1995). (b) Short-term memory (STM) in the honeybee is primarily composed of a sensitization component and decays within 3 min of training. Median-term memory (MTM) forms by approximately 7 min after training, peaking at 30 min and decaying within 24 h, while long-term memory (LTM) shows high retention from immediately following training up to the lifetime of the bee (at least 3 days). Based on data from Hammer and Menzel (1995).

results in no retention of the CS–US association. Interestingly, if bees are presented with two different odors, one paired with sucrose (CS^+) and the other not (CS^-), honeybees will show depressed PER probability to the CS^- while demonstrating an increased PER probability to the CS^+ odor (differential classical conditioning). Honeybees have also shown second-order conditioning of PER in that a previously sucrose-associated odor can later be paired with another stimulus and honeybees will respond with a PER following presentation of the second conditioned stimulus. Acquisition of the PER is inhibited if the CS is repeatedly paired with the absence of sucrose prior to training (inhibitory conditioning). In addition, PER is suppressed if sucrose is made available in the presence of an odor formerly paired with electric shock, demonstrating the honeybee's capability of learning both reward and aversively paired contingencies.

Different phases of memory have been described using the PER paradigm: short-term memory, median-term memory and long-term memory. Short-term memory is brief in duration (0–3 min) with rapid decay and a nonassociative sensitizing component due to the sucrose itself (Figure 4(b)). Median-term memory (approximately 7 min to 12 h) is sensitive to cooling or electrical stimulation of the brain following training and is produced following single-trial conditioning. The discrepancy between the time course of short-term and median-term memory suggests that honeybee memory is biphasic in nature, degrading quickly within a few minutes only to reappear a few minutes after that. Long-term memory is thought to occur only following repeated trials; it lasts from 3 days to the lifetime of the bee and is insensitive to cooling or electric shock. Interestingly, it has been shown that protein synthesis inhibition with the chemical cycloheximide does not disrupt long-term memory, although it has also been reported that inhibition of upstream protein synthesis activators results in its attenuation. As has been described for *Drosophila*, long-term memory is thought to be induced in parallel to median-term memory, as single-trial conditioning does not elicit long-term memory, while cooling and/or electrical stimulation disrupt median-term but not long-term memory.

In addition to associative paradigms employing PER, researchers have also coupled color stimuli with a sucrose reward in an open-field training environment and measured preference for certain stimuli. This type of conditioning allows for more ethologically relevant training and testing of honeybees. Honeybees have also been shown to be capable of complex maze learning as long as it is divided into simpler steps, for instance, if colors cue direction within the maze.

CONCLUSION

Research using simple systems has allowed researchers to gain new insights into some of the mechanisms of learning and memory. Different organisms have been useful for obtaining different types of information. *Aplysia* and *Hermissenda* have allowed detailed studies of experience-related changes in individual neurons; *Drosophila* and *Caenorhabditis elegans* have allowed investigations into the genes involved in learning and memory; studies using honeybees have demonstrated the ethological importance of learning and memory. Since the behavioral rules characteristic of simple forms of learning have been shown to be the same in simple invertebrates and in more complex mammals including humans, it is not surprising that many of the mechanisms of learning and memory found in invertebrates also have a role in learning and memory in vertebrates. Research in these diverse species indicates that long-term memory formation requires training to be spaced out over time, with rests between blocks of training sessions, rather than given all at once. Through research on learning in invertebrates we have learned about biochemical reactions that play a part in learning, and about proteins and genes that are involved in both learning and memory. The simple systems approach to understanding the mechanisms underlying learning and memory has been successful in advancing our understanding of these important processes.

Further Reading

Carew TJ and Sahley CC (1986) Invertebrate learning and memory: from behavior to molecules. *Annual Review of Neuroscience* **9**: 435–487.

Crow T (1988) Cellular and molecular analysis of associative learning and memory in *Hermissenda*. *Trends in Neurosciences* **11**: 136–141.

Davis D (1996) Physiology and biochemistry of *Drosophila* learning mutants. *Physiological Review* **76**: 299–317.

Dezazzo J and Tully T (1995) Dissection of memory formation: from behavioral pharmacology to molecular genetics. *Trends in Neurosciences* **18**: 212–218.

Dudai Y (1989) *The Neurobiology of Learning and Memory*. Oxford, UK: Oxford University Press.

Hammer M and Menzel R (1995) Learning and memory in the honeybee. *Journal of Neuroscience* **15**: 1617–1630.

Jorgensen EM and Rankin CH (1997) Neural plasticity. In: Riddle DL, Blumenthal T, Meyer BJ and Priess JR (eds) *C. elegans II*, pp. 769–790. New York, NY: Cold Spring Harbor Laboratory Press.

Rose JK and Rankin CH (2000) Behavioral, neural circuit and genetic analyses of habituation. In: Shaw CA and McEachern JC (eds) *Towards a Theory of Neural Plasticity*, pp. 176–192. Philadelphia, PA: Taylor & Francis.

Sahley C and Crow T (1998) Invertebrate learning: current perspectives. In: Martinez JL Eesner RP (eds) *Neurobiology of Learning and Memory*, pp. 177–209. San Diego, CA: Academic Press.

Tanchuk TL, Galloway JA, Peters KR and Rankin CH (1998) Memory in fruit flies and nematodes. *Encyclopedia of Life Sciences* Oct 27: 1–12. London, UK: Macmillan.

Learning, Psychology of

Introductory article

Geoffrey Hall, University of York, York, UK

CONTENTS

In its psychological usage, 'learning' refers to the process by which an animal (human or non-human) interacts with its environment and becomes changed by this experience so that its subsequent behavior is modified.

INTRODUCTION

Procedures for Demonstrating Learning

In order to demonstrate that learning has occurred, two observations of behavior must be made. There are two general procedures for doing this. In the first, a given individual is observed twice in the same situation, and if it behaves differently on the second occasion, we make the inference that there has been some change in the organism. In order to make this inference it is necessary that the test situation be identical to the initial training situation – otherwise any change in behavior could just as well derive from a change in the external conditions as from a change in the organism itself.

The change in the organism could occur as a consequence either of its initial experience of the training situation itself or of some experience that has occurred between the first observation of behavior and the second. An example of the former case is provided by the phenomenon of habituation in which, for example, the startle response evoked by a sudden loud noise is reduced in magnitude on the second presentation of the stimulus. We infer that the first presentation of the noise produced some change in the animal that caused it to respond differently to this stimulus on the second presentation. For an example of the latter case, consider the behavior of a student confronted with the question 'How is learning defined?' before and after reading this article. We may hope that the response to the question will be different on the two occasions and infer that the intervening activity (reading the article) has produced learning.

The second procedure is to compare the behavior of two individuals in a given test situation. Provided that these individuals are matched in other respects, differences in their behavior may be attributed to differences between them produced by their differing individual experiences. It seems likely that many of the individual differences exhibited by members of our own species are produced in this way, although, given that any two people (unless they are identical twins) will differ in their genetic make-up as well as in their experience, we cannot always be sure of this. Laboratory experiments allow more certainty. For example, if two laboratory rats of the same genetic stock differ in (say) their ability to perform a maze-running task when one has been raised in an enriched environment and the other has lived all its life in a standard cage, we may confidently attribute the difference in their behavior to learning.

An advantage of this procedure is that evidence for learning can be obtained even when the test situation is quite different from that used in training. A disadvantage is that it is not possible to know in which of the animals being compared the learning has occurred – does enrichment enhance the performance of the rat, or does an impoverished environment produce a change that retards performance? (Although learning will often make an organism better able to deal with the demands of its environment, there is nothing in our definition that requires that this be so.)

Qualifications

These basic definitions and procedures are subject to certain qualifications, of which the three most important are the following.

Firstly, although the demonstration of learning requires a change in behavior, it is possible that a given experience might produce a change in the animal (i.e. produce learning) that is behaviorally

silent under the conditions of the test. (A young child may learn from experience that the volume of a fluid is not changed when it is poured from one container to another, but may fail to show this when tested in the particular circumstances employed by the developmental psychologist.) The absence of a behavioral change cannot prove that no learning has occurred.

Secondly, some changes in behavior, although undoubtedly a consequence of interaction with the environment, are not usually regarded as instances of learning. For example, a rat will press a lever for the reward of a food pellet less readily on the second occasion than on the first if, in the meantime, it has been given free access to food. Such a change in behavior is attributed to a change in motivation rather than to learning. It is difficult to specify a set of rules for distinguishing between these two sources of behavioral change, but it is widely accepted that the changes that constitute learning are more permanent (or, at least, less easily reversed) than those classified as motivational. Similar considerations apply to the short-term behavioral changes produced by muscular fatigue or adaptation of the sensory system.

Thirdly, some of the changes in behavior shown by an individual over the course of its lifespan, although they satisfy the definition of learning offered above, are not usually regarded as such, but are attributed to a process of maturation. An example is the difference in behavior shown by a one-year-old when confronted with a flight of stairs and that shown by a two-year-old in the same situation. The input supplied by the environment over the intervening 12 months (even if it consists only of providing a good diet and the opportunity for exercise) is undoubtedly important in producing the change in the child that allows it to display enhanced stair-climbing ability. But whether this justifies our concluding that the ability has been 'learned' (in the sense in which we may say of an older child that he or she has learned to ride a bicycle) is a matter of debate.

EXAMPLES OF LEARNING

In spite of these qualifications, the psychologist's use of the notion of learning is much wider than that of the layperson. The latter is likely to restrict usage to cases in which new information is acquired (as in learning the properties of the elements in the periodic table) or in which a new skill is acquired (as in learning to drive a car). Both of these are certainly examples of learning under our definition, but so are the examples listed below. The list is not intended to be exhaustive; rather, its purpose is to give an indication of the variety of learning phenomena that have been studied by psychologists. Most of the cases cited below are from controlled laboratory studies (often from studies conducted on animal subjects rather than people), but their relevance to everyday instances of learning will be obvious. (*See* **Animal Learning**; **Comparative Psychology**)

Habituation

The magnitude of the response initially elicited by the presentation of a given stimulus will decline with repeated presentations of that stimulus. We have already cited the waning of the startle response to a loud noise, but the effect is found in many other response systems (e.g. a rat given a novel food will initially decline to eat it, but this neophobic response will decline with repeated presentations of the food). Habituation effects are not to be attributed to muscular fatigue or sensory adaptation: other tests reveal that an animal that has undergone habituation can still detect the stimulus and is still capable of performing the motor response, so that this behavior change should be categorized as learning.

Pavlovian Conditioning

A dog (the experimental subject used by the Russian physiologist Ivan Petrovich Pavlov in his early work on this phenomenon) given presentations of a neutral stimulus (such as the flashing of a light) immediately before the presentation of food will develop a tendency to salivate in response to the light. The light is then referred to as a conditioned stimulus (CS) and the response it evokes as a conditioned response. Food itself is an unconditioned stimulus (US), which, even without special training, is capable of evoking the unconditioned response (UR) of salivation. (*See* **Conditioning**; **Pavlov, Ivan Petrovich**)

The essential feature of this training procedure appears to be that the animal experiences paired presentations of two environmental events. Such pairings are effective in producing learning even when the detailed procedures used are very different from those used by Pavlov in his original studies. For example, a rat that experiences (experimentally induced) nausea after eating a novel food will develop a tendency to shun that food in the future (flavor aversion learning); a rat that experiences an electric shock while a tone is being sounded will develop a conditioned emotional

response (showing freezing and other signs of fear) when the tone is sounded again.

These effects are not confined to laboratory animals – the phobias (strong, seemingly irrational fears) exhibited by some people are plausibly interpreted as being the consequences of emotional conditioning that occurred early in life.

Instrumental (or Operant) Conditioning

If a hungry rat happens to perform some action, such as pressing a lever, that results in the delivery of a pellet of food, its behavior will change so that the rate at which it performs this action will increase. This form of learning is referred to as conditioning, but it differs from Pavlovian conditioning in that the paired events are an action performed by the animal and an environmental event, rather than two environmental events. It is called 'operant' because the animal operates upon its environment to produce an effect, or 'instrumental' because the animal's behavior is instrumental in producing that effect.

Other combinations of action and consequence will also produce learning. A response that removes the animal from a set of circumstances in which aversive events have previously occurred will tend to increase in frequency (a phenomenon known as avoidance learning); a response that is followed by an aversive consequence (as when a lever press produces an electric shock) will tend to decrease in frequency (a phenomenon known as punishment). The general principle underlying these forms of learning is that the likelihood of occurrence of a given form of behavior will be determined by the consequences that have followed that behavior in the past. Although the events that serve as rewards and punishments may be more subtle, it seems likely that much of our everyday behavior is governed by this principle – indeed, according to the American psychologist B. F. Skinner, operant conditioning is solely responsible for shaping all of the behavior that we conventionally refer to as 'voluntary'. (*See* **Reinforcement Learning: A Computational Perspective**)

Discrimination Learning

In discrimination learning, the subject learns to respond in different ways to different stimuli. A pigeon that receives food after pecking at a green disk but not after pecking at a red disk shows this form of learning as it comes to choose the green exclusively. Human subjects show the same sort of ability when they are required to sort a pack of cards, putting all those that bear a triangle (say) into one pile and all those that bear a circle into another.

The stimuli used in discrimination learning experiments may be much more complex than in these examples. In one study, pigeons were rewarded for pecking at each of 40 different pictures of oak leaves but not for pecking at pictures of other leaves. They not only solved the discrimination problem but also showed positive transfer when a new set of pictures of oak leaves replaced those used in initial training.

In this form of discrimination learning, the animal shows an ability to respond differentially according to the category to which the stimulus belongs. Category learning has been much investigated in our own species, for instance in studies of the way in which medical students become able to assign a particular disease name to the varying clusters of symptoms shown by individual patients. (*See* **Concept Learning and Categorization: Models**)

Perceptual Learning

Expert wine tasters are said to be able to make fine discriminations (e.g. distinguishing, by taste, between the top and bottom halves of a bottle) that are impossible for the rest of us. This ability is learned, depending on long experience of the relevant stimuli. The facilitation of discrimination by prior exposure to the stimuli has also been shown in laboratory studies. In one experiment, rats were exposed for many days in their home cages to cut-out geometrical shapes (triangles and circles). No explicit training was given at this stage; but when subsequently the animals were required to learn a discrimination in which one shape was associated with food and the other not, they were able to learn the task very readily. Thus, mere exposure to stimuli can modify the way in which they are perceived, making similar events more discriminable. This phenomenon is known as perceptual learning. (*See* **Perceptual Learning**)

Observational Learning

In one study, children were allowed to watch an adult behaving in an aggressive way towards a toy doll. Just watching was enough to produce learning, as was evidenced by the fact that the children subsequently repeated some of the actions of the adult and also behaved in a violent way to a range of other toys.

Learning by watching others may be a general phenomenon. An adult rhesus monkey confronted

by a snake (or a model of one) will show a characteristic set of fear responses; these responses are not normally shown by a laboratory-raised infant. But if the infant is allowed to watch its mother show the fear response, the infant will come to acquire the reaction.

Imitation constitutes a special case of observational learning in which an observer acquires some new skill from watching it being exhibited by a more competent demonstrator. Young primates who learn how to tackle some novel food may do so by imitating their elders. To some extent their performance may simply reflect the fact that the behavior of the demonstrator has drawn the attention of the observer to the relevant food object; but when the observer is seen to make use of the specific pattern of movement used by the demonstrator, true imitation may justifiably be inferred.

Verbal Learning

For the experimental investigation of learning in people, words (written or spoken) provide a convenient stimulus material. Verbal learning has been extensively studied. (*See* **Word Learning**)

A variety of procedures have been used. In the paired-associate procedure, the subject is exposed to a series of pairs of (usually unrelated) words. Learning is evidenced as the subject develops the ability to respond with the second word of the pair when presented just with the first.

In serial list learning, the subject is presented with a string of, say, 12 words, each being exposed for a few seconds before the next takes its place. With sufficient training the subject will acquire the ability not only to recall the items in the list but to put them in the correct serial order.

In one version of the procedure known as priming, subjects are presented in the training phase with a set of names of common objects. Later they are tested with questions for which the names of objects from the first list might be appropriate answers. Although the subjects may profess no memory of the original list, they are more likely to respond with an item from that list than are subjects not given the initial training (e.g. if 'ostrich' was in the first list they are likely to respond with this, rather than the more usual responses of 'canary' or 'sparrow', when asked to name a bird).

TYPES OF LEARNING

We have distinguished the examples of learning listed above largely in terms of their procedural or descriptive characteristics. Although these distinctions may be pragmatically useful, we may ask whether the various examples can be understood in terms of a smaller number of basic 'types' of learning that differ at a more fundamental level (in the mechanisms that produce them). This question has been vigorously debated since the early twentieth century, and has yet to be fully resolved.

Associative Learning

An extreme view, espoused by Pavlov himself, is that there is only one type of learning: that the process revealed by the Pavlovian conditioning procedure lies at the heart of all other examples of learning. This process was regarded by Pavlov as involving association formation, i.e. the formation of a newly functional link between the brain center that responded to the conditioned stimulus (CS) and the center sensitive to the unconditioned stimulus (US). This link was assumed to allow presentation of the CS to evoke activity in the US center – and thus to evoke behavior appropriate to the US (e.g. salivation to a light that had previously been paired with food) – even in the absence of the US itself.

Pavlov did not explain how such a process might result in, for example, the acquisition of a motor skill or the formation of a new concept; but in the absence of any well-specified alternative, it is difficult to rule out the possibility that the apparent complexity of these examples of learning might derive from the operation of a fundamentally simple associative mechanism.

The strongest arguments against Pavlov's unitary view came from those psychologists who studied another, seemingly simple, example of learning in laboratory animals, namely, instrumental learning. The essential feature of this procedure – that learning depends on the effect produced by the response – appears to have no parallel in Pavlovian conditioning. It was argued, therefore, that the two forms of learning depended on fundamentally different mechanisms: reward-produced strengthening of the response, in the instrumental case, and association formation produced by the contiguous occurrence of two stimuli in the Pavlovian case. It was further suggested that these two processes might operate selectively on different response systems: that Pavlovian conditioning works for simple reflex responses, whereas the modification of voluntary behavior depends on the instrumental learning process.

The distinction has, however, proved difficult to maintain given more recent demonstrations of the

instrumental conditioning of involuntary responses and of the fact that supposedly voluntary behavior (e.g. that shown by an animal in moving about its environment) can be modified by classical conditioning. The consensus now is that both are examples of associative learning, differing only in the nature of the events that become associated: two stimuli in the Pavlovian case; a response and its outcome in the instrumental case.

The notion that associations form between events that co-occur (be they neutral stimuli, motivationally significant events such as the delivery of food, or patterns of behavior emitted by the animal) provides a powerful explanatory tool that can be applied not only to classical and instrumental conditioning but to several others of the examples of learning given above. In discrimination learning, for instance, when the pigeon chooses one stimulus rather than another, this may simply reflect the fact that one has become associated with food and the other with the absence of food. The observational learning shown by an infant monkey may be interpreted as the formation of an association between the originally neutral stimulus and the aversive state engendered by the sight of its mother in distress.

In general, the fact that so many examples of learning involve presenting the animal with conjunctions of events makes the associative principle a plausible candidate for the explanation of all of them. But there remain examples in which learning occurs after exposure to just a single event (the most obvious cases are habituation and priming, but perceptual learning is also relevant here). It is difficult to see how associative mechanisms could be responsible for the behavior change seen in these procedures. We may therefore need to acknowledge the existence of at least two types of learning: associative learning (by which the animal learns what goes with what) and a nonassociative form of learning that allows the animal to learn something about the characteristics of an event to which it is exposed.

Implicit and Explicit Learning

Another distinction between different types of learning (which cuts across the distinction between associative and nonassociative learning) has recently been the subject of much attention. This is the distinction between explicit and implicit learning. When normal adult humans learn, they can often report the results of their learning verbally (they can tell you that, as a result of experience, they know that canaries are birds or that fire is hot). Sometimes they can report the details of a particular learning episode (they can tell you what they did on their twenty-first birthday). But not all learning is explicit in this way. Some people who have suffered damage to certain parts of the brain seem unable to acquire new facts or recall recent events from their everyday life. But although they may deny all knowledge of the episodes responsible, they are still capable of some forms of learning: they will show improvement when given practice at a new motor skill (e.g. learning to trace a pattern that is viewed only in a mirror); and they are sensitive to standard conditioning procedures (e.g. they will tend to blink to a neutral stimulus that has been paired with a puff of air to the eye). (*See* **Implicit Learning**)

Implicit learning phenomena are not confined to amnesic patients. The phenomenon of priming, described above, is a laboratory-based example. But anyone who practises a motor skill (as when learning to ride a bicycle) is likely to show an improvement from one session of training to the next while remaining largely unaware of the changed patterns of muscular coordination that produce the improvement (try explaining to a beginner exactly what he or she needs to do to ensure that the bicycle remains upright). And much of the learning that occurs as we acquire our native language remains implicit. Given a set of word strings we are usually able to distinguish those that are permitted by the rules of the language (those that are grammatical) from those that are not, even when we cannot specify the formal rules that justify the distinction. The learning of so-called artificial grammars may involve the same process. For example, people may be exposed to strings of letters that appear random but are in fact constrained by certain rules (e.g. when a J occurs it must always be followed by a V or an X, W can never be followed by itself). The subjects are not told these rules, and indeed are incapable of stating them, but nonetheless they prove able to make correct categorizations of new instances. (*See* **Motor Learning Models**; **Language Acquisition and Language Change**)

The importance of the distinction between implicit and explicit learning remains a matter of debate. One problem with the distinction is that it rests on the observation that the experimenter is unable to obtain from the subjects a verbal statement of what they have learned. Such a failure might be the fault of the experimenter rather than of the subject. Recent work has shown that in some cases more subtle interrogation will allow subjects to make explicit information that simpler questions (and the subject's own initial introspection) fail to reveal.

Secondly, even if it could be firmly established that some forms of learning produce changes that are truly implicit, this would not necessarily imply that the mechanisms involved are fundamentally different from those involved in explicit learning. We would still want to know why some of our memories are available to the conscious mind and some are not, but the difference between them might have no bearing on the nature of the physiological or psychological processes by which memories are formed.

MECHANISMS OF LEARNING

Neural Mechanisms

The changes that constitute learning almost certainly occur in the animal's nervous system; and one approach to investigating mechanisms of learning is to try to determine the nature of the neural processes involved. The dominant hypothesis (put forward by the eminent neuropsychologist D. O. Hebb in 1949 on the basis of very little evidence) has been that learning consists of a change in the properties of synapses (the structures by which one nerve cell makes contact with another). Hebb's proposal was that experience might render a nonfunctional synapse functional, so that activity in one nerve cell would become capable of inducing activity in another. (*See* **Hebb, Donald Olding**; **Hebb Synapses: Modeling of Neuronal Selectivity**; **Synaptic Plasticity, Mechanisms of**)

Recent work, principally on the simple forms of learning (e.g. Pavlovian conditioning) exhibited by invertebrates, has confirmed the validity of Hebb's conjecture, at least for the cases studied. In the mollusc *Aplysia*, the sensory neuron (nerve cell) that is sensitive to a touch on a structure known as the siphon makes a synapse with the motor neuron responsible for contraction of the gill. A light touch on the siphon is not usually enough in itself to excite the motor neuron and evoke the response. But pairing the touch with a more effective stimulus (a shock to the tail) produces a change in the synapse (a phenomenon known as presynaptic facilitation) so that a light touch now produces more neurotransmitter and gill contraction occurs. (*See* **Learning in Simple Organisms**)

An effect known as long-term potentiation provides evidence of synaptic plasticity in the mammalian brain. This phenomenon concerns the case in which two neurons (*A* and *B*) both have synapses with a third (*C*). Strong activation of one input (say *A*) will produce a near-permanent increase in the sensitivity of *C* to this input; and if *B* is activated (even if only weakly) at the same time as *A* is being strongly activated, the potentiation effect will spread to *B* so that it to will become more able to evoke activity in *C*. Direct links between this neural process and overt behavior have yet to be established; but there is a clear parallel with Pavlovian conditioning, which also depends on pairings of two stimuli, only one of which is initially capable of evoking the target response. (*See* **Long-term Potentiation, Discovery of**; **Long-term Potentiation and Long-term Depression**)

A Conceptual Nervous System

Although our knowledge of the neural mechanisms responsible for learning is still very incomplete, the psychological analysis of the phenomenon is relatively advanced. Such analysis proceeds by studying, usually in contrived laboratory preparations, how the nature of what an animal learns, and how readily this learning occurs, can vary according to conditions manipulated by the experimenter. It is possible to make deductions about the mechanisms that must be operating in order to produce the behavioral data obtained. The aim is to provide a specification of what has sometimes been called a 'conceptual nervous system'. The properties of this system will, it is hoped, be consistent with what we know about the functioning of the real nervous system; but the specification is usually given in rather general terms, as a high-level design for a machine that could just as readily be made of silicon and copper wire as of nervous tissue.

In fact, there is a widely accepted picture of the conceptual nervous system that fits very well with what we know of the functioning of the real nervous system. Many psychologists, who have approached the study of learning from a variety of different perspectives, have converged on the view that the mechanism responsible for learning should be viewed as consisting of a large set (a network) of interconnected units. Units may be inactive, or they may be activated to varying degrees, and activity in one unit will engender activity in another unit with which it has a functional connection. Learning is held to consist of changes in the strengths of connections, and thus of changes in the ease with which one unit can modulate activity in its neighbor. Within this scheme, the psychological analysis of learning consists in determining the nature of the units, their pattern of connections, and, most importantly, specifying the factors that determine how connection strengths will change.

These general principles (often referred to as connectionism) have been applied with some

success to a wide range of cognitive phenomena that involve complex information processing. But an appreciation of how they work may best be obtained by considering a simple form of learning such as Pavlovian conditioning. Pavlovian conditioning involves pairings of a CS with a US that evokes a UR. We begin by postulating sensory units, which are activated by presentation of the appropriate stimuli, and an output unit, whose activity generates the response. The US unit must be assumed to be unconditionally connected to the UR unit. The acquired ability of the CS to evoke a response is taken to indicate the formation of a new functional connection. The obvious hypothesis that a connection is formed between the CS unit and the UR unit has not been supported by experiments: conditioning has been demonstrated when the training conditions are such as to preclude the occurrence of the UR. This and other observations have led to the conclusion that Pavlovian conditioning reflects the formation of a connection that allows activity in the CS unit to engender activity in the US unit. (*See* **Connectionism**)

Other cases of conditioning require an associative structure more complex than the 'two units, one link' structure that serves for simple acquisition. For example, animals given training in which the US is paired with a compound stimulus (*A* and *B* presented together) along with separate presentations of *A* and *B* that are not accompanied by the US, will learn to respond just to the compound and not to the individual stimuli. One possible explanation of this accomplishment assumes a structure in which the units sensitive to presentations of *A* and *B* have connections with a third unit, which is activated when it receives inputs from both *A* and *B* (i.e. when the compound is presented). The strengthening of a connection between this third unit and the unit representing the US allows the animal to learn the discrimination and show a response to the compound stimulus. The presence of such hidden units (which are not directly activated by environmental events, but serve a purely computational function) greatly increases the explanatory power of the associative network. Networks employing such units have been applied to complex learning phenomena, such as category learning, with some success. (*See* **Bayesian and Computational Learning Theory**)

Laws of Association

The proposal that experience can produce changes in the strengths of the connections between units is central to the idea of this conceptual nervous system. Therefore, for a proper account of the mechanisms of learning we need to be able to state the conditions under which such changes will occur. A version of this question has engaged the minds of philosophers for several centuries; and they have proposed a variety of possible laws of association (principles determining the readiness with which one idea is able to 'call up' another). Examples include the law of contiguity (the principle that events that occur together in time and space will become associated) and the law of similarity (the principle that similar events are more likely to become associated than dissimilar events). Experimental study of these proposals has made use of simple learning preparations (particularly of Pavlovian conditioning) in which the nature and timing of the events to be associated can be easily manipulated. The magnitude of the response evoked by the CS is used as an index of the strength of the association that is formed under various conditions.

Conditioning experiments have largely confirmed the importance of contiguity. The conditioned response develops most readily when the CS and the US are presented close together. This supports the assumption that an excitatory connection will be strengthened when the units representing these stimuli are active concurrently. But contiguity may be neither necessary nor sufficient for learning to occur.

That contiguity may not be necessary for association formation is illustrated by the fact that in flavor aversion learning, nausea induced several hours after an animal has tasted a novel food will be effective in establishing some measure of aversion to that food. But whether this observation constitutes a fundamental challenge to the contiguity principle is not clear. Conditioning is readily established in more orthodox procedures when there is a short delay (of a second or two) between the CS and the US; and this observation is readily explained by supposing that the activity induced in the CS unit will persist for some time after the CS itself has ended. There is no reason why the same analysis should not be applied to the case of flavor aversion learning. Admittedly, the residual activity in the unit representing the taste of food is likely to be at a rather low level after an hour or so; but if there is any activity at all (and it is difficult to see how an association could be formed if there were not), then the principle of contiguity can be maintained.

It should be noted, however, that such long-delay learning is readily obtained only with certain combinations of events (such as taste and

nausea – indeed, it is difficult to establish any association at all with nausea as the US when the CS is an exteroceptive cue such as the sounding of a tone). The special ability of taste and nausea to become associated even over a long delay may indicate that another principle, in addition to the contiguity principle, is operating in this case. One hypothesis is that events concerned with maintaining the internal state of an animal have special propensity to become associated, thus allowing a taste unit that is only weakly activated to form a strong association with the nausea unit. No mechanism that might be responsible for this propensity has been specified. It may be an example of the more general principle that associations form with particular ease when the events to be associated are similar.

That contiguity is not sufficient to produce association formation is well supported by experimental evidence from a variety of procedures. Two examples will be given here. In one experiment, rats were given presentations of a CS that was followed with a certain probability by the US; pairings occurred often enough that a conditioned response was established, indicating that the association had been formed. Other rats received the same treatment except that, for them, further presentations of the US occurred in the interval between CS presentations. The probability of occurrence of the US was the same in the absence of the CS as in its presence. Rats in the latter condition (who received the same number of CS–US pairings as those in the former condition) did not acquire the conditioned response.

The second example comes from a procedure, much studied in recent years, known as blocking. In this, the subject is initially given training in which one CS (*A*) is paired with the US. The subject then receives a further phase of training in which a second CS (*B*) is added, and the compound (*A* and *B*) continues to be paired with the same US. A final test, in which *B* is presented alone, reveals that this stimulus will not evoke the conditioned response: apparently no *B*–US association is formed in these circumstances, in spite of the fact that the subject has been exposed to contiguous presentations of the two events a number of times during the second phase of training.

These two examples seem to show that contiguity is not enough: that conditioning will occur only when the CS supplies information about the likelihood of occurrence of the US. In the first example, learning fails to occur in the condition in which the US is as likely to occur when the CS is absent as when it is present; in the second, the subject fails to learn about a stimulus that supplies no new information about the likelihood that the US will occur (this being fully predicted by the presence of the pretrained stimulus, *A*).

One possible specification of the mechanisms responsible for these effects was proposed by the psychologists R. A. Rescorla and A. R. Wagner in the early 1970s. They accepted that an associative link between CS and US units will be formed when contiguous presentation of these events evokes activity of some sort in the relevant units. They went on to point out that the formation of the link means that presentation of the CS will be able to evoke activity in the US unit in advance of the occurrence of the US itself. Further strengthening of the link, they suggested, will be a function of the discrepancy between the level of activity evoked in the US node by the associative connection and that engendered by the application of the US itself.

This simple discrepancy principle (sometimes called the 'delta rule') proves to have wide explanatory powers. Its application to the phenomenon of blocking runs as follows. In the initial phase of training a strong connection is formed between the unit representing CS *A* and the US unit. This connection continues to be effective during the compound (*A* and *B*) trials, and thus fully activates the US unit during these trials. When *B* is presented, therefore, there is no discrepancy between the level of associatively produced activity and the level of activity produced by the US itself, and no connection involving CS *B* is established.

The essence of associative learning is that initially the organism does not know what the outcome of a given event will be but that with training it comes to do so. Once the relationship is well established there is no need for further learning. The Rescorla–Wagner principle provides a very simple mechanism by which associative learning may be achieved. It has been very widely accepted. This principle (or some version of it) has been a fundamental feature of all subsequent connectionist theories, and has allowed the successful application of these theories to a wide range of learning phenomena. It remains to be established whether these associative mechanisms can be extended to apply (as some psychologists have argued) to all instances of learning.

Further Reading

Anderson JR (1995) *Learning and Memory: An Integrated Approach*. New York, NY: Wiley.

Berry DC and Dienes ZD (eds) (1993) *Implicit Learning: Theoretical and Empirical Issues*. Hove, UK: Erlbaum.

Hall G (1991) *Perceptual and Associative Learning*. Oxford: Clarendon Press.

Hinton GE (1992) How neural networks learn from experience. *Scientific American* **267**(3): 104–109.
Kandel ER and Hawkins RD (1992) The biological basis of learning and individuality. *Scientific American* **267**(3): 52–60.
Klein SB and Mowrer RR (1989) *Contemporary Learning Theories*. Hillsdale, NJ: Erlbaum.
Mackintosh NJ (ed.) (1994) *Handbook of Perception and Cognition*, vol. IX 'Animal Learning and Cognition'. San Diego, CA: Academic Press.
Pearce JM (1997) *Animal Learning and Cognition*. Hove, UK: Psychology Press.
Rescorla RA and Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black A and Prokasy WF (eds) *Classical Conditioning*, vol. II, pp. 64–99. New York, NY: Appleton-Century-Crofts.
Shanks DR (1995) *The Psychology of Associative Learning*. Cambridge, UK: Cambridge University Press.
Wasserman EA and Miller RR (1997) What's elementary about associative learning? *Annual Review of Psychology* **48**: 573–607.

Lexical Acquisition

Intermediate article

Linda McCabe Smith, Southern Illinois University, Carbondale, Illinois, USA
Linda Bland-Stewart, George Washington University, Washington, DC, USA
Lewis Annette Carter, Southern Illinois University, Carbondale, Illinois, USA
Yvette Hyter, Western Michigan University, Kalamazoo, Michigan, USA
Tempi Champion, University of South Florida, Tampa, Florida, USA
Linda Campbell, St Louis University, St Louis, Missouri, USA

CONTENTS

Lexical development refers to the rules that children use to understand, produce, and create the meaning of individual words and their combined relationship to one another.

INTRODUCTION

The rapid growth of lexical development during the preschool years is indicated by the addition of at least five words daily to the child's meanings. Acquiring new meanings is a complex developmental stage for young children. For instance, a particular word has many characteristics that establish its meaning, and a child's task is to learn which factors are critical in acquiring the appropriate use of a word. As children learn words and their meanings, they are taking the first steps towards a complex sequence in the development of language. Thus, early lexicon development is a foundation for later language development. As children increase in age, the complexity, variety, and total number of words in their vocabulary increase, as do the individual relationships among words.

LEXICAL DEVELOPMENTAL KNOWLEDGE

Knowledge of lexical development refers to a young child's information concerning objects, people, and events with regard to function and attribute, as well as to the relationships between and among these factors (Bloom and Lahey, 1978; McLean and Snyder-McLean, 1978; Owens, 1996; Rice, 1984; Wells, 1985). For example, the frequency and variety of lexical categories that are present in a child's utterances as they progress from single to multiple word utterances represents important information about the emergence of language form, lexicon content, and the interaction between form and content. Early semantic meanings involve all aspects of language development, that is, syntax, pragmatics, and morphology that children apply to agents, actions, and objects within their vocabulary. One of the most important concepts to realize is that there appears to be semantic, syntactic, and pragmatic overlap in lexical development.

Critically, the semantic content of children's utterances is influenced by both pragmatic and cognitive constraints (Bloom and Lahey, 1978; Owens, 1996; Wells, 1985). Children's earliest words are pure performatives such as 'hi' or 'bye bye' (i.e. the word itself performs the action named) (Menyuk, 1974). Operation of reference follow (Leonard, 1976) includes nomination of substantive words, existence, non-existence, disappearance, recurrence, and negation. However, children are highly individualized and do not all follow the same developmental pattern.

At approximately 18–22 months of age, children have developed the cognitive capacity to separate entities from dynamic states and relationships. This developmental advance is reflected in the content and form of their productive language (Bloom and Lahey, 1978; Owens, 1996). For example, children now begin to combine words and increasingly follow simple word-order patterns in the production of their multiple word utterances that are based on semantic distinctions or 'semantic-syntactic rules' (Bloom and Lahey, 1978). At this level, the utterance 'dere truck' denotes the static location of the truck (i.e. 'the truck is there') – a 'locative state' utterance, whereas the utterance 'my truck' indicates possession or ownership of the truck (i.e. 'the truck belongs to me'), a 'possession' utterance.

Such two-word utterances also may include verbs, as in 'push truck', that designate the semantic category 'action', soon to be expanded into a full subject–verb–complement structure ('me push truck').

One of the first tactics in lexical development is a child's limited and faltering definition of words. That is, the child understands the concept that words have referents in the real world and that words stand for persons, places, objects, events, or things. When acquiring meaning, children tend to match up a set of semantic markers. For example, the semantic markers for the word 'dog' might include 'furry', 'four legs', 'bark', 'wagging tail'. The acquisition of a set of semantic markers for a given word may be described as a process of concept development.

CONCEPT DEVELOPMENT

One of the early developmental concepts that may assist children in acquiring meanings of words is object permanence. Object permanence refers to the understanding that things exist even when we are not currently experiencing them. Children are able have a mental picture of objects, persons, places or things even though these things are not in their immediate presence. If a child understands this basic concept, he or she will come to realize that mom leaving the room does not mean that she has disappeared entirely, but merely that she has disappeared from view.

This process of development can also be observed in the child who uses the word 'car' to refer only to the family's vehicle which is a 'station wagon', indicating a narrow sense of the word 'car'. Over a period of time the child will begin to apply the word to all cars, trucks, and tractors, and soon will label anything with four wheels a 'car'. Now the child's concept of car has become too broad and must be refined. The child will soon learn that there are different names for different types of vehicle.

Nelson (1974) writes that a young child carries around a set of concepts, or organized categories, referring to aspects of the environment. These basic terms consist of nouns, such as the child's own names for pets and family members, names of objects the child relates to directly (e.g. shoe/spoon), names of body parts (e.g. nose, ear), and foods that are liked by the child. As one can see from these samples the concepts are derived from the child's own environmental experiences. These concepts are further developed over time by the influence of the child's caregivers when they respond to the child's use of words; i.e. regarding them as right or wrong.

Eve Clark (1973) reviewed several diary studies, in which parents recorded their children's early communication attempts. The diaries showed that the children's first utterances seem heavily perceptual, i.e. based on sensory characteristics of objects. These organizations of sensory input, or precepts, might be responsible for the kind of overgeneralization of 'car' described above. Clark found that the most frequently used basis for perceptual categories is shape; other relevant attributes of reality include sound, size, movement, texture, and taste. She found no early categories based on color. Clark argued that these perceptual categories are universal: that is regardless of the language being acquired, there is something about children's nonlingusitic experience with the environment, and perhaps about the human sensory system itself, that causes children to organize reality around these characteristics.

SUMMARY

There appears to be disagreement about exactly what kinds of early, nonlinguistic experience lead to the meanings children express in their earliest words. Children's language is not meaningless. In young children, lexical development in context will progress from two-, three- and four-word utterances to more complex adultlike form and content. As the adult/child interactions continue, the child's contribution increases in meaningfulness. Increased vocabulary and relational terms enable the child to sustain and relate conversational-limited topics. Semantic meaning reflects the relations between words and concepts of reality. These aspects of language will change greatly throughout the school years and continue to be challenging to the study of lexical acquisition.

References

Bloom L and Lahey M (1978) *Language Development and Language Disorders*. New York: Wiley.

Clark EV (1973) What's in a word? One the child's acquisition of semantics in his first language. In: Moore TE (ed.) *Cognitive Development and the Acquisition of Language*, pp. 65–110. New York: Academic Press.

Leonard L (1976) *Meaning in Child Language*. New York: Grune and Stratton.

McLean J and Snyder-McLean L (1978) *A Transactional Approach to Early Language Training*. Columbus: Merrill.

Menyuk (1974) Early development of language: from babbling to words. In: Schiefelbusch R and Lloyd L

(eds) *Language Perspectives: Acquisition, Retardation, and Intervention*. Baltimore: University Park Press.
Nelson K (1974) Concept, word, and sentence: interrelations in acquisition and development. *Psychology Review* **81**: 267–285.
Owens RE (1996) *Language Development: An Introduction*. Needham Heights: Allyn and Bacon.
Rice ML (1984) Cognitive aspects of communicative development. In: Schiefelbusch R and Pickar J (eds) *The Acquisition of Communicative Competence*. Baltimore: University Park Press.
Wells G (1985) *Language Development in the Preschool Years*. New York: Cambridge University Press.

Further Reading

McLean J and Snyder-McLean L (1999) *How Children Learn Language*. San Diego, CA: Singular Publishing Group.
Naremore RC and Hopper R (1990) Development of meaning. In: Naremore RC and Hopper R (eds) *Children Learning Language: A Practical Introduction to Communication Development*, pp. 71–82. New York, NY: Harper and Row.
Olah LN (2000) How language comes to children: from birth to two years/how children learn the meaning of words. *Harvard Educational Review* **70**: 538–543.
Owens RE (1999) Specific intervention techniques. In: Owens RE (ed.) *Language Disorders: A Functional Approach to Assessment and Intervention*, pp. 302–309. Needham Heights, MA: Allyn and Bacon.
Owens RE (2001) Preschool pragmatic and semantic development. In: Owens RE (ed.) *Language Development: An Introduction*, pp. 292–302. Needham Heights, MA: Allyn and Bacon.
Rhea P (2001) *Language Disorders from Infancy through Adolescence*. St Louis, MO: Mosby.
Wilde J, Astington JW and Barriault T (2001) Children's theory of mind: how young children come to understand that people have thoughts and feelings. *Infants and Young Children* **13**: 1–12.

Lexical Ambiguity Resolution

Intermediate article

Cyma Van Petten, University of Arizona, Tucson, Arizona, USA

CONTENTS

Most words have several meanings, yet readers and listeners are able to determine a writer's or speaker's intent from the word's context. Research on the resolution of lexical ambiguity has developed not only theories on how comprehenders settle on a single meaning, but also experimental paradigms for examining this process.

INTRODUCTION

A small number of words in English are lexically ambiguous in that one spelling and/or one pronunciation is associated with two unrelated meanings – for example, 'spoke'. This small set of words has been the topic of hundreds of journal articles and book chapters since 1970. One reason for this great research interest is that almost all other words are 'polysemous' – they have multiple but related senses. Consider 'clear' as in 'passes light' versus 'easy to understand'; 'running' in 'running a marathon' versus 'running for election'; and 'paper' in 'wrapping paper' (substance) versus 'liberal paper' (institution). Britton (1978) reported that 44 percent of a random sample of English words had more than one dictionary definition. The more commonly a word is used, and the longer it has been part of the language, the more meanings it possesses (Zipf, 1945; Lee, 1990). Linguists and psycholinguists have wrestled with the proper treatment of the one-to-many mapping between word forms and word senses (Klein and Murphy, 2001), and many consider ambiguity to be the extreme end of a continuum of polysemy. Studies of ambiguity resolution thus concern a fundamental aspect of comprehension – how readers and listeners identify the contextually appropriate senses of words. Most lexical ambiguity research has been conducted in English, but ambiguity and polysemy also occur in the world's other languages. For example, one-to-many mappings between pronunciations and meanings are prevalent in spoken Chinese (Li, 1998), so that by one count the syllable '*yi*' (with dipping tone) has 90 different meanings in Mandarin (Yip, 2001). (*See* **Lexicon**; **Lexical Semantics**; **Word Meaning, Psychology of**)

MODELS OF AMBIGUITY RESOLUTION

In everyday language use, the meaning of words is clarified by their context. Laboratory studies have focused on how comprehenders use context to arrive at the relevant word sense, and have suggested three different models. The exhaustive access model proposes that comprehenders first activate all possible senses of a word, then use context to select one meaning. The ordered access model suggests that comprehenders initially activate the most common sense of a word, and proceed to less common senses only when the initial meaning does not fit the context. The selective access model suggests that appropriate prior context can direct comprehenders to the relevant meaning of a word immediately. Hybrid models are also possible, for instance that access is selective when the context is strong, but that the most frequent meaning is initially accessed with less constraining context. The following text provides a snapshot of different and often conflicting empirical findings. For a discussion of computational models of lexical ambiguity resolution, the reader is directed to articles by Kawamoto (1993) and Dixon and Twilley (1999).

THE CONTEXT–AMBIGUITY–PROBE PARADIGM

A commonly used paradigm for investigating ambiguity resolution consists of three stimuli: a semantic context biasing one sense of the ambiguity, the ambiguous word, and a probe word. The context may be a single word ('iron bar' versus 'gay bar'), a sentence frame providing semantic context ('She covered the floor of the stable with clean straw') or a sentence frame providing syntactic

context alone ('He needed to change' versus 'He needed some change'). Probe words are related to the contextually relevant sense of the ambiguity, to the irrelevant sense, or to neither (in the last example, 'clothes', 'coins' or 'water'). Responses to the probe word are the critical measure, in particular whether responses to irrelevant probes more closely resemble the relevant or the unrelated condition. Important factors include:

- the amount of time between the ambiguity and the probe
- the nature and strength of the context
- the dependent measure, or how the processing of the probe word is evaluated
- whether the sentence biases the more common (dominant) or less common (subordinate) meaning of the ambiguity (Simpson, 1994).

Two reaction time (RT) tasks are sensitive to semantic relationships: deciding whether a letter string is a word or a nonword (lexical decision time), and reading a written word aloud (naming latency). Table 1 shows a typical pattern of results in the context–ambiguity–probe (CAP) paradigm: when a short time elapses between the ambiguity and probe (around 200 ms), responses to the contextually irrelevant probes are faster than to unrelated words. When a slightly longer time elapses (around 700 ms), responses to the irrelevant probes are equivalent to those elicited by unrelated words (Swinney, 1979). These results have been taken as support for the exhaustive access model. The interpretation is that, when presented quickly enough, contextually irrelevant probes are processed as related because both meanings of the ambiguity are still active. A short time later, the second stage of semantic processing selects the contextually relevant sense of the ambiguity, and the contextually irrelevant probe is now processed as unrelated. (*See* **Language Comprehension, Methodologies for Studying**; **Lexical Access**; **Word Recognition**)

Table 1. Typical reaction time (RT) results from the probe word paradigm. With a short period between onset of the ambiguous word and onset of the probe word, each of the three RTs was significantly different from the other two. This pattern of a three-way difference between conditions is often observed in the probe word paradigm, although the difference between contextually relevant and irrelevant conditions has not been statistically significant in all studies. With the longer period between the stimuli, the RT in the contextually relevant condition was faster than the other two conditions, which were not significantly different. Data from Van Petten and Kutas (1987)

	Mean naming latency time (ms)	
Probe type	*200 ms SOA*	*700 ms SOA*
Contextually relevant	591 (73)	547 (71)
Contextually irrelevant	617 (77)	562 (69)
Unrelated	635 (85)	571 (71)

Values are in ms, standard deviation in parentheses. SOA, stimulus onset asynchrony, the time between the onset of the ambiguous word and the onset of the probe word.

Nature and Strength of the Context in the CAP Paradigm

The finding of faster RTs for contextually irrelevant probe words than for unrelated probe words has been replicated many times and serves as the central support for the exhaustive access model. However, two other experimental findings are problematic for this model. The first is that the majority of experiments showing faster RTs in the contextually irrelevant condition compared with the unrelated condition also show that RTs in the contextually irrelevant condition are slower than those in the contextually relevant condition. A gradient of RTs across conditions is incompatible with all three of the major models in their purest forms. One possibility is that one of the major models is true, and that the pattern of contextually relevant < contextually irrelevant < unrelated RTs is due to imperfect stimulus construction. For instance, a proponent of the exhaustive access model might suggest that contextually relevant probes receive additional contextual support directly from the sentence context, as well as from the ambiguity itself, and that the former contribution is of no interest. A proponent of the selective access model might suggest that some portion of the contexts used in a given experiment were, in fact, too weak to count as good contexts for the desired meaning of the ambiguity.

The second experimental finding that precludes accepting the pure version of the exhaustive access model is that some studies have found equivalent RTs for contextually irrelevant and relevant words – results supportive of the selective access model (Tabossi and Zardon, 1993; Simpson, 1994). Many arguments in this research area revolve around defining both the nature and the strength of context which can (sometimes) lead to such results. For instance, Seidenberg *et al*. (1982) suggested a three-way distinction between (1) syntactic context ('a rose' and 'he rose'), (2) single-word associative context (as in 'ship–deck' and 'card–deck'), and (3) more global semantic context which is

neither syntactic nor associative, but created by the overall meaning of a sentence. In this view, contexts of type 2 can indeed yield selective access because they work via direct word-to-word links within the mental lexicon, but contexts of types 1 and 3 are ineffective in blocking exhaustive access. A different view is that it is unnecessary to postulate different forms of context and that overall strength of context is the critical factor: strongly predictive contexts yield selective access, whereas weakly predictive contexts (such as 'a' or 'to') are much the same as no context at all (Paul *et al*., 1992).

The CAP Paradigm and Retroactive Context

Another important issue concerns the logic of the CAP paradigm itself. The standard interpretation of the paradigm is that reaction times to the probe reveal how the ambiguous word was initially interpreted, and that the probe itself has no additional impact. This interpretation assumes that the processing of sequential words is both serial and linear, so that each word can benefit only from what was presented earlier. However, several experiments with unambiguous words have shown that reaction times can be speeded, and error rates reduced, by the presentation of a related word shortly after the target word (Kiger and Glass, 1983; Dark, 1988; Van Petten and Kutas, 1991; Logan and Schulkind, 2000). These retroactive semantic context effects (sometimes called 'backward priming') are best understood in terms of temporal overlap in the processing of two words: if a second item is presented while the first is still being processed, both will benefit from their semantic relationship, much as a simultaneous pair of words are processed more rapidly if related than unrelated. A retroactive-context interpretation of the CAP paradigm suggests that access to a contextually irrelevant meaning is not a spontaneous response to reading an ambiguity, but is instead caused by presentation of the contextually irrelevant probe itself. Although retroactive context effects can clearly occur, the impact of this phenomenon in the CAP paradigm has been contentious (Burgess *et al*., 1989).

Event-related Brain Potential Measures in the CAP Paradigm

An undisputed finding of the CAP paradigm is that the results are sensitive to small differences in the timing of stimulus presentation, so that one appeal of the paradigm is the possibility of tracking the time course of language processing. A dependent measure that is particularly well suited to examining time course is the event-related potential (ERP), a measure of brain electrical activity that can be recorded as participants read or listen for comprehension. The ERP provides a continuous record of brain activity, so that one can describe the temporal onset and offset of experimental effects with some precision. One component of the ERP is sensitive to semantic context. The N400 (a negative voltage peaking 400 ms after stimulus onset) is elicited by all words, but is smaller in amplitude when a word fits with an established semantic context than when it does not (Van Petten and Kutas, 1994). (*See* **Event-related Potentials and Mental Chronometry**)

When a long period between the onset of one stimulus and the next (stimulus onset asynchrony, SOA) is used in the CAP paradigm, both contextually irrelevant and unrelated probe words elicit larger N400s than do contextually related probes (Figure 1a). These long SOA results are a perfect match with the typical pattern of reaction times. With a short SOA, the exhaustive access model predicts a large N400 only for unrelated probes, and equally small N400s for contextually relevant and irrelevant probes. The results in Figure 1b are somewhat different from this prediction. The semantic relationship between ambiguity and probe word influences the ERP beginning at 300 ms after the onset of the probe word. During the initial phase of the context effect, from 300 ms to 500 ms, the ERP to contextually irrelevant probes is similar to the response to completely unrelated words, indicating early access to only the contextually relevant meaning. Only some time later, from 500 ms to 1100 ms, does the contextually irrelevant response grow to resemble the contextually relevant response. This pattern of results is consistent with selective access, combined with a retroactive context effect. The contextually relevant sense of the ambiguity has a head start in influencing probe word processing, but temporal overlap in the processing of the contextually irrelevant sense and its related probe word is visible later.

EYE MOVEMENT MEASURES OF LEXICAL AMBIGUITY RESOLUTION

Gaze durations during reading have also been applied to the problem of ambiguity resolution. Readers generally spend longer fixating words that are contextually unpredictable or low in frequency of use. Gaze duration measures thus offer a naturalistic way of examining a reader's processing difficulty, without the need for an extra reaction

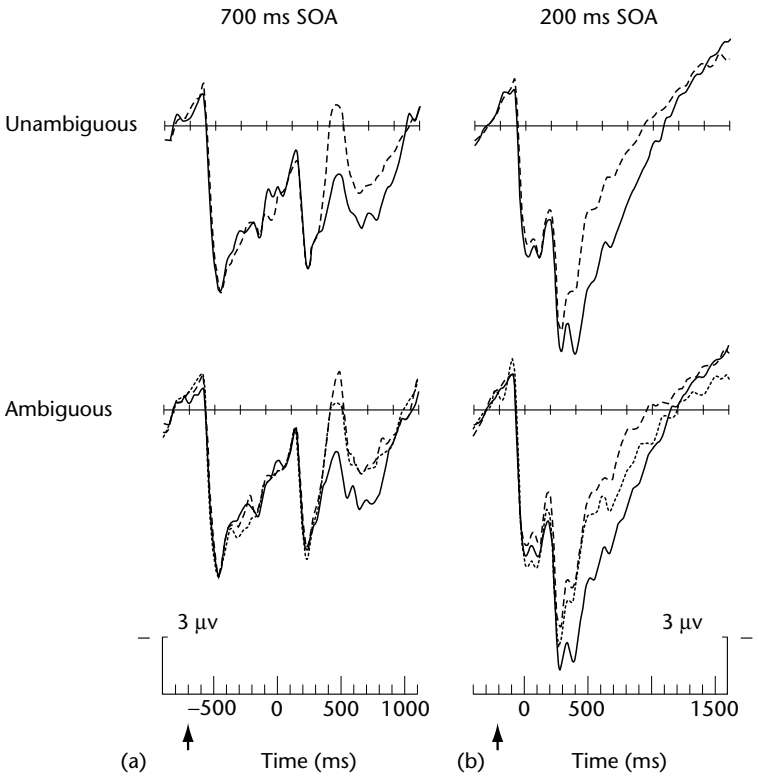


Figure 1. Grand average event-related brain potentials (ERPs) to ambiguous and unambiguous sentence terminal words and subsequent probe words which were unrelated (dashed line), contextually irrelevant (dotted line) or contextually relevant (solid line). Onset of the sentence terminal words is indicated by an arrow; onset of the probe words is at 0 ms. The ERPs were recorded at a midline central scalp site. Stimulus onset asynchrony (SOA) is 700 ms in (a), 200 ms in (b). Data from Van Petten and Kutas (1987); see also Van Petten (1995) for replication.

time task. Because they do not include probe words, these measures thus avoid the possibility of retroactive semantic context effects.

In the absence of prior semantic context, gazes are longer on ambiguous words with two equally strong meanings (balanced homographs) than on unambiguous words, or on homographs with one dominant meaning. If subsequent portions of a sentence favor the subordinate meaning of an unbalanced homograph, readers spend a particularly long time on disambiguating regions. These findings suggest that readers favor the more common sense of an ambiguity on first reading, and must do some extra work if the two meanings are in close competition, or if their initial choice of meaning turns out to be incorrect.

When prior context biases one meaning, gaze durations for balanced homographs are like unambiguous words, indicating that readers no longer suffer a conflict between two equally likely interpretations (an apparent case of selective access). However, when the prior context biases the subordinate sense of an unbalanced homograph, gaze durations are longer on the homograph than on unambiguous words (Pacht and Rayner, 1993). The latter finding suggests that readers may access the dominant meaning despite the context, supporting the ordered access model. However, others have argued that this result depends on relatively weak contexts, and that strong contexts can produce selective access for even subordinate meanings (Kellas and Vu, 1999). Overall, the gaze duration research has lent little support to the exhaustive access model, but has instead indicated that meaning frequency and strength of semantic context are critical factors in determining which (and how many) meanings of ambiguous words are considered by readers.

CONCLUSION

Disagreements in the domain of lexical ambiguity resolution reflect our imperfect understanding of the sequence of events that yields the comprehension of any word. Future language research that sheds light on the nature of semantic relationships

(thus allowing a better quantification of 'context'), and a detailed understanding of the time course of processing from sensory analysis to semantic integration are also likely to clarify how readers and listeners know when 'spoke' means 'part of a wheel' and when it means 'talked'. (*See* **Language Comprehension**; **Psycholinguistics**)

References

Britton BK (1978) Lexical ambiguity of words used in English text. *Behavior Research Methods and Instrumentation* **10**: 1–7.

Burgess C, Tanenhaus MK and Seidenberg MS (1989) Context and lexical access: implications of nonword interference for lexical ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**: 620–632.

Dark VJ (1988) Semantic priming, prime reportability, and retroactive priming are interdependent. *Memory and Cognition* **16**: 299–308.

Dixon P and Twilley LC (1999) Context and homograph meaning resolution. *Canadian Journal of Experimental Psychology* **53**: 335–346.

Kawamoto AH (1993) Nonlinear dynamics in the resolution of lexical ambiguity: a parallel distributed processing account. *Journal of Memory and Language* **32**: 474–516.

Kellas K and Vu H (1999) Strength of context does modulate the subordinant bias effect: a reply to Binder and Rayner. *Psychonomic Bulletin and Review* **6**: 511–517.

Kiger JI and Glass AL (1983) The facilitation of lexical decisions by a prime occurring after the target. *Memory and Cognition* **11**: 356–365.

Klein DE and Murphy GL (2001) The representation of polysemous words. *Journal of Memory and Language* **45**: 259–282.

Lee CJ (1990) Some hypotheses concerning the evolution of polysemous words. *Journal of Psycholinguistic Research* **19**: 211–219.

Li P (1998) Crosslinguistic variation and sentence processing: the case of Chinese. In: Hillert D (ed.) *Sentence Processing: A Crosslinguistic Perspective*, pp. 33–51. San Diego, CA: Academic Press.

Logan GD and Schulkind MD (2000) Parallel memory retrieval in dual-task situations: I. Semantic memory. *Journal of Experimental Psychology: Human Perception and Performance* **26**: 1072–1090.

Pacht JM and Rayner K (1993) The processing of homphonic homographs during reading: evidence from eye movement studies. *Journal of Psycholinguistic Research* **22**: 251–271.

Paul ST, Kellas G, Martin M and Clark MB (1992) Influence of contextual features on the activation of ambiguous word meanings. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**: 703–717.

Seidenberg MS, Tanenhaus MJ, Leiman JM and Bienkowski M (1982) Automatic access of the meanings of ambiguous words in context: some limitations of knowledge-based processing. *Cognitive Psychology* **14**: 538–559.

Simpson GB (1994) Context and the processing of ambiguous words. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 359–371. San Diego, CA: Academic Press.

Swinney DA (1979) Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* **18**: 645–659.

Tabossi P and Zardon F (1993) Processing ambiguous words in context. *Journal of Memory and Language* **32**: 359–372.

Van Petten C (1995) Words and sentences: event-related brain potential measures. *Psychophysiology* **32**: 511–525.

Van Petten C and Kutas M (1987) Ambiguous words in context: an event-related potential analysis of the time course of meaning activation. *Journal of Memory and Language* **26**: 188–208.

Van Petten C and Kutas M (1991) Electrophysiological evidence for the flexibility of lexical processing. In: Simpson GB (ed.) *Understanding Word and Sentence*, pp. 129–174. Amsterdam: Elsevier.

Van Petten C and Kutas M (1994) Psycholinguistics electrified: event-related brain potential investigations. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 83–144. San Diego, CA: Academic Press.

Yip MCW (2001) A preliminary study of subjective frequency estimates of words spoken in Cantonese. *Psychological Reports* **88**: 1253–1258.

Zipf GK (1945) The meaning-frequency relationship of words. *Journal of General Psychology* **33**: 251–256.

Further Reading

Fabiani M, Gratton G and Coles MGH (2000) Event related potentials. In: Cacioppo JT, Tassinary LG and Berntson GG (eds) *Handbook of Psychophysiology*, 2nd edn, pp. 53–84. Cambridge, UK: Cambridge University Press.

Gaskell MGG and Marslen-Wilson WD (2001) Lexical ambiguity resolution and spoken word recognition: bridging the gap. *Journal of Memory and Language* **44**: 325–349.

Kutas M, Federmeier KD, Coulson S, King JW and Munte TF (2000) Language. In: Cacioppo JT, Tassinary LG and Berntson GG (eds) *Handbook of Psychophysiology*, 2nd edn, pp. 576–601. Cambridge, UK: Cambridge University Press.

Rayner K and Sereno SC (1994) Eye movements in reading: psycholinguistic studies. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 57–82. San Diego, CA: Academic Press.

Small SI, Cottrell GW and Tanenhaus MK (1988) *Lexical Ambiguity Resolution. Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann.

Linguistic Relativity

Intermediate article

Lera Boroditsky, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

CONTENTS

Languages differ dramatically from one another in terms of how they describe the world. Does having different ways of describing the world lead speakers of different languages also to have different ways of thinking about the world?

DOES LANGUAGE SHAPE THOUGHT?

Humans communicate with one another using an amazing array of languages, and each language differs from the next in innumerable ways (from obvious differences in pronunciation and vocabulary to more subtle differences in grammar). For example, to say that 'the elephant ate the peanuts' in English, we must include tense – the fact that the event happened in the past. In Mandarin and Indonesian, indicating when the event occurred would be optional and couldn't be included in the verb. In Russian, the verb would need to include tense and also whether the peanut-eater was male or female (though only in the past tense), and whether said peanut-eater ate all of the peanuts or just a portion of them. In Turkish, on the other hand, one would specify (as a suffix on the verb) whether the eating of the peanuts was witnessed or if it was hearsay. It appears that speakers of different languages have to attend to and encode strikingly different aspects of the world in order to use their language properly (Sapir, 1921; Slobin, 1996). Do these quirks of languages affect the way their speakers think about the world? Do English, Mandarin, Russian, and Turkish speakers end up attending to, partitioning, and remembering their experiences differently simply because they speak different languages?

The idea that thought is shaped by language is most commonly associated with the writings of Benjamin Lee Whorf (Whorf, 1956). Whorf, impressed by linguistic diversity, proposed that the categories and distinctions of each language enshrine a way of perceiving, analyzing, and acting in the world. In so far as languages differ, their speakers too should differ in how they perceive and act in objectively similar situations. This strong Whorfian view – that thought and action are entirely determined by language – has long been abandoned in the field. However, definitively answering less deterministic versions of the 'does language shape thought' question has proven to be a very difficult task. Some studies have claimed evidence to the affirmative (e.g. Boroditsky, 2001; Bowerman, 1996; Davidoff *et al*., 1999; Gentner and Imai, 1997; Levinson, 1996; Lucy, 1992; Dehaene *et al*., 1999), while others report evidence to the contrary (e.g. Heider, 1972; Malt *et al*., 1999; Li and Gleitman, 2002).

In recent years, research on linguistic relativity has enjoyed a considerable resurgence, and much new evidence regarding the effects of language on thought has become available. This chapter reviews several lines of evidence regarding the effects of language on people's representations of space, time, substances, and objects.

SPACE

Languages differ considerably in how they describe spatial relations. Many such differences have been noted among English, Dutch, Finnish, Korean, and Spanish, among others (Bowerman, 1996). For example, English distinguishes between putting things into containers ('the apple *in* the bowl', 'the letter *in* the envelope') and putting things onto surfaces ('the apple *on* the table', 'the magnet *on* the refrigerator door'). Cross-cutting this containment/support distinction, Korean distinguishes between tight and loose fit or attachment. For example, putting an apple *in* a bowl requires a different relational term (*nehta*) from putting a letter *in* an envelope (*kitta*), because the first is an example of loose containment and the second an example of tight fit. Further, putting a letter

in an envelope and putting a magnet *on* the refrigerator are both described by *kitta* because both involve close fit.

To test whether these cross-linguistic differences are reflected in the way English and Korean speakers represent spatial relations, McDonough *et al*. (2000) showed scenes involving tight or loose fit to Korean- and English-speaking adults. After they had seen a few examples of either tight fit or loose fit, the subjects were shown an example of tight fit on one screen, and an example of loose fit on another. While Korean-speaking adults looked longer at the kind of spatial relation they had just been familiarized with, English speakers did not distinguish between the tight- and loose-fit scenes, looking equally long at the familiar and novel scenes. Further, when given several examples of tight fit and one example of loose fit (or vice versa), Korean adults could easily pick out the odd picture, but English speakers could not. Finally, McDonough *et al*. found that unlike adult English speakers, prelinguistic infants (being raised in both English-speaking and Korean-speaking households) distinguished between tight and loose fit in the looking-time test described above. This pattern of findings suggests that infants may come ready to attend to any number of spatial distinctions. However, as people learn and use language, the spatial distinctions reinforced by their particular language are the ones that remain salient in their representational repertoire.

Dramatic cross-linguistic differences have also been noted in the way languages describe spatial locations (Levinson, 1996). Whereas most languages (e.g. English, Dutch) rely heavily on relative spatial terms to describe the relative locations of objects (e.g. left/right, front/back), Tzeltal (a Mayan language) relies primarily on absolute reference (a system similar to the English north/south direction system). Spatial locations that are north are said to be downhill, and those south are said to be uphill. This absolute uphill/downhill system is the dominant way to describe spatial relations between objects in Tzeltal; no relational equivalents to the English terms front/back or left/right are available (Levinson, 1996).

To test whether this difference between the two languages has cognitive consequences, Levinson (1996) tested Dutch and Tzeltal speakers in a number of spatial tasks. In one study, participants were seated at a table and an arrow lay in front of them pointing either to the right (north) or to the left (south). They were then rotated 180 degrees to a second table which had two arrows (one pointing to the left (north) and one to the right (south)), and were asked to identify the arrow 'like the one they saw before'. Dutch speakers overwhelmingly chose the 'relative' solution. If the stimulus arrow pointed to the right (and north), Dutch speakers chose the arrow that still pointed to the right (though it now pointed south instead of the original north). Tzeltal speakers did exactly the opposite, overwhelmingly choosing the 'absolute' solution. If the stimulus arrow pointed to the right (and north), Tzeltal speakers chose the arrow that still pointed north (though it now pointed left instead of right). Thus, Tzeltal speakers' heavy reliance on absolute reference in spatial description appears to have affected their interpretation of (and performance on) a nonlinguistic orientation task.

Further studies of this task showed that English speakers (English is the same as Dutch in this respect) do not always favor relative responses; certain contextual factors can be used to induce English speakers to produce both absolute and relative responses on these tasks (Li and Gleitman, 2002). This is not surprising since English speakers use both absolute and relative forms in their language. It remains to be seen whether the same contextual factors can induce Tzeltal speakers to produce relative responses despite an apparent lack of relative terms in Tzeltal.

In summary, the evidence available so far suggests that reference frames and distinctions made available by one's language may indeed impose important constraints on one's spatial thinking.

TIME

Languages also differ from one another on their descriptions of time. While all languages use spatial terms to talk about time ('looking *forward* to a brighter tomorrow', 'proposing theories *ahead* of our time', 'falling *behind* schedule'), different languages use different spatial terms. For example, in English, we predominantly use front/back terms to talk about time. We can talk about the good times *ahead* of us, or the hardships *behind* us. We can move meetings *forward*, push deadlines *back*, and eat dessert *before* we're finished with our vegetables. On the whole, the terms used to order events are the same as those used to describe asymmetric horizontal spatial relations (e.g. 'he took three steps *forward*' or 'the path is *behind* the store'). In Mandarin, front/back spatial metaphors for time are also common (Scott, 1989). Mandarin speakers use the spatial morphemes *qián* (front) and *hòu* (back) to talk about time. What makes Mandarin interesting for present purposes is that Mandarin speakers also systematically use vertical metaphors

to talk about time (Scott, 1989). The spatial morphemes *shàng* (up) and *xià* (down) are frequently used to talk about the order of events, roughly translated into English as *last* and *next*. Earlier events are said to be *shàng* or 'up', and later events are said to be *xià* or 'down'. In summary, both Mandarin and English speakers use horizontal terms to talk about time. In addition, Mandarin speakers commonly use the vertical terms *shàng* and *xià*.

So, do the English and Mandarin ways of talking about time lead to differences in how people think about time? Specifically, are Mandarin speakers more likely to construct vertical timelines to think about time, while English speakers are more likely to construct horizontal timelines? A collection of studies showed that Mandarin speakers tend to think about time vertically even when thinking for English (Boroditsky, 2001). For example, Mandarin speakers were faster to confirm that March comes earlier than April if they had just seen a vertical array of objects than if they had just seen a horizontal array. The reverse was true for English speakers. Another study showed that the extent to which Mandarin–English bilinguals think about time vertically is related to how old they were when they first began to learn English. In another experiment native English speakers were taught to talk about time using vertical spatial terms in a way similar to Mandarin. On a subsequent test, this group of English speakers showed the same bias to think about time vertically as was observed with Mandarin speakers.

This last result suggests two things: (1) language is a powerful tool in shaping thought, and (2) one's native language plays a role in shaping habitual thought (how we tend to think about time, for example) but does not completely determine thought in the strong Whorfian sense (since one can always learn a new way of talking, and with it, a new way of thinking).

SHAPES AND SUBSTANCES

Languages also differ in the extent to which they make a grammatical distinction between objects and substances. For example, in English, objects like candles and chairs have distinct singular and plural forms (e.g. one candle versus two candles), but substances like mud and wax do not. Further, objects and substances are distinguished in English in counting. While one can say 'one candle, two candles, three candles' and so on, counting substances is a bit trickier. Instead of saying 'one mud, two muds', English speakers must specify the unit of measurement such as 'one mound of mud' or 'one cup of mud' (words like 'mound' and 'cup' here are called 'unitizers' because they specify the unit of measurement).

Unlike English, some languages do not have a grammatical boundary between objects and substances. In Yucatec Mayan, for example, all nouns act almost as if they refer to substances. All nouns require a unitizer when counting (usually specifying shape or form, for example 'one long thin unit'), and don't necessarily need to take distinct plural and singular forms (Lucy and Gaskins, 2001). This means that 'two candles' in English is more like 'two long thin units of wax' in Yucatec. Does talking about objects as if they were substances in their language lead Yucatec Mayans to attend more to the materials and substances that comprise the objects? Several studies suggest that this is indeed the case (e.g. Lucy and Gaskins, 2001). English speakers and Yucatec Mayans were shown an example object (e.g. a plastic comb with a handle) and asked to choose which of two other objects was more similar to this example. The two choices varied from the example either in shape (a plastic comb with no handle), or in material (a wooden comb with a handle). English speakers preferred the shape match, saying that the two combs with a handle were more similar (even though they were made of different materials). Yucatec Mayans, on the other hand, preferred the material match, saying that the two plastic combs were more similar (even though they differed in shape). These findings suggest that aspects of grammar can in fact shape the way speakers of a language conceptualize the shapes and materials of objects.

OBJECTS

Finally, languages also differ in how names of objects are grouped into grammatical categories. One such common feature of languages is grammatical gender. Unlike English, many languages have a grammatical gender system whereby all nouns (e.g. penguins, pockets, and toasters) are assigned a gender. Many languages only have masculine and feminine genders, but some also assign neuter, vegetative, and other more obscure genders. When speaking a language with grammatical gender, speakers are required to mark objects as gendered through definite articles and gendered pronouns, and often need to modify adjectives or even verbs to agree in gender with the nouns. Does talking about inanimate objects as if they were masculine or

feminine actually lead people to think of inanimate objects as having a gender?

A recent set of studies suggests that the grammatical genders assigned to objects by a language do indeed influence people's mental representations of objects (Boroditsky *et al*., in press). For example, Spanish and German speakers were asked to rate similarities between pictures of people (males or females) and pictures of objects (the names of which had opposite genders in Spanish and German). Both groups rated grammatically feminine objects to be more similar to females and grammatically masculine objects more similar to males. This was true even though all objects had opposite genders in Spanish and German, the test was completely nonlinguistic (conducted entirely in pictures with instructions given in English), and even when subjects performed the task during a verbal suppression manipulation (which would interfere with their ability to subvocally name the objects in any language). Other studies demonstrated that Spanish and German speakers also ascribe more feminine or more masculine properties to objects depending on their grammatical gender. For example, asked to describe a 'key' (a word masculine in German and feminine in Spanish), German speakers were more likely to use words like 'hard, heavy, jagged, metal, serrated, and useful', while Spanish speakers were more likely to say 'golden, intricate, little, lovely, shiny, and tiny'. To describe a 'bridge', on the other hand, (a word feminine in German and masculine in Spanish), German speakers said 'beautiful, elegant, fragile, peaceful, pretty, and slender', while Spanish speakers said 'big, dangerous, long, strong, sturdy, and towering'. These findings once again indicate that people's thinking about objects is influenced by the grammatical genders their native language assigns to the objects' names. It appears that even a small fluke of grammar (the seemingly arbitrary assignment of a noun to be masculine or feminine) can have an effect on how people think about things in the world.

CONCLUSION

Languages appear to influence many aspects of human cognition: evidence regarding space, time, objects, and substances has been reviewed in this article, but further studies have also found effects of language on people's understanding of numbers, colors, shapes, events, and other minds. Considering the many ways in which languages differ, the findings reviewed here suggest that the private mental lives of people who speak different languages may differ much more than previously thought.

Beyond showing that speakers of different languages think differently, these results suggest that linguistic processes are pervasive in most fundamental domains of thought. That is, it appears that what we normally call 'thinking' is in fact a complex set of collaborations between linguistic and nonlinguistic representations and processes. Further research into linguistic relativity may help uncover the exact nature of the interactions between these many processes in the service of complex cognitive function, as well as help us to establish what might be core or universal in human cognition.

References

Boroditsky L (2001) Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology* **43**(1): 1–22.

Boroditsky L, Schmidt L, and Phillips W (in press). In: Gentner D and Goldin-Meadow S (eds) *Language in Mind: Advances in the Study of Language and Cognition*. Cambridge, MA: MIT Press.

Bowerman M (1996) The origins of children's spatial semantic categories: cognitive versus linguistic determinants. In: Gumperz J and Levinson S (eds) *Rethinking Linguistic Relativity*, pp. 145–176. Cambridge, MA: Cambridge University Press.

Davidoff J, Davies I and Roberson D (1999) Colour categories of a stone-age tribe. *Nature* **398**: 203–204.

Dehaene S, Spelke E, Pinel P, Stanescu R and Tsivkin S (1999) Sources of mathematical thinking: behavioral and brain-imaging evidence. *Science* **284**: 970–974.

Gentner D and Imai M (1997) A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition* **62**(2): 169–200.

Heider E (1972) Universals in color naming and memory. *Journal of Experimental Psychology* **93**: 10–20.

Levinson S (1996) Frames of reference and Molyneux's question: crosslinguistic evidence. In: Bloom P and Peterson M (eds) *Language and Space*, pp. 109–169. Cambridge, MA: MIT Press.

Li P and Gleitman L (2002) Turning the tables: language and spatial reasoning. *Cognition* **83**: 265–294.

Lucy J (1992) *Grammatical categories and Cognition: a Case Study of the Linguistic Relativity Hypothesis*. Cambridge, UK: Cambridge University Press.

Lucy J and Gaskins S (2001) Grammatical categories and the development of classification preferences: a comparative approach. In: Bowerman M and Levinson S (eds) *Language Acquisition and Conceptual Development*, pp. 257–283. Cambridge, UK: Cambridge University Press.

Malt B, Sloman S, Gennari S, Shi M and Wang Y (1999) Knowing versus naming: similarity and the linguistic categorization of artifacts. *Journal of Memory and Language* **40**: 230–262.

McDonough L, Choi S and Mandler J (2000) Development of language-specific categorization of spatial relations from prelinguistic to linguistic stage: a preliminary study. Paper presented at the Finding the Words Conference at Stanford University, Stanford, California, April, 2002.

Sapir E (1921) *Language*. New York, NY: Harcourt, Brace, and World.

Scott A (1989) The vertical dimension and time in Mandarin. *Australian Journal of Linguistics* **9**: 295–314.

Slobin D (1996) From 'thought and language' to 'thinking for speaking'. In: Gumperz J and Levinson S (eds) *Rethinking Linguistic Relativity*, pp. 70–96. Cambridge, MA: Cambridge University Press.

Whorf B (1956) *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, edited by Carroll JB. Cambridge, MA: MIT Press.

Further Reading

Boroditsky L, Ham W and Ramscar M (2002) What is universal about event perception? Comparing English and Indonesian speakers. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Bowerman M and Levinson S (eds) (2001) *Language Acquisition and Conceptual Development*. Cambridge, UK: Cambridge University Press.

de Villiers JG (in press) Language and theory of mind: what is the developmental relationship? In: Baron-Cohen S, Tager-Flusberg H and Cohen D (eds) *Understanding Other Minds: Perspectives from Autism and Developmental Cognitive Neuroscience*. Cambridge, UK: Cambridge University Press.

Gentner D and Goldin-Meadow S (in press) *Language in Mind: Advances in the Study of Language and Cognition*. Cambridge, MA: MIT Press.

Gumperz J and Levinson S (eds) (1996) *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.

Hermer-Vasquez L, Spelke ES and Katsnelson AS (1999) Sources of flexibility in human cognition: dual-task studies of space and language. *Cognitive Psychology* **39**: 3–36.

Roberson D, Davidoff J and Shapiro L (in press) Squaring the circle: the cultural relativity of good shape. *Journal of Culture and Cognition*.

Masking

Intermediate article

Gregory Francis, Purdue University, West Lafayette, Indiana, USA

CONTENTS

Masking refers to a class of phenomena where presentation of one stimulus (the mask) can impair performance on some task that requires judgment about another stimulus (the target).

INTRODUCTION

Visual masking plays two roles in cognitive science. First, masking is used to investigate properties of the visual system. By identifying the way in which the target and mask stimuli influence each other, vision scientists are able to deduce details about the underlying mechanisms involved in visual perception. Second, visual masking is used to indirectly restrict systems involved in information processing of visual stimuli. The logic of this type of approach is that the mask can halt further processing of the target, and one can thereby explore the order and time course of many information-processing systems. A special subset of this approach is based on evidence that persons with various types of cognitive disorders may respond differently from normal subjects under some masking conditions. Thus, masking has been proposed as a simple method for detecting some disorders, and as a means of specifying the underlying mechanisms for those disorders.

FORWARD AND BACKWARD VISUAL MASKING

The distinction between forward and backward masking refers to the presentation order of the target and mask stimuli. In forward masking the mask stimulus onset precedes the target stimulus onset. In backward masking the mask stimulus onset follows the target stimulus onset. Of course, for some combinations of stimulus durations, the mask presentation may both precede the onset of the target and continue after offset of the target. Thus, it is not always appropriate to consider forward and backward masking as entirely distinct; they may contain many similar components. Forward masking is usually weaker than backward masking, and backward masking has more interesting properties and is more useful as a tool for investigating cognitive systems.

Both the target and mask stimuli are usually very brief (often less than 200 milliseconds). Despite its short duration, if the target stimulus is presented by itself, it is clearly visible and it is easy for observers to perform whatever judgment is required of them. What is interesting is that the subsequent presentation of a mask stimulus, even a hundred milliseconds after the target has turned off, can make the observer's task of judging something about the target exceedingly difficult. When this effect was first noted by Stigler (1910), it forced the field of perceptual psychology to realize that processing of visual information took time and could be interrupted. Backward masking has been used to identify details of the perceptual and cognitive processes involved in building percepts and judgments.

Types of Masks

Historically, masking has often been distinguished by the physical characteristics of the mask. For example, with fixed target stimuli and a fixed task (say, identification of a letter as D or O), the mask could be any of a number of stimuli. When the mask consists of a stimulus whose contours surround the target contours, but which does not overlap the target stimulus (e.g. an annulus around the target letter), the mask is called a metacontrast mask. (The term 'metacontrast masking' is more restrictive as it refers only to backward masking and not to a situation where the mask precedes the target, which is called paracontrast.) When the mask contours substantially overlap the contours of the target (e.g. a jumble of oriented lines), the mask is called a pattern mask. A noise mask is similar, but the mask is made of random noise (e.g. black and white dots). Finally, when the mask is a homogeneous field, the masking process is referred to as masking by light.

Pattern masking and metacontrast masking are the most commonly investigated types of masking. The former is often used as a means of curtailing information processing. The latter is often used to explore various properties of the visual system, particularly interactions among systems sensitive to image contours.

Masking Functions

Often the properties of the target and mask stimuli are held fixed, but the stimulus onset asynchrony (SOA) or the interstimulus interval (ISI) between the target and mask is varied. Performance of the observer for a given target-based task is then compared for different SOAs (or ISIs). The resulting set of data is often called a 'masking function'.

The masking function is interesting for two reasons. First, for those studies that use masking as a tool for exploring another cognitive process, it can partially reveal the time course of that process. For example, suppose that the observer's task is to distinguish between target letters D and Q. Suppose that when the mask appears right after offset of the target letter (ISI = 0), observers are simply guessing on the letter identity. Presumably, there is some perceptual or cognitive process that is computing information from the target stimulus. That process takes time, and if the mask interferes with that processing before it is completed, the observer will have to guess the identity of the letter. Thus, larger ISIs between the target and mask should eventually allow the observer to identify the target letter above a chance level. The ISI for which this happens can then, tentatively, be identified as a lower bound for the duration of information processing that limits the observer in this task. A masking function that corresponds to this idea would be monotonic increasing from ISI = 0, and is sometimes called a Type A function.

The second interesting characteristic of the masking function is somewhat problematic for the first. Namely, for some targets, masks, and tasks, the masking function is U-shaped. That is, for short ISIs (or SOAs) the target is clearly seen, and the task fairly easy to perform. For middle-duration ISIs (around 80 milliseconds, but it varies substantially), the target is harder to see and the task difficult to perform. For long-duration ISIs the target is seen and interpreted before the mask arrives, and task performance is again quite good. That going from a short-duration ISI to a medium-duration ISI should cause a larger detrimental effect on the observer's task is interesting because it implies that the effect of the mask is not just to interrupt the processing of the target. If the mask simply interrupted target processing, then increases in ISI would be expected to allow for more processing and so better performance on the task (or at least, not worse performance). But worse performance is what is found experimentally.

The U-shaped masking function is most commonly reported for metacontrast masking. However, U-shaped masking functions have also been reported for masking by light and pattern masking (Stewart and Purcell, 1974). For all types of masks, increasing the intensity, duration, and size of the mask tends to produce a monotonic masking function rather than a U-shaped masking function. It is for this reason that anyone intending to use masking as a means of investigating cognitive processes is advised to use strong masks. However, it remains unclear whether the existence of a monotonic-shaped masking function truly indicates that increases in ISI free various cognitive processes for longer durations.

VISUAL CHANNELS

Classically, there have been two theories of backward masking: interruption and integration. An interruption-based theory supposes that the mask stimulus interrupts the cognitive or perceptual processing of the target stimulus. Since that processing takes time, the appearance of the mask can hinder or weaken the observer's task of judging something about the target. An integration-based theory supposes that the mask and target stimuli are treated as a single combination stimulus, and this makes the observer's judgments about the target alone more difficult. These theories are not generally considered to be competing as a given masking situation may contain both interruption and integration effects. One difficulty in working with these theories is that they are described verbally rather than quantitatively. Without a quantitative specification of some underlying mechanisms, it is difficult to identify which theory is relevant for a given situation.

A more specific dual channel theory has developed out of the recognition that psychological theories of transient and sustained visual channels may be related to the magnocellular and parvocellular visual channels identified in animal neurophysiology (Breitmeyer and Ganz, 1976). This theory posits that the percept of a stimulus is largely related to representation of information in the sustained pathway. This representation tends to develop more slowly and last longer than the

corresponding representation in the transient pathway. Inhibition from the transient pathway to the sustained pathway is hypothesized to be an important mechanism for masking. In a backward masking situation, mask-generated transient signals inhibit target-generated sustained signals. In particular, because of the difference in onset times, the transient inhibition will often most strongly overlap the sustained representation when the target and mask are separated by a positive ISI. This is the basis of the U-shaped masking function. To account for other masking effects, this theory has been elaborated to include sustained–sustained inhibition and sustained–transient inhibition. Breitmeyer and Ögmen (2000) discuss some recent developments in this dual channel theory.

Although many parts of the transient–sustained theory correspond to neurophysiological findings, a full neurophysiological basis for masking remains elusive. This is primarily because neurophysiological studies of masking often find that the mask has no effect, or that the effect of the mask does not correspond to behavioral measures (e.g. Rolls *et al*., 1999).

Masking has been used to explore relationships between conscious and unconscious processing of visual information. Fehrer and Raab (1962) first reported that reaction time for responding to a target stimulus was unaffected by the presence or absence of a mask stimulus, even when the mask was so strong as to cause the observer to claim the target was invisible. This result led to an investigation of unconscious processing, which was hotly debated for many years. Recent work by Klotz and Neumann (1999) now seems to show convincingly that unconscious processing of the target does proceed, even when the mask renders it consciously absent. They demonstrated that a stimulus that was so completely masked that observers could not identify it could still have a priming effect on a subsequently presented stimulus by modifying the observer's reaction time. These results seem to be tantalizingly consistent with a theory of separate visual channels for perception and action (Goodale and Milner, 1992). In this theory, it would be possible for the visual channel related to action to process the target, even though the channel related to perception had its representations masked out. However, priming of masked stimuli seems to exist even when the priming information should be, in the theory, only in the perceptual visual channel (Klotz and Neumann, 1999).

BACKWARD MASKING PSYCHOPATHOLOGY

Some research suggests that backward visual masking may be a trait marker of various pathologies. For example, Merritt *et al*. (1986) showed that schizotypic observers (defined by other tests that suggest a person has a tendency to develop schizophrenic traits) tend to be more susceptible to backward masking than normal subjects or observers with other psychiatric conditions. Similar results are found in schizophrenia patients (e.g., Suslow and Arolt, 1998).

Masking differences between individuals with some disorder or tendency to a disorder and normal subjects have also been noted in studies of reading impairment (e.g., Boden and Brodeur, 1999), bipolar disorder (e.g., Tam *et al*., 1998), Type I diabetes (Muise *et al*., 1997), Alzheimer disease (Cronin-Golomb, 1995), and mania (Green *et al*., 1994).

The general application of masking to a range of disorders reflects the ability of masking to influence a wide variety of different cognitive functions. Thus, when those disorders influence different cognitive abilities, some type of masking study will be likely to reveal those differences.

References

Boden C and Brodeur D (1999) Visual processing of verbal and nonverbal stimuli in adolescents with reading disabilities. *Journal of Learning Disabilities* **32**: 58–71.

Breitmeyer B and Ganz L (1976) Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review* **83**: 1–36.

Breitmeyer B and Ögmen H (2000) Recent models and findings in visual backward masking: a comparison, review, and update. *Perception & Psychophysics* **62**: 1572–1595.

Cronin-Golomb A (1995) Vision in Alzheimer's disease. *Gerontologist* **35**: 370–376.

Fehrer E and Raab D (1962) Reaction time to stimuli masked by metacontrast. *Journal of Experimental Psychology* **64**: 126–130.

Goodale MA and Milner AD (1992) Separate visual pathways for perception and action. *Trends in Neurosciences* **15**: 20–25.

Green MF, Nuechterlein KH and Mintz J (1994) Backward masking in schizophrenia and mania: I. Specifying a mechanism. *Archives of General Psychiatry* **51**: 939–944.

Klotz W and Neumann O (1999) Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance* **25**: 976–992.

Merritt RD, Balogh DW and Leventhal DB (1986) Use of a metacontrast and paracontrast procedure to assess the visual information processing of hypothetically schizotypic college students. *Journal of Abnormal Psychology* **95**: 74–80.

Muise JG, Blanchard L, DesRosiers M, Watier C and Pelletier J (1997) Achromatic visual backward masking of colored stimuli in Type I diabetes. *Psychological Reports* **81**: 771–780.

Rolls ET, Tovee MJ and Panzeri S (1999) The neurophysiology of backward visual masking: information analysis. *Journal of Cognitive Neuroscience* **11**: 300–311.

Stewart AL and Purcell DG (1974) Visual backward masking by a flash of light: a study of U-shaped detection functions. *Journal of Experimental Psychology* **103**: 553–566.

Stigler R (1910) Chronophotische Studien über den Umgebungskontrast. *Pfügers Archiv für die gesamte Physiologie* **134**: 365–435.

Suslow T and Arolt V (1998) Backward masking in schizophrenia: time course of visual processing deficits during task performance. *Schizophrenia Research* **33**: 79–86.

Tam WC, Sewell KW and Deng H (1998) Information processing in schizophrenia and bipolar disorder: a discriminant analysis. *Journal of Nervous & Mental Disease* **186**: 597–603.

Further Reading

Alpern M (1952) Metacontrast: historical introduction. *American Journal of Optometry* **29**: 631–646.

Bachmann T (1994) *Psychophysiology of Visual Masking: The Fine Structure of Conscious Experience*. Commack, NY: Nova Science Publishers.

Breitmeyer B (1984) *Visual Masking: An Integrative Approach*. New York, NY: Oxford University Press.

Enns JT and Di Lollo V (2000) What's new in visual masking? *Trends in Cognitive Sciences* **4**: 345–352

Francis G (1998) Metacontrast masking. Visual Perception Online Laboratory [http://www.psych.purdue.edu/∼coglab/VisLab/]

Francis G (2000) Quantitative theories of metacontrast masking. *Psychological Review* **107**: 768–785.

Turvey MT (1973) On peripheral and central processes in vision: inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review* **80**: 1–52.

Weisstein N (1972) Metacontrast. In: Jameson D and Hurvich L (eds) *Handbook of Sensory Physiology*, vol. 7, no. 4, *Visual Psychophysics*. Berlin, Germany: Springer-Verlag.

Werner H (1935) Studies on contour: I. Qualitative analysis. *American Journal of Psychology* **47**: 40–64.

Mathematical Psychology

Introductory article

Richard A Chechile, Tufts University, Medford, Massachusetts, USA

CONTENTS

Mathematical psychology is broadly defined as the utilization of mathematical and computational modeling methods to measure, describe, and explain psychological processes.

EARLY PROGRESS TOWARDS A MATHEMATICAL PSYCHOLOGY

Gustav Fechner's 1860 psychological treatise was a pioneering demonstration of both mathematical psychology and experimental psychology. However, in the succeeding six decades after Fechner's work, the field of experimental psychology continued to develop, but there was relatively little utilization of mathematical models in psychology. In the 1920s, the field of psychometrics emerged under the leadership of Louis Thurstone, but these models were primarily designed to understand the latent structure of multi-item psychological tests and questionnaires and are different from the type of models that are associated with mathematical psychology. The formal field of mathematical psychology did not develop until after the Second World War.

During the Second World War, a number of psychologists interacted with engineers, physicists, and mathematicians, and this interaction helped foster a desire to create a more rigorous theoretical psychology that employed mathematical methods similar to those used in the physical sciences and in engineering. Moreover, the post-war period saw the development of new branches of applied mathematics such as control theory, cybernetics, information theory, system theory, game theory, and automata theory. These mathematical advances contained great promise for utilization within psychology.

By 1950 a number of psychologists were openly dissatisfied with the prevailing state of scientific psychology, in that it stressed behavioral and empirical fact-gathering within restricted domains of inquiry and did not emphasize the formulation of precise theories to account for experimental findings. There was also a desire to create a theoretical psychology that was different from the test theory emphasis in psychometrics. The pioneers of mathematical psychology in the 1950s sought to develop mathematical and computational models for the psychological processes studied in experimental psychology (e.g. learning, memory, and signal detection).

MATHEMATICAL PSYCHOLOGY AS A DISCIPLINE WITHIN PSYCHOLOGY

In the 1950s, 10 percent of the papers published in the *Psychological Review*, the primary theoretical journal in psychology, represented mathematical psychology research. Also, books appeared that were detailed monographs about mathematical psychology. Starting in the mid-1950s, there were regular summer workshops on the mathematical behavioral sciences held at Stanford University in California, and these meetings helped foster the exchange of information and the development of research collaborations in mathematical psychology. Early in the 1960s, there was an explosion of edited books with high-quality mathematical psychology papers. In 1964, the *Journal of Mathematical Psychology* was initiated with an editorial board of Richard Atkinson, Robert Bush, Clyde Coombs, William Estes, Duncan Luce, William McGill, George Miller, and Patrick Suppes. They selected Richard Atkinson as the first editor. In 1968, the first formal conference on mathematical psychology was held; this annual conference series has continued to the present. After the ninth meeting, an official society was created (the Society for Mathematical Psychology). Consequently, in North America the intellectual infrastructure was in place in the 1960s to support the field of mathematical psychology.

Outside North America, a parallel development in mathematical psychology occurred. In 1965, the *British Journal of Statistical Psychology* changed its name to the *British Journal of Mathematical and Statistical Psychology* in order to reflect the inclusion of mathematical psychology papers along with the more traditional psychometric papers. In 1967, the journal *Mathematical Social Sciences* was founded, with an editorial board including both North American and European social scientists. Furthermore, in 1971 a group of European mathematical psychologists held a conference in Paris. The annual European mathematical psychology meetings have continued to the present, and a number of edited books have ensued from the papers presented at this conference series. Moreover, in 1989 a group of Australian and Asian mathematical psychologists created their own series of research meetings (the Australasian Mathematical Psychology Conference).

RESEARCH IN MATHEMATICAL PSYCHOLOGY

During the 1990s, 39 percent of the papers in *Psychological Review* contained mathematical models. This statistic reflects the salience of mathematical psychology in the primary theoretical journal for psychology. It also reflects a wide range of research within psychology. Because of this breadth, it has not been possible to capture the content of mathematical psychology with any single volume. Researchers have come to specialize within mathematical psychology. Consequently, the research in mathematical psychology is perhaps best described through a series of selected examples of mathematical psychology models. The examples have been chosen to reflect a range of modeling techniques and research topics.

Signal Detection Theory

Signal detection theory (SDT) is perhaps the best known mathematical psychology model, and it is routinely used in experimental psychology. Although there are many variants of SDT, the model that is most widely used is from a seminal book in 1966 by Green and Swets. SDT emerged from the psychophysical framework of detecting a faint stimulus within a noisy background. On each trial of a typical experiment, the observer is presented either with a stimulus embedded in a noisy background or with just the noisy background without the stimulus. On trials where the observer responds positively, the response is either correct (a hit) if the stimulus was actually present or incorrect (a false alarm) if the stimulus is absent. Similarly, if the observer responds negatively, then the person is either correct (a correct rejection) if the stimulus is absent or incorrect (a miss) if the stimulus is present. For faint signals the discrimination is difficult and the responses will depend on the criterion used by the observer. For example, if a person adopts a strict criterion for responding positively, then the observer will commit few false alarms at the cost of missing many stimuli. In general, the data are affected both by the signal strength and by the decision criterion. The goal of SDT is to measure perceived signal strength.

The key concepts involved in SDT are illustrated in Figure 1. It is assumed that there is an underlying strength axis. It is also assumed that even on stimulus-absent trials, there is an activation on the psychological strength axis. The distribution of those responses across all the noise-alone trials is assumed to be characterized by a probability distribution. In the standard model, the probability distribution for the noise-alone condition is assumed to be a normal or Gaussian distribution (i.e. the distribution labeled N in Figure 1 which has mean μ_N and standard deviation σ_N). The noise is due to either random environmental sources or internal neural sources. The second distribution (the one labeled S + N) represents the probability distribution for the stimulus-present trials. It is assumed that the stimulus causes a shift in the underlying response distribution, and it can also cause a change in the standard deviation. The S + N distribution is also assumed to be Gaussian with the mean μ_S and the standard deviation σ_S. The total area under each distribution is 1.0. Areas under portions of a probability distribution correspond to probabilities. The signal sensitivity parameter d'_s is defined as the difference between the two means divided by the average of the standard deviation of the noise distribution and the signal-plus-noise distribution.

Also shown in Figure 1 is a point labeled A on the strength axis. This point represents a decision criterion (i.e. the observer responds positively only if the evoked strength is greater than the strength at point A). The proportion of hits are illustrated by the slanted-dashed section of the S + N distribution that is to the right of A. The proportion of false alarms are illustrated by the horizonal-shaded section of the N distribution (the section to the right of A). As the observer adopts different decision criteria, the hit and false alarm proportions will change accordingly, but the measure of the strength of the stimulus (d'_s) remains constant. It

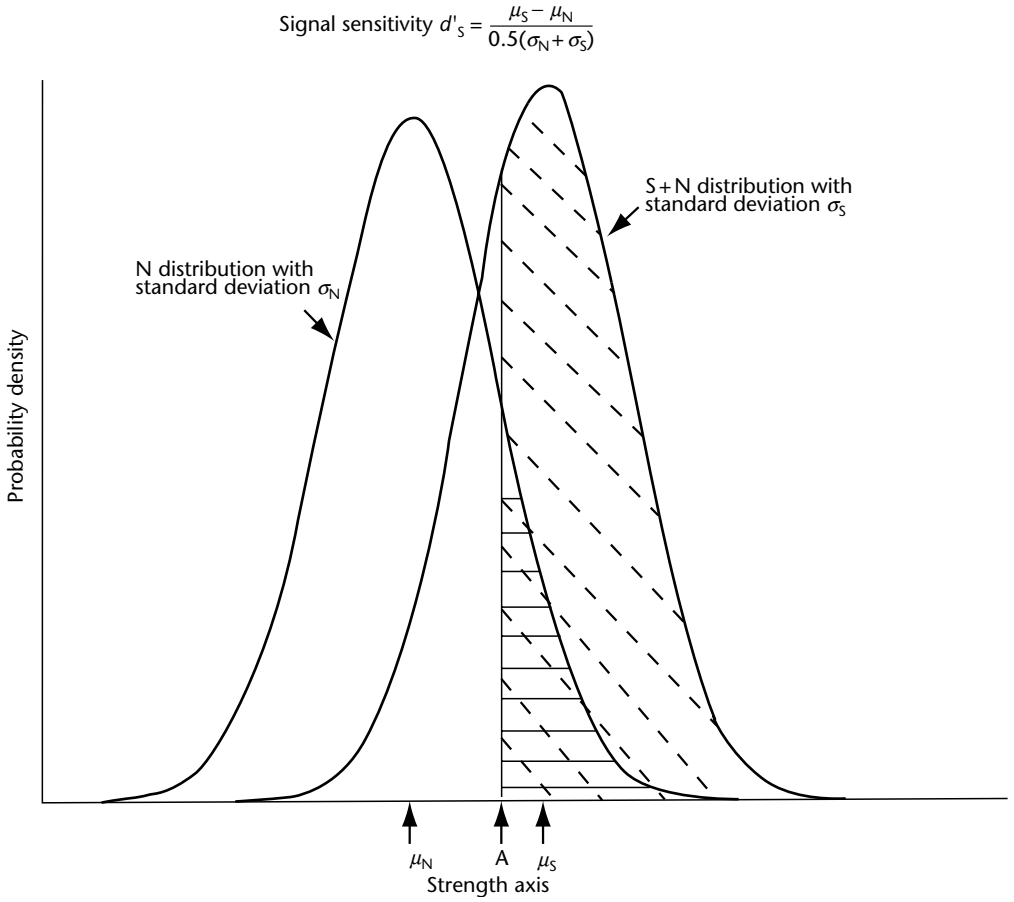


Figure 1. The assumed representation of response distribution according to signal detection theory. N = noise alone, S = stimulus present; μ = mean, σ = standard deviation. Slanted shading represents correct positive responses; horizontal shading represents 'false alarms'.

is possible to estimate both d'_s and the ratio σ_S/σ_N if the experimenter has hit and false-alarm data for more than one decision criterion.

Multinomial Process Tree Models

Multinomial process tree (MPT) models share with signal detection theory the measurement of underlying psychological processes that are not directly measurable. The model parameters are related to psychological processes that determine a set of categorical responses (i.e. the multinomial data structure). The psychological model is a detailed probability tree description as to how the latent processes result in the various proportions for the observed multinomial data. The mathematical psychologist must establish methods for statistically estimating the latent model parameters. William Batchelder and David Riefer have done considerable research on this class of psychological model.

An illustration of a multinomial process tree model is provided in Figure 2. This model is a form of the 1998 model developed by Chechile to obtain separate measures for the probability of sufficient storage, θ_S, and the probability of retrieving a sufficiently stored item from memory, θ_r. Sufficient storage is defined as the state of information representation that would enable the correct recall of the target event, provided that the person could gain access to that information. The probability of sufficient storage is the proportion of times that the memory representation is in this state of sufficient storage. However, even if a memory is sufficiently stored, it may not be recalled because of a problem of memory retrieval.

The model for the estimation of the storage and retrieval probabilities is designed for a very specific task. The participants in a memory experiment are previously shown a series of 'target' items to learn, and at a later time the people are asked to 'free recall' as many of the target items as they can

(a)
Multinomial cells

1	2	3	4	5
correct free recall	correct 4-afc	correct 3-afc	correct 2-afc	incorrect 2-afc

(b)
Tree model in terms of psychological parameters

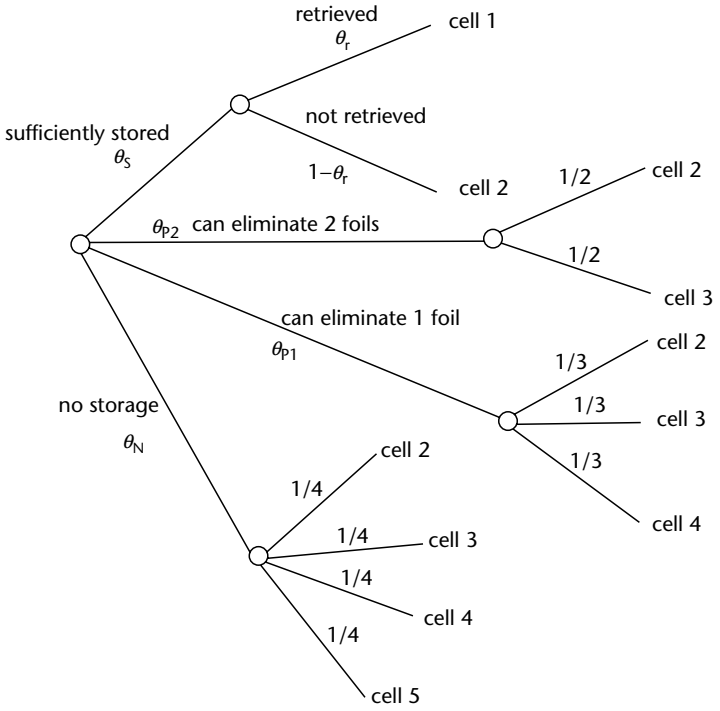


Figure 2. A multinomial process tree. The data representation is illustrated in (a) for the task of recall testing followed by forced choice recognition; (b) shows the latent process tree model for the task in terms of psychological parameters. Parameters are probabilities and are shown on branches of the tree.

remember. If a memory target is correctly recalled, then that is a response in cell 1 of Figure 2(a). After the recall test, the targets that were not recalled are further tested on a four-alternative forced-choice (4-afc) recognition test in which a single target is embedded in a string with three novel items. If a person is correct on the 4-afc task, that would be a response in cell 2 of Figure 2(a). If the person is incorrect, then the wrongly selected item is removed and the person must make a three-alternative forced-choice (3-afc) response. If the person is correct on the 3-afc task, then that is a response in cell 3. However, if the person is incorrect, then there is two-alternative forced-choice test for the last two items.

In Figure 2(b), the process tree for this task is shown. According to the model, the probability of correct free recall (cell 1) can occur only if the target is sufficiently stored with probability θ_S, and successfully retrieved with probability θ_r; hence the proportion of correct free recall is equal to $\theta_S\theta_r$. If the person has no representation of the target whatsoever, an outcome that occurs with probability θ_N, then the person will not be able to recall the target and will have a probability of 1/4 for responding in cells 2 through 5. In this case, where there is absolutely no target knowledge remaining in the memory system, the responses on the recognition tasks are at a pure guessing level. The other two states of insufficient storage shown in Figure 2(b) are related to two levels of partial storage of the target (i.e. being able to have enough knowledge of the target to eliminate two of the three novel items on the 4-afc task with probability θ_{P2} or being able

to eliminate only one of the novel items with probability θ_{P1}). The probability of partial target storage is $\theta_{P2} + \theta_{P1}$. If the target is sufficiently stored but not retrieved, then the person would have incorrect recall but would be correct on the 4-afc task (cell 2). Given data for the number of responses in the various cells of the multinomial, the underlying probabilities for storage, retrieval, partial storage, and no storage probabilities can be estimated, thus providing measures of the underlying psychological processes of interest.

With multinomial process tree models, the model is tightly linked to a specific experimental task. The mathematical psychologist usually must invent a task that provides a means to measure the psychological processes of interest. In the above model, the experimental task was novel and was created by the mathematical psychologist in order to obtain measures of complete as well as partial storage.

Information Processing and Reaction Time Models

Mathematical models have played a central role in the research in psychophysics, information processing, and cognition. Here the investigator is typically interested in explaining features of the dependent measures that are directly observable, such as the person's response time, percentage correct, or the tradeoff of time to complete a task with task accuracy. As with SDT and MPT models, these psychological models arise from a probabilistic framework.

One problem of particular interest is the differentiation between various information-processing architectures, for example, serial models versus parallel models. The differentiation between serial processing (processing one piece of information at a time in stages) and parallel processing (the concurrent processing of information) was a topic of particular interest in the experimental, memory search work by Saul Sternberg in the early 1960s. In experimental psychology, the signature for parallel processing was considered to be a flat reaction time function as task complexity was increased, whereas serial processing was expected to have an increasing time requirement for more complex tasks. The consistent experimental finding of a linearly increasing mean completion time for longer memory searches seemed to settle the matter that memory search was processed in a serial fashion. However, mathematical psychologists pointed out that this conclusion, based on mean response times, was premature. The problem was examined mathematically first in 1973 by James Townsend in a journal article and later in the 1983 book by James Townsend and Gregory Ashby. These analyses pointed out that a flat response time function would occur for parallel systems if the information-processing capacity (computations per unit of time) is unlimited. However, the more realistic assumption is that information-processing capacity is finite, and in this case the discrimination between serial and parallel architectures is more difficult because parallel processing models also require more time for tasks of greater complexity. Townsend and Ashby also showed that it is possible to differentiate between information-processing architectures with sufficiently rich response time data (i.e. examine many features of the response time distribution in addition to the mean task completion time).

Many psychological experiments require the subject to make a choice response; that is, the individual must decide if a stimulus is the same as or different from some reference standard. The statistical properties for the time to give a 'same' response is not equivalent to the time for a 'different' response. Mathematical psychologists have been interested in accounting for the entire response time distribution for each type of response. One successful model for response time was developed in 1975 by Steven Link and Richard Heath, and is called the relative-judgment, random-walk model. In general, random-walk models are characterized by a state variable (a real-value number that reflects the current state of evidence accumulation). The state variable is contained between two barriers. The state variable for the next time increment is the same as the old value plus or minus a random step value. Eventually the 'random walk' of the state variable terminates when the variable reaches either the upper or the lower barrier. The Link and Heath model specifies model parameters for the random walk, such as the location of the two barriers and the probability distribution for the incremental step size. The model results in a host of predictions for the same–different times.

Axiomatic Measurement Theories and Functional Equations

In contrast to the highly specific models that arise from a statistical framework (e.g. SDT and MPT), a great deal of research has also been conducted in mathematical psychology that deals with measurement at a very general level. This research focuses on the necessary and sufficient conditions for a

general type of measurement scale or a form of a functional relationship among variables. In this area of research, a set of general principles or axioms is considered, and the consequences of the assumed axioms are derived. If the theory is not supported, then at least one of the axioms is in error. One way to evaluate a theory is by testing the validity of the essential axioms.

An example of the axiomatic approach is the 1959 choice axiom formulated by Duncan Luce. To further clarify the axiom, let us consider a finite set of alternatives (e.g. potential candidates for a job). If there is a candidate who is never preferred over some another candidate, then that dominated candidate is removed from the set of alternatives because the remaining set of candidates includes the stronger alternative. This process is repeated until all dominated alternatives are removed. The remaining alternatives have pairwise choice probabilities P_{ij} which are neither 0 nor 1 (where P_{ij} denotes the probability that candidate i is preferred over candidate j). The key component of the choice axiom states that the probability that any candidate is preferred is statistically independent of the removal of one of the candidates. Luce shows that this axiom implies that there must exist nonnegative numbers v_i associated with each alternative, and that the choice probability is $v_i/(v_i + v_j)$. Other consequences of the choice axiom were developed by Luce.

Functional equation theory is another formal analysis method that is similar to the axiomatic measure approach. Functional equation theory is a branch of mathematics that dates back to d'Alembert, an eighteenth-century mathematician. In mathematical psychology, functional equation analysis has been utilized to determine, in a principled manner, the form of a mathematical relationship between critical variables. With this approach, theorems are developed that demonstrate that only one mathematical function satisfies a given set of requirements. One classic example of a functional equation analysis in mathematics is referred to as the Cauchy equation, and it is the proof that the only continuous function solution for the stipulation that $\mathrm{f}(x+y) = \mathrm{f}(x) + \mathrm{f}(y)$, where x and y are nonnegative, is $\mathrm{f}(x) = cx$, where c is a constant.

An excellent example in mathematical psychology of a functional equation analysis was provided by Luce. For this example, consider two different stimuli of intensities I and I', respectively, and let $\mathrm{f}(I)$ and $\mathrm{f}(I')$ be the corresponding psychological perception of the stimuli. From a functional equation analysis, if $I/I' = c$, where c is a constant, and if $\mathrm{f}(I)/\mathrm{f}(I') = g(c)$, then $\mathrm{f}(I)$ must be of the form of a power function, i.e. $\mathrm{f}(I) = AI^{\beta}$, where A and β are positive constants that are independent of I. With a functional equation analysis, the results are rationally established without the necessity of curve-fitting and statistical analysis.

Judgment and Decision-making Models

The topics of judgment and decision-making have been the focus of considerable research in mathematical psychology. Much of this work centered on the differences between normative models (i.e. rational models from a mathematical perspective) and descriptive models that characterize more accurately the behavior of individual decision-makers. For example, the perception of the psychological worth (or utility) of gambles is an area where there is a marked discrepancy between normative theory and experimental findings. Utility theory was formulated by the mathematician Daniel Bernoulli in 1738 as a solution to a gambling paradox. Prior to that time, the worth of a gamble was the expected value for the gamble; for example, given a gamble that has a 0.8 probability for a gain of \$10 and a probability of 0.2 for a loss of \$5, the expected value is $0.8(\$10) + (0.2)(-\$5) = \$7$. However, Bernoulli considered a complex gamble that was related to a bet-doubling system and demonstrated that the gamble had infinite expected value. Individuals did not perceive that gamble as having infinite value, and hence the paradox. To resolve this problem, Bernoulli replaced the monetary values with subjective worth numbers for the monetary outcomes – these numbers were called utility values.

In 1944, von Neuman and Morgenstern generated general axioms for an expected utility theory that became a theoretical cornerstone for economics. However, psychologists have provided numerous experimental demonstrations that this theory is not an accurate descriptive theory. In an effort to achieve more realistic theories for the perception of probability and the utility of gambles, mathematical psychologists and theoretical economists have formulated alternative models.

Models of Memory

The modeling of memory has been a vigorous area of research in mathematical psychology. These models typically have a number of parameters, often more parameters than are identifiable in a single condition of an experiment (i.e. given the data for a single condition, it might not be possible to estimate all the model parameters). However,

given values for the model parameters, it would be possible to account for the data obtained in that one condition as well as many other experimental conditions. In fact, a successful model is usually applied across a wide range of experiments and conditions without major changes in parameter values. In 1968, Richard Atkinson and Richard Shiffrin developed a model that explored the short-term retention of information by means of a multiple-store framework (a very short-term sensory register, a short-term store, and a permanent memory store). The model differentiated between hardware features of the memory system, such as the properties of the memory stores, and the strategies or control processes that a person could use with respect to the memory system. This model stimulated considerable research and is one of the best-known models in psychology.

Subsequent memory models have explored more fully the processes associated with recognition memory. One class of recognition memory models has come to be called 'global matching models'. Within this class of models, there are rather different conceptualizations as to the format of information in memory. Array models are a type of global matching model that treat the memory targets as separate *N*-dimensional vectors of attributes. A recognition memory probe activates each existing item in the memory system by an amount that is dependent on the similarity between the two items. For example, in Estes' 1994 array model for classification and recognition, there is a similarity function between any two *N*-dimensional memory vectors. The recognition decision is a function of the total similarity produced by the probe item with all the items in the memory system. In contradistinction to the array models, there are also distributed memory theories that utilize global matching. With distributed memory models, memory is composed of either a single matrix or a single vector. Consequently for the distributed models, there are no separate vectors for various items in memory. For example, in the TODAM (theory of distributed associative memory) model by Bennett Murdock, the totality of memory is contained in a single vector. Recognition is based on a vector function that depends only on the recognition probe and current state of the memory vector. As more items are added to distributed memory, previous item information may be lost.

Neural Network Models

Neural networks research is vast in scope and interest. Mathematical psychologists, such Stephen Grossberg, James McClelland, and David Rumelhart, have played a prominent role in this research area. Neural networks also have wide appeal outside psychology, and important contributions have been made by statisticians, engineers, and physicists.

In general, there is a distinction between real and artificial neural networks. By 'real' neural network is meant a model of brain structures such as the hippocampus, or the visual cortex. Real neural networks have a close linkage with research in neurology and physiological psychology.

Artificial neural networks are distributed computing systems that pass information and change the properties of links and nodes in the network. Artificial neural networks have proven to be valuable as models for learning and pattern recognition. There are many possible arrangements for artificial neural networks (e.g. the number of layers between the input of information and the output level, and the properties for revising linkage weights). A common feature of artificial neural networks is that the memory system is characterized by a large number of separate nodes that process information strictly in terms of information available to the node. These nodes are activated in parallel, but there can be multiple layers of nodes and inhibitory as well as excitatory activation.

Further Reading

Anderson JA (1995) *Practical Neural Modeling*. Cambridge, MA: MIT Press.

Batchelder WH and Riefer DM (1999) Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review* **6**: 57–86.

Chechile RA (1998) A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology* **42**: 432–471.

Estes WK (1994) *Classification and Cognition*. New York, NY: Oxford University Press.

Green DM and Swets JA (1966) *Signal Detection Theory and Psychophysics*. New York, NY: John Wiley.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: John Wiley.

Luce RD (1986) *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Luce RD (2000) *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. London, UK: Lawrence Erlbaum.

Luce RD, Bush RR and Galanter E (eds) (1963/1965) *Handbook of Mathematical Psychology*, vols I, II, and III. New York, NY: John Wiley.

Townsend JT and Ashby FG (1983) *Stochastic Modeling of Elementary Psychological Processes*. Cambridge, UK: Cambridge University Press.

Memory Distortions and Forgetting

Intermediate article

Asher Koriat, University of Haifa, Haifa, Israel
Morris Goldsmith, University of Haifa, Haifa, Israel
Ainat Pansky, University of Haifa, Haifa, Israel

CONTENTS

Forgetting refers to the failure to remember specific facts or events that took place in the past. In contrast, memory distortion includes many ways in which what a person does remember can deviate from what actually happened.

INTRODUCTION

Forgetting is perhaps the single most intrinsic aspect of the concept of memory; were there no forgetting, there would be little need for the concept of memory at all. For better or worse, however, forgetting is a ubiquitous phenomenon for all of us.

An influential memory metaphor that was proposed by the philosopher John Locke conceives memory as a storehouse: a place in which thoughts and experiences are initially stored and later retrieved. Such a conception implies a *quantity-oriented* approach to memory, in which the focus is on how much information is retained (remembered) and how much is lost (forgotten). This approach underlies much of the traditional experimental research on memory. More recently, however, a different conception has been gaining prominence, motivated primarily by real-life memory phenomena. In this *correspondence* conception, memory is treated as a representation or description of past events, and interest focuses on the extent to which that description faithfully portrays those events (see Koriat and Goldsmith, 1996a).

The contrast between the storehouse and correspondence conceptions of memory is useful for distinguishing two different notions of forgetting. The quantity-oriented storehouse view leads to a definition of forgetting in terms of omission, that is, the failure to remember specific facts or events. In contrast, the accuracy-oriented, correspondence conception leads to a focus on the many ways in which what a person does remember can deviate from what actually happened. For instance, we might 'remember' events that never occurred or distort those that did occur. In real-life situations, memory distortions are often more serious than omissions. For example, we would not expect an eyewitness to a crime to remember everything that happened at the time. We do, however, want to be able to depend on what he or she *does* report to be correct.

This article outlines and reviews some of what is known about the causes of omission errors, and about several different types of memory distortions.

FORGETTING

The first experimental investigation of forgetting was performed by Hermann Ebbinghaus in the late nineteenth century. Ebbinghaus studied lists of nonsense syllables until he achieved perfect recall, and then tested himself after different retention intervals. The resulting forgetting curves showed a great deal of forgetting within the first few hours after learning, quickly levelling off such that relatively little forgetting occurred thereafter. This basic pattern of decelerated forgetting over time has since been replicated repeatedly for various types of memory materials.

Ebbinghaus's approach implies a storehouse conception in which forgetting is defined as the loss of information over time. Such forgetting can derive from the spontaneous decay or weakening of memory traces, but it can also reflect the temporary inaccessibility of information that is otherwise available in memory. For example, experiments show that items that cannot be recalled at one point in time may be recalled (or recognized) on

subsequent memory tests, indicating that the memory traces of these items were not lost (this is often experienced by students who recall the 'forgotten' answers to exam questions just after leaving the exam room). Similarly, the 'tip-of-the-tongue' phenomenon, in which one feels that one knows the answer to a question (and really does) but is unable to retrieve it, is familiar to all of us. In fact, it is commonly held that the primary cause of forgetting is loss of access to stored information rather than loss of the information itself.

A major factor that can impair memory retrieval is interference. *Retroactive interference* occurs when newly acquired information interferes with the retrieval of previously learned information, whereas *proactive interference* occurs when previously learned information interferes with the acquisition and retrieval of new information. For example, remembering where I parked my car yesterday may impair my memory for where I parked it today (proactive interference) and vice versa (retroactive interference). Interference is especially likely to occur when the new and old pieces of information are similar.

People may also fail to retrieve a piece of information simply because the available retrieval cues are insufficient or ineffective. For example, we might fail to recollect the name of an acquaintance we are about to meet, but then her name may suddenly pop to mind when we see her face. Retrieval is especially likely to fail when retrieval cues do not match the way in which the information was initially encoded into memory, a principle known as 'encoding specificity' (Tulving and Thomson, 1973). For example, we may not recall who 'Debra Johnson' is, but then immediately remember her when a friend prompts us with: 'You know – Debbie!' Similarly, retrieval may also be impaired when the retrieval context differs from the encoding context. Thus, one's memory of an event may be enhanced by returning to the same place (external context) in which it occurred, or by re-experiencing the same mood or state of mind (internal context) that one was in at the time. State-dependent learning implies, for example, that if one were to study for an exam while drunk, it might actually be best to show up drunk to the exam as well!

In addition to the cognitive factors mentioned above, Sigmund Freud emphasized the importance of motivational factors that cause people to actively repress the memory of painful or traumatic personal events. Such repressed memories are held to remain active in the unconscious while being sealed off from consciousness. According to psychoanalytic theory, repressed memories cannot be wilfully retrieved, but they may emerge in the course of psychotherapeutic treatment (see below).

Finally, pathological memory disorders involving brain damage caused by injury or disease can obliterate the ability to recall portions of the person's autobiographical experience or to acquire new information. In *retrograde amnesia* a patient fails to recall past events, whereas in *anterograde amnesia* the patient has difficulty forming new memories.

MEMORY DISTORTION

Although a great deal of what we learn or experience is forgotten, it is perhaps more intriguing that what we do remember is not always veridical. Research on memory distortion and 'false memory' has important implications for real-life issues: for example, to what extent can we trust the memory of a courtroom witness? How reliable is the memory of a childhood traumatic event that is recovered years later in the course of psychotherapy? These questions concern the *accuracy* of what one remembers, rather than the amount.

Two basic principles can be used to explain many memory distortions and false memory phenomena:

1. What people remember depends not only on what actually happened, but also on constructive and reconstructive memory processes that people use to infer what might or should have happened.
2. What people remember depends on their ability to attribute remembered pieces of information to their proper source.

We now present phenomena and findings that illustrate these principles.

Memory Construction/Reconstruction Errors

Bartlett (1932) promoted the view of remembering as a dynamic, goal-directed 'effort after meaning'. Following his lead, a vast amount of research has shown that what is remembered is not simply a reproduction of the original input, but is an active construction or reconstruction based on inference and interpretation processes that are applied to that input – first when the information is initially encoded, and then again when the stored information is later retrieved.

These inference and interpretation processes are guided by one's general knowledge and expectations about the world. For example, one's cognitive

schema or *script* about what typically occurs in a restaurant may lead one to fill in and remember details that did not actually occur: that the host paid the bill, for instance, when in fact he walked out without paying. Such schema-based intrusions reflect a confusion between what one expects and what actually happened.

Reconstructive memory processes can also distort remembered details. Such distortions have been examined extensively in eyewitness research focusing on the effects of *leading questions* (Loftus, 1979). For example, the question 'how fast were the cars going when they *smashed into* each other?' was found to yield significantly higher speed estimates than more neutrally phrased questions, such as 'how fast were the cars going when they *hit* each other?' Apparently, people's memories can be contaminated by implications conveyed in a question's wording. Moreover, when questioned again a week later, witnesses who had previously been asked the leading question were more likely to falsely remember the presence of broken glass than witnesses who had received the more neutral question. In fact, simply using a definite article (e.g. 'Did you see *the* broken headlight?') rather than an indefinite article ('Did you see *a* broken headlight?') can bias witnesses into falsely remembering the specified object or event.

Reconstructive bias need not be externally induced, however. For example, one's current knowledge and beliefs about oneself (self-schemas) can distort one's memory for past beliefs, attitudes, and behaviors, often causing one to remember them as being more compatible with one's current self than they really are.

Finally, another type of reconstructive error derives from people's tendency to remember the general meaning or gist of experienced events rather than their exact details. Consequently, people often report information that is consistent with the gist of an event, though it may be inconsistent with its details. Even when relatively sterile word-list study materials are used, gist-based errors can appear, such as remembering 'canary' or 'bird' when 'sparrow' was actually studied.

Source/Reality Monitoring Errors

Many memory errors stem from a failure to identify correctly the source of retrieved information. For example, we may remember having called the doctor to cancel an appointment, but in fact we only thought about doing so. *Reality monitoring* – the ability to distinguish actual events from fantasy – is a special case of *source monitoring*: the ability to attribute experiences to their proper source (Johnson *et al*., 1993). Source-monitoring errors can result in confusions between details of events that were experienced in one situation and those that took place in another. A dramatic example is an incident that ironically involved a well-known memory researcher, Donald Thomson, who was wrongly identified by a rape victim as the rapist. Thomson's alibi both exonerated him immediately and helped explain the false accusation. He was giving a live television interview at the time of the rape: apparently, the victim had been watching the interview just before she was raped, and confused the memory of his image with that of the rapist. Thus, source-monitoring failures can often be more harmful than retrieval failures: fragments of real experience are accurately and vividly recalled, but are attributed to the wrong person, location, or time, resulting in false memory.

Source-monitoring errors may explain many false-memory phenomena. A prominent example again comes from eyewitness research. Studies indicate that wrong information presented to witnesses after the witnessed event (e.g. a statement or question that erroneously refers to an actual stop sign as a yield sign) can distort their subsequent memory for that event (e.g. remembering having *seen* a yield sign). This phenomenon may derive from deficient source monitoring: the post-event misinformation is more accessible than the original information and is wrongly attributed to the original event.

A second example comes from an experimental paradigm that has attracted much interest recently (Roediger and McDermott, 1995). In this paradigm, subjects study a list of related words (e.g. BED, REST, TIRED, DREAM), all converging on a particular 'lure' word (e.g. SLEEP) that is *not* presented for study. When tested later, the subjects tend to falsely recall or recognize the lure word. Interestingly, subjects are generally quite confident about these false memories, sometimes even claiming to remember the tone of voice in which the (nonpresented) word was spoken! Such errors may result from an incorrect inference regarding why the critical lure feels 'activated' or familiar – with the person misattributing the feeling to memory.

There has been a heated debate over the authenticity of memories of childhood sexual abuse that are recovered in adulthood (often through psychotherapy). The question is whether such recovered memories are accurate recollections that were repressed for years due to their traumatic nature, or were false memories induced during the process of therapy (by repeated imagination or compliance

with the therapist's suggestions). Can people 'remember' entire events that did not occur? Studies indicate that indeed memory for false events can be implanted: subjects who are urged to repeatedly imagine fictional childhood events subsequently tend to remember those events as real, and even provide additional details about them. Thus, people sometimes attribute to reality an episode that was only suggested to them, or only imagined by them, demonstrating extreme cases of faulty reality monitoring.

METACOGNITIVE PROCESSES AND THE PREVENTION OF MEMORY ERRORS

The preceding sections have indicated a variety of ways in which memory can go wrong. When a person has remembered and reported incorrect information, this implies not only a failure of memory, but also of *metamemory* processes; that is, a failure to realize that the remembered information is faulty. Conversely, sometimes information is omitted or 'forgotten' not because the information fails to come to mind, but because the person does not realize that the retrieved information is in fact correct. In this context, metamemory refers to what one knows about one's own memories, and how that knowledge is used to regulate what one reports.

To illustrate, consider a courtroom witness who is sworn to tell 'the whole truth and nothing but the truth'. To fulfill that goal, the witness must try to distinguish between correct and incorrect information that comes to mind, and report only (and all of) the correct information. The attempt to regulate one's memory reporting in order to provide as much information as possible but to avoid reporting wrong information seems to be an intrinsic aspect of remembering in real-life situations. Two types of strategic control over memory reporting have been examined (see Goldsmith *et al*., 2002; Koriat and Goldsmith, 1996b). The first, *report option*, involves the decision whether to report a remembered piece of information or to withhold it (e.g. to reply 'I don't know'). People tend to avoid reporting information that they feel unsure about, which generally enhances the accuracy of what they report, but may reduce the amount of correct information (i.e. increase omission errors) if people mistakenly screen out correct answers. Importantly, both the accuracy benefits and the quantity costs that ensue from the option of free report depend on two metacognitive factors: (a) monitoring effectiveness – people's ability to monitor the correctness of the information that comes to mind, and (b) control policy – the strictness or liberality of the confidence criterion that is set for volunteering answers.

Many of the cues that people use to monitor their memories have to do with source and reality monitoring (Johnson *et al*., 1993). Memories of witnessed events tend to be more vivid and include more perceptual detail than imagined events. Thus, people may utilize a *distinctiveness heuristic* to screen out false memories (Schacter *et al*., 1999), based on the awareness that the memory of true events should include recollection of distinctive details. Also, when the demands for accuracy are strong, the person may deliberately recruit additional corroborative information that helps verify the source of the retrieved events, or adopt a relatively strict criterion for reporting the information.

A second way in which rememberers regulate the amount and accuracy of the information that they report is by controlling the *grain size* of their answers, choosing a level of precision or coarseness at which they are unlikely to be wrong. Instead of reporting that the accident occurred precisely at 5:21 p.m. (which is likely to be wrong), one may choose to report that it occurred between 5:00 and 5:30, or even 'sometime in the late afternoon' (both of which are more likely to be correct). Of course, coarsely grained answers, though more likely to be correct, generally provide less information than more precise answers. Here too, rememberers tend to utilize their monitoring and control processes in a strategic manner, choosing a grain size that represents an expedient compromise between accuracy and informativeness (Goldsmith *et al*., 2002).

CONCLUDING REMARKS: HOW FAULTY IS MEMORY?

The focus on forgetting and memory distortion in this article could leave a pessimistic impression about the general faithfulness of human memory. But is human memory really as flawed as it seems? We think not. First, although some types of memory errors may appear to reflect flaws in the 'system design', they are in fact by-products of otherwise adaptive features of memory. Thus, for example, remembering the gist but forgetting the details of stories and events, or inferring information not actually present in the input, is often what is required in real-life situations. Second, when detrimental memory errors do occur, they appear to derive from the same memory processes that normally lead to accurate remembering. Thus,

although schema-based inferences are sometimes wrong, they are probably more often right – assuming that events that take place in the world ordinarily do agree with our general knowledge and expectations. Although perhaps it is natural for memory researchers to focus on the 'dark' side of memory in attempting to understand the causes of forgetting and distortion, it is amazing how much information people actually do remember, the vast majority of which is correct – or at least useful. Which brings us to one final point. Some of the functions of memory are expressed neither in its quantity nor in its accuracy, but rather, in its personal and social *utility* (Neisser and Winograd, 1988). Thus, for instance, our memories are important vehicles for preserving a sense of self, and in facilitating our interactions with others (e.g. storytelling and reminiscing). Such goals may be achieved despite (and perhaps because of) a certain amount of forgetting and distortion. Current work has begun to address these broader functions of memory, and how they are realized.

References

Bartlett FC (1932) *Remembering: A Study in Experimental and Social Psychology*. New York, NY: Cambridge University Press.

Goldsmith M, Koriat A and Weinberg-Eliezer A (2002) Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General* **131**(1): 73–95.

Johnson MK, Hashtroudi S and Lindsay DS (1993) Source monitoring. *Psychological Bulletin* **114**: 3–28.

Koriat A and Goldsmith M (1996a) Memory metaphors and the real-life/laboratory controversy: correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences* **19**: 167–228.

Koriat A and Goldsmith M (1996b) Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review* **103**: 490–517.

Loftus EF (1979) *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Neisser U and Winograd E (1988) *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*. New York, NY: Cambridge University Press.

Roediger HL and McDermott KB (1995) Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–814.

Schacter DL, Israel L and Racine C (1999) Suppressing false recognition in younger and older adults: the distinctiveness heuristic. *Journal of Memory and Language* **40**: 1–24.

Tulving E and Thomson DM (1973) Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80**: 359–380.

Further Reading

Baddeley AD (1990) *Human Memory: Theory and Practice*. Needham Heights, MA: Allyn & Bacon.

Goldsmith M and Koriat A (1999) The strategic regulation of memory reporting: mechanisms and performance consequences. In: Gopher D and Koriat A (eds) *Cognitive Regulation of Performance: Interaction of Theory and Application. Attention and Performance XVII*, pp. 373–400. Cambridge, MA: MIT Press.

Koriat A, Goldsmith M and Pansky A (2000) Toward a psychology of memory accuracy. *Annual Review of Psychology* **51**: 481–537.

Loftus EF (1992) When a lie becomes memory's truth: memory distortion after exposure to misinformation. *Current Directions in Psychological Science* **1**: 121–123.

Loftus EF (1997) Creating false memories. *Scientific American* **277**: 50–55.

Mitchell KJ and Johnson MK (2000) Source monitoring: attributing mental experiences. In: Tulving E and Craik FIM (eds) *The Oxford Handbook of Memory*, pp. 179–195. New York, NY: Oxford University Press.

Roediger HL III (1996) Memory illusions. *Journal of Memory and Language* **35**: 76–100.

Schacter DL (1999) The seven sins of memory. *American Psychologist* **54**: 182–203.

Schacter DL, Norman KA and Koutstaal W (1998) The cognitive neuroscience of constructive memory. *Annual Review of Psychology* **49**: 289–318.

Tulving E and Craik FIM (eds) (2000) *The Oxford Handbook of Memory*. New York, NY: Oxford University Press.

Memory: Implicit versus Explicit

Intermediate article

Neil W Mulligan, Southern Methodist University, Dallas, Texas, USA

CONTENTS

Human memory is expressed via conscious recollection (explicit memory) as well as through unconscious changes in behavior (implicit memory). Numerous dissociations indicate that these are distinct forms of memory mediated by different parts of the brain.

INTRODUCTION

We typically use the term 'memory' to refer to our ability to consciously and intentionally recollect past experience. We speak of remembering what we did last weekend or of trying to recall where we left our keys. In his landmark book, *The Principles of Psychology*, William James's definition of memory similarly emphasized conscious recollection:

> Memory… is the knowledge of a former state of mind after it has already once dropped from consciousness; or rather *it is the knowledge of an event*, or fact, of which meantime we have not been thinking, *with the additional consciousness that we have thought or experienced it before*. [italics in original] (James, 1890, p. 648)

However, researchers have also long supposed that memory for prior events can affect behavior even when people are not trying to remember and, indeed, when people are not aware that memory for prior events is operative (James, in his chapter on habit, was not insensitive to this point). Just such a consideration led Ebbinghaus (1885/1964) to develop the savings measure of memory. Ebbinghaus taught himself lists of nonsense syllables (e.g. ZUD) and later attempted to re-learn the same lists, measuring how much more quickly these lists could be learned the second time. This savings measure was designed to measure all memorial influences, not just those that give rise to conscious recollection. In modern psychology, a confluence of results from cognitive psychology, neuropsychology, and cognitive neuroscience has focused attention onto these divergent manifestations of memory, embodied in the distinction between *explicit* and *implicit* memory. Explicit memory refers to intentional or conscious recollection of prior experiences. Implicit memory, in contrast, refers to changes in behavior that are produced by prior experience and are unaccompanied by intentional or conscious recollection. Consequently, implicit memory is frequently described as unintentional or unconscious memory.

Research on anterograde amnesia provided a major impetus for modern interest in implicit memory. Anterograde amnesia was traditionally defined as an inability to retain new experiences coupled with preserved perceptual and intellectual abilities. Having been presented with a new piece of information, amnesics quickly lose the ability to recall it; that is, amnesics have a deficit in explicit memory. However, research in the 1970s and 1980s showed that amnesics are influenced byprior experience when memory is tested implicitly.

An early and very influential demonstration was provided by Warrington and Weiskrantz (1970). In this study, amnesics and normal control subjects were presented with a list of words. Their memory for the words was then tested explicitly, by asking them to recall the words, or implicitly, by asking them to identify fragmented words or to complete word stems (e.g. MET ___) (in which case some of the fragments or stems corresponded to studied words, such as METAL). In the latter tests, subjects were not asked to remember the words, yet the influence of the studied list can be observed by the increased likelihood of responding with studied words. The amnesics recalled many fewer words than the normal control subjects (consistent with the defining definition of the disorder). However, on the fragment and stem completion tests, amnesics demonstrated the same amount of improvement in performance as the normal control subjects. Thus, although deficient in conscious recollection of the past, the amnesics demonstrated an equivalent unconscious influence of past experience.

THE EXPERIMENTAL STUDY OF EXPLICIT AND IMPLICIT MEMORY

The Warrington and Weiskrantz (1970) experiment illustrates the standard approach for studying explicit and implicit memory. The typical memory experiment consists of two parts, a study episode followed by a memory test. In the study episode, the experimental subject is presented with some type of controlled experience, such as a series of words or pictures presented on a computer screen. The subsequent memory test measures how much of the studied information is retained. The memory test is either an explicit test, in which the subject is asked to think back and try to retrieve information about the study episode, or an implicit test, in which the subject is asked to perform a task that is nominally unrelated to the study episode. The operative distinction between explicit and implicit memory tests lies in the task instructions.

Examples of explicit memory tests include free recall, in which the subject is simply asked to report as much as possible about prior events ('Try to recall as many words from the study episode as possible'); cued recall, in which the subject is presented with cues to aid in recall ('Try to recall the name of fruits that were presented during the study episode' after having seen the word 'orange' during the study session); and recognition memory tests, in which the subject is asked to discriminate between previously experienced information and new information ('Which of these words were presented earlier?').

Examples of implicit memory tests include word-stem and word-fragment completion. In these tasks, the subject is presented with a series of word stems (e.g. ora___) or fragments (e.g. '_r_ng_') and asked to complete each with the first appropriate word that comes to mind (e.g. 'orange'). Some of the word stems/fragments correspond to words presented during the study episode and some do not, although the subject is not informed of this relationship. Memory for the study episode is inferred from enhanced performance for the studied words relative to a control set of unstudied words, a measure called *priming*.

DISSOCIATIONS OF EXPLICIT AND IMPLICIT MEMORY

Population Dissociations

The principles that govern performance on explicit and implicit memory tests differ in many ways. Warrington and Weiskrantz (1970) provided a striking example: compared to normal healthy adults, patients with anterograde amnesia are profoundly impaired on tests of explicit memory but show normal or near-normal levels of retention on implicit tests. In this example, the variable of neurological population has different effects on explicit and implicit memory tests, a pattern of results known as a *dissociation*. This dissociation has been observed many times for a number of different explicit and implicit memory tests, implying that conscious, recollective aspects of memory are impaired in organic amnesia, whereas unintentional, unconscious aspects of memory are unaffected by this disorder.

A number of other population dissociations have been reported. When healthy older and younger adults are compared, older adults generally produce worse memory on explicit tests such as recall and recognition. However, older adults often produce the same amount of priming on implicit tests such as word-stem and word-fragment completion. Likewise, patients with depression or schizophrenia have memory deficits in explicit remembering (as compared to healthy control subjects) but produce intact priming. In all of these cases, a population with deficient explicit memory produces normal or near-normal implicit memory.

Functional Dissociations

Experimental manipulations also produce dissociations of explicit and implicit memory tests, providing converging evidence for the separability of these forms of memory and providing insight into functional differences between explicit and implicit memory.

For example, manipulations of attention can produce divergent effects on explicit and implicit tests. In studies of attention and memory, some subjects are presented with a study list under full attention conditions, in which their sole task is to encode the stimuli. Other subjects are presented with the same study list but at the same time must carry out a secondary task, such as listening to a sequence of tones, and categorizing each as high, middle, or low. The secondary task is designed to distract subjects from the study list, reducing their attention to the study stimuli. Subjects in the divided attention condition typically recall or recognize fewer of the study items than subjects in the full attention condition. However, if memory is tested with implicit tests, such as word-fragment completion or perceptual identification, the two groups produce comparable levels of priming (Mulligan, 1998). Thus, dividing attention at encoding dissociates

performance on explicit and implicit tests, impacting performance on the former but not the latter. A similar dissociation is produced by varying the amount of semantic processing during the study episode, which has a marked effect on explicit memory tests (the levels-of-processing effect) but has little or no effect on such implicit memory tests as word-stem completion and perceptual identification.

Conversely, the similarity between the physical (or perceptual) features of the stimuli as presented at the time of study and test has a strong effect on many implicit memory tests but little or no effect on many explicit tests. An example is the effect of study modality. If some of the words on a study list are presented visually and some are presented aurally, later explicit memory for the words is typically unaffected. However, when memory is tested with the implicit tests of word-stem and word-fragment completion, study modality has a large impact; visually presented words lead to more priming than the aurally presented words.

Thus, some variables (divided attention, levels of processing) produce effects on explicit but not implicit tests, whereas other variables (e.g. study modality) produce the opposite pattern, affecting implicit but not explicit tests. Joining these is the read/generate manipulation which can produce directly opposite effects on explicit and implicit memory. This was initially demonstrated by Jacoby (1983) who had subjects either read individual words (e.g. 'cold') out of context ('xxx – cold'), or read words in a meaningful context ('hot – cold'), or generate words from context ('hot – ????'). Subjects' memory for the words was then tested with either an explicit (recognition memory) test or an implicit test (perceptual identification). For recognition, the generated words produced the greatest accuracy followed by words read in context and then words read out of context; this is the traditional finding that generating words leads to superior explicit memory (the generation effect). On the perceptual identification task, the opposite results obtained: reading words out of context produced the most priming, followed by reading words in context. The generate condition produced the least priming.

Pharmacological Dissociations

The benzodiazepines (including alprazolam, triazolam, diazepam, and midazolam) are a class of drugs used in anesthesia and to treat insomnia and anxiety disorders. These drugs also produce a powerful, but temporary and reversible, form of anterograde amnesia, in which information encountered after administration of the drug is poorly remembered. As the drug dosage wears off, so do its amnesic effects. Recent research indicates that these effects are restricted to explicit memory.

In a typical study, one group of subjects was administered a dose of a benzodiazepine (sufficient to affect memory but insufficient to induce stupor or sleep) and a second group was given a placebo. After an absorption period (allowing the drug to take effect in the drug group), both groups were presented with study materials. Following a retention interval sufficient to allow the effects of the drug to wash out, a memory test was administered. Compared to the placebo group, the group given benzodiazepine produced poor explicit memory for the study materials on tests such as free recall (thus evincing amnesia for information encountered while under the effects of the drug). However, if the subjects were presented with an implicit test, such as word-stem or word-fragment completion, then the drug and placebo group produced equivalent levels of priming.

Thus, pharmacological amnesia produces the same type of dissociation as produced by organic amnesia: it affects conscious recollection but appears to have no effect on unconscious influences of memory (see Curran, 2000, for a review).

THEORETICAL FRAMEWORKS

Activation View

An early account attributed implicit memory phenomena to the temporary activation of pre-existing knowledge representations. This view proposed that the initial processing of a stimulus automatically activated representations in long-term memory. Residual activation of these representations increased the probability and decreased the amount of time necessary for these representations to reach a threshold level of activation, at which time the information in the representations would be available in working memory (such as to support responses, decision-making, etc.). Thus, the speed and accuracy of processing would be enhanced for items that had recently been perceived compared with those that had not.

Although initially a viable account, several lines of evidence are contrary to this view. First, the activation view has difficulty accounting for the longevity of priming effects, which can last over periods of days, weeks, or even months. Such a time frame is inconsistent with standard notions of activation which assume that activation returns

to baseline within a period of minutes or hours, not days. Second, priming effects have been obtained for novel stimuli, such as nonsense words and line-drawings of nonsense objects. This is also contrary to the activation view, which predicts that only those stimuli with pre-existing representations (familiar, well-integrated stimuli) should exhibit priming. Third, the perceptual specificity of many priming phenomena (e.g. effects of study modality) appears incompatible with the activation view because this attributes priming to abstract, amodal representations of words and concepts.

Multiple Memory Systems

Neuropsychological and neuroscientific analyses of memory have often emphasized neurally distinct and functionally dissociable memory systems. A familiar distinction is that between short-term (or working) memory and long-term memory, supported by findings that patients with anterograde amnesia exhibit their deficit on delayed (explicit) memory tests but not on immediate memory tests such as digit span. The finding that amnesia dissociates performance on explicit and implicit memory has likewise been interpreted as reflecting the operation of distinct memory systems, in this case extending the fractionation of memory to long-term memory systems. Because anterograde amnesia is typically associated with damage to the hippocampus and medial temporal lobes, it is believed that this part of the brain is necessary for acquiring new experiences in ways that later produce conscious recollection.

Initially, multiple-memory system theories proposed dichotomous memory systems (e.g. declarative versus procedural, episodic versus semantic) to account for explicit and implicit memory, respectively. However, as evidence for dissociable forms of implicit memory mounted, the number of proposed long-term memory systems has increased. The dominant theoretical framework proposes four: episodic memory, semantic memory, the perceptual representation system (PRS), and procedural memory (Schacter *et al*., 2000). Episodic memory stores information about episodes in our personal past, enabling the experience of recollection. Semantic memory stores general knowledge about the world, including facts, conceptual information, and vocabulary. These two systems are sometimes jointly referred to as declarative memory, referring to knowledge that can be verbalized ('knowing that'). The PRS is a perceptual memory system that processes information about the form and structure of words and objects prior to the analysis of their semantic content. Finally, procedural memory represents knowledge of cognitive and motor skills ('knowing how'). Within this framework, explicit memory is assumed to be a product of the episodic memory system, whereas various forms of implicit memory are produced by the other systems.

Under this view, dissociations of implicit and explicit memory tests reflect the operation of distinct memory systems. Population and pharmacological dissociations provide critical support for the existence of multiple memory systems in the brain. For example, anterograde amnesia produces deficits on explicit memory, implying disruption to the episodic memory system, but typically not on verbal implicit memory tests (such as word-stem and word-fragment completion) or on tests of skill-learning, implying that the other systems are not disrupted. Perhaps the strongest support for the multiple-memory systems view is garnered by reports of double dissociations, dissociations in which two different patient groups (with damage to different parts of the brain) exhibit complementary dissociative patterns on implicit and explicit memory tests. For example, amnesic patients (with damage to medial temporal regions of the brain) have disrupted memory on explicit tests but not on implicit tests, as noted. The opposite dissociation occurs in patients with occipital-lobe lesions, who exhibit preserved explicit memory coupled with deficits in implicit memory for visual–perceptual information (Gabrieli *et al*., 1995). This provides strong support for the view that brain systems mediating performance on these two types of tests differ. Similar dissociative patterns support distinctions among the other imputed memory systems (Schacter *et al*., 2000).

Transfer-appropriate Processing

Although population and pharmacological dissociations strongly argue for the operation of multiple memory systems, patterns of functional dissociations have often been interpreted in terms of disparate processing requirements of different memory tests within a single memory system.

The dominant such account is the transfer-appropriate processing (TAP) framework (Jacoby, 1983; Roediger and McDermott, 1993). The TAP framework assumes that performance on a memory test benefits to the extent that cognitive processes carried out during initial learning are re-engaged at test. The TAP framework distinguishes between cognitive processes that are involved in the analysis of meaning (conceptual

processing) and processes that are involved in the analysis of perceptual or surface-level features (perceptual processing). The TAP view assumes that explicit and implicit memory tests often rely differentially on conceptual and perceptual processing, and consequently benefit from different types of initial learning. Many implicit memory tests require identification of degraded or ambiguous perceptual stimuli, such as word fragments, briefly presented words, or fragmented pictures. Under the TAP view, such tests rely heavily on perceptual processing and reflect the similarity between perceptual processes engaged at study and test. On the other hand, explicit memory instructions encourage the re-engagement of conceptual memory processes.

As an illustration, consider the results of Jacoby's (1983) read/generate experiment, reviewed earlier. One can characterize the encoding conditions as varying in the amount of perceptual or conceptual analysis required. Generating a word from a meaningful context requires conceptual processing, but provides no visual information for perceptual processing. Reading a word out of context requires perceptual processing of the word but presumably little conceptual processing. Reading a word in context occupies a middle ground in which both perceptual and conceptual processing contribute. According to the TAP view, encoding conditions that require more perceptual analysis should transfer well to implicit memory tests focusing on perceptual analysis, whereas encoding conditions encouraging conceptual processing should transfer well to explicit tests owing to their reliance on conceptual retrieval processes. The observed results are consistent with these expectations.

The distinction between perceptual and conceptual processing is not coextensive with the distinction between explicit and implicit test instructions. According to the TAP framework, dissociations of explicit and implicit tests are only indirectly attributable to test instructions; it is the memory cues and response requirements in conjunction with the test instructions which determine whether a test primarily re-engages perceptual or conceptual processing. Specifically, if the memory cues are conceptually related to the to-be-retrieved information, then conceptual processes are evoked. If the relationship between cues and responses is perceptual, perceptual processes are evoked. This theoretical view suggests that, with the appropriate choice of memory cues and response requirements, one can develop implicit memory tests which re-engage conceptual processes as well as explicit memory tests which re-engage perceptual processes. For example, consider the category production task. In this task, subjects are presented with the names of taxonomic categories and are asked to produce examples from each category. Examples from some of the categories are presented during the study episode and are consequently produced more often than examples which had not been presented previously, despite the fact that the subject is not informed of the relationship between the test and the study episode. Because the memory cues (the category names) bear a conceptual rather than a perceptual relationship to the target items (the category examples), this task would be an example of a conceptual implicit memory test.

The TAP framework predicts that dissociations are likely to occur between two implicit tests if one is perceptual and the other conceptual, and a number of such dissociations have been reported. For example, study modality affects perceptual implicit tests such as word-stem and word-fragment completion (as noted earlier) but not the conceptual implicit test of category production. Alternatively, levels of processing and divided attention have minimal impact on perceptual tasks such as word-fragment completion and perceptual identification, but produce substantial effects on conceptual priming tasks. Finally, the read/generate manipulation likewise dissociates perceptual and conceptual priming tasks; reading produces more perceptual priming but generating produces more conceptual priming.

The TAP framework enjoys great success accounting for functional dissociations between explicit and implicit tests, and among implicit tests of different types (perceptual and conceptual). Thus, the perceptual–conceptual processing dimension is an important element in accounting for implicit memory phenomena. However, it is unlikely to provide a complete explanation because, as noted by the progenitors of the TAP account, this view does not readily account for the population and pharmacological dissociations so important to the multiple-systems view. In particular, it has been generally found that explicit tests, whether relying on conceptual or perceptual retrieval cues, are affected by anterograde amnesia and drug treatments like the benzodiazepines. In addition, both organic amnesia and drug (benzodiazepine)-induced amnesia produce normal levels of priming on conceptual as well as perceptual implicit tests. Likewise, older compared to younger adults produce generally worse performance on explicit tests coupled with equivalent levels of conceptual and perceptual priming. These dissociations resist explanation in terms of perceptual versus conceptual

processing because, in these cases, perceptual and conceptual tests produce similar rather than different outcomes.

Component-processes Approach and Evidence from Neuroimaging

The complementary successes of multiple memory systems in accounting for population and pharmacological dissociations, and the TAP approach in accounting for functional dissociations, has produced the current view, which emphasizes multiple forms of priming and attempts to articulate the component processes that mediate performance on various memory tasks.

Recent neuroimaging research, using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), has supplemented traditional approaches in pursuing this goal. Neuroimaging techniques are used to elucidate the neural regions involved in explicit and implicit memory tests in normal control subjects. One interesting result is that there is substantial overlap in the active brain regions during explicit and implicit retrieval tasks. Despite this overlap, there are certain hallmarks of explicit and implicit memory.

Studies of explicit retrieval have consistently shown increased activity in anterior frontal (especially right prefrontal) lobe as well as activity in the medial–temporal regions, including the hippocampus. A common interpretation is that activity in the frontal lobe reflects a retrieval mode in which the individual is oriented towards the past and is intentionally trying to retrieve information. The medial–temporal activity is interpreted as reflecting the recollective experience itself, when the memory is successfully retrieved (Nyberg and Cabeza, 2000).

Priming on implicit tests is associated with decreased activity in various brain areas. This is believed to reflect a reduction in processing demands when a stimulus is processed a second time, which in turn produces increased speed and accuracy on the priming task. The brain regions involved depend on whether the implicit test is perceptual or conceptual, consistent with the TAP view. For perceptual tests such as word-stem or word-fragment completion, the decreased processing is found in visual cortex (in the posterior occipital lobe). In contrast, when items are reprocessed on conceptual implicit tests, decreased activity is found in the inferior frontal lobe and mid-temporal lobe. In general, these results imply that the same neural substrates responsible for the initial (perceptual or conceptual) processing are re-engaged at the time of test and exhibit the effects of the initial processing by their subsequent reduced activity.

References

Curran HV (2000) Psychopharmacological perspectives on memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook on Memory*, pp. 539–554. New York, NY: Oxford University Press.

Ebbinghaus H (1964/1885) *Memory: A Contribution to Experimental Psychology*. New York, NY: Dover.

Gabrieli JDE, Fleishman DA, Keane MM, Reminger SL and Morrell F (1995) Double dissociation between memory systems underlying explicit and implicit memory in the human brain. *Psychological Science* **6**: 76–82.

Jacoby LL (1983) Remembering the data: analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior* **22**: 485–508.

James W (1890) *The Principles of Psychology*. London, UK: Macmillan.

Mulligan NW (1998) The role of attention during encoding on implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 27–47.

Nyberg L and Cabeza R (2000) Brain imaging of memory. In: Tulving E and Craik FIM (eds) *The Oxford Handbook on Memory*, pp. 501–519. New York, NY: Oxford University Press.

Roediger HL and McDermott KB (1993) Implicit memory in normal human subjects. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 8, pp. 63–131. Amsterdam, Netherlands: Elsevier.

Schacter DL, Wagner AD and Buckner RL (2000) Memory systems of 1999. In: Tulving E and Craik FIM (eds) *The Oxford Handbook on Memory*. New York, NY: Oxford University Press.

Warrington EK and Weiskrantz L (1970) Amnesic syndrome: consolidation or retrieval. *Nature* **217**: 972–974.

Further Reading

Bowers JS and Marsolek CJ (eds) (in press) *Rethinking Implicit Memory*. Oxford, UK: Oxford University Press.

Foster JK and Jelicic M (eds) (1999) *Memory: Systems, Process, or Function*. New York, NY: Oxford University Press.

Graf P and Masson MEJ (1993) *Implicit Memory*. Hillside, NJ: Lawrence Erlbaum.

Kelley CM and Lindsay DS (1996) Conscious and unconscious forms of memory. In: Bjork EL and Bjork RA (eds) *Memory: Handbook of Perception and Cognition*. San Diego, CA: Academic Press.

Kihlstrom JF (1987) The cognitive unconscious. *Science* **237**: 1445–1452.

Moscovitch M, Vriezen E and Goshen-Gottstein Y (1993) Implicit memory in patients with focal lesions and degenerative brain disorders. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 8, pp. 133–173. Amsterdam, Netherlands: Elsevier.

Schacter DL and Badgaiyan RD (2001) Neuroimaging of priming: new perspectives on implicit and explicit memory. *Current Directions in Psychological Science* **10**: 1–4.

Memory Mnemonics

Intermediate article

John Best, Eastern Illinois University, Charleston, Illinois, USA

CONTENTS

Mnemonics are techniques that are used to improve both the amount of information that a person might retrieve from memory, and its accuracy. These techniques typically use highly overlearned organizational schemes and visual imagery.

INTRODUCTION

The term 'mnemonics' refers to a class of strategies for organizing material that a person believes he or she will want to retrieve. Thus, they have features in common with strategies used in other domains: their use is always conscious, deliberate, and effortful. The proponents of mnemonics believe, however, that the effort involved in using a mnemonic technique is amply rewarded by ease and accuracy at retrieval time.

Generally speaking, the effectiveness of most mnemonic techniques results from their exploitation of some other characteristic of the human cognitive system. For example, humans have long known they might improve their retrieval by incorporating some type of sensory information into the memory they wish to retrieve. Thus, when we use jingles and nursery rhymes to help us retrieve otherwise abstract things, we are exploiting the fact that the acoustic or phonetic properties of a stimulus can be used to help organize it. Similarly, the power of human imagination, with its ability to create unusual and hard-to-forget visual images, is even more likely to play a role in many formal mnemonic techniques.

Despite some impressive findings, the overall utility of mnemonic techniques has not yet been conclusively demonstrated. Similarly, it is not currently clear whether or not the astonishing feats of memory produced by modern mnemonists have any real implications for existing theories of 'normal' memory.

COMMON MNEMONIC METHODS

Loci

Loci is a Latin term pronounced 'low-sigh' and meaning 'locations' (singular: locus). The 'method of loci' thus refers to the notion of using a sequence of highly overlearned and easily visualized locations as a system for organizing memory stimuli. It is a technique whose origins lie in antiquity: the Greek bards of the fifth century BC used it to organize recitations of many hours' length.

To use this technique, you must know, and be able to visualize, a set of places in the exact order in which you might encounter them. For example, you might visualize the buildings you would encounter on your campus if you were to take a walk from your residence hall to the academic building where your lectures are held. Or, you might visualize the sequence of shops you would encounter at a nearby mall as you walked along one particular corridor. The important elements are that there is only one sequence of events, and that you are positive in your knowledge of that sequence.

To begin the method of loci, you construct a composite image consisting of the first stimulus you want to remember mentally placed in the first location of the sequence. Then you continue in the same fashion until all the elements on the list of stimuli to be learned have been imagined at their respective locations. To retrieve the elements, you 'take a mental walk' through the set of landmarks, mentally scanning the image you created for each landmark, and, if all goes well, retrieving the stimulus that you mentally placed there.

There are some principles that seem to enhance the technique's success. First, the likelihood of retrieval seems to be improved if the to-be-remembered element interacts in some way with the location. For example, if the memory task consisted of retrieving a list of words beginning with 'goat', a composite image of a goat simply standing passively on the doorstep of your residence hall might be less likely to be retrieved than would an image of a goat impatiently butting its head against the door. Some researchers have focused on the novelty, vividness, or 'bizarreness' of the composite image. So an image of a goat dressed up in the

uniform of doorperson, and opening the door for you as you leave your residence hall, might lead to still better retrieval. Neither the interaction nor the bizarreness is crucial for the technique to work, however. Some people experience an apparent loss of content addressability when using the method of loci. If you are like them, you may find that you cannot readily contact the seventh item on the list without mentally going through the first six. If then asked for the fifth item on the list, you may find that you cannot simply 'back up' two places from the seventh item, but rather must restart your mental walk.

Peg Mnemonics

'Peg' mnemonics exploit both the phonetic and imaginal properties of words. The peg is a concrete word that is readily imagined, and the peg word is constant. Each peg word in the system is used because it rhymes with the word for a particular ordinal position.

For example, the readily imagined word 'shoe' rhymes with the ordinal position word 'two', and so the peg word for two is 'shoe'. To learn an arbitrary list of objects that can be placed in an ordinal arrangement, the learner creates an image that incorporates both the object he or she is trying to recall with its associated ordinal position term. In our example, if you had a list of chores to do on Saturday morning, and the second chore involved getting the oil changed in your car, you might create an image of a shoe full of oil.

As with the method of loci, the learner must first master the rhyming list of peg words (Higbee, 1988). For example, the user first encodes the following list of associations:

One is a bun
Two is a shoe
Three is a tree
Four is a door
Five is a hive
Six is sticks
Seven is heaven
Eight is a gate
Nine is a line
Ten is a hen

To remember a set of objects, let's say a grocery list, each item on the list is associated in a visual image with the appropriate peg word. Thus, if the fourth item that you wanted to remember to get was a carton of milk, you might create an image of a door with a milk carton for a doorknocker. It's interesting to note that apparently several lists can be stored 'on' the same set of pegs, similar to hanging up more than one coat on a single peg. In other words, you might be able to remember the list of chores to be done, and the list of grocery items to be purchased, using the same list of peg words as shown above. Currently, it is not clear how our memory system avoids mixing the two lists up, despite the fact that it would seem very likely that such interference would occur.

Imagery

Interactive imagery is a technique that exploits the imaginal properties of words to create a single image in which unrelated elements are pictured together. For example, a person who wanted to remember a short list consisting of three unrelated words, 'farmer', 'cowboy', and 'Arkansas', might create an image of a farmer wearing a cowboy hat and being chased by a pig (that is, a razorback, the nickname of the University of Arkansas athletic teams). Apparently, rather lengthy lists of unrelated words can be retrieved using this technique, and it has an advantage over some other techniques in that much less prior work needs to be done to use it, compared to memorizing a list of peg words.

The creation of images is a common theme in most mnemonic techniques. However, the functional role of imagery in mnemonics continues to be debated. For example, Canellopoulou and Richardson (1998) tested the efficacy of image-based mnemonics in neurologically impaired patients. They found that those individuals who had intact central executive processing systems in working memory were able to profit from using the method of loci in a free recall task, whereas individuals who were impaired in this way were not able to use the method of loci, even when they were given imagery instructions. This finding suggests that it is not the image itself that is really 'doing the work' at retrieval time, but rather the ability of the working memory executive to simultaneously manage all the information that is needed to create the image in the first place.

SITUATIONS WHERE MNEMONIC METHODS ARE APPROPRIATE

Current thinking is that the number of retrieval situations that can be aided by mnemonics is actually rather limited (Searleman and Herrmann, 1994, p. 356). It is the case that the method of loci or a peg-word system can definitely enhance the retrieval of a list of unrelated words, but that is not a task we are often called upon to perform. Further, mnemonic techniques have not been shown to produce

dramatic improvements on the types of tasks we actually do. For example, mnemonic techniques do not generally lead to improvements in retrieving poems, stories, or lines in a play (the performances of the bards notwithstanding), nor do they lead to improved retrieval or execution of a series of physical actions such as in dance, gymnastics, or the playing of a musical instrument.

A finding reported by van Hell and Mahn (1997) is typical. College students who were familiar with the general principles of language learning were given concrete and abstract words to learn either by rote or by an imaginal-based keyword system. Experienced learners learned more of the novel words by rote than they did with the keyword system. This pattern was also observed among novice foreign language learners, with the added disadvantage that more time was required by the mnemonic users to retrieve the words compared to the rote learners.

Even in cases when mnemonic techniques could be used, such as going to the grocery store and retrieving all the items needed to make Thanksgiving dinner, most people find it much easier just to write the items down beforehand rather than burden themselves with learning a mnemonic.

THEORETICAL IMPORT OF MNEMONICS

It has long been known that organized material is easier to learn and remember than is unorganized material (Baddeley, 1990, p. 182), even when the organizational principle is not made clear to the participants, but is left for them to discover. For example, many years ago, Deese (1959) showed his subjects a list of 15 words that were either high associates of an initial starting word, or words that were all unassociated with each other. For a list beginning with the word 'butterfly', for example, high associates might include the words 'moth', 'insect', and 'cocoon'. Words on an unassociated list might include 'university', 'zebra', 'native', and so on. Subjects typically recalled 33 percent more from the high associates list. The improved retrieval of apparently unrelated words that is observed when subjects use mnemonic techniques supports this fundamental theoretical principle because, no matter what else mnemonic techniques accomplish, and by whatever mechanism they achieve it, nevertheless their use forces their users to organize incoming material. Moreover, the organization that the subjects subjectively apply to the material seems to be superior as a memory aid than any organization imposed by the researcher, and this condition is also met when using mnemonics.

The retrieval phenomena that are seen in the mnemonics literature are also very consistent with what is known about semantic activation and cuing (Reisberg, 1997, p. 262). For example, in a semantic network, a word that is semantically associated, even weakly, with a number of different words has a greater chance of being retrieved than does a word with only one association, even if that association is semantically very powerful. The reason is that a word that is semantically associated with a number of words – that is, a word enmeshed in a well-developed and rich semantic network – is more likely to be successfully cued because it is likely that any one of its many associates is activated by a memory search. On the other hand, a word associated with only one or a few other words has rather few 'routes' of activation that may succeed in cuing it. This explanation accounts for some of the inflexibility in mnemonic processes discussed earlier in the section on the method of loci. Because there are no semantic links from one item to the next on the list, a person using the method of loci cannot readily cue a list item from another list item.

References

Baddeley A (1990) *Human Memory: Theory and Practice*. Boston, MA: Allyn & Bacon.

Canellopoulou M and Richardson JTE (1998) The role of executive function in imagery mnemonics: evidence from multiple sclerosis. *Neuropsychologia* **36**: 1181–1188.

Deese J (1959) Influence of inter-item associative strength upon immediate free recall. *Psychological Reports* **5**: 305–312.

Hell JG van and Mahn AC (1997) Keyboard mnemonics versus rote rehearsal: learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning* **47**: 507–546.

Higbee KL (1988) *Your Memory,* 2nd edn. New York, NY: Prentice-Hall.

Reisberg D (1997) *Cognition: Exploring the Science of the Mind*. New York, NY: WW Norton.

Searleman A and Herrmann D (1994) *Memory from a Broader Perspective*. New York, NY: McGraw-Hill.

Further Reading

Herrmann DJ (1987) Task appropriateness of mnemonic techniques. *Perceptual and Motor Skills* **64**: 171–178.

Lorayne H and Lucas J (1974) *The Memory Book*. New York, NY: Stein & Day.

Luria AR (1968) *The Mind of a Mnemonist*. New York, NY: Basic Books.

Memory Models

Intermediate article

Bennet B Murdock, University of Toronto, Toronto, Canada

CONTENTS

Memory models are formal (analytic or simulation) models of human episodic memory. Their goal is to help in the understanding of human episodic memory, and the test of their adequacy is their ability to fit or predict experimental findings.

INTRODUCTION

Memory models are information-processing models which assume that the storage and retrieval of information in human memory is a dynamic, not a static process. New information is constantly entering the system and old information is becoming weaker or less available. These models attempt to specify the processes formally, so that quantitative predictions are possible and the models can be tested by comparing their predictions with the extensive experimental data on human memory. The models tend to focus on short-term episodic memory rather than long-term semantic memory because there is more data available and testing the models is easier.

MATRIX MODELS

The prototypical matrix model is the linear-associator model (LAM) of Anderson. It deals with associations, the building blocks of human memory. More particularly, it deals with an association between input and output. For instance, I hear the word 'echo' and I can then say the word 'echo'. Even though on paper the two 'echos' are the same, in actuality one is a stimulus (input) and the other is a response (output).

The model assumes that an item such as the word 'echo' can be represented by a vector of features where the features are normally distributed with mean zero and variance 1/N. (N is the dimensionality of the vectors: i.e. the number of features.) The model does not specify what these features are, but it does specify their underlying distribution (i.e. normal or bell-shaped).

The basic mechanism for LAM is auto-association where the association is represented by the outer-product matrix of the item vector with itself. The i,jth element is simply the product of the ith element of the item with the jth element of the same item, and it is stored in the outer-product matrix. For recall, if the item is later applied to one of the inputs then the echo will appear on the other input, which now functions as the output. Technically this involves pre- or post-multiplication of the item vector with the memory matrix depending on the input channel.

The matrix can store many auto-associations; one by one their auto-association is formed and added to the memory matrix (superposition). The items must be independent (uncorrelated) and the capacity depends on N, the dimensionality of the item vectors. The system can even do redintegration; a partial cue (a subset of the item features) can reinstate a reasonable approximation of the original item.

Even though the model can be (and has been) extended to hetero-associations (associating pairs of items), it is quite limited in scope. However, it is simple and elegant, requires no search process, and has served as the basis for much later development in the field.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) started in the late 1950s but these simple two-layer networks fell into disfavor when some of their limitations (basically linear independence problems) became known. However, there was a resurgence of interest when it was realized that multiple-layer networks could overcome these limitations. Most ANNs start with biologically realistic principles of neural function and they then develop an architecture to implement these functions, generally by means of computer software.

Many of the developments focus on getting computers to do interesting or useful things, but here I will discuss TECO (Sikstrom, 2000), a particular ANN which attempts to explain much psychological data on human memory. Although a very recent development, TECO (target, event, cue, and object) probably applies to a wider range of memory data than any other ANN. In many ways it is a modified version of a Hopfield net, a classic ANN which by a nonlinear iterative procedure can clear up a 'noisy' output and turn it into an exact copy of the original input.

In TECO, items and context are represented by independent nodes, which take on the binary values of 0 (inactive) or 1 (active) where the probability of being active is much smaller than the probability of being inactive. Each node is connected to every other node and these connections form the weight matrix that constitutes the memory of an ANN. The binding of item and context is achieved by auto-association as in LAM. Each time an item is presented in a context the weight matrix is updated by a Hebbian learning rule so again like other ANNs, this model uses superposition.

To explain habituation TECO assumes that there are inhibitory signals in the nodes and that the nodes are connected in a chain, where early parts correspond to perceptual areas in the brain and later parts to more semantic areas. Habituation increases with stimulus repetition, decays over time, and varies with the input frequency. The amount of encoding, the degree of forgetting, and the ease of retrieval is diminished by habituation. Including habituation in ANNs accounts for several empirical phenomena; for example, the primacy effect, the rate sensitivity of the primacy effect, long-term recency, and the shift from primacy to recency with delay.

The mirror effect is a current puzzle in the memory area; in item recognition there are fewer false alarms (saying 'old' to a new item) with rare words than common words, but more hits (saying 'old' to a studied word) with rare words. This is a puzzle if you believe recognition judgments are based on memory-trace strength; how can rare words (initially weaker) leapfrog over common words and become stronger following a single presentation?

TECO assumes that strength is based on the summed output of many neural units. If the strengths of the individual units are normally distributed, and units for common words have larger variance because they are embedded in many contexts, then the mirror effect and other related phenomena regarding the distributions of familiarity should always hold if the network is tuned to optimal performance. Optimal performance would result if the criterion were placed between the means of the old and the new items. This optimal tuning also yields a smaller variance of familiarity for new items and rare words that is consistent with experimental results.

TECO can also explain power-function forgetting curves, recognition failure of recallable words, serial-position effects in free recall, and many other memory phenomena. It is not yet clear whether TECO is subject to catastrophic interference. In a retroactive inhibition study, if subjects learn a list of A–B paired associates and then learn a second list with the same A items but new B items, they gradually forget the original B items as they learn the new B items. However, in many ANNs the networks show 'catastrophic' (i.e. very rapid) unlearning of the original B items, so these models are not consistent with experimental data.

MINERVA2

MINERVA2 (Hintzman, 1988), presumably named after a major Roman goddess, is a computer-simulation model which differs from LAM and TECO in that its memory is serially organized like a computer. Each item (example or exemplar) is stored in its own location (like books on a library shelf) and as each new item arrives it is placed at the bottom of the stack. Thus, there is no superposition; each item occupies its own discrete location.

As in TECO, items consist of binary features (+1 and −1, though unknown or forgotten features are set to 0) and, again like TECO, context and content features are concatenated (placed side by side in the same item vector). However, there is no auto-association. For paired associates, the context vector and the two single-item vectors are concatenated in a single memory trace.

When a probe is presented it is simultaneously compared to all items in memory (or in the search set which is determined by context) and an activation value is computed from the probe–item similarity values. The similarity value is the normalized dot-product of the probe with each item (the dot-product is the sum of the product of all feature values) and it is normalized by the number of active (i.e. non-zero) features in that item. The activation value is the cube of the similarity value, and the echo intensity (or resonance) is the sum of all the activation values. This echo intensity forms the basis for decision in a recognition-memory test.

MINERVA2 attempts to show how generic knowledge can develop out of episodic memory when there is only a single memory system with individual storage of exemplars (items). The model first focused on a schema-abstraction task in which subjects learn to classify selected exemplars of a small number of prototypes without seeing (or being told about) the prototypes themselves. A standard result is that old exemplars are classified more accurately than the prototypes on an immediate test, but this reverses over time.

Simulations of MINERVA2 show that it is able to reproduce not only the standard result but other results from the schema-abstraction paradigm (e.g. old exemplars are consistently classified better than new exemplars, and for new exemplars low-level distortions are consistently classified better than high-level distortions). Thus, the model shows one way that semantic memory could arise out of episodic memory, and it does not postulate a separate semantic memory.

Another accomplishment of MINERVA2 is to successfully model both judgments of frequency and recognition memory within the same basic theoretical framework. Since MINERVA2 is a multiplex model (each presentation of an item lays down a separate trace), the more often an item had been presented the greater the echo-intensity mean and variance. Item recognition (present or absent) is the special case (1–0) of frequency judgements. The model can also account for many basic findings (list-length effects, orienting tasks, and similarity effects). Even though the number of items stored in memory could be very large, in principle the judicious use of context could allow the model to focus on recent events such as the particular items presented in the most recent list.

SAM

The SAM model (search of associative memory) (Gillund and Shiffrin, 1984) is an elaboration of the earlier buffer model of Atkinson and Shiffrin, one of the early (and very influential) formal models of memory. The model has been applied primarily to item recognition and free recall, although it does make predictions about cued recall (paired associates) as well.

In SAM items are represented as images (nature not specified) and retrieval cues are used to activate (retrieve) these images. The efficacy of these cues depends on their strength, of which there are four kinds: context, item, associative, and background. These strengths depend on how long items or pairs of items have resided in the memory buffer. The memory (or rehearsal) buffer has a limited capacity (generally four single items or two pairs), and the image-strengths increase with the duration of buffer occupancy.

For recognition, the item(s) and the context cue are compared with all the items in the list, and a yes or no response is given according to whether or not the summed strength is above or below a predetermined criterion. The summed strength is the sum of the products of the item(s) and context cues with all the relevant images, so the model can apply to both item (old/new) or pair (intact/rearranged) probes. To introduce variability into the model a three-point variability scale is assumed so with this, the strength parameters, and the criterion value the model can be used to make quantitative predictions for experimental results obtained under specified conditions. However, unlike LAM, TECO, and MINERVA2 it is not a process model; rather it is a way of computing the predicted values by varying the parameter values to see if the model can produce results which match the experimental data.

For free recall, a task where the subject is asked to recall as many items as possible from the most recently presented list, there is a sampling and a recovery process. First all the items in the buffer are recalled (they are usually the last few items in the list), then using context as the cue, items are sampled until a new item is found and, if recovered (never certain), the context and recovered item are used to repeat the process until an arbitrary criterion is reached. This criterion determines the number of unsuccessful sample-and-recovery cycles the subject tolerates, and once this criterion is exceeded the process stops.

The SAM model has been very successful in accounting for a wide variety of data, at least at the qualitative level. In this regard, it is probably the best model we have. However, it has a large number of parameters (in the 1984 version, 13 for recognition and 18 for recall) and, depending on the application, one or more of these parameter values may be varied to make the model fit. In such cases it is the parameter(s), not the model, that is doing the work. On the other hand, the authors have often emphasized the qualitative rather than the quantitative predictions, and perhaps the parameters simply indicate the number of processes that must be involved.

OSCAR

OSCAR (Brown *et al.*, 2000), an acronym for oscillator-based associative recall, is different from the

models discussed so far because it deals primarily with the storage and retrieval of serial-order information. Serial-order information is what allows us to retrieve a list of items in order; for example, the days of the week, our access codes, how to spell words. This is probably the most challenging problem in the episodic memory area; there are many complex and puzzling phenomena that our models must explain.

Traditionally there have been two types of models, those based on item-to-item associations and those based on item-to-position associations. Item-to-item models are generally chaining models; if the list is ABCDE then A–B are assumed to be associated (linked), and B–C, C–D, up through D–E. Then if we can recall A (which can be a mystery in a chaining model) then A will give us B, B will give us C, and so on, until we have recalled the entire string. Thus, serial-order information is assumed to be based on linked pairwise associations.

Item-to-position models assume there is some fixed internal representation of ordinal position, perhaps like adjacent storage registers in a computer. Call these bins; then the first item goes in the first bin, the second item goes in the second bin, and so on. To retrieve the string we examine the bins in order and read out their contents. Quite apart from what the nature of these bins is, how can we remember at the same time short lists of items, such as 'IBM' or 'PC', and at the same time much longer strings, such as the letters of the alphabet or how to spell 'encyclopedia'?

OSCAR is basically an item-to-position model, and provides a simple and elegant answer to both these questions. It assumes there are a number of endogenous asynchronous oscillators in the brain that provide timing signals which provide the functional 'bins' required by any item-to-position model. That is, each item is yoked to a particular set of values of these timing signals, so to retrieve, say, a telephone number you have just seen or heard you reset the oscillators to their original initial value; then by recycling the oscillators you regenerate the telephone number.

Technically, items are represented as vectors of features (like LAM and TECO) and they are associated in an outer-product matrix with the context vector. The context vector is the output of the oscillators which is constantly changing. Recall can go in either direction, from item to context or from context to item. The model does not say much about recognition, but probably that would be easy to implement.

There are several sources of forgetting. Perhaps the learning context can be only partially reinstated at the time of recall. If only a subset of these oscillators are correctly reset then the others continue to evolve during recall and this degrades performance. Another possible source would be an inability to reinstate the oscillators to their correct initial value. This would presumably be more likely the longer the delay between study and test, and might account for the precipitous forgetting of short lists of items following a brief period of interpolated activity. Yet a third source of forgetting would be the interference among the multiple lists of items resulting from superposition in the common memory matrix.

According to the authors, one of the main reasons for preferring a model based on positional associations rather than sequential item-to-item associations is the effect of inter-item similarity on recall. If two list items, call them A and A′, are similar, and if A is followed by B but A′ is followed by D, then a chaining model would have to predict confusion in recall because, given A, there would be competing B and D associations. However, this does not seem to be the case; many studies have shown little or no difference between experimental and control lists where the control lists did not have the A–A′ similarity. OSCAR does not make this prediction because items are associated to context, not to their predecessors, so the similarity between A and A′ should not matter.

To give some idea of the explanatory power of this model, here are some of the various phenomena that OSCAR can model: separate item and order memory, serial learning, effects of item similarity on order memory, judgments of relative recency, judgments of absolute recency, probed serial recall, serial-position effects in recognition, grouping effects, grouping errors, group size effects, item and list memory, list-length effects, vocabulary size effects, and presentation rate effects. Computer simulations are run to see if the model generates results that mimic experimental data.

To give at least a very brief account of the necessary technical machinery, the processes in OSCAR (or any quantitative memory model) vary with the numerical values of the parameters just as the speed of a falling object depends upon the numerical value of the constant (parameter) for gravitational attraction. In OSCAR (again as in any other model) there are a number of parameters: the amount of change between successive learning contexts, the number of repeating learning-context vectors, the dimensionality of the vectors and their similarity, the learning rate, the learning-rate decay, the weight decay (cf. TECO), the proportion of learning context reinstatable at retrieval, the

vocabulary size, the amount of output interference, the inhibition of recalled items (present or absent), and the level of the output threshold. As many as possible of these parameter values are held constant across different applications but, as noted above, *ad hoc* changes in parameter values for specific applications weaken the explanatory value of the model.

Although OSCAR is a very impressive model of serial-order effects, questions can be raised about the context reinstatement mechanism. As the Greek philosopher Heraclitus noted (about 400 BC) you cannot step in the same river twice. Oscillators are often used in neural models of timing and counting, but there the operation is relatively automatic and they are not under voluntary control. Also, the model has not yet been applied to many of the phenomena dealing with the storage and retrieval of item and associative information, and these are important areas too.

TODAM2

TODAM2 (Murdock, 1997), the second version of a theory of distributed associative memory, is most like LAM in that it uses the same assumptions for item representation and the comparison process for recognition, but it uses a different formalism for storage and retrieval. TODAM2, like the original version, convolves item vectors so the association is still a vector, whereas LAM uses an outer product so the result is a matrix. While the vector formalism is noisier than the matrix formalism, it is more economical in terms of storage space and, more important, generalizes to multiple convolutions to explain serial-order effects. So TODAM2 can apply to item, associative, and serial-order information, and this is one of its main strengths.

Convolution and correlation are approximately inverse operations. If you convolve two items (two item vectors) and then correlate the result with one of them you recover an approximation to the other. It works either way, so 'forward' and 'backward' recall should be (as they are) equally good if the items forming the pairs are drawn from the sample pool of items.

Since the retrieved information is only an approximation to the target item in TODAM2, there are two processes involved in cued recall: correlation and deblurring (clean-up). This is like MINERVA2 except that the details of the deblurring process are not specified in TODAM2 as they are in MINERVA2. There are also two processes in recognition: a comparison process (the dot-product as in MINERVA2) and a decision process (as in SAM). The two-process recognition theories make it possible to use the classic signal detection theory to separate strength effects from interior effects, an essential requirement for any realistic theory of recognition memory.

To remember a list of items, TODAM2 assumes (like MINERVA2) that items and context are concatenated and then (unlike MINERVA2) auto-associated (auto-convolution) to bind items and context. Every auto-convolution is stored (superposition) in a common memory vector (like LAM), but context drifts slowly over the course of list presentation. It is this steady context drift that produces the extensive recency effect typically found in latency and accuracy measures in item recognition.

When a list of paired associates is presented, it is assumed that subjects store both the item and the associative information, again by superposition in a common memory vector. Each of the two items in the pair is concatenated with context and auto-convolved, and the two items in the pair are summed, concatenated, auto-convolved, and also added to the memory vector. This may sound unnecessarily complex, but it is necessary to explain some puzzling interactions (differential forgetting of item and associative information, and asymmetric effects of differential attention to items and pairs) that have recently been reported.

For serial order, TODAM2 assumes that subjects form chunks when they study a short list of items. A chunk is a sum of n-grams where an n-gram is the n-way auto-convolution of the sum of n items. While this results in a rather noisy memory trace, memory span is typically only four to five words with college students so the chunks do not have to be very big. If subjects also store item information in the common memory vector, then TODAM2 can also explain the characteristic primacy effects in ordered recall and recency effects in item recognition that are puzzling for many other models of serial-order effects.

DISCUSSION

The reader may have noticed many common mechanisms and processes in the memory models discussed here. In fact, the similarity is more real than apparent; very different terminology may obscure underlying similarities. While one can bemoan the lack of a standard unified model, perhaps the consensus is greater than it might seem. Perhaps we are starting to converge on a common model, and this will happen when the right person takes the right set of assumptions and puts them together in the right way.

References

Brown GDA, Preece T and Hulme C (2000) Oscillator-based memory for serial order. *Psychological Review* **107**(1): 127–181.

Gillund G and Shiffrin RM (1984) A retrieval model for both recognition and recall. *Psychological Review* **91**: 1–67.

Hintzman DL (1988) Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* **95**: 528–551.

Murdock BB (1997) Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review* **104**: 839–862.

Sikstrom S (2000) The TECO theory and lawful dependency in successive episodic tests. *Quarterly Journal of Experimental Psychology* **53**: 693–728.

Further Reading

Anderson JA (1995) An introduction to neural networks. Cambridge, MA: MIT Press.

Healy AF, Kosslyn SM and Shiffrin RM (eds) (1992) *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, vol. 1. Hillsdale, NJ: Lawrence Erlbaum.

Izawa C (ed.) (1999) *On Human Memory: Evolution, Progress, and Reflections on the 30th Anniversary of the Atkinson–Shiffrin model*. Mahwah, NJ: Lawrence Erlbaum.

Murdock BB (1974) *Human Memory: Theory and Data*. Potomac, MD: Lawrence Erlbaum.

Neath I (1998) *Human Memory: An Introduction to Research, Data, and Theory*. Pacific Grove, CA: Brooks/Cole.

Memory, Development of

Intermediate article

Patricia J Bauer, University of Minnesota Institute of Child Development, Minneapolis, Minnesota, USA
Dana L Van Abbema, St Mary's College of Maryland, St Mary's City, Maryland, USA

CONTENTS

Throughout infancy and early childhood there are remarkable changes in the ability to remember everyday and emotional events. More frequent and effective use of memory strategies, increases in speed of processing, and developments in metamemory all contribute to the changes.

INTRODUCTION

The development of the many facets of human memory has received much attention in recent decades, and deservedly so. In addition to being a cornerstone in the broader study of cognitive development, the study of memory has implications for our understanding of children's behavior and development in a number of other domains. Consider, for example, something as basic as how a 1-year-old child's ability to recognize familiar adults influences the manner in which the child interacts with them. The development of memory has implications for adult memory and cognition as well. Few would disagree that the mnemonic ability of an adult is significantly advanced compared with that of an infant, and furthermore that there appear to be differences in mnemonic competence between individuals. The study of memory during the intervening years can shed light on the mechanisms of developmental change and also on the nature and significance of individual differences.

For brevity, this discussion is confined to children's capacity for declarative memory, which captures most of what we think of when we refer to 'remembering'. Declarative memory includes explicit recognition and recall of names, objects and events, but not those abilities that are inaccessible to conscious awareness such as skill learning and priming.

CHILDREN'S MEMORY FOR EVERYDAY AND EMOTIONAL EVENTS

As adults, we seldom contemplate the pervasive role that memory plays in everyday life. Certainly it is essential to remember where one lives and adaptive to remember the names of one's in-laws; equally important, however, is one's memory for the wide range of past events that have been experienced. This is true for both personally meaningful events such as one's wedding day, and for events as mundane as a trip to the grocery store. Whereas it is primarily memories of the former type that accumulate to shape one's 'life history', memories of both types of events serve to influence future behaviors, in part by allowing for generalized scripts about the course of actions in a given event.

Perhaps because memory seems effortless, it is easy to underestimate the range of skills involved in remembering. Sharing a past event involves projecting back in time to 'locate' the event in question, organizing the thoughts and images associated with the event, and generating a coherent narrative about it. Somewhat miraculously, children are able to do this all by 3 years of age. This has led researchers to investigate the emergence of this ability in infancy, and how it develops in the years that follow.

Memory for Events in Infancy and Early Childhood

The paradigm of choice for research on declarative memory in older children and adults is simple – verbal report. This is not an option with very young children, however, and it was not until the establishment of a suitable nonverbal analogue to this method that the early foundations of declarative memory could reliably be assessed. In the

mid-1980s researchers began using deferred imitation as a test of declarative recall during the first 3 years of life (see Bauer *et al*., 2000, for a review). Originally suggested by Jean Piaget as a hallmark of the development of symbolic representation, deferred imitation as used in laboratories today involves using an object to produce a single action (or sequence of actions) and then inviting the infant or young child to imitate it immediately, after a delay, or both. Researchers examine recall by assessing whether children reproduce the modeled actions and whether they perform a sequence of actions in the correct temporal order. The use of this technique with children ranging in age from 6 months to 36 months affords a comprehensive picture of developments in mnemonic ability in children who are preverbal or newly verbal.

By the end of the first year of life infants shown two-step sequences demonstrate retention of event-related information over delays as long as 5 weeks, and many of them evidence recall that is temporally ordered (Carver and Bauer, 1999). Over the next 2 years of life a consolidation of declarative mnemonic processes appears to occur, as indicated by age-related changes in the reliability and robustness of long-term ordered recall. For example, more than half of a sample of children who experienced four-step sequences at 20 months of age evidenced ordered recall after delays of up to 12 months (Bauer *et al*., 2000). It is important to note, however, that the developments occurring over the second and third years of life do not reflect a simple 'growth chart' function for the length of time over which memory will persist. Research consistently shows that memory for past experiences is determined in part by many additional factors such as the nature of the temporal relations inherent in an event, the number and timing of exposures to the event, active participation in the event, and the availability of cues or reminders of the to-be-remembered event. For example, whereas older children on average remember events for longer periods than younger children, age differences are reduced when children are given cues or reminders of to-be-remembered events. These factors continue to affect the formation of long-term memories later in childhood and beyond, even as the mode of expression of changes from exclusively nonverbal to primarily verbal.

Development of Verbal Expression of Event Memory

By their third birthday children are relatively adept at talking about the 'here and now' and are increasingly able to participate in discussions about past events. Considering that in most memory conversations children are not only temporally but also spatially removed from the event being discussed, it is not surprising that talking about the past lags behind talking about the present and requires more 'scaffolding'. Initially children's talk about the past is confined to simple responses to specific prompts ('What did you do at the museum?') and is reflective of the more typical aspects of events experienced ('I ate a cookie'). With continued guidance from caregivers in how and why to talk about the past, children's memory responses gradually become longer and more informative. As children learn about what is important within the context of an event and about the structure in which to frame and share their experiences, the focus of their memory talk turns increasingly to the more unique and meaningful components of the occurrence and the overall account becomes more exhaustive. Notably, recall at all ages is largely accurate, and temporal order of the events tends to be preserved just as it was in earlier nonverbal 'reports'.

The growing body of research on children's autobiographical memory, or memory for personally meaningful events, consistently reveals remarkable mnemonic competence. An influential study by Hamond and Fivush (1991), in which children aged 3½–6 years were interviewed about a trip to a theme park that occurred up to 18 months earlier, revealed that regardless of age, children were able to provide thorough accounts of their experiences and that the overall amount of information recalled did not differ as a function of age. Older children did, however, offer a greater quantity of information spontaneously and included more detail than did younger children. This suggests that as children gain more control over retrieval processes they are less dependent on cues and are increasingly able to recall and verbalize details. After more significant delays of several years, a strong reliance on cues is apparent in children of all ages. For example, when Hudson and Fivush (1991) interviewed 11-year-old children about a trip to an interactive archeology exhibit when they were 5 years old, only one child commented on the experience in response to a general open-ended prompt; nearly all of them, however, provided accurate reports of the event in response to specific questions.

Children's Memory for Emotional Events

Whereas the content of events targeted for recall in studies of autobiographical memory typically has

been positive or at least neutral, not all events that children experience, unfortunately, are positive in nature. Because it is not clear whether conclusions drawn from this body of data can be applied to memory for all types of experiences, research targeting emotionally negative, stressful, and traumatic events is necessary. Thorough awareness of memory processes under such conditions has important implications for the understanding of children's adjustment following trauma and for the evaluation of children's credibility as eyewitnesses.

For practical and ethical reasons researchers cannot intentionally subject children to controlled laboratory-based emotional events. Emergency medical treatments and prescribed medical procedures, however, afford researchers opportunities to explore children's memory for stressful real-life experiences. Though socially sanctioned and performed for the child's own good, such measures nonetheless are perceived by children as unpleasant, painful, frightening, and executed against their will. In their examination of reports by children aged 3–10 years of a painful and embarrassing catheterization procedure, Goodman *et al*. (1994) found that older children recalled more information, answered more questions correctly and made fewer errors than the youngest children. When age was controlled, it was clear that children's emotional reactions and understanding of the event (as rated by the parent), as well as maternal communication and compassion, affected children's memories.

How can children's memories of stressful, traumatic events be expected to differ from those of benign events? In a review of the relevant research, Fivush (1998) noted that the little evidence available suggests that 'memories of trauma are at least as detailed if not more so than memories of more mundane experiences' (p. 709). She emphasized, however, that direct examination of memories for different types of events is uncommon and that comparisons across studies must be interpreted with caution. In a study that included the within-participant comparison necessary to evaluate differential reporting, we asked mother–child dyads to discuss the touchdown of a devastating tornado as well as an event that preceded and an event that followed the tornado. Mothers and children produced more complete verbal reports of their recollections of the tornado than of the other events; their reports about the tornado also were more narratively coherent. These findings indicate that although children's memories of events of a highly emotional nature probably follow a developmental course similar to that for more benign events, they are likely to differ in important ways as well. More research in which memories for benign and emotional events are directly compared is needed in order to identify characteristics that are unique to memories of emotional experiences and those that are common to all event memories.

EMERGENCE AND REFINEMENT OF CHILDREN'S DELIBERATE MEMORY STRATEGIES

Forming a memory of any type of event in which one is participating takes little conscious effort. If one is present, wakeful and attentive, at least some aspects of the experience are likely to linger in memory. In other situations we use memory in a more intentional way, as in repeatedly rehearsing an important telephone number when pen and paper are nowhere to be found. Research on children's use of deliberately implemented encoding strategies such as rehearsal, organization, and elaboration suggests that younger children generally have fewer memory strategies available to them and use them less efficiently than do older children; see Bjorklund and Douglas (1997) for a review. The developmental picture is not a simple one, however; although one might expect that children would use one strategy reliably on a given task and then replace it when they learn or discover a more effective one, in practice this rarely happens.

Early Strategies: Selective Attending and Rehearsal

For some time it was thought that preschool children were astrategic. When tasks are simplified and situated in a familiar context, however, it appears that such young children do indeed do things to help them remember, such as frequently looking towards the item or place that they have been instructed to remember. Though preschoolers may be more competent mnemonists than once believed, the bulk of development in strategy use nonetheless occurs in middle and later childhood, probably as a combined result of cognitive maturation and the expectations and training that accompany formal schooling. One simple yet effective strategy that appears in the early school years is that of rehearsal, or repetition of to-be-remembered information. With age, the use of rehearsal becomes more common, and also more efficient. For example, if instructed to remember the words in a list, it is not until about 12 years of age that children will consistently rehearse in a cumulative fashion, reciting previous words along with each new word

that is stated. Even in its simpler form, however, rehearsal appears to aid recall.

Later Strategies: Organization and Elaboration

One potentially more useful memory strategy is that of *organization*. Imagine that you are given the following list to remember: carrot, pencil, monkey, celery, fish, tomato, cat, chalk, crayon. If you merely rehearsed the words with little attention to their meaning, you might have difficulty remembering all nine of them. The task becomes more manageable, however, if you realize that the words fall into three categories – vegetables, animals, and writing implements – that can be used to aid storage and retrieval. In adults, conceptually clustered recall of words that were originally presented in random order is taken as evidence of organized encoding and is associated with higher levels of recall. At younger ages, organization is assessed through physical sort–recall tasks in which children are given a study period during which they could potentially sort the randomly presented items or pictures of items. Young children rarely demonstrate spontaneous sorting behavior and typically perform no better on groups with obvious categories than on groups without them. By about 10 years of age, however, children begin to exhibit spontaneous, effective organization and to experience a resultant improvement in memory performance.

The final memory strategy to be covered, known as *elaboration*, has typically been examined in the context of a paired associate task in which participants learn pairs of unrelated items (e.g., 'buffalo' and 'tuba') and are subsequently asked to recall one member of the pair when given the other. Because participants must impose some sort of elaborate association where none exists (in our example, a buffalo playing a tuba!) this is perhaps the most complex, and certainly the most creative, of the memory strategies. Not surprisingly, it is a latecomer to the strategic repertoire, rarely appearing before adolescence or even adulthood.

PRODUCTION VERSUS MEDIATION DEFICIENCIES

To form a complete picture of the development of strategy use, we must understand why children fail to use strategies. When young children do not spontaneously use more advanced memory strategies, is it because they lack the cognitive capacity necessary to do so? A wealth of research revealing children's abilities to use and benefit from new strategies when given appropriate instruction suggests the contrary. Failure on the part of children to spontaneously produce strategies that they are capable of using is known as a *production deficiency*. Evidence for production deficiencies has been found consistently in both cumulative rehearsal and organization strategies among young children and in elaboration among school-age children. By manipulating factors such as familiarity of context and metacognitive awareness (see below) that are thought to influence children's subsequent use of trained strategies, it is possible to gain insight into why children do not use manageable strategies when it is in their best interest to do so.

Even when children are taught a strategy it does not necessarily improve their performance. So-called mediation deficiencies are most frequently observed with regard to the more sophisticated strategies of organization and elaboration. The more recently identified *utilization deficiency* also is characterized by a lack of expected improvement in memory performance, but refers specifically to children's spontaneous rather than trained production of a new strategy, and typically occurs during the early phases of individual strategy acquisition. Put differently, children exhibiting a utilization deficiency can produce the given strategy on their own, but they experience little, if any, enhancement in task performance. One commonly offered explanation for this deficiency that has received considerable empirical support suggests that the mental effort needed to execute the new strategy is so immense that the resources remaining are insufficient for encoding *per se*. As is the case with mediation deficiencies, utilization deficiencies are most frequently apparent when children employ the strategies of organization and elaboration.

Clearly, the development of children's strategy use does not proceed in a tidy, step-like fashion from employment of the least to the most effective strategies. Considerable variability is to be expected across children, across tasks, across trials, and even within the course of a single trial. Robert Siegler's strategy choice model, developed primarily in the domain of arithmetic strategies, has much to offer the understanding of children's memory strategy use as well. According to Siegler (1996), children have a variety of available strategies in their repertoire that compete with one another for use, with the winner on any given trial being multiply determined by the learning context, task variables, and the child's general cognitive ability. With maturation and practice, more sophisticated and efficient strategies win increasingly often while more

primitive strategies win less and less – but nevertheless continue to stay in the game for some time.

AGE-RELATED CHANGES IN THE SPEED OF PROCESSING

One particular aspect of children's general cognitive ability that has been consistently nominated as a contributor to observed age-related differences in strategy use is the speed with which children are able to process information. Not surprisingly, the amount of time needed to execute cognitive operations, including the memory strategies just described, decreases reliably with age (Kail, 1993). Age-related developments in speed of processing appear to be influenced by at least two factors. First, with maturation children experience an increase in the myelinization of neurons in the associative areas of the brain. Myelin is a fatty substance that surrounds nerve cells and acts as a sort of insulation, facilitating the efficient transmission of impulses. Though myelinization of the sensory and motor areas of the brain is completed within the first years of life, that of the associative areas is not completed until adolescence.

Processing speed is not determined solely by maturation, however; a second factor that has been implicated in increased speed of processing is children's growing knowledge base. Unlike myelinization, which exerts its impact on processing speed across domains, the influence of the knowledge base is domain-specific. That is, improvements in processing can be expected only in those areas that have experienced a growth in knowledge. Whereas some domains of knowledge are developed on an individual basis (e.g. children who are experts in chess or dinosaur knowledge), others are likely to manifest rather predictably among children, as in the case of overall growth in knowledge and vocabulary across the school years. In addition to the speed with which information can be processed, individual and age-related differences in knowledge within a domain have an effect on other aspects of memory, such as the frequency and efficiency with which strategies are used. Generally speaking, the more one knows about a given topic or area, the easier it is to remember new information about it.

There are several ways in which changes in speed of processing could be expected to influence children's strategy use. Most simply, if the task at hand is one that involves continuing interference, as in the oral presentation of a list of words, the ability to quickly encode, recognize, and act upon each new word is essential for successful performance. The faster processing that is seen with age increases the likelihood of completing the necessary cognitive operations before the next item is presented, thus contributing to observed improvements. The role of the knowledge base in speed of processing comes into play when you consider that children's abilities to use the more sophisticated strategies of organization and elaboration depend on their understanding of the to-be-remembered items and familiarity with the relations between them. For example, it would be impossible to impose any sort of organization on the terms tibia, hippocampus, femur, amygdala, and patella if you had not yet acquired knowledge of the human brain and skeleton. With limited familiarity with the domain at hand, the increased effort needed to deploy a strategy could easily result in a utilization deficiency.

METAMEMORY

Another aspect of children's developing cognitive ability that merits coverage is metamemory, or knowledge of the contents, processes, and limitations of memory. Specifically, it includes awareness of one's memory strengths and weaknesses, appreciation of variation in task difficulty, understanding of potential memory strategies and of situations that warrant their use, and ongoing assessment of performance during a memory task.

One simple yet informative way to assess children's metamemory knowledge at different ages is to ask them. In an oft-cited comprehensive interview study by Kreutzer and colleagues, children in kindergarten and in the first, third and fifth grades of school were asked a series of questions such as whether they remember better than other people they know, if learning pairs of related words would be easier than learning pairs of unrelated words, whether a delay would affect their ability to remember, and how they would go about looking for a jacket that was lost at school (Kreutzer *et al*., 1975). Results suggest that although young children are competent in some respects, they have an overly optimistic view of their memory abilities and have difficulty evaluating the impact of assorted variables (e.g. delay) on memory performance. For example, kindergarten children know that it is easier to remember a few items than it is to remember many, but do not understand that memory abilities vary across individuals and across situations. Results of this and more recent work suggest that young children's rudimentary awareness of memory processes gives way to quite sophisticated understanding by

about 12 years of age and refinement in the years that follow.

An important question for researchers in this area is whether improved metamemory actually translates into more advanced memory behaviors and enhanced memory performance. Although the strength of such relations was once questioned, more recent research suggests a moderate yet reliable contribution, varying in part based on the child's age, the type and difficulty of task, and the nature of the metamemory assessment. The relation appears to be strongest for older children, but is evident at earlier ages when the task is simple and familiar and the metamemory processes questioned are closely tied to success on the task. Some researchers have proposed that the component of metamemory that is most closely related to behavior is not simply the knowledge that a particular memory strategy works, but rather why it works. Importantly, research evidence suggests that the relation between metamemory and memory behaviors is bidirectional; that is, improved metamemory leads to more effective strategy use, which in turn contributes to continued improvements in metamemory.

As you might imagine, metamemory plays an important role in reversing the deficiencies observed in strategy use. As children become increasingly aware of what strategies are available, and when strategies are required, production deficiencies are less likely to result. Furthermore, children who are able to engage in insightful reflection upon memory processes might continue using a difficult, sophisticated strategy despite little gain if they believed that it would be effective with additional practice, resulting in the eventual elimination of a utilization deficiency.

As this brief review suggests, the study of memory development is multifaceted. Whereas human infants remember specific past events by late in their first year, there are substantial and significant changes in memory capacity over the course of the second year of life and well beyond. Understanding the myriad developmental changes that occur is an important piece of the ontogenetic puzzle.

References

Bauer PJ, Wenner JA, Dropik PL and Wewerka S (2000) Parameters of remembering and forgetting in the transition from infancy to early childhood. *Monographs of the Society for Research in Child Development* **65**: 4.

Bjorklund DF and Douglas RN (1997) The development of memory strategies. In: Cowan N (ed.) *The Development of Memory in Childhood*, pp. 201–246. Hove, UK: Psychology Press.

Carver LJ and Bauer PJ (1999) When the event is more than the sum of its parts: nine-month-olds' long-term ordered recall. *Memory* **7**: 147–174.

Fivush R (1998) Children's recollections of traumatic and nontraumatic events. *Development and Psychopathology* **10**: 699–716.

Goodman GS, Quas JA, Batterman-Faunce JM, Riddlesberger MM and Kuhn J (1994) Predictors of accurate and inaccurate memories of traumatic events experienced in childhood. *Consciousness and Cognition* **3**: 269–294.

Hamond NR and Fivush R (1991) Memories of Mickey Mouse: young children recount their trip to Disney World. *Cognitive Development* **6**: 433–448.

Hudson JA and Fivush R (1991) As time goes by: sixth graders remember a kindergarten experience. *Applied Cognitive Psychology* **5**: 346–360.

Kail R (1993) The role of a global mechanism in developmental change in speed of processing. In: Howe ML and Pasnak R (eds) *Emerging Themes in Cognitive Development*, vol. 1, *Foundations*. New York, NY: Springer-Verlag.

Kreutzer MA, Leonard C and Flavell JH (1975) An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development* **40**: 1.

Siegler RS (1996) Cognitive variability: a key to understanding cognitive development. *Current Directions in Psychological Science* **3**: 1–5.

Further Reading

Bauer P, Kroupina M, Schwade J, Dropik P and Wewerka S (1998) If memory serves, will language? Later verbal accessibility of early memories. *Development and Psychopathology* **10**: 655–679.

Bjorklund DF (1987) How age changes in knowledge base contribute to the development of children's memory: an interpretive review. *Developmental Review* **7**: 93–130.

Cowan N (ed.) (1997) *The Development of Memory in Childhood*. Hove, UK: Psychology Press.

Howe ML and Courage ML (1997) The emergence and early development of autobiographical memory. *Psychological Review* **104**: 499–523.

Meltzoff AN (1995) What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. *Journal of Experimental Child Psychology* **59**: 497–515.

Miller PH and Seier WL (1994) Strategy utilization deficiencies in children: when, where, and why. In: Reese HW (ed.) *Advances in Child Development and Behavior*, vol. 25. New York, NY: Academic Press.

Nelson CA (2000) Neural plasticity and human development: the role of early experience in sculpting memory systems. *Developmental Science* **3**: 115–130.

Nelson K (1993) The psychological and social origins of autobiographical memory. *Psychological Science* **4**: 7–14.

Schneider W and Bjorklund DF (1998) *Memory*. In: Kuhn D and Siegler RS (eds) *Handbook of Child Psychology*, vol. 2, *Cognition, Perception, and Language*, 5th edn, pp. 467–521. New York, NY: John Wiley.

Schneider W and Pressley M (1997) *Memory Development between 2 and 20*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Memory, Long-term

Introductory article

Alan Richardson-Klavehn, Goldsmiths College, University of London, London, UK
Robert A Bjork, University of California, Los Angeles, USA

CONTENTS

Long-term memory is central to cognitive functioning. Taking a wide variety of forms, from skills to general knowledge to memory for personal experiences, it is characterized by dynamic interactions between encoding and retrieval processes and by constructive processes, and thus differs fundamentally from current human-made information storage systems.

DEFINITION AND CLASSIFICATION OF LONG-TERM MEMORY

The ability to retain information over long periods is fundamental to intelligent thought and behavior. Memory is the 'glue', in effect, that holds our intellectual processes together, from perception, attention, and language, to reasoning, decision-making, and problem-solving. Memory also plays a critical role in social and emotional functioning, because our sense of who we and other people are is distilled from factual and autobiographical information in our memories. The study of memory, therefore, occupies a central position in the cognitive sciences.

Broadly, memory can be broken into three stages of information processing: (1) encoding, the transformation of information into a form retainable in memory; (2) storage, the holding of information in memory across a time interval; and (3) retrieval, the accessing of information from storage after a time interval and the use of that information to guide thought and behavior. This distinction between stages is important but – as clarified later – encoding and retrieval processes are intimately interconnected and cannot be understood in isolation from each other. (*See* **Information Processing**)

Distinguishing between Short-term and Long-term Memory

In everyday discourse, long-term memory is usually distinguished from short-term memory in terms of the time that has elapsed since information was encoded. Moreover, it is not unusual to find memory that persists over days or weeks being described as short-term memory. In psychology, however, the terms long-term and short-term memory have come to have specialized meanings that stem from a distinction made by William James in 1890. James observed that our consciousness is not just of the immediate present: the information that we currently attend to lingers in consciousness for some period of time. He called this lingering consciousness primary memory, and distinguished it from secondary memory, which occurs when information has left consciousness but returns to it again later. Thus, secondary memory involves retrieval in a way that primary memory does not.

James's distinction is not simply one of retention interval. It would, theoretically, be possible to retain information in primary memory indefinitely as long as one's attention remained focused on that information (i.e., as long as the information was rehearsed). Conversely, information that leaves consciousness and then returns to it is retrieved from secondary memory, even if retrieval occurs only seconds later. Thus the distinction between short-term (primary) and long-term (secondary) memory as used by psychologists is a qualitative one, not a simple quantitative one based on retention interval. In recent years, Alan Baddeley has introduced the term working memory to refer to short-term memory, which emphasizes its role in manipulating – as well as maintaining in consciousness – a variety of kinds of information. (*See* **James, William**; **Working Memory**)

Varied Forms of Long-term Memory

As with the short- versus long-term memory distinction, the main distinctions between forms of long-term memory involve reference to consciousness. The distinction between declarative and

procedural memory originated in computer science, where stored data structures were distinguished from stored programs specifying how the data were manipulated. Psychologists borrowed these terms to capture a distinction made by the philosopher Gilbert Ryle in 1949, between *knowing that* and *knowing how*. Declarative memory involves knowing consciously *that* particular events happened in one's past, or *that* particular facts are true (e.g., Paris is the capital of France). Procedural memory, on the other hand, involves knowing *how* to manipulate mental or physical objects. Such knowledge is not necessarily consciously accessible and very difficult to communicate verbally. Explaining to someone how to ride a bicycle, for example, offers them scant assistance in learning that skill. Practicing such a skill is essential to its learning. The declarative/procedural distinction is closely associated with John R. Anderson and Larry Squire. (*See* **ACT**; **Skill Acquisition: Models**; **Automaticity**; **Implicit Learning**; **Skill Learning**; **Knowledge Representation, Psychology of**)

Within declarative memory, episodic memory is distinguished from semantic memory. Episodic memory involves awareness of particular events in one's personal autobiography, whereas semantic memory involves knowledge of language, categories and concepts, and facts. This distinction is closely associated with Endel Tulving. Within episodic memory, in turn, recollection is distinguished from familiarity. Recollection involves re-experiencing the particular contextual details of a past event, such as the tone of voice in which a statement was uttered in the kitchen at nine o'clock yesterday morning. Familiarity involves the knowledge that a current situation bears some relationship to a past event, without awareness of the particular contextual details of that event. For example, we sometimes experience the strong sense that we have met someone before, without being able to recollect where and when we met them, or anything else about them. The recollection/familiarity distinction is closely associated with George Mandler and Larry Jacoby. A related distinction, between remembering and knowing, has been made by Tulving and by John Gardiner. (*See* **Semantic Memory: Computational Models**; **Episodic Memory, Computational Models of**; **Autobiographical Memory**; **Knowledge Representation, Psychology of**)

A final important distinction is between explicit and implicit memory. Explicit memory refers to conscious awareness of events in one's personal past that accompanies deliberate attempts to think back to those events. Implicit memory refers to influences of past events on one's current behavior that occur involuntarily or unintentionally, often without any current awareness of the relevant prior events. This distinction, closely associated with Daniel Schacter and Peter Graf, can be traced back to similar distinctions by Hermann Ebbinghaus, who published the first experimental studies of memory in 1885, and by a number of other influential thinkers going back to René Descartes in 1649. (*See* **Descartes, René**; **Ebbinghaus, Hermann**)

It must be noted that none of the foregoing distinctions is universally accepted. First, none is entirely clear-cut. For example, Paul Kolers and Henry Roediger have questioned the procedural/declarative distinction, arguing that all forms of memory involve the modification of procedures for manipulating information. And Schacter, Alan Richardson-Klavehn and others have pointed out that the explicit/implicit memory distinction is blurred by cases when conscious awareness of events in one's personal past comes about without any deliberate attempt to retrieve those events, a phenomenon termed involuntary explicit memory or involuntary conscious memory.

Second, a controversial question is whether these distinctions imply different information-processing mechanisms, with different bases in the structure and function of the brain. Support for the latter view comes from research by Brenda Milner, Elizabeth Warrington, Lawrence Weiskrantz and others on the amnesia (memory loss) that results from damage to limbic system structures in the brain (the hippocampus, portions of the thalamus, and connected structures). This memory loss is selective, resulting in dissociations between different measures of memory. For example, short-term memory is largely spared, whereas the acquisition of new long-term memories is severely impaired, and the declarative and explicit forms of long-term memory are impaired much more than the procedural and implicit forms. Dissociations that are similar in some respects are observed in dementias such as Alzheimer and Huntington diseases, as well as in the memory loss that accompanies normal aging. Such dissociations have led some, including Schacter, Squire, and Tulving, to argue that the brain has distinct memory systems, which may have had different evolutionary histories. Others, such as Kolers, Roediger, Mary Sue Weldon, and Bruce Whittlesea, argue that a unitary memory system – in which similar information-processing mechanisms handle a wide variety of kinds of information – can account for such dissociations.

As argued by Morris Moscovitch and others, resolving these issues will involve clarifying the

extent to which the varied forms of memory involve different versus common information-processing components and on understanding the relationship between these components and brain structure and function. Recent advances in imaging the activity of the living brain (neuroimaging) are making an important contribution in these respects. Whatever their interpretation, however, the selective memory impairments that have fueled the current controversies offer a striking illustration of the complexity and variety of memory, and of the centrality of memory to intellectual and social functioning. Such memory impairments often have a catastrophic effect on an individual's ability to hold down a job, remain informed about ongoing affairs in the world, and maintain normal social relationships. (*See* **Human Cognition**; **Aging and Cognition**; **Memory: Implicit versus Explicit**; **Memory, Development of**; **Neural Basis of Memory: Systems Level**; **Amnesia**; **Alzheimer Disease**; **Huntington Disease**; **Neuroimaging**)

THE DYNAMIC CHARACTER OF LONG-TERM MEMORY

The remainder of this article focuses on the cognitive processes involved when new information is added to long-term memory and later retrieved. At first thought, libraries and computers might seem useful metaphors for understanding these processes. In computers, for example, files are created (encoding), held on disk (storage), and subsequently made active again (retrieval). Such metaphors, however, can be highly misleading. They suggest that encoding and retrieval are strictly sequential, and that encoding new information does not involve retrieving information that is already stored. With human memory, by contrast, encoding new information *depends* on retrieval of information already in memory. The computer metaphor also suggests that the act of retrieving an item makes that item no more and no less accessible in the future, and does not affect the accessibility of other items. With human memory, by contrast, retrieval renders the retrieved information more accessible in future, and can have either positive or negative effects on the retrievability of other information, depending on circumstances. Furthermore, such metaphors suggest that memories are stored in specific spatial locations, whereas human memories appear not to be stored in specific locations in the brain, but in distributed networks of brain cells (neurons), each of which participates in the storage of many memories.

The key to understanding the unique properties of human long-term memory is to appreciate that it has a dynamic character not shared by current human-made information storage systems. That is, the state of memory is constantly changing as a result of encoding and retrieval processes that are intimately interdependent. This unique character is a product of the properties of the brain as an information-processing device and may reflect the evolution of memory from the perceptual mechanisms of the brain.

The Interdependence of Encoding and Retrieval

Levels of processing, encoding specificity, and resonance

Research from the 1970s onwards has greatly enhanced our understanding of encoding and retrieval processes in long-term memory. One important principle to emerge is that the primary determinant of long-term retention is the level of cognitive processing when new material is encoded – irrespective of intention to learn, or amount of repetition or rehearsal, both of which have little impact on long-term retention. Shallow levels of processing involve attending to the physical characteristics (typically, appearance or sound) of material, whereas deep levels of processing involve attending to the meaning of material, with deep processing usually resulting in superior retention. Both shallow and deep processing involve retrieving pre-existing knowledge about appearance, sound, or meaning; the resulting memory trace is a by-product of the processing involved in retrieving that knowledge. This levels-of-processing principle originated with Fergus Craik and Robert Lockhart.

A second principle to emerge is that when encoded material is later retrieved, the stimuli present in the retrieval environment (retrieval cues) also play an important role in whether that material is retrievable. For successful retrieval, it is critical that the information provided by the retrieval cues matches the information in the memory trace, which will in turn reflect the type of processing engaged at encoding. This encoding specificity principle originated with Endel Tulving and Donald Thomson.

The interaction between level of processing at encoding and the cues present at retrieval is illustrated by the results of an experiment reported by Ronald Fisher and Fergus Craik. They asked people to study pairs of words, with the words in the pairs related either by meaning (e.g., *sleet–hail*) or by

sound (e.g., *pail–hail*). The meaning relationship led to a deep level of processing, whereas the sound relationship led to a shallow level of processing. Later, people's ability to recall the second word from each pair was tested via a cued recall test, in which cues were provided to assist with retrieval. The cues either involved the first word in the pair presented at encoding (e.g., *associated with sleet* and *rhymes with pail*, respectively), a similar word (e.g., *associated with snow* and *rhymes with bail*, respectively), or a different word (e.g., *rhymes with bail* and *associated with snow*, respectively).

When the original first word in the pair was presented as a retrieval cue, there was a substantial recall advantage for deep (meaning) over shallow (sound) processing at encoding (54 percent versus 24 percent), but when the cue was a similar word the advantage for deep processing was reduced (36 percent versus 18 percent), and when the cue was a different word the advantage was almost eliminated (22 percent versus 16 percent). This result demonstrates that both the level of processing at encoding and the cues present at retrieval are critical. If it were simply the case that deep processing produces stronger or longer lasting memory traces than shallow processing, then the advantage for deep processing would be observed regardless of retrieval conditions. And if, on the other hand, the only important factor is the match between retrieval cues and memory traces, then increasing the degree of match should benefit performance regardless of the level of processing at encoding. Instead this result suggests that deep processing produces memory traces containing information that is distinctive in comparison with the information contained in other memory traces. As a consequence, when a retrieval cue matches a trace that resulted from deep processing, only that trace is likely to be activated. By contrast, when a retrieval cue matches a trace that resulted from shallow processing, many other traces become active, because they also contain information matching the retrieval cue. In consequence, there is interference that impairs retrieval.

Combining the principles of levels of processing and encoding specificity leads to a more general principle that can be called the principle of selective resonance. The resonance idea is drawn from physics. A 440 Hz tone emitted near the undamped strings of a piano, for example, will lead to sympathetic resonance in the strings tuned to 440 Hz and, to a lesser extent, in the strings tuned to frequencies that have harmonic relationships to 440 Hz (e.g., 880 Hz, 220 Hz). Retrieval can be thought of as resembling resonance. Memory traces are 'tuned' to specific frequencies, based on the information encoded into them. At retrieval, they 'resonate' to the extent that they share information with retrieval cues. Retrieval succeeds when the resonance is unique to relevant traces, and not shared with irrelevant traces.

The selective resonance idea can be traced to the little-known memory theorist Richard Semon, who coined the term ecphory in 1921 to describe the process whereby memory traces (or engrams) resonate in response to retrieval cues, and to Hedwig Von Restorff, who demonstrated the importance of distinctiveness for memory in 1933. More recently, Roger Ratcliff has demonstrated that the resonance concept forms a realistic basis for formal mathematical models of retrieval. In addition, the principle of selective resonance is a natural property of recent models of memory that postulate networks of interconnected units analogous to networks of neurons in the brain (connectionist models). These models, developed by James A. Anderson, James McClelland and others, solve the conundrum of where memories are stored by postulating that they are stored in a distributed form, as changes in the connections between many units, with each unit participating in the storage of many memories. Finally, processes akin to selective resonance appear to be a fundamental property of neural information processing, starting with perception, where groups of neurons are 'tuned' to respond selectively to particular features of stimuli impinging on the senses. The resonance principle thus suggests that our memory capabilities may have evolved as an extension of our perceptual capabilities, with some of the underlying neural information-processing principles being carried through. (*See* **Human Cognition**; **Connectionism**; **Distributed Representations**; **Encoding and Retrieval, Neural Basis of**; **Learning and Memory, Models of**; **Pattern Vision, Neural Basis of**; **Perceptual Systems: The Visual Model**; **Memory Models**; **Perception: Overview**)

Perceptual memory tests and transfer appropriate processing

Deep processing at encoding is typically superior to shallow processing in supporting long-term retention but – as discussed earlier – such superiority can disappear when the cues at test mismatch those present at encoding. Might there then actually be circumstances in which shallow processing is superior to deep processing? The answer is yes. Tests such as cued recall, recognition memory, in which people are asked to discriminate previously presented material from material not previously

presented, and free recall, in which retrieval is not aided by external cues, show advantages for deep processing because succeeding on such tests requires semantic, meaning-related, processing. These tests are classified as conceptual tests. In special circumstances, however, when a memory test demands retrieval of information concerning the perceptual characteristics of previously encountered information (perceptual tests), the advantage of deep over shallow processing can be reversed. For example, Barry Stein showed that shallow processing at encoding was superior when people later had to recognize whether or not words had been shown to them in a particular configuration of upper and lower case letters. Such reversals can occur because shallow processing results in memory traces that contain more distinctive information about the perceptual – as opposed to semantic – characteristics of studied items than does deep processing.

In the light of such findings, John Bransford, Jeffrey Franks and others have argued that the levels-of-processing principle – which can be taken to imply universally superior retention for deep levels of processing – should be reformulated as the transfer appropriate processing principle. This principle retains Craik and Lockhart's fundamental insight that the content of the memory trace is a by-product of cognitive processing at encoding, but it asserts that the value of a particular level of processing at encoding is relative to the kind of processing that is later required at retrieval. (*See* **Memory: Implicit versus Explicit**)

Retrieval as an encoding event

The fundamental interplay of encoding and retrieval processes is further illustrated by experiments on the impact of retrieval on later memory performance. The generation effect, first systematically explored by Norman Slamecka and Peter Graf, is a particularly good example. Inducing people to actively generate items from semantically related cues (e.g., *horse-c_ _t*, the generated item being *cart*) produces better later memory for those items than does simply reading them (e.g., *horse-cart*). Generating, like deep processing, creates more semantically distinctive memory traces than does reading. Another illustration is the impact of retrieving newly acquired material on the later retrieval of that material. Thomas Landauer, Robert Bjork and others have demonstrated that testing people on newly acquired material – versus providing an additional exposure to the material – can result in superior later recall. Moreover, provided that recall succeeds, the more difficult or involved the recall is, the greater its positive effects on later recall. Thus retrieval modifies the state of memory, acting as an additional encoding event, such that retrieved material is rendered more accessible later. These positive effects of retrieval are known as test effects. (*See* **Memory**; **Learning Aids and Strategies**; **Education, Learning in**)

Environmental, mental state, and temporal context effects

The principle that reinstating the kind of cognitive processing engaged at encoding is critical for later retrieval extends to the influence on cognitive processing of environmental context (e.g., location and other ambient environmental stimuli) and of mental state context (e.g., mood states and drug states). Alan Baddeley, Eric Eich, Steven Smith and others have shown that retrieval is often less successful if these forms of context are changed between encoding and retrieval. However, such context effects are not always observed. They appear to be more likely when retrieval occurs in the absence of explicit cues, as in free recall tests; when people are unable to mentally reinstate the context present at encoding; and when the context either becomes explicitly associated with the to-be-remembered material at encoding or exerts an explicit or an implicit influence on the semantic interpretation of the material.

Changes in context can have positive as well as negative effects, as revealed in research by Smith, Arthur Glenberg, Robert Bjork and others. Material encountered on multiple occasions in different contexts is more retrievable later than is material always encountered in the same context. In addition, material encountered on multiple occasions is much more retrievable later when those occasions are spaced apart in time rather than massed together, a temporal context effect termed the spacing effect. These benefits appear to arise, in part, from a common mechanism: encoding variability. That is, changes in context across encounters vary the kind of cognitive processing involved when the material is encoded. In the case of spaced repetition, such variation occurs as a result of a drift in environmental and mental state context over time – an idea first formalized by William Estes in his 1955 stimulus fluctuation theory. Variable encoding benefits memory because it increases the likelihood that some aspect of the cognitive processing engaged at retrieval will match information in the memory trace.

Another factor in the enhanced retention that results from encoding variability is that retrieval is also an encoding event. On the second and

subsequent encounters with the material, it is necessary – if the repetition of the material is to count as such psychologically – that the material is recognized as having been encountered earlier. Such recognition is more difficult, and thus the retrieval involved more powerful as an encoding event, when context changes across the successive encounters with the material. This retrieval-based explanation of the benefit of encoding variability again illustrates the intimate relationship between encoding and retrieval processes. (*See* **Memory**; **Learning Aids and Strategies**; **Learning, Psychology of**; **Education, Learning in**; **Mathematical Psychology**)

Forgetting, Interference, and Inhibition

Any theory of memory must explain why information is often forgotten over time. The most obvious hypothesis relies on a further metaphor for memory, and the one that is perhaps most intuitive: memory traces are like characters engraved on a stone or wax tablet, or like footprints in the sand, and these imprints weather away over time. Stated more scientifically, the hypothesis is that memory traces have strengths that decay with time. As with the strength interpretation of level-of-processing effects, however, this hypothesis cannot explain forgetting. One illustration of the inadequacy of this strength–decay idea is the recognition failure of recallable words, a phenomenon discovered by Endel Tulving and Donald Thomson. They showed that words (e.g., *queen*) studied in the context of weakly related words (e.g., *woman–queen*) were often forgotten on a recognition test when presented in the context of strongly related words (e.g., *king–queen*). However, these forgotten items were often successfully retrieved on a subsequent cued recall test when the weakly related word presented at encoding (e.g., *woman*) was provided as the cue.

The importance of this finding is that the original recognition failure cannot be attributed to decay of trace strength, because it would then have been impossible for the items to be retrieved on the later – and nominally more difficult – cued recall test. Instead, the recognition failure occurred because the target word was presented in a changed associative context in the recognition test – a powerful demonstration of the encoding specificity principle. John McGeoch was the first to argue, in 1932, that forgetting from long-term memory, rather than being a consequence of the decay of memory traces, reflects an inability to retrieve those traces. He argued that such retrieval failures are caused by (1) changes in associative, environmental, and mental-state context over time that he termed altered stimulating conditions, and (2) interference between competing traces in memory. McGeoch's two-factor theory of forgetting has stood the test of time, and fits well with the resonance conception of retrieval described earlier.

Such interference effects were a major focus of research on human learning carried out in the behaviorist tradition (ca. 1900–1970). Conclusions from this research are (1) that interference is the greater the more the similarity in content between the interfering materials; (2) that new learning interferes with old learning (retroactive interference), but, that as time passes, old learning recovers to interfere with the new learning (proactive interference); and (3) that interference can take the form of competition between the sets of materials, evidenced as a confusion at retrieval about which set of materials is which, or as unlearning of the materials, evidenced as the inability to bring a particular set of materials to mind.

Recent research suggests, in addition, that the process of retrieving information from memory can itself cause forgetting. Successful retrieval, as discussed earlier, makes the retrieved material more accessible later, but at an apparent cost: other material associated to the cues guiding retrieval can become less accessible later. Michael Anderson, Robert Bjork, Elizabeth Bjork, and Barbara Spellman have demonstrated such retrieval-induced forgetting by showing that the repeated retrieval of particular members of a category of previously studied words can inhibit subsequent access to the other nonretrieved members of that category. Such inhibited access is apparently a consequence of the need to suppress – that is, not overtly respond with – those items during the earlier retrievals of the target items.

Retrieval-induced forgetting and related effects point to another dynamic property of memory. In order to avoid catastrophic interference as a result of the large amount of information stored, only a limited portion of the information can be accessible for retrieval at any given time. From an adaptive point of view, the portion of information that is most accessible at any given time should be the portion retrieved in the recent past, because that information is most likely to be relevant in the near future. When forgetting is viewed in this way, it can be seen to be essential to the efficient functioning of memory, and thus far from a negative phenomenon. This adaptive theory of interference, inhibition, and forgetting has been formulated as a new theory of disuse by Robert and Elizabeth

Bjork, integrating and extending ideas put forward by Edward Thorndike in 1914 and William Estes in 1955. (*See* **Learning, Psychology of**; **Rational Models of Cognition**)

THE CONSTRUCTIVE CHARACTER OF LONG-TERM MEMORY

The library and computer storage metaphors for human memory, which are misleading for the reasons suggested earlier, are misleading in another respect. They suggest that the storage capacity of human memory is gradually used up as more material is stored. By contrast, there is no known limit to the human capacity to acquire new information. Indeed, acquiring new information in the form of organized knowledge *creates* further capacity. This ever-expanding capacity reflects constructive processes that are unique to human memory. While these processes have the positive effect of enabling the retention of astounding quantities of information, they can also have serious negative effects, by leading to memory distortions and illusions. Such negative effects show that human memory – in contrast to a videotape recorder (another misleading metaphor) – is not a literal record of previously encountered information. Even vivid memories of personally experienced events can be attributions of currently experienced mental events to the past.

Organization, Chunking, and Expertise

Encoding, as discussed earlier, involves bringing pre-existing knowledge in memory to bear on the interpretation of new information. To understand our unlimited capacity to acquire new information, Fergus Craik has suggested that memory, rather than being thought of as a library, computer, or videotape, should be thought of as a scaffolding. The scaffolding is the organized information in memory, which forms a framework for the interpretation of new information, and which permits new information to be attached. It follows that the more scaffolding there is, the greater the capacity to attach (encode) new information. Such organized information plays an important role in retrieval as well: it permits reconstruction of the likely properties of the material.

There are many experimental demonstrations of the powerful positive effects of organizing new material in terms of existing semantic knowledge. For example, Gordon Bower and his colleagues gave people four opportunities to study and freely recall 112 words drawn from various semantic categories. When the words were presented separated into the semantic categories (e.g., *minerals–metals–alloys: bronze, steel, brass; minerals–stones–precious: sapphire, emerald, diamond*), recall of this very large number of words was almost perfect by the second study and recall attempt, and perfect by the third. By contrast, when the identical 112 words were presented in a randomly intermixed fashion for the same amount of study time, recall reached only around 60 percent by the fourth study and recall attempt.

When confronted with new material to learn, as a student for example, a common difficulty is that the material appears to lack organization, and is therefore meaningless. Effective learning requires abstracting the structure of to-be-learned material. One important component of such abstraction is 'chunking', a term coined by George Miller in 1956. A famous example of the importance of chunking, originating with Karl Lashley in 1951, is the following French sentence: *Pas de lieux Rhône que nous*. Even to a French speaker, this sentence is not memorable, because it is nonsensical. However, sounding the sentence out a few times (with the correct French pronunciation) soon leads to a reorganization of its constituent units that renders it instantly memorable – but as an English sentence. Most apparent examples of learning by rote repetition – for example, learning of scripts by actors, and even of sequences of steps by dancers – actually involve some form of chunking.

Chunking also plays a central role in creating differences in memory ability between experts and novices in a particular field. Adriaan De Groot, for example, showed that expert chess players remembered chess positions much more accurately than novice chess players, but that this advantage was not attributable to the experts having better overall memory capabilities: when the chess pieces were randomly arranged, the memory difference between experts and novices disappeared. Instead, the experts' superior memory depended on their ability to chunk groups of pieces according to their knowledge of positions that might be expected to occur in a game.

Exceptional and apparently astounding memory abilities can also be acquired by developing sophisticated chunking strategies. For example, William Chase and Anders Ericsson trained an individual to recall over 80 sequentially presented random digits, even though he could initially recall only an average seven. His ability reflected strategies for grouping the numbers into meaningful chunks, not a general increase in 'memory power' with training: When the task was switched to remembering random letter sequences, he could once

again recall only around seven. Naturalistic studies of memory experts who perform in public also usually confirm that they have learned sophisticated chunking strategies to perform their apparently photographic feats. Thus, memory is not like a muscle that can be 'strengthened' purely by repetitive exercise. (*See* **Lashley, Karl S.**; **Memory**; **Learning Aids and Strategies**; **Expertise**; **Memory Mnemonics**; **Education, Learning in**; **Learning and Memory, the Ecology of**)

Memory Distortions, Illusions, and Attributions

The inevitable role of existing world knowledge in the encoding and retrieval of new information has negative as well as positive consequences, as first reported by Frederick Bartlett in 1932. The then-prevailing experimental practice, originating with Ebbinghaus, was to attempt to examine memory for new information in isolation from pre-experimental knowledge by asking people to learn simple and often meaningless materials. Bartlett, by contrast, asked people to learn stories, which gave maximum scope for the influence of pre-experimental knowledge to be revealed. In recall of the stories, details were omitted, leaving memory for the gist, or main structural elements. Bartlett also found that new elements were introduced and existing elements distorted in accordance with knowledge – including cultural and social preconceptions – about the kind of events likely to have occurred in the story.

The kinds of general world knowledge that influenced and distorted the recollections of the participants in Bartlett's experiments are termed schemas. More recent experimental research with textual materials, as well as naturalistic research on memory for real-world events, conducted by Gordon Bower, John Bransford, Jeffery Franks, Ulric Neisser and others, has clarified the processes by which schemas produce these distortions and additions. At encoding, the specific information provided is elaborated in terms of schemas (e.g., assumptions and inferences are made), and these elaborations become part of the memory trace for the material. At retrieval, information provided by retrieval cues, by the schemas, and by specific information retrieved, which includes the elaborations made at encoding, combines to produce a reconstruction of the previously encountered information – one that can contain significant distortions. Once again, encoding and retrieval processes are intimately interwoven. (*See* **Bartlett, Frederic Charles**; **Schemas in Psychology**; **Memory Distortions and Forgetting**; **Learning and Memory, the Ecology of**)

Such distortions are also evident when people 'remember' recent well-circumscribed events that in fact never occurred. A striking recent example was reported by Henry Roediger and Kathleen McDermott, who updated an experimental procedure introduced by James Deese in 1959. People studied a number of lists of words, with the words within each list consisting of semantic associates (e.g., *mad*, *fear*, *hate*, *rage*, *temper*, *fury*) of a particular prototype word (e.g., *anger*), which was not itself studied. On a later recognition test, previously studied semantic associates were mixed with the nonstudied prototype words, and with other nonstudied words that were unrelated to the previously studied words (e.g., *bread*). People were well able to reject the unrelated words as not having been studied. But they were just as likely to endorse the nonpresented prototype words as having been studied as they were the actually presented words. Critically, they did not just guess that the nonstudied prototypes could plausibly have been studied. They not only 'recognized' them with high confidence, but also claimed to re-experience vividly the details of their prior occurrence (e.g., what they were thinking at the time). This false memory phenomenon, therefore, is not just a memory distortion, but a memory illusion.

This and other memory illusions illustrate that our understanding – or misunderstanding – of our own memory processes, which is termed metamemory, plays a critical role in long-term memory. There is considerable evidence, for example, that we monitor the fluency with which information is currently processed. When that fluency exceeds the fluency we would expect – based on our knowledge of how fluently that kind of information is normally processed – we face a problem of attribution. Where does that unexpected fluency come from? We could attribute it to current external conditions that make the information especially easy to process, the information being well established in memory, a recent encounter with the information, or some other factor.

These attributions about the source of current processing fluency are adaptive in the sense that they are usually valid. But they can sometimes be mistaken, causing striking memory illusions of various kinds. For example, Larry Jacoby and his colleagues found that increasing the identifiability of sentences spoken against a noise background, by presenting those sentences earlier in a supposedly unrelated task, causes a misattribution of that increased identifiability to a lowered level of the

background noise. And Lynne Reder and her colleagues found that prefamiliarizing key words in general knowledge questions, such as the words *golf* and *par* in the question *What is the term in golf for scoring one under par?*, increases the likelihood that people judge that they know the answer to the questions – but without improving their actual ability to answer them. Similar illusions of knowledge, based on general familiarity with a subject domain, can be evident in students' judgments of their comprehension and future memory of textual material, as shown by Arthur Glenberg, William Epstein and their colleagues.

Finally, with regard to the illusory recognition phenomenon described earlier, Bruce Whittlesea has recently theorized that there is unexpected fluency in the semantic processing of the nonpresented prototypes when they appear on the recognition test, caused by the prior presentation of their associates. This unexpected fluency results in the automatic construction of vivid 'recollections' of earlier encounters with the prototypes. Whether or not this particular theory stands the test of time, such attributional theories of memory are important more generally because they raise the possibility that all memories – whether veridical or illusory – are constructions based on current cognitive processing. (*See* **Memory**; **Learning Aids and Strategies**; **Metacognition**; **False Memory**; **Education, Learning in**)

CONCLUSION

The overall picture that emerges from just over a century of scientific research is that human long-term memory is exceptionally complex and sophisticated: it is varied, dynamic, and constructive, and quite unlike current human-made memory devices in virtually every important respect. The resulting capacities of human long-term memory are stunning, which can make its limitations – in terms of forgetting, distortions, and illusions – seem equally stunning. These limitations, however, are part and parcel of a neural information-processing system that is remarkably well adapted to cope with the demands of living in a constantly changing and ever more complex world.

Further Reading

Anderson JR (2000) *Learning and Memory: An Integrated Approach*. New York, NY: John Wiley.

Baddeley AD (1999) *Essentials of Human Memory*. Hove, UK: Psychology Press.

Bjork EL and Bjork RA (eds) (1996) *Handbook of Perception and Cognition*, vol. 10: *Memory*. New York, NY: Academic Press.

Bjork RA and Bjork EL (1992) A new theory of disuse and an old theory of stimulus fluctuation. In: Healy AF, Kosslyn SM and Shiffrin RM (eds) *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, vol. 2, pp. 35–67. Hillsdale, NJ: Lawrence Erlbaum.

Bjork RA and Richardson-Klavehn A (1989) On the puzzling relationship between environmental context and human memory. In: Izawa C (ed) *Current Issues in Cognitive Processes: The Tulane Flowerree Symposium on Cognition*, pp. 313–344. Hillsdale, NJ: Lawrence Erlbaum.

Buckner RL and Wheeler ME (2001) The cognitive neuroscience of remembering. *Nature Reviews: Neuroscience* **2**: 624–634.

De Groot AD, Gobet F and Jongman RW (1996) *Perception and Memory in Chess: Studies in the Heuristics of the Professional Eye*. Assen, Netherlands: Van Gorcum.

Draaisma D (2000) *Metaphors of Memory: A History of Ideas About the Mind*. New York, NY: Cambridge University Press.

Eichenbaum H and Cohen NJ (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. New York, NY: Oxford University Press.

Foster JK and Jelicic M (eds) (1999) *Memory: Systems, Process, or Function?* Oxford, UK: Oxford University Press.

Higbee KL (2001) *Your Memory: How It Works and How to Improve It*. New York, NY: Marlowe.

Naveh-Benjamin M, Moscovitch M and Roediger HL III (eds) (2001) *Perspectives on Human Memory and Cognitive Aging: Essays in Honor of Fergus Craik*. New York, NY: Psychology Press.

Parkin AJ (1999) *Memory and Amnesia*. Hove, UK: Psychology Press.

Parkin AJ (2000) *Memory: A Guide for Professionals*. New York, NY: John Wiley.

Richardson-Klavehn A and Bjork RA (1988) Measures of memory. *Annual Review of Psychology* **39**: 475–543.

Schacter DL (1996) *Searching for Memory: The Brain, the Mind, and the Past*. New York, NY: Basic Books.

Schacter DL (2001) *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Boston, MA: Houghton Mifflin.

Schacter DL (2001) *Forgotten Ideas, Neglected Pioneers: Richard Semon and the Story of Memory*. Philadelphia, PA: Psychology Press.

Squire LR (ed.) (1992) *Encyclopedia of Learning and Memory*. New York, NY: Macmillan.

Squire LR and Kandel ER (1999) *Memory: From Mind to Molecules*. New York, NY: Scientific American Books.

Tulving E and Craik FIM (eds) (2000) *The Oxford Handbook of Memory*. New York, NY: Oxford University Press.

Whittlesea BWA (1997) Production, evaluation, and preservation of experiences: constructive processing in remembering and performance tasks. In: Medin DL (ed.) *The Psychology of Learning and Motivation*, vol. 37, pp. 211–264. New York, NY: Academic Press.

Wilding J and Valentine E (1997) *Superior Memory*. Hove, UK: Psychology Press.

Mental Disorders, Computational Models of

Intermediate article

Brian F O'Donnell, Indiana University, Bloomington, Indiana, USA
Marcie AB Wilt, Indiana University, Bloomington, Indiana, USA

CONTENTS

Mental disorders frequently disrupt cognition. Computational models are mathematical representations of cognitive or neural processes. These models have been applied to simulate cognitive deficits, to relate cognitive deficits to disturbances of neuroanatomy and neurophysiology, and to make predictions of the effects of brain disturbance on cognitive function.

INTRODUCTION

Psychiatric and neurological disorders are frequently accompanied by specific disturbances of perception, cognition, affect or motor control. In addition, these disorders have been associated with abnormalities of neuromodulation or neural connectivity, or loss of neurons. Computational models have been applied to simulate and predict human cognitive performance. More recently, computational models of neural circuit architecture and function have been used to relate cognitive deficits to neural abnormalities in specific disorders. (*See* **Amnesia**; **Aphasia**; **Dyslexia**; **Memory Consolidation**; **Neurons, Representation in**; **Neurons, Computation in**; **Neural Development**; **Synapse**; **Hebb Synapse: Modeling of Neuronal Selectivity**; **Hebbian Cell Assemblies**; **Levels of Analysis in Neural Modeling**; **Learning and Memory, Models of**; **Attention, Models of**; **Neurotransmitters**; **Neural Degeneration**; **Synaptic Plasticity, Mechanisms of**; **Language Comprehension**; **Memory Models**; **Working Memory**; **Model Fitting**)

Computational models have been instantiated at three general levels in mental disorders. Some models have focused solely on the cognitive level, and do not attempt to relate the architecture of the model to the neural substrates of a disorder (e.g. Carter and Neufeld, 1999). A second type of model has been motivated by the role of specific regions of the brain in cognitive disorders, or the role of a neurotransmitter in specific functions (Cohen and Servan-Schreiber, 1992; Haarmann *et al*., 1997). A third approach uses experimental studies of the properties of neurons and neural circuits in animals to design neural circuit simulations, and then lesion components of this neural model to determine how the lesion affects information processing within the circuit (Grunze *et al*., 1996; Hasselmo and Wyble, 1996). Most models have used neural network models to simulate mental disorders. Because neural network models employ an architecture of parallel, interconnected processing units which resemble the architecture of the nervous system, they may be better suited to modeling brain disorder than serial, symbolic programs with a single central processing unit. The distributed processing units in neural networks allow great flexibility in the level of implementation of the model. For example, processing units can represent the properties of individual neurons within a circuit, or abstract representations such as visual features or words. In this article, neural network models which attempt to capture the cognitive or biological features of several major mental disorders will be reviewed. (*See* **Computational Models: Why Build Them?**; **Production Systems and Rule-based Inference**; **Connectionism**; **Connectionist Implementationalism and Hybrid Systems**; **Backpropagation**; **Working Memory, Computational Models of**; **Neural Behavior: Mathematical Models**; **Semantic Networks**; **Spreading-activation Networks**; **Adaptive Resonance Theory**; **Language Learning, Computational Models of**; **Semantic Memory: Computational Models**)

SCHIZOPHRENIA

Schizophrenia is a mental disorder characterized by cognitive deficits, hallucinations, delusions,

and disturbances of social and emotional behavior. Neuropathology suggests a reduction in neural connectivity, possibly secondary to abnormal synaptic pruning during development. In terms of neurotransmission, there is evidence of abnormal dopaminergic transmission, and of *N*-methyl-D-aspartate (NMDA) receptor function in schizophrenia. Notably, normal adults may develop transient psychotic symptoms similar to schizophrenia as a result of the use of amphetamines (which increase dopaminergic activity) or ketamine (a drug that blocks NMDA receptors). The application of neural network models has provided the first formal insights into the way in which these neural and cognitive features might be related (Grunze *et al*., 1996; Hoffman and McGlashan, 1997). Two models will be discussed here.

NMDA receptor function may be disrupted by schizophrenia. What would be the consequences of such a disturbance within local cortical circuits? At the circuit level, Grunze *et al*. (1996) studied the role of NMDA in modulating local circuit inhibition within the rate hippocampus. These neurophysiological findings were then used to develop a biophysical model to test the functional consequences of NMDA blockade. Grunze *et al*. found that inhibitory neurons in hippocampal circuits were highly sensitive to NMDA antagonists. Inhibitory neurons provide recurrent or feedback inhibition of the activity of excitatory, glutamatergic neurons. If NMDA receptors are dysfunctional in schizophrenia, recurrent inhibition would also be diminished, resulting in hyperexcitability within the circuit. A learning simulation using an auto-associative network showed that such a failure of recurrent inhibition would result in the spread of activation to irrelevant patterns during recall. This inappropriate spread of activation within a set of representations may contribute to such symptoms as 'loose associations' in speech, and disturbance of physiological measures of inhibition, such as P50 sensory gating and prepulse inhibition (Nestor and O'Donnell, 1998).

Dopaminergic neurotransmission is probably affected by schizophrenia as well. Dopaminergic modulation may play a role in prefrontal cortex maintenance of contextual information, and may affect gain in neural circuits. Cohen and Servan-Schreiber (1992) developed a back-propagation model of schizophrenic cognitive dysfunction which attempted to capture these features of schizophrenia, and they used the model to simulate deficits in task performance. The model had input, associative and output modules for learning. A context module affected the flow of information through the associative module. The context module provided functions similar to working memory (maintenance and manipulation of task-relevant information) and selective attention. Cohen and Servan-Schreiber found that reduction of gain in units in the context module resulted in degradation of internal representations in the context module. This may be analogous to the effects of dopamine dysregulation on prefrontal cortex function. The degradation of contextual modulation, in turn, resulted in deficits in network simulations of the Stroop task, the continuous performance test and a lexical disambiguation task. These simulated deficits were similar to those observed experimentally in schizophrenia. For example, individuals with schizophrenia show a bias towards the dominant meaning of a word, regardless of the prior context. Thus when provided with the prior statement 'You can't keep chickens', followed by 'without a *pen*', patients with schizophrenia were more likely to say that *pen* referred to a writing implement rather than to a fenced enclosure. When the function of the context module was degraded by reducing gain of units in the context module, the model showed the same pattern of dominant response errors as subjects with schizophrenia. Figure 1 illustrates the network model for this experiment.

Although the Cohen and Servan-Schreiber model did not attempt to model at the level of neural circuits, it did demonstrate how a model may incorporate pathophysiological data into simulations of cognitive deficits, and it identifies mechanisms which may be responsible

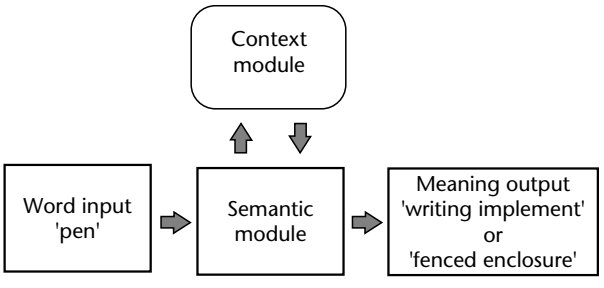


Figure 1. Diagram of connectionist model used in a lexical decision task implemented by Cohen and Servan-Schreiber (2001). The context module allows the system to disambiguate an input which could have either a dominant or a subordinate meaning. Decreasing the gain of the units in the context module produced a bias towards dominant meanings of an input, which was similar to the behavioral performance of patients with schizophrenia in lexical decision tasks.

for behavioral deficits in superficially dissimilar tasks.

HALLUCINATIONS

Auditory hallucinations are associated with a variety of mental disorders, such as affective psychosis and schizophrenia. Hallucinations are often characterized by hearing voices which an individual does not experience as his or her own, and which do not originate from auditory stimulation. In schizophrenia, such psychotic symptoms usually appear in adolescence or young adulthood, coinciding with the developmental interval in which large-scale synaptic pruning occurs in cortical areas. To investigate whether hallucinated speech could arise from pathological over-pruning, Hoffman and McGlashan (1997) developed a back-propagation network model of speech perception. The model had four layers, namely an *input layer* which represented phonetic features of words, a *hidden layer* that integrated feedforward projections from the inputs and recurrent projections from a temporary storage layer, a *storage layer* that represented working memory processes by storing a copy of previous hidden layer activation, and an *output layer* that represented semantic and syntactic features of words. To determine which word (if any) the model detected, output layer activation was passed through a decision algorithm. Before and after pruning, the authors counted the number of correct identifications, misidentifications and hallucinations for full, degraded and null (zero-valued) phonetic inputs. Hallucinations were defined as the presence of a detected word given a null input. Competitive pruning was modeled by modifying weights between the temporary storage and hidden layers.

Hoffman and McGlashan (1997) found that partial pruning of the network, which simulated normal developmental processes, enhanced the ability of the network to detect words. However, over-pruning impaired the network's ability to detect words, and it produced stereotyped hallucinations. Although this simulation did not attempt to model speech perception at the level of neural circuits, it does support the neurodevelopmental usefulness of selective synaptic elimination. Furthermore, the model proposes a plausible developmental mechanism for schizophrenia that is consistent with morphological findings of decreased synaptic density and that predicts perceptual and working memory disturbances which are also found in the disorder.

LANGUAGE DISORDERS

Computational approaches have also been used to model disturbances of language processing caused by stroke. Classic theories of aphasia are framed in terms of the differential role of cortical regions. For example, left temporoparietal lesions often produce Wernicke's aphasia, which is associated with severe language comprehension deficits. Computational models allow a formal representation of semantic and syntactic mechanisms that are affected by aphasia. Haarmann *et al*. (1997) postulated that left temporoparietal lesions cause sentence comprehension problems because they reduce the capacity of working memory for language. The authors tested this hypothesis in a hybrid computational model that combined the spreading activation of connectionist approaches with production rules (if…, then… statements). Their model contained three subsystems of production rules, namely a lexical access system, a parse tree system and a thematic role mapping system (Figure 2). The lexical access subsystem took the perceptual representation of single words as input, and activated word meanings, word syntax and verb structures as outputs. The parse tree subsystem applied grammar-based production rules to the syntactic output of the lexical access system to generate parse tree representations. Finally, the thematic role mapping subsystem combined word meanings and verb structures from the lexical system with parse tree representations to produce thematic role bindings (indicating who did what to whom). The performance of the model was determined by the activations of the thematic role binding units, which represented correct comprehension of parts of sentences. In the model, working memory capacity was defined as the total number of activation units that were available per unit of processing time. Activation spread at each processing step when conditions in the antecedent (if…) of production rules matched active elements in working memory.

In this model, aphasic performance was simulated by decreasing working memory capacity, or the amount of activation available per processing step. In situations where working memory demand (total activation summed across active antecedent conditions and production rule targets) exceeded capacity, target activations were reduced in order to maintain activations within the model's capacity. This modulation of working memory capacity resulted in a pattern of performance that mirrors aphasic deficits. The model demonstrated forgetting (which was instantiated by below-threshold

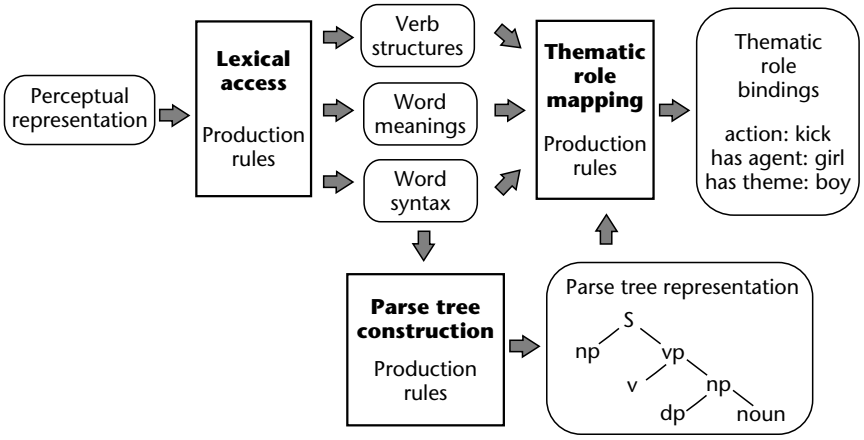


Figure 2. Computational model of semantic and syntactic mechanisms affected by aphasia (adapted from Haarmann *et al*., 1997). In this figure, production rule systems are represented by square boxes and active elements of working memory are denoted by rounded boxes. Different thematic roles include action, agent and theme. The input string 'The girl kicked the boy' activates the thematic role bindings as follows: action, kick; agent, girl; theme, boy. The working memory capacity reductions that are postulated to occur in aphasia are represented in the model by a reduction in the total number of elements that can be active at once.

activations), slower processing (which emerged because more processing steps were required to produce threshold activations) and partial comprehension (which corresponds to occasions when some processing steps never occurred because subthreshold activations did not engage production rules). These deficits closely resemble the patterns of forgetting, slower processing, and partial comprehension that are found in experimental studies of aphasia. The model also captures the interaction between sentence complexity and illness severity which is found in patients. Models with severely decreased working memory capacity performed like patients with severe aphasia – both showed progressively worse sentence comprehension performance with increasing sentence complexity.

ALZHEIMER DISEASE

Alzheimer disease is an adult-onset disorder associated with a progressive dementia. Initially, anterograde amnesia is the most severe deficit, but remote memories are also lost as the disease progresses. Neuropathological changes include loss of neurons and a reduction in cholinergic innervation of the cortex.

Hasselmo and Wyble (1996) used a neurophysiological model of pattern learning to show how cholinergic activity may play a central role in both the memory deficits and neural damage in Alzheimer disease. Within neural circuits, neurophysiological evidence and biophysical models suggest that acetylcholine suppresses association-fiber synaptic transmission, while direct afferent input remains effective. This allows learning of new patterns using the Hebb rule, while suppressing interference and runaway synaptic modification. If cholinergic activity was reduced, as occurs in Alzheimer-type dementia, the metabolic demands of the resulting runaway modification could result in the spread of neuronal damage and degeneration. Moreover, the functional consequences of excessive synaptic modification would include the disruption of learning and retrieval which has been observed in Alzheimer disease.

CONCLUSION

Computational models show promise for simulating experimental data and developing behavioral predictions with regard to cognitive processes in mental disorders. However, current models are limited due to their simplified representation of inputs, outputs and cortical circuits. The human cortex is estimated to contain billions of neurons, and each neuron may have thousands of connections. Current neural models typically use a small number of processing units and connection weights which cannot approach this level of biological complexity. In addition, because models have many parameters which can change as a function of initial weights, connections and learning, it is not clear how the fit of the model to experimental data or neural mechanisms can be rigorously tested. Nevertheless, the computational models

reviewed here often represent the first attempt to develop mathematical simulations of neural and cognitive deficits associated with mental disorders. The full potential of computational modeling of neural and cognitive systems may depend on the development of hardware and software which better represents the architecture of the brain.

References

Carter JR and Neufeld RWJ (1999) Cognitive processing of multidimensional stimuli in schizophrenia: formal modeling of judgement speed and content. *Journal of Abnormal Psychology* **108**: 633–654.

Cohen JD and Servan-Schreiber D (1992) Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review* **99**: 45–77.

Grunze HC, Rainnie DG, Hasselmo ME *et al*. (1996) NMDA-dependent modulation of CA1 local circuit inhibition. *Journal of Neuroscience* **16**: 2034–2043.

Haarmann HJ, Just MA and Carpenter PA (1997) Aphasic sentence comprehension as a resource deficit: a computational approach. *Brain and Language* **59**: 76–120.

Hasselmo ME and Wyble BP (1996) Does the spread of Alzheimer's disease neuropathology involve the mechanisms of consolidation? In: Reggia JA, Ruppin E and Berndt RS (eds) *Neural Modeling of Brain and Cognitive Disorders*, pp. 43–62. London: World Scientific Press.

Hoffman RE and McGlashan TH (1997) Synaptic elimination, neurodevelopment, and the mechanism of hallucinated 'voices' in schizophrenia. *American Journal of Psychiatry* **154**: 1683–1689.

Nestor PG and O'Donnell BF (1998) The mind adrift: attention dysregulation in schizophrenia. In: Parasuraman R (ed.) *The Attentive Brain*, pp. 527–546. Boston, MA: MIT Press.

Further Reading

Grossberg S (2000) The imbalanced brain: from normal behavior to schizophrenia. *Biological Psychiatry* **48**: 81–98.

Plaut DC, McClelland JL, Seidenberg MS and Patterson K (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115.

Reggia JA, Ruppin E and Berndt RS (eds) (1996) *Neural Modeling of Brain and Cognitive Disorders*. London: World Scientific Press.

Mental Models

Intermediate article

William F Brewer, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

CONTENTS

A mental model is a form of mental representation for mechanical–causal domains that affords explanations in these domains. The term 'mental model' is also used to refer to specific mental representations of static spatial domains, and for model-based reasoning about logic problems.

INTRODUCTION

Mental models are a form of knowledge representation. They have been hypothesized to play an important role in our interactions with the physical world. The term 'mental model' and its underlying psychological constructs have wide usage in the disciplines of cognitive psychology, cognitive development, science education, and human factors. There has been some confusion about the nature of mental models and the terms used to describe them. In particular, the term 'mental model' was used differently in two very influential books, both entitled *Mental Models* and both published in the same year (Johnson-Laird, 1983; Gentner and Stevens, 1983). This article will attempt to resolve some of the confusions in this area and to describe the aspects of the mental model construct that have made it valuable to researchers across a wide variety of disciplines.

MODELS IN SCIENCE

The introduction of mental models into psychology was inspired by their use in the physical sciences. The model-based approach to science reached its fullest development in British physics in the late nineteenth century (Klein, 1972). Scientists such as Kelvin, Thomson and Maxwell argued that for a scientific theory to be successful it had to provide a mechanical–causal model. Boltzmann (1902, p. 790) described the model-based approach to science as the belief that 'physical theory is merely a mental construction of mechanical models, the working of which we make plain to ourselves by the analogy of mechanisms we hold in our hands, and which have so much in common with natural phenomena as to help our comprehension of the latter'.

There are several important points to be noted in this description. Models are mental constructs in the minds of scientists. Models make use of mechanical–causal entities. The models in the mind have an analogical relation with the events that are to be explained in the world, and it is this relation that gives them their power. The models provide an understanding and explanation of events in the world. It should be noted that this usage of the term 'model' is quite different from its usage in the term 'mathematical model' in psychology, which involves the use of formal mathematical tools to describe some behavior.

With the advent of relativity theory and quantum mechanics, the model-based approach in physics was replaced with a formalist approach. The ability of physics to develop these very successful abstract approaches may give the impression that model-based approaches are not important in science. However, almost all the major developments in science before the twentieth century were model-based; and, apart from work in theoretical physics, many of the major advances in science in the twentieth century were model-based (e.g. plate tectonics, molecular biology). Indeed, most scientific theories of the natural world except for modern theoretical physics are (mechanical–causal) models.

JOHNSON-LAIRD MODELS

The work of Kenneth Craik formed a crucial link between the construct of models in the physical

sciences and the construct of models in psychology. In a far-sighted book, Craik (1943) used ideas from the physical sciences to develop a model-based approach to the mind. He suggested that human beings construct internal models of reality, and that these models have a 'relation structure' to the world, which allows individuals to make successful predictions about the future occurrence of events in the world by manipulation of the internal model.

Johnson-Laird's work on mental models represents one model-based approach in modern cognitive science. Johnson-Laird developed the construct of mental models as a new form of knowledge representation needed to understand language comprehension and logical reasoning. He stated that 'a model *represents* a state of affairs and accordingly its structure is not arbitrary like that of a propositional representation, but plays a direct representational or analogical role. Its structure mirrors the relevant aspects of the corresponding state of affairs in the world' (Johnson-Laird, 1980, p. 98). This model-based approach was quite different from most other proposals for the representation of knowledge (e.g. semantic features, semantic networks, propositions), in that the analogical assumption imposed constraints on what could be represented and how; so this form of representation was not as arbitrary and unconstrained as some other proposals. (*See* **Knowledge Representation, Psychology of**)

Johnson-Laird's model-based proposals have had a major influence on two areas of research: reasoning and text representation. Johnson-Laird used the idea of models to attack the view that people reason with some form of abstract mental logic. He and others have argued convincingly that people convert abstract logical problems into specific models and then manipulate the information in the models to generate conclusions. (*See* **Reasoning**; **Deductive Reasoning**)

Johnson-Laird's proposals have a number of powerful consequences in the area of mental representation of text information. He argued that many forms of text can be represented in terms of a mental model; and he often used the example of a text describing a spatial array. When Johnson-Laird was developing these ideas, most theories of text representation assumed that the structure of a text could be analyzed in terms of the concatenation of the sentences that made up the text. The mental model approach to texts describing spatial arrays made it obvious that one needed representations both for the linguistic form of the text itself and for the information conveyed by the text. This was a major advance in the study of linguistic discourse and has led to a wide range of experimental studies (e.g. Bower and Morrow, 1990). (*See* **Discourse Processing**)

Another important approach to the representation of text information at the time of the development of the mental model construct was the analysis of information in terms of schemata or generic knowledge structures. The use of schema representations remains a viable approach to the analysis of very stylized genres (e.g. soap operas, fairy tales). However, the mental model approach made clear the need for forms of knowledge representation that could account for the construction of a coherent mental representation from a text describing information that was new to the reader or hearer. Much of the early work on mental models of text focused on the processes involved in the construction of a specific mental model from a text describing an unfamiliar spatial array, with little discussion of the background knowledge that was used in the construction process (Brewer, 1987). (*See* **Schemas in Psychology**)

MENTAL MODELS IN HUMAN FACTORS

In the area of human factors there was a somewhat independent development of a construct of mental models (Wilson and Rutherford, 1989). In the early development of this area of research, theory tended to derive from stimulus–response psychology and the focus was on simple manual tasks. However, as researchers looked more deeply at what was going on in apparently simple tasks, and as they looked at complex process control tasks (e.g. someone controlling the operations of a chemical plant), they began to feel the need for new forms of mental representation. They needed a kind of knowledge structure that would allow them to account for the fact that operators were making (often correct) predictions about the future states of the various physical systems they were controlling. In fact, driven by the nature of the real-world tasks they were trying to understand, researchers in human factors began using the mental model construct 10 or 15 years before researchers in most other areas of cognitive science.

In the early period of the use of the mental model construct in human factors research it was often called an 'internal model', but later the term 'mental model' became standard. Rasmussen (1986, p. 141) provides an interesting definition of mental model as used in human factors research: 'A mental model of a physical environment is a

causal model structured in terms of objects with familiar functional properties. The objects interact in events, i.e., by state changes that propagate through the system.'

Most of the theorizing in this area has focused on mental models in long-term memory (roughly, the operators' theories of the tasks they are performing). However, much of the practical research has involved investigations of particular instantiations of mental models. For example, a researcher in human factors might study someone operating a plastics plant and find that the individual explains his or her actions at a specific moment by stating that he or she has just increased the temperature in one of the reactors because some of the feed material was not of good quality. Thus, human factors data typically relate to the dynamic moment-to-moment instantiation of the underlying mental model from long-term memory.

Because of the nature of human factors research, most of the data and theory in this area relate to the process of human beings interacting with human cultural artifacts. In many cases (e.g. flying airplanes, operating a nuclear power plant) one can think of the domain as having two aspects: the underlying mechanical–causal processes and the intentionally designed artifacts (e.g. control panels). This means that researchers must often consider the operator's mental model, the artifact designer's mental model, and the interaction of the two.

There has been considerable research on mental models for various types of computer-related tasks. It is not clear that mechanical–causal models are a reasonable representation for computers. This suggests either an inappropriate use of the model construct or the need to expand the mental model construct beyond mechanical–causal domains to deal with abstract–causal domains.

Researchers in human factors have carried out studies using techniques such as structured interviews and think-aloud protocols to attempt to describe the mental models used in various types of tasks. There have been two main lines of theoretical and empirical work. First, there is a strong belief that equipment should be designed so that it is consistent with or supports appropriate operator mental models. Second, there have been a number of studies of the role of mental models in instruction. There is considerable evidence that when the operator's task can be carried out in a script-like or procedural way (e.g. 'when the red alarm flashes turn the big switch off'), helping the operator develop an appropriate mental model for the task does not improve performance. However, most researchers believe that mental models improve performance when the operator has to solve a new problem (e.g. troubleshooting), and there is some limited empirical support for this.

MENTAL MODELS IN SCIENCE EDUCATION

The mental model construct was also developed somewhat independently in the area of science education. In the 1960s and 1970s a number of studies suggested that children showed 'misconceptions' in their understanding of certain natural phenomena. School children think that heat and cold move though objects like fluids, that plants obtain their food from the ground, and that boats made of iron should sink. Driver and Easley (1978) synthesized some of these studies and made the point that young children appear to develop 'alternative frameworks' for explaining many phenomena in the course of their interactions with the world and the adult culture. Driver and Easley's theory was loosely based on Thomas Kuhn's approach to the philosophy of science. They suggested that children's alternative frameworks showed how the same phenomenon could be explained by very different conceptual systems. They suggested that the children had simply developed an alternative theory of the phenomenon to that currently held in science.

Researchers in science education have not developed a common term for these types of beliefs held by children. They have been called misconceptions, alternative frameworks, children's science, students' conceptions, naive knowledge, etc. Many of these beliefs involve mechanical–causal explanations of phenomena. Indeed, some researchers have argued (e.g. Andersson, 1986) that many of the children's misconceptions are based on an overapplication of mechanical–causal theories. Thus, it would appear that many of these misconceptions should be considered as children's alternative mental models of the natural world.

Empirical studies in this area have focused on the domains of the natural world that are typically studied in school. The emphasis has mostly been on the mental models that the children have in long-term memory. Many researchers believe that instruction should consist of techniques that will facilitate the replacement of children's alternative mental models with the accepted scientific mental models. In other words, the goal is to have the children understand the world in ways closer to those of current scientific theories. Most investigators feel that knowing the child's current mental model should help in the design of instructional

techniques. However, children's self-generated mental models tend to be very resistant to change with most of the forms of instructional intervention that have been tried.

MENTAL MODELS IN DEVELOPMENTAL PSYCHOLOGY

Within the field of developmental psychology there is a research tradition that focuses on children's understanding of the natural world (Wellman and Gelman, 1992). Much recent work in this area has adopted the view that children are like little scientists and that from their observations of the natural world they develop conceptual understandings of natural phenomena. These mental representations of the natural world are often called 'naive theories'. Some of the research in this area has focused on topics such as children's theories about the minds of other people, and this work does not seem to fit well within the mechanical–causal framework of mental models. Some of the most explicit discussions of mental representation in children occur in this work on theory of mind; but these types of theories do not have an analogical relation to the physical world and do not make use of mechanical–causal mechanisms. Instead, they emphasize the role of abstract theoretical entities. Thus, these naive theories do not seem to be model-based. (*See* **Naive Theories, Development of**)

However, within the area of child development research there are a number of studies of mechanical–causal domains. For example, Vosniadou and Brewer (1994) have investigated young children's understanding of the day–night cycle. They found that children develop a variety of explanations based on mechanical–causal mechanisms, and they use the term 'mental model' to describe these mental representations.

GENTNER–STEVENS MODELS

It is difficult to characterize the similarities in the set of papers that were published in the important volume edited by Gentner and Stevens in 1983. These papers come from quite different research traditions: human factors, science education, and child development. However, they focus on mechanical–causal domains (including both human artifacts and the natural world); and most of them focus on mental representations that are in long-term memory, though often in a problem-solving context that includes a dynamic, constructed representation. This collection of papers is distinguished from most other discussions of mental models in that a number of them attempt to represent the mental models using representation schemes taken from the field of artificial intelligence.

IMAGES AND MENTAL MODELS

Many discussions of mental models describe them as having image properties. Participants in model-based experimental investigations often describe their phenomenal experience while carrying out the task as having an image of the situation. In the human factors literature a number of investigators have stated that individuals are able to solve model-based problems by envisioning a causal sequence 'in their mind's eye'. (*See* **Imagery**)

However, a number of theorists have argued that static images are not rich enough to account for performance on model-based tasks, and that some more dynamic knowledge-based form of representation is needed. For example, Johnson-Laird (1983, p. 157) suggests that images may be views of a richer underlying knowledge structure.

BARSALOU'S PERCEPTUAL SYMBOLS

Barsalou (1999) proposes a new form of knowledge representation which he calls 'perceptual symbols'. These representations have both spatial properties, derived from perception, and symbolic–generative properties, derived from frame representations. Barsalou argues that perceptual symbols are the underlying form of representation for a wide range of forms of knowledge, including very abstract knowledge. Barsalou does not focus on knowledge related to mechanical–causal domains, but his perceptual symbols have properties that make them a very plausible way to think about mental models in such domains.

ANALYSIS OF THE MENTAL MODEL CONSTRUCT

The different approaches to the mental model construct can be clarified by an analysis of mental models in terms of the domain of application and the type of knowledge that is assumed to reside in long-term memory. The classification of mental models by domain of knowledge is necessary because different forms of knowledge have different characteristics (for example in the static spatial domain, causality is not a relevant consideration). Their classification by information in long-term memory is justified by Brewer's (1987) argument that mental models as used by Johnson-Laird are best distinguished from schemata by the fact that

they are assumed to be constructed at the time of input, whereas schema representations are assumed to already be in long-term memory. This distinction will be applied across the different accounts of mental models discussed earlier.

Preformed Spatial Knowledge

Preformed spatial knowledge in long-term memory has sometimes been called a mental map or a spatial schema. When someone is exposed to an appropriate instance (e.g. a particular map of the USA) they use their generic mental map of the USA to form an instantiated location representation (e.g. a specific mental map in which Florida is green).

Constructed Spatial Knowledge

Constructed spatial knowledge in long-term memory consists of general principles of organization in Euclidean space, not generic preformed spatial information. When someone is exposed to an appropriate instance (e.g. a text description of an unfamiliar town) they can use the information in it, along with these general principles, to construct a new specific spatial representation. Johnson-Laird focuses on this process, and he refers to the new spatial representation as a mental model.

Preformed Mechanical–Causal Knowledge

Preformed mechanical–causal knowledge in long-term memory is an understanding of mechanisms (e.g. how filter coffee machines work or how the day–night cycle works). This type of knowledge has been given various names. Within human factors research it is often called a 'mental model'; within science education it has been called by various names including 'misconception' or 'alternative conception'; within developmental psychology it is usually called a 'naive theory'.

When someone is successful in dealing with a specific instance of the general information in long-term memory, then one typically says that the instance has been explained by the mechanical–causal model or theory. This class of phenomena covers much of problem solving in science education and troubleshooting in human factors.

Constructed Mechanical–Causal Knowledge

Constructed mechanical–causal knowledge in long-term memory consists of general principles, not a previously understood model or theory. This type of knowledge has been less well studied, and does not have a generally accepted name. In a wide variety of situations one deals with instances by constructing a new model. For example, this is what happens when one understands a Rube Goldberg cartoon (an unfamiliar and implausible causal chain). The same process may be operative when someone is taught a new theory to explain a particular instance of a phenomenon. In the most creative cases, this is what occurs when a child develops a naive theory or a scientist constructs a new theory to explain a particular mechanical–causal phenomenon.

PSYCHOLOGISTS' REPRESENTATIONS OF MENTAL MODELS

There is a difficult ontological problem in the psychological study of knowledge representation, which is particularly evident in the study of mental models. Are mental models something inside the heads of individuals, or are they a form of representation that psychologists as scientists are using to study individuals? This is a difficult question, and many researchers avoid it. For theorists with a realist view in the philosophy of science there are special difficulties. The problem is particularly acute for those who argue for some form of analogical relation. Is the analogical relation between the representation in the mind of the individual and the world, between the theorists' representation and the mind of the individual, or between all three? In practice, the term 'mental model' is typically used with a systematic ambiguity. For those researchers who are instrumentalists in the philosophy of science there is no problem: for them, mental models are just a construct used to account for the data and there is no issue of what is really in the head.

DEFINITION OF MENTAL MODELS

Analysis of the mental model construct as it is used across a wide range of fields shows that it is a complex construct, but suggests enough agreement to provide a general definition: a mental model is a form of mental representation for mechanical–causal domains that affords explanations for these domains. This type of mental model can be thought of as a subtype of naive theories of the physical world. Mental models contain mental representations of objects in space and the causal relations among the objects. The information in the mental

model has an analogical relation with the external world: the structure of the mental representation corresponds to the structure of the world. This analogical relation allows the mental model to make successful predictions about events in the world. Mental models give rise to the phenomenological experience of mental images, and the process of working out the causal relations of a mental model is often described as 'running the model in the mind's eye'. The term 'mental model' is used to refer both to general model-based knowledge in long-term memory and to temporary specific mental representations constructed in the course of understanding particular events in the world.

The term 'mental model' is also used to refer to mental representations of static spatial domains. In this usage, the term is typically restricted to the specific mental representations that are used to represent unfamiliar spatial arrays. These spatial mental models have an analogical relation with the spatial information in the external world and are typically experienced in terms of visual mental images. This usage of the term is also applied to the case where abstract logical problems are solved by converting them to spatial model-based formats.

References

Andersson B (1986) The experiential gestalt of causation: a common core to pupils' preconceptions in science. *European Journal of Science Education* **8**: 155–171.

Barsalou LW (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* **22**: 577–609.

Boltzmann L (1902) Models. *The New Volumes of the Encyclopaedia Britannica*, 10th edn, vol. xxx, pp. 788–791. London: Adam and Charles Black.

Bower GH and Morrow DG (1990) Mental models in narrative comprehension. *Science* **247**: 44–48.

Brewer WF (1987) Schemas versus mental models in human memory. In: Morris P (ed.) *Modelling Cognition*, pp. 187–197. Chichester, UK: Wiley.

Craik KJW (1943) *The Nature of Explanation*. Cambridge, UK: Cambridge University Press.

Driver R and Easley J (1978) Pupils and paradigms: a review of literature related to concept development in adolescent science students. *Studies in Science Education* **5**: 61–84.

Gentner D and Stevens AL (eds) (1983) *Mental Models*. Hillsdale, NJ: Erlbaum.

Johnson-Laird PN (1980) Mental models in cognitive science. *Cognitive Science* **4**: 71–115.

Johnson-Laird PN (1983) *Mental Models*. Cambridge, MA: Harvard University Press.

Klein MJ (1972) Mechanical explanation at the end of the nineteenth century. *Centaurus* **17**: 58–82.

Rasmussen J (1986) *Information Processing and Human–Machine Interaction*. New York, NY: North-Holland.

Vosniadou S and Brewer WF (1994) Mental models of the day/night cycle. *Cognitive Science* **18**: 123–183.

Wellman HM and Gelman SA (1992) Cognitive development: foundational theories of core domains. *Annual Review of Psychology* **43**: 337–375.

Wilson JR and Rutherford A (1989) Mental models: theory and application in human factors. *Human Factors* **31**: 617–634.

Further Reading

Campbell NR (1957) *Foundations of Science*. New York, NY: Dover. [First published in 1920 as *Physics: The Elements*. Cambridge, UK: Cambridge University Press.]

Driver R, Guesne E and Tiberghien A (eds) (1985) *Children's Ideas in Science*. Milton Keynes, UK: Open University Press.

Hesse MB (1963) *Models and Analogies in Science*. London: Sheed and Ward.

Johnson-Laird PN and Byrne RMJ (1991) *Deduction*. Hillsdale, NJ: Erlbaum.

Leatherdale WH (1974) *The Role of Analogy, Model and Metaphor in Science*. Amsterdam: North-Holland.

Moray N (1997) Models of models of … mental models. In: Sheridan TB and Van Lunteren T (eds) *Perspectives on the Human Controller*, pp. 271–285. Mahwah, NJ: Erlbaum.

Rickheit G and Habel C (eds) (1999) *Mental Models in Discourse Processing and Reasoning*. Amsterdam: North-Holland.

Rogers Y, Rutherford A and Bibby PA (eds) (1992) *Models in the Mind*. London: Academic Press.

Zwaan RA and Radvansky GA (1998) Situation models in language comprehension and memory. *Psychological Bulletin* **123**: 162–185.

Mental Rotation

Intermediate article

Yohtaro Takano, University of Tokyo, Hongo, Bunkyo-ku, Tokyo, Japan
Matia Okubo, University of Tokyo, Hongo, Bunkyo-ku, Tokyo, Japan

CONTENTS

Mental rotation refers to rotational transformation of an object's visual mental image. Mental rotation may indicate that, unlike the object's verbal description, the image shares some spatial and visual properties with the object itself.

INTRODUCTION

When psychological studies of mental imagery revived in the late 1960s, arguments centered upon the question of whether mental imagery should be distinguished from language as a separate system of mental representation. Although mental imagery was repeatedly shown to facilitate word retrieval, this did not convince many psychologists that mental imagery and language are different kinds of representation. Mental rotation attracted great attention because it appeared to attest that mental imagery has some spatial and visual properties that are lacking in language. (*See* **Mental Imagery, Philosophical Issues about**)

RELATION BETWEEN ANGULAR DISPARITY AND RESPONSE TIME

Mental rotation was first reported by Shepard and Metzler (1971). Participants in their experiment were shown a pair of objects like those in Figure 1(a), and asked to judge as quickly and accurately as possible whether they were the same or different. When the two objects were different, they were mirror images of each other as in Figure 1(a). Angular disparity between them was varied from 0 to 180 degrees. Response time was measured between presentation of the objects and a participant's response.

Mean response time for the 'same' pairs was found to be proportional to angular disparity (see 0–180 part of Figure 1(b)). Shepard and Metzler (1971) interpreted this finding as follows: to make mirror image discrimination, participants mentally rotated the image of one object at a constant speed until it is aligned with the other.

This interpretation triggered a lively debate as to whether an image was actually rotated or not. The most convincing evidence of rotation was provided by Cooper (1976), who estimated the speed of mental rotation (i.e. the slope of the linear function) for each participant in an experiment similar to Shepard and Metzler's (1971). In a subsequent experiment, the participant was asked to rotate the image of a presented standard object clockwise. A test object was later presented in the orientation that the standard object's image was expected to take at that very moment. The participant could not predict where the test object would be presented. Nevertheless, the results showed that the participant was able to make mirror image discrimination

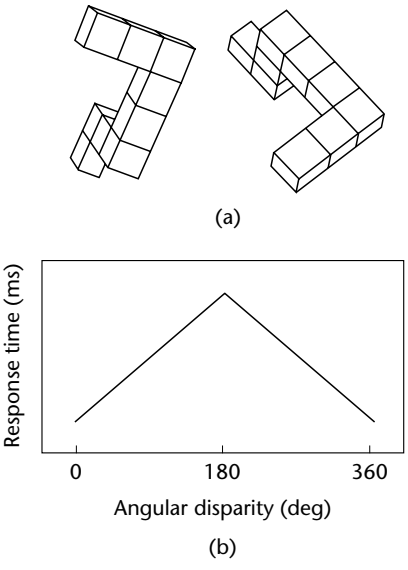


Figure 1. Materials and results in a typical mental rotation experiment: (a) an example of a 'different' pair of objects used by Shepard and Metzler (1971); (b) a schematic illustration of a typical mental rotation function.

at any angular disparity as fast as when the test and standard objects were presented in the same orientation. This strongly suggested that the image was actually present at the expected orientation. It follows that the image was actually rotated at least in that it passed through the intermediate rotational path.

Whether rotation is real or not, the systematic relation between angular disparity and response time provided strong evidence that mental imagery is different from language. In the case of language, the time to transform a sentence, for example 'The object is upright', into the sentence 'The object is tilted by X degrees' does not systematically relate to the value of X.

NON-MONOTONICITIES AT 180 DEGREES

When angular disparity exceeded 180 degrees, it was found that response time did not continue to increase monotonically but turned to decrease up to 360 degrees. In other words, it showed non-monotonicity at 180 degrees. As a result, a typical mental rotation function became an inverted V, as in Figure 1(b). This implies that it was possible to choose the shorter rotational path before starting the rotation.

The basis for identifying the shorter rotational path consists of those structural properties that are not affected by orientational change (Takano, 1989). In the case of the two objects in Figure 1(a), the human visual system detects the following structural properties irrespective of the objects' orientations: each object is composed of cubes, one end is composed of two cubes while the other end has only one, the two ends are orthogonal to each other, and so on. These orientation-free properties can be used to identify the corresponding parts between the two objects. Once the corresponding parts are identified, it is not difficult to determine the shorter rotational path.

INFLUENCES OF COMPLEXITY AND FAMILIARITY

Familiarity

It was found that more familiar objects could be rotated faster. To rotate unfamiliar objects like those in Figure 1(a) by 180 degrees, participants in Shepard and Metzler's (1971) experiment had to spend about 3000 ms. When objects were familiar alphanumeric characters, the time was reduced to 230–1200 ms.

The most dramatic demonstration was provided by Sayeki (1981) who told his participants to regard one of Shepard and Metzler's (1971) objects as a human body. In the case of Figure 1(a), the left object is a sitting person who is extending the left arm, whereas the right object is a person extending the right arm. Once an originally unfamiliar object was likened to a human body, it suddenly became a very familiar one. To rotate it by 180 degrees, Sayeki's (1981) participants needed only 180 ms.

Complexity

Early experiments found no effect of complexity on the speed of mental rotation. It turned out, however, that the reason for this absence was that participants reduced complex objects to simpler properties that were indispensable for making a same/different discrimination. When the task was changed so that the reduction was impossible, the speed of mental rotation was found to be slower for more complex objects.

It was also found that the speed of mentally rotating complex objects became faster as a result of practice (Heil *et al.*, 1998). Given that intensively practised objects become familiar, this effect of practice may be essentially identical to that of familiarity.

Mental rotation seems to have some limitation in its ability to deal with complex structures. Thompson's (1980) 'Thatcher illusion' is its most impressive demonstration: a face with some modifications looks almost normal when it is inverted, whereas it looks monstrous when it is upright. It is impossible to see the monstrous face by mentally rotating the inverted one. This suggests that mental rotation does not have sufficient capacity to transform such a complex structure as a human face while holding its structural details intact.

SITUATIONS THAT REQUIRE MENTAL ROTATION

A Basic Principle

In general, mental rotation occurs when a 'different' object is a mirror image of an original, whereas it does not when a 'different' object is of other types. Takano (1989) explained this by making a distinction between orientation-bound and orientation-free structural properties.

When two figures are mirror images of each other, they differ only in orientational relation between their constituent parts. In Figure 1(a), for

example, the left object's 'arm' extends to the *left* of its 'body', whereas the right object's 'arm' extends to the *right* of its 'body', if both objects are upright. This type of structural property is orientation-bound in that it is altered by a whole figure's orientational change. In Figure 1(a), in fact, the 'arm' of the right object which is not upright extends to the *left*, not *right*, of its 'body'. Therefore, it is impossible to compare orientational relations directly between two objects while they are in different orientations. To compare the orientational relations between the two objects, their orientations have to be aligned first. A common way of alignment is mental rotation.

When two objects are not mirror images of each other, there always exist structural differences that are orientation-free. Imagine, for example, that a non-mirror-image 'different' object is made by adding one more cube to the end of the left object's 'arm' in Figure 1(a). Then, it will be easy to tell whether a test object is the same or different, irrespective of its orientation. Mental rotation is not needed in this case.

Flexible Strategies

The human brain is an extremely complex and thus flexible system. Therefore, whether mental rotation is actually conducted or not sometimes differs from the above basic principle's prediction.

On the one hand, mental rotation could be conducted when it is unnecessary in principle. Suppose that you are looking for a toy. You have to discriminate it from other similar-looking objects, which could appear in any orientation for you. Although they may differ from the toy in some orientation-free properties, you cannot tell beforehand exactly what will be the differences. In such a situation, it may be wise to mentally rotate a similar-looking object to see whether it is the wanted toy. Thus, mental rotation may be conducted in natural settings as a convenient tool even if it is not indispensable. This sort of mental rotation was confirmed in an experiment, where participants failed to detect differences in orientation-free structural properties between compared objects because of their complexity. They did conduct mental rotation, though it was unnecessary in principle (Takano, 1989).

On the other hand, mental rotation may not be conducted when mirror images have to be discriminated. Mental rotation can be skipped because it is not the only way to discriminate between mirror images. If an object's form is remembered in various orientations, mirror image discrimination can be made without mental rotation (Tarr and Pinker, 1989) because the orientational relation in a presented object can be directly compared to that in the original object remembered in the same or neighboring orientation.

Another way of skipping mental rotation is to remember regularities in changing orientational relation. Suppose, for example, that part *A* is to the right of part *B* in an object. When the object is rotated by 90 degrees clockwise, *A* is below *B*; when the object is rotated by 180 degrees, *A* is to the left of *B*. If these regularities are utilized, mirror image discrimination can be made without mental rotation because the orientational relation in a presented object can be compared directly to that in the original object if the latter orientational relation is transformed adequately on the basis of these regularities (Takano, 1989).

THEORIES OF MENTAL ROTATION

Theoretical arguments about mental rotation centered upon the nature of mental representation of visual imagery. In the imagery (or analogue–propositional) debate, which continued for more than ten years, the analogue camp (e.g. Kosslyn, 1973) argued that the mental representation of a visual image is analogous to the corresponding physical object or its physical representation (e.g. a picture) with regard to some spatial and visual properties. The propositional camp (e.g. Pylyshyn, 1973) argued that the mental representation of an image is composed of propositions, a system of symbols that resembles language but is distinguished from it by different vocabulary and grammatical rules.

Mental rotation attracted a great deal of attention as the most convincing evidence for the analogue camp because mental rotation of a visual image closely resembled physical rotation of a physical object. The propositional camp showed that such properties as complexity and familiarity affected the 'speed' of mental rotation, and argued that a visual image differs from its corresponding physical object because a physical object's rotation speed is not affected by its complexity or familiarity. However, this type of criticism failed to show that the mental representation of an image must be propositions, because a representation could still be an analogue while it is different from its corresponding physical object in some respects (e.g., a picture of a person may be different from the real person in its size, material, temperature, and so on).

In the course of the imagery debate, the analogue camp was able to provide rational accounts for

experimental findings presented by the propositional camp; conversely, the propositional camp was also able to provide rational accounts for experimental findings presented by the analogue camp. Anderson (1978) claimed that the debate would not be resolved by any experimental data because any behavior of an analogue system would be mimicked by a propositional system equipped with appropriate processes that deal with propositional representation, and vice versa. This claim can be considered a special case of the Duhem–Quine thesis that any core assumption (e.g. propositional mental representation of imagery) can be made consistent with any empirical data by appropriately modifying peripheral assumptions.

Although the imagery debate thus has not reached any clear-cut resolution, a variety of experimental findings presented during the debate as to mental rotation provide an invaluable empirical basis to infer exactly what spatial and visual properties are manifested by visual mental imagery, whether they are implemented by analogue or by propositional representation.

References

Anderson JR (1978) Arguments concerning representations for mental imagery. *Psychological Review* **85**: 249–277.

Cooper LA (1976) Demonstration of a mental analog of an external rotation. *Perception & Psychophysics* **19**: 296–302.

Heil M, Rösler F, Link M and Bajric J (1998) What is improved if a mental rotation task is repeated – the efficiency of memory access, or the speed of a transformation routine? *Psychological Research* **61**: 99–106.

Kosslyn SM (1973) Scanning visual images: some structural implications. *Perception & Psychophysics* **14**: 90–94.

Pylyshyn ZW (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin* **80**: 1–24.

Sayeki Y (1981) 'Body analogy' and the cognition of rotated figures. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition* **3**: 36–40.

Shepard RN and Metzler J (1971) Mental rotation of three-dimensional objects. *Science* **171**: 701–703.

Takano Y (1989) Perception of rotated forms: a theory of information types. *Cognitive Psychology* **21**: 1–59.

Tarr MJ and Pinker S (1989) Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology* **21**: 233–282.

Thompson P (1980) Margaret Thatcher: a new illusion. *Perception* **9**: 483–484.

Further Reading

Kosslyn SM (1980) *Image and Mind*. Cambridge, MA: Harvard University Press.

Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.

Pylyshyn ZW (1984) *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Shepard RN and Cooper LA (1982) *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.

Tye M (1991) *The Imagery Debate*. Cambridge, MA: MIT Press.

Metacognition

Intermediate article

Petra Scheck, University of Maryland, College Park, Maryland, USA
Thomas O Nelson, University of Maryland, College Park, Maryland, USA

CONTENTS

Metacognition is 'cognition about one's own cognition'. Judgments about past or future memory performance can be examined with respect to their basis, their similarities to other judgments, and their accuracy at predicting memory performance.

INTRODUCTION

Introspective observations have been criticized as being unverifiable, and for a period in the middle of the twentieth century were largely eschewed by the field of psychology. Since the 1970s, experimental paradigms have been developed to examine the accuracy of introspections. Unlike previous introspective methods, these new methods of examining metacognitive accuracy called for specific judgments about performance that could later be compared with criterion measures.

Metacognitive judgments can be tested with regard to their accuracy at predicting memory performance. Research on metacognitive judgments is of interest because it informs us of when the monitoring of memory is accurate and of when the output from monitoring serves as input to control processes (e.g. rehearsal) that affect subsequent memory.

INTROSPECTION

Introspective judgments were investigated in the early years of psychology by researchers such as Wundt and Titchner. Their style of investigation involved observations and reports about tiny 'slices' of consciousness, with extended reports about a single introspective slice.

In the early twentieth century, introspection was attacked on two fronts. Firstly, Freud's theory of the unconscious postulated that not all mental activity was available to introspection, and that unconscious activity affected behavior. Secondly, radical behaviorists such as Watson and Skinner considered introspection to be irrelevant to the understanding of behavior, arguing that there were no reliably valid introspections. Because of these (and other) problems, the study of introspection fell out of favor, and (except in the field of perception) had little influence in psychology until the 1960s.

With the advent of cognitive psychology in the 1960s, introspection was revived. However, instead of assuming that introspections were accurate, cognitive psychologists compared them to relevant behavior. Such comparisons revealed situations in which metacognitive judgments have above-chance accuracy. For instance, investigators became interested in situations in which metacognitive judgments were very accurate (e.g. Nelson and Dunlosky, 1991), as compared to other situations in which they were mostly inaccurate, and an attempt was made to treat those findings as clues about the underlying mechanisms of metacognition.

FEELINGS OF KNOWING

Hart (1965) was one of the first researchers to investigate the accuracy of introspections about human memory. The 'feeling of knowing' (FOK) is an experience in which one has a feeling that a currently unretrieved item is nevertheless in memory. Hart investigated whether FOKs are accurate at predicting subsequent memory performance. He used the recall–judgment–recognition paradigm in which subjects receive a recall test, often consisting of general knowledge questions such as 'What is the capital of Australia?' or 'What star is called the North Star?' For answers recalled incorrectly, or when no answer is

produced, subjects are asked to report their FOK by indicating the likelihood that they would recognize the unrecalled answer.

Hart's experiments showed that a positive FOK was associated with higher probability of recognizing the correct answer than was a negative FOK, indicating that FOK judgments are at least somewhat accurate. Hart showed that subjects should be encouraged to guess on the recall test, even when they do not think they know the answer. This is important because when subjects are unsure of an answer, they may be reluctant to guess. Imagine a subject who retrieves an answer for a particular question, but fails to give the answer because he or she fears that it is incorrect (i.e. has low confidence in it). That item is scored as an omission, and thus will be given an FOK rating and included on the recognition test. The subject will probably give the item a high FOK rating, because although no answer was reported, one did come to mind, indicating that the subject has at least some (perhaps correct) knowledge about the question. Subsequently, if the unreported answer was correct, it will appear on the recognition test, and the subject will almost certainly recognize it as the correct answer. This will exaggerate FOK accuracy, because accuracy depends on the ability of subjects to use FOK judgments to discriminate between items that will be correctly recognized and those that will not. The item in question will probably be recognized, adding to the number of items given high FOKs that were subsequently correctly recognized.

Procedures for Investigating FOK Judgments

General knowledge questions are not the only kind of questions that have been used to study FOKs. Noun–noun paired associates (e.g. 'OCEAN – TREE') have also been used to investigate FOK judgments. In this procedure, subjects study a list of items, and during the recall test are asked to recall the second word when prompted with the first. As with general knowledge questions, subjects are asked to make FOK judgments for items that are not correctly recalled, and then a recognition test occurs for those items.

The difference between general knowledge questions and paired-associate learning is relevant. People are able not only to accurately monitor items that have been stored in memory for a relatively long time, but under some circumstances (Nelson and Narens, 1990) can also monitor items that are recently acquired.

Basis of FOK Judgments

Some researchers have investigated the factors on which FOK judgments are based. For example, Nelson *et al*. (1982) investigated whether the degree of learning affects the magnitude of FOK. Subjects learned paired associates to a criterion of one, two, or four correct recalls, and received a recall test four weeks later. Items not recalled during the test were ranked by the subject according to the likelihood that they would be correctly recognized. The results showed that subjects' FOK ranks increased with degree of learning, indicating that this is one factor on which FOK judgments may be based.

Reder and Ritter (1992) investigated the basis of very rapid FOK judgments that occurred prior to extended attempts at recall. Subjects were asked to solve mathematics problems. Some of the previously solved problems were presented again, while other problems consisted of the components of previously solved problems but with at least one critical change. For example, the subject may have previously solved the problem '6×19'. Subsequently, the 6 and 19 might be presented again, but with a different operand (e.g. '$6 + 19$'). Upon seeing the problem, subjects were asked to make an FOK judgment, estimating whether they could recall the answer or whether they would have to compute it anew. The results showed that higher FOKs were associated with familiar components of a problem even when the operand was changed, indicating that FOK judgments may be based partially on familiarity with the cue rather than on the subject's retrieval of the answer.

It is important to discover the factors that affect FOK accuracy in order to determine whether FOK judgments are monitoring unrecalled answers directly or whether they are monitoring information available from external cues and recalled portions of answers that are diagnostic of subsequent memory performance.

JUDGMENTS OF LEARNING AND ALLOCATION OF STUDY

Students studying for an examination make judgments about whether various facts have been learned sufficiently to be recalled in the exam or whether further study is required. This phenomenon is known as 'judgment of learning' (JOL). Judgment of learning is a kind of metacognitive monitoring process. The students also make decisions about which items to continue studying. This is known as 'allocation of study' and is a kind of metacognitive control process.

Paired Associates and JOL

Participants in a typical JOL experiment learn a list of items. The experimenter is able to control the duration of study, number of repetitions, order of presentation, and other factors that may affect performance. Paired associates are particularly well suited to the study of JOLs because they provide stimuli for a cued-recall test, which is useful for controlling effects caused by the order of recall.

Only the stimulus item of the pair should be present at the time of the judgment. Presence of the partial or entire response word at the time of the judgment has been shown to reduce the accuracy of the JOL in predicting recall (Dunlosky and Nelson, 1992).

Procedures for Investigating JOLs

After studying a given item, subjects may be prompted to make a JOL by giving a percentage confidence judgment that in about ten minutes they will be able to recall the second word of the pair when prompted with the first. Subjects make this judgment, known as an 'individual-item' JOL, for each pair.

An 'aggregate' JOL may also be made for the entire list. Here, subjects are asked to estimate how many of the items they will be able to recall in the test.

Kinds of Accuracy

Item-by-item JOLs can be used to investigate 'relative accuracy', which is the ability of people to distinguish items that will be correctly recalled at test from those that will not. For example, imagine that a person has studied two pairs, 'OCEAN – TREE' and 'DAFFODIL – BLOOD', and then made JOLs for both pairs, assigning the first pair a JOL of 80% and the second pair a JOL of 20%. If at test the person recalls 'TREE' when prompted with 'OCEAN' and fails to recall 'BLOOD' when prompted with 'DAFFODIL', the person can be said to have been accurate insofar as the item that received the greater JOL was also the item that had the better outcome during recall.

Another type of accuracy is 'absolute accuracy', which can be measured both for individual item JOLs and for aggregate JOLs. Absolute accuracy refers to the extent to which the cardinal value of the JOL corresponds to the percentage of correct recall. For example, if the person has studied and made item-by-item JOLs, and then, at test, recalls none of the items that had received a JOL of 0%, 20% of the items that had received a JOL of 20%, and 40% of the items that had received a JOL of 40%, then this person can be said to have perfect absolute accuracy. Likewise, for aggregate JOLs, absolute accuracy is the degree to which the aggregate JOL matches the percentage of recall.

Production of Accurate JOLs

The student studying for an examination will be interested in any strategy that will make JOLs more accurate, in order to know which items need further study and which are already learned well enough to be recalled later. Although individual-item JOLs made immediately after study are generally above-chance at predicting subsequent retention, they are far from perfectly accurate. JOLs have been shown to be very accurate when they are made at least 30 seconds after study (Nelson and Dunlosky, 1991). This is known as the 'delayed-JOL effect'. Although 30 seconds of filled activity is sufficient to produce a substantial increase in JOL accuracy, judgments made after a longer delay (e.g. five minutes) may be even more accurate (Kelemen and Weaver, 1997).

Monitoring and Control Processes

It is important to understand the way in which information acquired during metacognitive monitoring processes is used for metacognitive control. For instance, some researchers have investigated the interplay between JOLs and the allocation of subsequent study. In an experiment by Nelson *et al*. (1994), subjects studied, and made JOLs for, Swahili–English equivalents (e.g. 'ARDHI – SOIL'). Each subject then studied again either the items receiving the highest JOLs (from that subject) or those receiving the lowest JOLs (from that subject). Further study improved multi-trial learning more when it was devoted to the items that had received the lowest JOLs (but for boundary conditions, see Metcalfe and Son, 2000).

RETROSPECTIVE CONFIDENCE JUDGMENTS

Students leaving an examination may think about the questions, judging which of them were answered correctly. Such judgments are called 'retrospective confidence judgments' (RCJs). They may relate to individual items or to the aggregate (for example, an estimate of the percentage of answers that were correct).

Procedures for Investigating RCJs

In the laboratory, the procedures for investigating RCJs are similar to those for investigating other metacognitive judgments. Subjects are shown the cue for the judgment, which consists of a question (in the case of general knowledge questions) or the first word of a pair (in the case of paired associates). Sometimes the subject's answer is displayed along with the cue. Treadwell and Nelson (1996) showed that the accuracy of aggregate RCJs is not affected by the amount of information provided in the prompt for the judgment (cue alone, cue with subject's answer, or cue with answer and the list of choices). This is an important difference from JOLs, whose accuracy is affected by the content of the cue (as discussed above). Individual-item or aggregate judgments are elicited as described previously.

Accuracy of RCJs

Koriat *et al*. (1980) asked subjects to make RCJs for general knowledge questions. They found that when subjects were asked to give reasons for why their answer may be incorrect, the subjects gave more accurate RCJs than when they were asked to give reasons for why their answer was correct or when they were not asked to give any reasons. This finding suggests that subjects may search automatically for reasons for why their answers are correct – a phenomenon known as 'confirmation bias' – whereas searching for reasons why the answer may be wrong is not automatic, and must be done deliberately in order to increase the accuracy of RCJs.

FEELING OF WARMTH IN PROBLEM SOLVING

In an investigation of the feeling-of-warmth phenomenon, Metcalfe and Wiebe (1987) proposed two broad categories of problems, namely incremental problems and insight problems. Incremental problems are solved by degrees, with each step taking the person a little closer to the solution. Metcalfe and Wiebe predicted that people will make incremental feeling-of-warmth judgments for these types of problems, with ratings increasing as steps in the problem are completed. By contrast, insight problems are typically solved suddenly. Feeling-of-warmth judgments for insight problems were predicted to be fairly low and constant until the problem was solved, at which point they would increase suddenly. Metcalfe and Wiebe had subjects solve either incremental or insight problems, making feeling-of-warmth judgments every 15 seconds. The results confirmed the predicted patterns. These results suggest that the feeling of warmth is accurate only for incremental problems, and not for insight problems. With insight problems, a low feeling of warmth does not necessarily indicate that the problem is difficult or even that the person is far from the solution.

SUMMARY

Metacognitive judgments reflect the monitoring of one's own memory when they occur as predictions about subsequent performance on studied items (as in JOLs), as predictions about subsequent retrieval of currently unretrieved items (as in FOKs), as predictions of the accuracy of answers given on a test (as in RCJs), and as predictions of the imminence of solving problems (as in feeling-of-warmth judgments). In some circumstances, as in the case of delayed JOLs, metacognitive judgments are highly accurate; in other circumstances, as in the cases of immediate JOLs and feelings of warmth about insight problems, they are less accurate. It is important to know when metacognitive judgments are accurate, so that we may use them to control our cognitive processing effectively, and to know when they are inaccurate, so that we may discover ways to improve them.

References

Dunlosky J and Nelson TO (1992) Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory and Cognition* **20**: 374–380.

Hart JT (1965) Memory and the feeling-of-knowing experience. *Journal of Educational Psychology* **56**: 208–216.

Keleman WL and Weaver CA (1997) Enhanced memory at delays: why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**: 1394–1409.

Koriat A, Lichtenstein A and Fischhoff B (1980) Reasons for confidence. *Journal of Experimental Psychology: Learning, Memory and Cognition* **6**: 107–118.

Metcalfe J and Son LK (2000) Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory and Cognition* **26**: 204–221.

Metcalfe J and Wiebe D (1987) Intuition in insight and noninsight problem solving. *Memory and Cognition* **15**: 238–246.

Nelson TO and Dunlosky J (1991) When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the 'Delayed-JOL Effect'. *Psychological Science* **2**: 267–270.

Nelson TO, Dunlosky J, Graf A and Narens L (1994) Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science* **5**: 207–213.

Nelson TO, Leonesio RJ, Shimamura AP, Landwehr RF and Narens L (1982) Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition* **8**: 279–288.

Nelson TO and Narens L (1990) Metamemory: a theoretical framework and new findings. *Psychology of Learning and Motivation* **26**: 125–141.

Reder LM and Ritter FE (1992) What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory and Cognition* **18**: 435–451.

Treadwell JR and Nelson TO (1996) Availability of information and the aggregation of confidence in prior decisions. *Organizational, Behavioral and Human Decision Processes* **68**: 13–27.

Further Reading

Hacker DJ and Dunlosky J (1998) *Metacognition in Education Theory and Practice*. Mahwah, NJ: Erlbaum.

Mazzoni G and Nelson TO (eds) (1998) *Metacognition and Cognitive Neuropsychology: Monitoring and Control Processes*. Mahwah, NJ: Erlbaum.

Metcalfe J and Shimamura AP (eds) (1994) *Metacognition: Knowing About Knowing*. Cambridge, MA: MIT Press.

Nelson TO (ed.) (1992) *Metacognition: Core Readings*. Needham Heights, MA: Allyn and Bacon.

Reder LM (1996) *Implicit Memory and Metacognition*. Mahwah, NJ: Erlbaum.

Metaphor Processing, Psychology of

Introductory article

Dedre Gentner, Northwestern University, Evanston, Illinois, USA
Brian Bowdle, Northwestern University, Evanston, Illinois, USA

CONTENTS

A metaphor is a statement that characterizes one thing in terms of another thing, juxtaposing concepts from separate domains of experience. Metaphor can be used to describe abstract or unfamiliar topics, and to express ideas difficult to convey with literal language.

INTRODUCTION

A metaphor is a statement characterizing one thing in terms of another, where the two are normally considered to be unlike: for example, 'Time is a river'. Metaphors involve the juxtaposition of concepts from separate domains of experience; they ask us to think of something in terms of something else that is radically different. Aristotle regarded metaphor as the *master trope* – the figure of speech most associated with poetic genius. Such figurative juxtapositions of concepts can be expressed in a variety of ways.

Metaphors are closely related to similes, which have an explicit comparison term: for example, 'Time is like a river'. In a standard metaphor or simile, the first term ('time') is called the *topic* and the second term ('river') is called the *vehicle*. The interpretation of the metaphor is called the *ground*. In a good metaphor, the interpretation reveals something interesting about the topic, and sometimes about the vehicle as well.

Metaphors serve a number of cognitive and communicative functions. For instance, they can provide a compact and memorable way of expressing ideas that would be difficult to convey with literal language. Metaphors are often used to describe abstract or unfamiliar topics. For example, time (a relatively abstract dimension) is often described using metaphors drawn from space, as in 'The holidays lie before us' or 'Summer is coming fast'. Metaphors are common in literary and poetic contexts. They are also associated with new discoveries in scientific domains, as in the water wave metaphor for light. Metaphors are also common in everyday discourse. Systems of metaphors pervade our language and are often used to discuss abstract ideas. For example, people speak of 'life as a journey' with its 'pitfalls' and 'rough places' and occasional moments of 'coasting'. Cognitive linguists like Lakoff, Turner, and Fauconnier have analyzed systems of metaphors such as 'marriage as a journey' and 'politics as war'. There is evidence that some conceptual metaphoric systems, such as the space–time metaphors noted above, are not just ways of talking, but are also used in thinking.

METAPHOR AND LITERAL SIMILARITY

Literal similarity comparisons differ from metaphors in that, in literal similarity, many or most properties match, whereas in metaphor only a few properties match. As Ortony noted, the matching properties in a metaphor are often far more salient in the vehicle than in the topic. The metaphor acts to highlight otherwise unnoticed properties of the topic. Gentner and her colleagues found that many of these highlighted properties are relational: for example, 'Sermons are sleeping pills' conveys that they both put people to sleep. Because properties of the vehicle are used to illuminate the topic, metaphors are strongly directional. This directionality is a key diagnostic of literal versus metaphorical comparison. Whereas literal comparisons can typically be reversed – for example, 'A sweater is like a jacket/A jacket is like a sweater', a metaphorical comparison cannot – for example, 'Some jobs are jails/Some jails are jobs'.

THEORIES OF METAPHOR COMPREHENSION

A central question in research on metaphor is how metaphors are understood. In the past, metaphor was viewed as a peripheral aspect of communication, secondary in status to literal language. Early models of metaphor comprehension treated metaphors as deviations from proper literal language – as literally false expressions that violate the usual norms of communication. Current models view metaphor more positively, as a normal part of language. However, theories differ in exactly how metaphor is processed.

One long-standing approach is to view metaphor comprehension as property-matching. In this view, metaphors are understood by means of finding common properties, and the interpretation of a metaphor is the set of properties shared by the two terms. For example, 'The road was a silver ribbon' conveys the common property of a long thin silver line. This idea that metaphor comprehension involves a search for commonalities is intuitively appealing and widely accepted. However, it is not the whole story. In general, metaphors also convey new information that can be imported from the vehicle to the topic. For example, the metaphor 'That senator is a puppet' can be used to convey that the senator is being manipulated by someone else. Thus, metaphors do more than highlight existing commonalities – they create new insights about the topic.

Metaphors thus involve both highlighting common information and projecting new information from vehicle to topic. There are two current theories that attempt to explain both these aspects of metaphor: one likens metaphor to analogy, and the other likens metaphor to category inclusion. Taking the analogy view, Gentner and colleagues propose that metaphors are processed by means of the same structure-mapping processes that are used to understand analogies. Analogies are often used to explain or predict the behavior of an unfamiliar complex or abstract system by comparing it to another, better understood system: for example, 'Electricity is like water flow' or 'Poverty is a disease'. Further, the information conveyed by an analogy is typically relational information, rather than simple object properties. For example, the electricity/water flow comparison does not mean that electricity is wet or blue like water, but rather that it obeys the same relational principles: it flows from a high place (high voltage) to a low place (low voltage), it is impeded by obstacles (resistors), and so on.

On this view, metaphors are like analogies. They are comparisons between two situations that highlight common information and invite inferences from the base (the vehicle) to the target (the topic). For example, to understand a metaphor like 'A suburb is a parasite' the hearer first compares the topic and vehicle (base) representations, arriving at a common relational system: for example, 'A suburb uses the resources of a city just as a parasite uses the resources of an organism'. Once this structural match is established, any additional properties connected to the common relational system are projected as possible influences – for instance, the knowledge that parasites can sap the strength of an organism might be transferred to the topic concept, resulting in the inference 'Suburbs can sap the strength of a city'. By mapping the set of relations in the vehicle to the topic, one gains new insight into the topic.

Another prominent approach views metaphors as category statements. In the Attributive Category theory of Glucksberg and his colleagues, metaphors are understood as class inclusion statements. The idea is that in a metaphor one asserts that the topic is a member of the category of which the vehicle is a prototypical member: for example, the metaphor 'A suburb is a parasite' asserts that suburbs can be classified as parasites. Of course, suburbs do not fit the literal meaning of parasite – 'an organism that lives off another organism'. A metaphorical meaning such as 'something that lives off the resources of another entity without recompense' must be invoked or created from the vehicle. By assigning the topic 'suburb' to this metaphorical category, the properties of the metaphorical category derived from the vehicle can be attributed to the topic. On this account, metaphors are processed differently from literal statements. An open question for category theories is what signals the listener to create a metaphorical category instead of using the literal meaning of the vehicle.

PSYCHOLOGICAL EXPERIMENTS ON METAPHOR

Much early metaphor research was devoted to testing the claim that metaphors are deviant forms of language that require extra processing to be understood. According to the deviance view, hearers first try to derive a literal interpretation of

the expression. They then assess whether that interpretation is plausible, given the context. Only if the literal interpretation is anomalous or false does the listener start over again and derive a metaphoric interpretation. One implication of this approach to metaphor comprehension is that, because literal interpretation precedes metaphoric interpretation, metaphors should take longer to process than literal statements. A second implication of deviance models is that, because literal interpretations are taken to be obligatory, metaphoric interpretations should be sought only when literal interpretations are defective. Neither of these predictions has been reliably borne out in empirical studies. Most researchers now believe that the processes involved in comprehending metaphoric language are much the same as those used for literal language.

An influential piece of early research by Glucksberg and his colleagues dealt a conclusive blow to the two-stage deviance view. Their studies provided strong evidence against the view that people first attempt a literal interpretation and resort to metaphorical interpretation only if the literal interpretation is anomalous. Participants were simply asked to make *true or false* judgments. The materials included true category statements (e.g. 'Some birds are robins'), false category statements (e.g. 'Some birds are apples'), and metaphorical statements (e.g. 'Some jobs are jails'). Note that the answer is 'true' only for the first class; the other two are 'false'. The key question concerned how people would process the metaphors. According to deviance theory, people should have been fast to reject metaphors; they simply had to press 'false' as soon as they realized that the literal meaning was false. However, the results showed the reverse. Participants took much longer to reject metaphors than ordinary false statements, suggesting that the metaphorical meaning was noticed early and interfered with participants' ability to classify it as false. This finding dealt a serious blow to the dual-stage theory, for it showed that processing of metaphorical meanings begins *before* the literal judgment has occurred.

More recently, the metaphor interference effect has been used to trace the mechanisms by which metaphor is comprehended. Wolff and Gentner showed that the metaphor interference effect is equally strong for reversed metaphors (e.g. 'Some jails are jobs') as for forward metaphors (e.g. 'Some jobs are jails'). This suggests that metaphor processing begins with a symmetric alignment, as in the structure-mapping model, rather than by a directional projection from the vehicle to the topic.

'DEAD' AND LIVING METAPHORS

Recent evidence suggests an evolution in metaphor processing. Metaphors with novel vehicles are processed as comparisons, whereas conventional metaphors are processed as categorizations. This occurs because initially novel vehicles become conventionalized over time. If a given metaphoric base is used repeatedly in the same way, the abstraction it conveys becomes more and more accessible. Eventually the metaphoric meaning can be stored as a secondary word meaning. For example, 'goldmine' once referred solely to a shaft in the ground from which gold is excavated. But it has taken on a secondary metaphoric meaning – now listed in most dictionaries – as 'anything that is a source of something valuable' (as in 'A garage sale is a goldmine'). At this point the metaphor has a dual representation.

If this process of conventionalization continues, the metaphoric meaning can become quite stable and fixed. For example, the assertion 'My computer is a dog' conveys that the computer is no good, even if both speaker and hearer believe that dogs are loyal, intelligent, and reliable, because 'dog' has a stock metaphoric meaning. At this point the metaphor has become a stock metaphor and lost its early creative potential. Such metaphors are sometimes referred to as 'dead' metaphors.

If the conventionalization process continues still further, the metaphor may even lose its connection to the original literal meaning. For example, the term 'deadline' in the American Civil War meant a line around a prison camp; any prisoner crossing the line was shot. It was then metaphorically extended to a game of marbles, and then further extended from space to time: in newspaper parlance, it meant a time limit after which an article was unacceptable. Eventually, the literal meaning disappeared. The word 'deadline' now retains only its originally metaphorical sense of a time limit. In this way, metaphors can create new meanings.

Further Reading

Black M (1979) More about metaphor. In: Ortony A (ed.) *Metaphor and Thought*, pp. 19–43. Cambridge, UK: Cambridge University Press.

Gentner D (1988) Metaphor as structure mapping: the relational shift. *Child Development* **59**: 47–59.

Gentner D and Bowdle BF (2001) Convention, form, and figurative language processing. *Metaphor and Symbol* **16**(3&4): 223–247.

Gentner D, Bowdle B, Wolff P and Boronat C (2001) Metaphor is like analogy. In: Gentner D, Holyoak KJ and Kokinov BN (eds) *The Analogical Mind: Perspectives*

from Cognitive Science, pp. 199–253. Cambridge, MA: MIT Press.
Gentner D, Holyoak KJ and Kokinov BN (eds) (2001) *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.
Gibbs RW Jr (1994) *The Poetics of Mind: Figurative Thought, Language, and Understanding*. New York, NY: Cambridge University Press.
Glucksberg S, Gildea P and Bookin HB (1982) On understanding nonliteral speech: can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior* **21**(1): 85–98.
Glucksberg S and Keysar B (1990) Understanding metaphorical comparisons: beyond similarity. *Psychological Review* **97**(1): 3–18.
Kittay EF and Lehrer A (1981) Semantic fields and the structure of metaphor. *Studies in Language* **5**: 31–63.
Lakoff G and Johnson M (1980) *Metaphors We Live By*, pp. 3–34. Chicago, IL: University of Chicago Press.
Ortony A (1979) Beyond literal similarity. *Psychological Review* **86**: 161–180.
Ortony A (ed.) (1979) *Metaphor and Thought*. Cambridge, UK: Cambridge University Press.
Wolff P and Gentner D (2000) Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory and Cognition* **26**(2): 529–541.

Model Fitting

Introductory article

In J Myung, Ohio State University, Columbus, Ohio, USA
Mark A Pitt, Ohio State University, Columbus, Ohio, USA

CONTENTS

In model fitting, the computational model under investigation is evaluated for its adequacy in describing the underlying regularities of observed data.

INTRODUCTION

A measure of advancement in cognitive science, or in any other scientific discipline, is the discovery of general laws and principles that govern the cognitive phenomenon of interest, whether it be language comprehension or problem-solving. As these underlying principles are not directly observable, they are formulated in terms of hypotheses. Computational models in cognitive science are formal expressions of such hypotheses. At its most basic level, a computational model consists of a set of assumptions about the structure and functioning of the cognitive processes under investigation. The goal of modeling is to infer the form of this underlying process by assessing how well the model mimics human behavior. This article provides an overview of some of the key challenges in this endeavor.

PARAMETER ESTIMATION

Once a model is specified with its parameters and data have been collected, one can assess the model's goodness of fit (GOF) to the data. Goodness of fit refers to how well the model fits a particular set of observed data, which is often interpreted as a measure of the model's suitability: the better the fit, the more closely the underlying cognitive process is assumed to be approximated. GOF is measured by finding parameter values that provide the 'best' fit to the data in some defined sense. This process is called 'parameter estimation'.

There are two generally accepted methods of parameter estimation: least squares estimation (LSE) and maximum likelihood estimation (MLE). To describe these methods, let $X = \{X_1, X_2,\ldots,X_n\}$ denote a data sample of n observations and $Y_i(w)$ denote a model prediction for each X_i $(i = 1,\ldots,n)$ where $w = (w_1,\ldots,w_k)$ is a parameter vector. In LSE, we seek the parameter values that provide the most accurate prediction of the data, measured in terms of how closely the model fits the data. Formally, the sum of squared errors (SSE) between observations and predictions is minimized:

$$SSE(w) = \sum_{i=1}^{n} (X_i - Y_i(w))^2 \quad (1)$$

In MLE, on the other hand, we seek the parameter values that are most likely to have produced the data. This is obtained by maximizing the likelihood of the observed data sample. For computational ease, the log-likelihood is instead maximized in practice and is given as follows, assuming independent observations:

$$\ln L(w) = \sum_{i=1}^{n} \ln f(X_i|w) \quad (2)$$

In the equation, $L(w)$ is the likelihood function; $f(X_i \mid w)$ is the probability density function for observation X_i; ln denotes the natural logarithm of base e. If X is normally distributed, minimization of LSE is equivalent to maximization of MLE and, therefore, the same parameter values are obtained under either method. Otherwise, the two solutions tend to differ.

Finding the parameters that maximize the log-likelihood or minimize the sum of squared errors often requires use of a nonlinear optimization routine that attempts to find optimal parameters by trial and error. Optimization proceeds over repeated fittings that converge on the best set of parameter values. First, an initial set of parameters

is chosen either at random or by guessing, and the log-likelihood or the sum of squared errors for the chosen parameters is evaluated. On the next iteration, by taking into account the results from the previous iteration, a new set of parameters is obtained by introducing small changes to the initial parameters so that the new parameters are likely to lead to improved performance. Details of how this updating scheme is done define different optimization algorithms, such as the Gauss–Newton method, the Levenberg–Marquardt method, and the simplex method. The refitting process continues until improvements are judged to be negligible on an appropriately predefined criterion.

MODEL EVALUATION

On the face of it, GOF might seem like a good measure of a model's ability to capture the underlying cognitive process. This would be the case if the data reflected only the underlying process. Unfortunately, behavioral data contain random errors ('noise') because mental processes are stochastic in nature. An implication of noise-corrupted data is that a model's best fit to a data sample (the goodness of fit) can be decomposed into two portions, one that represents the model's ability to capture the underlying regularity, and one that arises from the model fitting random noise:

$$\begin{aligned} \textit{Goodness of Fit} = &\textit{Fit (underlying regularity)} \\ &+ \textit{Fit (random noise)} \end{aligned} \quad (3)$$

We are interested in only the first term on the right-hand side of the equation, but the second term is impossible to remove when fitting a single data set. This is because properties of a model can enable it to provide a good fit to the data for reasons that have nothing to do with the model's ability to approximate the underlying regularity but everything to do with fitting random noise. Such properties together refer to a model's 'complexity', and consist of the number of parameters and its functional form. A complex model is one with many parameters that are combined in a highly nonlinear fashion in the model equation. These properties enable it to absorb random error very easily. Consequently, a complex model often fits data better than a model with fewer parameters and a simpler functional form.

The solution to this problem of over-fitting a single data set is to change the model selection measure so that fit to random error is penalized rather than rewarded. The influence of random error can be minimized by measuring a model's fit to multiple data sets generated by the same cognitive process. To fit all data sets well (i.e. generalize), the model must be able to capture only the underlying regularity in the data and not the random noise. Generalizability, then, should be the measure of a model's suitability.

To illustrate, suppose that model M_1 provides 99% of the variance accounted for. In this case, for example, only 50% of the variance represents the model's true ability to capture the underlying process and the rest (49%) may be due to fitting random noise. In model M_2, which accounts for 80% of the variance overall, 70% is traced to the regularity whereas only 10% is traced to random error. M_2 should be selected over M_1 as the former generalizes better than the latter (70% versus 50%). On the other hand, if we were to choose between the two models based on overall goodness of fit, M_1 would be chosen instead, which is undesirable. This is because the improved fit of M_1 over M_2 can be traced to the former accounting for more of the nonsystematic variation in the data than the latter (49% versus 10%), suggesting that M_1 is more complex than M_2. This point is illustrated in Figure 1. Dots represent data points and curves

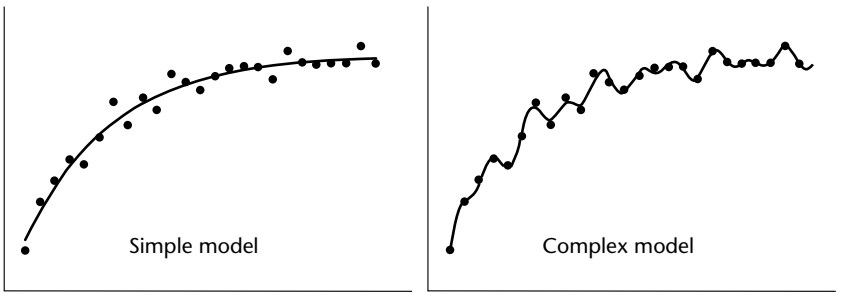


Figure 1. The effect of model complexity on goodness of fit. Two models (lines) were fitted to the same data set (dots). A simple model (left panel) provided a good fit to noisy data, capturing the general trend of the data pattern, whereas a complex model (right panel) over-fitted the same data.

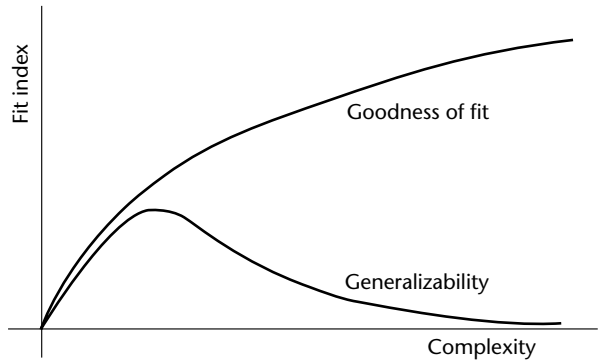


Figure 2. Illustration of the relationship between goodness of fit and generalizability as a function of model complexity.

represent best fits by two hypothetical models. The simple model captures the general trend in the data whereas the complex model enhances goodness of fit by fitting not only the general trend but also noisy patterns in the data.

To summarize, we should assess the adequacy of a model by its generalizability, not goodness of fit, by taking into account the complexity of the model. The intricate relationship among goodness of fit, generalizability, and model complexity is illustrated in Figure 2. Goodness of fit increases positively with complexity. Generalizability also increases with complexity but only up to the point where the model is complex enough to capture the regularities in the data. Beyond this point, generalizability of the model begins to drop as the additional complexity will only enable it to capture random error, but no longer the underlying regularity. An implication is that selecting among models based solely on goodness of fit will result in an overly complex model that generalizes poorly.

MODEL SELECTION METHODS: AIC AND BIC

The central tenet of model selection is achieving good generalizability by finding the best compromise between goodness of fit and complexity by trading off one for the other, thereby formalizing the principle of Occam's razor. As generalizability is not directly observable, it must be inferred from data. The overarching goal of many model selection methods has been the estimation of a model's generalizability. Here we introduce the two most commonly used selection criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In both, goodness of fit is measured by the log-likelihood, $\ln L(w_0)$, where w_0 represents the parameter vector that maximizes $L(w)$. However, they differ in how model complexity is measured: in AIC it is the number of parameters (k), whereas in BIC the sample size (n) is also taken into account. In short, the two criteria are defined as follows:

$$AIC = -2 \ln L(w_0) + 2k \tag{4}$$

$$BIC = -2 \ln L(w_0) + k \ln n \tag{5}$$

For normally distributed data, the first term can be rewritten in terms of the sum of squared errors as $\{n \cdot \ln (SSE\ (w_0)) + \text{constant}\}$. In each criterion, the first (lack of fit) term and the second (complexity) term together represent a measure of lack of generalizability in the sense that the lower the criterion value, the higher the generalizability. Accordingly, the model that minimizes a given criterion should be chosen.

AIC and BIC represent important progress in tackling the model selection problem. In these criteria, however, complexity is defined primarily as the number of parameters in a model. Another dimension of complexity that can also significantly affect model fit is its functional form, which refers to the way in which the parameters are combined in the model equation. For instance, two models, $y = ax + b$ and $y = ax^b$, may not be equally complex though both have the same number of parameters. The minimum description length method described below, which represents an improvement over AIC and BIC, considers functional form as well as the number of parameters.

MINIMUM DESCRIPTION LENGTH

Minimum description length (MDL) originated from coding theory in computer science and regards both model and data as codes. The idea is that any data set can be appropriately encoded with the help of a model, and further, that regularities or patterns in the data imply redundancy. The more we compress the data by extracting this redundancy, the more we uncover the regularities underlying the data. Thus, code length is directly related to the model's generalizability. From this standpoint, the best model is the one that provides the shortest description of the data by maximally compressing it. Formally, the description length of the data is given by the sum of the description length in bits of the data when encoded with the help of the model, $L(D \mid M)$, and the description length in bits of the model itself, $L(M)$: thus $MDL = L(D \mid M) + L(M)$. This definition is broad enough

to be applied for any well-defined models, even qualitative models. For a quantitative model, MDL takes the following form:

$$MDL = -\ln L(w_0) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int dw\sqrt{\det(I(w))} \quad (6)$$

where $I(w)$ is the Fisher information matrix of sample size 1 in statistics and $\det(I)$ is the determinant of the matrix. The first term of the MDL can be interpreted as a measure of lack of fit, and the second and third terms together are a measure of model complexity. Model evaluation in MDL is therefore carried out by trading off lack of fit for complexity.

The second term, $(k/2)\ln(n/2\pi)$, represents the effects of complexity due to the number of parameters k. The third, integral, term captures the effects of complexity due to functional form. To see this, note that the Fisher information matrix $I(w)$ is determined by the log-likelihood function $\{\ln L(w)\}$ which depends upon the form of the model equation; for example, whether $y = ax + b$ or $y = ax^b$. It is also worth noting that the second term increases logarithmically with sample size n whereas the third term is independent of sample size. This means that as sample size increases, the contribution of the effects due to functional form relative to those due to the number of parameters will gradually diminish. Thus, functional form effects will become negligible when sample size is sufficiently large, in which case the MDL is approximately equal to one-half of the BIC.

CONCLUSION

Parameter estimation is a fairly straightforward undertaking, making model fitting easy and accessible to members of the scientific community. The meaning of a good fit is far less straightforward, and can be misleading. A model can fit a data sample well for reasons other than providing a good description of the underlying process, which is why generalizability must instead be used when comparing competing computational models of cognition. MDL is one of the most promising tools to date that measures generalizability.

Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrox BN and Caski F (eds) *Second International Symposium on Information Theory*, pp. 267–281. Budapest, Hungary: Akademiai Kiado.

Kass RE and Raftery AE (1995) Bayes factor. *Journal of the American Statistical Association* **90**: 773–795.

Linhart H and Zucchini W (1986) *Model Selection*. New York, NY: John Wiley.

Myung IJ, Forster MR and Browne MW (eds) (2000) Special issue on model selection. *Journal of Mathematical Psychology* **44**.

Rissanen J (1996) Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* **42**: 40–47.

Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**: 461–464.

Motion Perception, Psychology of

Introductory article

George W Mather, University of Sussex, Brighton, UK

CONTENTS

Motion perception is important for figure–ground segregation, three-dimensional vision, and visual guidance of action. Specialized brain cells detect image motion. Adaptation in these cells leads to illusory motion, such as the motion after-effect.

INTRODUCTION

An essential attribute that distinguishes all animals from plants is their capacity for voluntary movement. Animals move to find mates, shelter, and food, and to avoid being eaten. But the ability to move brings with it the need to sense movement, whether to navigate through the world, or to detect the movement of other mobile animals such as approaching predators. For sighted animals, this means sensing movement in the visual image that is projected into the eye. The image is formed on a sheet of light-sensitive cells that line the inside of the eye – the retina. Specialized neural processes are required to detect the presence of movement in the retinal image.

USES OF MOTION INFORMATION

Surfaces, shapes, and objects in the scene under view create spatial patterns of light and dark in the retinal image. The image is very rarely still, as in a photograph. Instead, it is in a state of continuous change, due to the movement of objects in the scene (e.g. an approaching predator) or to shifts in the position of the observer's eyes, head, or body (e.g. while running away from the predator). Perception of movement in the image is crucial, because it can be used in a number of ways.

Figure–Ground Segregation

Shapes and objects that are invisible while static (e.g. camouflaged animals) are revealed as soon as they move relative to the background. Many animals have evolved special ways of moving, in an attempt to defeat figure–ground segregation. For example, prey animals such as lizards and rodents move in short, rapid bursts in between periods of complete stillness, in order to minimize the chances of detection by predators. Predators such as cats tend to move slowly and smoothly to avoid being seen by their prey.

Extraction of Three-dimensional Structure

When any solid object moves, the images of its various parts that are cast on the retina move relative to each other. Relative motion of this kind can be used to extract the three-dimensional structure of the object. For example, in a sideways view of a rotating globe, surface markings near the equator move across the field of view more rapidly than markings near the poles. In addition, markings near the equator follow almost a linear path, whereas those near the poles follow more elliptical paths. This highly structured variation in speed and direction is sufficient for the perception of the shape's three-dimensional structure.

Visual Guidance of Action

As an observer moves about, image detail 'flows' across the image on his or her retinas, to create a characteristic motion pattern known as 'optic flow'. A great deal of information can be extracted from optic flow, including the speed and direction of self-motion. For example, as you run through a wood, looking ahead, image details arising from rocks on the ground, and from trees ahead, appear near the center of your vision and then flow through your field of view as they disappear behind you, creating an expanding flow field. This pattern of motion flow allows you to navigate

a path through the wood without colliding with trees or misplacing your feet. The powerful movement effects experienced in 'Imax' movie theaters are due to optic flow.

MOTION CUES

The movement of objects under view, or of the observer through the world, cause the spatial pattern of light and dark in the image to fluctuate over time. For example, if you are looking across a relatively dark room, and a friend wearing light clothing moves across your line of sight, then the amount of light falling in the small region of the image at the center of your vision will suddenly increase just as the friend intersects your line of sight, and then decrease again once he has passed through. If the room is empty, and you switch on a light, then again the amount of light falling on the image at the center of your vision will increase. How can the brain distinguish between changes in image intensity due to movement and changes due to other causes, such as changing illumination? In order to solve this problem, the brain must combine information from several places in the image, rather than gathering information from just one place at a time. A change in illumination causes a change in intensity *everywhere* in the image simultaneously, whereas movement causes changes in only a very small part of the image at a time, as Figure 1 demonstrates.

The left-hand and middle panels of Figure 1 show two views of a scene containing a light human figure moving across a dark background. It is difficult to tell what movement has occurred between the two views by inspecting them individually. The right-hand panel shows the changes in light intensity that took place between the two views. Bright areas correspond to places where intensity increased over time from the first view to the second view, and dark areas correspond to places where intensity decreased from the first view to the second view. Gray areas were unchanged between the two views. Notice that the 'difference image' on the right effectively isolates the parts of the scene that contained movement. Stationary features disappear. This would allow the observer to detect the presence of movement, perhaps for figure–ground segregation.

Is it possible to infer the *direction* in which the figure was moving? Some parts of the scene increased in intensity over time (light in the right-hand panel), and other parts of the scene decreased in intensity over time (dark in the right-hand panel). Increases in intensity occurred where a bright edge belonging to the figure moved rightward into a region of the image that was previously dark (e.g. the shin of the leading leg). Decreases occurred where the edge of the figure moved out of a region of the image, returning that region to darkness (e.g. the calf of the leading leg). The brain can therefore infer the direction of a shape's movement by finding its edges, and then detecting whether the intensity of the image increases or decreases over time in the region of these edges.

Since the 1960s it has been known that the brain possesses specialized 'motion-detecting' neurons that respond specifically to movement. Each neuron responds only to movement in a specific direction over a small part of the image. Groups of these first-stage neurons are connected to second-stage neurons that integrate information over relatively large areas of the scene, in order to signal the movement of whole shapes and objects.

Figure 1. Cues for motion detection. The left and middle panels show two views of a human figure moving across the field of view, taken at slightly different times. The right panel shows the changes in light intensity that took place between the two views, created simply by subtracting the light intensities in the first image from the intensities in the second image. The 'difference image' effectively selects only the parts of the scene that moved in between the first view and the second view.

THE MOTION AFTER-EFFECT

The early Greeks were the first to discover a striking visual illusion now known as the motion after-effect (or MAE). The philosopher Aristotle noticed that if he stood in the middle of a river, and directed his gaze down at the fast-flowing water for a short time, when he shifted his gaze towards the riverbank the stationary scene appeared to flow backwards in the opposite direction to the river. This illusion has been rediscovered a number of times, most famously by Thomas Addams, a Scottish scientist. He visited the Falls of Foyers on the banks of Loch Ness, and noticed that if he fixed his gaze on the falling waters for a short time, and then looked at the rock face beside the falls, the rocks appeared to move upwards for a short time. For this reason the effect is also known as the waterfall illusion. It is powerful, robust, and easily demonstrated. A convenient way to experience the illusion today is to view the title credits of a TV program or movie. It is important to fix one's gaze steadily at the center of the screen rather than track the credits as they roll by. After about 30 seconds of adaptation, subsequently viewed scenes should appear to move in the opposite direction to the credits.

The MAE is thought to arise from adaptation in motion-detecting neurons in the brain, of the kind described in the previous section. While viewing an image containing contours moving in a particular direction, cells 'tuned' to respond to that direction will initially respond quite strongly. However, after prolonged exposure their ability to respond is reduced, and takes some time to recover back to normal levels.

Our perception of movement depends on a competition between cells tuned to different directions, rather like a tug-of-war, as shown in Figure 2. Normally, in the presence of a stationary image and without prior exposure to motion, the two opposing teams ('left' and 'right' in the top row of Figure 2) are well matched at a low level of activity, so we see no motion. While viewing a rightward-moving pattern the 'right' team is very active, and easily overcomes the 'left' team to win the competition, leading us to see rightward motion (middle panel). Afterwards, the 'right' team takes some time to recover, allowing the 'left' team to win even while not very active in the presence of a stationary pattern (bottom panel). As a result, illusory motion to the left is seen – the MAE. Once the adapted neurons recover, any bias in favor of one team disappears, so the illusion is no longer seen.

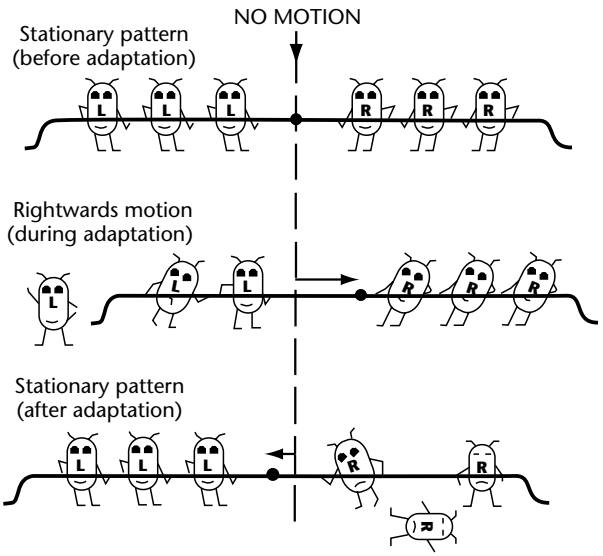


Figure 2. Explanation of the motion after-effect. Two teams of cells signal the direction of movement in the image. One team is active in the presence of leftward motion (L), and the other is active in the presence of rightward motion (R). The direction of perceived movement depends on the result of a tug-of-war competition between the teams. Previous adaptation suppresses the activity of one team, biasing the result of the competition.

Perceptual research on the properties of the MAE indicate that the illusion represents the combined effect of adaptation in at least two populations of cells in the brain, probably corresponding to the first-stage and second-stage motion-detecting neurons described above.

APPARENT MOTION

In natural images of real scenes, moving objects change position in the image in a smooth, continuous manner. If one could inspect the image over shorter and shorter time periods, the shift in position would become smaller and smaller, until at infinitesimal time intervals, the position shift would also be infinitesimal. For example, the left-hand panel of Figure 3 shows a disk drifting to the right over time. Space is plotted horizontally, and time is plotted vertically. The solid outlines represent the positions of the disk at times 1 and 16. The dashed outlines represent the positions of the disk at intermediate times. For obvious reasons, this kind of movement is called 'real movement'.

It is also possible to create the perception of movement in an image by changing the position

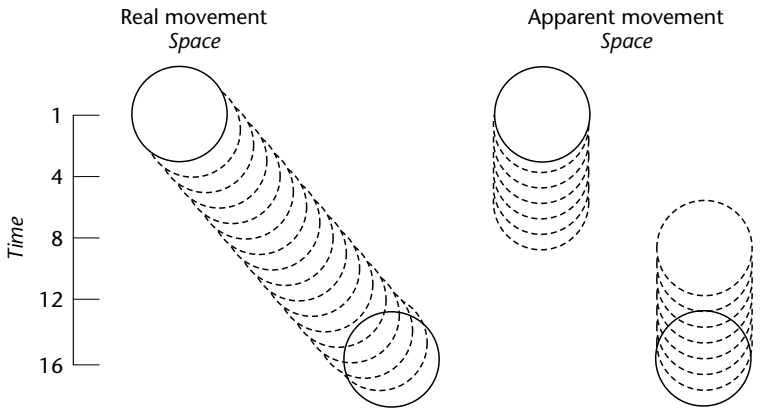


Figure 3. Real movement versus apparent movement. See the text for an explanation.

of an object suddenly over a relatively large distance. If one could inspect the image over sufficiently short time periods, there would be no shift in position. For example, the right-hand panel of Figure 3 shows a disk occupying just two discrete positions. As before, the solid outlines represent the positions of the disk at times 1 and 16. The disk shifts position only once, at time 8. So at all other times the disk remains stationary either at the first position or at the second (dashed outlines).

Observers do perceive movement in stimuli of this kind – it is the basis for the movement seen in TV and movies. TV images are displayed as a series of static images or frames presented very rapidly (50 frames per second in Europe, 60 frames per second in the USA). Any impression of movement seen in TV images is an illusion created by discrete changes in object position from one frame to the next in the display sequence. This kind of illusory perceived movement is called 'apparent movement' or 'phi movement', to distinguish it from the real movement seen in natural images.

Why is the apparent motion so effective and compelling? A common fallacy is that it results from the 'persistence of vision'. According to this explanation, each static image persists in our vision for a short time, so that successively presented static images blend together into one apparently continuous scene. However, it is known that visible persistence lasts only about one-tenth of a second, yet apparent motion can be seen between two stationary shapes even when the second shape appears half a second after the first shape has disappeared. Visible persistence may *contribute* to the perceived smoothness of apparent motion, but it cannot account for the perception of motion itself.

The effectiveness of apparent motion stimuli is almost certainly due to their ability to activate motion-detecting neurons in the brain. As described earlier, motion-detecting neurons rely on the systematic changes in image intensity created by a moving object. Apparent motion stimuli also create systematic changes in image intensity. Provided that the parameters of the apparent motion sequence are chosen carefully, it should excite motion-detecting neurons as effectively as real movement. As one would predict from this explanation, good apparent motion is indistinguishable from real movement. Movie and TV animations do seem very smooth and realistic, unless one sits in the very first row in front of a large movie screen. From this position the discontinuity of the movement, particularly in fast action sequences, can be seen easily.

However, responses in motion-detecting neurons are not the only explanation for apparent movement. It has also been argued that we can perceive motion independently of activity in neural detectors, as a result of perceptual inferences or of shifts in attention. According to the perceptual inference theory, apparent motion is the outcome of a perceptual inference to explain the otherwise mysteriously sudden appearance and disappearance of shapes in apparent motion displays. According to the attention-shift theory, apparent motion can also be perceived when mobile shapes or objects in the image capture and hold the attention of the viewer. As the objects change position in the image, one's focus of attention shifts to keep track of them. This shift in attention itself gives rise to the perception of apparent motion. Advocates of such high-level processes do not see them as inconsistent with the notion of lower-level neural motion detection, but rather as separate processes that co-exist with low-level detection. It therefore seems likely that the perception

of apparent movement is mediated both by low-level processes (motion-detecting neurons) and by high-level processes (inference and attention).

INDUCED MOTION

On a cloudy, moonlit night, as relatively large clouds move quickly across the face of the moon, it often seems as if the clouds are stationary but the moon is moving. This illusion is an example of 'induced motion': the appearance of motion in a stationary stimulus induced by the physical movement of another stimulus. Generally, induced motion is most effective when a large, slowly moving shape surrounds a smaller, stationary shape, such as the moon surrounded by clouds.

The simplest demonstration of induced motion consists of a small spot surrounded by a large frame, as shown in Figure 4. If the frame moves slowly sideways while the spot remains stationary (left-hand panel), observers tend to perceive the frame as stationary and the spot as moving (right-hand panel).

Induced motion highlights the problem of motion attribution. As mentioned earlier, movement in the visual image is detected by specialized motion-detecting neurons. Image motion can arise from two general sources, either movement of objects in the scene under view, or movement of the observer's body. It is obviously crucial to attribute motion to the correct source, but responses in motion-detecting neurons cannot distinguish between them.

The brain appears to use several strategies to solve the problem of motion attribution. These strategies are usually sufficient to arrive at the correct interpretation, but in certain situations the interpretation may be erroneous, leading to illusions such as induced motion.

One strategy is to make use of nonvisual information in order to determine whether the observer is moving through the scene. Movement of the eye, head, and body can be established using information from the muscles (or from commands to move the muscles), and from the balance (vestibular) sense. When movement in the image can be accounted for entirely by bodily movement, then no motion is perceived. For example, eye movements

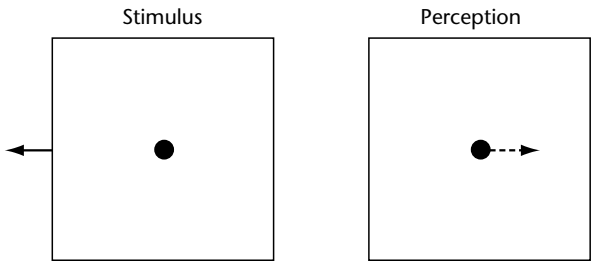


Figure 4. Induced motion. The physical stimulus consists of a small stationary spot surrounded by a slowly moving frame (left). Perceptually (right), the frame appears to be stationary while the spot appears to move.

create movement in the image. When the eyes turn to the left, the whole image translates to the right on the retina. This translation excites motion-detecting neurons in the brain, yet we do not perceive the world to move. The image motion that results from eye movement is correctly attributed to the eye movement, so no motion is perceived.

A second strategy, or heuristic, used in motion attribution relies on assumptions about the nature of the real world. In general, relatively large shapes and objects in a viewed scene tend to remain fixed in position, while small shapes and objects are likely to move. Large areas in a viewed scene may be filled by, for example, the wall of a building or the side of a hill. These shapes are extremely unlikely to move. Small areas, perhaps representing human or animal figures, or vehicles, are very likely to move. Consequently if there is relative movement between a small object and a large object, then the brain has a tendency to attribute the motion to the small object.

Further Reading

Mather G, Verstraten F and Anstis S (eds) (1998) *The Motion Aftereffect: A Modern Perspective*. Cambridge, MA: MIT Press.

Palmer SE (1999) *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Smith AT and Snowden RJ (eds) (1994) *Visual Detection of Motion*. London, UK: Academic Press.

Watanabe T (ed.) (1998) *High-level Motion Processing*. Cambridge, MA: MIT Press.

Motivation

Introductory article

Nira Liberman, Tel Aviv University, Tel Aviv, Israel

CONTENTS

The study of human motivation focuses on the objects of human pursuit, and how pursuit of these objects is structured and organized.

INTRODUCTION

The root of the word motivation is 'to move', and the study of human motivation is about what moves people. Two main questions have been addressed: what are the things that people typically consider worthy of pursuit; and what are the laws that describe the strength of motivation and the organization of goal pursuit? Motivation is a central research area both in basic psychological science and in applied disciplines that face a need to motivate people towards specific behaviors, such as the psychology of education, which is concerned with motivation towards acquiring knowledge and norms of behavior; the psychology of sales and marketing, which is concerned with motivation to purchase different products; and the psychology of human resources in organizations, which is concerned with work motivation.

DRIVES AND NEEDS

People are motivated to seek pleasure and to avoid pain. But what are the things that people find pleasurable and painful? The great variety of pleasurable and painful objects has been classified into a number of content categories, which correspond to different views of human nature.

Types of Need

Drives

Obviously, people, like other animals, seek to satisfy physical needs, s uch as the needs for food, water, sleep and sex. These motives, often called 'drives', are instigated periodically as a result of deprivation, and may be fulfilled by concrete entities (e.g. food, drinks). The need to relieve physical pain, although not instigated by deprivation, is also considered a drive. The study of drives, which has been conducted primarily with animals, dominated the early stages of motivational science.

It has been found, for example, that the intensity of a drive depends on the extent of deprivation, and that the intensity of the motivation to seek a drive-reducing stimulus (e.g. food) depends on the intensity of the drive (e.g. hunger), on the strength of prior learning (e.g. how strongly a given situation is associated with the possibility of obtaining food), and on the expected value of the outcome (e.g. how much food one expects to get). Drives have often been regarded as a general state of agitation, which motivates the organism to 'do something'. The drive-oriented study of motivation emphasizes the animal nature of humans.

The need for stimulation was added to the category of drives after studies on sensory deprivation dramatically demonstrated that both humans and animals seek sensual stimulation (e.g. sounds, sights, touch) even if it is not associated with any tangible reward (e.g. food). In these studies, participants are deprived of sensory input by being placed in an isolated quiet room wearing goggles that permit only homogeneous light. Very few people are able to stay in this state for longer than eight hours, even for a high payment. In one study, an audio recording of a stock market record was made available to participants. The intensity of the need for stimulation was revealed by the fact that participants requested to hear this quite uninteresting recording again and again.

Social needs

The need to be accepted within a social group, to form warm, continuous and rewarding relationships with friends and family, to love and to be loved are social needs. Subsumed within this general category are the motivations to make a favorable impression on others, to attract intimate

partners, to gain high social status, to avoid conflict with others, to distinguish one's own group from other groups, and sometimes to disparage other groups and their members. The study of social motivation suggests a view of people as social beings.

Mastery needs

The desire to control and predict oneself and one's environment, both social and physical, has been termed a need for 'mastery'. The motivations to uncover the truth, to enhance one's abilities and to manipulate the environment are subsumed under this general category. Mastery needs underlie scientific activity and activity that is directed towards technological advancement. Therefore, the study of mastery needs suggests an image of people as lay scientists.

Self-enhancement needs

People are motivated to see themselves and anything connected to them (e.g. the social groups they belong to, their possessions) in a favorable way. For example, people prefer positive feedback to negative feedback of equal informative value. The need to maintain a positive view is often at odds with the motive to have accurate knowledge, inasmuch as people often face unpleasant self-relevant information. When the self-enhancement motive prevails over the mastery motive, then people deny, avoid or seek to disqualify negative self-relevant information. For example, people tend to be more critical of the validity of ability tests on which they scored low than of tests on which they received a high score. The study of self-enhancement needs suggests an image of people as interested lawyers.

Approach and Avoidance Motivation

Within each of the needs specified above, approach motivations may be distinguished from avoidance motivations. Approach motivations are directed towards desirable outcomes (e.g. food, success), while avoidance motivations are directed away from undesirable outcomes (e.g. danger, pain). For example, a person may strive to succeed in an examination, or, alternatively, strive to avoid failure in an examination. Situations that involve advancement and improvement needs and maximal goals (i.e. a goal to achieve the best possible outcome) usually activate the approach system (e.g. looking for a luxurious car). On the other hand, safety and security needs and minimal goals (i.e. a goal to satisfy a minimal acceptable level) usually activate the avoidance system (e.g. seeking to move out of a dangerous neighborhood). Approach and avoidance create different types of positive and negative effect. Specifically, successful approach of a desirable outcome (e.g. finding a luxurious car) produces joy whereas successful avoidance of an undesirable outcome (e.g. moving out of danger) produces relaxation. A failure to approach produces disappointment, whereas a failure to avoid produces anxiety.

Individuals differ in their generalized tendency to approach or avoidance. Thus, the same situation may be conceived in terms of approach by some individuals (e.g. trying to get grade B or higher on an exam) and in terms of avoidance by others (e.g. trying to avoid getting grade C or lower). Because of the different types of emotions characteristic of approach and avoidance, these two personality types are associated with different types of pathologies. Depression, being the result of continuous failure of the approach system, would be characteristic of approach-oriented individuals. Anxiety, being the result of continuous failure in the avoidance system, would be characteristic of avoidance-oriented individuals.

Conscious and Unconscious Motivation

People are not always aware of, and are sometimes wrong about, the motivations that guide their behavior. The notion of unconscious motivation is of central importance in the psychoanalytic tradition, according to which people are guided by repressed sexual and aggressive motives. More modern notions of unconscious motivations are not restricted to these two motives: people's ability to introspect on their own psychological processes, including motivation, is limited and fallible, and any motivation can operate outside conscious awareness.

For example, studies on social influence often find effects of which the participants are unaware. In such studies, effects would be observed (e.g. participants would comply with a request to sign a petition more if it is presented by an attractive person than if it is presented by an unattractive person), which participants would not admit to nor be aware of (e.g. when asked to explain, either in private or in public, how they decided whether to sign the petition, participants would not mention the attractiveness of the person making the request). In this example, participants are unaware of the fact that some of their motivation in complying with the request is interpersonal (e.g. to win the

approval of, or make a favorable impression on, an attractive individual). Similarly, people tend to be unaware of (and ready to deny) the effects of self-enhancement needs on their behavior.

The existence of unconscious motivations implies that researchers of motivation cannot always rely on introspective self-reports, however confidential, but rather have to apply more indirect methods of study.

The Origin of Needs

There is little doubt that physiological, social, mastery and self-enhancement needs exist in humans. People spend a lot of money, time and energy to satisfy these needs. The origin of these needs is, however, controversial. A number of theoretical approaches have attempted to reduce the various needs to one or two basic motives.

Evolutionary theory postulates a single basic motive – spreading one's genes – and derives from it all other motives. According to evolutionary theory, people seek to attract opposite-sex partners in order to reproduce and thereby spread their genes. People seek to succeed (e.g. accumulate wealth) and to gain social approval in order to enhance their offspring's chances of survival. Even personal survival, in this view, is sought for the purpose of promoting reproduction and enhancing the chances that one's offspring would survive and further reproduce.

Freud's psychoanalytic approach reduced all human motives to sexual and aggressive drives. A somewhat related but more modern theoretical approach maintains that at the root of all human motives is the need to overcome the fear of death. This theory suggests that people strive to forget the fact of their inevitable physical demise and to extend their symbolic presence beyond it. In this view, scientific, artistic and religious activities are best understood as attempts to symbolically extend one's existence beyond physical death. It has been proposed also that one's culture or social group serve a similar function (i.e. provide symbolic existence beyond physical death), and therefore the motivation to be a valuable member of a valuable community are also explained as ways to overcome the terror of physical death.

The Relative Importance of Needs

The relative importance of the various needs has also been a subject of theoretical controversy. For example, some theorists contend that self-enhancement needs are subordinated to social needs, so that people strive to form positive views of themselves in order to be able to form more rewarding social relations. Other theorists hold the reverse view, namely, that social needs are subordinated to self-enhancement needs, so that people's ultimate goal in forming social relations is to feel good about themselves.

The relative strengths of different needs varies between individuals. For example, some individuals tend to assign more importance to social needs while other individuals tend to assign more importance to mastery needs. It was once assumed that women tend to have weaker mastery needs than men, but this assumption has been criticized on methodological grounds and is no longer believed to be true in general.

The relative strengths of the motives also varies between cultures. For example, it has been suggested that in collectivist cultures, such as exist in many communities in east Asia and South America, social needs are stronger and mastery needs are weaker than in Western, individualistic cultures. For example, a conflict between personal benefit and the benefit of one's group is more likely to be resolved in favor of the former motive in individualistic cultures than in collectivist cultures.

THE EXPECTANCY–VALUE MODEL

Motivation may be conceptualized as a force having direction and strength. Motivation is directed towards those things that people find worthy of pursuit (e.g. food, social approval, knowledge) or away from those things that they find aversive (e.g. pain, failure). But what determines the strength of motivation? In other words, what determines the amount of effort that people would be willing to invest in pursuing an outcome? For example, what would determine the intensity of a student's efforts in preparing for an examination? Several influential theoretical frameworks represent motivational strength as the product of the value of the goal towards which motivation is directed and the expectancy of achieving it:

$$\text{motivation} = \text{expectancy} \times \text{value} \quad (1)$$

In this equation, 'value' is the desirability of the goal (e.g. how important it is for the student to do well on the examination) and 'expectancy' is the goal's attainability. The multiplicative relation between value and expectancy means that to maintain motivation, both expectancy and value have to be nonzero. If expectancy approaches zero (e.g. if the goal is judged to be unattainable), the goal will

not be pursued, no matter how attractive it is. If value approaches zero (i.e. the goal holds very little attraction), it will not be pursued, no matter how easy it is to achieve.

Expectancy

Various constructs have been proposed for expectancy, most notably probability and controllability or efficacy (the extent to which increasing one's efforts increases the likelihood of attaining the goal or the magnitude of the attained outcome).

Probability is the expectancy of outcomes that are determined by chance (e.g. lottery draws) and of outcomes over which one has little or no control (e.g. getting flu). For example, one's motivation to buy a lottery ticket would increase as the value of the prize increases and as one's estimated chances of winning increase; and one's motivation to get a flu vaccination would increase the more one dislikes having flu and the higher one estimates one's chances of getting flu.

Controllability and efficacy refer to the belief that increasing one's motivation would result in a corresponding increase in either the likelihood of attaining the goal or in its magnitude: in other words, to the belief in the usefulness of one's efforts. For example, the motivation to study for an examination would increase the more important the examination is and the more one believes that studying increases one's chances of success. The motivation to adhere to a diet would increase the more important it is for one to lose weight and the more weight one expects to lose for each 'unit of dieting effort' (i.e. the more effective one believes the diet to be).

The expectancy–value model is widely used in various kinds of decision-making, and has the status of a normative theory, that is, a theory of how decisions should be made in order to achieve optimal outcomes. For example, in comparing alternative financial investments, a decision maker is advised to multiply the expected revenue of each alternative (e.g. in decisions on investing in developing new products, an estimation of how profitable would be each product) by the estimated likelihood of that revenue (how likely is successful development of each product), compare the product across alternatives, and pursue the alternative with the highest product. In comparing alternative treatments for a disease, the decision maker (in this case the physician) is advised to multiply the desirability of the outcome of each alternative treatment by the likelihood of the outcome, and pursue the alternative that yields the highest product. In many real-life situations, the alternatives have multiple possible outcomes (and corresponding probabilities), so the computations may be fairly complex. Computer programs may be used to support experts making decisions.

Value

Value is comprised of costs and benefits. For example, the value of installing an air conditioner in one's car may be thought of as the difference between the benefits and the costs associated with buying and maintaining it (e.g. more pleasant driving versus increased monetary costs and fuel consumption). Costs may refer to depletion of resources that could be used towards other ends (e.g. money, time, energy), or to negative aspects of outcomes (e.g. the noise of the air conditioner). Benefits (positive) and costs (negative) are multiplied by their respective expectancies (e.g. probabilities), and these products are added together to determine the overall motivation. If the sum of the products is positive, the person would be motivated to approach the outcome. If the sum of the products is negative, the person would be motivated to avoid the outcome.

The expectancy–value model of motivation has a number of interesting implications for both motivation and personality psychology. We will briefly examine three such implications below: probability-dependent value; distance-dependent expectancy and value; and personal beliefs in expectancy.

Probability-dependent Value

Sometimes the value of an outcome depends on the probability of achieving it. For example, doing well on a difficult examination is both more advantageous and less likely than doing well on an easy examination. We can model such situations by postulating that value $= k\,(1-p)$, where the expectancy p is between 0 and 1 and k is a constant. In this case, motivation may be expressed as $kp(1-p)$. Since this is maximized at $p=1/2$ motivation will be strongest at tasks with intermediate difficulty and would decrease when the task is made either too difficult or too easy. Indeed, many studies have documented a preference for achievement tasks of intermediate difficulty, and more intense efforts on such tasks, compared with tasks that are either too easy (and therefore are not valuable enough) or too difficult (and therefore too unlikely to yield success).

Distance-dependent Expectancy and Value

In many situations efficiency of effort (and thus also expectancy) increases as one approaches the goal. For example, people may notice that their studying efforts are more efficient closer to the examination because less of what they learn will be forgotten, and because failing to learn gets harder to compensate for. Hungry rats that are running towards a food incentive close the gap towards the goal by 50% with each step when they are two steps away from the goal, but only by 1% when they are 100 steps away from it. Thus, a unit of effort (i.e. a step) is more efficient closer to the goal.

Distance also influences the psychological intensity of value, so that people typically value imminent outcomes more highly than delayed ones. The tendency of people to be impatient about positive outcomes has been termed 'myopia' by economists. For example, people require a relatively large monetary compensation in order to agree to delay in receiving a payment or a product. Another example of myopia is when people choose to purchase cheaper products with relatively high maintenance costs (e.g. a cheaper air conditioner with higher electricity consumption). Presumably, they do so because saving on maintenance is removed in time, so that its psychological importance is discounted relative to saving on the buying price, which is immediate.

Research on discounting of value suggested that costs undergo steeper temporal discounting than benefits. For example, parachute jumping may be perceived as both fun (positive value) and scary (negative value). Both fun and fear would be less intense from a distant than from a close perspective, but the decrease in the intensity of fear would be steeper than the decrease in the intensity of fun. An approach-avoidance conflict results: a parachute jumping in the distant future may seem more fun than scary (and thus would be approached), but in a closer perspective it may seem more scary than fun (and thus would be avoided).

It follows from the expectancy–value model of motivation and from the increase in both value and expectancy closer to the outcome that motivation should increase the closer one gets to the goal. For example, students would be more motivated to study closer to the exam, and hungry rats would pull harder closer to a food incentive, both because their efforts get more efficient and because the incentive seems more intense. This phenomenon, known as the 'gradient of motivation', has been documented in many studies with both humans and animals.

Personal Beliefs in Expectancy

Expectancies may vary not only between situations (e.g. different tasks, different distances from the goal) but also between individuals. Research in personality has documented stable interpersonal differences between generalized beliefs about the controllability of significant life events, as well as between generalized beliefs about the efficacy of efforts towards achieving important outcomes. Thus, some people tend to believe that whether they succeed or fail depends on their efforts, whereas other people tend to believe that success and failure are determined by external events beyond their control. Similarly, some people believe that their efforts are efficient towards achieving desirable outcomes, whereas other people believe that investing effort may yield little outcome. Some people believe that important personal qualities, such as intelligence or social skills, are unchangeable entities, whereas other people believe that these qualities are malleable and may be acquired by learning.

The expectancy–value theory predicts that higher expectancy should result in higher motivation. Higher motivation, in turn, should yield, overall, better outcomes than lower motivation. Indeed, it has been found that beliefs in high expectancy are associated with high levels of motivation and achievement. For example, people who hold a generalized belief that they control events in their lives tend to perform better academically and to engage in more preventive health behaviors than those who do not hold such a belief. People who believe that academic ability may be acquired through learning, rather than being a fixed entity, show greater persistence in the face of difficult tasks and greater motivation to undertake academic tasks after failure. Finally, a large body of research has shown that attributing negative events to external, global and unchangeable causes (i.e. a belief that the expectancy of changing the cause of negative events is low) is associated with depression and with a tendency not to initiate attempts to change an undesirable situation.

SELF-REGULATION

While the objects of pursuit are usually easy to discern (e.g. one knows that eating a meal would satisfy one's hunger), the way to achieve these desirable outcomes may be less obvious. Often,

satisfying even relatively simple drives requires elaborate and extended action plans. For example, getting to a restaurant to satisfy one's hunger may involve choosing a place, getting into the car, driving, waiting, and so on.

The system of self-regulation is responsible for ensuring successful goal pursuit. The tasks of self-regulation include setting a goal, designing means towards achieving it, specifying the conditions under which pursuit has to begin, monitoring for these conditions and setting the system in motion when they materialize, monitoring progress, comparing progress to a standard, finding alternative means in case of insufficient progress, and, finally, disengaging from goal pursuit upon goal achievement or in case of a repeated failure to achieve sufficient progress. In addition, the self-regulation system resolves conflicts among coexisting goals when only one of them can be pursued.

Goal Hierarchies

It is convenient to conceptualize goals as organized in hierarchies of varying levels of abstraction (e.g. doing well academically, passing an examination, reading a book, following lines of print). Goal hierarchies are organized so that each goal (e.g. reading a book) is subordinated to a higher-level goal – which answers the question of why that goal is being pursued (e.g. to pass an exam) – and superordinate to lower-level goals, which constitute the means towards achieving the goal and answer the question of how the goal is to be pursued (e.g. following lines of print). At very high levels of the goal hierarchy are abstract goals, which typically specify general guidelines concerning the type of person one should or should not be (e.g. famous, kind, wealthy). Such abstract goals are often referred to as life tasks, values or self-guides. At very low levels of the goal hierarchy are simple and concrete actions (e.g. picking up the phone).

Cybernetic models of action control specify the process of translating goals into actions. They suggest that goals are translated into subordinated goals (i.e. the 'how' level) repeatedly until a level is reached at which the required action is obvious or automatic. For example, if the goal is studying for an examination, then one might think of reading a textbook as a means towards achieving it. If how to read is obvious, it will not be further translated into means (e.g. a person will not tell him- or herself 'I need to open a book', but rather will perform this action automatically). If, however, how to read is not obvious (e.g. if the book cannot be found), then reading would be further translated into subgoals. Difficulties during goal enactment also foster translation of the action into subgoals. For example, if reading proves difficult because of an ink blot on a page, the person will become conscious of the otherwise automatic action of following lines of print.

Monitoring Progress

Monitoring of progress towards the goal is performed throughout the process. If progress is found to be unsatisfactory, then alternative means for pursuing the same goal are sought. For example, while reading a textbook in order to prepare for an examination, a student would assess periodically whether her rate of reading suffices to ensure that she will cover the desired material. If the rate is too slow, she might try to study in a different way (e.g. ask a friend for her notes). If progress is too slow but no alternative means are available, then one may try to lower the level of the aspired goal (e.g. the student might decide to study only part of the material) or abandon it (give up the goal of doing well in the examination). Monitoring of sufficient or good progress elicits positive emotion, whereas monitoring of insufficient progress elicits negative emotion. It follows that goal attainment, which is naturally accompanied by slowing down of progress, may sometimes be accompanied by a decrease in positive emotion.

Many studies have documented the beneficial effects of easy monitoring of progress on motivation and performance. For example, it has been found in many workplace situations that concrete goals, set by either others or oneself (e.g. 'your goal is to sell five insurance plans tonight') increase performance, both relative to a situation in which no goals are set and relative to less specific goals (e.g. 'do your best'). It is believed that the beneficial effect of concrete goals is due to their facilitative effect on monitoring. Thus, a concrete goal makes it easy to estimate, at each point, how far one is from the goal and whether progress is sufficient to ensure goal attainment. If progress is judged to be insufficient, efforts may be increased until the appropriate level is found. Such feedback on the rate of one's progress is impossible with more general goals such as 'do your best'.

Disengagement

Disengagement occurs when goals are achieved or abandoned. (As noted above, goals are abandoned when judged unattainable.) In cybernetic systems,

disengagement is followed either by returning to the higher-level goal, to which the attained goal was subordinate, in order to receive further subgoals, or by setting a different, unrelated goal. For example, after finishing reading the book, the person will return to the superordinate goal of studying for the examination, and see if there is anything else that has to be done for studying (e.g. read one's notes). If no further action is required to meet the superordinate goal, then the superordinate goal will be disengaged. Alternative goals may then be pursued (e.g. the student may go out with friends to pursue the goal of socializing).

Unfulfilled goals are difficult to disengage, and thus can create a state of rumination – intrusions of goal-related thoughts, increased memory of and overall sensitivity to constructs that are related to the goal, and even the appearance of dreams related to the unfulfilled goal. Usually, after repeated feedback indicating lack of progress, people would disengage unfulfilled goals, or try to substitute other goals for them. For example, a person who failed to get a desired job would first ruminate about it, but gradually would cease to think of that job and set out to find another one. However, when the unfulfilled goal is important (e.g. because it is judged indispensable for satisfying a basic need), it is not likely to be abandoned or replaced with a substitute, and a state of rumination may persist for extended periods of time (e.g. months or even years). A familiar example of a state of rumination is that arising from an unrequited (and unsubstituted) love.

Goal Conflict

Because people strive to satisfy multiple needs, they often experience goal conflict: a motivation to pursue more than one goal at a time. Often, one needs to decide which goal is to be pursued and which has to be abandoned or 'put on hold' (i.e. await execution at a later time). It is theorized that people pursue the strongest of a set of simultaneously experienced motivations.

Of particular interest are conflicts between long-term and short-term goals, often referred to as a 'temptations'. For example, one may strive to lose weight but be tempted by a rich dessert. One may strive to obtain good grades in an examination, but be tempted to party on the evening before. Typically, a long time in advance people resolve the conflict in favor of the long-term goal (i.e. when offered a rich cake in a distant future situation, people readily reject it, indicating that they would rather keep their diet), but as the time of the short-term goal approaches, they become more likely to pursue it (i.e. people would be quite likely to take a rich cake that is offered immediately and break the diet). It is thus often suggested that the motivation to pursue short-term goals undergoes steeper decay over temporal distance than the motivation to pursue long-term goals. For example, the motivation to eat a fatty cake or to go out with friends is strong close to engaging in these activities and sharply decays as they are removed in time. By contrast, the motivations to diet or to study remain relatively constant over temporal distance.

People often attempt to overcome temptations by removing them either physically (e.g. avoiding buying rich cakes, disconnecting the phone on the night before the exam), or mentally (e.g. trying not to think about the cake). To minimize temptations, people may also assign in advance penalties for succumbing to temptations or rewards for successfully overcoming them (e.g. a person promising herself a vacation if she loses weight), mentally exaggerate the importance of the long-term goals (e.g. convincing herself that getting slimmer is the key to social success), or mentally exaggerate the negative consequences of succumbing to the temptation (e.g. convincing herself that eating a cake is more detrimental to the diet than it really is).

Research has shown that by five years of age some children develop effective strategies to delay gratification. They learn to wait for a larger reward (e.g. two marshmallows) later rather than succumb to an immediate smaller reward (e.g. one marshmallow). For example, five-year-olds typically know that in order to facilitate waiting they should cover the tempting reward or at least avoid looking at it. It has been found that effective strategies to delay gratification in young children are associated with higher achievement in adulthood. The ability to delay gratification is believed to be an important component of emotional intelligence.

INTRINSIC AND EXTRINSIC MOTIVATION

It is useful to distinguish between intrinsic and extrinsic motivations for performing an activity. When extrinsically motivated, people perform an activity in expectation of a reward (e.g. people work in order to get money) or a desirable outcome (e.g. students study in order to get a good grade in an examination). Extrinsically motivated behavior may be controlled by external rewards, because reducing (or increasing) the extrinsic reward results in a corresponding decrease (or increase) in the motivation to perform the activity. Such

behavior would not be performed in the absence of rewards.

When intrinsically motivated, people perform an activity for its own sake, usually because they derive pleasure from being engaged in it (for example, a person may enjoy playing music or a child may enjoy solving a puzzle) and perform these activities with no expectation of a tangible external reward. People typically report greater enjoyment and satisfaction when performing an activity for which they are intrinsically, rather than extrinsically, motivated. Intrinsic motivation is typically associated with greater commitment to the activity (e.g. greater chances of spontaneous resumption after an interruption). In addition, it has been suggested that intrinsic motivation enhances creativity.

Increasing Intrinsic Motivation

What are the antecedents of intrinsic motivation? In other words, what determines whether a person would be intrinsically or extrinsically motivated to perform an activity? It is believed that intrinsic motivation may develop via a process of internalization, whereby an activity that is originally performed because of an explicit request by another person (e.g. a parent or a teacher), who often offers rewards for performing it (or punishment for not performing it), becomes internalized, and is later performed in the absence of an external request. For example, children typically require an explicit request (and sometimes even the threat of tangible sanctions) in order to clean up their room, and they would not clean in the absence of an authority figure delivering the request. Gradually, however, (at least some) children would come to appreciate the value of a clean room, and eventually clean up in the absence of an explicit request, being motivated solely by the intrinsic value of the process or the outcome (i.e. because they enjoy cleaning or because they enjoy having a clean room).

Internalization is a gradual process, which follows a number of stages. An activity that was previously regulated externally may be 'introjected': it may be performed because the authority figure is imagined or expected, rather than being physically present, as in the fully externalized case. For example, a child imagines how angry the mother would be when she comes and sees the messy room, and starts to clean it up. The next stage, 'identification', comes with the recognition of the underlying value of a behavior or its outcome (e.g. a child recognizes that a clean room is valuable towards some end, such as making it easier to find things). Finally, at the stage of full internalization, the person comes to derive pleasure from the outcome or the process itself, rather than merely perceiving it to be useful to some end as in the case of identification. For example, when cleaning is fully internalized, a child enjoys cleaning or having a tidy room.

This analysis suggests that helping a person recognize the underlying value of actions should be conducive to internalization. For example, explaining the merits of studying when trying to motivate a person to study would increase the likelihood of a fast internalization of that behavior. It is also evident from this analysis that a subjective perception of free choice would be associated with increased intrinsic motivation. Indeed, research has demonstrated higher intrinsic motivation with tasks involving choice (e.g. when students could choose one of a number topics of study), compared with a no-choice situation (e.g. when the topics were assigned).

Undermining Intrinsic Motivation

Interestingly, giving an extrinsic reward for an intrinsically motivated activity decreases intrinsic motivation. For example, children typically enjoy solving puzzles or drawing, and would engage in such activities spontaneously. However, after they are offered a prize (i.e. an external reward) for doing those activities, their tendency to engage in them spontaneously, in the absence of an external reward, decreases. Monkeys exhibited a similar pattern of behavior: they would press a button to obtain a puzzle to manipulate, but after being offered a reward of food for solving the puzzle, they stopped using the puzzles on their own (i.e. in the absence of the food incentive). Setting contingencies such as 'you have to do your homework if you want to play outside' or 'you have to eat soup in order to get cake' has a similar effect of undermining intrinsic interest in the activity that is designated as a means towards another activity.

It is for this reason that increasing external rewards as a means to motivate people is not always a good strategy to achieve internalization and an enduring behavioral change. Increasing extrinsic rewards may be an effective method of behavioral control only if one plans to administer rewards each time the behavior in question is to be performed and does not expect it to be performed otherwise. For example, increasing payment for extra hours of work should be effective in increasing the worker's motivation to work extra

hours, and does not appear to be problematic inasmuch as workers are not expected to work extra hours with no compensation. However, rewarding homework externally could be problematic inasmuch as one does not intend to continue to reward this behavior in the future, but rather hopes that it will be internalized.

Further Reading

Carver CS and Scheier MF (1999) Themes and issues in the self-regulation of behavior. In: Wyer RS (ed.) *Advances in Social Cognition*, vol. XII, pp. 1–106. Mahwah, NJ: Erlbaum.

Gollwitzer PM and Bargh JA (eds) (1996) *The Psychology of Action*. New York, NY: Guilford Press.

Higgins ET (1997) Beyond pleasure and pain. *American Psychologist* **52**: 1280–1300.

Locke EA and Latham GP (1990) *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice-Hall.

Mischel W, Shoda Y and Rodriguez ML (1989) Delay of gratification in children. *Science* **244**: 933–938.

Sansone C and Harackiewicz JM (eds) (2000) *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*. San Diego, CA: Academic Press.

Weiner B (1989) *Human Motivation*. Hillsdale, NJ: Erlbaum.

Motor Control and Learning

Introductory article

Mark Mon-Williams, University of St Andrews, Fife, UK
James R Tresilian, University of Queensland, St Lucia, Queensland, Australia
John P Wann, University of Reading, Reading, UK

CONTENTS

Motor control is concerned with the processes that allow the central nervous system to interact with the world. How we learn to control movement is central to our understanding of all human behavior.

FITTS' LAW

Pressing a switch, posting a letter through a slot, hitting a nail with a hammer, and many similar tasks all involve moving something from one place to another over a certain distance. A task can vary in the distance to be covered and in the accuracy of placement required. For example, if you hit a nail with a hammer, the size of the nail head determines the accuracy with which you need to position the hammer in order to hit it. It is a matter of common observation that in tasks like these people move more slowly when they need to be more accurate. There is a basic principle at work here: people are limited in how fast they can perform a task while maintaining the requisite level of accuracy. People can be fast or they can be accurate, but they are limited in their ability to be simultaneously fast and accurate. Clearly, it is of practical importance to know what these limits are.

Paul Fitts was one of the first to systematically study this relationship in the laboratory. Using a simple task that involved repeatedly moving a hand-held stylus between two targets, Fitts discovered a mathematical form for the rule that relates speed of performance to the accuracy demands of the task. In later experiments, the same rule was found to apply to a range of other tasks such as single reaching movements to visible targets, placing pegs in holes, and picking up an object. In all these experiments the participants were instructed to perform as fast as possible while maintaining accuracy. These instructions were intended to ensure that the participants performed close to the limits of their ability. The results of performance under different requirements for placement accuracy were interpreted to reveal the way people trade off performance speed for accuracy in tasks requiring placement or positioning (the speed–accuracy trade-off). The measure of the speed of performance used in this kind of experiment is the time taken to complete a movement (movement time, abbreviated MT). The size of the target (denoted as either S or W) is used as the measure of required placement accuracy. The rule Fitts discovered relates MT to the the size of the target (S) and the distance to be moved (A for amplitude) and is written as follows:

$$\mathrm{MT} = a + b \log_2(2A/S) \qquad (1)$$

where a and b are constant parameters. This equation has come to be called Fitts' law: it captures quantitatively the everyday observation that greater accuracy (small S) results in movements of a slower speed, and that moving over greater distances takes longer than moving over smaller distances. Note that movements over different distances to the same-sized target are not performed at a slower speed when the movement distance is greater, despite taking longer; quite the opposite, in fact – people move at higher speeds when they move further.

The values of the parameters a and b in Fitts' law are constant for particular people performing a particular task. If you use Fitts' law to describe the experimental data from a group of people performing one task, say putting pegs in different-sized holes, you will require particular values for a and b. If the same people perform another task, say pressing switches of different sizes, different values for a and b will be required. Different values of a and b will also be required to describe the data

from different groups of people performing the same pegs-into-holes task.

Although Fitts' law has been found to describe the data from a large number of experiments it is necessary to be aware of the limits of its applicability. Since the law is an empirically derived relationship, it may apply only to data of the type collected in the experiments that support it. As mentioned above, there are many such experiments; these have investigated a variety of different tasks, task conditions, and groups of people. Three things about these experiments and the data collected are always very similar, however. First, the data are obtained from people who are required to perform close to the limits of their ability to be both fast and accurate. Second, the targets at which movements are directed are always stationary relative to the person performing the task. Third, the data are the average data from a group of several people. It has been found that Fitts' law often fails to adequately describe data obtained from a single person. It also fails to describe performance when the target is in motion and may also fail to describe data obtained in situations where people were not required to perform as fast and accurately as possible: you can, after all, move as slowly as you like to touch a large target.

Even for the types of data and experimental context from which the law was originally derived, there is some debate about whether the form of the relationship proposed by Fitts [eqn 1] provides the best description of the data. Fitts' law provides a good description of most data sets, typically accounting for over 90% of the variance. Other formulations of the speed–accuracy tradeoff – such as the power law formulation

$$\mathrm{MT} = \alpha(A/S)^{\beta} \qquad (2)$$

where α and β are constant parameters – have been found to provide slightly better descriptions of the data.

Whichever formulation provides the 'best' description of the data, we are still faced with an unanswered question: why is there a lawful relationship between movement duration, accuracy, and amplitude, and why does it take the quantitative form that it does – Fitts' law, [eqn 1]? More generally, we can ask why people move slowly when they need to be accurate. It is generally agreed that the basic reason for this behavior lies in the fact that human movement always has some intrinsic inaccuracy in it, partly because the processes in the nervous system and the muscles that generate movement are contaminated by noise. As a result, any movement you make will deviate slightly from the ideal movement. If you move to a target, these deviations may be large enough to cause you to miss the target. Clearly, this is more likely to happen if the target is small. Starting from this basic principle, two reasons for slowing down when greater accuracy is required have been proposed. First, if you move slowly you will have more time to detect that you are going to miss the target and to make some sort of correction to compensate for your inaccuracy before the movement is completed. Second, the faster a movement is performed the larger the deviations from the required movement tend to be, and so even without making any corrections, a slower movement might be more accurate. These two proposed reasons are not mutually exclusive – slower movements may be intrinsically more accurate and provide the time necessary to correct errors before the movement is over.

CONTROLLED AND BALLISTIC MOVEMENT

As mentioned above, one explanation for why people move more slowly when they need to be accurate is that slow movements allow people to correct for errors while the movement is being executed. When such corrections are made, the movement is said to be 'controlled'. While it is possible for a person to become aware of an error in a movement and respond with a consciously intended correction, skilled performance is characterized by correction processes that operate without conscious awareness.

In general terms, controlled movements are those in which sensory information obtained after the movement was initiated contributes to movement production. The term 'online' is often used to refer to this kind of sensory control of movement. There are a number of different ways in which sensory information can make an online contribution. In aiming and placement tasks sensory information has been shown to be used to make discrete corrections to the ongoing movement: a deviation from the required movement is detected and then a corrective response is produced that is incorporated into the ongoing movement. In other types of movement, sensory information is used to drive performance more continuously and directly. It is known, for example, that some types of eye movement can be driven in this way. These include tracking eye movements, made when you follow a moving object with your eyes, and the vestibular–ocular reflex that allows you to keep your gaze fixed on an object when you move your head

around. The idea that sensory information drives movement production in a continuous fashion has been suggested for other types of motor task such as catching a ball, but the empirical evidence is equivocal.

So far we have implicitly assumed the existence of processes that produce movement without online control by sensory information. Movements that are not influenced online by sensory information are often called 'ballistic' movements, by analogy with firing a projectile – once a projectile such as a bullet leaves the gun its trajectory is no longer influenced by the person who fired it. This analogy should not be taken too literally. A movement that is executed without an online contribution from sensory information is not necessarily executed without any control. Empirical work has shown that motor commands from higher centers in the nervous system can drive movements throughout the period of their execution even in the absence of any online sensory contribution. The production of such movements is not ballistic in the sense of a projectile – the brain does not cause a movement to be 'fired off' like a bullet. The nearest thing in the repertoire of human performance to a movement that is 'fired off' in this fashion is a saccadic eye movement (saccades are the rapid movements that you make when you shift gaze from one object to another). Because of these misleading implications of the term 'ballistic', it is more usual nowadays to refer to movements that are not controlled by sensory information as 'open-loop' movements. It should be clear that care is needed in using the terms 'controlled' and 'ballistic' to describe human movement: the distinction being drawn is between movements where sensory information contributes to performance in an online fashion, and movements where there is no online sensory contribution.

It is important to recognize that most motor tasks performed in everyday life are neither purely controlled by sensory information nor open-loop (ballistic), but rather combine both types of control. For example, in the many aiming and placement tasks that conform to Fitts' law, people initially make an open loop movement towards the target position. As the hand or object being aimed nears the target, the person can detect deviations and errors using vision and, based on this visual information, make online corrective adjustments to the initial open-loop movement. This idea was first proposed by Woodworth at the end of the nineteenth century, and it has since become an established principle of human motor control.

INVERSE KINEMATICS

Most of the tasks that have been discussed so far involve moving something from one place to another. Such tasks are fundamental to most of the things we do in everyday life, from making a cup of tea to playing a game of tennis. When we consider the performance of such tasks we often restrict our attention to the thing being moved – the hand when reaching for something, the finger when pointing at something, the hammer head when hammering in a nail, the racket head when hitting a tennis ball. The thing being moved is sometimes referred to as the 'end-effector', a term originally introduced in robotics.

Since it is the end-effector that executes the task – grasps the object, presses the switch, hits the ball – it is natural to suppose that the nervous system controls the movement of the end-effector, both the path it follows and the speed with which it moves. Thus, the idea arose that the nervous system preplans the trajectory of the end-effector: this plan is the basis for the open-loop component of aimed movements discussed earlier. The nervous system cannot directly control the movement of the end-effector; it can control it only indirectly by contracting muscles and moving body segments around joints. Thus, in order to make the end-effector move along a planned trajectory the nervous system must determine the rotational motions of the contributing body segments necessary for producing the planned motion of the end-effector and also the muscle forces necessary to cause the body segments to move that way.

The problem of determining the segment motions (often called joint motions) that will produce a required motion of the end-effector is called the 'inverse kinematics problem'. The term 'kinematic' means that the motions are considered without reference to the forces that cause them, and it is 'inverse' because you are going from the motion of the end-effector back to the motions around the joints that resulted in that end-effector motion. In order to solve this problem it is necessary to have information about the lengths of the body segments involved and their ranges of motion. When the nervous system possesses such knowledge, in either explicit or implicit form, it is said to possess a kinematic internal model of the body. Note that the accuracy of the model affects the accuracy of movement – if the model is not accurate, then any movement planned using the model will be inaccurate. This is another source of the inaccuracy of human movement

zmentioned at the end of the earlier discussion of Fitts' law.

In human movement a desired motion of the end-effector and a kinematic model of the body is generally insufficient to solve the inverse kinematics problem. That is, these factors alone are not sufficient to find a unique set of joint motions that produces the planned end-effector motion. This is due to the fact that many, possibly infinitely many, different motions of the same body segments can often produce the same end-effector motion. When a system can produce the same result in many different ways it is said to possess 'redundancy'. A redundant system is flexible and adaptable – if one way is blocked you can choose another. When you are free to choose, however, you do not want to be like Buridan's ass, trapped motionless between two or more apparently equal choices. What you need is something that helps you make a choice. Several proposals have been made concerning what this might be.

One popular idea, supported by empirical data, is that the nervous system establishes linkages between joints such that if one joint moves in a particular way, a linked joint is constrained to move in a related way. The existence of such linking constraints between joint motions means that some types of motion are not possible. Linkages of this kind have been called 'joint synergies'. An everyday analogy is the linkage that exists between the steered wheels of a motor vehicle. The two front wheels are linked together so that when the steering wheel is turned the two wheels move together. The linkage prevents the wheels moving independently. The existence of joint synergies reduces the number of possible solutions to the inverse kinematics problem – some logically possible solutions are excluded because of the existence of the synergies.

MOTOR FEEDBACK

Feedback is the name given to the situation in which the output of a system (what the system is achieving) provides an input to the same system. If the output is fed back as input, what the system is doing affects what the system does next. When we talk of feedback in the control of movement we are usually referring to the situation in which performance of movements at one moment affects the nervous system via the sensory systems and by this route can influence later performance. Sensory information about what you are currently doing or what you have done is referred to as 'sensory feedback'.

Sensory feedback about movement can be provided by different sensory systems. For example, feedback about the movements actually being made can be provided by the visual system – you need only look at what you are doing to obtain information about your performance. Such information can also be provided by the sensory organs embedded within the tissue of your muscles, joints, ligaments, and skin – often called 'kinesthetic feedback'. Both visual and kinesthetic feedback provide information about how your body is currently moving or about its current posture. Kinesthetic feedback can also convey information about the amount of force being developed by a muscle or being exerted on an object – information that the visual system cannot provide.

Sensory feedback about how a limb is moving or positioned and about the amount of force being developed can be used online to adjust performance through controlled movements. Feedback can also be about the outcome of a movement or action. This is the most familiar kind of feedback: it is obtained after you have completed your performance and informs you about what was achieved. For example, after you have played a shot in tennis you acquire visual feedback about the outcome – where did the ball land? Outcome feedback is clearly not the kind of feedback that can be used in the online control of movement: it cannot be used to make the kind of adjustments discussed above. Outcome feedback is nevertheless important for motor performance as it can be used to guide the process of motor learning. Many experimental studies of the use of outcome feedback in motor learning have attempted to control exactly what feedback was available to participants by preventing them directly observing performance outcomes; instead, the feedback is provided verbally by the experimenter or by a display device such as a meter or computer screen. In these cases the outcome feedback is referred to as 'knowledge of results'. Research using this technique for providing feedback has demonstrated its importance in the learning process.

SKILL AND LEARNING

Skill can be defined as a special ability acquired through training. Implicit within such a definition is the idea that skill is based upon learning. If one considers that most purposeful voluntary controlled movement is acquired after birth, it might be argued that all those who research motor control are concerned with the issue of learning at some level. In line with this, one of the major stimuli to

research in the area of motor control has been the desire to understand how humans reach certain levels of proficiency in tasks involving movement. The motivation behind this desire is straightforward – if we understand how skilled actions are learned, then it should be possible to design training programs that optimize the level of performance achieved by an individual in a particular task. Moreover, discovering how skills are learned would potentially allow individuals to achieve a set level of proficiency in the shortest possible time. The practical implications of optimized training are clear: athletes reaching the highest sporting achievements, increased work productivity in jobs involving manual control, and therapeutic strategies for conditions characterized by deficits in motor ability. Thus, it is not surprising that many studies in motor control have been concerned specifically with how people perform a given task under different learning conditions.

Initial attempts to understand the development of skilled behavior explored specific questions related to practical issues of whether it is better to practice the whole skill (e.g. typing bimanually) or components of the skill (e.g. typing unimanually), the influence of knowledge of results (see above), and the optimum intervals between training and rest, etc. There is no doubt that these approaches have yielded important practical information that can be used, for example, by sports coaches when devising specific training programs. The limitation of such approaches is that they tend to be task-specific and often fail to shed light on the underlying mechanisms that support skilled behavior. It is therefore worth considering the progress made in elucidating the mechanisms that underpin the learning of motor skills.

Modern attempts to understand motor learning have adopted computational approaches in order to clarify the problems faced by the nervous system and to explore possible solutions employed by the system. These approaches all advocate ways in which the system can acquire knowledge about the environment and suggest how the system can modify (adapt) its behavior in response to changes in the body or the environment. There are three major computational approaches to learning: supervised, unsupervised, and reinforced learning schemes. There are a number of different methods of implementing learning models (e.g. artificial neural networks). One of the ultimate goals of motor research is to use such models together with empirical investigation in order to identify the manner in which the human nervous system learns to make skilled movement.

SEQUENTIAL RESPONSE LEARNING

The control of human movement rests upon the nervous system maintaining and updating information about the relationship between the environment and the neuromuscular apparatus. Furthermore, the production of a number of complex behaviors rests upon the nervous system storing information about the relative sequence of a number of different actions. The acquisition of information about the correct sequence of events within a complex task is referred to as 'sequential response learning'.

The necessity of sequential response learning is well known to experienced pianists. A number of musical compositions require the pianist to play a lengthy sequence of notes that run up (or down) the piano keyboard. The difficulty is that the pianist quickly runs out of fingers and thumbs, requiring the seamless transition of the hand to the right (or left). One possible solution to this problem is known as 'fingering' and involves the pianist learning to move the thumb underneath the fingers at the appropriate point to play the subsequent note while simultaneously pivoting the hand in the desired direction. In order to become accomplished at this task the pianist must learn the correct sequence of digit movements (responses) for all of the different scales in all of the various keys. It will be noted that a countless number of everyday tasks also require sequential response learning.

A number of investigators have established that normal participants can learn complex patterns of sequences without being able to explicitly report the pattern of the sequence. These experiments show that participants' reaction time is decreased when asked to make repetitive responses to a complex series of stimuli. It has been argued from these experiments that sequential response learning involves implicit memory (where previous experience facilitates performance without the explicit recollection of that experience). It has been discovered that patients with dense anterograde amnesia are capable of showing sequential response learning. These findings have been used to argue that sequential learning does not require the brain systems that are responsible for explicit memory.

Further Reading

Desmurget M and Grafton S (2000) Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Science* **4**: 423–431.

Fitts PM (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* **47**: 381–391.

Heuer H and Keele S (eds) (1996) *Handbook of Perception and Action*. New York, NY: Academic Press.

Jordan MI and Wolpert DM (1999) Computational motor control. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Latash ML (1993) *Control of Human Movement*. Leeds, UK: Human Kinetics.

Plamondon R and Alimi AM (1997) Speed/accuracy trade-offs in target directed movements. *Behavioral and Brain Sciences* **20**: 279–349.

Rosenbaum DA (1990) *Human Motor Control*. San Diego, CA: Academic Press.

Rothwell JC (1987) *Control of Human Voluntary Movement*. London, UK: Croom Helm.

Schmidt R and Lee T (2000) *Motor Control and Learning: A Behavioral Emphasis*, 3rd edn. Leeds, UK: Human Kinetics.

Wolpert DM, Ghahramani Z and Flanagan JR (2001) Perspectives and problems in motor learning. *Trends in Cognitive Science* **5**: 487–494.

Motor Development

Introductory article

Karen E Adolph, New York University, New York, USA
Idell Weise, New York University, New York, USA
Ludovic Marin, New York University, New York, USA

CONTENTS

Motor development refers to changes in children's ability to control their body's movements, from infants' first spontaneous waving and kicking movements to the adaptive control of reaching, locomotion, and complex sport skills.

CLASSICAL AND CONTEMPORARY THEORIES OF MOTOR DEVELOPMENT

Early Pioneers

Researchers in the 1930s and 1940s provided the first detailed descriptions of change in infants' motor skills. Arnold Gesell, for example, identified 22 stages in the development of crawling, beginning when infants lifted their heads from a prone position and ending when they could crawl smoothly on their hands and feet. Myrtle McGraw described seven primary stages in the development of walking, progressing from newborns' reflexive stepping movements to independent walking at the end of infants' first year.

These early pioneers believed that motor development resulted from neuromuscular maturation – largely autonomous changes in infants' brains, muscles, and growing bodies. From this perspective, rich catalogs of motor milestones would yield insights into the maturation process. Normative descriptions of motor milestones were widely published in books, journals, and newspaper columns and are still the accepted guidelines for informing clinicians, doctors, and parents about the path of normal motor development.

However, the early pioneers may have done their job too well. Once neuromuscular maturation became the broadly accepted explanation for motor development and the major skills were amply cataloged, the urge for further research dwindled. From the 1950s to the 1980s, motor development was virtually ignored by developmental psychologists.

Contemporary Approaches

In the 1980s, interest in motor development was rekindled as new research methods and sophisticated recording technologies provided improved ways of measuring and analyzing infants' motor skills. More important, recent conceptual advances opened a new perspective for understanding developmental change. Neuromuscular maturation has lost its privileged status as the central impetus for motor development. Emphasis on the contributions of peripheral factors, perceptual information, and learning for adaptive control of movements have reinvigorated the field of motor development research.

A prominent influence on contemporary research is the dynamic systems approach inspired by the Russian physiologist Nicholai Bernstein, and popularized and expanded by developmental psychologist Esther Thelen. On the dynamic systems account, new motor skills may emerge from the confluence of many interacting factors, each with its own developmental trajectory. Each factor must be in place, sufficiently ripe and ready to go, but no factor enjoys privileged status compared with any other. Independent walking, for example, may emerge when infants have sufficient muscle strength, slimmed down body proportions, motivation to go some place, balance control, the appropriate environmental properties to support the action, as well as brain maturation.

A second prominent influence is the perception–action approach inspired by James and Eleanor Gibson. They argue that perception and movement are linked together. To be planned and executed adaptively, actions require perceptual information

about the relevant properties of the environment and the body, and the relationship between them. On the other hand, perceptual information typically requires movement to create the relevant structures in light, sound, and other ambient arrays of energy. For example, exploratory movements of the eyes, head, body, and extremities generate perceptual information in light, sound, muscles, and skin. Actions likewise generate more information for perceptual systems. Perceptual-motor learning is critical for discovering and honing exploratory movements and for discriminating and using the relevant information obtained from exploration.

Importance of Posture

On both classical and contemporary accounts, posture plays a central role in motor development. On the classical neuromuscular maturation account, infants' slow triumph over gravity as they acquire increasingly erect postures is evidence of greater cortical control of actions. On the contemporary dynamic systems and perception–action accounts, posture is the biomechanical foundation for action. Manual skills, locomotor skills, and even lifting or turning the head require a stable postural base. Each postural milestone in development (sitting, crawling, walking, etc.) requires learning about a new perception–action system.

The first step of learning to control a new postural system is to co-contract large muscle groups. This co-contraction frees up resources and allows attention to be focused on the goal-directed part of the movement. However, co-contraction results in jerky, energetically inefficient movements. After extended practice, muscles are activated in sequence and muscle forces are used sparingly so as to exploit gravitational and inertial forces which impel movements for 'free'.

MANUAL SKILLS: EXPLORING OBJECTS

Reaching and Grasping

Arm movements begin before birth. Fetuses wave their arms, produce isolated finger movements, and display coordinated arm–hand movements such as bringing their thumb to their mouth. After birth, infants must cope with gravity outside the buoyant, fluid-filled womb. Newborns' arm-flaps and jerky arm extensions become successful reaching to objects at four to five months. Careful motion analysis shows that infants initially solve the problem of getting their hand to a target in their own way. More sluggish infants must overcome gravity to move their arms from their sides, and more active infants must dampen inertial forces to control ongoing, spontaneous arm-flaps.

Like reaching, infants' first grasps are inefficient. They open their hand to the proper shape only after contacting the object. By about eight months, babies can use visual information about object size, shape, and orientation to adjust the shape of their hand before contacting the object. Infants need not look at their hand to guide its shaky path to the target because successful reaching and grasping in the dark occurs at the same time as in the light. However, visual information about the object's location and other properties is extremely important for planning arm and hand movements adaptively. A dramatic illustration is that young infants can intercept moving targets. They do so by moving the hand on the opposite side of their body to the appropriate location before the object arrives.

Apparently, coordination between manual skills and perceptual exploration is fundamental. In addition to hand–mouth behaviors, newborns turn their heads to keep an outstretched arm in view. Older infants use their reaching and grasping skills to bring objects (and fingers) to their mouths, and coordinate visual, tactile, and oral exploration by alternating between looking at the object, turning and fingering it, and sucking on it.

Even a simple reaching task requires balance control. When an arm is extended over the base of support, the center of gravity is displaced. Before babies' back and abdominal muscles are strong enough to support them in a sitting position, they must use their hands to maintain balance in a 'tripod' configuration. After onset of independent sitting at five to six months, their hands are free to reach, grasp, and manipulate small objects. Over the next several months, infants become skilled at coordinating reaching and leaning so as to prevent falling over.

Using Hand-held Implements as Tools

Once infants master the motor components of reaching and grasping, they can use these skills to extend their own abilities and to bring about rewarding outcomes. Babies bang rigid objects against rigid surfaces to make a noise, but cease banging squishy objects on squishy surfaces. When given a choice between different objects and surfaces, they test various combinations to find the object/surface combination that makes the most noise. Although nine-month-olds can hold a

spoon and use it to bang on their high-chair tray, it isn't until several months later that they can manage fine motor control and incorporate the spoon into a complex plan for transporting food to their mouth.

Swiss psychologist Jean Piaget was first to describe how older infants separate reaching and grasping into means for achieving ends. For example, after eight months or so infants reach and pull a cloth to bring a toy on the cloth into closer proximity. Similarly, older infants extend their reaching space by leaning forward with a stick in their hand to contact an object. Toddlers use canes as tools to rake in objects out of reach. They show understanding of the relationship between tool and target by turning the cane to the appropriate orientation.

BALANCE AND LOCOMOTION: EXPLORING THE LAYOUT

Interlimb Coordination in Crawling and Walking

Precursors of locomotion begin long before infants take their first steps. Fetal and newborn spontaneous arm and leg movements contribute to building and strengthening muscles necessary for later locomotion. Coordination between limbs may already be in place in newborns. When the constraints of muscle strength and balance control are removed, newborn infants produce the 'crawling', 'swimming', and 'stepping' reflexes. They display crawling movements when placed on a gentle downhill slope, swimming movements when put into a pool of water, and stepping movements when held upright on a table. Each of these movements resembles the coordination patterns of later appearing skills. And each shows a U-shaped developmental trajectory: the reflexive patterns displayed by newborns disappear and then reappear months later in altered form.

The classical explanation for this trajectory was neuromuscular maturation: as cortex becomes more refined and myelinated, the reflexive movements are inhibited and then reappear under cortical control. Recent research supports the contemporary dynamic systems account – that the central nervous system is not a privileged factor and that peripheral factors may play the pivotal role in development. When stepping infants' legs are weighted to simulate fat gain, they stop stepping. When non-stepping infants' legs are held in a tank of water, they step. When non-steppers are held in an upright position over a motorized treadmill, they step. Apparently, the peripheral factors of leg muscles and leg fat were the key factors in the U-shaped trajectory. Infants stop stepping when their legs are too fat and weak. Similarly, the treadmill compensates for leg strength by stretching the leg backward and allowing it to pop forward like a spring.

Crawling is usually infants' first success at independent locomotion. Many infants invent idiosyncratic styles of crawling at first, such as scraping along on their bellies. However, maintaining balance on hands and knees is highly constrained biomechanically and nearly all infants crawl the same way from their first week on hands and knees – a modified diagonal trot. Practice using idiosyncratic belly crawls is not wasted, however, because ex-belly crawlers are twice as fast and efficient as infants who skipped this phase, once both groups of babies begin crawling on hands and knees.

Independent walking typically appears after several upright transitional stages (pulling to a stand, balancing, and cruising sideways using furniture for support). There is a dramatic change in walking gait from infants' first steps to the toddler years and beyond. New walking is characterized by small steps with a wide space between the feet. Toes point outward and the legs are almost straight. Elbows are bent upward and the palms face the ceiling. Rather than a heel-to-toe progression like adults, infants walk on their toes or plant their whole foot down at one time. This strange gait pattern progressively improves so that by seven years of age children walk like adults.

Balance Control

Locomotion requires continual adjustment to disequilibrium by using compensatory sway to keep the center of mass over the base of support. Sometimes compensatory movements are obvious, as when trying to recover from a fall. Other times, as when just standing or walking on flat ground, it is not as apparent. However, detailed recording shows that the body is constantly swaying, even if movements are invisible to the naked eye. Keeping balance is especially difficult for infants because their body proportions are top-heavy and they fall faster given their short stature.

Children rely on several sources of information for balance control: somatosensory information from their muscles, joints, and skin, vestibular information from accelerations of the head, and visual flow information created by the body's movement. When visual flow is simulated by surreptitiously moving the surrounds (as in a flight

simulator), standing adults generate compensatory sways in accordance with the simulated visual flow. Walking infants also respond with compensatory movements, but they overcompensate and often stagger and fall. Even babies who cannot locomote independently show signs of sensitivity to visual information relevant for balance control by moving their heads in accordance with simulated flow.

Navigation over Variable Terrain

Adaptive locomotion requires learning. When newly locomotor infants are challenged with variable terrain, they show little ability to distinguish safe from risky ground. For example, at the edge of a cliff or steep slope, they plunge over heedlessly. (Infants are always protected in such experiments by covering the drop-off in safety glass or having a trained experimenter nearby to provide rescue.) After several weeks of locomotor experience, they avoid risky ground. Surprisingly, learning appears to be specific to each postural milestone in development. For example, experienced crawlers avoid risky slopes, but the same infants attempt the same risky ground a few weeks later after they begin walking.

CONCLUSION

Motor development does not stop after infancy. After mastering basic postural, manipulative, and locomotor skills, children acquire a host of more complex activities – writing, playing the piano, jumping, skipping, etc. As in infant development, later skill learning begins with stiff, wasteful, and uncoordinated movements and becomes progressively more rhythmical, smooth, and efficient.

Motor development is not an isolated domain. The newest research in motor development focuses on interactions between acquisition of new motor skills and developments in perceptual, cognitive, social, and affective domains.

Further Reading

Adolph KE (1997) Learning in the development of infant locomotion. *Monographs of the Society for Research in Child Development* **62** (3, serial no. 251).

Adolph KE, Eppler MA and Gibson EJ (1993) Development of perception of affordances. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, pp. 51–98. Norwood, NJ: Ablex.

Bernstein N (1967) *The Co-ordination and Regulation of Movements*. Oxford, UK: Pergamon Press.

Berthental BI and Clifton R (1998) Perception and action. In: Kuhn D and Siegler R (eds) *Cognition, Perception and Language*, pp. 51–102. New York, NY: John Wiley.

Campos JJ, Anderson DI, Barbu-Roth MA *et al*. (2000) Travel broadens the mind. *Infancy* **1**: 149–219.

Gesell A (1954) Maturation and the patterning of behavior. In: Carmichael L (ed.) *Manual of Child Psychology*, pp. 209–233. New York, NY: John Wiley.

Gibson EJ (1988) Exploratory behavior in the development of perceiving, acting and the acquiring of knowledge. *Annual Review of Psychology* **39**: 1–41.

Gibson EJ and Walk RD (1960) The 'visual cliff'. *Scientific American* **202**: 64–71.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Hofsten C (1993) Prospective control: a basic aspect of action development. *Human Development* **36**: 253–270.

Hofsten C von and Fazel-Zandy S (1984) Development of visually guided hand orientation in reaching. *Journal of Experimental Child Psychology* **38**: 208–219.

McGraw MB (1945) *The Neuromuscular Maturation of the Human Infant*. New York, NY: Columbia University Press.

Thelen E (1984) Learning to walk: ecological demands and phylogenetic constraints. *Advances in Infancy Research* **3**: 213–260.

Thelen E (1995) Motor development: a new synthesis. *American Psychologist* **50**: 79–95.

Thelen E and Smith LB (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Motor Learning Models

Introductory article

Daniel M Wolpert, University College London, UK
Zoubin Ghahramani, University College London, UK

CONTENTS

Motor learning models are adaptable systems within the central nervous system which are responsible for skilled motor behavior.

INTRODUCTION

Humans demonstrate a remarkable ability to produce skilled behavior effortlessly. Our own perception of the effort involved in such skill is in stark contrast with the complexity of the algorithms required to control robots even with their currently meagre performance. The basic requirement for such performance is to learn the transformations between sensory and motor variables (Figure 1). Such transformations are accomplished by the environment and by the musculoskeletal system; these physical systems transform descending motor commands from the central nervous system (CNS) into a new configuration of the body (state) which in turn generates sensory feedback from touch and position sensors in the body. It is also possible, however, to consider internal transformations, implemented by neural circuitry, that mimic this external motor-to-sensory transformation (Figure 1). Such internal transformations are known as internal forward models. Forward dynamic models, for example, predict the next configuration of the body (e.g. position and velocity of the arm) given the current configuration and the motor command. This is in contrast to internal inverse models, which invert the motor-to-sensory transformation, providing the motor command that will cause the desired change in configuration. As inverse models produce the motor command required to achieve some desired result they can be used by the CNS to control the motor system.

INVERSE MODELS FOR CONTROL

Several models have been proposed for how the motor commands are generated. The equilibrium point control model suggests that the CNS specifies spatial parameters and relies on the spring-like properties of muscles and reflex loop to move the limb. For example, a set of muscle activations defines a stable equilibrium position of the hand in space. Movement is achieved by playing out in time a set of such equilibrium positions along a desired trajectory. This model uses muscle and spinal cord properties as a feedback controller to pull the hand along the desired trajectory. Because of the dynamics of the system, the actual positions may not follow exactly the desired trajectory. An alternative proposal is that an inverse model is constructed to map desired states into motor commands. These two approaches can be contrasted by considering the problem of moving a ball around a circular path by specifying the forces acting on it. For inverse model control the equations of motion are solved and the forces on the ball applied to generate the desired acceleration. For equilibrium point control, the ball is attached by a spring to a control point which is simply moved around the circular path, with the ball following along behind.

FORWARD MODELS FOR PREDICTION

To control the body the CNS needs to know its configuration, that is its state. However, the CNS faces two problems. First, considerable delays exist in the transduction and transport of sensory signals to the CNS. Second, the CNS must estimate the configuration from noisy sensory signals. For example, consider a tennis ball we have just hit. If we simply used the retinal location of the ball to estimate its position, our estimate would be delayed by around 100 ms. A better estimate can be made by predicting where the ball actually is, using a forward model. This estimate can be improved by knowing how the ball was hit (i.e. the motor command), in conjunction with an internal

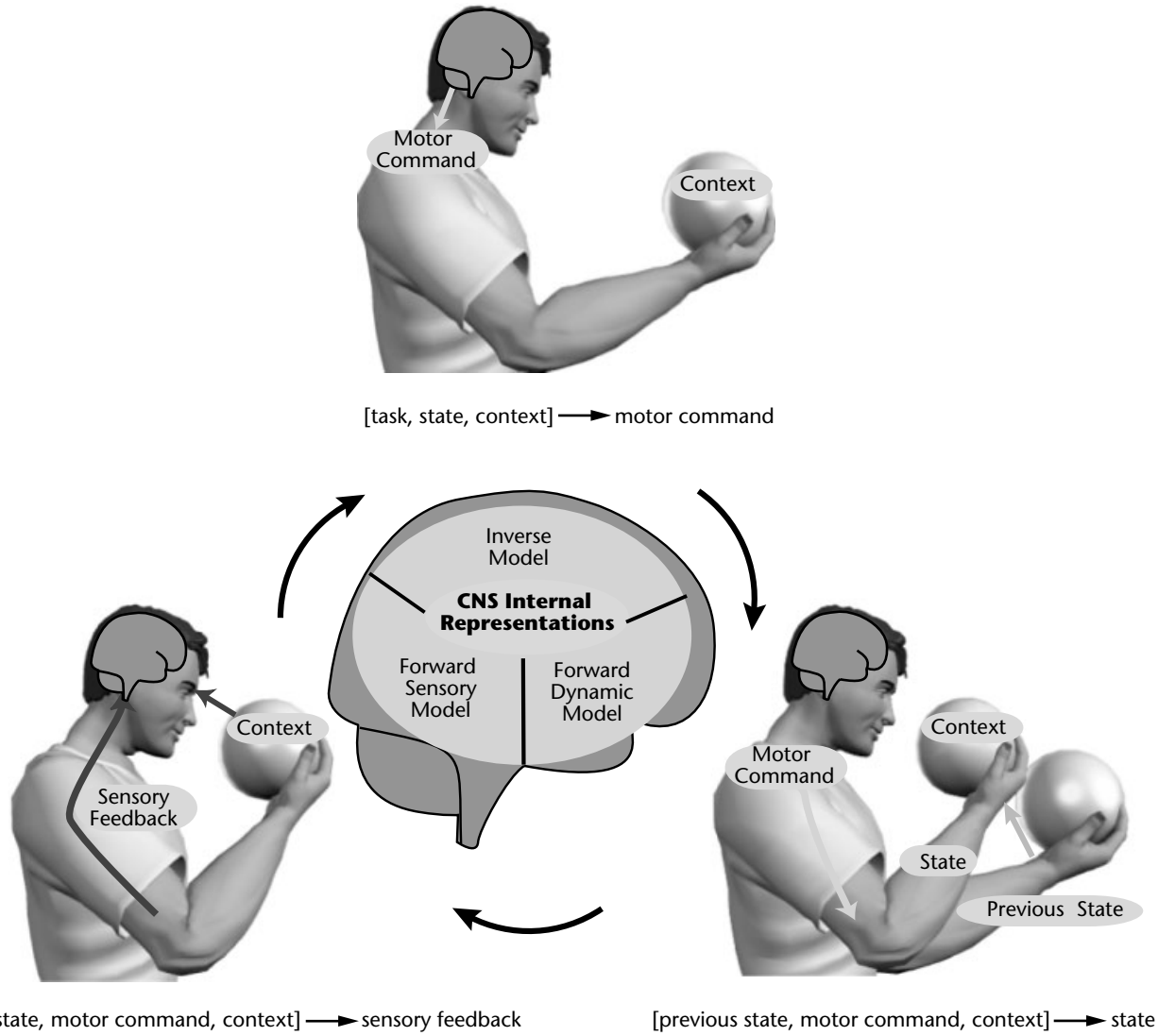


Figure 1. The sensorimotor loop and the relationship between motor commands, state, context, task and sensory feedback. The loop can be divided into three stages which govern the overall behavior of the sensorimotor system. The first stage specifies the motor command generated by the central nervous system (CNS) given the state and a particular task (top). The second stage determines how the state changes given the motor command (right). The third closes the loop by specifying the sensory feedback given this new state (left). These three stages are represented in the CNS as internal models – the inverse model, forward dynamic model and forward sensory model.

forward model of the ball's dynamics. This combination, using sensory feedback and forward models to estimate the current state, is known as an 'observer', an example of which is the Kalman filter. The major objectives of the observer are to compensate for the delays in the sensorimotor system and to reduce the uncertainty in the state estimate which arises due to noise inherent in both the sensory and motor signals. Such a model has been supported by empirical studies examining estimation of hand position, posture and head orientation. Damage to parietal cortex has been linked to inability to maintain such state estimates.

Using an observer it is also possible to predict the future. For example, by estimating the outcome of an action before sensory feedback is available, a forward model can reduce the deleterious effects of feedback delays in sensorimotor loops. Such a system is thought to underlie skilled manipulation. For example, when we move an object held in the hand, the fingers tighten their grip in anticipation to prevent the object slipping, a process shown to rely on prediction. State prediction can also be used

in mental simulation of intended movements, and damage to parietal cortex can lead to an inability to simulate mentally movements with the affected hand.

Sensory prediction can also be used to cancel out the effects of sensory changes induced by self-motion, thereby enhancing more relevant sensory information. Predictive mechanisms underlie our inability to tickle ourselves; the forward model predicts and attenuates the tickle sensation. It has been shown that the reduction of the felt intensity of self-applied tickle critically depends upon on the precise spatial and temporal alignment between the predicted and actual sensory consequences of the movement. Similarly, sensory predictions provide a mechanism to determine whether a movement is self-produced (and hence predictable), or produced externally. It has been proposed that a failure in this mechanism may underlie the delusions of control which can occur in schizophrenia, in which patients may feel that their body is being moved by forces other than their own. Interestingly, damage to the left parietal cortex can lead to a relative inability to determine whether viewed movements are one's own or not.

LEARNING INTERNAL MODELS

Internal models, both forward and inverse, capture information about the properties of the sensorimotor system. These properties are not static but change throughout life both on a short timescale, owing to interactions with the environment, and on a longer timescale, owing to growth. Internal models must therefore be adaptable to changes in the properties of the sensorimotor system. Learning can be considered as reducing an error measure associated with the internal model, for example a prediction error for a forward model. In order to learn, the system needs to measure this error signal and update parameters in the model, such as synaptic weights, to reduce future errors.

Since forward models predict the consequences of an action, the environment readily provides an appropriate training signal. The predicted outcome of an action can be compared with the actual outcome, and the discrepancy – the error signal – used to update the forward model.

Acquiring an inverse internal model through motor learning is generally a difficult task. Unlike forward models, the appropriate training signal for inverse models – the error in the motor command – is not directly available. When we fail to serve an ace no one tells us how our muscle activations should change to achieve this task. Instead, we receive error signals in sensory coordinates, and these sensory errors need to be converted into motor errors before they can be used to train an inverse model. An ingenious solution to this problem, called feedback–error–learning, proposes that a hard-wired, but not perfect, feedback controller exists which computes a motor command based on the discrepancy between desired and actual sensory feedback. The motor command is the sum of the feedback controller motor command and the output of an inverse model. If the feedback controller were to produce no motor command, then there would be no discrepancy between desired and actual sensory feedback, there would be no error in performance, and the inverse model would be performing perfectly. Based on this reasoning, the output of the feedback controller can be regarded as the error signal, and used to train the inverse model, an approach that is highly successful. Neurophysiological evidence supports this learning mechanism within the cerebellum for the simple reflex eye movement called the 'ocular following response'. This suggests that the cerebellum constructs an inverse model of the eye's dynamics.

Recent work on learning novel dynamics has focused on the representation of the inverse model. If people make point-to-point movements when the dynamics of their arms are changed, either by interaction with a robot or by Coriolis forces present in a rotating room, it has been shown that over time they adapt and are able to move naturally. Several theoretical questions have been addressed using this paradigm. The learning of dynamics generalizes in joint-based coordinates, learning depends on the states experienced but not on the order in which they are experienced, state-dependent fields are learned more efficiently than temporally changing fields, and during learning both forward and inverse models are simultaneously adapted.

MODULARITY OF INTERNAL MODELS

Human motor behavior demonstrates an enormous repertoire which enables us to interact with many different objects in a variety of different environments. No single internal model can capture the behavior of all the varieties of objects and environments that we interact with. One attractive hypothesis is that the sensorimotor system builds multiple internal models and switches between them when the context becomes appropriate. For example, the moment we grasp an object and lift it off a table, the dynamics of our arm change

abruptly and therefore a different controller is switched in. There are two challenges for the central nervous system in building a modular controller: the first is determining when to switch between different controllers; and second, the controllers must be adapted to cover the range of objects and environments the individual has to learn.

One approach to learning when to switch between different control modules is to use an architecture that has multiple paired forward and inverse models, each pair constituting a module. Each pair is tuned to a different context, for example to objects with different dynamics. Each forward model predicts the behavior of the motor system corresponding to its context. The array of forward models can then be used to determine which context we are acting in. Each forward model makes a prediction of the consequences of the descending motor command and these predictions are compared with the actual sensory feedback. The set of errors can be used to determine the likelihood of each context – the smaller the error, the more likely the context. This array of models therefore acts as a set of hypothesis testers. In addition, sensory information can be used to set the prior probability of each context; for example, sensory information can be used to select a module based on an object's visual appearance prior to physical interaction with the object. The likelihood and the prior probability can be optimally combined using Bayes' rule, which takes the product of these two probabilities and normalizes over all

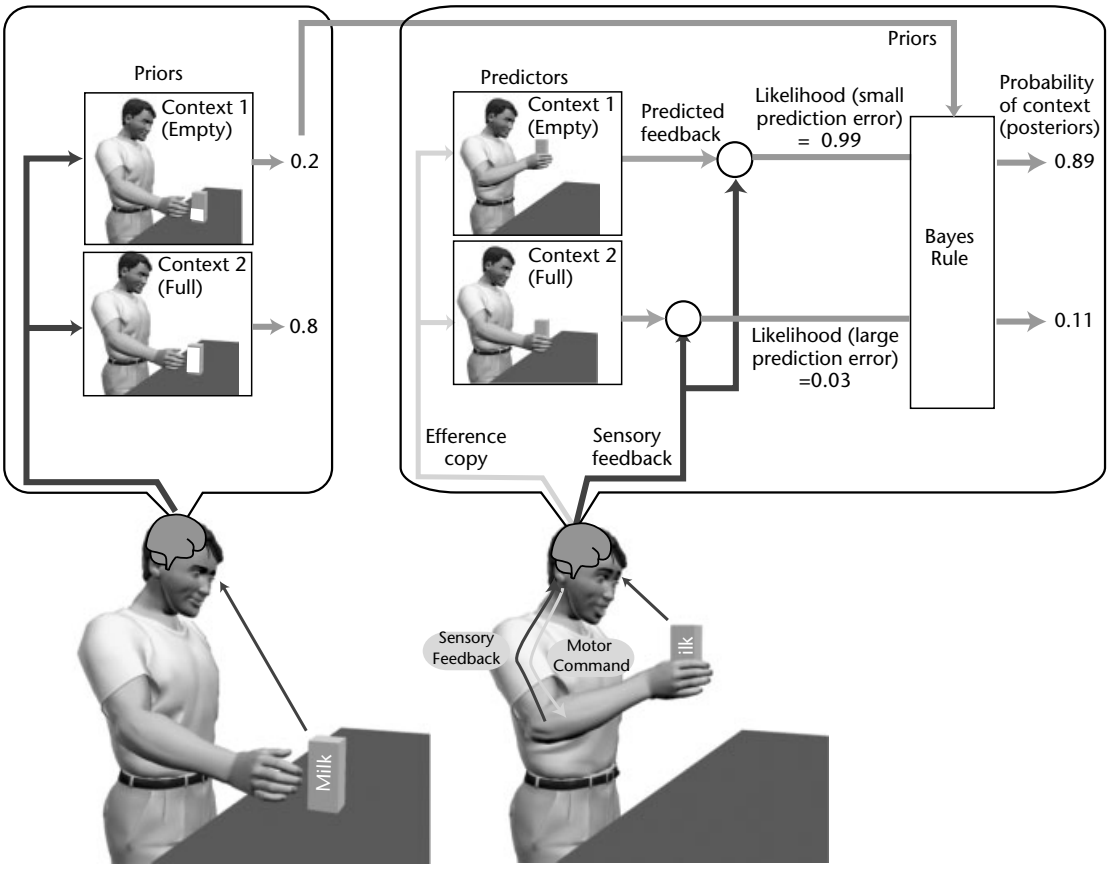


Figure 2. Context estimation with just two contexts: that a milk carton is empty or full. Initially sensory information from vision is used to set the prior probabilities of the two possible contexts; in this case, the carton appears more likely to be full. When the motor commands appropriate for a full carton are generated an efference copy of the motor command is used to simulate the sensory consequences under the two possible contexts. The predictions based on an empty carton suggest a large amount of movement compared with the full carton context. These predictions are compared with actual feedback. As the carton is in fact empty, the sensory feedback matches the predictions of the empty carton context. This leads to a high likelihood of an empty carton and a low likelihood of a full carton. The likelihoods are combined with the prior probabilities using Bayes' rule to generate the final probability of each context.

possible contexts, to generate a probability for each context. These probabilities are then used to switch between the controllers, with modules with higher probability contributing more to the final motor command. Although the set of predictors start off naive in this model, they compete to learn the experienced dynamics, differentiating over time. This competition is achieved by making the amount by which each module learns proportional to the probability of its context. This ensures that different models learn different contexts.

Figure 2 shows a person picking up what appears to be a full milk carton but is in reality empty. This shows how the predictive models correct on-line for erroneous prior probability assessment which initially weighted output of the controller for a full milk carton more than that for an empty one. Bayes' rule allows a quick correction to the appropriate control even though the initial strategy was incorrect. This example has two modules representing two contexts. However, the modular architecture can, in principle, scale to thousands of modules (that is, contexts). The interpretation of the processes necessary for context estimation is consistent with recent neurophysiological studies in primates showing that the central nervous system models the expected sensory feedback for a particular context as well as representing the likelihood of the sensory feedback given the context.

CONCLUSION

Internal models are fundamental for understanding state estimation, prediction, context estimation, control and learning in the motor system. These processes are the basis for human motor behavior.

Further Reading

Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* **9**(6): 718–727.

Mussa-Ivaldi FA (1999) Modular features of motor control and learning. *Current Opinion in Neurobiology* **9**(6): 713–717.

Wolpert DM and Ghahramani Z (2000) Computational principles of movement neuroscience. *Nature Neuroscience* **3**: 1212–1217.

Multidimensional Scaling

Intermediate article

Mark Steyvers, Stanford University, Stanford, California, USA

CONTENTS

Multidimensional scaling is a family of techniques for fitting spatial distance models to proximity data for the purpose of exploratory data analysis and/or analysis of the structure of mental representations.

ASSUMPTIONS OF MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) describes a family of techniques for the analysis of proximity data on a set of stimuli to reveal the hidden structure underlying the data. The proximity data can come from similarity judgments, identification confusion matrices, grouping data, same–different errors, or any other measure of pairwise similarity. The main assumption in MDS is that stimuli can be described by values along a set of dimensions that places these stimuli as points in a multidimensional space, and that the similarity between stimuli is inversely related to the distances of the corresponding points in the multidimensional space.

The Minkowski distance metric provides a general way to specify distance in a multidimensional space:

$$d_{ij} = \left[\sum_{k=1}^{n} |x_{ik} - x_{jk}|^r\right]^{1/r} \quad (1)$$

where n is the number of dimensions, and x_{ik} is the value of dimension k for stimulus i. With $r = 2$, the metric equals the Euclidian distance metric while $r = 1$ leads to the city-block metric. A Euclidian metric is appropriate when the stimuli are composed of integral or perceptually fused dimensions such as the dimensions of brightness and saturation for colors. The city-block metric is appropriate when the stimuli are composed of separable dimensions such as size and brightness (Attneave, 1950). In practice, the Euclidian distance metric is often used because of mathematical convenience in MDS procedures.

MDS can be applied with different purposes. One is exploratory data analysis; by placing objects as points in a low dimensional space, the observed complexity in the original data matrix can often be reduced while preserving the essential information in the data. By a representation of the pattern of proximities in two or three dimensions, researchers can visually study the structure in the data.

It has been used also to discover the mental representation of stimuli that explains how similarity judgments are generated. Sometimes, MDS reveals the psychological dimensions hidden in the data that can meaningfully describe the data. The multidimensional representations resulting from MDS are also often useful as the representational basis for various mathematical models of categorization, identification, and/or recognition memory (Nosofsky, 1992) or generalization (Shepard, 1987).

Some of the issues in MDS can be illustrated with the analysis of a face similarity judgment task. In Figure 1(a), the average of a group of subjects' similarity ratings is given for 10 faces shown in Figure 1 (c). The idea is to reveal some of the perceptual dimensions that subjects might have used when generating similarity judgments for these faces.

TECHNIQUES FOR MDS

There are many different MDS techniques to analyze proximity data and many issues in the analysis and interpretation of the results. First, there is the distinction between metric and nonmetric MDS. The goal of metric MDS is to find a configuration of points in some multidimensional space such that the interpoint distances are related to the experimentally obtained similarities by some transformation function (e.g. a linear transformation function). If the proximity data are generated with Euclidian distances for some stimulus configuration, then a procedure called classical metric MDS (Torgeson, 1965) can exactly recreate the configuration of points. Because a closed-form solution exists to find such a configuration of points, classical metric MDS can be performed efficiently on

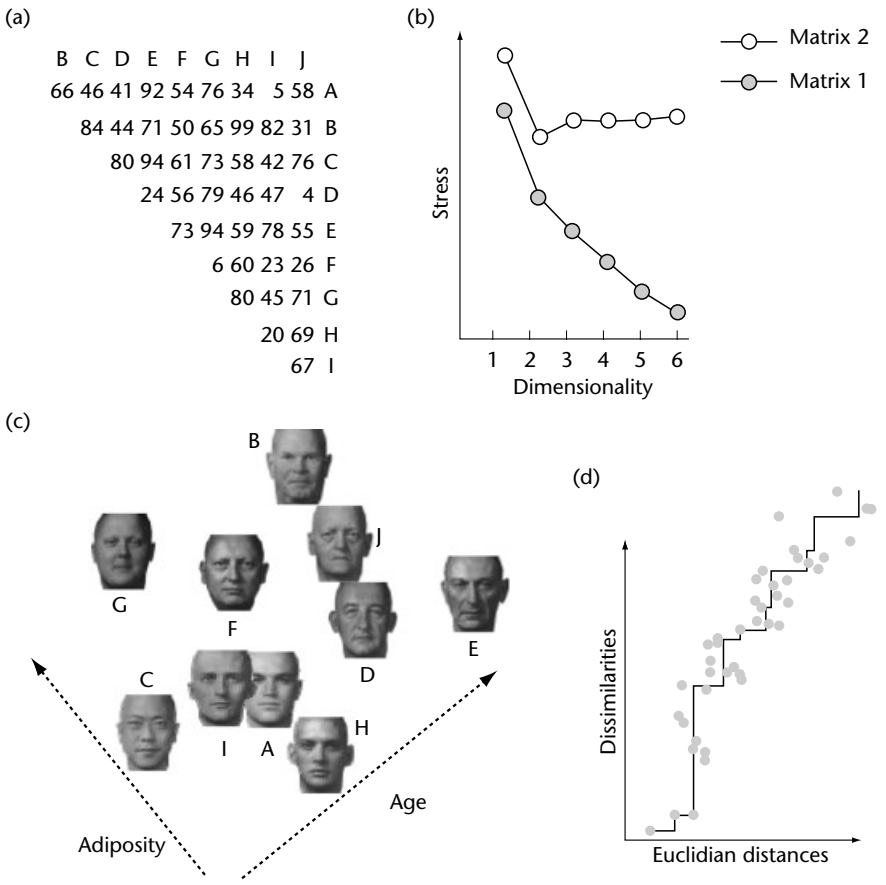


Figure 1. Analysis of a face similarity judgment task. (a) The average of a group of subjects' similarity ratings; (b) scree plot, for selecting dimensionality; (c) two-dimensional (age, adiposity) scaling solution for the 10 faces; (d) Shepard plot, showing the relationship between predicted distances and observed dissimilarities.

large matrices. In nonmetric MDS (first devised by Shepard in 1962), the goal is to establish a monotonic relationship between interpoint distances and obtained similarities. The advantage of nonmetric MDS is that no assumptions need to be made about the underlying transformation function; the only assumption is that the data are measured at the ordinal level.

Kruskal (1964) proposed a measure for the deviation from monotonicity between the distances d_{ij} and the observed dissimilarities o_{ij} called the *stress* function:

$$s = \sqrt{\frac{\sum_{ij} (d_{ij} - d_{ij}^*)^2}{\sum_{ij} {d_{ij}}^2}} \qquad (2)$$

Note that the observed dissimilarities o_{ij} do not appear in this formula. Instead, the discrepancy between the predicted distances d_{ij} and the target distances d_{ij}^* are measured. Based on the current configuration of points, the target distances d_{ij}^* are found by monotonic regression and represent the distances that are monotonically related to the observed dissimilarities o_{ij}. Several iterative minimization algorithms exist to move the object points in a multidimensional space in order to minimize stress (see Borg and Groenen, 1997).

In the face similarity example, Figure 1(d) displays what is known as the Shepard plot. It shows the relationship between predicted distances d_{ij} (for the two-dimensional scaling solution in Figure 1(c)) and observed dissimilarities as filled circles, and can serve to decide what metric transformation would be appropriate to relate one to the other. The line in the plot shows the relationship between the target distances d_{ij}^* found by monotonic regression and observed dissimilarities. Kruskal stress essentially is a measure based on the sum of the squared deviations between the filled circles and the line along the abscissa.

Another distinction in MDS is between weighted MDS, replicated MDS, and MDS on a single matrix

(Young and Hamer, 1994). In replicated MDS, several matrices of similarity data can be analyzed simultaneously. The matrices are provided by different subjects or by a single subject tested at multiple times, and a single scaling solution captures the similarity data of all matrices through separate metric or nonmetric relationships for each matrix. This approach can take individual differences in response bias into account. In weighted replicated MDS (e.g. INDSCAL: Carroll and Chang, 1970), the dimensions in the scaling solution can be weighted differently for each subject or subject replication, to model differences in attention or sensitivity for the different dimensions.

Finally, there is the distinction between deterministic and probabilistic MDS. In deterministic MDS, each object is represented as a single point in multidimensional space (e.g. Borg and Groenen, 1997) whereas in probabilistic MDS (MacKay, 1989), each object is represented as a probability distribution in multidimensional space. In understanding the mental representation of objects, this last approach is useful when representation of objects is assumed to be 'noisy' (i.e. the presentation of the same object on every trial gives rise to different internal representations).

An important issue in MDS is choosing the number of dimensions for the scaling solution. A configuration with a high number of dimensions achieves very low stress values but cannot easily be comprehended by the human eye, and is apt to be determined more by noise than by the essential structure in the data. On the other hand, a solution with too few dimensions might not reveal enough of the structure in the data.

A well-known method to select the dimensionality is the scree test (also known as the elbow test) where stress (or other lack-of-fit measure) is plotted against the dimensionality. Ideally, this choice is visually obvious from the 'elbow' in the scree plot where, after a certain number of dimensions, the stress is not reduced substantially. However, in many data sets, stress decreases smoothly with increasing dimensionality, making the choice of appropriate dimensionality very difficult with this method. In Figure 1(b), the filled circles show the scree plot for the face similarity data set. Note that a slight elbow is present at two dimensions, which suggests that a two-dimensional configuration might be appropriate.

A more salient indicator for the appropriate dimensionality can be obtained by cross-validation. The idea is to test how the configuration optimized to model the proximity data for one group of subjects can generalize to the proximity data of a different group of subjects. In Figure 1(b), the open circles show the stress value for a second group of subjects, with a clear rise in stress value after two dimensions while the stress continues to decrease for the first group of subjects. In this case, it seems reasonable to conclude that a two-dimensional configuration is appropriate because it can best generalize to other subjects. Lee (2001) has explored other techniques to determine dimensionality based on balancing the trade-off between model fit and model complexity.

Another important issue is the interpretation of the scaling solution resulting from MDS procedures. If the proximity data were generated by a function of the distances along some set of dimensions, then the resulting configuration of points in a scaling solution should reflect those dimensions. However, often Euclidian distances are used in scaling procedures so that the orientation of axes in the resulting configuration is arbitrary: any rotation of the axes would result in the same distances (and therefore stress). In such cases, the researcher can either visually scan the configuration in order to choose an orientation of axes that leads to interpretable results, or apply less arbitrary procedures by multiple regression analyses. In such analyses, the idea is to regress meaningful variables on the coordinates for the different dimensions and rotate the solution so as to maximize the interpretability. In Figure 1(c), the two-dimensional scaling solution is shown for the 10 faces. After visual inspection, the configuration can be interpreted as the perceptual dimensions of age and adiposity.

ADVANCES IN MDS

The success of the MDS approach arises in part from the simplicity of the underlying assumptions and the wide availability of computer software to create scaling solutions. Recent research has expanded the scope of the MDS approach in several directions. In Isomap (Tenenbaum *et al.*, 2000), stimuli are represented as points lying on a nonlinear manifold in some multidimensional space. Similarity is then computed as the geodesic distance on the manifold (i.e. the shortest distance along the manifold) as opposed to Euclidian distance in MDS. This technique is capable of discovering the nonlinear degrees of freedom that underlie complex data sets.

A drawback of most MDS algorithms is that a $N \times N$ matrix of similarity judgments is needed to scale N objects. Therefore, the number of similarity ratings needed depends quadratically on the number of objects, which leads to practical

limitations (e.g. subject time, number of subjects) on the number of objects that can be used in scaling studies. With modified MDS procedures, the amount of data that needs to be collected might be reduced. In the anchor point method (e.g. Buja *et al*., 1998), subjects rate all similarity pairs involving N objects and a smaller number of K anchor points that provide a representative sample of N objects. A modified MDS procedure then analyses the $N \times K$ similarity matrix in order to scale N objects. Future research will have to show how small K can become relative to N in order for this technique to make drastic savings possible in the similarity data collection.

Another recent advance in MDS models is the feature mapping approach (Rumelhart and Todd, 1992; Steyvers and Busey, 2000). In the traditional approach, the physical representation of the features comprising the stimuli is explicitly ignored. In such a purely top-down approach, the multidimensional representations are sometimes difficult to relate back to the physical stimulus by visual interpretation or regression analyses. In the feature mapping approach, in addition to the proximity data, additional physical measurement on the set of stimuli is available. The goal is to find a mapping between the set of physical measurements of a stimulus and the position of that stimulus in an abstract psychological space. With this approach, the multidimensional space is related directly to the physical dimensions of the stimuli.

CHALLENGES FOR MDS

Tversky and Hutchinson (1986) have argued that, for language-related stimuli that have conceptual as opposed to perceptual relations, geometric models based on MDS may fail to capture some aspects of the data and might therefore be inappropriate as a representational basis. For such stimuli, tree or graph-theoretic structures might be better suited than spatial/dimensional models based on MDS. Nevertheless, the geometric model of similarity has been applied to a wide variety of stimuli rich in conceptual structure. In 'latent semantic analysis' (Landauer and Dumais, 1997), words are placed in a high dimensional semantic space by a procedure related to MDS by analyzing the co-occurrence statistics for words appearing in contexts in a large corpus. In this semantic space, words with similar meaning are placed in nearby regions of the space. While the latent semantic analysis approach has been very successful in modeling human performance in a variety of semantic tasks, it remains to be seen to what degree a geometric model is appropriate to model language-related processes.

Also, for some highly structured objects, the simple geometric model for similarity as used in MDS may not be particularly revealing for analyzing the process of generating similarity judgments. In a geometric model, it is assumed that the set of objects can be described by a fixed collection of feature values where the process of generating similarity judgments is always based on the differences for the same set of features. However, for highly structured or complex objects, the features that play a role in the similarity judgments may differ depending on what objects are compared. In alignment models (Goldstone, 1994), similarity is assessed dynamically by an alignment process that analyzes which features play corresponding roles in the objects that are compared.

References

Attneave F (1950) Dimensions of similarity. *American Journal of Psychology* **63**: 516–556.

Borg I and Groenen P (1997) *Modern Multidimensional Scaling, Theory and Applications*. New York, NY: Springer-Verlag.

Buja A, Swayne DF, Littman M and Dean N (1998) XGvis: interactive data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*

Carroll JD and Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart–Young' decomposition. *Psychometrika* **35**: 238–319.

Goldstone RL (1994) Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**: 3–28.

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. I & II. *Psychometrika* **29**: 1–27, 115–129.

Landauer TK and Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 211–240.

Lee MD (2001) Determining the dimensionality of multidimensional scaling models for cognitive modeling. *Journal of Mathematical Psychology* **45**: 149–166.

MacKay DB (1989) Probabilistic multidimensional scaling: an anisotropic model for distance judgments. *Journal of Mathematical Psychology* **33**: 187–205.

Nosofsky RM (1992) Similarity scaling and cognitive process models. *Annual Review of Psychology* **43**: 25–53.

Rumelhart DE and Todd PM (1992) Learning and connectionist representations. In: Meyers D and Kornblum S (eds) *Attention and Performance*. Cambridge, MA: MIT Press.

Shepard RN (1962) Analysis of proximities: multidimensional scaling with an unknown distance function. I & II. *Psychometrika* **27**: 125–140, 219–246.

Shepard RN (1987) Towards a universal law of generalization for psychological science. *Science* **237**: 1317–1323.

Steyvers M and Busey T (2000) Predicting similarity ratings to faces using physical descriptions. In: Wenger M and Townsend J (eds) *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*. Makwah, NJ: Lawrence Erlbaum.

Tenenbaum JB, De Silva V and Lanford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500): 22.

Torgeson WS (1965) Multidimensional scaling of similarity. *Psychometrika* **30**: 379–393.

Tversky A and Hutchinson JW (1986) Nearest neighbor analysis of psychological spaces. *Psychological Review* **93**: 3–22.

Young FW and Hamer RM (1994) *Theory and Applications of Multidimensional Scaling*. Hillsdale, NJ: Lawrence Erlbaum.

Further Reading

Ashby FG (1992) *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Edelman S (1998) Representation is representation of similarity. *Behavioral and Brain Sciences* **21**: 449–498.

Medin DL, Goldstone RL and Gentner D (1993) Respects for similarity. *Psychological Review* **100**: 254–278.

Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* **210**: 390–398.

Tversky A (1977) Features of similarity. *Psychological Review* **84**: 327–352.

Music Cognition

Introductory article

Caroline Palmer, Ohio State University, Columbus, Ohio, USA
Melissa K Jungers, Ohio State University, Columbus, Ohio, USA

CONTENTS

Music cognition addresses the mental activities involved in perceiving, learning, remembering, and producing music.

INTRODUCTION

Music cognition addresses the mental activities involved in perceiving, learning, remembering, and producing music. Most of these mental activities are implicit and unconscious; even listeners without musical training can make sophisticated judgments about music. Psychological studies of music include perception of sound properties, understanding of pitch and rhythm, development of musical abilities, musical skill and performance, and parallels with language.

PERCEPTION OF SOUND PROPERTIES IN MUSIC

Four qualities of sound are especially important in music perception: pitch, duration, loudness, and timbre. Pitch arises from the perception of a sound wave's frequency (number of vibrations per second); loudness arises from the perception of amplitude or intensity (sound pressure change); duration is the perception of the time over which a sound is heard. Most musical instruments produce complex sound waves, which contain several frequencies (called harmonics) and amplitudes that change over time; timbre arises from the perception of these harmonics and their onsets (called attack transients). The lowest frequency in a complex tone, called the fundamental frequency, determines the perception of pitch.

UNDERSTANDING OF PITCH AND RHYTHM

Pitch and rhythm may be the most psychologically important dimensions of music. Most music in Western cultures, including classical, rock, and jazz styles, is based on 12 pitch classes, referred to as tones C, D, etc. (Music in most cultures uses pitch classes, but they are often tuned to different frequencies.) Tones whose frequencies differ by a factor of two, called an octave, are perceived as having the same pitch quality or 'chroma' and belong to the same pitch class. Each musical piece is composed primarily of tones from seven of the pitch classes, called a diatonic scale. Major and minor scales are defined for each pitch class by the intervals (number of pitch steps) between the tones. The key often refers to the pitch class from which the scale is built. There is a hierarchy of relative importance among tones in a key; tonality is the perception of pitches in relation to the tonic, the central prominent tone. Interestingly, each tone is perceived differently, depending on its key context. For example, the tone C is perceived as most important (the tonic) in the key of C and less important in the key of F.

Despite this complexity, listeners' understanding of pitch structure does not require explicit training. The statistical regularity with which tones, chords, and keys occur provide enough information for listeners to identify the tonic, or judge two tones as related. Tones that occur more often are perceived as more stable or less likely to change. Listeners can recognize a familiar melody based on the up-and-down pattern of pitch changes, called the melodic contour, even when the pitch classes have changed. With training, most listeners acquire the ability to name the pitch of a tone in relation to other pitches, called relative pitch. A few listeners have absolute pitch, or the ability to name pitches heard in isolation (without a reference pitch). Acquisition of absolute pitch may be related to the age at which musicians first learn to name pitches.

When listeners tap along with music, they are often responding to the rhythm: a temporal pattern that arises from many variables, including duration (from a tone's onset to its offset) and inter-onset interval (from one tone's onset to the next tone's onset). Meter and tempo also influence the perception of rhythm. Meter is a regular alternation of strong and weak beats in twos (such as marches) or threes (such as waltzes) at many hierarchical temporal levels. Tempo is the pace or rate of music; listeners show preferences for tempi around 300–900 ms per beat. People tend to tap and reproduce rhythms accurately at this tempo, suggesting that an internal pulse operating around this tempo guides perception and performance.

Music perception reflects a combination of pitch and rhythm at several hierarchical levels. Listeners tend to segment music into short groups: a group is a set of successive tones that are related. The perception of group boundaries is influenced by accents, or tones that stand out from others, caused by changes in pitch, duration, intensity, or timbre. Listeners also segment simultaneous or overlapping tones into perceptual parts or voices, called streams. Stream segregation is thought to reflect innate, 'bottom-up' perceptual processes, unaffected by general knowledge. Music perception is also influenced by 'top-down' knowledge: well-learned schemas, or acquired knowledge of tonal or metrical relationships. Schemas influence listeners' expectations and aid their memory for music. Schemas may also contribute to listeners' emotional response to music: schemas generate unconscious expectations for upcoming events, and musical events often occur that conflict with schematic expectations. Listeners experience arousal and an emotional response to the musically unexpected events.

DEVELOPMENT OF MUSICAL ABILITIES

Studies of musical development often focus on infants' responses, measured by attentiveness or listening preferences, because infants allow a test of which musical abilities are established prior to much musical exposure. Infants are typically exposed to certain types of music, including lullabies and play-songs sung by parents and other caregivers. Songs that adults sing to infants are higher in pitch, slower in tempo, and more expressive overall than the same songs sung to adults. By seven months old, infants are more attentive to these infant-directed songs than to other forms of singing.

Many basic perceptual abilities of infants resemble those of adults. For example, infants and adults are sensitive to the same pitch and rhythmic dimensions of music. Infants recognize the equivalence of pitches that are an octave apart. Around six months old, infants' babbling in response to music often preserves the melodic contour, and they respond the same way to melodies whose specific pitches are altered, as long as they retain the same melodic contour. By six months old, infants can produce body movements in response to the musical rhythm and they are sensitive to changes in tempo and rhythm. Finally, six-month-old infants prefer music that is segmented into natural phrase units more than music that is segmented in the middle of phrases (see example shown in Figure 1). Thus, infants attend to and group musical sequences according to the same perceptual principles as adults.

Other musical abilities develop with further musical exposure. By two to three years old, children spontaneously produce novel musical phrases in song, and can repeat short phrases accurately; by four or five years old, they can accurately imitate musical rhythms. By seven years old, children are sensitive to at least two metrical levels in musical rhythm; with musical training, adults show sensitivity to more metrical levels. Implicit knowledge of chord and key relationships also increases with musical exposure; the perception of hierarchical tonal relationships begins around the age of seven, but continues to develop into adulthood.

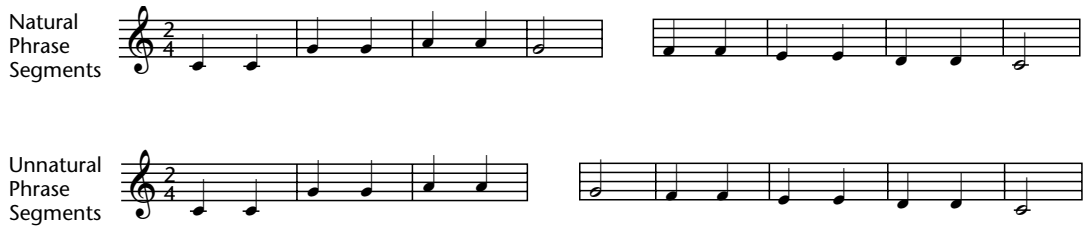


Figure 1. Example of natural (top) and unnatural (bottom) musical phrase segments in 'Twinkle, twinkle, little star'.

Interestingly, adults become less sensitive than infants to musical changes that preserve typical, schematic relationships of well-learned musical styles.

MUSICAL SKILL AND PERFORMANCE

Music performance is a rapid, fluent motor skill. Most people hum, clap, or perform in other ways that reflect sophisticated musical skills. Scientific studies of music performance, aided by computer-monitored musical instruments, focus on several cognitive and motor factors: segmenting and retrieving structural units from memory, coordinating motor movements, communicating structure and emotion with expressive features, and acquiring these skills.

Performers do not mentally prepare entire musical pieces at once, but instead segment them into smaller units such as phrases. In eye–hand span tasks, pianists are shown music notation briefly and then perform from memory; they recall musical segments that fit within short-term memory constraints. Errors in memorized performance indicate memory constraints on how far ahead performers can plan. Performance errors often result in tones that are similar to the intended tones in harmony, key, or meter. Thus, performers' musical knowledge reflects the same harmonic, diatonic, and rhythmic structures as listeners' knowledge.

After musical events are retrieved from memory, they must be transformed into appropriate movements. Internal clocks or timekeepers regulate and coordinate the production of different parts of a performance, such as hands in instrumental playing or performers in ensembles. Performers are more accurate at reproducing musical rhythms whose durations form simple ratios (1:1 or 2:1) that coincide with a simple internal timekeeper, than those whose durations form complex ratios (3:2 or 4:3). Musical motion is often compared to physical motion; for example, performers reduce the tempo near phrase boundaries at a rate similar to slowing down from a run to a walk.

Performers communicate musical structure and emotion to listeners through acoustic variations in frequency, amplitude, duration, and timbre, often termed 'expression'. Expression differentiates one performance from another, similar to how prosody (the rhythm and intonation of speech) differentiates one speaker from another. Expressive features of performance often coincide with structurally important events, such as metrical accents or phrase boundaries. Expressive features also signal emotional content; 'happy' tunes tend to be performed louder and faster than 'sad' tunes. Emotional content is linked to musical structure in performance: the expressive performance features that convey emotion tend to coincide with unexpected (less likely) structural features. (*See* **Prosody**)

How are performance abilities acquired? With short-term practice (hours), performers segment music into larger units, anticipate future events more, and increase expressive features. With long-term experience (years), performers monitor and adjust their performances better, refine expressive features, and extend their knowledge of how to perform one melody to unfamiliar melodies. Individual differences in performance skills are tied in part to accumulated practice over a lifetime. General memory and motor factors in performance increase most during the first few years of skill acquisition, whereas sensitivity to specific musical features such as phrase structure and meter increases across all skill levels.

PARALLELS BETWEEN MUSIC AND LANGUAGE

Music and language show interesting parallels in their structure and in how people perceive and produce them. They are found in all human cultures, and children spontaneously acquire both. Potentially universal mental activities in language and music include: auditory grouping/segmentation strategies; rhythmic structure in perception and production; and grammatical rules that influence understanding of music and language.

Listeners segment a continuous stream of sound into discrete units in speech and music. Two basic perceptual principles influence segmentation in both domains: first is sensitivity to acoustic changes or contrasts that define the boundaries of a unit, such as attack transients that occur at musical tone onsets or noise bursts at phoneme boundaries. Contrasts also mark larger units, such as changes in tempo (slowing down) that mark phrase boundaries in both music and speech. Second is sensitivity to periodic or repeating patterns, such as meter in music and in language. Memory limitations also influence segmentation: for example, the size of musical and verbal phrases is usually within short-term memory limits.

Rhythmic structure influences the relative importance that listeners attribute to individual elements in both music and speech, especially poetry. Tones (music) and syllables (speech) are accented in production with higher pitch, longer duration, and greater amplitude, which lead to the perception of rhythm. Music and speech commonly build

on a foundation of alternating strong and weak beats or pulses. Although human speech is not as rhythmically regular as most music, listeners tend to perceive alternating strong and weak stresses as rhythmically regular in both domains, and rhythmically regular music and language are better remembered.

Finally, humans have the capacity to understand an unlimited number of melodies or sentences: a grammar is a model of this capacity. Grammars, or a limited set of rules that generates an unlimited number of sequences, have been proposed for music as well as language. A compositional grammar can generate music in a particular style, such as counterpoint or jazz. A listening grammar applies a set of rules or implicit knowledge of the acceptable combinations of tones, chords, and rhythms in a musical style. The result is the perception of musical units such as phrases, hierarchical relationships among units, and harmonic or tonal tension and relaxation. A challenge for listening grammars is to accommodate different but equally correct segmentations of the same music. Thus, one difference between musical and linguistic grammars is that music is more flexible and less constrained in its grammar than language.

CONCLUSIONS

Scientific study of musical abilities has grown tremendously since the early 1990s, due to technological advances and theoretical overlap with related fields in cognitive science such as artificial intelligence, linguistics, neuroscience, and philosophy. Related topics include neural bases and computational models of musical behavior. Although music may be a specialized human ability, it is not special in its underlying perceptual, memory, and motor components.

Further Reading

Aiello R and Sloboda JA (eds) (1994) *Musical Perceptions*. New York, NY: Oxford University Press.

Butler D (1992) *The Musician's Guide to Perception and Cognition*. New York, NY: Macmillan.

Deutsch D (ed.) (1999) *The Psychology of Music*, 2nd edn. San Diego, CA: Academic Press.

Dowling WJ and Harwood DL (1986) *Music Cognition*. San Diego, CA: Academic Press.

Handel S (1989) *Listening*. Cambridge, MA: MIT Press.

Hargreaves D (1986) *The Developmental Psychology of Music*. Cambridge, UK: Cambridge University Press.

Jones MR and Holleran S (eds) (1992) *Cognitive Bases of Musical Communication*. Washington, DC: American Psychological Association.

Krumhansl CL (1990) *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Lerdahl F and Jackendoff R (1983) *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Meyer LB (1956) *Emotion and Meaning in Music*. Chicago, IL: University of Chicago Press.

Palmer C (1997) Music performance. *Annual Review of Psychology* **48**: 115–138.

Sloboda JA (1985) *The Musical Mind: The Cognitive Psychology of Music*. Oxford, UK: Clarendon Press.

Trehub S, Schellenberg G and Hill D (1997) The origins of music perception and cognition: a developmental perspective. In: Deliege I and Sloboda J (eds) *Perception and Cognition in Music*, pp. 103–128. Hove, UK: Psychology Press.

Naive Theories, Development of

Intermediate article

Susan A Gelman, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

A naive theory (also referred to as commonsense theory or folk theory) is a coherent set of knowledge and beliefs about a specific content domain (such as physics or psychology), which entails ontological commitments, attention to domain-specific causal principles, and appeal to unobservable entities.

INTRODUCTION

Children construct naive theories of physics, psychology, and biology by preschool age, with precursors in infancy. These theories provide a framework for children's developing knowledge, but also embody misconceptions that persist into adulthood.

ORIGINS OF NAIVE THEORIES AND DOMAIN KNOWLEDGE

Early in childhood, children begin to form bodies of knowledge, beliefs, and reasoning in several core domains, including language, physical objects, biological processes, psychological states, numerical reasoning, and spatial reasoning. What type of account best explains the development of such knowledge? The most prominent accounts are those referring to modules, to expertise, and to theories.

Modular approaches typically assume that domain-specific knowledge is innately constrained, biologically determined, and invariant in developmental outcome, and prime examples include the visual system or human language (syntax). Expertise approaches focus on the role of experience, attention, and practice in the gradual accrual of expertise in these domains, with attendant high variability in developmental outcomes. Example domains of expertise include reading, knowledge about dinosaurs, and chess; see, for example, Chi (1978). In contrast to both these positions, naive theories are characterized as undergoing important restructuring over time, and as informed (within broad constraints) by input and cultural beliefs. Candidate domains for theories that emerge early in childhood include psychology (Wellman, 1990), physics (McCloskey, 1983) and biology (Keil, 1994).

Evidence for naive theories (also known as commonsense or folk theories) includes ontological commitments, attention to domain-specific causal principles, appeal to unobservable entities, resistance to counterevidence, and coherence of beliefs (Carey, 1985). Ontological kinds reflect the basic categories of what sorts of entities there are in the world, and differ by domain (e.g. a folk theory of psychology concerns mental entities such as beliefs and desires, whereas a folk theory of physics concerns physical entities such as objects and substances). Naive theories also entail domain-specific causal explanations (e.g. the law of gravity cannot apply to mental states). Unobservable entities (e.g. gravity, beliefs, germs) are invoked to provide explanatory constructs that make sense of overt evidence (e.g. objects falling, people's actions, illness).

Naive theories are distinct from scientific theories in several major respects. Naive theories provide abstract frameworks for interpreting information, and thus are neither as precise nor as explicit as scientific theories. Thus, a child maintaining a naive theory of object motion and mechanical causation would be unable to state the assumptions or principles of the theory, although the child's expectations and predictions would be consistent with a set of implicit assumptions and principles. Indeed, the processes of reflection, awareness, and articulation appear to increase with development and themselves bring about theory change (Karmiloff-Smith, 1992). Naive theories are also not as accurate as scientific theories, though they may share striking similarities with them (e.g. Atran, 1990). Moreover, the process of naive theory construction

differs from the scientific process. Indeed, children have notorious difficulty with crucial scientific reasoning tasks, and for example fail to test systematically all possibilities to determine a causally relevant variable (Inhelder and Piaget, 1958). Changes in such capacities continue throughout childhood and into adulthood.

NAIVE PHYSICS

A naive theory of physics is concerned with the existence, movements, and interactions of physical objects, and the physical effects of such movements and interactions. Important components appear to be in place in early infancy. Four-month-old infants expect physical objects to move on paths that are connected (continuity constraint) and not to move through physical obstructions (solidity constraint). By 6 months of age, infants appreciate that the motion of inanimate objects requires direct physical contact. Other components develop later, with time and experience; for example, an expectation that objects will obey the law of gravity and a capacity to gauge the degree of contact needed for an object to be supported by a surface (e.g. Baillargeon *et al*., 1995).

Some of the experimental evidence that is interpreted as support for an early-developing naive physics stands in sharp contrast with traditional accounts of cognitive development, according to which infants are incapable of grasping even the most rudimentary physical principles. For example, Piaget's classic work on the object concept (Piaget, 1955) seemed to demonstrate that until about 12 months of age, children tend to assume that objects that disappear from sight no longer exist (a lack of 'object constancy'). In contrast, experiments indicate that infants as young as 3–5 months of age can track the existence of objects hidden behind a barrier and even reason about their size and numerosity (e.g. Baillargeon *et al*., 1995). These striking differences in research findings can be attributed to fundamental methodological differences: Piaget's work examined infants' overt search behaviors, include reaching and grasping; tasks finding early competence typically examine infants' gaze and dishabituation of gaze.

Naive physics continues to influence human reasoning throughout childhood and into adulthood. Adults appear to be remarkably resistant to input that would refute errors in their naive physical theories. For example, even those who have had college-level instruction in physics continue to err on some simple physical reasoning tasks, such as predicting the trajectories of balls rolling out of curved tubes (McCloskey, 1983).

NAIVE PSYCHOLOGY

A naive theory of psychology is often referred to as a 'theory of mind'. Interpreting behavior in a coherent, mentalistic framework is crucial for many tasks, including predicting and explaining the actions of others, engaging in deception, understanding and engaging in pretense, forming attachments to imaginary companions, and learning word meanings; see Wellman (1990) for a review. For example, if I know that you left your keys in the kitchen but I see you looking for them in the bedroom, then I can infer that you mistakenly believe that you left your keys in the bedroom.

Early accomplishments are found in children as young as 3 years of age, who accurately reason about mental states such as emotions and desires, understand links between perception and knowledge, and correctly distinguish mental from physical entities (Wellman, 1990). Between 3 years and 5 years of age, there is an important shift in children's reasoning about false beliefs. For example, consider a scenario in which the child is shown that a crayon box has unexpected contents (a toy truck instead of crayons), but another person is not shown the contents of the box. In other words, the child has access to information that the other person does not. When presented with a scenario of this sort, 3-year-old children typically make the error of attributing knowledge to the ignorant person (stating that the other person will say that the crayon box contains a toy truck), erroneously assuming that the person's beliefs will match objective reality. In contrast, 5-year-old children recognize that the other person does not have the same knowledge state, thus appropriately recognizing that subjective beliefs may conflict with objective reality. By 5 years of age, children are skilled in reasoning about a wide array of psychological tasks, and recognize that even potent mental states (e.g. visual imagery, dreams) are distinct from physical reality, that thoughts are subjective, and that psychological entities (in the form of beliefs and desires) explain and predict people's actions (Wellman, 1990). An exception to these general findings is that autistic children are considerably impaired in their naive psychological reasoning (Baron-Cohen *et al*., 1993).

Even infants appreciate some components of a naive psychology. By 18 months of age, infants distinguish intentional from accidental action when imitating an adult model (Meltzoff, 1995)

and learning a new word (Tomasello and Barton, 1994), and appreciate the subjectivity of desires and their link to emotions (Repacholi and Gopnik, 1997). Together, these findings contradict traditional claims by Piaget (1929) that young children are realists and are confused by the distinction between mental and physical phenomena.

NAIVE BIOLOGY

A naive theory of biology is concerned with classification of the world of living things (animals and plants), reasoning about biological processes such as growth, illness and reproduction, and distinguishing living from nonliving entities. There is controversy about when a naive theory of biology develops. Some have proposed that biology emerges as a new domain, out of psychology, in middle childhood (Carey, 1985). According to this view, young children account for biological processes by appealing to psychological explanations (explaining that people eat because they are hungry, rather than because their body needs nutrients). In contrast, others argue that biology is an independent theory domain from preschool age or earlier (Keil, 1994).

This debate seems to reflect the incomplete state of children's biological knowledge and thus leads to arguments over what constitutes a theory. At the very least, by preschool age children have certain clear understandings of how biological entities are distinct from both psychological entities and inanimate objects. For example, 4-year-old children recognize that biological processes (e.g. growth, heartbeat) are not under direct conscious control (Inagaki and Hatano, 1993) and that physical illnesses are contagious in a way that psychological disorders are not (Keil, 1994). Furthermore, preschoolers appreciate that there are numerous processes that apply only to living things, including growth, self-generated motion, illness, healing, and teleological action (Gelman and Opfer, 2002). Further evidence that naive biology is operating as a domain for young children is that preschoolers appeal to nonobvious constructs (namely, germs and invisible particles) to explain the biological processes of contamination and illness (e.g. Kalish, 1996). Furthermore, by 3 years of age children treat living kind categories as if they have a category essence (Gelman *et al*., 1994). Much less is known about biological reasoning in children below preschool age, including infants and toddlers.

As with naive theories of physics, naive theories of biology give rise to predictable errors and biases, including a tendency toward vitalism (Inagaki and Hatano, 1993) and creationism (Evans, 2000), and the persistence of certain folk-biological categories that have no scientific counterpart (Atran, 1990).

References

Atran S (1990) *Cognitive Foundations of Natural History*. Cambridge, UK: Cambridge University Press.

Baillargeon R, Kotovsky L and Needham A (1995) The acquisition of physical knowledge in infancy. In: Sperber D, Premack D and Premack AJ (eds) *Causal Understanding: A Multidisciplinary Debate*, pp. 79–116. New York, NY: Clarendon Press.

Baron-Cohen S, Tager-Flusberg H and Cohen DJ (eds) (1993) *Understanding Other Minds: Perspectives From Autism*. New York, NY: Oxford University Press.

Carey S (1985) *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chi MTH (1978) Knowledge structure and memory development. In: Siegler R (ed.) *Children's Thinking: What Develops?* pp. 73–96. Hillsdale, NJ: Lawrence Erlbaum.

Evans EM (2000) Beyond Scopes: why creationism is here to stay. In: Rosengren KS, Johnson CN and Harris PL (eds) *Imagining the Impossible: Magical, Scientific, and Religious Thinking in Children*. New York, NY: Cambridge University Press.

Gelman SA and Opfer J (2002) Development of the animate–inanimate distinction. In: Goswami U (ed.) *Handbook of Childhood Cognitive Development*. London, UK: Blackwell.

Gelman SA, Coley JD and Gottfried GM (1994) Essentialist beliefs in children: the acquisition of concepts and theories. In: Hirschfeld LA and Gelman SA (eds) *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 341–365. New York, NY: Cambridge University Press.

Inagaki K and Hatano G (1993) Young children's understanding of the mind–body distinction. *Child Development* **64**: 1534–1549.

Inhelder B and Piaget J (1958) *The Growth of Logical Thinking From Childhood to Adolescence*. New York, NY: Basic Books.

Kalish CW (1996) Causes and symptoms in children's understanding of illness. *Child Development* **67**: 1647–1670.

Karmiloff-Smith A (1992) *Beyond Modularity*. Cambridge, MA: MIT Press.

Keil FC (1994) The birth and nurturance of concepts by domains: the origins of concepts of living things. In: Hirschfeld LA and Gelman SA (eds) *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York, NY: Cambridge University Press.

McCloskey M (1983) Intuitive physics. *Scientific American* **248**: 122–130.

Meltzoff AN (1995) Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology* **31**: 838–850.

Piaget J (1929) *The Child's Conception of The World*. London, UK: Routledge & Kegan Paul.
Piaget J (1955) *The Child's Construction of Reality*. London, UK: Routledge & Kegan Paul.
Repacholi BM and Gopnik A (1997) Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental Psychology* **33**: 12–21.
Tomasello M and Barton M (1994) Learning words in nonostensive contexts. *Developmental Psychology* **30**: 639–650.
Wellman HM (1990) *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Further Reading

Au TK and Romo LF (1999) Mechanical causality in children's 'folkbiology'. In: Medin DL and Atran S (eds) *Folkbiology*. Cambridge, MA: MIT Press.
Au TK, Sidle AL and Rollins KB (1993) Developing an intuitive understanding of conservation and contamination: invisible particles as a plausible mechanism. *Developmental Psychology* **29**: 286–299.
Baldwin DA (1993) Early referential understanding: infants' ability to recognize referential acts for what they are. *Developmental Psychology* **29**: 832–843.
Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
Gopnik A and Wellman HM (1994) The theory theory. In: Hirschfeld LA and Gelman SA (eds) *Domain Specificity in Cognition and Culture*. New York, NY: Cambridge University Press.
Kim I and Spelke ES (1999) Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science* **2**: 339–362.
Klahr D, Fay A and Dunbar K (1993) Heuristics for scientific experimentation: a developmental study. *Cognitive Psychology* **25**: 111–146.
Kuhn D (1989) Children and adults as intuitive scientists. *Psychological Review* **96**: 674–689.
Lewis M, Stanger C and Sullivan MW (1989) Deception in 3-year-olds. *Developmental Psychology* **25**: 439–443.
Lillard AS (1993) Pretend play skills and the child's theory of mind. *Child Development* **64**: 348–371.
Medin D (1989) Concepts and conceptual structure. *American Psychologist* **44**: 1469–1481.
Pillow BH (1989) Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology* **47**: 116–129.
Premack D and Woodruff G (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* **4**: 515–526.
Rosen AB and Rozin P (1993) Now you see it, now you don't: the preschool child's conception of invisible particles in the context of dissolving. *Developmental Psychology* **29**: 300–311.
Siegal M (1988) Children's knowledge of contagion and contamination as causes of illness. *Child Development* **59**: 1353–1359.
Spelke ES, Phillips AT and Woodward AL (1995) Infants' knowledge of object motion and human action. In: Sperber D, Premack D and Premack AJ (eds) *Causal Understanding: A Multidisciplinary Debate*, pp. 44–78. New York, NY: Clarendon Press.
Taylor M (1999) *Imaginary Companions and the Children Who Create Them*. New York, NY: Oxford University Press.
Wellman HM and Gelman SA (1998) Knowledge acquisition. In: Kuhn D and Siegler R (eds) *Handbook of Child Psychology*, 5th edn, *Cognitive Development*, pp. 523–573. New York, NY: John Wiley.
Wynn K (1992) Addition and subtraction by human infants. *Nature* **358**: 749–750.

Natural Kinds and Artifacts

Intermediate article

Susan A Gelman, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

Natural kinds are categories of things found in nature (e.g. living things, natural substances); artifacts are categories of things created by humans (e.g. furniture, clothing). These two distinct sorts of concepts differ in their implications for human reasoning.

INTRODUCTION

Natural kind concepts and artifact concepts differ from one another in several important respects, with consequences for human reasoning. These differences are found in young children as well as adults.

PHILOSOPHICAL AND PSYCHOLOGICAL DISTINCTION

Human concepts are not all alike, and the ways in which they differ have important implications for human reasoning. One fundamental distinction is based on content domain, specifically, natural kinds versus artifacts. Natural kinds are categories of naturally occurring things (including animals, plants, naturally occurring substances such as gold, and natural formations such as stars); artifacts are categories of entities created by humans (e.g. tools, vehicles, clothing, toys, etc.). The distinction between natural kinds and artifacts rests in the *category* under consideration, and does not apply to objects *per se*. For example, a yellow, metal circle can be construed as either a substance (gold) or an object (ring). It is only when one considers the category to which the object is assigned that one can determine whether it is a natural kind (gold) or artifact (ring).

The distinction between natural kinds and artifacts has implications beyond content alone, although these differences are of degree and not absolute. Natural kind categories are hypothesized to differ from artifact categories in at least four major respects (see Schwartz, 1977). First, natural kinds have *rich inductive potential*, so that we can infer indefinitely many novel properties about members of a natural kind, based on information about just one or a few instances (Gelman, 1988). Thus, for example, upon learning a set of facts about one tapir (what it eats, where it sleeps, the number of bones in its skeleton, its resting body temperature, its method of hunting prey, the organization of the structures in its brain, the location of its spleen, etc.), we readily infer that other tapirs are likely to have the same properties. Although artifacts also have inductive potential, it tends to be more limited (consider, for example, the inductive potential of the categories CUP or HAT). Second, natural kinds are treated as having a hidden, nonobvious *essence* that is responsible for the other properties that category members share (e.g. DNA for animals; H_2O for water; Medin, 1989), and that enables constancy over transformations (e.g. from infant to adult; from caterpillar to butterfly). Artifacts do not possess an intrinsic physical essence in the same respect, although the intentions of the person who designed the object can be considered crucial to artifact identity (Bloom, 2000). Third, natural kinds are named in accord with a *division of linguistic labor*, such that meanings are not fully contained within a speaker but are determined in part by a community of experts. Thus, most individual speakers cannot distinguish diamonds from zircons, but instead rely on gem specialists (Putnam, 1975). Some artifacts also have such a division of labor (e.g. carburetor), although others do not (e.g. ball). Fourth, natural kind categories are universally organized into *taxonomic hierarchies* (e.g. beagles are a kind of dog; dogs are a kind of mammal; mammals are a kind of animal). Artifacts appear to fit less neatly into taxonomic hierarchies (e.g. Atran, 1990; but see Rosch *et al*., 1976).

Altogether, these domain distinctions reflect ordinary speakers' implicit assumptions that natural kind categories are 'real' (e.g. one discovers a natural kind category, such as tiger) whereas

categories of artifacts are artificial (e.g. one invents an artifact category, such as shoe). However, both natural kind and artifact categories are human constructions, and the properties listed above are *psychological* claims concerning these human constructions – not metaphysical claims concerning the structure of the world.

DIFFERING INDUCTIVE POTENTIAL

Induction is one of the central functions of categories. All categories permit inductive inferences concerning novel properties, which in turn allow us to expand our knowledge about the world. Even preverbal infants appreciate the inductive potential of categories (Baldwin *et al*., 1993). None the less, one of the most striking differences between natural kinds and artifacts is that the former allow a greater extent of inductive inferences, including inferences concerning non-obvious properties and inferences to atypical category members.

When research participants are given a choice between extending a novel fact on the basis of membership in a natural kind category versus outward appearances, they tend to favor category membership (Gelman and Markman, 1986). For example, adults learned that a particular flamingo has a right aortic arch and that a particular bat has a left aortic arch (novel and unfamiliar properties). When asked about the cardiac structure of a blackbird (physically more similar to the bat than to the flamingo), they inferred that it has a right aortic arch, like the flamingo, as both are members of the category 'bird'. Similarly, when 4-year-olds received the same task, but with simpler properties (e.g. the flamingo 'feeds its baby mashed-up food'; the bat 'feeds its baby milk'), they too tend to favor category membership. Further studies indicate that children as young as 2 and 3 years of age expect category members to share important, non-obvious similarities, even in the face of salient perceptual dissimilarities. For example, upon learning that an atypical exemplar is a member of a category (e.g. that a flying fish is a fish), children and adults draw novel inferences from typical instances to the atypical member (Gelman and Markman, 1986). By 7 to 8 years of age, children draw more inductive inferences within natural kind than artifact categories (Gelman, 1988).

ESSENTIALISM AND PSYCHOLOGICAL ESSENTIALISM

Philosophers have long suggested that members of a category share a non-obvious, immutable core (or essence) that confers identity and predicts a vast array of other properties. On this view all humans, for example, have something deep and hidden in common that makes them human. Specific essentialist construals can be found in concepts as divergent as *soul* and *DNA*, though essentialism may also be an unarticulated, placeholder notion – a belief *that* a category has a core, without knowing *what* that core is (Medin, 1989). For example, a child might believe that girls have some inner, non-obvious quality that distinguishes them from boys and that is responsible for the many observable differences in appearance and behavior between boys and girls, before ever learning about chromosomes or human physiology. Thus, psychological essentialism requires no specialized knowledge.

It is important to distinguish between essentialism as a philosophical position and psychological essentialism as a folk belief. The former addresses the nature of reality (a metaphysical question); the latter addresses the nature of people's ordinary belief systems (a psychological question). Psychological essentialism is a human reasoning bias that may not accurately characterize biological species (e.g. Sober, 1994). There is widespread agreement that essentialist construals of race, though ubiquitous, are deeply flawed (see Hirschfeld, 1996; Templeton, 1998).

Experimental studies of essentialism have documented that natural kind categories have the following properties: a non-obvious basis (natural kind categories are based on 'hidden' features that can yield classifications that contradict surface appearances, e.g. leading us to classify whales as mammals, to classify legless lizards as lizards instead of snakes, and to consider caterpillars and butterflies to be different stages of the same kind of animal), anomalous instances (e.g. an albino, dwarf tiger is still a tiger, even though it more closely resembles a domesticated cat), and constancy of identity over transformations.

One of the strongest pieces of evidence for the latter comes from Keil (1989), who asked research participants to consider hypothetical scenarios such as the following: an entity can be altered through surgical means so that in appearance and behavior it in no way resembles its original incarnation. For example, a raccoon is (hypothetically) operated upon so that it resembles a skunk in every outward respect (size, coloration, shape, odor, behavior). By 7 to 8 years of age, children who hear such stories report that the animal is still a raccoon, despite its appearance and behavior. Importantly, such transformation stories yield very different

results for artifacts. Thus, a coffee pot transformed into a bird feeder is reported to change in fact into a bird feeder. Adults likewise treat the animal transformations as immaterial, but the artifact transformations as affecting identity.

Further evidence for an essentialist bias can be found in studies indicating that young children (in some studies, as young as 4 years of age) judge nonvisible internal parts to be especially crucial to the identity and functioning of an item. These expectations are evident in children's concept learning, judgments of identity, and word learning (see Bloom, 2000, for review).

The finding that young children hold essentialist beliefs thus suggests that human concepts are not constructed atomistically from perceptual features. More generally, these findings illustrate the independence of similarity and category membership in human reasoning (Smith *et al*., 1998).

DEVELOPMENT OF SENSITIVITY

The possible roots of reasoning about natural kinds, particularly essentialism, have been much disputed. Some investigators have suggested that essentialism is a by-product of Western philosophy (dating back to Plato in the fourth century BC) or Western cultural traditions (Rorty, 1979), or a reflection of access to scientific processes and knowledge available to the lay audience (e.g. given the discovery of microscopic organisms, DNA, and other nonobvious entities). In contrast to these positions, the evidence briefly summarized above suggests that children are essentialists long before formal schooling. It is therefore unlikely that either Western philosophy or scientific theories are prerequisites to essentialism. The developmental findings provide particularly strong evidence, given children's well-documented focus on object appearances on many cognitive tasks (Inhelder and Piaget, 1964).

Recent research has begun to examine the roots of the natural kind–artefact distinction by examining a related distinction in infancy, the animate–inanimate distinction (see Rakison and Poulin-Dubois, 2001, for review). The animate–inanimate distinction appears fundamental in brain organization (Caramazza and Shelton, 1998), and may be the basis of an eventual distinction between natural kinds and artifacts. Preverbal infants distinguish animals from inanimate objects, making use of both featural information (e.g. faces, contour) and dynamic information (e.g. autonomous motion, contingent motion). The distinction is then recruited in infants' conceptual understanding of the world, with respect to socio-emotional understandings, theory of mind, and predictions of actions (see Gelman and Opfer, 2002, for review). In older children, the animate–inanimate distinction guides children's reasoning about a vast range of concepts, including contagion, teleology, language, and theory of mind (Gelman and Opfer, 2002). Thus, the animate–inanimate distinction, like the natural kind–artifact distinction, organizes much of human knowledge.

References

Atran S (1990) *Cognitive Foundations of Natural History: Towards an Anthropology of Science*. Cambridge, UK: Cambridge University Press.

Baldwin DA, Markman EM and Melartin RL (1993) Infants' ability to draw inferences about nonobvious object properties: evidence from exploratory play. *Child Development* **64**: 711–728.

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Caramazza A and Shelton JR (1998) Domain-specific knowledge systems in the brain: the animate–inanimate distinction. *Journal of Cognitive Neuroscience* **10**: 1–34.

Gelman SA (1988) The development of induction within natural kind and artifact categories. *Cognitive Psychology* **20**: 65–95.

Gelman SA and Markman EM (1986) Categories and induction in young children. *Cognition* **23**: 183–209.

Gelman SA and Opfer J (2002) Development of the animate–inanimate distinction. In: Goswami U (ed.) *Handbook of Childhood Cognitive Development*. Oxford, UK: Blackwell.

Hirschfeld L (1996) *Race in the Making*. Cambridge, MA: MIT Press.

Inhelder B and Piaget J (1964) *The Early Growth of Logic in the Child*. New York, NY: Norton.

Keil F (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: Bradford/MIT Press.

Medin DL (1989) Concepts and conceptual structure. *American Psychologist* **44**: 1469–1481.

Putnam H (1975) The meaning of 'meaning'. In: Putnam H (ed.) *Mind, Language, and Reality: Philosophical Papers*, vol. 2. New York, NY: Cambridge University Press.

Rakison DH and Poulin-Dubois D (2001) The developmental origin of the animate–inanimate distinction. *Psychological Bulletin* **127**: 209–228.

Rorty R (1979) *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press.

Rosch E, Mervis C, Gray W, Johnson D and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology* **8**: 382–439.

Schwartz SP (ed.) (1977) *Naming, Necessity, and Natural Kinds*. Ithaca, NY: Cornell University Press.

Smith EE, Patalano AL and Jonides J (1998) Alternative strategies of categorization. *Cognition* **65**: 167–196.

Sober E (1994) *From a Biological Point of View*. New York, NY: Cambridge University Press.

Templeton AR (1998) Human races: a genetic and evolutionary perspective. *American Anthropologist* **100**: 1–19.

Further Reading

Brace CL (1964) A nonracial approach towards the understanding of human diversity. In: Montagu A (ed.) *The Concept of Race*. New York, NY: The Free Press.

Diesendruck G, Gelman SA and Lebowitz K (1998) Conceptual and linguistic biases in children's word learning. *Developmental Psychology* **34**: 823–839.

Dupré J (1993) *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.

Hirschfeld LA and Gelman SA (1994) *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York, NY: Cambridge University Press.

Keil FC (1995) The growth of causal understanding of natural kinds: modes of construal and the emergence of biological thought. In: Premack A and Sperber D (eds) *Causal Cognition*. Oxford, UK: Oxford University Press.

Kornblith H (1993) *Inductive Inference and Its Natural Ground: An Essay in Naturalistic Epistemology*. Cambridge, MA: MIT Press.

Mayr R (1991) *One Long Argument: Charles Darwin and the Genesis of Modern Evolutionary Thought*. Cambridge, MA: Harvard University Press.

Osherson DN, Smith EE, Wilkie O *et al*. (1990) Category-based induction. *Psychological Review* **97**: 185–200.

Rips LJ (1989) Similarity, typicality, and categorization. In: Vosniadou S and Ortony A (eds) *Similarity and Analogical Reasoning*, pp. 21–59. New York, NY: Cambridge University Press.

Rothbart M and Taylor M (1990) Category labels and social reality: do we view social categories as natural kinds? In: Semin G and Fiedler K (eds) *Language and Social Cognition*. London, UK: Sage.

Neuropsychological Development

Introductory article

Michelle de Haan, University College London, London, UK
Mark H Johnson, Birkbeck College, London, UK

CONTENTS

Neuropsychological development concerns the growth in the structure and function of the brain and nervous system in relation to mind and behavior, and has been shown to involve an interaction of genetic inheritance and environmental experience.

INTRODUCTION

Humans' complex intellectual abilities are thought to be mediated by the extraordinary functional capacity of the cerebral cortex. It is often assumed that maturation of the cerebral cortex causes or allows specific advances in cognitive, perceptual, or motor abilities in infants and children. However, brain development is not merely the unfolding of a genetic plan that imposes itself on behavior in such a direct way. A striking characteristic of the development of the human brain is that it continues for a long period after birth. As a consequence, development is a period of plasticity during which the environment and the child's interactions in it can influence later phases of brain development. Thus, normal development of the brain depends on the interaction of genetic inheritance and environmental experience. (*See* **Early Experience and Cognitive Organization**)

THE ADULT CEREBRAL CORTEX

The cerebral cortex is a thin, flat sheet of cells (about 3–4 mm thick) which becomes increasingly convoluted, or folded, with both phylogenetic and ontogenetic development. Fitting the increasingly large brain into the skull causes the greater surface area to be compressed, like a crumpled sheet of paper, into that space. In humans, the cortex consists of six layers, each of which contains particular types of cells and has particular patterns of inputs and outputs. For example, much of the information perceived by our senses enters the cortex in layer IV. The mature cortex can also be divided into distinct functional areas based on variations in the cytoarchitectonics, or cell type and structure. The six layers are not equally thick in each of these areas, but rather each can vary in thickness according to the region's function. For example, layer IV, which receives sensory inputs, is thicker in primary visual cortex than in motor cortex.

The development of the cerebral cortex can be divided into four processes: (a) proliferation, or the dividing of cells to make new cells; (b) migration, or the movement of cells to their appropriate positions; (c) differentiation, or the acquisition by each cell of its unique characteristics such as shape and neurotransmitter type; and (d) myelination, or the covering of the neuron with a fatty sheath that speeds its communication of signals. As we shall see, these processes involve not only additive events, but also regressive events. (*See* **Neural Development**)

PRENATAL DEVELOPMENT OF THE CEREBRAL CORTEX

Proliferation of Neurons

Development of the nervous system begins from just a few cells in a sheet of tissue called the neural plate. By the third or fourth week of development, the edges of the neural plate begin to fold in to form a hollow structure called the neural tube. The ventricles of the brain and the spinal canal develop

from the hollow area inside the tube, while all of the cells that make up the brain develop from the layer of cells that line the tube. These neurons begin to rapidly divide, or proliferate, at approximately the sixth fetal week and they continue to divide until approximately the eighteenth week. It is commonly believed that after this time no new cortical neurons are ever formed, although some recent studies challenge this view and indicate that a relatively small number of neurons may continue to be formed in several areas of the adult cortex. (*See* **Neurogenesis**)

Neuron Migration

Neurons are not born in the exact position they will occupy in the adult cortex. Instead they must travel or migrate from the proliferative zone where they are formed to the position they will occupy in the mature cortex. This occurs through an active form of migration in which young cortical cells migrate past previously created cells towards the surface of the brain. Thus, neurons born earliest remain in the deepest layers of the cortex while neurons born later take positions closer to the surface, creating an 'inside-to-outside' pattern. Neurons find their way to the correct position with the help of support cells called radial glia, whose long fibres act like a climbing rope. Disruption of neuron migration can lead to disorders of brain structure and function. For example, in the migration disorder called lissencephaly, the cortex is abnormally thick, has only four layers, and shows an absence or decrease of surface folding. This condition is associated with profound mental retardation, seizures, and a short lifespan.

Over-production of Neurons and Programmed Cell Death

It is estimated that the average number of neurons in the adult human cortex is 21 billion. Yet during the course of development the number of neurons reaches an even higher level than that seen in adulthood. The normal development of the nervous system involves the programmed cell death of neurons to remove the excess cells. Certain neuron populations die through a process called apoptosis. In apoptosis, cells die as part of a gene expression-related program of cell differentiation. The balance within the cell of expression of 'pro-death' and 'anti-death' genes determines whether or not it will survive. Apoptosis is characterized by cell shrinkage, and thus is different from the process of necrotic cell death that occurs following brain damage, which is characterized by cell swelling.

There are several reasons why the seemingly wasteful process of programmed cell death may occur. One is that certain neurons may serve only a temporary purpose and die when they are no longer needed. For example, in the mammalian cerebral cortex, neurons located below the area where cortical neurons are forming provide a temporary target for subcortical neurons carrying sensory information. When the cortical neurons have formed and migrated to their proper positions, the subcortical projections retract their temporary connections to form new ones with these cortical neurons. The temporary target neurons are no longer needed and so are eliminated by programmed cell death. Another purpose for cell death is to allow the number of neurons innervating (providing input to) a target to match the size of the target. How many neurons survive is thought to be controlled by competition for a limited supply of growth factors supplied by their target tissues. A small target will produce a smaller amount of chemicals to promote the survival of neurons and thus would support relatively few neurons; in contrast, a larger target would produce more and thus would support more neurons. In this way, the number of neurons innervating a target can be precisely matched without this information having to be explicitly preprogrammed. For example, developing sympathetic neurons require a chemical called nerve growth factor for survival and die in its absence. Thus, targets could control their amount of innervation by sympathetic neurons through the amount of nerve growth factor they produce.

Programmed cell death is a normal part of prenatal brain development. There is some evidence that the elimination of surplus neurons may continue even into the first years of postnatal life; however, this is a continuing area of investigation. Inappropriate activation of the pathway for cell death in adults and children has been implicated in some neurological diseases, such as Parkinson's Disease and spinal muscular atrophy (an inherited childhood motoneuron disease).

POSTNATAL DEVELOPMENT OF THE CEREBRAL CORTEX

Once neurons have migrated to their final positions, they further differentiate to take on their mature characteristics. The differentiation of cortex into distinct areas or regions is not simply due to the reading off of a genetic 'map' in the cortical cells themselves, but rather is heavily influenced by factors outside the cortex.

Dendritic Arborization

The dendrites of a neuron are like antennae, picking up signals from many neurons and passing the signal down the axon and on to other neurons. The pattern of branching of dendrites is important, because it will affect the quantity and quality of signals the neuron receives. During cortical development one change that occurs is an increase in size and complexity of neurons' dendritic trees. Dendrites typically begin this process of differentiation once they have completed migration and are in their final position within a cortical layer. For example, by adulthood the length of the dendrites of neurons in the frontal cortex can increase to over 30 times their length at birth. Dendritic branching occurs at different times in different areas and layers of the cortex. For example, dendritic trees of cells in layer V of primary visual cortex are already at about 60 per cent of their maximum extent at birth; in contrast, the mean total length for dendrites in layer III is only at about 30 per cent of maximum at birth. Dendrite branching is one of the properties of neurons that can be influenced by the environment during development. For example, laboratory animals that receive environmental stimulation show more dendritic branching than those that do not.

Myelination

Myelin is a fatty sheath that surrounds neuronal pathways and increases the efficiency of neural transmission. The process by which this fatty sheath is formed around neurons, called myelination, begins before birth and continues for many years after birth. Studies of post-mortem brains and more recent studies using magnetic resonance imaging to visualize myelination indicate that not all regions of the brain are myelinated at the same time and the same rate. For example, myelination in cortical areas that process sensory information tends to occur earlier than myelination in areas that process motor information. The very last areas to complete myelination are the tracts of the cortical association areas, where this process is not complete until 20–30 years of age. Several investigators have noted that the areas that tend to myelinate first are those whose functions emerge earliest in life. For example, the visual cortical pathways myelinate relatively early, and visual function develops rapidly in the first year of life; in contrast, frontal cortical pathways myelinate later and over a more protracted period, and frontal cognitive functions develop over a more protracted period into adolescence. While myelination does facilitate synaptic transmission, under-myelinated connections in the young human brain are still able to transmit signals. Thus, even before myelination is completed, cortical areas are likely to be at least partially functional. Delays in myelination can occur during development, and can be associated with delayed maturation of function: for example, delayed myelination of motor cortex is associated with a delayed acquisition of motor milestones.

Synapse Overproduction

Synapses are the regions of communication between neurons. Studies of post-mortem tissue indicate that there is a steady developmental increase in the density of synapses in several regions of the human cerebral cortex to levels even higher than those observed in adults. This period of synapse overproduction is believed to play an important role in the enhanced plasticity, or sensitivity to experience, of the cortex during development. (*See* **Synapse**)

An increase in synaptogenesis begins around the time of birth for all cortical areas studied to date, but, at least in humans, the timing of the most rapid burst of synapse formation and the final peak density occur at different ages in different cortical areas. In the visual cortex there is a rapid burst of synapse formation at 3 to 4 months, and the maximum density of around 150 per cent of adult level is reached between 4 and 12 months. A similar time course is observed in the primary auditory cortex (Heschl's gyrus). Synaptogenesis also starts at the same time in a region of the prefrontal cortex; however, the density increases much more slowly and does not reach its peak until after the first year. Because the period of synapse overproduction is thought to be related to periods of developmental plasticity to experience, this pattern suggests that different cortical regions and the behavioral functions they mediate are most sensitive to environmental inputs at different times.

The differential time course of development of different cortical regions is also observable in the living human brain using a measure of the metabolic activity of the brain called positron emission tomography (PET). In infants under 5 weeks of age glucose uptake is highest in sensorimotor cortex, thalamus, brain stem, and the cerebellar vermis, while by 3 months of age there are considerable rises in the parietal, temporal, and occipital cortices, basal ganglia, and cerebellar cortex. Maturational rises are not found in the frontal and dorsolateral occipital cortex until approximately 6–8 months. These measures, like the measures of

synapse density, also show an increase above adult levels. There is a sharp rise in overall resting brain metabolism (glucose uptake) after the first year of life, with a peak of approximately 150 per cent of adult levels achieved somewhere around 4 to 5 years of age for some cortical areas. While the overall level remains above the adult level until 4 to 5 years, an adult-like *distribution* of resting activity within and across brain regions is observed by the end of the first year. (*See* **Neuroimaging**)

Synaptic Pruning

Following the overproduction of synapses, there is a period of synaptic loss. The elimination of synapses is a selective, and not a random, process. As with the timing of the bursts of synaptogenesis, the timing of the reduction in synaptic density is different in different cortical areas. For example, synaptic density in the visual cortex returns to adult levels between 2 and 4 years, while the same point is not reached until between 10 and 20 years of age for regions of the prefrontal cortex. One hypothesis is that the decrease in levels of brain metabolism to adult-like levels seen in the PET scan may reflect the decrease in synaptic contacts. In support of this idea, the peak of glucose uptake in the visual cortex of cats coincides with the peak in overproduction of synapses in this region. However, when the time course of synaptogenesis and glucose metabolism are compared in humans, it is apparent that the peak of glucose uptake occurs *after* the peak of synaptic density. Thus, the decrease in brain metabolism may reflect other factors (e.g. that less brain activity is required for certain skills once they are well learned).

EXPERIENCE-EXPECTANT AND EXPERIENCE-DEPENDENT GROWTH OF NEURAL CONNECTIONS

How is the brain able to achieve the very specific and adapted pattern of connections required for it to function normally? One possibility is that all of this information is prespecified in the genetic plan. However, given the sheer number of connections, many scientists have questioned whether the genome is even large enough to encode all of this information. Another possibility is that at least some aspects of this pattern are shaped by experience. In this view, the formation of the specific pattern of connections occurs by the overproduction and subsequent pruning of synapses, which allows information to be stored by selecting useful connections and eliminating surplus ones. This process is called 'experience-expectant' plasticity and is a mechanism that allows some aspects of experience to 'fine-tune' brain anatomy through selective synaptic loss. The synapses that are retained are those that respond to aspects of the environment that are common to all members of a species. For example, if the visual cortex is deprived of the input of patterned light due to a cataract, or clouding of the lens, connections in the visual cortex and visual function will not develop normally. This shows that the input of patterned light is necessary for normal structural and functional development of the visual cortex. Thus, aspects of brain organization can emerge commonly in most members of the species not because the organization is genetically encoded but because the brain is shaped by aspects of environment common to all species. (*See* **Innateness, Philosophical Issues about**; **Synaptic Plasticity, Mechanisms of**)

This time of plasticity does not last indefinitely but follows a developmental timetable. If certain synaptic connections are not laid down early in life, they are less likely to become established later in life. For example, in one study young and old owls were fitted with prisms that distorted their visual input. They were then tested to see whether they would be able to recalibrate the brain areas involved in relating visual and auditory information based on this altered visual input. Young owls were able to recalibrate in 3 weeks, but old owls never could. Moreover, if the prisms were replaced on the young owls once they were adults, they retained their ability to adjust to the altered experience. This study shows both that the effects of experience may be time-limited, and that the effects may persist into adulthood even if the experience itself was only temporary. Thus, once the number of neurons stabilizes at the mature level, changes in function may also stabilize or 'freeze' at a particular level.

While some experiences are typically common to all members of the species (e.g. hearing spoken language), other types of information may be useful only to individuals (e.g. whether that language is Spanish or English). A mechanism is also needed to store information that is unique to individuals. This second mechanism by which experience influences synapse development is called 'experience-dependent'. This refers to changes in the brain that can be specific to the individual and act to optimize adaptation to the particular characteristics of their environments. For example, rats show substantial changes in their brains if they are living in an 'enriched' social and physical environment compared to if they live in relative

isolation. These changes include production of new synapses and increases in complexity of the dendritic branching. Learning in humans may also involve creation of new synapses, although some scientists believe that in human adults the number of synapses remains stable and learning only modifies existing synapses by strengthening or remodeling them.

SUBCORTICAL VERSUS CORTICAL CONTROL OF BEHAVIOR

As the cortex develops and becomes increasingly mature, it can play an increasingly important role in regulating behavior. It allows more intentional, purposeful, behavior, in part by inhibiting the more automatic subcortical pathways. For example, newborns will automatically grasp an object placed in their hand, but this reflex disappears at 2 months of age when more controlled hand manipulations begin to emerge. The disappearance of this reflex and the emergence of the more purposeful hand movements are thought to be due to cortical development and the inhibition of brainstem neurons. The importance of the cortex for inhibiting this reflex can be demonstrated in adult humans or monkeys following lesions of particular regions of the frontal cortex: they show a reappearance of the newborn's palmar grasp reflex or 'forced grasping'. (*See* **Motor Development**)

The increasing influence of cortical control over behavior with development can also be observed in other domains, such as face recognition. Newborn infants will move their eyes further to follow a moving face than many other patterns. This preferential orienting response is thought to be primarily mediated by subcortical neural circuitry because of its developmental time course: it is present at birth and then declines sharply between 4 and 6 weeks after birth. The time course of disappearance of this response is similar to that of other newborn responses thought to be mediated by subcortical circuits, such as the grasp reflex described above. At the same time that this early tendency to visually follow faces decreases, a new response to faces emerges. At 6 to 8 weeks of age infants begin to show an increased visual attention to faces seen in the centre of vision. Both these changes are thought to be due to cortical development, which both inhibits subcortical reflexes and allows emergence of new functions. (*See* **Face Perception, Neural Basis of**; **Face Perception, Psychology of**)

The early subcortical reflexes of the infant may themselves play a role in cortical development. These reflexes may help ensure that infants receive certain inputs from the environment that are needed for the experience-expectant learning of the cortex. Thus, they serve to organize and stabilize information for cortical regions to incorporate.

EEG AND ERP TECHNIQUES

One very useful tool for studying cortical development in human infants is recording the electrical activity of the brain (a response caused by banks of neurons firing simultaneously). This activity can be recorded by means of electrodes that rest on the scalp surface. The electrical signals recorded at the surface are thought to reflect primarily cortical, rather than subcortical, activity. These recordings can be of either the spontaneous natural rhythms of the brain (electroencephalography – EEG), or the electrical activity induced by the presentation of a stimulus (event-related potential – ERP). EEG is ideally suited to study behaviors that occur over prolonged time periods (e.g. experience of emotion), while ERPs are more suitable for studying cognitive processing, because they retain precise information about the timing of cortical activation. (*See* **Event-related Potentials and Mental Chronometry**)

ERPs can be used to study several types of questions about cortical development: for example, ERPs can be used to investigate the timing of the emergence of cortical function. In studies of the brain activity related to making eye movements (saccades) in adults, a particular response called the spike potential occurs in the ERP just prior to the onset of the eye movement. This spike potential is not observed in young infants but is detectable by 12 months of age, suggesting that mature cortical control of saccadic eye movements does not emerge until at least that time. (*See* **Eye Movements**)

ERPs can also be used to investigate how experience shapes cortical development. For example, they have been used to study the influence of experience on the development of speech processing. Newborn infants discern differences between all the phonetic units used in the world's languages. This is an instance where infants' abilities are superior to those of adults: adults find it difficult to discriminate contrasts not found in their native language while infants do not (e.g. adult native speakers of Japanese find it difficult to discriminate the English /r/ and /l/, while both American and Japanese infants are quite able to make this discrimination). Infants' abilities change dramatically, however, between 6 and 12 months of age. This can be shown in the ERP in a deflection known as the mismatch negativity (MMN). The MMN reflects auditory processing by thalamocortical pathways. At 6 months

infants' MMN to contrasting phonemes reflects mainly the physical difference between the phonemes. However, by 12 months this is no longer the case: the MMN is enhanced if the contrast normally occurs in the native language but reduced if it does not. Thus, the experience of hearing the native language alters the cortical response to the auditory input. The functional significance of the MMN for later language processing is illustrated by the fact that children showing language-learning impairments show smaller amplitudes of MMN and show difficulty in discriminating speech contrasts. (*See* **Speech Perception, Development of**; **Auditory Event-related Potentials**)

IMMATURE BRAIN'S RESPONSE TO INJURY

The young brain's remarkable plasticity can also be observed in its response to injury. Damage to particular brain areas can have less serious consequences when the damage occurs during infancy than when it occurs during adulthood. One example is the effect on language of damage to the left hemisphere of the cerebral cortex. In the majority of normal adults, the left side of the brain plays a dominant role in language, and damage to the left perisylvian areas results in severe difficulties in speaking known as dysphasia or aphasia. In contrast, similar damage occurring in the first 5 years of life has much less devastating consequences. For example, in one study of 16 children who sustained similar injuries in the first 5 years of life, none had any detectable dysphasic symptoms. Thus it seems that before the age of five or six, brain areas other than those typically used are able to take on the functions of speech and language. However, by 6 years of age, as in adulthood, the left hemisphere of the normal child begins to dominate, so that left hemisphere injury that occurs at this age or later does produce a form of aphasia. The process of development can be thought of as an 'increasing restriction of fate'. What this means is that as the biological development of an individual proceeds, the range of options for further specification or specialization available to the organism decreases. The young organism shows plasticity in that there are still options available for alternative developmental pathways. As the brain develops and different regions become specialized for different functions, such as language, there are increasingly few or no options left for reorganization. (*See* **Aphasia**; **Brain Asymmetry**; **Language Development, Critical Periods in**; **Modularity in Neural Systems and Localization of Function**)

SUMMARY AND CONCLUSIONS

The normal development of the brain depends on the interaction of genetic inheritance and environmental experience. The genes provide the general structure of the nervous system, and the activity of this system as it interacts with the environment provides fine-tuning. Initially the young brain contains more components and connections than it will in adulthood, and the inputs it receives shape the elimination of this surplus. This provides a mechanism by which individuals can develop in similar ways even if the plan of development is not encoded specifically in the human genome. The brain continues to learn throughout the lifetime and to adapt to the unique characteristics of environments, but this process involves the addition to or strengthening of existing connections of the brain rather than their elimination.

Further Reading

Black JE (1998) How a child builds its brain: some lessons from animal studies of neural plasticity. *Preventative Medicine* **27**: 168–171.

Bourgeois J-P, Goldman-Rakic PS and Rakic P (2000) Formation, elimination and stabilization of synapses in the primate cerebral cortex. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 45–53. Cambridge, MA: MIT Press.

Johnson MH (1997) *Developmental Cognitive Neuroscience*. Oxford, UK: Blackwell.

Johnson MH (2001) Functional brain development in humans. *Nature Reviews Neuroscience* **2**: 475–483.

Johnson MH, Munakata Y and Gilmore R (2002) *Brain Development and Cognition: A Reader* 2nd edn. Oxford, UK: Blackwell.

Kalverboer AF and Gramsbergen A (2001) *Handbook of Brain and Behaviour in Human Development*. Amsterdam: Elsevier.

Nelson CA and Luciana M (2001) *Handbook of Developmental Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Rakic P (2000) Setting the stage for cognition: genesis of the primate cerebral cortex. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn, pp. 7–21. Cambridge, MA: MIT Press.

Shors TJ, Miesegaes G, Beylin A *et al*. (2001) Neurogenesis in the adult is involved in the formation of memory traces. *Science* **410**: 372–376.

Vargha-Khadem F and Mishkin M (1997) Speech and language outcome after hemispherectomy in childhood. In: Tuxhorn I, Holthausen H and Boenigk HE (eds) *Paediatric Epilepsy Syndromes and Their Surgical Treatment*, pp. 774–784. Montrouge, France: John Libbey & Company Ltd Medical Books.

Object Concept, Development of

Intermediate article

Renée Baillargeon, University of Illinois, Champaign-Urbana, Illinois, USA
Yuyan Luo, University of Illinois, Champaign-Urbana, Illinois, USA

CONTENTS

How do infants conceive of objects? In particular, do infants realize that objects continue to exist when hidden? Recent research indicates that although infants represent hidden objects from a very early age, their ability to reason about these objects improves dramatically during the first year of life.

PIAGET AND OBJECT PERMANENCE

As adults, we often observe events in which objects become hidden. For example, we might see a cup being pushed behind a teapot, an apple being lowered inside a bowl, or a fork being covered with a towel. Our representations of these events include both the objects that are directly visible – the teapot, the bowl, and the towel – and the objects that are not – the cup, the apple, and the fork. Piaget (1954) was the first researcher to ask whether infants, like adults, represent hidden objects. He concluded that it is not until infants are about 8 months of age that they realize that objects continue to exist when hidden. This conclusion was based primarily on analyses of infants' responses in manual search tasks. Piaget found that young infants typically do not search for objects they have observed being hidden: if a toy is covered with a cloth, for example, infants aged 5 to 7 months make no attempt to lift the cloth and grasp the toy, even though they are capable of performing these actions. Piaget speculated that for the young infant objects are not permanent entities that continue to exist when hidden; rather, they are transient entities that cease to exist when they cease to be visible.

PIAGET QUESTIONED

For several decades, researchers generally accepted Piaget's (1954) conclusion that young infants lack a notion of object permanence. This state of affairs began to change during the 1980s, however, when experiments conducted with novel, more sensitive methods yielded results that contradicted Piaget's conclusion. One such method was the violation-of-expectation method. In a typical experiment conducted with this method, infants see two test events, one consistent (expected event) and one inconsistent (unexpected event) with the expectation examined in the experiment. With appropriate controls, evidence that infants look reliably longer at the unexpected than at the expected event indicates that they (1) possess the expectation under examination; (2) detect the violation in the unexpected event; and (3) are interested or surprised by this violation.

The Rotating Screen Task

One often-cited example of a violation-of-expectation task used to examine whether young infants represent hidden objects is the rotating screen task (e.g. Baillargeon, 1987). In a series of experiments, 4.5- and 5.5-month-old infants were first habituated to a screen that rotated through a 180° arc in the manner of a drawbridge; the screen rotated continuously, first away from and then back toward the infants. Next, a box was placed behind the screen, and the infants saw an expected and an unexpected test event (see Figure 1). In both events, the screen began its rotation as before, progressively hiding the box. In the expected event, the screen rotated until it reached the hidden box, and then returned to its original position on the apparatus floor. In the unexpected event, the screen rotated through a full 180° arc as though the box was no longer present; the screen then returned to its original position, revealing the box standing intact behind it. The infants looked reliably longer at the unexpected than at the expected event, suggesting that they (1) believed that the box continued to exist after it disappeared from view; (2) expected the screen to stop against the hidden box; and (3) were surprised when this expectation was violated.

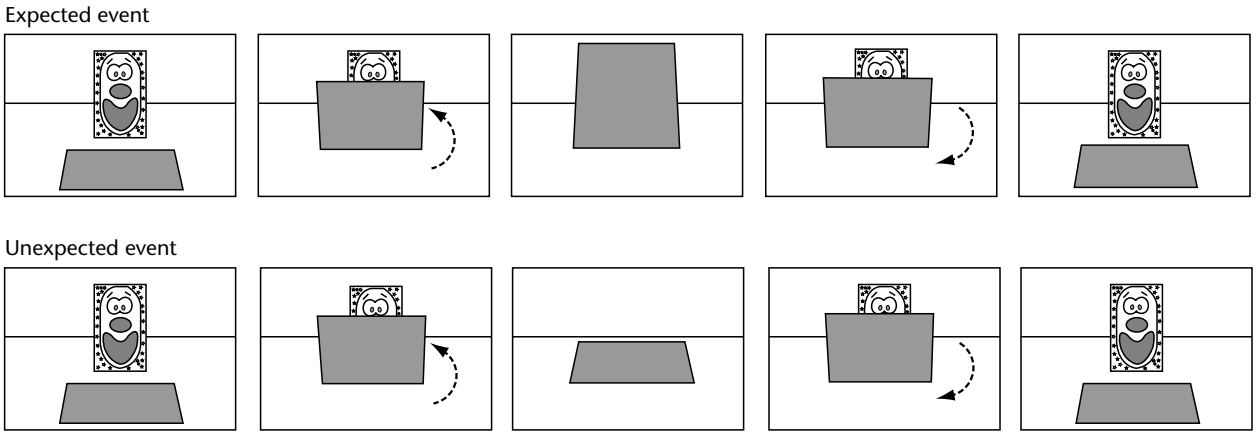


Figure 1. Schematic drawing of the test events used in Baillargeon (1987).

This interpretation was supported by control experiments in which the box was either absent or placed out of the screen's path; the infants in these experiments looked about equally at the two screen rotations.

These results were among the first to cast doubt on Piaget's (1954) conclusion that young infants lack a notion of object permanence. Since then, about three dozen reports have provided converging evidence that infants aged 2.5 to 7.5 months can represent hidden objects (for a partial listing, see Baillargeon, 2000). These reports were produced by different laboratories, using a wide variety of violation-of-expectation and other tasks. This substantial body of evidence has led many researchers to conclude that young infants, like adults, appreciate that objects continue to exist when hidden.

Perceptual Biases

A few researchers have recently questioned some of the initial results taken to reveal a notion of object permanence in young infants (e.g. Bogartz *et al.*, 2000). The concern is that these results could be explained by low-level perceptual biases in infants' processing of events. For example, it has been suggested that the infants in the rotating screen task could have looked reliably longer at the unexpected than at the expected event simply because they were habituated to the screen rotating through a 180° arc and preferred this familiar rotation in the test events. This and other similar alternative interpretations have been useful in reminding researchers to examine closely their data for possible confounds. Nevertheless, these interpretations have themselves become the subject of criticism, on several grounds (e.g. Aslin, 2000; Baillargeon, 2000). First, control experiments are generally ignored. For example, the alternative explanation just presented does not address the finding that control infants tested without a box behind the screen did not show a preference for the familiar 180° rotation. Second, a different perceptual bias is typically invoked for each separate result, giving rise to concerns about theoretical coherence and parsimony. Finally, the alternative interpretations proposed to date do not consider the many reports that have provided further evidence of object permanence in young infants.

DEVELOPMENTS IN INFANTS' RESPONSES TO HIDDEN OBJECTS

The evidence that infants aged 2.5 months and older are able to represent and to reason about hidden objects does not mean that these abilities undergo little or no development. Research over the past 10 years has uncovered many separate developments during the first year in infants' ability to reason about hidden objects. Four such developments are discussed below.

Predicting When Objects Should be Occluded

Recent evidence suggests that, although young infants recognize that an object continues to exist *after* it becomes hidden by a nearer object or occluder, they are rather poor at predicting *when* it should be occluded. In a series of experiments (e.g. Aguiar and Baillargeon, 1999, in press), infants aged 2.5 to 3.5 months saw an object move back and forth behind a screen; a portion of the screen was removed and the infants judged whether the object should remain hidden or become (at least partly)

visible when passing behind the screen. The results indicated that, by 2.5 months of age, infants have formed an initial concept of occlusion centered on a simple *behind/not behind* distinction. When the entire midsection of the screen is removed to form two separate screens, infants expect the object to become visible in the gap between them. However, if the screens remain connected either at the top or at the bottom by a short strip, infants no longer expect the object to become visible: they view the connected screens as a single occluder and expect the object to be hidden when behind it. Over the course of the next month, infants rapidly progress beyond their initial concept. At about 3 months of age, infants begin to consider the presence of a *discontinuity in the lower edge* of the screen. Although infants still expect the object to remain hidden when passing behind two screens that are connected at the bottom by a short strip, they now expect the object to become visible when passing behind two screens that are connected at the top by a short strip. Finally, at about 3.5 months of age, infants begin to consider the relative *heights* of the object and screen. When the object passes behind two screens that are connected at the bottom by a strip, infants expect the object to become partly visible if it is taller but not shorter than the strip.

Inferring the Presence of Occluded Objects

A number of experiments have examined young infants' ability to posit – rather than to merely represent – the presence of objects behind occluders. For example, Aguiar and Baillargeon (in press) showed 3- and 3.5-month-old infants a toy mouse that moved back and forth behind two screens connected at the top by a strip; the mouse did not appear in the gap between the screens. The 3-month-olds showed surprise at the event, but the 3.5-month-olds did not: these older infants apparently guessed that *two* identical mice were involved in the event, one traveling to the left and one to the right of the screens. Additional results supported this interpretation: for example, 3.5-month-olds again failed to show surprise at the event (1) when the screens were briefly lowered prior to the event to reveal a mouse and a small screen that was sufficiently large to hide a second mouse, but *not* (2) when the screens were lowered to reveal only one mouse. The infants thus produced a two-mouse explanation unless given information directly contradicting such an explanation. Together, these results indicate that, by 3.5 months of age, infants can posit hidden objects to make sense of (at least some) events that would otherwise violate their occlusion knowledge.

Reasoning About Occlusion and Other Events

New findings indicate that infants form distinct event categories and learn separately how each category operates (for a review, see Baillargeon and Wang, 2002). Because infants learn at different rates about events involving occluders, containers, and covers, striking décalages can be observed in their reasoning about these different event categories. For example, recent experiments compared infants' ability to reason about height information in occlusion and containment events (Hespos and Baillargeon, 2001). Infants aged 4.5 to 7.5 months saw a tall object being lowered either behind an occluder or inside a container; the height of the occluder or container was varied, and the infants judged whether the object could be fully or only partly hidden. The occlusion and containment events were perceptually very similar (e.g. in some experiments, the occluders were identical to the containers with their backs and bottoms removed). The results indicated that, at 4.5 months of age, infants are surprised to see a tall object become fully hidden behind a short occluder; however, it is not until infants are about 7.5 months of age that they are surprised to see a tall object become fully hidden inside a short container. Infants thus learn to consider the relative heights of objects and occluders several months before they do the same for objects and containers.

Mapping Objects From Occlusion to Other Events

Xu and Carey (1996) examined 10- and 12-month-old infants' ability to individuate the objects in an occlusion event – that is, to determine how many distinct objects were involved in the event. The infants first watched the following event sequence: an object (e.g. a ball) emerged from behind one edge of a screen and then returned behind it; next, a different object (e.g. a bottle) emerged from behind the other edge of the screen and again returned behind it. This sequence was repeated several times, and then the screen was lowered to reveal one or both of the objects. Only the 12-month-olds showed surprise at the one-object outcome. Xu and Carey concluded that the younger infants did not realize that two distinct objects were

involved in the occlusion event. They speculated that 10-month-old infants still lack specific object concepts such as ball and bottle, and that these concepts do not become available until the end of the first year, when word learning begins.

The finding that 10-month-old infants fail at the individuation task devised by Xu and Carey (1996) has been confirmed by many researchers. However, the interpretation of this finding has been questioned. First, rhesus macaques succeed at similar tasks, despite their lack of language (e.g. Uller *et al*., 1997). Second, prelinguistic infants also succeed at similar tasks when processing demands are reduced (e.g. Wilcox and Schweinle, 2002). For example, 5.5-month-old infants show surprise at a one-object outcome if the initial event sequence is considerably shortened (e.g. the first object disappears behind one edge of the screen, the second object emerges from behind the other edge, and the screen is immediately lowered).

Wilcox and Baillargeon (1998) proposed an alternative interpretation of the negative results obtained by Xu and Carey (1996). This interpretation rests on the distinction between event-mapping and event-monitoring tasks. In an event-mapping task, infants see events from *two* different event categories and judge whether the two events are consistent. In an event-monitoring task, infants see an event from *one* event category and judge whether successive portions of the event are consistent. Tasks such as that of Xu and Carey are event-mapping tasks: each test trial involves first an *occlusion* event (when the objects emerge successively from behind the screen) and then a *display* event (when one or both objects rest on the apparatus floor). To succeed, infants must (1) retrieve their representation of the occlusion event and (2) map the objects in this representation onto those in the display event. According to Wilcox and Baillargeon, young infants have difficulty completing this retrieval and mapping process.

Wilcox and her colleagues have tested several predictions derived from the preceding analysis (e.g. Wilcox and Chapa, in press; Wilcox *et al*., in press). One finding was particularly striking: 9.5-month-olds showed surprise at a one-object outcome when they viewed it through (1) a thin frame filled with a clear plastic, but not (2) an empty frame (the filled or empty frame was revealed when the screen was lowered). According to Wilcox, the infants tested with the empty frame faced an event-mapping task: they saw first an occlusion and then a display event, and they experienced the usual difficulty in mapping one event onto the other. In contrast, the infants tested with the filled frame faced an event-monitoring task: they saw a single, ongoing occlusion event involving first an opaque and then a clear occluder, and they easily kept track of the objects as the event unfolded. These results support the notion that event mapping poses special difficulties for young infants.

SUMMARY

Piaget (1954) was the first researcher to examine whether infants, like adults, represent hidden objects. Data from manual search tasks led him to conclude that infants younger than 8 months do not realize that objects are permanent entities that continue to exist when hidden. However, subsequent data obtained with violation-of-expectation and other tasks showed that even very young infants represent the continued existence of hidden objects.

In recent years, researchers have explored infants' ability to reason about hidden objects in various events. These experiments have brought to light significant developments in infants' ability (1) to posit the presence of occluded objects; (2) to predict when objects should be hidden in occlusion and other events; and (3) to map or keep track of objects across occlusion and other events.

References

Aguiar A and Baillargeon R (1999) 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology* **39**: 116–157.

Aguiar A and Baillargeon R (in press) Developments in young infants' reasoning about occluded objects. *Cognitive Psychology*.

Aslin RN (2000) Why take the cog out of infant cognition? *Infancy* **1**: 463–470.

Baillargeon R (1987) Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology* **23**: 655–664.

Baillargeon R (2000) Reply to Bogartz, Shinskey, and Schilling; Schilling; and Cashon and Cohen. *Infancy* **1**: 447–462.

Baillargeon R and Wang S (2002) Event categorization in infancy. *Trends in Cognitive Science* **6**: 85–93.

Bogartz RS, Shinskey JL and Schilling TH (2000) Object permanence in five-and-a-half-month-old infants? *Infancy* **1**: 403–428.

Hespos SJ and Baillargeon R (2001) Infants' knowledge about occlusion and containment events: a surprising discrepancy. *Psychological Science* **12**: 140–147.

Piaget J (1954) *The Construction of Reality in the Child*, translated by Cook M. New York: Basic Books. [Original work published in 1937.]

Uller C, Xu F, Carey S and Hauser MD (1997) Is language needed for constructing sortal concepts? A study with

nonhuman primates. In: Hughes E (ed.) *Proceedings of the 21st Annual Boston University Conference on Language Development*, pp. 665–677. New York: Oxford University Press.

Wilcox T and Baillargeon R (1998) Object individuation in infancy: the use of featural information in reasoning about occlusion events. *Cognitive Psychology* **37**: 97–155.

Wilcox T and Chapa C (in press) Infants' reasoning about opaque and transparent occluders in an individuation task. *Cognition*.

Wilcox T and Schweinle A (2002) Object individuation and event mapping: developmental changes in infants' use of featural information. *Developmental Science* **5**: 132–150.

Wilcox T, Schweinle A and Chapa C (in press) Object individuation in infancy. In: Fagan F and Hayne H (eds) *Progress in Infancy Research*, vol 3. Mahwah, NJ: Lawrence Erlbaum.

Xu F and Carey S (1996) Infants' metaphysics: the case of numerical identity. *Cognitive Psychology* **30**: 111–153.

Further Reading

Baillargeon R (1993) The object concept revisited: new directions in the investigation of infants' physical knowledge. In: Granrud CE (ed.) *Visual Perception and Cognition in Infancy*, pp. 265–315. Hillsdale, NJ: Erlbaum.

Baillargeon R (2002) The acquisition of physical knowledge in infancy: a summary in eight lessons. In: Goswami U (ed.) *Handbook of Childhood Cognitive Development*, pp. 47–83. Oxford: Blackwell.

Baillargeon R, Spelke ES and Wasserman S (1985) Object permanence in 5-month-old infants. *Cognition* **20**: 191–208.

Haith MM and Benson JB (1997) Infant cognition. In: Damon W (series ed.), Kuhn D and Siegler R (eds) *Handbook of Child Psychology*, vol. 2, pp. 199–254. New York: Wiley.

Meltzoff AN and Moore MK (1998) Object representation, identity, and the paradox of early permanence: steps toward a new framework. *Infant Behavior and Development* **21**: 201–235.

Munakata Y, McClelland JL, Johnson MH and Siegler R (1997) Rethinking infant knowledge: toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review* **104**: 686–713.

Needham A and Baillargeon R (2000) Infants' use of featural and experiential information in segregating and individuating objects: a reply to Xu, Carey and Welch. *Cognition* **74**: 255–284.

Santos LR, Sulkowski GM, Spaepen GM and Hauser MD (in press) Object individuation using property/kind information in rhesus macaques. *Cognition*.

Spelke ES, Breinlinger K, Macomber J and Jacobson K (1992) Origins of knowledge. *Psychological Review* **99**: 605–632.

Spelke ES and Hespos SJ (in press) Conceptual development in infancy: the case of containment. In: Stein N, Bauer P and Rabinowitch M (eds) *A Festschrift for Jean Mandler*. Hillsdale, NJ: Erlbaum.

Spelke ES, Kestenbaum R, Simons DJ and Wein D (1995) Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology* **13**: 1–30.

Xu F, Carey S and Welch J (1999) Infants' ability to use object kind information for object individuation. *Cognition* **70**: 137–166.

Object Perception, Development of

Intermediate article

Scott P Johnson, Cornell University, Ithaca, New York, USA

CONTENTS

Infants are born with the ability to detect visible surfaces as separate from one another, but not to perceive them as linked behind an occluder. This latter ability develops rapidly across the first several months after birth.

INTRODUCTION

When we look around us, we see a world that is occupied by objects at various distances, an experience that is made possible by light that is reflected from the object surfaces to the eye. What we *see directly* and what we *know* about objects, however, are quite different. The pattern of light reaching the eye from object surfaces is actually *fragmented* across space and time. That is, few objects are visible in their entirety from a single vantage point, because objects typically occlude one another, and any momentary view of the visual array is apt to change when either the observer or objects begin to move. Despite these potential challenges to the visual system, we do not experience a mercurial patchwork of disconnected image fragments, but instead perceive a stable layout of coherent objects that maintain their boundaries over time.

Objects are seen as separate from one another by virtue of the fact that reflectance characteristics of surfaces vary as a function of their composition, spatial arrangement, and movement. To perceive objects, we must attend to and utilize the visual information that specifies *surface segregation*: color, luminance, texture, orientation, contour, motion, and depth. But there is more to the problem of object perception. An object may be partially occluded, such that one part of its surface is visible in one part of the optic array, and another part visible somewhere else. To perceive this object accurately, the observer must unify these visible surfaces across the spatial gap, a process known as *unit formation*. The visual information that supports unit formation includes the common motion and similarity of disparate surfaces, as well as the alignment of their edges as they intersect the occluder and the shape formed by the joined regions. Surface segregation and unit formation, therefore, rely on the observer's ability to detect and accurately use numerous visual cues available in the optic environment.

Now consider the state of affairs for a neonate (a newborn infant). She has spent her entire life to that point in darkness (in the womb), and is suddenly confronted with the visual array, a kaleidoscope of colors, shapes, and motion. To perceive objects accurately, she must solve the problems outlined previously (detection and utilization of pertinent visual information, leading to surface segregation and unit formation), but she has had little or no exposure to patterned light until now. How do these problems come to be solved, with the onset of visual experience?

The question of the origins of object perception is an ideal candidate for framing in the terms of the classic nature–nurture debate: do infants perceive objects accurately from the start of postnatal life, suggesting that object perception is an innate capacity? Or is there some developmental timetable for this capacity, involving, for example, neural maturation or a period of experience watching and acting on objects? As we will see, this question has guided much of the research in the area, and still informs the debate. It is becoming increasingly clear, nevertheless, that a nativist, or nature-oriented, point of view cannot explain the bulk of the evidence that neonates do not perceive objects accurately. Instead, it appears far more likely that accurate object perception is a set of skills that emerges over the first several months after birth. These empirical facts, discussed in more detail below, shift the question to *mechanisms of development*: how, exactly, the infant's visual system changes to support accurate perception of objects. (*See* **Object Concept, Development of**)

PERCEPTION OF OBJECT UNITY

Young infants' perception of partly occluded objects has been investigated with displays that depict two rod parts, moving back and forth above and below an occluding box, against a textured or patterned background (Figure 1(a)), using a method developed by Kellman and Spelke (1983). Infants watch this display, and their looking towards the display is recorded. When looking times decline according to a preset criterion (e.g. half their original level), the infants are then shown two new test displays, each of which matches the first display in different ways. The display depicted in Figure 1(b), for example, matches a percept of the rod parts' unity (i.e. a 'complete' rod), and the display depicted in Figure 1(c) matches a percept of disjoint surfaces (i.e. a 'broken' rod). (Note that either 'interpretation' of the rod-and-box display is plausible, although adults are likely to perceive object unity in this display.) After infants repeatedly view any single stimulus until looking times wane, a process known as *habituation*, they typically will show a preference for a novel stimulus, relative to a familiar stimulus. Researchers have capitalized on this tendency towards posthabituation novelty preferences to probe the conditions under which infants will perceive object unity, in two kinds of investigation: the visual cues used by infants to perceive partly occluded objects, and the developmental origins of object perception. (*See* **Vision: Occlusion, Illusory Contours and 'Filling-in'**)

Exploring Infants' Use of Visual Information: Gestalt Principles?

Kellman and Spelke (1983) found that four-month-old infants looked longer at a broken rod, relative to a complete rod, after habituation to a three-dimensional rod-and-box display in which the rod parts were aligned across the occluder and underwent common motion. Results from a control condition revealed that there was no inherent preference for either test display, suggesting that longer looking at the broken rod was indeed a novelty preference, arising from the infants' habituation experience (i.e. watching the rod-and-box display and perceiving object unity). This effect generalized to a display in which a rod part above the box was paired with an irregular polygon shape below the box. The rod part and polygon underwent common motion but were mismatched in shape and surface texture, and the infants appeared to perceive them as unified. In contrast,

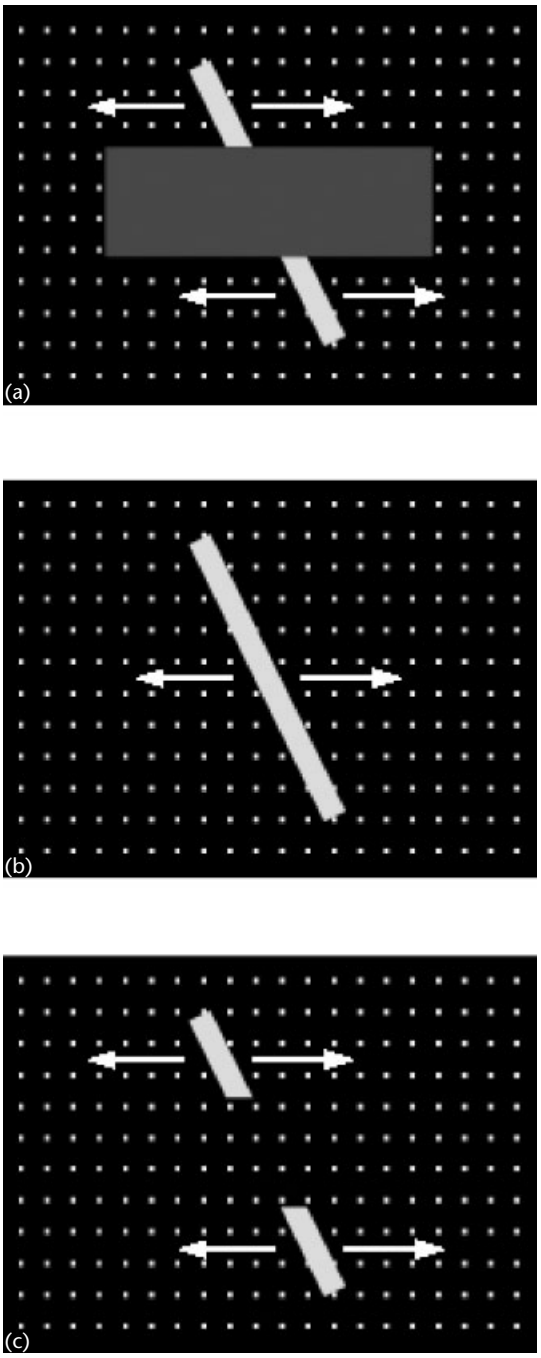


Figure 1. Displays used in investigations of infants' perception of object unity: (a) rod-and-box display, in which two rod parts are aligned and undergo common motion above and below an occluder; (b) complete rod test display; (c) broken rod test display. After habituation (repeated exposure) to (a), infants will typically look longer at (c) if they perceived the unity of the rod parts, and will look longer at (b) if they perceived the rod parts to be disjoint objects. Adapted from Johnson and Aslin (1996).

infants provided no evidence of unit formation in any stationary displays: no posthabituation preference was found under these conditions. Adults, in contrast, reported perception of object unity readily in these stimuli, presumably on the basis of 'static' information: the alignment of the rod parts, the fact that the conjoined rod segments comprise a simple, regular form, and so on.

These results raise intriguing questions about the nature of infants' perceptual development. It seems that some kinds of visual cue for occlusion are informative (e.g. motion) whereas others are not (e.g. alignment). For adults, the case is quite different: we exploit a range of cues in object perception tasks, including static information. This latter observation led the Gestalt psychologists during the last century (e.g. Koffka, 1935) to propose that the visual system is predisposed to organize the optic array into the simplest possible configuration. In general, this tendency accords with perception of objects as simple, regular forms with smooth contours rather than the fragmentary images that reach the eye. This idea was expressed more formally in terms of Gestalt 'principles', those visual cues that aided perceptual organization of our complex visual environment. These principles included *common fate* (or common motion), *good continuation* (or alignment), *good form*, *symmetry*, and *simplicity*, all of which are available in the rod-and-box display depicted in Figure 1(a), and thus may guide adults' percepts of unity. Because the visual system is inherently predisposed towards this kind of organization, according to the Gestalt view, it follows that infants should experience objects in like manner to adults, but we have seen that this is not the case: only a subset of the visual cues seem effective for infants' unit formation. (*See* **Perception, Gestalt Principles of**)

Edge-sensitive versus Edge-insensitive Processes

Kellman (1996) has suggested that young infants are 'edge insensitive', meaning that they fail to use alignment cues in unit formation tasks. Instead, infants younger than about six months are presumed to rely exclusively on motion information. After six months, an 'edge-sensitive' process is hypothesized to emerge, which takes advantage of static information, in addition to motion. This two-process theory is consistent with the findings presented thus far, in that four-month-olds perceive unity when two surfaces move together, even if dissimilar in shape and texture. When stationary, however, surfaces are not perceived as unified, even with other Gestalt information.

Recently, evidence has emerged for a kind of edge sensitivity in young infants that was not tested directly by Kellman and Spelke (1983). Johnson and Aslin (1996) presented four-month-olds with computer-generated (two-dimensional) rod-and-box displays in which the rod parts above and below the occluder were either misaligned (Figure 2(a)) or nonaligned (Figure 2(b)), and underwent common motion, until habituation of looking occurred. Following habituation, the infants viewed broken and complete rod test displays (these displays matched the configuration of the misaligned and nonaligned rod parts, not the aligned rod parts seen in Figure 1(a)). According to the Kellman (1996) two-process account of unit formation, four-month-olds are edge-insensitive, and would be predicted, therefore, to perceive the rod surfaces in both these displays as unified, and subsequently should look longer at the broken rod test displays relative to the complete rod. A different pattern of results was obtained, however. Infants who were habituated to the misaligned rod parts (Figure 2(a)) exhibited no posthabituation preference, and the infants who saw the nonaligned rod parts looked longer at the *complete* rod, the opposite result than the prediction. These results appear to reflect edge sensitivity in these infants: when the edges were misaligned, unity was indeterminate, and when edges were not aligned, the rod parts were perceived as disjoint surfaces.

In another experiment, Johnson and Aslin (1996) asked whether the presence of a textured background is necessary for young infants' unit formation, perhaps as a depth cue (i.e. the covering and uncovering of background texture provides information for the closer distance to the observer of the moving surface). Again, this manipulation should have no effect on infants' perception of object unity, under the two-process theory, but infants tested in a textureless rod-and-box condition (Figure 2(c)) showed no posthabituation preference. As in the case of the misaligned rod display, this outcome is most likely a result of an indeterminate percept: neither unified nor disjoint surfaces.

The Johnson and Aslin (1996) findings, then, begin to provide evidence for edge sensitivity in four-month-olds, and other cues are implicated in the process as well. Young infants attend to edge alignment in perception of object unity, just as they do common motion. What is the role of texture? Depth information is impoverished in these

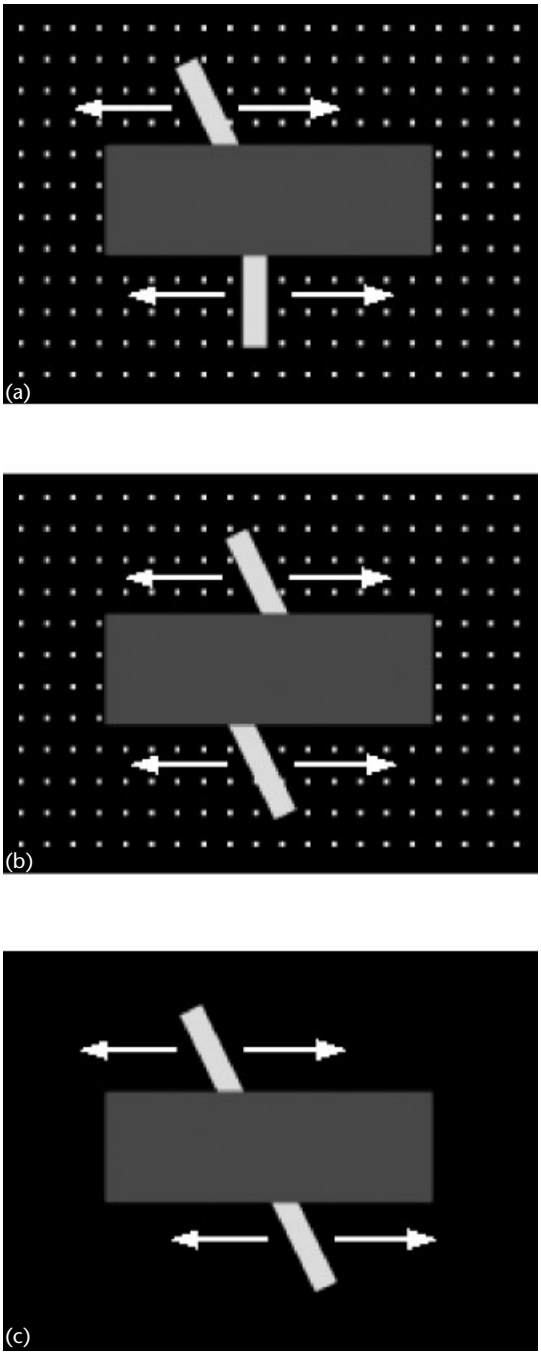


Figure 2. Displays used in investigations of the visual information used by four-month-old infants to perceive object unity: (a) misaligned rod display; (b) nonaligned rod display; (c) no texture display. In none of these displays did the infants appear to perceive object unity, suggesting that the common motion available in the displays was insufficient to specify connectedness. In addition to motion, infants also rely on edge alignment and depth cues to perceive unity. Adapted from Johnson and Aslin (1996).

two-dimensional displays, and we speculated that without the additional information for surface segregation that is provided in three-dimensional stimuli (i.e. placement of individual surfaces into their constituent depth planes relative to the observer), unit formation by the immature visual system may be precluded. The role of texture, therefore, is to provide an additional depth cue. Interestingly, this additional depth information is not needed by adults to perceive unity in the textureless display; apparently those cues that remain are sufficient to specify unity.

The Threshold Model

We have seen that the two-process theory of unit formation, while accurate in many respects, falls short of accounting for the range of evidence described this far concerning four-month-olds' perception of object unity. A *threshold model* may provide a preferable account of these data (Johnson, 1997). The threshold model posits that surface segregation and unit formation build on several subprocesses, and if any of these subprocesses are disrupted, accurate object perception can be precluded. Nakayama and Shimojo (1990) noted that in order to perceive unity in an occlusion display, the observer must determine in which depth plane each surface resides (*depth placement*), and determine which contours in the scene belong with which objects (*contour ownership*). Depth placement relies on depth cues, of course, and texture, for example, aids perceptual segregation of the rod and box surfaces into their constituent depth planes. Contour ownership may rely on edge alignment, and when rod edges are nonaligned, infants may perceive them as belonging to separate objects, as if the contours of the rod ended at the box. These results suggest that unit formation and surface segregation are multiply determined by independent sources of information, and, ultimately, object perception depends upon both the *sufficiency* of visual information and the *efficiency* of the observer's perceptual and/or cognitive skills. Unit formation and surface segregation, on this account, proceeds from an initial analysis of individual feature elements: edge orientations, surface intersections (at points of occlusion), and surface motions (Marr, 1982). From here, a description of the relative distances of surfaces is constructed, incorporating depth information from three-dimensional layout and other cues such as motion and texture. Young infants' skills at these subprocesses are somewhat compromised relative to adults, which explains improvements in performance with development.

Exploring Development of Perception of Object Unity: Core Principles?

Kellman and Spelke (1983) proposed that the roots of unit formation lie in an unlearned conception of what objects are like: 'Humans may begin life with the notion that the environment is composed of things that are coherent, that move as units independently of each other, and that tend to persist, maintaining their coherence and boundaries as they move' (p. 521). Infants' object perception has been posited to embrace certain 'core principles' that guide reasoning about objects from the start of postnatal life (Spelke and Van de Walle, 1993). These principles include cohesion (objects cannot move through the space occupied by another object) and contact (two surfaces that undergo a common, rigid motion tend to be connected). The contact principle is similar to the edge-insensitive process suggested by Kellman (1996), and as we have seen, does not provide an adequate account of recent evidence concerning four-month-olds' responses to object unity. Nevertheless, it might be that younger infants' object perception is guided by core principles.

Recall that the infants observed by Kellman and Spelke were four months of age, and that there was no direct evidence concerning origins of object perception. When neonates were tested using similar rod-and-box displays, the infants responded with the *opposite* looking-time pattern during test: a preference for the complete rod (Slater *et al*., 1990). Neonates, therefore, appeared to perceive disjoint rod surfaces in the rod-and-box display, disaffirming the nativist perspective suggested by Kellman and Spelke (1983) and Spelke and Van de Walle (1993). Core principles, therefore, do not seem to offer a plausible account of early object perception. On the contrary, neonates' responses to object occlusion seem to be quite inaccurate.

Exploring Development of Perception of Object Unity: Timing

The time between birth and four months, then, seems to be the period during which accurate responses to occlusion emerge. At what age can infants first perceive object unity? In the first study to explore this question, Johnson and Náñez (1995) reported that two-month-olds exhibited no preference for either a broken or a complete rod test display after habituation to a rod-and-box display. This suggests that two months of age represents a time of transition from perception of disjoint objects in the display (the neonates' response) to unit formation (the four-month-olds' response). Recall, however, the stipulations of the threshold model: it may be that we supplied insufficient visual information to support unit formation in a very young population that may have relatively ineffectual perceptual skills. This hypothesis was tested by presenting two-month-old infants with rod-and-box displays in which more of the rod was visible as it moved back and forth, either by reducing box height, or by incorporating gaps in the box (Johnson and Aslin, 1995). In each condition, the infants showed a consistent posthabituation preference for the broken rod relative to the complete rod, implying perception of object unity during habituation. Perception of object unity, therefore, may be a skill that is fragile in its earliest form, but nevertheless is available to even very young infants if given adequate perceptual support (see also Kawataba *et al*., 1999).

Our finding of two-month-olds' perception of object unity raises a vital question: will neonates perceive object unity as well, if given additional perceptual support? This possibility was investigated by Slater *et al*. (1996), who presented neonates with 'full-cue' three-dimensional rod-and-box displays that were rich in visual information: reduced occluder height, increased depth separation between rod, box, and background, and background texture (for additional depth information). Even with these added cues, however, the neonates provided no evidence of unit formation: they showed a clear and consistent posthabituation preference for the complete rod, relative to the broken rod.

Consider the implications raised by these experiments with neonates. No evidence emerged for perception of object unity. This result cannot be due to a general inability to distinguish between the surfaces in the display because of poor acuity, for example. Relative to adults, neonates' visual function is compromised in terms of acuity, contrast and color sensitivity, and so on. Even at birth, however, infants move their eyes volitionally, to (presumably) desired targets, most likely for closer inspection, and reflexively, in response to motion in the optic array. Neonates also exhibit visual preferences when differing stimuli are paired, looking longer, for example, at stripes versus a homogeneous gray, at curved versus rectilinear contours, at moving versus static stimuli, and others (Slater, 1995). Infants are born, therefore, with rudimentary visual skills, and do not scan the visual environment randomly. Instead, scanning patterns are structured from the start of postnatal life.

Neonates do not achieve unit formation, but decisive evidence was obtained for the other of our

object perception subskills outlined previously: surface segregation. Note that the neonates in the Slater *et al*. (1990, 1996) experiments looked longer at the complete rod during test. This implies perception of disjoint objects during habituation, when viewing the rod-and-box display. Neonates, therefore, are capable of perceiving the rod surfaces as separate from the occluder and background; in other words, they accomplished figure–ground segregation. If they had been unable to distinguish the rod surfaces from the occluder, say, there would have been no posthabituation preference, because both the broken and the complete rods would be equally novel. Responding to the partly occluded rod as consisting of disjoint objects suggests that the neonates perceived its boundaries to end at the point of intersection with the occluder.

MECHANISMS OF DEVELOPMENT

The research described in the previous sections outlines a timetable for the development of perception of object unity. At birth, infants respond to a partly occluded object as if it were composed of separate surfaces, and by two months, infants will perceive the unity of these surfaces under limited circumstances. By four months, the range of circumstances under which unity is perceived expands considerably, provided there is sufficient visual information, and there are improvements after this time as well (e.g. unit formation from static information). A bit of reflection reveals that this is one of the most profound changes that will occur in an individual's experience, which makes it one of the most intriguing questions for researchers in perceptual development. At birth, infants apparently perceive the world as a mosaic of disconnected shapes that must change continuously with every surface motion and with every movement of the self (including eye movements). There is a rapid turnabout of this state of affairs, such that within a few months the world must be experienced in a radically different way: a stable layout composed of coherent, bounded entities.

How does this expeditious and radical change occur? At present no single account encompasses the entire range of developmental evidence, but the threshold model holds promise in identifying important theoretical links that might help explain the emergence of unit formation. Recall that the model is based on the conjecture that improvements in information-processing skills underlie development of the ability to bind features into coherent surface and object percepts. Independent evidence is beginning to emerge that is consistent with this postulate, from observations of infants' eye movements, connectionist modeling, and neurophysiological development.

Eye Movements

A central tenet of the threshold model is the suggestion that with increased proficiency at information pickup, infants are more liable to detect and utilize information as appropriate in object perception tasks. Recording of eye movements can serve as an important tool to investigate this suggestion. Johnson and Johnson (2001) recorded scanning patterns in infants between 2 and 3.5 months as they viewed partly occluded rod displays, with a corneal-reflection eye tracker which provides extremely accurate data concerning the patterning and timing of eye movements. We predicted that older infants would scan more often in the rod's vicinity, scan more to both visible rod parts, and scan less in uninformative regions of the stimulus. Older infants produced a higher proportion of fixations per second than did younger infants, and scanned more extensively across the display, whereas younger infants scanned less often in the vicinity of the bottom rod part. Younger infants' fixations in the bottom region of the display were more frequent, however, when provided with longer stimulus presentations. We did not obtain direct evidence concerning perception of object unity in this study, but these results reveal important advances in scanning efficiency in the age range of interest to our question of the emergence of unit formation. (*See* **Visual Attention**)

Connectionist Modeling

Connectionist models are computer programs designed to respond to input stimuli and produce an output that reflects pattern recognition or a prediction about what might come next in a sequence. Connectionist models of developmental processes can provide important indicators concerning mechanisms of change, because the starting conditions and environmental context in which development occurs can be manipulated precisely in ways that are impossible with living systems. (*See* **Cognitive Development, Computational Models of**)

Mareschal and Johnson (2002) programmed connectionist models of the development of perception of object unity. These models were built with standard architectures and learning procedures (i.e. input, hidden, and output layers, and a back-propagation algorithm) and provided with

sensitivity to information that influences infants' perception of object unity (object orientation, motion, and background texture) and a transient memory (to retain stimulus information for brief durations). They were then trained with input representing partly occluded rod displays in which the rod moved back and forth behind an occluder, and emerged from either side, so that the object was both fully and partly visible during each translation. After varying amounts of training, the models were tested for perception of object unity with events in which the rod parts did not emerge from behind the occluder. The models perceived unity reliably in most of these test events. Learning efficiency was strongly dependent on the training environment: which cues were made available, and training duration. Surface binding, then, arose from an initial perceptual sensitivity combined with transient memory and experience in viewing objects that became occluded and again fully visible.

Developmental Neurophysiology

One way in which the visual system may bind perceptual features into coherent objects is with synchronized firing patterns across collections of neurons (Singer and Gray, 1995). If the neonatal cortex is incapable of achieving such coherent, synchronous activity (because, for example, of a general excess of 'noisy', chaotic firing), this might restrict the extent to which disparate surface fragments in the optic array could be bound into unified objects. Recent support for this possibility comes from evidence that synchronized neural activity shows marked improvements between six and eight months (Csibra *et al*., 2000). A second possible limitation may be deficiencies in long-range neural connections within and between areas of the immature visual system (Burkhalter *et al*., 1993), which, again, might impede the linking of spatially separate locations. There is evidence that these processes extend well beyond infancy, and into childhood (e.g. Kovács, 2000). (*See* **Binding Problem**; **Object Perception, Neural Basis of**; **Gamma Oscillations in Humans**)

CONCLUSION

The development of object perception has been investigated by assessing the extent to which young infants achieve perceptual completion in partly occluded object displays. These experiments lead to two conclusions. First, neonates are capable of figure–ground segregation, but do not perceive the unity of a center-occluded object; the ability to perceive object unity emerges over the first several postnatal months. Second, by four months, infants rely on a range of Gestalt visual information in perceiving unity, including common motion, alignment, and good form. This developmental pattern is hypothesized to rely on the increasing ability to detect and utilize appropriate visual information in support of the binding of features into surfaces and objects. Evidence from habituation experiments, changes in infant attention, computational modeling, and developmental neurophysiology is all consistent with this view. Specifically, the increasing ability of infants to perceive the world accurately appears to be rooted in a foundation of rudimentary visual skills that are present at birth, followed by a combination of visual experience and neural maturation.

References

Burkhalter A, Bernardo KL and Charles V (1993) Development of local circuits in human visual cortex. *Journal of Neuroscience* **13**: 1916–1931.

Csibra G, Davis G, Spratling MW and Johnson MH (2000) Gamma oscillations and object processing in the infant brain. *Science* **290**: 1582–1585.

Johnson SP (1997) Young infants' perception of object unity: implications for development of attentional and cognitive skills. *Current Directions in Psychological Science* **6**: 5–11.

Johnson SP and Aslin RN (1995) Perception of object unity in 2-month-old infants. *Developmental Psychology* **31**: 739–745.

Johnson SP and Aslin RN (1996) Perception of object unity in young infants: the roles of motion, depth, and orientation. *Cognitive Development* **11**: 161–180.

Johnson SP and Johnson KL (2001) Young infants' perception of partly occluded objects: evidence from scanning patterns. *Infant Behavior and Development* **23**: 461–483.

Johnson SP and Náñez JE (1995) Young infants' perception of object unity in two-dimensional displays. *Infant Behavior and Development* **18**: 133–143.

Kawataba H, Gyoba J, Inoue H and Ohtsubo H (1999) Visual completion of partly occluded grating in infants under 1 month of age. *Vision Research* **39**: 3586–3591.

Kellman PJ (1996) The origins of object perception. In: Carterette E and Friedman M (series eds) and Gelman R and Au T (vol. eds) *Handbook of Perception and Cognition: Perceptual and Cognitive Development*, 2nd edn, pp. 3–48. San Diego, CA: Academic Press.

Kellman PJ and Spelke ES (1983) Perception of partly occluded objects in infancy. *Cognitive Psychology* **15**: 483–524.

Koffka K (1935) *Principles of Gestalt Psychology*. London, UK: Routledge & Kegan Paul.

Kovács I (2000) Human development of perceptual organization. *Vision Research* **40**: 1301–1310.
Mareschal D and Johnson SP (2002) Learning to perceive object unity: a connectionist account. *Developmental Science*. (in press)
Marr D (1982) *Vision*. San Francisco, CA: Freeman.
Nakayama K and Shimojo S (1990) Toward a neural understanding of visual surface representation. *Cold Spring Harbor Symposia on Quantitative Biology* **40**: 911–924.
Singer W and Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–586.
Slater A (1995) Visual perception and memory at birth. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, vol. 9, pp. 107–162. Norwood, NJ: Ablex.
Slater A, Johnson SP, Brown E and Badenoch M (1996) Newborn infants' perception of partly occluded objects. *Infant Behavior and Development* **19**: 145–148.
Slater A, Morison V, Somers M *et al*. (1990) Newborn and older infants' perception of partly occluded objects. *Infant Behavior and Development* **13**: 33–49.
Spelke ES and Van de Walle G (1993) Perceiving and reasoning about objects: insights from infants. In: Eilan N, McCarthy RA and Brewer B (eds) *Spatial Representation: Problems in Philosophy and Psychology*, pp. 132–161. Oxford, UK: Blackwell.

Further Reading

Elman JL, Bates EA, Johnson MH *et al*. (1996) *Rethinking Innateness*. Cambridge, MA: MIT Press.
Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum.
Johnson MH (1997) *Developmental Cognitive Neuroscience*. Cambridge, MA: Blackwell.
Johnson SP (2001) Visual development in human infants: binding features, surfaces, and objects. *Visual Cognition* **8**: 565–578.
Kellman PJ and Arterberry ME (1998) *The Cradle of Knowledge: Perceptual Development in Infancy*. Cambridge, MA: MIT Press.
Mareschal D (2000) Object knowledge in infancy: current controversies and approaches. *Trends in Cognitive Sciences* **4**: 408–416.
Nakayama K, He ZJ and Shimojo S (1995) Visual surface representation: a critical link between lower-level and higher-level vision. In: Osherson DN (series ed.) and Kosslyn SM and Osherson DN (vol. eds) *An Invitation to Cognitive Science*, vol. 2: *Visual Cognition*, 2nd edn, pp. 1–70. Cambridge, MA: MIT Press.
Piaget J (1954) *The Construction of Reality in the Child*. New York, NY: Basic Books.
Richards JE (ed.) (1998) *Cognitive Neuroscience of Attention: A Developmental Perspective*. Mahwah, NJ: Lawrence Erlbaum.
Slater A (1998) *Perceptual Development: Visual, Auditory, and Speech Perception in Infancy*. Hove, UK: Psychology Press.
Spelke ES and Newport EL (1998) Nativism, empiricism, and the development of knowledge. In: Damon W (series ed.) and Lerner RM (vol. ed.) *Handbook of Child Psychology*, vol. 1: *Theoretical Models of Human Development*, 5th edn, pp. 275–340. New York, NY: John Wiley.

Perception: Overview

Introductory article

James R Pomerantz, Rice University, Houston, Texas, USA

CONTENTS

Perception is the complex sequence of processes by which we take the information received from our senses and then organize and interpret it, which in turn allows us to see and hear the world around us as meaningful, recognizable objects and events with clear locations in space and time.

INTRODUCTION

Definitions and Distinctions

Perception is the process by which we organize and interpret information about the world that has been collected by our sensory receptors. The story of perception begins outside the body with external stimuli – the physical energy in dabbles of light or pitches of sound – that happen to impinge on receptors in the eyes, ears, nose, tongue, skin and elsewhere. At those receptors the physical stimuli are converted into neural signals – into the language of our nervous system – by a process called transduction. The neural signals, which convey raw sensations, are in turn transformed into perceptions – that is, the images that we consciously experience and which are most often recognizable, meaningful and clearly placed in space and time. Without perception, we would no more recognize a familiar visual stimulus, such as the exterior of our house or the appearance of our grandmother, than does a camera, which can faithfully record patterns of light but has no understanding of what it is recording. Thus although both cameras and eyes detect and record patterns of light, it is the brain that analyzes these patterns into perceptions that tell us what is out there in the world around us.

Perception serves to interpret sensations in all of their various forms or modalities. Those modalities include vision (eyes), audition (ears), olfaction (nose), the tactile senses (skin), gustation (tongue) and the vestibular senses (inner ear). Note that each of these modalities itself embraces many components. For example, our eyes register not only the presence of light but also variations in light intensity (brightness), wavelength (color), location (edges and depth) and patterning over time (flicker and movement). Similarly, our skin registers pressure, temperature changes, pain, and so on.

Perception, aesthetics and survival

Perception is important because it is the source of virtually all that we know about the world around us – from the appearance of the evening sunset to the sound of a loved one's voice, and from the width of a pit we must leap across while on the run to the smell of food that has gone bad and should therefore not be eaten. Perception allows us to appreciate the joys and beauties of our environment, and so contributes to our aesthetic sense, including the appreciation of paintings and music. It also gives us some critical tools which we need in order to survive. Without perception, we would not last long in our sometimes hostile surroundings. Although certain of our perceptual abilities, such as color vision, may seem to be dispensable luxuries, in fact they have proved to be vital to our survival. Not only can we use color to identify foods that are safe to eat, but also color helps us to locate the edges of objects, which in turn speeds up our recognition of those objects and thus our reactions to them. Figure 1 shows how much easier it is to spot food in a colored image than in one containing only shades of gray.

Measurement

In any field of study, scientific observation requires a system of objective measurement – a scheme by which numbers can be assigned in meaningful ways to the structures and processes that we are studying. If perceptions are to be analyzed properly, they too need to be measured. If we want to know whether a human's eyes are as sensitive to

Figure 1. [*Figure is also reproduced in color section.*] Color helps us to differentiate between objects and to overcome camouflage. Notice how much easier it is to spot the edible food in the colored image.

light as are the eyes of a hawk, or whether they are as sensitive during the day as at night, we would need accurate measurements to answer these questions. If we want to know whether two bells that are rung simultaneously sound twice as loud as just one bell, again we need measurements. The complicating factor is that perceptions are private, subjective experiences which are locked up inside our individual minds. How then can we study and measure perception objectively?

Psychophysics

The branch of cognitive science that attempts to perform these measurements is called psychophysics, so named because it studies the mathematical link between physical entities and psychological responses. One of the earliest measurements ever to be attempted in the area of perception is called the threshold, which refers to the intensity level of the weakest stimulus (i.e. carrying the least amount of energy) that can be perceived by a human observer. Imagine a light bulb that is controlled by a rheostat (dimmer) whose knob can be rotated to increase or decrease the physical intensity of the light. In its simplest form, the experiment would involve the observer (you, for example) rotating the rheostat alternately clockwise, then anticlockwise, and so on until you were satisfied that you had found the exact point on the knob where the light crossed the boundary – the threshold – between visibility and invisibility. One would then measure the objective intensity of that just detectable light using physical instruments, and the resulting reading would give us the threshold measurement that we seek for our one observer (you).

Although they are simple in conception, these experiments become complicated in practice. If we repeat the knob-setting procedure many times with the same observer, we obtain slightly different answers each time, partly because our ability to detect light may change over time (e.g. as we adapt to being in the dark). Thus we must look at the entire distribution of measurements that we make and calculate an average. Moreover, the threshold that we measure depends on a number of additional variables. For example, the wavelength (color) of light that is used (500-nm wavelength is about the most detectable), where in the eye the light is shone (the fovea, or center of vision, is actually quite poor at detecting dim lights), and a host of other factors make a significant difference to the result that we obtain. In a classic experiment from the 1940s, it was discovered that under ideal circumstances the human eye is able to detect a light if as few as 10 rods (the commonest type of light-sensitive cell in the eye) are stimulated, and that absorbing merely a single quantum may be sufficient to activate one rod. In recent years, the idea of a strict threshold has been superseded by a different way of thinking, called *signal detection theory*, in which faint stimuli are proposed to occur against a background of ever-present sensory noise (think of random 'snow' on a television set tuned to a weak signal, or of static on a poor telephone connection), and in which processes akin to statistical decision-making are employed to help observers to determine whether they are perceiving genuine (albeit faint) signals or just the backdrop of noise.

Another question addressed by psychophysics involves scaling. The issue here is the rate at which a percept changes as the stimulus that produces it is changed. For example, if the physical intensity of a light were to be doubled, would it look twice as bright to the human eye? Experiments conducted over many decades have established that the answer to this is a firm no. Instead, the human perceptual systems are now known to be nonlinear. More precisely, they appear to follow a 'power-law' function, where the perceived brightness of

a light, for example, varies according to its physical intensity raised to the power 0.3. Although this may sound complicated, in fact it simply means that as the intensity of a stimulus is changed by some percentage, its percept changes by some other percentage. Thus as a light's physical intensity increases by a factor of 10, its perceived brightness increases by only a factor of 2 (10 raised to the power 0.3 is approximately 2). The exponents for different sensations vary widely. For example, for the perceived intensity of an electric shock the exponent is about 3.5, so that as the intensity of the stimulus is increased 10-fold, the perception of that shock increases by a factor of over 3000 (10 raised to the power 3.5 is 3162).

Eight Basic Facts About Perception

Although there is a virtually unlimited number of facts known about perception, they can be grouped into a smaller number of critical principles. The following eight facts cover the basics of what is known and widely accepted about perception today.

Perception is limited

Although it may seem that our senses allow us to perceive the complete world around us, in truth we perceive only the smallest fraction. For example, only the narrowest sliver of the electromagnetic spectrum (the band with wavelengths ranging from 400 to 700 nm), which we call light, is perceived by humans (Figure 2). The rest, including ultraviolet, infrared and radio waves, passes by us or through us undetected. The same is true for sound pitches with frequencies above 20 kHz or below 20 Hz, and with odors such as natural gas that our noses fail to detect. We are unable to detect events that occur either too slowly (e.g. the movement of the hour hand on a clock) or too quickly (e.g. a speeding bullet) to be noticed. The same is true of entities that are too small or too large to be seen with the naked eye (e.g. microorganisms that are only visible under a microscope).

Perception is selective

Despite these limitations with regard to what we can detect, our perception of the world is narrowed even further by the selectivity of our attention. At any one time we may be attending almost exclusively to one object in our environment (e.g. a person with whom we are speaking), while we ignore everything else (e.g. other conversations, the passing of nearby cars, the chirping of birds, the tightness of our socks on our feet). It often takes stimuli with special properties, such as the loud sirens and flashing lights on ambulances, to capture our attention. In some cases the selectivity of our perception is achieved by specialized systems or tools in our sensory systems. For example, eye movements help to direct our gaze, and eyelids help us to block out vision altogether. However, in other cases our selectivity is achieved largely in our brains, as for example in audition (where we have no ear movements or 'earlids' to help to direct our attention).

Perception refers to the distal stimulus, not the proximal stimulus

The distal stimulus is the physical object or event that is typically separated from you by some distance. The proximal stimulus is the image of that object or event (e.g. in light or sound patterns) that arrives at your sensory receptors and whose characteristics are partly determined by chance (e.g. viewing angle and illumination conditions). When you look at your grandmother wearing a red sweater, she is the distal stimulus. By contrast, the image of her that strikes your eye and is projected on to your retina is the proximal stimulus. As she

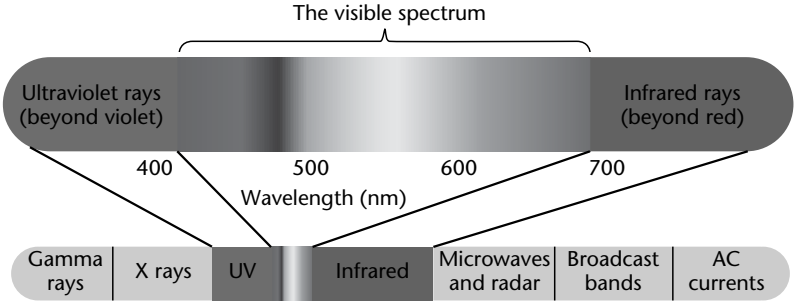


Figure 2. [*Figure is also reproduced in color section.*] The visible spectrum. Humans can see wavelengths in the range 400–700 nm, but they fail to detect energy outside that range. Note that only a small portion of the range of electromagnetic radiation gives rise to visible light. Also note that the hue depends on the particular wavelength.

walks towards you, her image expands on your retina, but you do not see her growing larger. As she walks out through the door into the sunlight, you do not see the door changing shape as it swings through a series of trapezoidal projections to your eyes. As the sun falls on your grandmother, you do not see her sweater changing brightness or color (although you may see the color more clearly). So important is our ability to keep our percepts constant despite changes in the proximal stimuli which we receive that we shall visit this matter again in the section on constancies.

Perception requires time

It seems as if we perceive objects and events in the world in real time (i.e. at the very moment when they happen or appear), but this is not the case. Instead, we perceive the world with a slight time lag, as it was a moment earlier. For example, the sluggishness of our visual system can be demonstrated by considering fluorescent lamps or television screens, which appear to be illuminated continuously but in fact are flickering (e.g. at 50 Hz in the UK and at 60 Hz in the USA). A flashlight whirled rapidly in the dark appears to form a complete circle of light – another side-effect of the sluggishness of vision. A particularly striking demonstration of this phenomenon, known as metacontrast, requires specialized equipment, but in summary it involves flashing a disk (a filled circle) of light, pausing briefly, and then flashing an annulus (ring- or doughnut-shaped) of light just encircling the point where the disk had appeared. When the timing is set just right, observers see only the annulus, and not the disk. Thus a later event prevents you from seeing an earlier event – something that could never happen if we saw the world in real time. Given that perception involves a sequence of processes, it should not be surprising that it takes time to operate.

Perception is not entirely veridical

The key word here is *veridical*, which means accurate, faithful or with high fidelity. Our perception of the world is so laced with inaccuracies (generally called illusions) that it may seem surprising that humans have survived as well as we have! Some of our misperceptions of the world are not the fault of our sensory or perceptual systems, but rather are caused by physical phenomena. For example, a pencil that is placed in a glass of water appears to bend (Figure 3). We should call this an optical illusion rather than a visual illusion, because it is caused by optics – light rays are being bent (refracted) when they pass from water into air. A camera would record this bending in exactly the same way as our eyes do. However, other illusions are clearly caused by properties of our perceptual system. Dozens of powerful geometric illusions have been discovered, such as that shown in Figure 4, but many illusions can be seen outside textbooks and laboratories. For example, the moon looks larger to the human eye when it is low in the sky, near the horizon, than when it is high in the sky, near the zenith.

Figure 3. A bent pencil in a glass of water: an 'optical' illusion whose cause lies in optics and not in faulty perception.

Many scientists believe that illusions should not be regarded simply as mistakes, but rather should be viewed as inevitable side-effects of shortcuts and assumptions made by our perceptual system that normally serve us well.

Perception requires memory

Although it may not be immediately obvious, we would not be able to perceive the world very well if we could not store our perceptions in memory. Consider the act of recognition (which means 'knowing again'). How could we possibly recognize our grandmother if we had no memory of what she looked, sounded and felt like? How could we see an object as moving or hear a pitch

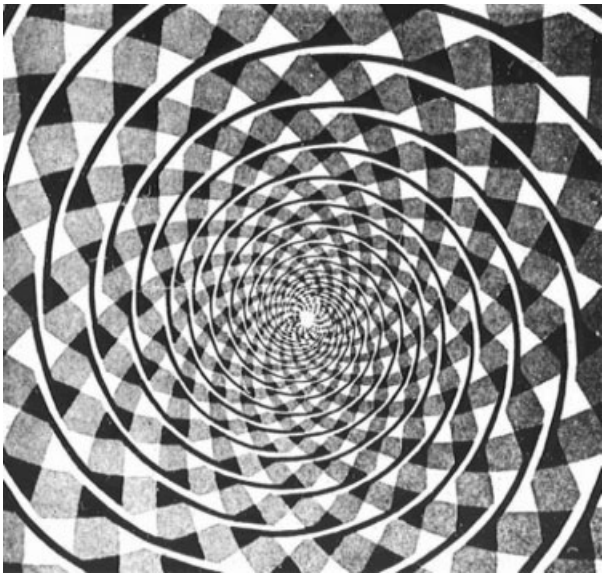


Figure 4. An example of a visual illusion that is not optical but originates in our sensory or perceptual system. The illusion here is that the apparent spirals are really concentric circles.

as rising if we had no moment-to-moment memory of where the stimulus had been a second earlier? We would be unable to compare objects for differences, to determine trajectories of moving objects, or to perceive speech or melodies if we had no perceptual memory. According to the most widely accepted theories, recognition requires the matching up of incoming sensory information with information that was stored earlier in our memory system. This could involve either an explicit memory store or alternative neural mechanisms that implicitly hold past information (as a sheet of paper 'remembers' a fold), but somehow information about the past is retained.

Perception requires internal representations

Early philosophers believed that when we look at an object, a tiny copy of that object enters our bodies through our senses. We now know better, but we also know that some process or structure in the system (probably a neural structure or circuit) is required for us to perceive the external world, and we call this an internal representation. According to this view, our sensory signals are processed to the point where the internal representation is activated. These internal representations can be regarded as the perceptual memories that were discussed in the previous paragraph.

Consider the perceptual imagery that many of us experience when we close our eyes and try to recall how many windows there were in the house where we grew up. Many scientists believe that such recollection activates the same internal representations that would be stimulated if we were to tour the house itself with our eyes open. According to prevailing theories of perception, objects and their sensory qualities are represented by the particular neurons (nerve cells) that they trigger in the brain. Although it is difficult to imagine that the conscious experience of the color green, for example, corresponds to nothing more than the activation of certain neurons in the brain, this nonetheless is the consensus view among researchers.

Perception is influenced by context

Stimuli rarely if ever appear in isolation. Instead, they appear in a context of other stimuli distributed in time and space, and those other stimuli can exert a powerful influence on our percepts. One familiar example is how short the referees in basketball games appear to be. In fact they are often taller than average, but when surrounded by the unusually tall players of this game, they certainly seem to be short. Similarly, a particular musical note may sound pleasant in the context of one chord, but grating (discordant) in another. In gustation, orange juice can taste delicious, but not immediately after brushing one's teeth with toothpaste (the added sugar of which temporarily masks the sweetness of the orange juice, thus making its sourness more prominent). These effects are so important to our understanding of perception that they will be discussed in a later section of this article (on context effects).

DIRECT PERCEPTION

Several theoretical frameworks have been proposed for understanding how perception works, but two of them dominate the field, namely *direct perception* and *unconscious inference*. Direct perception, which is most closely associated with the psychologist J. J. Gibson, may agree more closely with our intuitions or personal introspections. Briefly, it holds that our perception of the world around us is direct in the sense that there are no intermediate steps, no inferences, and no drawing on learned knowledge or thought-like processes for us to perceive the world. We simply see the world, the theory claims, rather than inferring or making guesses about it. In Gibson's framework, the observer and the environment are tightly coupled in such a way that changes in the latter are relayed faithfully and directly to the former. Gibson argued that there are structural invariants in the

environment that contain a wealth of information that an observer can use to determine the structure of the world.

One example Gibson used for a structural invariant is the texture gradient (Figure 5). When looking at such gradients, even when they are pictured on the flat pages of a book, observers tend to see the texture receding into the distance. Is this the result of an inference that the shift in grain size of the texture in the picture signals depth (perhaps an inference resulting from learning and experience)? Gibson did not think so. Instead he argued that texture gradients have consistent geometric properties which can be picked up directly by an observer and used to determine directly not only that the texture is receding in depth but also the precise angles and distances which that entails. One can judge that the two cylinders in Figure 5 are of about the same size because they cover about the same amount of brick on which they sit, proportionally speaking.

This example is rather abstract, so let us consider another that is more concrete. Suppose you are driving straight down a road and you see another automobile approaching yours from a side road running perpendicular to your stretch of road. How can you determine whether this car is on a collision course with yours as you head towards the intersection? Before you start reaching for your calculator and struggling to recall the appropriate formulae from trigonometry, be advised that there is an easier, 'direct' way to arrive at the answer. As you drive down the road, if the angle from your car to the other car does not change, this means that you are going to collide. It is a simple geometrical fact, and one that Gibson might argue typifies the way in which our perceptual system has evolved to pick up on these invariants (the above rule works regardless of the distances involved, the angle formed between the two roads, or the constant speeds of the two cars) and allow direct, immediate and accurate perceptions.

Figure 5. A texture gradient of the type described by Gibson.

UNCONSCIOUS INFERENCE

The other well-known theoretical framework for perception is unconscious inference, which is most closely associated with Herman von Helmholtz (and more recently with Irvin Rock, Richard Gregory and Julian Hochberg). In contrast to direct perception, unconscious inference holds (as its name implies) that perception takes place through various intermediate steps, and that those steps involve inference, guessing and other thought-like processes (although Helmholtz believed that we are unaware of these inferences, and he therefore described them as unconscious).

According to unconscious inference, percepts are constructed from evidence provided by our senses. This evidence is too vague and incomplete to serve as anything more than a set of clues so, like a detective, our perceptual system works from these clues to develop hunches or hypotheses about the world. These hypotheses then lead to implications that the perceptual system can further test to determine whether a hunch is right or wrong.

We are all familiar with situations in which sensory information is far from complete because viewing conditions are poor. Consider lying in bed in a darkened room and seeing a shadow on your wall, which you take to be an intruder in your room. Alarmed, you leap out of bed and switch on your light, only to discover that it was merely the faint shadow of an overcoat being cast on your wall that you were seeing. This would be a case of an incorrect hypothesis being generated from flimsy perceptual evidence (and moreover being evaluated by a drowsy detective). According to unconscious inference, this is how perception operates all of the time, even when the conditions for perception are excellent.

The evidence in support of unconscious inference comes from several sources, including experiments which show that what we perceive depends partly on what we expect to see. Those findings will be discussed later in this article. For now, it is useful to note that unconscious inference and direct perception are not necessarily mutually exclusive, as one could accurately describe how some perceptual functions work, while the other describes the remaining functions.

TOP-DOWN AND BOTTOM-UP PROCESSING

Top-down and *bottom-up* processing are terms that describe two ways in which perception could operate. In all likelihood human perception involves a mixture of the two. Let us begin with bottom-up processing, which is the more straightforward of the two.

Bottom-up Processing

Bottom-up processing begins with distal stimuli, which project proximal stimuli to the senses. The senses transduce (i.e. convert) the stimuli into neural codes, as for example when rods and cones in the retina convert incoming light into voltage differences within the photoreceptors (light-sensitive cells) – differences that ultimately result in neural firings (action potentials). These neural responses feed into other neurons that are organized to detect specific patterns. For example, if three rods arranged in a straight line on the retina activate one neuron, then if a strip of light is shone into the eye and falls exactly on those three rods, this would activate the neuron as much or more than would any other visual stimulus. We could then regard this neuron as a 'detector' for (or internal representation of) line segments of just the right orientation and position, and in fact neurons of this type have been discovered in the visual system of primates. In any case, perception in this scheme operates strictly bottom-up processing in that the proximal stimulus triggers a chain of events that reliably leads to an accurate perception. The flow of information processing proceeds strictly in the bottom-up direction, from the sensory receptors toward higher centers in the brain, with no signals traveling back in the top-down direction.

Top-down Processing

In top-down processing, the flow of information is reversed, at least in part, from bottom-up processing. Imagine closing your eyes and forming an image of your childhood home or of your grandmother. (Admittedly not all of us will be able to do this, but most can with little trouble.) If you succeed in this effort, your resulting experience – better regarded as imagery than as perception – will clearly be from the top down, since your sensory receptors (in the retina) are receiving no stimulation. In dreams, the images can be so vivid and convincing that we may even cry out in our sleep when the plot in the dream takes a terrifying turn. This tells us that top-down processing can sometimes yield results similar to and as realistic as those obtained from bottom-up processing.

In practice, top-down processing need not be so extreme as in dreams or hallucinations, where none of the information comes from our senses. Instead, consider a system that processes a little information from the senses, makes an inference about what may be out there in the world, and then proceeds to (1) fill in some of the missing information (recall the models of dinosaurs in museums of natural history, where most of the skeleton is plaster of Paris) and (2) generate hypotheses about what other sensory information might confirm or refute the current hypothesis or inference.

There is much evidence to support top-down processing, but none is more impressive than the subjective contours discovered by Gaetano Kanizsa and illustrated in Figure 6. Here we can see shapes that are not objectively present in the figure, such as the central 'square' that we can detect in this figure. We can also see brightness differences that are not objectively present (the central 'square' appears to be darker than its surround).

According to the principle of unconscious inference, these figures contain clues that are consistent with the presence of a square – clues such as dots coming to an abrupt halt (being truncated) in the middle of space, and edges that are separated but collinear (i.e. they line up exactly despite a physical separation). Although subjective contours could be detected by bottom-up mechanisms such as collinearity detectors, some subjective contours appear or disappear when viewers are given different instructions on how to interpret the image. Perhaps you can make the subjective square in Figure 6 vanish for yourself simply by prolonged viewing of or staring at certain areas. Our perceptual system may have evolved to treat such cues as evidence for the existence of a central figure, and now it proceeds with confidence to fill in the rest of the square, creating edges and brightness differences along the way. This 'filling in' is a good example of top-down processing that occurs in normal, awake

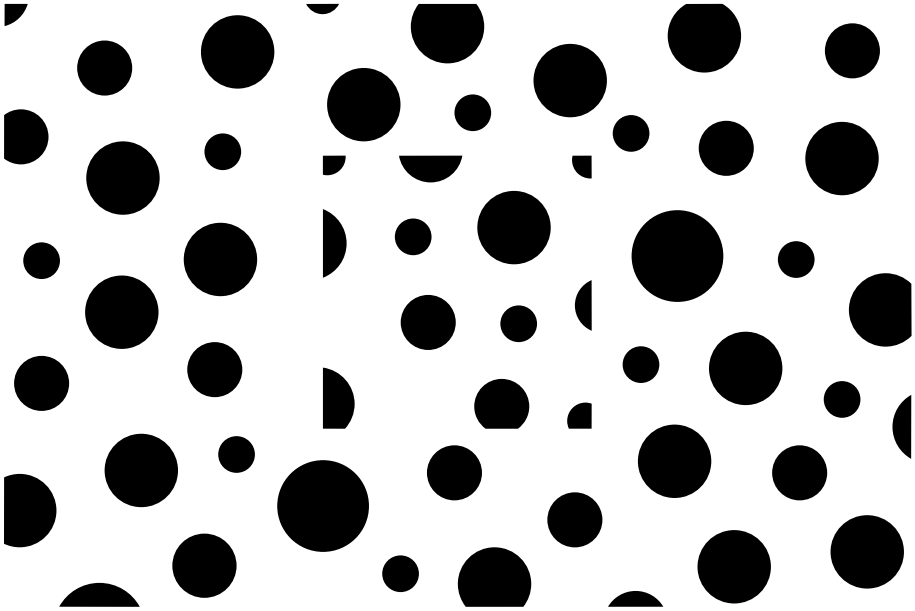


Figure 6. Subjective contours, discovered by Gaetano Kanizsa. The square patch floating in the center of this painting is illusory.

individuals perceiving the world around them. Visual experiences with the eyes closed, as in dreams or hallucinations, are more purely top-down in origin.

CONSTANCIES

As perceivers, we are more often interested in the constant, enduring properties of the distal stimulus than in the fluctuating and often accidental properties of the proximal stimulus. That is, we care more about the 'true' color of grandmother's sweater than about the hue it happens to assume when she walks out of bluish fluorescent light into yellowish natural sunlight. We care more about the objective shape of a door – normally a rectangle – than about the momentary trapezoidal shapes it projects to our eyes as we look at it from various angles.

The human perceptual system is equipped with clever and useful compensatory mechanisms to help us to achieve these constancies. These mechanisms embody a built-in understanding of the various couplings that pervade our perceptual world (e.g. that as an object moves away from us, the size of the image that it projects to our eyes diminishes; that as an object rotates in depth before us, the shape that the object projects changes in a regular manner; that as the illumination falling on an object increases, the amount of light it reflects to our eyes also increases). Thus if we stand next to a lamp and watch as a book is carried towards us, we are not fooled into thinking that the book is getting larger or brighter, nor are we deceived into thinking that the book is changing shape if it is rotating in depth as it approaches us.

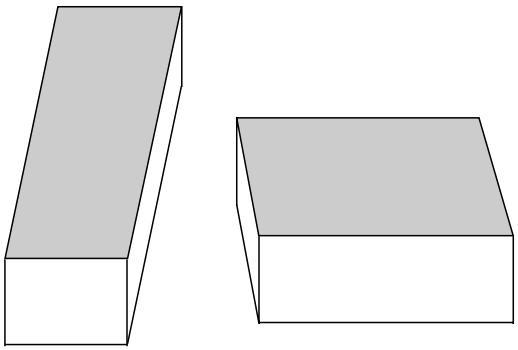


Figure 7. The Shepard box top illusion, caused by the functioning of our shape constancy mechanisms.

The constancies can exert powerful effects on our perceptions, as is illustrated in Figure 7. Look at the two objects depicted in this figure and compare the shapes of their top surfaces. Although they appear to be quite different, they are in fact identical in shape. The reason why they appear so different is that constancy mechanisms in our visual system allow us to look beyond the shape that three-dimensional objects project onto our two-dimensional retinas and focus instead on what the

original three-dimensional shape must actually be. Thus if one were to go about constructing two physical boxes that, if photographed, would produce a picture like that shown in Figure 7, one would discover that the two boxes would indeed need to have quite different-shaped tops. In a sense, our perceptual systems help us to work backwards from the proximal stimulus to our real object of interest, namely the distal stimulus.

CONTEXT EFFECTS

One of the most pervasive properties of human perception is that the appearance of any one stimulus is strongly dependent on the context in which it appears. As we noted above, most people would look short when surrounded by tall basketball players. However, context effects go far beyond perceptions of size. They influence almost every property to which we are sensitive.

Let us consider some examples. The first, which dates back at least to John Locke in 1690, involves the context of time rather than of space. In this demonstration, you begin by filling three bowls with water – one with hot (but not too hot!) water, a second with cold water, and a third with a lukewarm mixture of hot and cold water. Next, put one hand in the hot water and the other in the cold water, and leave them there for a minute or two, just long enough for them to adapt to the temperature. Then quickly plunge both of your hands into the lukewarm bowl. If you are like others, you will find that the lukewarm water feels cool to the hand that has been in hot water, whereas it feels warm to the hand that has been in cold water. The result is a paradox of sorts – you know that both of your hands are immersed in the same water, but it feels quite different to the two hands, because each hand perceives its current temperature in the context of what came before.

A second example involves context in space rather than in time, and is illustrated in Figure 8. Here the same gray patch looks lighter when placed on a dark background, and darker when placed on a light background. As with the two bowls demonstration, context leads to enhanced contrast, magnifying the difference and making us more aware of changes in our surroundings.

Figure 9 shows some additional examples of context effects. In Figure 9(a), a given stimulus can look like the letter B or the number 13, depending on whether you attend to the surrounding digits or the surrounding letters. In Figure 9(b), a curved line can look like the letter C or like a nose, depending on what surrounds it. In Figure 9(c), we are more likely to perceive a diamond as a tilted square when it is placed in a particular configuration of other diamonds.

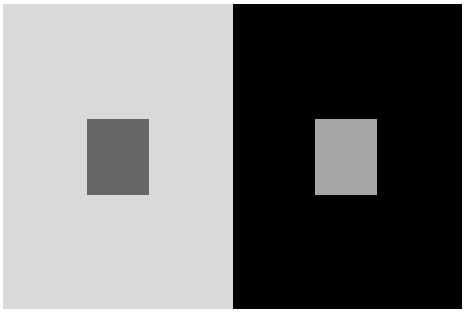


Figure 8. Simultaneous contrast, in which a given shade of gray looks darker on a light background and lighter on a dark background.

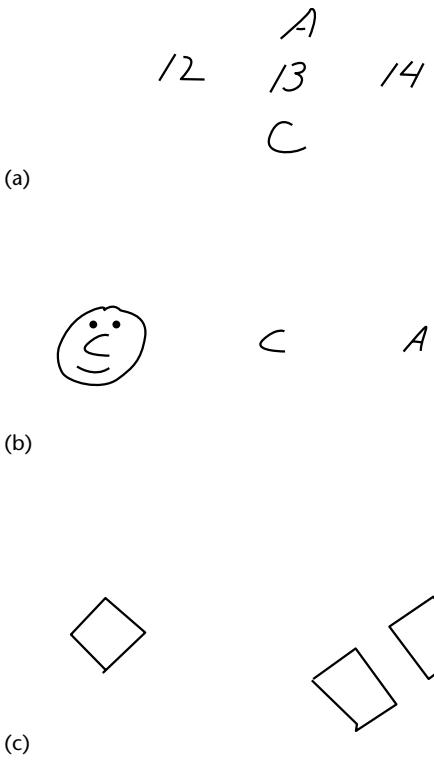


Figure 9. Context effects in visual perception. (a) A given figure can look like a letter or a number depending on whether one attends to the vertical elements or the horizontal elements of the display. (b) A given curve can look like the letter C or a nose, depending on the context. (c) A diamond can look like a tilted square in the right context.

One attempt to explain such context effects is known as Gestalt psychology, which is noted for claiming that, in perception, the whole is more than the sum of its parts. (More accurately, Gestalt psychology claims that the whole is simply *different*

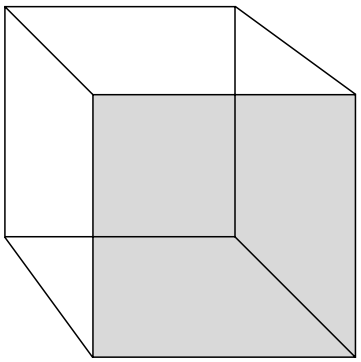


Figure 10. The Necker cube, which can be seen in either of two orientations.

to the sum of its parts.) According to the Gestalt approach, our primary perceptual experience is the configuration of or interrelationships between elements in the stimulus, rather than the elements themselves. Consider a melody as a set of relationships between musical notes. If we transpose a melody into a different key, the melody will remain the same even though all of the notes will have been changed.

Gestalt psychologists developed a number of principles that were intended to explain certain compelling perceptual phenomena. These include perceptual grouping (where physically separate elements in our field of vision appear to form clusters or groups), figure–ground segregation (where we must decide which side of an edge represents an object near to us and which side represents the distant background) and multistability (in which a pattern such as the Necker cube in Figure 10 appears to flip as one looks at it for an extended period of time). One general principle of Gestalt psychology is that we tend to organize our percepts into the simplest possible configuration (e.g. seeing four dots arranged in a square configuration as indeed forming a square or a circle, rather than forming more complex shapes that might also connect the four dots). Although there is much evidence to support a simplicity principle in perception, much evidence also points to an alternative notion that originated with Helmholtz. That notion, known as the likelihood principle, holds that our percepts are organized so as to maximize the chances of their being accurate, whether they are simple or not.

DISORDERS AND TREATMENTS

Our perceptual systems are sometimes described as 'transparent' because when they are operating correctly we hardly notice that they are there. We see through them as if they were clear glass. It is only when they fail us or act abnormally that we become acutely aware of their presence instead of taking them for granted. Losses in vision range from common and readily corrected refraction (focusing) errors such as myopia or astigmatism to partial or complete blindness. Likewise, losses in audition range from mild hearing loss and tinnitus (a ringing in the ear) to profound deafness.

In addition to these losses of our ability to perceive or resolve faint stimuli, there are other types of blindness that involve selected subsystems of perception. Those maladies in turn have taught us much about the way in which these systems work in healthy, intact individuals. Perhaps the commonest of these losses is color blindness, a genetically linked disorder that mainly affects males. Color-blind people rarely lack color vision altogether. Rather, they fail to see certain colors properly or at all. The most frequent type is red–green color blindness, but there are over a dozen other forms of this disorder. Importantly, red and green wavelengths of light also mix to form white (although red and green paints do not mix to form white paint, because paints are pigments that absorb light, so their combination is subtractive rather than additive as with lights). If we stare for a long time at a bright red square and then look at a plain white page, we see a green after-image (and conversely we see a red after-image after staring at a green square). The accumulation of evidence from diseases and disorders such as color blindness alongside the evidence from illusions such as colored after-images has contributed greatly to our scientific understanding of the way in which normal color vision works. With recent improvements in neuroimaging that allow us to identify areas of brain activation, researchers in cognitive neuroscience have learned much about the way in which vision and the other senses operate in the brain.

In addition to color blindness, researchers and physicians have documented blindness for motion, for stereoscopic depth, and even for stimuli as specific as faces. In the latter disorder, known as prosopagnosia, individuals with otherwise normal vision have great difficulty in recognizing the faces of people, even if they are famous or family members. Recent research using neuroimaging techniques (e.g. positron emission tomography and functional magnetic resonance imaging) supports the notion that specialized centers in the visual areas of the brain underlie the perception of motion and of faces, and that damage to those

areas can produce these forms of blindness. (*See* **Prosopagnosia**)

In addition to these disorders, there are others that are even more surprising and revealing with regard to the underlying processes of perception. For example, patients with blindsight can point accurately to the location of objects that they claim they cannot see. In typical cases, patients have an area of their visual field, known as a scotoma, where they cannot see light. In some respects scotomas are like the blind spot that we all possess in each eye, namely a tiny region lacking photoreceptors in which we cannot detect light, but a scotoma can be much larger, taking up half of the visual field or more. When their vision is tested, these patients detect perfectly well lights that are flashed in the functioning areas of their field of vision, but they fail to detect lights that are flashed in their scotoma. If they are asked to guess where these lights appeared, they are understandably reluctant to do so (the very question seems ridiculous), but their judgments turn out to be quite accurate. This suggests that the portion of their visual system that determines where objects are situated is separate from the portion that identifies objects and produces the conscious experience of perception.

Further Reading

Farah MJ (1999) *The Cognitive Neuroscience of Vision*. Oxford: Blackwell Science.

Gibson JJ (1966) *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Gregory RL (1997) *Eye and Brain: The Psychology of Seeing*, 5th edn. Princeton, NJ: Princeton University Press.

Helmholtz H von (1896) Concerning the perceptions in general. In: *Physiological Optics*, vol. III, section 26, pp. 1–36 (translated from the third German edition by JPC Southall). New York: Optical Society of America.

Hochberg JE (1978) *Perception*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Hubel DH and Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* **195**: 215–243.

Kanizsa G (1979) *Organization in Vision*. New York: Praeger.

Köhler W (1929) *Gestalt Psychology*. New York: Liveright.

Pomerantz JR and Kubovy M (1986) Theoretical approaches to perceptual organization. In: Kaufman L and Thomas J (eds) *Handbook of Perception and Human Performance*, pp. 36.1–36.46. New York: John Wiley & Sons.

Rock I (1983) *The Logic of Perception*. Cambridge, MA: MIT Press.

Perception: The Ecological Approach

Introductory article

M T Turvey, University of Connecticut, Storrs, Connecticut, USA

CONTENTS

The ecological approach to perception is a theoretical perspective on how animals, including humans, can be aware of their surroundings. It emphasizes the relevance of activity to defining the environment to be perceived.

INTRODUCTION

Perception refers to how animals, including humans, can be aware of their surroundings. The ecological approach to perception refers to a particular idea of how perception works and how it should be studied. The label 'ecological' reflects two main themes that distinguish this approach from the establishment view. First, perception is an achievement of animal–environment systems, not simply animals (or their brains). What makes up the environment of a particular animal – cliffs, caves or crowds – is part of this theory of perception. Second, perception's main purpose is to guide activity, so a theory of perception cannot ignore what animals do. The kinds of activities that a particular animal does – how it eats, moves and mates – are part of this theory of perception.

Including the environment and behavior as important parts of perceptual theory, rather than as afterthoughts, is clearly different from theories that start inside the eye (or the ear or the skin); but it is not necessarily controversial. None the less, the ecological approach is considered controversial because of one central claim: perception is direct. To understand the claim, and why some might consider it troubling, we have to contrast it with the more traditional view. Most scientists believe that perception begins with faulty input. For example, when objects in the world reflect light, the pattern of light that reaches the back of the eye (the retina) has lost and distorted much of the detail. The job of perception, then, becomes one of correcting the input and adding meaningful interpretations to it so that the brain can make an educated guess, an inference, about what caused the input in the first place. This traditional view is labeled as indirect perception because the animal's awareness of the world is a result of these intermediary steps. A theory of direct perception, in contrast, argues that the intermediary steps are only needed if the scientist has described the input incorrectly. Including the environment and activity in the theory of perception allows a better description of the input, a description that shows the input to be richly structured by the environment and the animal's own activities. This means that the intermediary steps are not needed and perception is direct.

GIBSON'S CONTRIBUTION

The ecological approach to perception originated in the work of the American psychologist James J. Gibson (1904–1979). Two biographical facts were important to shaping his theory. As the young son of a railroad employee in America's Midwest, he had spent many hours on trains watching the world flow by. He noticed that from a vantage point in the locomotive looking forwards the flow was outwards; from the caboose looking rearwards the flow was inwards. Here was the seed of his notion that the light that comes to our eyes is reliably structured by activity – structured light can be rich and meaningful. Many years later, after he had established a career as a perceptual psychologist, Gibson took time away from his college teaching position to spend 4 years as a scientist with the aviation psychology program during the Second World War. There he realized that the practical problems of takeoffs and landings, and pursuit and evasion, which could not only be mastered by 18-year-old pilot trainees but also performed routinely by birds and bees, had little to do with the physiology of the eyeball. Here was the seed of his notion that perceptual theory should try to explain

real-world behaviors (and not simply human behaviors). (*See* **Gibson, James J.**)

Upon resuming his job as a college professor, Gibson set about challenging the assumptions that he thought sat unexamined in most laboratory work, including his own. He argued that perceivers are aware of the world, not of their own sensations, and perception theory should respect that. Gibson's reformulation of the problem of distance perception illustrates the essence of what developed into his ecological approach. For centuries, scientists believed that 'distance is not perceivable by eye alone'. Indeed, if objects are treated as isolated points in otherwise empty space, then their distances on a line projecting to the eye are indistinct: each stimulates the same retinal location. Gibson dubbed this formulation 'air theory' (A in Figure 1) and argued that it was inappropriate for addressing how we see. His alternative was 'ground theory', which emphasized the contribution of a continuous background surface to providing rich visual structure (G in Figure 1). The simple step of acknowledging that points do not float in the air but are attached to a surface introduces a higher-order property, the gradient, which opened up the new possibility that perception might be veridical, that is, about facts of the world.

Gibson began to emphasize the enriching role of movement in perception. Once more, an ecological solution to an old problem is instructive. This problem concerned how a perceiver could distinguish object motion from his or her own motion. The puzzle arises because of the traditional assumption that the cue to perceived motion is the stimulation of successive retinal locations. An object moving from left to right fixated by a stationary eye will stimulate retinal receptors A, B, C, then D. That same object fixated by an eye moving from right to left will also stimulate retinal receptors A, B, C, then D. Since the retinal input is ambiguous, it must be compared with other input having to do with whether any muscle commands had been issued to move the eyes, the head or the legs. In the absence of counteracting motor commands, object motion is concluded; in the presence of such commands, the retinal signals would be counteracted, allowing the alternative conclusion of self-motion. Another possibility is that the observer is moved passively under somebody else's power (as in a train) so that other input and even knowledge must be taken into account. Gibson suggested an elegant alternative solution: overall, or global, change in the pattern of light is specific to self-motion; local change against a stationary background is specific to object motion. This simple insight (echoing the experience of the young Gibson riding the rails) opened a new field of research devoted to uncovering the structure in changing patterns of light: optic flow.

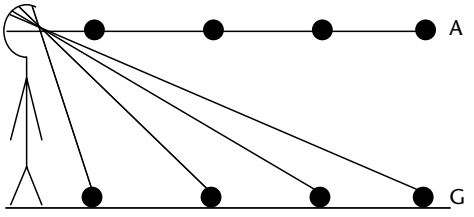


Figure 1. The distances of four points in the air (A) are indistinct because their projections are onto the same retinal location. As texture elements on the ground (G), the gradient of their projections distinguishes their relative depth.

OPTIC FLOW AND THE VISUAL GUIDANCE OF LOCOMOTION

Optic flow refers to the patterns of light, structured by particular animal–environment settings, available to a point of observation. The goal of optic flow research is to discover particular reliable patterns of optical structure, called invariants, relevant to guiding activity. Outflow and inflow are distinct forms of optic flow (distinct flow morphologies) that tell us whether we are moving forwards or backwards. As scientists consider how that flow is structured by the variety of clutter that we encounter as we move around – doorways, hillsides and the like – they discover invariants specific to those facts as well (Figure 2).

In order to effectively guide their activities, animals need to know more than simply what they are approaching. They also need to know how they are approaching – are they moving too fast? – and whether they need to adjust that approach: should they slow down? Turn? As a busy waiter rushes towards the swinging door of the restaurant kitchen, he makes subtle adjustments to his behavior in order to control his collision. He needs to maintain enough speed to push through the door but not so much that he crashes into it. Effective behavior requires that he should know when a collision will happen (so he does not slow down too early) and how hard the collision will be (so that he slows down enough). Optical structure relevant to these facts has been identified and provides examples of quantitative invariants. The optical quantity called τ is specific to when a point of observation will contact an upcoming surface. Consider the rectangular contour in Figure 2(a). As the waiter approaches it, its optical projection

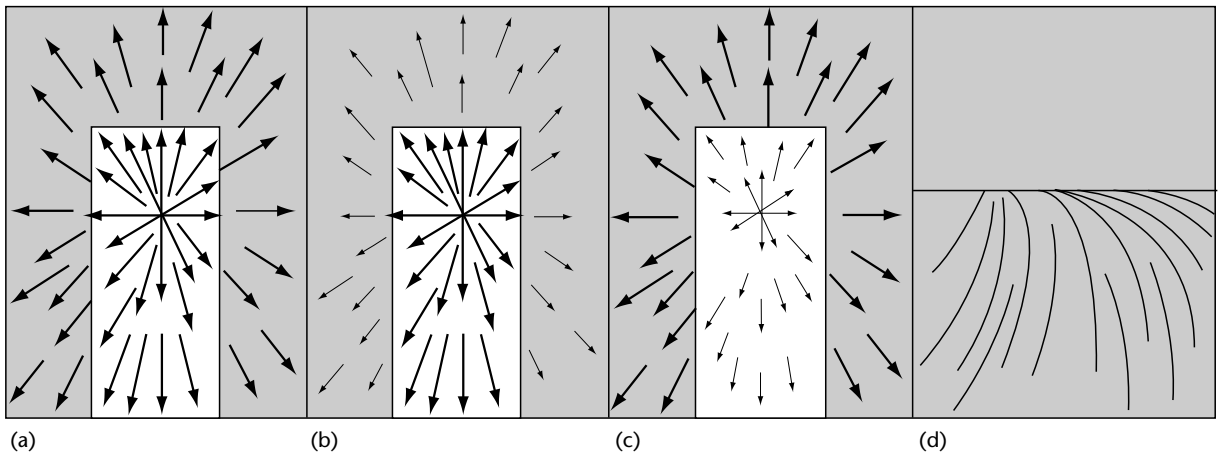


Figure 2. (a) Global optical expansion, which accompanies forward locomotion, is represented as a velocity vector field radiating from a focus of expansion. Smooth flow specifies approach to a flat surface. (b) Faster expansion inside a contour specifies an obstacle. (c) Faster expansion outside a contour specifies an opening. (d) Curvilinear flow specifies direction of heading.

magnifies. The speed of approach affects the rate of expansion, that is, the change in optical area per unit time. The quantity τ is given by the inverse of the relative rate of this expansion – how long will it take until there are no units of time left. As he slows down (or speeds up), the rate at which τ approaches zero changes. The rate of this change (that is, the derivative of τ) is specific to how severe the collision will be. It essentially quantifies for a moving observer (the waiter) whether his kinetic energy is being reduced (by braking) at a rate sufficient to stop movement before contact occurs.

The preceding descriptions of global optical structure refer to situations in which the observer is approaching a surface. However, they are also relevant to a surface, such as a projectile, approaching the point of observation. Local disturbances of optical structure relevant to the guidance of interceptive behavior can also be described in terms of τ and its derivative, specific to when and how hard a collision will be. Other optical quantities are relevant to moving the perceiver (or the perceiver's hand or racket) into a position to intercept (or avoid) the projectile. Moreover, the same invariants are available to the family dog catching a ball, a chickadee landing on a bird feeder, and a bumblebee searching for pollen. Although these creatures have obviously different visual systems and brains, the information relevant to guiding their behaviors is the same. Invariants of tissue deformation and sound compression waves reveal the same richness that has been found in optical structure. Of course, while all creatures need to perceive things such as openings and obstacles to locomotion, what counts as an opening necessarily differs. How perception is 'personalized' is addressed in Gibson's notion of affordances.

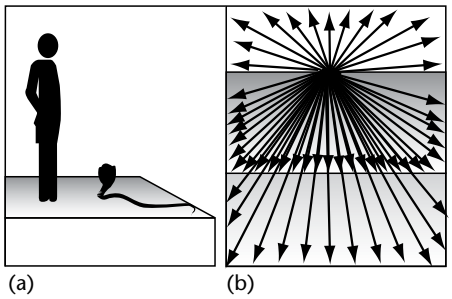


Figure 3. (a) A brink in the path of locomotion presents different behavioral possibilities to animals of different sizes and locomotory styles. (b) For an animal with legs, the ratio of flow above the horizontal contour of the brink relative to the rate of flow below scales the brink to the animal's eye height, specifying whether it can be descended safely.

AFFORDANCES

In highlighting the relevance of optical structure to whether an activity is possible ('Is the ball catchable?') the preceding examples introduce what might be the most radical contribution of this theory: behavioral possibilities are perceived. Gibson coined the word 'affordances' to refer to the possibilities for action of a particular animal–environment setting. They are what an arrangement of surfaces means to an animal. Affordances are usually described as '-ables', as in 'catchable', 'pass throughable', 'climbable', and so on. Whether

a ledge, for example, is a stepping-down place or a falling-off place is not determined by its absolute size or shape but how it relates to a particular animal, including that animal's size and agility, and style of locomotion (Figure 3(a)). Gibson proposed that such activity-relevant animal–environment relations are specified in the optics (Figure 3(b)).

Through their research on affordances, Gibson and his followers 'ecologize' the traditional problems of size, distance and shape perception. They manipulate aspects of the environment (such as the height of a flat surface, or the distance of a target) and ask people to evaluate whether a certain behavior would be possible. Instead of asking for estimates of size or distance in absolute units, they ask (for example) whether the surface could be climbed or the target could be reached. Ideally, experimenters would prefer that people's responses be as natural as possible (that is, without making conscious decisions). When the technology allows, therefore, active behavioral adjustments are recorded. Either way, we have learned that children, adults and the elderly all perceive possibilities for, and restrictions on, activity. They are well aware of behavioral category boundaries (this is climbable; that is not). Even though the absolute boundary – measured in centimeters – differs for people of different sizes, the functional boundary – scaled by something like the person's eye height – is the same. If the eye height is manipulated in some way (e.g. by having the person wear platform shoes or by surreptitiously raising the floor of the room they view through a window) it has predictable consequences for their affordance evaluations.

In summary, the theory of affordances is radical because it treats relational properties as objective: the relation between the ledge size and animal size in Figure 3 exists even when it is not being perceived. Moreover, such relational properties are considered more fundamental to describing environments and understanding perception than the objective primary properties studied so far by physics. According to this ecological theory, meanings are directly perceptible. The wide-ranging species inhabiting our planet, from simplest bacterium to human, do not invent meanings (by conception), they discover them (by perception).

Further Reading

Gibson JJ (1986) *The Ecological Approach to Visual Perception*. Mahwah, NJ: Erlbaum.

Reed ES (1988) *James J. Gibson and the Psychology of Perception*. New Haven, CT: Yale University Press.

Perception, Gestalt Principles of

Intermediate article

James T Enns, University of British Columbia, Vancouver, British Columbia, Canada

CONTENTS

Gestalt principles of perception are those aspects of perception that cannot be defined in terms of smaller, constituent elements. Gestalt *is a German word which translates in English as whole, configuration, or pattern.*

WHOLES AND SUMS OF PARTS

The visual world we experience is of people, trees, buildings, coffee cups, and other objects. Yet, light shining onto the eye from these objects produces regions of abrupt changes in luminance and color (edges), regions of gradual changes (shading), and regions of relative uniformity (shapes). The problem of how the brain analyzes these sensory events so as to provide us with the experience of objects is the problem of *perceptual organization*.

Perceptual organization was first identified as a central problem by the Gestalt school of psychology in the 1920s. Among the most influential in this school were Max Wertheimer, Kurt Koffka, and Wolfgang Köhler. They emphasized that the perceived object is different from, and sometimes even more than, the sum of its sensory parts. Modern researchers tend to divide perceptual organization into the smaller subproblems discussed below.

Edge Perception

Edges are the building blocks of vision. When eye movements are prevented, so that edges are no longer stimulating the eye over time, we quickly lose all visual experience. Figure 1 allows you to experience something like a stabilized image by staring at the center dot for about one minute. As you stare at the dot, the fuzzy gray border will begin to disappear because it stimulates a part of your eye that is sufficiently distant from the high-resolution centre that it is unable to register this fuzzy edge. As far as your brain is concerned, the fuzzy border no longer exists, and so the brain concludes that the light gray region (which is defined by a sharp border) is continuous.

Edges interact in cooperative and competitive ways (Field and Hayes, 1993). This can be seen at work in Figure 2, where a large O-shape has been drawn in two ways. In Figure 2(a), the easily seen O is defined by line segments whose extensions form a continuous curved arc. In Figure 2(b) the same O-shape has been drawn, using the same number of line segments, but the line segments now lie at a 90-degree angle to the arc of the curve. The O is now very difficult to see and you will have to inspect individual segments to confirm it. This illustrates the cooperation that exists between nearly continuous edges and the competition that exists between edges of different orientation.

Feature Extraction

When observers are asked to respond rapidly to image features, it is apparent that certain features

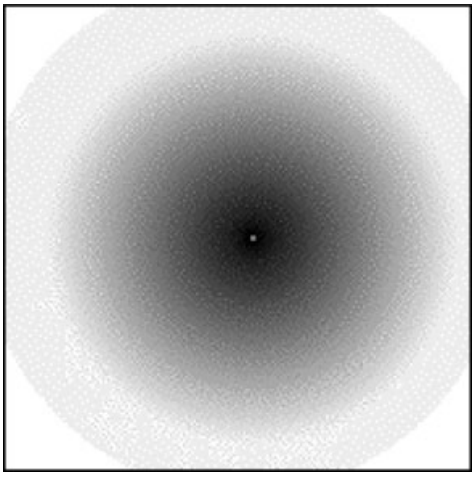


Figure 1. Stabilizing a fuzzy edge on the retina. Stare at the central dot for about a minute with one eye. The fuzzy inner edge will soon disappear and be replaced by a uniform light gray.

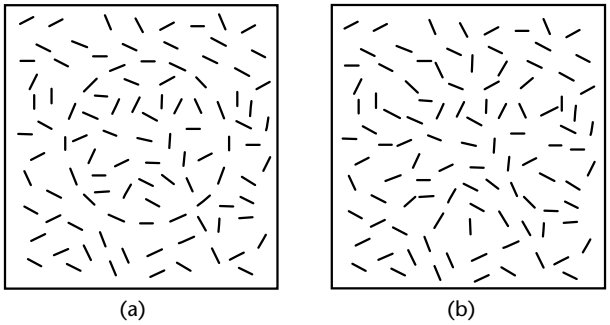


Figure 2. O-shaped configurations defined by (a) nearly continuous line segments, and (b) the same number of line segments oriented at 90 degrees to the continuous curve.

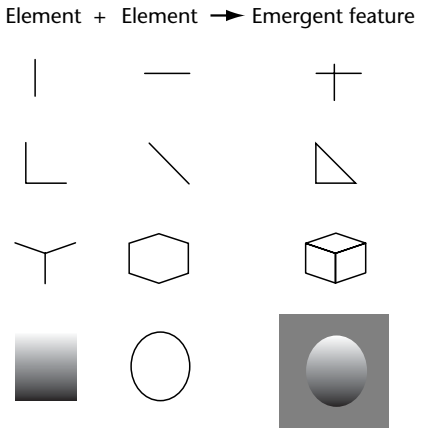


Figure 3. Component elements that create emergent features when combined.

are more basic, including luminance, color, the principal axis of a shape, and shape curvature. Tasks that require observers to respond rapidly to arbitrary combinations of these features are generally done slowly and with many errors (Beck, 1982; Julesz, 1984; Treisman, 1986). However, there are some combinations that do permit rapid analysis. These are *emergent features* since they form *gestalts* that cannot be predicted from the component features (Pomerantz, 1986). Examples are shown in Figure 3 (Enns and Rensink, 1991; Nakayama and Silverman, 1986; Ramachandran, 1988).

Figural Assignment

The assignment of some shapes as belonging to objects that are important to the observer and other shapes as belonging to the background is one of the crowning achievements of vision. It permits actions that are appropriate to objects in the

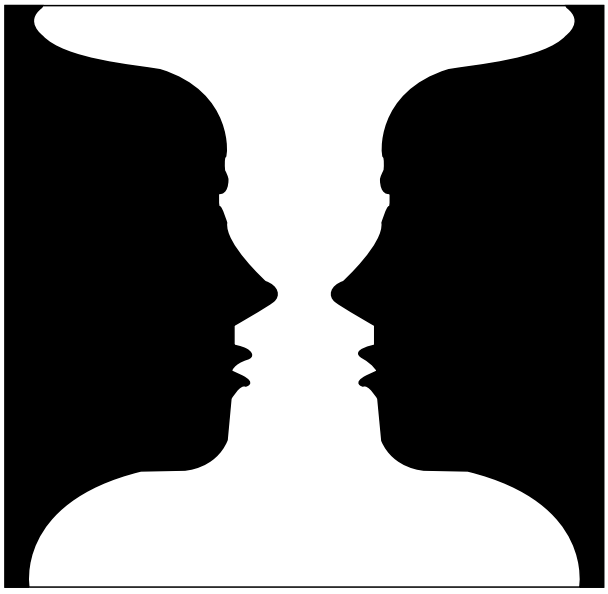


Figure 4. A display that can be seen either as a pair of black faces or as an ornate white vase.

environment, based solely on vision. Some of the complexities involved in the analysis are illustrated in Figure 4, a version of the ambiguous face–vase figure made famous by Rubin (1921). Observers of this display may see a pair of silhouette faces gazing at each other or they may see an ornate vase. The two interpretations also alternate in consciousness over time, illustrating that the assignment of *figure* and *background* is a product of brain processes, and not one determined by the image.

Three-dimensional Structure

Once a three-dimensional interpretation has been made of the figures in an image, the object can be acted upon by the motor systems of the organism. The leading theories in this area have proposed that shapes in the image are assigned to *volumetric primitives*, or simple, convex solids that comprise the three-dimensional building blocks of vision (Biederman, 1987).

PERCEPTUAL FIELD

Perceptual field has two distinct meanings for Gestalt perception. An older sense derives from the early Gestalt theories of brain processes as analogous to electrical force fields. As illustrated in Figure 5(a), Gestalt theories claimed that a single dot viewed on an otherwise blank page, in analogy to a charged particle, had a force that spreads out uniformly in all directions. They hypothesized that

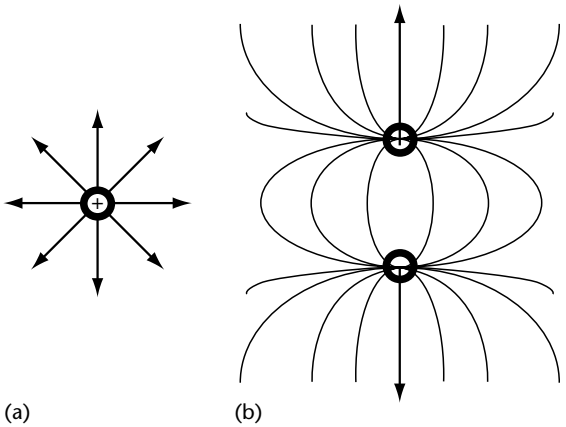


Figure 5. (a) A single particle with a positive electrical charge. (b) The same particle along with a nearby particle with a negative charge. Lines illustrate the electrical field forces.

if a second dot were added, as in Figure 5(b), the two fields of energy would interact. The Gestalt psychologists hoped that such neural interactions would begin to explain perception. However, this theory failed because neuroscientists were unable to find brain mechanisms that followed these patterns. Instead, the dominant metaphor for neurons is the electronic circuit, which means that neurons can perform simple arithmetic operations. In the modern view, neurons do interact with one another in excitatory and inhibitory ways but these interactions do not resemble the spread of energy around magnetic dipoles.

A second sense of *perceptual field* refers to the spatial region over which vision functions. Only a subset of the entire field of view is at any moment contributing to the performance of the task. Some researchers refer to this as the *useful field of view*. Consider how the perceptual field differs depending on the task. If the task is to detect a bright light flashed anywhere in the visual field, the visual field is usually estimated to be almost 180 degrees wide and 100 degrees high. However, if the task is to read letters in a standard row of text, without moving the eyes, then the perceptual field may shrink to only 4–6 degrees wide.

Modern research emphasizes the role of *attention* in perception. Where an observer attends, and what the observer expects to see there, has a powerful effect on what is seen. This dynamic aspect of perception is studied in tasks where observers are asked to detect changes in scenes that are separated by a shift in viewpoint (Simons, 2000) or a short temporal interval (Rensink *et al*., 1997). The main finding is that observers are 'blind' to changes in objects they were not attending to. In other studies, observers perform a visually demanding task and then are suddenly and briefly shown a surprise stimulus. The detection of these unannounced stimuli is so infrequent that the term *inattentional blindness* has been coined (Mack and Rock, 1998). These studies emphasize the 'piecemeal nature' of perception (Hochberg, 1982) and that very little of the available information is consciously registered (Shapiro, 1994).

GROUPING

Which parts of an image belong together and which belong to different objects? This is the problem of grouping that early Gestalt psychologists are best known for. Wertheimer (1923) went so far as to call his answers 'laws', since he believed them to be as reliable as other physical laws. Today reseachers tend to use the term *principle* to reflect their important, but not absolute, influence on perception.

The classic Gestalt principles were illustrated using the elements of dots and line segments. They included the principle of *proximity* (elements that are relatively close tend to be seen as belonging to the same perceptual unit), *similarity* (elements that are relatively similar tend to be seen as belonging together), *good continuation* (elements following a simple curve will tend to be grouped together), *common fate* (elements with common motion direction will tend to cohere), and *closure* (elements which enclose a region will tend to be seen as belonging to the same figure).

Although these principles operate reliably when applied to the drawings of Gestalt psychologists, they are difficult to apply in other settings. There are several reasons for this failure. First, the 'elements' of perception were never defined. Unfortunately, natural images are not coded in the sharply defined dots and lines used in the drawings of Gestalt researchers. Second, the features to which 'similarity' and 'common direction' should be applied were not defined, making the basis of similarity unclear. Third, the spatial region over which the laws applied was not defined. It was simply assumed that each drawing in its entirety would submit to the principles in a coherent manner. This ignores the importance of attentional limits on perceptual field size (Hochberg, 1982).

Some limitations of the classic Gestalt view are being addressed by modern research. For instance, one proposal defines the elements of an image as regions of *common brightness* (Palmer and Rock, 1994). Elements so defined are then grouped further based on a principle of *common region*

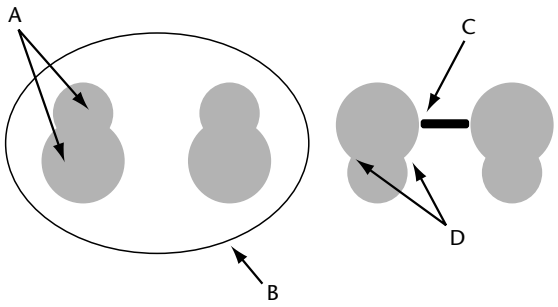


Figure 6. Illustrations of several new principles of grouping (Palmer and Rock, 1994). (A) Common brightness: the two 'bubbles' are grouped because they share brightness. (B) Common region: the two pairs of bubbles are grouped because they are enclosed by a common outline. (C) Element connectedness: the pairs of bubbles are grouped because they are joined by a bar. (D) Image segmentation based on concavities: the two bubbles are seen as separate objects because of the concavities in the surrounding edge.

(elements located within the same enclosed region of space will tend to be grouped) and *element connectedness* (elements in physical contact with other elements will tend to be grouped). However, the initial elements defined by common brightness may also be parsed in greater detail, using the heuristic of segmentation at *regions of deep concavity*. Examples of these principles are shown in Figure 6.

PROXIMITY

An important question for modern research is the perceptual space in which grouping occurs. For example, is the relevant distance for grouping by proximity the retinal image or distance in three-dimensional space? This question has been studied by creating a dot lattice using luminous beads attached to threads, as shown in Figure 7 (Rock, 1997). If observers see the dots as lying in a plane perpendicular to the line of sight, then grouping is influenced primarily by the proximity between dots as measured in the retinal image. On the other hand, if they perceive the dots as part of a surface receding in depth, then the apparent proximity of the dots on the surface determines whether row or column grouping is seen. Related experiments have studied the relations between perceived proximity and the perceptual completion of partially occluding objects. In all cases, it is clear that grouping mechanisms based on proximity work in cooperation with other processes such as depth perception. Similar findings have been reported for grouping by temporal proximity (Lee and Blake, 1999).

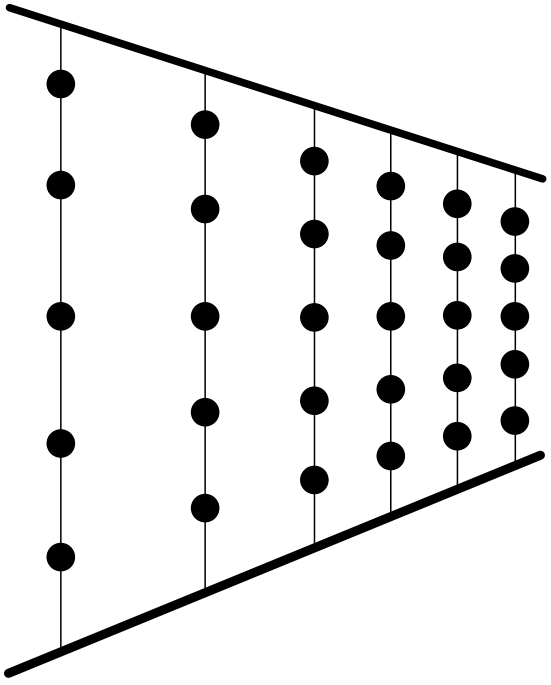


Figure 7. Grouping by proximity depends on whether lattices such as these are seen as lying perpendicular to the line of sight or as receding in depth.

SIMILARITY

Whereas early Gestalt researchers assumed that similarity was based on such obvious properties as color, brightness, and shape, more recent studies have shown that this too is task-dependent. Consider, for example, whether an L-shape or a tilted T-shape is more similar to an upright T-shape. When observers are shown this triplet of elements in isolation, they tend to agree that the tilted and upright Ts are more similar to each other than either one of these two is to the L-shape. However, as shown in Figure 8, when large numbers of these shapes are used to form a texture, borders are formed where regions of uniform elements meet one another. Now it is clear that the strongest border lies between the upright Ts and the tilted Ts (Beck, 1982). This demonstrates that similarity varies with the visual task. Seeing a dense texture of elements leads to an emphasis on differences in element orientation, whereas seeing elements in isolation permits an analysis of the spatial relations among component segments.

In studies of perceived brightness and similarity, it has been shown that grouping depends on the

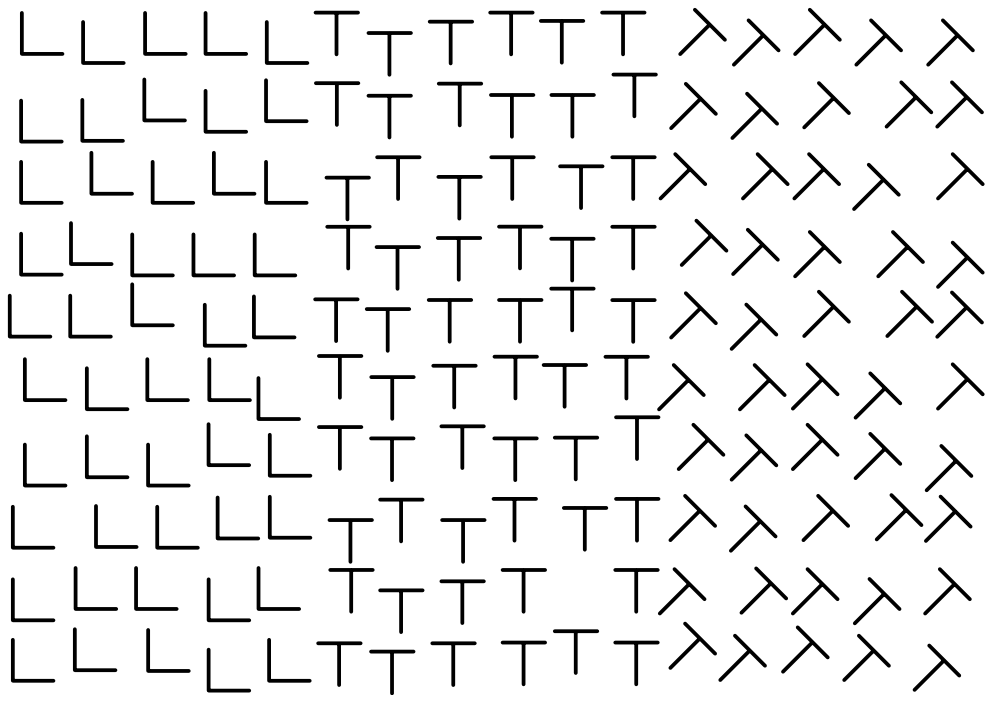


Figure 8. Grouping by similarity in dense textures follows different rules from grouping by similarity of individual texture elements.

apparent brightness, not on the physical luminance intensities of elements (Adelson, 1993; Rock, 1997). This means that observers take into account the role of shadows and assumed light sources in their evaluation of similarity based on luminance. As such, it also demonstrates once again that visual grouping does not operate in isolation from other perceptual processes, especially those involved in the interpretation of an image as being generated by light reflecting from three-dimensional surfaces.

References

Adelson EH (1993) Perceptual organization and the judgment of brightness. *Science* **262**: 2042–2044.

Beck J (1982) Textural segmentation. In: Beck J (ed.) *Organization and Representation in Perception*, pp. 285–318. Hillsdale, NJ: Lawrence Erlbaum.

Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychological Review* **94**: 115–147.

Enns JT and Rensink RA (1991) Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review* **98**: 335–351.

Field DJ and Hayes A (1993) Contour integration by the human visual system: evidence for a local 'association field'. *Vision Research* **33**: 173–193.

Hochberg J (1982) How big is a stimulus? In: Beck J (ed.) *Organization and Representation in Perception*, pp. 191–218. Hillsdale, NJ: Lawrence Erlbaum.

Julesz B (1984) A brief outline of the texton theory of human vision. *Trends in Neuroscience* **7**: 41–45.

Lee S-H and Blake R (1999) Visual form created solely from temporal structure. *Science* **284**: 1165–1168.

Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

Nakayama K and Silverman GH (1986) Serial and parallel processing of visual feature conjuctions. *Nature* **320**: 264–265.

Palmer SE and Rock I (1994) Rethinking perceptual organization: the role of uniform connectedness. *Psychonomic Bulletin & Review* **1**: 29–55.

Pomerantz JR (1986) Visual form perception: an overview. In: Schwab EC and Nusbaum HC (eds) *Pattern Recognition by Humans and Machines*, vol. 2: *Visual Perception*, pp. 1–30. Orlando, FL: Academic Press.

Ramachandran VS (1988) Perceiving shape from shading. *Scientific American* **259**: 76–83.

Rensink RA, O'Regan JK and Clark JJ (1997) To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* **8**: 368–373.

Rock I (1997) *Indirect Perception*. Cambridge, MA: MIT Press.

Rubin E (1921) *Visuell wahrgenommene Figuren*. Copenhagen, Denmark: Glydendalske.

Shapiro KL (1994) The attentional blink: the brain's eyeblink. *Current Directions in Psychology* **3**: 86–89.

Simons DJ (2000) Attentional capture and inattentional blindness. *Trends in Cognitive Science* **4**: 147–155.

Treisman AM (1986) Features and objects in visual processing. *Scientific American* **255**: 114B-125.

Wertheimer M (1923) Principles of perceptual organization. Abridged translation by M Wertheimer. In: Beardslee DS and Wertheimer M (eds) *Readings in Perception*, pp. 115–137. Princeton, NJ: Van Nostrand-Reinhold. [Original work published 1923, *Psychologische Forschung* **41**: 301–350.]

Further Reading

Coren S, Ward LM and Enns JT (1999) *Sensation and Perception*, 5th edn. New York, NY: Harcourt Brace.
Marr D (1982) *Vision*. San Francisco, CA: WH Freeman.
Palmer SE (1999) *Vision Science*. Cambridge, MA: MIT Press.

Perception, Haptic

Introductory article

Roberta L Klatzky, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
Susan J Lederman, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

CONTENTS

Haptic perception is that derived from the sense of touch.

INTRODUCTION

The word 'haptic' comes from a Greek term meaning 'able to lay hold of'. Another word of Greek origin, 'stereognosis', is often used synonymously, particularly in medical contexts. Haptic perception is based on combined sensory inputs from the skin, muscles, tendons, joints and mucosae exposed to the environment (particularly in the mouth). Although these inputs could arise passively, as when an object is pressed against the skin, more commonly they result from active, purposive touch. Special receptors in these sites respond to stimulation from external surfaces or from a person's own movement. Receptors that lie within the skin (cutaneous) include mechanoreceptors, which respond to pressure; thermal receptors, which respond to thermal changes; and nociceptors, which respond to high-intensity, noxious stimulation such as sharp pricks, extreme pressure or very hot or cold surfaces. The word 'kinesthesis' is used to describe perception based on mechanoreceptors in muscles, tendons and joints (and also in the skin of the hand), which give rise to perceived movement, position or strain in body parts.

The cutaneous receptors comprise four types, as defined by the size of the skin surface over which they detect stimulation (receptive field size) and by how quickly they cease firing when stimulation remains constant (rate of adaptation). For example, a class of mechanoreceptors called slowly adapting type I (SA-I) has a small receptive field and adapts slowly, so that it identifies a small area of skin and gives a sustained response when pressure is applied there. These attributes enable it to provide information about small raised patterns that might be touched with the fingertip, like a letter of the Braille alphabet. In contrast, another class of mechanoreceptor, called fast-adapting type II (FA-II), has a large receptive field and stops responding quickly unless mechanical contact changes. This class of receptor responds best to stimuli that repeat rapidly, like the throat of a cat as it purrs. The two remaining classes of mechanoreceptor are rapidly adapting with small receptive fields (FA-I), and slowly adapting with large receptive fields (SA-II). The receptors in muscles and tendons appear to provide information about the positions and movements of the joints. Receptors in the tongue and oral cavity allow people to discriminate food texture, viscosity and temperature, and contribute to the enjoyment of eating.

Much psychological research has investigated the minimum amount of cutaneous stimulation that leads to a conscious sensation – the threshold. Thresholds have been measured, for example, to determine sensitivity to warmth or pain, presumably tapping the inputs from thermoreceptors and nociceptors, respectively. The spatial resolution threshold can be measured by the minimum detectable gap between two points pressed into the skin. Cutaneous thresholds vary somewhat with factors such as air temperature, whereabouts on the skin the stimulus occurs, and the area of skin contacted. Kinesthetic thresholds such as the minimally detectable change in the position of a finger have also been determined. Like most sensory modalities, the haptic system exhibits a general decrease in sensitivity, or heightened thresholds, as people age. Threshold-level responses have implications for haptic perception more generally, as they indicate the intensive, spatial and temporal sensitivities of the system.

HAPTIC PERCEPTION OF SPACE WITHIN REACH OF THE HANDS

In general, perception is not veridical; systematic errors arise. People visually perceive objects at a

distance as nearer than they are (foreshortening). Similarly, various illusions and distortions have been found when people feel shapes in the space within reach of their hands. Some illusions found in touch have counterparts in visual space perception. An example is the Müller–Lyer illusion, in which a horizontal line looks shorter when arrow heads are presented at both ends than when the arrow fins are presented at both ends. Another illusion found in both vision and touch is the oblique effect: it is easier for people to judge the orientation of a rod that is oriented vertically or horizontally than when it is oriented at an oblique angle (such as 45° from the vertical). The analogy between these effects across vision and touch suggests that they might be based on a common underlying representation of space, which can be formed from viewing objects or from feeling them. However, the haptic perceptual system appears capable of producing a variety of representations for any one spatial object, because the way in which the stimulus is explored can have a pronounced effect on the extent of these illusions.

The movements used in exploring space can directly produce some distortions in its perception. For example, when people move their finger from a starting point to a stopping point along an indirect path, the length of their movement affects their estimate of the straight-line distance between the two points. The longer the indirect path from start to end, the longer the direct distance between these points is estimated to be. Another movement-based distortion in haptic space perception is the radial/tangential illusion. Movements towards and away from the body (radial movements) are estimated as being longer than equally long movements made from side to side (tangential movements). One cause of this distortion may be that movements in the two directions take different lengths of time to produce.

HAPTIC PERCEPTION OF TWO-DIMENSIONAL PATTERNS

Most people are familiar with the Braille symbols that can be found in printed media for people with visual impairment and often in elevators and other public facilities. These are examples of two-dimensional patterns that can be presented through touch. Tangible patterns can be relatively small in scale, fitting within a single fingertip, or larger, in which case the pattern must be explored by moving the fingers over it.

Fingertip-size patterns include raised stationary forms, such as Braille, or vibratory stimuli presented by pins that move up and down against the skin. The ability to identify vibratory patterns depends on several factors, including the skin site that is stimulated (the finger or the back, for example), the number of pins in the display, and – when two patterns are presented in succession – by the temporal and spatial relationship between them. Larger patterns, which are traced by the fingers, have often been used as graphic displays for people with poor sight. For example, a line drawing of an object can be converted to a raised, tangible form by a special duplicating machine. The ability to interpret raised drawings of common objects through touch is far below the level that can be achieved with vision. One reason is that the contours being felt must be stored in memory and integrated over time and space. Experience with raised pictures appears to contribute to success in identifying them, as does past visual experience. One strategy that is used to identify tangible patterns without sight is to convert them to a spatial image, which appears to be easier for people who were previously sighted.

PERCEPTION OF OBJECTS AND THEIR PROPERTIES

People are capable of naming familiar, common objects by touch alone with virtually no error, within a few seconds of contact. They do so by perceiving the object's properties – its shape, hardness, and so on. They interpret what they perceive in the context of their knowledge about objects and expectations about what is likely to occur in a given situation.

The nature of people's physical interaction with the world – how they explore and how objects contact them – is critical to what is perceived about an object. People use specialized patterns of touching, called 'exploratory procedures,' in order to determine particular object properties. For example, in order to perceive the roughness of an object's surface, people typically would rub the object. Haptic exploratory procedures have been identified for a variety of other object properties, such as weight and hardness (Figure 1). The exploratory procedure that people choose to use voluntarily, when trying to learn about some property, is also generally the most effective means because it enhances the relevant perceptual information. Rubbing an object, which people spontaneously do when trying to perceive roughness, has been found to heighten the responses of mechanoreceptors that are known to provide information about surface texture.

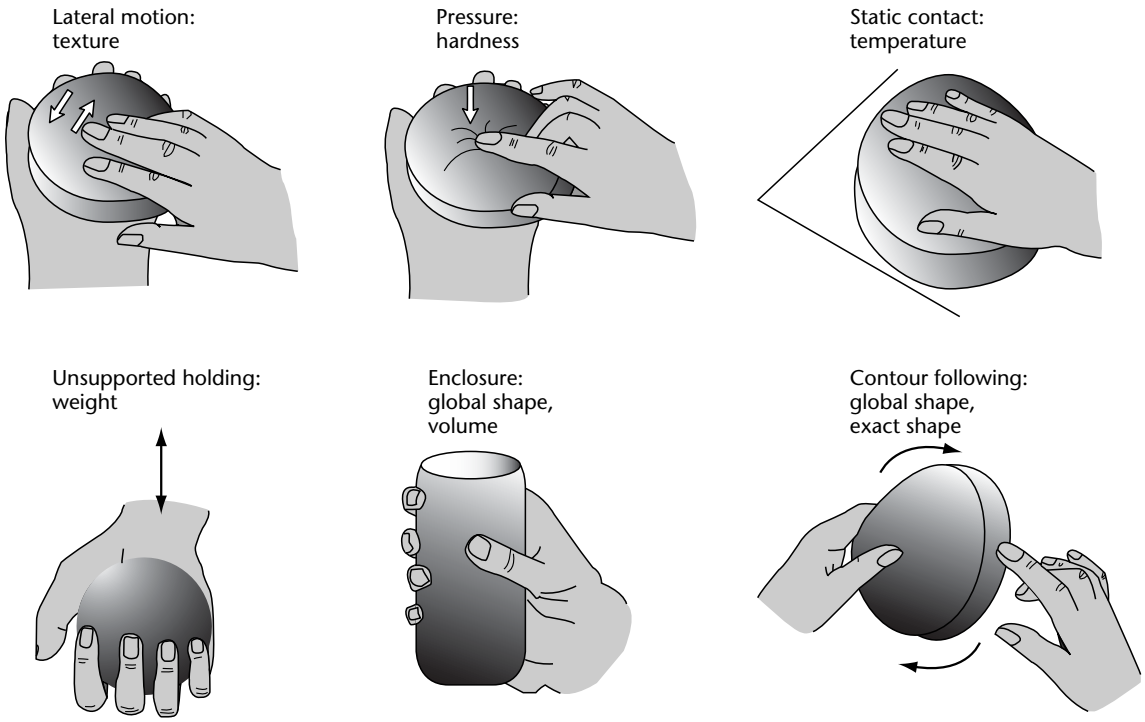


Figure 1. Exploratory procedures and associated object properties. From Lederman and Klatzky (1987); copyright Academic Press, with permission.

Not all properties of an object are equally available through touch. In general, vision and touch provide complementary information about the world, the visual system being more effective at providing information about an object's geometry (size and shape), and the haptic system being more effective at providing information about an object's material properties (roughness, hardness and apparent temperature). The relative availability of object properties can be measured with a haptic search task, in which participants indicate whether a target feature can be found in a display presented to the fingertips. The display incorporates not only the target, if present, but a number of distractor features. For example, the target might be a rough surface presented among smooth surfaces. The time to decide if the target is present is a measure of how quickly the property can be perceived. Material properties tend to be detected relatively quickly, and the time it takes is independent of the number of fingers being stimulated. Such properties are said to 'pop out' of the display. In contrast, haptic searches that involve discrimination between spatial arrangements on the fingertip, such as searching for a horizontal bar among vertical bars, tend to show a pattern of slower response time which increases with the number of fingers stimulated. This finding indicates that processing resources must be directed to each finger.

HAPTIC MEMORY

Two types of memory task have been studied extensively by cognitive psychologists. Explicit memory tasks require people to say directly what they remember about some past event. Implicit memory tasks look for a change in performance in some task, due to the individual having experienced the past event. Research on memory for words and pictures has shown that some manipulations that profoundly affect explicit memory tasks do not have an effect on implicit tasks. For example, thinking about the meaning of a visually presented word while studying it enhances later performance in an explicit test in which studied words must be recalled. However, if the test is implicit – for example, participants must identify words that are flashed briefly – words that were previously viewed will be identified better than unstudied words, regardless of the manner of study.

Comparable distinctions between implicit and explicit memory have been demonstrated in haptic memory experiments. In one experiment,

participants were asked to study a series of tangible forms by touch and then were given either a recognition test (explicit), in which they had to discriminate previously studied forms from novel items, or a test in which they were asked to feel and then draw a series of forms as accurately as possible. The recognition test was affected by whether the participant studied the form in a meaningful way (thinking up some use for an object with that shape) or a nonmeaningful way (counting the number of horizontal and vertical lines in the form). However, the drawing test, which measured implicit memory, was not affected by the task used at study. Evidence of implicit memory has also been found cross-modally – that is, when the participant studied an object by touch and the test object was visual, or vice versa. This indicates that there is a memory trace that can change performance implicitly and is accessible by both modalities.

APPLICATIONS OF RESEARCH

Research into haptic perception leads to several areas of application. Haptically distinguishable cues can be incorporated into the design of handles, knobs or dials, helping to make instruments easier to operate by touch. Haptic stimulators have been used in devices for deaf people, with the goal of helping them disambiguate lip movements. More recently, devices that send forces and vibrations to the hand have been connected to computers. These devices can be used to create haptic virtual worlds, where the forces simulate objects and events that do not physically exist. Similar force-feedback devices can be used to convey the results of real but remote manipulation, such as would occur during telesurgery.

Further Reading

Clark FJ and Horch KW (1986) Kinesthesia. In: Boff KR, Kaufman L and Thomas JP (eds) *Handbook of Perception and Human Performance,* vol. 1 *Sensory Processes and Perception*, pp. 13/1–13/62. New York: John Wiley.

Craig JC and Rollman GB (1999) Somesthesis. *Annual Review of Psychology* **50**: 305–331.

Heller MA and Schiff W (eds) (1991) *The Psychology of Touch*, pp. 91–114. Mahwah, NJ: Erlbaum.

Klatzky RL and Lederman SJ (1999) The haptic glance: a route to rapid object identification and manipulation. In: Gopher D and Koriat A (eds) *Attention and Performance XVII. Cognitive Regulation of Performance: Interaction of Theory and Application*, pp. 165–196. Mahwah, NJ: Erlbaum.

Klatzky RL and Lederman SJ (2002) Touch. In: Healy AF and Proctor RW (eds) *Experimental Psychology,* vol. 4*, Handbook of Psychology*. New York: John Wiley.

Lederman SJ and Klatzky RL (1987) Hand movements: a window into haptic object recognition. *Cognitive Psychology* **19**: 342–368.

Loomis JM and Lederman SJ (1986) Tactual perception. In: Boff KR, Kaufman L and Thomas JP (eds) *Handbook of Perception and Human Performances,* vol. 2*, Cognitive Processes and Performance*, pp. 31/1–31/41. New York: John Wiley.

Schiff W and Foulk E (eds) (1982) *Tactual Perception: A Sourcebook*. Cambridge, UK: Cambridge University Press.

Sherrick CE and Cholewiak RW (1986) Cutaneous sensitivity. In: Boff K, Kaufman L and Thomas J (eds) *Handbook of Perception and Human Performance*, pp. 1–70. New York: John Wiley.

Srinivas K, Greene AJ and Easton RD (1997) Implicit and explicit memory for haptically experienced two-dimensional patterns. *Psychological Science* **8**(3): 243–246.

Turvey MT (1996) Dynamic touch. *American Psychologist* **51**(11): 1134–1152.

Perceptual Learning

Introductory article

Manfred Fahle, University of Bremen, Bremen, Germany

CONTENTS

Perceptual learning leads to relatively permanent and often very specific improvements in solving perceptual tasks as a result of preceding experience and training. It seems to involve even rather peripheral parts of the central nervous system, such as the primary sensory cortices.

INTRODUCTION

Training and learning improve the performance of humans in a large number of cognitive, motor, and perceptual tasks, as has been known for centuries. Here, we will deal with improvements of *perception* caused by learning. Perceptual learning is defined as a rather permanent and specific change in the perception of and reaction to external stimuli based on preceding training or experience with these stimuli. One factor discriminating perceptual learning from other types of learning is its high specificity, discovered only in the early 1990s. A second difference is that improvement in perceptual learning is generally not based on conscious insight that could be communicated to others, but seems to happen subconsciously, in contrast to explicit forms of learning. Perceptual learning may be regarded as a new subtype of implicit learning and seems to involve a number of different levels of the cortical hierarchy, from higher, cognitive levels down to primary sensory areas, with strong involvement of attention and other top-down influences. The involvement of the low levels of cortical information processing sets perceptual learning apart from most other forms of nonmotor learning and has revolutionized our views regarding the plasticity of these early levels in adult humans.

DEFINITIONS: LEARNING, ADAPTATION, PLASTICITY, AND RELATED TERMS

Learning, of course, may be used as a general notion for all lasting changes in behavior or sensation based on prior experience and/or insight, based on the acquisition and storage of information. Perceptual learning is a specific case, characterized by being restricted to sensory information and by proceeding usually without subjective insight into what exactly is changing during the process, as is typical for implicit forms of learning. This distinguishes perceptual learning from insightful, cognition-based changes in the understanding of facts or relations and from explicit forms of learning. 'Plasticity' stands for changes on both the functional and the anatomic level of the central nervous system which lead to a better adjustment to the outer world or to compensation of defects. The term 'adaptation' denotes a short-term adjustment of the working range of a sensory system, often without any long-term consequences, which is its main difference from perceptual learning. An example is the adaptation to the prevailing light intensity. Two special cases of adaptation are 'saturation' and 'habituation': they denote transient decreases of sensitivity as a result of (usually repeated and/or strong) stimulation. In a way, after-effects are a third special case of adaptation. They are either of retinal origin (after-images) or else cortical origin, mostly due to neuronal fatigue and often caused by antagonistic neuronal mechanisms. Other forms of adaptation are 'sensitization', an increased response after a recent strong stimulation, and 'priming', a latent facilitation of a stimulus–response pair. 'Maturation' and 'development'

relate to changes in function and nervous system structure that – unlike learning – are based mainly on genetics and not on the individual's interaction with the environment.

EXPERT PERCEPTUAL SKILLS

Perceptual learning is not just a phenomenon occurring under restricted laboratory conditions – though it is most often studied this way to better isolate the influence of specific parameters. In everyday life, too, training will improve performance in a multitude of perceptual skills, ranging from visual over acoustical to gustatory (taste) and olfactory (smell) abilities. Expert perceptual skills reach astonishing levels and give testimony of what subtlety of discrimination can be obtained through training, for example in discriminating between simple (such as vernier targets – see below) and complex visual objects (such as CT-scans of normal versus pathological brains), attributing a few bars of music to the corresponding composer, or determining the chateau and year of origin of a wine. All these feats result from long training periods and would be impossible for anyone untrained.

VERNIER ACUITY AND OTHER HYPERACUITIES

To monitor the progress of perceptual learning, it is advisable to use a perceptual function that is easily quantified, gives reliable results, and is very sensitive to change of performance level. These requirements favor perceptual tasks that are complex enough not to be ultimately limited by the optics of the eye or the retinal receptor mosaic, but that are genuine tasks of the cortical machinery – yet mostly achieved by relatively peripheral parts of this cortical machinery, in order to be as independent as possible from higher cognitive processes. Even untrained observers yield amazingly low discrimination thresholds for a number of visual spatial tasks that are known under the heading of hyperacuities. Examples include stereoscopic depth perception, curvature detection, bisection (Figure 1(e)), and vernier discrimination (Figure 1(b)). For example, in vernier discrimination, where subjects have to discriminate between offsets to the right versus offsets to the left in lines with tiny breaks, even thresholds of untrained observers are usually around 20 arcsec, that is below the diameter and spacing of foveal photoreceptors (i.e. the light-receiving elements at the center of the retina). Trained observers may achieve thresholds down to about 2 arcsec, hence vernier discrimination is quite finely tuned but nevertheless open to considerable improvement through training. Hence it is a sensitive measure of perceptual learning, and therefore a large body of experiments has investigated perceptual learning using these hyperacuity tasks.

SPECIFICITY OF LEARNING

It has been known since 1980 that improvement in many of these tasks is rather specific for basic stimulus attributes such as stimulus orientation or position in the visual field. Fiorentini and Beradi, in 1980, found that the improvement through training in discriminating between complex grating stimuli disappeared after rotation of the stimuli by 90 degrees (Figure 1(a)). The same was true for a vernier discrimination task (Figure 1(b) (c)), and even a two-dot stereoscopic depth discrimination task improved for horizontally separated dots through training with these stimuli, only to return to baseline performance when the dots were separated vertically rather than horizontally.

The improvement through training was, moreover, specific for the visual field position that was used during training, both for vernier discrimination (Figure 2(a)) and for texture segregation. Training of a vernier discrimination task at an eccentric field position improved performance significantly within less than 1 hour of training, but if the stimulus was transferred to another position of equal eccentricity, improvement of performance disappeared.

Improvement obtained through training was even (at least partly) specific for the eye used during training: in a texture discrimination task where the subjects had to detect a figure on a background based on the orientation of stimulus elements (Figure 2) no transfer of improvement occurred after monocular training to the eye covered during training. The same specificity for the trained eye holds true for vernier discrimination tasks.

The possible artefact in some of these experiments is that observers may undergo motor learning: they might improve their fixation stability or/and the quality of accommodation. To test this possibility, observers were sequentially tested with a three-dot vernier (Figure 1(d)) and a three-dot bisection task (Figure 1(e)). In both of these tasks, the position of the middle one of three dots differs by less than one photoreceptor spacing from its position in the other task. Observers improved for both of these tasks, but there was no transfer of

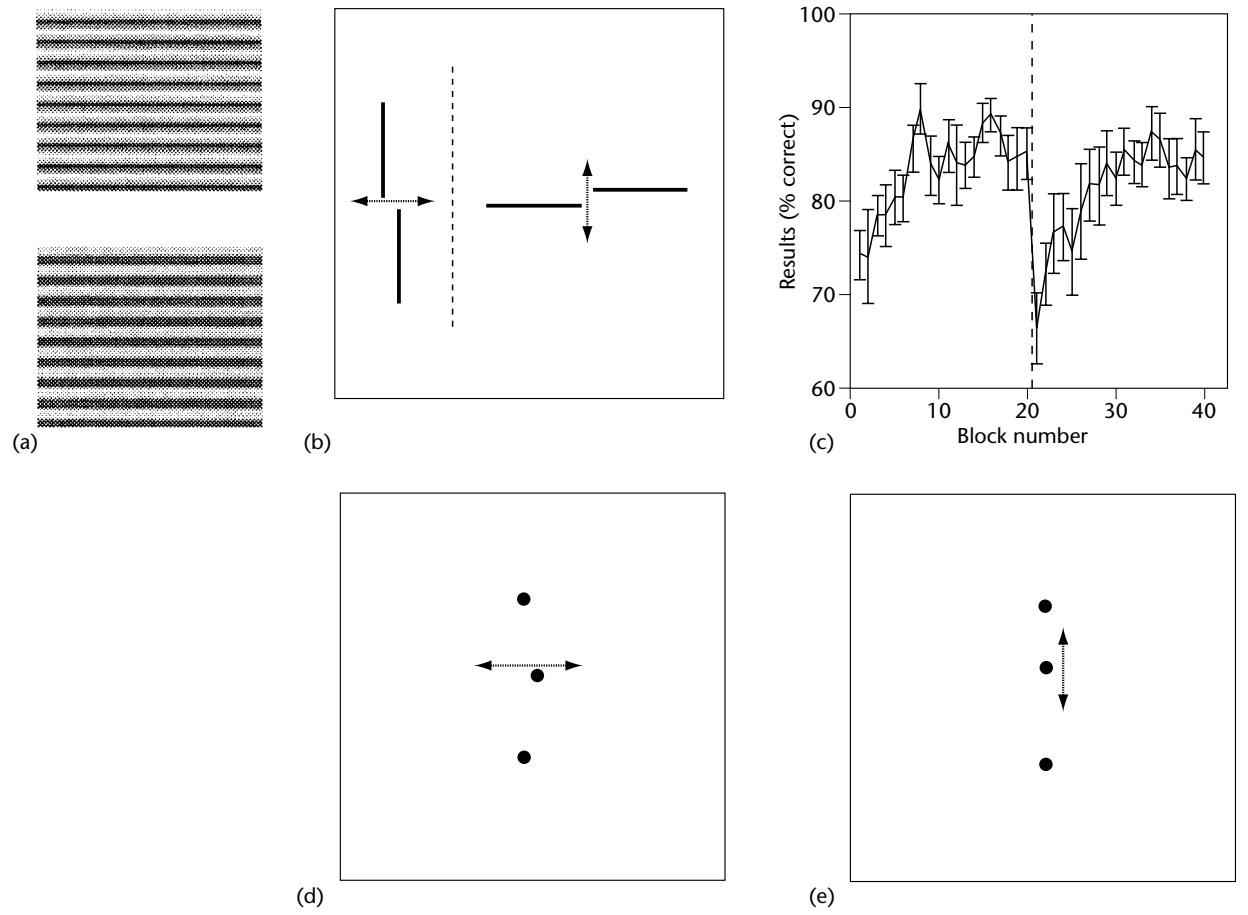


Figure 1. (a) Complex grating patterns as used by Fiorentini and Beradi (1980) to investigate orientation specificity of perceptual learning. Observers learned to discriminate between the two gratings displayed. (b) A vernier discrimination task: is the lower (or left-hand) segment offset to the right or to the left (or up or down) relative to the upper (or right-hand) one? (c) Improvement through training in a vernier discrimination task, within one hour of training, disappears after stimulus rotation by 90 degrees (after Poggio *et al*., 1992). (d) Three-dot vernier task: is the middle dot offset to the left or to the right relative to an imaginary line through the endpoints? (e) Three-dot bisection task: is the middle dot closer to the upper or the lower dot?

improvement, indicating that perceptual learning is not (mainly) based on motor learning, since any form of better fixation or accommodation should have transferred between these tasks. Hence, the specificity for both orientation and the trained eye indicates that the neuronal changes underlying perceptual learning may occur on the level of the primary visual cortex, since only there, neurons are specific both for the eye and for stimulus orientation, while they are not orientation-specific for all more peripheral levels and generally not eye-specific for all more central levels. Specificity of improvement is strongly diminished or disappears altogether for relatively easy tasks. A plausible explanation is that for easy tasks, there is no need to modify peripheral levels of sensory information processing; rather, it suffices to make better use of the information supplied by these 'early' levels.

ROLE OF ATTENTION

Given the above statement that perceptual learning is based most likely on changes on rather peripheral cortical levels of information processing such as the primary sensory cortex, the influence of attention not only on performance but also on speed of learning may be surprising at first sight, but it is a fact. Several independent investigations found that mere presentation of a stimulus that is not attended to improves long-term performance only marginally, or not at all. Hence, there must be strong top-down influences from more 'upstream' parts of the cortical hierarchy that select which

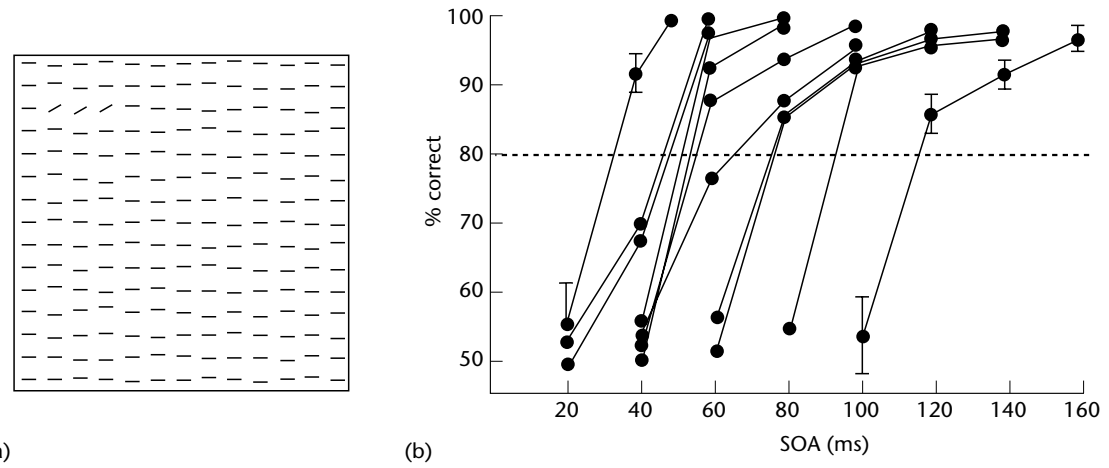


Figure 2. Improvement through perceptual learning in a figure–ground discrimination task based on orientation differences (after Karni and Sagi, 1991). (a) Observers had to discriminate the target defined by the three elements of deviating orientation. (b) Initially, stimulus onset intervals (SOAs) between the stimulus (a) and a subsequent mask had to be longer than 100 ms for the observer to discriminate the target, while less than 40 ms was sufficient after training.

features of the stimulus are attended to, and only those features seem to be effectively learned. Another indication for top-down influences is the fact that after a few clearly above-threshold presentations of the critical feature (hence after stimuli that are easily categorized), observers seem to gain some form of knowledge about which features of the stimulus to attend to and considerably improve performance within a short time.

ROLE OF FEEDBACK

Stimuli usually contain several features, and the number of features increases with the complexity of the stimulus displayed. How does the visual system know which features to attend to? As we just saw, top-down influences such as 'attention' probably play a major role here. Another top-down influence is error feedback. While perceptual learning is certainly possible even without error feedback, it seems to be faster when error feedback is provided. This feedback can be rather sparse, for example, taking the form of a percentage score of correct responses after each block of presentations. Here, the feedback cannot serve as a teacher signal – as in most neuronal network models – since it does not supply information about individual trials, and hence the observer cannot reliably classify stimuli as correctly or incorrectly answered on the basis of this block feedback. A second argument against error feedback serving as a teacher signal is the finding that training is still effective even if half of the error signals are omitted, i.e. feedback is only partial; if the feedback system served to categorize individual stimuli, half of the stimuli misclassified initially by the observer would be 'ratified' by the (missing) error signal – causing large problems for simple neuronal networks. Manipulated error feedback, for example random feedback signals completely unrelated to the observer's performance, can effectively block improvement, and the observer's performance fluctuates strongly but without any clear net improvement.

NEURONAL MECHANISMS AND MODELS OF PERCEPTUAL LEARNING

In the late twentieth century, the primary visual cortex was considered to be plastic only during childhood when it was optimized to become a perfect first filtering stage for visual information processing. In adults, however, the primary visual cortex should be 'hard-wired' in order to prevent changes made in response to one perceptual task from decreasing efficiency of this first filtering stage for other types of stimuli. How then can the statement above be true, that perceptual learning is so specific that it has to (at least partly) rely on changes as early as the primary visual cortex? The answer has also been given above, at least implicitly: the changes on the level of the primary visual cortex might be realized by top-down influences. Thus, the adaptation could be switched on for the appropriate class of stimuli only and hence would not interfere with the processing of other types of

stimuli. Following this hypothesis, plasticity of the primary visual cortex in adults would not interfere with formerly learned tasks.

Apart from the psychological evidence, such as stimulus specificity, there is also electrophysiological evidence for a change of the primary visual cortex following training. On the level of sum-potentials in humans, field distributions over the occipital skull change after training in a vernier discrimination task, especially for the evoked potentials with very short latencies. And on the single cell level, animal studies demonstrate pronounced changes of receptive field positions of neurons in cortical area 17 after lesions of the retinal input.

In summary, the experiments on perceptual learning demonstrate that the primary visual cortex, even of adult humans and animals, is still capable of plastic changes, that these changes underlying perceptual learning seem to be controlled mainly via top-down influences, and that perceptual learning involves several levels of cortical information processing. The results thus indicate that information processing in the visual system is not realized by strictly feedforward connections but that feedback loops play an important role for optimal perceptual performance.

Further Reading

Ahissar M and Hochstein S (1997) Task difficulty and learning specificity: reverse hierarchies in sensory processing and perceptual learning. *Nature* **387**: 401–406.

Fahle M, Edelman S and Poggio T (1995) Fast perceptual learning in hyperacuity. *Vision Research* **35**: 3003–3013.

Fahle M and Morgan M (1996) No transfer of perceptual learning between similar stimuli in the same retinal position. *Current Biology* **6**: 292–297.

Fahle M and Poggio T (eds) (2001) *Perceptual Learning*. Cambridge, MA: MIT Press.

Fahle M and Skrandies W (1994) An electrophysiological correlate of learning in motion perception. *German Journal of Ophthalmology* **3**: 427–432.

Fiorentini A and Beradi N (1980) Perceptual learning specific for orientation and spatial frequency. *Nature* **287**: 43–44.

Gilbert C and Wiesel T (1992) Receptive field dynamics in adult primary visual cortex. *Nature* **356**: 150–152.

Herzog MH and Fahle M (1997) The role of feedback in learning a vernier discrimination task. *Vision Research* **37**: 2133–2141.

Karni A and Sagi D (1991) Where practice makes perfect in texture discrimination: evidence for primary cortex plasticity. *Proceeding of the National Academy of Sciences USA* **88**: 4966–4970.

Poggio T, Fahle M and Edelman S (1992) Fast perceptual learning in visual hyperacuity. *Science* **256**: 1018–1021.

Performance

Introductory article

Dirk Kerzel, Justus-Liebig-University, Giessen, Germany
Wolfgang Prinz, Max Planck Institute for Psychological Research, Munich, Germany

CONTENTS

Human performance is a field of research focusing on the study of elementary sensorimotor skills in well-defined task environments that require fast and accurate responses. Among other things, processes related to the initiation and interruption of responses in single and dual tasks, such as motor planning, response selection, response execution, and response inhibition are investigated.

INTRODUCTION

Human performance is a field of research in which elementary sensorimotor skills are studied. These building blocks of human behavior involve sequences of an external stimulus and an observable action in response to the stimulus, with a well-defined relation between stimulus and action. That is, people are instructed to emit a specified response once a certain stimulus appears. Additionally, they are put under pressure to respond rapidly and without errors. Typically, the experimenter measures both the speed and the accuracy of responses in different conditions. As these aspects of human behavior are relevant to industrial productivity and safety, research on human performance was initially closely related to applied questions such as the design of dashboards or workplaces.

STARTING AND STOPPING ACTIONS

Indicators of Response Preparation

There are many indicators that bear witness to cognitive processes related to a response before that response is visible to an observer. If the electrical activity of the brain is measured by electrodes attached to the scalp, it can be observed that the brain potential on the side of the head opposite to where a response is about to be executed becomes more negative than the potential on the same side. For instance, if a person is instructed to press a left key, the electroencephalogram shows that the brain potential on the right side is more negative than that on the left. This phenomenon is called 'lateralized readiness potential' (LRP) because it has been shown to be involved in response preparation. Notably, the LRP may occur even slightly before the person is consciously aware of intending to execute the action. Thus, physiological events in the brain (at least occasionally) precede our awareness of action plans. At a peripheral level, the electromyographic activity at the muscle rises about 50 ms before the muscle actually moves. Finally, the observable response is emitted. The time between stimulus onset and onset of the associated response is the reaction time (RT). It is common practice to draw conclusions about motor preparation or planning from the analysis of psychophysiological indicators and RT. Any one of these indicators is incomplete as it never captures all aspects of motor planning; however, as a sum, they may be used to derive models of cognitive motor planning. By far the most widely used of these indicators of human performance is the RT. Different types of reaction time may be distinguished. If there is only one possible stimulus and only a single response associated with it, the observed latencies are simple RTs. If there is more than one stimulus, and several associated responses, choice RTs are measured. Typically, choice RTs are longer than simple RTs.

Simple Reaction Time

Simple RT is usually of the order of 200 ms, but the exact time depends on a number of factors. On the stimulus side, it has been shown that the intensity of the signal affects simple RT. The higher the intensity of the stimulus, the faster the response. For instance, bright targets yield faster RTs than dim targets, and loud sounds yield faster RTs than soft ones. Also, stable characteristics of the person affect

simple RTs. Simple responses of young and alert persons are faster than those of older and less alert persons. The degree of preparation is a further determinant of simple RT. Generally, increasing the level of preparation by presenting a warning signal speeds up simple RTs. However, the ability to maintain a high degree of preparation is limited. When the time between the warning signal and the stimulus is long, the beneficial effects of the warning signal disappear. Further, simple RT increases with uncertainty about the time when the stimulus will appear. Responses are faster if the person is sure about when the signal will occur than when the time of signal presentation is uncertain. Finally, simple RT depends on the effector that is executing the response. The larger the mass of the effector, the longer the response latency. For instance, lifting the finger is faster than flexing the elbow, and flexing the elbow is faster than flexing the shoulder. The reason for these differences may be electromechanical factors related to the size of the effectors, not the programming of the response: the time between stimulus presentation and the first electromyographic activity at the peripheral muscles is the same for all effectors, such that the central processes related to motor programming do not differ. Only the time between the first electrical activity at the muscle and the muscle movement differs as a function of effector.

Choice Reaction Time

Choice RT differs from simple RT in that a response has to be selected from an array of alternative responses (two at the minimum). A large variety of responses is available. For instance, the experimenter may ask for key presses on a keyboard with four keys arranged in a square, or for joystick movements to the left, right, up and down, or for verbal responses such as 'yes' and 'no'. In choice tasks, the stimuli are mapped onto responses: that is, the stimulus determines which response has to be executed. Usually, the mapping of stimuli onto response is conveyed by verbal instruction. For instance, the digit 1 appearing on a computer screen may require a joystick movement to the left, and the digit 2 may require a joystick movement to the right. Choice reaction time is longer than simple reaction time because the response in a choice task cannot be as well prepared as the response in a simple task.

The number of response alternatives in choice tasks has been shown to affect directly choice reaction time. Choice reaction time increases with the number of response alternatives: when people are asked to press one of n keys laid horizontally out in front of them whenever one of n lights located directly above the keys is turned on, RTs increase with n. The relation between n and choice reaction time follows a nonlinear function that has become known as the Hick–Hyman law. It states that choice reaction time increases linearly with the logarithm to the base 2 of the number of response alternatives (n). With an increase of 1 in the logarithm, the response time increases by about 150 ms. (*See* **Action**)

Stimulus–Response Compatibility

It is commonly observed that the speed of a choice response depends on how responses are mapped onto stimuli. Generally, the more natural or intuitive the mapping, the faster the responses. For instance, responses are fast when a left stimulus has to be responded by a left response and a right stimulus by a right response, but slow when a left stimulus has to be responded by a right response and a right stimulus by a left response. The former mapping is referred to as a compatible mapping. Stimulus–response compatibility may lead to violations of the Hick–Hyman law. When the assignment of responses to stimuli is extremely compatible, for instance when the tactile stimulation of a finger has to be responded to by depressing the stimulated finger, no increase of choice RT with the number of alternatives is noted. (*See* **Action**)

Sequences of Responses

The previous sections were concerned with a single action as a response to a stimulus. In this section, sequences of actions that follow an external stimulus are discussed. *A priori*, two ways of handling motor programming of a sequence of movements may be distinguished. First, the programming may be done in advance, and the program is only executed once the response is initiated. In this case, a motor output buffer is needed that retains the sequence of motor commands. Assuming that motor commands have to be loaded into the buffer one after the other, the time to load the buffer with commands increases with the required number of commands. Second, movement programming may be restricted to one movement at a time, such that in a sequence of actions, programming and execution alternate. In this case, no effect of number of movements on the reaction time of the first response should occur. The empirical research

suggests that the truth lies in between these extreme views. In support of prior programming, it was observed that the reaction time to initiate a movement sequence increases with the number of movements that have to be executed. For instance, lifting a finger is faster than grasping a tennis ball, which in turn is faster than grasping the tennis ball twice. Note that the question is how fast these movement sequences are initiated – that is, how fast a home key is released before the sequence is executed – not how long the sequence lasts. A similar slowing with increasing number of movements has been observed with sequences of key presses, spoken syllables and written letters. In support of the ongoing programming view, the reaction time of individual movements in a sequence varies as a function of their serial position. At some serial positions, reprogramming of the upcoming movements is necessary, which slows down the response. Thus, sequences of movements cannot be fully pre-programmed, suggesting that the motor output buffer has a limited capacity.

Stopping an Action

Many situations require people to stop their ongoing actions. The successful operation of a stop process may be just as important for survival as the ability to initiate an action quickly. In stop-signal experiments, participants are given a primary task to perform and a stop signal is intermittently presented telling them not to respond on that trial. One may think of the processes related to the stop signal and those related to the primary task as independent sets of processes that are contestants in a race: if the primary task process finishes before the stop-signal process, the response is executed, but if the stop-signal process finishes before the primary task process, the response is inhibited. Studies using the stop-signal paradigm revealed that the stopping of actions does not differ substantially between individuals, strategies or tasks. The latency of inhibition is of the order of 200 ms. Also, the stop-signal paradigm allows for the evaluation of ballistic processes involved in action execution. Ballistic processes cannot be inhibited once they begin; rather, they must run to completion. Little evidence for ballistic processes was obtained. For instance, one might assume that during typing, whole words are programmed, and the typing of the word sequence is ballistic. This is hardly the case. When typists were given an auditory stop signal during typing, they were able to stop typing after a single keystroke, regardless of word boundary (the only notable exception being 'the').

ERRORS OF PERFORMANCE

Errors are usually not independent of the speed with which an action is performed. When errors and reaction time both follow the same pattern across conditions, it might be concluded that performance was affected by the experimental variation. For instance, in an easy condition, fast responses and few errors may be observed, whereas in a difficult condition, slow responses and many errors occur. In this case, one might attribute differences in performance to the difficulty of the conditions. However, if errors and reaction time do not follow the same pattern, strategic decisions of the participant, not the properties of the experimental conditions, might account for the observed performance. If people opt for a daring response strategy, their responses are fast, but they make many errors. Alternatively, they may choose a cautious response strategy, implying that their responses are slow and few errors occur. In other words, participants may trade off speed and accuracy.

At least two types of errors may be distinguished: choice errors and serial order errors. If people select the wrong response from an array of possible responses, this is called a choice error. Serial order errors occur when a correct response is selected, but it is emitted at the wrong position. The exchange of speech sounds such as 'The queer old dean' instead of 'The dear old queen' show that we sometimes confuse the serial order of initial consonants. Also, these errors tell us that we store motor programs for a series of actions. In sum, errors of performance are another measure to characterize human performance in addition to reaction times.

COMBINING BEHAVIORS

Coarticulation

In a stream of actions, it is often efficient not to wait for one action to end before another is initiated. Coarticulation refers to simultaneous motion of effectors that makes a series of movements more efficient. It denotes that an upcoming action is prepared while another action is being performed. Such a strategy reduces latency of actions, and it leads to the combination of behaviors that are adjacent in time. The term 'coarticulation' was coined with respect to speech production, but the phenomenon encompasses other types of movements as well. For instance, finger movements during typing show that the fingers are not at rest until the previous keystrokes have been finished; rather, the hands move to their target positions ahead of

time, allowing for fast typing rates. (*See* **Speech Production**)

Bimanual Rhythmical Movements

In the case of coarticulation, the simultaneous motion of two effectors is a by-product of efficient motor performance. However, it is also common to intentionally perform two movements at the same time. In particular, the production of rhythmical patterns involves simultaneous bimanual movements. However, research shows that our ability to perform two independent rhythmical movements with two effectors is very limited. A famous example may illustrate this point. The index fingers are flexed or extended resulting in a movement that resembles tapping (e.g. on a table top). Two basic relations between the movement of the two index fingers may be distinguished: left and right index fingers are flexed and extended simultaneously, which is referred to as in-phase pattern. In the anti-phase pattern, one index finger flexes as the other extends. When in-phase and anti-phase movement patterns are performed at a moderate speed, and the speed is subsequently increased, a striking difference emerges: whereas the in-phase pattern is stable across different movement speeds, the anti-phase pattern breaks down at fast speeds and a switch into the in-phase pattern occurs because it is easier in terms of the emerging perceptual or motor pattern. Thus, performance in tasks that involve continuous action of two effectors is severely limited.

DOING MORE THAN ONE THING AT ONCE

Dual Task Performance

In everyday life it is common to see people perform two different activities at the same time. Driving a car does not prevent people from having a conversation, and eating potato crisps does not keep them from watching television. Usually people do not experience any difficulty in concurrently carrying out two tasks, unless the tasks are intellectually challenging or physically impossible. Contrary to this belief, laboratory studies show that performing even simple tasks concurrently causes interference between the two tasks – that is, one of the tasks is performed less efficiently in terms of response time or proportion of errors.

The Psychological Refractory Period

In one important experimental design used to investigate dual task performance, observers are presented with two stimuli in succession, S1 and S2. The stimuli may be auditory, visual or even tactile, and they do not have to be presented within the same modality (e.g. an auditory S1 may be combined with a visual S2). The stimuli S1 and S2 are temporally separated by a variable stimulus onset asynchrony (SOA). Typically, the SOA varies between 0 ms and 1000 ms. Before the experiment starts, the participant is instructed to respond to S1 with a response R1, and to S2 with a response R2 (Figure 1). By far the most common responses are key presses because they can be easily recorded on a keyboard. Other responses include vocal responses, eye movements or foot movements. Again, R1 and R2 do not have to originate from the same effector (e.g. a manual R1 may be combined with a vocal R2). A large number of mappings between stimulus and response are possible

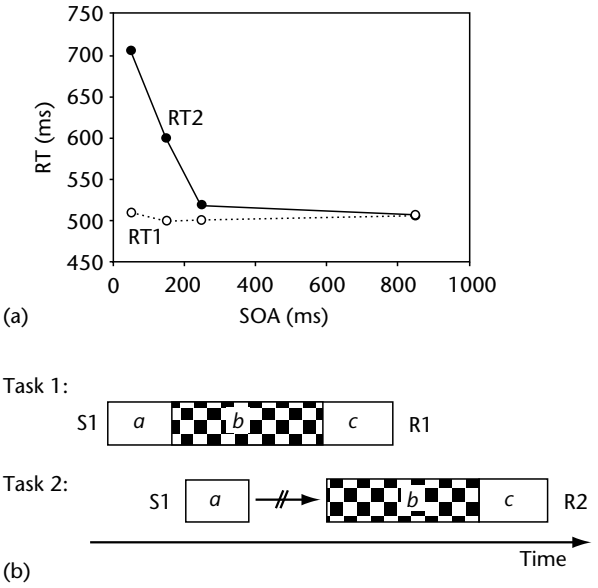


Figure 1. (a) The psychological refractory period effect. The reaction time for the first task (RT1) is unaffected by the stimulus onset asynchrony (SOA), whereas the reaction time in the second task (RT2) is longer at shorter SOAs. (b) The central bottleneck model of dual task performance. Cognitive processes *a*, *b* and *c* intervene between stimulus presentation S and response execution R, in two tasks. Process *b* (checkerboard region) is a bottleneck because it cannot begin in task 2 until the corresponding part of task 1 is complete. Process *b* has been associated with response selection.

and a fair percentage of them have already been tested. However, the main finding is robust across a large number of possible combinations of stimuli and responses: the time elapsing between the onset of S1 and the execution of R1 (RT1) varies little as a function of SOA between S1 and S2. In contrast, the time between onset of S2 and execution of R2 (RT2) increases dramatically when the SOA between S1 and S2 is made short. This finding is referred to as the psychological refractory period (PRP) effect. The PRP effect is the slowing of performance in one (or sometimes both) of two speeded tasks when the two tasks are performed at approximately the same time.

Other Dual Task Situations

Contrary to our everyday experience, the PRP effect shows that performance in two apparently easy tasks is slowed if they are performed at the same time. However, there are also dual task situations in which no interference between the two tasks occurs. When participants are engaged in a task without response uncertainty, such as repetitive finger-tapping or saying 'the' repetitively, there is almost no slowing of concurrent speeded responses in a different modality or of mental operations such as counting. Also, when there is no speed pressure on a perceptual judgment, it does not suffer from a speeded response in another modality. For instance, unspeeded visual identification of a letter is not affected by the approximately simultaneous response to an acoustic signal. Taken together, these findings suggest that the important distinction between dual task situations that show indications of interference and those that do not is whether both tasks or only one task requires a speeded response. Generally, whenever a rapid decision about a response has to be made in two approximately simultaneous tasks, interference results; otherwise, dual task interference is less likely and may be attributable to factors that are unrelated to the execution of the two tasks.

Models of Dual Task Performance

A widely accepted account of dual task performance, and in particular the PRP effect, is based on the assumption of a central bottleneck. A mental process *b* is considered a central bottleneck if (1) *b* is necessary for the completion of both tasks involved in a dual task situation, and (2) *b* may only be used by one task at a time. In other words, when a task (T1) lays claim on *b*, *b* cannot be used by another task (T2). Rather, T2 has to wait for *b* to be released from T1. Only after *b* has been released from T1 can T2 be completed. A central processing bottleneck would explain the PRP effect in the following way: when S1 is presented, the execution of the first task, T1, is initiated. Because S2 is presented after S1, execution of T2 is initiated after T1. Therefore, T1 claims *b* before T2 tries to. Consequently, with short intervals between S1 and S2, T1 may still occupy *b* when T2 is ready to use *b*. As a result, responses to S2 are delayed. With long intervals between S1 and S2, responses to S2 are faster because T1 has already released *b*, such that T2 has free access to it.

Considerable debate has focused on the nature of the central bottleneck. During the execution of a task that requires a particular response when a particular stimulus is presented, different processing stages may be identified: perceptual identification of the stimulus, selection of the correct response, and execution of the response. The central bottleneck may be located at any of these stages. Over the years, evidence has accumulated that the central bottleneck is located at the stage of response selection – that is, at the stage at which we decide whether to press the left or the right key, or whether to say 'a' or 'b' in response to a given stimulus.

The hypothesis of a central bottleneck in response selection has been criticized, on the grounds that the instructions in experiments on the PRP effect gave performance in T1 priority over performance in T2. That is, participants were asked to emit R1 before R2, possibly forcing them to use a bottleneck mechanism to postpone information processing for T2, so that responses to S2 did not occur before those to S1. Alternatives to the central bottleneck approach assume that the scheduling of processes in the PRP paradigm is under strategic control. That is, participants voluntarily decide to complete stimulus identification and response selection for T1 before tackling T2 because of the higher priority of T1, and not because of a structural bottleneck that forces them to do so. If participants had been given different instructions, they might just as well have scheduled response selection for T1 and T2 to occur at the same time (in violation of the second condition for a central bottleneck).

Further Reading

Gazzaniga MS, Ivry RB and Mangun GR (1998) Motor control. In: Gazzaniga MS, Ivry RB and Mangun GR (eds) *Cognitive Neuroscience*, pp. 371–422. New York: Norton.

Hommel B and Prinz W (1997) *Theoretical Issues in Stimulus-response Compatibility*. Amsterdam: North-Holland.

Jeannerod M (1997) *The Cognitive Neuroscience of Action*. Cambridge, UK: Blackwell.

Pashler H (1994) Dual-task interference in simple tasks: data and theory. *Psychological Bulletin* **116**: 220–244.

Rosenbaum DA (1991) *Human Motor Control*. San Diego: Academic Press.

Sanders AF (1998) *Elements of Human Performance: Reaction Processes and Attention in Human Skill*. Mahwah, NJ: Lawrence Erlbaum.

Schmidt RA and Lee TD (1999) *Motor Control and Learning*. Champaign, IL: Human Kinetics.

Wickens CD (1992) *Engineering Psychology and Human Performance*, 2nd edn. New York: HarperCollins.

Phenomenology, Psychological

Advanced article

Michael Kubovy, University of Virginia, Charlottesville, Virginia, USA

CONTENTS

Phenomenology is the study of how the world (material, mental, or cultural) appears to me. Given the common view of scientific observation as objective, unaffected by a point of view, it is natural that controversy exists over the role of phenomenology in science.

INTRODUCTION

The debate over the role of phenomenology in science (Roy *et al*., 1999) can be understood with the help of a diagram (Figure 1) that shows the three facets of cognitive theory – a theory of the brain, a theory of how the mind performs its computations, and a theory of experience – and the links between them. The controversy has focused on the link between the computational and the phenomenological mind. Skeptics have claimed that there is an unbridgeable 'explanatory gap' (a term coined by Levine (1983)) in this link. An early exponent of this position was Thomas Nagel (1970), who eloquently raised doubts about the possibility of ever incorporating phenomenological data into cognitive science.

Those who believe that the explanatory gap is unbridgeable give two reasons. The first reason is that theories of the phenomenological mind are inevitably based on private data: I cannot tap into the experience of another person. In contrast, theories of the computational mind are, generally speaking, based on public data that we obtain from psychological experiments. The two can therefore never meet. But the argument from privacy alone is not sufficient to warrant the skeptical view of the explanatory gap. If you had reasons to believe that another person's experiences were the same as yours, you would not care that they are inaccessible. The argument is parallel to the following one. Because you have reasons to believe that the text of a book remains the same when the book is closed, you do not care that it is inaccessible when it is closed.

Thus, to argue for the explanatory gap, the skeptic requires a second argument: the argument from subjectivity. It claims that different sentient beings must have different experiences. So even if you could tap into a bat's experience of the world, the experience would be unfathomable, since bats, for example, catch insects so differently from the way you do. But could we sidestep skepticism by capitulating on the bat front only, while denying that the argument from subjectivity holds among humans? The skeptic would reply that humans too experience things differently from each other, because they are of different sexes, because they grew up in different cultures, or simply because they are different people.

We will not try to resolve this issue on philosophical grounds. Instead, we present a phenomenological analysis to illustrate how a cognitive psychologist might deal with these issues. In working through this example we will follow Husserl (see Moran, 2000, chapters 2–5), as interpreted by Sartre (1948) and Ihde (1977)).

MENTAL IMAGERY

In the 1970s and 1980s, following the groundbreaking work of Shepard and Metzler (1971) (see also Shepard and Cooper, 1982), there raged a debate over the nature of mental imagery. We will follow Kosslyn's (1994) account. The two camps in this debate defended two different kinds of mental representations that could underlie mental imagery (Kosslyn, 1994, p. 5):

- *Propositional:* 'A propositional representation is a "mental sentence" that specifies unambiguously the meaning of an assertion. Such a representation must contain a relation … [which] ties together two or more … arguments.'
- *Depictive:* 'A depictive representation is a type of picture, which specifies the locations and values of configurations of points in a space. … [D]epictive representations convey meaning via their resemblance to an object, with parts of the representation corresponding to parts of the object.'

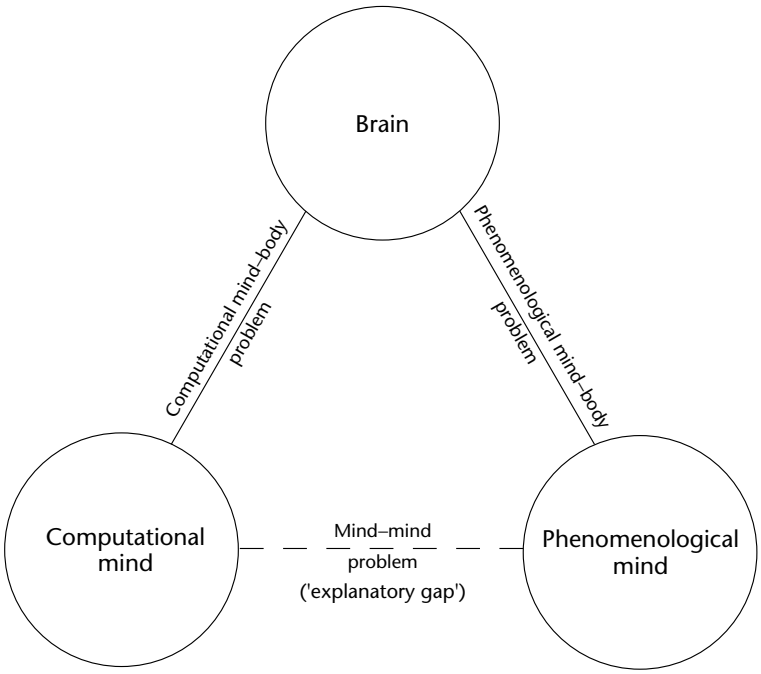


Figure 1. The three facets and the three linkage problems of cognitive science, and the explanatory gap. (Adapted from Jackendoff, 1987.)

At the end of his survey, Kosslyn concludes that mental images 'rely … on depictive representations'. He also effectively sidesteps the following challenge offered by Pylyshyn (1973, 1981). If a mental representation is depictive, it must be depictive for someone or something, perhaps the eye of a homunculus. But if we accepted that, we would then fall into an infinite regress, because we would have to explain how this homuncular eye works. Kosslyn replies that he is under no obligation to answer Pylyshyn's question about imagery. We know that the visual system contains depictive representations: the existence of multiple retinotopically organized maps has been well established (Wandell, 1999; Zeki, 1993; Livingstone and Hubel, 1988). Yet no one asks who or what looks at these retinotopic maps. Kosslyn argues that imagery relies on depictive representations because, firstly, 'imagery and perception share common mechanisms', and secondly, the mechanisms underlying perception are depictive. He concludes that Pylyshyn's homunculus argument about imagery is no more justified than such an argument about visual perception.

Now for a little exercise in imagery. Think of the home you live in. How many windows can one see on the facade of the house? This is a question that most people answer willingly. (Many people make small pointing gestures while they count.) Once you have counted, ask yourself: where were you standing when you counted? Most people answer with no hesitation.

Before we examine our exercise, let us mention a fundamental proposition of phenomenology: all experience is the experiencing of something (Moran, 2000, pp. 16–17). This 'aboutness principle' suggests that the structure of an experience is '*experiencing* → *experienced*', where the → represents the 'aboutness' relation.

If you think back on what you just did, you will recognize that during the exercise you were conscious of the facade, not of an image of the facade. In this respect, imagining is no different from perceiving. How, then, do they differ? As Sartre (1948) observed, when you perceive the facade, you experience it as present, but when you imagine it, you experience it as absent. Just as Jastrow's (1900) drawing (Figure 2) remains unchanged whether you see it as a duck or as a rabbit, the facade is the same, whether you see it, and therefore experience it as existing, or imagine it, and therefore experience it as absent.

What we have learned (following Ihde, 1977, chap. 2) can be summarized by two schemata: '*seeing* → *facade*' and '*imagining* → *facade*'. Thus both experiences are experiences of the same thing, the facade.

Our observation regarding the difference between seeing and imagining the facade can be

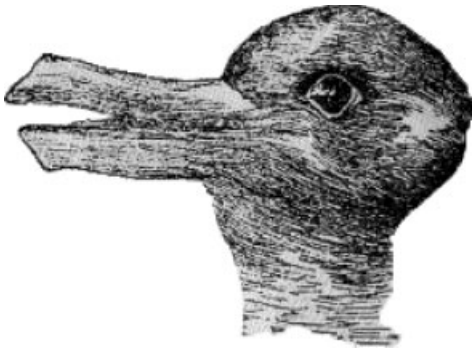


Figure 2. The duck-rabbit. (Adapted from Jastrow, 1900.)

represented by the contrast between '*seeing: experiencing as present → facade*' and '*imagining: experiencing as absent → facade*'.

Finally, having observed that you knew where you stood while counting the windows, we infer that this act of imagining contains an experiencer, which we denote '(I)'. So we can write: '*(I) imagining: experiencing as absent → facade*'.

A theory of mental imagery must take account of these observations. But current theories do not. Regarding the fact that you knew where you stood when you counted windows, perhaps we need a theory of embodied imagination (perhaps in the direction proposed by Ballard *et al*. (1997)), according to which imagining a visible object is a partial reenactment of many of the bodily activities involved in perceiving something, not just an activation of the visual system.

INTERSUBJECTIVITY AND EXPERIMENTAL PHENOMENOLOGY

The issue of subjectivity would not come into play in our phenomenological analysis unless someone who did our exercise challenged our observations. For example, you could tell me that you actually did not know where you stood while you were counting the windows; that when I asked, you inferred your position from the fact that you could see the whole facade. Of course, such a challenge would also bring to the fore the problem of the privacy of experience: I could not challenge your assertion. It is precisely here that other methods of research would come into play. One could not resolve a disagreement over the source of people's phenomenological intuitions within the domain of phenomenology. It is only by assuming that the links among the three facets of cognitive science are intact that progress can be made. To exclude phenomenological methods just because

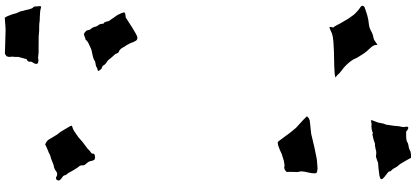


Figure 3. Grouping. (Adapted from Köhler, 1947.)

they do not bear a seal of infallibility is to ensure the existence of an explanatory gap.

Thus the key to the use of phenomenological observation is 'intersubjectivity' (Husserl, 1967), i.e. an agreement among individuals about the nature of their experience. The Gestalt psychologists – who were the ones who imported phenomenology into psychology – took it for granted that phenomenology was not a study of the subjective. For example, Köhler (1947, p. 142, emphasis mine) writes of Figure 3:

> The reader has before him two groups of patches. Why not merely six patches? Or two other groups? Or three groups of three members each? When looking casually at this pattern *everyone* beholds the two groups of three patches each.

Phenomenology has given rise to a research methodology, sometimes known as 'experimental phenomenology', which differs from conventional psychological experiments as summarized in Table 1 (Bozzi, 1989, chap. 7). This methodology is appropriate for the kind of phenomenological exploration we carried out above.

OPPOSITES

For an example of the application of Bozzi's method, we turn to the research of Savardi and Bianchi (2000) on psycholinguistics and the experience of space. They claim that opposites structure our experience, no less than unity, identity, or similarity.

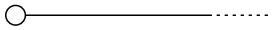
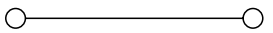
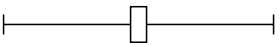
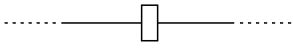
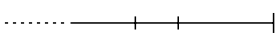
The research was conducted with a group of undergraduate students of industrial design at the Politecnico di Milano (the Milan Institute of Technology). They met in groups of three once a week in three-hour sessions to exchange observations about their experience of space. They were asked to find all the words (in everyday Italian) they needed to exhaustively describe any space or spatial relation. They produced 74 adjectives or adverbs. Upon examination of these words, Savardi and Bianchi found that each word had its opposite within the set (see Table 2).

Once they had obtained these opposites, Savardi and Bianchi assumed that these were the endpoints

Table 1. Comparison of conventional and phenomenological experiments. (Adapted from Bozzi, 1989.)

	Conventional	*Phenomenological*
Environment	isolated (e.g. a laboratory)	any (preferably not a laboratory)
Participants	kept naive about the topic or purpose of the research	told everything
Task	well-defined	jointly defined by participant and researcher
Participants' response	often the first that comes to mind	may transcend their first impression, and thus provide information about their solution space
	may not be modified	may be reconsidered
	either correct or incorrect	always valid
	unambiguous, or filtered into a set of mutually exclusive and collectively exhaustive *a priori* categories	classified only after all the data have been examined

Table 2. Thirty-seven pairs of opposites, grouped by cluster. A disk represents a singular pole. A solid line turning into a dashed line represent an unbounded polarity. A solid line that ends with a vertical tick represents a bounded polarity. A rectangle represents a single intermediate state. Short vertical ticks between the endpoints represent multiple intermediate states (Adapted from Savardi and Bianchi, 2000.)

Dimension	*Representation*	*Dimension*	*Representation*
Cluster 1		*Cluster 2*	
regular–irregular		moving–still	
symmetric–asymmetric		open–closed	
complete–incomplete			
straight–curved			
ordered–disordered			
supported–unsupported			
bounded–unbounded			
Cluster 3		*Cluster 4*	
full–empty		obtuse–acute	
standing–lying		uphill–downhill	
vertical–horizontal		convex–concave	
top–bottom		divergent–convergent	
beginning–end		right–left	
floating–sunken		rounded–angular	
upright–upside down			
inside–outside			
Cluster 5		*Cluster 6*	
above–below		far–near	
in front of–behind		long–short	
		wide–narrow	
		thick–thin	
		high–low	
		broad–narrow	
		many–few	
		large–small	
		dense–sparse	
		fat–thin	
		simple–complex	
		deep–shallow	

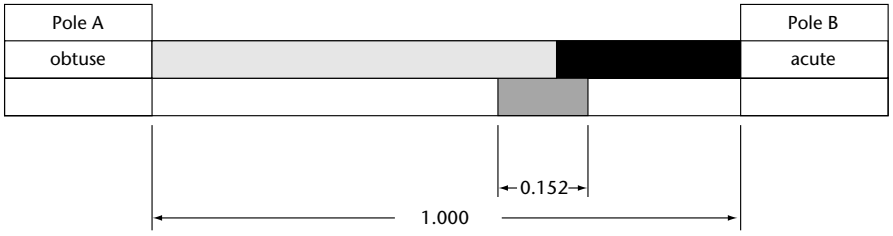


Figure 4. Fragment of the table used by Savardi and Bianchi (2000). It is shown as it might look after having been filled by participants.

of dimensions. To understand these dimensions, they explored the characteristics of things that are at neither pole of these dimensions. They gave the participants a table (Figure 4) that listed the 37 dimensions defined by their endpoints, or poles. About the 'obtuse–acute' dimension they were told to think of the space between the poles as a scale that represents 'all possible visual experiences of things that you would call "obtuse" or "acute" or anything in between'. For each pair, there were two scales. On the first scale, they were then told to mark the boundary between things they would call 'obtuse' and things they would call 'acute'. On the second scale they were asked to indicate the bounds of a range that would straddle the boundary marked on the first scale. This range was to represent the things they would neither call 'obtuse' nor 'acute'. The participants were reminded that the boundaries they were asked to place should not represent the number of things that are obtuse, acute, or in between, but rather the number of ways there are to be obtuse, acute, or in between.

Having obtained these judgments, they computed the proportion of the dimension covered by the 'in between' region (15.2% in Figure 4), and then computed what fraction of the remainder was given to each pole. They found that, on the average, only 12% of our experience of space is intermediate. They then performed a hierarchical cluster analysis of these judgments, and found that the dimensions fell into three classes: strongly polarized (19 dimensions), moderately polarized (13 dimensions), and weakly polarized (5 dimensions). For the strongly polarized dimensions, such as 'moving–still', each polar adjective either holds or not, there are no names for intermediate states, and the poles cover the entire range of possible states. For the moderately polarized dimensions, such as 'vertical–horizontal', there are names for intermediate states. Finally, for the weakly polarized dimensions, such as 'full–empty', each polar adjective either holds or not, there are no names for intermediate states, and the poles cover only the extreme states.

Savardi and Bianchi then performed a hierarchical cluster analysis of the asymmetry of the dimensions, and further classified the dimensions, as shown in Table 2. After several more steps, they produced a taxonomy of spatial concepts, also shown in this table.

Two remarks should be made here. Firstly, phenomenological research is similar to protocol analysis, which involves the use of verbal reports as data, and is often used in research on problem-solving (Simon and Kaplan, 1989), and to other non-experimental research methods (Smith *et al.*, 1995). Secondly, phenomenology should not be confused with introspectionism as practiced by early psychologists such as Titchener. In fact, any account of the introspectionists' methods (e.g. Lyons, 1986), and their methodological assumptions, will confirm that introspectionism is closer to contemporary experimental psychology than to phenomenology. Indeed, overall, introspective research is well described by Bozzi's list of features of conventional psychological experiments (Table 1).

The application of phenomenology to cognitive science is not restricted to the type of exercise we engaged in earlier and the type of work Bozzi described in his valuable discussion of 'Interobservation as a Method of Experimental Phenomenology', as summarized in Table 1 (Bozzi, 1989, chap. 7). We therefore turn to a discussion of a methodology that draws upon psychophysics and phenomenology, which Kubovy and Gepshtein (2002) called 'phenomenological psychophysics'.

PHENOMENOLOGICAL PSYCHOPHYSICS

Palmer (2002) in a discussion of demonstrations of grouping (such as Figure 3) writes that, they are 'a useful, but relatively blunt instrument for studying perceptual organization' because they have a

subjective basis. Pomerantz and Kubovy (1981, p. 426) had similar concerns:

> [T]he pragmatic streak in American psychology drives us to ask what role … experiences, however compelling their demonstration, play in the causal chain that ends in action. Thus we ask whether such phenomenology might not be a mere epiphenomenon, unrelated to behavior.

The standard solution to this problem is to use objective behavioral tasks:

> [I]f we can set up situations in which we ask subjects questions about the stimulus that have a correct answer, and if organizational processes affect their judgments (and so their answers), then the experimentalists' skepticism about the importance of organizational phenomena should be dispelled.

Can one determine whether an experience is epiphenomenal? For example, can perceptual organization be studied by embedding it in an experimental task for which responses can be judged to be correct or incorrect, i.e. a traditional psychophysical task? This involves, as Palmer says, 'changing what is actually being studied from subjective grouping to something else'. What does this transformation of one task into another entail, and what does it achieve?

As an illustration of traditional psychophysics applied to a problem in perceptual organization, consider the experiments in which Palmer and Bucher (1981) studied the pointing of equilateral triangles (Figure 5(d)). The equilateral triangle at the bottom of the left panel of Figure 5(d) appears to point about equally often at 60°, 180°, or 300°. Suppose you were shown this triangle, and asked to report which way it is pointing. The bottom section of the left panel of Figure 5(d) shows the triangle you see. Its background is white, indicating that it represents a public event. The middle section represents the three possible ways in which you could see the triangle pointing. Its background is gray, indicating that it represents a private event, your experience. The top section represents your response. Since you could see the triangle in one of three ways, you are asked to choose one of these alternatives, i.e. to make a three-alternative forced choice (3AFC). The bottom, middle, and top sections of the panel are connected by white arrows; they represent a presumed temporal and causal order. The structure of this task is summarized in Figure 5(b); it is the structure of a phenomenological psychophysics task.

Now suppose you were asked a different question: is the triangle pointing to the right or left? This is what Palmer and Bucher (1981) asked. In the left panel of Figure 5(c), the bottom section shows the triangle, and the left side of the middle section shows the ways you could see it. But since you were not asked which way it was pointing, the phenomenology is somewhat more involved than in the first example. If you spontaneously saw the triangle pointing horizontally, you could answer the question just by deciding whether it was pointing left or right, which requires some effort. But if you spontaneously saw it pointing in another direction, to answer the question you would have to relinquish that view of the triangle, so that you could see which way it would point were it pointing horizontally (right side of the middle section of the panel). This is not always easy to do. We can call this effort 'perceptual work'. This work is represented by the doubled-headed arrow in the middle section. Here you are asked to make a two-alternative forced choice (2AFC), and your reaction time (RT) is recorded. The structure of this task is summarized in Figure 5(a); it is the structure of a traditional psychophysics task.

As opposed to phenomenological psychophysical tasks, traditional psychophysical tasks are indirect. This idea is illustrated in Figures 5(a) and 5(b). In natural viewing conditions, as well as in the tasks used in phenomenological psychophysics, certain aspects of the visual scene ('stimulus' in the figure) lead to a corresponding percept by means of a private perceptual process. The latter is labeled as a 'spontaneous perceptual process' in the figure to emphasize that the process occurs naturally, just as it does when the observer views a stimulus outside the laboratory. The experimental phenomenologist strives to devise experimental conditions that ask observers to give reports that are as close as possible to the way they would describe their experiences outside the laboratory.

We now come to the crux of the Palmer and Bucher experiment, and to the resolution of the issue of epiphenomenality. If you align three equilateral triangles along a common axis of mirror symmetry tilted at 60° (Figures 5(c) and 5(d), right panels), they appear to point most often at 60°. Palmer and Bucher were interested in the effect of this context on the pointing of triangles. They used the traditional psychophysics task described in the preceding paragraph, and found that observers were slower to decide whether the axis-aligned triangles point to the right or to the left than to decide whether the isolated triangle does. The pointing induced by the common axis forced the observers to do more perceptual work than in the case of the single triangle. It is this perceptual work that persuades us that neither pointing nor

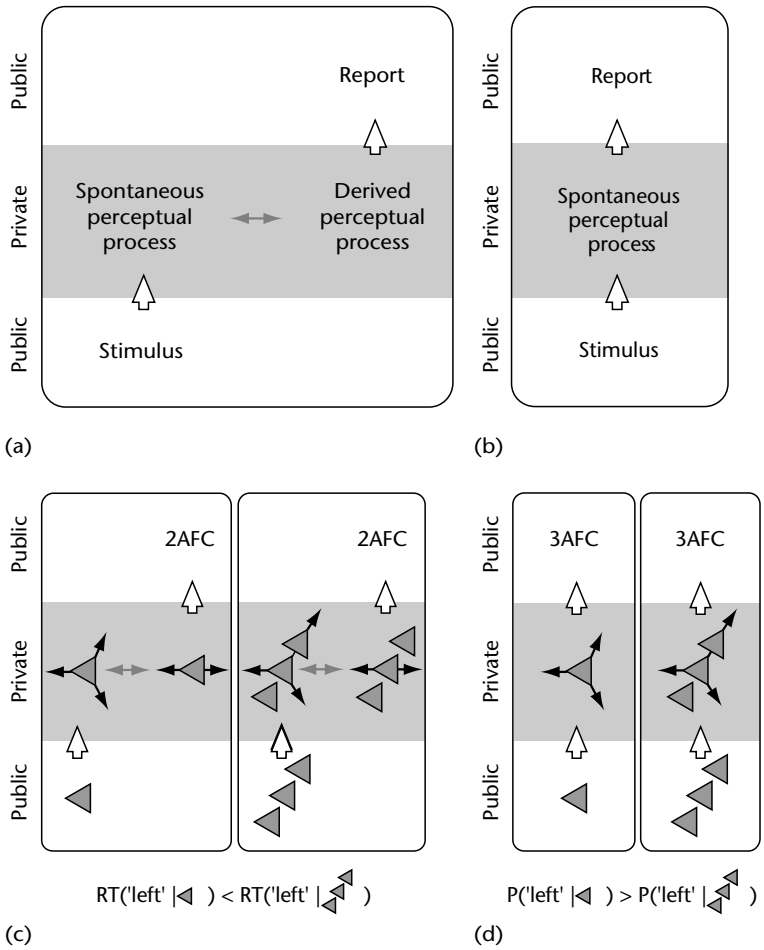


Figure 5. Comparison of the processes that take place in an observer engaged in different types of experimental procedures. The hypothetical events that are not public are marked by a gray band. The procedures of traditional psychophysics force the observer to do 'perceptual work' (horizontal arrows), i.e. to transform their experience in order to meet the requirements of the procedure. Thus they engage perceptual processes additional to the procedures of phenomenological psychophysics. In that sense, the latter are more direct than the former. (a) Traditional psychophysical procedure. (b) Phenomenological psychophysics procedure. (c) The experiment of Palmer and Bucher (1981) as a traditional psychophysical procedure (2AFC). (d) Hypothetical study of pointing using phenomenological report (3AFC).

the effect of a common axis is epiphenomenal (or purely subjective).

Epiphenomenality cannot be refuted with phenomenological methods. However, having established that the effect of a common axis on pointing is not epiphenomenal, one can explore the effect directly, i.e. phenomenologically, without forcing observers to do perceptual work (Figure 5(c), right). For example, one could use a phenomenological psychophysics procedure with a three-alternative forced choice (3AFC) in which the observer's task is to report (by pressing one of three keys) in which direction the middle (or single) triangle is pointing (Figure 5(c): '$P(X)$' stands for the probability of percept X). This is a phenomenological report because the three report categories offered to the observers agree with the three likely spontaneous organizations of the stimulus.

In traditional psychophysics the natural perceptual experience is transformed. It is transformed by asking observers to judge certain aspects of the stimulus, thus engaging mechanisms not normally involved in the perception of natural scenes. Or the perception of the stimulus may be hindered, either by adding external noise to the stimulus or by presenting the stimulus at the threshold of visibility. It is not clear whether such transformations of perceptual experience are indispensable in studies of perception.

The inferences involved in the interpretation of psychophysical studies of perception would not make sense without assuming the existence of a

spontaneous perception, which could have led to a phenomenological report. They are an indirect assessment of perceptual processes. We do not suggest that tasks that have correct and incorrect responses are useless after one has established that a perceptual phenomenon is not epiphenomenal. These tasks may give us important information about the underlying process. But indirect psychophysical methods have no intrinsic advantage over phenomenological methods.

Acknowledgement

Supported by NEI grant R01 EY 12926-06.

References

Ballard DH, Hayhoe MM, Pook PK and Rao RP (1997) Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* **20**: 723–767.

Bozzi P (1989) *Fenomenologia sperimentale*. Bologna, Italy: Il Mulino.

Husserl E (1967) *Cartesian Meditations* (translated by D. Cairns). The Hague: Nijhoff. [First published 1931.]

Ihde D (1977) *Experimental Phenomenology: An Introduction*. New York, NY: Putnam.

Jackendoff RS (1987) *Consciousness and the Computational Mind*. Cambridge, MA: Bradford Books/MIT Press.

Jastrow J (1900) *Fact and Fable in Psychology*. Boston, MA: Houghton Miffin.

Köhler W (1947) *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. New York, NY: Liveright. [Revised edition. First published 1921.]

Kosslyn SM (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: Bradford Books/MIT Press.

Kubovy M and Gepshtein S (2002) Grouping in space and in space-time: an exercise in phenomenological psychophysics. In: Behrmann M and Kimchi R (eds) *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Mahwah, NJ: Erlbaum.

Levine J (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–361.

Livingstone M and Hubel D (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240**: 740–749.

Lyons W (1986) The disappearance of introspection. Cambridge, MA: MIT Press.

Moran D (2000) *Introduction to Phenomenology*. London: Routledge.

Nagel T (1970) What is it like to be a bat? *Philosophical Review* **79**: 394–403.

Palmer SE (2002) Understanding perceptual organization and grouping. In: Behrmann M and Kimchi R (eds) *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Mahwah, NJ: Erlbaum.

Palmer SE and Bucher NM (1981) Configural effects in perceived pointing of ambiguous triangles. *Journal of Experimental Psychology: Human Perception and Performance* **7**: 88–114.

Pomerantz JR and Kubovy M (1981) Perceptual organization: an overview. In: Kubovy M and Pomerantz J (eds) *Perceptual Organization*, pp. 423–456. Hillsdale, NJ: Erlbaum.

Pylyshyn ZW (1973) What the mind's eye tells the mind's brain. *Psychological Bulletin* **80**: 1–24.

Pylyshyn ZW (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **87**: 16–45.

Roy J-M, Petitot J, Pachoud B and Varela FJ (1999) Beyond the gap: an introduction to naturalizing phenomenology. In: Petitot J, Varela FJ, Pachoud B and Roy J-M (eds) *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, pp. 1–80. Stanford, CA: Stanford University Press.

Sartre J-P (1948) *The Psychology of Imagination*. New York, NY: Philosophical Library. [First published 1940.]

Savardi U and Bianchi I (2000) *L'Identità dei Contrari* Verona, Italy: Cierra.

Shepard RN and Cooper LA (1982) Mental images and their transformations. Cambridge, MA: MIT Press.

Shepard RN and Metzler J (1971) Mental rotation of three-dimensional objects. *Science* **171**: 701–703.

Simon HA and Kaplan CA (1989) Foundations of cognitive science. In: Posner MI (ed.) *Foundations of Cognitive Science*, pp. 1–47. Cambrdige, MA: MIT press.

Smith JA, Harré R and Langenhove LV (eds) (1995) *Rethinking Methods in Psychology*. London: Sage.

Wandell BA (1999) Computational neuroimaging of human visual cortex. *Annual Review of Neuroscience* **22**: 145–173.

Zeki S (1993) *A Vision of the Brain*. Cambridge, MA: Blackwell.

Piagetian Theory, Development of Conceptual Structure

Introductory article

Kurt Fischer, Harvard Graduate School of Education, Cambridge, Massachusetts, USA
Ulas Kaplan, Harvard Graduate School of Education, Cambridge, Massachusetts, USA

CONTENTS

Jean Piaget established an important approach to the study of mental development. He distinguished a series of stages of cognitive development and outlined a process of change through which children actively create a new type of intelligence at each stage. New tools, in contextual and cultural analysis, dynamical systems theory, and neuroscience, are helping to realize his goal of connecting developmental stages with processes of growth.

INTRODUCTION

Jean Piaget advanced one of the most important endeavors in cognitive science: the analysis of the development of the mind. Working in what he called genetic epistemology (study of the genesis of knowledge), he described cognitive development richly and argued compellingly that children and adults build understanding and knowledge through their own activities in the world. In his constructivist view, people stand between nature and nurture when they act on objects, events, and people: through coordinating those activities, people drive the development of their own individual intelligences, as well as the collective intelligence of human cultures. Piaget's emphasis on people as agents constructing their own reality contributed to the birth of cognitive science and established firmly the argument that children build their intelligence through their own activities. They are neither 'tabulae rasae', shaped entirely by their environment, nor unfolding flowers determined entirely by their genes, but dynamic agents connecting themselves with their worlds.

In the 1920s, Piaget began the effort to explain human action and thought with analysis of how children act on things in their environment and thus learn about world and self. In this early work, and in the extensive developmental research that followed it, he emphasized two principles: firstly, that each individual actively constructs knowledge; and secondly, that the organization of that construction moves systematically through stages based on qualitative transformations of the mind. Piaget's work engendered and influenced much research and theory in the second half of the twentieth century, some building upon his work and some opposing it. By the end of the twentieth century, new findings and concepts about developmental processes, context, culture, and brain functioning had made it possible to study the transformation and development of concepts and skills in ways that built on Piaget's work and went beyond it. Areas that Piaget could only speculate about have become focuses of active research.

THE BASIS OF KNOWLEDGE IN ACTION ON THE WORLD

Jean Piaget and his wife Valentine, a psychologist, created the foundation of Piaget's theory of cognitive development by building on their detailed observations of how their own infants acted with objects and events. Piaget's book *The Origins of Intelligence in Children*, published in 1936, focused on his first principle: how infants construct their activities and knowledge with real objects in particular contexts, learning, for example, how to make a mobile move in the crib or how to make the mother return after she walks away. In *The Construction of Reality in the Child* published in 1937, Piaget emphasized his second principle: how infants' skills develop through a sequence of stages that are defined by patterns of action and appear across different domains, especially those of objects, space, causality, and time. Piaget's

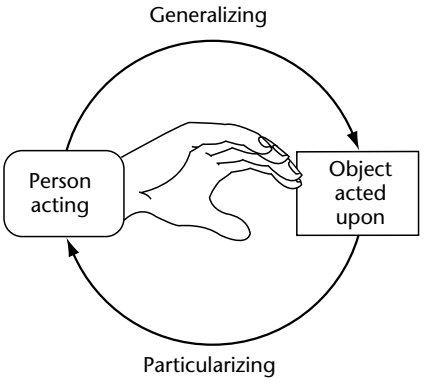


Figure 1. Piaget's 'knowing circle': understanding objects through acting on them. A person acts on an object, event, or person, and thus comes to know it by adapting his or her actions to it. Piaget called the generalizing aspect of a knowing action 'assimilation', and the particularizing aspect 'accommodation'.

long-term goal was to integrate the two principles, bringing together person and environment in a single framework.

Piaget's model for the process of development was equilibration, whereby children's activities in the world involve a continual reciprocal collaboration between person and object, as shown in Figure 1. A child acts on an object in a specific context: for example, picking up a toy rattle in her hand from her crib, shaking it, and hearing a rattling noise. In this action – a 'scheme', in Piaget's terminology – she uses her grasping, manipulating, and listening in a general way to act on the rattle, and she adapts those activities to the particular properties of the rattle, including its graspability and its capacity to make noise. The generalizing role of action is what Piaget called 'assimilation', and the particularizing role of adapting to the object is what he called 'accommodation'. Figure 1 shows how assimilation and accommodation work together simultaneously in action to produce knowing. This equilibration enables a person to function in and adapt to an environment, or context. The person and context exist together as integral parts of each other in intelligence.

An infant builds up each action scheme by repeatedly trying out an action with a particular object, such as the rattle. Piaget called this repetitive process a 'circular reaction', following the American psychologist James Mark Baldwin. Through repetition and variation, the baby learns how to control the action in that context, and then works to extend it to other contexts and build a more generalized scheme. Eventually this process leads to equilibrium, in which a knowledge scheme fits an object and context so well that activities are stable and well adapted, as when an infant can manipulate a rattle at will to produce its rattling sound.

Piaget hypothesized that the strongest form of equilibrium is structured by logic, which children gradually construct from their activities, producing a series of qualitatively distinct stages of intelligence. According to Piaget, logic has an intrinsic stability to it because it involves coordination of diverse activities in a single system without contradiction. At the same time, all knowledge schemes – even logical ones – include the seeds for further development. An existing scheme leads a child to recognize limitations in his or her knowledge, pointing to things that he or she does not understand. Stimulated by such 'disequilibria', people move on to build new knowledge upon the foundations of existing knowledge.

QUALITATIVE CHANGES IN INTELLIGENCE PRODUCED BY ACTION ON THE WORLD

Piaget described a series of 'stages' or 'periods' in the development of intelligence (see Table 1). He also described substages within each stage. When the stages and substages are used as tools for describing the qualitative changes in intelligence, their number depends on how fine-grained a description is required; but when analyzed in terms of logic, they are more precise and fixed.

These broad stages are often called periods. Three of the four periods involve different types of logic. The periods build on each other, forming a hierarchy in which each higher form of intelligence includes the earlier ones – similar to the way in which a body is composed of organs, organs are composed of cells, and cells are composed of organic molecules. Piaget concentrated on knowledge about the physical world in his research, but other researchers have elaborated Piaget's concepts to apply to knowledge of people as well. For example, Lawrence Kohlberg described the development of moral judgment through stages related to Piaget's.

In the first period, characterized by sensorimotor intelligence, infants understand their worlds through direct actions and perceptions, such as grasping, moving, looking, and listening. The stages of infant development show gradual coordination and differentiation of these actions to produce successively more sophisticated systems, which culminate at 1 to 2 years of age in what Piaget called *the logic of action*: infants can find

Table 1. Piaget's four main stages of intelligence

Stage	*Schemes*	*Problems dealt with*	*Type of logic*	*Age*
Sensorimotor	Actions	Acting on world, including objects, space, causality, and time	Logic of actions, such as movement in space	Infancy: 0–2 years
Preoperational	Representations	Symbols, including thought, language, imitation, and play	No logic, but egocentric, illogical thought	Early childhood: 2–6 years
Concrete-operational	Operations on representations	Classes, relations, and numbers	Logic of mental actions on representations (concrete operations)	Middle childhood: 6–12 years
Formal-operational	Operations on operations	Abstract and hypothetical thinking	Logic of the hypothetical and possible (formal operations)	Adolescence and adulthood: beyond 12 years

objects that are hidden, recognizing their permanence; and they can find their way around in space, relating multiple paths to get back to the same location in a system of spatial displacements. Piaget proposed that this logic of action is the culmination of development in infancy and that it creates a powerful equilibrium in intelligence. At the same time it launches a new form of intelligence, and many disequilibria.

In the second period, young children use this new form of intelligence: preoperational intelligence, the capacity to represent objects, people, and events in the mind. Representations are actions that children do in their minds without needing to carry out the overt actions or perceptions, such as when a 3-year-old girl pretends that a doll is her friend Jennifer taking a bath and also describes the activity in words ('Jennifer is wet in the bathtub'). Although these representations are based in the logic of action, the logic is only in the actions, not in the representations. The thought of such children is mostly illogical, involving frequent contradictions, and egocentric. This leads to disequilibria, which in turn lead to a new kind of logic and intelligence.

Concrete-operational intelligence is the logic of concrete operations on representations, in which a child can combine several representations in his or her mind, coordinating characteristics of an object that can change or compensate for each other. For example, a child with preoperational intelligence believes that when a ball of clay is elongated into a sausage shape, it becomes bigger. A child with concrete-operational intelligence understands that the reduced width of the sausage compensates for its increased length, so that the total amount remains the same. Similarly, children become able to take other people's concrete perspectives and coordinate them with their own, moving from egocentrism to 'perspective taking'.

In the fourth and final period, the adolescent or adult with formal-operational intelligence is no longer limited to the actual and concrete but can mentally coordinate the world of possibilities, creating a logic of the hypothetical by representing operations on concrete operations. Like the scientist performing an experiment, the formal-operational thinker can systematically consider all possible causes and outcomes of a situation in relation to a hypothesis or question. Such a person may be able to solve problems like getting a stalled car to start, explaining the results of a combination of chemicals, or evaluating the morality of hurting one person in order to help another.

VARIATION, DYNAMICS, AND CATASTROPHES

There are many variations in cognitive development that go beyond Piaget's descriptions. Much of what people know and believe does not seem to fit the logical structures that Piaget described, and even for tasks that do involve logic, people show wide variability in development – much more than Piaget's theory can explain. Many researchers have described patterns of *décalage* or unevenness, in which, for example, understanding of conservation of amount of clay develops at a different age from other kinds of conservation understanding. In particular, the skills of young infants sometimes show knowledge of objects, space, causality, and number that are more advanced than Piaget described.

Variability is also evident within individuals. The stages that a person demonstrates in one assessment session greatly depend on the immediate contextual support, showing a wide 'developmental range'. The term 'levels' is often used instead of 'stages' to reflect this variability. For instance, young adults may exhibit advanced formal-operational thinking when supported by questions and prompts from a knowledgeable teacher, but revert to concrete-operational thinking without that support. Research inspired by James J. Gibson, Lev Vygotsky and others has shown how the physical and social worlds respectively support cognition and development. This work explains much of the variability that posed problems for Piagetian theory.

Until recently, the main focus of research and theory has been the stages of development rather than the processes of equilibration. The new understanding of contextual support is one of several new tools that make it possible to integrate processes of development with stages. Mathematical techniques for describing nonlinear dynamical change (including theories of chaos, catastrophe, and complexity) provide powerful tools for analyzing growth processes, such as Paul van Geert's nonlinear mathematical model of Piaget's equilibration. Under many different growth conditions, the model produces three or four stages of development, as Piaget posited; at the same time, it produces pervasive and interesting patterns of *décalage* and variation.

Other research has established clear empirical criteria for emergence of a stage, and conditions under which development shows stages. Kurt Fischer has found that children's skills show sudden spurts with the emergence of a new stage or level, but only under conditions of high contextual support, not in most ordinary activities. Contrary to Piaget's hypothesis, developing skills do not all fit a single logical structure, but instead they develop independently along separate strands (domains) in a developmental 'web', as illustrated in Figure 2. However, concurrent change to a new level does occur across many strands for activities with high contextual support (as shown by the time zones marked by boxes in Figure 2). The changes can include not only spurts in skill but also joining of domains or separating of strands into distinct domains.

Although no structures have the universal applicability that Piaget sought for logic, Robbie Case has shown that some 'central conceptual structures' generalize broadly across relevant, limited domains. The best-documented example is

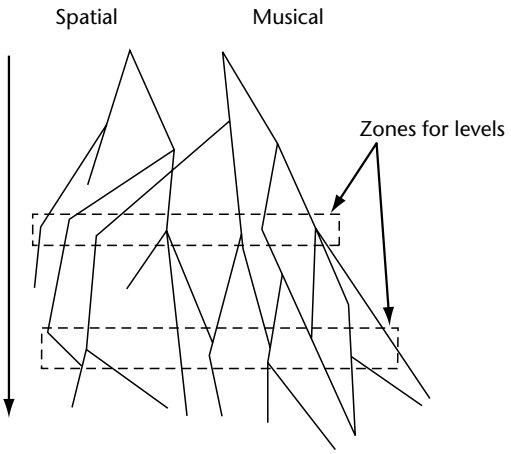


Figure 2. Developmental web showing emergence of two levels in two domains. People develop along many independent strands even within a coherent domain, as illustrated for spatial skills and musical skills. Simultaneously, performance under optimal conditions shows clusters of discontinuous changes across many strands within a zone, marking a new developmental stage or level, as indicated by the boxes.

the number line, which frames number in terms of a geometric line divided into segments by units, a structure that is strongly supported in language and cultural practices.

Fischer, Case and others have also shown that children develop through several substages for each general stage, with substages marked by spurts under high support. These substages explain many of the findings that abilities develop earlier or later than Piaget thought. Neuroscience research suggests that neural networks in the brain are reorganized at each substage and stage, as reflected by changes in patterns of electrical activity and blood flow in the cerebral cortex. Mathematical models of neural networks have many of the properties of development that Piaget postulated, including dependence on activity, sensitivity to contextual support, and occurrence of relatively abrupt reorganizations.

CONCLUSION

Piaget characterized development of intelligence in terms of two broad principles: that children construct knowledge and skill through acting on objects, people, and events in the world; and that their intelligence moves through a series of qualitative reorganizations that build on each other hierarchically and are markedly different from each other. The processes of change involve equilibration, whereby a person acts on objects using

existing schemes (activity structures), gradually achieving relatively stable knowledge states. Each such state leads to recognition of things that are confusing or not understood, and thus drives the person towards new knowledge.

Piaget's broad principles of equilibration and stages continue to frame most current research in cognitive development. Although Piaget dealt only broadly with important issues such as context, dynamical systems, and the brain, his framework is remarkably consistent with recent directions and findings in cognitive science and developmental neuroscience. Piaget's analysis of the development of intelligence still defines most of the basic questions and concepts that drive research in these fields.

Further Reading

Case R (ed.) (1991) *The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge*. Hillsdale, NJ: Lawrence Erlbaum.

Elman JL, Bates EA, Johnson MK *et al*. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: Bradford Books.

Fischer KW and Bidell TR (1998) Dynamic development of psychological structures in action and thought. In: Lerner RM (ed.) and Damon W (series ed.) *Handbook of Child Psychology*, vol. I: *Theoretical Models of Human Development*, 5th edn, pp. 467–561. New York, NY: John Wiley.

Fischer KW and Rose SP (1994) Dynamic development of coordination of components in brain and behavior: a framework for theory and research. In: Dawson G and Fischer KW (eds) *Human Behavior and the Developing Brain*, pp. 3–66. New York, NY: Guilford Press.

van Geert P (1998) A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review* **105**: 634–677.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Inhelder B and Piaget J (1958) *The Growth of Logical Thinking From Childhood to Adolescence*, translated by A. P. S. Seagrim. New York, NY: Basic Books. [First published 1955.]

Kohlberg L (1969) Stage and sequence: the cognitive developmental approach to socialization. In: Goslin DA (ed.) *Handbook of Socialization Theory and Research*, pp. 347–480. Chicago, IL: Rand McNally.

Piaget J (1952) *The Origins of Intelligence in Children*, translated by M. Cook. New York, NY: International Universities Press. [First published 1936.]

Piaget J (1954).*The Construction of Reality in the Child*, translated by M. Cook. New York, NY: Basic Books. [First published, 1937.]

Piaget J and Inhelder B (1969). *The Psychology of the Child*. New York, NY: Basic Books. [First published 1966.]

Vygotsky L (1978) *Mind in Society: The Development of Higher Psychological Processes*, translated by M. Cole, V. John-Steiner, S. Scribner and E. Souberman. Cambridge, MA: Harvard University Press.

Post-traumatic Stress Disorder

Introductory article

Rachel Yehuda, Sinai School of Medicine, New York, New York, USA

CONTENTS

Post-traumatic stress disorder is an anxiety disorder in which an individual's ability to function is impaired by emotional responses to memories of a traumatic event.

INTRODUCTION

The diagnosis of post-traumatic stress disorder (PTSD) first appeared in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* in 1980. This disorder develops in response to a terrifying event or ordeal that has been experienced, witnessed or learned about. The event is usually life-threatening or capable of producing bodily harm, and typically involves interpersonal violence or disaster. Examples of such events include rape, assault, torture, car or plane crashes, or being exposed to earthquake, tornadoes or floods. What these events have in common is their ability to cause feelings of intense fear, horror or helplessness. This response can lead to a cascade of adverse psychological reactions that can result in substantial disability.

In the field of mental health, professionals have been slow to recognize that the psychological effects of traumatic experiences can be long-lasting. Prior to the establishment of the diagnosis of PTSD, stress-related symptoms tended to be viewed as transient, and not requiring intensive treatment. This view was in keeping with the pervasive feeling in society that, with time, people ought to be able to 'get over' the effects of a traumatic experience, and to move on without noticeable impairment. Accordingly, those who did develop long-term symptoms following trauma were perceived as being constitutionally vulnerable. The diagnosis of PTSD has permitted greater understanding of the impact of traumatic events in precipitating long-term psychological symptoms, and has created a model that allows systematic hypothesis testing about the nature of adverse consequences following trauma. In the 1990s scientific developments led to a better understanding of the etiology of PTSD, the course of the disorder, and the biological changes associated with its development.

CLINICAL FEATURES

Post-traumatic stress disorder is classified as an anxiety disorder according to the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), and defines a syndrome in which a person is unable to modulate emotional responses to memories of a traumatic event. The disorder can affect many aspects of a person's life, particularly day-to-day functioning, quality of life, and relationships. Three symptom clusters are associated with PTSD:

- Reexperiencing the event through distressing images, unwanted memories, nightmares or flashbacks of the event that cause distress and attendant physical symptoms such as palpitations, shortness of breath and other manifestations of panic.
- Avoidance of reminders of the event, including people, places or things associated with the trauma, and becoming emotionally numb, constricted or generally unresponsive to the environment.
- Hyperarousal, reflected by physiological symptoms such as insomnia, irritability, impaired concentration, hypervigilance and increased startle responses.

The disorder can cause impairment, and in some cases substantial disability, in social, occupational and interpersonal domains.

People who develop PTSD usually begin to show signs of anxiety and emotional withdrawal in the immediate aftermath of the traumatic event. Neither trauma survivors nor their families are particularly alarmed by such symptoms immediately after the event, believing them to be normal responses that will pass with time. Indeed, most trauma survivors wait for several weeks or months

before contacting a physician in order to give themselves time to resolve their traumatic reactions on their own. Ultimately, the failure of post-traumatic symptoms to improve spontaneously, and the increasing disability caused by sleep disturbance, panic and depression, cause people to seek professional help.

The response that is perhaps most detrimental in the long run to trauma survivors is that of avoiding talking or thinking about (i.e. 'processing') the event. Most experts assert that giving trauma survivors a forum to discuss the event and their resultant feelings about it promotes recovery by minimizing avoidance. However, since talking about the event may be distressing, both survivors and those who are there to listen to them are often content to avoid such discussions. Yet PTSD symptoms can be ameliorated once the survivor actively confronts the feelings of fear and vulnerability associated with the traumatic event.

ETIOLOGY

Studies of the prevalence of PTSD have demonstrated that this condition is the fourth most common psychiatric disorder, afflicting 7–14 percent of the population at some time in their lives. The nature of the trauma experienced seems to be the single most significant factor in the etiology of PTSD. Events involving interpersonal violence, such as torture, rape, assaultive violence and combat, are more potent agents in the causation of PTSD than experiences such as motor vehicle accidents and natural disasters. The former events produce PTSD in as many as 50–75 percent of trauma survivors, whereas the latter result in PTSD approximately 10–25 percent of the time. For specific events, dose–response relationships have been observed between degree of exposure and incidence of PTSD.

The observation that for any given trauma only a subset of those exposed will develop chronic PTSD has led to the search for other risk factors that contribute to the development of (or recovery from) this disorder. In addition to the nature and severity of the traumatic event, previous exposure to stress or trauma, particularly in childhood, a history of psychological and behavioral problems, and familial factors such as parental PTSD and family history of anxiety and depression, have been noted as risk factors. Gender also appears to be a potent risk factor for the development of this disorder, with studies consistently demonstrating it to be doubly prevalent in women.

Epidemiological studies have identified clusters of risk factors that are clearly interrelated. For example, lower levels of education and income, differences in ethnicity, poverty, and lower intellectual functioning have been identified as risk factors for the development of PTSD in the wake of a traumatic event. However, these variables are also associated with a greater exposure to some kinds of traumatic events, such as assaultive violence.

A history of family instability is associated with increased incidence of PTSD, and numerous studies have indicated that familial psychiatric history may place an individual at higher risk. In particular, parental psychopathology and/or PTSD appears to be a highly specific risk factor for the development of PTSD in offspring. It is not clear whether the tendency to develop PTSD is genetically inherited. An intriguing finding in a study on soldiers, examining PTSD in monozygotic (identical) twins, demonstrated that as much as 30 percent of some PTSD symptoms were present in both the combat-exposed twin and his nonexposed co-twin. However, this finding leaves open the possibility that either genetic or shared environmental influences may contribute to vulnerability for this disorder.

The development of PTSD may also be associated with how the survivor interprets the traumatic event in the context of preexisting ideas about personal safety. Individuals who believe that they can no longer be safe after experiencing a traumatic event, for example, may be more likely to develop chronic PTSD following a trauma. In addition, interpreting initial symptoms such as intrusive thoughts and hypervigilance as a sign of falling apart or being permanently altered for the worse may serve to maintain them. Since ideas about the world and one's ability to cope with adversity are often shaped by prior experience, this may in part explain why exposure to earlier adversity places an individual at risk at greater risk for developing PTSD following exposure to a later event. Unfortunately, little is known about resiliency factors that prevent the development of PTSD or increase recovery once this condition develops.

COURSE OF THE DISORDER

The symptoms of PTSD do not necessarily become more severe as time passes. However, the failure of the trauma-related symptoms to remit often results in a cascade of secondary behavioral, emotional and personality problems. People who develop PTSD are also more likely to develop other psychiatric disorders – such as mood, personality and eating disorders – and substance dependence.

Post-traumatic stress disorder can remit spontaneously or following treatment. However, the likelihood of developing a recrudescence of symptoms increases upon further exposure to stress or trauma.

The course of PTSD may thus vary considerably from person to person, taking acute, chronic or intermittent forms. Although in the majority of PTSD cases symptoms show immediate onset following the trauma, sometimes onset may be delayed. Delayed onset is usually stimulated by an environmental trigger. For example, a woman who was sexually abused as a 5-year-old might not develop PTSD at the time of the event but as an adult, triggered by her daughter's fifth birthday. A combat veteran might develop symptoms in response to hearing about the outbreak of a war decades later. Retirement can be a trigger of post-traumatic symptoms in elderly people traumatized decades earlier, if the survivor had used a lifelong pattern of keeping busy with work as a way of delaying thinking about the traumatic event and resultant feelings associated with it.

Community studies have indicated that the nature of the symptoms experienced changes over time, with intrusive symptoms being particularly prominent in earlier stages of the disorder, and avoidance and arousal being more significant later on. Treatment can influence the course of PTSD. Many treatment approaches have been demonstrated as effective in reducing PTSD symptoms, even to the point of achieving remission. Psychotherapeutic or counseling methods, such as cognitive-behavioral therapy including exposure and anxiety management treatments, have been shown to be effective: these methods encourage trauma survivors to talk about their traumatic experiences, and to view direct confrontation of frightening emotional memories as essential for symptom resolution. Medications such as antidepressant and anticonvulsant drugs have also been shown to be effective, and the use of two selective serotonin reuptake inhibitors has been approved by the US Food and Drug Administration for the treatment of PTSD. Usually some combination of psychotherapy and medication is required for the best possible treatment outcome.

NEURAL AND HORMONAL CORRELATES

Alterations have been observed in hormonal systems involved in the body's response to stress in PTSD. These alterations include increased concentrations of noradrenaline (norepinephrine) and an amplified reactivity of noradrenergic receptors, increased thyroid hormone levels, and increased reactivity of the hormones of the hypothalamic–pituitary–adrenal (HPA) axis. Anatomical changes occur in two major brain structures – the amygdala and hippocampus – that are directly involved in fear responses and in the acquisition of traumatic and emotional memories. The amygdala and hippocampus are important target organs in both initiating and terminating biological stress responses; changes in their activity in PTSD may be linked to other chronic, stress-related alterations in hormones and neurotransmitters.

It is important to understand that the alterations observed in PTSD do not uniformly resemble those associated with stress. One of the more unexpected findings has been the observation of low cortisol levels in some studies of chronic PTSD, even decades after trauma exposure. These and other findings have led to the idea that the development of PTSD may be facilitated by a failure to contain the normal stress response at the time of the trauma, resulting in a cascade of biological alterations that lead to intrusive, avoidance and hyperarousal symptoms.

The normal fear response is characterized by a series of biological reactions, initiated by the amygdala. Activation of the sympathetic nervous system and the release of adrenaline (epinephrine) allow the organism to increase its physiological capacity for 'fight or flight' in response to threat, and facilitate consolidation of the threat memory. The simultaneous activation of the HPA axis, culminating in the release of cortisol, helps contain the neural, defensive stress reactions. As stress-activated biological reactions are restricted, elevated cortisol levels also suppress the further release of cortisol through negative feedback inhibition on the pituitary, hypothalamus, hippocampus and amygdala.

Prospective biologic studies have begun to demonstrate that individuals who develop PTSD or PTSD symptoms appear to have smaller cortisol increases in the acute aftermath of a trauma than those who do not develop this disorder. Furthermore, those who develop PTSD show higher heart rates in the emergency room and 1 week after the trauma compared with those who ultimately recover, suggesting a greater degree of sympathetic nervous system activation. These findings imply that the actual biological response to acute trauma may be different in people who develop PTSD, some changes being typical and others atypical of the 'normal' stress response. These data offer an intriguing opportunity for developing hypotheses about processes in the early aftermath of trauma

that appear to forestall recovery and prevent the natural restitution of the stress response.

One model explaining the development of PTSD following trauma proposes that the increased sympathetic nervous system activity leads to an exaggerated response by this system to the trauma, manifested by an increased concentration of adrenaline. This in turn initiates a process in which traumatic memories become 'overconsolidated' or inappropriately remembered owing to an exaggerated level of distress. The primary mechanism through which adrenaline facilitates memory formation is by maintaining organisms at a high level of arousal. If cortisol failed to adequately shut down adrenaline production this arousal might be prolonged, and the consolidation of the memory facilitated. The increased distress every time there are traumatic reminders would further activate stress-responsive systems, resulting in secondary biological alterations associated with anxiety and hyperarousal. Although there may be other biologic pathways leading to PTSD, the above appears to be a promising candidate.

Further Reading

Breslau N, Kessler RC, Chilcoat HD *et al*. (1998) Trauma and posttraumatic stress disorder in the community: the 1996 Detroit Area Survey of Trauma. *American Journal of Psychiatry* **55**: 626–632.

Davidson JRT and Foa EB (1993) *Posttraumatic Stress Disorder: DSM-IV and Beyond*. Washington, DC: American Psychiatric Press.

Foa EB, Steketee G and Rothbaum BO (1989) Behavioral/cognitive conceptualizations of post-traumatic stress disorder. *Behavior Therapy* **20**: 155–176.

Foa EB, Keane TM and Friedman MJ (2000) *Effective Treatments for PTSD*. New York: Guilford Press.

Yehuda R (2002) Current concepts: posttraumatic stress disorder. *New England Journal of Medicine* **346**: 109–114.

Yehuda R, Shalev AY and McFarlane AC (1998) Predicting the development of posttraumatic stress disorder from the acute response to a traumatic event. *Biological Psychiatry* **44**: 1305–1313.

Prejudice

Intermediate article

William von Hippel, University of New South Wales, Sydney, Australia
Steven Fein, Bronfmann Science Center, Williamstown, Massachusetts, USA

CONTENTS

Prejudice is conscious or unconscious animosity towards members of other groups. Although far fewer people openly express dislike of other groups today than in the past, unconscious forms of prejudice are still common, and have a subtle influence on how people treat each other.

ROOTS OF PREJUDICE

There are many different forms of prejudice, and many causes of it. Prejudice is often connected with stereotypes and inter-group attitudes. (*See* **Stereotypes**; **Attitudes**)

One of the most important roots of prejudice, and perhaps the principal reason why prejudice is so universal and resistant to change, is that people are often motivated to see other groups in a negative way. This motivation may stem from a desire to gain from, or maintain the tangible rewards associated with, power over other groups. For example, when two groups are competing against each other, one group may derogate the other in order to justify taking aggressive action against it. Members of a dominant group may exhibit prejudice towards a weaker group to reinforce the status quo, whereas members of the weaker group may use hostility towards the dominant group in order to rally their group to challenge the status quo. In a classic field experiment conducted at a summer camp, the psychologist Muzafer Sherif and his colleagues (Sherif *et al*., 1961) found that creating competition between two groups of boys quickly led to intense hostility between the groups. Once these groups were induced to share common goals, however, the hostility just as quickly subsided.

Even if there is no conflict between groups, inter-group prejudice can spring from group members' desire to feel good about themselves. Social identity theory (Tajfel and Turner, 1986) proposes that in addition to the sense of self people derive from their personal qualities, they also gain a sense of self from their group memberships. Thus, people tend to feel better about themselves when their groups are held in high esteem. Because prejudice towards other groups results in the perceived relative elevation of one's own group, social identity theory suggests that prejudice can make people feel better about their groups, and therefore about themselves.

Despite these positive effects of prejudice on self-esteem, in contemporary times it is often the case that awareness of one's own prejudice can actually make one feel worse, rather than better, about oneself. It is the current social unacceptability of prejudice that often (although not always) causes prejudice to express itself through subtle, unconscious processes. The remainder of this article focuses on these processes.

UNCONSCIOUS PREJUDICE

Much like the bacteria that have mutated in response to the use of antibiotics, prejudice has changed dramatically in western society over the past fifty years. At a conscious level, people have become less likely to show 'old-fashioned' prejudice. For example, few now advocate the exclusion of minorities from their workplace or country (Schuman *et al*., 1997). Instead, most people voice egalitarian ideals; and if they do show negativity toward minorities, it is expressed in terms of resentment at unfair advantages, for example.

In contrast, unconscious prejudice is quite common (e.g. Fazio *et al*., 1995). This prevalence of unconscious prejudice in the absence of the conscious variety has led some psychologists to propose that most North Americans and Europeans are secretly bigots. Although this characterization is probably untrue, research does support the notion that prejudice tends to be learned at a young age and can be very difficult to change (Devine *et al*., 1991). Indeed, when psychologists Anthony Greenwald and Mahzarin Banaji

developed a measure of unconscious prejudice and created websites that enabled people to test themselves on their measure, they found that well over half of the hundreds of thousands of white respondents showed unconscious prejudice towards blacks (Nosek *et al*., 2002). Their measure relies on differences in the time it takes to make certain judgments (the difference between prejudiced and unprejudiced individuals is measured in milliseconds). The test is available on the internet (Nosek *et al*., 2002). Unconscious prejudice emerges with similar prevalence in other experiments, with other measures, and with samples that are not web-based or self-selected.

These findings do not mean that the advances in race relations and civil rights that have occurred since the Second World War are illusory. As Philip Tetlock (1994) noted, 'we have come a long way from the Selma, Alabama of 40 years ago if we now have reached the point in our history when white racism must be measured in milliseconds'. Tatlock is undeniably correct, as almost every indicator of race relations today shows a dramatic reduction in prejudice compared with the recent past. It is now clear that Americans have largely accepted the fact that their egalitarian credo must apply to blacks as well as whites. The depth of this change in American attitudes is apparent from the words of Ellis Cose, writing in *Newsweek*: 'If you are touched at all by American culture, your idol is likely to be black.'

In a very important way, however, Tetlock is also wrong. Just because prejudice must often be measured in milliseconds does not mean that it has only negligible effects. On the contrary, the sort of subtle, unconscious prejudice that exists today can be almost as disruptive as its more blatant precursor. For example, research indicates that people who test high on measures of unconscious prejudice are more likely to be selectively rude to minorities than people who test low on these measures (Sekaquaptewa *et al*., 2002). Such individuals deny that they are behaving rudely – indeed, they probably do not know it themselves – but a variety of behavioral measures show important differences in how they respond to members of different groups.

THE ROLE OF AMBIGUITY

These findings might lead one to ask whether we should really be concerned about the possibility that an unconscious 'bigot' might be a little unfriendly when interacting with minorities. After all, this seems much preferable to a world where a person who openly hates blacks, for example, can engage in socially sanctioned beatings and lynchings. No one is required to be friendly to everyone, so why should we worry if some whites unintentionally show a little extra rudeness or displeasure towards blacks?

Unfortunately, in some ways, the subtle effects of unconscious prejudice are more pernicious than the obvious effects of blatant prejudice. Society at large tends to recoil from the latter, and we feel sympathy for its victims (this response to blatant prejudice is what made the civil rights movement in the United States so effective). On the other hand, when prejudice is subtle and its effects ambiguous, we tend to doubt or deny it, and grow weary of claims (and claimants) of its existence.

Consistent with this argument, a number of experiments have shown that when people are in situations in which their negative behavior could easily be interpreted in terms of prejudice, they go out of their way to behave in an egalitarian fashion, often unintentionally favoring members of minority groups (Dovidio and Gaertner, 1998). This sort of 'bending over backwards' – whereby people are biased in favor of minorities – is quite common, as people use such circumstances to demonstrate to themselves and others that they harbor no prejudices.

These experiments also demonstrate, however, that unconscious prejudice influences behavior when the situation provides the necessary excuse. As a consequence, those who regularly are on the receiving end of prejudice learn that it can be hard to identify. For example, imagine a black student who is evaluated negatively by a white teacher. The student is confronted with two possible explanations: 'I did a poor job' versus 'the teacher is biased'. Even when the student is rated positively, the student is again confronted with two possible explanations: 'I did a good job' versus 'the teacher is bending over backwards so as not to look prejudiced'. As psychologists Brenda Major and Jennifer Crocker (1993) have demonstrated, because there is usually no way to distinguish between these alternative explanations, blacks are often unsure whether they can believe the evaluations of whites. The result is a lack of trust between blacks and even well-meaning whites. This lack of trust can become especially acute when whites are in the role of supervisor or teacher, and blacks are in the role of employee or student, leading to a breakdown in communication that both sides find frustrating and difficult to overcome. See Cohen *et al*. (1999) for one solution to this problem.

SITUATIONAL DETERMINANTS OF UNCONSCIOUS PREJUDICE

Lest one conclude that unconscious prejudice only ever manifests in subtle ways, it is worth noting that it can also have dramatic effects. Because unconscious prejudice is usually stationed only just outside the spotlight of attention, it is ready to manifest itself in conscious thought and behavior on a moment's notice when it becomes socially sanctioned or psychologically salubrious. Consider, for example, the reaction of most United States citizens to the terrorist attacks of 11 September 2001. Within moments of learning who the perpetrators were, many found themselves possessing and expressing a strong dislike for Arabs and Muslims. Because this enmity was now perceived to be socially acceptable, racist statements that would otherwise have remained hidden became commonplace. Indeed, the numbers of verbal and physical attacks on Arabs and Muslims increased in many western countries. Clearly such discriminatory behavior would hardly be tolerated or expressed during peacetime, but, as many experiments and natural situations have shown, conflict brings unconscious prejudices and generalized dislike of other groups to the surface.

It is not only dramatic events like wars that bring unconscious prejudices to the fore: mundane situations can have the same effect. For example, people are particularly likely to denigrate others when they are feeling bad about themselves. This behavior can be functional for people who are feeling low, as it changes their point of reference. In order to feel better about oneself (in relation to others), one can either raise oneself above others or lower others below oneself. Because it is easier to denigrate others than to improve oneself, the method of putting other people down in order to restore one's own self-esteem is popular and effective. Furthermore, this process can be particularly successful when the denigrated target is an entire group of people, because under this circumstance, looking down on others leads a person to feel better than an entire segment of the population.

A series of experiments examining this idea has demonstrated that even people who show no prejudice under normal circumstances will discriminate against a variety of different groups (e.g. gays, Jews, Asian Americans, African Americans) when they are led to feel bad about themselves (Fein and Spencer, 1997; Spencer *et al*., 1998). In these experiments, college students (predominantly liberal) were told either that they had performed very well or that they had performed very poorly on an intelligence test. When the students thought they had done well, they were magnanimous in their evaluations of others. In contrast, when they thought they had done poorly, they became selectively nasty towards members of other groups – for example, showing dislike toward blacks and Jews but not towards whites and Christians. In effect, the students were consoling themselves with thoughts like 'I may not have done well on that test, but at least I'm not Jewish.' Indeed, the more prejudice they showed, the more they restored their sense of self-worth.

These findings suggest that prejudice may be difficult to eradicate in part because it can be rewarding, making people feel better about themselves at the expense of others. However, to the extent that people internalize goals of egalitarianism, and see connections and similarities between their own groups and other groups, the rewards of prejudice may diminish, compared with the rewards of fairness.

References

Cohen GL, Steele CM and Ross LD (1999) The mentors' dilemna: providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin* **25**: 1302–1318.

Devine PG, Monteith MJ, Zuwerink JR and Elliot AJ (1991) Prejudice with and without compunction. *Journal of Personality and Social Psychology* **60**: 817–830.

Dovidio JF and Gaertner SL (1998) On the nature of contemporary prejudice: the causes, consequences, and challenges of aversive racism. In: Eberhardt J and Fiske ST (eds) *Racism: The Problem and the Response*, pp. 1–32. Newbury Park, CA: Sage.

Fazio RH, Jackson JR, Dunton BC and Williams CJ (1995) Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of Personality and Social Psychology* **69**: 1013–1027.

Fein S and Spencer S (1997) Prejudice as self-image maintenance: affirming the self through derogating others. *Journal of Personality and Social Psychology* **73**: 31–45.

Major B and Crocker J (1993) Social stigma: the consequences of attributional ambiguity. In: Mackie DM and Hamilton DL (eds) *Affect, Cognition, and Stereotyping: Interactive Processes in Group Perception*, pp. 345–370. San Diego, CA: Academic Press.

Nosek BA, Banaji MR and Greenwald AG (2001) Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics* **61**: 101–115. http://buster.cs.yale.edu/implicit/

Schuman H, Steeth C, Bobo L and Krysan M (1997) *Racial Attitudes in America: Trends and Interpretation*. Cambridge, MA: Harvard University Press.

Sekaquaptewa D, Espinoza P, Thompson M, Vargas P and von Hippel W (2002) Stereotypic explanatory bias: implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*.

Sherif M, Harvey LJ, White BJ, Hood WR and Sherif CW (1961) *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Middletown, CT: Wesleyan University Press.

Spencer SJ, Fein S, Wolfe CT, Fong C and Dunn MA (1998) Automatic activation of stereotypes: the role of self-image threat. *Personality and Social Psychology Bulletin* **24**: 1139–1152.

Tajfel H and Turner JC (1986) The social identity theory of intergroup behavior. In: Worchel S and Austin WG (eds) *Psychology of Intergroup Relations*, pp. 7–24. Chicago, IL: Nelson.

Tetlock PE (1994) Political psychology or politicized psychology: is the road to scientific hell paved with good moral intentions? *Political Psychology* **15**: 509–552.

Further Reading

Allport G (1954) *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Brewer MB (1999) The psychology of prejudice: ingroup love or outgroup hate? *Journal of Social Issues* **55**: 429–444.

Crocker J, Major B and Steele C (1998) *Social stigma*. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, 4th edn, pp. 504–553. New York, NY: McGraw-Hill.

Fiske S (1998) Stereotypes, prejudice, and discrimination. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, 4th edn, pp. 357–411. New York, NY: McGraw-Hill.

Jones JM (1997) *Prejudice and Racism*, 2nd edn. New York, NY: McGraw-Hill.

Oskamp S (ed.) (2000) *Reducing Prejudice and Discrimination: The Claremont Symposium on Applied Social Psychology*. Mahwah, NJ: Lawrence Erlbaum.

Priming in Psychopathology

Introductory article

Brendan Maher, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

Priming is a process involved in memory and in the preparation of responses to coming events. Anomalies in priming have been identified in some kinds of psychopathology, especially in the schizophrenias. Studies of priming by psychopathologists have been primarily interested in priming as a technique for the investigations of the influence of the activation of associations to a presented stimulus upon responses to subsequent stimuli. In this respect their purposes have differed somewhat from those of cognitive scientists interested in dissecting the components of normal cognition.

INTRODUCTION

When the occurrence of a specific stimulus event (A) indicates an increased probability that it will be followed by a second specific event (B), the second event is likely to be perceived and responded to more rapidly and accurately than if it had occurred without the prior occurrence of A. This sequence is an example of 'priming'. The appearance of A (the 'prime') affects the readiness of the individual to respond to B (the 'target'). B may be a repetition of A. It is not necessary that the individual have consciously noticed the appearance of A. It is sufficient to demonstrate priming if we can show that the response to B is faster, more accurate, and more adaptive than it is when it is not preceded by A.

Priming occurs on the basis of predictable sequences that have been experienced in the past and have been stored in memory. The sequences may consist of spoken or written words, pictures, signs, or indeed any events that can be perceived, processed, and stored. Research into priming and psychopathology is conducted on the basis of hypotheses about how sequences are stored in the brain and how deviant brain functioning may affect priming. Defective priming can lead to maladaptive behavior, and for this reason the process has been studied intensively in patients with serious psychopathology. These include patients with major psychotic disorders such as schizophrenia, as well as patients with known brain pathology including Alzheimer's disease.

INHIBITION

Excitation and Inhibition in the Brain

The activity of the neurons, the nerve cells of the brain, can be classified in many ways. One fundamental distinction is between excitatory and inhibitory activity. The neurons of the brain are constantly active when we are awake and when we sleep. For the brain to function effectively it is essential that patterns of excitatory activity that are appropriate to the tasks of the moment can operate without disruption from irrelevant surrounding activity. This is achieved by the action of inhibitory neurons that suppress the activity of potentially disruptive excitatory cells. Adaptive regulation of inhibitory and excitatory activity appears to be located in the dorsolateral prefrontal cortex of the brain (Brodmann areas 9 and 46). Patients with lesions in the prefrontal cortex have difficulty in excluding distracting stimuli in the environment. They are therefore impaired in their ability to maintain sustained attention to a specific task, particularly if the task requires a time delay before the appropriate response can be made. There are resulting disturbances in language, the control of motor behavior, and in decision-making.

The Association Network

Our present understanding of this process of inhibition in relation to priming can be illustrated with the concept of the association network. Words (pictures, odors, music, etc.) are connected by interlinked associations. When I ask somebody to tell me the first word that comes into her head when I say 'cat' she may respond with 'mouse' or some other relevant associated word. For her, the word 'cat' has primed the association to 'mouse'. The

response will be rapid and direct and usually unaccompanied by conscious intervening reflection. The linkage is robust. But if the same individual takes her cat to the veterinarian to report that it is sick, she does not say 'My cat mouse is sick'. The strong association does not disruptively enter into her consciousness and language because the now irrelevant cat–mouse link has been inhibited. Inhibition of this kind does not occur as a conscious act but is an integral component of the production of organized thought and speech. If the inhibitory process were defective such intrusions might well enter into actual speech, which would then become incoherent.

It is useful to consider the association network around a particular word, picture, or other class of stimulus as having a central node (e.g. a specific word) and a series of branching links extending in three dimensions connecting at various points with other networks. Any one element may be a member of several networks. The content of a network is determined by individual experience; the fact that members of a culture will have many common experiences means that normative associative links may be found with some frequency in different persons. This makes it possible to study priming experimentally in the general population and people with psychopathological conditions.

FACILITATION

Semantic Priming

One method commonly used to study priming employs the technique of semantic priming, and particularly the task of lexical decision. In the lexical decision paradigm the researcher presents the participating patient with a rapid sequence of two separate strings of letters appearing successively on a computer screen. The first string of the pair is the prime. It is replaced by the second, the target. The patient is to decide whether or not the target is a real word. Pressing a keyboard key marked 'Yes' signals that it is a word, pressing one marked 'No' that it is not. The reaction time (R/T) and accuracy of response are recorded. Typical sequences might be TABLE – KROD (word – nonword), GRAX – BOAT (nonword – word). Word–word pairs are either associated, as in DOCTOR–NURSE, or nonassociated, as in TAXI–NURSE.

Studies of the general population show that R/T in response to the target in associated word pairs is significantly greater than in nonassociated pairs. For example, the word NURSE would be recognized more quickly following DOCTOR than following TAXI. The gain is interpreted as due to the automatic activation of the associated target by the prime, making recognition more immediate. The resulting difference in recognition time between the two conditions is termed 'facilitation'. Facilitation scores are computed by subtracting the mean R/T to associated pairs from that for the nonassociated pairs. The greater the difference, the more facilitation. Facilitation is most marked when the time interval between the appearance of the prime and the target is brief. Short intervals – typically of 250 milliseconds are more likely to produce facilitation than longer intervals, such as 800 milliseconds. Variations in the lexical procedure include the use of phonological ('clang') associations, such as PANG – SANG, or orthographic similarity such as CONFESS – CONFUSE.

Schizophrenia

Clinical observers of the behavior of schizophrenia patients have long pointed out that associations often intrude into their spoken and written language. One proposed explanation of this is that inhibitory processes that normally prevent this are defective. Thus activated associations are not inhibited but disrupt the coherence of the utterance, as in the example 'My cat mouse is sick' mentioned above. Intrusion of associations into utterance is extremely rare in normal speech or writing; presumably this reflects the greater effectiveness of inhibition in the normal case. In the light of this, we might expect that the patients will show greater facilitation (hyperfacilitation) than nonpatient controls. Research has shown this to be the case, particularly with patients who exhibit disturbed language. The evidence also suggests that the hyperfacilitation is more characteristic of patients the onset of whose illness has been recent. With increasing years of illness, the effect disappears and, in some patients, reverses itself – the associated pairs producing slower R/Ts than the nonassociated. This effect appears to be independent of the increasing chronological age of the patients.

Other priming paradigms for lexical decisions exist. One approach presents a degraded target word for recognition. The patient presses the key to receive increasing components of the target until the patient recognizes it correctly. Schizophrenia patients correctly identified the associated target at lower levels of clarity than were required by the control participants. There was no difference in performance under other condition. The hypothesis behind this investigation proposes that as the task is essentially one of matching the degraded

pattern of elements of the prime against the stored image of an actual word, the schizophrenia patients had the advantage of matching it against the actual association activated by the prime, while the controls did not. Another method employs word pronunciation (WP), where the patient reads the target word aloud, rather than simply pressing a key. This technique produces lower levels of facilitation generally and is less likely to elicit hyperfacilitation in patients.

Negative Priming

In some settings the appearance of an initial event (A) has the effect that the next event in sequence (B), one that would ordinarily elicit a specific response, is followed by a different or diminished response or no response at all. In effect, this requires that the normal response to B must be inhibited (not facilitated) when A has preceded it. In semantic priming this would be demonstrated when responses to associated prime–target sequences are slower or less accurate than to associated prime–target sequences.This is found in some chronic schizophrenia patients and in some patients with bipolar affective disorder.

INDIRECT PRIMING

Our understanding of associational networks is that they may be linked to each other through a series of connecting overlapping associations. For example, if the word RIVER primes the association WATER, and WATER primes WASH, which then primes SOAP, we have a sequential series of associative links which are moving progressively further away from the original prime. It is relatively easy to elicit this kind of chaining by the simple instruction to the participant to associate sequentially to his own responses. We may also demonstrate the pervasiveness of such links by asking a participant to connect two unrelated words by a chain of plausible intermediate associations. Given the task of finding a link between, let us say, TIME and GRASS, we might get TIME–CLOCK–HAND–FOOT–YARD–GRASS.

The existence of such sequences raises the question whether a prime might facilitate a target word that is two or more links away from it. In the example above, would the prime word CLOCK facilitate the target word FOOT? Psychopathologists refer to this as 'indirect priming'. The psychopathologist Eugen Bleuler first described the phenomenon in clinical cases in 1911. He used the term 'mediated' association, and this term is in current use in investigations of 'implicit' memory in normal populations. In contemporary experimental studies of psychopathological samples, the term 'indirect priming' is commonly employed. When it occurs it is regarded as evidence of relatively uninhibited excitatory activity. In schizophrenia patients indirect priming has been demonstrated in this way, but it is rare in nonpsychiatric comparison groups or in depressed patients.

Polysemous Priming

Many words or icons have more than one meaning. Such words are termed 'polysemous'. Their different meanings may be quite unrelated and each meaning has its own network of associations. English is replete with such examples. Consider the various meanings of such words as 'bank', 'stock', 'seal', 'pen', 'bark', and 'yard'. Effective thinking and communication requires that we do not confuse the possible meanings when we are trying to understand a speaker or trying to frame an utterance for others to understand. Under these circumstances the meaning (and its associations) that is not relevant to the utterance are inhibited.

In many polysemous words one meaning is more common than the others. If we ask for first associations to 'Stock' from stockbrokers we will likely get a different response than if we ask in a farming community. If, however, we have defective inhibitory processes the association to the dominant meaning may intrude into an utterance intended to convey a nondominant meaning. This happens in some utterances by schizophrenia patients, and produces what is termed 'tangential thinking', as in 'The notary put a seal on a document but it swam away'. It is evident that a chain of indirect priming has been activated by the polysemous word 'seal'. This is also found in some schizophrenia patients, suggesting again a defect in selective inhibition.

SUMMARY

Disturbances in the balance of excitatory and inhibitory processes in the neurons give rise to disturbances in language and thought in some forms of psychopathology. These disturbances may appear as the intrusion of associations into spoken or written utterances and/or difficulty in following a connected sequence of discourse. The effects seem to arise from a defect in the processes of selective inhibition and have been reported most often in patients with schizophrenia. They may appear in patients with bipolar affective disorders.

Further Reading

Kwapil TR, Hegley DC, Chapman LJ and Chapman JP (1990) Facilitation of word recognition by semantic priming in schizophrenia. *Journal of Abnormal Psychology* **99**: 215–221.

Maher BA, Manschreck TC, Hoover TM and Weisstein CC (1987) Thought disorder and measured features of language production in schizophrenia. In: Harvey PD and Walker EF (eds) *Positive and Negative Symptoms in Psychosis*, pp. 195–215. Hillsdale, NJ: Erlbaum.

Maher BA, Manschreck TC, Redmond D and Beaudette S Length of illness and the gradient from positive to negative semantic priming in schizophrenic patients. *Schizophrenia Research* **22**: 127–132.

Meyer DE and Schvaneveldt RE (1971) Facilitation in recognizing words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* **90**: 227–234.

Neely JH (1991) Semantic priming effects in visual word recognition: a selective review of current findings and theories. In: Besner D, Humphreys GW (eds) *Basic Processes in Reading and Visual Word Recognition*, pp. 264–333. Hillsdale, NJ: Erlbaum.

Spitzer M, Braun U, Maier S, Hermle L and Maher BA (1993) Indirect semantic priming in schizophrenic patients. *Schizophrenia Research* **11**: 71–80.

Vinogradov S, Ober BA and Shenaut GK (1992) Semantic priming of word pronunciation and lexical decision in schizophrenia. *Schizophrenia Research* **8**: 171–181.

Priming

Introductory article

James H Neely, University at Albany, State University of New York, New York, USA

CONTENTS

Priming is a change in the response to a stimulus (the target) due to a recent exposure to it or a similar stimulus (the prime).

INTRODUCTION

This article will consider how our responses to the multitude of stimuli that we encounter every day depend not only on our accumulated learning about those stimuli, but also on our most recent exposure to them or to related stimuli. Priming is defined as a change in the response to a stimulus (the target) due to a recent exposure to it or a related stimulus (the prime). It has been studied extensively with diverse procedures.

1. When completing a word fragment such as *a---e* with an English word, people are more likely to respond with *amble* if they have recently seen that word. This occurs even if days separate the presentation of the prime *amble* and the target word fragment. (*See* **Memory: Implicit versus Explicit**)
2. When solving the problem $3 \times 8 = ?$, people are faster if two to nine problems ago they received $3 \times 8 = ?$ as a prime, but are slower if they received the related problem $3 \times 6 = ?$ as a prime, relative to when they had been primed with totally unrelated problems.
3. When asked to describe a picture of a dog chasing a cat, people are more likely to say 'A cat is being chased by a dog' if they recently heard the word 'cat', and are more likely to say 'The dog is chasing the cat' if they recently heard the word 'dog'.

SHORT-TERM PRIMING IN WORD-RECOGNITION TASKS

Because studies of priming have been so diverse and extensive, this article will focus on short-term priming effects in word-recognition tasks. In these tasks, a reader responds as quickly as possible to a single target letter string either by pronouncing it aloud or by pressing one of two keys to indicate (1) whether it is a word or a non-word (a lexical decision task) or a noun or a verb, or (2) whether its referent is pleasant or unpleasant, animate or inanimate, larger or smaller than a soccer ball, or is a member of a specific semantic category such as *fruit*. Because errors are rare, researchers focus on the person's reaction time (RT) to the target. The prime that precedes the target can be either (1) a clearly visible word, presented for 0.05–2 s, that the reader is told to read silently or (2) a word that is presented for a very brief duration (< 70 ms), and which is immediately followed by a visual mask. With such masking the reader cannot identify the prime, and even claims to be unaware of its presentation. The time interval between the prime's onset and the target's onset, known as the stimulus onset asynchrony (SOA), is typically varied between 100 and 2000 ms to control the length of time for which the person can think about the prime before the target appears. (*See* **Word Recognition**)

With these procedures, priming effects are computed by subtracting RTs to targets when the prime and target are somehow related from RTs to targets that follow an unrelated prime. When people attend to the prime, most priming effects are positive (i.e. RTs are quicker after related than after unrelated primes). However, when people ignore the prime and attend (respond) to some other simultaneously presented stimulus, priming effects are often negative. This article will focus on studies in which the prime is attended to and the prime and target are presented alone.

A VARIETY OF PRIME–TARGET RELATIONS PRODUCE PRIMING

Most short-term word-priming studies have used pronunciation or lexical (word/non-word) decision

tasks. Unsurprisingly, repetition (e.g. *dog dog*) priming effects are positive and larger than other priming effects, and even with masked primes they are large. In orthographic priming, the prime and target share many of the same orthographic (letter) units, whereas in phonological priming their pronunciations share many of the same phonological (sound) units. Although most orthographically related primes and targets are also phonologically related, and vice versa (e.g. *tribe tripe*), one can find 'pure' orthographically related items that are phonologically unrelated (e.g. *dough cough*), and 'pure' phonologically related items that are orthographically unrelated (e.g. *eight ate*). Interestingly, orthographically related and phonologically related primes facilitate target processing to a greater extent when they are masked. In fact, when an attended unmasked prime is orthographically/phonologically related to the target, priming can be negative for lexical decisions, especially when the prime is less common than the target (e.g. *waif wait*). For unmasked primes, negative priming can also occur for 'pure' orthographically related items (e.g. *dough cough*).

In semantic priming, the prime and target are similar in meaning (e.g. *lion tiger*). In associative priming, they are associatively related (e.g. *blood red*), which means that when people are asked to free associate (i.e. give the first word that comes to mind) to the prime, they often give the target. As with orthographic and phonological relationships, it is difficult to isolate semantic and associative relationships. However, some researchers have claimed to do so by using 'pure' semantically related items such as *goat lion* (when given *goat*, few people give *lion* as an associate) and 'pure' associatively related items such as *snow ball*. However, the concept of a 'pure' semantic relationship is questionable because mediating associations may exist for semantically related items (e.g. 'animal' for *goat* and *lion*). Semantic/associative priming effects are consistently positive for unmasked primes, but sometimes do not occur (or are even negative) for masked primes, even under conditions that yield large positive repetition priming effects.

There are two other less extensively investigated types of associative priming. In backward priming (e.g. *baby stork*), the prime (*baby*) is often given as an associate of the target (*stork*), but the target (*stork*) is rarely given as an associate of the prime (*baby*). In mediated priming, the prime and the target are neither semantically related nor directly associated in either direction, but are linked via their sharing an association with a third mediating word (e.g. *lion stripes*, with 'tiger' as the mediating associate). At long SOAs, positive backward priming occurs for lexical decisions but not for pronunciation, whereas positive mediated priming is easier to obtain for pronunciation than for lexical decisions.

Morphologically related primes and targets share the same base morpheme. A morpheme is the smallest possible unit of meaning. For example, *unpainted* contains the base morpheme *paint* plus two morphemic affixes (i.e. *un* (not) and *ed* (past participle)). When morphologically related primes and targets are also orthographically, phonologically and semantically related (e.g. *heavy* and *heavier*), it is difficult to know whether priming is based solely on shared morphology. However, morphological priming is not merely an orthographic, phonological or semantic priming effect, because priming occurs for irregular morphologically related primes and targets that are not orthographically or phonologically related (e.g. *buy bought*), even under conditions that eliminate semantic/associative priming. (*See* **Morphology**)

In the rarely studied syntactic priming effect, RTs are faster when the related prime and target form a syntactically well-formed pair (e.g. *angry cup*) than when they do not (e.g. *anger cup*).

FACILITATION AND INHIBITION AND THE TIME COURSE OF PRIMING

RTs are faster when *dog* is primed by *cat* rather than by *wall* either because *cat* facilitates the processing of *dog*, or because *wall* is instead (or also) inhibiting the processing of *dog*. To measure facilitation and inhibition, one needs a 'neutral' prime such as ******, *XXXXXX* or the word *ready*. If RTs for targets following a word prime are faster than RTs for targets following the neutral prime, facilitation has occurred. If they are slower, inhibition has occurred. Facilitation and inhibition effects have most often been examined for semantic/associative priming effects. Related primes typically produce facilitation, and at SOAs as short as 50 ms. However, inhibition from unrelated primes occurs only when most of the related primes and targets in the experiment are related but not strongly associated, and only at SOAs of 300 ms or longer. Only repetition and morphological priming effects consistently occur when up to 2–3 min and many unrelated words intervene between a single prime and the target. Although there have been a couple of reports of semantic/associative priming from multiple primes lasting that long, semantic/associative priming typically survives for only a few seconds when unrelated words intervene between a target and its single related prime.

AUTOMATIC AND STRATEGIC PRIMING

Priming is said to be automatic if it (1) occurs quickly (at SOAs of 250 ms or less), (2) yields facilitation from related primes but no inhibition from unrelated primes and (3) occurs unintentionally without the prime even being attended to or entering awareness. In contrast, strategic priming (1) occurs only at SOAs of 300 ms or longer, (2) often yields inhibition from unrelated primes as well as facilitation from related primes and (3) depends on the reader being aware of the prime and intentionally using it to aid the processing of related targets. Most experiments designed to separate automatic and strategic mechanisms have examined semantic/associative priming. (*See* **Automaticity**)

At SOAs of 250 ms or less, which do not allow the reader enough time to use the prime to invoke strategies, semantic/associative priming seems to be automatic for the following reasons.

1. As the proportion of related prime–target pairs increases, priming does not increase as it would if the person was becoming more likely to use the prime to invoke strategies to aid the processing of the more likely related targets.
2. Priming effects are due to facilitation from related primes, but not to inhibition from unrelated primes.
3. Even when word targets that follow a category-name prime such as *body* are highly likely to be words selected from the category *building parts*, the *body* prime produces facilitation for rarely occurring 'unexpected' related targets such as *arm*, but not for frequently occurring 'expected' but unrelated targets such as *door*.

At SOAs of 300 ms or longer, which provide enough time for strategic processes to be engaged, priming seems to be strategic for the following reasons.

1. Priming now increases as the proportion of related prime–target pairs increases.
2. Facilitation from related primes when most of the related primes are not strongly associated.
3. When building part targets are highly likely to follow the prime *body*, the *body* prime now produces inhibition rather than facilitation for 'unexpected' related targets such as *arm*, and facilitation for 'expected' but unrelated targets such as *door*.

Although the distinction between automatic and strategic priming is well accepted, there is controversy over whether automatic semantic priming occurs when people are truly unaware of the prime.

MODELS OF PRIMING

According to one popular account, automatic semantic priming is mediated by activation that spreads within a lexical network of interconnected memory nodes corresponding to the individual words that a person knows. A prime word's presentation automatically activates its node, and this activation spreads via associative links to preactivate the node of a semantically related target, making it easier to process when it is presented. The concept of spreading activation is used in many theories designed to explain a large number of cognitive phenomena in addition to priming. (*See* **Semantic Networks**; **Spreading-activation Networks**; **Lexicon, Computational Models of**)

Strategic semantic priming occurs when a person uses the prime to generate an expectancy set containing potential targets that are related to it. The target is first searched for in this expectancy set. If it is not found there, the complete search of lexical memory that always occurs following neutral primes is then initiated. If most of the related primes and targets are strongly related, the expectancy set is small. Thus facilitation from related primes is large because the target is found quickly in a small expectancy set, and inhibition from unrelated primes is small because the unsuccessful search of a small expectancy set only slightly delays the search of lexical memory. When most of the primes and targets are weakly related, the expectancy set is large. By the previous reasoning, this causes facilitation to be small and inhibition to be large.

Spreading-activation and expectancy accounts of semantic priming assume a search process that operates on individual nodes which correspond to each word that a reader knows. More recently, connectionist 'neural-net' models of lexical memory have been developed in which orthographic, phonological and semantic 'features' are connected to each other and to response features via other 'content-free features' (called hidden units) through excitatory and inhibitory connections. However, because these models have only very recently been applied to a very limited number of priming phenomena, their adequacy as models of priming remains unclear. (*See* **Connectionism**; **Distributed Representations**)

PRIMING AS A METHODOLOGICAL TOOL

Priming is also commonly used as a methodological tool. For example, because automatic

priming presumably occurs only for prime–target relationships that are directly represented in lexical memory, priming has been used to study the representational structure of bilingual lexical memory. It has also been used to study individual differences. For example, in some cases cognitive deficits due to normal aging, stroke, Alzheimer disease's and psychopathology seem to be due to a strategic processing deficit, while in other cases they appear to be due to a deterioration of the lexical network itself. Finally, priming has been used to examine differences in the brain's two hemispheres with regard to the types of information that they have directly represented in them and the types of strategic processing mechanisms that they control. (*See* **Language and Brain**; **Brain Asymmetry**; **Priming in Psychopathology**)

Further Reading

Chiarello C (1998) On codes of meaning and the meaning of codes: semantic access and retrieval within and between hemispheres. In: Beeman M and Chiarello C (eds) *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience*, pp. 141–160. Mahwah, NJ: Erlbaum.

Dalrymple-Alford EC and Marmurek HHC (1999) Semantic priming in fully recurrent network models of lexical knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition* **25**: 758–775.

Klinger MR, Burton PC and Pitts GS (2000) Mechanisms of unconscious priming. I. Response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory and Cognition* **26**: 441–455.

McNamara TP and Holbrook JB (2002) Semantic memory and priming. In: Healy AF and Proctor R (eds) *Comprehensive Handbook of Psychology. Vol. 4. Experimental Psychology*. New York, NY: John Wiley & Sons.

Masson MEJ (1999) Semantic priming in a recurrent network: comment on Dalrymple-Alford and Marmurek. *Journal of Experimental Psychology: Learning, Memory and Cognition* **25**: 776–794.

Neely JH (1991) Semantic priming effects in visual word recognition: a selective review of current findings and theories. In: Besner D and Humphreys GW (eds) *Basic Processes in Reading: Visual Word Recognition*, pp. 264–336. Hillsdale, NJ: Erlbaum.

Ober BA and Shenaut GK (1995) Semantic priming in Alzheimer's disease: meta-analysis and theoretical evaluation. In: Allen PA and Bashore TR (eds) *Age Differences in Word and Language Processing*, pp. 247–271. Amsterdam: Elsevier.

Plaut DC and Booth JR (2000) Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review* **107**: 786–823.

Problem Solving, Psychology of

Introductory article

Laura R Novick, Vanderbilt University, Nashville, Tennessee, USA

CONTENTS

Problem-solving is the process by which a person represents the goal to be accomplished, determines a procedure for reaching the goal, executes that procedure, and evaluates the success of the solution attempt.

INTRODUCTION

Much of human activity involves problem-solving. In school, the algebra teacher assigns the end-of-chapter problems for homework. On family game night, players must decide what move to make in chess, what word to write in Scrabble, or whether to build houses in Monopoly. In the world of work, a travel agent may need to coordinate plane flights for people living in two different states in the USA who wish to fly to Scotland together; a cognitive psychologist may want to design an experiment to study some aspect of problem-solving; and an architect may attempt to design an award-winning building. What all these examples have in common is that there is a goal to be achieved (solving the algebra problem, writing a high-scoring word, finding convenient and reasonably priced plane flights), and the solution that achieves that goal is not already known. If the solution can be retrieved from memory (e.g. if an adult is asked to multiply 5×7), there is no problem to be solved.

MODEL OF PROBLEM-SOLVING

Problem-solving consists of three general stages: understanding, production, and judgment. Each of these stages, in turn, involves several specific processes.

Understanding

The outcome of the understanding stage is a representation of the problem. This representation reflects the solver's understanding (correct or incorrect) of the information given and the goal to be achieved. The importance of constructing a good representation for successful problem-solving will be considered in a later section. The processes involved in understanding the problem include: (a) identifying the initial state of the problem; (b) identifying the solution criteria; (c) determining the constraints imposed on the solution attempt; and (d) comparing the current problem with problems encountered previously.

Consider the algebra word problem shown in Figure 1(a). Letting N denote the price of the notebook and P the price of the pencil, the initial state of the problem is: $N = 4P$, $P = N - 30$. In this domain, the solution criterion is straightforward: the correct

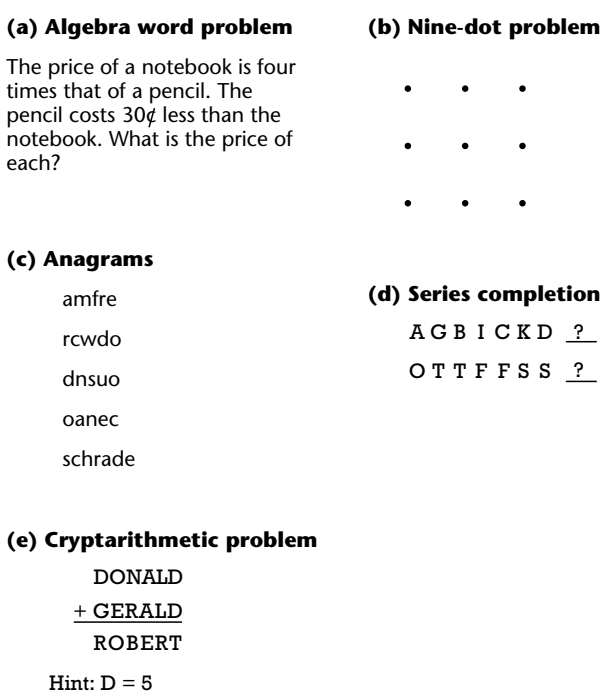


Figure 1. Examples of five types of problems.

answer has been found if both equations yield valid arithmetic statements when the derived values of N and P are substituted for the variables. Figure 3 at the end of the article gives the answer to this and all other problems in Figures 1 and 2.

To think about task constraints, it is useful to consider another example – the nine-dot problem illustrated in Figure 1(b). The goal is to connect the nine dots with exactly four straight lines without lifting one's pencil from the paper. Try to solve this problem before reading further.

The difficulty most people encounter when they try to solve this problem is that they assume that the lines cannot go outside the boundary of the imaginary box surrounding the dots. With this constraint, which is incorrectly imposed by the solver and is not part of the definition of the problem, the problem cannot be solved.

One's understanding and solution of a current problem may be guided by its similarity to a problem encountered previously. In such cases, psychologists talk about people solving problems by analogy. This topic will be considered in a later section.

Production

The outcome of this stage is a candidate solution to the problem. How the solution is generated depends on the type of problem to be solved. Two general strategies – search and analogy – will be considered in later sections. Broadly speaking, the solver applies operators (e.g. moving pegs, adding numbers) to the elements of the problem in order to transform or rearrange the information given into the solution or to induce the structure of the information given. For example, in an anagram (see Figure 1(c)), one must rearrange the letters to form a word in English (or another language). In a series completion problem (see Figure 1(d)), one must figure out the pattern in order to predict the next item in the sequence.

Often, generating a solution requires one to retrieve relevant facts and procedures from long-term memory. Consider the cryptarithmetic problem shown in Figure 1(e). The goal is to substitute a different number (0–9) for each letter to create a valid addition problem. A hint is given that the letter D has the value of 5. To solve this problem, one must retrieve addition facts (e.g. $5 + 5 = 10$) and procedures (e.g. only the one's digit of an answer is written below the line; the ten's digit is written at the top of the next column as a carry).

Generating a solution is more difficult, and the solution attempt more error-prone, when one must do all the work in one's head, in what cognitive psychologists refer to as working memory. It is harder to perform operations on the contents of working memory and maintain the outcomes of prior operations in working memory than to do the same on paper. To illustrate this, imagine being given a 'running' arithmetic problem in which you must perform the arithmetic operations in the order they are stated. For the first problem, $8 + 16 \times 3$, you are allowed to use pencil and paper. So, you add 8 and 16 and write down the answer, 24. Then, again on paper, you multiply 24 by 3 to get 72. For the second problem, $5 + 18 \times 4$, you must perform all the computations in your head (i.e. in working memory). Clearly, the second problem is harder, and you are more likely to get it wrong.

Many times, solvers store in long-term memory the procedure used to solve a problem. This information may enable them to use the current solution attempt to help solve related problems in the future, through analogical problem-solving.

Judgment

The outcome of this final stage is a decision that the problem has been solved or that more work is needed. Two processes are involved. One is to choose a rule for determining that a sufficient match exists between the candidate solution produced in the production stage and the goal specified in the understanding stage. Often this is fairly trivial. For anagrams, for example, the rule is that the anagram has been solved if the letter rearrangement the solver generated matches a word in his or her mental dictionary. For some problems, however, it is much more difficult to choose a decision rule. This will be the case when the goal is ill-defined – for example, to design an award-winning building – or when an optimum solution is not possible – for example, if one has a limited amount of money and cannot afford all of the car features one wants.

The second process is to compare the candidate solution to the solution criteria. Continuing with the anagram example, one compares the generated solution to words in one's mental dictionary. If a match is obtained, the anagram is solved. Otherwise, one tries to find a different arrangement of the letters that in fact is a word.

REPRESENTATION

Problem representations depict a solver's understanding of the structure of the problem to be

solved. They can be distinguished in a variety of ways. They can be constructed and stored internally in working memory, or they can be constructed and stored externally – for example, as a drawing on paper. As noted earlier, problem-solving is facilitated when representations can be stored and manipulated externally. A second important distinction is that representations may resemble their physical referents (e.g. a drawing of a pulley system is similar in appearance to the object it represents), or they may be more abstract and symbolic (e.g. a hierarchy diagram depicting the sequence of games in a basketball tournament looks nothing like the tournament). Finally, representations may be diagrammatic/pictorial or verbal.

It is often possible to represent the information in a problem in multiple ways. Consider the problem stated in Figure 2(a). Figures 2(b)–2(d) provide examples of three types of external representations that solvers might construct for this problem: (b) verbal shorthand (partially complete); (c) Venn diagram; and (d) network (partially complete).

(a) Problem statement:

A third-grade teacher asked her students to raise their hands if they collect rocks. She repeated the question for shells and for key chains. The following students raised their hands for each question:
Rocks: Harry, Francesca, Aaron, Joseph, Marcus, Elana, Quentin, Naomi, Rebecca, Stephanie, William, Derek, Gillian, Courtney.
Shells: William, Oliver, Kayla, Gillian, Elana, Aaron, Teondria, Stephanie.
Key chains: Victoria, Courtney, Leah, Quentin, Francesca, Elana, Marcus, Stephanie, Oliver, Benjamin
Three students never raised their hand: Isaac, Ursula, Pavel.
What percentage of the class collects rock and shells but not key chains?

(b) Verbal representation

rocks only: Harry, Joseph, Naomi, Rebecca, Derek
shells only: Kayla, Teondria
key chains only: Victoria, Leah, Benjamin
rocks & shells only: Aaron, William, Gillian
etc.

(c) Venn diagram representation (d) Network representation

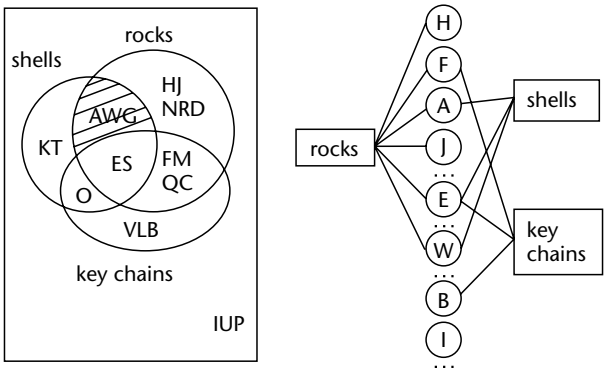


Figure 2. An illustration of different types of representations constructed for a single problem.

Many studies have shown that the type of representation constructed for a problem affects the ease or likelihood of solution: performance is facilitated when the structure of the representation corresponds well with the structure of the problem. Thus, a critical aspect of successful problem-solving is selecting, and then constructing, the most appropriate type of representation for the problem at hand. Often, solvers devote insufficient attention to this step, jumping straight from reading the problem to trying to get the answer – that is, to the production stage.

According to folk wisdom, visual representations are superior to verbal ones. As the oft-repeated Chinese proverb tells us, a picture is worth ten thousand words. Research supports the folk wisdom, as many studies have documented the benefits of pictures and diagrams for successful problem-solving. Abstract diagrams, in particular, are important instruments in the toolbox for thought. Such diagrams include, among others, Venn diagrams, hierarchies, matrices, and networks (sometimes called path diagrams).

PROBLEM SPACE, SEARCH, MEANS–ENDS ANALYSIS

Allen Newell and Herbert Simon referred to the solver's internal representation of a problem (and related context) as a problem space. In their view, a problem space has four major components: (a) an initial state of knowledge, specifying what the solver knows about the problem before trying to solve it; (b) a goal state, specifying the final state of knowledge to be reached; (c) operators, which produce new knowledge states when applied to existing knowledge states; and (d) a set of intermediate knowledge states, which result from applying the operators. Given this type of representation, solving a problem involves searching for a path through the problem space that connects the initial state to the goal state, going through some sequence of intermediate states. (*See* **Newell, Allen**; **Simon, Herbert A.**)

Consider the Tower of Hanoi problem. There are three pegs sticking up from a board. On the leftmost peg are three disks of differing sizes, with the largest disk on the bottom and the smallest disk on top. The goal is to move this tower to the rightmost peg. There are two constraints on the moves: (a) only one disk can be moved at a time, and (b) a larger disk cannot be placed on top of a smaller

disk. There are 27 different ways the three disks can be arranged on the three pegs. Solving this problem can be conceptualized as searching for a path through these 27 knowledge states that connects the initial state to the goal state. Try to find the shortest solution path, which involves only seven moves, before reading further.

Solvers may use any of a variety of search strategies to decide what operator to apply next. The simplest, but least effective, strategy is to pick an operator randomly. A more sophisticated strategy is known as hill-climbing, in which, for each move, the solver picks the operator that generates a new knowledge state that most closely resembles the goal state. This strategy is better than random guessing, because it allows the solver to look one move ahead. Often in problem-solving, however, one must backtrack a little to move forward, and in these cases hill-climbing will not lead to successful solution. An even more sophisticated strategy is means–ends analysis with subgoals.

The basic idea with means–ends analysis is to compare the current state to the goal state in order to select an operator that will reduce the difference between them. A critical aspect of this strategy is the setting of subgoals. Often, the solver is blocked from applying the best operator (e.g. because of the constraints of the problem). In that case, the solver sets a subgoal to remove the block. The current state is then compared to the subgoal state, and an operator is selected to reduce the difference between them. Additional subgoals may need to be set before an appropriate operator can be applied. (*See* **Means–Ends Analysis**)

Consider again the Tower of Hanoi problem. Because the largest disk must be placed on the bottom of the rightmost peg to build the goal tower, the best operator to apply is to move that disk from the leftmost peg to the rightmost peg. However, the two disks resting on top of that disk prevent that course of action. Therefore, the solver must set a subgoal to move the top, two-disk tower to the middle peg. To satisfy this subgoal, the solver must move the medium-sized disk to the middle peg. But, the smallest disk is in the way and prevents application of this operator. So another subgoal is set to move that disk out of the way. This subgoal can be satisfied directly, by moving the smallest disk to the rightmost peg, which then enables the remaining subgoals to be satisfied in reverse order. Continuing in this manner, the problem is easily solved.

ANALOGY

General-purpose strategies such as hill-climbing and means–ends analysis work well when solvers do not have specific, domain-relevant knowledge to guide their solution attempt. Often, however, solvers do have more specific knowledge available. This may be because they have solved a similar problem in the past or because they have acquired some expertise in the domain. Analogical problem-solving will be considered in this section and expertise in the next.

Analogical problem-solving involves four subprocesses: retrieval, mapping, inference, and adaptation. First, one must retrieve a relevant example problem. Solvers usually select a problem whose story line is similar to that of the current, or target, problem. For example, algebra students given Target Problem 1 in the middle of Table 1 for homework might look for a related example problem encountered earlier in the chapter, such as the first problem in the table. (Solvers who are relatively expert in a domain can often identify analogous problems without such overt similarity between them.) Then, the mapping process identifies correspondences between elements in the two problems. For example, 'washing the windows' maps onto 'mowing the lawn', '3.75 hours' maps onto '42 minutes', and 'Max' maps onto 'Sarah'.

The solver uses the mapping, along with links connecting elements of the example problem to elements of that problem's solution, to generate inferences concerning the solution to the target problem. For example, the fact that one multiplies the reciprocal of 42 times the unknown number of minutes worked in the example problem suggests that one should similarly multiply the reciprocal of 3.75 times the unknown number of hours worked in the target problem. If the two problems are structurally identical (i.e. isomorphic), like the example problem and Target Problem 1, these inferences should enable solution of the target problem. But if the two problems are only similar, the adaptation process will be needed to modify some of the inferences or to generate new inferences to account for the unique aspects of the target problem's structure. For example, two adaptations are needed to solve Target Problem 2 in Table 1 by analogy to the example problem. First, the right-hand side of the equation should be 1/2 rather than 1, because the couple works together to complete only half the job. Second, distinctive labels must be used to represent the amount of time worked by

Table 1. Algebra word problems to illustrate analogical problem-solving

Example problem with solution

It takes Sarah 42 minutes to mow the lawn, whereas her younger brother, Adam, requires 63 minutes to do that same job. How long will it take to mow the lawn if the two children work together?

$$\left(\frac{1}{42}\right)X + \left(\frac{1}{63}\right)X = 1 \rightarrow \left(\frac{3}{126}\right)X + \left(\frac{2}{126}\right)X = 1$$

$$\left(\frac{5}{126}\right)X = 1 \rightarrow 5\,X = 126 \rightarrow X = 25.2 \text{ minutes}$$

Target Problem 1 – isomorphic to example problem

It takes Max 3.75 hours to wash the windows of his house. His wife, Jessica, needs 4.25 hours to do that same job. How long will it take to wash the windows if the two of them work together?

Target Problem 2 – similar to example problem

It takes Max 3.75 hours to wash the windows of his house. His wife, Jessica, needs 4.25 hours to do that same job. Jessica washes half the windows on Saturday. On Sunday, they work together to complete the job, but Jessica starts half an hour after Max. How long will each person work on Sunday?

Max and by his wife (e.g. 'X' and 'X − 0.5', respectively), because they do not work for the same amount of time.

EXPERTISE

When cognitive psychologists moved from studying problem-solving using puzzles such as the Tower of Hanoi and those illustrated in Figures 1(b)–1(e) to using subject-matter domains such as mathematics and physics (e.g. Table 1), the importance of considering domain expertise became clear. Research on expertise in many domains (e.g. physics, mathematics, computer programming, X-ray diagnosis, chess, basketball) has identified several important differences between experts and novices that have an impact on problem-solving.

Perhaps most importantly, experts and novices in a domain differ with respect to the types of features they focus on in their representations of problems in that domain. Novices tend to focus on what are referred to as superficial features of the domain, such as the particular objects and terms mentioned in the problem and the way the question happens to be phrased. Thus, for example, when asked to sort a set of mechanics (physics) problems according to how they would be solved, undergraduates who have done well in an elementary physics course sort by the types of objects described, making separate piles for spring problems, pulley problems, inclined plane problems, etc. In contrast, experts focus on the underlying structure of the problems, in particular, how the objects are related to each other, regardless of what the objects are. Given the same sorting task, physics graduate students group the problems according to the physics principle required for solution – e.g. conservation of energy or Newton's second law – even if that means putting a spring problem, a pulley problem, and an inclined plane problem together. Thus, acquiring expertise in a domain involves learning the important distinguishing features of problems in that domain.

This representational difference has important ramifications for problem-solving. On the one hand, because the structure of a problem is what determines how to solve it, experts are more successful than novices at devising solution procedures from scratch. They are also more successful than novices at analogical problem-sol ving, because they are: (a) more likely to retrie ve a structurally similar (i.e. relevant) example problem; (b) more successful at constructing the mapping between the target and example problems; and (c) more successful at adapting the example problem's solution procedure, if necessary, to fit the unique requirements of the target problem. On the other hand, experts' solution attempts are less affected than those of novices by superficial characteristics of problems. For example, although both highly skilled and less skilled anagram solvers require more time to solve pronounceable than unpronounceable anagrams (e.g. *aftin* versus *infta, koech* versus *oekch, tarms* versus *rtmsa*), the negative impact of this superficial characteristic is much greater for novices (the solutions are 'faint', 'choke', and 'smart', respectively).

Experts also have better memory for, and can more quickly recognize, meaningful patterns of information within their domain of expertise. Finally,

experts and novices often differ in how they approach the task of problem-solving. Novices tend to reason backwards from the desired goal (e.g. the conclusion to be proved in a geometry problem or the variable to be solved for in a physics problem), often using a general-purpose strategy such as means–ends analysis. In contrast, experts tend to reason forwards from the information given (i.e. the initial state), based on their knowledge of the structural relations among important concepts in the domain, relations that have been learned through prior experience reasoning backwards. It makes sense, therefore, that experts revert to reasoning backwards when they encounter novel problems.

The transition from novice to expert is gradual. Problem-solvers at intermediate levels of expertise typically show patterns of performance that are intermediate between those of novices and experts.

INSIGHT

The discussion so far has considered problem-solving to be a deliberate, effortful, and consciously accessible means for proceeding incrementally from the initial state to the goal state. Although this characterization captures people's intuitions concerning many of their problem-solving attempts (e.g. solving the algebra word problems in Table 1), there are exceptions. Sometimes, people have the impression that a solution has simply popped into mind suddenly, seemingly out of nowhere, without any accompanying awareness of having made incremental progress towards the goal. The phenomenological experience in such cases is one of 'aha!' Perhaps you generated such a solution to one of the anagrams in Figure 1(c). Problems that often lead to 'aha' solutions (e.g. anagrams), or that lead to such solutions after the solver has restructured the problem in an appropriate way (e.g. the nine-dot problem in Figure 1(b)), have been referred to as *insight* problems.

Psychologists have long been interested in whether insight (sometimes called 'pop-out') solutions are fundamentally the same as or qualitatively different from noninsight solutions. The Gestalt psychologists explained the sudden emergence of the insightful solution and the accompanying 'aha!' sensation by the sudden reorganization of one's understanding of the problem. For example, they suggested that solution of the nine-dot problem could be obtained easily if one's understanding of the problem were restructured to allow the lines to go outside the boundary of the dots.

More generally, there are four components to the Gestalt view of insight problem-solving: (a) solution to the problem is prevented initially by fixation on an inappropriate organization (i.e. representation) of the problem; (b) removal of the fixation leads to a sudden restructuring of the problem; (c) the new (restructured) representation leads to the sudden emergence of the solution; (d) the restructuring process is different in kind from the type of process used to solve noninsight (e.g. search) problems. That is, what happens outside of conscious awareness mimics solvers' conscious experience.

Modern views have questioned whether restructuring necessarily leads to immediate solution, that is, to insight. For example, for the nine-dot problem, even after realizing that the lines may extend outside the boundary of the dots, one must still search in the new problem space for a solution, much like one searches for a solution to noninsight problems (e.g. the Tower of Hanoi). Modern research has also called into question the Gestalt claim that the restructuring arises full-blown. Although the solution may pop into mind suddenly, solution-relevant information appears to accumulate gradually outside of awareness. For example, highly skilled anagram solvers can reliably determine whether a string of five letters can be unscrambled to form an English word (e.g. 'dnsuo', 'injto', and 'amfre' can, but 'ynsil', 'ocjeu', and 'cofen' cannot) within half a second, and their performance on this solvability judgment task improves as the time available to make a decision increases, from half a second to a second, even when the available time is too short to enable them to unscramble the solvable items (i.e. the anagrams).

The results of recent research suggest that although both pop-out and search solutions rely on the gradual accumulation of partial information, there may be some merit to the Gestalt claim that pop-out solutions are different in kind from search solutions: pop-out solutions may result from the solver attempting to satisfy the many, sometimes conflicting, constraints on problem solution *in parallel* (i.e. simultaneously) and outside of awareness; in contrast, search solutions appear to result from trying to satisfy the same constraints *sequentially* (i.e. one after the other) and consciously.

Recent research also suggests that there may be a link between insight solutions and expertise: at least in some domains (e.g. anagrams), pop-out solutions are much more common among experts. And there may be a transition from primarily sequential to greater parallel processing with increasing expertise.

CONCLUSION

Problem-solving consists of three general stages: (a) understanding, which results in a representation of the problem; (b) production, which results in a candidate solution; and (c) judgment, which results in a decision as to whether the problem has been solved or more work is needed. Understanding is arguably the most important stage, because the type of representation constructed affects the ease or likelihood of solution. Among external representations, visual ones are often superior to verbal ones. Problem representations held in memory are sometimes referred to as problem spaces. Problem-solving is often conceptualized as a process of searching through a problem space for a path connecting the initial state to the goal state. An important search strategy is means–ends analysis with subgoals. Sometimes, solvers attempt to solve a current problem by making an analogy to a structurally similar problem encountered previously. Experts in a domain have a problem-solving advantage over novices because (a) their problem representations highlight solution-relevant structural features rather than solution-irrelevant superficial features, and (b) they more quickly recognize meaningful patterns within their domain of expertise. Sometimes, problem solutions seem to pop into mind suddenly, without any accompanying awareness of the process by which they were generated. Current research suggests that such insight solutions, like noninsight solutions, are preceded by the gradual accumulation of partial information relevant to solution.

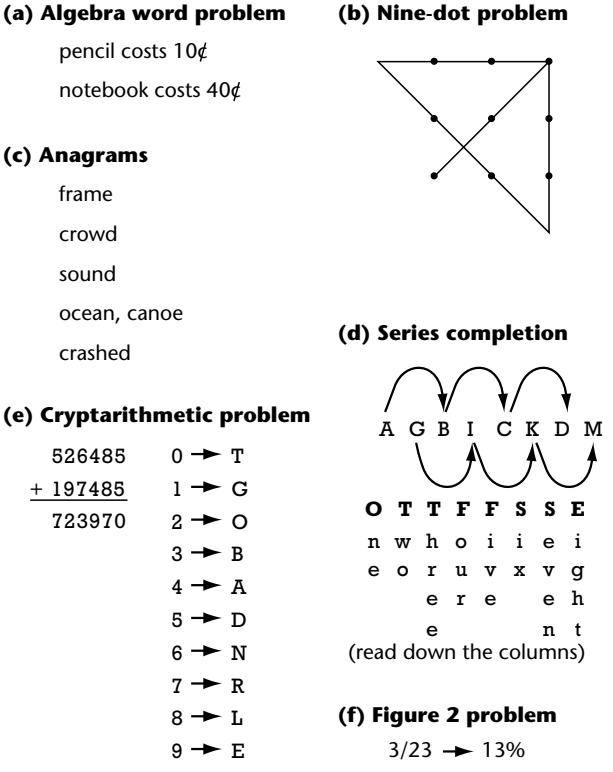


Figure 3. Answers to the problems shown in Figures 1 and 2.

Further Reading

Day RS (1988) Alternative representations. In: Bower GH (ed.) *The Psychology of Learning and Motivation*, vol. 22, pp. 261–305. San Diego, CA: Academic Press.

Duncker K (1945) On problem-solving. *Psychological Monographs* **58** (Whole No. 270).

Durso FT, Rea CB and Dayton T (1994) Graph-theoretic confirmation of restructuring during insight. *Psychological Science* **5**: 94–98.

Ericsson KA and Smith J (eds) (1991) *Toward a General Theory of Expertise*. Cambridge, UK: Cambridge University Press.

Greeno JG and Simon HA (1988) Problem solving and reasoning. In: Atkinson RC, Herrnstein RJ, Lindzey G and Luce RD (eds) *Stevens' Handbook of Experimental Psychology*, 2nd edn, vol. 2, pp. 589–672. New York, NY: Wiley.

Holyoak KJ and Thagard P (1995) *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Simon HA (1978) *Information-processing theory of human problem solving*. In: Estes WK (ed.) *Handbook of Learning and Cognitive Processes*, vol. 5, pp. 271–295. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg RJ and Frensch PA (eds) (1991) *Complex Problem Solving: Principles and Mechanisms*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wertheimer M (1996) *Productive Thinking* (enlarged edn). Chicago, IL: University of Chicago Press. [Original work published 1959.]

Prototype Representations

Intermediate article

James A Hampton, City University, London, UK

CONTENTS

Prototype representations represent classes in terms of their central or most typical example – the prototype – rather than in terms of an explicit definition of the class boundary. Classification is based on whether similarity to the prototype is above some threshold criterion value.

INTRODUCTION

Psychological models of conceptual knowledge and knowledge of word meaning both require an internal representation of the world in the mind. Models or systems of representation vary in their level of complexity. At the simplest level are exemplar storage systems, in which individual experiences of objects are each stored, and categorization of novel instances proceeds by comparison to previously encountered exemplars. At the most complex level are systems for representing complex knowledge incorporating causal explanations of how individual attributes of an object class relate to each other. Prototype representations lie in the middle of this complexity dimension. The central notion is that we abstract a generic representation or prototype of a class (such as *fruit*) from our experience with many examples. This representation may be more or less structured, but does not contain information about specific individuals. Deciding that a novel instance is of a particular type involves a decision about how closely it matches the prototype for the class.

This form of representation gives rise to two phenomena. First, the membership of the class is graded in terms of typicality: those instances that match the prototype well (such as an apple) will be considered more representative or typical than those that match poorly (such as a coconut). Second, because the borderline of the class is not well specified, there may be instances which are neither clearly in, nor clearly out of the class (such as tomatoes or olives). This vagueness in the application of the category is a key aspect of prototype representations since it maps onto the acknowledged vagueness of the meaning of most nouns and verbs in natural language.

HISTORICAL DEVELOPMENTS

Early theories of semantic and conceptual representation tended to assume that a conceptual category was defined in terms of a set of singly necessary and jointly sufficient defining features – a view known as the Classical View of concept structure. It can be seen in analyses of the meaning of such words as 'Uncle' or 'Cousin', which can be defined in terms of a small set of dimensions.

One of the first appearances of the notion of prototype was a classic study by Posner and Keele (1968). Their study investigated the learning of novel stimulus classes through trial and error. Rather than investigating the learning of different forms of definitional rule, Posner and Keele taught people to differentiate stimulus classes that were based on small distortions of a random array of dots. In general terms a prototype can be thought of as a point in 'stimulus space'. Each exemplar of a category can be described in terms of its position along a number of orthogonal dimensions (for example its size, orientation, color, etc.). By plotting the dimensions in a multidimensional space, the exemplars can be placed in the space according to their coordinates on each dimension, and distance in the space can then be mapped onto the similarity between exemplars. This spatial representation is the stimulus space. The prototype is then defined as the centre of gravity of the set of exemplars in the space – the (possible) exemplar that has maximum overall similarity to all of the surrounding exemplars.

(Where a stimulus is not describable in terms of continuous or dichotomous dimensions, then this spatial model breaks down, and other ways of defining the prototype are used, such as taking the prototype to be the possible exemplar that has the most commonly occurring values on each feature or dimension.)

Prototype representations have in common that they are limited to representing classes that are linearly discriminable. That is to say that it is possible to discriminate between members and nonmembers in terms of some simple additive combination of the features or dimensions. Probability of an item being categorized in a prototype class should be expressible as a positive function of its similarity to the prototype and as a negative function of its similarity to the prototypes of other contrasting categories. This constraint has been used to argue that natural biological categories are not represented by prototypes. For example, it is claimed, a creature such as a whale is more similar to the prototype for fish than is a creature such as a seahorse, and yet while the whale is not a fish, the seahorse is.

The primary development of prototype theory is largely due to the work of Eleanor Rosch and Carolyn Mervis in the 1970s. Rosch (working initially as Heider) first developed the notion of 'natural' prototypes in the context of color and geometric figures. She showed that in the color spectrum there were certain hues that were naturally considered 'good examples' of a color term such as 'red', and others that were atypical. In her work with a primitive society – the Dani of New Guinea – Rosch was able to show that even though the group had no language term for 'red', they found color categories based on a 'good' red easier to learn than color categories centred around a 'poor' or atypical red. (More recent research has questioned this result.) According to Rosch, certain visual forms and colors form natural (universal) cognitive reference points or prototypes. Work in comparative linguistics confirms this notion. Whereas the boundaries between color terms vary widely across languages, the choice of the most central hue for a color term shows much closer agreement.

FAMILY RESEMBLANCE STRUCTURES: WITTGENSTEIN AND ROSCH

In subsequent work with Mervis, Rosch extended the notion of prototype concept to include more complex natural categories such as biological and artifact kinds (birds, fish, fruits, tools, furniture etc.). In developing the notion, they referred to a philosophical analysis given by Wittgenstein (1953) in his *Philosophical Investigations*. Wittgenstein spent many years worrying over the relation between word meaning and the underlying logic of thought and language. Towards the end of his life he came to the view that words do not correspond directly to the logical terms of propositions, but that the meaning of a word is defined instead in terms of the complex pattern of use that it has in language. He pointed out, for example, that one cannot specify just what all the activities that we call 'game' have in common, since for any possible relevant defining feature (e.g. 'games are all competitive') clear counterexamples could be found (ring-a-rosy is a children's game that is not competitive). Games, he said, are like members of a family. There are clear resemblances amongst the members of a family, but there may be no single distinguishing feature that would pick out the whole family from everyone else.

Rosch and Mervis (1975) took this notion and developed it into the Prototype Theory of Concepts (PTC). According to PTC, a concept such as 'game' is defined in terms of a number of features, such as *competitive* or *has teams*, and membership in the class of games involves possessing a sufficient number of these features. Their theory was equivalent to a classification system known as *polythetic* classification in theoretical taxonomy. The theory proposes that because different features or dimensions of objects are correlated within a domain, objects naturally fall into similarity clusters. For example, the domain of creatures having a beak is correlated with laying eggs and flying, while having a scaly skin is correlated with having cold blood and teeth. Prototype representations capitalize on these intercorrelations amongst features by drawing clusters of intercorrelated features together into a prototype of, say, birds or reptiles. However, because the correlations are imperfect, knowing that an object is in a particular class would not tell you just which set of the prototypical features it would possess.

Evidence for their theory was obtained as follows. They presented subjects with members of categories such as birds or fruits, and had them list properties or features of those members. Other subjects then decided the extent to which each member possessed each feature. A tally was computed of the degree to which each member shared features with each other member. This 'family resemblance score' was shown to correlate well with independent judgments of the typicality of each member. Using a more direct method of

interrogating subjects about the features that they thought relevant to defining each category, Hampton confirmed that the number of shared features predicts typicality. He also showed that it could be used with some degree of accuracy to determine whether an item was considered to be a category member or not, and the speed with which the decision was made (Hampton, 1979).

The novel feature of PTC as applied to word meanings is that, unlike traditional lexical semantic analyses of meaning, a feature may be part of the 'definition' of a term even though it is not true of all the things covered by the term. Analyses of prototype effects in any domain typically involve showing four types of effect: (a) there is no explicit definition to be given in terms of a conjunction of defining features; (b) features (e.g. that birds can fly) are listed as important to the concept's meaning even though they are not in fact common to all members of the category; (c) there are clear differences in the 'representativeness' or typicality of different category members; and (d) there are differences in the degree to which different items may actually be considered to belong in the category – that is to say, the category borderline is vague or fuzzy.

Rosch, Mervis, and others extended the exploration of prototype effects into many different areas. Prototype effects have been found in the learning of novel categories based on prototypes, the speed and accuracy of categorization, the strength of inductive inferences based on a typical as opposed to an atypical category member, and the differential build-up and release from proactive interference in short-term memory studies. An example of the predictive power of PTC is a study of categorization by Hampton (1982). One consequence of defining category membership in terms of a 'sufficient number' of the prototype features is that one may find a hierarchy of classes (A is a kind of B; B is a kind of C) for which the transitive inference (A is a kind of C) does not hold. Hampton confirmed the existence of such sets. For example 'car-seat' was categorized as a chair, and 'chair' was categorized as furniture, but 'car-seat' was not considered to be furniture.

Another closely related example of the use of prototype representations in reasoning is the conjunction fallacy, in which people erroneously use similarity to a prototype to estimate probabilities of class membership.

PTC as proposed by Rosch is not in itself a psychological model of concept representation, but is more a way of drawing together a large set of phenomena. More precise models for representing and learning prototype concepts have since been developed. While PTC has undoubtedly had a great influence on the understanding of concepts and word meaning, it is important to note that few theorists still regard it as adequate as a basis for conceptual representation. Murphy and Medin (1985) developed a critique of the theory from the point of view of the lack of constraints that it provides on the notion of feature or similarity. They argue that one cannot define similarity without making prior assumptions about the relevant dimensions of difference amongst a set of stimuli, and that these assumptions come from a much more sophisticated understanding of the domain in question. Concepts have to be seen in the context of a broader domain 'theory' in which they play an explanatory role. This view, sometimes known as the 'theory theory', has been widely endorsed by developmentalists studying children's concepts (e.g. Carey, 1985; Keil, 1989).

IMPORTANCE OF PROTOTYPE THEORY FOR INVESTIGATIONS INTO WORD MEANING

Evidence for prototype effects in word use and word meaning is easy to find. In addition to the intransitivity of categorization, the perceived strength of inductive arguments is also affected by typicality differences. Even when told that all members of a broad category have some property, people feel more confident about concluding that a typical subclass would have it, than that an atypical subclass would. People are aware that the more typical an example, the more likely it is to possess all the prototypical features of the class.

Demonstrations of the vagueness of category borderlines are particularly important in studies of word meaning. It can be argued that all natural language terms (that is, excluding terms in axiomatically defined systems such as mathematics) are vague to some degree. Judgments on whether to call a man tall, whether to call a geological formation a mountain, or whether to call a particular organism a dog, may all be a matter of debate. There is some evidence, however, that with biological kinds as opposed to artifacts or social categories, people are less willing to agree that the boundaries of the class may be vague, and instead assume that there is some essential constitutive property, known perhaps only to experts. For example, people may assume that 'arthritis' refers to some well-defined condition with a unique identifiable cause, whereas in fact it simply means inflammation of the joints.

A simple way to account for the ubiquitous vagueness of natural language terms is through Wittgenstein's basic insight that, as Rosch puts it, we can 'judge how clear a case something is and deal with categories on the basis of clear cases in the total absence of information about boundaries' (Rosch, 1978). Prototype representations can be held independently of a rule for determining the category boundary.

Take an example such as 'murder'. Most people's understanding of the meaning of this word in English is based on prototypical examples – the classical murder mystery type of murder in which one individual deliberately and intentionally kills another through their own direct physical action, with some clear motive such as revenge, jealousy, or personal financial gain. A prototype analysis of the concept would involve gathering a list of all such features from a sample of English speakers and putting them together to generate a prototypical or paradigm case. From this prototype one can then invent different possible scenarios in which one or more of the standard features are missing. These might include cases where the killing occurs through failing to act, cases where the victim may be considered a borderline case of being a person (as in abortion), or cases where the motives are not self-serving (as in euthanasia). The doubt and debate that is engendered in this series of 'moral dilemmas' are evidence of the multidimensional nature of the concept itself and the vagueness that results from representing clear cases, but not clearly representing the class boundary.

The problems of vagueness in language use provide a good argument for a clear differentiation between concepts and word meanings. Osherson and Smith (1981) and a number of philosophers have argued that our ability to understand and to use logic in our thinking and speaking is itself evidence that concepts cannot be vague prototypes. Osherson and Smith, for example, argue that there is no workable logic for handling logical combinations of vague concept terms such as 'striped apple' or 'pet fish'. Typicality in these complex concepts is not a simple function of typicality in the original sets, since a guppy or a goldfish may be a clear example of a pet fish, but atypical as either a pet or a fish. Work on conceptual combination arising from Osherson and Smith's paper has shown that the combination of prototypes does obey some constraints, although there is also evidence that broader world knowledge is involved in determining the meaning of complex noun phrases. The difficulty is that if concepts are to be components of thoughts, then the way in which they combine should follow the simple rules of logic. Since prototypes are mostly noncompositional, it is argued that they cannot serve the necessary function of being the building blocks for a compositional theory of thought.

Another important critique of PTC was offered by Armstrong *et al*. (1983), who asked subjects to give typicality ratings to exemplars of well-defined categories such as 'even number'. The degree of inter-subject agreement on the typicality of numbers such as 2, 18, or 574 was as great as that for the typicality of different fruits or different items of furniture. From this result they argued that typicality effects *per se* were not strong enough evidence to support the conclusion that a concept had a prototype representation. At the very least, their demonstration makes the point that determining the membership of a class need not involve the same information as judging what is typical of that class.

Prototype theory has been applied to the analysis of a wide range of semantic domains. Labov demonstrated how the use of simple terms such as 'cup' and 'bowl' could be mapped into a stimulus space involving dimensions of size, shape, and use of containers (Labor, 1973). Cantor and colleagues have applied the analysis with success to person perception, personality traits, psychological situations, and psychiatric diagnosis (Cantor and Mischel, 1977). Coleman and Kay showed graded structure in the types of speech act that would be categorized as 'lying' (Coleman and Kay, 1981), and Hampton applied it to abstract terms such as 'Art' and 'Science', although interestingly some other terms like 'Rule' and 'Instinct' did not show prototype structure (Hampton, 1981).

PROTOTYPES IN LINGUISTICS, AND RADIAL CATEGORIES IN WORD MEANING

Cognitive linguistics has embraced the notion of prototypes for analysis of word usage and word meaning. Most notable has been the work of George Lakoff, whose book *Women, Fire, and Dangerous Things* (Lakoff, 1987) presents a detailed account of prototype effects, which are viewed as reflecting the underlying idealized cognitive models that we use to represent the world. According to Lakoff, prototype effects can be identified not only in lexical meaning, but also in phonology, morphology, and syntax. For example, even syntactic classes such as noun may be based around prototypes. Some nouns (typically referring to concrete objects) show a wider range of

allowable syntactic manipulations than do others. Membership in the category appears to be graded – some words are more 'nouny' than others.

Lakoff distinguishes a number of different sources of prototype effects in lexical semantics. One case is where a number of related cognitive models cluster within the same domain. For example, the concept 'mother' really involves a cluster of concepts including giving birth, being the genetic parent, nurturing the child, and playing the relevant family role. Surrogate, adoptive, foster, and egg-donor mothers fit one or another of these models. Prototype effects result from the different applicability of the cluster of models to different cases.

Lakoff takes the analysis further with his discussion of 'radial categories'. These are clusters of meaning in which the application of a term has come to be extended to a range of other cases through a nonarbitrary but yet unpredictable process of chaining. For example, in Japanese, there is a noun classifier 'hon' which is used most commonly to describe long thin objects. By radial chaining it is also used for associated nouns such as martial arts contests using staffs or swords, hits in baseball, telephone calls through long thin wires, injections using long thin needles, and so forth. According to Lakoff the best explanation for the diverse range of nouns that take 'hon' is in terms of chains of association in which a central case (e.g. a long thin staff) becomes extended to a secondary case (e.g. a martial arts contest) and from there to a case at third remove such as a judo contest (similar to a martial arts contest, but now lacking the long thin staff). The process of chaining has also been shown to affect categorization in the learning of novel categories.

One consequence of radial category structure is that the original notion of representing a concept with a single prototype and an allowable degree of distortion is no longer adequate. Because the radial extensions take the meaning along chained paths in unpredictable ways, one can no longer expect the application of a term to occupy a linearly discriminable region of the stimulus space. A good demonstration of the complexity of radial categories is Malt's analysis of the use of the word 'water'. Malt asked a group of students to judge the extent to which a range of liquids such as lemonade, dishwater, rain, etc. contained H_2O. There was surprisingly little correlation between the estimated proportion of H_2O and the appropriateness of the term 'water' for referring to the liquid. Use of the term was affected by a number of other factors such as how the liquid was used, and whether other more appropriate labels for it existed (Malt, 1994).

As another example of prototype effects in language, Lakoff also reports an extensive analysis by Brugman of the different uses in English of the spatial preposition 'over'. Although at first sight a word with only one meaning, a careful analysis reveals a host of different but related senses – moving above and across (jumping over), being above (hovering over), covering up (painting over), to name but three. It is argued that the way in which word meanings form an interconnected network of related senses shows a particular form of prototype structure that is endemic to natural languages.

ACQUISITION OF PROTOTYPE CATEGORIES

The acquisition of prototype categories has been studied in two different ways. First there have been extensive studies of how adults (and children) learn novel categories based around a prototype structure. The domains used include a wide range of materials such as stick figures, random shapes, simple geometric figures, lists of disease symptoms, random strings of letters, or schematic scenes. Prototype models make the specific prediction that even if not presented in training, the prototype stimulus itself will always be at least as fast and as accurately classified as other category members. The main competitor to prototype theory for category learning is the Generalized Context Model (GCM) which assumes that individual exemplars are stored without any abstraction of the prototype. Across a range of experiments, the GCM has been developed into a highly successful predictor of a range of results, and frequently outperforms prototype models in predicting behavior. However, the two approaches are perhaps best seen as variants of a more general model. Both learn the statistical properties of the stimulus input, in a way that could easily be modeled by a neural net with a hidden layer of nodes between the input of features and the output of category membership. As the hidden layer becomes more restricted, so the ability to retain individual exemplar information is lost and abstraction of a prototype becomes more important.

The second way in which the acquisition of prototypes has been studied is through the study of children's early use of words. Keil and Batterman (1984) presented children with a range of concepts such as 'island' or 'uncle', and tested how they would categorize novel examples. Younger

children aged 4–5 years tended to categorize on the basis of surface appearance – for example, an 'uncle' is an adult male who gives you presents on your birthday. Older children changed to using more definitional information – for example, allowing that if your grandmother had a young male child then that could still be your uncle. Keil accounted for this change by suggesting that children start out forming similarity-based prototype concepts for the meaning of these terms based on the examples that they have experienced. It is only later, as they start to develop causal explanatory principles for organizing their knowledge, that they switch to the correct adult usage of the terms.

CROSS-LINGUISTIC ANALYSES

Schwanenflugel *et al*. (1991) review the influence of cultural and linguistic differences on word meanings and concepts. The culture and language into which a child is born present her with a system of cutting up the world and labeling it which requires attention to the correct attributes and dimensions, and learning of the appropriate underlying theories and models of different domains.

Studies of ethnobiological terms (words for biological kinds) suggest that all cultures divide up the natural world in similar ways, and have labels corresponding roughly to the level of species. In other domains there are, however, important cross-cultural differences in how language divides up the world. For example, Polish has no word corresponding to the English word 'disgust', and English no word corresponding to the German 'Gemütlichkeit' (although it could be translated as 'comfortableness', 'cosiness', or 'amiability' to convey the idea). Although sometimes taken as evidence that thought is thereby constrained by language, the fact that we frequently adopt useful terms from foreign languages into our own vocabulary suggests that the constraint can be overcome. A language that lacks a term for a particular concept is very similar to an individual language speaker whose vocabulary in her native language is limited. There is always the possibility of extending the expressive range of one's language.

A study by Malt *et al*. (1999) investigated the domain of container names in English, Spanish, and Chinese Mandarin. In one task, participants were given a set of photographs of a wide variety of different containers (boxes, cartons, bottles, jars, etc.) and had to say what they would call them. The results showed almost no relationship between one language and another. Three different containers may all be given the same label in one language, yet have three different labels in another. However, when another group of participants rated the similarity of each container to the others, there was far greater agreement between the different language groups. It is therefore clear that the effects of language and the effects of the cultural environment in which a person lives may be independent of each other. Rated similarity was based on a shared experience of the appearance and use of a particular container, whereas the label given to it was highly idiosyncratic to the particular language being spoken.

Another case in which languages differ markedly is in the use of spatial prepositions such as 'in', 'over', or 'through' in English, which according to Lakoff form radial categories of meaning. Even European languages with common roots such as French, Italian, and Spanish have very different ways in which these terms map onto the world, and learning their use requires attention to subtly different conceptual distinctions in each language.

References

Armstrong SL, Gleitman LR and Gleitman H (1983) What some concepts might not be. *Cognition* **13**: 263–308.

Cantor N and Mischel W (1977) Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology* **35**: 38–48.

Carey S (1985) *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Coleman L and Kay P (1981) Prototype semantics: the English word 'lie'. *Language* **57**: 26–44.

Hampton JA (1979) Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior* **18**: 441–461.

Hampton JA (1981) An investigation of the nature of abstract concepts. *Memory and Cognition* **9**: 149–156.

Hampton JA (1982) A demonstration of intransitivity in natural categories. *Cognition* **12**: 151–164.

Keil FC (1989) *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Keil FC and Batterman N (1984) A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior* **23**: 221–236.

Labor W (1973) The boundaries of words and their meanings. In: Bailey CJ and Shuy R (eds) *New Ways of Analysing Variation in English*. Washington, DC: Georgetown University Press.

Lakoff G (1987) *Women, Fire and Dangerous Things*. Chicago, IL: University of Chicago Press.

Malt BC (1994) Water is not H_2O. *Cognitive Psychology* **27**: 41–70.

Malt BC, Sloman SA, Gennari S, Shi M and Wang Y (1999) Knowing vs naming: similarity and the linguistic categorization of artifacts. *Journal of Memory and Language* **40**: 230–262.

Murphy GL and Medin DL (1985) The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316.
Osherson DN and Smith EE (1981) On the adequacy of prototype theory as a theory of concepts. *Cognition* **11**: 35–58.
Posner MI and Keele SW (1968) On the genesis of abstract ideas. *Journal of Experimental Psychology* **77**: 353–363.
Rosch E (1978) Principles of categorization. In: Rosch E and Lloyd B (eds) *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.
Rosch E and Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.
Schwanenflugel P, Blount BG and Lin P (1991) Cross-cultural aspects of word meanings. In: Schwanenflugel P (ed.) *The Psychology of Word Meanings*, pp. 71–90. Hillsdale, NJ: Lawrence Erlbaum Associates.
Wittgenstein L (1953) *Philosophical Investigations*. New York, NY: Macmillan.

Further Reading

Berlin B, Breedlove DE and Raven PH (1973) General principles of classification and nomenclature in folk biology. *American Anthropologist* **75**: 214–242.
Goldstone RL (1994) The role of similarity in categorization: providing a groundwork. *Cognition* **52**: 125–157.
Hampton JA (1995) Testing prototype theory of concepts. *Journal of Memory and Language* **34**: 686–708.
Hampton JA (1997) Psychological representation of concepts. In: Conway MA (ed.) *Cognitive Models of Memory*. Hove, UK: Psychology Press.
Kamp H and Partee B (1995) Prototype theory and compositionality. *Cognition* **57**: 129–191.
Keefe R and Smith P (1997) Theories of vagueness. In: Keefe R and Smith P (eds) *Vagueness: A Reader*, pp. 1–57. Cambridge, MA: MIT Press.
Rosch E and Lloyd BB (1978) *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Schwanenflugel PJ (ed.) (1991) *The Psychology of Word Meanings*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: Harvard University Press.
Van Mechelen I, Hampton JA, Michalski RS and Theuns P (eds) (1993) *Categories and Concepts: Theoretical Views and Inductive Data Analysis*. London, UK: Academic Press.

Psychoactive Drugs

Introductory article

Deepak Cyril D'Souza, Yale University School of Medicine, New Haven, Connecticut, USA
John Harrison Krystal, Yale University School of Medicine, New Haven, Connecticut, USA

CONTENTS

Psychoactive drugs are chemical substances that influence the mind or mental processes by their effects on brain chemical messengers.

INTRODUCTION

The psychoactive effects of a drug are the sum of its effects on the brain, the characteristics of the individual and the setting in which the drug is used, resulting in wide interindividual differences in subjective drug effects. The immediate and long-term effects of a drug differ, as do the effects of short-term and chronic use. The descriptions that follow focus on the immediate effects of short-term psychoactive drug use.

Despite differences in their mechanism of action, chemical structure and behavioral effects, most psychoactive drugs that are misused share the capacity to increase levels of the chemical messenger (neurotransmitter) dopamine in the nucleus accumbens, a region that is believed to be a critical part of the reward pathway of the brain.

DOPAMINERGIC STIMULANTS

Dopaminergic stimulants include amphetamine, related compounds (such as methylphenidate and ephedrine) and cocaine. Cocaine blocks the reuptake of the neurotransmitters dopamine, serotonin and noradrenaline (norepinephrine); amphetamines increase the release of dopamine, noradrenaline and serotonin in addition to blocking the reuptake of these neurotransmitters. Thus, the common net effect of amphetamine and cocaine is an increase in the availability of synaptic dopamine, serotonin and noradrenaline. This explains why these drugs produce similar effects.

Amphetamines

Amphetamines can be used orally, by smoking or by injection. Amphetamines produce a 'rush' – an orgasm-like state, followed by a state of euphoria, a feeling of well-being, alertness, and increased energy and self-confidence that could enhance vocational or social activities, which could last for hours following a single dose. As an individual increases the use of the drug and expands the number of settings in which the drug is used, conditioning occurs – that is, the drug's effects (euphoria and increased energy) become associated with the setting in which the drug was consumed. Conditioning has been postulated to contribute to drug craving and relapse to addiction. With continued repeated consumption, dependence may develop. Some individuals start to binge – bursts of repeated consumption spanning hours to a few days – and this pattern of use may be associated with the symptoms of paranoid schizophrenia. Bingeing is followed by a 'crash' phase or withdrawal syndrome characterized by depression, agitation, anxiety and low energy followed by insomnia and fatigue, culminating in increased appetite and hypersomnolence. In the immediate abstinence period, individuals will experience effects opposite to those of amphetamine along with an intense craving to use the drug. With extended abstinence from the drug, conditioned urges and craving undergo gradual extinction.

Cocaine

Cocaine is derived from the coca plant and is used in the form of a hydrochloride salt or as a free base ('crack'). It is consumed intranasally by 'snorting' or sniffing a 'line', by smoking the free base, or

by intravenous injection. The duration and onset of the effects of cocaine depend on its route of administration. Smoking or intravenous injection produces rapid and accentuated effects in comparison with intranasal use. Since the euphoric effects of a single dose of cocaine last only about an hour, individuals will repeat its use several times in succession within a day. These properties of cocaine are probably the basis of its extremely high reinforcing properties. The effects of cocaine are similar to those of amphetamines except that they are of much shorter duration. As a result of the differences in duration, cocaine is associated with greater repeated use and amphetamine is associated with a longer withdrawal syndrome. So powerful is the direct stimulation of the reward pathway by cocaine that everything else in life becomes irrelevant to the cocaine user. The reinforcing effects of cocaine are strong enough to severely disrupt the social and vocational functioning of the individual.

The dopaminergic stimulants are associated with tolerance, intense psychological dependence and withdrawal. In addition, amphetamines and cocaine may be associated with sensitization.

CAFFEINE

Caffeine, the principal active ingredient of coffee, is perhaps the most widely used psychoactive drug. Caffeine is also present in a wide range of beverages and foods (tea, cocoa, chocolate, colas), prescription medications, over-the-counter analgesics and appetite suppressants. Caffeine produces its effects by blocking adenosine receptors.

At moderate single doses (100 mg) caffeine produces a state of enhanced attention and alertness. Higher (> 250 mg) or repeated doses can produce restlessness, nervousness, excitement, insomnia, agitation, increased energy, a feeling of inexhaustibility, muscle twitches, diuresis (urination), palpitations, increased heart rate and blood pressure, and tremor. Caffeine is associated with physical dependence and a withdrawal syndrome; the latter begins 24–48 h after the last dose and is characterized by lethargy, irritability and headache.

NICOTINE

Nicotine, the active ingredient of tobacco, produces its psychoactive effects through agonist effects at brain nicotinic acetylcholine receptors. Tobacco can be smoked in the form of cigarettes or cigars, or used as snuff which is either insufflated or held between the cheek and gums.

Nicotine has effects that are in some aspects similar to the classic dopaminergic stimulants. In nicotine-dependent individuals (smokers), nicotine produces euphoric and anxiolytic effects, and increases arousal and alertness within seconds. Nicotine-dependent individuals liken its effect to those of amphetamines, but milder. In individuals not dependent on nicotine (e.g. novice smokers), nicotine produces symptoms of dizziness, light-headedness, vertigo and at higher doses nausea and vomiting. It may increase the speed and accuracy of motor activity and may improve the capacity of the brain to process information. Nicotine is associated with physical dependence and a withdrawal syndrome that begins within hours of the last dose and is characterized by mood changes (irritability, anger, anxiety, depression), cognitive changes (drowsiness, fatigue, difficulty concentrating) and physical symptoms (insomnia, hunger, constipation, sweating).

CLASSICAL SEROTONERGIC HALLUCINOGENS

Drugs such as lysergic acid diethylamide (LSD), dimethyltryptamine (DMT), mescaline, derived from the peyote cactus, and psilocybin and psilobin, both derived from 'magic mushrooms' (*Psilocybe mexicana*) produce their psychoactive effects by partial stimulation of the 5-HT_{2A} subtype of serotonin (5-hydroxytryptamine, 5-HT) receptors in the brain. While these drugs may differ in the onset, duration and intensity of effects, their acute behavioral effects are similar. Initially LSD produces somatic symptoms including paresthesias (tingling of the extremities), dizziness, weakness and tremor. This is followed by perceptual alterations including blurring of vision, illusions, visual hallucinations, less discriminant hearing, and altered time perception. Dream-like imagery may develop when eyes are closed. Sensory inputs sometimes become mixed, so that people report 'seeing' sounds or 'feeling' colors (synesthesia). This is accompanied by changes in emotions: several emotions may occur at once, or emotions may become intensified and change rapidly. Depersonalization (the feeling of being separated from oneself) and difficulty separating self from environmental stimuli may occur. Finally, cognitive disturbances are generally mild and rarely include impaired memory, difficulty with attention and concentration. Individuals may report the capacity to gain greater introspection and reveal new 'insights' about themselves, leading to the reputation that these drugs are 'mind revealing' (or 'psychedelic').

Tolerance to the physical effects of these drugs develops rapidly. However, these drugs are not associated with dependence or withdrawal and have low reinforcement liability relative to stimulants.

SUBSTITUTED STIMULANT-HALLUCINOGENS

Drugs such as methylenedioxymethamphetamine (MDMA, 'ecstasy'), *N*-ethyl-3,4-methylenedioxyamphetamine (MDEA), 2,5-dimethoxy-4-methylamphetamine (DOM), methylenedioxyethylamphetamine (MDA, or 'Eve'), and methylenedioxyamphetamine (MDA) are synthetic compounds based on the amphetamine molecule; they are both stimulants and hallucinogens. Though intranasal use has been reported, these drugs are typically ingested orally. The psychoactive properties of these drugs appear to be related to some combination of effects on both serotonin and dopamine release, and partial agonist effects at 5-HT_{2A} (serotonin) receptors. The psychoactive effects begin with a mild 'rush' followed by other stimulant-like effects including euphoria, increased self-confidence and decreased appetite. The use of MDMA reportedly makes people feel more 'open' or more 'connected' to others. Like serotonergic hallucinogens, these drugs produce perceptual alterations (heightened sensory perception, illusions, hallucinations), difficulty concentrating and changes in mood (calmness, or anxiety and panic). Other effects include ataxia (instability of gait), insomnia, jaw clenching (bruxism) and muscle rigidity. It is possible that MDMA and MDA cause irreversible damage to brain serotonin neurons.

There is little information about whether these drugs are associated with tolerance, dependence or withdrawal. However, since they share some properties with amphetamines, one would predict tolerance, dependence and withdrawal to their stimulant-like effects.

CANNABIS

One of the oldest and most widely used psychoactive drugs, cannabis is derived from the flowers and leaves of the plant *Cannabis sativa*. Delta-9-tetrahydrocannabinol (THC) is the principal active ingredient, but there are more than eighty other cannabinoids present in cannabis which may also contribute to its effects. While usually smoked, cannabis is sometimes incorporated into confections and eaten. The application of hydroponics and the use of cloning appear to have increased the THC content of cannabis by up to ten-fold. Delta-9-tetrahydrocannabinol produces its psychoactive effects by actions at brain cannabinoid receptors.

Cannabis intoxication begins about 5–10 min after smoking and includes changes in mood, perception and cognition. The effects of cannabis may last 1–2 h, but its constituents remain in the body for several days. The initial phase of intoxication consists of euphoria which may be accompanied by uncontrollable laughter and a feeling of hilarity. This is followed by a phase of feeling calm, relaxed, mellow, passive, apathetic and drowsy. Sometimes individuals may feel anxious and, rarely, may experience panic attack-like symptoms. The perceptual alterations induced by cannabis include a heightened sensitivity to all sensory modalities; colors and sounds seem brighter and richer, and details previously unappreciated become obvious. This may also be accompanied by a sense of greater insight or self-revelation. There is distortion of time sense (time feels slowed down), spatial sense, and body image. Individuals feel as if they are separated from the experience, watching themselves as distant observers (depersonalization). Hallucinations and synesthesias are rare. The cognitive effects include impairment of short-term memory, free recall, difficulty concentrating, difficulty keeping track of thoughts, rapid flow of ideas and poor judgment. Paranoia (irrational suspiciousness or distrustfulness) is not infrequently observed, but frank psychosis is uncommon. Other effects include motor incoordination, increased appetite, increased heart rate and conjunctival injection (redness of eyes).

There appears to be some evidence suggesting tolerance to some (though not all) of the effects of cannabis, and increasing amounts of evidence suggesting the presence of a mild and subtle, but definite, cannabis withdrawal syndrome.

OPIOIDS

The opioids include the natural compounds (opium, morphine, heroin) isolated from the poppy plant (*Papaver somniferum*), and synthetic compounds such as pethidine, fentanyl, oxycodone and codeine. Opioids produce their psychoactive and physiological effects through actions at delta (δ), kappa (κ) and mu (μ) opioid receptors located in the brain and other systems in the body. The prototypical opioids are agonists for the μ opioid receptors. However, they vary in their affinity for the various subtypes of opioid receptors

and for their agonist/antagonist effects at these receptors.

Opioids can be smoked, used orally, or intranasally, or injected. Especially when injected, they produce an intense 'rush' followed by euphoria, a reduction in anxiety, elevation of mood, drowsiness, and a feeling of tranquillity. At high doses consciousness may be impaired as well as regulation of respiration. Opioids have powerful analgesic (pain-killing) effects and also depress the brain center that regulates respiration. Other effects include nausea and vomiting, narrowing of the pupils, alterations in temperature regulation and disruption of hormone release. This state lasts about 3–5 h. At typical doses, opioids do not appear to significantly impair cognition or perception. Sometimes heroin is combined with cocaine ('speedball') to produce a 'rush' which addicts claim feels better than either drug alone.

Tolerance to the psychoactive effects of opioids develops rapidly with repeated dosing. The need to avoid unpleasant withdrawal symptoms (dependence) is significant. The combination of tolerance and dependence make opioids extremely reinforcing. Approximately 8–10 h after the last dose,

Table 1. Properties of psychoactive drugs

Drug	*Mechanism of action*	*Route of administration*	*Tolerance*	*Physical dependence*	*Withdrawal*
Dopaminergic stimulants					
Amphetamine, methylphenidate	Stimulates dopamine and noradrenaline release and blocks dopamine, serotonin and noradrenaline reuptake	PO, S, IV	++	+	+/−
Cocaine	Blocks dopamine, serotonin and noradrenaline reuptake	PO, S, IN, IV	++	−	+/−
Caffeine	Blocks adenosine receptors	PO	+	+	+
Nicotine	Stimulates nicotinic acetylcholine receptors	IN	+	+	+
Classical serotonergic hallucinogens					
LSD, psilocybin, mescaline	Partially stimulate 5-HT_{2A} receptors	PO	+	−	−
Substituted stimulant-hallucinogens					
MDMA (ecstasy), MDEA (Eve), MDA	Stimulate dopamine and serotonin release, and partially stimulate 5-HT_{2A} receptors	PO	+/−	?	?
GABA-mediated sedative-hypnotics					
Benzodiazepines, barbiturates	Facilitate $GABA_A$ receptor function	PO	+	++	++
Alcohol (ethanol)	Facilitates $GABA_A$ receptor function, inhibits NMDA receptor function	PO	+	++	++
NMDA antagonists					
PCP (angel dust), ketamine (special K)	Inhibits NMDA receptors		−	−	−
Cannabinoids					
Cannabis, hashish	Stimulate cannabinoid receptors	PO, IN	+	−	+
Opioids					
Opium, heroin, morphine	Stimulate μ, κ and δ opioid receptors	PO, S, IN, IV	+	++	++

GABA, γ-aminobutyric acid; 5-HT_{2A}, 2A subtype of serotonin receptor; IN, intranasal; IV, intravenous; LSD, lysergic acid diethylamide; MDA, 3,4-methylenedioxyamphetamine; MDEA, *N*-ethyl-3,4-methylenedioxyamphetamine; MDMA, *N*-methyl-3,4-methylenedioxyamphetamine; NMDA, *N*-methyl-D-aspartate; PO, per oral; PCP, phencyclidine; +, present; −, absent; +/−, equivocal; ?, questionable; S, smoking.

symptoms of opioid withdrawal syndrome set in. The withdrawal syndrome, which can persist for over a week, is characterized by dysphoria (feeling 'down'), yawning, lacrimation (tearing), rhinorrhea (sniffles), sweating, gooseflesh, chills, nausea, vomiting, abdominal cramps, diarrhea, muscle and bone pains, high blood pressure, increased body temperature and restless sleep. A secondary phase of this withdrawal syndrome lasting up to 30 weeks is characterized by low blood pressure, reduced heart rate, dilatation of the pupils, and decreased body temperature.

NMDA ANTAGONISTS

Drugs such as phencyclidine (PCP, angel dust), ketamine (special K) and thienyl phencyclidine (TCP) were originally developed as anesthetic agents. These drugs produce their effects by blocking the *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptors. Interestingly, alcohol is also thought to produce some of its psychoactive effects by a similar mechanism – however, since it also facilitates γ-aminobutyric acid (GABA) receptor function, it is discussed in the next section. Phencyclidine can be inhaled, smoked, consumed orally or injected. The effects occur within minutes of smoking the drug, and include initial euphoria, followed by perceptual alterations (feelings of detachment, distortion of body image, heightened sensitivity to sounds and images, illusions, hallucinations) and cognitive deficits (disruption of memory, difficulty in planning, impairment of judgment, difficulty with abstract thinking, difficulty concentrating and maintaining a train of thought). Other effects include numbness, paresthesias, unsteady gait and slurred speech. The effects of PCP have been likened to the symptoms of schizophrenia, and in rare instances PCP use may be associated with a psychotic syndrome that outlasts intoxication.

There is little evidence to suggest that the misuse of these drugs by humans is associated with tolerance, dependence or withdrawal.

GABA-MEDIATED DRUGS

The drugs referred to collectively as 'sedative hypnotics' include alcohol, barbiturates, benzodiazepines, meprobamate and chloral hydrate among others. These drugs are usually consumed orally and the effects tend to be biphasic: at low doses or soon after consumption the effects are predominantly stimulatory (euphoria, reduction in anxiety, disinhibition of impulses, lability of mood, talkativeness), whereas later on or at higher doses, depressant (drowsiness, amnesia) effects emerge. These drugs also disrupt memory and in some instances can induce black-outs or periods of total amnesia. Other effects include slurred speech, incoordination, unsteady gait, and nystagmus. At high doses or when used in combination these drugs can depress the brain center that regulates respiration and can also lead to coma.

The GABA-mediated drugs are associated with tolerance, physical and psychological dependence, and a clear withdrawal syndrome. The latter is characterized by anxiety, tremulousness, nausea and vomiting, transitory illusions and hallucinations, weakness, sweating, changes in blood pressure and heart rate, headache and insomnia, which begin within hours of the last drink. The withdrawal syndrome may progress to delirium tremens 2–3 days after drinking stops and may culminate in convulsions.

The properties of the above-mentioned drugs are summarized in Table 1.

Further Reading

Davis KL, Charney D, Coyle JT and Nemeroff C (eds) (2002) *Neuropsychopharmacology: The Fifth Generation of Progress*, sect. 10, Substance Use Disorders, pp. 1355–1591. Philadelphia: Lippincott Williams & Wilkins.

Koob GF and Nestler EJ (1997) The neurobiology of drug addiction. *Journal of Neuropsychiatry Clin Neurosci* **9**: 482–497.

Krystal JH, Abi-Dargham A, Laruelle M and Moghaddam B (1999) Pharmacologic model psychoses. In: Charney DS, Nestler E and Bunney BS (eds) *Neurobiology of Mental Illness*, pp. 214–224. New York: Oxford University Press.

Psycholinguistics

Introductory article

Morton Ann Gernsbacher, University of Wisconsin, Madison, Wisconsin, USA
Michael P Kaschak, University of Wisconsin, Madison, Wisconsin, USA

CONTENTS

Psycholinguistics is the scientific study of the production and comprehension of language.

INTRODUCTION

Psycholinguistics is a branch of cognitive psychology that deals with the production and comprehension of language. It emerged in the early part of the 1950s, as the thinking of psychologists was shifting from behavioristic views towards the 'cognitive revolution'. Although psycholinguists initially concerned themselves with exploring the psychological aspects of Noam Chomsky's work in linguistics, the field has grown to embrace a wide range of methodologies and theoretical approaches to studying the development of language (how do children learn language?), the comprehension of language (how do we understand conversations, novels, lectures, and more?), the production of language (how do we produce spoken and written language), and the loss of language (such as occurs with brain damage or stroke).

PSYCHOLINGUISTIC METHODS

Psycholinguists are first and foremost scientists who use experimental methods to test their ideas about how language works. The following are the more common methods used in the field.

Self-report/Behavioral Methods

These methods are commonly referred to as 'offline' measures because the measurement is not taken directly while the participants are comprehending or producing language. One example of such a task is a 'cloze procedure' task (e.g. fill in the blank: 'Megan walked _____.').

Reading Time/Reaction Time Methods

Time-based measures (i.e. measuring how rapidly a participant can perform a task) are by far the most common set of methods used by psycholinguists. Typically, researchers use computers to display words or sentences to research participants, and then time how long it takes for the research participants to read what was presented on the screen. This type of experiment can be used to measure how long it takes a person to read each word in a sentence, or how long it takes to read the entire sentence. By looking at which words and sentences are read comparatively quickly or slowly, researchers can make inferences about the nature of the processing involved in reading what was presented.

Variants of this methodology time responses to particular stimuli, rather than recording the time it takes to read, *per se*. Two widely used variants on this method are the probe response task and the lexical decision task. In the former case, research participants are often presented with a word and asked to determine as quickly as possible if it is related to a sentence or story they had just read. In the latter case, research participants are presented with stimuli, and they must decide as quickly as possible if the stimulus is a word or not (e.g. 'bank' is a word, whereas 'blick' is not). As before, the pattern of response times can be used to test predictions generated by different theories. Both reading time and reaction time methodologies (as well as all the methods discussed subsequently) are considered 'online' measures because the measurement is taken while language is being processed.

Priming Experiments

Priming experiments are a special type of reaction time experiment. In these experiments, subjects are presented with a stimulus (called a 'prime'), which is hypothesized to speed their response to a subsequently presented stimulus (called a 'target'). For example, presenting the word 'doctor' (prime) will

speed responding to the word 'nurse' (target), but will not speed responding to the word 'tiger'. Priming experiments are often used to test hypotheses about how word meanings relate to each other, and how information is organized in the brain.

Eye-tracking Methods

Since the 1980s, a popular method in the field has been tracking the movement of a research participant's eyes as the participant reads words, sentences, or passages, or inspects visual displays. Eye-tracking involves the use of lasers. These methods allow researchers to measure how long readers spend viewing each word they read, and when and if readers go back to re-read earlier parts of sentences. Eye-tracking also allows researchers to measure what objects in the environment research participants look at when they are listening to language. This method helps explain what happens at the earliest stages of language comprehension.

Event-related Potential (ERP) Methods

One way of determining what the brain is doing as it processes language is to use ERP methods. ERP studies examine the patterns of electrical activity in the brain as the subject is presented with various types of stimuli. These studies allow researchers to get a general sense of what parts of the brain are active when language is processed, and to determine what types of language present the brain with processing difficulty.

Neuroimaging Methods

Neuroimaging methods encompass a family of techniques that can be used to determine somewhat precisely what areas of the brain are active during the processing of different types of language, and at different stages of language processing. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) are common brain-imaging methods. Most of the techniques in this category exploit the increased flow of blood to active brain regions to determine which parts of the brain are used in particular types of processing. As ERP and neuroimaging techniques have become more widely available, the field has witnessed a growth in the number of researchers who use several methods cooperatively to draw more precise conclusions about how language is processed.

HISTORICAL DEVELOPMENT OF THE FIELD

Origin in the 1950s

Although many consider the 1957 publication of Noam Chomsky's book *Syntactic Structures* to mark the birth of psycholinguistics, the field was formally christened 4 years earlier. In 1953, Cornell University hosted a conference of linguists and psychologists who wished to bring their fields together to address issues of common concern. It was here that the term 'psycholinguistics' was coined. Despite the resolve of the members of this conference, however, it wasn't until the publication of Chomsky's work that psycholinguistics gained wide recognition as a field in psychology.

Chomsky's early work in linguistics was critical for the development of psycholinguistics in two ways. First, his linguistic theory (called 'transformational generative grammar') used a mathematical system of notation to create a precision previously unknown in the field. The importance of this idea cannot be underestimated, as one crucial factor that kept 'cognitive' terms out of behavioristic theory was the lack of precision with which many terms (such as 'attention') could be defined. With a precise linguistic theory in hand, psychologists could now explore the workings of language in a more exact, scientific manner. Chomsky's second contribution to the rise of psycholinguistics was his review of noted behaviorist B. F. Skinner's work *Verbal Behavior*. In the review, Chomsky took Skinner to task for presenting an oversimplified view of language acquisition based on the behaviorist principles of reward and punishment. This review made it clear to psychologists that a new approach to the study of language, and to the study of cognitive processes, was required.

1960s: Psycholinguists Work to Verify Linguistic Analyses

Through the latter part of the 1950s and much of the 1960s, psycholinguists worked to develop theories of language processing based on Chomsky's generative grammar. To illustrate, consider Chomsky's proposal that each sentence has a *deep structure* (a structure that translates the meaning of the sentence into a syntactic form) that can be transformed in a number of ways to produce various *surface structures* (roughly, the actual order of the words in the sentence). A deep structure such as 'Mike kicked the ball' could be produced without transformation (leaving it as it is), or with one

transformation, rendering the sentence as 'The ball was kicked by Mike'. Both versions of the sentence have the same deep structure (and, hence, the same meaning), but are produced with different surface structures. In pioneering psycholinguistic work, George Miller and colleagues demonstrated that the response time to simple sentences like those presented above was related to the number of transformations required to get from the surface form to the deep structure. Thus, all things being equal, it takes longer to read a sentence with one transformation ('The ball was kicked by Mike') than to read a sentence with no transformations ('Mike kicked the ball').

1970s: Separation from Goals of Linguistic Theory

Towards the end of the 1960s, psycholinguistics began to broaden in scope. Inspired by the theory of information processing that was developing throughout cognitive psychology, psycholinguists began to move away from the notion that language processing was best understood with respect to linguistic theory, and towards the notion that language processing could be characterized in terms of the types of information processing that computers performed. The predicate calculus system of representation presented in Miller and Johnson-Laird's seminal work *Language and Perception*, as well as proposition-based theories of language processing, were characteristic of this new approach. To contrast this new approach with that taken in the 1960s, consider again the difference between the two versions of the 'Mike …' sentence presented earlier. The early psycholinguists attributed the longer reading time of the passive version of the sentence to the presence of a transformation that needed to be performed in order to recover the deep structure of the sentence; the more recent theorists proposed that the difference in reading time was due to the more complicated propositional structure associated with the passive sentence.

1980s: Preoccupation with Modularity and 'Early Is Interesting'

Although theories of language processing moved away from a strict reliance on linguistic and philosophical analysis, central issues from those fields continued to be important to psycholinguists. A prime example is the question of whether language could be considered a 'modular' cognitive structure. Jerry Fodor presented a strong statement that certain aspects of language processing, such as the recovery of the syntactic form of a sentence, are performed by specialized cognitive machinery without influence from other structures or sources of information. Thus, the syntactic form of a sentence would be computed without recourse to information such as the meaning of the words in the sentence. This idea, which was known as 'modularity', was hotly debated throughout the latter part of the 1970s, and the early to mid-1980s.

Interest in the modularity of language was fueled in part by the advent of technology that allowed researchers to get finer-grained measures of the time it takes to read words and sentences than had been possible before, as well as by the advent of devices that allowed one to track the eye movements that individuals made while reading. With these powerful tools, psycholinguists attempted to ascertain whether sentences were initially processed with respect to syntax only (the 'language is modular' position), or with respect to multiple sources of information, such as syntax and meaning. As the debates intensified, there was great interest in obtaining data on how people processed sentences at earlier and earlier stages of the comprehension process, based on the belief that questions about modularity could be answered conclusively by figuring out what people did first when reading.

1990s: Investigation of Constraint Satisfaction Models and Other Approaches to Psycholinguistics

As the 1980s ended, interest in the modularity issue waned. This was in part because the imagination of psycholinguists was captured by the investigation of constraint satisfaction theories of language processing. Constraint satisfaction theories proposed that language comprehension involves the rapid use of mutually constraining sources of information, such as knowledge about syntactic structures and word meanings. With the development of computer-based neural networks, it became possible to develop sophisticated models of constraint satisfaction theories, and much effort was directed towards developing computer simulations of language processes. These new theories emphasized the notion that language processing is nonmodular. They also diverged from past work in psycholinguistics by drawing not from Chomskyan theories of linguistics, but from alternative approaches to characterizing the structure of language (e.g. head-driven phrase structure grammar).

One further development in the 1990s deserves mention. Several researchers are now approaching language processing based on the idea of embodied cognition. While only in its infancy, this work has shifted the focus away from studying how a reader generates the structure of the language, and back towards how meaning is generated from language. This recent development, along with constraint satisfaction theories and attempts to use neuroimaging to study language processing, promise to move psycholinguistics in interesting new directions in the future.

Further Reading

Chomsky N (1959) Review of B. F. Skinner's *Verbal Behavior*. *Language* **35**: 26–58.

Fodor JA (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Just MA and Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. *Psychological Review* **87**: 329–354.

Kintsch W (1988) The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review* **95**: 163–182.

MacDonald MC, Pearlmutter NJ and Seidenberg MS (1994) Lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703.

Miller GA (1965) Some preliminaries to psycholinguistics. *American Psychologist* **20**: 15–20.

Miller GA and Johnson-Laird PN (1976) *Language and Perception*. Cambridge, MA: MIT Press.

Mitchell DC (1994) Sentence parsing. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 375–409. San Diego, CA: Academic Press.

Pinker S (1994) *The Language Instinct*. New York, NY: HarperCollins.

Skinner BF (1957) *Verbal behavior*. New York, NY: Appleton-Century-Crofts.

Psychology: Experimental Methods

Introductory article

Robert W Proctor, Purdue University, West Lafayette, Indiana, USA
EJ Capaldi, Purdue University, West Lafayette, Indiana, USA
Kim-Phuong L Vu, Purdue University, West Lafayette, Indiana, USA

CONTENTS

Experimental methods in psychology are the procedures used to isolate the effects of manipulations on behavioral measures.

USE OF EXPERIMENTS IN PSYCHOLOGY

Experimental methods have been associated with psychology since the field's beginnings. The founding of psychology as a science is usually dated to 1879, when Wilhelm Wundt established the first laboratory devoted to experimental investigation of psychological phenomena. This association of psychology with the experimental method and its emphasis on control of the environment is what enabled the discipline to make the claim of being a science. The scientific approach provides a more objective method for establishing facts and evaluating alternative possible explanations. Throughout its history, experimentation has remained the central method of psychology, although it has not been without its critics, and non-experimental methods such as naturalistic observation and survey research have come into increasingly wide use.

Experiments can be conducted with humans or animals. The specific population that is studied will depend on several factors, including the topic with which the research is concerned, the theoretical predispositions of the researcher, the specific methods that are feasible with a particular population (e.g. humans cannot be lesioned, but non-humans cannot provide verbal reports), and the fact that more control can be exerted over a laboratory animal's history and environment than can be extended over a human's. Much psychological research in the late nineteenth century used human subjects, in part because researchers had an interest in the subjective experience of perceptual events. Beginning in the early twentieth century with the behaviorist movement, the use of animals increased. Because the learning and conditioning principles studied by the behaviorists were considered to be generalizable across species, much of the research focused on rats and pigeons, animals that can be studied easily.

Research on humans continued to be conducted throughout this period, but a major renewal of interest occurred with the advent of contemporary cognitive psychology in the 1950s. Most experimental research on humans is conducted in laboratory settings with undergraduate psychology students. One concern with such research is the extent to which the principles derived from it generalize to other populations and beyond the laboratory. Therefore, human populations other than undergraduate students are also sometimes tested. For example, there is now a large literature on expert–novice differences, in which performances and strategies of experts in a domain such as chess are compared with those of novices. Comparisons across different human and non-human populations are essential in many areas of research. Developmental psychologists compare experimental results across different age groups or over a period of development to evaluate the course of development throughout the lifespan. Comparative psychologists study a range of species, including humans, using similarities and differences across species to draw conclusions about general and species-specific behaviors.

In the last quarter of the twentieth century there was increasing concern with ethical principles in experimentation. All research must first be

approved by an institutional review board. This board determines whether the costs and benefits of a proposed research project justify its being conducted. If an experiment poses more than minimal risk for humans or hazard to non-humans, the procedure must be strongly justified by the researcher, who must ensure that precautions will be taken to minimize the risk to the subjects. There are guidelines that prescribe ethical principles for human and non-human research. Humans must give their informed consent to participate after being told about the general nature of the experiment, and they must be provided with a more detailed debriefing about the purpose of the study upon completion of their participation.

CONTROL IN PSYCHOLOGY EXPERIMENTS

Many properties vary simultaneously in the world, making it difficult to determine the causal relationships among them. In an experiment, the basic idea is to control those properties, or variables, that are not of interest to the experimenter and to isolate the critical ones. One or more variables (the 'independent' variables) are manipulated, and their effects on behavioral or psychophysiological variables (the 'dependent' variables) are measured. The question of interest is whether, and how, the conditions defined by the values of the independent variables influence the dependent variables. Often a researcher wants to test implications about hypothetical constructs that are not directly observable, and these constructs must be mapped to observable variables. This mapping yields 'operational definitions' of the constructs in terms of the operations the researcher uses to manipulate and measure them. For example, if the goal of an experiment is to determine whether learning of a repeated sequence of events occurs without awareness, the researcher must decide on appropriate measures of 'learning' (e.g. faster response speed on a repeated sequence than on a random series) and 'without awareness' (e.g. inability of the participant to verbalize whether a sequence is present). Debates in many areas of research center around the appropriateness of various operational definitions of the theoretical constructs.

When conducting research, the values of the variables must be measured accurately. All measures consist of two components: true measure and error. For example, on a recognition memory test, in which a 'yes' or 'no' response must be given regarding whether each test item was seen in a previous list, the proportion of correct responses reflects both the accuracy of memory and whether lucky guesses are made when the person is unsure. Two people with the same exact accuracy of memory may show different proportions of correct responses on the memory test. A reliable measure will give the same or similar values every time. A valid measure will reflect what it is intended to measure. For example, if repeated measurements were made of the duration of a stimulus that is displayed for exactly 1 second, a timing device that consistently read 1.5 s would be reliable, but it would not be a valid measure of the stimulus duration. The concept of two components of measures, true and error, is central to the logic of the inferential statistical methods that are used to decide whether independent variables affect dependent variables.

A good experiment controls extraneous variables that could produce differential effects on the dependent variable across conditions. When the extraneous variables are not controlled, they are said to be confounded with the independent variables. When a systematic confounding is present, the experimental results could be a consequence of the confounded variables and not the independent variables. Control of extraneous variables may be accomplished in several ways, such as holding a variable constant or distributing its effects equally across conditions. It is important to eliminate confounds because a researcher can only claim that the independent variable caused any observed differences on the dependent measure when there are no alternative explanations. A well-controlled experiment with no obvious confounds is said to have high internal validity because the researcher can be relatively confident that the independent variable caused any observed effect on the dependent variable.

An example of a poorly designed experiment with many confounds is the single-group pre- and post-test design. With this design, the researcher measures the dependent variable before and after manipulating the independent variable and compares the pre-test and post-test values. The goal is to attribute any differences in the dependent measure to the effect of the independent variable.

At first glance, this design seems like a good one, because performance on the post-test can be compared with performance on the pre-test as a control condition. However, this design does not control variables adequately. Suppose a researcher tests the hypothesis that consuming caffeine (a stimulant) prior to an examination decreases performance. The researcher has the students in a class

take one version of an examination and, afterwards, drink a cup of caffeinated coffee. A second version of the examination is then administered, and the scores on it are found to be 20 percent lower than those on the first version. Is it appropriate for the researcher to conclude that consuming caffeine before an examination decreases performance? The answer is 'no' because the experimental design did not control for confounding variables. Several factors other than consumption of caffeine could have caused the lower scores on the second examination: the second version may have been more difficult than the first; the students may have been more fatigued by the time they took the second examination, and consequently may not have tried as hard; the act of consuming the coffee itself may have affected the students' performance. A better experimental design would use a second group of students who received decaffeinated, rather than caffeinated, coffee. This group could provide a control for the confounds described above. In a double-blind design, neither the researcher conducting the experiment nor the student knows whether they are receiving caffeinated or decaffeinated coffee. In addition, half of the students should receive the second version of the examination for the pre-test and the first version for the post-test, so that the averages for the pre- and post-tests would reflect equal contributions from the two versions. With this two-group pre- and post-test design, the three alternative causes described above for the difference between pre- and post-test results can be ruled out if the group that received the caffeinated coffee shows the difference in performance between the pre- and post-tests but the group that received the decaffeinated coffee does not.

For control procedures of this type to be effective, the independent variable must be the only variable that systematically distinguishes the two groups. In other words, the groups must be equivalent in other respects, subject to the laws of probability. If this is the case, then statistical tests can be performed that enable the researcher to evaluate whether the mean difference between the groups can be attributed entirely to chance or is large enough to suggest a systematic difference due to the independent variable.

BETWEEN-SUBJECTS AND WITHIN-SUBJECTS DESIGNS

An independent variable can be manipulated either between subjects or within subjects.

For a between-subjects design, different groups of participants receive each condition, or value of the independent variable. There are two permissible methods for assigning the participants to the different groups. The simplest is random assignment, in which each person is arbitrarily assigned to a condition. The justification for random assignment is that the groups will be equivalent within chance limits for all subject variables (i.e., subject characteristics) because only chance determines who gets assigned to which condition. The second method is to match the groups according to subject variables known to be correlated with the behavior of interest. For example, one might wish to match groups or individuals in a memory experiment according to intelligence quotient (IQ) scores to ensure that the groups are closely equated on this characteristic.

To illustrate these two methods, we will assume that there are only two groups, *A* and *B*, and that 40 participants are being tested. With random assignment, only chance determines whether a participant will be put in group *A* or group *B*, with the constraint that 20 participants should be in each group. To match the two groups on IQ, participants are ranked from high to low and usually matched in pairs. The researcher randomly assigns one of the two members of each pair to group *A* and the other to group *B*. Thus, as with random assignment, there are 20 participants in each group, but each is participant is paired with another who scored similarly on the IQ test. Both procedures allow the researcher to assume that the two groups are equivalent within chance and that the independent variable is the only factor that systematically distinguishes them. The major benefit of the matching procedure is that it reduces the chance, or error, component of the measurements. In other words, if IQ score is highly correlated with memory ability, then the average memory ability for the subjects in the two groups is more likely to be highly similar if the matching procedure was used than if only random assignment was employed. A consequence of this reduction in chance variability is that it is easier to detect any effect of the independent variable on the dependent measure.

Regardless of which method is employed, if a statistically significant difference between the two groups is detected, this difference can be attributed to an effect of the independent variable because it is the only factor systematically distinguishing the two groups.

For within-subjects, or repeated-measures, designs, each participant is tested in all experimental

conditions. In the example above, 20 participants would be tested, rather than 40, to obtain 20 scores in each condition. This illustrates one advantage of using a within-subjects design: fewer participants are needed because each participates in all conditions. More importantly, this design is more sensitive than a between-subjects design to differences between conditions, because it effectively matches all conditions precisely on all possible subject variables (i.e. each subject contributes a score to each condition).

There are also disadvantages of using within-subjects designs. There may be order effects, whereby the order of the conditions affects the dependent variable. If these are likely to be present then a within-subjects design would be inadvisable. In addition, there can be general practice or fatigue effects, whereby previous exposure to the task affects the dependent measure. The general effects can be controlled by counterbalancing the order in which the conditions are presented, so that, in the example, half of the participants receive condition *A* first and half condition *B* first. With complete counterbalancing, all possible orders of presentation are included in the experiment, thus averaging out general practice and fatigue effects.

Thus, there are trade-offs between the different research designs. The choice of design should be determined by a clear understanding of the objectives of the experiment. Whether a between- or within-subjects design is preferred depends on the goals of the experimenter and on the nature of the experiment being conducted. Within-subjects designs require fewer participants and are more sensitive to effects of the independent variable. However, within-subjects designs cannot be usedwhen there is a possibility of differential 'carry-over' effects, or effects that are influenced by exposure to the previous conditions. In addition, within-subjects designs cannot be used when any experimental manipulation causes permanent changes in the subject (e.g. brain lesion).

COMPLEX DESIGNS AND INTERACTIONS

Most research uses complex designs in which two or more independent variables are manipulated. The conditions comprise all combinations of the values of the independent variables. All variables can be manipulated either between subjects, in which case the number of groups will equal the number of conditions, or within subjects, in which case there will be only one group. In mixed designs, one or more variables is between subjects and one or more is within subjects. The caffeine example above was a mixed design in which the type of coffee (caffeinated or decaffeinated) was a between-subjects variable and whether the examination was pre-test or post-test was a within-subjects variable. In mixed designs, the number of groups is the number of between-subjects conditions.

The advantage of manipulating more than one independent variable is that interactions can be studied. This is important because the effects of a variable often depend on other variables. For example, consuming a stimulant like caffeine, which increases levels of arousal, may enhance performance of a simple task but degrade performance of a complex task. In this case, the effect of caffeine interacts with that of task difficulty. In a design with two independent variables, regardless of whether they are manipulated between or within subjects, one can look at main effects of each variable as well as their interaction. A variable has a main effect if there is a difference in the dependent measure as a function of its values when averaged across all values of the other independent variable. An interaction occurs if the effect of one variable is different in magnitude or direction for different values of the other variable. If a variable has a main effect but no interaction, then this effect generalizes across values of the other independent variable. If it has a main effect but also an interaction with the second independent variable, the effect of the variable must be examined at each value of the other variable.

STATISTICAL SIGNIFICANCE TESTING

The purpose of statistical tests in experimental psychology is to determine whether observed differences in the dependent measure for each condition are due to the manipulation of the independent variable or to chance error. Significance testing is used to determine whether the difference in the dependent measure is larger than that expected due to chance. In significance testing, there are usually two hypotheses: the 'null' hypothesis and the alternative one. The null hypothesis usually states that the manipulation of the independent variable had no effect on the dependent one. The alternative hypothesis usually states that the manipulation of the independent variable had an effect on the dependent one. If there is evidence that the difference between groups is larger than that due to chance alone, then the null hypothesis is rejected in favor of the alternative one.

Before conducting an experiment, the researcher selects an 'alpha' level, which is the probability

level required for significance. Most researchers use an alpha level of .05, which means that if there is less than a 5 percent chance that the results were due to random error, the researcher concludes that the difference is significant, or that there is an effect of the independent variable. If the probability is greater than .05 that the difference between conditions could result from chance alone, the null hypothesis is not rejected. Because of the arbitrariness of the choice of alpha, some psychologists have criticized the use of null hypothesis testing and recommended other methods, such as the use of confidence intervals for reaching decisions.

There are four possible outcomes of significance testing (see Figure 1). First, the null hypothesis that there is no difference between the groups may be accepted (or, more accurately, not rejected), when in reality there is no difference (correct decision). Second, the null hypothesis may be rejected when in reality there is not a difference between the groups ('Type I error'). Third, the null hypothesis may be accepted when in reality there is a difference ('Type II error'). Finally, the null hypothesis may be rejected when in reality there is a difference (correct decision). When the .05 alpha level is used and the analysis shows that there is a significant difference between groups, the decision to reject the null hypothesis will lead to a Type I error 5 times out of 100. The probability of a Type II error occurring when the statistical test leads to acceptance of the null hypothesis is more difficult to specify because it depends on factors such as the size of the measurement error and the number of subjects tested. If an effect size of interest, and error variability for the analysis, can be estimated, a power analysis can be performed to determine the sample size needed to be reasonably sure of detecting an effect of the independent variable if one is really present.

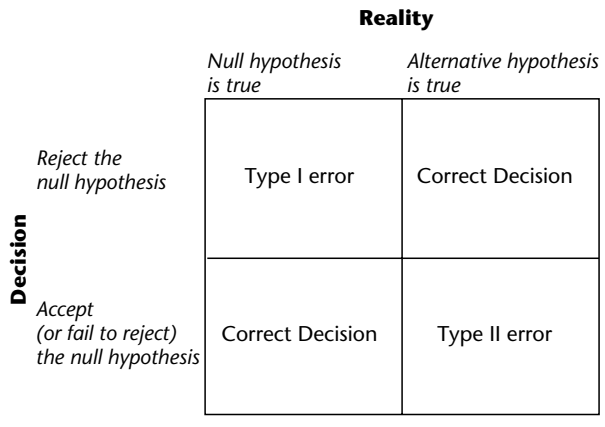


Figure 1. Decision matrix for statistical significance testing.

VARIETIES OF SCIENTIFIC METHODS

According to a popular point of view, the major feature that distinguishes science from other activities is its method, the scientific method. The scientific method is commonly identified with hypothesis testing. According to this approach, first introduced by William Whewell in about 1850, an empirically testable hypothesis is formulated on some basis. A prediction from the hypothesis is derived and then tested by designing a suitable experiment. If the experimental finding is consistent with the prediction, the hypothesis is retained. Otherwise, the hypothesis is rejected. This formulation may be found in virtually all of the widely used methodology textbooks written for undergraduate psychology students.

However, hypothesis testing is not as simple and clear-cut as is suggested by most methodology textbooks in psychology, because of what has come to be known as the Duhem–Quine thesis. This thesis rightly suggests that when a hypothesis is subjected to experimental test, numerous additional propositions are simultaneously under test. Thus, failure to confirm the hypothesis may be due to one or more of these additional factors. For instance, is the measuring instrument (Geiger counter, maze, paper and pencil test) an appropriate vehicle for testing the hypothesis? It may be insufficiently sensitive to test the hypothesis. Is the measuring instrument in good working order? It may be improperly calibrated. Is the prediction drawn from the hypothesis a valid one? For example, by emphasizing some neglected aspect of the hypothesis, a somewhat different prediction may follow. Can the prediction be rendered consistent with the data by making an additional modest assumption? As many historians of science have suggested, much good science has resulted from individuals resolutely holding on to a supposedly disconfirmed hypothesis by making such additional assumptions.

It is possible to object to equating science with hypothesis testing for the following reason: other methods have led to successful and useful scientific theories. For example, some useful theories, called explanatory theories, have been devised by explaining already known phenomena, rather than predicting new ones. A recent example of this is plate tectonics in geology. That theory, which was based exclusively on already known phenomena when initially formulated, was viewed quite

unsympathetically by American geologists who favored the hypothesis testing approach. European geologists, who were much less wedded to hypothesis testing, were much quicker to embrace plate tectonics.

Behaviorism is a set of broad assumptions that has led to the formulation of many individual theories, ranging from Tolman's cognitive behaviorism to Skinner's radical behaviorism. More recently, cognitive psychology has similarly given rise to a variety of theories. Broad approaches such as behaviorism and cognitive psychology have been called, variously, paradigms, research programmes, and research traditions. Needless to say, perhaps, broad research traditions such as behaviorism and cognitive psychology have had a great influence on the theory and practice of experimental psychology. Yet at the time of their introduction they were not able to explain many interesting phenomena; nor did they give rise to many novel, confirmed predictions. Their original acceptance could be likened to the acceptance of a promissory note, which promises a big pay-off in the future. Clearly, the ability to explain existing phenomena or to predict new ones is not the only factor in shaping interesting scientific research. At least as important is the capacity of an approach to attract adherents because of its promise of leading to fruitful discoveries.

Thus, major components of the scientific method include, in addition to hypothesis testing, explanatory theories that apply themselves, at least in their initial stages, to already known phenomena, and broad viewpoints, known variously as paradigms, research programs, or research traditions, that are accepted at first on the basis of their perceived ability to solve interesting scientific problems. If we view science in this broader perspective, it becomes apparent that different scientists will have different thresholds for accepting or rejecting theoretical propositions. For example, a few scientists will be quick to accept, at least tentatively, a new theory that appears to them to be promising; while other scientists may hold onto a theory long after most others have abandoned it. Of course, many intermediate positions are possible. History teaches that which possibility turns out to be correct in a particular case will depend on a variety of circumstances that can only be determined on the basis of additional research.

Further Reading

Anderson CA, Lindsey JJ and Bushman BJ (1999) Research in the psychological laboratory: truth or triviality? *Current Directions in Psychological Science* **8**: 3–9.

Capaldi EJ and Proctor RW (1999) *Contextualism in Psychological Research? A Criticial Review*. Thousand Oaks, CA: Sage.

Chalmers AF (1999) *What Is This Thing Called Science?* 3rd edn. Indianapolis, IN: Hackett.

Garner WR, Hake HW and Eriksen CW (1956) Operationism and the concept of perception. *Psychological Review* **63**: 149–159.

Kantowitz BH, Roediger HL and Elmes DG (2001) *Experimental Psychology: Understanding Psychological Research*, 7th edn. Belmont, CA: Wadsworth/Thomson Learning.

Kirk RE (1999) *Statistics: An Introduction*. Fort Worth, TX: Harcourt Brace.

Maccorquodale K and Meehl PE (1948) On a distinction between hypothetical constructs and intervening variables. *Psychological Review* **55**: 95–107.

Proctor RW and Capaldi EJ (2001) Improving the science education of psychology students: better teaching of methodology. *Teaching of Psychology* **28**: 173–181.

Scarborough D and Sternberg S (eds) (1998) *An Invitation to Cognitive Science*, **vol. IV** 'Methods, Models and Conceptual Issues'. Cambridge, MA: MIT Press.

Shaughnessy JJ, Zechmeister EB and Zechmister JS (2000) *Research Methods in Psychology*, 5th edn. Boston, MA: McGraw-Hill.

Psychophysics

Introductory article

Daniel Algom, Tel Aviv University, Ramat-Aviv, Israel

CONTENTS

Psychophysics is the scientific study of the relation between stimulus and sensation, and therefore its study concerns fundamental questions of psychology and cognitive science.

INTRODUCTION

Psychophysics is the scientific study of the relation between stimuli and sensations. According to the consensual view in psychophysics, the human perceptual system is a measuring instrument whose sensitivity to changes in the environment can be quantitatively analyzed. Of the many aspects of impinging stimuli and the corresponding sensations, three stand out and have received much attention. First, the system's response must exceed that triggered by a critical stimulus intensity – threshold – for any sensation at all to be experienced. Second, when a stimulus more intense than the threshold impinges on the sense organ, its intensity must be increased by a critical amount – the difference threshold – for the person to sense a just noticeable difference (JND) in sensation. Finally, of paramount importance is specifying the functional relation between stimulus magnitude as assessed by the instruments of physics and sensation magnitude as assessed by the *scaling* of people's perceptions.

Before describing each of these traditional problem areas, the fundamental finding of psychophysical research should be appreciated: a human being is not a perfect measuring instrument, infinitely sensitive to changes in the impinging stimuli. High levels of refinement are thus required of the methods and theories of psychophysics in their quest to unravel the operational characteristics of the human perceptual system.

ABSOLUTE THRESHOLD AND SENSITIVITY

The absolute threshold, the smallest amount of stimulus energy necessary to produce sensation, specifies the sensitivity in a given sensory modality. Despite the appealing simplicity of the concept of threshold, its measurement poses difficult problems. First, the absolute threshold is not a rigidly fixed value because both sensitivity and the intensity of a nominally invariant stimulus fluctuate irregularly over time (the latter variation refers to quantal fluctuations). Second, the threshold also varies with changes in the conditions of stimulation. For example, to produce threshold sensation in vision requires substantially more energy at one wavelength than at another. In general, the senses are not uniformly sensitive over their respective ranges of detectable energy.

There is no answer to the simple question 'how sensitive is the eye to light?' The sensitivity of the eye cannot be gauged by a single threshold because it depends on the conditions of stimulation such as prior exposure of the eye to light, and the wavelength, area, and duration of the stimulus. In contrast, the sensitivity of the eye under optimal conditions of stimulation can be determined. In order to see, it is necessary for only one quantum of light to be absorbed by a single molecule of photochemical pigment in each of five to 14 receptors in the retina. The maximum sensitivity of the eye is constrained only by the limit imposed by the nature of light. The ear nearly matches that remarkable sensitivity: under optimal conditions, the eardrum has to move a distance less than the diameter of a hydrogen molecule for a sound to be heard. Given its dependence on the conditions of

stimulation, the absolute sensitivity of a perceptual system is best described by examining the relations between the threshold and the various stimulus dimensions that affect its value.

Given the variability, the threshold must be measured statistically. For instance, it can be calculated as that level of energy detected on half the stimulus presentations. This definition is adopted in the *method of constant stimuli* in which fixed stimulus levels are presented many times in an irregular order. In the *method of limits*, stimuli are presented in order of increasing or decreasing magnitude until the observer modifies her or his responses from 'not perceived' to 'perceived' in an ascending series, or from 'perceived' to 'not perceived' in a descending series. The average value of the reversals defines the threshold. Finally, in the *method of adjustment* the observer sets the level of the stimulus so that it is just perceptible. The threshold is the average of several settings. The three methods, known as the classical psychophysical methods, have been in widespread use ever since their inception by Gustav T. Fechner, the founder of psychophysics in the nineteenth century.

Novel methods include variations on the classical ones. In the *staircase method*, a variation on the method of limits, stimulus levels around the threshold are made weaker when the observer detects them and made stronger when they go undetected. Hence, the method is adaptive and saves time because stimuli much below or above threshold are never presented. In the *threshold tracking method*, the observer maintains just perceptible intensity of the stimulus as it continuously changes on another dimension such as frequency. Individuals differ greatly in how much they tend to report the presence of sensation when no stimulus is present, a tendency referred to as 'response bias'. Methods of *forced choice*, in which the observer *must* select on each trial that time interval or location that contains the stimulus, control for response bias, as do the methods based on the theory of signal detectability (TSD) that include presentations of *catch trials* containing no stimuli. Within the framework of TSD, separate and pure measures have been developed for sensitivity and for response bias.

DIFFERENCE THRESHOLD, SENSORY RESOLVING POWER, AND WEBER'S LAW

The difference threshold (DL for the German *Differenz Limen*), the smallest detectable change in energy, specifies the resolving power of a given sensory system. It is the stimulus increment, ΔI, required to produce a JND in sensation. Because the difference threshold is as variable as the absolute threshold, the same methods of measurement are adapted to specify its value. Using the *method of constant stimuli*, for instance, the observer reports on each trial whether a stimulus (selected from a set of predetermined levels) is stronger than the invariant standard stimulus. The difference threshold can be defined as half the distance between the levels reported stronger than the standard on a quarter and on three-quarters of the presentations. The stimulus level that is perceived to be stronger than the standard as often as not, is defined as the point of subjective equality (PSE), and the difference between the PSE and the standard stimulus is a psychophysical quantity called the 'constant error'.

Applying the method of constant stimuli (or any other method) recurrently, DLs can be specified for several stimuli taken as standards. Does the DL remain invariant at different intensities along a given continuum? Ernst H. Weber discovered that it does not. According to Weber's law, the DL depends on the starting intensity in a linear manner such that $\Delta I = cI$, where ΔI stands as before for the stimulus increment necessary to produce a JND when added to the starting stimulus level I; c is a constant known as Weber's fraction. Dividing both sides of the equation by I gives what is perhaps the better-known form of Weber's law, $\Delta I/I = c$, where $\Delta I/I$ is again Weber's fraction. Weber's law states that the minimum change in stimulus intensity that can be noticed is a constant fraction of the starting intensity of the stimulus. To be discriminable, the intensities of two stimuli must differ by an amount that is proportional to their absolute level.

Weber's fraction differs for different sensory reactions, which means that sensory systems differ in their resolving power. The constant is less than 1 percent for pain, about 4 percent for visual length and heaviness, approximately 8 percent for brightness and loudness, and can be as high as 20 percent for saltiness. Weber fractions thus vary over an order of magnitude across the full range of human sensory systems.

Because the Weber fraction is a dimensionless number, sensory systems can be directly compared on resolving power (it is impossible to perform a meaningful comparison of sensitivity because the absolute thresholds do carry various physical dimensions). Gauged by the Weber fraction, pain is more acute than the perception of length, which in turn is more acute than either brightness or loudness.

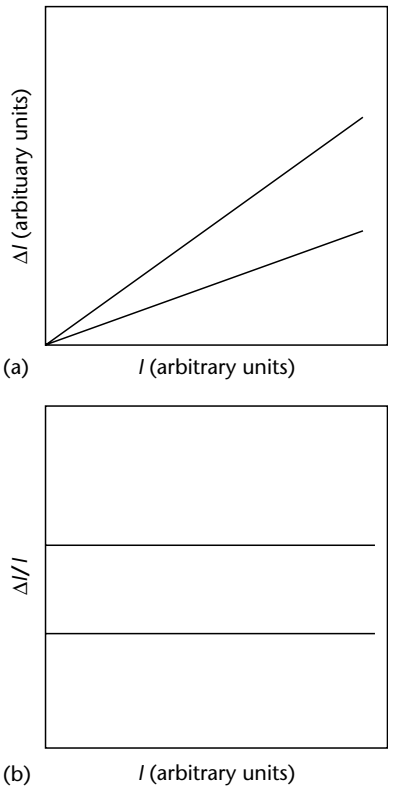


Figure 1. Two characterizations of Weber's law for two sensory continua: (a) the DL (or ΔI) as a function of stimulus intensity I, or (b) the Weber fraction, $\Delta I/I$, as a function of I. In both characterizations, the lower functions mark finer resolving power (i.e. smaller Weber fractions).

Figure 1 shows two characteristic renditions of Weber's law. From its inception in the nineteenth century, Weber's law has been repeatedly tested, and has been to shown to hold remarkably well over most of the dynamic ranges of the respective sensory continua. Deviations occur at very low intensities (in the vicinity of the absolute threshold) and at extremely high intensities. Weber's law remains the oldest, broadest, and most useful empirical generalization in the behavioral sciences.

PSYCHOPHYSICAL SCALING: FECHNER'S LAW AND STEVENS' LAW

Fechner used Weber's relativity principle, namely that DLs are proportional to stimulus intensity, in deriving his psychophysical law, the first explicit, quantitative statement relating sensations to stimuli.

Fechner complemented Weber's law by assuming that all JNDs comprise equal increments in sensation magnitude regardless of the size of the DL in physical units. Therefore, the JND can serve as a unit of sensation. According to Weber's law, pairs of stimuli discriminable by single DLs are separated by different physical increments, although the ratios, I_2/I_1, $I_3/I_2, \ldots I_n/I_{n-1}$, are equal. According to Fechner, these stimulus ratios correspond to equal *increments* in sensation because the respective JNDs are equal in magnitude. As a result, a geometrically spaced series of values on the physical continuum gives rise to an arithmetically spaced series of values on the psychological continuum. This relation defines the logarithmic function, and Fechner's law accordingly is $R = M \log (I/I_0)$, where R is sensation magnitude, M is the constant of proportionality, and I_0 is the threshold. Fechner's law implies that equal *ratios* of stimulus magnitude produce equal *differences* in subjective magnitude; hence, as is apparent in Figure 2, sensation magnitude increases as a negatively accelerated function of stimulus intensity.

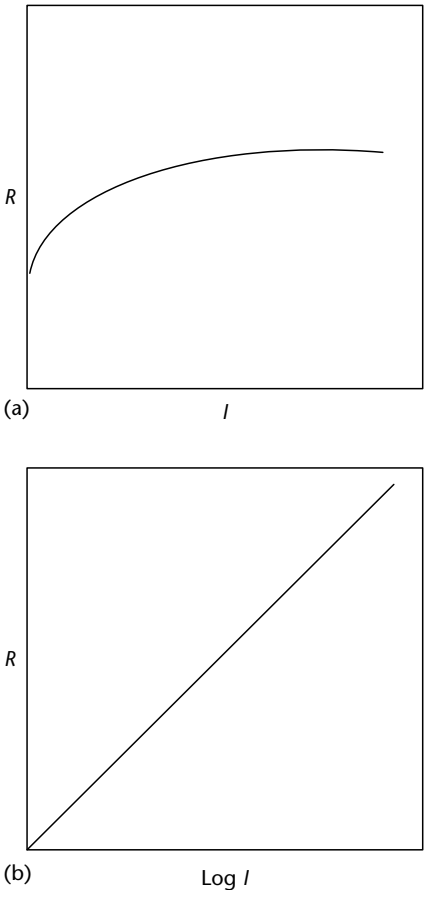


Figure 2. Two characterizations of Fechner's law. (a) Sensation magnitude R increases as a negatively accelerated function of stimulus intensity I. (b) When stimulus intensity is plotted logarithmically the function appears as a straight line.

Following Fechner's law, actual scaling requires determining I_0 and ΔI in the laboratory through the classical psychophysical methods. For suprathreshold scaling, the observer has to arrange stimuli in equal-appearing intervals, or to rate or classify them in equally spaced categories. The hallmark of these methods is that the observer merely matches or orders stimuli on a continuum (by responses such as 'greater', 'smaller', or 'equal'). These responses avoid many of the pitfalls associated with 'direct' numerical responses. Their validity is guaranteed by their extreme familiarity and simplicity.

It is important to distinguish between the laws of Weber and Fechner (there exists no Weber–Fechner law). Weber did not refer to a concept of sensation magnitude or sensation difference in his law. The sole psychological component in Weber's law is the observer's indication of when two stimuli are discriminably different. Yet Weber's law is silent on the crucial question: 'What sensation is felt at a given JND?' It was Fechner who created the notion of sensation magnitude and assumed that all JNDs were subjectively equal, thereby conceiving a truly 'psycho-physical' relation.

Fechner's logarithmic law has had a profound influence in science, from promoting the decibel scale, which is a logarithmic scale of (sound) energy, to aiding in the development of measures of pain such as the dol scale, to quantifying the results of electrical recording from receptors. Nevertheless, the universal validity of Fechner's law has been challenged. For one objection, reports by expert observers – acoustical engineers – judging the relative loudness of sounds did not agree with the logarithmic function. For another, the adaptive value of a logarithmic function for pain is questionable, given the paramount importance of good discriminability at high levels (the logarithmic function, in contrast, implies good discriminability at low levels). Finally, applying direct numerical estimates of the magnitudes of sensations typically yields power functions of intensity rather than logarithmic functions of intensity. Capitalizing on these results, S. S. Stevens has proposed the power function as *the* psychophysical law, replacing Fechner's logarithmic formulation.

According to Stevens, $R = kI^b$, where b and k are constants. The size of the exponent b varies from one continuum to another, subject, of course, to the conditions of stimulation. Exponents can take on values that vary from much smaller than unity (0.33 for brightness) through near unity (1.0 for perceived length) to much greater than unity (3.5 for the perceived intensity of alternating electric current). The power functions for such continua are depicted in Figure 3; the same functions become straight lines when plotted in logarithmically spaced axes because the logarithmic form of the power law is $\log R = \log k + b \log I$. The slopes of these linear functions correspond to the values of the respective exponents.

Assuming that a rule like Weber's law also holds for sensation results in the psychophysical power law. According to Ekman's law, JND is not constant but rather a linear function of sensation magnitude such that $\Delta R = gR$, where g is a constant, Ekman's fraction (which may or may not equal Weber's fraction for a particular modality), and ΔR is the increment in sensation that is just noticeable when added to starting sensation R. Assuming the validity of Weber's law for values of the stimulus and

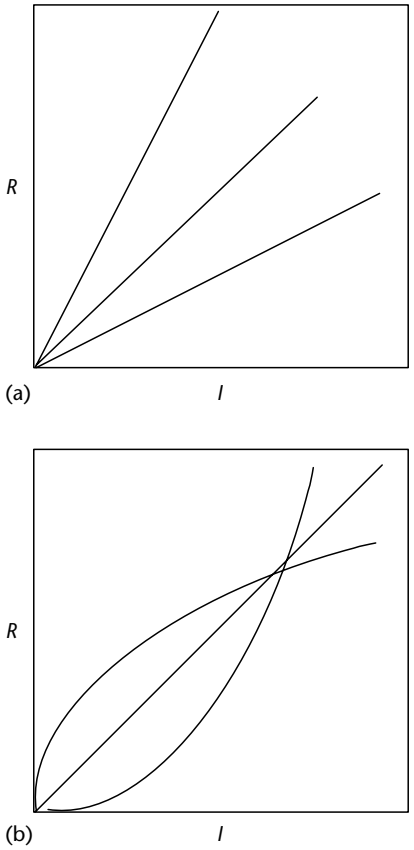


Figure 3. Two characterizations of Stevens' power law. (a) The power functions appear linear on double logarithmic coordinates, with the slopes corresponding to the respective exponents of the power functions. (b) Plotted on linear coordinates, the form of the functions is greatly influenced by the size of the exponent. An exponent of 1.0 corresponds to a linear function; an exponent smaller than 1 corresponds to a concave down function; an exponent greater than 1 corresponds to a concave upward function.

the validity of Ekman's law for values of the sensations, a geometrically spaced series of values on the physical continuum gives rise to a geometrically spaced series of values on the psychological continuum. This relation defines the power function with the exponent b reflecting the ratio of the respective fractions of Ekman and Weber.

In contrast to the Fechnerian approach, the methods of scaling advocated by Stevens are direct. Chief among the latter is *magnitude estimation* in which the observer is required to make direct numerical judgments of the stimuli in proportion to their sensory magnitudes. If one stimulus feels twice as strong as does another, the observer should give them numbers standing in a 2 to 1 ratio. The experimenter may provide the starting stimulus (standard) and assign it a certain numerical value (modulus) such that numbers are assigned to subsequent stimuli relative to the value of the modulus. The most popular method, however, is free magnitude estimation in which stimuli are randomly presented and the observer assigns numbers to them in proportion to the magnitudes of the respective sensations. No standard is provided and the observer is able to establish her or his own modulus. Other direct scaling procedures include *magnitude production*, in which the observer adjusts the level of the stimuli in proportion to numbers called out by the experimenter, and *absolute magnitude estimation*, in which the observer is asked to assign a number to each stimulus in a unique fashion, independent of the values of the other stimuli (hence, the resulting scale is an absolute one that cannot be transformed in any way).

In *cross-modality matching* (CMM) the observer is not required to make numerical judgments. The task is to set the magnitude of sensation in one modality equal to that presented in another modality. For instance, the observer might be asked to adjust the intensity of a vibration on the fingertip to that produced by a sound presented by the experimenter. If power functions are valid descriptions for vibration and for sound, then plotting the matching stimulus levels of the two modalities through CMM results in a power function with an exponent that is the ratio of those for vibration and sound. This result has often been confirmed in the laboratory. Stevens sought to use CMM to validate magnitude estimation and the associated power law. However, results of CMM predictable on the basis of the power law are better considered successful tests of transitivity or internal consistency. A good case can be made for magnitude estimation comprising a special case of CMM in which the observer matches number magnitude to that of the stimuli on the tested continuum. Consequently, any continuum could be substituted for numbers as the standard continuum to measure sensation magnitude on all of the other continua.

CONCLUSION

The implicit assumption that people are able to assign numbers to stimuli in a manner proportional to their inner sensations has been increasingly challenged in modern psychophysics. If it is not the case, then it is illegitimate to treat R as true numbers, let alone introduce them to quantitative, functional relations. Stevens' power law conflates the function relating stimulus intensity to sensation magnitude (the psychophysical function) and the function relating sensation magnitude to the observable verbal response R (the response function). Specifying the former as a power function strongly depends on the unjustified assumption that the latter is linear with zero intercept. The psychophysical function thus is indeterminable when a single stimulus factor is considered, as is usually the case in standard magnitude estimation scaling. Advanced multidimensional models have been developed in modern psychophysics to provide the necessary constraints for deriving and validating the 'true' psychophysical function.

Further Reading

Algom D (ed.) (1992) *Psychophysical Approaches to Cognition*. Amsterdam, Netherlands: Elsevier/North-Holland.

Baird JC and Noma E (1978) *Fundamentals of Scaling and Psychophysics*. New York, NY: John Wiley.

Gescheider GA (1997) *Psychophysics: The Fundamentals*. Mahwah, NJ: Lawrence Erlbaum.

Laming D (1997) *The Measurement of Sensation*. Oxford, UK: Oxford University Press.

Marks LE (1974) *Sensory Processes: The New Psychophysics*. New York, NY: Academic Press.

Marks LE and Algom D (1998) Psychophysical scaling. In: Birnbaum MH (ed.) *Measurement, Judgment, and Decision Making*, pp. 81–178. New York, NY: Academic Press.

Stevens SS (1975) *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. New York, NY: John Wiley.

Quantitative Reasoning

Introductory article

Kristy vanMarle, Yale University, New Haven, Connecticut, USA
Karen Wynn, Yale University, New Haven, Connecticut, USA

CONTENTS

Quantitative reasoning refers to human and non-human animals' sensitivity to and ability to represent different types of numerical information, including both discrete and continuous quantities, and to perform mental operations over these representations.

INTRODUCTION

The ability to mentally represent number and quantity is one of the human species' greatest assets. It has led to the development of our formal system of mathematics, which has enabled the advancement of our species throughout much of its recent history. As an abstract, logical system, mathematics allows us to represent both physical and abstract entities. This article will discuss how numerical information is represented in the human mind, and how the mind uses these representations to carry out basic numerical computations such as simple addition and subtraction. Recent research suggests that our sensitivity to number may be an evolved capacity and one that we share with other animal species.

THE NEUROPSYCHOLOGY OF NUMBER PROCESSING

The psychologist Stanislas Dehaene formulated a model of number processing, which currently dominates the study of numerical cognition. According to this model, humans' number knowledge is comprised of three representational systems, occupying distinct areas of the brain. Although these systems represent numerical information differently, they are interconnected, and function together, under normal circumstances, to support the numerical abilities of humans.

The 'verbal number' system is located in the language area of the left hemisphere of the brain. It represents numerical information linguistically, including both spoken and written number words. It also represents well-learned arithmetical facts, such as addition and multiplication tables learned by rote, and supports multidigit calculations, which require both the recall of memorized facts and the visuospatial representation of numbers. The 'visual Arabic numeral' system, located in the occipito-temporal region of both hemispheres, represents the visual forms of Arabic numerals. Note that not all cultures use the Arabic form of numbers in their written language. Presumably this area of the brain would represent the written form of numerals whether it be Arabic or otherwise. However, as the known data implicating this area of the brain have been tested only with Arabic numerals, we will continue to refer to it here as the 'visual Arabic numeral' system. A 'magnitude' system, located in the inferior parietal cortex regions of both hemispheres, represents numerical information in the form of magnitudes. It gives us our sense of quantity – the meanings of numbers – and therefore supports our ability to compare two numbers, perform approximate calculations, and so on. Empirical findings support the idea that such processes rely on a magnitude representation of number. For example, in the 'distance effect', adults become faster at saying which of two numbers is larger as the proportionate difference between the numbers increases. This is consistent with a magnitude representation of number, which predicts (as explained below) that as the proportionate difference between two numbers becomes smaller, it becomes more difficult to compare them.

Since these three systems are interconnected, activating the representation of, say, 'seven' in the verbal number store automatically activates both the visual Arabic form '7' and the magnitude representation of the number 7. Nevertheless,

neuropsychological evidence clearly indicates that these systems are both functionally and anatomically distinct: they are disrupted by damage to different areas of the brain, and if one system is damaged, the functioning of the remaining two systems is often preserved.

Studies of patients with brain damage support the existence of distinct neural circuitry underlying each of these different systems. For example, there are patients with left-hemisphere injuries affecting their language areas who show deficiencies in understanding number words and performing simple rote calculations such as 'eight plus nine' or 'six times four'. However, they remain able to recognize Arabic numerals, make magnitude judgments (e.g. whether 3 is larger than 7), and even compute approximate results of numerical calculations – for example, judging correctly that '$2+2=9$' is false, while erroneously judging that '$2+2=5$' is correct. In contrast, there are patients with damage to the inferior parietal cortex (where the magnitude number system is located) who can recognize written and spoken number words and Arabic digits, but have lost their sense of magnitude. These patients can recite arithmetic facts, such as multiplication tables, but cannot say, for example, whether 3 is larger or smaller than 7.

The fact that patients with left-hemisphere damage may show impaired ability to recognize and produce verbal number words, yet still be able to judge approximate magnitudes and visually recognize Arabic numerals, suggests that the damaged (but not the preserved) abilities rely on areas of the brain within the left hemisphere. Similarly, the fact that patients with damage to the inferior parietal cortex show impaired performance on tasks requiring magnitude judgments, but retain the ability to recognize verbal and written numbers as well as Arabic numerals, suggests that the damaged (but not the preserved) systems reside in the inferior parietal cortex.

Brain imaging studies with normal adults reveal that tasks involving exact calculation recruit different brain areas from those involving approximate calculation. Specifically, the left and right parietal lobes show more activation for approximate than for exact calculation, while the left inferior frontal lobe shows the opposite pattern. And studies with bilingual subjects indicate that precise number facts are represented in the language in which they are learned, while approximate number facts, which engage subjects' sense of magnitude, seem to be represented in a format that is independent of language.

Taken together, this evidence provides strong support for the existence of distinct systems of numerical knowledge and processes. Note that two of the components in the model, the verbal number system and the Arabic numeral system, consist of knowledge that must be learned. The magnitude number system, however, is thought to be an innate, evolved capacity. If so, we would expect to find evidence of it in non-human animals and prelinguistic infants.

NUMERICAL ABILITIES OF ANIMALS AND HUMAN INFANTS

Many studies have documented extensive numerical abilities across a wide range of warm-blooded vertebrate species. Many animals, including rats, pigeons, parrots, raccoons and chimpanzees, are able to respond on the basis of number of stimuli, whether the stimuli are visual or auditory events, objects in the world, or actions of the animal itself (such as presses of a lever), and whether the stimuli are simultaneously or sequentially presented. Moreover, animals trained to respond to number of one kind of stimuli show generalization to stimuli of other kinds. A model of a magnitude number mechanism was developed by Warren Meck and Russell Church, to account for animals' numerical abilities. On this model, the magnitude number mechanism can be thought of as a small container into which units of water can be added – one unit for each item counted. The subsequent fullness of the container represents the total number of items counted. The discriminability of two numbers relies on the proportionate, rather than the absolute, difference between their values. Consequently, it is easier to discriminate smaller numbers (e.g. 2 and 3) than larger numbers with the same absolute difference (e.g. 12 and 13).

Predictions derived from this model have been tested with prelinguistic humans to investigate the nature of their numerical abilities. The majority of infant studies conducted recently have used looking time as a dependent measure. There are two basic paradigms: habituation, and violation of expectation. In habituation studies, infants are presented repeatedly with a particular stimulus until they become familiar with it. When an infant's looking time decreases, according to a predetermined criterion, the infant is considered to be 'habituated', and is then presented with new instances of the habituated stimulus interspersed with instances of a novel stimulus; the infant's looking time to each is measured. If infants can discriminate between the habituated and the novel stimuli,

then their looking time should recover (i.e., increase) for the novel stimulus, but not for the habituated stimulus.

In violation-of-expectation studies, there is no habituation phase: the infant is simply presented with trials consisting of series of events, the outcome of which is sometimes an 'expected' one (e.g. '$1+1=2$') and sometimes an 'unexpected' one (e.g. '$1+1=1$'). If infants have expectations about the outcome that should obtain in a particular event, then they should look longer at outcomes that do not match their expectation.

If human infants possess the same magnitude number system as do non-human animals, they should show similar numerical abilities. In particular, infants' ability to discriminate two numbers should depend on their proportionate, not their absolute, difference; and infants' numerical abilities should apply to a wide range of entities (auditory, visual, etc.).

Experiments have shown that infants are sensitive to numerical quantity. They can discriminate between both small and large numbers of items, and, as predicted by the model, their ability to discriminate two numbers depends on their proportionate difference – infants can discriminate 2 items from 4 items, 8 from 16, and 16 from 32, but under similar conditions fail to discriminate 8 items from 12 items or 16 from 24. Moreover, their numerical abilities are abstract: infants can enumerate visual objects, collections of objects, events (e.g. jumps of a puppet), and auditory stimuli.

Recent studies recording infants' event-related potentials – which measure activity across different areas of the brain in response to stimuli – reveal that when making numerical discriminations, infants' parietal cortex is highly activated. This is the area of the brain that is responsible for the magnitude sense of number in adults. Further studies have shown that infants can do more than discriminate numbers: their numerical representations can also be used to perform simple numerical operations such as addition and subtraction. When 5-month-old infants are shown an object placed on a stage and then hidden behind a screen, and then shown another object placed behind the screen, they expect two objects to be revealed when the screen is removed, and will look for longer if one or three objects are revealed. Similarly, if shown two objects that are then hidden behind a screen, and then shown one of them being removed, infants expect one object to remain and will look for longer if two are revealed behind the screen.

Similar studies conducted with non-human primates show that they possess similar abilities. Both rhesus macaques and cotton-top tamarins respond as do human infants when shown addition and subtraction situations as described above: they look for longer at incorrect outcomes than at correct ones. For example, when rhesus monkeys were shown one eggplant placed out of sight in a box, and then shown another eggplant also placed in the box, they expected two eggplants to be in the box, and looked for longer if only one was revealed.

ABILITY TO REASON ABOUT CONTINUOUS QUANTITY

Humans – and other animals – can represent continuous as well as discrete quantity: the height of a table, the volume of a piece of cake, the weight of a child, the duration of a sound. Research indicates that the ability to represent continuous quantities is also present from an early age. Infants under 6 months of age can discriminate between objects that differ in their amount of surface area. They can also represent the height of an object, the distance traversed by an object, and the duration of an event. Indeed, some experiments intended as investigations of infants' numerical discriminations (e.g. between different numbers of items) may have inadvertently been testing their ability to measure and distinguish continuous quantities (e.g. total area): in these studies, continuous quantities were confounded with discrete quantities (because a greater number of items always had a greater total surface area). However, recent studies testing infants' numerical abilities while strictly controlling for continuous variables have shown clearly that these abilities are distinct. Infants can represent both continuous quantities, such as the total surface area of elements in a display, and discrete quantities, such as the number of elements in a display.

Thus, from a very early age – long before any formal schooling, and even before any language comprehension – humans have distinct systems for quantitative reasoning. Empirical results with infants suggest at least two basic representational systems for quantity (discrete and continuous), but there may well be many more. It is implausible that all of our representations of continuous quantities are subserved by the same mental structures. For example, we would intuitively expect that our processes for representing an object's speed, the height of an object, the duration of an event, and the weight of an object are quite distinct. However, little is known about the cognitive systems that represent and reason about these kinds of continuous quantities. To obtain a richer understanding of

the full range of humans' quantitative abilities, and the relationships between these different quantitative systems, will require rigorous empirical investigation of these abilities.

CONCLUSION

In conclusion, there is a great deal of evidence supporting at least three components of number processing. Neuropsychological evidence shows that verbal number knowledge, visual recognition of the Arabic numerals, and our sense of numerical magnitude, while interconnected in normal humans, are functionally and anatomically distinct. Moreover, the magnitude number mechanism is special in that it is an evolved mechanism that humans share with other animals, and that is independent of language and learning. It gives us our sense of quantity, allows us to understand relationships between different quantities of items, and underlies our ability to perform simple numerical operations. Empirical findings with adults, prelinguistic infants, and nonhuman animals converge to support this claim. Further research is necessary to determine in more detail the functional parameters of the analog magnitude mechanism, and how all three systems interact throughout development to form human adults' numerical abilities.

Further Reading

Broadbent H, Church R, Meck W and Ratikin B (1993) Quantitative relationships between timing and counting. In: Boysen S and Capaldi E (eds) *The Development of Numerical Competence: Animal and Human Models*, pp. 171–187. Hillsdale, NJ: Lawrence Erlbaum.

Carey S and Xu F (2001) Infants' knowledge of objects: beyond object files and object tracking. *Cognition* **80**(1–2): 179–213.

Dehaene S (1997) *The Number Sense: How the Mind Creates Mathematics*. New York, NY and Oxford, UK: Oxford University Press.

Hauser MD (2000) *Wild Minds: What Animals Really Think*. New York, NY: Henry Holt.

Rational Models of Cognition

Intermediate article

Nick Chater, University of Warwick, Coventry, UK
Mike Oaksford, Cardiff University, Cardiff, UK

CONTENTS

Rational models of cognition attempt to explain the function or purpose of cognitive processes.

CONSTRAINTS ON MODELS OF COGNITION

A scientific explanation of psychological, biological, or social phenomena can take one of two complementary forms. The first is mechanistic: phenomena are explained by analysing their internal causal structures. The second is purposive: phenomena are explained in terms of their purpose, what problems they solve.

In biology, purposive explanation concerns the function of biological structures and processes (e.g. the function of the heart is to pump blood). The same style of explanation is applied to animal behavior (e.g. the function of building nests is to provide a safe shelter for eggs). In the social sciences, 'rational choice' explanation views people as having the purpose of maximizing their 'utility', given the constraints imposed by their environment. Moreover, in everyday life, we explain each other's behavior by giving reasons for why this behavior 'makes sense', given our desires and our beliefs.

In cognitive science, however, mechanistic explanation has been predominant. Computational models, whether symbolic or connectionist, have focused on specifying architectures and algorithms for cognition; and experimental work has been oriented towards mechanistic questions, such as the limits of human memory, or the number of, and interconnections between, memory stores. The picture of the cognitive system that emerges from this focus on mechanistic explanation is as an assortment of apparently arbitrary mechanisms, subject to equally arbitrary limitations, with no apparent rationale or purpose.

By downplaying purposive explanation of cognition, cognitive science may have been missing an essential source of constraints on cognitive models: namely, that in many domains, cogntion appears to be extremely well adapted to the challenges that it faces. In perception, motor control, language processing, common-sense reasoning and decision-making, the cognitive system reliably (though not infallibly) handles perceptual and cognitive problems of great complexity, typically under conditions of uncertainty. The cognitive system can learn to deal with a remarkably broad range of challenges, both natural and artificial, from unicycling to backgammon to musical composition. And the cognitive system acquires, stores and retrieves a rich understanding of the everyday world. It seems plausible that, as for other biological structures, this success is not accidental. It seems more likely that the cognitive system is superbly adapted to serve practical and computational ends. Thus, cognitive models should, ideally, not just fit the empirical data, but also, where possible, make sense as solutions to adaptive problems that the cognitive system faces.

CHARACTERIZING RATIONALITY IN HUMAN COGNITION

Rational models of human cognition aim to explain the function or purpose of human behavior or the cognitive processes underlying it. An idealized methodology for providing such explanation is given in Anderson's (1990) notion of 'rational analysis'. This methodology has six steps:

1. *Goals*. Specify precisely the goals of the cognitive system.
2. *Environment*. Develop a formal model of the environment to which the system is adapted.
3. *Computational limitations*. Make minimal assumptions about computational limitations.
4. *Optimization*. Derive the optimal behavior function.

5. *Data*. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed.
6. *Iteration*. Repeat, iteratively refining the theory.

The idea is that a rational model explains behavior as an optimal (or nearly optimal) attempt (step 4) to achieve certain goals (step 1), in the context of a particular environment (step 2), and with possibly limited computational resources (step 3). The project is empirical, in two senses. First, the goals, environment, and computational limitations can only be determined empirically. Second, the goal of a rational analysis is to explain patterns of empirical data. So, an optimal system for some aspect of categorization or reasoning is only of interest if it captures empirical data on how people do categorize or reason. As with any empirical scientific project, there may be a continuous adjustment of all the elements of the explanation, in order to obtain the most compelling relationship between theory and data (step 6).

How can this 'rational' style of explanation relate to, and potentially constrain, a mechanistic cognitive model? The answer is that the mechanistic cognitive model can implement the computations specified by the rational model (or, at least, some approximation to them). Thus, building a rational model complements, rather than displaces, traditional mechanistic modeling in cognitive science.

Rational models have been developed, more or less independently, in a number of contexts. One tradition, mentioned above, is 'rational choice' explanation, which, in its classical form, assumes that individuals make decisions in order to maximize their expected utility (or, in some biological contexts, to maximize their number of viable offspring). Rational choice explanation is the foundation of modern economics, and has applications in animal behavior, sociology, and political science. In cognitive science, rational models have been developed for specific cognitive processes in perception, categorization, reasoning, problem solving, memory, and language processing, rather than for the whole individual. We will discuss work in this tradition below; related approaches have also been developed independently in the study of vision (e.g. likelihood and simplicity models in perceptual organization, ideal observer models, and the computational level of explanation (Marr, 1982; Pomerantz and Kubovy, 1986)).

Many rational models, including those described below, use a particular theorem of probability, Bayes' theorem. Given two states A and B, the *joint* probability $P(A \,\&\, B)$ is the probability that both A and B are true; and the *conditional* probability of A given B, written $P(A \mid B)$, is the proportion of the probability associated with B that is also associated with A. So by definition, $P(A \mid B) = P(A \,\&\, B)/P(B)$ and $P(B \mid A) = P(A \,\&\, B)/P(A)$. Putting these together, and rearranging, we obtain Bayes' theorem:

$$P(B|A) \;=\; \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

This simple theorem has considerable application, not only in building rational models of cognition, but also in statistics and the philosophy of science.

BAYESIAN MODELS OF CATEGORIZATION

Formulating a rational model of categorization requires specifying a goal or purpose which categorization is presumed to serve. Anderson (1991) makes the natural assumption that the goal of categorization (step 1) is to predict unknown features of objects from known features. He assumes, further, that the environment consists of classes characterized by a probabilistic relationship with a set of features (step 2). Specifically, given a class C_i, the assumption is that for each feature F_j, there is a probability $P(F_j \mid C_i)$ that a item of category C_i has the feature F_j; and, crucially, that this probability is *conditonally independent* of the other features that item has (formally, this means that $P(F_1 \,\&\, \ldots \,\&\, F_n|C_i) \;=\; \prod_{j=1}^{n} P(F_j \mid C_i)$). Here, we shall assume that step 3 is null: no specific computational constraints are needed. Given these assumptions, what is the optimal way of predicting unknown features from known features (step 4)?

Rather than follow Anderson's precise formulation, for clarity we follow a simpler analysis. Suppose we know that an item possesses a set of features $F_1, \ldots F_n$, and want to know $P(C_i \mid F_1 \,\&\, \ldots \,\&\, F_n)$ for each category C_i. That is, we want to know the probability that the item belongs to category C_i. Bayes' theorem gives:

$$P(C_i \mid F_1 \,\&\, \ldots \,\&\, F_n) = \frac{P(F_1 \,\&\, \ldots \,\&\, F_n \mid C_i)P(C_i)}{P(F_1 \,\&\, \ldots \,\&\, F_n)} \tag{2}$$

$$= \frac{P(C_i) \prod_{j=1}^{n} P(F_j \mid C_i)}{P(F_1 \,\&\, \ldots \,\&\, F_n)} \tag{3}$$

where the simplification follows because of Anderson's crucial assumption that features are conditionally independent. Finally, suppose we

want to predict an unknown feature F_{n+1}. If we knew that the item belonged to category C_i, then the probability of F_{n+1} would simply be $P(F_{n+1} | C_i)$. But we know $F_1, \ldots F_n$, rather than the category, so we must predict F_{n+1} by summing these conditional probabilities, weighted by the probability of each C_i, given the known features $F_1, \ldots F_n$. Thus,

$$P(F_{n+1}|F_1 \,\&\, \ldots \,\&\, F_n) = \sum_i P(C_i|F_1 \,\&\, \ldots \,\&\, F_n) P(F_{n+1}|C_i) \quad (4)$$

Bayesian models of categorization, of various forms, have been used to capture empirical data on categorization (rational analysis step 5) (Anderson, 1991), as well being widely applied in artificial intelligence and machine learning.

BAYESIAN MODELS OF BELIEF REVISION

Bayesian models are also widely used in understanding reasoning and belief revision. In artificial intelligence, there has been a substantial shift from logical to Bayesian views of how beliefs should be revised in the light of new knowledge. According to the logical viewpoint, knowledge is encoded as a set of axioms and their deductive consequences. New knowledge (for example, derived from perception or language) is encoded in new axioms; and the new knowledge state consists of the larger set of axioms and their deductive consequences. This approach runs into difficulties where new and old knowledge appear inconsistent, because, in most logical systems, all propositions (and their negations) follow from a contradiction, leading to potential inferential chaos. There have been numerous ingenious attempts to combat this difficulty. But the Bayesian approach aims to avoid it entirely, by assuming that 'knowledge' is only probabilistic – or more accurately, by modeling belief revision in terms of probability theory. In the probabilistic framework, outright contradictions need not occur (what was previously probable simply becomes much less probable). Pearl (1988) and others have shown how to build parallel distributed computational mechanisms for probabilistic reasoning for belief revision. These models depend, crucially, on making independence assumptions between pieces of information, in just the way that we assumed above that features were conditionally independent given the relevant category. For example, effects are typically viewed as conditionally independent given their causes.

A similar shift from logic to probability theory has been advocated in the psychology of reasoning. It has been argued that various apparent experimental demonstrations of irrationality can be reinterpreted. For example, Oaksford and Chater (1994) have argued that searching instances which confirm a conditional rule 'if A then B' is rational from a probabilistic perspective, because a confirming instance can substantially raise the probability that the statement is true. Yet on a traditional viewpoint in the psychology of reasoning, searching for confirmatory evidence is misguided, because general statements cannot be logically derived from their instances – the next observation could always be a refutation. Thus, the human tendency to seek confirming evidence may appear irrational from a logical perspective, but entirely rational according to a Bayesian rational analysis. (*See* **Reasoning**)

EMPIRICAL EVIDENCE FOR AND AGAINST RATIONALITY

The case noted above highlights the difficulty of interpreting empirical evidence for or against rationality: the interpretation depends on the theoretical perspective adopted. But it might seem that rational models of cognition do not usefully contribute to the debate on whether people are rational, because they seem to assume the idea of the rationality of cognition from the outset. The approach seems to presuppose rationality, regardless of any empirical evidence that might be collected. The picture is, however, not so straightforward.

First, the dictates of a rational cognitive model will typically only be implemented approximately. These approximations will result in irrational behaviour. For example, Chater and Oaksford have given a Bayesian rational model of how people reason with syllogisms (e.g., 'all X are Y, all Y are Z, therefore, all X are Z'). Where there is a probabilistically valid conclusion for a syllogism, the heuristics generally generate it successfully; but they also generate other conclusions, giving 'irrational' answers for syllogisms where no conclusion follows.

Second, there is an important distinction between the rationality of specific cognitive processes and the rationality of the whole person, which is comprised of the interaction of innumerable cognitive processes. For example, the tendency of the cognitive system to pay attention to relative rather than absolute magnitudes may be highly adaptive in encoding information about the external world

(because many aspects of the world are 'scale-invariant' (Chater and Brown, 1999)). But this may give rise to irrationality in risky decision-making, where, for example, the difference between prizes of \$0 and \$10 may be viewed as far less significant than the difference between prizes of \$90 and \$100 (Kahneman *et al*., 1982) – even though the differences are objectively the same. In general, we might conjecture that specialized cognitive processes might exhibit greater 'rationality' than the whole individual. This is because specialized processes need only be adapted to some relatively narrow class of tasks (e.g. interpreting stereoscopic disparities between the two eyes, segmenting the visual field) which has been encountered throughout an individual's life, and perhaps also through millions of years of evolutionary history. The whole person, on the other hand, must cope with an endless variety of tasks (e.g. making financial decisions), for which neither experience nor evolution may provide much guidance. If this is the case, then rational choice explanation, as described above, may seek support from human rationality just where it is weakest – a disturbing reflection from the point of view of the foundations of economics.

Third, the attempt to apply rational models of cognition can be viewed as a way of measuring the degree of rationality of the cognitive system. The rationality of thought and behavior can only be assessed against a standard of 'correct' performance. But to choose an appropriate standard of correct performance, we need to have decided what computational function the cognitive system is attempting to perform – and this is the goal of rational analysis. We cannot merely stipulate the standards against which cognition should be measured. If we do so, we run the risk of, for example, condemning people as irrational because they fail to reason logically, when they are reasoning quite rationally according to the dictates of probability, as noted above. Thus, far from presupposing human rationality, the project of building rational models of cognition should provide a test for when and to what degree people are rational.

References

Anderson JR (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson JR (1991) The adaptive nature of human categorization. *Psychological Review* **98**: 409–429.

Chater N and Brown GDA (1999) Scale invariance as a unifying psychological principle. *Cognition* **69**: B17–B24.

Kahneman D, Slovic P and Tversky A (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Marr D (1982) *Vision*. San Francisco, CA: Freeman.

Oaksford M and Chater N (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* **101**: 608–631.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*. Palo Alto, CA: Morgan Kaufman.

Pomerantz JR and Kubovy M (1986) Theoretical approaches to perceptual organization: simplicity and likelihood principles. In: Boff KR, Kaufman L and Thomas JP (eds) *Handbook of Perception and Human Performance*, vol. II 'Cognitive Processes and Performance'. New York, NY: Wiley.

Further Reading

Anderson JR (1991) Is human cognition adaptive? *Behavioral and Brain Sciences* **14**: 471–517.

Anderson JR (1994) *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Cheng PW (1997) From covariation to causation: a causal power theory. *Psychological Review* **104**: 367–405.

Oaksford M and Chater N (1998) *Rationality in an Uncertain World*. Hove, UK: Psychology Press.

Oaksford M and Chater N (eds) (1998) *Rational Models of Cognition*. Oxford, UK: Oxford University Press.

Shanks DR (1995) Is human learning rational? *Quarterly Journal of Experimental Psychology* **48A**: 257–279.

Shepard RN (1987) Towards a universal law of generalization for psychological science. *Science* **237**: 1317–1323.

Reaction Time

Intermediate article

James T Townsend, Indiana University, Bloomington, Indiana, USA

CONTENTS

Reaction time is, along with accuracy, the most important dependent variable employed in experimental cognitive psychology and perhaps in all of experimental psychology.

INTRODUCTION

Reaction time (RT) – also called response time – refers to the time taken by a person (called a 'subject' or 'participant') to perform some task in an experiment. The experimental psychologist must manipulate aspects of the experiment in order to discover important characteristics of behavior and psychological processes, and to test theories or models of performance. Variables in the experiment that are manipulated by the experimenter are called 'independent variables'. Examples include the various aspects of the stimulus (something presented to the participant), the details of the task requirements, and even environmental variables such as the illumination in the experimental chamber.

Reaction time is, in contrast, a dependent variable, because it (like the accuracy of the response) is recorded by the experimenter and used to draw conclusions of psychological interest. Reaction time is used, for instance, to study the characteristics of the psychological system within the participant that is performing various kinds of perceptual, mental and motor tasks. Professional RT analyses depend on an understanding of statistics, experimental design, and often mathematical modeling. This article will attempt only to give a feel for the topic. This entails some sacrifice of detail and precision.

STATISTICAL PROPERTIES OF RT

One of the very first problems that a psychologist must face is that human behavior is virtually always probabilistic, or statistical; that is, it varies according to the laws of chance. This chance, probabilistic, or equivalently, statistical, aspect is present in everything from the behavior of a single neuron all the way up to complex mental operations and actions. Hence, RTs are never the same from trial to trial. The scientist collects a series of them over the experimental session and plots the results in the form of a so-called *frequency function*. The underlying variable is called the 'variate'. In the present case, the variate is RT. The RTs are segregated into small bins, say of about 0.01 seconds each, and then the number of RTs within each time bin is plotted (usually standardized by dividing by the total number of RTs accumulated in the experiment). The result is the RT frequency function.

The *cumulative frequency function*, where the frequencies are summed from the smallest value of RT up to an arbitrary level, is also very useful. Thus, the value of the cumulative frequency function at, say, $RT = 1\,s$, measures how many RTs were at or below 1 second.

These functions show the likelihood of a particular RT for a given experimental condition. They provide the basis for all other analyses and conclusions regarding the experimental results (Wenger and Townsend, 2000).

The *mean*, giving the central tendency of a frequency function, and the *variance*, indicating the variability in the data, are the most important and often-employed statistics. They are used, along with standard statistical assumptions, to investigate hypotheses about mental processes, often employing statistical techniques such as the *t*-test, analysis of variance, and so on. For instance, an experimental group given a special kind of treatment in a memory experiment (e.g., a drug, or extra learning trials) may perform a memory task faster than the control group, as revealed by its mean RT being lower in a statistically significant fashion. Such statistical inference, involving only means and variances, provides the basis for much of modern cognitive psychology and other branches of psychology.

However, there are other statistics or statistical functions of the variate that are considerably more

powerful (Townsend, 1990; Townsend and Ashby, 1983). In fact, there exists a hierarchy of statistical functions organized according to their logical strength. The 'hazard function' reveals the chance that the subject will react in the next instant, given that he or she has not ret reacted. The 'likelihood function' is a ratio of two separate frequency functions evaluated at any particular value of RT. We will next illustrate this and the other statistics within a realistic example.

Suppose that a psychologist is interested in testing the effectiveness of a drug intended to facilitate memory search. Naturally, she wants to make sure the treatment is more effective than no treatment or a placebo. She administers the real drug and a placebo to two different groups, measures their RTs in a memory-search task, and plots the frequency function. Then, the various statistical functions noted – mean, cumulative frequency function, hazard function, and likelihood function – differ in their power to discriminate between the two groups. The mean is, of course, simply the arithmetic average of the sampled RTs. It can be written as $M = \Sigma_i P_i \times RT_i$ where P_i is the proportion of RTs that fall into the i^{th} time bin and RT_i is the reaction time at the i^{th} bin. The greek letter sigma (Σ) simply tells us to sum up the values of the multiplication just to the right ($P_i \times RT_i$). And, the subscript 'i' is the index over which we form the sum. Next, the cumulative frequency function, $F(RT_i)$, is the sum of the frequencies from the smallest time bin up to and including the j^{th}, $F(RT_i) = \Sigma_i^j P_i$, where i runs from 1 up to j, where j is less than or equal to the last index value (standing for the largest value of RT found in the current experiment). The hazard function at time RT_i is just $h(RT_i) = P_i / (1 - F(RT_i))$.

All of these statistics can be compared for the real drug vs. the placebo group. In fact we could write M(DRUG) and M(PLACEBO), $F(RT_i;DRUG)$ and $F(RT_i;PLACEBO)$, and $h(RT_i;DRUG)$ and $h(RT_i;PLACEBO)$, and then see which is bigger, the real drug statistics or the placebo statistics, or if they are basically the same. If M for the real drug group is smaller than that for the placebo group then that suggests that the drug group is performing faster than the placebo group. However, this inference of the real drug group being faster than the placebo group will be supported in an even stronger manner if F is always bigger for the real drug group and even stronger than either M or F, if h is larger for the real drug group than for the placebo group. Finally, the likelihood function is composed of a ratio of P_i for the real drug group over the P_i for the placebo group, which might be expressed as $P_i(DRUG)/P_i(PLACEBO)$. If that ratio, increases as RT grows (the same thing as the index i getting bigger) then that provides the strongest evidence of all for the true drug group being faster than the placebo group, indicating effectiveness of the pharmaceutical. Again, it is important to understand that the likelihood increasing implies the hazard function ordering, that is if $P_i(DRUG)/P_i(PLACEBO)$ increases as i increases, then it follows that $h(RT_i;DRUG) > h(RT_i;PLACEBO)$, where '>' is the 'greater than' relation, for every value of i. Similarly, if $h(RT_i;DRUG) > h(RT_i;PLACEBO)$, then the result that $F(RT_i;DRUG) > F(RT_i;PLACEBO)$ for every value of i is forced to occur. Any of these findings implies that $M(DRUG) < M(PLACEBO)$. Finally, as mentioned earlier, none of these implications works in reverse. For instance, the ordering in h does not force the likelihood function to be increasing and so on. That is what we mean by the indicators differing in the power of what they say about the data. Such results are not only important for ordinary inference about experimental conditions: internal cognitive mechanisms make predictions that permit their testing against other hypothetical mechanisms only if their variables are ordered in a relatively strong way (Townsend, 1990a).

MODEL TESTING

Model testing can be roughly divided into two categories, which we shall discuss. We must begin by defining the terms 'parameter' and 'free parameter'. A parameter is a variable in a quantitative model that must be given an exact value in order for the model to make numerical predictions. A simple example is the prediction that the dependent variable y is a linear function of the independent variable x: $y = ax + b$. Here, the slope a and the intercept b are the parameters. Before they are given exact values, they are considered free parameters. After being assigned values, they are no longer free.

Fitting the Model to the Data

The most common approach to model testing involves estimating free parameters so that the predictions of the model are as close as possible to the data. Then, the fit of the model – that is, how well the model predicts the data – is assessed.

There are almost no cases in psychology where parameters can be assigned numerical values before running the experiment. With some methods of fit, the assessment of how well a

model predicts the data can be done quantitatively, with the null hypothesis of the predictions not being significantly different from the data being explicitly tested. However, in a number of important cases, there is no way to statistically test the degree of fit. Another problem with fit strategies is that the more sparse the data (e.g., fewer trials or subjects), the more likely the model is to fit acceptably, even if the method of fit allows a statistical test. This makes it difficult to adequately test a model. The reasons are fundamentally the same as for the fact that a small sample size gives less power to reject the null hypothesis in ordinary statistical inference. One helpful strategy is to compare the fits of two or more competing models. This is especially useful where no statistical assessment of fit is available.

Within the model-fitting approach, one must distinguish between the simple positing of a particular probability or frequency function for the processing time of one or more hypothesized subtasks, and a processing model that is based on higher-order psychological hypotheses. Frequency functions that have been proposed for processing times, and sometimes for the complete RT probability law, include the 'ex-Gaussian' (the convolution of an exponential with a Gaussian frequency function), the exponential itself, and the gamma or general-gamma (identical with a convolution of exponentials with the same or different parameters, respectively) (Luce, 1986).

When constructing a process model for a psychological phenomenon, one may employ particular frequency functions in the model, but usually one also has to hypothesize an architecture of separate individual sub-processes, rules of interaction, and input and output. One of the most basic questions is whether processing is serial or parallel – that is, are any two sub-processes operating at the same time, or only one at a time? For instance, it has been proposed that when people search their memories for specific information, the search involves examination of each of the individual items in memory one at a time – in other words, serially (e.g., Sternberg, 1969). However, others have suggested that instead, this type of search might take place in a parallel fashion (see, e.g., the review in Townsend (1990)). In this type of search, all the items could be examined simultaneously, that is, in parallel.

Suppose two sub-processes are arranged in a serial fashion with processing durations T_1 and T_2 and that there is also a set of other residual subprocesses that we can combine into the single variable T_0. Note that these must be random variables, since RTs are never perfectly constant from trial to trial. Then, the overall RT can be expressed as $\mathrm{RT} = T_0 + T_1 + T_2$. If, in addition, these random variables are probabilistically independent, then the frequency function of their sum is formed by a mathematical operation called the *convolution* of the individual frequency functions. Furthermore, it can be shown that the mean or expectation of RT is just $\mathrm{E(RT)} = \mathrm{E}(T_1) + \mathrm{E}(T_2) + \mathrm{E}(T_0)$.

Suppose, on the other hand, that the two subprocesses of interest are processed in parallel, but again the residual times are in series with them. Then the total time for completing both, plus the residual time, is the maximum of the two separate times plus T_0; that is, $\mathrm{RT} = \max(T_1, T_2) + T_0$, and $\mathrm{E(RT)} = \mathrm{E}(\max(T_1, T_2)) + \mathrm{E}(T_0)$. Thus, we have already formed two possible models for a mental task that involves two separate sub-processes. Of course, there are other issues or questions about system architecture and processing that can arise in model construction (e.g., Townsend and Ashby, 1983), and the theorist may need to form models of much greater complexity (e.g., Schweickert, 1978; Schweickert and Townsend, 1989).

Qualitative Testing

The second, somewhat less common, approach to model testing is to explore and determine qualitative characteristics that are predicted by a model and then to probe whether and to what extent these are exhibited by the data. By 'qualitative' is meant, for example, inequalities, the general form that functions or graphs should take, and other relationships among various aspects of data. Suppose we are interested in testing a model of choice-responding that says that the mean RT increases with the number of choice alternatives n, in proportion to the logarithm of n plus a constant. That is, $\mathrm{E(RT)} = k\log(n) + C$. Now, one might actually fit this simple model to the RT data and assess the fit, or one could start by checking a qualitative prediction of the model: whether the data curve is concave down, like the logarithm function (and many other functions). If the data curve is not concave down, then the logarithm model (as well as many other models) is rejected (disconfirmed). If the curve is concave down, then it may be reasonable to attempt more precise data fits.

A more powerful kind of qualitative distinction can be found in the discussion above of various types of statistical functions. For instance, a model might predict that the cumulative frequency functions are ordered in a certain way. In fact, it has been found that in memory search experiments, the cumulative frequency functions are indeed ordered in

terms of the number of items that must be searched in memory (e.g., Townsend, 1990b, Figure 3). This ordering implies that in a relatively strong statistical sense (stronger than would follow from an ordering based on mean RT alone, for instance), more items do indeed take longer to search through. Interestingly, it turns out that both serial models and many parallel models can make this prediction.

There is one qualitative approach that was developed in the late 1960s and has evolved into a rather impressive arsenal of related techniques intended to reveal the underlying architecture. The approach began as the 'additive factors method' (Sternberg, 1969; Sanders, 1983). The basic idea is to assume that processing is serial and that each sub-process in a series of sub-processes engaged by a certain task may be associated with an experimental factor that can lengthen or shorten the duration taken by that sub-process to finish its part of the task in the series. For instance, an early sensory sub-process might be affected by stimulus intensity, while a late motor sub-process might be affected by response difficulty. (*See* **Information Processing**)

In an additive factors study, the experimenter manipulates two or more factors intended to target certain sub-processes of interest. Because of the assumed seriality the manipulation of the experimental factors should influence RT in an additive fashion. We can then write $\mathrm{RT}(x_1, x_2) = T_1(x_1) + T_2(x_2) + T_0$, and we have the prediction that $\mathrm{E}(\mathrm{RT}) = \mathrm{E}(T_1(x_1)) + \mathrm{E}(T_2(x_2)) + \mathrm{E}(T_0)$. This prediction can readily be tested in the RT data by standard statistical procedures. If additivity is found then the conclusion is that there do indeed exist two psychological sub-processes, that they operate in series, and that they are suitably affected by the experimental factors. On the other hand, if additivity is not found, then one or more of the experimental factors must affect more than 'its own' sub-process (e.g., stimulus intensity might affect both an early and a late sub-process), or that there are not two separate serially-arranged sub-processes.

The additive factors method was extended over the years by a number of authors to include parallel processes as well as much more complex mental networks (e.g., Schweickert and Townsend, 1989). The point to be made here is that the predicted additivity by serial models, and various kinds of non-additivity by other types of architectures, are all of a qualitative nature. It is not necessary to know exactly what the RTs are, only that, for instance, they obey the additivity property. Of course, it is possible, once one has affirmed, say, the additivity, to attempt to fit particular serial models to the numerical data.

It appears that RT will continue to be one of the most important dependent variables used in studying cognitive processes.

References

Luce RD (1986) *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Sanders AF (1983) Towards a model of stress and human performance. *Acta Psychologica* **53**: 61–97.

Schweickert R (1978) A critical path generalization of the additive factor method: analysis of a stroop task. *Journal of Mathematical Psychology* **18**: 105–139.

Schweickert R and Townsend JT (1989) A trichotomy method: interactions of factors prolonging sequential and concurrent mental processes in stochastic PERT networks. *Journal of Mathematical Psychology* **33**: 328–347.

Sternberg S (1969) Memory scanning: mental processes revealed by reaction time experiments. *American Scientist* **57**: 421–457.

Townsend JT (1990a) The truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. *Psychological Bulletin* **108**: 551–567.

Townsend JT (1990b) Serial vs. parallel processing: sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science* **1**: 46–54.

Townsend JT and Ashby FG (1983) *The Stochastic Modeling of Elementary Psychological Processes*. Cambridge, MA: Cambridge University Press.

Wenger MJ and Townsend JT (2000) $H(t)$ and $C(t)$: basic tools for attention and general processing capacity in perception and cognition. *Journal of General Psychology: Visual Attention* **80**: 67–99.

Further Reading

Ashby FG, Tein JY and Balakrishnian JD (1993) Response time distributions in memory scanning. *Journal of Mathematical Psychology* **37**: 526–555.

Myung IJ, Forster MR and Browne MW (eds) *Journal of Mathematical Psychology* **44**(1). [A technical survey of various important approaches to model testing.]

Smith PL (1995) Psychophysically principled models of visual simple reaction time. *Psychological Reviews* **102**: 567–593.

Townsend JT (1990) Serial vs. parallel processing: sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science* **1**: 46–54.

Van Zandt T (2000) How to fit a response time distribution. *Psychonomic Bulletin and Review* **7**: 424–465.

Van Zandt T and Ratcliff R (1995) Statistical mimicking of reaction time distributions: mixtures and parameter variability. *Psychonomic Bulletin and Review* **2**: 20–54.

Reading, Psychology of

Introductory article

Simon Garrod, University of Glasgow, Glasgow, UK
Meredyth Daneman, University of Toronto, Toronto, Canada

CONTENTS

The psychology of reading investigates the process by which readers extract visual information from written text and make sense of it.

INTRODUCTION

Reading is the process by which a reader extracts visual information from a piece of written text and makes sense of it. Psychologists are interested in questions such as how readers extract this visual information, what writing is, how it relates to speech, and precisely how a reader makes sense of the text during reading. Because children have to be taught to read in a way that they do not have to be taught to speak, psychologists are also interested in how we learn to read and what underlies individual differences in reading skill.

Consider the following passage:

> Clyde did not want to arouse suspicion. So he sat down in the waiting room with his hand over his holster and smiled politely at the other occupants of the room. He thumbed deliberately through the heaps of reading matter on the table until he spotted the latest *Newsweek*. He opened the magazine and then carefully counted the number of bullets it held, waiting to be fired. When the office door opened, Clyde was poised and ready. The magazine was full.

This passage is made up of a series of graphical symbols (letters, words, sentences) that represent a word-by-word transcription of its spoken form. Although all writing systems use graphical symbols to represent the spoken language, they do so in different ways. For example, in the Japanese Kana writing system, the symbols correspond to whole syllables (e.g. the *wait* + the *ing* sounds in 'waiting'), whereas in English, letters or groups of letters correspond to smaller sound units or *phonemes* (e.g. the *w ai t i ng* sounds in 'waiting'). These different systems pose different problems for readers and those learning to read.

WRITING SYSTEMS AND ORTHOGRAPHIC DEPTH

The three main kinds of writing system are *logographic* systems (as in Chinese and Japanese Kanji), *syllabic* systems (as in Japanese Kana), and *alphabetic* systems (as in English, Greek, and Russian). Logographic systems use symbols to represent word meanings, so the symbols – called 'characters' in Chinese – do not map directly onto any sound segments in the spoken language. As a result, readers have to learn thousands of different characters – almost one for each word – to become literate in Chinese. Syllabaries and alphabets have symbols that map onto whole syllables in the first case or phonemes in the second. Because there are many more syllables than phonemes, syllabaries tend to have more symbols (typically 100) than do alphabets (typically 25 to 30). However, even with alphabetic systems, such as written English or Italian, the transparency of the mapping between the letters in the alphabet and the spoken phonemes varies considerably. For example, in Italian there is a very direct and consistent mapping between *graphemes* (letters and patterns of letters) and phonemes (individual speech sounds), whereas in English there is not. Across English words, the same combination of letters may map onto a range of distinct speech sounds (compare the *ou* sound in the words c*ou*nted and thr*ou*gh in the Clyde passage; now consider en*ou*gh and th*ou*gh).

There are various reasons for this difference between the written languages. One reason is that languages like English have tended to import foreign words and retain their original spellings (e.g. 'villa', 'centre'). This means that the grapheme-to-phoneme mappings may be inappropriate for English pronunciation. Another reason is that in some written languages the spelling sometimes reflects the *morphemic* or meaning structure of words

as opposed to their pronunciation. For example in English, 'vine' and 'vineyard' share the same letter pattern *VINE*, which corresponds to the common meaning component of the two words. However, this pattern of letters is pronounced quite differently in the two spoken words. The degree to which the written language reflects morphology as opposed to phonology is called *orthographic depth*. Written languages like Italian have shallow orthographies, whereas written English has a deep orthography. In general, it is more difficult to learn to read in languages with deep orthographies as opposed to shallow orthographies. Also, alphabetic systems tend to be more difficult to learn than syllabaries. (*See* **Reading and Writing**)

PHONOLOGICAL RECODING DURING SILENT READING

Most readers are aware of covertly pronouncing printed words as they read. Many even report experiencing an inner speech that has the appropriate stress, pauses, and intonation patterns for the text being read. The process of translating print to sound has been called *phonological recoding* (also *phonological encoding*, *speech recoding*, *subvocalization*, *inner speech*). It has been argued that phonological recoding limits reading speed to the rate at which we can covertly articulate the words. Indeed, commercial speed reading courses lure their clients by boasting that they can dramatically increase reading speed at no cost to comprehension simply by teaching people to eliminate phonological recoding. However, some psychologists would argue that phonological recoding serves an important, or even necessary, information-processing function in silent reading so that preventing or limiting inner speech would have an adverse effect on reading comprehension. (*See* **Phonological Encoding of Words**)

Psychologists have proposed two possible roles for phonological recoding. According to the first view, readers recode a printed word into the way it sounds as a way of gaining access to the meaning of that word held in memory. So for example, by applying spelling-to-sound rules, the letters *s-a-t* could be mentally converted into their corresponding phonemes /s/ /æ/ /t/, and the phonological code /sæt/ could then be used to access the word's meaning, much as is done during listening comprehension. Evidence for the use of phonological codes to access word meanings comes from demonstrating *homophone* (i.e. words that sound the same) confusion effects. For example, if readers use sound to access meaning, they should be more likely to mistakenly judge a phrase to be sensible if it sounds correct (e.g. 'He thumbed deliberately threw the heaps of reading matter …') than if it does not sound correct ('He thumbed deliberately throw the heaps of reading matter …'). Psychologists agree that *prelexical* phonological codes (i.e. codes accessed before recovering the meaning of the word) are particularly important for beginning readers and unskilled readers who will frequently encounter words that are unfamiliar to them in print (but familiar if recoded into sound). However, psychologists are in some disagreement as to whether fluent adult readers generate prelexical phonological codes during silent reading, or whether they access the meanings of words directly on the basis of the words' visual features. Some researchers argue that skilled readers use the direct visual pathway, others argue that skilled readers use the indirect phonologically mediated pathway, and yet others argue for dual (alternate) pathways to the lexicon.

Regardless of the extent to which skilled readers engage in prelexical phonological recoding, at least some of the inner speech of a skilled reader must be generated *after* lexical access. This is because there is no way to know the appropriate stress and intonation patterns without first interpreting the sentence at several levels, such as determining what the individual words mean and doing at least some preliminary analysis of how the words are semantically and syntactically related to one another. The second view of phonological recoding is that phonological codes are generated after lexical access to help the reader to retain information in working memory (temporary memory) long enough for the higher-level semantic integration to occur. During reading, sequences of words must he held in a temporary storage buffer while the comprehension processes integrate them into a meaningful conceptual structure that can be stored in a more permanent memory.

For example, when first encountering the Clyde passage, most readers probably initially interpreted the word 'magazine' to mean 'reading periodical' because this is the meaning that was strongly primed by the preceding context ('… he spotted the latest *Newsweek*'). However, this meaning is inconsistent with the subsequent text ('the number of bullets it held, waiting to be fired'), and a resolution of the apparent inconsistency requires a reinterpretation of 'magazine' to mean 'bullet chamber'. If recently processed information cannot be stored at least temporarily, the reader would be continually backtracking to reread parts or even whole sentences and passages.

It has been argued that the most stable code in working memory is a sound-based one. By generating speech-based (phonological) codes that are less vulnerable to memory loss, the reader can keep track of exact words rather than rough meanings. Thus, according to this second view, phonological recoding plays an important role in reading by facilitating the storage and integration of successive ideas in a text. Evidence for this view comes from demonstrating that comprehension suffers when readers are required to engage in a concurrent task that interferes with the generation or maintenance of phonological codes in working memory. For example, repeating an irrelevant word such as 'cola, cola, cola' during reading has adverse effects on comprehension, whereas engaging in an equally effortful nonverbal finger-tapping task does not. (*See* **Working Memory**)

METHODS FOR STUDYING READING

There are two approaches to the study of the reading process itself. The first is to investigate what is remembered after reading a text. This can tell us what information a reader gleans from a text and how long it is retained. For example, memory tests show that readers quickly forget the precise wording of a text as they read it, yet they retain the gist. Memory-based measures of reading are called 'offline measures' because they cannot address directly the moment-by-moment decisions a reader makes as he or she is actually reading. For this reason offline measures are generally used only in combination with other more direct measures to investigate the reading process itself.

The second approach is to measure the time readers spend looking at the words and sentences as they read them. Recording a reader's eye-movement patterns is the most direct method. Less direct methods use self-paced reading procedures, which let the participant determine the rate at which written material is presented. The reader might be required to pace himself or herself sentence by sentence, phrase by phrase, or word by word. For example, in the word-by-word procedure, a word is presented, and as soon as the reader has understood it, he or she presses a key to trigger presentation of the next word. The sequence is then repeated until all the text has been read and the time taken to read each word is recorded. Eye-tracking and self-paced reading produce 'online measures', because they reflect the moment-by-moment processing of the text.

During reading the eye moves in a systematic way. There are brief fixations in which gaze stays on the same letter for between 100 and 450 milliseconds. These are interspersed with fast movements called 'saccades' during which the gaze moves to another letter or word of the text. Typically, saccades take only 30 milliseconds to execute and change the position of gaze by about eight letter spaces in the text. For a skilled reader, nine out of ten saccades move the gaze from left to right to sample new material from the text, whereas one out of ten saccades return the point of gaze to previously read material (these are called 'regressions'). Thus, skilled readers typically view all the content words in a text and typically view them in the order that they would occur if spoken. The duration of fixations and the length and direction of saccades (forward or backward movement of the gaze) directly reflect the ease or difficulty of the reading process. Furthermore, they indicate the precise word in the text that is causing reading difficulty because attention is given only to the word currently fixated. (*See* **Visual Attention**; **Eye Movements**)

The limited span of attention during reading can be demonstrated using the *moving window* technique, in which a computer program controls dynamically the window of text presented to the reader as a function of where the subject is fixating. For example, with an asymmetric 12-character window, the four characters to left and the eight characters to the right of where the reader is fixating will be displayed as normal, whereas all the remaining text will be converted into random letters. The window of text together with its surround of random letters then changes as the point of fixation changes (see Figure 1). The window changes do not interfere with normal reading because they occur during saccadic eye movements when no information is taken in from the visual stimulus. As a consequence, readers do not notice that the text in front of them is changing as their eyes move across it.

With the moving window technique, one can reduce the size and form of the text window and measure when it begins to affect reading rate. It turns out that normal reading is quite possible when the window contains only the word currently fixated plus the first three letters of the next word on the line. However, there is a proviso that the material around the window must retain the spaces between the words in the original text. When the window arrangement is reversed so that the window contains random letters and the surround contains the normal text, readers encounter difficulty. With a reverse window of only 11 letters in width, reading becomes almost impossible.

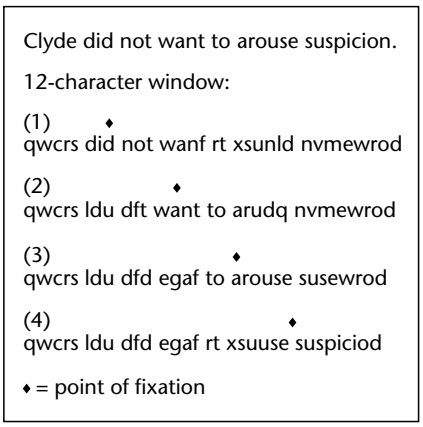


Figure 1. Sequence of presentations for an asymmetric 12-character moving window.

Moving window studies indicate that readers take in information from only a very limited region at any time during reading. This means that any extra time spent fixating the region must reflect processing difficulty associated with that region of text or previous regions not completely processed but still held in memory. Current theories of eye-movement control assume that the programming of where the fixation is to land next is determined on the basis of information about the gross shape of words and spaces to the right of the fixation. This explains why it is important to retain the spaces outside the moving window. In contrast, the decision as to when to launch the eye movement is determined by more immediate processing considerations, such as recognition of the word currently being fixated.

Eye-movement studies confirm that the basic identification of each word and its meaning often occurs during the very first fixation on the word. For example, the first fixation duration is affected by the frequency of the word and whether or not it is ambiguous.

SEMANTIC INTERPRETATION AND INTEGRATION

Moving window experiments demonstrate that words are identified and semantic information extracted before the eye moves on to the next word in a text. But is all this information integrated into the current interpretation immediately or are some processes delayed? The relationship between sampling and processing time is captured by what is called the *immediacy hypothesis*. This states that each word in a text is processed to the deepest level possible (in relation to its meaning, how it relates to the context, and so on) before the reader goes on to the next word.

In general, this hypothesis has been supported. Evidence for immediate contextual effects on semantic interpretation comes from eye-tracking studies in which people read ambiguous words in different contexts. Consider, for example, interpreting the ambiguous word 'magazine' in the Clyde passage. 'Magazine' has a dominant meaning (periodical) and a less dominant meaning (bullet chamber). Eye-tracking studies show that when the prior context favors the less dominant meaning (bullet chamber), readers spend longer fixating the word than when the context favors either the dominant meaning or neither meaning in particular. This suggests that the contextually appropriate meaning of the word is identified before the reader moves on to the next word in the text. Other findings relate to interpreting what are called *anaphors*. These are expressions such as the pronoun 'he' in the Clyde passage whose interpretation depends upon something that has been previously mentioned in the text (i.e., the antecedent mention of 'Clyde'). The difficulty of establishing the appropriate antecedent is reflected in the time spent reading the anaphor itself and the words that immediately follow it. This indicates that at least some aspects of the interpretation of the anaphoric pronoun occur as soon as it is encountered in the text. (*See* **Anaphora, Processing of**)

However, there is also evidence that some semantic processing and integration is left until the reader has completed a whole clause or sentence of text. Hence, readers tend to spend a little extra time looking at the words at the end of a sentence before proceeding to the beginning of the next sentence in the text. This process is sometimes referred to as 'sentence wrap-up'. (*See* **Lexical Ambiguity Resolution**)

INDIVIDUAL DIFFERENCES IN READING ABILITY

There are large individual differences in how well people read. Some adults can read only 150 words a minute whereas others can read 400 words a minute or more. Differences in comprehension ability can be just as large. Good readers not only understand the literal facts in a passage, but they also make the appropriate inferences, attend to how the passage is organized, and appreciate the author's tone and style. By contrast, poor readers may read an entire passage without understanding or retaining even the main point. What accounts for the enormous differences in how fast and how

accurately people can read? Because reading is a complex cognitive skill that draws on many component processes and resources, any of the component processes has the potential for being a source of individual differences in reading ability. Many – but not all – are. Here we consider some factors that do and do not account for the range of reading ability differences that might be encountered in a typical school or university classroom.

Eye-movement control is one component of reading that does not appear to account for individual differences in reading ability. It is not that poor readers display the same pattern of eye movements and fixations as good readers. On the contrary; they make more and longer fixations than do good readers, as well as many more regressions. However, training poor readers to make the eye-movement patterns of good readers does not lead to improvements in their comprehension. Consequently, the erratic and inefficient eye movements of poor readers are thought to be the *result* rather than the *source* of their reading problems. Low-level visual-perceptual processes also do not appear to account for individual differences in reading ability because good and poor readers do not differ in the amount of information they can extract during a single eye fixation. Indeed, almost all reading problems are due to difficulties in recognizing words and comprehending language.

There is relatively strong evidence that word recognition skills contribute to overall reading ability. Poor readers are slower and less efficient at recognizing written words, slower at accessing word meanings from memory, and less skilled at deriving phonology from print. However, the relationship between word recognition skills and reading ability is much stronger for young readers than it is for adults. For example, the speed with which readers can access word meanings from memory accounts for only about 10 percent of the variance in reading ability found in a typical university classroom, and is more related to reading fluency than to reading comprehension.

On the other hand, the higher-level language comprehension processes common to both reading and listening account for much more of the variance in reading ability that one finds in a typical university classroom. Poor readers are at a particular disadvantage when they are required to execute a process that involves integrating newly encountered information with information that was encountered earlier in the text or that must be retrieved from long-term memory. For example, poor readers have problems interrelating successive topics and integrating information to derive the overall gist or main theme of a passage. They have more difficulty identifying the antecedent referent for a pronoun (e.g. determining that the pronoun 'it' in 'it held' refers to the magazine of the gun that Clyde opened); they have more difficulty making inferences (e.g. that Clyde's holster contained a gun; that he removed his hand from the holster); and they tend to make fewer thematic inferences spontaneously during reading (e.g. that Clyde may have wanted to murder someone, that he was in a doctor's office, or perhaps a lawyer's office). All in all, poor readers tend not to demand informational coherence and consistency in a text, and they often fail to detect, let alone repair, semantic inconsistencies (such as the inconsistency between the initially accessed 'reading periodical' meaning for 'magazine' and the subsequent phrase 'held, waiting to be fired…').

Two mechanisms have been proposed to account for why poor readers have difficulty with the integration processes of reading: *working memory capacity*, and *background knowledge*. According to working memory theories of reading ability, poor readers are at a disadvantage at all of the processes that require the integration of newly encountered information with information encountered *earlier in the text* because they have less capacity to keep the earlier information active in temporary storage. According to knowledge-based theories of reading ability, poor readers are at a disadvantage at all of the processes that require the integration of newly encountered information with information that must be *retrieved from long-term memory*, either because they have less background knowledge of the topic being read, and/or because they have less ability to access that knowledge when required. (*See* **Language Comprehension and Verbal Working Memory**)

The existence of individual differences in reading ability has far-reaching educational implications. Because reading is the major medium for acquiring knowledge and skills, poor readers will experience difficulty not only in a literature class, but also in classes as diverse as history, economics, and science.

DYSLEXIA

Dyslexia is the term applied to individuals who have a much more severe reading disability than the poor readers described in the previous section, in the sense that these individuals struggle to decode or recognize even the simplest of words.

Individuals who were previously competent readers but who suffered an impairment of that ability due to brain injury are said to have *acquired dyslexia*. Individuals who failed to attain normal reading skills in the first place are said to have *developmental dyslexia*. (*See* **Dyslexia**; **Developmental Disorders of Language**)

Acquired dyslexia comes in a number of different forms. Individuals with *surface dyslexia* appear to have damage to the direct visual pathway to the lexicon, and so they have to rely on the indirect route of assembling the phonological code by applying spelling-to-sound (grapheme-to-phoneme) conversion rules. They can read regular words (e.g. 'sat' and 'hand') and pronounceable nonwords (e.g. 'flum' and 'pib'), but they have difficulty reading words with irregular spelling patterns (e.g. 'thumbed', 'through'). On encountering the irregular word 'listen', one surface dyslexic misread it as 'liston' (pronouncing the *t* which should be silent), and then added 'the famous boxer' (referring to Sonny Liston, a former boxing champion).

The fact that surface dyslexics have a tendency to regularize irregular words (e.g. to say 'liston' for 'listen') and then to ascribe meaning to the words based on how they sound rather than how they look (to interpret 'listen' as 'Sonny Liston, the boxer') is consistent with the idea that surface dyslexics rely on the indirect phonological route to the lexicon. In contrast, individuals with *phonological dyslexia* appear to have lost their capacity for grapheme-to-phoneme conversions and so they are unable to read pronounceable nonwords and unfamiliar words. However, their direct visual pathway is intact, so they can read (and understand) regular and irregular real words, as long as the words are already familiar to them.

Like phonological dyslexics, *deep dyslexics* have lost the capacity for grapheme-to-phoneme conversion (they cannot read nonwords), and must rely on the direct visual route (they can recognize very familiar words). However, deep dyslexics are thought to have multiple sources of damage because they have a number of other difficulties as well. These include difficulty reading familiar function words (e.g. 'the', 'in'), difficulty reading abstract words (e.g. 'suspicion'), and the propensity to make striking semantic substitution errors (they might read 'ape' as 'monkey' or 'blood' as 'pressure').

It has been estimated that approximately 4 percent of school-aged children have developmental dyslexia in that they have severe difficulty learning to read despite adequate intelligence, vision, and opportunity to learn. Many people are under the misconception that developmental dyslexia is a visual disturbance that manifests itself in the tendency to read letters and words backwards (e.g. to read *d* as *b*, 'was' as 'saw'). However, there is little evidence to support this notion because dyslexic readers make reversal errors no more frequently than beginning readers do. There is still considerable disagreement among psychologists as to the nature of the deficit(s) underlying developmental dyslexia. Nevertheless, there is an emerging consensus that developmental dyslexics do not constitute a homogeneous population, but rather fall into a number of distinct subgroups. Indeed, some researchers have drawn parallels to the different subgroups of acquired dyslexia. For example, approximately 60 percent of developmental dyslexics are like the phonological dyslexics described earlier, in that they have difficulty assembling a phonological code by applying grapheme-to-phoneme conversion rules. However, there are other developmental dyslexics who are more like the surface dyslexics described earlier, in that they are relatively competent at reading phonologically regular words but have severe problems reading irregular words.

CONCLUSION

Reading is a complex skill and particularly so with alphabetic written languages such as English. Although skilled readers can process as many as six words a second, phonological recoding imposes an upper limit on reading rate. Even skilled readers fixate almost all the words in a text and can attend to only one word at a time. Yet, each word is identified rapidly, and, in most cases, its meaning is integrated immediately into the meaning of the passage as a whole.

Since reading is such a complex skill for which we have to receive instruction, it is perhaps not surprising that there are striking individual differences in reading ability. Individual differences arise both within the normal range of readers and in relation to those with specific disabilities such as dyslexia.

Further Reading

Daneman M (1991) Individual differences in reading skills. In: Barr R, Kamil ML, Mosenthal P and Pearson PD (eds) *Handbook of Reading Research*, vol. 2, pp. 512–538. White Plains, NY: Longman.

Garrod S and Sanford AJ (1994) Resolving sentences in a discourse context: how discourse representation affects language understanding. In: Gernsbacher M (ed.)

Handbook of Psycholinguistics, pp. 675–698. New York, NY: Academic Press.
Just MA and Carpenter PA (1987) *The Psychology of Reading and Language Comprehension*. Newton, MA: Allyn & Bacon.
Klein RM and McMullen PA (1999) *Converging Methods for Understanding Reading and Dyslexia*. Cambridge, MA: MIT Press.
Rayner K and Pollatsek A (1989) *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice-Hall.

Reasoning

Introductory article

Mike Oaksford, Cardiff University, Cardiff, UK

CONTENTS

Reasoning is a mental process by which given information is transformed into a new and more useful form.

WHAT IS REASONING?

Reasoning is an ever-present feature of human life. People are so dependent on the processes involved that they tend to go by unnoticed. But it is simple to demonstrate how much of human behavior depends on reasoning processes. Suppose you see your neighbor arriving home. She passes her garage and looks in to see that the car is gone. When she reaches the door, instead of ringing the doorbell, she takes out her key and opens the door. Why your neighbor broke her habitual pattern of behavior can be explained in an instant: she sees that the car is gone; she therefore infers that someone has driven it away; because she knows that only her partner has the keys, she infers that her partner has driven it away; she further infers that if he is in the car he is not in the house and hence he would not open the door if she rang the bell; consequently she takes out her key and opens the door herself.

In this example, the given information is derived directly from perception (the car was gone) and from prior knowledge (e.g. if the car is gone someone has driven it away). This information is combined, in an inference, to yield new information: someone has driven the car away. The given information can be regarded as the premises, and the new information as the conclusion, of a passage of reasoning. The subsequent steps that lead her to the final conclusion that she must use her key to get in to the house can all be characterized in the same way. For example, the second step involves the inference from the premises

Only if you have the keys you can drive the car.
Only her partner has the keys. (1)

to the conclusion

Her partner drove the car. (2)

So it would seem that even the most mundane passages of human behavior involve complex reasoning processes that involve using given information (premises) to infer new information (conclusions). Since Aristotle, this ability to reason has been taken as a defining quality of humans: humans are rational animals.

However, not every process of moving from given information to new information is reasoning. For example, conclusion 2 above could be replaced with 'Her partner had a cream tea'. This may be new information for her, but it does not seem to be related to the premises in the right kind of way. So the process that leads to the conclusion in a passage of reasoning must depend in some rational way on the premises: if you believe the premises then somehow you must believe the conclusion. Reasoning may be studied in terms of formal logical systems, which mark the beginnings of a true science of cognition. (*See* **Inductive Reasoning, Psychology of**)

FORMAL LOGICAL SYSTEMS

Systems of formal logic began to be developed in the mid-nineteenth century to provide a mathematical theory of how mathematicians should reason. Mathematics is about manipulating symbols according to rules. Thus, someone following the rules of arithmetic knows that if they see '2 + 2' then they can replace those symbols with '4'. This rule can be applied without knowing what the symbols '2' and '+' mean. Systems of such rules are called *formal* systems. Mathematicians and philosophers like Boole, Frege, and Russell realized that the same formal treatment could be applied to passages of reasoning such as that described above. So using the symbol p to stand for 'the car is gone' and q to stand for 'someone has driven it

away', the first inference in the example can be expressed formally as 'if p then q; p; therefore q'. This logical rule is called *modus ponens* (MP). Another similar rule is: 'if p then q; not q; therefore not p'. (Thus, if the car is not gone, you can infer that no one has driven it away.) This is called *modus tollens* (MT). (*See* **Deductive Reasoning**)

These rules can be justified by what the sentences mean. The sentence 'if the car is gone, someone has driven it away' is only definitely false when although the car is gone, no one drove it away. In general, assuming that 'if p then q' is false only when p is true and q is false (and is true otherwise), the formal logical rules MP and MT must be truth-preserving: that is, if the premises are true then the conclusion must be true. Formal logical systems provide a standard against which reasoning can be measured: logic provides a theory of the inferences people should and should not make.

This account of formal logic resolves the problem alluded to above. The process that leads necessarily from premises to conclusion is truth-preserving. The truth of 'if the car is gone, someone has driven it away' and 'the car is gone' convey no information about the truth of 'her partner had a cream tea'; so that conclusion cannot be inferred from these premises alone.

Early psychologists researching human cognition assumed that human adults reasoned logically. Indeed, Piaget placed formal logical reasoning at the pinnacle of his stage theory of cognitive development: according to that theory, logically competent adults began to emerge at about age 11. However, since the work of Peter Wason in the early 1960s, research into the psychology of reasoning has frequently produced results that seem to cast doubt on the ability of human adults to reason logically.

ARE PEOPLE LOGICAL?

Research into the psychology of reasoning has proceeded by developing tasks that have apparently straightforward logical solutions. People's performance on these tasks can then be compared with the logical solution. The most famous such task is Wason's selection task.

Wason's Selection Task

In the selection task, people assess whether evidence is relevant to the truth or falsity of a conditional rule. In the abstract version, the rule concerns cards that have a number on one side and a letter on the other. For example, a typical rule might be 'if there is an "A" on one side (p) then there is a "2" on the other side (q)'. Four cards are placed before the participant, so that just one side is visible, showing an 'A' (p), a 'K' ($\neg p$), a '2' (q) and a '7' ($\neg q$) (Figure 1). Participants select those cards they must turn over to determine whether the rule is true or false.

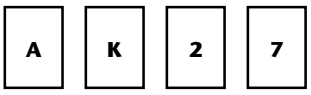


Figure 1. The four cards in a version of Wason's selection task. Each card has a letter on one side and a number on the other. The rule to be tested is: 'if there is an "A" on one side then there is a "2" on the other side'. Subjects must select which cards they need to turn over to determine whether the rule is true or false.

According to logic, only a card with an 'A' on one side but without a '2' on the other side would render this rule false. There are only two cards that could possibly be of this type: the 'A' card could have a number other than '2' on the other side, and the '7' card could have an 'A' on the other side. So people should select the 'A' and the '7' cards to turn over, but not the 'K' or the '2' cards.

These selections are rarely observed in the experimental results. Typical results are: 'A' and '2' (46%); 'A' only (33%); 'A', '2', and '7' (7%); 'A' and '7' (4%), other (10%). Thus, only 4% of participants make the response predicted by logic. Participants typically select cards that could confirm the rule, i.e., the p and q cards. But the choice of the q card is illogical. This is an example of 'confirmation bias'. It would appear that people are not usually logical.

Content Effects

Further experiments have shown that more logic-like performance is observed when real-world content is used. People seem to make more logical (p and not-q) card selections for rules like:

If Johnny travels to Manchester, he takes the train. (3)

If you use a second-class stamp, you must leave the envelope unsealed. (4)

If you are drinking beer, you must be over 21 years of age. (5)

However, this seems at odds with the idea that people use a formal logic that applies independently of content; i.e. regardless of what the symbols mean.

Moreover, not all contents facilitate the logical response. Rules like 3 lead to logical responses less reliably than rules like 4 and 5. The reason for this appears to have little to do with logic. Rules like 4 and 5 are 'prescriptions' for how people should or should not behave; they are not descriptions of the world or of how someone behaves in the world. Discovering that Johnny has traveled to Manchester by car may cast doubt on the truth of rule 3, but finding someone who is drinking beer under the age of 21 does not cast doubt on the truth of rule 5. Rule 5 is in force regardless of the number of people found to be violating it.

Rule 3 is an example of an 'indicative' conditional; rules 4 and 5 are examples of 'deontic' conditionals. Only deontic conditionals appear to reliably produce logic-like performance. However, this is only because the task is no longer one for which standard logic provides the solution. Thus, it still appears that people are not logical.

Everyday Reasoning

Many of the inferences that people need to make in their everyday lives do not conform to logic. Consider the first inference in our example. The premises are

If the car is gone, someone has driven it away.
The car is gone. (6)

The conclusion is

Someone has driven it away. (7)

According to logic, if the premises are true then the conclusion must be true; i.e. the conclusion cannot be false. But this seems to be wrong. For example, although the premise in (6) seems reasonable, the car may have been towed away, disassembled, or even driven away by a trained chimpanzee. Any one of these additional pieces of information would defeat the conclusion that someone drove the car away. Everyday inferences like this are called 'defeasible'. Defeasible inference is at odds with logic, in which additional information cannot overturn the conclusion of a valid inference. (*See* **Frame Problem, The**; **Non-monotonic Logic**)

THEORIES OF REASONING

In the cognitive science of human reasoning there have been many theoretical responses to these findings, each seemingly guided by a different insight. The main theoretical approaches are described below.

Abstract Rule Theories

There are several different 'abstract rule' theories, also known as 'mental logic' theories. These accounts are close in spirit to Piaget, who argued that normal adult human thought was just the operation of formal logic. According to abstract rule accounts, logic and symbolic computation must be at the heart of a well-specified theory of reasoning. However, in modern accounts, people need not possess all the logical inference rules. This should not limit the set of inferences that can be drawn, though some inferences might be more complicated than they would be with a complete set of rules. Furthermore, modern accounts appeal to the distinction between the logic and the control procedure that decides which rules to apply and when. For example, in constructing a mathematical proof it is up to the mathematician to make these decisions. A proof may be shorter or more elegant, depending on the order in which the rules are applied. A computational model of human reasoning must model this flow of control. These control procedures do not have to be as rigid as the logical rules themselves. They might include heuristics (rules of thumb) that suggest applying certain rules at certain stages of a proof.

Thus, abstract rule theories have considerable scope for variation in how they explain people's reasoning behavior. In the Wason selection task (Figure 1), people try to apply logical inference rules to the card sides that they can see to predict what logically should be on the other, unseen side. If people possessed both the inference rules MP and MT, they would turn the p card ('A') expecting to see a q card ('2') by MP, and they would turn the $\neg q$ card ('7') expecting to see a $\neg p$ card by MT. However, if people did not possess the MT rule, then they would only select the p card, as do about 33% of participants.

People often interpret conditionals as 'biconditionals'. For example, they may interpret 'if you have the keys you can drive the car' as entailing also 'if you can drive the car you must have the keys'. This amounts to testing two rules: 'if p then q' and 'if q then p'. The logical rules MP and MT can then also be applied to the second conditional. With only the MP rule available, participants will now select the p and the q cards, as do about 46% of participants. Thus, abstract rule theories appear to be able to explain the main pattern of responses in the selection task.

Abstract rule theories may appear to be less able to deal with content effects and everyday reasoning; but this is not necessarily the case.

Logicians have formulated abstract rule theories that deal with the 'modal' terms, such as 'must', that feature in deontic conditionals. Everyday inference appears more problematic because it seems to rely on an ability to draw inferences using large amounts of world knowledge stored in long-term memory. Logically this means that very large numbers of premises may be involved, leading to an exponential increase in computational cost. Abstract rule theories generally avoid this problem by suggesting that everyday inference is really a form of 'inductive' inference, to do with changing one's beliefs about the world, rather than 'deductive' inference which is dealt with by logic.

Mental Models

Mental models also assume that people are at least in principle capable of logical reasoning. However, rather than reason by applying formal rules to mental sentences, mental models theory argues that people reason over pictorial representations of what those sentences are about. These representations are then manipulated in a search for counterexamples to the conclusion they initially seem to support. For the conditional, this means representing the conditions where it is true. Because people have a limited 'working memory' capacity, they do not represent the complete meaning of a conditional; rather, they have a preferred initial mental model that may be 'fleshed out' if required. (*See* **Mental Models**)

Figure 2(a) shows the initial representation for the conditional 'if p then q'. The three dots indicate that the model may be incomplete. The square brackets indicate that p cannot be paired with any term other than q. Figure 2(b) shows the fully fleshed-out representation, with all the conditions that make the rule true. Figures 2(c) and 2(d) illustrate the corresponding representations for the biconditional.

According to mental model theory, in Wason's selection task, given representation 2(a) only the p card should be turned (because it must have a q on it). The q card need not be turned because it may be paired with other terms that are not represented. However, given representation 2(c) both the p and the q cards should be turned. Given representation 2(b) the p and the $\neg q$ cards should be turned; and given representation 2(d) all four cards should be turned. Thus, mental models theory appears to be able to explain performance in the selection task.

Mental models theory has also been applied to reasoning with deontic conditionals. This involves marking certain mental models as necessary or possible to capture the modal terms used in the rules. Whether mental models can apply to everyday reasoning is less certain. It is argued that the construction of the initial model is determined by world knowledge, and that moreover it may be altered to accommodate additional information. But the processes involved are not clear. This is problematic because recent experiments appear to show that people usually only construct an initial model. If this is the case, then most of the action in human reasoning would appear to be in the construction of this initial model and not in the subsequent manipulation of mental models.

Although mental models may be less precise than abstract rule theories, they give good coverage of the data. Indeed, mental models theory seems to provide the most plausible current account of a range of phenomena in human reasoning. The fact that mental models often only represent what is true leads to the prediction of a range of 'illusory' inferences that people make apparently because of failing to consider the cases that can make a claim false. Mental models theory is currently the most active area of research in human reasoning. (*See* **Spatial Representation and Reasoning**)

Domain-specific Theories

Domain-specific theories are largely inspired by the observation of content effects. The fact that people seem to reason better with deontic conditionals

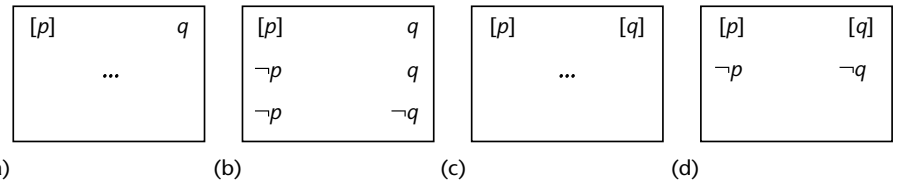


Figure 2. Mental models for the conditional 'if p then q'. Three dots indicate that the model may be incomplete. Square brackets around a term indicate that the term cannot be paired with any term other than the one shown; '¬' indicates 'not'. (a) A simple representation of the conditional. (b) A fully fleshed-out representation of the conditional. (c) A simple representation of the biconditional. (d) A fully fleshed-out representation of the biconditional.

than with indicative conditionals is taken to suggest that people possess domain-specific reasoning mechanisms. There are two varieties of this type of theory. 'Social contract theory' takes its lead from evolutionary psychology. To be part of a social group requires conforming to rules of behavior like rules 4 and 5 above. According to social contract theory, because reasoning about social contracts was so important in early human evolution an innate module evolved to deal with this type of reasoning. One source of evidence for this view is the early emergence of correct deontic reasoning in children. (*See* **Evolutionary Psychology: Theoretical Foundation**)

According to 'pragmatic reasoning schema theory' people possess many domain-specific schemata for reasoning. Each of these schemata contains rules that operate only in its specific domain. For example, rules 4 and 5 both have the same underlying form: they relate a precondition for performing an action to the action. A pragmatic reasoning schema for deontic reasoning contains rules concerning how to draw inferences about preconditions and actions. Such a schema is only triggered if the antecedent and consequent of a conditional can be interpreted as a precondition and an action, respectively. Pragmatic reasoning schemata are generally considered to be learned rather than as innate.

Neither of these two theories has much to say about the standard Wason selection task using indicative conditionals, or about everyday reasoning. This may be because they derive from other theoretical frameworks whose primary concerns lay elsewhere. Social contract theory is seen mainly as a contribution to evolutionary psychology, and researchers in this area are not primarily concerned with creating general theories of reasoning. Pragmatic reasoning schema theory was originally part of a more general theory of inductive reasoning. In the future these theories may be extended to deal with more phenomena in human reasoning. However, one problem is that they do not adopt a theory of the inferences people should make, such as that provided by logic in abstract rule theories and in mental models. Subjects' responses are not shown to be 'correct' (or 'incorrect') according to some theory of the domain (such as logic), so it is difficult to see how they can be generalized to new contexts of reasoning.

Dual Process Theories

Dual process theories have a long history. They suggest a two-way partition in reasoning abilities. This is somewhat similar to the distinction between deductive and inductive reasoning involved by some abstract rule theories. Typically, dual process theories suggest that we do have a (perhaps limited) ability for explicit logical reasoning, which may be embodied in a mental logic or in mental models. However, a lot of reasoning goes on implicitly and is independent of these logical processes. In one version of this type of theory, reasoning in the abstract selection task is largely heuristic. It has been observed that if a rule like 'if there is an "A" on one side (p) then there is not a "2" on the other side ($\neg q$)' is used, people still select the 'A' (p) and the '2' (q) cards; i.e. they appear to ignore the negation. It is argued that people are guided by a heuristic that says that '2' is still the topic of the consequent 'there is not a "2" on the other side', so, regardless of the negation, the '2' card is still selected. This makes sense in everyday discourse, where mentioned information is often still the topic even in a negated sentence: for example, the lateness of the train is still the topic of 'the train is not late'.

Another distinction has been drawn between two types of rationality. According to this account, people are rational in one sense when their reasoning conforms to a normative standard like logic. They are rational in another sense when they reason so as to achieve their goals in the world, regardless of whether their reasoning conforms to a normative standard. Different mental processes are involved in these forms of reasoning.

Much of the evidence seems to support the view that some reasoning is under the guidance of abstract rules while other reasoning is achieved by processes that can be captured in 'neural networks'; i.e., networks of massively interconnected small processors, which imitate neurons in the brain.

According to these theories, most of the reasoning seen in Wason's selection task, content effects, and everyday reasoning are under the guidance of implicit cognitive processes. Perhaps the main problem with this approach is that the nature of the implicit reasoning processes is left obscure. Although one suggestion is that neural networks are involved, no neural network model of any of the passages of reasoning they are supposed to explain currently exists.

Probabilistic Theories

Probabilistic theories start from the intuition that everyday inference cannot be based on logic. According to these accounts, the problems of everyday reasoning arise because people have to

reason about the uncertain world in which they live. Consequently, rather than assume that people are trying to reason logically but have imperfect mental mechanisms for inference, as in abstract rule theories and mental models, it is argued that people are reasoning according to a probabilistic standard. For example, rather than applying logical rules such as MP and MT, people will endorse inferences to a degree depending on the probability of the conclusion given the categorical premise. For MP, this is the conditional probability of q given p ($P(q/p)$). Conditional probabilities for all other inferences can be derived from $P(q/p)$, $P(p)$, $P(q)$. (*See* **Reasoning under Uncertainty**)

Thus, selecting the p and the q cards in the Wason selection task may be the rational thing to do. The only assumption that needs to be made is that $P(p)$ and $P(q)$ are small (this is called the 'rarity assumption'), which has received independent verification. According to probabilistic theories, each card has some probability of being selected, which varies with $P(p)$ and $P(q)$. This can account for the major patterns of card selection. (*See* **Rational Models of Cognition**)

Probabilistic theories have also been applied to content effects. Here it is argued that the materials influence how much people value certain outcomes, like finding people who are trying to violate a rule. The expected value of looking at a card can then be calculated using what is called 'subjective expected utility' theory. This account again predicts the main pattern of results.

Probabilistic theories are relatively new. However, like mental models and abstract rule theories, they have been generalized to the main inference tasks investigated in the psychology of reasoning. These theories deal directly with the uncertainty of everyday inference. Moreover, probability theory may be the right language in which to connect high-level theories of reasoning with the operation of neural networks and ultimately with neural processes in the brain. However, the mental processes that actually implement such models have yet to be specified; and the empirical scope of this approach is currently limited, at least in comparison with mental models accounts. (*See* **Reasoning and Thinking, Neural Basis of**)

CONCLUSION

Reasoning is a vital human ability. It involves transforming given information into a new and more useful form. Computational approaches to cognition provide a mechanistic account of reasoning. However, results in the psychology of reasoning seem to show that people do not reason logically. There have been a variety of theoretical approaches to explain this, each of which begins from some intuition. Some deal with problem cases by drawing distinctions between different kinds of reasoning. It remains to be seen which of these intuitions is the most important in accounting for human reasoning, and hence which theoretical framework will prevail.

Further Reading

Braine MDS and O'Brien DP (1998) *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum.

Evans JBT and Over DE (1996) *Rationality and Reasoning*. Hove, UK: Psychology Press.

Fiddick L, Cosmides L and Tooby J (2000) No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition* **77**: 1–79.

Johnson-Laird PN and Byrne RMJ (1991) *Deduction*. Hove, UK: Lawrence Erlbaum.

Manktelow KI (1999) *Reasoning and Thinking*. Hove, UK: Psychology Press.

Oaksford M and Chater N (1998) *Rationality in an Uncertain World*. Hove, UK: Psychology Press.

Rips L (1994) *The Psychology of Proof*. Cambridge, MA: MIT Press.

Sloman SA (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* **119**: 3–22.

Sperber D, Cara F and Girotto V (1995) Relevance theory explains the selection task. *Cognition* **57**: 31–95.

Religious Thought

Intermediate article

E Thomas Lawson, Western Michigan University, Kalamazoo, Michigan, USA

CONTENTS

Religious thought is a type of cognitive processing that exploits, and is parasitic on, ordinary cognitive resources by minimally modifying standard conceptions of action and agency. Its cultural transmission is determined by its symbolic evocativeness, its subtle combination of intuitive and counter-intuitive features, and its inferential potential.

INTRODUCTION

While a number of social scientists, psychologists, socio-biologists, philosophers, historians, and cultural critics have occasionally turned their attention to an analysis and explanation of religious thought in the last hundred years or so, only recently have cognitive scientists begun to examine religion. Earlier inquirers in the tradition of William James and his heirs (see Etzel Cardena *et al*., 2000) have focused on the nature, causes, and consequences of religious experience. These studies take such experience to be primarily an affective and emotional matter, quite often far removed from rational thought and in a significant sense antithetical to it. Religious thought, from this point of view was nothing more than a reflection on religious experience, often serving as a rationalization. The real story was to be found at the noncognitive level. One recent example of the 'experiential approach' can be found in the work of James H. Austin (1999), who has investigated the brain processes that putatively account for extraordinary (mystical) religious experiences, often viewed as altered states of consciousness. While such research appears to place the examination of religious thought on a firm scientific footing, it does not have any obvious connection with the widespread generation and reception of religious ideas. Nor does it tell us much about the practice of religious ritual acts, even by those who lay no claim to such mystical encounters.

More recent work in cognitive science has analyzed religious thought through a variety of concepts:

- as an example of an autonomous symbolic mechanism,
- as a system for the representation of action coupled with a conceptual scheme,
- as a set of concepts emerging out of an intuitive ontology that is memorable because of its minimally counter-intuitive properties,
- as a transmissible set of concepts whose transmission is made highly probable because of its frequency correlated with its emotional intensity,
- as an instance of anthropomorphism generated by a hyper-active agency detector,
- and as a difference between on-line and off-line thinking.

Recently experimental studies have put some of these claims to empirical test.

SYMBOLIC THOUGHT

Hints of a cognitive approach to religious thought that focuses upon its symbolic character start with the groundbreaking work of the cognitive scientist Dan Sperber (1975) who argued that religious thought was one aspect of a more general cognitive process of symbolism which he labeled 'creative competence'. Sperber regards symbolism as an autonomous mechanism that, together with perceptual and conceptual processes, contributes to the production of knowledge and the dynamics of memory. Going against the semiological grain fashionable at the time, Sperber argued that symbolism is a typical instance of an inferential process common to all people that is neither analytic nor synthetic. In effect he was claiming that symbolism is a perfectly natural 'instinct'. From Sperber's point of view, symbolism, of which religious thought is an instance, is to be examined as a set of cognitive processes with quite specific products related to but distinct from what he has called 'dictionary' and 'encyclopedic' knowledge. The major objective of Sperber's approach has been to emphasize the continuities between ordinary modes of thinking and symbolic thinking. His

point has been to show that the inferential processes that lead to religious and other symbolic representations, rather than being radically different from the products of our ordinary cognitive resources, engage the same inferential engines that are called into play when presented with half-understood beliefs. Sperber argues that religious thought is a quite specific cognitive process of evocation that triggers specific memories scattered across a number of domains and connects them to each other in quite unusual ways. According to Sperber, then, when human minds are exposed to certain cultural phenomena such as statues of deities, strange behavior, or apparently bizarre beliefs – phenomena which fail to meet the criteria of either definition or description – a symbolic process of evocation becomes activated which establishes unique connections among the representations encoded in human memory. These connections are not required by either dictionary definitions or encyclopedic descriptions, but restless human minds interpret them anyway. By doing so they make the beliefs plausible, the behavior relevant, and the objects salient.

THE REPRESENTATION OF ACTION AND RITUAL FORM

Inspired by Sperber's theory of symbolism, Lawson and McCauley (1990) established the cognitive science of religion by developing an analogy between competence theorizing in linguistics and ritual competence in religious contexts, thus giving theoretical and empirical bite to Sperber's more general notion of 'creative competence'. They argue that just as people possess a tacit knowledge of the structure of their language, so religious participants possess a tacit, noncultural knowledge of their religious ritual system. This tacit knowledge is demonstrated by the people's ability to make judgments about religious ritual forms even when they do not consciously entertain questions about their structure, how well formed they are, their efficacy, repeatability and their reversibility, and clearly have not been instructed in these features by members of their society. Some of Lawson and McCauley's claims have recently been experimentally confirmed (Barrett and Lawson, 2001).

Lawson and McCauley argue that while cognitive scientists have paid a great deal of attention to questions about the cognitive representation of agency (Leslie, 1995), they have paid far less attention to representations of action. Lawson and McCauley think that representations of action are crucial for understanding not only religious ideas but the actions that follow from them. Hence they propose that human beings possess a set of cognitive processes collectively known as an 'action representation system' which, when coupled with a culturally variable conceptual scheme (which provides the content in the action representation system for agent, action, and patient with all their various properties), is sufficient to generate a set of formal descriptions of religious ritual structure.

Lawson and McCauley claim that there are three basic types of ritual forms: special agent, special instrument, and special patient rituals (McCauley and Lawson, 2002). They also argue that ritual actions are systematically connected to each other. Almost invariably any particular ritual act presupposes acts already performed: for example, integration as an adult into a specific cultural community presupposes prior rituals such as initiation and marriage. So the structural description of a ritual would have to include the 'embedded' rituals. A ritual's full structural description would contrast with its immediate, surface description. Whereas in ordinary reasoning such 'embedding' can go on indefinitely (involving either causal reasoning or concatenation), religious reasoning, while engaging the same representational resources, typically involves an endpoint. The 'buck stops' with the gods. Lawson and McCauley also show that rituals in which the 'buckstopping agent' (a culturally postulated superhuman agent) is the actor rather than the patient turn out to be crucial for understanding the differences among rituals. These differences in ritual form play a crucial role in accounting for aspects of cultural transmission.

INTUITIVE ONTOLOGY

Pascal Boyer (1994) has been investigating the underlying properties of the conceptual scheme that informs the action representation system by showing how the variability of religious concepts can be explained by paying attention to the underlying intuitive ontology in terms of which all human beings interpret the world. Boyer argues for the 'naturalness' of religious ideas. Religious ideas are natural in the sense that they exploit our ordinary cognitive machinery for representing the world. He shows how the operation of the intuitive ontology that all human beings employ in their traffic with the world can quite comfortably explain the generation and transmission of religious ideas. These operations involve the violation of aspects of the default assumptions (by either breaching them or transferring them from one

category to another) that our ontological categories typically possess. For example, the category 'artificial object' normally is thought of as having physical properties but not biological or mental qualities. Boyer shows that by transferring the property of intentionality normally associated with the category of 'person' to the 'artificial object' category, we can develop the concept of an artificial object, such as a statue that listens to prayers, that has a psychology. Or, given the category 'person', the default assumptions involving intentionality, biological characteristics, and physical features – violating only the physicality assumption (while maintaining biological and intentional properties) – generate the concept of a person who can be thought of as being alive and having intentions but no body: namely, a spirit. Boyer's point is to show that with only minimal alterations our ordinary ontological concepts are readily available for generating religious concepts. Such minimally counter-intuitive representations are different enough to be interesting and therefore transmissible. Multiple violations, on the other hand, would make the concepts more difficult to represent and, therefore, be less likely to be transmitted.

THE FREQUENCY HYPOTHESIS

Harvey Whitehouse, a cognitive anthropologist who has been doing fieldwork in Papua New Guinea, also has evinced an interest in developing a cognitive account of religious thought. He is particularly interested in how religious systems change, how such changes can be accounted for, what role memory plays in the transmission of religious concepts and the behavior they inform, and the processes by which such concepts are encoded. In Memorable religions (1992), he signaled his interest in giving a cognitive account of cultural transmission. He suggests that frequency is an important variable that accounts for aspects of the transmissibility of religious concepts. The more frequently a religious ritual is performed, accounting for its transmission becomes less of a problem; while if a ritual is performed less frequently, emotional factors are more likely to be involved in ensuring its transmission. His fieldwork seems to confirm his hypothesis and in *Inside the Cult* (1995), he presents a full account of his reasoning.

ANTHROPOMORPHIC THOUGHT

Stuart Guthrie (1993), a cognitive anthropologist, hypothesizes that people explain gods by a hyperactive agency detection device bequeathed to humans by evolutionary forces. He thinks that being able to postulate unseen, hidden agents with very special qualities provides an evolutionary advantage. The interpretation of events in anthropomorphic terms, i.e. as the work of superhuman agents, demonstrates a 'better be safe than sorry' attitude that can enhance the possibility of survival. Hence it is quite understandable that humans not only see faces in the clouds but interpret ordinary events in anthropomorphic ways.

While Sperber, Lawson and McCauley, Boyer, Whitehouse and Guthrie depend primarily on ethnography for empirical support, cognitive psychologists interested in religious thought but with experimental inclinations, and aware of the work of the above-mentioned cognitive scientists, have begun to subject such claims to empirical tests in controlled studies.

EXPERIMENTAL EXPLORATIONS

Cognitive psychologists Justin Barrett and Frank Keil (1996) turn from theoretical analysis to experimental exploration of the empirical dimensions of the cognitive science of religion by proceeding to demonstrate experimentally in a rather clever way that an important distinction needs to be made between theological and religious thought. They demonstrate that theological thought is a far more abstract, 'off-line', mode of reasoning, which requires conscious reflection and explicit instruction for its transmission from generation to generation (much like scientific knowledge). Barrett and Keil show that religious thought is anthropomorphic, 'on-line', and much more directly related to the way that people generally conceive of agents and actions. Thus religious thought is much more like folk science. Their experiments demonstrate that even when people have developed highly sophisticated theological conceptions of deity (including such puzzling notions as atemporality and omniscience), they nevertheless, when required to make rapid judgments, tend to reason about the properties of the gods in the same way that they reason about the properties of the rest of us. The point of the Barrett and Keil study is to show that our ordinary, psychological processes constrain our religious representations. This is one of those areas of thought where not 'anything goes'. Even the gods are represented as being temporally constrained.

Since these initial empirical and experimental studies, further cross-cultural investigations by cognitive psychologists, cognitive anthropologists, and some comparative religionists are being

pursued in order to ensure that investigators overcome unconscious biases about human modes of thought based upon Western examples. Particularly notable for their empirical and experimental cross-cultural studies of the structure of religious thought and its modes of cultural transmission are those of Boyer and Ramble (in press) and Barrett (1998).

Boyer and Ramble's study shows that participants more accurately recall items which violate category-level assumptions than those items which conform to these assumptions. In addition they demonstrate that there is very little cross-cultural difference in the sensitivity to these assumptions.

Barrett's (1998) replication in India of the Barrett and Keil (1996) study shows that whether the subjects are in northern New York or northern India, representations of deities are subject to the cognitive assumptions that govern all intentional agents. Even though the explicit concepts attribute to the deities nontemporal properties endorsed by a given theological tradition, people in both cultural situations nevertheless, in their everyday, 'on-line' reasoning situations, attribute to the gods the same psychological and physical qualities that they attribute to ordinary intentional agents.

References

Austin JH (1999) *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. Cambridge, MA: MIT Press.

Barrett JL (1998) Cognitive constraints on Hindu concepts of the Divine. *Journal of the Scientific Study of Religion* **37**: 608–619.

Barrett JL and Keil FC (1996) Conceptualizing a non-natural entity: anthropomorphism in god concepts. *Cognitive Psychology* **31**: 210–247.

Barrett JL and Lawson ET (2001) Ritual intuitions. *Journal of Cognition and Culture* **1**(2): 183–201.

Boyer P (1994) *The Naturalness of Religious Ideas: A Cognitive Theory of Religion*. Berkeley, CA: University of California Press.

Boyer P and Ramble C (in press) Cognitive templates for religious concepts: cross-cultural evidence for recall of counterintuitive representations. *Cognitive Science*.

Cardena E, Krippner SC and Lynn SJ (eds) (2000) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Guthrie S (1993) *Faces in the Clouds: A New Theory of Religion*. New York, NY: Oxford University Press.

Lawson ET and McCauley RN (1990) *Rethinking Religion: Connecting Cognition and Culture*. Cambridge, UK: Cambridge University Press.

Leslie AM (1995) A theory of agency. In: Sperber D, Premack D and Premack AJ (eds) *Causal Cognition: A Multidisciplinary Debate*. Oxford, UK: Clarendon Press.

McCauley RN and Lawson ET (2002) *Bringing Ritual to Mind: Psychological Foundations of Cultural Forms*. Cambridge, UK: Cambridge University Press.

Sperber D (1975) *Rethinking Symbolism*. Cambridge, UK: Cambridge University Press.

Whitehouse H (1992) Memorable religions: transmission, codification and change in divergent Melanesian contexts. *Man* (N.S.) **27**: 777–797.

Whitehouse H (1995) *Inside the Cult: Religious Innovation and Transmission in Papua New Guinea*. Oxford, UK: Clarendon Press.

Whitehouse H (2000) *Arguments and Icons: The Cognitive, Social and Historical Implications of Divergent Modes of Religiosity*. Oxford, UK: Oxford University Press.

Further Reading

Andresen J (ed.) (2001) *Religion in Mind: Cognitive Perspectives on Religious Belief, Ritual and Experience*. Cambridge, UK: Cambridge University Press.

Barrett JL (2000) Exploring the natural foundations of religion. *Trends in Cognitive Science* **4**(1): 29–34.

Bloch M (1992) *Prey into Hunter: The Politics of Religious Experience*. Cambridge, UK: Cambridge University Press.

Boyer P (ed.) (1993) *Cognitive Aspects of Religious Symbolism*. Cambridge, UK: Cambridge University Press.

Hirschfeld LA and Gelman SA (eds) (1994) *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York, NY: Cambridge University Press.

McCauley RN (2000) The naturalness of religion and the unnaturalness of science. In: Keil F and Wilson R (eds) *Explanation and Cognition*, pp. 61–85. Cambridge, MA: MIT Press.

Newberg AB and D'Aquili EG (1998) The neurophysiology of spiritual experience. In: Koening HG (ed.) *Handbook of Religion and Mental Health*, pp. 75–94. San Diego, CA: Academic Press.

Sperber D (1996) *Explaining Culture: A Naturalistic Approach*. Oxford, UK: Blackwell.

Representation Formats in Psychology

Introductory article

Arthur B Markman, University of Texas, Austin, Texas, USA

CONTENTS

Representation formats are specific methods used by the mind/brain to store information for later use.

MODES OF INFORMATION STORAGE

Knowledge representation is at the core of psychological theories. In brief, a representation is some internal state of a cognitive system that carries information that the system uses for thought and action. Representations consist of an internal state (known as the *representing world*) which corresponds to some *represented world*. The format in which knowledge is represented (the way the representing world is set up) has a significant impact on what the cognitive system finds easy to do and what it finds hard to do. For this reason, psychologists have examined a number of different representational formats in order to understand their properties and their suitability for theories of cognitive processing.

This article is organized around two issues. First, proposals about representation have explored how the structure of the representing world influences the ease with which aspects of the represented world can be encoded. Two important dichotomies that have been explored in this context are the distinctions between analog and propositional representations and between localist and distributed representations. Second, research on representation has explored how representational structures allow a cognitive system to go beyond the information present in the environment. This issue will be discussed in the context of scripts and schemas.

ANALOG AND PROPOSITIONAL REPRESENTATIONS

The distinction between analog and propositional representations focuses on how the representing world comes to reflect the structure of the represented world. Analog representations are representing worlds whose natural structure mirrors that of the represented world. For example, an analog clock represents time as the angular distance traveled by a hand on the clock. Time is a quantity, and there are certain relationships that hold between quantities. For example, amounts of time have a transitive structure, so that if event A is longer than event B, and event B is longer than event C, event A must be longer than event C. The angular distance traveled by a hand on a clock is also a quantity, and thus it also naturally displays a transitive structure. For example, if the angular distance traveled by the minute hand of a clock during event A is larger than the angular distance traveled by the minute hand during event B, and the angular distance traveled by the minute hand during event B is larger than the angular distance traveled by the minute hand during event C, then the angular distance traveled by the minute hand during event A must be larger than the angular distance traveled by the minute hand during event C. Thus, angular distance and time share some relationships. Not every relationship in the domains is shared, as angular distance may be traversed clockwise or counterclockwise, but time moves in only one direction.

Analog representations have been used in theories of visual perception and mental imagery. On this view, the representation of visual information has the same structure as the three-dimensional space being represented. Studies examining whether mental images have an analog representation have generally tried to demonstrate that transformations of mental images have the same properties as transformations of real images. For example, classic studies by Shepard and his colleagues found that the amount of time required to rotate a mental image increases with the angle of

rotation, just as the time required to rotate a physical image increases with the angle of rotation.

Propositional representations differ from analog representations in that the relationships among aspects of the representing world are arbitrary. Thus, in order for the representing world to capture relationships in the represented world, these relations must be explicitly defined. For example, a digital clock represents time using numbers as a representing world. The structure of the number system is set up by convention. There is nothing inherent in the shapes of the numbers ('1', '2') that require them to stand for certain quantities. Furthermore, there is nothing in the way numbers are written that determines that 7 is less than 8 or that 8 is less than 9. Thus, the domain of digits does not have a natural transitive structure the way time and angular distance do. In order to use numbers to represent quantities, the relationships among the numbers need to be defined explicitly.

Propositional representations encode complex items such as natural objects or situations using relations among parts. For example, an image of a person might be described as having a head on top of a body. By using this type of representation, the relative positions of the parts can be represented, even when the same object is viewed from different perspectives. Studies examining whether mental representations have a propositional structure typically focus on whether complex objects or scenes are broken down into predictable subcomponents and relations between them. This work suggests that people do tend to divide objects into subcomponents.

In research on visual representation, proposals for analog representations are typically *viewer-centered*. That means that the representation of an object will change when the perspective of the viewer changes. In contrast, proposals for propositional representations are typically *object-centered*, meaning that the parts of objects are represented relative to each other rather than relative to the viewer. Thus, an object-centered representation will not change when the perspective of the viewer changes.

Given that some evidence from visual perception favors analog representations and some favors propositional representations, which type is correct? In all likelihood, both are correct. Analog representations facilitate storage of information about metric properties of objects such as the distances between parts. Propositional representations facilitate storage of relational properties such as the relative position of two parts of an object. Kosslyn has suggested that the hemispheres of the brain may differ in the types of representations they favor, with metric information being stored by the right hemisphere and relational information being stored by the left hemisphere.

LOCALIST AND DISTRIBUTED REPRESENTATIONS

Representations may also be localist or distributed. A localist representation involves a collection of discrete symbols that are used to represent properties. For example, a dog might be described by the properties 'furry, four-legged, barks, loyal'. Each of these features is a symbol that stands for some property that is true of dogs. Each feature is a localist representation, because it has a particular meaning associated with it, and that meaning is distinct from the meanings of other symbols.

At first glance, localist representations appear to be the only obvious way of encoding information. This intuition is driven by the fact that words are symbols just like those in localist representations, and we often derive our intuitions about psychology on the basis of what is easy to verbalize. Indeed, localist representations are important in many theories of language processing.

Despite this strong intuition, even a cursory examination of the brain suggests a different way of representing information. The brain is an organ consisting of a number of different types of cells. One important type, called the *neuron*, is capable of sending electrical signals from one cell to another. The brain carries information on the basis of the pattern of signaling (or *firing*) of a group of neurons rather than in the signals sent by a single neuron. For example, classic work on the visual cortex by Hubel, Wiesel, and their colleagues suggests that there are neurons that fire most strongly (i.e. they send the most signals) for lines of a particular orientation. These neurons will fire somewhat less strongly in the presence of lines that deviate slightly from their preferred orientation, and they will fire very little for lines that deviate a lot from their preferred orientation. Thus, at any given moment, there are a large number of directionally sensitive neurons in visual cortex firing, with some cells sending a lot of signals and others sending very few. The orientation of the line in the environment is determined by the pattern of activation across the set of neurons rather than by the activation of a particular neuron.

There are two advantages of this distributed coding over a localist coding. One is fairly obvious. Cells in the brain are fragile. If a cell dies, the brain's ability to represent information does not

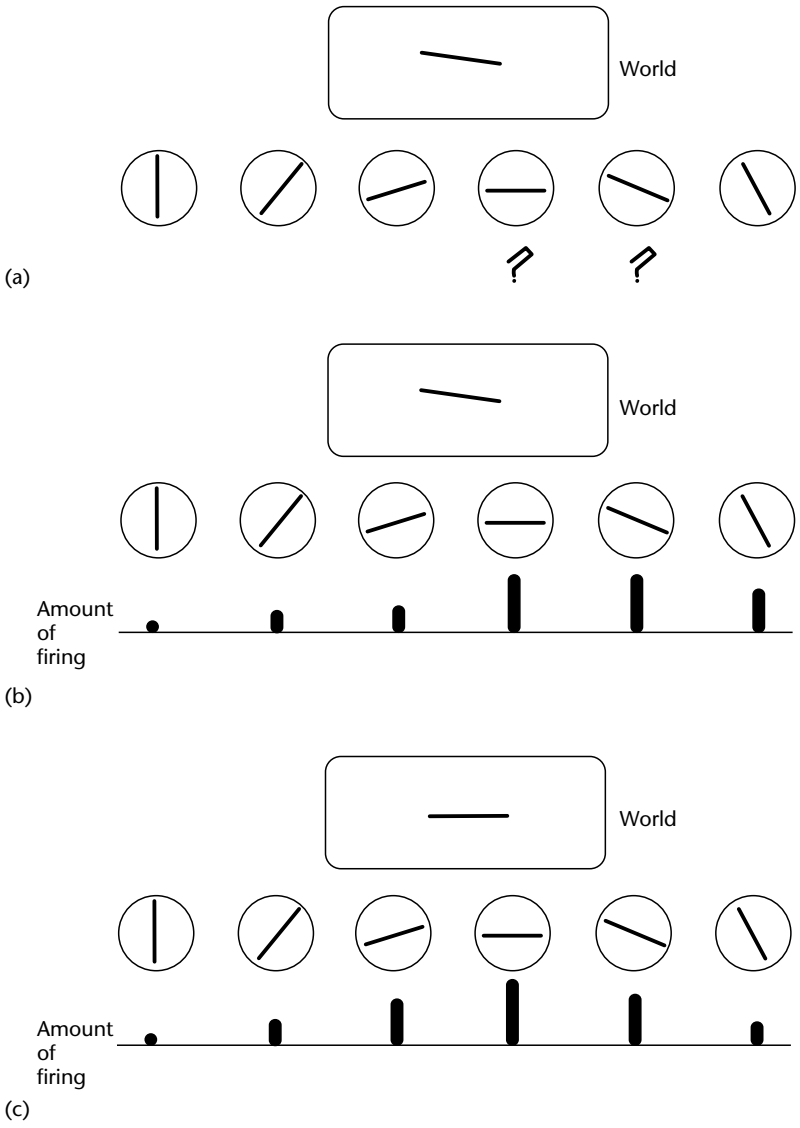


Figure 1. Examples of localist and distributed coding. The circles are representational units for orientation of a line. The preferred orientation of each unit is shown in the circle. In the localist coding (a), each unit represents a different orientation. The maximum resolution of the representation depends on the degree of difference in preferred orientation of neighboring representations. In a distributed coding, (b) and (c), each unit fires to some degree depending on the similarity of the input to the preferred orientation of the unit. The pattern of firing across the units changes even with small changes in the orientation of the line in the world. Thus, the maximum resolution of the distributed representation is actually finer than the difference in preferred orientation of neighboring units.

suffer much, because the pattern of firing across a large set of neurons carries information. This robustness in the face of cell loss is called *graceful degradation*. The second advantage of a distributed coding is more subtle. Imagine a localist coding of orientation like that shown in Figure 1(a). When a line is presented in the world, one (localist) unit will be active. This active unit describes the orientation of the line in the world. Because there are only six different units, the orientation of the line cannot be represented with a high degree of accuracy. The resolution of this representation is only as good as the difference in preferred orientation of neighboring units. In contrast, in a distributed representation, like those in Figures 1(b) and 1 (c), every unit fires to some degree when a line is present, but the amount of firing depends on the similarity of the line to the preferred orientation of the unit. Thus, even small changes in the orientation of a line in the world lead to different patterns of firing of the units. In this way, the system can represent orientations of lines to a finer degree of

resolution than the difference in preferred orientation of neighboring units.

GOING BEYOND THE INFORMATION GIVEN: SCRIPTS AND SCHEMAS

The representations discussed so far are constructed on the basis of information in the environment. In most complex situations, the information needed to understand what is happening in the world must be derived from a combination of what is available in the world and what can be inferred from past experience. For example, consider the following story:

> Sarah went to the restaurant and ordered a salad. She ate it and left.

After reading this story, people typically assume that Sarah was sitting while she ordered and ate, that her order was taken by an adult, and that she paid for the salad. This information is not contained in the story, but rather follows straightforwardly from our previous experiences with restaurants.

There have been many proposals for overarching knowledge structures that can be used to interpret the world. These knowledge structures are called 'schemas' or 'scripts'. At heart, these proposals have a core set of characteristics. First, it is assumed that these structures are built up on the basis of past experience (both personal experience and second-hand experience through narratives). Second, these structures describe the temporal order of events in a situation, which enables people to predict what will occur next. Third, schemas contain causal relations among events that describe not only what sub-events tend to occur in an event, but also why those sub-events occur. These causal relations are important for understanding violations of the typical sequence of events, such as:

> Sarah went to the restaurant and ordered a salad. She took one bite and left without paying.

In this case, most people assume that Sarah was dissatisfied with her order, because paying for the food at the end of the meal is part of the transaction that occurs when going to a restaurant. These deviations from expectations are often the most interesting parts of stories.

In sum, schemas and scripts allow people to go beyond the information that is obviously available in the environment in order to comprehend a new situation. These knowledge structures allow people to make predictions about what will happen in a new situation and also provide causal information that is useful for reasoning about deviations from the typical course of events.

CONCLUSION

Three issues need to be considered when positing a representation in a theory of cognitive processing. First, a theorist must decide whether there is a representational system with a structure analogous to that of the represented world that should be used. Second, the theorist must determine whether it would be advantageous to use a distributed representation of the information. Finally, the theorist must determine whether constructing a representation in a domain will involve going beyond the information that is available in the environment. If so, then the theory must incorporate knowledge structures that allow the system to construct plausible inferences.

Further Reading

Bower GH, Black JB and Turner TJ (1979) Scripts in memory for text. *Cognitive Psychology* **11**: 177–220.

Hinton GE and Anderson JA (eds) (1981) *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Hubel DH and Wiesel TN (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* **28**: 229–289.

Hummel JE (2000) Where view-based theories break down: the role of structure in human shape perception. In: Dietrich E and Markman AB (eds) *Cognitive Dynamics*, pp. 157–186. Mahwah, NJ: Lawrence Erlbaum.

Kosslyn SM (1994) *Image and Brain*. Cambridge, MA: MIT Press.

Markman AB (1999) *Knowledge Representation*. Mahwah, NJ: Lawrence Erlbaum.

Reed SK (1974) Structural descriptions and the limitations of visual images. *Memory and Cognition* **2**: 329–336.

Rumelhart DE and McClelland JL (eds) (1986) *Parallel Distributed Processing*, vol. 1. Cambridge, MA: MIT Press.

Schank RC and Abelson RP (1977) *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Shepard RN and Cooper LA (eds) (1982) *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.

Representations, Abstract and Concrete

Intermediate article

Joan Gay Snodgrass, New York University, New York, USA

CONTENTS

Abstract and concrete representations are symbols that refer to an object, event or idea that has existed, does exist or might exist in the real world. Concrete representations bear a physical resemblance to their referents, whereas abstract representations do not. Both come to be learned through the process of categorization.

CONCRETE VERSUS ABSTRACT REPRESENTATIONS

A representation is a symbol that stands for something else. Usually the 'something else' is an object, event or idea that has existed, does exist or might exist in the real world. A concrete representation is one that bears a physical resemblance to the real-world object it represents. This representation could be visual, as in a picture of an object; auditory, as in the sound the object makes; tactual, as in the shape of an object felt through the skin; or olfactory or gustatory, as in the smell or taste of an object. In contrast, an abstract representation is one that bears no resemblance to the object being represented. The abstract representation *par excellence* is a word or combination of words in a language. Almost everything that can be experienced or thought about can be captured in a word or series of words in a speaker's language.

This article is confined to concrete representations that are pictorial symbols. These are most useful for representing objects: for example, line drawings of common objects and animals, such as those shown in Figure 1, are easily recognized as representing objects in the real world, as evidenced by the fact that they can be easily named by both children and adults, and are given comparable names across languages (e.g. the chair is consistently named 'chair' in English, 'chaise' in French and 'sedia' in Italian). Line drawings are easily recognized as representing objects even by children who have not been previously exposed to drawings.

Other pictorial symbols have been used to represent more abstract ideas. A good example is the international system of road signs, which uses a set of easily understandable symbols for road commands (although such symbols are not universally understandable without some training – for example, does the picture of an upright palm of a hand mean 'stop' or 'no entrance'?) Another example is the icons introduced by Apple Computer and incorporated into the Microsoft Windows operating system and software for lessening the memory load on users.

In terms of concrete versus abstract representations, we can imagine a continuum in which pictures in all their manifestations, including photographs, paintings and line drawings, can best represent the surface appearance of the world; while words in all their manifestations, including single words representing a common object, words representing abstract ideas, sentences expressing a complex idea, and poems representing inexpressible thoughts and feelings, can best express the abstract nature of the world. These different types of representations lie along a continuum, as shown in Figure 2. However, as we shall see, we could also consider a drawing of a generic object, such as the chair in Figure 1, to be abstract in some sense.

COMMON OR SEPARATE CODES?

A great deal of research has centered on the middle section of the abstract/concrete continuum of Figure 2, in which concepts can be represented both by pictures and by the words that name them. Two competing theoretical positions have been a focus of controversy in this research area. One, the common code model, argues that there is an underlying conceptual representation of an idea which can be accessed equally well by either of the two surface forms – the picture of an object or its

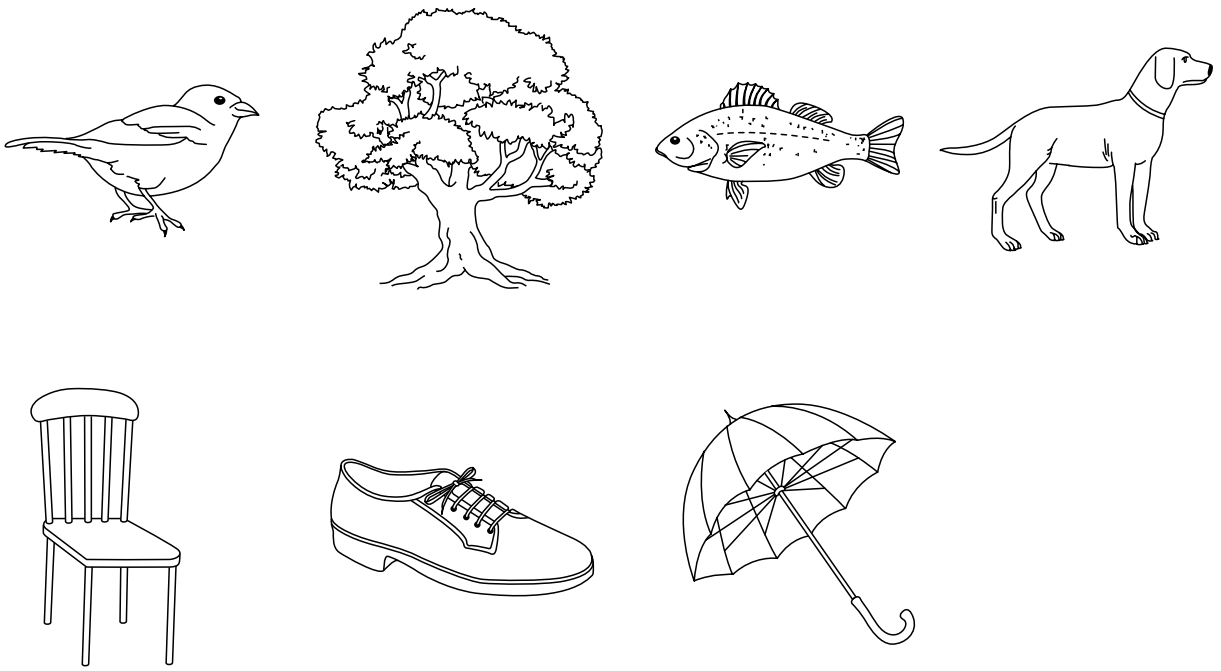


Figure 1. Examples of line drawings of common objects and animals (from Snodgrass and Vanderwart, 1980).

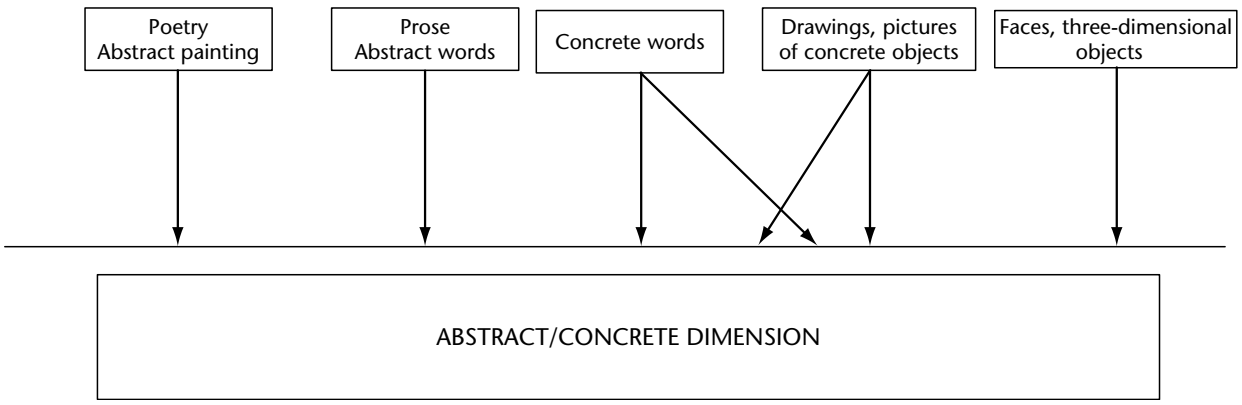


Figure 2. Abstract/concrete continuum illustrating the forms of representation best suited to abstract concepts (to the left) and those best suited to concrete concepts (to the right). The intermediate section, including both concrete words and pictures of common objects, is the representational domain most studied in cognitive science.

name. The second, the separate or dual-code hypothesis, argues that knowledge about the world is organized according to the way in which that knowledge was acquired, and thus different underlying representations or systems of meaning will be accessed by the two surface forms.

This dual-code hypothesis was first formulated by Paivio (1971), who proposed that images corresponding to objects or pictures of objects are kept in an image store, while verbal codes corresponding to the names of objects are kept in a verbal store. The two stores are associatively linked, so that pictures may be named and referents of words may be imaged, but the stores are separate and the two types of code are specialized for different functions. Both pictures and names of concrete objects may be dually encoded (Figure 2).

A picture is virtually certain to produce an image code, and may also produce a verbal code if the picture is named by the observer. Similarly, a word is virtually certain to produce a verbal code and may also produce an image code if the observer forms a mental visual image of the referent of the word. However, Paivio assumed that pictures are

more likely to be spontaneously dually encoded than words are. In addition, Paivio argued that the image code produces a stronger memory trace than the verbal code. These two assumptions – that pictures are more likely to be dually encoded than are words, and that the image code is a better memory trace than the verbal code – were used by him to account for the picture 'superiority effect' in memory: pictures are generally remembered better than words, regardless of whether the memory test is recall or recognition (Figure 3).

Paivio also proposed that image and verbal codes differ in the processes for which they are specialized. Image codes are specialized for representing spatial properties such as size and location, while verbal codes are specialized for representing temporal properties such as order. As support for this, he showed that size differences between objects in the real world were more quickly discriminated when the objects were represented by pictures than by words (Paivio, 1975).

Although Paivio's assumptions about differences between image and verbal codes have generally been accepted on the basis of experimental evidence, his assumption that the core concepts or meaning underlying image and verbal codes reside in separate stores has not. Instead, a hybrid model conceptualizing separate stores for image and verbal codes and a common store for their underlying representations has evolved (Figure 4).

This hybrid model contains an intermediate level which corresponds to Paivio's dual-coding model. The verbal code is assumed to be an acoustic image of the word in question (the phonological code), and the visual code is assumed to be a visual image of the picture which preserves the spatial

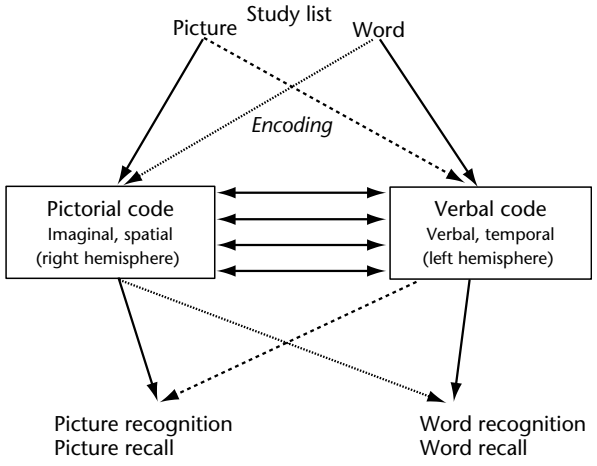


Figure 3. Paivio's dual-code model.

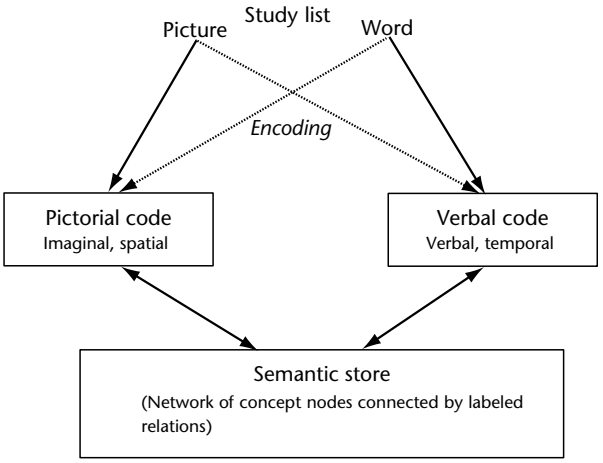


Figure 4. The common store model.

relationships among the elements of the picture but may not preserve every detail. This intermediate level corresponds to the contents of conscious thought, which consists of a series of acoustic images (we hear ourselves think) and a series of visual images (we visualize past, present and future events). Although not pictured in Figure 4, this model – described more fully by Snodgrass (1980) – assumes that the acoustic and visual image stores contain information about how the typical object looks and the typical word sounds. Because most of the words and pictures we experience are familiar, we assume that each has a corresponding acoustic and visual code that can be accessed and used to generate acoustic and visual images. So we can hear the word 'apple' spoken in our mind's ear, and see the visual image of an apple in our mind's eye by generating the corresponding image from the image store.

A presented word or picture may not always correspond to the stored prototypical image. For example, a word printed in mixed-case letters (ApPlE) or spoken with a foreign accent will produce a mismatch between its physical code and the prototypical store and accumulate in a mismatch accumulator, which detects discrepancies between the heard or seen word or picture and its prototype. It is much more common, however, for a picture to produce a mismatch (because pictures of objects can be drawn in so many different ways), so pictures are more likely than words to accumulate mismatch information. This accumulation of mismatching information for pictures is assumed to provide another way in which pictures may be remembered better than their names, by making them more distinctive. This 'distinctiveness hypothesis' for the picture superiority effect in

memory was first proposed by Nelson and his colleagues (Nelson *et al*., 1977; Nelson, 1979).

The common store accessed by both pictures and their names is the semantic store (Figure 4). This is an abstract set of nodes and interconnections among nodes, which differ qualitatively and could be labeled by such relationships as 'is a member of the category', 'has the property', and so on. Although we use language to talk about the properties of the semantic store, semantic memory has no language or visual images and its operations are not accessible to consciousness. Both words and pictures have access to the semantic store, but words referring to abstract concepts will access parts of the semantic store inaccessible to pictures, and pictures referring to indescribable concepts will access parts of the semantic store inaccessible to words.

The reality of visual imagery, whether produced as mental representations of a visual object or scene, or generated by an observer to a verbal description, has been well documented in both behavioral and cognitive neuroscience. Attention has increasingly focused on similarities and differences in processing of pictures and words in particular, and in concrete and abstract representations more generally.

GROUNDING OF ABSTRACTIONS

An abstract representation is one that bears no resemblance to what it represents. However, there are degrees of abstraction, from the most abstract (such as a word describing an idea, e.g. 'truth', which has no material referent in the world) to the most concrete (such as an object in your home). The chair pictured in Figure 1 is meant to represent a typical chair, and is easily understood as a chair even though it may not represent the exact form of any chair in a person's home. The name 'chair' is even more abstract in the sense that it includes properties that are not visible in a picture, such as the fact that a chair is constructed of rigid material, or that a chair is a member of the superordinate category furniture, or that a chair may be bought in a furniture store. Even more abstract concepts such as truth and beauty are not represented directly in objects we encounter in the world, but rather come to be inferred from our experience with the world.

To return to the concept of chair, we can think of this concept as a category of things representing an intermediate level of abstractness, ranging from a thing in the world, to artefact, to furniture, to chair, to dining-room chair, to the chair in my dining room with the nick on its right leg. How do we decide on the level of abstractness with which to describe our world? Of course, it depends upon the purpose; if I want the nick on my dining-room chair repaired, then I will speak of a specific chair. However, in common discourse we generally choose an intermediate level of abstractness which Rosch *et al*. (1976) have termed the 'basic level'. The basic level of categorization is a unique level of abstraction which provides an optimum level of specificity so that human information-processing abilities are not overtaxed, yet with sufficient specificity so that accurate communication can occur. Rosch and her colleagues used a variety of techniques to show that the basic level was optimum for communication. These included asking observer participants to list features for various levels in a hierarchy, determining which level showed the most commonality of motor movements, and determining which level showed the greatest commonality of shape. They showed that the basic level had the greatest number of features that distinguished one basic level object from another (known as cue validity), provided the greatest commonality of motor movements (the motor movements for sitting on a chair are the same regardless what the chair looks like), and provided the greatest commonality of shape.

Priority of Specific or Concrete Information

Many investigators have concluded that concrete representations such as pictures of objects are processed more efficiently than abstract representations such as the names of objects. These two surface forms have usually been compared in tasks in which existing knowledge (semantic memory) is required, so they are known as semantic memory tasks. In semantic memory tasks, participants are assumed to enter the laboratory in possession of this general cultural knowledge, such as the names of pictures, the pronunciation of words, or the categories to which objects belong. Because participants do not need to learn anything new to perform these tasks, they have also been called 'online processing tasks'.

Two types of tasks have been used to compare online processing differences between pictures and words: naming or reading tasks, and categorization tasks. In the naming task, participants are shown either pictures or words and asked to vocally name them. The experimenter measures both naming time (time from the presentation of the stimulus to voice onset) and, particularly for brain-damaged individuals, naming errors. More than a century

ago, Cattell (1886) noted that naming pictures took longer than naming words. He attributed this word advantage to habit – to the closer association between concept and expression built up through experience. Theios and Amrhein (1989) reviewed a number of theories for the word naming advantage and showed that habit could not completely account for the superiority, rather that one needed to include the possibility that words can be named phonologically. Thus, the fact that word reading is faster than picture naming does not mean that pictures are less readily understood than words, but rather that word naming benefits from overlearning and from correspondence between the graphemes in a printed word and rules for their pronunciation.

A task that is more generally agreed to index the speed with which a representation is understood is the categorization task. Here, participants are asked to determine the superordinate category of an object, represented as either a picture or a word, and their categorization speed is determined. There are a number of ways in which categorization may be performed: participants may be asked to name the superordinate category (superordinate naming), to decide whether two exemplars are from the same or different categories (same/different tasks), or to make a binary decision about whether an exemplar is from one or the other of two categories (binary decision task). Across tasks, the general finding is that pictures are categorized faster than their names, and this has been taken to mean that pictures are understood faster than words (Potter and Faulconer, 1975; Pelligrino *et al*., 1977).

There are a number of reasons why pictures might be understood more quickly than words. First, 'form follows function' – the shape of objects often represents their function. For example, animals locomote, and their manner of locomotion, legs, are often a prominent feature of their pictorial representation; whereas tools are manipulated by hand, so a handle is often a prominent feature of *their* pictorial representation (Snodgrass and McCullough, 1986). Indeed, a basic dichotomy between living things and artefacts has been proposed on the basis of a number of studies of people with brain damage. Several patients have been reported to show category-specific naming deficits which differentially affect living things as opposed to artefacts; see Saffran and Schwartz (1994) for a review. This has been taken as evidence that different brain regions subserve understanding of living things versus artefacts. The next section considers how categories develop, and what models of categorization have been proposed.

THEORIES OF CATEGORIZATION

Categorization is a basic human cognitive process which is indispensable in structuring both experience and thought. Although some categories are arbitrary and must be learned by rote (e.g. letters of the alphabet) and still others are physiologically based and appear to be universal across cultures (e.g. color and form), many categories of natural objects and artefacts, such as those representable by both pictures and words, appear to be mediated by physical and functional similarity in which some members of the category are more central or typical than other members of the category.

The basis of virtually all theories of categorization is similarity. In the classical theory, which goes back to Aristotle, each member of a category is held to possess a set of features or properties which taken together are singly necessary and jointly sufficient to define the category. There are certain categories, such as geometric figures and members of the two gender classes male and female, which are classical in that sense. However, most other categories such as natural things and manufactured objects have graded representations. Consider, for example, the category of 'bird'. All birds have feathers and wings; most 'typical' birds are small, can fly, build nests, lay eggs and eat insects. However, ostriches are large and cannot fly, hawks eat meat, and bats, although not birds, have many of the visible features of birds. In fact, people make distinctions among exemplars of categories which reflect how closely they resemble the typical member of a category. The most typical member of a category is called a 'prototype'. The prototype of a category may not exist, but simply possess most of the important diagnostic features characteristic of the set of exemplars.

The importance of prototypes in structuring naturally occurring categories has been demonstrated in many ways. First, people agree on their ratings of prototypicality – 'robin' is rated as a more typical bird than 'ostrich'. Second, a more prototypical member of the category is categorized more quickly than a less prototypical member. Finally, in experiments in which artificial categories are constructed and taught to participants, the prototype of the constructed category, which may never be shown during the learning process, is usually classified as a member of the category quickly and with relatively little error in a transfer test.

How do people come to form abstractions or categories based upon experience with the world? How does a child, having experienced several furry objects that have been called 'dog', come to

recognize a new object as a dog? This question has been asked, for adults, by giving them experience with artificial categories. This permits the investigator to control a participant's experience with the category. For a variety of reasons, a relatively sparse representation has been commonly used, in which objects can belong to only one of two categories (A or B), the exemplars of the two categories differ in only a few features, and there is no single feature that can be used to distinguish members of category A from those of B. A particularly influential category structure was first introduced by Medin and his colleagues (Medin and Schaffer, 1978), and has been used by many investigators since. This structure is shown in Table 1. The ones and zeros represent the presence and absence, respectively, of a particular feature. Note that Table 1 represents the entire population of the 16 exemplars which can be constructed from four dimensions varying on only two values (i.e. $2^4 = 16$).

The categories are linearly separable, which means they can be separated by a hyperplane so that it is possible to learn to classify the objects into each category without error, although participants do not always learn to do so. In general, category A exemplars have more positive values on dimensions, and in particular more positive values on dimensions 1 and 3, whereas the opposite is true for category B. Note, too, that the prototype for category A is transfer stimulus 11 (T11) and the prototype for category B is training stimulus 9 (B9).

Table 1. The 5–4 category structure introduced by Medin and Schaffer (1978) (adapted from Smith and Minda, 2000)

Stimulus	*Dimension 1*	*Dimension 2*	*Dimension 3*	*Dimension 4*
Training stimulus				
Category A				
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1
Category B				
B6	1	1	0	0
B7	0	1	1	0
B8	0	0	0	1
B9	0	0	0	0
Transfer stimulus				
T10	1	0	0	1
T11	1	0	0	0
T12	1	1	1	1
T13	0	0	1	0
T14	0	1	0	1
T15	0	0	1	1
T16	0	1	0	0

The experimental procedure is to have participants go through several runs of the nine training stimuli, with feedback, and then test them on the transfer stimuli. The dependent variable that categorization models attempt to predict is the proportion of errors made on the training and transfer stimuli, usually after several epochs of training.

Medin and Schaffer (1978) showed that prototype theory failed to predict the pattern of errors in this task, and instead they presented a new theory – context theory – which proposed that participants used the similarity between the transfer stimulus and stored exemplars in order to make their categorization responses. This exemplar model of categorization has been very influential and has led to many refinements and a virtual explosion of research. Exemplar theory states that people store memories of exemplars of each member of a category, and then classify a new exemplar according to its similarity to the set of stored exemplars. This provides a natural way for children to learn new concepts. After seeing several furry objects that adults in their environment have named 'dog', the child is able to call a new object 'dog' if that object shares enough similar features with enough stored exemplars.

One problem with exemplar theory is that it would appear to require massive amounts of memory, as specific exemplars of each of the many categories of knowledge we possess would seem to be required in order to categorize each new object. This theory would also seem to make the existence of semantic memory unnecessary and redundant, as what we have been calling semantic memory would simply be based upon episodes of experience in episodic memory.

Smith and Minda (2000) analyzed the bases on which exemplar models have succeeded, and prototype models have failed, in accounting for data from the category structure in Table 1. They concluded that the successes of exemplar models rest on their ability to account for superior categorization of the trained exemplars, and that when prototype models are given that same facility, they are able to fit the data as well as the context (exemplar) model. They also suggested that because of the sparseness of the category structure, participants may have been motivated to memorize the training examples rather than to search for a classification rule, thus producing better performance on the training exemplars than would be

predicted by their performance on the transfer exemplars. If nothing else, this analysis suggests that, in category formation, the roles of familiarity and experience are important.

CONCLUSION

This article has compared two types of representations – concrete representations of pictures and abstract representations of the words that name them – and has argued for a common code model which relates the two surface forms. Consideration has been given to ways in which abstract representations are built up out of the overpowering specificity of the world.

References

Cattell JM (1886) The time it takes to see and name objects. *Mind* **11**: 63–65.

Medin DL and Schaffer MM (1978) Context theory of classification learning. *Psychological Review* **85**: 207–238.

Nelson DL (1979) Remembering pictures and words: appearance, significance, and name. In: Cermak LS and Craik FIM (eds) *Levels of Processing in Human Memory*, pp. 45–75. Hillsdale, NJ: Lawrence Erlbaum.

Nelson DL, Reed VS and McEvoy CL (1977) Learning to order pictures and words: a model of sensory and semantic encoding. *Journal of Experimental Psychology: Human Learning and Memory* **3**: 485–497.

Paivio A (1971) *Imagery and Verbal Processes*. New York, NY: Holt.

Paivio A (1975) Perceptual comparisons through the mind's eye. *Memory and Cognition* **3**: 635–647.

Pellegrino JW, Rosinski RR, Chiesi HL and Siegel A (1977) Picture–word differences in decision latency: an analysis of single and dual memory models. *Memory and Cognition* **5**: 383–396.

Potter MC and Faulconer BA (1975) Time to understand pictures and words. *Nature* (London) **253**: 437–438.

Rosch E, Mervis CB, Gray WD, Johnson DM and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology* **8**: 382–439.

Saffran EM and Schwartz MF (1994) Of cabbages and things. Semantic memory from a neuropsychological perspective – a tutorial review. In: Umilta C and Moscovitch M (eds) *Attention and Performance 15: Conscious and Nonconscious Information Processing*, pp. 507–536. Cambridge, MA: MIT Press.

Smith JD and Minda JP (2000) Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory and Cognition* **26**: 3–27.

Snodgrass JG (1980) Toward a model for picture–word processing. In: Kolers PA, Wrolstad ME and Bouma H (eds) *Processing of Visible Language*, vol. 2. New York, NY: Plenum.

Snodgrass JG and McCullough B (1986) The role of visual similarity in picture categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition* **12**: 147–154.

Snodgrass JG and Vanderwart M (1980) A standardized set of 260 pictures: norms for naming agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* **6**: 174–215.

Theios J and Amrhein PC (1989) Theoretical analysis of the cognitive processing of lexical and pictorial stimuli: reading, naming, and visual and conceptual comparisons. *Psychological Review* **96**: 5–24.

Further Reading

Caramazza A and Shelton JR (1998) Domain-specific knowledge systems in the brain: the animate–inanimate distinction. *Journal of Cognitive Neuroscience* **10**: 1–34.

Paivio A (1986) *Mental Representations: A Dual Coding Approach*. New York, NY: Oxford University Press.

Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Snodgrass JG (1984) Concepts and their surface representations. *Journal of Verbal Learning and Verbal Behavior* **23**: 3–22.

Tulving E (1972) *Episodic and semantic memory*. In: Tulving E and Donaldson W (eds) *Organization of Memory*, pp. 381–403. New York, NY: Academic Press.

Rule-based Thought

Introductory article

Ulrike Hahn, University of Wales, Cardiff, UK

CONTENTS

Rule-based thought involves cognitive processes based on the applications of general statements which have been derived from past experiences.

INTRODUCTION

The principle that cognition is based on the application of mental rules was once virtually axiomatic. However, difficulties with rule-based approaches subsequently gave rise to a wide range of alternative models of thought, the chief ones drawing on similarity-based processes. This prompted a large experimental literature seeking to determine empirically whether or not particular cognitive processes are rule based.

RULE- VERSUS SIMILARITY-BASED THOUGHT

The extent to which human cognition is based on rules is an issue of longstanding interest. In the early days of artificial intelligence it was axiomatic that human thought is rule based. Rules no longer have this general dominant role, but there are still rule-based accounts of many tasks, whether these be language, problem solving or classification. However, the notion of rule is one of the most confused within cognitive science, and conceptual clarification is necessary if claims that a particular behavior is rule based are to have substance. The crucial distinction that must be made is between behavior which is *guided* by rules and behavior that is merely *described* by rules. Corresponding to this distinction, there are two different types of appeals to 'rules' in the literature, regarded as 'weak' and 'strong', respectively. An example of weak usage of the term 'rule' in explaining behavior is a general behavioral claim that a language learner has succeeded in 'mastering the rules of English', or the assumption that infants are born with 'rules for looking' which guide their exploration of the visual environment. Statements such as these use the term 'rule' to refer to an external regularity (of English) or an internal constraint without making a claim about mental architecture (i.e. without wishing to endorse a particular view about how the external regularity or the innate constraint are internally represented by the agent). Such a weak usage of the term 'rule' in a cognitive context is *not* the focus of the debate about mental rules, which revolves around the 'strong' use of 'rule'. According to the strong reading, speaking of an agent as possessing a rule is a statement about cognitive architecture. It is the claim that an agent is making use of mental representations of a particular representational format, namely 'rules', which can be distinguished from other types of mental representations. This stronger, more specific claim lies at the heart of debates in which rule-based models are contrasted with similarity-based accounts such as exemplar models. Crucially, the strong use claims an agent-internal role for the rule. Stating that an agent possesses a particular rule is not merely saying that this agent's behavior displays a particular regularity, but rather it is saying that this 'rule' has a causal role in producing this behavior – that is, the behavior has the regularity that it does *because* the agent possesses the rule in question.

This is commonly phrased in terms of the distinction between rule-guided or 'rule-following' behavior and behavior which is merely conveniently described by rule. Rule following is exemplified by legal systems and their effect, where documents that encode the law cause particular behaviors, such as paying certain amounts of tax. Merely rule-describable but not rule-guided behavior is exemplified by planetary motion – planets' orbits are well described by physical laws, but planets do not themselves consult these laws in order to guide their behavior. In the latter example, the statement of the rules of planetary motion does not function causally in the generation of the planets' behavior, as do legal rules. This is readily apparent from the fact that planetary motion is unaffected by the presence or lack of knowledge of the appropriate physical laws, whereas legal

rules must be known for them to produce the behaviors in question.

Confusion readily arises because cognitive theories speak of rule-guided behavior even when the rules are entirely agent internal and not consciously accessible. Terminologically, it helps to avoid confusion if 'rules' are distinguished from mere 'regularities.' A 'rule', as invoked by the strong usage of rules, is not the regularity itself but rather a *statement* of this regularity. In other words, a rule is a representation of the regularity which follows a particular representational format. Any strong claim about the rule-based nature of a particular behavior is claiming that an appropriate *statement of a regularity in a rule-based format has a causal role in the production of the behavior* in question, whether or not the cognitive agent is aware of this.

This can be illustrated by a classic example. In English, the vast majority of past tenses of verbs are formed by adding /ed/ to the stem of the word (e.g. walk → walked, talk → talked, kiss → kissed). This constitutes a regularity governing the English past tense. One way in which this regularity might be exploited by the cognitive system is through the acquisition, on the part of the language learner, of an internal mental rule, the content of which would be 'stem + /ed/ → past tense', and which is applied (albeit unconsciously) whenever a past tense is to be formed. This would constitute an instance of rule-guided behavior. However, an internal mental rule of this kind need not be the only way in which this regularity might be exploited. A statistical device such as a standard connectionist network, or a similarity-based approach in which past tenses are formed analogously to those of the most similar-sounding known words, might give rise to the same behavior. These would not constitute examples of rule-based cognition.

In many if not all areas in which rule-based explanations of cognitive performance have been advanced, they have been contrasted by explanations based on similarity or analogy. Another example of this type concerns the way in which we classify objects in day-to-day life. For example, when learning about dogs we might have abstracted a general rule, say 'If a creature is furry, has four legs and barks, then it is a dog', which we apply in order to determine whether or not something is a dog. Alternatively, we might simply store in memory lots of individual dogs that we have encountered (Lassie, the neighbor's dog Rex, etc.), and classify as a dog anything which we find to be sufficiently similar to those dogs we have in memory. The key difference between these two approaches is that the application of the rule requires that all of its criteria are met. In other words, matching of the rule to the instance to which it is to be applied is *strict*, whereas similarity-based classification is a matter of *partial* matching – that is, the instance one is seeking to classify need not match a stored instance in all respects, but may only match it more or less. A direct consequence of the fact that rule application is strict is that rules must be more *abstract* than the instances to which they are applied, if they are to match more than one instance. By contrast, the partial matching involved in similarity comparisons means that abstraction is not required.

Strict application versus partial matching, and the need for abstraction versus independence from abstraction for generalization, constitute the core criteria of rule-based and similarity-based thought, respectively. In addition, further criteria might be specified before one would be willing to regard an internal representation as a 'rule', or a partial matching process as a 'similarity comparison', but these general criteria and their implications are all that need be assumed for most empirical efforts to distinguish rules from similarity.

HALLMARKS OF RULE USE

The fact that rule application involves the strict application of an abstraction gives rise to the critical prediction that performance on items to which the rule is applied should be *uniform*.

In other words, idiosyncratic individual properties of instances which are not referenced by the rule should not affect behavior. Typically, the idiosyncratic properties of items which are experimentally manipulated in empirical tests for rule-based thought are the items' similarity to each other. Similarity between items is expected to affect performance on similarity-based accounts but not on rule-based accounts. For example, in classification, only similarity-based accounts suggest that one might expect items with high similarity to others of the class to be classified more confidently (as indeed is typically the case), whereas to the rule, 'all items are created equal'. The general logic present here can be exploited experimentally in numerous ways – for example, by examining behavioral responses (confidence, response times, etc.) in familiar instances, or through the examination of generalization performance on previously unseen items which vary in their similarity to known items. Although most of the empirical studies that seek to provide evidence for or against rule-based thought are instances of this general approach, other strategies have also been used. Because rule- and similarity-based models require

different learning strategies (rule induction versus instance storage), they might produce quite different learning profiles – the time course of learning can differ, as may what is easy and what is hard. One might be more tolerant of 'noise' in the data, and so on. *Any* such attributes could be called upon. However, the one that has probably received most attention is the presence of discontinuities in performance as learning progresses. The classic example here is the so-called 'U-shaped learning' found in (among other areas) language learning, whereby a period of correct performance on exceptions to the general regularity (e.g. the past tense forms come → came, or sing → sang) is followed by sudden over-regularization of these exceptions ('comed', 'singed'), which only gradually subsides. The onset of regularization has been explained as reflecting a transition from instance storage to rule use on the part of the learner, although non-rule-based explanations for such discontinuities have also been found.

ADVANTAGES AND DISADVANTAGES OF RULE USE

A prime advantage of rule-based models is that they are conceptually straightforward to generate, a fact which has presumably greatly contributed to their popularity. Specifically, once a regularity has been observed, we need only posit a representation or statement of that regularity which is agent internal and we have the core of a cognitive account. Given the observation that the vast majority of English words form a past tense by adding the suffix /ed/ (the regularity), one can immediately derive a cognitive account by positing a mental representation of this regularity (roughly 'for past tense add /ed/') as an internal rule which speakers are using to generate the appropriate forms, and one has a cognitive theory of past tense production.

However, rule-based theories have other advantages. A single rule can cover many, often disparate instances, thus enabling generalization across a wide range of circumstances. This power becomes extremely valuable where the rule can be communicated, thereby allowing the solution to many different examples to be acquired in one fell swoop. Practical examples of this are found in educational and legal settings. For example, the single simple rule 'No vehicles allowed in park' suffices to convey that cars, lorries, buses, bicycles, tricycles and motor scooters are all banned. However, this same generality can pose problems if the rule has to be induced (i.e. learned from experience). This is because there are typically many alternative rules which fit an existing set of data. For example, faced with three brown dogs which bark, we could infer the rules that 'all dogs bark', 'all brown dogs bark', 'all brown creatures bark', 'all four-legged creatures bark', and so on. A similar indeterminacy arises if a rule has to interact productively with already existing rules. When the predictions of a body of rules turn out to be false, there are typically many potential solutions to the question of which rule(s) to adjust. For example, if while in possession of two rules, first that 'all spiders make webs' and second that 'all eight-legged land creatures are spiders,' one discovers an eight-legged land creature which does not make webs, one could adjust either the first rule or the second one in order to eliminate the conflict with one's observation. The observation data itself does not pinpoint any unique solution.

By contrast, learning and revision in instance-based models are trivial – items are simply stored in memory. New conflicting experiences are dealt with in exactly the same manner, which means that adjustments are always local, and consequently problems like that of deciding which of many interacting rules to adjust simply do not arise. However, in order to achieve adequate performance on complex problems, large numbers of instances are typically required, or alternatively the type of analogizing performed must be of considerable sophistication and complexity, which typically requires a background knowledge of some sort. Consequently, rule- and similarity-based reasoning might be seen as complementary in their strengths and weaknesses, which suggests that cognition might do worse than resort to both.

Further Reading

Hahn U and Chater N (1998) Similarity and rules. Distinct? Exhaustive? Empirically distinguishable. *Cognition* **65**: 197–230.

Hahn U and Chater N (1998) *When is Behavior Rule-Guided?* Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society, pp. 466–471.

Newell A and Simon H (1972) *Human Problem-Solving*. Englewood Cliffs, NJ: Prentice Hall.

Rumelhart DE and McClelland JL (1986) On learning past tenses of English verbs. In: Rumelhart DE and McClelland JL (eds) *Parallel Distributed Processing. Vol. 2. Psychological and Biological Models*, pp. 216–271. Cambridge, MA: MIT Press.

Sloman S (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* **119**: 3–22.

Smith EE, Langston C and Nisbett RE (1992) The case for rules in reasoning. *Cognitive Science* **16**: 1–40.

Schemas in Psychology

Introductory article

Evan Heit, University of Warwick, Coventry, UK

CONTENTS

A schema is a source of general knowledge that tells us what to expect from observations, as well as providing a structure for understanding them.

INTRODUCTION

One of the most important contributions of cognitive psychology is the idea that people do not just passively collect information. Instead they bring their prior knowledge to bear to help them to learn about new things. For example, someone learning a new computer game would rely on previous knowledge of other computer games. By this account, a new computer game that is very similar to known games should be especially easy to learn, and indeed, the prior knowledge might help to highlight what is especially new about this new game.

BARTLETT'S STUDIES

The idea that prior knowledge or schemas guide our learning from observations has a distinguished history in psychology. Frederic Bartlett's work early in the twentieth century was a response to the Ebbinghausian tradition of studying memory through supposedly meaningless stimuli such as nonsense syllables. Bartlett argued that it is impossible to separate any stimulus from its subjective meaning, and crucially, the influences of people's knowledge are central to the nature of memory. Bartlett conducted an extensive series of studies documenting the influences of schemas on memory. (*See* **Ebbinghaus, Hermann**)

These studies showed that remembering pictures as well as written material such as stories depended highly on reconstruction, and that veridical or perfectly accurate memory was rare. For example, some participants were shown drawings of scenes such as a closed gate in a field, with a sign next to it. They later recalled that the sign said 'Trespassers will be prosecuted' even though the sign was not legible in the original picture.

Using his method of repeated reproduction, Bartlett showed the pervasive influence of schemas on memory for stories. Bartlett had British people read a story from another culture, such as an American Indian folk tale known as 'The War of the Ghosts'. These participants were required to recall the story repeatedly over a period of days or weeks. Many details such as proper names were omitted from the recalled stories. Participants tended to recall the gist of the story rather than the exact words, but even the gist of the story was distorted or 'rationalized' in light of the participants' own cultural conventions. For example, this supernatural story about a battle with ghosts was sometimes recalled as a battle with a clan of men known as the Ghosts.

SLOTS, FILLERS, AND DEFAULTS

Central to the idea of schemas is that they do not simply express people's beliefs, or tell them what to expect, but represent this information in a highly structured fashion. According to schema theory, schemas can be thought of as having a hierarchical structure, with more general schemas including more specific variants. For example, within the building schema there is a house schema, an office building schema, a church schema, and so on. When a person approaches some building for the first time, this person would need to observe the building directly, but would also need to retrieve a specific schema for understanding this building, depending on whether it is a house or an office or another kind of building. Merely retrieving a general building schema would not be informative enough. However, retrieving the wrong schema would also lead to difficulties. For example, an incorrect assumption that someone's private house is actually a place of worship would lead to

further incorrect assumptions about whether it is all right to enter without knocking.

Another way that schemas have structure is that they are used to organize knowledge into compartments or slots. For example, a house schema would have slots for the house's address, the name of its owner, what color it is painted, how many floors it has, and so on. Learning about some particular house can be thought of as obtaining fillers for these slots, such as filling in the address and the owner's name. (This process is also known as instantiating the schema.) In this way, schemas guide us to look for certain information that is relevant. Information that does not correspond to an existing slot can easily be lost. For example, if the house has its own roller skating rink then details about this (such as its color and size) could be forgotten because there are no slots for roller skating rink details in the house schema.

Perhaps the greatest value of schemas is that they provide defaults, that can fill slots with expected values, even before direct observations are made. For example, it can be assumed by default that rooms have ceilings, kitchens have ovens, and bathrooms have baths even without checking all these details. In this way, schemas can potentially save a lot of time. Imagine the inconvenience if each time someone entered a room, it could not be assumed that the room had a ceiling and a floor. However, these default values can be overridden, so that for example it is easy to note that a bathroom does not actually have a bath but has a shower instead.

SCHANK'S SCRIPTS

When schemas refer to events they are usually known as scripts. For example, according to script theory people have scripts for going to the doctor, buying groceries, going to a restaurant, and so on, that represent their general knowledge about these everyday events. Roger Schank was interested in how people apply these scripts to understanding events. He was also interested in characterizing this human activity in a formal manner that could be expressed as a grammar or a computer program. If people's use of schemas can be formalized, then this characterization could be used to help build more intelligent computer programs that can understand stories or draw common-sense inferences.

For example, consider the following story:

> John went to a restaurant. He ordered chicken. He left a large tip.

Crucial to understanding this story is filling in the blanks, for example that John also sat down, he read the menu, the waiter brought the chicken, John ate the chicken, and he paid the bill. These inferences depend on the causal knowledge that would be contained in scripts, for example that ordering food causes the waiter to bring the food.

Script theory also provides a means for embedding scripts within each other. For example, eating chicken can be thought of as filling a slot within the restaurant script, but eating chicken can also be thought of as a script-based activity in itself, which indeed is a specialized version of a more general eating script.

EVIDENCE FOR THE USE OF SCHEMAS

The influence of Bartlett's studies can be seen in later psychological experiments in the 1970s and 1980s, that provided further evidence for schema theory and script theory. Brewer and Treyens looked at how people remembered directly observed scenes, such as what they recalled after waiting in a graduate student's office in a university laboratory. It appeared that people have an office schema that is applied to observations of real offices. This schema had a number of effects on what was later remembered about the office. One effect was that the schema seemed to support memory for details that were part of the schema: for example, people accurately recalled that the office had a desk and chairs. However, the schema also acted as a filter: participants left out certain details, such as a skull sitting on a shelf, that did not fit with the office schema. Another effect was that people used their schemas to draw (incorrect) inferences: people recalled having seen books on the shelves although there were actually no books in the office. It is quite natural to think of these results in terms of slots, fillers, and defaults for the office schema.

There is also good experimental evidence supporting the basic elements of script theory. For example, Bower, Black, and Turner asked people about their knowledge about going to a restaurant. It was found that people generally agreed about the key elements of this script, with a stereotypical sequence of being seated, looking at the menu, ordering the meal, eating, paying the bill, and leaving. This general knowledge about restaurants affected what participants remembered after they read stories about people in restaurants. As in the other studies, participants' previous knowledge seemed to affect what they remembered as much

as what they actually saw in the story. The use of inference was common; for example, participants would recall that the people in the story ate their food, even if this was not stated in the story. A more dramatic effect was that script knowledge led to changes in the order of remembered events. For example, when people in the story paid the bill before reading the menu, the story was often recalled with these events in the more conventional order.

PDP APPROACH

Some of the more recent work in schema theory has involved further efforts to implement schemas in a form that can be run by a computer program. Rumelhart and colleagues took a different approach than Schank, implementing schema theory with connectionist or PDP (parallel distributed processing) networks. These networks represent knowledge in terms of a pattern of associations between various units, which could correspond to default values for various schema. For example, the schema for a kitchen may be represented as associations between units for oven, cupboard, coffeepot, fridge, and so on.

PDP networks can represent positive as well as negative associations. So the same network could represent positive associations between sofas, easy chairs, and televisions, as well as a negative association between sofa and fridge. In this way, the network could embody multiple schemas such as the kitchen schema and the living room schema. Furthermore, the network could be used to draw inferences, such as that when a sofa is present a chair is also likely to be present, but a fridge and oven are unlikely to be present. Implicitly, the network would draw on its knowledge of the living room schema to make this inference. Although not capturing all the causal knowledge assumed by Schank's script theory, the PDP approach still provides an elegant way of implementing mathematical models that can be used to recognize schemas and apply them to observations. In particular, the PDP approach gives a successful account of constraints between fillers of different slots.

CONCLUSION

In the present day, schema theory in itself is not as much a direct object of research as it was during earlier periods in the history of psychology. Indeed, modern researchers do not necessarily assume that schemas exist in the exact forms that have been proposed. However, the influence of schema theory on current psychological research is thorough and pervasive. It is now taken for granted that modern theories within cognitive psychology, addressing memory, reasoning, comprehension, and so on, must take into account people's use of previous knowledge. Furthermore, it is accepted that schematic knowledge is richly structured, for example, with different levels of generality and with interrelations between different features of what we may observe.

Further Reading

Alba JW and Hasher L (1983) Is memory schematic? *Psychological Bulletin* **93**: 203–231.

Bartlett FC (1932) *Remembering*. Cambridge, UK: Cambridge University Press.

Brewer WF and Nakamura GV (1984) The nature and functions of schemas. In: Wyer RS and Srull RK (eds) *Handbook of Social Cognition*, vol. 1. Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart DE and Ortony A (1977) The representation of knowledge in memory. In: Anderson RC, Spiro RJ and Montague WC (eds) *Schooling and the Acquisition of Knowledge*. Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart DE, Smolensky P, McClelland JL and Hinton GE (1986) Schemata and sequential thought processes in PDP models. In: McClelland JL and Rumelhart DE (eds) *Parallel Distributed Processing*, vol. 2. Cambridge, MA: MIT Press.

Schank RC and Abelson RP (1977) *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Scientific Thought

Introductory article

Kevin Dunbar, Dartmouth College, New Hampshire, USA

CONTENTS

'Scientific thought' refers to the thought processes involved in the wide field of scientific activity, and in the process of radical or mundane revision of scientific concepts through a variety of reasoning techniques and problem-solving heuristics.

INTRODUCTION

Scientific thought refers to the thought processes involved in scientific activities, ranging from theory building and experimental design to data interpretation and scientific argumentation. More deeply considered, scientific thought involves a process of conceptual change whereby individuals and entire scientific fields both add to and radically change scientific concepts using conceptual and empirical processes. This process of conceptual change is accomplished using categorical, analogical, distributed, and causal reasoning as well as specific problem-solving heuristics.

Scientific thought is obviously a product of the human mind and uses the same cognitive processes that all human beings use in other domains such as football, art, and business. What differentiates science from these other human activities are the particular ways that the cognitive processes are combined; rather than merely using a particular cognitive process, such as induction, many cognitive processes are used and combined in scientific thought. For example, where one inductive reasoning process, analogy, might be used on its own in art, business, war, or football, scientists frequently use analogy along with causal reasoning and induction by generalization when they encounter unexpected findings, or analogy along with causal reasoning and deduction when designing experiments. Thus, while scientific thought is composed of the whole range of human thought processes, these processes are used in very specific ways in science; further, the ways that these cognitive processes are combined can vary even within disciplines, such as in physics when the thinking of experimental and theoretical physicists are compared.

Five main approaches have been used to investigate scientific thought: (1) focusing on the reasoning strategies of important scientists such as Faraday and Einstein; (2) conducting detailed investigations of the evolution of concepts in particular fields, such as the concept of disease; (3) investigating the acquisition of particular scientific concepts in adults and children by conducting experiments in the cognitive laboratory ('*in vitro*' cognition); (4) analyzing scientists *in vivo* as they think and reason at work in their laboratories; and (5) building computational models of key scientific reasoning heuristics and discoveries. Taken together, these different approaches to investigating scientific thought have converged upon a set of key mechanisms that are at the core of the scientific mind. This article will thus focus on the underlying mechanisms rather than the different types of approaches.

SCIENTIFIC CONCEPTUAL CHANGE

With the shift in philosophy of science from purely analytic methods to an understanding of the conceptual structures in science and the origins of scientific revolutions, there has been a shift to understanding the changes in the ways that scientific knowledge is represented in entire scientific fields, in particular scientists, and in children. These analyses have revealed that often conceptual change is little more than an addition or deletion of a certain fact to a theory or model. More rarely, an entire conceptual structure has to be reorganized. This latter type of change is what is most commonly referred to as radical conceptual change.

Radical conceptual change is evident in entire fields, in the work of particular scientists such as Faraday, and in children acquiring new scientific knowledge such as that involved in physics and biology. Computationally, this type of far-reaching

conceptual reorganization has been captured in models such as 'Explanatory Coherence by Harmany Optimization' (ECHO) and 'Scientific Discovery as Dual Search' (SDDS). In ECHO, explanatory coherence is the driving force behind theory change in science. In SDDS theory, change is proposed to involve shifts in frames or schemas as a function of obtaining findings inconsistent with a particular hypothesis. These models also capture the more mundane conceptual changes that are most common in science. The mechanisms underlying scientific conceptual change range from inductive processes, such as generalization and abductive reasoning, to the use of deductive reasoning heuristics.

One important question for research on scientific thought is whether the types of mental processes underlying radical conceptual change are the same as or different from those underlying more mundane conceptual changes. Some authors have argued that the processes involved in these different kinds of change are fundamentally different. Others have argued that they are the same. Answers to this question are important as they have many implications for models of scientific thinking. Thus, this is not merely an academic question, as the goal of much work in science education, and science itself, is to foster conceptual change in scientific concepts. Currently, detailed analyses of children's and scientists' subtle changes in concepts over time are being conducted and should make it possible to both understand and harness conceptual change in development, to aid practising scientists, and to build computational tools that foster both radical and mundane conceptual change in areas such as bioinformatics and cosmology.

ANALOGICAL REASONING IN SCIENCE

Scientists frequently refer to the important roles that analogy has had on their research. There are numerous anecdotes of particular analogies being the impetus for a conceptual change. Kekulé's analogy to snakes being the origin of the benzene ring is a frequently given example of the role of analogy in science. As with Kekulé's snake analogy, many scientists have talked about analogies where the source and the target are from radically different domains, such as Francois Jacob saying that by drawing an analogy between the genes he was working on and his son's toy he was able to propose the first theory of genetic regulation. Many theories of creativity in science have argued for the importance of these distant analogies.

However, recent analyses of the historical uses of analogy in science and detailed analyses of scientists' reasoning live in their laboratories reveals that the use of distant analogies in science is usually, though not always, limited to explanations. Instead, these analyses reveal that analogies to concepts and methods from the *same* domain are frequently the source of new hypotheses and are important in conceptual change. When scientists have been examined live, or *in vivo*, it has been found that a whole sequence of different analogies from the same domain can produce conceptual change. Rather than one distant analogy being the source of a conceptual change, many analogies can be the source of different aspects of the conceptual change.

Analogies in science are often used in the service of particular goals such as discovering why a certain unexpected finding was obtained. The ways that analogy is used in science are thus highly related to the goals of the scientist. If the goal is to fix an experimental problem, the analogy will be to similar experiments. If the goal is to formulate hypotheses, the analogy is to something in a related domain that shares underlying sets of relations. However, if the goal is to explain a concept to the general public or to nonspecialists in an area, then analogies are often made to very different domains.

SCIENTIFIC CAUSAL REASONING

One important component of scientific thought is the construction of causal models. These models frequently provide a particular mechanism for going from a cause to an effect, such as a mechanism for how HIV causes AIDS, and also captures statistical information about the co-occurrence of cause and effect. Causal reasoning is ubiquitous in science and is involved in experimental design, theory building, and accounting for unexpected findings. While the causal reasoning literature has tended to focus on either mechanistic or statistical information as the main source of information on causal reasoning, scientists use both statistical and mechanistic information to constrain their causal models.

Research on experimental design and hypothesis testing has frequently shown that people use underlying knowledge of causal mechanisms to interpret patterns of data. In certain circumstances scientists use their causal models to override the statistical information that they receive. This has been referred to as 'confirmation bias', and has been frequently documented *in vitro* (standard

psychological experiments) in the cognitive laboratory. However, expert scientists appear sensitive to both the scope of their statistical information (the number of samples that it comes from) and the underlying base rates when interpreting their data. Thus, while some spectacular examples of confirmation bias exist, such as the discovery of cold fusion, the actual causes and frequency of this phenomenon have been difficult to determine.

One important place where causal reasoning is used is in the design of experiments. Both *in vitro* and *in vivo* research on experimental design has shown that the possibility of error in an experimental result is a key factor that dictates the types of experiments conducted and the interpretation of results. When scientists reason at their lab meetings, they spend a considerable amount of time planning controls into their experiments that specifically control for the possibility of error. However, some research has shown that when people are told that there is the possibility of error in experimental results, this allows them to insulate their hypotheses against disconfirmation and allows them to maintain disproven hypotheses. One way that scientists deal with the issue of error is to build various different types of controls into their experiments that allow them to isolate a particular cause of the error.

DISTRIBUTED SCIENTIFIC THOUGHT

The image of a lone scientist toiling away under a naked light bulb is one that has implicitly dominated cognitive science since the 1960s. However, recent historical and *in vivo* analyses of science, as well as situated models of cognition, have highlighted the fact that much of scientific thinking and reasoning takes place in groups. Often scientists build models and theories, design experiments, and explore entire conceptual spaces in groups, and the reasoning is distributed over many people rather than being inside the head of one person. This distributed reasoning in science is usually collaborative, with members of a group changing parts of the representation of a concept by adding, deleting, and replacing components of a conceptual structure. Occasionally, groups of scientists can converge on completely new conceptual structures that can change an entire field. The exploration of the interaction of situational, cognitive, and social factors in distributed scientific thought is a new emerging direction for the field of scientific thinking.

CONCLUSION

Scientific thought involves the entire gamut of thinking and reasoning processes that human beings use. It makes use of a combination of various cognitive processes such as problem-solving, concept change, and categorization, as well as analogy and causal and visual reasoning. Accounts of scientific thought focus on the way these different aspects of thought are combined and are used in the service of scientific goals such as theory building, experimental design, and scientific explanation. The ultimate goal of research on scientific thought is to understand the mechanics and development of the scientific mind and the relationship between cognition of science and its social context, to aid scientists in their work, and to help educate new generations of scientists on strategies that can be used to make discoveries.

Further Reading

Boden MA (1991) *The Creative Mind: Myths and Mechanisms*. London, UK: Weidenfeld & Nicolson.

Carruthers PM, Stich S and Siegal M (eds) (2002) *The Cognitive Basis of Science*. Cambridge, UK: Cambridge University Press.

Dunbar K (1997) How scientists think: online creativity and conceptual change in science. In: Ward TB, Smith SM and Vaid S (eds) *Conceptual Structures and Processes: Emergence, Discovery and Change*. Washington, DC: APA Press.

Dunbar K (2001) What scientific thinking reveals about the nature of cognition. In: Crowley K, Schunn CD and Okada T (eds) *Designing for Science: Implications from Everyday, Classroom, and Professional Settings*. Hillsdale, NJ: Lawrence Erlbaum.

Dunbar K and Blanchette I (2001) The in vivo/in vitro approach to cognition: the case of analogy. *Trends in Cognitive Sciences* **5**: 334–339.

Feist S and Gorman M (1998) The psychology of science: review and integration of a nascent discipline. *Review of General Psychology* **2**(1): 3–47.

Giere RN (ed.) (1992) *Minnesota Studies in the Philosophy of Science*, vol. 15: *Cognitive Models of Science*. Minneapolis, MN: University of Minnesota Press.

Klahr D with Dunbar K, Fay A, Penner D and Schunn C (2000) *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.

Langley P (2000) The computational support of scientific discovery. *International Journal of Human–Computer Studies* **53**: 393–410.

Langley P, Simon HA, Bradshaw GL and Zytkow JM (1987) *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, MA: MIT Press.

Magnani L, Nersessian N and Thagard P (eds) (1999) *Model Based Reasoning in Scientific Discovery*. New York, NY: Kluwer.

Nersessian N (1990) *Faraday to Einstein: Constructing Meaning in Scientific Theories*. Dordecht, Netherlands: Martinus Nijhoff.
Simon HA (1979) *Models of Thought*. New Haven, CT: Yale University Press.
Thagard P (1999) *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.
Tweney RD and Chitwood ST (1995) Scientific reasoning. In: Newstead SE and Evans JSBT (eds) *Perspectives on Thinking and Reasoning: Essays in Honor of Peter Wason*, pp. 241–260. Hillsdale, NJ: Lawrence Erlbaum.
Tweney RD, Doherty ME and Mynatt CR (eds) (1981) *On Scientific Thinking*. New York, NY: Columbia University Press.

Selective Attention

Introductory article

Asher Cohen, The Hebrew University, Jerusalem, Israel

CONTENTS

A fundamental aspect of our cognitive activity is selection, by attentional mechanisms, of a portion of the vast amount of information we are confronting at any moment.

INTRODUCTION

A fundamental empirical phenomenon in human cognition is its limitation. At any moment in time, a vast amount of information impinges upon our senses. Many studies show that we cannot fully process all this information, and that some of it appears to be lost. One trademark of a limited system is its need for selection. Given that not all the impinged information can be processed, it is mandatory to select which portion of it will be preferred. In theory, the selection can be random. However, people are able to perform a nonrandom selection. For example, drivers at a junction with traffic lights are able to focus on the lights rather than on other stimuli present in the scene. The mechanism in charge of the selection is termed 'selective attention'.

Any type of selection presupposes the availability of some information in order to perform the very selection. Thus, some 'pre-attentive' processing must be performed prior to the operation of selective attention, and its output is used for the selection. The distinction between pre-attentive and attentive processing is essential in the study of selective attention.

Extensive research over the last 50 years has explored the basic properties of selective attention. Many issues related to attention have been clarified, but questions concerning its operation are still debated among researchers. Perhaps the most basic question concerns the point in the processing stream of information where attention begins to operate. This issue, developed into a controversy known as 'early versus late selection', is reviewed first. We then review studies suggesting that there may be multiple levels in which selective attention operates.

EARLY VERSUS LATE SELECTION

It is useful to consider task performance as a stream of information processing starting with input (usually via our senses) and ending with output (usually, some behavioral action). A major question concerns the locus of processing at which selection is performed. As mentioned before, there is some initial pre-attentive processing at the input side, but up to what point? At what stage of processing does selection (and selective attention) begin? Because pre-attentive processing is by definition unlimited, whereas post-attentive processing is limited, one may answer this question by uncovering the point where limitation is first evident. This point is often called 'the bottleneck'. Two classes of studies, focal attention and divided attention, were used to explore this question.

Focal Attention Studies

In focal attention studies, subjects are required to focus on a subset of the stimuli presented to them and ignore all other stimuli in the scene. We focus on one such paradigm, the dichotic listening paradigm, but findings from other paradigms are similar. In a typical dichotic listening experiment, two auditory messages (e.g. two stories) are played simultaneously. Subjects are asked to monitor one message, usually by shadowing it (i.e. repeating verbatim), and ignore the other. Studies with this paradigm revealed an important difference between two types of task. In one type, the two messages differ by a physical property. For example, the messages may differ in their intensity or by their pitch (male versus female). In the second type, the messages differ by semantic content. For example, words denoting animate and inanimate objects are played simultaneously, and subjects are required to shadow the animate words.

Early studies showed a dramatic difference between these two types of studies. When the messages differed in semantic content, subjects simply

failed to perform the shadowing task. When the messages differed in a physical property, subjects could perform the shadowing task. This shadowing ability, however, was coupled with a profound inability to report the content of the ignored message. In one study the ignored message was repeated 35 times. In another study, the language of the ignored message was changed from English to German. Both changes were not noticed by the subjects. In contrast, subjects did notice a change of a physical property in the ignored message. For example, subjects immediately notice when the gender of the speaker of the ignored message changes. There is then a fundamental difference between processing of physical and semantic properties of stimuli. Moreover, processing of semantic information in a rejected/ignored message is dramatically limited.

To capture these findings, Donald Broadbent proposed his influential early selection model of selective attention. According to this model, physical properties in the scene are processed in parallel and without any limitation. To process any semantic content, attention selects a physical property and acts like a filter: the semantic content carried by the selected physical property is recovered by higher level processes. Semantic information not carried by this physical property is lost. For example, attention may select a range of pitches corresponding to a female voice. Consequently, the semantic content carried by the female voice is processed, but other kinds of semantic information in the scene is lost. It is an 'early selection' model because selection is done early in the stream of information processing, at a point where only physical properties are available.

Subsequent studies, however, indicated that pre-attentive processing is not as limited. Although subjects do not generally notice the content of ignored messages, they sometimes detect in them important information (e.g. their own name). Anne Treisman proposed her attenuation model to accommodate these findings. This model resembles that of Broadbent, with an important modification: the attention filter attenuates rather than blocks other stimuli. The implications of this modification are best understood with another important general assumption of standard cognitive models concerning the way we represent stimuli in our brain and how such representations reach consciousness. Namely, each representation has a variable level of activation. To reach consciousness, a representation has to accumulate a high level of activation. The resting level of activation of most representations is low and thus outside consciousness. Through perceptual processes, a stimulus impinging on our senses causes an increase at the level of activation of its representation that eventually leads to its conscious recognition. The resting level of different representations differs. Important representations or representations relevant to the current cognitive context have a higher resting level of activation and consequently need a smaller additional activation to be consciously recognized. We can now appreciate the difference between Treisman's and Broadbent's models. According to the attenuation model, non-selected stimuli can be processed as well, albeit to a lesser extent. However, if the resting level of activation of their representations is sufficiently high, the attenuated processing may still cause the representation to reach consciousness. This explains how subjects sometimes notice their name, a representation with a presumably high resting level of activation, in the ignored message.

The models of Broadbent and Treisman were based on studies using subjects' conscious report of ignored messages. Other methods, however, revealed that stimuli may be processed and affect behavior indirectly even when subjects cannot report them consciously. For example, some studies showed that words presented in the ignored message and not reported by the subjects may nonetheless affect the interpretation of the attended message. Late selection models were proposed to capture these findings. According to these models, semantic information is also processed pre-attentively. The bottleneck is between extensive pre-attentive (physical and semantic) processing and conscious report rather than between physical and semantic information.

The early versus late selection debate has not been settled with focal attention studies, partly because most such studies have an inherent problem: we cannot be certain that subjects indeed focus their attention on 'attended' messages. The ability of subjects to process portions of ignored messages may be explained as occasional shifts of attention to them. Studies based on direct report of subjects are problematic for another reason: ample evidence suggests that stimuli are often processed without being consciously reported. Late selection models can dismiss poor direct report of ignored messages as a reflection of interfering processes that mask the pre-attentive processing of these messages.

Divided Attention Studies

Focal attention studies led to 'bottleneck' theories. Another class of explanations emerged from

divided attention studies in which subjects are typically asked to perform two tasks simultaneously. The performance in this dual task situation is compared to that of each of the individual tasks. Limitation is revealed by a decrement in the dual task performance relative to that of the individual tasks.

Initial findings with this paradigm have led to two generalizations. First, performance in the dual task is generally poorer, indicating that the cognitive system is limited. Second, subjects can, upon instructions, prefer one task over another in a semicontinuous fashion. Subjects instructed to invest 50 percent, or 60 percent, or 70 percent, and so on in one of the tasks, perform progressively better in this task relative to the other. These generalizations are readily captured by a 'limited resources' theory, stating that people have a limited but flexible amount of cognitive resources (or energy). Dual task performance is limited because of limited resources: two tasks cannot receive simultaneously the same amount of resources as individual tasks. Because the resources are flexible, their division among the tasks can vary, leading to a better performance in the task with the added resources.

Later studies, however, revealed that a simple resources theory is not adequate. The main problem is that dual task performance is better when the two tasks are dissimilar to each other. For example, when the input to the two tasks is both auditory or both visual, performance is worse than when the inputs to the two tasks are visual and auditory respectively. To account for these and similar findings, the resource theory had to assume that there are several independent pools of resources, each of which is limited. When two tasks draw on the same pool of resources, performance is limited. The more the two tasks draw on different pools of resources, the less limited is the dual task performance. There are few studies in which no decrement in dual task performance was observed. The multiple resource theory explains these findings by stating that the tasks in these studies draw upon entirely different pools of resources. One persistent problem with this theory, however, is that it has proven impossible to identify the nature of the independent pools of resources.

Summary

Although bottleneck and resource theories were primarily designed for focal and divided attention studies respectively, both were proposed as general accounts, and each theory was used as an explanation for both paradigms. Resources theories claim that the reduced processing of unattended messages is caused by allocation of fewer resources for this purpose. Bottleneck theories claim that decrements in dual task performance arise when both require a processing unit that can only be used for one task at a time.

The debate among the theories has not yet been settled. One difficulty is methodological: it turned out to be exceedingly difficult to control tightly the subjects' attention, a prerequisite for an unambiguous interpretation of the findings. Another possible reason is that there may be more than one source of limitation (or selection) in the stream of processing. The next section addresses this latter possibility.

MULTIPLE LEVELS OF SELECTION

The models described so far assume a single selective attention mechanism. The early versus late selection debate concerns the locus of that single mechanism. The literature, however, suggests that there may be at least two distinct levels in which selection may take place, with distinct mechanisms operating in each of these levels. There is a high-level selection used for strategic choices such as a preference of one task over another, or a shift from one task to another. It is often stated that strategic selection is performed by a set of processes called 'executive functions'. There is also a second, lower-level selection mechanism that may even be modality-specific.

Selection by Executive Functions

We are constantly facing strategic cognitive choices in our everyday life. At a larger scale we decide on the activities in which we want to be involved. At a smaller scale we are often faced with several possible tasks and need to decide which has a higher priority or when to shift from one task to another. For example, we can be engaged in driving, listening to the radio, and talking to a friend. We may decide to carry out all these activities simultaneously, but we often assign different priorities to the tasks, and can shift these priorities with changing conditions. The executive functions perform these control activities.

We focus on one paradigm, known as the psychological refractory period (PRP) paradigm, because it is relevant for the question concerning the locus of selection. In this paradigm subjects are required to perform two different tasks in succession. The input to the two tasks is presented in succession as well. Subjects respond to the two tasks as fast as possible with the constraint that the

response to one task (T1) is performed before the response to the second task (T2). The main manipulation is the temporal gap, called stimulus onset asynchrony (SOA), between the presentation of the inputs to T1 and T2. Subjects are usually able to conform to the instructions, presumably by the use of executive functions: the response to T1 is committed first, and is not affected by the SOA. The response to T2, however, is typically dramatically affected by the SOA. At very short SOAs (e.g. 50 milliseconds), response to T2 is very slow. With longer SOAs, response to T2 improves progressively. Around SOA of 300–400 milliseconds, the response becomes similar to that of T2 by itself.

Why is the response to T2 affected by its temporal proximity to T1? Much evidence suggest that subjects do not assign stimuli to responses for two tasks simultaneously, a process called response selection. That is, while response selection is performed for T1, it cannot be done for T2. Instead, subjects 'wait' until response selection for T1 is complete and only then proceed to the response selection of T2.

These findings led some researchers to suggest that a 'bottleneck' in the stream of processing occurs at the response selection stage. Others claim that this apparent bottleneck at the response selection stage is only a reflection of decisions by the executive functions to allocate all the resources to response selection of T1. In other words, both 'bottleneck' and 'resources' theories can explain these findings. Regardless, we see how selection of T1 in the expense of T2 is presumably done by selective attention mechanisms related to the executive functions.

Selection within Modalities

The preceding section showed how attention selects between tasks. Attention also selects within a single task, even when the task is exceedingly simple. We focus on the visual modality because most of the research in this domain used visual tasks. Selections also take place in other modalities.

Imagine a task where you are required to make a single response to a stimulus when it appears inside one of two boxes located to your right and left. The location of the stimulus is not relevant to your response because the same response is required for the two locations. Much research suggests that if you are cued in advance that the target is likely to appear in one side, your response is faster when the target indeed appears in the cued location and is slower when the target appears in the other side, relative to situations with no cueing. The costs and benefits from the cueing are ascribed to the operation of visual attention. Attention operates by selecting the location (or the box) of the cued area, leading to facilitation in the response to targets within the selected area.

As in other phenomena of selection, there are disagreements concerning the locus in which attention affects processing. It could affect perception of the target, or it could affect response decisions for the target, or more resources are assigned to the selected area. Note that this selection is observed even for exceedingly simple tasks. There may be additional types of selection required for more complex task. These, however, are more controversial and will not be reviewed here.

Relation between the Two Types of Selection

The selection between two tasks, and the selection within a single task appear quite different. Indeed, there are behavioral and neuropsychological studies that support the separation of these two types of selection. Behaviorally, there is evidence that, although (as noted above) subjects in the PRP paradigm do not select responses to task 2 when performing task 1, they are able to shift their visual attention for task 2 during the performance of task 1. This suggests that response selection, done by the executive functions, and visual attention are distinct. Data from neurologically impaired patients and from imaging techniques suggest the existence of a posterior system of attention, presumably dedicated to lower-level selections (e.g. visual attention), and an anterior system dealing with executive functions.

Further Reading

Duncan J (1980) The demonstration of capacity limitation. *Cognitive Psychology* **12**: 75–96.

Johnston JC, McCann RS and Remington RW (1995) Chronometric evidence for two types of attention. *Psychological Science* **6**: 365–369.

Kahneman D and Treisman AM (1984) Changing views of attention and automaticity. In: Parasuraman R and Davies DR (eds) *Varieties of Attention*, pp. 29–61. Orlando: Academic Press.

Lavie N (1995) Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance* **21**: 451–468.

Pashler (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Posner MI (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* **32**: 3–25.

Posner MI and Petersen SE (1990) The attention system of the human brain. *Annual Review of Neuroscience* **13**: 25–42.

Shiffrin RM (1988) Attention. In: Atkinson RC, Herrnstein RJ, Lindzay G and Luce RD (eds) *Steven's Handbook of Experimental Psychology: Volume 2, Learning and Cognition*, pp. 739–811. New York: John Wiley & Sons.

Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* **87**: 127–190.

Yantis Y and Johnston JC (1990) On the locus of visual selection: evidence from focused attention tasks. *Journal of Experimental Psychology: Human Perception and Performance* **16**: 135–149.

Self, Psychology of

Introductory article

Roy F Baumeister, Case Western Reserve University, Cleveland, Ohio, USA

CONTENTS

Social psychology of self is the study of structures and processes through which people know and evaluate themselves, present themselves to others, and exert control.

Psychology's efforts to understand the self date to its earliest thinkers, including William James, Sigmund Freud, and Carl Jung, and in recent decades it has been a pervasive theme among researchers. Under the broad umbrella of studying the self, many different subtopics have flourished. Interest has been sustained by the importance of many findings and questions, by lively debates among researchers with different theories, and by the fundamental fascination with the topic, as captured in the ancient Greek philosopher's injunction to 'Know thyself!'

The vast quantity of research on the self can be conveniently grouped into three broad categories. First, the cognitive processes include self-concepts and other representations of self, self-awareness, self-esteem, and other links between the self and the processing of social information. Second, interpersonal processes include managing how the self is perceived by others, identification of self as part of social groups, and altering the self to fit the communications and expectations of others. Third, the executive function involves how the self acts and decides, as well as how the self regulates and controls itself. These three categories roughly cover the entire field and can be used to organize the available information, even though inevitably there is some overlap between them.

SELF-CONCEPT AND COGNITIVE PROCESSES

Although experts believe that rudimentary signs of self-awareness can be observed in other species (especially primates), the capacity for being aware of self is much greater and more developed in human beings. Awareness of self makes it possible for people to form mental representations and concepts of themselves and to regulate their own behavior. In particular, self-awareness typically involves comparing oneself to goals, ideals, norms, other people, and various other standards.

The self-concept refers to a mental representation of a person by that same person. Typically this encompasses one's social identity in the sense of one's roles, as well as one's personality traits and to some extent one's major values and beliefs. The term is used loosely by researchers such that any knowledge or beliefs about the self can be considered to be part of that person's self-concept.

The simple term 'self-concept' is misleading in some ways, because it implies a single, unified, consistent representation of the self. Contrary to that view, people seem to have multiple beliefs about themselves, and these are not necessarily integrated or consistent with each other. To be sure, people do make efforts to resolve obvious contradictions, but many inconsistencies can survive without being recognized as such. Different versions of the self-concept may dominate at different times and in different contexts, such as when a young person thinks and feels differently when attending church with his parents as opposed to partying with fraternity brothers. Some theorists have taken this inner diversity as a reason to speak of having 'multiple selves', but it is better understood as a multiplicity of conceptions or of versions of the same self.

In any case, it is clear that different conceptions of the self may dominate at different times. People have a great stock of self-knowledge, and at any given time they use only a small part of it. That is why they can have inconsistent and even contradictory views of self.

Three main motives dominate the acquisition, memory, and use of information about the self. First, the positivity motive involves the desire to think well of oneself and hence exerts an ongoing preference to learn favorable things that reflect well on the self. This motive is reflected in people's

desire to obtain positive feedback and in their tendency to distort memory so as to portray the self in a more flattering light. Second, the consistency motive refers to a wish to maintain a stable, unchanging view of self, and so it exerts a preference for information that confirms what the person already believes about the self. People report skepticism and even rejection of information that disconfirms what they believe about themselves. Last, the diagnosticity motive signifies a desire to learn accurate, correct information about the self. Sometimes people show an interest in performing tasks (such as those of intermediate difficulty) that will provide the clearest and most unambiguous information about the self and what it can do.

Because accurate knowledge would seemingly be the most useful of the three types of information, there is reason to regard the diagnosticity motive as the most adaptive and important of these motives. However, research consistently finds that the positivity motive is the strongest, with consistency second and diagnosticity the weakest. All three motives are genuine and exert some influence, however.

Partly because of conflicting motives, people end up with self-concepts that are not strictly accurate. In particular, there is ample evidence that people hold favorable, flattering, and comforting assessments of themselves. More precisely, self-knowledge shows three persistent patterns of distortion. First, people overestimate their good qualities, achievements, and strengths while underestimating or downplaying their faults and misdeeds. Second, people overestimate their degree of control over the events that affect their lives. Third, people are unrealistically optimistic and hence prone to expect that good things are more likely (and bad things more unlikely) to happen to them than to the average person. These distortions may sound like narcissistic delusions of grandeur, but in fact they are common and pervasive among normal, psychologically healthy individuals. They are less common among people with low self-esteem and among depressed people, whose self-assessments tend to be more even-handed and accurate.

Self-esteem refers to the favorability of the self-concept. People with high self-esteem think well of themselves. Low self-esteem does not however usually indicate a strongly negative view of self, but more commonly low self-esteem refers simply to the absence of favorable views (thus indicating a neutral, intermediate view of self). People with low self-esteem also tend to exhibit 'self-concept confusion' in the sense that their beliefs about themselves are likely to be inconsistent, contradictory, tentative, and uncertain. A difference in broad approach to life is also apparent: people with high self-esteem seek to enhance themselves by accomplishing great things and winning new friends and admirers, whereas people with low self-esteem seek first to protect themselves against failure, rejection, and embarrassment. People high in self-esteem seek to identify their strengths and capitalize on them, whereas people low in self-esteem try to identify their weaknesses and remedy them.

The benefits of high self-esteem have been debated. High self-esteem appears to confer confidence, which can help performance, although in most studies people with high self-esteem do not perform any better (or much better) than people with low self-esteem – they merely regard their performance as better. It confers a stock of good feelings that can be useful to help people recover from setbacks and traumas and to help them persist in the face of discouraging failure. People with low self-esteem are more prone to give up easily. Thinking well of oneself does not appear to produce much else in the way of benefits, such as being better citizens, avoiding self-destructive or interpersonally harmful acts, or accomplishing more in life. Some forms of high self-esteem, such as a narcissistic belief that one is superior to others and entitled to special, preferential treatment, have been associated with aggressive and destructive tendencies. By and large, self-esteem has at best weak correlations with success in life, and even when such links have been found, there is reason to think that self-esteem is the outcome rather than the cause of success.

INTERPERSONAL PROCESSES

The self is not just a stock of information but rather also comprises a tool for interacting with others. Hence the self is strongly shaped by interpersonal processes and will be modified by social interactions.

Social identity is formed by recognizing oneself as a member of various groups and categories. From early in life, self-knowledge involves being a member of a particular family and having a particular gender. Identification with various ethnic groups, nationalities, organizations, and other categories continues to be important throughout life. For some (but not all) people, self-esteem and self-knowledge are heavily intertwined with these collective aspects of the self (such as racial pride). Immersion of self in a group sets up a

conflict between the motive to be distinctive in some way versus the motive to fit in to the group, and there is some evidence that cultures differ in the relative emphasis they put on them. Specifically, Asian cultures appear to emphasize the collective aspects of the self and people try to fit in with others, whereas Western cultures foster more individuality and people expend more effort cultivating independence and distinctiveness. Indeed, this idea was proposed in a highly influential article by Hazel Markus and Shinobu Kitayama (1991) and has become one of the most heavily studied theories in cross-cultural work on the psychology of self. In support of it, Westerners tend to emphasize aspects of self-concept that distinguish them from others, whereas Asians may focus on aspects of self that involve relationships to others and common traits that link one to one's social groups.

Undoubtedly much self-knowledge is distilled from social interactions. In fact, symbolic interactionist theories emphasized that the inner self is essentially derived from interpersonal, social events. The simplest version of this view is that people learn about themselves from others, but the evidence has not supported this simple version. That is, people's self-concepts do not closely match the way they are perceived by their friends and family. Two significant processes prevent people from coming to see themselves as others see them. First, people do not communicate all their impressions accurately, especially when it comes to telling someone his or her faults. Therefore people do not learn the full story about what others think of them. Second, individual self-knowledge is subject to motivated biases, and so people discount what others may tell them if they do not want to hear it.

People also seek to mold how others perceive them (impression management, also known as self-presentation). Often they try to portray themselves in a very positive manner, such as by concealing their faults and emphasizing or exaggerating their good qualities. Sometimes this is a simple matter of trying to conform to the expectations of others, such as when a person might conceal habits of laziness or drunkenness during a job interview. More broadly, people try to present themselves to others as a way of claiming desired identities for themselves. Theorists dating back at least to the symbolic interactionist school have emphasized that identity requires social validation, and so before someone can claim to own a specific identity it is necessary to persuade or induce some other people to perceive oneself as having that identity.

EXECUTIVE FUNCTION

The executive function refers to how the self exerts volition, such as by making choices and decisions, initiating action (as opposed to being passive), and regulating the self. This operates primarily by means of deliberate, conscious control. It is obvious that in most other species behavior can occur steadily and automatically without the involvement of anything resembling a human self. Even among human beings, most behavior may well occur as a result of automatic processes, habit and routine, natural impulses, and situational pressures. The operation of the executive function is akin to the traditional concept of 'free will' (in which the person defies impulses and external pressures in order to act in a carefully chosen manner) and encompasses only a small proportion of all behavior – but nonetheless a very powerful and influential one. People are motivated to seek and maintain control over what happens to them, even though they do not use this control all the time.

The regulation of self (akin to the popular notion of self-control) is an important manifestation of the executive function. The regulation of self is accomplished by means of a feedback loop. This conscious process compares the self to a given standard, and if there is an unwanted discrepancy, it initiates some action to remedy this. Once the discrepancy is resolved (as indicated by comparing the perceived self to the goal or standard once again), it terminates the process.

The process by which the self changes itself appears to involve some form of energy, not unlike the traditional notion of 'willpower'. This force is used to override the self's initial response (such as in resisting temptation) and instead perform some alternative, more acceptable act. It appears further that other acts by the executive function, such as making choices and decisions, also draw on this same resource. These acts of self-control and choice can deplete the self's energy, leaving it less able function effectively until it is replenished. In that depleted state, people are reluctant to make decisions, assert themselves, or resist temptation. The expenditure of energy is one reason that the executive function can only be used sparingly, and so most of life has to be lived according to routine, habit, and automatic responses.

Although the executive function is undoubtedly a powerful tool for fostering the rational pursuit of enlightened self-interest, people do sometimes perform self-defeating acts (defined as acts that thwart the self's best interests) as well. Typically, these

self-defeating acts involve seeking short-term gain despite long-term risks and costs, such as in smoking cigarettes or abusing drugs. Indeed, resisting the temptation of short-term benefits in order to pursue long-term ones may be a crucial reason that the executive function (with its capacity for self-control and future orientation) may have evolved in the first place.

Further Reading

Baumeister RF (1998) The self. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, 4th edn, pp. 680–740. New York: McGraw-Hill.

Campbell JD (1990) Self-esteem and clarity of the self-concept. *Journal of Personality and Social Psychology* **59**: 538–549.

Carver CS and Scheier MF (1981) *Attention and Self-Regulation*. New York: Springer-Verlag.

Higgins ET (1987) Self-discrepancy: a theory relating self and affect. *Psychological Review* **94**: 319–340.

Markus HR (1977) Self-schemata and processing information about the self. *Journal of Personality and Social Psychology* **35**: 63–78.

Markus HR and Kitayama S (1991) Culture and the self: implications for cognition, emotion, and motivation. *Psychological Review* **98**: 224–253.

Shrauger JS and Schoeneman TJ (1979) Symbolic interactionist view of self-concept: through the looking glass darkly. *Psychological Bulletin* **86**: 549–573.

Steele CM (1988) The psychology of self-affirmation: sustaining the integrity of the self. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, vol. 21, pp. 261–302. New York: Academic Press.

Taylor SE and Brown JD (1988) Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin* **103**: 193–210.

Turner RH (1976) The real self: from institution to impulse. *American Journal of Sociology* **81**: 989–1016.

Self-organizing Systems

Introductory article

Scott Camazine, Boalsburg, Pennsylvania, USA

CONTENTS

Self-organizing systems are physical and biological systems in which pattern and structure at the global level arises solely from interactions among the lower-level components of the system. The rules specifying interactions among the system's components are executed using only local information, without reference to the global pattern.

WHAT IS SELF-ORGANIZATION?

Self-organization is a process whereby pattern at the global level of a system emerges solely from interactions among the lower-level components of the system. The rules specifying the interactions among the system's components are executed using only local information, without reference to the global pattern. Examples of self-organization include a wide range of pattern formation processes in both physical and biological systems: sand grains assembling into rippled dunes, chemical reactants forming swirling spiral patterns, the patterns on sea shells, or fish swimming in coordinated schools (Figure 1). 'Pattern' is used here in a broad sense to refer not only to a particular arrangement of objects in space, but also to structure and organization in time. An example is the remarkable synchronous flashing that sometimes develops among aggregations of thousands of fireflies in southeast Asia. In neurobiology, self-organization contributes to temporal structure and anatomical organization in systems ranging from central pattern generators in simple invertebrates to cognition in humans.

In self-organizing systems, pattern and organization develop through interactions internal to the system, that is, without the intervention of external influences, such as a 'leader' who directs or oversees the process. The pattern is an emergent property of the system itself, rather than a property imposed upon the system by an external supervisory influence.

EMERGENT PROPERTIES IN A SELF-ORGANIZING SYSTEM

The term 'emergence' refers to a process by which a system of interacting elements acquires qualitatively new pattern and structure that cannot be understood simply as the superposition of the individual contributions. Although the term may suggest that something mysteriously or magically materializes within the system, this is not the case. The human mind is generally poor at predicting the properties of systems that consist of multiple components with complex, dynamic interactions. Thus, even if one has a full knowledge of the system's elements and their mode of interaction, the collective properties of a self-organizing system often seem to arise unexpectedly.

HOW DOES A SELF-ORGANIZING SYSTEM WORK?

An example may make this abstract description of self-organization and emergent properties clearer. Striped and mottled patterns are found throughout nature – on a zebra's coat, on a fish's skin, and in the ocular dominance columns of the brain (Figure 2). Experimental and theoretical work suggests that these patterns develop from a few simple rules that are continually iterated among the components of the system. Suppose, for example, that each pigment cell on a zebra's coat could either produce a dark pigment or not, depending on a certain chemical activation above or below a certain threshold level. Further suppose that the cells in the skin produced both a chemical activator and an antagonistic inhibitor (called 'morphogens'), which both diffused through the skin. The rules regulating the state of each cell – either 'on' (producing pigment) or 'off' (not producing pigment) – depend on the relative strengths of the activation and inhibition,

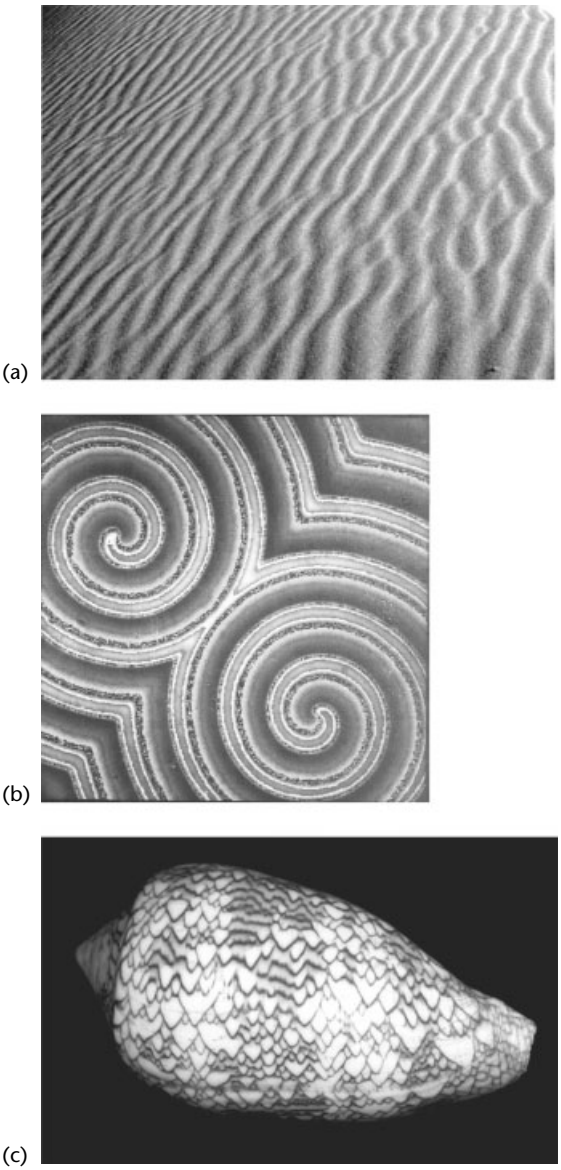


Figure 1. Examples of self-organized pattern formation in physical, chemical, and biological systems. (a) Sand dune stripes. (b) Belusov–Zhabotinsky chemical reaction (image courtesy of Stefan C. Müller). (c) A cone shell from Ceylon.

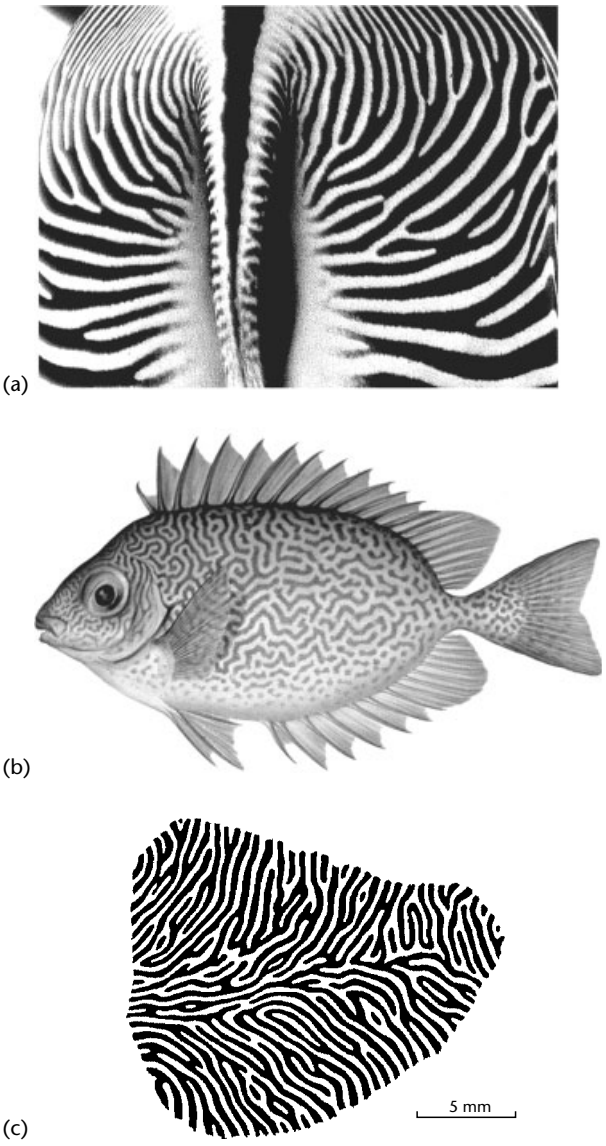


Figure 2. Striped and mottled patterns found in biological systems. (a) Alternating stripes on a zebra's coat (*Equus grevii*). (b) Mottled pattern of pigments on the skin of a vermiculated rabbitfish (*Siganus vermiculatus*). (c) Ocular dominance stripes in the visual cortex of a macaque monkey. Regions receiving inputs from one eye are shown in black, and regions receiving inputs from the other eye are shown in white. Adapted from: Hubel DH and Wiesel TN (1977) Functional architecture of the macaque monkey visual cortex. *Proceedings of the Royal Society, Series B* **198**: 1–59.

their diffusion rates, the initial distribution of the cells, and their thresholds for pigment production.

In 1952, Alan Turing first suggested the general scheme for this mechanism of self-organized pattern formation. In 1972, A. Gierer and H. Meinhardt developed a model as shown in Figure 3. Their system has a series of sites that are the source of a short-range activator, which has two functions: to promote its own productions (autocatalysis), and to cause an increase in the production of an antagonist, the inhibitor. Since the inhibitor diffuses rapidly into the surroundings, the result is a local increase in the activation and a long-range antagonistic effect that restricts the self-enhancing reaction and keeps it localized.

SIMULATION OF SELF-ORGANIZING SYSTEMS

Because of the difficulty of predicting the behavior of these systems, computer simulations are a useful means of performing 'thought experiments' and for better understanding how these systems work. One method of modeling these systems is by the use of nonlinear differential equations. Another method is to simulate the system by means of cellular automata.

A cellular automaton is a simulation that is discrete in time, space, and state. Typically, the components (cells) of the system are arranged on a two-dimensional grid or lattice. Each cell is characterized by its location on the grid and its condition (state). Cells interact with each other according to a set of simple rules which take into account their proximity to neighboring cells, their own state, and the states of their neighbors. The rules specify the transition of the cell from one state to another as the system evolves over time.

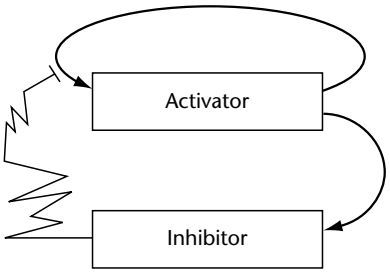


Figure 3. Reaction scheme for pattern formation by autocatalysis and long-range inhibition. The two arrows denote activation, with the activator stimulating both its own production (autocatalysis) and that of the inhibitor. The jagged line shows the effect of the inhibitor which provides negative feedback by inhibiting the effect of the activator. Adapted from: Meinhardt H (1995) *The Algorithmic Beauty of Sea Shells*. Berlin, Germany: Springer.

Consider the example of animal coat patterns presented above. This can be implemented as a cellular automaton model that consists of a set of cells laid out on a grid. Each cell is initially assigned a state randomly, 'on' or 'off'. Each 'on' cell is assumed to produce a specified amount of activator and a specified amount of inhibitor that diffuse at different rates across the grid. In the simulation, each 'on' cell is represented as black and each 'off' cell is represented as white. At each timestep, the program calculated the net amount of activation at each site on the grid. This is determined as the difference between the sum of all the activation from the cells in the neighborhood and the sum of all the inhibition from those cells in the neighborhood. If this total is above a prespecified threshold level, then the cell at that site is assigned the 'on' state; otherwise, it is assigned the 'off' state. In this manner, cells switch from one state to another according to a single rule. The program continually iterates the rule, causing a pattern to emerge from the initial random array of 'on' and 'off' cells, as shown in Figure 4. For one set of diffusion rules, an irregular mottled pattern develops. When the

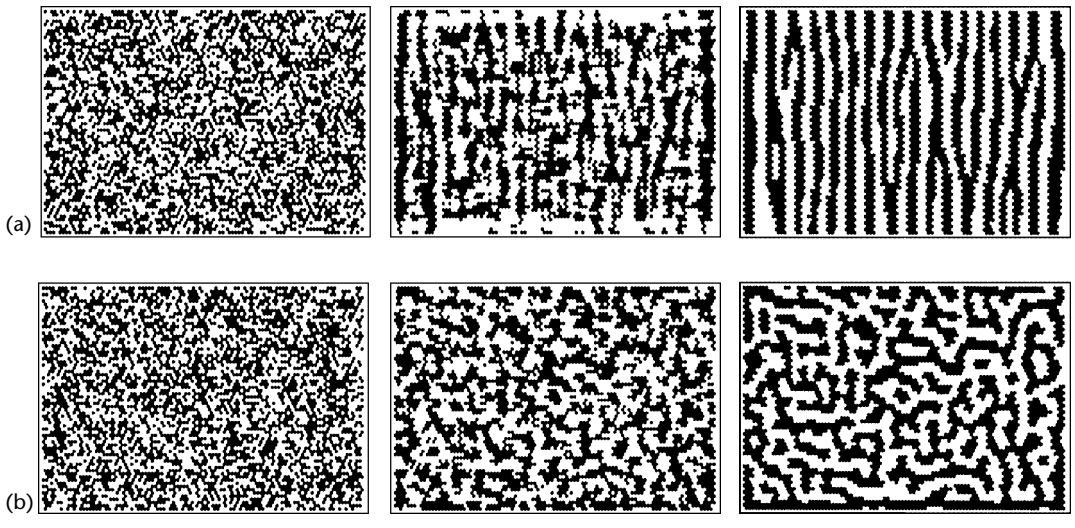


Figure 4. Cellular automaton simulations of pattern formation according to an activation–inhibition model. In each example, the first grid shows the initial random state of the system, the second grid shows an intermediate state, and the third grid shows the final stable pattern. (a) Time sequence showing a striped pattern formation, as in Figures 2(a) and 2(c). (b) Time sequence showing a mottled pattern formation, as in Figure 2(b).

conditions are changed slightly, a zebra-stripe pattern develops. The only differences between the two examples shown are that in the zebra-stripe pattern the diffusion of the activator and inhibitor is greater in one direction than the other, and that the relative strengths of the activator and inhibitor are different in the two cases.

SELF-ORGANIZATION IN THE NEURAL AND COGNITIVE SCIENCES

To understand the brain is one of the greatest challenges in biology. The brain of an insect such as a honey-bee contains relatively few neurons – approximately one million. In isolation, each neuron is essentially a simple switch. When stimulated sufficiently, an impulse is fired, and a brief electrical event called the *action potential* moves through the cell from one end to the other. Although the insect brain is miniscule, with relatively few neurons, compared with that of birds or mammals, it nonetheless coordinates very sophisticated behaviors. The honey-bee is arguably more complex than any computer. This tiny insect can navigate by the sun, fly to a food source, make decisions, communicate with other honey-bees, and perform many other complex activities.

The brain achieves this complexity largely through the connectivity of its elements and their interactions. Each neuron is connected to others through synapses, which form a vast network of dense interconnections. In ways that we are just beginning to understand, this connectivity is the basis of the brain's enormous complexity.

Neuroscientists are beginning to understand both how these connections among the neurons develop and how their interactions make cognition possible. Both of these processes rely, in large part, on self-organization. One of the great mysteries of biology is how the enormous morphogenic, physiological, behavioral, and cognitive complexity of an organism can be achieved with the limited amount of genetic information contained within the genome. It is inconceivable that the pattern of connections for each neuron in the brain could be genetically coded. Rather, there must exist special mechanisms for economizing on the amount of information that must be coded within the genes. Self-organization is such a mechanism. For example, the pattern of ocular dominance stripes in the visual cortex of the brain (Figure 2(c)) is a characteristic morphogenic feature of neuroanatomical organization. The neural inputs from each eye to the visual cortex in the back of the brain consist of a series of alternating stripes. This architecture is believed to play an important role in how the brain organizes and interprets visual information received by the retina. This functional architecture can be seen by injecting radioactive proline into one eye, and making autoradiographs of sections of the cortex. The resulting pattern is reminiscent of the stripes seen on a zebra's coat or the ridges of a sand dune. These patterns are believed to arise through a self-organizing activation–inhibition mechanism similar to that described above.

Studies such as these suggest that through natural selection, organisms can evolve mechanims that rely on relatively simple sets of rules – algorithms economically encoded in the genome. Through self-organizing processes these algorithms can generate the enormous complexity seen in biological systems. The result has been the evolution of complex morphological and physiological adaptations and behavioral and cognitive abilities.

Further Reading

Camazine S, Deneubourg JL, Franks N *et al*. (2001) *Self-Organization in Biological Systems*. Princeton, NJ: Princeton University Press.

Gierer A and Meinhardt H (1972) A theory of biological pattern formation. *Kybernetik* **12**: 30–39.

Hubel DH and Wiesel TN (1977) Functional architecture of the macaque monkey visual cortex. *Proceedings of the Royal Society, Series B* **198**: 1–59.

Meinhardt H (1995) *The Algorithmic Beauty of Sea Shells*. Berlin, Germany: Springer-Verlag.

Miller KD, Keller JB and Stryler MP (1989) Ocular dominance column development: analysis and simulation *Science* **245**: 605–615.

Swindale NV (1980) A model for the formation of ocular dominance stripes. *Proceedings of the Royal Society, Series B* **208**: 243–264

Turing A (1952) The chemical basis for morphogenesis. *Philosophical Transactions of the Royal Society* **237**: 37–72.

Semantic Memory: Computational Models

Intermediate article

Timothy T Rogers, Medical Research Council, Cambridge, UK

CONTENTS

Semantic memory encompasses knowledge about the meanings of words, objects, and events. Computational models of semantic memory describe explicit mechanisms for the storage, representation, and retrieval of semantic information. Three modeling approaches are discussed in this entry: hierarchical processing models, similarity-based categorisation models, and connectionist models.

INTRODUCTION

Semantic memory is typically defined as memory for the meaning of words and objects. It can encompass knowledge about object properties (e.g. dogs have fur, hearts, and bones) and arbitrary facts (e.g. Napoleon was defeated at Waterloo), as well as the meanings of verbs, abstract nouns, and events, though most research in the domain has focused on understanding human knowledge about concrete objects and their properties. Semantic memory is *explicit*, in that its contents are accessible to conscious awareness, and *declarative*, in that these contents may be described verbally. In contrast to episodic long-term memory, semantic memory is not typically associated with a particular place and time – for example, although most everyone knows the meaning of the word 'dog', they do not tend to remember specifically where and when they first learned this information.

The basic research question posed by most theories of semantic memory is: how do human beings know which objects in the world have which properties? For example, upon encountering an unfamiliar dog for the first time, we are able to make a broad range of inferences about it: we know that it is called a 'dog', that it barks, it likes to chase sticks, it may bite, it has internal organs, it can grow, etc. We know these things despite having never previously encountered this particular dog. What accounts for this remarkable human capacity?

One view with a long provenance in philosophy and cognitive science posits that semantic memory is subserved by a process of categorization. We can make appropriate inferences about the properties of an unfamiliar object when we know what class it belongs to, because we have stored information about these classes generally. On this view, semantic task performance is a process of logical inference: in the classical syllogism, Aristotle infers that Socrates is mortal by categorizing him as a *man*, and consulting his knowledge about the class (Socrates is a man; all men are mortal; therefore, Socrates is mortal). For much of the history of the study of cognition, the dominant view has been that semantic knowledge is stored in memory as a system of concepts and propositions of this sort. In the twentieth century this view has begun to change, thanks in part to rapid advances in computational modeling of cognitive and neural processes during the past thirty years.

Computational approaches to semantic memory fall into three broad classes, which will be reviewed in this entry: *hierarchical processing models*, such as the spreading activation model proposed by Collins and Quillian (1969); *similarity-based categorisation models*, which include category prototype theories as well as exemplar or instance-trace theories; and *connectionist* or *parallel distributed processing* (PDP) models.

HIERARCHICAL PROCESSING MODELS

One of the earliest and most influential computational theories comprised an effort to directly implement semantic memory as a system of logical inference involving stored concepts and propositions (Collins and Quillian, 1969). In this model, shown in Figure 1, different concepts are represented as discrete nodes, whereas different

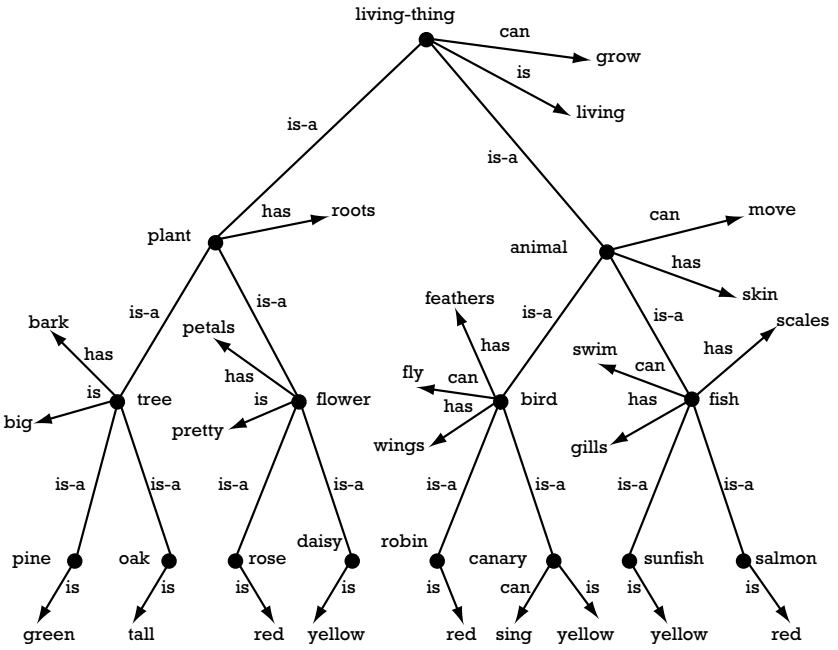


Figure 1. Collins and Quillian's (1969) hierarchical spreading-activation model. Individual concepts are represented as nodes, and propositions describing the relations between concepts are represented as links between nodes. Reprinted from Figure 4 of McClelland et al. (1995).

propositions are represented as labeled links. Semantic knowledge is represented in the model by the particular set of nodes and configuration of links that constitute the 'semantic network'. For example, the nodes labeled *robin* and *bird* in the illustration indicate that the system has stored knowledge about the class of birds and robins separately. The link between the *robin* and *bird* nodes (labelled *isa*) indicates knowledge that robins are a subset of the class *bird*. The link between *robin* and the concept *red* (labelled *is*) indicates that items in the class *robin* are likely to be red.

The diagram provides a concise description of the contents of semantic memory, but it was also intended to illustrate a processing model of semantic storage and retrieval. Information could be stored simply by adding new nodes and links to the network, or by altering the configuration of links between nodes. To store a new fact such as 'a wren is a kind of bird', one would create a new node corresponding to the concept *wren*, and attach it to the *bird* node with an *isa* link. To search memory, the authors proposed a *spreading activation* mechanism, by which the activation of a concept in memory would spread outward along stored propositional links, resulting in the activation of related concepts (and hence retrieval of semantic information stored with these nodes).

One appeal of the Collins and Quillian model stems from the observation that category membership for more inclusive categories (such as *animal*) entails properties true of more specific categories (such as *dog* or *german shepard*). These properties need only be stored with the more general category node, and will be inherited by more specific concepts as a consequence of the spreading-activation mechanism. For example, a property such as *has blood* need not be stored separately with each different animal concept (e.g. *goat*, *pig*, *dog*, *bird*) but can be stored once with the more inclusive concept *animal* and will extend to these more specific concepts by virtue of stored class inclusion links in the semantic network.

Quillian wrote a computer program that implemented the proposed architecture and memory-search procedure for semantic networks similar to the one shown in Figure 1. The implemented model allowed the authors to make empirical predictions about the processing time that would be needed to perform different kinds of semantic tasks. For example, they devised a sentence verification task in which participants listened to simple propositions such as 'all birds have wings', and had to decide as quickly as possible whether the proposition was true or not. On the basis of the model, Collins and Quillian predicted that sentences would take longer to verify when their terms were distal (i.e. separated by many links) in the semantic network than when they were proximal. Early work seemed to confirm this prediction – participants

were faster and more accurate to verify propositions separated by a single link (e.g. 'all robins are birds') than those separated by many links (e.g. 'all robins are animals'), just as was the computer program. Collins and Quillian and others used the theory to explain further data from a broad range of semantic tasks, including priming experiments, category fluency studies, and several different permutations of the sentence verification paradigm.

However, some results from the sentence-verification procedure were not explained by the original spreading-activation theory, and others directly contradicted the model's predictions. For instance, subjects are faster and more accurate at verifying the properties of typical category members (e.g. a robin is a bird) than those of atypical members (e.g. a penguin is a bird). Since both penguin and robin concepts are attached directly to the bird node in the semantic network, there is no reason why propositions about one should take longer to verify than propositions about the other. Moreover, subjects are sometimes faster to categorise atypical items at a superordinate rather than an intermediate level – verifying the sentence 'a chicken is an animal' more quickly than the sentence 'a chicken is a bird'. This result was not compatible with Quillian's implemented model, because there was no way for activation to spread in the hierarchy from *chicken* to *animal* without passing through *bird*. Finally, the empirical phenomena that were accounted for by the theory turned out to be susceptible to confounding factors other than the distance between terms in a semantic network, such as concept familiarity and typicality. Once these factors are taken into account, it is not clear that distance in the semantic network adds any explanatory power. Responding to these challenges, Collins and Loftus (1975) adapted the original framework by adding link strengths, activation thresholds for concept nodes, and various other constructs. Though they did not implement these in a computational model, these ideas have been incorporated into subsequent implemented spreading-activation-style models. (*See* **ACT**; **Semantic Networks**)

The strong influence of Collins and Quillian's (1969) theory is due in part to the specific and testable empirical predictions derived from Quillian's implemented model. Challenges to these predictions set the stage for further study. For example, where the spreading-activation theory concerned itself with how inferences were made once a given category representation had been activated in memory, it neglected the important question of how objects in the world were categorized in the first place. With the data mined from the property-verification task, several interesting theoretical questions presented themselves. Why were typical category exemplars easier to categorize than atypical exemplars? How did familiarity and frequency exert their influence on storage and retrieval processes? How was information about different classes extracted from the environment in the first place? These questions lead to an increasing focus on understanding mechanisms of categorization.

SIMILARITY-BASED CATEGORIZATION MODELS

Categorization-based theories of semantic memory come in two varieties. *Prototype theories* propose that knowledge about classes is stored in summary representations that describe the properties typical of category members (Rosch and Mervis, 1975). Usually prototypes take the form of a vector of attributes. For example, the prototype for the category *dog* might include features such as *is furry*, *isa pet*, *has four legs*, *can bark*, etc. To make appropriate inferences about a novel object, prototype models compare the object's observed properties to the known properties of stored prototypes, and assign the item to the category with the best fit according to some measure of similarity. Other properties stored with the prototype are then retrieved and attributed to the object. One appeal of prototype theories is that they offer an intuitive means of explaining the influence of typicality on semantic task performance. By definition, typical exemplars of a category have many properties in common with other members of the class and few distinguishing characteristics. Hence, typical members will more closely match the category prototype than will atypical members, and are faster and easier to categorize.

Like category prototype theories, *exemplar* or *instance-trace* (e.g. Nosofsky, 1988) theories propose that information about objects and categories is stored in discrete feature-vector representations, and is accessed by measuring the similarity between a probe stimulus and stored representations. Rather than storing summary representations of different categories, however, instance-trace theories propose that a separate representation (or memory-trace) is stored for each individual object or event experienced in the environment. During retrieval, a probe stimulus activates each trace in parallel in proportion to its similarity to the probe; for example, upon encountering a novel dog in the world, the system will activate stored traces of

other encounters with particular dogs. Properties common to these traces (e.g. *has fur*, *has four legs*, etc.) are stored with every individual trace and hence are likely to be retrieved successfully. The idiosyncratic properties of each particular dog are supported by only a small number of traces, and hence are not likely to be attributed to the unfamiliar dog.

Like prototype theories, instance-trace theories can explain typicality effects in categorization and property-verification tasks: because typical exemplars share many properties with other category exemplars, they will be similar to many stored instance traces and will activate all of these in parallel, resulting in rapid retrieval of properties shared by the class. However, instance-trace theories can also explain experimental data showing that knowledge about individual familiar instances can influence information processing speed and accuracy in certain experimental paradigms (see Whittlesea, 1987). Category prototype theories have difficulty with such findings, because they assume that information about individual items is subsumed in the prototype representation and is not retained in memory – thus processing speed and accuracy should not be influenced by similarity to the particular training instances. Prototype theories and instance-trace theories represent a trade-off between explanatory power on the one hand and representational capacity on the other: though instance-trace theories may explain a broader range of data, category prototype theories offer a more economical means of storing information in memory.

Both kinds of similarity-based models encounter computational difficulties that neither has yet resolved. In both cases, the attributes of an object are usually assumed to contribute independently to the similarity metric that determines their match to representations in memory. However, the weight that should be given to a particular attribute can depend upon the category of the object, or on the other attributes that constitute the stimulus description. For example, color is an important attribute when considering whether a particular fruit is an orange or a grapefruit, but it is irrelevant when considering whether a particular car is a Ford or a Chrysler. Whether or not color should be considered when categorizing a particular object, then, depends upon whether the object is a car or a fruit – but the category cannot be determined until the similarity between the probe and the representation has been computed.

This difficulty can be overcome in a model with hierarchical processing structure. For example, suppose that a particular object can be classified as a kind of fruit on the basis of its shape and texture. In a hierarchical model, knowledge about how other properties (such as color) should be weighted can be stored with the *fruit* representation, and used to constrain more specific category assignments (e.g. *orange* or *grapefruit*). However, the apparent simplicity of this idea is belied by the difficulty of implementing such a theory in a computational model.

CONNECTIONIST NETWORKS

The third computational approach offers a quite different means of understanding semantic memory storage and retrieval processes. *Connectionist models* propose that semantic knowledge is stored in the configuration of weights in a neural network which performs mappings between perceptual representations of words and objects, and internal semantic representations. In the connectionist paradigm, representations take the form of patterns of activity across a set of simple, neuron-like processing units. Different pools of units encode different kinds of representations. For example, the phonological structure of a spoken word might be encoded by a pattern of activity across one pool of units; the visual appearance of an object might be coded by a pattern of activity across another; and an object's semantic representation might be coded by patterns of activity across a third pool. Just as neurons in the brain communicate with one another by sending chemical signals across synapses, the units in connectionist networks communicate their activation states by means of weighted connections with other units in the system. Information processing proceeds by the successive updating of unit states in response to inputs. The way that downstream units respond to a given input depends upon the particular configuration of interconnecting weights: when the weights are set so that the system can produce correct outputs in response to probes, the system can be said to 'know' the domain. (*See* **Connectionism**)

Hinton's Distributed Semantic Network

One of the seminal connectionist models of semantic memory was proposed by Hinton (1981). Hinton's model, shown in Figure 2, consists of several banks of simple neuron-like processing units, grouped into layers and connected as indicated in the illustration. The architecture of the model reflects the structure of a simple proposition, with one bank of units (labeled *Object*) representing the

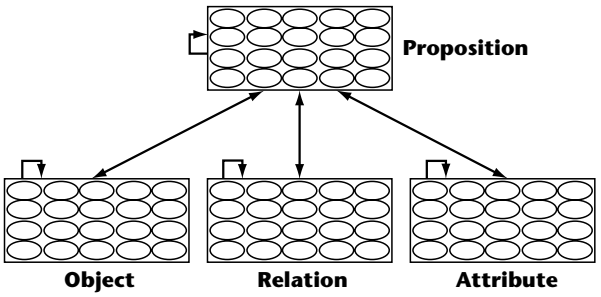


Figure 2. Hinton's (1981) PDP model of semantic memory. Each term in a simple proposition (e.g. 'Clyde isa elephant') is represented by a pattern of activity across the corresponding pool of units. Units in different pools interact with one another by means of weighted connections with units in the pool labeled *Proposition*. Individual propositions can be stored by adjusting the weights within and between pools, so that the patterns of activity corresponding to the terms of the proposition are stable. Redrawn with alterations from Hinton (1981). Permission pending.

first term, one bank (labeled *Relation*) representing the proposition, and one bank (labelled *Attribute*) representing the second term. Different fillers for these three slots are coded by different patterns of activity across the corresponding units.

All three banks of units send projections to (and receive projections from) a fourth layer, labeled *Proposition* in the figure. The sign and magnitude of each weight determines how the activity of the sending unit will influence the state of the receiving unit. When an input is presented to the network, each unit updates its state by taking a weighted sum across all units from which it receives projections (with the contribution of each sending unit weighted by the strength of the intervening connection). Thus when a proposition is given to the network as input, activity in the *Object*, *Relation*, and *Attribute* layers gives rise to a pattern of activation across the *Proposition* units. These units in turn send new signals to the *Object*, *Relation*, and *Attribute* units, which update their states accordingly. The process iterates until the unit states stop changing, at which point the network is said to have *settled* into a steady state. Hinton demonstrated that individual propositions could be stored in the network, by adjusting the weights to make the patterns representing the proposition stable. Each stored proposition would then be represented in the network by a unique pattern of activity across the *Proposition* units, which simultaneously activated and received support from the input patterns.

Like the spreading-activation model, the network was capable of completing stored propositions when given an appropriate cue as input. For example, if the model was taught the proposition *Clyde isa elephant*, and was then given the incomplete input *Clyde isa*, it would settle into a steady state in which the pattern representing the correct completion of the proposition (*elephant*) was observed across the *Attribute* units. This retrieval mechanism also provided an account of knowledge generalization quite different from that embodied in the Collins and Quillian model. If related objects (such as various individual elephants) were represented by overlapping patterns of activity across the *Object* units, they would contribute similar inputs to the *Proposition* units. Thus, the entire network would tend to settle into an appropriate steady state (corresponding to the most similar stored proposition) when given a novel input that overlapped with familiar, stored patterns. For example, if the network had stored the proposition *Clyde is gray*, and was then given the inputs *Elmer is* in the *Object* and *Relation* units, it would settle to a state in which the pattern corresponding to *gray* was observed across *Attribute* units – provided that the representations of *Clyde* and *Elmer* were sufficiently similar.

Extensions of Hinton's Approach

Where the architecture of Hinton's model reflected the structure of simple propositions, subsequent approaches have employed architectures that more closely capture the functional organisation of the cortex. One common architecture is shown in Figure 3. Visual representations of objects are represented by patterns of activity across *Visual* units; the perceptual structure of spoken words or statements is represented by activity in the *Verbal* layer; and semantic representations are encoded in the activity of the *Semantic* units. Words with similar sounds (such as 'cat' and 'bat') give rise to similar patterns of activity across *Verbal* units, whereas objects with similar appearances (e.g. a bat and a bird) give rise to similar patterns of activity across *Visual* units. Units in the *Semantic* layer perform the mappings between phonological and visual layers, just as did the *Proposition* layer in Hinton's model. When semantically related objects are represented with similar patterns of activity in this layer, the architecture provides a mechanism for the generalization of stored knowledge to novel items on the basis of semantic relatedness.

In this framework, storage of semantic knowledge involves the gradual adjustment of the weights that connect units within and between pools in the system, so that activation of a given

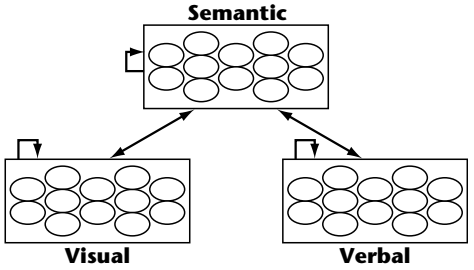


Figure 3. A common connectionist architecture for semantic memory which is more consistent with the gross functional anatomy of cortex. Verbal units code the perceptual structure of speech sounds; Visual units code the visual structure of a perceived object; and Semantic units code the semantic similarity relations existing among objects in the world.

Visual or *Verbal* representation will give rise to the appropriate pattern of activity across other pools. Several powerful learning algorithms have been proposed for finding weights that will allow models like the one in Figure 3 to perform the correct mappings.

Acquisition of Semantic Representations

Whereas the models described above assign prespecified semantic representations to objects, other approaches have explored how such representations might be acquired through learning in a connectionist network. For example, Rumelhart and Todd (1993) demonstrated that the propositional content contained in Collins and Quillian's hierarchical model could also be coded in the distributed representations acquired by a connectionist network trained with backpropagation. Their model is shown in Figure 4: entities in the environment are represented by individual input units in the layer labeled *Item*; different propositional relations are represented by individual units in the layer labeled *Relation*; and the various completions of different propositions are represented by units in the layer labeled *Attribute*. In order to answer queries about various items, the network must find a configuration of weights that will produce the correct states across output units when a single *Item* and *Relation* are provided as inputs. However in this case, the mediating representations are unspecified. Instead, weights throughout the model are initially set to small, random values, and the network is trained with backpropagation. (*See* **Backpropagation**).

The activation of a single input unit generates a distributed pattern of activity across *Representation* units; and when the model's weights are small and random, every input produces a similar pattern of activity throughout the network. Because semantically related items share many attributes in the output, they give rise to similar error signals early in learning, and produce similar weight changes throughout the network. Consequently the network continues to generate similar patterns of activity across the *Representation* units for related items as it learns, but gradually differentiates its representations of unrelated items. Ultimately, the internal representations acquired by the network capture the semantic similarity relations that exist among the objects in its environment; and this learned similarity structure can provide the basis for knowledge generalization.

Neuropsychological Models

A further advantage of the connectionist framework is that, because the learning and information processing principles it adopts are grossly consistent with those known to operate in the brain, models developed in this tradition offer a means of understanding how the functional anatomy of cortex supports semantic cognition. Consequently, connectionist models of semantic memory have probably had their strongest impact in the domain of neuropsychology.

To understand neuropsychological impairment in a connectionist framework, theorists begin with a working model of the phenomenon in question, and investigate how the model's behavior changes when the knowledge stored in its weights is disrupted by simulated lesions. For example, several studies have shown that semantic knowledge about living and nonliving things can be doubly-dissociated as a consequence of neural trauma. Warrington and Shallice (1984) suggested that this pattern might arise as a consequence of the sensory-motor organisation of the cortex, assuming first that knowledge about the sensory and functional properties of objects are stored in separate areas of cortex that may be damaged independently; and second, that functional information is more important for the representation of nonliving things, whereas the reverse is true living things. Under these conditions, damage to areas of cortex subserving functional versus perceptual knowledge would differentially affect knowledge of nonliving and living domains.

Farah and McClelland (1991) proved the feasibility of this hypothesis in a computational model

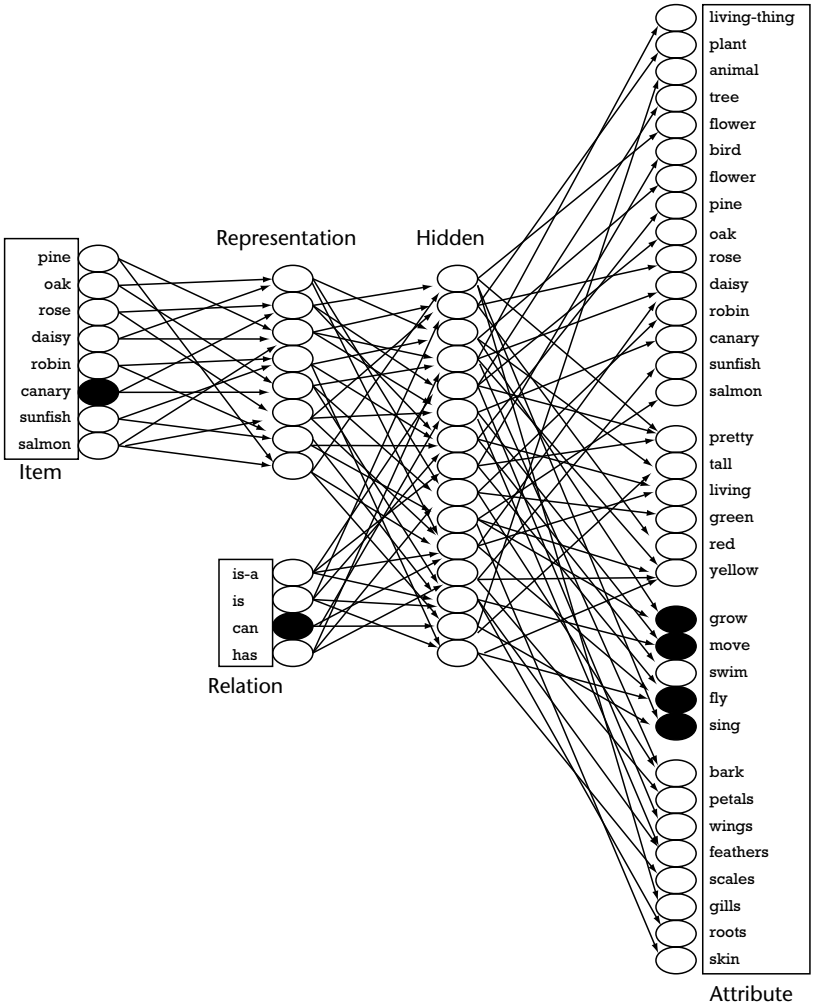


Figure 4. A feed-forward model of semantic memory described by Rumelhart and Todd (1993). As it learns to complete various propositions, the network constructs distributed internal representations for the objects in its environment that capture their semantic relations. Based on the network depicted in Rumelhart and Todd (1993).

with an architecture similar to that shown in Figure 3. They created *Verbal, Visual*, and *Semantic* representations of objects by generating patterns of activity across each of these pools. However, they also assumed that some of the units in the *Semantic* layer would be responsible for encoding primarily functional information (*functional-semantic* features) whereas others would encode primarily perceptual information (*perceptual-semantic* features). To capture the intuition that living things rely to a greater extent on perceptual information than do nonliving things, they created semantic representations of living things that had many perceptual-semantic features, and representations of nonliving things that had relatively few. The model was then trained to associate visual, verbal, and semantic representations, using an error-correcting learning algorithm.

Once the model had learned to perform the correct mappings, the authors investigated its behavior under simulated lesions to either the functional-semantic units or the perceptual-semantic units. They found that, when perceptual-semantic units were lesioned, the model was frequently unable to produce the correct name in response to a visual input – but that this deficit was much more apparent for living things than nonliving things. When functional-semantic units were lesioned, the reverse was true: the model had a more difficult time retrieving the correct name for nonliving things.

The Farah-McClelland model challenged the conclusion often drawn from double-dissociations in the case literature, that the pattern of data reflects the existence of independent subsystems that are specialized to each dissociated function. In this

case, the model showed that knowledge of living and nonliving things might be doubly dissociated, without there being independent systems for representing each domain. It also demonstrated that the sensory-functional hypothesis had some appealing and counter-intuitive implications. For example, although the model showed worse performance for living things relative to nonliving things when its perceptual-semantic units were damaged, its performance with nonliving things was not perfect. The same is also true of patients: so-called category-specific deficits are almost invariably accompanied by milder impairment of knowledge for objects in the relatively spared domain.

FUTURE DIRECTIONS

The models reviewed in this entry mark a progression from formal logical approaches to semantic cognition, involving the storage of discrete concepts and propositions, toward somewhat more graded and distributed schemes in which the similarity structure of internal representations guides semantic knowledge generalization. Connectionist models provide a means of linking intuitions about similarity-based representation to principles of information processing known to operate in the brain, and to the functional neuroanatomy of cortex at a gross scale. Such models offer great promise for a detailed and mechanistic account of semantic knowledge about concrete objects. However, the scope of semantic knowledge extends far beyond concrete object knowledge. Very little computational work has been done to understand knowledge about the meanings of verbs and abstract words, relations between objects, events, actions, or complex scenes. One appeal of propositional approaches to conceptual knowledge is that they may extend well to these domains: under such theories, knowledge about abstract words and verbs may be stored as sets of concepts and propositions, just as is knowledge about concrete objects. A challenge for similarity-based and connectionist models in future will be to account for such knowledge using the same principles of knowledge acquisition and processing that explain concrete object knowledge.

References

Collins AM and Quillian MR (1969) Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour* **8**: 240–248.

Collins AM and Loftus EK (1975) A spreading-activation theory of semantic processing. *Psychological Review* **82**: 407–428.

Farah MJ and McClelland JL (1991) A computational model of semantic memory impairment: modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* **120**(4): 339–357.

Hinton GE (1981) Implementing semantic networks in parallel hardware. In: Hinton GE and Anderson JA (eds) *Parallel models of associative memory*, pp. 161–187. Hillsdale, NJ: Erlbaum.

McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**(3): 419–437.

Nosofsky RM (1988) Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**: 700–708.

Rosch E and Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.

Rumelhart DE and Todd PM (1993) Learning and connectionist representations. In: Myer DE and Kornblum S (eds) *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Warrington E and Shallice T (1984) Category specific semantic impairments. *Brain* **107**: 829–853.

Whittlesea BWA (1987) Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**: 3–17.

Further Reading

Anderson JR (1991) The adaptive nature of human categorization. *Psychological Review* **98**(3): 409–426.

Hinton GE and Shallice T (1991) Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review* **98**(1): 191–243.

Plaut DC and Shallice T (1993) Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**(5): 377–500.

Rumelhart DE, McClelland JL and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Smith EE and Medin DL (1981) *Categories and Concepts*. Cambridge, MA: MIT Press.

Sentence Processing

Introductory article

Michael K Tanenhaus, University of Rochester, Rochester, New York, USA

CONTENTS

Sentence processing is the study of the representations people form as they understand a sentence or utterance and the mechanisms underlying the component processes. These include recognizing the words in a sentence, determining the syntactic and semantic relationships among these words, and interpreting the sentence with respect to the relevant linguistic and non-linguistic context.

INTRODUCTION

The goal of research in sentence processing is to understand the representations people form as they understand a sentence and the processes involved in developing these representations. Sentence processing involves recognizing the words in a sentence, determining the syntactic and semantic relationships among these words, and interpreting the sentence with respect to the relevant linguistic and nonlinguistic context. These processes draw upon specifically linguistic knowledge as well as knowledge about the world and how people use language for purposes of communicating their ideas, goals, and intentions. The interdisciplinary community that addresses issues in sentence processing includes linguists, computer scientists, and psychologists, and it draws upon theoretical ideas and methodological tools from each of these disciplines.

During comprehension, readers and listeners construct a mental model or discourse model. In order to build the model, the comprehender must recognize each of the words in the sentence and determine the syntactic (and semantic) relationship among them, as defined by the grammar of the language. The process of assigning syntactic relationships is generally referred to as 'parsing', though this term is sometimes used to include both the assignment of syntactic relationships and interpretation. As a result of parsing the sentence, the comprehender can determine the 'message' expressed by the sentence, i.e. who did what to whom.

As the propositional content of the sentence is extracted, new events and entities are introduced into the model and reference is made to those that have already been introduced. Referring expressions known as 'anaphors' and other contextually dependent expressions play a primary role in this process. In addition, readers and listeners may routinely incorporate some inferences that are triggered by information in the sentence, the context of the sentence or utterance, and general world knowledge.

As an example, consider the sentence, *The student spotted by the proctor was expelled*. Syntactically, *student* is the head noun in the definite noun phrase *the student spotted by the proctor*, which is the subject noun phrase for the verb phrase *was expelled*. *The student* is modified by a relative clause (*spotted by the proctor*). Both the main clause and the embedded clause are in the passive voice and both are in the past tense.

Semantically, the sentence describes two events: a spotting event and an expelling event, each of which takes place in the past. The proctor and the student are discourse entities who each participate in the spotting event in different ways or modes, commonly referred to as 'thematic roles'. The proctor is the agent of the spotting event – the one doing the spotting – and the student is the theme or patient – what is being spotted. The student is also the theme or patient of the expelling event. Thematic roles are closely linked to the syntactic relationships among verbs (and other relational words) and their syntactic complements or arguments.

The syntactic and semantic relationships in a sentence are in large part determined by

constraints defined by the structure of the language. However, the way that they are recovered in sentence processing is strongly influenced by the sequential nature of the input. The input is sequential in spoken language because auditory stimuli are composed of transient acoustic events. The input is sequential in written language because readers typically fixate successively on each word in a sentence, processing only a limited amount of information from the periphery.

IMMEDIACY OF COMPREHENSION

Many aspects of sentence processing are remarkably time-locked to the unfolding linguistic input. Access to the semantic representations of spoken words begins as soon as phonetic input is encountered. By the middle of the first vowel in *student*, potential lexical candidates, such as student, stool, and stoop are becoming active in memory, along with their associated syntactic and semantic representations. In the absence of strongly constraining context, convergence on the most likely lexical candidate occurs shortly after enough phonetic input is received to distinguish the input from other likely alternatives – often well before the word ends. Constraints based on prior syntactic, semantic, and discourse constraints further increase the speed with which a word is identified, suggesting that semantic integration takes place in parallel with lexical access. Syntactic processing occurs equally quickly. When a sentence becomes syntactically anomalous, processing effects are observed immediately after the word where the anomaly occurs. Moreover, when a sentence that is briefly ambiguous between two or more syntactic alternatives is disambiguated in favor of its less preferred alternative, processing consequences are observed as soon as the disambiguating information is encountered. Similar effects are observed for semantic and pragmatic anomalies.

Methodological Consequences

The speed with which sentences are comprehended has led researchers to focus on 'on-line' experimental measures that are closely time-locked to the input. The most widely used methodologies include:

1. monitoring tasks in which a participant subject is timed while monitoring the incoming linguistic input for various linguistic units, e.g. a target phoneme, word, member of a category or a rhyme;
2. probing and priming tasks in which the developing representation is probed by examining response times to targets that occurred in the sentence or are related to a word or a phrase in the sentence;
3. participant-controlled or self-paced reading in which reading times are measured while subjects read text in presentation 'windows' of various sizes;
4. monitoring event-related potentials (ERPs) as subjects read or listen to sentences; and
5. monitoring eye fixations as participants read text or listen to spoken utterances in the context of visual displays containing potential referents.

AMBIGUITY

Nearly all sentences contain numerous temporary ambiguities as they unfold over time. While ambiguity exists at all levels of language comprehension, from categorizing a phoneme to recognizing a speaker's intentions, researchers in sentence processing have been primarily concerned with how readers and listeners resolve the structural ambiguity that arises in assigning grammatical relationships to words and phrases as a sentence unfolds.

Structural ambiguity arises because of several characteristics of natural language. Linguistic forms such as words and morphemes are frequently ambiguous with respect to their syntactic category, and tense and voice (e.g. *spotted* could be either an adjective or a verb, and as a verb it could be a past tense form of a passive participle). Syntactic structure is hierarchical and, at least partially, recursive, resulting in frequent dependencies between non-adjacent words and phrases. Returning to the example, *The student spotted by the proctor was expelled*, the noun phrase *the student spotted by the proctor* contains a sentence embedded within a noun phrase. Consequently, *the student* is the subject of the verb *expelled* even though several words intervene, including another noun phrase, *the proctor*. Similarly the verb *was* agrees in number with the noun *student*, and not the most local noun *proctor*.

Ambiguous forms and non-adjacent grammatical dependencies often combine to result in fragments that are temporarily ambiguous among multiple syntactic structures. For example, *the student spotted …* is ambiguous between a past tense main clause in the active voice (e.g. *The student spotted the proctor*) and a relative clause (e.g. *the student spotted by the proctor …*).

The alternative syntactic structures for a temporarily ambiguous fragment typically have different semantic and discourse consequences. As a result, information at these higher levels could provide constraints that are relevant to resolving ambiguities at lower levels. For example, in a main clause, *the student* would be assigned the agent thematic

role (i.e. the agent of the spotting event, whereas it would be assigned the theme role in a relative clause. Therefore, the thematic fit of the noun phrase to each of these potential roles could be useful in resolving the syntactic ambiguity. *Student* is both a good agent and a good theme for a spotting event; however, another noun such as *crib sheet* is a good theme but a poor agent. Thus, thematic fit might be used to help disambiguate a fragment such as *the crib sheet spotted …* in favor of a relative clause at the verb.

A past tense main clause and a (restrictive) relative clause also differ with respect to how they would be integrated into the discourse model. A main clause is used to introduce a new event into the model, whereas a restrictive relative clause can be used to refer to an already established event. Thus the felicity with which each type of event could be incorporated into the model in specific discourse contexts is another potential source of constraint.

Beginning with the classic work by Bever in 1970, sentences with local syntactic ambiguities have served as a primary empirical base for developing and testing models of syntactic processing.

Processing modularity

The question of how and when the processing system exploits correlated constraints has played a central role in research on sentence processing. Two broad classes of approaches to these questions have been explored, each of which explaining the speed and efficiency of sentence processing differently. One approach, which has its roots in Marslen-Wilson's early work, assumes that the processing system is able to rapidly and optimally integrate different types of information by taking advantage of correlated constraints to evaluate the evidence for multiple alternatives. In these constraint-based or interactive approaches, use of correlated constraints is what underlies the speed and efficiency of on-line comprehension. As a consequence, processing at different levels is closely intertwined. For example, lexical and syntactic processing are interleaved with the interpretation and construction of a discourse model. Constraint-based approaches are increasingly drawing upon constructs drawn from neural network models.

A second approach posits informationally encapsulated subsystems or modules, each of which is responsible for processing distinct types of representations. In modular models, the speed and efficiency of on-line comprehension arises, in part, because certain processes are insulated from others. Thus correlated constraints are ignored in initial processing. Fast mandatory processes are often viewed as restricted to recovering the syntactic and perhaps semantic relationships defined by the grammatical structure of the sentence. These processes are insulated from higher-level interpretative and inferential processes that relate a sentence to its context.

These contrasting perspectives lie at the center of the modularity debate that played a prominent theoretical role in sentence processing research, especially research on lexical and syntactic processing during the 1980s and early 1990s. The modularity issue is closely intertwined with questions about the nature of linguistic representation and how grammatical knowledge is accessed and used in processing. Together these issues have played a central role in sentence comprehension research, especially research examining syntactic ambiguity resolution.

PSYCHOLOGICAL REALITY OF LINGUISTIC STRUCTURE

Some of the most influential early research in sentence processing was aimed at establishing the plausibility of controversial ideas drawn from theoretical linguistics. These ideas included the importance of abstract syntactic structure. Nearly all current research in sentence processing incorporates constructs drawn from linguistic theory. However, a subset of research is primarily focused on using experimental data to evaluate competing linguistic hypotheses. This approach typically contrasts predictions generated by different grammatical frameworks. In recent years, work in this tradition has focused on structures in which there are syntactic and semantic dependencies between non-adjacent constituents – structures that some, but not all, syntactic frameworks analyze by postulating phonologically unrealized (empty) categories that result when a constituent has been moved at some level of syntactic representation.

MODELS OF SENTENCE PROCESSING AND AMBIGUITY RESOLUTION

When a sentence containing a temporary ambiguity is resolved in favor of the less preferred alternative, people often experience a feeling of having been led down the 'garden path'. Preferences for particular interpretations of locally ambiguous

sentences are systematic; there is a strong tendency for sentences with similar structures to exhibit similar preferences.

Although the presence of systematic preferences for temporarily ambiguous sentences is well documented, models of sentence processing differ in how they account for these preferences. Models of ambiguity resolution can be divided into classes along two interrelated dimensions. First, models differ in whether they assume that a single syntactic alternative is initially considered (serial models) or whether multiple alternatives are evaluated in parallel. Second, models differ in what information is used when – in the case of serial models to determine the initial analysis, in the case of parallel models to determine the relative viability of the alternatives.

At one end of the continuum are models in which a restricted domain of information, typically syntactic constraints or a subset of syntactic constraints, plays a privileged role in initially structuring the input or ranking the alternatives. For example, in the garden-path model, an encapsulated syntactic processor initially structures the linguistic input, making a provisional commitment to a single structure using decision principles based primarily on structural complexity. Other encapsulated subsystems or modules are assumed to be responsible for other aspects of sentence processing, including lexical access, reference resolution, and assignment of thematic roles. Information from these modules does not inform initial syntactic decisions but is used to evaluate and, if necessary, revise initial syntactic commitments.

At the other end of the continuum are constraint-based models in which rich lexical representations make available multiple syntactic alternatives that are weighted by the frequency of lexical forms and their argument structures in specific syntactic environments. The alternatives are continuously evaluated, using relevant linguistic and nonlinguistic constraints such as the semantic/thematic fit between a phrase and a potential argument position, and the effects of information from the discourse context. A central claim of these models is that the complex patterns of structural preferences and interactions with discourse and local semantic context arise from simple, domain-independent integration mechanisms, without appeal to syntactic complexity as an explanatory primitive. Other models fall somewhere in between these two classes in the degree to which they rely on structural complexity, parallel analysis, and use of multiple constraints.

COMPUTATIONAL MODELS

In contrast to many other areas of cognitive science, including some domains within psycholinguistics, implemented computational models have played a surprisingly limited role in guiding theoretical and empirical research in sentence processing. For example, there are no computationally explicit realizations of the influential class of models that combine explicit assumptions about linguistic structure with decision principles designed to predict the processing environments that will result in explicit syntactic misanalysis and the nature of the recovery process when an initial analysis is incorrect. One reason is that the complexity of the component processes in sentence processing does not lend itself well to developing models that make close contact with empirical data without making numerous ancillary assumptions.

Gibson has developed an explicit model that accounts for word-by-word processing difficulty across a wide range of sentence types using two factors: (1) the number of syntactic predictions that are resolved at a word or phrase and (2) the number of intervening discourse entities over which the predictions must be maintained. These two factors increase difficulty because they require processing resources.

Most constraint-based models assume that probabilistic constraints across multiple levels of representation are continuously combined to weight alternative lexical and structural alternatives according to Bayesian principles. Models incorporating constraint-based ideas are increasingly realized within connectionist architectures. In hybrid models symbolic representations are computed using competition-based connectionist algorithms. For example, Stevenson has developed such a parser that accounts for many behaviorally observed patterns of processing complexity, while Vosse and Kempen adopt a related approach using competition with a lexicalized grammer.

One promising class of learning-based models combines lexicalized grammars with connectionist-based learning principles to predict preferences for locally ambiguous sentences. Another approach uses simple recurrent networks (SRNs) in which the task of the network is to predict the next upcoming word. The input to the network is a sequence of sentences presented one word at a time that is generated from a probabilistic-finite state grammar that mimics the frequency of occurrence of particular words and structures, as determined by prior corpus analyses. SRN models have simulated the relative difficulty of different types of

embedded sentences, including effects traditionally attributed to individual differences in working memory capacity as well as complex patterns of interactions between lexical, structural, and contingent frequency effects.

SOURCE OF CONSTRAINTS

There is an emerging consensus that at least four classes of constraints influence ambiguity resolution and the difficulty of processing unambiguous sentences in reading and listening:

- structural complexity that draws upon limited computational resources
- lexical constraints
- frequency-based constraints
- discourse and pragmatic context.

Computational Resources

The idea that limited capacity working memory resources place important constraints on sentence processing has played a central role in most models of sentence processing since the seminal work of Miller and Chomsky. Most generally, the assumption of limited capacity resources motivates a link between structural complexity and processing complexity. Although the precise nature of these capacity limits has been relatively unexplored, there is an ongoing debate about whether resource capacity limits reflect general or language-specific resource limits.

Miller and Chomsky hypothesized that sentences with multiple center embeddings such as *The student the proctor spotted cheated* are difficult because the *the student* has to be held in working memory while the embedded constituent *the proctor spotted* is processed. *The student* must then simultaneously be assigned the grammatical roles of object of *spotted* and subject of *cheated*. Working memory constraints motivated the need for recoding at clause boundaries in the clausal model proposed by Fodor *et al*. and they provide the rationale for why parsing is serial in the garden-path model. Memory constraints also motivate the two influential parsing principles that guide its syntactic decisions: minimal attachment (choose the syntactically simplest analysis) and late closure (adopt the analysis that incorporates a word or phrase into the most recent constituent). Memory and capacity constraints provide the motivation for differences in processing difficulty in Gibson's and Lewis's models. Similar patterns emerge from simple SRNs trained on representative corpora. These models have capacity limits, and a recency bias, but they do not have a dedicated working memory system.

Lexical Constraints

An ongoing issue in sentence processing is how and when syntactic processing makes use of constraints defined over grammatical categories (e.g. noun verb, preposition) or draws upon constraints that are tied to specific lexical items, such as the difference between the verb *put*, which requires three arguments (e.g. *John put the car in the garage*), an agent, a theme and a location, and the verb *died*, which allows only a single argument (e.g. *John died*). Syntactic and semantic aspects of verb-argument structure are accessed as soon as a verb is recognized and verb-specific constraints are rapidly used in resolving local ambiguities and in projecting upcoming structure, including potential arguments. However, the extent to which lexically based representations are used in parallel with and/or supercede structural constraints defined over syntactic categories (e.g. minimal attachment) remains controversial. The importance of lexically specific constraints can be illustrated by comparing the following two sentences, each of which has the same syntactic structure and sequence of syntactic categories:

The raft floated down the river sank. (1)

The salmon released in the stream spawned. (2)

Most readers find the first sentence confusing because it is difficult to identify *the raft floated down the river* as a relative clause (compare *the raft that floated down the river sank*). However, the second sentence is much easier to understand.

Frequency

The difference in processing difficulty between the example sentences in (1) and (2) is largely due to the frequencies with which the component are used in the syntactic structures in these sentences. Sentences are easier to process when words and structures are used in familiar ways. There is a strong correlation between frequency of use as a transitive verb and the frequency with which a verb is used as a past participle in a relative clause. *Floated* is frequently used intransitively, whereas *released* is nearly always used transitively.

Frequency effects are diagnostic of exposure-based influences, i.e. learning effects. As a consequence, they play an important role in motivating

constraint-based approaches, especially those approaches in which structural constraints are emergent properties of learning-based systems. Other approaches treat frequencies as biases that affect the probabilities with which different structures are chosen. An important issue for such approaches is how to define and constrain the 'grain' at which frequencies are stored by the processing system.

Discourse and Pragmatic Context

Research in language processing has been divided into two very different traditions. The dominant tradition in sentence processing, the 'language-as-product tradition', emphasizes the individual cognitive processes by which speakers create, and listeners recover, linguistic representations – the 'products' of language production and comprehension.

Within the product tradition, studies of context have focused primarily on how discourse referents affect the difficulty of sentence processing, typically focusing on temporally ambiguous sentences. Definite noun phrases presuppose a uniquely identifiable referent (e.g. *the raft*). Many garden-path sentences are temporarily ambiguous between a structure in which a new phrase is introducing an argument and a structure in which that phrase is modifying an existing argument. The typical preference for an argument interpretation is reduced or eliminated when the context provides several possible referents for the definite noun phrase, and the new phrase refers to salient context that disambiguates the referent.

The second tradition, the 'language-as-action' tradition, emphasizes interactive conversation as the most basic form of language use. A central tenet in work within this tradition is that utterances can only be understood within a particular context, which includes the time, the place, and the participant's conversation goals, as well as the collaborative processes that are intrinsic to conversation.

The advent of lightweight head-mounted eye-tracking methodologies has enabled researchers to investigate moment-by-moment comprehension processes in contexts with real-world referents, using tasks in which the language is grounded in well-defined behavioral goals for the speaker and the listener.

When listeners listen to spoken utterances in task-relevant visual contexts, saccades to potentially relevant objects are closely time-locked to the referring expressions in the utterance. Listeners dynamically update mental representations combining the unfolding linguistic input with task-relevant properties of real-world referents that affect the earliest moments of reference resolution and syntactic ambiguity resolution. Whether lexical access, parsing, and reference resolution also require the listener to take into account the common ground between the speaker and the listener is more controversial.

CURRENT TRENDS AND FUTURE DIRECTIONS

For most of the last two decades, research in sentence processing has been (productively) focused on accounting for a somewhat limited set of phenomena involving temporarily ambiguous sentences, with English as the primary target language, and reading as the preferred mode. Important work on these topics continues. However, the field is rapidly expanding to include a wider range of phenomena, including cross-linguistic investigations and investigations of sentence processing in preliterate children, and its development. As investigations of spoken language increase, there is more focus on the role of prosody in comprehension. Work at the interface between syntax, semantics, and pragmatics is increasingly seeking to bridge the language-as-product and language-as-action traditions, often making reference to ideas developed in studies of dialogue systems. Finally, researchers are beginning to examine how people process the types of natural, often disfluent, utterances that occur in spontaneous face-to-face conversational settings.

Further Reading

Altmann GTM and Steedman M (1988) Interaction with context during human sentence processing. *Cognition*, **30**: 191–238.

Bever TG (1970) The cognitive basis for linguistic structures. In Hayes JR (ed.) *Cognition and the Development of Language*, pp. 279–362. New York, NY: Wiley.

Christiansen MH and Chater N (2001) Connectionist psycholinguistics: capturing the empirical data. *Trends in Cognitive Science* **5**: 82–89.

Clark HH (1992) *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Clark HH (1996) *Using Language*. New York, NY: Cambridge University Press.

Fodor JA (1983) *Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor JA, Bever TG and Garrett MF (1974) *The Psychology of Language*. New York, NY: McGraw Hill.

Fodor JD (1995) Comprehending sentence structure. In: Gleitman LR and Liberman M (eds) *An Invitation to*

Cognitive Science, *Volume 1*: *Language*, 2nd edn, pp. 209–246. Cambridge, MA: MIT Press.
Fodor JD and Ferriera F (1998) *Reanalysis in Sentence Processing*. Dordrecht, The Netherlands: Kluwer.
Frazier L (1987) Sentence processing: a tutorial review. In: Coltheart M (ed.) *Attention and Performance XII: The Psychology of Reading*. London, UK: Erlbaum.
Frazier L and Clifton C (1996) *Construal*. Cambridge, MA: MIT Press.
Garnham A (2001) *Mental Models and the Interpretation of Anaphora*. Philadelphia, PA: Psychology Press/Taylor and Francis.
Gibson T (1998) Linguistic complexity: locality of syntactic dependence. *Cognition* **68**: 1–76.
Gibson T and Pearlmutter N (1998) Constraints on sentence comprehension. *Trends in Cognitive Science* **2**: 262–268.
Gorrell P (1995) *Syntax and Parsing*. Cambridge, UK: Cambridge University Press.
Johnson-Laird PN (1983) *Mental Models: Towards a Cognitive Science of Language*. Cambridge, MA: Harvard University Press.
Jurafsky D (1996) A probabilistic model of lexical and syntactic disambiguation. *Cognitive Science* **20**: 137–194.
Just M and Carpenter P (1992) A capacity theory of comprehension: individual differences and working memory. *Psychological Review* **99**: 122–149.
Lewis RL (2000) Specifying complete architectures for language processing: process, control, and memory in parsing and interpretation. In: Crocker MW, Pickering M and Clifton C (eds) *Architectures and Mechanisms for Language Processing*. Cambridge, UK: Cambridge University Press.
MacDonald MC, Pearlmutter N and Seidenberg MS (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703.
Marslen-Wilson W (1975) Sentence perception as an interactive parallel process. *Science* **189**: 226–228.
Miller GA and Chomsky N (1963) Finitary models of language users. In: Luce RD, Bush RR and Galanter E (eds) *Handbook of Mathematical Psychology*. New York, NY: Wiley.
Mitchell DC (1994) Sentence processing. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 375–409. New York, NY: Academic Press.
Pickering ML, Clifton C and Crocker MW (2000) Architectures and mechanisms in sentence comprehension. In: Crocker MW, Pickering M and Clifton C (eds) *Architectures and Mechanisms for Language Processing*. Cambridge, UK: Cambridge University Press.
Rayner K (1998) Eye movements in reading and information processing; 20 years of research. *Psychological Bulletin* **124**: 372–422.
Seidenberg MS and MacDonald MC (1999) A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* **23**: 569–588.
Stevenson S (1994) Competition and recency in a hybrid model of syntactic disambiguation. *Journal of Psycholinguistic Research* **23**: 295–322.
Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM and Sedivy JE (l995) Integration of visual and linguistic information in spoken language comprehension. *Science* **268**: 632–634.
Tanenhaus MK and Trueswell JC (1995) Sentence comprehension. In: Miller J and Eimas P (eds) *Handbook of Cognition and Perception*. San Diego, CA: Academic Press.
Townsend DJ and Bever TG (2001) *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.
Trueswell JC, Sekerina L, Hill NM and Logrip ML (1999) The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition* **73**: 89–134.
Vosse T and Kempren G (2001) Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* **75**: 105–143.

Sex Differences in Cognition

Intermediate article

Doreen Kimura, Simon Fraser University, Burnaby, British Columbia, Canada

CONTENTS

Men and women are known to differ reliably on a number of cognitive and motor abilities. Some biological influences on such abilities are reviewed.

INTRODUCTION

Men and women have, for several decades now, consistently shown average differences in their cognitive strengths. For example, men excel on some types of spatial abilities, and on mathematical reasoning. Women excel on memory for verbal material (words), and on a function called 'perceptual speed', where rapid identity comparisons must be made, as well as on some measures of verbal fluency. In the motor sphere, men are superior in targeting ability as manifested in throwing accuracy, and women in fine motor skills emphasizing finger dexterity.

Until fairly recently these sex differences were usually attributed to the differing environments that men and women experienced. The claim that the differences were not seen until after puberty was interpreted as supporting that position, though it is equally consistent with a hormonal influence. Since the 1980s, however, research on sex differences has intensified, and it has become clear that exposure to sex hormones early in life has an important influence on adult cognitive patterns. Moreover, some of these differences are present as early as ages three and four (and possibly earlier, could they be appropriately tested), and some of them appear in widely differing cultures. The former overriding salience of socialization as an explanation has thus been significantly diminished.

In fact, most current researchers on sex differences in behavior view them in terms of the long-standing division of labor between males and females during our evolutionary history. Men were responsible for both short- and long-distance hunting and scavenging, and for manufacture and use of weapons. Women foraged near home and cared for the home and for infant children. According to this schema, men were selected for spatial navigational ability and targeting skills, and women for perceptual abilities sensitive to small changes in children or in the home base. Though such theories are difficult to test, they provide a useful interim heuristic for understanding how sex differences might have come about.

WHAT ARE THE DIFFERENCES?

The review below is not exhaustive of all the sex differences in cognition that have been reported. Moreover, although this article outlines several tests on which either men or women are superior, it must be kept in mind that there is a great deal of overlap between the sexes on most tests. Abilities on which there is less overlap – that is, where the differences are quite large – are mental rotation, throwing accuracy, and verbal memory, described below.

Male-favoring Abilities

Spatial abilities

The kinds of spatial tasks on which men excel include both simple and complex abilities. One relatively simple task requires only that the slope of a line be correctly matched to its counterpart in an array of lines. More complex tasks include the ability to make a correction for a change in the orientation of an object, referred to as 'mental rotation' (Figure 1). A somewhat different kind of spatial ability favoring males requires that one imagine what a depicted object will look like when manipulated – for example, a flat surface when folded and then unfolded in various ways, a function often called 'spatial visualization'. Performance on a mental rotation test has been found to

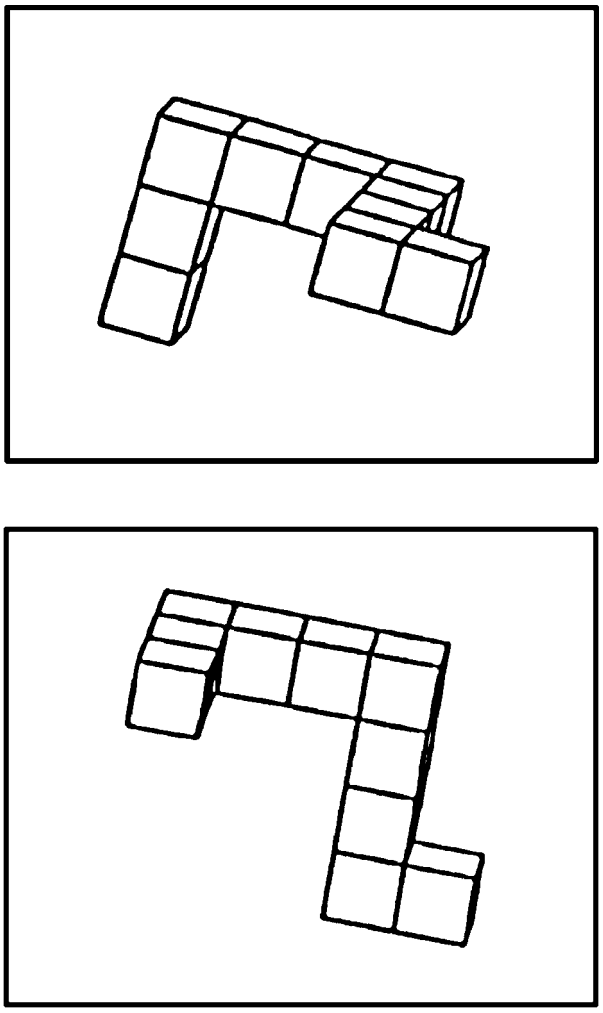


Figure 1. An example of a difficult 'mental rotation' task. The subject must decide whether the two figures are different, or are simply the same figure, rotated.

be closely related to that on a computerized labyrinth, suggesting that mental rotation ability contributes substantially to navigation (Moffat *et al*., 1998).

Mathematical reasoning

Ever since the invention of mathematical aptitude tests, boys have outperformed girls. In contrast, on school marks, which reflect learned solutions or achievement, girls do at least as well as boys in maths. It appears that boys are better able to use their maths experience to solve new problems. Although some have claimed that such sex differences are disappearing, the sex difference on the Scholastic Aptitude Test – Mathematics remained fairly steady from 1964 to 1981, with an effect size (a measure of the difference in standard deviation scores) of approximately 0.40 (Donlon, 1984).

The superiority of males becomes more marked as the tests become more demanding. So even in adolescents chosen for maths talent, the scores of males on the SAT–Math is higher than that of females, and at the highest end of the scores, males outnumber females by more than 10 to one (Benbow, 1988). On an even more demanding maths test, the Putnam competition in North America, less than 5 percent of the top scorers in a recent year were women. These findings may help explain why fewer women choose maths-intensive fields of study, such as physics, engineering, and maths itself (Lubinski and Benbow, 1992), whereas their entry into biological sciences approaches that of men.

Targeting accuracy

Men are much more accurate on average in hitting a target with a missile than women are. This difference appears early in life, before there is a wide divergence of sports activities, and in any case has been shown not to be accounted for by sports history (Watson and Kimura, 1991). Men are also more accurate in intercepting a launched missile, even if they merely have to touch it in flight. It appears that they are better able to process the spatiomotor requirements of such tasks.

Female-favoring Abilities

Verbal memory and verbal fluency

Women are better at recalling words, whether in a spoken list or a meaningful paragraph. This difference is present in children and lasts into old age. It is possible that verbal memory contributes to another task on which women outperform men – object location memory (Figure 2) (Eals and Silverman, 1994). Although men may be better at recalling locations *per se*, women appear to excel in recall of locations of specific objects when presented in an array. To the extent that the labeling of objects assists in recall of their locations, better verbal memory might provide an advantage on the task.

Women are also somewhat better at verbal fluency, by which is meant an ability to generate words with some constraint on the letters they contain, such as generating words beginning with a specified letter. It is well established that females are also faster at naming a series of patches of common colors, though it is still unclear whether this rapid naming advantage is restricted to colors or holds for forms as well.

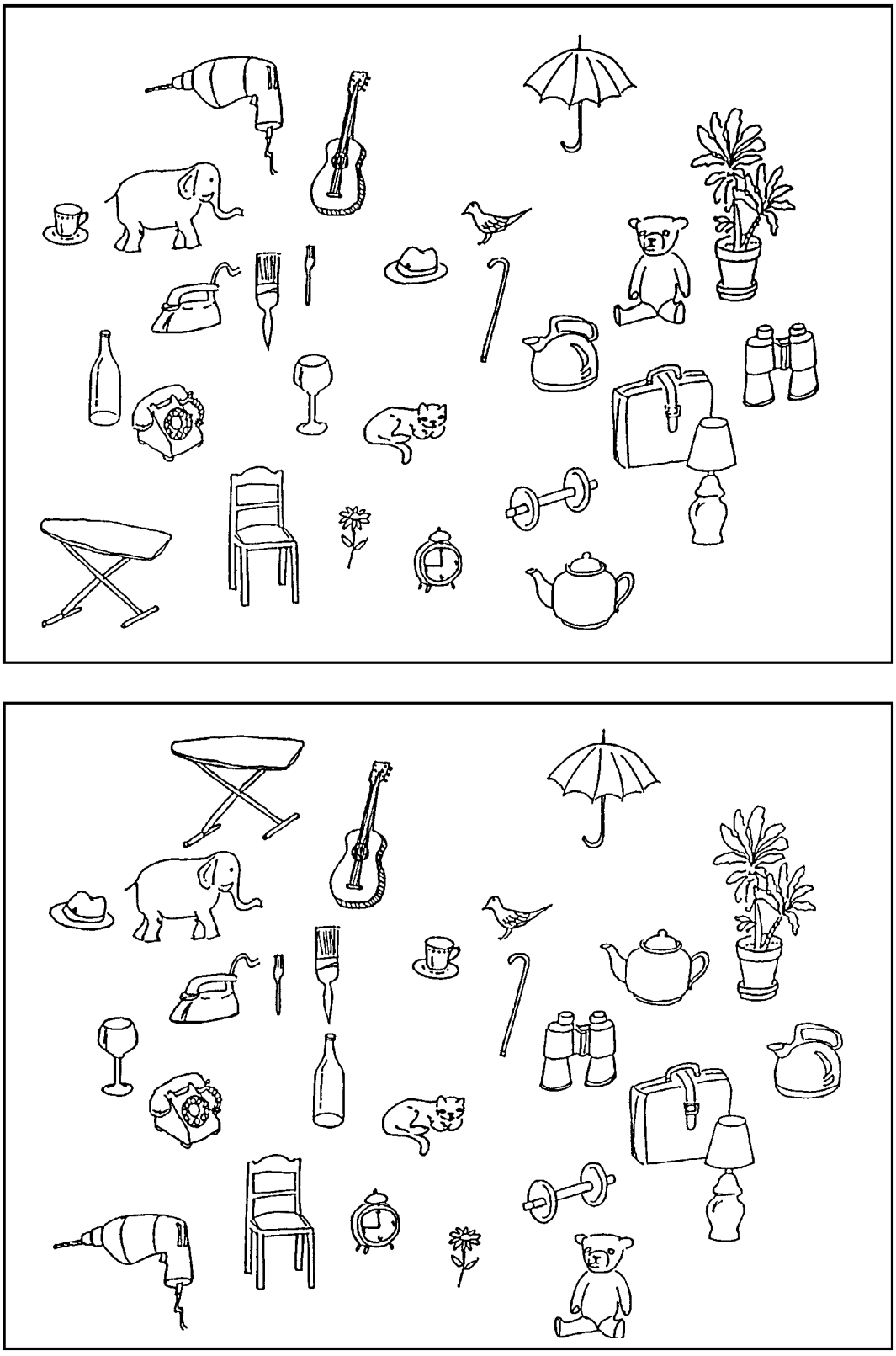


Figure 2. An object location memory task. The subject must decide which of the articles in the top array have shifted position in the bottom array. Reproduced by permission of the authors (Silverman and Eals, 1992).

Perceptual abilities

Women are faster at matching identical forms. For example, in a series of items like that in Figure 3, women complete more items in a specified time period. They also show better fusion of the images to the two eyes to yield the perception of depth, at least within personal body space. This enhanced depth perception may be related to the kind of fine motor skills women also enjoy.

Motor skills

Women excel at certain motor tasks requiring precise finger movements. For example, they can place pegs into a series of holes in a board more quickly than men. It has been suggested that this is due to their smaller fingers, but that does not appear to account for all instances of their greater fine motor skill.

HORMONAL MECHANISMS

General Effects

The effects of sex hormones on behavior are usually classified as *organizational* or *activational* (Goy and McEwen, 1980), concepts derived primarily from animal research. 'Organizational' refers to a demonstrated effect of early exposure to hormones (prenatal or immediate postnatal) on adult behavior, and has been seen chiefly in the influence of androgens in masculinizing the brain and behavior. 'Activational' refers to an effect of current hormone levels. The two are not entirely independent. The influence of hormones is seen not only in reproductive behavior but so far has appeared in almost all behaviors that differentiate males and females.

Cognitive Effects

Early exposure to androgens

In rats it has been shown that the male 'strategy' of using geometric cues to solve mazes, in contrast to the female strategy of using landmark cues, is due to the early exposure of males to androgen derivatives (Williams *et al*., 1990). Several studies suggest that the greater ability of human males to solve spatial problems of the kind shown in Figure 1 is due to a similar mechanism. The most convincing of such studies has shown that girls with congenital adrenal hyperplasia (CAH), who experience hyperexposure to adrenal androgens prenatally but not after birth due to corrective therapy, are superior to unaffected girls on such spatial tasks (Berenbaum *et al*., 1995; Hampson *et al*., 1998).

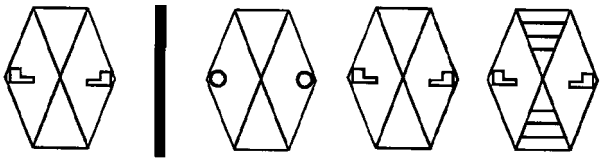


Figure 3. An example of a perceptual matching test. The subject must decide which of the three figures on the right is identical to the one on the left. A large series of such problems is presented, with a time limit.

Relation to current hormone levels

The current adult level of testosterone, the chief androgen, has been shown to relate systematically to scores on spatial tests similar to those enhanced in CAH girls. In young men with lower normal levels, spatial scores are higher than in young men with higher testosterone levels. In women, just the reverse occurs. This finding has been interpreted to mean that there is an optimal level of testosterone for spatial ability, and it is in the low normal male range. Below or above this, the spatial ability is poorer.

Cognitive patterns also vary within individuals with changing hormone levels. In men, testosterone levels change across the year and throughout the day, and spatial ability changes concomitantly. In women, estrogen, considered a female hormone, varies across the menstrual cycle, and women's cognitive pattern fluctuates with it. When estrogen levels are high, women are relatively better on tasks at which females typically excel, such as verbal fluency and fine motor skill; whereas on male-favoring spatial tasks, scores are relatively worse (Hampson, 1990). When estrogen levels are low, the reverse pattern appears.

These, as well as studies on hormone treatment in older men, and in the course of sex-change hormone therapy, combine to suggest that, although our cognitive strengths and weaknesses are strongly influenced by hormones early in life, they continue to be affected by current hormone levels in adulthood.

BRAIN SEX DIFFERENCES

Since the brain mediates all behavior, it must follow that consistent group differences in cognitive pattern, of the kind we see between men and women, will be reflected in brain differences of some kind. However, the search for brain correlates of cognitive sex differences has so far met with limited success. Nevertheless, we can confidently expect advances in this field in the coming years. This section will review two main features of brain or-

ganization that have been investigated as bases for such differences: functional brain asymmetry, and interhemispheric connections.

Brain Asymmetry

It is well established that the left and right cerebral hemispheres of the human brain have different functions, with the left hemisphere more concerned with communicative and complex motor ability, and the right with perceptual and spatial ability. This fact has been determined initially and most convincingly by studying the effects of damage to the left or right hemisphere, and later by perceptual techniques that can sample the function of one or other hemisphere, as well as by brain imaging techniques.

It has frequently been postulated that such functional brain asymmetry is more marked in men than in women, and that this is one basis for their differing cognitive patterns. When one examines the brain-damage evidence for sex differences in brain asymmetry, however, it seems that the left hemisphere is no less critical for basic speech functions in women than in men. That is, speech disturbances (aphasias) are no more frequent after right-hemisphere pathology in women, as would be predicted by the bilaterality hypothesis. However, the pattern of organization for speech *within* the left hemisphere is different, with speech more anterior-based in women, and more posterior-based in men (Kimura, 1999). In contrast to these more basic speech abilities, verbal functions such as vocabulary or defining words, and verbal fluency, do appear to be more bilaterally organized in women.

It has similarly been claimed that the right hemisphere develops earlier and is more specialized for spatial functions in men than in women. One would then expect that right-hemisphere pathology would be more disruptive of functions such as spatial rotation in men than in women, compared to left-hemisphere pathology, but this has not proved to be the case. Women's spatial ability is at least as affected by right-hemisphere strokes as is men's (Kimura, 1999).

Interhemispheric Connections

The largest commissural connection between the hemispheres is the corpus callosum, with a lesser role played by the anterior commissure, among others. The anterior commissure has been shown to have a larger cross-sectional area in women than in men, and the corpus callosum has been variously claimed to be larger in women in its posterior portion, called the splenium. A greater number of commissural fibers might be expected to result in better connections between the hemispheres, but so far no cognitive feature has been related to sex differences in the commissural area.

CONCLUSION

The differing cognitive patterns of men and women are known to be influenced by both early (pre- and possibly perinatal) and current levels of sex hormones. While such differences must obviously be based in the nervous system, the precise brain mechanisms are still to be determined.

References

Benbow CP (1988) Sex differences in mathematical reasoning ability in intellectually talented preadolescents: their nature, effects, and possible causes. *Behavioral & Brain Sciences* **11**: 169–182.

Berenbaum SA, Korman K and Leveroni C (1995) Early hormones and sex differences in cognitive abilities. *Learning and Individual Differences* **7**: 303–321.

Donlon TF (1984) Predictive validity of the ATP tests. In: *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*, pp. 141–170. New York, NY: College Examination Board.

Eals M and Silverman I (1994) The hunter-gatherer theory of spatial sex differences: proximate factors mediating the female advantage in recall of object arrays. *Ethology and Sociobiology* **15**: 95–105.

Goy RW and McEwen BS (1980) *Sexual Differentiation of the Brain*. Cambridge, MA: MIT Press.

Hampson E (1990) Estrogen-related variations in human spatial and articulatory-motor skills. *Psychoneuroendocrinology* **15**: 97–111.

Hampson E, Rovet JF and Altmann D (1998) Spatial reasoning in children with congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Developmental Neuropsychology* **14**: 299–320.

Kimura D (1999) *Sex and Cognition*. Cambridge, MA: MIT Press.

Lubinski D and Benbow CP (1992) Gender differences in abilities and preferences among the gifted: implications for the math–science pipeline. *Current Directions in Psychological Science* **1**: 61–66.

Moffat SD, Hampson E and Hatzipantelis M (1998) Navigation in a 'virtual' maze: sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior* **19**: 73–87.

Watson NV and Kimura D (1991) Nontrivial sex differences in throwing and intercepting: relation to psychometrically-defined spatial functions. *Personality & Individual Differences* **12**: 375–385.

Williams CL, Barnett AM and Meck WH (1990) Organizational effects of early gonadal secretions on sexual differentiation in spatial memory. *Behavioral Neuroscience* **104**: 84–97.

Further Reading

Ankney CD (1992) Sex differences in relative brain size: the mismeasure of woman, too? *Intelligence* **16**: 329–336.

Breedlove SM (1994) Sexual differentiation of the human nervous system. *Annual Review of Psychology* **45**: 389–418.

Halpern DF (1997) Sex differences in intelligence. *American Psychologist* **52**: 1091–1102.

Kimura D (1999) Sex differences in the brain. *Scientific American*, Summer Quarterly, **10**: 26–31.

Lynn R (1994) Sex differences in intelligence and brain size: a paradox resolved. *Personality & Individual Differences* **17**: 257–271.

Pool R (1994) *Eve's Rib*. New York, NY: Crown Publishers.

Silverman I and Eals M (1992) Sex differences in spatial abilities: evolutionary theory and data. In Barkow JH, Cosmides L and Tooby J (eds) *The Adapted Mind*, pp. 533–549. New York, NY: Oxford University Press.

Sexual Arousal

Intermediate article

John Bancroft, Kinsey Institute, Indiana University, Bloomington, Indiana, USA

CONTENTS

Sexual arousal is described as an emotional and motivational state involving complex interaction between information processing of sexual stimuli, central arousal, genital response, and behavior. The state motivates the individual to increase sexual stimulation and pleasure.

INTRODUCTION

The term 'sexual arousal', though used in a variety of ways, generally refers to a state motivated towards the experience of sexual pleasure and possibly orgasm, and characterized by genital response and other manifestations of general arousal. It shares some features with emotional states, and others with more clearly motivated states such as hunger. Although this emotional/motivational system is biologically organized to serve reproduction, in the human and a number of other primates nonreproductive exploitation of sexual pleasure is widespread, though susceptible to a variety of social constraints.

As with other emotional/motivational states, it is not yet possible to do more than infer the central state or 'drive' underlying sexual arousal. In such circumstances it is helpful to work back from the assumed behavioral objective. In the male, the fundamental reproductive goal of sexual arousal is coitus with male ejaculation into the female's reproductive tract. The motivational state driving this behavior is not dependent on any physiological deprivation, as with motivational states such as hunger or thirst, but relies in some way on the anticipation of and desire for sexual pleasure. In the male, an important aspect is that orgasm and ejaculation, once it occurs, is followed by an active inhibition of the capacity for further sexual arousal, producing a transient phase of refractoriness. This may show some parallels with satiation following excessive food intake, but there may also be important differences.

The situation in the female is less straightforward. In rodents, for example, coitus requires the female to express a reflexive lordosis response which involves active inhibition of other motor responses. It is far from clear to what extent this response pattern results from a motivation for sexual pleasure. In humans, the female may be more similar to the male, but orgasm in the female does not have the fundamental reproductive significance of male orgasm and ejaculation, and the capacity for female orgasm probably exists because there is no evolutionary need for it to be suppressed during development (Beach, 1976). Possibly because of the lesser reproductive significance of female sexual response, and a greater tendency for it to be inhibited in the female than in the male, the sexuality of women has been much more subject to social repression and constraint, particularly during certain historical periods. Whereas orgasm may be a motivational goal in women, it is clearly less crucial as an organizing factor in women's sexuality than it is in men.

Much of the confusion and inconsistency in the use of the term 'sexual arousal' stems from issues in operationalizing the definition for research purposes; and the particular and varied challenges involved have depended on the species being studied (Sachs, 2000). The underlying mechanisms are highly complex and our scientific observations are restricted to certain 'windows' into this complexity. The 'cognitive' window is both more accessible and relevant to the human. Invasive manipulations allow windows into brain mechanisms in animals, which are not available to human research. The types of physiological monitoring of sexual arousal used for human and animal research again involve very different 'windows'. This article focuses principally on the human, though its reference to underlying brain mechanisms depends to a large extent on evidence from animal studies.

We can thus view sexual arousal through four principal 'windows': (1) information processing of sexual stimuli, (2) central arousal or activation, (3) the elicitation of genital response such as penile erection in the male or clitoral tumescence and vaginal lubrication in the female, and (4) behavior, which is under voluntary control, aimed at intensifying sexual stimulation and likely, at least in the male, to result in orgasm. We can reasonably infer that these four aspects are both integrated and interactive with each other.

INFORMATION PROCESSING

The appraisal of a stimulus as sexual is usually the first step in initiating the interactive process leading to sexual arousal. Learning plays a major part in determining what stimuli are identified as sexual, though to some extent the 'sexual significance' of a stimulus may be innately determined and independent of previous experience. Processing of sexual stimuli operates at two levels: (1) 'automatic', fast, unconscious processing requiring little attention, and (2) 'controlled' (or attentional) cognitive processing (Janssen *et al*., 2000). 'Automatic' processing is regarded as fundamental in the appraisal of emotional events in general, leading to appropriately rapid 'response sets'. The 'controlled' processing allows an additional level of appraisal which may either augment or reduce the initial automatic response, and which may be activated and focused by the emotional arousal accompanying it.

Sexual stimuli are not only external; internal imagery can initiate the process. Automatic processing of spontaneous genital response or tactile stimulation is a further source. It is noteworthy that such 'positive feedback' of genital response is more predictable in men than in women.

AROUSAL

The brain responds to novel stimuli with an increase in alertness and attention. Such activation is a key component of emotional states, affecting both cortical activity and peripheral mechanisms such as blood pressure and heart rate. In general there is a pattern of generalized arousal, including peripheral autonomic responsiveness needed for motor response, which is relatively nonspecific, occurring in a variety of emotional states including fear, anger, and sexual arousal. What is more specific and also less well understood is the focusing of attention on to stimuli relevant to the specific type of arousal, and in the case of sexual arousal, the activation of genital response.

GENITAL RESPONSE

This is the most specifically sexual component of 'sexual arousal', which is essential for the completion of the reproductive sequence. In the male, penile erection is crucial, and controlled by a balance between excitatory and inhibitory signals originating from higher in the central nervous system (Steers, 2000). Specialized mechanisms in the penis involve arterial vasodilation, increasing blood flow into the erectile tissues, relaxation of smooth muscle in the sinusoidal spaces, allowing distension by inflowing blood, and venous constriction which prevents escape of blood from the erectile spaces. Vasodilation and smooth muscle relaxation depend mainly on cholinergic and nitrergic (nitric oxide) mediation, and are counterbalanced by the constricting effects of norepinephric mediation (Rehman and Melman, 2001).

Genital responses in the female are less well understood. Transudation of fluid through the vaginal wall and tumescence of the vulva and vaginal introitus serve to facilitate entry of the erect penis during coitus. Tumescence of the clitoral bodies is in many respects homologous to penile erection and may be an important source of sexual pleasure for some women. Apart from its motivating potential, clitoral response has no obvious reproductive function.

MOTOR RESPONSE

In animals, sexually relevant motor responses are typically divided into 'appetitive' (e.g. gaining access to receptive partner) and 'consummatory' (e.g. mounting and intromission) responses. In humans, while motor responses are a central part of sexual activity they do not fall easily into these two categories, though clearly certain types of human sexual activity are 'consummatory' (e.g. masturbation or coitus). In general, when experiencing sexual arousal there will be a motivation to 'do something about it' which can be seen as reflecting the motor component.

INTEGRATING THE SYSTEM

To what extent can we infer central mechanisms that integrate the various phenomena observed through our 'scientific windows'? The three monoaminergic systems originating in the brain stem are obviously crucial, although in each case their role is

relevant to aroused and motivated states in general, not just sexual arousal (Role and Kelly, 1991).

The norepinephric (NE) system is central to the arousal component. This has two parts which originate in two brainstem nuclei. One, the locus ceruleus (l.c.), has widespread ascending projections to the basal forebrain, hypothalamus, and cortex among other parts of the brain, and descending projections to sensory nuclei in the brainstem and to the spinal cord. The other, the lateral tegmental nucleus (l.t.n.), has even more diffusely projecting neurons which also include brainstem, spinal cord, and cortex. However, whereas the l.c. is activated by novel sensory input, the l.t.n. is involved in integration of autonomic function in brainstem and spinal cord nuclei. We can therefore reasonably assume that together they are a key part of the interface between emotionally relevant external stimuli and a generalized arousal response.

The dopaminergic (DA) system is more discretely organized into several subsystems, three of which are relevant to sexual arousal, one of them specifically and the other two nonspecifically. The nigro-striatal tract, degeneration of which causes Parkinson disease, is involved in the organization of motor behavior, particularly initiation of motor responses and 'readiness to respond'. This includes copulatory but also many other integrated motor patterns. The meso-limbic tract promotes 'appetite' for a variety of appetitive behaviors including sexual. More specifically, the dopaminergic input to the medial pre-optic area (MPOA) from the A14 periventricular system is involved in the orchestration of genital responses and associated motor patterns such as mounting or thrusting.

The serotonergic system, the most extensive of the three monoaminergic systems, has wide-ranging and predominantly inhibitory effects.

There is as yet no clearly identified center or system for producing genital response; the closest is the nucleus paragigantocellularis in the brain stem which is involved in the inhibition of erection (McKenna, 2000).

What is particularly unclear is how these various nonspecific systems in the brain become recruited to serve the specific state of sexual arousal. Testosterone (T) may have a role in this respect. T receptors are found in the NE system and may be involved in the sexual focusing of central arousal. T receptors are also found in the MPOA and in other sites and are likely to be involved in sexual focusing of the DA system. But at present we can only speculate on such interactions. Other neurotransmitters and neuromodulators, such as excitatory amino acids and neuropeptides, are also involved at some level in the development and integration of sexual arousal. This is well illustrated in the case of post-ejaculatory refractoriness and 'sexual satiation', a state induced in male rodents by allowing repeated access to receptive females in a short period of time. After seven or eight such copulations a state of sexual exhaustion (but not general exhaustion) becomes evident which may remain for several days. This state has been shown to be dependent on the integrity of the NE arousal system but also involves peptidergic, serotonergic, and dopaminergic mechanisms (Rodriguez-Manzo and Fernandez-Guasti, 1995).

Given the difficulties in explaining specific response patterns in this complex system, it is useful to think in terms of 'conceptual systems', which infer organizing mechanisms of functional relevance and which lend themselves to empirical testing.

One such conceptual system is the 'dual control' of sexual response, which postulates that there are both excitatory and inhibitory systems operating to determine whether sexual arousal and response occurs in any specific situation (Bancroft, 1999). Appraising stimuli as sexual and nonthreatening activates the excitatory system; appraisal of threat associated with a sexual stimulus activates the inhibitory system, reducing the likelihood that a sexual response will occur. Such a balanced mechanism can be seen as adaptive; while sexual behavior is biologically important and inherently rewarding it is also potentially dangerous, exposing the sexually aroused individual to a variety of hazards. Similarly, when faced with a nonsexual threat it is important to inhibit other response patterns, such as sexual or eating, which might distract from the avoidance of threat. Such a system may explain how central arousal may be channeled to sexual arousal in some situations and to fearful arousal in others. It is also inherently likely that individuals will vary in their propensity for both excitation and inhibition. Thus a person with low propensity for inhibition may become sexually aroused in the presence of a sexual stimulus even when threat is involved. Conversely, an individual with high propensity for inhibition may be unable to respond in relatively unthreatening situations, leading to sexual dysfunction.

This conceptual system also offers an explanation for the process described as 'excitation transfer', where arousal induced by one type of stimulus becomes recruited to activate the 'arousal response' to another type of stimulus (Zillmann, 1983). It has been shown experimentally that an individual experiencing anger may have an augmented response

to a subsequent sexual stimulus; or conversely, someone in a sexually aroused state may respond with greater anger to a subsequent provocation. It is unlikely that such patterns are found in everyone and this may be an important type of individual variability. The 'dual control' model postulates that such transfer of arousal, from say a frightening or threatening situation, may augment the response to a coexisting sexual stimulus in those individuals who have a low propensity for inhibition of sexual response. For the average or high-propensity individual, appraisal of threat would effectively ensure through inhibitor mechanisms that such excitation transfer did not take place, and in most situations that would be a more adaptive response pattern.

CONCLUSION

Sexual arousal involves the orchestration and integration of complex processes in the brain reactive to the presence of sexual stimuli and resulting in an emotional state which motivates the aroused individual towards obtaining increased sexual stimulation. The underlying mechanisms are only partially understood and individual differences in responsiveness of different aspects of this system are likely to contribute to the considerable variability of human sexual expression.

References

Bancroft J (1999) Central inhibition of sexual response in the male: a theoretical perspective. *Neuroscience and Biobehavioral Reviews* **23**: 763–784.

Beach FA (1976) Cross-species comparisons and the human heritage. *Archives of Sexual Behavior* **5**: 469–485.

Janssen E, Everaerd W, Spiering M and Janssen J (2000) Automatic processes and the appraisal of sexual stimuli: towards an information processing model of sexual arousal. *Journal of Sex Research* **37**: 8–23.

McKenna KE (2000) Some proposals for the organization of the central nervous system control of penile erection. *Neuroscience and Biobehavioral Reviews* **24**: 535–540.

Rehman J and Melman A (2001) Normal anatomy and physiology. In: Mulcahy JJ (ed.) *Male Sexual Function: A Guide to Clinical Management*, pp. 1–46. Totowa, NJ: Humana.

Rodriguez-Manzo G and Fernandez-Guasti A (1995) Participation of the central noradrenergic system in the reestablishment of copulatory behavior of sexually exhausted rats by yohimbine, naloxone, and 8-OH-DPAT. *Brain Research Bulletin* **38**: 399–404.

Role LW and Kelly JP (1991) The brain stem: cranial nerve nuclei and the monoaminergic systems. In: Kandel ER Schwartz JH and Jessell TM (eds) *Principles of Neural Science*, 3rd edn, pp. 683–699. Norwalk, CT: Appleton & Lange.

Sachs BD (2000) Contextual approaches to the physiology and classification of erectile function, erectile dysfunction, and sexual arousal. *Neuroscience and Biobehavioral Reviews* **24**: 541–560.

Steers WD (2000) Neural pathways and central sites involved in penile erection: neuroanatomy and clinical implications. *Neuroscience and Biobehavioral Reviews* **24**: 507–516.

Zillmann D (1983) Transfer of excitation in emotional behavior. In: Cacioppo JT and Petty RE (eds) *Social Psychophysiology*, pp. 215–240. New York, NY: Guilford Press.

Further Reading

Andrew RJ (1974) Arousal and the causation of behavior. *Behaviour* **51**: 135–165.

Bancroft J (1989) *Human Sexuality and Its Problems*. Edinburgh, UK: Churchill Livingstone.

Janssen E and Everaerd W (1993) Determinants of male sexual arousal. *Annual Review of Sex Research* **4**: 211–245.

Pfaff DW (1999) *Drive: Neurobiological and Molecular Mechanisms of Sexual Motivation*. Cambridge, MA: MIT Press.

Pfaus JG (1999) Revisiting the concept of sexual motivation. *Annual Review of Sex Research* **10**: 120–156.

Rosen RC and Beck JG (1988) *Patterns of Sexual Arousal*. New York, NY: Guilford Press.

Zillmann D (1984) *Connections between Sex and Aggression*. Hillsdale, NJ: Lawrence Erlbaum.

Signal Detection Theory

Introductory article

Justin A MacDonald, Purdue University, West Lafayette, Indiana, USA
JD Balakrishnan, Purdue University, West Lafayette, Indiana, USA

CONTENTS

Signal detection theory is one of the first attempts by psychologists to model the processes involved in elementary perceptual recognition tasks. The theory's main feature is a rigorous distinction between performance limits due to the quality of information provided by the senses (sensitivity), and limits arising from decision-making strategies (bias).

INTRODUCTION

Imagine the following situation: a pathologist has been given a tissue sample from a male child to check for the presence of cancer. If she incorrectly concludes that the sample is cancerous, the child will unnecessarily undergo a battery of costly and potentially risky interventions. On the other hand, if she incorrectly concludes that there is no cancer present, the disease may be allowed to develop beyond the ability of medical science to treat it. Because this disease is relatively rare in children, the pathologist knows that the probability that the cells are cancerous is small. Before looking at the tissue, therefore, she might be fairly confident that it is noncancerous. Unfortunately, the signs of cancer in a tissue sample are not clear-cut. Every sample is unique; sometimes there are obvious, strong indications that cancer is present, at other times the signs are moderate or weak. How does the pathologist combine the complex visual information from examining the tissue sample with her prior knowledge about the likelihood of a cancer being present? How do we evaluate the performance of these expert decision-makers to determine whether they are making the best possible use of their knowledge about the probability of the disease in this situation and the visual perceptual information they glean from examining the sample?

SIGNAL DETECTION THEORY

The pathologist's dilemma is one of many examples of what psychologists call the 'signal detection problem'. Are the cells in the tissue sample cancerous (a signal) or is it a false alarm (noise)? There are two possible events (signal and noise) and two possible decisions (detect and reject).

During the Second World War, engineers working on radar detection systems developed a model to describe the processes involved in this particular type of decision-making problem. They used this model to study the behavior of radar operators who were responsible for distinguishing random blips on the radar screen from blips caused by an actual ship or aircraft (the signal detection problem). This mathematical model was soon noticed by psychologists and developed into a set of more detailed models and procedures, which have come to be referred to collectively as 'signal detection theory' (SDT).

According to this general theory, there are two distinct stages in the detection process, assimilation of information from the environment (encoding) and selection of an appropriate response (decision-making). Detection errors occur either because the output of the encoding process is misleading (due to noise in the outside world or in the brain) or because the detector's decision-making strategy is flawed.

Although the basic principles of SDT are general enough to encompass more complex decision-making tasks, most applications have focused on the original problem, in which there are only two possible stimuli and two possible responses. In such situations (often called discrimination tasks), the four possible outcomes of a given trial during the experiment are 'hits' (correctly detect a signal),

'misses' (fail to detect a signal), 'correct rejections' (reject a non-signal event) and 'false alarms' (incorrectly identify a non-signal event as a signal event).

Since there are only two possible decisions, detect or reject, the proportion of hits on signal trials plus the proportion of misses on signal trials must equal 1. Similarly, the proportion of false alarms on noise trials plus the proportion of correct rejections on noise trials must equal 1. The performance of a subject can therefore be reduced to two values, one for each stimulus condition, and it is customary to report just the hit and false alarm rates.

SENSITIVITY AND BIAS

One of the basic assumptions of SDT is that all of the sensory information that could conceivably be used to come to a decision on a given trial can be represented as a single point on a one-dimensional continuum called the 'decision axis'. Points near the left extreme of the axis correspond to sensory information that constitutes compelling evidence that only background noise was present; and points near the right end are compelling evidence for a signal. Somewhere between these two extremes lies the point at which the sensory information is entirely inconclusive (the two stimuli or conditions are equally likely). In the pathologist's situation, the axis might be the number of what appear to her to be questionable cells in the tissue sample: many questionable cells might constitute very strong evidence that the sample is cancerous while very few questionable cells might indicate a noncancerous sample. The decision-maker must combine this sensory information with other information, such as the history of the patient.

Because of the random variability that is almost sure to exist either in the environment or in the brain (or both), the sensory effect of a stimulus should be expected to vary from trial to trial even if the stimulus stays constant (in the pathology example, different cancerous tissue samples would look different). However, some effects should be more common than others when they come from a single type of source. For technical reasons, the most popular version of SDT assumes that these relative frequencies for a given stimulus follow a standard probability distribution, the 'normal' or 'bell-shaped' curve, as illustrated in Figure 1.

Notice that each stimulus has its own distribution, because each stimulus should have its own set of effects that it will produce often (effects near the center, or mean, of the distribution) and its own set of effects that are possible but rare (effects in the tails of the distribution). The degree to which the distributions for two different stimuli overlap is a measure of the 'sensitivity' of the senses to the physical differences between the two stimuli. In general, sensitivity should decrease with increasing physical similarity of the stimuli being discriminated.

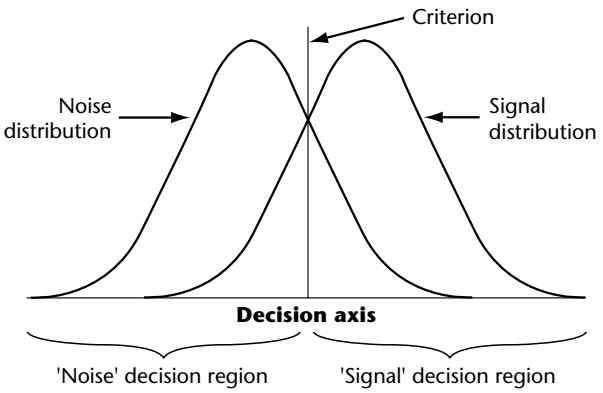


Figure 1. The distributions of effects on signal and noise trials in SDT. Notice that the two distributions have the same variance but different means. According to SDT, the observer changes his or her response strategy by shifting the criterion along the decision axis, thereby changing the boundary between the 'signal' and 'noise' response regions. In this example, the criterion is placed at the point of intersection between the two distributions and is therefore unbiased. Response biases affect the placement of the criterion (shifting it to the left or right), but not the distributions themselves.

Decision-making Strategies

The fact that the two distributions in Figure 1 overlap is a fundamental problem for the decision-maker: for any given sensory effect, the decision-maker cannot be sure which stimulus was presented. From the SDT point of view, this is the main reason why subjects make errors in discrimination tasks. On the other hand, the sensory effect does provide some information: some effects are more likely to have been caused by the signal stimulus and others are more likely to have been caused by the noise stimulus. The decision-maker's task is therefore to find a 'decision rule' that maps each effect to its most probable cause. SDT assumes that, to accomplish this, a 'criterion' is set somewhere on the decision axis.

Intuitively, the best place to put this criterion might seem to be the point at which the two distributions intersect; i.e. where the relative frequency of the sensory effect is identical on signal and noise trials. This is valid if the signal and noise trials

occur with equal frequency during the experiment (i.e., the 'base rates' of the stimuli are equal). However, if the base rates are unequal, the optimal placement (maximizing the percentage of correct responses) will be somewhere to the left or right of the point of intersection of the two distributions, depending on which stimulus is presented more often.

To see why the criterion should shift in accordance with the base rates, consider the case in which the signal stimulus will never be presented (when its base rate is zero): the optimal decision-maker would always respond 'noise', no matter what sensory effect is present. In terms of the model, the criterion in this case would be shifted to the extreme right of the decision axis so that all of the sensory effects are mapped to the noise response.

For each pair of base rates, there is one and only one position on the decision axis that would lead to optimal performance: all of the others are suboptimal to some degree. Any criterion placed at any point other than the point of intersection (which may or may not be suboptimal, depending on the base rates) is called a 'biased' decision rule.

CALCULATING SENSITIVITY

For any set of base rates, the SDT model illustrated in Figure 1 can be fitted to the data (i.e. to the hit and false alarm rates) to estimate the two parameters of the model, which are the distance d' between the means of the two distributions (in standard deviation units) and the position X_C of the criterion. Assuming a standard deviation of 1, the formulas for these two parameters are:

$$d' = \Phi^{-1}(p_{\text{hit}}) - \Phi^{-1}(p_{\text{fa}}) \quad (1)$$

$$X_C = -\Phi^{-1}(p_{\text{fa}}) \quad (2)$$

where p_{hit} and p_{fa} are the hit and false alarm rates, respectively, and Φ^{-1} is the inverse cumulative distribution function for the standard normal distribution. The values of these inverse functions can be obtained from tables or from statistical software packages.

ROC CURVES

In principle, there is no reason why the two distributions of effects should always be normal, or should differ only in their mean values. In fact, empirical studies often suggest that the signal distribution is more variable than the noise distribution. Partly for these reasons, detection theorists often estimate sensitivity using a different approach that incorporates the same basic principles of the detection model but does not assume anything about the particular shapes of the distributions. To do this, it is necessary to estimate the hit and false alarm rates for several different positions of the decision criterion. These pairs of estimates are then presented in a scatter diagram, with the false alarm rate on the abscissa and the hit rate on the ordinate. An example of such a 'receiver operating characteristic' (ROC) curve is shown in Figure 2. In addition to illustrating how different degrees of bias towards one of the two responses would affect a decision-maker's performance, such plots can also provide a convenient estimate of sensitivity, because the area underneath a plot will generally increase as the sensitivity of the subject increases.

Of course, in order to estimate the ROC curves at more than one point, the investigator needs more estimates of the hit and false alarm rates than are available from a standard discrimination task. One approach is to run several experiments using different base rates, presumably inducing the subjects to shift their decision criteria to the left or right depending on the relative frequency of the signal. However, since this approach is more costly in both time and effort, most investigators prefer an alternative method, which is to run a single condition with equal base rates but to ask the subjects to report both the stimulus that they think was presented and their degree of confidence in this. In effect, this is equivalent to asking the subjects to report the position of the sensory effect on the decision axis as well as the response they wish to

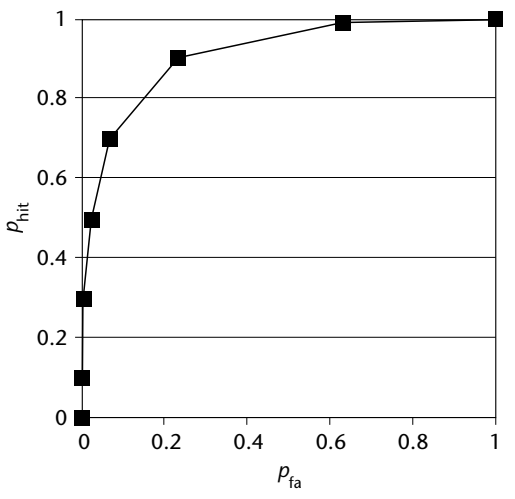


Figure 2. ROC curve from a hypothetical discrimination experiment. Points along the curve correspond to conditions with different signal base rates.

assign to this effect. From this extra information it is possible to estimate the ROC curve.

APPLICATIONS OF SDT

SDT has been adopted as a framework for studying human performance in many different areas of basic and applied research, including perception and memory, decision-making, radar monitoring, airport security, and personnel evaluation. Because the need to discriminate between two possible states of the physical world from imperfect (but not useless) information arises in so many situations in the laboratory and in the workplace, it is important to have a rigorous means of describing these kinds of activities and the effects that decision-making biases can have on them. When its underlying assumptions are known to be satisfied, SDT can be a useful tool to help assess decision-making styles and distinguish these styles from other aspects of the skills involved in a discrimination task. In these cases, it provides a legitimate method of identifying and helping to correct potentially harmful biases of decision-makers in critical decision situations. In addition, SDT is often used as a basis for testing theories that make predictions about factors that either should or should not affect the inherent difficulty of a task (rather than merely the strategies that subjects will choose to employ), and to summarize in an efficient way the effects of these factors on performance.

However, these applications of the model have little or no scientific merit if the model's representation of behavior is inaccurate. It is important, therefore, to test the model's assumptions in a rigorous way before accepting them. Some of the most convincing evidence in favor of the model comes from manipulations that according to SDT should be directly related to the decision-making process (the placement of the criterion) and unrelated to the sensory encoding process (and hence the sensitivity level). In particular, if the noise and signal trials are not presented with equal frequency (the base rates are unequal), or if a 'miss' is, say, more costly than a 'false alarm', the optimal decision-maker will shift the criterion away from the point of intersection. If the criterion is shifted to the right, the hit and false alarm rate should both decrease; and many studies have shown that this is precisely what occurs when the frequency of the noise stimulus (or the penalty for a false alarm) is increased. In other words, this fundamental prediction of the SDT model is confirmed by empirical data.

SDT also makes some other predictions, however, that have been shown to be violated. If the criterion is biased towards, say, the signal response, as in the example shown in Figure 3, the theory predicts that the relative frequency of low-confidence signal responses (effects immediately to the right of the criterion) should be lower on signal trials than on noise trials. This prediction is

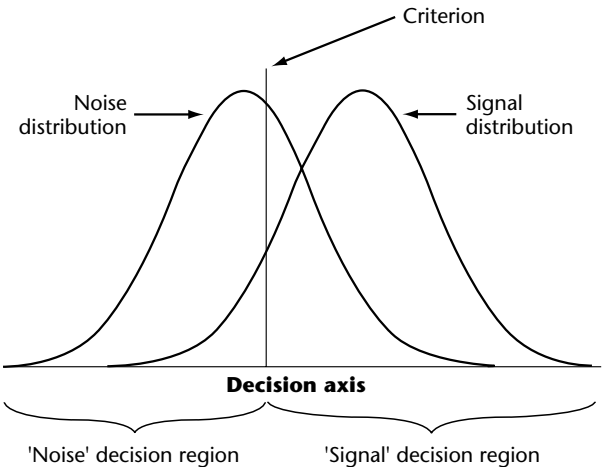


Figure 3. An SDT model with a biased decision rule. The criterion has been shifted to the left of the point of intersection between the two distributions, indicating a bias towards the signal response (compared to the unbiased rule, more of the effects on the decision axis are mapped to the signal response).

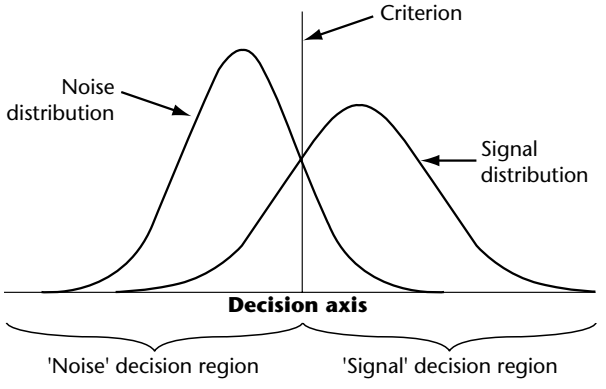


Figure 4. A modification of the SDT model that is consistent with empirical data. Instead of the decision criterion shifting, the relative variances of the two distributions change as the bias increases. Higher base rates of a stimulus lead to a distribution with smaller variance. In this example, increasing the base rate of the noise stimulus decreased the variance of the noise distribution and increased the variance of the signal distribution. The decision rule is unbiased, but the change in variance accounts for the relationship between hit and false alarm rates when base rates are manipulated.

consistently violated in discrimination experiments. These two relative frequencies convergence to the same value as the subjects' confidence decreases (the likelihood ratio at the criterion is 1).

In order to explain the covariance of the hit and false alarm rates under base rate manipulations, while also explaining this direct empirical evidence that the decision rule is always unbiased, the SDT model needs to be modified so that the shapes of the distributions change as the base rates are manipulated. An example is shown in Figure 4. Instead of the criterion shifting, the relative variances of the two distributions change, with a higher base rate for a given stimulus leading to lower variance in its associated distribution of sensory effects. Obviously, this is a very different kind of bias, and one which eliminates the 'shifting criterion' representation of decision-making processes in SDT. It is not yet clear exactly what causes the distributions to change shape in the manner shown in Figure 4, or what consequences this effect has for measurement of discrimination skills. More research is needed to answer this important question. Meanwhile, applications of the traditional SDT models should be regarded sceptically.

Further Reading

Green DM and Swets JA (1966) *Signal Detection Theory and Psychophysics*. New York, NY: John Wiley.

Balakrishnan JD and MacDonald JA (2001) Alternatives to signal detection theory. In: Karwowski W (ed.) *International Encyclopedia of Ergonomics and Human Factors*. London, UK: Taylor & Francis.

Swets JA (1998) Enhancing diagnostic decisions. In: Hoffman RR, Sherrick MF *et al*. (eds) *Viewing Psychology as a Whole: The Integrative Science of William N. Dember*. Washington, DC: American Psychological Association.

Similarity

Intermediate article

Ulrike Hahn, University of Wales, Cardiff, UK

CONTENTS

Similarity can be viewed as the degree of resemblance between two objects or events. This is thought to influence a wide variety of cognitive processes.

INTRODUCTION

Similarity is a ubiquitous concept within cognitive science. Though not without critics, it forms part of the explanation of cognitive processes as diverse as memory retrieval, categorization, visual search, problem-solving, learning, language processing, reasoning, and social processes. Several distinct approaches to similarity have emerged; each of these can be viewed both as a tool for the measurement of similarity and as a theoretical statement about its nature.

SPATIAL MODELS

Spatial models seek to represent similarity in terms of distance in a psychological space (see Shepard, 1980, for an overview). An item's position is determined through its coordinate values along the relevant dimensions; nearby points thus represent similar items, whereas distant items are psychologically very different (see Figure 1). Typically, the relevant space is derived from matrices of 'proximity' data such as pairwise confusability or similarity ratings between items through the application of multidimensional scaling (MDS) – a statistical procedure for dimensionality reduction. Spatial models have been used widely for data visualization. They have also formed the heart of very detailed cognitive process models, for example, of recognition and classification.

FEATURAL MODELS

In contrast to the continuous-valued dimensions of spatial models, featural models assume binary features as their representational basis. In other words, a continuous dimension such as 'size' might be reduced to a single psychological feature 'tall' or 'not-tall'. (Often cumbersome schemes for translating dimensions into features and vice versa have been proposed in Tversky (1977).) The most famous featural account of similarity is Tversky's *contrast model*. This account was developed to address perceived limitations of spatial models which seem to bring these into conflict with behavioral data (see below).

In the contrast model, similarity is a function both of the features common to the pair of items under comparison and of the features distinctive to each. In full, the model is specified as:

$$\mathrm{Sim}(\mathrm{D},\mathrm{E}) = \alpha f(\mathrm{D} \cap \mathrm{E}) - \beta f(\mathrm{D} - \mathrm{E}) - \gamma f(\mathrm{E} - \mathrm{D})$$

where D and E are both objects or events each of which is represented in terms of a set of characteristic features; $(\mathrm{D} \cap \mathrm{E})$ represents the set of features common to both D and E, whereas $(\mathrm{D}-\mathrm{E})$ and $(\mathrm{E}-\mathrm{D})$ represent the distinguishing features of D and E, respectively. Each of these three feature sets is goverened by two kinds of parameters (for full details see Tversky, 1977): f represents a scaling parameter reflecting the salience or prominence of different features; α, β, and γ are weighting parameters that allow the model to capture attentional differences present in so-called directional similarity comparisons (e.g. 'how similar is D to E', as opposed to 'how similar *are* D and E'). They enable the model to capture asymmetries in similarity judgments, to be discussed in more detail below, which have been put forth as important empirical evidence against spatial models.

ALIGNMENT AND TRANSFORMATION MODELS

More recently, it has been argued that spatial and featural models share a fundamental limitation in that they define similarity over very specific – and very simple – kinds of representations: points in

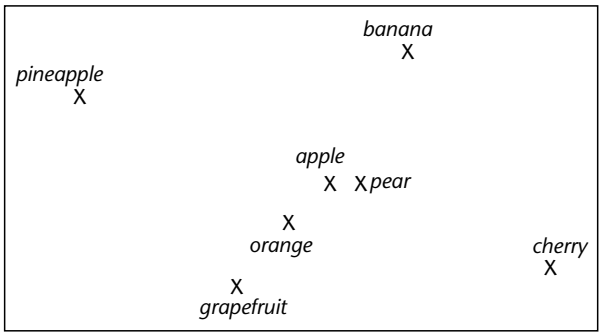


Figure 1. A sample MDS-derived spatial representation of the similarity relations between various kinds of fruit.

space or feature sets. Arguably, most theories of the representation of natural objects, whether they be faces, visual or auditory scenes, or sentences, assume that they cannot be represented in line with these restrictions. Instead, they seem to require *structured representations*: complex representations of objects, their parts and properties, and – crucially – the interrelationships between them. Such descriptions cannot be boiled down to either lists of features or points in space. This can be illustrated with a simple relation such as MOTHER-OF(,) and its arguments *Jane* and *Jim Jnr* which, as MOTHER-OF(*Jane, Jim Jnr*), provides a represention of the fact that Jane is the mother of Jim Jnr in a classic scheme for structured representations – the language of first-order predicate logic.

This representation scheme distinguishes relations and their arguments so both can be accessed and compared independently: the MOTHER-OF relation might be found to be semantically similar to the relation FATHER-OF(,), and the state of affairs represented by MOTHER-OF(*Jane, Sally*) would be classed as closely related as well. However, on a featural or spatial scheme, relations and their arguments are necessarily bound together into single atomic units which represent objects or states of affairs as a unitary whole. Such a single feature representing the composite fact that Jane is Jim Jnr's mother does not allow independent access of the relation and arguments in question because these are representationally bound together into a single unit. Consequently, the feature representing the fact that Jane is Jim Jnr's mother is completely distinct from a feature representing the fact that Jim is Jim Jnr's father or that Jane is also Sally's mother. None of the features representing these events will match exactly (otherwise they couldn't represent different events), and the fact that they represent events with overlapping components is inaccessible because featural schemes are by definition not componential.

The structural alignment account of similarity (e.g. Gentner and Markman, 1994), which has its roots in the literature on analogical reasoning, operates over structured representations. The structural alignment process begins with semantic similarities between two comparison objects or events and then seeks a structurally consistent mapping. The semantic comparison requires that at least some of the predicates (that is, relations such as ABOVE(x,y)) are identical across the comparison. These identical predicates are placed in correspondence and the alignment process then seeks to build maximal structurally consistent matches between the two representations. Structurally consistent matches are ones that respect the constraints of *parallel connectivity* and *one-to-one mapping*. Parallel connectivity implies that wherever relations are placed in correspondence, their arguments are placed in correspondence also. The one-to-one mapping constraint requires that each element in one representation must match at most one element in the other representation. (In Figure 2, parallel connectivity means that the two 'function' predicates will be placed in correspondence, as will their arguments 'sitting' and 'eating'.) Structural alignment has been implemented in a variety of computational models

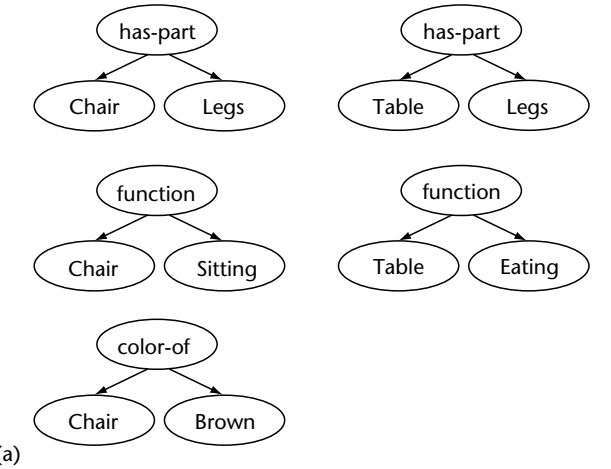


Figure 2. An example of structural alignment (after Markman, 2001).

(see Markman, 2001) which have been used to capture behavioral data.

The capacity to deal with structured representations is also central to recent transformational approaches to similarity (Hahn *et al.*, 2001). In the transformational framework, the similarity between two object representations is seen to be a function of the 'effort' required to transform one representation into the other. For example, an image of a robin could be more readily transformed into an image of a blackbird than into an image of a dog, reflecting the greater similarity between the two birds. The basic idea of a transformation-based measure of similarity is also present in the edit-distance measures popular in psycholinguistics. Here, the similarity between two sound sequences depends on the number of insertions, deletions, and substitutions required to turn one sequence into the other. The words 'bet' and 'bent', for instance, are separated by only a single insertion ('n') and sound far more similar than do 'bet' and 'mat' which require two substitutions in order to be made identical.

EMPIRICAL DISTINGUISHABILITY

Experimental attempts to reveal the psychological function governing similarity initially focused on the so-called metric axioms which underlie any spatial scheme (see Tversky, 1977). Several apparent violations of these axioms in behavioral data were discovered, and were taken as evidence against the psychological validity of spatial accounts. Because the contrast model is not subject to the metric axioms such violations were presented as evidence in its favor, but they equally support any other approach not subject to these assumptions.

The two most important metric axioms in this context are (1) the symmetry constraint, and (2) the so-called triangle inequality.

Distances in the psychological space posited by spatial models are necessarily symmetrical: the psychological distance of item D to item E necessarily equals that of E to D. However, there is considerable evidence for asymmetries in human similarity data, whether these data be explicit ratings or confusability data (Tversky, 1977). In particular, typical members of a category may be perceived as less similar to atypical members than vice versa; for example, the similarity of penguins to robins might be rated higher than that of robins to penguins.

The triangle inequality, by contrast, imposes a constraint on the possible relationships between more than two items. Specifically, for any three distances d between the items a, b, and c:

$$d_{ab} + d_{bc} \geq d_{ac}$$

An informal apparent counterexample is provided by Tversky's example of Jamaica, Cuba, and (then Soviet) Russia. Jamaica (a) seems very similar to Cuba (b), as does Cuba (b) to Russia (c), yet Russia and Jamaica seem very dissimilar indeed; in other words, the psychological distance (ac) between Russia and Cuba is likely to exceed the sum of the other two distances (ab) and (bc). Ingenious experiments showing violations of this metric axiom were presented by Tversky and Gati (1982).

Neverthless, the demonstration of these various violations of fundamental constraints of spatial models has been less damaging than might have been expected. This is largely due to the fact that the most successful applications of the spatial paradigm to cognitive modeling embed these in more encompassing models; Nosofsky's spatially based *generalized context model* of identification, recognition, and classification, for example, involves an exponential function which converts the basic distances of the underlying spatial model into a measure of psychological similarity. This allows apparent violations of the triangle inequality. Similarly, the model incorporates a bias term which allows it to capture asymmetries (see Nosofsky, 1991 for complete details).

Potentially more damaging to spatial models – and equally threatening to featural accounts – is the evidence for the importance of structural information in human similarity judgments. One prediction of the structural alignment account is of a cognitive difference between *alignable* and *unalignable differences*. Unalignable differences between two representations are representational elements of one object which have no correspondence in the other; alignable differences, by contrast, are representational elements which correspond across the two representations but which are nonidentical. (In Figure 2, 'color' for which no information is included in the representation of 'table' constitutes an unalignable difference between the two objects. The different functions 'eating' and 'sitting' constitute alignable differences.) This distinction has received a fair amount of empirical support, both in that participants typically list more alignable than nonalignable differences in free listing tasks, and in that alignable differences seem to affect perceived similarity more strongly than do unalignable differences. Relational structure does seem to matter in the context of similarity, and this would seem to count against both spatial and featural accounts.

At present, spatial and featural models survive, not only owing to their simplicity but because spatial models, in particular, have very successfully provided detailed models of cognitive performance in a variety of tasks. This might be because the particular stimulus materials used simply did not require relational information to be processed; in this case, models should be shown to falter with more naturalistic materials. It is also possible that the cognitive system makes use of both fairly 'cheap' computations well approximated by spatial models and computationally more 'expensive' processing which incorporates relational information for higher-level tasks. Only further research, not only into similarity but also into the nature of supposedly similarity-based cognitive processes, will be able to tell.

References

Gentner D and Markman AB (1994) Structural alignment in comparison: no difference without similarity. *Psychological Science* **5**: 152–158.

Hahn U, Chater N and Richardson LB (2001) Similarity: a transformational approach. In: *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pp. 393–398. Hillsdale, NJ: Erlbaum.

Markman AB (2001) Structural alignment, similarity, and the internal structure of category representations. In: Hahn U and Ramscar M (eds) *Similarity and Categorization*, pp. 109–130. Oxford, UK: Oxford University Press.

Nosofsky RM (1991) Stimulus bias, asymmetric similarity and classification. *Cognitive Psychology* **23**: 94–140.

Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* **210**: 390–397.

Tversky A (1977) Features of similarity. *Psychological Review* **84**: 327–352.

Tversky A and Gati I (1982) Similarity, separability, and the triangle inequality. *Psychological Review* **89**: 123–154.

Further Reading

Gentner D (1989) The mechanisms of analogical learning. In: Vosniadou S and Ortony A (eds) *Similarity and Analogical Reasoning*, pp. 199–241. Cambridge, UK: Cambridge University Press.

Goldstone R (1994) The role of similarity in categorization: providing a groundwork. *Cognition* **52**: 125–157.

Hahn U and Chater N (1997) Concepts and similarity. In: Lamberts K and Shanks D (eds) *Knowledge Concepts and Categories*, pp. 43–92. Hove, UK: Psychology Press/MIT Press.

Nosofsky R (1986) Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**: 39–57.

Tversky A and Hutchinson J (1986) Nearest neighbor analysis of psychological space. *Psychological Review* **93**: 3–22.

Skill Learning

Introductory article

Richard A Carlson, Pennsylvania State University, University Park, Pennsylvania, USA

CONTENTS

Skill learning is the acquisition and improvement of mental or physical abilities through practice.

INTRODUCTION

Skill refers to an acquired ability that has improved as a consequence of practice. Skills may be primarily physical or *motor* in nature (for example, riding a bicycle) or largely mental or *cognitive* (playing chess, solving mathematical problems). Skill learning obeys similar principles across motor and cognitive domains. It can be distinguished from other types of learning such as conditioning, acquiring knowledge of facts, or learning a first language, though skill learning and these other types of learning may share some underlying mechanisms. *Expertise* usually refers to a broad repertoire of knowledge and skills in some domain, and can thus also be distinguished from skill. Studies comparing experts and novices in particular domains have, however, been important in advancing understanding of skill learning. Finally, skill should be distinguished from *talent*, an innate capacity to acquire certain types of skill or knowledge.

PHENOMENA OF SKILL ACQUISITION

The most obvious feature of skill learning is improvement with *practice*, the repeated performance of the same activity. All modern theories of skill learning share the view that practice results in the acquisition and improvement of relatively specific knowledge, in contrast to historical views that practice could improve abstract, general mental abilities.

A key feature of practice is regularity or *consistency* – something must stay the same from trial to trial in order for performance to improve. Improvement in performance may be measured in a variety of ways: by a reduction in errors or increase in the likelihood of success, by more precise coordination of component actions with one another or with the environment, by reduction in the effort required for performance, or by more rapid performance.

Each of these approaches to assessing skill might be seen as one aspect of *fluency*, the ability to perform an activity readily, smoothly, and effectively. Of course, which aspect is most important depends on the particular skill being considered – it might be argued that faster is always better in solving mathematical problems, but it would be odd to argue that faster is always better in performing music (though the possibility of performing faster indicates a higher level of at least some components of musical skill). Research has established a number of empirical phenomena that indicate the ways in which fluency increases as a consequence of practice.

Increases in Speed

Research on skill learning has often emphasized speed of performance. Across a wide range of mental and physical skills, the effect of practice on speed of performance follows a similar function: large speed-ups are observed over the first few attempts, with additional trials resulting in progressively smaller increments in speed. Researchers have observed this pattern with skills as disparate as solving mathematical problems, typing, rolling cigars, and even writing novels.

This pattern of speed-up with practice is usually well described by a mathematical *power function*, which states that the time required to perform a task is a function of the number of trials raised to some power. This function is represented by the equation $T = a + bN^{-c}$, where T is the time needed to perform the task, a is the asymptotic (shortest possible) time, b is the total amount of speed-up possible, N is the number of practice trials, and the exponent c is the rate of learning. The power function implies that speed-up will continue indefinitely, though after extensive practice each additional trial will produce only very small speed-ups

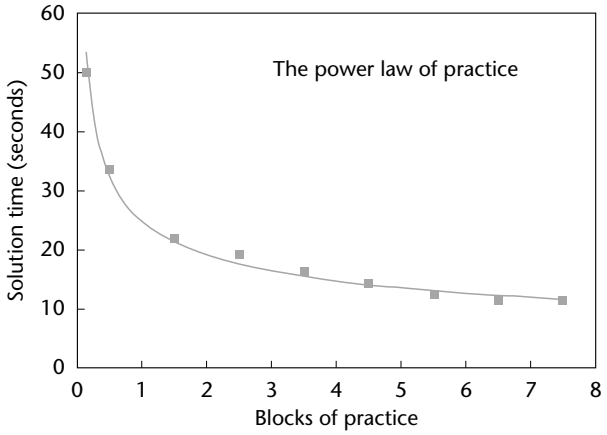


Figure 1. Time required to solve multiple-step arithmetic problems as a function of practice. A power function curve is fitted to the data.

(see Figure 1). This relation between practice and performance time has been described as the *power law of practice*, though some authors have argued that alternative functions (e.g. exponential functions) are better mathematical representations of the practice–speed relation.

Regardless of the exact mathematical representation, it is clear that learning curves typically show greater improvement early in practice than late in practice. Recent analyses suggest that smooth, continuous power-law speed-up will be observed only for particular strategies or component skills, and that much of the speed-up of complex skills depends on the combined effect of component skill speed-up and reorganization of component skills.

Restructuring and Chunking

Another consequence of practice is *restructuring*, the reorganization of activity to generate superior performance. A simple example is the acquisition of arithmetic skill: most theorists agree with the common observation that children learning addition shift from relatively laborious, inefficient counting procedures to direct, single-step retrieval of addition facts from memory. This kind of restructuring makes it possible to skip steps that may have been necessary in novice performance. Restructuring may depend in many cases on the acquisition of *perceptual chunks*, perceivable patterns that are meaningful in the context of a particular skill. For example, chess experts are superior to novices in identifying and remembering particular arrangements of chess pieces, if those arrangements are meaningful in the context of chess games.

Some researchers concerned primarily with complex physical skills have suggested that skill learning may involve the discovery of new *coordinative modes*, new ways of relating component actions to apply the system of components more effectively. Running and walking, for example, involve two coordinative modes for locomotion, organizing the movements of various parts of the body differently.

Another form of restructuring is *information reduction*. Skilled performers are more efficient in selecting information from the environment, reducing the amount of information that must be considered in order to perform an activity. For example, skilled pilots have efficient strategies for scanning cockpit instruments to guide their decision-making and control of flying.

Automaticity

A particularly interesting consequence of practice is *automaticity*, the ability to perform a task rapidly with minimal effort, little need for deliberate control, and sometimes little or no memory of having performed the task. Acquired automaticity has been documented by showing that, after extensive practice, individuals can perform tasks with little interference from other concurrent tasks, and with little effect of factors such as the number of alternative stimuli or responses. For example, anovice performing the laboratory task of searching a computer display for a particular target letter will be slower, the more items are displayed. Given sufficient practice with consistent search – that is, a letter that is a target is always a target – individuals in some circumstances can learn to search displays equally rapidly regardless of the number of letters displayed. Familiar examples of acquired automaticity include such skills as driving a car.

Such examples make clear that most real-life skills involve a combination of automatic and deliberate or controlled components. Automaticity is a matter of degree, and large amounts of practice are typically required for high levels of automaticity. Learning component skills to a high level of automaticity may be important for combining those skills in performing more complex tasks. For example, the ability to retrieve arithmetic facts automatically may be important for skilled solving of mathematical problems.

Retention and Transfer

The retention of skill over time, and the possibility of *transfer* from particular practice conditions to new situations or tasks, are important for both

practical and theoretical reasons. Retention of well-learned skills is often very good, with little decline in skill over months or years of disuse. Larger amounts of practice may increase retention, even when practice is continued to the point that improvements in performance are very small. Interestingly, retention often cannot be predicted effectively from practice performance. Conditions such as reduced feedback or interference among multiple skills may lead to poorer performance during practice but to superior retention.

Skill learning also typically exhibits *specificity of practice* – when the conditions for performing a skill are changed only slightly, the benefits of practice may be lost or dramatically reduced. For example, practice in generating lines of computer code may provide little benefit in evaluating what similar lines of code do. Another way to describe this phenomenon is to say that *transfer* – the application of knowledge or skill acquired in one context to a new situation – is typically quite narrow. Transfer can usually be improved by designing practice so that learners focus on aspects of a skill that are sufficiently abstract to apply to new situations, but specificity of practice is a major practical limit on skill training.

Dissociations of Declarative and Procedural Knowledge

Skill learning often demonstrates a dissociation between *declarative* and *procedural* knowledge of the skill. Declarative knowledge is knowledge that can be explicitly expressed ('declared') or consulted, whereas procedural knowledge ('knowing how') can only be performed. A commonplace observation is that knowing how to ride a bicycle does not imply an ability to describe how one rides a bicycle.

Researchers have documented this dissociation in various ways. For example, some patients with brain damage that produces near-total amnesia (and thus an inability to explicitly recall their experience of learning) can nevertheless acquire new skills. In some laboratory tasks, learners appear to acquire skills *implicitly* in the sense that their performance of tasks improves even when they cannot describe the regularities that allow them to learn the task. One consequence of this dissociation is that highly skilled performers are often unable to reflect on or talk about how they achieve their skilled performance.

Individual Differences

Individuals differ both in the ease with which they acquire skills and in the final level of skill they achieve. The existence of highly specific talents for learning particular skills is controversial, and many individual differences may result from differences in motivation and experience rather than innate talent. For many skills, though, individual differences in general ability, perceptual speed, and psychomotor ability do affect the course of skill acquisition. General ability is typically most important in the early stages of skill learning, whereas perceptual speed and psychomotor ability are more important in later stages of learning and the final level of skill achieved.

STAGES OF SKILL LEARNING

In many situations, skill learning appears to proceed through a series of characteristic stages. In acquiring a new skill, the first problem for the learner is to perform the skill for the first time, beginning the practice that will finally result in a fluent skill. This initial performance is often guided by verbal instruction (possibly self-instruction generated by the learner), by analogy to an existing skill, or by the use of an example or model to be imitated. A learner may continue to rely on such information for some time, perhaps refining his or her self-instruction or using examples or models in a more focused way. This initial stage of skill acquisition has been described as *cognitive, directed*, or *declarative*.

There has been some controversy over the best characterization of this initial stage, as some authors have argued that it necessarily involves declarative knowledge whereas others have pointed to such phenomena as *implicit learning* as evidence that skill can be acquired in the absence of relevant declarative knowledge. It seems clear, however, that initial performance of a new skill requires some kind of direction, usually external, that is not needed once the skill has been extensively practised. During the cognitive stage of skill learning, performance is limited both by lack of knowledge and by limits on working memory, the ability to temporarily hold information needed to support performance. Depending on the particular skill considered, this initial cognitive stage of learning may last for only a few trials, or may extend for some time.

Relatively modest amounts of practice are typically sufficient for a learner to move beyond the initial cognitive stage of skill learning. However, restructuring of the skill may continue for a long time, resulting in increasingly refined strategies and procedures. This intermediate stage of practice has been described as the *associative* or *knowledge*

compilation stage. During this stage, learners may begin to skip some steps that were required in the cognitive stage as they acquire more efficient procedures.

A third stage of skill acquisition involves gradual continuing improvement in the performance of a skill that is already stable and well-learned. This continuing improvement has been called the *autonomous, procedural*, or *routine* stage of skill learning. The routine stage of skill learning may extend for many thousands of trials, with performance continuing to show small improvements in fluency.

CONDITIONS OF PRACTICE

The success and efficiency of skill learning depends on the conditions of practice. Simple repetition may suffice to improve performance of some skills, but the effectiveness of practice depends on a variety of factors.

External Support

Skill acquisition, especially in the initial cognitive stage, often depends on external support for the initially unskilled performance. This external support may be provided by a teacher or coach who guides the learner and provides feedback. Increasingly, external support for skill learning is provided technologically by means such as videotape or computer software.

Some form of guided learning is necessary in at least the initial stage of learning most skills. However, research suggests that strict guidance, such as that provided by the keystroke templates in many computer tutorials, results in less effective learning than discovery learning, in which the learner must discover how to perform the task. External support that helps learners overcome the working memory difficulties typically encountered in complex mental skills may also be important in allowing multiple skill components to be held in mind together in order that links among them can be learned.

The initial performance of a skill is often guided by an example, model, or analogy. For cognitive skills such as solving mathematics or science problems, learners often make use of worked examples that illustrate the solution of similar problems. For perceptual-motor skills, learners often rely on observing a model, a skilled performer whose actions can be imitated. The issues involved in using examples or models are similar: learners must be able to correctly interpret and analyze the example or model, and must already possess component skills adequate to follow the example or model. Examples and models are most effective when they convey the hierarchical structure of a task domain. For example, students learn to solve mathematical problems most effectively when example problems highlight particular subgoals and their relations to the overall goal of solving a problem.

Many skills involve interacting with a mechanical apparatus (for example, piloting an airplane) or a symbolic medium (for example, solving a mathematical problem with pencil and paper or on a computer). These devices or media provide external support for memory and for monitoring performance. Of course, all skills are performed in some environment that provides external support for some of the psychological aspects of the skill. However, practice need not depend on external support. For many skills, *mental practice* can improve performance, though usually not as much as practice in the actual task environment.

Feedback

An important aspect of practice is the *feedback* available to the learner, information that specifies to what degree the skill is performed accurately and fluently. Feedback may result from the learner's own comparison of his or her performance to a mental standard, or may be provided explicitly by a teacher or by some mechanical means such as a software program. Research on the effects of feedback has focused for practical reasons on explicit feedback.

Feedback may vary in a number of ways. First, it may convey information primarily about the final outcome of a performance (*knowledge of results*) such as whether the answer to a problem is correct, or about the dynamic characteristics of the performance (*knowledge of performance*). Second, feedback may be available immediately on every attempt to perform a skill, or it may be delayed, intermittent, or summarized over multiple attempts. Finally, feedback may vary in its precision or in the aspects of outcome or performance about which it is informative.

Feedback is usually most effective when it is specific and immediate, allowing the learner to focus on those aspects of a particular performance that need improvement. However, learning may be better when feedback is not provided on every trial,

because learners given continual specific feedback may come to rely on that feedback to guide their performance and thus perform poorly when feedback is no longer available.

Practice Schedules

Another important aspect of practice is how that practice is distributed over time. *Distributed* or *spaced practice*, in which a skill is practised on multiple occasions spread over an extended period of time, generally results in better retention than does *massed practice* in which the skill is practised repeatedly in a relatively short period of time. The benefits of distributed practice have been demonstrated for a wide variety of skills, ranging from retrieval of arithmetic facts to complex perceptual-motor skills. A related phenomenon is the *contextual interference effect*, observed in verbal memory as well as in both motor and cognitive skills. When multiple skills are to be learned, practising the skills in a random schedule in which tasks are presented in a mixed fashion results in poorer performance during learning, but superior retention and transfer, relative to practising each skill in a separate block of trials.

Motivation

An important feature distinguishing effective from less effective practice is the learner's motivation. Motivation is of course necessary for practice to occur at all, a fact well known by teachers, parents, and coaches. Given that a learner does practise, however, the nature of the motivation with which they approach practising has a large effect on the level of skill attained. *Deliberate practice*, practice undertaken with the goal of improving performance, involves the selection of specific tasks and attention to particular cues that focus on aspects of performance to be improved. Research suggests that it is the amount of deliberate practice that largely determines the level of skill achieved in real-life domains such as music and sports.

Individuals who come to stand out as experts generally do so as a result of high motivation and large amounts of deliberate practice. However, individual differences in motivation may also explain differences in more modest levels of achievement. Some individuals seem to be characteristically motivated for immediate achievement, whereas others are characteristically motivated for learning. These differences in motivation result in differences in how learners approach tasks, in both laboratory and educational contexts. Individuals motivated for learning are likely to benefit more from practice, especially with difficult tasks.

THEORIES OF SKILL ACQUISITION

Behavioral Principles

Much theorizing about skill learning has involved primarily the accumulation of empirical principles, based on the generalizations described above, that provide guidance for effective practice. For example, it is safe to say that more practice is usually better, that appropriate consistencies make practice more effective, and that practice reduces the effort required to perform a task. A general principle governing the transfer of skill to new situations is known as *identical elements theory*, the idea that practice will transfer to the extent that the new skill shares identical elements – stimuli, rules, or responses – with the practised skill.

Cognitive–Symbolic Theories

Beginning in the 1970s, a number of researchers proposed and tested theories of skill acquisition based on the information-processing or computational approach to psychological theory. These theories explain skill learning by considering changes in the representation of knowledge as a function of practice. For example, one way to explain increasing skill in arithmetic is to describe the learning of 'number facts' as the acquisition of specific memories associating problems and answers. These memories can then be used in place of more laborious procedures (such as multiplying by repeated addition), resulting in faster solution of arithmetic problems.

The most highly developed cognitive–symbolic theories take the form of *production systems*, computational systems that represent knowledge in the form of productions, or if–then rules, that specify the conditions under which particular actions will be taken. Productions represent a fine-grained analysis of skill, such that between five and 100 production rules would be executed each second during skilled performance. In production-system theories, the effects of practice are represented by the acquisition of new production rules, by the combining of small sets of production rules into single, more efficient production rules, and by the speed-up in application of individual rules.

Cognitive–symbolic theories have been successful in explaining many phenomena of skill learning, especially in cognitive domains, and continue

to support active research programs. These theories have generally de-emphasized the perceptual-motor details of how skills are performed.

A related idea in the domain of motor skill is *motor program theory*, which suggests that the acquisition of skill involves learning and tuning a motor program that directs movement. According to this theory, the knowledge underlying a motor skill is analogous to a computer program. The effect of practice, according to this theory, is to adjust parameters of the program for optimal performance.

Dynamical Systems Theories

Dynamical systems theories of skill learning have recently become prominent among researchers concerned with motor skills. A starting point for these theories is the observation that motor skills involve coordinating the movements of multiple muscles and joints to accomplish movement goals, and that movement goals can generally be accomplished in more than one way. For example, grasping a glass requires coordinating movements of the fingers, wrists, and arm, and one may successfully reach for and grasp the glass in a variety of ways.

Dynamical systems theory treats skill learning as the process of discovering ways of coordinating activity, and focuses on the physics of movement as a source of constraints on coordination. From this point of view, an important characteristic of learners is that their musculo-skeletal systems, together with the environment in which they act, can be seen as *self-organizing* in the sense that the parts of these systems mutually constrain one another.

In contrast to cognitive–symbolic theories, dynamical systems theories de-emphasize mental representations but may provide substantial detail concerning how skills are actually performed. The control of skilled performance is thus viewed as a joint function of the performer and the environment.

Neurological Theories

Recent advances in methods for studying the brain are providing new insights into the neural substrates of skill learning. One important finding is that the brain appears to have multiple systems for learning, involving different neural circuits and different means of representing knowledge. Some neural systems may be involved in many kinds of skill; for example, the cerebellum appears to be involved in perceiving and producing sequential activity in both mental and perceptual-motor skills. Other neural systems may be involved primarily in specific kinds of skills; for example, portions of the temporal cortex seem to be active mostly in the performance of skills that have verbal or symbolic components.

COMPLEX SKILLS: MUSIC AND SPORTS

Most laboratory research on skill learning has focused on relatively simple skills, or on intellectual skills such as scientific problem-solving. The study of some complex real-world skills, such those involved in music and sports, raises additional issues. These skills typically combine cognitive, perceptual-motor, and emotional or esthetic aspects.

For example, skilled musical performance involves thought about musical structure, skills for manipulating an instrument, and judgments about performance details that convey expressive aspects of the music. Skilled performance in sports requires not just high levels of perceptual-motor skill, but skill in making strategic and tactical decisions in game settings. Issues such as anticipation, flexibility in recruiting component skills, and precise tuning to the environment (including the actions of fellow players or, in the case of sports, opponents) are thus more prominent in music and sports than in most laboratory research on skill.

It is widely believed that talent plays a greater role in acquiring music and sports skills than it does in other domains, and this belief dominates the literature on skill in other esthetic domains such as painting and drawing. However, research provides little support for this view. For example, a high degree of sensitivity to musical structure appears to be nearly universal, though acquiring substantial musical skill requires a large amount of deliberate practice. Motivation and deliberate practice appear to account for much of the difference between outstanding performers in sports and music and those who are less skilled.

Domains such as music and sports involve public performance. The best performers are highly valued by society, and thus represent a highly selected sample characterized by considerable levels of skill. Research suggests, however, that the principles of skill learning observed in laboratory research also apply to these complex and culturally situated skills. The scientific study of skill learning thus promises to explain even the highest levels of skill.

Further Reading

Anderson JR (ed.) (1981) *Cognitive Skills and Their Acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Carlson RA (1997) *Experienced Cognition*. Mahwah, NJ: Lawrence Erlbaum.

Ericsson KA and Smith J (eds) (1991) *Toward a General Theory of Expertise*. Cambridge, UK: Cambridge University Press.

Healy AF and Bourne LE Jr (eds) (1995) *Learning and Memory of Knowledge and Skills: Durability and Specificity*. Thousand Oaks, CA: Sage.

MacKay DG (1982) The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychological Review* **89**: 483–506.

Newell KM (1991) Motor skill acquisition. *Annual Review of Psychology* **42**: 213–237.

Proctor RW and Dutta A (1995) *Skill Acquisition and Human Performance*. Thousand Oaks, CA: Sage.

Rosenbaum DA, Carlson RA and Gilmore RO (2001) Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology* **52**: 453–470.

Schmidt RA and Lee TD (1999) *Motor Control and Learning – A Behavioral Emphasis*, 3rd edn. Champaign, IL: Human Kinetics.

Singley K and Anderson JR (1989) *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.

Sleep and Dreaming

Introductory article

J Allan Hobson, Harvard Medical School, Boston, Massachusetts, USA

CONTENTS

Sleep and dreaming have an important role in the regulation of cognition. New investigative techniques suggest that brain activity during sleep is fundamental to learning.

INTRODUCTION

The cognitive neuroscience of sleep and dreaming has recently enjoyed a welcome growth spurt owing to a confluence of new results from neuropsychology, experimental psychology and fundamental neuroscience. Neuropsychology has contributed new brain imaging data regarding selective activation and deactivation of brain regions during normal sleep and complementary findings regarding the effects of brain damage upon dreaming in neurological patients. Experimental psychology has shown that sleep is not only permissive but essential to certain kinds of learning. Fundamental neuroscience has provided new models for understanding how the various stages of sleep are controlled by the brainstem and linked to the circadian rest–activity cycle controlled by the hypothalamus. The result is a coherent picture that illustrates the dynamic and highly functional role of sleep and dreaming in the regulation of cognition as well as other adaptive responses. For a discussion of the implications of these findings for theories of consciousness. (*See* **Consciousness, Sleep, and Dreaming**)

STAGES OF SLEEP

The rich differentiation of sleep into its component phases has been revealed by laboratory studies of humans and other mammalian subjects using the electroencephalogram (EEG) for recording brain waves; the electromyogram (EMG) for recording postural muscle tone; and the electrooculogram (EOG) for recording eye movement (Table 1, Figure 1). In essence the sleeping brain fluctuates between epochs of relative inactivation (quiet or slow-wave sleep) and epochs of activation (active or fast-wave sleep). These epochs can be distinguished by the presence or absence of rapid eye movements (REM). Nonrapid eye movement (NREM) or slow-wave sleep is associated with a decline of cortical electrical activity and of postural muscle tone. The NREM phase occupies about 75% of total sleep time and can be further subdivided into four stages according to EEG characteristics (Table 2). The degree to which NREM sleep is dominated by stages III and IV declines over the night along with the depth of sleep and the level of associated mental activity, although the period and length of each cycle is constant (Figure 1b). Stages III and IV are associated with energetically restorative processes as evidenced by the release of growth hormone and the secondary sex hormones follicle stimulating hormone and luteinizing hormone, as well as by low levels of sympathetic autonomic activity: decreased blood pressure, respiratory rate, etc. (Figure 1c). People are difficult to rouse from stage III or IV and often experience confusion, disorientation and confabulation on arousal. Reports of dreaming are rare, although sleep-walking, sleep-talking and nightmares can occur. As the night progresses the NREM epochs become less sharply differentiated from the REM epochs and mental activity often indistinguishable from REM is increasingly reported.

At intervals of about 90 min throughout the night the NREM phase is interrupted by the spontaneous EEG activated eye movements and muscle tone suppression of REM. This REM phase is typically short early in the night but lengthens as sleep progresses so that it occupies about 25% of total sleep time. Rapid eye movement sleep is correlated with increases in sympathetic autonomic activity and with penile erection; and arousals yield long and detailed reports of dreaming.

This cyclical organization of sleep into NREM and REM phases is seen in all mammals at all

Table 1. Electrographic differentiation of waking and sleep

	Awake	*NREM*	*REM*
EEG	+	−	+
EMG	++	−	0
EOG	+	−	+

EEG, electroencephalogram; EMG, electromyogram; EOG, electrooculogram; NREM, nonrapid eye movement; REM, rapid eye movement [sleep].

Table 2. Electroencephalographic characteristics of NREM sleep stages

NREM sleep stage	*EEG frequency range (Hz)*	*EEG rhythm*
I	4–8	Theta
II	12–15	Sigma (spindles)
III	1–4 + 12–15	Spindles plus slow waves
IV	1–4	Delta (slow waves)

EEG, electroencephalogram; NREM, nonrapid eye movement.

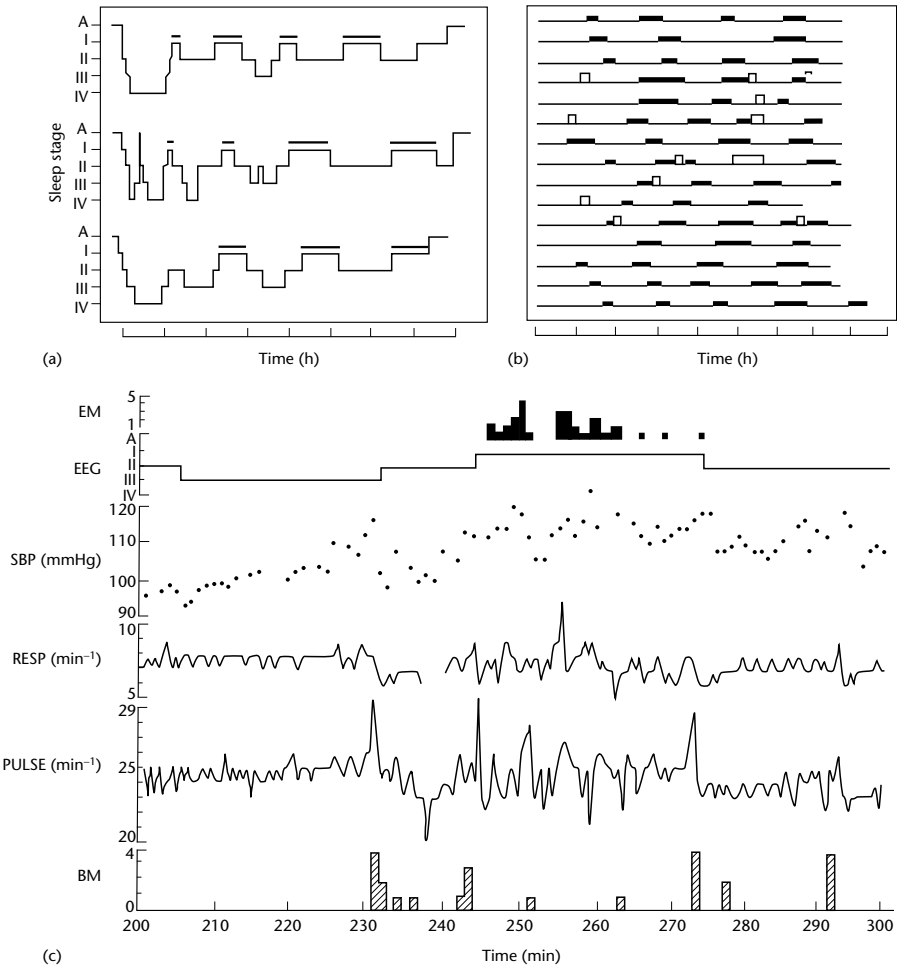


Figure 1. Sleep cycle with periodic activation. (a) Detailed sleep-stage graphs of three human subjects: note the preponderance of the deepest stages of NREM sleep in the first two or three cycles of the night; REM sleep is correspondingly brief (in subjects 1 and 2) or even aborted (subject 3). During the last two cycles, NREM is restricted to the lighter stage II, and REM periods occupy proportionally more time, with individual episodes often exceeding 60 min. (b) REM sleep periodograms of 15 human subjects show the same tendency to increase REM sleep duration. (c) Fluctuations in sympathetic autonomic activity. EEG, electroencephalogram; RESP, respiratory rate; SBP, systolic blood pressure.

stages of development but the period length of the cycle varies as a function of size and age (the larger and older, the longer), and the characteristics of the NREM and REM phases also change with development. Newborns have abundant REM sleep indicating its possible importance for early brain

development. Adolescents have abundant stages III and IV sleep, signaling its importance for psychosexual and intellectual maturation. With advancing age both sleep length and depth typically decline together with the gradual deterioration of the brain and cognitive function.

Normal Variations in Sleep Length and Depth

Like all other biological variables, the length and depth of sleep show large individual differences. Some people, 'larks', awake early and easily in the morning; others, 'owls', are able to stay up late at night without feeling tired. Whatever their times of sleep onset and offset, some are refreshed by as little as 4 h sleep while others need 8–10 h and may still feel listless by day. Short sleepers tend to be more energetic, ambitious, optimistic and productive, while long sleepers are more easily fatigued, accomplish less, and experience pessimistic and even depressed moods. Threshold to arousal also varies normally: some can sleep in daylight, with the windows uncovered and fire engines roaring by, while others have difficulty staying asleep in the depth of night, in the total silence and comfort of a well-insulated bedroom.

Whether or not the extremes of normal sleep variability should be considered pathological is an important clinical judgment, particularly in view of the risks of long-term sedative and stimulant use. Individuals who have too little or too much sleep for their social and physical comfort should know and follow certain simple rules before turning to medication. The first is that the timing of sleep is very sensitive to the time of waking; getting up early is the best way to promote falling asleep at the desired time of night. The second is that sleep is exquisitely carefully self-regulated by the brain; this means that debts are carefully accounted and paid back. The third is that sleep is impeded by stress, anxiety, and obsessive thinking, but enhanced by physical work and exercise.

The fact that cognition (depending on its content) can either benefit or impede sleep is a corollary of the functional principle that sleep, depending upon its content, can benefit or impede cognition. The worst thing that a poor sleeper can do is think about it, because obsessive reflection makes sleep worse. On top of that, poor sleepers tend to grossly overestimate the time spent awake, hence both exaggerating their insomnia and applying the resulting anxiety to their sleep loss worries. This vicious circle responds to its treatment of choice: cognitive behavioral therapy.

SLEEP DEPRIVATION

Following the discovery of REM sleep by Aserinsky and Kleitman in 1953 and its correlation with dreaming by Dement and Kleitman in 1957, researchers began to investigate the psychological consequences of selective deprivation of REM. While the hypothesis that REM deprivation ('dream deprivation') might have a specific deleterious effect on cognition was not confirmed, the experiments did reveal that both REM and NREM sleep were rigorously regulated. When either was prevented by instrumental awakenings, the participants' sleep intensified so as to restore the losses, and it became increasingly difficult to activate the deprivation as it was repeated. The quantitative relationship between the amount of sleep lost and that recovered later indicated that sleep was homeostatically controlled by highly reliable brain mechanisms and that sleep was not only functionally valuable but indispensable to adaptation and to life.

Humans who are subjected to REM, NREM or total sleep deprivation are not merely sleepy; they are also inattentive, emotionally labile, and unable to perform complex analytic tasks. These cognitive symptoms can be seen early in even modest sleep curtailment regimens if the individual is sensitive, middle-aged to aged, or is engaged in intellectually demanding work. Critical monitoring tasks, such as automobile, aircraft, ship or railroad train operation, also require remedial attention. Spectacular anecdotal examples like the *Exxon Valdez* oil spill catastrophe have prompted quantitative epidemiologic studies showing clearly that accident incidence is correlated with sleepiness and substantial loss of vigilance.

When carried to greater lengths, sleep deprivation can within days precipitate hallucinations, delusions, or frank psychosis in humans. In animals deprivation of sleep leads to death within 4 weeks. En route, experimenters observe a breakdown of the integrity of the skin, voracious appetite, and powerfully driven heat-seeking behavior in sleep-deprived rats. These signs of overdriving and exhaustion of the sympathetic nervous system ultimately result in an inability to maintain body weight or body temperature. However, if the deprived animal is then allowed to sleep, even for a few hours, these functions are immediately

restored. If not, they lead to death by sepsis when gastrointestinal bacteria invade the bloodstream and overwhelm weakened immune defenses.

REM SLEEP AND DREAMING

Although the association of dreaming with REM sleep is not exclusive, REM is the sleep phase that provides the optimal conditions for dreaming in its most fully realized forms. It is thus in REM sleep neurophysiology that cognitive scientists can most profitably seek the brain conditions of dreaming. Dream reports have the formal characteristics shown in Table 3. Dreams in REM sleep are mental experiences that are dominated by internally generated percepts (hallucinations) of which the predominant domain is visuomotor, so that dreamers vividly imagine themselves to be moving through a visually detailed world. So vivid are the dream percepts that people rarely recognize that they are not awake but dreaming. The deficient self-reflection leading to this delusional belief is all the more surprising in view of the bizarre cognition, the discontinuities and the incongruities that characterize the hallucinated dream world. Contributing to the delusional acceptance of the dream world as real is the presence of compellingly strong emotions (such as anxiety, elation and anger) and the activation of instinctually organized motor programs (such as sex, fight or flight) that conspire in the elaboration of gripping dream scenarios.

On the cognitive side the deficient self-reflection is compounded by a marked reduction in the prevalence of thinking of any sort and by a loss of memory capacity that makes recognition of the orientational instability of dreams so difficult for us. Our dreams change the times, places and persons of our dream plots without notice, and they may change, even radically, without our noticing it. Our inability to think critically is matched by our inability to evaluate, choose, and execute a plan of action. We do not will dreams; instead, they just happen to us.

Table 3. Formal characteristics of rapid eye movement sleep dreams, as compared with waking

Intensified	*Diminished*
Hallucination	Thought
Bizarreness	Self-reflection
Delusion	Memory
Emotion	Volition
Instinctual activity	

Two important changes in our way of thinking about dream cognition result from this formal analytical approach. The first is the replacement of the simplistic 'REM equals dreaming' model with a more quantitative and statistical statement of probability relating each formal aspect of dreaming to its physiological underpinnings. Thus, for example, the probability that mental activity will be hallucinating increases progressively as the brain moves from wake, through sleep onset, to the NREM phase, and reaches its peak in REM. Conversely, the probability that mental activity will be thought-like progressively and reciprocally declines (Figure 2). The second is a change in the interpretive agenda from an attempt to account for the formal features described above in terms of a content analytic schema (such as Freud's disguise censorship model) with physiologically based models (such as the activation–synthesis model) that attribute the cognitive changes – such as the increased probability of hallucinations and the decreased possibility of thought – to brain physiology, rather than findings of psychology. This change allows the search for the emotional salience or 'meaning' of dreams to proceed in a more straightforward, commonsense manner without the encumbrance of Freud's convoluted psychological theory.

THE NEW NEUROPSYCHOLOGY OF SLEEP AND DREAMING

Complementing the basics of sleep and dream science are data from three methodological innovations that revolutionized the field during the 1990s. The first is the development of home-based data collection methods that allow subjective experience to be quantified around the clock: humans can record the vicissitudes of consciousness that accompany sleep and waking in the natural settings of their lives. This is accomplished by combining the technique of experience sampling with novel sleep monitoring devices which can discriminate between waking, sleep onset, NREM and REM sleep by algorithmic analysis of head and eyelid movement (Figure 3). Using this approach it is possible to study people over many successive days and nights and to obtain thousands of reports of mental activity for analysis in terms of the brain states that underlie them. It is from such studies that the reciprocal relationship between hallucinations and thinking has been discovered, and it is in such data that other formal mental-state features like emotion and bizarreness can be explored in great detail. The simplicity and low cost of the

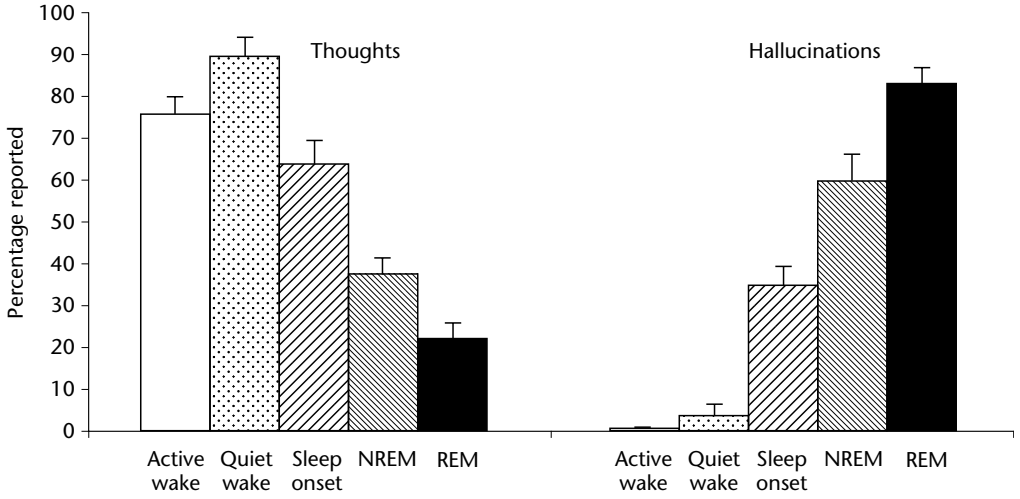


Figure 2. Reciprocal variation in thoughts and hallucinations: percentage of reports in each wake–sleep state.

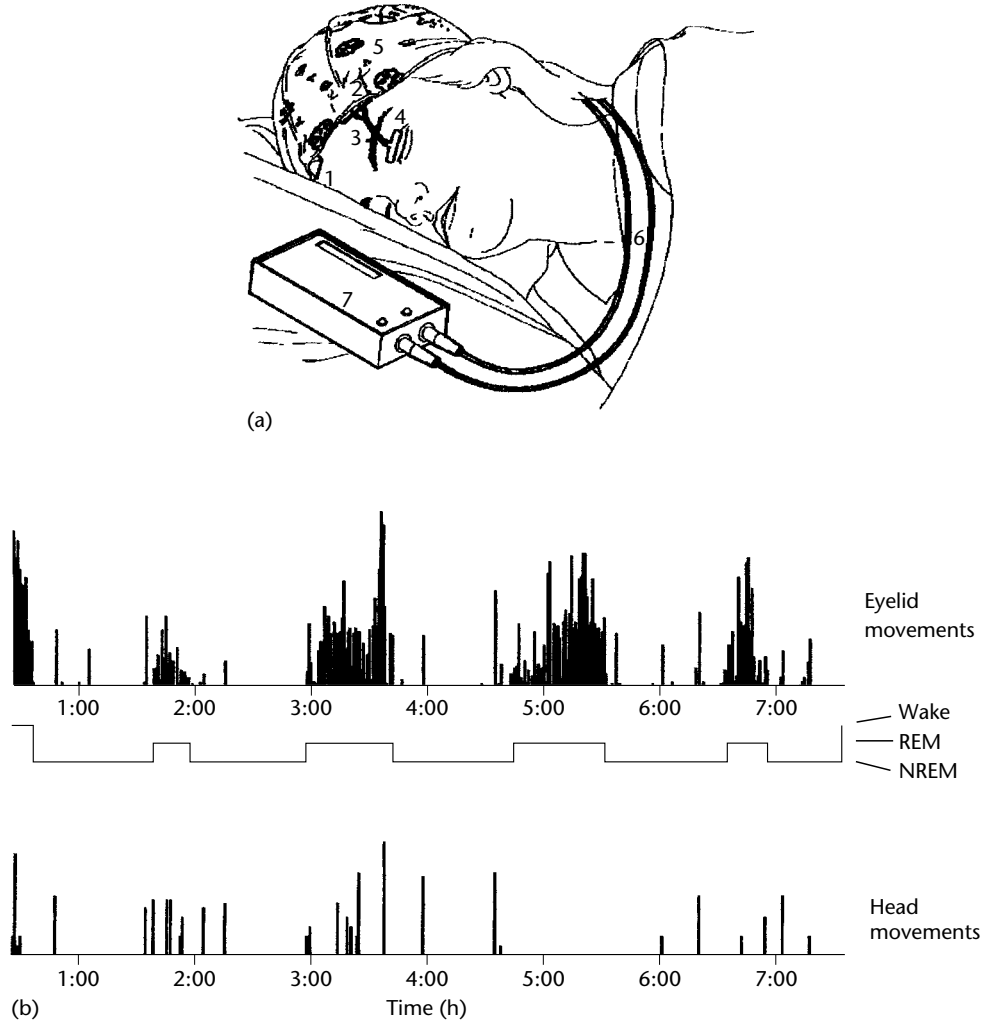


Figure 3. Sleep monitoring device, (a), recording the various stages of sleep through algorithmic analysis of head and eyelid movements, (b). 1, head movement sensor; 2, eyelid movement sensor mount; 3, eyelid sensor lead; 4, eyelid sensor with adhesive backing; 5, bandana (worn 'pirate' style); 6, wires from sensors to recording unit; 7, recording unit.

method puts sleep and dream research in the hands of any cognitive scientist wishing to study sleep–wake phenomena *per se* or the changes in memory or mood that accompany them in normal or experimental populations.

The second methodological innovation is human brain imaging, which reveals the regional differences in activation that are associated with sleep. Data derived from positron emission tomography (PET) is cross-sectional but using functional magnetic resonance imaging (fMRI) it will soon be continuous. The revelations from these images are very informative: the deactivation of the brain reflected in the EEG slowing of NREM sleep is associated with global declines in regional blood flow; the reactivation that is evinced in REM is sharply differentiated, with some structures getting more blood flow, and others less, than in waking.

The third innovation is the report of changes in the dream experiences of patients with regional brain damage. A global loss of dreaming occurs after damage to the parietal operculum or deep frontal white matter, two of the brain regions shown by brain imaging to be selectively activated in REM.

The net result of these three new approaches is a remarkably coherent picture of how the brain determines the formal nature of mental experience. For either waking or dreaming to occur, the forebrain needs to be activated by the brainstem. This activation is regionally and chemically distinct in ways that help us better understand the cognitive differences between waking and dreaming. In waking, the activation is global and aminergic neuromodulation is maximal: as a consequence the waking mind can perceive, attend to, critically evaluate and remember both externally and internally generated data. In REM sleep the activation is selective and the neuromodulation is minimal: the dreaming mind perceives only internally generated data in an unfocused, uncritical, and forgetful way. In NREM the activation is globally reduced to its low point and aminergic modulation declines by half; the deeply sleeping mind is capable of only low to nonexistent levels of information processing.

Functional Significance of Sleep and Dreaming

To what end does the brain undergo such radical changes in sleep? Before considering the mounting evidence that learning and memory are the surprising beneficiaries of their own apparent dissolution in sleep, it is important to stress the evidence that sleep is essential to life: the fact that prolonged sleep deprivation leads inevitably to death has been convincingly demonstrated in rats. En route to their demise, the sleep-deprived animals show defects in such vital physiological functions as body temperature control, dietary energy regulation, body weight control, integrity of the skin, and resistance to infection. These profound and severe defects match the mounting evidence that sleep is an integral aspect of the circadian rhythm controlled by the supracharismatic nucleus, which also regulates body temperature and rest–activity cycles. The circadian clock is intimately linked to other hypothalamic control systems via increasingly well-defined synaptic pathways to and from the neuromodulatory systems of the lower brainstem. In other words, sleep, energy regulation, sexual development and immune responsiveness are all coordinated by shared sets of anatomical circuits and physiological mechanisms in the subcortical brain.

Memory and Learning

Now that the complex and highly organized brain systems uniting sleep with other vital biological processes have been defined, it should come as no surprise to learn that cognition – that most highly adaptive attribute of the mammalian (and especially primate) brain – is also facilitated by sleep. Early work on learning and memory focused on REM sleep but more recent studies have shown that NREM sleep, for all its deficiencies in sustaining ongoing conscious experience, is also involved in the enhancement of learning. In fact, it now appears that learning of information may be a two-stage process by which information acquired in waking is consolidated in NREM sleep and then integrated in REM sleep. A new theory has even proposed that the fluctuating and regionally selective activation and aminergic-cholinergic neuromodulatory processes described earlier underlie transfer of data back and forth from hippocampus to cortex on each successive NREM–REM cycle so that associative memory can be increasingly fine-tuned over the night (Figure 4).

The evidence for this exciting hypothesis comes from four distinct sources. The first is the intrinsic dynamism of hippocampal, cortical and brainstem neuromodulatory activity as revealed by single cell and molecular level microdialysis recording in animal experiments. Although the brain is off-line in sleep, it is in constant action, and engaged in differential communications with itself even in the deepest stages of NREM sleep when conscious experience is at its nadir. This establishes the

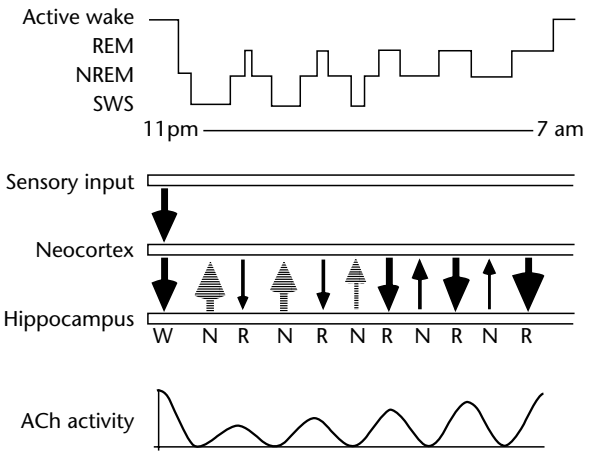


Figure 4. Neuromodulatory activities during sleep; transfer of data back and forth between the cortex and hippocampus during successive sleep cycles is believed to enhance associative memory. ACh, acetylcholine; NREM, non-REM; REM, rapid eye movement; SWS, slow-wave sleep.

plausibility of the model of sleep-learning. The second is the empirical evidence, from both single cell studies in animals and cognitive studies in humans, that recently acquired information (such as place-cell orientation data in rats, or video game-induced imagery in humans) is repetitively replayed in NREM sleep. This data corresponds to the iterative reading out of stored hippocampal data to the cortex. The third is the empirical evidence from human studies that both NREM and REM sleep are essential to optimal performance on tests of visual discrimination learnt to criterion prior to sleep. These data indicate that the two sleep phases serve not just to consolidate learning, but actually to enhance it: participants who slept deeply early in the night (NREM) and remained asleep late in the night (REM) did better when tested in the morning than they did at their peak prior to sleep. The fourth is the empirical evidence from animal studies that training on a variety of tasks leads to a subsequent increase in REM sleep that may persist at intervals from 1–6 weeks after training. Deprivation of REM sleep early in the post-training period may block the improved task performance that emerges later.

CONCLUSION

The hypothesis that sleep facilites cognition finally appears to be on a solid conceptual, neurobiological and empirical footing. By combining experimental techniques with the new neuropsychological approaches, it may soon be possible to enunciate a unified theory of brain–mind status that accounts for both phenomenological and cognitive data with the same set of neurocognitive models.

Further Reading

Aserinsky E and Kleitman N (1953) Regularly occurring periods of ocular motility and concomitant phenomena during sleep. *Science* **118**: 361–375.

Dement W and Kleitman N (1957) The relation of eye movements during sleep to dream activity: an objective method for the study of dreaming. *Journal of Experimental Psychology* **53**: 339–346.

Fosse R, Stickgold R and Hobson JA (2001) Brain mind states: reciprocal variation in thoughts and hallucinations. *Psychological Science* **12**(1): 30–36.

Frederickson CJ and Rechtschaffen A (1978) Effects of sleep deprivation on awakening thresholds and sensory evoked potentials in the rat. *Sleep* **1**(1): 69–82.

Hobson JA and Pace-Schott EF (2002) *Handbook of Neuropsychopharmacology: A Decade of Progress*.

Hobson JA, Pace-Schott EF and Stickgold R (2000) Dreaming and the brain: toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences* **23**(6): 793–842 [discussion 904–1121].

Maquet P (2001) The role of sleep in learning and memory. *Science* **294**(5544): 1048–1052.

Roffwarg HP *et al*. (1962) Dream imagery: relationship to rapid eye movements of sleep. *Archives of General Psychiatry* **7**: 235–238.

Solms M (1997) *The Neuropsychology of Dreams: A Clinico-anatomical Study*. Mahwah, NJ: Lawrence Erlbaum.

Stickgold R, Malia A, Maguire D, Roddenberry D and O'Connor M (2000) Replaying the game: hypnagogic images in normals and amnesiacs. *Science* **290**: 350–353.

Social Cognition

Introductory article

Kimberly A Quinn, Northwestern University, Evanston, Illinois, USA
C Neil Macrae, Dartmouth College, Hanover, New Hampshire, USA
Galen V Bodenhausen, Northwestern University, Evanston, Illinois, USA

CONTENTS

Social cognition refers to the mental processes by which we make sense of our social worlds. The basic claim of the social cognition perspective is that accounting for the complex dynamics of social behavior requires an understanding of the cognitive structures and processes that shape the individual's understanding of the social situation.

INTRODUCTION

Defined broadly, social cognition refers to those aspects of mental processing that are shaped by social interaction, real or imagined, and which in turn influence subsequent social behavior. Defined more narrowly, social cognition refers to a research orientation that employs cognitive principles to analyze and investigate social psychological topics such as social inference, the self, and social perception. Social-cognitive research, with its adherence to the information-processing metaphor, is fundamentally the study of process; that is, social cognition is the part of social psychology that deals with the psychological mechanisms that mediate the individual's responses to the social environment. As such, the nature of mental representation and the dynamics of information processing are central topics of social-cognitive inquiry. (*See* **Information Processing**)

SOCIAL INFORMATION PROCESSING

A basic issue in social cognition research concerns the nature of impression formation. Interpersonal behavior is assumed to be dictated by the character of the impressions that people form of others. The impression-formation process has generally been assumed to proceed as follows. The social perceiver first identifies salient attributes of the target person; then searches memory for category representations that are similar to the detected attributes of the target; selects the most appropriate category representation; uses the content of that representation to draw inferences about the individual; and stores the resultant impression or evaluation in long-term memory. Thus, there is an initial 'bottom-up' or 'data-driven' process in which the features of the target trigger applicable material in memory; 'top-down' or 'theory-driven' processes then guide the perceiver's understanding of the person along particular dimensions. (*See* **Causal Perception, Development of**)

The Mental Representation of Social Information

Social cognition can loosely be broken down into two main elements: the mental structures that are used to represent social information, and the processes that operate on these representations. Broadly speaking, a mental representation is a record of the experienced past that can be constructed, retained in memory, and accessed and used by perceivers in the course of their dealings with others.

Social representations have frequently been assumed to take the form of schemata, associative networks, or prototypes. Although these theoretical viewpoints vary in terms of the internal organization assumed to characterize social knowledge and the degree of interconnectedness thought to exist among these structures, they all share the assumption that social information is represented in the form of general knowledge rather than knowledge of episodes bound to particular times and contexts. These models assume that representations are composed of the individual features that describe the 'typical' category member. That is, there is a generic summary representation of the social target, and the representation that best matches the target is used to guide the

categorization and inference processes. (*See* **Schemas in Psychology**; **Prototype Representations**)

However, evidence exists that people are able to use memory for specific episodes or individuals when making judgments, leading many social-cognitive researchers to endorse an exemplar-based account of knowledge representation. Exemplar representations, in contrast to more generic representational formats, consist of memory traces for specific stimuli or episodes. There is no summary representation for any given collection of targets, and different subsets of exemplar representations can be activated by different targets or contextual cues. According to this viewpoint, stereotypes, for example, do not exist as independently stored knowledge structures but, rather, are created in certain contexts when perceivers summarize the features of a collection of activated exemplars.

The abstraction–exemplar debate within social cognition has taken a similar form to the debate within cognitive psychology, focusing on whether each type of model can account for the effects presumed to be mediated by the other form of representation. More recently, mixed models of representation, in which the perceiver stores both details of specific episodes and generalities across episodes, have been proposed. Some evidence has suggested that the nature of the mental representation of a social group depends on the perceiver's degree of experience with the group in question, such that greater experience is associated with the use of generic knowledge representations (i.e. prototypes). (*See* **Representations, Abstract and Concrete**)

Nonetheless, even with expertise perceivers are able to recruit and use specific exemplars in social judgment, suggesting that the predominance of prototype versus exemplar use in social judgment may not necessarily reflect the manner in which social concepts are represented in memory but, rather, the default processing strategies that are engaged when social perceivers deal with familiar versus unfamiliar targets. Moreover, this general class of models has been criticized for failing to provide a parsimonious account of knowledge representation and memory function. More recently, connectionist (i.e. parallel distributed processing – PDP) accounts of knowledge representation have been proposed as an alternative to the more traditional symbolic approach for understanding social cognition.

Like the more traditional symbolic models, PDP models view representations as networks of interconnected units, and assume that activation spreads along these connections. However, they contend that representations, rather than being discrete, are distributed and superposed: meaning derives from the pattern of activation across many units, and the same set of units can represent different concepts depending on the pattern of activation across the units. This form of representation has been likened metaphorically to a television screen: no pixel has any specific meaning by itself, but by taking on different patterns of illumination the entire array of pixels can constitute a large number of meaningful 'pictures' or representations. In this way, distributed representation seems to be an efficient means of capturing knowledge. The distributed representation is a mechanism that both processes and stores information. This enables both greater context sensitivity and greater storage efficiency. In the area of social cognition, PDP models have been developed to explain stereotype representation, impression formation, and causal explanation.

It is important to note that all of the models described thus far are based on the probabilistic view of categorization, where targets are categorized (and judged) as a function of how similar their attributes are to the features stored in the representation. In recent years, however, a number of cognitive psychologists have argued that the feature-based probabilistic approach is insufficient and that a 'theory-based' approach to concept representation is more fruitful. Proponents of this approach have accrued evidence that categorization depends critically on factors other than similarity matching, and that similarity itself is context-dependent. Perhaps the strongest indictment of the feature-based approach is evidence that, although similarity-based approaches seem appropriate when the perceiver cannot generate an explanation for why a target belongs to a particular category, they do not seem to have strong predictive power when an explanation for category membership is available.

Although few social–cognitive psychologists have explicitly adopted the theory-based approach, evidence in the social psychology literature supports the role of theory and causal explanation in the construction of social representations. (*See* **Similarity**)

Automaticity and Control in Social Cognition

The issue of representational format aside, how does the existence of these knowledge structures influence information processing and behavior? Much of the social psychology research conducted

prior to the mid-1970s assumed, explicitly or implicitly, that people were aware of the cognitive processes underlying their judgments and behaviors and were capable of monitoring and controlling these processes. With the advent of the cognitive revolution, however, evidence emerged to suggest that the perceiver has little introspective access to higher-order cognitive processes and can be completely unaware of the role that various factors play in influencing judgments and preferences. As a result, social psychologists became increasingly interested in processes that occur outside of awareness, thereby evading the perceiver's attempts to understand and control his or her own behavior. The outcome of this revolution was an expansive literature suggesting that much of mental life unfolds in an automatic manner. Our judgments, feelings, and behaviors can be influenced by factors of which we are unaware, by factors of which we were once aware but can now no longer recall, and by factors that we can still recall but whose influence escapes our detection.

Automaticity has been observed in a variety of social judgment domains. For example, it appears that when we observe a person's behavior, we automatically make inferences about the person's underlying traits ('spontaneous trait inferences'). The mere apprehension of people, events, or objects elicits evaluative responses to these targets, and these responses are pre-conscious and automatic. Stereotypic information also appears to be activated automatically, in that it is triggered without the individual's awareness of consent. Recent research has demonstrated that even complex social behaviors can be automatic at times. Nonconscious activation of stereotypes can sometimes lead the individual to behave in accordance with those stereotypes, even if she is not a member of the relevant social category (the 'perception–behavior link'). (*See* **Automaticity**)

That these processes occur automatically is not trivial. Our automatic reactions can guide our decisions and judgments and can influence our thoughts about other people, even if we are not consciously aware of these reactions. Our evaluations of others may be inadvertently influenced by our goals, by our moods, by our stereotypes, by aspects of the situation, and by a multitude of recent experiences, without our recognition that these influences even exist.

Historically, automaticity has been defined in terms of four features: awareness, intentionality, controllability, and efficiency. Unlike controlled processes, automatic processes occur outside awareness, are carried out without intention, are uncontrollable in the sense that we are unable to stop them, and are highly efficient in that they require no attention. Most interesting mental phenomena, however, are of sufficient complexity to be composed of some automatic and some controlled features. Thus, although it appeared in earlier studies that trait inferences were automatic, further research demonstrated that the process could be circumvented by the imposition of a cognitive load, suggesting that trait inferences are conditionally rather than fully automatic. In the domain of stereotyping, some evidence suggests that imposing a processing limitation does not impede the perceiver's categorization of a target into a social group; it does, however, potentially impede the activation of stereotypes associated with that social group. Research on stereotype application suggests further that stereotypes can be automatically inhibited if the stereotypes are at odds with the perceiver's goals. Finally, it appears that the perception–behavior link is moderated by goals: to the extent that the perceiver's goals conflict with the primed concept, the perceiver will not act in accordance with that concept.

One of the most intriguing examples of how a cognitive process can be composed of both automatic and controlled components comes from the theory of 'ironic' mental control. According to this theory, the successful suppression of undesired thoughts requires the conjoint operation of an intentional, controlled search for distracters and an automatic search for the unwanted thoughts (so that they can be suppressed by distracters). When cognitive resources are scarce, the controlled search for distracters is disrupted, but the automatic search for the unwanted thoughts continues unabated, resulting in the hyperaccessibility of the unwanted thoughts. Research has demonstrated the application of this model in the context of stereotyping: actively trying to suppress stereotype use can actually lead the perceiver to rely more on the stereotype than would have been the case in the absence of these suppression attempts.

THE INTERSECTION OF MOTIVATION AND COGNITION

Not surprisingly, how we feel and what we desire can color our judgments. Beyond the level of mere knowledge activation, most cognitive activity is goal-dependent; that is, it is initiated by a perceived discrepancy between an actual and a desired state. Motivation can influence social cognition in a number of important ways. Motivational factors can determine the degree of cognitive effort

expended to process relevant information, as well as the direction that the process takes. Motivation can affect the direction of processing by facilitating the activation of goal-relevant cognitive categories, in a sense determining the 'theories' that are applicable for interpreting the available data. Finally, motivation can also affect the extent of information processing, based on how important the perceiver's goals are and how much cognitive effort (e.g. elaboration, distortion, inconsistency resolution) is required to make the current situation match the desired situation.

Cognition, of course, can also influence motivation. Motivation has a cognitive aspect, in that goals may be thought of as knowledge structures, governed by the same processes and mechanisms that govern other cognitive structures. Cognitive capacity, for example, constrains the extent to which motivation can exert its influence: to the extent that the perceiver can draw on all of her cognitive resources, motivation will have stronger qualitative and quantitative influences on processing; however, to the extent that the perceiver's resources are depleted – by virtue of distraction, anxiety, circadian rhythms, and so on – she will be less able to control both the direction and the magnitude of processing.

Affect and Cognition

The intersection of motivation and social cognition has been most clearly represented by theory and research on affect and cognition. This research has yielded several findings that suggest that encoding, elaboration, and judgment are mediated by the recall of mood-congruent information stored in memory. Ambiguous information tends to be encoded in terms of concepts of the same valence as the perceiver's current mood. Inferences that perceivers draw are often matched in valence to their current mood. Moreover, perceivers are more likely to remember information if it is affectively matched to their current mood state.

In terms of the relation between affect and cognition, three general frameworks have been proposed. The first approach views affect as an emotional state and adopts a functionalist approach to the emotion–cognition relation. Proponents of this framework assert that to make predictions regarding the direction and magnitude of cognitive activity, it is necessary to consider the adaptive significance of the emotion in question. For example, it appears that happiness, which signals to the perceiver that all is well, leads to a decline in processing activity (unless such activity is intrinsically enjoyable) – presumably because the perceiver either feels no need to engage in deep processing or does not want to risk the decline in mood that could accompany such effort. Sadness, in contrast, presumably signals to the perceiver that something is amiss. Sad perceivers tend to engage in deeper processing, perhaps to distract themselves from, or to find a remedy for, their emotional state. Interestingly, these patterns translate into greater stereotyping by happy perceivers and less stereotyping by sad perceivers, relative to perceivers in a neutral mood. High-arousal emotions such as anger and anxiety also lead to greater stereotyping; in this case, however, this appears to be a function of the capacity-diminishing nature of arousal, rather than of any appraisals of emotional significance.

The second framework is exemplified by the 'mood as information' view of affect. Proponents of this approach suggest that feelings may serve informative functions, and that perceivers use their apparent affective responses to targets as a source of information in evaluating those targets. The impact of feelings on these evaluative judgments depends on their perceived informational value: to the extent that affective reactions seem to offer relevant information, they will influence evaluations of the target; however, to the extent that affective reactions are deemed irrelevant, they will not be used as a basis for judgment.

Unlike the frameworks in which affective information can become linked to the cognitive representation of some target, a third framework proposes that emotions themselves provide a framework within which targets may be categorized and represented. That is, emotions do not simply activate congruent information; rather, they actually lead individuals to reorganize conceptual space according to emotional equivalences. This reorganization then determines how people perceive similarities and differences among objects and events and how they respond to them. Thus, emotions affect not only memory but also category construction and use. Emotional response categorization is assumed to be functional: a category of things that have elicited a particular emotion enhances the perceiver's understanding of the meaning of that experience in terms of his or her own personal learning history. This, in turn, facilitates the perceiver's ability to imagine the consequences of reactions to new objects.

THE SELF

Social cognition theorists assume that social behavior is mediated not only by mental representations

of others, but also by actors' currently active representations of themselves. Much of the research in social cognition that concerns the analysis of the self has focused on the person's mental representation of his or her own personality attributes, social roles, past experiences, and future goals, and how these representations influence social inference and social judgment. (*See* **Self, Psychology of**)

Self-knowledge

People differ in which attributes they consider central and self-defining. For each of their most central attributes, individuals may develop elaborate self-schemata (i.e. integrated sets of memories and beliefs about their relevant behaviors). Although people have many stable and enduring memories about the self, their working self-concepts (that is, their sense of self at a given moment) vary from one occasion to another, as different subsets of self-knowledge become activated. People are 'self-schematic' on dimensions that are important to them, on which they think of themselves as extreme, and on which they are certain that the opposite is not true.

Information pertaining to the self has implications for both self- and other-perception. People demonstrate a self-reference effect, such that information relating to the self is processed more thoroughly and deeply, and hence is remembered better, than other information. People who are schematic on a given trait can make judgments about their standing on that trait very rapidly, can back up these judgments with extensive personal examples, and are reluctant to accept evidence that questions these self-views. They also possess more general expertise about this trait, which they draw upon to make sense of others' behavior. We often evaluate other people by comparing their behaviors and traits to our own.

The context-specificity and flexibility of the self-concept has also been a topic of interest to social cognitive researchers in the domain of intergroup relations. Researchers in the self-categorization tradition, for example, have addressed how the self is shaped by the social context. These researchers have argued that the self is not a fixed mental structure; rather, it is viewed as the expression of a dynamic process of social judgment. Thus, self-perception and self-definition do not reflect the activation of preformed self-concepts but, rather, a flexible, constructive process of judgment in which varying self-concepts are constructed to fit the perceiver's relationship to the current social environment. These self-concepts have implications for the perceiver's inferences about other individuals and social categories: how perceivers define themselves – in relation to the other individuals or groups present in the current situation – affects the goals, beliefs, and expectancies that they bring to the situation.

Self-regulation

Although the goal of the self-categorization research was to make the case that the content of the self could vary as a function of the intergroup context, the perspective also highlighted the fact that self-concept flexibility has functional utility. Having a concept of one's self – and especially a flexible conception of one's self – is integral to social functioning. It permits the perceiver to relate to people and to be an active agent and decision-maker. That is, the self-concept does not merely provide the person with self-knowledge, it also allows for self-regulation.

In recognition of the functional utility of the self, recent social cognitive research has turned to investigations of the 'executive function' of the self. The self-concept summarizes information about oneself as an object in the world in order to serve self-regulatory functions. The 'self-digest' summarizes a person's relations to his or her world and the personal consequences of these relations. In this framework, knowledge about oneself as an object in the world is represented to the extent that it is functional in self-regulation, in agentic decision-making and behavior. The self-digest, then, helps the person fulfill needs and achieve goals when interacting with the world.

The Self as a Nonprivileged Concept

Whether or not the self merits the status of a privileged concept has been a matter of debate, largely stimulated by the phenomenal experience of the self and of self-relevant information as 'special'. Most recent social–cognitive theorizing on the self, however, has accorded it no privileged status, arguing that the extensive processing associated with self-relevant information is due to the self being a highly familiar and well-organized body of knowledge. The self-concept may be the most central, the most important, and the most complex concept available to the person, but the processes through which it is developed and through which it exerts its influence have been regarded, to date, as largely the same processes involved in the representation and use of other social (and perhaps non-social) concepts.

PERCEIVING PERSONS AND GROUPS

Perhaps the most central topic to the field of social cognition is that of impression formation, the process by which the perceiver integrates information about and evaluates target individuals. Indeed, the impetus for virtually all social–cognitive research on memory and information processing stems from interest in understanding how the social perceiver makes sense of others.

Person Perception and Impression Formation

The process of impression formation has been debated since the inception of social psychology. Early theorists assumed that the full range of information known about the target individual was integrated into one's impression of that person. From a Gestalt perspective, the perceiver was assumed to merge the diverse features of the target person into a coherent, unitary impression that took into account the meaning of individual features as well as their interrelationships. From an elemental perspective, the perceiver was assumed to assess the implications of each piece of information about the target person and then combine them algebraically into a summary impression.

More recently, models of impression formation have distinguished between top-down and bottom-up processes. These newer approaches have assumed that it is necessary to distinguish between the influences of stereotypic information on the one hand, and attribute-based or individuating information on the other. Two such models – the dual-process model and the continuum model – have received particular attention. Although the two approaches differ in the extent to which they allow for stereotypic and individuated processing to operate in tandem, both assume that perceivers first engage in stereotype-based processing and then, depending on motivation and ability, correct their impressions on the basis of individuating information. Moreover, both assume that the use of stereotypic and individuating information involves fundamentally different processes.

Recent approaches have criticized these influential models. A 'parallel constraint satisfaction' model of impression formation has been proposed and has postulated that stereotypic and individuating information are processed simultaneously and given equal weight in the impression-formation process (within certain limiting conditions). More recently, other criticisms have noted the possibility that individuated impressions may rely on a conjunction of sterotypic and idiosyncratic information, and that reliance on stereotypes may actually facilitate the perceiver's ability to simultaneously process individuating information.

Spontaneous trait inferences and effortful attributional analysis

People tend to make trait inferences spontaneously when they observe trait-relevant behaviors, even when they have no explicit intention of doing so. The term 'correspondent inferences' was coined to refer to the tendency of social perceivers to infer that observed behaviors correspond to underlying traits. Early work tended to assume that these correspondent inferences were, in fact, dispositional inferences; that is, they assumed that perceivers spontaneously infer that an actor's behavior was indicative of an underlying personality. More recent research, however, has challenged this view. At this point, it remains unclear as to whether these spontaneous trait attributions are in fact dispositional inferences. Empirical evidence suggests that perceivers do spontaneously (automatically) generate inferences regarding the trait meaning of observed behaviors, but that they do not necessarily generalize from this inference to beliefs about the stable disposition of the actor. It appears that the perceiver automatically infers the trait meaning of observed behaviors; more controlled processes then either encourage or discourage the perceiver from making judgments about the actor's chronic disposition.

Social Categorization and Stereotyping

The mere categorization of individuals into social groups initiates cognitive processes that function to promote the perception of within-group similarities and between-group differences. The most effective of these processes is undoubtedly the activation and application of stereotypes. (*See* **Stereotypes**)

Stereotype activation

By endorsing the view that semantic priming is an inevitable consequence of mere apprehension of a stimulus in the environment, social psychologists have concluded that stereotype activation must be an unconditionally automatic process. Indeed, ample evidence has emerged to suggest that once the target's group membership has been identified, the relevant stereotype is activated without intent or awareness.

As noted earlier, however, very few processes satisfy the criteria for unconditional automaticity.

Research on stereotype activation in recent years has begun to accumulate evidence that the process is only conditionally automatic. Mere exposure to a stereotyped target, then, may be insufficient to trigger category activation.

Two factors appear to moderate the activation of stereotypes: processing goals and attitudes. Goal states can function not only to interfere with stereotype activation, but also to promote stereotype application. Recent empirical work, for example, has demonstrated that participants who are motivated to view a target in a particular way are able to simultaneously activate the stereotype that favors their desired impression and inhibit the stereotype that contradicts that impression, and that these processes occur spontaneously.

Perceivers' chronic beliefs about social groups also appear to moderate the activation of categorical thinking, a finding that is at odds with conventional thinking on the dynamics of the categorization process. Until relatively recently, it has been widely accepted that both prejudiced and egalitarian individuals activate stereotypes to the same degree when they encounter members of stereotyped social groups. In fact, empirical evidence now demonstrates that egalitarians display little or no evidence of stereotype activation when presented with categorical priming stimuli. These findings suggest that stereotype activation, rather than being fully automatic, is a conditionally automatic process.

Stereotype application

Stereotype application can take two forms. First, stereotypes can serve as frameworks for the assimilation and integration of expectancy-consistent information, leading the perceiver to emphasize stereotype-consistent information to a greater extent than he or she would have in the absence of categorical information. At the same time, stereotypes can also sensitize the perceiver to unexpected information, leading to a greater emphasis on stereotype-inconsistent information following stereotype activation.

A functional analysis of stereotyping suggests that the perceiver can accrue benefits from the application of stereotypes. Stereotype-based expectancies provide a framework that facilitates the identification and comprehension of consistent information, such that the processing of that information requires little deliberative attention or thought. The fluency of expectancy-consistent information means that substantial attention may be redirected to other concurrent tasks, including the encoding of inconsistent information. The benefits of stereotype application in demanding environments are thus twofold: first, expectancy-consistent information can be processed in a relatively effortless manner; second, remaining attentional resources can be redirected to unexpected information, enabling the perceiver to process and remember this potentially important individuating information. (*See* **Attention**)

A variety of motivational factors also seem important to stereotype application. The use of stereotypes can be overridden by accuracy motivation, but can be enhanced by ego-defensive motivations by providing a basis for downward social comparisons. 'Social judgability' concerns also play a role, and perceivers are unlikely to report stereotypic judgments unless they believe there is a legitimate informational basis for such a judgment.

Stereotype suppression

As we have already discussed, several studies have documented the ironic consequences of stereotype suppression for perceivers' evaluations of, memory for, and behavior towards stereotyped targets. Notwithstanding these demonstrations, doubt remains over the generality of these effects. Low-prejudice participants, for example, are apparently not susceptible to 'rebound' effects. It also seems that perceivers may be more consistent in their efforts to avoid stereotyping for sensitive social groups, thereby preventing the emergence of rebound effects. This suggests that stereotype suppression can be effective to the extent that perceivers are motivated by concerns of egalitarianism. (*See* **Thought Suppression and Mental Control**)

Entitativity: Perceiving Persons and Groups

Although research on both impression formation and stereotyping have long and rich histories within social cognition, little research has been directed towards understanding how the two processes might be similar or different. In both cases, research is concerned with how a perceiver comes to develop a conception of a social target, either a person or a group. But do the same mechanisms and processes govern social perception in these two domains?

Recent research on 'entitativity', or the extent to which a target is seen as coherent and unified, suggests that default expectancies for individuals versus groups has implications for how the perceiver forms impressions of individuals and develops conceptions of groups. The fundamental postulate of this research is that the social perceiver

assumes unity in the personalities of others: persons are seen as coherent entities, and the perceiver's impression of a target person should reflect that unity and coherence. In general, however, perceivers do not expect the same degree of unity and coherence among members of a group as they expect in the personality of an individual.

As a result of holding this assumption, the perceiver seeks to draw inferences about the dispositional properties constituting the core of the person's personality, but not about the 'disposition' of the group. The result is that the perceiver draws inferences quickly when the target is a person, but not so rapidly when the target is a group. The expectation of consistency in the traits and behaviors of individuals, rather than groups, also leads the perceiver to strive to resolve inconsistencies in the information acquired about the target person, but to tolerate inconsistencies in the information acquired about different group members.

This is not to suggest that processing information about individuals will necessarily be different from processing information about groups. Although research on entitativity does indeed demonstrate that the default assumption is that individuals form more coherent entities than do groups, groups can also vary in their perceived entitativity. Empirical evidence exists to suggest that perceivers process information about groups and individuals in a similar manner when those groups are perceived to be high in entitativity. The nature of the social target, person or group, then, is not the crucial element in determining the impression-formation process. When expectancies of unity, consistency, and coherence are controlled or equated, the processes and outcome of impression formation are very similar for individual and group targets.

CONCLUSION

Any attempt to define and summarize a broad domain must necessarily be non-exhaustive and this article is no exception. Social cognition, with its broad mandate to investigate the mental processes that mediate relations between the individual and his or her social world, encompasses not only the topics reviewed here, but also many other mental processes central to social functioning, such as goal-setting, decision-making, and heuristics and biases in social judgment. Also important are the emerging field of implicit social cognition and recent work on the diversity of the processes (e.g., perceptual and conceptual) believed to characterize the complexity and flexibility of the social perceiver's mental life. Ultimately, the goal of social cognition is to explain how all of these processes interact to determine social behavior.

Further Reading

Chaiken S and Trope Y (1999) *Dual-process Theories in Social Psychology*. New York, NY: Guilford Press.

Fiske ST (1998) Stereotyping, prejudice, and discrimination. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*, 4th edn, vol. 2, pp. 357–411. New York, NY: McGraw-Hill.

Forgas JP (2001) *Handbook of Affect and Social Cognition*. Mahwah, NJ: Lawrence Erlbaum.

Gilbert DT (1998) Ordinary personology. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*, 4th edn, vol. 2, pp. 89–150. New York, NY: McGraw-Hill.

Kruglanski AW (1996) Motivated social cognition: principles of the interface. In: Higgins ET and Kruglanski AW (eds) *Social Psychology: Handbook of Basic Principles*, pp. 493–522. New York, NY: Guilford Press.

Kunda Z (1999) *Social Cognition: Making Sense of Others*. Cambridge, MA: MIT Press.

Macrae CN and Bodenhausen GV (2000) Social cognition: thinking categorically about others. *Annual Review of Psychology* **51**: 93–120.

Sherman JW (2001) The dynamic relationship between stereotype efficiency and mental representation. In: Moskowitz GB (ed.) *Cognitive Social Psychology: The Princeton Symposium on the Legacy and Future of Social Cognition*, pp. 177–190. Hillsdale, NJ: Lawrence Erlbaum.

Smith ER (1998) Mental representation and memory. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*, 4th edn, vol. 1, pp. 391–445. New York, NY: McGraw-Hill.

Wegner DM and Bargh JA (1998) Control and automaticity in social life. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*, 4th edn, vol. 1, pp. 446–496. New York, NY: McGraw-Hill.

Social Processes, Computational Models of

Introductory article

Andrzej Nowak, University of Warsaw, Warsaw, Poland
Robin R Vallacher, Florida Atlantic University, Boca Raton, Florida, USA

CONTENTS

Computational models of social processes formalize social processes in terms of computer algorithms. Such models, commonly investigated by means of computer simulations, provide a means by which theories of social processes can be precisely expressed, and they enable investigators to study non-obvious consequences of the assumptions underlying social theories.

INTRODUCTION

In the 1970s, social scientists began to conceptualize social processes in terms of computer algorithms, and to investigate the consequences of these computational models using computer simulations. Computational models have since become an important component of social science research. They can reveal properties that cannot be investigated by analytical (e.g. traditional mathematical) means. Computational models provide a precise formalism for social processes and enable one to test a theory's internal consistency and completeness. The computational approach, then, has added considerable precision and rigor to social science theory and research. It has also proven relevant to the solution of real-world social problems.

Computational models were first used to study large-scale social processes (pertaining to groups and societies). These models are formulated as sets of variables describing the basic properties of the modeled system. Mathematical equations specify how each variable changes in time as a function of other variables. Computer simulations are then used to study these changes. This approach is especially prevalent in the study of economic and demographic processes. However, the exclusive focus on group-level properties does not enable one to study the relationships between different scales, from individual agents through groups and organizations to entire societies. Most computational models now attempt to relate individual agents and aggregate social entities (e.g. social groups).

BUILDING SOCIETIES FROM INDIVIDUALS

To investigate how individual-level processes give rise to group-level properties, computational models specify a set of individuals, referred to as agents, and a set of rules describing the relations (e.g. mutual influence, interdependence) among the agents (and sometimes between agents and the environment). Simulation models differ with respect to the representation of agents and the rules of interaction. Some models employ very simple representations of individual agents: in the extreme, an agent may be represented by a single variable representing a single characteristic. The idea is to identify the minimal assumptions necessary to produce specific group-level properties.

For example, in modeling the emergence of public opinion, individuals are represented as cells on a two-dimensional lattice. Each individual has one of two opinions ('pro' or 'con') on a particular issue. Individuals also differ in the strength of their advocation of their opinion. Before interaction, each individual's opinion is unrelated to that of his or her neighbors. In each 'round', the individual assesses the opinions of his or her neighbors, giving greatest weight to the opinions of the strongest advocates. If this assessment leads to the conclusion that an opinion contrary to his or her own is the prevailing opinion, the individual will adopt this opinion and advocate it in the next round. Individuals maintain the same physical location throughout the simulation. Figure 1 shows

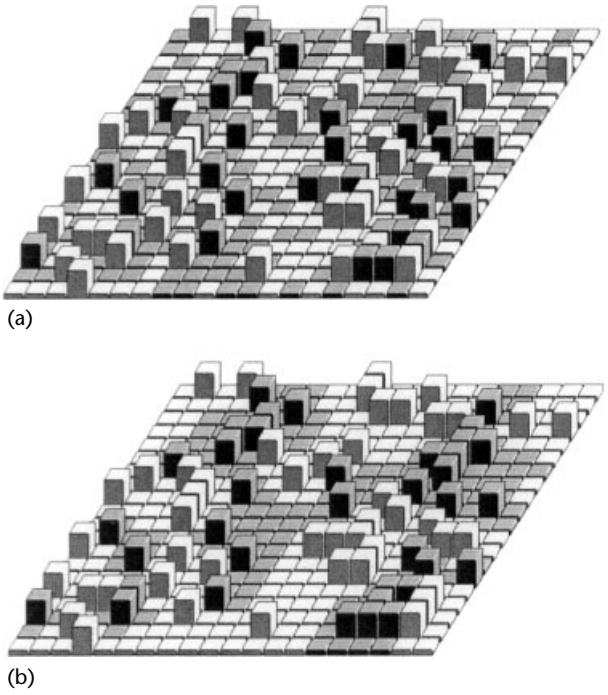


Figure 1. A simple model of emergence of public opinion. The tint of a cell denotes the individual's opinion ('pro' versus 'con'), and the height of a cell represents the strength of the individual's advocacy of his or her opinion. (a) Before running the simulation (random distribution of opinions). (b) After several iterations (polarization and clustering emerge).

the distribution of opinions in a social group before and after several rounds of such interaction. Note that the initial minority opinion declines in popularity and clusters spatially into like-minded groups. Thus, this simple simulation rule defining influence on an individual level causes the emergence of two basic group-level phenomena, polarization and clustering.

Other computational models, employing the tools of artificial intelligence (AI), take a more complex view of agents. Each agent has a repertoire of actions, and may also have other characteristics (e.g. knowledge, goals, values). The simulation rules specify how agents' actions and characteristics depend on those of other agents. The computer simulations investigate the consequences of these rules as the individuals interact with each other and with the environment. Agents are considered to be adaptive because they can improve their 'fitness' in the process of learning or evolution. Fitness typically refers to an agent's ability to maximize his or her resources (e.g. to acquire capital) in a social dilemma situation. In variations of this approach, agents may represent higher-level entities, such as social groups, organizations, or societies.

DYNAMICS OF SOCIAL INTERACTION

In computational models, social interactions are modeled as rules specifying how agents influence one another's states and actions, and the environment in which they are embedded. The dynamic behavior of the system is produced by repeated iteration of the interaction rules. After each iteration, the state of each agent is updated according to the model's rules. These updated states, in turn, provide the input for the next iteration of the program, and so on. For example, an individual may change his or her opinion in a given simulation round as a result of social influence, and subsequently influence others to adopt the new opinion.

Computational models also specify the pattern of social interactions. Although some models assume interactions among all agents, most assume that each agent interacts only with a specified subset of other agents. In models defined in terms of intelligent agents, the subsets can change in accordance with the nature of the interaction. Thus, one type of information might be transmitted only to agents with whom the agent is linked in an organizational network (direct communication), while another type of information may be made available to all other agents (broadcasting). In models that concentrate on the evolution of social relationships, the structure of social interactions is open to change.

There are two main approaches to modeling the patterns of interaction among agents. In the spatial approach, each agent resides in a particular location in space, usually a two-dimensional grid, representing the spatial patterning of social interactions (see Figure 1). Each agent interacts only with other agents in nearby locations (e.g. adjacent cells). Social relations are therefore represented as spatial proximity, and changes in social relations are represented as spatial movement. The spatial approach allows the visualization of patterns on the global level that develop as a result of local interactions.

In the network approach, social relations are conceptualized as links, representing friendship patterns, communication networks, organizational structure, and other patterns of interpersonal relations. Some network models assume that a single network of links defines all interactions among the agents. In other models, different networks may define different forms of interaction. Thus, informal communication (e.g. gossip) may take place through one network, while formal communication

(e.g. the flow of documents in an organization) may take place through a different network. The dynamics of social relations are represented by changes in the links among agents.

In both spatial and network models, the dynamics of social interaction may be represented by changes in the properties and actions of individual agents, and by changes in the interaction patterns among agents.

EMERGENT SOCIAL PROCESSES

The new theory of complexity has shown that models composed of very simple interacting elements can display highly complex behavior on the system level. The low-level interactions are often nonlinear, so that their consequences cannot be derived by analytical means. Computational models involving multiple interacting agents are complex adaptive systems. They have been fruitfully applied to emergent phenomena in diverse areas of science. In these models, behavior at the aggregate level is not directly programmed into the model, but rather emerges as a consequence of interactions between agents. These models exhibit self-organization, since regularities, patterns and complex properties are produced by interactions among system elements without the supervision of higher-order agents. For example, in a cellular automaton such as that illustrated in Figure 1, spatially coherent groups of like-minded individuals emerge from social influence rules specified on the level of individuals. The group-level phenomena (polarization and clustering of opinions) were not programmed into the model, but rather emerged from the local interactions among agents.

Two variations of this general approach have been employed in the social sciences. In one, individuals' characteristics change as a result of 'updating' rules. This type of model is useful for understanding changes in attitudes and opinions as a result of social interaction. The emergence of public opinion illustrated in Figure 1 exemplifies this approach.

In 'migration' models, on the other hand, individuals' characteristics do not change, but individuals may change their physical location. For example, in modeling social segregation, if the individual is surrounded by a local majority of individuals who are different in a relevant characteristic (e.g. race), he or she will move to a different spatial location. Migration models are used to investigate the emergence of spatial patterns on the basis of stable values and preferences.

In both of those approaches, iteration of the updating rules often produces regularities and patterns at the group level that were not directly programmed into the agents.

EVOLUTION OF COOPERATION

One of the most interesting examples of emergence in the social sciences concerns the evolution of cooperation among agents who are motivated to maximize their personal gain. The 'prisoner's dilemma' game is often used in simulations of this process. In each round, each agent chooses whether to 'cooperate' or 'defect' in his or her interactions with another agent. This choice is regulated by individual strategies, which evolve by means of genetic algorithms. Each agent begins with a random sequence of genes that generates choices reflecting some strategy. By analogy with natural selection, those individuals who acquire the least profit in their interactions with many others are eliminated from the population, while the rest are allowed to produce offspring. New combinations of genes are produced through mutations and the exchange (crossover) of genes between individuals.

Initially, there is usually a sharp increase in the frequency of defection, since defection produces higher pay-offs regardless of the choice made by an interaction partner. After some time, however, the frequency of cooperative choices increases and becomes prevalent. Although cooperative choices produce lower outcomes in any single interaction, in the long run this strategy may induce cooperation in one's partner, yielding higher profits for both. Defection strategies, on the other hand, are maladaptive in the long run because they turn partners into defectors. The elimination of such strategies corresponds with a growth in the proportion of strategies capable of both cooperation and, when faced with defection, retaliation.

SIMULATING ORGANIZATIONS

Unlike informal groups, organizations attempt to achieve specific objectives and are characterized by a better-defined structure of interactions among agents. The agents within organizations perform distinct roles defined in terms of tasks and decision-making. Each agent is capable of performing a set of actions, such as requesting information, preparing a report, helping others, wasting time, and so forth. These actions are performed in response to a specific set of conditions, such as input from other agents (e.g. orders, information, requests for help) or the appearance of new tasks. Agents are usually

heterogeneous with respect to possible actions, knowledge, and goals. Organizational dynamics thus reflect the interactions among intelligent autonomous agents, who are modeled using the tools of AI.

The computational approach is useful in investigating the mechanisms of known organizational phenomena (e.g. poor decision-making, development of trust, team effectiveness, group cooperation). It has also been employed to search for optimal organizational structures and procedures. Distributed AI models, for example, are useful in modeling effective group decision-making. Each agent is represented as a goal-oriented problem-solver (e.g. using the SOAR model), with specific knowledge. Agents can share data, judgments, and decisions. Different team decision schemes are tested for their effectiveness in reaching an optimal decision. Computational models have proven to have considerable practical value, and are used to analyze existing organizations and provide advice for the improvement of organizations.

RELEVANCE OF MODELS OF SOCIAL ORGANIZATION

Computational models have contributed to our understanding of pressing societal concerns. For example, models in which agents play the prisoner's dilemma game have generated insights into the emergence and resolution of conflict at different scales of social reality (e.g. between individuals, groups, and nations). This approach has identified the different conditions that promote cooperation and competition. It has informed policy makers about effective strategies for resolving conflicts of interest involving limited resources, and about situations that can lead to the escalation of conflicts (e.g. spiraling retaliation). This approach has also proven useful in modeling the formation of solidarity networks and various economic processes.

Models of social influence based on cellular automata have revealed the mechanisms responsible for the formation of public opinion, social change, the emergence of ideologies, and the spread of innovations. Models based on intelligent agents have proven especially relevant to issues in organization function and design. This approach has also been used to test whether hypothesized patterns of social life match the patterns known from archeological artifacts in past societies.

Computational models also play a prominent role in testing military strategies. They have proven useful in resolving Cold War and post-Cold War conflicts.

SUMMARY

Computational models of social processes have become a vital tool in social science theory and research. Social processes are represented as computer algorithms and investigated with computer simulations. Computational models provide insight into the emergence of aggregate-level properties from simple rules specifying the interactions among lower-level elements of a social system. These models are constructed as assemblies of interacting agents, which can be defined at different scales (e.g. individuals, groups, organizations, societies). The computational approach has generated insight into the nature of diverse phenomena in social science, and has been successfully applied to practical issues concerning interpersonal and inter-group dynamics.

Further Reading

Axelrod R (1997) *The Complexity of Cooperation: Agent-Based Models of Competition and Cooperation*. Princeton, NJ: Princeton University Press.

Axelrod R and Cohen MD (1999) *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. New York, NY: Free Press.

Epstein J and Axtell R (1997) *Growing Artificial Societies*. Cambridge, MA: MIT Press.

Gilbert N and Troitzsch KG (1999) *Simulation for the Social Scientist*. Buckingham, UK: Open University Press.

Kohler TA and Gumerman GJ (eds) (2000) *Dynamics in Human and Primate Societies: Agent Based Modelling of Social and Spatial Processes*. Oxford: Oxford University Press.

Liebrand WBG, Nowak A and Heselmann R (eds) (1998) *Computer Modelling of Social Processes*. London: Sage.

Nowak A and Vallacher RR (1998) *Dynamical Social Psychology*. New York, NY: Guilford Press.

Prietula M, Carley KM and Gasser L (eds) (1998) *Simulating Organizations: Computational Models of Institutions and Groups*. Menlo Park, CA/Cambridge, MA: AAAI Press/MIT Press.

Sterman J (2000) *Business Dynamics: Systems Thinking and Modelling for a Complex World*. New York, NY: McGraw-Hill.

Space Perception, Development of

Intermediate article

Albert Yonas, University of Minnesota, Minneapolis, Minnesota, USA

CONTENTS

The development of space perception refers to the changing processes that take place within an organism that make possible the control of action based on knowledge of the three-dimensional layout of the environment. Development refers to both the consequences of experience and the maturation of sensory and perceptual mechanisms.

INTRODUCTION

Human infants, unlike more precocious species, develop the ability to use vision to perceive the layout of the external environment over a period of months. Over this period, sensitivity to sensory properties, such as contrast and orientation, also improves as neural mechanisms in the eye and brain develop. This article focuses on changes in perception of the external world rather than on the development of sensory processes. While mechanisms early in the visual system respond to the orientation of contours and make it possible to detect that contours converge to a vanishing point, the detection of perspective convergence on the retina does not mean that the viewer perceives that the region in which the contours converge is more distant. For depth to be perceived as such, a higher-level perceptional process is needed.

There are many cues that provide independent information for depth. They can be placed in three classes: motion-carried information, binocular information and static monocular or pictorial information. In humans, the ability to perceive the layout of the environment develops in stages as responsiveness to motion-carried, binocular, and static monocular cues develops. There is evidence that newborns use retinal motion produced when they move their heads to perceive depth. By one month of age, infants respond to optical expansion information for impending collision, and by two months they act as if they perceive some aspects of object shape from motion. Perception of depth from binocular cues has been demonstrated in four-month-olds. Perception of spatial layout from static monocular cues appears to develop last and has only been found in infants over six months of age.

Although efforts to describe depth cues have a long history, and the list of known depth cues may seem extensive, the task is not finished. It is likely that in the future researchers will discover new cues for depth that are available in static and moving displays. Only then will research on the development of sensitivity to those cues be possible.

INFERRING DEPTH PERCEPTION FROM INFANTS' RESPONSES

Certain spatial behaviors, such as crawling (which can reveal that infants avoid moving over the 'deep side' of a visual cliff (Walk and Gibson, 1961)) and precise reaching, are excellent indicators of depth perception. However, human infants are rarely able to crawl until 7 months of age, and do not begin to reach with accuracy until they are about 5 months old, so these behaviors cannot be used as measures of depth perception with younger infants. Other responses, such as upward head rotation (which occurs in the first month of life), do not have the specifically spatial character of crawling and are open to multiple interpretations. The behavior that is most commonly used in infant research is preferential looking at one display rather than another. Looking behavior can tell us about the infant's ability to discriminate between the displays on a sensory level, but it does not necessarily indicate whether the infant perceives differences in the spatial meaning specified by the displays.

SENSITIVITY TO MOTION-CARRIED INFORMATION FOR DEPTH

Motion-carried information for depth can be obtained when observers move their heads or

bodies, or when a viewed object moves. Such information is sometimes called 'motion parallax'.

A. Slater has suggested that motion parallax gives newborn infants some degree of size and shape constancy (Slater *et al.*, 1990). Size constancy is the ability to perceive the unchanging physical size of an object although its retinal size changes when its distance is changed. Shape constancy is the ability to perceive the unchanging physical shape of an object when its orientation is changed. It is also possible that the angle between the two eyes is the source of the depth information that underlies size and shape constancy in newborns (Kellman and Arterberry, 1998).

Optical Expansion and Contraction: Perceiving Impending Collision

When an object approaches a perceiver, it projects an expanding flow pattern that serves as information for its approach. When the approach is constant in speed, the rate of retinal expansion increases and is explosive at the last moment, indicating impending collision (Schiff, 1965). Retinal contraction is information for withdrawal. Initial research found that 1-month-olds defend against collision by moving their heads back and bringing their hands between their face and the object when it is approaching them, whether it is real or simulated on a screen. Another possible explanation for these actions is that infants track the rising contour of the expanding pattern and as a result rotate their heads upward. When the head rotates upward and control of head posture is lost, the head falls backward and the young infant's arms are raised. This latter interpretation was supported by studies that varied upward optical motion and information for collision (Yonas, 1981).

Researchers have observed that in the early stages of development, infants blink when a hand is moved close to their eyes. There is a slow increase in responsiveness over the first months of life. Recent studies have found that although blinking to an approaching object is infrequent during the first month, it is even rarer when an object withdraws. The claim that maturation determines responsiveness is supported by a study which compared two groups of infants. One group was born 3 to 4 weeks after their due dates while the other was full term at birth. The former group responded more consistently than the latter group (Pettersen *et al.*, 1980).

Infants at 1 month appear to be at a lower level of perceptual specificity in perceiving impending collision than are infants between 3 and 4 months. Young infants blink equally to a display corresponding to an object slowing as it approaches and to a display that presents explosive expansion; older infants respond only to the explosive display. It is clear that 1-month-olds do respond to motion-carried information for approach, since they blink more frequently to expansion displays that specify approach than to control displays that do not (Nanez and Yonas, 1994).

Motion-Carried Information for Occlusion

Boundary flow

As described in Gibson (1966), Michotte *et al*. (1964) demonstrated that by manipulating the shape of a disk, one could make it appear to go behind the edge of an occluding surface (see Figure 1).

The horizontal boundary of the black disk in Figure 1 is not perceived as part of the disk; rather, it is perceived as the upper edge of the surface that the disk moves behind. The information that suggests this is that as the disk moves downward, the horizontal contour does not move. There is no information for depth other than the transformation of the curved and straight parts of the display. Normally, objects and their edges move in the same direction at the same speed, while objects

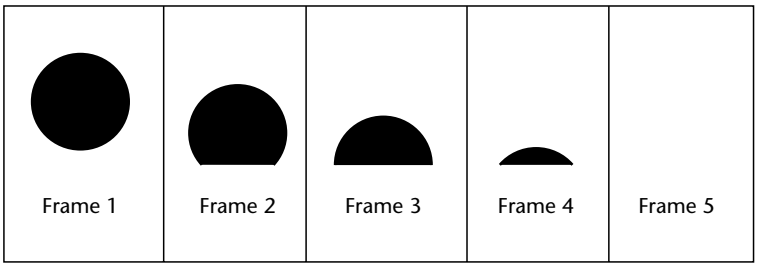


Figure 1. Five frames in which the nature of the disappearance of the disk indicates that it is moving behind a horizontal surface. (Adapted from Gibson (1966) and Michotte *et al.* (1964)).

and background do not share common motion. This assumption is the basis of the motion-carried cue for depth order called 'boundary flow'.

Accretion and deletion of texture

James Gibson (1966) proposed another potential cue for depth order, the accretion (appearance) and deletion (disappearance) of texture, which take place when one surface covers and uncovers another. Experiments that have tested the effectiveness of this cue have also included boundary flow information (Kaplan, 1969). Infants aged 5 months reach to the apparently closer side of a screen on which random dots move as if one side were a surface occluding the other. As in the prior adult studies, both cues were presented. When displays were created in which there was a gap between the texture and the boundary, so that texture did not appear or disappear, 5-month-olds consistently reached to the apparently closer side of the display (Kellman and Arterberry, 1998). Younger infants may also be sensitive to this information, but so far this has not been tested.

Motion-Carried Information for Shape

A field of random dots is perceived as flat when it is motionless. The same dots can be made to appear to be on the surfaces of a three-dimensional cube, and other shapes, by using computer animation to simulate the proper motion. Infants at 2 and 4 months are able to discriminate between different three-dimensional shapes when presented with displays in which motion is the only depth cue. A control condition rules out the possibility that discrimination is based on a two-dimensional property, differences in the velocity of the dots (Arterberry and Yonas, 2000).

BINOCULAR INFORMATION FOR DEPTH

Binocular Convergence

'Binocular convergence' refers to the angle that results when the viewer turns both eyes to fixate a target. Within a distance of about 2 meters, convergence functions as a depth cue for adults. Although 1-month-olds adjust convergence to the distance of near targets, evidence that infants can use binocular convergence to perceive depth has only been shown in 5-month-olds, who reach for an object at the distance specified by convergence (Kellman and Arterberry, 1998).

Binocular Disparity

Binocular disparity is depth information that results from the fact that our two eyes view the world from different points in space. A difference in the relative positions of objects in the images projected to the two eyes is a cue that the objects are at different distances. The tendency of infants to fixate a display that presents differences in disparity rather than one that does not (i.e. that looks flat to adults) has been used to investigate the onset of binocular sensitivity. Most infants begin to show sensitivity at between 3 and 4 months (Held *et al.*, 1980). Infants who show evidence of sensitivity to disparity in a preferential looking task also transfer the perception of shape from motion to displays in which only binocular cues specify shape. This suggests that convergence and binocular disparity function as effective cues for depth in some 4-month-olds (Kellman and Arterberry, 1998).

PICTORIAL DEPTH CUES

Pictorial depth cues, which artists have used to depict space in paintings, are also effective when we view the world around us. When we close one eye and observe the world without moving, it does not become a flat mosaic.

As the distance between an observer and an object decreases, the size of its image on the retina increases. If one assumes that the object is fixed in size, its apparent expansion can be used to infer that it is approaching (as described above). Pictorial cues rest on similar assumptions.

Linear Perspective

The assumption that lines in the world are parallel makes it possible to infer from the convergence of lines in an image that they extend away from the observer. This convergence of lines to a vanishing point is called 'linear perspective'.

Relative Retinal Size

If objects are assumed to be of similar size, differences in retinal size indicate that the smaller image is projected by the more distant object.

Familiar size

Among all the depth cues described here, familiar size is special because it must be acquired through experience. For particular objects that have only one size, there is a relationship between retinal

size and distance. If this relationship is learned, it is possible to perceive the distance of that familiar object from its retinal size.

Interposition

Information for the order of surfaces in depth is given by the way lines are joined. In the proper context, a T-shaped line junction indicates that the stem of the T is occluded by a surface that is bounded by the top of the T. This cue is sometimes called 'interposition'.

Surface Contour

While the cues of linear perspective and relative retinal size appear to be based on the assumption of equality, the surface contour cue requires an assumption that contours in an image are caused by straight lines in the world (see Figure 2).

Shading

The cue of shading can indicate shape, although it requires several assumptions. If a surface in an image is constant in reflectance, illumination is even, and the light source is above, a bulge in a wall will be lighter on its upper surface than on its lower surface. The location of cast shadows can also indicate the position of an object in space. The presence of a gap between an object and its shadow indicates that it is raised above the background, while a shadow that is connected to an object indicates that the object is resting on the background.

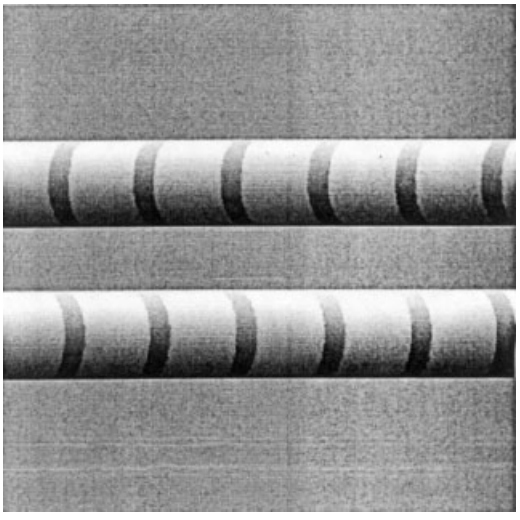


Figure 2. Surface contours indicate that the ends of the cylinders differ in depth.

Research on the development of sensitivity to pictorial cues has generally used reaching as a measure, because most infants, from 5 months, will reach to the closer of two objects when binocular depth information is present. Infants may view a display in which there are no actual depth differences, but in which pictorial cues create, for adults, an illusion of depth when one eye is covered. When viewed with two eyes, the illusion is reduced. A difference in reaching, between one- and two-eyed presentations, constitutes evidence that a depth cue is effective. Studies of pictorial depth cues with 7-month-olds have found more frequent reaching to the apparently closer part of the display with monocular presentation. In contrast, 5-month-olds do not seem to be sensitive to these cues (Kellman and Arterberry, 1998). An account for this change is needed. An interesting possibility is that, at about 6 months, long-distance neural connections become more effective. Thus, the infant becomes able to use pictorial cues and information requiring recognition and memory to direct reaching.

References

Arterberry ME and Yonas A (2000) Perception of three-dimensional shape specified by optic flow by 8-week-old infants. *Perception and Psychophysics* **62**(3): 550–556.

Gibson JJ (1966) *The Senses Considered as a Perceptual System*. Boston, MA: Houghton Mifflin.

Held R, Birch E and Gwiazda J (1980) Stereoacuity in human infants. *Proceedings of the National Academy of Sciences* **77**: 5572–5574.

Kaplan GA (1969) Kinetic disruption of optical texture: the perception of depth at an edge. *Perception and Psychophysics* **6**: 193–198.

Kellman PJ and Arterberry ME (1998) *The Cradle of Knowledge: Development of Perception in Infancy*. Cambridge, MA: MIT Press.

Michotte A, Thinès G and Crabbé G (1964) *Les Complements Amodaux des Structures Perceptives*. Louvain: Publications Universitaires de Louvain. [Described in (Gibson, 1966).]

Nanez N and Yonas A (1994) Effect of luminance and texture motion on infant defensive reactions to optical collision. *Infant Behavior and Development* **17**: 165–174.

Pettersen L, Yonas A and Fisch RO (1980) The development of blinking in response to impending collision in preterm, full term, and postterm infants. *Infant Behavior and Development* **3**: 155–165.

Schiff W (1965) The perception of impending collision: a study of visually directed avoidant behavior. *Psychological Monographs* **79** (whole no. 604).

Slater A, Mattock A and Brown E (1990) Size constancy at birth: newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology* **49**(2): 314–322.

Walk RD and Gibson EJ (1961) A comparative and analytical study of visual depth perception. *Psychological Monographs* **75** (whole no. 15).

Yonas A (1981) Infants' responses to optical information for collision. In: Aslin RN, Alberts JR and Pettersen MR (eds) *The Development of Perception: Psychobiological Perspectives*, vol. II 'The Visual System', pp. 313–334. New York, NY: Academic Press.

Further Reading

Goldstein EB (1999) *Sensation and Perception*, 5th edn. Pacific Grove, CA: Brook/Cole.

Hochberg J (1971) Perception II. Space and movement. In: Kling JW and Riggs LA (eds) *Woodworth and Schlosberg's Experimental Psychology*, pp. 475–550. New York, NY: Holt, Rinehart and Winston.

Kellman PJ and Banks M (1998) Infant visual perception. In: Kuhn D and Siegler R (eds) *The Handbook of Child Psychology*, 5th edn, vol. II, pp. 103–146. New York, NY: Wiley.

Yonas A and Owsley C (1987) Development of visual space perception. In: Salapatek P and Cohen L (eds) *Handbook of Infant Perception*, vol. II, pp. 79–122. Orlando, FL: Academic Press.

Spatial Cognition, Models of

Introductory article

Tom Hartley, Institute of Cognitive Neuroscience, University College London, UK
Neil Burgess, Institute of Cognitive Neuroscience, University College London, UK

CONTENTS

Neurophysiological research has led to detailed mechanistic theories about how spatial locations and headings are represented in the brain and used in memory and navigation.

INTRODUCTION

The term 'spatial cognition' covers processes controlling behavior that must be directed at particular locations, or responses that depend on the location or spatial arrangement of stimuli. There are many circumstances in which such processes are required for adaptive behavior, and they are exhibited in creatures as diverse as bees, birds, rats, and primates. At the most basic level, an organism must be able to flee from a dangerous location to a safer place. It may also need to return to a location where food is abundant or has been stored; to act upon a stimulus at one location, while temporarily ignoring other stimuli; or to navigate from one place to another by an efficient route, avoiding obstacles. All of these behaviors seem to demand some sort of spatial representation; a neural code that distinguishes one place or spatial arrangement of stimuli from another. Models of spatial cognition describe these representations and the nature of the processes that operate on them to give rise to spatial behavior.

NEURAL REPRESENTATIONS OF SPACE

Models of spatial cognition are constrained by experimental evidence from cognitive psychology, neuropsychology, neuroimaging, and neurophysiology. The picture emerging from this evidence is that spatial cognition can be divided into two modes, which are to some extent separated in the mammalian brain.

Broadly speaking, processes involved in action, attention, and perceptual constancy involve the parietal neocortex. The importance of parietal processes in spatial attention and action is illustrated by the well-known neuropsychological phenomenon of hemispatial neglect, in which patients with lesions of the right parietal cortex show an attentional bias towards the right. The presence of stimuli on the patient's right side tends to extinguish any response to a stimulus on the left, so that for instance, a patient may shave only the right side of his face, or copy only the right side of a picture. (*See* **Neglect**)

Processes involved in long-term spatial memory, orientation, and navigation take place in the hippocampus and adjacent cortical and subcortical structures. Patients with damage to these regions, especially in the right hemisphere, are impaired in a range of topographical memory tasks such as drawing maps or judging the distance between locations. Neuroimaging studies have shown that the right hippocampus is activated during the verbal recall of routes and during navigation in a virtual-reality town, and that its activation correlates with success in the navigation task. (*See* **Navigation and Homing, Neural Basis of**)

This division of labour is something of an oversimplification. Many tasks do not fall comfortably into either memory or action categories, but involve elements of both. Additionally, other brain areas are involved in some spatial tasks; for example prefrontal cortex is implicated in tasks demanding planning, while reflexes and stereotyped or overlearned spatial behavior will involve subcortical regions such as the basal ganglia and superior colliculus and cerebellum. Additionally, the hippocampus and parietal cortex have important functions that go beyond those outlined above; for example, in humans the hippocampus plays a more general role in memory for personally experienced events. However, the generalization outlined

above is useful, because it makes clear some of the important constraints on spatial processing that may pertain to different brain regions. (*See* **Planning: Neural and Psychological**; **Basal Ganglia**; **Amnesia**; **Hippocampus**; **Neural Basis of Memory: Systems Level**)

First, the processing modes differ in the spatial and temporal scales over which they operate. Hippocampal processes are concerned with large distances and long timescales, whereas parietal processes are more concerned with short timescales and the space immediately surrounding the body.

Second, the processing modes differ in the forms of spatial representation they demand. Parietal processes controlling action in the immediate environment use egocentric representations of space (i.e., locations are represented in terms of their relation to the subject). So, for instance, the firing rates of neurons in the medial interparietal area of monkey parietal cortex, which fire when the animal is about to reach for an object, vary depending on the position of the object relative to the monkey's hand. This is an example of a neural representation in an egocentric (in this case hand-centered) reference frame. Such representations are clearly useful for guiding action over the short term, or where the stimuli whose locations are to be encoded are immediately available to the perceptual system. (*See* **Parietal Cortex**)

Egocentric representations have the disadvantage that, in order to remain valid over the long term, they must be actively updated to reflect changes in the subject's location and heading. Unless corrected by new sensory information, any errors in this updating process will be cumulative, so that egocentric representations of location are unreliable for long-term storage.

In contrast, processes demanding long-term memory of a location should make use of representations that relate locations to each other and to landmarks in the environment, rather than to the subject. Such representations are called world-centered or 'allocentric'. They are map-like in the sense that there is no privileged location to which all others are related. Instead, they provide a basis from which one's current location and orientation can be computed from one's relationship to sensory cues in the environment. A set of locations represented in an allocentric framework can be thought of as a 'cognitive map'. (*See* **Animal Navigation and Cognitive Maps**)

A cognitive map has several advantages in the context of long-term memory. Over a period of days, months or years, a given place may be approached from different directions on different occasions; the viewpoint-independence of an allocentric representation will thus be useful in navigating towards or recognizing locations over such timescales. Furthermore, locations represented in an allocentric reference frame do not need to be continuously updated in the way that locations in an egocentric reference frame must be. An allocentric form of representation is thus not prone to the cumulative error inherent in such an updating process, making it particularly suitable in tasks where behavior has to be directed towards a location that is not immediately available to perception. This might occur either because it is far away (as in the case of navigation in large-scale space; for example, returning to a familiar nesting site after a protracted foraging expedition) or because it is hidden (e.g., returning to the location of a hidden food store).

For long-term memory, an allocentric representation of space could provide a solution to some of the shortcomings of egocentric representations. However, it also raises some new questions, not least of which is how a map-like representation can be abstracted from the egocentric information available to sensory systems. What form does the allocentric 'cognitive map' take, and how could it support navigation? In order to understand how allocentric representations are formed, it is useful first to consider the ways in which transformations between different reference frames might be achieved in the parietal cortex. Such transformations are certainly involved in immediate action-orientated processes, but may also occur in the encoding or retrieval of long-term spatial memories.

TRANSFORMATION BETWEEN REFERENCE FRAMES

Sensory cortices encode stimulus location egocentrically. For action, it is generally necessary to transform information about the location of a stimulus into a reference frame appropriate to the effector system involved in the response. For instance, visual information about the location of an object is encoded retinotopically in visual cortex (reflecting location in an eye-centered reference frame) and must be transformed into a hand-(or arm-) centered coordinate system in order that it can be used to direct a reaching response.

In fact, most actions are likely to require multiple reference frames, as they demand the coordinated action of many effectors moving in concert. There is neurophysiological evidence of the existence of such multiple egocentric reference frames in the parietal cortex of monkeys. As well as cells representing object locations relative to the monkey's

hand (mentioned previously), other populations of cells have been discovered where firing rates are determined by the relationship of objects to the monkey's trunk, arm, head, and so forth. (*See* **Parietal Cortex**)

How does the parietal cortex transform locations represented in one reference frame to another? A three-layer feed-forward neural network can be trained to perform such transformations (e.g., between eye-centered and head-centered coordinates). A layer of input neurons represents locations on the retina and the gaze direction (the angle of the eye in the head), and a layer of output neurons represents locations relative to the head. Between the input and output layers is a 'hidden' layer of processing neurons. During training, the strengths of connections to and from the hidden layer are gradually changed, so that the output layer produces increasingly accurate transformations of the input. Ultimately, each output neuron has a firing rate determined by stimulus location relative to the head – the connections to and from the hidden layer of neurons translate the eye-centered representation of the stimulus location into a head-centered reference frame, taking into account the gaze direction.

What is interesting about this model is the way in which the transformation is achieved. After training, neurons in the middle processing layer between the input and output layers show responses that are modulated by both the location of a stimulus relative to the head, and by its retinal location. This type of encoding is referred to as a 'gain field'. It is a useful form of representation because the combined influences of stimulus location in two different reference frames on the neural response would allow the location of the stimulus to be represented in either reference frame in subsequent processing layers. Gain field responses have been found in neurons in the interparietal sulcus of the monkey posterior parietal cortex. It is plausible that the multiple egocentric representations of location in the posterior parietal cortex are linked by intermediate gain field representations that mediate the translation of location information from one reference frame to another.

NEURAL BASIS OF SPATIAL MEMORY AND ORIENTATION

Place Cells

Experiments with rats provided some of the first evidence that the hippocampus was involved in spatial memory and the allocentric representation of space. A popular laboratory task demanding spatial memory is the Morris water maze, in which rats learn to escape from a pool of cloudy water by navigating to a platform hidden beneath its surface. Rats with lesions to the hippocampus are unable to learn the location of the platform in the water maze. The anatomy of the hippocampus is substantially similar in rats and primates, including humans, so there are grounds for believing that the mechanisms of spatial memory and navigation in these diverse species may also be similar. But how does the hippocampus represent locations, such as the location of the hidden platform?

Place cells, in areas CA1 and CA3 of the hippocampus, encode the rat's location independently of its heading. They are cells whose firing rates vary depending on where the animal is. In a typical experiment the firing of hippocampal pyramidal cells is recorded as the animal freely explores its environment (typically a low-walled box or a raised platform). A place cell might fire, for instance, whenever the rat is in the northeast corner of the environment, but not elsewhere. Each place cell responds in a different part of the environment, referred to as the cell's place field. Together these firing fields cover the entire area of the box. Each place in the box produces a different pattern of firing in the place cells. By looking at the firing rate of several place cells at once (say 30), one could know to within a few centimeters where in the box the rat is at any time.

Place cells could clearly be of importance in representing behaviorally important locations, and their discovery prompted O'Keefe and Nadel to propose that the hippocampus functions as a cognitive map. The mechanisms underlying the neural representation of place are likely to be important in understanding cognitive processes such as navigation and long-term spatial memory, and we will return to the question of how place fields are formed later.

Head Direction Cells

Whereas place cells encode the rat's location independently of where it is heading, a complementary system of 'head direction cells' represent the animal's heading independent of its location. Each has a preferred direction (e.g., north) and fires whenever the rat is facing in that direction. Head direction cells are found in the mamillary bodies, anterior thalamus and presubiculum, parts which together with the hippocampus are connected to form an anatomical circuit. Over short periods, populations of head direction cells function as a

neural compass, tracking changes in heading with a remarkable accuracy. Recorded over longer periods, they do not maintain fixed compass directions: the direction to which a particular cell is tuned depends on sensory (especially visual) cues present in the environment. Rotating all such cues causes a matching change in the head direction cell's response, and if visual cues are absent, the tuning of a cell may gradually drift. Interestingly, the responses of simultaneously recorded head direction cells remain locked together (always maintaining the same angle between preferred directions).

A neural compass could clearly be of great use in navigation, but it may also fulfil another, related role in spatial memory. As noted above, an important part of the process of forming a long-term representation of a location may involve transforming egocentric representations of the environment into an allocentric form. One of the essential properties of such a representation is that it is orientation-neutral (i.e., independent of the heading direction at the time of encoding). This means rotating sensory information about the location to be encoded, so that directions are represented in a reference frame that is fixed with respect to the world (i.e., as compass directions) rather than one that is fixed with respect to the direction one is currently facing. The transformation between orientation-specific perceptual and orientation-neutral mnemonic representations could, in principle, be achieved through a 'gain field' mechanism (see above), and would require a representation of heading (analogous to the gaze angle information in the transformation between eye- and head-centered reference frames). Thus the head direction system could be involved in the encoding of long-term memories. This may explain why the anatomical circuit involved in head direction is also implicated in amnesia.

MODELS OF ALLOCENTRIC SPATIAL REPRESENTATIONS

We now turn to the question of how allocentric representations of the kind described above might be abstracted from egocentrically encoded sensory information: what makes a place cell fire where it does?

A place field's location does not depend on the rat's position relative to any single cue. This can be demonstrated by changing various aspects of the environment while the rat is absent. For instance, the walls making up the perimeter of the box can be swapped around, or the floor of the box replaced, but a place cell will still fire in the same location. Therefore the cell does not fire in response to the distinctive smell of the wall or floor near its firing field. Similarly, removing subsets of visual cues need not affect place cell firing.

One way of explaining the independence of place fields to any single external cue is that internal information (e.g., self motion, vestibular and motor efference information) is used to track changes in location over time, with external sensory information simply serving to calibrate this system. This mechanism is often referred to as 'path integration'. Another possibility is that external cues are used, but that a conjunction of several features is required to drive a place cell to fire at a given location. As the hippocampus receives inputs from many areas of sensory cortex, both external cues and path integration information could play a part in controlling place cell activation. Experimental evidence indicates that visual cues are the most important determinants of place cell firing under normal circumstances. However, visual information is not essential; congenitally blind rats have apparently normal place fields.

The geometry of the environment is particularly important in determining place field locations. This can be demonstrated by varying the size and shape of the experimental environment. The locations of place fields remain fixed with respect to some of the walls of the environment (usually the nearer walls; Figure 1) even when the box is moved within the laboratory (and thus with respect to many distant visual cues visible over the low walls of the box). Place field locations are not affected when smaller objects placed within the enclosure are moved around, but if the same objects are placed close together so that they form a more substantial barrier to movement, they do affect place field location.

These results suggest that the distances and directions of boundaries are represented in the cortical inputs to the hippocampus. These distances and directions are probably determined by external sensory systems, as they would be difficult to compute on the basis of path integration alone. Because place cell firing rates in open environments tend to be independent of the rat's heading, it would appear that the directions of these critical features are represented in an orientation-neutral reference frame (north, south, east, etc.) rather than one specific to the current heading (left, right, ahead, etc.). This suggests that initially egocentric representations of the distance and direction of geometric features of the environment are transformed into an allocentric directional framework, by taking into

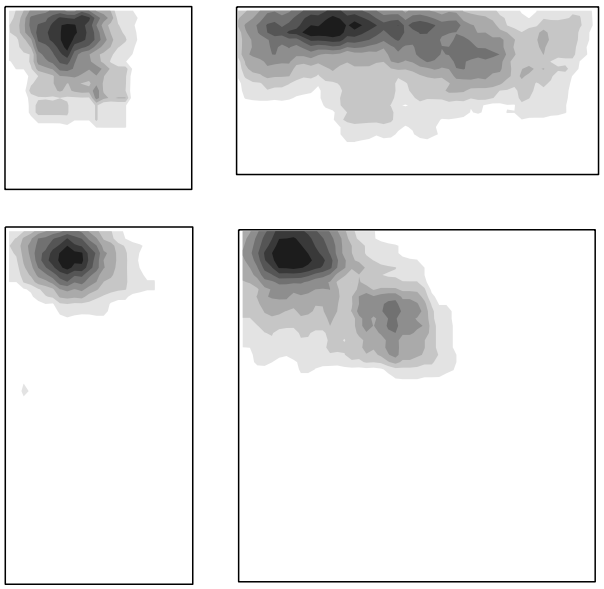


Figure 1. Place cells in area CA3 of the rat hippocampus have firing rates that vary systematically according to the rat's location. The region where the cell fires is called a place field. The shading indicates the place fields of one cell, recorded as the rat explored four different rectangular boxes. In each case, the cell fires most strongly in the northwest corner of the box (i.e., the location of the field remains fixed relative to the north and west walls). Adapted from O'Keefe J and Burgess N (1996) Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**: 425–428.

account the rat's heading, as discussed previously. Consistent with this idea, manipulations that affect the orientation of the head direction system also affect place field locations. For instance, if distant visual cues are rotated about the center of the rat's environment, place fields are rotated through a corresponding angle.

A Simple Feedforward Model of Place Field Formation

The simplest models of place field formation are feedforward models. Many of these assume that the inputs to the hippocampus are sensitive to landmarks at particular distances and/or directions, but as we have seen some features of the environment (such as distant visual cues or walls) are more important than others.

As an example of a feedforward model, let us assume that the inputs to the hippocampus include some that are dependent on the geometry of the environment and in particular its boundaries (Figure 2). These hypothetical cells have the property that they fire strongly when a wall is at a given distance and allocentric direction to the rat. A place cell could receive inputs from many such cells, with the place field reflecting the overlap between the firing fields of its inputs. For instance, a place cell with a firing field in the Northeast corner of a box might have one input which responds whenever there is a wall a few centimeters North of the rat, and another that fires when there is a wall a few centimeters to its East. Either stimulus on its own might be insufficient to drive the place cell to fire (so that it does not fire in the Northwest corner, for instance), but the combined inputs of cells responding to the North and East walls drive the place cell to fire in the Northeast corner. If we change the geometry of the box (for instance by doubling its size), the cell will still fire in the Northeast corner. Note that in this model the only requirement for path integration or similar processes is in the tracking of heading.

The above model explains data concerning place field locations in boxes of different shapes and sizes, and predicts that there are cells upstream of the hippocampus, each of which fires in response to a boundary at a particular distance and bearing. It also predicts (in line with experimental observations) that even radical alterations of the shape of the environment should result in fields fixed with respect to the nearby walls of the environment. The model makes no mention of learning, and only requires that each place cell be connected to a random selection of inputs in order to account for the observed patterns of place field location in differently shaped environments. Indeed, one can use place fields obtained from the same cell recorded in differently shaped environments to infer which geometric inputs are driving the cell, and thus to predict the behaviour of the cell in a novel environment.

One potential problem with this simple model is that it makes no mention of the abundant lateral connections that are known to exist between cells in area CA3, and is at odds with data showing that markedly different place fields can sometimes be established in different boxes. This phenomenon, known as 'remapping', has been taken as evidence that the place cell system functions as an 'attractor' network.

Attractor Networks of Place and Orientation Representations

Attractor networks are neural networks with recurrent connections (i.e., connections between neurons in a single processing 'layer'). These make it possible for any neuron in the interconnected layer to affect the activation of any of the

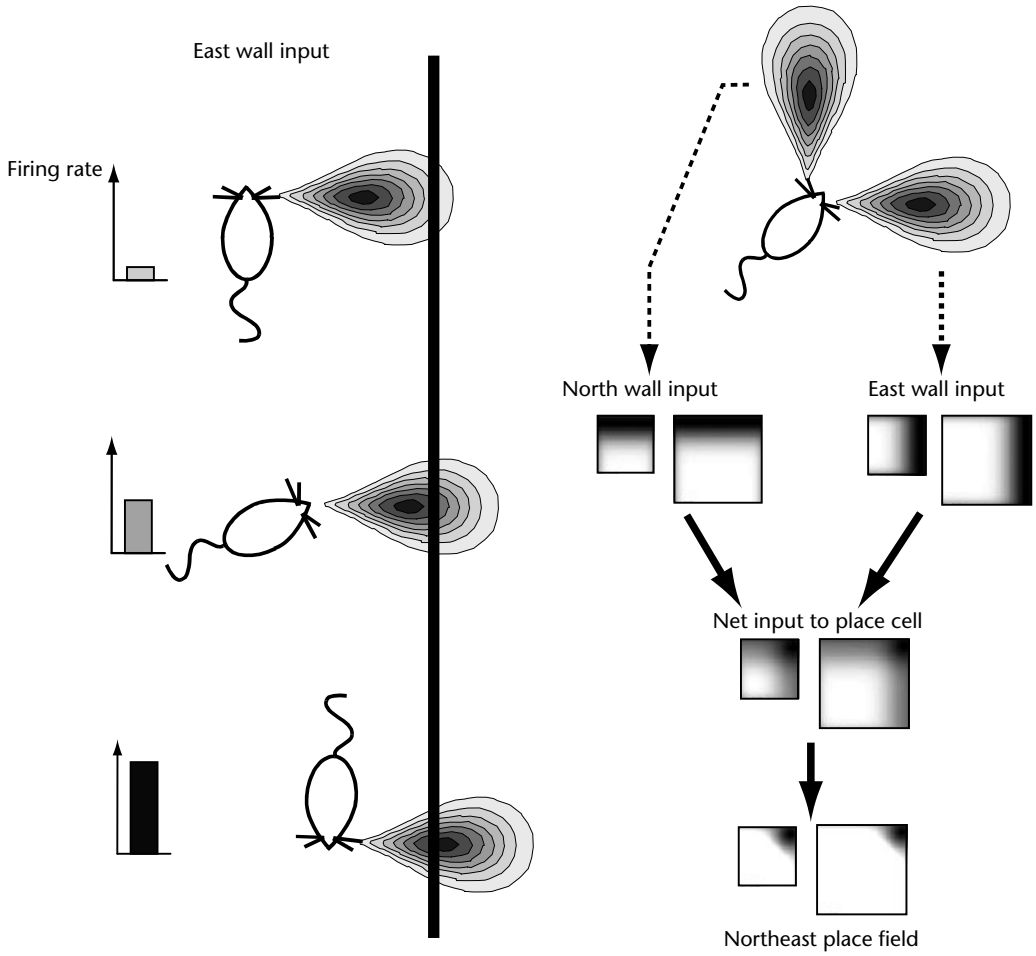


Figure 2. A simple feedforward model of place field formation. In this model place cells have inputs that respond to walls or other boundaries at particular distances and directions from the rat. The left panel shows the receptive field of one such cell (sensitive to a boundary a short distance to the East), and illustrates how the cell's firing rate varies as the rat moves nearer to the wall. Note that the direction that the input cell is tuned to does not depend on the rat's heading. Place fields could be formed by combining several such inputs. The right panel shows how a place field in the Northeast corner of two different-sized square boxes can be modeled by combining inputs responsive to North and East walls. The two input cells fire close to the North and East walls respectively (in both boxes). The net input to the place cell is greatest when the rat is in the Northeast corner of either box. The place cell's firing rate is a thresholded function of its net inputs: the cell fires only in the Northeast corner of the box. Adapted from Hartley T, Burgess N, Lever C, Cacucci F and O'Keefe J (2000) Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**: 369–379.

others to which it is directly or indirectly connected, a property that makes their dynamics potentially rather complex. However, with some constraints, such as symmetry in the strength of connections between neurons (connections from A to B are of equal strength to those from B to A), an attractor network is guaranteed to settle into one of a number of possible stable states (attractor states).

Place Representations

In an attractor model of place cell firing, spatial inputs from sensory neocortex activate some of the hippocampal pyramidal cells. Activation then flows between cells along the recurrent connections. This gradually changes the pattern of firing in the pyramidal cells (indirectly activating some cells that are not driven directly by the neocortical input and inhibiting some of those that are) until it settles at an attractor state.

In 'continuous attractor' networks, the connection strengths are constrained so that rather than having a number of distinct attractor states, the network will move smoothly between attractor states as the input is changed. One way this could be achieved in place cells is for cells with nearby

place fields to have strong connections, and for place fields with well-separated fields to have weak (or inhibitory) connections.

If the inputs to the hippocampus are spatially modulated, and the hippocampal pyramidal cells are connected together to form a continuous attractor network, then one would expect to find a population of place cells that shows the same pattern of activation whenever the animal visits a particular location, with some cells firing strongly, others only weakly. At another location one would see a quite different pattern of activation. As the pattern of active cells changes smoothly as the rat moves from one place to another, the firing rates of individual cells will also vary smoothly between locations; i.e., each cell will have a smooth firing field with some spatial extent – a place field. With appropriate lateral connections, any spatially modulated input could be sufficient to produce place fields, so path integration could play a more significant role in this type of model. However, it would still be difficult to account for the geometric constraints on place field location in terms of a model whose spatial inputs were exclusively internal.

A continuous attractor model requires a particular pattern of connection strengths between place cells. These connection strengths might be learned on exposure to a new environment. However, place fields appear to be present on the first exposure to a new environment (within a few minutes) and are stable for at least several days. This suggests that either learning is extremely rapid or that preexisting lateral connections are sufficient to produce the observed fields.

Orientation Representations

The head direction system can also be modeled as a continuous attractor network with inputs corresponding to visual cues to orientation and vestibular responses to left–right rotation. In this model, connections between head direction cells are such that, for any input, the network will settle into a stable state with a few cells (representing the current heading) firing, and the rest silent: cells representing opposite directions inhibit one another, while there are weak excitatory connections between cells representing neighboring directions. One can imagine the head direction cells as arranged in a circle, with excitatory connections between neighboring cells, and inhibitory connections between more distant cells. At any time a cluster of active cells will indicate the direction of the neural compass. By virtue of the hard-wired connections each head direction cell receives from the vestibular system, the activation peak can be driven clockwise or counterclockwise. The visual inputs function to calibrate the system and tie it to the external world.

MODELS OF NAVIGATION

Together, the two types of allocentric representation described above – place and heading – provide the basic tools for spatial memory and navigation. However, it is important to note that they do not in themselves constitute a cognitive map. A map also has to identify important locations and store them for later use: without some representation of an intended destination, information about current location and heading will be of little use in navigation. Furthermore, a cognitive map must also include information about the way in which different places are connected to one another (topology), so that it is possible to plan a route between places that avoids obstacles. Here we must be more speculative about the neural mechanisms, because to date no neurophysiological study has identified neurons that encode either destinations or topology.

The recurrent collaterals in region CA3 also figure prominently in models of navigation. Hebb's ideas on the modification of functional connections between cells, and the apparently related process of long-term potentiation of synapses, indicate that strong connections should develop between cells that fire simultaneously. For place cells, this would produce strong connections between cells with nearby fields (as required to form a continuous attractor, see above). However, it would also lead to the topology of space being encoded in the strength of the recurrent connections; in other words, the strength of a connection between place cells could indicate, on average, how long it took the rat to travel between the corresponding place fields. Thus the places along the shortest path between two locations could be found in principle, although how this would work in practice is less clear.

Some experiments have shown a temporal asymmetry in long-term potentiation: connections from one cell to another are strengthened most when the first cell fires before the second. This asymmetry would allow the learning of the path taken by an animal as it moved from one place to another. Place cells with firing fields near the beginning of the route will develop strong connections to place cells with fields further along the route. These directional connections could, in principle, be used to store the route but, again, it is less clear how they would be used to guide behavior in practice.

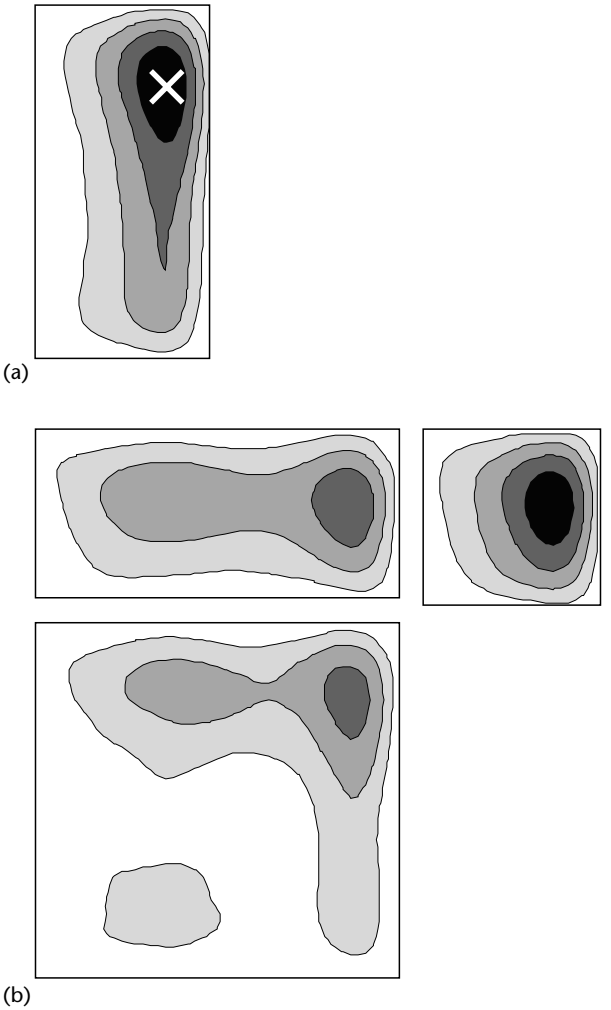


Figure 3. (a) A simple goal cell model of navigation to a stored goal location. When reward is received at location X, a goal cell is activated. Connections from active place cells to the goal cell are strengthened. The goal cell's firing rate will now vary depending on the rat's proximity to X. To return to the stored location the animal must move so as to maximize the firing rate of the goal cell. The firing rate map was made by simulating a large number of place cells each having a random selection of geometric inputs and thus producing a varied selection of place fields which overlapped and covered the entire space of the box. The goal cell firing rate is calculated as the similarity of the place cell firing rates at each location to the stored pattern (firing rates at X). (b) Where will the rat search if the shape of the box is changed? Because we can simulate the firing of the same place cells in different boxes, we can also predict the search locus in environments of different shape. Such predictions will allow us to test the model in behavioral experiments. Adapted from Hartley T, Burgess N, Lever C, Cacucci F and O'Keefe J (2000) Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**: 369–379.

As with models of spatial representation, simple feedforward models may suffice for modeling navigation. Perhaps the simplest model of spatial memory posits that receiving a reward at a particular location causes a 'goal cell' immediately downstream of the hippocampus to fire. This in turn strengthens synaptic connections to it from the place cells active at that location. The subsequent firing of the goal cell would indicate the proximity of the goal. As the animal moves nearer the goal location, more of the place cells with strong connections to the goal cell will fire, increasing the net current input to the goal cell. Thus the animal will be able to return to the goal simply by moving so as to increase the firing rate of the goal cell. More generally, this type of model predicts that the likelihood of searching for a remembered location in a particular place increases with the similarity of the place cell representations of the goal and the place (Figure 3).

Other models have looked at the generation of movements in navigation in more detail, enabling the animal to head directly to the goal, rather than hunting around to ascertain the right direction. These models confront the problem of translating allocentric hippocampal information (such as 'the goal is to the north') into egocentric movements (such as 'turn left'). This translation is assumed to occur in the parietal cortex or basal ganglia, again making use of the head direction system to translate between orientation-neutral and orientation-specific reference frames.

CONCLUSION

Thanks to neurophysiological data gathered over more than three decades we now have fairly detailed mechanistic theories of the processes involved in forming spatial representations, in translating between representations in different reference frames and in storing them and using them to guide behavior. These models have now progressed to the point that we can make links between physiology and complex behaviors such as navigation, and thus develop an ever more detailed understanding of spatial cognition.

Further Reading

Burgess N, Recce M and O'Keefe J (1995) Spatial models of the hippocampus. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks*, pp. 468–472. Cambridge, MA: MIT Press.

Burgess N, Jeffery KJ and O'Keefe J (eds) (1999) *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford, UK: Oxford University Press.

Burgess N, Becker S, King JA and O'Keefe J (2001) Memory for events and their spatial context: models and experiments. *Philosophical Transactions of the Royal Society (London) Biological Science* **356**: 1493–1503.

Muller RU, Stead M and Pach J (1996) The hippocampus as a cognitive graph. *Journal of General Physiology* **107**: 663–694.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford, UK: Oxford University Press.

O'Mara SM (ed.) (2000) Special issue on the nature of hippocampal–cortical interaction: theoretical and experimental perspectives. *Hippocampus* **10**: 351–499.

Samsonovich A and McNaughton BL (1997) Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* **17**: 5900–5920.

Skaggs WE, Knerim JJ, Kudrimoti HS and McNaughton BL (1995) A model of the neural basis of the rat's sense of direction. In: Tesauro D, Touretzky DS and Leen TK (eds) *Advances Neural Information Processing Systems 7*, pp. 173–180. Cambridge, MA: MIT Press.

Taube JS (1998) Head direction cells and the neuropsychological basis for a sense of direction. *Progress in Neurobiology* **55**: 225–256.

Trullier O, Wiener SI, Berthoz A and Meyer JA (1997) Biologically based artificial navigation systems: review and prospects. *Progress in Neurobiology* **51**: 483–544.

Zipser D and Andersen RA (1988) A backpropagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* **331**: 670–684.

Spatial Cognition, Psychology of

Introductory article

Barbara Tversky, Stanford University, Stanford, California, USA

CONTENTS

People think about space in terms of the elements in space and their spatial relations, not in metric terms. Different spaces are conceptualized differently: the space of the body, the space around the body, the space of navigation, and the metaphoric or miniature space of graphics.

CONCEPTUALIZATIONS OF SPACE

We live in space. All of our perceptions and actions are situated in space. From the moment of birth, perhaps even before, and for every minute thereafter, we are engaged in spatial cognition, in implicit awareness of or active interaction with the surrounding space. We engage in spatial cognition in our imaginations as well, in ways ranging from cognizance of the things around us that we cannot currently perceive, to judgments about environments too large to be perceived, to mental modeling of imaginary worlds. Spatial knowledge comes both directly, from perceptual experience (not only from vision, but also from audition, proprioception, olfaction, and other senses), and indirectly, from language, diagrams, and maps.

Our conceptions of space are not geometric. In mathematical geometry, space itself is primary, and things in space are located with respect to the external coordinates of the space. This view of space is unitary: it applies to all spaces, regardless of what is contained in them. Measurement is quantitative. For people, the things in space are primary, and they are located, mentally, with respect to each other and to a psychologically salient reference frame. Different spaces are conceived of differently, with different reference objects and frames, depending on the roles they play in people's lives. Our spatial measurement is qualitative and piecemeal, even inconsistent.

Knowledge of space is not just internalized perceptions of space. Nor is it simple averages or combinations of memories of space. Rather, spatial knowledge is constructed. These constructions are based on conceptions that are based on experience with space, rather than internalized perceptions of experience. Although the constructions differ for different spaces, their essential ingredients are elements and the spatial relations among them. This article will discuss some of our psychological spaces: in particular, the space of the body, the space around the body, the space of navigation, and the space of external representations.

ELEMENTS OF SPACE

Elements in space have spatial extent in themselves. Thus, although under some perspectives elements in space are treated as points, they may also be treated as spatial entities. Usually, what we refer to as 'objects' or 'entities' have integrity in the sense that their parts are connected, hence they move in concert; and they are smaller than we are, things we can manipulate. This is in contrast to 'environments', the spaces we move in or that surround us. One of the remarkable features of the human mind is that it can conceive of environments as objects – even as points, as when we represent cities in maps.

Just as different spaces are treated differently by the mind, so different elements in space are treated differently. Objects seem to be recognized by their shapes or contours. Shapes, in turn, are determined by parts in configuration. Thus our conceptions of object parts, as well as of objects, play a role in recognition. Faces are a special kind of object. Technically, they are part of a larger object, a human body. But when we refer to face recognition, we usually mean recognition at the level of the individual (e.g., recognition that it's John F. Kennedy, not Bobbie or Ted), rather than recognition at the level of the category (e.g., recognition that it's a table and not a chair, an elephant and not a giraffe). Since all faces have the same general contour,

contour is not useful for identifying faces at the level of the individual. Rather, the shapes and exact configuration of internal features seem to be critical for identifying faces. Different regions of visual cortex are differentially sensitive to faces and objects, suggesting that recognition of these different classes of stimuli differs computationally. There are brain-damaged patients who can easily recognize some classes of objects (e.g., living things), but have difficulties recognizing other classes of objects (e.g., artifacts).

Even more significant for spatial cognition is evidence that the parahippocampal place area is selectively involved in recognizing places – that is, three-dimensional configurations of walls, furniture, buildings, streets, trees and so on. This area is selectively active under viewing of scenes, and patients with damage to this area experience severe difficulties acquiring spatial knowledge of new places.

THE SPACE OF THE BODY

Bodies can be regarded as objects. But bodies have a special status as objects in that we know them from the inside as well as from the outside. Bodies can also be regarded as spaces, as in cases of sunburn or rashes. How is the space of the body conceived? Are the parts of the body most readily accessed the larger parts, or the most perceptually salient parts, or the most important parts? To address this question, observers viewed from the side bodies in many positions and orientations with one body part highlighted. The set of parts that could be highlighted was drawn from the parts that are most commonly named across languages: head, arm, hand, chest, back, leg, and foot.

Three types of theory have been proposed to explain representations of the space of the body. According to 'imagery' theories, larger parts should be identified faster; this would mean that chest and back would be identified faster than hand or head. Theories of 'object recognition by components' suggest that parts that are perceptually salient, that is, parts that extend from the object's contour, should be most rapidly identified, so arm and leg would be identified more rapidly than back. Finally, parts rated as 'good' enjoy both perceptual salience and functional significance. The third type of theory posits that parts that have functional significance should be most accessible. Perceptual distinctiveness and functional significance are correlated in artifacts and natural kinds. Parts have a distinctive appearance as well as a distinctive function so that they serve as a bridge between perception and function. For body parts, too, salience and significance are correlated, but not completely. The major exception is the chest, which has relatively low contour discontinuity but relatively high functional significance, perhaps because it protects crucial organs of the body, perhaps because it represents the forward direction privileged in perception and action.

The observers in the experiments described above had to decide if pairs of stimuli referred to the same or different parts. The comparison stimulus was always a depiction of a body with a part highlighted. The other stimulus was either a name of a body part or another depiction of a body (oriented differently) with a part highlighted. Half the comparisons were of the same body part, half of a different body part.

The results for the two tasks differed in an interesting way. For body–body comparisons, perceptual salience was the best predictor of identification reaction times. However, for name–body comparisons, functional significance was the best predictor. When two depictions of bodies appear on the screen, they can be compared as objects, even as forms devoid of meaning. The process of comparing a name to a depiction apparently involves meaning, and the meaning of a body part includes its functional significance.

In fact, rankings of salience and significance of body parts are correlated with each other; they are negatively correlated with part size; and part size is negatively correlated with identification times. Thus, larger parts are relatively less salient and significant. The space of the body, at least for these tasks, is not conceived of in terms of size; rather, it is conceived of in terms of salience and significance.

The space of the body, then, is decomposed into natural body parts. Size is not of primary importance in mental representations of the space of the body. The parts that are most accessible are not the largest parts but rather the parts that enjoy perceptual distinctiveness and functional significance, which are correlated. Perceptual distinctiveness is more important when the body is regarded as an object, but when language is involved, functional significance seems to be more important.

THE SPACE AROUND THE BODY

As we move about the world we maintain awareness of the things around us that are no longer in view. This knowledge appears to be a consequence of mental models that we form of the world around us, which we update as we move or

the surroundings change. Moreover, it seems that such mental models can be induced by language describing movements and things in the world as well as by actually moving in and perceiving the world. Our ability to turn descriptions of the world into mental models of the world underlies some of the power of literature and poetry. Research suggests that such mental models are schematic, rather than rich detailed images; and indeed, that even active perception of the extant world is schematic and lacks rich detail: major elements of viewed scenes can be changed without viewers' noticing.

In experiments designed to investigate the nature of these mental models, participants read descriptions of themselves in environments such as a hotel lobby or opera house, surrounded by objects to their front, back, left and right, and above and below them. Once they had learned such an environment, they were verbally reoriented to face another object in the scene. Then they were probed by direction terms ('front', 'back', 'left', and so on) for the objects currently in those directions from the body. Repeatedly, participants were reoriented and probed again. Performance was essentially perfect; that is, participants readily reoriented themselves in these described environments. The data of interest were the times to access objects in different directions from the body.

It might be expected that all directions should be equally accessible, as the content of the narratives did not privilege any regions of space or any objects over others. However, access times differed considerably, depending on the direction. One way in which participants might perform the task would be to mentally image the environment and then mentally scan the environment for the objects in the probed directions. This model also failed to account for the pattern of retrieval times.

The data were, however, consistent with a 'spatial framework' account. According to this account, participants construct a mental spatial model from extensions of the three body axes and attach the objects to them, updating as needed. Accessibility depends on characteristics of the body axes as well as on characteristics of the axes of the spatial world. For an upright observer, the head–feet axis is most accessible because it is an asymmetric, hence easily discriminable, axis of the body and because it is aligned with the only asymmetric axis of the world, the axis of gravity. Front–back is next, as it is asymmetric; and the least accessible axis is left–right, which lacks salient asymmetries. The situation changes when the observer reclines and no axis of the body is aligned with gravity. Then accessibility depends only on the body axes: front–back is fastest to access, as that axis separates the world that can be perceived and manipulated from the world that cannot; head–feet is next; and left–right is slowest, as before.

These patterns of accessibility of directions from the body obtain not only from worlds learned from narratives, but also from worlds learned from experience, models, and diagrams. Thus, however the knowledge is acquired, the space around the body is conceived in three dimensions, extensions of the body axes.

THE SPACE OF NAVIGATION

The space of navigation is too large to be seen from one viewpoint; rather, it is pieced together from different viewpoints, and often from different occasions and different types of experience – for example, perception, language, and maps.

Route and Survey Perspectives

Language is often used to establish mental models of spaces larger than the ones around the body. Spontaneous descriptions of the space of navigation use one of two perspectives, or a combination of both. A 'route' perspective takes the listener (addressed as 'you') on a mental tour through an environment, locating landmarks with respect to the traveler in terms of 'front', 'back', 'left', and 'right'. A 'survey' perspective takes a viewpoint above the environment and locates landmarks with respect to each other in terms of 'north', 'south', 'east', and 'west'. A route perspective is egocentric, a survey perspective allocentric.

Do the mental models established from the two perspectives differ? For well-learned restricted environments, apparently not. In a series of experiments, participants studied descriptions of spaces, such as small towns or convention centers from either a route or a survey perspective. They then answered 'true or false' questions from both perspectives, either verbatim from the texts or inferable from the texts. Response speed and accuracy were greater for verbatim statements from the studied text than for inference statements, indicating that participants retained the exact wording of the text. However, for inference statements, there were no differences in speed or accuracy between the studied perspective and the other perspective in four experiments. Maps sketched at the end of the experiments were highly accurate for both perspectives. These findings suggest that readers

formed perspective-free representations, something akin to an architect's model, which can be viewed from many different perspectives.

The environments of which people readily formed mental models were, however, schematic and restricted. They included the relative spatial relations of only a dozen or so environmental elements. For larger environments, or even smaller ones that are studied only briefly, the way the environment is learned does affect the mental representations used. Switching the perspective of a description of an environment from route to survey or vice versa slows down reading times during acquisition of the environment. Representations of complex environments can show lasting effects of the mode of acquisition. For example, participants who learned a building complex from experience were better at route distance estimates than participants who learned from a map; while participants who learned a building from a map were better at imagining adjacent rooms that were not directly linked. The latter investigations found equally strong effects of task goal, to learn the layout or to learn routes, on later mental representations. Moreover, research using brain imaging shows that route learning and retrieval activate different brain areas from map learning and retrieval.

Cognitive Maps and Cognitive Collages

In the experiments described above, the spatial mental models established from descriptions, though accurate, were highly schematic. Indeed, they were acquired from schematic summaries – either descriptions or sketch maps, both of which abstract and structure information, in similar ways. The models were also limited to small, well-learned environments.

People's knowledge of less constrained environments – which may have been acquired from experience, maps, descriptions, or some combination of these – is also schematic, but is often piecemeal and scattered. It does not seem that people have 'cognitive maps', in the sense of coherent mental representations of all the environments they have encountered, stored in memory and ready to be consulted. Rather, it seems that when people need information about an environment – perhaps to answer queries about directions or distances, perhaps to navigate, perhaps to make inferences about weather or population migrations – they retrieve whatever information seems relevant. That information might be in a variety of formats; it might be incomplete; it might be schematic; it might be difficult or impossible to integrate into a coherent whole. Because such representations are put together 'on the fly' from a multitude of media, a more apt metaphor than 'cognitive map' for the functional mental representation is 'cognitive collage'.

Much of the evidence for this conclusion comes from research demonstrating systematic errors in cognitive 'maps'. These errors can reveal the perceptual and conceptual organizing processes that people use to represent and integrate spatial information. One of the first processes in perception is distinguishing figures from ground. Once figures have been isolated, they are oriented – a critical process in their identification and located with respect to each other and a frame of reference. Systematic errors of alignment and rotation reflect each of these processes.

Errors of alignment

When asked the direction between Chicago and Monaco, a majority of people report, incorrectly, that Monaco is south of Chicago. Similarly, when asked the direction between Boston and Rio, most people report, again incorrectly, that Boston is east of Rio. In organizing the geography of the world, people seem to mentally align large regions like the US and Europe or North and South America. The US and Europe are remembered at similar latitudes and North and South America at similar longitudes. In fact, if presented with a choice between a true map of the world or the Americas and a map that has been altered so that the US and Europe or North and South America are more aligned, a majority of people choose the incorrect aligned map as the 'true' map. The alignment effect also appears in memory for artificial maps and for meaningless blobs, showing that it is a general perceptual strategy, not restricted to maps and environments.

Errors of rotation

Other systematic geographical errors show that people organize large geographic entities relative to frames of reference as well as to each other. When Stanford students were asked to indicate the direction from Stanford to Berkeley, they erroneously indicated that Stanford was west of Berkeley. They also indicated that Santa Cruz on the coast was west of Palo Alto, again erroneously. People apparently conceive of the southern region of the San Francisco Bay area to be aligned north–south, when in fact it is aligned northwest–southeast. A similar error is observed when people are asked to orient the

shape of South America in a north–south–east–west frame: they 'upright' South America, suggesting that, in their mental representation, the 'natural' axes of the figure are aligned with the axes of the surrounding frame. Like the alignment effect, the rotation effect is a general perceptual organizing principle, appearing for artificial maps and meaningless blobs.

Errors due to hierarchical organization, perspective, and landmarks

There are other systematic errors in spatial memory. The space of navigation is organized hierarchically. For spaces the size of a piece of paper or a sandbox, adults and children make location errors in the direction of the central tendency of the spatial category. Other things equal, the preferred spatial category is a quadrant around a point. For real environments, the spatial categories are often geographic units such as cities, states, and countries. These geographic units affect judgments of distance and direction. Distances within units are underestimated relative to distances between units; directions of entities like cities within units like states are distorted towards the overall direction between the larger units. The perspective from which an environment is conceived also affects distance estimates. In one experiment, people imagining themselves on the east coast of North America estimated the distance between New York City and Pittsburgh to be larger, and the distance from San Francisco to Salt Lake City smaller, than people imagining themselves on the west coast. In both cases, the participants were actually situated in Ann Arbor, so it is the imagined point of view, not the actual point of view, that affects the judgments.

Another striking refutation of the view that maps in the head are like conventional maps on paper comes from the finding that distance estimates from a landmark to an ordinary building are smaller than distance estimates from an ordinary building to a landmark. That is, imagined distances between two points can be asymmetric. This error even extends to metaphoric distances. For example, the similarity of North Korea to China is judged to be less than the similarity of China to North Korea. Likewise, sons are viewed as more similar to fathers than are fathers to sons. Such errors violate an elementary assumption about metric spaces, whereby the distance from A to B must be the same as the distance from B to A.

Together, these errors indicate that the human mind does not contain a library of maps from which distance, direction, route, and other information can be read off. Rather, it seems that a query directs a search for relevant information. That information may turn out to be more or less accurate, complete, consistent, or relevant. Why has the mind developed mechanisms to remember and use spatial information that are guaranteed to produce error? If the mind is an optimizing machine, it must be optimizing many things at once. A mechanism that produces efficient memory – in particular, schematization, both as a way of conserving memory and as a way of integrating disparate items of information – may not optimize accuracy. Erroneous, even self-defeating, actions occur in many domains of human behavior.

Navigation in the Wild

If mental representations of environments are so distorted and incomplete, how is it that people ever succeed in navigating? Actual navigation depends on more than spatial cognition. It is situated in environments that present constraints. If a turn is remembered as a right angle, but the actual angle is less or more, the roads themselves will allow only the correct angle. Similarly, if the distances between landmarks are incorrectly remembered, the landmarks themselves will correct the error. Environments serve as cues, which may elicit information and behaviors that the imagination fails to evoke.

Furthermore, other senses, as well as other information, are operative. Proprioceptive and vestibular information serve actual navigation, but are unlikely to be accessible to cognition. Path integration – that is, maintaining a sense of one's position by integrating distances and turns – while not perfect, can also help travelers. These mechanisms are used by other species as well. Even ants, bees, and rodents correct errors resulting from biased path integration by referring to features of their environments, such as landmarks.

THE SPACE OF EXTERNAL REPRESENTATIONS

Creating external representations as deliberate aids to cognition seems to be a uniquely human behavior. Such representations range from trail markers, cave drawings, tallies, and maps to architects' sketches, economic graphs, organization charts, decision trees, and tables of chemical elements. These cognitive tools capitalize on human adeptness with space, mapping abstract as well as inherently spatial elements and relations onto other spatial elements and relations in meaningful ways. Diagrams and charts use likenesses, icons,

and symbols to represent elements. Examples include icons in computer interfaces and airport signs, and picture-writing in many languages. Diagrams and charts also use space to convey relations among elements, distance in the diagrammatic space representing distance on some other dimension, such as strength or preference. External representations augment the power of the mind by relieving memory and processing, while taking advantage of the mind's ability to understand space, represent it, and make spatial inferences.

CONCLUSION

There are many kinds of spatial thinking. A great deal of human experience is spatial. Each mental space is based on and serves a particular kind of human spatial experience, extracting and schematizing information that is useful for it. These mental spaces are so useful that they subserve thinking in many other domains, including those of emotion, interpersonal interaction, and science. We feel 'up' or 'down'; one nation's culture or language 'invades' or 'penetrates' another; inertia, pressure, and unemployment 'rise' or 'fall'. Even at its most lofty, the mind rests on the concrete.

Further Reading

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press. [On navigation.]

Golledge RG (ed.) (1999) *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. Baltimore, MD: Johns Hopkins Press. [On navigation.]

Kitchin R and Freundschuh SM (eds) (2000) *Cognitive Mapping: Past, Present and Future*, pp. 24–43. London, UK: Routledge. [On cognitive maps.]

Tversky B (forthcoming) Functional significance of visuospatial representations. In: Shah P and Miyake A (eds) *Handbook of Higher-Level Visuospatial Representations*. Cambridge, UK: Cambridge University Press. [On spatial cognition.]

Speaking and Listening Skills

Advanced article

Leonard Abbeduto, University of Wisconsin-Madison, Madison, Wisconsin, USA

CONTENTS

Speaking and listening are conceptualized as goal-directed, collaborative activities that reflect social practices and constraints inherent in the human information processing system. Research has focused on how these activities develop and vary across individuals and cultures.

THE SOCIAL NATURE OF SPEAKING AND LISTENING

Views of speaking and listening have changed dramatically over the past 40 years. In the 1960s, research on speaking and listening was strongly influenced by the linguistic approach of Noam Chomsky (Smith, 1999). In the Chomskyan approach, abstract descriptions of the formal (phonological, syntactic, and semantic) properties of sentences were created to describe the knowledge of language assumed to be 'in the head' of all language users. For Chomsky, speaking and listening were linguistic performances that provided a window into the language user's knowledge, or competence, but they were not in and of themselves interesting. Indeed, the window provided by speaking and listening was thought to be imperfect because they were subject to the limitations of the human information processing system, such as limited memory.

In the 1970s, dissatisfaction with the linguistic approach grew on a number of fronts. Particularly important was dissatisfaction with the exclusion of the social dimensions of speaking and listening. This led to an interest in pragmatics, which is the study of the social uses of language (Levinson, 1983). Early work, or stage one pragmatics, assumed that speakers and listeners followed socially oriented rules to translate their knowledge of phonology, syntax, and semantics into real instances of speaking and listening (Duchan *et al*., 1994). Importantly, stage one pragmatics accepted many of the Chomskyan assumptions (Abbeduto and Short-Meyerson, 2002).

Stage one pragmatics is illustrated nicely by work on speech acts (Levinson, 1983). A speech act is essentially the social function that an utterance is intended to perform, such as requesting information or making an assertion. Speech acts and language forms are not isomorphic. 'That's mine', for example, could request the return of an object or make a statement of fact. In explaining how listeners arrive at different speech acts, stage one theorists (Levinson, 1983) assumed that any utterance has associated with it a literal, or direct, interpretation that is solely a function of the words it contains and the linguistic rules governing their combination. Nonliteral, or indirect, interpretations were thought to be derived from the literal interpretation according to context-dependent pragmatic rules. For example, listeners assume that speakers adhere to the rule 'an utterance must supply new information' and thus, 'that's mine' would be interpreted as a request (its indirect interpretation) rather than as a statement (its direct interpretation) if it was already obvious to the speaker and listener that the object in question belonged to the speaker (Abbeduto *et al*., 1988).

Stage one pragmatics was valuable because it focused on the social nature of speaking and listening. This work, however, suffered from a number of limitations. First, it assumed that knowledge of pragmatic rules is acquired, represented, and accessed independently of knowledge of other components of language despite evidence that pragmatic considerations motivate and shape the acquisition and use of many linguistic forms and vice versa (Abbeduto *et al*., 2001). Second, many so-called pragmatic rules reflect statistical tendencies rather than invariant patterns and, thus, they do not meet the usual definition of a rule. Third, there is a considerable amount of psychological activity that is ignored in the stage one account, such as how listeners determine what contextual information is relevant to deciding between a direct and an indirect interpretation of a sentence. Fourth, many

stage one rules have their origins in constraints on human behavior and social interaction rather than in some independently acquired and represented pragmatic knowledge (Prideaux, 1991). The 'rule' that no more than one person speaks at a time, for example, might simply reflect the difficulty of talking and listening simultaneously. And finally, stage one rules were thought to operate on sentences or smaller linguistic units. In conversation, however, a speaker's goals often can be realized only over an extended sequence of utterances (Abbeduto and Short-Meyerson, 2002).

These criticisms have led to a reconceptualization of speaking and listening, or to stage two pragmatics (Duchan *et al*., 1994). In stage two models, speaking and listening are viewed as reciprocal processes in which the participants collaborate to achieve their goals through sequences of behaviors that are inextricably bound to the dynamically evolving context in which the talk occurs and the practices of the communities to which the participants belong (e.g. Clark, 1996). This reconceptualization of speaking and listening has led researchers to focus on how participants recognize and respond to each other's goals, check on each other's background assumptions and knowledge, and solicit and supply feedback concerning the ways in which each other's utterances have been interpreted (Clark, 1996). It has also led to an expanded interest in the types of information that can serve as relevant contexts for the decisions that speakers and listeners make (Graesser *et al*., 1997).

DEVELOPMENTAL ISSUES

The development of speaking and listening skills is a protracted process that continues well into adolescence (Abbeduto and Short-Meyerson, 2002). Nevertheless, even toddlers can participate successfully in talk across a range of settings and partners, although often this success is due to the support, or scaffolding, provided by adult partners (Ninio and Snow, 1996). In this section, we focus on the developmental course of those aspects of speaking and listening emphasized by stage two models and on the skills and experiences needed to achieve proficiency.

Developmental Course

By two years of age children express and recognize many of the goals that can be achieved through language (Ninio and Snow, 1996). Even elementary school-age children, however, have difficulty recognizing another person's goals when those goals are implied rather than explicitly stated in his or her utterances (Abbeduto *et al*., 1992). There also is improvement well into adolescence in children's ability to use language to achieve their goals when doing so requires that they first achieve a subsidiary goal, such as securing the listener's attention (Anderson *et al*., 1994).

Children are able to solicit information about their partner's background assumptions and knowledge in situationally appropriate ways by the time they are three or four years of age (Short-Meyerson and Abbeduto, 1997). For example, preschoolers are less likely to talk about background experiences (e.g. 'You bake cookies before?') when they and their partners have extensive knowledge of the topic than when either of them has limited knowledge. The extent to which children engage in such talk, however, increases with age (Short-Meyerson and Abbeduto, 1997).

By the age of two years children are able to solicit and supply verbal feedback about the ways in which the utterances of the participants have been interpreted (Abbeduto and Short-Meyerson, 2002). For example, toddlers and preschoolers signal noncomprehension of their partner's messages and respond to such signals from others. However, they are more likely to respond to some types of signals of noncomprehension (e.g. 'What?') than to other types (e.g. 'Which book?'), and not all of their responses are informative. In fact, progress here continues throughout the school years (Abbeduto *et al*., 1997).

Considerable research has demonstrated that children use context to make a variety of decisions about language by two or three years of age, although they become increasingly more sensitive to an ever broader range of contextual information throughout the school years (Abbeduto and Short-Meyerson, 2002). Thus, toddlers and preschoolers tailor the complexity, politeness, and level of detail of their language according to the ability, authority, familiarity, and knowledge of the partner. They also use their partner's immediately preceding utterances as a basis for decisions about what the 'current' utterance means. Toddlers and preschoolers, however, often fail to use context that is not explicitly stated or immediately perceptible (Ackerman *et al*., 1990). There also is steady improvement with age in the effectiveness with which children track changes in context during the interaction (Ricard, 1993).

When considering the development of speaking and listening, it is important to recognize that different genres of talk require different skills and have different developmental trajectories (Ninio

and Snow, 1996). A particularly important genre is narration, or story telling. The development of facility with narrative is thought to provide the foundation for a variety of social and academic skills (Hyter and Westby, 1996). Narrative development begins early but continues into adulthood, with a particularly important shift in early adolescence from a preoccupation with talk about observable events, or a 'landscape of action', to an interest in psychological states, or a 'landscape of consciousness' (Bruner, 1986).

Necessary Skills and Experiences

Developmental progress in speaking and listening is driven in part by achievements in other domains of psychological functioning (Abbeduto and Short-Meyerson, 2002). First, acquisitions in linguistic knowledge (i.e. phonology, syntax, and semantics) ensure that the child has a fuller array of tools for collaborating in talk. Indeed, individuals who have delays in acquiring any or all forms of linguistic knowledge (e.g. those with mental retardation) have pervasive delays in speaking and listening (Abbeduto *et al*., 1998). Second, achievements in speaking and listening depend on children's increasing knowledge of the world and the increasing efficiency of their information-processing systems. For example, children achieve more success in dyadic interactions involving routine events for which they have rich scripts rather than impoverished scripts (Furman and Walden, 1990). Third, the ability to reason about other people's mental states provides the foundation for many achievements in speaking and listening, such as the ability to consider the audience's perspective in narration (Ninio and Snow, 1996) and the use of behaviors for soliciting feedback about how an utterance has been interpreted (Abbeduto *et al*., 1998).

Developmental progress also depends on participating in talk in a variety of settings and with a variety of other people. Particularly important early in development is talk with adults (Ninio and Snow, 1996). Adults provide structure and support that circumvents some of the limitations of young children, thereby ensuring the interaction's success and exposure to 'model' behaviors. Gradually, adult support allows the child to perform behaviors that were previously beyond his or her ken. These interactions with adults are complemented by participation in talk with peers and siblings which allows the child to practice speaking and listening skills that are called for less often in interactions with adults (Dunn and Kendrick, 1982).

INDIVIDUAL AND CULTURAL DIFFERENCES

In addition to age-related differences in speaking and listening, there are differences between individuals of the same age. Not surprisingly, these can be traced in part to differences in other domains of ability that are recruited for use in speaking and listening, such as working memory and theory of mind. In some instances, a child may deviate from average expectations to such a degree that he or she is classified as having a problem that warrants special educational or therapeutic attention, as in specific language impairment and mental retardation. It is interesting to note that in both of these, dissociations in speaking and listening are possible, with speaking usually being more impaired. In the case of mental retardation, the extent and nature of the dissociation varies across etiologies; for example, individuals with Down syndrome are more likely to show poorer speaking than listening skills, whereas such a dissociation is less likely in fragile X syndrome despite the fact that both syndromes can be associated with similar levels of intellectual impairment (Abbeduto *et al*., 2001).

Individual differences in speaking and listening can also be related to cultural differences (Ninio and Snow, 1996). There is evidence, for example, that children speaking African-American English are more likely to convey referential information in their narratives through gesture and prosody than are European-American children, although the narratives of both are equally rich in referential detail (Hyter and Westby, 1996). It is important to recognize that such differences in children's speaking and listening reflect more than exposure to a different language or dialect. They also reflect differences in various practices endorsed or sanctioned by the culture. Crago and her colleagues (e.g. Crago and Eriks-Brophy, 1994), for example, found that parents and teachers in the rural Inuit communities of northern Quebec are more likely to rely on nonverbal modeling than talk when interacting with children compared to adults in French-Canadian or European-American communities. Importantly, such cultural differences do not reflect differences in the adequacy of children's learning environments but rather the simple fact that they must grow up to be members of cultures with different values and practices.

References

Abbeduto L, Davies B and Furman L (1988) The development of speech act comprehension in mentally

retarded individuals and nonretarded children. *Child Development* **59**: 1460–1472.

Abbeduto L, Evans J and Dolan T (2001) Progress in understanding language and communication problems in mental retardation and developmental disabilities. *Mental Retardation and Developmental Disabilities Research Reviews* **7**: 45–55.

Abbeduto L, Nuccio J, Al-Mabuk R, Rotto P and Maas F (1992) Interpreting and responding to spoken language: children's recognition and use of a speaker's goal. *Journal of Child Language* **19**: 677–693.

Abbeduto L, Pavetto M, Kesin E, Weissman M, Karadottir S, O'Brien A and Cawthon S (2001) The linguistic and cognitive profile in Down syndrome: evidence from a comparison with fragile X syndrome. *Down Syndrome Research Quarterly* **7**: 9–15.

Abbeduto L and Short-Meyerson K (2002) Linguistic influences on social interaction. In: Goldstein H, Kaczmarek L and English KM (eds) *Promoting Social Communication in Children and Youth with Developmental Disabilities*, pp. 27–54. Baltimore: Brookes.

Abbeduto L, Short-Meyerson K, Benson G and Dolish J (1997) Signaling of noncomprehension by children and adolescents with mental retardation. Effects of problem type and speaker identity. *Journal of Speech, Language, and Hearing Research* **40**: 20–32.

Abbeduto L, Short-Meyerson K, Benson G, Dolish J and Weissman M (1998) Understanding referential expressions: use of common ground by children and adolescents with mental retardation. *Journal of Speech, Language, and Hearing Research* **41**: 348–362.

Ackerman BP, Szymanski J and Silver D (1990) Children's use of the common ground in interpreting ambiguous referential utterances. *Developmental Psychology* **26**: 234–245.

Anderson AH, Clark A and Mullin J (1994) Interactive communication between children: learning how to make language work in dialogue. *Journal of Child Language* **21**: 439–464.

Bruner JS (1986) *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Clark HH (1996) *Using Language*. New York: Cambridge University Press.

Crago MB and Eriks-Brophy A (1994) Culture, conversation, and interaction: implications for intervention. In: Duchan JF, Hewitt LE and Sonnenmeier RM (eds) *Pragmatics: From Theory to Practice*, pp. 43–58. Englewood Cliffs, NJ: Prentice-Hall.

Duchan JF, Hewitt LE and Sonnenmeier RM (1994) Three themes: stage two pragmatics, combating marginalization, and the relation of theory to practice. In: Duchan JF, Hewitt LE and Sonnenmeier RM (eds) *Pragmatics: From Theory to Practice*, pp. 1–9. Englewood Cliffs, NJ: Prentice-Hall.

Dunn J and Kendrick C (1982) *Siblings' Love, Envy, and Understanding*. Cambridge, MA: Harvard University Press.

Furman LN and Walden TA (1990) Effects of script knowledge on preschool children's communicative interactions. *Developmental Psychology* **26**: 227–233.

Graesser AC, Millis KK and Zwaan RA (1997) Discourse comprehension. In: Spence JT, Darley JM and Foss DJ (eds) *Annual Review of Psychology*, pp. 163–189. Palo Alto: The Annual Reviews, Inc.

Hyter YD and Westby CE (1996) Using oral narratives to assess communicative competence. In: Kamhi AG, Pollock KE and Harries JL (eds) *Communication and Development and Disorders in African American Children*, pp. 247–284. Baltimore: Brookes.

Levinson SC (1983) *Pragmatics*. Cambridge, UK: Cambridge University Press.

Ninio A and Snow CE (1996) *Pragmatic Development*. Boulder: Westview Press.

Prideaux GD (1991) Syntactic form and textual rhetoric: the cognitive basis for certain pragmatic principles. *Journal of Pragmatics* **16**: 113–129.

Ricard RJ (1993) Conversational coordination: collaboration for effective communication. *Applied Psycholinguistics* **14**: 387–412.

Short-Meyerson K and Abbeduto L (1997) Preschoolers' communication during scripted interaction. *Journal of Child Language* **24**: 469–493.

Smith N (1999) *Chomsky*: *Ideas and Ideals*. Cambridge, UK: Cambridge University Press .

Speech Perception, Development of

Intermediate article

Peter W Jusczyk, Johns Hopkins University, Baltimore, Maryland, USA

CONTENTS

To acquire a language infants must be able to perceive its basic speech sounds. Infants are born with excellent abilities to discriminate a wide range of speech sounds. During the course of their first year, these capacities become more specialized for processing the sound patterns of their native language.

INTRODUCTION

Prior to 1970, most researchers identified the beginnings of language acquisition with infants' first attempts at babbling, or even with their first words. Thanks to the development of methods for studying the perceptual capacities of young infants, it is clear that infants categorize speech sounds long before they have begun to produce speech. Early studies sought to characterize the nature of infants' speech perception capacities. Knowledge of these basic capacities led to investigations how and when infants learn about the sound organization (or phonology) of their native language. Recent evidence indicates that in a brief period (the latter half of their first year) infants learn much about native language sound structure from spoken language input.

DISCRIMINATIVE CAPACITIES OF YOUNG INFANTS

The first indication that infants can distinguish speech sounds was reported by Eimas *et al* (1971). They used the high-amplitude sucking procedure to determine whether 1-month-old infants perceive a contrast in two syllables, [ba] and [pa], differing by a single phonetic feature (voicing). Infants were trained to suck on a pacifier to hear a particular speech sound such as [ba]. This training produced a marked increase in sucking rates for several minutes until infants habituated to this sound. At this point, some infants (the control group) continued to hear the same syllable for the next 4 min; for two other groups, a new speech syllable was introduced. For the between-category group, the voicing characteristics of the new syllable were changed so it sounded like [pa] to adults. For the within-category group, an equivalent voicing change was made, but the resulting syllable was one adults perceive as [ba]. Compared with the control group, only the between-category group showed significant increases in sucking for the new syllable. Thus, even 1-month-old infants discriminate the voicing distinction between [ba] and [pa]. Moreover, like adults, infants' discrimination of voicing differences is categorical – infants detect voicing differences between sounds from different phonetic categories, but not voicing differences between sounds from the same phonetic category.

This finding raised questions about the range of infants' abilities and the mechanisms underlying their discriminative capacities. Many subsequent studies investigated infants' abilities to discriminate other types of phonetic contrasts. Infants were shown to discriminate place of articulation contrasts involving stop consonants such as [bae] and [dae] (Eimas, 1974) and fricatives, such as [fa] and [Qa] (Levitt *et al*, 1988), and manner of articulation contrasts between stops and glides such as [ba] and [wa] (Eimas and Miller, 1980a) and stops and nasals such as [ba] and [ma] (Eimas and Miller, 1980b). Moreover, the capacity for discriminating consonant contrasts is present in newborns (Bertoncini *et al*, 1987). Other findings indicate that infants perceive vowel contrasts such as [a] and [i] (Trehub, 1973) and [i] and [I] (Swoboda *et al*, 1976). With respect to the latter contrast, just as for adults, infants' perception of this contrast is continuous, as opposed to categorical. In other words, infants perceive within category differences in vowels.

Infants' discrimination of speech sounds is similar to that of adults in another way, in that it is

robust with respect to acoustic variability in the production of speech. Because of differences in the size and shape of the articulatory systems of different speakers, their productions of the same speech sound will differ acoustically. Moreover, the same speech sound produced by an individual on different occasions varies, depending on such factors as speaking rate and emotional state. Thus, identifying the same word spoken on different occasions entails ignoring the kinds of acoustic differences among speech sounds that are irrelevant for making phonetic distinctions. Kuhl (1979) first demonstrated that 6-month-old infants are able to make phonetic distinctions, even when these are produced by a variety of different male and female speakers. Furthermore, infants as young as 2 months can make phonetic distinctions, even in the face of acoustic variability produced by different speakers (Jusczyk *et al*, 1992) or by changes in speaking rate (Eimas and Miller, 1980a).

SENSITIVITY TO NATIVE AND NON-NATIVE CONTRASTS

Since the first indications of speech discrimination by infants, researchers have been interested in the extent to which infants' discriminative capacities are driven by experience. Because fetuses in the last trimester respond to sounds, and because some speech information is transmitted through the womb, even newborn infants cannot be said to lack all experience of speech sounds from their native language. Thus, assessing the effects of experience requires the use of speech sound differences that are not present in the native language. Streeter (1976) reported that 2-month-old infants from Kikuyu-speaking environments can discriminate a [ba]–[pa] difference, even though these sounds do not occur in their native language. Similarly, Trehub (1976) found that infants 1–4 months old from English learning environments discriminated two other non-native contrasts, [pa]–[pã] and [r&a]–[za]. Findings from these and other investigations (e.g. Lasky *et al*, 1975; Aslin *et al*, 1981) suggest that prior to 6 months, infants are able to discriminate non-native speech contrasts.

However, experience plays a more important role in the latter half of the second year. Specifically, infants show a declining sensitivity to certain non-native contrasts. This developmental change in speech discrimination abilities was first noted by Werker and Tees (1984). They found that although English learners could discriminate non-native contrasts from Hindi and Nthlakapmx at 6–8 months, by 10–12 months they no longer discriminated these contrasts. This decline in sensitivity to non-native contrasts has also been observed for other contrasts (Tsushima *et al*, 1994; Best *et al*, 1995). However, discrimination of some types of non-native contrasts, such as Zulu clicks, is maintained into adulthood (Best *et al*, 1988). Why lack of experience causes a decline in sensitivity to some, but not all, non-native contrasts is not fully understood. However, an important factor appears to be how similar the non-native contrasts are to ones in the native language (Best, 1995). Contrasts that map to distinct native language phoneme categories tend to be easily discriminated, whereas non-native contrasts that map to the same phoneme category are poorly discriminated.

DEVELOPMENT OF WORD SEGMENTATION ABILITIES

Because words in fluent speech are usually produced continuously without clear pauses between them, some ability to segment words is critical for understanding language. It appears that listeners use information about the sound organization of their native language to locate the boundaries of words in speech. Among the possible cues for locating word boundaries are (a) consistent word stress patterns (i.e. prosodic cues); (b) the frequency with which certain combinations of phonetic segments can occur at the beginning and end of words (i.e. phonotactic constraints); and (c) restrictions on the positions that particular variants of a phoneme can occupy within a word (i.e. allophonic cues). The relative importance of each of these cues varies across languages depending on their particular sound organization.

Jusczyk and Aslin (1995) reported that English-learning infants aged 7.5 months but not those aged 6 months, display some capacity for segmenting words from fluent speech. The 7.5-month-olds gave evidence of segmenting words, regardless of whether they were familiarized with a pair of words presented in isolation and then tested on their recognition of the words in passages, or whether they were familiarized with the passages first, then tested on isolated words. Subsequent investigations have explored the source of infants' word segmentation abilities. Prior research had shown that, between the ages of 6 and 9 months, English learners develop sensitivity to prosodic and phonotactic cues that could potentially signal word boundaries (Jusczyk *et al*, 1993 a,b). When beginning to segment words, English-learning 7.5-month-olds rely heavily on prosodic cues (Jusczyk *et al*, 1999b). Hence, they segment words

with the predominant word stress pattern of English (initial stressed syllable followed by an unstressed syllable), but they do not segment words with less frequent stress patterns until the age of 10.5 months. The latter ability depends on the infants' use of other information about word boundaries, such as phonotactic and allophonic cues. English learners can use phonotactic cues at about 9 months (Mattys *et al*, 1999; Mattys and Jusczyk, in press) and allophonic cues at about 10.5 months (Jusczyk *et al*, 1999a).

TRACKING STATISTICS AND LEARNING PATTERNS

In order to learn which sequences of sounds are possible in words in a language, infants must track the frequency of occurrence of particular combinations of sounds. As noted in the previous section, 9-month-old infants distinguish between sequences of sounds that occur frequently in the input versus those that only occur infrequently (Jusczyk *et al*, 1994). However, what is remarkable is how rapidly infants can learn about statistical regularities in the input. Saffran *et al* (1996) presented 8-month-olds with a continuous stream of speech (about 2.5 min long) composed of four different trisyllables, which were randomly ordered with the exception that the same item never followed itself in the sequence. Within each trisyllable (e.g. 'golatu'), the order of each syllable was perfectly predictable, i.e. the first syllable 'go' always was followed by the same second syllable 'la', which was always followed by the same third syllable 'tu'. However, the predictability of the syllable that followed the third syllable was only 0.33 (because each trisyllable was followed equally often by one of the other three trisyllables). This low point in the transitional probabilities of adjacent syllables provides cues to the organization of elements within the speech stream. During the test phase, infants heard two of the original trisyllables and two new trisyllables (composed of the last syllable of one trisyllable and the first two syllables of another). The infants treated the original trisyllabic items as familiar, suggesting that they had segmented the speech stream using the statistical cues. A follow-up study indicated that infants can also track statistical regularities for sequences of nonspeech sounds such as tones (Saffran *et al*., 1999). Other studies show that infants can detect relations more abstract than those having to do with how often one particular item follows another. In particular, Marcus *et al* (1999) found that 8-month-olds learn about relations among elements of a three-item sequence, even when the particular elements in the sequence change across trials. One-year-olds can even detect patterns in longer strings that vary in their overall length, just as the number of words in a sentences vary (Gomez and Gerken, 1999).

PROCEDURES FOR STUDYING INFANT SPEECH PERCEPTION

Several procedures have been used to investigate infants' speech perception capacities. The high-amplitude sucking procedure provides information about infants' discriminative capacities between birth and 4 months. Infants suck a sterilized nonnutritive nipple connected to a pressure transducer, allowing experimenters to monitor sucking rates in response to speech sounds. Among the limitations of this procedure are that it allows for only one type of contrast to be presented per experimental session and that it yields measures of group, rather than individual, performance. The operant head-turn procedure is typically used between 6 months and 12 months of age. Infants are trained to turn their heads in the direction of a visual display box whenever they detect a sound change in a stream of repeating syllables. Training and testing may require several test sessions, but this procedure can provide a measure of individual performance. One limitation of this procedure is only brief stimuli (about 1 s) can be used. The head-turn preference procedure is used between the ages of 4.5 months and 24 months. Infants are seated in a three-sided enclosure with a light mounted on the panels to the left and right of the infant. If infants turn to the light when it flashes, speech samples are played either to completion or until the infant turns away for 2 s. Advantages of this procedure are that it allows for the presentation of varied types of materials and longer speech samples (e.g. 45 s) Measures of individual performance can be obtained. One limitation of the procedure is that lack of a significant listening preference may be due either to a failure to discriminate the materials or to a lack of interest in them. For further details of these three procedures, see Polka *et al* (1995). Another method used with infants aged 1–2 years is the preferential looking procedure. Infants view a side-by-side video display of two objects. Infants' looking responses to each object are monitored in response to the presentation of an audio stimulus, such as a word naming one of the objects. This measure yields information about individual performance and has proved useful in studies of word learning and word recognition in

older infants; see Fernald *et al* (1998) for further details.

Acknowledgement

The publishers would like to thank Richard Aslin, Professor of Brain and Cognitive Sciences, University of Rochester, for his help in proofreading and revising the text of this article.

References

Aslin RN, Pisoni DB, Hennessy BL and Perey AJ (1981) Discrimination of voice onset time by human infants: new findings and implications for the effects of early experience. *Child Development* **52**: 1135–1145.

Bertoncini J, Bijeljac-Babic R, Blumstein SE and Mehler J (1987) Discrimination in neonates of very short CV's. *Journal of the Acoustical Society of America* **82**: 31–37.

Best CT (1995) Learning to perceive the sound patterns of English. In: Rovee-Collier C and Lipsitt LP (eds) *Advances in Infancy Research*, pp. 217–304. Norwood, NJ: Ablex.

Best CT, McRoberts GW and Sithole NM (1988) Examination of the perceptual re-organization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance* **14**: 345–360.

Best CT, Lafleur R and McRoberts GW (1995) Divergent developmental patterns for infants' perception of two non-native contrasts. *Infant Behavior and Development* **18**: 339–350.

Eimas PD (1974) Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics* **16**: 513–521.

Eimas PD and Miller JL (1980a) Contextual effects in infant speech perception. *Science* **209**: 1140–1141.

Eimas PD and Miller JL (1980b) Discrimination of the information for manner of articulation. *Infant Behavior and Development* **3**: 367–375.

Eimas PD, Siqueland ER, Jusczyk PW and Vigorito J (1971) Speech perception in infants. *Science* **171**: 303–306.

Fernald A, Pinto JP, Swingley D, Weinberg A and McRoberts GW (1998) Rapid gains in the speed of verbal processing by infants in the second year. *Psychological Science* **9**: 228–231.

Gomez RL and Gerken LA (1999) Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition* **70**: 109–135.

Jusczyk PW and Aslin RN (1995) Infants' detection of sound patterns of words in fluent speech. *Cognitive Psychology* **29**: 1–23.

Jusczyk PW, Pisoni DB and Mullennix J (1992) Some consequences of stimulus variability on speech processing by 2-month old infants. *Cognition* **43**: 253–291.

Jusczyk PW, Cutler A and Redanz N (1993a) Preference for the predominant stress patterns of English words. *Child Development* **64**: 675–687.

Jusczyk PW, Friederici AD, Wessels J, Svenkerud VY and Jusczyk AM (1993b) Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* **32**: 402–420.

Jusczyk PW, Luce PA and Charles Luce J (1994) Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* **33**: 630–645.

Jusczyk PW, Hohne EA and Bauman A (1999a) Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics* **61**: 1465–1476.

Jusczyk PW, Houston D and Newsome M (1999b) The beginnings of word segmentation in English-learning infants. *Cognitive Psychology* **39**: 159–207.

Kuhl PK (1979) Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America* **66**: 1666–1679.

Lasky RE, Syrdal-Lasky A and Klein RE (1975) VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology* **20**: 215–225.

Levitt A, Jusczyk PW, Murray J and Carden G (1988) The perception of place of articulation contrasts in voiced and voiceless fricatives by two-month-old infants. *Journal of Experimental Psychology: Human Perception and Performance* **14**: 361–368.

Marcus GF, Vijayan S, Rao SB and Vishton PM (1999) Rule learning by seven-month-old infants. *Science* **283**: 434–435.

Mattys SL and Jusczyk PW (in press) Phonotactic cues for segmentation of fluent speech by infants. *Cognition*.

Mattys SL, Jusczyk PW, Luce PA and Morgan JL (1999) Word segmentation in infants: how phonotactics and prosody combine. *Cognitive Psychology* **38**: 465–494.

Polka L, Jusczyk PW and Rvachew S (1995) Methods for studying speech perception in infants and children. In: Strange W (ed.) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research*, pp. 49–89. Timonium, MD: York Press.

Saffran JR, Aslin RN and Newport EL (1996) Statistical learning in 8-month-old infants. *Science* **274**: 2926–1928.

Saffran JR, Johnson EK, Aslin RN and Newport EL (1999) Statistical learning of tone sequences by human adults and infants. *Cognition* **70**: 27–52.

Streeter LA (1976) Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature* **259**: 39–41.

Swoboda P, Morse PA and Leavitt LA (1976) Continuous vowel discrimination in normal and at-risk infants. *Child Development* **47**: 459–465.

Trehub SE (1973) Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology* **9**: 91–96.

Trehub SE (1976) The discrimination of foreign speech contrasts by infants and adults. *Child Development* **47**: 466–472.

Tsushima T, Takizawa O, Sasaki M *et al* (1994) Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6–12 months: language specific developmental changes in speech perception abilities. *International Conference on Spoken Language Processing*, Yokohama, Japan, pp. 1695–1698.

Werker JF and Tees RC (1984) Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* **7**: 49–63.

Speech Production

Introductory article

Gabriella Vigliocco, University College London, London, UK
David P Vinson, University College London, London, UK

CONTENTS

'Speech production' refers to the processes by which speakers make their thoughts and intentions known to listeners through the use of language. This includes retrieving information from memory (e.g. words), integrating different parts of information into a coherent whole (e.g. sentences), and physically producing the sounds that are required to realize an utterance.

INTRODUCTION

How can humans, by exercising their vocal organs, utter words in such a manner as to convey a meaning to a listener? This is the basic question addressed by speech production research. More specifically, the question concerns the cognitive representations and processes that underscore our ability to put a thought, an intention we want to convey, into action (a sequence of sounds that expresses that intention). Although producing language is most commonly achieved by using our vocal cords, language may be produced by other forms of action; for example, sign languages use a visual instead of an acoustical modality for their expression. This article addresses the cognitive processes involved in going from the intention to preparing a plan for articulation; motor development and execution of this plan are not addressed.

RESEARCH METHODS IN SPEECH PRODUCTION

Adult English speakers know between 30,000 and 80,000 words. The average for high school graduates has been estimated at 45,000 words. In sentences these words can be arranged in a large, possibly infinite number of ways conforming to the linguistic conventions (both grammar and phonology) of English. Sentences are built in under two seconds, and spontaneously occurring errors during production may be as rare as one out of 1000 words, indicating that production is both efficient and accurate. However, errors and dysfluencies do occur: average speakers spend half of their speaking time not speaking but hemming, hawing, and pausing; such disruptions occur between three and 12 times per minute.

Important theoretical developments in our understanding of the cognitive machinery that underlies speech have been achieved since the early 1970s. The research methods that have marked these developments can be grouped in two camps: naturalistic observations and experimental methods.

Naturalistic Observations

Two spontaneously occurring phenomena have been the focus of research: slips of the tongue, and dysfluencies. Some of the most famous and most quoted speech errors were made by Reverend William Spooner (1844–1930). He once told his congregation that 'the Lord is a shoving leopard' and his students that they had 'tasted the whole worm'. Although it is unclear whether Spooner's errors were genuine, true errors (also referred to as slips of the tongue) do occur; pioneers of speech production such as Victoria Fromkin and Merrill Garrett have collected large bodies or corpora of slips.

Such investigations assume that errors occur when mapping between representations goes wrong. By investigating the regularities in such failed mappings we can learn about the nature of the representations that are involved.

Compare '… something for my *shoulders*' [intended: elbows] to '… sat on an *envelope*' [intended: elephant]. In both cases a word is erroneously replaced by another, but in the first, the target and intruding words are similar in meaning, while in the second they are similar only in their sound. Because word substitution errors are similar in either meaning or form, researchers have

postulated that words are retrieved in two steps when we speak: first we look for the meaning we want to express, then we look for a word form corresponding to that meaning.

As another example consider 'like a *l*ilting *w*illy' [intended: a wilting lily] and 'a *mother* to my *letter*' [intended: a letter to my mother]. In these errors, two elements are exchanged, but the type of unit is different: phonemes (/w/ and /l/) in the first error, and words ('letter' and 'mother') in the second. Different constraints govern these different exchanges: sound exchanges occur between similar phonemes from the same syllabic position; word exchanges occur between words of the same grammatical class (e.g. nouns). Converging evidence has also been obtained by analyses of the speech of speakers who have become aphasic as a consequence of brain damage; errors by these speakers parallel those by non-language-impaired speakers. From the observation that different types of linguistic information are integrated in different ways, researchers have inferred that there are different levels at which stored linguistic information is integrated to form sentences.

Errors are not the only type of disruption of speech: dysfluencies, including filled pauses such as 'um' and 'uh', empty pauses, hesitations, and repetitions are also well represented. Can dysfluencies tell us something about production? The assumption here is that dysfluencies reflect ongoing difficulties, and speakers' anticipation of upcoming difficulties in utterances they are preparing. The properties of dysfluencies, particularly the points at which they occur, and how they are resolved, can reveal the nature of the elements that are involved in particular levels of language processing. For example, pauses in speaking have been found to occur not only when a speaker reflects upon what to say next, but also sometimes as an indicator of difficulty in retrieving an upcoming word.

Experimental Studies

Naturalistic data are difficult to come by, given the overall accuracy of fluent speech; the corpora of slips mentioned above took years to collect and, because they occur spontaneously, we cannot control the factors determining them. In order to investigate language production in more controlled settings, naturalistic data are complemented by laboratory experimentation.

First, following the logic of investigating the properties of naturally occurring slips of the tongue and dysfluencies, one line of experimentation is inducing errors in laboratory settings. This can be accomplished by asking participants to repeat tongue twisters, or to perform speaking tasks under time pressure. Errors produced in such settings exhibit similar patterns of constraints to spontaneously occurring errors; researchers also have the advantage of knowing precisely what the intended utterance was in each case. Most crucially, experimental error induction allows experimenters to systematically manipulate those aspects of utterances that are thought to be relevant to a particular error type.

Experimental research on language production also allows investigation of when various elements of language processing come into play. This is generally carried out through vocal reaction time experiments, in which participants are presented with a stimulus (usually a picture of an object or a printed word) and asked to name (or read) it aloud, and the time between presentation and the beginning of the spoken response is measured. By presenting such naming tasks within a sentential context, it can be determined whether, and to what extent, the context affects the time it takes to produce a word. Such effects can be facilitatory (speakers are fast in naming a pictured spoon after reading a predictable sentence context like 'He ate the soup with a …') or inhibitory (speakers are slower in naming a pictured object, e.g. 'dog', if presented in the context of a meaning-related distractor word, e.g. 'cat').

Within the realm of producing phrases and sentences, Kathryn Bock introduced experimental techniques to 'prime' speakers, that is, to induce preferences in speakers to use a particular type of structure (e.g. instead of 'Lightning struck the church', to say 'the church was struck by lightning'), demonstrating the influence of various factors on the flexibility speakers have to express a given idea. For example, priming a word often results in that word appearing earlier in the sentence (priming 'church' is more likely to elicit a passive sentence like 'The church was struck by lightning'), suggesting that words' availability influences the syntactic structure that is ultimately produced.

THE ARCHITECTURE OF THE PRODUCTION SYSTEM

Using these different tools, researchers converge in depicting speech production as a process that involves both retrieving stored information from memory and integrating this information into utterances at different levels, each characterized by specific linguistic properties. Figure 1 provides an

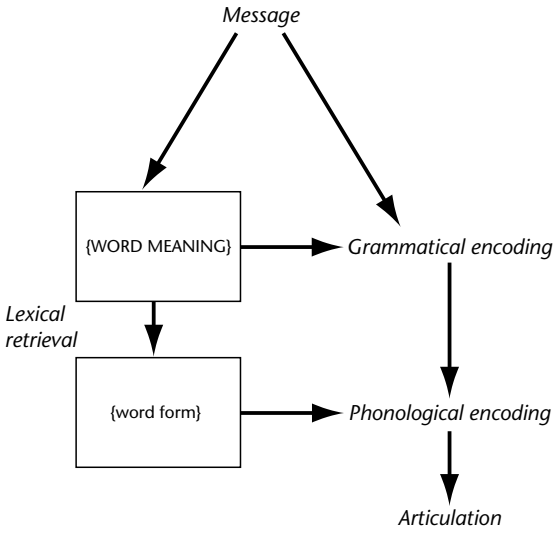


Figure 1. An overview of the speech production system.

outline of the processing steps mediating between intention and articulation. The processes are broadly divided into *grammatical* and *phonological* encoding. As Kathryn Bock described it, grammatical encoding can be conceived as building a skeleton for a sentence; phonological encoding, as fleshing the skeleton out by putting sounds into it.

Retrieving Information from Memory

It is uncontroversial that linguistic information is stored in memory; production research asks what is stored and how the retrieval process proceeds.

Within the stored information is included the intended word's meaning and its sound pattern (or word form); but because we use words in ways that must conform to the grammar of the language, syntactic information (i.e. information on how to use the word appropriately in a sentence) must also be included. How is the retrieval of these different types of information orchestrated in real time?

Retrieving words from memory involves two steps: during a first step a word corresponding in meaning to the intended concept (as represented at the message level) is retrieved. This retrieval process is competitive: other words similar in meaning are also activated and occasionally intrude, substituting for the intended word (e.g. saying 'shoulders' instead of 'elbows'). It is also generally assumed that syntactic properties of words are retrieved as a consequence of this process. During the second step, the word's phonological make-up is retrieved; when this second step goes wrong, a word similar in form is mistakenly retrieved (e.g. saying 'envelope' for 'elephant').

Further evidence for the sequentiality of the two retrieval steps comes from studies in which participants are asked to name a picture of an object, while they are presented with a distracting word. When the distracting word is presented immediately before the picture, it interferes with the retrieval of the name of the picture only if it is related in meaning; words related in form have no effect at this point in time. When the distracting word is presented immediately after the picture, it facilitates the retrieval process only if it is related in form; words related in meaning have no effect.

Such sequentiality is further demonstrated by another type of derailment in speech – the tip-of-the-tongue (TOT) phenomenon: the common experience of knowing a word but being unable to say it. Speakers experiencing TOTs know the meaning they want to express, and often also have partial information about the word's sound (e.g. it starts with a 't') although they cannot retrieve the full word form. Speakers experiencing TOTs can also report syntactic information about the word even when unable to report any word-form information; this finding suggests that syntactic information is retrieved before word form.

Besides information concerning whole words, other types of linguistic information are also assumed to be stored. This includes the specific phonemes of a given language (e.g. /m/, /u:/, /n/ as in 'moon', etc.). A further type of stored information may be the syllable. The number of different syllables in a language is not so high; for example, in English approximately 6600 syllables are sufficient to cover 38,000 different word-types. According to Willem Levelt, this fact suggests that syllables are stored as a whole. In what format would they be stored? As we will see below, this is most likely to be in terms of the action patterns, so-called articulatory gestures (movements of the lips and jaw) necessary to produce syllables.

How is the retrieval of all these different types of information orchestrated? It is agreed upon that the retrieval process is sequential: initial retrieval is meaning-based, followed by information about syntactic properties, then the word form is retrieved, then the phonemes, and finally the corresponding articulatory gestures. Theories, however, differ with respect to whether the flow of information from one level to the next is strictly unidirectional; such debate has arisen as a consequence of finding form influences on meaning-based

retrieval, as described by Gary Dell. One such influence has been observed in substitution errors. Mixed errors, in which the intended word and the produced word are similar in *both* meaning and form (e.g. 'pass me a *melon*' [intended: 'lemon']) seem to be more common than we should expect by chance. Such errors suggest that the form information can affect meaning-based retrieval, or in other words, that the retrieval process is interactive.

Integrating the Stored Information

Retrieving stored information is necessary but not sufficient to ensure the production of utterances. This information needs to be integrated in ways constrained by the current speaker's intentions and by linguistic conventions. One core linguistic convention is the grammar of a language, which, broadly speaking, can be considered as procedural knowledge concerning how words can be used in sentences.

All languages have a grammar, but different languages have different specific grammatical requirements. For example, consider the sentence 'The ferocious dog bites the boy'. In English the subject ('dog') must precede the verb ('bites') and the object ('boy'), and adjectives must precede nouns ('ferocious dog'). In Italian, however, the subject of the sentence may precede or follow the verb, and adjectives must follow nouns; hence in Italian the example above could be the equivalent of 'bites the dog ferocious the boy', which is strictly ungrammatical in English.

In parallel to the sequence of steps in retrieving different types of linguistic information from memory, a sequence of different encoding processes is also generally agreed upon by speech production theories. Because speakers produce two to three words per second in a fluent and efficient manner, it is important also to consider how such a system can ensure fluency and efficiency.

In order to speak fluently, a speaker must produce words, phrases, and sentences arrayed in time, flowing smoothly from one to the next. If a speaker must prepare each utterance completely in advance before producing it, speech would be marked by frequent 'planning pauses'. Yet such pauses are remarkably rare; they can be avoided through incremental and parallel processing of information; that is, encoding processes can begin as soon as a minimal amount of information is available, and multiple levels of processing can be in operation at once, to allow a speaker to begin uttering a sentence before its entire contents are planned. If this is to be achieved most efficiently, the units of information at a given level of processing should be as small as possible; such retrieved information is combined at several distinct levels of processing.

Mapping from cognition to syntax: grammatical encoding

The first step in going from 'mind to mouth' entails building a syntactically specified frame starting from the events specified in the message.

Assuming that the speaker wants to express the event 'The ferocious dog bites the boy', the stored meaning-based representations for 'ferocious', 'dog', 'to bite', and 'boy' are retrieved. These representations would further specify that 'dog' and 'boy' are nouns (hence usable as subject and object), 'to bite' is a verb (hence specifying a relation between the nouns), and 'ferocious' is an adjective (hence it can modify a noun). These representations also specify that 'dog' and 'boy' are common, countable nouns, and may be pluralized, and that 'to bite' may be used to refer to the past, present, or future, depending upon the speaker's intentions. The message-level information controls the assignment of the different lexical elements to their intended grammatical functions (e.g. 'dog' is assigned to be the subject, and 'boy' as the object and not vice versa, which would result in the unintended sentence: 'The boy bites the ferocious dog'). Message-level information also ensures that the appropriate specification of number for the noun 'dog' (singular) and tense for the verb 'to bite' (present) are passed along.

Functional-level processes are further responsible for establishing relationships among words corresponding to relationships among the participants in the event. In the example, at this level agreement is computed between the (singular) subject of the sentence and the verb (which consequently must also be singular). The intermediate representation for the utterance at this point does not strictly specify the order in which words will appear in the sentence, nor is it phonologically specified. This latter claim is motivated by the fact that exchange errors at this level are not constrained by the words' phonological specifications, while they are constrained by the words' grammatical properties. Grammar has been introduced as procedural knowledge, somewhat separable from information about words. Evidence for this distinction comes from the finding that syntactic structures can persist (i.e. speakers tend to reuse the same grammatical functions, regardless of the lexical content). For

example, if a speaker has just produced the sentence 'Jane gave the girl a present' it is more likely that he or she will subsequently produce 'John lent the mechanic money' instead of the perfectly grammatical alternative 'John lent money to the mechanic'. These processes, part of what we label 'grammatical encoding', are illustrated in Figure 2. This figure illustrates grammatical encoding processes involved in one possible way of expressing an event in which a dog bites a boy.

We speak one word at a time, and for each word, one phoneme at a time. Bridging from the domain of meaning and syntax to the domain of phonology, the order of words in the to-be-uttered sentence must first be specified. In English this requires that the adjective 'ferocious' be placed before the noun, while it must be placed after the noun in Italian. Evidence compatible with such a process comes from studies showing that word order can persist independently from grammatical functions. At this level, which we consider as part of grammatical encoding, as illustrated in Figure 2, a frame is built for a to-be-uttered sentence, specifying closed class morphemes (e.g. determiners, auxiliaries, prepositions, and also inflections, such as *-s* for a plural noun, *-ed* for a past tense verb). Open class morphemes (e.g. nouns, verbs, adjectives) are inserted in this frame. This distinction between

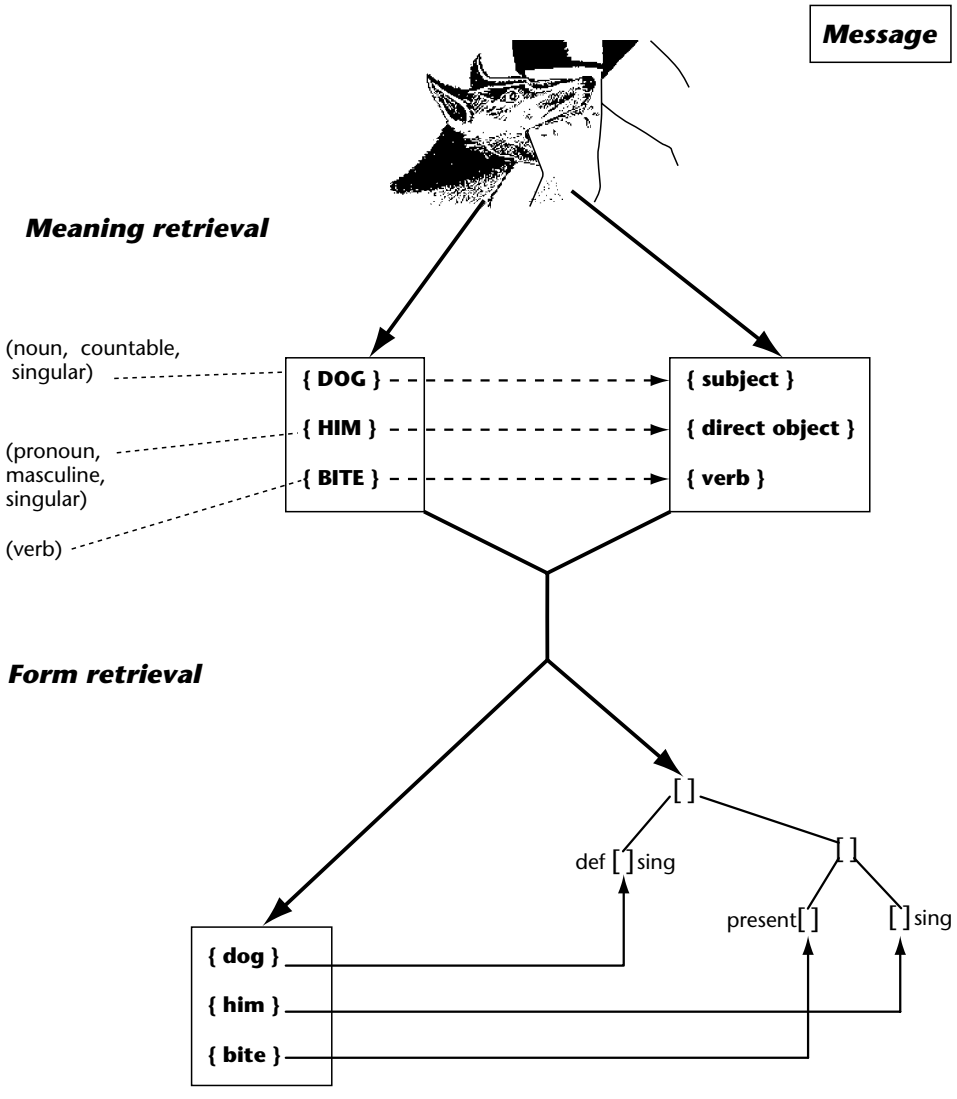


Figure 2. Grammatical encoding. On the basis of the speaker's message, abstract units specifying words' meanings are retrieved and assigned to the grammatical functions to be expressed. A sentence's syntactic structure is created on the basis of these functions, and word order is spelled out. At this point, abstract units representing word forms are filled into the sentential slots, allowing phonological encoding to begin.

vocabulary type (open/closed class) is supported by the finding that the two types of vocabulary can break down selectively in aphasia. The frame-content (or slot-filler) nature of the process is supported by exchange errors such as '… cut *rains* in the *tree*' [intended: cut trees in the rain], in which 'tree' and 'rain' swapped position. However, the plural marker *-s* of 'trees' stayed in place, suggesting that it is part of a frame in which morphemes are inserted.

Mapping from syntax to sounds: phonological encoding

Once the frame is filled, the next step consists of interpreting it in phonological terms, as illustrated in Figure 3. This involves both specifying the sounds to go in each word (the content), as well as specifying sound characteristics for the whole sentence (the frame).

Details of the frame's specification are controlled by the speaker's intentions and by linguistic conventions. For example, do I shout or whisper? Fast or slow? Fast speech is more likely to use contractions such as 'I've' instead of 'I have'. It is influenced by the speaker's emphasis in the sentence (compare 'are you coming?' to 'are you COMING?'). Furthermore, the specific words to be used force differential realization of syllables. For example, in 'the dog bites the boy', 'bites the' is pronounced with each word in its own syllable, 'bites-the', while in the sentence 'the dog bites him', the phrase 'bites him' may resyllabify into the single phonological word 'bite-sim'. Rising and declining intonation marks each phonological word, as well as larger (phrasal) chunks at this level.

Why would speakers go through the trouble of resyllabifying? First, syllables mark the rhythm of the unfolding utterance, which gives speech its specific 'musical' quality which has been argued to be particularly important in determining turn-taking during speaking. Without resyllabification, the rhythm would be broken. Second, syllables may be stored as articulatory gestures to reduce the burden of executing utterances.

The developed phonological frames are then filled with the specific phonemes. The process of filling the phonological frames with sounds is similar in nature to the process of filling the

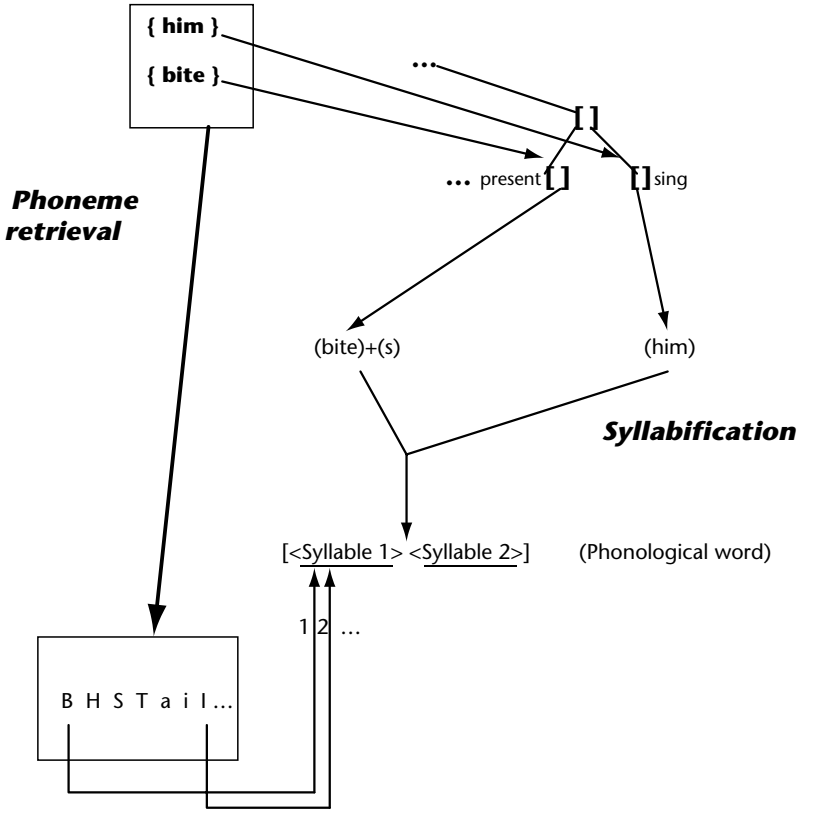


Figure 3. Phonological encoding. Abstract representations specifying word forms are inserted into the respective sentential positions. Appropriate phonemes are assigned to their appropriate positions once syllabification is carried out, and the subsequent phonologically specified syllables are then output for articulation.

syntactic frame with words as described above. A main source of evidence comes, once again, from exchange errors. Errors like '*h*eft *l*emisphere' [intended: 'left hemisphere'] indicates that phonemes can be misplaced during the filling process. Where they end up is not random; it is constrained by the elements' syllabic position. In the example, both phonemes occur as syllable (and word) onsets. In parallel to the grammatical class constraint for word substitution errors, the syllable constraint for sound errors suggests that phonemes' stored representations include information about where those phonemes can be used within a syllable. Note that error evidence suggests that retrieved phoneme representations are still abstract at this point in processing. In the error '*sp*ace' [intended: 'place'], 'p' has a slightly different sound in the error context; 'p' has accommodated to the new environment.

How does the slot-filling process proceed? Evidence from reaction time studies suggests that it proceeds in a strictly left-to-right order, consistent with the temporal requirements of speaking.

ARTICULATION

Syllables are comfortably articulated at a rate of six or more per second, calling on more muscle fibers (approximately 100) than may be required for any other mechanical performance of the human body. These muscles are distributed over three anatomically distinct structures, all of which subserve other functions beyond speech alone.

The *respiratory system* controls the steady outflow of air during speech. It has its own mode of functioning during speaking that differs markedly from non-speaking breathing. The *laryngeal system*, with the vocal folds as its central part, controls voicing and loudness of speech. During voicing it generates a periodic train of air puffs which provides a large frequency spectrum on which resonation builds. The *supralaryngeal system* (vocal tract) contains the chambers in which resonation develops, in particular the nasal, oral, and pharyngeal cavities. Their shapes determine the timbre or tone quality of vowels and consonants. Moreover, the vocal tract is the main determinant of the articulation of speech segments. Constriction of the vocal tract in different places (compare 'pa' and 'ga') and in different manners (compare 'pa' and 'ba') provides us with the large variety of speech sounds we produce.

Articulation is context-dependent. It was mentioned above that it is generally agreed that when sounds are inserted into phonological words they are abstract, namely their physical properties are not specified for their insertion environment. But context-dependency is clear. For example, compare the way in which the phoneme /t/ is pronounced in 'top' and in 'stop'. Such context-dependency does not only occur within words (if this were the case, it could be stored) but also arises between words in a sentence (note the variation in the sound of 't' of 'take' between 'Let's take it' and 'He did take it'), and goes in both directions: a sound is pronounced differently depending upon both the preceding and the following context (compare the differing sounds of 't' in 'went by' and 'went from'). In technical terms, such context-dependency has been referred to as 'co-articulation'.

A number of theories have been developed concerning motor control of speech. In these theories, major questions concern the motor commands and how they are executed. The existence of context-dependency has played a large role in this theoretical development. However, some degree of context-independence is argued for by most theories. At a general level some degree of context-independence is desirable, because otherwise we could not explain how speakers immediately adapt their motor commands to successfully speak while having food in the mouth. Context-dependency arises as an adjustment to the physical environment during the execution of a motor plan. For example, producing a 'b' may be specified as a command to close the lips with a certain force – however, *how* to reach this goal, depending upon both jaw and lip movement, is not specified and may depend on whether the speaker has something in his/her mouth. If this is the case, then the question concerns the characterization of the abstract motor commands.

Further Reading

Bock JK (1995) Sentence production: from mind to mouth. In: Miller JL and Eimas PD (eds) *Speech, Language and Communication: Handbook of Perception and Cognition*, vol. 11, pp. 181–216. San Diego, CA: Academic Press.

Bock JK and Levelt WJM (1994) Language production: grammatical encoding. In: Gernsbacher MA (ed.) *Handbook of Psycholinguistics*, pp. 945–984. San Diego, CA: Academic Press.

Caramazza A (1988) Some aspects of language processing revealed through the analysis of acquired aphasia: the lexical system. *Annual Review of Neuroscience* **11**: 395–421.

Cognition (1992) Vol. 42: Special issue on lexical production (ed. WJM Levelt).

Dell GS (1986) A spreading activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–321.

Fowler CA (1996) Speaking. In: Heuer H and Keele SW (eds) *Handbook of Perception and Action*, vol. 2: *Motor skills*, pp. 503–560. San Diego, CA: Academic Press.

Garrett MF (1993) Errors and their relevance for models of language production. In: Blanken G *et al*. (eds) *Linguistic Disorders and Pathologies*. Berlin, Germany: de Gruyter.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Statistical Methods: Overview

Introductory article

John K Kruschke, Indiana University, Bloomington, Indiana, USA

CONTENTS

Statistical methods specify how confident we can be to confirm or deny a claim.

THE PURPOSE OF INFERENTIAL STATISTICS

Our knowledge about the world is based on a finite set of observations of it. We cannot know with certainty about the true underlying state of the world; instead, we must infer what might be true, or what is probably false, given the limited data we have. The purpose of inferential statistics is to give precision to our confidence in claims about the world.

Examples of the Need for Inferential Statistics

Suppose that someone claims that Monica is a better tennis player than Conchita. How can the claim be tested? If we could somehow directly ascertain the absolute expertise of both players, by applying some kind of 'expertisometer' to their foreheads, then we would have proof about the claim, either for or against. But we have no such direct method. Instead, we can only let the players compete in a finite number of games, and observe which player wins more games.

Suppose the competitors play a single game, and Monica wins. Does this outcome prove that Monica is better? No, of course not. There are numerous random influences in any given game, such as wind gusts, noises from the audience, small bumps on the court surface, and so on. To try to equalize the influence of these random factors on both players, we have them play several games on different courts.

Suppose that, after seven games, Monica has won four. Is Monica definitely the better player? A supporter of Conchita could argue that Monica just got lucky and happened to win four out of seven, even though Conchita is actually just as good a player. But if this were the case, just how lucky did Monica get? What is the probability that Monica could win four games out of seven, despite being no better a player? If the probability is fairly high, then Conchita's supporters can retain their hope. But if the probability is extremely low, then they may have to admit that Monica is the better player.

The role of statistical inference is to determine the probability of observable events (e.g., the probability of winning four games out of seven) from hypothesized underlying states of the world (e.g., Monica and Conchita having the same level of expertise). If the hypothesized state of the world cannot easily generate the observed events, then the hypothesis is probably wrong.

Most claims about the world are of the same status as claims about tennis expertise. Consider claims that might be made in cognitive science, such as 'my dog knows the command "stay"'. The same claim might be made instead of a robot, or of a human participant in a psycholinguistics experiment. How do we test such claims? We can only say the word 'stay' several times and observe what the listener does, whether the listener is a robot, a dog, or a human. From these observations, we infer whether it is tenable to claim that the listener knows the command.

The Logic of Hypothesis Testing: Population and Sampling Distributions

Suppose we make a claim about the world. This claim amounts to a specific description of the underlying population from which our observations are sampled. Given this model of the population, we determine the probability of getting any particular sample of observations. That is, we determine the distribution of samples that we would get from the hypothesized population. From this sampling distribution, we can ascertain the kind of sample we should expect to obtain, and we can

determine the specific probabilities of getting samples that deviate from the expected sample. (Examples of this process are given below.) If our actual set of observations deviates wildly from the expected sample, then the hypothesized population is probably not a good description of the actual world. Standard statistical inference is just the mathematical formalization of this logic for different types of populations.

POPULATION MODELS WITH DISCRETE CATEGORIES

In many situations we are concerned with discrete states of the world. Familiar examples include a tossed coin coming up heads or tails (two discrete states), or the hair color of a randomly selected person being black, brown, blond or red (four discrete states). This type of situation can be contrasted with situations in which the states of the world are ordered, or ranked (e.g., grades in an examination) or interval-scaled (e.g., response times). Situations with scaled (continuous) variables are discussed later in this article.

The Binomial Sampling Distribution

Suppose we want to test somebody's claim that a certain coin is fair; i.e., that it comes up heads 50% of the time. This is a situation in which the world has two states: heads or tails. If the coin is fair, then when we toss it a number of times, it should come up heads in about half of those tosses. For example, if we toss the coin eight times, it should come up heads about four times. Suppose that it actually comes up heads six times out of eight tosses. Is this deviation from the expected four heads large enough for us to reject the hypothesis that the coin is fair? What if the coin comes up heads seven times out of eight, or eight times out of eight? By how much can the sample deviate from what we expect before we reject the claim that the coin is fair?

To answer this question, we must calculate the probability that a fair coin would generate so many heads in eight tosses. Figure 1 shows these probabilities. The population distribution at the left assumes a fair coin; i.e., a coin for which the probability of heads is 50%. The graph shows two bars, one labeled 'H' for heads and the other labeled 'T' for tails. The height of each bar is 0.5, indicating that we are supposing that the probability of getting a head is 0.5 and the probability of getting a tail is also 0.5. Imagine repeatedly tossing the coin eight times, as exemplified in the middle of Figure 1. Usually a fair coin will come up heads about four times out of eight, but occasionally we would expect more extreme outcomes. The exact probability of each possible outcome can be mathematically derived, and the result, called a *binomial distribution*, is plotted in the graph on the right side of Figure 1. We see that the probability of getting four heads out of eight tosses is about 27%, the probability of getting five heads is about 22%, the probability of getting six heads is about 11%, and so on. Thus, a hypothesis about something we cannot directly observe (i.e., the fairness of the coin) makes specific predictions about the probabilities of what we can observe (i.e., getting so many heads out of eight tosses).

Suppose we now actually toss the suspected coin eight times, and the result is that it comes up heads seven times. From the binomial sampling distribution, we see that the probability of getting seven heads is about 3%. The probability of getting eight heads is less than 1%. Thus, the probability of getting a result this deviant or more from the expected outcome is less than 4%.

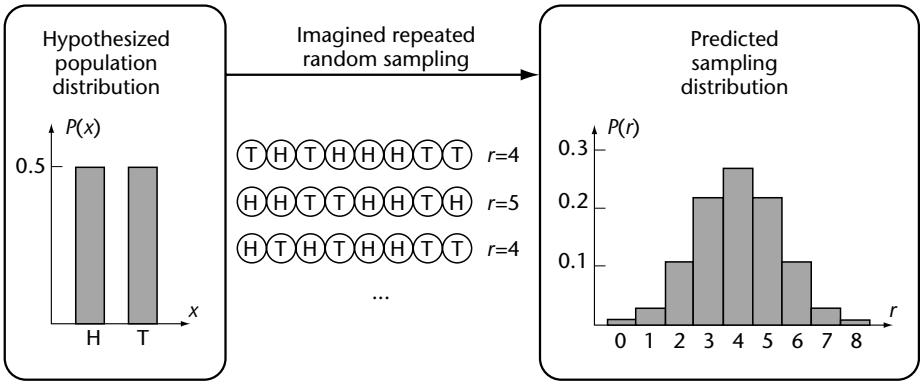


Figure 1. Illustration of a binomial sampling distribution generated by a fair coin tossed eight times. The predicted sampling distribution on the right shows the probability of getting *r* heads, for values of *r* between 0 and 8.

Therefore, because the result is so unlikely if the coin is fair, we reject the hypothesis that the coin is fair. The coin still might be fair, but it is rather unlikely given the result of seven heads out of eight flips. Note also that this is a 'one-tailed' test. A two-tailed test (in which we would have to add the probabilities of getting no heads or one head) is generally more appropriate in this and other similar situations.

The Chi-squared Distribution

An analogous procedure is applied when we have a situation with more than two discrete states. For example, suppose we want to know if four hair colors – black, brown, blond and red – are equally likely to occur in the general population. We hypothesize that each of the four colors has a probability of 0.25 in the population. We then imagine repeatedly sampling from this population. Suppose we select 40 people at random in each sample. If the four colors are equally likely, then we would expect to get about 10 people of each hair color, but by chance we would often get a few more or a few less of each color. For any given sample, a number called chi-squared (χ^2) describes how much the sample deviates from what we expect. For every possible sample of 40 people, there is a corresponding value of chi-squared which measures how deviant that sample is from the expected 10 people in each category. In most cases the samples will have moderate chi-squared values, but on rare occasions the samples will have extreme chi-squared values. The hypothesis of equally likely colors thereby generates a predicted sampling distribution for the chi-squared values. This sampling distribution shows the probability of getting each possible chi-squared value from the hypothesized population.

We now get an actual sample of 40 randomly selected people, and we determine the chi-squared value of this sample. If this actual chi-squared value is very extreme in the predicted sampling distribution, we reject the hypothesis.

The chi-squared method can also be applied to the situation of just two discrete states, and therefore subsumes the binomial distribution. Nevertheless, the binomial is often used in research because of its simplicity.

POPULATION MODELS WITH CONTINUOUS VARIABLES

The methods described above dealt with discrete variables; e.g. heads versus tails or category of hair color. The same logic can be applied to continuously scaled variables such as temperature, response time, blood pressure, etc.

The *t* Distribution

Suppose someone claims that a certain drug decreases blood pressure. (Notice that blood pressure is measured on a scale with meaningful intervals, not just discrete categories such as yellow or green.) How do we test this claim about the drug? One approach is to randomly assign people either to a group that takes the drug or to a comparison group that does not take the drug. (Usually the comparison group will be given some innocuous replacement substance, called a placebo, in a so-called 'double blind' procedure, so that neither the subject nor the person administering the drug knows whether it is really the drug or the placebo. This equalizes any psychosomatic effects in the two groups.) Then we measure the blood pressure of everyone in the two groups, and compare the mean blood pressures of the two groups. If the drug-treated group has lower mean blood pressure than the control group, are we certain that the drug works as claimed? Not necessarily. It could be that the drug has no real effect, but that the people in the drug-treated group happened to have lower blood pressure because of other random influences. So we need to determine just how probable it would be for the drug-treated group to have a mean that much lower, if the actual effect of the drug were zero.

First we create a model of the hypothesized population in which the drug has no overall effect. We conceive of the control group as having a distribution of blood pressures, some higher than average and some lower. We conceive of the drug-treated group as having the very same distribution of blood pressures: that is, the drug has no effect.

What shape of distribution should we assume? The usual assumption is a normal distribution – the well-known bell curve, as shown in the left panel of Figure 2. There are several motivations for using the normal distribution. One is that it has convenient mathematical properties for deriving probabilities in the sampling distribution. Another is that the central limit theorem shows that many different distributions become approximately normal when the processes that generate them are repeated many times and the results averaged. Finally, many distributions in the real world are approximately normal. The normal curve shows the relative probability of each blood pressure. Thus, the most probable blood pressure is the value under

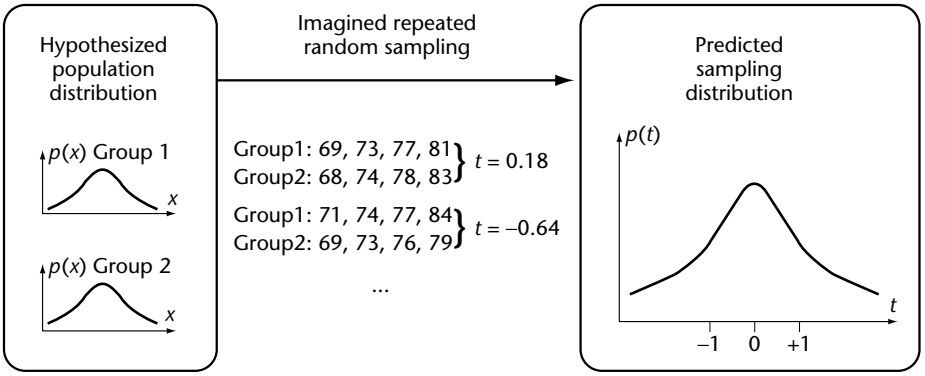


Figure 2. Illustration of the t sampling distribution generated by comparing four scores sampled from each of two identical normal distributions.

the peak of the normal curve. Blood pressures higher or lower than this average are less probable.

Having hypothesized this model of the population, we then imagine repeatedly sampling from it. Suppose we randomly select four people from each group. The mean of each sample will be about the same as the mean of the population, but the two means will not be exactly the same because of the random sampling. This difference between the group means is not very meaningful by itself, because its value depends on the scale units in which it is measured (e.g., a temperature difference of 3 degrees has different meanings depending on whether the scale is Fahrenheit or Celsius). Therefore the difference between group means is divided by the expected dispersion of this difference across samplings, and the resulting ratio is called the t statistic. Usually the difference between groups means, and hence the value of t, will be close to zero, on the hypothesis that we are sampling from a population with identical groups. Sometimes, however, it will be significantly above or below zero. The sampling distribution for t is illustrated in the right panel of Figure 2. Thus, a hypothesis about something we cannot directly observe (i.e., the equality of the distributions for the two groups) makes specific predictions about the probabilities of what we can observe – values of the t statistic.

Suppose we now actually test the drug on four randomly selected people, with four other randomly selected people getting the placebo. Suppose that we get a t value of 2.45. This value implies that the difference between the group means is fairly large compared with the expected variation of this difference. From the sampling distribution of t, it can be determined that the probability of getting a t value this extreme is less than 5%. Therefore, because this extreme value of t is so unlikely if the two groups have the same underlying population, we reject the hypothesis that the drug has no effect. Note that the drug might nevertheless have no effect, but it is rather unlikely given the result that $t = 2.45$.

The *F* Distribution and Analysis of Variance

An analogous method can be applied when there are more than two groups. For example, there might be several different drugs that are claimed to be effective in reducing blood pressure. To test whether the drugged and the control groups differ in their overall mean blood pressures, we can measure the overall dispersion among the group means relative to the dispersion within the groups. This ratio is called the F statistic. The logic of using it is the same as for other statistics. First, hypothesize identical normal populations for the groups. Then imagine repeated random sampling from these populations, each time computing the value of the F statistic. This generates a predicted sampling distribution for F. Now get actual random samples of blood pressures from people who have taken the various drugs or the placebo, and compute the F statistic. If the actual F statistic deviates greatly from what would be expected from the hypothesis of equality, then we reject that hypothesis.

This procedure is called *analysis of variance* (ANOVA) because the overall variance among the individual scores is decomposed (analyzed) into (1) the variance between group means and (2) the variance within groups. The ratio of these two components of the overall variance is the F statistic.

ANOVA can be applied when there are just two groups, and so it subsumes the t statistic. Nevertheless the t statistic is often used in research because of its simplicity.

Linear Regression and the General Linear Model

Many types of claims can be expressed as mathematical hypotheses about a population. We can imagine more and more complicated hypotheses, yet apply the same logic as in the previously discussed situations.

For example, suppose that we believe that the blood pressure drug should yield a decrease in blood pressure proportional to the amount of drug administered. We might also imagine that exercise increases blood pressure, in proportion to the rate of exertion. This type of hypothesis can be expressed mathematically as a so-called *linear model*. Denote predicted blood pressure by $\hat{y}$, drug quantity by x_d, and exercise rate by x_r.Then a linear relationship between blood pressure, drug and exercise can be described as $\hat{y} = \beta_0 + \beta_d x_d + \beta_r x_r$, where β_0 is base-line blood pressure and β_d and β_r are numbers that reflect how big an influence each factor (drug and exercise) has on blood pressure. If β_d is a large negative number, then predicted blood pressure drops rapidly as drug dosage increases. If β_r is a large positive number, then predicted blood pressure increases rapidly as exercise rate increases.

We can construct a series of hypothetical populations in which one or more of the factors (drug dosage or exercise rate) has no effect. For example, if both β_d and β_r are zero, then we are hypothesizing no effect of the drug and no effect of exercise on blood pressure. We can then imagine repeatedly randomly sampling from this hypothesized population, and for each sample computing the value of a statistic analogous to F. From this process we generate a sampling distribution predicted by the hypothesis. Then we can get an actual sample, and determine whether the hypothesized population could generate this actual sample with reasonable probability.

If the hypothesis that neither factor has an effect can be rejected, then we can hypothesize that one or both of the factors does have an influence; that is, we allow one or both of β_d and β_r to be nonzero. Suppose we hypothesize that the drug has no effect, so $\beta_d = 0$, but exercise does have an effect, so $\beta_r \neq 0$. We want to test whether this hypothesis can be rejected, but what should be the exact value of β_r for our population model? We set it to whatever value best fits the data, to give the hypothesis its best chance of being confirmed. The degree to which a model fits data is usually measured as the 'sum squared error' $\sum_i (y_i - \hat{y}_i)^2$ between actual data points y_i and predicted values $\hat{y}_i$, where $\sum_i$ indicates summation over all the data points. The value of β_r that minimizes this quantity is used for the population hypothesis. Then we proceed with the imagined sampling from the hypothesized population, to assess the probability that our actual data could have come from such an underlying population.

This type of procedure, using linear models, is called *multiple linear regression*. It turns out that models of group differences in ANOVA can all be expressed as particular cases of linear models. Therefore, the general linear model subsumes multiple linear regression and ANOVA. In much research, however, ANOVA is used because of its conceptual simplicity.

The Model Comparison Approach

It is often the case that complicated hypotheses about the world can be modeled as extensions of simpler models. More complicated models will involve more free parameters, like the coefficients in the linear model described above. As parameters are added to a model, it will usually fit data better, but only at the cost of being a more complex model. We want the model to be as simple as possible while also fitting the data well. Does the inclusion of particular new parameters improve the fit of the model significantly?

To answer this question, we use the same logic as has been described above. Assume that the 'simple' model, the one with fewer parameters, is a correct description of the world. (The parameter values in this simple model are set to whatever best fits our present data set.) The simple model incorporates an assumption about the distribution of values in the population; typically the population is assumed to be normally distributed around means specified by a linear function. We then imagine repeatedly sampling from this 'simple' population. For every sample, we compute the improvement in fit that is gained by including the extra parameters. This measure of improvement in fit is called the *generalized F ratio*. Note that a given sample from the simple population will almost surely be fitted better by a model with more parameters, because the sample will deviate randomly from the simple population's form, and this random deviation can be partially accommodated by the extra parameters.

The repeated sampling generates a sampling distribution of improvements in fit achieved by the more complex model, even when the simpler model is correct. Once we know this sampling distribution, we can see whether the improvement in

fit for our actual data set lies near the expected value of this sampling distribution. If it does, then the simpler model may be adequate. On the other hand, if the actual improvement in fit deviates greatly from what we would expect from the simpler model, then we reject the simpler model.

All of the methods described in this section are special cases of this model comparison approach. The model comparison approach is conceptually powerful and helps unify and generalize the various traditional statistical tests. However, the specific methods, such as t tests, ANOVA, and multiple linear regression, continue to be popular.

RESAMPLING METHODS

All of the methods described so far use the same basic logic: hypothesize a specific form of the population distribution (i.e., start with a model of the population) and then derive the predicted sampling distribution of whatever statistic seems useful. For example, the t statistic was sampled from normally distributed populations.

But what if we do not have a specific population distribution in mind? For example, what if a normal distribution seems inappropriate for modeling our particular population? We could just hypothesize a different, more appropriate distribution – but what if we have no theoretical commitment to any particular distribution? To answer this question, we ask another question: what is the best information we have about the underlying population? It is our actual sample. So we let the data themselves serve as the population, from which we then resample to generate a sampling distribution.

For example, suppose we are measuring scores on a statistics examination. Three students who read an encyclopedia article about statistics scored 58, 70, and 89. Three other students who did not read the article scored 50, 55, and 68. On average, the group who read the article did better, but maybe this was merely a random effect of the sampling. Maybe the two groups really come from the same underlying population, with no genuine difference between them. If we do not want to assume any specific form for this mutual underlying population, we can just let the six scores themselves be the population from which we sample. So the question becomes this: if we repeatedly sample from this population of six scores, how likely is it that a random sampling would yield a difference between groups as large as what we found in our actual sample? It turns out that there are 20 ways of sampling two groups of three from the six scores, and of these 20, there are 4 ways that yield group differences as large or larger than the one actually obtained. Because 4 out of 20 is not very low, we would probably not reject the hypothesis that the two groups come from the same underlying population.

Resampling methods are becoming increasingly popular because of the increasing availability of computer software. Their main advantage is that they do not require specific assumptions about the shape of the population distribution. A disadvantage is that important statistical concepts (such as power and confidence interval) are difficult to quantify by these methods, because to do so would require specification of alternative hypothetical populations.

BAYESIAN APPROACHES

All the methods described above are based on determining the probability of the data given a hypothesis. The logic is that if the data are very unlikely to have come from the hypothesized model, then the model is probably wrong. But this logic is incomplete. The probability of data D given a hypothesis H is not necessarily the same as the probability of the hypothesis given the data. In mathematical notation, usually $P(H|D) \neq P(D|H)$, where $P(A|B)$ means the probability of A given B. The methods described above provide $P(D|H)$, but we would really like to know $P(H|D)$.

For example, consider a deck of playing cards. If I select a card at random, and tell you that it is a king, then what is the probability that it is also a heart? The answer is 1/4, because of the four kings in the deck, one is a heart. If instead I tell you that the card I selected is a heart, then what is the probability that it is also a king? The answer is 1/13, because of the 13 hearts in the deck, one is a king. Thus $P(\text{heart}|\text{king}) \neq P(\text{king}|\text{heart})$.

The relationship between $P(H|D)$ and $P(D|H)$ is given by Bayes' theorem: $P(H|D) = P(D|H)p(H)/\sum_i P(D|H_i)P(H_i)$, where the terms H_i refer to all possible hypotheses about the situation. At first it might seem impossible to consider all possible hypotheses, and to establish the prior probability $P(H_i)$ that each is true. But in many situations this can be done in reasonable ways.

For example, suppose that we are questioning the fairness of a coin. We believe that there is a 70% chance that the coin is fair; i.e., $P(H_{.50}) = 0.70$. We believe that there is a 25% chance that the coin is biased to yield heads 35% of the time; i.e., $P(H_{.35}) = 0.25$. And we believe that there is a 5% chance that the coin is biased to yield heads 10% of the

time; i.e., $P(H_{.10}) = 0.05$. Suppose we actually toss the coin four times, and it comes up heads zero times; this result is our data, denoted D. The binomial distribution indicates that $P(D|H_{.50}) = 0.0625$, $P(D|H_{.35}) = 0.1785$, and $P(D|H_{.10}) = 0.6561$. Thus, the probability that the data could be generated by a fair coin is pretty low, and we might even want to reject the hypothesis that the coin is fair. We also see that the most biased of the three hypotheses has the highest probability of generating the data.

The most biased hypothesis, however, is also the hypothesis that we believed, before flipping the coin, was very unlikely to be true. So how much should we increase our belief in this unlikely hypothesis? Using Bayes' theorem, we obtain $P(H_{.10}|D) = P(D|H_{.10})P(H_{.10})/[P(D|H_{.10})P(H_{.10}) + P(D|H_{.35})P(H_{.35}) + P(D|H_{.50})P(H_{.50})] = 27\%$. Similarly, we find that $P(H_{0.35}|D) = 37\%$ and $P(H_{0.50}|D) = 36\%$. Thus, before tossing the coin, we thought that $P(H_{.10})$ was only 5%, but after tossing the coin, we think that the probability is 27%, more than five times as great. Moreover, before tossing the coin, we believed that $P(H_{.50})$ was 70%, but after tossing the coin, we think that the probability is 36%, only about half as likely. Nevertheless, despite the data, the chances that the coin is severely biased are still not as high as the chances that the coin is unbiased, or only moderately biased. To revise our prior beliefs more radically, we would need more convincing data.

There are several advantages of a Bayesian approach to inference. Firstly, it forces the theorist to consider multiple hypotheses and the prior probability of each. This encourages breadth of theorizing. Secondly, it takes into account the prior probabilities of hypotheses, so that bizarre hypotheses with small prior probabilities are unlikely to be accepted even if they could generate the data, and very plausible hypotheses with high prior probabilities are not necessarily rejected even if a particular data set is not easily generated by them. Thirdly, the framework can be extended to take into account the costs or pay-offs of different kinds of correct or incorrect decisions about hypotheses. On the other hand, a disadvantage of Bayesian statistics is that it is often impossible to specify the prior probabilities of all possible hypotheses.

CONCLUSION

For statistical methods to be applied, society must provide two conditions. Firstly, people must have the liberty to doubt claims. If claims about the world are dictated by authorities, without permission to doubt the claims, then there is no room for inferential statistics. Secondly, people must have the liberty to gather data that are randomly sampled and representative of the underlying population about which the claim is made. If a claim is made that Monica is better than Conchita, but the only games you are allowed to observe have the wind against Conchita, then your observations are not a good test of the claim. Thus, without liberty, statistical methods cannot be meaningfully applied, and when statistical methods have been appropriately applied, liberty has been exercised. *Ubi dubium, ibi libertas*: where there is doubt, there is liberty. Doing a statistical test is a political act, and an expression of the culture in which we work.

Liberty should not be confused with license. Scientific research depends on the liberty to apply statistical methods to random, representative samples. But care must be taken that the methodological strictures of statistical method and experimental design do not override the ethical principles of human beings. The methods of science are an exquisitely refined expression of human intellect combined with human awe and curiosity about the mysteries of nature. Statistical methods must be applied with humility and responsibility; they must not violate the rights of the very beings whose culture created the methods.

Further Reading

Bronowski J (1973) Knowledge or certainty. In: *The Ascent of Man*, chap. XI, pp. 352–377. Boston, MA: Little, Brown.

Huff D (1954) *How to Lie With Statistics*. New York, NY: W. W. Norton.

Judd CM and McClelland GH (1989) *Data Analysis: A Model-Comparison Approach*. San Diego, CA: Harcourt Brace Jovanovich.

Maxwell SE and Delany HD (2000) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum.

Wickens TD (1998) Drawing conclusions from data: statistical methods for coping with uncertainty. In: Scarborough D and Sternberg S (eds) *Methods, Models and Conceptual Issues: An Invitation to Cognitive Science*, vol. IV, pp. 585–634. Cambridge, MA: MIT Press.

Stereotypes

Introductory article

Steven Fein, Williams College, Williamstown, Massachusetts, USA
William von Hippel, University of New South Wales, Sydney, New South Wales, Australia

CONTENTS

Stereotypes are consensual beliefs about group characteristics that influence the perception, interpretation, and evaluation of others, sometimes blatantly but often in a manner so subtle that they are outside awareness. Because they serve basic cognitive and motivational functions, stereotypes are highly resistant to change.

INTRODUCTION

On 4 February 1999, just after midnight, 22-year-old Amadou Diallo entered his apartment building. He was spotted by members of the Street Crime Unit, an elite corps of New York City police officers who had been extraordinarily successful in reducing crime, but were often criticized for being too aggressive. In particular, in a practice known as 'racial profiling', the Street Crime Unit was routinely selecting African American and Hispanic men to be stopped and frisked. That winter night in 1999, Amadou Diallo was targeted to be one of them. Four white, plainclothes police officers from the unit spotted Diallo, who matched the general description of a suspected rapist. The police thought that Diallo looked suspicious as he appeared to duck into his building to avoid them. As they approached and ordered him to freeze, he reached into his pocket and began to pull out a wallet. Thinking that the wallet was a gun, the police opened fire. Nineteen of their 41 shots hit Diallo, and he lay dead in the vestibule. Diallo was unarmed.

In the days and weeks that followed, numerous protests erupted in New York City and around the country. The controversial use of 'racial profiling' came under national attention and renewed attack. The four officers responsible for the shooting of Diallo were acquitted of any wrong-doing, but many felt that their stereotypes of African Americans played a critical role in their misperception of his wallet as a gun, and their decision to fire so many bullets. In contrast, various politicians, columnists, and citizens defended the police, noting how difficult it is to make life-or-death decisions in the blink of an eye. Although we'll never know what role stereotypes played in the Diallo shooting, the research discussed in this article makes it clear that things might have turned out very differently if Diallo had been white.

The Diallo tragedy, and much of the research discussed in this article, focus on race. Nevertheless, race is but one kind of group membership that can influence people's thoughts, feelings, and actions towards others. Stereotypes, prejudice, and discrimination also emerge as a function of people's gender, sexual orientation, age, physical appearance, economic class, religion, and a variety of other social categories. Watch the news and you might see stories about genocide in Rwanda, 'ethnic cleansing' in Bosnia, neo-Nazi violence in Germany, and hate crimes against gays, Jews, and blacks in the USA. Despite the magnitude and prevalence of all this violence, it is really just the tip of the iceberg, as stereotypes change the way that people interact with each other in countless subtle ways all over the globe.

This article is divided into four parts. First, we consider the origins and formation of stereotypes. We then examine how stereotypes are maintained, often in the face of inconsistent information from the environment. Next, we discuss how and when stereotypes are applied, with special attention to the consequences of stereotype application for those who are the targets of stereotyping. We conclude with a discussion of ways that stereotypes can be changed.

STEREOTYPE FORMATION

Stereotypes are consensual beliefs about the characteristics and traits of a group of people. Although stereotypes are not necessarily negative (e.g. blacks

are stereotyped as athletic), even positive stereotypes tend to take on negative connotations when used to describe members of other groups (e.g. black athleticism is seen as a sign of a primitive nature). This is one of the ways in which stereotypes and prejudice (negative attitudes towards members of other groups) are interrelated. There are numerous origins of such stereotypes. From a historical perspective, it has been argued that slavery in America gave rise to the portrayal of blacks as inferior, just as the surprise attack on Pearl Harbor in the Second World War fostered a belief that the Japanese are sneaky. From a political perspective, stereotypes are viewed as a means by which groups rationalize war, religious intolerance, and economic oppression. From a social/cultural perspective, it has been argued that different social roles and real differences between groups contribute to perceived differences.

The formation of stereotypes involves two related processes. The first is *categorization,* by which people sort themselves and others into groups. The second is a process by which people perceive groups to which they belong (ingroups) as being different from groups to which they do not belong (outgroups). These two processes reflect not only basic cognitive operations but also cultural and motivational factors. Because these two processes are fundamental to social interaction, stereotypes form very early in life. For example, children in the USA have been shown to endorse stereotypes concerning blacks by the age of five, and to endorse gender stereotypes at an even earlier age.

Social Categorization

People routinely sort objects into groups rather than thinking of each as unique. Just as we categorize a new piece of furniture as a chair, and thereby know how to properly interact with it, we also sort each other into groups on the basis of gender, race, age, and other attributes. In a manner similar to object categorization, social categorization is natural and adaptive. By grouping people the way we group foods, furniture, and other objects, we form impressions quickly and use past experience to guide new interactions. In this way we are able to 'go beyond the information given', making inferences about people whom we've never met without expending a great deal of energy or effort.

There are, however, serious drawbacks to the information gained and energy saved through social categorization. First of all, our categorizations may rely on erroneous beliefs, or may lump together people who have little in common. Second, even in cases in which the stereotype is associated with real group differences, categorizing people leads us to overestimate the differences between groups and to underestimate the differences within groups.

Third, once people are sorted into categories, they often are evaluated and remembered with reference to the category. Thus, people tend to notice minority group members more than majority group members, one consequence of which is that negative behavior from a minority group member tends to stand out in people's minds more than negative behavior from a majority group member. This process is known as *illusory correlation*, and it leads people to favor majorities over minorities even when the overall pattern of behaviors is identical across groups.

Finally, because people can't choose many of the groups they belong to (e.g. gender and race), all members of the group tend to be associated with the stereotypic labels whether these labels accurately describe them or not. This process is often unfair and demeaning to those who are targets of stereotypes, and as the Diallo case shows, it can be dangerous to members of groups who are marginalized in society or perceived as violent.

Ingroups versus Outgroups

The second process that promotes stereotyping follows directly from the first. Although grouping humans is much like grouping objects, there is a critical difference. When it comes to social categorization, the people doing the categorizing are members or nonmembers of the categories they use. The tendency to carve the world into ingroups and outgroups, or 'us' and 'them', has a number of important consequences. Perhaps the most important consequence is ingroup bias, or the nearly universal tendency to favor members of one's own group over members of other groups. Ingroup bias is so ingrained that the simple act of placing people into randomly determined groups can create it, and indeed the mere mention of the words 'us' and 'them' leads automatically to associated positive and negative emotions.

Social, Cultural, and Motivational Factors

Social categorization and ingroup/outgroup distinctions reflect basic cognitive processes; they are

by-products of how people think. They are also influenced, however, by situational factors, such as the motivations that people have in particular settings and the cultural context in which they live. For example, people are more likely to rely on stereotypes when they are feeling bad about themselves, as stereotypes help them denigrate others and thereby feel relatively better by comparison. Stereotypes and prejudice are also magnified when people are in conflict with each other, as the act of stereotyping facilitates dehumanization, which in turn enables people to be more ruthless with each other than they might otherwise be. It is also clear that people learn stereotypes through role models, conformity to group norms, and immersion in their culture more generally. As with hairstyles and taste in music, stereotypes are affected by peers, family, and immediate culture.

STEREOTYPE MAINTENANCE

People tend to perceive and explain events differently as a function of whether the events are consistent or inconsistent with their stereotypes, and these perceptual and explanatory processes are biased in favor of stereotype maintenance. A large number of experiments show that even when members of two groups behave identically, people who hold different stereotypes of the two groups will typically see them as different from each other in ways that are stereotype-consistent. In addition, subtyping and self-fulfilling prophecies are two important processes that help maintain stereotypes in the face of what otherwise might be disconfirming evidence.

Subtyping

One of the unnerving paradoxes of stereotyping is that people often manage to hold negative views about a certain group even when they like individual members of the group. Gordon Allport recognized this phenomenon many decades ago, when he wrote: 'There is a common mental device that permits people to hold prejudgments even in the face of much contradictory evidence. It is the device of admitting exceptions. … By excluding a few favored cases, the negative rubric is kept intact for all other cases.' Confronted with a woman who does not seem particularly warm and nurturing, for example, people can either develop a more diversified image of females or toss the mismatch into a special subtype – say, *career women*. To the extent that people create this subtype, their existing stereotype of women-in-general will remain intact.

Self-fulfilling Prophecies

Stereotypes not only influence perceptions of other groups, they can also influence how other group members actually behave. Through the mechanism of self-fulfilling prophecies (whereby people hold a belief that causes them to change their behavior, which in turn causes their original belief to come true), stereotypes can bring about their own reality.

For example, when teachers think that their poor or minority students are more likely to be disruptive, and less likely to perform well academically, they tend to challenge them less and discipline them more. The consequence of this pattern of treatment is that poor and minority children don't learn as much, and because they aren't intellectually stimulated they tend to act out in the classroom. This behavioral pattern only confirms the teachers' original stereotypes, leading to more discipline and less mental challenge. The unfortunate consequence of this spiraling behavioral sequence is that poor and minority children tend to perform much worse in school than their well-to-do and majority counterparts, in the USA, the UK, and around the world. Self-fulfilling prophecies also emerge in a variety of settings outside the classroom, and all groups are occasionally victims and perpetrators of the process.

STEREOTYPE APPLICATION

Stereotypes often color the interpretation of events, particularly events that are ambiguous or learned second-hand. For example, imagine learning that a nurse got into a fight at work. Now imagine learning that a construction worker got into a fight at work. What images of these actions come to mind? Research shows that people interpret 'getting into a fight' differently as a function of who did it, and then falsely remember that they learned rather than imagined the stereotypic information (e.g. that the nurse got into an argument and the construction worker got into a fist-fight). This is a fundamental effect of stereotyping: people are likely to think of others as more stereotypic than they actually are, and often do not recognize that they are interpreting behaviors in a stereotypic fashion.

The Importance of Ambiguity

Because most people do not want to admit (either to themselves or to others) that they are stereotyping, they tend to rely on their stereotypes only when the situation provides them with ambiguity about the cause of their behavior. As an illustrative

example, consider a classic experiment conducted by Melvin Snyder and his colleagues at Dartmouth College. Snyder and colleagues brought people into the laboratory under the auspices of completing a questionnaire, and asked them to have a seat in a nearby room to fill it out. In reality, Snyder wasn't interested in the questionnaire; he only cared where the people sat down. Only two seats were available, one of which was next to a physically disabled person and one of which was next to a person who wasn't disabled. In front of each empty chair sat a television; half the time the two TVs were presenting the same program and half the time they were showing different programs. When the two TVs were showing the same program, most people sat down next to the disabled person. But when the two TVs were showing different programs, most people avoided the disabled person (and it didn't matter which TV was showing which program).

The reason that the television programming was so important is that when the two TVs were showing different programs, they provided an excuse for participants to avoid the disabled person. The different programming on the TVs created ambiguity (for both self and other) about whether the choice of seating had anything to do with the person in the next seat, or was really caused by the program being shown. In contrast, when the two TVs were airing the same program, if the participant avoided the disabled person the meaning of this behavior would be crystal clear to both self and other. Under this circumstance, there would be no ambiguity about whether people were avoiding the disabled person, a behavior that most of us would be embarrassed just for considering.

These results suggest that stereotypes and prejudice will influence behavior towards members of other groups only when other aspects of the situation could legitimately be the cause of the behavior. Numerous other studies have supported this general idea that people will typically show evidence of stereotyping only when the situation provides an excuse for what would otherwise be obviously stereotypical judgments or behavior. Because the everyday world provides numerous situational details that could potentially be the source of what is really stereotypical behavior, these results suggest that subtle stereotyping should be quite common, and indeed it is. From shopping malls to employment agencies to the courtroom to housing and schools, blacks and other minorities face stereotypes that limit their opportunities in ways that are so subtle that frequently they themselves are unaware that stereotyping is at work.

Automatic Stereotyping

Because most cultures are suffused with stereotypes, people often automatically activate their stereotypes when they are exposed to members of groups for which popular stereotypes exist. Just as most of us automatically think of 'butter' when someone says 'bread', we also tend to automatically think of concepts relevant to a stereotype when we think of members of a stereotyped group. We can try to prevent the stereotype from influencing our judgments or behaviors (and nonprejudiced people do just that), but because we are often unaware that a stereotype has been activated, it can affect us despite our best intentions to the contrary.

That being said, there are people for whom automatic stereotype activation is less likely, and there are circumstances in which automatically activated stereotypes are more or less likely to be applied. In particular, people are likely to form stereotypic impressions when they're busy, pressed for time, or unable to think carefully (e.g. due to exhaustion or intoxication) about the unique attributes of the person encountered. In contrast, people often manage to inhibit or replace their stereotypic thoughts, and even prevent their activation, when they are highly motivated to form an accurate impression or be egalitarian in their judgments.

Stereotype Application from the Target's Perspective

In a provocative theory that has attracted a great deal of attention, Claude Steele has proposed that in situations in which a negative stereotype can apply to someone, people may fear being seen 'through the lens of diminishing stereotypes and low expectations'. Steele calls this predicament *stereotype threat*, because it hangs like 'a threat in the air' when the individual is in the stereotype-relevant situation. This predicament can be particularly threatening for individuals whose identity and self-esteem are invested in domains in which the stereotype is relevant. Steele argues that stereotype threat plays an important role in diminishing the performance and identification of stereotyped group members.

According to Steele's theory, stereotype threat can hamper achievement in two ways. First, the 'threat in the air' can directly interfere with performance, by increasing anxiety and triggering

distracting thoughts in performance situations. Second, if stereotype threat is chronic in a particular domain, it can cause people to *disidentify* from that domain – to dismiss the domain as no longer relevant to their self-esteem and identity.

To illustrate, imagine a black and a white student who enter high school equally qualified in academic performance. Imagine that, while taking a particularly difficult test at the beginning of the school year, each student struggles on the first few problems. The white student may begin to worry about failing, but the black student may also have a large set of additional worries about appearing to confirm a negative stereotype of blacks. Even if the black student doesn't believe the stereotype at all, or doesn't believe that it describes himself, the threat of being reduced to a stereotype in the eyes of those around him can trigger anxiety and distraction, impairing performance. And if he experiences this threat in school frequently – perhaps because he stands out as one of only a few blacks in the school, or perhaps because he is treated stereotypically by others – the threat can eventually wear him down. To buffer himself against the threat, he may learn to disidentify with school; if so, his academic performance will become less relevant to his identity and self-esteem. In its place, some other domain of life, such as social success or a particular nonacademic talent, will become a more important source of identity and pride.

The unfortunate consequence of this process is that once people disidentify with school, their academic performance tends to suffer because they no longer put the same time and energy into academic activities.

STEREOTYPE CHANGE

Intergroup Contact

Modern stereotypes are difficult to overcome because they manifest themselves in indirect ways, and even the perpetrators are often unaware that they are relying on stereotypes. Is there a solution to this problem?

According to Gordon Allport's contact hypothesis, one way to reduce stereotyping is to bring members of different groups into contact with one another. The contact hypothesis states that four conditions facilitate the positive effects of intergroup contact: people should have equal status, common goals, and cooperative means to achieve those common goals, and intergroup contact should be sanctioned by relevant authorities. The results of hundreds of experiments in dozens of countries have generally confirmed this hypothesis, although this research has also shown that intergroup contact can make stereotypes worse if people feel anxious or threatened in a contact situation.

The Jigsaw Classroom

As is noted above, cooperation and shared goals are important for intergroup contact to be successful. They can break down the psychological barrier between groups, leading members to recategorize the two groups into one and reducing ingroup favoritism: 'they' become part of 'us'. Yet the typical classroom is filled with competition, a factor that usually leads to increased stereotyping.

To combat this problem in the classroom, Elliot Aronson and his colleagues developed a cooperative learning method called the *jigsaw classroom*. In newly desegregated public schools in Texas and California, they assigned fifth-graders to small racially and academically mixed groups. The material to be learned within each group was divided into subtopics, much like the way a jigsaw puzzle is broken into pieces. Each student was responsible for learning one piece of the puzzle, after which all members took turns teaching their material to one another. In this system, everyone – regardless of race, ability, or self-confidence – needs everyone else if the group as a whole is to succeed.

This method produced impressive results. Compared with children in traditional classes, those in jigsaw classrooms grew to like each other more, liked school more, were less prejudiced, and had higher self-esteem. What's more, academic test scores improved for minority students and did not diminish for majority students. Much like an interracial sports team, the jigsaw classroom offers a promising way to create a truly integrated educational experience. It also provides a model of how to use interpersonal contact to promote greater tolerance of diversity.

Undoing Automatic Stereotype Activation

The situations in which people find themselves have an important influence on the degree to which they rely on stereotypes, but long-term personal decisions not to stereotype others can be important as well. Low-prejudice people in particular are often successful at regularly bypassing their stereotypes, as they seem to focus on personal information about individual members of

stereotyped groups and thus have an easier time keeping stereotypic thoughts out of mind.

This may be the best strategy for avoiding the influences of stereotypes: rather than trying to suppress thoughts about a stereotyped group, try instead to activate thoughts about the individual who happens to be a member of that group. Keep in mind, however, that unlearning stereotyping is much like trying to break other bad habits; it takes a long time, needs lots of practice, and you should not be discouraged if you slip up on occasion. Indeed, accidentally relying on stereotypes and thereby being unfair to someone is one of the most important motivators that helps low-prejudice people maintain vigilance so that they don't make such mistakes in the future.

In addition to a personal commitment to breaking the stereotyping habit, there are two other factors that may help put the brakes on stereotype activation. These are:

Take the perspective of a member of a stereotyped group. For example, in one experiment people spent a few minutes writing about a day in the life of an elderly man. Some of the participants were asked to suppress their stereotypes about the elderly while writing, and others were asked to take the perspective of the elderly man himself while writing. Both groups were equally good at not using stereotypes in their stories. But on a subsequent task in which their activation of the elderly stereotype was measured, those who had earlier tried to suppress the stereotype showed strong evidence of stereotype activation, but those who had taken the perspective of the elderly man did not.

Be motivated to be fair and egalitarian towards other groups. A number of experiments have shown that this motivation can be very effective in reducing stereotype activation. This goal is constant in some people, but it also can be induced temporarily in others, such as when people are made to feel that they recently have *not* been fair to others. When people feel that they have violated their own personal or cultural standards of fairness and morality, they become less likely to activate negative stereotypes of others. Conversely, if they feel that they have demonstrated their fairness and virtue, they may ironically become more likely to activate negative stereotypes.

CONCLUSION

A variety of cognitive, motivational, and social/cultural factors come together to cause people to rely on stereotypes even when they would rather not. This combination of factors contributes to the pervasiveness and perniciousness of stereotyping. When situations or personal standards and motivations cause people to see and value others as individuals rather than group members, however, they often can put stereotypes aside. Learning and applying stereotypes tend to be all too easy; resisting stereotypes, in contrast, requires effort, practice, and motivation. Nevertheless, the benefits to society that such resistance can bring are enormous, and thus researchers continue to study when and how stereotypes can be changed, weakened, and undone.

Further Reading

Allport G (1954) *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Bargh JA (1999) The cognitive monster: the case against the controllability of automatic stereotype effects. In: Chaiken S and Trope Y (eds) *Dual-process Theories in Social Psychology*, pp. 361–382. New York, NY: Guilford Press.

Bodenhausen GV and Macrae CN (1998) Stereotype activation and inhibition. In: Wyer RS (ed.) *Stereotype Activation and Inhibition: Advances in Social Cognition*, vol.11, pp. 1–52. Mahwah, NJ: Lawrence Erlbaum.

Crocker J, Major B and Steele C (1998) Social stigma. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, 4th edn. pp. 504–553. New York, NY: McGraw-Hill.

Devine PG and Monteith MJ (1999) Automaticity and control in stereotyping. In: Chaiken S and Trope Y (eds), *Dual-process Theories in Social Psychology*, pp. 339–360. New York, NY: Guilford Press.

Fiske S (1998) Stereotyping, prejudice, and discrimination. In: Gilbert DT, Fiske ST and Lindzey G (eds) *Handbook of Social Psychology*, 4th edn. pp. 357–411. New York, NY: McGraw-Hill.

Steele CM (1997) A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist* **52**: 613–629.

Theory of Mind

Introductory article

Andrew Whiten, University of St Andrews, Fife, Scotland, UK

CONTENTS

Theory of mind refers to the everyday psychology that we use to understand and explain our own and others' actions by reference to mental states, such as 'desiring', 'knowing' and 'believing'.

INTRODUCTION

The expression 'theory of mind' (ToM) was introduced into psychology by David Premack and Guy Woodruff in 1978. Asking, 'Does the chimpanzee have a theory of mind?', they described experiments to assess whether the primate most closely related to us shares our tendency to interpret others' actions by attributing to them 'states of mind' (or 'mental states'). Premack and Woodruff referred to these attributions as an everyday 'theory' because they are rather like scientific theories, first in being about theoretical entities – mental states, in this case – that are not directly observable; and second, in generating testable predictions about how others will behave, according to the state of mind they are presumed to be in.

These authors' attempts to investigate whether chimpanzees construe others as 'wanting' or 'intending' things are now generally seen as inconclusive, but they laid the foundations for what is now a large and influential area of research. Psychologists realized that the question of how humans – let alone chimpanzees – achieve 'everyday mindreading' remained largely unanswered. Developmental psychologists in particular saw that the growth of the child's ToM was a topic ripe for investigation, and during the 1990s they completed hundreds of experimental investigations. Meanwhile, the topic continued to preoccupy those studying our closest primate relatives, and also became a natural focus for new work by philosophers. As a result, ToM has become a topic of highly productive interdisciplinary debate and investigation in several branches of the cognitive sciences.

FOLK PSYCHOLOGY

Various terms have been used as rough synonyms for ToM, sometimes with slightly different connotations. For example, unlike ToM, 'mindreading' provides a useful verb – to mindread – and tends to be preferred by those who dislike talk of a 'theory' when we are just talking about people's everyday ideas about mind, rather than the science of psychology. Of course, in the present context 'mindreading' does not imply telepathy, as it can in common discourse! Another related expression is 'folk psychology'. Again, this is sometimes used simply as an alternative way of talking about ToM, but can imply something either broader or narrower. The broader sense of folk psychology refers to all the ways in which people act like amateur psychologists, explaining and predicting how and why people behave as they do. This may include certain rules expressed in behavioral terms, rather than appealing to the mind, as ToM specifically does; there is a folklore, for example, about how men and women might typically act differently in certain situations. In this respect folk psychology can include more than ToM. However, folk psychology is often taken to include only the folklore about the mind that people consciously and verbally express. In contrast, at least some conceptions of ToM include elements that a person is unaware of (a useful analogy is with the grammar that an infant acquires and applies, without being able to verbalize its rules), and from this perspective folk psychology may be seen as narrower than ToM.

DEVELOPMENT OF A THEORY OF MIND

Philosophers commenting on the original chimpanzee study by Premack and Woodruff suggested that clearer evidence of ToM would be the attribution to another individual of a false belief, because

this requires recognizing that another mind might represent the world in a way different from how one knows it truly to be. Developmental psychologists studying ToM in children accordingly made this achievement the priority in their research. In what has become one of the most frequently used tests for this ability, the child watches a scene in which Sally hides a special object (here a ball) in a covered basket, then leaves (Figure 1). While she is out, Anne transfers the ball to a closed box. Sally then returns and the child is asked the critical question: where will Sally look for her ball? By the age of 5 years, a child typically answers in the adult way, saying that Sally will look in the basket. Three-year-old children, however, are more likely to say that Sally will look where they themselves know the ball to be, in the box. They seem unable to appreciate that Sally will search under the guidance of a false belief, which is different to the child's own belief about the ball's location.

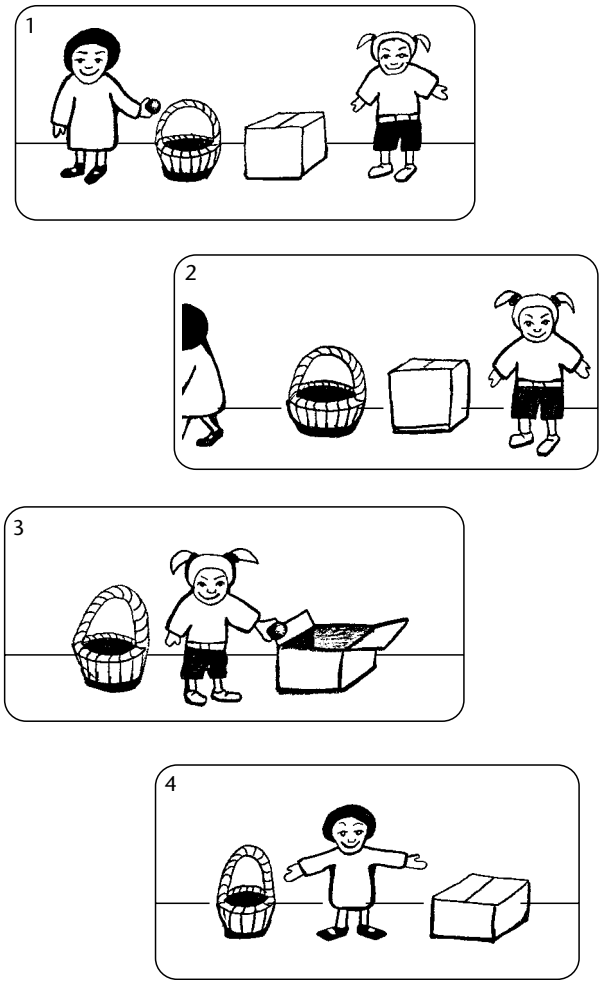


Figure 1. The 'Sally–Anne' test. Sally puts her ball in the basket (1) and leaves (2). Anne takes the ball from the basket and puts it in the box (3). Sally returns (4). Having watched the story unfold, a child is asked: 'Where will Sally look for her ball?' A three-year-old will tend to answer 'in the box,' a five-year-old 'in the basket.'

A major step is thus made between these ages in the operation of a theory of mind, and it can be detected in several other social contexts. One that has important ramifications in the child's everyday social life is the ability to create false beliefs – to lie and deceive. Although at earlier ages the child may have learned some social tactics that work by misleading others, only with this theory of mind 'watershed' at around 4 years of age is there a true understanding about why such tactics work. The child is a mentalist long before this, however. A variety of ingenious experiments and careful observations have provided evidence that 2-year-old children typically understand and talk about the rudiments of 'seeing' (whether something is hidden or visible to someone else, for example), as well as aspects of emotions, desires, attention and intentions. These states of mind do not require a recognition of the possibility of misrepresenting reality and they appear easier for young children to understand than false belief. More controversial evidence points to an ability to recognize goal-directed action at even earlier ages.

AUTISM AND THE 'THEORY OF MIND' MODULE

Individuals with autism, a condition in which social interaction is typically disturbed, have been found to fail the 'Sally–Anne' test for recognition of false beliefs even when their overall level of mental functioning is quite high. It has been suggested that this autistic 'mind blindness', as Simon Baron-Cohen has called it, has much to do with the social difficulties characteristic of the condition. A more general conclusion drawn from this result is that theory of mind is an example of a mental module, which develops and operates somewhat independently of other components of the mind. The hypothesis is that it is a device that evolved specifically for recognizing states of mind, so it is possible for it to become a relatively isolated malfunctioning unit even in an otherwise intelligent mind. The malfunctioning probably begins well before the stage at which the child would normally begin to attribute false beliefs, however. There is evidence it can often be detected earlier in a failure to engage in sharing the attentional focus of companions, an activity common in normal 2-year-olds and one generally regarded to indicate

the beginnings of engaging with the mental perspective of others.

Conversely, difficulties also tend to remain at later stages of development. Some individuals with autism eventually pass tests like the Sally–Anne one, that require first-order attribution of mental states, but yet fail to make more complex second-order attributions that would normally be expected for their mental age. Where a first-order attribution might be that Sally thinks her marble is in the basket, a second-order attribution embeds Sally's thought in another mental state, such as John believing that Sally thinks her marble is in the basket. This illustrates an important outcome of possessing a theory of mind: because such embedding can in principle continue indefinitely, mindreading can come to involve great social complexity (for example, 'I think you realized I didn't want you to believe what I said yesterday') that taxes even its most proficient users. In the case of autism, the findings about first-order and second-order difficulties have suggested that in this condition ToM suffers delayed development, rather than some fixed deficit.

DOES THE CHIMPANZEE HAVE A THEORY OF MIND?

The possibility that ToM might be an ability shared with other primates gained added support in the context of the 'Machiavellian intelligence' hypothesis, that the special intelligence of apes and monkeys is due more to the complexities of their social environments than anything else. This hypothesis is supported by evidence that the social complexity characteristic of a species is correlated with the relative size of their neocortex. If social intelligence provides the reproductive advantage this theory suggests, it might have encouraged the evolution of at least some elements of a ToM that could help primates to exploit possibilities for deception, for example. However, by comparison with the number of studies of children's ToM, studies on primates remain few and have concentrated on chimpanzees. The earliest studies offered some support for chimpanzee ToM, but then the pendulum of evidence swung the other way. In particular, chimpanzees failed to match young children even in the relatively elementary attribution of 'seeing', choosing randomly between two people from whom to request food, even when one had their eyes covered by an opaque screen and the other did not. More recently, chimpanzees placed in a more natural competitive situation with other chimpanzees did discriminate between situations in which their opponent could or could not see a piece of food, and they even did this in relation to what the opponent would or would not have seen earlier, which appears close to attributing 'knowledge' to the other. The question of primate ToM thus remains one of active research and controversy. What the data suggest is that if nonhuman primate ToM exists, it is not the robust and routine social tool that becomes so evident in human childhood.

HOW IS MINDREADING DONE?

Two rather different ways in which mindreading might operate have been distinguished. One is that connoted by use of the term, 'theory' in 'theory of mind' (which gives this theory the unfortunate name of the 'theory theory'): an individual becomes a mindreader by observing and experimenting in the social world rather like a little scientist, and comes up with a theory-like recognition that imputing specific states of mind (e.g. that Sally wants her doll, and believes it is in the cupboard even though it is not) predicts and explains actions rather well (Sally searches in the cupboard, only to be disappointed). In the case of child development, simple ToMs are superseded by more sophisticated ToMs as the former are found wanting, in a fashion somewhat analogous to the progress of science.

The main alternative idea is that instead of constructing such a 'theory', mindreaders put themselves in the position of the other, and use their own mind to simulate what the other might think or believe. Accordingly this is known as the 'simulation' theory. So far, there has been little evidence that can decide between this and the 'theory' theory in practice. A finding that 'theory' theory adherents have claimed in support comes from comparing children's mental attributions to self and others. One test begins with a child answering, 'Smarties' when asked what is inside a well-known sweet package. Then the child is shown that the package actually contains a pencil. When asked what another, approaching child will think is inside, a 5-year-old will tend to say 'Smarties', a 3-year-old 'pencil', consistent with the Sally–Anne results outlined earlier. Interestingly, however, each child will give similar answers about what they had originally thought was in the package (i.e. the young child will say 'pencils'). This has been taken to indicate that children would not find simulation helpful as a way of ever advancing the sophistication of their mindreading, because it appears to show they 'know their own mind' no better than that of others.

THE BRAIN BASIS OF THEORY OF MIND

How mindreading is done by the brain has begun to be investigated through three main kinds of evidence. First, at a crude level, it is known that damage to frontal cortex can impair social behavior. Second, brain scanning methods have been applied to people performing ToM tasks like the Sally–Anne one outlined above. The different methods converge on implicating medial prefrontal cortex in ToM operations, along with other areas including lateral inferior frontal cortex. The third kind of evidence comes from single neuron recordings in monkeys that focus on the coding of information about social actions likely to form the foundations for ToM, such as gaze direction and goal-directed action. These studies show activities in regions of the brain such as the superior temporal sulcus and area F5, adjacent to those regions identified with ToM in humans. We are finally beginning to glimpse how actions are processed in what are necessarily complex ways, such that 'brains can read minds'.

Further Reading

Baron-Cohen S, Tager-Flusberg H and Cohen DJ (eds) (2000) *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*. Oxford: Oxford University Press.

Call J (2001) Social cognition in chimpanzees. *Trends in Cognitive Sciences* **5**: 388–393.

Carruthers P and Smith PK (eds) (1996) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Frith CD and Frith U (1999) Interacting minds – a biological basis. *Science* **286**: 1692–1695.

Gopnik A (1993) How we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**: 1–14.

Mitchell P (1997) *Introduction to Theory of Mind: Children, Autism and Apes*. London: Arnold.

Perner J (1991) *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Premack D and Woodruff G (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* **1**: 515–526.

Sperber D (2000) *Metarepresentations: A Multidisciplinary Perspective*. Oxford, UK: Oxford University Press.

Whiten A (1997) The Machiavellian mindreader. In: Whiten A and Byrne RW (eds) *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge, UK: Cambridge University Press.

Williams JHG, Whiten A, Suddendorf T and Perrett DI (2000) Imitation, mirror neurons and autism. *Neuroscience and Biobehavioural Reviews* **25**: 287–295.

Theory-of-mind in Two-person Experiments

Introductory article

Mary Rigdon, George Mason University, Fairfax, Virginia, USA

CONTENTS

The theory of mind or folk psychology is appealed to by game theoretic concepts and explains behavior in a variety of bargaining environments.

THEORY OF MIND AND COMMON KNOWLEDGE IN GAMES

As cognitive agents we routinely predict and explain the behavior of others by ascribing to them a cornucopia of mental states *See* **Theory of Mind**. How we do this – what the information processing mechanisms are for these ascriptions, and how these ascriptions are used – is still very much an open question. Another interesting question is what role theory of mind (ToM) or folk psychology plays in other aspects of our cognitive lives. One such aspect is the class of strategic interaction contexts which can be modeled using game theory. ToM shows up in two different ways in economics experiments. First, even the simplest solution concepts of game theory tacitly appeal to agents' folk psychological abilities. Second, facts about ToM constrain those theories which attempt to explain how people behave in strategic environments.

Consider first how ToM is appealed to by even the simplest solution concepts in game theory. All classical solution concepts assume that the structure and rationality of the players are items of common knowledge among the players in the game. This information then is sufficient to derive predictions about plays of the game. Consider, in particular, the solution concept of rationalizability. The concept is that players predict the play of others given the assumption of common knowledge of rationality. So if a player i's action A_i is part of a rationalizable strategy, we know that i has made a prediction about what each player $j \neq i$ will do. But what is the prediction supposed to be based upon? Common knowledge of rationality, no doubt, but can we unpack this further? Suppose we say that a player is rational only if i best responds to what she justifiably believes the other players will do. Common knowledge of this fact would entail that player i's predictions about j are based on a host of mental states: beliefs about j's desires, beliefs, and so on. The same applies to j with respect to i's mental states. In fact, game theory assumes that both players will expect that each possesses self-interested intentions. The bottom line is that rationalizability depends tacitly on folk psychology. The same sort of inspection of other solution concepts – e.g. subgame perfection – leads to the same conclusion: players are assumed to have beliefs about and be motivated by the beliefs, desires, and preferences of other players in the game. This is just to say that these solution concepts tacitly appeal to agents' folk psychological abilities.

EVIDENCE FOR THEORY OF MIND IN ECONOMICS EXPERIMENTS

Now let us consider more concrete, empirical effects of ToM in strategic environments. One hypothesis is that an agent's ability to read the intentions of others plays a central role in achieving cooperative outcomes in personal exchange. One question is whether agents possess a 'friend-or-foe' mental mechanism for evaluating the intentions of another person. This module can be thought of as providing the needed information about one's counterpart in order to determine a course of action. If someone is detected as a friend, then this increases the likelihood of positive reciprocity from this person following a trusting action; whereas if someone is detected as a foe, then a different strategy may be followed to avoid a defection outcome. Perhaps such a mechanism can be primed in an experimental setting via simple changes in the instructional protocol.

Consider the following sequential decision problem; see Figure 1. In this decision problem, Player 1 begins by choosing between the outcome ($7, $14)

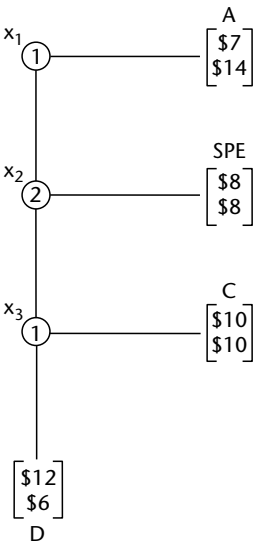


Figure 1. Trust game (extensive form).

or passing the game to Player 2, his counterpart. Player 2 then has a choice to move right, resulting in the outcome of ($8, $8) or passing the game back to Player 1, her counterpart. If the game is passed back to Player 1, then he decides between ending the game at the cooperative outcome of ($10, $10) (i.e. a move right at x_3) or ending the game by defecting from his counterpart, yielding the outcome of ($12, $6) (i.e. a move down at x_3).

Assume that it is common knowledge that each player acts in his or her own self-interest. So if Player 2 were to choose down at x_2, Player 1 would move down, ending the game at ($12, $6). The cooperative outcome would never be reached. Given that Player 2 knows this is the case, Player 2 will choose to move right at x_2 because the outcome ($8, $8) will be reached for sure, and $8 is certainly better than $6. Player 1 will therefore move down at x_1, preferring $8 to $7. The ($8, $8) outcome is predicted to be reached by agents using backward induction and eliminating dominated strategies. This is called the subgame perfect equilibrium (SPE) of the game.

The experimental protocol used in the laboratory involves 12 or 28 subjects participating in a session. Each receives $5 for showing up on time, is randomly assigned the role of Player 1 or 2, and randomly paired with another individual. The session begins with a single play of the sequential decision problem and then subjects are paid according to their decisions. The experimenter announces that there is another experiment where all of the above is the same except that the decision problem is repeated 10 times, each time with a new counterpart. At the completion of the second experiment, subjects are privately paid their earnings in cash. A total of 156 pairs have participated.

We can manipulate how subjects think about each other by using the descriptive labels of 'opponent' or 'partner' in the instructions, rather than the more neutral description of 'counterpart'. This small variation in the reference to the other individual in the pairing has a significant impact on behavior in the above game. Players 2 are more trusting (i.e. choose down at x_2) in the Partner treatment than in the Opponent treatment: 29 percent versus 21 percent. This difference is statistically significant ($p < 0.001$). Furthermore, Players 2 are twice as likely to be trustworthy (i.e. choose the cooperative outcome) in the Partner condition than under the Opponent: 68 percent versus 33 percent. This difference is also statistically significant ($p < 0.04$). These results demonstrate that by providing subjects with a subtle cue to aid mind reading in the decision problem, behavior moves toward more cooperation.

NORMAL VERSUS EXTENSIVE FORM GAMES

In the above experiment, agents make their decisions sequentially – Player 1 moves first, followed by Player 2's move, until an outcome is reached. Player 2 sees the move of Player 1 before Player 2 chooses to move. This is known as the *extensive form* of the game. In the *normal form* of the same game, each player chooses a move at each node without knowing whether that node will be reached in the move sequence. That is, players make their decisions simultaneously, not knowing the choice of the other.

To represent the trust game in Figure 1 in the normal form (see Figure 2), we express the payoff consequences of all possible sequences of moves. Each row represents a strategy for the first player. Similarly, each column represents a strategy for the second player. The first row contains the strategy choose right at x_1, ending the game. The strategy is indicated by R^*. The outcome reached is *(7, 14)*

	r	*d*
*R**	(7, 14)	(7, 14)
D, R	(8, 8)	(10, 10)
D, D	(8, 8)	(12, 6)

Figure 2. Normal form representation.

regardless of what the second player chooses. The second row contains the strategy down at x_1 and right at x_3, *D,R*. If the second player chooses right (*r*) then the outcome is (*8, 8*), but if she chooses down (*d*), then the outcome reached is the SPE *(10, 10)*. Lastly, the third row contains the strategy down at x_1 and down at x_3, *D,D*. If the second player chooses right (*r*) then the outcome is cooperative *(10, 10)*, but if she chooses down (*d*), then the outcome reached is defection *(12, 6)*.

The standard game-theoretic prediction is that rational behavior will be equivalent regardless of the form; behavior is invariant to the form of the game. However, when decisions are made sequentially, it allows intentions to be detected from the actual move. Only upon observing an actual move down at x_2 can Player 1 interpret Player 2's intentions; namely, that both parties can be better off at the cooperative outcome relative to the SPE. As a result, if the reading of intentions is an important component to whether or not cooperation is achieved, then we arrive at a different prediction – namely, cooperation levels in the extensive form (see Figure 1) will be greater than those achieved under the same game in the normal form (see Figure 2).

A total of 26 subjects participated in the extensive form trust game (see Figure 1) and a total of 29 participated in the normal form trust game (see Figure 2). The data being reported here is across-subject comparisons and are for experiments where the subjects participate *once* and only once in the decision problem (data are also available for repeated matching with the same counterpart).

Individual behavior does differ between the two game forms. Table 1 reports the frequencies of decisions. The proportion of ($10, $10) cooperative outcomes being reached is higher in the extensive form than in the normal form: 50 percent in the extensive form, and only 14 percent in the normal form. Using a one-tailed binomial test, the null hypothesis that the proportion of cooperative outcomes is the same under both game forms can be rejected at the 95 percent level of significance. Also the proportion of ($8, $8), subgame perfect, outcomes reached is higher in the extensive form than in the normal form: 100 percent in the extensive and a smaller 82 percent in the normal. Again using a one-tailed binomial test, the null hypothesis that the proportion of SP outcomes is the same can be rejected at the 95 percent level of significance. These results suggest that the extensive form allows better coordination among non-cooperators as well. Both results are not predicted by the standard game-theoretic analysis, but are indeed explained by a reciprocity interpretation which relies on ToM and intentionality detection.

Table 1. Frequencies of play in extensive versus normal form

Outcome	*Extensive form*	*Normal form*
($7, $14)	$\frac{0}{26} = 0.0$	$\frac{0}{29} = 0.0$
Down by P2	$\frac{12}{26} = 0.46$	$\frac{7}{24} = 0.29$
($8, $8)	$\frac{6}{12} = 0.50$	$\frac{6}{7} = 0.86$
($10, $10)	$\frac{6}{12} = 0.50$	$\frac{1}{7} = 0.14$

CONCLUSION

If intentions and ToM were unrelated to behavior in experimental exchange games, then these differences would be puzzling. That we see such differences across variations in labeling (partners versus opponents) and in game form (normal versus extensive) suggests that ToM and folk psychology play an important role in how agents solve game theoretic problems.

The experimental data reported here are from undergraduates with a variety of majors attending the University of Arizona. Another interesting population of individuals, whose behavior has yet to be explored in these particular exchange games, is autistics. Individuals with autism suffer from 'mindblindness', blind to the existence of beliefs, intentions, and other mental states. Given this, perhaps their behavior would be significantly less trusting and less trustworthy. Experiments have been run with children using a variety of bargaining environments to explore whether trust and reciprocal tendencies change with age. Results suggest that this is the case.

Standard solution concepts and the concept of common knowledge rely tacitly on folk psychology anyway, but the ToM standard theory the latter relies on is impoverished. By making the relationship between ToM and economic behavior explicit, we can better understand how real economic agents solve problems of strategic interaction.

Further Reading

Baron-Cohen S (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Baron-Cohen S, Tager-Flusberg H and Cohen D (eds) (2000) *Understanding Other Minds*. New York: Oxford University Press.

Burnham T, McCabe K and Smith V (2000) Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization* **43**: 57–73.

Cosmides L and Tooby J (1992) Cognitive adaptations for social exchange. In: Barkow J, Cosmides L, and Tooby J (eds) *The Adapted Mind*, pp. 163–228. New York: Oxford University Press.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Fehr E and Gächter S (2000) Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives* **14**: 159–181.

Harbaugh W, Krause K, Kiday S and Vesterlund L (in press) Trust in children. In: Ostrom E and Walker J (eds) *Trust, Reciprocity, and Gains from Association: Interdisciplinary Lessons from Experimental Research*. New York: Russell Sage Foundation.

McCabe K, Houser D, Ryan L, Smith V and Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* **98**: 11832–11835.

McCabe K, Rigdon M and Smith V (in press) Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*.

McCabe K, Smith V and LePore M (2000) Intentionality detection and 'mindreading': why does game form matter? *Proceedings of the National Academy of Sciences of the USA* **97**(8): 4404–4409.

Osbourne M and Rubinstein A (1994) *A Course in Game Theory*. Cambridge, MA: MIT Press.

Thought Suppression and Mental Control

Advanced article

Daniel M Wegner, Harvard University, Cambridge, Massachusetts, USA

CONTENTS

Consciously attempting not to think about something is a mental control strategy known as thought suppression. This strategy can be successful under certain conditions, but it often promotes an increase in the accessibility of the thought to consciousness, and along with this, a number of ironic processes and unwanted effects.

INTRODUCTION

In the fifth act of *Macbeth*, the king is distraught over his wife's apparent mental illness. He asks the doctor:

> Canst thou not minister to a mind diseased,
> Pluck from the memory a rooted sorrow,
> Raze out the written troubles of the brain
> And with some sweet oblivious antidote
> Cleanse the stuff'd bosom of that perilous stuff
> Which weighs upon the heart?

The doctor replies:

> Therein the patient
> Must minister to himself.

Shakespeare's observation in the voice of this doctor still holds true some four centuries after it was written. In the matter of unwanted memories, sorrows, troubles, and weights upon the heart, we are often alone in life – left with no one who can clear our minds and solve our problems for us. Instead, we must somehow deal with these things by ourselves.

When unwanted thoughts arrive, people often minister to themselves by trying not to think about them. They attempt to exert mental control. The realization that everyone does this led to much of Freud's insight into human psychology, and more recently has led to a number of inquiries into the nature and effectiveness of people's attempts to control their own minds (Wegner, 1989; Wenzlaff and Wegner, 2000). The general finding from research on this topic to date is that thought suppression is sometimes possible, but that it can produce a host of side effects that are sometimes more damaging, and certainly more far-reaching, than the pain the person may experience on allowing the unwanted thought to dwell in the conscious mind.

UNWANTED THOUGHTS

What are the thoughts that people want to keep out of their minds? The thought of an old love affair that went wrong, the thought of a cake that will break one's diet, the thought of a feared event in the future, the thought of a secret one is hoping not to divulge: these are all examples of thoughts that normal individuals might not want. And people who are suffering from psychopathological disorders such as depression, anxiety, phobia, or obsessive–compulsive disorder, often find that the central feature of their psychological problem is the struggle to avoid a particular set of thoughts.

One approach to avoiding such thoughts is simply to try not to think of them. People who are asked to suppress a thought in the laboratory while they report their stream of consciousness usually mention selecting distractor thoughts to think about instead.

REBOUND AND HYPERACCESSIBILITY

In many laboratory studies, people have been asked to suppress thoughts in just this way. Individuals were prompted not to think about a white bear (something once mentioned by Dostoevsky as impossible to keep out of mind) (e.g. Wegner *et al*., 1987). In think-aloud recordings taken over the course of five minutes, people continued to mention white bears about once per minute. These participants were then asked to go ahead and think

about a white bear for a subsequent five-minute session. Their reports of the thought became more frequent over this expression period, but in a pattern radically unlike that of other participants who had been asked to think about a white bear without prior suppression. Those who were invited to express the thought after suppression appeared to become preoccupied with it – exhibiting a post-suppression rebound effect. This effect has since been observed when people are asked to suppress thoughts of pain, to suppress unhappy thoughts that are spoiling their mood, or to suppress thoughts of a lost love whose absence they grieve. In each case, initial suppression increases the frequency of return of the thought once suppression is discontinued.

Suppression appears to yield even more intense levels of preoccupation with a thought than does concentration. This is apparent not only after suppression is released, but even during suppression when the person is working under stress or mental load. People trying not to think about a target thought show hyperaccessibility – the tendency for the thought to come to mind more readily even than a thought that is the focus of intentional concentration – when they are put under an added mental load or stress. Trying not to think about a target word under conditions of mental load (while rehearsing a long phone number, for instance) makes people unusually slow at identifying the color in which the target word is presented (e.g. Wegner and Erber, 1992). The word seems to jump into the mind before the color, and interferes with the task of naming it. By this measure, unwanted thoughts are more accessible even than thoughts on which a person is intentionally concentrating.

IRONIC PROCESSES OF THOUGHT

These observations can be explained by a theory of ironic processes. The attempt to suppress a thought seems to conjure up an ironic psychological process which then works automatically against the very intention that set it in motion. The suppressed thought is brought to mind in sporadic intrusions because of this sensitivity. Later, when suppression is over, the automatic and intrusive return of the thought apparently continues, in a post-suppression rebound.

Why might such ironic processes occur? One explanation is that ironic processes are part of the machinery of mental control (Wegner, 1994). It may be that in any attempt to control our minds, two processes are instituted: an 'operating' process that works consciously and effortfully to carry out our desire, and an 'ironic' process that works unconsciously and automatically to check whether the operating process is failing and needs renewal. In the case of thought suppression, the operating process involves the conscious and effortful search for distractors (as we try to fasten our minds on anything other than the unwanted thought), whereas the ironic process is an automatic search for the unwanted thought itself. The ironic process is a sort of monitor, which determines whether the operating process is needed, but which also has a tendency to influence the accessibility of conscious mental contents. It ironically enhances the sensitivity of the mind to the very thought that is being suppressed.

An ironic process theory can explain more than the paradox of thought suppression. Such processes may be involved in almost everything we try to do with our minds. If an ironic process is inherent in the control system whereby we secure whatever mental control we do enjoy, then it ought to be evident across many domains in which we do have some success in controlling our minds. Because the operating process requires conscious effort and mental resources, it can be undermined by distraction, and evidence of ironic processes will then arise. When people undertake to control their minds while they are burdened by mental loads – such as distractors, stress, or time pressure – the result should, according to this model, often be the opposite of what they intend. Studies have uncovered evidence of many ironic effects. Ironic mood effects occur, for example, when people attempt to control their moods while they are under mental load. Individuals following instructions to try to make themselves happy become sad, whereas those trying to make themselves sad actually experience a happier mood.

Ironic effects occur in the self-control of anxiety. People trying to relax under load show psychophysiological indications of anxiousness, whereas those not trying to relax show fewer indications.

Ironic effects occur in the control of sleep. People who are encouraged to 'fall asleep as quickly as you can' as they listen to raucous, distracting music stay awake for longer than those who are not given such encouragement.

Ironic effects occur in the control of movement, arising when people try to keep a handheld pendulum from moving in a certain direction, or when they try to keep from overshooting a golf putt. In both cases, an imposition of mental load makes individuals more likely to commit exactly the unwanted action.

Ironic effects also arise from thoughts of death. After people have been asked to reflect for a while on their own death, they spontaneously suppress the thought. Those who are then distracted with stressful tasks show high levels of accessibility of death-related thoughts.

Ironic effects have also been observed in person perception. When people are put under mental load while they are forming impressions of a person, they project a personality trait onto the target when they are suppressing thoughts of that trait – whether they are suppressing in response to suppression instructions, or spontaneously because they dislike the trait in themselves.

Ironic effects also occur when people try to control their prejudices. Bodenhausen and Macrae (1998) report, for example, that people who are trying not to stereotype a skinhead as they form an impression of him show greater stereotyping under mental load. Individuals in this circumstance have been found to avoid even sitting near the skinhead. And people under mental load who are specifically trying to forget the stereotypical characteristics of a person (in a directed forgetting study) have been found to be more likely to recall those characteristics than people without such load.

THE ABILITY TO CONTROL ONE'S MIND

These studies illustrate how things can go awry when we 'minister to ourselves'. People often begin on the path towards ironic effects when they try to exercise good intentions – to behave effectively, to avoid prejudice, to be happy, to relax, to clear their minds of negative thoughts or thoughts of personal shortcomings, or even just to sleep. The intention is to minister to oneself, but it may be the first step towards ironic effects.

The next step towards ironic effects is the pursuit of such goals in the face of a shortage of mental resources. When there is insufficient time and thought available to achieve the intention, people do not merely fail to produce the mental control they desire. Rather, the ironic process goes beyond 'no change' to produce an actual reversal. The opposite of the desire happens. Ironic effects are precipitated when we try to do more than we can with our minds.

Why would we do this? At the extreme, we do this when we are desperate: we will try to achieve a particular sort of mental control even though we are mentally exhausted. These circumstances are very reminiscent of the circumstances of many people suffering from various forms of psychological disorder. People who are anxious, depressed, traumatized, obsessed, or suffering from disorders of sleep, eating, movement, and so on, might frequently try to overcome their symptoms – and might be inclined to attempt such control even under adverse conditions of stress or distraction. Evidence from correlational studies suggests a possible role for ironic processes in several such forms of psychopathology.

Another line of evidence suggesting a role for ironic processes in the beginnings of some disorders comes from studies of what happens when mental control is rescinded. When people are encouraged to express their deepest thoughts and feelings aloud or in writing, and so to suspend any suppression of these thoughts, they experience subsequent improvements in psychological and physical health (Pennebaker, 1997). Expressing oneself in this way involves relinquishing the pursuit of mental control, and so eliminates a basic requirement for the production of ironic effects. The motive to keep one's thoughts and personal characteristics secret is strongly linked with mental control. Disclosing these things to others, or even in writing to oneself, is a first step towards abandoning what may be a futile quest to control one's own thoughts and emotions. In the pursuit of mental control, we may sometimes be most successful when we choose not to minister to ourselves.

References

Bodenhausen GV and Macrae CN (1998) Stereotype activation and inhibition. In: Wyer RS (ed.) *Advances in Social Cognition*, vol. XI, pp. 1–52. Mahwah, NJ: Lawrence Erlbaum.

Pennebaker JW (1997) *Opening Up: The Healing Power of Expressing Emotions*. New York, NY: Guilford Press.

Wegner DM (1989) *White Bears and Other Unwanted Thoughts*. New York, NY: Viking/Penguin.

Wegner DM (1994) Ironic processes of mental control. *Psychological Review* **101**: 34–52.

Wegner DM and Erber R (1992) The hyperaccessibility of suppressed thoughts. *Journal of Personality and Social Psychology* **63**: 903–912.

Wegner DM, Schneider DJ, Carter SR and White TL (1987) Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology* **53**: 5–13.

Wenzlaff RM and Wegner DM (2000) Thought suppression. In: Fiske ST (ed.) *Annual Review of Psychology*, vol. LI, pp. 59–91. Palo Alto, CA: Annual Reviews.

Time Perception

Introductory article

Russell M Church, Brown University, Providence, Rhode Island, USA

CONTENTS

People and other animals are able to perceive the duration of intervals between events, and the accuracy of their perceptions can be assessed. In situations in which there are many different time intervals, these can be combined for the assessment of the typical interval. Associative learning is dependent upon time perception, and the mechanisms of time perception involve an internal clock.

PERCEIVED TIMING OF EVENTS

A clear distinction should be made between the perception of time as a psychological dimension and the measurement of time as a physical dimension. For the perception of the time interval between two events, a person does not need to use any special timing devices such as a watch. Without any external aids, a person can report reasonably accurately the duration between two events (the method of estimation) or make two responses separated by approximately the same interval that separated the two events (the method of reproduction). Such methods are known as 'psychophysical methods' because they have a physical input (the physical duration between the two events) and a psychological output (the reported duration or the interval between two responses). Such outputs could equally well be called 'behavioral', because the person makes some verbal, motor, or other response. (*See* **Psychophysics**)

The accuracy of the perception of time measured in absolute units, such as seconds, generally decreases as the physical duration of the interval increases. But the accuracy of the perception of time measured in relative units (the proportion of the physical duration) is roughly constant over a wide range of intervals from fractions of a second to hours. The increment in physical time required to produce a particular change in perceived time in some experiments with human participants has been found to be about 5 percent of the physical time. This ratio has been called the Weber fraction, and the rule that it is approximately constant is an example of Weber's law. Deviations from constancy, such as a higher Weber fraction at very short intervals and low fractions at certain intervals such as the circadian range of about 24 hours, has provided evidence regarding the biological mechanisms responsible for the perception of time.

Because the psychophysical methods used for the study of the perception of time involve measurements of behavior as a consequence of physical input, such as the interval between two events, they can be used with nonverbal animals. For example, a rat or a pigeon can be readily trained to make one response if a noise stimulus is short (such as two seconds) and another response if a noise stimulus is long (such as eight seconds). The training consists of rewarding a correct response, but not an incorrect response. Then stimuli of intermediate durations can be added. The probability of the long response (the response that was rewarded following an eight-second stimulus) increases as a function of the duration of the stimulus. This is evidence that animals perceive the duration between two events – in this case, the onset and termination of the noise stimulus. Similar experiments with human participants have produced similar results. The similarities between humans and other animals are more than superficial because the relationships between the stimulus duration and the response have many of the same characteristics. For example, these functions are approximately the same for different stimulus ranges when time is examined in relative units (Weber's law); the duration that individuals are equally likely to call short or long is approximately at the geometric mean between the original training durations. Such similarities have led to the hypothesis that humans and other animals use similar mechanisms for the perception of time. (*See* **Animal Cognition**; **Animal Learning**)

TEMPORAL INTEGRATION

With the ability to estimate time intervals, people and other animals are able to choose rewards that occur immediately in preference to those that occur only after some delay. For example, a rat in a maze with two paths will choose the path that leads immediately to food rather than the one in which it receives food only after some delay interval; and a pigeon in a box with two plastic disks will peck the one that leads to immediate rather than delayed delivery of food. Of course, the preference for immediate food can be reduced by delivering substantially more food if the animal chooses the delayed reward. The choice of a larger, more delayed, reward has been called 'self-control', and the conditions influencing it have been studied extensively.

In foraging situations, an animal may obtain many rewards spread over time in some irregular manner. In some patches the rewards may occur at a higher rate than in others, and the animals tend to choose the patch with the higher reward rate. The behavior is often well characterized by the 'matching law' such that the proportion of time spent in each patch is equal to the proportion of rewards obtained in each patch. The basis for this behavior may be related to the way in which animals combine the many times between successive rewards. For example, animals choose a situation in which food occurs at random times over one in which food occurs at fixed times at the same average frequency. One proposal is that they average the local rates of food delivery, rather than the times between food deliveries.

TIMING AND ASSOCIATION LEARNING

Until recently, the study of time perception and classical conditioning have had quite separate histories. The origin of the study of time perception was in the psychophysical laboratories, in which investigations were made of the effect of various conditions on the accuracy of temporal judgments. The conditions included variations in the stimulus to be judged, other task variables, and drug effects. Most of the research was conducted with human participants. In contrast, the origin of the study of classical conditioning was in the learning laboratories of Pavlov, in which investigations were made of the conditions affecting the acquisition of performance of a particular response. The research was conducted on animals. Many of the animal studies of association learning involve time perception. (*See* **Conditioning**; **Learning, Psychology of**)

Pavlov's experiments were based on salivary conditioning of dogs. They involved the presentation of stimuli (visual, auditory, or tactile conditioned stimuli) and reinforcements (unconditioned stimuli, such as food), and the recording of responses (salivary responses). Temporal conditioning involved the presentation of the reinforcements at fixed times. This procedure led to salivation primarily near the end of the interval – a result that indicated that the time interval between successive reinforcements was serving as a stimulus. Another procedure, delayed conditioning, involved the presentation of a reinforcement after a stimulus had been ongoing for a particular length of time. This procedure also led to salivation primarily near the end of the interval – a result that indicated that the time interval from the onset of the stimulus serves as a stimulus. These, and many other, procedures used by Pavlov provided convincing evidence that a stimulus change can serve as a time marker and that the behavior of animals can be related to the time following a time marker. The conclusions of Pavlov have been verified and greatly extended by subsequent research with other animals (mostly rats and pigeons, but also many other species) on a wide variety of classically conditioned responses.

OSCILLATORS, INTERNAL CLOCKS, AND OTHER METHODS OF KEEPING TIME

Although people can approximate the time without the use of external aids, they can do so much more accurately with the use of clocks. As needs have arisen for synchronization of activity and other purposes, increasingly accurate and smaller timekeeping devices have been developed and become available to people. Although these are human inventions designed to solve particular problems, it is possible that these engineering solutions used some of the same features that have evolved into an internal, biological, clock.

One type of internal clock that has been proposed consists of a pacemaker, a switch, and an accumulator. The pacemaker emits pulses with some mean rate and distribution form; the switch is either open or closed, and when the switch is closed the pulses are sent to the accumulator. The perceived duration that the switch was closed is represented by the number of pulses in the accumulator. Another type of internal clock that has been proposed consists of an oscillator that has some mean period. The perceived duration is the phase of the oscillator (or the phase difference, if the oscillator is not reset

at the onset of the time marker). Any process that regularly changes as a function of time following a time marker provides the basis for an internal clock, and many other processes have been proposed. In addition to a single accumulator, oscillator, or other function, a time marker may elicit multiple processes such that the perceived duration is the state of all of these functions.

The relationship between an internal clock for the perception of durations between arbitrary intervals and a circadian clock is uncertain. The circadian clock is an oscillatory process with a period of about 24 hours that can be entrained to light that occurs within a few hours of this period, and that can run freely at about that period even in a constant environment without periodic cues. The phase of a circadian clock provides a time-of-day that can be used as a discriminative stimulus, and such periodic processes may be involved in the timing of arbitrary intervals. (*See* **Circadian Rhythms**)

An internal clock mechanism provides a representation of the time, but it is not sufficient for timed performance. This requires a temporal memory and decision process. On the basis of previous experience, temporal memory stores the perceived time of previous reinforcement. The decision to respond is based on a comparison of the perceived time (from the internal clock) and the remembered time (from temporal memory). If these two values are close enough, a response is made; otherwise it is not.

Quantitative models of timing involve specific assumptions about the perception of time, temporal memory, and decision processes. One example is *scalar timing theory* that makes explicit assumptions about the internal clock as a pacemaker–switch–accumulator system, temporal memory as a memory of specific examples, and a decision process that involves a ratio comparison of the current value in the accumulator and a random sample of a value from temporal memory. Another example is the *behavioral theory of timing* that makes explicit assumptions about the representation of time as a behavioral state and temporal memory as based on the strength of each of the behavioral states due to their frequency of reinforcement. The goal of these theories is to provide a simple and accurate description of behavior in a wide range of conditions.

Further Reading

Bradshaw CM and Szabadi E (eds) (1997) *Time and Behaviour: Psychological and Neurobehavioural Analyses*. Amsterdam, Netherlands: Elsevier.

Fraisse P (1963) *The Psychology of Time*, translated by J Leith. New York, NY: Harper & Row.

Gallistel CR (1990) *The Organization of Learning*. Cambridge, MA: MIT Press.

Herrnstein R (1997) *The Matching Law: Papers in Psychology and Economics*. New York, NY: Russell Sage Foundation.

Landes DS (1983) *Revolution in Time: Clocks and the Making of the Modern World*. Cambridge, MA: Harvard University Press.

Macey SL (ed.) (1994) *Encyclopedia of Time*. New York, NY: Garland.

Moore-Ede MC, Sulzman FM and Fuller CA (1982) *The Clocks that Time Us: Physiology of the Circadian Timing System*. Cambridge, MA: Harvard University Press.

Pavlov IP (1927) *Conditioned Reflexes*. Oxford, UK: Oxford University Press.

Rachlin H (2000) *The Science of Self-control*. Cambridge, MA: Harvard University Press.

Richelle M and Lejeune H (1980) *Time and Animal Behaviour*. Oxford, UK: Pergamon.

Roeckelein JE (2000) *The Concept of Time in Psychology: A Resource Book and Annotated Bibliography*. Westport, CT: Greenwood.

Rosenbaum DA and Collyer CE (eds) (1998) *Timing of Behavior: Neural, Psychological, and Computational Perspectives*. Cambridge, MA: MIT Press.

Whitrow GJ (1988) *Time in History: Views of Time from Prehistory to the Present Day*. Oxford, UK: Oxford University Press.

Topology and Cognition

Advanced article

Roberto Casati, Centre National de la Recherche Scientifique, Paris, France

CONTENTS

Spatial representation as registered in vision and language appears to be sensitive to topological properties, which are among the most general spatial properties.

INTRODUCTION

We can see that a smiling face changes smoothly into a sad face; we can appreciate the difference between stretching an object and tearing it apart; we can distinguish between an object's being inside or outside another; we see the difference between a pretzel and a donut, or between a broken and a whole glass; we can understand that in a sense the letter B has the same configuration as the digit 8 and that both differ from the digit 2. In all these cases, we acknowledge the topological structure of the scene we perceive or think about. Some important topological properties and relations are those of continuity, interior, exterior, tangential part, boundary, and the presence and number of holes. Such properties have attracted the interest of cognitive scientists, as they may also implicitly enter into descriptions or explanations of cognitive performances. In a visual scene, our ability to distinguish between an area corresponding to a figure and an area corresponding to the background is related to the fact that the figural area is assigned the boundary. Sensitivity to subtle topological differences has been proposed as an explanation of our understanding of spatial prepositions such as 'across' or 'between'. Some psychological theories explain children's understanding of physical objects in terms of sensitivity to topological properties.

A consideration of topological properties seems to be required to account for some of the contents of spatial representation. However, there has so far been no systematic study of topology in cognition. This article will concentrate on some aspects of current research on cognition where topological notions and properties play a key role, whether as subject matter of explicit judgment, or as hidden components of psychological explanation. Although the visual and the linguistic domains will be our main focus, topology enters into the study of other faculties. For instance, haptic continuity has been described, and in the domain of action it has been proposed that reaching is guided by the hypothesis that objects move on continuous paths. In the auditory domain one can find loose analogies for the notion of contact (such as collisions, which may be heard).

We will not discuss the various applications of topology in cognitive science, such as the topology of neural networks, or the topology of color and other quality spaces: these are instruments for studying cognition, not objects of cognition.

WHAT IS TOPOLOGY?

Topology as a branch of mathematics focuses on basic abstract spatial properties. For example, topologically, a square and an ellipse are equivalent. Roughly, topological properties are those that are not altered by 'distortion' and 'stretching'.

In general, we can classify the geometric properties of a shape by noting whether they are preserved in some classes of transformations. Thus, copying or rotating a square will preserve all of its shape properties, such as the length of its sides or the size of its angles. Stretching a quadrilateral may change its angles and the lengths of its sides; therefore, these are not topological properties of the quadrilateral. The transformation that carries a square onto an ellipse seems to destroy all geometrical properties of the square. The square is no longer 'recognizable'; yet some of its properties are preserved in the ellipse. Both the square and the ellipse are closed curves. The projection on the ellipse of a set of points with a certain order on the square will have the same order on the ellipse.

The order of points is not altered through distortion and stretching, and counts as a topological property of the square.

Cognitive science deals with relatively simple and intuitive topological properties of objects, such as being in one piece, or having holes. The mathematical theory of topology encompasses situations that are not as easily graspable as the above examples, but the basic idea is simple. A topological characterization of a set S of 'points' is defined in terms of subsets of S. Subsets of S allow one to separate points of S.

Formally, a 'topology' on a set S is a set T of subsets of S, containing both S itself and the empty set which is closed under arbitrary unions and finite intersections.

A 'topological space' is a set S with an associated topology T. A given set S can have many different topologies, with a different discriminatory power. The 'discrete topology' consists of all subsets of S, and is able to discriminate all points in S.

Continuous functions between topological spaces may be defined in terms of 'neighborhoods' (a generalization of the metric notion of distance). In this way, the notion of a topological space becomes a tool for classifying remarkable properties of figures. A topological space is 'connected' if in its topology no two disjoint nonempty sets exist whose union is the whole space. A topological space S is 'path-connected' if for any two points a and b in S there is a continuous path (that is, a continuous function from the closed unit interval [0,1] to S) whose extremes are a and b. In general, there will be many paths connecting any two points of a path-connected surface. These paths can be grouped into 'homotopy' classes (two paths in the same homotopy class can be transformed into one another via a continuous mapping). If there is only one homotopy class, the space is said to be 'simply connected'.

For instance, a sphere is simply connected. Consider paths whose ends coincide (i.e. loops). Many loops will pass through a given point on a sphere, but they are all homotopic, as they can all be continuously shrunk to a point. On the other hand a toroidal (donut-shaped) surface is not simply connected: not all loops through a given point can be shrunk to a point. Nor is a plane from which a finite portion has been removed.

In general, surfaces can be classified in terms of the homotopy classes of the loops we can draw on them. The presence of distinct homotopy classes in a space intuitively indicates the presence of features that prevent loops from shrinking. These features are typically holes and handles.

RELEVANCE OF TOPOLOGY TO COGNITION

Topology enters cognitive science as the object of a special, indeed fundamental, case of spatial representation. Topology-related spatial representation can be isolated within the class of 'non-immersed' spatial representations, as opposed to immersed representations such as 'to the right of', whose interpretation depends on a viewpoint embedded in the representation (be it object- or view-centered). For instance, 'is one meter long' or 'is inside the box' are non-immersed representations, as their applicability is not affected by changes in point of view. This ball is or is not inside this box no matter how we look at these two objects, whereas this ball is to the right of this box only from a certain viewpoint.

Within the class of non-immersed spatial representations one can further distinguish the representations that are size- and shape-related from those that are not. We can perceive and make judgments about spatial facts without necessarily having to assess the size and distance of the objects involved. One can, for instance, judge that John is between Paul and Mary without accessing any specific metric information concerning the relative sizes and distances of these three people. Typically, topological representation abstracts from size and shape. Moreover, the output of a topological judgment is discrete (e.g. 'yes' or 'no') whereas the output of a metric judgment is continuous.

Topological properties appear to be required to describe some of the contents of spatial representation. But do we possess an implicit knowledge of topology? At what level is this knowledge accessed, and how is it used? Are topological relations computed 'online' at some level?

We can distinguish here two types of evidence: direct evidence, from personal or sub-personal assessment of topological facts, and indirect evidence, stemming from an analysis of data that do not concern an explicit topological representation but seem to require some access to topological representation or processing.

Direct Evidence for Topological Analysis

Consider an object such as the Möbius strip (a long rectangular strip one of whose ends has been twisted by 180° and glued to the other end) (Figure 1). If one is asked to predict the result of cutting it along its median line (a line running

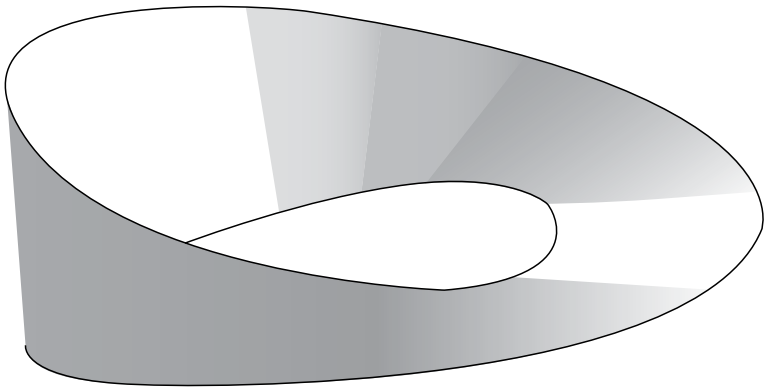


Figure 1. The Möbius strip.

parallel to its borders), one may predict that two strips will result. Actually, one gets one doubly-twisted Möbius strip. The border of a Möbius strip is strange. It contravenes the intuitive principle that a border must separate two sides, so that one cannot reach one side from the other without crossing the border.

Exposure to some technical topology does not, by itself, correct one's propensity to misjudge some topological facts. Those who understand and master simple topological tasks, such as the detection of the topological equivalence between a sphere and a cube, or of a donut and a cup of coffee, may be unwilling to accept that a cylindrical surface (such as the portion of a tube) and a disk with an internal portion removed (such as a CD) are topologically equivalent. (Popular topology books emphasize the oddity of the topological equivalence among objects that look utterly different.) The situation is reminiscent of intuitive physics. It has been found that subjects' understanding of Newtonian physics is blocked by their implicit and automatic reliance on an intuitive theory that is at odds with Newtonian physics. Analogously, there may exist an 'intuitive topology' that prevents us from appreciating actual topological equivalences (or non-equivalences).

Intuitive topological classification may be a by-product of independent criteria of classification that emphasize some other spatial aspects of objects, such as the presence of holes, conceived as objects in themselves (Casati and Varzi, 1994, 1999), or else of higher-order physical principles that dictate plausible shapes for physical objects. Sub-personal direct evidence is methodologically problematic, because of the generality of topological properties. Experimental designs may be insufficiently fine-tuned to detect sensitivity to topological structure. For instance, in experiments about numerosity, a test display with two black spots may follow a habituation setting in which a single black spot is present. Sensitivity to the change can be related to a sensitivity to numerosity (implying a sensitivity to disconnection, hence to a topological property of the setting) but it may as well reflect a sensitivity to the overall quantity of blackened space or to the overall length of the perimeter of the two spots. The experimental setting should control for some of these factors, but topological properties are so basic and ubiquitous that almost any factor interferes with them.

There is evidence that perceptual priming can be independent of size, location, and orientation (Biederman and Cooper, 1990, 1992). Therefore, recognition processes can abstract from these spatial features of objects. Does abstraction go all the way down to the most basic, topological features? Experiments so far have provided little evidence. Chen (1982, 1990) proposes that the extraction of topological properties is one of the primary functions of the visual system. This claim may be supported by the experimental finding that subjects are better at discriminating pairs of topologically distinct figures than pairs of topologically similar but metrically distinct figures, and by the occurrence of illusory conjunctions in displays that involve holes. However, it has been pointed out (Rubin and Kanwisher, 1985) that other factors (such as differences in luminous flux and in the overall perimeter of the figures) could account for most of these results.

Indirect Evidence for Topological Analysis

If sensitivity to topological structure is difficult to assess directly in relation to configurational properties of perceptual displays, it is nevertheless

possible to postulate that some mechanism for evaluating the topology of the display is activated in visual perception.

Some cognitive abilities do not relate directly to topology, but turn out to presuppose sensitivity to the topological parsing of a scene. We shall consider, in turn: the assessment, at personal level, of object unity; the sub-personal segregation of perceptual units; part–whole parsing; infants' mastering of the notion of a physical object; and representational advantages. Further evidence related to language will be discussed afterwards.

Personal assessment of object unity

The display in Figure 2 could be described in two mutually incompatible ways: as two black objects, one on the left and one on the right; or as one black object, partly on the left and partly on the right.

Although the second description sounds utterly artificial, there are no obvious reasons for ruling it out. It seems that in judging what counts as one object in the display we rely on background assumptions concerning some topological properties of the display itself: there are two mutually disconnected black patches, each of which is maximally self-connected. This suggests that we use connectedness as a criterion of unity, and that the visual system is able to compute it. If topology biases the choice, this in turn entails that cognition takes topology into account.

However, topological self-connectedness is not always interpreted as a necessary and sufficient condition for unity. Two juxtaposed objects may count as two even though they are connected; and some kinds of causal unit (such as two objects in coherent motion) may not require connection.

Segmentation of the visual field

Gestalt psychologists have proposed an explanation for perceptual organization (e.g. figure–ground articulation) of the visual field in terms of our tendency to group perceptual units according to principles such as similarity or relative proximity. Perceptual units are discrete elements that are put together and considered as parts of superordinate units in virtue of some organizing factors. However, this account takes for granted the formation of the perceptual units that get grouped, the 'atoms' of the visual field. Palmer and Rock (1994) claim that these units are regions of the visual field that are 'uniformly connected'. Uniform connectedness is not a principle of grouping, insofar as it does not presuppose units but creates them, and so it must operate prior to any grouping activity.

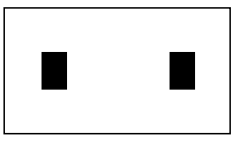


Figure 2. Two black objects or one disconnected black object.

For example, consider Figure 3, in which dotted lines demarcate various regions of the display (they are not part of the image). In (e) the demarcated region is the total black area, and in (f) it is the total white area.

What makes the left black rectangle in Figure 3(a) a "unit", whereas the areas delineated in Figure 3(c) and 3(e) do not correspond to unitary objects? A simple answer is that the area in 3(a) is both uniform (it is all black) and connected (it is a piece). The area indicated in Figure 3(c) is connected, and that indicated in 3(e) is uniform, but neither is a perceptual unit. The areas indicated in Figures 3(b) and 3(d) are uniformly connected, but even they are not perceptual units. One must add the condition of maximality: an image is partitioned into maximal self-connected regions of uniform quality. Note that the complement of the total black area in Figure 3(f) is maximally uniformly connected, and so counts as a unit.

The notions of object unity and unit formation are subtle. Should two regions that are tangent on a point be taken as one region or two? Shape may sometimes influence judgments about unity. Moreover, a 'unit' can still have clearly demarcated parts depending on its shape (if we connected the two black regions with a black line, we would obtain a unitary object with salient parts).

Part–whole parsing

Parsing is the labeling of some perceptual items as salient parts of other items. (Not all parts of a perceptual unit need be salient at all.) Thus, parsing is subsequent to unit segregation. It may follow topological principles, though topological properties and relations are cognitively intertwined in a complex way with part–whole relations. Attempts to reduce topological connectedness to part–whole relations have only shown once more the bias towards topological unity. One may think that the two boxes in Figure 2 are disconnected because any 'bridging' object that has parts in common with them will have parts in common with their complement. But this is so only if the bridging object is a connected unit. If one accepts disconnected units, one may have

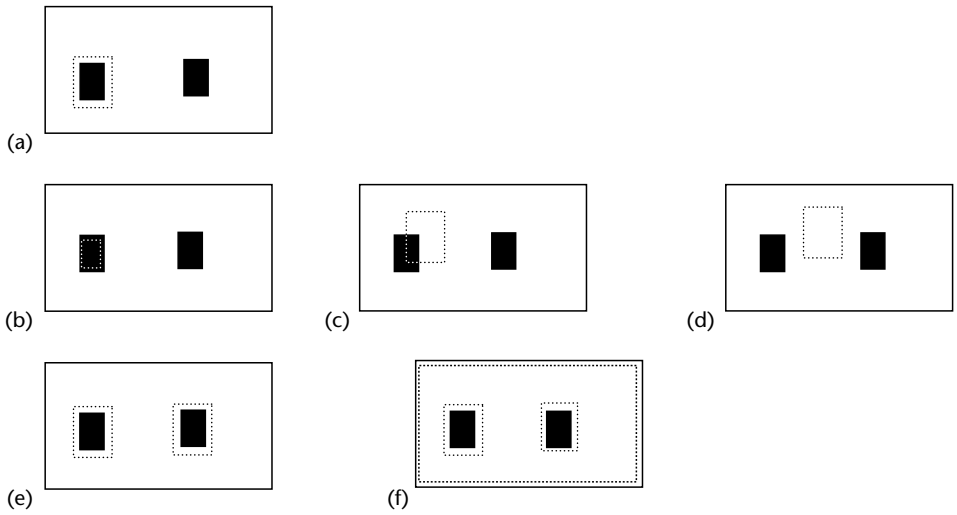


Figure 3. Various candidate 'perceptual units' of a display. Dotted lines demarcate regions of the display (they are not part of the image). In (e) the demarcated region is the total black area; in (f) it is the total white area.

a disconnected bridging object that is exactly composed of the two boxes.

Part–whole relations hold between entities that may not be labelled as salient parts, such as a square and its top half. Surely, a disconnected object such as a broken glass has as many salient parts as there are self-connected units composing it. But a self-connected object may still have salient parts that are individuated by the geometry of the object, typically by concave discontinuities on an object's contour (Hoffman and Richards, 1984). A V-junction is a topologically unitary object and is topologically equivalent to a U; but unlike the U it is parsed as the conjunction of two elements. Similarly, the topological identity of a Y-junction and a T-junction can be overruled by their configurational differences, so that we do make a difference and consider the Y as composed of three bars and the T of two bars.

Physical objects in infant cognition

Infants appear to divide up the world into objects according to some basic principles (Spelke, 1990). In particular, objects are assumed to move along continuous paths, and to interact if and only if they are in contact. These principles require that infant cognition take into account topological features, such as contact and continuity, though not necessarily an explicit perceptual representation. For example, a hidden connection may be inferred from a display if there is an occluder in view.

Same-object advantages

Subjects are faster and more reliable at assessing local changes within a unitary object than they are across disconnected objects. This 'same-object advantage' means that topological unity is relevant at the sub-personal level.

TOPOLOGY IN LANGUAGE

Linguistic data provide a further body of indirect evidence for topological evaluation. English contains a vast number of geometry-related terms for spatially describing objects ('square', 'circle') and properties ('circular'). In principle, we can describe countless shapes by simply building a predicate that mentions an object having that shape ('star-shaped'). Some linguistic uses, though, presuppose an appreciation of rather abstract geometric properties. For instance, the correct use of the preposition 'in' largely abstracts from the shape and size of the objects that stand in the corresponding relation. A fish can be in a pool and in the same sense of 'in' a submarine can be in the ocean.

A broad distinction between two general classes of linguistic items, the open or lexical class and the closed or grammatical class, is believed to reflect two levels of cognitive generality (Talmy, 2000). Membership of the open class is subject to great variability, whereas membership of the closed class is rather fixed. Grammatically significant items (prepositions, suffixes, prefixes, etc.) cluster in the closed class. For instance, prepositions rarely get added to or disappear from English ('betwixt'). A popular hypothesis is that closed-class items tend to capture general facts of cognitive import. This would explain the stability of the class over time. In particular, only certain concepts are

expressed by items in the closed class, and these concepts have a cognitive structuring role. Language does not have grammatical expressions for size, shape, or color. There is no preposition that can be used to convey the idea that an object is square or red. By contrast, the presence of a plane so curved as to define a portion of space is conveyed by uses of the preposition 'in', and these uses abstract from size ('in the thimble', 'in the volcano') and shape ('in the well', 'in the trench').

An important subclass of these 'structuring' concepts concern space and spatial properties and relations that are shape- and size-neutral, as in the example of 'in' above. Similarly, in the sentence 'I walked through the woods', the preposition 'through' expresses a concept that is indifferent to the shape and size of the path followed. These abstract concepts come close to the abstract notions of topology; and it has been suggested (Talmy, 2000) that there could exist a language-based topology whose notions are akin to those of mathematical topology. The notions of this language-based topology would include the notions of point, linear extent, locatedness, withinness, region, side, partition, singularity, plurality, sameness, difference, and adjacency of points. It can be argued that topology as a mathematical theory is a normative refinement of these notions. However, not all topological relations are lexicalized. Thus, there are words for three-place relations, such as 'between', but not for four-place relations.

THEORIES OF TOPOLOGICAL REPRESENTATION

Topological Classification

Topology as a mathematical theory provides a classification of objects' shapes. But cognition does not appear to match closely the classification provided by mathematical topology, except for some very simple surfaces (as demonstrated by the counter-intuitiveness of many topological equivalences). So far there has been no systematic study of the classifications endorsed by cognition. One hypothesis is that shapes are classified in terms of the number and position of holes (tunnels, indentations), where holes are counted as parts of the object or as objects in their own right.

Schematic Spatial Representation

Linguistic representation appears to be relatively independent of the size, the shape, and often the type of the object represented. It has been proposed that the language system may access a particular level of spatial representation, the level of geometric schematization, which is also interfaced with the visual system (Herskovits, 1997). At this level objects are assumed to be represented by simplified geometric schemata that may or may not take into account their metric properties. The structure of prepositions can be used as a heuristic for the study of the level of schematization. How does language access this level? According to Herskovits, prepositions do not refer to schemata, but express constraints on allowable schemata. It is an open question how a proposition such as 'the ball is in the box' is evaluated. Possibly the evaluation relies on a computation performed at the schematization level that answers a 'command' from the linguistic system.

Online Computing

If topological facts are computed at some level in spatial representation, the problem arises of finding the mechanisms that do the computing (Ullman, 1996). Various algorithms have been proposed. For instance, a simple routine to find out whether two marks lie on the same line or on different lines may explore the line starting from one mark (first in one direction and then in the opposite direction) and stop at an endpoint or at the other mark. In order to determine whether a given point is within or without a certain closed region, a 'filling-in' algorithm can color the field starting from the point until the boundaries of the shape are met.

Algorithms of this sort may be too powerful, delivering definite answers where the visual system does not. For instance, in Figure 4, we may not be able to decide whether point *a* is inside or outside the closed curve, or whether point *b* is on the same line as point *c*, whereas a filling-in

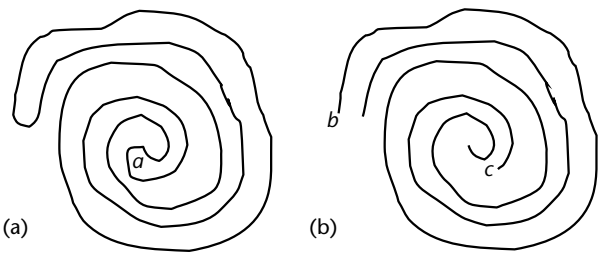


Figure 4. The visual system may not be able to decide (a) whether the point *a* is inside or outside the closed curve, or (b) whether points *b* and *c* are on the same line.

algorithm or line-exploring algorithm, respectively, would deliver a definite answer.

From the fact that topological properties are basic and ubiquitous it does not follow that their determination is simply accomplished. Minsky and Papert (1988) argue that evaluating topological properties may be a complex task. They discuss the limitations of perceptrons, devices capable of assessing whether a given configuration satisfies a given predicate by computing the linear (purely additive) results of collections of weighted predicates, each of which is true or false locally (say, a nonmaximal cell assembly on a retinal sheet). They prove that some spatial predicates, such as 'is convex', can be easily computed by perceptrons, but others, such as 'is connected', are resistant to computation for most classes of perceptrons, and may be computed only by fairly complex systems.

References

Biederman I and Cooper EC (1990) Evidence for complete translational and reflectional invariance in visual object priming. *Perception* **20**: 585–593.

Biederman I and Cooper EC (1992) Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 121–133.

Casati R and Varzi AC (1994) *Holes*. Cambridge, MA: MIT Press.

Casati R and Varzi AC (1999) *Parts and Places*. Cambridge, MA: MIT Press.

Chen L (1982) Topological structure in visual perception. *Science* **218**: 699–700.

Chen L (1990) Holes and wholes: a reply to Rubin and Kanwisher. *Perception and Psychophysics* **47**: 47–53.

Herskovits A (1997) Language, spatial cognition, and vision. In: Stock O (ed.) *Spatial and Temporal Reasoning*, pp. 155–202. Dordrecht, Netherlands: Kluwer.

Hoffman DD and Richards WA (1984) Parts of recognition. *Cognition* **18**: 65–96.

Minsky ML and Papert SL (1988) *Perceptrons*. Cambridge, MA: MIT Press. [Expanded edition. First published 1969.]

Palmer S and Rock I (1994) Rethinking perceptual organization: the role of uniform connectedness. *Psychonomic Bulletin and Review* **1**: 29–55.

Rubin JM and Kanwisher N (1985) Topological perception: holes in an experiment. *Perception and Psychophysics* **37**: 179–180.

Spelke ES (1990) Principles of object segregation. *Cognitive Science* **14**: 29–56.

Talmy L (2000) *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Further Reading

Halligan PW and Marshall JC (1993) When two is one: a case study of spatial parsing in visual neglect. *Perception* **22**: 301–312.

Hilbert D and Cohn-Vossen S (1990) *Geometry and the Imagination*, 2nd edn, translated by P. Nemenyi. New York, NY: Chelsea. [A study of spatial imagination in mathematics, first published in 1932.]

Mendelson B (1990) *Introduction to Topology*. New York, NY: Dover. [Technical but accessible introduction to mathematical topology.]

Piaget J and Inhelder B (1967) *The Child's Conception of Space*. New York, NY: WW Norton.

Scholl BJ (2001) Objects and attention: the state of the art. *Cognition* **80**: 1–46. [The state of the art on various types of object advantages.]

Ullman S (1996) *High-Level Vision*. Cambridge, MA: MIT Press.

Vandeloise C (1994) Methodology and analyses of the preposition 'in'. *Cognitive Linguistics* **5**: 157–184.

Vision: Early Psychological Processes

Intermediate article

Patrick J Bennett, McMaster University, Hamilton, Ontario, Canada

CONTENTS

Early visual processes encode the spatial and temporal distributions of light intensity in the two retinal images that carry information about the arrangement of surfaces in a scene. These early processes are important for visual perception because they constrain higher-level processes such as object recognition.

INTRODUCTION

Vision is a process whereby the information contained in the two retinal images is used to create representations of the layout of surfaces in a scene. These representations are derived from various sources of information (e.g., the spatial arrangement of image contours, patterns of retinal motion, gradients of texture, etc.) and are used for a variety of purposes, including object recognition, regulation of posture, and the control of reaching and grasping movements.

This diversity of tasks and information sources means that it is unlikely that visual processing is performed with a single set of algorithms or representations. Instead, there is ample evidence to support the view that the primate visual system consists of a collection of modules, each specialized to use a particular source of information to perform a specific task (Marr, 1982). Nevertheless, it is important to note that all such modules ultimately depend on an accurate representation of the structure in retinal images. The primary function of early visual processes is to create this initial representation by first detecting changes in intensity that produce contours in the retinal image, and then encoding the size, orientation, and motion of these contours. These early processes must create a rich and detailed representation of image structure, because information that is lost in these initial stages will constrain higher-level processes such as object recognition.

SENSITIVITY OF THE EYE

Sensitivity to changes in intensity depends on many factors, including stimulus size, duration, wavelength, position on the retina, and intensity of the background. In a classic paper (Hecht *et al.*, 1942), Hecht and colleagues adjusted these factors to create viewing conditions that maximized sensitivity, and they found that human observers could reliably detect spots of light that delivered 50–150 quanta to the surface of the cornea. Taking into account the number of quanta that were reflected and absorbed by the pre-retinal media, this just visible stimulus results in 5–15 quanta being absorbed by rhodopsin, which is the photopigment contained in rod photoreceptors. Because the retinal image of the spot covered approximately 500 rods, the probability of any single rod being stimulated by two or more quanta was very low, so Hecht and colleagues concluded that a single quantum is sufficient to produce a response in a rod. It was not until 35 years later that this startling result was confirmed by physiological measurements.

Such exquisite sensitivity offers tremendous advantages in conditions of dim illumination, which probably explains why rods that respond to a single quantum are common among vertebrates. However, it also presents a problem, namely how the large range of visual inputs can be represented by sensitive neurons that have a limited dynamic range. The range of light intensities that we encounter in our natural environment is enormous (approximately 12 log units), but visual neurons respond differentially to light only over a much narrower range. To overcome this problem, the visual system has multiple mechanisms for adjusting sensitivity to match ambient light levels. These mechanisms underlie the phenomena of dark and light adaptation.

ADAPTATION

If you walk into a dimly lit theater after being outside in bright light, everything appears dark. However, after a few minutes your eyes adapt to the darkness and it is surprisingly easy to perceive objects. Conversely, you can be dazzled by bright sunshine after emerging from the theatre, but your eyes adjust rapidly. Recovery of sensitivity in total darkness is referred to as dark adaptation, whereas recovery of sensitivity to a non-zero background of light is called light adaptation. In both cases, the time needed for recovery is a function of the change in illumination, but dark adaptation is typically a much slower process. Complete dark adaptation occurs within several minutes after exposure to moderate intensities of light, but requires more than 30 minutes after exposure to a very bright flash. The amount of adaptation can be substantial. After exposure to a bright light, the detection threshold can drop by more than a factor of 10 000 after about 30 minutes in the dark (Figure 1).

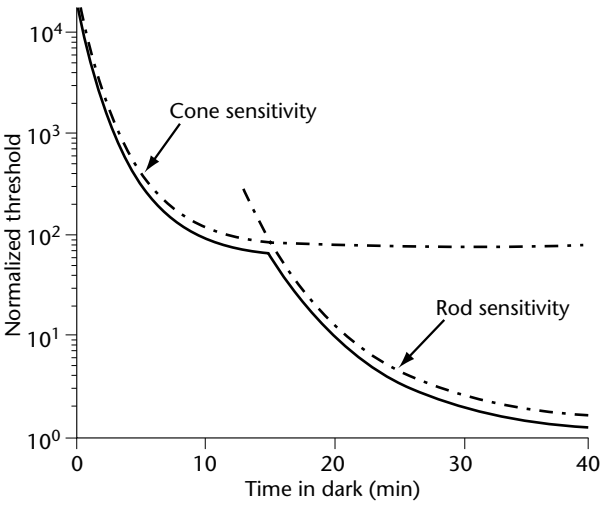


Figure 1. The effects of dark adaptation on detection threshold after exposure to a very bright light. A typical experiment begins with the observer looking at a large bright uniform light for several minutes. At time zero, the uniform light is extinguished and the observer must detect a small spot of light presented against a dark background. The detection threshold is measured periodically, and the solid curve illustrates how it changes as a function of time in the dark. In this graph, the threshold has been normalized so that the lowest threshold has a value of 1. Notice that the threshold declines rapidly during the first 5 minutes, levels off at around 10 minutes, and then drops again after approximately 15 minutes. Over a 40-minute period, the threshold declines by more than four log units. Occasionally it is possible to test observers who lack rod photoreceptors (and who therefore have vision mediated solely by cones), and observers who lack cones (and who therefore have vision mediated entirely by rods). Dark adaptation curves for these two types of observers are illustrated by the broken curves labeled *cone sensitivity* and *rod sensitivity*, respectively. It can be seen that cone-mediated vision adapts quickly but levels off at a high threshold, whereas rod-mediated vision adapts relatively slowly but eventually reaches a much lower threshold. Dark adaptation curves for normal observers contain both cone- and rod-mediated components.

Mechanisms of Adaptation

The pupil area decreases by a factor of 16 as illumination is changed from very dim to very bright levels. This change in size contributes to adaptation, but it is much too small to compensate for the enormous changes in illumination that are encountered in natural conditions. Clearly, other mechanisms must contribute to adaptation. One obvious possibility, outlined by Selig Hecht in a series of papers in the 1930s, is that the visual threshold is proportional to the amount of bleached photopigment in the retina. During Hecht's time it was known that absorption of light causes a photopigment molecule to change shape, thereby rendering it insensitive to light, and that a series of biochemical processes converted photopigment from this bleached state back to a light-sensitive state. With proper adjustment of these rate parameters, which were unknown at the time, Hecht's photochemical theory was able to account for a wide range of adaptation data. However, we now know that the photochemical theory is incorrect. Measurements of the proportion of bleached photopigment in the living eye have shown that threshold and bleached pigment are related logarithmically, not linearly as predicted by the photochemical theory. In other words, very small changes in the proportion of bleached photopigment produce very large changes in the visual threshold (Cornsweet, 1970).

The failure of optical and photochemical factors to account for adaptation led to the search for neural adaptation mechanisms. One way of solving the dynamic range problem is to stagger multiple neural mechanisms that differ in sensitivity. The visual system certainly adopts this strategy, with rods mediating vision in dim light (i.e., scotopic conditions) and cone photoreceptors operating in brighter light (photopic conditions). Dark adaptation experiments provided evidence for this duplex retinal organization by showing that the threshold does not drop monotonically as a function of time in the dark. Instead, it first drops rapidly and then levels off. After about 10–15 minutes,

the threshold begins to decline once more. A great deal of research with human observers who lack rods or cones, as well as experiments using techniques that isolate rods or cones in normal observers, has demonstrated that the first and second stages of dark adaptation reflect changes in cone and rod sensitivities, respectively (Figure 1).

Although the duplex retinal organization contributes to adaptation, it cannot be the whole story because significant adaptation occurs within the scotopic and photopic regimes. Thus theories of adaptation have postulated one or more gain control mechanisms that adjust the sensitivity of the scotopic and photopic systems. At least part of this gain control exists in the retina, because adaptation can occur separately for each eye. Physiological recordings in the cat have shown that changes in illumination shift the dynamic range of retinal ganglion cells, so that the cells continue to give differential responses across a wide range of light intensities. Physiological recordings suggest that rods and cones in the mammalian retina do not adjust their dynamic range in response to changes in illumination, so post-receptoral mechanisms must contribute to adaptation in ganglion cells. The details of this mechanism are still unknown.

Weber's Law

The high sensitivity reported by Hecht and colleagues (Hecht *et al*., 1942) occurs only when the target is presented against a dark background. As the intensity of the background (I_b) increases, adaptation processes reduce the visual sensitivity so that the change in intensity (ΔI) necessary to detect an object is proportional to the background: $\Delta I = kI_b$. This relationship between threshold and background is known as Weber's law, and it implies that the visibility of an object depends on the ratio of object to background intensities. A common measure of stimulus contrast is $\Delta I/I_b$, so another way of stating Weber's law is to say that pattern visibility depends on contrast, not intensity.

At first glance, this law appears to make perception more difficult, because a just noticeable difference in intensity increases with background intensity. However, Weber's law in fact makes it easier to recognize objects under different levels of illumination. To illustrate this point, consider what happens when we change the intensity of illumination that is falling on an object presented against a uniform background. Let the intensities of the light coming from the object and background be I_o and I_b, respectively, and the intensity difference ($\Delta I = I_o - I_b$) be set to detection threshold. Changing the intensity of illumination alters the object intensity (I_o), the background intensity (I_b) and the intensity difference (ΔI), but the stimulus contrast ($\Delta I/I_b$) remains constant. Thus the fact that sensitivity obeys Weber's law means that contour visibility remains approximately constant across a wide range of illumination levels.

CONTRAST SENSITIVITY FUNCTION

Classical studies of visual sensitivity used very simple stimuli (usually a spot presented against a uniform background) to probe early visual processes. However, it is known that the detection threshold is strongly dependent on the spatial distribution of contrast in a stimulus. For example, the visibility of circular and rectilinear black and white stripes differs dramatically even when the two types of patterns have the same contrast (Kelly and Magnuski, 1975). If the threshold depends on the shape of a pattern, then how can we develop a general characterization of visual sensitivity? One common index of the visual system's sensitivity to spatial patterns is the contrast sensitivity function (CSF). The CSF relates the visibility of a sine-wave grating (a repeating, striped pattern whose intensity varies sinusoidally) to the grating's spatial frequency (Figure 2). For each spatial frequency, threshold is measured by adjusting the grating's contrast until the stripes are just detectable. The CSF plots contrast sensitivity, defined as the reciprocal of the contrast detection threshold, as a function of a grating's spatial frequency (Figure 3).

Sensitivity for sine-wave gratings would seem to have little application to the perception of other patterns. However, such a view would be mistaken. Fourier's theorem implies that any pattern can be represented as the sum of a unique set of sine-wave gratings. Thus the latter can be thought of as a set of elements from which all other patterns can be constructed, with different bands of spatial frequencies corresponding to different types of structure in complex images (Figure 4). Moreover, there is a rich mathematical framework (referred to as linear systems analysis) that can be used to relate the CSF to the perception of arbitrary patterns (Cornsweet, 1970).

Strictly speaking, the application of linear systems analysis requires that the system under study must satisfy certain assumptions. Although the visual system does not satisfy these assumptions in detail, under certain conditions the assumptions are met with sufficient precision for the CSF to be used to characterize the perception of a variety of patterns. For example, the shape

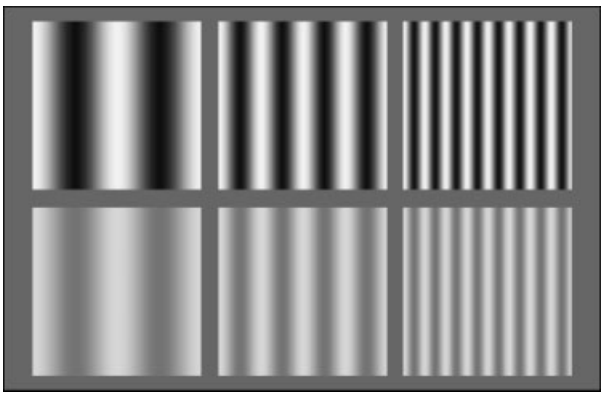


Figure 2. Illustrations of sine-wave gratings at three spatial frequencies and two contrasts. Spatial frequency (the number of stripes per unit distance) is 2, 4 and 8 cycles per image in the left, middle and right panels, respectively. Sine-wave grating contrast is commonly defined as $(I_{max} - I_{min})/(I_{max} + I_{min})$, where I_{max} and I_{min} are the maximum and minimum grating intensities, respectively. High and low contrasts are shown in the top and bottom rows, respectively. A grating with zero contrast is a uniform gray field. It is important to note that changing contrast does not alter the total amount of light in the stimulus, and therefore grating detection depends on encoding the pattern of light, rather than total intensity.

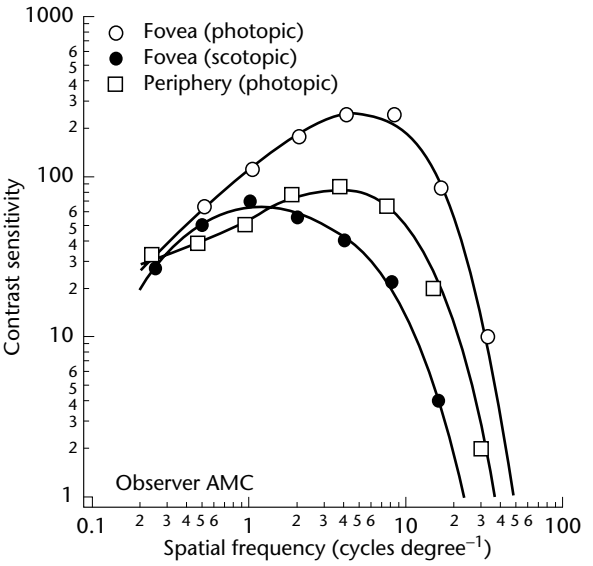


Figure 3. Contrast sensitivity functions measured in three conditions in one observer. The patterns were sine-wave gratings presented within a circular aperture (diameter = 4 degrees of visual angle), presented either directly on the line of sight, so that the retinal images were centered on the fovea, or 5 degrees below the line of sight in the peripheral visual field. Gratings were presented either on a background that was bright enough to stimulate cone photoreceptors (photopic conditions) or on a dim background that was 2 log units lower in intensity and primarily stimulated rods (scotopic conditions). Contrast sensitivity is defined as the reciprocal of contrast detection threshold. For patterns presented in the fovea and in bright light, contrast sensitivity peaks at approximately 6 cycles degree^{-1} and drops significantly at higher and lower spatial frequencies. Presenting the patterns in the peripheral visual field or against a dim background reduced contrast sensitivity significantly at high frequencies, but only slightly at low frequencies.

of the CSF can be used to predict the image structure that is visible in different viewing conditions. Many studies have shown that presenting patterns in the peripheral visual field, or viewing them in dim light, reduces contrast sensitivity significantly at high but not low spatial frequencies (Figure 3). High spatial frequencies correspond to fine details (Figure 4), so reduced sensitivity at high frequencies implies that only coarse structure will be visible in those conditions. This prediction is consistent with our visual experience, as detailed features of objects are visible only when we fixate them, and visual tasks that depend on fine spatial detail (e.g., reading) are difficult to perform in dim light. It is possible to go beyond these qualitative predictions and use the CSF to make accurate quantitative predictions about detection and discrimination thresholds for a variety of patterns (Campbell and Robson, 1968; Campbell *et al*., 1969; Kelly and Magnuski, 1975). The shape of the CSF has also been important for identifying the constraints imposed by optics and retinal mechanisms on pattern sensitivity (Banks *et al*., 1987, 1991).

FEATURE EXTRACTION

Recovering the three-dimensional structure of visual scenes from images requires more than merely detecting the presence of patterned stimulation. In addition, it is necessary to encode the size, orientation, and motion of pattern elements. There is now abundant evidence that these dimensions are encoded, at least initially, by sets of multiple, feature-selective channels. For example, recordings of visual cortical neurons in monkeys have demonstrated that each location in the visual field is encoded by multiple cells that respond selectively to different spatial frequencies, orientations, and direction and speed of motion. Evidence from many psychophysical studies in humans also suggests that pattern information is processed by feature-selective channels which have properties closely resembling those of cortical cells (DeValois and DeValois, 1988).

It is instructive to consider how one psychophysical technique, based on pattern-selective

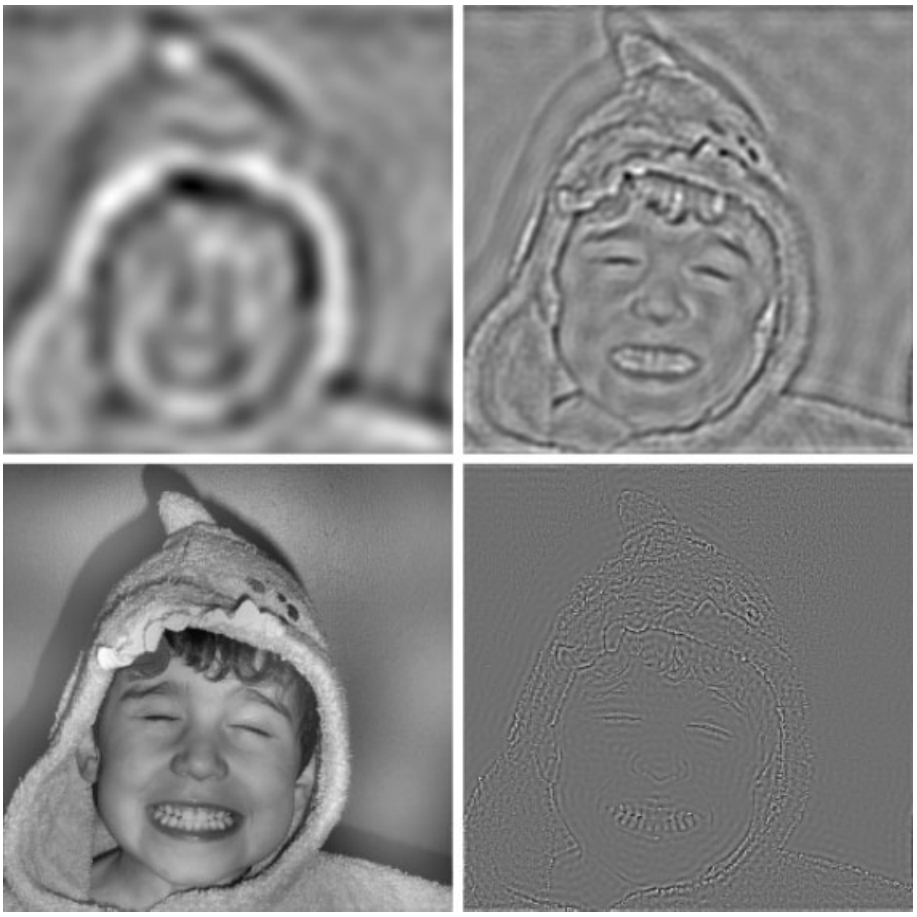


Figure 4. Illustration of the types of spatial structure in natural images that are carried by different bands of spatial frequency. The spatial frequency content for each pattern (moving clockwise from the upper left) is 3–12, 12–48, 48–192 and 3–192 cycles per image. All of the images have the same average intensity, which is referred to as the zero-frequency component. Note that low spatial frequencies carry information about large coarse features, whereas high spatial frequencies carry information about fine details.

adaptation, can be used to probe the properties of feature-selective channels. Blakemore and Campbell (1969) compared CSFs measured before and after an adaptation period during which observers viewed a high-contrast sine-wave grating. They found that adaptation reduced contrast sensitivity significantly, but only for spatial frequencies similar to that of the adaptation stimulus. Reduction in sensitivity after adaptation is a common phenomenon, and presumably occurred because prolonged viewing of a high-contrast grating fatigued the mechanisms that were responding to that stimulus. However, the fact that adaptation was frequency specific implies that the adaptation stimulus evoked responses in some mechanisms but not in others. In other words, frequency-specific adaptation implies that patterns are encoded by frequency-tuned mechanisms. Furthermore, reduced sensitivity was found only for gratings at the same orientation as the adaptation stimulus, so pattern-encoding mechanisms must also be tuned to particular orientations. Finally, Blakemore and Campbell (1969) found that adaptation in one eye reduced sensitivity in the other eye, which suggests that adaptation occurs at or beyond the primary visual cortex, as that is where signals from the two eyes first interact. All of these psychophysical findings – spatial frequency tuning, orientation tuning, and binocularity – are consistent with the known properties of neurons in the primary visual cortex.

These kinds of results suggest that properties such as size, orientation, and motion are represented by distributions of responses across populations of feature-selective mechanisms, and that different populations are responsible for coding different pieces of the retinal image. Any given visual stimulus might activate many mechanisms, but the distribution of activation across

mechanisms will be more or less unique for a particular stimulus. It might seem that there would not be sufficient numbers of cells to construct population codes for all of the information in the retina, but recent analyses suggest that there are enough cells in layer IV of the primary visual cortex to create many complete representations, at least for the part of the retinal image near the fovea. Population codes have several useful properties. They are robust (in the sense that damage to individual cells or channels does not greatly alter the population response), and they make it possible to discriminate much smaller changes in size, orientation, and motion than would be possible by using the output of a single mechanism. Computational advantages such as these might explain why population codes are found in many parts of the nervous system.

CONCLUSION

Visual mechanisms are exquisitely sensitive in dim light, but adaptation processes act to reduce sensitivity with increasing illumination. An important consequence of adaptation is that the visibility of a pattern depends on the relative amounts of light falling in different regions of the image (i.e., contrast), rather than on the total amount of light. However, the visibility of a pattern depends not just on contrast magnitude, but also on the spatial distribution of contrast. The contrast sensitivity function provides an economical account of this dependence of sensitivity on pattern shape, and has been used to predict thresholds for a variety of stimuli. The findings of psychophysical studies suggest that aspects of image contours such as size, orientation, and motion are encoded by multiple, feature-selective channels, the properties of which are strikingly similar to the properties of cells in the primary visual cortex. The result of these early visual processes is a representation of image structure near the fovea that is sufficiently rich to support higher-level visual processes.

References

Banks MS, Geisler WS and Bennett PJ (1987) The physical limits of grating visibility. *Vision Research* **27**: 1915–1924.

Banks MS, Sekuler AB and Anderson SJ (1991) Peripheral spatial vision: limits imposed by optics, photoreceptors and receptor pooling. *Journal of the Optical Society of America* **8**: 1775–1787.

Blakemore C and Campbell FW (1969) On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology* **203**: 237–260.

Campbell FW and Robson JG (1968) Application of Fourier analysis to the visibility of gratings. *Journal of Physiology* **197**: 551–566.

Campbell FW, Carpenter RHS and Levinson JZ (1969) Visibility of aperiodic patterns compared with that of sinusoidal gratings. *Journal of Physiology* **204**: 293–298.

Cornsweet TN (1970) *Visual Perception*. New York, NY: Academic Press.

DeValois RL and DeValois KK (1988) *Spatial Vision*. Oxford, UK: Oxford University Press.

Hecht S, Schlaer S and Pirenne M (1942) Energy, quanta and vision. *Journal of General Physiology* **25**: 819–840.

Kelly DH and Magnuski HS (1975) Pattern detection and the two-dimensional Fourier transform: circular targets. *Vision Research* **15**: 911–915.

Marr D (1982) *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: WH Freeman.

Further Reading

Barlow HB (1981) The Ferrier Lecture, 1980. Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London (B): Biological Sciences* **212**: 1–34.

Baylor DA, Lamb TD and Yau KW (1979) Responses of retinal rods to single photons. *Journal of Physiology* **288**: 613–634.

Braddick O, Campbell FW and Atkinson J (1978) Channels in vision: basic aspects. In: Held R, Leibowitz HW and Teuber HL (eds) *Handbook of Sensory Physiology*, vol. 8, pp. 3–38. New York, NY: Springer-Verlag.

Graham CH (ed.) (1965) *Vision and Visual Perception*. New York, NY: John Wiley.

Kelly DH (1976) Pattern detection and the two-dimensional Fourier transform: flickering checkerboards and chromatic mechanisms. *Vision Research* **16**: 277–287.

Kelly DH (1977) Visual contrast sensitivity. *Optica Acta* **24**: 107–109.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer Associates.

Sekuler R (1974) Spatial vision. *Annual Review of Psychology* **25**: 195–232.

Shapley R and Enroth-Cugell C (1984) Visual adaptation and retinal gain controls. In: Osborne N and Chader G (eds) *Progress in Retinal Research*, vol. 3, pp. 263–346. London, UK: Pergamon.

Spillman L and Werner J (eds) (1990) *Visual Perception: the Neurophysiological Foundations*. San Diego, CA: Academic Press.

Vision: Form Perception

Intermediate article

Donald D Hoffman, University of California, Irvine, California, USA
Manish Singh, Rutgers University, New Brunswick, New Jersey, USA

CONTENTS

Human vision constructs the perceived two-dimensional and three-dimensional shapes of visual objects and scenes from retinal images that are inherently ambiguous. In the process it consults a variety of sources of information, including motion, shading, texture, occlusion, binocular disparity, and image contours.

INTRODUCTION

The human eye focuses an image onto a light-sensitive sheet of neural tissue called the retina (Dowling, 1987). This image is captured by a discrete array of cells in the retina, called photoreceptors. Each photoreceptor generates a signal which varies in time with the discrete number of photons of light that the photoreceptor catches. This discrete array of time-varying signals is the starting point of vision. The only information this array makes explicit is the varying number of photon catches at each individual photoreceptor. It does not make explicit lines, curves, two-dimensional regions, three-dimensional shapes or any other aspect of the visual forms of objects and their environments. The perception of visual forms is the consequence of sophisticated processes of construction which engage literally billions of neurons and trillions of synaptic connections between neurons. Every line, curve, 2D region or 3D shape that we see is a construction of our visual system, created rapidly starting with just the photon catches at the retina. Vision researchers have made substantial progress in describing the constructive processes underlying the perception of visual form. (*See* **Pattern Vision, Neural Basis of**)

CONTOUR DETECTION

The construction of contours occurs early in the flow of visual processing. In many mammals the primary visual cortex, which is the first stage of cortical visual processing, begins the construction of contours with neurons called 'simple cells', which construct oriented line segments (Hubel and Wiesel, 1959, 1962; Hubel, 1995). These cells construct lines such as the short black lines shown in Figure 1. Computational theories of vision also place the construction of contours early in the flow of visual processing. In the influential theory of David Marr, the flow of visual processing leads to a sequence of visual representations: the primal sketch, the 2½D sketch, and the 3D model (Marr, 1982). The primal sketch makes explicit structure and grouping in the 2D image, the 2½D sketch represents the surface geometry of objects relative to the viewer, and the 3D model represents the volume of objects in coordinates centered in the objects.

The primal sketch contains contours as a key element of the representation. In Marr's theory these are constructed by first linearly filtering the retinal image with a spatial filter whose shape is defined mathematically as the Laplacian of a 2D Gaussian (Marr and Hildreth, 1980). This filter is circularly symmetric, and its shape resembles a Mexican hat. Edges correspond to those places where the values of the filtered output pass through zero. Marr and Hildreth proposed that simple cells in the primary visual cortex are in fact detectors of these zero crossings. Their computational theory of edge detection has been superseded by later approaches that filter the image with spatial filters that are not circularly symmetric, but are designed to optimize the signal-to-noise ratio in the edge construction process (Canny, 1986; Deriche, 1987).

Linearly filtering an image is a key first step in the construction of contours. The steps that come after filtering are complex and highly nonlinear. This is also illustrated in Figure 1: in square 1, one sees not only the short black lines, but also a square with clear edges. The square appears to be slightly

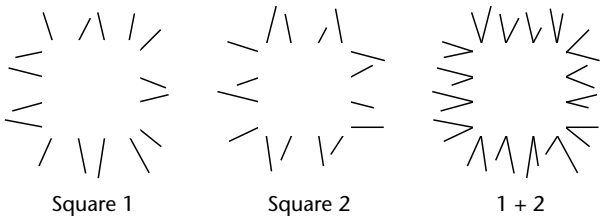


Figure 1. Constructing visual contours is not a simple linear process. In square 1 and square 2 we see a square with clear edges. When these two figures are linearly superimposed, as in the right-hand image, we no longer see a square with clear edges: this is the opposite of what one would expect from a linear process.

brighter than the background. In fact, measurement of actual light intensities would indicate that the white region inside the square is identical to the white region outside the square, and that no edge has been drawn around the square. The brightness of the square, and the clear edges that mark its border, are entirely constructs of the visual system. The square region is called a subjective (or illusory) surface and its edges are called subjective (illusory) contours (Petry and Meyer, 1987). Square 2 also appears to have a subjective square region surrounded by clear subjective contours. If we superimpose square 1 on square 2 so that their subjective squares are perfectly aligned, we obtain the image labeled '1 + 2'. Here the subjective square and subjective contours are weak or missing altogether; this is the opposite of what one would predict if the construction of contours were an entirely linear process (Albert and Hoffman, 2000).

Neurons that signal subjective contours have been found in the primary visual cortex, area V1, of the macaque monkey (Grosof *et al*., 1993) and also in area V2, the next stage of cortical visual processing (von der Heydt *et al*., 1984). Several artificial neural network models have been proposed to explain the properties of these cells (von der Heydt and Peterhans, 1989; Francis and Grossberg, 1996).

THREE-DIMENSIONAL SHAPE FROM IMAGE CONTOURS

The visual system not only constructs 2D contours, it also constructs the 3D shapes of objects and represents these 2D contours and 3D shapes in terms of parts and their spatial relationships. In the process of constructing 3D shape, human vision consults a variety of sources of information, including shading, texture, motion, occlusion, binocular disparity, and 2D contours. (*See* **Object Perception, Neural Basis of**)

Figure 2a illustrates the construction of 3D shapes from 2D contours. The figure uses just a few contours to depict a doughnut shape. The contours are called 'occluding contours' because they depict points where the visible portions of the 3D shape just begin to occlude the hidden portions. Yet these few occluding contours are sufficient to trigger the visual system to construct the smooth 3D shape of the doughnut. Human vision uses several geometric facts and assumptions in the process. The qualitative shape of a 3D object can be described at each point as being convex, concave, cylindrical, or saddle-shaped. Convex regions of a doughnut are denoted by plus signs in Figure 2a, and saddle-shaped regions by minus signs. As can be seen from Figure 2a, the outer occluding contours are in convex regions, whereas the inner occluding contours are in saddle-shaped regions. Jan Koenderink showed that human vision can, in principle, infer the qualitative shapes of smooth 3D surfaces from their projected occluding contours (Koenderink, 1984). The visual system also assumes in this case that the surface varies smoothly between the occluding contours. Since the qualitative shape changes from convex to saddle between these contours, it must be cylindrical at the boundary between convex and saddle. The cylindrical points are indicated by the dashed circle on top of the doughnut. This gives a complete qualitative description of the 3D shape of the doughnut, which Koenderink has shown can be derived entirely from its occluding contours and some built-in knowledge of geometry and projection.

Human vision organizes 2D and 3D shapes into parts as an aid to recognition, manipulation, and naming. It often uses concave regions, and especially points of highest magnitude of curvature

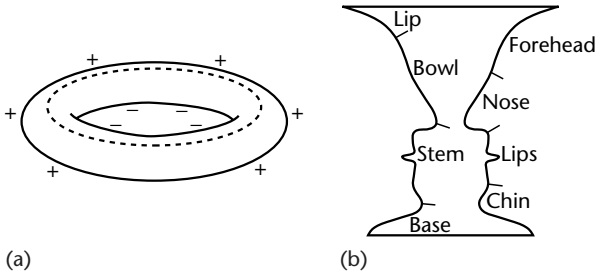


Figure 2. (a) Human vision can construct 3D shapes from simple 2D drawings of contours, as illustrated by the doughnut in this figure. (b) Human vision has rules for carving 2D curves and 3D shapes into parts. In the face–goblet ambiguous figure these rules divide the faces into parts corresponding to the forehead, nose, lips and chin; they divide the goblet into a lip, bowl, stem and base.

within these concave regions, to divide shapes into parts: this is called the 'minima rule' (Hoffman and Richards, 1984; Hoffman and Singh, 1997; Singh *et al*., 1999). This is illustrated in Figure 2b, which shows the well-known face–goblet ambiguous figure. If the faces are taken as the object, then the extrema of curvature within the concave regions of the faces are used by human vision as the boundaries between parts. These extrema are indicated by the short dashes on the right-hand side of the figure, and divide the faces into parts corresponding to the forehead, nose, lips and chin. If the goblet is taken as the object, then the regions of the curves that are concave and convex are reversed from when the faces are taken as the object. Since only concave regions are used to create part boundaries, human vision creates a new set of part boundaries for the goblet that are different from the part boundaries of the faces. These are shown on the left-hand side of Figure 2b, and divide the goblet into a lip, bowl, stem and base. Neurophysiological studies of visual area V4 have found many cells that signal the extrema of curvature of 2D contours (Pasupathy and Connor, 1999).

SHADING

Consider Figure 3a. It contains a pattern of gray-level intensity values distributed on a flat sheet of paper. However, it is quite difficult to see this figure as simply a two-dimensional pattern. One cannot help but see a light-colored wrinkled surface extended in three dimensions, lit from the left. The ability of the visual system to construct a 3D surface shape from 2D shading information is a striking computational feat. When viewing a surface, the light intensity at any given location on the retina is a combined function of (at least) three different variables: the reflectance properties of the surface, the local orientation of the surface, and the position and intensity of the light sources. In order to construct the percept of a 3D surface, the visual system must somehow decompose the pattern of luminance values into the separate contributions of surface reflectance, surface shape, and illumination.

In principle a unique solution to the shape-from-shading problem can be derived under restrictive conditions: the reflectance function of the surface is known, the surface is smooth, and the illuminant is a single point source (Horn, 1977). These assumptions allow variations in image luminance to be attributed entirely to gradual changes in surface orientation. Humans, however, are able to perceive shape from shading in many situations where the reflectance is unknown, the surface contains

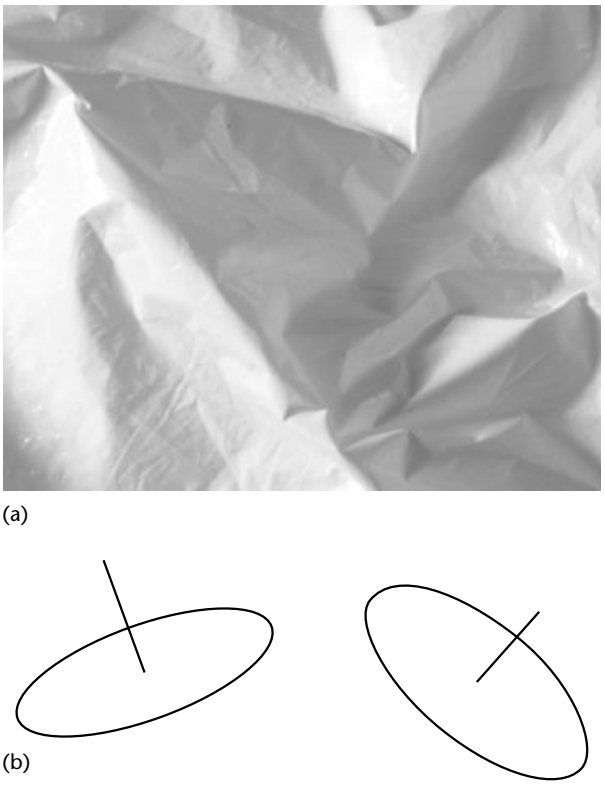


Figure 3. (a) Human vision can construct three-dimensional surface shape from two-dimensional images depicting shading information. To do so, it must separate the contributions of surface shape from those of surface reflectance and illumination (photograph by Marc Talusan). (b) The gauge figure is used to probe the surface structure perceived by observers (Koenderink *et al*., 1992). Observers adjust the 3D orientation of the gauge figure until it 'fits' a depicted surface locally.

tangent discontinuities, and is lit by more than one light source. This suggests that human vision brings to bear other constraints in computing 3D shape from shading information – although it is fair to say that the precise constraints used by human vision remain largely unknown.

Koenderink and colleagues have probed psychophysically the human perception of surface relief from gray-level images (these typically contain both shading and occluding contours). One technique uses a gauge figure – the projection of a circular disk oriented in 3D, with a short line segment sticking perpendicularly out of its center (Figure 3b). In a typical experiment, a gauge figure is superimposed at a large number of locations on a shaded image, and the observer adjusts its perceived 3D orientation using a trackball, in order to match the perceived local orientation of the surface (Koenderink *et al*., 1992). These local measurements can then be integrated to construct the global

surface structure perceived by the observer (this presupposes internal consistency in the observer's responses, which is typically the case). Although this perceived surface structure is qualitatively similar to the depicted surface, it is not quantitatively the same. In particular, surface reliefs seen by different observers differ by depth scalings (i.e. flattenings and elongations in depth), and similarly, surface reliefs seen by the same observer under different illumination conditions differ by depth scalings. Moreover, different parts of the depicted surface are often seen depth-scaled by different amounts (Todd *et al*., 1996), providing further evidence that human vision represents surface shapes in terms of parts.

Qualitatively similar surface reliefs are obtained using a different technique in which pairs of points are presented at different locations on a shaded image, and observers are asked to indicate which of the two points appears closer in depth (Koenderink *et al*., 1996). This method, however, yields results that are less precise than those obtained by the gauge figure method by an order of magnitude. This, in itself, is a significant finding: it suggests that a depth map – a pointwise specification of relative depth at each image location (e.g. Marr, 1982) – is not the primary way human vision represents surface structure, from which it then derives local surface orientation. Rather, surface orientation itself appears to be a perceptually fundamental variable. Moreover, the fact that quantitatively different surface reliefs are obtained with different experimental methods suggests that the visual system may not have a single representation of surface shape, but rather may invoke distinct representations depending on the task at hand (Koenderink, 1998).

CONCLUSION

Human vision starts with just the shower of photons that hit the retina of each eye, and proceeds to construct two-dimensional contours and three-dimensional forms by consulting various sources of information, including shading, texture, motion, occlusion, binocular disparity and image contours. In the process it uses many rules of construction, some of which are based on laws of geometry, reflectance, lighting and projection.

References

Albert MK and Hoffman DD (2000) The generic-viewpoint assumption and illusory contours. *Perception* **29**: 303–312.

Canny JF (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**: 769–798.

Deriche R (1987) Using Canny's criteria to derive an optimum edge detector recursively implemented. *International Journal of Computer Vision* **2**: 167–187.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Harvard University Press.

Francis G and Grossberg S (1996) Cortical dynamics of form and motion integration: persistence, apparent motion, and illusory contours. *Vision Research* **36**: 149–173.

Grosof DH, Shapley RM and Hawken MJ (1993) Macaque-V1 neurons can signal illusory contours. *Nature* **365**(6446): 550–552.

Hoffman DD and Richards WA (1984) Parts of recognition. *Cognition* **18**: 65–96.

Hoffman DD and Singh M (1997) Salience of visual parts. *Cognition* **63**: 29–78.

Horn BKP (1977) Image intensity understanding. *Artificial Intelligence* **8**: 201–231.

Hubel DH (1995) *Eye, Brain, and Vision*. New York, NY: Freeman Press.

Hubel DH and Wiesel TN (1959) Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology* **148**: 574–591.

Hubel DH and Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–154.

Koenderink JJ (1984) What does the occluding contour tell us about solid shape? *Perception* **13**: 321–330.

Koenderink JJ (1998) Pictorial relief. *Philosophical Transactions of the Royal Society of London Series A* **356**: 1071–1086.

Koenderink JJ, van Doorn AJ and Kappers AML (1992) Surface perception in pictures. *Perception and Psychophysics* **52**: 487–496.

Koenderink JJ, van Doorn AJ and Kappers AML (1996) Pictorial surface attitude and local depth comparisons. *Perception and Psychophysics* **58**: 163–173.

Marr D (1982) *Vision: A Computational Investigation into The Human Representation and Processing of Visual Information*. San Francisco, CA: WH Freeman.

Marr D and Hildreth EC (1980) Theory of edge detection. *Proceedings of the Royal Society of London Series B* **207**: 187–217.

Pasupathy A and Connor CE (1999) Responses to contour features in macaque area V4. *Journal of Neurophysiology* **82**: 2490–2502.

Petry S and Meyer GE (1987) *The Perception of Illusory Contours*. New York, NY: Springer.

Singh M, Seyranian GD and Hoffman DD (1999) Parsing silhouettes: the short-cut rule. *Perception and Psychophysics* **61**: 636–660.

Todd JT, Koenderink JJ, van Doorn AJ and Kappers AML (1996) Effects of changing viewing conditions on the perceived structure of smoothly curved surfaces. *Journal of Experimental Psychology: Human Perception and Performance* **22**: 695–706.

Von der Heydt R and Peterhans E (1989) Cortical contour mechanisms and geometrical illusions. In: Lam DM and Gilbert CD (eds) *Neural Mechanisms of Visual Perceptions*, pp. 157–170. Woodlands, TX: Portfolio Publishing.

Von der Heydt R, Peterhans E and Baumgartner G (1984) Illusory contours and cortical neuron responses. *Science* **224**: 1260–1262.

Further Reading

Farah MJ (2000) *The Cognitive Neuroscience of Vision*. Malden, MA: Blackwell Publishers.

Gregory RL (1997) *Eye and Brain*, 5th edn. Princeton, NJ: Princeton University Press.

Hoffman DD (1998) *Visual Intelligence: How We Create What We See*. New York, NY: WW Norton.

Horn BKP and Brooks MJ (1989) *Shape from Shading*. Cambridge, MA: MIT Press.

Kellman PJ and Arterberry ME (1998) *The Cradle of Knowledge: Development of Perception in Infancy*. Cambridge, MA: MIT Press.

Koenderink JJ (1990) *Solid Shape*. Cambridge, MA: MIT Press.

Marr D (1982) *Vision: A Computational Investigation into The Human Representation and Processing of Visual Information*. San Francisco, CA: WH Freeman.

Palmer SE (1999) *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Regan D (2000) *Human Perception of Objects: Early Visual Processing of Spatial Form Defined by Luminance, Color, Texture, Motion, and Binocular Disparity*. Sunderland, MA: Sinauer.

Wandell BA (1995) *Foundations of Vision*. Sunderland, MA: Sinauer.

Zeki S (2000) *Inner Vision: An Exploration of Art and the Brain*. Oxford, UK: Oxford University Press.

Vision: Object Recognition

Intermediate article

Michael Tarr, Brown University, Providence, Rhode Island, USA

CONTENTS

Object recognition is the process whereby observers are able to recognize three-dimensional objects despite receiving only two-dimensional input that varies greatly depending on viewing conditions. Theories explaining this process differ in the nature of the underlying features, varying from two-dimensional image properties to three-dimensional parts.

INTRODUCTION

One of the most impressive feats of perception is the interpretation of the ever-changing pattern of light that falls on our retinas. From this undifferentiated pattern of illumination, we extract meaning in the form of objects. Although there is basic agreement about the steps necessary for the brain to solve the problem of visual object recognition, the specifics of this process have led to one of the more spirited debates in the study of vision (Biederman and Gerhardstein, 1995; Tarr and Bülthoff, 1995). Unquestionably, recognition requires the matching of representations of visual input to like representations in memory. Moreover, there is a general sense that the form of the representation includes information about surfaces and contours extracted from the image. However, there is disagreement regarding how such information is organized into high-level object representations. The image-based approach posits that collections of viewpoint-dependent surfaces and contours constitute the basic features of object representation and recognition. In contrast, the structural-description approach posits that the basic features are viewpoint-invariant, three-dimensional (3D) volumes that are derived from surfaces and contours.

THE PROCESS OF OBJECT RECOGNITION

Early and mid-level vision takes the unstructured array of visual stimulation that falls on the retina and organizes it into coherent perceptual entities that form the bases of high-level vision. The end results of these processes – object representations of visual input – are then compared with like representations encoded in long-term visual memory. This process of object recognition begins with perceptual organization in which meaningful units such as local contours and surfaces are inferred from variations in the intensity array (Marr, 1982). These local features are then subjected to further, more global, perceptual organization which results in the mental representations that we rely on for visual recognition. Because these representations of the input and representations of familiar objects in visual memory are assumed to be in the same format, a search or indexing process can be conducted to establish the most probable matches between input and memory. The most likely candidate representations are then directly compared with the representation of the input, with the best match providing information regarding the identity (or familiarity) of the object in question.

Although this recognition process may seem straightforward, two factors make finding the 'best match' a difficult problem. First, objects can vary in their appearance from one moment to the next. An object can rotate, shift position, or change configuration; an observer can move; and the lighting of the scene can change. All of these transformations can produce dramatic changes in the image of an object presented to our visual system. However, the goal of object recognition is object constancy – that is, we want to be able to recognize a given object regardless of how its appearance changes over different viewings. Second, objects can vary at the level they are recognized. An object can be visually identified at the entry ('dolphin'), subordinate ('bottlenose dolphin') or individual ('Flipper') levels (Jolicoeur *et al*., 1984; Tarr and Gauthier, 2000). Thus, the question of what constitutes an appropriate match and what information within the representation is relevant to that match depends on the task at hand as well as the

information currently available in the environment. For example, information across the representations of many specific objects might play a role in entry-level recognition, while information within a single object representation might be relevant to individual-level recognition. The issues of object constancy and recognition at varying levels of specificity are two of the most important problems addressed by different models of object recognition.

IMAGE-BASED AND STRUCTURAL DESCRIPTION MODELS

One of the largest sources of variation for an image of an object is viewpoint change. Given that rotations in depth alter an object's visible geometry and surfaces, how does one recognize that a new view of an object depicts a familiar object seen from a different vantage point? This question is perhaps the most-studied in the object recognition literature, in part because competing models offer different predictions regarding how the human visual system compensates for variations in viewpoint.

Image-based models postulate that an object's features are encoded as they appeared in the original image. Thus, image-based object representations, sometimes referred to as 'views', are viewpoint-dependent to the extent that they capture the appearance of objects from specific viewpoints. One important implication of viewpoint dependency is that a single view is unlikely to adequately represent the appearance of a 3D object; therefore, multiple views must be learned for each object in order to support recognition from many different vantage points (Tarr and Pinker, 1989). A second implication is that, owing to changes in the viewpoint of the observer or in the orientation of the object, the currently observed image may still not correspond directly to any known view; therefore, mental transformation or generalization mechanisms must be included to support the alignment of familiar and unfamiliar views (Jolicoeur, 1985; Tarr and Pinker, 1989; Ullman, 1989; Poggio and Edelman, 1990; Perrett *et al*., 1998). Although not a strict constraint, nearly all generalization mechanisms are assumed to be viewpoint-sensitive in that as the magnitude of normalization increases (i.e. larger viewpoint differences) the costs associated with recognition increase as well (i.e. longer response times and higher error rates; Shepard and Metzler, 1971). Taken together, these two implications delineate a viewpoint-dependent model of object recognition in which observers represent oft-seen (Tarr and Pinker, 1989) or geometrically regular views of objects (Palmer *et al*., 1981) that are matched to new, unfamiliar views using viewpoint-sensitive generalization procedures.

In contrast to image-based models, most structural description models postulate that object features are encoded in a viewpoint-invariant manner and that the same representation of an object is derived regardless of the viewpoint of the observer or the pose of the object. Thus, a single structural description should match a known object from a wide range of viewpoints (Marr and Nishihara, 1978; Biederman, 1985) and there is no need to posit either multiple views or generalization mechanisms. Behaviorally such models predict that the costs associated with recognizing an object in a given viewpoint should be independent of whether that viewpoint is familiar or unfamiliar (Biederman, 1985). Note, however, that some structural description models do not assume complete viewpoint invariance. Consider what happens if we view the front and the back of an object: the visible features and parts for the two views will be different and, in some structural description models, will give rise to different object representations. Consequently, although structural description models typically posit viewpoint invariance, they do so only over a limited range of viewpoints for each distinct structural description of an object. Such models must also assume that new structural descriptions are encoded for views of the same object that show different features or parts (Biederman, 1985; Biederman and Gerhardstein, 1993). In some sense these structural description models are also viewpoint-dependent, but only at a coarse level relative to image-based models.

BEHAVIORAL EVIDENCE

How do we evaluate these competing pictures of object recognition? At the simplest level one could name familiar objects at their most common, upright orientation or at less familiar orientations (e.g. upside down) and observe whether recognition performance is viewpoint-dependent or viewpoint-invariant. Such experiments often reveal a pattern of viewpoint dependency, with longer naming times and more naming errors as objects are rotated away from their most familiar orientations either in the picture plane (Jolicoeur, 1985) or in depth (Palmer *et al*., 1981), although there are other studies in which orientation had little effect on naming performance. One problem with such results is that, regardless of familiarity, some views of objects are more readily recognized than other views: that is, views that provide stable

information about an object's shape are processed more efficiently, leading to faster responses and lower errors compared with less stable or less informative views of the same object. Thus, if the most familiar viewpoints are also views that are easy to recognize, then viewpoint dependency might result from the difficulty of perceiving the object's shape, not from viewpoint-dependent recognition processes. A second concern is that most of us have already seen many different objects in many different viewpoints. Therefore, a participant in an experiment using familiar objects might have previously acquired either multiple views or a structural description for each object – both types of representations would support viewpoint-invariant recognition (Tarr and Pinker, 1989). Accordingly, it is difficult to draw any strong conclusions about the nature of object representations with respect to viewpoint from any behavioral results obtained with familiar objects.

To address these concerns, several studies were designed in which participants had to learn novel 2D or 3D objects in a small number of viewpoints (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Biederman and Gerhardstein, 1993; Tarr, 1995; Hayward and Tarr, 1997; Tarr *et al*., 1998). For example, in a study by Tarr and Pinker (1989; see also Tarr, 1995) participants named visually similar shapes until recognition performance was equivalent at several orientations. As with familiar objects, this viewpoint invariance might have arisen from learning either multiple views or a viewpoint-invariant description. These two alternatives were tested by presenting the shapes in never-before-seen orientations. For these new orientations recognition performance was monotonically dependent on the distance from a familiar orientation. This pattern indicates that participants encoded each shape at each familiar orientation (an image-based 'view') and then generalized from unfamiliar orientations to the nearest known view. Similarly, there is evidence that neural representations of objects are viewpoint-dependent, with individual neurons being 'view-tuned' (Logothetis *et al*., 1995). That is, neurons in the inferotemporal cortex (IT) that are selective for a particular object are also maximally responsive to specific viewpoints of that object, with different neurons coding for different views. Thus, the most diagnostic task we can use, the recognition of novel objects, is often viewpoint-dependent. At the same time, there is some evidence for viewpoint-invariant mechanisms in recognition. However, the conditions under which such evidence has been obtained are quite constrained. Indeed, the best-known model of viewpoint-invariant recognition (Biederman, 1985; Biederman and Gerhardstein, 1993) places specific limits on when viewpoint invariance is predicted.

FEATURES OF REPRESENTATION

The nature of the features encoded in the representation is a second dimension along which different models may be compared, and is critical to how objects can be recognized at multiple levels. Image-based representations are generally considered to be composed of large numbers of local measures of the image: orientation, shape, surface slant, bounding contour, color, luminance, shading gradient, and potentially many more (Edelman, 1993; Mel, 1997; Riesenhuber and Poggio, 1999; Ullman and Sali, 2000). Because the fundamental units of object representation are a natural consequence of earlier perceptual processing (e.g. Livingston and Hubel, 1987), image-based representations capture not only object shape, but also surface properties such as texture, color, shading, and local depth. This rich feature set supports recognition at multiple levels of access and the same approach can account for entry, subordinate, and individual-level visual recognition.

In contrast, structural descriptions are typically thought to be composed of 3D volumes, such as parameterized generalized cylinders (Marr and Nishihara, 1978), or restricted sets of qualitatively defined 3D parts. The best-known version of the latter approach assumes that configurations of local shape features are combined to specify particular 3D parts known as 'geons' (Biederman, 1985). The geon model makes two critical assumptions that may be empirically evaluated. First, geon structural descriptions must be recovered by processing steps beyond those normally associated with mid-level perception. That is, the shape features that specify geons must be identified, combined, and remapped into the appropriate 3D primitives. One consequence of this recovery process is a many-to-one mapping of feature configurations into geons whereby many visually similar objects come to be represented by the same geon structural description. Therefore, a theory based on geons is necessarily restricted to explaining only entry-level recognition – information relevant to subordinate-level or individual-level recognition is lost in the process of redescribing an image. Consequently, mechanisms different from those that rely on structural descriptions must be posited to account for both subordinate-level and individual-level recognition abilities. Second, geon structural

descriptions rely on 3D parts, not separable image features: that is, so long as the visible part configuration of an object does not change, the same description should be derived. However, Hayward and Tarr (1997) found that, despite the same geons being visible in every image, viewpoint sensitivity was determined by changes in the visible qualitative image features (e.g. curved versus straight).

CONCLUSION

Image-based and structural description models have emerged as potential explanations for our ability to recognize objects. Empirical evaluation of these competing approaches suggests that current versions of both models leave something to be desired. For example, the geon structural description model is limited in the kinds of recognition tasks that it can explain. Similarly, image-based models should incorporate sensitivity to qualitative features above and beyond metric variations. Thus, more sophisticated models must be developed before we have a complete understanding of the complex process of visual object recognition.

References

Biederman I (1985) Human image understanding: recent research and a theory. *Computer Vision, Graphics, and Image Processing* **32**: 29–73.

Biederman I and Gerhardstein PC (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance* **19**: 1162–1182.

Biederman I and Gerhardstein PC (1995) Viewpoint-dependent mechanisms in visual object recognition: reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance* **21**: 1506–1514.

Bülthoff HH and Edelman S (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science of the USA* **89**: 60–64.

Edelman S (1993) Representing three-dimensional objects by sets of activities of receptive fields. *Biological Cybernetics* **70**: 37–45.

Hayward WG and Tarr MJ (1997) Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance* **23**: 1511–1521.

Jolicoeur P (1985) The time to name disoriented natural objects. *Memory and Cognition* **13**: 289–303.

Jolicoeur P, Gluck M and Kosslyn SM (1984) Pictures and names: making the connection. *Cognitive Psychology* **16**: 243–275.

Livingstone MS and Hubel DH (1987) Psychophysical evidence for separate channels for the perception of form, color, movement and depth. *Journal of Neuroscience* **7**: 3416–3468.

Logothetis NK, Pauls J and Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. *Current Biology* **5**: 552–563.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: WH Freeman.

Marr D and Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B Series* **200**: 269–294.

Mel B (1997) SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation* **9**: 977–804.

Palmer S, Rosch E and Chase P (1981) Canonical perspective and the perception of objects. In: Long J and Baddeley A (eds) *Attention and Performance*, vol. IX, pp. 135–151. Hillsdale, NJ: Lawrence Erlbaum.

Perrett DI, Oram MW and Ashbridge E (1998) Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition* **67**(12): 111–145.

Poggio T and Edelman S (1990) A network that learns to recognize three-dimensional objects. *Nature* **343**: 263–266.

Riesenhuber M and Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature Neuroscience* **2**: 1019–1025.

Shepard RN and Metzler J (1971) Mental rotation of three-dimensional objects. *Science* **171**: 701–703.

Tarr MJ (1995) Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review* **2**(1): 55–82.

Tarr MJ and Bülthoff HH (1995) Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance* **21**: 1494–1505.

Tarr MJ and Gauthier I (2000) FFA: A Flexible Fusiform Area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* **3**: 764–769.

Tarr MJ and Pinker S (1989) Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology* **21**(28): 233–282.

Tarr MJ, Williams P, Hayward WG and Gauthier I (1998) Three-dimensional object recognition is viewpoint-dependent. *Nature Neuroscience* **1**: 275–277.

Ullman S (1989) Aligning pictorial descriptions: an approach to object recognition. *Cognition* **32**: 193–254.

Ullman S and Sali E (2000) Object classification using a fragment-based representation. In: Lee SW, Bülthoff HH and Poggio T (eds) *Biologically Motivated Computer*

Vision, vol. 1811, pp. 73–87. Berlin, Germany: Springer-Verlag.

Further Reading

Edelman S (1999) *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.

Logothetis NK and Sheinberg DL (1996) Visual object recognition. *Annual Review of Neuroscience* **19**: 577–621.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford, UK: Oxford University Press.

Pinker S (1984) Visual cognition: an introduction. *Cognition* **18**: 1–63.

Tarr MJ and Bülthoff HH (1998) *Object Recognition in Man, Monkey, and Machine*. Cambridge, MA: MIT Press.

Ullman S (1996) *High-Level Vision*. Cambridge, MA: MIT Press.

Vision: Occlusion, Illusory Contours and 'Filling-in'

Intermediate article

Philip J Kellman, University of California: Los Angeles, Los Angeles, USA

CONTENTS

Visual interpolation processes allow humans to perceive complete objects despite input that is fragmentary in space and time owing to occlusion. Mental representations of objects and surfaces, and their correspondence to real objects, depend heavily on interpolation processes that create both contour and surface connections among visible areas.

INTRODUCTION

Perceptual systems provide information about the world that is used to guide thought and action. For detailed knowledge of objects and the spatial layout, obtained at a distance, vision is preeminent. Vision is studied at many levels. A rich scientific tradition, for example, has sought to characterize the eye as an optical instrument, investigating its sensitivity and responses to light. To understand vision, however, we need to address another set of questions involving how visual information leads to descriptions of objects, spatial arrangements and events.

The ease of seeing conceals many complexities. A fundamental one is the problem of fragmentation. Whereas both the world and our representations of it contain coherent objects and continuous surfaces, the input from the world to our eyes is fragmentary. Most objects are partly occluded by other objects: only parts of their surfaces reflect light to the eyes of an observer. A building viewed through foliage consists at the retina of many separate patches of building separated by the projections of branches and leaves. How does the visual system assign hundreds of visible patches to a representation of a single, continuous, and complete building, separate from the foliage? These problems grow even more complicated when observers or objects move, as patterns of occlusion continuously change. Fortunately, human vision employs processes that overcome the problems of fragmentation and occlusion. These segmentation and grouping processes produce perception of objects and surfaces from the fragmentary input.

PHENOMENA OF VISUAL INTERPOLATION

Visual interpolation is the connecting of contours and surfaces across gaps. Figure 1 illustrates this phenomenon. Figure 1a shows an example of partial occlusion: six noncontiguous gray regions appear, yet the observer's visual system connects them into a single object extending behind the black occluder. The object's overall shape is apparent. Perceptual organization of this scene also leads to the perception of circular apertures in the black surface, through which the gray object and a more distant white surface are seen. Figure 1b illustrates the related phenomenon of illusory contours or illusory objects. Here, the visual system connects contours across gaps such that these interpolated contours appear in front of other surfaces in the array. Illusory contours were discovered by Schumann (1904), but they have become an important topic of vision research since the seminal work of Kanizsa (1979). In Figure 1c, an additional effect, one of transparency, can be observed in the figure that is created across the gaps. Finally, in Figure 1d, a homogeneous black region splits into two visible figures, a phenomenon sometimes referred to as "self-splitting objects".

These phenomena are formally similar in that equivalent or similar visible contours and surface fragments become linked to form objects, despite intervening gaps. Those in parts a, b and c of Figure 1 are actually even more similar: the completed object in each case is defined by the same collection of physically specified and interpolated contours.

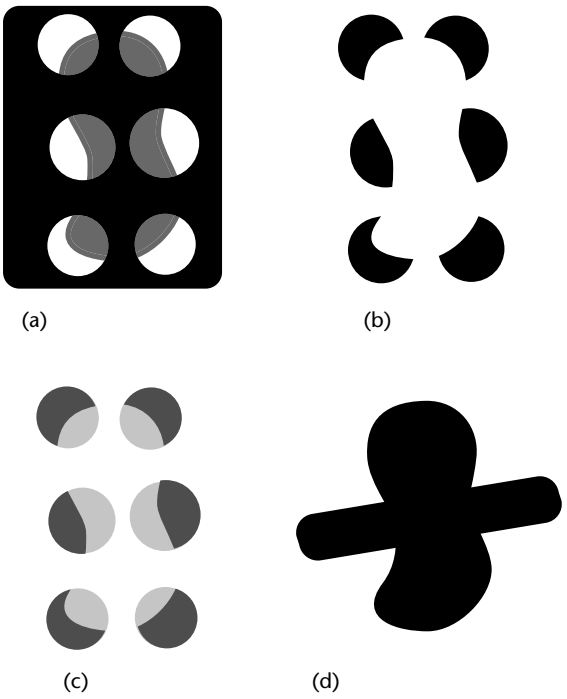


Figure 1. [*Figure is also reproduced in color section.*] Visual interpolation phenomena. In each case edges and surfaces are connected perceptually across gaps in the input: (a) partly occluded object; (b) illusory object; (c) transparent object; (d) self-splitting object.

PROCESSES OF VISUAL INTERPOLATION

How do visual filling-in processes occur? The answer begins with the available information in light reflected from objects. Broadly speaking, the visual system uses relationships among visible contour segments and relationships of surface properties in visible areas.

Contour Processes

Visible contours provide the primary information used to segment scenes into objects. Contours are one-dimensional entities marking abrupt changes in some visible property. Researchers have most often considered changes in lightness or color; however, contours based on differences in texture, depth or velocity also provide important information. The importance of contours derives from the fact that objects will tend to be relatively homogeneous in their composition (and thus in their lightness, color, texture, depth, etc.) and will often differ from nearby objects and surfaces. Visible contours often mark the locations of object boundaries. Computations using abrupt changes in visible properties across a scene to locate contours are an important starting point of object perception, both in biological and artificial vision systems.

Because of occlusion, however, visible contours seldom delineate a whole object. Most objects are opaque and three-dimensional; in projecting to a two-dimensional retina, they occlude parts of themselves. Moreover, in normal scenes, most objects are partly occluded by others. Contour interpolation processes connect visible contours across gaps caused by occlusion to produce perceptual units that correspond more accurately to the actual objects in a scene. These computations begin with contour intersections or junctions. Whereas points along contours have unique orientations (tangents), junctions are points where there are ordinarily two orientations – those of two intersecting contours. As can be seen in Figure 1, interpolated contours in all cases begin and end at these tangent discontinuities in the image. These points include all of the places where edges go out of sight and need to be interpolated. It has been proposed that 'end-stopped' cells in visual cortex (cells sensitive to the contours that end or abruptly change direction within their receptive fields) underlie junction detection (e.g. Heitger *et al*., 1993).

Interpolated contours agree with the slopes of visible contours at their connection points (Figure 1). Contour interpolation can be summarized by saying that visual system interpolates according to a smoothness principle, called contour relatability, such that interpolated contours are smooth (differentiable at least once) and monotonic (singly inflected), and at their end points they match the slopes of real contours (Kellman and Shipley, 1991). The notion of relatability in perceptual grouping is not the same as (but is related to) the principle of good continuation, proposed as one of several principles of perceptual organization in the early twentieth century by the Gestalt psychologists (Wertheimer, 1923).

The contour interpolation phenomena of Figure 1 have different appearances. Occluded objects, illusory objects and other interpolation cases involve different assignments of depth and boundary ownership relative to adjacent surfaces. Despite these differences in the final appearance, research indicates that an identical interpolation process – connecting visible edges across gaps – operates in all of these cases (Kellman *et al*., 1998; Ringbach and Shapley, 1996). At an early processing stage, oriented edge fragments signaled by cortical cells sensitive to small regions are integrated into visible contour fragments. Interpolation processes produce connections among these, based on the

geometry of relatability. The final perceived arrangement depends on constraints imposed, both early and late in processing, by boundary ownership, depth, and scene consistency.

Figure 2 illustrates two phenomena that have provided insight into a common interpolation process. Figure 2a consists of two images which if shown to the two eyes (by free-fusing or using a stereoscope) produce a single display with parts at different depths. The central circle in the figure has four interpolated contours, each of which is an illusory contour along part of its length and an occluded contour along another part. That these can join to form a single perceived contour (or rather, that interpolation occurs between these two different kinds of end points) suggests an interpolation process that accepts either type of input. Figure 2b (after Petter, 1956) shows a homogeneously black display that, because of interpolation processes, splits into two perceived objects. Where these objects cross in the image, one appears in front, having illusory contours, and the other appears to pass behind, having occluded contours.

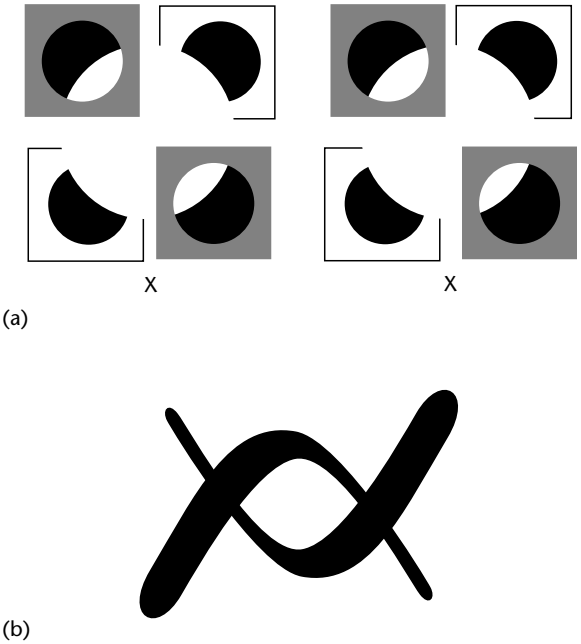


Figure 2. Examples of a common contour interpolation process in illusory and occluded object perception. (a) The central circle appears by virtue of interpolated contours that combine occluded and illusory portions. The effect is visible in each of the paired images. Viewing the display as a stereo pair, in which one image is shown to each eye, reveals the effect more vividly. The display may be free-fused by crossing or diverging the eyes, or viewed using a stereoscope. (b) Example of Petter's effect.

Petter noticed an interesting regularity in such displays. Where interpolated boundaries cross, the pair crossing the smaller gap appears in front, having illusory contours, whereas the other object's boundaries are seen as occluded in that region. Petter's claim, which has been confirmed experimentally, provides a strong argument for an interpolation mechanism common to illusory and occluded contours. The reason is that the determination of illusory versus occluded appearance here depends on a comparison of interpolated contours. That the outcome is determined by the relative lengths of crossing contours implies that the system must determine the locations and extents of contour interpolation prior to the determination of final appearance as illusory or occluded.

Surface Processes

Contours are not the only means by which the visual system links visible areas; surface relations also play a part (Figure 3). Some of the circles in the display, such as the gray ones, appear as spots on the surface. In contrast, most of the black circles appear to be part of a single, occluded, black figure, visible through holes. The white spots also appear to be holes rather than spots; through them, the white background surface is seen. These perceptual experiences arise from the surface interpolation process. Visible regions are connected across gaps in the input based on the similarity of their surface qualities (e.g. lightness, shade and texture). These connections cannot be given by contour interpolation, as the circles have no contour junctions. Certain rules govern surface interpolation; for example, it is confined by real and interpolated edges (Yin *et al*., 2000). In Figure 3, note that the rightmost black circle does not link up with the occluded object: this is because that circle does not fall within the real or interpolated contours of the black object. Whereas contour interpolation processes are relatively insensitive to relations of lightness or color, the surface process depends crucially on these. Notice that the gray dot on the lower left does not appear as part of the occluded object, despite being within the interpolated and real contours of the black object.

This phenomenon of surface interpolation under occlusion appears to be one of a family of surface spreading or 'filling-in' phenomena. A related spreading phenomenon, studied by Yarbus (1967), involved images stabilized on the retina. When an image or part of one is stabilized on the retina, it disappears in a few seconds. In Yarbus's displays, a red circle was shown on a blue background. The

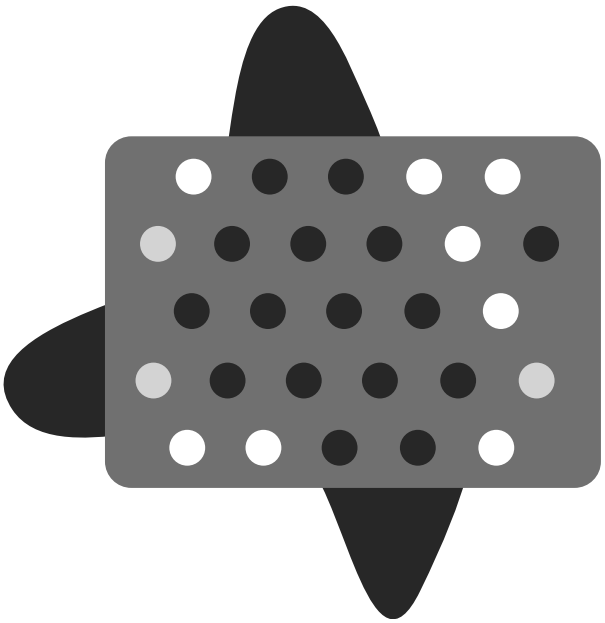


Figure 3. [*Figure is also reproduced in color section*.] Surface interpolation process.

contour between the red and blue areas was retinally stabilized, although the border of the blue background was not. After a few seconds, the central circle disappeared. Instead of the homogeneous gray that observers would see if an entire image were stabilized, the center circle took on the color of the surround – it appeared as blue. With the stabilized boundary 'removed', the surrounding surface color appeared to spread into the central region.

Another example is completion across the blind spot, an area of the retina about 15° off center (toward the nose) where blood vessels and nerves pass through the retina and no receptors exist to pick up light information. We do not ordinarily notice the blind spot; no hole appears in the visual field when we stare at a wall. This fact may seem remarkable if one knows that the blind spot has a diameter that is approximately 5° of the visual field. (In contrast, the fovea, responsible for all of our finely detailed vision, is only 1–2° in diameter.) Similarly, in certain neurological injuries or acute events, including migraine headaches, there may be a functional scotoma, or hole in the visual field (such that no local information is received there); here too, no gap is observed visually, as surrounding surfaces fill in.

These filling-in phenomena suggest general processes that serve to represent continuous surfaces in the world despite gaps in the local information.

THREE-DIMENSIONAL AND MOTION INFORMATION

Most studies of filling-in phenomena have involved static, two-dimensional displays, as these are more easily created and manipulated. Given that the world is three-dimensional and contains moving objects, we might expect that contour and surface processes use three-dimensional and motion-carried information and produce representations of three-dimensional objects. Study of these aspects has made it clear that both expectations are correct.

Visual contour and surface interpolation processes use the orientations and relations of visible object parts in all three spatial dimensions. Figure 4 shows two examples. Both are stereo pairs containing two X-shaped images. Each pair may be free-fused by crossing the eyes until a single image appears. Because of depth information given by differences in the images shown to the two eyes, the resulting displays in Figure 4a and 4b have very different outcomes in terms of their segmentation into objects and interpolated contours and surfaces. In Figure 4a, two separate, vertically oriented objects are seen, with illusory contours in the

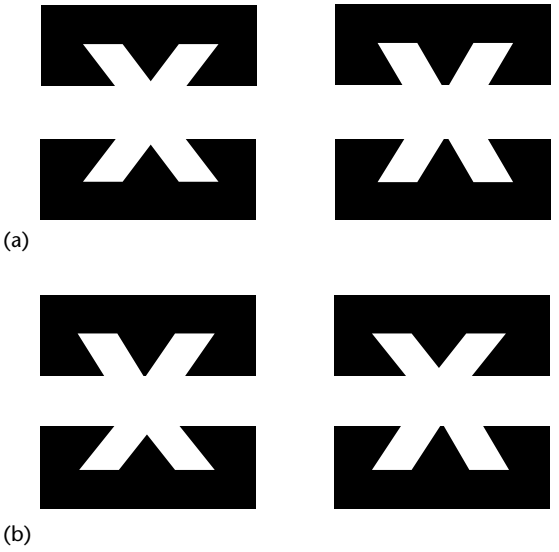


Figure 4. Three-dimensional interpolation. Each display is a stereo pair, in which the left and right images are separate views to be shown to the two eyes (the display is designed to be free-fused by crossing the eyes). The displays in (a) and (b) have the same visible parts; however, their positions have been shifted in the two eye's views to create binocular disparity, giving them different apparent positions in depth. (a) Three-dimensional relations of the visible parts favor completion as two upright, curved objects, separated horizontally. (b) Three-dimensional relations of the visible parts favor completion as two crossing bands curving in depth.

middle of the right-hand object and occluded contours for the left-hand object. In Figure 4b, two diagonally crossing objects are seen, one bowing outward with illusory contours and the other receding behind, having occluded contours. Experimental studies confirm that interpolation processes use three-dimensional information and produce three-dimensional representations of contours and surfaces.

The fourth dimension – time – is also important to our perception of coherent objects. Moving observers looking through foliage or other occluding objects see a constantly changing set of patches from objects behind. Visual processes handle this kind of situation by holding visible fragments in a brief memory buffer, along with information about their motion relative to the observer. Motion information is used to extrapolate the positions of these stored fragments over time. As new contours and surface patches become visible, these are connected to the previously viewed pieces following the same geometry of relatability, here applied to currently visible and previously stored fragments in updated spatial positions.

CONCLUSION

Many aspects of these processes and their implementation in the brain have yet to be explained. Although these remarkable interpolation abilities pose complex challenges for researchers, their seamless operation in daily life empowers thought and behavior by allowing us to perceive complete objects from fragmentary views of the world.

References

Heitger F, Rosenthaler L, von der Heydt R, Peterhans E and Kübler O (1993) Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research* **32**: 963–981.

Kanizsa G (1979) *Organization in Vision*. New York: Praeger.

Kellman PJ and Shipley T (1991) A theory of visual interpolation in object perception. *Cognitive Psychology* **23**: 141–221.

Kellman PJ, Yin C and Shipley TF (1998) A common mechanism for illusory and occluded object completion. *Journal of Experimental Psychology: Human Perception and Performance* **24**: 859–869.

Petter G (1956) Nuove ricerche sperimentali sulla totalizzazione percettiva. *Rivista di Psicologia* **50**: 213–227.

Ringach DL and Shapley R (1996) Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research* **36**: 3037–3050.

Schumann F (1904) Einige Beobachtungen uber die Zusammenfassung von Gesichlseindrucken zu Einheiten. *Psychol. Stud*. **1**: 1–32.

Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung* **4**: 301–350. [Abridged version entitled 'Laws of organization in perceptual forms' appears in Ellis WD (translator) *A Source Book of Gestalt Psychology*, London: Routledge & Kegan Paul, 1938.]

Yarbus AL (1967) *Eye Movements and Vision*. New York: Plenum Press.

Yin C, Kellman PJ and Shipley TF (2000) Surface integration influences depth discrimination. *Vision Research* **40**(15): 1969–1978.

Further Reading

Meyer G and Petry G (eds) (1987) *The Perception of Illusory Contours*, pp. 151–164. New York: Springer.

Shipley TF and Kellman PJ (eds) (2001) *From Fragments to Objects: Segmentation and Grouping in Vision*. Amsterdam: Elsevier.

Vision: Top-down Effects

Intermediate article

Mary A Peterson, University of Arizona, Tucson, Arizona, USA

CONTENTS

'Top-down effects' are effects that originate in high levels of the hierarchy of visual processes and exert an influence at lower levels. Examples of visual information thought to reside at, or arise from, high levels in the hierarchy are: the perceiver's intentions, expectations, attentional goals, and memories established on the basis of past experience.

INTRODUCTION

It is now generally acknowledged that the perceiver's past experience, intentions, expectations, and attention influence visual perception. These are called 'top-down' effects because they are thought to originate at high levels within the hierarchy of visual processes and to exert their influence on lower levels in the hierarchy. In contrast, the sources of 'bottom-up' effects are thought to reside at low levels in the visual hierarchy and to be immune to influences from higher levels.

Following the Gestalt revolution (*circa* 1920–1940), most theorists believed that much of visual perception occurred in a bottom-up fashion and was immune to influences from higher levels (Fodor, 1983). High-level processes were thought to exert their effects only postperceptually – that is after the products of early perceptual processes were available (on decision processes, for instance). Since the 1980s, however, improved and varied methods have revealed that high-level processes can affect perceptual processes *per se*, and not just postperceptual processes. Plausible mediators for these top-down effects have been identified through both computational modeling efforts and neuroanotomical and neurophysiological experiments.

This article discusses both interactive activation models that allow for top-down effects on perception, and empirical evidence indicating that the perceiver's past experience, perceptual intentions, expectations, and attention affect visual perception.

INTERACTIVE ACTIVATION MODELS

Traditionally, visual perception and cognition were thought to be related hierarchically, with the higher stage (cognition) operating on the completed outputs of the lower stage (visual perception). As an alternative, McClelland and Rumelhart (1981) proposed an interactive activation model (IAM). In the IAM, processes residing at higher levels operate on the partial outputs of lower levels; they need not wait for processing at lower levels to be completed. Further, the partial outputs from higher-level processes are fed back to influence lower-level processes, thereby allowing for interactions between levels.

The first of many behavioral effects to be modeled by the IAM was the word superiority effect – that observers can tell which of two letters has been shown in a brief exposure more accurately when the target letter is presented as part of a word rather than alone. Word superiority effects are observed even when there is no reason to expect the word context to be useful (e.g., when the words 'hit' and 'hot' serve as the context for the target letters 'i' and 'o'). In the IAM, at least two stages of processing are involved in letter discrimination: a letter identification stage and a higher-level word recognition stage. Partial outputs from the letter stage are fed forward to the word stage. At the word level, representations compete. Over time, one word dominates; consequently feedback from that word enhances processing of the target at the letter stage, increasing discrimination accuracy. More recent experiments have revealed evidence of interactions between higher and lower levels.

As originally proposed the IAM was considered to be biologically implausible, but recent refinements have increased its plausibility. The discovery of ubiquitous feedback connections in the brain identified a potential neural substrate for top-down effects. Furthermore, behavioral research demonstrated that top-down effects from past

experience, attention, intention, and expectation can affect perception *per se*, rather than postperceptual processes, as described in the following sections.

PAST EXPERIENCE

In this section, three types of past experience effects on perception will be discussed: effects on figure assignment arising from previous exposure to objects, the role of past experience in determining the parts of an object, and how seeing objects in their typical context affects perception.

Object Memory Effects on Figure Assignment

A process that is thought to occur early in the visual processing hierarchy is figure–ground assignment. Figure–ground assignment occurs when two adjacent regions share a border. The border is typically assigned to one of the two regions and not to the other. The region to which the border is assigned has a definite shape, and is called the 'figure'. The region to which the border is not assigned lacks a shape near the border, and typically appears simply to continue behind the figure; this region is called the 'ground'. Thus, figure assignment partitions the visual field into shapes; until this organization is accomplished, no shaped entities are perceived.

It was long thought that figure–ground organization had first to partition the visual field into shaped figures and shapeless grounds before memories of previously seen shapes or objects could be accessed. Following figure–ground assignment, object memories were thought to be accessed by figures and not by grounds. This 'figure assignment first' assumption was based on the Gestalt argument that past experience needed an organized substrate to operate upon; the unorganized input was insufficient to constrain past experience. Shaped entities were deemed the necessary substrate for the type of past experience coded in shape or object memories. The belief that shapes necessarily serve as the substrate for matches to object memories rests on a naive understanding of the nature of object memories, however, inasmuch as object memories are not themselves shapes. Another reason to hold the 'figure assignment first' hypothesis is the belief that early vision must provide an accurate representation of the external world, and it can only do so if it is unaffected by past experience. The faulty assumption that top-down effects necessarily dominate bottom-up processes is implicit in this reasoning, however.

Contrary to the traditional view, recent research has revealed that at least partial memories of previously seen objects are accessed before figure assignment and can affect its outcome (Peterson, 1994). For instance, the white regions in Figure 1 are more likely to be seen as shaped figures when the displays are upright (as in Figure 1) rather than inverted (as can be seen by viewing the page upside down). Changing the orientation of Figure 1 from upright to inverted does not change the bottom-up factors known to influence figure assignment (e.g., the depth cues and the Gestalt configural cues of symmetry, enclosure, smallness of relative area, and/or convexity). However, it takes longer for object memories matching the white region to reach some threshold activation level for inverted displays compared to upright displays. This delay is sufficient to diminish or remove the influences from object memories on figure assignment.

Peterson (1994) and Peterson *et al*. (2001) proposed that these effects are mediated by edge-based access to partial object memories; thus, neither shaped entities nor whole regions are necessarily the substrate for access to object

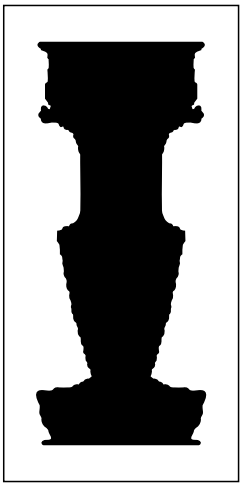


Figure 1. A figure–ground display drawn by Julian Hochberg and used by Peterson and colleagues in which two adjacent regions – black and white – share a boundary. The Gestalt configural cues of symmetry, enclosure, and smallness of relative area favor the interpretation that the black central region is the shaped figure. The white surround portrays portions of two women shown in partial silhouette from head to foot. Thus, object memory cues, effective for upright but not inverted displays, favor the interpretation that the white regions are the shaped figures. Reprinted from Peterson, Harvey and Weidenbacher (1991) with permission from the American Psychological Association.

memories. Furthermore, Peterson and colleagues showed that object memory cues do not always dominate bottom-up cues to figure assignment; object memory cues seem to be simply one of an ensemble of cues to figural status. Therefore, concerns that top-down effects might reduce the ability for early visual processes to model the external world are unfounded.

Are Parts of Objects Innate or Learned?

Object memories store past experience with objects. But what are the critical features or parts of object memories? Are they innate or are they determined by past experience (i.e., learned)? Some theorists propose that a small, fixed alphabet of geometrical components is used to represent all objects (e.g., Biederman, 1987). Other theorists argue that the features, or parts, used to represent objects are flexible and depend upon the observers' past experience or current goals. Consistent with the latter idea, Goldstone and his colleagues (e.g., Goldstone *et al*., 2000) asked two different groups of observers to use different rules to classify generated novel outline objects. The different rules required observers to attend to different parts of the same objects. In both cases, the parts on which the classification response depended were perceptually good parts that were highly visible to both groups of observers. After observers learned the classification task, they performed a different task in which they judged whether a part displayed after a whole object had been present in the object. Observers' 'present' responses were faster for parts that had been diagnostic for the classification task they had learned rather than the other classification task, even though the parts were equally good parts *a priori*. Thus, top-down effects instantiated in the participant's classification strategy have a powerful influence on perception by selecting the parts used to code objects in memory.

Context Effects on Object Perception

Visual perception can also be influenced by knowledge gained from past experience with scenes where objects were likely to be found. For instance, observers' ability to perceive objects is affected by knowledge regarding physical laws such as support relations (e.g., couches are likely to be found resting on the floor rather than floating in the sky). The ability to identify objects is also influenced by the likelihood of finding an object in a given context (e.g., airplanes are more likely to be found in the sky than in the living room). When pictures of scenes are presented briefly, objects are more likely to be detected or identified if they are portrayed in appropriate contexts (or in appropriate locations within a context) rather than in neutral or inappropriate contexts or locations (Biederman *et al*., 1974). Such context effects might be 'top-down' effects in that they reflect feedback from higher, semantic, levels of processing. Alternatively, context effects might reflect the statistical regularities in the environment, extracted over repeated exposures to similar scenes and used to guide attention to the location of a target (Chun and Jiang, 1998). It is not yet clear where knowledge regarding statistical regularity is stored; it may not be stored at a sufficiently high level in the hierarchy of visual processing to be considered a top-down effect.

Summary

In summary, behavioral evidence indicates that past experience can affect a variety of visual processes, including those thought to reside at very low levels in the processing hierarchy and, consequently, to be immune to effects of past experience. Past experience can change the manner in which objects are segregated from backgrounds and the manner in which parts are segregated from each other. However, it is not clear whether past experience necessarily operates in a top-down fashion. Effects of past experience might be stored at many levels in the visual processing stream. A challenge for future research will be to determine which effects of knowledge and past experience reflect top-down influences and which are coded at low levels of the hierarchy.

INTENTIONS, EXPECTATIONS, AND ATTENTION

Effects of perceivers' intentions and goals would seem to be truly top-down effects. Such top-down effects on perception have been observed in a number of settings. For instance, Hochberg and Peterson (1987) showed that an observer's perceptual intentions can affect the perceived depth organization of two-dimensional (2-D) and three-dimensional (3-D) cubes. Observers viewed small cubes that were biased near one corner toward one of the two possible depth organizations (facing upwards and to the right or downwards and to the left) but remained unbiased near another, nearby, corner. A sample is shown in Figure 2. Observers intentions were manipulated through instructions to try to see the cube facing downwards and to the left on some trials, and upwards and to the right on

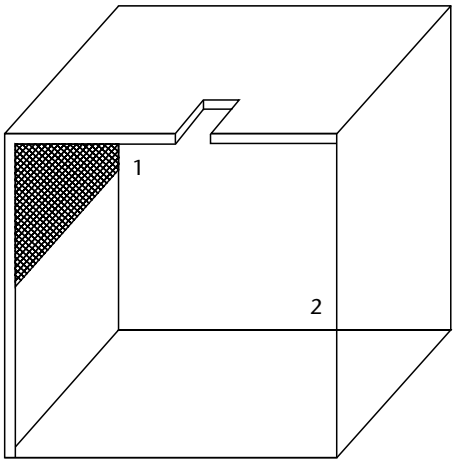


Figure 2. A schematic of a 2–D display used by Hochberg and Peterson. The cube is biased at its upper left corner toward the interpretation that it faces downwards and to the left rather than upwards and to the right but remains unbiased at the lower right corner. Point 1 marks the intersection that was fixated near the upper left corner; point 2 marks the intersection that was fixated near the lower right corner. Reprinted from Peterson and Hochberg (1983) with permission from the American Psychological Association.

other trials. The critical finding was that, despite the proximity of the biased corner, observers' intentions exerted a strong influence on perception when they fixated on the unbiased corner. For both 2-D and 3-D cubes, instructed intentions affected the perceived ordering of depth planes, even though fixation remained constant. (The viewer's intentions did not eliminate reversals, however, consistent with the idea that top-down factors do not dominate all other factors. In other words, intention affects perception but does not dominate it.)

These results could not be interpreted in terms of postperceptual processes because reports about perceptually coupled variables agreed with direct reports about perceived depth. Perceptual-couples are variables that co-vary in perception, even if they do not co-vary in the stimulus. The knowledge coded in these couplings is implicit but not explicit. Hence, reports regarding perceptually coupled variables can be used to index whether viewers' reports correspond to what they truly perceive. For example, for moving observers viewing a stationary 3-D cube, illusory concomitant motion is coupled to a perceived depth reversal. Similarly, for stationary viewers, the perceived direction of rotation of an oscillating cube is coupled to perceived depth organization. Experiments showed that viewers' reports about these perceptually coupled variables varied as expected if perceptions were altered by their attempts to follow the intention instructions, providing strong evidence that intention can affect perceived depth organization (e.g., Hochberg and Peterson, 1987).

Recent investigations have focused on whether the viewer's goals and task expectations determine which aspects of a display will attract attention or whether salient stimulus features can 'capture' attention regardless of the viewer's current goal. Evidence suggests that observers may be unaware of stimulus features unrelated to their task. Yet the same features are easily detected when they are task-relevant. For instance, Mack and Rock (1998) asked participants to determine which of the two arms of a cross was longer. The task was difficult because the cross was displayed briefly and the arms were only slightly different in length. After just a few trials, Mack and Rock changed something about the background against which the cross was presented. In one such experiment, they exposed a simple geometric shape (a square, a circle, or a triangle) in one of the four quadrants sketched by the cross. Of their observers, 25 percent failed to detect the unexpected shape; indeed, they failed to notice that anything unusual had occurred on that trial. Thus, it seemed that the shape onset did not necessarily capture attention if observers were focusing on another task.

Visual information unrelated to the observers' goals might be processed outside of awareness; the critical question is to what extent the perceiver's expectations close off access to awareness. A recent model proposed by DiLollo *et al*. (2000) posits that attention plays an important role in maintaining low-level activation long enough to allow top-down processes (feedback or 're-entrant' processes) to confirm expectations established at higher levels. With this model, expectations unconfirmed by re-entrant processes do not enter awareness.

CONCLUSION

Behavioral evidence indicates that past experience, intentions, expectations, and attention interact with bottom-up processes to structure visual perception. Neural connections capable of subserving these effects have been identified and biologically plausible computational theories have been proposed to account for them. A task for the future is to delimit which types of knowledge and past experience can affect vision; some types clearly can, whereas other types cannot (Peterson *et al*., 1991). It will also be important to determine where various types of knowledge are stored in the visual hierarchy and where various tasks are accomplished. Only by

doing so can we determine which effects reflect within-level versus between-level (top-down) interactions.

References

Biederman I (1987) Recognition by components: a theory of human image understanding. *Psychological Review* **94**: 115–147.

Biederman I, Rabinowitz JC, Glass AL and Stacy EW (1974) On the information extracted from a glance at a scene. *Journal of Experimental Psychology* **103**: 597–600.

Chun MM and Jiang Y (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology* **36**: 28–71.

DiLollo V, Enns JT and Rensink RA (2000) Competition for consciousness among visual events: the psychophysics of reentrant processes. *Journal of Experimental Psychology: General* **129**: 481–507.

Fodor JA (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Goldstone RL, Steyvers M and Spencer-Smith KA (2000) Interactions between perceptual and conceptual learning. In Dietrich E and Markman AB (eds) *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*, pp. 191–228. Mahwah, NJ: Lawrence Erlbaum.

Hochberg J and Peterson MA (1987) Piecemeal organization and cognitive components in object perception: perceptually coupled responses to moving objects. *Journal of Experimental Psychology: General* **116**: 370–380.

Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception. 1: An account of basic findings. *Psychological Review* **88**: 375–407.

Peterson MA (1994) Shape recognition can and does occur before figure–ground organization. *Current Directions in Psychological Science* **3**: 105–111.

Peterson MA, Harvey EH and Weidenbacher HL (1991) Shape recognition inputs to figure–ground organization: which route counts? *Journal of Experimental Psychology: Human Perception and Performance* **17**: 1075–1089.

Peterson MA and Hochberg J (1983) Opposed-set measurement procedure: a quantitative analysis of the role of local cues and intention in form perception. *Journal of Experimental Psychology: Human Perception and Performance* **9**: 183–193.

Further Reading

Hochberg J (1968) In the mind's eye. In: Haber RN (ed.) *Contemporary Theory and Research in Visual Perception*, pp. 309–331. New York, NY: Holt, Rinehart, Winston.

O'Reilly RC (1998) Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences* **2**: 455–462.

Palmer SE (1975) The effects of contextual scenes on the identification of objects. *Memory and Cognition* **3**: 519–526.

Reingold EM and Jolicouer P (1993) Perceptual versus post-perceptual mediation of visual context effects: evidence from the letter-superiority effect. *Perception & Psychophysics* **53**: 166–178.

Schyns PG, Goldstone RL and Thibaut J-P (1998) The development of features in object concepts. *Behavioral and Brain Sciences* **21**: 1–17.

Visual Art

Introductory article

Robert L Solso, University of Nevada, Reno, USA

CONTENTS

Many neurological pathways and interactions are involved in viewing art. Our perceptions are based not only on visual stimuli but also on higher-order cognitive associations which give meaning to what we see.

INTRODUCTION

When we humans look at art a fascinating sequence of neurological, perceptual, and cognitive phenomena emerge in which the art work is seen and understood in less time than it takes to read these words. Cognitive scientists have unraveled many of the strands of the tangled neurological pathways and interactions involved in the process of perceiving art. Here, some of the main ideas regarding cognitive science and art are reviewed.

PSYCHOLOGICAL PRINCIPLES AND ART

Looking at art (as well as visual perception) engages two major systems: nativistic perception and directed perception.

Nativistic Perception

Perception of visual events is based on the idea that people have certain inborn ways in which visual stimuli, including art, are initially organized and perceived. Sometimes nativistic perception is called 'natural' perception as it is the way people look at visual events intrinsically, or as we might view them without learning. These perceptual qualities are genetic and affect all human perception uniformly. Casually speaking, they are 'hard-wired' in our neural system, by which it is meant that the way visual information is initially processed starts with an analysis of basic visual elements. Thus, when one looks at a particular painting the eye and brain distinguish one object from another, detect different intensities of reflected light, see colors, and discern basic shapes and forms. Nativistic perception shares many features with 'bottom up' perception in which basic objects, lines, and geometric primitives are detected and then combined to make more complex, meaningful forms.

The full comprehension of art involves more than the fundamental identification of basic features. These include the sociology of the period, the meaning of the subject, and the personal relevance of the piece among other factors – a topic called directed perception, to which we now turn.

Directed Perception

Directed perception is based on the idea that we interpret the world through our past learning experience. In the early halcyon days of behaviorism a radical view suggested that all human characteristics were the consequence of environmental experiences and learning. The concept was related to the idea that children are born into the world as a *tabula rasa*, or blank page upon which experiences are recorded.

This extreme position has been largely discarded but the view that past experiences influence the perception of objects, especially meaningful objects, has an important role in the psychology of art. Sometimes this approach to the viewing of art has been called a 'top down' process, as it implies that the observer begins with a schema or plan as to what might be seen in a work of art. Thus, in directed perception, the viewer attends to art with a rich background of personal experiences. Directed perception has been conceptualized as 'seeing the world with a thousand hypotheses', as one looks for meaning, relationships, and intentions when viewing art. While the different theories represent different views, most believe that viewing art incorporates features from both ideas – an approach called the 'holistic' perception of art.

Holistic Perception

Holistic perception is a comprehensive approach to the viewing and understanding of art which engages both nativistic and directed perception. During the initial stages of processing visual events, basic features, contours, figures, and colors are sensed and perceived. These perceptions are then blended with one's knowledge of the world, information about the art object, and personal beliefs, which leads to a 'deeper' understanding of the viewed object.

ART PERCEPTION: A COGNITIVE VIEW

The modern period in psychology has emphasized an information processing model in which external events, such as a work of art, are sensed, perceived, and processed in a series of stages as we think about the object and interpret its meaning. The advantage of an information processing model over older models (such as behaviorism or psychoanalysis) is that it provides an analytic system in which various stages may be studied more or less independently of other stages. Applying this model to the perception and understanding of art might follow the following sequence (with physiological structures in parentheses):

1. Sensing and perceiving the object (retina and primary visual cortex).
2. Identification of primitive structures (primary visual cortex).
3. Processing of basic content such as location, forms, color, and faces (subsequent dorsal and ventral streams in the brain).
4. Derivation of complex meaning such as social or political significance (many parts of the brain including the associative centers in the frontal regions).

Both nativistic and directed perception take place in the various stages of information processing, as shown in the following example.

Take a look at the reproduction of the painting by Raphael shown in Figure 1. What do you see? Before you go on reading this section, take a moment to think about what you see and how you see it.

The painter Raphael (1483–1520), one of the most celebrated and talented artists of the Renaissance, produced this exquisite painting in 1505. From a cognitive or perceptual view Raphael shows great sensitivity to the way a scene might be parsed by a human viewer; from an artistic view it is an example of superb technical skill; and from a religious view, it is a poignant representation of the Madonna, St John the Baptist (on the left) and Jesus (on the right).

Figure 1. [*Figure is also reproduced in color section.*] *Madonna of the Meadows* by Raphael (detail). Used by permission of the National Gallery of Art.

The human eye and brain look for stable elements in the perceptual field and, in the case of this picture, tend to organize the scene by means of a triangle (Figure 2). When viewing this painting our eyes are drawn to the face of Mary which forms the apex of a perfect triangle that subtends an angle embracing the principal figures of the piece. Furthermore, the triangle is one of the most pleasingly stable of all geometric forms, which satisfies our natural yearning to organize visual scenes into stable, discernible views of the world. Metaphorically, the triangle in Christian symbolization represents the Holy Trinity – God the Father, the Son, and Holy Ghost. In this painting three people (Mary, John the Baptist, and Jesus) represent important people in Christendom. Note that the left foot of Mary is brightly portrayed and thus frames the view of the observer. John is presenting a bamboo cross to the infant Christ (years before the crucifixion) which animates an otherwise static picture. Mary's gentle hands caress Jesus, which lends a tender, calming atmosphere to the scene. An analytic, information processing examination of the painting reveals both nativistic and directed perception which is carried out by human neurological structures and processes.

THE BRAIN AND ART

Investigations of the neural pathways in the brain activated by visual perception have broadened our

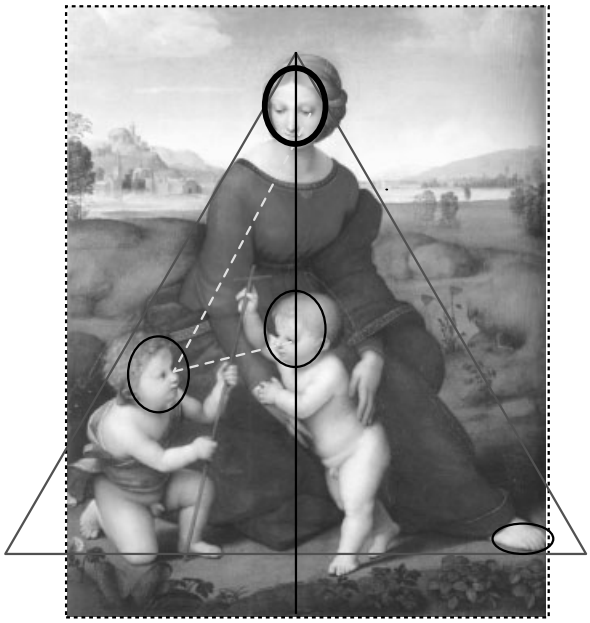


Figure 2. [*Figure is also reproduced in color section.*] *Madonna of the Meadows* (detail) showing major perceptual-organizational features. We tend to organize the painting in terms of basic shapes (triangle) and balance (ellipses). Art perception is influenced by the natural inclination to parse perceptual scenes into basic shapes, colors, and patterns.

understanding of how art is perceived and processed by human observers. These studies have been greatly enhanced by the use of imaging technology such as functional magnetic resonance imaging and positron emission tomography which give accurate views of the regions of the brain active during information processing. From the moment the retina in the eye senses energy within its level of receptiveness, a series of predictable routes of neurological information processing are followed. Two types of early processing of art have been found in the pathways in the brain. The ascending route, the dorsal stream, is associated with location and motion and has been called the 'where is it?' pathway, whereas the descending route, the ventral stream, is associated with color, form, detail, and faces and has been called the 'what is it?' pathway. The neurological pathways implicated in visual perception are extremely complicated, and there appears to be considerable 'cross-talk' between the streams.

The viewing of and understanding of art initially engages both of these pathways and many more. From our knowledge of these routes it would appear that two important questions about the world in general, and the perception of art in particular, are addressed by the brain's neural circuitry: where is the object, and what is it? These neurological mechanisms originated during our long evolutionary history for the purpose of survival and procreation. Now, such pathways are used in the processing of art as well as other forms of visual processing.

PAST EXPERIENCE AND ART PERCEPTION

The viewing and understanding of art is influenced by one's past experience and current inclination. Thus, someone interested in biology might see the details of plant and animal life in a painting, while another person interested in psychology might attend to the people shown in the same painting. Succinctly stated, we 'see' what we find interesting or are encouraged to see. As an example of this idea, students of art who already know about the theme of a particular painting, its style, the techniques used, its historical setting, and other important features, will probably see more details in the piece than a novice. In addition, those who view art with an 'intelligent' eye will comprehend a piece in a larger intellectual context than one unschooled in art. (The same is true of experts in other fields – radiologists, for example, who examine an X-ray are able to see in a glance much more than a novice could see in minutes.)

Such effects of experiences have been discussed since the beginning of psychology. William James, America's foremost pioneer psychologist, suggested in 1890 that if several people of different backgrounds visited Europe one might bring home picturesque impressions, while another might come back with impressions of the cost of the trip, and yet another would see and remember the restaurants and theaters. In modern terms, we conceptualize the way people organize the impressions of the world, including their perception of art, in terms of personal schemas. Experimental evidence has been presented which confirms the idea that we see and remember components of a visual scene according to the schema that is operating. In one study people were asked to write a paragraph about what a member of an occupational group would do on a typical day: for example, the day of a nurse, an architect, or a police officer. With these induced schemas the groups were shown a series of pictures like the one in Figure 3. Afterwards they were asked to recall as much as they could about the pictures. In general, the group who had written about a nurse saw matters related to health and caring, the architect group saw things dealing with the structure of the building, and the

Figure 3. [*Figure is also reproduced in color section*.] Scene in an art gallery. People with different vocational schema 'see' different aspects. An architect might attend more to the physical characteristics of the room such as the absence of windows, while a police officer might attend to the surveillance camera. Art perception is influenced by who we are and what we are encouraged to see. Photograph by R. Solso.

police officer group saw details related to criminal activities. From studies of this type we can see that the perception of art is critically affected by who we are, what we know, and what situational schema is activated. From studies of naturalistic perception we also know that art perception is influenced by the natural organization of forms, shapes, colors, and the like.

Modern cognitive science has shown that the perception and understanding of art are based on our basic apprehension of primitive stimuli and accessing higher-order cognitive associations about the contents of a painting which gives meaning to the things we sense.

CONCLUSION

Studies in cognitive science have shown us that the viewing of art is dependent upon two systems which work together to provide an integrative appreciation of art. Upon viewing a work of art the visual signals are initially processed by the brain in featural analysis. These basic elements are combined with our knowledge, interest, and viewing proclivity to construct an overall understanding of the art object.

Further Reading

Arnheim R (1974) *Art and Visual Perception*. Berkeley, CA: University of California Press.

Gombrich EH (1982) *The Image and The Eye*. Oxford, UK: Phaidon.

Gregory RL (1987) *The Oxford Companion to the Mind*. Oxford, UK: Oxford University Press.

Shepard RN (1990) *Mind Sights*. San Francisco, CA: Freeman.

Solso RL (1994) *Cognition and the Visual Arts*. Cambridge, MA: MIT Press.

Solso RL (2003) *The Psychology of Art and the Evolution of the Conscious Brain*. Cambridge, MA: MIT Press.

Visual Attention

Intermediate article

Ronald A Rensink, University of British Columbia, Vancouver, British Columbia, Canada

CONTENTS

Visual attention is the factor controlling the selective access and integration of visual information.

INTRODUCTION

Although much of vision appears to be effortless and all-encompassing, nevertheless there are limits to what it can do. For example, consider air traffic control, where it is imperative to keep track of all moving items in a display (corresponding to the airplanes in an airspace). If only a single item is present, it can generally be tracked without problem. It is also possible to track four or five items simultaneously, although some effort is needed to do so. However, for 20 or 30 items, even a maximal effort will not suffice, and the task must be shared among several controllers. What appears to be happening in such cases is that visual perception is constrained by a consciously controlled factor within the observer – a factor that enables certain types of processing to take place, but which is limited in the extent to which it can be applied. This factor is termed *visual attention*.

Interestingly, although most observers immediately know what to do when asked to 'pay attention' to a stimulus, it has been rather difficult to give this an objective characterization. Indeed, until recently there was no general consensus on the basic function of attention, and at various times it was associated with such things as clarity of perception, intensity of perception, consciousness, and selection.

SELECTION

During the past few decades, considerable progress has been achieved by focusing on *selection* as the basic function of visual attention. Two types of selection are of particular importance. The first is *selective access* (i.e., allowing only certain parts or properties to be sent on to later processes). It was originally believed that selective access protected processors at higher levels from being overwhelmed by too much information. However, more recent research has tended to view selective access as a way to delimit control of various actions (e.g., focusing on the locations of items that are to be grasped).

The second type of selection is *selective integration* (i.e., combining selected parts or properties into structures that then form the basis of further processing). For example, three adjoining lines could be combined into a complete figure. This figure (and not the lines themselves) might then provide the basis for subsequent control of grasping. It was initially believed that such integration had to be selective in order to make good use of a limited amount of processing 'resource'. However, more recent research has tended to view selective integration in terms of the selective *coordination* of the outputs of multiple processes.

According to this more recent view, therefore, visual attention is not a unitary faculty. Instead, it is simply the selective control of information in the visual system, achieved in various ways by various processes. When considered from this perspective, several of the unresolved issues in earlier treatments of visual attention simply vanish. One such example is the issue of whether selection is 'early' or 'late' (i.e., whether it acts on simple, precategorical structures or more complex ones). Given that selective control may be carried out by a number of systems, there may not be a single site where attention acts, so this issue becomes meaningless.

Since a complete understanding of visual attention is still a long way off, this article will survey only the major behavioral techniques that have been used for its exploration and several of the more important results that have been obtained. Furthermore, it will focus entirely on the purely visual aspects of the processes involved (i.e., on how the stimuli themselves are handled, rather

than how responses to them are generated). Issues such as the sequencing of multiple responses are considered to involve central control at higher levels, so will not be discussed here. (*See* **Attention**)

VISUAL ORIENTING

Consider the situation where a driver is at a red traffic light, anxiously waiting for it to turn green. Here the stop light becomes the center of the driver's perceptual world, and when the light turns green the driver will respond almost immediately. This is an example of *visual orienting*. More generally, orienting can be defined as the alignment of the perceptual apparatus to allow optimal perception of what is happening (or expected to happen). Judging by the near universality of its occurrence, orienting is highly important to our survival in the world.

Orienting is usually studied by determining how well an observer can use advance information about a target item (e.g., position) to improve performance. Typically, the observer tries to detect, identify or locate the target. Two performance measures are generally used, namely *accuracy* and *response time* (RT). The effectiveness of orienting is measured by comparing performance when the observer has advance knowledge with performance when no such knowledge is available (Figure 1).

Overt Orienting

Perhaps the simplest form of orienting is the *overt orienting* of the eyes towards a stimulus. This is not usually a matter of eye movements alone, for the torso and head also contribute to it. Thus, for example, there is a reflexive orienting of the head towards items that suddenly appear in peripheral vision; this behavior is so basic that it is found even in newborn infants.

The net result of these movements is that the eye fixates on some part of the world. Since the resolution of a human retina is highest at its center, estimates of shapes or positions are most accurate for those items at the fixated location. As such, overt orienting is essentially a form of selective access. (*See* **Eye Movements**)

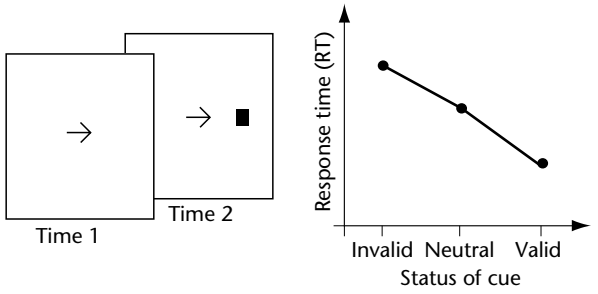


Figure 1. Covert orienting. When a cue indicates the location of the stimulus in a subsequent display, responses are faster and more accurate if the cue is valid (i.e., if the arrow points correctly to the location of the stimulus) than if the cue is uninformative (if it contains only the stem of the arrow). This improvement is considered to be due to the covert orienting of attention. Note that performance is impaired if the cue is invalid (the arrow points to the wrong location), presumably because attention needs to be reoriented from the incorrect location.

Covert Orienting

Observers can detect a target faster and more accurately if they are presented with a cue containing advance information about its location. Given that eye movements can be prevented, such improvement indicates the existence of a *covert orienting* performed by neural mechanisms. Enhancement begins within 50 ms of the cue and increases thereafter, reaching a peak about 200 ms after cue onset.

One explanation for this facilitation is that location is selected via a *spotlight of attention* (Posner *et al.*, 1980), which allows input only from the area that it 'lights'. Studies of this have been largely based on interference caused by irrelevant items, which can cause performance to degrade if near the target. Interference effects indicate that the spotlight covers about 1° of visual angle in central vision, although it can 'zoom out' to cover a larger area if necessary. The minimum area increases with eccentricity from the fovea, and appears to be greater in the upper visual field.

The relationship between covert and overt orienting is not a direct one. Attention does not have to be given to the fixated location – people can move their attention without moving their eyes. Moreover, although some form of attention is needed to select the target of an eye movement, this selection need not be accompanied by a withdrawal of attention from other items.

Space-based Versus Object-based Selection

Although there is considerable agreement that orienting is concerned with selection, there is less agreement about what exactly is being selected. For overt orienting, the situation is straightforward – orienting involves aligning the eyes to a particular two-dimensional location in space, and if the eyes are properly coordinated, they can select a

particular depth as well. Similarly, covert orienting need not be limited to a two-dimensional location in the visual field, but can also be based on three-dimensional depth.

An important issue is whether covert orientation selects for a particular location in space (that happens to be part of some object), or for a particular object (that happens to be at some location), or for both. When two overlapping items are presented, it is easier to report two properties from a single item than to report one property from each of the pair, indicating that covert orienting can be influenced by object structure (Duncan, 1984). The extent to which selection depends on spatial location and object structure has not yet been fully elucidated.

VISUAL SEARCH

Another task that is important in everyday life is *visual search*. For example, when looking for a friend in a crowd, an observer must check each person in turn until the friend is seen. More generally, a search task involves scanning through various items (cars, keys, people, etc.) until the desired item is found.

In studies of visual search, observers typically attempt to detect, identify, or locate a given (target) item among a set of other (distractor) items in a visual display. One of two performance measures are generally used: (1) accuracy on briefly presented displays, or (2) RT on displays that are continually visible. The central issue is the way in which performance is affected by the properties of the items in the display (Figure 2).

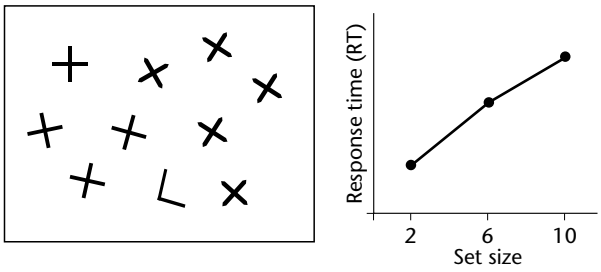


Figure 2. Visual search. The observer is asked to detect, identify, or locate a unique target placed among a set of distractor items. When the target contains a unique feature, performance is largely independent of set size (the number of items in the display). When the target is a strong conjunction of the distractor's features (when it contains the same basic features, such as the same line segments), performance is strongly dependent on the number of elements.

Pop-out

It is easy to see a single yellow dot among an array of blue dots. More generally, a target with a unique property can often be detected rapidly and with little dependence on the number of items in the display. Such *pop-out* is believed to indicate the presence of a distinctive property (or *feature*) in the target item. According to this view, various features (e.g., color or orientation) are computed rapidly and in parallel across the visual field. A unique feature will be *salient*, and so attract attention to its location, causing it (and its properties) to be selectively accessed. Salience can also arise via *differences* between adjacent features. An item will pop out if it is the only one with an orientation that differs greatly from those of its neighbors, even if its orientation occurs elsewhere in the display.

Support for this view comes from the existence of *search asymmetries*, where a switch in the role of target and distractor items strongly affects performance. For example, a 'Q' among a set of 'O's pops out, whereas an 'O' among a set of 'Q's does not. This can be explained by the 'Q' having a unique feature (the tail) that is not present in the 'O's, and thereby becoming salient. Since 'O' is distinguished from 'Q' by the absence of a feature, it can never be salient. Search asymmetries yield a powerful method for identifying features – pop-out if some feature in the target does not exist in the distractors, and a slower search otherwise.

Emergent Features

Many features have been found that are simple properties of the image (e.g., orientation, size, and color). However, other features are *emergent*, being derived from the image in a relatively complex way. For example, when targets are triangles and distractors are forked-shaped items constructed of the same line segments, the triangles pop out, indicating that some property of the line arrangement (presumably closure) acts as a feature. The existence of such features is due to processes operating *preattentively* (i.e., prior to the application of selective integration).

Several types of preattentive process are known to exist. These include grouping between items, grouping within items, and completion of occluded items. Features have also been found that are based on the (recovered) three-dimensional scene rather than the two-dimensional retinal image (e.g., surface convexity and three-dimensional orientation).

Although preattentive processes can make search easy under some conditions (i.e., if the target

contains an emergent feature), they can also sometimes make it more difficult. For example, if target and distractors are both line configurations with distinctive line segments, search can still be difficult if the overall length of the configurations is the same. This indicates that selective access is easiest not for simple attributes of the image (e.g., individual line segments), but rather for the more complex structures formed by preattentive processes.

Conjunctions

An important type of search is the *conjunction task*, where the target and distractor sets contain the same features but differ in how they are assembled. Conjunctions can be either *weak* (at the level of sets) or *strong* (at the level of items). An example of a weak conjunction is search for a red vertical line among green vertical lines and red horizontal lines. Here the set of distractors contains all of the features in the set of targets. However, at the level of individual items, the target remains unique. In contrast, in a strong conjunction each target item would contain the same features as a distractor item (e.g., an 'L' among 'T's).

If highly discriminable properties are involved, search for weak conjunctions can be relatively easy. This has been explained in terms of *guided search*, where a guidance mechanism either inhibits items with non-target features, or excites items with target features. Once these mechanisms have reduced the number of possible items, the remainder can be searched more easily – if a target has a unique feature, it will pop out. Thus, for example, if searching for a red vertical line, all green items can be inhibited, leaving the single remaining (red) vertical line to pop out. According to this view, therefore, search for weak conjunctions may simply involve selective access.

In a strong conjunction, items differ only in the way in which their features are arranged, and so guidance cannot be used. Search for strong conjunctions is often difficult, requiring at least 30–50 ms for each item. This has been taken to indicate the serial application of a spotlight of attention that binds the features of each item together, integrating them into an *object file* (i.e., a coherent collection of features) (Kahneman *et al*., 1992). As such, search for strong conjunctions is primarily an issue of selective integration.

Several issues with regard to strong conjunctions are currently unresolved. Although it has been proposed that selective integration acts on one item at a time, at a rate of 50 ms per item, it has also been proposed that it might act on clusters of four or five items, at a rate of 200–300 ms per cluster. Indeed, a serial mechanism may not even be involved – search could be carried out via a parallel process with a processing speed dependent on the number of items in the display. It is also not known what happens to object files after attention has been withdrawn. Although it has been suggested that object files exist for some time, it may be that they dissolve almost immediately (Wolfe *et al*., 2000).

INDUCED FAILURES OF PERCEPTION

Suppose that we are looking for a pencil, and we believe that it is yellow. If the pencil is indeed yellow, the attentional control based on our belief can facilitate its perception. However, if the pencil is blue, this same control can be detrimental – indeed, we may completely fail to see a blue pencil that is directly in front of us. More generally, failures of selective access and selective integration can provide important insights into the operation of the mechanisms involved. Such studies can also provide information about what aspects of perception continue when selective access and selective integration fail.

Change Blindness

A number of studies have examined the ability of an observer to detect, identify, or localize the occurrence of a change in a display. Performance is measured in one of two ways, either by accuracy on a pair of displays containing a single change, or by RT on displays that continually alternate between an original and a modified image. In both cases, observers often experience great difficulty in reporting the presence of a change that is made simultaneously with another event, such as an eye movement, blink, or flash (Figure 3).

This *change blindness* has been regarded as evidence that attention is needed in order to see change (Rensink, 2000). According to this view, when the local signals due to the change are swamped (or otherwise neutralized) so that they no longer capture attention, a time-consuming attentional scan of the display is needed. The observer will be blind to the change until the appropriate item is attended.

Although a selective process of some type is involved in perceiving change, the nature of this selectivity has not yet been established. One possibility is that a complete representation of the original display is formed, and that change blindness results from a limited ability to compare this representation with the current image. According to this view, the key factor is selective access to a

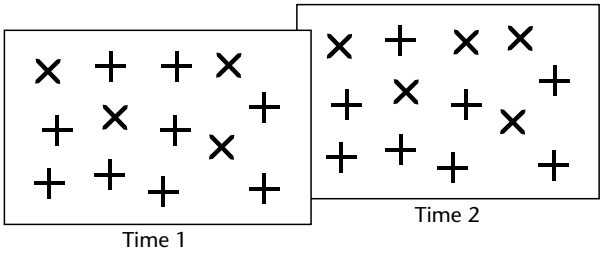


Figure 3. Change blindness. The observer is asked to detect, identify, or locate a unique change among a set of distractor items. This change can either be shown just once, or it can be continually presented by alternating between the two displays. Performance is extremely poor, provided that successive displays are separated by an interval of at least 80 ms.

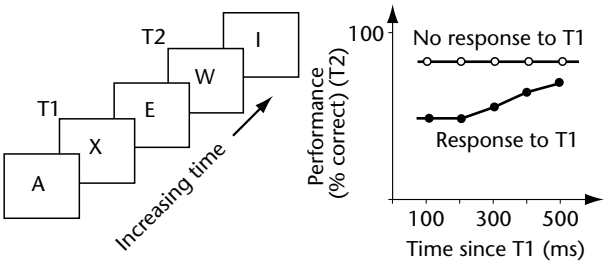


Figure 4. Attentional blink. When the observer is asked to report on two items (e.g., the colors of the consonants in a stream of letters), performance on the second of these (T2) is impaired if it is made within several hundred milliseconds of the occurrence of the first (T1). However, this degradation does not occur if no response is required of the first letter (e.g., if the observer reports only the color of the W).

comparison mechanism. Another possibility is that a complete representation of the display is never formed – representations of unattended items are volatile, and are simply replaced by representations of subsequent stimuli. Here attention is believed to act much as it does in visual search for strong conjunctions, integrating selected items over time and space so that they have a degree of spatiotemporal continuity, which then allows them to be seen to change.

Attentional Blink

Another phenomenon involving the failure to perceive stimuli is the *attentional blink*. Here a stream of successive items (usually letters) is presented at a location, and the observer attempts to report the presence of an item that has been designated in some way (e.g., by its color or its identity). Performance is measured by the accuracy of response. Whereas it is easy to detect a target when items appear at a rate of less than about 10 items per second, it becomes much more difficult to do this if the observer also has to respond to a target (T1) appearing earlier in the stream. Somehow, responding to T1 induces a 'blink' that makes it difficult for the observer to see – or at least respond to – a second target (T2) for the next few hundred milliseconds (Figure 4).

Control experiments have shown that this blink is not due to perceptual, memory or output limitations. Instead, it appears to result from attention being given to T1, leaving subsequent stimuli unattended and thus vulnerable to replacement by the items that follow them. Evidence in favor of this explanation is provided by the fact that T2 is difficult to report only if it is followed by another item. (Note that this replacement mechanism is similar to that proposed for change blindness.) Unseen T2 items can also facilitate the processing of semantically related items (Shapiro *et al*., 1997), providing further evidence that a considerable degree of processing can take place in the absence of awareness (and thus presumably in the absence of attention).

The time course of the blink itself has also been taken to correspond to an attentional *dwell time* (the time needed to integrate the properties of an attended object into a coherent form). According to this view, once attentional processing has started on T1, it cannot be simply halted when T2 is encountered, but must run its course. The duration of this blink appears to be about 300–400 ms.

Inattentional Blindness

A different approach to exploring attention involves asking observers to attend to a given event, and then introducing an unexpected item at some point. Performance is measured by the accuracy of responses to questions about the intrusive stimuli. Interestingly, observers often have great difficulty in reporting such stimuli, even though the items are easily seen if they are expected. This failure to report unexpected – and therefore unattended – stimuli is called *inattentional blindness* (Figure 5).

Early studies were *selective*, requiring the observer to attend to a subset of the stimuli (e.g., to attend only to the white-shirted or black-shirted players in a basketball game). Observers often had difficulty noticing the appearance of the unexpected stimulus under these conditions. Although superimposed images were used initially, later studies showed that these failures occurred also even when the stimuli were elements of a single scene and the unexpected stimulus was a person

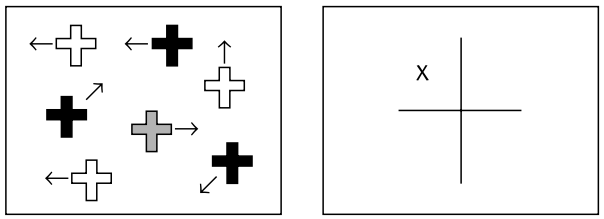


Figure 5. Inattentional blindness. (a) In a *selective* task, the observer must select a subset of the items in the display (i.e., track a set of moving black crosses among a set of white ones). While engaged in such selection, the observer will generally be blind to the occurrence of an unexpected item, even if it is unique (e.g., a gray cross moving in a unique direction). (b) In a *nonselective* task, the observer is presented with a pair of lines and asked to judge which of the two lines is longer. After a number of such tests, a display is presented that contains an extra element (e.g., a letter). Again, detection of these unexpected elements is generally quite poor.

dressed as a gorilla (Simons and Chabris, 1999). For selective tasks, inattentional blindness appears to be at least partly due to observers inhibiting the features of the irrelevant stimuli. If this were so, it would illustrate a failure of selective access.

Mack and Rock (1998) introduced a *nonselective* variant, where attention could be given to all stimuli that were present before the appearance of the unexpected item. Here observers attended to an overlapping pair of lines (one horizontal and one vertical) and had to judge which line was longer. Again they often failed to see the unexpected item under these conditions, even when it was at the center of fixation. It is still unknown whether inhibition is also an important factor in nonselective tasks.

Although observers may not report seeing an unexpected item, such items can still influence conscious perception. For example, surrounding lines can induce a length illusion in the test lines that are perceived, even if the surrounding lines themselves are not reported. Again this indicates that representations of considerable sophistication are constructed in the absence of awareness (and thus presumably in the absence of attention).

ATTENTIONAL CONTROL

Whereas much of the research on visual attention has centered on the nature of the mechanisms involved (e.g., their speed, or what they select), an equally important issue concerns the way in which these mechanisms are controlled. Two types of control appear to exist, namely goal-driven *direction* and stimulus-driven *capture*. These behave quite differently, with direction corresponding to a slow, sustained process, and capture being faster and more transient. It has been argued that these mechanisms control different types of attentional process, direction being involved with selective access, and capture being involved with selective integration (Briand and Klein, 1987).

Attentional direction occurs, for example, when observers in an orienting task engage their attention on the stimulus to which an arrow is pointing. It also occurs in guided search, where observers select the features to be enhanced or suppressed. In both cases, observers voluntarily select the appropriate locations or features in order to facilitate performance, and they are able to refrain from this if performance is adversely affected (e.g., if cues are misleading).

In contrast, attentional capture is largely involuntary, and interference can result from any features consistent with an *attentional control setting*. For example, visual search for a unique orientation is impaired if one of the distractors has a unique color, even though color is irrelevant to the task. In this case, capture occurs because the control setting can be set only for a *difference* in some feature; the relevant type of difference (e.g., orientation) cannot be represented. Indeed, a unique feature of any kind will capture attention when this control setting has been selected.

The sudden appearance of an item has a privileged status, in that capture can occur regardless of the setting. Interestingly, the relevant factor is not the motion signal that accompanies the appearance, but rather the appearance of the object itself (Yantis and Hillstrom, 1994). However, this type of capture has its limits, in that it appears to be effective only if attention is not already engaged on some task.

RELATIONSHIP TO CONSCIOUSNESS

Although recent research views attention primarily in terms of selection, an older tradition (stemming from William James) viewed it primarily in terms of conscious perception. Traces of this older tradition persist in two functions that are currently ascribed to attention, namely selective access or integration for a process that eventually affects conscious perception, and selective entry into consciousness itself. With regard to the first of these, the relationship between attention and consciousness is unproblematic for a process that is under conscious control. However, there is increasing evidence that many actions are performed without

any involvement of consciousness. Given that such actions require selection for effective operation (e.g., manual grasping may need to select items with a horizontal orientation), there may be selective mechanisms acting on processes that never involve consciousness, at least in terms of immediate control. Whether these mechanisms should be described as 'attentional' would seem to be a matter of convention. (*See* **Motor Control and Learning**)

With regard to the entry of stimuli into consciousness, there may well be selective processes with exactly this function. For example, change blindness, the attentional blink, and inattentional blindness are all phenomena in which the failure to report an otherwise highly visible stimulus could be attributed to a failure of consciousness to access the appropriate representation. However, it is not yet clear whether this is actually the case – although a failure to report could be due to a failure to consciously see a stimulus, it could also be due to a failure to remember it. It is also not yet clear whether conscious experience necessarily arises via attention – non-attentional processes may exist that can provide conscious experience of at least some aspects of the visual field. As with many of the issues relating to attention, a clearer understanding of these matters must await future developments.

References

Briand KA and Klein R (1987) Is Posner's bean the same as Treisman's glue? On the relation between visual orienting and feature integration theory. *Journal of Experimental Psychology: Human Perception and Performance* **13**: 228–241.

Duncan J (1984) Selective attention and the organization of visual information. *Journal of Experimental Psychology: General* **113**: 501–517.

Kahneman D, Treisman A and Gibbs B (1992) The reviewing of object files: object-specific integration of information. *Cognitive Psychology* **24**: 175–219.

Mack A and Rock I (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

Posner MI, Snyder CR and Davidson BJ (1980) Attention and the detection of signals. *Journal of Experimental Psychology: General* **109**: 160–174.

Rensink RA (2000) Seeing, sensing and scrutinizing. *Vision Research* **40**: 1469–1487.

Shapiro K, Driver J, Ward R and Sorensen RE (1997) Priming from the attentional blink: a failure to extract visual tokens but not visual types. *Psychological Science* **8**: 95–100.

Simons DJ and Chabris CF (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* **28**: 1059–1074.

Wolfe JM, Klempen N and Dahlen K (2000) Postattentive vision. *Journal of Experimental Psychology: Human Perception and Performance* **26**: 693–716.

Yantis S and Hillstrom AP (1994) Stimulus-driven attentional capture: evidence from equiluminant visual objects. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 95–107.

Further Reading

Allport A (1992) Attention and control: have we been asking the wrong questions? A critical review of twenty-five years. In: Meyer DE and Kornblum S (eds) *Attention and Performance XIV*, pp. 183–218. Cambridge, MA: MIT Press.

Braun J and Sagi D (1990) Vision outside the focus of attention. *Perception and Psychophysics* **48**: 45–58.

Cave KR and Bichot NP (1999) Visuospatial attention: beyond a spotlight model. *Psychonomic Bulletin and Review* **6**: 204–223.

He S, Cavanagh P and Intriligator J (1997) Attentional resolution. *Trends in Cognitive Sciences* **1**: 115–121.

Laberge D (1995) *Attentional Processing: the Brain's Art of Mindfulness*. Cambridge, MA: Harvard University Press.

Pashler HE (ed.) (1998) *Attention*. Philadelphia, PA: Taylor & Francis.

Simons DJ (ed.) (2000) *Change Blindness and Visual Memory*. Hove, UK: Psychology Press.

Treisman A (1985) Preattentive processing in vision. *Computer Vision, Graphics and Image Processing* **31**: 156–177.

van der Heijden AHC and Bem S (1997) Successive approximations to an adequate model of attention. *Consciousness and Cognition* **6**: 413–428.

Wright RD (ed.) (1998) *Visual Attention*. New York, NY: Oxford University Press.

Visual Evoked Potentials

Intermediate article

Brian F O'Donnell, Indiana University, Bloomington, Indiana, USA

CONTENTS

Visual evoked potentials are deflections in the electroencephalogram which are evoked by a visual stimulus, time-locked to stimulus onset, and generated by neural activity.

INTRODUCTION

Since the development of experimental psychology in the late 1800s, behavioral measures such as reaction time and accuracy have been used to test models of perceptual, cognitive and motor processes. While behavioral measures have proved highly successful in testing models, they have several limitations. First, the processing of stimuli or events that do not elicit behavioral responses is difficult to characterize. Second, the timing of mental processes between stimulus and response can only be indirectly measured. Visual evoked potentials (VEPs) provide a direct, noninvasive measure of neurophysiological activation which can complement behavioral measures in the study of cognitive mechanisms. Because of the high temporal resolution of VEPs (milliseconds), they are particularly well suited for investigation of the temporal dynamics of cognition.

EVOKED POTENTIALS

In 1929 Hans Berger reported that electrical activity from the human brain could be noninvasively recorded using scalp electrodes, and that this activity could be altered by sensory stimulation. This recording was called the electroencephalogram (EEG), popularly known as 'brain waves'. The EEG represents the synchronized activity of several separate populations of neurons which is conducted to the scalp. Stimulation may result in subsequent positive and negative deflections in the EEG, referred to as evoked potentials (EPs) or event-related potentials. Magnetoencephalography (MEG) has also been used to obtain measures of event-related neuromagnetic activity. Because of their smallness relative to the background EEG, evoked potentials are often averaged together over several stimulus presentations or trials to increase the signal-to-noise ratio. This technique, called signal or transient averaging, was applied by Dawson in 1951. It results in noise reduction that is proportional to the square root of the number of trials, and can resolve signals in the microvolt range. Signal averaging requires the assumption that the EP signal is stationary from trial to trial. Other signal processing techniques, such as adaptive filters or response locked averaging, can be used when the latency of the EP signal varies from trial to trial. (*See* **Electroencephalography (EEG)**)

TRANSIENT VERSUS STEADY STATE VEPS

Visual evoked potentials have been used since the 1960s to investigate sensory and cognitive processing. Unlike reaction time or accuracy, VEPs are spatiotemporal measures which vary over time and the surface of the head. Transient VEPs are elicited by a single stimulus, and consist of a series of deflections in the EEG which usually return to baseline. Steady state VEPs are elicited by synchronization of the EEG to the temporal frequency of a periodic visual stimulus, such as a flickering pattern, producing a repetitive waveform in phase with the stimulus frequency.

A wide variety of signal processing techniques have been used to measure VEP phenomena, which can be broadly categorized into time and frequency domain measures (Regan, 1989). Transient VEPs are usually analyzed in the time domain: measures are obtained from the averaged transient waveform elicited by a stimulus onset or other time-locked event, usually displayed with time on the x axis

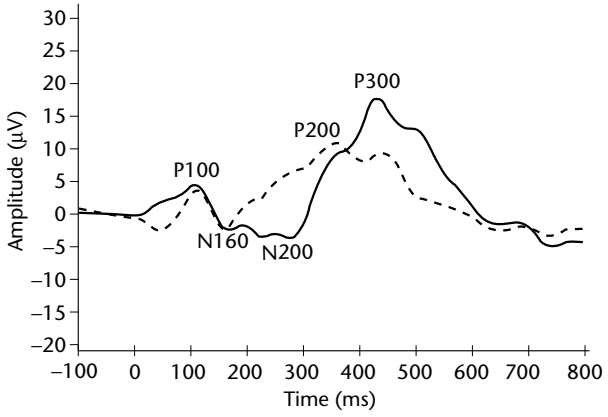


Figure 1. Transient visual evoked potentials elicited by frequently presented nontarget line segments (solid line) and infrequently presented (probability 0.15) target line segments (dashed line). The segments differed in orientation, and the participant pressed a button to the target segments. The frequent, nontarget stimuli elicit a negative deflection about 160 ms after stimulus onset (N160) and a positive deflection that follows it (P200). The target stimuli elicit a prolonged, negative shift in the event-related potential, sometimes called the selection negativity, followed by a large positive deflection, the P300 component. Because a difficult discrimination was used in this paradigm, the P200, N200, and P300 peaks are delayed in time of appearance.

and voltage on the *y* axis (Figure 1). The simplest approach to measuring the waveform is 'peak–picking', in which the latency and amplitude of a peak voltage is measured. Because the peak voltage in a given latency range may vary from recording site to site, average voltages over a latency window may be used. The dependent measures for the transient VEP are latency, amplitude (voltage) and topography (variations in voltage across recording sites on the scalp). Steady state VEPs are usually analyzed in the frequency domain, and have been extensively used in studies of sensory processes. These techniques usually require transforming the time domain measures into the frequency domain using Fourier analysis. The Fourier analysis transforms a time domain waveform into a sum of sinusoidal waveforms differing in power and phase. The power spectrum, which displays power in a segment of EEG as a function of frequency, is shown in Figure 2 for a steady state waveform.

COMPONENTS

Cognitive scientists often describe a transient VEP waveform in terms of 'components' associated with a specific polarity, latency and scalp topography. Components are typically labeled in terms of their polarity and latency relative to a sensory or motor event. Polarity is indicated by 'P' for positive and 'N' for negative. Latency is indicated in milliseconds. For example, 'N400' indicates a component with negative polarity which occurs about 400 ms after stimulus onset. Other techniques such as principal components analysis or independent component analysis have been used to identify components on the basis of covariation of voltage among recording sites or conditions. Because the characteristics of VEP components are highly dependent on stimulus properties and task demands, components are usually implicitly or explicitly associated with specific types of experimental protocols. In the next section, VEP components and methods are described which have been applied to the investigation of visual processing in humans. These include the characterization of mechanisms underlying perception, selective attention, language comprehension, recognition memory, and perceptual integration. In neurological disorders, VEPs are highly sensitive to lesions that affect conduction along the retinal-striate pathway, such as neuritis of the optic nerve or multiple sclerosis. In addition, VEP components are sensitive to illnesses that affect visual attention, such as neurodegenerative illness or schizophrenia.

VISUAL CHANNELS

Psychophysical studies suggest that early stage vision entails the operation of channels tuned to such simple stimulus features as spatial frequency, motion, contrast and color. Both psychophysical and cellular evidence suggests the existence of two major visual channels or pathways in humans. The magnocellular, transient, or broadband system is sensitive to contrast, high temporal frequencies, and low spatial frequencies. The parvocellular, sustained, or color opponent channel is sensitive to color, low temporal frequencies and high spatial frequencies. (*See* **Vision: Form Perception**)

Steady state and transient VEP studies have provided neurophysiological support for multiple visual channels in humans. For example, studies using flickering gratings indicate that low spatial frequencies yield the largest response amplitudes at high temporal frequencies, while high spatial frequencies show the largest amplitudes at low temporal frequencies (Regan, 1989). Transient EPs to patterns modulated at low temporal frequencies (about 1 Hz) have been shown to be sensitive to luminance, contrast, visual field location, and possibly color. Selective attention to particular

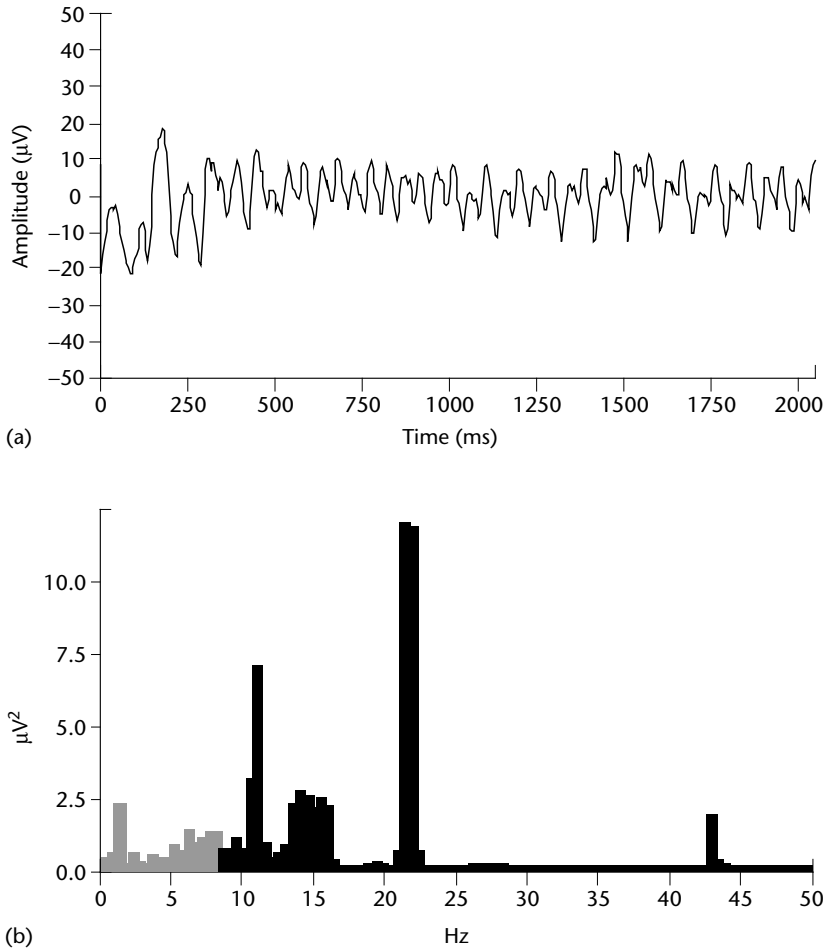


Figure 2. Time and frequency domain representations of a visual evoked potential (VEP). (a) Averaged VEP elicited by a flickering white and black patch, presented at a flicker rate of 21 Hz. Notice the synchronization of the VEP with the oscillating stimulus. (b) A Fourier transform was applied to the time domain data to obtain a power spectrum, which shows peak power at the stimulation rate (21 Hz).

stimulus features such as orientation, size, location, and color can have marked effects on VEP components, as described below (O'Donnell *et al.*, 1997; Luck and Hillyard, 2000).

ATTENTION

Visual EPs are highly sensitive to attentional demands. Most VEP studies of attentional mechanisms have characterized VEP components in the transient, or time, domain (O'Donnell *et al.*, 1997; Luck and Hillyard, 2000). Typical VEPs to nontarget and target stimuli in a discrimination task are shown in Figure 1. Selective attention effects may be apparent in the VEP as early as 80 ms after stimulus onset, and continue to affect processing after execution of a response. From an electrophysiological standpoint, then, selective attention is neither an early nor a late process, but subsumes a variety of mechanisms which span the perceptual, semantic and working memory domains. (*See* **Attention, Models of**; **Attention, Neural Basis of**; **Visual Attention**)

Spatial attention towards a specific visual region enhances the P100 and N160 components to stimuli presented in that region. This effect is largest over posterior electrode sites contralateral to the attended location. Under conditions of high stimulation rate and task demands, P100 and N160 show a graded diminution of amplitude as the target becomes more distant from the attended region. Since P100 effects may be detected to both target and nontarget stimuli, and probably occur prior to stimulus identification, it may reflect spatial-filtering mechanisms engaged before categorization of stimuli.

Target stimuli in a discrimination task usually elicit two long-latency components, the N200 and

P300 components, particularly to low-probability, novel stimuli (Donchin and Coles, 1988). Both N200 and P300 appear to index operations involved in selective attention. The amplitude of P300 is inversely proportional to global stimulus probability in a sequence, as well as the sequence of immediately preceding stimuli. Both N200 and P300 can be elicited by omitted stimuli in a sequence, indicating that these components can arise from internal representations or operations in the absence of a stimulus. The N200 and P300 latencies are proportional to difficulty of discrimination and classification. Both P300 amplitude and latency are sensitive to the mental workload imposed in a dual-task paradigm, suggesting that it may provide a physiological measure of the attentional resources required by a task. The P300 amplitude is usually maximal over the parietal region, with a symmetric distribution for both auditory and visual stimulation, while N200 voltage topography varies with visual task demands and stimulus properties. These findings suggest that the N200 component reflects activation of cortical tissue directly involved in the representation and comparison of a sensory input and the working memory representation of a target stimulus; P300, on the other hand, is much less sensitive to the physical characteristics of stimuli, and may represent a higher-order process which is minimally influenced by perceptual representation, such as expectancies, context updating, and stimulus meaning.

LANGUAGE

Event-related potential components have been used to probe the temporal course of semantic and syntactic processes. In 1980, Kutas and Hillyard (Kutas and Van Petten, 1994) showed that a large negative deflection occurred when an incongruent word terminated a sentence read by a subject. For example, the sentence 'He spread the warm bread with —' could be terminated with a word that is congruent with the preceding context, such as 'butter', or instead with an incongruent word, such as 'socks'. The incongruent word elicited a negative deflection at about 400 ms, the N400 component, which was smaller or absent in the VEP to the congruent or appropriate word. This effect was not observed when a word differed in font size, but was semantically appropriate. The N400 component can also be evoked in word-pair priming paradigms; the deflection is larger to the second word of a pair if the words are semantically unrelated, such as 'doctor – goose', than in the case when the second word of the pair is related to the first, such as 'doctor – nurse'. The amplitude of the N400 component is inversely related to the predictability of the word from the preceding context. Within a sensible sentence, N400 amplitude is larger for words at initial positions than at later positions, again suggesting that it directly reflects contextual constraint. At early sentence positions, N400 amplitude is also affected by word frequency, in that low-frequency words produce larger N400 amplitudes than high-frequency words. Word priming effects have suggested that N400 indexes automatic spread of activation through semantic networks. In terms of sentence processing, it may also reflect specific mechanisms in working memory which integrate a new word into the preceding context. The N400 component is often followed by a large positive component at about 600 ms, the P600 component. The functional significance of the P600 component is still unclear, but it may be more sensitive to syntactic than semantic anomalies (Kutas and Van Petten, 1994; Osterhout *et al*., 1997). (*See* **Syntax and Semantics, Neural Basis of**; **Natural Language Processing, Disambiguation in**; **Lexical Ambiguity Resolution**)

MEMORY

Visual evoked potentials are highly sensitive to the occurrence of novel stimuli and target stimuli in a sequence, showing sensitivity to mechanisms involved in recognition memory. Evoked potentials may also reflect encoding processes (Rugg, 1995). Visual EPs to words to be subsequently recalled or recognized show greater positivity than to words that are not subsequently remembered. This effect usually occurs after several hundred milliseconds of processing. Repetition effects have also been noted, although the relative contribution of implicit versus explicit memory in these effects is still under investigation. Words that are repeated in a series of words show a more positive VEP than those that are seen for the first time, similar to repetition priming observed on behavioral measures.

NEURAL SYNCHRONIZATION

Cellular studies suggest that neural activity in the gamma frequency range (> 30 Hz) reflects the synchronization of neural assemblies involved in 'binding' or integration of various features of an object within a sensory modality, across modalities, and across time. Both EEG and MEG methods have been used to measure human neural synchronization during perceptual and cognitive processing. In experimentally induced binocular rivalry, for

example, MEG coherence across widely separated sensor sites was largest to the temporal frequency of the stimulus which was consciously perceived (Srinivasan *et al*., 1999). While still controversial, this finding supports the hypothesis that neural synchronization may be an important mechanism for late stage integration of percepts. (*See* **Neural Oscillations**; **Gamma Oscillations in Humans**)

References

Donchin E and Coles MGH (1988) Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences* **11**: 357–373.

Kutas M and Van Petten C (1994) Psycholinguistics electrified: event-related brain potential investigations. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*, pp. 83–144. New York: Academic Press.

Luck SJ and Hillyard SA (2000) The operation of selective attention at multiple stages of processing: evidence from human and monkey electrophysiology. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn. Cambridge, MA: MIT Press.

O'Donnell BF, Swearer JM, Smith LT, Hokama H and McCarley RW (1997) A topographic study of ERPs elicited by visual feature discrimination. *Brain Topography* **10**: 1–11.

Osterhout L, McLaughlin J and Bersick M (1997) Event related potentials and human language. *Trends in Cognitive Neurosciences* **1**: 203–209.

Regan D (1989) *Human Brain Electrophysiology*. New York: Elsevier.

Rugg MD (1995) ERP studies of memory. In: Rugg MD and Coles GH (eds) *Electrophysiology of Mind: Event-Related Potentials and Cognition*, pp. 132–170. New York: Oxford University Press.

Srinivasan R, Russell DP, Edelman GM and Tononi G (1999) Increased synchronization of neuromagnetic responses during conscious perception. *Journal of Neuroscience* **19**: 5435–5448.

Further Reading

Chiappa KH (ed.) (1997) *Evoked Potentials in Clinical Medicine*, 3rd edn. New York, NY: Lippincott-Raven.

Luck SJ and Vogel EK (1997) The capacity of visual working memory for features and conjunctions. *Nature* **390**: 279–281.

Makeig S, Jung TP, Bell AJ, Ghahremani D and Sejnowski TJ (1997) Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences USA* **94**: 10979–10984.

Näätänen R (1992) *Attention and Brain Function*. Hilldale, NJ: Erlbaum.

Polich J and Kok A (1995) Cognitive and biological determinants of P300: an integrative review. *Biological Psychology* **41**: 103–146.

Rodriguez E, George N, Lachaux JP *et al*. (1999) Perception's shadow: long-distance synchronization of human brain activity. *Nature* **397**: 430–433.

Visual Scene Perception

Introductory article

Helene Intraub, University of Delaware, Newark, Delaware, USA

CONTENTS

When studying a scene, viewers can make as many as three or four eye fixations per second. A single fixation is typically enough to allow comprehension of a scene. What is remembered, however, is not a photographic replica, but a more abstract mental representation that captures the 'gist' and general layout of the scene along with a limited amount of detail.

INTRODUCTION

The visual world exists all around us, but physiological constraints prevent us from seeing it all at once. Eye movements (called saccades) shift the position of the eye as quickly as three or four times per second. Vision is suppressed during each saccade, and then resumes during the next eye fixation (the period of time that the eye pauses to receive visual input). It has been postulated that the mental representation of a scene that is maintained across a saccade ('transsaccadic memory') is surprisingly sparse. This information, in conjunction with expectations about upcoming layout, provides a means for integrating successive views into a coherent representation of a scene. (*See* **Perception: Overview**)

SCENE COMPREHENSION AND ITS INFLUENCE ON OBJECT DETECTION

Perception occurs so quickly and automatically that individuals cannot discern the amount of time they need to understand a scene. Are multiple eye fixations necessary for scene perception, or are we able to perceive the meaning of a scene right away, based on the first eye fixation? One way to address this question is to ask whether viewers can understand unrelated scenes presented in rapid succession at a rate that mimics the rapid pace at which we normally make eye fixations (e.g. three per second). The reason for using unrelated scenes is to allow an assessment of what can be gleaned from a single glimpse without any bias from prior context. Immediately after viewing a rapid sequence like this, viewers participate in a recognition test. The pictures they just viewed are mixed with new pictures and are slowly presented one at a time. The viewers' task is to indicate which pictures they remember seeing before. Following rapid presentation, memory for the previously seen pictures is rather poor. For example, when 16 pictures were presented in this way, moments later, viewers were able to recognize only about 40 to 50 percent in the memory test. One explanation is that they were able to perceive only about 40 to 50 percent of the scenes that had flashed by and those were the scenes they remembered. However, contrary to this plausible explanation, many viewers adamantly claimed that they had indeed momentarily perceived most of the scenes but that the onslaught of new scenes interfered with their ability to remember what they had momentarily grasped. (*See* **Memory Consolidation**; **Memory Distortions and Forgetting**)

To determine whether viewers were good at momentarily grasping the meaning of a scene under these conditions, it was necessary to design a task that could tap into the early stages of scene processing at a point before forgetting is likely to take place. Subjects were shown the same rapid sequences as in the previous experiments, but in this case a brief description of one of the pictures (e.g. 'a road with cars') was provided in advance. They were instructed to look for a picture matching that description and to immediately press a response key as soon as they saw it. Because the detection task required them to respond immediately (without waiting until the end of the sequence), this minimized the likelihood of forgetting. At speeds as fast as three and four pictures per second, the ability to correctly detect a scene based on a description was excellent, usually reaching about 90 percent correct or better. This performance far exceeded the 40 to 50 percent

correct obtained from subjects who took the recognition test. Other experiments were conducted in which the description provided at the outset was nonspecific (e.g. 'a type of furniture'), and viewers were again better at detecting pictures than remembering them. These results indicate that scene perception is very rapid: a single 'fixation' is frequently sufficient to allow identification, even under the demanding conditions of rapid serial visual presentation.

Scenes, however, usually contain multiple objects. There are two ways that the identification process might proceed: (a) first, objects are identified, and then viewers begin to understand what kind of scene they are looking at, or (b) first, the scene's general meaning and layout are perceived holistically, and then identification of specific objects follows. Although a controversial topic, many studies suggest the second alternative. In a series of experiments, outline drawings of scenes were each briefly presented (e.g. 50–150 milliseconds) one at a time. Prior to the presentation of each scene, viewers were provided with the name of an object (e.g. fire hydrant), and immediately after each presentation a dot appeared on the screen at the location previously filled by one of the objects in the scene. Viewers had to say whether or not the object named at the outset (in this example, a fire hydrant) had been present at that location. In some cases the object did fit with the general meaning of the scene (e.g. a fire hydrant in a street scene) and in others it did not (e.g. a fire hydrant in a kitchen scene). If objects are identified before the scene's meaning is grasped, then the ability to detect the object should be unaffected by whether or not it fits with the context of the scene. However, contrary to this possibility, the object's relation to the scene as a whole did affect the speed and accuracy of the response. Other relations of the object to the scene, such as whether its relative size did or did not fit the scene, or its location in the scene was plausible or implausible, had a similar effect on the response. This suggests that we may grasp the meaning and general layout of a scene rapidly enough to affect our ability to detect specific objects in the scene. (*See* **Vision: Object Recognition**)

A similar conclusion has been drawn from other experiments in which eye movements were monitored. Pictures were presented for relatively long intervals that allowed viewers to make many eye fixations. Prior to the picture's onset, viewers fixated the center of the screen. Frequently, the first saccade brought the eyes to an object that didn't belong, and subjects maintained longer fixation times on that object, as if they were trying to understand it. Again, this suggests that a scene's meaning and layout is understood very rapidly in processing, occurring prior to identification of all its objects. How does this information become integrated with new information obtained from the next eye fixation?

TRANSSACCADIC MEMORY

Initially, many researchers who studied the relation between eye movements and cognition thought that, after each eye fixation on a scene, a detailed sensory record of the fixated region was stored in a very short-term memory system called an 'integrative buffer'. Within the buffer, information from a new fixation would be integrated (i.e. knitted together) with the stored information from the previous fixation. By integrating the details of successive views, the system was thought to provide the viewer with a seamless, detailed perception of the world – like piecing together parts of a jigsaw puzzle. However, research on memory for briefly glimpsed scenes, and research on reading and eye movements, has led to a different proposal about how views are integrated across saccades. The idea is that our perception of a detailed, continuous world is to some degree illusory. Instead it is proposed that 'transsaccadic memory' (memory for information that is maintained across a saccade) includes the scene's meaning, along with the general layout, and only some detail.

Transsaccadic memory allows the visual system to relate the information obtained from one eye fixation to the contents of the next – but not by integrating detailed photograph-like views. It is somewhat surprising to think that our perception of the world is built up from relatively sparse information, but there is a lot of evidence to support this claim. You can get a sense of this yourself if you look at a complex visual scene (e.g. a bookshelf with books of different sizes and colors) and then close your eyes and recount all the details. You will probably find that your visual memory is missing a lot of information. Of course, the claim is best supported by controlled experiments that carefully address the viewer's ability to detect changes.

Various types of change detection experiments have been conducted in which a change is made in a display at exactly the same time that the eyes make a saccade. Using computer technology, a viewer's eyes are monitored, and when a saccade is launched, a change is made before the eyes land again and begin the next fixation. Therefore, the change in the scene takes place while vision is suppressed. When the eyes land, the change has

already occurred. In these situations, many kinds of changes go unnoticed both in scenes and in text. For example, in one case, researchers presented sentences in an AlTeRnAtInG cAsE on a computer screen. Each time the viewer made an eye movement, the case was reversed (capital letters became small and small letters became capitals). This did not disrupt the ability to read, and understand the text. Amazingly, the viewers didn't even notice the changes.

PERCEPTION OF 'GIST' VERSUS DETAILS

Although the research described earlier shows that viewers can rapidly understand the general meaning or 'gist' of a scene, they are remarkably poor at noticing changes from one look to the next: a phenomenon called 'change blindness'. One of the best examples of change blindness uses a presentation technique (the 'flicker' technique) that is very similar to the rapid picture presentation technique described earlier. So it is interesting to make a comparison. (*See* **Object Perception, Neural Basis of**; **Change Blindness**; **Visual Attention**; **Illusions**)

In preparation for the study, the experimenter selects scenes and then uses computer graphics to edit each one, changing a particular feature (e.g. a lamp disappears or changes color; diagonal stripes on a wall are reversed). The original and the altered version are each presented for a brief duration that mimics a single eye fixation. The two versions keep alternating throughout the sequence with a blank interval in between presentations. This interval is meant to simulate the suppression of vision during a saccade. The viewers' task is to identify the change. Even though they can clearly identify the scene every time it appears, they do not notice the change right away. The most difficult cases required more than 80 alternations (more than 50 seconds) before even large changes were detected. What is most interesting is that the changes were actually very easy to see if the viewer received a hint about the location, or just fixated the location. This demonstrates that memory for a scene, just a fraction of a second later, does not provide a detailed, picture-like representation – but does provide sufficient information to allow the viewer to recognize that the same basic scene is being repeated.

BOUNDARY EXTENSION

How does the visual system integrate the relatively sparse information held in transsaccadic memory from one fixation to the next, thus providing a coherent representation of the scene? A phenomenon called 'boundary extension' provides one answer to this perplexing question. It also shows that although the representation of a scene lacks detail, it has an overabundance of another type of information that may serve to facilitate integration of views. After looking at photographs for as long as 30 seconds each, people tend to make an interesting error. They remember having seen beyond the edges of the picture! They remember seeing information that was *not* in the picture but that was likely to have existed just outside the camera's range of view.

This overinclusive memory is so convincing, that when they see the same photographs again (in a recognition test) they reject them as being the same: they claim that the test picture doesn't show as much of the scene as had the 'original' picture they studied earlier. Boundary extension can also be seen when other viewers draw the photographs from memory. The first two pictures in Figure 1 (panels (a) and (b)) show a close-up view of a scene and a viewer's drawing from memory. Notice that the drawing extends the boundaries of the view. Although the trashcans, fence, and lid are all cropped by the photograph's edges, the subject remembered seeing them as whole, and also

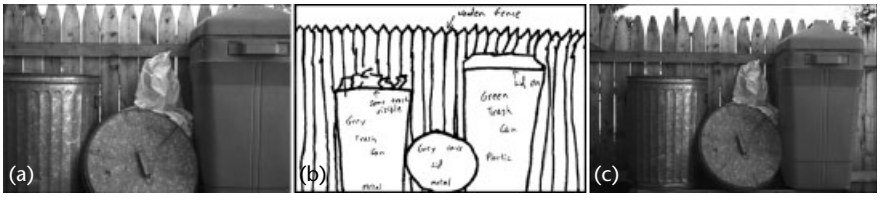


Figure 1. (a) A photographic close-up of a scene, (b) a viewer's drawing of the close-up from memory (note the extended boundaries), and (c) a more wide-angle photograph of the same scene. (Based on H Intraub and M Richardson (1989) Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory and Cognition* **15**: 179–187).

remembered seeing parts of the fence on the left and right of the scene, some sky above the fence, and more of the scene at the bottom. It is tempting to think of this as just an error, but if you look at a wider-angle photograph of the same scene in panel (c), you will see that the viewer's 'error' actually provides an excellent prediction of what really did exist outside the boundaries of the close-up. The drawing looks more like the wide-angle view than the close-up that the subject had actually studied. This effect appears to be the rule rather than the exception in memory for scenes (particularly close-ups). In one experiment it occurred in 95 per cent of the 133 drawings made by 20 different people. Other research showed that even when people tried hard not to make this mistake, they couldn't prevent it from happening.

Boundary extension reveals the remarkable ability of the visual system to predict the continuation of scene layout and may serve to facilitate the integration of views by (a) 'priming' the visual system to see the upcoming layout, and (b) placing the view within a larger context. To determine whether boundary extension would occur rapidly enough to aid integration of views during visual scanning, rapid sequences of pictures were again used in an experiment. Subjects viewed three pictures in rapid succession, and one second later, one picture was repeated and remained on the screen. Viewers rated the picture on a 5-point scale to indicate whether the view was the same, was more wide-angle, or more close up than the view they had seen a second earlier. Subjects tended to rate the repeated picture as showing less of the scene than the 'original' picture – thus indicating that they had remembered it as showing a more wide-angle view. The visual system very rapidly extrapolated the picture's layout – so that the viewer remembered having seen more of the scene than they actually did. Recent research has shown that the same anticipatory error occurs in memory for real three-dimensional scenes that are viewed through a window. People remember seeing expected information from just outside the view. (*See* **Imagery**; **Memory Distortions and Forgetting**; **Vision: Occlusion, Illusory Contours and 'Filling-in'**)

Taken together, experiments on scene perception and memory suggest that the visual system strikes a compromise between the speed of scene comprehension and the amount of detail that will subsequently be stored. Rapid serial visual presentation of pictures in conjunction with detection tasks shows how quickly complex scenes can be identified, whereas change detection research suggests that viewers cannot retain all the details in a scene – even a fraction of a second later. Although memory from one fixation to the next is less detailed than a photograph, this 'sparseness' may aid integration of successive views as the eyes visually scan a scene. Indeed, it has been argued that we don't need to have a detailed visual memory because if we want to see a detail in our environment, we can rapidly fixate the region in question – an act requiring only a fraction of a second. Research on boundary memory indicates that the exact location of the borders of each view is not retained. The visual system apparently systematically extrapolates beyond those borders, so that the viewer remembers not only the layout that was actually seen but also the expected layout just beyond the edges of the view. This combination of rapid identification, limited detail retention and predictive extrapolation would seem to allow considerable economy in the way we perceive and remember the visual world. (*See* **Spatial Cognition, Models of**; **Social Processes, Computational Models of**; **Vision: Top-down Effects**)

Further Reading

Biederman I (1981) On the semantics of a glance at a scene. In: Kubovy M and Pomerantz JR (eds) *Perceptual Organization*, pp. 213–253. Hillsdale, NJ: Lawrence Erlbaum.

Henderson JM and Hollingworth A (1999) High-level scene perception. *Annual Review of Psychology* **50**: 243–271.

Intraub H (1999) Understanding and remembering briefly glimpsed pictures: implications for visual scanning and memory. In: Coltheart V (ed.) *Fleeting Memories*, pp. 47–70. Cambridge, MA: MIT Press.

Intraub H (2002) Anticipatory spatial representation of natural scenes: momentum without movement? *Visual Cognition* **9**: 93–119.

Irwin DE (1991) Information integration across saccadic eye movements. *Cognitive Psychology* **23**: 420–456.

McConkie GW and Zola D (1979) Is visual information integrated across successive fixations in reading? *Perception & Psychophysics* **25**: 221–224.

O'Regan JK, Rensink RA and Clark JJ (1999) Change blindness as a result of 'mudsplashes'. *Nature* **398**: 34.

Potter MC (1999) Understanding sentences and scenes: the role of conceptual short-term memory. In: Coltheart V (ed.) *Fleeting Memories*, pp. 13–46. Cambridge, MA: MIT Press.

Simons DJ and Levin DT (1997) Change blindness. *Trends in Cognitive Science* **1**: 261–267.

Wolfe JM (1999) Inattentional amnesia. In: Coltheart V (ed.) *Fleeting Memories*, pp. 71–94. Cambridge, MA: MIT Press.

Williams Syndrome

Introductory article

Howard M Lenhoff, University of California, Irvine, California, USA

CONTENTS

Williams syndrome is a rare genetic condition caused by the absence of a small portion (containing about 20 genes) from one copy of chromosome 7. Individuals with Williams syndrome have a mean IQ of 55 and exhibit a strange array of impairments and behaviors. Their abilities in language and music intrigue both cognitive scientists and geneticists.

INTRODUCTION

Individuals having the neurodevelopmental congenital condition called Williams syndrome (WS) share a number of features, including 'elfin' facial traits, talkative and friendly behavior, an average intelligence quotient (IQ) of 55, and a number of physiological and motor impairments. Williams syndrome is presumed to be present in 1 out of every 20 000 births. This syndrome is of special interest to cognitive scientists for a number of reasons. First, in addition to their cognitive and physical defects, people with WS also possess a range of unusual abilities, and even talents. Unlike most populations of cognitively disabled people, those with WS show particular strengths in using language, in recognizing and discriminating among faces, and in their proclivity for and abilities in music. Second, people with WS are genetically homogeneous, exhibiting a common loss of a particular group of genes from one strand of chromosome 7. In this way they differ from other atypical populations who show unusual abilities, such as people with autism, who are poorly defined genetically. Third, new research on the presence and relatively high incidence of absolute (perfect) pitch among people with WS shows promise of being useful in understanding links between musical abilities and the acquisition of language by young children, and the relationship of specific genes to those behaviors.

Scientific knowledge of WS, first described in 1961, is relatively new. In 1993 we learned that most individuals affected by WS have a common microdeletion of about 20 genes from the band region q11.23 of chromosome 7. Three of the missing genes (*LIMK1*, *FZD3* and *WSCR1*) are active in the normal brain – a sign that they could influence brain development and function. In accord with the findings of the human genome projects, it would seem that the unusual behaviors found in people with WS are polygenic in origin, i.e. they are influenced by many genes during the development of the brain and related structures.

SYMPTOMS

People with WS tend to look like one another, especially in the years before puberty. Examination of their faces shows eyes that appear puffy and relatively close together. They have a small, upturned 'pug' nose, a wide mouth with full lips, and a small chin with receding jaw (Figure 1). At birth many of these babies possess minor to severe cardiovascular problems, some requiring surgery. Other common medical problems in infancy are difficulty with feeding and severe stomach pains, sleeplessness, and (in some cases) high levels of blood calcium. They experience delays in motor development, and begin walking at about 21 months, usually with an awkward gait. As the children grow older they show further delayed physical and mental development. Many of the childhood physical problems gradually become less severe, whereas musculoskeletal problems become more noticeable. Problems with connective tissue, such as diverticuli of the bladder or colon, may develop later in life. Their hair turns gray and their skin ages prematurely.

PERSONALITY TRAITS AND TALENTS

Cognitive Profile

Most people with WS share a number of personality and behavioral traits. Their average IQ is 55,

Figure 1. Typical facial characteristics of children with Williams syndrome.

some scoring 80. They are generally weak in basic academic and cognitive skills such as visual closure and spatial tasks. They have short attention spans (although not to music), are overly anxious, and have poor relationships with peers who do not have WS. On the other hand, they are generally viewed as friendly, outgoing and loquacious; they have been documented as good storytellers and as possessing good language abilities compared with others of similar IQ. They also are reported to have good visual recall skills and to perform well on tasks of face recognition and on those requiring recall of verbal material. Most possess the trait of hyperacusis, an extreme sensitivity to loud sounds, which they find painful.

In the early 1980s researchers compared the performances of people with WS with that of the general population and of people with Down syndrome. People with WS, in contrast to their generally weak performance on overall tests of cognitive ability, commonly used well-formed grammar in their spontaneous speech. They performed significantly better than those with Down syndrome on tasks of grammatical comprehension and production. They also had vocabularies larger than expected for their mental age: when asked to list some animals, for example, they often chose exotic examples rather than simply listing monosyllabic names such as 'cat' or 'cow'. They are also said to have strengths in prosody, i.e. the meter and tone of speech that carries meaning.

Children with WS seem more expressive than other children. When asked to provide a story for a series of pictures, as they told their tale, they often altered their pitch, volume, length of words or rhythm to enhance the emotional tone of the story. Unfortunately, these traits often mislead teachers and counselors into thinking that the children have better reasoning skills than they actually possess; thus, some do not receive the academic support they need.

People with WS do poorly on visual processing tasks, such as copying drawings. They fail, however, in different ways from people with Down syndrome, suggesting that the deficits in the two groups stem from different abnormalities in brain anatomy or processing. People with WS give more attention to details of images but fail to recognize overall patterns, whereas people with Down syndrome are more likely to perceive the global organization but overlook details. Individuals with WS also show strengths in recognizing and discriminating among photographs of faces previously unfamiliar to them.

Neurological Studies

Examination of brains by magnetic resonance imaging and at autopsy supports the probability that in Williams syndrome the brain is altered in complicated ways. For example, anatomical changes (such as abnormal clustering of neurons in visual areas) occur, which might lead to deficits in visuospatial abilities. Research linking specific structures in the brain with specific cognitive functions will potentially give us a better understanding of how differences in the brain in WS account for some of the unusual behaviors in this syndrome. In people with WS the brain volume is about 80% that of normal individuals, but we do not know how this deficiency affects specific behaviors.

MUSICAL ABILITIES

The peaks and valleys in the cognitive abilities of individuals with WS constitute a sort of mental asymmetry. Some major peaks, first observed by parents and music teachers, represent the unusual musical interest and abilities of these individuals. Their attention span for listening to and participating in musical activities is surprisingly long. People with WS have strengths in rhythm, many showing excellent skills on the drums. Many are able to retain the lyrics and melodies of complex music (some in a variety of languages) for periods of years. Those who speak and sing in languages other than their own have near-perfect accents; one is known to sing in thirty foreign languages. Most of those examples are anecdotal, however, and have not been extensively studied or quantified by researchers.

Absolute Pitch

Research on absolute pitch – the rare capacity to recognize, name and produce the pitch of a

musical note without a reference pitch – shows promise of elucidating a link between music and language, and of providing insights into how genes affect brain development to account for the unique mental asymmetry in WS. Absolute pitch occurs in 1 in 10 000 normal individuals in Western populations and in higher proportions among some Asian populations. Despite their low IQs, five individuals with WS participating in the research exhibited near-ceiling levels of absolute pitch, scoring, out of 1084 trials, 97.5% correct as a group; well-trained normal musicians with absolute pitch averaged 84.3% correct on similar tests. The exceptional abilities of people with WS support the view that they excel in a major component of 'musical intelligence', as defined by Gardner. That the five participants with WS tested all possessed absolute pitch suggests that its incidence in the WS population may be greater than that found in the general population of the Western world. The number of people with WS in North America on record is 4500; even if none of the others had absolute pitch, its incidence in this population in North America would be about 1 in 1000, which is still 10 times greater than the general population.

Critical Period

Cognitive scientists believe that in developing brains of normal individuals there is a critical period, a 'window of opportunity', for individuals to possess absolute pitch in adulthood. During that period, which covers infancy through age 6 years, normal people need to have had intense musical training. In contrast, most people with WS to date did not begin their study of music until later in life (including four of those in the pitch study) and their musical training is usually not rigorous because most do not read musical notes or name them. In WS the genes affecting brain development may damage not only those parts of the brain dealing with normal cognitive function, but also the brain mechanism for closing the window allowing for absolute pitch in adulthood. This change in brain development may also account for some of their other remarkable musical abilities.

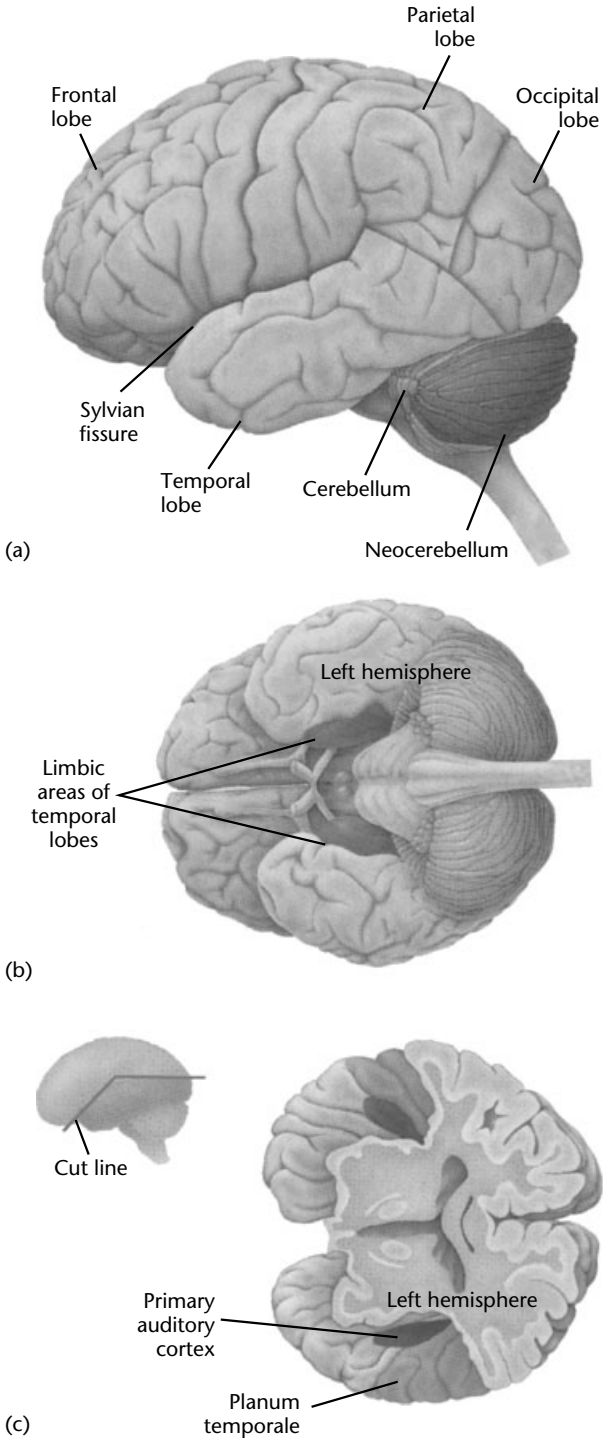


Figure 2. [*Figure is also reproduced in color section.*] Anatomy of the brain. (a) Side of brain; (b) underside of brain; (c) cut in plane of sylvian fissure.

Brain Anatomy

Preliminary anatomical analyses of the brain have identified features that could help explain the presence of musical talent in this syndrome. The primary auditory cortex (located in the temporal lobe) and an adjacent auditory region, the left planum temporale, thought to be important to language as well as to musicality, are relatively enlarged in musicians having absolute pitch. This

left–right asymmetry of the planum temporale is also observed in the few WS brains examined so far (Figure 2).

Much research on correlating parts of the brain with processes of learning language and music has studied sporadic cases of individuals whose brains have been damaged in specific regions by an accident, stroke or surgery. People with WS, whose brains have been specifically damaged in the neurodevelopmental processes affected by the lack of specific genes in chromosome 7, may offer cognitive scientists a different method to pinpoint areas of brain function in cognition.

Pitch and Language

What is the evolutionary advantage to the species of having absolute pitch? A possible answer comes from two different lines of research: one shows that infants are more attracted to absolute pitch than to relative pitch, whereas the opposite is true of adults; the other study shows that absolute pitch is relatively high in peoples who speak tonal languages, such as Mandarin and Vietnamese, where the same word can have different meanings dependent upon the tone used. Because a third of the world's people speak in tonal languages, researchers propose that absolute pitch allows humans to acquire languages and accents. Because the critical period for the development of absolute pitch appears to remain open in WS, adults with this syndrome may also offer opportunities for understanding the brain processes involved in the normal acquisition of language.

CONCLUSION

Research into the cognitive processes in people with WS may help elucidate the neuronal mechanisms by which the brain is involved in the acquisition of language. It is also making investigators see individuals with learning disabilities in a new light. Close study of WS has shown that low IQ scores can mask the existence of such capacities as the possession of absolute pitch. It also signals that other such individuals could have untapped potentials waiting to be uncovered if only researchers and society would take the trouble to look for and cultivate them.

Further Reading

Bellugi U, Lichtenberger L, Mills D, Galaburda A and Korenberg J (1999) Bridging cognition, the brain and molecular genetics: evidence from Williams syndrome. *Trends in Neurosciences* **22**: 197–207.

Lenhoff HM (2002) Williams syndrome. In: *Encyclopedia of the Life Sciences*, vol. 19, pp. 540–547. London: Macmillan. [http://www.els.net]

Lenhoff HM, Perales O and Hickok G (2001) Absolute pitch in Williams syndrome. *Music Perception* **18**: 491–503.

Lenhoff HM, Wang PP, Greenberg F and Bellugi U (1997) Williams syndrome and the brain. *Scientific American* **26**: 42–47.

Udwin O, Davies M and Howlin P (1996) A longitudinal study of cognitive abilities and educational attainment in Williams syndrome. *Developmental Medicine and Child Neurology* **38**: 1020–1029.

Word Learning

Intermediate article

Paul Bloom, Yale University, New Haven, Connecticut, USA
Erika Nurmsoo, Yale University, New Haven, Connecticut, USA

CONTENTS

The ability to learn the meanings of words is central to the development of language, and requires rich mental capacities – linguistic, conceptual, and social – that interact in complicated ways.

THE PROBLEM OF WORD LEARNING

Word learning is a remarkable feat. The communication systems of nonhumans contain nothing akin to words, and attempts to teach words to primates in captivity have met with, at best, limited success – even by the most enthusiastic estimates, none of these primates has attained the vocabulary size of a normal two-year-old human. And there is as yet no computer or robot that can do something as apparently simple as learning a word like 'dog' or 'mommy'. The unique power of humans to learn words has profound consequences, since word learning is obviously essential to language learning, and, more controversially, may play an important role in cognitive development.

What makes word learning so difficult? Developmental psychologists often cite an example by Quine (1960), who imagines a linguist regarding a rabbit and hearing it called 'Gavagai'. What does *Gavagai* mean? Quine notes that there exists an infinity of possible interpretations of this new word: it could refer to rabbits, to mammals, to animals; to tails or legs; to whiteness or furriness; to running or moving. It could, from a logical point of view, refer to the top half of the rabbit, or to its outer surface, or even to undetached rabbit parts. Since all of these interpretations are consistent with how the word is used, how does the linguist – or the child – ever learn what this word, or any word, means?

The answer is that the hypothesis space of the word learner must be somehow constrained; the learner must be biased, either innately or through prior experience, to favor some interpretations over others. Much research in word learning concerns the nature of these constraints and how they emerge in the course of development.

In fact, Quine's example actually understates the difficulty of word learning, as it assumes that the linguist is regarding the rabbit at the moment the word is being spoken. Words are not usually presented in such 'transparent' circumstances. In some cultures, there is no explicit labeling of objects, and even in Western societies, the mapping between a word and what it refers to is far from obvious. Gleitman (1990) argues that this is particularly the case for verbs. Most of the time, for instance, that the verb 'opening' is spoken, nothing is being opened, and most of the time something is being opened, 'opening' is not spoken – and yet children have no special problem realizing that 'opening' means opening. Plato was perhaps the first to point out the even more serious puzzles that arise when one considers the learning of abstract terms, such as those for numbers or ideal geometrical forms.

FIRST WORDS AND THE TIME COURSE OF WORD LEARNING

Despite the problems, children begin to understand the meanings of some words as early as nine months of age, and start to produce words on or soon after their first birthday. Early words, regardless of the culture, consist mainly of names for important individuals ('Mommy'), social routines ('bye-bye'), common nouns ('cookie', 'milk'), and hard-to-classify expressions such as 'more', 'up', and the ubiquitous 'no'. Later on, verbs and adjectives start to appear, and, still later, function words such as determiners ('the') and conjunctions ('or') appear. While children sometimes get the precise meaning of a word wrong (calling a cat 'a dog', for instance), serious confusions (calling a cat 'a cookie', say) are virtually nonexistent.

Vocabulary growth starts slowly. One-year-olds learn less than one word a day, and two-year-olds learn about two or three words a day. This rate gradually increases and eventually, in part through the onset of literacy, children come to learn well over 12 new words a day. Early on, there is a correlation between vocabulary size and syntactic knowledge, even with age factored out (Fenson *et al*., 1994), but the reason for this is unclear – it could be because syntactic knowledge facilitates word learning (see below), because word learning facilitates syntactic development, or both.

It is often said that a sudden shift in word learning occurs at about 18 months of age, or when a child has about 50 words in his or her vocabulary; this is often called a 'word spurt' or 'vocabulary spurt'. Surprisingly, however, this might well be a myth; more recent studies find, for the most part, that the increase in word learning speed tends to be gradual, without any periods of sudden acceleration.

CONCEPTUAL CONSTRAINTS AND WORD LEARNING

One key factor in how children learn words involves conceptual constraints. While learning words may sometimes lead to the development of new concepts, there is considerable evidence that much of word learning can be understood as the process of mapping pre-existing concepts onto the sounds and signs that people produce.

Indeed, children can grasp aspects of the meanings of new words with very few exposures, without training or feedback and without ostensive naming – a process called 'fast mapping'. For instance, if a three-year-old hears an object being referred to in passing as a 'koba', over a month later she will tend to remember which object (from a group of ten) received this novel label (Markson and Bloom, 1997). Children under the age of two can fast map new nouns (e.g. Waxman and Markow, 1995), and the meanings of these early acquired words seem to be much the same as they are for adults. This suggests that word learning is supported by a pre-existing conceptual repertoire, one that includes the notion of rabbits, but not the notion of undetached rabbit parts. This position is supported by research with prelinguistic infants showing that they possess a rich understanding of objects, actions, and other ontological kinds.

Specific products of these pre-existing conceptual structures may be the 'whole object bias' and the 'taxonomic bias' (Markman and Hutchinson, 1984). These work together to predispose children to treat new words as referring to kinds of objects – and so 'Gavagai' is naturally thought of as referring to the rabbit, not to the tail, or to the top half of the rabbit. An alternative view, however, is that these biases are the product of innate mechanisms that are dedicated to the task of word learning (e.g. Waxman and Markow, 1995).

SOCIAL KNOWLEDGE AND WORD LEARNING

A second system central to word learning involves children's appreciation of the mental states of other people, what is sometimes called 'naive psychology' or 'theory of mind'. There is abundant evidence that children will take a word as referring to a given object if and only if there is evidence that the speaker intended to refer to that object. In one study, children were given one object to play with while another object was put into a bucket that was in front of the experimenter. When a child was looking at the object in front of her, the experimenter looked at the object in the bucket and said a new word, such as 'It's a modi!'. Eighteen-month-olds looked at the experimenter and redirected their attention to what she was looking at, in this case, at the object in the bucket. And when later shown the two objects and asked to 'find the modi', they assumed that the word referred to the object the experimenter was looking at when she said the word – not the object that the child herself was looking at (Baldwin, 1991).

More sophisticated intentional capacities are displayed by 24-month-olds. In one study, an adult announced her intention to find an object – 'Let's find the toma!' – and then picked up and nonverbally rejected (by frowning) two other objects before picking up a third object and smiling. Despite the temporal gap, children inferred that this third object was what 'toma' referred to. In another study, an adult used a novel verb to declare her intention to perform an action (e.g. 'I am going to blork!'), proceeded to carry out an action 'accidentally' (saying 'Whoops!') and then performed another action, with satisfaction (saying 'There!' with a pleased expression). Children connected the verb with the action the speaker seemed satisfied with, not the accidental one.

Such studies indicate that even very young children infer the intention of the speaker (through attention to cues that include line-of-regard and emotional indications of satisfaction) when determining the referent of a new word, for both nouns and verbs (Tomasello and Barton, 1994).

SYNTACTIC CUES AND WORD LEARNING

Young children attend to the syntax of a word when determining what the word means. The classic study showing this was done by Roger Brown (1957), who showed preschoolers a picture of a strange action being performed on a novel substance using an unfamiliar object. One group of children was told: 'Do you know what a sib is? In this picture, you can see a sib' (count noun syntax); a second group was told: 'Have you seen any sib? In this picture, you can see sib' (mass noun syntax); and a third group was told: 'Have you seen sibbing? In this picture, you can see sibbing' (verb syntax). The preschoolers tended to construe the count noun as referring to the object, the mass noun as referring to the substance, and the verb as referring to the action.

Subsequent research has found that syntactic cues can help even younger children learn words. For example, 2-year-olds who hear 'This is ZAV' expect the word to refer to a specific individual, as with a proper name like 'Fred' (Katz *et al*., 1974). Two-year-olds hearing 'John ZAVS Bill' expect the word to have a meaning similar to that of 'hit', while those who hear 'John and Bill ZAV' expect it to have a meaning similar to that of 'stand' (Naigles, 1990). It might well be that a sensitivity to syntactic cues is what guides children to appreciate the meanings of words such as verbs and prepositions, explaining why these words are not present at the very onset of vocabulary development.

References

Baldwin DA (1991) Infants' contribution to the achievement of joint reference. *Child Development* **62**: 875–890.

Brown R (1957) Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology* **55**: 1–5.

Fenson L, Dale PS, Reznick JS *et al*. (1994) Variability in early communicative development. *Monographs of the Society for Research in Child Development* **59**(5, serial no. 242).

Gleitman LR (1990) The structural sources of word meaning. *Language Acquisition* **1**: 3–55.

Katz N, Baker E and Macnamara J (1974) What's in a name? A study of how children learn common and proper names. *Child Development* **45**: 469–473.

Markman EM and Hutchinson JE (1984) Children's sensitivity to constraints on word meaning: taxonomic versus thematic relations. *Cognitive Psychology* **16**: 1–27.

Markson L and Bloom P (1997) Evidence against a dedicated system for word learning in children. *Nature* **385**: 813–815.

Naigles LR (1990) Children use syntax to learn verb meanings. *Journal of Child Language* **17**: 357–374.

Quine WVO (1960) *Word and Object*. Cambridge, MA: MIT Press.

Tomasello M and Barton M (1994) Learning words in non-ostensive contexts. *Developmental Psychology* **30**: 639–650.

Waxman SR and Markow DB (1995) Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology* **29**: 257–302.

Further Reading

Baldwin DA (2000) Interpersonal understanding fuels knowledge acquisition. *Current Directions in Psychological Science* **9**: 40–45.

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Carey S (1978) The child as word-learner. In: Halle M, Bresnan J and Miller GA (eds) *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

Fodor JA (1981) The current status of the innateness controversy. In: Fodor JA (ed.) *Representations*. Cambridge, MA: MIT Press.

Gleitman LR and Gleitman H (1997) What is language made out of? *Lingua* **100**: 29–55.

Golinkoff RM, Mervis CB and Hirsh-Pasek K (1994) Early object labels: the case for a developmental lexical principles framework. *Journal of Child Language* **21**: 125–155.

Hall DG, Waxman SR and Hurwitz WM (1993) How two- and four-year-old children interpret adjectives and count nouns. *Child Development* **64**: 1651–1664.

Macnamara J (1982) *Names for Things: A Study of Human Learning*. Cambridge, MA: MIT Press.

Markman EM (1990) Constraints children place on word meanings. *Cognitive Science* **14**: 57–77.

Woodward AL and Markman EM (2000) Early word learning. In: Damion W, Kuhn D and Siegler R (eds) *Handbook of Child Psychology*, vol. 2, pp. 371–420: *Cognition, Perception, and Language*. New York, NY: John Wiley.

Word Meaning, Psychology of

Introductory article

Paula J Schwanenflugel, University of Georgia, Athens, Georgia, USA
Susan J Parault, University of Georgia, Athens, Georgia, USA

CONTENTS

The psychology of word meaning concerns how words are organized, mentally depicted, and processed cognitively.

INTRODUCTION

The study of the psychology of word meaning has been decidedly interdisciplinary. Linguistics, philosophy, artificial intelligence, anthropology, education, and psychology have all contributed major ideas. Key questions concern how words are organized, mentally depicted, and processed cognitively. The *mental lexicon*, our mental dictionary, is thought to contain upwards of 40,000 words in the average high school graduate. However, the mental lexicon appears to be organized more like a thesaurus than a regular dictionary because it is organized both conceptually as well as alphabetically (or, rather, according to the words' sound characteristics or phonologically, and orthography in literate people). This organization seems to be designed to connect speakers and listeners to the right words and conceptual domains quickly and effortlessly. There are a number of views regarding the characteristics of this mental lexicon.

FEATURE THEORIES

The *feature comparison view* proposes that word meanings are represented as sets of *semantic features*. Semantic features are simple word-meaning parts which taken together serve to comprise the word's meaning, much like basic elements join together to form whole molecules. The feature comparison view suggests that there are two different types of semantic features inherent in concepts. The first type is *defining features*, which are in themselves necessary and sufficient for defining the concept: that is, something must have all of the defining features to be considered an example of the concept. The second type is *characteristic features*, which are more optional but which tend to be present in good examples of the concept. For example, the concept *bachelor* has the defining features [unmarried], [male], and [human]. Technically speaking, any person possessing these features might be considered a bachelor including male babies, priests, and centenarians. However, good examples of bachelors also [are marriageable], [are 20–50 years old], and [date women].

Presumably, to decide that something is an example of a concept (such as when one decides whether a robin is a *bird*), the features of the example are compared to those of the concept. If they share most of the features, as in the case of robins and birds, then one can quickly decide that it is an example of the concept. If they share few features (such as when one compares a brick to a bird), then one can quickly decide that it is not an example of the concept. However, occasionally, this match-up is intermediate, such as in deciding whether a bat is a bird. Bats share several features in common with birds such as the fact that they [fly], [have wings], and [have two feet], but they cannot be classified as birds. In such cases, one uses only the defining features to decide whether something is an example of the concept. However, going through these defining features takes time. Therefore, perhaps the best evidence for this model is the fact that people take longer to decide that bats are not birds than to decide that bricks are not birds.

The *analytic prototype view* is similar, but does not make the qualitative distinction between defining and characteristic word features. Instead, word features are said to be merely weighted in their importance. This model of lexical organization suggests that concepts that match the central

tendency of a category are considered prototypes. Concept features, then, are compared to those of the category prototype to determine category membership. Perhaps the best evidence for this view of lexical organization comes from the fact that children learn prototypical prior to atypical examples of a concept (that robins are birds sooner than that ostriches are birds).

THEORY-THEORY VIEW

The *theory-theory view* proposes that representations of concepts are rooted in our knowledge about how the world works. Murphy and Medin have argued that important and coherent concepts in a domain are embedded in deeper, underlying theories that people have about the domain. These theories determine the features that are used to distinguish one concept from another. Theories also serve to relate concepts in the domain to each other. They determine which features are key to the domain.

Thus, most people know that the feature [burnable] refers to both *money* and *wood*, but the feature would appear in our representation of wood only because of its use as a fuel. [Burnable] plays no similar role in our concept of money, and, in fact, other features are more important to the role that money plays in our lives. The theory surrounding [burnable] would then relate wood to other fuels such as gasoline and other items for which its ability to be burned is a key feature.

This view extends feature theories by providing a rationale for which features become attended to, how they derive their importance and are combined to produce coherent concepts. Perhaps the best evidence for this view is the finding that a developmental change in the theory underlying a domain is accompanied by a reorganization of and appearance of new concepts in the domain. For example, when children come to view the mind as intentional, effortful, and constructive, the concepts of *divided attention* and *inhibition* emerge (which are effortful subtypes of attention), and they reorganize other attention processes according to the effort they entail.

NETWORK THEORIES

The *spreading activation network view* is qualitatively different from the feature comparison, analytic prototype, and theory-theory views of lexical organization. This view of word meaning depicts words as nodes connected to other concepts through association. Nodes are merely abstract representations of words that serve to identify the location of the word in memory. Word meaning is said to be the sum of concepts with which a given word is associated. The key difference between this and feature theories is that meaning is represented by associative links between nodes representing the relationship between words (such as *is a* and *has*) and not overlapping word features, as in the feature theories.

A second aspect of this view is that this interlinked network is designed to permit one concept to activate others associated with it. In this view, hearing or seeing the word 'robin' would activate the node for *robin* in one's memory. The activation of this node would then send activation along the associative links to other concepts connected with it such as *bird*, *feathers*, etc. This process is called *spreading activation*. In this view, the length or strength of the links between nodes represents the degree of association between concepts. Shorter or stronger links represent greater association and faster spreading activation between concepts. For example, the link between *robin* and *bird* is probably stronger than the link between *ostrich* and *bird*. Because spreading activation moves across strong links faster than weak ones, it is faster to decide that a robin is a bird than that an ostrich is.

Perhaps the best evidence for the spreading activation view is the finding of *mediated association priming*. The spreading activation model assumes that when words are activated, activation spreads rather far and wide to related concepts and their associates. For example, *lion* and *tiger* are both associatively linked. *Tiger* and *stripes* are also associatively linked. However, *lion* is associatively linked to *stripes* only through *tiger*. *Stripes* can be activated by *lion* only if *lion* activates *tiger* first. It has been shown that people are faster at recognizing the word *stripes* when preceded by *lion* than when preceded by some completely unassociated word, suggesting that activation spreads through mediating links from *lion* to *tiger* to *stripes*.

Distributed connectionist models of the lexicon view the representation of concepts as comprising subsymbolic nodes rather than word nodes. These subsymbolic nodes are said to work together like computer bits that are turned either on or off. The lexicon is said to contain different types of nodes that correspond to the letter features (orthographic units), sound features (phonological units), and word-meaning features (semantic units) of words. Like the spreading activation network view, these node types are interlinked in the network with varying strengths between their

connections. Unlike spreading activation views of word meaning, these aspects of lexical representation are thought to be distributed throughout the memory system, rather than localized with specific information around a word node itself. In fact, the concept of a *word* in these models is merely the pattern of activation of units that are activated when a particular word is presented. With each presentation of a new word comes a change in the strength of the links connecting the orthographic, phonological, and semantic feature units for that word, making that pattern of units more easily activated the next time the word is presented.

This kind of system is said to be dynamic. That is, partial activation of one type of node causes the resulting partial activation of another type of node. For example, when a person is presented with a written word such as 'girl', the orthographic units *g*, *i*, *r*, and *l* become partially activated and quickly begin to send activation corresponding to both the phonological units (*g*, *er*, *l*) and semantic units (*female*, *human*, *young*) associated with the word pattern. When the phonological and meaning units become partially activated, they each send activation back to the other two types of units, combining their influences on each other until the person recognizes the word. In this model, related words (such as *girl–boy*) tend to benefit the processing of each other because the semantic units of the two words have similar patterns of activation (that is, both words activate *human* and *young*). After reading or hearing the first of two related words, the shared meaning units of the two words remain activated upon presentation of the second word. This enables the pre-activated meaning units of the second word to send activation to the phonological and orthographic units, so that the person recognizes the word quickly.

Perhaps the best evidence for this kind of model comes from experiments where a single unrelated word is interspersed between two related words (*girl–book–boy*) and the person has to name the final word. This model predicts that, while reading these three words, the meaning units associated with *girl* will be activated at first, but then become deactivated after reading the second word *book*. Consequently, the meaning units of *boy* will have to become reactivated anew when *boy* is presented for naming, eliminating any benefit from meaning-unit activation from *girl*. In fact, studies show that presenting an unrelated word between two related words disrupts the processing benefits normally seen between related words.

FRAME (SCHEMA) VIEWS

Frame (sometimes called *schema*) views describe word meaning as comprising 'slots' and 'fillers' organized to represent basic regularities and expectations that people have for concepts. Slots are a kind of variable that needs to be filled; fillers provide value to the slots. For example, the word 'goldfish' might be represented with slots that represent basic properties of goldfish such as *habitat* and *eats*, which might be typically filled with the concepts *aquarium* and *fish food*, respectively. Such typical slot fillers are said to form default values for the concept. However, slots usually have alternatives to these default fillers that may represent the knowledge that, say, *goldfish* might also live in *ponds* and eat *insect larvae*.

The information represented by the frame view differs from the relations represented by the spreading activation model in that they represent a larger number and a more concept-specific set of relations than semantic network models. Further, frames represent more extensive knowledge about a concept because of the representation of typical defaults and other permissible fillers. In this sense, the knowledge contained in frames for a word is encyclopedic in nature, rather than precise or analytic.

One of the strongest arguments for this form of representation comes from the fact that verbs seem to require complex, frame-like knowledge to represent them adequately. For example, any context using the verb *sell* needs to provide some indication of (or filler for) the *seller*, *buyer*, and *goods sold*, suggesting that all three of these are part of the frame for this verb.

SOME IMPORTANT PROBLEMS FOR THE PSYCHOLOGY OF WORD MEANING

One interesting problem for the psychology of word meaning is *polysemy*. Polysemy refers to the issue that two words that mean something quite different can sometimes sound and look the same. For example, the word 'duck' is polysemous because it describes both the bird and the act of lowering quickly to evade something. It is generally agreed that we possess two different semantic representations for such words to which the orthographic and phonological features are attached.

One intriguing question is whether we always retrieve both meanings for ambiguous words when we read or hear them. For example, if we heard the sentence 'To avoid the tree branch, Carlos

had to duck', would we actually retrieve the *bird* meaning as well as the *lowering* meaning? Most researchers think that for a very short time after reading or hearing an ambiguous word (around 300 milliseconds), we actually retrieve both meanings as long as both meanings are reasonably common interpretations for the word. After this short time, we quickly select the meaning that makes the most sense in the context and lose awareness that the alternative meaning was ever activated.

A second problem for the psychology of word meaning is how words are represented in *bilinguals*. The bilingual lexicon is interesting because it requires a distinction between the actual words from each language and the concepts with which they are associated. One way the concepts might be represented is to assume that there is an associative link between words in a bilingual's first language and those of his or her second language. To figure out the meaning of a word in a second language, one would first have to translate a second language word into a corresponding word in one's first language and then retrieve that word's meaning. Another way the concepts might be represented is to assume that words in each language are indirectly connected to each other through the common meanings that they share. Word meaning would be accessed directly from words of each language, and there would be no need to translate first. In fact, translation from one language to the other would be performed by going through the word meaning system first. There is some evidence that beginning bilinguals might operate according to the word association depiction and advanced bilinguals according to the concept mediation view, but this has not yet been fully determined.

Further Reading

Anglin JM (1977) *Word, Object, and Conceptual Development*. New York, NY: Norton.

Collins AM and Loftus EF (1975) A spreading activation theory of semantic processing. *Psychological Review* **82**: 407–428.

Fillmore C and Atkins BT (1992) Toward a frame-based lexicon: the semantics of RISK and its neighbors. In: Lehrer A and Kittay EF (eds) *Frames, Fields, and Contrasts*, pp. 75–102. Hillsdale, NJ: Lawrence Erlbaum.

Kroll JF and deGroot AMB (1997) Lexical and conceptual memory in the bilingual: mapping form to meaning in two languages. In: deGroot AMB and Kroll JF (eds) *Tutorials in Bilingualism: Psycholinguistic Perspectives*, pp. 169–199. Mahwah, NJ: Lawrence Erlbaum.

Masson MEJ (1995) A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory and Cognition* **21**: 3–23.

McNamara TP (1992) Theories of priming, I: Associative distance and lag. *Journal of Experimental Psychology: Learning. Memory and Cognition* **18**: 1173–1191.

Murphy GL and Medin DL (1985) The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316.

Nagy WE and Herman PA (1987) Breadth and depth of vocabulary knowledge: implications for acquisition and instruction. In: McKeown MG and Curtis ME (eds) *The Nature of Vocabulary Acquisition*, pp. 19–35. Hillsdale, NJ: Lawrence Erlbaum.

Parault SJ and Schwanenflugel PJ (2000) The development of conceptual categories of attention during the elementary school years. *Journal of Experimental Child Psychology* **75**: 246–262.

Rosch E and Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605.

Smith EE, Shoben EJ and Rips LJ (1974) Structure and process in semantic memory: a featural model for semantic decision. *Psychological Review* **81**: 214–241.

Tanenhaus MK, Leiman JM and Seidenberg MS (1979) Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior* **17**: 143–154.

Word Recognition

Introductory article

Jonathan Grainger, University of Provence, Aix-en-Provence, France

CONTENTS

Psychological research on visual word recognition aims to describe the information processing involved in identifying a given string of letters as a specific word during the reading process.

INTRODUCTION

Visual word recognition is one of the most highly researched areas of cognitive psychology. Researchers in this field apply the information processing paradigm adopted by cognitive psychologists since the 1950s: this paradigm is used to understand how the light pattern reflected by a written word is transformed from activity in the retinas of our eyes to meaning in our heads, by specifying the mental representations involved, and the processes that operate on these. Although strictly speaking the study of visual word recognition is concerned only with the extraction of meaning from strings of letters (i.e. silent reading), the term is often extended to include the act of reading aloud (i.e. how a specific pronunciation is assigned to a given string of letters). The following discussion is limited to the case of silent reading of isolated words, and is also restricted to alphabetic writing systems (as opposed to non-alphabetic languages such as Chinese); it is based on research performed primarily in English. (*See* **Lexicon, Computational Models of**; **Language Comprehension**; **Reading, Psychology of**)

PRINTED WORDS

Written words come in all shapes and sizes, yet skilled readers manage to abstract away from such variation in surface form (e.g. word, WORD, wOrD) to reach the common underlying meaning. This skill could be analogous to our ability to recognize other familiar visual patterns, such as everyday objects. However, printed words are very special visual patterns composed of a small number of clearly identifiable subunits, their component letters. Unlike spoken language, printed alphabetic script marks word boundaries plus the boundaries of individual letters in the word. Another important property of printed words in alphabetic languages is that the individual letters are linked to a word's component sounds in a nonarbitrary way.

METHODS

Following the general methods of experimental psychology, researchers vary certain properties of the stimulus while recording the responses of human participants who are requested to react in some way to these stimuli. The way the participants' performance varies systematically as a function of variations in stimulus properties is used to infer the information processing that would generate the observed response pattern. In studying visual word recognition, various properties of printed words are manipulated, and the selected sample of words presented to participants who are asked to perform a given task.

Lexical Decision

In the standard lexical decision task, participants see a random mixture of word and nonword stimuli one at a time on a computer screen, and must press one of two response keys as fast as possible to say whether the stimulus is a word or not. The time between stimulus onset and the participant's response is measured (response time, RT) as is the number of errors made.

Perceptual Identification

In contrast to the lexical decision task, perceptual identification tasks use impoverished stimulus presentation conditions in order to make words harder

to recognize than usual. In this type of task it is the percentage of correct identifications that is the measure of interest. In some variants the stimulus gradually becomes visible over a short period (typically 1–2 s), and the time taken to recognize the word is recorded. This has the advantage of providing a measure of performance (RT) on each trial rather than having performance averaged over several trials.

Semantic Categorization

In semantic categorization participants are asked to classify words as belonging to a predefined semantic category or not. These categories can either be general (e.g. animate versus inanimate) or more specific (e.g. flower, body part). After being informed of the target category, participants are presented with a series of words, to which they must respond as rapidly as possible, indicating which category the word belongs to.

Priming

The priming technique involves presentation of two successive stimuli, the prime and the target. The goal is to see whether performance on a given target word varies as a function of the type of prime stimulus that precedes the target. Will the word 'butter' be recognized more quickly when preceded by the prime 'bread' than when it is preceded by 'train'? This particular question addresses the issue of semantic priming, but the same technique can be used to examine all possible relations that exist across words of a given language. In a variant of the priming paradigm, prime stimuli are presented so briefly that they cannot be identified by the participants ('subliminal' priming). The influence of such subliminal primes on target word recognition is thought to reflect the operation of fast, automatic and largely unconscious processes engaged in identifying printed words.

BASIC PHENOMENA

Research on visual word recognition provides an example of science at its best. Phenomena in this research area have been clearly defined, and hypotheses have been developed to explain them and predict the existence of new ones. Here we will examine some of the key phenomena, from low-level visual and orthographic processing to higher-level lexical and semantic operations.

Orthographic Processing

The identities of a word's component letters must be coded for successful word recognition. Evidence for the role of individual letter representations has been provided by orthographic priming studies that vary the number of letters shared by prime and target, all other factors being held equal. These orthographic priming effects occur over and above effects of visual similarity, being unaffected by changes in case and type font. However, letter identity information is a necessary but not a sufficient condition for correct word identification. One also needs to know the position of these letters. Current evidence suggests that the initial coding of letter position is somewhat approximate, providing only relative position information (e.g. the letter 'b' in 'table' is somewhere in the middle, just after 'a' and just before 'l') and not absolute position information (e.g. the letter 'b' in 'table' is the third letter in a string of five letters).

Orthographic processing is complicated by the fact that on any given eye fixation in a word, letter visibility varies as a function of letter position relative to the fixation point (higher visibility closer to fixation), and whether the letter is the first, last or a middle letter (better visibility for external letters). These variations in letter visibility determine how accurately a word can be identified at a single glance. As eye fixation in the word shifts from the first to the last letter one obtains an inverted U-shaped function of word recognition accuracy. However, the function is not symmetric, with a definite advantage for fixations to the left of the center of the word (at least for languages that are read from left to right). This asymmetry is also evident with lateralized presentation of stimuli. Words presented to the right of fixation are easier to recognize than words presented to the left. This lateralization effect is thought to be due to the fact that brain areas involved in word recognition are located in the left hemisphere of the brain, and this hemisphere receives information directly from the right visual field.

Phonological Processing

A key finding in word recognition research is that a word's phonology (its component sounds) influences the recognition process. For example, when presented with the target word 'make' in a masked priming experiment, participants found it easier to recognize following the subliminal prime 'mayk' (a nonword called a pseudohomophone because it

can be pronounced like a real word) than following the prime 'malk'. Our ability to recognize target words is influenced by the amount of phonology they share with prime stimuli over and above any orthographic overlap. This phonological priming arises about 30 ms after orthographic priming, showing how fast the brain has been able to translate orthographic information into a phonological code. Phonological influences have also been demonstrated in semantic categorization tasks without priming. It is harder for participants to say that 'rows' is not a flower than to respond negatively to the word 'robs'. Indeed, homophony (when different words share the same pronunciation) has a systematically negative influence on the process of printed word perception.

Lexical Processing

The extraction of orthographic and phonological information from a string of letters eventually allows the human brain to identify the stimulus as a known word and to recover the meaning associated with that particular form. The speed with which this occurs depends on that word's frequency in a given language. Word frequency is measured by counting the number of times a given word occurs in large corpuses of text. Thus, in one particular corpus of the English language, the word 'table' occurs 198 times per million words, while the word 'tablet' occurs 3 times per million. Although this measure of printed word frequency is often partially confounded with other possible measures of word usage, such as the age at which a word is learned (age of acquisition), it is generally assumed that it exerts the most significant independent contribution to word recognition fluency. Word frequency can be situated on a general continuum of stimulus familiarity, with unfamiliar nonword stimuli situated at one extreme. The familiarity of letter strings determines how easy it is to recognize the component letters of these stimuli. Letters in words are easier to identify than letters in nonwords (the 'word superiority' effect).

Semantic Processing

The priming paradigm has been extensively applied to investigate semantic processing in visual word recognition. Three kinds of semantic priming effect have been isolated: categorical, associative, and similarity-based priming. Categorical relations are defined *a priori* by taxonomists. Associative and semantic similarity relations are obtained by normative measures. When asked to provide the next word that comes to mind on hearing the word 'table', the majority of native speakers of English will reply 'chair'. The associative norms derived from such a procedure are then used to create associatively related primes and targets in a semantic priming experiment. Semantic similarity measures are obtained by rating judgments – how semantically similar are these two words? All these measures of semantic relatedness have been shown to facilitate target processing in semantic priming experiments.

MODELS

Faced with these various phenomena to explain, word recognition theorists have come up with a number of different approaches. One early account considered visual word recognition as analogous to how we look up a new word in a dictionary. The idea is that we have a mental dictionary (mental lexicon) listing all the words we know, and given some minimal information about the stimulus (i.e. a few letter identities) we can guide our search to the appropriate section of the dictionary and begin a detailed verification process comparing information extracted from the stimulus with information stored in memory.

More recent approaches have abandoned the notion of search for the concept of activation (or parallel information accumulation). On presentation of a printed word, orthographic, phonologic, and semantic codes are activated in an interconnected network of simple processing units. Activation builds up in the network until a stable state is achieved and the system has settled on a unique combination of orthography, phonology, and semantics. Much modeling work today is concerned with describing how the brain has learned to map orthographic codes onto phonological codes on the one hand, and onto semantic codes on the other. There is increasing evidence that such mapping occurs in a highly interactive system: orthographic codes send activation onto phonological and semantic codes which mutually activate each other, and send activation back to the orthographic level.

CROSS-LINGUISTIC RESEARCH

Languages vary in the type of visual code they use to represent speech. Compare, for example, the logographic code of Chinese with the alphabetic code of English. Cross-linguistic research in visual word recognition has also revealed interesting

differences across languages that share the same alphabet. One such difference is the way in which individual letters are associated with the component sounds of a given language. The amount of ambiguity in the mapping of orthography onto phonology and vice versa (one-to-one, one-to-many, many-to-one) varies considerably across languages. This has a major impact on the speed with which children learn to read in these different languages, and subsequently influences the nature of the sublexical representations involved in printed word perception. Degree of consistency in the mapping of orthography to phonology is also thought to determine the division of labour across an orthographic–semantic processing route and an orthographic–phonological–semantic route in silent reading.

Finally, individuals vary in terms of the number of languages that they know, and this affects processing within each language. Current evidence suggests that people who are bilingual cannot 'switch off' the irrelevant language (when reading a book in one language, for example). There have been a number of reports of cross-language interference during word recognition in bilingual participants. The amount of interference depends on how similar the target word is to words of the other language. Thus words from different languages interact in a way that mimics within-language interactions. Nevertheless, bilingual people can, to some extent, control these between-language interactions by use of knowledge about which language each specific word belongs to.

Further Reading

Balota D (1994) Visual word recognition: the journey from features to meaning. In: Gernsbacher MA (ed.) *The Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Besner D and Humphreys G (1991) *Basic Processes in Reading: Visual Word Recognition*. Hillsdale, NJ: Erlbaum.

Frost R and Katz L (1992) *Orthography, Phonology, Morphology, and Meaning*. Amsterdam, Netherlands: North-Holland.

Grainger J and Dijkstra T (1996) Visual word recognition. In: Dijkstra T and De Smedt K (eds) *Computational Psycholinguistics: Symbolic and Connectionist Models of Language Processing*. Hemel Hempstead, UK: Harvester Wheatsheaf.

Harley T (2001) *The Psychology of Language*, 2nd edn. Hove, UK: Psychology Press.

Rayner K and Pollatsek A (1989) *The Psychology of Reading*, chap. 3. Englewood Cliffs, NJ: Prentice-Hall.

Working Memory

Intermediate article

Barbara A Dosher, University of California, Irvine, California, USA

CONTENTS

Working memory is the capacity to manipulate and maintain information over short periods (2–15 seconds) in order to support simple memory tasks such as remembering a telephone number, or more general cognitive tasks such as problem solving, simple reasoning, or reading. Working memory consists of several distinguishable memory capacities, together with executive functions that manage information retrieval, reactivation, and transformation.

THE ROLE OF WORKING MEMORY IN INFORMATION PROCESSING

Human memory – that is, the retention of information over short or long delays – is critically important for cognitive function. *Working memory* is the capacity to manipulate and maintain information over short temporal delays. Working memory supports cognitive activities such as solving mathematical problems, evaluating spatial layouts, or comprehending sentences, and it may be critical in language learning. It is one of several closely related short-duration memory systems. Working memory is of longer duration than the relatively unprocessed and very-short-duration (< 2 seconds) visual sensory memory (iconic memory) and auditory sensory memory (echoic memory). It maintains information over a period of several seconds or longer through proactive control and rehearsal mechanisms. Thus working memory goes beyond the short-term or immediate memory function that briefly maintains information to include attention and information management – the capacity for manipulation and transformation of information that is required during cognitive tasks.

WORKING MEMORY FRAMEWORK

According to one classic view, working memory is a system of three interacting modules (Baddeley, 1986). These include an *executive control module*, a *visual–spatial sketchpad*, and an *articulatory loop* (Figure 1). Evidence for the separable visual and verbal modules arises from the differential impact of visual distraction on memory for visual information, and the differential impact of verbal distraction on memory for verbal information. The executive control module manages task goals, manipulation of information, and complex rehearsal. The visual–spatial sketchpad maintains visual information (visual content) and spatial information (spatial layout). The articulatory loop maintains temporally ordered verbal information. The sketchpad and the loop are subservient to the executive control module.

The modular working memory framework was originally developed partly to account for capacity limitations in verbal tasks that require correctly ordered recall. In verbal recall tasks, a written or spoken list of verbal items is presented in temporal sequence, and the list is then recalled in order. Several phenomena in serial ordered recall were critical in the development of the modular working memory framework. First, lists of phonologically more dissimilar or distinct items are recalled more accurately than lists of phonologically similar items. This *phonological similarity effect* implies that a phonological or articulatory representation is dominant in working memory. Secondly, the number of items that can be recalled depends upon word length. Longer lists (up to 7–9 items) can be correctly recalled if those items are 'short', with a smaller number of syllables or a shorter time required to pronounce them, but recall is reduced for lists of items that are 'long' (a *word-length effect*). The sensitivity of working memory performance to word length suggested the existence of a cyclic rehearsal loop in which items were refreshed or rehearsed in turn in order to maintain memory. This emphasizes the importance of control processes. Thirdly, the repeated speaking of a distracting word or phrase during the memory task dramatically reduces the length of list that can be

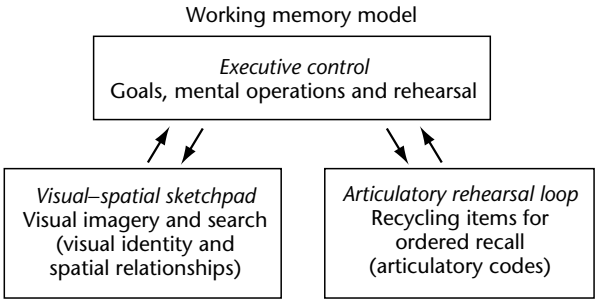


Figure 1. The modular working memory model (Baddeley, 1986) consists of three modules. The executive control module manages rehearsal and other executive functions to transform and maintain information in working memory. The visual–spatial sketchpad maintains visual (content) and spatial (layout) information for forms, patterns, and faces. The articulatory loop maintains item and order information through verbal, articulatory codes for digits, letters, words, and sentences.

recalled. This is known as the *articulatory suppression effect*. Articulatory suppression also eliminates phonological similarity effects for visual presentation.

These phenomena and their complex interplay suggested the operation of an articulatory/phonological rehearsal mechanism, verbal recoding of visual inputs, and the importance of control or rehearsal activities – all critical concepts in the modular working memory framework. For a review of these and related phenomena see Neath (1998).

Other alternative frameworks and computational models based on quite different principles may also account for a wide range of these data. One alternative view inspired by human cognitive neuroscience and neural network representations is that short-term memory is the currently active subset of long-term memory, and that working memory is the system of short-term memory plus the strategy and attention functions that reactivate fading short-term memory traces (Cowan, 1995). Other formal models have been developed for several specific tasks (Page and Norris, 1998; Henson, 2001), or to specify more precisely the representation of the abstract and modality-specific features of items in working memory (Neath, 1998). As the brain areas relevant to working memory function are identified, and the temporal properties of the responses in these regions are more fully understood, the description of the modular working memory framework may be transformed into a more detailed specification of the computational implementation and the brain circuitry, consisting of processing structures and more complex computational network representations. Nonetheless, the modular working memory architecture (Baddeley, 1986) provides a structural-level description of memory function, and has provided a framework for an extensive body of behavioral and physiological studies of working memory in a range of memory tasks. The important distinctions between classes of working memory tasks will be considered in the following sections.

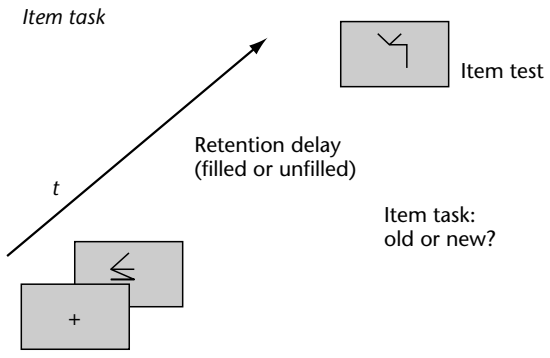
FORMS OF WORKING MEMORY

Working memory operates somewhat differently in tasks with differing complexity or cognitive processing demands. One task that is used extensively in animal models and in physiological analyses requires a single item to be maintained through a delay period. More demanding working memory tasks involve the full recall of a list of items during complex distracting activities (Daneman and Carpenter, 1980). Some tasks focus on visual coding, while others focus on verbal coding. Each form of working memory task places different demands on information retrieval and executive function.

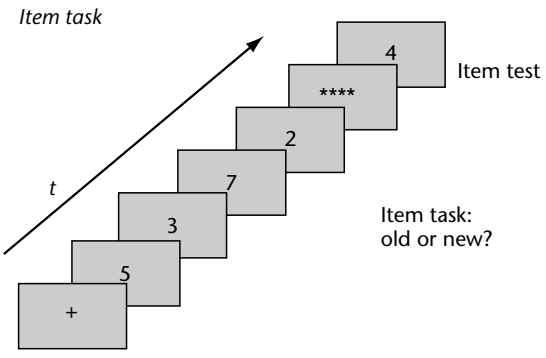
Working memory tasks are of several major types (Figure 2). *Item memory tasks* require the evaluation of an individual item or individual position, although the memory set may consist of multiple items or positions. *Relational memory tasks* require the evaluation of some relational property, such as relative temporal order or the spatial location of several items. Finally, *complex memory tasks* require extended reproduction of a memory set with a sequence of temporally or spatially ordered responses. Item, relational, and complex tasks vary further in difficulty depending on the level of the memory load or of distraction.

Item Working Memory Tasks

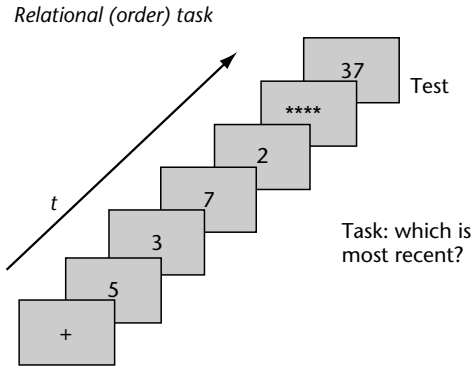
Item working memory tasks test memory for a single item, possibly from among a set of items, following a brief delay. In the simplest item task (Figure 2(a)), recognition memory for a single item is tested after a retention delay, usually of between 1 and 60 seconds (Figure 3(a)). Recognition of the item declines with increasing delay. This task has been important for studying short-term retention in animals (the 'delayed match to sample' task). It has also been important in studies of delay period activity of cortical neurons, and sustained neural activity during the retention interval, that has been associated with working memory (Chafee and Goldman-Rakic, 1998). For humans, delayed



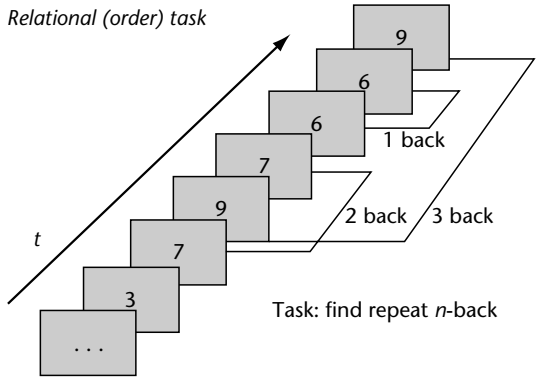
(a) Retention delay task



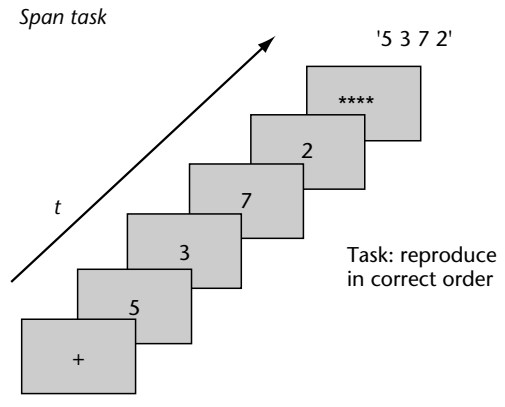
(b) Temporally ordered working memory load



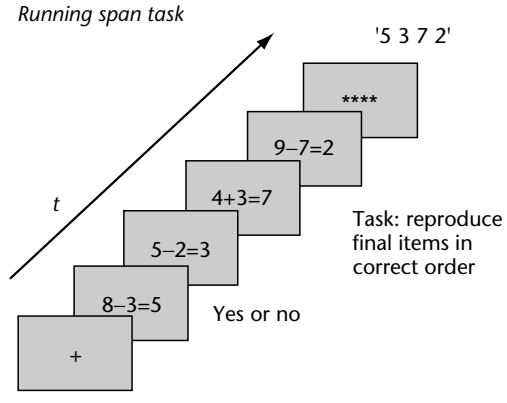
(c) Temporal order task



(d) *n*-back task



(e) Temporally ordered working memory load



(f) Temporally ordered working memory load

Figure 2. Item, relational, and complex working memory tasks are successively more demanding of working memory. (a and b) Examples of item working memory tasks. (a) The delayed recognition task (a simple item working memory task) measures forgetting over short time delays. In this example, the stimuli require visual rather than verbal coding. (b) Item recognition with memory load task requires memory for an individual item from a memory set, or recognition after filled delays. In this example, the stimuli utilize verbal coding. (c and d) Examples of relational (order) working memory tasks. (c) In the temporal order task, the most recent of two items from a working memory set is selected. (d) In the *n*-back task, an item repetition is detected at a specified delay, 1-back (immediate repetition), 2-back (repetition with one intervening item). (e and f) Examples of complex working memory tasks. (e) In memory span tasks, items are recalled in the correct order. (f) In running span tasks, identified items from a series of mental tasks (e.g., simple mathematical problems) are recalled in the correct order. Reproduced with permission of B. Dosher.

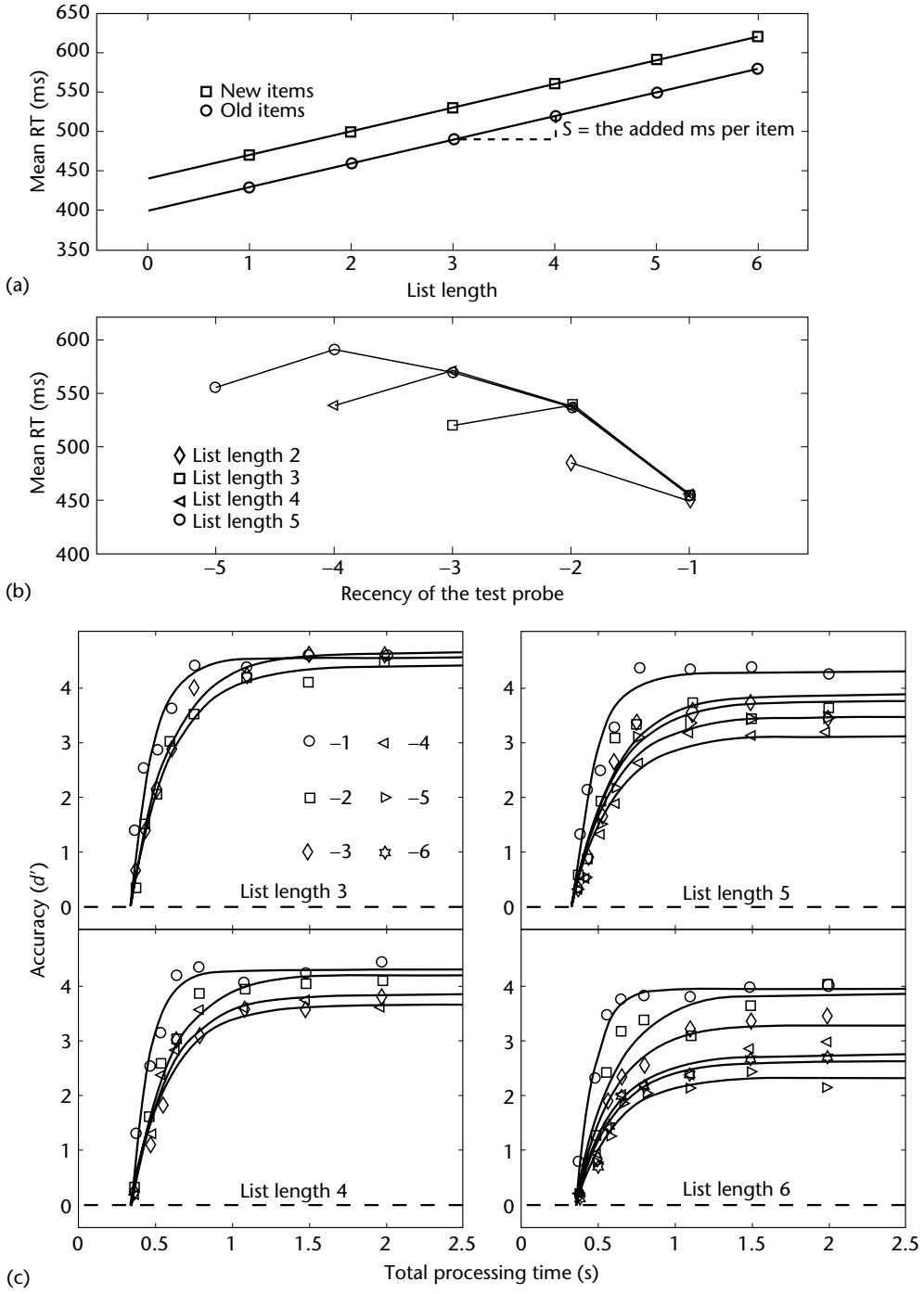


Figure 3. Item tasks are characterized by simple decay functions and load-independent retrieval. Schematic results of a delayed item recognition task are shown. (a) Mean reaction time (RT) to recognize a single item in working memory increases with the memory load or list length. These aggregate results reflect faster response times for more recent list items. (b) Reaction time (and accuracy) of item recognition depends upon recency. Response times are shortest for list-final (most recent) positions, and they increase as items become less recent. (c) Recognition accuracy (bias-free strength d') is measured as a function of retrieval time by interrupting recognition after various amounts of time for working memory loads (list lengths) of from 4 to 6 items (data from McElree and Dosher, 1989). Limiting accuracy is higher for more recent items. However, the time course of item recognition reflects parallel processing – it does not depend on working memory load or item recency. Reproduced with permission of B. Dosher.

recognition of a single study item may be too simple to tax working memory, especially for simple verbal items, so task complexity is increased by increasing the memory load, or the number of items to be remembered. Alternatively, the complexity or demand of the memory load might be increased.

One of the most extensively studied working memory tasks tests recognition of one item from a multi-item memory load presented over time (Figure 2(b)). Item recognition takes place immediately following a study list when error rates are relatively low, and response time is used to index memory. The memory load may consist of new items, in which case the judgment may reflect familiarity, or it may consist of highly repeated items from a small set, in which case the judgment must reflect not familiarity with the items themselves, but specific list membership. The average time taken to recognize an item as a member of the memory load increases approximately linearly with the load size (Figure 3(a)), which has often been claimed to reflect a serial process in retrieval from working memory (Sternberg, 1975). Recent evidence instead suggests that in simple item recognition tests, familiarity or activation is processed fairly directly, with accuracy and activation times reflecting item recency (the delay interval). Memory lists of different sizes reflect different mixtures of recency effects on both recognition time and accuracy (Figure 3(b)). In fact, under many conditions, retrieval of single items from working memory is achieved with the same time course regardless of the size of the working memory load or the study position of the item tested for recognition (Figure 3(c)). Information starts to become available at the same time and increases at the same rate independent of working memory load or temporal position (Dosher and McElree, 1992). The slower response times for larger working memory loads reflect lower memory strength for less recent items.

Simple, relatively rapid mnemonic evaluation of single items, independent of memory load, is a characteristic of item working memory. Item working memory tasks may exhibit parallel retrieval independent of the memory load. In contrast, the patterns of retrieval in relational working memory tasks exhibit substantial load effects and slower and load-dependent retrieval.

Relational Working Memory Tasks

Relational working memory tasks involve relational judgments, such as judgments of temporal or spatial order, over two or more items. These judgments place greater demands upon retrieval and rehearsal than do simple item tasks. For example, the relative order task (Figure 2(c)) requires the selection of the most recent of two test items. In comparison with the relatively rapid and load-independent retrieval exhibited for item tasks, processing of the relative temporal order (recency) is considerably slower than item retrieval, and is strongly dependent on the temporal positions of the items. The time course of judgments is rapid for memory probes including the most recent item, perhaps comparable to that of retrieval of single-item information. However, the time course of temporal order judgments is successively slower for probes with items experienced less recently, consistent with a sequential recovery of less recent items. Order judgments induce a more effortful, time-consuming, and serial process of retrieval (McElree and Dosher, 1993) (Figure 4).

Another form of relational, or order, working memory task that has been extensively studied is the *n*-back task (Figure 2(d)), which is a continuous performance task in which items repeated at a particular delay are detected. In a 1-back condition, immediate repetitions are detected, while in a 2-back condition, repetitions with one intervening item are detected (Cohen *et al*., 1997; McElree, 2001). In related tasks, a stream of successive items is inspected for a particular ordered sequence, such as '472'. Like other order judgment tasks, the *n*-back tasks require the maintenance, retrieval, and processing of order information as well as rehearsal and information manipulation to continuously update the task-relevant working memory set. The processing demands in these tasks have also been associated with attention (Engle *et al*., 1999). The *n*-back order tasks show load-dependent activity in functional brain activation, as studied by functional magnetic resonance imaging (fMRI) (Cohen *et al*., 1997), reflecting these higher retrieval and rehearsal demands.

In contrast to item working memory tasks, even the simplest relational or order tasks place increased demands on retrieval processes, with relational judgments often reflecting serial processing. More demanding forms of the relational or order tasks, such as the *n*-back tasks, also require extensive manipulation of the memory set ('relational plus' tasks), or a high demand on executive function.

Complex Working Memory Tasks

Complex working memory tasks involve sequences of responses. The complex tasks, like relational or order tasks, emphasize temporal order or spatial

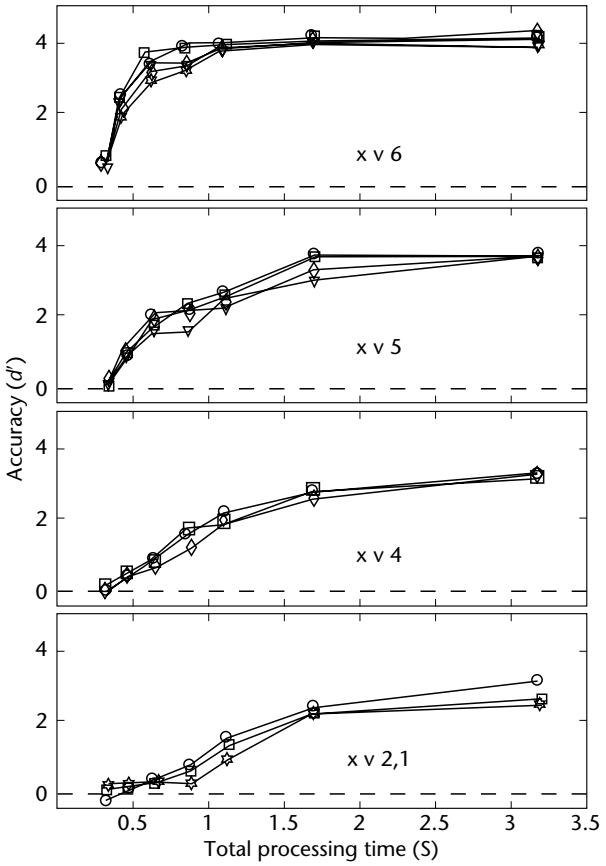


Figure 4. A relational task (judgment of recency) (Figure 2(c)) requires slow and serial processing. The task is to decide which of two item probes is the most recent in the working memory load of six items. Accuracy (d') is expressed as a function of processing time for conditions grouped by the most recent item. The time course of relational judgment depends on the recency of the most recent item, reflecting a serial retrieval process. The upper panel shows fast retrieval for all tests involving the last (sixth) item (1v6, 2v6, etc.). The other panels show successively slower retrieval for tests in which the second to last (fifth) item was most recent, and the middle (third) item was most recent (data from McElree and Dosher, 1993). Reproduced with permission of B. Dosher.

order of the memory set in the response sequence. Among the most classical working memory tasks is the span task (Figure 2(e)). In this task, a working memory load of digits, letters, or words is presented, usually in temporal sequence, and is then recalled under instructions to reproduce the list exactly (serial ordered recall). The more demanding form, known as 'running span', interleaves processing distraction with the input of the working memory set (Figure 2(f)).

As was mentioned earlier, the modular view of working memory was developed and tested using the ordered word span task. One key observation was that the time taken to pronounce words determined the number of words that were successfully recalled in the correct order (the word-length effect) (Baddeley *et al*., 1975). Based on these observations, the duration of working memory was estimated to be approximately 1.5–2 seconds – the time required to pronounce a list of span length (that length of list which is perfectly recalled on average half the time). The idea is that if (subvocal) rehearsal can be completed before item traces are lost from working memory, then the list could be perfectly maintained (Schweickert and Boruff, 1986). One problem with this account is that the actual time taken to recall a span-length list may be 5–8 seconds, which is two to three times the estimate of 1.5–2 seconds (Dosher and Ma, 1998). Either the verbal memory trace is maintained in a phonological loop over intervals of up to 8 seconds, or complex attention strategies maintain items in working memory during the extended act of recall. In common with the relational or order judgment tasks, the simple span tasks require the maintenance and retrieval of order information in working memory. In addition, these complex tasks require the coordination of a succession of recall products, and may involve complex rehearsal and information management.

Complex memory span (Figure 2(f)) makes the same demands as simple serial ordered recall, but each item in the working memory set is separated by a distracting task (Daneman and Carpenter, 1980). Perhaps because these tasks make extremely high demands on mental manipulation and transformation of information, they have been especially important in demonstrating the interrelationship between working memory capacity or efficiency in individuals and the general cognitive functioning of those individuals (Engle *et al*., 1999).

MODELS OF SERIAL MEMORY

Performance in serial memory tasks is characterized by a large and systematic set of phenomena. This does not just include the word-length effect (in which fewer longer words than shorter words can be remembered) and the reduction of successful recall under articulatory suppression (repetition of a distracting phrase) discussed in relation to the modular working memory framework (Baddeley, 1986). Additional phenomena include better memory for spoken than for written presentation, error rates that increase with serial position, and the dominance of transposition (order inversion) errors, as well as many others.

A number of precise computational models have been developed to account for performance in serial memory tasks. Different models are based, in some cases, on quite different principles and assumptions about representation, the causes of information loss, and information retrieval. Three examples of specific mechanistic accounts of working memory in sequential recall tasks are the feature model of Neath and Nairne (Neath, 1998), the primacy model of Page and Norris (1998), and the distributed associative memory model of Lewandowsky and Murdock (1989). This list is by no means exhaustive, but it is selected to illustrate the range of possible representations and processes in models of working memory. Future research must link these representational and process assumptions to the consequences of brain function and coding (for a network example, see O'Reilly *et al*., 1999).

In the feature model, items are represented by collections of modality-independent and modality-dependent features, with the items that are represented stored in order of input as a linked list. As each new item is stored, it over-writes similar features of the immediately previous item or items. In ordered recall, each item is recovered in order from an incomplete or noisy representation. Spoken presentation leads to the storage of a larger number of modality-specific features than visual presentation. Auditory similarity reduces memory due to over-writing of similar auditory features, while articulatory suppression over-writes verbal modality-specific features. Word length has an effect on performance because long words are stored in a sequence of syllabic structures. In contrast, in the primacy model, order is recoded as strength – items are stored in memory with higher activation strength for earlier list items. These strengths then undergo forgetting, or loss. At the time of recall, items are produced in the (noisy) order of strength. Rehearsal resets strength. Longer words take longer to produce or rehearse and experience more forgetting. Finally, in the associative model, individual items and associations between temporally adjacent items are stored in a single composite memory. Items, associations, and the composite memory are represented as a vector of feature values. Recall occurs by recovering successive items from the composite trace through chaining of recovered representations. Similarity between items may increase the error rates due to confusion between similar traces. Although each of the models fails to account for some aspect of the data, each of them also provides a good account of many of the phenomena in working memory that are exhibited in serial recall tasks.

As is evident, each of these models assumes a quite different form of representation and makes quite different assumptions about the nature of memory. Yet each of the models incorporates a mechanism for sequential readout of items from memory, and a mechanism of over-writing or interference to account for forgetting. As these and other quantitative models are developed further, they should provide an increasingly accurate explanation of the maintenance functions of working memory. Precise structural and functional descriptions will complement information about biological implementation.

BIOLOGICAL SUBSTRATES OF WORKING MEMORY

Working memory functions are now thought to reflect activity in a network of brain regions, including regions of the prefrontal cortex and more posterior regions in association cortex (Petrides, 1994). Important brain regions have been identified using a range of methods. Cellular recording of activity in non-human primates during various delayed response tasks has identified the continued activation during the delay (retention) period in certain prefrontal areas with storage of that information (Fuster, 1973). The time course of the delay activity has been shown to depend upon both response accuracy and the duration of the retention interval. Although initial demonstrations demanded spatial memory in responses, delay-period activity in the prefrontal cortex has also been demonstrated in a variety of nonspatial working memory tasks.

Regions of the prefrontal cortex have also been identified in a variety of brain imaging studies in humans (Cohen *et al*., 1997). However, the majority of these studies involved more complex forms of working memory tasks, such as the *n*-back tasks, which make extensive demands upon manipulation and transformation of the information to be remembered – executive function or control architectures. Some researchers have argued that activation in the dorsolateral prefrontal cortex reflects manipulation and transformation, rather than storage or retention (Owen *et al*., 1998). Some evidence from early lesion studies of patients with specific disorders of working memory and also from imaging studies has identified regions of posterior parietal cortex with simple registration and storage. The registration, storage, retrieval, and manipulation functions have yet to be segregated in analyses of working memory function as measured by brain imaging in humans. This will require the

measurement of brain activation with reasonable temporal resolution. In particular, this will necessitate the examination of tasks with clearly temporally segregated retention periods, and the comparison of different tasks with different demands for information manipulation. It will also require the measurement of brain activity in each of the major classes of working memory tasks (Figure 2). Critical new evidence segregating the activation patterns of these brain regions during specific time intervals, isolating storage, manipulation and retrieval in each of these task types, should be available within the next few years with improved methods and technology.

WORKING MEMORY, SHORT-TERM MEMORY, ATTENTION, AND EXECUTIVE FUNCTION

Understanding the functions and interrelationships between modules in working memory is necessary to a full understanding of the brain mechanisms involved, and to the construction of a complete information-processing model. Some theorists (Cowan, 1995) define working memory as the functional interrelationship between short-term memory, processes of attention that reactivate the short-term memory set, and executive function related to transformation, manipulation, and strategy selection. A substantive task analysis is necessary for the segregation of these functions in behavioral analysis as well as in the analysis of brain activation. Future research instantiating working memory tasks within models of perceptual motor limitations, memory stores, and executive function (O'Reilly *et al*., 1999) may provide a precise theoretical structure for distinguishing these subfunctions of working memory.

WORKING MEMORY AND GENERAL COGNITIVE PERFORMANCE

The importance of working memory in general cognitive function is directly supported by the relationship between measures of working memory and performance in a variety of cognitive tasks (Baddeley, 1986). Measures of working memory retrieval time (Figure 3) have been correlated with general cognitive indices such as aptitude scores (Engle *et al*., 1999), and have been shown to differ for different developmental and other populations (Sternberg, 1975). The simpler modules of working memory, such as the phonological loop, appear to play a vital and quite specific role in developmental functions such as language learning (Baddeley *et al*., 1998), where the ability to repeat and maintain a phonological representation may be especially important in learning new words during development, or in learning new languages when supportive semantic information may be unavailable. Performance on working memory tasks, especially complex working memory tasks such as running span, with high demands on executive function (Figure 2(f)), is also correlated with performance on a wide range of other tasks, such as reading or problem solving. Thus, executive functions of working memory may be especially relevant for general cognitive function.

Working memory serves a basic human intellectual function. Elucidating the behavioral function of working memory, the brain activity that supports working memory function, and its relationship to general cognitive function represents a central component of understanding intelligent human activity.

References

Baddeley AD (1986) *Working Memory*. Oxford, UK: Oxford University Press.

Baddeley AD, Gathercole S and Papagno C (1998) The phonological loop as a language learning device. *Psychological Review* **105**: 158–173.

Baddeley AD, Thomson N and Buchanan M (1975) Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* **14**: 575–589.

Chafee MV and Goldman-Rakic PS (1998) Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. *Journal of Neurophysiology* **79**: 2919–2940.

Cohen JD, Perlstein WM, Braver TS *et al*. (1997) Temporal dynamics of brain activation during a working memory task. *Nature* **386**: 604–608.

Cowan N (1995) *Attention and Memory: an Integrated Framework*. Oxford, UK: Oxford University Press.

Daneman M and Carpenter PA (1980) Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* **19**: 450–466.

Dosher BA and McElree B (1992) Memory search: retrieval processes in short-term and long-term recognition. In: Squire LR (ed.) *Encyclopedia of Learning and Memory*, pp. 398–406. New York, NY: Macmillan.

Dosher B and Ma J-J (1998) Output loss or rehearsal loop? Output time vs. pronunciation time limits in immediate recall in forgetting-matched materials. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 316–335.

Engle RW, Kane M and Tuholski S (1999) Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence and functions of the prefrontal cortex. In: Miyake A and Shah P (eds) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*,

pp. 102–134. New York, NY: Cambridge University Press.
Fuster JM (1973) Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology* **36**: 61–78.
Henson R (2001) Serial order in short-term memory. *Psychologist* **14**: 70–73.
Lewandowsky S and Murdock BB (1989) Memory for serial order. *Psychological Review* **96**: 25–57.
McElree B (2001) Working memory and focal attention. *Journal of Experimental Psychology* **27**: 817–835.
McElree B and Dosher B (1989) Serial position and set size in short-term memory: the time course of recognition. *Journal of Experimental Psychology* **118**: 346–373.
McElree B and Dosher B (1993) Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General* **122**: 291–315.
Neath I (1998) *Human Memory: an Introduction to Research, Data and Theory*. Pacific Grove, CA: Brooks/Cole.
O'Reilly RC, Braver TS and Cohen JD (1999) A biologically based computational model of working memory. In: Miyake A and Shah P (eds) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, pp. 375–411. New York, NY: Cambridge University Press.
Owen AM, Stern CE, Look RB *et al*. (1998) Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. *Proceedings of the National Academy of Sciences of the USA* **95**: 7721–7726.
Page MPA and Norris D (1998) The primacy model: a new model of immediate serial recall. *Psychological Review* **105**: 761–781.
Petrides M (1994) Frontal lobes and working memory: evidence from investigations of the effects of cortical excisions in nonhuman primates. In: Boller F and Grafman J (eds) *Handbook of Neuropsychology*, vol. 9, pp. 59–82. Elsevier.
Schweickert R and Boruff B (1986) Short-term memory capacity: magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory and Cognition* **12**: 419–425.
Sternberg S (1975) Memory scanning: new findings and current controversies. *Quarterly Journal of Experimental Psychology* **27**: 1–32.

Further Reading

Fuster JM (1973) Unit activity in prefrontal cortex during delayed response performance: neuronal correlates of transient memory. *Journal of Neurophysiology* **36**: 61–78.
Hunt E, Fronst N and Lunneborg C (1973) Individual differences in cognition: a new approach to intelligence. In: Bower G (ed.) *Advances in Learning and Motivation*, vol. 7, pp. 87–122. New York, NY: Academic Press.
Jonides J, Smith EE, Koeppe RA *et al*. (1993) Spatial working memory in humans as revealed by PET. *Nature* **363**: 623–625.
McElree B and Dosher B (1989) Serial position and set size in short-term memory: the time course of recognition. *Journal of Experimental Psychology: General* **118**: 346–373.
Mayes AR (1988) *Human Organic Memory Disorders*. Cambridge, UK: Cambridge University Press.
Neath I and Nairne JS (1995) Word-length effects in immediate memory: overwriting trace decay theory. *Psychonomic Bulletin and Review* **2**: 429–441.
O'Reilly RC and Munakata Y (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.
Smith EE, Jonides J, Koeppe RA *et al*. (1995) Spatial versus object working memory: PET investigations. *Journal of Cognitive Neuroscience* **7**: 337–356.